CAPSTONE PROJECT REPORT

ON

CLUSTERING NEIGHBOURHOODS IN BERLIN CITY (GERMANY) TO FIND OUT WHICH NEIGHBOURHOODS REQUIRE MORE PUBLIC PARKS, BASED ON PUBLIC PARKS AVAILABLE TO RESIDENTS

SUBMITTED BY:

GAURI TOSHNIWAL

CONTENTS:

1. INTRODUCTION

1.1**. BACKGROUND**

City parks play a vital role in the social, economic, and physical well-being of cities and their residents. City parks provide access to recreational opportunities, increase property values, spur local economies, combat crime, and protect cities from environmental impact. City parks encourage active lifestyles and reduce health costs also parks help clean the air and improve public health.

1.2. **PROBLEM**

Parks can only reach their peak potential if every household can access such green space and that access shouldn't be tied to a private vehicle or an unreasonable walking distance. At the city level, researchers have long confirmed that park access is unequal. City residents may also remain disconnected from parks depending on their income, education, and race. Unequal park access is creating another form of environmental injustice.

There is an urgency to understand park access at the metropolitan scale. In response, metropolitan areas need regional plans to bring parks closer to all residents. Increased walkability and urban connectivity are required. Local planners and real estate developers lack data-driven insights which will help them to create a plan which will reduce spatial and economic barriers tied to park access.

1.3**. INTEREST**

It will be useful to city planners, real estate developers and the government as working on these insights will lead to the betterment of the city. Analyzing and finding out which areas require more attention and urgency towards park development will help in decision making and ultimately implementation of theses insights will benefit all the residents of the city.

2. DATA ACQUISITION AND CLEANING

2.1. **DATA SOURCES**

2.1.1. **Berlin Neighbourhood Data:**

Data will be scraped from the following Wikipedia page https://en.wikipedia.org/wiki/Boroughs_and_neighborhoods_of_Berlin

The data includes 96 localities in Berlin city.

2.1.2**. Location Coordinates for each Neighbourhood:**

The Geocoder Python package will be used to get latitude and longitude of each Neighbourhood.

2.1.3. **Berlin Venue Recommendations from FourSquare API**

(FourSquare website: www.foursquare.com)

The FourSquare API will be used to explore neighbourhoods in Berlin City. The Foursquare explore function will be used to get parks in each neighbourhood, and then use this as one of feature to group the neighbourhoods into clusters. The following information will be retrieved:

· Venue Name

· Coordinates: Latitude and Longitude

· Category Name

2.2. **DATA CLEANING**

Data was scraped from the mentioned resources which included localities in Berlin City. This data was in the form of lots of data frames(Every Borough had a separate table on the Html page) which had the following columns:

- Locality
- Area in Km2
- The population as of 2008
- Density inhabitants per Km2
- Map

All these data frames had similar columns and so they were combined in a single data frame by taking Union of all the data frames.The Map column was deleted as it was unnecessary and a proper index was set.

Next using Nominatium from geopy latitudes and longitudes for all Localities was fetched. These data frames consisting of coordinates were then attached to the main data frame. Finally, the resulting data frames had these following columns:

- Locality
- Area in Km2
- The population as of 2008
- Density inhabitants per Km2
- Latitude
- Longitude

This data frame was finally stored in a CSV file named berlin_localities. Now, this data frame was used to get venues in surrounding of each neighbourhood this was done using Foursquare API the URL was formed so that all parks near every locality could be fetched. This data was in JSON format it was then converted into structured data frames which consisted of following columns:

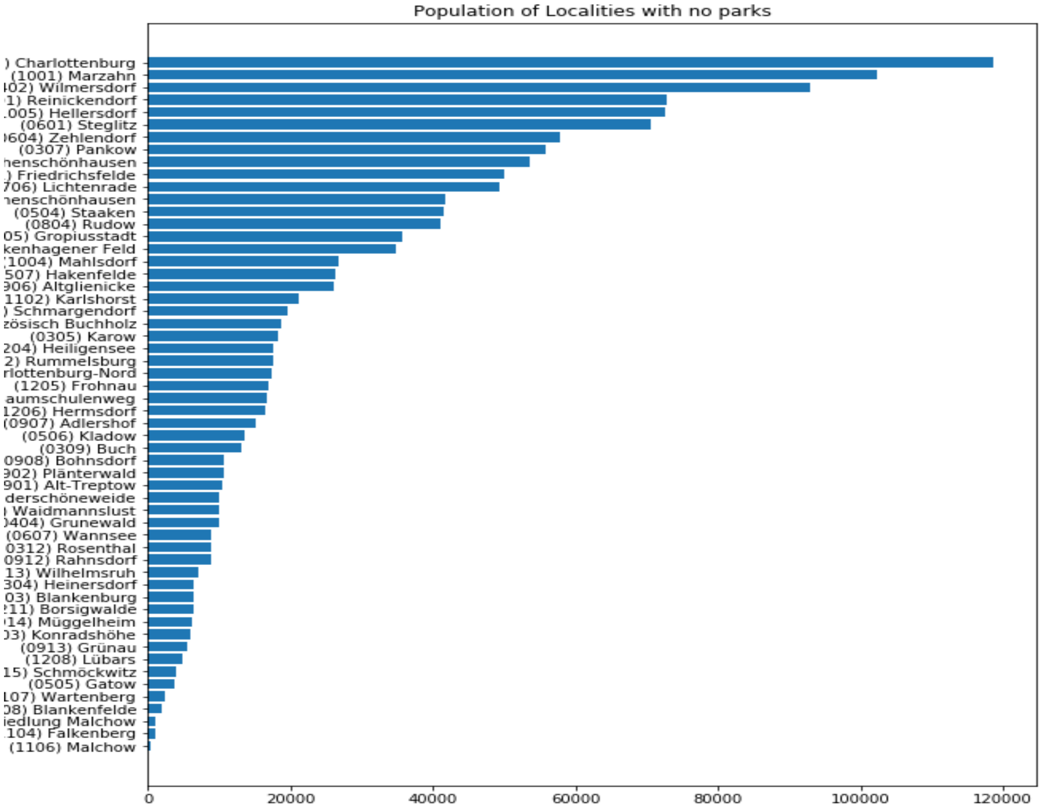
- Neighbourhood
- Neighbourhood latitude
- Neighbourhood Longitude
- Venue
- Venue Latitude
- Venue Longitude
- Venue Category

This data frame was stored in a CSV file named berlin_venues_refined. After this Values in the venue category columns were checked any extra categories other than park, garden, a playground was removed. This resulted in 66 rows indicating a total of 66 parks. Then the parks were grouped according to the neighbourhoods to find the number of parks each neighbourhood had.

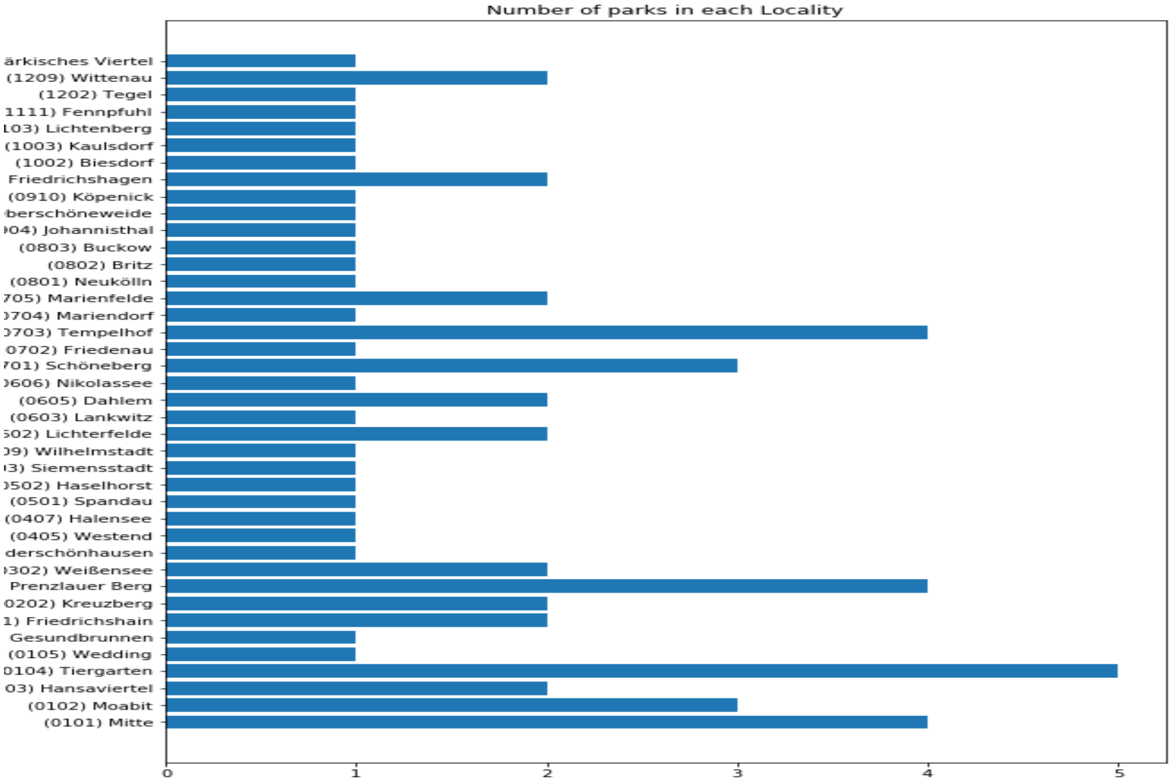
3. METHODOLOGY

3.1. EXPLORATORY DATA ANALYSIS

It was found that out of total 96 Localities 40 Localities have parks and 56 Localities don't. The localities with no parks were further analyzed based on their population, 5 localities were found which had more than 70,000 residents each and no parks.



Further Localities with parks were explored to find the distribution of parks in different localities.

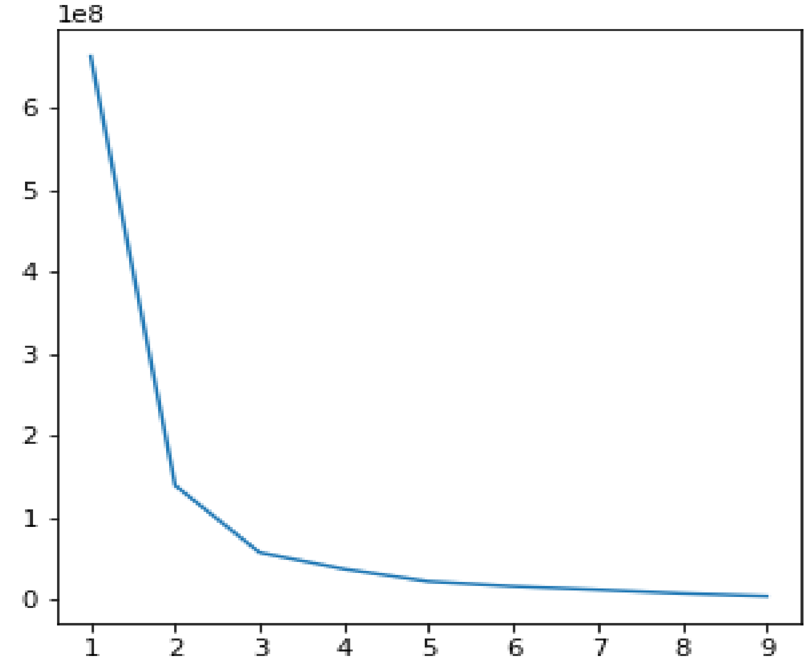


So 66 parks were found to be distributed over 40 localities with the number of parks ranging from 1 to 5 per locality.

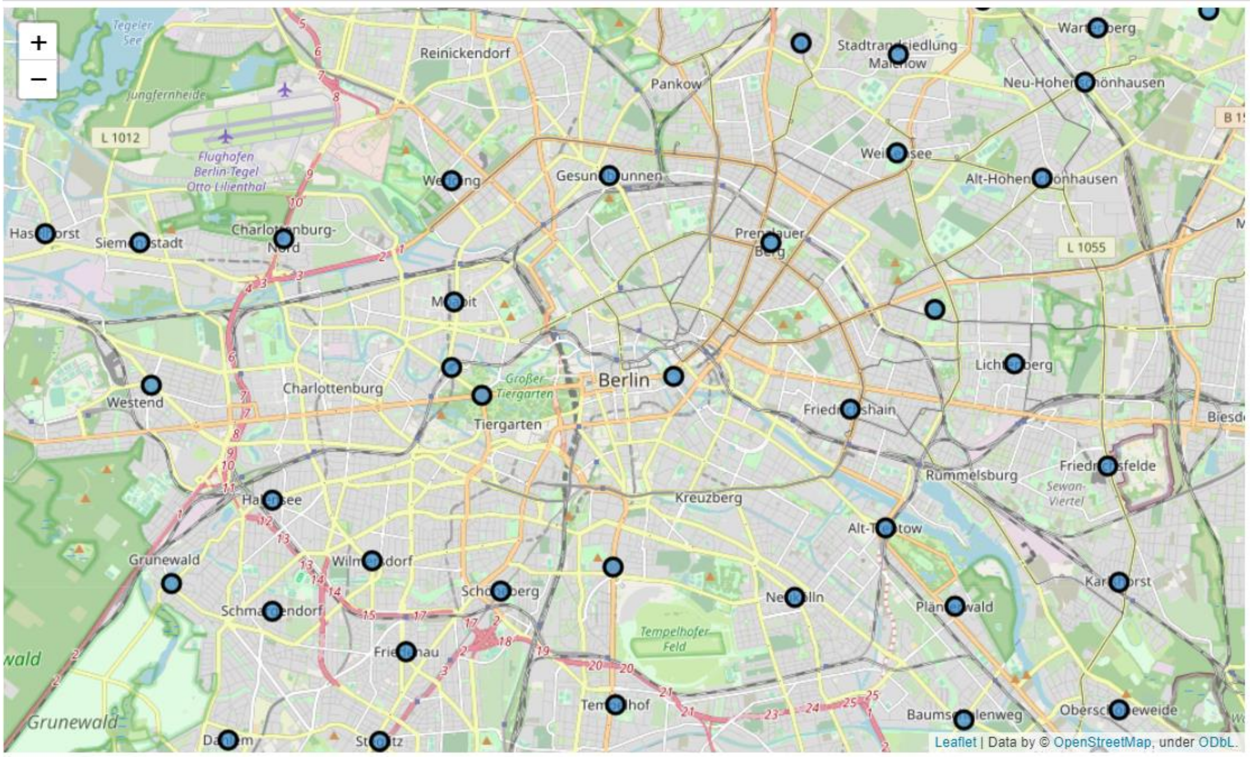
3.1. CLUSTERING

Clustering was used for summary generation. Partition based clustering was used to segment all localities into different segments depending on the area of the locality and population to parks ratio i.e. finding if sufficient parks are available depending on the number of residents of the locality.

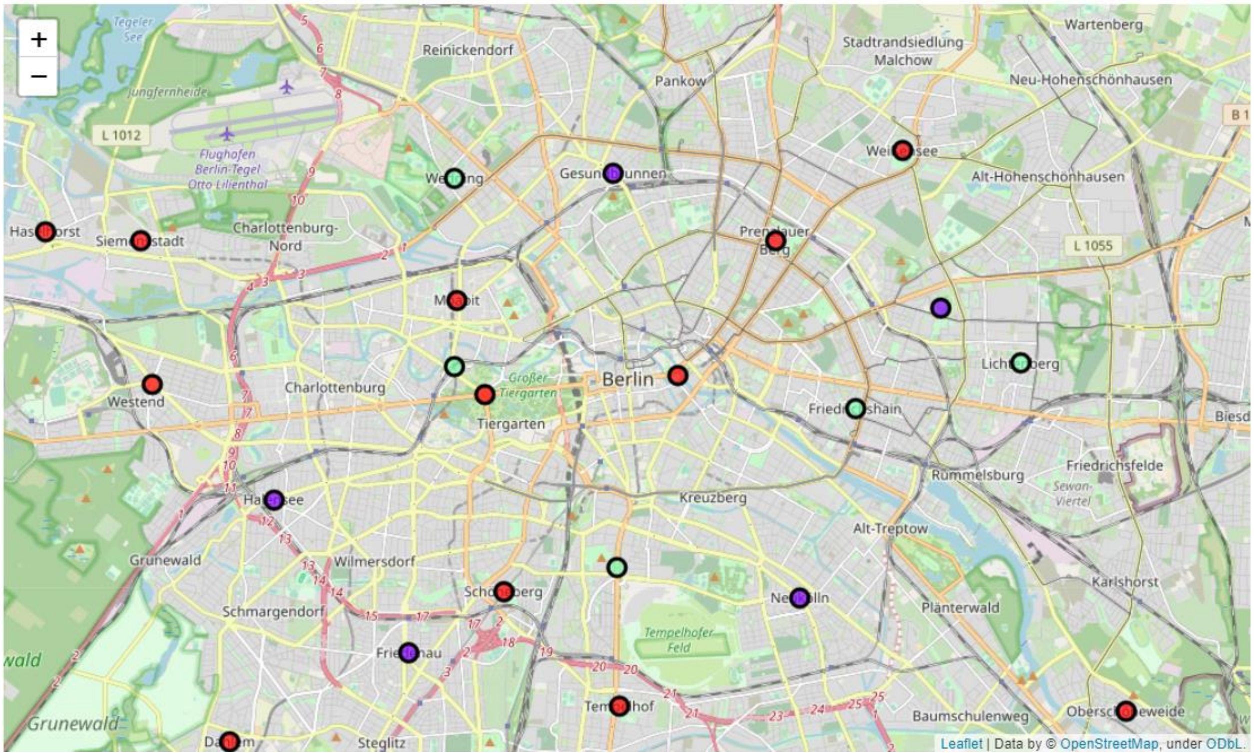
k-means algorithm was used for clustering. To choose the k value, k ranging from 1 to 10 was considered and distance from centroids was calculated and plotted to find the elbow point, therefore, k=3 was chosen.



All localities which have at least one were then assigned clusters based on the area of locality, population and number of parks present.



The below figure show the clustered localities:



4. RESULTS

Hence all localities were finally segmented into different groups according to the distribution of parks. Four segments are formed which include parks with no localities at all and localities with good, average and excellent conditions as far as the number of parks depending on the residents is concerned.

5. OBSERVATION AND RECOMMENDATION

i. Cluster 0 has a total of 24 Localities and these can be categorized as excellent localities based on the number parks present for the residents in these localities i.e. according to the number of residents sufficient parks are present and do not require immediate attention as far as building more parks is concerned. But Maintaining these parks is a must.

ii. Cluster 1 has a total of 6 Localities and these can be categorized as good localities due to the small area of these localities as parks are present in all of these but they are not sufficient for the residents. Also, these are densely populated regions with a smaller area and bigger population so due to smaller size many parks cant be built but a few more parks are a must.

iii. Cluster 2 has a total of 10 Localities and these can be categorized as average localities based on the number parks present for the residents in these localities the area of these localities is sufficiently moderate and the number of parks is low compared to the number of residents.

iv. Lastly, there are 56 localities with No parks and out of these 5 localities have a population of more than 70,000 these need urgent efforts in planning and building of parks.

6. CONCLUSION

Hence building more parks has the following priority according to the localities:

i. First, the localities with no parks at all also, among them the top 5 with a huge population.

ii. Second priority is cluster 2 as the area is sufficient to build more parks and population to parks ration is moderately high.

iii. Third will be cluster 1 due to population to parks ration is high but the area of these localities is small

iv. Fourth are the 24 localities in cluster one having an optimum number of parks!!

7. REFERENCES

[1] "impact of public parks on human life: a case study"
https://www.researchgate.net/publication/299018101_IMPACT_OF_PUBLIC_PARKS_ON_HUMAN_LIFE_A_CASE_STUDY

[2]"Parks make great places, but not enough Americans can reach them"
https://www.brookings.edu/blog/the-avenue/2019/08/21/parks-make-great-places-but-not-enough-americans-can-reach-them/