ANALYSIS OF BERLIN

Ijlal Muneeb

IBM DATA SCIENCE CAPSTONE PROJECT

07/01/21

Table of Contents

1. INTRODUCTION

1.1. BACKGROUND

Berlin is the capital of Germany and is a fast-paced city full of digital innovation and entrepreneurship. It is to Germany what Silicon Valley is to San Francisco. In this modern-day digital age, in Germany, Berlin attracts the top minds in technology and innovation. So, it provides a lucrative opportunity for tech enthusiasts and investors alike.

1.2. PROBLEM

A hypothetical investor looking to invest in an up-and-coming business, is looking for an area for business activities and to set up office with the following attributes:

- Buzzing with people, for visibility and accessibility
- Safety
- Close to the city center
- Social activities

1.3. INTEREST

Although, this project offers an overview of the city from a business point of view, it still provides significant relevant information for people looking for an area to live in, to buy a house, real estate in, etc. This project provides the customer with key insights about Berlin with a direct what-to-expect attitude.

2. DATA

2.1. DATA SOURCES

With these attributes in mind, we'll try to carry out a generalized analysis for Berlin which gives our client an overview of Berlin and helps him make an informed decision to choose a place for his office. Since we have certain amenities and social activities to consider, we will leverage the Foursquare API to our advantage to get local venues available in the different neighborhoods of Berlin.

We'll be using a variety of Data sources for this analysis. Their general use purpose is mentioned along with their source below.

1. Github Repository consisting of GeoData files for Berlin:
2. (https://github.com/funkeinteraktiv/Berlin-Geodaten)
3. Boroughs and neighborhoods of Berlin:
4. (https://en.wikipedia.org/wiki/Boroughs_and_neighborhoods_of_Berlin#Boroughs)
5. For collecting venues of Berlin: (https://foursquare.com)
6. Crime Dataset of Berlin from Kaggle: (https://https://www.kaggle.com/danilzyryanov/crime-in-berlin-2012-2019)

2.2. DATA CLEANING

2.2.1 Wikipedia Tables

I used the Wikipedia page on Boroughs and Neighborhoods in Berlin to extract Data Tables consisting of the Borough names, populations, area, and population density. After scraping the tables with python library BeautifulSoup, I put it in a pandas Dataframe. Then, using Geopy I collected the geographical coordinates of all the localities in Berlin and combined them in the Boroughs dataframe.

2.2.2. Berlin GeoData

The Geodata files for Berlin were acquired from the aforementioned github repo. It allowed me to plot a map of Berlin and show various attributes on it which will be shown later in the analysis section. One critical point in using this particular file was that the 'key-on' feature for properties.name is 'name' instead of 'Borough' or 'Neighborhood'. So, I had to rename the index column from 'Borough' to 'Name' in order to key the geodata together on the folium map.

2.2.3. Berlin Crime Data

The crime data acquired from Kaggle spanned from 2012 to 2019. But to save on speed and memory power, I decided to only use the data from 2019, so I split this data into a new dataframe.

2.3. FEATURE SELECTION

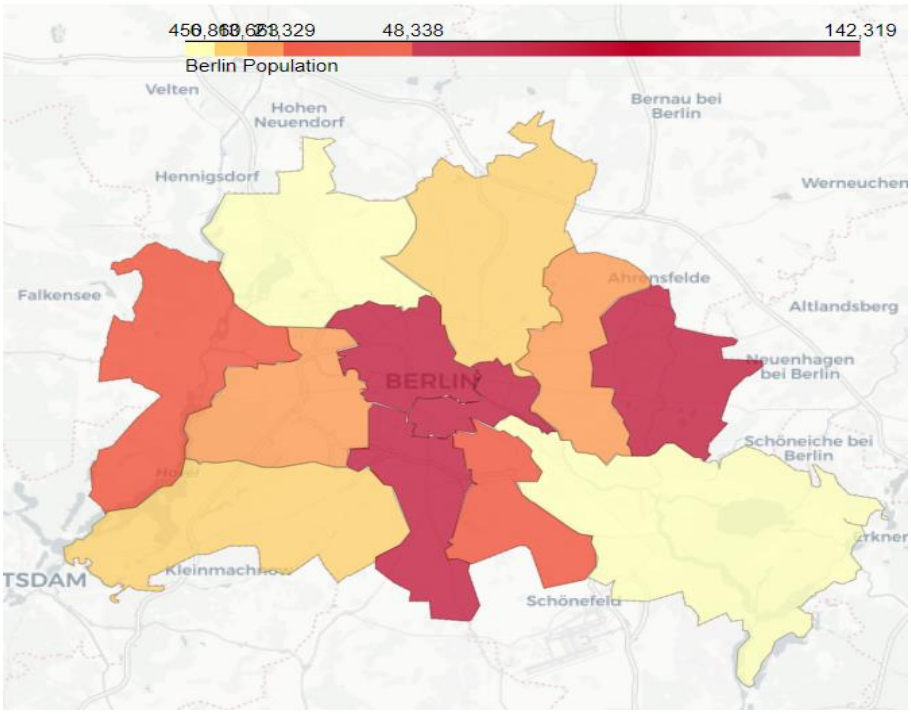
For this analysis, we take a look at the population metrics, and the crime metrics for every neighbourhood in Berlin. Later on, we use the Neighbourhood locations and onehot encode them so they can be used for Machine Learning.

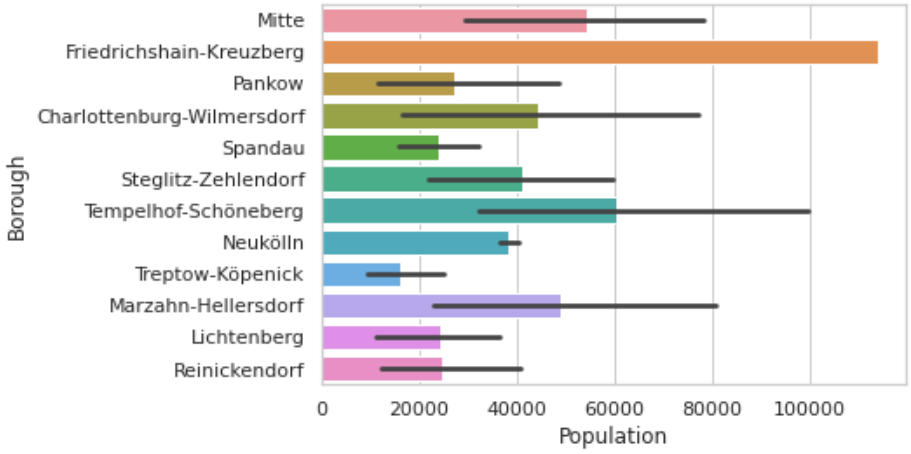
3. Methodology

3.1. EXPLORATORY DATA ANALYSIS

3.1.1 Geo Statistics

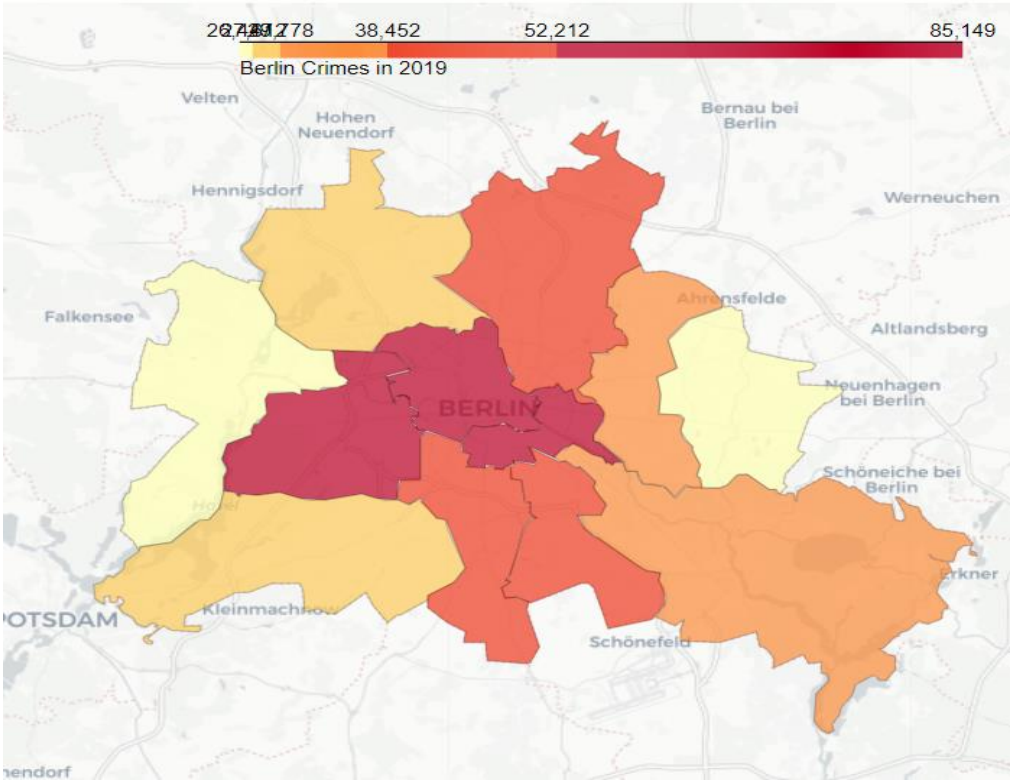
To start off, I first analyzed the geography of Berlin. It consists of 12 Boroughs with 86 total Neighborhoods. The population metrics of the boroughs can be seen in the figures below:

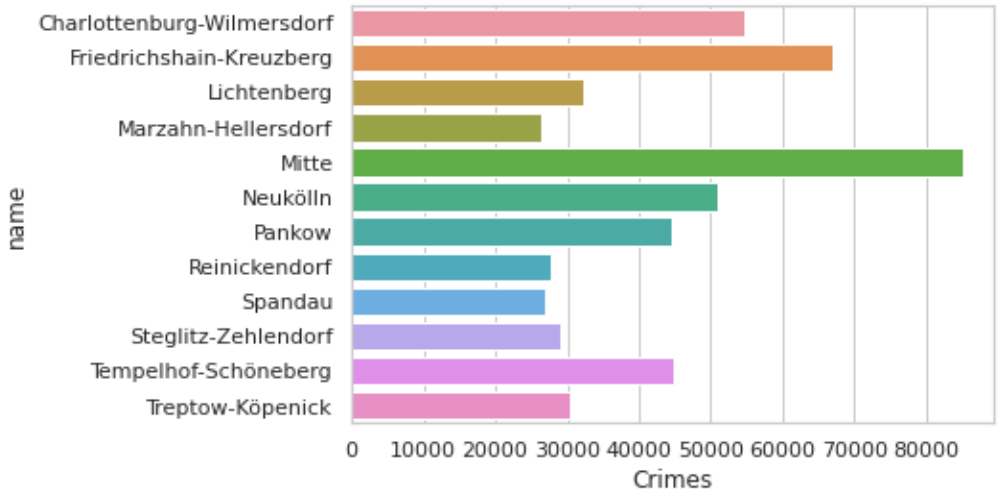




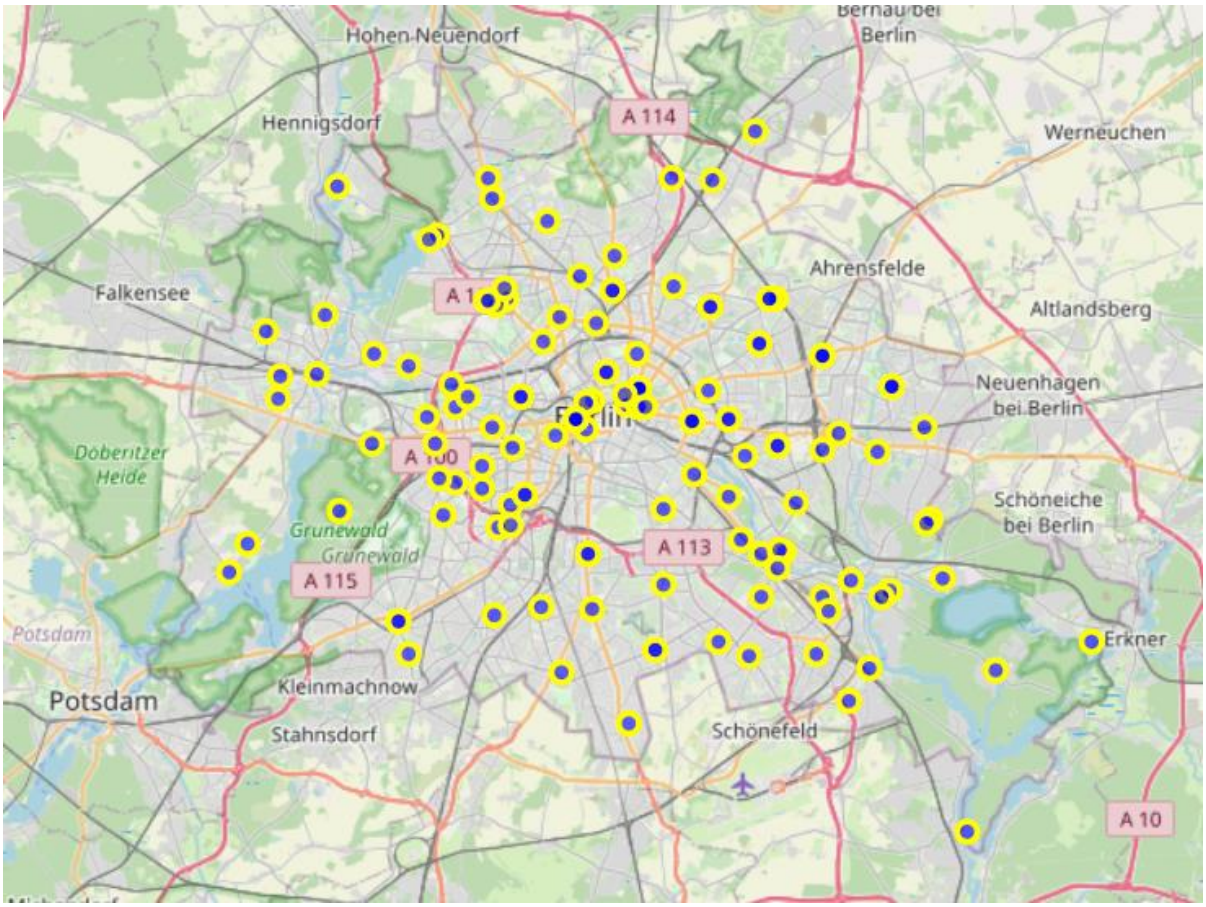
So, the largest amount of people reside in Friedrichshain-Kreuzberg, Mitte, and Tempelhof respectively.

Now, looking at a similar figure of crime data we see that there seems to be a general relation of population to crime. Larger the population, higher the crime rate. The highest recorded crime rates belong to Mitte, and Friedrichshain-Kreuzberg, two of the most populated boroughs of Berlin.


38,452
52,212
85,149
Berlin Crimes in 2019
Velten
Hohen Neuendorf
Bernau bei Berlin
Hennigsdorf
Werneuchen
Falkensee
Altlandsberg
BERLIN
Neuenhagen bei Berlin
Schöneiche bei Berlin
Kleinmachnow
Schönefeld


Charlottenburg-Wilmersdorf
Friedrichshain-Kreuzberg
Lichtenberg
Marzahn-Hellersdorf
Mitte
Neukölln
Pankow
Reinickendorf
Spandau
Steglitz-Zehlendorf
Tempelhof-Schöneberg
Treptow-Köpenick
name
0
10000
20000
30000
40000
50000
60000
70000
80000
Crimes

Having formed a general idea of these statistics, I then plotted the crime data on a map of Berlin as dots. This shows us whether the crime is spread out all over the city or are there any crime hotspots?



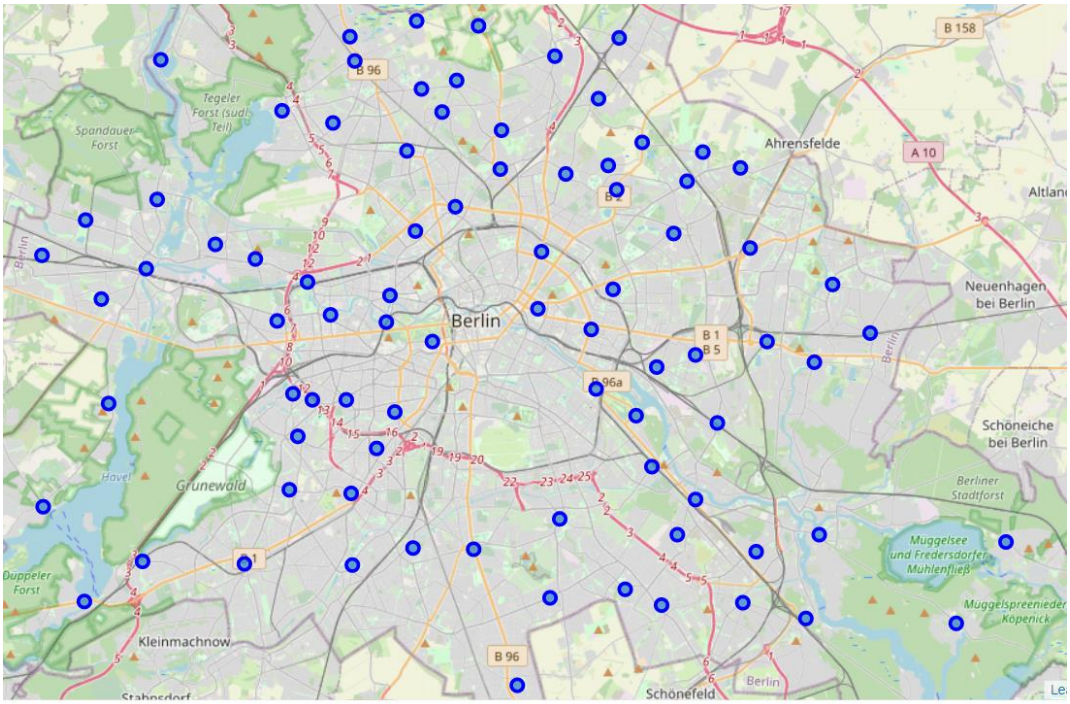
As we can see, apart from the epicenter of the city, the rest of the crime data is pretty evenly spread. The center of the city is generally prone to crime due to seeing high numbers of people every day such as travelers and tourists.

3.1.2. CLUSTERING

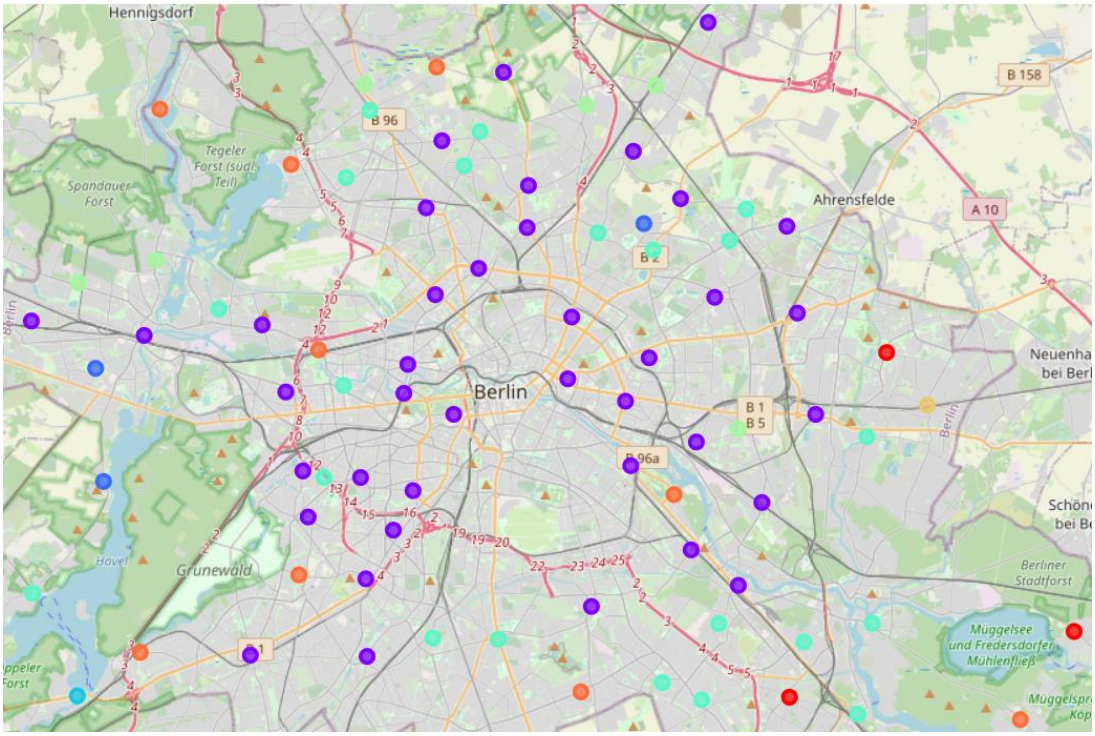
Moving on to the next phase of the analysis. I used the Foursquare API to get venue data for all the neighborhoods in Berlin. After performing onehot encoding of all the categorical variables, i.e. Venue names and categories, I then proceeded to group similar neighborhoods using the KMeans algorithm. Then I combined the clusters into a pandas dataframe that shows the top ten venues of every neighborhood.

This map provides our client a simple view to decide on selecting a neighborhood for his business based on other factors discussed before. One final factor is the availability and accessibility of social amenities.

Below we see all the neighborhoods plotted on a map.



The KMeans algorithm takes care of that and provides us with color coded clusters of neighborhoods comprising of a similar surrounding social amenities landscape. The map of clusters is shown below:



There are 8 clusters in total. Color-coded for similarities among them.

Given below are a portion of each of the clusters formed:

CLUSTER 1:

Neighbourhood	Latitude	Longitude	Cluster Labels	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
Altglienicke	52.42006	13.53969	0	Supermarket	Bowling Alley	Discount Store	Yoga Studio	Electronics Store	Food & Drink Shop	Flea Market	Fishing Store	Fish Market	Financial or Legal Service
Rahnsdorf	52.44093	13.68891	0	Supermarket	Tram Station	Café	Yoga Studio	Eastern European Restaurant	Flea Market	Fishing Store	Fish Market	Financial or Legal Service	Fast Food Restaurant
Hellersdorf	52.52945	13.59068	0	Supermarket	Bowling Alley	Bus Stop	Drugstore	Cable Car	Yoga Studio	Falafel Restaurant	Food & Drink Shop	Flea Market	Fishing Store

CLUSTER 2:

Neighbourhood	Latitude	Longitude	Cluster Labels	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	10t Co
Stadtrandsiedlung Malchow	52.57038	13.46342	2	Playground	Bus Stop	Eastern European Restaurant	Food & Drink Shop	Flea Market	Fishing Store	Fish Market	Financial or Legal Service	Fast Food Restaurant	F
Gatow	52.48884	13.18070	2	Bus Stop	Hotel	Italian Restaurant	Supermarket	Harbor / Marina	Eastern European Restaurant	Flea Market	Fishing Store	Fish Market	Fi c
Wilhelmstadt	52.52462	13.17707	2	Bus Stop	Supermarket	Tree	Yoga Studio	Eastern European Restaurant	Flea Market	Fishing Store	Fish Market	Financial or Legal Service	Fa Res

CLUSTER 3:

Neighbourhood	Latitude	Longitude	Cluster Labels	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
Mitte	52.52119	13.42414	1	Nightclub	Supermarket	Plaza	Post Office	Bike Shop	Restaurant	Fountain	Furniture / Home Store
Moabit	52.52570	13.34005	1	Bar	Café	Ice Cream Shop	Organic Grocery	Vietnamese Restaurant	Gym / Fitness Center	Doner Restaurant	Drugstore
Hansaviertel	52.51679	13.33835	1	Café	Hotel	Bakery	Hookah Bar	Plaza	Farmers Market	Flea Market	Roof Deck
Tiergarten	52.50993	13.36393	1	Art Museum	Concert Hall	Breakfast Spot	Memorial Site	Hotel	Hotel Bar	Museum	Plaza
Wedding	52.54781	13.35473	1	Café	Drugstore	Burger Joint	Turkish Restaurant	Doner Restaurant	Pizza Place	Fast Food Restaurant	Supermarket
Gesundbrunnen	52.55619	13.37710	1	Doner Restaurant	Bakery	Turkish Restaurant	Bar	Soccer Field	Metro Station	Mediterranean Restaurant	Shopping Mall
Friedrichshain	52.51402	13.45403	1	Café	Pub	Middle Eastern Restaurant	Thai Restaurant	Italian Restaurant	Bakery	Doner Restaurant	Thrift / Vintage Store

CLUSTER 4

Neighbourhood	Latitude	Longitude	Cluster Labels	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
Wannsee	52.42046	13.16709	3	Harbor / Marina	Supermarket	Yoga Studio	Eastern European Restaurant	Food & Drink Shop	Flea Market	Fishing Store	Fish Market	Financial or Legal Service	Fast Food Restaurant

CLUSTER 5:

Neighbourhood	Latitude	Longitude	Cluster Labels	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
Weißensee	52.56203	13.46839	4	Supermarket	Music Venue	Bus Stop	Laser Tag	Bank	Yoga Studio	Empanada Restaurant	Food & Drink Shop
Heinersdorf	52.56770	13.43980	4	Electronics Store	Currywurst Joint	Supermarket	Gym / Fitness Center	Pet Store	Tram Station	Farm	Empanada Restaurant
Rosenthal	52.59977	13.37776	4	Tram Station	Hotel	Supermarket	German Restaurant	Eastern European Restaurant	Flea Market	Fishing Store	Fish Market
Wilhelmsruh	52.58887	13.36949	4	Asian Restaurant	Post Office	Supermarket	Bakery	Flea Market	Fishing Store	Fish Market	Financial or Legal Service
Charlottenburg	52.51915	13.30639	4	Supermarket	Bakery	Italian Restaurant	Middle Eastern Restaurant	Brewery	Beer Garden	Café	Russian Restaurant
Halensee	52.49005	13.29602	4	Bakery	Supermarket	Trattoria/Osteria	Dessert Shop	Business Service	Bar	Italian Restaurant	German Restaurant

CLUSTER 6:

Neighbourhood	Latitude	Longitude	Cluster Labels	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
Karow	52.61418	13.46976	5	Supermarket	Light Rail Station	Doner Restaurant	Drugstore	Eastern European Restaurant	Food & Drink Shop	Flea Market	Fishing Store	Fish Market	Financial or Legal Service
Französisch Buchholz	52.60822	13.43368	5	Supermarket	Garden Center	Drugstore	Yoga Studio	Electronics Store	Food & Drink Shop	Flea Market	Fishing Store	Fish Market	Financial or Legal Service
Hakenfelde	52.55877	13.20831	5	Supermarket	Bus Stop	Turkish Restaurant	Drugstore	Yoga Studio	Eastern European Restaurant	Flea Market	Fishing Store	Fish Market	Financial or Legal Service
Falkenhagener Feld	52.55191	13.16802	5	Supermarket	Snack Place	Drugstore	Yoga Studio	Eastern European Restaurant	Flea Market	Fishing Store	Fish Market	Financial or Legal Service	Fast Food Restaurant
Friedrichsfelde	52.50527	13.51326	5	Supermarket	Drugstore	Stadium	Tram Station	Gastropub	Yoga Studio	Electronics Store	Flea Market	Fishing Store	Fish Market
Hermsdorf	52.61490	13.31783	5	Supermarket	Beer Garden	Drugstore	Yoga Studio	Empanada Restaurant	Food Court	Food & Drink Shop	Flea Market	Fishing Store	Fish Market

CLUSTER 7:

Neighbourhood	Latitude	Longitude	Cluster Labels	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
Mahlsdorf	52.51297	13.61185	6	Platform	Greek Restaurant	Bakery	Food & Drink Shop	Flea Market	Fishing Store	Fish Market	Financial or Legal Service	Fast Food Restaurant	Farmers Market

CLUSTER 8:

Neighbourhood	Latitude	Longitude	Cluster Labels	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
Charlottenburg-Nord	52.53048	13.29371	7	Platform	Metro Station	Museum	Boat or Ferry	Deli / Bodega	Falafel Restaurant	Food & Drink Shop	Currywurst Joint	Flea Market
Dahlem	52.45874	13.28344	7	German Restaurant	Pizza Place	Bus Stop	Farm	Metro Station	Gym	Garden Center	Yoga Studio	Fishing Store
Nikolassee	52.43411	13.19973	7	Hotel	Light Rail Station	Intersection	German Restaurant	Tennis Court	Food & Drink Shop	Flea Market	Fishing Store	Fish Market
Lichtenrade	52.39159	13.41247	7	Asian Restaurant	Bus Stop	Supermarket	Restaurant	Café	Yoga Studio	Electronics Store	Flea Market	Fishing Store
Buckow	52.42168	13.43099	7	Discount Store	Diner	Bakery	Bus Stop	Restaurant	German Restaurant	Gastropub	Pizza Place	Farmers Market
Plänterwald	52.48439	13.47957	7	Beer Garden	Boat Rental	Playground	Boat or Ferry	Seafood Restaurant	Restaurant	Falafel Restaurant	Farm	Farmers Market
Müggelheim	52.41300	13.66050	7	American Restaurant	Bus Stop	Café	Harbor / Marina	German Restaurant	Yoga Studio	Electronics Store	Flea Market	Fishing Store

4. RESULTS

The clustering of neighborhood has provided us insight into many things. One being that most of the neighborhoods in Berlin have a very similar atmosphere in terms of social amenities. Secondly, the crime data coincides with the amount of population in that area. This means that crime in Berlin is under control and no areas exists that could be considered particularly dangerous or crime hotspots.

5. DISCUSSIONS

Having completed this analysis, I have to say that there is plenty of room to add in other factors to the research using new data sources such as, schooling locations, hospital data, traffic flow, etc. One caveat in this analysis is in the fact that the population metrics used are rather old while the crime data is new. So, it does not provide a 100% accurate picture of the population vs crime metric.

6. CONCLUSION

In this study, I analyzed the population of Berlin, segmented the geography of Berlin into Boroughs and neighborhoods. I explored the neighborhoods using the Foursquare API to get venue data. I compared the various types of data such as crime, population, social amenities etc., to paint an overview of Berlin. I built a Clustering model to group similar neighborhoods together. After reading this report, along with the infographics provided, I hope the reader will be able to make an informed decision on where to setup a certain type of business in Berlin.