**Introduction/Business Problem:**

Road traffic crashes are one of the world's largest public health and injury prevention problems. The problem is all the more acute because the victims are overwhelmingly healthy before their crashes. According to the World Health Organization (WHO), more than 1 million people are killed on the world's roads each year. A report published by the WHO in 2004 estimated that some 1.2 million people were killed and 50 million injured in traffic collisions on the roads around the world each year and was the leading cause of death among children 10–19 years of age.

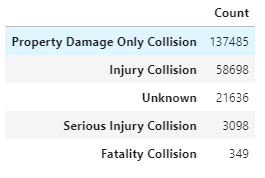
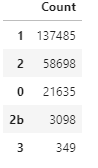
The project's objective is to predict the probability and severity of a car accident before it happens in order to intimate the driver to avoid such incidents through the use of historical car accidents data and how aspects like weather, road conditions, traffic and many other factors affect the probability of an accident and how severe it can be.

The insights of this project can be helpful to the following parties:

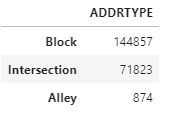
* **Drivers**: They can get to know the probability and severity of the accident that can happen, they might consider driving more carefully or even choose a safer route.
* **Emergency Services/Road Safety/Traffic Department:** They can get to know the areas where the probability of an accident to happen is more. So, they can have more teams near those areas.

**Data understanding:**

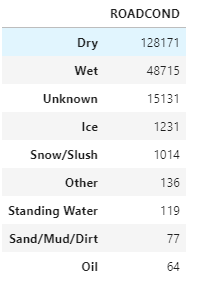
1. **Data Description:**The data is taken from the Seattle Department of Transportation and recorded by Traffic Records Group. It covers the annual collisions data from 2004 to the present. The dataset can be downloaded from [here](https://www.kaggle.com/jonleon/seattle-sdot-collisions-data/download) and the metadata from [here](https://www.seattle.gov/Documents/Departments/SDOT/GIS/Collisions_OD.pdf).
2. **Data Exploration:**   
   The dataset consists of **221266 rows and 40 columns**. The target column is labeled **SEVERITYCODE**. We can potentially use the other 39 columns to train the algorithm such as Road Condition, Weather Condition or Light Conditions. There are many columns that are not relevant for this project or they contain non-standardized data and missing values hence the data needs to be cleaned up.  
     
   For this project I’m using the dataset from kaggle instead of the one provided by Coursera because:   
    1. It contains collisions of all severity types and not just 1 and 2.  
    2. It has more features and records.  
   Following are some features which can be useful for predicting the severity of a collision:   
     
   **SEVERITYCODE**: This is the target column and it corresponds to the severity of the collision.  
   **SEVERITYDESC**: A detailed description of the severity of the collision.



* Around 62% of the collisions resulted in Property Damage while 0.15% resulted in Fatality Collsion.

**ADDRTYPE**: Collision address type.  


* Around 66% of the collisions occurred in Block while 0.4% ocured in Alley.

**ROADCOND:** The condition of the road during the collision.  


* Around 65% of the collisions occurred in Dry while 0.03% ocured in Oil.