Capstone Project - Battle of the Neighborhoods in Gauteng, South Africa

Luka Beverin

February 2020

1 Introduction

1.1 Background

Relocating doesn't only mean to find a new house, you have to find the perfect neighborhood that suites your needs. This is no small feat, especially if you're moving to a new area and can't assess the various aspects that will affect your life. When moving homes, the typical individual will look at nearby schools, market price, hospitals and entertainment. However, the most important feauture of any neighborhood should be safety. If you don't feel safe in your own home, you're not going to be able to enjoy living there. For this reason it important to research these factors before planning our next move in life.

1.2 Problem statement

The crime statistics of South Africa rank amongst the worst in the World. South Africa's reported crimes are dominated by murder, gender-based violence and hijacking of vehicles. Unfortunately, this makes South Africa one of the most dangerous countries to live in and special caution should be taken when choosing an area to live in. Many people do not know the prevalence of a crime in a certain neighborhood until it's too late. On the brighter side, South Africa is an extremely vibrant and diverse country that is waiting to be explored. For example, the Maboneng Precint in central Johannesburg is filled with restaurants, rooftop bars and art exhibits. All factors should be considered when looking to relocate to a new neighborhood.

This project aims to cluster neighborhoods in Gauteng, South Africa using k-means clustering. These cluster will make use of data on total crimes in each area and the 10 most common venues in each neighborhood. Once we have segmented the neighborhoods, we will analyse the top 5 safest areas with respect to the 10 most common attractions nearby. This report will be targeted at people who are looking to relocate to Gauteng, South Africa.

1.3 Interest

This project is geared towards individuals and families who are looking to relocate to Gauteng or are purely interested in the crime statistics and venues of their own neighborhood.

2 Data Acquisition and Preparation

Based on the definition of our problem, factors that will influence our decision are the total number of crimes committed in each of the areas during the latest year available and the most common venues in each of the neighborhoods in Gauteng.

2.1 Data Acquisition

Following data sources will be needed to extract/generate the required information:

1. Preprocessing a real world data set from Kaggle showing the South African Crimes from 2005 to 2016: A dataset consisting of the crime statistics of each Ploice Station in South Africa is obtained from Kaggle. The Kaggle data set also contains shape files with geolocations of the Police Stations.
2. Creating a data set of the venues which are in each neighborhood of Gauteng: Using the Foursquare API to extract venues in neighborhoods based on their co-ordinates.

Part 1: Preprocessing a real world data set from Kaggle showing the South African Crimes from 2005 to 2016.

About this file:

- **Province:** The Province name in South Africa
- **Station:** The name of the Police Station that the crime is being reported to.
- **Category:** The type of crime being reported
- **2005 -2016:** Years with reported crime counts

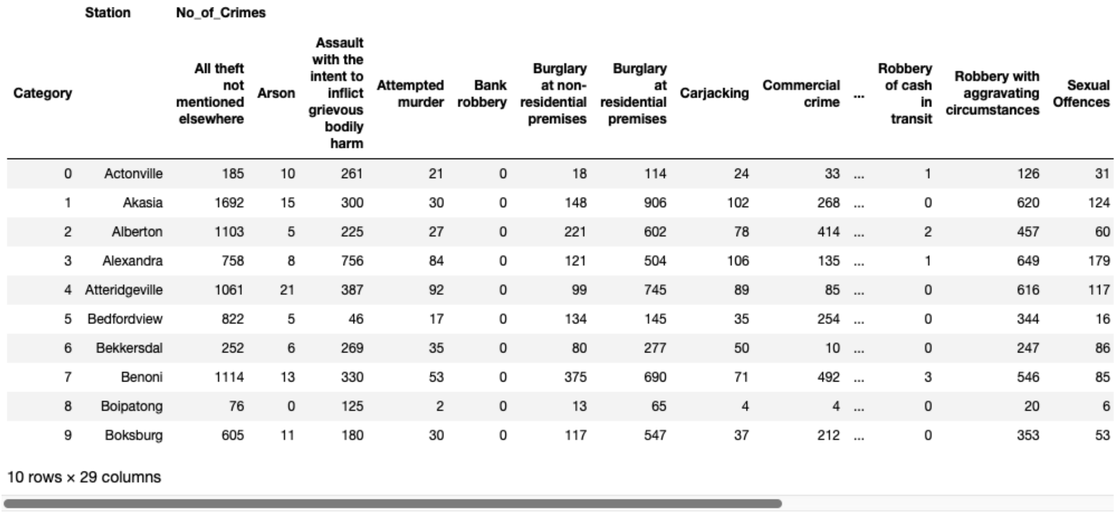
Data set URL: https://www.kaggle.com/slwessels/crime-statistics-for-south-africa

Part 2: Extract venues in each neighborhood using the Foursquare API. The Foursquare credentials are obtained through a developer account, which can be accessed via their website.

2.2 Data Preperation

Part 1

Firstly, using the Pandas library, we read in the data that we have downloaded from Kaggle. This file does not contain any geographical data, only Provinces, Stations and number of crimes reported per year. We are only interested in the last year available, so we only retain the latest year which is 2015-2016. This is done by dropping columns specified by their column number. Then we delete all empty rows. For this project, we are only interested in the Gauteng Province, thus we keep location where the Province is equal to Gauteng. Now that we have a data frame for the year 2015-2016 and Gauteng Province, we change the last column name to 'No_of_crimes' in order to clarify the data set. Next we pivot the table to view the number of crimes for each category of crime reported at each Police station in Gauteng.



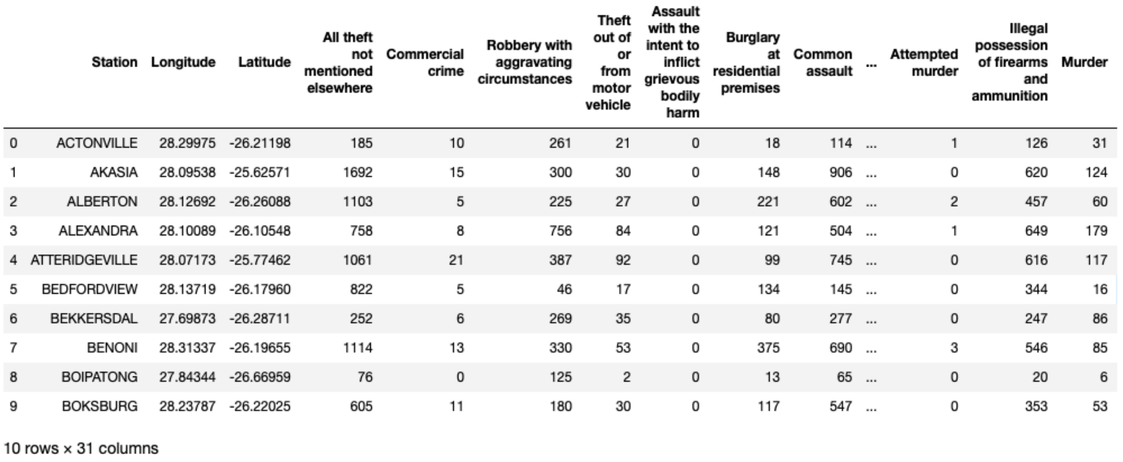
	Station	No_of_Crimes												
Category		All theft not mentioned elsewhere	Arson	Assault with the intent to inflict grievous bodily harm	Attempted murder	Bank robbery	Burglary at non-residential premises	Burglary at residential premises	Carjacking	Commercial crime	...	Robbery of cash in transit	Robbery with aggravating circumstances	Sexual Offences
0	Actonville	185	10	261	21	0	18	114	24	33	...	1	126	31
1	Akasia	1692	15	300	30	0	148	906	102	268	...	0	620	124
2	Alberton	1103	5	225	27	0	221	602	78	414	...	2	457	60
3	Alexandra	758	8	756	84	0	121	504	106	135	...	1	649	179
4	Atteridgeville	1061	21	387	92	0	99	745	89	85	...	0	616	117
5	Bedfordview	822	5	46	17	0	134	145	35	254	...	0	344	16
6	Bekkersdal	252	6	269	35	0	80	277	50	10	...	0	247	86
7	Benoni	1114	13	330	53	0	375	690	71	492	...	3	546	85
8	Boipatong	76	0	125	2	0	13	65	4	4	...	0	20	6
9	Boksburg	605	11	180	30	0	117	547	37	212	...	0	353	53

10 rows × 29 columns

Figure 1: Data frame of crimes in Gauteng

Finally we calculate the total crimes for each Station in Gauteng (create a new column at the end of the table called 'Total'). The end result is visible in the figure above. The next step in the data cleaning/preperation phase is to remove the multi index so that it will be easier to merge tables.

Now we would like to read the dbf file from Kaggle which contains the geolocations of the Police stations in South Africa. The goal is to merge the two tables such that we have a new table which contains the Stations, reported crimes and the co-ordinates of the Police stations. These co-ordinates will be helpful in visualising the clusters. Reading in the dbf file is done by the using the 'dbfread' library. Once we have the table, we rename the columns and drop columns that are not needed. We are only left with the Police Stations and their respective co-ordinates. Now we merge the two data frames so that we have the crimes for each station, as well as the geographical co-ordinates.



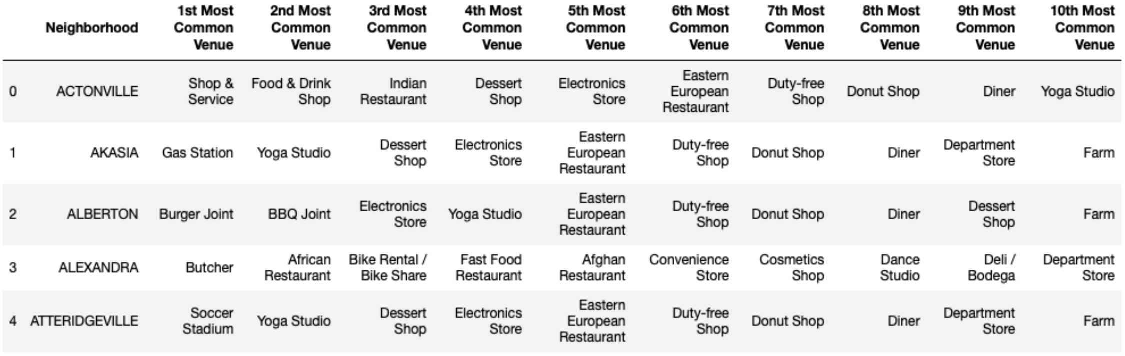
	Station	Longitude	Latitude	All theft not mentioned elsewhere	Commercial crime	Robbery with aggravating circumstances	Theft out of or from motor vehicle	Assault with the intent to inflict grievous bodily harm	Burglary at residential premises	Common assault	...	Attempted murder	Illegal possession of firearms and ammunition	Murder
0	ACTONVILLE	28.29975	-26.21198	185	10	261	21	0	18	114	...	1	126	31
1	AKASIA	28.09538	-25.62571	1692	15	300	30	0	148	906	...	0	620	124
2	ALBERTON	28.12692	-26.26088	1103	5	225	27	0	221	602	...	2	457	60
3	ALEXANDRA	28.10089	-26.10548	758	8	756	84	0	121	504	...	1	649	179
4	ATTERIDGEVILLE	28.07173	-25.77462	1061	21	387	92	0	99	745	...	0	616	117
5	BEDFORDVIEW	28.13719	-26.17960	822	5	46	17	0	134	145	...	0	344	16
6	BEKKERSDAL	27.69873	-26.28711	252	6	269	35	0	80	277	...	0	247	86
7	BENONI	28.31337	-26.19655	1114	13	330	53	0	375	690	...	3	546	85
8	BOIPATONG	27.84344	-26.66959	76	0	125	2	0	13	65	...	0	20	6
9	BOKSBURG	28.23787	-26.22025	605	11	180	30	0	117	547	...	0	353	53

10 rows × 31 columns

Figure 2: Data frame including co-ordinates

Part 2

The goal of this section of the data preparation phase is to start utilizing the Foursquare API to explore the neighborhoods and segment them. First, we create the GET request URL. We name our URL 'url' and use our personal Foursquare API credentials. Next we borrow the 'get_category_type' function from the Foursquare lab. We are then ready to clean the json and structure it into a pandas data frame. After creating a function that repeats the same process for all the Stations in Gauteng, we can analyse each station/area. Next, we group rows by area/neighborhood and by taking the mean of the frequency of occurrence of each category. Finally, we create a new data frame that displays the top 10 venues for each neighborhood. The final data frame generated by the Foursquare API is shown below.



	Neighborhood	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	ACTONVILLE	Shop & Service	Food & Drink Shop	Indian Restaurant	Dessert Shop	Electronics Store	Eastern European Restaurant	Duty-free Shop	Donut Shop	Diner	Yoga Studio
1	AKASIA	Gas Station	Yoga Studio	Dessert Shop	Electronics Store	Eastern European Restaurant	Duty-free Shop	Donut Shop	Diner	Department Store	Farm
2	ALBERTON	Burger Joint	BBQ Joint	Electronics Store	Yoga Studio	Eastern European Restaurant	Duty-free Shop	Donut Shop	Diner	Dessert Shop	Farm
3	ALEXANDRA	Butcher	African Restaurant	Bike Rental / Bike Share	Fast Food Restaurant	Afghan Restaurant	Convenience Store	Cosmetics Shop	Dance Studio	Deli / Bodega	Department Store
4	ATTERIDGEVILLE	Soccer Stadium	Yoga Studio	Dessert Shop	Electronics Store	Eastern European Restaurant	Duty-free Shop	Donut Shop	Diner	Department Store	Farm

Figure 3: Data frame of popular venus

3 Exploratory Data Analysis

3.1 Areas with highest amount of crime

The most dangerous areas in Gauteng are located around Jhb Central, Honeydew, Pretoria Central, Hillbrow and Sandton. The crimes reported in Jhb Central are substantially higher than Hillbrow and Sandton, even though both are ranked in the top 5 most dangerous areas. Families with young children should avoid these areas.

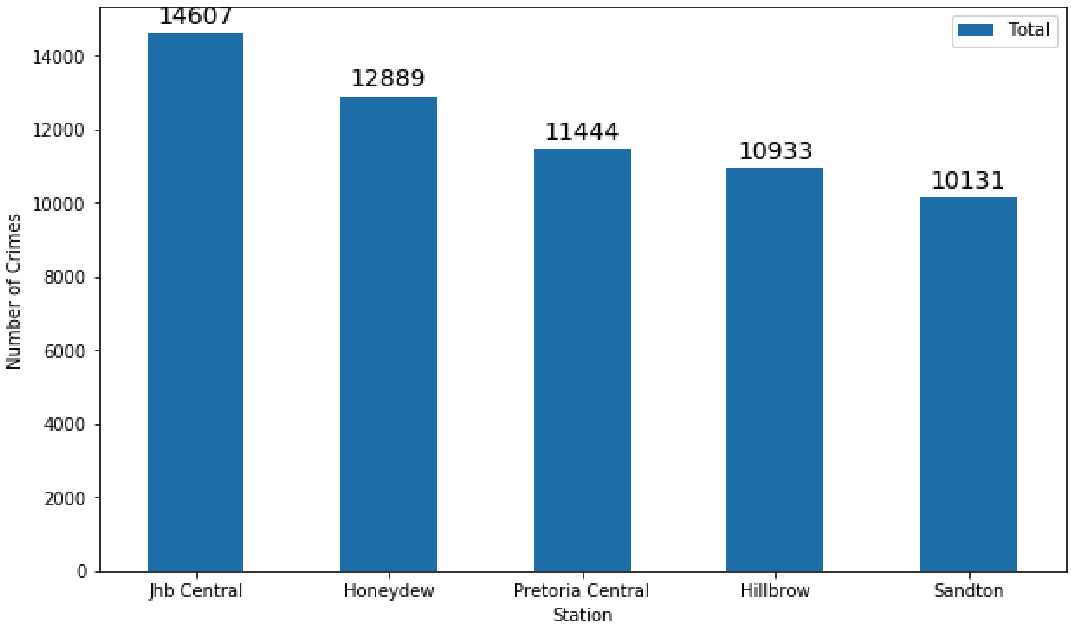
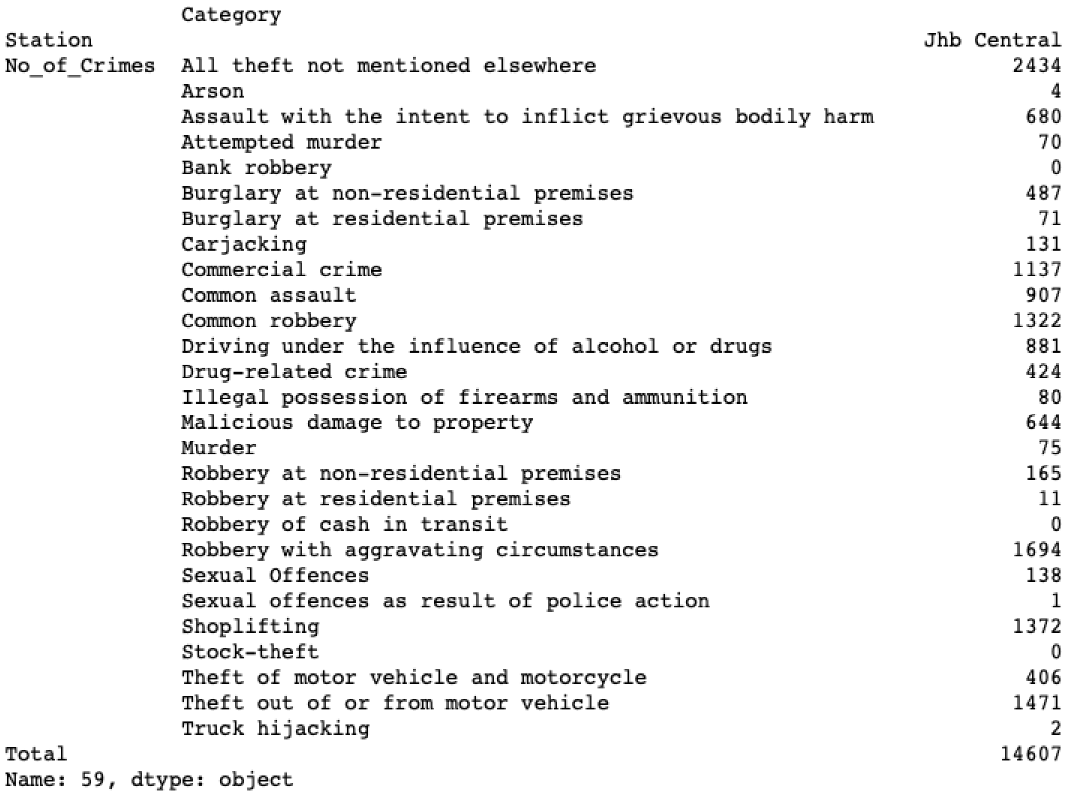


Figure 4: Gauteng stations with the highest no. of crime

Johannesburg is the robbery capital of South Africa and the most dangerous metro in Gauteng. We have a look at the summary of Jhb Central crimes to understand which types of crimes are most prevalent. The most common crimes are robbery with aggravating circumstances, theft, shoplifting and common robbery. It is not yet quite understood why Johannesburg faces such high crime rates, but it is believed that inequality is key driver, with the metro facing rapid population growth. These are factors that need to be considered by individuals and families before relocating to a neighborhood. Surprisingly, zero bank robberies were reported. This may be due to robberies being reported to a different station or errors in data handling.



```
             Category
Station                                                             Jhb Central
No_of_Crimes All theft not mentioned elsewhere                             2434
             Arson                                                            4
             Assault with the intent to inflict grievous bodily harm        680
             Attempted murder                                                70
             Bank robbery                                                     0
             Burglary at non-residential premises                           487
             Burglary at residential premises                                71
             Carjacking                                                     131
             Commercial crime                                              1137
             Common assault                                                 907
             Common robbery                                                1322
             Driving under the influence of alcohol or drugs                881
             Drug-related crime                                             424
             Illegal possession of firearms and ammunition                   80
             Malicious damage to property                                   644
             Murder                                                          75
             Robbery at non-residential premises                            165
             Robbery at residential premises                                 11
             Robbery of cash in transit                                       0
             Robbery with aggravating circumstances                        1694
             Sexual Offences                                                138
             Sexual offences as result of police action                       1
             Shoplifting                                                   1372
             Stock-theft                                                      0
             Theft of motor vehicle and motorcycle                          406
             Theft out of or from motor vehicle                            1471
             Truck hijacking                                                  2
Total                                                                     14607
Name: 59, dtype: object
```

Figure 5: Summary of crimes in Jhb Central

3.2 Areas with Lowest Amount of Crime

The least dangerous areas in Gauteng are located around Vaal Marina, Wedela, Hekpoort, Devon and Dube. The Vaal Marina has an exceptionally low crime rate. This could be due to a small population or a strict policing system. If an individual had to base their relocation decision on the area with the lowest crime rate, their choice without a doubt should be the Vaal Marina.

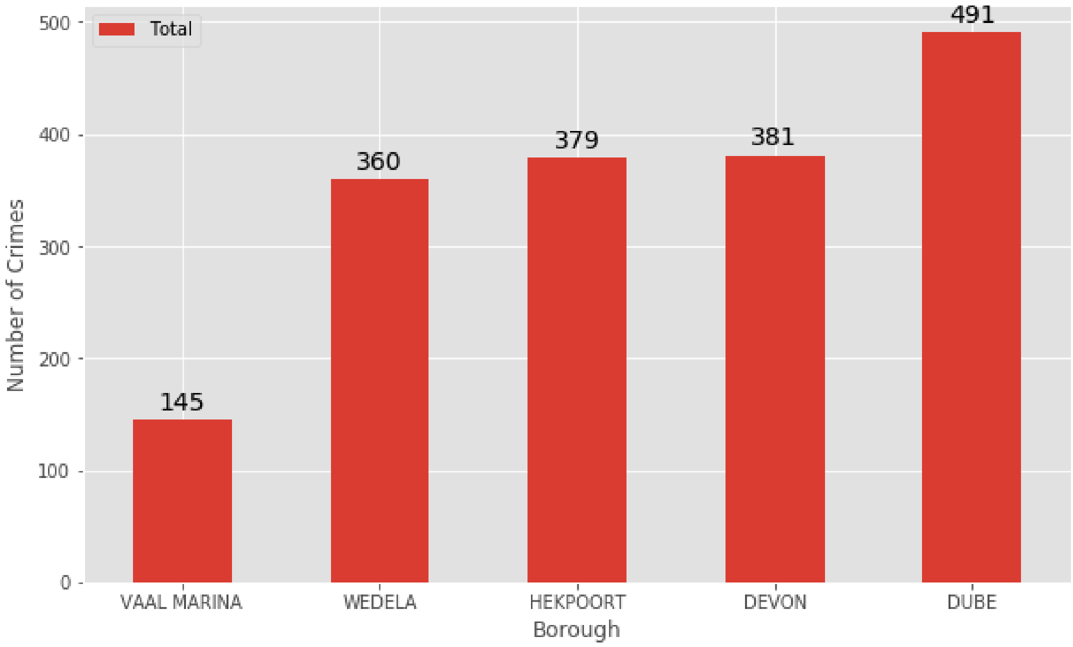


Figure 6: Gauteng stations with the lowest no. of crime

4 K-means Clustering

K-means algorithm is an iterative algorithm that tries to partition the data set into K pre-defined distinct non-overlapping subgroups (clusters) where each data point belongs to only one group. It tries to make the inter-cluster data points as similar as possible while also keeping the clusters as different (far) as possible. It assigns data points to a cluster such that the sum of the squared distance between the data points and the cluster's centroid (arithmetic mean of all the data points that belong to that cluster) is at the minimum. The less variation we have within clusters, the more homogeneous (similar) the data points are within the same cluster.
We run k-means to cluster the neighborhood into 5 clusters.

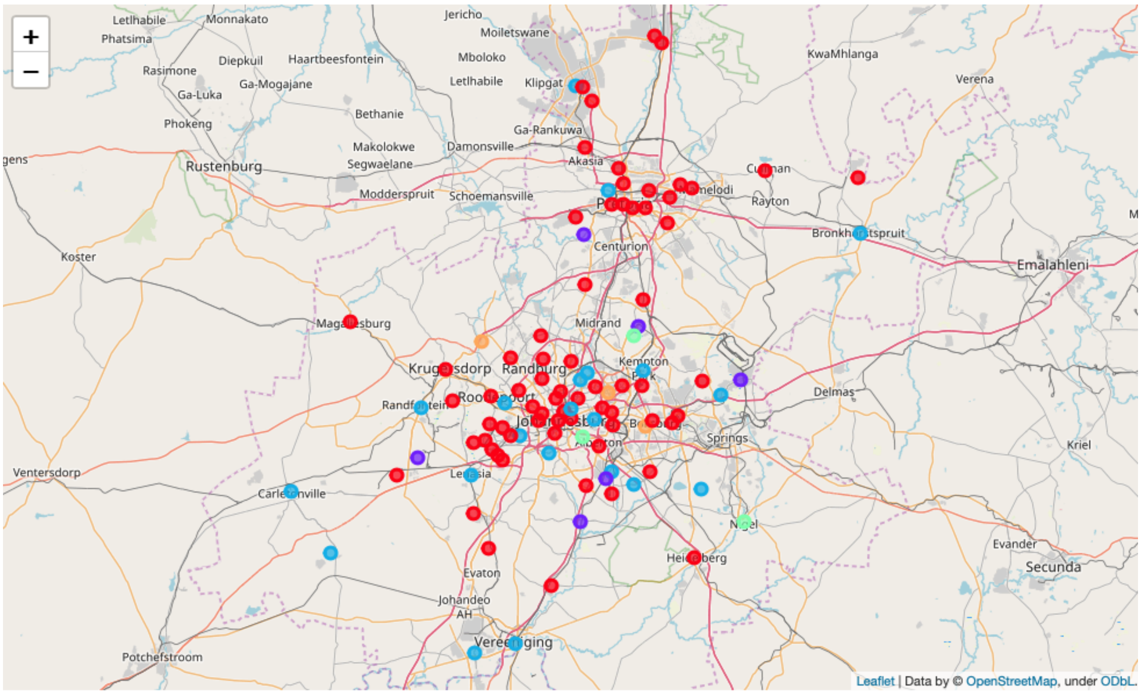


Figure 7: Cluster of neighborhoods

Cluster 1

The first cluster has the by far the most neighborhoods with 68.

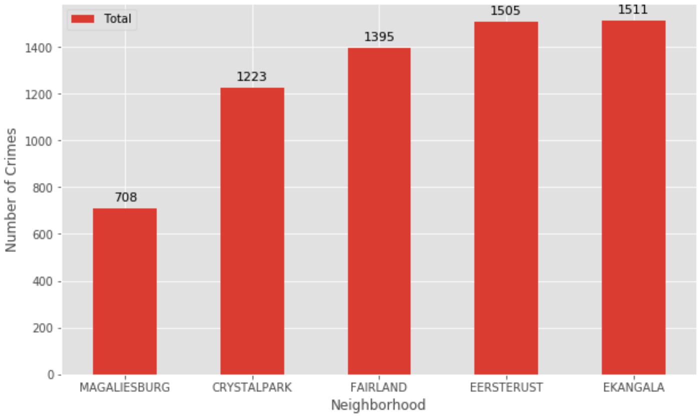


Figure 8: Neighborhoods with the least no. of crime in cluster 1

Magaliesburg is the safest area in cluster 1 and has popular venues such as arts and crafts store, yoga studio, dessert shop and electronics store.

Cluster 2

The second cluster only has 6 neighborhoods (purple dot).

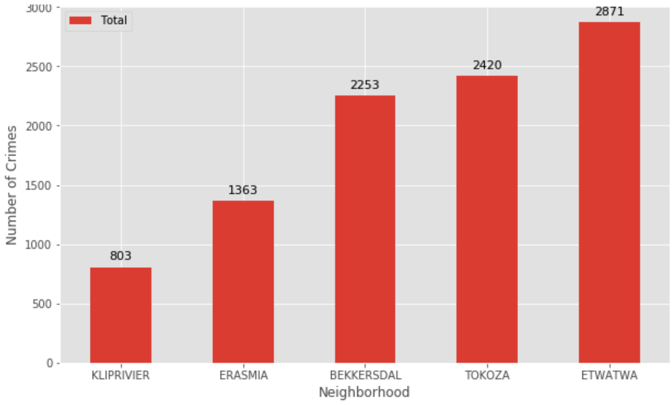


Figure 9: Neighborhoods with the least no. of crime in cluster 2

Kliprivier is the safest area in cluster 2 and has popular venues such as a burger place, convenience store, yoga store and electronics store.

Cluster 3

The third cluster has 21 neighborhoods (blue dot).

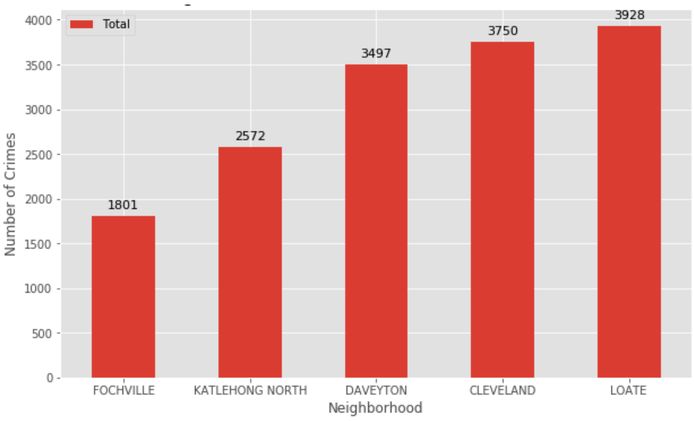


Figure 10: Neighborhoods with the least no. of crime in cluster 3

Fochville is the safest area in cluster 3 and has popular venues such a fast food restaurant, lawyer and yoga studio.

Cluster 4

The fourth cluster has 3 neighborhoods (green dot). Most of the popular venues in this cluster includes repair and department stores.

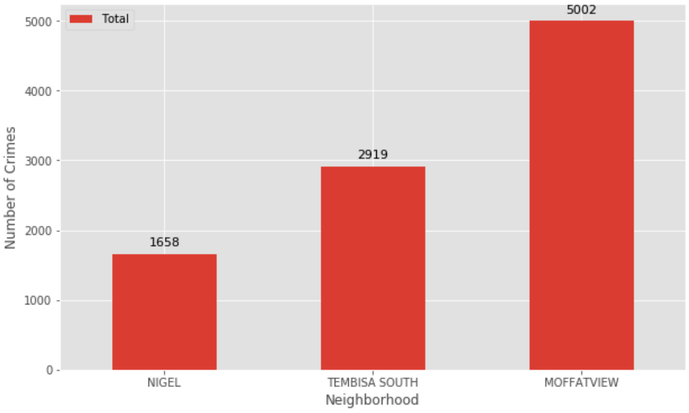


Figure 11: Neighborhoods with the least no. of crime in cluster 4

Nigel is the safest area in cluster 4 and has popular venues such as a repair shop, yoga studio and department store.

Cluster 5

The fifth cluster has 3 neighborhoods (orange dot).

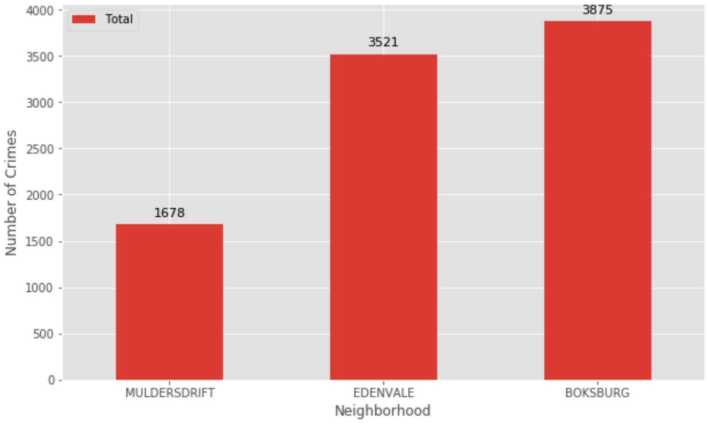


Figure 12: Neighborhoods with the least no. of crime in cluster 5

Muldersdrift is the safest area in cluster 5 and has popular venues such as a coffee Shop, dessert shop and electronics store.

5 Discussion

The aim of this project is to help individuals and families choose a neighborhood based on safety and nearby popular venues. The venues are clustered in terms of common venues and crime statistics. The first cluster contains larger neighborhoods with popular venues such as gas stations, shopping malls and restaurants. This type of neighborhood is usually where offices are located but the downside is that there are safety concern. The safest area in cluster 1 is Magaliesburg. Cluster 2 is filled with outer city neighborhoods and is associated most with convenience stores. If a person is looking for a neighborhood with lots of restaurants and entertainment then cluster 3 should be considered. The safest area in cluster 3 is Fochville. Cluster 4 is similar to cluster 2 with repair shops being popular. It is a very small cluster and Nigel is the safest neighborhood. Cluster 5 is characterised by coffee and cafe shops, which may suite a retired individual.