Student Name: Monishankar Hazra

Applied Data Science Capstone

This document contains:

Project Name:

In Search of a Safe place to Stay with Family in Chicago, IL

(The Battle of Neighborhoods)

Week4 Part 1: Introduction/Business Problem

Chicago

Chicago, officially the City of Chicago, is the most populous city in the U.S. state of Illinois, and the third-most-populous city in the United States. With an estimated population of 2,693,976 in 2019, it is also the most populous city in the Midwestern United States. Located on the shores of freshwater Lake Michigan, Chicago was incorporated as a city in 1837 near a portage between the Great Lakes and the Mississippi River watershed. Chicago is an international hub for finance, culture, commerce, industry, education, technology, telecommunications, and transportation. Chicago is home to several Fortune 500 companies, including Allstate, Boeing, Caterpillar, Exelon, Kraft Heinz, McDonald's, Mondelez International, Sears, United Airlines Holdings, US Foods, and Walgreens.

The Problem

The city's overall crime rate, especially the violent crime rate, is higher than the US average. Chicago was responsible for nearly half of 2016's increase in homicides in the US, though the nation's crime rates remain near historic lows. The reasons for the higher numbers in Chicago remain unclear.

Crime in Chicago has been tracked by the Chicago Police Department's Bureau of Records since the beginning of the 20th century.

In this situation, a family with a school-going kid, needs to move to Chicago, and **find out few safe Community Areas to live with good amenities in the Neighborhoods**. The project aims to analyze the Community Areas of Chicago using Data Science, using K-Means Clustering techniques, and finally suggest an appropriate Community Area of Safe living.

The project aims to do :

a. Importing Datasets
b. Cleaning the Data
c. Data frame manipulation
d. Web-scrapping
e. Gathering GeoLocation using geopy
f. Remote call to FourSquare to get neighbour-list
g. Summarizing the Data
h. Visualization
i. Clustering

Week4 Part 2: Data section

2. **Data Acquisition and Cleaning**

2.1 Data Sources:

a. Chicago_Crimes_2012_to_2017
   Data Source: Kaggle

b. Chicago_Public_Schools_-_Progress_Report_Cards__2011-2012
   Data Source: IBM

c. Census_Data_Selected_socioeconomic_indicators_Chicago_008_2012
   Data Source: IBM

d. List of Community Areas in Chicago
   https://en.wikipedia.org/wiki/Community_areas_in_Chicago

e. FourSquare.com
   https://api.foursquare.com

Details about the Datasets:

Crime Data
ID
Primary Type
Arrest
Community Area
Year

School Data:
School ID
COMMUNITY_AREA_NAME
SAFETY SCORE

Census Data:
COMMUNITY_AREA_NUMBER
PERCENT HOUSEHOLDS BELOW POVERTY
PER_CAPITA_INCOME

A. **Data Cleaning**:

1. Inspect the Crime data
    a. Unnecessary cols to be dropped
    b. Zero and Null Values to be managed

2. Corroborate the Community Number and Name with Wiki-Community Areas by WebScraping

B. **Feature Selection**

a. Consider Crime Data with Arrest = YES for Year = 2012
b. Consider Average School Safety Score for each Community
c. Consider Percent Households Below Poverty Level for for each Community
d. Consider Per Capita Income for each Community

C. **Run Correlation**, for each Community Area, among:

- Total Number of Crimes
- Average School Safety Score
- Percent Households Below Poverty Level
- Per Capita Income for each Community

D. **Identify the Safest Communities with above features**

E. Send calls to Foursquare.com and find out 100 neighborhoods of the identified Community

F. **Classification of Five Clusters**, which gives brief understanding of the characteristics of the safest Communities in Chicago