IBM Coursera Capstone: Week 5 Assignment

Problem description:

Berlin is the capital of Germany and a very fast growing city. Due to this constant growth prices for houses and flats have reached new heights and it gets more and more expensive to live in the city. The city is also a hotspot for creative internet startups like Delivery Hero, Zalando and N26. All these companies offer interesting jobs for Data Scientists.

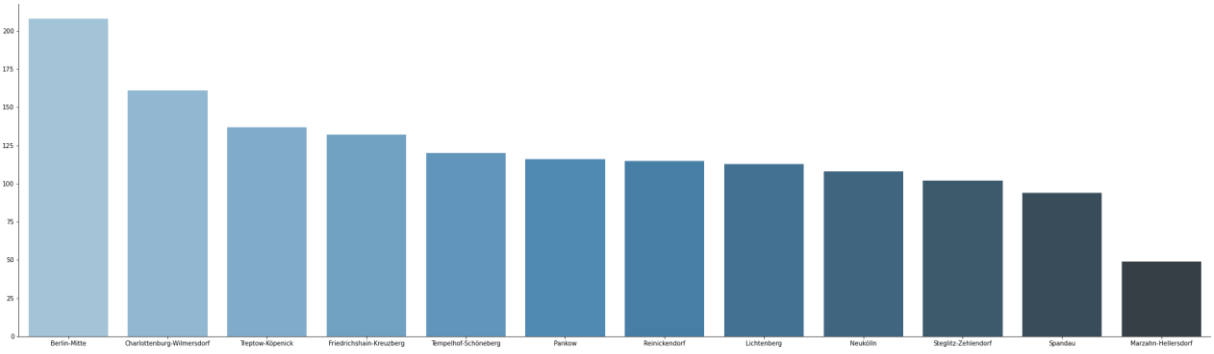
The goal of this project is to combine data about the housing market with data about venues from Foursquare to see how homogeneous the city is and if it possible to have similar venues in a cheaper neighborhood compared to the most expensive ones.

Data:

Four main datasets have been used in this project:

1. Data about Berlins Neighborhoods and Boroughs from Kauperts. Kauperts is a famous creator of plans for tourists that come to Berlin to show them the city.
2. Data about prices for new homes in Berlin. These data set is from statista.
3. Data about the population in the Boroughs and Neighborhoods of Berlin from Citypopulation.
4. Foursquare data to show the venues in the Neighborhoods of Berlin.

The data from Kauperty, Citypopulation and Foursquare was directly accessible via BeautifoulSoup or the Foursquare API. The data from statista can be found as a csv in the folder on GitHub. The data was downloaded and merged to get an overview over the m² price for new homes in the different boroughs. Additionally, the number of people living in each borough was added.



The data of the different Neighborhoods were used to get the longitude and latitude data. With this data the Foursquare API was called. In total 1455 venues were found for the 96 different Neighborhoods.

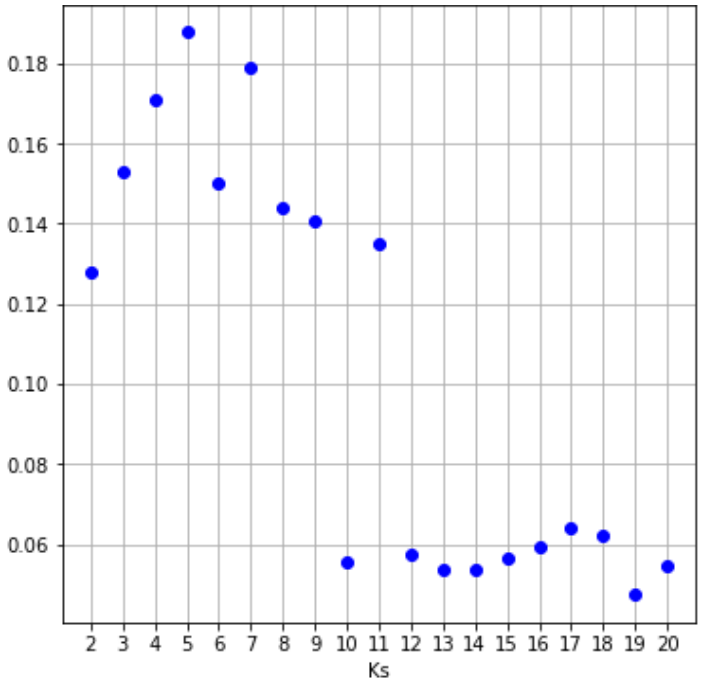
Methodology:

The price per m² for a new home and the number of people living in a certain Neighborhood were scaled with StandardScaler to make them comparable

Borough	Price	People	Price_Scaled	People_Scaled
Charlottenburg-Wilmersdorf	4856	343592	1.41	0.61
Berlin-Mitte	4693	385748	1.19	1.48
Friedrichshain-Kreuzberg	4555	290386	1.01	-0.49
Pankow	4467	409335	0.90	1.96
Steglitz-Zehlendorf	4068	310071	0.38	-0.08
Tempelhof-Schöneberg	4031	350984	0.33	0.76
Neukölln	3758	329917	-0.03	0.33
Lichtenberg	3585	294201	-0.25	-0.41
Reinickendorf	3053	266408	-0.94	-0.98
Treptow-Köpenick	3013	273689	-1.00	-0.83
Spandau	2751	245197	-1.34	-1.42
Marzahn-Hellersdorf	2500	269967	-1.67	-0.91

After the Foursquare data had been downloaded for the different Neighborhoods OneHotEncoder was applied to the data to make it ready for the KMeans algorithm.

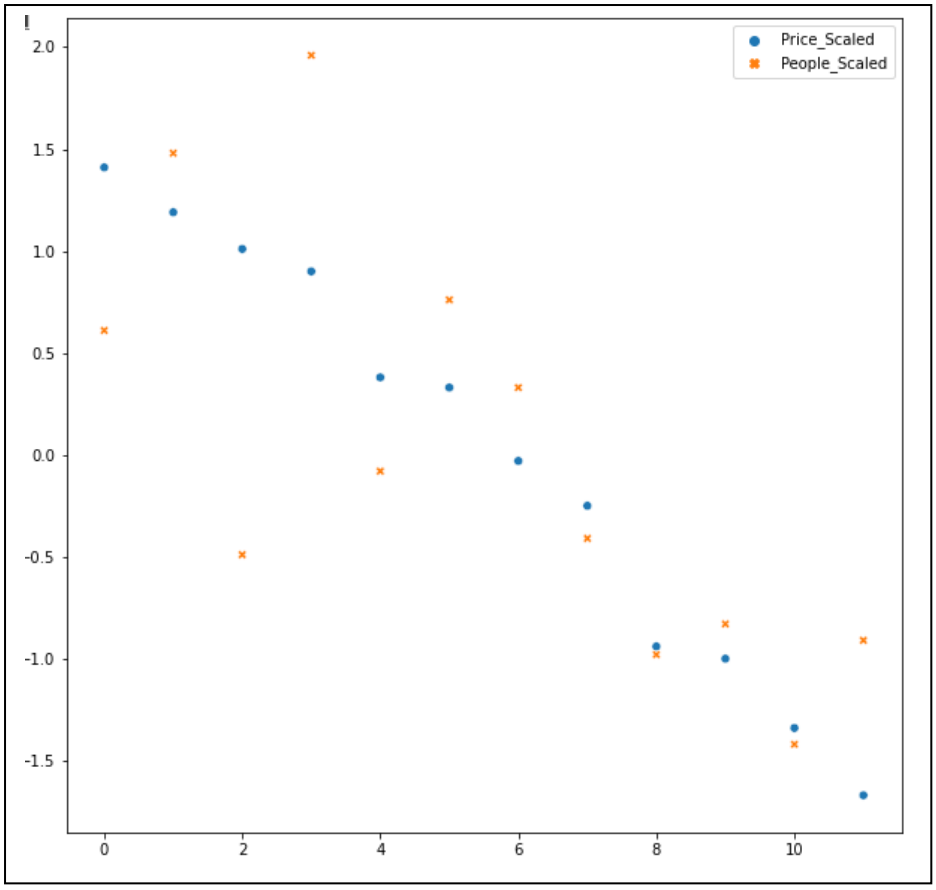
To determine the optimal value of K in KMeans the silhouette score was calculated for different Ks from 2 to 20.



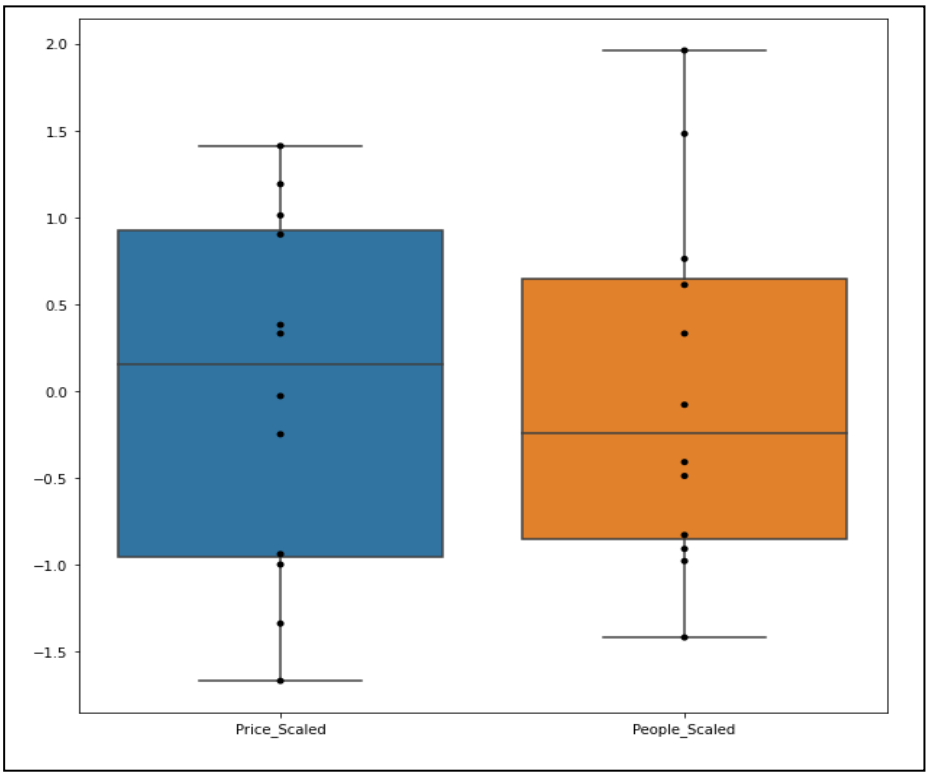
After this KMeans was applied on the whole data set to see how the similar the venues are across the different Neighborhoods

Results:

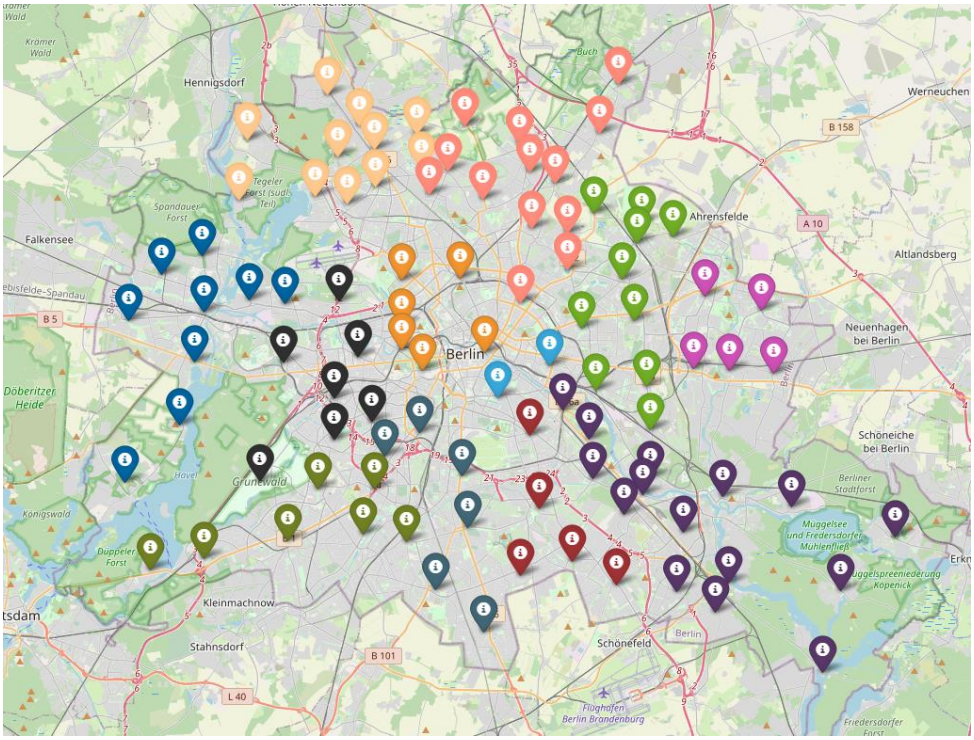
Based upon the scaling of price per m² and people living in a certain Borough there is a relationship between price and how many people live in that area. The more people want to live in a Borough the higher is the price.



The Boxplot shows that 50% of the prices are in a scaled range from 1.01 to -0.94 with outliers at 1.41 and -1.67. For people 50% are in a scaled range from 0.61 to -0.83 with outliers at 1.96 and -1.42.



The different data points visualized on a map of Berlin look like this while each color represents a Borough of the city.

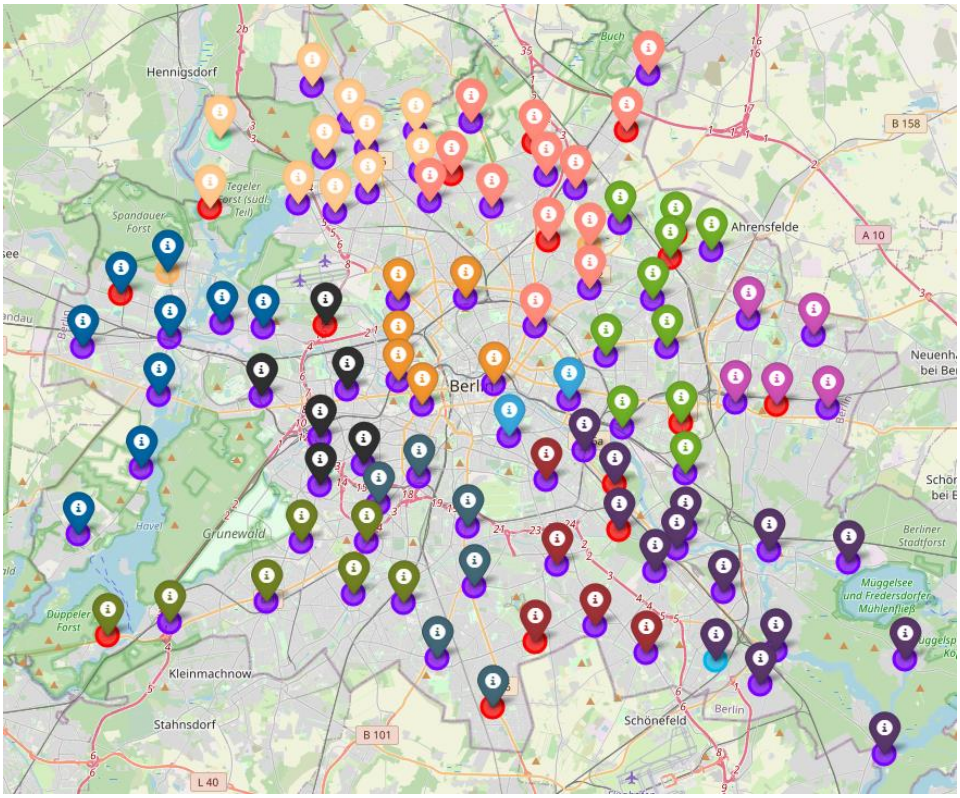


According to the silhouette Score the algorithm was only working at an accuracy level of about 18%. With this the relatively low (it should be okay as this is more about showing what we have learned I guess). Therefore, results should be treated carefully. As K was set to 5 the distribution of the different clusters looks like this:

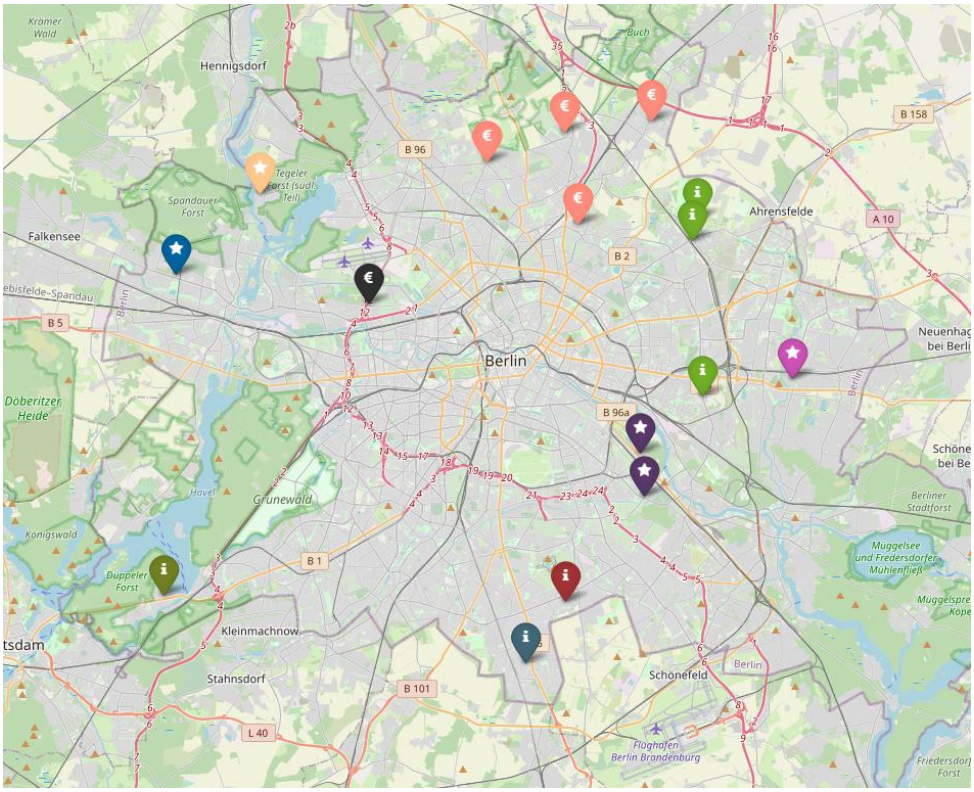
	Cluster Labels
0	16
1	74
2	1
3	1
4	2

It looks like Berlin is very homogeneous based upon the ventures returned by Foursquare because the algorithm assigned the huge majority to cluster 1 and some to cluster 0. It seems like cluster 2, 3 and 4 are very unique.

To visualize this each cluster is represented by a circle in the following map.



To answer the question which price a person needs to pay for living that wants to live in a certain cluster, cluster 0 and 1 have been looked at seperatly. A map for cluster 0 looks like this:



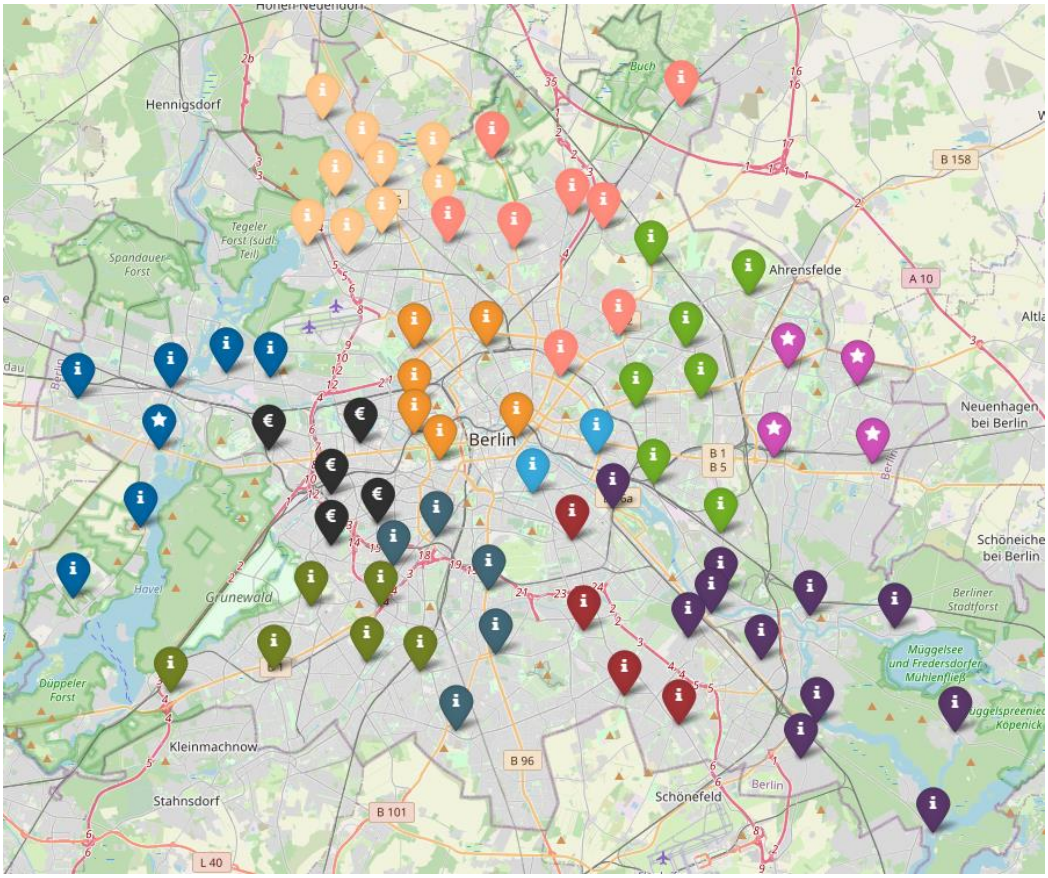
The Euro-symbol shows the top 5 expensive neighborhoods while the star shows the cheapest ones. The info symbol shows everything in-between and the color of the marker represents the borough.

Neighborhood	Borough	Price	People	Price_Scaled	People_Scaled	Latitude	Longitude	Cluster Labels
Kaulsdorf	Marzahn-Hellersdorf	2500	269967	-1.67	-0.91	52.510132	13.580990	0
Falkenhagener-Feld	Spandau	2751	245197	-1.34	-1.42	52.552403	13.166894	0
Plaenterwald	Treptow-Koepenick	3013	273689	-1.00	-0.83	52.479544	13.478808	0
Baumschulenweg	Treptow-Koepenick	3013	273689	-1.00	-0.83	52.461694	13.481548	0
Konradshoehe	Reinickendorf	3053	266408	-0.94	-0.98	52.585684	13.223198	0

Neighborhood	Borough	Price	People	Price_Scaled	People_Scaled	Latitude	Longitude	Cluster Labels
Rosenthal	Pankow	4467	409335	0.90	1.96	52.598319	13.375519	0
Karow	Pankow	4467	409335	0.90	1.96	52.615087	13.486276	0
Heinersdorf	Pankow	4467	409335	0.90	1.96	52.572825	13.437015	0
Franzoesisch-Buchholz	Pankow	4467	409335	0.90	1.96	52.610513	13.428110	0
Charlottenburg-Nord	Charlottenburg-Wilmersdorf	4856	343592	1.41	0.61	52.540525	13.296266	0

This clearly shows a difference in the price that a person who wants to live in cluster 0 must pay in Berlin. The person could either move to Kaulsdorf in Marzahn-hellersdorf and should be able to find a more attractive price compared to Charlottenburg Nord in Charlottenburg Wilmersdorf.

The same has been done for cluster 1:



Neighborhood	Borough	Price	People	Price_Scaled	People_Scaled	Latitude	Longitude	Cluster Labels
Hellersdorf	Marzahn-Hellersdorf	2500	269967	-1.67	-0.91	52.536854	13.604774	1
Mahlsdorf	Marzahn-Hellersdorf	2500	269967	-1.67	-0.91	52.508699	13.613162	1
Marzahn	Marzahn-Hellersdorf	2500	269967	-1.67	-0.91	52.542948	13.563142	1
Biesdorf	Marzahn-Hellersdorf	2500	269967	-1.67	-0.91	52.510992	13.555013	1
Wilhelmstadt	Spandau	2751	245197	-1.34	-1.42	52.513921	13.191452	1

Neighborhood	Borough	Price	People	Price_Scaled	People_Scaled	Latitude	Longitude	Cluster Labels
Wilmersdorf	Charlottenburg-Wilmersdorf	4856	343592	1.41	0.61	52.487115	13.320330	1
Westend	Charlottenburg-Wilmersdorf	4856	343592	1.41	0.61	52.513399	13.255842	1
Schmargendorf	Charlottenburg-Wilmersdorf	4856	343592	1.41	0.61	52.478902	13.292996	1
Halensee	Charlottenburg-Wilmersdorf	4856	343592	1.41	0.61	52.497226	13.292999	1
Charlottenburg	Charlottenburg-Wilmersdorf	4856	343592	1.41	0.61	52.515747	13.309683	1

Discussion & Conclusion:

The project showed that machine learning can be used to decide on where to buy a room to live based upon similarity of the ventures in the Neighborhood and that similar ventures are available in areas that have a significant price difference. The project could be further improved by gathering more detailed data on the housing market in Berlin. The data source used here only distinguishes by Borough and not by Neighborhood and even with in the same Borough there can be significant price differences between Neighborhoods.