Where can be the next Supermarket, Grocery Store in Cologne Germany!!

Santhosh Vasudevan

31-March-2021

1. Introduction

1.1. Background & Motivation for New Supermarket

FMCG Fast Moving Consumer Goods are the products that sell quickly at relatively low cost. They are purchased for consumption by the average consumer who are part of normal common man population. They are basic essentials for sustenance of life products eg., Bread, Dairy, Fruits, Vegetables, Frozen foods, Meat and of course a huge list of household essentials. A supermarket is the first thing we look for in our neighborhood.

Having a supermarket that carries fresh fruit and veg is the number one indicator of the health of a neighborhood. Supermarkets are important indicators of financial and physical health.

Such a supermarket needs to be as close as possible to vast majority of the population. **This project is aimed at identifying densely populated areas which have relatively lesser Supermarket in the region and suggesting the optimal location for next hit Supermarket in the City of Cologne Germany**.

1.2. Business Problem Description

Modern communication has brought the order and delivery systems. Order online and have it delivered or pick it up. They don't charge extra or change the prices for this service so either the food is already marked up enough to cover it or the added competitive edge is worth the extra cost.

The Success of supermarket depends on lot of other factors like, the team working in the market will determine the level of customer service, how inventory is managed throughout the store, and general culture i.e., how the team in the store feels about upper managemnet, their jobs etc). Most supermarkets now sell the same organic -/conventional products so Supermarket needs to have a difference in business and customer experience than other competitors.

However, the above limitations will not downplay the necessity of new supermarket in a city with population of 1 million. **This project will address the Supermarket Optimum location that will cater the needy population looking to walk, drive or travel less for visiting the Supermarket.**

2. Data Sources and Description

2.1. Data Sources

2.1.1. https://www.cybo.com is a popular Global Business Directory which covers a vast majority of countries of the world. The intention is to scrape the information available about Cologne, Germany. https://postal-codes.cybo.com/germany/cologne/ Information like, postal codes, Neighborhood, Population, Area, Male Population, Female Population, Median Age, Latitude, Longitude information can be sourced.

	Information
Timezone	Central European Time
Area	281.7 km²
Population	977,691 (More Details)
Male Population	473,826 (48.5%)
Female Population	503,864 (51.5%)
Median Age	41.3
Postal Codes	50667, 50668, 50670 (43 more)
Area Codes	2203, 221

Figure 1 Generic Information on Cologne City

	Information
Population	977,691
Population Density	3,470 / km²
Male Population	473,826 (48.5%)
Female Population	503,864 (51.5%)
Median Age	41.3
Male Median Age	40.5
Female Median Age	42
Businesses in Cologne	85,523
Population (1975)	702,101
Population (2000)	892,156
Population change from 1975 to 2015	+39.3%
Population change from 2000 to 2015	+9.6%

Figure 2 Further General Information with past Population Data

Postal Code	City	Administrative Region	Population	Area	Area Codes	Neighborhoods	Latitude	Longitude	Male Population	Fema Populati
50667	Cologne	North Rhine-Westphalia	4,512	1.031 km²	2203, 221	Altstadt-Nord, Innenstadt	50.940088623517646°	6.957684918323491°	2,187 (48.5%)	2,3 (51.5
50668	Cologne	North Rhine-Westphalia	7,717	1.692 km²	2203, 221	Altstadt-Nord, Innenstadt, Kunibertsviertel	50.950695435874°	6.965450051394539°	3,740 (48.5%)	3,9 (51.5
50670	Cologne	North Rhine-Westphalia	8,612	1.9 km²	2203, 221	Agnesviertel, Innenstadt	50.95117220931733°	6.950957439590489°	4,174 (48.5%)	4,4 (51.5
50672	Cologne	North Rhine-Westphalia	4,657	1.027 km²	2203, 221	Friesenviertel, Innenstadt	50.943440989539745°	6.938476189303402°	2,257 (48.5%)	2,4 (51.5
50674	Cologne	North Rhine-	6,643	1.389	2203, 221	Innenstadt,	50.93358384190404°	6.937040819338811°	3,219	3,4

Figure 3 DataFrame of Scraped Table from source

2.1.2. Cologne City Administration publishes many of the public data https://www.offenedaten-koeln.de of which a GeoJSON file for the city boundaries and boundaries of postal codes are used to identify the correct location of neighborhoods. https://www.offenedaten-koeln.de/dataset/postleitzahlgebiete-koeln

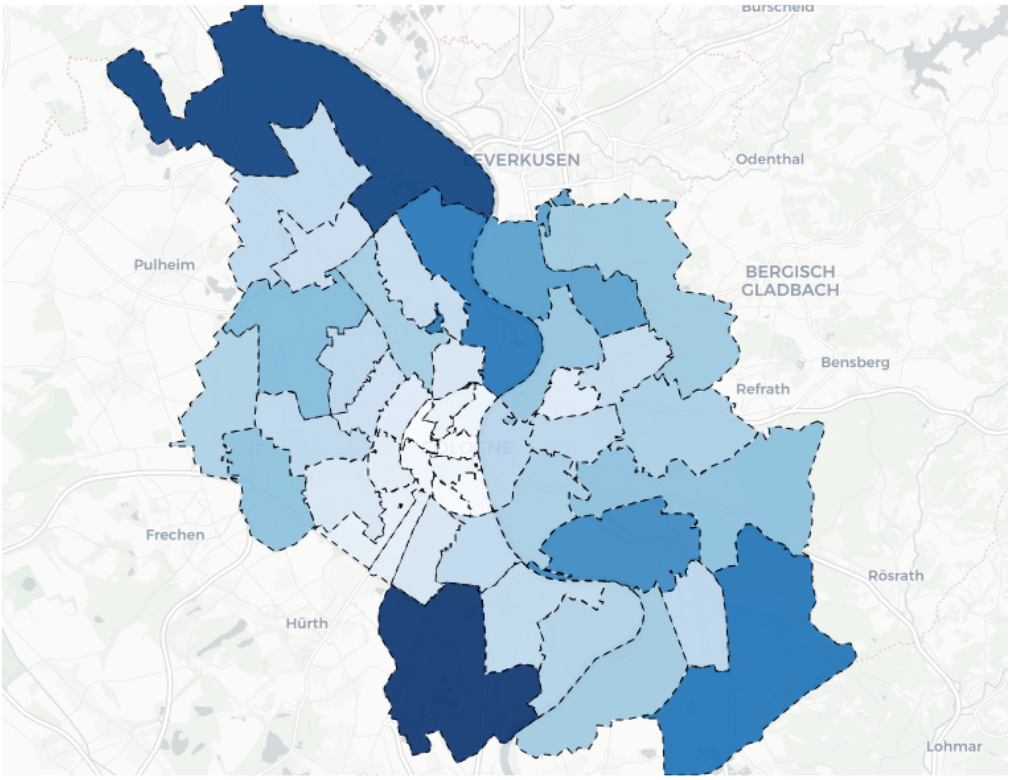

Figure 4 Cologne Postal Codes Boundaries showing Population in regions

2.1.3. Foursquare API will be used to fetch the venue information, venue categories by passing in the information obtained from above sources.

Postal Code	PostCode Latitude	PostCode Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
51149	50.906367	7.048909	Packstation 132	50.903793	7.058213	Shipping Store
51149	50.906367	7.048909	Rangierbahnhof Köln-Gremberg	50.899960	7.052129	Train Station
51149	50.906367	7.048909	A.T.U Köln - Gremberghoven	50.911754	7.056019	Automotive Shop
51149	50.906367	7.048909	Kienbaum Consultants International	50.913817	7.050480	Business Service
51149	50.906367	7.048909	CANCOM Pironet	50.914348	7.048726	Business Service

Figure 5 Venue Information from Foursquare API

2.1.4. Folium is a Data Visualisation library that lets us manipulate data in python and visualize the data in leaflet.js library. Folium will be used for Data Visualization.

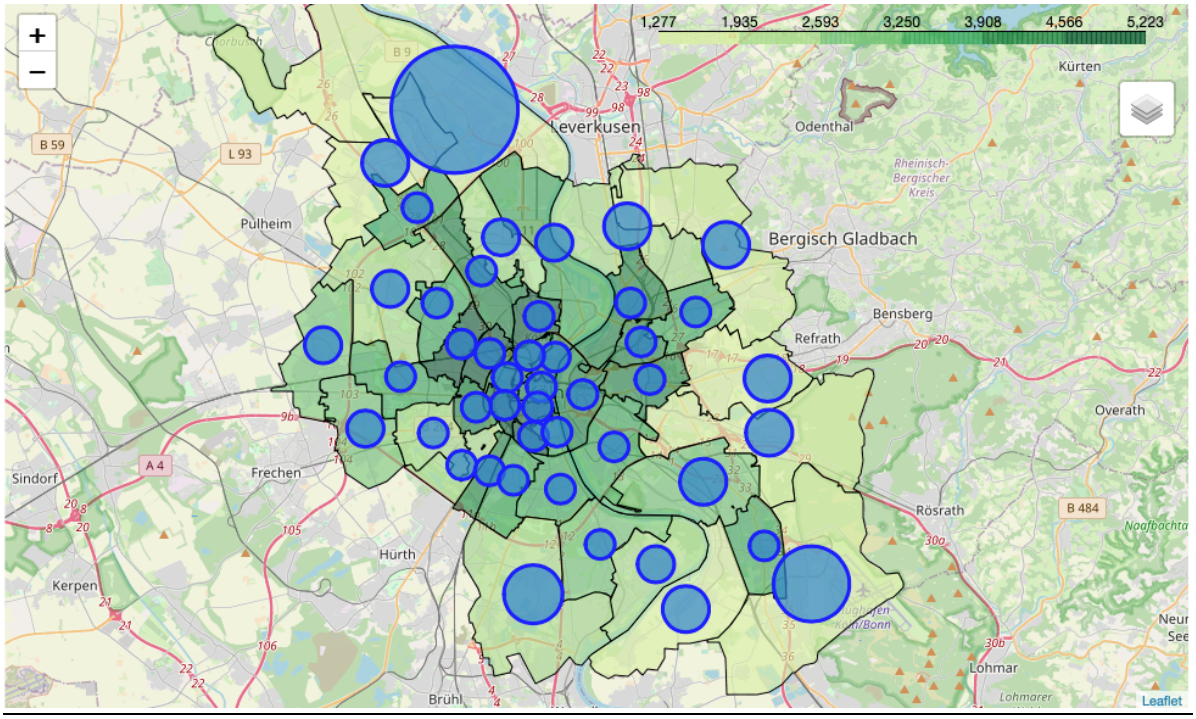

Figure 6 Folium Map showing used Radius

2.2. Data Explanation

2.2.1. The scraped data from cybo, will be used as all the information available about Cologne, primarily Postal Codes, Neighborhoods names, Population in each neighborhood and its Area along with Latitude and Longitude.

2.2.2. The boundaries GeoJSON file is obtained from Open Data of Cologne website which will help in visualization of region under lens and shadow regions.

2.2.3. Foursquare API is used to fetch the venue information in the region which is restricted to 100 venues in a query and a radius is chosen dynamically.

2.2.4. Folium is used to plot the map data available

2.2.5. Machine Learning techniques will be used to find the optimum location for Supermarket as part of solution to the business problem.

3. Methodology

3.1. Data Cleaning

The obtained data in its raw form cannot be processed. Hence some data cleaning steps are required

I. Population data is represented in textual form with thousands comma separator. This needs to be converted to integer format
II. The Area column is in text form and contains km^2 in all the values. This needs to be converted to integer
III. The Latitude and Longitude data contains ° (degrees notation). This is removed and converted to textual form.
IV. Population Density is calculated by dividing Population by Area of the postal code region.

3.2. Feature Extraction

After Data Cleaning process below features are extracted from the scraped data.

I. Postal Code
II. Population
III. Area
IV. Neighborhoods
V. Latitude
VI. Longitude
VII. Male Population
VIII. Female Population
IX. Population Density

Postal Code	City	Administrative Region	Population	Area	Area Codes	Neighborhoods	Latitude	Longitude	Male Population	Female Population	Median Age	Male Median Age	F M
50667	Cologne	North Rhine-Westphalia	4512	1.031	2203, 221	Altstadt-Nord, Innenstadt	50.940089	6.957685	2187	2325	41.3 years	40.5	
50668	Cologne	North Rhine-Westphalia	7717	1.692	2203, 221	Altstadt-Nord, Innenstadt, Kunibertsviertel	50.950695	6.965450	3740	3977	41.3 years	40.5	
50670	Cologne	North Rhine-Westphalia	8612	1.900	2203, 221	Agnesviertel, Innenstadt	50.951172	6.950957	4174	4438	41.3 years	40.5	
50672	Cologne	North Rhine-Westphalia	4657	1.027	2203, 221	Friesenviertel, Innenstadt	50.943441	6.938476	2257	2400	41.3 years	40.5	
50674	Cologne	North Rhine-Westphalia	6643	1.389	2203, 221	Innenstadt, Rathenauviertel	50.933584	6.937041	3219	3424	41.3 years	40.5	

Figure 7 Feature Table for Analysis

3.3. Exploratory Data Analysis

3.3.1. Figure 7 and Figure 8 shows the Population choropleth plot based on Postal Code regions and boundaries of each Postal Code. The Darker color regions have high population than lighter color regions. It can be seen that central part of city is shown to be with relatively low population. Certainly this could be also because the area of the region could be low and hence low population, contrarily the region with high area has high population

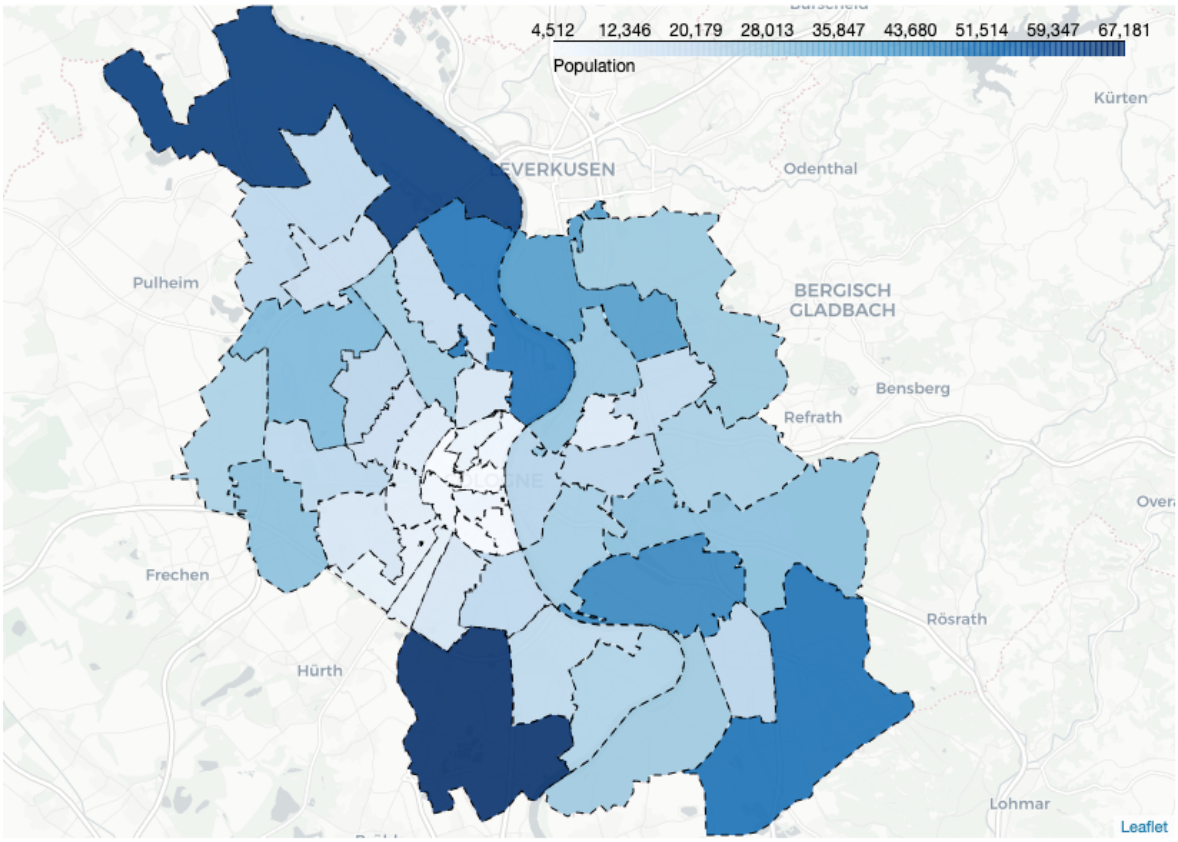

Figure 8 Population Map on region

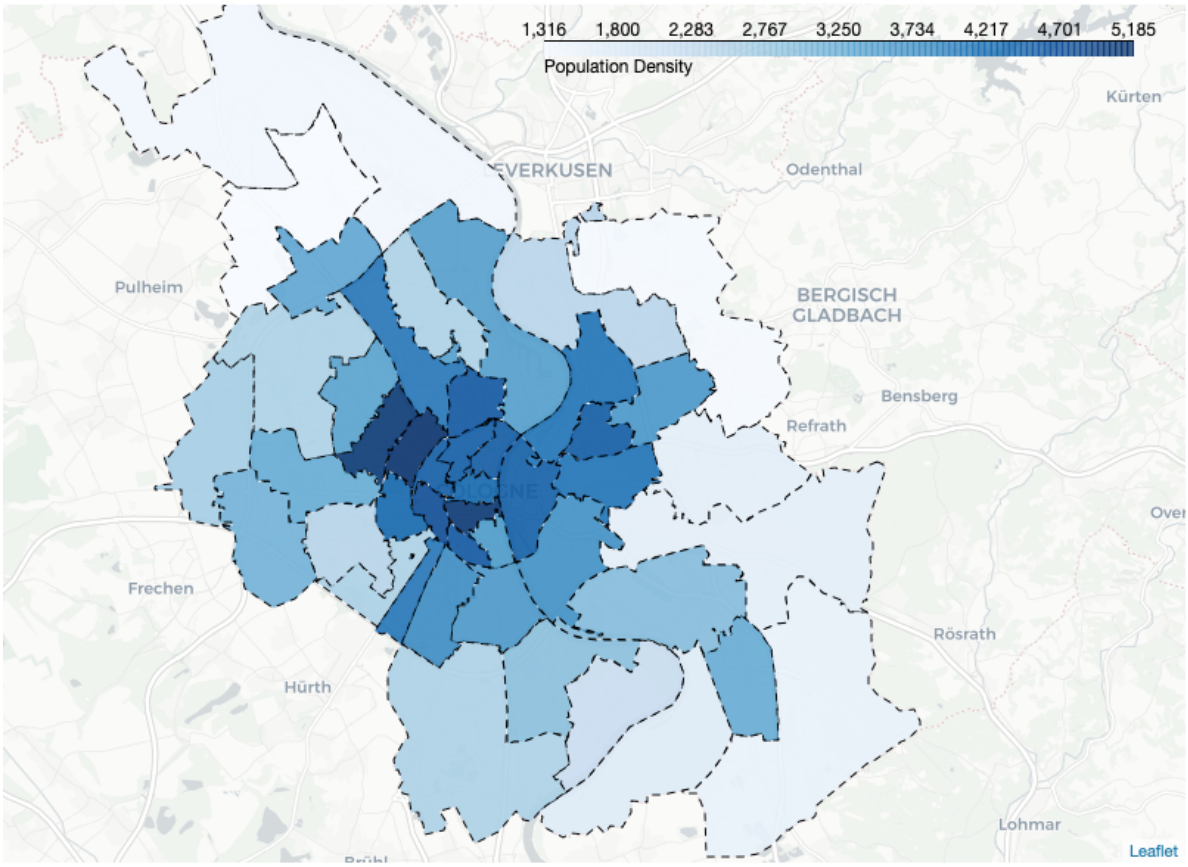

Figure 9 Map of Population Density

3.3.2. Male vs Female Population of each postal region is plotted below. The result measures that Female Population are considerably higher in the Cologne City.

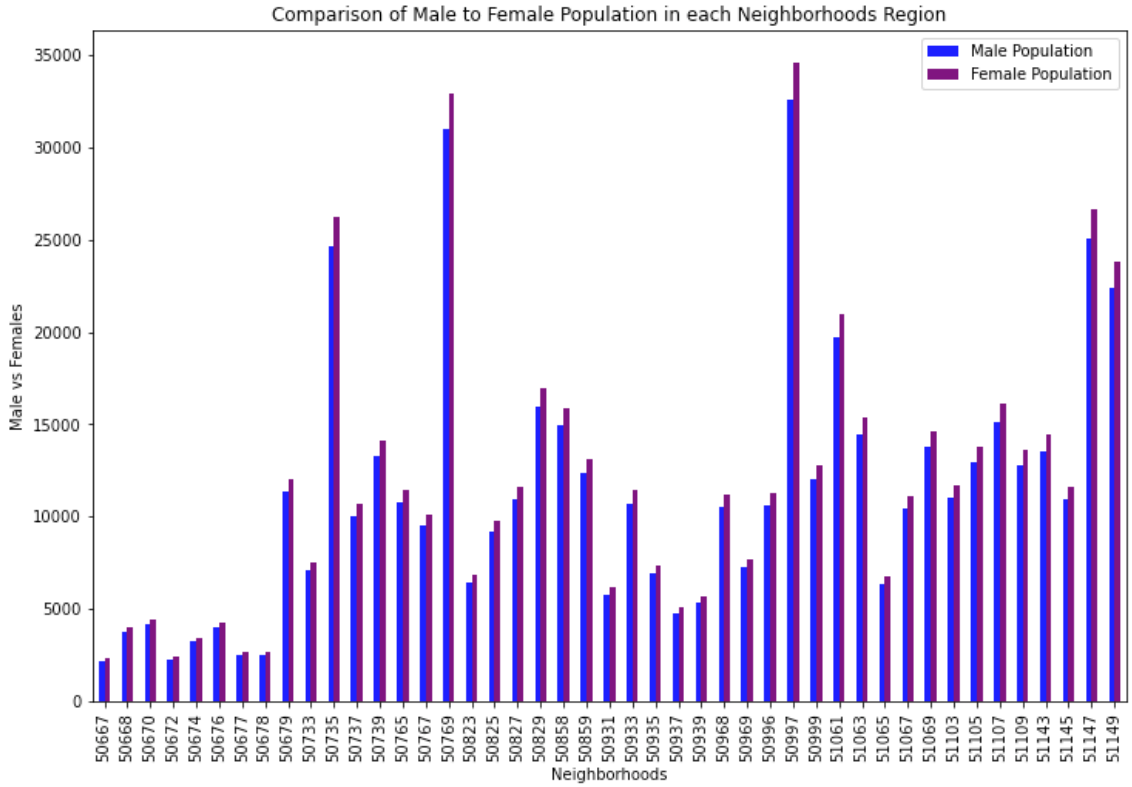

Figure 10 Male- Female Population by region

3.4. Core Methodology

3.4.1. Venue Category Identification

Venues can be found based on Foursquare API, where along with login credentials, the intended Latitude, Longitude and Radius are passed on. The result from Foursquare API is in json format. Venues Name, Venue Category and Venue Location are obtained.

```
There are about 1467 venues returned by the FourSquare API in Co
logne, Germany

There are about 242 unqiue venue categories in Cologne, Germany
```

3.4.2. Venue Category Filtering

The intention of the project is to find best location for the next upcoming Supermarket. Once the venues are stored in dataframe. The required Target Category “Supermarket”,”Grocery Store” and “Organic Grocery Store” are filtered. The top 5 rows are shown below.

	Postal Code	Supermarket	Grocery Store	Organic Grocery
0	50667	1	1	0
1	50668	1	0	0
2	50670	0	0	0
3	50672	1	0	0
4	50674	0	0	0

Figure 11 Filtered Dataframe for Analysis

3.4.3. K-Means Clustering

Target Feature variables are preprocessed by Min Max Scaler, the result of which is helpful in analysing the regions with availability of Supermarkets directly.

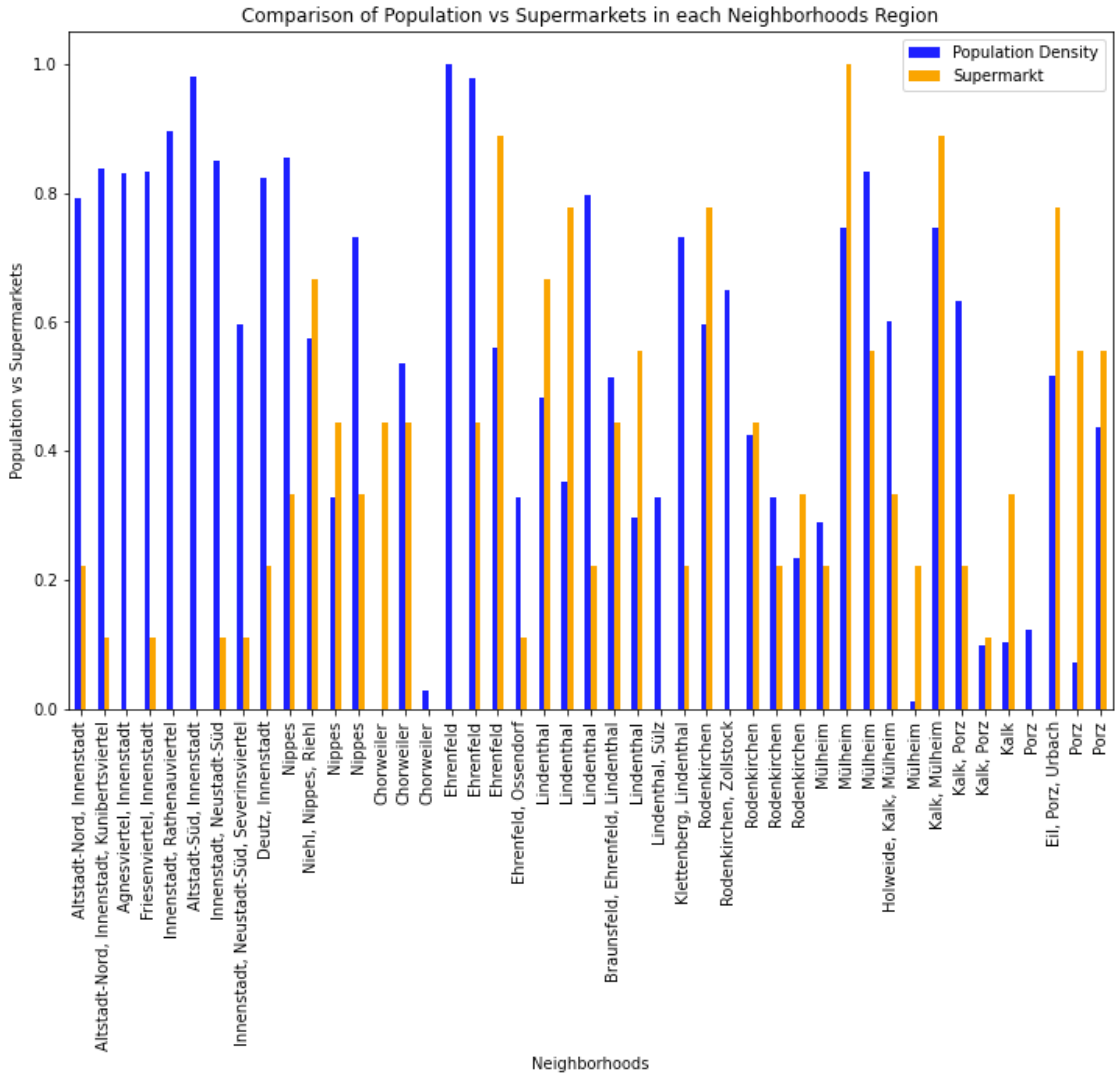

Figure 12 Co-relation between Population Density to Supermarket

The above chart can be used to directly compare the neighborhood with availability of supermarket. We can apply K-means clustering approach to group the supermarkets in the city as say k clusters. Each cluster will have similarity in terms of its availability in the region. We will compare this with Population Density and come to a conclusion of which neighborhood could be the better location.

3.4.4. Optimum K Value

How many group of clusters will be optimum in cologne supermarket dataset depends on the K-Means inertia factor. Inertial is recognized as a measure of how internally coherent clusters are. This can be determined by utilizing the scikit library. Below graph is plotted which shows the cluster where inertia is least.

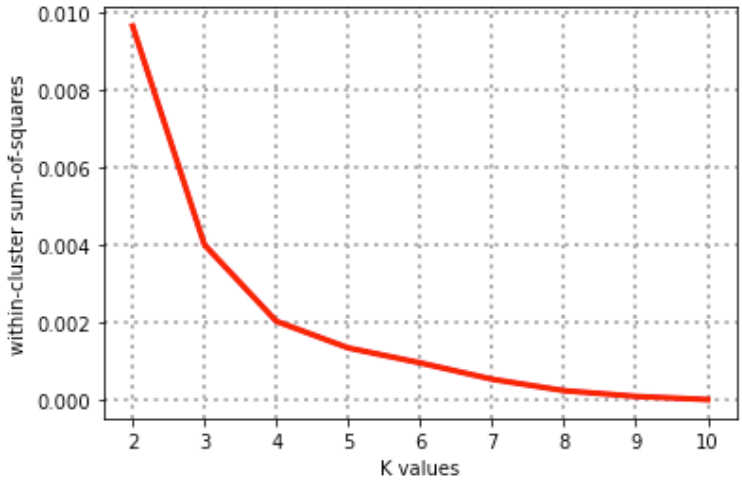

Figure 13 Cost Funciton of different K-Values

From the above graph it can determined that the least error is at value 8 after which it converges. Hence for simplicity we can choose 8 as the ideal k value for the number of clusters

3.4.5. Observations

With the k-value of 8. Following observations are made for 8 clusters of supermarket regions. A plot between Population Density and Supermarket is plotted as bar chart.

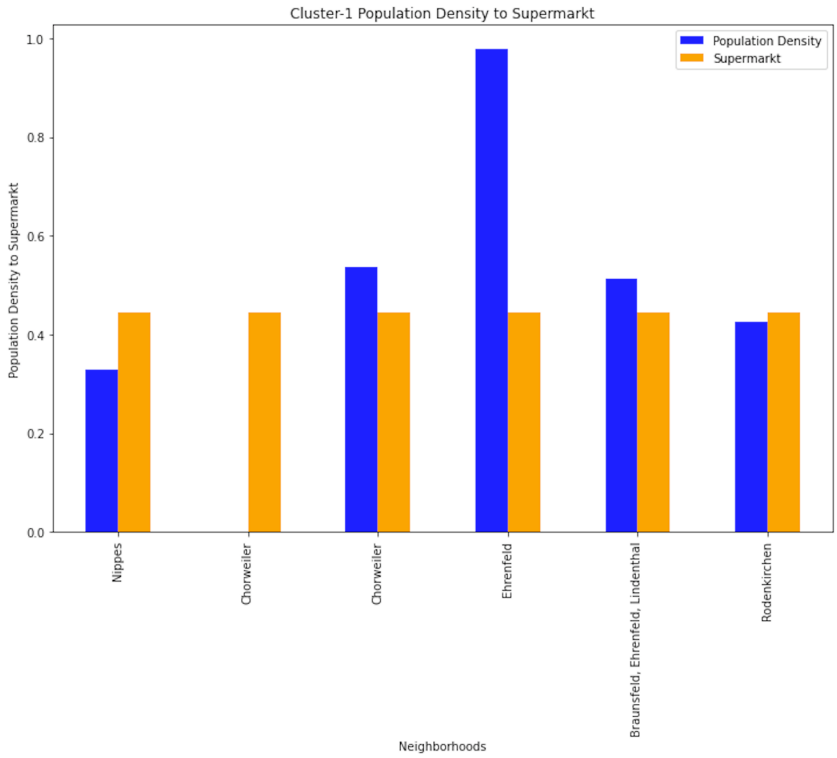

Figure 14 Cluster 1 Result

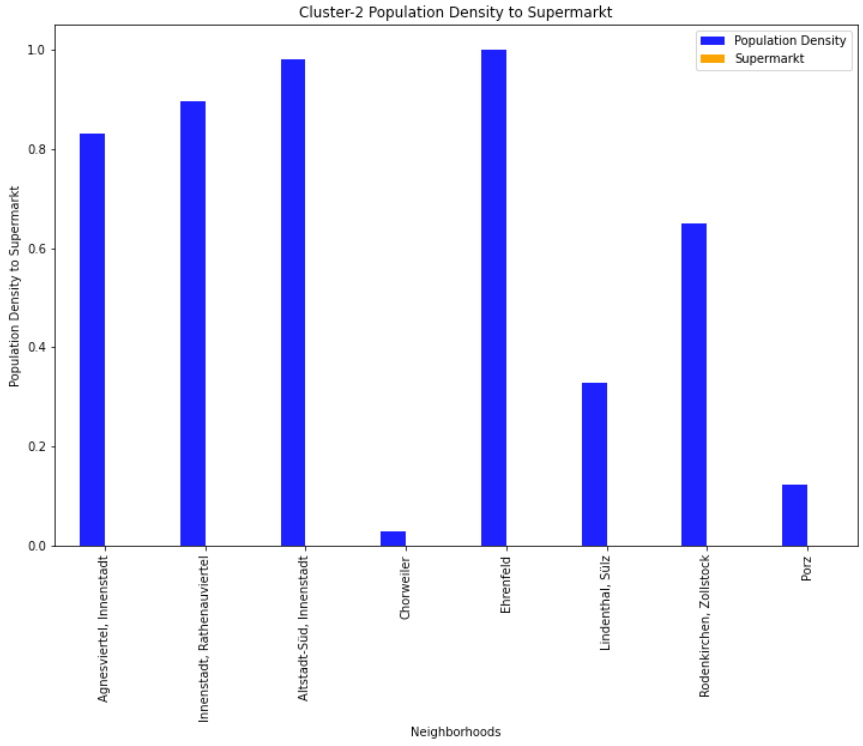

Figure 15 Cluster 2

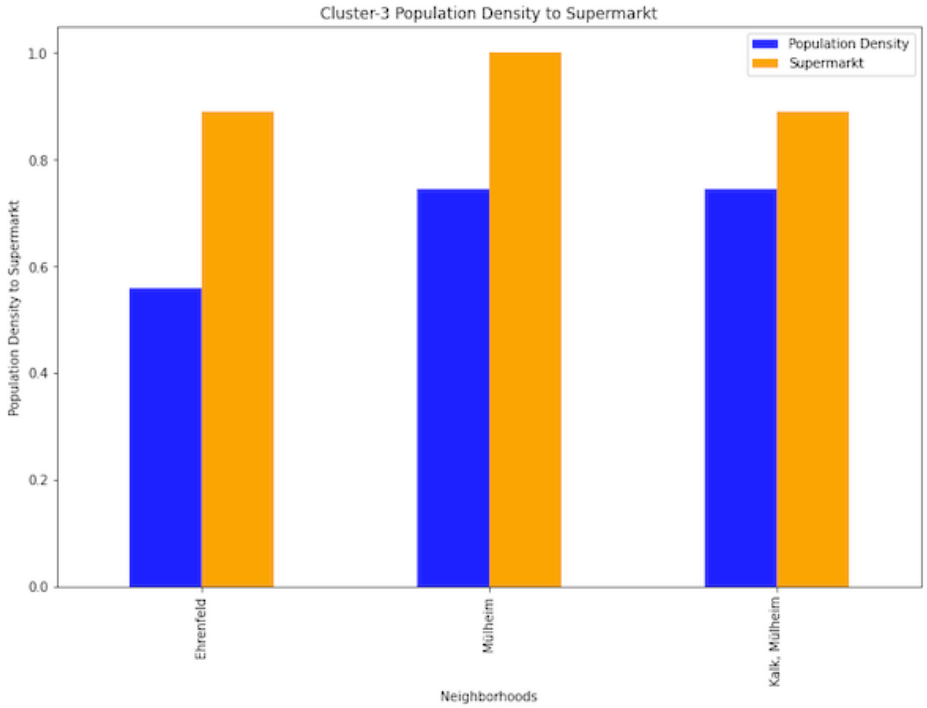

Figure 16 Cluster 3

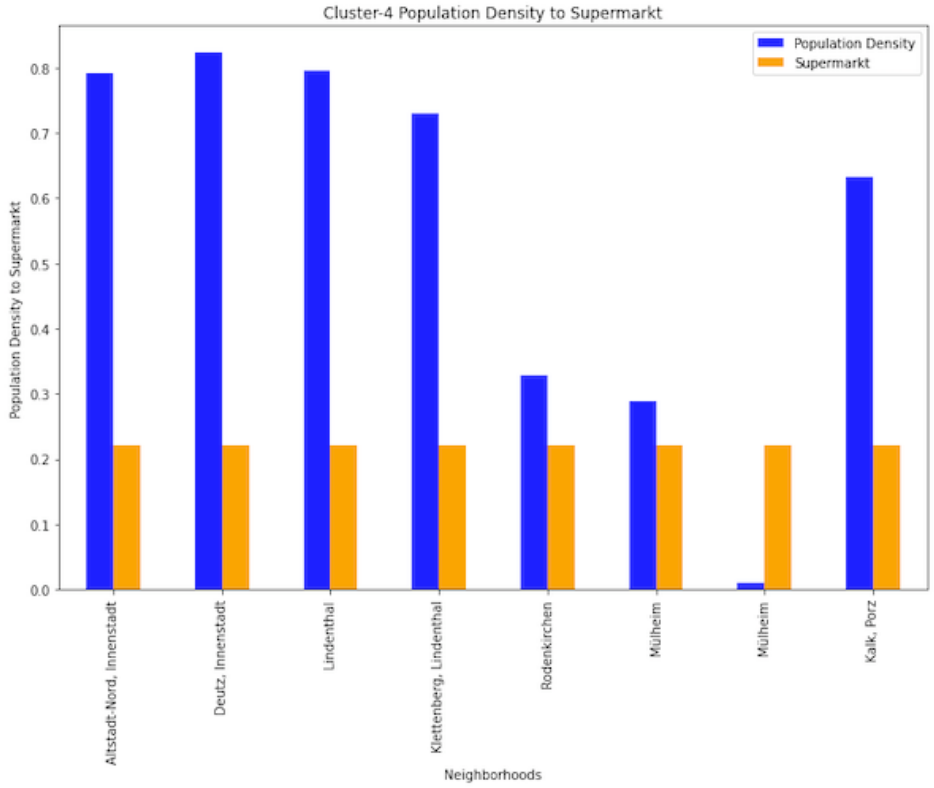

Figure 17 Cluster 4

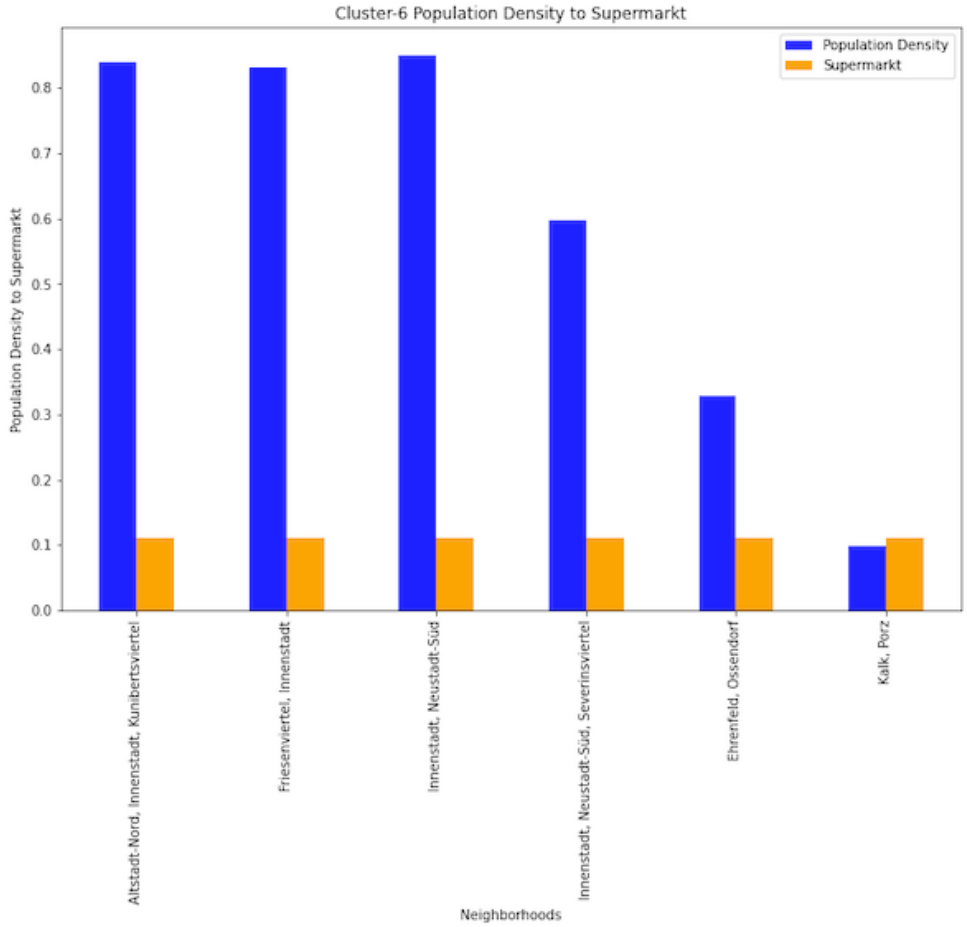

Figure 18 Cluster 6

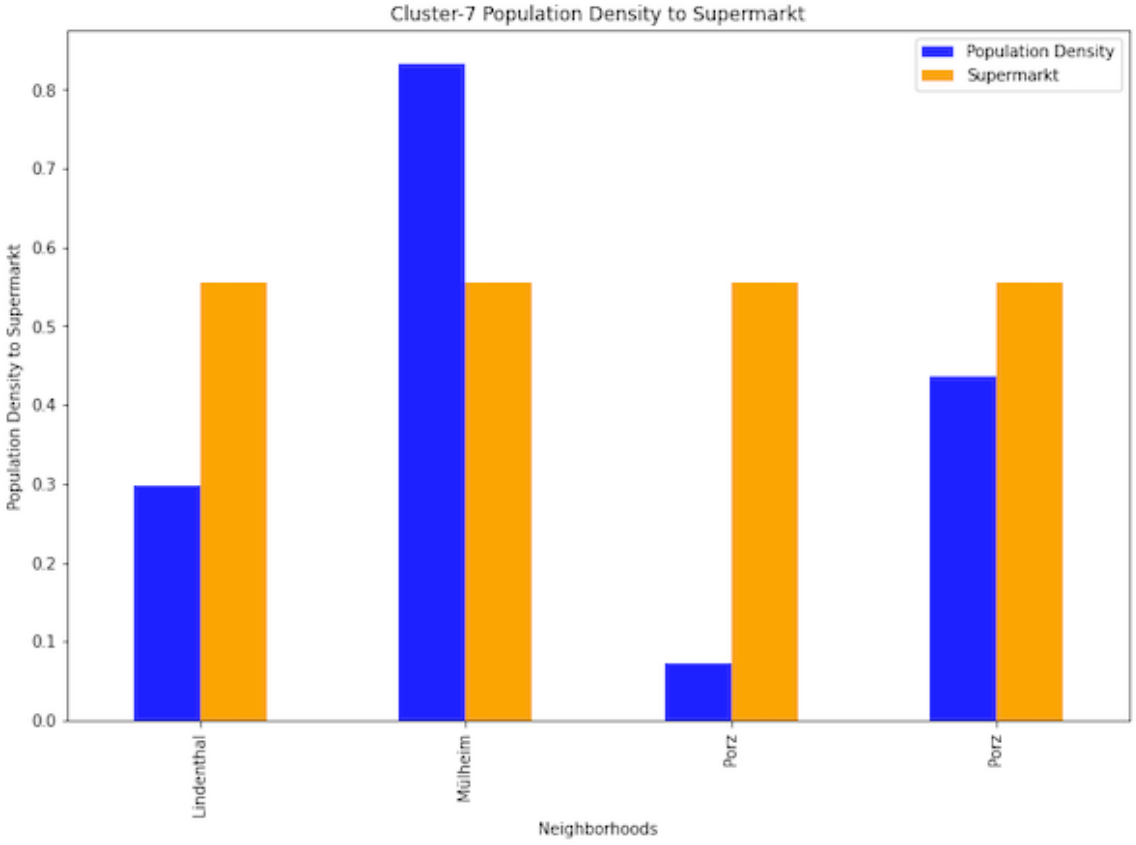

Figure 19 Cluster 7

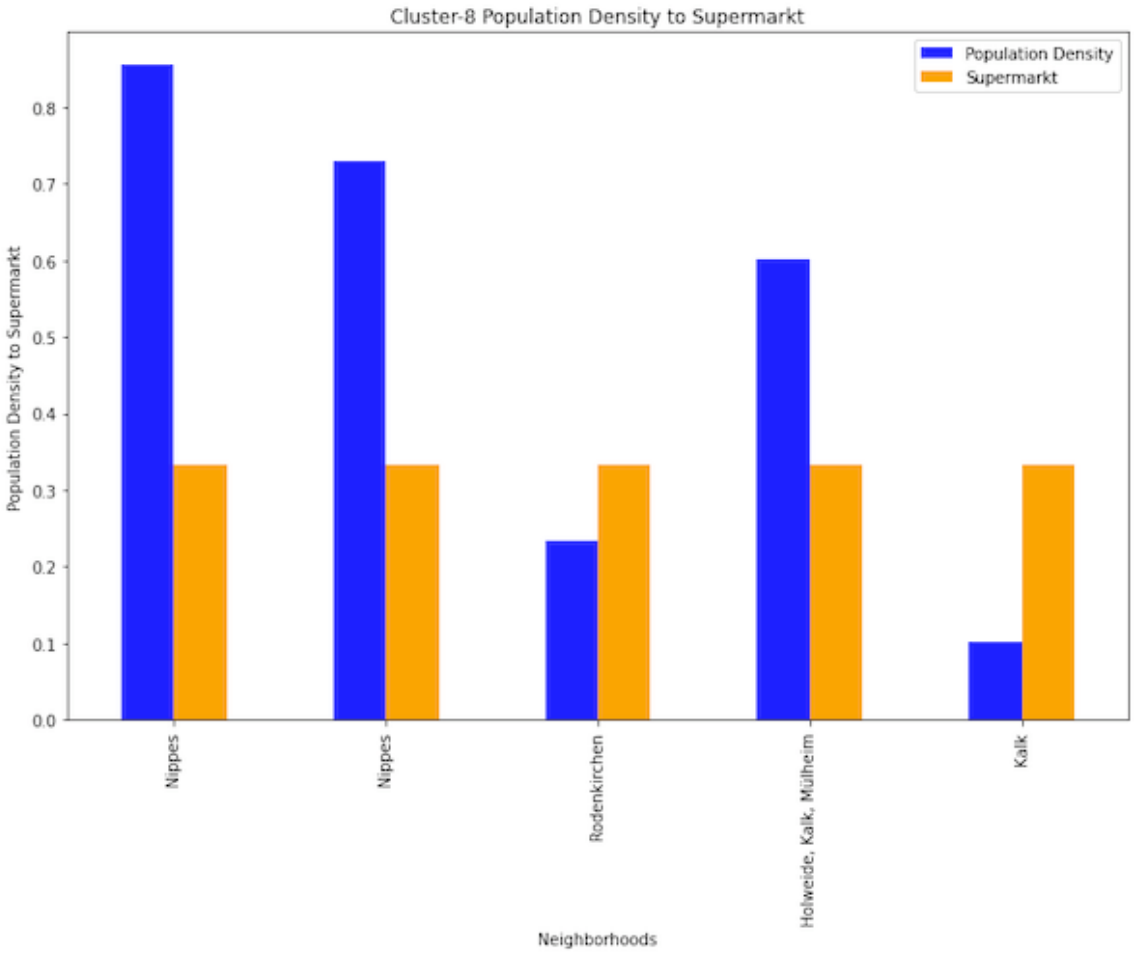

Figure 20 Cluster 8

4. Results

Examining all the clusters we can analyze results below

4.1. Cluster 1

There are 6 region in the Cluster 1, of which 50825 Ehrenfeld region can be considered as suggestion since it has very high population density with low supermarkets around it.

Considered: 50825 Ehrenfeld

4.2. Cluster 2

Cluster 2 is interesting which shows almost no supermarkets in the region. Forcing us to think this could be the best location for Supermarkets. However, for anyone who lives in cologne can guess that, that's because Cologne downtown region is always happening region which attract mostly tourist and are filled with other interesting places to see. Probably the Foursquare Limit of 100 venues did not show a minor portion of Supermarkets.

This Cluster shows us the limitation of the analysis as insufficient data from Foursquare. Also considering space availability and cost of setup in this neighborhood it can be ruled out that this cluster will not be good suggestion for our business.

4.3. Cluster 3

This cluster shows 3 regions, which already has higher the amount of Supermarkets compared to population density. This can be ruled out.

Considered: None

4.4. Cluster 4

This cluster gives us some usefull results, where the population density is high, but the number of supermarkets is low. There are in total 8 regions in cluster. Some of these can be considered as potential region for our analysis.

Considered:

I. 50667 Altstadt Nord, Innenstadt
II. 50679 Deutz, Innenstadt
III. 50931 Lindenthal
IV. 50939 Klettenberg, Lindenthal
V. 51105 Kalk Porz

4.5. Cluster 5

Cluster 5 has five regions, which already has high number of supermarkets compared to the population density. So this cluster can be ignored.

Considered: None

4.6. Cluster 6

Cluster 6 has six regions which can be potential locations, because a very high population density region has few supermarkets. Some of these can be considered.

Considered:

I. 50668 Altstadt-Nord Innenstadt
II. 50672 Friesenviertel
III. 50677 Innenstadt Neustadt Süd

4.7. Cluster 7

Cluster 7 has four regions. Three of these have high supermarket compared.

Considered:

I. 51065 Müllheim

4.8. Cluster 8

Cluster 8 has 5 regions of which two can be considered as choice.

Considered:

II. 50737 Nippes
III. 50739 Nippes

5. Discussion

This project has set out and brought in interesting analytical insights with the available data. However yes there are always scope of improve as with any machine learning models. Here are few that can be discussed.

This analysis result is performed with radius of 1600m with given latitude and longitude of the regions. Inspite of this radius there are many regions which are not covered a simple illustration of this is shown in the Figure. These overlaps have been taken care by ensuring duplicate entries are removed. Also there are more regions that can be covered by adopting some better algorithms.

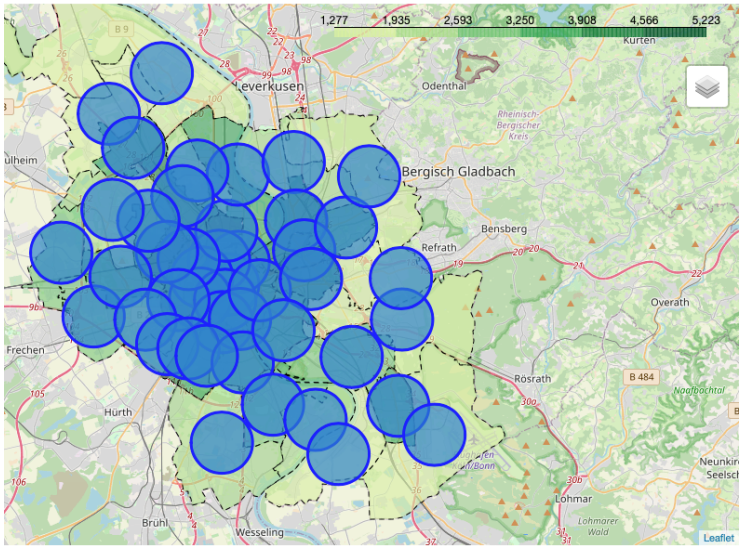

Figure 21 1600m Radius circles

Also the Foursquare API venue limit factor of 100 venues could have hidden the supermarkets. Apart from these

Further more, better data on several other factors, like Availabilty of space, prices can be used to improve the model prediction. Further demographic data can be further drilled down to average income of the population in the given region. This can help in further focusing the product suggestion. Of course having good team at supermarket making customer find what the best quality with cost product are other important factors.

6. Conclusion

Regarding the business problem, Where can be the next best location for Supermarket in Cologne. The answer is of course yes and there are 12 regions which are having high population density but less number of supermarkets, grocery stores.
Cluster 4, Cluster 6 and Cluster 8 are all good choices from the results of this study.