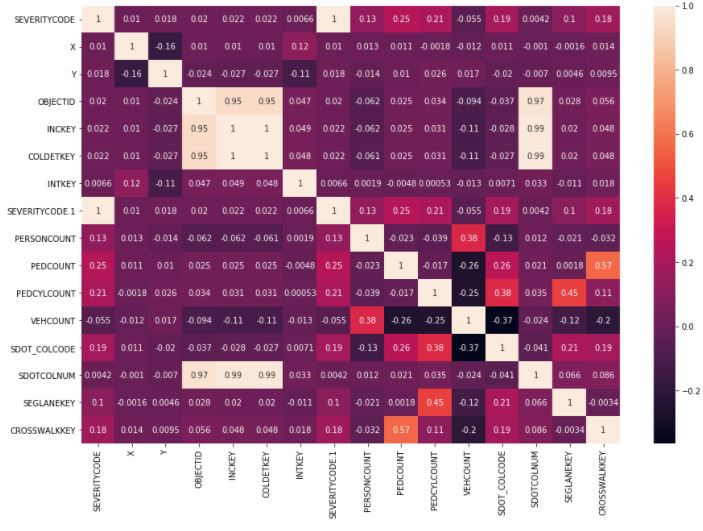
I am using the example data set which has been provided in the course itself. Below are some of the observations that I made:

1. The dataset has 194673 rows and 38 columns. Out of these 38 our target variable is ‘SEVERITYCODE’, all the other variables are independent ones which will help in predicting the value of the target variable.
2. The columns have different data types including int, float, and object.
3. 19 columns have null values (NaN) with the percentage of null values varying greatly among these columns. The columns having 70% or more (such as ‘SDOTCOLNUM’, ‘SPEEDING’ and more) null values will be dropped while creating a model.
4. A number of columns (such as ‘OBJECTID’) have continuous values. These values will serve as an obstacle while using classification models. Keeping this in mind, the columns having continuous values have been dropped.
5. The correlation matrix is shown below showing that the correlation between most of the columns wand our target variable is not high.



1. To test out different classification models, I will be dropping columns based on the criterion described in the above points, will be using dummy variables for the selected columns (feature variables), and then splitting the dataset into train and test sets.
2. Finally, after cleaning the data, our final data frame has 182895 rows and 15 columns (the number of columns will increase after creating dummy variables)