CAPSTONE PROJECT- THE BATTLE OF THE NEIGHBORHOODS

Data Science Certification by IBM/Coursera

Contents

Introduction

Business Problem

"Prospects of starting a Restaurant/Bar by inspecting the Borough of Berlin"

Berlin is both a city and one of Germany's federal states (city state). Since the 2001 administrative reform, it has been made up of twelve boroughs or districts, each with its own local government, though all boroughs are subject to Berlin's city and state government. Each borough is governed by a council with five councilors and a borough mayor. The borough council is elected by the borough assembly. The borough governments' power is limited, and subordinate to the Berlin Senate. The borough mayors form a council of mayors, which advises the Senate. It's the largest city of Germany by both area and population. Its '3748148' inhabitants make it the second most populous city proper of the European Union after London. Area of Berlin is 891,1 sq km. Which is divided into 12 Boroughs.

While looking for places to open a business, we need to select the busiest zones in Berlin where a constant crowd is guaranteed. In a city like Berlin there will be a huge competition for businesses. Keeping this in mind, the surrounding of the selected Borough should not have a lot of similar businesses. Analyzing the office areas of the Borough, it is expected that there will be a lot of restaurants.

The Business Problem can be stated as:

"What is the best place to open a Restaurant/Bar in Berlin?"

Interest

Some People spend a lot of time in search for perfect opportunity and place to open a restaurant in any city. Hence understanding how good a neighborhood is for business it's one of the most important thing. As the population continues to increase, cities grow with that too with more opportunity for businesses. A decision if taken in pressure or without proper investigation and data can cause a lot of trouble specially in the loss of money. So, finding a good location for business is crucial and time-consuming process. This study will give a glimpse for selection of neighborhood for restaurant.

Target Audience:

1. The primary target audience for this project are the entrepreneurs who want to open up a new business
2. Investors who want to invest in good business ideas
3. Students who are exploring Data Science and are trying to learn the art of telling a story by training, analyzing and learning from a data

Data:

Requirements and collection

Zones Data (along with Coordinates)

- **Requirement**: There are 12 boroughs in Berlin. The basic data required to start this project is the names of all these Boroughs along with their coordinates
- **Collection**: Web scape the data of Borough of Berlin using 'BeautifulSoup'. Use 'Python Geocoder' to get the latitude and longitude values of these Boroughs.

Professional Venue Data

- **Requirement**: From these 12 boroughs we need to find out which borough have the most professional venues like offices, hospitals, industries, factories etc. In other words, we need to know in which zones we will have a constant flow of people (customers).
- **Collection**: Using 'Foursquare' by giving a specific category ID we can find the most frequent professional venues in these 12 boroughs.

Nearby Venues Data

- **Requirement**: We need to have an idea about the competition before we open a business. So, we need data about the most frequent venues nearby each selected zone.
- **Collection**: Explore the zones using 'Foursquare'

Pricing Data

- **Requirement**: Pricing data will help us in two ways: By giving us an estimate of the price values if you want to buy the land or rent it for the business. By giving us an idea about what kind of resident customers we are dealing with
- **Collection**: Websites have pricing data for boroughs of Berlin. (It is generally difficult to find accurate pricing data.) During the collection of data, Exploratory Data Analysis (Week2) is performed simultaneously to get more insights on it. Step by step lets understand the data.

Methodology:

Exploratory Data Analysis:

From the 456 professional venues obtained from all 12 boroughs was processed to find the most frequent venue and the top professional boroughs.

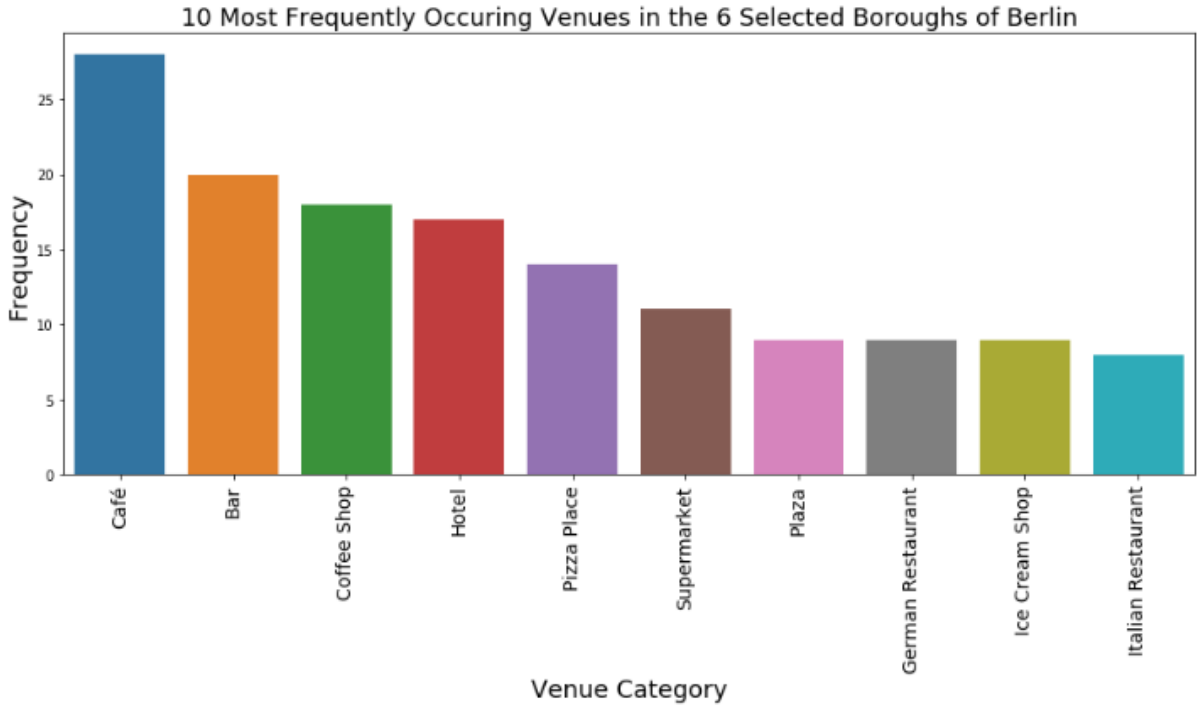
	Venue_Category	Count
0	Office	64
1	Doctor's Office	44
2	Convention Center	25
3	Dentist's Office	22
4	Building	20
5	Coworking Space	19
6	Medical Center	18

	Zone(Location)	Count
0	Mitte	50
1	Friedrichshain-Kreuzberg	50
2	Neukölln	50
3	Charlottenburg-Wilmersdorf	49
4	Tempelhof-Schöneberg	48
5	Marzahn-Hellersdorf	43
6	Steglitz-Zehlendorf	43

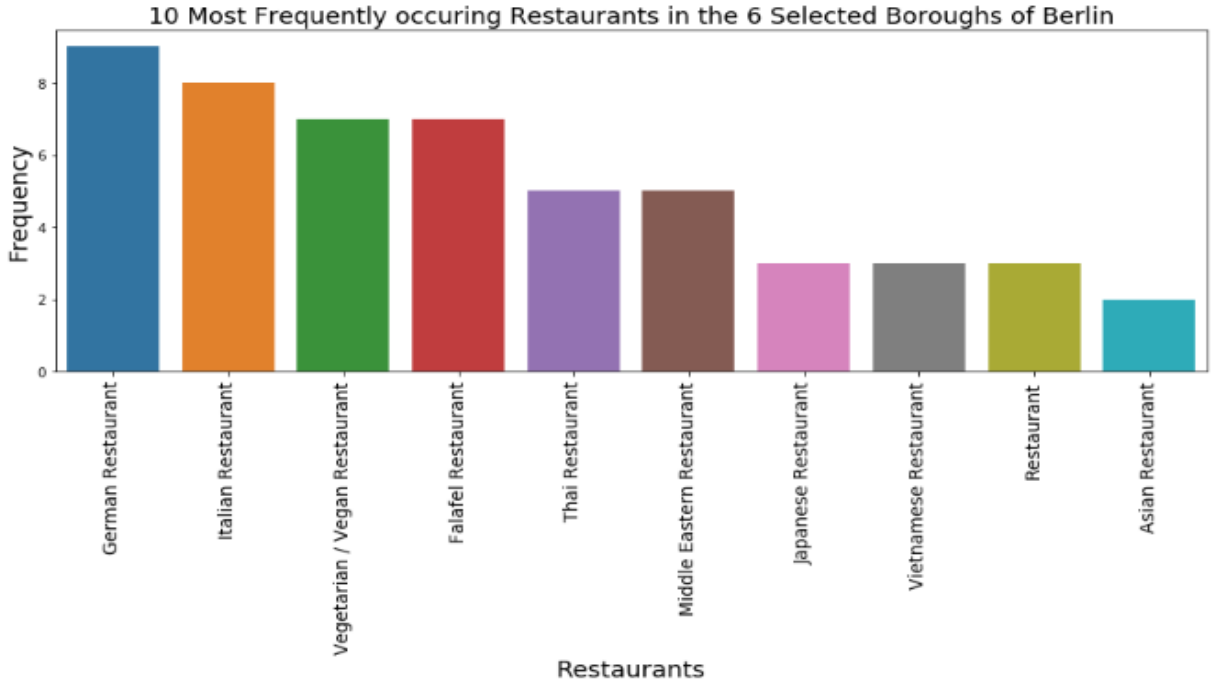
From this analysis 6 top boroughs are selected for further analysis.

	Location	Latitude	Longitude
0	Charlottenburg-Wilmersdorf	52.507856	13.263952
1	Friedrichshain-Kreuzberg	52.515306	13.461612
2	Marzahn-Hellersdorf	52.522523	13.587663
3	Mitte	52.517690	13.402376
4	Neukölln	52.481150	13.435350
5	Tempelhof-Schöneberg	52.440603	13.373703

143 venues nearby these selected boroughs were explored using foursquare and their details were collected.



Exploring this data, it was found that the most frequent place in these selected boroughs is 'Café'. To find the most frequent restaurant all the restaurants data was analyzed.



This project requires us to find the business or professional boroughs of Berlin and explore these boroughs to find out the frequent venues of these boroughs. All of these is done so that we choose a borough which has more demand for our new business Firstly, we have collected all the required data and have done some exploratory data analysis to find the top 6 professional Boroughs of Berlin based on the professional venues' frequency in that borough. We found that the most frequent Professional Venue in all the boroughs combined is an "Office". Frequent venues were explored in these selected zones and it was found that venue category of "German Restaurant" is the most frequent venue nearby these selected zones but if compared with café or bar it's far less. From this it is clear who our potential customers are and what they prefer. Secondly, we need to analyze the data a little more to get insights into the venue category. This can be done by using one-hot encoding. Thirdly. We will use a machine learning method called K-Means Cluster to cluster the zones into groups depending how similar or dissimilar they are.

Analysis

Using one hot encoding the venue categories are analyzed.

```
=========Friedrichshain-Kreuzberg=========
              Venue  Freq
0            Office  0.16
1  Doctor's Office  0.12
2     Tech Startup  0.10
3      Post Office  0.08
4  Language School  0.06
```

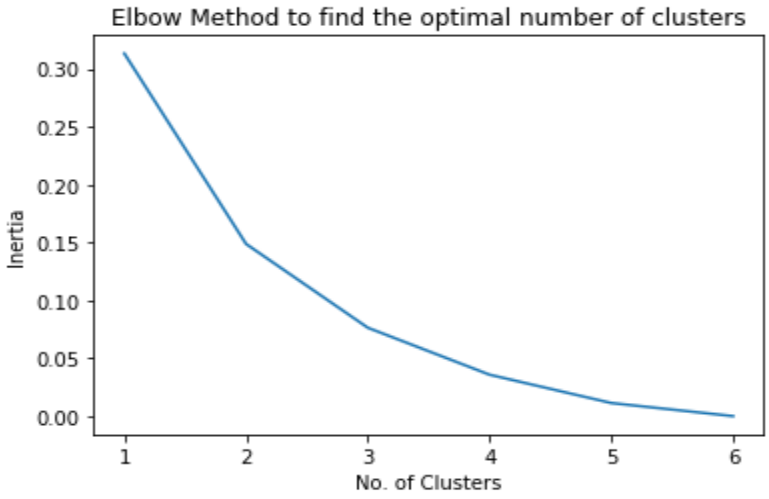
```
==========Marzahn-Hellersdorf==========
                           Venue  Freq
0                Doctor's Office  0.23
1                 Medical Center  0.09
2                    Post Office  0.07
3                         Office  0.07
4  Professional & Other Places  0.05
```

```
==========Mitte==========
                 Venue  Freq
0               Office  0.16
1         Tech Startup  0.16
2             Building  0.10
3  Monument / Landmark  0.08
4      Language School  0.06
```

```
==========Neukölln==========
              Venue  Freq
0  Coworking Space  0.18
1  Doctor's Office  0.08
2  Language School  0.06
3           Office  0.06
4        City Hall  0.06
```

```
=========Tempelhof-Schöneberg=========
             Venue  Freq
0           Office  0.27
1  Medical Center  0.06
2           Church  0.06
3         Building  0.04
4       TV Station  0.04
```

After choosing the number of clusters using Elbow method, the Boroughs were clustered into 4 clusters using K-means Clustering.



A final dataframe including the cluster label and average price per sqm of each Borough was made.

	Location	Cluster Label	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	Avg Price per sqft
5	Tempelhof-Schöneberg	1	Bakery	Bus Stop	Supermarket	4500
0	Charlottenburg-Wilmersdorf	2	Café	Italian Restaurant	German Restaurant	4500
1	Friedrichshain-Kreuzberg	2	Café	Pizza Place	Bar	5000
4	Neukölln	2	Bar	Coffee Shop	Café	4500
2	Marzahn-Hellersdorf	3	Supermarket	Drugstore	Metro Station	3000
3	Mitte	4	Hotel	History Museum	Plaza	5800

Result and Discussion:

Results

- Cluster Results Analysis
    - Cluster 1 contains zones whose common venues are not restaurants
    - Cluster 2 contains zones with top 2 most common venues being Cafe
    - Cluster 3 contains zones with whose common venues are of other categories
    - Cluster 4 again contains zones whose common venues are of other categories
- Cafe or Bars are the most frequent venues near the selected Borough.
- 'Marzahn-Hellersdorf' has the least average price per sqm, followed by 'Perungudi', among the selected zones.

Discussion

Based on the clustering and exploratory data analysis with maximum frequency touristic places are in the area ‘Mitte’ seems like a potential zone to open up our Restaurant or Cafe or Bar. The pricing data also seems less favorable to this due to many other commercial buildings in the area. Clustering also shows these venues in cluster 2 which represents the cluster with restaurants as the frequent venues. Although the results seem promising as Mitte is an area with a lot of touristic places in the city of Berlin, further analysis needs to be done based on the wards in these Boroughs to get a more accurate location to open the business. Since the clustering is done based on only the common venues obtained from Foursquare the results will need more refining. But this preliminary analysis will be of great help in the beginning stages of the business plan.

Conclusion

The main objective of this project was to understand how to deal with real life data science projects using some of the popular Python packages such as seaborn, folium, BeautifulSoup and geocoders. I have also got a glimpse of how web scraping is done and how FourSquare can be used to acquire data of frequent venues in a selected area.

The idea of opening a Restaurant or Cafe or Bar in an area which has a hub for touristic places is an interesting and a potential idea to try in Berlin. Although the analysis is very preliminary and requires a lot of refining based on the data used (refined ward data per each Borough, pricing data), this analysis helped me understand Berlin more.