National Tourism Promotion in Portugal

- Data Science Group project

(Group 29)

Matej Federic (m20210118)
Abdul Ghaffar (m20190690)
Johar Shakil Yousuf (m20210753)

Foreword

This document was prepared as a supplement to the group project using the CRISP-DM model to find solutions to the problems faced by the Portuguese tourism sector. The project follows CRISP-DM methodology and contains several key sub-chapters with extra citations and information.

Report background

Our team, acting as business consultants to the Portuguese National Tourism Board Organization (NTBO), was asked to understand visitors' patterns and analyze if a pandemic has played any role in those patterns. Moreover, we were asked to identify, on the basis of data, key competitors of the Portuguese tourism sector in Europe so as to successfully set the Portuguese tourism sector against them and attract tourists to Portugal. The finding will help us in devising right marketing strategies to revive growth in the tourism sector to the pre-pandemic levels, at least.

The report contains all the reasoning and description of the CRISP-DM model that helped our team to the conclusion of the analysis.

Table of content

BUSINESS UNDERSTANDING

One of the most critical factors in the successful data analysis projects is understanding the business objective of the project. As rightly mentioned by Abbott (2015), “A possible consequence of neglecting this step is to expend a great deal of effort producing the right answers to the wrong questions”. In this case, our team put effort into understanding the business case of the course project before other steps of the CRISP-DM model.

It should be kept in mind, however, that the team was not given much information/consultation with the Portuguese National Tourism Board Organization (NTBO) therefore, many of the business cases are the hypotheses developed by the team itself. Team did make an effort, though, by consulting external research, official tourism reports & strategies and other important literature.

Determine Business objectives

Background

Our client is the Portuguese National Tourism Board Organization (NTBO), which is based in Lisbon Portugal. The Board is responsible for promoting Portugal as a tourism attraction in the world and is mandated to devise strategies for branding Portugal as the best destination for tourism.

Although the team has not been provided with complete information as what are the specific goals of the Board for reviving tourism in the country yet we have hypothesized many of the objectives, partly, from our own understanding and, partly, from reading recent reports and official press briefings on reviving tourism in the country.

Business Objectives

Primary business objective of the National Tourism Board Organization (NTBO) is to bring back tourism in the country to the pre-pandemic level. The Board will share the findings of the study with the marketing department to devise data-driven marketing strategies to attract tourists to the country.

Although the team was not given much information from the client yet the team empathized, and put themselves in the client’s shoes, to best guess the key business questions and thus, derive business objectives for the study. As a team responsible for the economic growth, through tourism we would set our business objectives as following:

- Revive tourism back to where it was at its best
- Identify key factors/patterns that tourists deem important in the regions they visit
- What is it that we can learn from and compare to our competitors especially their post-pandemic tactics

We would shed more light on the business objectives in the data exploration section of this project report.

Business Success Criteria

The challenging part of the project is to set business success criteria or KPIs to satisfy the needs of the client because the team does not know precisely the business objectives of the study at hand. Due to this shortcoming, the team put together some of the KPIs that will resemble the business

success criteria for the client and help them with their business endeavors.The success criteria may include the following:

- Increase the tourism in Portugal back to pre-pandemic levels
- Set Portugal as a more competitive tourist destination in Europe
- Revive employment levels in other sectors that depend on tourism such as; restaurants & hotels, transportation, and others

Assess Situation

Inventory of Resources

There are several dimensions of resources necessary for conducting a business project. This may include human resources, information resources e.g; data resources, and other resources such as computational & software resources. Below we have briefly described each type of resources that are required for the successful completion of the project:

Human Resource: Our team (Matej, Abdul, Johar) who carry important tasks in the project + NOVA IMS instructors. No resource was hired from the National Tourism Board Organization.

Information Resource: Some of the data for the project was provided by the Board while other data was retrieved from online tourism websites such as; booking.com and others.

Computational Resources: Apart from personal computers no other hardware was required and used during the project.

Software Resources: Team used Python and MS Excel were used for the analysis and completion of the project. Resources did not require any additional costs for the team.

Requirements, Assumptions & Constraints

An important requirement for the project is that it must be completed within the stipulated time of the course and needs to be delivered before the end of the academic semester.

Another key requirement is that the model must be developed and run several times to test the validity of the results. In other words, the models need to be run easily and maintain credibility if and when used by the stakeholders of the project.

The team made the assumption that the results generated will be easy to comprehend by all stakeholders involved.

There are no foreseeable constraints and the team expects there will be no additional costs in running the model.

Risks & Contingencies

Although the team will try to minimize and avoid all types of risks involved in the project however, there are some of the risks that might arise in the future. Following are the potential risks, and contingency plans to deal with the risk, if they arose:

Business/organization risk: This is the risk that may arise due to incomplete information while setting the business case for the study. Setting wrong/improper questions, in the absence of key stakeholders, will result in setting the study in the wrong direction, and therefore, no use of the results for the client.

Marketing and Financial risk: The client may face marketing and financial risks if the results and recommendations of the study are implemented as it is.

Technical risk: There are problems associated in transferring data from one computer to another, from one software to another and from one format to another. The unwanted problems may arise because data will be transferred several times between team members.

Outcome risk: Data science models seldom work perfectly when the subset that is provided is not fully representative.

Terminology

The report does not contain a glossary of terminology as the team will present it to subject experts on the matter and therefore does not require a glossary of terminology. Had the audience/readers been laymen the team would have definitely provided them with the glossary of important terminology so as to make it easy to comprehend by them.

Costs & Benefits

The study is limited to finding the key patterns and factors that are considered important by tourists in any tourist area and it has not done any analysis on the monetary costs and benefits for the client. There are no estimates given in the study of the costs related to marketing budgets needed for reviving tourists and associated financial benefits from those tourists.

Determine Data Mining goals

Data Mining Goals

In order to provide a better understanding of the key patterns and factors in a tourist region the team's main data mining goal is to retrieve data (qualitative and quantitative) that best help unearth these variables. Moreover, to set the best possible business case for the study the team aimed at collecting reports, press briefings and other online data that can give a hint about Portugal national tourism strategy especially in the post-pandemic situation.

The main goal is to provide organization stakeholders with meaningful and insightful information about visitors of Portugal, compared to other countries. Better the data is prepared, the more accurate information it holds.

Data Mining Success Criteria

Speaking about Data Mining success criteria, for some of the modelings such as Association Rules for example, a statistically significant sample must be possessed. Also, minimum support and confidence values need to be met in order to provide strong marketing recommendations.

Produce Project Plan

Project plan

The project is part of an academic course and therefore must be completed within the stipulated time period of the academic semester. Some of the deliverables of the project are:

1. Jupyter Notebook script
2. CRISP-DM report
3. Final presentation slides deck including results and recommendations

Entire study is based on the CRISP-DM model and the entire report is based on the 6 phases/steps of the model.

Although each phase/step of the Model is critical yet some of the steps required more attention than others such as; data preparation and modelling stages took more percentage time than other steps of the Model. The reason for this is that the project team had to go forth and back several times to finally arrive at the final data set. Similarly, modelling required more time to not only develop but also check it several times to check the accuracy of the results.

Initial Assessment of Tools & Techniques

The tools selected are Python, performed via Jupyter Notebook and MS Excel. There were no other tools used in the completion of this project. No other tools are selected.

DATA UNDERSTANDING

Collect Initial Data

Data collection report

EuropeTop100Attractions.xlsx is the main dataset that our team will work with. This dataset involves 2 separate sheets (tables)that would be merged together based on the common variables which is the local ID. The data obtains both numerical and categorical variables that are explored in further sections of this report and holds records of TripAdvisor reviews on top 100 European attractions between January 2019 and August 2021.

For the sake of our team's convenient work, some few necessary edits were made before loading this initial data into the Jupyter notebook.

- The mentioned EuropeTop100Attractions.xlsx dataset was already split into two different excel files (based on the two sheets it holds), followed by the merge function of these sheets in Python. We saw this step as crucial based on the mentioned common variable that enables us to explore data (follow next section) in a better and more interesting way (for e.g. Local ID with MAG001 = Sagrada Familia).

- Secondly, also for convenience reasons, we corrected the data in the attraction sheet (already as a separate excel file). This step was done via pivot tables and following data “errors” were issued:
    - 2x the same name as “Old Town”. We found out that this belongs to Warsaw and Dubrovnik, so we made name adjustments to that.
    - MAG04, the Warsaw old town, was associated with the wrong ISO code (HR). We fixed that to PL (Poland)
    - The Vatican was assigned to Italy but with a different ISO code (VA). This was assigned to IT (Italy)
    - Both Scot and Scotland were identified, which was clearly just a typo error. We changed it to Scotland. Also, as Scotland holds the UK ISO code (same as England for ex.), we will have a disproportion of number of countries and ISO codes.

The final delivery of our team includes the edited excel sheet but also the one where the following observations were made. Methods of identifying those are included. As unveiled, Python and Excel are the main programs through which data is analysed and modelled.

Also, our team was given the Holiday.csv dataset. This dataset has not been collected and loaded in the working Jupyter as our team has not considered it as crucial for the analysis. Re-consideration may happen once we see a reason for it.

Describe Data

Data description report

Attachment 1 displays the meaningful description of variables from the EuropeTop100Attractions.xlsx dataset. From those, we consider several ones to be interesting and very important for the data analysis such as:

- LocalID + Name of the Attraction + Country
- reviewRating
- tripType
- reviewVisited (just month) & reviewWritten (concrete day)
- userLocation

Thanks to the describe function in Python, the summary statistics on the merged dataset is provided as the attachment 2. However, please bear in mind that there has been a lot of jumping back and forth between Data Understanding and Data preparation phase, so only initial summary statistics is provided to give us some first ideas. Later in the “exploration” part another summary of the statistics of the final dataset is provided once again. From the initial dataset however, we highlight the following insights:

- 92 120 observations
- We have some missing values on User Location, triptypes, Name (together with Country and ISO)
- Only 98 unique names but should be 100 as Local IDs
- ISO / Countries (24/23) does not match, probably because, as mentioned, country Scotland belongs to the UK, same as England
- Too many user locations (12 613)

Explore data

Data exploration report

The Exploration part unveils several interesting pieces of information from the whole dataset. Yet, we also explore the first difference (if any) between Portugal and the rest of the dataset. This is due to our team's task to look into Portugal mainly.

Attachment 3 explores the most rated countries, ordered ascendingly. Spain is by far the country that was visited the most, followed by England, Italy and Portugal. Attachment 4 explores the top 20 most visited attractions, led by Sagrada Familia and Tower of London and attachment 5 illustrates the bar chart (based on count function) showing differences across trip types held. By far, most of the visits, and so reviews, come from couple holidays, followed by family and friends.

When looking at histogram of reviews written on time scale from 2019 up to August 2021, there the strong impact of Covid-19 may be spotted. The reviews written dropped drastically from April 2020 onwards. When comparing Portugal through a density plot (Kernel Density Estimate), we observe an almost identical trendline, above all, meaning that the drop of reviews written in Portugal (green colour) came at the same time as in every other country (blue). This is illustrated in attachment 6.

Further and more detailed exploration, unveiling crucial insights would be generated after the Data preparation phase is finished with more accurate data.

Verify data quality

Data quality report

The provided database also includes some of the errors including missing values, duplicated rows and also some wrongly imputed categorical variables. All of them are described and identified in this data quality report.

Starting with missing values, the dataset holds many missing values, especially across trip type and user location. Assuming that these two fields were not required, this comes as no surprise. Also, about 5 500 missing values appear to be across Name, and so country and ISO codes. However, this is investigated further as these variables come from the merge attractions table. Find the information about missing values across all columns, including the query as an attachment 7.

The missing values across Name opened the data quality issue that is further investigated by using the Python query to unveil all of the unique LocalIDs. From attachment 8 that looks into it, it is quite obvious that MAG005 and MAG006 are missing, or better said, stored under the wrong name (u, genis). This has probably caused missing values from the previous paragraph but can be easily fixed in the data preparation phase.

Lastly, we perform the check control on duplicated rows across the dataset. From the simply duplicated.sum() query we have not found any duplicated rows. However, this query only checks all rows together but what if there are duplicated rows with the same user name and name of the attractions? We investigated it and found 7478 duplicated rows which we double checked with one of the duplicated usernames. Find it as an attachment 9.

DATA PREPARATION

Select Data

Our team keeps the provided datasets and no other data sources are analyzed in the Jupyter notebook.

Clean Data

Firstly, our dataset needs to be fixed by the right LocalIDs namings as illustrated in the previous stage. Therefore, we re-named “u” and “genis” LocalIDs with the correct ones. However, as this type of error appeared to be in the reviews sheet before, we were not being assured that by now, the missing values from Names, ISO and countries disappeared, rather the opposite. The merge of the two tables (reviews and attractions) happened before this clean-up, so the expected change in the data can not be observed. For this reason, we needed to perform the merge of these two tables once again. Therefore, we erased original columns from the attractions table (Name, Country, ISO) and then, redid the merge once again. For the double check of this effort, we ran the summary statistics to find out whether the missing values observed from the Data verification part have been truly fixed or not. Luckily, they were.

Secondly, we erased the duplicated rows and only kept the last record. For assurance reasons, this clean-up activity was double-checked too.

Construct Data

In this part, further data operations are done to make the dataset better for the modelling part. First of all, we see a lot of columns that we consider as meaningless to keep. For that reason, we go for the removal of the following variables:

- ISO - country gives us the same information
- Extraction date - two dates only, not giving us meaningful information
- Position on ranking - not related towards or analyses
- Sites on ranking - not related towards our analyses
- Total reviews - one value per attraction - not saying us much either
- Review language - all hold english value, meaningless information

Derived features

Our team elaborates on creating new variables that can provide meaningful information for further analyses. For such reasons, the following derived features are created:

- Covid_time - with values “before_covid” and “after_covid”, based on whether the review visited was prior to the 1st of March 2020 (when we considered the start of Covid 19, especially in Portugal but also around europe (1)
- Visits_together_country - total visits per country. This is analysed later on to compare visits increase per country compared to the pre Covid time

- Day_of_week_review_given - day at which the review was given.
- User_visits_sum - number of visits in the given country assigned per username (notice that this is only applicable for the following countries: Portugal, Spain, Italy, France. We explain this step further below)

After these variables were created, our team desired to have two more datasets. One would be for Portugal only and another one consisting of the following four countries: Portugal, Spain, France and Italy. These countries were selected because we believe that those can be primarily compared to Portugal. All are culturally very similar and most importantly, they all belong to the most visited countries in Europe (2). We think that the mix of these factors can form such countries as example countries that Portugal can be inspired by.

First, all countries have created their own dataset. Later, to each one of them, we create the mentioned variable “user_visits_sum”, to make sure that if someone has visited more than one country, this user name won't be assigned total visits across all countries but rather just the given one. This is done this way for the analyses explained later on. After this, our team joined all these four dataset back together, through the .concat function (3). For better understanding of the data preparation explained in the last 2 paragraphs, please follow attachment 10 that illustrates the code method.

Lastly, for the Portuguese dataset, we tried to apply a string contain method that could possibly group together the majority of user locations under a common umbrella (like England, Scotland belong to the UK). As already observed from the data exploration part, many user locations consist of a city and then the country (especially US). With the mentioned method, we tried to group together as many possible locations as possible, meaning London, UK or Manchester, UK under a common variable - United Kingdom. Those rows that had a missing value, we just assigned them a value to this new column as “Missing location - NaN” whereas for those that we were not able to assign an updated location we wrote “different country” and grouped them together.

Data Formatting

Any data formatting elaboration was made up to this point and neither is expected in the further elaborations with data.

Data Integrating

No additional external data was found to be useful and therefore, integrated to our prepared dataset.

Data exploration - part two

All the data preparation explained in the previous paragraphs were constructed not only for the modelling reasons. Another important reason was to be able to provide more information from the data exploration part. Our team gathered some more key information extracted from the data and looked into Portugal separately, as well as the comparisons between Portugal and the mentioned competitive countries such as Italy, France and Spain.

Attachment 11 illustrates the histogram of ReviewVisited in Portugal. While we consider this graph as not really descriptive to make some final conclusion on the tourists' distribution across months (simply just few data and not having data for the past 5 years at least), we noticed a weird bar change between july and august 2019. It looks like in August 2019, when there was still no covid, there was a huge drop in reviews visited. When comparing it to nearby countries such as Spain or Italy (attachment 12) through density plot, the same happened in Spain and a little bit in Italy too. We did not find any significant reason why, but one assumption was made that it may be because of longer summer vacations (2 months) during which tourists may visit sights but all check in the first month of their arrival.

From attachment 11, we obviously strongly noticed that the Covid19 start has had a huge impact ever since it reached Portugal. From April to May 2020, the drop has reached almost zero and there was never a chance of getting the initial numbers back. The very same happened in every other country.

Insight 1 - Spain with the heaviest drop of visitors after the pandemic hit

Attachment 13 displays crosstabs on each Portuguese attraction with two columns: before and after covid. The goal was to identify not only the numeric difference (as sum of reviews before and after covid), but also to find drops in the avg review given. This could show potentially that some attractions were doing better after the pandemic compared to others. Even though the differences are noticeable, when looking at the absolute number of reviews given after covid per each attraction, the statistical significance with minimum sample size of 100 (4) is not presented. Yet, attachment 14 presents another column representing the increase of visitors per each attraction after covid compared to the period before the pandemic. It was calculated as follows: **(nr of visits after covid / nr of visits before covid) * 100**. . However, no noticeable increase was spotted either.

The same formula was applied for Portugal, Spain, Italy and France separately and the first meaningful comparison insight may be spotted through the visualization of it at attachment 15. It shows that while France, Italy and Portugal were visited relatively equally, a noticeable drop was spotted in Spain. This can be due to the heavy pandemic hit on Spain caused in spring 2020 which resulted in the Spanish government decision to shut down the borders (5).

Insight 2 - Portugal and Spain with better proportion of visits per visitor

Attachment 16 displays the proportion of visits per visitors for Spain, Italy, France and Portugal. Clearly, Spain and Portugal hosted more than 40% of visitors that visited more than just one attraction in the given country. On the other hand, Italy and France have this proportion much more different, hosting 80%+ percent of visitors that paid a visit to only one attraction in their country.

Even though this comparison may seem interesting, our team concludes that it is not a strong evidence of Portugal, nor Spain being better at attracting visitors to visit more than one attraction. For example, many Portuguese attractions are close to its capital city Lisboa and naturally people might visit more than just one attraction there. Yet, this may not be the

case for other countries, especially France or Italy which are bigger countries and their attractions may be far from each other.

Attachment 17 shows the proportion of visits per visitor across Portuguese attractions. The rationale behind it was to identify attractions where visitors with 2 or more overall visits across Portugal go to. For example, more than 60% of visitors of Cais de Ribeira visit another Portuguese attraction. From the marketing point of view, this may be the best promotion place for other attractions.

Insight 3 - Couple being by far the most popular triptypes

Attachment 18 compares trip types and clearly shows that couples are being the most represented types of trip. 50% of this type representation is not only across all Portuguese attractions, but also when compared to other countries Portugal is compared to.
These factors are strong enough to build some marketing recommendations on. For example, some ticket promotions can be done specifically made for couples like get one, get another one with 50% off. Also, the online ads of Portuguese attractions can be targeted to such people more.

Attachment 19 provides a look into the trip type categories across the time. The pink shaded colour represents solo trip types and all other colours represent Couples, Friends and Family. Business trip type was disregarded due to a small number of occasions. It comes as no surprise that the solo trip type did not drop drastically as other categories, even though the drop is very significant also. This information just confirms the assumption that solo trip type is the most reasonable in the pandemic times. Having this assumption confirmed, a little bit more budget may be recommended to allocate for the online ads targeting reaching solo travellers.

Insight 4 - Portuguese as the tourists with significantly less drop in Portugal after covid

The Covid-19 pandemic has caused a lot less drop in local travelling, rather the opposite. This is clearly visible at attachment 20. Based on the updated user location constructed in the data preparation phase, we can see that tourists from Portugal (represented by the blue shaded colour) have travelled less in Portugal but way more differently than for example. Australians, represented by the black line in this case. Tourists from other countries, represented by different colour lines, have the similar trend line on the density plot, maybe except Spain (pink colour). Visitors from Spain have not drastically dropped, probably because of the fact that it is the only country that shares borders with Portugal and therefore, the access there is faster and more simple compared to other countries.

The provided density plot has unveiled what many of us think and that local travelling has been crucial to keep the travel economy running and bleed less. For that reason, one of our recommendations would be to try to promote local attractions not only to the international audience, but also to local Portuguese people. Even eventually to provide them with specific discounts to drive the visits higher.

Other observations

Attachment 21 provides a horizontal bar chart showing the most popular days at which attractions in Portugal receive TripAdvisor reviews. There are no significant differences, yet Sunday is the most popular day for writing reviews. We further investigated whether the day of the review given has any impact on the ratings. This can be observed at attachment 22 with results that are not so different one day from another. Sunday, however, was the day when people tended to give slightly better ratings. Also, when looking at this graph by countries (see attachment 23), this assumption is neither proved right or wrong. From a marketing point of view, even though it is not a significant marketing activity, we would recommend notifying visitors about the reviews more on the happy days such as Sunday, rather than Monday.

MODELLING

Select modelling technique

Algorithm selection

Even though the data exploration revealed a lot of information already, our team thinks that the Association Rules modelling through the Apriori algorithm should be the main modelling technique to use.

RFM segmentation, neither similarity nor similarity matrices are included in the report and presentation. Our team tried to run those several times too but after analysing the results, we do not think that these modelling techniques provide any sort of valuable information value.

RFM in particular, would have been interesting if the data was different. From previous graphs, it is clear to notice that the recency variable would be unimportant as the majority of visits were done prior to the pandemic. Segmenting visitors based on whether they visited Portugal 20 or 22 months ago does not make sense to us. Please, see attachment 24 that displays the distribution histogram of the RFM. It is clear to see that the recency (nr of days from the last visit) and frequency (80% of visitors only visited once) would not be strongly distinguishing the segments. We thought of using user contribution as a monetary value, but in the end, this variable would not give us any meaningful information either. For these reasons, RFM is excluded.

Similarity and dissimilarity would be interesting to do and so can be seen in the Jupyter notebook. Yet, even beforehand, our team thought whether it can bring us any good information to work with but unfortunately, this model does not seem relevant either.

Modelling assumptions

No specific modelling assumption needs to be done as long as a separate table (dataset) with user name and visited attraction is provided.

Generate test design

After the exploration of data we found 959 unique users have visited at least 2 attractions in Portugal so this number has a good combination of attractions in the dataset. Its significant record to expect valuable information.

However, it is not much to do the test design on a small sample first and then go for the final modelling. Therefore, for the model we consumed the whole dataset.

Build Model

1. First, we create a new dataset that only contains two columns: user name and name of the attraction visited. Please see attachment 24 to see this dataset based on which the Apriori algorithm would work.
2. Secondly, we create a simple pivot table containing the mentioned two variables. The final outcome can be seen as attachment 24, displaying each username on a separate row with 1s and 0s. If 1 is displayed, that means that the user visited the given attraction. If 0, then no visit was recorded.
3. After that, we apply the apriori algorithm with minimum support of 1%. See the figure below.

Figure 1 - Apriori algorithm

```
portugal_frequent_namesets = apriori(portugal_associations_pivot, min_support=0.01, use_colnames=True)
```

4. Then, we are able to generate association rules based on all three key components namely Confidence, Lift and Support. All three to be shown in the Evaluation part of the report. Figure 2 shows the block of code used for such generation.

Figure 2 - Query to generate association rules by support

```
# Generate the association rules - by support
rulesSupport = association_rules(portugal_frequent_namesets, metric="support", min_threshold=0.001)
rulesSupport.sort_values(by='support', ascending=False, inplace=True)
rulesSupport.head(5)
```

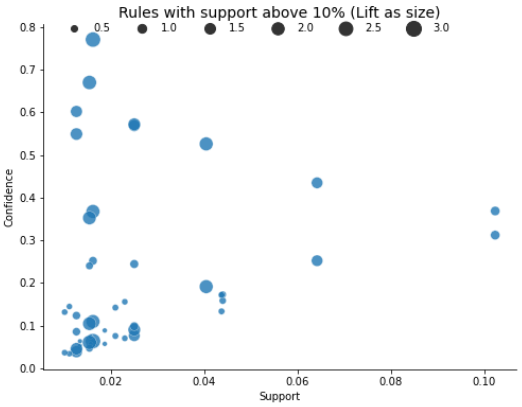
Except the minimum threshold set, there are no other criteria our team followed when generating association rules. This is due to a relatively small number of overall records.

Asses model

For the assessment of the model we used a simple scatter plot. On the X axis we placed the support, for the Y axis we used confidence and the size of the dots depends on the lift criteria. Such visualization does not provide us with some general information only, but most importantly, it can unveil how statistically significant and accurate the association rules are.

Based on the following figure, our team is able to provide how significant the association rules model can be.

Figure 3 - Scatter plot of generated Association Rules

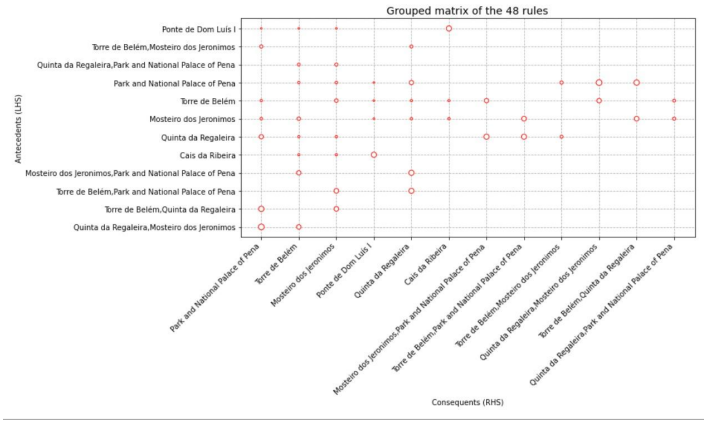


The graph from above illustrates relatively low quality results. First and foremost, only two sets of attractions combined have support of around 10%, meaning they are present in 10% of the all results together. Even though such support value is quite high, their confidence is in between 30-40% and the size of their dots illustrate that their lift value is low, definitely not higher than 1,5. On the other hand, we have some combinations with high confidence and potentially high lift value too, but unfortunately with small support.

The presented scatter plot indicates that the model itself might not be presenting high-quality and valuable data information.

Figure 4 gives us the first idea of what the best combinations may be.

Figure 4 - Group matrix of 48 rules



EVALUATION

Evaluate results

Assessment of Data Mining results

For a better overview, we provide the following tables that represent outputs of Association Rules modelling.

Figure 2 - Top 5 Association by Support

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
12	(Torre de Belém)	(Mosteiro dos Jeronimos)	0.327774	0.277483	0.102350	0.312259	1.125327	0.011399	1.050566
13	(Mosteiro dos Jeronimos)	(Torre de Belém)	0.277483	0.327774	0.102350	0.368852	1.125327	0.011399	1.065086
17	(Park and National Palace of Pena)	(Quinta da Regaleira)	0.254486	0.147587	0.064190	0.252234	1.709061	0.026631	1.139947
16	(Quinta da Regaleira)	(Park and National Palace of Pena)	0.147587	0.254486	0.064190	0.434932	1.709061	0.026631	1.319334
6	(Park and National Palace of Pena)	(Mosteiro dos Jeronimos)	0.254486	0.277483	0.043973	0.172790	0.622707	-0.026643	0.873439

Some twins are identified, yet when looking at the confidence and especially lift values, it already tells us that these combinations are not really strong. Confidence around 30% may be fine, but then lift value tells us about how greater this combination is compared to the “as-is-now” situation. Moreover, these are the attraction combinations (first 4 rows) that are situated close to each other, so these results would be kind of expected and already in consideration. For example, Torre de Belem website already provides tickets for both its tower plus Mosteiro dos Jeronimos.

Figure 3 - Top 5 Association by Confidence

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
25	(Quinta da Regaleira, Mosteiro dos Jeronimos)	(Park and National Palace of Pena)	0.020975	0.254486	0.016174	0.771084	3.029971	0.010836	3.256720
42	(Torre de Belém, Quinta da Regaleira)	(Park and National Palace of Pena)	0.022997	0.254486	0.015416	0.670330	2.634056	0.009563	2.261393
38	(Quinta da Regaleira, Mosteiro dos Jeronimos)	(Torre de Belém)	0.020975	0.327774	0.012636	0.602410	1.837884	0.005761	1.690751
30	(Torre de Belém, Park and National Palace of P...	(Mosteiro dos Jeronimos)	0.043720	0.277483	0.025019	0.572254	2.062305	0.012887	1.689128
32	(Mosteiro dos Jeronimos, Park and National Pal...	(Torre de Belém)	0.043973	0.327774	0.025019	0.568966	1.735849	0.010606	1.559565

Even though here the confidence levels are quite high, the probability of someone visiting 2 places already is pretty low (no more than 2.5%) and untrackable, unlike in the online environment.

Figure 4 - Top 5 Association by Lift

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
28	(Park and National Palace of Pena)	(Quinta da Regaleira, Mosteiro dos Jeronimos)	0.254486	0.020975	0.016174	0.063555	3.029971	0.010836	1.045469
25	(Quinta da Regaleira, Mosteiro dos Jeronimos)	(Park and National Palace of Pena)	0.020975	0.254486	0.016174	0.771084	3.029971	0.010836	3.256720
47	(Park and National Palace of Pena)	(Torre de Belém, Quinta da Regaleira)	0.254486	0.022997	0.015416	0.060576	2.634056	0.009563	1.040002
42	(Torre de Belém, Quinta da Regaleira)	(Park and National Palace of Pena)	0.022997	0.254486	0.015416	0.670330	2.634056	0.009563	2.261393
27	(Quinta da Regaleira)	(Mosteiro dos Jeronimos, Park and National Pal...	0.147587	0.043973	0.016174	0.109589	2.492206	0.009684	1.073692

Here, the support is extremely low, meaning on very few occasions such visits of presented attractions happen.

Association rules are usually used for the following actions: Cross-sell, up-sell, Product placement, Affinity promotion and customer behaviour. As already mentioned, presented results show what was expected: People visit nearby attractions also (such as Torre de Belem and Mosteiro dos Jerónimos).

In terms of cross-selling recommendations, our team still thinks it is a good idea to make sure each attraction promotes the other one with the highest level of confidence. However, some of the attractions already do that as mentioned.

We do believe that information from the data exploration part two also unveiled a good level of information which some further marketing recommendations can be based on. For example, thanks to digital analytics, attractions can target young couples and solo travellers, especially from Portugal.

Approved models

From a data mining perspective, association rules did not meet necessary criteria, especially due to the low number of attraction combinations with strong confidence levels, accompanied with decent lift or support values. Therefore, we consider the Association rules Model as not a success.

Review process

Review of process

At the beginning the dataset looked to be promising and actually was ok to fulfil the purpose of answering some of the visitation patterns of Portuguese attractions. However, lack of data has caused low quality results of the modelling techniques.

Our team believes that some interesting patterns were found, as well as descriptive comparisons with Portuguese main tourism competitors. However, the quantity part of the dataset became an issue, since very few recordings were made after covid and that caused small statistical significance for us to make further interesting analyses. For example, we

were not able to provide and prove how given attractions adapted to the new needs after the pandemics. Crosstabs analysis with average ratings before and after pandemic simply could not be referred to as statistically relevant. As explained before, neither RFM modelling could have been done in a right way that would eventually let us understand visitors of Portugal more.

Having a larger datasource of Portugal would definitely help the process. Also, some quality variables such as age, marital status etc. were missing and robbed us from further analyses and understanding of the customers and their segmentation. User location was very hard to analyze and would have been much easier if only the states were provided.

Lastly, some additional information from attractions would be welcomed as well. For example, knowing the date ranges when a given attraction was closed due to the pandemics or price of the tickets (maybe for correlation analyses) would come useful as well.

Determine next steps

List of possible actions

Before going for final marketing recommendations resulting from the possible modelling techniques (RFM for ex.), our team would challenge to obtain another dataset that the Portuguese National Tourism Board Organization might possess.

Also, if the stakeholders are satisfied with the current results and analyses, it can approach further by recommending given attractions to provide recommended marketing suggestions written in this report.

Decision

Decision lies with the National Tourism Board Organization together with the team.

DEPLOYMENT

Plan Deployment

As the provided recommendations are for the separate Portuguese attractions, it is mostly on NTBO to make sure our marketing suggestions are delivered. And then, it is up to the single attractions' stakeholders to implement recommended suggestions.

Plan Monitoring and Maintenance

Unfortunately, through a third party provider (tripAdvisor) an A/B test is unlikely to happen. However, there are ways in which attractions can measure the impact of recommended suggestions. This can be done via a digital analytics team in each of the attraction entities. If some attraction decides to apply any of the targeting suggestions, it can be tracked via analytics tools to measure success.

Also, long-term wise, the dataset with updated numbers for 2022 could be analysed once again to monitor changes in visitors patterns. However, this may be a more challenging plan, as our team was mostly asked to do a descriptive analysis.

Produce Final Report & Review Project

Besides this final report, our team provides a presentation on February 3rd. On this date, main results from this analysis, as well as the feedback on this project will be discussed.

ATTACHMENTS

Attachment 1 - Description of variables

***EuropeTop100Attractions_ENG_20190101_20210821* dataset description**

Sheet Reviews

Reviews published in Tripadvisor from January 1, 2019 to August 21, 2021, in English, for the top 100 tourist attractions in Europe.

- **localID**: string - ID of the attraction
- **extractionDate** - date - date when the review was extracted
- **globalRating** - numeric - global rating of the attraction at the time of the review extraction (reviews in Tripadvidor are in a scale from 1 to 5 stars)
- **positionOnRanking** - numeric - position in TripAdvisor's regional ranking at the extraction date
- **sitesOnRanking** - numeric - total number of attractions in TripAdvisor's regional ranking at the extraction date
- **totalReviews** - numeric - total reviews written for the attraction at the time of the review extraction
- **userName** - string - user name of the TripAdvisor user who posted the review. The user name is composed of two parts (first@second). The first is the public name of the user. The second is the TripAdvisor unique identifier of the user.
- **userLocation** - string - location of where the user who posted the review lives. This is not a mandatory field, so many users to not provide their location
- **userContributions** - numeric - how many reviews have the user wrote in TripAdvisor at the moment of the extraction of the review
- **tripType** - string - type of trip type. This is not a mandatory field
- **reviewWritten** - date - date when the review was published
- **reviewVisited** - date - date when the customer visited the attraction. The day is always 1 because Tripadvisor only ask users to describe the year and the month, not the day
- **reviewRating** - numeric - quantitative rating assigned by the user (1 star - bad to 5 stars - excellent)
- **reviewLanguage** - string - language the review was written (in this case should be always "en" for english)
- **reviewFullText** - string - full text of the review (since this course does not address Text Mining the use of this field is completely optional and its use will not be considered for grading)

Sheet Attractions

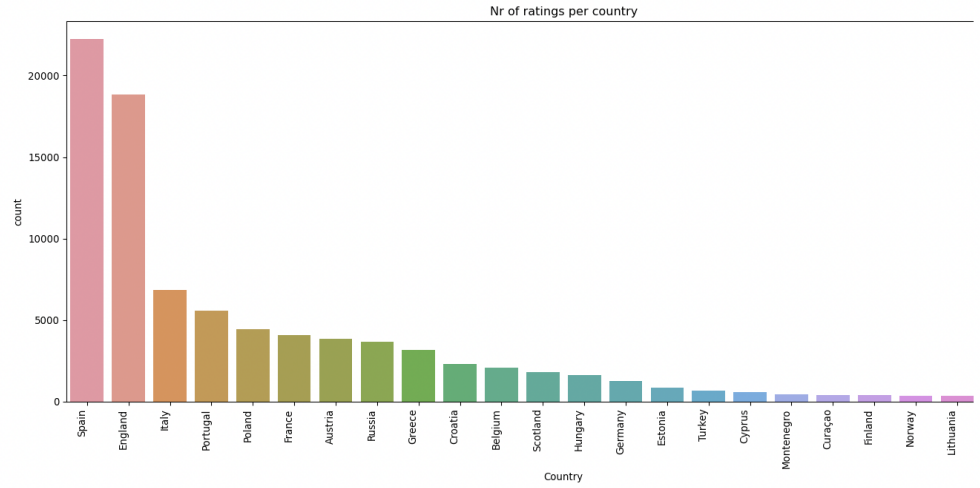
Information about the attractions.

- **ID**: string - ID of the attraction
- **Name**: string - name of the attraction
- **Country**: string - name of the country or region
- **ISO**: string - ISO code of the country or region

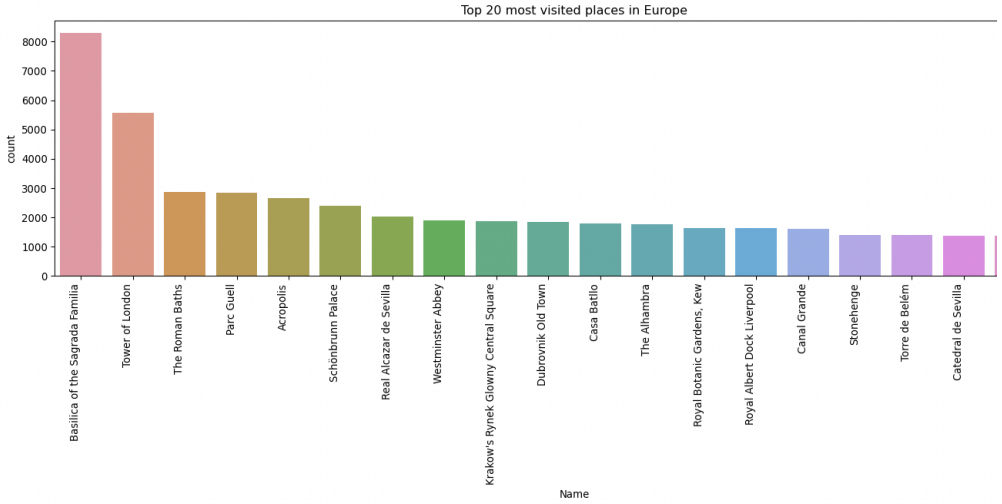
Attachment 2 - Original dataset - summary statistics

	count	unique	top	freq	first	last	mean	std	min	25%	50%	75%
localID	92120	100	MAG001	8309	NaT	NaT	NaN	NaN	NaN	NaN	NaN	NaN
extractionDate	92120	91896	2021-08-20 09:00:54.185000	2	2021-08-20 08:24:40.077	2021-08-21 16:27:17.026	NaN	NaN	NaN	NaN	NaN	NaN
globalRating	92120.0	NaN	NaN	NaN	NaT	NaT	4.485166	0.178085	4.0	4.5	4.5	4.5
positionOnRanking	92120.0	NaN	NaN	NaN	NaT	NaT	3.91459	4.843013	1.0	1.0	2.0	6.0
sitesOnRanking	92120.0	NaN	NaN	NaN	NaT	NaT	748.263537	802.742304	5.0	154.0	484.0	1186.0
totalReviews	92120.0	NaN	NaN	NaN	NaT	NaT	40556.601813	42914.381014	5179.0	14152.0	24454.0	51324.0
userName	92100	65785	Malgorzata@Margo7850p	31	NaT	NaT	NaN	NaN	NaN	NaN	NaN	NaN
userLocation	78652	12613	London, UK	3710	NaT	NaT	NaN	NaN	NaN	NaN	NaN	NaN
userContributions	92120.0	NaN	NaN	NaN	NaT	NaT	477.52056	7270.518677	0.0	20.0	66.0	215.0
tripType	63052	5	Couples	31702	NaT	NaT	NaN	NaN	NaN	NaN	NaN	NaN
reviewWritten	92120	934	2019-10-09 00:00:00	473	2019-01-01 00:00:00.000	2021-08-21 00:00:00.000	NaN	NaN	NaN	NaN	NaN	NaN
reviewVisited	91410	57	2019-09-01 00:00:00	8497	2015-10-01 00:00:00.000	2021-08-01 00:00:00.000	NaN	NaN	NaN	NaN	NaN	NaN
reviewRating	92120.0	NaN	NaN	NaN	NaT	NaT	4.578658	0.792693	1.0	4.0	5.0	5.0
reviewLanguage	92120	1	en	92120	NaT	NaT	NaN	NaN	NaN	NaN	NaN	NaN
reviewFullText	92120	85088	Is a nice place to visit when you visit Barcel...	3	NaT	NaT	NaN	NaN	NaN	NaN	NaN	NaN
Name	86560	98	Basilica of the Sagrada Familia	8309	NaT	NaT	NaN	NaN	NaN	NaN	NaN	NaN
Country	86560	24	Spain	22232	NaT	NaT	NaN	NaN	NaN	NaN	NaN	NaN
ISO	86560	23	ES	22232	NaT	NaT	NaN	NaN	NaN	NaN	NaN	NaN

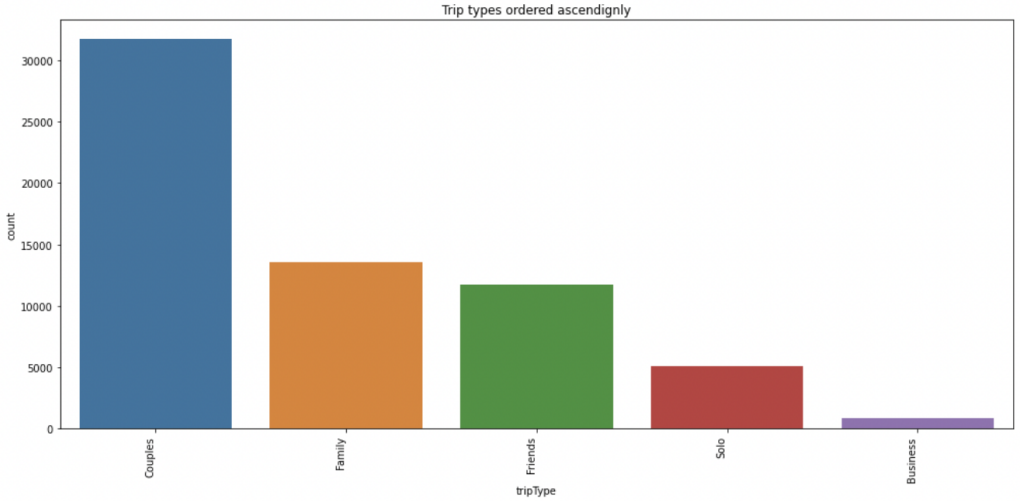
Attachment 3 - Most rated countries



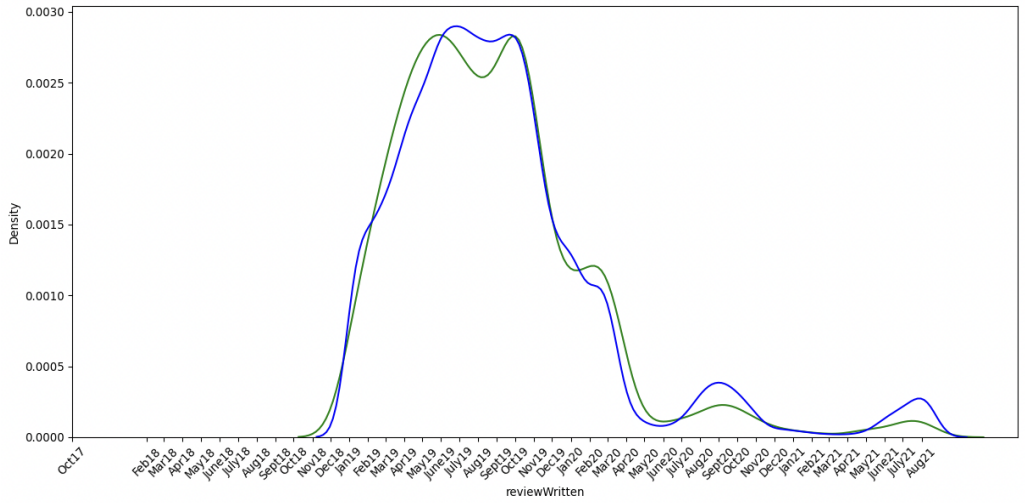
Attachment 4 - Top 20 most visited attractions



Attachment 5 - Most represented trip types



Attachment 6 - DENSITY PLOT Portugal (green) vs Rest of countries (blue)



Attachment 7 - Missing values across all columns

```
localID                 0
extractionDate          0
globalRating            0
positionOnRanking       0
sitesOnRanking          0
totalReviews            0
userName               20
userLocation        13468
userContributions       0
tripType            29068
reviewWritten           0
reviewVisited         710
reviewRating            0
reviewLanguage          0
reviewFullText          0
Name                 5560
Country              5560
ISO                  5560
dtype: int64
```

Attachment 8 - Wrong localIDs

```
ds_one.localID.unique()
```

```
array(['MAG001', 'MAG002', 'MAG003', 'MAG004', 'genis', 'u', 'MAG007',
       'MAG008', 'MAG009', 'MAG010', 'MAG011', 'MAG012', 'MAG013',
       'MAG014', 'MAG015', 'MAG016', 'MAG017', 'MAG018', 'MAG019',
       'MAG020', 'MAG021', 'MAG022', 'MAG023', 'MAG024', 'MAG025',
       'MAG026', 'MAG027', 'MAG028', 'MAG029', 'MAG030', 'MAG031',
       'MAG032', 'MAG033', 'MAG034', 'MAG035', 'MAG036', 'MAG037',
       'MAG038', 'MAG039', 'MAG040', 'MAG041', 'MAG042', 'MAG043',
       'MAG044', 'MAG045', 'MAG046', 'MAG047', 'MAG048', 'MAG049',
       'MAG050', 'MAG051', 'MAG052', 'MAG053', 'MAG054', 'MAG055',
       'MAG056', 'MAG057', 'MAG058', 'MAG059', 'MAG060', 'MAG061',
       'MAG062', 'MAG063', 'MAG064', 'MAG065', 'MAG066', 'MAG067',
       'MAG068', 'MAG069', 'MAG070', 'MAG071', 'MAG072', 'MAG073',
       'MAG074', 'MAG075', 'MAG076', 'MAG077', 'MAG078', 'MAG079',
       'MAG080', 'MAG081', 'MAG082', 'MAG083', 'MAG084', 'MAG085',
       'MAG086', 'MAG087', 'MAG088', 'MAG089', 'MAG090', 'MAG091',
       'MAG092', 'MAG093', 'MAG094', 'MAG095', 'MAG096', 'MAG097',
       'MAG098', 'MAG099', 'MAG100'], dtype=object)
```

Attachment 9 - Duplicated rows

```
ds_one[ds_one[['Name', 'userName', 'userLocation']].duplicated() == True]

# looks like as those, almost 7,5K rows are like that. In the data preparation phase, we only keep last records
```

	localID	extractionDate	globalRating	positionOnRanking	sitesOnRanking	totalReviews	userName	userLocation	userContributions
209	MAG001	2021-08-20 08:26:28.691	4.5	1	1186	163828	michael.t@michaeltast	Kolkata (Calcutta), India	19
213	MAG001	2021-08-20 08:26:34.067	4.5	1	1186	163828	MerylStrepp@MerylStrepp	Henley-on-Thames, UK	57
270	MAG001	2021-08-20 08:27:02.989	4.5	1	1186	163828	insertname@jadedbear	Singapore, Singapore	10
271	MAG001	2021-08-20 08:27:03.020	4.5	1	1186	163828	James V@jamesv841	Odessa, TX	4
272	MAG001	2021-08-20 08:27:03.035	4.5	1	1186	163828	Ashleigh1505@Ashleigh1505	Essex, UK	48
...	...	...	...	...	...	...	...	...	...
91994	MAG098	2021-08-21 16:25:49.098	4.5	1	58	5327	Globetrotter008@Globetrotter008	United Kingdom	1135
91995	MAG098	2021-08-21 16:25:49.223	4.5	1	58	5327	Linda R@l1275STlindar	Brighton, UK	9
91996	MAG098	2021-08-21 16:25:49.360	4.5	1	58	5327	Juanchoborda@Juanchoborda	Tampa, FL	109
91997	MAG098	2021-08-21 16:25:49.558	4.5	1	58	5327	FlyingMike@mikevE9651PM	NaN	32
92049	MAG099	2021-08-21 16:26:27.357	5.0	1	73	5345	See you around Alice@alicemierzchala	Nancy, France	389

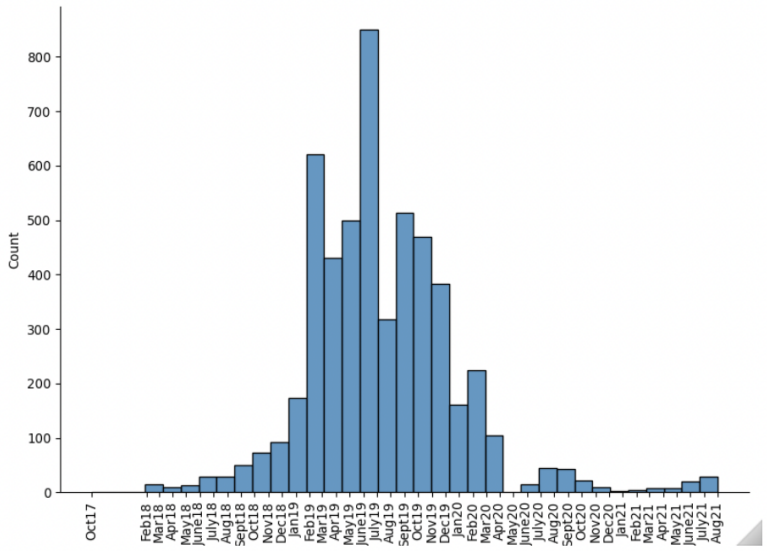
7478 rows × 18 columns

Attachment 10 - .concat function code

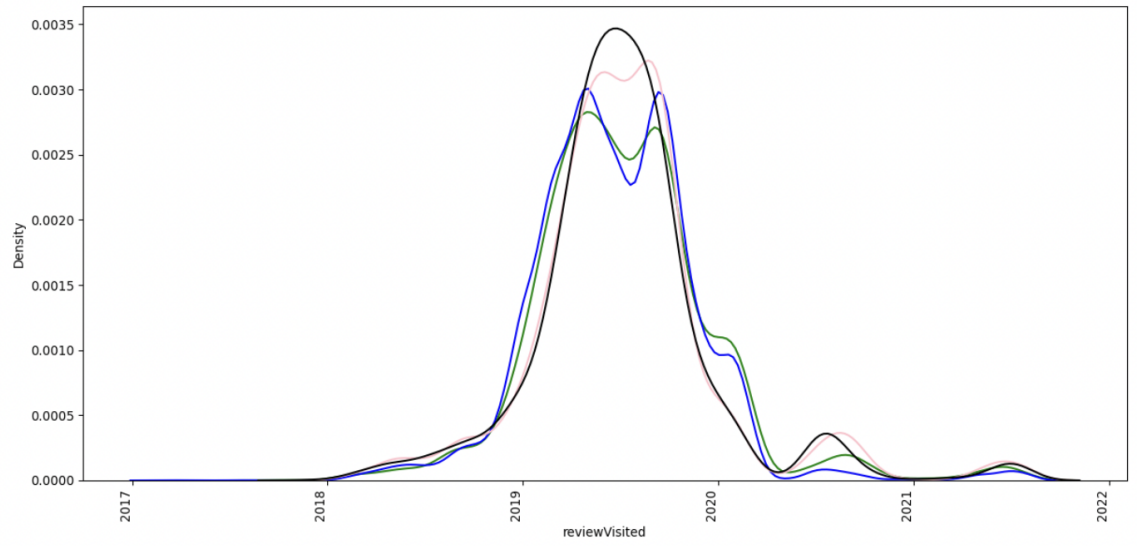
```
# adding user visits for all countries into separate columns
Portugal['user_visits_sum'] = Portugal['userName'].map(Portugal['userName'].value_counts())
italy['user_visits_sum'] = italy['userName'].map(italy['userName'].value_counts())
spain['user_visits_sum'] = spain['userName'].map(spain['userName'].value_counts())
france['user_visits_sum'] = france['userName'].map(france['userName'].value_counts())
```

```
# merging dataset together
pr_it_es_fr_visits = [Portugal, italy, spain, france]
pr_it_es_fr_visits_final = pd.concat(pr_it_es_fr_visits)
pr_it_es_fr_visits_final
```

Attachment 11 - Review Visited histogram - Portugal



Attachment 12 - Density plot PRT (green), ESP (blue), FRA (black), ITA (pink)



Attachment 13 - Crosstabs on every Portuguese attraction

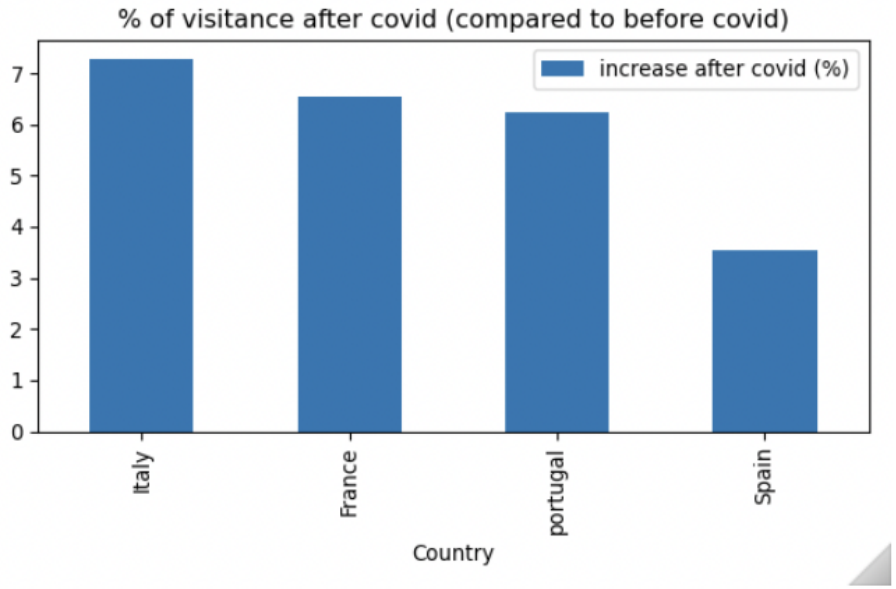
Covid_time	After Covid	Before Covid
Name		
Bom Jesus do Monte	11	122
Cais da Ribeira	18	286
Mosteiro dos Jeronimos	54	1044
Park and National Palace of Pena	54	951
Ponte de Dom Luís I	54	782
Quinta da Regaleira	37	547
Torre de Belém	81	1216

Covid_time	After Covid	Before Covid
Name		
Bom Jesus do Monte	4.181818	4.770492
Cais da Ribeira	4.666667	4.587413
Mosteiro dos Jeronimos	4.370370	4.444444
Park and National Palace of Pena	4.555556	4.176656
Ponte de Dom Luís I	4.777778	4.695652
Quinta da Regaleira	4.540541	4.804388
Torre de Belém	4.160494	4.229441

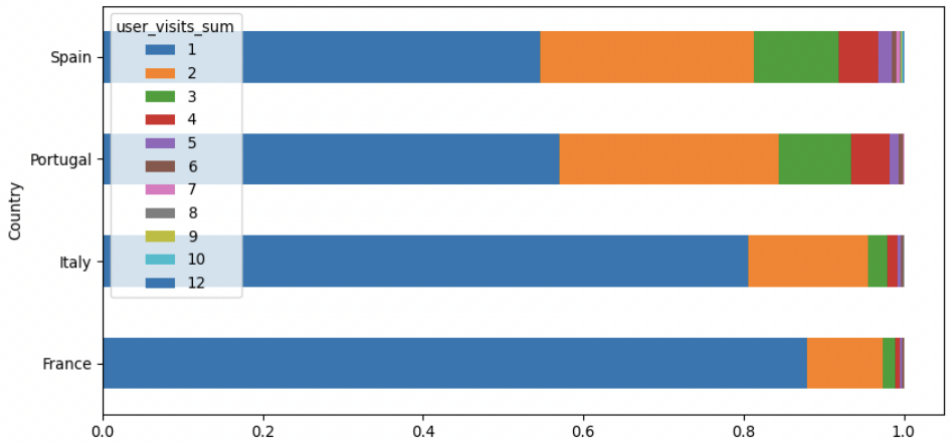
Attachment 14 - Increase of visitors after covid for each of Portuguese attractions

Name	Covid_time After Covid	Before Covid	AC growth after before covid
Bom Jesus do Monte	11	122	9.016393
Cais da Ribeira	18	286	6.293706
Mosteiro dos Jeronimos	54	1044	5.172414
Park and National Palace of Pena	54	951	5.678233
Ponte de Dom Luís I	54	782	6.905371
Quinta da Regaleira	37	547	6.764168
Torre de Belém	81	1216	6.661184

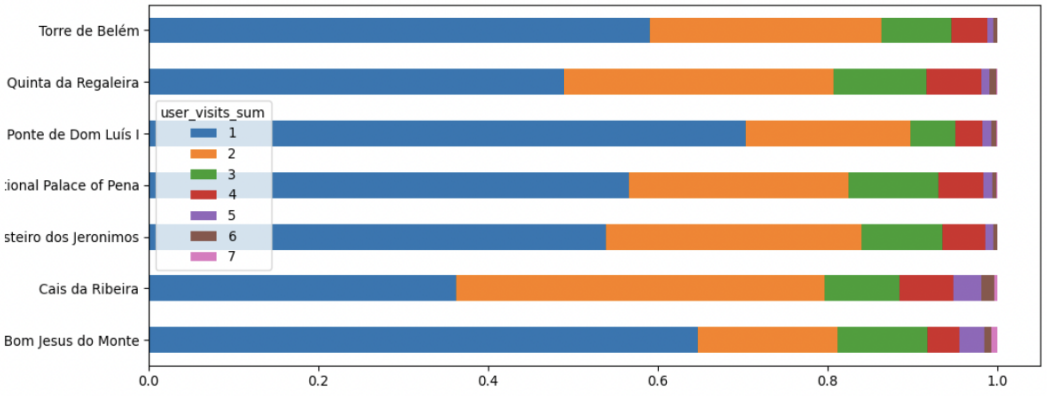
Attachment 15 - Increase of visitors after covid for each country



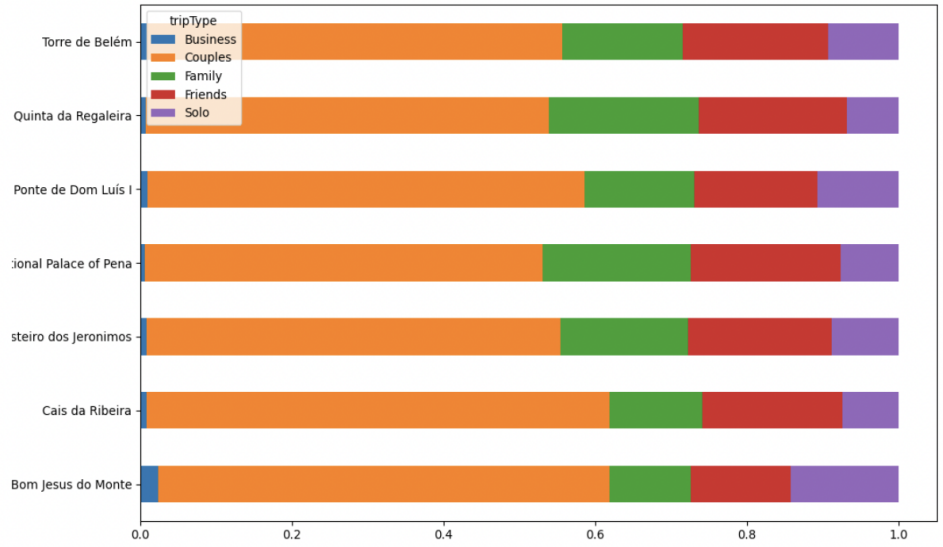
Attachment 16 - Proportion of visitors based on number of visits per country



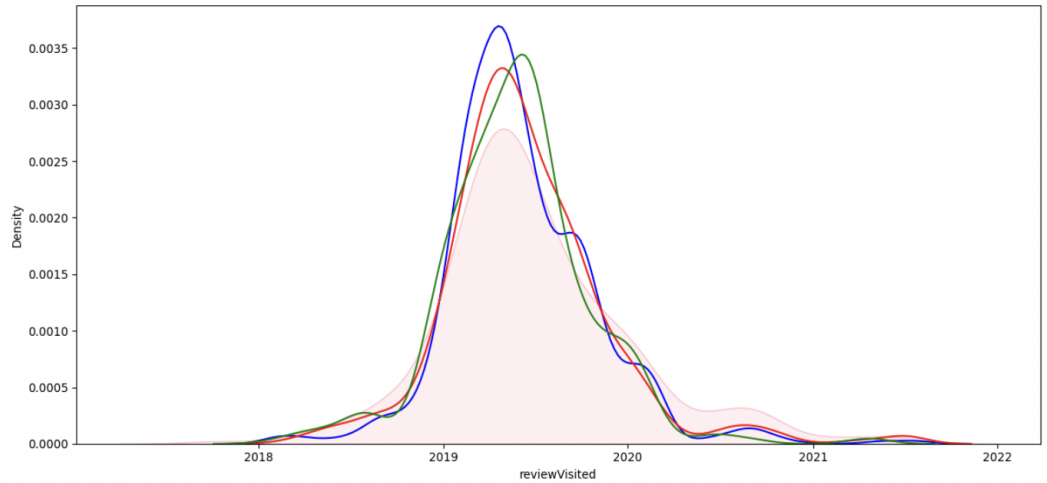
Attachment 17 - Proportion of visitors based on number of visits per PRT attraction



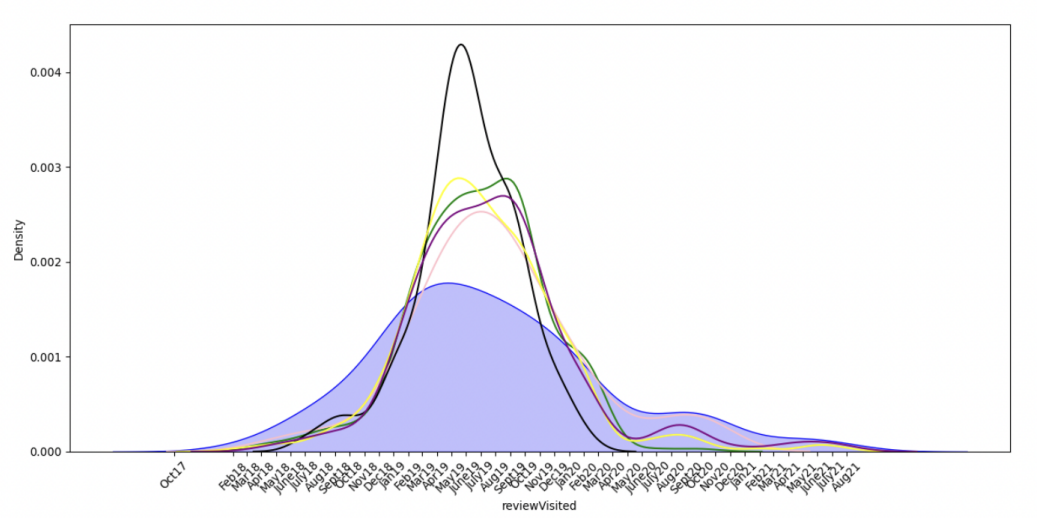
Attachment 18 - Proportion of trip types per Portuguese attraction



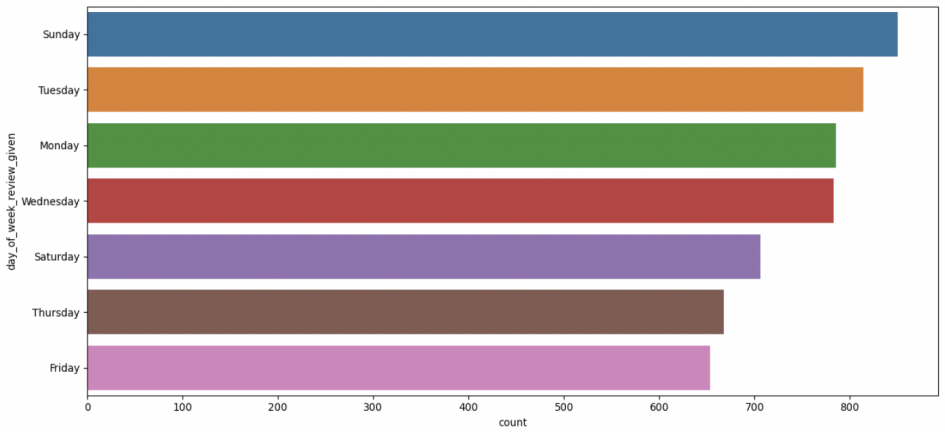
Attachment 19 - Density plot split by trip types in Portugal (solo = pink (shaded), friends = red, family = green, couple = blue)



Attachment 20 - Density plot split by updated user location (Portugal = blue (shaded), Australia = black, Spain = pink, Canada = green, Different country = yellow, missing = purple)



Attachment 21 - Popular days by reviews given (Portugal)



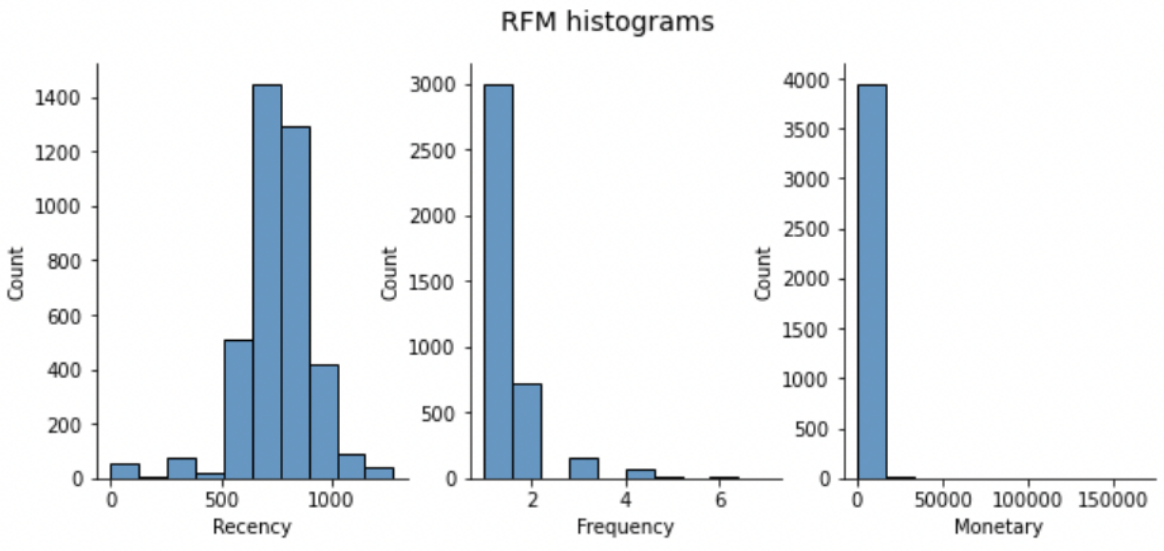
Attachment 22 - AVG rating per day of review given (Portugal)

day_of_week_review_given	Friday	Monday	Saturday	Sunday	Thursday	Tuesday	Wednesday
Country							
Portugal	4.401225	4.394904	4.432011	4.484706	4.425150	4.455774	4.450830

Attachment 23 - AVG rating per day of review given (Portugal, Spain, France, Italy)

day_of_week_review_given	Friday	Monday	Saturday	Sunday	Thursday	Tuesday	Wednesday
Country							
France	4.459641	4.510476	4.445190	4.473445	4.427451	4.382906	4.458716
Italy	4.632184	4.632236	4.684524	4.659229	4.654777	4.655819	4.625626
Portugal	4.401225	4.394904	4.432011	4.484706	4.425150	4.455774	4.450830
Spain	4.600734	4.575082	4.617228	4.616027	4.580573	4.605664	4.623188

Attachment 24 - Distribution histogram of the RFM model



REFERENCES & SOURCES

Main reference points from the report

1. https://en.wikipedia.org/wiki/COVID-19_pandemic_in_Portugal#:~:text=On%202%20March%202020%2C%20the,19%20was%20reported%20in%20Portugal.
2. https://www.schengenvisainfo.com/travel-guide/top-10-most-visited-european-countries/
3. https://towardsdatascience.com/joining-datasets-with-pythons-pandas-ed832f01450c
4. https://tools4dev.org/guide/how-to-choose-a-sample-size/
5. https://english.elpais.com/society/2020-03-16/spain-closes-its-borders-to-contain-coronavirus.html

Other sources

1. https://towardsdatascience.com/4-methods-for-changing-the-column-order-of-a-pandas-data-frame-a16cf0b58943
2. https://stackoverflow.com/questions/17709270/create-column-of-value-counts-in-pandas-dataframe
3. https://towardsdatascience.com/joining-datasets-with-pythons-pandas-ed832f01450c
4. https://stackoverflow.com/questions/36653419/str-contains-to-create-new-column-in-pandas-dataframe
5. https://stackoverflow.com/questions/26577516/how-to-test-if-a-string-contains-one-of-the-substrings-in-a-list-in-pandas
6. https://seaborn.pydata.org/generated/seaborn.lineplot.html
7. https://pandas.pydata.org/docs/reference/api/pandas.io.formats.style.Styler.background_gradient.html