**Data Science Final Project – San Francisco Data Analysis**

**Introduction**

**Background:**

The purpose of this project is to help the people wanting to immigrate to major cities of the world with research on good neighborhoods and ultimately make a smart decision.

The settlement decision will be made based on crimes count and access to the colleges.

Essentially this project will provide a detailed information of the neighborhoods in a city including the various catergorical venues including yoga centers, coffee shops, restaurants etc.

Assuming that an immigrant wants to setup a restaurant business, this project will be providing with a list of neighborhoods with least number of restaurants for someone to setup as a business.

In totality it will help people get an awareness on a new city. It will provide a comparative analysis of various neighbor hoods for settlement and opening up a particular business.

**Problem Statement:**

The purpose of this project is to provide information on:

* Good neighborhoods with accessible venues
* Good secondary school location and information
* Good neighborhood to setup a restaurant

**Location to be analyzed:**

San Francisco, CA is a very popular destination for immigration for staying and setting up business. Therefore, for this project the SF city, its neighborhoods, school districts and various categorical venues will be used.

**K-Means clustering:**

The similarities or dissimilarities between two neighborhoods in a city could be visualized by segmenting them into various clusters utilizing the k-means clustering machine learning algorithm. So, this project aims at clustering and making sense of data obtained from this clustering technique

**Location Data Analysis - Four Square API:**

Forsquare is one of the location data providers. A developer account has been created and credentials have been obtained. Due to a limited number of accesses for this developer account, there will be some restrictions on the radius and count of venues search.

**Data Analysis and Libraries:**

Along with the Foursquare API, Python and its associated libraries and packages will be used for data analysis and visualization.

Pandas: Dataframe creation and manipulation

Numpy: Mathematical Analysis

Matplotlib: Python plotting module

Folium: Interactive leaflet map creation

Scikit Learn: Implementing k-means clustering algorithm

JSON: Handling JSON files

Geocoder: Retrieving location data

**Data Section**

The SF crimes data is obtained form:

<https://data.sfgov.org/Public-Safety/Police-Department-Incident-Reports-2018-to-Present/wg3w-h783>

The SF colleges data:

<https://data.sfgov.org/Economy-and-Community/Colleges-in-San-Francisco-2011-/8r3f-pc6a>

Also, the data set with SF neighborhood spatial data is obtained from:

<https://data.sfgov.org/Geographic-Locations-and-Boundaries/Analysis-Neighborhoods/p5b7-5n3h>

## Data obtained from Four Square API:

## Four Square us a location data provider with information on venues nearby a neighborhood. It provides:

## Neighborhood

## Neighborhood Latitude

## Neighborhood Longitude

## Venue

## Name of the Venue

## Venue latitude

## Venue longitude

## Venue Category

## Determining the location to live

## The immigrants look for lesser crimes and easy access to good colleges in the neighborhood they want to settle.

## Based on crimes, the top 10 neighborhoods with least crimes is provided.

## 

## 

## Then the college data is obtained and map visualizations are provided.

## 

## Based on the college locations and lesser crimes, the best neighborhoods for settlement is provided.

## 

## Determining the best neighborhood to open restaurant business

## The decision is made on the basis that the restaurant is is not in top 3 venues for any particular neighborhood.

## The top 10 neighborhoods based on crimes are considered to display the venues.

## Foursquare API is used to get all the information of venues and locations around the neighborhoods.

## Venue category is an important column for us to determine the neighborhoods without coffee shop as their top venues.

## Then, a table is created with all the neighborhoods along with their top 5 venues.

## 

## 

## Determining the best neighborhood to open restaurant business and settle

## Based on the analysis on crimes, colleges and venues data, we can come to a conclusion that:

## The best neighborhoods to settle:

## 

## The best neighborhoods for restaurant business:

## 

## K-Means Clustering – Analyzing the Neighborhoods and Venues using Machine Learning

## First, the cluster labels are created and cluster labels are added to the table with neighborhood and venues information.

## Based on this cluster label we can map them to visualize the location distinction.

## Then we can analyze them label by label to understand how the algorithms has segregated the city data.

## 

## 

## Each color dot represents one cluster, for this example only 8 clusters are used.

## Analyzing each cluster separately.

## Cluster 0 represents the neighborhoods with restaurants as their most common venues. None of the neighborhoods to setup the business are in this cluster.

## 

## The top neighborhoods to open a restaurant business are in other clusters.

## 

## Conclusion

## The San Francisco neighborhood data, crimes data and college data is collected and analyzed to decide on the neighborhood to settle. The venues data obtained from foursquare API along with the settlement analysis has given the best location to open a restaurant business.

## The KMeans clustering is used to prove the hypothesis correct regarding the restaurant business.