# Part 1 - Project Description

This project demonstrates an analysis of neighborhoods using heterogeneous data sources and data science methods by the example of Port Elizabeth And Pretoria City In South Africa. This requires the extraction, load, transformation and analysis of all data sources contained in the following notebook.

# Introduction / Business Problem

When thinking about relocating to a new city or country for work purposes or to start a new life, or to go for a holiday destination people tend to research areas before moving. This research includes population rate, average house price, school ratings, crime rates, weather conditions, recreational facilities, holiday destinations-tourism, Carnivals and Sports events/activity. etc.
Based on the above, a search engine algorithm would be an efficient tool to use that will allow users to enter cities and get the neighborhood name that best suits their lifestyle or living conditions.
In this project, we will study in detail the area classification using foursquare data and machine learning segmentation and clustering. And segment areas of two cities based on the most common places captured from Foursquare.
-This could be done using an algorithm (Using segmentation and clustering) that will perform an extensive analysis on

1. The similarities and dissimilarities between neighborhoods in the two cities of the user's search criteria, and
2. Determine which neighborhoods best suits their lifestyle.

For this project, I will be developing a recommendation system using the Port Elizabeth and Pretoria cities in South Africa as my search criteria:

## Brief Information About: Port Elizabeth and Pretoria:

Port Elizabeth and Pretoria are two major cities in South Africa Both the cities become a center of attention for residential, holiday destinations-tourism, education, job employment, shopping and sports activity. Both cities are well known in South Africa and become the top choice for local and foreign communities. Also, for the best holiday destinations in the world because of its Mediterranean climate, vibrant nightlife, Michelin Star restaurants, scenic coastal drives, staggering mountain landscape and friendly people; the latter holds the top spot as a family-friendly destination and for its wonderful beaches and water activities.

## Port Elizabeth

Port Elizabeth or The Bay[2] (Xhosa: iBhayi; Afrikaans: Die Baai [di ˈbɑːi]) is one of the major cities in South Africa; it is situated in the Eastern Cape Province, 770 km (478 mi) east of Cape Town. The city, often shortened to PE and nicknamed "The Windy City", stretches for 16 km along Algoa Bay, and is one of the major seaports in South Africa. Port Elizabeth is the southernmost large city on the African continent, just farther south than Cape Town. Port Elizabeth was founded as a town in 1820 to house British settlers as a way of strengthening the border region between the Cape Colony and the Xhosa. It now forms part of the Nelson Mandela Bay Metropolitan Municipality, which has a population of over 1.3 million.
(Source - https://en.wikipedia.org/wiki/Port_Elizabeth )

# Pretoria

Pretoria (/prɪˈtɔːriə/; Xhosa: E-Pitoli) is a city in the northern part of Gauteng province in South Africa. It straddles the Apies River and has spread eastwards into the foothills of the Magaliesberg mountains. It is one of the country's three capital cities, serving as the seat of the administrative branch of government, and of foreign embassies to South Africa. Pretoria has a reputation for being an academic city with three universities, the Council for Scientific and Industrial Research (CSIR) and the Human Sciences Research Council. The city also hosts the National Research Foundation and the South African Bureau of Standards making the city a hub for research. Pretoria is the central part of the Tshwane Metropolitan Municipality which was formed by the amalgamation of several former local authorities including Centurion and Soshanguve.
(Source - https://en.wikipedia.org/wiki/Pretoria )

Target Audience

Through this project we are expecting following people to benefit out of the findings.
People migrating city for work.
Business person looking for new location to start office etc.
Tourist.
Restaurants to finalized menu based on the type or people, their likings based on feedbacks etc.
Sports Events, Activities Organizers and many more.

# Data

From the view of the data, this project contains various data sources. and this required data can be gathered from:
Port Elizabeth and Pretoria City information, including districts and neighborhoods, can be obtained from Wikipedia:
(Source- https://en.wikipedia.org/wiki/Port_Elizabeth )

(Source -https://en.wikipedia.org/wiki/Pretoria )

Photos and Picture of Port Elizabeth and Pretoria City used for Presentation from
(Source - https://afrotourism.com/travelogue/)
First, the location data consists of The data used for this project will be acquired from Source - ( https://www.sapostalcodes.info/) . The datasets consist of the postal codes and suburb names of each city.

Second, the Foursquare API provides a database of more than 100 million places, globally. In order to obtain venues and their categories we will use Foursquare API search feature [FOURSQUARE] (Source - https://foursquare.com/ ) to collect neighborhood venue information as well as the longitude and latitude details of each suburb. Details about local venues and locality will provide insight into the qualities of a neighborhood.

# Methodology and Python libraries

From the methodical point of view, this project utilizes a collection of various data sources from web APIs like FOURSQUARE. In addition to Foursquare, various python packages will be used to create maps and machine learning models to gather further insights and provide efficient recommendations and results into our neighborhood battle project.

These packages includes:

1. Pandas - Library for Data Analysis
2. NumPy – Library to handle data in a vectorized manner
3. JSON – Library to handle JSON files
4. Geopy – To retrieve Location Data
5. Geocorder - For geolocation of neighborhoods
6. Requests – Library to handle http requests
7. Matplotlib – Python Plotting Module
8. Sklearn – Python machine learning Library
9. Folium – Map rendering Library

# Basic Work Flow followed as :

HTTP requests would be made to this Foursquare API server using postal codes of Port Elizabeth Suburbs and Pretoria Suburbs to pull out the latitude and longitude which will be used for creation of the map as well data analysis.
Using credentials Foursquare API search feature would be enabled to collect the nearby places of the suburbs. Due to http request limitations, the number of places per suburb parameter would be set to 100 and the radius parameter would be set to 700.
Folium- Python visualization library would be used to visualize the suburbs cluster distribution of Port Elizabeth and Pretoria over an interactive leaflet map.
Extensive comparative analysis of two suburbs world be carried out to derive the desirable insights from the outcomes using python's scientific libraries Pandas, NumPy and Scikit-learn.
Unsupervised machine learning algorithm K-mean clustering would be applied to form the clusters of different categories of places residing in and around the neighborhoods. These clusters from each of those two chosen suburbs would be analyzed individually collectively and comparatively to derive the conclusions.

# 3. Methodology

A Jupyter Notebook developed in order to process data and segment the neighborhoods. Following steps are implemented:

**1. Import Libraries**

The notebook requires the following libraries. And we have installed it

Pandas - Library for Data Analysis

NumPy – Library to handle data in a vectorized manner

JSON – Library to handle JSON files

Geopy – To retrieve Location Data

Geocorder - For geolocation of neighborhoods

Requests – Library to handle http requests

Matplotlib – Python Plotting Module

Sklearn – Python machine learning Library

Folium – Map rendering Library

**2. Build neighborhoods list**

A list of Suburb and Postal code information is obtained from http://www.sapostalcodes.info/ for Port Elizabeth, and Pretoria city of South Africa That list contains the names of the neighborhoods for both the cities. As output a dataset containing a list of "city, suburb" is build.

**3. Neighborhoods geolocation**

Every element in the neighborhoods dataset is geolocated using Python Geolocator and two columns are updated Containing the latitude and the longitude coordinates of each city, neighborhood. Also the Geographical coordinate of Port Elizabeth, found out.

**4. Find Geographical Coordinates and No of Suburbs**

As a next stage, the Geographical coordinate of Port Elizabeth, and Pretoria city found out. Also, the no of suburbs are found out for both the cities.

**5. Venues compilation**

As next step Foursquare services are used for obtaining venues for every neighborhood. The output is a new dataset with many records for every neighborhood containing the venues found for every one of them.

**6. Neighborhoods Segmentation**

The problem in hand is a case of unsupervised segmentation and, from the possible machine learning algorithms, K-means was choosen. Taking in account that the venues information obtained from Foursquare is categorical, it must be previously processed in order to be handled by K-means algorithm. For this _"pandas.getdummies" is used for dummies variables.

Therefore For this unsupervised machine learning algorithm K-mean clustering applied to form the clusters of different categories of places residing in and around the neighborhoods.
Next step is build the segmentation dataframe, composed of the top venues for every neighborhood plus a segment label determined by K-means.
**7. Segments analysis**

Every segment is printed individually, were different characteristics can be observed for each group. These clusters from each of those two chosen suburbs would be analyzed individually collectively and comparatively to derive the conclusions.
Next section describes the results.

## 4. Results -

## 1. Outcomes – Port Elizabeth, South Africa

The K-means method was used to cluster the suburbs of Port Elizabeth city into 5 clusters. The details of the clusters are as follows:

1. **Cluster 1** -15 Suburbs

Common Venus include Restaurants, Golf Course, Yoga, Garden, Gym ,Coffee Shops, Store, N and Shopping Mall
2. **Cluster 2** - 13 Suburbs

Common Venues include Fast Food Restaurant, Coffee Shops, Bookstores, Restaurants, Clothing Store, and Electronics Store
3. **Cluster 3 – 04** Suburbs

Common Venues include Convenience Store, Coffee Shops, Thai Restaurant, Fast Food Restaurant Pubs, Accessory Stores and Electronics Store
4. **Cluster 4** - 14 Suburbs

Common Venues include Fast Food Restaurants, Winery, Beaches and Cafes, Thai Restaurant, Grocery Store, Department Store
5. **Cluster 5** -1 Suburbs

Common Venues include Thai Restaurant, Grocery Store, Fried Chicken Joint, Fast Food Restaurant, Electronics Store, and Department Store
**Fig – Visualization - The Resulting Clusters Of Port Elizabeth City** PAGE 9

## 4. Results -

### 1. Outcomes – Port Elizabeth, South Africa

The K-means method was used to cluster the suburbs of Port Elizabeth city into 5 clusters. The details of the clusters are as follows:

1. **Cluster 1** -25 Suburbs

   Common Venus include Restaurants, Coffee Shops, Grocery Store, Nightclubs and Shopping Mall

2. **Cluster 2** - 13 Suburbs

   Common Venues include Fast Food Restaurant, Coffee Shops, Bookstores, Restaurants, Clothing Store, and Electronics Store

3. **Cluster 3** – **04** Suburbs

   Common Venues include Convenience Store, Coffee Shops, Thai Restaurant, Fast Food Restaurant Pubs, Accessory Stores and Electronics Store

4. **Cluster 4** - 14 Suburbs

   Common Venues include Fast Food Restaurants, Winery, Beaches and Cafes, Thai Restaurant, Grocery Store, Department Store

5. **Cluster 5** -1 Suburbs

   Common Venues include Thai Restaurant, Grocery Store, Fried Chicken Joint, Fast Food Restaurant, Electronics Store, and Department Store

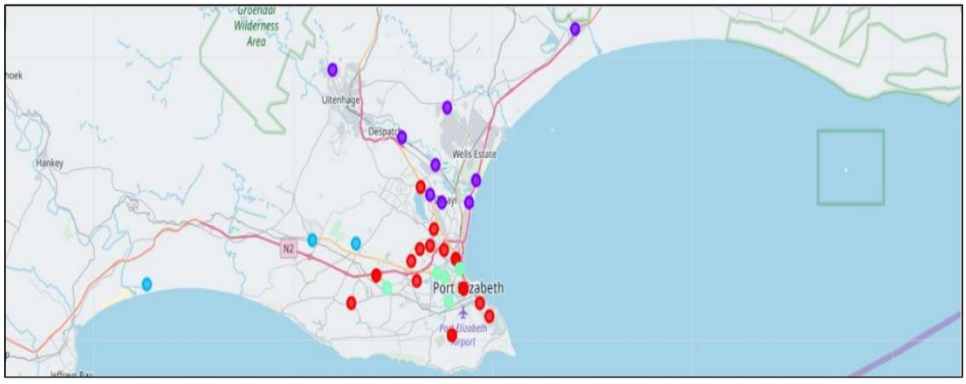


**Fig – Visualization - The Resulting Clusters Of Port Elizabeth City**

## 2. Outcomes – Pretoria, South Africa

The K-means method was used to cluster the suburbs of Pretoria into 5 clusters. The details of the clusters are as follows:

1. **Cluster 1** -23 Suburbs

Common Venues include Hotels, Lawyer, Water Park, Gift Shop, Electronics Store, Fast Food Restaurant, Flea Market, Furniture / Home Store, Gas Station, Gastropub, Golf Course.

2. **Cluster 2** -2 Suburbs

Common Venue is Lawyer, Water Park, Gift Shop, Electronics Store, Fast Food Restaurant, Flea Market, Furniture / Home Store, Gas Station, Gastropub, Golf Course.

3. **Cluster 3** - 3 Suburbs

Common Venues include Gastro pub, Department Store, Electronics Store, Fast Food Restaurant, Flea Market, Furniture / Home Store, Gas Station, Gift Shop.

4. **Cluster 4** - 25 Suburbs

Common Venues include Water Park, Gift Shop, Electronics Store, Fast Food Restaurant, Flea Market, Furniture / Home Store, Gas Station, Golf Course.

5. **Cluster 5** - 4 Suburbs

Common Venues include Water Park, Gift Shop, Electronics Store, Fast Food Restaurant, Flea Market , Furniture / Home Store, Gas Station, Golf Course

**Fig – Visualization - The Resulting Clusters Of Pretoria City** PAGE 10

# 5. Discussion section

**Port Elizabeth** has **47 suburbs** with **76 venues.** In addition, the geogr apical coordinate of Port Elizabeth, South Africa are **-33.9617051, 25.6 207519.**
. The best suburb to stay in is **ESCOMBE STREET** with the following ven ues:
**Suburb ESCOMBE STREET**

- 1st Most Common Venue Golf Course
- 2nd Most Common Venue Restaurant
- 3rd Most Common Venue Coffee Shop
- 4th Most Common Venue Mexican Restaurant
- 5th Most Common Venue Yoga Studio
- 6th Most Common Venue Deli / Bodegat
- 7th Most Common Venue gym
- 8th Most Common Venue Greek Restaurant
- 9th Most Common Venue Garden
- 10th Most Common Venue Frozen Yogurt Shop

**Pretoria** has **69 suburbs with 132 venues.** In addition, the geographic al coordinate of Pretoria, South Africa are -29.861825, 31.009909. T he best suburb to stay in is **PRETORIA** DOORNPOORT
with the following venues:

**Suburb PRETORIA DOORNPOORT**

- 1st Most Common Venue Golf Course
- 2nd Most Common Venue Water Park
- 3rd Most Common Venue Gift Shop
- 4th Most Common Venue Electronics Store
- 5th Most Common Venue Mexican Restaurant
- 6th Most Common Venue Flea Market
- 7th Most Common Venue Furniture / Home Store
- 8th Most Common Venue gym
- 9th Most Common Venue Gastropub
- 10th Most Common Venue Airport

Many of the neighborhoods are homogenous and are very similar to each other. Both **Port Elizabeth** and **Pretoria City** consist of suburb clusters that contain majority of the suburbs.

## 6. Conclusion section

**Pretoria** had a significant more number of suburbs and venues than **Port Elizabeth** th erefore it would be the better option to relocate to Pretoria, specifically **PRETORIA DOORNP OORT** as the most efficient choice. Pretoria offers a variety in choices for restaurants, gy ms, grocery stores, and Water Park, golf course extracurricular activities for individuals a nd families.