Capstone Project
The Battle of Neighborhoods

for the IBM Data Science Professional Certificate

by Tobias Bertschinger

tobias.bertschinger@gmail.com

15 May 2020

Table of Contents

Introduction

The selected topic for this Capstone Projects analyses top visited tourist attractions around the world. The goal is to analyse whether the venues around those attractions share common features and whether it is possible to cluster attractions according to the nearby recommended venues.

Such findings may provide helpful insights for both tourism directors of the most visited attractions and to directs of attractions which are not currently visited by a comparable amount of people. Other interested parties could be business or restaurant owners as the analysis should provide insight into what visitors expect from well known and frequently visited tourist attractions at the venues such as cafés and restaurants nearby.

Data

The website Love Home Swap published an article called “The world’s 50 most visited tourist attractions” on 1 January, 2015 (Love Home Swap, 2015). Data for this list has been gathered from various sources such as the US National Park Services or websites from the attractions. Data driven travel blog Travel Stats Man has analysed the venues and published easily accessible data tables (Travel Stats Man, 2019).

Each entry includes the attraction name, the city and country of the attraction, annual visitors in millions of visitors and the rank in the top 50 list. As an example, the most visited attraction was The Strip in Las Vegas, USA, with an annual visitor count of 39.6 million.

In a first step, I will download this data and convert it into a pandas dataframe using python. The second step will involve downloading location data for all the attractions. For this, we will use the package geopy for python. As an example, for the most visited attraction mentioned before, we find latitude = 36.2859033 and longitude = -115.0071652. Checking this location on google maps reveals that the location is somewhat off The Strip. Therefore, we will either double check locations in the second part of the assignment or use the google maps API to solve this part of the problem.

After fetching the respective latitude and longitude data for each attraction, we can then use Forsquare’s Places API. Using the venues endpoint group and the explore endpoint, we will be able to fetch nearby locations.

We will then analyse recommended places nearby attractions according to their group (e.g. coffee shop) and calculate each group’s relative frequency. This will help us to sort the most common venues near each attraction.

Finally, we will be able to cluster attractions using k-Means clustering to analyse and visualize similarity between the recommended venues near each attraction.

Methodology

We start with an analysis of the top 50 most visited attractions. As it can be seen in Figure 1: Top 50 Attractions, the top 50 attractions have annual visitors ranging from roughly 5 to 40 million. The most visited attractions is The Strip in Las Vegas. The mean is roughly 12m visitors p.a. or 1m visitor per month.

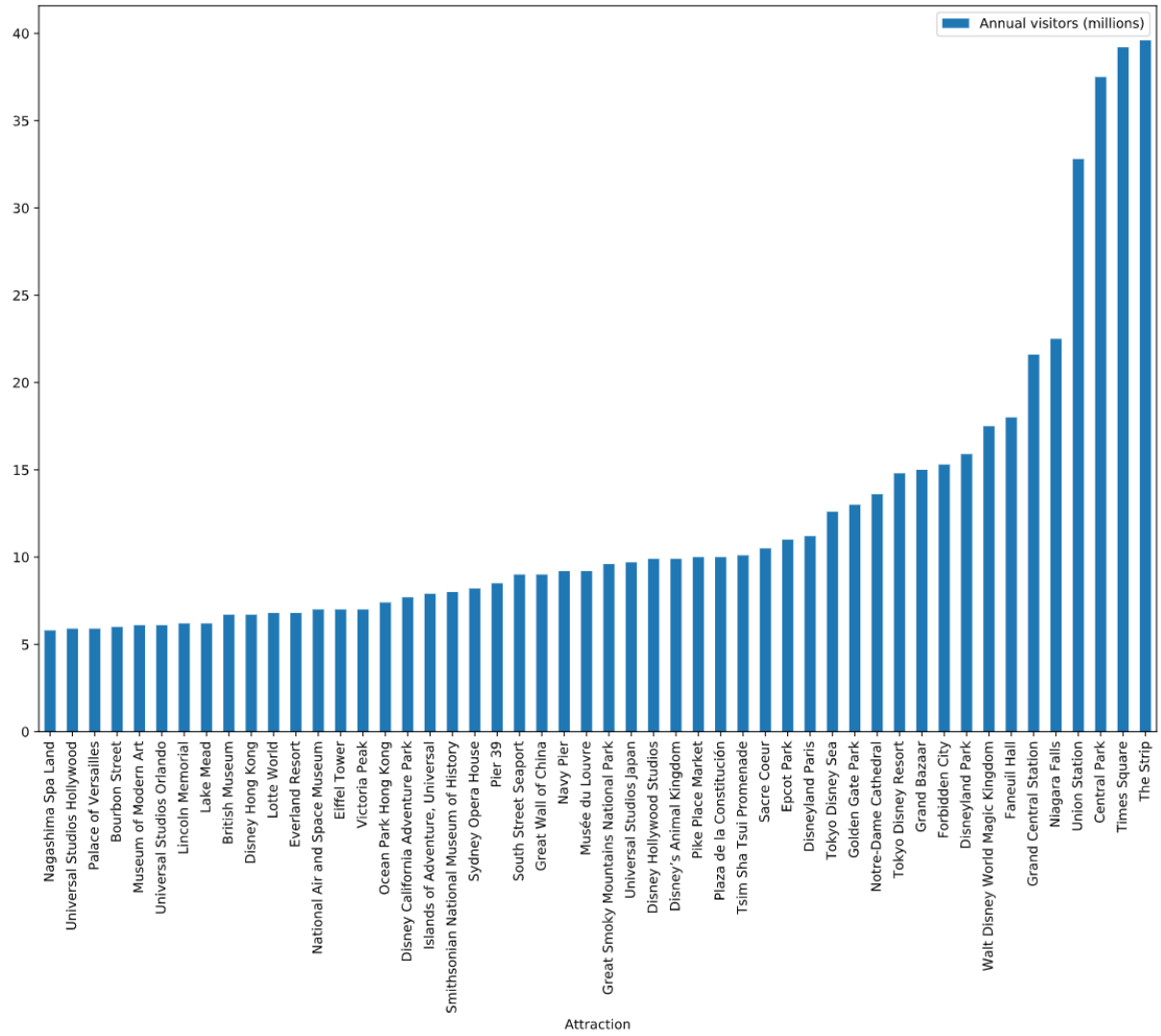

Figure 1: Top 50 Attractions

The distribution of annual visitors in millions is skewed towards lower numbers.

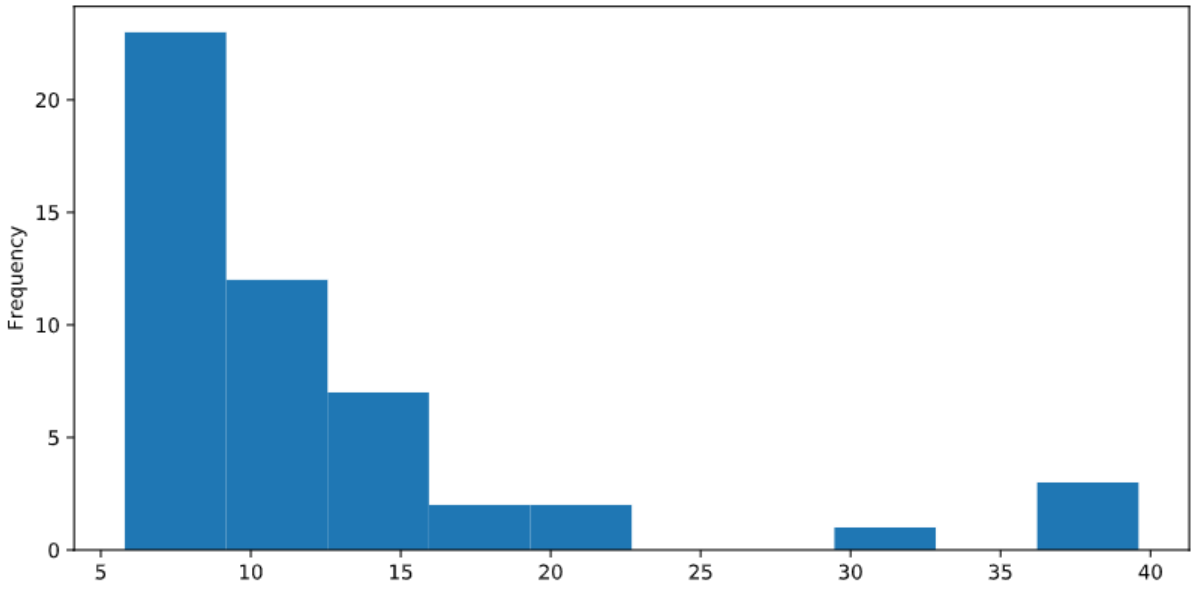

Figure 2: Histogram of Annual visitors in millions

A breakdown by country as shown in Figure 3: Top 50 by Country reveals that most of the top 50 attractions are in the US, representing 50% of the attractions in the data set. This can of course have various reasons such as a bias in collecting the data, general availability of visitor number or a definition of a sight. The strip in Las Vegas is much larger with more available activities than e.g. the Eiffel Tower in Paris.

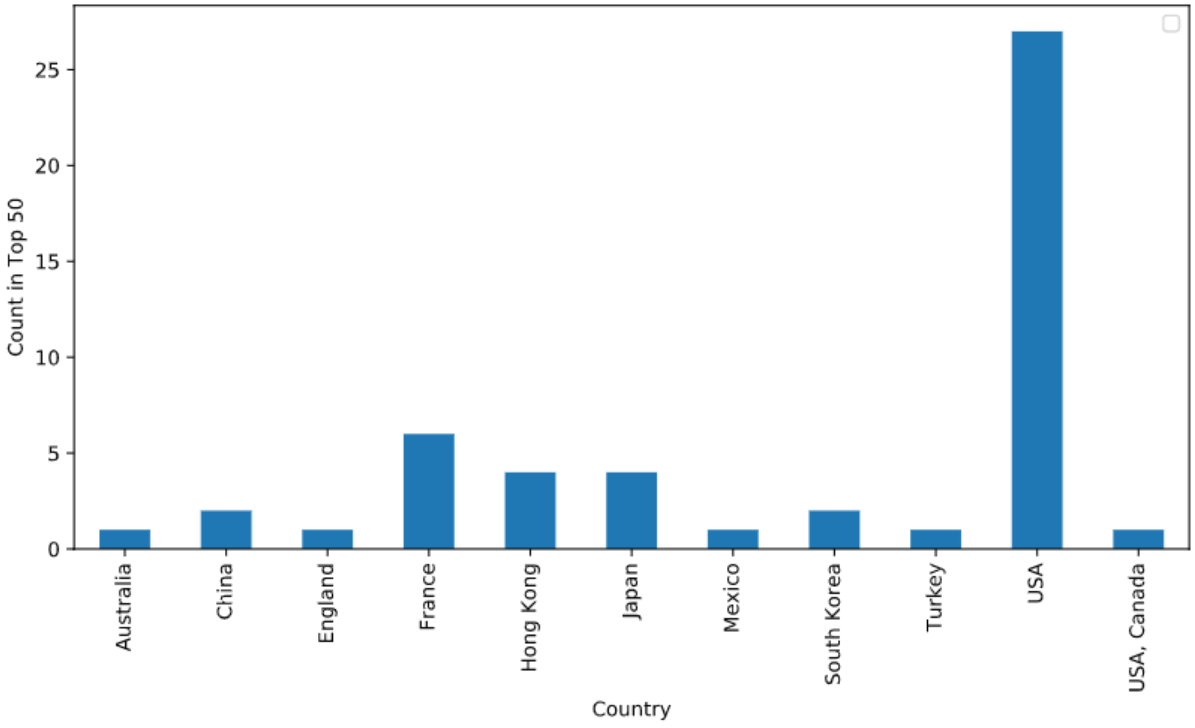

Figure 3: Top 50 by Country

Figure 4: Map of Top 50 Attractions

Using the Forsquare API, we download recommended nearby venues. Because of the different kind of locations (e.g. national park), we changed the radius for the API request for some of the attractions (see code for more details).

The downloaded venues have more than 380 different categories. Because of the huge range for e.g. different types of restaurant, we group all restaurants into one category.

We then count the relative frequency of each category type for each attraction. For example, we find that around the Eiffel Tower, 39% of recommended venues are restaurants, followed by 9% of hotels and 6% art museums.

For each attraction we calculate the top 10 most common venues.

To find similarities and differences between the top 50 tourist attractions, we use a k-Means clustering algorithm to classify attractions into one of 5 categories. We generate another world map using different colours for the clusters as it can be seen in Figure 5: Map of Clustered Top 50 Attractions.

Results

We find distinct groups based on the clustering of recommended nearby venues of the top 50 most visited tourist attractions.

Cluster A: Theme Parks such Universal Studios with mainly recommended attractions from the parks, restaurants and gift shops.

Cluster B: Large downtown attractions such as the Eiffel Tower with recommendations for mainly nearby restaurants, hotels and coffee shops.

Cluster C: Outdoors attractions such as Great Smoky Mountains NP.

Cluster D: Chinese attractions (Great Wall, Forbidden Cities).

Cluster E: Various other attractions such as Pier 39 in SF or Museums surrounded by restaurants and a wide variety of other venues.

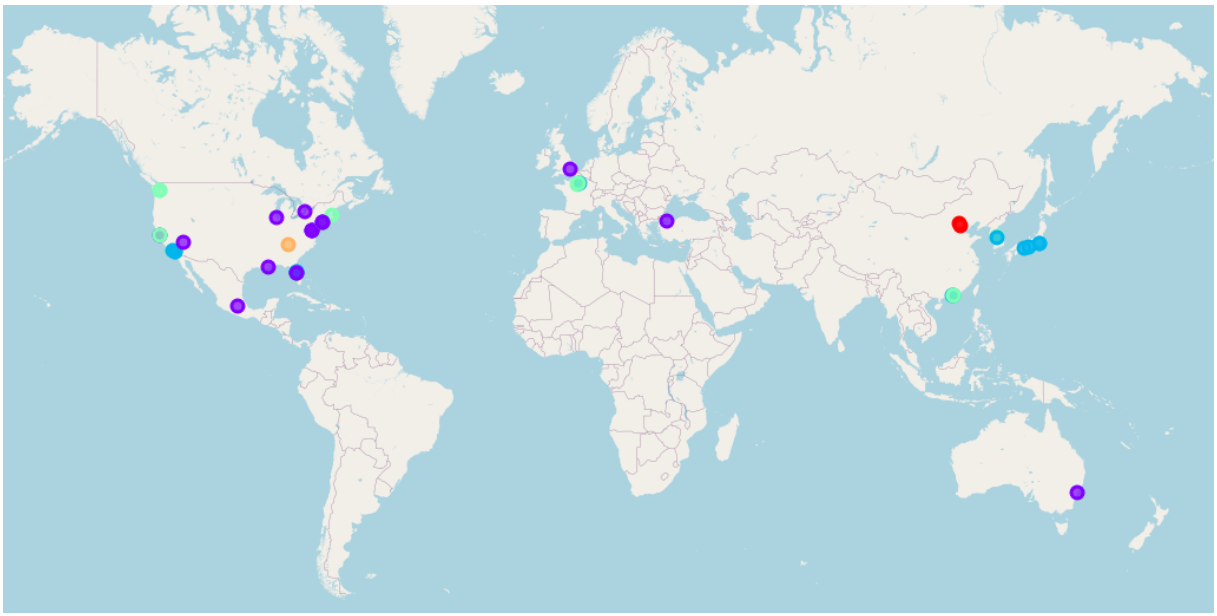

Figure 5: Map of Clustered Top 50 Attractions

Discussion

We found that even for very different types of most visited tourist attractions, we can group them into clearly distinct clusters based on their surrounding recommended venues.

However, it is rather unclear what the use of such clusters can be for e.g. people working in the tourism sector. The analysis could be used to compare attractions which don't belong into the top 50 group and to draw conclusion on what venues should be opened near a tourist attraction to make such venue even more popular. Business owners could use the results to find attractive locations for their business, because there is e.g. a lack of restaurants at a specific attraction.

For better results, one could invest more hours into cleaning the venues data and grouping venues into fewer categories. Further work could also be done on the amount of clusters leading to more granular clusters. Additionally, the data of the top 50 attractions could be challenged and broadened to include further places.

Important note: This project was done as a final capstone project for IBM Data Science Professional Certificate Capstone Project. As we had to use Forsquare data and come up with a problem by our own, it is not a real problem which I would try to solve using the process above in the real world.

References

Love Home Swap. (2015, January 1). Retrieved from https://www.lovehomeswap.com/blog/latest-news/the-50-most-visited-tourist-attractions-in-the-world

Travel Stats Man. (2019, January 7). Retrieved from https://www.travelstatsman.com/07012019/the-worlds-most-visited-tourist-attractions/