Alexandre Miguel Guerreiro Contreiras

PREDICTIVE ANALYTICS FOR SALES FORECASTING IN SMEs:

A MACHINE LEARNING AND BI INTEGRATION

2024

Alexandre Miguel Guerreiro Contreiras

PREDICTIVE ANALYTICS FOR SALES FORECASTING IN SMEs:

A MACHINE LEARNING AND BI INTEGRATION

MASTERS IN SME MANAGEMENT

Dissertation made under the supervision of:

Ph.D. Célia Maria Quitério Ramos

2024

PREDICTIVE ANALYTICS FOR SALES FORECASTING IN SMEs:

A MACHINE LEARNING AND BI INTEGRATION

Work Authorship Declaration

I declare to be the author of this work, which is unique and unprecedented. Authors and works consulted are properly cited in the text and are included in the listing of references.

Alexandre Miguel Guerreiro Contreiras

Copyright

ACKNOWLEDGEMENTS

This thesis marks the end of a long and challenging journey, and I would like to express my heartfelt gratitude to everyone who supported me throughout this process.

First and foremost, I am sincerely grateful to my supervisor, Ph.D. Célia Ramos, for her invaluable guidance, expertise, and encouragement. Her insightful feedback and support throughout this work were crucial to the successful completion of this thesis.

I would also like to extend my thanks to the studied company, which has chosen to remain anonymous, for providing the essential data and collaborating with me during this project. Your contribution was fundamental to the practical implementation of this research.

To my friends, especially Filipe, thank you for your encouragement and companionship. Our quiz nights and shared moments of laughter helped me stay grounded throughout this journey, and I will always cherish those memories.

To my parents, thank you for helping make the beginning of this academic journey and my education possible. To my sisters, thank you for always bringing joy and laughter into my life. And to my brother, your memory continues to inspire and motivate us all.

Finally, to my love, Rita, your endless support, patience, and love have been my constant source of strength. I can't begin to express how deeply grateful I am for your endless encouragement and understanding. Through every challenge, your belief in me has been the fuel that kept me going. Your love has been, and always will be, my greatest inspiration.

RESUMO

A previsão de vendas é um desafio significativo para empresas do setor turístico, uma vez que depende de vários elementos, como a sazonalidade, o comportamento dos turistas, as condições macroeconómicas e inclusive as condições meteorológicas. O Algarve é uma região central para o turismo em Portugal, atraindo turistas de diversas partes do mundo devido à sua extensa costa e condições favoráveis para atividades marítimo-turísticas.

Este trabalho tem como propósito investigar a integração de análise preditiva com técnicas de *machine learning* para a previsão de vendas numa pequena e média empresa no setor do turismo náutico. O estudo centra-se numa empresa que opera principalmente na oferta de passeios marítimo- turísticos, localizada no Algarve. Para tal, desenvolveu-se um modelo preditivo que utiliza dados históricos de *bookings*, bem como outros conjuntos de dados suplementares, tais como chegadas ao aeroporto de Faro, estadias no distrito de Faro e tráfego no website da empresa, integrando-os num software de *Business Intelligence*, o Power BI.

O principal objetivo deste estudo foi criar um modelo de previsão de vendas capaz de criar previsões precisas e de fácil interpretação referente aos lugares nas embarcações dos passeios da empresa em estudo. Estas previsões de vendas visam apoiar os gestores nas suas decisões operacionais e estratégicas. Com todo o desenvolvimento de modelos e análises dos *datasets*, não se pretende apenas realizar uma previsão de vendas, mas também fornecer informações úteis para a otimização de recursos e planeamento financeiro, que são fundamentais para melhorar a eficiência de pequenas e média empresas neste setor altamente competitivo. Assim, este estudo procura uma abordagem dinâmica que combina técnicas de *machine learning* com ferramentas de visualização de dados, permitindo que os gestores da empresa em estudo consigam compreender e aplicar conclusões em casos práticos na gestão da mesma.

A metodologia deste estudo baseou-se no CRISP-DM (Cross-Industry Standard Process for Data Mining), uma metodologia amplamente utilizada no ramo de ciência de dados e *data mining*. Este processo compreende seis fases: *Business Understanding*, *Data Understanding*, *Data Preparation*, *Modeling*, *Evaluation* e *Deployment*. A fase inicial envolveu uma análise aprofundada do setor e operadores do turismo náutico no Algarve,

bem como uma revisão da literatura sobre previsão de vendas e técnicas de *machine learning* aplicadas ao turismo.

A fase de *Data Understanding*, envolveu a recolha de dados provenientes de várias fontes. O *dataset* principal, consistiu em registos de reservas da empresa em estudo, recolhidos entre 2017 e 2024. Este conjunto de dados foi complementado com três *datasets* secundários: dados de chegadas ao Aeroporto de Faro, dados de estadias (dormidas) no distrito de Faro e dados referentes ao website da empresa. A integração destes conjuntos de dados proporcionou um contexto mais completo sobre os fatores que influenciam a procura turística, permitindo uma modelação preditiva mais robusta.

Durante a fase de preparação dos dados, foram aplicadas várias técnicas de extração, transformação e carregamento de dados, para garantir que estes apresentassem a qualidade desejada e demonstrando ser consistentes e prontos para análise. Além disso, foram criadas novas variáveis, como *Lead Time* (número de dias entre a data de reserva e a data de marcação) e *Total PAX* (número total de passageiros por reserva), para enriquecer o modelo. O tratamento de valores vazios e a remoção de *outliers* garantiram que o modelo estivesse preparado para gerar previsões fiáveis.

A fase de *Modeling* envolveu a aplicação de várias técnicas de *machine learning*, incluindo *Support Vector Regression*, *Random Forest* e *XGBoost*. Os dados foram divididos em três granularidades de previsão: diária, semanal e mensal. Para cada escala temporal, foram desenvolvidos modelos separados e os seus desempenhos foram avaliados utilizando métricas como MAE, RSME, WAPE e R^2.

Os resultados mostram que o modelo *Support Vector Regression* apresenta o melhor desempenho na previsão diária, oferecendo previsões mais precisas para intervalos de curto prazo. O *XGBoost*, por outro lado, destacou-se nas previsões de vendas semanais e mensais, demonstrando a sua capacidade de capturar padrões mais amplos e de longo prazo no comportamento das vendas. Estes resultados demonstram que, ao combinar diferentes modelos de *machine learning*, é possível obter previsões mais precisas dependendo do intervalo temporal em análise.

Os modelos desenvolvidos foram implementados através do Power BI, permitindo a criação de quatro *dashboards* interativos. Cada um destes *dashboards* foi desenhado para responder a diferentes necessidades de gestão da empresa, fornecendo insights cruciais.

O primeiro *dashboard* apresenta uma visão geral da performance da empresa. O segundo *dashboard* é focado na parte operacional da empresa. O terceiro *dashboard* auxilia na análise de desempenho financeiro, permitindo a visualização de receitas. Por fim, o quarto dashboard está focado nos resultados dos modelos, apresentando as métricas de desempenho e gráficos com as previsões geradas. Adicionalmente, foi integrado um gráfico de simulação de vendas que permite visualizar o impacto da alteração do preço da tour. O dashboard também inclui gráficos sobre as variáveis externas, proporcionando uma visão mais abrangente dos fatores que podem influenciar as reservas.

Esta integração com o Power BI facilita a interpretação dos resultados pelos gestores da empresa, permitindo-lhes ajustar rapidamente as suas operações com base nas previsões de vendas e comportamento do cliente, melhorando assim a eficiência operacional e a capacidade de resposta às mudanças no setor.

Uma das principais contribuições deste estudo foi a combinação de *machine learning* com *data storytelling* através de ferramentas de visualização interativas, como o Power BI. Esta abordagem não só forneceu previsões de vendas, como também tornou os resultados acessíveis e fáceis de interpretar para os gestores. Ao traduzir dados complexos em narrativas visuais, este estudo promoveu uma compreensão mais intuitiva das tendências de vendas e das necessidades operacionais. Este método contribui para o desenvolvimento de modelos preditivos mais sofisticados, que podem ser aplicados não só no turismo náutico, mas também em outros setores do turismo e hospitalidade.

Em conclusão, este estudo demonstra o potencial das técnicas de *machine learning* para transformar o processo de previsão de vendas em PME do setor turístico, fornecendo ferramentas mais eficazes para a tomada de decisões. O uso de ferramentas de BI e análise preditiva oferece às empresas a capacidade de ajustar a sua estratégia de forma proativa, com base em dados, melhorando assim a competitividade num mercado em constante mudança.

Palavras-Chave: Previsão de Vendas, Análise Preditiva, Turismo Náutico, *Machine Learning*

ABSTRACT

Sales forecasting is a significant challenge for companies in the tourism sector, as it depends on various factors such as seasonality, tourist behavior, macroeconomic conditions, and even weather conditions. The Algarve is a central region for tourism in Portugal, attracting tourists from various parts of the world due to its extensive coastline and favorable conditions for maritime tourism activities.

This thesis examines the integration of predictive analytics and machine learning for sales forecasting in a small and medium-sized enterprise within the nautical tourism industry. The primary objective is to develop a robust forecasting model that helps the studied maritime tourism operator in the Algarve optimize resources and make informed business decisions.

The methodology is based on the application of analysis and feature engineering, leveraging historical booking data along with secondary datasets, including district airport arrivals, overnight district stays, and website traffic. The data was processed, and the forecasting model, plus booking data, was integrated with Power BI to facilitate monitoring and operational planning. The results indicate that the SVR model is the most accurate for daily forecasting, while the XGBoost model demonstrates superior performance for weekly and monthly sales predictions. These models provide forecasts that support decision-making processes related to resource allocation and pricing strategies. The study also highlights the role of data storytelling in transforming raw data into actionable insights, allowing decision-makers to easily interpret and apply forecast results. This innovative approach contributes to the enhancement of sales forecasting models within the tourism sector, emphasizing the value of machine learning in improving financial efficiency and operational agility.

Keywords: Sales Forecasting, Predictive Analytics, Nautical Tourism, Machine Learning

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABREVIATIONS AND ACRONYMS

AI – Artificial Intelligence

AMPIC - Área Marinha Protegida de Interesse Comunitário

CCMAR - Centro de Ciências do Mar

CNN - Convolutional Neural Networks

EDA - Exploratory Data Analysis

ETL – Extract, Transform, Load

EU – European Union

Fundação Oceano Azul – OA

GDP - Gross Domestic Product

GVA – Gross Value Added

IQR – Interquartile Range

KPI – Key Performance Indicators

LSTM - Long Short-Term Memory

MAE - Mean Absolute Error

MAPE- Mean Absolute Percentage Error

ML – Machine Learning

MT – Maritime Tourism

MTO - Maritime Tourism Operator

NT - Nautical Tourism

RF – Random Forest

RMSE - Root Mean Squared Error

R^2 - R-squared

SME - Small and medium-sized enterprise

SVM – Support Vector Machine

SVR – Support Vector Regression

UK – United Kingdom

XGB – Extreme Gradient Boosting

WMAPE – Weighted Root Mean Squared Error

CHAPTER I - INTRODUCTION

Sales forecasting plays a pivotal role in optimizing business operations and strategically planning the future of a company. In recent years, predictive analytics, powered by ML (Machine Learning), has emerged as a disruptive innovation, enabling companies to become more agile, reduce waste, and improve profitability.

In the context of tourism, SMEs (small and medium-sized enterprises) with domain expertise can provide valuable insights for accurate sales predictions. By leveraging ML techniques, businesses can harness historical data to make informed decisions about resource allocation, inventory management, and marketing strategies.

This chapter will provide an outline of the thesis' motivation (Section 1.1), objectives (Section 1.2) and conclude with a thesis overview (Section 1.3).

1.1. Motivation

In the fiercely competitive business environment, the capacity to predict market trajectories based on robust data analysis and subsequently shape strategic responses is key for success (Danese & Kalchschmidt, 2011). For SMEs, this translates into a significant advantage in their competition within their touristic region. Precise sales forecasting isn't just a tool; it's a strategic asset empowering these enterprises to optimize resources, tailor services, and adapt swiftly to market fluctuations. The research seeks to develop tailored forecasting model that caters to the specific needs of the enterprises in study.

The motivation driving this research is rooted in the crucial role of sales forecasting within enterprise management (Mentzer & Moon, 2005). The object of study will be an SME nestled within the Algarve's tourism sector.

Central to the motivation of this research is the acknowledgment that beyond accurate forecasting models, the art of data storytelling holds immense value. Recognizing that raw data can be complex and overwhelming, the ability to distill insights into compelling narratives facilitates easier comprehension and more informed decision-making. By weaving data into narratives that resonate with stakeholders, these enterprises can comprehend market trends, customer behaviors, and industry shifts more intuitively.

In essence, the core motivation behind this research is to equip Algarve's SMEs with advanced predictive analytics, integrating ML and data storytelling. By blending

predictive analytics with narrative-driven insights derived from data, this study aims to make complex information more accessible for these enterprises. This fusion bridges the gap between raw data and actionable insights, empowering SMEs to intuitively interpret market trends and consumer behaviors. The goal is to provide a comprehensive decision-making framework, fortifying these enterprises within the dynamic and competitive tourism industry of the Algarve.

1.2. Objectives

This thesis aims to develop and evaluate a sales forecasting model using ML for a MTO (Maritime Tourism Operator) in the Algarve, Portugal. This model will be integrated with Power BI for monitoring. The primary goal is to provide a relevant sales forecast model that can support management in making informed decisions and help improving the operational and financial efficiency of the SME under study.

1.3. Thesis Overview

After this introduction, Chapters 2 and 3 of this thesis present a summary of the key contextual information that serves as the foundation for the research. These chapters explore nautical tourism in Portugal and the importance of sales forecasting, discussing relevant methodologies and techniques for this sector. They also introduce the concept of data storytelling as a critical tool for communicating sales forecasts to decision-makers.

Chapter 4 outlines the methodology used in this study, detailing the data collection process, feature engineering, and the development and evaluation of the forecasting models. Chapter 5 provides an in-depth analysis of the results, discussing the performance of the forecasting models across different time scales and their practical implications for operational planning and decision-making within the business.

Finally, the thesis concludes with a summary of the key contributions and findings, followed by the limitations encountered during this work and suggestions for future research.

CHAPTER II - NAUTICAL TOURISM

The tourism sector, a fundamental driver of Portugal's economy, has witnessed substantial diversification in recent years. Among the various niche segments, NT (Nautical Tourism) has emerged (EY-AM&A, 2019) as a significant contributor to the nation's tourism industry, particularly in the region of Algarve (EY-AM&A, 2019). This section delves into the dynamics of Maritime Tourism (MT) operations in the Algarve, examining its intricate interconnections with the broader Portuguese tourism industry.

2.1. Tourism in Portugal:

Portugal, renowned for its rich cultural heritage and stunning landscapes, has solidified its position as a sought-after destination for tourists worldwide. Portugal's tourism sector is integral to the country's economic development, representing 15.80% of the country's GDP (Gross Domestic Product) in 2022 (INE, 2023).

According to the latest statistics from INE (2023), Portugal experienced a significant economic recovery in 2022, with a GDP growth of 6.70%, the highest since 1987. This growth was largely fueled by strong domestic demand, especially in private consumption, which increased by 5.80%. Despite slower growth in public consumption and investment, exports of goods and services surged by an impressive 80.90%, driven primarily by the booming tourism industry. This led to a positive net external demand contribution of 1.90 percentage points, reflecting a higher international demand for Portuguese goods and services compared to the domestic demand for foreign products (INE, 2023).

The tourism sector played a crucial role in this economic rebound, with 28.9 million visitors and 77.20 million overnight stays in 2022, marking increases of 80.70% and 81.10%, respectively, over the previous year (INE, 2023). Domestic tourism accounted for 35.6% of total overnight stays, growing by 22.20%, while international tourism saw a substantial rise of 146.90%, though it remained below 2019 levels. Revenue from accommodation establishments also saw a significant increase, reaching 5.00 billion euros, a 115.2% rise from the previous year (INE, 2023).

When compared to other Southern European countries, Portugal's recovery in the tourism sector has been commendable, although there is still room for improvement to reach pre-pandemic levels (European Travel Commission, 2023).

Looking ahead, the future of Portugal's tourism sector appears promising. With the government's continued support and initiatives to promote sustainable and responsible tourism, coupled with the sector's demonstrated resilience, Portugal is well-positioned to further capitalize on its tourism potential (WTTC, 2022). However, it will be important to monitor global trends and challenges, such as the ongoing international conflicts and the increasing importance of sustainable practices in tourism (Ministério da Economia e da Inovação, 2017).

The global tourism industry has been witnessing a shift towards experiential travel, with tourists seeking unique and immersive experiences (World Economic Forum, 2022). Within this expansive industry, NT has emerged as a compelling choice for travelers (Vázquez et al., 2021).

2.2. Nautical Tourism in Portugal

NT, a subset of the broader tourism industry, encompasses three primary segments: cruises, yachting, and MT (Ministério da Economia e da Inovação, 2007). Benefiting from Portugal's extensive coastline and favorable weather conditions, the nation has emerged as an ideal destination for nautical enthusiasts (Kovačić & Silveira, 2018). In addition to the vast territorial resources allocated to these activities (Pereira et al., 2015), the maritime dimension also plays a key role in shaping the country's international image (Martins, 2015; Ministério da Economia e da Inovação, 2017)

In 2018, tourism contributed 8.00% to GVA (Gross Value Added). For the same year, the "Recreation, sport and tourism" sectors accounted for 71.90% of jobs and 69.80% of GVA (91,000 jobs and 2.30 billion euros) in the total Economy of the Sea (Ministério do Mar, 2021). Despite the positive economic impact, NT exerts pressures on ecosystems, marine species, coastal and underwater heritage, and navigation safety, particularly at the local level (Ministério do Mar, 2021).

This concern aligns with Portugal's broader dedication to the EU (European Union) and the European Agenda for Tourism 2050, emphasizing innovative and sustainable tourism. According to the 'Estratégia Nacional para o Mar 2021-2030' by Ministério do Mar of Portugal, it aims to foster an ecologically sustainable ocean, promoting blue growth – a long term strategy to support sustainable growth in the marine and maritime sectors as a whole (De Vet et al., 2016) - enhancing the quality of life for the Portuguese populace, and

positioning Portugal as a leader in ocean governance, all based in scientific understanding. This strategic viewpoint underscores the national responsibility to strike a balance between economic growth and environmental preservation, ensuring the durability of maritime operations and position with international sustainability objectives.

Moreover, the priority in this sector is to promote a diversified infrastructure network supporting nautical recreation, including maritime-touristic activities (Vázquez et al., 2021). This involves enhancing marinas, docks, moorings, passenger boarding facilities, and support structures, valuing coastal and maritime cultural heritage, and engaging local communities. All of this, while, simultaneously, developing an integrated regulatory framework for maritime recreational activities, which is crucial (Ministério do Mar, 2021).

2.3. Maritime Tourism Operators in the Algarve

As NT gains traction, MTOs play a critical role in facilitating and organizing maritime activities for tourists. These operators serve as intermediaries between tourists and the diverse array of nautical experiences available in the region.

The Decree-Law No. 108/2009 of September 3rd defines that "touristic animation activities developed through the use of boats for profit are designated as maritime touristic activities". These include, for example, the following activities: maritime tourist tours; boat rental with/without crew; services performed by taxi ferry; tourist fishing; among others.

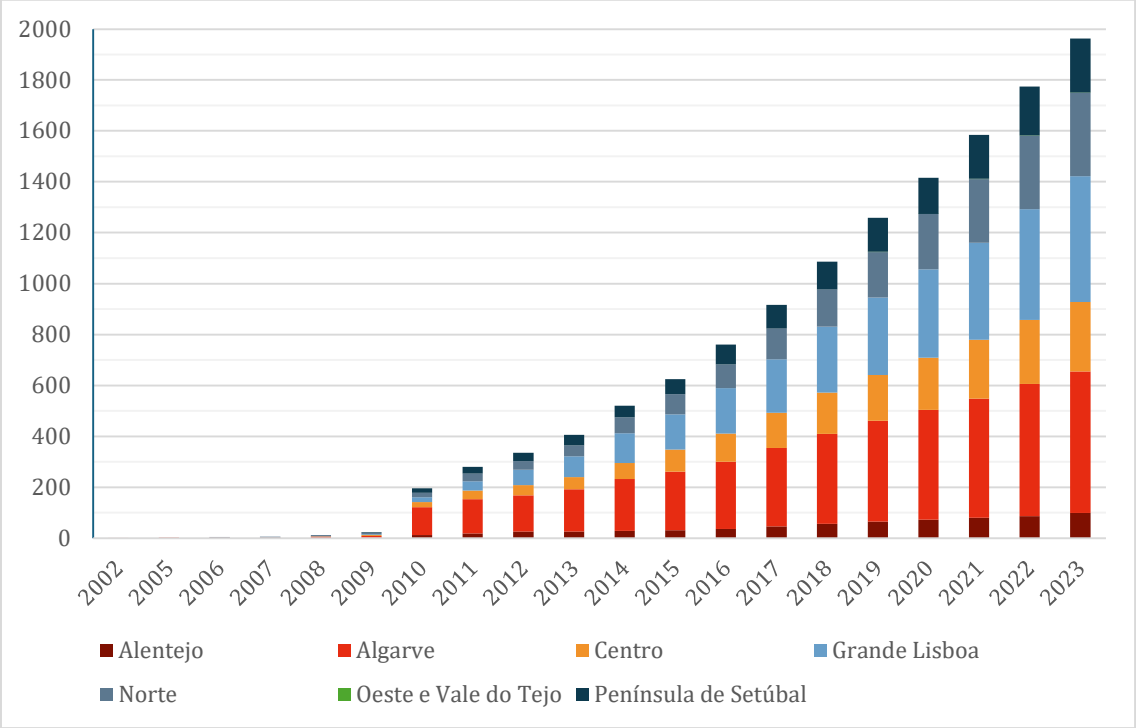
Understanding the operational dynamics, market trends, and challenges faced by MTOs is essential for comprehending the broader context of the Algarve's MT sector.

The total number of registrations for companies that practice this type of activity has been increasing exponentially, as demonstrated by Figure 1. In 2023, the total number of registrations was 1962, while in 2013 it was 406 (almost five times more) – representing an average annual growth of 17.0% (RNAAT, 2024).

The Algarve region has the most registrations of MTOs, demonstrating the importance of this region for this type of activities in Portugal.

Figure 1

Total Number of MTOs Registered in Portugal from 2002 to 2023



RNAAT (2024)

In the survey characterizing the demand for tourist animation agents, conducted in 2021 by Turismo de Portugal, MT activity is again highlighted. Regarding the demand for tourist animation activities (where MT activity is included), outdoor activities were particularly popular, mainly those integrated into nature tourism (33.00%), and maritime-tourist activities (30.00%). (Turismo de Portugal, 2022).

In 2020, research to implement an AMPIC (Área Marinha Protegida de Interesse Comunitário) on one of the biggest costal reefs of Portugal, between Farol de Alfanzina and Marina de Albufeira, was conducted by the CCMAR (Centro de Ciências do Mar) and funded by the OA (Fundação Oceano Azul). Three types of economic activity with high social and economic relevance to that region were studied: commercial fishing, recreational fishing, and MT. Despite the study's incidence on a particular region of the Algarve, the AMPIC, (opposed to the entirety of its region), due to its depth and quality of the research, some of its findings can be generalized to a broader perspective of the Algarve region.

This study estimated that, in 2018, the MT sector supported by AMPIC attracted a total of 993,909 visitors – in the same year, the Algarve received 4.7M guests (INE,2019) –

meaning that of these, 21.00% practiced MT activities in AMPIC alone. Most of these visitors made, on average, one MT outing per stay.

According to the information provided by the interviewed operators (*N*=53) and extrapolated to the total number of MTOs using AMPIC (*N*=74), the MT industry generated a total of €40M in gross revenues (Ressurreição et al., 2020). Among the different modalities, coastal tours emerged as the most popular, contributing approximately €30 million in revenue in 2018. Cetacean observation followed was the second most lucrative MT activity, generating around €10 million. Diving accounted for approximately €0.5 million, while recreational fishing with operators contributed around €300,000 to the overall revenue.

In 2018, the MT industry, in AMPIC, also supported 1051 direct jobs, of which 290 were permanent and 761 seasonal. Additionally, 80.00% of MT operators' managers declared that this occupation was their only professional occupation.

The vast majority of customers for MT activities at AMPIC are foreign tourists. Diving was the activity that attracted the highest percentage of domestic tourists (23.00%). The main countries of origin of AMPIC's MT clients are the United Kingdom (UK), France, Ireland, Germany, the Netherlands, Belgium, Spain, and Italy. This is in line with the number of overnight stays by people from these countries. According to Turismo de Portugal, the UK (37.00%), Germany (13.10%), France (7.30%) and Spain (6.60%) were the countries with the most overnight stays (by foreigners) in the Algarve in 2018 (Turismo de Portugal, 2023).

MTOs use several ways to advertise their services: contracts with travel agencies, hotels, and ticketshops, as well as online platforms such as Get Your Guide, Facebook, Google Ads, and web pages were the most used forms.

Following the conclusion of the survey, MTOs were asked to delineate the primary challenges they encounter in the advancement of their activities Figure 2. Challenges associated with an excess of companies/boats, burdensome fees and bureaucratic hurdles, lack of training for skippers, and the absence of an institutional strategy to promote these activities were unanimously identified by operators across all MT sectors.

Predominantly, operators cited the deficiency in infrastructure (e.g., boarding pontoons, restroom facilities in marinas) as a significant hindrance to the development of

these activities. Additionally, the absence of integrative management within this sector, alongside a lack of coordination and unity among operators and other entities (e.g., maritime law enforcement, local municipalities, tourism institutions), were underscored as limiting factors (Ressurreição et al., 2020).

Figure 2

Current Challenges Faced by the Different Maritime Activities

Diving	**Recreational fishing**
Other users do not respect signaling buoys	Too many companies/boats
Too many companies/boats	Too many fees and bureaucracies
Too many fees and bureaucracies	
Lack of promotion of this activity	
Lack of supporting infrastructure (e.g. pontoons, WCs, etc)	
Cetacean watching	**Coastline trips**
Mass tourism mostly interested in going to the beach	Mass tourism, poorly differentiated
Too many companies/boats, putting the economical and ecological sustainability of this activity at risk	Too many companies/boats
Lack of training for professionals (e.g. Skippers)	Too many fees and bureaucracies
Lack of institutional promotion for these activities	Lack of qualified professionals
Lack of integrated management of the sector	Lack of promotion of these activities
Too many fess and bureaucracies	Lack of cooperation between operators
Lack of supporting infrastructure (e.g. pontoons, WCs, etc)	Lack of supporting infrastructure (e.g. pontoons, WCs, etc)
Lack of cargo capacity evaluation for these activities	Unfair competition
Lack of coordination and unity between operators	Overcrowding in Benagil
Unlicensed companies	
Unfair price competition	
Overcrowding in Benagil	

Ressurreição *et al.* (2020, p.122)

To exemplify this situation, the Benagil Caves Working Group was established in 2023 to address the uncontrolled growth of MT in the Algarve, specifically in the Benagil Caves (CCDR Algarve, 2023). The group aims to determine the human carrying capacity and access conditions for the caves located on Benagil Beach in Lagoa, Portugal.

While the economic and social importance of the activity was emphasised, ensuring the safety of practitioners due to overcrowding of boats and companies is a top priority (CCDR Algarve, 2023). Moreover, the sector must adapt to the new rules and regulations on sustainability and Blue Growth outlined in the Estratégia Nacional para o Mar 2021-2030,

as its rapid growth poses challenges for environmental sustainability and the management of maritime space (Ministério do Mar, 2021).

The characteristics of the sector and market trends indicate that there is still ample room for development (Vázquez et al., 2021), meaning that MTOs will have to be prepared for any circumstances and adversities (Almeida & Silva, 2020). For example, the COVID-19 pandemic forced many companies in Portugal, including those in the maritime sector, to adopt digital platforms for bookings and sales. Costa *et al.* (2022) highlighted how digital tools helped small-scale fisheries (a part of the maritime sector), to maintain operations during the pandemic by shifting to online sales and auctions. This trend towards online transactions was further supported by Plácido *et al.* (2021), who observed a significant increase in e-commerce in Portugal during the pandemic, reflecting how businesses across various sectors adapted to survive.

Additionally, in the NT sector, the 2020 boat rental season saw significant adjustments, with companies leveraging online bookings and digital marketing to attract local customers in the absence of international tourists. Many operators reported higher-than-expected demand during certain periods as travellers sought more isolated and flexible vacation options (YachtSys, 2020). Globally, research by Makasarashvili and Giguashvili (2021) and Tairov and Petrova (2022) demonstrated that transitioning to online operations allowed businesses to improve efficiency and increase revenues, a trend not limited to Portugal.

As part of this forward-thinking approach, the use of tools such as sales forecasting can enable MTOs to remain agile and competitive in a constantly evolving market (Jiao & Chen, 2018). By analysing marketing dynamics and customer trends, companies can adapt swiftly to changing circumstances (Pelham, 2000). The increasing reliance on online sales channels, particularly accelerated by the COVID-19 pandemic, has further expanded the volume of data available. Utilizing this data, as explored in the study on the use of customer reviews for demand distribution and sales forecasting, businesses can capitalize on big data analytics (See-To & Ngai, 2018). This strategic approach not only enhances operational efficiency but also enables SMEs to capitalize on emerging opportunities (Cadavid et al., 2018), where the role of sales forecasting in NT would play a part, helping explore how it can drive business success and ensure sustained growth in this dynamic industry.

CHAPTER III - SALES FORECASTING IN NAUTICAL TOURISM BUSINESS

Sales are an essential aspect of a company's operations, serving as a fundamental revenue-generating function (Donaldson, 1998). In tourism literature, the role of sales in enhancing revenue and supporting industry expansion is well-documented, applying to various segments including nautical and maritime tourism.

3.1. The Role of Sales in Nautical Tourism

In addition to generating revenue, sales operations contribute significantly to various aspects of organizational growth and development (Djatmiko et al., 2018). For MTOs, this could play a significant role in their market expansion and penetration, as the efforts of the sales team are directed towards acquiring new customers and expanding the company's market presence (Terho et al., 2015), enabling companies to reach new customer segments and geographical markets (Zoltners et al., 2006).

Moreover, it can be argued that sales efforts are of vital importance in establishing brand awareness and promoting brand loyalty (Haas et al., 2012), both of which are key for sustaining enduring customer connections and augmenting brand equity. In a 2022 study that looked at the brand equity of different NT firms in the Centro Region of Portugal, Santos *et al*. (2022) concluded that brand equity is a significant variable that exerts a positive influence on corporate performance and economic sustainability.

By adopting a well-planned sales strategy, MTOs, as businesses, can proactively interact with their customers, gain insights into their requirements (Guenzi & Troilo, 2007), and adeptly customize their products or services to meet necessary needs (Terho et al., 2012).

In general, the impact of sales strategy on companies' performance is also significant. Sales performance metrics provide insights into the effectiveness of the sales strategy and can influence strategic decision-making at the company level, it can also directly influence overall company performance (Panagopoulos & Avlonitis, 2010), with high sales performance often translating into increased revenue. MTOs that employ sales strategies and metrics would be able to benefit from these factors.

Ultimately, an effective sales strategy, empowers a company to stand out in the market, ensuring a sustained competitive edge through informed decision-making and strategic

agility (Terho et al., 2015), meaning that in the context of NT, the alignment of sales strategies with organizational objectives is crucial in determining the overall success and performance of MTOs in the market.

3.2. Sales Forecasting in Nautical Tourism

Sales forecasting is the process of estimating the quantity of specific products and services that an enterprise is likely to sell in a specific time in the future (Mentzer & Bienstock, 1998), based on past sales data and methodical forecasting models (Shiman, 2023). It is a multifaceted field that plays a critical function in business strategy (Makridakis, 1996), resource allocation (Stein, 1997), and decision-making (Hyndman & Athanasopoulos, Forecasting: principles and practice, 2018).

3.2.1 Importance of Sales Forecasting in Nautical Tourism

Sales forecast is key to production, transportation, and workforce decisions at all levels of a supply chain (Verstraete et al., 2020). By anticipating future sales trends and demand, NT companies can align their goals, manage resources efficiently (Stein, 1997), and optimize operational efficiency (Oliva & Watson, 2009).

Forecasting sales can directly impact financial projections, whether it's projecting revenue, cashflow, or profit margins, accurate sales estimates are essential (Pan et al., 1977). Investors, lenders, and stakeholders rely on these projections to assess a company's financial health and growth prospects (Abor, 2017).

As referenced earlier, international tourism is growing, driven by social, economic, political and technological factors. This trend introduces both opportunities and challenges as destinations compete for limited resources (Lim, 2006), this created the need to companies such as airlines, tour operators, hotels, cruise ship lines, and recreation facility providers to monitor and anticipate demand for their products by tourists (Song & Turner, 2006).

According to Song & Turner (2006), "the success of many businesses depends largely or totally on the state of tourism demand, and ultimate management failure is quite often due to the failure to meet market demand". It is crucial to distinguish between overall tourism demand and a company's specific sales forecast. However, while specific literature on sales forecasting in NT is limited, the sector's principles of demand forecasting provide valuable insights.

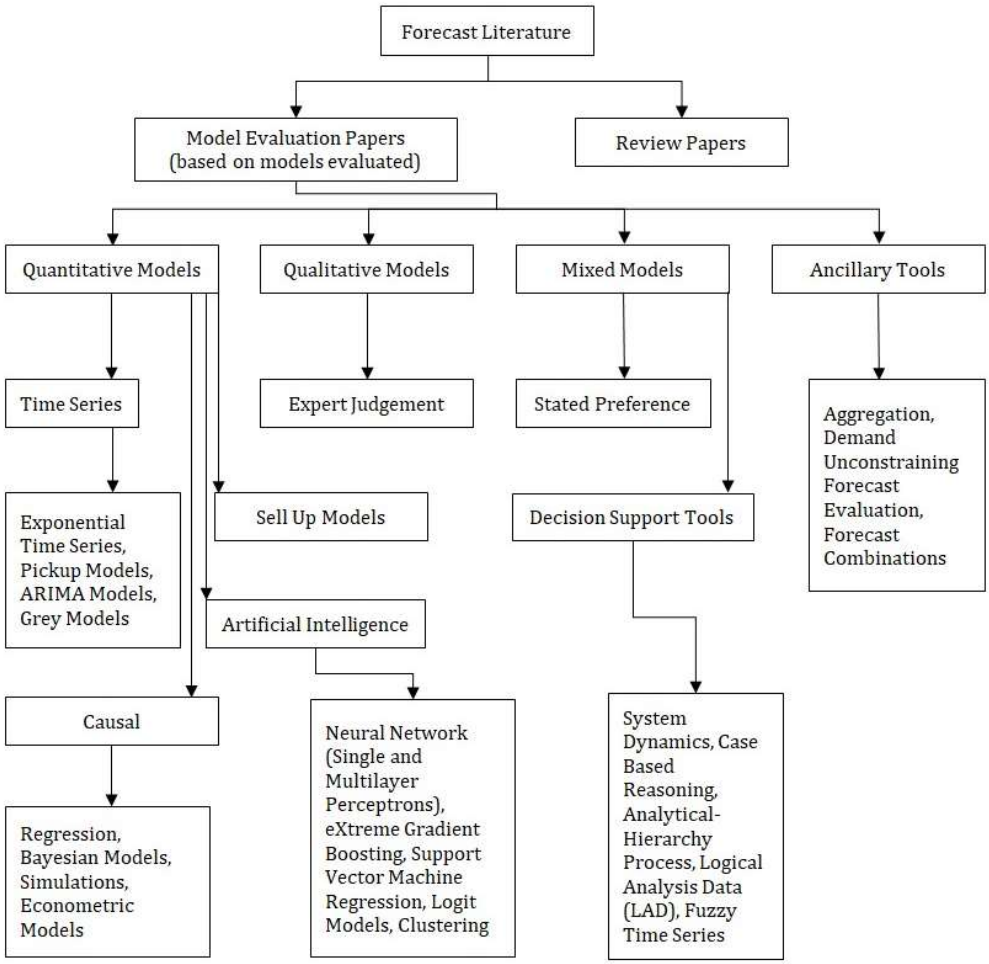
Applying these principles MTOs, can create accurate sales forecasting which enables companies to navigate the complex interplay of global tourism trends and local operational requirements. This foresight allows MTOs to make informed decisions that enhance their competitive advantage and ensure sustainable growth within the sector.

3.2.2 Forecasting Methods

Over the years, businesses have employed diverse forecast techniques (Morlidge & Player, 2010), ranging from conventional methods grounded in statistical analysis and expert judgment to contemporary ML algorithms adept at handling extensive datasets (Banerjee et al., 2020). Forecasting methods can be broadly classified into qualitative and quantitative approaches (Figure 3), each with distinct benefits and challenges.

Figure 3

Forecasting Methods



Adapted from Banerjee *et al.*(2020, p.799)

3.2.2.1 Qualitative Methods of Forecasting

In scenarios where historical data is unavailable, incomplete, or inconsistent, qualitative forecasting offers valuable insights for NT operators, since it does not rely on numerical data but rather on expert knowledge, intuition, and judgment (Shiman, 2023).

Expert Judgment:

Expert judgment relies on insights, opinions, and forecasts from individuals possessing domain expertise and industry knowledge (Blair et al., 2010). In the context of MTOs, experts such as seasoned tour operators and sales managers, industry analysts, and subject matter experts contribute qualitative inputs based on their understanding of market conditions, competitor strategies, economic indicators, and emerging trends (Shiman, 2023). While inherently subjective, expert judgment complements quantitative forecasting methods by providing nuanced insights into uncertain or ambiguous situations (Mukhopadhyay et al., 2007), such as sudden changes in tourism demand due to weather patterns or geopolitical events.

3.2.2.2 Quantitative Methods of Forecasting

Quantitative forecasting is a methodical approach that uses mathematical models and historical data to predict future outcomes (Armstrong, 2001). It's a fact-based method that relies on numerical data and statistical analysis, contrasting with qualitative forecasting which is more subjective and relies on expert opinions (Meneghini et al., 2018). In the context of NT, where data trends such as seasonal demand fluctuations and external factors like weather and economic indicators play a crucial role, these methods provide a solid foundation for accurate sales forecasting.

Time Series Analysis:

Time series analysis requires the examination of historical sales data across time intervals to recognize patterns, trends, and seasonality (Chatfield, 2000). It uses models which are commonly employed to analyze time series data and can be used to project future sales figures. In markets with stable demand, these methods can be executed with ease and have demonstrated strong performance (Chopra & Meindl, 2013). Table 1 provides an overview of key time series models, along with their practical applications and detailed descriptions, relevant to forecasting in NT.

Table 1

Time Series Models

Model	*Description*	*References*
Exponential Time Series (ETS)	Forecasts based on past data where older observations have exponentially decreasing weight. Often used as a benchmark.	Samagaio and Wolters (2010) (airport transit), Riedel and Gabrys (2003) (comparative forecasting for demand).
Pickup Forecasting	Utilizes a booking matrix to project cumulative forecasts based on progressive bookings. Effective in short-term forecasting.	Wickham (1995) (airline), Sun, Gauri, and Webster (2011) (cruise line).
Autoregressive (AR)	Regresses a variable against its past values. AR (1) uses the immediate past, AR (2) goes back two steps, and so on.	Riedel and Gabrys (2003) (flight-specific demand forecasting), Sickles *et al*. (1998) (strategic demand in transportation).
Moving Average (MA)	Forecasts based on past forecast errors. Often used in conjunction with AR models.	Riedel and Gabrys (2003), Sickles *et al*. (1998).
ARMA	Combines AR and MA models for more comprehensive forecasting.	Riedel and Gabrys (2003) (comparative study with multiple models), Sickles *et al*. (1998).
ARIMA & SARIMA	Adds differencing to ARMA models to account for trends, with SARIMA incorporating seasonality.	Dutta and Ghosh (2012) (railways), Cyprich, Konecny, and Kilianova (2013) (bus ridership).
Grey Model	Transforms data to a monotonic series for forecasting, then reverse-transforms it.	Wang *et al*. (2013) (yearly passenger volume), Carmona Benitez *et al*. (2013) (nautical transport).
GARCH	Measures conditional variance to capture periods of volatility in data, especially useful for revenue management.	Wang (2016) (cruise line revenue management), Nieto (2018) (airline forecasting)

Adapted from Banerjee *et al* (2020, p.803)

Causal Models:

Causal Models (see Table 2) focus on identifying causal relationships between sales and external factors such as economic indicators, demographic trends, and industry-specific variables (Lee et al., 2008). In NT, techniques such as Regression, Logit, Econometric and Bayesian models would leverage statistical techniques to quantify the impact of independent variables and generate forecasts based on projected values of the predictors (Banerjee et al., 2020).

Table 2

Causal Models

Model	*Description*	*References*
Regression	Identifies the relationship between a scalar dependent variable and one or more explanatory variables. Commonly used for demand forecasting when explanatory variables are provided.	(Sa, 1987) (passenger transport demand), (Dutta & Marodia, 2015) (revenue management in transportation).
Logit models	Regression models where the dependent variables are categorical. Commonly used in transportation forecasting for predicting passenger choice and demand.	(Liu & Li, 2012) (high-speed railway systems), (Cipriani *et al*., 2014) (passenger choice in intercity flights), (Leng *et al*., 2015) (short-term demand forecasting).
Econometric models	May include multinational regressions or gravity models to forecast travel interactions between regions. Suitable for long-term planning and travel interactions.	(Cassetta & Coppola, 2014) (demand in railway), (Tsekeris & Tsekeris, 2011) (demand in transport systems)
Bayesian models	Uses observed data in time series to estimate state-space models for forecasting. Applicable in complex forecasting systems such as NT demand predictions.	(Jiao *et al*., 2016) (demand in transport systems), (Tsekeris & Tsekeris, 2011) (transport forecasting), (Aston & Koopman, 2006) (transport forecasting).

Adapted from Banerjee *et al* (2020, p.803)

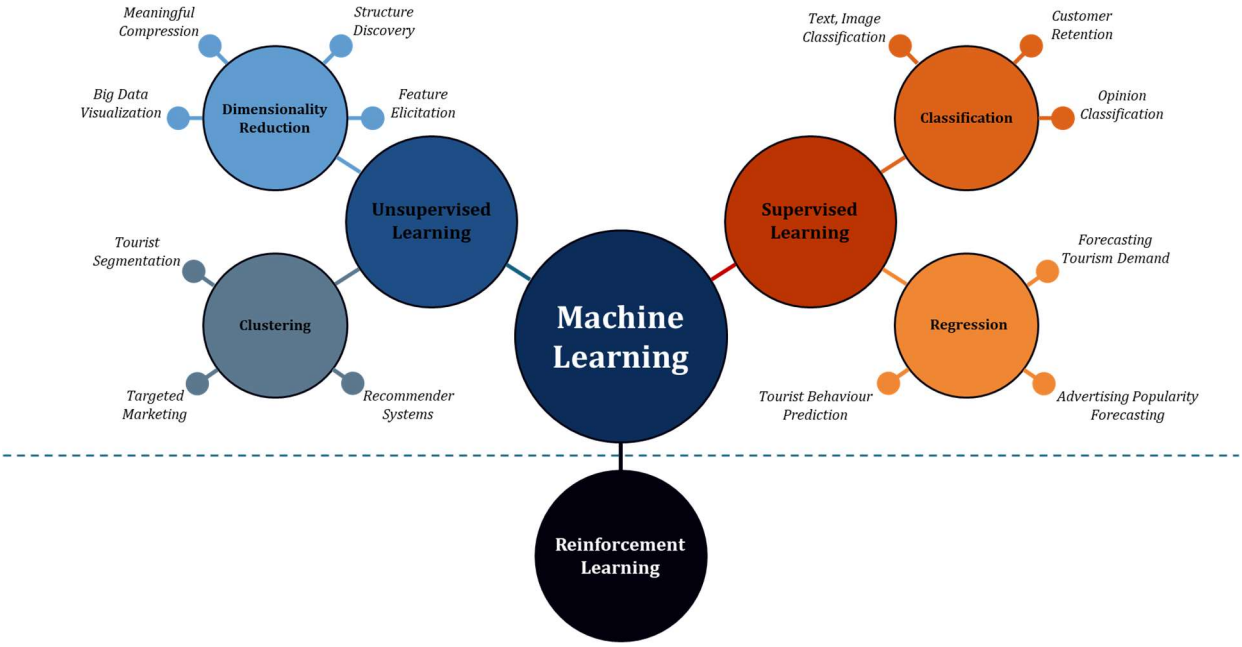
3.2.3 Machine Learning for Predictive Data Analytics

Artificial Intelligence (AI) has instigated a significant paradigm shift in the field, emerging as a powerful tool in sales forecasting (Loureiro et al., 2018). ML, a subset of AI, utilizes algorithms to learn from data, enabling the development of predictive models and adaptive systems. As an automated process that identifies patterns in data, it employs advanced computational techniques to analyze large datasets and make accurate predictions (Kelleher et al., 2020).

ML is a broad field, encompassing a variety of algorithms and approaches (see Figure 4). These can be broadly categorized into three types: unsupervised learning (clustering and dimensionality reduction), supervised learning (classification and regression), and reinforcement learning (Jamal et al., 2018). Reinforcement learning, while a significant area of ML, has low relevance in the context of tourism when compared to the Supervised and Unsupervised Learning (Egger, 2022).

Figure 4

ML paradigms and application examples



Adapted from Egger (2022, p90)

ML algorithms can be further differentiated based on the type of data they handle and the availability of labels for the dependent variable. Algorithms can process either continuous or discrete dependent variables. If these variables are labeled, supervised methods are employed. In the absence of labels, unsupervised methods are utilized (Egger, 2022).

Supervised ML is often used to build models for predictive data analytics applications (Osisanwo et al., 2017). This technique learns the relationship between descriptive and target features from historical examples, which can then be used to predict outcomes for new instances (Kelleher et al., 2020).

The nature of the target variable (continuous or discrete) determines whether it is a regression or classification problem. Labeled training data is used to create a model that can accurately predict the output for new data (Awad & Khanna, 2015), meaning that the quality of the training data has a significant impact on the model results.

3.3.2.1 Machine Learning in Nautical Tourism

In the context of NT, ML has been leveraged to enhance various aspects of maritime operations and tourism overall forecasting. These applications underscore the versatility

and potential of ML in improving efficiency, safety, and customer experience within the NT sector.

One significant application is in forecasting vessel flows. Deep learning methods, such as convolutional neural networks (CNN) and long short-term memory (LSTM) networks, have been employed to predict maritime vessel inflow and outflow. These methods significantly enhance the accuracy of these predictions compared to traditional techniques, aiding in congestion management, emission reduction, and improved route planning (Zhou et al., 2020).

ML techniques have also been instrumental in tourism demand forecasting. By analyzing search engine queries, algorithms such as SVMs (Support Vector Machines) and neural networks outperform traditional models in terms of forecasting accuracy and robustness. These advanced approaches provide valuable insights for tourism planning and policymaking (Sun et al., 2019).

In addition, ML models have been developed to estimate ocean-wave conditions, which are crucial for NT activities like sailing and marine navigation. These models demonstrate high accuracy in predicting wave heights and periods, thereby enhancing safety and operational planning (James et al., 2018).

Optimized ML approaches have been applied to forecast Chinese cruise tourism demand, effectively utilizing big data and search query data to improve prediction performance. This is critical for investment decision-making and planning in the cruise industry (Xie et al., 2021).

Moreover, advanced ML techniques, such as the Perceptron-based Feature and Kriging Gradient Boost Classification, have been used to enhance the accuracy and efficiency of marine weather forecasting. This application is vital for the planning of NT activities and maritime operations (Anbarasi & Radha, 2023).

Overall, the integration of ML into NT has proven to be transformative, offering enhanced forecasting capabilities that improve operational efficiency, safety, and customer satisfaction within the industry.

3.2.4 Traditional Methods vs. Machine Learning Approaches

Traditional forecasting methods, while well-established and often effective in stable environments, struggle to adapt to rapid or unforeseen macroeconomic shifts and changes in the business landscape, which can drastically influence sales patterns (Verstraete et al., 2020). Industries like NT, where multiple external factors—such as seasonality, macroeconomic trends, weather, and customer behavior—affect sales, traditional models find it difficult to capture the complexity and non-linear patterns inherent in such environments (Anesti et al., 2024).

According to Law *et al*. (2019),traditional forecasting models can encounter difficulties when processing large volumes of search intensity indices used in predictive tourism models. Similarly, Tripathi *et al*. (2019) point out that although traditional models are favored in many industries due to their simplicity, they often fall short in accurately capturing the intricate patterns in sales data, particularly when faced with non-linear trends and high-dimensional datasets.

ML approaches provide a more sophisticated alternative, capable of handling the complexities of modern sales forecasting (Anesti et al., 2024). Unlike traditional methods, ML models can process and analyze vast amounts of data, including categorical, numerical, text, and even image data (Choi et al., 2018), enabling them to provide a richer and more comprehensive forecast (Cadavid et al., 2018). This versatility is especially useful in complex scenarios where sales are driven by multiple variables, including customer behavior, seasonality, and macroeconomic indicators (Cadavid et al., 2018).

One key advantage of ML models is their ability to detect and model non-linear relationships between variables, which is often missed by traditional approaches (Helmini et al., 2019). For instance, XGB (Extreme Gradient Boosting), a popular tree-based ML model, has been shown to outperform traditional methods like ARIMA (Ke et al., 2017) (Ke et al., 2017)and SARIMA (Atanda et al., 2024) by accurately capturing complex, non-linear relationships between sales drivers. By integrating diverse data sources and automatically adjusting to new trends, ML models provide more flexible and adaptive forecasts, crucial in industries like NT.

Furthermore, ML models are far better at handling large, high-dimensional datasets. Unlike traditional methods that often struggle with complex relationships and changing

patterns, ML algorithms can process large volumes of data, capture intricate relationships between variables, and adapt to new trends, making them increasingly popular for sales prediction (Pustokhina & Pustokhin, 2023).

Despite its advantages, ML-based forecasting comes with certain challenges (Shiman, 2023). One of the key barriers is the need for substantial computational resources (Choi et al., 2018), especially for deep learning models or large-scale ensemble methods. Companies must invest in both infrastructure and technical expertise to implement ML effectively, which may be prohibitive for SMEs (Cadavid et al., 2018).

Data quality and preprocessing are other critical factors that can significantly impact the performance of ML models (Egger, 2022).Therefore, substantial effort must be put in data cleaning, transformation, and feature engineering to ensure that the model accurately reflects real-world sales patterns rather than noise (Hyndman & Athanasopoulos, 2021)

In summary, while traditional forecasting methods remain valuable in predictable, stable environments, ML models offer a superior alternative for dynamic, data-rich scenarios. By integrating diverse data sources and capturing complex relationships, ML can significantly improve forecast accuracy. However, organizations must weigh the trade-offs between computational requirements, technical expertise, and model interpretability when choosing between traditional and ML-based approaches. Equally important is the ability to translate these complex forecasts into clear, actionable insights for decision-makers—a process where effective data storytelling becomes essential, enabling stakeholders to understand and leverage the full potential of the models.

3.3. Data Storytelling

3.3.1 Definition and Concepts

Data storytelling is a multidisciplinary approach that combines data analysis, visualization, and narrative techniques to communicate complex information and insights in a compelling manner (Errey et al., 2024). It recognizes the innate human tendency to understand and remember information through stories, making data more accessible and engaging (Kosara & Mackinlay, 2013). Data storytelling involves transforming raw data into meaningful narratives that captivate the audience, enhance comprehension, and drive action (Ryan, 2018).

The key concepts in data storytelling include data analysis, visualization, narrative structure, and audience-centricity. Data analysis involves extracting insights, patterns, and correlations from datasets using statistical methods and analytical techniques (Segel & Heer, 2010). Visualization techniques, such as charts, graphs, and infographics, are employed to represent data visually and facilitate understanding (Fry, 2007). Narrative structure provides a framework for organizing the story, incorporating an introduction, a clear objective, supporting evidence, and a conclusion (Segel & Heer, 2010). Lastly, an audience-centric approach tailors the storytelling to the target audience's level of data literacy, domain knowledge, and preferences (Dykes, 2020).

Advancements in data analytics and big data have significantly contributed to the rise of data storytelling. The abundance of data generated in various domains necessitates effective communication of insights and findings. Data storytelling emerged as a response to this need, offering a structured and engaging approach to convey data-driven information to diverse audiences (Zhang et al., 2022).

3.3.2 Practical Applications

Data storytelling finds practical applications across multiple domains. In business, it helps organizations present KPIs (key performance indicators), market trends, and customer insights to support decision-making and strategic planning (Segel & Heer, 2010). Through data storytelling, organizations can bridge the gap between data insights and real-world outcomes by ensuring that stakeholders understand not only what the data says but why it matters.

3.2.3.1 Uses of Data Storytelling in Nautical Tourism

In the context of NT, data storytelling can transform data into actionable insights for various management functions:

Marketing and Customer Insights: Analyzing social media and customer feedback to understand tourist behavior and preferences, which helps in tailoring marketing campaigns and improving customer service Treboux *et al.* (2016) demonstrated the potential of social data mining and visualization in informing marketing decisions.

Operational Efficiency: The maritime industry benefits from data visualization, which makes complex information more accessible (Dykes, 2020). By analyzing operational data—such as fuel consumption, maintenance schedules, and crew performance—companies can optimize their processes, reduce costs, and enhance safety. Notably, big data analytics play a crucial role in maritime logistics, aiding in areas like logistics optimization and safety enhancement (Mirović et al., 2018).

Strategic Planning: Providing insights into market trends, competitor analysis, and financial performance supports evidence-based decision-making and helps in setting realistic and achievable business goals. Miah *et al*. (2017) presented a big data analytics method for supporting strategic decision-making in tourism destination management.

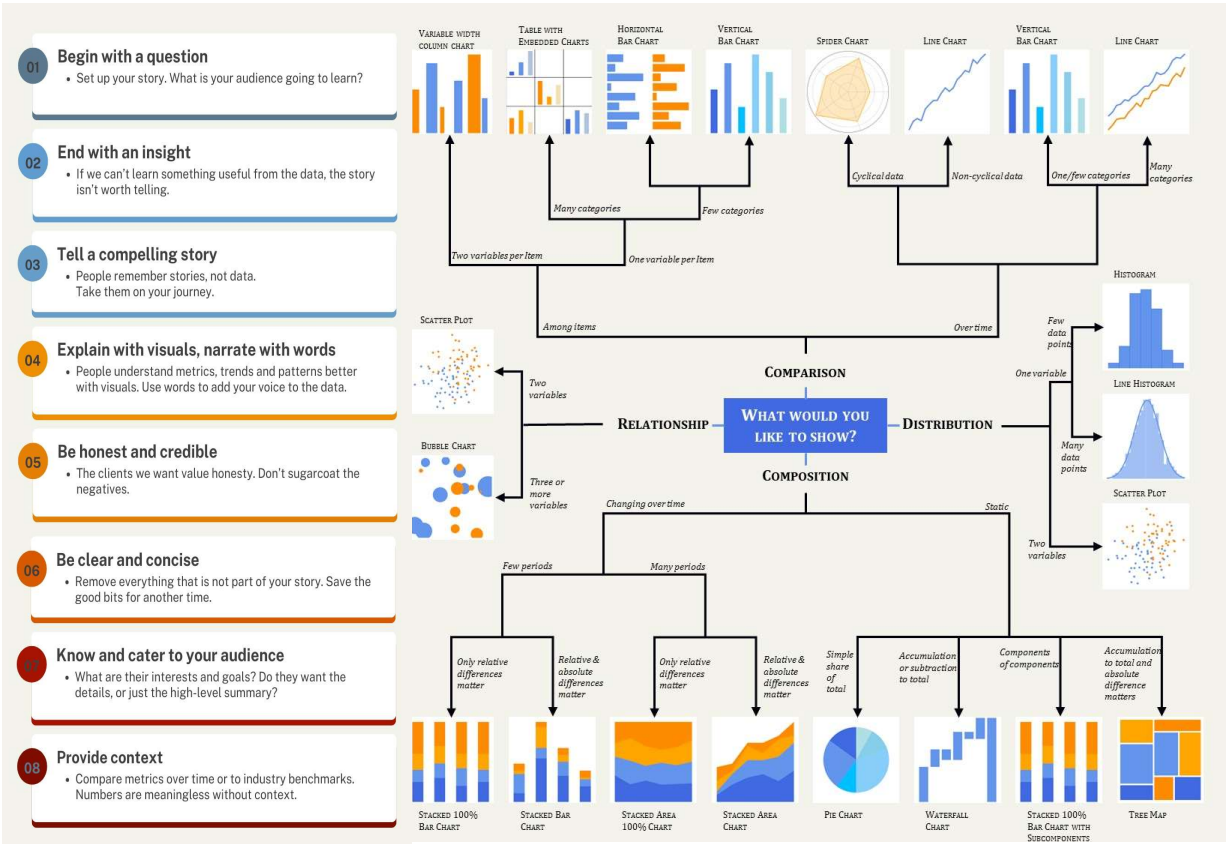
Sales Forecasting with Data Storytelling: The presentation of these forecasts can significantly impact decision-making. Research found that graphical presentations of data improve the accuracy of forecasts compared to tabular formats (Harvey & Bolger, 1996). Similarly, Segel and Heer (2010) highlighted the potential of narrative visualization to enhance the understanding and communication of data stories through adequate visualizations.

3.2.3.2 Choosing Appropriate Visualizations

To effectively utilize data storytelling, several steps are typically followed. These include identifying the story's objective, selecting relevant data sources, conducting EDA (exploratory data analysis), deriving insights, and structuring a coherent narrative

Figure 5

Data Storytelling & Charts Guide



Adapted from Two Octobers (2024) & A. Abela *(2024)*.

Furthermore, the choice of appropriate graphs and visualizations is crucial in delivering information accurately and engagingly (Dykes, 2020). As analyzed in Figure 5, graphs should align with the nature of the data and the message to be communicated, considering factors such as data types, relationships and trends (Iliinsky & Steele, 2011).

Popular tools like Power BI, Tableau, and others provide interactive and user-friendly interfaces for creating visually compelling data stories. These tools offer a range of customizable visualizations and storytelling features that facilitate the creation of engaging narratives and interactive dashboards.

The versatility and effectiveness of data storytelling make it a valuable tool in NT, enabling effective communication of insights, influencing decision-making, and driving positive outcomes. For companies in the Algarve, integrating data storytelling into their operations can enhance marketing strategies, improve customer experiences, and optimize resource management, ultimately contributing to the region's tourism growth and sustainability.

CHAPTER IV - METHODS

The purpose of this work is to develop a sales forecast model for a MTO using historical booking data. This model will attempt to predict the number of seats sold (during a determined frame of time), which will then be multiplied by the predetermined average price per seat. Notably, these prices remain fixed throughout the year, allowing for consistent financial planning.

This chapter follows the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology, which provides a structured approach to data mining and model development. The CRISP-DM process consists of six key phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment (Chapman et al., 1999).

The Business Understanding phase was covered in the previous chapters, where context was given for the studied industry and for sales forecasting. Data Understanding is represented in Section 4.1, Section 4.2 and Section 4.3. Data Preparation is addressed in Section 4.4. The Modeling phase is covered in Section 4.5, which outlines the selection and training of the forecasting models. Evaluation of the models is discussed in Section 4.6, where the performance of each model is assessed. Finally, the Deployment phase is described in Section 4.7, which details how the forecasting model is integrated into Power BI for operational use.

The data used in this study encompasses order records from various tours offered by a MTO in the Algarve region. This includes both booking data and website activity data. The analysis aims to understand booking trends, identify significant sales patterns, and develop a predictive model for forecasting future sales. To safeguard sensitive business information and maintain confidentiality, the company has opted to remain anonymous in this study, given the nature of the data involved.

The entire data analysis, from data extraction and transformation to analysis and modeling, is conducted using Python environment using the main packages *pandas*, *sklearn*, *numpy*, *keras*, *matplotlib*, *seaborn* and *ploty* (for figures)- to conduct the analysis, Jupyter Notebook was used. The results are then visualized using Power BI, ensuring that managers can easily interpret the insights and apply them to operational and strategic decisions.

4.1. Data Collection

The primary dataset (**"*Bookings*"**) used in this study consists of tour bookings (or orders) made over the past eight years. This data was extracted from the booking software of the studied MTO in June 2024, covering the period from January 2017 to June 2024. The dataset includes comprehensive records of all bookings during this timeframe, which form the core of the sales forecast model.

In addition to the primary dataset, three secondary datasets were collected to provide supplementary context and enhance the robustness of the forecast model:

- **"*Airport*": ** Monthly data from January 2017 to June 2024 on arrivals at Faro Airport (Data source: INE, 2024). This dataset serves as a proxy for tourist volume, which may correlate with booking trends.
- **"*Overnight Stays*": ** Monthly statistics from January 2017 to June 2024 on overnight stays in Faro District (Data source: INE, 2024). This data helps gauge accommodation trends that can influence booking patterns.
- **"*Webvisits*": ** Daily records from January 2017 to June 2024 of website visits to the company's site (Data source: Company Data, 2024). This dataset provides insights into online engagement and its potential impact on booking behavior.

By integrating these secondary datasets with the primary booking data, this research aims to achieve a more significant understanding of the factors affecting demand for the company's services, thereby improving the accuracy and reliability of the sales forecast model.

4.2. Extract, Transform, Load (ETL)

4.2.1 Extraction

All datasets used in this study were downloaded in CSV format and imported into Python using *pandas* package (McKinney, 2010), which allows for efficient data handling and preparation for further analysis.

4.2.2 Transformation

The data transformation phase was integral to preparing the raw datasets for analysis and model development. This phase involved several critical steps to ensure the data was clean, consistent, and structured appropriately for the forecasting model.

At the outset of the transformation process, the primary dataset and supporting datasets were described in Table 3.

Table 3

Description of the Primary and Supporting Datasets

Dataset	*Number of rows*	*Number of columns*	*Needs transformation?*
Bookings	81,257	85	YES
Airport	88	2	NO
Overnight Stays	90	2	NO
Webvisits	2,738	2	NO

The three secondary datasets— *"Airport"*, *"Overnight Stays"*, *"Webvisits"*—were assessed for completeness and accuracy. These datasets were straightforward, each containing only two columns, which minimized the complexity of the data, meaning they were ready to use without further transformation.

In contrast, the primary dataset *"Bookings"*, which includes detailed tour bookings information, required comprehensive transformation. This dataset was more complex and contained multiple columns with various types of data.

Checking for Duplicates (Order number)

No duplicates were detected on column *Order number* showing dataset integrity.

Checking for Missing Values

Missing values can significantly impact the accuracy and reliability of data analysis. Therefore, a thorough examination was conducted to identify and address these gaps in the dataset.

The inspection revealed that several columns contained missing values, so for key columns with a high percentage of missing values, the following measures were taken: For key columns with missing values, the following measures were taken:

- **Numerical Columns:** Empty numerical fields were filled with zeroes.

- **Categorical:** Non-numeric columns were analyzed individually to determine the most appropriate method for addressing missing values. Where feasible, missing entries were filled with relevant and correct information.

This approach ensured that the dataset remained focused on critical features while minimizing the impact of incomplete data. Specific attention was given to the handling of key columns, so that essential columns contained complete data, which meant that no rows were removed based on missing values.

Column Transformation

Certain demographic fields were notably empty. This lack of data made impossible the creation of customer profiles. The absence of demographic information can be attributed to the efficiency-oriented practices of sales representatives, who often did not complete these fields during order creation, as well as to, online customers who frequently chose not to disclose personal details. Consequently, these demographic columns were excluded from the feature set for model development.

In addition to addressing missing values, data anonymization was undertaken to protect sensitive information and ensure compliance with privacy standards. Columns containing personal client information were removed, and identifiers related to the company, such as seller names, boat names, and product names, were replaced with generic labels (e.g., "Seller 1", "Boat 1", "Product Kayak"). This step was crucial to anonymize the data and focus on the analysis without compromising confidentiality.

Columns that were not required for analysis were also removed and relevant columns were retained or aggregated. This approach ensured that the dataset focused on the most impactful entries while reducing noise from less relevant data.

Data Formatting

Date columns were converted to datetime format, to standardize date handling, while other columns were formatted as integers for accurate numerical analysis. These adjustments ensured data consistency and accuracy for subsequent processing.

Outlier Detection and Removal

Outliers were addressed in the dataset to ensure a more accurate representation of typical bookings. These values often represented unusually high values due to group

bookings or private tours, which were not typical of the general dataset. For instance, high values in *Order total amount* and *Adult* count columns were primarily due to large group bookings or private tours, which skewed the data.

To address this, outliers were identified and removed using the IQR (Interquartile Range) method. This statistical approach facilitated the exclusion of extreme values, thereby improving the reliability of the results. The IQR method is widely recognized for its efficiency in removing outliers, and research has shown that it has a sustainable impact on improving data correlation with target variables (Mallikharjuna Rao et al., 2023).

New columns

Some columns were created to allow further analysis, these columns are:

- *Booked/Order hour*: Both representing the hour of either *Booked/Order date*;
- *Total PAX* – which represents the sum of number of *Adults*, *Children* and *Infants* of each order;
- *Lead time*: which represents the number of days between the *Order date* and *Booked date*.

4.2.3 Loading

After completing the data extraction and transformation, the prepared datasets were loaded for further examination and modeling. The cleaned and structured data were stored as CSV files to maintain compatibility with the tools used in subsequent stages, including Python and Power BI. This finalized data is now ready for the next phase.

4.3. Exploratory Data Analysis (EDA)

EDA is a critical step in the data analysis process, allowing for a thorough understanding of the datasets used in this study. The main objective of EDA is to identify patterns, detect anomalies, validate assumptions, and explore relationships among variables, all of which are critical for developing a robust and accurate sales forecast model.

Tables 4, 5, 6 and 7 description of the key variables within each dataset used in this study: *"Bookings"*, *"Airport"*, *"Overnight Stays"*, and *"Webvisits"*:

Table 4

Description of the Key Variables - "*Bookings*" Dataset

Variable	*Type*	*Description*	*Source*
Booked Date	Date	Date booking was confirmed.	Tour Operator Software
Order Date	Date	Date order was placed.	Tour Operator Software
Booked Hour	Date	Tour hour booking (GMT).	Calculated
Order Hour	Date	Hour order was placed (GMT).	Calculated
Order Number	Categorical	Unique booking identifier.	Tour Operator Software
Lead Time	Numeric	Days between order and booking.	Calculated
Product	Categorical	Tour or product booked.	Tour Operator Software
Status	Categorical	Order Status	Tour Operator Software
Adult	Integer	Number of adults in booking.	Tour Operator Software
Infant	Integer	Number of infants in booking.	Tour Operator Software
Child	Integer	Number of children in booking.	Tour Operator Software
Total PAX	Integer	Total number of people in booking.	Calculated
Order Total Amount	Numeric	Total booking amount.	Tour Operator Software
Paid to Agent	Numeric	Amount paid to agent.	Tour Operator Software
Cash	Numeric	Amount paid in cash.	Tour Operator Software
Credit Card	Numeric	Amount paid by credit card.	Tour Operator Software
Promo Code	Numeric	Discount from promo code.	Tour Operator Software
Other Payment	Numeric	Combined other payment methods.	Aggregated
Agent	Categorical	Booking agent.	Tour Operator Software
Created by	Categorical	Creator of the booking.	Tour Operator Software
Resource Name	Categorical	Assigned resource for booking.	Tour Operator Software

Table 5

Description of the Key Variables - "*Airport*" Dataset

Variable	*Type*	*Description*	*Source*
Date	Date	Date of passenger data.	Faro Airport Data
Passengers	Integer	Number of airport passengers.	Faro Airport Data

Table 6

Description of the Key Variables - "*Overnight Stays*" Dataset

Variable	*Type*	*Description*	*Source*
Date	Date	Date of overnight stay data.	Faro District Data
Overnight Stays	Integer	Number of overnight stays.	Faro District Data

Table 7

Description of the key variables - "*Webvisits*" dataset

Variable	*Type*	*Description*	*Source*
Date	Date	Date of website visit data	Company Data
Website Visits	Integer	Number of visits to the website	Company Data

4.3.1 Descriptive Statistics

Descriptive statistics provide a foundational understanding of the data being studied. This section presents an overview of the descriptive statistics for the key variables in the datasets.

Supporting Datasets

The supporting datasets provide valuable context for understanding the broader environment in which the main dataset operates. These datasets include variables related to general tourism metrics such as number of *Passengers*, *Overnight Stays*, and *Webvisits*. Analyzing these variables can help identify trends in tourism demand and web engagement over time, complementing the analysis of the main booking data.

Regarding the monthly *"Airport"* dataset (Table 8), 88 observations were detected, from 1 March 2017 to 1 June 2024. The mean number of *Passengers* per period is approximately 302,239, with a minimum of 51 and a maximum of 641,550 *Passengers*. The IQR shows a broad distribution, with 50% of the data falling between 130,606 and 506,708 *Passengers*. This wide range indicates varying levels of passenger traffic over time, reflecting Algarve's seasonal peaks.

Table 8

Summary Statistics - "*Airport*"

Variable	*count*	*mean*	*min*	*25%*	*50%*	*75%*	*max*
Date	88	2020-10-15	2017-03-01	2018-12-24	2020-10-16	2022-08-08	2024-06-01
Passengers	88	302,238.86	51.00	130,605.75	223,674.00	506,707.75	641,550.00

The monthly *"Overnight Stays"* dataset (Table 9) covers a similar period, covering 90 observations, also spans from 1 January 2017 to 1 June 2024. The mean number of *Overnight Stays* is 1,428,524, with a minimum of 25,614 and a maximum of 3,439,271.

Table 9

Summary Statistics – "*Overnight Stays*"

Variable	*count*	*mean*	*min*	*25%*	*50%*	*75%*	*max*
Date	90	2020-09-15	2017-01-01	2018-11-08	2020-09-16	2022-07-24	2024-06-01
Overnight Stays	90	1,428,523.76	25,614.00	590,794.00	1,216,942.50	2,087,024.50	3,439,271.00

"Webvisits" daily data (Table 10), collected over 2,738 observations from January 1, 2017, to June 30, 2024, indicates a mean of 508 visits per period. The dataset reveals considerable variability, with visits ranging from 62 to 4,054. Median web traffic hovers around 421, suggesting that while typical periods see moderate engagement, certain times experience significant spikes, likely driven by marketing efforts or increased interest in particular events.

Table 10

Summary Statistics - "*Webvisits*"

Variable	*count*	*mean*	*min*	*25%*	*50%*	*75%*	*max*
Date	2738	2020-09-30	2017-01-01	2018-11-16	2020-09-30	2022-08-15	2024-06-30
Webvisits	273	508.09	62.00	261.00	420.50	715.00	4,054.00

After reviewing the supporting datasets, the analysis proceeded to the main dataset, *"Bookings"*, beginning with an exploration of the distribution of booking statuses for each order.

Distribution of Booking Status

The Status variable identifies the state of each booking, with categories such as *CONFIRMED*, *CANCELLED*, and *ABANDONED_CART*, the latter referring to incomplete online bookings. By examining the distribution of *Status* (Table 11), we gain a clearer understanding of both customer behavior and the efficiency of the booking process.

In this dataset, most bookings (89.15%) are confirmed, 9.48% have been cancelled, and 1.37% were abandoned during the online booking process. Although nearly 1 in 10 bookings are cancelled, this is likely attributable to the MT industry, where factors like weather conditions can significantly impact decisions to proceed with or cancel tours.

Table 11 – Distribution of Booking Status

Distribution of Booking Status

Status	*Frequency*	*Percentage*
CONFIRMED	67,983	89.15%
CANCELLED	7,229	9.48%
ABANDONED_CART	1,048	1.37%
Total	76,260	100.00%

To ensure the analysis accurately reflects sales performance, the dataset was filtered to include only *Status* = *CONFIRMED*. This filtered dataset will be used for all subsequent analyses, ensuring that the insights are based on finalized transactions.

Summary Statistics for Date Variables

The dataset includes date-related variables such as *Booked Date*, *Order Date*, *Booked Hour*, and *Order Hour*. These variables provide insights into the timing and distribution of bookings over the period covered by the data.

***Booked Date*:** Ranging from 12 April 2017 to 30 June 2024, with a mean date of 22 May 2021, the booking dates are well-distributed across this span. This suggests a broad temporal coverage, enabling the analysis of long-term trends in booking behavior.

***Order Date*:** Spanning from 11 April 2017 to 30 June 2024, the order dates have a mean of 17 May 2021. The close alignment of mean order and booked dates indicates that bookings are generally taken soon after orders are placed, reflecting efficient customer conversion.

***Booked Hour* and *Order Hour*:** The most frequent booking hour is 11:15 AM, while orders are most placed around 12:22 PM. The wide range of booking hours suggests varied customer engagement times, which could be influenced by factors such as time zones or customer demographics.

***Lead Time*:** This variable measures the number of days between the order date and the booked date, with a mean of 4.27 days. *Lead Time* ranges from 0 (min.) to 67 days (max), with at least 50% of values being between 0 and 1 days, indicating that most bookings are made within a short period following the order.

As shown in Table 12, the date variables demonstrate a well-spread dataset over several years, with a concentration in recent years, suitable for identifying both long-term and seasonal trends in customer behavior.

Table 12

Summary Statistics for Date Variables

Variables	*unique*	*top*	*freq*	*mean*	*min*	*25%*	*50%*	*75%*	*max*
Booked Date	1,536	2023-08-09	172	2021-05-20	2017-04-12	2019-07-26	2021-09-23	2023-05-06	2024-06-30
Order date	1,919	2022-08-02	177	2021-05-16	2017-04-11	2019-07-22	2021-09-21	2023-05-01	2024-06-30
Booked Hour	93	11:15	10,481						
Order Hour	39,726	12:22	10						
Lead time	86	1	21,195	4.27	0	1	1	3	67

Summary Statistics for Numerical Variables

Numerical variables in the dataset were analyzed to understand their distribution. Table 13 presents the key descriptive statistics for these variables.

The number of adults per booking averages around 2.50, with a standard deviation of 1.32, indicating a typical booking involves a moderate number of adults but can vary widely. Children and infants are less common in bookings, with average numbers of 0.22 and 0.05, respectively. The variability in these figures suggests occasional family bookings with multiple children or infants. *Total PAX*, which represents the total number of passengers per booking, has a mean of 2.77 and a standard deviation of 1.49, indicating most bookings involve a small group of passengers.

Table 13

Summary Statistics for Numerical Variables

Variable	*mean*	*std*	*min*	*25%*	*50%*	*75%*	*max*
Adult	2.50	1.32	0.0	2.0	2.0	3.0	17.0
Child	0.22	0.62	0.0	0.0	0.0	0.0	5.0
Infant	0.05	0.16	0.0	0.0	0.0	0.0	2.0
Total PAX	2.77	1.49	0.0	2.0	2.0	4.0	18.0
Order total amount	90.37	55.17	0.0	56.0	73.5	112.0	504.0
Paid to agent	42.83	60.08	-60.0	0.0	0.0	70.2	400.0
Cash	11.44	28.71	0.0	0.0	0.0	0.0	196.0
Credit card	16.60	39.93	0.0	0.0	0.0	0.0	274.0
Promo code	3.72	8.75	0.0	0.0	0.0	0.0	60.0
Other Payment	2.74	19.84	-456.0	0.0	0.0	0.0	504.0

Examining the financial data shows that an average order total amount is €90.37, with a standard deviation of €55.17, reflecting a broad range of booking values from €0 to

€504. Payments to agents average €42.83, with a significant variability shown by a standard deviation of €60.08. *Cash* payments have an average of €11.44, while *Credit Card* payments average €16.60. Both types of payments exhibit considerable variability, with cash ranging up to €196 and credit card payments up to €274. Discounts from promo codes average €3.72, and other payments, which include various payment methods, have a mean of €2.74. The broad range of other payments reflects some significant adjustments and refunds.

Summary Statistics for Categorical Variables

Categorical variables in the dataset reveal significant insights into the booking process, including details on products, resources, agents, and creators (Table 14).

For *Product* categories, 'Kayak Tour' emerges as the most popular option, with 17,744 bookings. This is closely followed by "Coast Cruise 1" with 13,930 bookings. Other notable products include "Grotto Tour 1" with 11,709 bookings and 'Coast Cruise 2' with 10,699 bookings. This distribution highlights a strong preference for certain types of tours, which could be crucial for targeted marketing and product development strategies.

Regarding *Resource* utilization, 'Catamaran 1' leads with 24,885 bookings, indicating its significant role in the tour operations. 'Kayak' is also a popular choice, with 17,748 bookings, while 'Catamaran 2' and 'Grotto 1' have 5,891 and 4,915 bookings respectively. These results suggest a clear hierarchy in resource booking, which can be explained by tour popularity and resource capacity.

A tour can be booked through three sources: 'MARKETPLACE_PREF_RATE' indicates that the booking was made through an external seller or agent, 'INTERNAL' refers to bookings made through the company's own sellers, and 'ONLINE' represents bookings made through the company's website. Most bookings are made through external agents.

In terms of booking agents, 'NONE' (meaning it's an internal sale) is the most common category, representing 25,732 bookings. 'GetYourGuide' follows with 23,538 bookings, showing its strong market presence. 'Viator' and 'other' categories account for 8,548 and 7,988 bookings respectively, underscoring the dominance of specific booking platforms and their importance in the overall booking strategy.

Creator categories show that 'EXTERNAL' is the leading contributor with 45,821 bookings, while 'other' accounts for 12,061. Specific sellers such as 'Seller 2' and 'Seller 4' have 4,003 and 2,169 bookings respectively. This distribution emphasizes the predominant role of external sources in the booking process, along with the varying impact of different sellers.

Table 14

Summary Statistics for Categorical Variables

Variable	*Count*	*Unique*	*Top Counts (4)*
Product	67,983	15	"Kayak Tour": 17,744, "Coast Cruise 1": 13930, "Grotto Tour 1": 11,709, "Coast Cruise 2": 10,699
Resource Name	67,983	9	"Catamaran 1": 24,885, "Kayak": 17,748, "Catamaran 2": 5,891, "Grotto 1": 4,915
Source	67,983	3	"MARKETPLACE_PREF_RATE": 3,3813, "INTERNAL": 2,2158, "ONLINE": 12,012
Agent	67,983	6	"NONE": 25,732, "GetYourGuide": 23538, "Viator": 8,548, "other": 7,988
Created by	67,983	6	"EXTERNAL": 45,821, "other": 12,061, "Seller 2": 4,003, "Seller 4": 2,169

4.3.2 Data Visualization

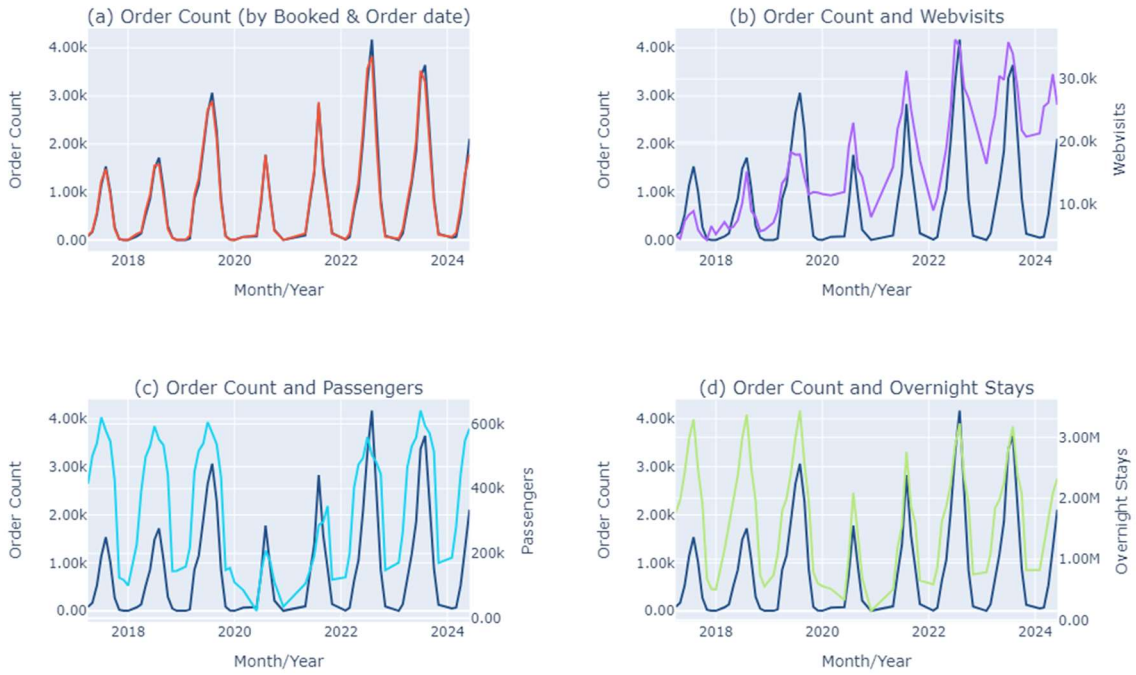
Data visualization is a crucial step in EDA, enabling the identification of patterns, trends, and anomalies within the dataset. Through various visual representations, complex data can be more easily interpreted, facilitating better decision-making. This section outlines the key visualizations used to explore the dataset.

Temporal Patterns

Time series analysis is essential for identifying trends, seasonality, and cycles in booking data. This analysis helps in understanding how bookings fluctuate over time (Figure 6).

Figure 6

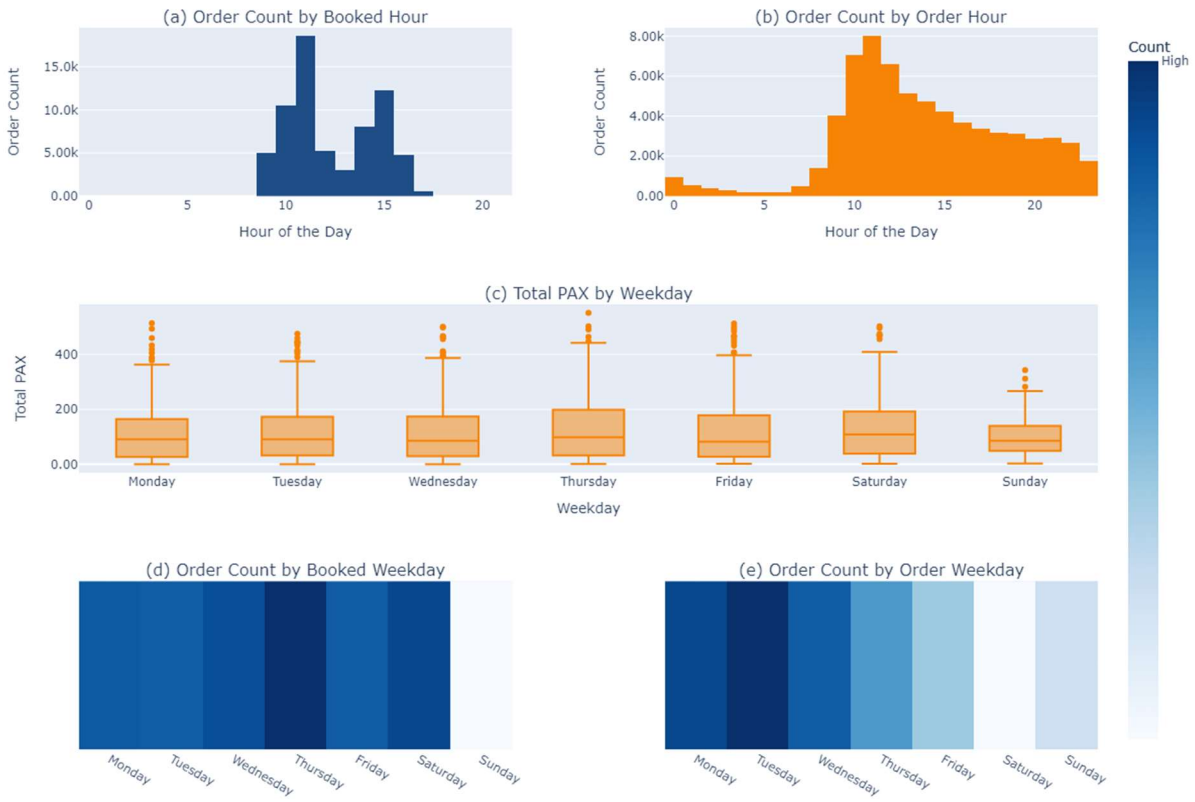
Comparison of Different Metrics Over Time



The visualizations depict a consistent, cyclical pattern across several metrics, including a - order counts (by both *Booked* and *Order Date*), b- *Webvisits*, c - *Passengers* arrivals at Faro Airport, and d - *Overnight Stays* from 2017 to 2024. Each metric exhibits recurring peaks, particularly aligning with high-demand seasons (from June to August), demonstrating a clear seasonality in these datasets. The strong correlation between order counts and external factors like webvisits, passenger volume, and overnight stays suggests that these external variables may be key drivers or indicators of order volume fluctuations. This indicates that these metrics can be predictive of future performance, helping in forecasting and resource planning.

Figure 7

Hourly Patterns of Bookings and Orders



In Figure 7a, located in the top left corner, reveals that bookings consistently peak between 11 AM and 3 PM. This surge indicates a high booking rate for tours during the late morning hours and right after lunch. In contrast, graph 7b demonstrates a steady flow of orders throughout the day, with a notable peak around lunch hour.

When analyzing the center box plot (graph 7c), we observe the distribution of total PAX per weekday. While the overall pattern remains consistent throughout the week, Fridays and Saturdays exhibit slightly higher outliers.

Lastly, the bottom heatmaps (7d and 7e) provide valuable insights into the distribution of order counts by weekday. Thursdays and Fridays emerge as the busiest days for bookings, while actual order fulfillment tends to peak earlier in the week.

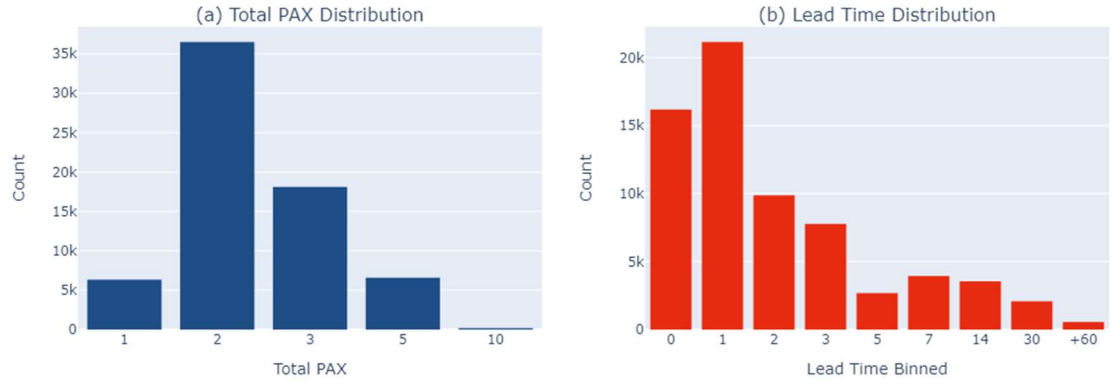
In summary, the data highlights key behavioral patterns among customers. Most tours booked happen during late-morning hours, with the actual order peak is around lunch

hour. Thursdays and Fridays stand out as popular booking days, while order fulfillment is more evenly distributed across the entire week.

Distribution of Operational Variables

Figure 8

Distribution of *Total PAX* and *Lead Time*



Upon analyzing the Total PAX distribution, it's observed that most orders involve small groups of 2. Following closely are orders with a PAX size of 3. In contrast, orders with only 1 PAX or larger groups occur less frequently. Regarding Lead Time, it's noticed that most orders have a lead time of 1 day or less. The next common lead times are 2 or 3 days. Beyond 7 days, there is a sharp decline in the number of orders, with very few having lead times exceeding 60 days (Figure 8).

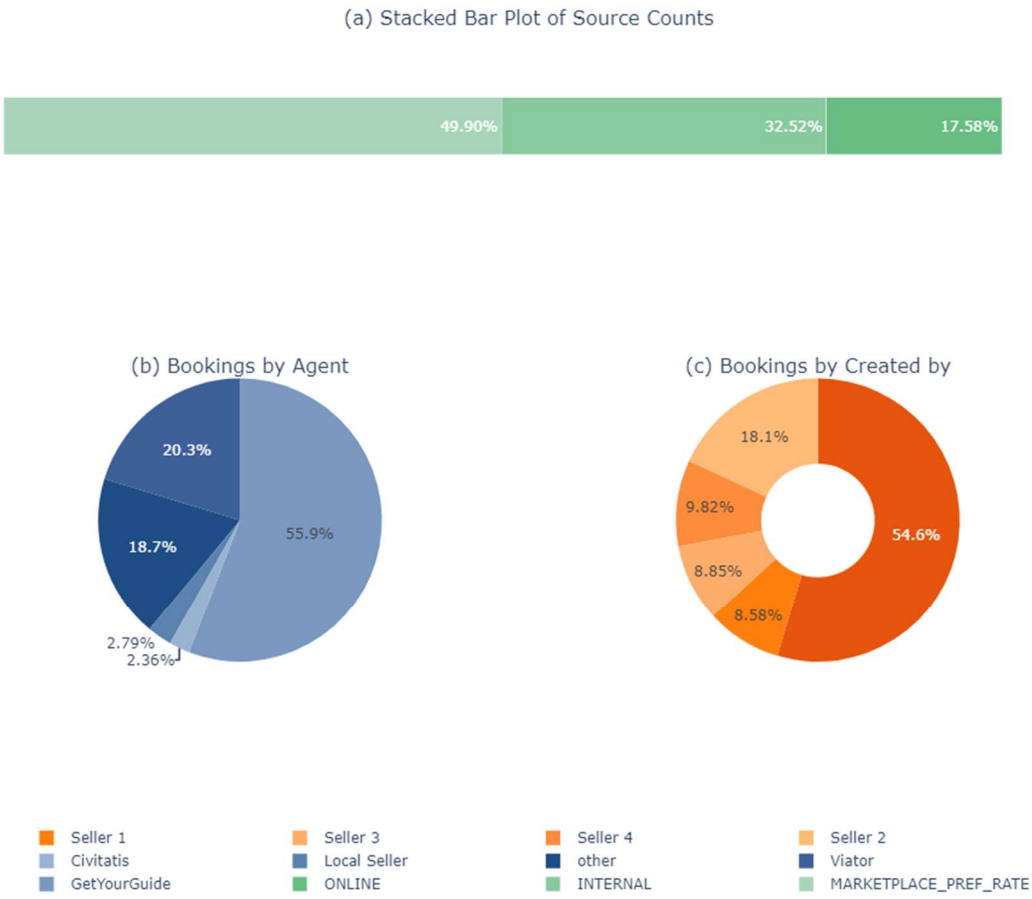
This means that orders primarily consist of small groups (typically 2-3 people), and most bookings occur with short lead times (0-1 day). This pattern suggests that customers often make reservations with minimal advance notice, possibly reflecting spontaneous or last-minute booking behavior.

The dataset contains several distinct sources of bookings. The most common sources include MARKETPLACE (49.90%), INTERNAL (33.52%), and ONLINE (17.58%). This distribution indicates that marketplaces and direct salesperson bookings are the primary drivers of sales (Figure 9a).

In Figure 9b the pie chart on the left breaks down marketplace bookings by agent, where GetYourGuide is the dominant agent, responsible for 55.90% of bookings. Viator ranks second at 20.30%, with other agents, including Civitatis and smaller sellers, completing the distribution.

Figure 9 - *Source*, *Agent* and *Created by* Distribution

Source, Agent and *Created by* Distribution



On the right (Figure 9c), the donut chart focuses on the distribution of internal bookings by company seller. Seller 2 is the most influential, driving 18.10% of all internal bookings, followed by Seller 4 with 9.82%. The top 4 sellers represent almost half of all internal sales.

This analysis highlights the significant influence of GetYourGuide and Seller 2 in the booking process, with Viator and Seller 4 also playing crucial roles.

Figure 10

Total PAX by *Resource Name*

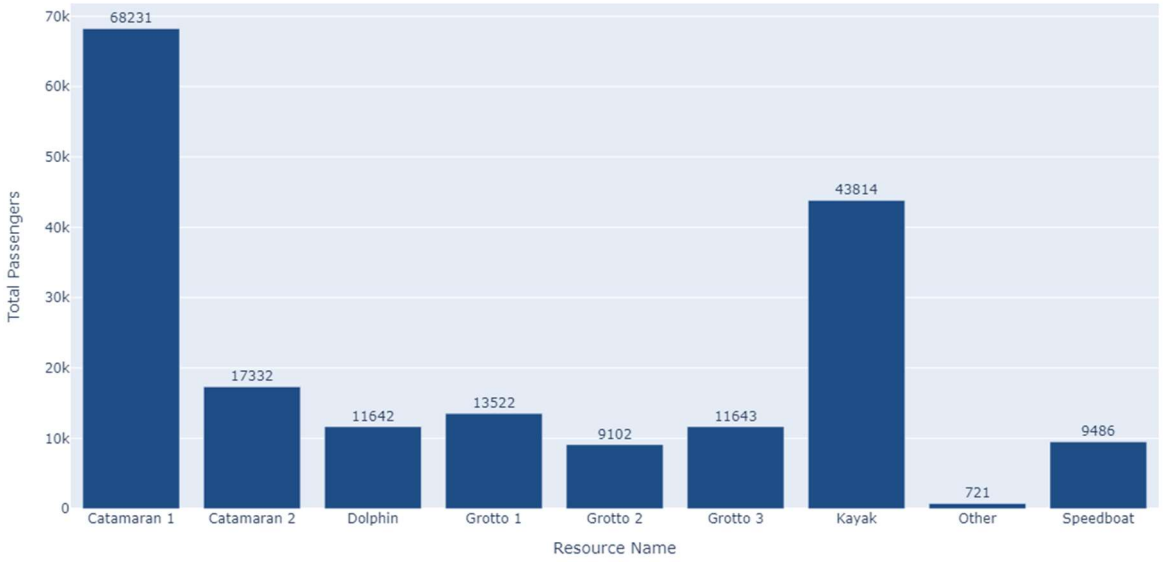
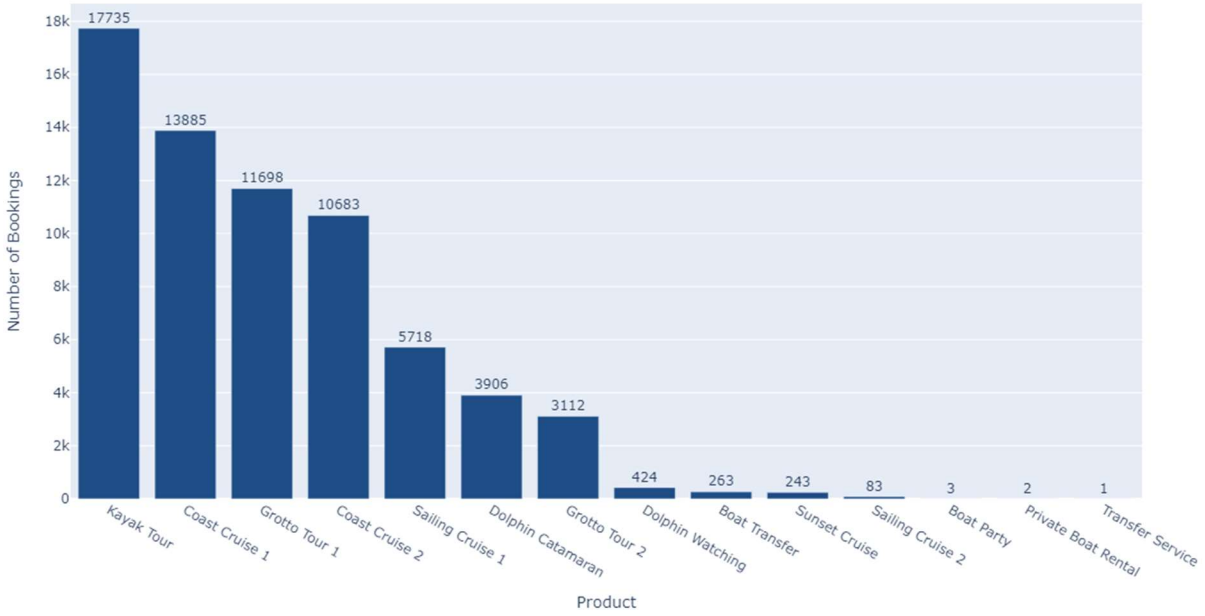


Figure 10 illustrates the total number of passengers (PAX) for each resource name, indicating the popularity of various services or excursions. Catamaran 1 shows the highest passenger count at 68,231, highlighting its significant contribution to total activity. This is followed by Kayak, with 43,814 passengers, which also represents a substantial portion of the activity. Other resources, such as Catamaran 2 and Grotto 1, contribute 17,332 and 13,522 passengers, respectively, while resources like Dolphin, Grotto 2, and Speedboat have smaller but still notable figures, ranging from 9,102 to 11,642 passengers. Other resources, at 721 passengers, demonstrate minimal engagement, signifying their marginal role in overall operations. This distribution emphasizes the predominance of certain activities over others, particularly the dominance of Catamaran 1 and Kayak.

Figure 11

Booking Count by *Product*



The bar chart depicted in Figure 11 shows the number of bookings across different products, providing insight into the most popular services. The Kayak Tour is the leading product with 17,735 bookings, significantly ahead of other offerings. Coast Cruise 1 and Grotto Tour 1 follow with 13,885 and 11,698 bookings, respectively, further indicating their appeal to customers. Meanwhile, products such as Coast Cruise 2 and Sailing Cruise 1 garner moderate interest, with 10,683 and 5,718 bookings, respectively. Products like Dolphin Catamaran, Grotto Tour 2, and Dolphin Watching exhibit smaller numbers, with 3,906 and fewer bookings, indicating less popularity compared to the top services. Less common products, including Sunset Cruise and Sailing Cruise 2, are characterized by very low booking counts, under 100. This distribution reflects significant disparities in product demand, with a few tours dominating customer interest while others see limited engagement.

Financial Variables Visualization

Figure 12 depicts the monthly total order amount in blue and the average order amount in yellow from 2017 to 2024. The total order amount exhibits significant fluctuations with notable peaks around the end of each year, which could correlate with seasonal demand spikes such as holiday seasons. The average order amount shows a more stable trend, yet it also peaks during these high-demand periods, indicating a

potential increase in the value of orders during these times. The cyclical pattern observed in both metrics suggests a strong seasonality in the business, where certain times of the year consistently drive higher sales volumes.

Figure 12

Monthly *Total Order Amount* and Average Order Amount

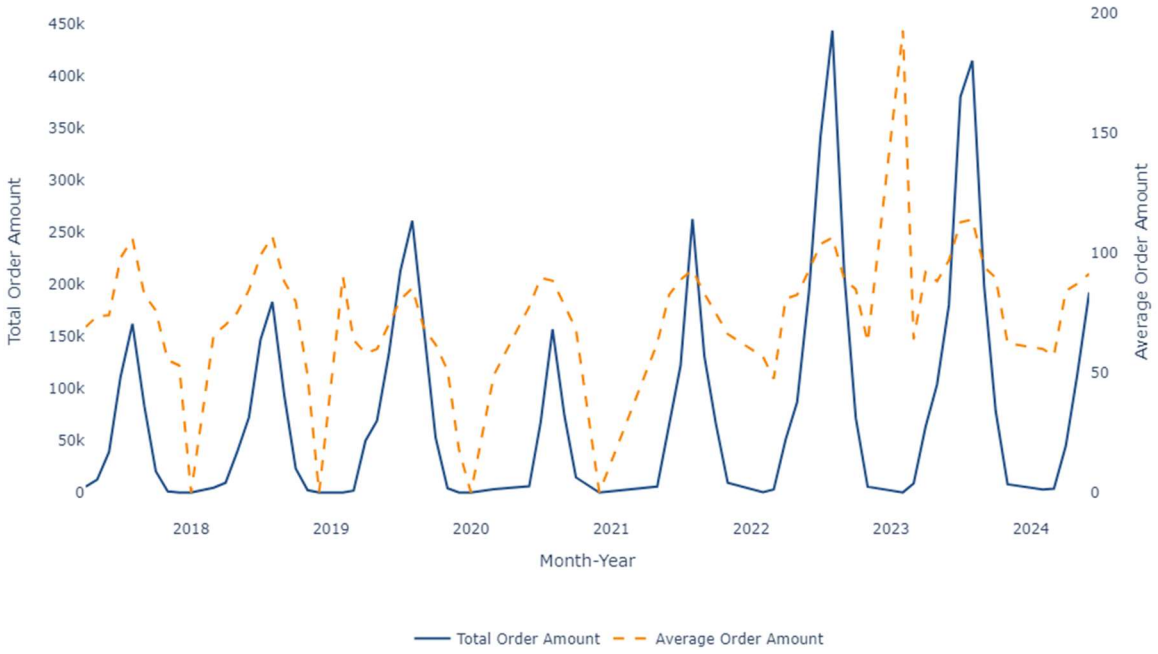
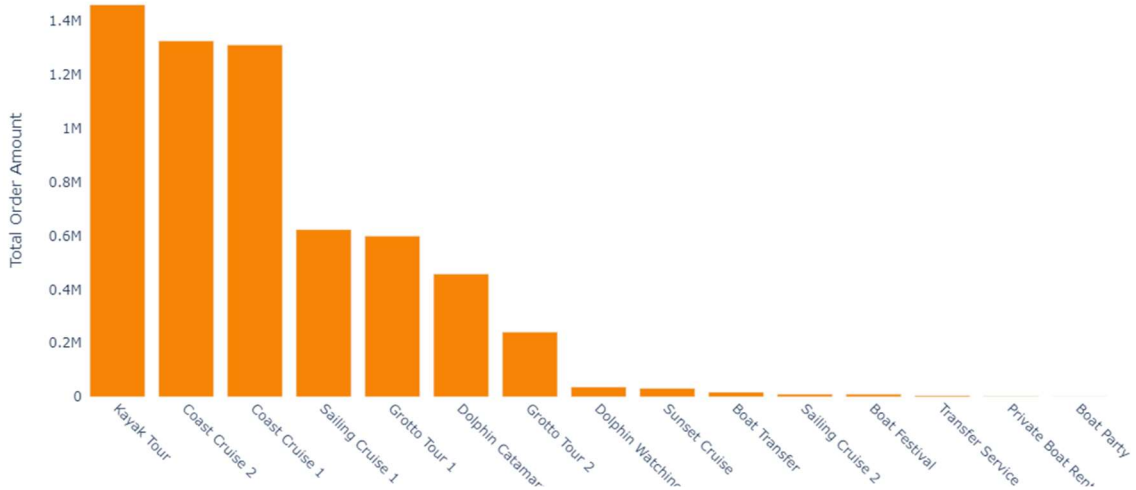


Figure 13 - *Total Order Amount* by *Product*

Total Order Amount by *Product*



The bar plot in Figure 13 represents the total order amount segmented by different products. The heights of the bars indicate the cumulative sales generated by each product over the analyzed period. The data suggests that certain products significantly outperform others, contributing more to the overall revenue. The variation in total order

amounts across products emphasizes the differing levels of demand and popularity among the product offerings. This visualization provides a clear indication of which products are driving the majority of sales, allowing for strategic insights into product performance and inventory management.

Figure 14 showcases the total order amount categorized by resource name, providing insight into the performance of different resources in generating sales. The varying heights of the bars highlight the differing contributions of each resource to the overall revenue. Some resources clearly dominate in terms of total sales, indicating either higher demand or more effective utilization. This distribution underscores the importance of these key resources, potentially guiding future resource allocation and optimization strategies to maximize revenue.

Figure 14

Total Order Amount by *Resource*

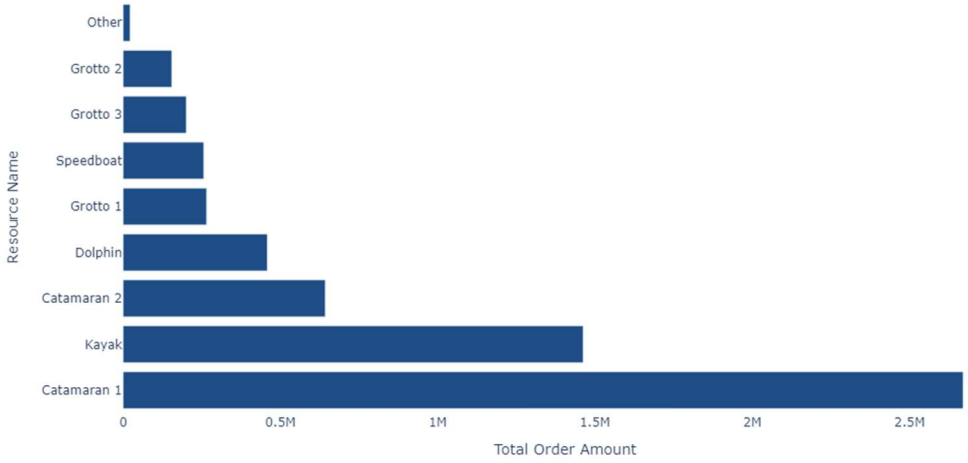


Figure 15

Distribution for Different Payment Methods

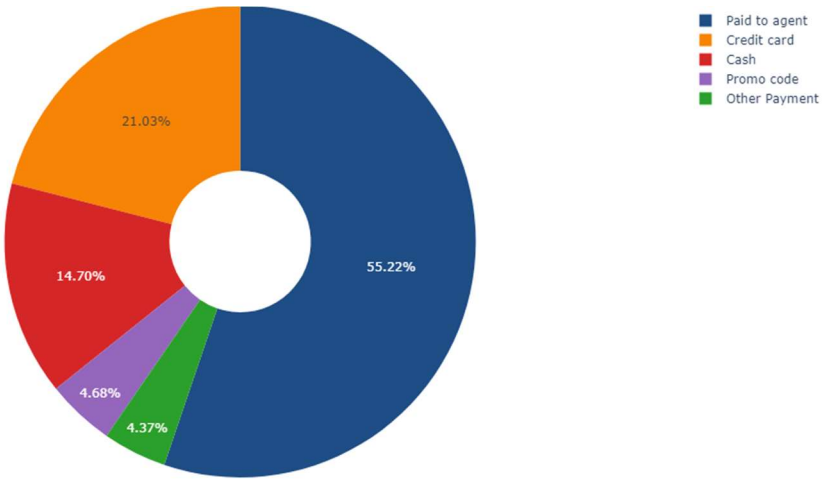


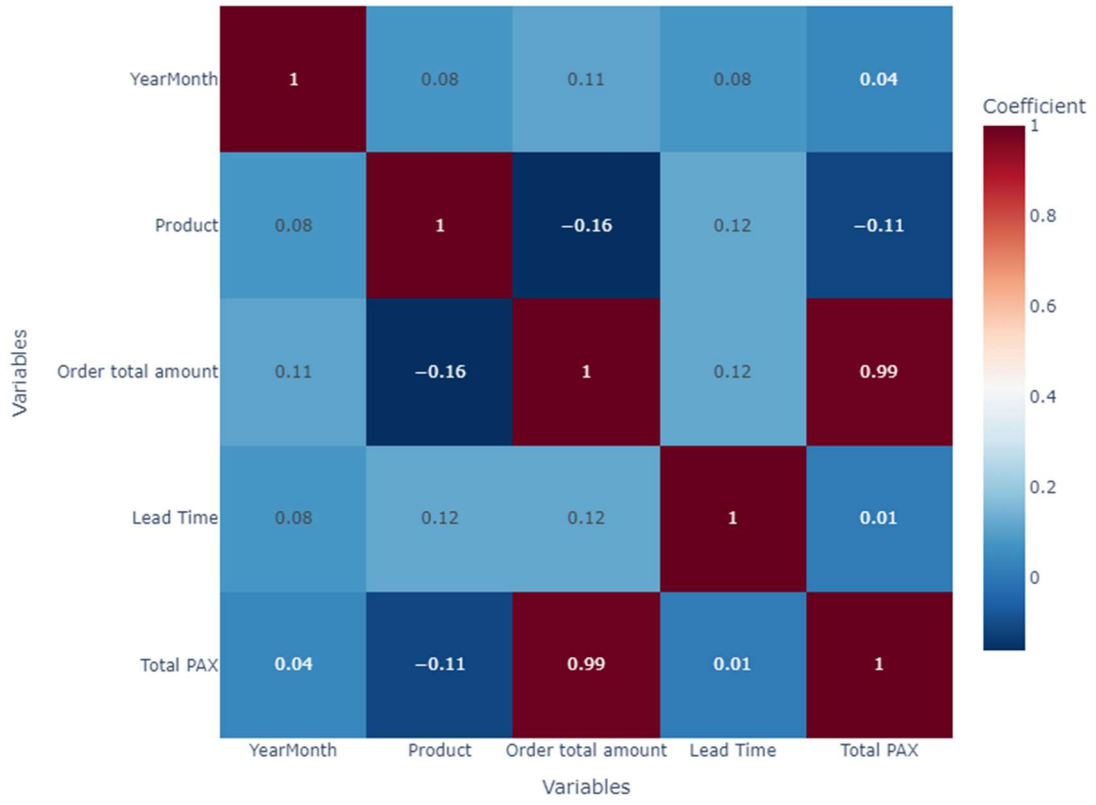
Figure 15 illustrates the distribution of payment methods used by customers. A substantial majority, 55.22%, prefer payments made to agents, indicating a strong reliance on intermediaries or agents in the transaction process. Credit card payments account for 21.03%, followed by cash at 14.70%, which suggests that while digital transactions are gaining traction, cash still holds a significant share. The remaining payment methods, including promo codes and other forms, constitute a smaller fraction, 4.68% and 4.37% respectively, underscoring the less frequent use of alternative payment options. This distribution highlights the diverse payment preferences of the customer base, with a pronounced tendency towards traditional and agent-mediated transactions.

Correlation Analysis Visualization

The correlation matrices shown in Figures 16 and 17 provide insights into the relationships between various metrics in the dataset. *Order Total Amount* shows a strong positive correlation of 0.99 with *Total PAX*, naturally higher passenger counts are associated with increased order values.

Figure 16

Correlation Matrix – "*Bookings*"



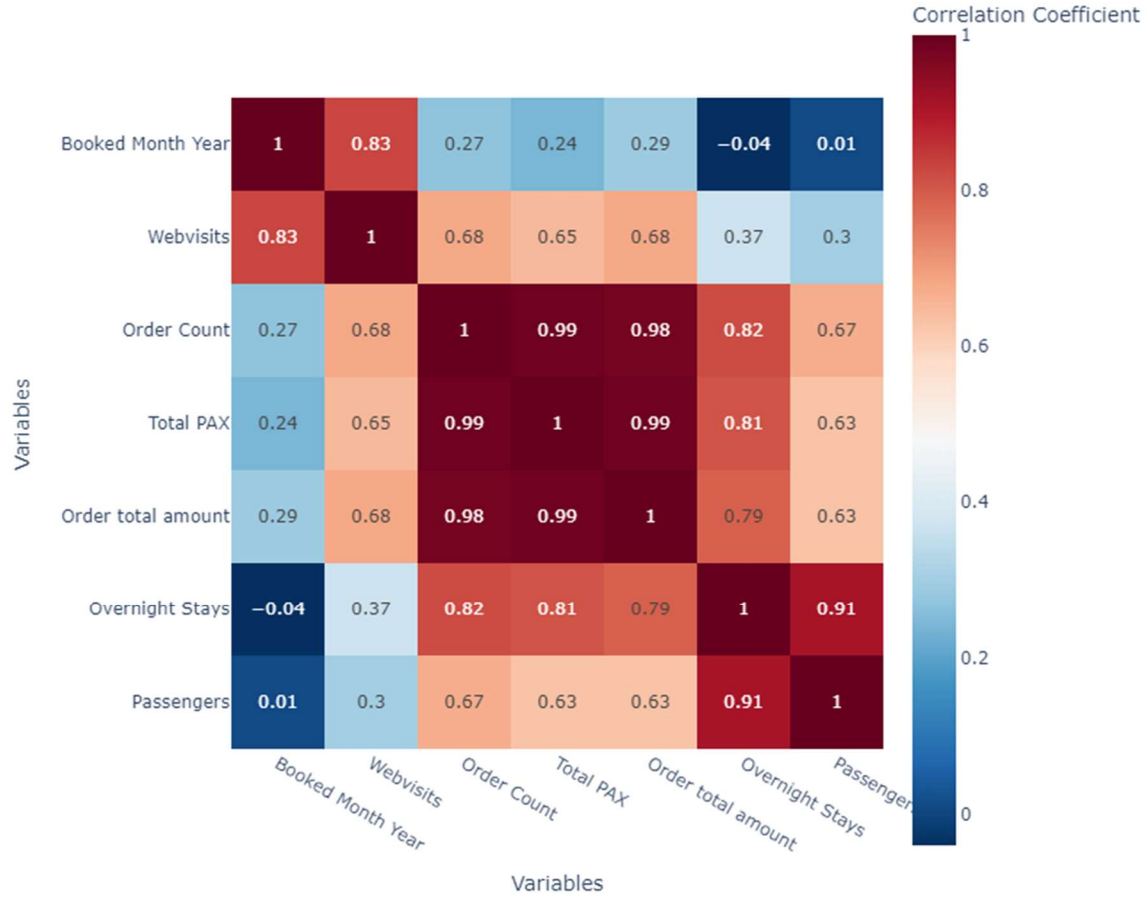
On the other hand, Product displays a negative correlation of -0.16 with *Order Total Amount*, suggesting that changes in product categories have a minor inverse impact on the total order value. This could imply that certain products might be less impactful on overall revenue.

Lead Time shows negligible correlations with other variables, including Order Total Amount and Total PAX, suggesting that the time between booking and the actual service does not significantly influence the order amount or passenger count.

Overall, these correlations highlight key relationships, particularly the strong link between order total and passenger count, while suggesting that lead time have minimal effects on the order values.

Figure 17

Correlation Matrix – Supporting Datasets



The correlation matrix provides a comprehensive view of the linear relationships between various key metrics. Notably, 'Webvisits' and 'Order Count' exhibit a strong positive correlation of 0.68, signifying that increased web traffic is closely associated with

higher order volumes. Similarly, 'Webvisits' correlates strongly with 'Order Total Amount' at 0.68, suggesting that higher web engagement translates to greater revenue.

The relationship between 'Total PAX' and 'Order Total Amount' is strong at 0.99, underlining that the total number of passengers is a significant driver of total order revenue. Additionally, Overnight Stays and 'Passengers' demonstrate a high correlation of 0.91, suggesting that variations in sleep patterns are closely linked to passenger numbers.

Overall, these correlations highlight critical patterns in the data, emphasizing the impact of web traffic and passenger volume on both order counts and revenue, and the interconnected nature of these variables in the context of bookings and sales performance.

4.4. Feature Engineering

Feature Selection

The initial feature selection was conducted through a correlation matrix analysis combined with domain expertise, ensuring that only the most relevant and impactful variables were retained. This approach allowed the model to focus on core drivers of sales. Key features such as historical sales data, resource availability (e.g., boat or seat capacity), and time-related variables (e.g., day of the week, month) were identified as important. These variables are naturally strong predictors as they capture both seasonal trends and operational constraints that directly influence the model.

Feature Creation

To improve the model's understanding of sales dynamics, new features were created. One significant addition was the total number of seats available feature. This new feature (*Total seats*) accounts for the number of seats available at any given time, reflecting the company's capacity and growth potential. By incorporating this feature, the model can better capture variations in sales due to changes in operational scale, such as the acquisition of new assets or expansion of services.

In addition, time-related features (such as month, season, week number etc.) were introduced to enhance the model's ability to recognize patterns over time. A key feature created was the *Workday* feature, which checks whether seats were available on a specific day. If the number of available seats was greater than 0, *Workday* was set to 1; otherwise, it was set to 0. This feature enables the model to track the number of days tours were

operational within a given week or month, thereby improving its ability to capture the actual frequency of tour occurrences. This is especially valuable for the weekly and monthly datasets, as it could help identify periods of downtime or operational adjustments that influence sales.

Moreover, three distinct datasets were constructed for daily, weekly, and monthly forecasting, enabling the model to analyze sales patterns at multiple time scales. This approach allows for better detection of trends and seasonality at different levels of granularity. For instance, monthly datasets focus on long-term trends and macroeconomic effects, which may not be apparent in daily or weekly data. This hierarchical data structuring aligns with best practices in time series forecasting, which suggest that models should incorporate different time frequencies to capture both short-term fluctuations and long-term patterns (Hyndman & Athanasopoulos, 2021).

Additionally, lag features were introduced to capture temporal dependencies in the data. Lag intervals, such as weekly, monthly, and annual lags were tested. This is a common practice in time series forecasting as it helps models learn from past behaviors (Shmueli & Lichtendahl Jr., 2016).

Feature Scaling

Normalization was applied to standardize features across different scales. For instance, features with widely varying scales, such as the month number (ranging from 1 to 12) and total seats sold (in thousands), were normalized using Min-Max Scaling. It transforms a feature to a range of [0,1] based on the following formula:

$$X_{\text{scaled}} = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \quad (1)$$

where $X_{\min}$ and $X_{\max}$ represent the minimum and maximum values of the feature. This normalization process ensured that all features contributed equally to model training, improving overall model performance (Han et al., 2012).

In order to interpret the model's predictions in the original units of the features, for the model evaluation phase, the data was reverse normalized to its original scale using the inverse of the Min-Max Scaling transformation.

4.5. Model Development

The selection and development of predictive models for sales forecasting are critical to ensuring the accuracy and robustness of the forecasts. The model development process involves selecting suitable ML algorithms, training and evaluating their performance. This section outlines the criteria for model selection, provides an overview of the chosen models (Table 15), and details the model training and evaluation procedures.

4.5.1 Model Selection

Selecting the right model is crucial for addressing the unique challenges of sales forecasting. For this study, XGB, Random Forest (RF), and Support Vector Regression (SVR) were chosen based on their proven effectiveness since these models are commonly used for predictive tasks and have demonstrated strong performance in the forecasting domain (Atanda *et al*., 2024; Jain *et al*, 2015; Dairu & Shilong, 2021; Lin & Lee, 2013).

The selection was based on the following considerations:

- **Non-linearity**: The ability of the model to capture complex relationships between data and external factors.
- **Seasonality**: The model's ability to account for seasonal trends in sales data.
- **External Features**: The flexibility to integrate additional variables present in supporting datasets, such as district airport arrivals, district overnight stays and company website visits, which are crucial for improving forecast accuracy.

XGBoost (Extreme Gradient Boosting):

XGB is a powerful gradient boosting framework that sequentially builds models to minimize residual errors. Its ability to handle structured and sparse data, while incorporating L1 and L2 regularization, makes it ideal for preventing overfitting in sales forecasting (Chen & Guestrin, 2016). Research consistently demonstrates that XGB outperforms traditional models in sales forecasting by capturing complex patterns and improving prediction accuracy (Atanda et al., 2024). Its iterative approach, which focuses on correcting errors in each subsequent model, enhances its effectiveness in predictive modeling (Suryawanshi et al., 2024).

Random Forest (RF):

Random Forest (RF) is a widely recognized ensemble learning model that builds multiple decision trees and aggregates their predictions to enhance its resilience and capacity for generalization (Suryawanshi et al., 2024). Its robustness makes it especially suitable for sales forecasting where patterns can be erratic due to seasonality or unpredictable market behavior (Huang et al., 2018).

Support Vector Regression (SVR):

Support Vector Regression (SVR) is a powerful machine learning technique that maps input data into a high-dimensional space using kernel functions, enabling it to capture complex, non-linear relationships. The model penalizes complexity using a regularization term, which balances the trade-off between model flexibility and generalization (Schölkopf & Smola (2001); Schölkopf & Smola (2003)). This makes SVR particularly effective for sales forecasting where demand patterns often exhibit non-linear behavior (Honga et al., 2011). In tourism forecasting, SVR has been successfully applied to model complex sales relationships with high accuracy (Claveria et al., Regional Forecasting with Support Vector Regressions: The Case of Spain. ERN, 2015). Though computationally intensive, its flexibility with different kernel functions, such as Gaussian kernels, makes SVR suitable for medium and long term forecasting (Claveria et al., 2016).

Table 15

Overview of Selected Models

Model	*Methodological Foundations*	*Advantages*	*Limitations*	*Sources*
XGB	- Gradient boosting to reduce residuals - L1 & L2 regularization - Efficient handling of sparse data.	High performance with large datasets - Flexible with multiple objective functions - Provides feature importance insights.	- Complex hyperparameter tuning - Less interpretable than simpler models	(Chen & Guestrin, 2016) (Egger, 2022)
RF	- Bagging with random data and feature subsets - Aggregates predictions across trees	- Robust against overfitting - Assesses feature importance - Easy to use with minimal tuning	- Resource-intensive - Less interpretable due to ensemble nature	(Tyralis & Papacharalampous, 2017) (Egger, 2022) (Genuer & Poggi, 2020)
SVR	- Kernel trick for higher-dimensional transformations - Epsilon-insensitive loss function - Regularization (C parameter)	- Effective for complex non-linear relationships - Flexible with different kernels	- Computationally expensive - Sensitive to hyperparameters	(Schölkopf & Smola, 2003) (Claveria et al., 2016) (Egger, 2022)

The distinct strengths of each model are summarized in Table 16, which provides an overview of the methodological foundations, advantages, limitations, and common applications of each model. This comparison highlights how these models address the complexities of sales forecasting, such as handling non-linear data, managing seasonality, and incorporating external factors.

4.5.2 Model Training

Training Data Preparation

The training data preparation involved several key steps to ensure the accuracy and reliability of the forecasting models. Proper data preparation is essential for minimizing biases and enhancing model performance (Han et al., 2012). The following steps outline the actions taken to preprocess, split, and organize the data for model training. The training data preparation involved several key steps:

1. **Data Preprocessing**: The data underwent preprocessing as described in previous sections, resulting in the creation of three distinct datasets for daily, weekly, and monthly forecasting. Additionally, during the model development process, it was observed that the models performed significantly better when predictions were made on a tour-by-tour basis, rather than combining all tours. Consequently, the data was segmented by the major tours: "Kayak Tour", "Grotto Tour 1", "Coast Cruise 2", "Grotto Tour 2", "Coast Cruise 1", "Sailing Cruise 1", "Dolphin Catamaran", and "Sailing Cruise 2".

2. **Data Splitting**: The data was divided using the convenience sampling approach. Given that this is a time series forecasting problem, it's appropriate that the data split is made chronologically (in blocks) rather than randomly (Reitermanová, 2010).The split was determined to occur on 1 January 2023, separating the training and test datasets, meaning the model will be developed to predict the number of seats sold between the data split until 30 June 2024.

Training Procedure and Hyperparameters

The models (XGB, RF, and SVR) were trained on the daily, weekly, and monthly datasets. For each model, multiple hyperparameters were tested using both grid search and random search techniques to optimize performance. Table 16 summarizes the key hyperparameters tested during the training process.

Table 16

Models Overview

Model	*Hyperparameter*	*Description/Function*
XGB (Chen & Guestrin, 2016) (Egger, 2022)	n_estimators	Number of trees in the model.
	max_depth	Controls tree depth, limiting model complexity.
	learning_rate	Reduces tree contribution. Lower values need more trees.
	subsample	Fraction of samples used per tree. Helps prevent overfitting.
	colsample_bytree	Fraction of features used per tree. Reduces overfitting.
	alpha	L1 regularization term. Prevents overfitting by adding sparsity.
	lambda	L2 regularization term. Penalizes large weights.
	min_child_weight	Minimum instance weight to create a node. Controls complexity.
	gamma	Minimum loss reduction. Makes model conservative.
	objective	Defines the learning task and loss function. 'reg' for regression.
RF (Tyralis & Papacharalampous, 2017) (Egger, 2022) (Genuer & Poggi, 2020)	n_estimators	Number of trees. Higher increase runtime.
	max_depth	Controls tree depth, limiting complexity.
	min_samples_split	Minimum samples needed to split a node. Controls overfitting.
	min_samples_leaf	Minimum samples required at a leaf node. Reduces overfitting.
	max_features	Number of features considered per split. Reduces overfitting.
	bootstrap	Uses bootstrapped samples to improve generalization.
SVR (Schölkopf & Smola, 2003) (Claveria et al., 2016) (Egger, 2022)	C	Regularization - balances training error and complexity.
	epsilon	Specifies the margin of tolerance for error.
	kernel	Maps data into higher dimensions or for linear relationships.
	gamma	Kernel coefficient. 'auto' sets 1/n_features.

4.6. Model Evaluation

The evaluation of ML models is a critical step in determining their performance and generalization capabilities. A combination of cross-validation techniques and evaluation metrics were employed to assess the predictive power of the models across daily, weekly, and monthly forecasts.

4.6.1 Cross-Validation Techniques

To thoroughly evaluate the generalization performance of the models and detect potential issues like overfitting or underfitting, k-fold cross-validation was employed. Specifically, 10-fold cross-validation was used to ensure that each model was trained and validated on different subsets of the data, providing a comprehensive assessment of model performance. This method is particularly useful in mitigating overfitting by ensuring that the model's performance is evaluated on multiple partitions of the data (Hastie et al., 2009). In ML, overfitting happens when a model learns the training data too well, including random noise, making it unable to perform well on new, unseen data. This results in poor predictions for future observations (Hastie et al., 2009).

4.6.2 Evaluation Metrics

To evaluate the performance of the predictive models, a suite of metrics was calculated across the various forecasting horizons (daily, weekly, and monthly). These metrics provide a comprehensive assessment of both the accuracy and the robustness of the models, being:

- **Mean Absolute Error (MAE)**: MAE measures the average magnitude of the errors in a set of predictions, without considering their direction. It is calculated as the average of the absolute differences between predicted and actual values:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (2)$$

  MAE provides a straightforward interpretation of prediction error in the same units as the original data.

- **Root Mean Squared Error (RMSE)**: RMSE is the square root of MSE and provides a metric in the same units as the target variable, making it easier to interpret:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (3)$$

  It is particularly useful when large errors are undesirable, as it penalizes large deviations more heavily.

- **R-squared (R^2)**: R^2 measures the proportion of the variance in the dependent variable that is predictable from the independent variables:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (4)$$

  It is a widely used metric that indicates how well the model explains the variation in the data, with values closer to 1 indicating better model performance.

- **Weighted Mean Absolute Percentage Error (WMAPE)**: WMAPE is a weighted version of MAPE, calculated as the total absolute error divided by the sum of the actual values:

$$\text{WMAPE} = \frac{\sum_{i=1}^{n}|y_i - \hat{y}_i|}{\sum_{i=1}^{n}|y_i|} \quad (5)$$

  It was decided to use WMAPE instead of MAPE (Mean Absolute Error) due to the presence of predictions close to zero and sometimes even 0 when forecasting the number of seats sold. In such cases, MAPE tends to produce disproportionately large errors or become undefined, making WMAPE a more suitable metric.

4.7. Model Implementation

The implementation of the forecasting model in the company environment is critical to ensuring that the model provides timely and accurate predictions to support decision-making. This section outlines the deployment strategy, data flow, and integration into the company's operations, ensuring that the model's results are effectively used by stakeholders.

4.7.1 Deployment Strategy

The deployment of the forecasting model is designed to operate in a Python environment, where the model is run periodically. Since the data is not updated in real-time, the forecasting model will be triggered manually, typically before key strategic meetings or resource allocation decisions. This allows for flexibility in the forecasting process and ensures that the most up-to-date data is utilized in making predictions.

Python Environment

The model will run in a Python environment hosted on a local machine or server. The Python environment contains all necessary dependencies, libraries and modules for both data preprocessing and model execution.

Data Access and Updating

Since this forecasting model is dependent on both internal data (from the company's booking system) and external supporting data, a manual ETL process will be used to update the datasets before running the model. This process involves exporting new data from the company's booking system and loading it into the Python environment.

The ETL pipeline is fully integrated into the Python code, ensuring seamless preprocessing and data transformation. The code has been designed to accommodate new data imports from the booking system and external sources. Once the data is updated and preprocessed, it is passed into the model for forecasting. This process ensures that managers and the sales team work with the most recent data available when reviewing the forecasts.

Running the Model

The model is not set up to run automatically, as the forecasts are typically generated before meetings or strategic discussions. Managers or data analysts manually trigger the model execution through the Python environment.

Output and Power BI Integration

Once the model has been executed and the best model is selected for each granularity, the forecasts are saved in CSV files. These CSV files serve as the input for the company's Power BI dashboard, which is used by managers and the sales team to visualize the results. The dashboard is directly connected to the CSV files, allowing for easy updates whenever the forecasting model is rerun.

Model Adjustability and Monitoring

The forecasting system consists of nine models in total (three models for each granularity: daily, weekly, and monthly). After each run, the model with the best performance metrics is selected and its results are exported. However, the system is designed to remain highly flexible and adjustable over time, allowing the company to fine-tune the model as business needs evolve.

Implementation in the Company Environment

The overall implementation strategy is designed to be simple and manageable within the company's existing operations. While the model requires manual execution, it fits into the company's workflow by providing forecasts before key meetings and decision points. In practice, the company's internal team or an external contractor can handle the data extraction, preprocessing, and model execution.

Monitoring and Continuous Improvement

To ensure ongoing performance and relevance, the model's accuracy is monitored regularly by comparing predicted values against actual sales data. If the model's performance starts to degrade, it can be retrained with new data or adjusted by modifying the hyperparameters or adding new features. This ensures that the forecasting model stays in line with evolving market conditions and company needs. By maintaining an ongoing feedback loop, the model will continue to provide accurate and actionable forecasts that support strategic decision-making.

4.7.2 Power BI Integration

The integration of the forecasting model with Power BI provides an intuitive and interactive platform for the company's management and sales teams to make data-driven decisions. The dashboard is designed to present the forecasting model's outputs in a clear and actionable way, divided into four main sheets: *Overview*, *Operational*, *Financial*, and *Forecasting*. Each sheet serves a distinct purpose, allowing stakeholders to focus on specific aspects of the business:

- "**Overview"**: Summarizes overall sales and performance.
- **"Operational"**: Supports daily resource management and adjustments.
- **"Financial"**: Provides insights into sales performance and resource performance.
- **"Forecasting"**: Compares model performance for sales projections.

This integration of Power BI ensures that forecasts are not only accessible but also actionable, supporting strategic decision-making across operational, financial, and sales departments.

CHAPTER V - RESULTS & DISCUSSION

In this chapter, the results of the forecasting models are presented and analyzed, focusing on their performance and implications for the studied MTO. The models were trained and tested using historical data, with predictions generated for various time granularities, including daily, weekly, and monthly forecasts.

The integration of the model results into the company's business intelligence system, Power BI, plays a key role in ensuring that these forecasts are not only generated but also easily accessible and interpretable by decision-makers.

This chapter will first outline the overall performance of the models, followed by a discussion of the metrics, comparisons across models and time granularities, and the practical implications for the business.

5.1. Overview of Model Performance

The evaluation of the forecasting models centers on their ability to predict sales across different time granularities—daily, weekly, and monthly. The three models tested, XGB, RF, and SVR, each offer unique approaches to predictive modeling. The performance of these models is assessed using key metrics such as MAE, RMSE, R^2, and WMAPE. These metrics provide insight into how well each model can generalize and predict future sales patterns.

To ensure a more focused and clear analysis, this section will present the performance results of the forecasting models applied to the Kayak Tour. The decision to concentrate on this particular tour stems from the fact that it consistently generates the highest number of bookings, as shown in Section 4.3 (EDA), making it the most representative of the dataset. Additionally, the Kayak Tour is the second highest in terms of seats sold, and the stability and consistency of its data ensure robust model performance. Although similar trends are observed across other tours, the data for those tours can be slightly more volatile and less consistent, resulting in minor instabilities in the model predictions. Therefore, the results presented here can be seen as indicative of the general performance across all tours.

5.2. Model Performance Results

5.2.1 Daily Forecasting Results

The daily forecasts were generated using XGB, RF, and SVR models. Each model was evaluated based on the metrics chosen. Below is a summary of the performance for each model (Table 17).

Table 17

Daily Forecasting Results

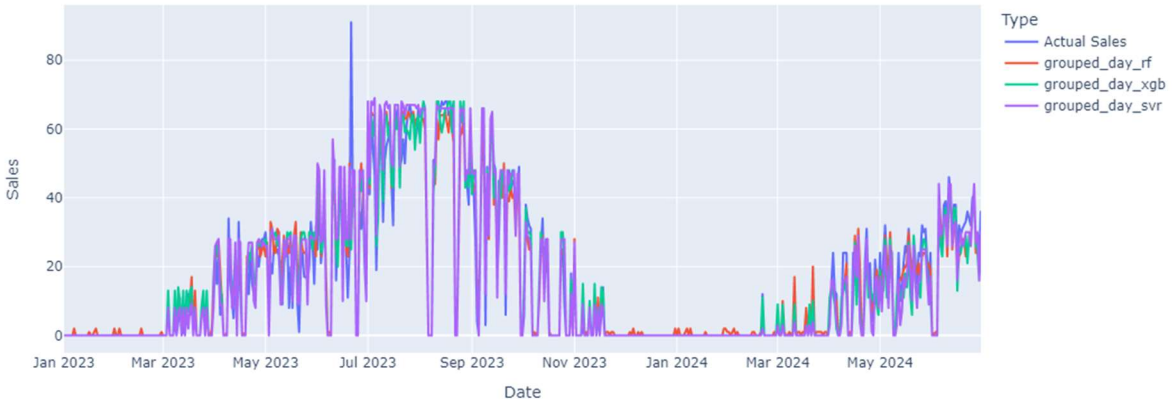
Daily Model	*MAE*	*RMSE*	*WMAPE*	R^2
XGB	2.93	5.69	0.185	0.922
RF	3.16	5.76	0.199	0.920
SVR	2.91	6.09	0.183	0.910

The SVR model outperforms the other models in terms of MAE and WMAPE, providing the most accurate daily forecasts. However, XGBoost offers a competitive alternative with a slightly higher R^2 score and lower RMSE, indicating it explains a comparable amount of variance and is similarly robust at predicting day-to-day sales.

Additionally, two visualizations were created: a time series plot comparing actual and predicted sales, and a parity plot showing predicted values against actual sales.

Figure 18

Daily Time Series Plot: Actual vs. Predicted Sales



The daily time series plot (Figure 18) shows actual sales alongside the predicted values from each model. The SVR model closely follows the actual sales, particularly during peaks and troughs, indicating higher precision. The XGB model performs similarly but shows

minor deviations during periods of high sales. The Random Forest model is more volatile, especially during periods of low and high sales, where its predictions deviate more from the actual values.

Figure 19

Daily Parity Plot: Actual vs. Predicted Sales

The parity plot (Figure 19) evaluates how well predictions align with actual sales. Ideally, points should fall along the 45-degree line, indicating perfect accuracy. The SVR model shows the closest alignment, although slight over-predictions are seen for higher sales. The XGB model is also well-aligned but displays more variance, particularly in mid-range sales.

5.2.2 Weekly Forecasting Results

The performance of the models for weekly forecasting was assessed using the same metrics as the daily analysis. Table 18 provides a comparison of the results for each model.

Table 18

Weekly Forecasting Results

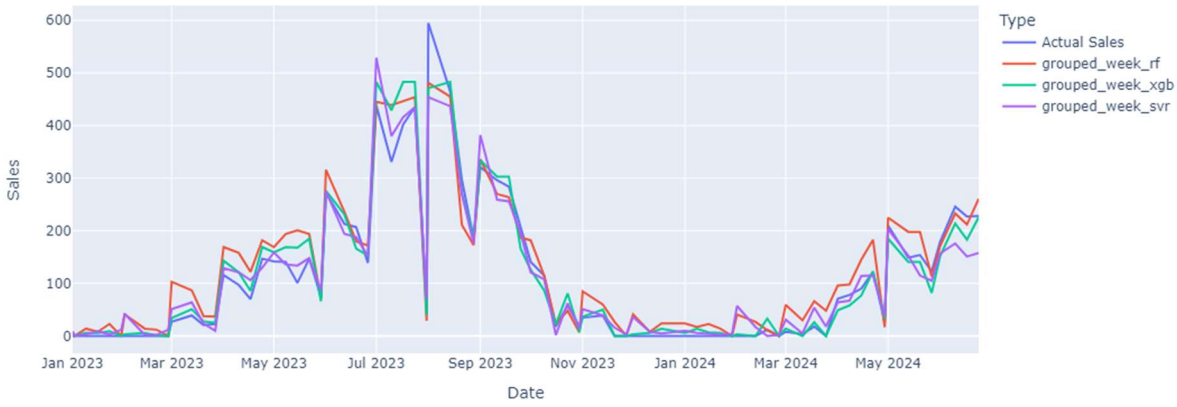
Weekly Model	*MAE*	*RMSE*	*WMAPE*	R^2
XGB	17.77	28.22	0.160	0.955
RF	29.09	38.18	0.262	0.918
SVR	20.12	30.89	0.181	0.946

In terms of weekly forecasting performance, the XGB model delivers the most accurate results, achieving the lowest MAE, RMSE, and WMAPE values, along with the highest R^2. This demonstrates XGB's superior capability in capturing weekly trends and patterns. The SVR model follows, showing competitive performance with relatively low error values

and a high R^2. However, despite having a good R^2 result, it underperforms, with noticeably higher errors and lower predictive accuracy compared to the other two models.

Figure 20

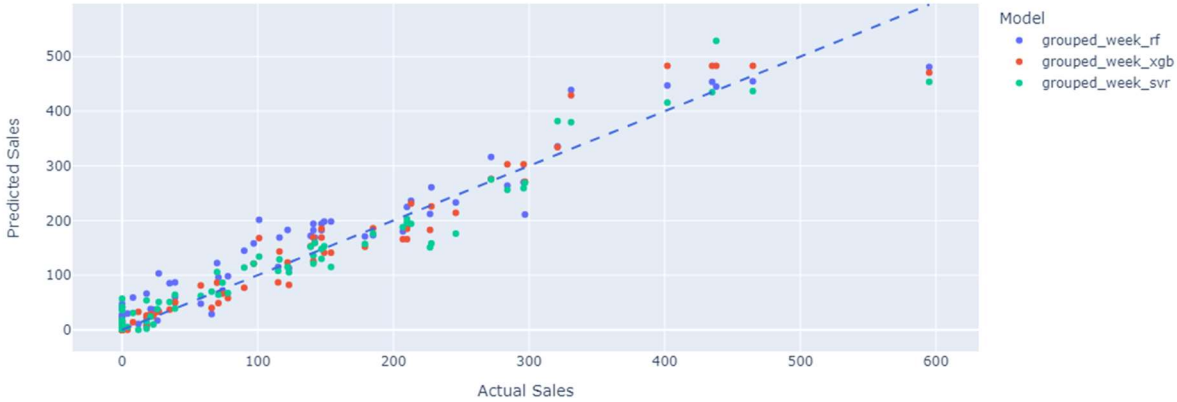
Weekly Time Series Plot: Actual vs. Predicted Sales



The time series plot (Figure 20) illustrates how each model tracks the actual weekly sales values over the forecasting period. In this case, XGB demonstrates a strong ability to follow the weekly sales patterns, particularly in periods of high volatility and fluctuating sales, where it closely mirrors the peaks and troughs in the data. SVR also performs well but exhibits minor deviations during some of the sales peaks, where it underpredicts sales to a greater extent than XGB.

Figure 21

Weekly Parity Plot: Actual vs. Predicted Sales



The parity plot (Figure 21) illustrates the alignment between actual and predicted weekly sales. XGB performs well, with predictions following the diagonal, suggesting accurate predictions across a range of sales values. SVR also aligns well while RF exhibits a more stable spread of point at mid to high sales when compared to the other models.

5.2.3 Monthly Forecasting Results

The monthly forecasting performance of each model was evaluated using the same set of error metrics as used in the previous sections, resulting in Table 19.

Table 19

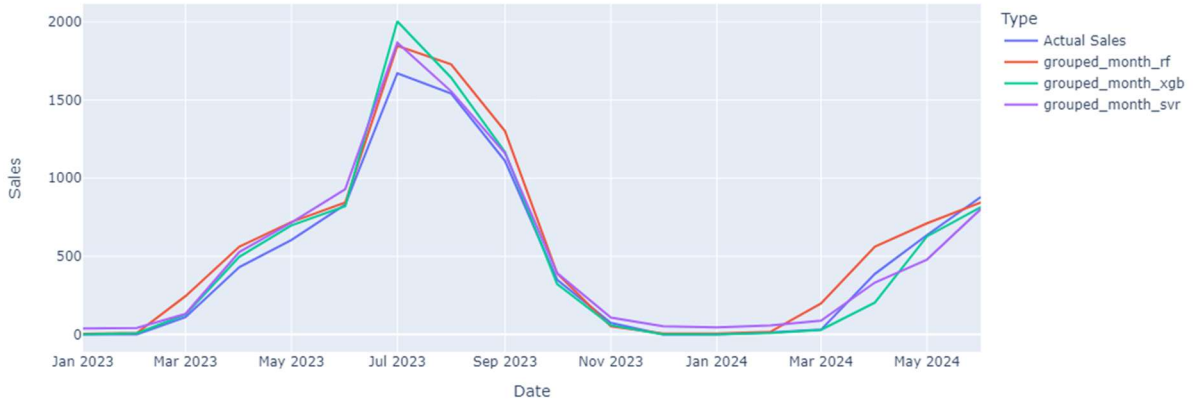
Monthly Forecasting Results

Monthly Model	*MAE*	*RMSE*	*WMAPE*	R^2
XGB	54.89	98.89	0.114	0.964
RF	82.83	110.19	0.172	0.956
SVR	68.39	82.56	0.142	0.975

For monthly forecasting, XGB demonstrates the lowest MAE and WMAPE, indicating higher precision and reliability in predicting long-term sales. It also provides a competitive RMSE score, which reflects a lower magnitude of large errors, making it the most consistent model overall. SVR, on the other hand, achieves the highest R^2, which suggests that it captures a greater proportion of variance in the data, despite slightly higher error metrics.

Figure 22

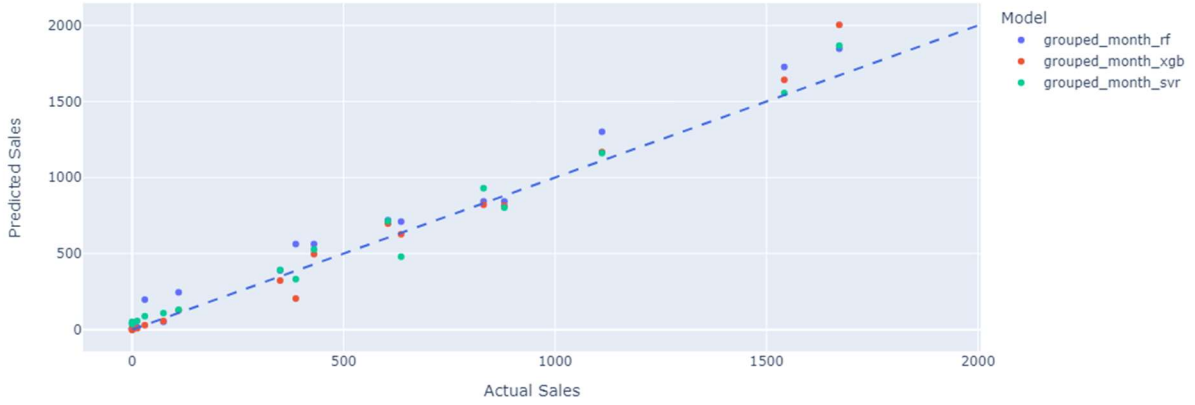
Monthly Time Series Plot: Actual vs. Predicted Sales



The time series plot (Figure 22) compares actual monthly sales against the predicted values from each model over the forecasting period. XGB closely tracks the actual sales curve, especially during periods of both rapid increases and decreases, demonstrating its ability to capture long-term trends. However, it significantly overestimates sales in July 2023, where it predicts higher values than observed, marking the most pronounced deviation among the models during this period, thus justifying the high RMSE value when compared to the SVR model. Similarly, SVR also aligns well with the actual data, while RF, tends to overshoot in certain months and struggles with the sharp peaks and troughs, leading to more pronounced errors compared to the other two models.

The parity plot (Figure 23) provides insight into how well the predicted monthly sales match the actual values. In an ideal model, points would align along the 45-degree dashed line, representing perfect prediction accuracy.

Figure 23 -
Monthly Parity Plot: Actual vs. Predicted Sales



In this plot, XGB shows the closest alignment with the diagonal, especially for lower sales values, indicating a strong fit to the actual data. SVR also performs well, though its predictions tend to cluster slightly below the diagonal line in mid sales ranges, suggesting minor underestimation. RF exhibits greater spread across all sales levels, particularly at higher values, reflecting its challenges in accurately predicting monthly sales.

5.2.4 Comparative Analysis

In evaluating and comparing the performance of the models across various time granularities, several important trends and patterns begin to emerge, offering insights into how each model behaves depending on the forecasting horizon.

- **XGB:** Excels in weekly and monthly forecasts, delivering high accuracy, particularly when capturing long-term patterns. Its predictive power makes it ideal for scenarios where both accuracy and long-term trends are essential, though caution should be exercised in cases where extreme fluctuations are expected.
- **SVR:** Demonstrates superior performance in daily forecasts, achieving the lowest error metrics and proving to be the most accurate model for short-term predictions. Although it remains competitive in weekly and monthly forecasting, its accuracy slightly diminishes as the time horizon extends, despite its ability to capture underlying data patterns effectively.
- **RF:** While not as precise as XGB or SVR, offers an interpretable and simple approach to forecasting. Its higher error metrics across all time granularities indicate that it may not be the best choice when predictive accuracy is the primary goal. However, its ease of use, simplicity and model transparency make it a viable option in contexts where these factors are prioritized.

5.3. Discussion of Model Performance

5.3.1 Overfitting and Underfitting

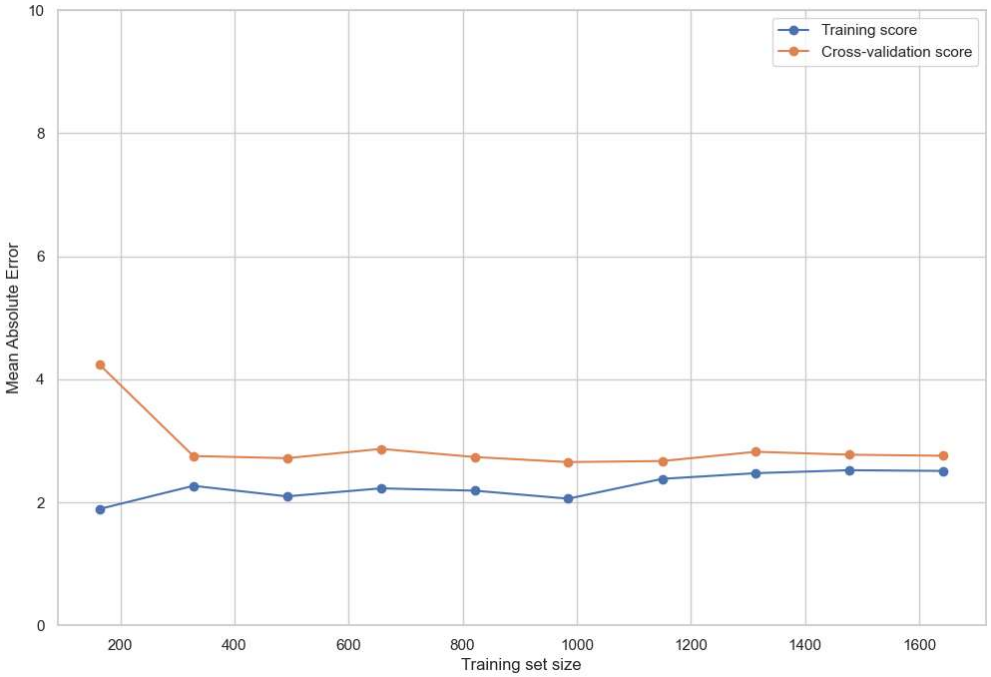
The following section examines the potential overfitting or underfitting of the models based on the learning curves generated for the best-performing models across different granularities—daily, weekly, and monthly forecasts. Each of these learning curves plots the training error and cross-validation error as the size of the training set increases, offering insights into how the models generalize to unseen data.

Daily Forecasting: SVR Learning Curve

The learning curve for SVR on daily sales forecasting (see Figure 24) indicates that the model avoids significant overfitting. As the training set size increases, the training error remains relatively stable and consistently low, suggesting that the model is not over-adjusting to the training data. Simultaneously, the cross-validation error gradually decreases, indicating the model's ability to generalize well to unseen data.

Figure 24

Daily Forecasting: SVR Learning Curve



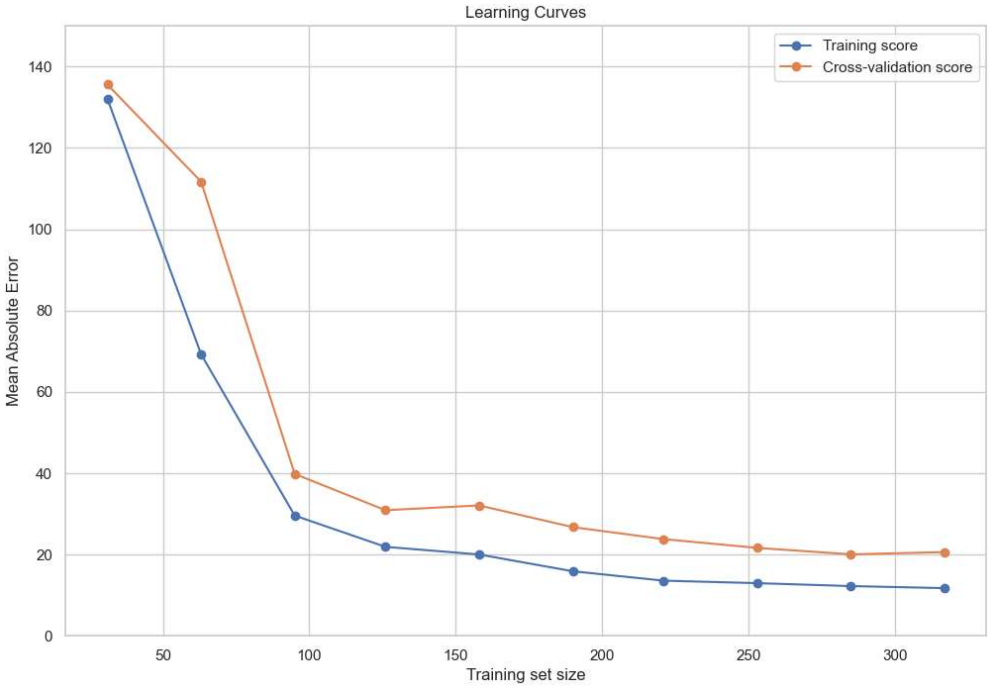
The steady performance across larger datasets, with little fluctuation in errors, suggests that the SVR model is well-suited to the problem and unlikely to suffer from overfitting. Importantly, the relatively flat slope of the learning curves at larger training sizes indicates that the model has reached a stable state, where additional data may bring marginal improvements.

Weekly Forecasting: XGBoost Learning Curve

For the weekly forecasting model using XGB (see Figure 25), the learning curve shows a clear trend where both the training and cross-validation errors decrease as the training set size increases. Initially, the cross-validation error and training error are significantly high at smaller data sizes. However, as the training set increases, the two errors converge, indicating improved model generalization.

Figure 25

Weekly Forecasting: XGB Learning Curve



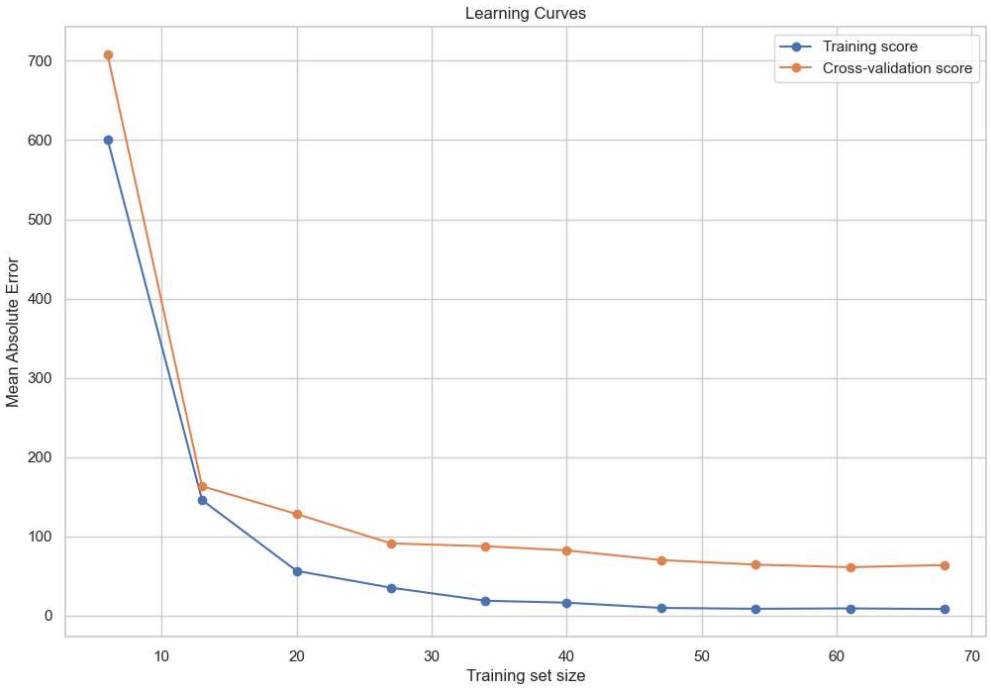
The convergence of these two errors in XGB indicates a well-balanced model that effectively learns from the data. However, as the training set size increases beyond 150 samples, the cross-validation error plateaus, with minimal improvement despite the addition of more data. This stabilization indicates that the model has likely reached the limits of its current configuration, and adding more data does not substantially reduce the validation error. While the model demonstrates a good ability to learn from the training data, as reflected by the steady decline in training error, the gap between the two curves suggests room for further optimization.

Monthly Forecasting: XGBoost Learning Curve

For the monthly forecasting task, the XGB model (see Figure 26) displays a similar pattern to the weekly forecast but with a more pronounced drop in both the training and cross-validation errors. Initially, both the training and validation errors are relatively high, especially with a smaller dataset. As the dataset size increases, both errors drop significantly and stabilize.

Figure 26

Monthly Forecasting: XGB Learning Curve



This behavior suggests that while XGB initially struggles to capture the monthly trends with smaller datasets, it rapidly improves as more data is introduced, eventually converging to a stable point. The convergence of the training and cross-validation errors demonstrates that the model generalizes well without overfitting. However, the gap between the two curves suggests potential for more improvement.

5.3.2 Features Importance

In this study, feature importance for the XGB models was evaluated using two key metrics: gain and weight, both of which are provided by the model's package. These metrics provide insight into the contribution of each feature to the model's predictions. *Gain* represents the improvement in prediction accuracy when a feature is used in a decision split, reflecting the feature's overall contribution to reducing prediction error. On the other hand, *Weight* refers to the frequency with which a feature is used in the decision-making process (Abu-Rmileh, 2019).

To assess feature importance in the SVR model, permutation importance was applied. Features that cause a larger drop in performance when shuffled are deemed more important (Breiman, 2001). This method is particularly well-suited for non-linear models like SVR.

Table 20

Feature Importance Results

Feature	***Daily SVR***	***Weekly XGB***		***Monthly XGB***	
	Permutation (%)	*Gain (%)*	*Weight (%)*	*Gain (%)*	*Weight (%)*
Total Seats	71.44%	70.24%	24.07%	64.74%	24.25%
Avg Value	4.65%	12.05%	21.11%	4.09%	20.17%
Webvisits	6.55	8.45%	48.52%	N/A	N/A
Workday	10.26%	9.26%	6.30%	5.38%	12.66%
Month_Sin	7.10%	N/A	N/A	N/A	N/A
Season_Summer	N/A	N/A	N/A	1.36%	2.36%
Passengers	N/A	N/A	N/A	9.47%	16.31%
Overnight Stays	N/A	N/A	N/A	14.96%	24.25%

The feature importance analysis in Table 20 reveals key insights into the predictive performance of the models used for daily, weekly, and monthly sales forecasting. Across all three models, *Total Seats* available emerges as the most important feature, consistently showing high gain and weight percentages in the XGB models and in permutation importance in the SVR model. This is expected, as total seat capacity is a direct constraint on the maximum number of bookings, making it the primary driver of sales. The dominance of this feature across all time granularities highlights its critical role in improving prediction accuracy.

A significant improvement came from separating the tours and running the models for each tour individually, rather than forecasting all tours aggregated. This separation allowed the models to capture the unique dynamics of each tour, greatly enhancing predictive accuracy.

Workday and *Webvisits* also contribute, particularly in daily and weekly forecasting, though their impact is subtle. In monthly forecasting, external features such as *Overnight Stays* and *Passengers* (from the supporting datasets) showed some importance in the XGB model. During the model development phase, non-relevant features were dropped to streamline the models.

5.3.3 Hyperparameter Tuning Results

The hyperparameter tuning process led to the identification of optimal settings for each model. Table 21 provides a summary of the best hyperparameters identified for each model.

Table 21

Hyperparameters Selected

Model	*Granularity*	*Best Hyperparameters*
SVR	Daily	kernel = 'rbf', C = 7, epsilon = 0.01, gamma = 'auto'
XGB	Weekly	n_estimators = 150, max_depth = 8, learning_rate = 0.2, subsample = 0.8, colsample_bytree = 0.5, alpha = 0.01, lambda = 3, min_child_weight =6, gamma = 0.01, objective = 'reg:absoluteerror'
XGB	Monthly	n_estimators = 125, max_depth = 4, learning_rate = 0.4, subsample = 0.8, colsample_bytree = 0.5, alpha = 0.01, lambda = 10, min_child_weight = 4, gamma = 0.05, objective = 'reg:absoluteerror'

5.4. Practical Implications for the Business

The practical implications of the forecasting models and the Power BI dashboard extend beyond merely predicting future sales, since the primary goal of this study is to support the SME under study in decision-making and optimize the operational and financial efficiency. The actionable insights provided by the forecasts can have a tangible impact on operational planning and sales strategies. By using data storytelling when visualizing and interacting with the analyzed data in Power BI, the studied MTO will be able to make informed decisions that optimize resource allocation, improve revenue, and adjust pricing dynamically.

5.4.1 Insights from Power BI Dashboard

The integration of data into the Power BI dashboard not only enhances the company's ability to interpret forecast results but also provides valuable insights from the booking software data. Organized into several key sections—Overview, Operational, Financial, and Forecasting—the dashboard offers the MTO an intuitive, interactive platform for data exploration. This setup enables a more comprehensive understanding of sales trends, operational needs, and future projections, ultimately driving more informed decision-making across the organization.

Overview Sheet

The Overview Sheet (Figure 27) provides a comprehensive summary of KPIs that deliver insights into the MTO's overall metrics (in the current year). This dashboard integrates several interactive visualizations that facilitate monitoring of business performance. Key metrics such as Total Orders, Seats Confirmed, and Total Revenue are prominently displayed, offering an immediate snapshot of the company's overall performance.

Additionally, the dashboard includes visual representations of sales and seat occupancy data, presenting the confirmed seats for each tour alongside occupancy rates, which helps track both demand and operational efficiency. A time series chart further illustrates historical sales data while projecting future trends

Figure 27

Dashboard Overview Sheet

To enhance its functionality, the dashboard incorporates interactive filters that allow users to explore data at various levels of granularity, including by year, month, day, tour, category, and location. This interactivity ensures that the dashboard can be adapted to specific business queries, making it a flexible and valuable tool for both operational and strategic decision-making.

Operational Sheet

The Operational Sheet (Figure 28) offers a focused view of daily performance metrics that are crucial for managing resource allocation, staffing, and short-term operations. Key indicators such as Seat Occupancy Rate, Total Seats Confirmed, and Total Orders are presented at the top, providing an immediate snapshot of the current operational status.

Figure 28

Dashboard Operational Sheet

This dashboard includes a line chart that tracks fluctuations in seat occupancy rates and a bar chart showing daily confirmed bookings. A table offers detailed insights into seat utilization across various tours by displaying total available and confirmed seats. The line chart on webvisits helps correlate online traffic with booking trends. Interactive filters enable users to refine data by year, month, day, tour, category, and resource, enhancing the dashboard's adaptability. Additionally, the weather forecast offers context for interpreting performance, especially for weather-dependent tours. This dashboard facilitates informed decision-making by allowing the company to dynamically respond to changes in demand and operational factors.

Financial Sheet

The Financial Sheet (Figure 29) provides a detailed overview of the company's financial performance. Key financial indicators such as Total Revenue and Average Booking Value are prominently displayed, offering immediate insight into overall

business profitability. Below these metrics, a time series chart tracks Average Revenue per Tour and Total Revenue by Year & Month, allowing users to analyze long-term trends in revenue generation and fluctuations across seasons.

Figure 29

Dashboard Financial Sheet

A donut chart visually represents the Distribution of Revenue by Tour, offering a clear breakdown of which tours contribute the most to total revenue. This is complemented by two detailed tables: one displaying Average Booking Value and Average Revenue per Session by Location, and the other by Category, enabling a more granular understanding of revenue sources across different regions and product types.

Interactive filters at the top of the dashboard allow users to explore financial data by year, month, day, tour, category, and location, enhancing its adaptability and usability for different financial analyses.

Forecasting Sheet

The Forecasting Sheet (Figure 30) provides a detailed analysis of sales projections, allowing the company to simulate potential revenue outcomes based on different pricing strategies.

Figure 30

Dashboard Forecasting Sheet

In the Current Pricing graph, sales forecasts for the upcoming months are displayed, based on the existing tour prices. The Price Simulation section introduces an interactive element, where users can adjust the tour price using a What-If Parameter. This feature dynamically updates the value of forecasted sales based on the new tour price, providing a quick and intuitive method for evaluating how pricing changes might impact total sales and revenue, based on the predicted seats by the models.

At the bottom of the dashboard, a performance metrics table presents key indicators for the daily, weekly, and monthly models. Additionally, *Passengers* & *Overnight Stays* graph gives a broader view of macro trends, allowing for cross-referencing between

5.4.2 Using Forecasts for Operational Planning

The operational planning process within the company can be significantly optimized by integrating forecast and booking data into the Power BI dashboard. This integration enables managers to make informed decisions regarding staffing, resource allocation, and maintenance schedules, all based on anticipated sales trends.

Staffing allocation is a key example of how forecasts enhance operational planning. By predicting high-sales periods, such as for the Kayak Tour, managers can schedule additional staff during peaks. Sales forecasts enable managers to allocate boats,

equipment, and other assets effectively, preventing overbooking and underutilization. Additionally, maintenance schedules can be planned during slower periods, minimizing disruptions and ensuring resources are ready for peak seasons.

The monthly forecasting models also aid in seasonal planning. By analyzing forecasted sales, managers can better prepare for both peak and off-peak periods, informing promotional strategies, resource planning, and other initiatives to maximize opportunities and mitigate challenges.

5.4.3 Price Simulation

The Forecasting Sheet includes a price simulation feature that allows managers to explore the potential impact of different pricing strategies on sales figures. While this tool does not alter the model's forecast, it offers valuable financial projections by simulating how price changes could affect revenue. This simulation is particularly useful, as the company typically maintains fixed prices throughout the year, meaning the impact on actual sales figures (rather than forecast) can be studied.

Another important application is in discount campaigns. The dashboard allows managers to simulate different discount scenarios and evaluate their potential impact on revenue. By visualizing these different scenarios, the company can find the right balance between volume and pricing to meet overall revenue targets.

Overall, the price simulation capability empowers the sales and financial teams to experiment with and evaluate various pricing strategies and resource management decisions, significantly reducing the risk of negative outcomes from untested approaches.

CHAPTER VI - CONCLUSION

6.1. Summary of Findings

The primary objective of this thesis was to develop and evaluate predictive sales forecasting models that could provide accurate, actionable insights to support decision-making in a MTO in the Algarve region. The models focused on predicting sales across daily, weekly, and monthly granularities for the studied SME.

To achieve these objectives, several ML algorithms were employed, including XGB, RF, and SVR. These models were evaluated using various performance metrics such as MAE, RMSE, WMAPE, and R^2, with the best-performing models integrated into the company's Power BI dashboard for analysis and visualization.

The Power BI dashboard helps in the effective interpretation of model outputs, since stakeholders and managers are able to interact with the data, explore different time horizons, and make informed decisions related to staffing, resource allocation, and pricing strategies. The dashboard's ability to simulate pricing changes and project their impact on revenue further extended the practical utility of the models in daily operations.

The key results of this thesis demonstrated that:

- XGB performed best overall, especially in weekly and monthly forecasting, due to its ability to handle non-linear relationships and seasonal patterns.
- SVR was particularly effective in daily forecasts, managing short-term fluctuations well.

The Power BI dashboard was essential for transforming raw model outputs into actionable insights, providing a visual framework for decision-making across operational and financial areas.

6.2. Limitations and Challenges

Despite the promising results, several limitations and challenges were encountered throughout the study. These limitations may have impacted the accuracy and generalizability of the models.

6.2.1 Data Limitations

One of the primary limitations of the project was the availability and quality of data. While some tours provided a consistent dataset, others exhibited more volatile patterns and fewer bookings, which led to difficulties in building accurate models for those tours. Additionally, the availability of macroeconomic data posed a challenge: such data is typically released monthly and has monthly granularity , which limits the model's ability to detect short-term changes in economic conditions that could affect booking trends. This delay in macroeconomic updates can create a lag in incorporating external economic factors into the forecasting models, potentially reducing their real-time accuracy.

Moreover, external factors such as weather conditions and sea conditions were not considered in the analysis. Incorporating data on daily weather forecasts, including temperature, precipitation, and wind conditions, as well as sea conditions (e.g., wave height, currents), could enhance the predictive power of the models. These variables can significantly impact customer behavior, especially in maritime tourism, where bookings may fluctuate depending on favorable or unfavorable conditions. Including these environmental factors could provide more accurate short-term forecasts, especially for activities like kayaking and other sea-related tours.

6.2.2 Model Limitations.

All the models employed in this thesis have inherent limitations that could affect their overall performance. One of the key challenges is the computational complexity of the models, which can become significant when dealing with larger datasets. As the volume of data grows, the models may require considerable computational resources, leading to longer processing times and potentially limiting their practicality in real-time forecasting environments. Additionally, the models often demand careful hyperparameter tuning to achieve optimal performance. Without this, there is a risk of overfitting, particularly with smaller datasets, where the models may capture noise rather than meaningful patterns, reducing their ability to generalize to unseen data.

Despite these challenges, the models produced acceptable results. However, the performance metrics suggest that some models could benefit from better parameter tuning. This would likely improve their accuracy and ability to capture more complex relationships within the data, particularly in more volatile periods or when external factors change rapidly.

Lastly, real-time forecasting presents another challenge. The models, as configured in this study, require manual data updates and re-training, which limits their ability to provide real-time insights. In fast-moving business environments, where timely and accurate forecasts are crucial, this delay can hinder decision-making and reduce the models' effectiveness in adjusting to rapid market changes or shifts in external conditions, such as weather or economic updates.

6.2.3 Power BI Dashboard Limitations

While the Power BI dashboard has proven to be an effective tool for visualizing and interpreting forecast results, it does have certain limitations. One limitation is related to real-time updates: the dashboard requires manual updating with new data exports from the booking system, which restricts its ability to provide continuous, real-time insights. Additionally, scalability is another challenge. The current configuration of the dashboard may require adjustments if the forecasting models are to be applied to a broader range of tours or services. As more data sources are added or the volume of data increases, the existing system could become overwhelmed, necessitating further optimization to maintain its performance.

6.3. Future Work

The findings from this thesis point to several opportunities for further research and model refinement. Future efforts could focus on improving model accuracy, expanding the application of forecasting, and enhancing the Power BI dashboard to support more complex analysis.

6.3.1 Enhancements to Models

To improve forecasting accuracy, several enhancements can be considered. One option is to explore alternative machine learning algorithms that may better capture complex relationships in the data, offering improved accuracy for different forecasting horizons.

Another improvement would be the incorporation of external data, such as weather conditions, economic indicators, or dynamic pricing, which could help enhance the model's predictive power. These external factors often influence customer behavior and demand, especially in tourism, and their integration could lead to more accurate forecasts.

Finally, it is essential to study additional model parameters to further optimize performance. By carefully adjusting these parameters, the models could provide better results and a more precise fit to the data, ensuring improved generalization to new, unseen data.

6.3.2 Enhancing the Power BI Dashboard

The Power BI dashboard could be enhanced in several ways to improve functionality and insights. Automating the data integration process would enable real-time data updates from the booking system, reducing the need for manual intervention and ensuring continuous forecasting updates.

Additionally, incorporating advanced analytics, such as real-time predictive models and more in-depth what-if simulations, would provide deeper insights into potential future scenarios and their impact on sales.

Finally, adding interactive tools for adjusting key parameters like price or demand elasticity within the dashboard would allow decision-makers to simulate different business scenarios more easily, helping them make more informed operational adjustments.

6.3.3 Final Thoughts

The forecasting models developed in this thesis were primarily focused on a MTO, a SME operating within the NT sector in the Algarve region. Incorporating models to test price elasticity could help simulate the impact of price changes on demand, offering valuable insights for pricing strategies within the highly competitive tourism industry. Furthermore, implementing revenue forecasting would enable the MTO to predict future income based on fluctuating demand, aiding in financial planning and resource allocation.

This thesis has demonstrated the potential value of using machine learning models and data visualization tools like Power BI to support data-driven decisions, especially for an SME in the nautical tourism sector. While the current models provide foundational insights, adding functionalities such as price elasticity and revenue forecasting could further enhance the MTO's ability to make informed operational decisions. The Power BI dashboard has proven instrumental in transforming raw data into actionable insights, allowing decision-makers in this SME to better respond to changing market conditions.

As businesses in the tourism sector increasingly adopt data-driven decision-making, the integration of advanced forecasting models and visualization tools will play a critical role in optimizing operations and ensuring competitiveness, especially for SMEs. The work presented here lays a solid foundation for further developments and extensions, ultimately helping the MTO improve its efficiency, pricing strategies, and financial planning in a dynamic and evolving market landscape.

BIBLIOGRAPHY

Abela, A. (2024, June 1). *7. Charts*. Retrieved from Extreme Presentation: https://extremepresentation.com/design/7-charts/

Abor, J. (2017). Financial Planning and Forecasting. 199-224. https://doi.org/10.1007/978-3-319-34021-0_9

Abu-Rmileh, A. (2019, February 8). *The Multiple faces of 'Feature importance' in XGBoost*. Retrieved 07 14, 2024, from Medium: https://towardsdatascience.com/be-careful-when-interpreting-your-features-importance-in-xgboost-6e16132588e7

Almeida, F., & Silva, O. (2020). THE IMPACT OF COVID-19 ON TOURISM SUSTAINABILITY: EVIDENCE FROM PORTUGAL. *Advances in Hospitality and Tourism Research*, 440-446. https://doi.org/10.30519/ahtr.775340

Anbarasi, J., & Radha, V. (2023). CLASSIFICATION FOR BIG DATA DRIVEN MARINE WEATHER FORECASTING USING MACHINE LEARNING TECHNIQUES. *International Journal of Scientific Research in Engineering and Management, 7*. https://doi.org/10.55041/ijsrem17712

Anesti, N., Kalamara, E., & Kapetanios, G. (2024). Forecasting with Machine Learning methods and multiple large datasets. *Econometrics and Statistics*. https://doi.org/10.1016/j.ecosta.2024.08.003

Armstrong, J. (2001). *PRINCIPLES OF FORECASTING: A Handbook for Researchers and Practitioners.* Kluwer Academic Publishers. https://doi.org/10.1007/978-0-306-47630-3

Atanda, O., Adebiyi, M., Adewumi, D., Abiodun, M., Awodoye, O., Adebiyi, A., & Lawrence, M. (2024). Intelligent Sales Forecasting System Using Arima, Sarima, and Xgboost Models. *International Conference on Science, Engineering and Business for Driving Sustainable Development Goals (SEB4SDG)*, (pp. 1-8). Omu-Aran, Nigeria. https://doi.org/10.1109/SEB4SDG60871.2024.10629780

Awad, M., & Khanna, R. (2015). *Books for professionals by professionals. Efficient learning machines: Theories, concepts, and applications for engineers and system designers.* Apress Springer Science+Business Media. https://doi.org/10.1007/978-1-4302-5990-9

Banerjee, N., Morton, A., & Akartunalı, K. (2020). Passenger demand forecasting in scheduled transportation. *European Journal of Operational Research, 286*, 797-810. https://doi.org/10.1016/j.ejor.2019.10.032

Benítez, R., Paredes, R., Lodewijks, G., & Nabais, J. (2013). Damp trend Grey Model forecasting method for airline industry. *Expert Systems with Applications, 40*(12), 4915-4921. https://doi.org/10.1016/j.eswa.2013.02.014

Blair, A., Mandelker, G., Saaty, T. L., & Whitaker, R. (2010). Forecasting the resurgence of the U.S. economy in 2010: An expert. *Socio-Economic Planning Sciences, 44*, 114-121. https://doi.org/10.1016/j.seps.2010.03.002

Breiman, L. (2001). Random Forests. *Machine Learning, 45*, 5-32. https://doi.org/10.1023/A:1010933404324

Cadavid, J., Lamouri, S., & Grabot, B. (2018). Trends in Machine Learning Applied to Demand & Sales Forecasting: A Review. *7th International Conference on Information Systems, Logistics and Supply Chain ILS Conference 2018.* Retrieved from https://hal.science/hal-01881362

Cascetta, E., & Coppola, P. (2014). High Speed Rail (HSR) induced demand models. *Procedia-Social and Behavioral Sciences, 111*(5), 147-156. https://doi.org/10.1016/j.sbspro.2014.01.047

CCDR Algarve. (2023). *Relatório de progresso do GT das grutas de Benagil.* CCDR Algarve & CCMAR.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (1999). *CRISP-DM 1.0 - Step-by-step data mining guide.* SPSS. Retrieved from https://web.archive.org/web/20220401041957/https://www.the-modeling-agency.com/crisp-dm.pdf

Chatfield, C. (2000). *Time-Series Forecasting* (1st ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9781420036206

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (pp. 785-794). New York, NY, USA: ACM. https://doi.org/10.1145/2939672.2939785

Choi, T.-M., Wallace, S., & Wang, Y. (2018). Big Data Analytics in Operations Management. Production and Operations Management. *Production and Operations Management Society, 27*(10), 1868-1883. https://doi.org/10.1111/poms.12838

Chopra, S., & Meindl, P. (2013). *Supply Chain Management* (5th ed.). Pearson Education, Inc.

Cipriani, E., Crescenzi, L., & Nigro, M. (2014). Behavioral Models for the Estimation of the Air Transport Demand: The Case Study of Rome-London Flight Connection. *Procedia - Social and Behavioral Sciences, 111*(5), 78-87. https://doi.org/10.1016/j.sbspro.2014.01.040

Claveria, O., Monte, E., & Torra, S. (2015). *Regional Forecasting with Support Vector Regressions: The Case of Spain. ERN.* ERN: Neural Networks & Related Topics. https://doi.org/10.2139/ssrn.2945533

Claveria, O., Monte, E., & Torra, S. (2016). Modelling tourism demand to Spain with machine learning techniques. The impact of forecast horizon on model selection. *Revista de Economia Aplicada, 24*(72), 109-132. https://doi.org/10.48550/arXiv.1805.00878

Costa, A., Soares, J., Salas-Leiton, E., Bordalo, A., & Costa-Dias, S. (2022). The COVID-19 as a Driver for Alternative Trade Networks in the Small-Scale Fisheries: Portugal as a Case Study. https://doi.org/10.3390/su14116405

Cyprich, O., Konečný, V., & Kiliánová, K. (2013). Short-Term Passenger Demand Forecasting Using Univariate Time Series Theory. *Promet - Traffic&Transportation, 25*(6), 533-541. https://doi.org/10.7307/ptt.v25i6.338

Dairu, X., & Shilong, Z. (2021). Machine Learning Model for Sales Forecasting by Using XGBoost. *2021 IEEE International Conference on Consumer Electronics and*

Computer Engineering (ICCECE), (pp. 480-483). Guangzhou, China. https://doi.org/10.1109/ICCECE51280.2021.9342304

Danese, P., & Kalchschmidt, M. (2011). The role of the forecasting process in improving forecast accuracy and operational performance. *International Journal of Production Economics, 131*(1), 204-214. https://doi.org/10.1016/j.ijpe.2010.09.006

De Vet, J. M., Edwards, J., & Bocci, M. (2016). *Blue Growth and Smart Specialisation: How to catch maritime growth through 'Value Nets'.* S3 Policy Brief Series No. 17/2016. Retrieved from https://s3platform.jrc.ec.europa.eu/documents/20125/248836/Blue+Growth+and+Smart+Specialisation.pdf/f2ed7c31-80b0-a62b-e4b7-7e7e8a192085?t=1621268542601

Djatmiko, M., Husain, A., Maulani, G., & Nirmalasari, L. (2018). Analyze and Record a Series of Corporate Sales Transactions On Web Based Accounting Online System. *Aptisi Transactions on Management*. https://doi.org/10.33050/atm.v1i2.691

Donaldson, B. (1998). The Role of Selling and Sales Management in Marketing. 3-24. https://doi.org/10.1007/978-1-349-26354-7_1

Dutta, G., & Ghosh, P. (2012). A passenger revenue management system (RMS) for a National Railway in an Emerging Asian Economy. *J Revenue Pricing Manag, 11*, 487-499. https://doi.org/10.1057/rpm.2012.10

Dutta, G., & Marodia, D. (2015). *Comparison of forecasting techniques in revenue management for a national railway in an emerging Asian economy. International Journal of Revenue Management, 8*(2), 130-152. https://doi.org/10.1504/IJRM.2015.070000

Dykes, B. (2020). *EFFECTIVE DATA STORYTELLING - How to Drive Change with Data, Narrative, and Visuals.* John Wiley and Sons, Inc.

Egger, R. (2022). *Applied Data Science in Tourism: Interdisciplinary Approaches, Methodologies, and Applications.* Springer. https://doi.org/10.1007/978-3-030-88389-8

Errey, N., Liang, J., Leong, T., & Zowghi, D. (2024). Evaluating narrative visualization: a survey of practitioners. *International Journal of Data Science and Analytics, 18*, 19-34. https://doi.org/10.1007/s41060-023-00394-9

European Travel Commission. (2023). *European Tourism: Trends & Prospects. Quarterly report (Q3/2023).* Brussels: A reported produced by: Tourism Economics. Retrieved from https://etc-corporate.org/uploads/2023/11/ETC-Quarterly-Report-Q3_Public.pdf

EY-AM&A. (2019). *A Economia do Mar em Portugal.* Retrieved from https://ind.millenniumbcp.pt/pt/negocios/financiamento/Documents/Economia-do-Mar-em-Portugal-Apresentacao201811.pdf

Fry, B. (2007). *Visualizing data: Exploring and explaining data with the processing environment.* O'Reilly Media, Inc.

Genuer, R., & Poggi, J.-M. (2020). *Random Forests with R.* Springer. https://doi.org/10.1007/978-3-030-56485-8

Guenzi, P., & Troilo, G. (2007). The joint contribution of marketing and sales to the creation of superior customer value. *Journal of Business Research, 60*, 98-107. https://doi.org/10.1016/J.JBUSRES.2006.10.007

Haas, A., Snehota, I., & Corsaro, D. (2012). Creating value in business relationships: The role of sales. *Industrial Marketing Management, 41*, 94-105. https://doi.org/10.1016/J.INDMARMAN.2011.11.004

Han, J., Kamber, M., & Pei, J. (2012). *Data Mining Concepts and Techniques* (3rd ed.). Morgan Kaufmann.

Harvey, N., & Bolger, F. (1996). Graphs versus tables: Effects of data presentation format on judgemental forecasting. *International Journal of Forecasting*(12), pp. 119-137. https://doi.org/10.1016/0169-2070(95)00634-6.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). Springer Series in Statistics.

Helmini, S., Jihan, N., Jayasinghe, M., & Perera, S. (2019). Sales forecasting using multivariate long short term memory network models. *PeerJ Prepr, e27712v1*. https://doi.org/10.7287/peerj.preprints.27712v1

Honga, W., Dong, Y., Chen, L., & Wei, S. (2011). SVR with hybrid chaotic genetic algorithms for tourism demand forecasting. *Applied Soft Computing, 11*(2), 1881-1890. https://doi.org/10.1016/j.asoc.2010.06.003

Huang, W., Xiao, Q., Dai, H., & Yan, N. (2018). Sales Forecast for O2O Services - Based on Incremental Random Forest Method. *15th International Conference on Service Systems and Service Management (ICSSSM).* https://doi.org/10.1109/icsssm.2018.8465121

Hyndman, R., & Athanasopoulos, G. (2018). *Forecasting: principles and practice* (2nd ed.). OTexts. Retrieved from https://otexts.com/fpp2/

Hyndman, R., & Athanasopoulos, G. (2021). *Forecasting: principles and practice* (3rd ed.). Melbourne, Australia: OTexts. Retrieved June 25, 2024, from OTexts.com/fpp3

Ilić, M., & Šimičević, D. (2022). FUTURE OF TOURISM AND COVID-19 CRISIS: THE PARADIGM SHIFT. *The Seventh International Scientific Conference - THE FUTURE OF TOURISM.* https://doi.org/10.52370/tisc22484mi

Iliinsky, N., & Steele, J. (2011). *Designing data visualizations: Representing informational Relationships.* O'Reilly Media, Inc.

INE. (2023). *Estatísticas do Turismo - 2022.* Lisboa: Instituto Nacional de Estatística, I.P. Retrieved from https://www.ine.pt/xurl/pub/72384938

INE. (2024). *Dormidas (N.º) nos estabelecimentos de alojamento turístico - Faro Jan/2017 - Jun/2024.* Retrieved from https://www.ine.pt/xportal/xmain?xpid=INE&xpgid=ine_indicadores&indOcorrCod=0009808&contexto=bd&selTab=tab2

INE. (2024). *Passageiros desembarcados (N.º) nos Aeroportos - Faro Jan/2017 - Jun/2024.* Retrieved from https://www.ine.pt/xportal/xmain?xpid=INE&xpgid=ine_indicadores&contecto=pi&indOcorrCod=0000862&selTab=tab0&xlang=pt

Jain, A., Menon, M., & Chandra, S. (2015). Sales forecasting for retail chains. *California: UC San Diego Jacobs School of Engineering.* San Diego. Retrieved from https://cseweb.ucsd.edu/classes/wi15/cse255-a/reports/fa15/004.pdf

Jamal, S., Goyal, S., Grover, A., & Shanker, A. (2018). Machine Learning: What, Why, and How? In A. Shanker, *Bioinformatics: Sequences, Structures, Phylogeny* (pp. 359-374). Springer Nature Singapore. https://doi.org/10.1007/978-981-13-1562-6_16

James, S., Zhang, Y., & O'Donncha, F. (2018). A machine learning framework to forecast wave conditions. In *Coastal Engineering* (Vol. 137, pp. 1-10). Elsevier. https://doi.org/10.1016/j.coastaleng.2018.03.004

Jiao, E., & Chen, J. (2018). Tourism forecasting: A review of methodological developments over the last decade. *Tourism Economics, 25*, 469-492. https://doi.org/10.1177/1354816618812588

Jiao, P., Li, R., Sun, T., Hou, Z., & Ibrahim, A. (2016). Three Revised Kalman Filtering Models for Short-Term Rail Transit Passenger Flow Prediction. *Mathematical Problems in Engineering, 16*. https://doi.org/10.1155/2016/9717582

Ke, J., Zheng, H., Yang, H., & Chen, X. (2017). Short-term forecasting of passenger demand under on-demand ride services: A spatio-temporal deep learning approach. *Transportation Research Part C: Emerging Technologies, 85*, pp. 591-608. https://doi.org/10.1016/j.trc.2017.10.016

Kelleher, J., Mac Namee, B., & D'Arcy, A. (2020). *Fundamentals of Machine Learning for Predictive Data Analytics - Algorithms, Worked Examples, and Case Studies* (2nd ed.). Cambridge: The MIT Press.

Kosara, R., & Mackinlay, J. (2013). Storytelling: The Next Step for Visualization. In *Computer* (Vol. 46, pp. 44-50). IEEE Computer Society. https://doi.org/10.1109/MC.2013.36

Kovačić, M., & Silveira, L. (2018). *Nautical Tourism in Croatia and in Portugal in the Late 2010's.* Pomorstvo. https://doi.org/10.31217/P.31.2.13

Law, R., Li, G., Ka Chio Fong, D., & Han, X. (2019). Tourism demand forecasting: A deep learning approach. *Annals of Tourism Research, 75*, 410-423. https://doi.org/10.1016/j.annals.2019.01.014

Lee, C.-K., Song, H., & Mjelde, J. (2008). The forecasting of International Expo tourism using quantitative and qualitative techniques. *Tourism Management, 29*(6), 1084-1098. https://doi.org/10.1016/j.tourman.2008.02.007

Leng, N., Nie, L., Guo, G., & Wu, X. (2015). Passenger flow forecasting for chinese high speed rail network. *Proceedings of the 2015 International Conference on Mechatronics, Electronic, Industrial and Control Engineering (MEIC-15).* Atlantis Press. https://doi.org/10.2991/meic-15.2015.153

Lim, C. (2006). A survey of tourism demand modelling practice: Issues and implications. In *International handbook on the economics of tourism* (pp. 45-72). Edward Elgar Publishing Limited. https://doi.org/10.4337/9781847201638.00008

Lin, C., & Lee, T. (2013). Tourism demand forecasting: Econometric model based on multivariate adaptive regression splines, artificial neural network and support vector regression. *Advances in Management and Applied Economics, 3*(6), 1-18. Retrieved from https://www.scienpress.com/Upload/AMAE/Vol%203_6_1.pdf

Liu, J., Liu, C., Zhang, L., & Xu, Y. (2019). Research on sales information prediction system of e-commerce enterprises based on time series model. *Liu, J., Liu, C., Zhang, L., & Xu, Y. (2019). Research on sales information prediction system of e-commerInformation Systems and e-Business Management, 18*, 823-836. https://doi.org/10.1007/s10257-019-00399-7

Liu, R., & Li, A. (2012). Forecasting high-speed rail ridership using a simultaneous modeling approach. *Transportation Planning and Technology, 35*(5), 577-590. https://doi.org/10.1080/03081060.2012.701816

Loureiro, A., Miguéis, V., & Da Silva, L. (2018). Exploring the use of deep neural networks for sales forecasting in fashion retail. *Decision Support Systems, 114*, 81-93. https://doi.org/10.1016/j.dss.2018.08.010

Makasarashvili, T., & Giguashvili, G. (2021). THE MAIN CHALLENGES OF DIGITAL COMMERCE IN THE CONDITIONS OF "COVID-19". *Globalization and Business*. https://doi.org/10.35945/gb.2021.11.020

Makridakis, S. (1996). Forecasting: its role and value for planning and strategy. *International Journal of Forecasting, 12*(4), 513-537. https://doi.org/10.1016/S0169-2070(96)00677-2

Mallikharjuna Rao, K., Saikrishna, G., & Supriya, K. (2023). Data preprocessing techniques: emergence and selection towards machine learning models - a practical review using HPA dataset. In *Multimed Tools Appl* (pp. 37177–37196). https://doi.org/10.1007/s11042-023-15087-5

Martins, P. (2015). Sea bathing and seaside tourism in Portugal in the nineteenth and twentieth centuries: an overview. *Journal of Tourism History, 7*, 246 - 267. https://doi.org/10.1080/1755182X.2015.1114685

McKinney, W. (2010). Data structures for statistical computing in python. *Proceedings of the 9th Python in Science Conference. 445*, pp. 51–56. SCIPY 2010. https://doi.org/10.25080/Majora-92bf1922-00a

Meneghini, M., Anzanello, M., Kahmann, A., & Tortorella, G. (2018). Quantitative demand forecasting adjustment based on qualitative factors: case study at a fast food restaurant. https://doi.org/10.20985/1980-5160.2018.v13n1.1188

Mentzer, J., & Bienstock, C. (1998). *Sales Forecasting Management: Understanding the Techniques, Systems and Management of the Sales Forecasting Process.* SAGE Publications, Inc.

Mentzer, J., & Moon, M. (2005). *Sales Forecasting Management: A Demand Management Approach.* https://doi.org/10.4135/9781452204444

Miah, S., Vu, H., Gammack, J., & McGrath, M. (2017). A Big Data Analytics Method for Tourist Behaviour Analysis. *Inf. Manag, 54*, pp. 771-785. https://doi.org/10.1016/j.im.2016.11.011

Ministério da Economia e da Inovação. (2007). *Plano Estratégico Nacional do Turismo.*

Ministério da Economia e da Inovação. (2017). *Estratégia Turismo 2027.*

Ministério do Mar. (2021). *Estratégia Nacional para o Mar 2021-2030.*

Mirović, M., Miličević, M., & Obradović, I. (2018). Big Data in the Maritime Industry. *NAŠE MORE : znanstveni časopis za more i pomorstvo, 65*(1), pp. 56-62. https://doi.org/10.17818/NM/2018/1.8

Morlidge, S., & Player, S. (2010). *Future ready: How to master business forecasting.* John Wiley & Sons.

Mukhopadhyay, S., Samaddar, S., & Colville, G. (2007). Improving Revenue Management Decision Making for Airlines by Evaluating Analyst-Adjusted Passenger Demand Forecasts. *Decision Sciences, 38*, 309-327. https://doi.org/10.1111/j.1540-5915.2007.00160.x

Nieto, M., & Benitez, R. (2018). ARIMA + GARCH + Bootstrap forecasting method applied to the airline industry. *Journal of Air Transport Management, 71*(5). https://doi.org/10.1016/j.jairtraman.2018.05.007

Oliva, R., & Watson, N. (2009). Managing Functional Biases in Organizational Forecasts: A Case Study of Consensus Forecasting in Supply Chain Planning. *Production and Operations Management, 18*(2), 138-151. https://doi.org/10.1111/j.1937-5956.2009.01003

Osisanwo, F., Akinsola, J., Awodele, O., Hinmikaiye, J., Olakanmi, O., & Akinjobi, J. (2017). Supervised Machine Learning Algorithms: Classification and Comparison. *International Journal of Computer Trends and Technology*, 128-138. https://doi.org/10.14445/22312803/IJCTT-V48P126

Pan, J., Nichols, D., & Joy, O. (1977). Sales Forecasting Practices of Large U.S. *Industrial Firms. Financial Management, 6*(72). https://doi.org/10.2307/3665259

Panagopoulos, N., & Avlonitis, G. (2010). Performance implications of sales strategy: The moderating effects of leadership and environment. *International Journal of Research in Marketing, 27*, 46-57. https://doi.org/10.1016/J.IJRESMAR.2009.11.001

Pataropura, A., Riki, R., & Saputra, A. (2019). Sales Analysis Using the Forecasting Method. *bit-Tech*. https://doi.org/10.32877/BT.V1I3.79

Pelham, A. (2000). Market Orientation and Other Potential Influences on Performance in Small and Medium-Sized Manufacturing Firms. *Journal of Small Business Management, 38*, 48.

Pereira, E., Mascarenhas, M., Flores, A., & Pires, G. (2015). Nautical small-scale sports events portfolio: a strategic leveraging approach. *European Sport Management Quarterly, 15*, 27-47. https://doi.org/10.1080/16184742.2015.1007883

Plácido, B., Proença, S., Moreira, D., Boução, L., Branco, F., & Au-Yong-Oliveira, M. (2021). E-commerce During the COVID-19 Pandemic: An Exploratory Study Involving Portuguese Consumers. pp. 422-434. https://doi.org/10.1007/978-3-030-72654-6_41

Pustokhina, I., & Pustokhin, D. (2023). A Comparative Analysis of Traditional Forecasting Methods and Machine Learning Techniques for Sales Prediction in E-commerce. *American Journal of Business and Operations Research (AJBOR), 10*(2), 39-51. https://doi.org/10.54216/AJBOR.100205

Reitermanová, Z. (2010). Data splitting. (MATFYZPRESS, Ed.) *WDS'10 Proceedings of Contributed Papers, Part I*, 31-36.

Ressurreição, A., Rangel, M., Oliveira, F., Monteiro, P., Bentes, L., Pontes, J., . . . Andrade M.; Horta e Costa B.; Gonçalves J.M.S. (2020). AMPICvalue. *Área Marinha Protegida de Interesse Comunitário (AMPIC)*.

Riedel, S., & Gabrys, B. (2003). Adaptive Mechanisms in an Airline Ticket Demand Forecasting System. *EUNITE'2003 Conference: European Symposium on Intelligent Technologies, Hybrid Systems and their implementation on Smart Adaptive Systems.* Oulu,Finland. Retrieved from https://eprints.bournemouth.ac.uk/8533/

RNAAT. (2024). *Registo Nacional de Agentes de Animação Turística*. Retrieved from Turismo de Portugal: https://rnt.turismodeportugal.pt/RNT/Pesquisa_AAT.aspx

Ryan, L. (2018). *Visual data storytelling with Tableau.* Addison-Wesley Professional.

Sa, J. (1987). Reservations forecasting in airline yield management. *M.S. thesis, Massachusetts Institute of Technology, Dept. of Aeronautics and Astronautics*(16877756). Retrieved from http://hdl.handle.net/1721.1/68058

Sagaert, Y., Aghezzaf, E., Kourentzes, N., & Desmet, B. (2018). Tactical sales forecasting using a very large set of macroeconomic indicators. *Eur. J. Oper. Res, 264*, 558-569. https://doi.org/10.1016/j.ejor.2017.06.054

Sagaert, Y., Aghezzaf, E., Kourentzes, N., & Desmet, B. (2018). Temporal Big Data for Tactical Sales Forecasting in the Tire Industry. *Interfaces, 49*, 121-129. https://doi.org/10.1287/inte.2017.0901

Samagaio, A., & Wolters, M. (2010). Comparative analysis of government forecasts for the Lisbon Airport. *Journal of Air Transport Management, 16*(4), 213-217. https://doi.org/10.1016/j.jairtraman.2009.09.002

Santos, E., Lisboa, I., Crespo, C., Moreira, J., & Eugenio, T. (2022). Evaluating Economic Sustainability of Nautical Tourism Through Brand Equity and Corporate Performance. *Transcending Borders in Tourism Through Innovation and Cultural Heritage 8th International Conference* (pp. 105-118). Hydra, Greece: Springer Proceedings in Business and Economics. https://doi.org/10.1007/978-3-030-92491-1_6

Schölkopf, B., & Smola, A. (2001). Learning with Kernels - Support Vector Machines, Regularization, Optimization and Beyond. *The MIT Press*. https://doi.org/10.7551/mitpress/4175.001.0001

Schölkopf, B., & Smola, A. (2003). A Short Introduction to Learning with Kernels. *Advanced Lectures on Machine Learning. Lecture Notes in Computer Science. 2600.* Berlin, Heidelberg: Springer. https://doi.org/10.1007/3-540-36434-X_2

See-To, E., & Ngai, E. (2018). Customer reviews for demand distribution and sales nowcasting: a big data approach. *Big Data Analytics in Operations & Supply Chain Management, 270*, 415-431. https://doi.org/10.1007/s10479-016-2296-z

Segel, E., & Heer, J. (2010). Narrative Visualization: Telling Stories with Data. In *IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS* (Vol. 16, pp. 1139-1148). IEEE Computer Society. https://doi.org/10.1109/TVCG.2010.179

Shiman, X. (2023). Comparison of Sales Prediction in Conventional Insights and Machine Learning Perspective. *Psychology Research, 13*(3), 146-156. https://doi.org/10.17265/2159-5542/2023.03.005

Shmueli, G., & Lichtendahl Jr., K. (2016). *Practical Time Series Forecasting with R: A Hands-On Guide* (2nd ed.). Axelrod Schnall Publishers.

Sickles, R., Good, D., Postert, A., & Getachew, L. (1998). A model of world aircraft demand. *In Airport Facilities: Innovations for the Next Century. Proceedings of the 25th International Air Transportation Conference.* American Society of Civil Engineers. Retrieved from https://trid.trb.org/View/506354

Sohrabpour, V., Oghazi, P., Toorajipour, R., & Nazarpour, A. (2020). Export sales forecasting using artificial intelligence. *Technological Forecasting and Social Change*. https://doi.org/10.1016/j.techfore.2020.120480

Song, H., & Turner, L. (2006). Tourism demand forecasting. In *International handbook on the economics of tourism.* (pp. 89-114). Edward Elgar Publishing Limited. https://doi.org/10.4337/9781847201638.00010

Stein, J. C. (1997). Internal capital markets and the competition for corporate resources. *Journal of Finance, 52*, 111-133. https://doi.org/10.2307/2329558

Su, A. (2021). Tourism Marketing Innovation Management Model Based on Big Data. *Journal of Physics: Conference Series, 1744*. https://doi.org/10.1088/1742-6596/1744/4/042141

Sun, S., Wei, Y., Tsui, K., & Wang, S. (2019). Forecasting tourist arrivals with machine learning and internet search index. In *Tourism Management* (Vol. 70, pp. 1-10). Elsevier. https://doi.org/10.1016/j.tourman.2018.07.010

Sun, X., Gauri, D., & Webster, S. (2011). Forecasting for cruise line revenue management. *J Revenue Pricing Manag, 10*, 306-324. https://doi.org/10.1057/rpm.2009.55

Suryawanshi, R., Musale, S., & Bhosa, S. (2024). Comparative Analysis of use of Machine Learning Algorithm for Prediction of Sales. *J. Electrical Systems, 20*(3), 851-863. https://doi.org/10.52783/jes.1383

Tairov, I., & Petrova, M. (2022). Challenges and Opportunities in Electronic Business During the Covid-19 Pandemic. *Tairov, I., & Petrova, M. (2022). Challenges and Opportunities in EleIEEE 9th International Conference on Problems of Infocommunications, Science and Technology (PIC S&T)*, (pp. 231-235). https://doi.org/10.1109/PICST57299.2022.10238561

Terho, H., Eggert, A., Haas, A., & Ulaga, W. (2015). How sales strategy translates into performance: The role of salesperson customer orientation and value-based selling. *Industrial Marketing Management, 45*, 12-21. https://doi.org/10.1016/J.INDMARMAN.2015

Terho, H., Haas, A., Eggert, A., & Ulaga, W. (2012). 'It's almost like taking the sales out of selling'—Towards a conceptualization of value-based selling in business markets. *Industrial Marketing Management, 41*, 174-185. https://doi.org/10.1016/J.INDMARMAN.2011.11.011

Treboux, J., Cretton, F., Evéquoz, F., Calvé, A., & Genoud, D. (2016). Mining and Visualizing Social Data to Inform Marketing Decisions. *EEE 30th International Conference on Advanced Information Networking and Applications (AINA)*, 66-73. https://doi.org/10.1109/AINA.2016.130

Tripathi, S., Yadav, S., & Kumar, H. (2019). Hardesh Kumar. A comparative analysis of traditional time series models and deep learning approaches in sales prediction: Strengths and limitations. *The Pharma Innovation Journal, 8*(2), 20-24. https://doi.org/10.22271/tpi.2019.v8.i2Sa.25244

Tsekeris, T., & Tsekeris, C. (2011). Demand Forecasting in Transport: Overview and Modeling Advances. *Economic Research-Ekonomska Istraživanja, 24*(1), 82-94. https://doi.org/10.1080/1331677X.2011.11517446

Turismo de Portugal. (2022). *Animação Turística em Portugal | 2021 - Inquérito de caracterização da procura dos agentes de animação turística.* Retrieved from https://travelbi.turismodeportugal.pt/animacao-turistica/animacao-turistica-em-portugal-2021/

Turismo de Portugal. (2023). *Turismo em Números | 2018.* Retrieved from https://travelbi.turismodeportugal.pt/turismo-em-portugal/turismo-em-numeros-2018/

Two Octobers. (2024, June 6). *8 Data Storytelling Concepts with Examples*. Retrieved from Two Octobers: https://twooctobers.com/blog/8-data-storytelling-concepts-with-examples/

Tyralis, H., & Papacharalampous , G. (2017). Variable Selection in Time Series Forecasting Using Random Forests. *Algorithms, 10*(114). https://doi.org/10.3390/a10040114

Vázquez, R., García, J., & Valenciano, J. (2021). Analysis and Trends of Global Research on Nautical, Maritime and Marine Tourism. *Journal of Marine Science and Engineering*. https://doi.org/10.3390/JMSE9010093.

Verstraete, G., Aghezzaf, E., & Desmet, B. (2020). A leading macroeconomic indicators' based framework to automatically generate tactical sales forecasts. *Computers & Industrial Engineering, 139*, 106-169. https://doi.org/10.1016/j.cie.2019.106169

Wang, B. (2016). Research on cruise pricing strategy based on Forecast. *In 2015 5th International Conference on Computer Sciences and Automation Engineering (ICCSAE 2015)* (pp. 99-102). Atlantis Press. https://doi.org/10.2991/iccsae-15.2016.21

Wang, Y., Chen, X., Han, Y., & Guo, S. (2013). Forecast of Passenger and Freight Traffic Volume based on Elasticity Coefficient Method and Grey Model. *Procedia - Social and Behavioral Sciences, 96*, 136-147. https://doi.org/10.1016/j.sbspro.2013.08.019

Wickham, R. (1995). *Evaluation of forecasting techniques for short-term demand of air transportation.* (Doctoral dissertation, Massachusetts Institute of Technology).

World Economic Forum. (2022, May 24). *"People want to travel": 4 sector leaders say that tourism will change and grow.* Retrieved from World Economic Forum: https://www.weforum.org/agenda/2022/05/future-of-travel-tourism-sector-leaders-report/

WTTC. (2022, June 13). *Portugal's Travel & Tourism Could Surpass Pre-pandemic Levels in 2023.* Retrieved from World Travel & Tourism Council: https://wttc.org/news-article/portugals-travel-and-tourism-could-surpass-pre-pandemic-levels-in-2023

Xie, G., Qian, Y., & Wang, S. (2021). Forecasting Chinese cruise tourism demand with big data: An optimized machine learning approach. In *Tourism Management* (Vol. 82, p. 104208). Elsevier. https://doi.org/10.1016/j.tourman.2020.104208

YachtSys. (2020, December 01). *Takeaways from the Boat Rental Season 2020*. Retrieved from https://www.yachtsys.com/Takeaways-from-the-Boat-Rental-Season-2020.html

Zhang, Y., Reynolds, M., Lugmayr, A., Damjanov, K., & Hassan, G. (2022). A Visual Data Storytelling Framework. *Informatics, 9*(73). https://doi.org/10.3390/informatics9040073

Zhou, X., Liu, Z., Wang, F., Xie, Y., & Zhang, X. (2020). Using Deep Learning to Forecast Maritime Vessel Flows. *Sensors, 20*. https://doi.org/10.3390/s20061761

Zoltners, A., Sinha, P., & Lorimer, S. (2006). Match your sales force structure to your business life cycle. *Harvard business review, 84*, 80-89,188.