**Introduction/Business Problem**

The science and technology has changed the world transportation sector. Now people love to travel in cars intercity or long distances with greater comfort and speed. However, the problem of car accidents are still suffer us a lot. The USA Highway and Traffic Administration department said there is a loss of about 900 billion dollars in a year due to car accidents besides loss precious human lives. In 2017, Washington traffic fatalities increased by 5.4 percent (from 536 to 565). Nationally, traffic deaths remained largely unchanged, decreasing less than half a percent and leveling off the steepest two-year increase in 50 years. With the rebounding economy and increases in vehicle miles traveled, Washington must require to implement new innovations and modern data analyzation to realize its vision of zero traffic fatalities and serious injuries. This Data Science Project aims to analyze the traffic accidents data and find the factors that caused these accidents. The main stake holders involved in this are 1- Public Development Authority Seattle 2- Drivers

**Data:**

The latest accident dataset of accidents has been taken from the city of Seattle, Washington from 2004 to 2020. The data contains accidents details such as time, condition of weather, location, severity, junction type, road conditions, speeding etc. There are more than 30 features in the dataset. The data set used can be downloaded from [there.](https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv) The propose model will try to predict an accident severity, the feature accident severity is in the form of 1( Car or property damage) and 2 (physical injury) that are encoded into 0 (property damage only) and 1( Physical Injury) using One hot Encoder. The 1 denote higher severity as due to Physical damage than 0 which indicates property damage only. However, there would be many cases in which the condition will be very serious such as loss of lives which are unique values for every variable which were either Other or Unknown. To solve this, the arrays were made for each column which are encoded equal proportions of elements as original columns. These arrays were used on the original columns of which had Others or Unknown in them to fill the missing data. This entire process of cleaning and filling of data led to a loss of almost 5000 rows which had redundant data, whereas other rows with unknown values were filled earlier.

