Road Mishap Risk Assessment

Instructor: Prof. Ralph Lano
Recent Trends and Technologies
(First Results Summary)

Monday, 8 June 2020

Created By:

1. Fakkiragouda J. Patil
Matrikel-Nr: 00013920

2. Manas Ranjan Chhotray
Matrikel-Nr: 00015920

Exploratory Data Analysis

- **What is Exploratory Data Analysis?**

Exploratory data analysis, or EDA, is a (mainly) visual approach and philosophy that focuses on the initial ways by which one should explore a data set or experiment. Two main aspects of EDA are:

Openness. A person exploring the data should be open to all possibilities prior to its exploration.

Skepticism. One must ensure that the obvious story the data tells is not misleading.

- **What is EDA Used For?**

EDA is used for:

- Catching mistakes and anomalies
- Gaining new insights into data
- Detecting outliers in data
- Testing assumptions
- Identifying important factors in the data
- Understanding relationships

- **Steps for Performing EDA:**

1. Examine the Data
2. Clean The Data
3. Data Wrangling
4. Build Data Profile Tables & Plots
5. Explore Data Relationships

1. Examine the Data:

We have the used different python library functionalities to do the below activities:

a) List Out the Data types

We have used the python pandas library functionalities to list down the data types of the data types :

2. Data Types

A. List out the Data Types

```
df.dtypes
```

```
Accident_Index                 object
1st_Road_Class                 object
1st_Road_Number               float64
2nd_Road_Class                 object
```

b) List Down the categorical attributes

```
df.Accident_Severity.unique()
```

```
array(['Serious', 'Slight', 'Fatal'], dtype=object)
```

c) Get the statistics according to the Data distribution.

1. List down the Categorical Attributes

```
df.nunique(axis=0)
```

```
Accident_Index                 2047256
1st_Road_Class                       6
1st_Road_Number                   7160
```

```
df.groupby('Accident_Severity').Road_Type.describe()
```

Accident_Severity	count	unique	top	freq
Fatal	26369	6	Single carriageway	20085
Serious	286339	6	Single carriageway	226950
Slight	1734548	7	Single carriageway	1280847

2. <u>Clean The Data</u>

Data cleaning, also called data cleansing or scrubbing, deals with detecting and removing errors and inconsistencies from data in order to improve the quality of data.

Some of the most important steps to be followed to achieve Data Cleaning are:

a) Removing Redundant Variables

b) Attribute Selection

c) Removing Outliers

d) Removing rows with null values

e) Removing Duplicate Rows

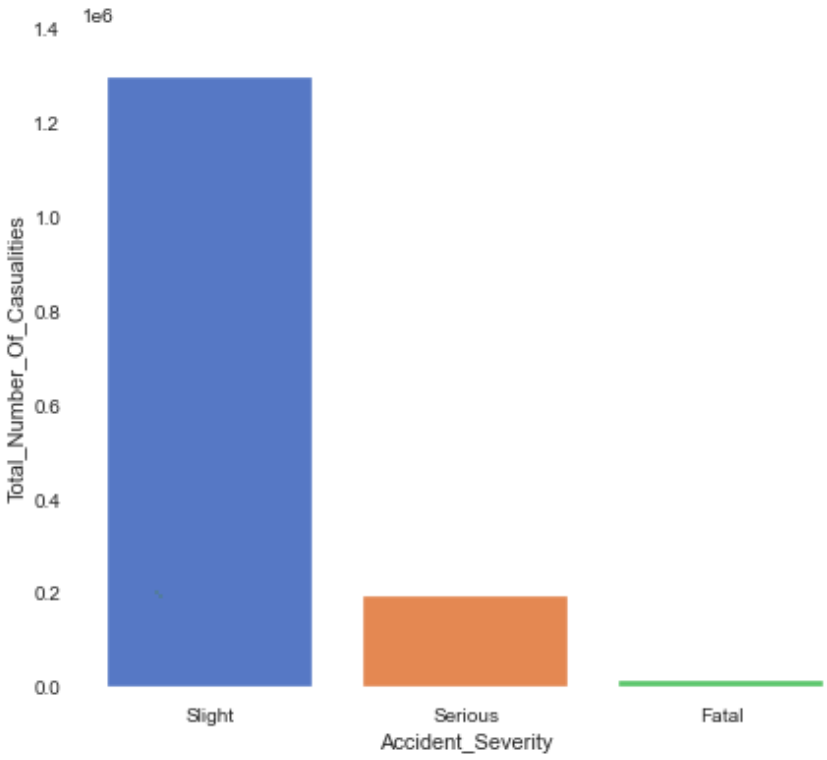
3. Data Wrangling

Data wrangling, sometimes referred to as data munging, is the process of transforming and mapping data from one "raw" data form into another format with the intent of making it more appropriate and valuable for a variety of downstream purposes such as analytics. A data wrangler is a person who performs these transformation operations.

4. Build Data Profile Tables & Plots

a) Analysis for Numerical Datas

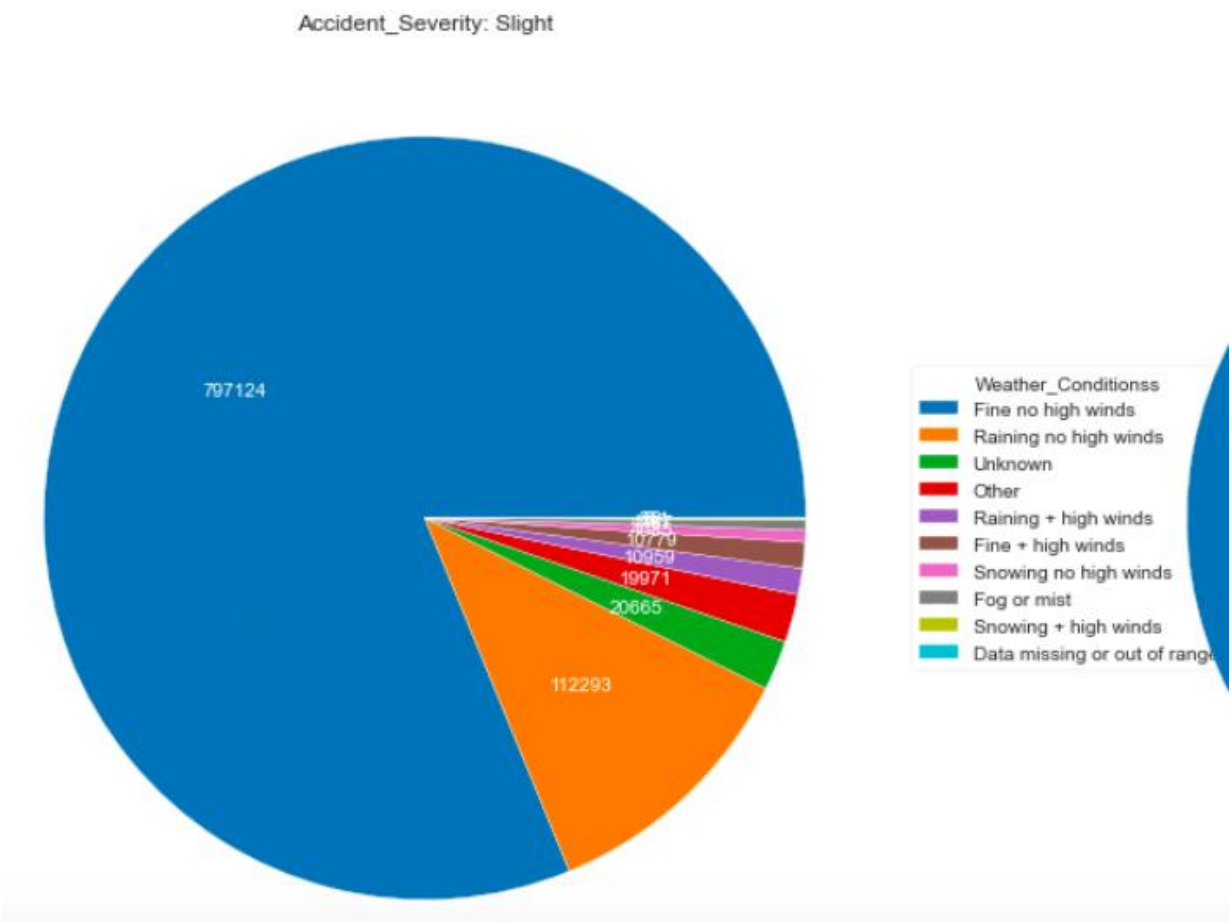
We have used line plot, Histogram and scatter plots to successfully analyse the Numerical Data's. We've used different aggregation functions such as Count, Max, Min, Average to understand the data in several possible directions, which can come in handy while selecting and performing the feature engineering.



For example, the above bar graph showing count of Total number of **Casualities(Y-Axis) vs Accident Severity(X-Axis)**

b) Pie Chart Analysis for Categorical Data

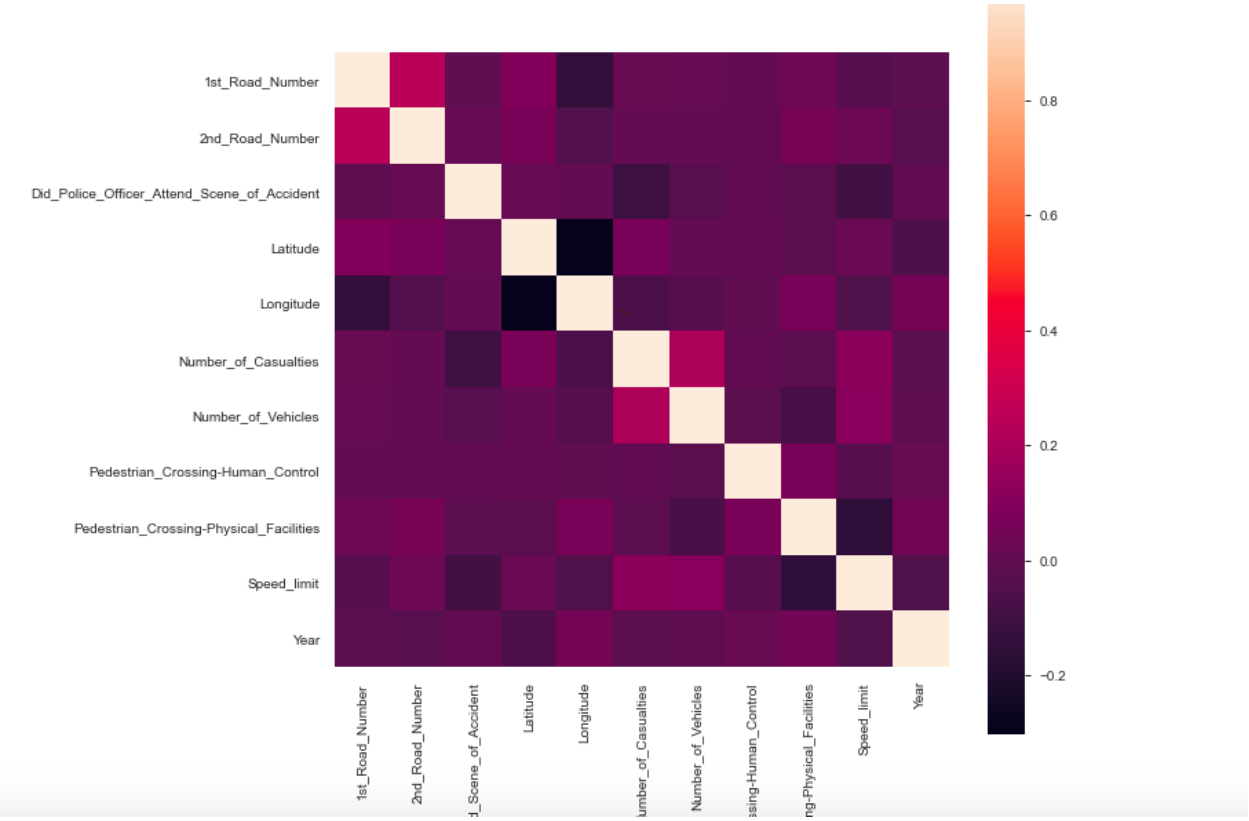
pie chart to compare the proportion of data in each category. A pie chart is a circle ("pie") that is divided into segments ("slices") to represent the proportion of observations that are in each category.



For Example in our case the above pie chart shows the proportions of accident counts w.r.t. Accident_Severity="Slight" classifications at various weather conditions.

c) Heat Map and Correlation-Matrix

A correlation matrix is a table showing correlation coefficients between variables. Each cell in the table shows the correlation between two variables. A correlation matrix is used to summarize data, as an input into a more advanced analysis, and as a diagnostic for advanced analyses.

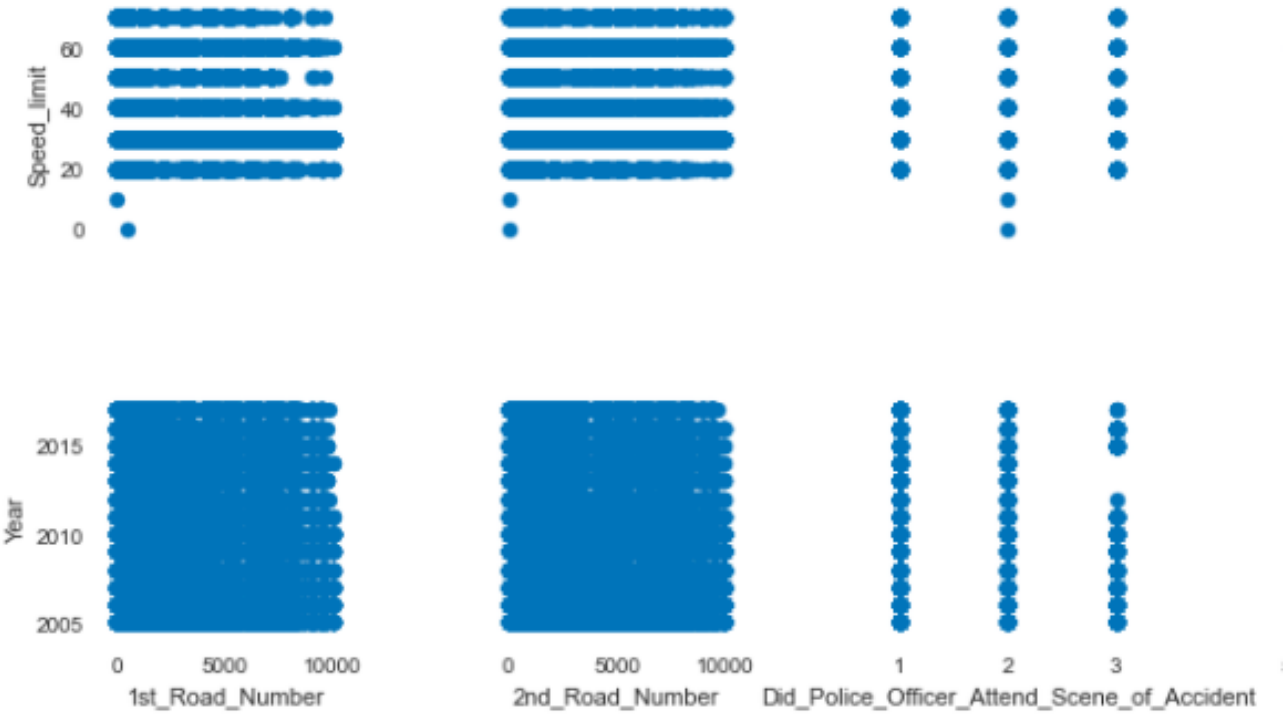


In our case we have computed a Pearsons Product moment correlation co-efficient matrix and its showing very low correlation between each and every attribute because most of the datas in our case are categorical.

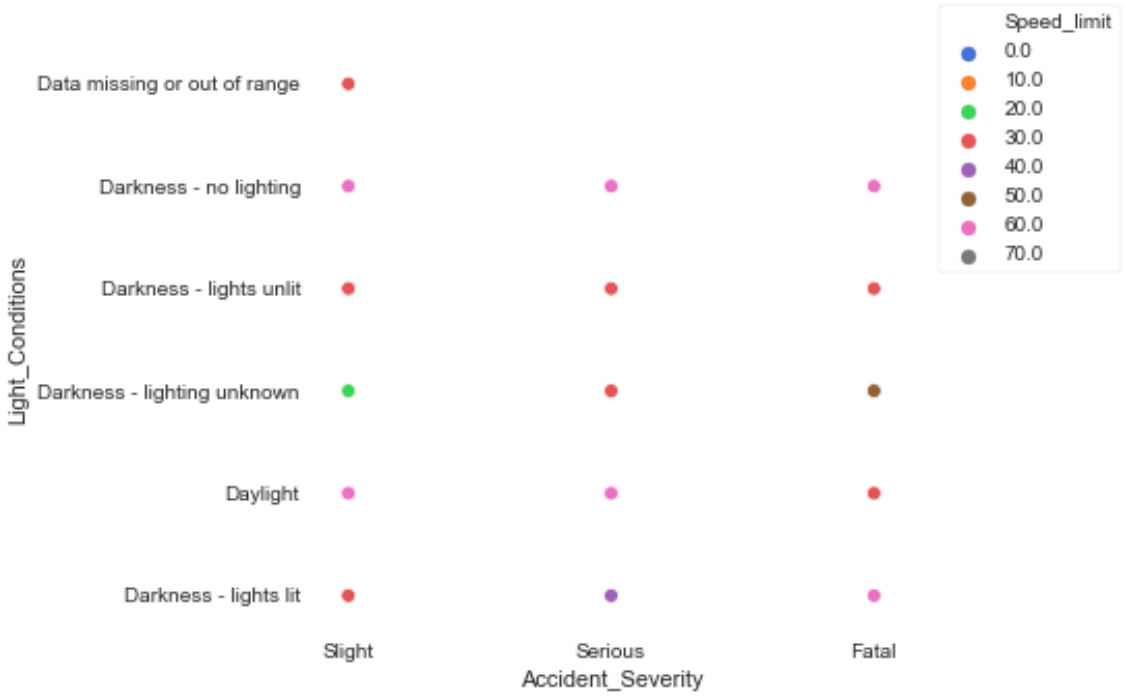
d) Scatter Plot

A scatter plot (also called a scatterplot, scatter graph, scatter chart, scattergram, or scatter diagram) is a type of plot or mathematical diagram using Cartesian coordinates to display values for typically two variables for a set of data.

2 Variable Scatter Plot:-



3 **Variable Scatter Plot:-**



In Our case we've used both 2 variable and 3 Variable scatter plots to detect hidden patterns more accurately. For example In the above 3 Variable scatter plot for some of the cases we can easily determine the **accident severity** level by taking **speed_limit** and **Light_Conditions** into deciding factors.

5. Explore Data Relationships

Data exploration is the fifth step in Exploratory Data Analysis and typically involves summarising the main characteristics of a data set, including its size, accuracy, initial patterns in the data and other attributes. Here we have plotted several relationship models using Python pandas, matplotlib and seaborn libraries to know how many cases are there in the data set, what variables are included, how many missing values there are and what general hypotheses the data is likely to support. An initial exploration of the data set can help answer these questions by familiarising analysts with the data with which they are working.

Once data exploration has uncovered the relationships between the different variables, we can continue the data mining process by creating and deploying data models to take action on the insights gained.

Feature Engineering

Feature engineering is the process of using domain knowledge to extract **features** from raw data via data mining techniques. These **features** can be used to improve the performance of machine learning algorithms. **Feature engineering** can be considered as applied machine learning itself.

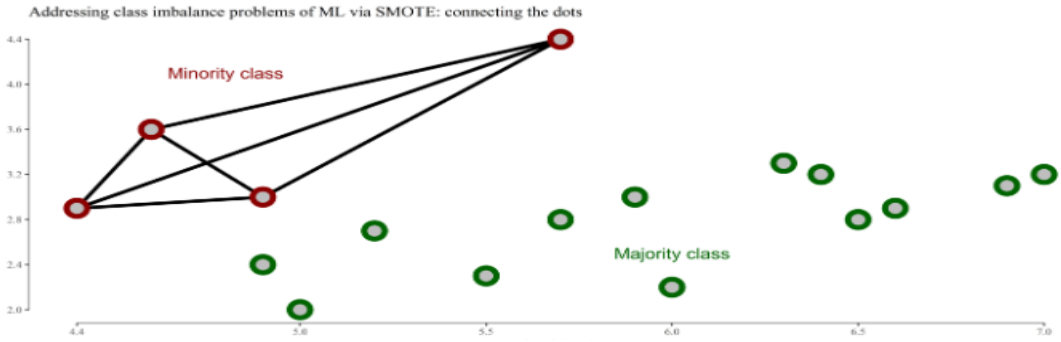
In our first Data experiment we analysed the data and identified that there are millions of records available, but after analysing we understood that there are around 33 columns of labelled data which are having duplicate set of values and there is less collinearity between the data. Currently our is to predict the accident severity by looking into the out target variable we identified that data is biased towards minority classes which is not good, so we applied SMOTE technique to do create the balanced dataset as PCA to reduce the dimensionality of the data, in future implementations we are going to apply K-Fold cross validation method to identify the best fit model to our data. The above mentioned are going to explain in detail in the following section with visualization images.

Synthetic Minority Oversampling Technique (SMOTE):

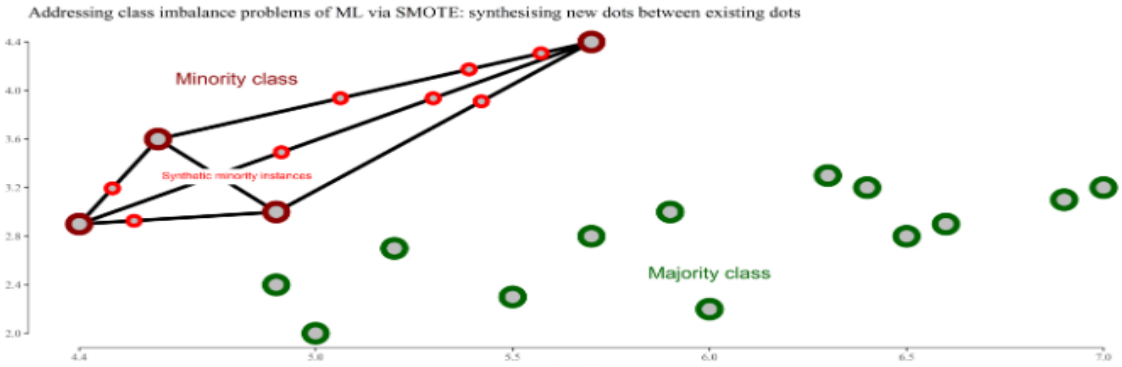
Definition: This is a statistical technique for increasing the number of cases in your dataset in a balanced way.

How it works:

SMOTE synthesizes new minority instances between existing (real) minority instances. Imagine that SMOTE draws lines between existing minority instances like this.



SMOTE then imagines new, synthetic minority instances somewhere on these lines.

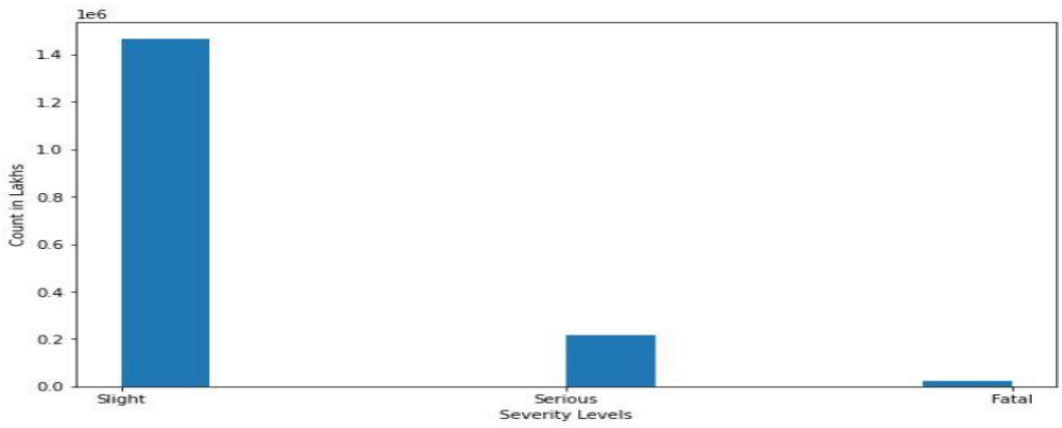


After synthesising new minority instances, the imbalance shrinks from 4 red versus 13 green to 12 red versus 13 green. Red flowers now dominate within the ranges typical for red flowers on both axes.

Our Experiment results on UK Accident Data:

```
# view previous class distribution
print('Before Applying SMOTE Technique:'), print(target['Accident_Severity'].value_counts())
```

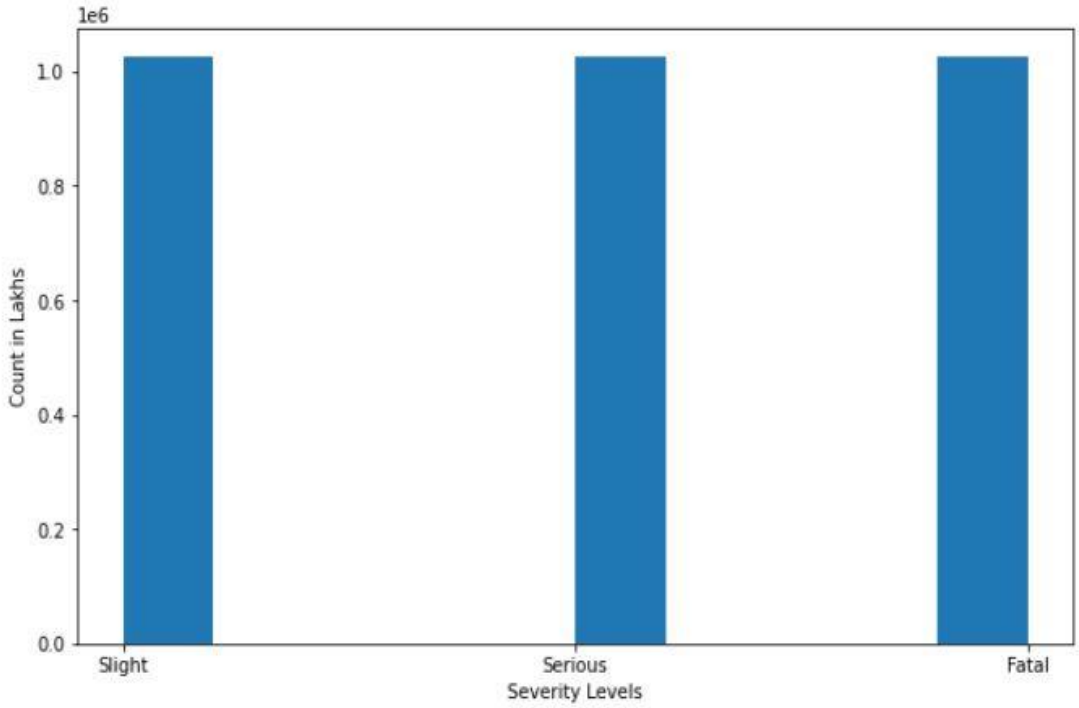
```
Before Applying SMOTE Technique:
Slight     1464234
Serious     216801
Fatal        23018
Name: Accident_Severity, dtype: int64
```



```
# resample data ONLY using training data
X_resampled, y_resampled = SMOTE().fit_sample(X_train, y_train)
```

```
# After SMOTE
print("After Applying SMOTE Technique:")
for i in y_resampled.columns:
    x = y_resampled[i].value_counts()
    print(x)
```

```
After Applying SMOTE Technique:
Slight     1025021
Serious    1025021
Fatal      1025021
Name: Accident_Severity, dtype: int64
```



Principal Component Analysis:

In our dataset we have 33 columns and relatively large dataset of 20 million records, so we tried to reduce the dimensionality of the data by applying PCA.

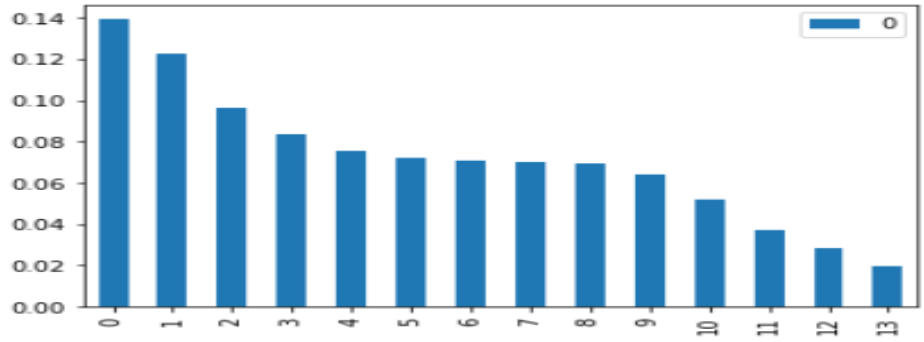
Definition: PCA is a technique used to emphasize variation and bring out strong patterns in a dataset. Its often easy to explore and visualize.

PCA creates the linear combination of original predictor variables which captures the maximum variance in the dataset and create the component one is PC1 and Second Principal component captures the remaining variance and so on.

We have performed the PCA on our dataset and identified that PCA is not a perfect fit because it contains lot of categorical and duplicate variables, so using PCA reducing dimensionality did not help us much but some of the insights are draws from the PCA and following are the visualization.

```
pca.explained_variance_ratio_
```

```
array([0.13949662, 0.12285003, 0.09658726, 0.08345402, 0.07516442,
       0.07195879, 0.07096458, 0.06978849, 0.06942289, 0.06373536,
       0.05158661, 0.03724324, 0.02837567, 0.019372  ])
```




```
pca.explained_variance_ratio_.sum()
```

```
0.999999999999999
```

Model Selection: Here we want to give further steps in identifying the best fit model to our data using K fold cross validation.

Definition: Cross-validation is a resampling procedure used to evaluate machine learning models on a limited data sample.

How it works?

1. Shuffle the dataset randomly.
2. Split the dataset into k groups
3. For each unique group:
    - Take the group as a hold out or test data set
    - Take the remaining groups as a training data set
    - Fit a model on the training set and evaluate it on the test set
4. Retain the evaluation score and discard the model
5. Summarize the skill of the model using the sample of model evaluation scores

So, in the future development will follow the above steps and identify the best fit model.