**Data Science Final Project : Maryland Traffic Violation Data**

Problem Statement

A high proportion of traffic violations lead to traffic accidents.

We need to identify the driver behavior and profile which lead to accidents and develop mitigation measures to reduce them. Mitigation measures could include higher insurance premiums, increased enforcement etc.

Hypothesis/Assumptions

Traffic violations are a small subset of inappropriate driver behavior; they must be caught by the police before entering the database!

There is a strong correlation between traffic violations and traffic accidents, and human error is the sole cause of these accidents (i.e. NOT machine failure or act of God).

There is a certain type of driver profile and driver behavior that are likely to lead to traffic violations and subsequently traffic accidents.

Null Hypothesis : There is a correlation between a certain driver profile/behavior and traffic violations.

Alternative Hypothesis : There is no correlation between a certain driver profile/behavior and traffic violations.

Goal/Success Metrics

The features within the dataset needs to be subdivided into different subsets to provide a clear distinction between different sorts of driver behavior and profile.

Two separate classification models will be created for driver behavior and risk profile; to see whether they correlate to traffic accidents.

Risk/Limitations

Although driver profile and behavior are independent of each other, including them within the same model would increase the complexity unnecessarily and subsequently increased variance.

The assumption that all traffic accidents are caused by traffic violations may not hold true.

There is already a sampling bias introduced when data is analysed from a local county police database.

Datasets Required

Kaggle’s Maryland Traffic Violation Data set. We will use this dataset to generalize traffic violations worldwide.

Features available to create the model:

1. **Driver Behavior**

‘Description’ Column – Filter out the top 3 at-risk behavior and create dummy variables.

1. **Driver Profile**

‘Gender’ Column

‘Vehicle Make’ Column

‘Vehicle Colour’ Column