# Car Accident Severity Analysis

(Applied Data Science Capstone)

The project aims to study factors which play a role in severity of accidents using Machine Learning Models

# 1. Introduction

## 1.1 Background

Seattle, also known as the Emerald city, is Washington State's largest city, with home to a large tech industry with Microsoft and Amazon headquartered in its metropolitan area. As of 2020, it has a total metro area population of 3.4 million (www.macrotrends.net). The total number of personal vehicles in Seattle in the year 2016 hit a new high of nearly 444,000 vehicles. In one South Lake Union census tract, the car population has more than doubled since 2010 (www.seattletimes.com). The increase in car ownership rates can lead to higher numbers of accidents on the road because of a simple probability. Worldwide, approximately 1.35 million people die in road crashes each year, on average 3,700 people lose their lives every day on the roads and an additional 20-50 million suffer non-fatal injuries, often resulting in long-term disabilities.

## 1.2 Problem

The world suffers due to car accidents, including the USA. National Highway Traffic Safety Administration of the USA suggests that the economical and societal harm from car accidents can cost up to $871 billion in a single year. According to 2017 WSDOT data, a car accident occurs every 4 minutes and a person dies due to a car crash every 20 hours in the state of Washington while Fatal crashes went from 508 in 2016 to 525 in 2017, resulting in the death of 555 people. The project aims to predict how severity of accidents can be reduced based on a few factors.

## 1.3 Stakeholders

The reduction in severity of accidents can be beneficial to the Public Development Authority of Seattle which works towards improving those road factors and the car drivers themselves who may take precaution to reduce the severity of accidents.

# 2.0 Understanding Data

## 2.1 Data Analysis

Dataset provided has 194673 responses and 38 variables. Below is the basic information about data, and each variable data type. Second figure shows variable with blanks in them, many of them have large no of blanks.

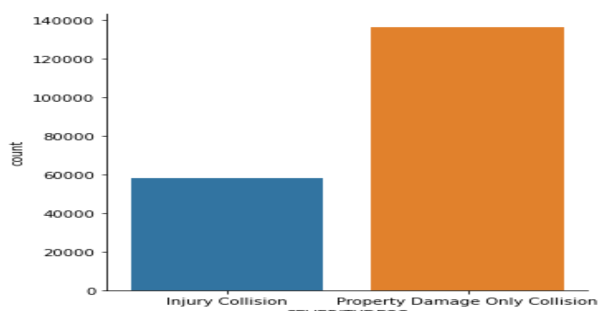
```
 #   Column          Non-Null Count   Dtype
---  ------          --------------   -----
 0   SEVERITYCODE    194673 non-null  int64
 1   X               189339 non-null  float64
 2   Y               189339 non-null  float64
 3   OBJECTID        194673 non-null  int64
 4   INCKEY          194673 non-null  int64
 5   COLDETKEY       194673 non-null  int64
 6   REPORTNO        194673 non-null  object
 7   STATUS          194673 non-null  object
 8   ADDRTYPE        192747 non-null  object
 9   INTKEY          65070 non-null   float64
 10  LOCATION        191996 non-null  object
 11  EXCEPTRSNCODE   84811 non-null   object
 12  EXCEPTRSNDESC   5638 non-null    object
 13  SEVERITYCODE.1  194673 non-null  int64
 14  SEVERITYDESC    194673 non-null  object
 15  COLLISIONTYPE   189769 non-null  object
 16  PERSONCOUNT     194673 non-null  int64
 17  PEDCOUNT        194673 non-null  int64
 18  PEDCYLCOUNT     194673 non-null  int64
 19  VEHCOUNT        194673 non-null  int64
 20  INCDATE         194673 non-null  object
 21  INCDTTM         194673 non-null  object
 22  JUNCTIONTYPE    188344 non-null  object
 23  SDOT_COLCODE    194673 non-null  int64
 24  SDOT_COLDESC    194673 non-null  object
 25  INATTENTIONIND  29805 non-null   object
 26  UNDERINFL       189789 non-null  object
 27  WEATHER         189592 non-null  object
 28  ROADCOND        189661 non-null  object
 29  LIGHTCOND       189503 non-null  object
 30  PEDROWNOTGRNT   4667 non-null    object
 31  SDOTCOLNUM      114936 non-null  float64
 32  SPEEDING        9333 non-null    object
 33  ST_COLCODE      194655 non-null  object
 34  ST_COLDESC      189769 non-null  object
 35  SEGLANEKEY      194673 non-null  int64
 36  CROSSWALKKEY    194673 non-null  int64
 37  HITPARKEDCAR    194673 non-null  object
dtypes: float64(4), int64(12), object(22)
memory usage: 56.4+ MB
```

```
X                   5334
Y                   5334
ADDRTYPE            1926
INTKEY            129603
LOCATION            2677
EXCEPTRSNCODE     109862
EXCEPTRSNDESC     189035
COLLISIONTYPE       4904
JUNCTIONTYPE        6329
INATTENTIONIND    164868
UNDERINFL           4884
WEATHER             5081
ROADCOND            5012
LIGHTCOND           5170
PEDROWNOTGRNT     190006
SDOTCOLNUM         79737
SPEEDING          185340
ST_COLCODE            18
ST_COLDESC          4904
dtype: int64
```

We need to predict severity of accident, using ‘SEVERITYCODE’. However, dataset is unbalanced. We could use technique like SMOT to balance the data.

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

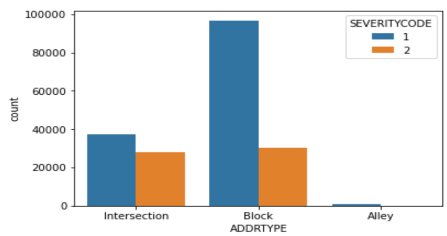
```
1    136485
2     58188
Name: SEVERITYCODE, dtype: int64
```




- ADDRTYPE – Data shows Block has most collisions. Intersection has very high Injury Collision with respect to number of collisions occur in Intersection.

```python
#Intersection more percentage of INJURY accident, block has very number of property damage accidents
sns.countplot(x='ADDRTYPE',hue='SEVERITYCODE',data=accident_df)
pd.pivot_table(accident_df,index=['SEVERITYDESC'],columns=['ADDRTYPE'],values='SEVERITYCODE',aggfunc='count')
```

| ADDRTYPE | Alley | Block | Intersection |
|---|---|---|---|
| SEVERITYDESC | | | |
| Injury Collision | 82 | 30096 | 27819 |
| Property Damage Only Collision | 669 | 96830 | 37251 |

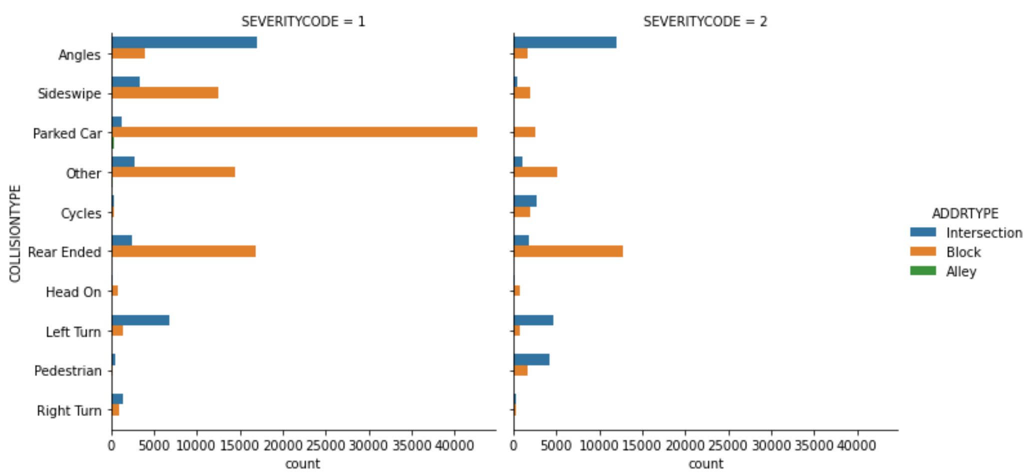



- COLLISIONTYPE – Parked car face most collisions, which mostly occur in Block. Rear ended have high number of injury collision. Cyclist and Pedestrian have injury collisions

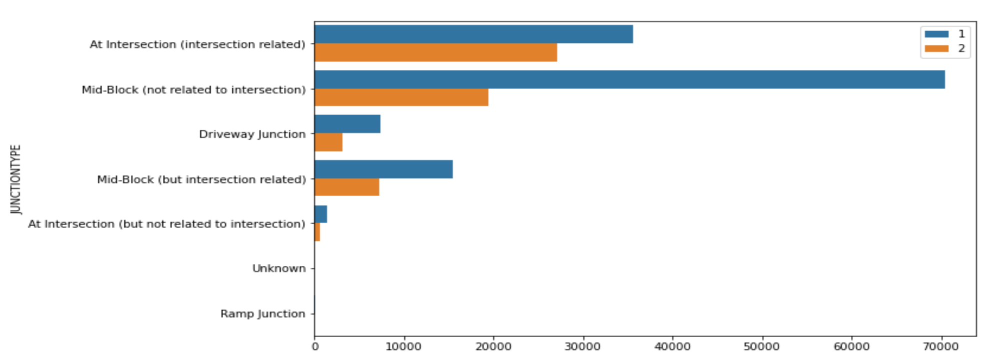
```python
#Cyclist and Pedestrian have injury collisions
#Parked Car collosion with property damage mostly occur in block
#Rear ended have high number of injury collision
plt.figure(figsize=(10,10))
sns.catplot(y='COLLISIONTYPE',col='SEVERITYCODE',hue='ADDRTYPE',data=accident_df,kind='count',aspect=1)
```

```
<seaborn.axisgrid.FacetGrid at 0x1d85cb99be0>

<Figure size 720x720 with 0 Axes>
```




- JUNCTIONTYPE – "Mid-Block (not related to intersection)" and "At Intersection (intersection related)" have high number of collisions.



- SDOT_COLDESC – PEDACYCLIST/PEDESTRIAN face mostly Injury Collision. Trying group as category list is large.

| SEVERITYDESC | Injury Collision | Property Damage Only Collision |
|---|---|---|
| **SDOT_COLCODE** | | |
| **0** | 708 | 9079 |
| **1** | 43079 | 115253 |
| **2** | 8836 | 1108 |
| **3** | 3872 | 10472 |
| **4** | 21 | 251 |
| **5** | 19 | 100 |
| **6** | 1481 | 216 |
| **7** | 172 | 6 |

```
accident_df['SDOT_COLCODE_New']=accident_df['SDOT_COLCODE'].copy()
accident_df['SDOT_COLCODE'].replace([11,12,13,14,15,16],1,inplace=True)  #MOTOR VEHICLE STRUCK MOTOR VEHICLE
accident_df['SDOT_COLCODE'].replace([18,21,22,23,24],2,inplace=True)     #MOTOR VEHICLE STRUCK PEDALCYCLIST/PEDESTRIAN
accident_df['SDOT_COLCODE'].replace([25,26,27,28,29],3,inplace=True)     #MOTOR VEHICLE SELF
accident_df['SDOT_COLCODE'].replace([31,32,33,34,35,36],4,inplace=True)  #DRIVERLESS VEHICLE STRUCK MOTOR
accident_df['SDOT_COLCODE'].replace([44,46,47,48],5,inplace=True)        #DRIVERLESS VEHICLE SELF
accident_df['SDOT_COLCODE'].replace([51,52,53,54,55,56,58],6,inplace=True)  #PEDALCYCLIST STRUCK
accident_df['SDOT_COLCODE'].replace([61,64,66,68,69],7,inplace=True)      #PEDALCYCLIST SELF
accident_df['SDOT_COLCODE'].replace(0,0,inplace=True)                     #NOT APPLICABLE
```

- WEATHER – Mostly collisions occur when weather is clear.

| SEVERITYDESC | Injury Collision | Property Damage Only Collision |
|---|---|---|
| **WEATHER** | | |
| **Blowing Sand/Dirt** | 15 | 41 |
| **Clear** | 35840 | 75295 |
| **Fog/Smog/Smoke** | 187 | 382 |
| **Other** | 116 | 716 |
| **Overcast** | 8745 | 18969 |
| **Partly Cloudy** | 3 | 2 |
| **Raining** | 11176 | 21969 |
| **Severe Crosswind** | 7 | 18 |
| **Sleet/Hail/Freezing Rain** | 28 | 85 |
| **Snowing** | 171 | 736 |
| **Unknown** | 816 | 14275 |

- ROADCOND – Most collisions take place when road condition is Dry.

```
Dry               124510
Wet                47474
4                  15210
Ice                 1209
Snow/Slush          1004
Standing Water       115
Sand/Mud/Dirt         75
Oil                   64
Name: ROADCOND, dtype: int64
```

- LIGHTCOND – Most collisions take place in Daylight condition.

```
Daylight                    116137
Dark - Street Lights On      48507
4                            13708
Dusk                          5902
Dawn                          2502
Dark - No Street Lights       1537
Dark - Street Lights Off      1199
Dark - Unknown Lighting         11
Name: LIGHTCOND, dtype: int64
```

- INATTENTIONIND, PEDROWNOTGRNT, SPEEDING, UNDERINFL – Have Y and blank row. Replacing blank with N. Replacing Yes with 1 and No with 0.

```
accident_df['INATTENTIONIND'].value_counts()
```

```
0.0     164868
1.0      29805
Name: INATTENTIONIND, dtype: int64
```

```
accident_df['PEDROWNOTGRNT'].value_counts()
```

```
0.0     190006
1.0       4667
Name: PEDROWNOTGRNT, dtype: int64
```

```
accident_df['SPEEDING'].value_counts()
```

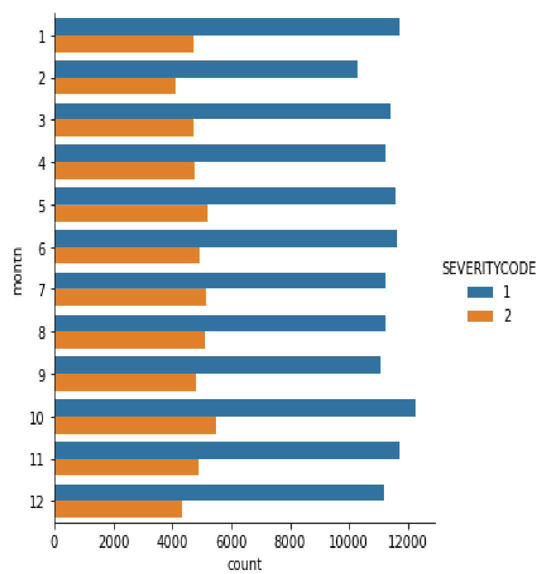
```
0.0     185340
1.0       9333
Name: SPEEDING, dtype: int64
```

```
accident_df['UNDERINFL'].value_counts()
```

```
0     185552
1       9121
Name: UNDERINFL, dtype: int64
```

- INCDTTM – Converting it to Datetime, to run additional date time analysis. Like checking Month, Year, Hour of the day, week of the day with collision to find any pattern.

| SEVERITYDESC | Injury Collision | Property Damage Only Collision |
|---|---|---|
| year | | |
| 2004 | 3647 | 8218 |
| 2005 | 4450 | 10665 |
| 2006 | 4350 | 10838 |
| 2007 | 4017 | 10439 |
| 2008 | 3767 | 9893 |
| 2009 | 3378 | 8356 |
| 2010 | 3245 | 7563 |
| 2011 | 3099 | 7820 |
| 2012 | 3467 | 7440 |
| 2013 | 3290 | 7287 |
| 2014 | 3490 | 8351 |
| 2015 | 3752 | 9243 |
| 2016 | 3714 | 7945 |
| 2017 | 3419 | 7454 |
| 2018 | 3358 | 7061 |
| 2019 | 3062 | 6350 |
| 2020 | 683 | 1562 |



## 2.2 Feature Selection

Finally selecting the below variables in the analysis along with SEVERITYCODE, which is our target variable.

| Feature Variables | Description |
|---|---|
| ADDRTYPE | Collision address type:<br>• Alley<br>• Block<br>• Intersection |
| JUNCTIONTYPE | Category of junction at which collision took place |
| INATTENTIONIND | Whether or not collision was due to inattention. (Y/N |
| UNDERINFL | Whether or not a driver involved was under the influence of drugs or alcohol. |
| WEATHER | A description of the weather conditions duringthe time of the collision. |
| ROADCOND | The condition of the road during the collision. |
| LIGHTCOND | The light conditions during the collision. |
| PEDROWNOTGRNT | Whether or not the pedestrian right of way was not granted. (Y/N) |
| SDOT_COLCODE | A description of the collision corresponding tothe collision code. |
| SPEEDING | Whether or not speeding was a factor in the collision. (Y/N) |

Post selecting the variables and then checking for blanks. Approximately 6% of data is lost in the process to remove blanks.