**Data Analysis – Traffic Accidents**

**INTRODUCTION**

As a part of the course work for Data Analysis, I have selected a dataset on US Accident, obtained from the website <https://smoosavi.org/datasets/us_accidents>.

This data set on US Accidents can be used for numerous applications such as real-time accident prediction, studying accident hotspot locations, casualty analysis and extracting cause and effect rules to predict accidents, or studying the impact of precipitation or other environmental stimuli on accident occurrence.

The purpose of this analysis is to analyze the data and find out what are the key variables that impact the severity of the traffic accidents that happened in US and ultimately predict the severity of the accidents based on given variables through data visualization and regression analysis using Python.

**DATASET**

 In this analysis I used data that is collected from February 2016 to June 2020 for the [Contiguous United States](https://en.wikipedia.org/wiki/Contiguous_United_States). In this dataset, we have about **3.5 million** traffic accidents. The dataset covers 49 states of the US. This dataset consists of 49 attributes. As part of the analysis some attributes have been excluded in data cleansing process.

**METHODOLOGY USED FOR ANALYSIS:**

## Step 1: Data cleaning:

Created a subset with variables using for this project. Creating factors of variable and removing the unknown and null values.

In this step we will check the dataset: preview data and some summary statistics; check columns for NaN (data missing); remove unnecessary columns that will not be required for analysis.

Necessary manipulations of dataset will be performed to lay the path for Data Analysis.

## Step 2: Data Visualization:

In this step the cleaned Dataset will used for various Data Visualizations which will guide our analysis to conclusions.

Data will be grouped by Severity. The codes for Severity in the dataset are as follows:

|  |  |
| --- | --- |
| Severity | Description |
| 1 | indicates the least impact on traffic  (i.e., short delay as a result of the accident) |
| 2 | indicates the moderate impact on traffic  (i.e., short delay as a result of the accident) |
| 3 | Indicates a moderate impact on traffic (i.e., moderate delay as a result of the accident). |
| 4 | Indicates a significant impact on traffic (i.e., long delay as a result of the accident). |

A number of Data Visualizations are performed:

* Plot data to visualize year over year traffic accident trend
* Plot the data by **Severity**and by**Year** to show the trend of accidents by year and by accident severity
* Plot the data by weather related variables have impact on the severity of the accidents
  + Temperature Vs Severity
  + **Humidity** Vs **Severity**
  + Wind Chill Vs Severity
  + By time of Sunrise & Sunset
  + By Months of Year
  + By the length of the road extent affected by the accident.
  + Using the heat map to know for each month, which day of the month are more likely to have more accidents.

## Step 3: Regression analysis:

Selecting best subset of variables to perform regression analysis by using regularized Logistic regression method since the dependent variable is Severity, which is a categorical variable.

The independent variables are selected based on the previous analysis. Split data into two parts train and test. Finally fit the Logistic regression model using train data set and test this model with test dataset and finally conclusion. Finally use Confusion Matrix to see how many of the prediction are correct and how many are incorrect.