The Battle of Neighborhoods – Final Report

Philipp Spangenberg

September 24th, 2020

1. Introduction

1.1 Background

The Nuremberg Metropolitan Region comprises 3.5 million people on 21,800 square kilometers. It consists of the cities of Nuremberg, Fürth, Erlangen, Bayreuth and Bamberg and is one of Germany's strongest economic areas. Due to a decline in historically prevalent industry, such as consumer electronics the area has lacked behind in economic development compared to other more famous German regions, such as Munich or Stuttgart.

However, this is also means that real estate and wages are lower compared to its contemporaries. Thus, potential investors find a large pool of well-educated workers, consumers and relatively cheap real estate.

1.2 Business Understanding/Problem Description

The optimal location for an investor would maximize population density, while minimizing real estate prices and competition. These values vary significantly from district to district and from city to city.

Therefore, we want to create a map, which charts all areas according to its real estate values, population and venue density. The features targeted in this model are the following:

- The density of competitors (number of venues)
- Investment expenses (Average price per SQM)
- Potential customers (total population & population density)

Afterwards, each district is clustered according to the density of venues and business opportunities.

2. Data Description

The following data sources were identified to tackle the business problem:

- The number of venues within the certain radius of each district (Foresquare API)
- The net income per citizen per district. Source: http://www.boeckler.de/pdf/wsi_vm_verfuegbare_einkommen.xlsx
- The population and the population density of the district. Source: http://www.daten.statistik.nuernberg.de/geoinf/ia_bezirksatlas/atlas.html
- The housing prices per district. Source: https://www.wohnungsboerse.net/mietspiegel-Nuernberg/2176
- The coordinates of each district. Source: Open Street Map https://nominatim.openstreetmap.org/ui/search.html?q=nuremberg

2.1 Dataset

GitHub was utilized to store the data and versions of the final code (Jupyter notebook). A variety of data frames was extracted from public information or web scrapping. The data sets imported, using the python pandas library. Initially, the data frames were split into “Districts”, “District Coordinates”, “District Government Information” and “Nuremberg Rent Index 2020”.

After cleaning and preparing the data, the following data frame (master) was created, by merging either on the district or on the name of the Borough:

	District	Latitude	Longitude	Borough	Size (ha)	Population	Population below 18 in %	Number of Households	Population with Employment	Unemployed	Finished Houses	Price per Sqm	Population Density
0	1	49.447654	11.081863	Altstadt, St. Lorenz	86.7	5275.0	7,5	3 605	2 334	227	14	12.58	61.0
1	2	49.449398	11.090167	Marienvorstadt	60.0	1338.0	11,1	919	591	57	11	NaN	22.0
2	3	49.444268	11.070317	Tafelhof	64.7	1312.0	17,8	676	557	72	-	16.27	20.0
3	4	49.449685	11.059096	Gostenhof	51.8	9462.0	16,4	5 166	3 525	593	-	12.50	183.0
4	5	49.451141	11.063438	Himpfelshof	65.4	6193.0	13,6	3 616	2 614	196	-	12.45	95.0

Figure 2.1.1 Master Data Frame

Since a total of 93 Boroughs were available, the decision was made to reduce the dataset to Borough’s with meaningful population. Therefore, the dataset was reduced to those where the population was larger than 5000. The emerging data frame showed the following characteristics:

	Latitude	Longitude	Size (ha)	Population	Price per Sqm	Population Density
count	49,00	49,00	49	49	49	49,0
mean	49,44	11,08	210	8.900	11,10 €	80,1
std	0,03	0,04	291	3.183	1,14 €	60,7
min	49,35	11,02	43	5.039	8,76 €	7,0
25%	49,42	11,06	80	6.197	10,37 €	29,0
50%	49,44	11,08	125	8.285	10,92 €	61,0
75%	49,46	11,11	252	10.401	11,75 €	118,0
max	49,49	11,21	1.943	19.927	14,25 €	237,0

Figure 2.1.2 Master Data Frame – Descriptive Statistics

2.2 Neighborhood

The Foursquare API to explore the boroughs and segment them. I designed the limit as 300 venues and the radius of 500 meter for each borough from their given latitude and longitude information's. Here is a head of the list Venues name, category, and latitude and longitude information's from the Foursquare API:

	Borough	Borough Latitude	Borough Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Altstadt, St. Lorenz	49.447654	11.081863	Sangam	49.448187	11.081314	Indian Restaurant
1	Altstadt, St. Lorenz	49.447654	11.081863	Kokoro	49.447690	11.080558	Sushi Restaurant
2	Altstadt, St. Lorenz	49.447654	11.081863	Wurstdurst	49.447971	11.078928	Currywurst Joint
3	Altstadt, St. Lorenz	49.447654	11.081863	Atelier-Bar	49.447632	11.082496	Hotel Bar
4	Altstadt, St. Lorenz	49.447654	11.081863	Park Plaza	49.447502	11.083238	Hotel

Figure 2.2 Sample of extracted venues

The venues were subsequently counted and ordered by top categories of venues. This was done in order to identify potentially missing types of business, later on in the analysis:

	Borough	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	Almoshof	Construction & Landscaping	Scenic Lookout	Hill	Trattoria/Osteria	Castle	Motel	Forest	Food Truck	Food & Drink Shop	Fishing Store
1	Altenfurt Nord	Hotel	Restaurant	Greek Restaurant	Chinese Restaurant	Café	Czech Restaurant	Dance Studio	Forest	Food Truck	Food & Drink Shop
2	Altenfurt, Moorenbrunn	Hotel	Supermarket	German Restaurant	Bank	Electronics Store	Franconian Restaurant	Forest	Food Truck	Food & Drink Shop	Fishing Store
3	Altstadt, St. Lorenz	Hotel	Bar	Drugstore	Café	Burger Joint	Hotel Bar	Plaza	Irish Pub	Rock Club	Sushi Restaurant
4	Altstadt, St. Sebald	Café	German Restaurant	Franconian Restaurant	Bar	Plaza	Italian Restaurant	Coffee Shop	Ice Cream Shop	Vietnamese Restaurant	Tapas Restaurant

Figure 2.3 Most common venues by Borough

3. Methodology

After cleaning and preparing the data, the task of creating a model was broken down into several steps, in order to identify the optimal districts.

Firstly, basic exploratory analysis was applied to the data. This was done by creating bar charts for the most meaningful features of the dataset. Secondly, the number features in data frame was either reduced, or replaced by more reasonable data. Finally, a cluster analysis was to find the best cluster of boroughs with meaningful features.

The python folium library was used to visualize geographic details of Nuremberg and its Borough's. After testing the latitude and longitude from the master data frame and sample testing several boroughs against google maps, the following visual was saved for the later clustering analysis:

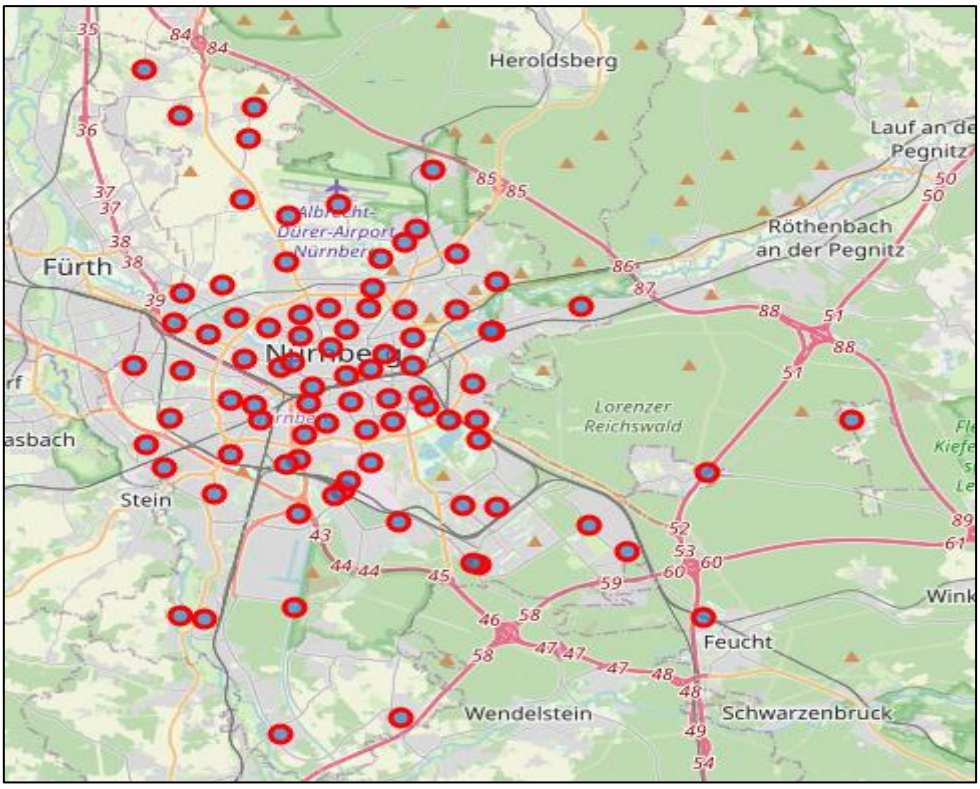

Figure 3. Nuremberg map with centers of the Boroughs

3.1 Exploratory Data Analysis

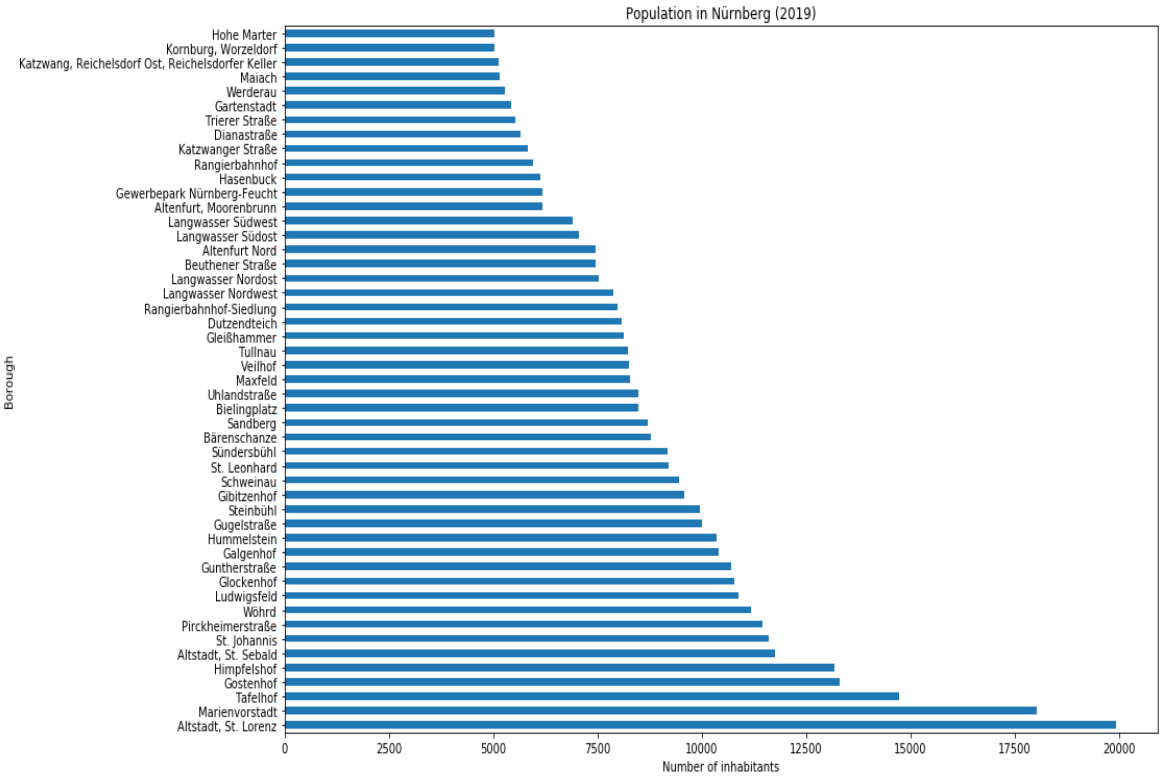

Figure 3.1.1

Figure 3.1.1 shows that the Alstadt and several adjacent Boroughs has the highest population and therefore the greatest number of potential customers. After examining the data, it was decided to drop data points with a population of less than 5000 people per Borough.

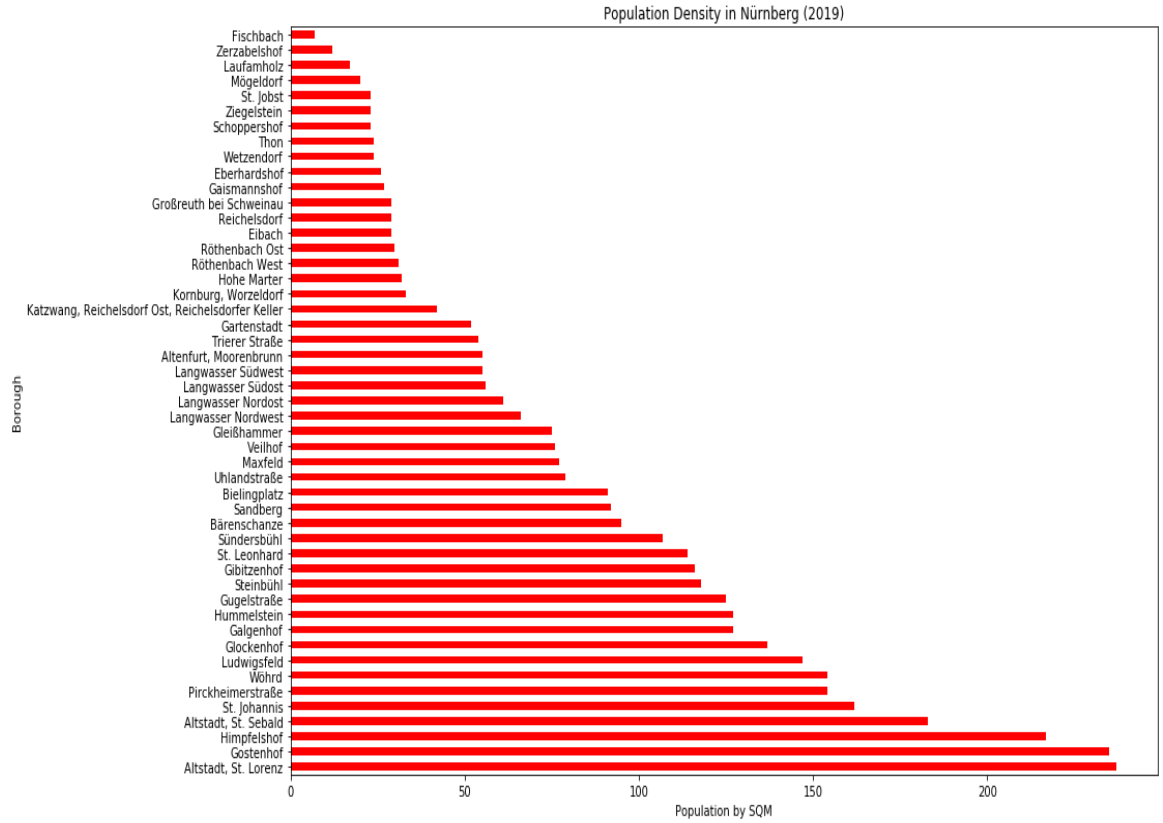

Figure 3.1.2

Figure 3.1.2 is more meaningfull as it shows the population density. Apart from the inner city, the Boroughs of Gostenhof shows that the Alstadt and several adjacent Boroughs has the highest population and therefore the greatest number of potential customers.

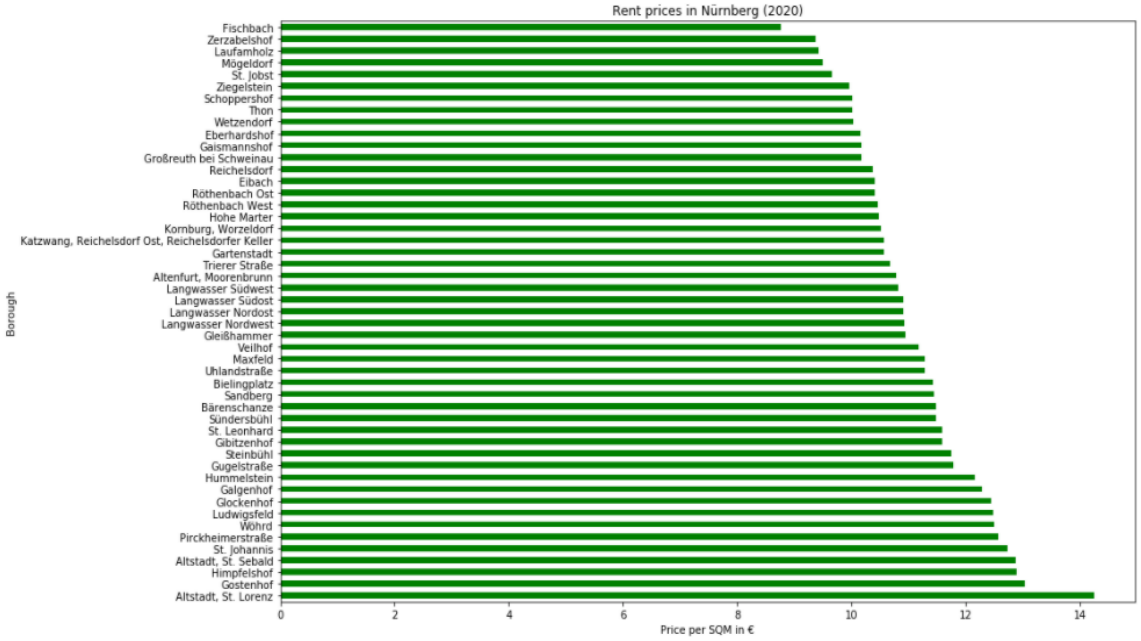

Figure 3.1.3

As can be seen, the most populated Boroughs are also the most expensive ones. Therefore, a balance between potential customers and expenses has to be the target goal.

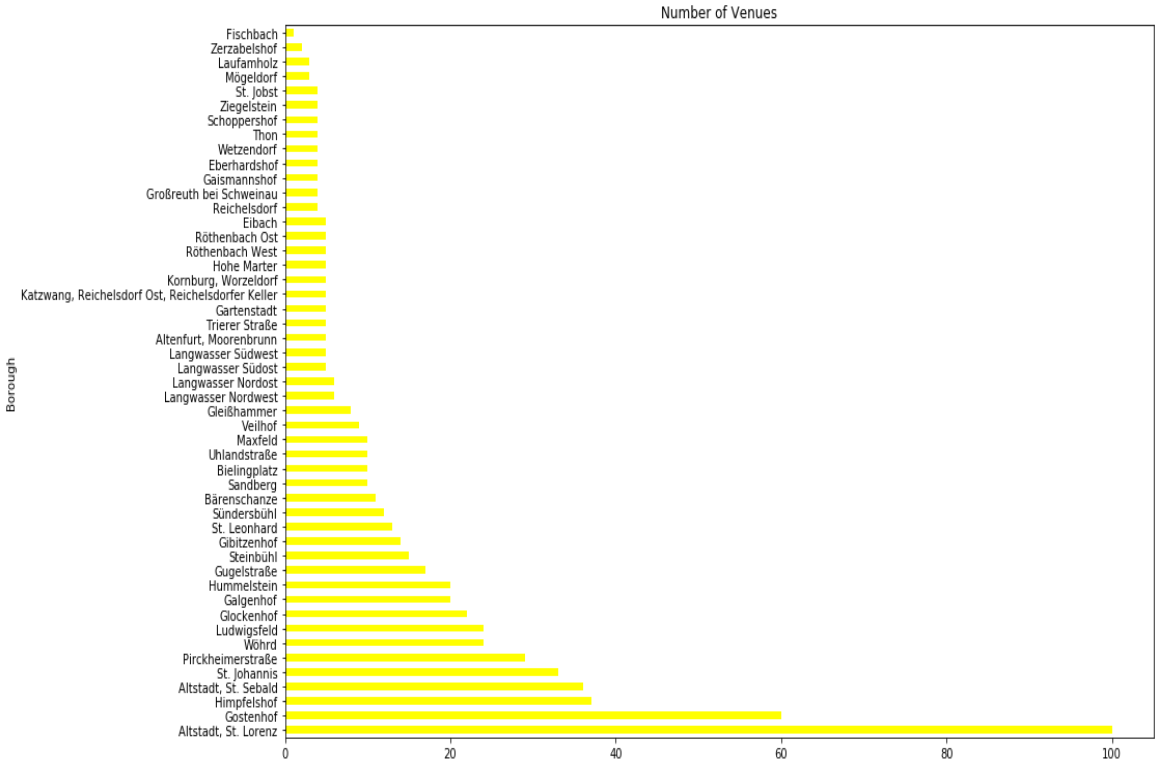

Figure 3.1.4

The large count of venues in St. Sebald and St. Lorenz makes them rather undesirable, due to the large amount of competition.

3.2 Clustering

In order to identify groups (clusters) with similar characteristics, the unsupervised machine learning method of clustering (K-Means) will be applied to the data set.

To reduce dimensionality, the “number of venues variable was replaced by “Venues per 1000 people”:

	Price per Sqm	Population Density	Number of Venues per 1000 people
0	12.58	61.0	11.374408
3	12.50	183.0	3.910378
4	12.45	95.0	3.229453
5	12.74	116.0	10.873111
6	11.59	137.0	4.052560

Figure 3.2.1

In addition, as many of the independent variables have different scales (population density & price per Sqm), the variables were transformed and normalized by using the standard scaler of the Sklearn library.

The elbow method was used to identify the optimal number of clusters:

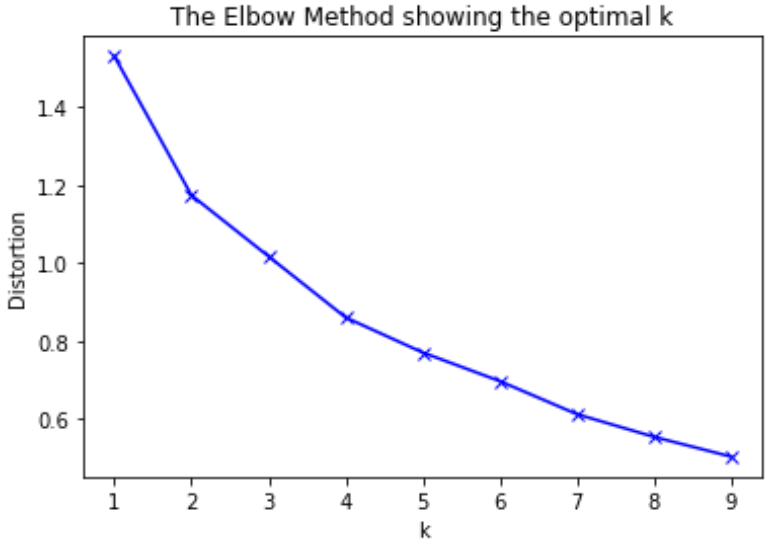

Figure 3.2.2

As can be seen in the graph, viable choices would be 2 or 4 clusters. After testing both choices, it was decided that 4 clusters yielded superior results.

Finally, the folium map from section 3.0 was combined with the resulting cluster labels. The result is a map which splits the area into clusters of similar characteristics (Population, Price Level and Count of Business per 1000 inhabitants):

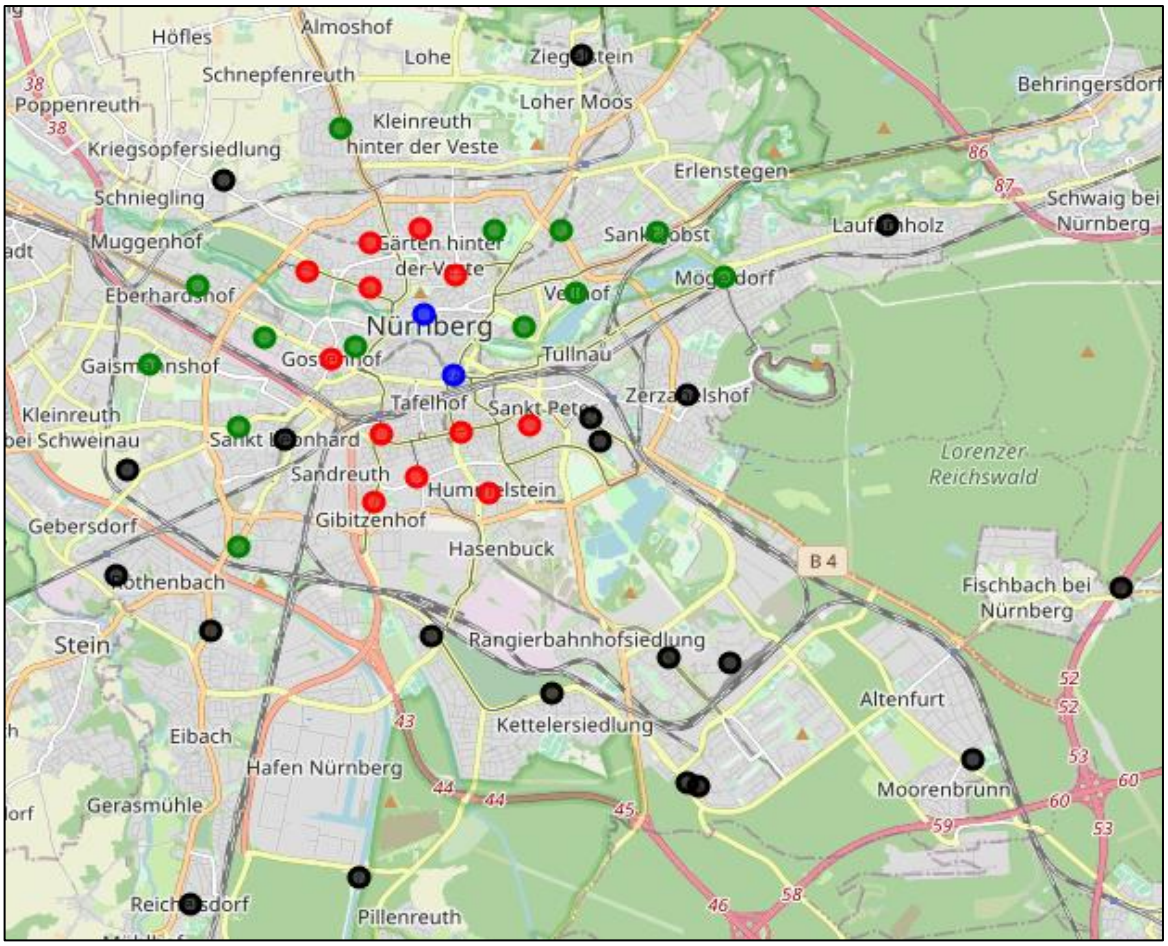

Figure 3.2.3 Nürnberg Map with clustered Boroughs

As can be seen in the map, the city center consists of two circles (blue), the outer areas (black) represent the suburbs. The green and red areas are a mixture of residential and commercial districts.

In order to compare clusters on the independent variables, scatter plots were created. (Purple = Red; Blue = Green; Black = Blue; Green = Yellow). The grey circle represents the centroid of each cluster. The data is normalized; therefore, it does not indicate real values.

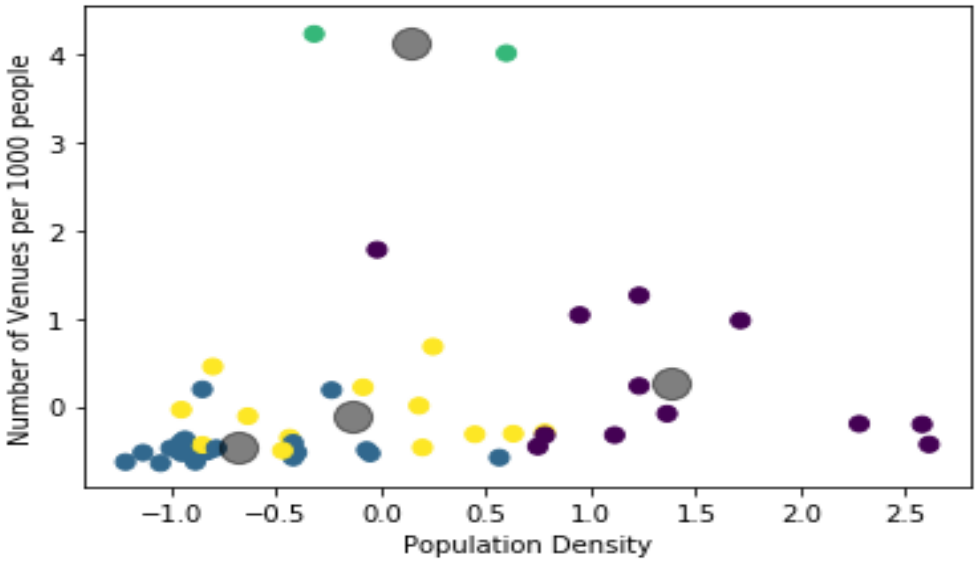

Figure 3.2.4

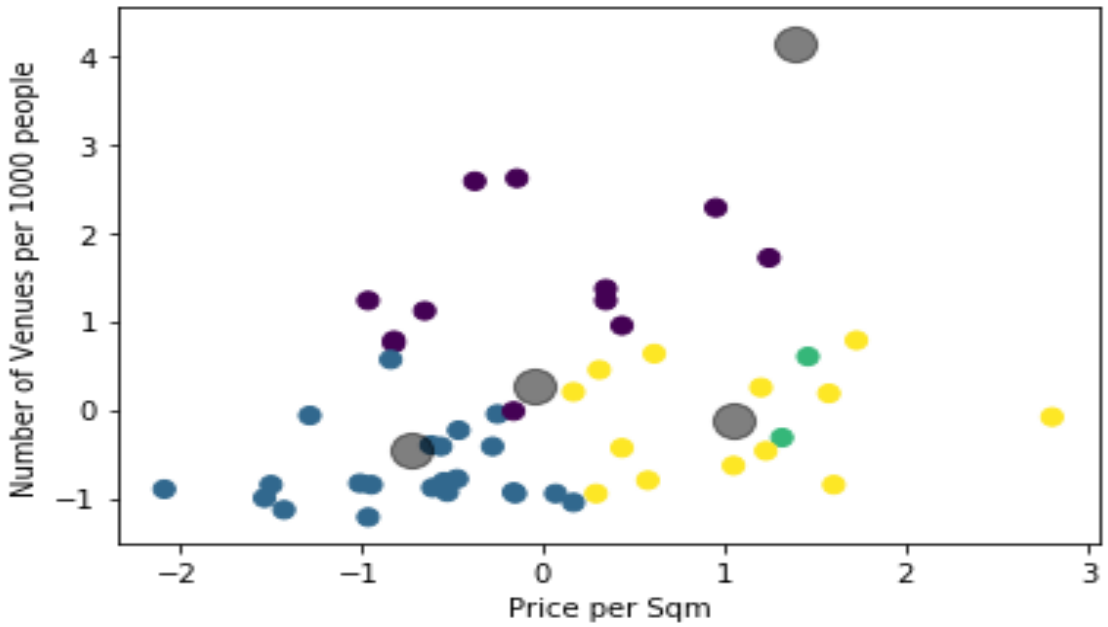

Figure 3.2.5

Positive outliers are the purple (red in the map) Boroughs, they show a medium amount of venues, a large population and an average price per Sqm.

In order to better understand the indivual clusters the most common venues per Borough were added to a final data frame:

Borough	Size (ha)	Population	Price per Sqm	Population Density	Count of Venues	1st Most Common Venue_x	2nd Most Common Venue_x	3rd Most Common Venue_x	4th Most Common Venue_x	5th Most Common Venue_x
Gostenhof	51.8	9462.0	12.50	183.0	37	Café	Bakery	Gastropub	Greek Restaurant	Bar
St. Johannis	59.4	8143.0	11.59	137.0	33	Café	German Restaurant	Greek Restaurant	Thai Restaurant	Park
Pirckheimerstraße	51.2	7890.0	11.49	154.0	36	Castle	Café	Bar	Coffee Shop	Vietnamese Restaurant
Glockenhof	83.0	18042.0	12.17	217.0	22	Supermarket	Bakery	Italian Restaurant	Sporting Goods Shop	Dance Studio
Galgenhof	84.8	19927.0	10.68	235.0	24	Bar	Supermarket	Drugstore	Hotel	Bakery

Figure 3.2.6

This combines all gathered and relevant information for each cluster and each Borough. As can be seen Glockenhof and Galgenhof have a high population density of over 200 people per Hektar and are comparatively cheap.

4. Result

During the analysis, four clusters were defined. The first (blue) consists of two areas and represents the city centre. It is expensive and sports a high amount of potential competition. The second (black) consists of 22 areas and is mostly residential.

This leaves the red and the green cluster as potential areas to open a business. Between the two, the red cluster has a lower average price per SQM and a lower amount of venues. However, the green cluster has a population density of 163 (compared to 72). Therefore, is has a greater pool of customers, while only being slightly more expensive.

Furthermore, the red clusters are close to the centre and well connected by public infrastructure. In order to identify the best most promising area among the red cluster and identify the optimal type of business, additional analysis is required. This would involve a clustering of the most common venues, based on their string values. This is not easy to perform in unison with the normalized float values of the quantitative components (such as population).

5. Discussion

A point of improvement would be full access to the foursquare library, in order to escape current limitations. Furthermore, foursquare doesn't represent the full amount of venues, since many venues are not on the listed. Therefore, another map could be utilized such as Google map or Open street map. Moreover, districts have a complex geometry, thus defining the closest venues within the certain radius brings additional error to our analysis.

Lastly, the income and therefore purchasing power of the population was not publicly available by cluster. Adding this feature would greatly improve the model.

6. Conclusion

To conclude, the data sets were scrapped from public sources, loaded into data frames, cleaned and combined into usable data frames ready for analysis. Exploratory data analysis was done in order to understand the data and find angles for later analysis. The data was then reduced to its most important components and clustered, in order to find the optimal areas for business opportunities. Lastly, the red cluster was chosen as a pool of potential locations. On first glance, the Boroughs of Glockenhof or Galgenhof seem promising.

7. Reference

[1] https://en.wikipedia.org/wiki/Nuremberg_Metropolitan_Region

[2] http://www.daten.statistik.nuernberg.de/geoinf/ia_bezirksatlas/atlas.html

[3] http://www.boeckler.de/pdf/wsi_vm_verfuegbare_einkommen.xlsx

[4] Forsquare API

[5] https://nominatim.openstreetmap.org/ui/search.html?q=nuremberg