# The Battle of Neighborhoods – Final Report

Philipp Spangenberg

September 4th, 2020



1. Introduction
   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. 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.

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

1. 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>

* 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:

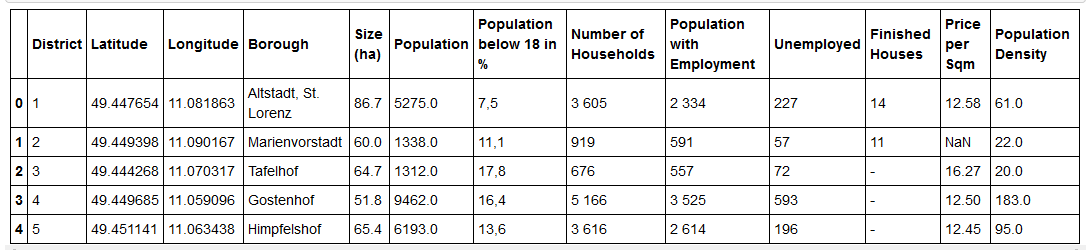


Figure 2.1.1 Master Data Frame

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:



Figure 2.1.2 Master Data Frame – Descriptive Statistics

* 1. 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:

Figure 2.2 Sample of extracted venues

1. Methodology

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

After cleaning and preparing the data, let us identify the steps, that have to be performed in order to find the most optimal boroughs .Firstly, we will apply some basic exploratory analysis to our data. For that, let's find the location of each borough on the map. Then we can visually inspect some values in our data with the help of bar charts. Secondly, we have the possibility to reduce the number features in data frame by replacing them with more reasonable data. Finally, we will perform cluster analysis to find the best cluster of boroughs with meaningful features.

* 1. Exploratory Data Analysis
  2. Clustering

1. Result
2. Discussion

During the analysis, three clusters were defined. One cluster, that consists of only one area, has been defined as the outsider, due to the high number of competitors, which means that the placement of beer restaurant in that area is too risky venture. Two other groups were clustered according to the amount income per person. It is obvious, that the cluster with highest average income per person has the highest priority for us (Cluster 2). Stuttgart-Sud and Stuttgart-Nord are the most attractive options in terms of distances to the center of their own cluster and relatively high value of income per person. However, one can perform further analysis of this particular cluster with additional features, such as distance to the center of city or to the center of cluster.After defining a borough,one can perform deeper analysis to find the best exact location of the restaurant taking into account factors such as number of parking places in the vicinity of the spot or distances to the main streets. What could be done better? Foursquare doesn't represent the full picture, since many venues are not on the list. For that reason, another map could be utilized such as Google map or Openstreet map. Boroughs have too complex geometry, thus defining the closest venues within the certain radius brings additional error to our analysis.

1. Conclusion

To conclude, the basic data analysis was performed to identify the most optimal boroughs for the placement of the beer restaurant in the city of Stuttgart.During the analysis, several important statistical features of the boroughs were explored and visualized. Furthermore, clustering helped to highlight the group of optimal areas. Finally, Stuttgart Sud and Stuttgart Nord were chosen as the most attractive options for the further analysis.

1. 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](https://developer.foursquare.com/" \t "_blank)

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

Links:

<https://www.wsi.de/de/einkommen-14582-17976.htm>

<https://www.capital.de/immobilien-kompass/nuernberg>

<http://www.daten.statistik.nuernberg.de/geoinf/ia_bezirksatlas/atlas.html>

<https://www.metropolregionnuernberg.de/daten-fakten>

<https://www.boeckler.de/pdf/wsi_vm_verfuegbare_einkommen.pdf>

<https://github.com/zhumazik/IBM-Capstone-Project/blob/master/Report.pdf>

<https://github.com/zhumazik/IBM-Capstone-Project/blob/master/Capstone_project.ipynb>

<https://github.com/rupeshvins/IBM-Applied-Data-Science-Capstone/blob/master/Data%20Science%20Project%20Week%205th%20-%20IBM.pdf>

<https://www.linkedin.com/pulse/housing-sales-prices-venues-data-analysis-ofistanbul-sercan-y%C4%B1ld%C4%B1z/>

For the second week, the final deliverables of the project will be:

1. A link to your Notebook on your Github repository, showing your code. (**15 marks**)
2. A full report consisting of all of the following components (**15 marks**):

* Introduction where you discuss the business problem and who would be interested in this project.
* Data where you describe the data that will be used to solve the problem and the source of the data.
* Methodology section which represents the main component of the report where you discuss and describe any exploratory data analysis that you did, any inferential statistical testing that you performed, if any, and what machine learnings were used and why.
* Results section where you discuss the results.
* Discussion section where you discuss any observations you noted and any recommendations you can make based on the results.
* Conclusion section where you conclude the report.

3. Your choice of a presentation or blogpost. (**10 marks**)