# Accident Severity Analysis

**Exploratory Data Analysis Project**

ADITYA MADDALI

# Introduction

The objective of this exploratory data analysis project is to evaluate the severity and frequency of accidents in the Seattle area based on multiple factors like environment and driver condition. The analysis is based on accident data from 2003 to Dec-2020 provided by Seattle Police Department and Seattle Department of Transportation (SDOT).

The dataset has 223,304 entries of accidents each with 40 attributes. SDOT also provided the metadata to understand the attributes. These attributes include various accident information such as: environmental conditions, SDOT codes, location, date and time etc.

The results will be of interest to law enforcement, emergency response services and other agencies that are interested in understanding the role of below mentioned factors on accident severity. Armed with this understanding, these agencies can perform two functions:

1. Mitigate accident severity by reducing the impact and/or probability of contributing factors
2. Provide better timely response to accidents by understanding risk prone areas.

The factors that will the severity and frequency of accidents can be broadly classified into two categories:

1. External Factors

External Factors are environmental conditions in which the accident takes place. These are: location, light conditions, road conditions, weather conditions etc.

2. Internal Factors

Internal Factors are a sum total of the state of the driver when the accident took place. That is, if the driver was attentive or if the driver was under the influence of substances etc.

# Exploratory Data Analysis

1. First step in data analysis to understand the data overview and statistics:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 223304 entries, 0 to 223303
Data columns (total 40 columns):
 #   Column           Non-Null Count   Dtype
---  ------           --------------   -----
 0   X                215802 non-null  float64
 1   Y                215802 non-null  float64
 2   OBJECTID         223304 non-null  int64
 3   INCKEY           223304 non-null  int64
 4   COLDETKEY        223304 non-null  int64
 5   REPORTNO         223304 non-null  object
 6   STATUS           223304 non-null  object
 7   ADDRTYPE         219570 non-null  object
 8   INTKEY           72598 non-null   float64
 9   LOCATION         218689 non-null  object
 10  EXCEPTRSNCODE    102901 non-null  object
 11  EXCEPTRSNDESC    11865 non-null   object
 12  SEVERITYCODE     223303 non-null  object
 13  SEVERITYDESC     223304 non-null  object
 14  COLLISIONTYPE    196371 non-null  object
 15  PERSONCOUNT      223304 non-null  int64
 16  PEDCOUNT         223304 non-null  int64
 17  PEDCYLCOUNT      223304 non-null  int64
 18  VEHCOUNT         223304 non-null  int64
 19  INJURIES         223304 non-null  int64
 20  SERIOUSINJURIES  223304 non-null  int64
 21  FATALITIES       223304 non-null  int64
 22  INCDATE          223304 non-null  object
 23  INCDTTM          223304 non-null  object
 24  JUNCTIONTYPE     211276 non-null  object
 25  SDOT_COLCODE     223303 non-null  float64
 26  SDOT_COLDESC     223303 non-null  object
 27  INATTENTIONIND   30195 non-null   object
 28  UNDERINFL        196391 non-null  object
 29  WEATHER          196180 non-null  object
 30  ROADCOND         196260 non-null  object
 31  LIGHTCOND        196089 non-null  object
 32  PEDROWNOTGRNT    5222 non-null    object
 33  SDOTCOLNUM       127205 non-null  float64
 34  SPEEDING         10004 non-null   object
 35  ST_COLCODE       213891 non-null  object
 36  ST_COLDESC       196371 non-null  object
 37  SEGLANEKEY       223304 non-null  int64
 38  CROSSWALKKEY     223304 non-null  int64
 39  HITPARKEDCAR     223304 non-null  object
dtypes: float64(5), int64(12), object(23)
memory usage: 68.1+ MB
```

The original dataset has many columns that are administrative in nature and are therefore irrelevant to the analysis.

2. Second step was to extract relevant data.
   Columns of interest (shown below) was extracted into a new data frame.
   'X', 'Y', 'INCKEY', 'SEVERITYDESC', 'COLLISIONTYPE', 'INJURIES', 'SERIOUSINJURIES', 'FATALITIES', 'UNDERINFL', 'WEATHER', 'ROADCOND', 'LIGHTCOND'
   Some columns of interest had to dropped due to poor quality of data e.g.: SPEEDING, PEDROWNOTGRNT

```
print(accdf.shape)
print(accdf.info())
accdf.head()
```

```
(223304, 12)
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 223304 entries, 0 to 223303
Data columns (total 12 columns):
 #   Column           Non-Null Count   Dtype
---  ------           --------------   -----
 0   X                215802 non-null  float64
 1   Y                215802 non-null  float64
 2   INCKEY           223304 non-null  int64
 3   SEVERITYDESC     223304 non-null  object
 4   COLLISIONTYPE    196371 non-null  object
 5   INJURIES         223304 non-null  int64
 6   SERIOUSINJURIES  223304 non-null  int64
 7   FATALITIES       223304 non-null  int64
 8   UNDERINFL        196391 non-null  object
 9   WEATHER          196180 non-null  object
 10  ROADCOND         196260 non-null  object
 11  LIGHTCOND        196089 non-null  object
dtypes: float64(2), int64(4), object(6)
memory usage: 20.4+ MB
None
```

This is a manageable dataset and includes only columns of interest.

3. Third step was to analyze the quality of data present in this dataset by column.

```
Property Damage Only    138663
Injury                   59233
Unknown                  21910
Serious Injury            3140
Fatality                   358
Name: SEVERITYDESC, dtype: int64
Parked Car    48728
Angles        35837
Rear Ended    34888
Other         24770
Sideswipe     18995
Left Turn     14204
Pedestrian     7717
Cycles         5982
Right Turn     3038
Head On        2212
Name: COLLISIONTYPE, dtype: int64
```

```
N    105099
0     81663
Y      5399
1      4230
Name: UNDERINFL, dtype: int64
Clear                       115531
Raining                      34153
Overcast                     28728
Unknown                      15131
Snowing                        919
Other                          877
Fog/Smog/Smoke                 632
Sleet/Hail/Freezing Rain       116
Blowing Sand/Dirt               56
Severe Crosswind                26
Partly Cloudy                   10
Blowing Snow                     1
Name: WEATHER, dtype: int64
```

```
Dry                 129527
Wet                  48930
Unknown              15160
Ice                   1233
Snow/Slush            1014
Other                  136
Standing Water         119
Sand/Mud/Dirt           77
Oil                     64
Name: ROADCOND, dtype: int64
```

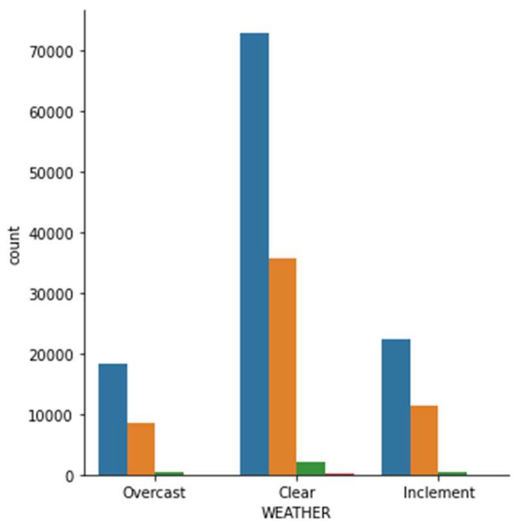
```
Daylight                    120260
Dark - Street Lights On      50442
Unknown                      13547
Dusk                          6119
Dawn                          2620
Dark - No Street Lights       1583
Dark - Street Lights Off      1240
Other                          245
Dark - Unknown Lighting         33
Name: LIGHTCOND, dtype: int64
```

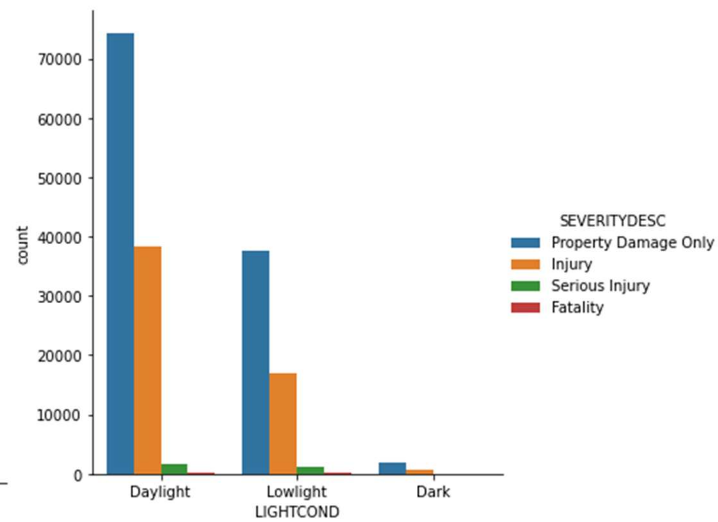
From the .value_counts() method it can be seen that the data is varied and needs to be cleaned

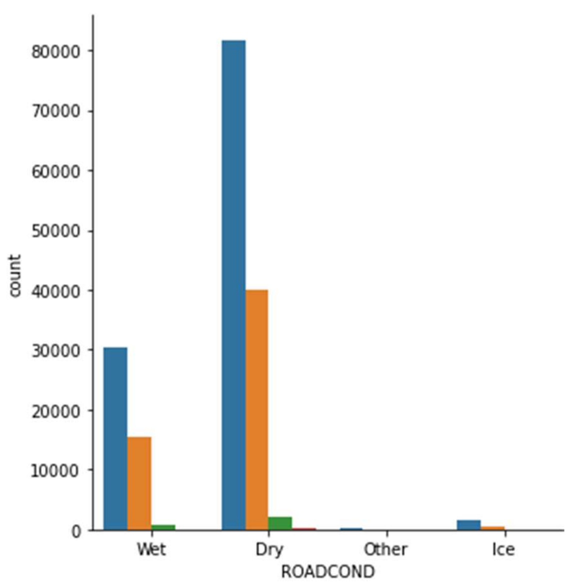
4. The data in each of these columns was then cleaned up and standardized to help with further analysis. Then it was filtered to remove blanks and Unknown values which will not help the analysis. Some of the values of categorical variables have been changed to reduce variations. This reduced the dataset to 172,841 entries.
5. The data was then visualized to gain quick insights.

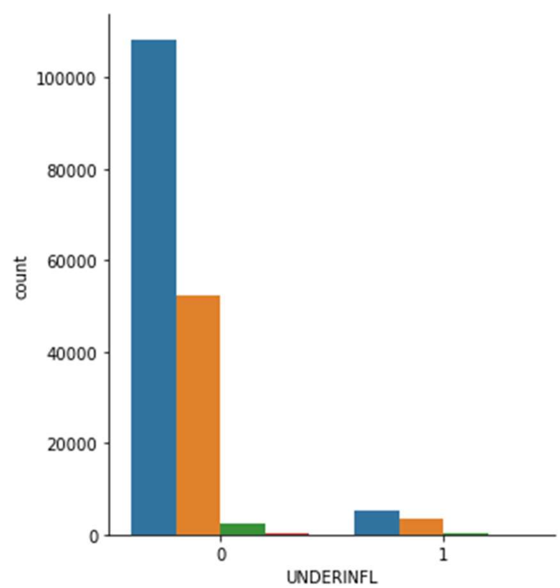
# Results

These four plots provide a quick insight into the role internal and external environmental factors play in the frequency and severity of accident occurrence.




The results seem very counter-intuitive. The only possible explanation is that the accident frequency and severity seem to be reduced when driving conditions are bad probably because fewer people venture out onto the roads. If this data was normalized for total number of vehicles on the road, we will get a better picture of how these factors affect accidents.

# Hypothesis Testing

1. Weather: It is intuitive to think that more accidents occur during inclement weather conditions.
   Null Hypothesis: More accidents occur when weather conditions are bad.
2. Light: It is intuitive to think that more accidents occur when the lighting is poor.
   Null Hypothesis: More accidents occur in bad lighting.
3. Road: It is intuitive to think that more accidents occur when the road conditions are unfavorable.
   Null Hypothesis: More accidents occur in bad road conditions.
4. DUI: It is intuitive to think that more accidents occur when the driver is intoxicated.
   Null Hypothesis: More accidents happen when the driver is intoxicated.

   From the data below and the graphs it is obvious that we can reject all null hypotheses (p-value of 5%) stated above as the frequency and severity of accidents is far greater when the driving conditions are conducive.

```
0.0    186762
1.0      9629
Name: UNDERINFL, dtype: int64
Clear        115531
Inclement     36780
Overcast      28738
Unknown       15131
Name: WEATHER, dtype: int64
Dry        129527
Wet         49049
Unknown     15160
Ice          2247
Other         277
Name: ROADCOND, dtype: int64

Daylight    120260
Lowlight     59181
Unknown      13825
Dark          2823
Name: LIGHTCOND, dtype: int64
```

# Conclusion

Quality of the data can be definitely improved by ensuring that the blank fields are filled and that they are filled consistently. The fields that have many missing or unknown values are:

1. Under the Influence
2. Inattention
3. Pedestrian right of way not granted
4. Speeding

These attributes can help with more analyses.

Also, adding more information about traffic or the total number of cars on the road will aid in data normalization that will give better insights into the effect of poor driving conditions on accidents.

Further data analysis can be done by following these steps:

1. Create a metric: Accident Severity Indicator that encapsulates the total damage due to the accident. This metric will include the severity of the accident as defined by SDOT and the number of people involved. Once established, the data will be analyzed for factors that will affect this metric.
2. For the factors that are categorical variables, one hot encoding can be performed to see if there are any patterns.