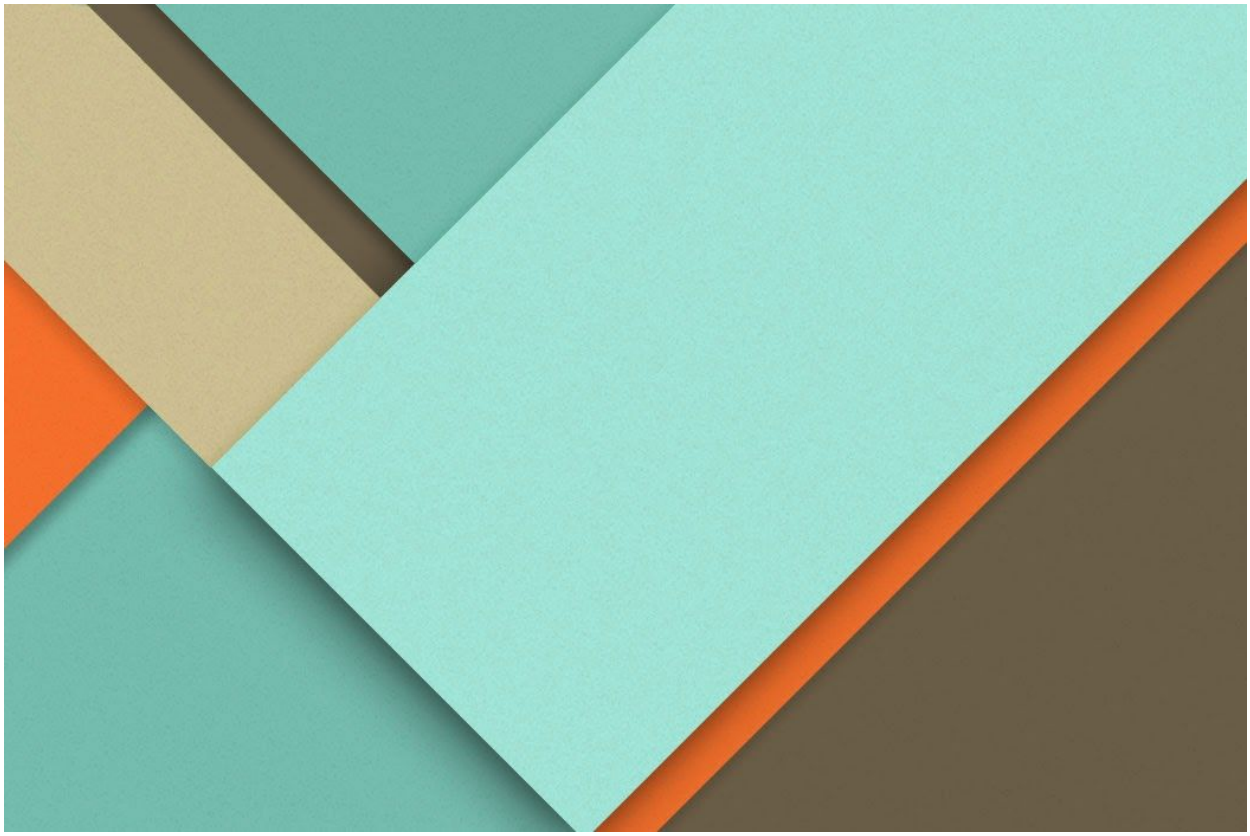


Capstone Project - Car Accident Severity

IBM - Data Science Specialization

- Sandeep krishna Donepudi

About the Dataset

The data for this project is taken from an open source website -

https://data-seattlecitygis.opendata.arcgis.com/datasets/5b5c745e0f1f48e7a53acec63a0022ab_0/data

The data consists of 39 independent variables and 221525 rows. The dependent variable, “SEVERITYCODE”, contains codes that correspond to different levels of severity caused by an accident.

```
df['SEVERITYCODE'].value_counts()
```

```
1     137671
2      58783
0      21615
2b      3105
3        350
Name: SEVERITYCODE, dtype: int64
```

Severity codes are as follows:

- 0 : Unknown
- 1 : Property Damage Only Collision
- 2 : Injury Collision
- 2b : Serious Injury Collision
- 3 : Fatality Collision

```
df['SEVERITYDESC'].value_counts()
```

```
Property Damage Only Collision    137671
Injury Collision                   58783
Unknown                            21616
Serious Injury Collision            3105
Fatality Collision                   350
Name: SEVERITYDESC, dtype: int64
```

Furthermore, because of the existence of null values in some records, the data needs to be preprocessed before any further processing.

Data Preprocessing

The dataset in the original form is not ready for data analysis. In order to prepare the data, first, we need to drop the non-relevant columns. In addition, most of the features are of object data types that need to be converted into numerical data types. We have to convert the **SEVERITY CODE** our target variable into numerical data type too.

For this we have updated the value of **SEVERITYCODE** '2b' to '4' and then updated the column data type.

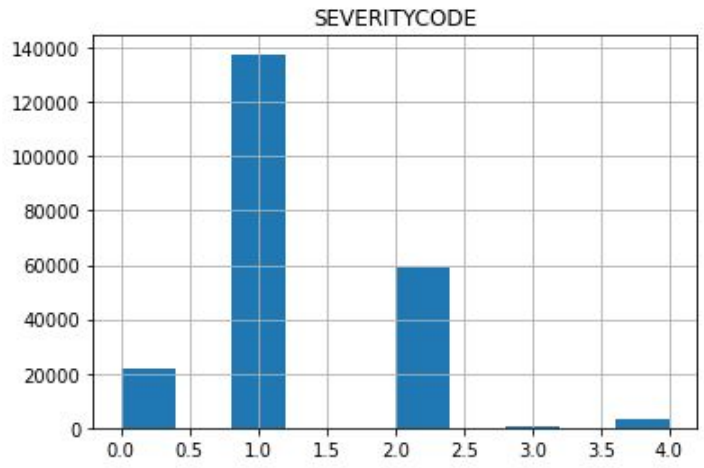
```
df['SEVERITYCODE'].value_counts()
```

```
1    137671
2     58783
0     21615
4      3105
3       350
Name: SEVERITYCODE, dtype: int64
```

To get a good understanding of the dataset, I have checked different values in the features. The results show the target feature is imbalance, so we use a simple statistical technique to balance it.

```
df.hist(column='SEVERITYCODE')
```

```
array([[<matplotlib.axes._subplots.AxesSubplot object at 0x000001A62DC566D0>]],
      dtype=object)
```



To get a good understanding of the dataset, I have checked different values in the features. The results show the target feature is imbalance, so we use a simple statistical technique to balance it.