WEEK 1

This capstone project will be graded by your peers. This capstone project is worth **70%** of your total grade. The project will be completed over the course of **2 weeks**. Week 1 submissions will be worth **30%** whereas week 2 submissions will be worth **40% of your total grade**.

For this week, you will be required to submit the following:

1. A description of the problem and a discussion of the background. (**15 marks**)
2. A description of the data and how it will be used to solve the problem. (**15 marks)**

1. Background

South Africa is divided into 9 Provinces; these provinces each have one or more cities. Gauteng Province hosts the economic hub of Africa, by implication cities within this province have most of the economic activities than any other cities in the country. The province is also mainly urbanized than the other 8 provinces.

The South African citizen's perception is therefore that if you need to shop the best of the best you need to do it in one of the cities in Gauteng. Of all the cities in the Gauteng Province, Johannesburg is held to the highest regard than the others, but the rest of the cities in Gauteng also remain attractive to prospective business from international and locally.

2. Problem Statement

Construction Company XYZ Investments is interested in building a massive shopping mall that will be housing mostly international and a number of local businesses, the mall will also have mostly high end restaurants. The mall should be located in one of the Gauteng cities but shouldn't be too close to an area where other malls are popular as this will increase competition especially on prices.

As a Data Scientists, the team is required to identify a conducive city or cluster of neighborhoods where the mall can be built with a potential for good business.

3. Data Description

The data that will be used is geographical codes and coordination of South African Provinces, this data will be vital in identifying the cities/neighborhoods across the provinces.

https://dirkstrauss.com/south-african-cities/

The second data will be coming from the Foursquare site; this data will be utilized to identify business types across the cities/neighborhoods. The data will also assist in clustering cities/neighborhoods according to the types of businesses.

https://foursquare.com/

4. Data Utilization

The South African Cities codes and coordinates will be utilized to create the initial map of how the cities are clustered across each province. This data will then be sliced to reflect the targeted province (Gauteng) and the cluster of the cities/neighborhoods.

The Foursquare data will be utilized to identify business across each city in the targeted province, the business will be superimposed on the cities/neighborhoods. A clustering model will then be utilized to create clusters so as to assist in identifying a suitable city for building the mall.