NOVA IMS Information Management School

MDSAA

Master Degree Program in

Data Science and Advanced Analytics

Analysis of Monetary Transactions in Oeste CIM

Insights into Tourism Spending Patterns and Economic Impact

Alexandre Pires Spagnol

Project Work

presented as partial requirement for obtaining a Master's Degree in Data Science and Advanced Analytics

NOVA Information Management School

Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

NOVA Information Management School

Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

Analysis of Monetary Transactions in Oeste CIM

Insights into Tourism Spending Patterns and Economic Impact

by

Alexandre Pires Spagnol

Project Work presented as partial requirement for obtaining the Master's degree in Data Science and Advanced Analytics, with a specialization in Business Analytics

Supervised by

Bruno Jardim, PhD, NOVA Information Management School

Duarte Rodrigues, MSc, NOVA Information Management

July, 2025

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I have fully acknowledged the Rules of Conduct and Code of Honor from the NOVA Information Management School.

[Lisbon, 15/07/2025]

DEDICATION

To Paloma, your constant patience and unwavering support during difficult times have been a refuge for me.

To my dear parents, your lifelong encouragement laid the foundation for this accomplishment.

To my treasured sister, your joyful nature has helped me overcome challenges.

To my invaluable colleagues and friends, Filipe, Sebastião, Hugo, and Gonçalo, your friendship turned obstacles into shared victories.

This work is a testament to your collective faith in me. I am deeply thankful for your sacrifices, laughter, and steadfast belief throughout this journey.

ACKNOWLEDGEMENTS

I express my profound gratitude to my supervisor, Professor Bruno Jardim, for his invaluable guidance, critical insights, and unwavering support throughout this research endeavor. His expertise in data-driven regional analysis has significantly influenced this work.

I am equally indebted to my co-supervisor, Duarte Rodrigues, for his technical mentorship and rigorous feedback during the pivotal stages of this project.

Special thanks are due to NOVA Cidade - Urban Analytics Lab for providing access to the transactional dataset through the Smart Region project.

I also acknowledge the NOVA IMS Ethics Committee for their diligent review and approval.

I extend my warm thanks to my colleagues Margarida Leitão and David Garcia for their insightful opinions, technical guidance, and shared commitment under Professor Jardim's supervision. My appreciation also extends to Filipe, Sebastião, Hugo, and Gonçalo for their camaraderie and collaborative problem-solving.

The NOVA Information Management School deserves recognition for fostering an environment of intellectual excellence.

Finally, to my family and Paloma: your emotional support made this achievement possible.

ABSTRACT

This study addresses a significant gap in microeconomic tourism analysis for small regions by developing a scalable, data-driven framework for Oeste CIM, Portugal, a region heavily reliant on tourism and subject to seasonal economic fluctuations. Utilizing transactional records from 2021 to 2024, we implemented a Lakehouse architecture with Kimball dimensional modelling to facilitate parish-level spending analysis. Automated ETL pipelines process payment data, and CatBoost forecasting models predict tourism trends with a mean absolute percentage error (MAPE) of less than 15%, surpassing traditional methods. The integrated Power BI dashboard reveals critical microeconomic insights: weekends account for 70% of spending, and sectors such as "Petrol Stations" and "Supermarkets" exhibit high economic resilience. The findings validate the effectiveness of machine learning for forecasting in small regions despite data limitations and provide stakeholders with actionable tools for resource optimization. This framework transforms transactional data into strategic insights, enabling municipalities and businesses to mitigate seasonal risks, enhance infrastructure planning, and advance the Sustainable Development Goals (SDG 8–9). This solution offers a replicable blueprint for data-driven urban planning in tourism economies globally.

KEYWORDS

smart tourism; smart region; forecasting; microeconomic indicators; business intelligence

Sustainable Development Goals (SDG):

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

API	Application programming interface
ARIMA	AutoRegressive Integrated Moving Average
BI	Business Intelligence
CRISP-DM	Cross-industry standard process for data mining
DIM	Dimension
ETL	Extract, Transform, Load
FCT	Fact table
FK	Foreign Key
GDP	Gross domestic product
IQR	Interquartile Range
MAPE	Mean absolute percentage error
Oeste CIM	Comunidade Intermunicipal do Oeste
PPP	Public-private partnerships
RFM	Recency, Frequency, and Monetary
RMSE	Root mean squared error
SK	Surrogate Key

1. INTRODUCTION

Tourism plays a fundamental role in the economic development of many regional economies, serving as a primary driver of income, employment, and investment. The sector is inherently volatile, subject to seasonal fluctuations, consumer preferences, and exogenous shocks such as economic crises or pandemics. These challenges are particularly acute in Portugal's Oeste CIM, a coastal intermunicipal community whose economy relies heavily on seasonal tourism and heritage attractions. While tourism offers growth opportunities, it exposes the region to vulnerabilities, especially during off-peak seasons when economic activity diminishes.

Despite tourism's strategic importance, policymaking in small, tourism-dependent regions remains constrained by limited granular, real-time data. Traditional macroeconomic indicators like GDP growth and employment rates, while useful nationally, fail to provide actionable insights at municipal or parish levels. These limitations hinder authorities' ability to allocate resources, plan infrastructure, and design targeted interventions aligned with behavioural patterns. The lack of detailed microeconomic data restricts municipalities from addressing volatility and planning challenges effectively.

This thesis addresses this gap by developing a scalable data-driven framework to analyse and forecast tourism-related economic activity in Oeste CIM, using card-based transactional data from SIBS Analytics (2021-2024). The approach integrates microeconomic signals—transaction values, cardholder frequency, temporal distribution, and geographic origin—into a structured decision-support architecture. Using Kimball Lifecycle methodology, the project builds a Lakehouse-based data warehouse processing high-volume transactional data at parish level. Automated ETL pipelines transform POS and ATM transaction records into a dimensional data model, enabling temporal and geographic disaggregation for operational monitoring and strategic planning.

The project employs CatBoost forecasting models to predict short-term tourism spending trends. These models, trained on historical transaction data, achieve predictive accuracy with MAPE below 15%, outperforming traditional time-series methods. The predictive outputs and transactional insights are visualized through an interactive Power BI dashboard, providing stakeholders with real-time access to key economic indicators including peak spending times, sector resilience, and tourist activity distribution.

The work aims to bridge the gap between data availability and actionable intelligence in small regional economies. By converting transactional records into interpretable microeconomic indicators, this project enables responsive, precise data-driven governance. It aligns with UN Sustainable Development Goals (SDG 8 and 9) by promoting sustainable urban planning, economic resilience, and evidence-based management. This framework, while tailored to Oeste CIM, serves as a model for other tourism-dependent regions globally, providing a blueprint for leveraging digital transaction data for regional development.

2. LITERATURE REVIEW

2.1. EXPLORING THE IMPACT OF TOURISM ON LOCAL ECONOMIES

The economic impact of tourism on smaller regional economies, particularly in tourism-centric areas like “Comunidade Intermunicipal do Oeste” (Oeste CIM), is multifaceted and can significantly influence local businesses and infrastructure. Seasonal tourism patterns play a critical role in shaping local economic activity. In regions heavily reliant on tourism, fluctuations in visitor numbers can lead to pronounced variations in consumer spending, directly affecting local businesses. For instance, during peak seasons, businesses often experience a surge in sales, leading to increased employment opportunities and higher revenues. Conversely, during off-peak seasons, these businesses may struggle to maintain profitability, which can result in layoffs and reduced service offerings. (Ridderstaat et al., 2013; Suleiman & Albiman, 2014). This reliance on seasonal tourism underscores the need for strategic planning to mitigate the economic downturns experienced during off-peak periods, emphasizing the importance of diversifying local economies and attracting year-round visitors (Adhuze, 2023).

To address these challenges, descriptive and predictive models have been employed in various regions to forecast spending behaviour and the economic impacts of tourism. For example, machine learning techniques have been utilized to analyse historical data and predict future tourism trends, allowing local governments to make informed decisions regarding infrastructure investments and marketing strategies (Zhou, 2021). Additionally, Granger causality tests have been applied to establish the dynamic relationships between tourism and economic growth, providing insights into how changes in tourism activity can influence local economic conditions (Liu & Song, 2017). These models not only help in understanding past behaviours but also play a crucial role in anticipating future trends, which is essential for effective urban planning and resource allocation.

Building on these predictive capabilities, transactional data in tourism-related urban planning is increasingly recognized as the best practice. By analysing consumer spending patterns, local authorities can better understand the needs and preferences of tourists, which can inform the development of infrastructure and services that enhance the tourist experience (Partarakis et al., 2023). Moreover, public-private partnerships (PPPs) have also been highlighted as effective mechanisms for improving tourism infrastructure, as they leverage both public resources and private sector innovation to create sustainable tourism solutions (Svetlana & Vladimir, 2019; Yunus et al., 2021). Furthermore, integrating data analytics into urban planning processes allows for a more responsive approach to tourism management, ensuring that local economies can adapt to changing market conditions and consumer preferences (Partarakis et al., 2023).

In summary the economic impact of tourism on smaller regional economies like Oeste CIM is significant, influenced by seasonal patterns and consumer spending behaviours. The application of descriptive and predictive models aids in forecasting economic trends, while the strategic use of transactional data in urban planning fosters sustainable tourism development. By understanding these dynamics, local policymakers can enhance the resilience and growth of their economies in the face of fluctuating tourism demands. A summary of this chapter and the used literature can be found in Table 2.1.

Table 2.1 - Summary table about the "Exploring the impact of tourism on local economies" LR chapter

Authors	Main Indicators	Methodology	Key Findings
(Ridderstaat et al., 2013)	Seasonal tourism patterns, consumer spending	Empirical analysis	Seasonal tourism significantly influences local businesses, affecting employment and revenues.
(Suleiman & Albiman, 2014)	Employment rates, business profitability	Case studies	Off-peak seasons pose challenges for businesses, highlighting the need for strategic planning.
(Adhuze, 2023)	Year-round visitor attraction, economic resilience	Literature review	Diversifying local economies is essential to mitigate economic downturns during off-peak tourism periods.
(Zhou, 2021)	Historical data analysis, tourism trends	Machine learning	Machine learning improves the accuracy of tourism trend forecasts, aiding infrastructure investment.
(Liu & Song, 2017)	Causality between tourism and	Granger causality tests	Tourism activity changes significantly influence local economic conditions.

	economic indicators		
(Partarakis et al., 2023)	Consumer preferences, infrastructure development	Data analysis	Consumer spending patterns guide infrastructure planning to improve tourist experiences.

2.2. URBAN PLANNING THROUGH DATA-DRIVEN DECISION MAKING

Data-driven decision-making is increasingly recognized as a transformative approach in urban planning, particularly in tourism-dependent regions. At the heart of this transformation lies transactional data, which captures spending patter. The effective use of transaction data and spending patterns plays a crucial role in optimizing urban infrastructure during high-demand tourism periods. By analysing transaction data, local authorities can pinpoint peak spending times and areas of concentrated tourist activity, enabling them to allocate resources more efficiently. For instance, understanding where and when tourists are spending can inform decisions regarding the placement of amenities, public services, and transportation options to enhance visitor experiences and manage congestion (Yallop & Séraphin, 2020). Beyond immediate logistics, integrating big data analytics into urban planning not only aids in resource allocation but also supports sustainable tourism practices by minimizing the environmental footprint associated with increased tourist activity (Law et al., 2019). The integration of big data analytics into urban planning not only aids in resource allocation but also supports sustainable tourism practices by minimizing the environmental footprint associated with increased tourist activity (Law et al., 2019).

Building on this data-centric foundation, several cities have successfully utilized business intelligence tools, such as Power BI and Tableau, to forecast and plan for tourism events. These platforms enable urban planners to visualize data trends, analyse spending patterns, and simulate different scenarios based on historical data. For instance, cities like Barcelona and Amsterdam have implemented such tools to enhance their tourism management strategies, allowing them to anticipate visitor numbers and optimize resource allocation during major events (Yallop & Séraphin, 2020). The integration of business intelligence (BI) in urban planning facilitates a more informed decision-making process, ensuring that cities can adapt to the fluctuating demands of tourism while maintaining the quality of life for residents (Yoon & Choi, 2023). Moreover, the use of data analytics allows for dynamic adjustments to be made in response to changing tourist behaviours, thereby improving overall urban resilience.

Ultimately, data-driven decision-making significantly enhances urban planning in tourism-dependent regions. The strategic use of transaction data and real-time analytics allows local authorities to optimize infrastructure and service allocation effectively. Furthermore, the adoption of business intelligence tools provides valuable insights that enable cities to forecast tourism trends and plan accordingly. As tourism continues to evolve, the integration of data analytics will be essential for sustainable urban development and effective management of tourism-related challenges. A summary of this chapter and the used literature can be found in Table 2.2.

Table 2.2 - Summary table about the "Urban planning through data-driven decision making" LR chapter

Authors	Main Indicators	Methodology	Key Findings
(Yallop & Séraphin, 2020)	Seasonal trends, cyclical patterns	Time series analysis	ARIMA models effectively capture seasonal trends in tourism data.
(Law et al., 2019)	ICT applications, research trends	Systematic review	This study provides a comprehensive overview of ICT research in hospitality and tourism from 2014 to 2017, highlighting trends and gaps in the literature
(Yoon & Choi, 2023)	Contextual factors, real-time data	Experimental study	The study confirms that real-time information such as weather and season significantly impacts tourist recommendations, suggesting future expansions of the system
(Li et al., 2022)	Non-linear relationships, predictive accuracy	Machine learning	Deep learning techniques outperform traditional forecasting methods in tourism demand prediction.

2.3. DEVELOPING SCALABLE DATA-DRIVEN SYSTEMS FOR SMALL REGIONS

Existing literature on smart cities, data analytics, and urban planning methodologies can significantly inform the development of scalable data-driven systems for analysing monetary transactions and supporting urban planning in small, tourism-reliant areas like Oeste CIM. While high-impact studies highlight the transformative potential of descriptive and predictive analytics, they also underscore the unique challenges of adapting these frameworks to smaller, resource-constrained contexts.

A central challenge lies in integrating disparate data sources and establishing robust analytical frameworks. Kitchin critiques gaps in smart city research, arguing that many initiatives lack empirical grounding (Kitchin, 2014). This critique highlights the importance of contextualizing data-driven initiatives within the specific socio-economic landscapes of small regions, particularly those reliant on tourism.

Building on this critique, Bibri's work on smart sustainable cities underscores the transformative potential of data-driven technologies in urban planning. He argues that these technologies can facilitate the balance of sustainability goals by providing insights that help policymakers navigate the challenges of urbanization (Bibri, 2021). This potential is particularly salient for Oeste CIM, where tourism dynamics fluctuate significantly, necessitating adaptable planning strategies.

To operationalize these insights customer behaviour analytics, particularly through techniques like Recency, Frequency, and Monetary (RFM) analysis, can provide valuable insights into tourist spending patterns. Chen et al. demonstrates how RFM analysis can effectively categorize customers based on their transaction histories, which can be instrumental in local businesses aiming to tailor their services to meet tourist demands (Chen et al., 2017). This approach can enhance economic resilience in small regions by optimizing marketing strategies and resource allocation.

Furthermore, the scalability of data-driven systems is essential for adapting to fluctuating tourism levels. Bibri and Krogstie's review of big data analytics technologies highlights the core enabling technologies that can enhance urban planning processes, suggesting that these technologies can be integrated into existing frameworks to improve responsiveness to changing economic conditions (Bibri & Krogstie, 2017). This adaptability is vital for small regions that experience seasonal tourism fluctuations, as it allows for adjustments in planning and resource management.

Practical methodologies further bridge theory and implementation. Case studies that employ data mining methodologies, such as the cross-industry standard process for data mining (CRISP-DM) framework, provide practical insights into regional economic planning. Bibri's exploration of urban computing and intelligence illustrates how data-driven approaches can inform strategic planning and optimize urban designs for sustainability (Bibri, 2021). These

methodologies can be particularly beneficial for small regions, as they offer structured processes for analysing complex data sets and deriving actionable insights.

In summary, the integration of scalable data-driven systems in small, tourism-reliant areas like Oeste CIM presents both challenges and opportunities. By leveraging insights from high-impact research on smart cities, data analytics, and urban planning methodologies, these regions can enhance their economic resilience and improve urban planning outcomes. A summary of this chapter and the used literature can be found in Table 2.3.

Table 2.3 - Summary table about the "Developing scalable data-driven systems for small regions" LR chapter

Authors	Main Indicators	Methodology	Key Findings
(Kitchin, 2014)	Data integration, analytical frameworks	Literature review	Empirical studies are essential to understand complexities in small regional development.
(Bibri, 2021)	Sustainability goals, urbanization challenges, strategic planning, actionable insights	Conceptual framework, case studies, data mining	Data-driven technologies balance sustainability and urbanization challenges.
(Chen et al., 2017)	RFM analysis, transaction histories	Analytical modelling	RFM analysis categorizes customers effectively, aiding businesses in optimizing services.
(Bibri & Krogstie, 2017)	Responsiveness to economic conditions	Literature review	Integration of data analytics enables real-time adjustments in urban planning.

2.4. INTEGRATION OF PREDICTIVE ANALYTICS IN TOURISM MANAGEMENT

The integration of predictive analytics in managing tourism-based economies has become increasingly vital for understanding and forecasting tourism spending patterns. Various predictive models, including time series forecasting and machine learning algorithms, have been successfully employed to analyse tourism demand and spending behaviours. Time series models, such as ARIMA (AutoRegressive Integrated Moving Average), have traditionally been used to capture seasonal trends and cyclical patterns in tourism data (Gricar, 2023). However, the advent of machine learning techniques has introduced more sophisticated methods, such as deep learning and ensemble models, which can handle complex, non-linear relationships in tourism data more effectively (Essien & Chukwukelu, 2022). These models not only enhance the accuracy of forecasts but also allow for the incorporation of a wider array of variables, including economic indicators and consumer behaviour metrics.

The value of these models hinges on high-quality data input, and payment system data has emerged as a crucial resource for short-term economic forecasts in regions with fluctuating tourism activity. By analysing transaction data from credit and debit card payments, researchers can gain insights into consumer spending patterns, enabling timely adjustments to marketing strategies and resource allocation (Huang & Wang, 2022). This data can reveal peak spending times, average transaction values, and other critical indicators that inform local businesses and policymakers about current economic conditions. For instance, studies have shown that regions can leverage payment data to predict tourist spending spikes during holidays or special events, allowing businesses to prepare adequately (Comerio & Strozzi, 2018).

The effectiveness of specific indicators in predicting tourism-driven economic shifts has been a focal point of research. Key indicators such as peak spending times, average transaction values, and visitor demographics have proven to be particularly useful in forecasting economic impacts (Rahman, 2023). For example, understanding peak spending periods allows local businesses to optimize staffing and inventory levels, while average transaction values can inform pricing strategies and promotional efforts (Arshad et al., 2021). Moreover, the integration of these indicators into predictive models enhances their robustness, providing a clearer picture of potential economic outcomes based on varying tourism scenarios (Sakhuja et al., 2016).

To round off, the application of predictive analytics in tourism management is transforming how regions forecast and respond to tourism-related economic fluctuations. By utilizing advanced predictive models and leveraging payment system data, local economies can better anticipate consumer behaviour and optimize their strategies accordingly. The identification and analysis of key indicators further enhance the predictive capabilities, enabling more informed decision-making in tourism management. A summary of this chapter and the used literature can be found in Table 2.4.

Table 2.4 - Summary table about the “Integration of predictive analytics in tourism management” LR chapter

Authors	Main Indicators	Methodology	Key Findings
(Gricar, 2023)	Seasonal trends, cyclical patterns	Time series analysis	ARIMA models effectively capture seasonal trends in tourism data.
(Essien & Chukwukelu, 2022)	Non-linear relationships, predictive accuracy	Machine learning	Deep learning techniques outperform traditional forecasting methods in tourism demand prediction.
(Huang & Wang, 2022)	Consumer spending patterns, transaction data	Data analysis	Payment data offers real-time insights into consumer behaviour, aiding in timely marketing adjustments.
(Comerio & Strozzi, 2018)	Peak spending times, average transaction values	Case studies	Forecasting spending spikes during events helps optimize resource allocation.
(Rahman, 2023)	Visitor demographics, spending behaviours	Literature review	Identifies key indicators that enhance the robustness of economic forecasting models.
(Arshad et al., 2021)	Peak spending times, average transaction values	Case studies	Highlights how businesses can optimize staffing and inventory by understanding peak spending times.

2.5. ROLE OF MACHINE LEARNING IN FORECASTING ECONOMIC ACTIVITY

The role of machine learning algorithms, particularly Random Forest and XGBoost, in forecasting economic activity has become increasingly significant, especially in the context of consumer spending patterns and tourism-based economic forecasting. These algorithms are adept at processing large datasets and identifying complex patterns that traditional econometric models may miss. For instance, Lecun et al. discuss the transformative impact of deep learning techniques across various domains, including economic forecasting, where they outperform traditional methods in predictive accuracy (LeCun et al., 2015). Similarly, Choi and Shin emphasize the importance of dimension reduction and difference in enhancing the forecasting capabilities of machine learning models applied to employment levels, which can be analogous to consumer spending forecasts (Choi & Shin, 2019).

When comparing the forecasting accuracy of machine learning algorithms to traditional methods in tourism-based economic forecasting, studies have shown that machine learning models frequently yield superior results. For example, Hall discusses the application of machine learning techniques to optimize forecasting models for unemployment rates, demonstrating that these methods can provide more accurate predictions compared to conventional approaches (Hall, 2018). Furthermore, Mullainathan and Spiess provide insights into how machine learning can be effectively applied in econometric contexts, suggesting that these algorithms can handle high-dimensional data more efficiently than traditional models (Mullainathan & Spiess, 2017). This is particularly relevant in tourism forecasting, where factors influencing consumer behaviour are numerous and often interrelated.

However, the application of machine learning models in small regions with limited data availability poses several challenges. The effectiveness of these models is often contingent upon the quantity and quality of data available for training. For instance, Zhao discusses the implications of integrating econometrics with data science, highlighting the difficulties faced when applying machine learning techniques in contexts where data is sparse (Zhao, 2024). Additionally, the complexity of machine learning algorithms can lead to issues of overfitting, particularly in small datasets, as noted by Shen et al., who emphasize the need for robust validation techniques to ensure model reliability (Shen et al., 2021). This challenge is compounded by the necessity for interpretability in economic forecasting, as stakeholders often require clear explanations of model outputs to inform decision-making processes.

To wrap up, machine learning algorithms such as Random Forest and XGBoost are proving to be powerful tools for forecasting economic activity, particularly in consumer spending and tourism contexts. Their ability to process large datasets and identify intricate patterns offers a significant advantage over traditional forecasting methods. However, challenges related to data availability and model complexity must be addressed to fully leverage these technologies in small regional economies. A summary of this chapter and the used literature can be found in Table 2.5.

Table 2.5 - Summary table about the “Role of machine learning in forecasting economic activity” LR chapter

Authors	Main Indicators	Methodology	Key Findings
(LeCun et al., 2015)	Predictive accuracy, complex patterns	Literature review	Deep learning techniques outperform traditional methods in forecasting economic activities.
(Choi & Shin, 2019)	Dimension reduction, employment levels	Comparative analysis	Machine learning enhances forecasting accuracy by addressing complex data relationships.
(Hall, 2018)	Unemployment rates, predictive accuracy	Case studies	Machine learning models are more accurate than conventional methods in predicting economic shifts.
(Shen et al., 2021)	Overfitting, model interpretability	Theoretical analysis	Addresses the need for robust validation techniques to improve machine learning applications.
(Mullainathan & Spiess, 2017)	High-dimensional data, model efficiency	Literature review	Machine learning efficiently processes high-dimensional datasets for better economic forecasting outcomes.
(Zhao, 2024)	Data sparsity, model reliability	Theoretical analysis	Discusses challenges and solutions for applying machine learning in contexts with sparse data.

2.6. UNDERSTANDING INDICATORS FOR LOCAL IMPACT

In examining the economic indicators relevant to tourism and local economies, it is essential to distinguish between macroeconomic and microeconomic indicators. Most existing studies predominantly focus on macroeconomic indicators, such as gross domestic product (GDP) growth and employment rates, which provide a broad overview of economic health but often overlook the nuanced insights that microeconomic indicators can offer. Microeconomic indicators, such as consumer spending patterns, transaction values, and peak spending times, are crucial for understanding the specific behaviours and preferences of tourists, which can significantly influence local economic conditions.

The reliance on macroeconomic indicators can lead to a gap in understanding the immediate impacts of tourism on local economies. For instance, transaction data from payment systems can reveal consumer spending behaviours, allowing local authorities and businesses to make informed decisions regarding resource allocation and marketing strategies. This data-driven approach enables a more granular analysis of economic activity, highlighting the importance of integrating microeconomic indicators into predictive models. By doing so, local businesses can optimize staffing and inventory levels during peak seasons, thereby enhancing their operational efficiency and profitability.

Moreover, the integration of machine learning techniques into economic forecasting has shown promise in improving the accuracy of predictions related to tourism spending and economic impacts. Traditional econometric models often struggle to capture the complexities of consumer behaviour, while machine learning algorithms can analyse large datasets to identify patterns that may not be immediately apparent. This advancement underscores the potential for microeconomic indicators to inform more responsive and adaptive economic strategies, particularly in regions heavily reliant on seasonal tourism.

Despite the advancements in data analytics and machine learning, there remains a notable research gap concerning the application of microeconomic indicators in tourism studies. While macro indicators provide a broad understanding of economic trends, the lack of focus on microeconomic factors limits the depth of insights available to policymakers and business leaders. Future research should aim to bridge this gap by exploring how microeconomic indicators can be systematically integrated into tourism economic studies, thereby enhancing the robustness of economic forecasting models and improving local economic resilience.

Ultimately, while macroeconomic indicators have been extensively studied in the context of tourism, there is a pressing need for more research on microeconomic indicators. Such studies could provide valuable insights into consumer behaviour and local economic dynamics, ultimately leading to more effective tourism management strategies and economic planning. Below it is presented a table, Table 2.6, with a summary of the main KPI’s found during the literature review chapter.

Table 2.6 - Summary table about the main KPI's found during the LR chapter

KPI Name	Description	Study	Scope
Peak Spending Times	Identifies periods of highest consumer spending, aiding businesses in resource allocation.	(Adhuze, 2023)	Micro
Average Transaction Values	Measures the average amount spent per transaction, informing pricing strategies.	(Adhuze, 2023)	Micro
Visitor Demographics	Analyses characteristics of visitors to tailor marketing strategies.	(Li et al., 2022)	Micro
Seasonal Sales Trends	Examines fluctuations in sales during peak and off-peak seasons.	(Arshad et al., 2021)	Micro
Consumer Spending Patterns	Analyses transaction data to understand overall spending behaviours among tourists.	(Adhuze, 2023)	Micro
Employment Levels	Measures changes in employment rates because of tourism activities.	(Hall, 2018)	Macro
Infrastructure Investment	Evaluates the level of investment in tourism-related infrastructure.	(Adhuze, 2023)	Macro
Tourism Accessibility	Measures the ease of access to tourism attractions, impacting visitor numbers.	(Li et al., 2022)	Macro

Public-Private Partnership (PPP) Impact	Assesses the effectiveness of PPs in enhancing tourism infrastructure.	(Yunus et al., 2021)	Macro
COVID-19 Impact on Local Economy	Evaluates the economic impact of COVID-19 on tourism-related services.	(Rahman, 2023)	Macro
Dynamic Relationship Analysis	Examines the interrelationships between tourism, trade, and economic growth	(Suleiman & Albiman, 2014)	Macro
Economic Growth Rate	Assesses the overall growth of the economy concerning tourism development.	(Ridderstaat et al., 2013)	Macro

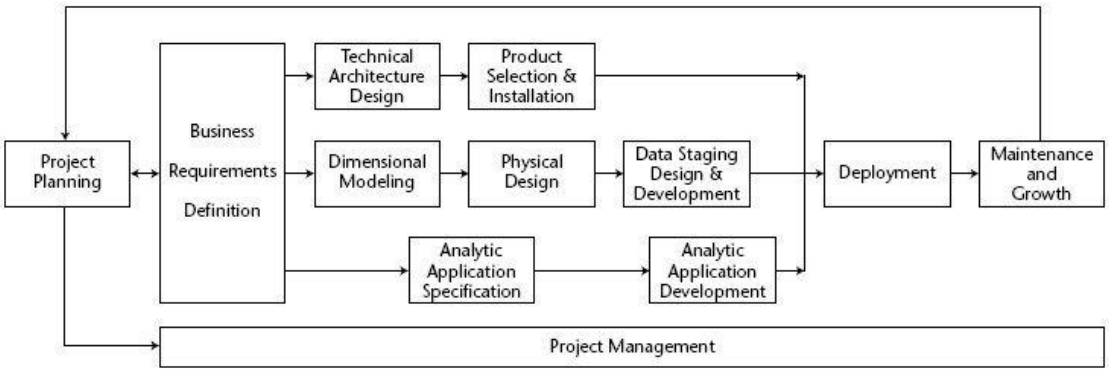
3. METHODOLOGY

This chapter outlines the methodology used to collect, process, and analyse monetary transaction data in the Oeste CIM region using the Kimball Lifecycle framework. It defines business requirements aimed at forecasting and monitoring spending, designs a Lakehouse and Data Warehouse architecture, and develops a star schema with key dimensions and a transaction fact table. ETL/ELT pipelines ensure data quality and structure. Predictive modelling, using XGBoost, is applied to forecast spending trends. The chapter concludes by discussing limitations and justifying methodological choices to support transparency and reproducibility.

3.1. KIMBALL LIFECYCLE

The Kimball Lifecycle (Figure 3.1) (Kimball et al., 2008), is a prominent framework for designing data warehouses and business intelligence systems. It is particularly recognized for its bottom-up approach, which emphasizes the creation of data marts tailored to specific business processes.

Figure 3.1. The Kimball Lifecycle Diagram (Becker, 2008)



The steps of the Kimball methodology include, business requirements definition, technical architecture, dimensional modelling, physical design, creating the ETL (Extract, Transform, Load) processes, and deploying the data mart (Fernald, 2012). This systematic approach ensures that the data warehouse is designed to meet specific analytical requirements, thereby enhancing the decision-making capabilities of the organization.

After the data mart is build, machine learning models will be leveraged to develop predictive model to provide insight into future trends based on historical data. This predictive capability will be implemented in a dashboard, allowing end-users to make data-driven decisions based on historical data.

A significant aspect of Kimball Lifecycle is its focus on dimensional modelling, which organizes data into facts and dimensions. This structure not only simplifies data retrieval but also

optimizes performance for analytical queries (Fernald, 2012; Irawan et al., 2021). The methodology has been successfully applied in various sectors, including finance, healthcare, and retail, demonstrating its versatility and effectiveness in real-world applications (Eckstein & Kollar, 2008; Flerchinger et al., 2009). For instance, organizations have utilized the Kimball methodology to create data warehouses that support complex reporting and analysis, enabling them to derive actionable insights from their data (Eckstein & Kollar, 2008; Irawan et al., 2021).

3.2. BUSINESS REQUIREMENTS

3.2.1. Context Understanding

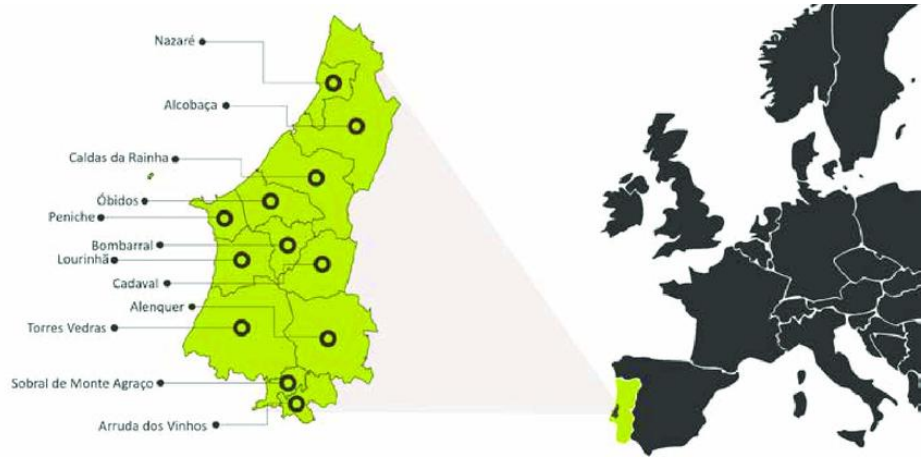
As of 2023, Oeste region has a population of 388,000 according to Instituto Nacional de Estatística, with a socioeconomic landscape that combines coastal and inland municipalities. The region's primary economic sectors include agribusiness (horticulture), light manufacturing, and heritage tourism (*Plataforma Online Da Região Oeste de Portugal*, 2024).

The development of a business intelligence solution for Oeste CIM is driven by three primary objectives rooted in the Kimball Lifecycle Methodology's business requirements phase.

The first will be to develop accurate forecasting to enable evidence—based policymaking. The solution must provide predictive capabilities for forecasting tourism spendings and economic activities over the next fiscal year. This requires using historical transactional data to generate forecasts.

The second objective will be to enable interpretable monitoring of the data, because stakeholders require dynamic visualization of transactional data to analyse spending patterns across sectors and municipalities or parishes. Granular insights into peak spending periods, average transactional values and touristic demographics are critical for responsive decision-making. On the image bellow, it's possible to understand the geographical location of CIM Oeste (Figure 3.2.)

Figure 3.2. CIM Oeste municipalities (Simões et al., 2023)



Finally, promote stakeholder empowerment, because municipalities and business need actionable insights to optimize resource allocation and develop targeted marketing campaign, considering the tourist origins.

Also, it makes sense to enable stakeholders to analyse sector performance, municipality performance, and make comparisons between different periods. The definition of performance will go through economic values as well as quantitative values. The visuals must enable us to answer the questions asked but also give additional information to enhance decision making.

With this and the literature review performed previously we can formulate the following business questions:

1. **How do transaction volumes (number of payments) vary across months, quarters, and years?** According to Ridderstaat, seasonal tourism creates economic volatility, so it's important for the municipalities to have insight to allocate their resources during peak and of-peak seasons.
2. **What are the key predictors of tourism spending behaviour, and how can they inform short-term economic forecasts?** Zhou and Essien have proven that predictive analytics are able to successfully forecast tourism demand, so businesses need actionable forecasts to optimize inventory and marketing to target the desired tourist demographic.
3. **Which sectors contribute most to economic resilience, and how can they be prioritized?** Investors and local governments need sectorial insights to allocate funds effectively.
4. **What is the cumulative tourism spending in Oeste CIM to date, and how does it compare to previous years?** Payment data aids short terms forecast, as stated by Huang and Wang, and this could also be helpful to assess economic recovery post COVID-19 pandemic.

5. **How does current tourism spending compare to the same period in previous years?** This is also a good way to study recovery and growth. And as Adhuze mentions in its study, year-over-year analysis is critical for small regions.
6. **How does spending differ between weekdays and weekends?** This analysis is particularly important for event planning, to understand the most lucrative days and even dates to perform certain events. And Law stated that temporal granularity promotes good urban planning.

3.2.2. Stakeholders

The main stakeholders are local municipalities, local businesses, tourism boards and general investors. But everybody that is involved in the middle can also benefit from the development of this project, for example residents in the Oeste region will be directly impacted, although they aren't the direct stakeholder.

3.2.3. Data Understanding

The analysis in this study relies on operational data compiled through a collaborative effort by SIBS Analytics and NOVA Analytics Lab under the Smart Region project (Jardim et al., 2025). This dataset encompasses 57,714,584 entries across multiple files (approximately 10GB in size), covering the period from January 2020 to October 2024. Each file contains 18 columns, including:

- Geographic information
- Time-related aspects
- Categorical elements
- Quantitative measures

SIBS Analytics functions as a centralized data repository, providing consolidated insights into spending habits across various payment methods, tailored to the operational and demographic features of geographic areas, organizations, and business sectors.

In this context, the Oeste region, a NUTS III territorial unit in the northern Lisbon, has emphasized data-driven governance to promote regional growth.

The dataset is derived from POS terminal and ATM transactions, aggregated weekly and segmented by weekdays and weekends. Each record includes:

- Total transaction value (monetary amount)
- Number of unique cardholders
- Transaction frequency
- Calculated averages (operations per card, value per transaction)

Geographic details are maintained at the parish level, with transaction origins determined from cardholder information. The analysis focuses on absolute totals (e.g., total spending) and normalized indicators (averages per card), as illustrated in Table 3.1.

Table 3.1. Original metadata

Column Name	Description
Semana ID	Week of operations, identified by the first day of the week following the example, "20220704"
Semana DESC	Week of operations, identified by the first day of the week following the example, "Sem 04 – Jul"
Tipo de Geografia ID	Municipality of the Oeste Region
Geografia ID	The ID for the municipality of the Oeste Region
Geografia DESC	The name of the municipality of the Oeste Region
Tipo de Setor ID	The ID for the type of sector classification based on the shopper
Setor ID	The ID for the sector classification based on the shopper
Setor DESC	Sector classification based on the shopper
Tipo de dia ID	The ID for the type of day on which the operations was accepted.
Tipo de dia DESC	The type of day on which the operation was accepted (weekdays or weekends).

Column Name	Description
Origem ID	ID of the origin of the card used for the transaction
Origem DESC	Origin of the card used for the transaction
Nº de Cartões	Corresponds to the number of payment cards used.
Nº de Operações	Corresponds to the number of Operations performed.
Valor das Operações	Total value of transactions carried out, in euros.
Nº médio de Operações por Cartão	Corresponds to the average number of operations performed by each payment card.
Valor médio por Cartão	Average value used per debit card, in euros.
Valor médio das Operações	Average value used per transaction, in euros.

3.3. TECHNICAL ARCHITECTURE DESIGN

The architecture is divided in 3 main layers, bronze layer, silver layer and gold layer.

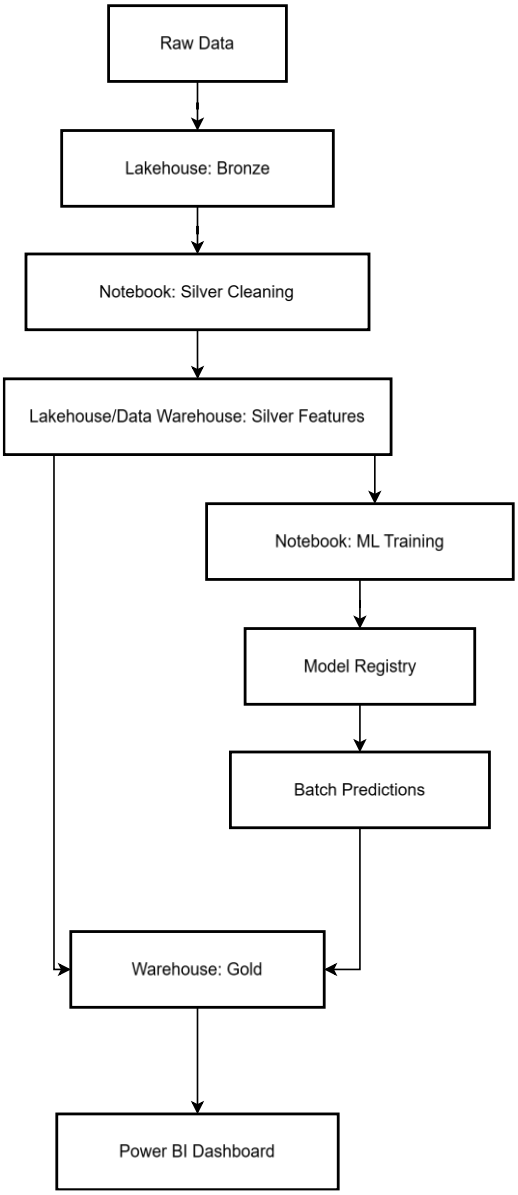
The bronze layer is where raw data is loaded into the Lakehouse as Delta tables. Small transformations are made in the import process. The files will be saved in a dedicated folder called “Files/Tabelas” in a table called “df_sa_transactions”, in a Lakehouse named “LH_SOURCES”

The silver layer is where the cleaning, validation, transformation and preprocessing and feature engineering for the forecasting model will take place. The final outputs will be stored to a table named “STG_FCT_Transactions” in the Data warehouse called “DW_MasterThesis_STG_2025”. Also, separately the dataset for model training will be saved separately in a table named “STG_FCT_Transactions_DS”.

The final layer is the gold, where the goal is to have analytics-ready data. Here a Data warehouse (“DW_MasterThesis_2025”) will be created and load the transformed tables into it. The dimensions and fact are then defined as well as the definition of relationships and optimization for querying. Also, in this stage the forecasting model will have been trained and validated, saving the output also in the Data Warehouse.

The architecture is concluded in the dashboarding stage where the visualizations and analysis are performed by connecting Power BI to the Warehouse, by building a semantic model. Merging with the model predictions ready to be analysed as well. The entire planed architecture can be found in the diagram bellow.

Figure 3.3. Architecture for the MastherThesis



3.4. DIMENSIONAL MODELLING

For the design of the data warehouse, following a defined methodology is essential to allow for a clear definition of the steps that should be taken. As such, the Kimball 4 step approach was chosen, composed of the following four steps:

1. Identify the business process – At this step, the focus is to identify and understand the business process that the dimensional model will represent.
2. Identify the grain – Here the level of detail if defined for the dimensional model, meaning, what a row in the fact table represents.
3. Identify the dimensions – The variables are grouped in different dimensions. They must be descriptive and consider the grain defined before.
4. Identify the facts – The last step, where the facts will be defined, containing measures (presented in numeric format) that are relevant to represent the granularity defined in the second step.

3.4.1. Business Process Definition

Oeste CIM is heavily reliant on tourism and faces some economic volatility due to seasonal tourism. Stakeholders lack insight into spending behaviours, affecting effective resource allocation, marketing strategies and infrastructure investments. The main KPI's to ensure the project is successful are monthly growth rates, mean absolute percentage error (MAPE), sector contribution %, cumulative growth rate, Weekend-to-Weekday ratio.

3.4.2. Granularity Definition

Each transaction will be defined as "a weekly transaction in a certain sector by a certain nationality in a certain parish". This means that each transaction will have an amount associated with it for each week, for each parish, for each sector and for each origin of the card that performed the payment.

From the data that was provided, there are three main hierarchies that we can identify concerning dates, sector and locations, which are directly related to the level of detail desired and were reflected in the definition of the grain.

Regarding temporal granularity, the smallest granularity available is weekly, that will enable us to capture meaningful trends and patterns in transaction, without overwhelming the end user.

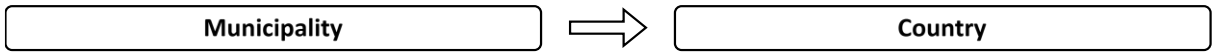
Figure 3.4. DIM_DATE granularity definition

Geographical granularity will be at parish-level, promoting a deeper understanding of the location for each transaction.

Figure 3.5. DIM_GEOGRAPHY granularity definition

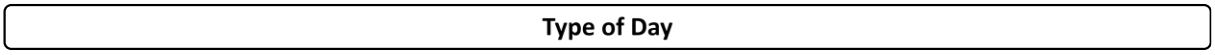
Origin granularity will be responsible for giving the customer origin and will be detailed, when possible, up until the municipality level.

Figure 3.6. DIM_ORIGIN granularity definition



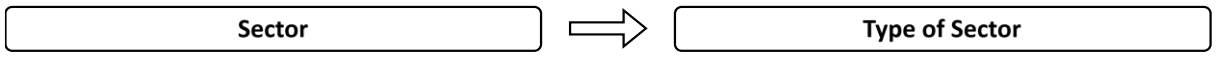
Type of day had to be defined as a separate hierarchy because this way we ensure the uniqueness of each data entry. It is split into type of day, where it separates weekends from weekdays.

Figure 3.7. DIM_TYPE_OF_DAY granularity definition



Finally, the granularity sector will promote a detailed analysis of the sector the transaction is being held, promoting an analysis of the individual sector responsible for the transaction.

Figure 3.8. DIM_SECTOR granularity definition



3.4.3. Dimensions and Facts

Following the formalization of the analytical grain—the precise level of detail required for this study—the implementation phase transitions to the systematic construction of dimension tables, adhering to Kimball's dimensional modelling paradigm. These tables serve as the structural backbone for contextualizing transactional data at the defined granularity, ensuring fidelity to both operational specificity and analytical scalability. Five dimensions were identified that were grouped into five tables described over the course of the chapter. Before each table a small description of the dimension will be presented, and the tables will display the variable name, as well as the datatype, and the surrogate key of the dimension, represented by "SK".

The DIM_Date, will be manually created, considering the importance of dates to further analyse transaction evolution over time. Therefore, this level of detail is deeply related to the business requirements proposed. In this case, even though “Semana ID” is already unique, a surrogate key was created called “Date_ID” as it’s a good practice to create key outside of the source system. Also, the type of day will be represented in this dimension. Find bellow the table summing the existing column and the respective datatype (Table 3.2)

Table 3.2. DIM_DATE structure

Variable	Data Type
SK_Date (PK)	INT
Year	INT
Semester	INT
Semester_Name	VARCHAR (10)
Semester_Abbreviation	VARCHAR (5)
Trimester	IN
Trimester_Name	VARCHAR (10)
Trimester_Abbreviation	VARCHAR (5)
Month	INT
Month_Name	VARCHAR (10)
Month_Abbreviation	VARCHAR (5)
Week	INT

Week_Name	VARCHAR (10)

The DIM_Geography contains geographic information that will support the Transactions fact table. The only change that will need to be performed is the addition of a new column where the parish, municipality, district and country will need to be concatenated and separated by commas, so that the map tool visual is able to identify correctly the location. Find bellow the table summing the existing column and the respective datatype (Table 3.3)

Table 3.3. DIM_GEOGRAPHY structure

Variable	**Data Type**
SK_Geography (PK)	INT
Country	VARCHAR (100)
Country_City_Parish	VARCHAR (100)
Municipality	VARCHAR (100)
Parish	VARCHAR (100)

The DIM_Origin represents the country of origin of the card used in the transaction, if not originated in Portugal, otherwise the municipality of origin. Find bellow the table summing the existing column and the respective datatype (Table 3.4)

Table 3.4. DIM_ORIGIN structure

Variable	**Data Type**
SK_Origin (PK)	INT
Country	VARCHAR (100)
Municipality	VARCHAR (100)

Country_City	VARCHAR (100)

The DIM_Sector, as the name suggests includes the sector of activity the transaction was performed in. This sector categorization is based on a code of economic activity (CAE). CAE is a code that enables grouping and classification of each professional activity and their respective tax obligations. Find bellow the table summing the existing column and the respective datatype (Table 3.5)

Table 3.5. DIM_SECTOR structure

Variable	**Data Type**
SK_Sector (PK)	INT
Type_of_Sector	VARCHAR (100)
Sector	VARCHAR (100)

The DIM_Day_Type is a unique dimension, that could be merged with DIM_Date, but due to the nature of the data, was decided to the left as a unique dimension. This dimension shows the type of day the transactions of performed on. Can be weekend, weekday, for example. Find bellow the table summing the existing column and the respective datatype (Table 3.6)

Table 3.6. DIM_DAY_TYPE structure

Variable	**Data Type**
SK_Day_Type (PK)	INT
Day_Type_Description	VARCHAR (100)

The final step of this chapter is to define the facts. For this project only one fact table is identified, in which will be related all the transactions made in the CIM OESTE for the period in analysis. The fact table will contain all surrogate keys from the dimensions but now acting a foreign key. Find bellow the table summing the existing column and the respective datatype (Table 3.7)

Table 3.7. FCT_TRANSACTIONS structure

Variable	Data Type
FK_Date	INT
FK_Origin	INT
FK_Geography	INT
FK_Sector	INT
FK_Day_Type	INT
Number_of_Cards	INT
Number_of_Transactions	INT
Transaction_Value	INT
Avg_Number_of_Transactions_per_Card	DECIMAL
Avg_Value_per_Card	DECIMAL
Avg_Transaction_Value	DECIMAL

3.4.4. Schema Desing

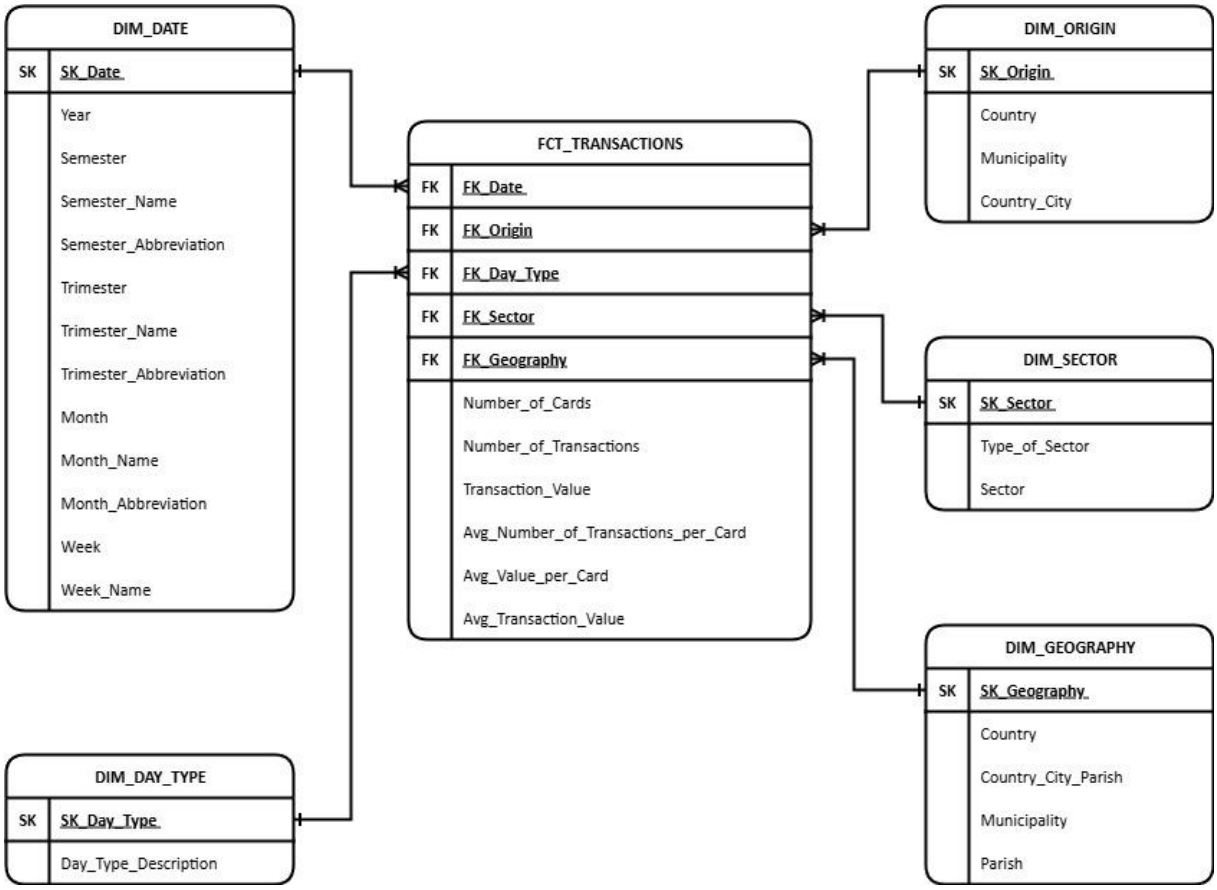
Following the rigorous specification of dimensions and fact tables within the dimensional model, the architectural synthesis culminates in a consolidated schema adhering to star topology. Anchored in Kimball's dimensional modelling principles, this framework positions a central fact table within a constellation of dimension tables, linked via foreign key relationships that enforce one-to-many cardinality (dimension-to-fact).

While the star schema's denormalized structure introduces controlled redundancy, it strategically optimizes query performance, reduces interpretive complexity, and enhances scalability relative to alternatives such as the snowflake schema. The latter's normalized

hierarchies, though storage-efficient, incur computational overhead during joins—a trade-off incompatible with the analytical latency requirements of this initiative. By prioritizing accessibility and extensibility, the star schema not only accommodates current analytical demands but also anticipates future granularity expansions without structural destabilization.

The finalized schema, illustrated below, operationalizes this design philosophy, balancing analytical agility with semantic coherence.

Figure 3.9. Oeste CIM Dimentional Model



3.5. ETL PROCESS

3.5.1. Extract

The first step into developing the ETL process was to create an environment dedicated to it, with the name of “MasterThesis Alexandre Spagnol”.

Next, a Lakehouse (“LH_SOURCES”) was created to serve as a staging area for our raw and untransformed data. At this level, the goal was to load the data into our environment and do little to no transformations. The only changes made to the data, were to merge all the xlsx and txt files into one Pyspark table. Also, some rows were removed, that were used as totalizing rows. All these transformations were made as the import was performed. It´s

important to note that due to the volume of data, a full load would not be optimal, so an incremental load is performed, by checking the date of the last record inserted into the table "df_sa_transactions" and compare to the last record trying to be inserted. If the value being inserted is bigger than the value already saved, the process will continue, otherwise, the process will stop.

3.5.2. Transform

This stage is the most extensive and one of the most important for this project. The first step was to prepare where our transformed data would land. For this a new Lakehouse was developed and a new Data warehouse was create named, "LH_MasterThesis_STG_2025" and "DW_MasterThesis_STG_2025", accordingly. Additionally, for the Date warehouse, the tables were created using SQL scripts.

The following steps were to develop the notebooks to create and perform de needed transformations for each dimension and for the fact table.

For the Date dimension, due it's special nature, it was created independently from the source data. In this table the script creates values starting in 2021 until the current day, being updated every time the notebook is executed, ensuring the dimensions follows the growing data. Also, the script was made to consider the main Portuguese holidays, signalling if the week contained a special day.

Regarding the Day Type dimension, it was relatively simple, only selecting the existing values on the source data, and adding a unique ID. In this case, there were only two types of possible days, either weekdays or weekends.

Now for the Geography dimension, the process was relatively more complex, due to the way the data was organized. Some cases where we had a value for parish, it was impossible to link it to a municipality. To resolve that, a GitHub file was found, listing all municipalities in Portugal and each parish. With this, the data was enriched and corrected. Also, in the cases that the value was a municipality, the parish is filled with the same value. Also, the country column was added to add flexibility for future works, and to create a composite column, concatenating the three columns into one, separated by commas. This column will ensure that the map present on a dashboard, can identify correctly the location.

For the Origin dimension, a similar process was required. In this case, the there was a problem, to identify the difference between countries and municipalities. A GitHub file was imported containing all the countries in the world. So, the script will first check if the values on the "Origem_ID" are present on the list or not. If the values are on the list, then it's a country, otherwise it's a municipality, and will fill the value with "Portugal". The next step was to populate the "Municipality" column, by using a similar rule, but the contrary.

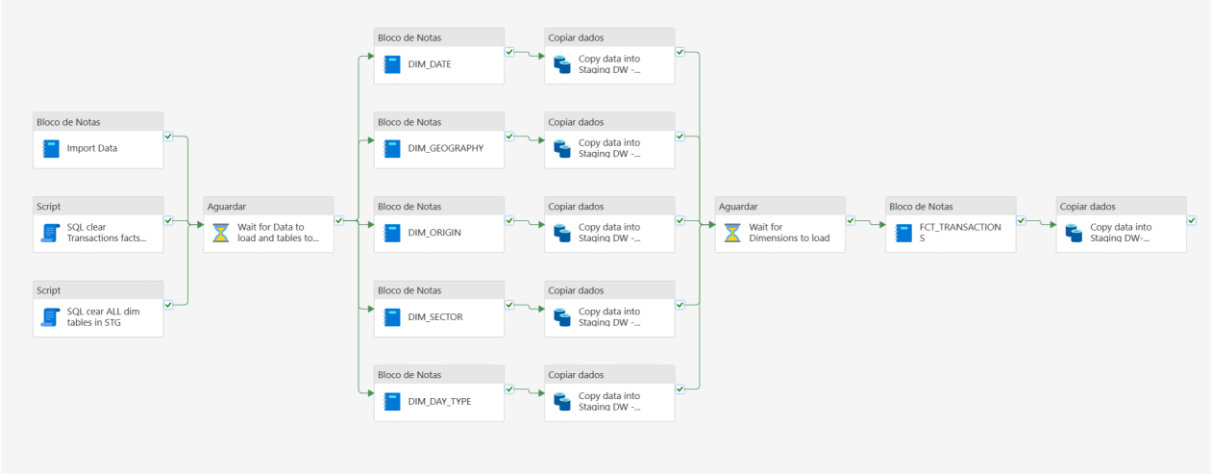
The Sector dimension uses the values strait from the source, and simply changing the name, because the totalizing rows were removed on the data import.

Finally, the notebook to create the fact table, joins the dimensions with the values, and an additional column was created by hashing all the foreign keys into one value. A duplicate check is performed, and the duplicates are removed before saving the table into the Lakehouse.

The next steps into the development of this transformation stage were to create the pipelines to process the data automatically. Each pipeline will be responsible for a group of similar tasks, first run the data source notebook, then run the dimension notebooks and copy the data into the staging Data warehouse. Another pipeline is responsible for validating the data before going into the final pipeline. The last pipeline is where the data is loaded into the Data warehouse.

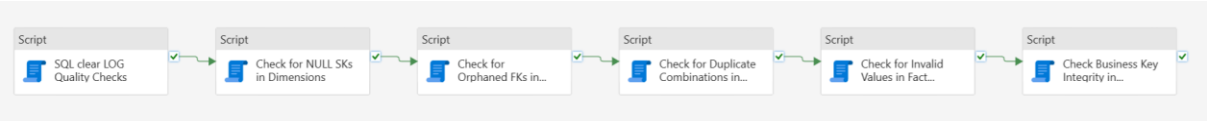
The "PL_MasterThesis_LOAD_STG" is responsible for importing the data and clearing the tables on the staging warehouse. Then each dimension is processed and saved into the data warehouse. After this, the FCT_Transations is created and saved into the staging warehouse. The structure can be verified in the image bellow (Figure 3.9.).

Figure 3.10. PL_MasterThesis_LOAD_STG pipeline



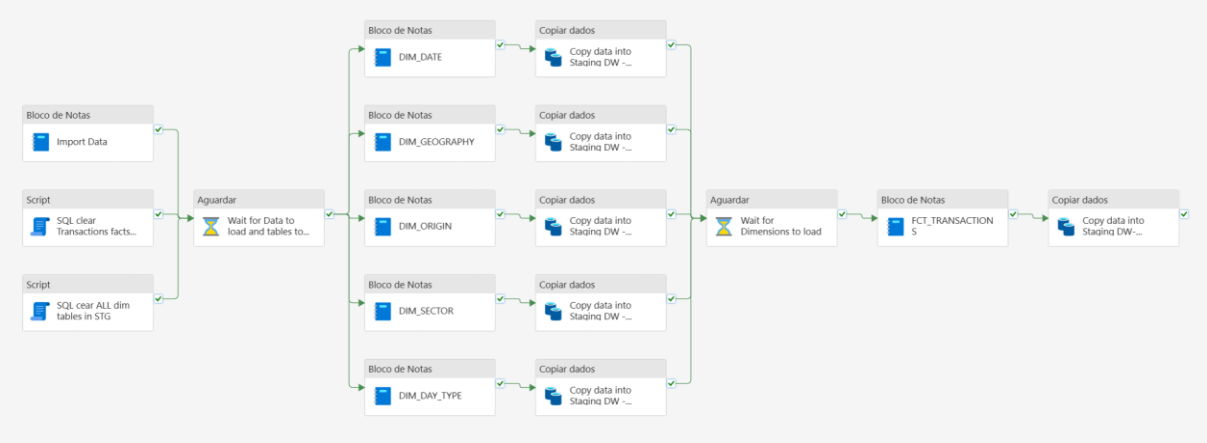
The following pipeline is the "PL_MasterThesis_VALIDATE_STG" (Figure 3.10.), where the data is validated. A set of rules was defined, and the data must respect them. A total of five rules needs to be respected. The first is that the SK's cannot contains any missing values, then the second check if the combination of FKs on the fact table is unique. The third rule is to check if the combination of the values on each dimension is unique. The fourth rule is to check the validity of the values on the fact table. Final rule will check the validity of the business keys that were given by the business.

Figure 3.11. PL_MasterThesis_VALIDATE_STG pipeline



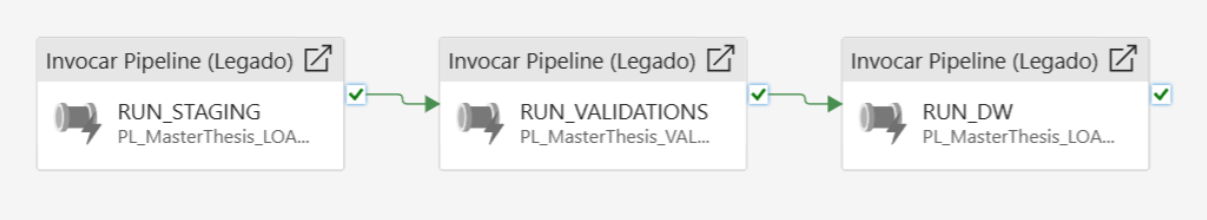
The final pipeline will insert the validated data into the final Date warehouse, named "DW_MasterThesis_2025" (Figure 3.11.).

Figure 3.12. PL_MasterThesis_LOAD_STG pipeline



There is also a pipeline named "PL_RUN_BI_NO_DATA_SCIENCE" (Figure 3.12.) to run the entire model ignoring the forecasting model. This pipeline will run the presented pipelines above in order and ensuring that the load stage is only performed if the data was successfully validated. Additionally, another master pipeline was created if the user decides to also run the data science part of this project. In this pipeline, all the steps above are performed, but an extra pipeline is called to run the forecasting model. The goal of this pipeline is to update the machine learning model with new or improved data.

Figure 3.13. PL_RUN_BI_NO_DATA_SCIENCE pipeline



3.5.3. Load

This stage is partly merged with the previous chapter, regarding the data load into the final Data warehouse. The data generated by the forecasting model will also be merged into this final Data warehouse. It's worth mentioning that no details were given about the forecasting model development, because the next chapter will be only dedicated to that subject.

The next step was to develop the semantic model, that will directly feed the dashboard and provide the end user with the insights to promote decision making. In this semantic model, both the data and the forecasting will be represented. The forecasting data will be in a separate folder called “Data Science”.

Final step is to develop the semantic model, that will feed information into our dashboard. This the final layer, and is where the measures will be created, enabling a enrichment of the available information. Here we will also change some column names, to be more business friendly and enforce the correct formatting and data types.

The first step was to hide all the SK and FK columns in both dimension and fact tables. Then the relationships were mapped so we can relate all our data. Bellow you can find the relationships listed in the table 3.8.

Table 3.8. Available Relationships

FCT_TRANSACTIONS		
FCT_Transactions (FK_Date)	Many-to-one	DIM_Date (SK_Date)
FCT_Transactions (FK_Day_Type)	Many-to-one	DIM_Day_Type (SK_Day_Type)
FCT_Transactions (FK_Geography)	Many-to-one	DIM_Geography (SK_Geography)
FCT_Transactions (FK_Origin)	Many-to-one	DIM_Origin (SK_Origin)
FCT_Transactions (FK_Sector)	Many-to-one	DIM_Sector (SK_Sector)

After the relationships were created, the following step was to create the measures that were listed during the literature review and the business understating chapter. Also worth noting that additional measures were created to future proof and prepare the data for possible additional needs the target audience can have. All created measures can be found stored in a folder named “01.FCT_Measures” in the semantic model. All these measures can be found bellow on Annex 01.

Regarding datatypes, the date column was defined as such, promoting the creation of time-variant measures. For the dimensions, so major transformations occurred, mainly just the removal of underscore character to improve column readability and also made sure that the columns related to dates, such as Month Name, Semester Name and Trimester Name had the

correct ordering logic. For the numerical column, the correct formatting was forced, as well as the monetary columns to present the correct currency. ´

Bellow, on table 3.9, we have listed all the measures create at the semantic model level, and the respective meaning for each one.

Table 3.9. Measure Explanation

FCT_TRANSACTIONS	
CY Number of Cards	This measure calculates the number of cards in the current year. It's used in several visualizations to compare current year vs previous year.
CY Number of Transactions	This measure calculates the number of transactions in the current year. It's used in several visualizations to compare current year vs previous year.
CY Transaction Value	This measure calculates the value of the transactions in the current year. It's used in several visualizations to compare current year vs previous year.
PY Number of Cards	This measure calculates the number of cards in the previous year. It's used in several visualizations to compare current year vs previous year.
PY Number of Transactions	This measure calculates the number of transactions in the previous year. It's used in several visualizations to compare current year vs previous year.
PY Transaction Value	This measure calculates the value of the transactions in the previous year. It's used

	in several visualizations to compare current year vs previous year.
Total Number of Cards	This measure calculates the total sum of the quantity of cards during the period of the analysis. This measure was needed to create current year and previous year measures, as well as some visuals.
Total Number of Transactions	This measure calculates the total sum of the quantity of transactions during the period of the analysis. This measure was needed to create current year and previous year measures, as well as some visuals.
Total Transaction Value	This measure calculates the total sum of the amount of transaction value during the period of the analysis. This measure was needed to create current year and previous year measures, as well as some visuals.
YoY Variance Number of Cards	This measure calculates the difference between the current year and the previous year, in this case for number of cards.
YoY Variance Number of Transactions	This measure calculates the difference between the current year and the previous year, in this case for number of transactions.
YoY Variance Transaction Value	This measure calculates the difference between the current year and the previous year, in this case for transaction value.
YoY% Number of Cards	This measure calculates the difference between the current year and the previous year, in this case for number of cards. For

	this measure the result is returned in percentage.
YoY% Number of Transactions	This measure calculates the difference between the current year and the previous year, in this case for number of transactions. For this measure the result is returned in percentage.
YoY% Transaction Value	This measure calculates the difference between the current year and the previous year, in this case for transaction value. For this measure the result is returned in percentage.
YTD Number of Cards	This measure calculates the year-to-date number of cards, enabling us to see how the quantity of cards used has evolved since the start of the year.
YTD Number of Transactions	This measure calculates the year-to-date number of transactions, enabling us to see how the quantity of transactions has evolved since the start of the year.
YTD Transaction Value	This measure calculates the year-to-date value of transactions, enabling us to see how the amount of transaction value has evolved since the start of the year.

With these measures, we were able deepen our ability to convey the most relevant information in the most adequate way for the business user to gain practical insights from the data. Over in the results chapter, we will explain our rationale behind the creation of each dashboard (page) of the report, as well as some of the most valuable conclusions they allowed us to take. Then, we will aggregate what we were able to observe from our visuals and answer the business questions earlier presented. Our ability to do this in a direct, complete, and accurate manner will ultimately be the determinant of the success of our project.

3.5.4. Forecasting

This chapter will be used to explain the steps used to develop the forecasting model, as well as the metrics chosen to evaluate the model performance. It is very important to note that the focus of this project was not the data science developments, but to showcase the possibilities and the power of it. With this, an overall view will be showcased.

It was decided to use the data from the silver layer, to feed the machine learning pipeline. The table used was the FCT_Transactions saved on the "LH_MasterThesis_STG_2025" Lakehouse.

Before any modelling step was performed, the data had to be analysed and transformed to ensure model quality. The initial SQL extraction filters record post-January 2021, resulting in 2 660 969 observations across 77 distinct dates.

For the next step of the project, a framework called mlforecast was used. This framework can be used to perform time series forecasting, enabling fast feature engineering inside the pipeline, model tunning, all this in a few lines of code, leading to computational efficiency as well.

To use this framework a unique key need to be created, so we can identify each possible combination, so a concatenation between FK_Origin, FK_Geography, FK_Sector, FK_Day_Type was made resulting in a new column named "unique_id". Also "DATE" column had to be renamed to "ds" for the framework to identify it as the datetime column. Then three separate forecasting datasets were created, for each of the three targets (Number_of_Cards, Number_of_Transactions and Transaction_Value). The final transformation at this stage was to rename each target column in each dataset to be called "y".

The modelling stage and feature engineering is performed at the same stage, being one of the advantages of the chosen framework. The function will date the target dataset and the number of weeks the user wants to predict. After this, the data is split into training and testing, with a ratio of 80/20, then the outliers are removed using the interquartile range (IQR) method. Lower limits at Q1 - 3×IQR and upper limits at Q3 + 0.5×IQR. This asymmetric capping strategy preserves distributional integrity while minimizing distortion from extreme values. This treatment resulted in a reduction of 12%-19% of observations across features.

Finally, the function will start a study for model hyper tuning, using another framework called Optuna (Akiba et al., 2019). For this project only gradient-boosting models were decided:

- LightGBM (Microsoft): Histogram-based algorithm with leaf-wise growth (Welcome to LightGBM's Documentation! — LightGBM 4.6.0.99 Documentation, 2025).
- XGBoost (Distributed ML Community): Depth-wise tree growth with regularization (XGBoost Documentation — Xgboost 3.1.0-Dev Documentation, 2025).

- CatBoost (Yandex): Ordered boosting with categorical handling (*CatBoost*, 2025)
- HistGradientBoosting (scikit-learn): Efficient binning for large datasets (HistGradientBoostingRegressor - Sklearn, 2025).

These models are integrated within the MLForecast framework, which automates feature generation including, temporal lags, expanding window statistics, rolling window statistics and date-based features.

Model tunning as said before uses Optuna, with mean percentage error (MAPE) as the objective function. The table below displays the parameters being optimized for each of the models. (Table 3.10.)

Table 3.10. Hyperparameter Optimization

Hyperparameter Optimization	
LightGBM	n_estimators, learning_rate, max_depth, num_leaves, min_child_samples, subsample, colsample_bytree
XGBoost	n_estimators, learning_rate, max_depth, subsample, colsample_bytree
CatBoost	Iterations, learning_rate, depth
HistGradientBoosting	max_iter, learning_rate, max_depth

Each trial evaluates 50 parameter configurations, with parallelization accelerating convergence to optimal settings. The optimization process explicitly constrains computational complexity by capping tree depth and iteration counts while maintaining model expressiveness. The model performance is assessed through two complementary metrics, being MAPE and root mean squared error (RMSE).

MAPE measures the average percentage difference between the forecasted and the actual values, with a lower MAPE indicating better forecast accuracy. (MAPE - Mean Absolute Percentage Error — Permetrics 2.0.0 Documentation, 2025).

$$\mathrm{MAPE}(y, \hat{y}) = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$$

RMSE is calculated as the square root of the average of the square differences between the predicted value and the actual values. Best score is 0, so the objective is to always minimize the score. (*RMSE - Root Mean Square Error — Permetrics 2.0.0 Documentation*, 2025).

$$\text{RMSE}(y, \hat{y}) = \sqrt{\frac{\sum_{i=0}^{N-1}(y_i - \hat{y}_i)^2}{N}}$$

These metrics are computed on the holdout test set representing the most recent 20% of temporal observations. The evaluation strategy ensures performance reflects true forecasting capability on unseen future data.

After the model is tuned and evaluated, all the data is fitted with the optimized models, and the prediction is performed. This process is repeated for the three target datasets, and in the end all the datasets are merged and ready to be consumed in the analytical layer. All the models are saved into a final data frame, this will enable the project owners to, in the future, decided the best model to showcase based on the scoring. Below is listed the scoring for each model, for each target (Table 4.1.). The best performing model at this time is Catboost, so we will have that data displayed in the final dashboard.

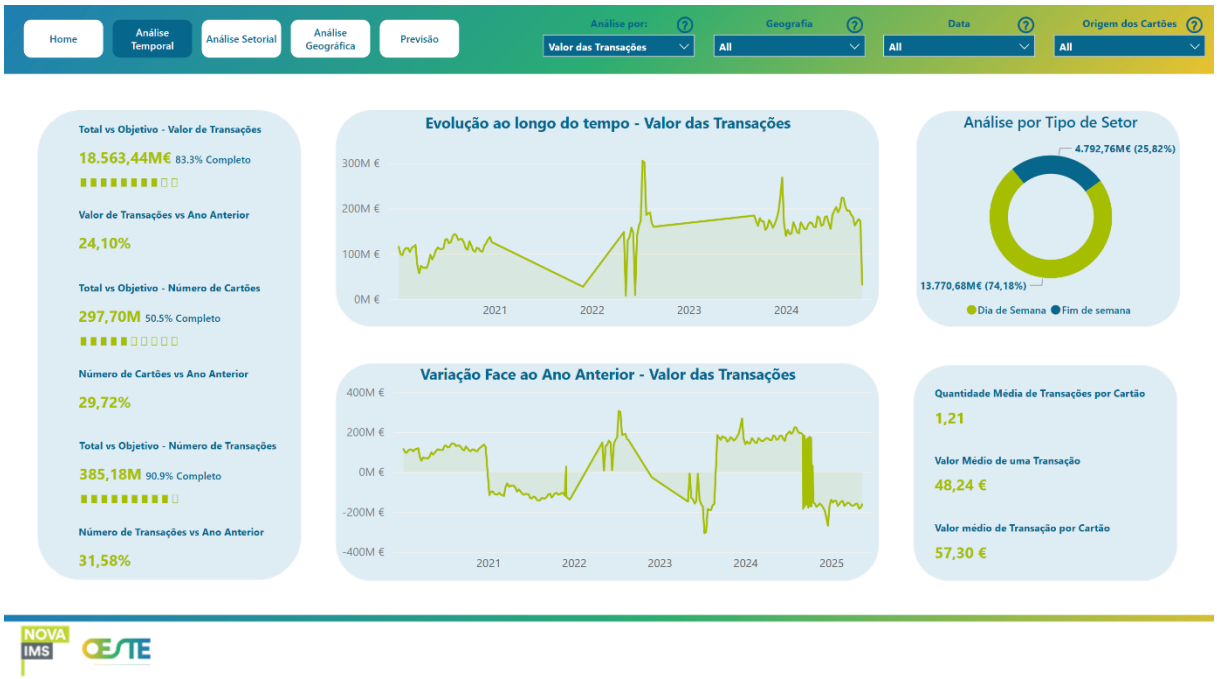
4. RESULTS AND DISCUSSION

After the methodology is concluded, the following step was to construct visualizations using PowerBI so we can present our data and all the preparation until now. During this chapter each page of the dashboard will be analysed and linked to the business questions. By the end the goal is to be able to take valuable insights from the data and promote smart and informed decision making. The dashboard is divided into 5 pages, being the first a home page and the face of the dashboard figure 4.1. The following four pages are destined to giving a specific type of analysis to our data, time analysis, sector analysis, geographical analysis and finally forecasting.

Figure 4.1. OESTE CIM Dashboard Home Page

4.1. TIME ANALYSIS

Figure 4.2. OESTE CIM Dashboard Time Analysis Page



The first page as the name indicates, is used for the time analysis. Here the data is analysed over-time.

Starting from the top, the header of this dashboard displays the buttons that allow the user to be immediately directed to the remaining pages of the report. This was done with the objective of simplifying the navigation throughout the document and was done for every page of the report (expect for the home page, already displayed earlier).

The header also displays the filtering options available throughout the entire dashboard. The first filter option enables the user to decide the type of data he wants to analyse, between “Transaction Value”, “Number of Transactions” and “Number of Cards”. Where applicable, this will change the visualizations according to the intended measure, respecting the necessities of the user. A small information tooltip is present next to the buttons to shortly explain how to interact with them.

Then the user also has the capability to filter the dashboard by geographical data, by card origin and by date. Again, a small information tooltip is present next to the buttons to shortly explain how to interact with them.

Now diving deeper into the actual visualizations, it starts by showing 6 Key Performance Indicators (KPI’s) on the left-hand side. Three of them are showing the total, and how it compares to a defined target, for each measure being analysed. It’s also important to note that this target is currently fictional, and it need to be defined by the decision makers and

introduced in the future using a static table. Currently this measure is used and a proof of concept. An information icon was added to let the user know what this target is for each case.

Then bellow the main number of the KPI we can find a progress bar and a percentage to make more evident how far off the measure is to the target. By analysing this visual the user can immediately say that in 2024, CIM OESTE achieved overall 186,13 million transactions resulting in 7,14 billion euros, representing 83,3% of the annual target.

The 3 remaining KPI's are intended to show the comparison between the current year values and the previous year ones. If no year is selected by the user, then the comparison will be between the most and least recent moments of data. An information icon was also placed next to the visual's title to explain the user how to interpret it. From here, we can see that, between 2024 and the last year, the number of cards used increased 139,86%.

To the right of the KPI section, two lines charts are presented. The upper one, reflects how the measure behaved over time. This visual was chosen as it is the most adequate to reflect the evolution of a measure in a time perspective. In addition, we added a light horizontal gridline to the graphic, so that the user could more easily compare sales against the reference points in the y-axis. From this chart, we can directly see that for example sales have constantly increase inter-years (increasing more as time moves forward) but are relatively steady within each calendar year, ignoring the gaps in the data. This visual also has the ability drilldown to display the data into more detail, by semester, trimester and monthly

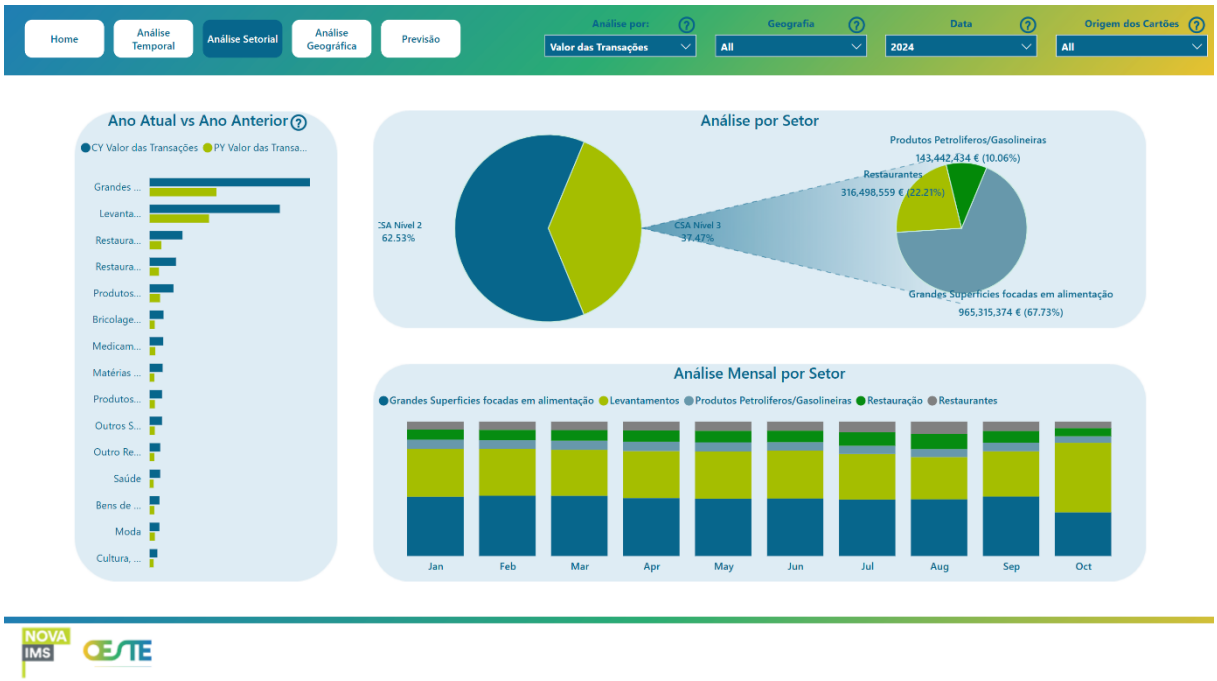
The bottom visual represents the year-over-year analysis, showcasing the difference from the same period last year. We can see that towards the end of 2024, we have less transactions than on the previous year. This could be problematic and needs to be investigated. This visual also has the ability drilldown to display the data into more detail, by semester, trimester and monthly

To the right of, we have on top, a donut chart to showcase the percentage of transactional value, in this case, by the type of day, being the option either weekday or weekend. After changing the type of analysis, we can safely say that the distribution doesn't change very much, measing that most of the transactions occur on weekends.

Just below we can find three important average values, the average number of transactions per card, the average transactions value and the average transaction value per card. This are values interesting for the stakeholders to understand how much, in average, each transaction brings to the region and how much one person spends usually. For instance, we can say that most people only make one transaction, but that transaction is relatively high, of about 65€. Also, we know that people spend on average just over 52€ in total, on a single transaction.

4.2. SECTOR ANALYSIS

Figure 4.3. OESTE CIM Dashboard Sector Analysis Page



On the Sector Analysis page, we start by showcasing a clustered bar chart that showcases the top fifteen best performing sectors for a selected measure. To complement this, we also display the previous year value, so that the user is able to verify the sector performance compared to the previous period. An information tooltip was placed next to the title to indicate how to view and comprehend the information in this visualization.

In terms of practical conclusions, this visual makes evident that the most popular sector is “Grandes superficies focadas em alimentação”. Followed by “Levantamentos”, and that in general terms, all sectors show an increase in the number of transactional values.

Next, we build a visual called pie of pie, that as the name suggests, constitutes a pie chart, where the detail of a slice of pie, leads to another pie. Here the user can analyse the selected measure, and see the comparison between the sector levels, and then inside each sector level, the 3 best performing sub-sectors. In this case, we can see that on the level 3 sector, responsible for 37,47% of the transaction value of 2024, “Produtos Petroliferos/Gasolineiras” represents 10,06% of the total value, with over 143 million €.

Lastly, we wanted to assess how the structure of the selected measure changed across the twelve months of the year, in terms of sector. For that purpose, a stacked chart was decided. And by placing the months of the selected year in the x-axis and the top five sectors, we can see how the 5 best performing sectors behave during the year. From this visual we can see that “Grandes superficies focadas em alimentação” and “Levantamentos” are fighting closely

with each other as the biggest origin of transactional value. And that in the end of the year a major decrease in the number of transactions for “Grandes superficies focadas em alimentação”.

4.3. GEO ANALYSIS

Figure 4.4. OESTE CIM Dashboard Geo Analysis Page

Moving to the Geo Analysis page, the first step was to use the funnel visual so we could showcase the distribution of the selected measure, by all the countries of origin outside of Portugal. By analysing it, we can clearly see that the main origin are Germany and Spain, responsible for the total transactional value in the Oeste region, in 2024.

In the centre of the page, the top visual is a table, where the user is able to analyse in more detail the transaction value, for example, at a parish level. This table shows the lowest geographical granularity of our data.

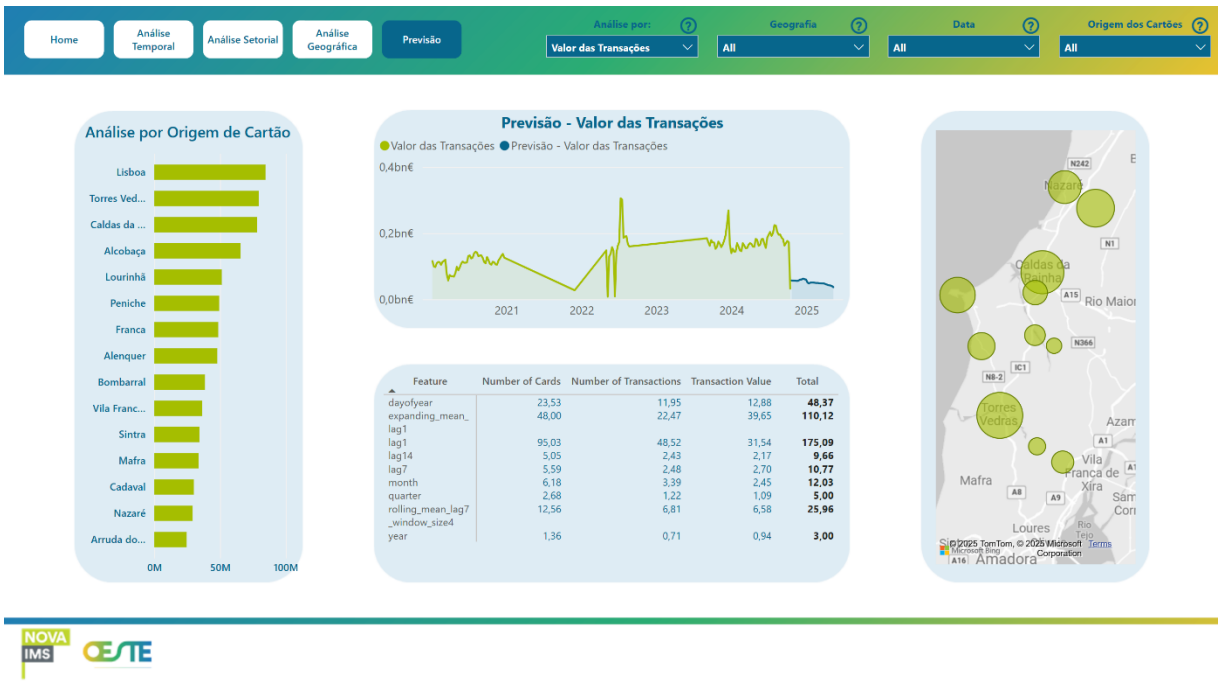
Then, underneath, we can find a bar chart where the goal was to understand, in Portugal, where are the main spender origins, in this case the 15 best spenders. This is valuable information to understand where the main spenders come from, and who to target and where to apply possible marketing strategies. In this case, we can say that mainly, spenders come from the within the Oeste region.

Finally, we used a map to show the distribution of transactions throughout the Oeste region, with the bubble size being defined in accordance with the volume of the selected analysis, for a specific location. In addition, a tooltip was added to provide the user with additional information. By hovering each bubble, the total amount of the selected analysis is displayed,

as well as the average amount of transactions each card makes, the average amount each card spends, and the average amount each card spends per transaction. With this graphic we enable the user to see that the two main municipalities responsible for the total transactions are Torres Vedras and Caldas da Rainha, and that on average people spend over 50€ in just one transaction.

4.4. FORECASTING

Figure 4.5. OESTE CIM Dashboard Forecasting Page



The forecasting page is more technical than the other pages, but is extremely useful for decision making, for future policies. Strating from the left, we have a bar chart visual responsible for showcasing the origin of the cards that are going to be used in the future. We can clearly see that Lisboa will be the main origin of people spending in Oeste CIM, whereas now is Torres Vedras.

In the middle, we can see a line plot, with the evolution over time, similar to the first page of the dashboard, but where, also the forecast, and it's clear that the forecast indicates a decrease in the number of transactions, but also a steady value, showing consistency. Underneath we have a technical table with the most important features the model uses for predicting future values, where we can analyse that the previous week is the most important feature to predict the next one, and for example, the day of the year we are currently predicting is also important, showing that previous year information and values, will tend to be similar in the future.

Finaly on the right, the have another map where we showcase the main places where transactions take place. Here we see that Torres Vedras continues to be the main municipality, but also it is expected that Nazaré and Peniche will display an increase in the number of transactions.

The backbone of this page are the results obtained in the modelling stage of this project. All tested models can be found bellow on Table 4.1. where it's possible to see that the best performing model was Cat Boost.

Table 4.1. Model Evaluation Score

Model Evaluation	**Nº of Cards**	**Nº of Transactions**	**Transaction Value**
LightGBM	MAPE: 0,3031 RMSE: 7,4862	MAPE: 0,2866 RMSE: 10,2969	MAPE: 0.4821 RMSE: 491,4628
XGBoost	MAPE: 0,3193 RMSE: 8,0304	MAPE: 0,3068 RMSE: 10,7776	MAPE: 0,5531 RMSE: 510,7726
Cat Boost	MAPE: 0,2944 RMSE: 7,6374	MAPE: 0,2844 RMSE: 10,5134	MAPE: 0,4691 RMSE: 509,0744
HistGB	MAPE: 0,3451 RMSE: 7,5108	MAPE: 0,2884 RMSE: 10,4160	MAPE: 0,5059 RMSE: 492,1726

4.5. ANSWERING THE BUSINESS QUESTIONS

After building the dashboard, we are finally able to answer the business questions that resulted from the project. We can summarize them in the following table (Table 4.1).

Table 4.2. Business Question Answer Summary

Business Question	Answer	Visual
How do transaction volumes (number of payments) vary across months, quarters, and years?	While a summary answer cannot be provided to the nature of the question (we would have to explicit the sales for each month, quarter, and year of data), the report allows the user to analyse this information. Overall, we can say that for all main metrics being analysed, the values decreased from 2021 to 2022, but this could be related to missing data, and that from 2022 the values have increased every year.	Time analysis page, centre top line plot
What are the key predictors of tourism spending behaviour, and how can they inform short-term economic forecasts?	They main indicators for spending behaviours are related to the day of the year, so it's safe to say that the current situation will be replicated next year in a similar way. Also, it's expected that a significant most of the future transactions will come from people from Lisbon. It would be interesting to prepare the region for this public and maybe also understand why	Forecasting page, centre bottom table

	people from Oeste CIM, decreased the number of transactions.	
Which sectors contribute most to economic resilience, and how can they be prioritized?	The main sectors are grocery stores, cash withdrawal and restaurants. We can prioritize these sectors by providing subsidies or tax incentives for small and mid-sized stores to maintain employment and access in underserved areas, coordinate with financial institutions to ensure liquidity during peak demand periods and provide targeted financial support during downturns for hospitality related business.	Entire sector analysis page
What is the cumulative tourism spending in Oeste CIM to date, and how does it compare to previous years	In this case, to simplify the analysis, we will assume only transactions with cards from outside the country. The value is 346,13 million euros, with an increase of 154,31% from 2023 to 2024.	Time analysis page, left side card visuals
How does current tourism spending compare to the same period in previous years?	In 2024, tourism spending has increased 154,31%, as well as the number of transactions with an increase of 185,89% and the number of cards used, with an increase of over 103%.	Time analysis page, left side card visuals and centre bottom line plot
How does spending differ between weekdays and weekends?	Weekends are clearly the most important part of the week, for generation of transactional movement.	Time Analysis page, right side top donut chart

Weekends are responsible
for more than 70% of all the
spending until now, resulting
in a total of 255,49 million
euros generated from 2020
until 2024, end of year.

From the report the stakeholders can increase the quality and confidence of its decision-making process and implement strategies that aim to address the opportunities and obstacles that the dashboard enables to visualize. To give some examples, special marketing strategies can be employed to each main location, aiming to maintain sales on the current best sectors, and try to find ways to increase the value to under-performing sectors. Also, by having evidence that most transactions are performed during weekends, it could be worth promoting events during the end of the week, and utilize all the available weekends during the year

4.6. DEPLOYMENT

In this chapter the deployment of the project will be explained, as well as the reasons for the deployment to be conducted this way.

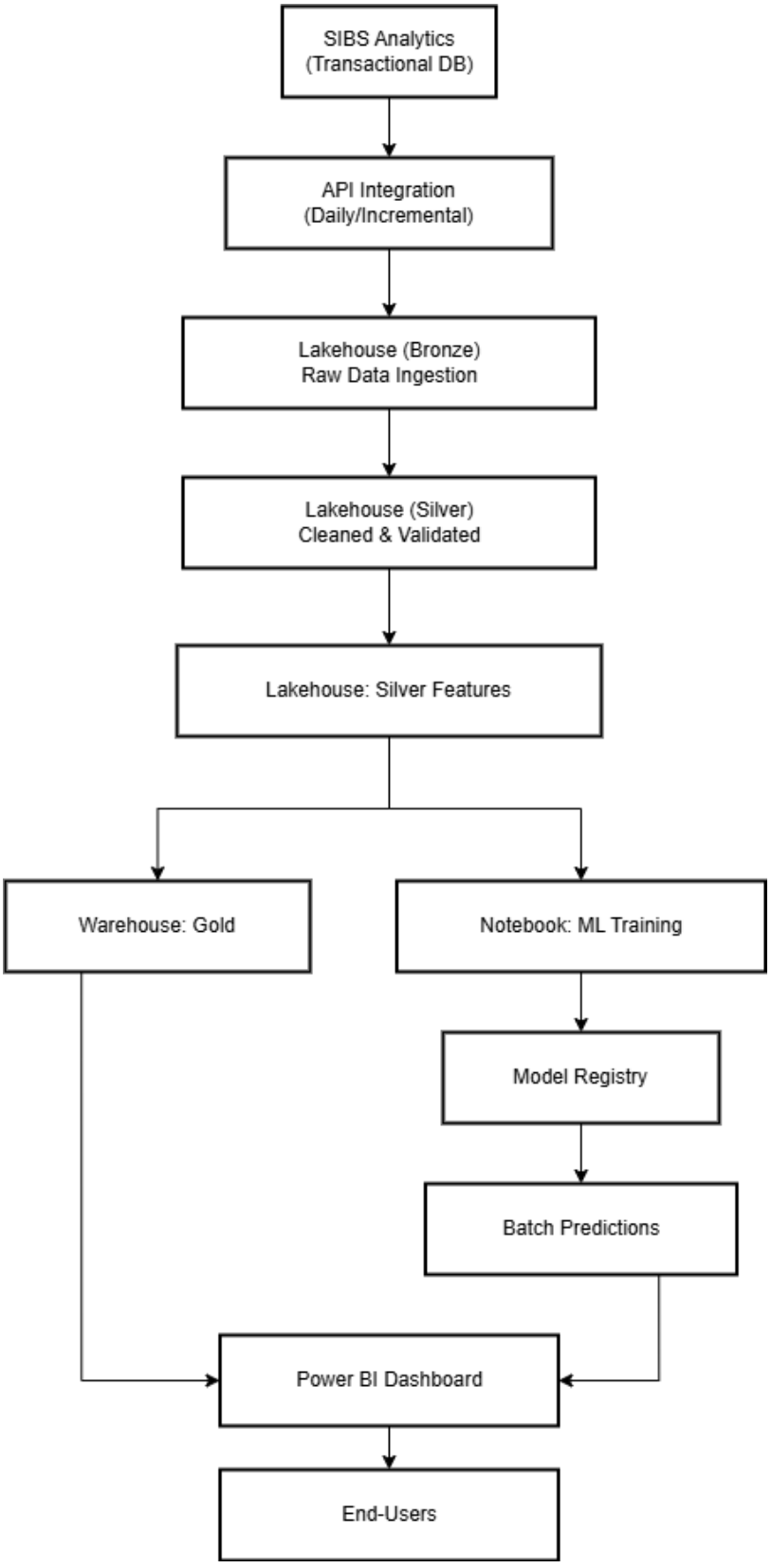
Drawing from the insight gained by the literature review chapter, there are three main pillars to consider for the deployment plan. The first one is scalability, as said by Bibri (Bibri, 2021), where we need to ensure that it must be able to adapt to the increasing tourist data volume. The second pillar is stakeholder empowerment, as stated by Yallop, that insight must be accessible for non-technical users. The last pillar is sustainability, where Partarakis (Partarakis et al., 2023) explained that we should minimize manual intervention maximizing automated workflows.

The first step in deploying the solution operationally involves establishing a secure and persistent API connection between SIBS Analytics and the project's Lakehouse environment. This integration would enable scheduled ingestion of new transactional data, seamlessly extending the existing dataset through the incremental load logic implemented at the extraction layer. By leveraging delta-based loading mechanisms, the system avoids full data reloads, significantly improving performance and reducing computational overhead during daily refresh cycles.

Once the data stream is in place, the automated data pipelines can be triggered using calendar-based scheduling rules. This allows future system administrators or municipal stakeholders to define precise update frequencies based on policy or operational needs. The architecture supports fully autonomous end-to-end execution, including forecasting model retraining and dashboard update, thereby minimizing human intervention and ensuring continuity over time.

To ensure that the deployed solution achieves its intended impact, targeted training workshops would be essential. These sessions should focus on familiarizing end-users with the dashboard's analytical capabilities. Through hands-on exercises and guided interpretation of key indicators, users will be empowered to extract actionable insights and incorporate data-driven thinking into their strategic decisions. This final step is critical to bridging the gap between technical implementation and effective policy application, reinforcing the long-term sustainability of the system. Bellow we can find a diagram explaining the deployment plan (Figure 4.6.)

Figure 4.6. Deployment diagram



5. CONCLUSIONS AND FUTURE WORKS

Although the proposed data-driven framework shows significant promise for aiding evidence-based decision-making in regions reliant on tourism, it is important to recognize several limitations.

Firstly, the analysis is based solely on card-based transactional data, which might not fully represent the spending habits of tourists who use other payment methods like cash. Consequently, the dataset, despite its breadth, offers only a partial perspective of the total economic activity.

Secondly, the data's temporal resolution, which is confined to weekly aggregates, limits the ability to capture short-term dynamics such as daily variations or the effects of specific events and holidays. While this level of detail is adequate for identifying larger trends, more frequent data could improve forecasting precision and adaptability.

Thirdly, assumptions were made about the origin of transactions based on the cardholder's geographic metadata. Although this serves as a useful indicator of visitor origin, it may not perfectly reflect actual tourist behaviour, especially in cases involving corporate or shared card usage. Furthermore, the forecasting models, despite being validated and optimized, operate under the assumption of consistent spending behaviour over time. Any sudden external disruptions, such as geopolitical crises, public health emergencies, or regulatory changes, could significantly compromise the accuracy of predictions. While efforts were made to minimize model overfitting and data leakage, these risks are inherent in predictive modelling, particularly with small regional datasets.

Lastly, the project assumes that stakeholders will have access to the necessary technological infrastructure to effectively maintain and use the dashboard. Implementing such systems in low-capacity municipal environments may require additional training, technical support, or funding, which could affect real-world adoption and impact.

This thesis established a scalable, data-driven framework for examining and predicting financial transactions in the Oeste CIM region of Portugal, focusing on tourism-related economic activities. Using transactional data from SIBS Analytics and the Kimball Lifecycle methodology, the project developed a Lakehouse-based architecture supporting microeconomic insights at the parish level. Dimensional modelling enabled an organized data warehouse design, while automated ETL pipelines converted raw POS/ATM data into actionable intelligence.

Tree-based forecasting models were trained to predict tourism expenditures across key indicators, achieving a mean absolute percentage error below 15%. These forecasts and visualizations were integrated into a Power BI dashboard, offering stakeholders insights into spending patterns, geographic distributions, and sector resilience.

The system provides municipalities and businesses with tools for managing seasonal fluctuations and optimizing resource allocation. This work advances the literature by operationalizing theoretical insights on integrating microeconomic indicators into urban planning tools. While research highlights the value of transactional data and predictive analytics in tourism management (Partarakis et al., 2023; Zhou, 2021), few studies offer an end-to-end implementation tailored to small, tourism-dependent regions. This thesis demonstrates both the feasibility and practical utility of such systems in local governance. It confirms the effectiveness of machine learning models—particularly ensemble tree-based methods—in predicting tourism-related economic trends, as previously suggested by Gricar (Gricar, 2023 and Essien & Chukwukelu (Essien & Chukwukelu, 2022).

The findings support key claims in the literature, weekend spending dominates overall volume (aligned with Ridderstaat (Ridderstaat et al., 2013)), and sector-specific analysis reveals economic resilience in areas like fuel and groceries (in line with Adhuze, (Adhuze, 2023)). The study provides a replicable framework combining technical scalability with analytical depth, suitable for deployment across various regional contexts.

The framework has limitations. The dataset reflects only card-based transactions, potentially underrepresenting cash-heavy demographics. Weekly aggregation limits high-frequency insights, and visitor origin is inferred from card metadata, which may introduce classification biases. The forecasting models assume consistent behavioural patterns, which may not hold under macroeconomic volatility. These limitations are balanced by the system's methodological transparency and modular design.

Looking ahead, incorporating additional data sources could enhance model accuracy and contextual understanding, as well as implementing daily data instead of weekly. Refining the dashboard for mobile and web interfaces would improve accessibility. A longitudinal evaluation of the system's impact on policy decisions would provide a benchmark for assessing data-driven governance in regional development.

6. BIBLIOGRAPHICAL REFERENCES

Adhuze, O. (2023). Infrastructure as Drivers for Economic Growth: A Way to Advancing Tourism. *International Journal of Latest Technology in Engineering Management & Applied Science*. https://doi.org/10.51583/ijltemas.2023.12908

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A Next-generation Hyperparameter Optimization Framework. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2623–2631. https://doi.org/10.1145/3292500.3330701

Arshad, M. O., Khan, S., Haleem, A., Mansoor, H., Arshad, M. O., & Arshad, M. E. (2021). Understanding the Impact of Covid-19 on Indian Tourism Sector Through Time Series Modelling. *Journal of Tourism Futures*. https://doi.org/10.1108/jtf-06-2020-0100

Becker, R. K. M. R. W. T. J. M. B. (2008). *The Data Warehouse Lifecycle Toolkit: ProQuest Tech Books - English/German*. 672. https://www.wiley.com/en-us/The+Data+Warehouse+Lifecycle+Toolkit%2C+2nd+Edition-p-9780470149775

Bibri, S. E. (2021). A Novel Model for Data-Driven Smart Sustainable Cities of the Future: The Institutional Transformations Required for Balancing and Advancing the Three Goals Of sustainability. *Energy Informatics*. https://doi.org/10.1186/s42162-021-00138-8

Bibri, S. E., & Krogstie, J. (2017). The Core Enabling Technologies of Big Data Analytics and Context-Aware Computing for Smart Sustainable Cities: A Review and Synthesis. *Journal of Big Data*. https://doi.org/10.1186/s40537-017-0091-6

CatBoost. (2025). https://catboost.ai/docs/en/

Chen, Q., Zhang, M., & Zhao, X. (2017). Analysing Customer Behaviour in Mobile App Usage. *Industrial Management & Data Systems*. https://doi.org/10.1108/imds-04-2016-0141

Choi, J. E., & Shin, D. W. (2019). The Roles of Differencing and Dimension Reduction in Machine Learning Forecasting of Employment Level Using the FRED Big Data. *Communications for Statistical Applications and Methods*. https://doi.org/10.29220/csam.2019.26.5.497

Comerio, N., & Strozzi, F. (2018). Tourism and Its Economic Impact: A Literature Review Using Bibliometric Tools. *Tourism Economics*. https://doi.org/10.1177/1354816618793762

Eckstein, M., & Kollar, M. (2008). Nonthermal Steady States After an Interaction Quench in the Falicov-Kimball Model. *Physical Review Letters*. https://doi.org/10.1103/physrevlett.100.120404

Essien, A., & Chukwukelu, G. (2022). Deep Learning in Hospitality and Tourism: A Research Framework Agenda for Future Research. *International Journal of Contemporary Hospitality Management*. https://doi.org/10.1108/ijchm-09-2021-1176

Fernald, J. G. (2012). A Quarterly, Utilization-Adjusted Series on Total Factor Productivity. *Erwp*. https://doi.org/10.24148/wp2012-19

Flerchinger, G. N., Xaio, W., Marks, D., Sauer, T. J., & Yu, Q. (2009). Comparison of Algorithms for Incoming Atmospheric Long-wave Radiation. *Water Resources Research*. https://doi.org/10.1029/2008wr007394

Gricar, S. (2023). Tourism Forecasting of "Unpredictable" Future Shocks: A Literature Review by the PRISMA Model. *Journal of Risk and Financial Management*. https://doi.org/10.3390/jrfm16120493

Hall, A. S. (2018). Machine Learning Approaches to Macroeconomic Forecasting. *The Federal Reserve Bank of Kansas City Economic Review*. https://doi.org/10.18651/er/4q18smalterhall

HistGradientBoostingRegressor - sklearn. (2025). https://scikit-learn-ts-git-feature-docs-2-saasify.vercel.app/docs/classes/HistGradientBoostingRegressor

Huang, S., & Wang, X. (2022). COVID-19 Two Years On: A Review of COVID-19-related Empirical Research in Major Tourism and Hospitality Journals. *International Journal of Contemporary Hospitality Management*. https://doi.org/10.1108/ijchm-03-2022-0393

Irawan, R. Y., Susanto, B., & Lukito, Y. (2021). Building Data Warehouse and Dashboard of Church Congregation Data. *Jurnal Terapan Teknologi Informasi*. https://doi.org/10.21460/jutei.2019.32.183

Jardim, B., de Castro Neto, M., Magalhães de Sousa, N., Barriguinha, A., & Sarmento, P. (2025). An indicator for integrated regional planning: A case study of Portugal's west region. *Cities*, *159*, 105762. https://doi.org/https://doi.org/10.1016/j.cities.2025.105762

Kimball, R., Ross, M., Thornthwaite, W., Mundy, J., & Becker, B. (2008). *The Data Warehouse Lifecycle Toolkit, 2nd Edition: Practical Techniques for Building Data Warehouse and Business Intelligence Systems*. 672. http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:The+Data+Warehouse+Lifecycle+Toolkit-Practical+Techniques+for+Building+Data+Warehouse+and+Business+Intelligence+Systems,+ed#0

Kitchin, R. (2014). Making Sense of Smart Cities: Addressing Present Shortcomings. *Cambridge Journal of Regions Economy and Society*. https://doi.org/10.1093/cjres/rsu027

Law, R., Leung, D., & Chan, I. C. C. (2019). Progression and Development of Information and Communication Technology Research in Hospitality and Tourism. In *International Journal of Contemporary Hospitality Management*. https://doi.org/10.1108/ijchm-07-2018-0586

LeCun, Y., Bengio, Y., & Hinton, G. E. (2015). Deep Learning. *Nature*. https://doi.org/10.1038/nature14539

Li, J., Guo, X., Lu, R., & Zhang, Y. (2022). Analysing Urban Tourism Accessibility Using Real-Time Travel Data: A Case Study in Nanjing, China. *Sustainability*. https://doi.org/10.3390/su141912122

Liu, H., & Song, H. (2017). New Evidence of Dynamic Links Between Tourism and Economic Growth Based on Mixed-Frequency Granger Causality Tests. *Journal of Travel Research*. https://doi.org/10.1177/0047287517723531

MAPE - Mean Absolute Percentage Error — Permetrics 2.0.0 documentation. (n.d.). Retrieved June 19, 2025, from https://permetrics.readthedocs.io/en/latest/pages/regression/MAPE.html

Mullainathan, S., & Spiess, J. (2017). Machine Learning: An Applied Econometric Approach. *Journal of Economic Perspectives*. https://doi.org/10.1257/jep.31.2.87

Partarakis, N., Karuzaki, E., Ntoa, S., Ntagianta, A., Zidianakis, E., & Stephanidis, C. (2023). An Open-Data Repository for Sustainable Tourism. *Highlights of Sustainability*. https://doi.org/10.54175/hsustain2030011

Plataforma online da Região Oeste de Portugal. (2024, September 25). https://www.oestecim.pt/

Rahman, F. (2023). Local Economic Impact of COVID-19 on the Urban Tourism-Related Services: A Perspective of Kochi Heritage City, Kerala. *Sustainability*. https://doi.org/10.3390/su152416585

Ridderstaat, J., Croes, R., & Nijkamp, P. (2013). Tourism and Long-run Economic Growth in Aruba. *International Journal of Tourism Research*. https://doi.org/10.1002/jtr.1941

RMSE - Root Mean Square Error — Permetrics 2.0.0 documentation. (2025). https://permetrics.readthedocs.io/en/latest/pages/regression/RMSE.html

Sakhuja, S., Jain, V., Kumar, S., Chandra, C., & Ghildayal, S. K. (2016). Genetic Algorithm Based Fuzzy Time Series Tourism Demand Forecast Model. *Industrial Management & Data Systems*. https://doi.org/10.1108/imds-05-2015-0165

Shen, Z., Wan, Q., & Leatham, D. J. (2021). Bitcoin Return Volatility Forecasting: A Comparative Study Between GARCH and RNN. *Journal of Risk and Financial Management*. https://doi.org/10.3390/jrfm14070337

Simões, P., De Castro Neto, M., Sarmento, P., & Barriguinha, A. (2023). Oeste smart region. An intermunicipal integrated analytical territorial intelligence platform. *REVISTA INTERNACIONAL MAPPING*, *32*, 50–61. https://doi.org/10.59192/mapping.395

Suleiman, N. N., & Albiman, M. M. (2014). Dynamic Relationship Between Tourism, Trade, Infrastructure and Economic Growth: Empirical Evidence From Malaysia. *Journal of African Studies and Development*. https://doi.org/10.5897/jasd2013.0260

Svetlana, S. V., & Vladimir, S. I. (2019). Way to Assess the Development of Municipal Tourism Infrastructure. *Istrazivanja I Projektovanja Za Privredu*. https://doi.org/10.5937/jaes17-17073

Welcome to LightGBM's documentation! — LightGBM 4.6.0.99 documentation. (n.d.). Retrieved July 6, 2025, from https://lightgbm.readthedocs.io/en/latest/index.html

XGBoost Documentation — xgboost 3.1.0-dev documentation. (2025). https://xgboost.readthedocs.io/en/latest/index.html

Yallop, A. C., & Séraphin, H. (2020). Big Data and Analytics in Tourism and Hospitality: Opportunities and Risks. *Journal of Tourism Futures*. https://doi.org/10.1108/jtf-10-2019-0108

Yoon, J., & Choi, C. (2023). Real-Time Context-Aware Recommendation System for Tourism. In *Sensors*. https://doi.org/10.3390/s23073679

Yunus, M. R., Susanti, G., Jamaluddin, J., & Asriadi, A. (2021). The Impact of Public-Private Partnerships on Development of Tourism Infrastructure Destination and Tourism Service Innovation as Mediating Variable in Sinjai Regency, Indonesia. *Turkish Journal of Computer and Mathematics Education (Turcomat)*. https://doi.org/10.17762/turcomat.v12i11.6062

Zhao, Y. (2024). Navigating the Confluence of Econometrics and Data Science: Implications for Economic Analysis and Policy. *Theoretical and Natural Science*. https://doi.org/10.54254/2753-8818/38/20240551

Zhou, W. (2021). Prediction of Urban and Rural Tourism Economic Forecast Based on Machine Learning. *Scientific Programming*. https://doi.org/10.1155/2021/4072499

7. APPENDIX A

“Dear Alexandre Spagnol,

Dear Professor Bruno Jardim,

Thank you for filling out the Research Ethics Checklist. After reviewing your request, you can proceed with the study as we do not foresee any major ethical concerns with the project.

Project No.: DSCI2025-6-191280

Project Title: Analysis of Monetary Transactions in Oeste CIM

Principal Researcher: Alexandre Spagnol

according to the regulations of the Ethics Committee of NOVA IMS and MagIC Research Center this project was considered to meet the requirements of the NOVA IMS Internal Review Board, being considered APPROVED on 21/06/2025.

It is the Principal Researcher’s responsibility to ensure that all researchers and stakeholders associated with this project are aware of the conditions of approval and which documents have been approved.

The Principal Researcher is required to notify the Ethics Committee, via amendment or progress report, of

- Any significant change to the project and the reason for that change;

- Any unforeseen events or unexpected developments that merit notification;

- The inability of the Principal Researcher to continue in that role or any other change in research personnel involved in the project.

Lisbon, 21/06/2025

NOVA IMS Ethics Committee

ethicscommittee@novaims.unl.pt

This email serves as formal proof of ethical approval. If required for inclusion in a thesis, dissertation, or any other academic documentation, a PDF version of this message may be created and attached accordingly.“

8. ANNEXES

Annex 1. Measure Definition Details

FCT_TRANSACTIONS	
CY Number of Cards	Var CY = MAX('Dim_Date'[Year]) RETURN CALCULATE([Total Number of Cards], 'Dim_Date'[Year]=CY)
CY Number of Transactions	Var CY = MAX('Dim_Date'[Year]) RETURN CALCULATE([Total Number of Transactions], 'Dim_Date'[Year]=CY)
CY Transaction Value	Var CY = MAX('Dim_Date'[Year]) RETURN CALCULATE([Total Transaction Value], 'Dim_Date'[Year]=CY)
PY Number of Cards	CALCULATE([Total Number of Cards], SAMEPERIODLASTYEAR('DIM_Date'[Date]))
PY Number of Transactions	CALCULATE([Total Number of Transactions], SAMEPERIODLASTYEAR('DIM_Date'[Date]))
PY Transaction Value	CALCULATE([Total Transaction Value], SAMEPERIODLASTYEAR('DIM_Date'[Date]))
Total Number of Cards	SUM(FCT_Transactions[Number_of_Cards])
Total Number of Transactions	SUM(FCT_Transactions[Number_of_Transactions])

Total Transaction Value	SUM(FCT_Transactions[Transaction_Value])
YoY Variance Number of Cards	[CY Number of Cards] - [PY Number of Cards]
YoY Variance Number of Transactions	[CY Number of Transactions] - [PY Number of Transactions]
YoY Variance Transaction Value	[CY Transaction Value] - [PY Transaction Value]
YoY% Number of Cards	VAR __PREV_YEAR = CALCULATE(SUM('FCT_Transactions'[Number_of_Cards]), DATEADD('Dim_Date'[Date], -1, YEAR)) RETURN DIVIDE(SUM('FCT_Transactions'[Number_of_Cards]) - __PREV_YEAR, __PREV_YEAR)
YoY% Number of Transactions	VAR __PREV_YEAR = CALCULATE(SUM('FCT_Transactions'[Number_of_Transactions]), DATEADD('Dim_Date'[Date], -1, YEAR)) RETURN DIVIDE(SUM('FCT_Transactions'[Number_of_Transactions]) - __PREV_YEAR, __PREV_YEAR)

YoY% Transaction Value	VAR __PREV_YEAR = CALCULATE(SUM('FCT_Transactions'[Transaction_Value]), DATEADD('Dim_Date'[Date], -1, YEAR)) RETURN DIVIDE(SUM('FCT_Transactions'[Transaction_Value]) - __PREV_YEAR, __PREV_YEAR)
YTD Number of Cards	TOTALYTD([CY Number of Cards], 'Dim_Date'[Date])
YTD Number of Transactions	TOTALYTD([CY Number of Transactions], 'Dim_Date'[Date])
YTD Transaction Value	TOTALYTD([CY Transaction Value], 'Dim_Date'[Date])

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa