# Capstone Project

IBM Applied Data Science Capstone Project

## Analysis of Gas Stations in Bay Area


By: Vivek Advani

Date: 04/10/202

# Introduction

Gas stations have gained popularity over the years, as the demand of crude oil tends to increase exponentially. Many economists still refer crude oil as "black gold" and it remains the single most important commodity in the world for primary source of energy production. With respect to changing patterns of consumer behaviour, the prices of gas have touched all-time high. Now to qualify as a good gas station it is required to provide certain retail service such as fast-moving consumer goods shopping and fresh corners. Apart from just making a gas station a better place do we need to answer following concerns:

Are there enough gas stations in our neighbourhood?
Does our favourite gas station require us to wait in long queue?
Should population density be correlated to the number of gas stations?

# Business Problem

The object of this project is to analyse which neighbourhood in Bay Area requires a gas station. With the help of Data Science Methodology and machine learning techniques, we can certainly build an analysis to provide solution for following questions.

# Methodology

Analytic Approach The analytic approach for this problem is to perform unsupervised learning technique such as K-means Clustering. This will help to identify various patterns based on neighbourhoods in Bay Area.

Data Requirements We would require data such as list of Boroughs and Neighbourhood of San Francisco Bay Area, also the census data for Bay Area)

Data Collection Once we have noted the data requirements, the next step is data collection. We need to scrape data from the online websites using libraries such as Beautiful Soup. Also, using Foursquare.

Data Understanding and Preparation The data understanding, and preparation part would be the most difficult as the collected data would not be clean. Goal here is to clean the data.

Modelling and Evaluation We would use K-Means algorithm to create K clusters. There is no such accuracy metric for Unsupervised Learning algorithm, we would use elbow graph to select the best K.

# Data Source

"https://en.wikipedia.org/wiki/List_of_cities_and_towns_in_the_San_Francisco_Bay_Area"

The San Francisco Bay Area, commonly known as the Bay Area, is a metropolitan region surrounding the San Francisco Bay estuaries in Northern California. According to the 2010 United States Census, the region has over 7.1 million inhabitants and approximately 6,900 square miles (18,000 km2) of land.The region is home to two major cities: San Francisco and Oakland and, the largest city, San Jose.

"https://foursquare.com/city-guide"

Using Foursquare to get 100 venues for each neighbourhood in Bay Area and then select only Gas Stations to build model.