Coursera Capstone

IBM Data Science Capstone Project

Strategic Locations for Supermarket-Chain in Berlin, Germany

By: Helal Chowdhury
April 2020

1. Background

Almost every single adult living in the world needs to visit supermarkets or marketplaces at least few times in a month. Supermarket is a place where you can buy most the things necessary for everyday-life. A traditional supermarket offers groceries, utensils, toiletries cosmetics etc. In addition, few supermarkets are multipurpose offering furniture, clothing etc. and many more. Considering demands from every part of the city, often we can see the appearance of one or more supermarkets. Berlin is a well-developed and capital city of Germany, with lots of business opportunities and business friendly environment attracting many different players into the market. To be specific, there are some hundred supermarkets in the city of Berlin. Since the overall population in Berlin is on the rising trend, new supermarkets or new branches of an existing chain are appearing on a regular basis. This is an opportunity for any new company or an existing one offering new facilities in different parts of Berlin. However, the market is highly competitive, for any business decision, opening a number of new locations needs serious consideration and is a lot more complicated than usual. Any new business decision or expansion endeavor from a venture needs to be reviewed carefully and strategically so that the return on investment will be sustainably reasonable, and more importantly can be considerably less risker. Particularly, strategic locations are of paramount importance in this case. Selecting few random locations or only central locations might not lead to success always.

2. Business Problem

Berlin is one of the densely populated cities in Germany, with more than 3.6 million residents and an average of almost 1.2 million visitors each month [1]. Let’s say a large supermarket chain is interested in opening some new branches in Berlin and therefore looking for some strategic locations. The objective of this project is to analyze and select such locations, say five for example, in the capital city of Berlin. Using data analysis and machine learning techniques like clustering, this project aims to provide answer to the business question: In the city of Berlin, if a local or foreign supermarket giant looks for opening few new spots, which strategic locations should be preferred considering business potential?

3. Target Audience

This project is particularly useful for local or foreign supermarket chains or for a new venture willing to do similar business in the capital city of Berlin. The objective is to locate and recommend to the management of interested parties which set of neighborhoods will be the best choice to start their services. The management also expects to understand the rationale of the recommendations in the report. This project is timely as the city is currently shows an increasing trend of population, especially due to the flood of new people from war-hit zones. The success criteria of this project will be a good recommendation of the neighborhood choice to the management based on some key factors: higher population, less competition, higher density per km^2 and neighborhood similarities.

4. Working with data

To find the answer of the above business problem, we will use the following data:

- List of neighborhoods including borough in Berlin from Wikipedia [2].
- Latitude and longitude of all neighborhoods. This is required for getting venue data and for the visualization of neighborhoods. Related data will be obtained using Python geocoder package.
- Population and density of each neighborhood with which we will shortlist the candidate locations, from Wikipedia [2].
- Venue data with the density of supermarkets with which we will perform clustering on the neighborhoods in order to identify strategically similar locations.

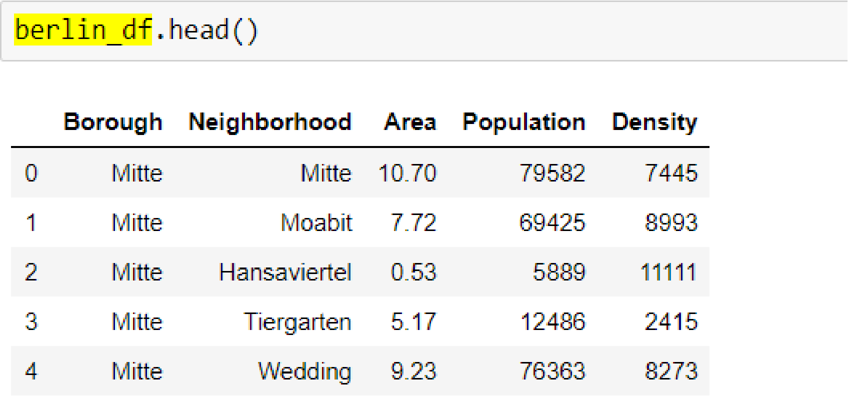
The Wikipedia page in [2] contains all basic information of Berlin including borough, neighborhood, population, area, density etc. for each locality. The page shows that there are 12 borough and 96 neighborhoods. We will use web scraping techniques to extract the data from the Wikipedia page, with the help of Python Pandas. Then we will get the geographical coordinates of the neighborhoods using Python Geocoder package which will give us latitudes and longitudes. After that, we will use APIs from one of the popular sources Foursquare to get the venue data from all neighborhoods. Foursquare has one of the largest databases of 105+ million places and is used by over 125,000 developers [3]. One of the features of Foursquare APIs is to provide a list of venues within a specific location, based on latitude/longitude coordinates and a radius. By passing the proper parameters via an HTTP request, we get the required data. The *location* object contains the coordinates of each venue, which will be used to associate it with its respective neighborhood. The *categories* array will be used to categorize the neighborhood. Basically, we will count how many venues from all available categories are found on each neighborhood, and then use that information to compare neighborhoods in Berlin. Foursquare API provides many categories of the venue data; we are particularly interested in similar neighborhoods considering supermarket category in order to solve the business problem.

The preprocessed data will help the K-means algorithm to segment and cluster these neighborhoods so that we can group them together to understand their similarities and what best we can do for these types of neighborhoods. With all these features, techniques and data, we will then be able to come up with a best recommendation to the management of interested companies, that what are the best neighborhoods set for them to start their services? For an example, we will not want to enter a neighborhood whereby there is already a high concentration of grocery stores available or there is a high trending of such stores upcoming in the neighborhood; we will like to recommend a neighborhood where we know that there will be a higher demand of such delivery service due to the lack of supply in that area. This project requires many data science skills, from web scraping (from Wikipedia), working with API (of Foursquare), data cleaning, data wrangling, to machine learning (K-means clustering) and map visualization (using Folium).

5. Methodology

This section is presented in six steps. First step is related to data preprocessing, while step 2-5 are dedicated to exploratory data analysis and primary visualization of the neighborhoods. At the end of step 5 we will have a rough primary selection of the top candidate-locations. The clustering algorithm is described in the final step.

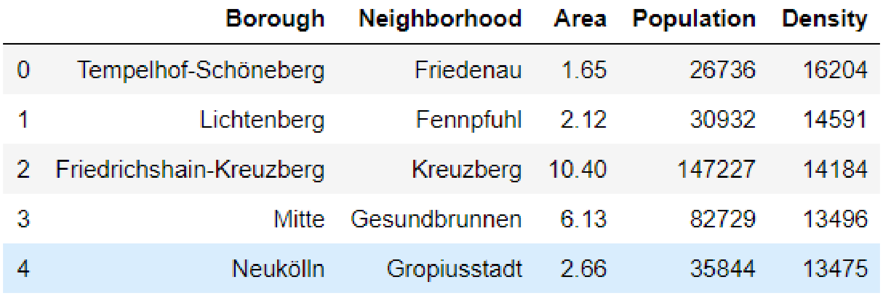
- **Step-1: Extracting Borough-Neighborhood-Area-Population-Density (96 rows)** Web scraping has been applied considering the wiki page in [2]. Using ***pandas.read_html()*** function all 17 tables have been copied to a list of dataframes. Among them 12 tables for 12 boroughs are necessary. Each one contains neighborhood name, area, population, density and corresponding map as NaN. These 12 dataframes are combined into one with 96 neighborhoods, then borough name of each neighborhood has been attached, and Map column has been dropped. Unfortunately each neighborhood name comes with an extra area code, which is unnecessary. So the area codes from each name have been removed. The processed combined dataframe is named as ***berlin_df*** as shown below:



```
berlin_df.head()
```

	Borough	Neighborhood	Area	Population	Density
0	Mitte	Mitte	10.70	79582	7445
1	Mitte	Moabit	7.72	69425	8993
2	Mitte	Hansaviertel	0.53	5889	11111
3	Mitte	Tiergarten	5.17	12486	2415
4	Mitte	Wedding	9.23	76363	8273

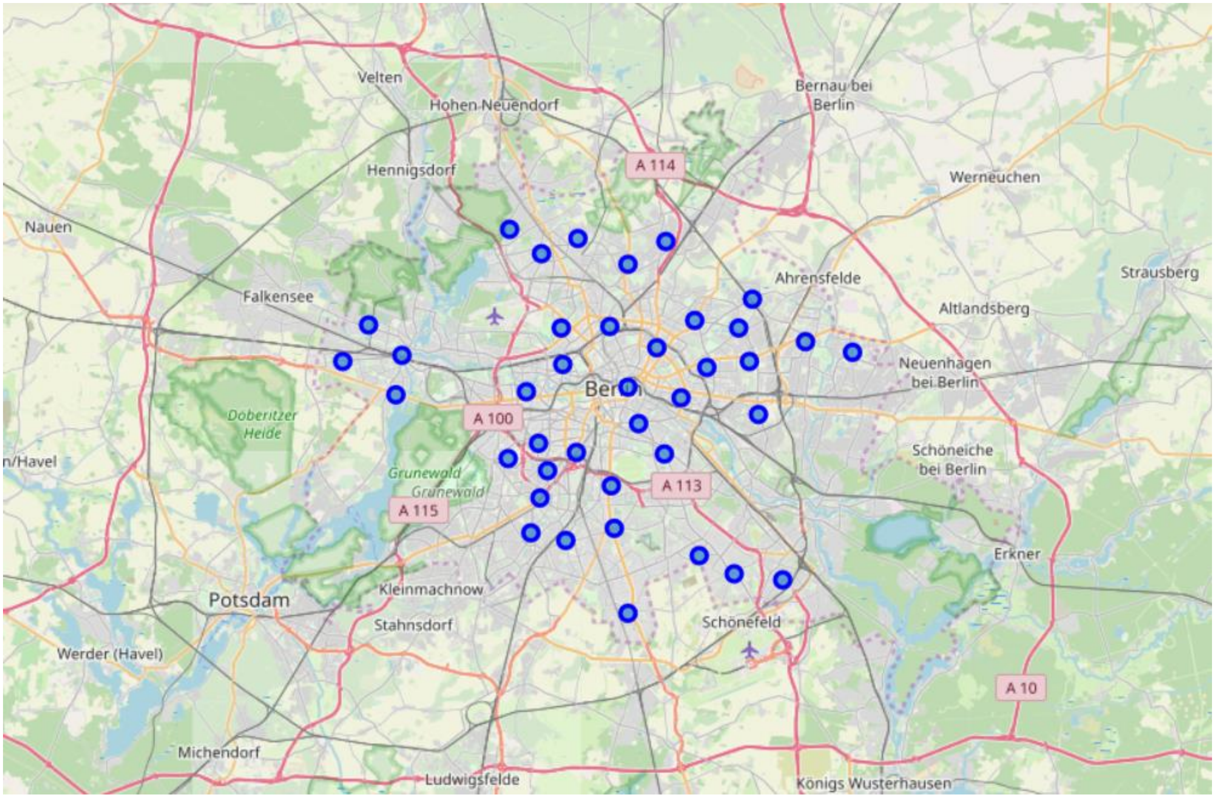
- **Step-2: Reducing candidate neighborhood to 40**
There are 96 neighborhoods in ***berlin_df,*** which is reduced to 40 based on higher population and higher density. After sorting according to population, 46 rows with less population have been dropped, and then 10 more rows have been excluded based on lower density per km^2. The assumption here is that the regions with very lower number of population and lower density do not hold the promise of good supermarket business. Finally we have 40 neighborhoods in ***berlin_df*** as shown below***:***



	Borough	Neighborhood	Area	Population	Density
0	Tempelhof-Schöneberg	Friedenau	1.65	26736	16204
1	Lichtenberg	Fennpfuhl	2.12	30932	14591
2	Friedrichshain-Kreuzberg	Kreuzberg	10.40	147227	14184
3	Mitte	Gesundbrunnen	6.13	82729	13496
4	Neukölln	Gropiusstadt	2.66	35844	13475

- **Step-3: Adding latitude/longitude to each neighborhood and visualization**
Using the function ***Nominatim()*** from ***geopy.geocoders*** package latidute/longitude data have been extracted for each neighborhood and added to ***berlin_df*** . Using ***Folium*** package all 40 neighborhoods have been shown in the below map. This map allows us to perform a sanity

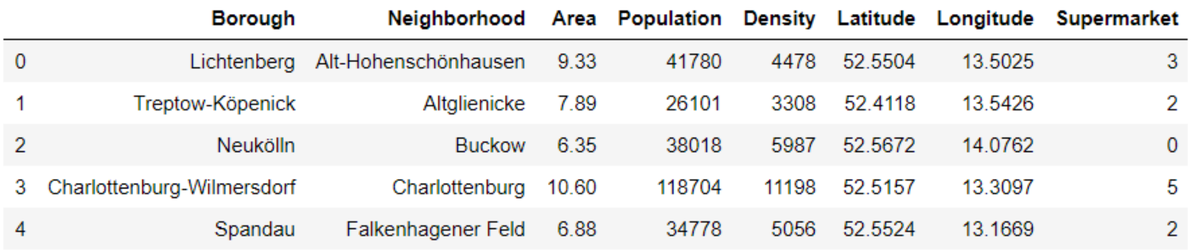
check to make sure that the geographical coordinate's data returned by Geocoder are correctly plotted in the city of Berlin.



- **Step-4: Adding number of supermarkets to 40 candidate neighborhoods**

For using Foursquare APIs we need to register a Foursquare Developer Account in order to obtain the Foursquare ID and Foursquare secret key. We then make API calls to Foursquare passing in the geographical coordinates of the neighborhoods in a Python loop. Foursquare will return the venue data in JSON format and we will extract the venue name, venue categor y, venue latitude and longitude. With the data, we can check how many venues were returne d for each neighborhood and examine how many unique categories can be curated from all the returned venues. Then, we will analyze each neighborhood by grouping the rows by neighborhood and taking the frequency of occurrence of each venue category.

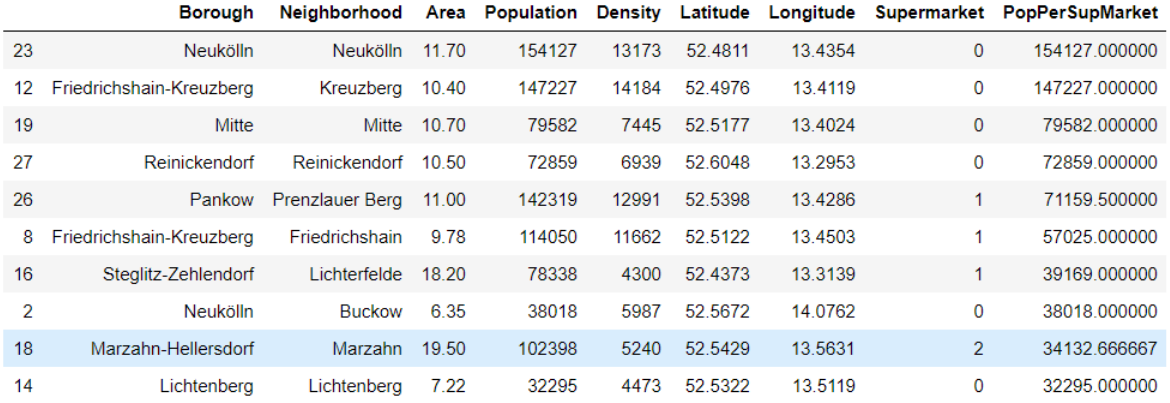
Using Foursquare APIs maximum 100 venues were attempted for each neighborhood within a circle of 1 km radius. A total of 1964 venues were collected where unique categories were 289. Out of all vanues 145 have been identified as supermarket. Then one-hot-encoding has been applied for each category considering all 40 candidate neighborhoods and grouped the dataset in order to add respective number of supermarkets to ***berlin_df***:



	Borough	Neighborhood	Area	Population	Density	Latitude	Longitude	Supermarket
0	Lichtenberg	Alt-Hohenschönhausen	9.33	41780	4478	52.5504	13.5025	3
1	Treptow-Köpenick	Altglienicke	7.89	26101	3308	52.4118	13.5426	2
2	Neukölln	Buckow	6.35	38018	5987	52.5672	14.0762	0
3	Charlottenburg-Wilmersdorf	Charlottenburg	10.60	118704	11198	52.5157	13.3097	5
4	Spandau	Falkenhagener Feld	6.88	34778	5056	52.5524	13.1669	2

- **Step-5: Assessing population per supermarket in each candidate neighborhood**

Since we have population and number of supermarkets for each candidate neighborhood, we can easily calculate population per supermarket (PopPerSupMarket) in each candidate region projecting an additional supermarket for each. Thus the dataframe ***berlin_df*** is sorted according to the ascending order of population-per-supermarket, and further reduced to 30 from 40, where top 10 candidates' looks:



	Borough	Neighborhood	Area	Population	Density	Latitude	Longitude	Supermarket	PopPerSupMarket
23	Neukölln	Neukölln	11.70	154127	13173	52.4811	13.4354	0	154127.000000
12	Friedrichshain-Kreuzberg	Kreuzberg	10.40	147227	14184	52.4976	13.4119	0	147227.000000
19	Mitte	Mitte	10.70	79582	7445	52.5177	13.4024	0	79582.000000
27	Reinickendorf	Reinickendorf	10.50	72859	6939	52.6048	13.2953	0	72859.000000
26	Pankow	Prenzlauer Berg	11.00	142319	12991	52.5398	13.4286	1	71159.500000
8	Friedrichshain-Kreuzberg	Friedrichshain	9.78	114050	11662	52.5122	13.4503	1	57025.000000
16	Steglitz-Zehlendorf	Lichterfelde	18.20	78338	4300	52.4373	13.3139	1	39169.000000
2	Neukölln	Buckow	6.35	38018	5987	52.5672	14.0762	0	38018.000000
18	Marzahn-Hellersdorf	Marzahn	19.50	102398	5240	52.5429	13.5631	2	34132.666667
14	Lichtenberg	Lichtenberg	7.22	32295	4473	52.5322	13.5119	0	32295.000000

From this step we can have a rough idea of top potential candidate-neighborhoods. Top 10 locations shown above out of 30 can roughly be considered for high business potentialities based on higher population, higher density and higher PopPerSupMarket . However, till now we did not use any machine learning techniqe like clusering, so the assessement would not be considered as final one. Rather we will verify and refine our top candidates.

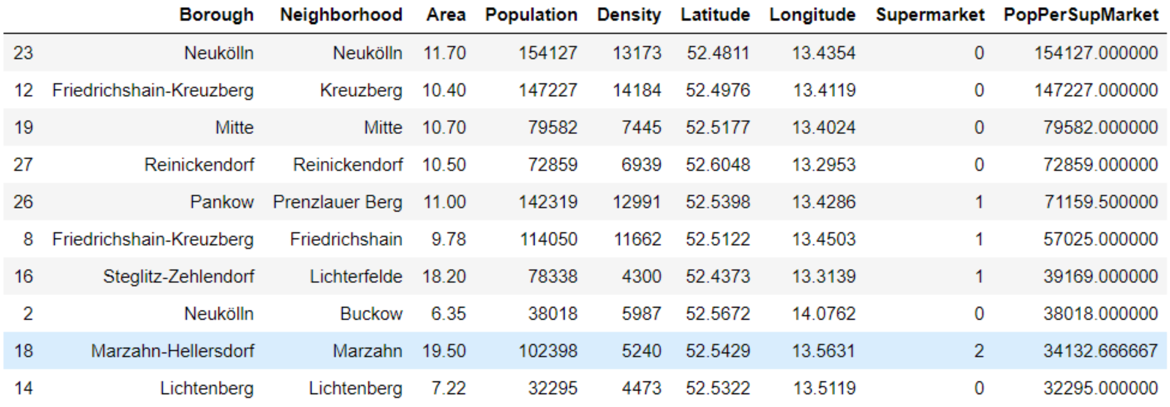
- **Step-6: Clustering with K-means and choosing top 5 candidates**

  K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine learning algorithms and is particularly suited to solve the problem of current project. In this step we will apply K-means clustering algorithm in order to see pontential neighborhoods with similar characteristics. We will see whether K-means considers the primary top-10 locations into similar cluster. We will cluster the neighborhoods into 3 clusters mainly based on their frequency of occurrence for 'Supermarket'. The results will allow us to identify which neighborhoods have higher concentration of supermarket and which neighborhoods have fewer numbers. Based on the occurrence in different neighborhoods, it will help us to answer the question as to which neighborhoods are most suitable to open new branches. Based on top 10 candidates from step-5 and clustering of K-means, we will try to find best set of 5 candidates that are supposed to be more refined and certain.

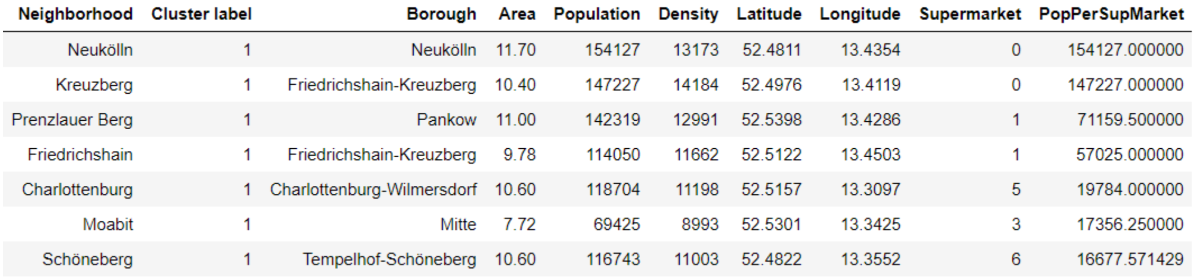
6. Results

After applying K-means with k=3, we see the three categories of high, medium and low or no density cluster of supermarkets. We can also observe that first cluster consists of 22 neighborhoods where top common vanues are supermarket. These 22 locations can be considered highly competitive with abadance of supermarkets. There are 7 neighborhoods in the 2nd cluster and only one in the 3rd. So our candidate neighborhoods hve been reduced from 30 to 8 after excluding 22. On the other hand, we have seen top 10 primarily selected candidated from exploratory analysis from step-5. Now our task would be to choose the best set of five locations considering 10 (from step 5) and 8 (from cluster 2 and 3) neighborhoods.

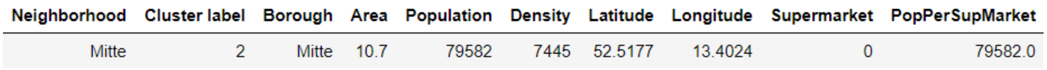
Interestingly, we find that top 5 neighborhoods out of 6 from step-5 (i.e Neukolln, Kreuzberg, Mitte, Prenzlauer Berg and Friedrichshain) are common in the non supermarket-rich clusters. Related dataframes are shown below for better visualization and comparison.



	Borough	Neighborhood	Area	Population	Density	Latitude	Longitude	Supermarket	PopPerSupMarket
23	Neukölln	Neukölln	11.70	154127	13173	52.4811	13.4354	0	154127.000000
12	Friedrichshain-Kreuzberg	Kreuzberg	10.40	147227	14184	52.4976	13.4119	0	147227.000000
19	Mitte	Mitte	10.70	79582	7445	52.5177	13.4024	0	79582.000000
27	Reinickendorf	Reinickendorf	10.50	72859	6939	52.6048	13.2953	0	72859.000000
26	Pankow	Prenzlauer Berg	11.00	142319	12991	52.5398	13.4286	1	71159.500000
8	Friedrichshain-Kreuzberg	Friedrichshain	9.78	114050	11662	52.5122	13.4503	1	57025.000000
16	Steglitz-Zehlendorf	Lichterfelde	18.20	78338	4300	52.4373	13.3139	1	39169.000000
2	Neukölln	Buckow	6.35	38018	5987	52.5672	14.0762	0	38018.000000
18	Marzahn-Hellersdorf	Marzahn	19.50	102398	5240	52.5429	13.5631	2	34132.666667
14	Lichtenberg	Lichtenberg	7.22	32295	4473	52.5322	13.5119	0	32295.000000

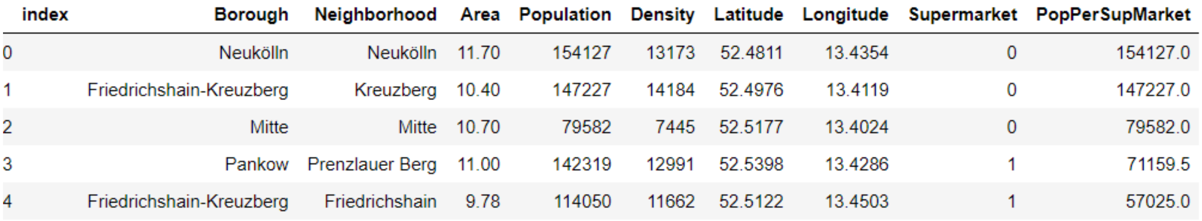


Neighborhood	Cluster label	Borough	Area	Population	Density	Latitude	Longitude	Supermarket	PopPerSupMarket
Neukölln	1	Neukölln	11.70	154127	13173	52.4811	13.4354	0	154127.000000
Kreuzberg	1	Friedrichshain-Kreuzberg	10.40	147227	14184	52.4976	13.4119	0	147227.000000
Prenzlauer Berg	1	Pankow	11.00	142319	12991	52.5398	13.4286	1	71159.500000
Friedrichshain	1	Friedrichshain-Kreuzberg	9.78	114050	11662	52.5122	13.4503	1	57025.000000
Charlottenburg	1	Charlottenburg-Wilmersdorf	10.60	118704	11198	52.5157	13.3097	5	19784.000000
Moabit	1	Mitte	7.72	69425	8993	52.5301	13.3425	3	17356.250000
Schöneberg	1	Tempelhof-Schöneberg	10.60	116743	11003	52.4822	13.3552	6	16677.571429



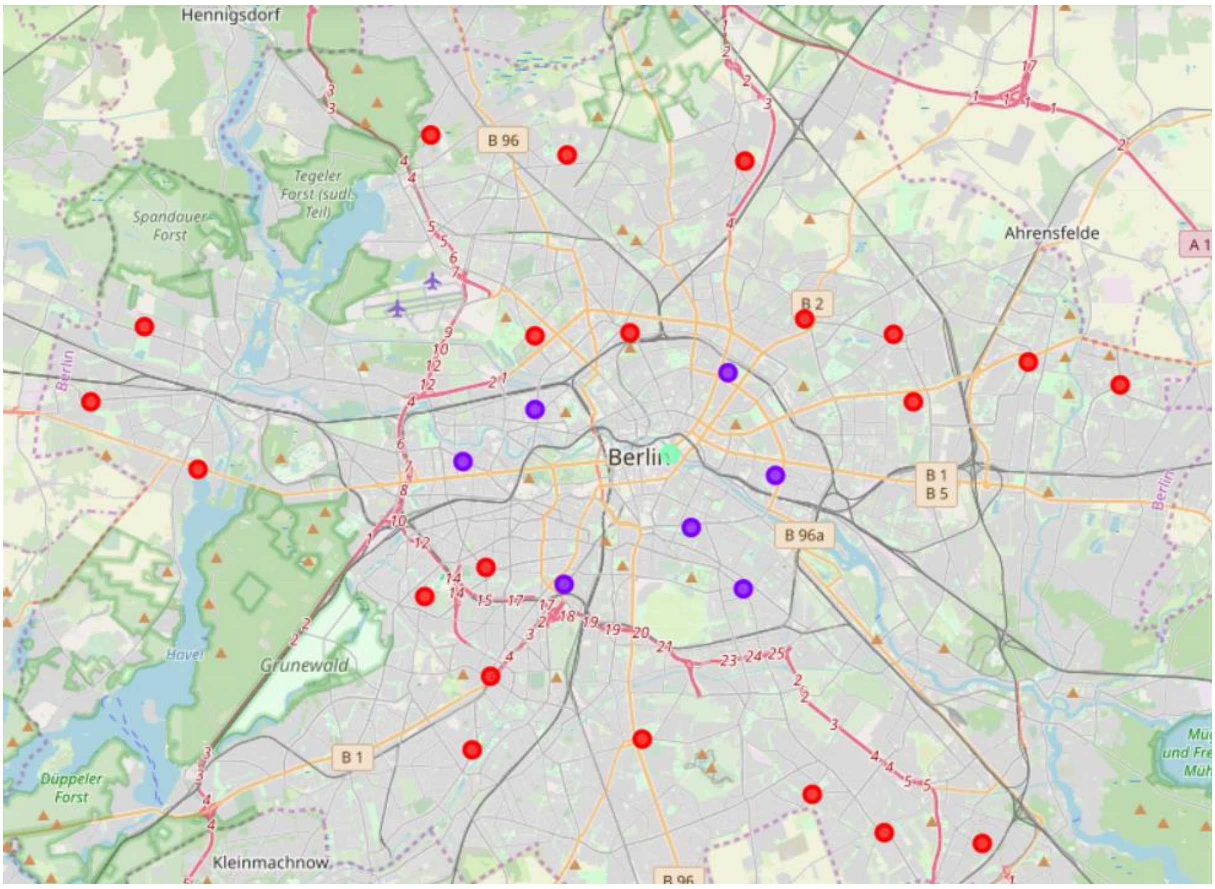
Neighborhood	Cluster label	Borough	Area	Population	Density	Latitude	Longitude	Supermarket	PopPerSupMarket
Mitte	2	Mitte	10.7	79582	7445	52.5177	13.4024	0	79582.0

The K-means just reaffirms the top candidates that were assessed beforehand based on their higher population, higher density and higher PopPerSupMarket. Both analyses give similar top-candidates as summarised below:



index	Borough	Neighborhood	Area	Population	Density	Latitude	Longitude	Supermarket	PopPerSupMarket
0	Neukölln	Neukölln	11.70	154127	13173	52.4811	13.4354	0	154127.0
1	Friedrichshain-Kreuzberg	Kreuzberg	10.40	147227	14184	52.4976	13.4119	0	147227.0
2	Mitte	Mitte	10.70	79582	7445	52.5177	13.4024	0	79582.0
3	Pankow	Prenzlauer Berg	11.00	142319	12991	52.5398	13.4286	1	71159.5
4	Friedrichshain-Kreuzberg	Friedrichshain	9.78	114050	11662	52.5122	13.4503	1	57025.0

Selected best set of five cadidates are located around the vicinity of the city center as shown in the below figure (green and purple).



7. Discussion

If we see further detail we can observe that K-means did not consider Reinickendorf (4th candidate from step-5) as a top potential candidate. Even though population and PopPerSupMarket are higher, and competition is less in Reinickendorf; nevertheless it has the lowest density among top 6 candidate from step-5. More importantly the location of Reickendorf is little far from the city center, for which K-means groups it to the first cluster. Probably the later reason is more logical considering the map.

The segmented cluster 1 has very high number of supermarkets in the neighborhoods. These 22 neighborhoods are likely suffering from intense competition due to oversupply and high concentration of supermarkets, and hence no neighborhood is chosen from this cluster. On the other hand, cluster 2 and 3 represent a great opportunity and high potential areas to open new shops as there is very little to no competition. So based on higher population-per-supermarket, population and density we would like to recommend the top 5 spots as Neukolln, Kreuzberg, Mitte, Prenzlauer Berg and Friedrichshain.

One question might come to the reader of this report that why K-means and exploratory analysis recommend best locations around the city center only? Or how will it be possible that all neighborhoods around the city center are less competitive? This is because the study only considered 30 localities of high importance and excluded 66 neighborhoods (out of 96). Even before applying K-means clustering neighborhoods were excluded based on lower density, lower population and lower

PopPerSupMarket. So the less importance locations around the city center were already disregarded and hence are not visible in the cluster-map.

8. Conclusion

In this project we have applied exploratory data analysis and K-means clustering algorithm to find the best set of five neighborhoods in Berlin. We would like to conclude that the best possible set for interested supermarket-chains in the neighborhood of Berlin is Neukolln, Kreuzberg, Mitte, Prenzlauer Berg and Friedrichshain. Consideing factors like higher demand, lower competition, higher population, higher density and PopPerSupMarket we have reached into this conclusion.

However, it should be mentioned that specific street within a recommended neighborhood has not been suggested as many more data would be necessary for that. Besides, we emphasized on the frequency of occurrence of supermarkets. There are other factors such as income of residents that could influence the location. Future research could devise a methodology with much more data to be used in the clustering algorithm to determine the preferred locations to open a new supermarket-chain. In addition, this project made use of the free Sandbox Tier Account of Foursquare API that came with limitations as to the number of API calls and results returned. Future research could make use of paid account to bypass these limitations and obtain more results.

9. References

[1] https://www.statista.com/statistics/568463/tourism-arrivals-berlin-germany-by-origin/
[2] https://en.wikipedia.org/wiki/Boroughs_and_neighborhoods_of_Berlin
[3] https://developer.foursquare.com