Data Processing and Understanding

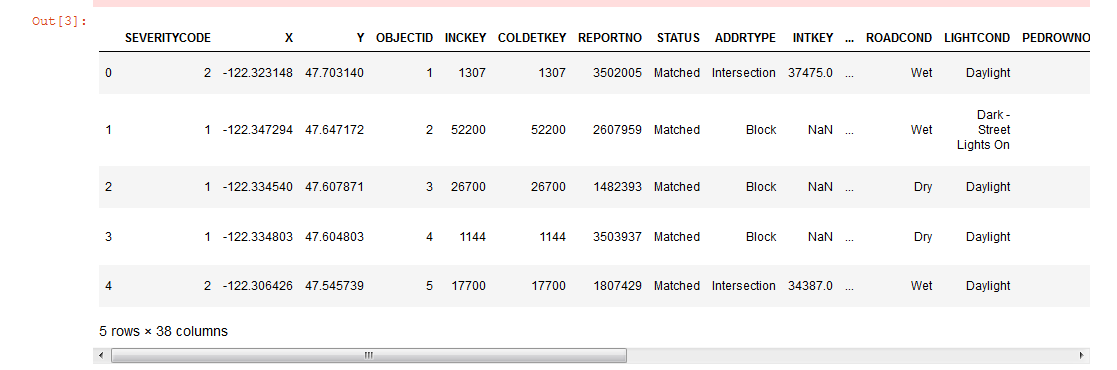
# ****Data****

This is an extensive data set from the Seattle Police Department with almost 35 parameters like collision type, Person Count, Vehicle count, weather , Speed etc. To accurately build a model to prevent future accidents and/or reduce their severity, I have decided to use few parameters and drop the remaining parameters depend on the data

# ****Methodology****

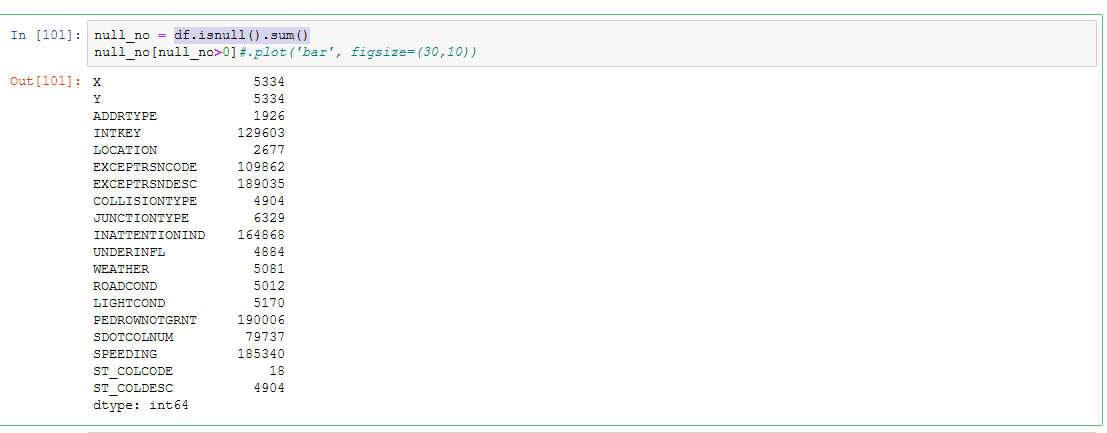
I used Jupyter Notebook to conduct that analysis and imported all the necessary Python libraries like Pandas, Numpy, Matplotlib and sklearn.

I started by downloading the csv file as given in week 1 details and saved as accident\_data.csv.



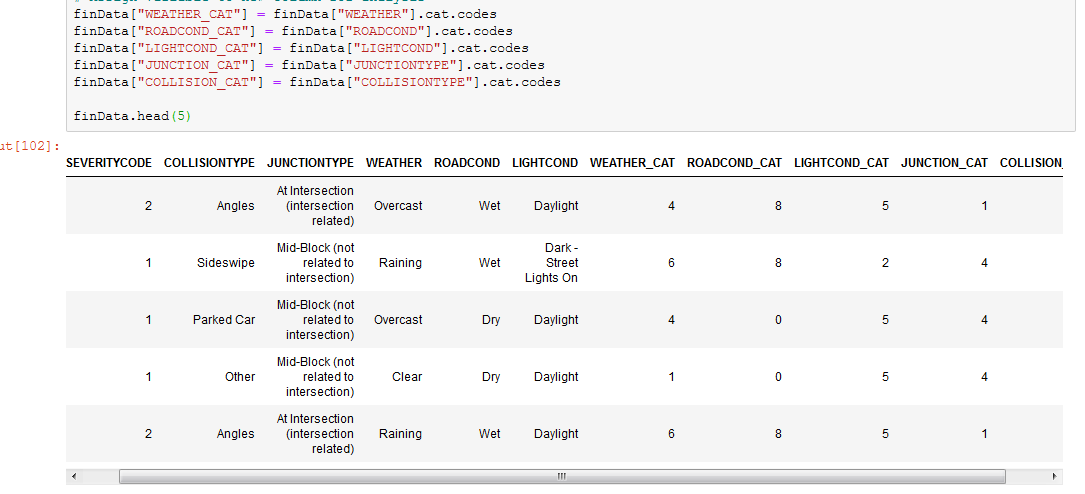
The database considered around 35 parameters and the output label was severity code based on these parameters

To prepare the data, so that I can drop columns we do not need from the dataset, i.e., columns that do not have values or where the values are unknown I used the commands df.info( ) , df.describe( )and df.isnull( ).sum( ) .



By looking at the table, I decided to incorporate 5 parameters weather, road condition, light condition , Junction type and collision category. Even though speeding is an important parameter, we have to drop speeding entirely because it is missing over 180,000 values and this can hamper the results.

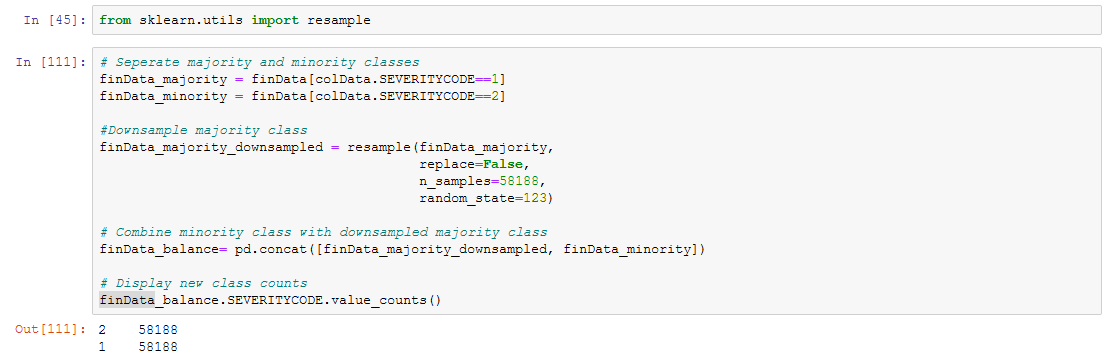
From the main dataframe df, I dropped the unwanted parameters and kept only the necessary parameters and changed their data type to int64



When we analysed the severity code we found out that data is unbalanced as Severity code 1 was approximately three times larger than Severity code 2.



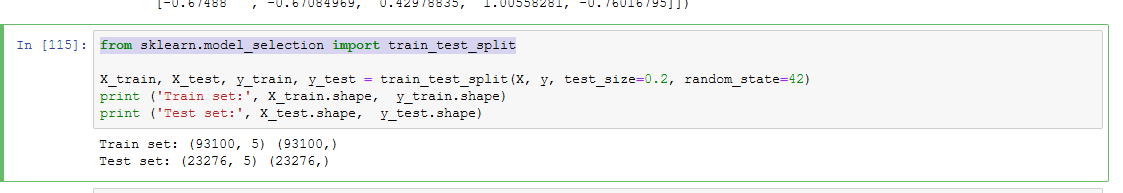
So we have done down sampling for severity code 1 class with sklearn's resample tool. We down sampled to match the severity code 2 exactly with 58188 values each.



With this I have completed data cleaning and balancing work.

Now I have to split the data into training data and testing data with a ratio of 80:20 which I have completed using the command

from sklearn.model\_selection import train\_test\_split



So the train data set consists of 93100 samples with 5 parameters and 93100 output labels and the test data consists of 23276 samples with 5 parameters and 23276 output labels.

Since we now we are ready with all necessary data , our next step is to build the models and analyse the performance of the system.