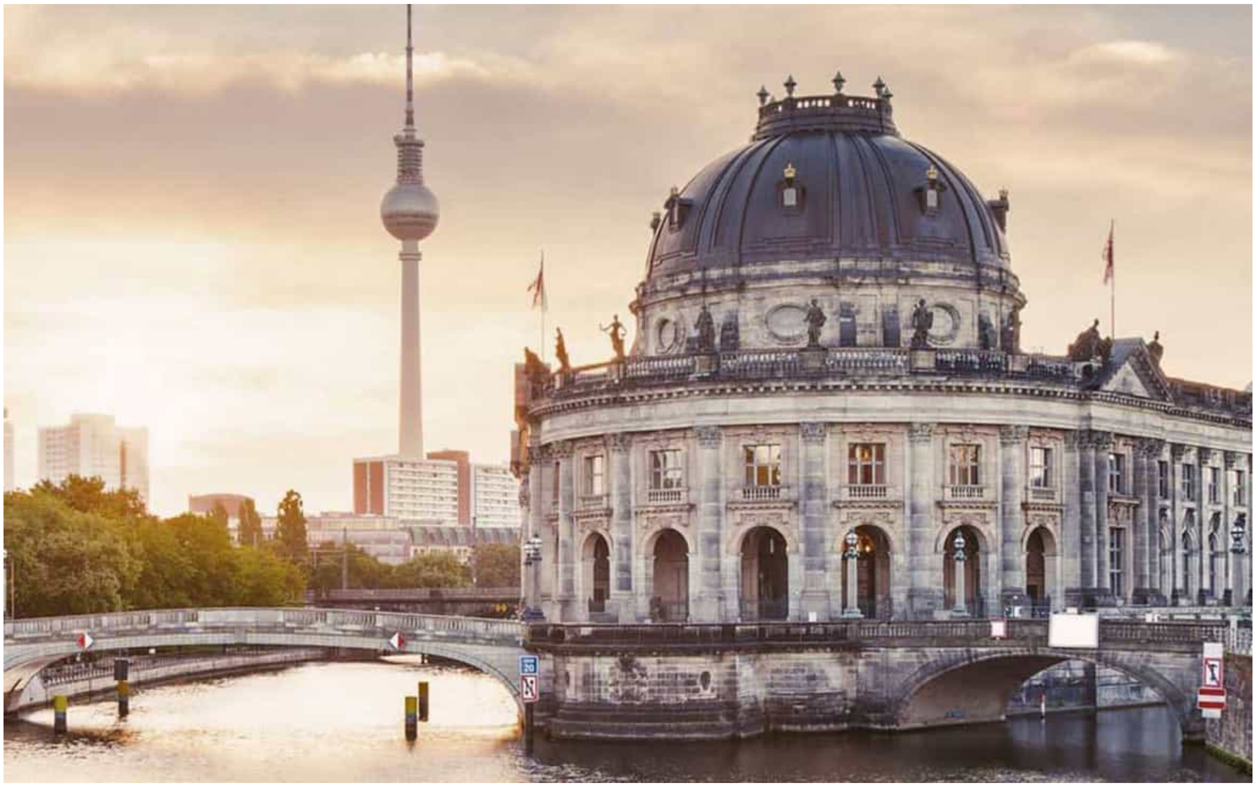
Coursera Capstone

IBM Applied Data Science Capstone

Opening a Mexican restaurant in Berlin, Germany



By: David Briones

Introduction

Everyone loves restaurants. They are great for social interaction and obviously, they are great for getting a kind of food that you cannot cook at home. Nonetheless, it turns out that restaurants are also good for urban development. They can revive an abandoned neighborhood and turn it into a hot location. They can attract young professionals to rundown areas and turn them into beautiful neighborhoods. They can bring other parts of the world closer to you. Restaurants are not merely a place to eat, but a place to congregate, interact, and enjoy life.

The Problem

Opening a restaurant is a challenging venture for many reasons. It requires a hefty capital investment, a great menu, a reliable supply chain, a food safety system, and a great location. However, in this report we will focus on one of these challenges: location. Location is incredibly important because a great restaurant placed in the incorrect area cannot possibly succeed. The problem is that choosing the right location appears to be almost impossible for people. How can you consider every aspect? This is where Data Science can come in. An entrepreneur can use machine learning algorithms to aid in the search for the right location. In this project we will consider the fictional Mexican restaurant Avo-Cado and their search for a great location in Berlin. By using a K-Means algorithm and Python programming we will help Avo-Cado’s team to choose the best location.

Data

In order to solve this problem, we obviously need data, and this is what we will need:

- A list of neighborhoods or localities in Berlin. This will allow us to limit the scope of the project to restaurant in Berlin.
- Latitude and longitude coordinates for these localities. This will allow us to properly create clusters and then map them in order to display the data visually.
- Restaurant data regarding the type of restaurant. This will allow us to adequately create clusters of localities and determine which ones have a lot of restaurants but lack Mexican food.

Sources and Methods of Extraction

A list of localities in Berlin can be found in Wikipedia at https://en.wikipedia.org/wiki/Category:Localities_of_Berlin. Using the BeautifulSoup Python package we will scrape this website and extract the necessary data. Then we will use the Geocoder Python package to find the geographical coordinates for these localities.

Once we have the necessary data, we will use the Foursquare API to get locational data regarding the restaurants in these localities. Once we have done this, we will use a K-Means Clustering algorithm to determine which one these neighborhoods are a prime location for opening a Mexican restaurant.

Data Extraction and Cleaning

The data was extracted from a list of localities in Berlin using BeautifulSoup. However, when I first put the data into a list, I noticed that a lot of these localities had either "(Berlin)" or "(Locality)" in front of them, so I used a couple of loops to get rid of these. Once the data was in a list, I turned this list into a Pandas data frame that allowed me to manipulate the data better.

I then used the data frame with all the localities to find the coordinates to each of these areas by using Geocoder. After I had both of these set of data, I merged them into one data frame that included the locality and its coordinates. Lastly, I plotted a map with folium that included the localities to make sure that it looked accurate.

Methodology

In order to determine what area in Berlin would be a good place to open a Mexican restaurant I decided there were two factors that were important. I wanted an area that had a lot of restaurants, because this meant that people go here to eat more than to other areas, but it also had to be an area that didn't yet have too many Mexican restaurants, since this would mean that competition would be low. To do this I decided to use a K-Means algorithm to group localities into similar clusters, so I could then identify the clusters that had a lot of restaurants but not many Mexican food places.

I started by using the coordinates of each locality to identify top 100 venues that are within a radius of 3000 meters from the locality. Once I had this list, I created a data frame which I then used to identify areas that had a lot of restaurants and also determine whether an area had a high number of Mexican restaurants. To do this, I added one hot encoding and I then group the rows (localities) by taking the mean of the occurrence of each category. In other words, I determined a value for each category within each respective neighborhood.

Once I had this, I plotted the squared error against the values of K, or the number of clusters, for the K-Means algorithm. This helped me determine that the most optimal value of K would be 4, meaning I would end up with 4 clusters. After this, I ran the K-Means algorithm so that I could then examine the clusters and determine the results of my project.

Results

I first took a look at the mean and median values of the full clusters. In other words, I grouped the clusters by taking the mean and median of occurrence of each category (Restaurants and Mexican Restaurants). These results showed that clearly cluster 0 was the worst cluster to open a Mexican restaurant since it had barely any restaurants. Cluster 1 was the best place to open a restaurant because of the high number of restaurants and the low number of Mexican restaurants. Clusters 2 and 3 had better results than cluster 0, but worse results than cluster 1. This conclusion was further supported by looking at each cluster by itself. Localities in cluster 1 had a lot of restaurants, but none of them had a Mexican restaurant, while most of the localities in cluster 0 had neither a restaurant nor a Mexican restaurant. Overall, I would suggest the owners of Avo-Cado to start by taking a look at the neighborhoods in cluster 1 and determine which one of those is best suited to open the restaurant.

Discussion

There are two main things that I must acknowledge, and I would like to consider during a much longer and in-depth study. First, I would like to consider whether a type of restaurant that is successful in one area is related to another type of restaurant being successful as well. For instance, I wonder if a neighborhood where Pizza places are successful is also a neighborhood where Mexican food will be successful. In this study I wasn’t able to consider this because there are a lot of categories related to food and restaurants that do not fall under restaurants, and in order to be able to consider them all, I would have to spend a significant amount of time identifying a way to group them all together. Secondly, I must acknowledge that simply because an area has a lot of restaurants, it does not necessarily mean that other restaurants will do well there. For my next project, I would like to only choose certain neighborhoods that have the right environment for the specific restaurant in question. For instance, if we want to identify the perfect place for a fast food restaurant, then we could consider areas that have a large population of young adults or students.

Conclusion

Overall, I believe that the project was successful and that we were able to identify a few areas that might be worth looking into as a location for Avo-Cado. I would recommend the owners to start by taking a look at the localities within cluster 1 and start examining which ones fit the theme and the style of the restaurant better. We could then have another study to determine which area would be the most profitable and further determine the eligibility of a locality.