Retail Analysis in Berlin

Juan Acevedo

1. Introduction

1.1 Background

With more than 3.5 million inhabitants, Berlin is one of the most important markets for Germany's retailers. In grocery stores, four big players dominate the market: Edeka, Schwarz group (Lidl and Kaufland), Rewe, and Aldi (Nord & Sud). Based on the size, most of the stores can be categorized either in supermarkets or grocery stores. Usually, these players are looking for new locations to expand and increase revenues. Therefore, retailers need to identify new business opportunities to keep growing and improving their financial statements.

1.2 Problem

Data could contribute to determining which borough is oversaturated by competitors and which one needs more convenience stores. Based on the number of residents in each borough and the number of grocery stores, it could be possible to determine where retailers can open a new store. Consequently, this project aims to determine where retailers can open new stores in the city of Berlin.

1.3 Objectives

- Cluster retail branches using Kmean
- Identify the most popular retailer in the city of Berlin
- Identify zones very competition is low and new stores can be open

1.4 Data

Supermarkets data from Foursquare and demographic data from Wikipedia related to the Berlin population will be used to identify market opportunities. Additionally, the API Foursquare will be used to get the retailers per borough. Overall, this project aims to calculate the number of stores per 100.000 habitats in a specific ratio to determine market opportunities.

2. Methodology

2.1 Preprocessing

To accomplish the proposed objectives, data from Wikipedia was used to identify all boroughs and their population in the city of Berlin. This data was stored in an excel file and then imported, cleaned, saved as df_berlin using pandas (Image 1). Using ArcGis, the boroughs from df_Berlin were used to obtain their coordinates and saved in a new data frame. Next, the coordinates were merged with df_berlin (image 2).

	Borough	Quarter	Population 2019
0	Berlin-Mitte [Berlin-Center]	Gesundbrunnen	95175
1	Berlin-Mitte [Berlin-Center]	Hansaviertel	5926
2	Berlin-Mitte [Berlin-Center]	Mitte	102465
3	Berlin-Mitte [Berlin-Center]	Moabit	80495
4	Berlin-Mitte [Berlin-Center]	Tiergarten	14881

Figure 1 - Data from Berlin first five rows

	Borough	Quarter	Population 2019	Latitude	Longitude
0	Berlin-Mitte [Berlin-Center]	Gesundbrunnen	95175	52.55619	13.37710
1	Berlin-Mitte [Berlin-Center]	Hansaviertel	5926	52.51679	13.33835
2	Berlin-Mitte [Berlin-Center]	Mitte	102465	52.52119	13.42414
3	Berlin-Mitte [Berlin-Center]	Moabit	80495	52.52570	13.34005
4	Berlin-Mitte [Berlin-Center]	Tiergarten	14881	52.50993	13.36393

Figure 2- Data from Berlin first five rows

2.2 Data from supermarkets

The next step was to obtain the list of supermarkets from Foursquare's API, using category id supermarkets as a filter. The list was then processed to identify the most common supermarket per Neighborhood (figure 3) and then cluster using Kmean into 5 clusters to identify similarities inside the city.

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Adlershof	REWE	ALDI NORD	Netto Filiale	Lidl	NETTO
1	Alt-Hohenschönhausen	REWE	Netto Marken-Discount	Lidl	ALDI NORD	Netto Filiale
2	Alt-Treptow	Lidl	ALDI NORD	REWE	Netto Marken-Discount	PENNY
3	Altglienicke	ALDI NORD	Lidl	Netto Filiale	REWE	Netto Marken-Discount
4	Baumschulenweg	Lidl	ALDI NORD	Netto Marken-Discount	REWE	NETTO

Figure 3- Most common supermarket per quarter

Moreover, Foursquare's data was used to establish the most popular supermarket per quarter and then displayed using Folium. Similarly, using the demographic data from Wikipedia, it was possible to calculate the number of clients per store in each quarter(figure 4).

	Borough	Quarter	Population 2019	Latitude	Longitude	Number of Retails	Customer per Store
0	Berlin-Mitte [Berlin-Center]	Gesundbrunnen	95175	52.55619	13.377100	81	1175
1	Berlin-Mitte [Berlin-Center]	Hansaviertel	5926	52.51679	13.338350	94	63
2	Berlin-Mitte [Berlin-Center]	Mitte	102465	52.52119	13.424140	100	1024
3	Berlin-Mitte [Berlin-Center]	Moabit	80495	52.52570	13.340050	75	1073
4	Berlin-Mitte [Berlin-Center]	Tiergarten	14881	52.50993	13.363930	100	148
5	Berlin-Mitte [Berlin-Center]	Wedding	86806	52.54781	13.354730	79	1098

Figure 4- Data Customer per store

3. Results

The cluster map shows how the supermarkets are clustered in the city of Berlin. The most common cluster is blue and green around the city center, while outside the city center cluster red, orange and Purple are more common. The map indicates that retailers' distribution in the city center is similar and varies on the edges of the city.

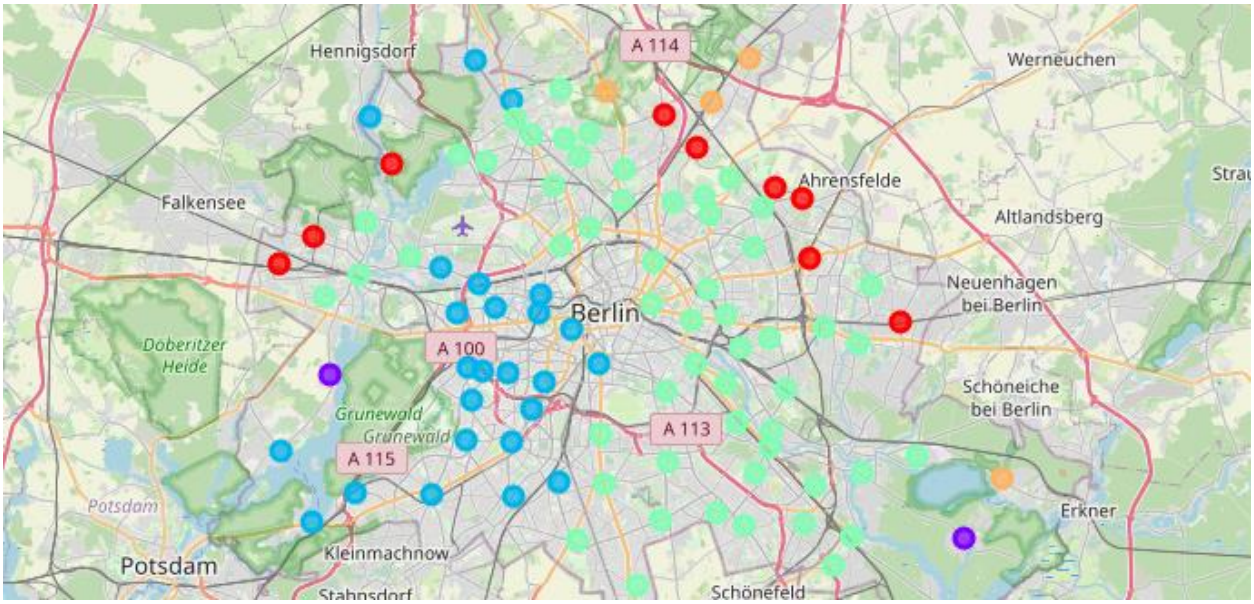

Figure 5- Data Clusters

The following maps illustrate the most common retail per quarter by color. The distribution of colors was assigned as follows: blue = Lidl, green = Aldi, red = Rewe, yellow = Edeka, grey = others. As illustrated, each part of the city is dominated by different retailers, Lidl dominates the north and south, Aldi the southwest and southeast, and Rewe the northeast. The most diverse area of the city is the northwest, where all colors can be seen.

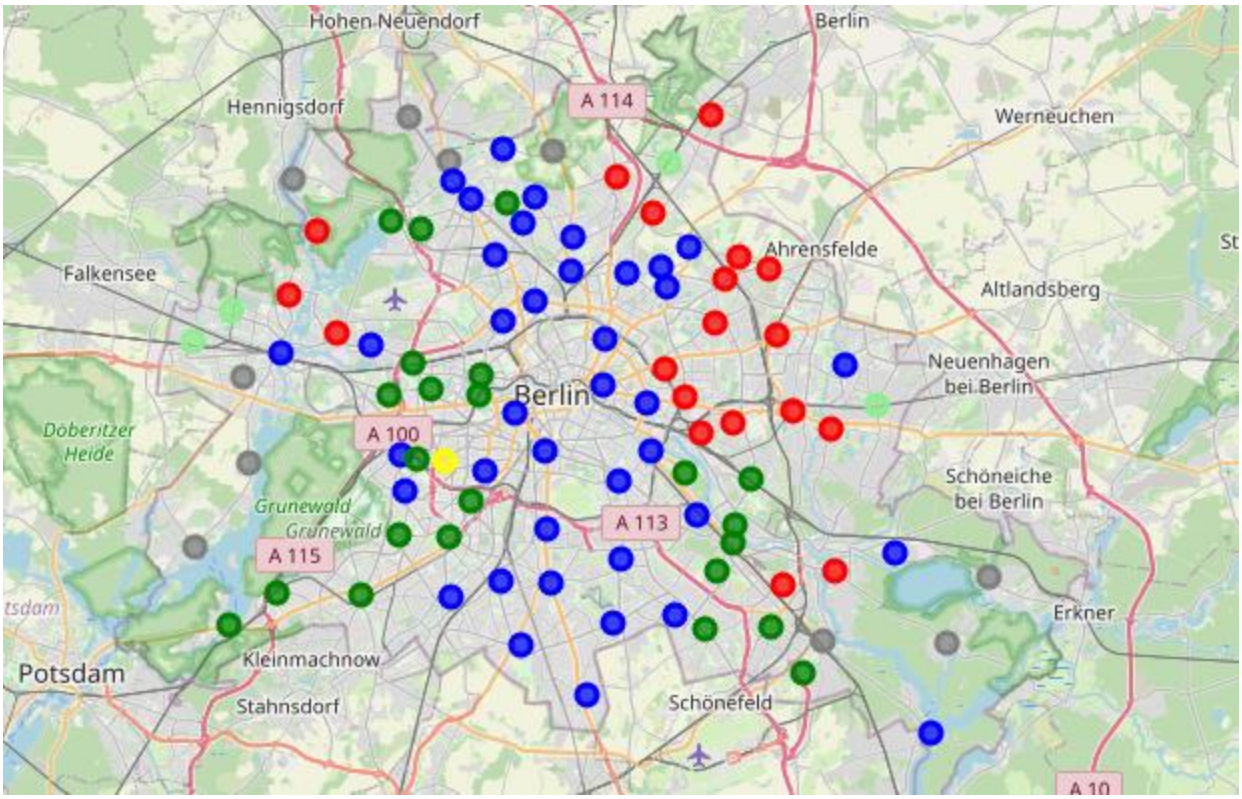

Figure 6- Most popular retaile by quarte - Blue = Lidl, Green = Aldi, Red = Rewe, Yellow = Edeka, Gray = Others

The final map illustrated the number of customers per store. The number of customers per store was calculated dividing the number of inhabitants in the quarter by the number of stores in the same quarter. Then the results were colored as

follows: red = less than 50 customers per store, orange = between 51 and 200 customers per store, yellow = between 201 and 1000 customers per store, green = between 1001 and 2001 customers per store, and purple = more than 2000 customers per store. The red color and orange color indicates an oversaturated market where stores may not be profitable. The yellow stores where the number is ok but with a low number of customers. Green indicates areas where stores must be profitable due to the high number of customers and purple new business opportunities where stores can be open.

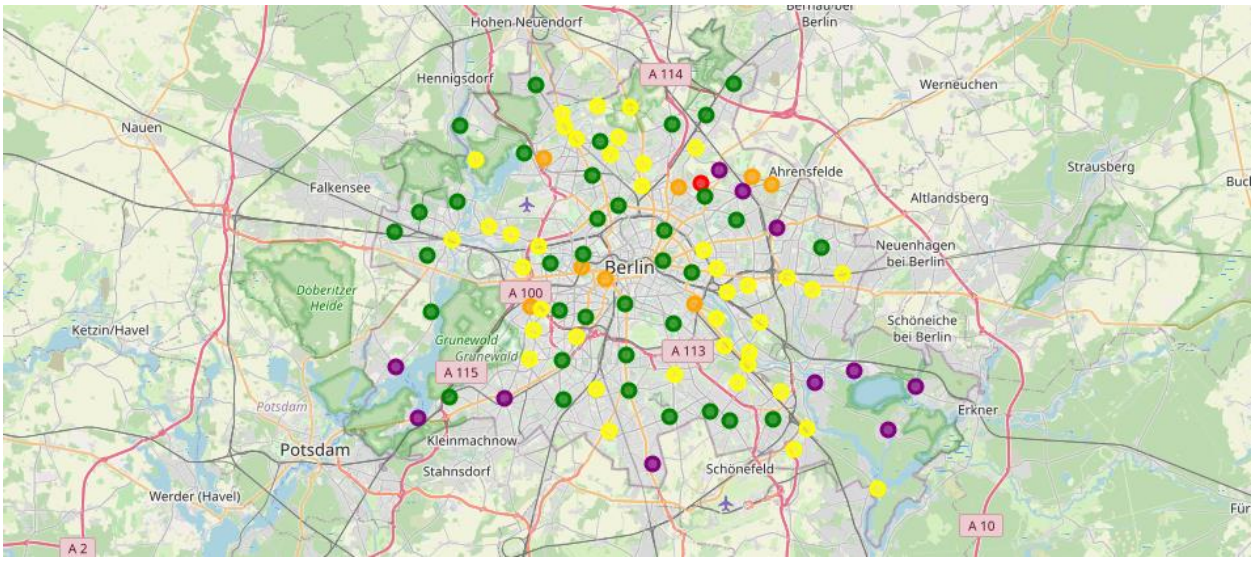

Figure 7- Business Opportunities – Red =Extremly high competition, Orange and Yellow = High competition, Green = Good level of competition, Purple = new business opportunity

4. Conclusion

In this project, it was possible to determine that the most the distribution of retail branches is very similar in the city center and varies in the edges of the city. Additionally, retailers dominate specific areas of the city except for the northwest, where several players dominate each quarter. Additionally, it was possible to determine that new business opportunities can be found in near the borders of the city.