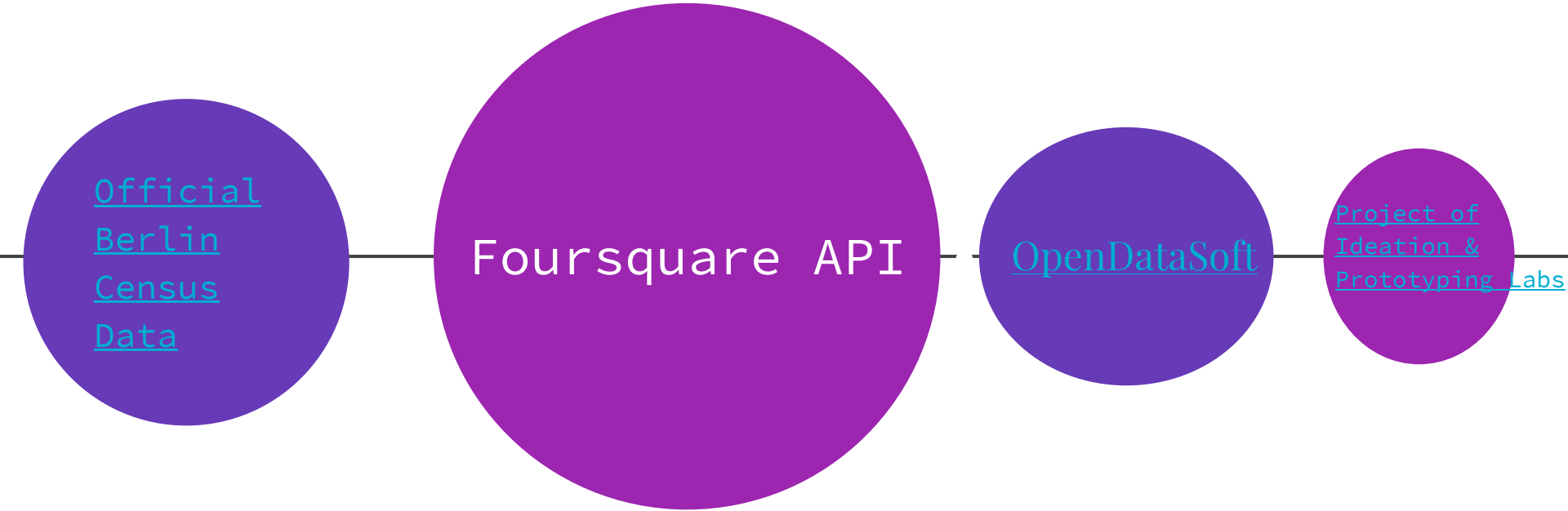
Capstone Project

Segmentation and Clustering of neighborhoods in Berlin

Segmentation of Berlin's neighborhoods is valuable for:

- 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
- 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

Various data sources have been used for data acquisation



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:

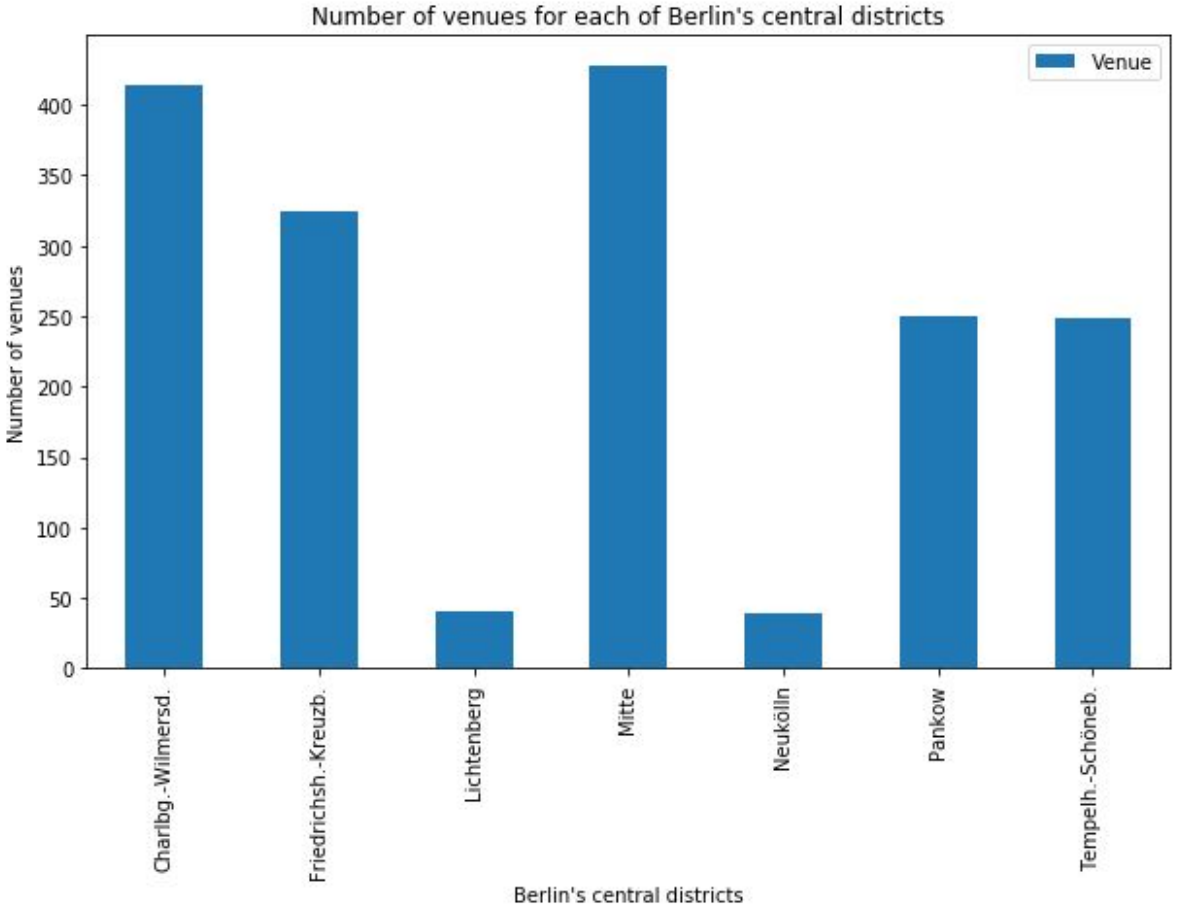
- Berlin_data.csv: Berlin's population and age (for each postal area)
- Berlin_tourist_2019.csv: tourist numbers for each district of Berlin
- berlin_geocoordinates.csv: longitude and latitude data for each postal code area of Berlin

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

Usage of Foursqare API

- Foursquare API was used for retrieving most common venues of each search area
- search criteria: 100 as limit for number of venues returned and 300 m as search radius
- The end result of the search was 1746 venues which I grouped according to the districts they belong to.

Three districts have the majority of venues, while Lichtenburg and Neukölln have the fewest

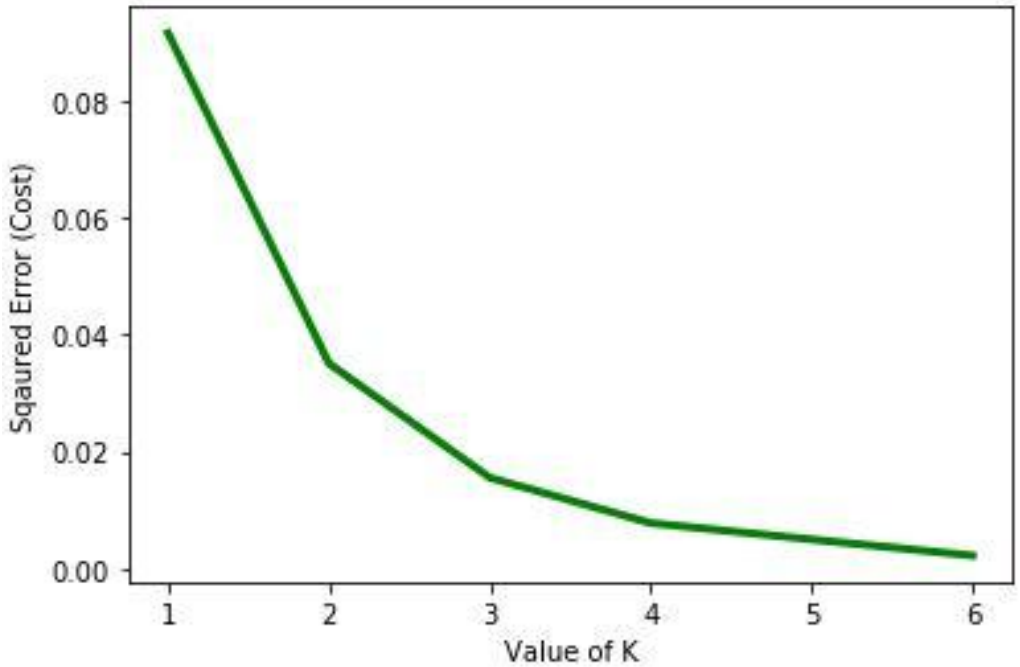


Overview of the top 5 venues for each district area

	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

Choose K Value for K-Means Clustering Algorithm

Using the Elbow Method, K should be set to value 4



Most common venues in Cluster 1 and Cluster 2

	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

	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

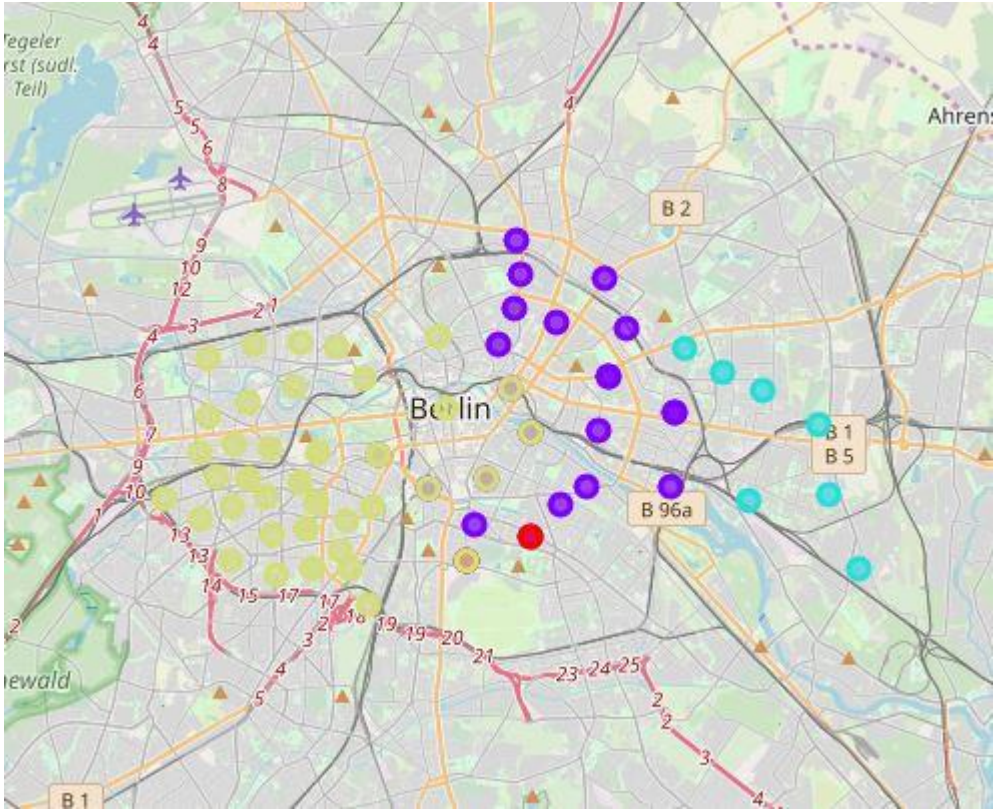
Most common venues in Cluster 3 and Cluster 4

	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

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Commo
0	Mitte	Hotel	Café	Italian Restaurant	German Restaurant	Coff
1	Mitte	Hotel	Café	Italian Restaurant	German Restaurant	Coff
2	Mitte	Hotel	Café	Italian Restaurant	German Restaurant	Coff
5	Mitte	Hotel	Café	Italian Restaurant	German Restaurant	Coff
7	Mitte	Hotel	Café	Italian Restaurant	German Restaurant	Coff
24	Mitte	Hotel	Café	Italian Restaurant	German Restaurant	Coff

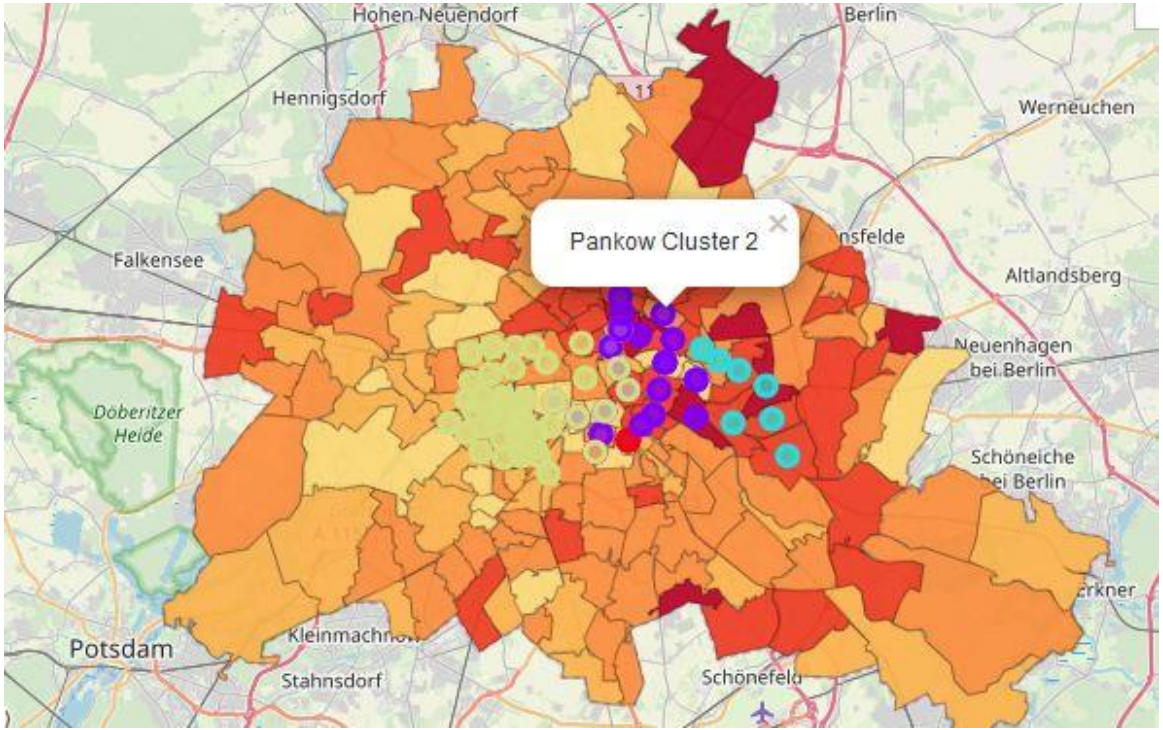
Cluster Map of Berlin

Cluster 1 (food/restaurants) is red
Cluster 2 (night and social life) is purple
Cluster 3 (hotels, touristic and social venues) is yellow
Cluster 4 (residential area) is blue



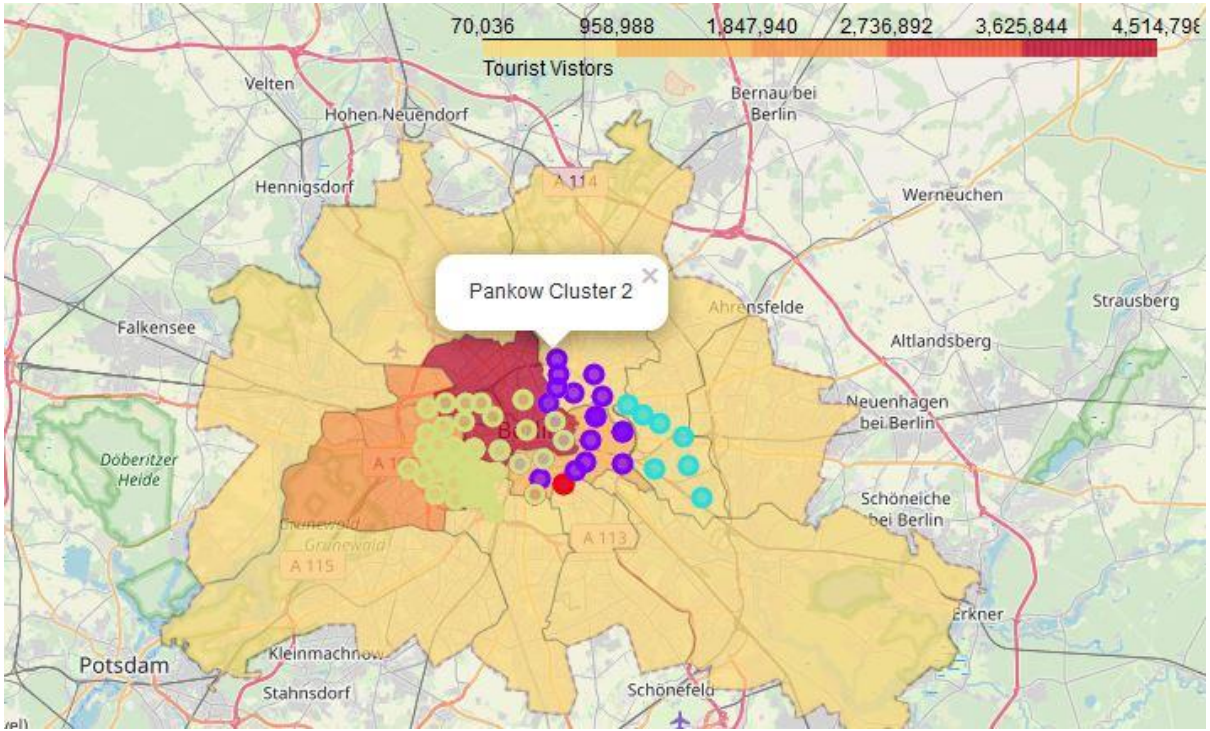
Choropleth map of Berlin's population density, combined with cluster labels

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.



Choropleth map of Berlin’s tourist visitors combined with cluster labels

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.



Conclusions and future directions

- 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.
- Foursquare has limitations (user-generated data might not be complete and up-to-date and does not take into account venue's size)
- If popular venues are retrieved from more popular apps such as Google, the analysis will be more precise and expand to bigger area