**Introduction**/**Business Problem**

The purpose of this Capstone Project is to help people to get awareness of accidents occurrences and causes of it every year in a city. This project will help to predict the severity of accidents due to changing environmental conditions and vehicle conditions.

According to preliminary estimates from National Highway Traffic Safety Administration (NHTSA), 36,120 people died in motor vehicle crashes in [2019](https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812946), down 1.2 percent from 36,560 in 2018. Reducing traffic accidents is an important public safety challenge around the world. Accident prediction is important for optimizing public transportation, enabling safer routes, and cost-effectively improving the transportation infrastructure, all in order to make the roads safer. It will help people making smart and efficient decision on selecting safe road routes to avoid accidents and be cautious.

The goal of ‘how to deal with the accidents data’- accident prediction is usually to provide a measure of the risk of accidents at different points in time and space. The severity of an accident is the label used to train the model which describes the fatality of an accident, and the proposed model can be used to identify where and when the risk of accident is significantly higher than average in order to take actions to reduce that risk.

This Capstone Project aim to analyse accident forecast basing on fatality of an accident. The severity of an accident is the label used to train the model which describes the fatality of an accident, and the proposed model can be used to identify where and when the risk of accident is significantly higher than average in order to take actions to reduce that risk.

**Data Description**

Data : Data –Collisions from Applied Data Science course

Data link: [https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv](%20%20https:/s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv)

Metadata Link: <https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Metadata.pdf>

An effort to reduce the frequency of car collisions in a community, an algorithm must be developed to predict the severity of an accident given the current weather, road and visibility conditions. When conditions are bad, this model will alert drivers to remind them to be more careful.

For that a proactive approach is required, that includes a collision prevention, meaning, preventing a potential unsafe road conditions from occurring in the first place. By recognizing the key factors that influence accident severity, the solution may be of great utility to various Government Departments/Authorities. The results of analysis and modeling can be used by these Departments to take appropriate measures to reduce accident impact and thereby improve traffic safety.

In order to mitigate the impact of data size on analysis and prediction, we present a new dataset, by using labels encoding to covert the features to our desired data type. Timeframe: 2004 to Present. The data has 194674 rows and 37 columns with wide range of attributes including Location, Severity Code, Vehicle Count, Injuries, Fatalities, Junction Type, Person Count, Weather, Road Condition, collision type, address type, speeding, Collisions type etc.

Our predictor or target variable will be 'SEVERITYCODE' because it is used measure the severity of an accident within the dataset. Attributes used to weigh the severity of an accident are 'WEATHER', 'ROADCOND' and 'LIGHTCOND' and few other attributes which do not correlate and have impact on target in an regression model.

In its original form, this data is not fit for analysis. Firstly, there are many columns that we will not use for this model. secondly, most of the features are of type object, when they should be numerical type. We must use label encoding to covert the features to our desired data type, also might need to do some feature engineering to improve the predictability of your model.

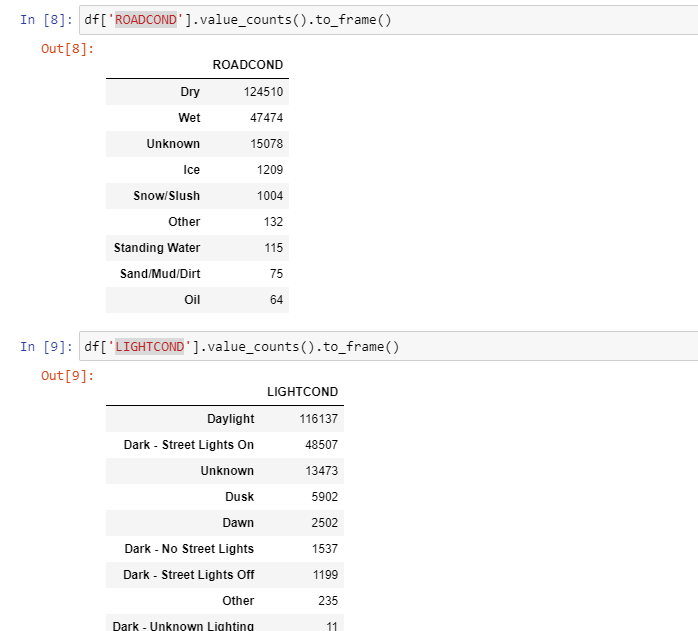
**Data Preparation:**

The data preparation includes all the required activities to construct the final dataset which will be fed into the modeling tools. Data preparation can be performed multiple times and it includes balancing the labeled data, transformation, filling missing data, and cleaning the dataset.

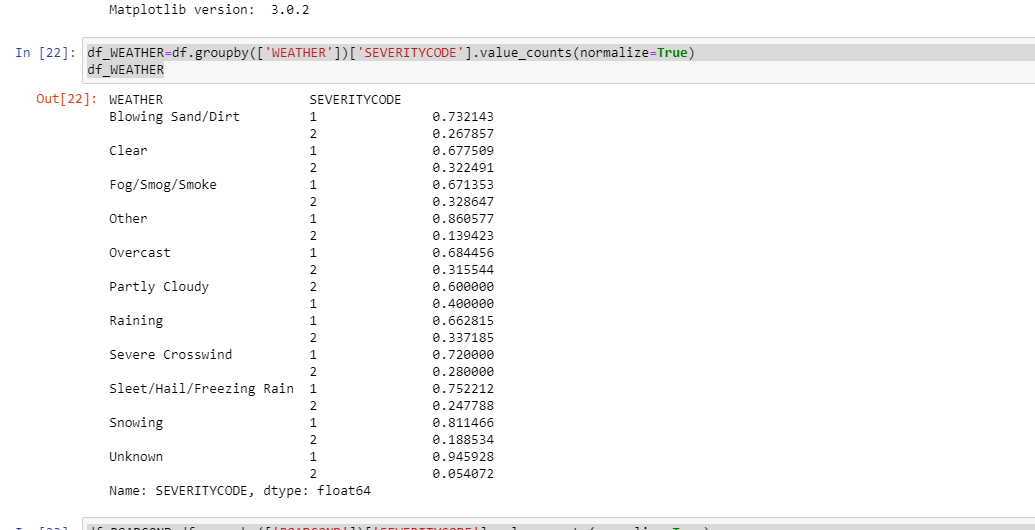
Features such as SPEEDING, "ADDRTYPE","WEATHER" ,"LIGHTCOND" ,"ROADCOND" ,"UNDERINFL" ,"STATUS" are chosen as independent variables which may imact the severity of accident.

Severitycode is the target variable. Predicting the target variable with the help of independent variables will provide a efficient and effective model car accident model.

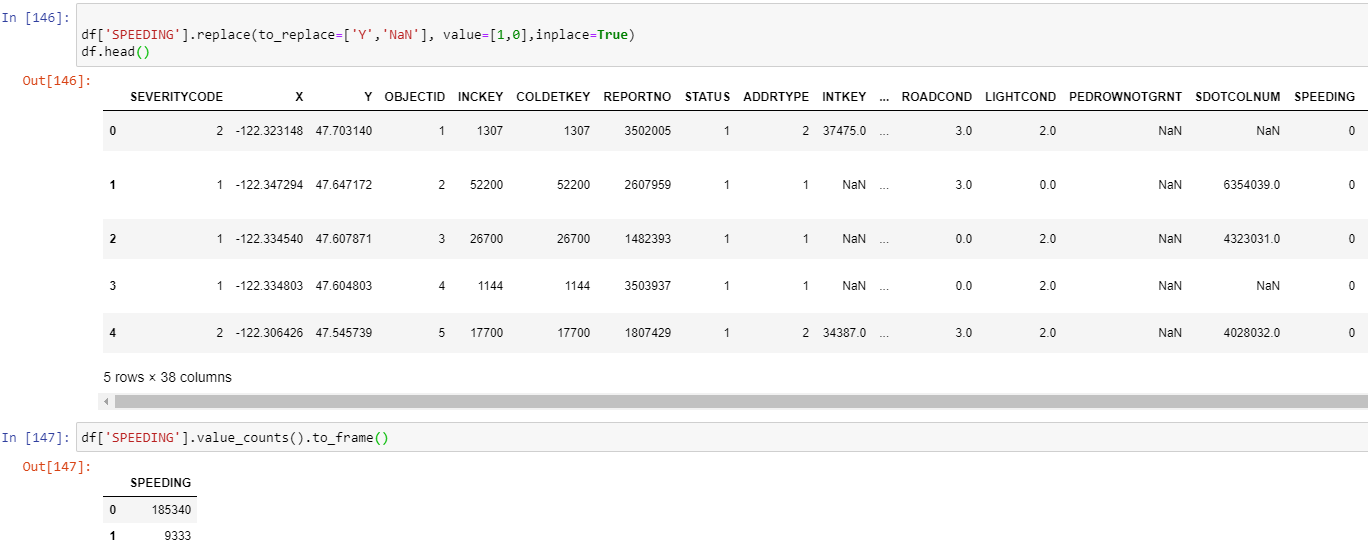
While preparing the data, each attribute is checked to get the brief idea of varients.



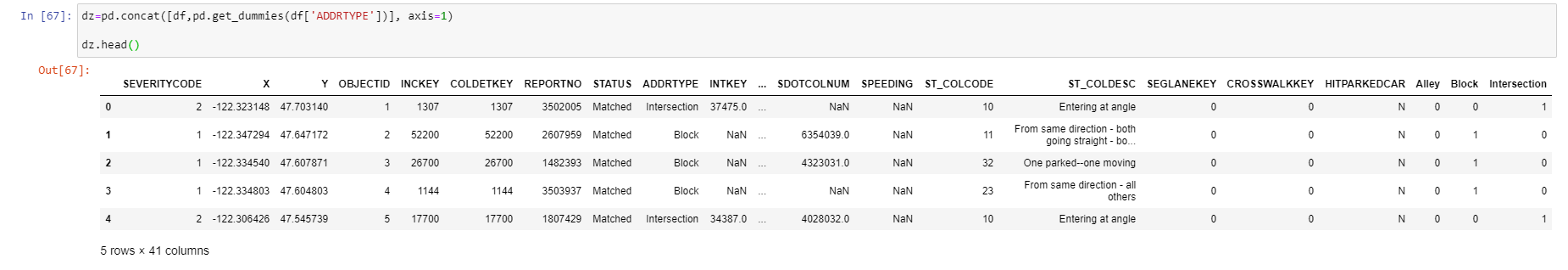
As the categorical variables need to be transformed into numeric data for modelling purpose, initially these attributes are grouped by severitycode and normalized to check the contribution of each subvarient of the attributes.



The transformation of categorical variables to numeric variables is done as follows

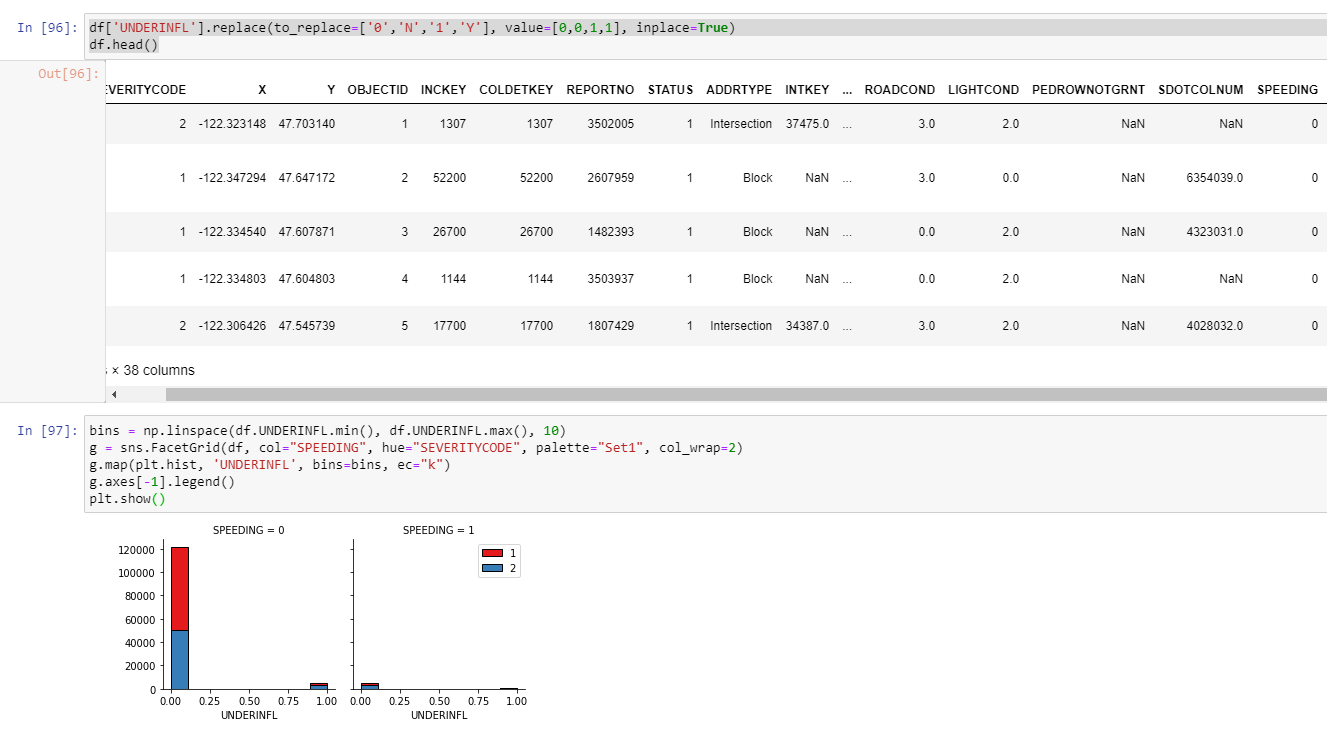


Creation of dummy variables



A quick visualization of transformed variables, showing the impact on Severatity code

A



Adjusting the null values in the dataset to frequency adjustment and following the classification model .

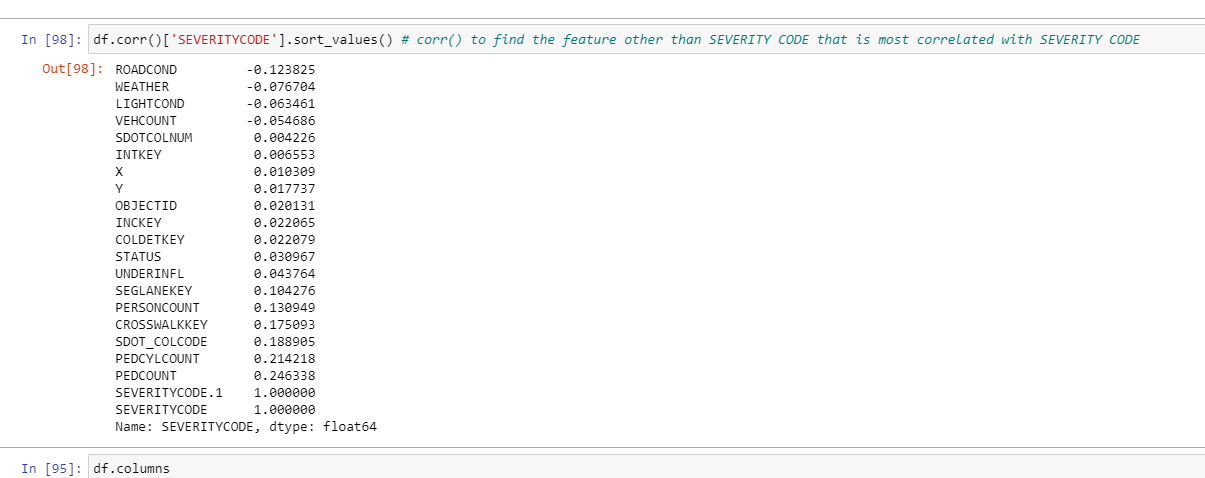
The data is prepared and is set for modelling .

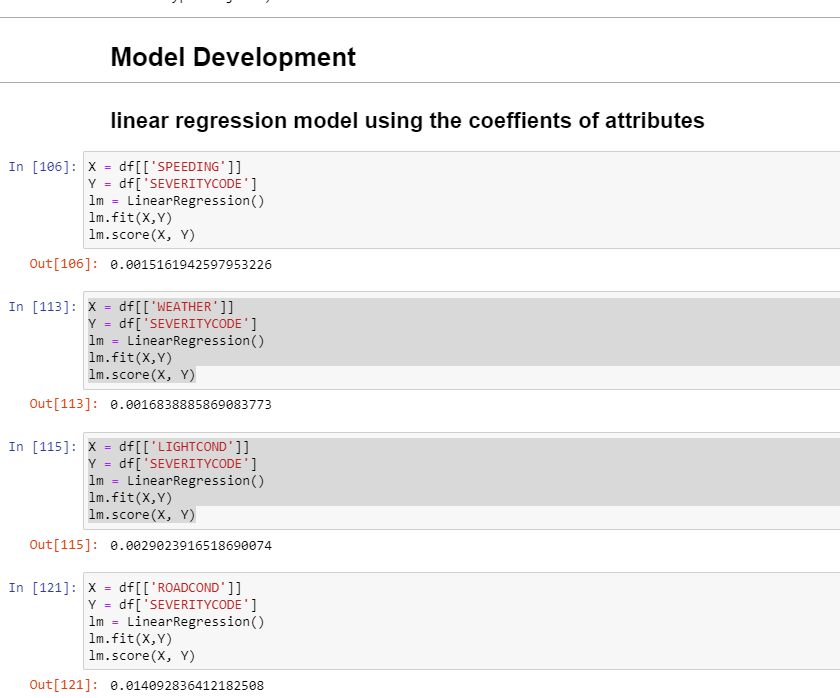
### ****Modeling:****

In this phase, various algorithms and methods can be selected and applied to build the model including supervised machine learning techniques.

Descriptive analysis is conducted for the prediction of this model

Correction is calculated to see the impact of individual attribute

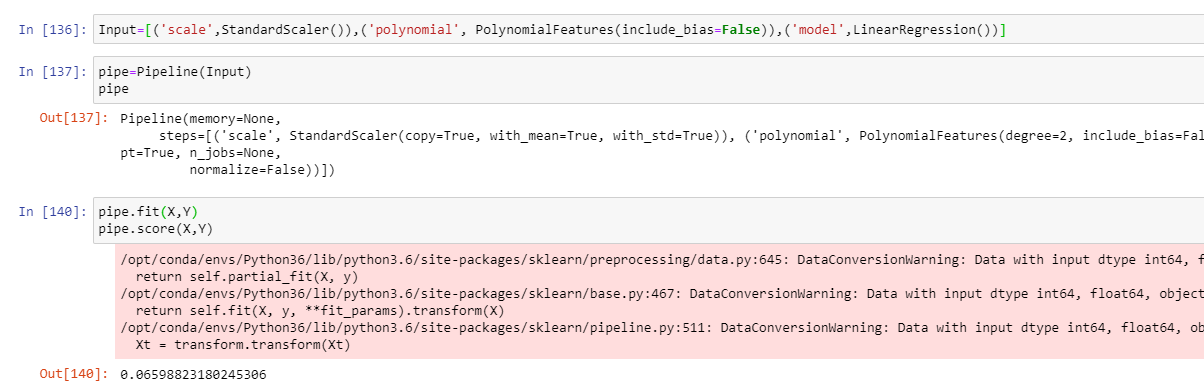




The out121 specifies the contribution of variable road Conditions to severitycode which is the highest of all attributes.



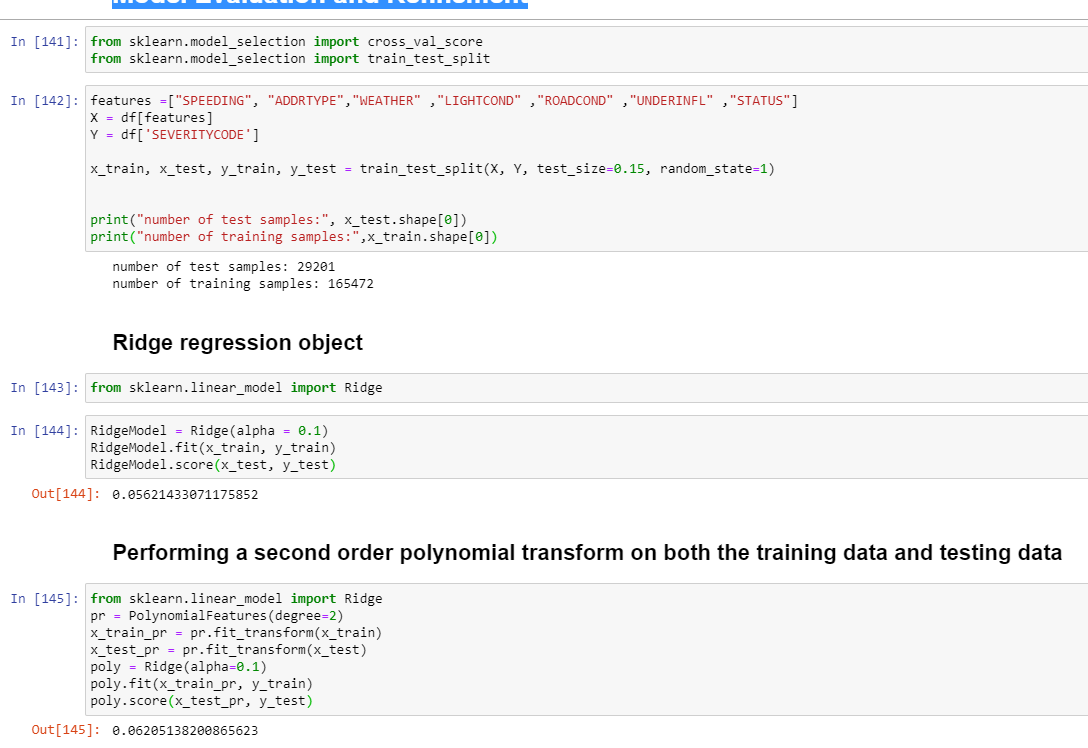
R^2 specifies how good is the model. Higher the R^2 value higer is the fitness of the model.



# Model Evaluation and Refinement

### ****Evaluation****:

Before proceeding to the deployment stage, the model needs to be evaluated thoroughly to ensure that the business or the applications' objectives are achieved. Certain metrics can be used for the model evaluation such as accuracy, recall, F1-score, precision, and others.



Trinin and Test sets have similar R^2 values. So the predictability of model is efficient.

### ****Deployment:****

The deployment phase requirements vary from project to project. It can be as simple as creating a report, developing interactive visualization, or making the machine learning model available in the production environment. In this environment, the customers or end-users can utilize the model in different ways.