TABLE OF CONTENTS

1. BACKGROUND

The COVID-19 pandemic has significantly impacted the global tourism sector, presenting meaningful challenges over the past few years. As is well known, Portugal is the 7th best destination in Europe and 9th in the world, being recognized for its breathtaking landscapes and rich cultural heritage. However, this sector has experienced notable changes, especially in tourist preferences and behaviors. To effectively manage our country's promotional efforts, the Portuguese National Tourism Board Office (NTBO) recognizes the importance of a comprehensive analysis of the evolving tourism environment. Therefore, this project aims to use data analytics to compare Portugal's leading tourist destinations with top competitors across Europe, providing insightful conclusions about traveler characteristics and behavioral patterns. By using TripAdvisor reviews and travel trend data, this study seeks to uncover underlying trends, traveler behaviors, and potential strategies to boost Portugal's attractiveness as a leading travel destination. With the main goal of supporting Portugal's tourism resilience and sustained growth, the findings from this research will guide NTBO's marketing strategies and align them with the current post-pandemic travel reality.

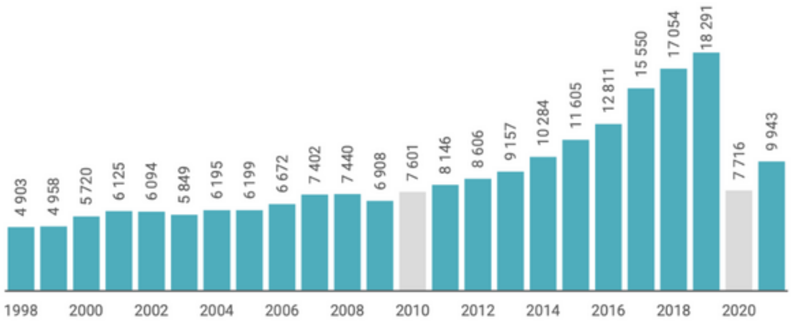
2. BRIEF DESCRIPTION OF THE PROBLEM

The objective of this project is to analyze how the COVID-19 pandemic has impacted tourism patterns at major Portuguese attractions. We aim to provide the Portuguese National Tourism Board Office (NTBO) with a detailed, data-driven analysis of visitor reviews and other tourism-related data to help them understand how preferences, demographics, and satisfaction levels have evolved. The challenge has two main components: first, to capture the current state of Portuguese tourism by examining how the pandemic has influenced visitor demographics, behaviors, and satisfaction levels; and second, to assess Portugal's position in the tourism market by comparing these findings with those from other European countries. The resulting insights will inform the NTBO's strategic and marketing decisions, enabling them to develop targeted campaigns and recovery initiatives that align with the preferences of today's travelers. This work is crucial for establishing a resilient, long-term tourism strategy for Portugal that can endure future market crises.

3. BUSINESS UNDERSTANDING

The tourism industry is a **driven force** behind the Portuguese economic growth. In 2023, tourism revenue represented 9.55 of the GDP, breaking records that had been established in 2019, a year before the pandemic for overnight stays (+10.0%) and guests (+10.7%). Moreover, "30.0 million guests were registered, of which 18.3 million were foreigners, an increase of 13.3% and 19.1% respectively, when compared to 2022." (Tourism in Portugal, 2024). Considered to be a safe and secure country with a unique blend of culture, natural beauty, all- year good weather and a fabulous gastronomic journey. Portugal is a country that presents several reasons for people to visit and spend their holidays. Thus, as being one the biggest financiers of the country, Portugal relies severely in this performance of this sector.

In 2020, the **COVID-19 viruses caused a horrific impact provoking a crisis for the tourism industry and the Portuguese economy**. As it can be seen Pandemic led to a downfall in tourism which forced small businesses to close their businesses. Moreover, with restrictions several companies and events, such as festivals and concerts weren't allowed to happen. Consequently, employees for several sectors had to be laid off, since the workload didn't correspond to the need for a high number of employees, corresponding to a monetary incapacity for the companies to support. As can be seen in *Figure 1,* in 2020, in nominal terms, the national GDP recorded a decrease of 6.7%, reflecting the negative impact of the COVID-19 pandemic, in which the tourism sector suffered a 58% decline in revenue in that same year.



Fonte: Banco de Portugal

Figure 1: Evolution of Overnight Stays in Tourist Accommodation Establishments (2000-2019)

In *Figure 2*, presented by Statistica, there was a tremendous downfall in the total contribution of travel and tourism to the gross domestic product in 2020. On the other hand, it can also be observed a recovery of over 30% in 2021, and by the year 2023, Portugal was able to surpass 2019 numbers, contributing 56 billion U.S. dollars to GDP.

3. BUSINESS UNDERSTANDING

To rebuild tourism, there were significant investments allocated to the tourism industry, backed by National Tourism Board Organizations (NTBO) and government efforts. These efforts have concentrated on improving Portugal's infrastructure and maintaining a high level of hospitality to draw in and serve an international audience. Ending 2024 and starting 2025 it can be seen the positive impacts of the effort and how it has amplified tourism numbers in Portugal. Nevertheless, the comprehensive analysis that will be performed will allow to Portuguese National Tourism Board's strategic decisions, regarding how can Portugal continue increasing and improving in the tourism sector compared to its competitors, analyzing new opportunities of how Portugal can be positioned as a leading European destination.

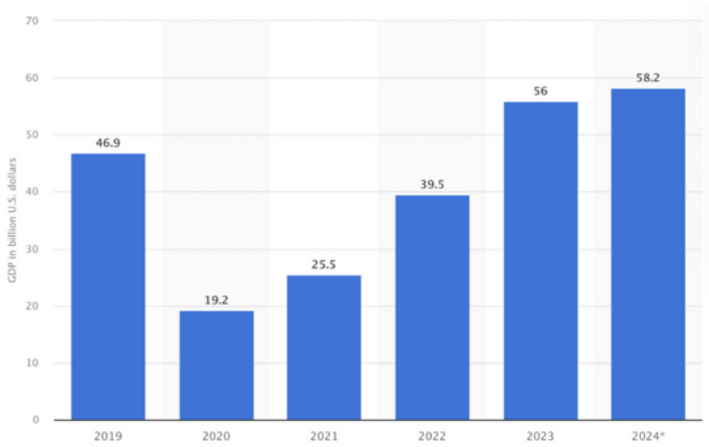


Figure 2: Total contribution of travel and tourism to the gross domestic product in Portugal from 2019 to 2024 (in billion U.S dollars)

Through *Figure 3* it can be observed the tendency for overnight stays had a significant decrease, only showing an improvement in 2022 recovering almost entirely from the pandemic impact. The imposition of several restrictions, and the fear of contracting the COVID-19 virus led to an uncomfortable feeling about traveling. Moreover, with the reinforcement of work remotely, traveling to work events or congresses, since they could do it from the comfort of their homes.

3. BUSINESS UNDERSTANDING

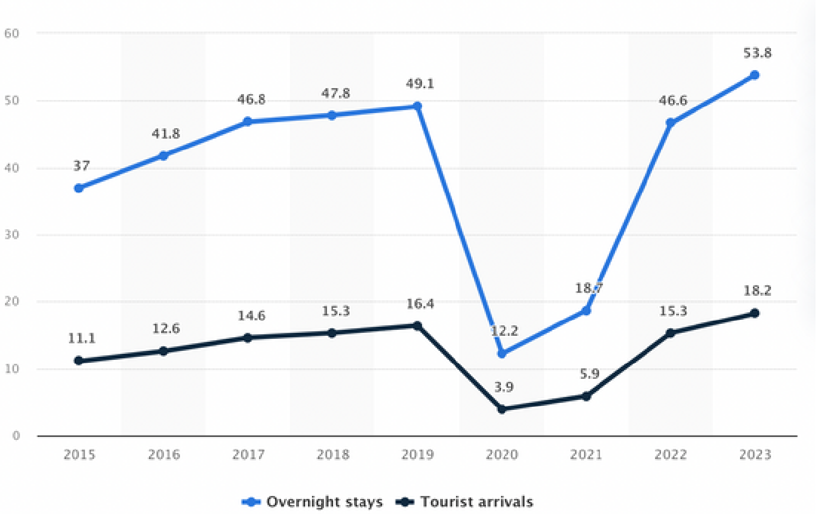


Figure 3: Number of international tourists' arrivals and overnight stays in Portugal from 2015 to 2023 (in millions)

On a final analysis of our business understanding, we focus on determining the main competitors of Portugal in the tourism, and after looking at business journals and other papers related to tourism we identified **Spain, Italy, and Greece as our main competition**. This decision was made because these 3 counties with Portugal are the ones that represent the South of Europe.

To rebuild tourism, there were significant investments allocated to the tourism industry, backed by National Tourism Board Organizations (NTBO) and government efforts. These efforts have concentrated on improving Portugal’s infrastructure and maintaining a high level of hospitality to draw in and serve an international audience. Ending 2024 and starting 2025 it can be seen the positive impacts of the effort and how it has amplified tourism numbers in Portugal. Nevertheless, the comprehensive analysis that will be performed will allow to Portuguese National Tourism Board's strategic decisions, regarding how can Portugal continue increasing and improving in the tourism sector compared to its competitors, analyzing new opportunities of how Portugal can be positioned as a leading European destination.

4. DATA UNDERSTANDING AND PREPARATION

4.1. ATTRACTIONS DATASET

Started this process by checking the information available in the data frame, which corresponded to a table with 100 columns and 4 rows, regarding the ID, Name, Country Name, and ISO. In the dataset, the number of duplicates with the same name, statistics, unique ISOs, countries, and missing values was verified. Following these verifications, it was noticed that the Country name Scotland was misspelled suffering a change from Scot to Scotland. As research was done regarding the country's ISOs the “UK” ISO was changed to ‘GB’, since those are the initials representing all the countries from the United Kingdom. Moreover, Curaçao despite being considered to a part of the Netherlands the code was changed from “NL” to “CW”. To ensure that the ISO codes from the data frame matched the ISO codes from the dictionary a new column was created, which was ‘ISO_Check’. For the rows that the ISO wasn’t matching with the dictionary, those were altered. After the corrections, since it wasn’t needed anymore the ‘ISO_Check’ column was deleted.

By creating a horizontal bar, *Figure 4* with the numbers provided from the data frame and having those presented in ascending order, it was able to recognize that **Spain, Italy, and Great Britain were the top 3 countries with the highest attractions**.

4.2. REVIEWS DATASET

On the reviews dataset, we encountered a table with 15 columns and 92,120 rows.
On our first analysis, we identified that the variable tripType has 31.55% missing values, userLocation accounts for 14.62%, reviewVisited has 0.77% missing values, and “username” shows the lowest proportion, with only 0.02% of values missing. Because of the importance userLocation and tripType columns for our analysis we decided replace those missing values with Unknown.

We chose to eliminate all rows with missing values in the remaining columns because conducting our analysis without the userName of the reviewer would be impractical. Additionally, we observed in the reviewVisited column that a small number of reviews were dated before 2019. Since our primary analysis focuses on more recent data, we excluded rows that weren't ritghen after 01/01/2019. While checking the LocalID column, we noticed there was a strange value identified as "u", so, we just decided to remove the row entirely. Another important step in the data preparation was to verify if any of the reviews were written before the actual date of the visit, which we inicitialy encounter by random occasion while just exploring the excel file. We identified 20 rows that exhibited this inconsistency, and were subsequently removed to maintain the reliability of our data.

The “userContributions” column displayed a distribution with positive skewness,*Figure 5*, which could interfere with the accuracy of statistical or predictive models. To address this, we applied a logarithmic transformation.

On our final examination of all the variables in the sheet, we also decided to keep "positionOnRanking" and "siteOnRanking" ,as in our perspective at time, could offer valuable insights into attraction rankings .

The final analysis involving the data sheet assessed whether the numeric columns in the dataset were chosen for proper application of statistical calculations. First, kurtosis for these variables was calculated, indicating the shape of distribution to identify characteristics such as extreme values or long tails. From there, we created a correlation matrix to assess the strength and direction of linear relationships between our numeric variables.

In the correlation matrix, *Figure 6,* these are the main findings that are observable: a moderate positive correlation between the variables totalReviews and sitesOnRanking, with r = 0.4, and a weak negative correlation between globalRating and positionOnRanking, with r = -0.2; for most other variables the correlation is weak or insignificant, r values below 0.1, suggesting little direct relationship between them.

For the final visualisation method, we created a set of histograms, to helps record the frequency distribution of the variables, making it easier to identify patterns, asymmetries, and outliers.

On our histograms, *Figure 7*, the main findings we observed are: **globalRating and reviewRating are highly concentrated at high values**; variables positionOnRanking and sitesOnRanking show positively skewed distributions; **the distribution of totalReviews is strongly asymmetric**; the userContributions variable shows a more symmetrical distribution after the logarithmic transformation, demonstrating that the normalization process was effective in reducing skewness

For our categorical variables we first examine the usercountry column by creating a dictionary, that was given to us by ChatGPT, with values that corresponded to the original names of the countries. After the implementation we can see that the most strong countries are the UK and the USA followed by Unkwon, wich is a problem that we will deal with later.

For our examination of our last categorical variable, that could have the most relevant information for our project, the reviewvisited, we group up this variable by each of the four different seasons of the year. On *Figure 8 we* can see that the times **of the year in wich the reveiwers traveled were in the summer, followed by spring and autumn and lastly winter.**

4. DATA UNDERSTANDING AND PREPARATION

4.3. HOLIDAY DATASET

At first, we started by importing the dataset from Holidays.csv file, using a semicolon as the delimiter. Then, just like in the last two datasets, we created a copy of the original dataset, so we could work on it without affecting the original data. At this stage, we began the Data Understanding itself and computed some descriptive statistics, concluding that the dataset appeared to have 69,557 rows and 7 columns, the last featuring 7 categorical variables, including the 'date' column, which was also stored as an object. By performing a comprehensive analysis to the descriptive statistics we found out that the 'countryOrRegion' column had more values than the 'ISO', indicating that some countries lacked corresponding codes. Also, the 'isPaidTimeOff' column had a lot of missing values. In the end of the Data Understanding stage, we checked if the columns 'holidayName' and 'normalizedHolidayName' had the same values in all rows, as the name could suggest that they were the same, with the possibility of being a duplicated column. We then concluded that they had some differences, which were explored in the next stage.

Afterward, we proceeded with the Data Preparation phase. To enhance clarity, we started by renaming the 'Unnamed: 0' column to 'ID', as it is a column with unique values that can be used as an holiday ID column.

Then, we checked the unique differences between the 'holidayName' column and the 'normalizedHolidayName'. Through this, we found out that there were a lot of **misspellings**, specially in the first column which sometimes contained **irrelevant content** inside parenthesis, like the word 'Observed' and the country where the holiday happens, being **redudant** as we already had the 'countryOrRegion' variable. Therefore, we decided to clean and use the 'normalizedHolidayName' column and drop 'holidayName'.

As we previously saw, there were two variables with missing values: 'countryRegionCode' and 'isPaidTimeOff', with 7.19% and 94.34% respectively. For the first column, we renamed it from 'countryRegion Code' to 'ISO', so it would match the name in the previous datasets. We perceived that the null values in 'ISO' were associated to specific countries, more precisely Scotland, Northern Ireland, England and Wales, which are consolidated under the 'GB' code.

We addressed this by filling the missing values with 'GB'. Further, our analysis revealed that the 'isPaidTimeOff' field had too many null entries, limited to few countries like US, GB and India. We decided to drop this column, since we were not able to impute values and creating an 'Unknown' category wouldn't allow us to draw any relevant conclusions.

As we mentioned at the beginning, the column ‘date’ was a string (object), so we converted to a datetime format ('%d/%m/%Y'), representing day, month, and year, respectively (no hours included). Then, we filtered the rows between 01/01/2019 and 31/12/2021, which were the ones relevant to our analysis.

Following this, we checked how many holidays we had before treating the ‘normalizeHolidayName’ column, resulting in **438 different holidays.**

This column contained numerous unique entries, and underwent a series of processing steps. First, we perceived that multiple holiday names were concatenated, and separated by commas. For instance, "Christmas, New Year's Eve" was split into ["Christmas", "New Year's Eve"]. Therefore, we employed the 'explode' method in pandas, transforming each element of these lists into individual rows. For example, ["Christmas", "New Year's Eve"] resulted in the separate rows "Christmas" and "New Year's Eve". Proceeding, we used the 'str.strip' method to remove any leading or trailing whitespace surrounding the holiday names, ensuring organized and consistent entries. Lastly, we removed any content within square brackets [], which contained English translations that were not necessary for our analysis, stripping the extra whitespace again. After this, and within the data range that we implemented, we had 342 different holidays, 38 countries/regions and 34 ISO codes (we have 4 more countries than ISO codes since Scotland, Northern Ireland, England and Wales all belong to 1 ISO code: GB).

Additionally, to ensure a universal understanding of the dataset, we **standardized** the values in the ‘normalizeHolidayName’ field by replacing the holiday names presented in foreign languages with their English equivalents, through a mapping dictionary provided by ChatGPT. For example, instead of ‘Año Nuevo' we have ‘New Year’, making the dataset easier to read and presenting it in a much simpler and more concise structure. Lastly, for a better perception, the ‘normalizeHolidayname’ column name was modified to just ‘HolidayNames’. We ended our Data Preparation phase with only 236 unique holidays, being Sunday, New Year and Christmas the most frequent ones.

Finally, we performed some Data Visualization, starting by the Total Holidays by Country (*Figure 9*), which showed us Sweden, Norway and the United Kingdom emerge as the countries boasting the greatest number of holidays, around 204, 192, and 163 holidays, respectively. On the contrary, India, Australia, Switzerland, Germany and Spain rank among the countries with the **fewest holidays**, each having a total of fewer than 30 holidays. We also conclude that **December** is the month with the **highest number of holidays**, followed by April and May (*Figure 10*).

4. DATA UNDERSTANDING AND PREPARATION

4.4. MERGING ATTRACTIONS AND REVIEWS

During this General Analysis, it decided to have a greater focused on the countries, Spain, Italy, and Greece, that had been Portugal's main tourism competitors. To obtain a greater understanding of the impact of the pandemic it was first decided to perform a quarter analysis in order to understand the differences in the number of visits across the different quarters of the year. Therefore, we divided the months of the year by the corresponding quarter and performed a visualization of the distribution of the number of visits through each quarter and conclude that the third quarter which encompasses July, August and September, is the quarter people travel more and in the fourth quarter (October, November and December) people travel less. This analysis can be linked to the fact that these quarters correspond to summer and winter respectively, leading to a noticeable discrepancy between them every year.

A map*, Figure 11*, was designed for a review dataset with the months corresponding to the quarters. In this one, it was able to understand that there was a **predominance in the third quarter** with the highest visit count, representing July, August, and September.

Following this, it was created a new map to understand the number of visits to the attractions by quarters for Portugal and its competitor countries, Spain, Italy, and Greece. For a more comprehensive analysis, we wanted to perceive if the conclusion regarding the total number of visits per quarter is applicable to Portugal and its competitors. The visual representation, *Figure 12*, leads us to understand that in general people **don’t tend to travel** to these countries during October, November and December (Quarter 4). On the other hand, Quarter 3 of the year represents the months of election for people to travel the most, representing a higher competition for Portugal to attract tourists.

Throughout the years Portugal has been gaining more popularity, especially after the pandemic hit. As a country that presents weather, tourist sites, and sports destinations it’s interesting to understand how Portugal can remain consistent throughout the entire year, compared to Italy and Greece. Moreover, despite how close and similar we can be to Spain, this country is one of the strongest competitors also presenting a strong consistency and high visit numbers.

As it can be seen in *Figure 13*, Portugal in the fourth quarter, is the one where we got less visitors, just like in Spain, Italy and Greece. The four countries have more visitors in the second and third quarters, as they are renowned for their warm spring and summer climates rather than the cold days of winter.

To have a greater understanding of the pandemic impact on tourism, more specifically in Portugal and its competitors, Spain, Italy, and Greece conducted a more in-depth analysis. Consequently, it made an analysis focused on the periods, pre-pandemic and after the pandemic, analyzing the ISO and the review count for the countries in question. It merged the pre-pandemic and the post-pandemic reviews on 'country' to do a comparison by having on the suffixes the 'pre' and 'post'. For all the countries that were missing reviews, these were substituted with a 0, to allow a smother calculation for the percentage change in the willingness that people had to write reviews regarding their holiday experience.

Through *Figure 14*, it can be observed people's frequency of writing reviews pre-Covid and post-Covid. For this one, it used values regarding the ISO, Review Count, and Time Period. Through the visual representation of the results, a strong tendency has been seen for people to be enthusiastic about writing reviews before the pandemic. The drastic change testified led us to conclude that the **impact of this virus significantly changed people's perspectives on writing reviews**.

To find the maximum percentage for both datasets it was used the same scale was created the *Figure 15*, with two subplots, one to plot the pre-Covid data for competitors, and analyze the country and percentage datasets, and the other for post-Covid, going through the country name and the percentage of visits. Comparing the values presented it can be seen that the virus **didn't have an impact on preference** regarding these 4 countries. Spain remains the election country for people to visit, compared to the other 3, Portugal, Italy, and Greece, this country suffered a significant negative impact on tourism.

Taking a closer look at an analysis of Portugal. It was created as a table, in which all locations that presented missing values and were represented by the name 'Unknown' were excluded. Filtering and merging all the data allowed us to understand where those tourists lived (country-wise) and how many tourists were flying from each country. Thus, the user country and the visit count were grouped. The values were reorganized to be in descending order to simplify the analysis and understand who the top visitor countries were. Following this, it was selected the top 10 visitor countries to create a table representing the top 10 visitor origins for Portugal. Through *Figure 16* it was able to understand that the top 3 were the United Kingdom, USA, and Canada.

As the journal Portugal Resident states, “To put the numbers into perspective, the number of US tourists staying at Portuguese tourist accommodation establishments between January and June 2022 was just over 584,000, meaning the number has nearly doubled in just two years.” (Bruxo, 2024). After the Pandemic the numbers have been increasing consistently, breaking records. This is associated with the discovery of this country that offers a rich culture, delicious food, and places to visit and relax. As this American attraction is a surprise, for the English to be ranked 2 it’s not. Portugal is a country that offers good weather year long and has been consistently considered to be one of the greatest destinations for golf in the world, as it was this year by the World Golf Awards, this is the perfect place for the United Kingdom travelers to come to. In third place, Canadians have been electing Portugal as its holiday country and like the growth noticeable with the other two countries this one is no different. In October of 2024, “the Canadian market stood out with the biggest annual increase (+15%) among the ten main tourist-sending markets, and, simultaneously, resident overnight stays reversed the downward trend recorded in September, rising 1.2%, to a total of 1.9 million.” (Lusa, 2024).

This virus has severely impacted several sectors, tourism being one of those. The implementation of safety measures, travel restrictions, and the fear of catching the virus led people to feel anxious and insecure about traveling. However, as time passed the desire to go back to their routines and lifestyles was bigger. Tourism in Portugal pos- covid has been significantly increasing presenting yearly consistencies in the increase of tourism and people wanting to visit the country.

5. MODELING

5.1. RFM MODEL

The RFM Model is a widely used lookup model, considered the easiest form of customer segmentation, used to analyze customer purchasing behaviors. The acronym "RFM" stands for Recency, Frequency, and Monetary Value, which are the three core factors considered in this model. Each of these dimensions provides valuable insights into different aspects of customer behavior. Based on this concept, we developed four distinct RFM Models for our project:

- **RFM for All Users:** This model segments users by their usernames, using reviewWritten as the Recency metric, userContributions for Frequency, and reviewRating as the Monetary measure.
- **RFM for Portuguese Users:** Same measures as we did in the RFM for all users but this time we filtered results to only get Portuguese users.
- **RFM for Portuguese Attractions:** This RFM model segments Portuguese attractions based on user reviews, using reviewVisited for Recency, totalReviews for Frequency, and reviewRating for Monetary. Recency is calculated as the number of days since the most recent review for each attraction, Frequency represents the total number of reviews for each attraction, and Monetary is measured as the average review rating.
- **RFM by Portugal Visits (Country of Origin):** This RFM model will segment the users that have visited Portugal by their original country (usercountry). On this model we used reviewVisited for Recency, totalReviews for Frequency and reviewRating for monetary.

These steps were designed to provide a comprehensive analysis of customer behavior and segment performance, ensuring insightful and actionable models.

5. MODELING

5.1.1. RFM FOR ALL USERS

Starting with the RFM for all users, first we calculated the Recency by determining the days since each user's most recent review , stored in ‘MostRecentView’ variable, Frequency by counting their total contributions and renaming it ‘TotalContributions’, and Monetary by averaging their ratings and save the values in the ‘AverageRating’ column. The quantiles (25th, 50th, and 75th percentiles) for each metric were calculated and we assigned a score from 1 to 4 for each dimension based on where their values fall within these ranges.Finally, we combined the scores resulting in a RFMScore, sorting users based on this score.

The following histograms show the distribution of Recency, Frequency, and Monetary scores across all users. From them, we can’t take much conclusions, as Recency and Frequency scores are evenly distributed. On the contrary, the Monetary score is skewed heavily, with the absence of users with 3 and 4 values of MScore.

Therefore, we decided to compute three bar plots showing the Average Recency by Rscore, the Average Frequency by FScore and the Average Monetary by MScore. This way, the results were much more readable and insightful. The first graph (*Figure 18*) gives us the Average Recency (‘MostRecentView’), being the average number of days between each review of the same user. We can perceive that **the bigger the Average Recency, the worst the RScore** since it is better to have smaller periods of time between each review. The second figure (*Figure 19*) shows us the average Frequency for each Fscore. Contrarily to the Average Recency, we want this value to be 4, as we can see the more contributions a user makes the better the FScore is going to be.

Finally, as we previously saw, the MScore has no users with 3 and 4 values. This is given to the fact that the rating has a limited value between 1 and 5, meaning that there is no space to separate the values for each quartile of the RFM model, since the average lowest value is 3.48 and the highest average value is 4.96. However, we can perceive that the higher the Average Rating, the better for the MScore.

The following figure (*Figure 21*) shows the treemap we built to get the Total Number of Users per Segment. As we can see, the **Bronze segment has the most users** with a total of approximately 48 thousand users, the Silver segment totals around 6.6 thousand users and the Gold around 2.8 thousands of users. This not unusual, as RFM segments typically follow a pyramid structure with the majority of users falling into lower-value segments.

5. MODELING

5.1.2. RFM FOR PORTUGUESE USERS

Moving on to RFM for Portuguese Users, first we first created filtered dataset where we stored only users that have their “UserCountry” identified as Portugal. Then we started the process of creating the RFM Model, and we started by the Recency, which we calculated by the number of days since each user's most recent review, and rename it “MostRecentView”. Next, we analyzed Frequency, by extracting how many contributions a user has made, and rename it “TotalContributions”. And last we did Monetary by averaging their ratings and save the values in the ‘AverageRating’ column. The quantiles (25th, 50th, and 75th percentiles) for each metric were calculated and we assigned a score from 1 to 4 for each dimension based on where their values fall within these ranges. Finally, we combined the scores resulting in a RFMScore, sorting users based on this score.

The histograms generated illustrate the distribution of Recency, Frequency, and Monetary scores across all users. Our first graph, *Figure 22,* show us Average Recency, where we can observe that **the bigger the average recency, the worst the RScore will be**, which corresponds to shorter periods between reviews being an indicator of engagement.

The second graph, *Figure 23,* shows us the relationship between FScore and Average Frequency. From here we can see that the **rows with a higher FScore have a much bigger force in the contribution aspect**, evidenced by the high value in Score 4.

The final graph, *Figure 24,* shows us the MScore aplicated to the Average Frequency. Just like in the previous RFM Model, the MScore doesn't apply to the 3 and 4 values because of the scale of 1-5, that limitates our input. Asides that, the only valuable infromation we can take from the graph is that **the bigger our average rating the better our MScore will be**.

By directly comparing the results of the graphs, we can easily observe that they are **almost identical to the results we got for the RFM for all users**, evidencing that the behavior of two groups are not distinct from each other. This is even more evident when we put the data into an treemap, as exemplifeid in *Figure 25.*

5. MODELING

5.1.3. RFM FOR PORTUGUESE ATTRACTIONS

The Average Recency by RScore graphic represents the time between each review for each Portuguese attraction. This metric is calculated by grouping the attractions by their RScore and calculating the average recency (time between reviews). As seen in the *figure 26*, a longer gap between reviews corresponds to a lower RScore performance, suggesting that more frequent reviews lead to better results in terms of the recency score.

The Average Frequency by FScore represents the total number of reviews each attraction has received, grouped by its FScore. The data indicates that **attractions with a higher number of reviews tend to have a better FScore**, suggesting that more frequent reviews contribute to a higher frequency score.

Finally, the Average Monetary by MScore, which is measured by the reviewRating. This metric is calculated by grouping the attractions by their MScore and calculating the average rating. It shows that a higher average rating tends to correspond with a higher MScore, implying that better-rated attractions are associated with a higher monetary score.

The treemap visualization *(Figure 29*) highlights the distribution of Portuguese attractions across two segments being **six attractions in the Bronze** one and **one attraction in the Gold** (Quinta da Regaleira). The gold segment indicates that the attraction is consistently well-reviewed, attracts a frequent number of visitors, and receives high ratings, positioning it as one of the more elite attractions. On the other hand, the Bronze, while having a greater number of attractions, represents those that have room for improvement in one or more of the RFM criteria, such as lower visitor engagement or less frequent reviews.

5. MODELING

5.1.4. RFM BY PORTUGAL VISITS (COUNTRY OF ORIGIN)

As for our last use of the RFM Model, we will use it to analyze the users by Country of origin that have visited Portugal.
We started by organizing the country's names and correct wrong or repeated values. On an initial check on the RFM Data we saw that our FScore was chosing sines of skew. Analyzing it in more detail we found that the distribution of the data revealed an uneven spread. To resolve this problem we redefine the bins to make them reflect the data range and distribution so that it can better representative all the distintc groups. After this proceeder we started the analysis.

Examining the RScore by country, *Figure 30,* offers a more detailed perspective, as it measures the interval between visits. **Over a quarter of the countries show strong performance**, indicating a consistent pattern of frequent visits to Portugal. This analysis provides valuable insight into visitor trends, revealing the nations with a notable affinity for returning to Portugal regularly.
Now analyzing the FScore for the frequency of the visits, *Figure 31,* we can observe a significant portion of countries falling into the higher categories, showing us that they pose as consistent visitors.

On the MScore analysis, on *Figure 32,* show´s to us the great monetary value that 3 and 4 represent to Portugal, while also demonstrating that 1 and 2 can pontencial be important to the country given their similarity whtih the other two.

When segmenting the results by RFMScore, we observed on *Figure 33* that all countries fall into the Bronze segment. It is important to emphasize that this classification does not indicate that visitors from Portugal are undesirable customers. Rather, it highlights that the metrics used for Frequency and Monetary aspects do not provide meaningful differentiation when analyzed at the user Country level.

5. MODELING

5.2. ASSOCIATION RULES

Associations Rules allow to unveil the frequency of hidden patterns and relationships between variables in large datasets. For this project it was decided to apply the Apriori algorithm, one of the most frequently used algorithms in l arge datasets. In this one the bottom- up and delete relabel approach was applied to ensure that is uncover frequent patterns, eliminating infrequent item sets from consideration. In the algorithm it was analyzed the relationships between the attractions for how frequently visited together by tourists in Portugal and the most common countries of origin of Portugal's visitors.

5.2.1. ASSOCIATION RULES BY ATTRACTIONS

To understand what tourists, prefer to visit and if there is a pattern in the other that they prefer to visit the Portuguese touristic places, it was analyzed the associations between attractions.

Through this process, three tables were created to evaluate the support, confidence, and lift relationships to create associations that will give valuable insights to the Portuguese National Tourism Board Office (NTBO). This analysis started by creating a reviews pivot table. For each row the index was represented by the “userName”, each attraction by the “LocalID”, if there wasn’t there wasn’t a relationship between the tourists with the countries it was values a 0. If there was an association it would show 1. This table allowed to start the **Apriori Algorithm** with a cleaner database for evaluation. In the support table, a threshold of 0.01 was established, with 10 item sets, implying that these associations would only need to happen once to be valuable for this analysis. For the confidence and lift tables, the threshold was modified to 0.05%. Despite this change, the same 4 relationships remained for evaluation presenting enough frequency, likeliness to occur and probability to be considered strong associations to evaluate.

As can be seen in *Figure 34*, where it presents the 3 tables, when MAG014 as the antecedent and MAG010 as consequent, and MAG010 as the antecedent and MAG014 as consequent the support and lift values are the same. The same thing happens to the MAG032 and MAG047 relationships. However, the antecedent, consequent support, and confidence level are different, having an impact on how the values are ranked for each table.

Before proceeding with this explanation, it's important to establish the names for these LocalIDs:
MAG010: Torre de Belém
MAG014: Mosteiro dos Jeronimos
MAG032: Palacio da Pena
MAG047: Quinta da Regaleira

The support table shows a 9.97% possibility of frequency between the MAG014 and MAG010 and a 6.34% possibility of frequency for the MAG032 and MAG047 relationship. Regarding the confidence level, all the outputs present different numbers showing a higher likeliness of occurrence for the {MAG047} -> {MAG032} relationship and the {MAG014} -> {MAG010} one. It's important to note that despite the antecedent values, if the consequent happens often, the confidence level for the rule will naturally be high, regardless of what happens in the antecedent, which can be verified on the {MAG047} -> {MAG032} relationship. Lastly, the Lift table reveals that the highest probability of it happening, with the value of 1.71 is: **if a tourist visits Palacio da Pena, they are most likely to visit Quinta da Regaleira attraction as well**. Nevertheless, since all the relationships present a value higher than 1, there is a strong association that when visiting one of these attractions, tourists will visit the other one.

From this study we were able to understand that tourists have a preference to visit these 4 attractions in Portugal, comparing to other ones Ponte de Dom Luís I, Bom Jesus do Monte and Cais da Ribeira. Nonetheless, the **level of proximity** between the relationships and **publicity** created can be considered the strongest reasons to present stronger associations.

5.2.2. ASSOCIATION RULES BY COUNTRIES

To have an understanding which were the attractions between the places that each country enjoys traveling too it was analyzed the associations between countries. The coding process was executed like it was done for the associations for attractions. It was created a pivot table tourists_pivot_table. Each row represented a user location, column the country of attraction. If the values presented were, then there was at least one review from the user location for a country, if here wasn't an association then it was presented the number 0.

For the **Apriori Algorithm,** it was done with a minimum support of 0.6. Regarding the minimum threshold, it was increased from support, with 0.5, confidence increased to 0.20, and lift to 1.20, this allowed us to have a greater understanding of the strength between these connections.

As shown in *Figure 35*, for the relationships observed between the same countries, the strength and lift present the same values. On the other hand, the antecedent, consequent support, and confidence level are different. For support the relationship between Spain and England has the greatest frequency, appearing in the dataset with a value of 0.10. In terms of confidence, the strongest association was 0.637 when flying from Russia to Spain, followed by flying from Austria to Spain. In the lift table, the strongest association between countries was between flying from Spain to Russia, followed by Russia to Spain, with a value of 1.82. All the relationships presented show a value over **1.20** showing a significant and valued relationship to take into consideration and evaluation.

Overall, Spain stands out as a country of preference for travel. Russia is still present in these tables as one of the countries of preference for flying to or from; however, due to the active war between Russia and Ukraine, the country is not a favorable place to travel to or in an economic situation for tourism. Lastly, comparing to our main competitors, Spain, Italy and Greece, Portugal was considered the second country with the greatest amount of associations.

6. HYPOTHESIS TESTING

To conclude this study a hypothesis testing was created to evaluate the number of reviews and visits presented for the 4 seasons in Portugal, by performing an A/ B test, more specifically the Kruskal- Walli's test, known to compare independent samples for sample sizes. In this one, it was evaluated the number of visits per season to perform the test. However, the p-value corresponded to approximately 0.392. Consequently, it implied that throughout the seasons **there wasn’t a significant difference** in the number of written reviews that could present evidence to reject the null hypothesis.

7. DEPLOYMENT

In this final step of the data mining process, this part will focus on providing marketing suggestions and business benefits that can be achieved in the long run. These will 4 suggestions will allow NTBO to solidify Portugal’s position as a leading travel destination. After implementing these strategies, it’s important to be aware that this is a **cyclical process**, implying that these data mining results will most likely suffer changes. Thus, there is a need to have a consistent evaluation of the new numbers, updating them, and doing new assessments. A steady analysis will allow you to understand if the strategy implemented is working and have a greater understanding of consumer behavior.

7.1. RECOMMENDATIONS FOR PORTUGAL

Throughout this research, key factors were studied that allowed us to have a greater understanding of the decision-making process when choosing a country to spend holidays. Thus, having a solo focus on how we can improve the tourism experience in Portugal and incentivize more people to decide to visit the country was designed with marketing ideas based on the research done.

Here are the 4 marketing suggestions to encourage people to choose Portugal as their election country to spend their holidays:

1. **Create Attraction Itineraries:** It was seen that people have a greater preference to visit attractions that are closer to each other. Thus, it would be pertinent to create Itineraries based on the distance from each attraction. Another option, to make it more fun it could be created a “Lisbon Bucket List Challenge”. This guidance will allow tourists to visit several attractions in a day in an order focused on how close those are from each other. By the end of the day, they will have visited several places, the first one being extremely far from the last attraction, however, since they were always visiting new places, it doesn't give them that sense. Moreover, when deciding where to travel to, having this Itinerary already created and displayed can incentivize during the decision-making process, since people won't have to do extensive research on places to visit. This one could be prepared for tourists that only 1 day to tourist, 2 days and 3.
2. **Incentives and rewards to have more reviews:** From the study, it was noticed a significant decrease in the willingness for people to write reviews regarding their holidays, post-COVID. Complementing the previous idea, to increase these numbers, incentives could be created for people to write more reviews. As people will follow the Portuguese attraction guide, that one could also present coffee places, restaurants, and other interactive places (not considered to be monuments) for tourists to go to. To increase the number of reviews, the challenge would be presented, such as "Have 10% off your tickets to visit the Quake, the Lisbon earthquake experience! If you visit the Mosteiros dos Jerónimos, write a review on the website and show it when purchasing the tickets”. This type of opportunity would appear several times. Not only would incentivize people to write more reviews, but would also to visit more places, such as Portuguese small businesses.
3. **Personalized Itinerary:** Following the two ideas provided, it could be created Itineraries for each season of the year. For each, it should be emphasized the best activities to do and places to visit during that time of the year. This will allow people to understand all the possibilities that Portugal has to offer throughout the year. Also, depending on the traveler profiles, itineraries can be created with a solo focus for families, and solo or work-trip travelers, to allow a wider audience appeal.

4. **Promote Portugal's Winter Sunshine:** Portugal is a country that is fortunate to have good weather all year around. Thus, to boost winter travel numbers, this advantage needs to be advertised more frequently. The National Tourism Board Organizations (NTBO) should create specific advertising campaigns reflecting how the good weather should be a key-factor in the process of selecting which country to spend the winter holidays in.

8. CONCLUSION

In conclusion, the comprehensive analysis present in this report offers actionable insights and recommendations tailored for the Portuguese National Tourism Board Office (NTBO). By using the CRISP-DM methodology, a six-staged framework that comprises business understanding, data understanding, data preparation, modeling, evaluation, and deployment, our analysis explored tourist behaviors and attraction preferences in Portugal during the period from January 1st, 2019 to December 31st, 2021.

Portugal's tourism sector has shown promising signs of recovery despite the challenges posed by the COVID-19 pandemic, since the most recent available data reveals that indicators, such as overnight stays and revenue, have surpassed pre-pandemic levels. The country’s recovery is also supported by stronger engagement from markets like the UK and USA, as tourists from these countries are visiting more and more, and at the same time, Portugal’s popular attractions keep appealing them, boosting this recovery.

Additionally, the use of techniques such as RFM analysis and Association Rules enabled us to identify key patterns that support the NTBO in post-pandemic recovery efforts. By employing these techniques, we have identified key visitor segments and preferences, which provided helpful advice for targeted marketing efforts. Attractions such as Quinta da Regaleira, Torre de Belém, and Mosteiro dos Jerónimos remain crucial to Portugal’s appeal, and marketing based on these key locations would enhance visitors’ experience.

Lastly, this report emphasises the importance of data-driven decision-making in revitalizing Portugal’s tourism industry. The country’s unique cultural resources offer significant opportunities for growth, so Portugal can continue to strengthen its position as a **leading travel destination in Europe**. With a focused approach, the NTBO can improve its strategies to align with evolving tourist expectations, ensuring sustained long-term success and growth.

APPENDIX

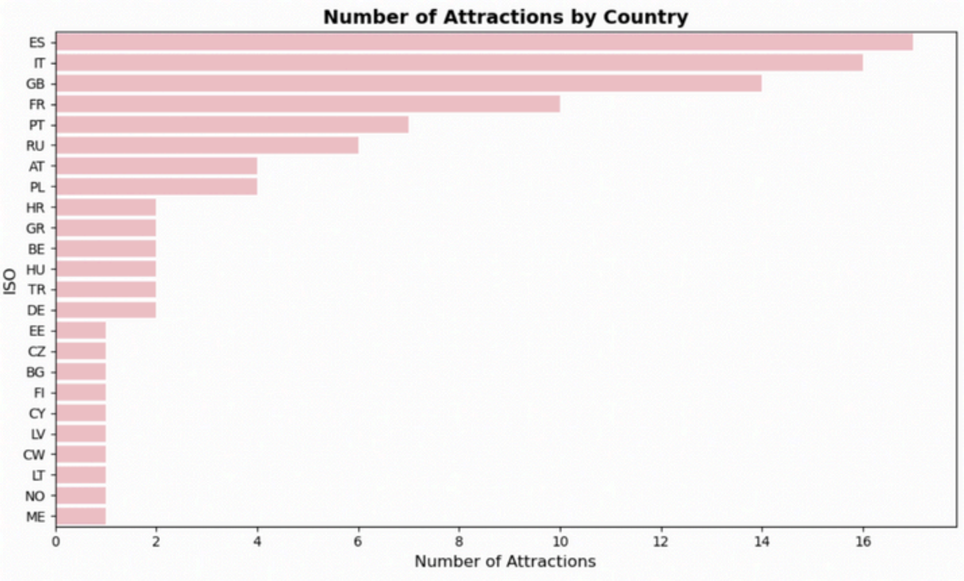
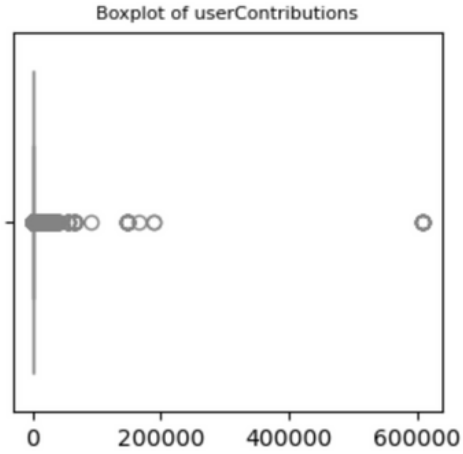
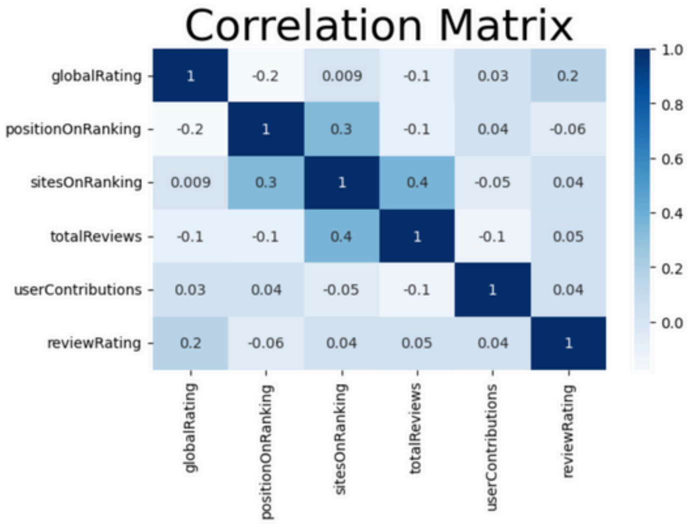


Figure 4: Number of Attractions by Country





	globalRating	positionOnRanking	sitesOnRanking	totalReviews	userContributions	reviewRating
globalRating	1	-0.2	0.009	-0.1	0.03	0.2
positionOnRanking	-0.2	1	0.3	-0.1	0.04	-0.06
sitesOnRanking	0.009	0.3	1	0.4	-0.05	0.04
totalReviews	-0.1	-0.1	0.4	1	-0.1	0.05
userContributions	0.03	0.04	-0.05	-0.1	1	0.04
reviewRating	0.2	-0.06	0.04	0.05	0.04	1

Figure 5: Boxplot of usercontributions

Figure 6: Correlation Matrix of Reviews

APPENDIX

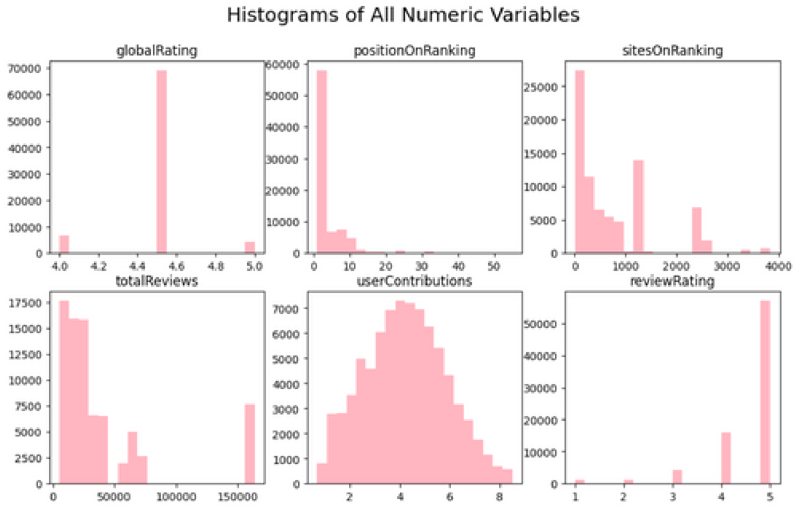
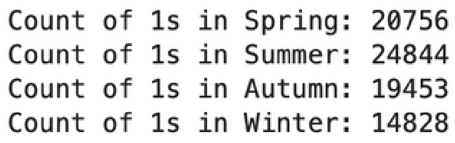


Figure 7: Histogram off all the numeric variables



```
Count of 1s in Spring: 20756
Count of 1s in Summer: 24844
Count of 1s in Autumn: 19453
Count of 1s in Winter: 14828
```

Figure 8: Visits by season of the year

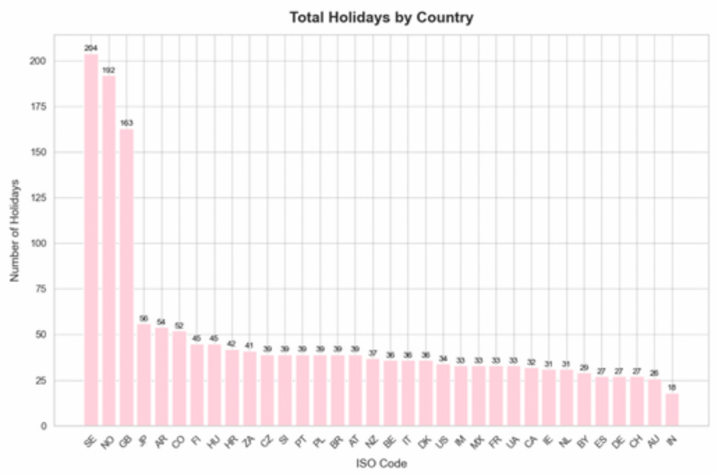


Figure 9: Total Holidays by Country

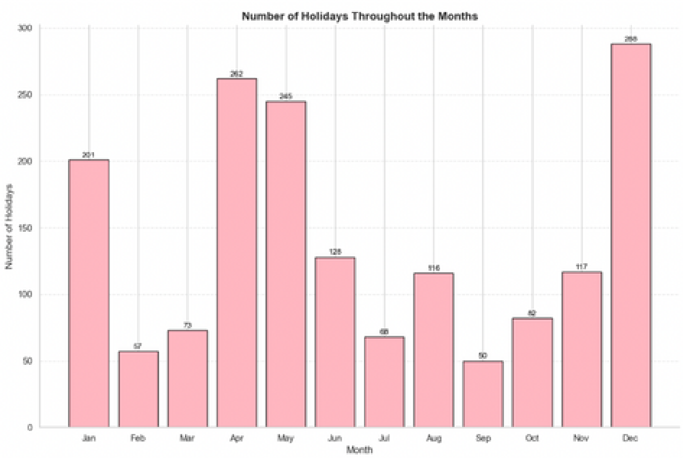
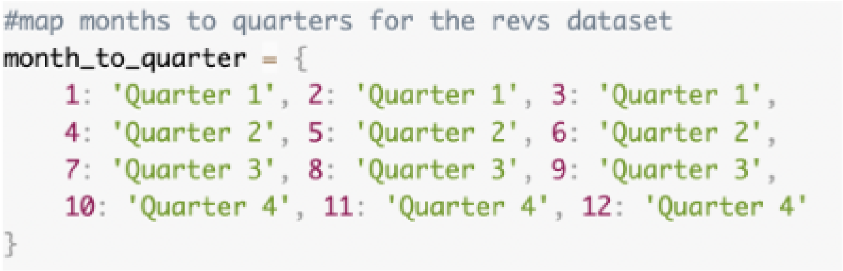


Figure 10: Total Holidays by Month

APPENDIX



```
#map months to quarters for the revs dataset
month_to_quarter = {
    1: 'Quarter 1', 2: 'Quarter 1', 3: 'Quarter 1',
    4: 'Quarter 2', 5: 'Quarter 2', 6: 'Quarter 2',
    7: 'Quarter 3', 8: 'Quarter 3', 9: 'Quarter 3',
    10: 'Quarter 4', 11: 'Quarter 4', 12: 'Quarter 4'
}
```

Figure 11: Months to quarters for the Review's dataset

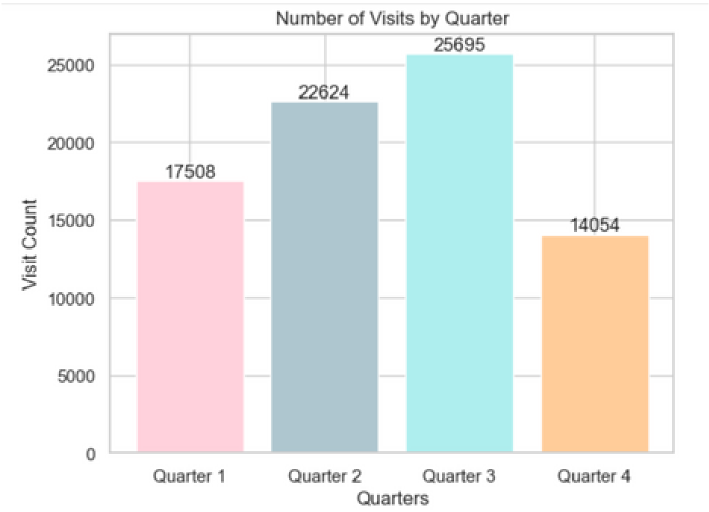


Figure 12: Number of Visits by Quarter

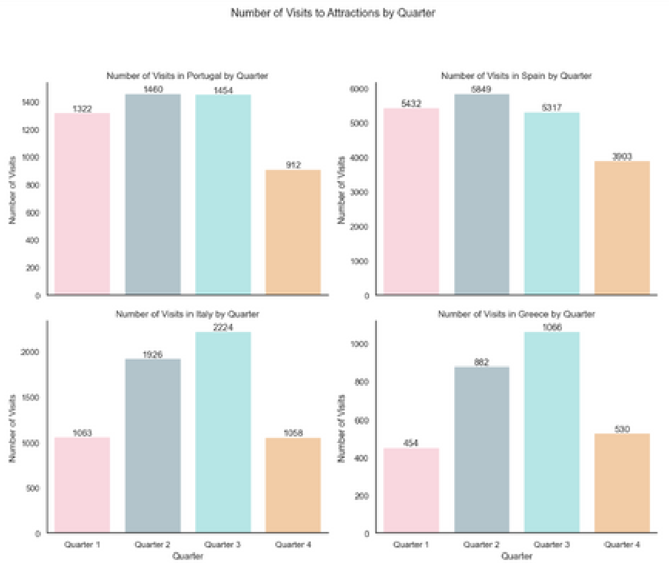


Figure 13: Number of Visits by Quarter and by Portugal's Competitors

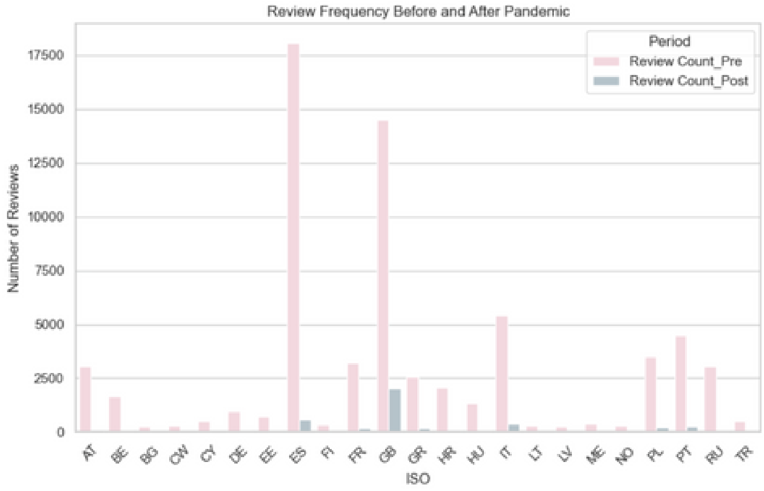


Figure 14: Review Frequency Before and After Pandemic

APPENDIX

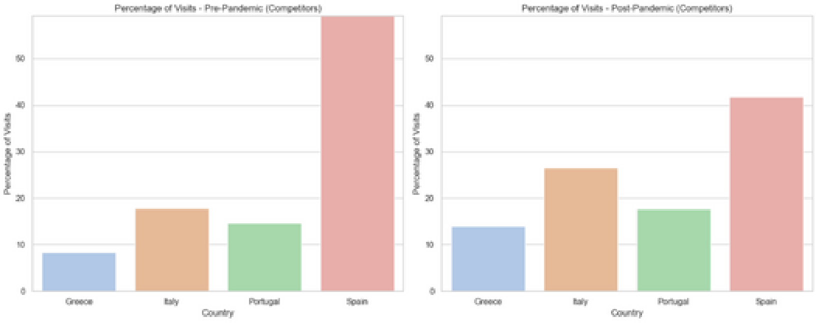


Figure 15: Percentage of Visits Pre and Post Pandemic of Portugal and its competitors

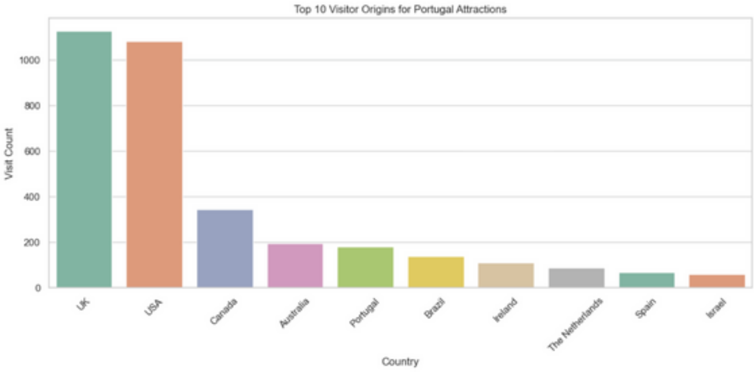


Figure 16: Top Visitor Origins for Portugal Attractions

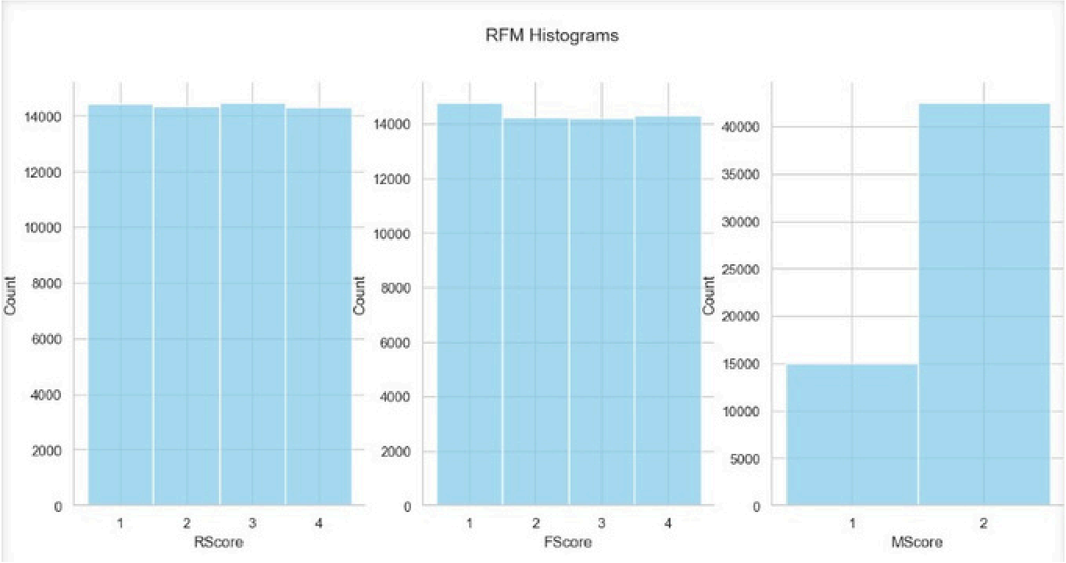


Figure 17: RFM Histograms of RScore, FScore and Score

APPENDIX

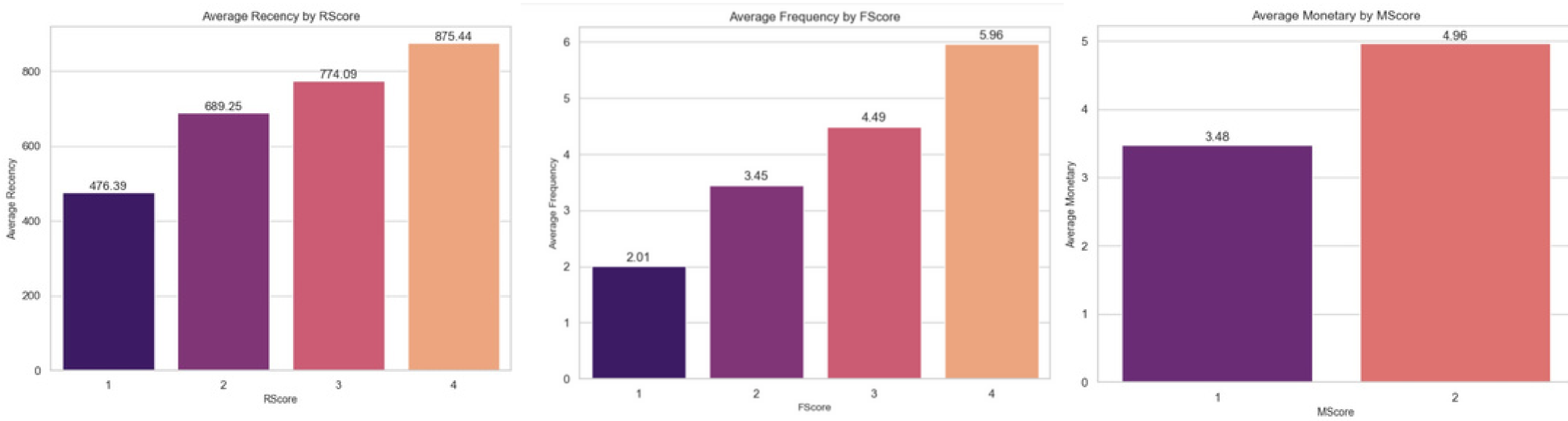


Figure 18, 19 and 20: RFM for All Users RScore, FScore and MScore, respectively

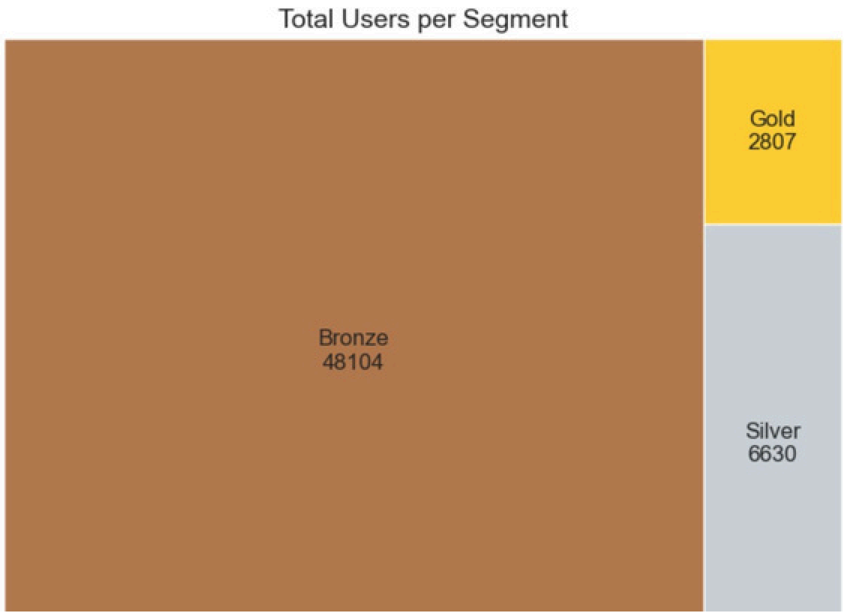


Figure 21: Treemap With Total Users per Segment

APPENDIX

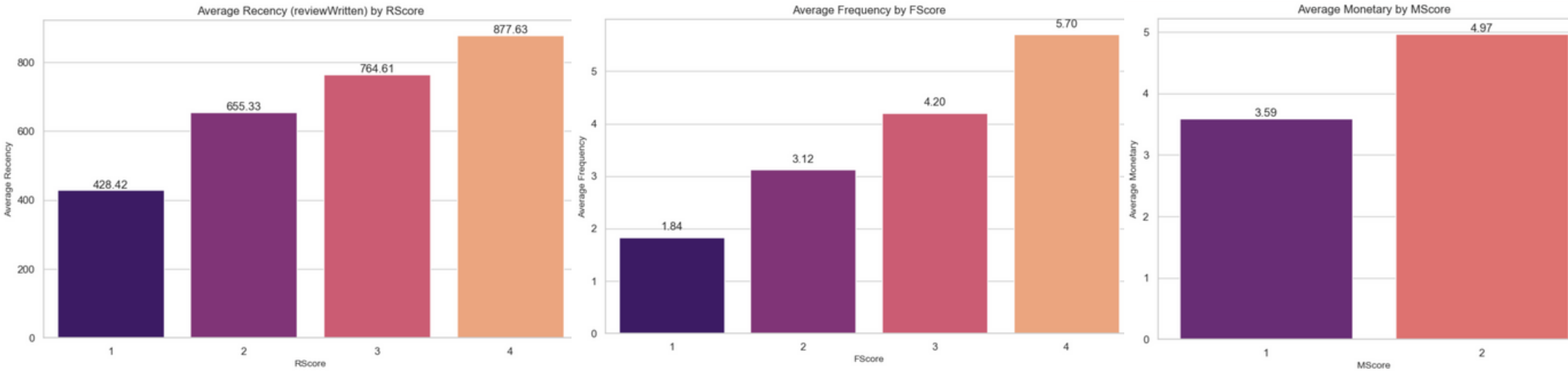


Figure 22, 23 and 24: RFM for Portuguese Users RScore, FScore and MScore, respectively

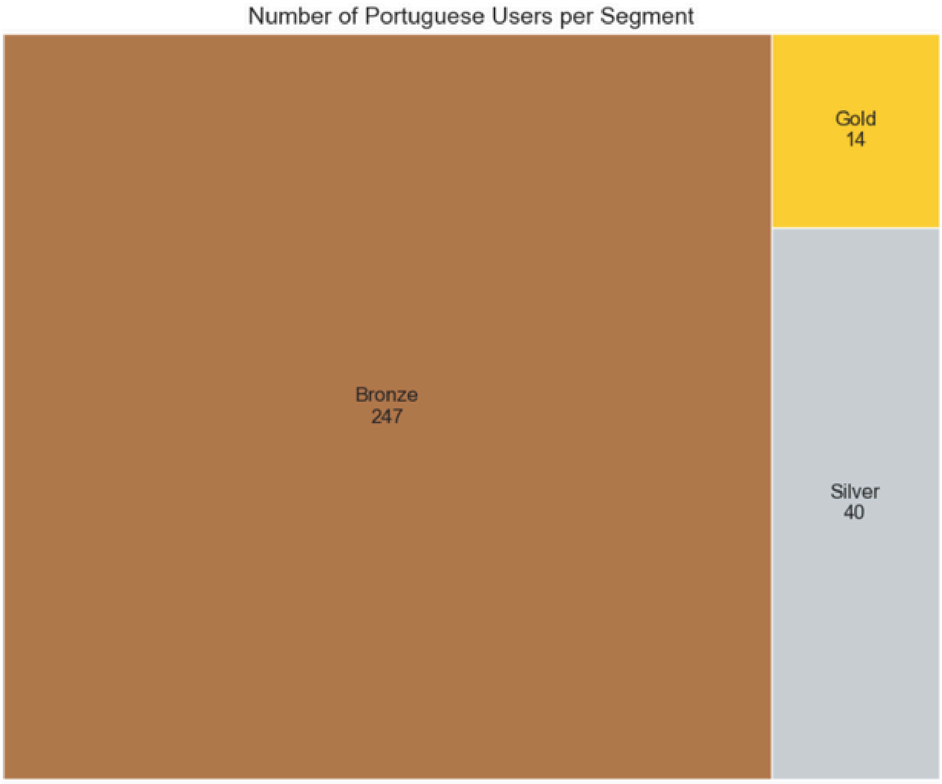


Figure 25: Treemap With Portuguese Users per Segment

APPENDIX

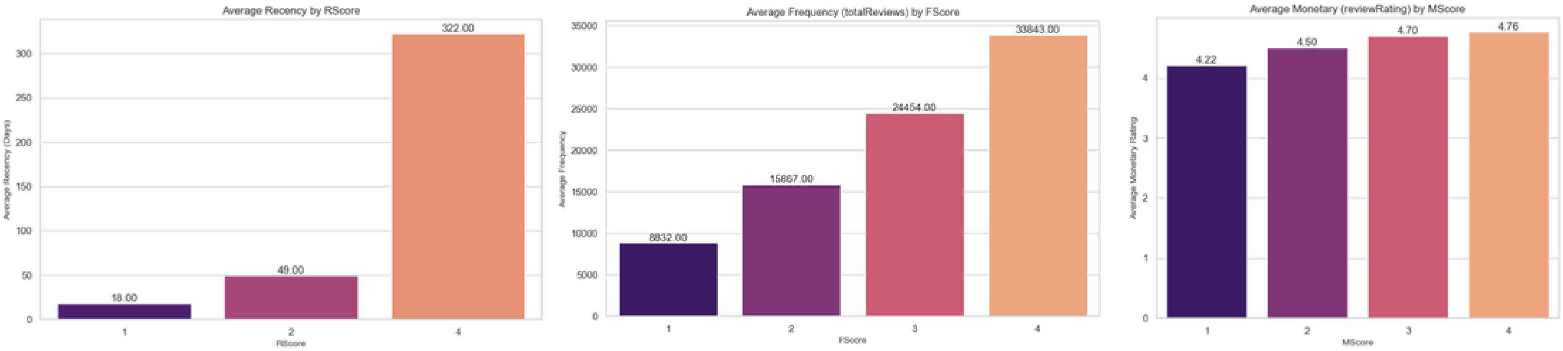


Figure 26, 27 and 28: RFM for Portuguese Attractions RScore, FScore and MScore, respectively

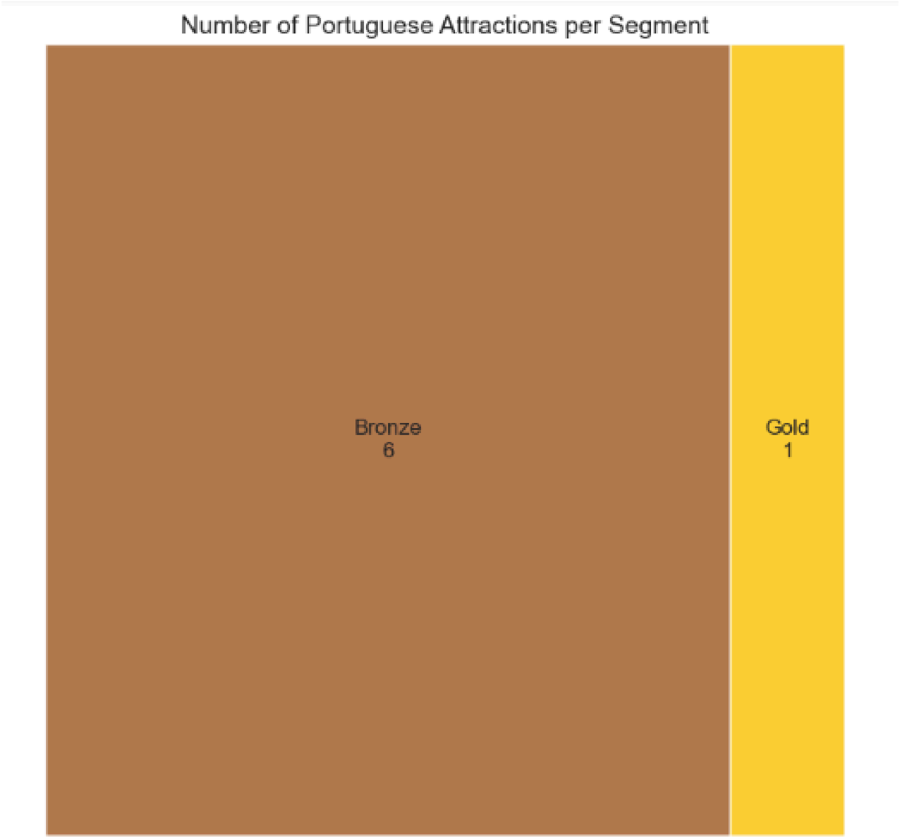


Figure 29: Treemap With Portuguese Attractions per Segment

APPENDIX

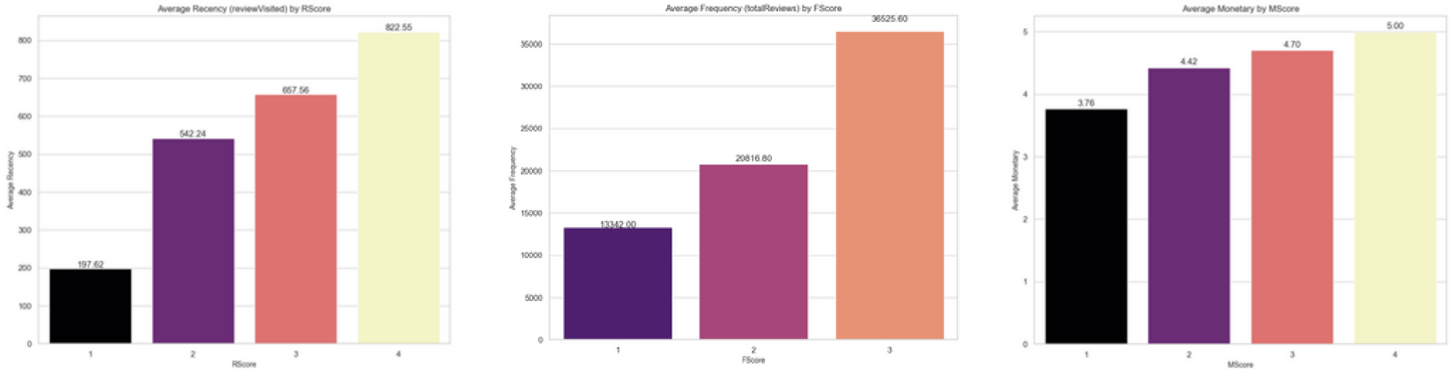


Figure 30, 31 and 32: RFM for Portugal Visitors RScore, FScore and MScore, respectively

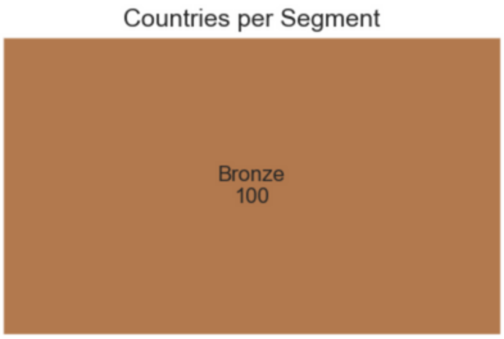
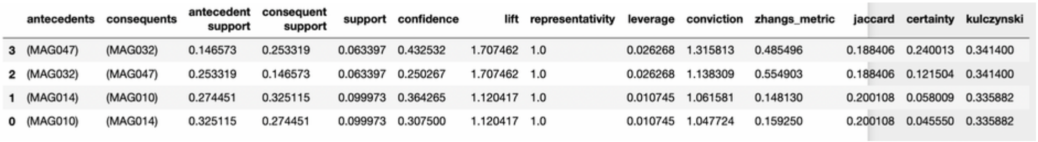
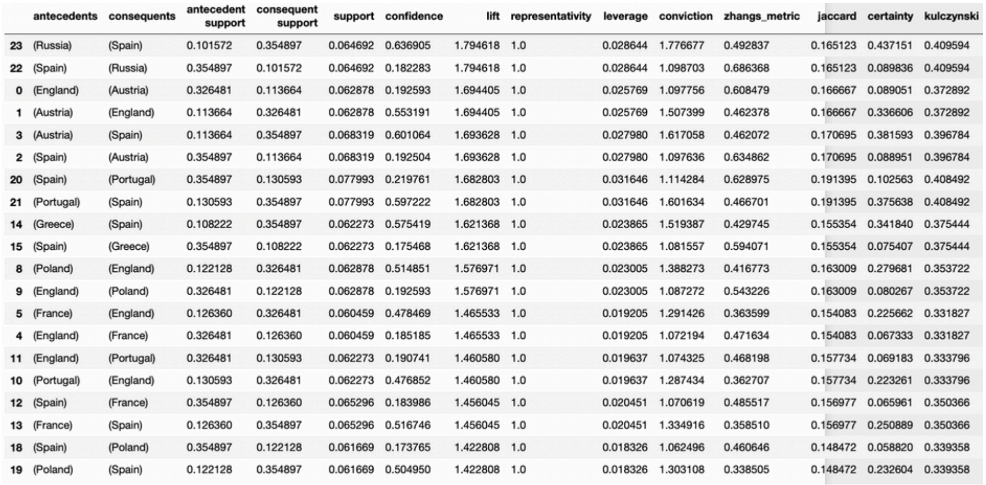


Figure 33: Treemap With Portugal Visitors per Segment



	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	representativity	leverage	conviction	zhangs_metric	jaccard	certainty	kulczynski
3	(MAG047)	(MAG032)	0.146573	0.253319	0.063397	0.432532	1.707462	1.0	0.026268	1.315813	0.485496	0.188406	0.240013	0.341400
2	(MAG032)	(MAG047)	0.253319	0.146573	0.063397	0.250267	1.707462	1.0	0.026268	1.138309	0.554903	0.188406	0.121504	0.341400
1	(MAG014)	(MAG010)	0.274451	0.325115	0.099973	0.364265	1.120417	1.0	0.010745	1.061581	0.148130	0.200108	0.058009	0.335882
0	(MAG010)	(MAG014)	0.325115	0.274451	0.099973	0.307500	1.120417	1.0	0.010745	1.047724	0.159250	0.200108	0.045550	0.335882

Figure 34: Associations Rules for Attractions Table by Lift



	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	representativity	leverage	conviction	zhangs_metric	jaccard	certainty	kulczynski
23	(Russia)	(Spain)	0.101572	0.354897	0.064692	0.636905	1.794618	1.0	0.028644	1.776677	0.492837	0.165123	0.437151	0.409594
22	(Spain)	(Russia)	0.354897	0.101572	0.064692	0.182283	1.794618	1.0	0.028644	1.098703	0.686368	0.165123	0.089836	0.409594
0	(England)	(Austria)	0.326481	0.113664	0.062878	0.192593	1.694405	1.0	0.025769	1.097756	0.608479	0.166667	0.089051	0.372892
1	(Austria)	(England)	0.113664	0.326481	0.062878	0.553191	1.694405	1.0	0.025769	1.507399	0.462378	0.166667	0.336606	0.372892
3	(Austria)	(Spain)	0.113664	0.354897	0.068319	0.601064	1.693628	1.0	0.027980	1.617058	0.462072	0.170695	0.381593	0.396784
2	(Spain)	(Austria)	0.354897	0.113664	0.068319	0.192504	1.693628	1.0	0.027980	1.097636	0.634862	0.170695	0.088951	0.396784
20	(Spain)	(Portugal)	0.354897	0.130593	0.077993	0.219761	1.682803	1.0	0.031646	1.114284	0.628975	0.191395	0.102563	0.408492
21	(Portugal)	(Spain)	0.130593	0.354897	0.077993	0.597222	1.682803	1.0	0.031646	1.601634	0.466701	0.191395	0.375638	0.408492
14	(Greece)	(Spain)	0.108222	0.354897	0.062273	0.575419	1.621368	1.0	0.023865	1.519387	0.429745	0.155354	0.341840	0.375444
15	(Spain)	(Greece)	0.354897	0.108222	0.062273	0.175468	1.621368	1.0	0.023865	1.081557	0.594071	0.155354	0.075407	0.375444
8	(Poland)	(England)	0.122128	0.326481	0.062878	0.514851	1.576971	1.0	0.023005	1.388273	0.416773	0.163009	0.279681	0.353722
9	(England)	(Poland)	0.326481	0.122128	0.062878	0.192593	1.576971	1.0	0.023005	1.087272	0.543226	0.163009	0.080267	0.353722
5	(France)	(England)	0.126360	0.326481	0.060459	0.478469	1.465533	1.0	0.019205	1.291426	0.363599	0.154083	0.225662	0.331827
4	(England)	(France)	0.326481	0.126360	0.060459	0.185185	1.465533	1.0	0.019205	1.072194	0.471634	0.154083	0.067333	0.331827
11	(England)	(Portugal)	0.326481	0.130593	0.062273	0.190741	1.460580	1.0	0.019637	1.074325	0.468198	0.157734	0.069183	0.333796
10	(Portugal)	(England)	0.130593	0.326481	0.062273	0.476852	1.460580	1.0	0.019637	1.287434	0.362707	0.157734	0.223261	0.333796
12	(Spain)	(France)	0.354897	0.126360	0.065296	0.183986	1.456045	1.0	0.020451	1.070619	0.485517	0.156977	0.065961	0.350366
13	(France)	(Spain)	0.126360	0.354897	0.065296	0.516746	1.456045	1.0	0.020451	1.334916	0.358510	0.156977	0.250889	0.350366
18	(Spain)	(Poland)	0.354897	0.122128	0.061669	0.173765	1.422808	1.0	0.018326	1.062496	0.460646	0.148472	0.058820	0.339358
19	(Poland)	(Spain)	0.122128	0.354897	0.061669	0.504950	1.422808	1.0	0.018326	1.303108	0.338505	0.148472	0.232604	0.339358

Figure 35: Associations Rules for Countries Table by Lift

RESOURCES

Bruxo, Michael. (2024, August 27). American tourism boom in Portugal continues in 2024. Portugal Resident. Retrieved from https://www.portugalresident.com/american-tourism-boom-in-portugal-continues-in-2024/

Lusa. (2024, November 29). Canadians boosting autumn tourism. The Portugal News. Retrieved from Bruxo, Michael. (2024, August 27). American tourism boom in Portugal continues in 2024. Portugal Resident. Retrieved from https://www.portugalresident.com/american-tourism-boom-in-portugal-continues-in-2024/

Turismo de Portugal. (2024, November 4). Tourism performance. Business Turismo de Portugal. Retrieved from https://business.turismodeportugal.pt/en/Conhecer/Apresentacao/Desempenho_Turistico/Pages/default.aspx

Statista. (2024, April). Number of international visitors and overnight stays in Portugal from 2006 to 2021. Retrieved from https://www.statista.com/statistics/398360/number-of-international-visitors-and-overnight-stays-in-portugal/

Statista. (2024, June). Travel and tourism's total contribution to GDP in Portugal from 2012 to 2022. Retrieved from https://www.statista.com/statistics/770057/travel-and-tourism-s-total-contribution-to-gdp-in-portugal/

https://chatgpt.com