Analysis of Tourist Arrivals in Cities Around the World

Massimiliano Bortolato

January 07, 2021

1. Introduction

1.1 Background

Tourism is travel for pleasure or business; also the theory and practice of touring, the business of attracting, accommodating, and entertaining tourists, and the business of operating tours. Increasing the number of annual tourist arrivals leads to an improvement in the economy of cities and states.

1.2 Problem

Data that might contribute to improve the number arrivals in the world cities can be land area, number of hotels, bars, historical sites, restaurant, museum. This project aims to divide the 100 cities with the high number of tourist arrivals in the world in three categories High number of arrivals, Medium number of arrivals and Low number of arrivals, create a model that can predict in which category a city belongs and give a list of the most important features that if improved can increase the number of annual tourists.

1.3 Interest

Obviously, city tourist management agencies can be very interested to have a tool that they can use to increase the number of annual tourists and improve the economy of their city.

2. Data acquisition and cleaning

2.1 Data sources

LIST OF CITIES 1: I use the "List of cities by international visitors" (https://en.wikipedia.org/wiki/List_of_cities_by_international_visitors) classified by the Euromonitor Rank. I'm intersted in the "City" and "2018 Arrivals" columns. I scrape the page and make a Pandas Dataframe. For evey city i will find Latitude an Longitude with geolocator and add this 2 columns to the DataFrame.

LIST OF CITIES 2: I use the list of city by polpulation (csv file: https://worldpopulationreview.com/world-cities) and make a Pandas data frame.

LIST OF CITIES 3: I use the list of city by area (http://www.citymayors.com/statistics/largest-cities-area-125.html). I scrape the page and make a Pandas Dataframe

I merge the three dataframe by the city in the LIST OF CITIES 1. the Data frame will have the following columns: City, Population, Area, Arrivals, Latitude, Longitude, Population, Land Area, Arrivals.

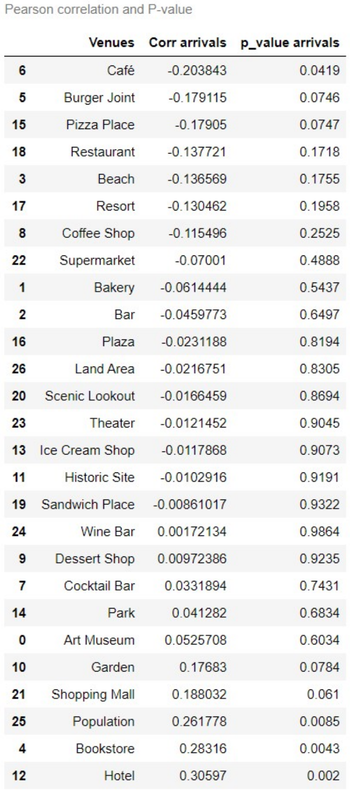
FINAL DATASET: I will add to this dataset the number of venues (restaurant, hotel, etc.) for each city, getting the data from Foursquare.

1.1 Data cleaning and feature selection

Data downloaded or scraped from multiple sources were combined into one table. There were a lot of missing values, mainly geographic and demographic data. I have to fill the missing values manually. After data cleaning and feature selection there were 100 samples and 28 features in the data.

2. Data Analysis

Pearson correlation and relative p-value were calculated, as shown in below table, related to the feature Arrivals

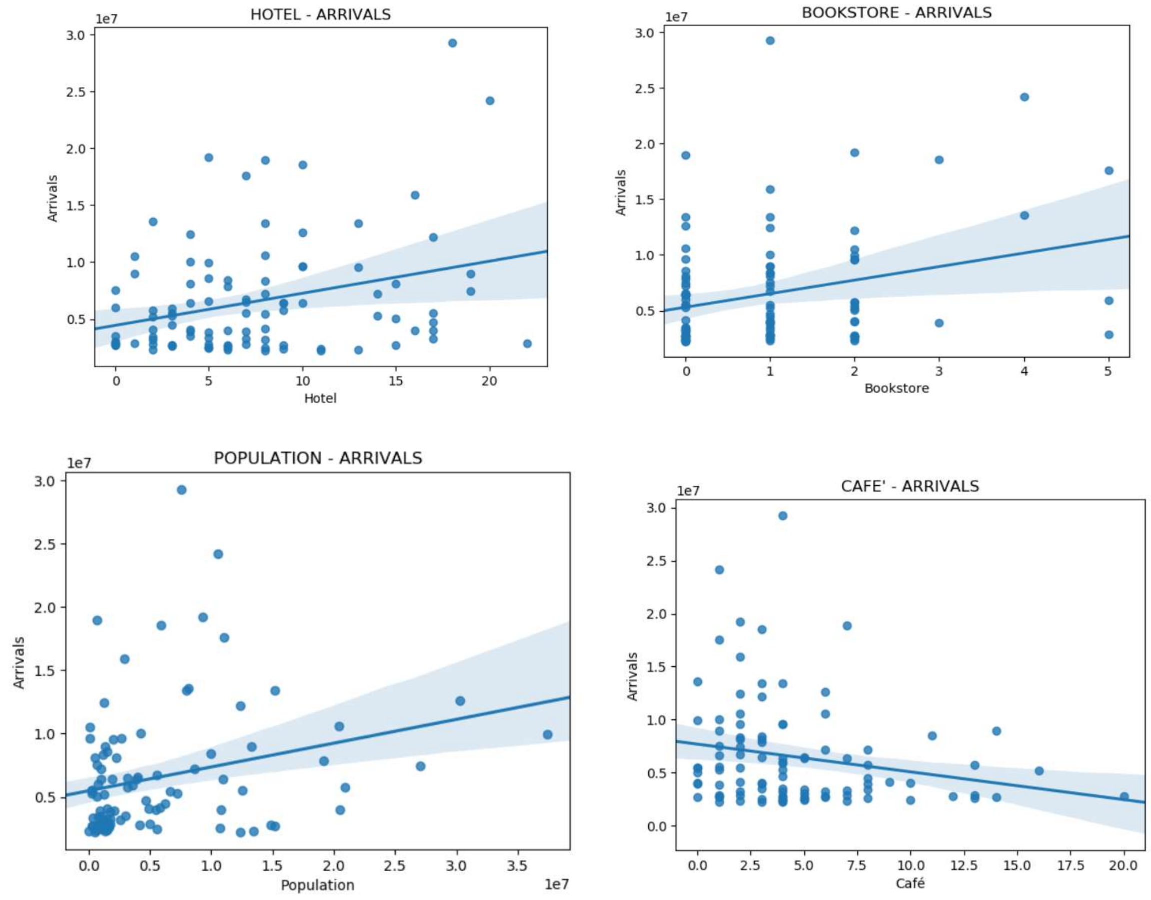


Pearson correlation and P-value

	Venues	Corr arrivals	p_value arrivals
6	Café	-0.203843	0.0419
5	Burger Joint	-0.179115	0.0746
15	Pizza Place	-0.17905	0.0747
18	Restaurant	-0.137721	0.1718
3	Beach	-0.136569	0.1755
17	Resort	-0.130462	0.1958
8	Coffee Shop	-0.115496	0.2525
22	Supermarket	-0.07001	0.4888
1	Bakery	-0.0614444	0.5437
2	Bar	-0.0459773	0.6497
16	Plaza	-0.0231188	0.8194
26	Land Area	-0.0216751	0.8305
20	Scenic Lookout	-0.0166459	0.8694
23	Theater	-0.0121452	0.9045
13	Ice Cream Shop	-0.0117868	0.9073
11	Historic Site	-0.0102916	0.9191
19	Sandwich Place	-0.00861017	0.9322
24	Wine Bar	0.00172134	0.9864
9	Dessert Shop	0.00972386	0.9235
7	Cocktail Bar	0.0331894	0.7431
14	Park	0.041282	0.6834
0	Art Museum	0.0525708	0.6034
10	Garden	0.17683	0.0784
21	Shopping Mall	0.188032	0.061
25	Population	0.261778	0.0085
4	Bookstore	0.28316	0.0043
12	Hotel	0.30597	0.002

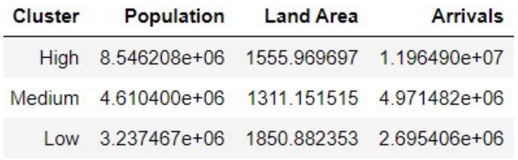
As can be seen from the table we have a week correlation for all the features. The features with the higher correlation coefficient and lower p-values is ‘Hotel’ (number of hotels). That seems correct, intuitively to a greater number of tourists will correspond a greater number of hotels.

The p-values are generally very high. That means the associated correlation coefficients did not have statistical significance. In the picture below there is the scatter plot of the features that have a p-value below 0.05.



3. Cluster Creation

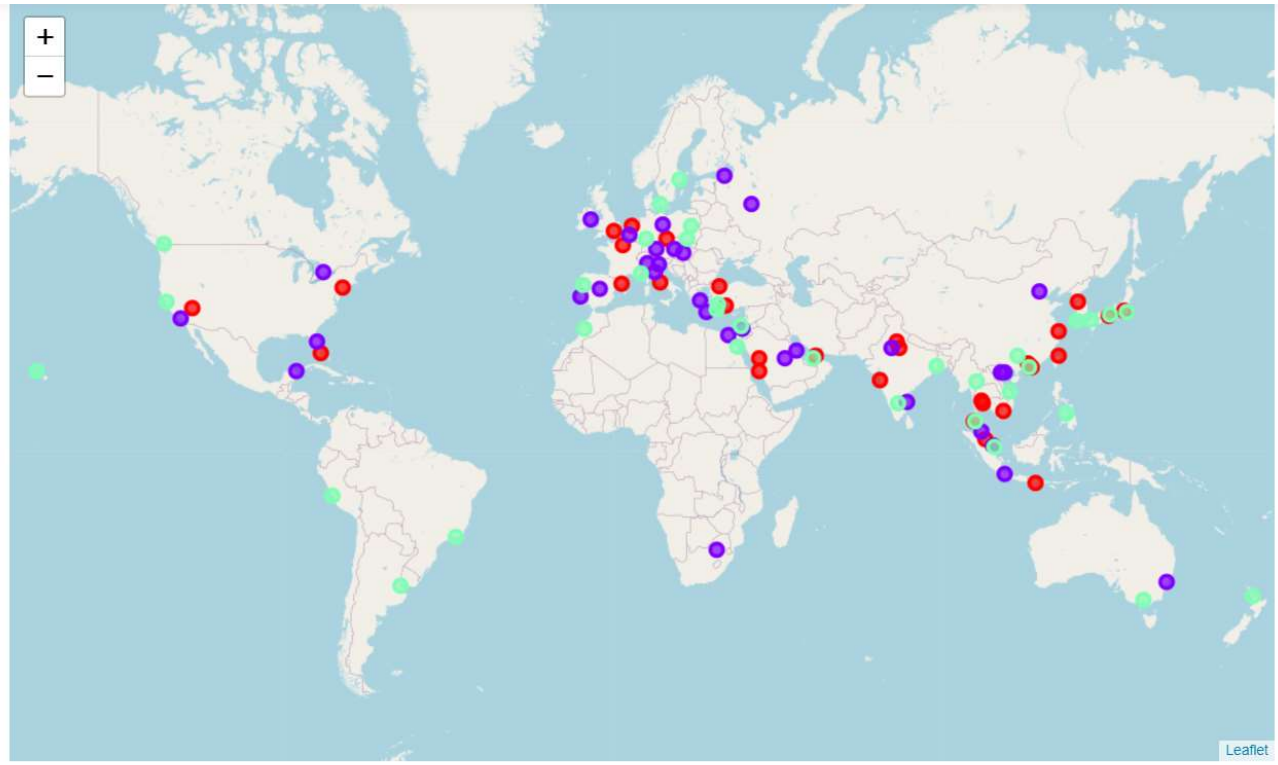
The database samples were sorted in descending order by the feature 'Arrivals' with Pandas DataFrame tool and assigned to three different clusters: High arrivals (first 33 samples), Medium arrivals (second 33 samples), Low Arrivals (last 34 samples) in order to have some sort of starting classification. In the table below the average geographic and demographic data of the three clusters.



Cluster	Population	Land Area	Arrivals
High	8.546208e+06	1555.969697	1.196490e+07
Medium	4.610400e+06	1311.151515	4.971482e+06
Low	3.237467e+06	1850.882353	2.695406e+06

The table shows that increasing arrivals and population are related: more population more arrivals. Land area is not related with the number of arrivals.

In the picture below the geographical distribution of the 3 clusters.



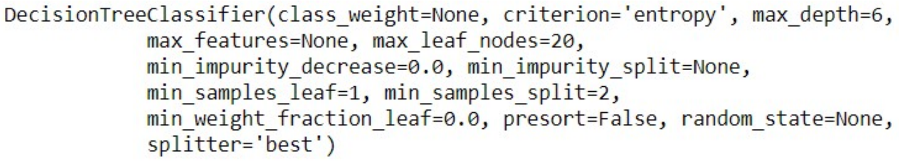
4. Predictive Modeling

Classification can be used to predict the cluster of a city. In this case a decision tree model is used because it can provide insights such as the importance of the features. The library used is sklearn.

The following steps have been done:

- Creation of features matrix and feature scaling (Z-score)
- Creation of target array
- Data splitting in train and test set (70% train set, 30% test set)
- Model creation and hyper-parameter tunning
- Model Evaluation (Accuracy on test set)
- Features with importance = 0 has been deleted

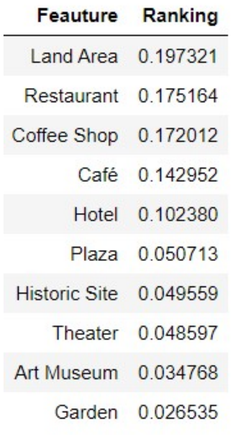
The best decision tree model has the following parameters:



```
DecisionTreeClassifier(class_weight=None, criterion='entropy', max_depth=6,
            max_features=None, max_leaf_nodes=20,
            min_impurity_decrease=0.0, min_impurity_split=None,
            min_samples_leaf=1, min_samples_split=2,
            min_weight_fraction_leaf=0.0, presort=False, random_state=None,
            splitter='best')
```

Model evaluation: ACCURACY = 0.89

The table below shows the importance of the features selected.



Feauture	Ranking
Land Area	0.197321
Restaurant	0.175164
Coffee Shop	0.172012
Café	0.142952
Hotel	0.102380
Plaza	0.050713
Historic Site	0.049559
Theater	0.048597
Art Museum	0.034768
Garden	0.026535

The features can be divided in three blocks. Land Area belong to the first block. The Restaurant, Coffee Shop, Café and Hotel belong to the second block and they are the necessary structures if you want to do tourism. The Plaza, Historic Site, Theater, Art Museum and the garden belong to the third block, they represent the attractions that are on the site.

5. Conclusions

Excluding the land area, this study shows that services (hotels, restaurants, etc.) are the structures that have the most influence on tourism numbers. The tourism management agencies they must therefore take into account that more than the attractions of the place (parks, historical sites, etc.) are not enough to increase the number of tourists, but they need to improve the services.