# IBM Data Science Capstone:

# Car Accident Severity Prediction

## 1. Introduction

Traffic accidents are a significant source of deaths, injuries, property damage, and a major concern for public health and traffic safety. Accidents are also a major cause of traffic congestion and delay. Effective management of accident is crucial to mitigating accident impacts and improving traffic safety and transportation system efficiency. Accurate predictions of severity can provide crucial information for emergency responders to evaluate the severity level of accidents, estimate the potential impacts, and implement efficient accident management procedures. By recognizing the key factors that influence accident severity, the solution may be of great utility to various Government Departments/Authorities like DOT and Police. The results of analysis and modeling can be used by these Departments to take appropriate measures; such as early warning system to drivers; to reduce accident impact and thereby improve traffic safety. It is also useful to the Insurers in terms of reduced claims and better underwriting as well as rate making.

## 2. Data Understanding

The dataset come from City of Seattle Open Data Portal that contains all types of collisions from 2004 to Present. This raw dataset consists of 221,266 cases and 40 attributes. The attributes is described in this link https://www.seattle.gov/Documents/Departments/SDOT/GIS/Collisions_OD.pdf

The dependent variable is SEVERITYCODE because it is used to measure the severity of the traffic accident.

| | | |
|---|---|---|
| SEVERITYCODE | Text, 100 | A code that corresponds to the severity of the collision:<br>• **3**—fatality<br>• **2b**—serious injury<br>• **2**—injury<br>• **1**—prop damage<br>• **0**—unknown |

## Acquiring the data

[1]:
```
!wget -O collisions.csv https://opendata.arcgis.com/datasets/5b5c745e0f1f48e7a53acec63a0022ab_0.csv
```

```
--2020-09-09 15:30:02--  https://opendata.arcgis.com/datasets/5b5c745e0f1f48e7a53acec63a0022ab_0.csv
Resolving opendata.arcgis.com (opendata.arcgis.com)... 34.202.76.40, 34.235.215.225, 54.152.131.176
Connecting to opendata.arcgis.com (opendata.arcgis.com)|34.202.76.40|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: unspecified [text/csv]
Saving to: 'collisions.csv'

collisions.csv          [             <=>    ]  80.92M  17.2MB/s    in 4.6s

2020-09-09 15:30:07 (17.6 MB/s) - 'collisions.csv' saved [84855377]
```

## Load collisions data into dataframe

[ ]:
```
df = pd.read_csv('collisions.csv')
df.head()
```

[11]:
```
print("Number of cases: %d and number of attributes: %d" %(df.shape[0], df.shape[1]))
```

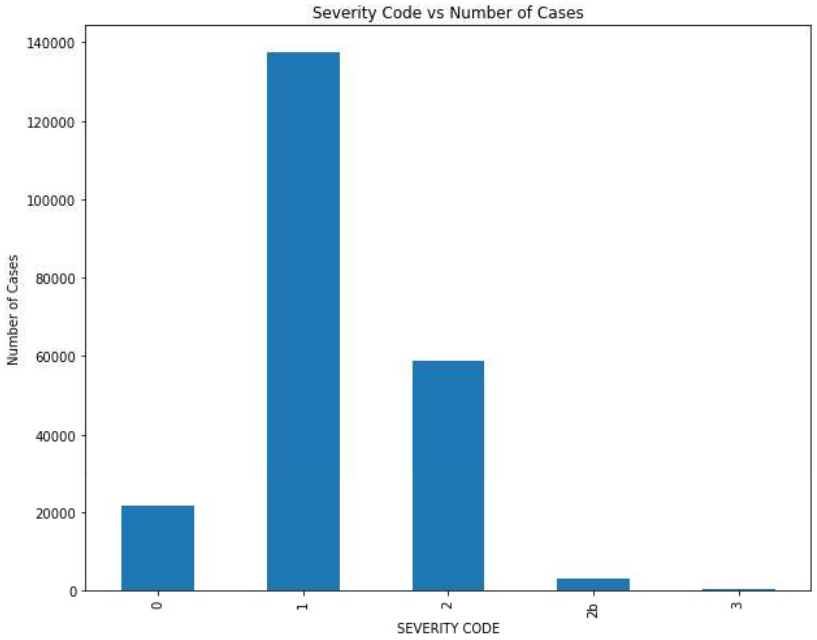
```
Number of cases: 221266 and number of attributes: 40
```

[13]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 221266 entries, 0 to 221265
Data columns (total 40 columns):
 #   Column           Non-Null Count   Dtype
---  ------           --------------   -----
 0   X                213797 non-null  float64
 1   Y                213797 non-null  float64
 2   OBJECTID         221266 non-null  int64
 3   INCKEY           221266 non-null  int64
 4   COLDETKEY        221266 non-null  int64
 5   REPORTNO         221266 non-null  object
 6   STATUS           221266 non-null  object
 7   ADDRTYPE         217554 non-null  object
 8   INTKEY           71823 non-null   float64
 9   LOCATION         216680 non-null  object
 10  EXCEPTRSNCODE    100863 non-null  object
 11  EXCEPTRSNDESC    11775 non-null   object
 12  SEVERITYCODE     221265 non-null  object
 13  SEVERITYDESC     221266 non-null  object
 14  COLLISIONTYPE    194767 non-null  object
 15  PERSONCOUNT      221266 non-null  int64
 16  PEDCOUNT         221266 non-null  int64
 17  PEDCYLCOUNT      221266 non-null  int64
 18  VEHCOUNT         221266 non-null  int64
 19  INJURIES         221266 non-null  int64
 20  SERIOUSINJURIES  221266 non-null  int64
 21  FATALITIES       221266 non-null  int64
 22  INCDATE          221266 non-null  object
 23  INCDTTM          221266 non-null  object
 24  JUNCTIONTYPE     209299 non-null  object
 25  SDOT_COLCODE     221265 non-null  float64
 26  SDOT_COLDESC     221265 non-null  object
 27  INATTENTIONIND   30188 non-null   object
 28  UNDERINFL        194787 non-null  object
 29  WEATHER          194578 non-null  object
 30  ROADCOND         194658 non-null  object
 31  LIGHTCOND        194490 non-null  object
 32  PEDROWNOTGRNT    5188 non-null    object
 33  SDOTCOLNUM       127205 non-null  float64
 34  SPEEDING         9913 non-null    object
 35  ST_COLCODE       211853 non-null  object
 36  ST_COLDESC       194767 non-null  object
 37  SEGLANEKEY       221266 non-null  int64
 38  CROSSWALKKEY     221266 non-null  int64
 39  HITPARKEDCAR     221266 non-null  object
dtypes: float64(5), int64(12), object(23)
memory usage: 67.5+ MB
```

## Displaying Dependent Variable

[43]:
```
plt.figure(figsize=(10,8))
skw = df.SEVERITYCODE.value_counts().reindex(['0','1','2','2b','3'])
skw.plot(kind='bar')
plt.xlabel('SEVERITY CODE')
plt.ylabel('Number of Cases')
plt.title("Severity Code vs Number of Cases");
```




Severity Code:
0 - Unknown
1 - Prop Damage
2 - Injury
2b- Serious Injury
3 - Fatality

This is going to be a **multiclass classification problem**. This code will be change to int data type and replace code of 2b, 3 to become 3, and 4 respectively

New Severity Code 0 - Minor Prop Damage
1 - Prop Damage
2 - Injury
3- Serious Injury
4 - Fatality

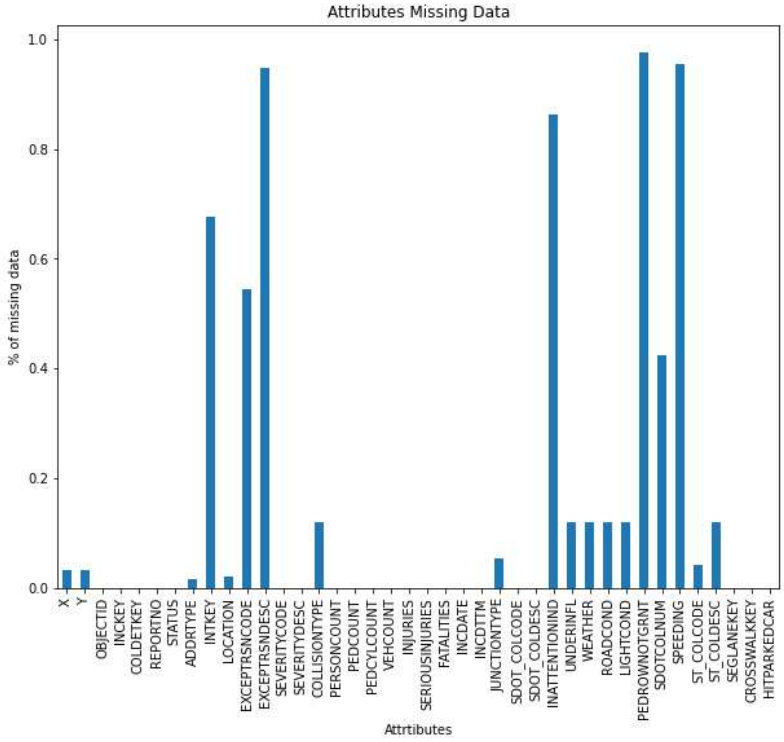
Also, the dataset negatively skewed (unbalanced).

---

### Display Attributes Missing Data

[15]:
```
absent_data = df.isnull().sum(axis=0)/df.shape[0]
```

[19]:
```
plt.figure(figsize=(10,8))
absent_data.plot(kind='bar')
plt.ylabel("% of missing data")
plt.xlabel("Attrtibutes")
plt.title("Attributes Missing Data");
```




Not all attributes is going to be used as independent variables. Non relevant attributes such as **OBJECTID** will be removed as well attributes with high percentage of missing data such as **SPEEDING**.

---

## 3. Data Preparation