Finding the Best Spot in Berlin for Opening a Café cum Restaurant | Data Science Capstone Project

Alpesh Laxman Vora

July 29, 2020

Since the last few months, I have been working to achieve an IBM Data Science Professional certification. Through this course, I have learned about data science and also acquired various skills and in-depth understanding about different tools that the data scientists need to solve the given problems. As a Capstone for the certification, I am required to define busines problem in the city of my choice, solve it by scraping data from the web and location data from Foursquare, and produce full report. This report contains every process - from problem description to conclusion, including data processing, methodology, analysis, results, and discussion - for finding the best spot in Berlin to open a Café cum Restaurant. Detailed python code for this project is located on GitHub-Capstone Project Code or Jupyter Notebook - Capstone Project.

1 Introduction

1.1 Background Discussion

Berlin is the largest city of Germany by both population and area. Its 3.77 million registered inhabitants also make it the most populous city of the European Union. Since 1990 after the fall of the Wall, Berlin is the capital of Germany and it established itself as a city of unlimited possibilities for travelers. The mixture of its historical significance, various festivals, diverse architecture, spectacular nightlife, and contemporary arts makes this city a magnet for tourists. Berlin's 3.77 Million inhabitants are hosts to nearly 34 million overnight stays and 14 million visitors in year 2019 [1], which makes Berlin a top European destination – ranked third after London and Paris.

In addition, Berlin is also Europe's leading economic force and its economy is growing rapidly for several years. In 2019, the nominal gross domestic product in Berlin was € 153.3 billion [2], which is 3% higher compared to the previous year and the number of people employed increased by 2.4% to more than 2 million [3].

All in all, the city's rapidly growing economy, office businesses and tourisms make Berlin one of the best places to start up a new business, especially in Restaurants as the office businesses and the tourism are an important driver for restaurant industry.

1.2 Business Opportunity (Problem Description)

Now a days, most people, especially business employees, are eating more meals out of the home, because the work and lifestyle pressure often do not allow sufficient time for most of employees to prepare their own cuisine. Therefore, they rely on the Cafés and restaurants for their dietary needs, especially for breakfast and lunch. Similarly, tourist or business visitors also rely on the cafés and restaurants for their every meal. Berlin, where both the office businesses and tourism are growing rapidly since several years, provides huge opportunities for restaurant business. Reasonably priced restaurant in popular tourist and highly dense office places can attract both audiences – office employees and tourists. With given this scenario, we will analyze various factors to determine the best place in Berlin to open a café cum restaurant, where breakfast cum lunch or even brunch can be served. This report outlines some basic assumptions, data sets and data processing, and detailed analysis which can help us to select the best spot in Berlin for opening a café cum restaurant. In this process, we have assumed that money is not an issue for stating the business.

1.3 Target Audience

The key audience of this report would be anyone who wants to invest or open a restaurant in Berlin, or anyone in Berlin looking for a delicious breakfast, brunch, or lunch. The analysis will also help to office employees to find reasonable breakfast/lunch/brunch place close to office area.

2 Data – Acquisition and Cleaning

In order to find best spot in Berlin, various set of data like boroughs, neighborhoods, area, population, price of rent or buying property, number of registered companies, and number of visitors, etc. are required. Unfortunately, all sets of data are not available from any single source. Hence, the various sets of data are retrieved from different sources and Foursquare API is used to retrieve the venue data.

Altogether, below 5 sets of data are used for our analysis:

1. Berlin Boroughs, Neighborhoods, Area and Population Data
2. Berlin Registered Business Data
3. Berlin Registered Tourist Data
4. Rent or buying price of property in Berlin
5. Foursquare API

2.1 Berlin Boroughs, Neighborhoods, Area and Population Data

This set of data is retrieved by scrapping the Wikipedia Page - "Verwaltungsgliederung Berlins" [4] and placed it into a pandas data frame using python. This data frame contains lists of every neighborhoods of Berlin with their boroughs, size of area, and population in each neighborhood. This set of data will help us to figure out the densely populated

neighborhoods or boroughs in Berlin. In the scrapped raw data, many columns contained the trailing newlines (\n), and also commas as decimal separators and dots as thousand separators due to German conventions. After cleaning and processing the raw data frame by coding in python, we get resulting data frame with standard conventions as follow:

```
Data types of each columns:
 Neighborhoods          object
Boroughs               object
Area_sqkm             float64
Population              int32
Population Density    float64
dtype: object

The shape of dataframe is: (96, 5)
Total numbers of Boroughs in Berlin are: 12
Total numbers of Neighborhoods in Berlin are: 96
```

Out[3]:

	Neighborhoods	Boroughs	Area_sqkm	Population	Population Density
1	Mitte	Mitte	10.70	101932	9526.0
2	Moabit	Mitte	7.72	79512	10299.0
3	Hansaviertel	Mitte	0.53	5894	11121.0
4	Tiergarten	Mitte	5.17	14753	2854.0
5	Wedding	Mitte	9.23	86688	9392.0
...	...	...	...	...	...
92	Waidmannslust	Reinickendorf	2.30	10958	4764.0
93	Lübars	Reinickendorf	5.00	5174	1035.0
94	Wittenau	Reinickendorf	5.90	24306	4120.0
95	Märkisches Viertel	Reinickendorf	3.20	40258	12581.0
96	Borsigwalde	Reinickendorf	2.00	6826	3413.0

96 rows × 5 columns

As we can see there is a total 12 boroughs and 96 neighborhoods in Berlin.

2.2 Berlin Registered Business Data

This data is downloaded from *"Amt für Statistik Berlin-Brandenburg"* website [5] in csv format. This downloaded csv file is placed into a pandas data frame using python. This data frame contains the number of registered businesses in each borough of Berlin. For each borough, the numbers of registered businesses are also divided into four groups according to number of employees in each business. But, we would like to have the total

number of registered businesses in each borough because it can give us a rough indication of most busy areas of berlin by business employees, at least on the workdays. After processing the dataset by coding in python, we get the resulting data frame with total number of registered businesses in each borough as follow:

```
Data types of each columns:
 Boroughs            object
Total Businesses     int64
dtype: object

The shape of dataframe is: (12, 2)
Total numbers of registered businesses in Berlin are: 192199
```

Out[6]:

	Boroughs	Total Businesses
1	Charlottenburg-Wilmersdorf	29324
2	Mitte	28553
3	Pankow	22628
4	Friedrichshain-Kreuzberg	20978
5	Tempelhof-Schöneberg	18926
6	Steglitz-Zehlendorf	14400
7	Neukölln	12511
8	Treptow-Köpenick	11316
9	Reinickendorf	9966
10	Lichtenberg	8233
11	Spandau	7692
12	Marzahn-Hellersdorf	7672

We can see that there are total 192,199 businesses in Berlin. The Charlottenburg-Wilmersdorf and Mitte boroughs have significantly more business. It is also seen that there are four boroughs in which the total number of businesses are more than 20,000.

2.3 Berlin Registered Tourist Data

This data is downloaded from *"Amt für Statistik Berlin-Brandenburg"* website [5] in csv format. This downloaded csv file is placed into a pandas data frame using python. This data frame contains the number of visitors and overnight stays visitors in each borough of Berlin for year 2019. This data can give us a rough indication of most busy areas of berlin by tourists. After processing the dataset by coding in python, we get the resulting data frame with total number of tourists in each borough as follow:

```
Data types of each columns:
 Boroughs                 object
Total Visitors in 2019    int64
dtype: object

The shape of dataframe is: (12, 2)
Total numbers of tourists visited Berlin in year 2019 are: 48087709
```

Out[9]:

	Boroughs	Total Visitors in 2019
1	Mitte	20809270
2	Charlottenburg-Wilmersdorf	9157433
3	Friedrichshain-Kreuzberg	6495909
4	Tempelhof-Schöneberg	2948106
5	Pankow	1924337
6	Lichtenberg	1791645
7	Neukölln	1331199
8	Treptow-Köpenick	951862
9	Spandau	877379
10	Reinickendorf	732466
11	Steglitz-Zehlendorf	711672
12	Marzahn-Hellersdorf	356431

We can see that Berliner inhabitants hosted nearly 48 million visitors in year 2019. Out of 48 Million tourists, nearly 21 Million tourists have visited or stayed overnight only in Mitte borough.

2.4 Rent or buying price of property in Berlin

Even though we have assumed that money is not an issue, the information about the rent or buying price of property can help us to give rough idea of how expensive it will be to maintain a café cum restaurant business in each neighborhoods or borough. Fortunately, we found the average rent price in € per month-m^2 in year 2019 for each neighborhoods of Berlin from "Homeday – mein Immobilienmakler" website [6]. From there, the data is retrieved by scrapping the web page [6] and placed it into a pandas data frame using python. The scrapped web-data contained commas as decimal operators due to German conventions. After cleaning and processing the scrapped web-data by coding in python, we get resulting data frame, with standard conventions, contained the rent price in each neighborhood of berlin:

```
Data types of each columns:
 Neighborhoods                          object
Rent Price in 2019 [€/m²-month]    float64
dtype: object

The shape of df_berlin dataframe is: (96, 2)
```

Out[12]:

	Neighborhoods	Rent Price in 2019 [€/m²-month]
1	Tiergarten	14.0
2	Friedrichshain	13.6
3	Grunewald	13.0
4	Rummelsburg	13.0
5	Moabit	12.6
...	...	...
92	Neu-Hohenschönhausen	8.0
93	Hellersdorf	7.9
94	Falkenberg	7.5
95	Marzahn	7.3
96	Wartenberg	7.2

96 rows × 2 columns

3 Methodology and Exploratory Data Analysis

3.1 Merging Datasets and Importing the Geospatial Data

As discussed in Section-2, we have now total four sets of data, which are retrieved from different sources and placed them into different data-frames. For further analysis, these sets of data must be merged into one Pandas data frame. In order to merge these datasets into one data-frame, all sets of data must be in same shape or form, either for each neighborhood or for each borough. But out of these four datasets, two datasets are based on

neighborhoods, whereas other two datasets (business and tourist data) are only available for each borough. Hence, the neighborhood-based datasets are needed to transform to borough-based datasets by grouping down them on boroughs. After transforming datasets, they are merged on boroughs in the one data frame and unnecessary columns (Population density column) are removed.

After merging the datasets, the geographical coordinates (latitude and longitude) of each borough are imported by using Python client GeoPy geocoders, which converts the address or any landmarks into geographic coordinates. But some of the co-ordinates returned by GeoPy for each borough are slightly differed from the google search, and one of the co-ordinates (Lichtenberg borough) are completely wrong. Hence, the differed co-ordinates for each borough are replaced by actual co-ordinates. After updating the geographical coordinates, the final data-frame looks like as below:

	Boroughs	Neighborhoods	Area_sqkm	Population	Total Businesses	Total Visitors in 2019	Rent Price in 2019 [€/m²-month]	Latitude	Longitude
1	Charlottenburg-Wilmersdorf	Charlottenburg, Wilmersdorf, Schmargendorf, Gr...	64.62	342332	29324	9157433	13.0	52.507856	13.263952
2	Friedrichshain-Kreuzberg	Friedrichshain, Kreuzberg	20.18	289762	20978	6495909	13.6	52.515306	13.461612
3	Lichtenberg	Friedrichsfelde, Karlshorst, Lichtenberg, Falk...	52.02	291452	8233	1791645	13.0	48.921296	7.481227
4	Marzahn-Hellersdorf	Marzahn, Biesdorf, Kaulsdorf, Mahlsdorf, Helle...	61.71	268548	7672	356431	9.8	52.522523	13.587663
5	Mitte	Mitte, Moabit, Hansaviertel, Tiergarten, Weddi...	39.48	384172	28553	20809270	14.0	52.517690	13.402376
6	Neukölln	Neukölln, Britz, Buckow, Rudow, Gropiusstadt	44.91	329691	12511	1331199	9.3	52.481150	13.435350
7	Pankow	Prenzlauer Berg, Weißensee, Blankenburg, Heine...	103.26	407765	22628	1924337	12.1	52.597637	13.436374
8	Reinickendorf	Reinickendorf, Tegel, Konradshöhe, Heiligensee...	89.40	265220	9966	732466	10.0	52.604763	13.295287
9	Spandau	Spandau, Haselhorst, Siemensstadt, Staaken, Ga...	91.90	243977	7692	877379	10.0	52.535788	13.197792
10	Steglitz-Zehlendorf	Steglitz, Lichterfelde, Lankwitz, Zehlendorf, ...	102.47	308697	14400	711672	12.4	52.429205	13.229974
11	Tempelhof-Schöneberg	Schöneberg, Friedenau, Tempelhof, Mariendorf, ...	53.08	351644	18926	2948106	11.8	52.440603	13.373703
12	Treptow-Köpenick	Alt-Treptow, Plänterwald, Baumschulenweg, Joha...	165.70	271153	11316	951862	11.6	52.417893	13.600185

3.2 Narrowing Down Boroughs and Visualizing using Folium

Berlin has total 12 boroughs. For detail exploration and analysis, we can narrow down our options to 5 boroughs by examining each relevant parameter (businesses, tourists, rent price) from our data sets. These all parameters are visualized in Figure 1-3.

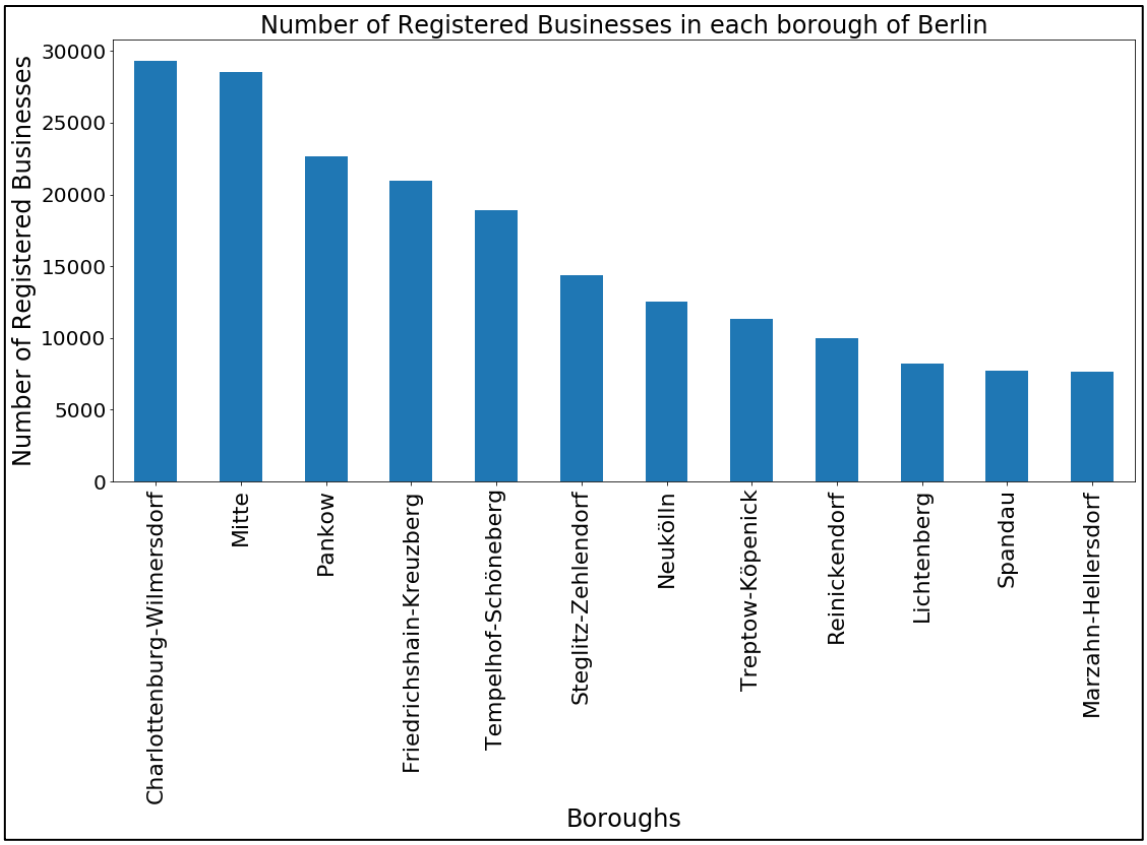

Figure 1. Number of registered businesses in each boroughs of Berlin.

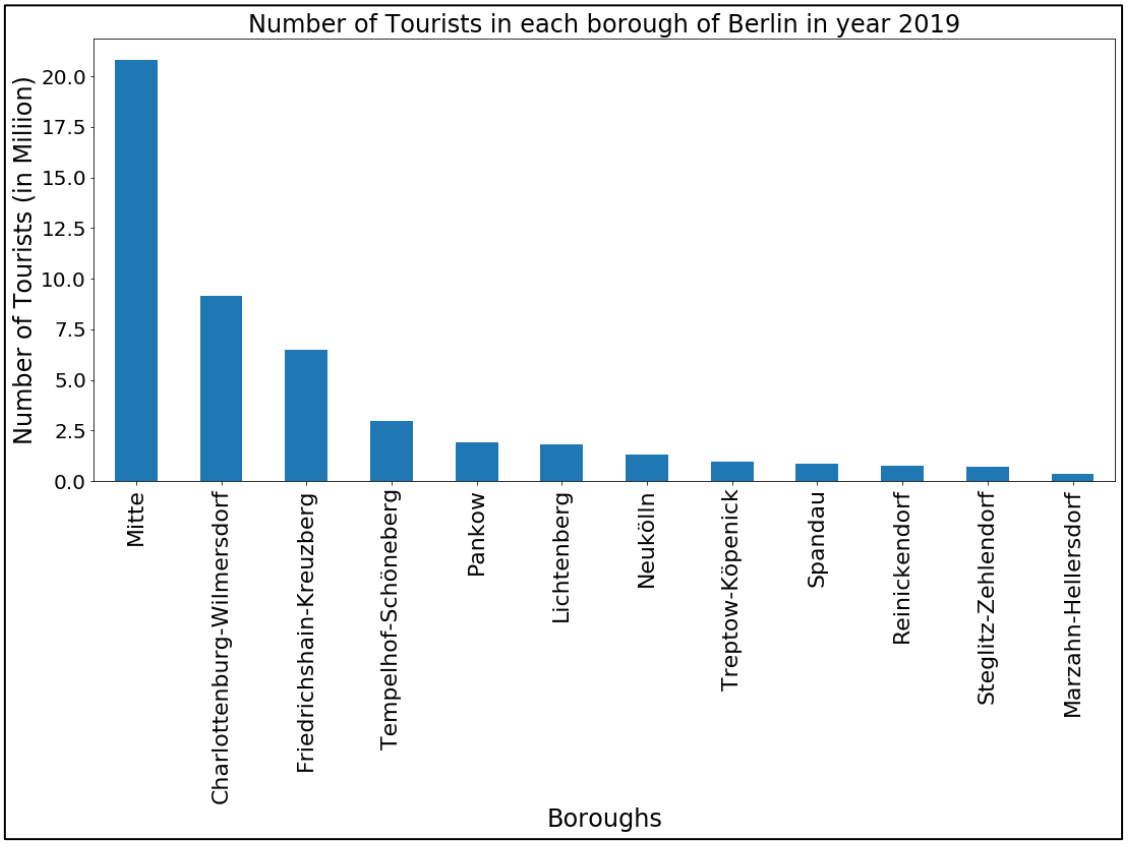

Figure 2. Number of visitors and overnight stays tourists in each boroughs of Berlin.

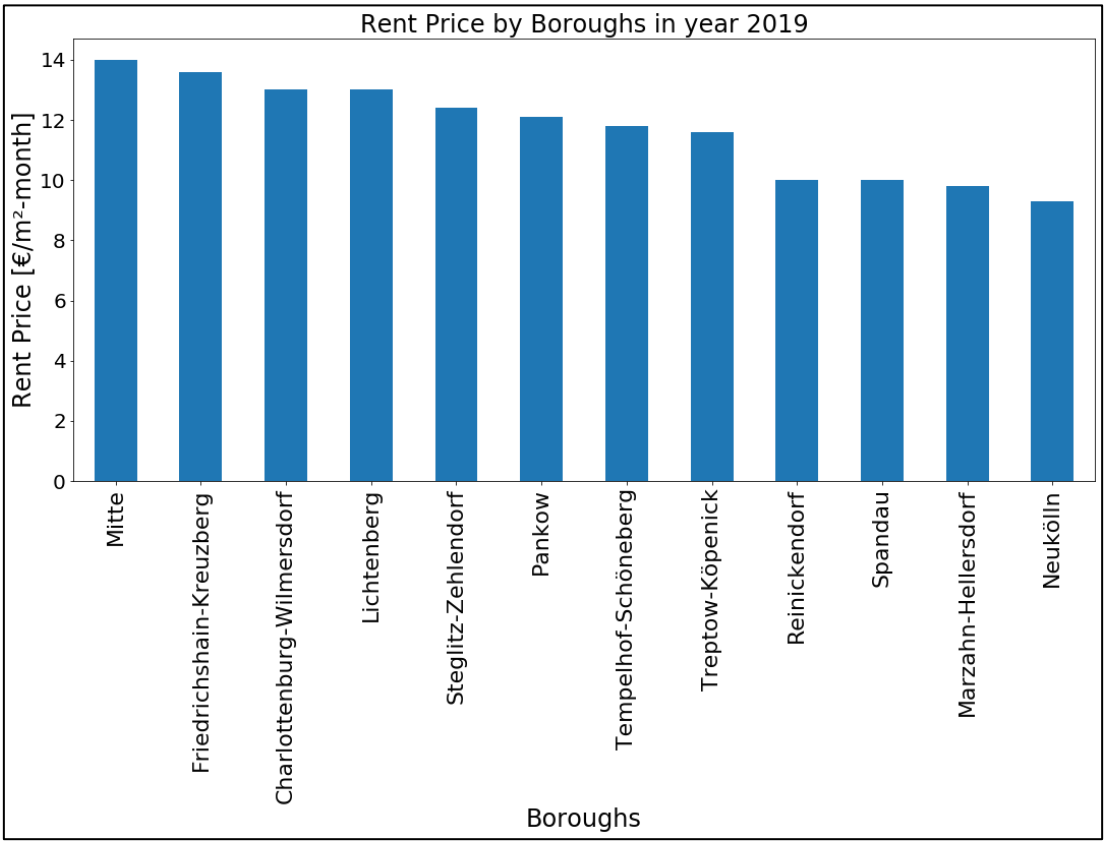

Figure 3. Rent price of property in each boroughs of Berlin.

As mentioned earlier that money is not an issue, hence, our selection of top 5 boroughs will be only based on number of Businesses and Tourists in each borough. Based on businesses and tourists' datasets, one can clearly see two points:

- 62.7% of total businesses are existing only in 5 boroughs. Out of 5 boroughs, both Charlottenburg-Wilmersdorf and Mitte boroughs have 30% of total businesses, whereas Pankow, Friedrichshain-Kreuzberg and Tempelhof-Schöneberg boroughs cover another 31.7% of total businesses.
- 86% of total Tourists have either visited or stayed overnight in 5 Boroughs. Out of those tourists, 43% of tourists are only hosted by Mitte borough, whereas another 43% are hosted by another four boroughs, namely, Charlottenburg-Wilmersdorf, Friedrichshain-Kreuzberg, Tempelhof-Schöneberg and Pankow boroughs.

Coincidentally, the top 5 boroughs which cover 62.7% of total businesses, and the top 5 boroughs which are hosts to 86% of total tourists are the same. Hence, it seems that we have narrowed down our options to 5 boroughs and from now onwards we will consider only follow 5 boroughs for further analysis:

1. Charlottenburg-Wilmersdorf

2. Mitte
3. Pankow
4. Friedrichshain-Kreuzberg
5. Tempelhof-Schönberg

The resultant panda data frame with 5 narrowed down boroughs is as follow:

```
Data types of each columns:
 Boroughs                            object
Neighborhoods                        object
Area_sqkm                           float64
Population                            int32
Total Businesses                      int64
Total Visitors in 2019                int64
Rent Price in 2019 [€/m²-month]     float64
Latitude                            float64
Longitude                           float64
dtype: object

The shape of dataFrame is: (5, 9)
```

	Boroughs	Neighborhoods	Area_sqkm	Population	Total Businesses	Total Visitors in 2019	Rent Price in 2019 [€/m²-month]	Latitude	Longitude
1	Charlottenburg-Wilmersdorf	Charlottenburg, Wilmersdorf, Schmargendorf, Gr...	64.62	342332	29324	9157433	13.0	52.498889	13.284917
2	Mitte	Mitte, Moabit, Hansaviertel, Tiergarten, Weddi...	39.48	384172	28553	20809270	14.0	52.516667	13.366667
3	Pankow	Prenzlauer Berg, Weißensee, Blankenburg, Heine...	103.26	407765	22628	1924337	12.1	52.568889	13.402222
4	Friedrichshain-Kreuzberg	Friedrichshain, Kreuzberg	20.18	289762	20978	6495909	13.6	52.500000	13.433333
5	Tempelhof-Schöneberg	Schöneberg, Friedenau, Tempelhof, Mariendorf, ...	53.08	351644	18926	2948106	11.8	52.466667	13.383333

By using Folium library, these 5 boroughs are shown on leaflet map (Figure 4).

Figure 4. A map of berlin with each of our 5 selected boroughs marked with circle

3.3 Foursquare Data Analysis

Foursquare API allows us to retrieve information about the most popular venues in each boroughs of Berlin. Foursquare API returns the data in as a JSON file, which can be converted into a data frame in the python notebook for further analysis.

To start the Foursquare analysis, a function is implemented to search the most popular venues within 1 km radius of selected boroughs. For our selected boroughs, this function has returned the total 500 nearby different venues (100 venues for each of our selected boroughs) with their addresses, geographical coordinates, and their categories.

```
Charlottenburg-Wilmersdorf
Mitte
Pankow
Friedrichshain-Kreuzberg
Tempelhof-Schöneberg


Total 500 nearby venues were returned by Foursquare.
Shape of the nearby venues Dataframe:  (500, 7)
```

	Boroughs	Boroughs Latitude	Boroughs Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Charlottenburg-Wilmersdorf	52.498889	13.284917	Lietzenseepark	52.505199	13.286676	Park
1	Charlottenburg-Wilmersdorf	52.498889	13.284917	Fleischerei Bünger	52.496390	13.292842	Butcher
2	Charlottenburg-Wilmersdorf	52.498889	13.284917	Aspria Berlin Ku'damm	52.500597	13.294459	Hotel
3	Charlottenburg-Wilmersdorf	52.498889	13.284917	Cups	52.497388	13.291307	Café
4	Charlottenburg-Wilmersdorf	52.498889	13.284917	Scoopy doo	52.498309	13.290653	Ice Cream Shop
...	...	...	...	...	...	...	...
495	Tempelhof-Schöneberg	52.466667	13.383333	Café Olé	52.454252	13.381650	Restaurant
496	Tempelhof-Schöneberg	52.466667	13.383333	Eckert	52.476677	13.364705	Newsstand
497	Tempelhof-Schöneberg	52.466667	13.383333	Attilaplatz	52.455340	13.375917	Plaza
498	Tempelhof-Schöneberg	52.466667	13.383333	MediaMarkt	52.455747	13.385719	Electronics Store
499	Tempelhof-Schöneberg	52.466667	13.383333	Le Crobag	52.475956	13.365191	Sandwich Place

In 500 nearby venues, there are 157 unique venues categories:

```
In [24]: # Unique categories from all the returned venues
         print('There are {} unique categories.'.format(len(BerlinTop5Boroughs_nearby_venues['Venue Category'].unique())))
         print(BerlinTop5Boroughs_nearby_venues['Venue Category'].value_counts())

         There are 157 unique categories.
         Café                            36
         Hotel                           24
         Italian Restaurant              22
         Park                            16
         Supermarket                     16
                                         ..
         Bistro                           1
         Pizza Place                      1
         Cheese Shop                      1
         Tea Room                         1
         Paper / Office Supplies Store    1
         Name: Venue Category, Length: 157, dtype: int64
```

Out of 157 unique venues categories, the most top 10 popular venue categories across 5 selected boroughs are plotted in Figure 5.

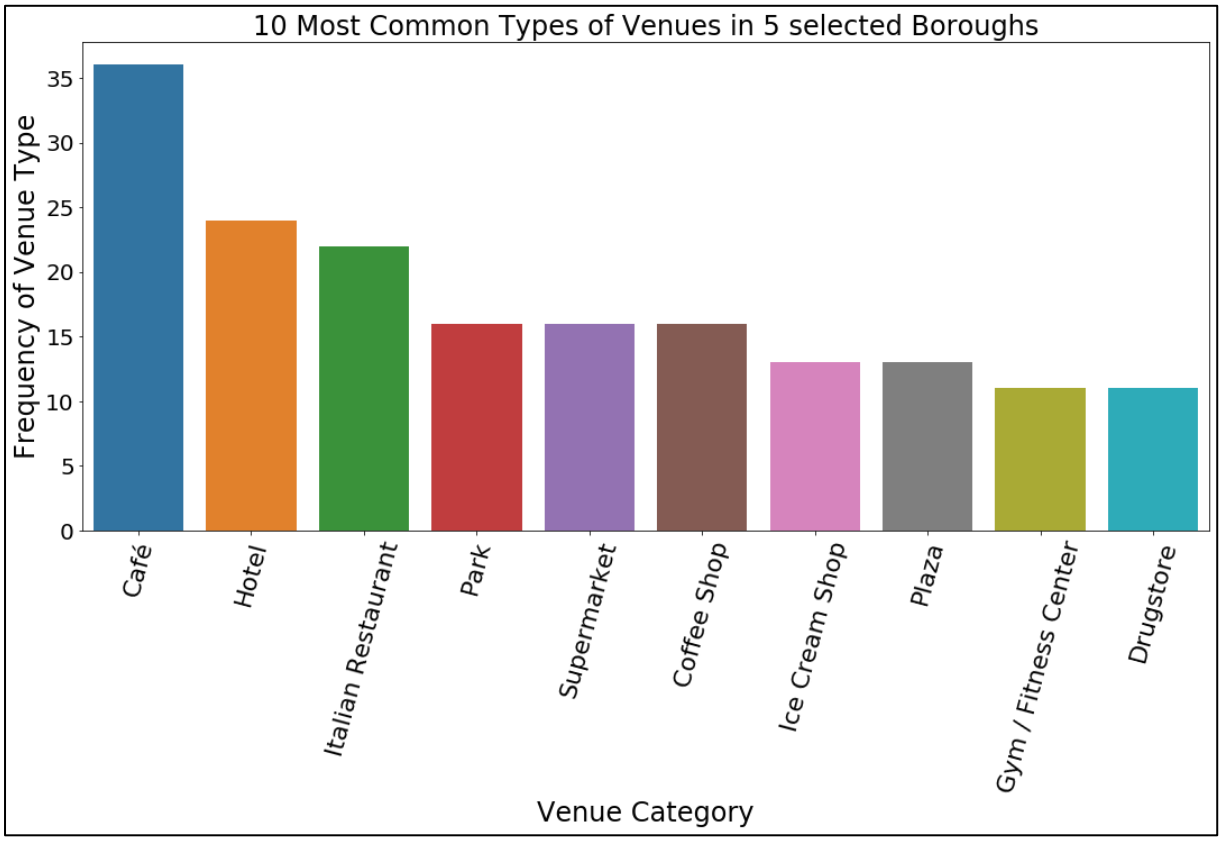

Figure 5. Most frequent venues around our 5 selected boroughs of berlin

From figure 5, it is clearly seen that the most common popular type of venue in our selected boroughs overall is Café, and then followed by Hotel, Italian Restaurant, Park, Supermarket and so on.

Similarly, we can also obtain information about the top 5 venues for each selected borough by performing few steps as follow:

- Create a data frame for venue categories with pandas one hot encoding
- Group by on boroughs and calculate the mean of the venue categories
- Transpose the data frame and arrange in descending order

After implementing one hot encoding and performing above steps in Panda, we get top 5 venues for each selected borough as follow:

```
========== Charlottenburg-Wilmersdorf ==========
                  Venue  Freq
0     Italian Restaurant  0.10
1                   Café  0.09
2                  Hotel  0.06
3      Trattoria/Osteria  0.06
4  Vietnamese Restaurant  0.05


========== Friedrichshain-Kreuzberg ==========
             Venue  Freq
0     Coffee Shop  0.09
1             Bar  0.06
2  Ice Cream Shop  0.05
3            Café  0.05
4     Yoga Studio  0.03


========== Mitte ==========
                  Venue  Freq
0                 Hotel  0.12
1                 Plaza  0.06
2         Concert Hall  0.05
3           Art Museum  0.04
4  Monument / Landmark  0.04


========== Pankow ==========
                Venue  Freq
0                Café  0.15
1         Supermarket  0.06
2                Park  0.05
3  Italian Restaurant  0.04
4           Drugstore  0.04


========== Tempelhof-Schöneberg ==========
                Venue  Freq
0                Park  0.08
1         Supermarket  0.07
2                Café  0.05
3  Italian Restaurant  0.05
4    Doner Restaurant  0.04
```

The above data is very important because it gives us information about top 5 popular venues in each borough, which help us to differentiate the individual boroughs. It also helps us to find the particular venue category in each borough. As our focus is only on Restaurant and Café categories, the python code is implemented to find the number of restaurants and café in each selected borough and the obtained result is presented in Figure 6 with table.

	Boroughs	Number of Restaurants	Number of Cafes	Number of Hotels
1	Charlottenburg-Wilmersdorf	37	9	6
2	Friedrichshain-Kreuzberg	24	5	4
3	Mitte	7	2	14
4	Pankow	20	15	2
5	Tempelhof-Schöneberg	21	5	1

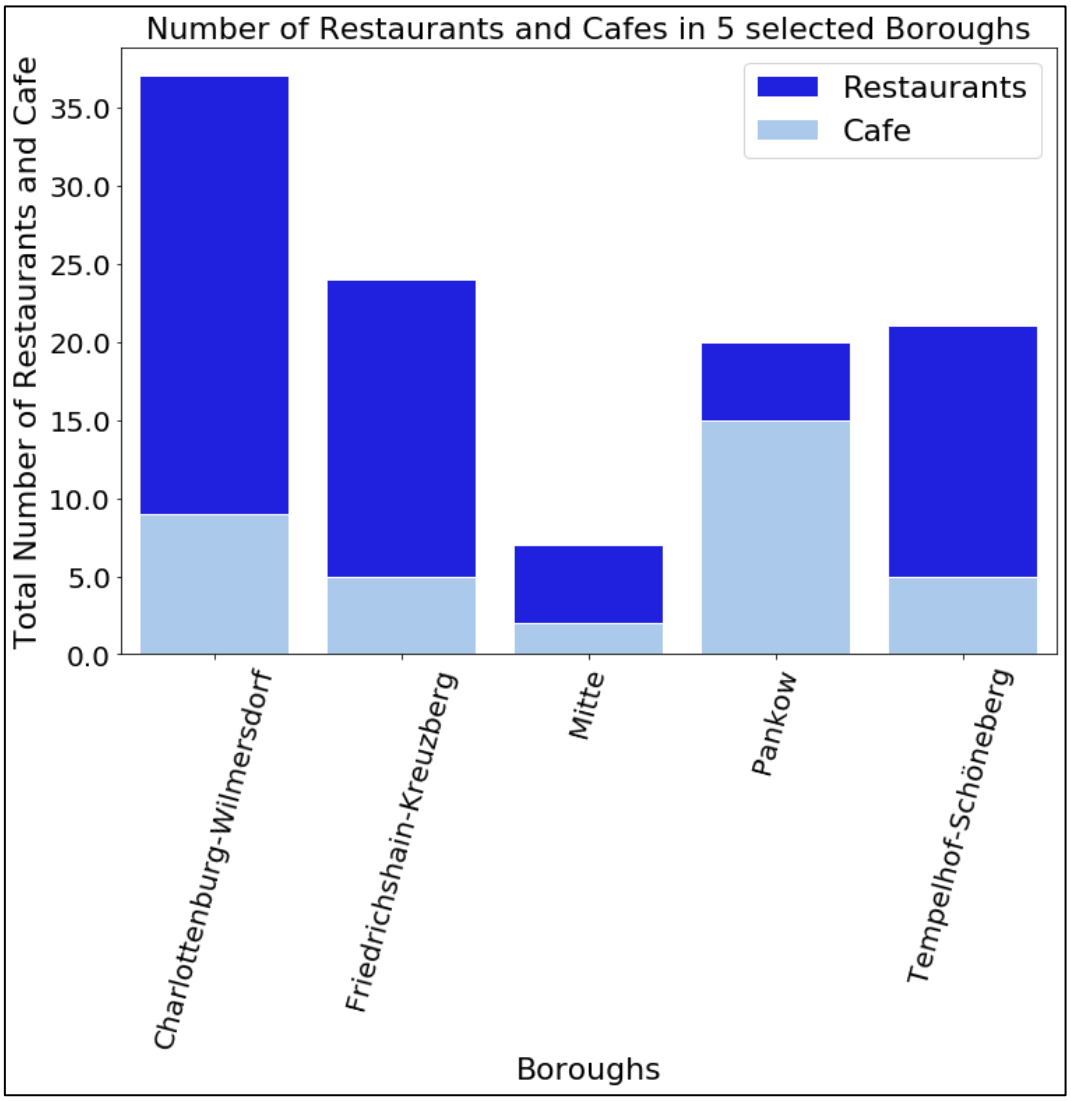

Figure 6. Number of restaurants and cafés as most common venues in our 5 selected boroughs of berlin.

By using Folium library, most visited restaurant and cafés in our selected 5 boroughs can be displayed on a leaflet map (Figure 7-8).

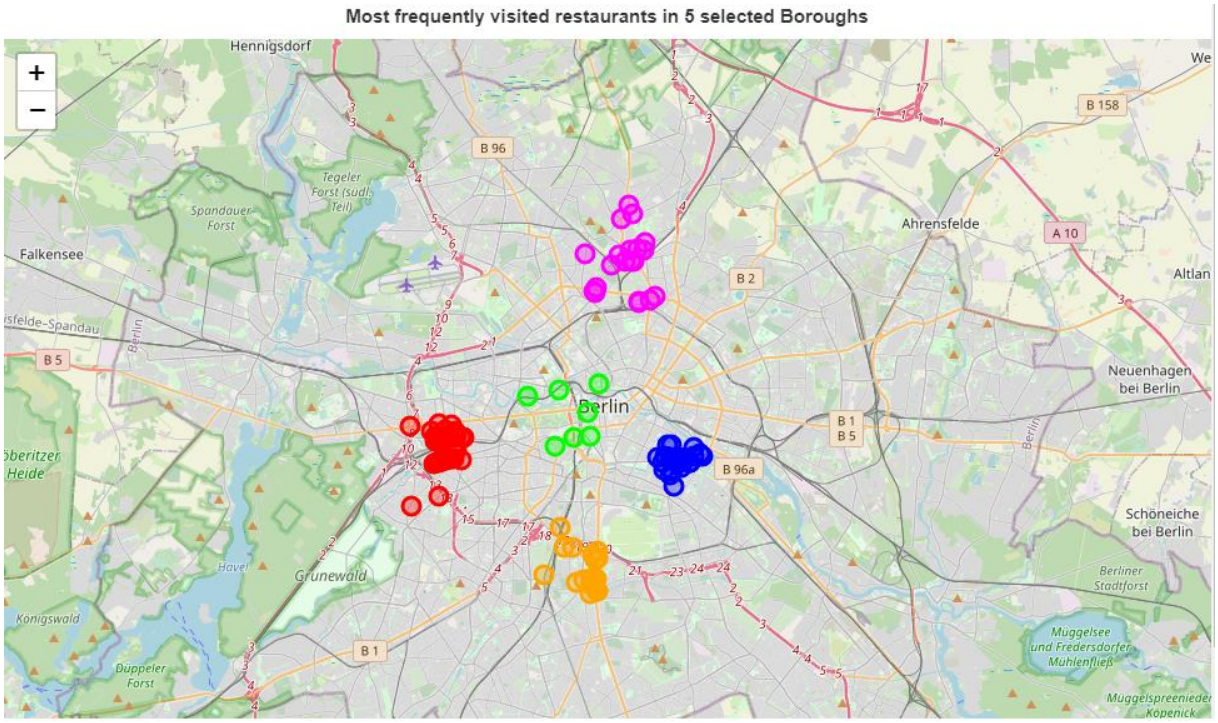

Figure 7. Most frequently visited restaurants (marked with circles) in our 5 selected boroughs (Charlottenburg-Wilmersdorf, Mitte, Pankow, Tempelhof-Schöneberg, Friedrichshain-Kreuzberg)

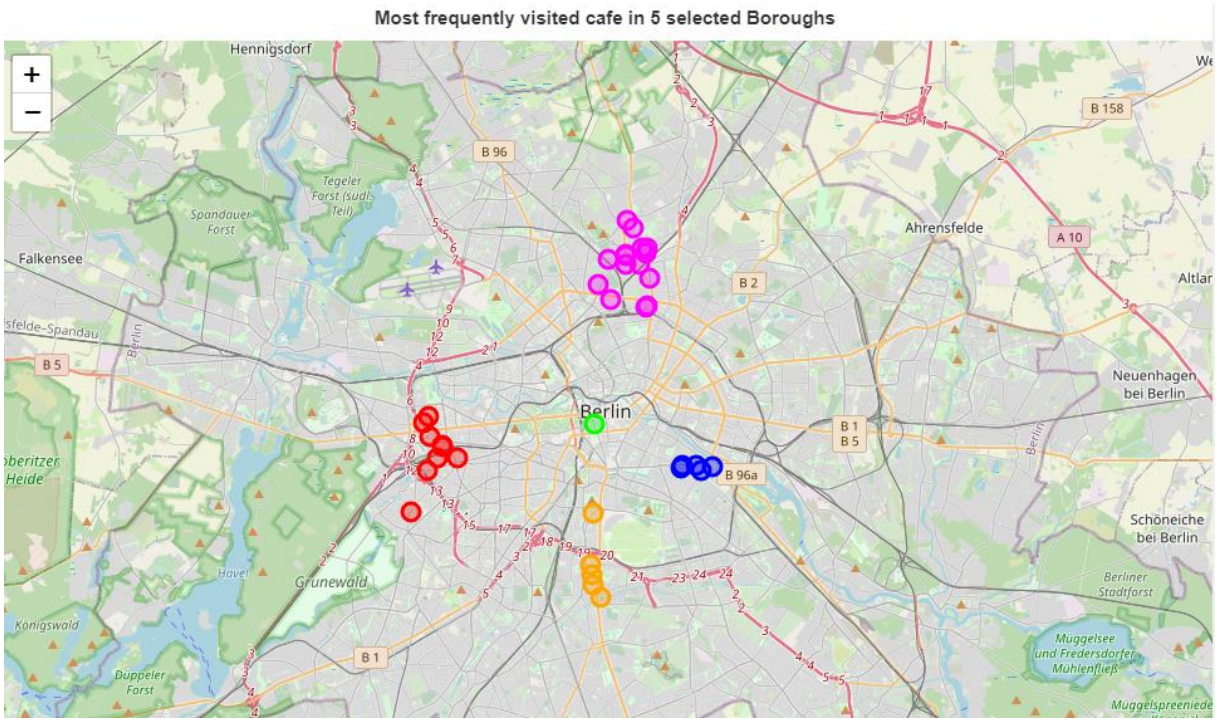

Figure 8. Most frequently visited Cafés (marked with circles) in our 5 selected boroughs (Charlottenburg-Wilmersdorf, Mitte, Pankow, Tempelhof-Schöneberg, Friedrichshain-Kreuzberg)

3.4 Borough Clust ering using K-Means Clustering

Finally, we can cluster these 5 boroughs based on their popular venue categories, which will create the groups of similar type of boroughs based on the similarity of venue categories in each borough. For clustering, K-Means clustering is used to group the 5 selected boroughs into 3 clusters. The detail code is shown below:

Clustering the selected boroughs for more insights using K-Means clustering

```
# set nubmer of clusters
k_clusters = 3

berlin_grouped_clustering = BerlinTop5Boroughs_nearby_venues_grouped.drop('Boroughs', 1)

# Run the K-Means clustering
kmeans = KMeans(n_clusters = k_clusters, random_state = 0).fit(berlin_grouped_clustering)

print('Generated cluster labels', kmeans.labels_)

# add the clustering labels
if 'Cluster Labels' in BerlinTop5Boroughs_nearby_venues_sorted.columns:
    BerlinTop5Boroughs_nearby_venues_sorted.drop(columns=['Cluster Labels'], axis=1, inplace=True)
    BerlinTop5Boroughs_nearby_venues_sorted.insert(0, 'Cluster Labels', kmeans.labels_)
else:
    BerlinTop5Boroughs_nearby_venues_sorted.insert(0, 'Cluster Labels', kmeans.labels_)


berlin_nearby_merged = df_berlinTop5Boroughs

# merge berlin_nearby_merged with berlinTop5Boroughs data to add latitude/longitude for each Boroughs
berlin_nearby_merged = berlin_nearby_merged.join(BerlinTop5Boroughs_nearby_venues_sorted.set_index('Boroughs'), on='Boroughs')

# Also add the calculated restaurant, coffee shops and hotels for each boroughs
berlin_nearby_merged = berlin_nearby_merged.join(df_berlinTop5Boroughs_cafeRestaurantHotel.set_index('Boroughs'), on='Boroughs')

# Print DataType, shape and DataFrame called df_berlin
#print("Data types of each columns:\n", berlin_nearby_merged.dtypes)
print("\nThe shape of dataFrame is:", berlin_nearby_merged.shape)
berlin_nearby_merged # check the last 3 columns!
```

```
Generated cluster labels [0 2 1 0 0]

The shape of dataFrame is: (5, 23)
```

	Boroughs	Neighborhoods	Area_sqkm	Population	Total Businesses	Total Visitors in 2019	Rent Price in 2019 [€/m²-month]	Latitude	Longitude	Cluster Labels	...	4th Most Common Venue	5th Most Common Venue	6th M Comm Ve
1	Charlottenburg-Wilmersdorf	Charlottenburg, Wilmersdorf, Schmargendorf, Gr...	64.62	342332	29324	9157433	13.0	52.498889	13.284917	0	...	Trattoria/Osteria	Vietnamese Restaurant	G Fit Ce
2	Mitte	Mitte, Moabit, Hansaviertel, Tiergarten, Weddi...	39.48	384172	28553	20809270	14.0	52.516667	13.366667	1	...	Monument / Landmark	Art Museum	Hote
3	Pankow	Prenzlauer Berg, Weißensee, Blankenburg, Heine...	103.26	407765	22628	1924337	12.1	52.568889	13.402222	0	...	Italian Restaurant	Drugstore	Ba
4	Friedrichshain-Kreuzberg	Friedrichshain, Kreuzberg	20.18	289762	20978	6495909	13.6	52.500000	13.433333	2	...	Café	Yoga Studio	I
5	Tempelhof-Schöneberg	Schöneberg, Friedenau, Tempelhof, Mariendorf, ...	53.08	351644	18926	2948106	11.8	52.466667	13.383333	0	...	Italian Restaurant	Drugstore	G Fit Ce

5 rows × 23 columns

By using Folium library, 3 clusters can be displayed on a leaflet map (Figure 9).

As the total numbers of restaurants and café in each borough are calculated, 3 clusters can also be displayed on a leaflet map using Folium library as function of the total numbers of restaurants and cafés as shown in Figure 10.

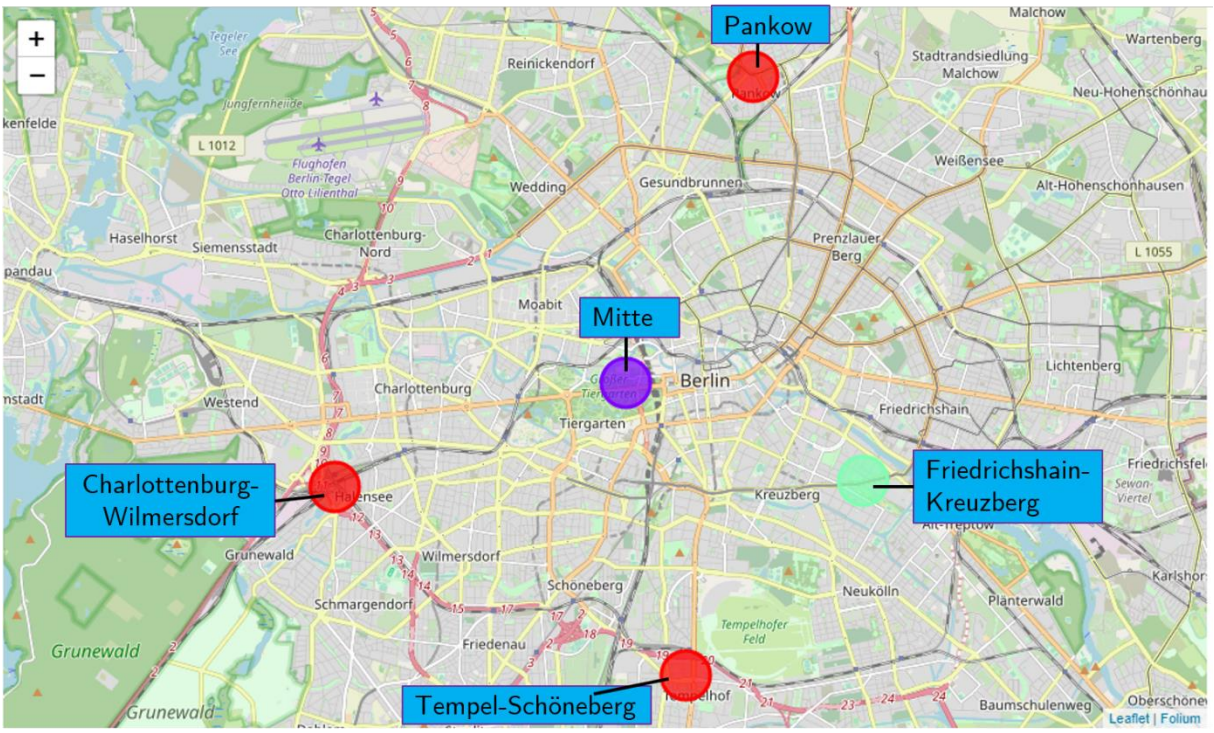

Figure 9: A map of Berlin with each of our selected 5 boroughs clustered into 3 groups based on their popular venue categories in each borough.

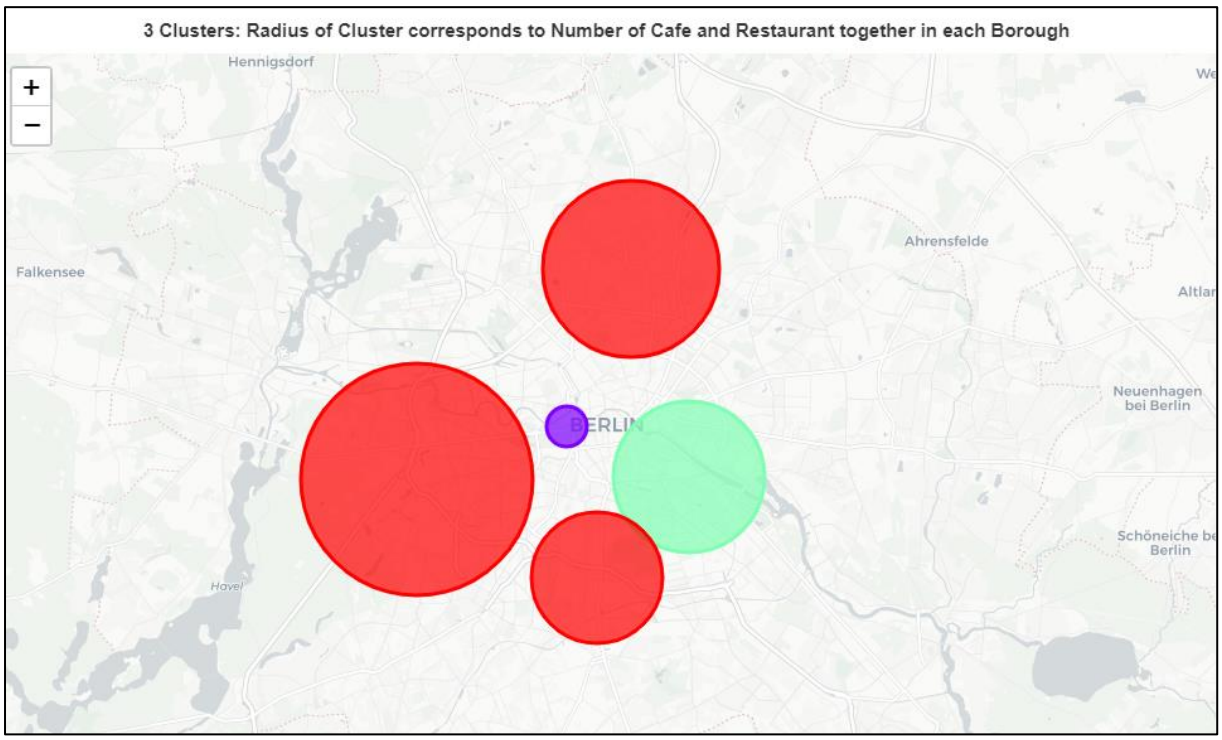

Figure 10: A map of Berlin with each of our selected 5 boroughs clustered into 3 groups based on their popular venue categories in each borough. The size of each circles represents the number of restaurants and cafés together listed as popular venues.

4 Results and Discussion

To find a best spot in Berlin for opening a café cum restaurant, we have retrieved Berlin's geographic, business, tourist and rent price datasets for all 12 boroughs of berlin from various sources, and then explored these datasets for the preferred 5 boroughs and analyzed mainly on the restaurant and café. At the end, our analysis shows us that:

- 43% of total tourists in Berlin are hosted by only Mitte borough, whereas another four boroughs together host 43% of total tourist in Berlin.
- 30% of total businesses of berlin are in Charlottenburg-Wilmersdorf (15.26%) and Mitte (14.86%) boroughs; whereas Pankow, Friedrichshain-Kreuzberg and Tempelhof-Schöneberg boroughs cover another 31.7% of total businesses.
- The most common venues in the preferred 5 boroughs are Café, followed by Hotel, Italian Restaurant, Park, Supermarket and Coffee Shops.
- Charlottenburg-Wilmersdorf, Friedrichshain-Kreuzberg, Pankow and Tempelhof-Schöneberg boroughs have restaurants and cafés as popular venues, whereas Mitte borough is dominated by Hotel.
- Boroughs clustering based on their most popular venues grouped Charlottenburg-Wilmersdorf, Pankow with Tempelhof-Schöneberg into a cluster, and Mitte and Friedrichshain-Kreuzberg as their own independent clusters.
- According to Homeday.de report, the Mitte, Friedrichshain-Kreuzberg and Charlottenburg-Wilmersdorf boroughs are expensive places to live with rent price 14, 13.6 and 13 €/m^2-month, respectively.
- Tempelhof-Schöneberg and Pankow boroughs are more affordable with 11.8 and 12.1 €/m^2-month, respectively.

According to this analysis, Mitte and Charlottenburg-Wilmersdorf boroughs are more favorable to starting a café cum restaurant compare to the other boroughs despite of having a high cost of rent. Because, both boroughs have a very high foot traffic by business employees and tourists compared to any other boroughs. Out of Mitte and Charlottenburg-Wilmersdorf, Mitte borough will provide least competition for café cum restaurant as the frequency of café-restaurant as common venue is very low compared to the Charlottenburg-Wilmersdorf. Moreover, the number of tourists in Mitte borough is almost 2.3 times than Charlottenburg-Wilmersdorf. Hence, Mitte borough can be best choice to open a café cum restaurant.

Some major drawbacks of this analysis - the boroughs clustering is completely based on the most popular venues obtained from Foursquare data; the business and tourists data for each boroughs rather than the neighborhood; average rent price of boroughs – all these play major role in the analysis.

5 Conclusion

In this data science project, I have retrieved, explored, and analyzed the sets of Berlin's business, tourist, and geographical data to find out the best spot for opening a café cum restaurant; and discussed the results in great detail. Through this project, I have made to use various analytical tools like the use of python libraries to scrape and manipulate the data sets, Foursquare API to explore the neighborhoods, and Folium leaflet map to visualize the clusters. This experience of use of these analytical tools helped me to understand that how a real-life business problem can be solved by using various data science tools.

6 References

[1]. Berlin.de, The Official Website of Berlin, Germany, accessed 21 July 2020, https://www.berlin.de/sen/wirtschaft/wirtschaft/branchen/tourismus/tourismus-in-zahlen/

[2]. Berliner Wirtschaft 2019 stark gewachsen, Wirtschaft aktuell, Senatsverwaltung für Wirtschaft, Energie und Betriebe, Germany, April 2020.

[3]. Economic Development, The Official Website of Berlin, Germany, accessed 21 July 2020, https://www.berlin.de/en/business-and-economy/economic-center/5611367-4011028-economic-development.en.html

[4]. Verwaltungsgliederung Berlins, accessed 21 July 2020. In Wikipedia. Retrieved from https://de.wikipedia.org/wiki/Verwaltungsgliederung_Berlins

[5]. Amt für Statistik Berlin-Brandenburg; Germany, accessed 21 July 2020, https://www.statistik-berlin-brandenburg.de/datenbank/inhalt-datenbank.asp

[6]. Homeday.de, Homeday mein Immobilienmakler, accessed 25 July 2020, https://www.homeday.de/de/blog/mietpreise-berlin-2019/