INTRODUCTION

According the World Health Organization, "there were 1.35 million road traffic deaths globally as at 2016, with millions more sustaining serious injuries and living with long-term adverse health consequences. Globally, road traffic crashes are a leading cause of death among young people, and the main cause of death among those aged 15–29 years. Road traffic injuries are currently estimated to be the eight leading cause of death across all age groups globally, and are predicted to become the seventh leading cause of death by 2030." Source: WHO

Can we determine the common causes of road accidents, and develop ways to prevent them from happening? This analysis would be using accident data from the Seattle City Police Department as a case study to enable road safety officials such as the police and other law enforcement officials understand better the causes of road accidents and predict the possibility and severity of an accident, based on the daily weather patterns and the road conditions so as to enable them take the necessary safety precautions.

DATA ACQUISITION AND CLEANING

Data sources: The data from the Seattle Police Department covers a period over 15+ years with 190,000 observations made during the said period. Data was obtained from the data links on the coursera website in the form of a CSV file and a Metadata file. The data set was made up of 194,674 rows and 38 attributes. To accurately build a model to better understand the common causes of road accidents and help predict future accidents and reduce the severity, the following attributes would be selected: ADDRTYPE, PERSONCOUNT, VEHCOUNT, WEATHER, and ROADCOND.

```
[6]: df.shape

[6]: (194673, 38)
```

Figure 1: Data Frame Shape

Data cleaning

The Jupyter Notebooks was chosen because of its usage easy. Pandas, Numpy, Matplotlib, Sklearn and Seaborn were determined as necessary tools for this analysis, they were therefore imported into the Python Library. Noticing that the data to be used for this analysis were already in category format, using a graphical illustration such as histogram was chosen to show the correlation between variables.

The first step was importing the CSV file containing the data into the Notebook. After determining the total number of rows and columns, I proceeded to drop all rows and columns that would be irrelevant to the analysis. Missing data's were also found out and dropped from the data set.

```
[15]: #Drop data that are either irrelevant or the True value is more than 20%

to_drop =['SPEEDING', 'PEDROWNOTGRNT', 'INATTENTIONIND', 'INTKEY', 'SDOTCOLNUM',
          'INATTENTIONIND', 'JUNCTIONTYPE', 'EXCEPTRSNCODE', 'X', 'Y', 'OBJECTID',
           'COLDETKEY', 'EXCEPTRSNDESC', 'INCDATE', 'INCDTTM', 'SDOT_COLCODE',
          'SDOT_COLDESC', 'SDOTCOLNUM', 'ST_COLCODE', 'ST_COLDESC', 'SEGLANEKEY',
           'CROSSWALKKEY', 'INTKEY', 'REPORTNO', 'STATUS', 'HITPARKEDCAR', 'LOCATION',
           'SEVERITYDESC', 'COLLISIONTYPE', 'INCKEY', 'PEDCOUNT', 'PEDCYLCOUNT',
           'SEVERITYCODE.1', 'UNDERINFL', 'LIGHTCOND']
df.drop(to_drop, axis = 1, inplace = True)
```

Figure 2: Dropped Data

While speed should always be considered when carrying out analysis on vehicular accidents, for this analysis, speed was dropped from the data set because it had a total of 185,340 missing data. Retaining such data in the data set would have adverse effect in the analysis.

15. "SPEEDING": 185340 missing data

Figure 3: Missing Speed Data

Feature selection

After dropping the irrelevant data from the dataset, it was discovered that WEATHER and ROADCOND contained values of "unknown" data of 15,091 and 15,078 respectively. These also were dropped because of the potential effect it would have on the analysis.

```
[139]: # Remove values from ROADCOND because they are unknown
       df = df [df['ROADCOND'] != 'Unknown']

[140]: # Remove values from WEATHER because they are unknown
       df = df [df['WEATHER'] != 'Unknown']
```

Figure 4: Unknown Data

A further analysis of the dataset showed that some of the attributes contained blank cells. This had to be dropped from the dataset.

```
[144]: df.isnull().sum(axis = 0)

[144]: SEVERITYCODE       0
       ADDRTYPE         874
       PERSONCOUNT        0
       VEHCOUNT           0
       WEATHER         5071
       ROADCOND        5001
       dtype: int64

[145]: #Drop all null values
       df.dropna(inplace=True)

[146]: df.isnull().sum(axis = 0)

[146]: SEVERITYCODE    0
       ADDRTYPE        0
       PERSONCOUNT     0
       VEHCOUNT        0
       WEATHER         0
       ROADCOND        0
       dtype: int64
```

Figure 5: Showing total number of blank cells

With an initial 194,673 rows and 38 columns that was uploaded, after engaging in the cleaning process, we were left with 172,242 row and 6 columns for the analysis.

```
[147]: df.info()

<class 'pandas.core.frame.DataFrame'>
Int64Index: 172242 entries, 0 to 194672
Data columns (total 6 columns):
 #   Column         Non-Null Count   Dtype
---  ------         --------------   -----
 0   SEVERITYCODE   172242 non-null  int64
 1   ADDRTYPE       172242 non-null  object
 2   PERSONCOUNT    172242 non-null  int64
 3   VEHCOUNT       172242 non-null  int64
 4   WEATHER        172242 non-null  object
 5   ROADCOND       172242 non-null  object
dtypes: int64(3), object(3)
memory usage: 9.2+ MB
```

Figure 6 Final Data Frame Info

EXPLORATORY DATA ANALYSIS

Having completed the cleaning process, I proceeded to understanding the relationship between vehicles involved in an accident and the type of accidents that they were involved in. Other analysis carried out where:

- Find out the number of persons involved in an accident and the type of accident
- Where most likely an accident might occur
- The number of accidents during different types of weather
- The number of accidents under different road conditions

Number of vehicles in an accident:

Matplotlib and Seaborn were used in presenting the graphical illustration in the form of histogram while the data used was VEHCOUNT. According to the ArcGIS Metadata Form, this indicated the number of vehicles involved in the collision. From the histogram chart, I could see that out of a total of 172,242 accidents, over 120,000 accidents occurred between 2 -3 vehicles.

```
[49]: bins = np.arange(df.VEHCOUNT.min(),8,1)
      plt.hist(df.VEHCOUNT, bins = bins)

      plt.title('No of Vehicles In An Accidents')
      plt.ylabel('Number of Accidents')
      plt.xlabel('Number of Vehicle')
```

```
[49]: Text(0.5, 0, 'Number of Vehicle')
```

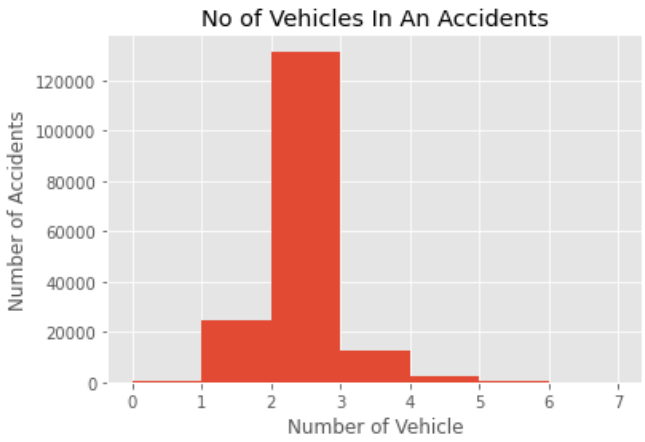

Figure 7: Number of Vehicles Involved In An Accident

Number of persons involved:

From a total of 172,242 accidents recorded, over 120,000 accidents recorded between 2 – 4 persons as victims of accidents. A further probe into this number saw that 95,947 of these accidents had just 2 persons as victims. While 34,189 of the 120,000 accidents recorded 3 persons as victims of the accidents.

```
[57]: bins = np.arange(df.PERSONCOUNT.min(),17,2)
      plt.hist(df.PERSONCOUNT, bins = bins)

      plt.title('No of People In Accidents')
      plt.ylabel('Number of Accidents')
      plt.xlabel('Number of People')

[57]: Text(0.5, 0, 'Number of People')
```

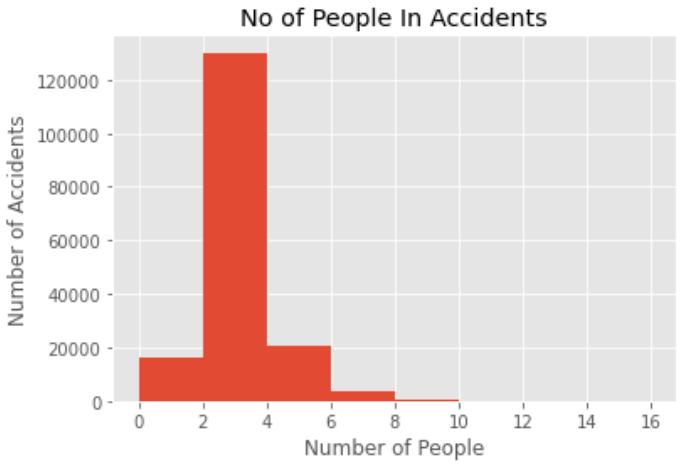

Figure 8: Number of Persons Involved In An Accidents

ACCTYPE:

Under the ArcGIS Metadata Form, ACCTYPE was defined as the collision address area. This was either at a Block, Alley or Intersection. The idea behind this analysis was, identifying the most like area that an accident would occur. From the total sample dataset, vehicular accidents are more likely to occur at an intersection. Accidents recorded at an intersection were over 100,000 while 60,000 accidents were recorded at a block.

```
[43]: X = df.ADDRTYPE.unique()
      Data = df.ADDRTYPE.value_counts()
      plt.bar(X, height=Data)
      plt.xlabel('Location')
      plt.ylabel('No of Accidents')
      plt.title('No of Accidents In Reltions to Locations')

[43]: Text(0.5, 1.0, 'No of Accidents In Reltions to Locations')
```

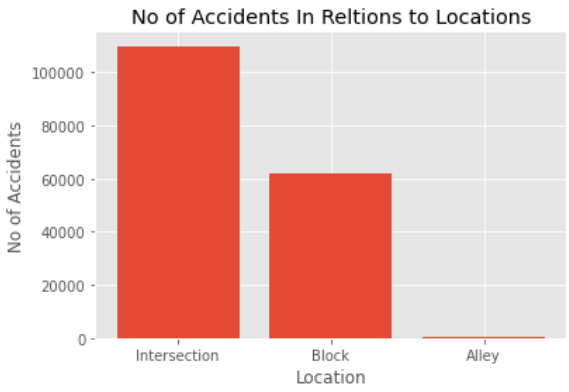

Figure 9: ACCTYPE Occurrence

Weather:

When asked to guest under which weather condition would more accidents be recorded? It won't be out of place to expect such answers as during a rainy day or snowy day or evening during a foggy day because of the seemly poor visibility during those periods. From the histogram chart, it was found that an overcast weather witnessed over 100,000 accidents, while a rainy and clear skies days would both witness over 20,000 but less 40,000 accidents

```
Data = df.WEATHER.value_counts()
plt.bar(X, height=Data)
plt.xlabel('Weather')
plt.ylabel('No of Accidents')
plt.title('No of Accidents In Reltions to Weather')
plt.xticks(rotation= 90)
plt.show()
```

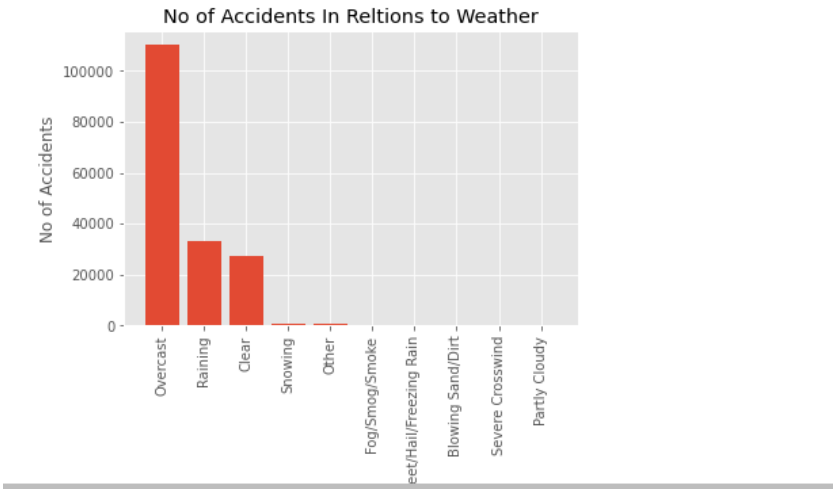

Figure 10: Number of Accidents Occurring During Various Weather Conditions

ROADCOND:

The condition of road can't be ignored. Under what road conditions were accidents recorded mostly? From the histogram chart, with total of over 120,000 accidents recorded, accidents a most likely to occurred under a wet road condition than under any other condition. A road condition of dry which was the closes had a record of less than 50,000 accidents.

```
Data = df.ROADCOND.value_counts()
plt.bar(X, height=Data)
plt.xlabel('Road Condiction')
plt.ylabel('No of Accidents')
plt.title('No of Accidents In Reltions to Road Condiction')
plt.xticks(rotation= 90)
plt.show()
```

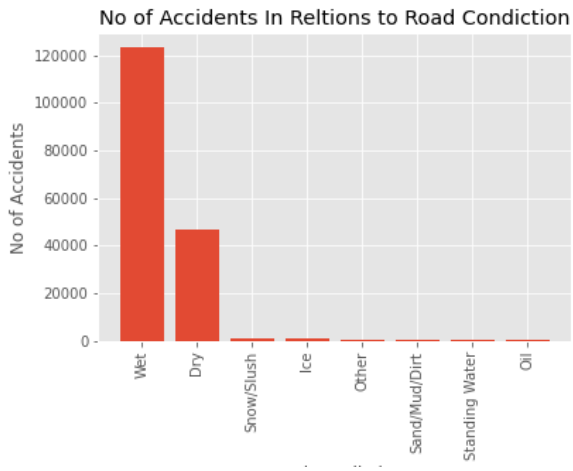

Figure 11: Number of Accidents Occurring Under Various Road Conditions

ACCIDENT SEVERITY:

What then are severity of the accidents in our dataset? From the attribute of our data set, we examined the severity of accident locations and persons involved in in accidents. What I noticed was that accident that occurred in alley's witnessed very low property damage and treat to life of persons involved in the accidents. On the other hand, with over 60,000 vehicle and over 80,000 person count accident at a Block, accidents occurring at this location have high rate of damaging properties. This was indicated by red color for level 1 and the blue color for level 2. While recorder accident at an intersection is more likely to lead to injuries to individuals than property damage.

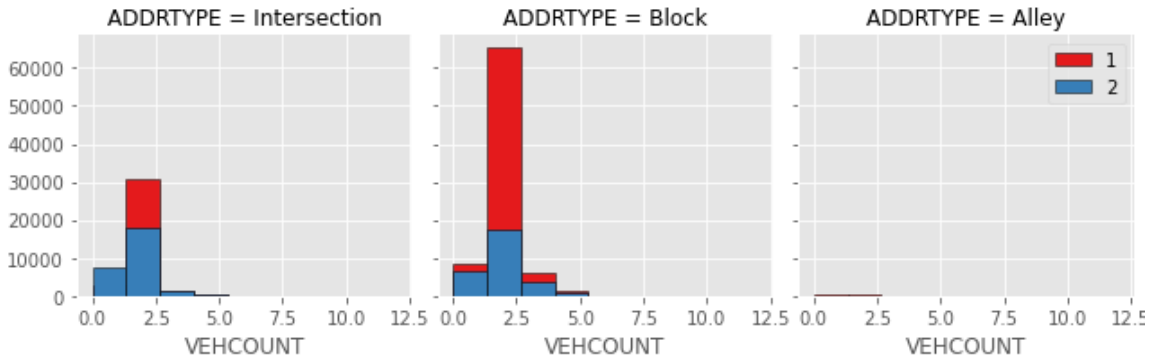

Figure 12: ACCTYPE SEVERITY

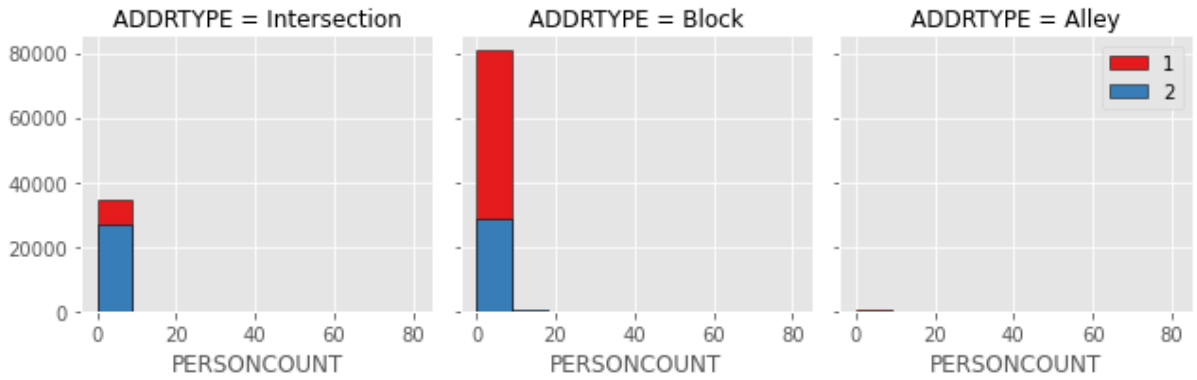

Figure 13: Person's Count Severity

DISCUSSION

Based on road, weather conditions, and other factors, I tried to figure out the seriousness and occurrence of road accidents, at the beginning of this analysis. Even though our data was a good size, there were a number of missing elements that needed to be cleaned in order to get a good result. 'SPEED' had to be dropped because there were too many missing elements but I think speed is an important factor that should be considered. The analysis shows that most accidents occur at intersections involving solo drivers, due to bad weather, on wet roads, and are minor in nature. Such information could be helpful to the police department in deciding where to add cameras at intersections, and install more stop signs, to induce people to slow down. We also live in a

technologically friendly world so maybe we can develop some inbuilt technology in our cars that warn us when the road and weather conditions are bad, or the car is approaching a stop sign.

CONCLUSION

Despite this analysis, there needs to be a closer inspection of certain other variables for insights. There is a considerable amount of property loss, however a lot of these accidents are minor and avoidable according to the analysis. These findings could be helpful to the Seattle Police Department in enforcing some new measures to prevent future accidents.