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

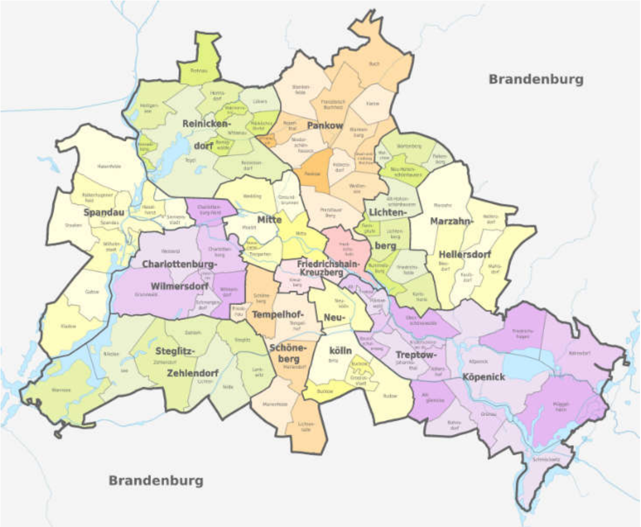
[3] https://mytaxi.com/de/

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 :

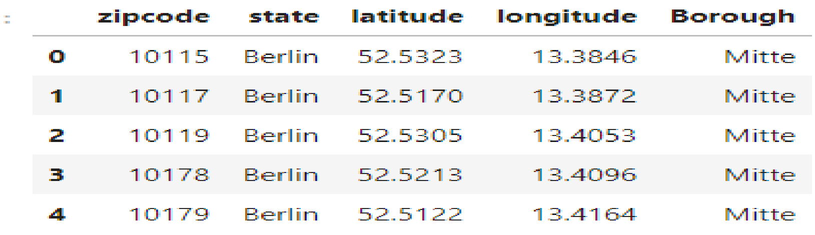
Due to difficulty of working in neighbourhood level in such a metropole, insignificance of finding a pivot solution for a taxi based on neighbourhoods and unavailability of some portion of neighbourhood level data for Berlin; segmenting and clustering analysis in this study is performed on BOROUGH level. 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]

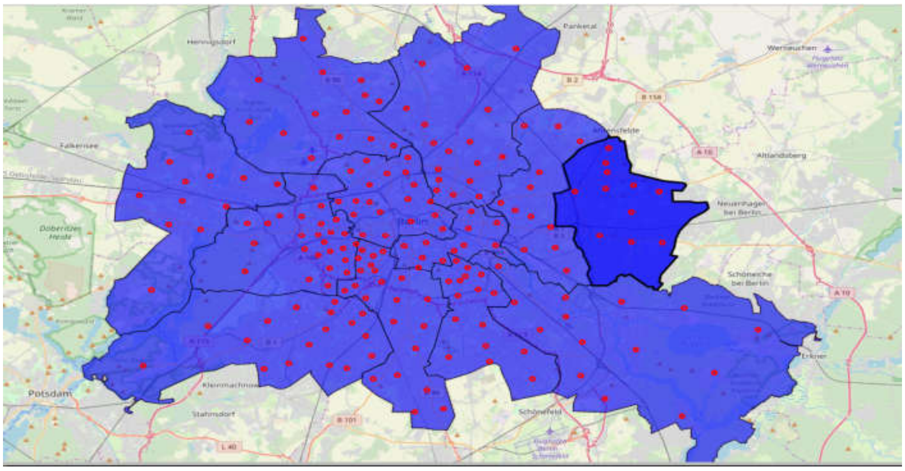
b. Data Wrangling :

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

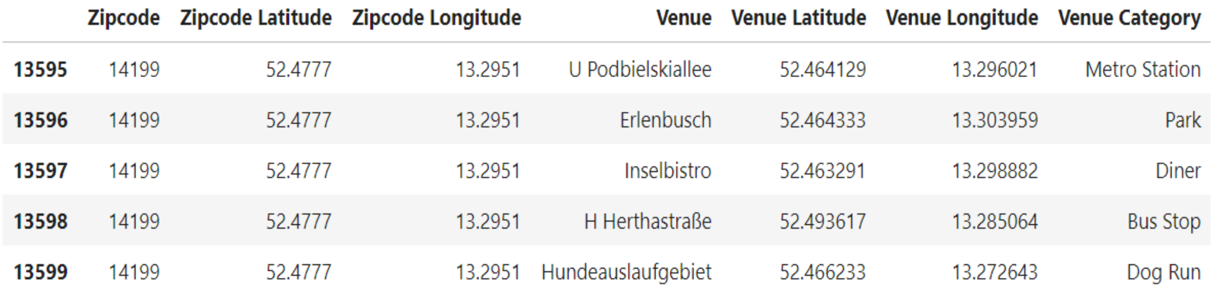


	zipcode	state	latitude	longitude	Borough
0	10115	Berlin	52.5323	13.3846	Mitte
1	10117	Berlin	52.5170	13.3872	Mitte
2	10119	Berlin	52.5305	13.4053	Mitte
3	10178	Berlin	52.5213	13.4096	Mitte
4	10179	Berlin	52.5122	13.4164	Mitte

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 2000 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 13600 total venues and 373 unique categories in the data set.



	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

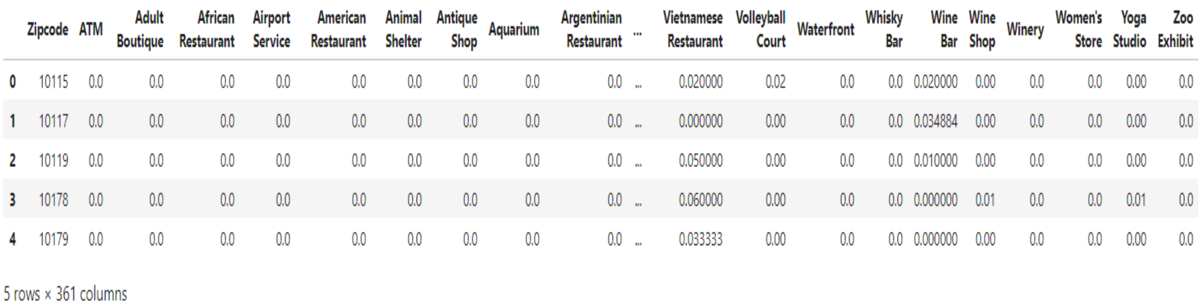
The parameters “radius” and “number of venues” are reasonable choices for finding number of venues within all boroughs. Yelp API provides data concerning top-rated venues at given coordinates. This information might interest cab drivers since those venues are plac

[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

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.

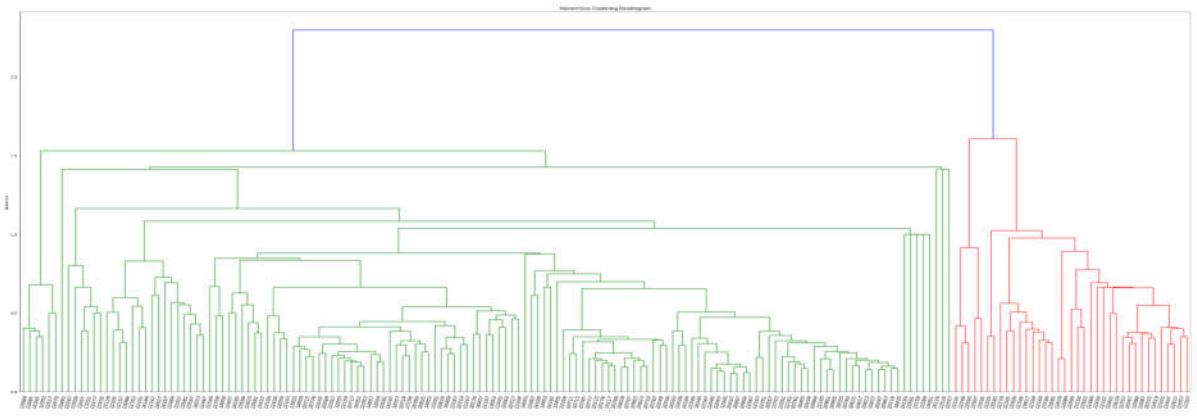
These frequencies of venues per category, called "category feature" serve as a measurement of dissimilarity of distinct locations. The cluster analysis groups locations with similar "category features" into a cluster and separates locations with more diverse features.



	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

A dendrogram shows the distances between the "category features" in order to determine a plausible number of clusters. Therewith, a hierarchical cluster algorithm provides the cluster labels for the zipcodes. These derivers cluster of similar locations within boroughs of Berlin.



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.