Capstone Project - Los Angeles Motor Traffic Collisions during Year 2018 (Week 1)

IBM professional Data Science Capstone by IBM/Coursera

Introduction: Problem

In this use case, lets analyze Los Angeles traffic collision data and identify & cluster similar venue categories near to the collision locations. So that Los Angeles Dept of Transportation authorities can better understand the public needs and plan accordingly to ease up traffic congestions in each locality around the popular venues, which will help in minimizing the traffic collision rates.

Background

Los Angeles has one of the highest motor collision rate among United States metropolitan cities.

Los Angeles traffic congestion is also rated as one of the highest in United States and probably in the world. Even though, LA has mass public transport system including public buses, light Rail and subway, ridership rates are decreasing annually and residents are taking on to roads to move around the city, which is resulting in high fatality collision rates.

There are several contributing factors for high collision rates like increasing population rates, aging public transit infrastructure, increasing economy, popular venues or events etc.

Popular venues may inadverently contributes to traffic congestions and collisions. Popular venues includes music arenas, parks, cafes, yoga studios etc. Analyzing venues near to collision locations helps in understanding traffic footnote and public needs. Thus helping LA Dept of Transportation to plan better in order to ease up congestions and minimize collisions.

Data

Following are the data sources:

- Under open data policy, **LAPD** (Los Angeles Police Department) collects and regularly updates Los Angeles traffic collision data every week.
- Los Angeles traffic collision data is provided in LA city open data website. https://data.lacity.org/A-Safe-City/Traffic-Collision-Data-from-2010-to-Present/d5tf-ez2w
- Los Angeles Police Department reporting divisions geo location data is provided in LA city open data website. https://geohub.lacity.org/datasets/031d488e158144d0b3aecaa9c888b7b3_0/data
- Venues and venue categories data wil be obained from **Foursquare API** using their APIs.

Following are the observations made about the collision data provided to the public:

- Los Angeles traffic collision data is provided from year 2010 to till current date.
- Data is provided in various formats like csv, json (via api) etc.
- Provided data is sanitized to scrub all PII (personally identifiable information).
- Geo location coordinates are provided for all collisions.

Data Consumption is done the following way:

- Los Angeles traffic collision data will be read in csv format.
- Los Angeles police reporting divisions data will be read in geojson format.
- Foursquare Venue exploration data will be read in json format.

Data Usage is done the following way:

- Los Angeles traffic collision data for year 2018 will be used to analyze the collisions.
- Collision geo location coordinates will be used as input to obtain nearby venue data from Foursquare.
- Obtained Foursquare venue data will be used to cluster collision locations
- LAPD reporting divisions geo json data will be used for choropleth maps.

Data features:

- LA Traffic collision data has 24 features and 466242 rows till date.
- Following are the features of the dataset.
    - DR_Number

    - DR_Number
    - Date_Reported
    - Date_Occurred
    - Time_Occurred
    - Area_ID
    - Area_Name
    - Reporting_District
    - Crime_Code
    - Crime_Code_Description
    - MO_Codes
    - Victim_Age
    - Victim_Sex
    - Victim_Descent
    - Premise_Code
    - Premise_Description
    - Address
    - Cross_Street
    - Location
    - Zip_Codes
    - Census_Tracts
    - Precinct_Boundaries
    - LA_Specific_Plans
    - Council_Districts
    - Neighborhood*Councils*(Certified)
- Feature 'Date_Occurred' will be used to select collisions happened in year 2018.
- Feature 'Location' contains geo coordinates for collisions.
- Feature 'Area_Name' is LAPD Division.