Full Report: Car Accident Severity
By
Reda Belhaj

I- Introduction

Road Accident is the most undesirable and unexpected thing to occur to a road user, though they happen quite often. Unfortunately, we can see a minatory rise of road accidents conspicuously highroad accidents over the past few years. It has a massive impact on society as well as in the economy as there is an immense cost of fatalities and injuries. according to WHO, the economic cost of road accidents to a developing country like us is 2-3% of GDP, which is a significant loss for a country like ours. Moreover, reducing this loss has become a great matter of concern for our country now.

Problem

Predicting Car accident severity might help fixing and reducing the cost of accident and its impact on society and economy.

Data Science might help to Predict Car accident Severity this project aim to do so.

II- Data acquisition and cleaning

1. Data sources

The data is qll collisions provided by SPD and recorded by Traffic Records of Seattle this includes all types of collisions. Collisions will display at the intersection or mid-block of a segment.

The timeframe: 2004 to Present.

2. Data Cleaning

For the data our investigation led to only interested in the columns Severity Code, Weather Elements, Light Condition, Adress Type, Road Condition and Collission Type.

We have identified

4904 missing values for the collision Type column

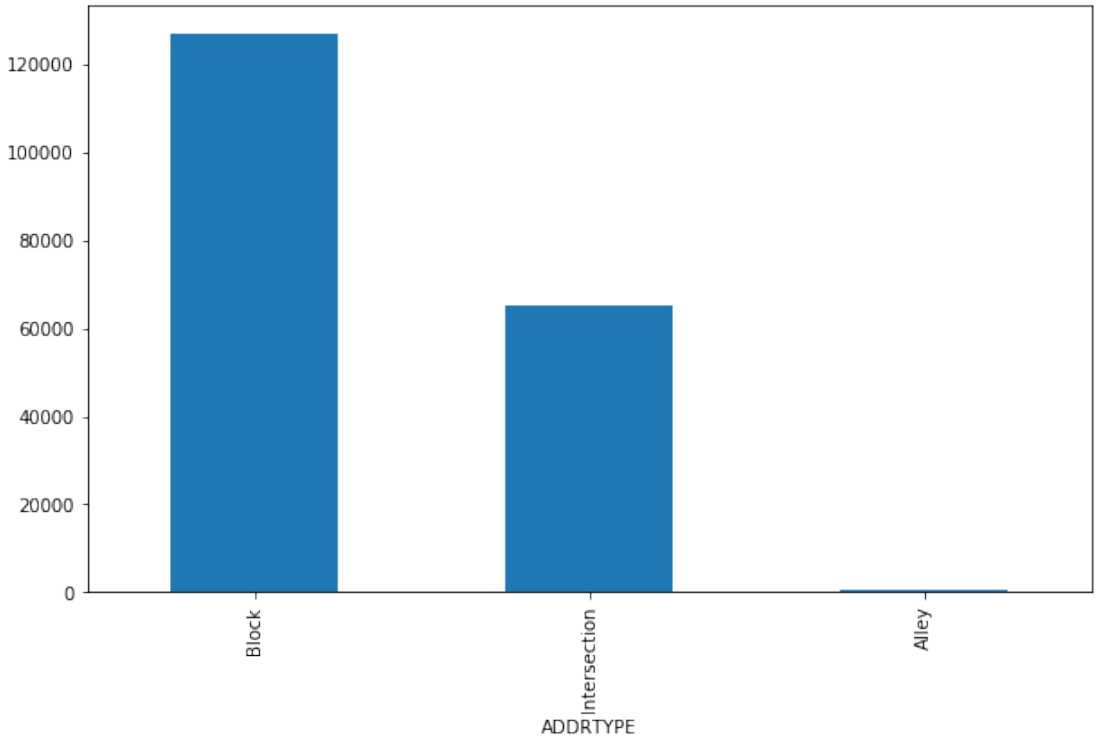
1926 missing values for address type column

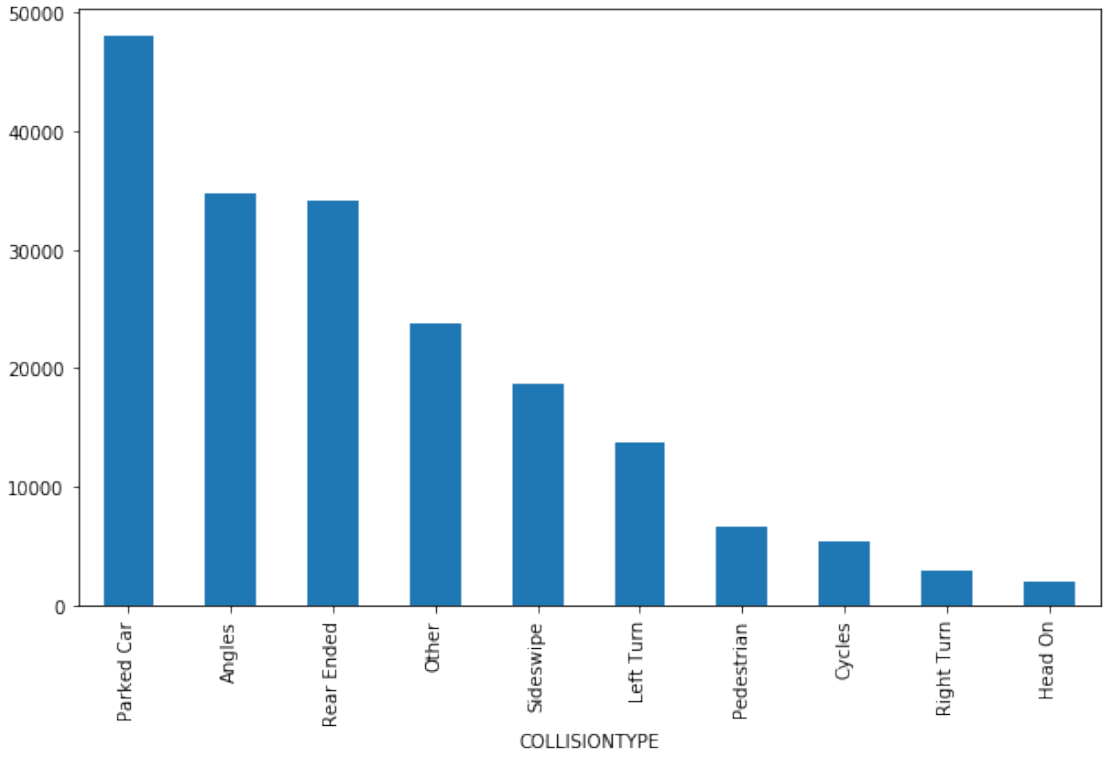
5081 missing values for weather column

5012 missing values for Road condition column

5170 missing values for Light condition

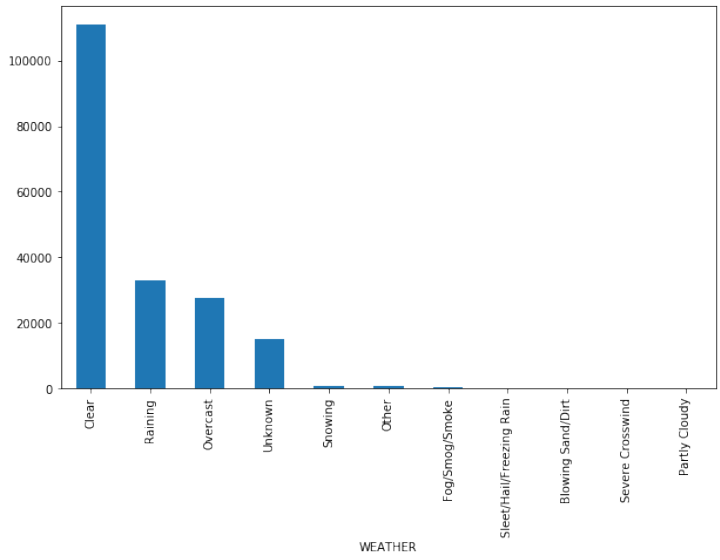
For Collision Type & **address type column the distribution of the data is as indicated by these bar charts**

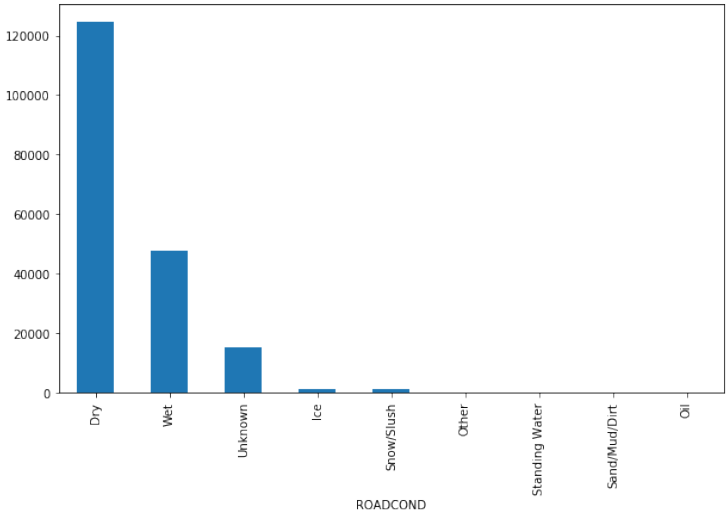


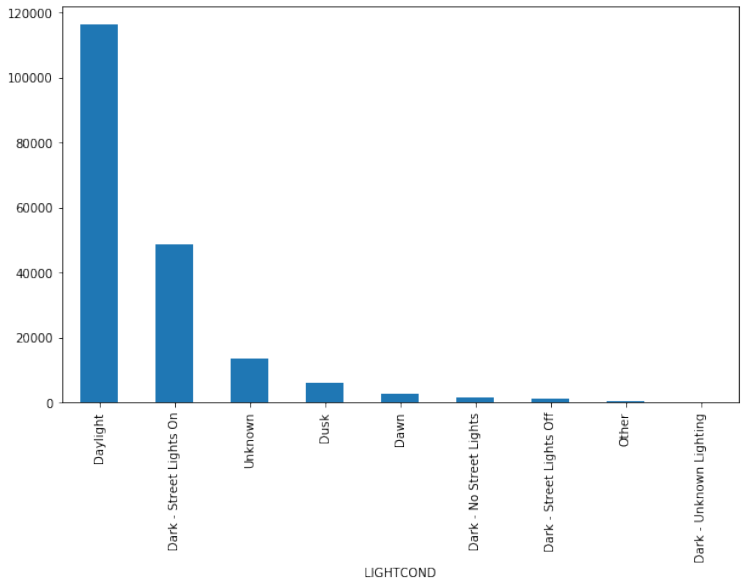


I concluded that there is no clear mode to replace the missing values with There for I will just drop the rows with these missing values.

For Weather & **Road Condition and Light Condition columns the distribution of the data is as indicated by these bar charts**







There are some clear modes in each of the three columuns

So I decided to replace the mission data as follow

For the weather column I will replace the missing values with ‘Clear’

For the Road Condition column I will replace the missing values with ‘Dry’

For the light Condition column I will replace the missing values with ‘Daylight’

For the accurate prediction of the severity of accidents, a considerable number of traffic accident records with full information is required to train by using the proposed approaches.
The entire dataset will split into two parts- Training Dataset and Test Dataset. 70% of the whole dataset has been chosen randomly by using a python library as a training data set and the remaining 30% has been used as our test dataset. We have used the 70-30 ratio for splitting dataset because of its proven accuracy.

A. Methodology

As a database, I used GitHub repository in my study. My master data which has the main components **Severity Code, Weather Elements, Light Condition, Adresse Type, Road Condition and Collission Type**.
After cleaning the data we have 187950 traffic accidents to work on.

```
In [37]: data.head()
```

Out[37]:

	SEVERITYCODE	COLLISIONTYPE	ADDRTYPE	WEATHER	ROADCOND	LIGHTCOND
0	2	Angles	Intersection	Overcast	Wet	Daylight
1	1	Sideswipe	Block	Raining	Wet	Dark - Street Lights On
2	1	Parked Car	Block	Overcast	Dry	Daylight
3	1	Other	Block	Clear	Dry	Daylight
4	2	Angles	Intersection	Raining	Wet	Daylight

```
In [38]: data.shape
```

```
Out[38]: (187950, 6)
```

Discussion

We observe that most of the accidents in our dataset have the severity code as “prop damage” and “injury” for the other three classes there is none.

For the accurate prediction of the severity of accidents, a considerable number of traffic accident records with full information is required to train by using the proposed approaches.
The entire dataset will split into two parts- Training Dataset and Test Dataset.

70% of the whole dataset has been chosen randomly by using a python library as a training data set and the remaining 30% has been used as our test dataset. We have used the 70-30 ratio for splitting dataset because of its proven accuracy.

III- Exploratory Data Analysis

I will be selecting for for this project two machine learning algorithms

Decision tree

And k-nearest neighbor

1- **Decision Tree to predict Accident Severity**

I have split the data into two part training set which is 70% of the whole dataset and testing set which is 30%.
After preprocessing the data into numerical labels I run the algorithm and tested the accuracy of the model

```
predTree = AccTree.predict(X_testset)
```

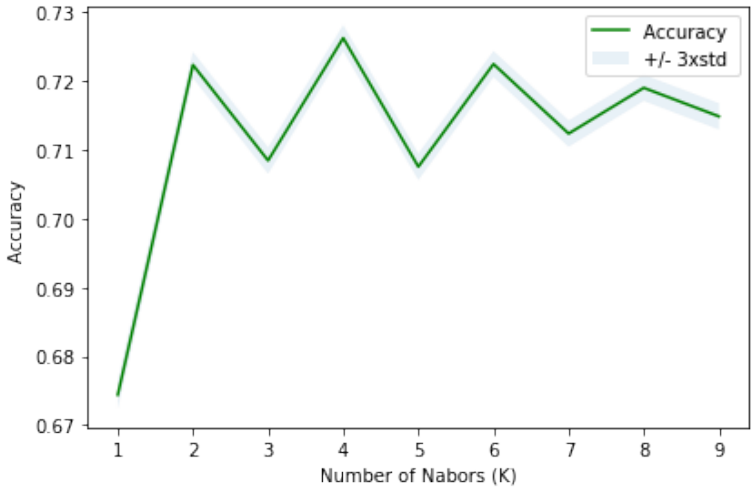
```
from sklearn import metrics
import matplotlib.pyplot as plt
print("DecisionTrees's Accuracy: ", metrics.accura
```

```
DecisionTrees's Accuracy:  0.7224793828145784
```

2. K Neighbors

There for the accuracy for the decision tree to find the severity code for the accident is **0.7225**

2- K-Nearest Neighbor
For this machine learning model I have run this algorithm with different value for k and tested the accuracy for each value



It is clear that the best k value for this model is k=4.
With this the accuracy of the model is **0.7149.**