Project: Data Science for Marketing

Group Number 40

Group Members:

Emna Bouassida_20221740

Ludovico Toscano_20220044

Marta Manevska_20220056

Antonina Filatova_20221104

Contents

National Tourism Promotion, Group Project

Group 40

CRISP-DM. Business understanding. Goals and objectives

Goal

Provide business recommendations to National Tourism Board Organizations (NTBO) of Portugal in order to increase tourism activity in Portugal.

Objectives

- Characterize and describe the patterns of visitants of Portuguese attractions
- Compare Portuguese attractions to main tourism competitors
- Identify changes in touristic behavior caused by pandemic

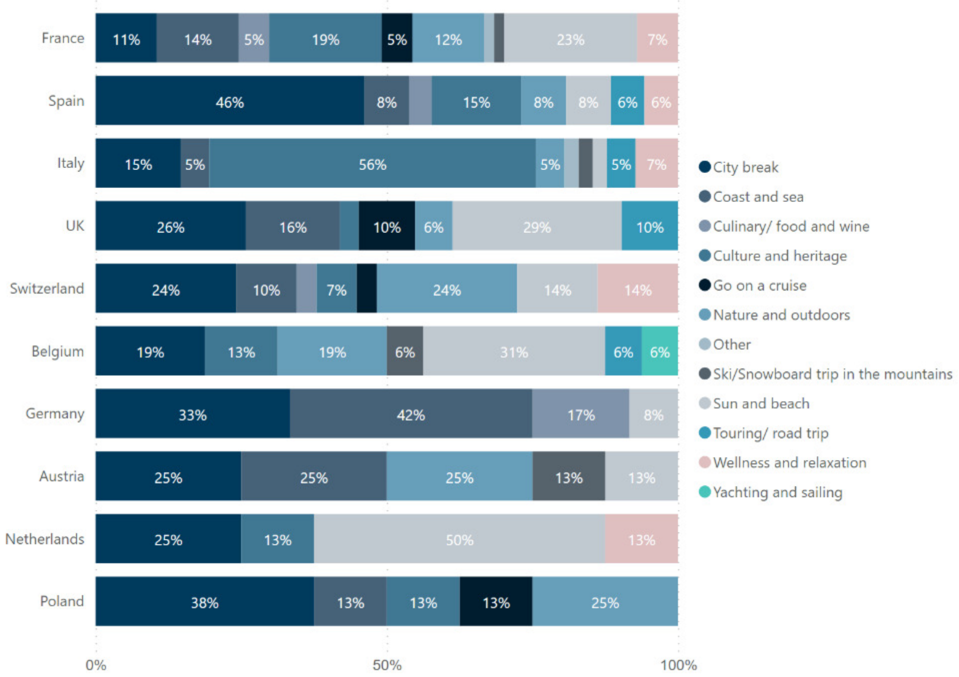
Data provided: dataset of reviews published in TripAdvisor from January 1st, 2019, to August 21st, 2021, in English, for the top 100 tourist attractions in Europe.

Research on tourism in Portugal

Business understanding

About tourism in Portugal

We tried to see if there are countries competing with Portugal, and we found that the competitions are happening in different **niches**. For example, there are people that travel to experience food, others that travel to experience surfing, and so on. So, we should focus on niches, and not on countries. Which is fine because every attraction is competing with **similar attractions** all around the world.



Source: TravelBI-Turismo de Portugal

We have also found a table thatshows the country of origin of people that visited Portugal the most, as well as the number and the percentages of these visitors. This can be useful in order to see if there are **patterns** between these groups of people.

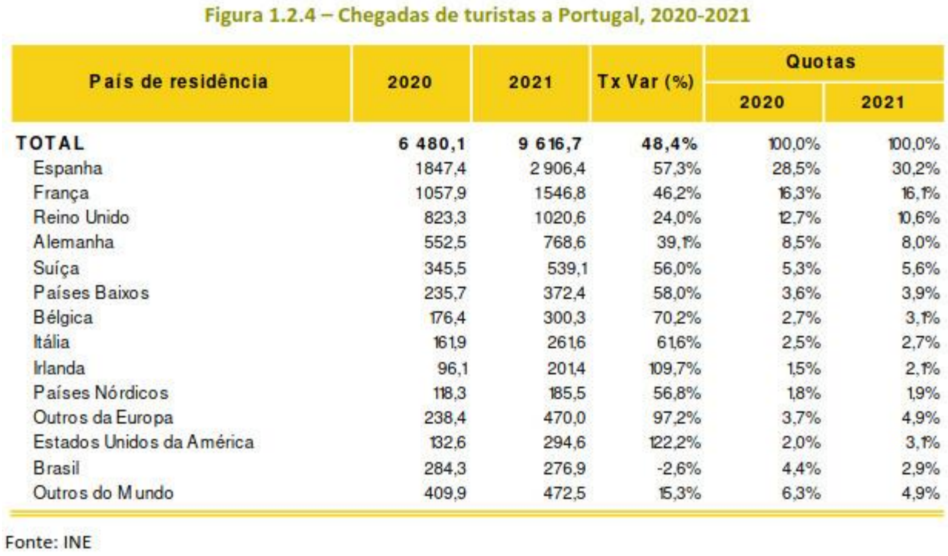


Figura 1.2.4 – Chegadas de turistas a Portugal, 2020-2021

País de residência	2020	2021	Tx Var (%)	Quotas	
				2020	2021
TOTAL	**6 480,1**	**9 616,7**	**48,4%**	100,0%	100,0%
Espanha	1847,4	2 906,4	57,3%	28,5%	30,2%
França	1057,9	1546,8	46,2%	16,3%	16,1%
Reino Unido	823,3	1020,6	24,0%	12,7%	10,6%
Alemanha	552,5	768,6	39,1%	8,5%	8,0%
Suíça	345,5	539,1	56,0%	5,3%	5,6%
Países Baixos	235,7	372,4	58,0%	3,6%	3,9%
Bélgica	176,4	300,3	70,2%	2,7%	3,1%
Itália	161,9	261,6	61,6%	2,5%	2,7%
Irlanda	96,1	201,4	109,7%	1,5%	2,1%
Países Nórdicos	118,3	185,5	56,8%	1,8%	1,9%
Outros da Europa	238,4	470,0	97,2%	3,7%	4,9%
Estados Unidos da América	132,6	294,6	122,2%	2,0%	3,1%
Brasil	284,3	276,9	-2,6%	4,4%	2,9%
Outros do Mundo	409,9	472,5	15,3%	6,3%	4,9%

Fonte: INE

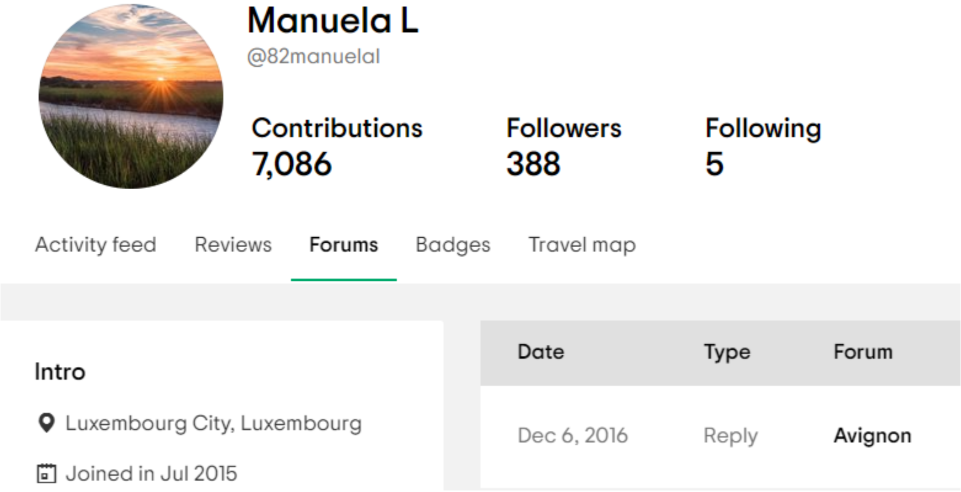
Source: Instituto Nacional De Estatística

Trip advisor

First, we tried to understand why we have **outliers** in the user contribution field. The first thing that we did was going to a page where you can find explanations about how the platform works.

Although it was very useful, the information still wasn't enough for us. So, we went to the page of **the annual reports**, to see if there was something about the outliers. With the report that we found, we were able to track the growth of the company in terms of reviews, number of attractions and so on. Furthermore, we could find information regarding the growth of TripAdvisor in terms of reviews, activities and attractions, which could be useful for our analysis.

Then we moved to another investor file, and we discovered that , on February 1st 2022, the user with more **reviews** was a girl @82manuelal from Luxenburg with more than 7000 reviews. This result was contradictory to another discovery that we made while exploring the dataset, which states that the highest number of 'UserContribution' goes to a user named 'Neil K'.



Manuela L

@82manuelal

Contributions	Followers	Following
7,086	388	5

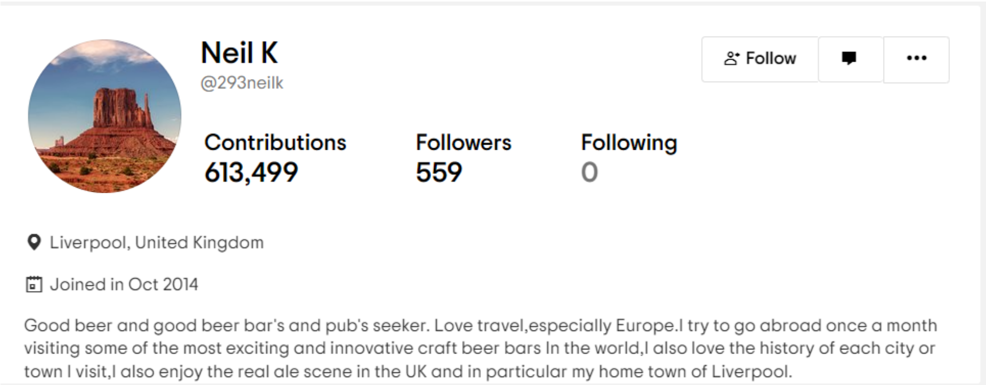
Intro

Luxembourg City, Luxembourg

Joined in Jul 2015

Date	Type	Forum
Dec 6, 2016	Reply	Avignon

This made us check both TripAdvisor user profiles, which led us to a very important insight into the 'UserContribution' variable's definition. We discovered that contrarily to definition 'userContributions - how many reviews have the user wrote in TripAdvisor at the moment of the extraction of the review' from the project file, 'userContribution' in fact shows **all activity of the user**, including forum posts, ratings and reviews. In our dataset a user with username *Neil K@293neilk* has the highest 'userContribution'.



Neil K

@293neilk

Contributions	Followers	Following
613,499	559	0

Liverpool, United Kingdom

Joined in Oct 2014

Good beer and good beer bar's and pub's seeker. Love travel,especially Europe.I try to go abroad once a month visiting some of the most exciting and innovative craft beer bars In the world,I also love the history of each city or town I visit,I also enjoy the real ale scene in the UK and in particular my home town of Liverpool.

Heading to TripAdvisor.com we found the profile of this user and discovered that he has been very active on the platform, posting dozens of photographs and couple of reviews daily, thus he has reached 613K of contributions since the **extraction**

date of our dataset! This user has been posting repetitive content, and sharing the same pictures, nevertheless, we made sure that it was a real user, not a bot.

So, finally we found a post on the trip advisor forum regarding the **user contributions.** What we discovered is that these users are not bots because the posts on the forum, photos and other types of content are counted as contributions.

Here are some statistics that we found during the research on the investor annual report:

TripAdvisor **was launched in Europe** in 2005/2006

2006: 5 million reviews, 220000 hotel and attractions

2007: 17 million reviews

2011: 60 million reviews, 20 million member, 900000 restaurants and attractions

2012: 100 million reviews, 1,3 restaurant and attractions avg 30 million reviews this year 44 million members

2013: 125 million reviews, 400000 attractions

2014: 200 million reviews, 500000 attractions

2015: 320 million reviews, 625000 attractions

2016: 500 million reviews, 760000 activities and attractions

2017: 600 million reviews, 915000 activities and attractions

2018: 730 million reviews, 1 million activities

2019: 859 million reviews, 1,2 million activities

2020: 884 million reviews, “activities” is missing

2021: 1 billion reviews, “activities” is missing

CRISP-DM. Data Understanding

Variables

The original dataset provided includes the following variables explanations:

- **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

Some discoveries and modifications will be brought up to certain variables during the Data Modeling phase.

userLocation

We observed that the userLocation variable was not **formatted** in the same way for each entry. For example, most locations contain two parts: the name of a **city** and a name of the **country**, like **"Malaga, Spain"**, however those who related to the US, as a second part have the **code of the state**, instead of a country, i.e "Atlanta, GA", and some entries have numerous parts, i.e **"London, England, United Kingdom"**. The data of the userLocation is not a **mandatory field f**or the user to give this information in the Trip Advisor platform and this is why this section appears uncomplete – there are 78 thousand rows out of 92 thousand rows overall data.

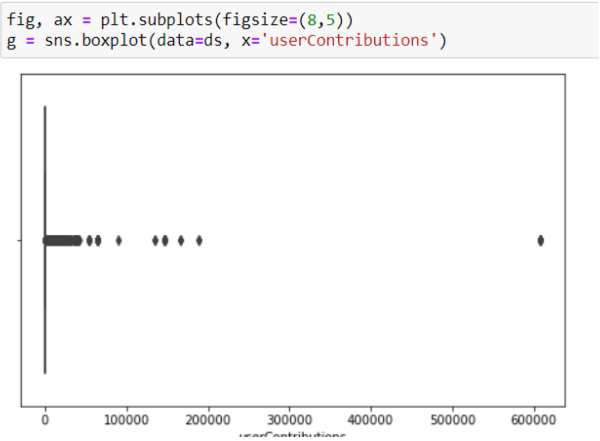
TripType

The variable TripType is available in 63 thousand entries out of 92 thousand rows, being **not mandatory** for the users to fill as well. There are five types of trip types: **friends, family, couple, solo, and business.** On a more advanced step, we will perform analysis on the tripType variable.

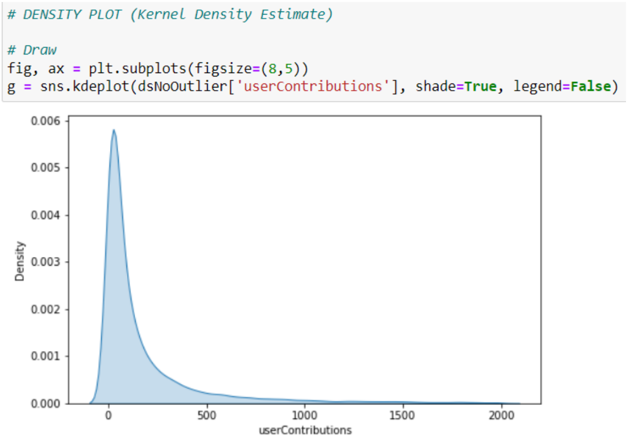
UserContribution

Our next challenge was the variable 'userContribution' with the **maximum value of 607732**. This value looks suspicious, even considering that this user ('Neil K') has been using TripAdvisor since October 2014, (according to his TripAdvisor Profile) and our data was extracted in August 2021; such contributions imply an extraordinary number of reviews **per day,** up to 86819 reviews per day.

In the beginning, we suspected it to be a possible **bot activity** and attributed this issue to **outliers** of the dataset. In our first iteration, based on **Kernel density plot** and **box plot visualization,** and by relying on commonsense we set a **threshold of 2000 contributions,** just to visualize the data distribution. We ended up deciding not to remove it, even after discovering that there's no bot activity.



```
fig, ax = plt.subplots(figsize=(8,5))
g = sns.boxplot(data=ds, x='userContributions')
```

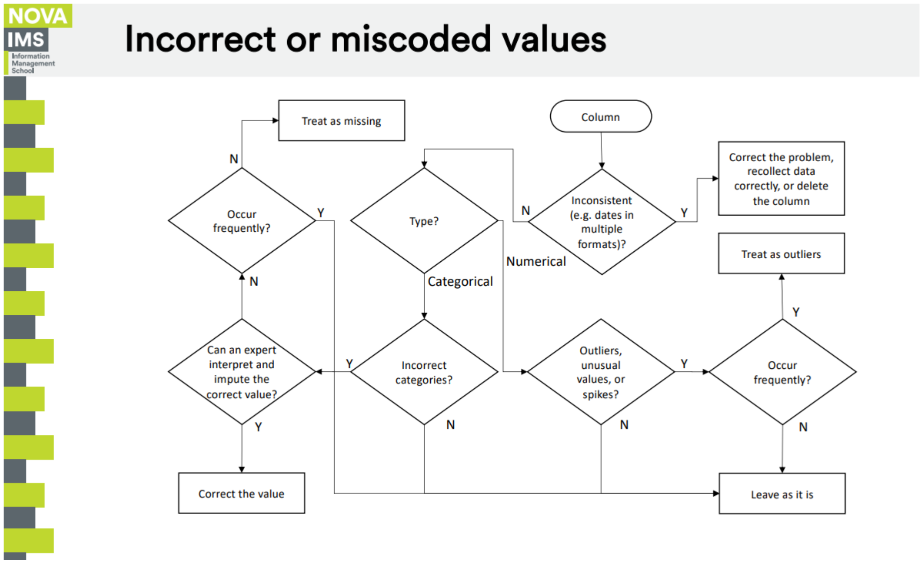


```
# DENSITY PLOT (Kernel Density Estimate)

# Draw
fig, ax = plt.subplots(figsize=(8,5))
g = sns.kdeplot(dsNoOutlier['userContributions'], shade=True, legend=False)
```

CRISP-DM. Data preparation

Our first step in the data preparation stage was to overview the dataset in **Excel**.



Incorrect or miscoded values

NOVA
IMS
Information Management School
Column
Inconsistent (e.g. dates in multiple formats)?
Y
Correct the problem, recollect data correctly, or delete the column
N
Type?
Numerical
Categorical
Incorrect categories?
Y
Can an expert interpret and impute the correct value?
Y
Correct the value
N
Occur frequently?
N
Treat as missing
Y
Outliers, unusual values, or spikes?
Y
Occur frequently?
Y
Treat as outliers
N
Leave as it is
N
N

Incorrect and missing values

We used the guide from Lecture 7 to treat incorrect data and missing values.

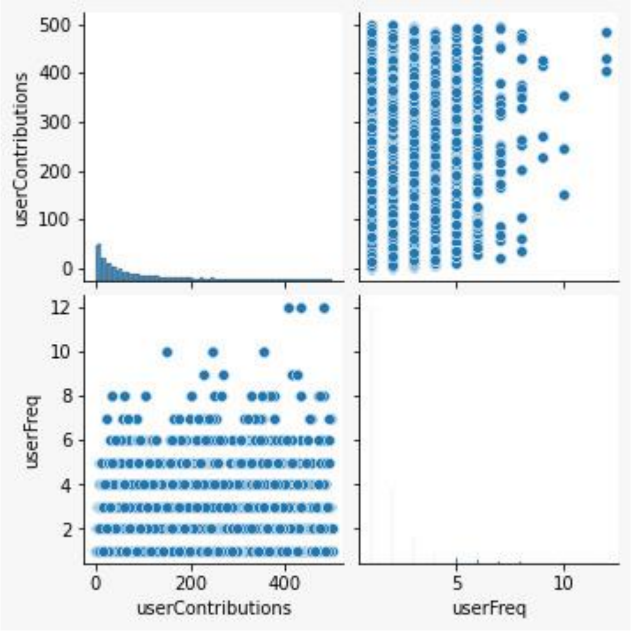
- We saw some incorrect variables as well, two entries in **'localID'** were not codified in a standard way: **'genis'** and **'u'**. These inputs were not numerous, so we chose to correct them manually: reading through review text we discovered to which attractions they belong, **Staromestske namesti** and **Edinburgh Castle** respectively.

- Next, we saw that some rows were repeated in the dataset. This is due to data being scraped, and when you use this technique, some comments might be scraped twice. So, there is no reason to keep duplicated, rows. For that our dataset was reduced from 92120 to 85068 rows.
- Later we found out that in some cases text of the review (**‘reviewFullText’**) was repeated by the same users (same ‘userName’), but the row contained different trip types, so the rows were not seen as duplicated. We finally decided to keep the second observation because the column "triptype" was filled unlike the first one that is why we deleted it.
- We saw that there were some **missing values** in variables ‘userLocation’, ‘tripType’, ‘reviewVisited’. We could not replace them, so we left the dataset as it was.
- While manipulating the ISO codes of the attractions we discovered that the attraction **“Queen Emma Pontoon Bridge”** with the ISO code NL is in Curaçao, o**utside of Europe**. This attraction had 422 entries and was removed.
- On the sheet **“Attractions”** of the Excel document the country of the attraction “Calton Hill” was defined as **“Scot”** instead of **“Scotland”.** This single entry was corrected manually using Excel. We also changed VA which is Vatican into IT which is Italy. We did this because Vatican City is small and there is no border between Italy and Vatican (the borders were closed during the pandemic). Later, for the ISO data modeling, a place in Poland with an ISO code of Croatia was changed to ISO code PL.

Creation of new variables

For further data manipulation new variables were created:

- **reviewLenght**: int – number of characters in a review.
- **same_iso**: bool – tell if the review is about the same country where the person live.
- **attractionCountry:** bool - this is about users that live in a country where there are attractions.
- **averageRating:** float - is the average of all the ratings for the given user.
- **diffAvgRating:** float - the difference between the average of the user and the rating of a specific review.
- **holiday**: bool – true, if the review is made within three days before or after a holiday in the same country as userLocation.
- **userFreq**: int - number of reviews present in the dataset made by a user.
- **attractionFreq:** int - sum of reviews per attraction in the dataset.
- **clusterFreq**: int - sum of reviews per cluster in our dataset.
- **beforeAfter**: str - 'reviewVisited' with ‘datetime’ earlier than '2020/03/01’ is assigned to before.
- **userISO**: - str - this is the ISO code of th.e place where the user live.
- **localISO**: - str - this is the ISO code of the attraction.
- **cluster**: - int- this is the number of the cluster which the attraction belong too.



Correlation between "userContributions" and "userFreq"

The graph above shows that there is a low correlation between userContributions and userFreq, meaning that one variable can't predict the other one.

Before and After pandemic subsets

In order to address the question about the change in tourist behavior due to Covid-19 in further analysis, the dataset was divided into **two subsets**. **March 2020** was chosen as the starting month of the pandemic since most European countries introduced restrictions in March. Rows with 'reviewVisited' with 'datetime' earlier than '2020/03/01 00:00:00' were assigned to "Before" subset and later dates formed "After" subset. However, we need to mention that "After" subset stands for "after start of the pandemic" as it includes data from March 2020 untill August 2021, which technically is a period during pandemic.

CRISP-DM. Data modelling

UserISO variable

In order to understand **travelling behavior**, it is important to know the **home countries** of the tourists as the travelling choices are influenced by **culture** and by **distance** of the destination. This part of data modelling is dedicated to formatting the home countries (**userLocation** variable). As discussed earlier, the initial format of this variable was not unified. It was decided to choose **ISO code as a unified format** for userLocation and to attribute codes for each row. To complete this data transformation two libraries were used: **pycountry** and **Nominatim api** by geopy.geocoders over several iterations.

First, we subset the data frame to 'userName' and 'userLocation' columns and then split the 'userLocation' in order to have the last part of the entry, indicating the country. In case the last entry was composed of two characters, it either stood for ‘UK’ and then it was left as it is, and for a code of the American state a US code was attributed. Otherwise, **pycountry** library was used to generate ISO codes. Some values were not recognized so the second iteration of search was performed using **Nominatim API,** which is a more extensive database. **Nominatim,** however, returns the address including name of the state (and not ISO), so to the newly recognized values in the list the final iteration of **pycountry** was applied. By the end of this operation the list of ISO codes standing for userLocation was obtained, which was used for further analysis. The variable **userISO** was created.

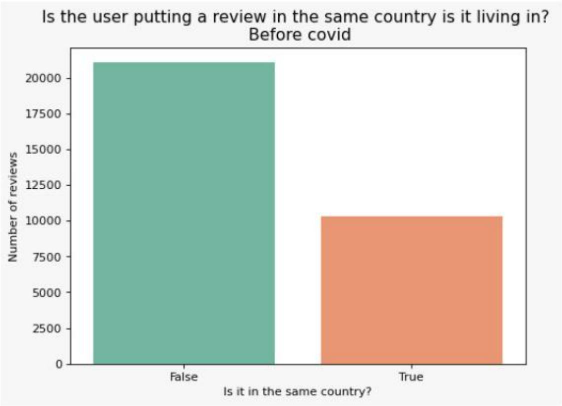
Analysis based on userISO

Having created userISO in **ISO code format,** we were able to test the hypothesis about **tourist behavior** that we developed during the business understanding and data understanding stages. Since European countries have introduced travel restrictions and maintained closed borders during the pandemic, it is suggested that they switch to **domestic travel** and the number of international trips has reduced. **The language of all reviews in the dataset is English**, therefore, it is suggested that **most of the users in the dataset are English-speaking**. The most frequent country of attractions reviewed is the UK, it is suggested that most of the users are from the UK.

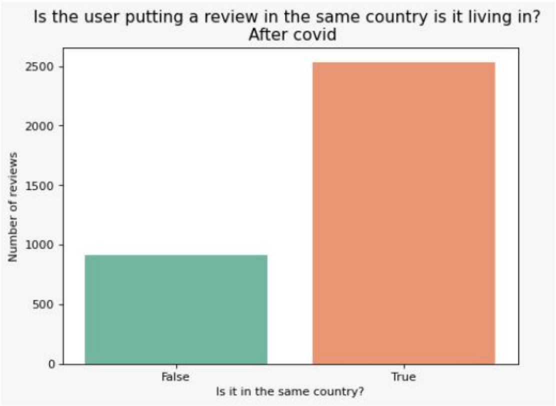
- H1: Percentage of cases when **userISO = localISO** is higher in ‘After’ subset than in ‘Before’ subset.
- H2: Top countries in the most frequent userISO list are English-speaking countries.
- H3: The UK is the most frequent country of userISO.

Before and After Covid home country tourism

To test the first hypothesis a Boolean variable **“sameISO”** was created using, equal to True, when a user visited an attraction in the country where they lived. “SameISO” was defined for each data subset “Before” and “After”.



sameISO Before

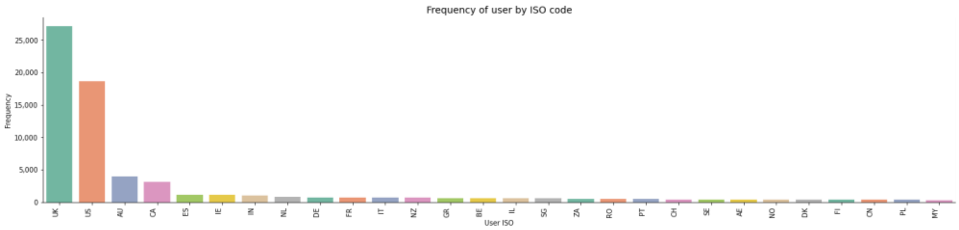


sameISO After

The graph above shows how many times users visited the home country attraction. It can be seen that before the pandemic only in 32% cases users traveled in their home countries while after the pandemic this number has risen to

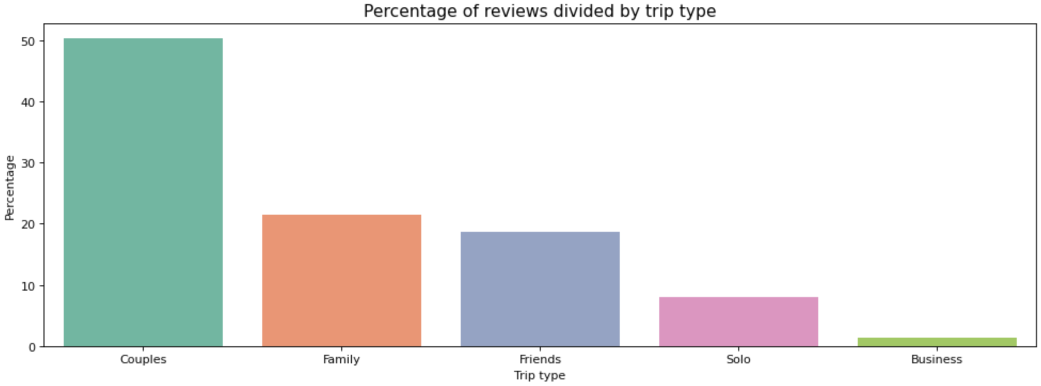
73%, which is proving H1 => **Covid-19 restrictions have changed touristic behavior in favor of traveling within their home countries.**

To test the second hypothesis the frequency for each userISO was counted.

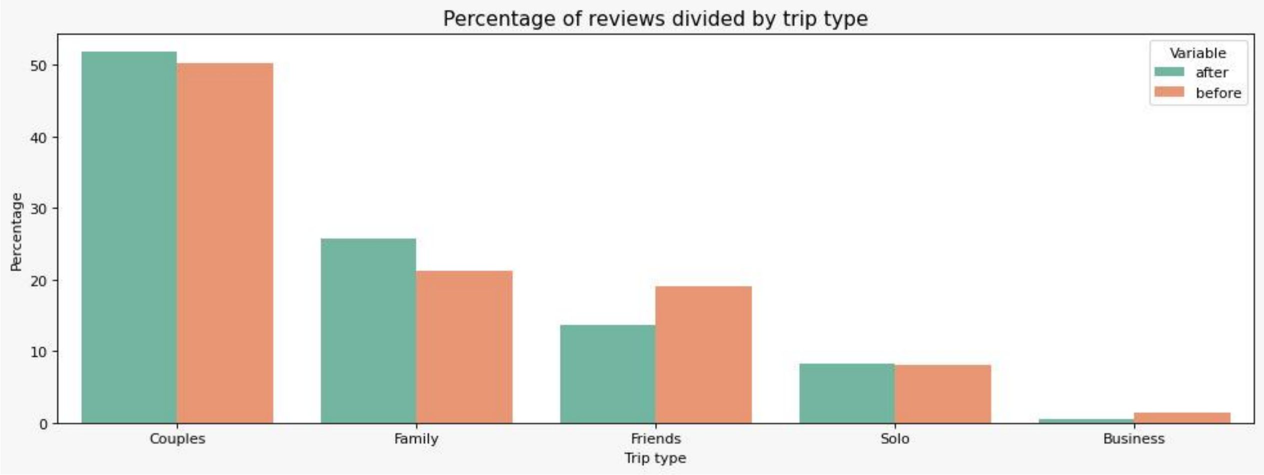


The graph above shows the top ISO codes out of 274 ISO codes (these codes appear more than 250 times) sorted by userISO frequency. Both H1 and H2 are confirmed: four countries at the **top of the list, the UK, the US, Australia, and Canada, are all English-speaking.** Accounting for 27 thousand entries, **the UK is by far the most frequent home country of the users. Confirmation of H2 and H3 stands for the fact that the dataset is biased.**

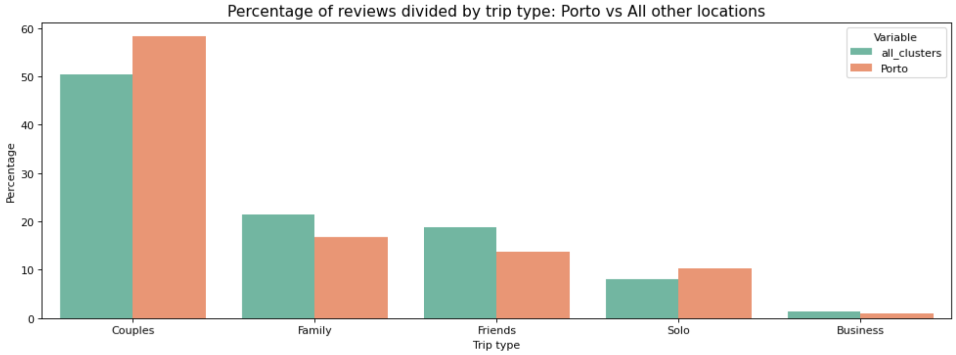
T*ripType*



The distribution of visitors' percentages per tripType in the dataset

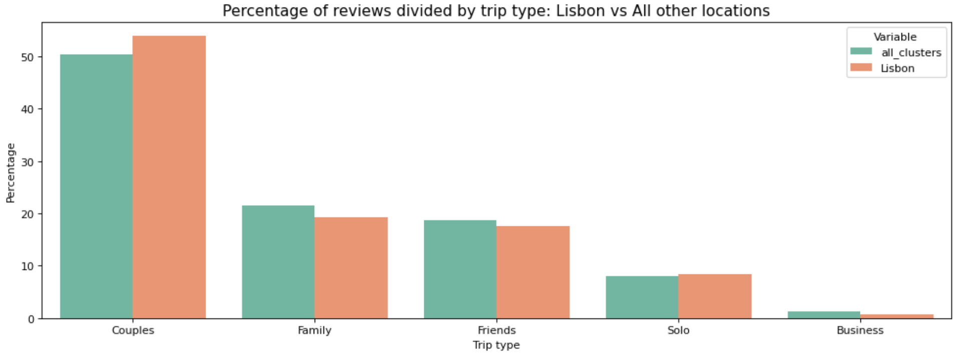


We used the variable tripType to analyze the visitors' behavior before and after COVID. We discovered that couples, family, and solo trips increased after COVID, in contrast to Friends and Business trips, which decreased considerably.



The distribution of visitors' percentages per tripType in PORTO

We can see here that couples and solo travelers go a lot more to Porto in comparison to all other clusters.



The distribution of visitors' percentages per tripType in LISBON

This graph shows that Lisbon is also a preferred destination relatively by couples and solo travelers comparing to other clusters.

The Holidays Dataset

After a thorough examination of the Holidays Dataset, we decided that its use won't be relevant for our analysis. The problem here is that we have marked as holiday dates that are within three days of a holiday (where there is holiday in the same country of the user). However, in some countries a lot of people go on holiday in August because offices are closed, and we are not measuring this data. Then some countries like France for example have holiday after a fixed number of weeks, also this is not measured. Without this data we can only say that the holiday dataset isn't enough.

The reviewLength variable

By creating the variable **reviewLenght**: int – number of characters in a review we noticed that there is a difference in the length of the reviews; the reviews after Covid were bigger than before the start of Covid. In the data 'Before' the mean was 360 characters and median - 255 characters. In the data 'After' the mean was 389 characters and the median 268 – characters, which stands for slightly longer reviews.

CRISP-DM. Data modelling. Association rules

Data clustering

When people go to visit an attraction in a city, they also go and visit another attraction in the same city, but that does not mean that they will write a review; this is why we think that clustering is important. **K-means clustering** was performed to get clusters of attractions in Europe based on **latitude** and **longitude**. K-means was not the ideal one for geographical clustering but in our case, we decided to keep it even though we ended up with having together two locations of Turkey with one in Grece and one in Bulgaria.

In order to get the coordinates of attractions we used **Nominatim api**. Some of the coordinates were identified incorrectly or were not found because the attraction name was not present in the database, we checked them manually with the help of **OpenStreetMap**, a service which is as well based on Nominatim database. After several **iterations** we were able to match every location with the correct address.

K-means was chosen over other clustering algorithms because there are 99 attractions and 25 countries the right number of k (k is the number of clusters) should be in between of those values. The result obtained was **40 clusters.** Clusters are the attractions which are located close to each other on city level.

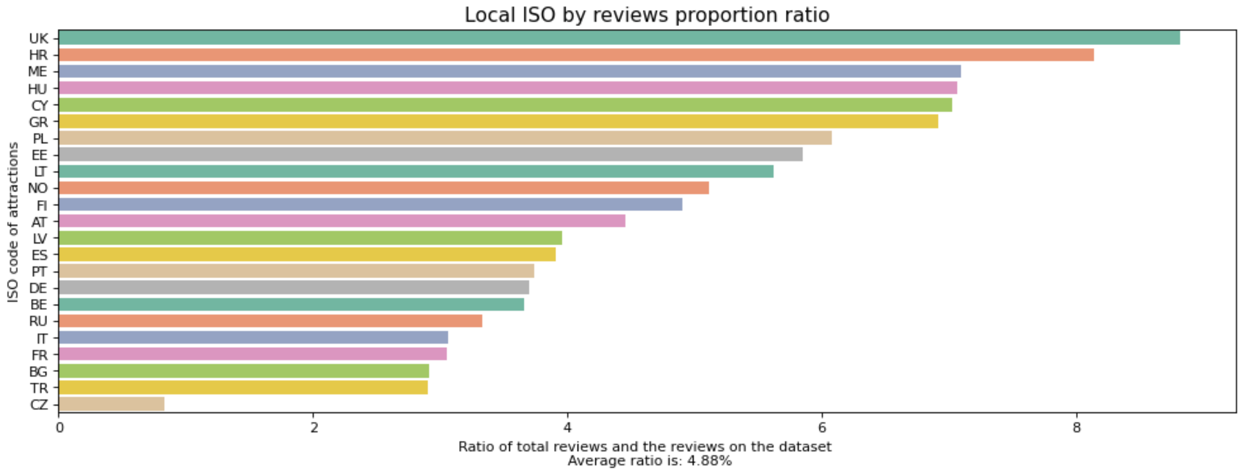
Using the clustering analysis, we can conclude that we have **two clusters in Portugal**; the first one is **Porto** and the second one is **Lisbon with Sintra.** First one is with the destinations: Funicular do Bom Jesus do Monte, Cais da Ribeira and Ponte de Dom Luís I. The second cluster contains destinations in Lisbon and Sintra which Quinta da Regaleira, National Palace of Pena, Mosteiro dos Jeronimos and Torre de Belém.

Apriori Grouping

It is important to mention that the dataset does not provide information about single trips but shows all tourists' records throughout two years. And since most of the users have one or two reviews it is not possible to attribute reviews to specific trips. Unlike a single trip where a tourist is choosing between limited options, a two-year period implies that the travels choices can be independent on different trips. If we have a limited amount of time, for example three days in a holiday, the user will visit just a few places (one instead of another) and we can use the Apriori for better understanding the frequent itemset of association, unlike having two years. Apriori algorithm is based on the assumption of a choice between limited and competing options, which is different from the tourist choosing behavior in the dataset.

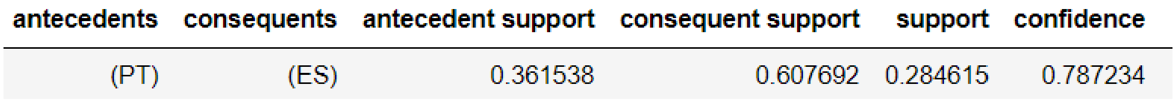
On this picture, we have the proportions per country of total reviews in the data set that were provided with, and all the data from the Trip Advisor. In the Apriori Grouping, the bias issue is due to **some countries being overrepresented**, **and some countries are underrepresented.**

We can assume from the average proportion ratio, which is 4.88%, that the UK is overrepresented, and it will always have a higher support than other countries on the entire TripAdvisor dataset. For example, Italy Is underrepresented it will have a lower support.

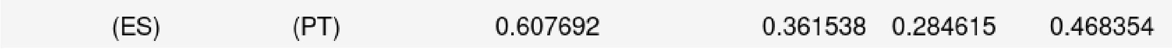


Knowing that the data is biased we must be careful. We created a dataset with users that have **more than seven reviews written** in order to get the relations between the users traveling through different countries. We have done that by using Apriori to compute the Association rule. The reason for putting seven reviews is that the more reviews per user the better the Apriori works. The lift for countries is expected to be near one and with putting less than seven reviews we get different lift values. What we can conclude from the support in the picture below that 28% of the reviewers rated both Portugal and Spain and from the confidence rule we can say that if a user has a new review about Portugal with 78% chance this user had posted (or is going to post) a review about Spain. Other way around a user with a review about Spain will post a review about Portugal with 46% probability. We can also see that Portugal appear 36% of the time when Spain is there, in the first example by the 'antecedents support'.

The results about Portugal and Spain are like this because they are not only **competitors** but also **neighboring countries**. From the **business understanding** part, we have seen that **30% of tourists came in Portugal from Spain**.

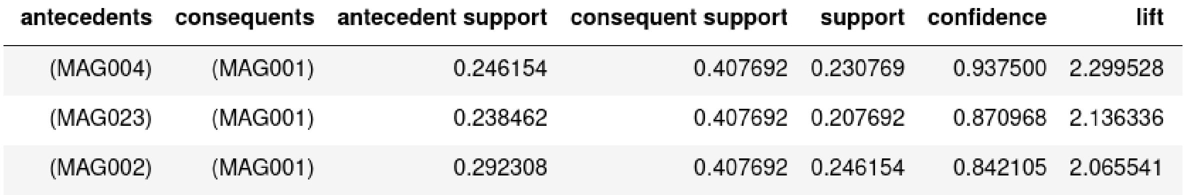


antecedents	consequents	antecedent support	consequent support	support	confidence
(PT)	(ES)	0.361538	0.607692	0.284615	0.787234



(ES)	(PT)	0.607692	0.361538	0.284615	0.468354

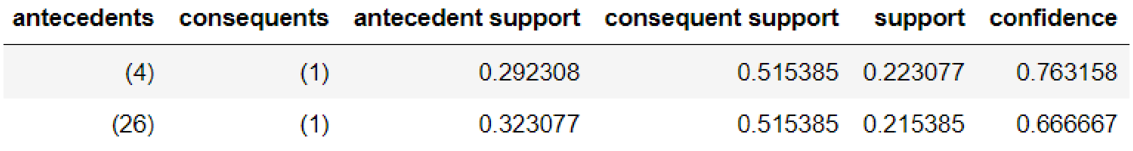
From the table below we can conclude that every time we tried to run the Apriori with the attraction we had results with the places in the same city. Our main goal was to compare Portuguese places with other places. That is why we had to subset the data and create clusters for further analysis.



antecedents	consequents	antecedent support	consequent support	support	confidence	lift
(MAG004)	(MAG001)	0.246154	0.407692	0.230769	0.937500	2.299528
(MAG023)	(MAG001)	0.238462	0.407692	0.207692	0.870968	2.136336
(MAG002)	(MAG001)	0.292308	0.407692	0.246154	0.842105	2.065541

We have **cluster number 4** which represents Catedral de Sevilla, Mezquita Cathedral de Cordoba and Real Alcazar de Sevilla in Spain and **cluster number 26** which represents Torre de Belém, Mosteiro dos Jeronimos, National Palace of Pena and Quinta da Regaleira.

Cluster 26 which are places on Portugal and cluster 1 which are places in Spain appear together 21% of the time and if people put review on cluster 26 the chances are 66% we have review for cluster 1. We discovered that people travel to places that are near to each other in a regional level.



antecedents	consequents	antecedent support	consequent support	support	confidence
(4)	(1)	0.292308	0.515385	0.223077	0.763158
(26)	(1)	0.323077	0.515385	0.215385	0.666667

Similarity

We subset the data with users that have more than 3 reviews to find the cosine similarity based on the difference between the rating of the attraction and the average rating per person, this is named **centered cosine**. We used more than three reviews per person, and we have the average of 4.5 reviews overall. The more review per user we have the more the similarity would be accurate. The solution for this is either asking the users to write more reviews, use an alternative feature or predict the feature. Since we cannot do any of these, we tried to use a threshold of 7 ratings in order to see if there will be any difference. The only change was the order of the of the similar attraction, but in this way, we can lose a lot of users' ratings.

This result should be taken carefully because we have a small number of reviews per user.

We started with **Torre De Belem** and we can conclude that people that visit La Lonja de la Seda, Villa d'Este, Canterbury Cathedral, Catedral De Burgos, Palais des Papes have put the same rating. In the list below are the Portuguese attractions and the attractions that are similar:

- **Monastero dos jeronimos:** Museum Island, Avila walls, Carcassonne, bryggen Hanseatic wharf, alcazar Segovia

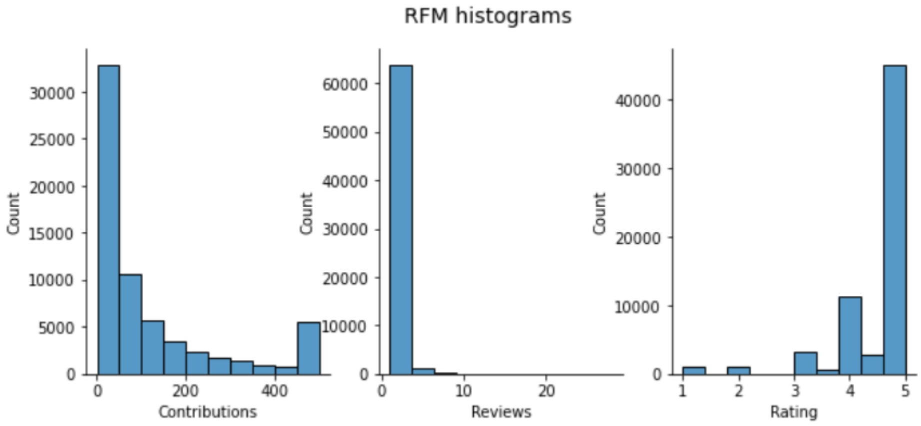
- **Ponte de dom luis**: bom Jésus do monte, Kato Paphos archeological park, Moscow kremlin, quinta de regaleira, la lonja de la seda
- **Palace of pena**: walls of Avila, cathedral de Burgos, arena di Verona, royal Albert Dock Liverpool, pont du Gard
- **Quinta de regaleira**: royal botanic garden, cathedral de Compostela, ponte de dom luis, abbaye du mont saint michel, palace of Catalan music
- **Casa da rebeira**: cathedral de Cordoba, schonbrunn Gardens, acropolis, Sevilla cathedral, red square
- **Bom Jésus do Monte**: Ponte de Dom Luis, Kato Paphos Archeological Park, Historic Center of Brugge, Seine River, Warsaw Old Town

We didn't use similarity with the cluster because **the rating cannot be aggregated**. We also tried to see if there are different similarities between users in different countries, but we don't have enough reviews for the countries.

RFM

The "RFM" in RFM analysis stands for recency, frequency and monetary value. RFM analysis is a way to use data based on existing customer behavior to predict how a new customer is likely to act in the future. It's also a great tool to segment your current customers in order to know where to invest marketing efforts and what kind of approaches to do based on the segment's importance and type. We transformed the UserContributions that are higher than 500, to a value of 500 to gain an advantage with **visualization.** At the same time **most of the distribution would not be in the first quartile in this way**.

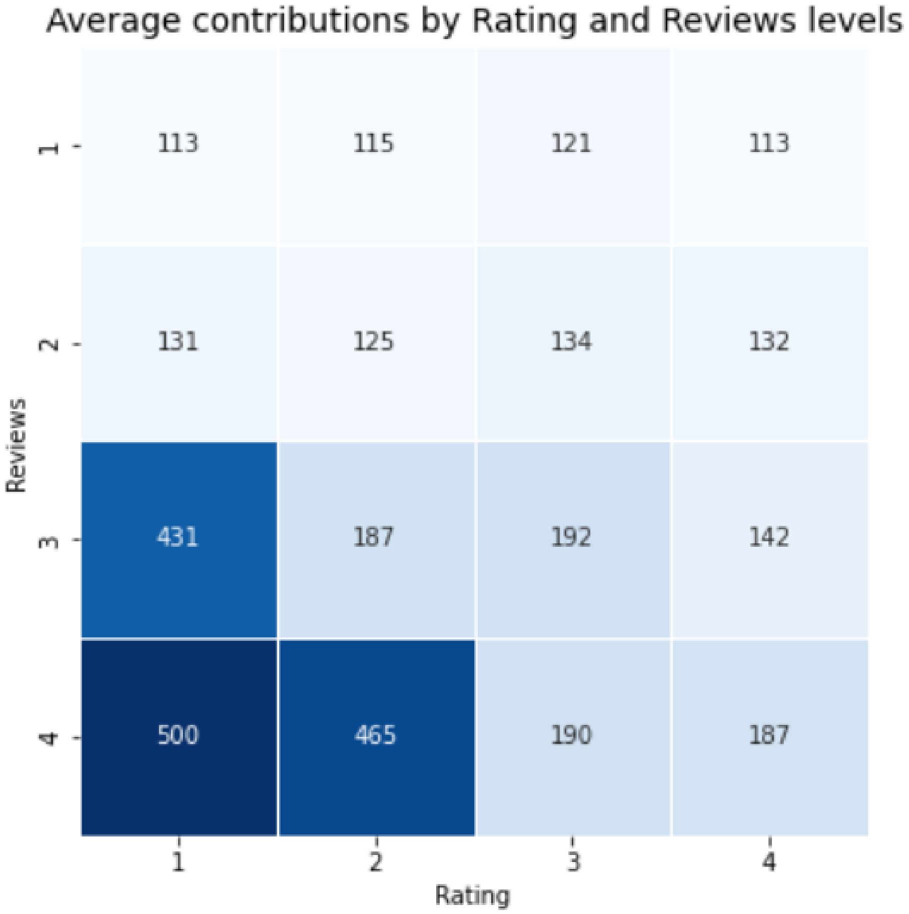
We are inspired by the RFM model using **userContributions** for the Recency, **userFreq** for the Frequency and **reviewRating** for Monetary value. We decided to use these values because the more contributions and reviews user has the more content he will put if he visits Portugal. Since User contribution is Reviews plus other types of contribution and since they are not highly correlated, we decided to keep both values. The higher the average rating value the higher the likehood to put high rating in Portugal attractions.



This is the distribution of the data that we have used for the RFM. As you can see on the histogram below the last bin of contributions have a lot of observations. This is caused by the way we dealt with outliers as we said before. We can notice that there were more users with ratings of 5, 4 and 3 stars, than users with ratings of 3.5 and 4.5. This is because 62% of the users have written only 1 review, and because the only rating options to select are discrete numbers such as 1, 2, 3, 4 or 5.

For the userContributions we decided to use quartiles, instead of hardcoding like userFreq and reviewRating. For reviewRating we have four segmentations hardcoded which are above 3.5, above 4, above 4.5 and lower than 3.5. We have done that way because most of the reviews have ratings of 4.5 and 5 stars and if we don't segment it manually every review would be in the second or the first quartile. The same is applyed for the userFreq. This is why we put the average of 3.5. The same is applied for the userFreq, since most of the people have 1 review in our dataset. For that variable we decided to use 2, 4 and 6 because there aren't a lot of people with more than 6 reviews.

The heatmap shows that people with more than 3 reviews behave in a different way to the others when it comes to rating. We can notice that the rating starts to reach its lowest values when the number of contributions exceeds 190. Keep in mind that rating 1 here is 3.5 stars which is already a good score.



CRISP-DM. Deployment. Business recommendations

Improvement of data quality. The first conclusion worth mentioning is that the given dataset is biased, which was proved trough different analysis approaches: the UK is most frequent country of user locations, and most reviews are written by users from English speaking countries, presumably since the language of all reviews is English. The recommendation for improving the quality of data analysis is to include other languages of reviews in order to balance the proportion of user countries.

Differentiating offers for tourist by trip type. It was observed that distribution of trip types inside Portuguese clusters are different from the dataset and between each other. Porto is visited by couples (58%) and solo travelers (10%) much more often than Lisbon (53% and 8% respectively). Lisbon is more popular as a destination for friends (19%) and family (17%) trips. Therefore, it is recommended to pay attention to offers for these tourist groups, e.g. launch an advertisement of Porto as a "romantic destination", make discounts for groups of young people in Lisbon (that most likely fall into 'friends' trip type) or come up with an offer of Lisbon attraction tickets for families.

Targeting tourists that like similar attractions. During the analysis the list of attractions, like Portuguese attractions were detected. It is advised to target the users that have already visited those attractions because they might be interested in visiting Portuguese attraction.

Promoting Portuguese attractions in neighboring cities Spain. According to the open-source statistics, the most of tourists visiting Portugal live in Spain. Apriori algorithm analysis confirmed that the tourists are likely to visit places that are close to each other, like attractions in Porto and Santiago de Compostela cathedral in Spain. Thus, it is recommended to advertise Portugal in Spanish bordering cities, for instance spreading cross-promotion information about Lisbon attractions in tourist points of Sevilla.

Targeting TripAdvisor influencers. Having completed RFM model analysis it is possible to segment TripAdvisor users, and the most attractive segment are the users that have the highest review number, ratings and a userContribution not higher than 190. It is advised to incentivize these users to visit Portuguese attractions and publish reviews about them, which can be done with discounts and special offers.

Annex

Annual Reports | Tripadvisor. (n.d.). Ir.tripadvisor.com. https://ir.tripadvisor.com/financial-information/annual-reports

Expedia Group. (n.d.). Www.expediagroup.com. https://www.expediagroup.com/investors/financial-information/annual-reports/default.aspx

How the site works - Tripadvisor. (n.d.). Www.tripadvisor.com. https://www.tripadvisor.com/pages/service_en.html

Lecture 43 — Collaborative Filtering | Stanford University. (n.d.). www.youtube.com

Lecture 43 — Collaborative Filtering | Stanford University

Manuela L | @82manuelal | Profile on Tripadvisor. (n.d.). Www.tripadvisor.com. Retrieved January 15, 2023, from https://www.tripadvisor.com/Profile/82manuelal?tab=reviews&fid=4ac36e46-dedc-4184-b158-b6a0a71de152

Neil K | @293neilk | Profile on Tripadvisor. (n.d.). Www.tripadvisor.com. Retrieved January 15, 2023, from https://www.tripadvisor.com/Profile/293neilk?fid=98f26a85-daeb-42fa-bbe9-5a5d99374c37

Reviews and Contributions - Tripadvisor Support Forum. (n.d.). Www.tripadvisor.com. Retrieved January 15, 2023, from https://www.tripadvisor.com/ShowTopic-g1-i12105-k7644070-Reviews_and_Contributions-Tripadvisor_Support.htm

Statistics Portugal - Web Portal. (n.d.). Www.ine.pt. https://www.ine.pt/xportal/xmain?xpid=INE&xpgid=ine_publicacoes&PUBLICACOESpub_boui=22122921&PUBLICACOESmodo=2

TravelBI | Relançamento. (n.d.). Business.turismodeportugal.pt. Retrieved January 15, 2023, from https://business.turismodeportugal.pt/pt/noticias/Paginas/travelbi.aspx

Travelers Push Tripadvisor Past 1 Billion Reviews & Opinions! | Tripadvisor. (2022). Tripadvisor. https://ir.tripadvisor.com/news-releases/news-release-details/travelers-push-tripadvisor-past-1-billion-reviews-opinions