Capstone Project - The Battle of Neighborhoods in Berlin

Coursera Applied Data Science with Capstone

Source: https://media.globalchampionstour.com/cache/750x429/assets/berlin.jpg

Table of contents

1. Introduction

A café owner would like to open a new coffee shop in Berlin, Germany. She already has two cafés in Lisbon, Portugal, and Dublin, Irland. Her new café will add up to her current business. The choice of the city was based on two factors: a big city and the hipster community. Furthermore, Berlin is well-known with low rent prices and a variety of international representatives. The café is a bio vegan café. It will offer not only a fresh brew bio coffee but also bio pastry entirely made by vegan products. The owner is looking at where will be the best neighbourhood in Berlin to open her new business venture. With a variety of areas, Berlin offers a wide range of cuisine and refreshment places. Therefore, it is essential that the café is in the right neighbourhood, where it could reach the target clients.

There are several factors under consideration when someone wants to open a new business. Some of them are the population, the price of the rent, the average salary of the residents, the lifestyle of the residents, and so on. One of the factors is the number of competition which offers similar services. In our case, every other place which serves primarily coffee products and high snacks is under that category. Therefore, such a place to open a new café should be in an area where the competition is not dense.

A comparison between Berlin's neighbourhoods will be conducted. Based on the venue category related to coffee places, the top choices in each district will be presented by using the Foursquare API. Using k-means clustering, an unsupervised machine learning method, which will split the n number of samples into k number clusters, where each sample belongs to one of the clusters. Then, the results from the analysis will be presented, and a recommendation of the best place for opening a bio vegan café will be proposed. The proposal will be shortly discussed concerning the business goal.

2. Data

2.1. Data Sources

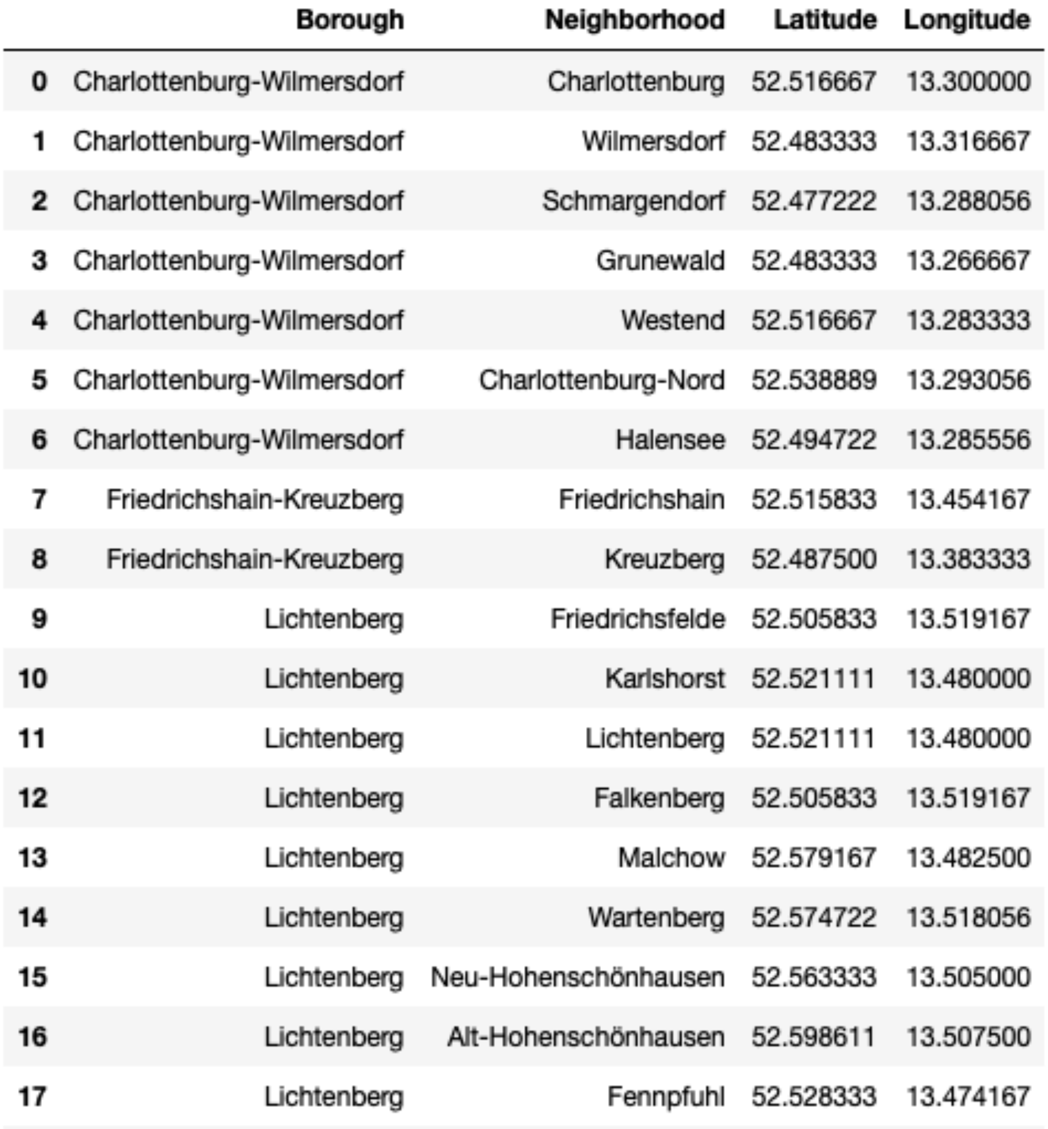
Berlin has a total of 12 boroughs and 96 neighbourhoods. A data set, containing their names and coordinates, was created. The coordinates were based on information posted on Wikipedia. A pandas dataframe created from the data table is shown below.

Following data sources will be needed to extract/generate the required information:

- List of boroughs and their neighbourhoods in Berlin collected from Wikipedia
- Number of coffee places and location in every neighbourhood will be obtained using Foursquare API

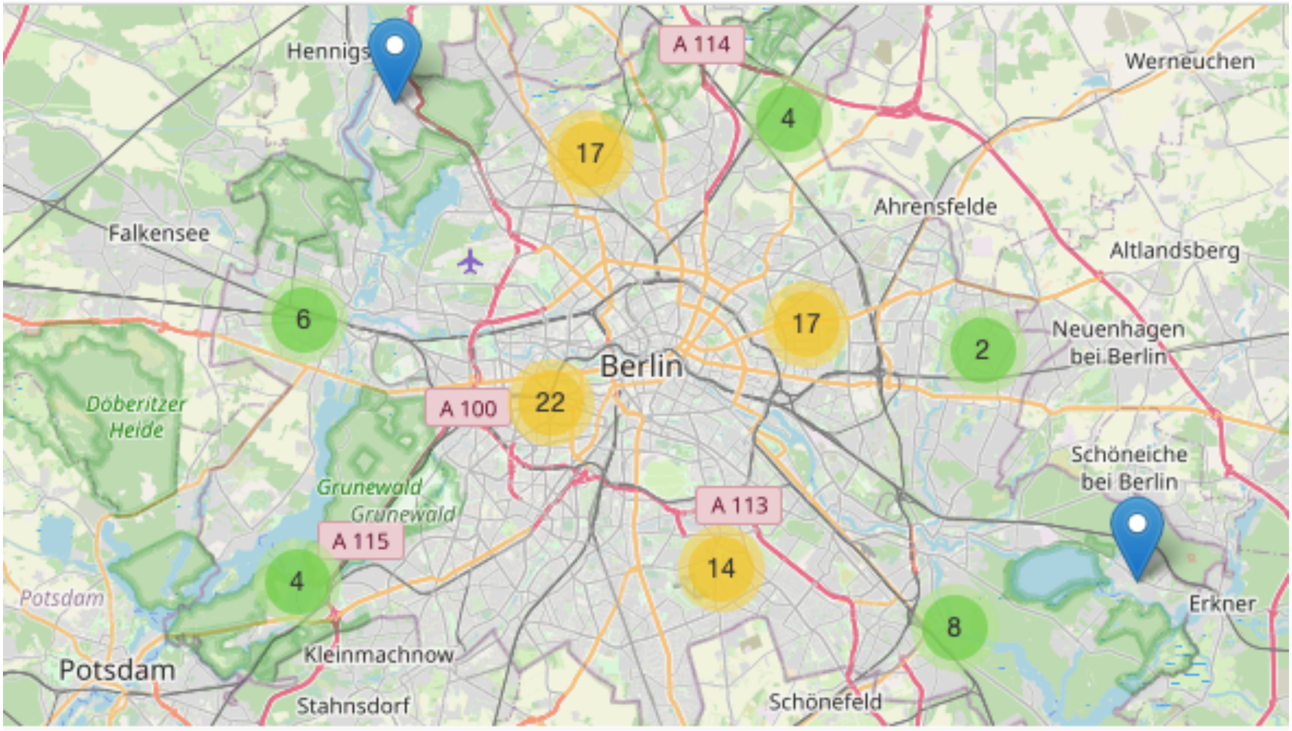
2.2. Data Preparation

No Wiki page was found which counting the coordinates of all neighbourhoods in Berlin, Germany. Therefore, a new table was created and imported as csv file in the IBM Cloud Notebook. The table contains the borough, the neighbourhood, the latitude, and the longitude.



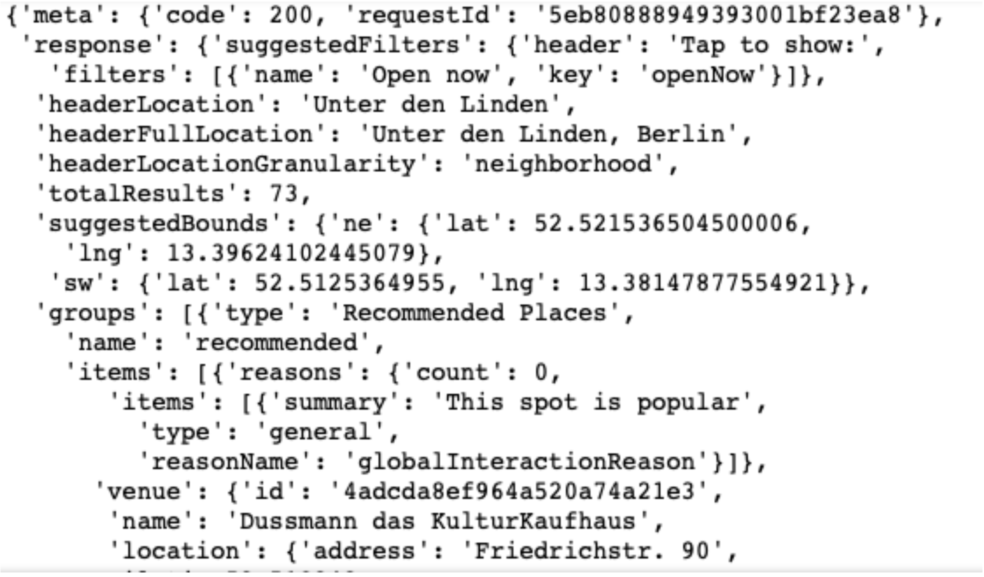
	Borough	Neighborhood	Latitude	Longitude
0	Charlottenburg-Wilmersdorf	Charlottenburg	52.516667	13.300000
1	Charlottenburg-Wilmersdorf	Wilmersdorf	52.483333	13.316667
2	Charlottenburg-Wilmersdorf	Schmargendorf	52.477222	13.288056
3	Charlottenburg-Wilmersdorf	Grunewald	52.483333	13.266667
4	Charlottenburg-Wilmersdorf	Westend	52.516667	13.283333
5	Charlottenburg-Wilmersdorf	Charlottenburg-Nord	52.538889	13.293056
6	Charlottenburg-Wilmersdorf	Halensee	52.494722	13.285556
7	Friedrichshain-Kreuzberg	Friedrichshain	52.515833	13.454167
8	Friedrichshain-Kreuzberg	Kreuzberg	52.487500	13.383333
9	Lichtenberg	Friedrichsfelde	52.505833	13.519167
10	Lichtenberg	Karlshorst	52.521111	13.480000
11	Lichtenberg	Lichtenberg	52.521111	13.480000
12	Lichtenberg	Falkenberg	52.505833	13.519167
13	Lichtenberg	Malchow	52.579167	13.482500
14	Lichtenberg	Wartenberg	52.574722	13.518056
15	Lichtenberg	Neu-Hohenschönhausen	52.563333	13.505000
16	Lichtenberg	Alt-Hohenschönhausen	52.598611	13.507500
17	Lichtenberg	Fennpfuhl	52.528333	13.474167

A map of Berlin was plotted using Folium library to show the neighbourhoods location. The neighbourhood markers were clustered for better visualisation.



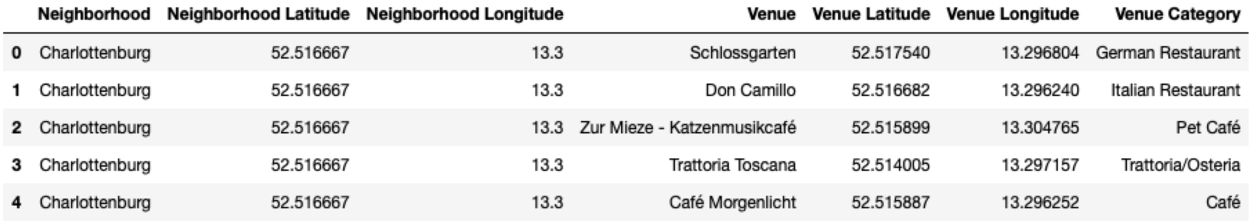
2.3. Foursquare API usage

The API offered by Foursquare was used to get more information on venues in Berlin. A query was created to get the top 100 venues within a radius of 500 m from all neighbourhoods. The result was a json file. Sample of the file is shown below.



```
{'meta': {'code': 200, 'requestId': '5eb808889493930 01bf23ea8'},
 'response': {'suggestedFilters': {'header': 'Tap to show:',
   'filters': [{'name': 'Open now', 'key': 'openNow'}]},
  'headerLocation': 'Unter den Linden',
  'headerFullLocation': 'Unter den Linden, Berlin',
  'headerLocationGranularity': 'neighborhood',
  'totalResults': 73,
  'suggestedBounds': {'ne': {'lat': 52.521536504500006,
    'lng': 13.39624102445079},
   'sw': {'lat': 52.5125364955, 'lng': 13.38147877554921}},
  'groups': [{'type': 'Recommended Places',
    'name': 'recommended',
    'items': [{'reasons': {'count': 0,
       'items': [{'summary': 'This spot is popular',
         'type': 'general',
         'reasonName': 'globalInteractionReason'}]},
      'venue': {'id': '4adcda8ef964a520a74a21e3',
       'name': 'Dussmann das KulturKaufhaus',
       'location': {'address': 'Friedrichstr. 90',
```

After that the json file was cleaned and structured into a pandas dataframe contains only the relevant information such as venue’s name, category, and location.



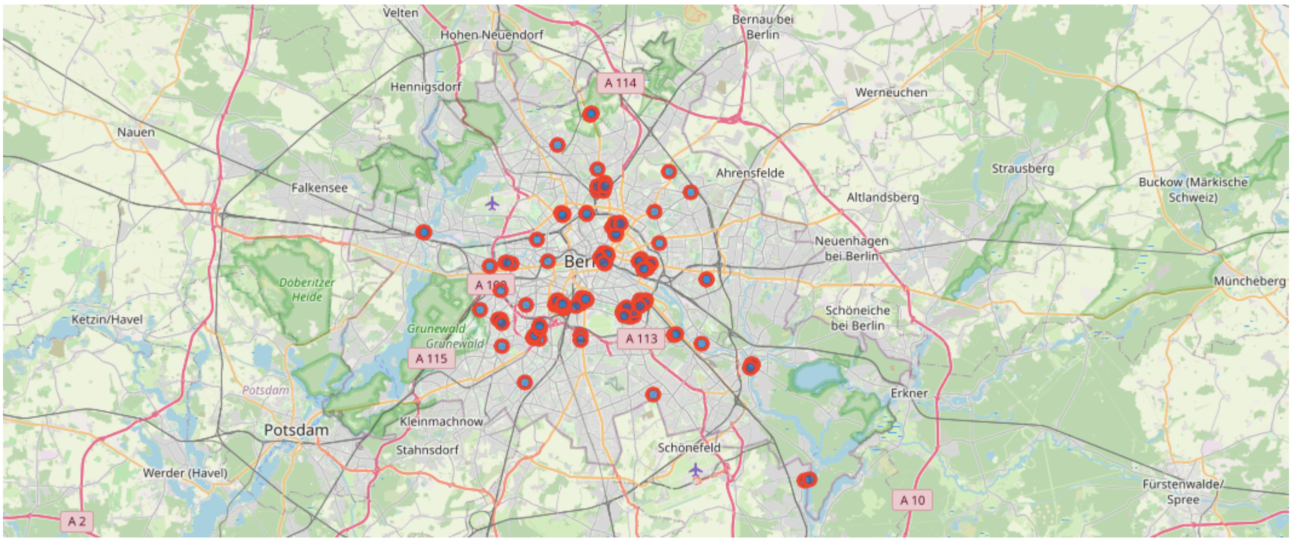
	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Charlottenburg	52.516667	13.3	Schlossgarten	52.517540	13.296804	German Restaurant
1	Charlottenburg	52.516667	13.3	Don Camillo	52.516682	13.296240	Italian Restaurant
2	Charlottenburg	52.516667	13.3	Zur Mieze - Katzenmusikcafé	52.515899	13.304765	Pet Café
3	Charlottenburg	52.516667	13.3	Trattoria Toscana	52.514005	13.297157	Trattoria/Osteria
4	Charlottenburg	52.516667	13.3	Café Morgenlicht	52.515887	13.296252	Café

3. Methodology

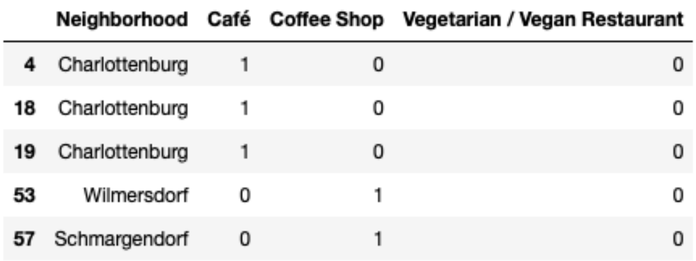
3.1. Exploratory analysis

The dataset was filtered on neighbourhoods that had venues such as “Café”, “Coffee Shop”, and “Vegetarian/Vegan Restaurant”. The last word was not randomly chosen. Foursquare has not different categories of coffee places such as vegan or organic. Therefore, some of the owners in Berlin put their vegan cafés under the category of Vegetarian/Vegan Restaurant. Most likely, not so many owners categorised in such manner their business, but to more complete the analysis, I added this category as well. As the new café will serve as well as light snacks, it can be put under this category.

A map displaying only the venues which contained the three categories described above.

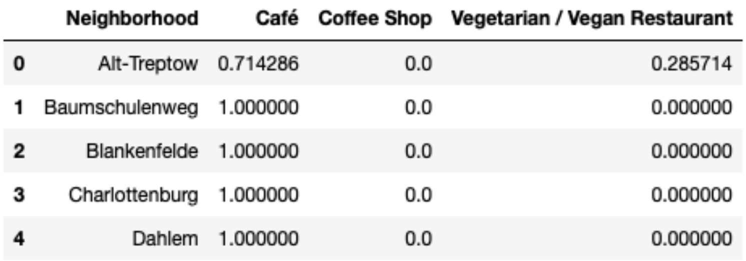


A one hot encoding was used to set all categories to binary values.



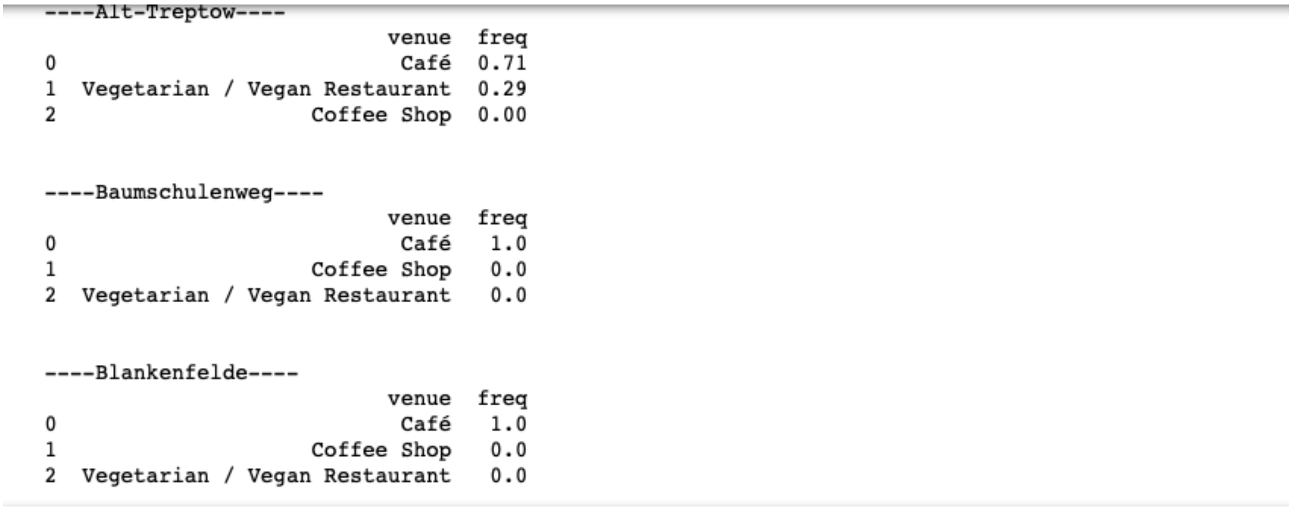
	Neighborhood	Café	Coffee Shop	Vegetarian / Vegan Restaurant
4	Charlottenburg	1	0	0
18	Charlottenburg	1	0	0
19	Charlottenburg	1	0	0
53	Wilmersdorf	0	1	0
57	Schmargendorf	0	1	0

After that, a data frame was created which grouped the rows by Neighborhood and by taking the each category's mean frequency.



	Neighborhood	Café	Coffee Shop	Vegetarian / Vegan Restaurant
0	Alt-Treptow	0.714286	0.0	0.285714
1	Baumschulenweg	1.000000	0.0	0.000000
2	Blankenfelde	1.000000	0.0	0.000000
3	Charlottenburg	1.000000	0.0	0.000000
4	Dahlem	1.000000	0.0	0.000000

For each one of the neighbourhood the frequency of the coffee places was extracted.

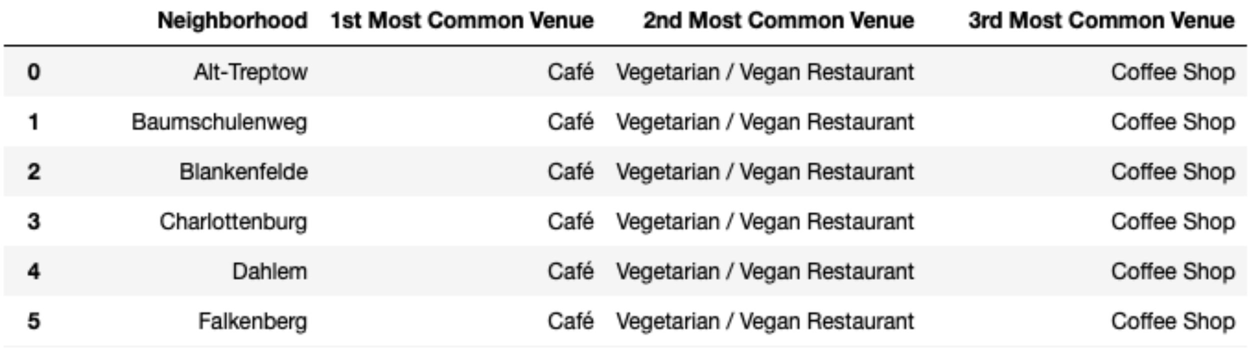


```
----Alt-Treptow----
                          venue  freq
0                          Café  0.71
1  Vegetarian / Vegan Restaurant  0.29
2                   Coffee Shop  0.00


----Baumschulenweg----
                          venue  freq
0                          Café   1.0
1                   Coffee Shop   0.0
2  Vegetarian / Vegan Restaurant   0.0


----Blankenfelde----
                          venue  freq
0                          Café   1.0
1                   Coffee Shop   0.0
2  Vegetarian / Vegan Restaurant   0.0
```

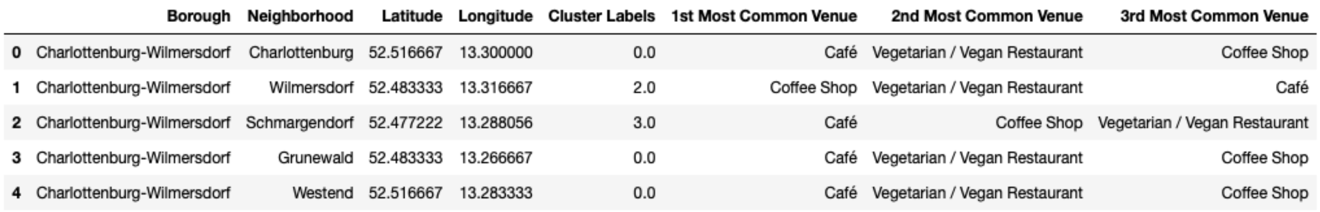
Using the information from the neighbourhoods and the venues, the top 3 venues for each neighbourhood was presented. Dues to the limited number of categories, the variation in the recommendation was not high.



	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
0	Alt-Treptow	Café	Vegetarian / Vegan Restaurant	Coffee Shop
1	Baumschulenweg	Café	Vegetarian / Vegan Restaurant	Coffee Shop
2	Blankenfelde	Café	Vegetarian / Vegan Restaurant	Coffee Shop
3	Charlottenburg	Café	Vegetarian / Vegan Restaurant	Coffee Shop
4	Dahlem	Café	Vegetarian / Vegan Restaurant	Coffee Shop
5	Falkenberg	Café	Vegetarian / Vegan Restaurant	Coffee Shop

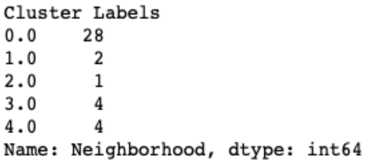
3.2. K-means clustering

The number of clusters was set to 5 clusters. In order to make the recommendations, a popularity recommendation approach was used. A new data frame was created combining the top 3 venues, neighbourhood, and their locations with the cluster labels. The biggest cluster was the first one, Cluster 0, which contained most of the neighbourhoods.



	Borough	Neighborhood	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
0	Charlottenburg-Wilmersdorf	Charlottenburg	52.516667	13.300000	0.0	Café	Vegetarian / Vegan Restaurant	Coffee Shop
1	Charlottenburg-Wilmersdorf	Wilmersdorf	52.483333	13.316667	2.0	Coffee Shop	Vegetarian / Vegan Restaurant	Café
2	Charlottenburg-Wilmersdorf	Schmargendorf	52.477222	13.288056	3.0	Café	Coffee Shop	Vegetarian / Vegan Restaurant
3	Charlottenburg-Wilmersdorf	Grunewald	52.483333	13.266667	0.0	Café	Vegetarian / Vegan Restaurant	Coffee Shop
4	Charlottenburg-Wilmersdorf	Westend	52.516667	13.283333	0.0	Café	Vegetarian / Vegan Restaurant	Coffee Shop

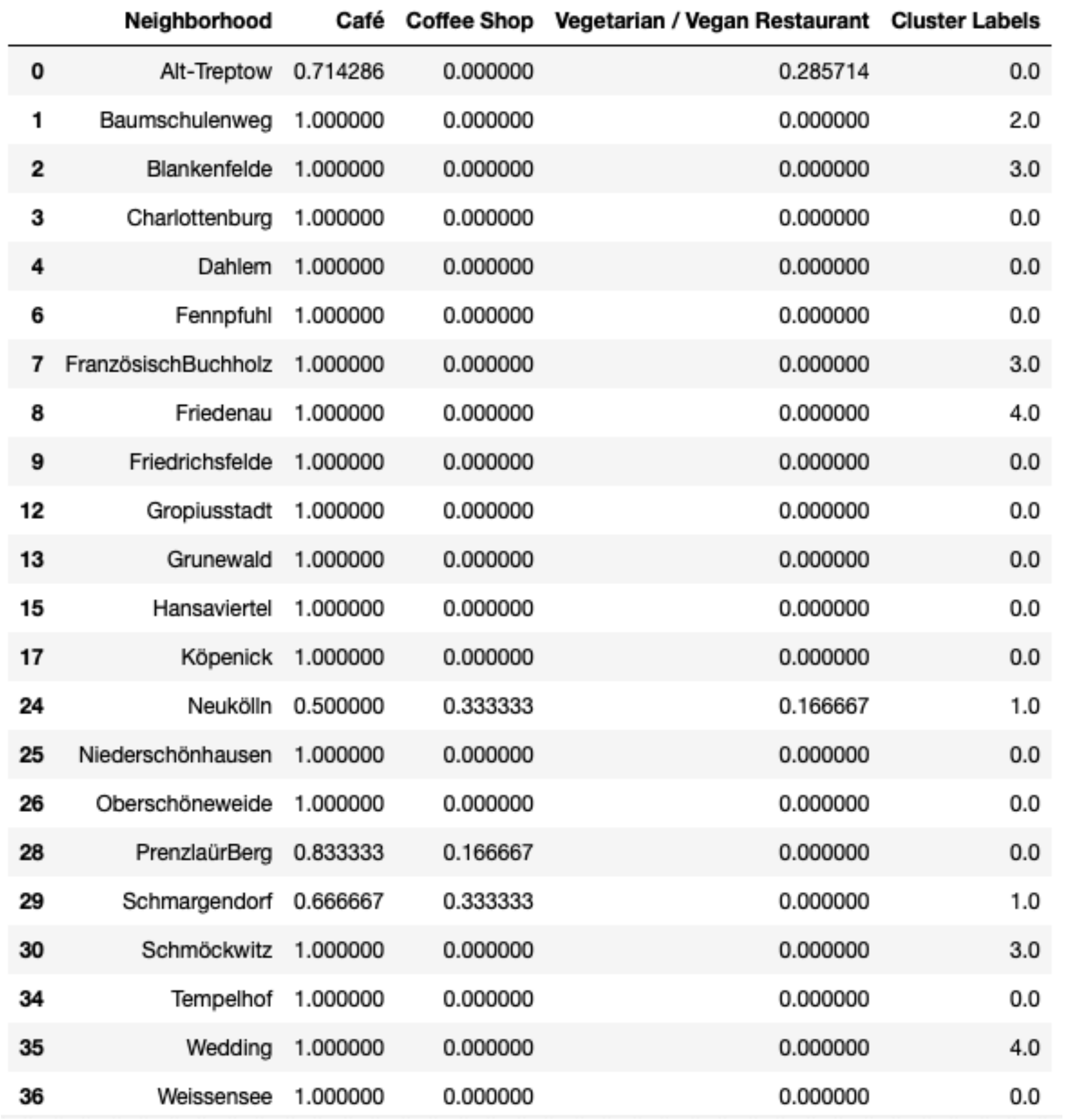
The number of the neighbourhood by cluster was calculated.



```
Cluster Labels
0.0    28
1.0     2
2.0     1
3.0     4
4.0     4
Name: Neighborhood, dtype: int64
```

4. Results and Discussion

It was observed from the frequency analysis on coffee places that the most popular place from the three categories was Café. However, in some neighbourhoods such as Gesunfbrunnen, Mitte, Prenzlaürberg, and Schmargendorf, the popularity was shared among Café and Coffee Shop. Furthermore, there were some neighbourhoods such as Friedrichshein and Neukölln, where the Vegan Restaurant was also a popular place.

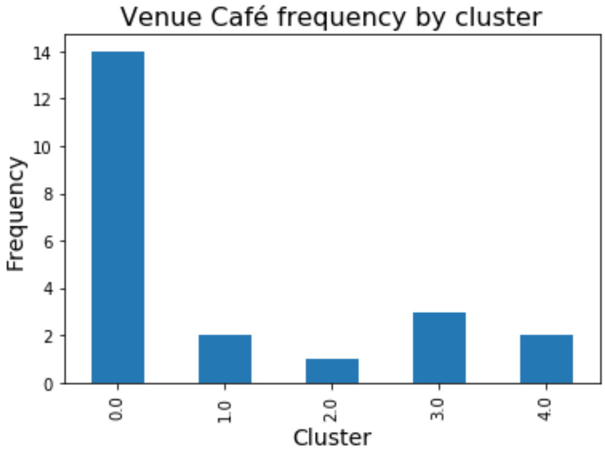


	Neighborhood	Café	Coffee Shop	Vegetarian / Vegan Restaurant	Cluster Labels
0	Alt-Treptow	0.714286	0.000000	0.285714	0.0
1	Baumschulenweg	1.000000	0.000000	0.000000	2.0
2	Blankenfelde	1.000000	0.000000	0.000000	3.0
3	Charlottenburg	1.000000	0.000000	0.000000	0.0
4	Dahlem	1.000000	0.000000	0.000000	0.0
6	Fennpfuhl	1.000000	0.000000	0.000000	0.0
7	FranzösischBuchholz	1.000000	0.000000	0.000000	3.0
8	Friedenau	1.000000	0.000000	0.000000	4.0
9	Friedrichsfelde	1.000000	0.000000	0.000000	0.0
12	Gropiusstadt	1.000000	0.000000	0.000000	0.0
13	Grunewald	1.000000	0.000000	0.000000	0.0
15	Hansaviertel	1.000000	0.000000	0.000000	0.0
17	Köpenick	1.000000	0.000000	0.000000	0.0
24	Neukölln	0.500000	0.333333	0.166667	1.0
25	Niederschönhausen	1.000000	0.000000	0.000000	0.0
26	Oberschöneweide	1.000000	0.000000	0.000000	0.0
28	PrenzlaürBerg	0.833333	0.166667	0.000000	0.0
29	Schmargendorf	0.666667	0.333333	0.000000	1.0
30	Schmöckwitz	1.000000	0.000000	0.000000	3.0
34	Tempelhof	1.000000	0.000000	0.000000	0.0
35	Wedding	1.000000	0.000000	0.000000	4.0
36	Weissensee	1.000000	0.000000	0.000000	0.0

Then a new data frame was created, where each venue category was a different Series and the cluster labels was another Series as well. Each neighbourhood was presented by a different cluster label. As the biggest cluster, Cluster 0 contained as well the most results from the Café venue.

5. Conclusion

The best area to open a new café is Cluster 2. There is only one Café venue there. They were followed by Cluster 1 and Cluster 4.



A quick reference to the table above showed that in Cluster 1, the other venue, Coffee Shop, is popular as well. It is important to be pointed out that a venue Café is a popular place in almost every neighbourhood as many people visit this type of venue in their free time. Unfortunately, until now, there is no clear way to know, based only on the Foursquare information, which café is vegan or offers only organic products. Therefore, further analysis should be carried out to create more precise recommendations for the new business venture.