1. Introduction

Berlin, the capital of Germany and one of the most important cities in Europe, it is attractive for tourists and investors for many reasons. Among them belong its vibrant culture, fabulous food, intense night-life and centuries-old history. Although the city is pretty big, there are some key areas which are of most interest for business owners and investors.

The classification and segmentation of Berlin's central neighborhoods with respect to popular venues, as well as the analysis of population demographics and tourist attractiveness is aimed to support finding those areas and thus the decision making of following stakeholders:

- Any contractor or entrepreneur who is aiming to open new business (e.g. hotel, restaurant, cafe), and while looking for the right location has to take into consideration factors such as target group or possible surrounding businesses and competitors. Tourist visit ratio and most common venues of given area are important parameters and are subject of the analysis presented in next chapters.
- Investors in commercial real estate. Investing in the right property includes understanding of local demographics, having overview of the type of successful businesses in the area or it attractiveness for tourism.
- City planners when developing the design and monitoring the development of public spaces, parks, transport, community infrastructure in each neighborhood.

2. Data acquisition and cleaning

2.1. Data sources

To address and analyze the problem of interest presented above, data was drawn from three separate data sources:

- Official Berlin Census Data for up-to-date demographic and touristic data for the city of Berlin The information is part of the Creative Commons Namensnennung 3.0 Deutschland Lizenz 2019.

  The official population statistics of Berlin consist of various statistics, each of which has its own function. Our main focus is the number of residents in Berlin divided by postal code (PLZ), district (Bezirk) and age groups. This information is available in xslx-format under following link: Link to file

  To illustrate domestic tourism, accommodation facilities with ten or more guest beds and all campsites for holiday camping with ten or more parking spaces are surveyed monthly by the official authorities in Germany. The results are then published and publicly available on the internet. The tourist data of interest for this project is the number of guest arrivals and overnight stays for each district (Bezirk) of Berlin for the

period January 2019 till September 2019. It can be downloaded in XLSX format rom the following link: Link to file

- OpenDataSoft to get the geo coordinates (longitude and latitude) of each postal code area. This data is available under the Open Database License. For more information and credits visit "© OpenStreetMap contributors"

  I downloaded from the following link the CSV file containing postal codes of all neighborhoods within Berlin with the corresponding gps coordinates which are required for executing requests and retrieve information from Foursquare API:
  Link to File

- Project of Ideation & Prototyping Labs to get the geo coordinates of the borders of each postal code area, as well of each district (Bezrik) in Berlin. This data is available under the Lizenz CC-BY-3.0 (https://creativecommons.org/licenses/by/3.0/de/). For more information and credits visit the author of this data: Amt für Statistik Berlin-Brandenburg
  I downloaded from the following link the geojson file containing the information required to create maps with borders of each postal code area of Berlin:
  Link to File
  I downloaded from the following link the geojson file containing the information required to create maps with borders of each district area of Berlin:
  Link to File

- Foursqaure Developer to get the most popular venues for each postal code area of interest using Foursquare technology and its user-generated content data.

2.2. Data Cleaning and preparation

Data downloaded from Official Berlin Census Data and from OpenDataSoft was saved locally and the excel tables were manually changed so that desired data format has been reached, namely three csv files with following information:

- Berlin_data.csv: Berlin's population and age (for each postal area)
- Berlin_tourist_2019.csv: tourist numbers for each district of Berlin
- berlin_geocoordinates.csv: longitude and latitude data for each postal code area of Berlin

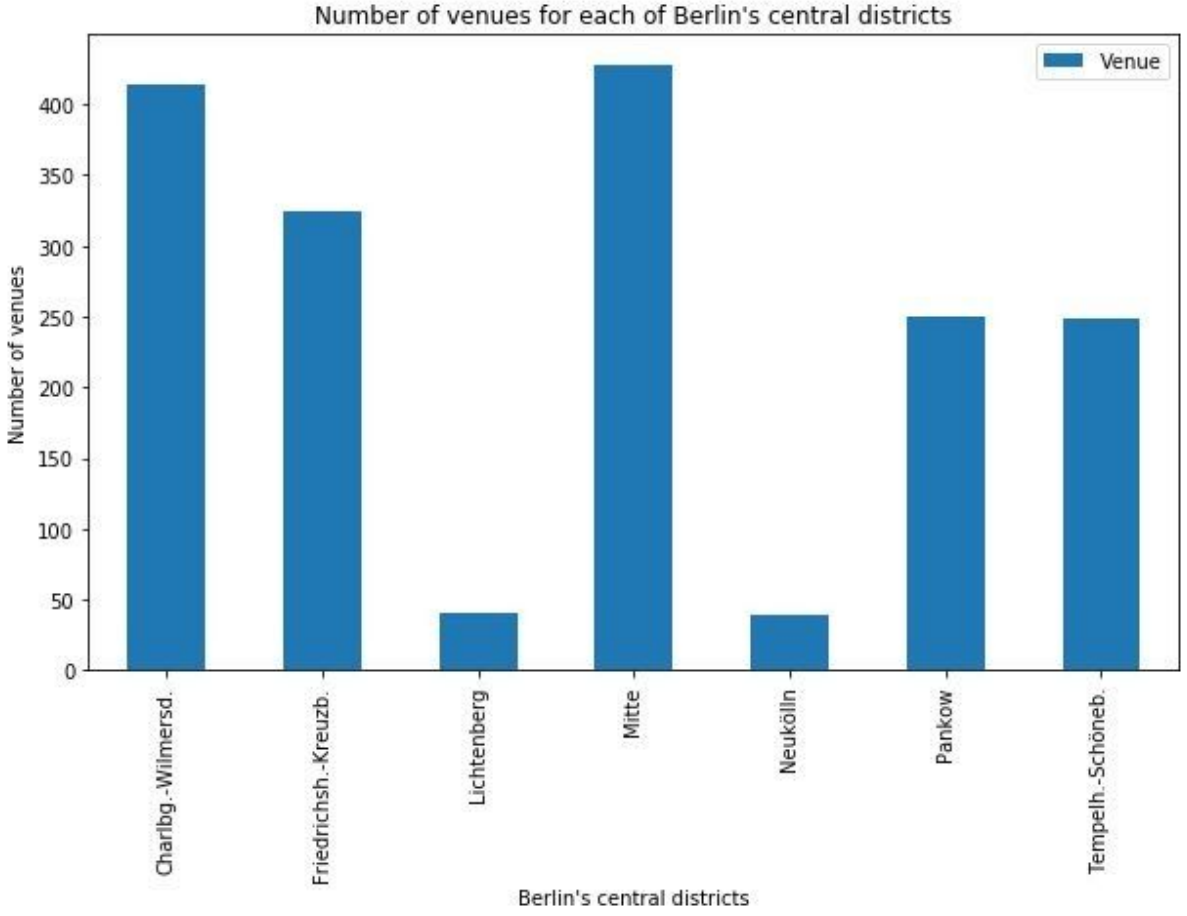
After loading those tables into panda dataframes, additional data cleaning and wrangling of missing, incorrect or not required data was performed.

It is important to clarify that Berlin consists of 12 districts (Bezirke) and each of them is further divided into smaller areas called Stadtteile. Additionally the city can be divided by postal code areas where one postal code area can belong to one or more than one Stadtteile but to only one Bezirk. Since Berlin is covering big area and Foursqaure Developer Api provides information mostly for the central part of Berlin, I have decided to analyse only the central area of Berlin with postal codes starting with 10xxx. Analysis of areas for which we do not have enough input would not have delivered plausible results.

3. Methodology

First step after loading, preparing and cleaning the population data, the tourist data and the geo coordinates data, was to start exploring Berlin's central neighborhoods by retrieving venues for each postal code area using Foursquare API. I have decided to set as search criteria 100 as limit for number of venues returned and 300 m as search radius taking into consideration the approximate average size of each postal code area. Please note that the search for venues is within postal code area and not within districts. Main reason for this is that districts are very vast and different in size, and therefore would have needed different values for different districts when setting the search radius.

The end result of the search was 1746 venues which I grouped according to the districts they belong to and the results is shown on the figure below. We can clearly see that three districts have the majority of venues, while Lichtenburg and Neukölln have the fewest.

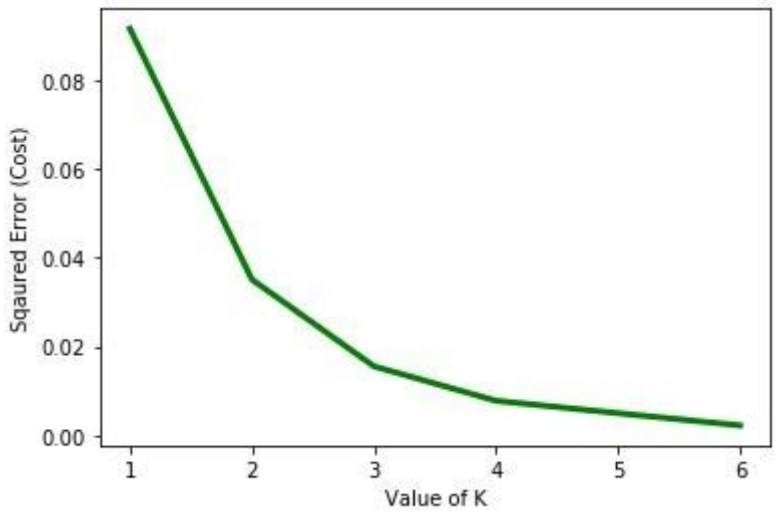


Afterwards I estimated that there are 231 unique categories which were returned by Foursquare API, and I created an overview of the top 5 venues for each district area, the result is shown in the table below.

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Charlbg.-Wilmersd.	Hotel	Italian Restaurant	Café	Supermarket	German Restaurant
1	Friedrichsh.-Kreuzb.	Bar	Café	Italian Restaurant	Vietnamese Restaurant	Hotel
2	Lichtenberg	Supermarket	Drugstore	Plaza	Gym / Fitness Center	Italian Restaurant
3	Mitte	Hotel	Café	Italian Restaurant	German Restaurant	Coffee Shop
4	Neukölln	Café	Italian Restaurant	Asian Restaurant	Vietnamese Restaurant	Chinese Restaurant
5	Pankow	Café	Bakery	Bar	Italian Restaurant	Hotel
6	Tempelh.-Schöneb.	Café	Hotel	Italian Restaurant	Supermarket	Zoo Exhibit

As you can see many of the categories are repeated for different districts. In order to be able to segment Berlin's neighborhoods, K-Means clustering algorithm was applied using python's sklearn library.

K-Means clustering algorithm is an unsupervised algorithm and it is used to segment (or partition) given data into K-clusters or parts based on the K-centroids. Choosing the right value for K is important for the correct execution of this algorithm. One of the most common and reliable methods is the so called Elbow Method. The main idea is to calculate the Within-Cluster Sum of Squared Error for different values of k, and choose the k for which the Sum of Squared Error starts to diminish. The figure below shows the result of plotting the Sum of Squared Error versus different K Values and for K - value equals to 4 to error starts to diminish. That is why I used K = 4 for the K-Means clustering algorithm.



Another library which I used for exploring Berlin's neighborhoods is folium library for interactive map visualizations. I created choropleth maps in which areas of Berlin were shaded in different colors depending on population density or on tourists numbers, and combined this information with the clustering results from the K-Clustering algorithm.

4. Results

The tables below show the results after applying the clustering algorithm. Analyzing the five most common venues for each cluster following conclusions can be made:

	Neighborhood	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
73	Neukölln	0	Café	Italian Restaurant	Asian Restaurant	Vietnamese Restaurant	Chinese Restaurant
76	Neukölln	0	Café	Italian Restaurant	Asian Restaurant	Vietnamese Restaurant	Chinese Restaurant

Cluster 1: most common venues are cafes, followed by Italian restaurants and Asian restaurants. This segment is attractive for food-tourists and food businesses.

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venu
3	Pankow	Café	Bakery	Bar	Italian Restaurant	Hote
4	Friedrichsh.-Kreuzb.	Bar	Café	Italian Restaurant	German Restaurant	Vietnamese Restauran
6	Friedrichsh.-Kreuzb.	Bar	Café	Italian Restaurant	German Restaurant	Vietnamese Restauran
8	Friedrichsh.-Kreuzb.	Bar	Café	Italian Restaurant	German Restaurant	Vietnamese Restauran
9	Friedrichsh.-Kreuzb.	Bar	Café	Italian Restaurant	German Restaurant	Vietnamese Restauran
10	Friedrichsh.-Kreuzb.	Bar	Café	Italian Restaurant	German Restaurant	Vietnamese Restauran
11	Pankow	Café	Bakery	Bar	Italian Restaurant	Hote
12	Friedrichsh.-Kreuzb.	Bar	Café	Italian Restaurant	German Restaurant	Vietnamese Restauran
13	Pankow	Café	Bakery	Bar	Italian Restaurant	Hote
21	Pankow	Café	Bakery	Bar	Italian Restaurant	Hote
22	Pankow	Café	Bakery	Bar	Italian Restaurant	Hote
23	Pankow	Café	Bakery	Bar	Italian Restaurant	Hote
25	Pankow	Café	Bakery	Bar	Italian Restaurant	Hote
26	Pankow	Café	Bakery	Bar	Italian Restaurant	Hote
27	Pankow	Café	Bakery	Bar	Italian Restaurant	Hote
56	Friedrichsh.-Kreuzb.	Bar	Café	Italian Restaurant	German Restaurant	Vietnamese Restauran

Cluster 2: most common venues are cafes and bars, followed by bakeries, restaurants, hotels. This segment of Berlin is connected with nightlife and social venues.

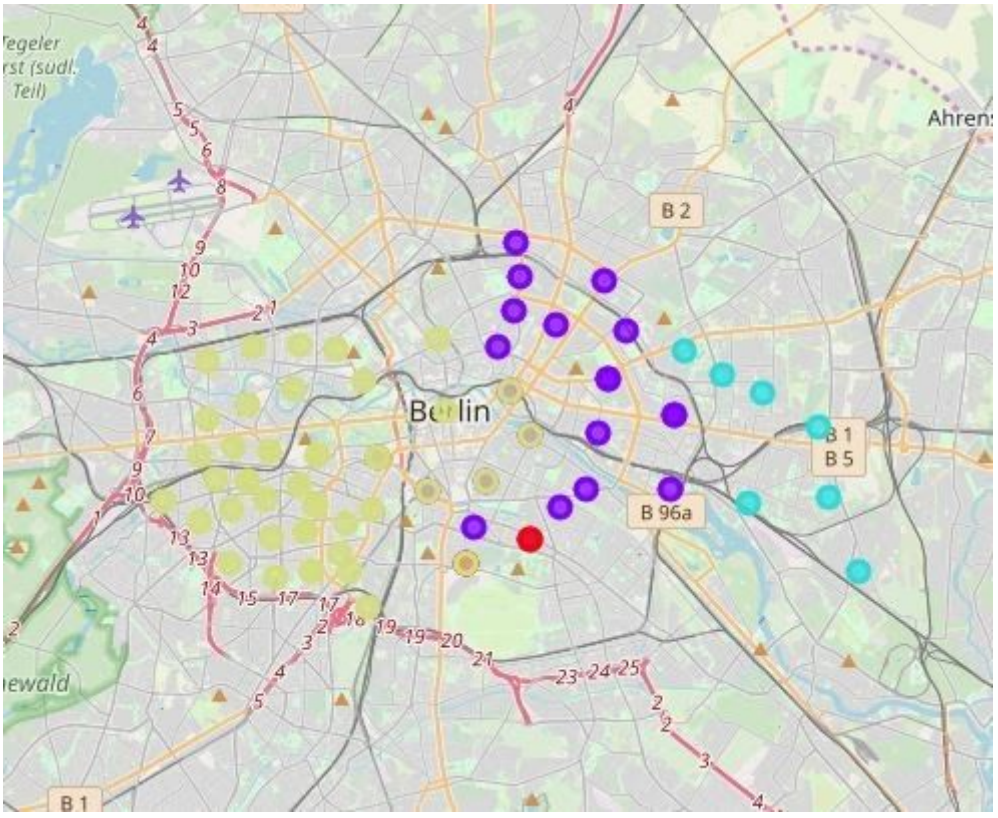
	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
14	Lichtenberg	Supermarket	Drugstore	Tram Station	Italian Restaurant	Soccer Field
15	Lichtenberg	Supermarket	Drugstore	Tram Station	Italian Restaurant	Soccer Field
16	Lichtenberg	Supermarket	Drugstore	Tram Station	Italian Restaurant	Soccer Field
17	Lichtenberg	Supermarket	Drugstore	Tram Station	Italian Restaurant	Soccer Field
18	Lichtenberg	Supermarket	Drugstore	Tram Station	Italian Restaurant	Soccer Field
19	Lichtenberg	Supermarket	Drugstore	Tram Station	Italian Restaurant	Soccer Field
20	Lichtenberg	Supermarket	Drugstore	Tram Station	Italian Restaurant	Soccer Field

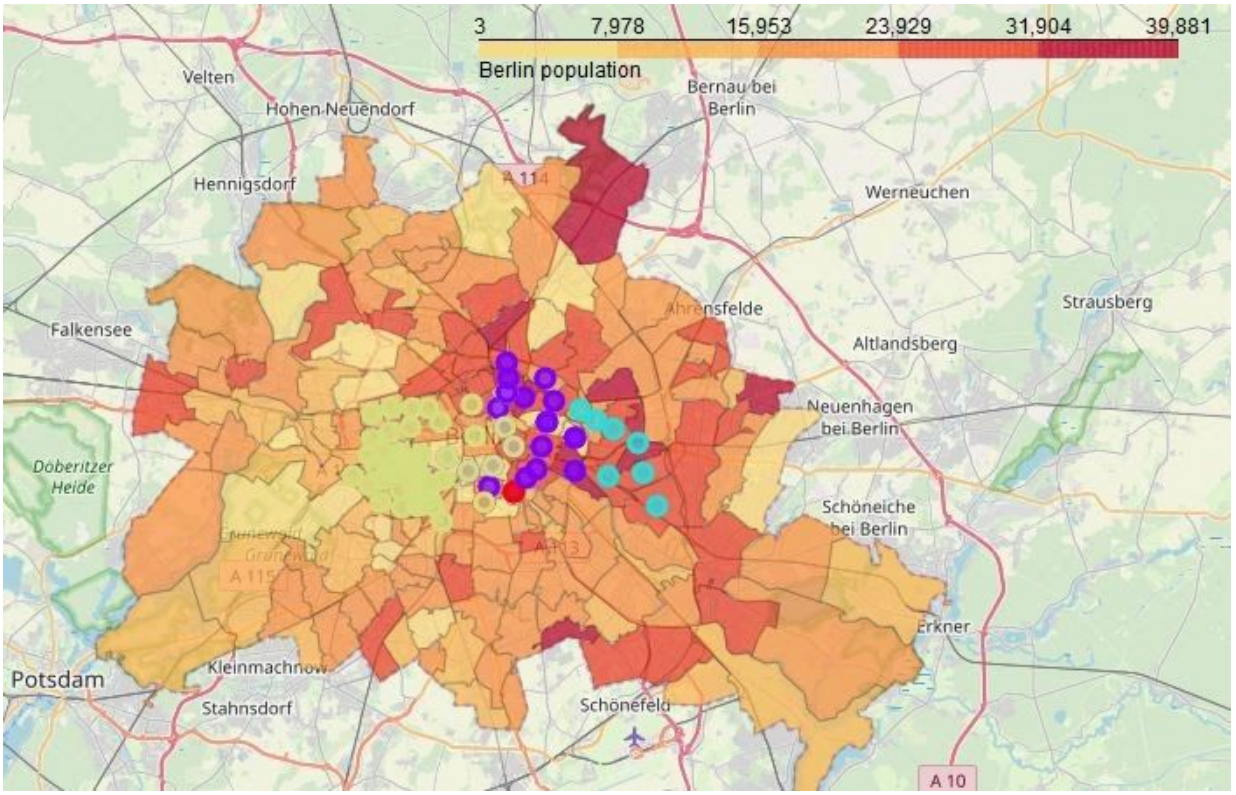
Cluster 3: most common venues are Supermarket, Drugstores, Tram stations. Therefore this segment of Berlin is mostly residential.

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Mitte	Hotel	Café	Italian Restaurant	German Restaurant	Coffee Shop
1	Mitte	Hotel	Café	Italian Restaurant	German Restaurant	Coffee Shop
2	Mitte	Hotel	Café	Italian Restaurant	German Restaurant	Coffee Shop
5	Mitte	Hotel	Café	Italian Restaurant	German Restaurant	Coffee Shop
7	Mitte	Hotel	Café	Italian Restaurant	German Restaurant	Coffee Shop
24	Mitte	Hotel	Café	Italian Restaurant	German Restaurant	Coffee Shop
28	Mitte	Hotel	Café	Italian Restaurant	German Restaurant	Coffee Shop
29	Mitte	Hotel	Café	Italian Restaurant	German Restaurant	Coffee Shop
30	Mitte	Hotel	Café	Italian Restaurant	German Restaurant	Coffee Shop
31	Mitte	Hotel	Café	Italian Restaurant	German Restaurant	Coffee Shop
32	Mitte	Hotel	Café	Italian Restaurant	German Restaurant	Coffee Shop
33	Charlbg.-Wilmersd.	Italian Restaurant	Café	Hotel	German Restaurant	Supermarket
34	Charlbg.-Wilmersd.	Italian Restaurant	Café	Hotel	German Restaurant	Supermarket
35	Charlbg.-Wilmersd.	Italian Restaurant	Café	Hotel	German Restaurant	Supermarket
36	Charlbg.-Wilmersd.	Italian Restaurant	Café	Hotel	German Restaurant	Supermarket
37	Charlbg.-Wilmersd.	Italian Restaurant	Café	Hotel	German Restaurant	Supermarket
38	Mitte	Hotel	Café	Italian Restaurant	German Restaurant	Coffee Shop

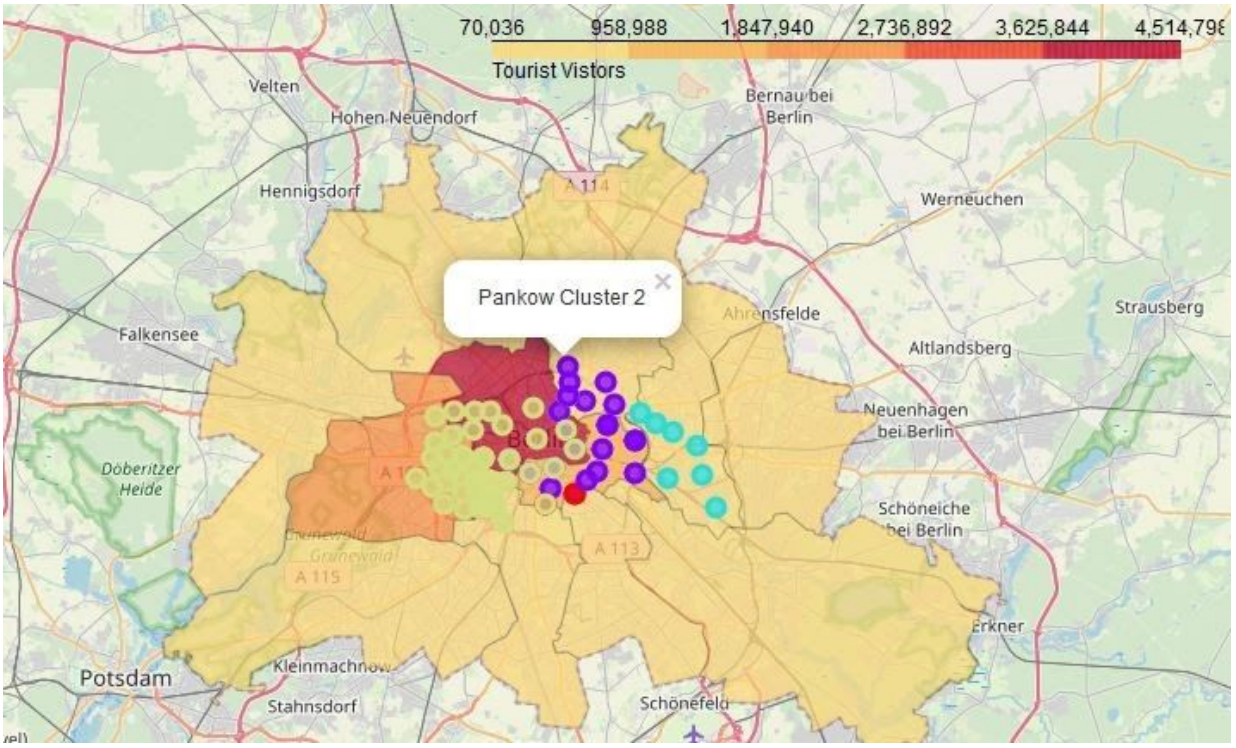
Cluster 4: most common venues are hotels, followed by cafes, restaurants, coffee shops. This part of Berlin is obviously the most attractive one for tourists.

Additionally, the segmentation is visualised on the cluster map of Berlin below. Cluster 1 (food/restaurants) is red, cluster 2 (night and social life) is purple, cluster 3 (hotels, touristic and social venues) is yellow and cluster 4 (residential area) is blue.





The Choropleth map above displays the Berlin's population density in different postal code areas, combined with the clustering labels from K-Means clustering algortihm. We can cleary see that Cluster 2 (dark blue colour: Night life and social venues) and Cluster 3 (light blue colour: Residential area) are located in areas with higher population density, while cluster 4 (yellow colour: venues that are most attractive for tourists) is with much less population density.



Last part of the analysis is to visualize the number of tourists, staying in Berlin's districts, combined with the cluster labels. The result is shown on the map above: cluster 4 (yellow colour) is covering the area with biggest tourist density while cluster 2 (dark blue) and especially cluster 3 (light blue: residential area) are covering area with much less tourist visitors.

5. Discussion

Berlin is attractive and important city, growing and changing continuously. Understanding the direction of development of each part of the city is important but not that easy task. With the help of the segmentation algorithm and the visualisation of the results, I have shown that Central West Berlin is the most attractive part for tourists, while Central East Berlin is attractive for nightlife, or that the outskirts of East Berlin are mostly residential.

The project relied on user-generated data of Foursquare website, which is not a popular app in Germany and therefore the venues data might not be up-to-date or complete. Another limitation of the Foursquare data is that the highest number of venues are in the Food and Shop categories., it also does not take into account venue's size (city park and bakery are both venues with the same weight). Using user-generated data for venues from Google, which is way more popular, might improve the quality of the analysis in future studies. Due to the limitations of Foursquare I have only analysed the central parts of Berlin. Future works should expand the scope of the analysed area.

6. Conclusion

In this project, I analyzed Berlin's central neighborhoods and using K-Means clustering algorithm and Python's folium library for visualisation was able to segment Berlin's central area into 4 parts and visualise the results. The clustering model and the visualised results can be used by business owners or investors when looking for the right location when planning to open new restaurants or invest in new properties.