Battle of Neighbourhoods; Finding The Best Pivot Area For Cab Drivers in Berlin

1. Introduction: Business Problem

Berlin is the capital city of the Federal Republic of Germany. According to 2017 census data this city has approximately 3.65 million inhabitants and as far as population concerned Berlin is the biggest city in Germany. Berlin is evidently a metropole and moreover it is an overwhelming, fascinating and a growing city. Until 2030, the population of Berlin is supposed to grow around 7.5%, predominantly because of increasing migration[1].

Among the major sectors in Berlin include the tourism, the creative and cultural industries, the biotechnology and healthcare industry with medical and pharmaceutical industries, information and communication technologies, the construction and property industry, trade, optoelectronics, energy technology and Trade fair and congress industry. The city is a European hub of rail and air traffic. Berlin is one of the emerging international centers for innovative business start-ups, with high annual growth in the number of employed people[2].

Let's think about a person who wants to dive in to taxi business in Berlin. He wants to work as a cab driver in Berlin. Since he has a limited budget, his goal is profit maximization. Because of high mobility and better standards compared to conventional taxi business, he wants to work under the organisation of 'UBER' or being a member of 'MyTaxi' organisation. We can call our friend from now on an unconventional cab driver in order to differentiate him from normal taxi drivers who are not directed from a center and generally have yellow taxis. 'UBER' is one of the legal and valid organisations in Germany in taxi business. Although it is sometimes controversial and prone to some limitations because of conventional taxi driver's protest, UBER is generally available both in Germany and Berlin. Companies like MyTaxi are also providing similar service in Berlin[3].

In a routine day, as part of their routine life, many people commute in Berlin. The season (sommer, winter etc), the day type (weekday or weekend) and even the time period of the day (morning, noon, afternoon, night) has great effect on the total number of German people commuting in Berlin and their movement patterns respectively. We can talk about trends when people commute in such a metropole. Biggest people movement is always observed to be mainly from accommodation places to respective working places and vice versa. And also from working place to recreation areas or venues. Weekend periods are always exceptions. People don't work at weekends; they rest instead, and try to enjoy their free times. Therefore, they move from their respective homes to recreation areas or to possible venues available for spending time and vice versa. Sommer period is also a different case . In particular, for a touristic place like Berlin.

---

[1] https://de.wikipedia.org/wiki/Berlin

[2] https://en.wikipedia.org/wiki/Economy_of_Berlin

[3] https://mytaxi.com/de/

Both Uber and MyTaxi work on internet based applications (over cell phone or tablet) and these applications function based on the logic of finding the closest cab driver available for any client or passenger. When someone registered to one of this companies as a cab driver; first the application is installed and then the calls for cab comes from clients over this application. And it is so evident that mobility of people influences the possibility of finding client. If a cab driver is in the right place, he can easily find a client and earn money.

For a person working as a non-conventional cap driver minimizing the costs and maximizing the revenues is definitely the main goal. And being in the right place in the right time is the key to success in this business. A cab driver must find more clients in order to maximize the revenues and must be close to the best client source areas. We can call them live areas where people are. A cab driver can maximize profits by staying within a certain distance of the most live venues or recreation areas. Moreover, a predetermined center or pivot point that a cab driver has to return after each trip with client is quite logical to find more clients. Finding the best pivot neighbourhood for cab drivers will really help this labor group. This pivot area can basically be defined as the area with biggest number of venues such as restaurants, cinemas, theatres, hotels, museums, parks, working places etc. Because venues are serving people.

There may of course be other factors in effect such as the number of people living in this neighbourhood and the number of people working there. Moreover, public transportation service availability is also a factor for the revenues of non-conventional cab drivers. But in general one can state that people around venues need cabs to move to different places. We have an optimisation problem at hand and in order to find the best borough or neighbourhood for any cab driver as a pivot area to turn around, we need to segment and cluster the city of Berlin based on Foursquare API data. In addition, in order to find best client source, we need to find the top rated venues and calculate the total number of venues in each cluster.

There are 12 boroughs (Charlottenburg-Wilmersdorf, Friedrichshain-Kreuzberg, Lichtenberg, Marzahn-Hellersdorf,Mitte, Neukölln, Pankow, Reinickendorf, Spandau, Steglitz-Zehlendorf, Tempelhof-Schöneberg, Treptow-Köpenick) and 96 neighbourhoods in city of Berlin. Determining the right area (borough or neighbourhood) is our goal and we will use a Machine Learning algorithm to segment and cluster the neighbourhoods. Finally, we will make a recommendation to our cab drivers as a solution to their pivot area finding problem.

Boroughs of Berlin[4]

2. Data :

a. Data Sources/Web Scraping :

Due to difficulty of working in neighbourhood level in such a metropole, insignificance of finding a pivot solution for a taxi based on neighbourhoods and unavailability of some portion of neighbourhood level data for Berlin; segmenting and clustering analysis in this study is performed on BOROUGH level. Coordinates of borough zipcodes are used to find number of venues within a defined area. Venue requests are made based on borough information and mainly following data sources are used. The location data consists of the boroughs, their

4 https://upload.wikimedia.org/wikipedia/commons/6/68/Berlin%2C_administrative_divisions_%28%2Bdistricts_%2Bboroughs_-pop%29_-_de_-_colored.svg

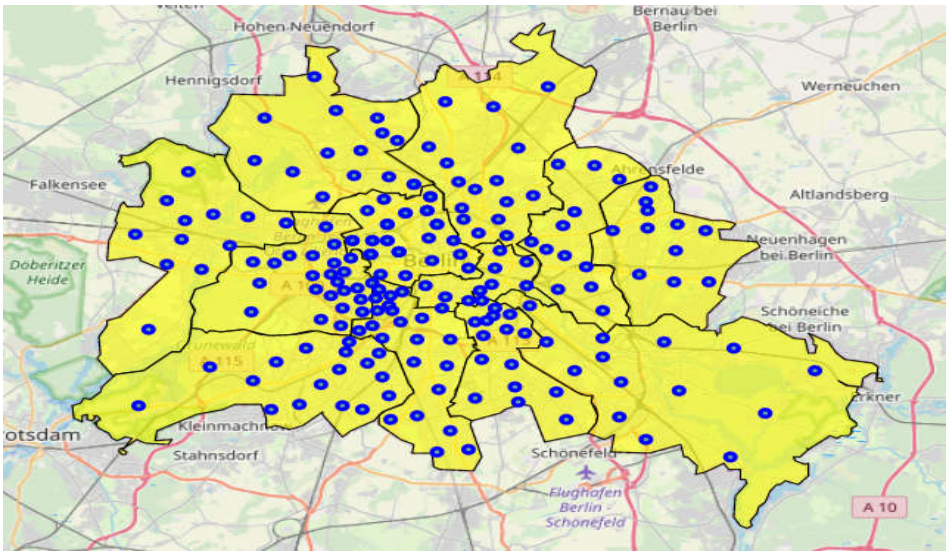
zipcodes, and coordinates[5]. I have scraped that data from 2 web pages and acquired necessary info. Zipcodes served as the index and refer to coordinates, borough names, etc. And I used location data in Github to describe the shape and centers of the boroughs as it is in geojson format [6]

b. Data Wrangling :

After some data processing and wrangling I had below data format.

	zipcode	state	latitude	longitude	Borough
0	10115	Berlin	52.5323	13.3846	Mitte
1	10117	Berlin	52.5170	13.3872	Mitte
2	10119	Berlin	52.5305	13.4053	Mitte
3	10178	Berlin	52.5213	13.4096	Mitte
4	10179	Berlin	52.5122	13.4164	Mitte

Later, I created a map of Berlin with all boroughs and zipcodes of all neighbourhoods superimposed on them by using Folio.



This data is combined with Foursquare API data. We explore the areas around the collected zipcodes(postal codes) in Berlin. Therewith, we perform location search and gather the most famous venues within a circle of 1500 meters radius for each zipcode. By doing so we will try to almost cover all areas within a specific borough. Four square data looked like below. As a result, I got 12678 total venues, and 394 unique venue categories in our dataset.

	Zipcode	Zipcode Latitude	Zipcode Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
13595	14199	52.4777	13.2951	U Podbielskiallee	52.464129	13.296021	Metro Station
13596	14199	52.4777	13.2951	Erlenbusch	52.464333	13.303959	Park
13597	14199	52.4777	13.2951	Inselbistro	52.463291	13.298882	Diner
13598	14199	52.4777	13.2951	H Herthastraße	52.493617	13.285064	Bus Stop
13599	14199	52.4777	13.2951	Hundeauslaufgebiet	52.466233	13.272643	Dog Run

[5] http://www.statistik-berlin-brandenburg.de/produkte/verzeichnisse/zuordnungderbezirkezupostleitzahlen.xls
https://raw.githubusercontent.com/TrustChainEG/postal-codes-json-xml-csv/master/data/DE/zipcodes.de.csv
[6] https://raw.githubusercontent.com/m-hoerz/berlin-shapes/master/berliner-bezirke.geojson

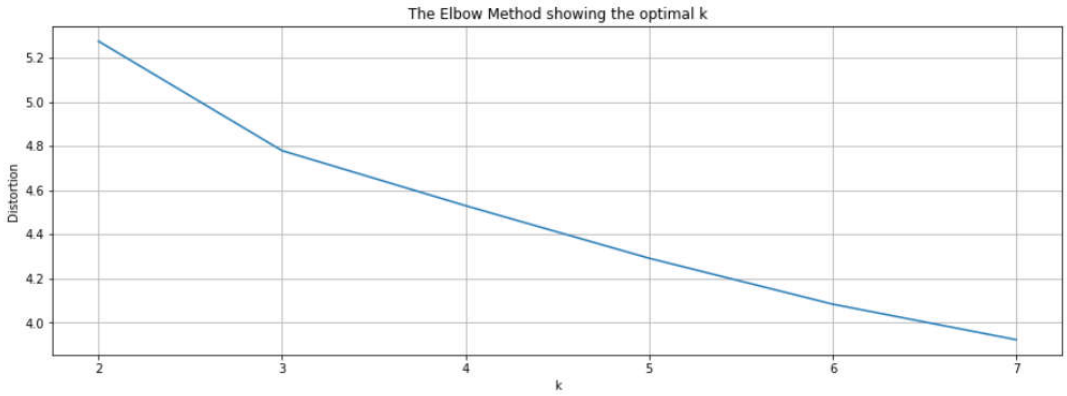
The parameters “radius” and “number of venues” are reasonable choices for finding number of venues within all boroughs. Yelp API provides data concerning top-rated venues at given coordinates. This information might interest cab drivers since those venues are plac e of attractions for cab clients. Further, to obtain a systematic view on the structure of the boroughs, a cluster analysis facilitated a comparison of the locations. Therefore, the venues and their categories were collected at each zipcode in order to compare the relative frequencies of venues per category at each zipcode.

These frequencies of venues per category, called "category feature" serve as a measurement of dissimilarity of distinct locations. The cluster analysis groups locations with similar "category features" into a cluster and separates locations with more diverse features.

	Zipcode	ATM	Adult Boutique	African Restaurant	Airport Service	American Restaurant	Animal Shelter	Antique Shop	Aquarium	Argentinian Restaurant	...	Vietnamese Restaurant	Volleyball Court	Waterfront	Whisky Bar	Wine Bar	Wine Shop	Winery	Women's Store	Yoga Studio	Zoo Exhibit
0	10115	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.020000	0.02	0.0	0.0	0.020000	0.00	0.0	0.0	0.00	0.0
1	10117	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.000000	0.00	0.0	0.0	0.034884	0.00	0.0	0.0	0.00	0.0
2	10119	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.050000	0.00	0.0	0.0	0.010000	0.00	0.0	0.0	0.00	0.0
3	10178	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.060000	0.00	0.0	0.0	0.000000	0.01	0.0	0.0	0.01	0.0
4	10179	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.033333	0.00	0.0	0.0	0.000000	0.00	0.0	0.0	0.00	0.0

5 rows × 361 columns

In order to determine optimum number of clusters to be used, I deployed Elbow method before K-means Clustering. When k increases, the centroids are closer to the cluster centroids. Here the distortion, mean sum of squared distances to centers, decreases to the optimum point. The improvements will decline, at some point rapidly, creating the elbow shape. That point is the optimal value for k in the image above, k=3



Equipped with these data and tools, one can select some interesting locations. In order to explore the selected location in more detail, we present top-rated venues at given zipcodes.

The analysis mainly applies the following Python libraries:

- Pandas, Numpy – Libraries for data storage, manipulation and array computing

- Scipy – Library for dendrogram and hierarchical cluster analysis

- Matplotlib, Folium – Libraries for representing numeric and locational data

- Geopy – Library to retrieve locational data

- Json – Library to handle JSON files

- Requests, Urllib – Libraries to retrieve data and handle http exchange with the Foursquare API and Yelp API.

3. Methodology :

a. General Methodology

In this part the data is explored by using visualization and basic data analysis approach to understand which parts of the data is more meaningful for my solution and try to understand which variables can help me to define the best pivot area for cab drivers. Descriptive statistics and visualization will be followed by machine learning.

With the use of data acquired, I applied K-means Clustering Machine learning algorithm to group the zipcodes and get some insight which area or combination of zipcodes is more suitable for our cab driver, in terms of minimizing distances of travelling without a client and costs. While on the other hand, improving the revenues.

b. Explore Data Set/ Neighborhoods in Berlin

I will in this part give some summary information over the zipcodes that we gathered, try to explore and later to analyze data to get insight for the solution of our business problem. Since I would like to determine a pivot area for a cab driver then the number of venues, which means clients for a cab, in each zipcode area will be valuable. This table summarizes number of venues within 1500 meter of particular zipcode latitude and longitude. And this total number venues are limited to 100 as an upper limit.

[16]:

Zipcode	Zipcode Latitude	Zipcode Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
10115	100	100	100	100	100	100
10117	100	100	100	100	100	100
10119	100	100	100	100	100	100
10178	100	100	100	100	100	100
10179	100	100	100	100	100	100
10243	100	100	100	100	100	100
10245	100	100	100	100	100	100

c. Preprocessing Venues

In this section I created a new DataFrame and displayed the top 10 venue categories per zipcode, measured by the relative frequency of venues per category.

	Zipcode	ATM	Adult Boutique	African Restaurant	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	Amphitheater	Animal Shelter	Antique Shop	Aquarium	Arcade	Argentinian Restaurant	Art Gallery	Art Museum	Arts & Crafts Store	Asian Restaurant	Athletics & Sports	Austrian Restaurant	Auto Dealership	Auto Garage	Auto Workshop	Automotive Shop	BBQ Joint	Baby Store	Bagel Shop	Bake
0	10115	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.03	0.01	0.00	0.00	0.02	0.00	0.0	0.0	0.0	0.0	0.01	0.0	0.0	0.
1	10117	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.01	0.02	0.00	0.01	0.00	0.01	0.0	0.0	0.0	0.0	0.00	0.0	0.0	0.
2	10119	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.01	0.00	0.00	0.00	0.00	0.00	0.0	0.0	0.0	0.0	0.00	0.0	0.0	0.
3	10178	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.02	0.02	0.00	0.01	0.00	0.00	0.0	0.0	0.0	0.0	0.00	0.0	0.0	0.
4	10179	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.02	0.00	0.01	0.00	0.00	0.00	0.0	0.0	0.0	0.0	0.00	0.0	0.0	0.

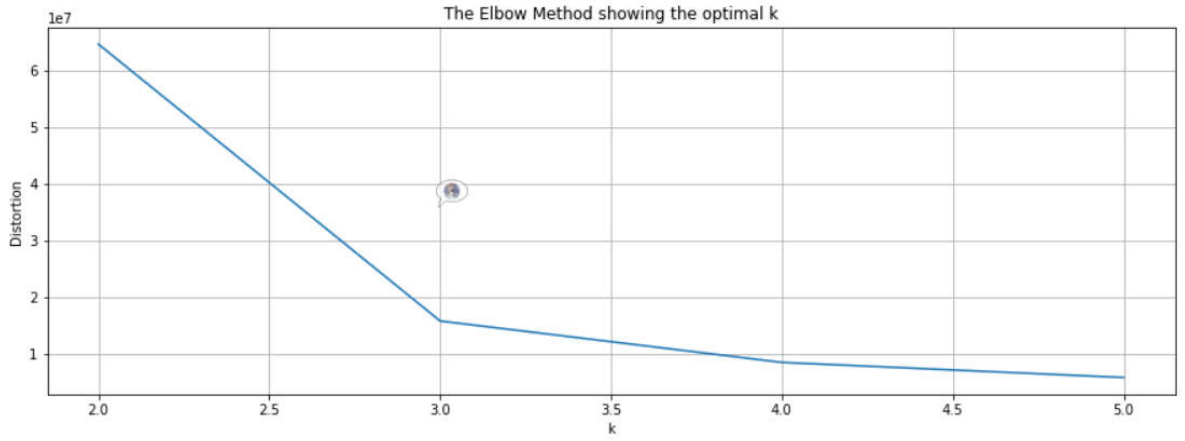
Then I looked for most common venues for each zipcode.

	Zipcode	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	10115	Coffee Shop	Hotel	Art Gallery	Hostel	Restaurant	Cocktail Bar	Boutique	Bookstore	Italian Restaurant	Nightclub
1	10117	Hotel	History Museum	Theater	Clothing Store	Monument / Landmark	Chocolate Shop	Cocktail Bar	Bookstore	Gourmet Shop	Drugstore
2	10119	Italian Restaurant	Coffee Shop	Ice Cream Shop	Bookstore	Beer Bar	Vietnamese Restaurant	Café	Tea Room	Hotel	Burger Joint
3	10178	Hotel	Café	Indie Movie Theater	Coffee Shop	Vietnamese Restaurant	Ice Cream Shop	Optical Shop	History Museum	Gym / Fitness Center	Gift Shop
4	10179	Hotel	Nightclub	Coffee Shop	Restaurant	German Restaurant	Café	Italian Restaurant	Garden	Bakery	Museum
5	10243	Italian Restaurant	Vegetarian / Vegan Restaurant	Nightclub	Café	Bar	Bakery	Hotel	Hostel	Coffee Shop	Record Shop
6	10245	Café	Nightclub	Vegetarian / Vegan Restaurant	Bar	Vietnamese Restaurant	Ice Cream Shop	Gym / Fitness Center	Falafel Restaurant	Coffee Shop	Beer Garden
7	10247	Café	Pizza Place	Bar	Falafel Restaurant	Coffee Shop	Ice Cream Shop	Bakery	Vegetarian / Vegan Restaurant	Wine Bar	Bookstore
8	10249	Café	Italian Restaurant	Vietnamese Restaurant	Park	Hotel	Ice Cream Shop	Indie Movie Theater	Bar	Pizza Place	Nightclub
9	10315	Tram Station	Supermarket	Bakery	Hotel	Zoo Exhibit	Drugstore	Bus Stop	Cafeteria	Coffee Shop	Park
10	10317	Supermarket	Park	Beer Garden	Bakery	Italian Restaurant	Hotel	Boat Rental	Boat or Ferry	Harbor / Marina	Café
11	10318	Supermarket	Tram Station	Organic Grocery	Light Rail Station	Bakery	Greek Restaurant	German Restaurant	Gastropub	Forest	Flower Shop
12	10319	Supermarket	Zoo Exhibit	Drugstore	Bus Stop	Bakery	Pizza Place	Italian Restaurant	Tram Station	Dog Run	Stadium
13	10365	Supermarket	Park	Vietnamese Restaurant	Hotel	Coffee Shop	Bakery	Asian Restaurant	Smoke Shop	Shopping Mall	Hardware Store
14	10367	Supermarket	Italian Restaurant	Café	Coffee Shop	Tram Station	Park	Fast Food Restaurant	Bakery	Pizza Place	Asian Restaurant
15	10369	Supermarket	Pizza Place	Hotel	Park	Vietnamese Restaurant	Bar	Tram Station	Café	Soccer Field	Nightclub
16	10405	Café	Ice Cream Shop	Vietnamese Restaurant	Cocktail Bar	Wine Bar	Coffee Shop	Italian Restaurant	Beer Bar	Bar	Park
17	10407	Café	Vietnamese Restaurant	Park	Bakery	Italian Restaurant	Hotel	Supermarket	Pizza Place	Drugstore	Bar

After all, the Dataset is Ready for any Machine Learning Algorithm now.

d. Machine Learning/ Clustering Neighborhoods

I deployed K-Elbow Method to find optimum K . And it is found to be 3.



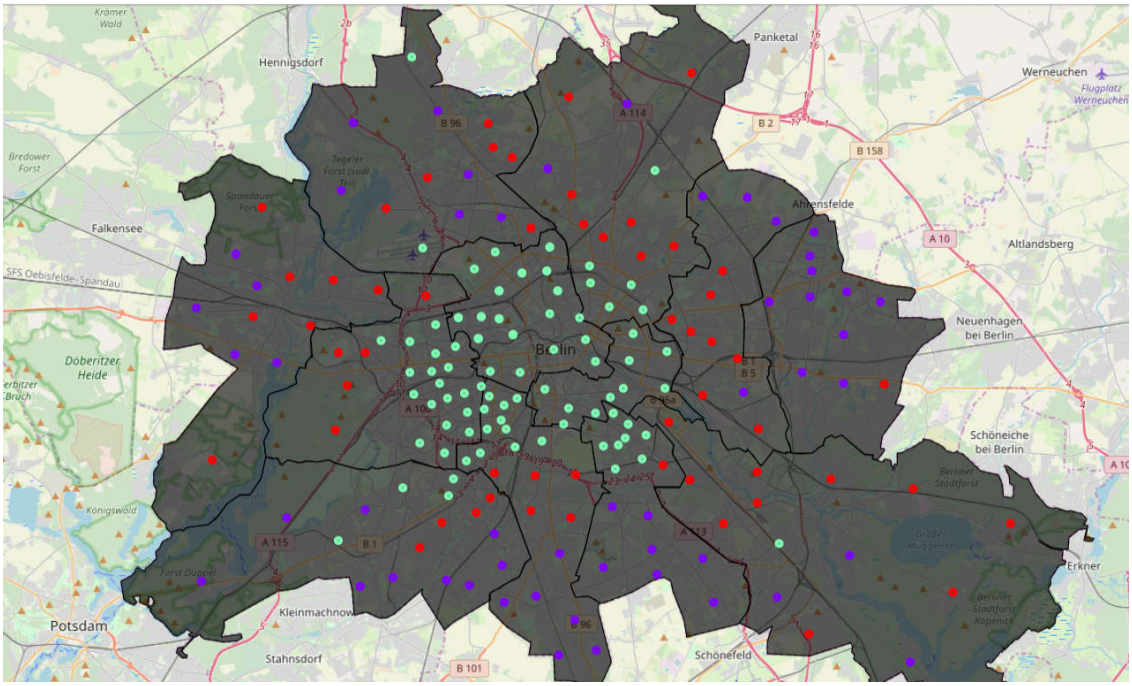
When K increases, the centroids are closer to the clusters centroids. Here the distortion, mean sum of squared distances to centers, decreases to the optimum point. The improvements will decline, at some point rapidly, creating the elbow shape. That point is the optimal value for k. In the image above, k=3. We run now k-means to Cluster Zipcodes into 3 Clusters. Below table shows results of this K-means clustering together with 10 most common venue in each cluster.

	Zipcode	Cluster Labels	latitude	longitude	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	10115	2	52.5323	13.3846	Mitte	Coffee Shop	Hotel	Art Gallery	Hostel	Restaurant	Cocktail Bar	Boutique	Bookstore	Italian Restaurant	Nightclub
1	10117	2	52.5170	13.3872	Mitte	Hotel	History Museum	Theater	Clothing Store	Monument / Landmark	Chocolate Shop	Cocktail Bar	Bookstore	Gourmet Shop	Drugstore
2	10119	2	52.5305	13.4053	Mitte	Italian Restaurant	Coffee Shop	Ice Cream Shop	Bookstore	Beer Bar	Vietnamese Restaurant	Café	Tea Room	Hotel	Burger Joint
3	10178	2	52.5213	13.4096	Mitte	Hotel	Café	Indie Movie Theater	Coffee Shop	Vietnamese Restaurant	Ice Cream Shop	Optical Shop	History Museum	Gym / Fitness Center	Gift Shop
4	10179	2	52.5122	13.4164	Mitte	Hotel	Nightclub	Coffee Shop	Restaurant	German Restaurant	Café	Italian Restaurant	Garden	Bakery	Museum
5	10243	2	52.5123	13.4394	Friedrichshain-Kreuzberg	Italian Restaurant	Vegetarian / Vegan Restaurant	Nightclub	Café	Bar	Bakery	Hotel	Hostel	Coffee Shop	Record Shop
6	10245	2	52.5007	13.4647	Friedrichshain-Kreuzberg	Café	Nightclub	Vegetarian / Vegan Restaurant	Bar	Vietnamese Restaurant	Ice Cream Shop	Gym / Fitness Center	Falafel Restaurant	Coffee Shop	Beer Garden
7	10247	2	52.5161	13.4656	Friedrichshain-Kreuzberg	Café	Pizza Place	Bar	Falafel Restaurant	Coffee Shop	Ice Cream Shop	Bakery	Vegetarian / Vegan Restaurant	Wine Bar	Bookstore
8	10249	2	52.5238	13.4428	Friedrichshain-Kreuzberg	Café	Italian Restaurant	Vietnamese Restaurant	Park	Hotel	Ice Cream Shop	Indie Movie Theater	Bar	Pizza Place	Nightclub
9	10315	3	52.5132	13.5148	Lichtenberg	Tram Station	Supermarket	Bakery	Hotel	Zoo Exhibit	Drugstore	Bus Stop	Cafeteria	Coffee Shop	Park

4. Results (where you discuss the results)

a. Examine Clusters

The following map visualizes the distribution of clustered zipcodes (based on similar venues) in Berlin:



The map of the clustered locations show some geographical structure. When we look at the map closely we briefly see that;

1. Cluster 2 **(LIGHT GREEN)** lie around the center of Berlin. The boroughs it encompasses together with the portions are; Mitte (% 100), Friedrichshain-Kreuzberg (% 100), Charlottenburg Wilmersdorf (% 45), Tempelhof Schöneberg (% 25), Neukölln (% 20),Pankow (% 25), Steglitz Zehlendorf (% 15).

2. Cluster 3 (RED) is mainly located towards outwards of center of Berlin. It lies as an outer boundry around Cluster 2. The boroughs it encompasses together with the portions are: Treptow-Köpenick (% 75) Lichtenberg (% 60) Pankow (% 40) Reinickendorf (% 20) Charlottenburg-Wilmersdorf (% 20) Spandau (% 30) Steglitz-Zehlendorf (% 10).

3. Cluster 1 (BLUE) is mainly located more outwards from Berlin center than Cluster 3 . It lies as an outer bound around both Clusters 2 and 3. The boroughs it encompasses together with the portions are: Marzahn-Hellersdorf (% 90) Lichtenberg (% 15) Pankow (% 5) Reinickendorf (% 45) Spandau (% 30) Steglitz-Zehlendorf (% 40) Tempelhof Schöneberg (% 35 Neukölln (% 35) Treptow-Köpenick (% 15) Marzahn-Hellersdorf (% 80)

b. Describe the clusters

In this section, I present the clusters and determine the venues categories that distinguish the clusters.

Cluster1 (BLUE)

Cluster1 lies close to the outer boundaries of Berlin. Therefore one might expect less venues for tourism, food, coffee, or other recreational places compared to more central clusters. The modes show the most frequent venue categories per rank over all zipcode without duplicates. The most important venue categories in this Cluster are: 'Big Box Store',

'Bus Stop', 'Drugstore', 'Electronics Store', 'German Restaurant', 'Park', 'Supermarket'. This cluster includes 52 zipcodes. When we consider overall potential clients for cabs. This cluster is not so fruitfull compared to other 2 Clusters. For details see table "Top ten venue categories".

	Zipcode	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	10319	Supermarket	Zoo Exhibit	Drugstore	Bus Stop	Bakery	Pizza Place	Italian Restaurant	Tram Station	Dog Run	Stadium
1	12107	Supermarket	Steakhouse	Bakery	Diner	Chinese Restaurant	Bus Stop	Italian Restaurant	Hotel	Lottery Retailer	Greek Restaurant
2	12207	Supermarket	Bakery	Bus Stop	Organic Grocery	Light Rail Station	Drugstore	Hotel	Café	German Restaurant	Fast Food Restaurant
3	12209	Supermarket	Fast Food Restaurant	Drugstore	Bus Stop	Organic Grocery	Bakery	German Restaurant	Tennis Stadium	Big Box Store	Taverna
4	12247	Supermarket	Park	Bus Stop	Italian Restaurant	Drugstore	Liquor Store	Greek Restaurant	Tennis Court	Taverna	Ice Cream Shop
5	12249	Supermarket	Italian Restaurant	Bus Stop	Drugstore	History Museum	Optical Shop	Movie Theater	Fast Food Restaurant	Taverna	Park
6	12277	Supermarket	Bus Stop	Light Rail Station	Chinese Restaurant	Fast Food Restaurant	Big Box Store	Italian Restaurant	Clothing Store	Arts & Crafts Store	Tennis Court
7	12279	Supermarket	Bus Stop	Italian Restaurant	Taverna	Restaurant	Chinese Restaurant	Soccer Field	Park	Fast Food Restaurant	Recreation Center
8	12305	Supermarket	Bus Stop	Italian Restaurant	Greek Restaurant	Chinese Restaurant	Electronics Store	Organic Grocery	Taverna	Fast Food Restaurant	Big Box Store
9	12307	Supermarket	Bakery	Doner Restaurant	Soccer Field	Gas Station	Miscellaneous Shop	Mobile Phone Shop	Electronics Store	Pharmacy	Light Rail Station
10	12309	Supermarket	Doner Restaurant	Soccer Field	Bus Stop	German Restaurant	Gas Station	Miscellaneous Shop	Bakery	Gym / Fitness Center	Mobile Phone Shop

Individual Top Ten Categories in Cluster-1

Cluster2 (LIGHT GREEN)

Cluster2 lies around the centre of Berlin. It is so evident that this cluster area includes majority of the live places or recreational areas in Berlin. The modes show the most frequent venue categories per rank over all zipcode without duplicates. This cluster includes 80 zipcodes and all of them lie around city center. Most important venue categories in this cluster are: 'Big Box Store', 'Bus Stop', 'Drugstore', 'Electronics Store', 'German Restaurant', 'Park', 'Supermarket'. It is so evident that Cluster 2 has much more cab client potential than other Clusters. For details see table "Top ten venue categories".

	Zipcode	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	10115	Coffee Shop	Hotel	Art Gallery	Hostel	Restaurant	Cocktail Bar	Boutique	Bookstore	Italian Restaurant	Nightclub
1	10117	Hotel	History Museum	Theater	Clothing Store	Monument / Landmark	Chocolate Shop	Cocktail Bar	Bookstore	Gourmet Shop	Drugstore
2	10119	Italian Restaurant	Coffee Shop	Ice Cream Shop	Bookstore	Beer Bar	Vietnamese Restaurant	Café	Tea Room	Hotel	Burger Joint
3	10178	Hotel	Café	Indie Movie Theater	Coffee Shop	Vietnamese Restaurant	Ice Cream Shop	Optical Shop	History Museum	Gym / Fitness Center	Gift Shop
4	10179	Hotel	Nightclub	Coffee Shop	Restaurant	German Restaurant	Café	Italian Restaurant	Garden	Bakery	Museum
			Vegetarian / Vegan								

Individual Top Ten Categories in Cluster-2

Results: Cluster 3 (RED)

Cluster3 lies between other two clusters as a pillow. It seems that this cluster has more cab client potential than cluster 1 but however less potential than cluster 2. The modes show the most frequent venue categories per rank over all zipcode without duplicates. This cluster includes 54 zipcodes and majority of them lie between the other two clusters (Cluster 1 and 2). Most important venue categories in this cluster are: 'Bakery', 'Big Box Store', 'Café', 'Chinese Restaurant', 'Drugstore', 'Greek Restaurant', 'Park', 'Supermarket'. It is so evident that Cluster 3 has much more cab client potential than Cluster 1 but less than Cluster 2. For details see table "Top ten venue categories".

	Zipcode	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	10315	Tram Station	Supermarket	Bakery	Hotel	Zoo Exhibit	Drugstore	Bus Stop	Cafeteria	Coffee Shop	Park
1	10317	Supermarket	Park	Beer Garden	Bakery	Italian Restaurant	Hotel	Boat Rental	Boat or Ferry	Harbor / Marina	Café
2	10318	Supermarket	Tram Station	Organic Grocery	Light Rail Station	Bakery	Greek Restaurant	German Restaurant	Gastropub	Forest	Flower Shop
3	10365	Supermarket	Park	Vietnamese Restaurant	Hotel	Coffee Shop	Bakery	Asian Restaurant	Smoke Shop	Shopping Mall	Hardware Store
4	10367	Supermarket	Italian Restaurant	Café	Coffee Shop	Tram Station	Park	Fast Food Restaurant	Bakery	Pizza Place	Asian Restaurant
5	10369	Supermarket	Pizza Place	Hotel	Park	Vietnamese Restaurant	Bar	Tram Station	Café	Soccer Field	Nightclub
6	12057	Supermarket	Bus Stop	Bar	Bakery	Breakfast Spot	Café	Rental Car Location	Doner Restaurant	Light Rail Station	Fast Food Restaurant
										Vietnamese	

Individual Top Ten Categories in Cluster-3

c. Exploring Top Rated Venues

As far as taxi business concerned, finding a pivot area may really help. But on the other hand, pinpointing most live places within this pivot area also helps more. Therefore we "ll try to find top rated places within our most fruitful cluster. This will facilitate our cab driver"s patrol area problem and alleviate client finding problem.

Now I want to survey top rated venues in Cluster 2. Since Cluster 2 seems to be the best pivot area, I want to find most rated venues in this pivot area to tell more to our cab driver about his cab client finding problem. As we have mentioned and seen in the map, Cluster 3 encompasses Boroughs such as: Mitte, Friedrichshain-Kreuzberg, Charlottenburg Wilmersdorf, Tempelhof Schöneberg, Neukölln, Pankow, Steglitz Zehlendorf. Therefore it is good to find most rated venues in these boroughs. For the sake of simplicity I have chosen first zipcode from each Borough. By using same algorithm all top rated venues for all zipcodes can be found.

Sample Survey Result of Top Rated Venues in Each Borough Based on Zipcodes

5. Discussion

In order to recommend the best pivot area for a cab driver, cluster analysis is used in this project. Based on similarities of venues, Berlin is segmented and clustered into 3 different clusters. Since a venue is a place where people being served and spend time, those venues are always potential client sources for cab drivers. And in such a metropole like Berlin people frequently commute to or from venues. Therefore, the aim of this project was to find not only the most crowded venue cluster but also to find the top-rated venues within these clusters.

After all one can easily say that in such a metropole, in any case, the places around city center is strongly excepted to be the best solution. And that is a solid fact and this solution can most probably be declared as valid. However, with such a metropole foursquare data (12678 total venues and 394 uniques categories) it is not always easy to say top rated venues and also it is not easy to delimit a particular pivot area.

In addition, within the scope of this project, a methodology has been followed and with the help of a scientific approach a well-known fact has been acknowledged. Finally an anticipation has come to truth. At the end of the day, we can now more powerfully state that the area around city center is the most fruitful area as far as top rated venues and potential cab clients concerned.

We can imitate this methodology for different metropoles and to some content may improve it. Since taxi business is very seasonal and very vulnerable to weather conditions, implementing real time meteorological data to the analysis really contributes to the issue. For example, in unpredicted bad weather conditions, the demand for cabs highly increases. And being in the right place in the right time requires more information. Besides, when mobility is concerned more timely data may contribute better to the problem. In another words a cab driver needs a tool showing real time people flood in such a metropole.

There are a lot of different things affecting cab demand in big cities. Availability of public transport is also a great variable of this topic and can be included in future variations of this analysis.

Yelp API is also used within the scope of this study in order to support our pivot area with top rated venues. It is also a quite necessary information for a cab driver patrolling in a determined pivot area. If he knows top rated places then he can anticipate more to find more clients around these places.

Kmeans algorithm is used as a part of this clustering study. As a result of the Elbow method, 3 is found as k value. Data analysis and visual exploration is also used to see the differences between zipcodes and try to find trends with regards to most fruitful venues.

Finally, offering a cluster area as a pilot area does not mean that a cab driver can never find clients out of this area. There is always a chance to have client in any part of the city. But the probability of getting a client in cluster 2 is expected to be higher than the others.

6. Conclusion

Findings of this project illustrates how measures of data science support a systematic analysis of neighbourhoods. Within the scope of this project, various data sources and analysis methods have been implemented. The results show differences in a lot of respects. A hierarchical cluster analysis classifies zipcodes into three distinct clusters. A cluster is grouped by zipcodes with a similar structure of venues. A comparison of clusters show clear differences in the venues structures that can be related to specific areas in Berlin.

Moreover, in broader sense, the goal of the project is to support peoples who consider to be a cab driver in Berlin. A solution area has been delimited as a result of k-means clustering and the solution is supported with top rated venues. The next step for this project can be integration of particular data such as weather conditions, public transport availability to the analysis data set.

In conclusion cluster 2 is recommended to be the solution pivot area. It includes Mitte, Friedrichshain-Kreuzberg, Charlottenburg Wilmersdorf , Tempelhof Schöneberg, Neukölln , Pankow , Steglitz Zehlendorf boroughs and encompasses the city center. This part of city of Berlin is full of recreational areas and touristic sites, including the famous "Berlin Wall". The second alternative is found to be the next circular area around cluster 2. That is Cluster 3 and the last one is Cluster 1. When we look at the number of zipcodes within each cluster we observe similar trend. Cluster 2 has 80, Cluster 3 has 54 and Cluster 1 has 52 zipcodes.

In conclusion, I recommend cab drivers in Berlin to use Cluster 2 area to catch clients or passengers by paying more attention to the top-rated areas within this cluster. And hope the best for them.

REFERENCES:

1.https://de.wikipedia.org/wiki/Berlin
2.https://en.wikipedia.org/wiki/Economy_of_Berlin
3.https://mytaxi.com/de/
4.https://upload.wikimedia.org/wikipedia/commons/6/68/Berlin%2C_administrative_divisions_%28%2Bdistricts_%2Bboroughs_-pop%29_-_de_-_colored.svg
5.http://www.statistik-berlin-brandenburg.de/produkte/verzeichnisse/zuordnungderbezirkezupostleitzahlen.xls
https://raw.githubusercontent.com/TrustChainEG/postal-codes-json-xml-csv/master/data/DE/zipcodes.de.csv
6.https://raw.githubusercontent.com/m-hoerz/berlin-shapes/master/berliner-bezirke.geojson

Guray CERMAN
Grycrmn.de@gmail.com
https://www.linkedin.com/in/garuy-carmen-259063176/
https://github.com/crmngry