Choosing a suburb to stay in Johannesburg

Moloi Mokete

August 2019

1. Introduction

1.1. Background

The title of the project is *'Choosing a suburb to stay in Johannesburg'*. Johannesburg, also known as Joburg, is the largest city in South Africa and one of the 50 largest urban areas in the world. It is the provincial capital and largest city of Gauteng, which is the wealthiest province in South Africa, this makes Johannesburg attractive for immigrants, investors, and people hoping for a better life. In 1950, the population of Johannesburg was 910,550 and as of 2019, Johannesburg has a population of about 5,635,127. Immigration accounts for a significant 17% of SA's population growth. For a newcomer to the big city, it can be a daunting task to determine which part of Johannesburg to settle in depending on a person's preferences.

1.2. Problem

Deciding which suburb to move into in Johannesburg, should be trouble-free. People look for different things to determine a suitable place to stay depending on their needs, priorities and likings. Some of the factors people consider to determine neighborhoods to stay in includes crime rate, schools, healthcare availability, jobs, cost of living, etc. A platform, either mobile application or website, that allows people to screen suburbs in Johannesburg on the mentioned factors would make a huge difference and save people time , energy and money that goes into choosing the most favorable place to settle into.

At the present moment choosing the right suburb to settle into involves a lot of work and stress on individuals. This includes searching on the information internet, making calls and visiting different suburbs physically. A website, http://www.teleport.org provides information about the whole city of Johannesburg for people willing to relocate to the city but the site does not analyze individual suburbs within Johannesburg.

In response to this problem my project proposes to come up with a solution. My solution uses machine learning, clustering in particular, to analyze profiles of suburbs in Johannesburg and categories them. The project considers three factors which are crime rate, school availability and healthcare availability in each suburb. The suburbs are then categorized into three clusters, from best to worst based on crime rate, school availability and healthcare availability.

2. Data cleaning and acquisition

2.1. Data sources

The dataset for this project was built from the following sources:

1. Crime Stats SA: https://www.crimestatssa.com/index.php
2. City of Johannesburg Municipality Wikipedia Page: City of Johannesburg

3. Foursquare location data: https://foursquare.com

City of Johannesburg Municipality Wikipedia Page:

This page contains list of suburbs that are under the City of Johannesburg Municipality as well as the links to individual Wikipedia pages for each suburb, which then has coordinates of each suburb. The City of Johannesburg Municipality is divided into 7 regions (Region A – Region G) and each suburb belongs to a particular region.

For example below is the screenshot of the Alexandra suburb Wikipedia page:

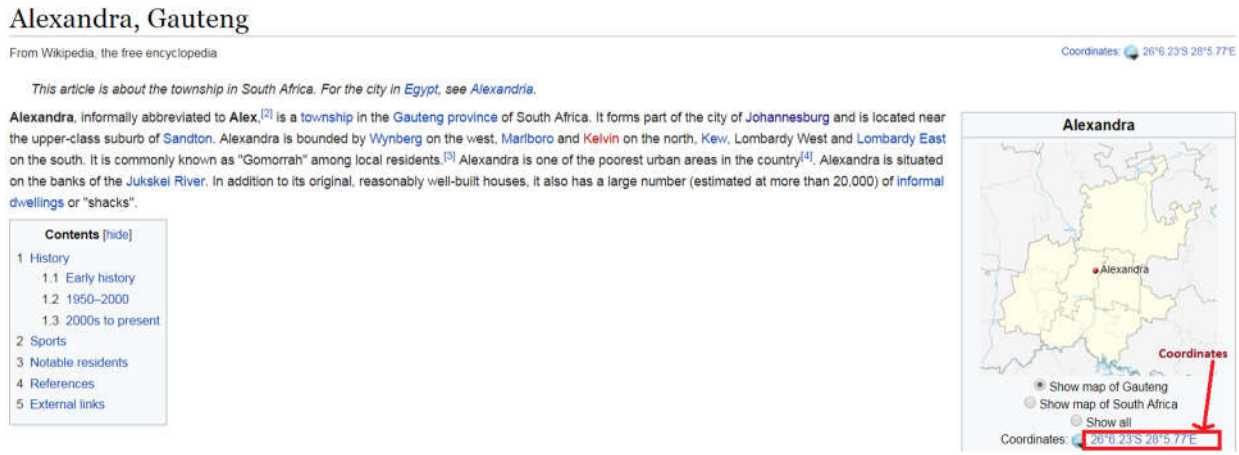

Figure 1 Wikipedia Page of Alexandra, Suburb.

The name of the suburb was be used to identify the suburb and the coordinates was used in the Foursquare API and also for visualizing the results of the project on a map of Johannesburg.

Crime Stats SA:

Crime stats is a website that displays the latest and historic South African crime statistics in an easy-to-understand format. The website has crime data for areas all over South Africa. Total crime data is a sum of 39 different crime categories for each suburb.

For example below is the screenshot, from the Crime Stats SA website, showing crime data for Sandton, Johannesburg. The crime data was used to determine the crime rate per suburb.

Precinct	Murder	Sexual Offences	Attempted murder	Assault with the intent to inflict grievous bodily harm	Common assault	Common robbery	Robbery with aggravating circumstances	Arson	Malicious damage to property
	2018	2018	2018	2018	2018	2018	2018	2018	2018
Sandton	15	59	11	107	395	206	826	0	154
Total	15	59	11	107	395	206	826	0	154

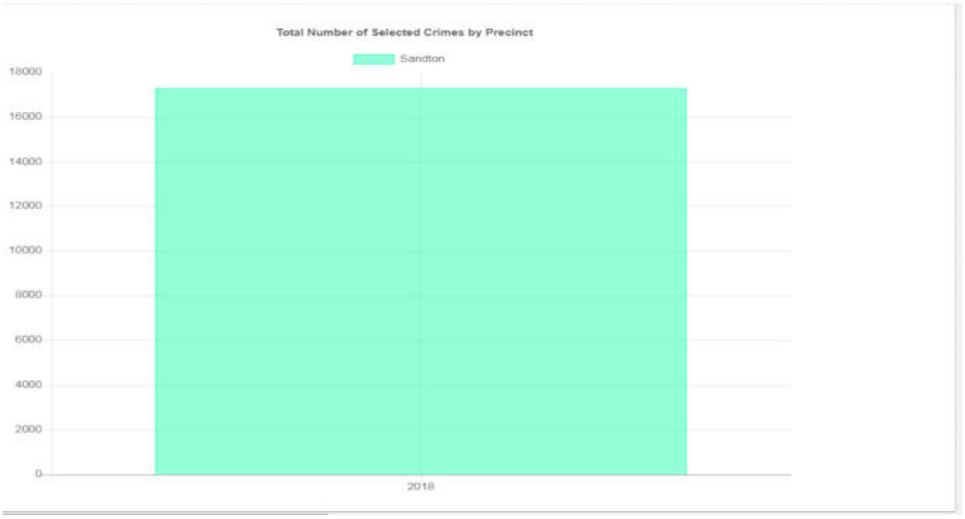

Figure 2. Total Crime, Sandton 2018

Foursquare Location Data:

Foursquare is a technology company that built a massive dataset of accurate location data. Foursquare location data was used to find number of available schools and healthcare facilities within a determined radius for each suburb. For schools I concentrated on primary schools, elementary schools, high schools and colleges, but for healthcare facilities I considered all healthcare establishments within a suburb.

2.2. Data Cleaning

Data scraped from https://en.wikipedia.org/wiki/City_of_Johannesburg_Metropolitan_Municipality) and https://www.crimestatssa.com/index.php was combined into one table. The problem with the dataset was that some suburbs from the City of Johannesburg dataset did not appear in the Crime Stats SA dataset, so I decided to only keep suburbs that appear in both datasets. The number of suburbs that remained was 29.

3. Methodology

The dataset before using Foursquare API to get number of schools and healthcare facilities per suburbs had components; '*Suburb*', *'Total Crimes', 'Region', 'Latitude' and 'Longitude'.*

	Suburb	CrimeRate	Schools	Healthcare	Latitude	Longitude
0	Alexandra	12104.0	4	3	-26.103833	28.096167
1	Bramley	7357.0	8	22	-26.124167	28.081667
2	Brixton	8290.0	4	42	-26.183333	28.000000
3	Diepkloof	10179.0	3	14	-26.249000	27.946000
4	Diepsloot	9723.0	2	0	-25.934722	28.012500

Figure 3. Dataset after scraping data from the internet.

I then utilized the Foursquare API to find the number of schools and healthcare facilities in each suburb. I used the *'categoryID'* attribute in the Foursquare API URL to limit search for schools venues to primary schools, elementary schools, high schools and colleges as well as to search for all healthcare venues without limitation in each suburb. I designed the limit as 100 venues and radius as 2000m for each suburb. Below is the list of each suburb, along with crime rate, number of available schools and healthcare facilities for the suburb.

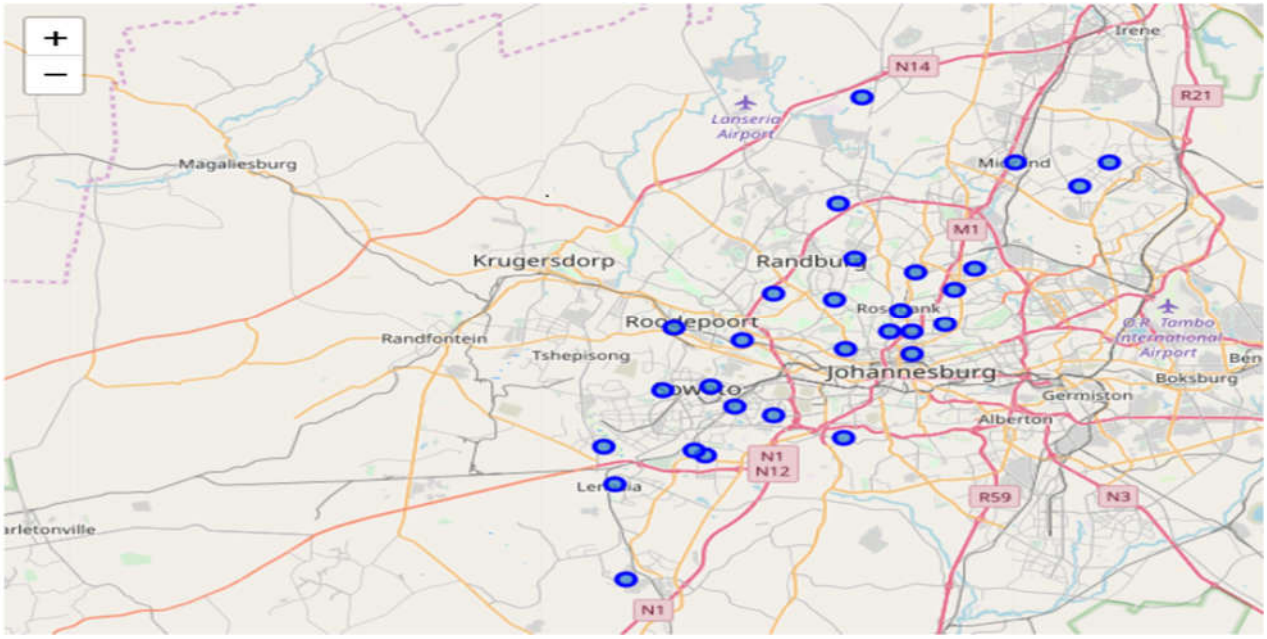

Figure 4. Map of Johannesburg with suburbs superimposed on top.

From the graphs below it is shown that Region D suburbs have the highest crime rate, while Region G suburbs have the lowest crime rate. It is also shown that Region B has the highest number of schools and healthcare facilities in its suburbs. Suburbs in Region A and Region G have very few schools and healthcare facilities.

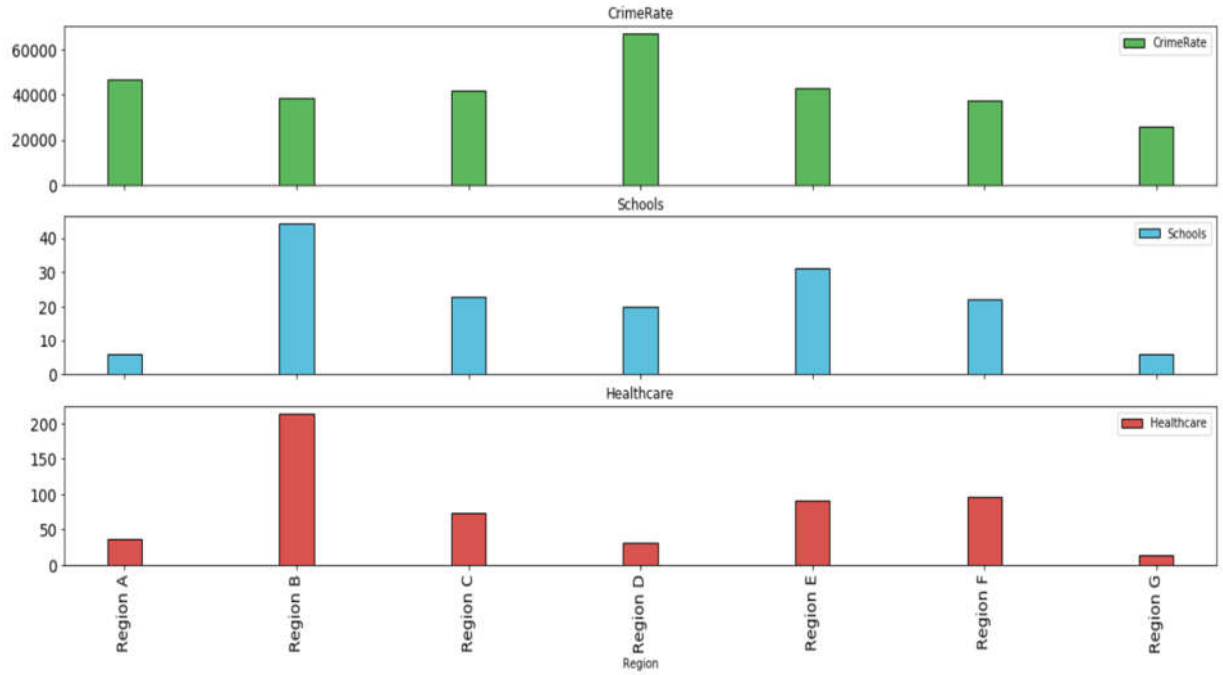

Figure 5. Comparison of Regions

Looking at the suburbs, it can be observed that suburbs like Hillbow and Sandton have very high number of schools and healthcare facilities as well as a high crime rate. It is particularly interesting to see that Sandton, regarded as one of the richest suburbs in South Africa, has a very high crime rate.

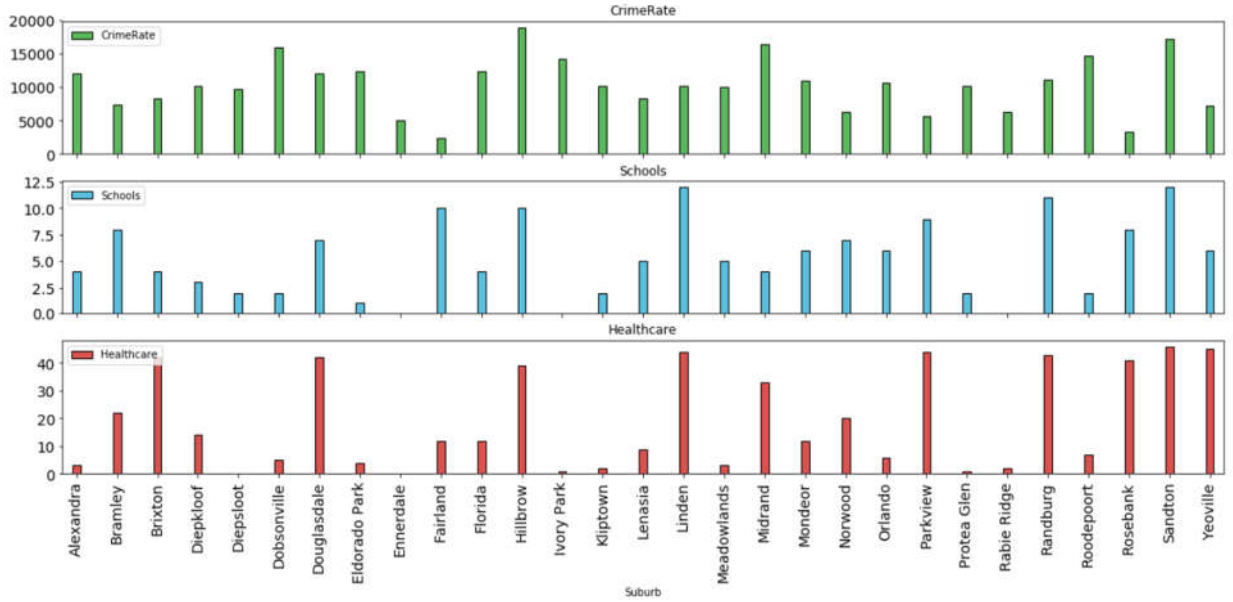

Figure 6. Comparisons of Suburbs.

Clustering:

I applied an unsupervised learning K-means algorithm to cluster the suburbs which had common characteristics based on crime rate, schools and healthcare availability. The K-means cluster algorithm used has three clusters because when I analyzed the K-Means with elbow method it ensured me the 3 degree for optimum k of the K-Means.

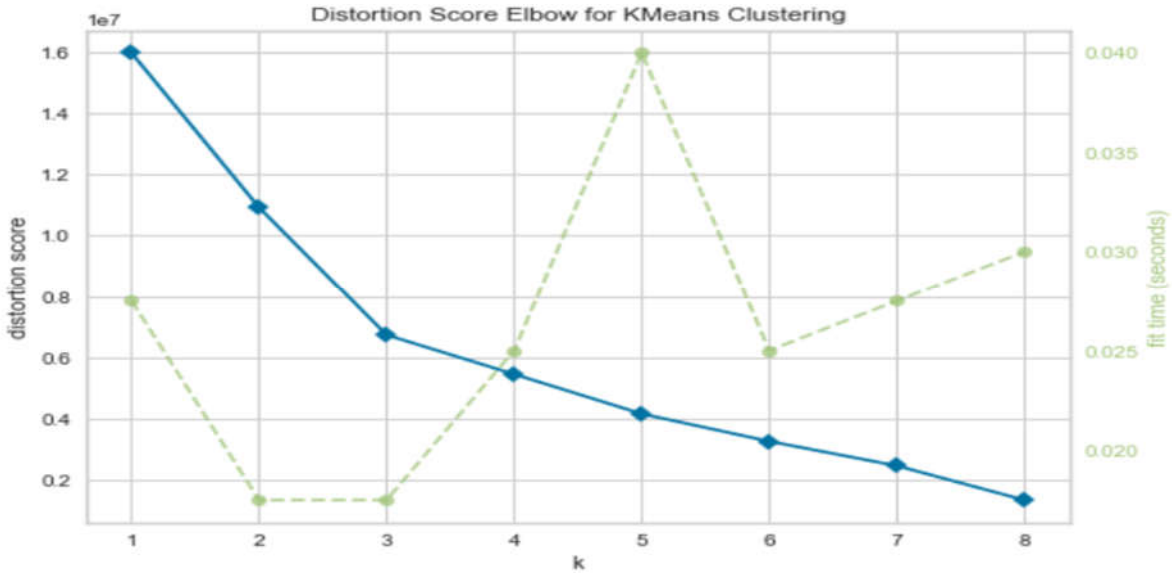

Figure 7. Elbow method to detrmine optimum k for k-means

4. Results

The table below shows the merged dataset of the suburbs with the clusters as well.

	Cluster Labels	Suburb	CrimeRate	Schools	Healthcare	Latitude	Longitude
0	0	Alexandra	12104.0	4	3	-26.103833	28.096167
1	2	Bramley	7357.0	8	22	-26.124167	28.081667
2	2	Brixton	8290.0	4	42	-26.183333	28.000000
3	0	Diepkloof	10179.0	3	14	-26.249000	27.946000
4	0	Diepsloot	9723.0	2	0	-25.934722	28.012500

Figure 8. Table showing clusters along with suburb data.

I then averaged the features in each cluster so as to compare the centroids of each feature per cluster. The bar chart shows that we can label the clusters as follows:

- Cluster 0: Suburbs with 'High crime rate, low availability of schools and healthcare facilities'
- Cluster 1: Suburbs with 'High availability of schools and healthcare facilities but also very high crime rate'

- Cluster 2: Suburbs with ‘High availability of schools and healthcare facilities, and also very low crime rate’.

Figure 9. Cluster comparisons.

I also used python folium library to visualize the clusters of the suburbs with the suburbs superimposed on top of the map Johannesburg. The red circles represent Cluster 0, purple ones Cluster 1 and green ones Cluster 2. It can be observed that most suburbs fall under Cluster 0. An interesting observation is also that Cluster 1 and Cluster 2 seem to be centered on the city center.

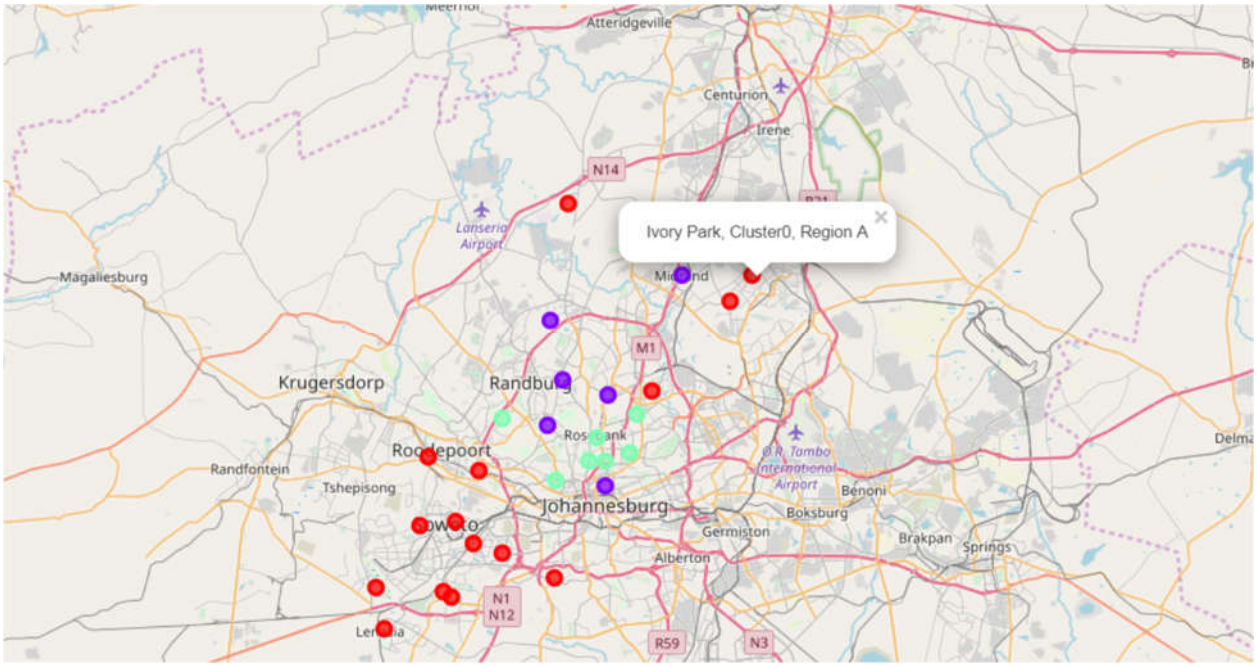

Figure 10. Map showing which clusters each suburb belongs to.

5. Discussion

The most difficult challenge I was faced with was that the dataset that I needed for the project was not readily available and I had to scrape a few websites to build the dataset required for the project. However, this presented another problem as the data from different websites was not synchronized e.g. areas regarded as suburbs by the City of Johannesburg Municipality were not all present in the Crime Stats SA websites, and such suburbs had to be dropped from the dataset, leaving only 29 suburbs. This greatly reduced the number of suburbs that were analyzed. Using the same suburbs as designated by City of Johannesburg Municipality to record crime data by Crime Stats SA would substantially increase number of suburbs in the dataset and provide wider coverage of the city.

For clustering I used the K-means algorithm because it is computationally faster than hierarchical clustering if k remains small and also because it produces tighter clusters. I used the elbow method to determine the optimum value of k, and set the k value to 3. I believe more accurate optimization of the algorithm can be attained with the increase of number of suburbs analyzed.

On the way forward, the model can be made more compressive and broad by adding more features to cluster suburbs including job availability, housing prices and rent, traffic, internet access, outdoor activities etc.

Furthermore an interactive web or mobile application, with maps, can be developed to help Johannesburg newcomers make informed decision about where they would like to live, improving on visualization map of Johannesburg developed in this project. This could made further useful by advertising available housing in an area of interest.

6. Conclusion

It is estimated that 6.5 million people will be living in Johannesburg in 2040, a population growth of about 66% from now. As a result this project could save new people coming to Johannesburg a lot of time, money and stress that goes into choosing a neighborhood to stay in depending on each individual's preferences.

Moreover the project could also help the city authorities to predict suburbs that are likely to be more populated. This will help to strategize on how the suburbs in the city are developed to cater and accommodate increasing population.