Syracuse University

Memo

|  |  |
| --- | --- |
| To: | Dr. Landowski |
| From: | Sathish Kumar Rajendiran (Team Members: Prasad Kulkarni / Sathish Kumar Rajendiran) |
| Date: | 08/01/2020 |
| Re: | Project Proposal |

**Topic: LOS ANGELES TRAFFIC COLLISION ANALYSIS**

As the second largest in the United States, Los Angeles has traffic challenges due to a large and growing population and an increase in the number of cars. A better understanding of the factors that contribute to accidents can help government officials, companies, citizens and other interested parties to understand how to make the city safer and more drivable.

**Data Sources:**

1. **Structured data** 
   1. **Kaggle:-**  <https://www.kaggle.com/cityofLA/los-angeles-traffic-collision-data>
2. **Semi Structured data**
   1. **Twitter data**

**Data Description:**

**LA Traffic data:**

The Los Angeles Traffic Collision Data is publicly available from Kaggle.com is owned by the City of Los Angeles. The contains 481,568 incidents from 2010 to 2019.

|  |  |  |  |
| --- | --- | --- | --- |
| **Column** | **Description** | **Data Type** | **Range of Values** |
| **Area Name** | The 21 geographic areas or Patrol Divisions given a name based on landmark or surrounding community it is responsible for | Object | ‘Devonshire’, ‘West Valley’, ‘Topanga’, ‘Mission’, |
| **Time Occurred** | Time of collision | Integer | Time values |
| **Victim Age** | Age of victim of car collision | Integer | Age values from 0-99 |
| **Victim Sex** | Sex of the victims | Object | F - female M - male |
|  | Genders called “H” and “N” were ignored in analysis since no indication what they represented from Kaggle website and also represented a very small amount |  | X - unknown |
| **Victim Descent** | Ethnicity of victim of collision | Object | A - Other Asian B - Black C - Chinese D - Cambodian F - Filipino G - Guamanian H - Hispanic/Latin/Mexican I - American Indian/Alaskan Native J - Japanese K - Korean L - Laotian O - Other P - Pacific Islander S - Samoan U - Hawaiian V - Vietnamese W - White X - Unknown Z - Asian Indian |
| **Premise Description** | Indicates type of location where collision occurred | Object | 42 unique values such as ‘STREET’, ‘PARKING LOT’, ‘FREEWAY’. |
| **Address** | Street address of collision | Object | Streets |
| **Cross Street** | Nearest intersection street to Address | Object | Cross street |
| **Location** | GPS coordinates of collision with longitude and latitude | Object | Latitude and Longitude coordinates |
| **Zip Codes** | Zip code of collision | Object | 5-digit number |
| **Council Districts** | Council District number of collisions | Integer | Values from 1-15 |

**Twitter Data:**

Using Tweepy (python library) and developer API account from Twitter - We are also trying to get supplemental tweets data related to climate change discussions, overall sentiment around accidents and how these factors play a role into the incidents during 2010 to 2019.

|  |  |  |  |
| --- | --- | --- | --- |
| **Column** | **Description** | **Data Type** | **Range of Values** |
| **Created\_at** | UTC time when this tweet was created | String | Time values |
| **Id** | Unique Identifier for this tweet | Integer | Ex. “id:105011842332” |
| **Text** | Actual UTF-8 Twitter text | String | Ex. “David Reese:#WaterWorld #Federalism SoCalGas sues California over climate change policy - Los Angeles Times” |
| **Coordinates** | Geographic location of this tweet as reported | Coordinates | Ex.  "coordinates":  {  "coordinates":  [  -75.14310264,  40.05701649  ],  "type":"Point"  } |
| **Retweet\_count** | Number of times this tweet has been retweeted | Integer | Ex."retweet\_count":160 |
| **Retweeted** | Whether this tweet has been retweeted | Boolean | Ex. "retweeted":false |
| **Lang** | Language Identifier | String | Ex. "lang": "en" |
| **Favorite\_count** | Approximately how many times this tweet has been liked | Integer | Ex. "favorite\_count":295 |

We are also trying to get supplemental data related to weather conditions and avg income do some analysis of how people involved in collision are affected and how factors like weather conditions play a role into the incidents.

**Research Questions:**

The goal is to explore the trends and correlations between the data to provide useful information that can help answer our proposed analysis questions:

* What are the most dangerous intersections?
* What are the most common collision areas in Los Angeles?
* What is the best/worst times of the day for accidents? Best/worst month?
* What is the demographic makeup of victims in collisions?
* What is the relationship between income and collision victims?
* Does climate change play a factor? (ex. rainy day with more accidents)
* What is the general sentiment?

**Data Preparation Plan**

1. Load the data into Python.
2. Merge tweets related to climate change and accidents in LA with the Collision data.
3. Convert data using Date Time and broken up into months, weekdays, and year columns.
4. Explore the data for anomalies, missing data, adjusted duration, etc.
5. Depending on the amount of data missing, decide whether to replace data with average or median values.
6. Break location details using longitude and latitude columns to check feasibility of doing analysis with map visualization.
7. Include plots for visualization.