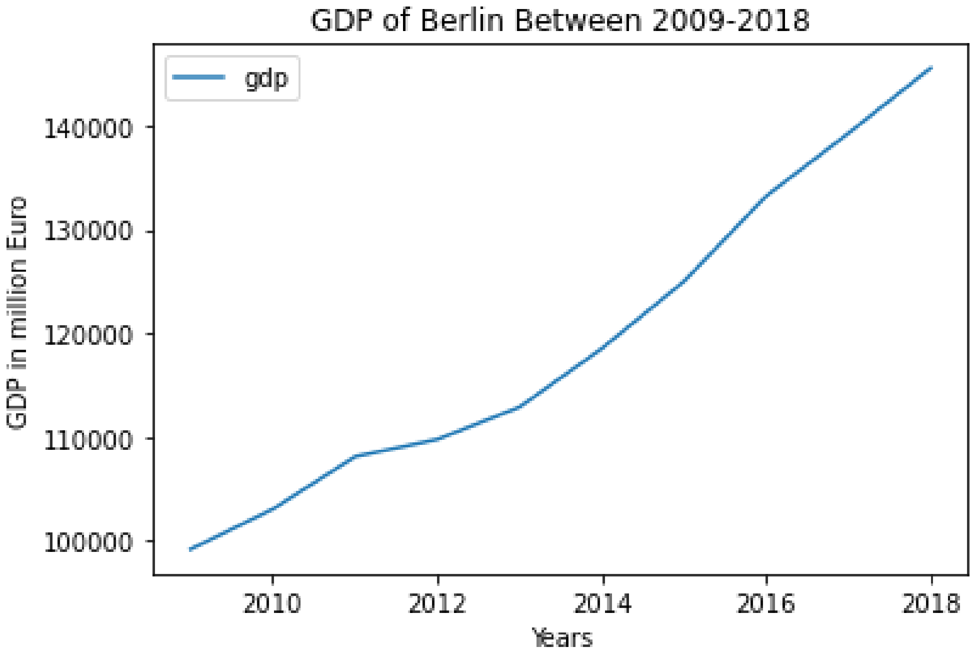
Berlin : A Data Analysis and Visualization Project for Investors

Introduction

Berlin is the capital and largest city of Germany by both area and population. With its approximately 3.5 million inhabitants, there might always be good business and entrepreneurship opportunities for investors. As of 2020, Berlin has €154.6 Billion GDP and 3.3% economic decline due to Covid-19 pandemic throughout the year.[1] Berlin's economy is almost based on services sector, 84% of companies conducting business in services sector. [2]

Even though there is a slight decline in Berlin's GDP in 2020 due to Covid-19 pandemic, after widespread vaccinations started, the restrictions and measures are expected to be gradually lifted and the economy will be booming again for Berlin. That is why I imagined that I am working for an investing company and they asked me to do feasibility work and business analysis for possible services sector investments for future. As we can see from Figure 1, there is a steady growth in Berlin's economy since last 20 years which makes Berlin a good candidate for new investment prospects. However, there are thousands of business models, companies, small scale shops, cafes, restaurants. Which business is the best type and where should it be located in the city ? I will try to answer these two key questions with various data science methods that I have learned on the IBM Data Science Professional Certificate.



[3]

[1] "Amt für Statistik Berlin Brandenburg – Statistiken". *Amt für Statistik Berlin-Brandenburg*

[2] https://en.wikipedia.org/wiki/Berlin#Economy

[3] https://www.statistikportal.de/de/vgrdl/ergebnisse-laenderebene/bruttoinlandsprodukt-bruttowertschoepfung

Data Description

Given the concept above, here is the list of data that I used :

- Venues data with 500 meters radius extracted from Foursquare API,
- GeoJSON file for geographical boundries' coordinates of 96 neighborhoods of Berlin from Statistische Ämter des Bundes und der Länder
- I created a dataset from Wikipedia pages 1 and 2 for names of Berlin's boroughs and neighborhoods and from a Github user, I used and adapted a function to obtain geographical center coordinates of these neighborhoods.
- I scraped this Wikipedia page (https://de.wikipedia.org/wiki/Verwaltungsgliederung_Berlins) to retrieve population and density data of neighborhoods and boroughs.

I will use the Foursquare API to fetch the most common venues in every neighborhood with 500 meters radius. With these venues I am going to use k-means clustering algorithm to cluster the neighborhoods to determine venue categories that distinguish each other.

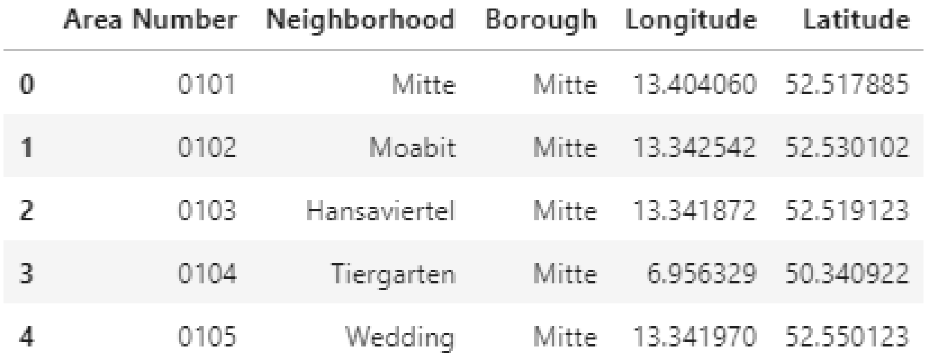
I will use the GeoJSON file with Python package Folium to visualize each neighborhood's density and try to decide where could be the most promising area to invest money.

Lastly, I will try to create my own dataset with for neighborhoods' center coordinates.

Methodology

Data Acquisition, Cleaning and Creating

First of all, I used Jupyter Notebook as my coding environment. I started creating my own data by loading and reading a Geojson file that I found on Amt für Statistik Berlin-Brandenburg. I cleaned that data from unnecessary column and renamed all the column due to dataset was in German. Then I defined a function to obtain geographical center coordinates of every neighborhood in Berlin. After getting neighborhood, borough and coordinates data, my data frame was looking like this :



	Area Number	Neighborhood	Borough	Longitude	Latitude
0	0101	Mitte	Mitte	13.404060	52.517885
1	0102	Moabit	Mitte	13.342542	52.530102
2	0103	Hansaviertel	Mitte	13.341872	52.519123
3	0104	Tiergarten	Mitte	6.956329	50.340922
4	0105	Wedding	Mitte	13.341970	52.550123

Figure 1:Berlin Neighborhood DataFrame

Then I used Folium to show each neighborhood superimposed on top of Berlin map :

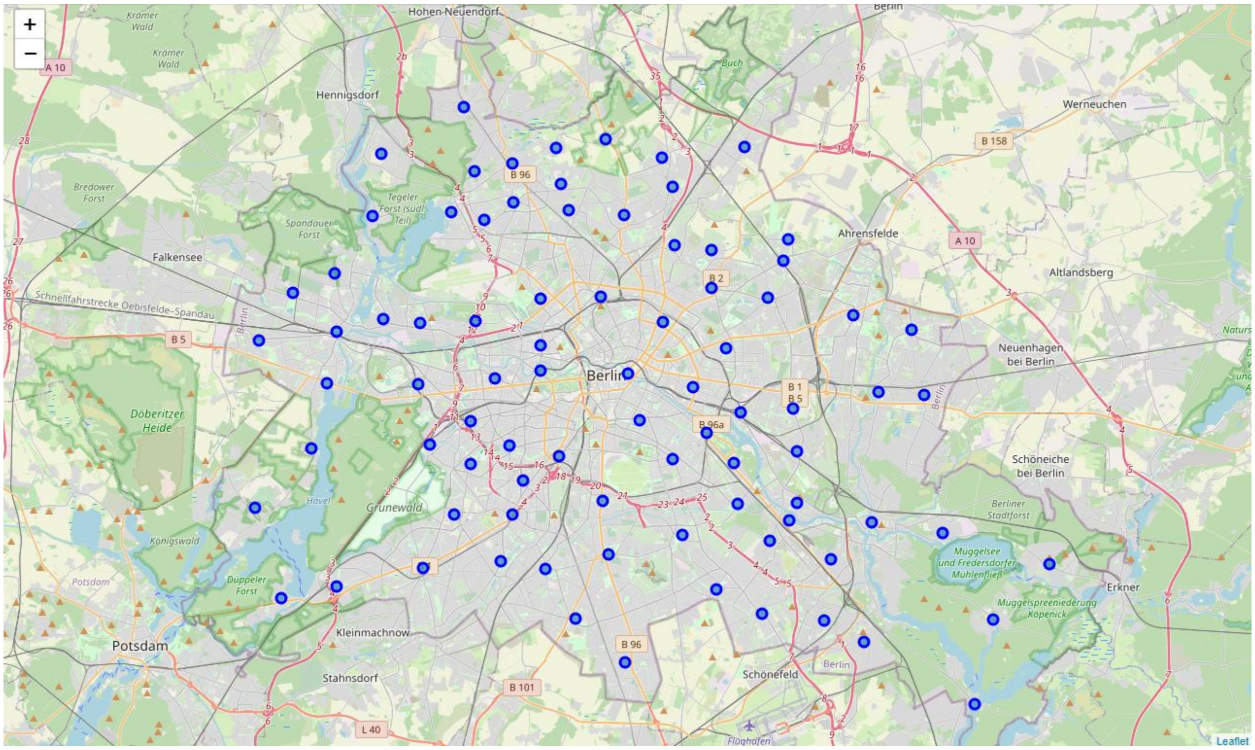
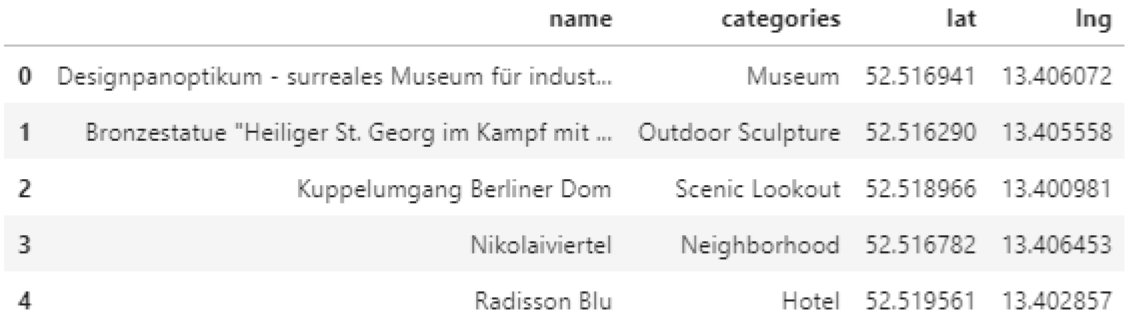


Figure 2:Berlin neighborhoods superimposed on map

I used Foursquare API to get the Berlin's neighborhoods' venues, categories and their latitude and longitude data inside the radius of 750 meters. Here is how the dataframe looks like :



	name	categories	lat	lng
0	Designpanoptikum - surreales Museum für indust...	Museum	52.516941	13.406072
1	Bronzestatue "Heiliger St. Georg im Kampf mit ...	Outdoor Sculpture	52.516290	13.405558
2	Kuppelumgang Berliner Dom	Scenic Lookout	52.518966	13.400981
3	Nikolaiviertel	Neighborhood	52.516782	13.406453
4	Radisson Blu	Hotel	52.519561	13.402857

Figure 3:Berlin Venue Categories

Foursquare API returned 2386 venues in all 96 neighborhoods and 12 boroughs of Berlin. There were 302 unique categories of venues.

Using this data frame, I wanted to visualize top 10 venues in terms of numbers in Berlin with a pie chart to see which ones are most popular across Berlin :

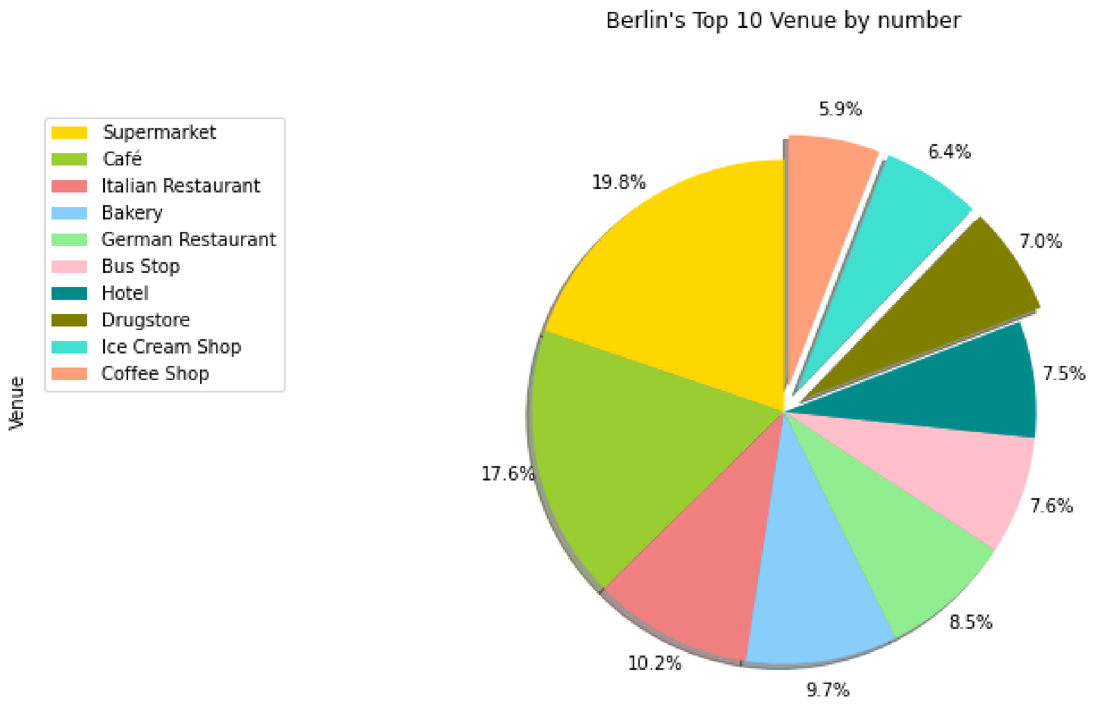
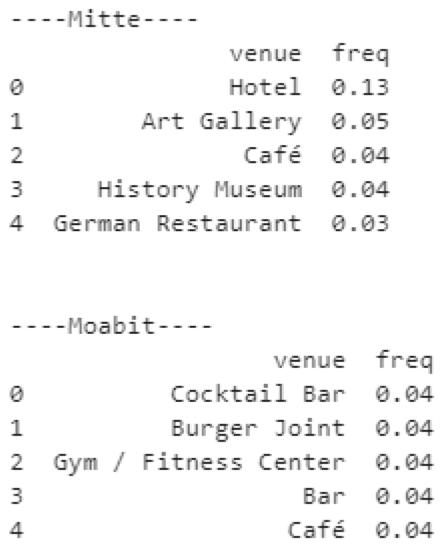


Figure 4: Berlin's most popular 10 venues

Analyzing Each Neighborhood

I assigned every venue in Berlin 1s and 0s so I can calculate which venues is the most common in each neighborhood. Then I took the mean of these values to see how frequent the top 5 venues are in each neighborhood :

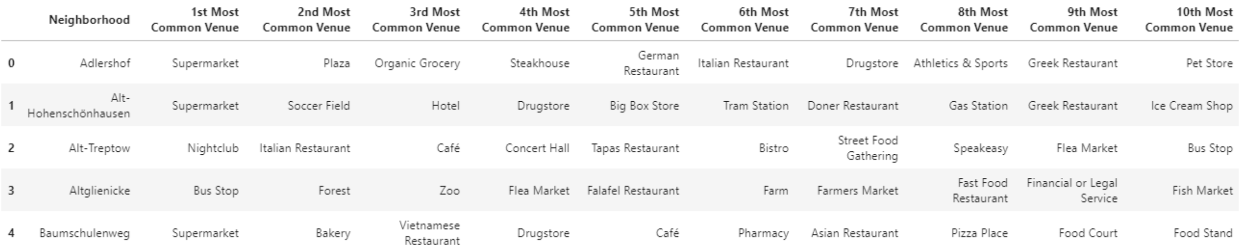


```
----Mitte----
               venue  freq
0              Hotel  0.13
1        Art Gallery  0.05
2               Café  0.04
3     History Museum  0.04
4  German Restaurant  0.03


----Moabit----
                  venue  freq
0          Cocktail Bar  0.04
1          Burger Joint  0.04
2  Gym / Fitness Center  0.04
3                   Bar  0.04
4                  Café  0.04
```

Figure 5:Top 5 venues' frequency of each neighborhood

Then I defined a function to create a data frame to see the most common venues of Berlin. Here is what I have :



	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Adlershof	Supermarket	Plaza	Organic Grocery	Steakhouse	German Restaurant	Italian Restaurant	Drugstore	Athletics & Sports	Greek Restaurant	Pet Store
1	Alt-Hohenschönhausen	Supermarket	Soccer Field	Hotel	Drugstore	Big Box Store	Tram Station	Doner Restaurant	Gas Station	Greek Restaurant	Ice Cream Shop
2	Alt-Treptow	Nightclub	Italian Restaurant	Café	Concert Hall	Tapas Restaurant	Bistro	Street Food Gathering	Speakeasy	Flea Market	Bus Stop
3	Altglienicke	Bus Stop	Forest	Zoo	Flea Market	Falafel Restaurant	Farm	Farmers Market	Fast Food Restaurant	Financial or Legal Service	Fish Market
4	Baumschulenweg	Supermarket	Bakery	Vietnamese Restaurant	Drugstore	Café	Pharmacy	Asian Restaurant	Pizza Place	Food Court	Food Stand

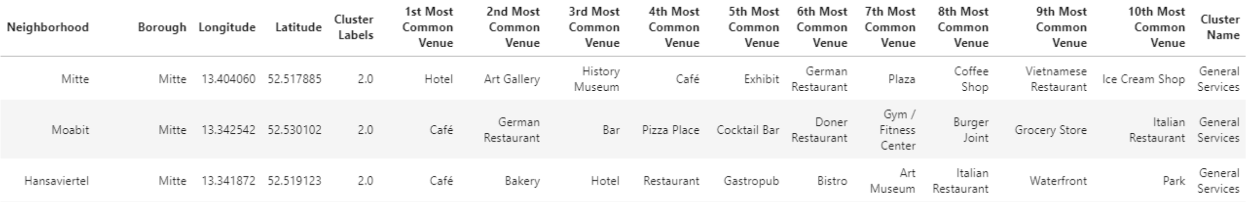
Figure 6:Berlin's most common venues in each neighborhood

Clustering Neighborhoods

I used k-means clustering machine learning algorithm to segregate neighborhood. I determined k value as 5 after multiple times of running the algorithm. It returned the most meaningful results for my dataset. According to results I had three main clusters and two other less common clusters. Namely :

- Cluster 1 : Bus Stop is the most common venue, we can categorize that as general public services
- Cluster 2 : Retail store cluster such as supermarkets, stores
- Cluster 3 : Food and Beverage Outlets such as cafes, restaurants, bakeries
- Cluster 4 : Historic sites
- Cluster 5 : Zoo, event spaces

I named each cluster label and added label name column to the data frame :



Neighborhood	Borough	Longitude	Latitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue	Cluster Name
Mitte	Mitte	13.404060	52.517885	2.0	Hotel	Art Gallery	History Museum	Café	Exhibit	German Restaurant	Plaza	Coffee Shop	Vietnamese Restaurant	Ice Cream Shop	General Services
Moabit	Mitte	13.342542	52.530102	2.0	Café	German Restaurant	Bar	Pizza Place	Cocktail Bar	Doner Restaurant	Gym / Fitness Center	Burger Joint	Grocery Store	Italian Restaurant	General Services
Hansaviertel	Mitte	13.341872	52.519123	2.0	Café	Bakery	Hotel	Restaurant	Gastropub	Bistro	Art Museum	Italian Restaurant	Waterfront	Park	General Services

This is how map is looking like after I integrate clusters to the map :

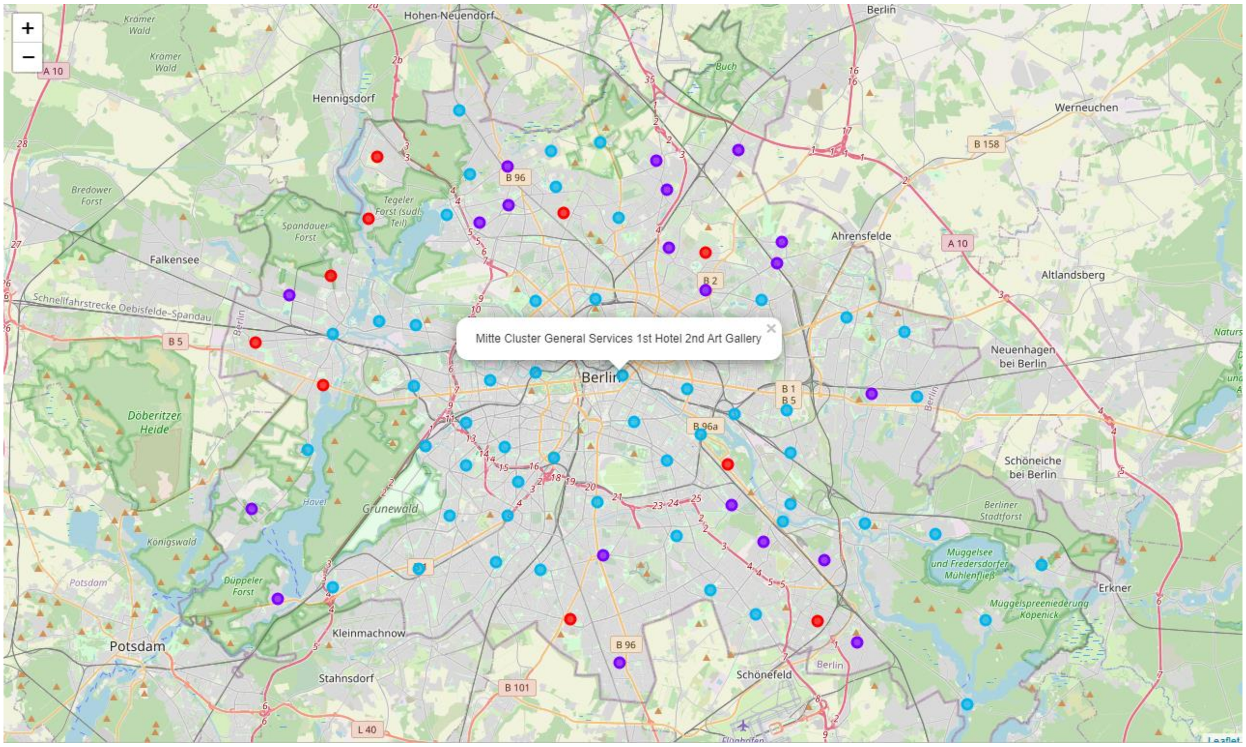


Figure 7: k-means Clustered map of Berlin

Afterwards I scraped, cleaned and read into data frame the neighborhoods’ density table from the Wikipedia page of Verwaltungsgliederung Berlins. There are useful information such as density, population and area about Berlin’s neighborhoods. I used this table to create a choropleth map of Berlin based on neighborhoods’ density. Finally, I utilized Amt für Statistik Berlin-Brandenburg’s Lifeworld oriented spaces (LOR) dataset to determine and visualize geographical boundaries of neighborhoods. After all done, I merged clustered map of Berlin and choropleth map to see and evaluate my results.

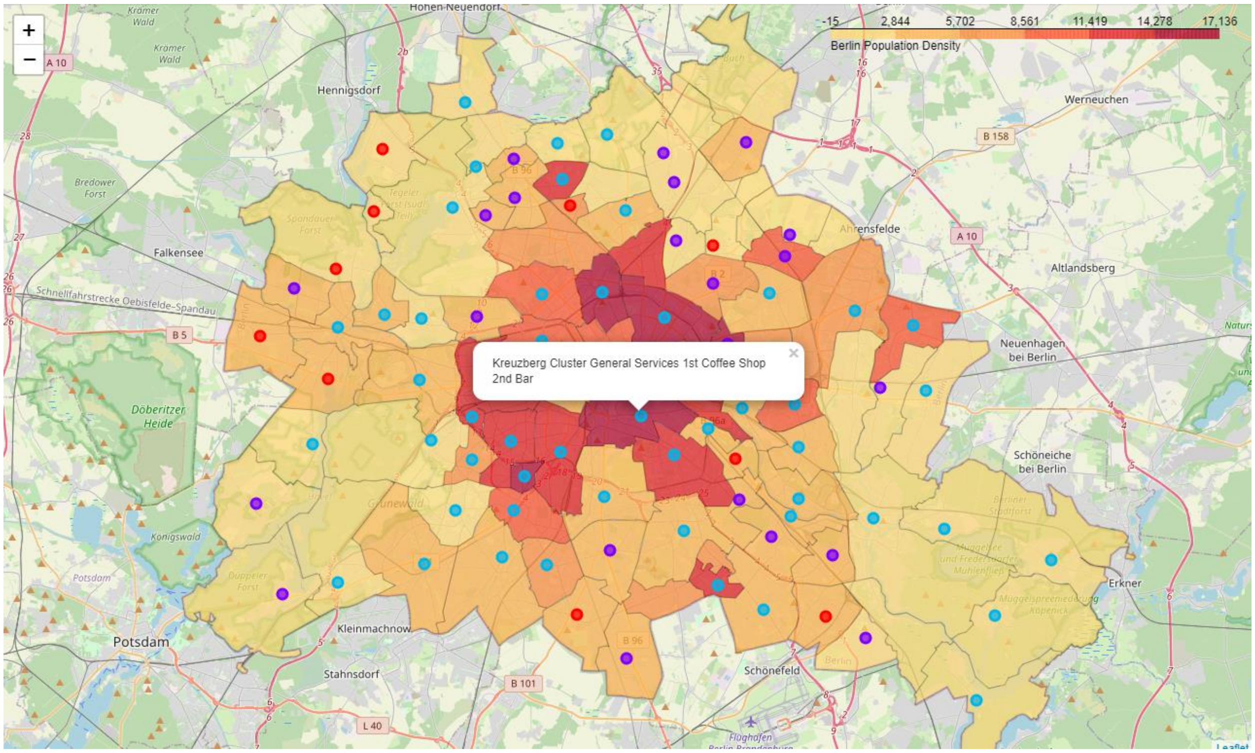


Figure 8: Berlin's cluster and density based choropleth map

Results

The main conclusion that we reach is that Berlin has three main venue clusters and these three different categories of clusters are segregated slightly in three different areas of the city. We may confidently say the first and by far important cluster is cluster 3 due to it settled evidently at the heart of the city which is the most populated are of Berlin. The cluster contains venues such as cafes, restaurants, hotels, bakeries. That is why I called it "General services". In my opinion, it would be fair to consider these highly tourist attraction neighborhoods are the most promising areas in terms of investment feasibility. Due to its highly dense population, it would be safe to say that these areas might be profitable when we consider Berlin's steadily growing gross domestic product.

Second cluster is cluster 2. I named this cluster "Retail stores" because it consist of supermarkets, pharmacies etc. This cluster is sort of surrounding the city center and is more active in less dense population neighborhoods. It should be remembered that as shown in figure 4, supermarkets are the most numbered venues of Berlin which points out that it might be still profitable to invest in retail sector anyway. It should be evaluated further with other data science tools to be more precise on this type of investment.

The last main cluster is cluster 1 and this mainly consists of bus stops which is a public service that why I labeled it as such. Due to public services are not scope of this research I left it out for commenting.

The other two clusters have only one venue for each and they should be excluded from analysis to due lack of information.

Discussions

Since the world is still fighting with covid-19 pandemic, it should be stated that it might still be too risky to invest money anything physical such as cafes or stores. Because these types of investments are not very easy to liquidate. A pre-market analysis should carefully be made and check there are enough demand for such an investment.

Other than this, we should accept the fact that this analysis is very narrow and should be supported other deeper and detailed data as well such as migration, age and gender qualifications of population etc.

Conclusion

In this study, I tried to analyze the types of venues of Berlin and how they are clustering on map to see the possible new investment opportunities. I used k-means clustering algorithm and found 5 different clusters from the total of 2385 venues in 96 neighborhoods. There are 236 unique categories of venues so we can say there are plenty of diversity in Berlin's venue types. I tried to associate population density with venue clusters in these neighborhoods in my study.

References

1. Foursquare API
2. Statistische Ämter des Bundes und der Länder
3. https://juanitorduz.github.io/germany_plots/
4. https://de.wikipedia.org/wiki/Verwaltungsgliederung_Berlins
5. https://daten.odis-berlin.de/de/dataset/lor_prognoseraeume/
6. Python libraries : Beautiful Soup, Numpy, Pandas, Json, Geopy, Geopandas, Requests, Matplotlib, Sklearn, Folium