Introduction

The goal is to try to reduce both the number and severity of car collisions in Seattle. We are given a dataset and try to both qualitatively and quantitatively highlight the drivers of number and severity of car collisions as to help drivers avoid catastrophic situations.

This dataset can be used by a wide array of constituents. Individual citizens who are trying to be careful, public planning officials, and first responders.

Data

We start with the CSV file provided. The dependent variable is accident severity ranked 0 to 5 with 5 being the most severe. We will use all other columns as independent variables to predict 1) accidents and 2) severity of accidents.

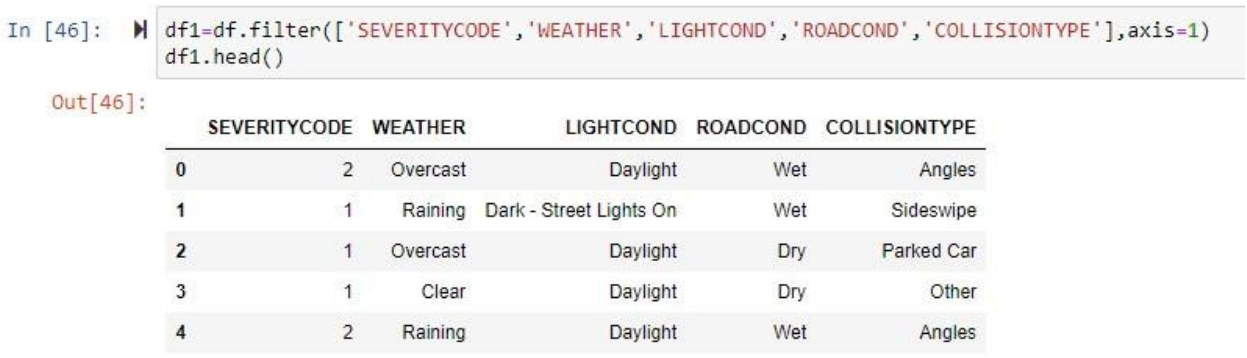
The table contains far too many columns that could theoretically be independent variables. I picked 4 independent variables to analyze that seemed the most relevant:

a) Road conditions
b) Light conditions
c) Weather
d) Collison type

Methodology

I primarily used Pandas to convert the CSV file to a dataframe. I also used SKLearn to balance the severity outcomes as it was an unbalanced dataset leaning to less severe accidents.

First, I narrowed the table down to a smaller dataframe to evaluate only 4 independent variables.



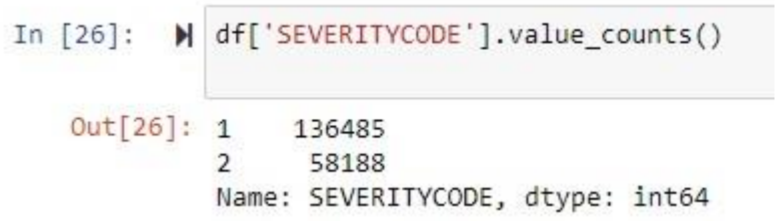
In [46]:

```
df1=df.filter(['SEVERITYCODE','WEATHER','LIGHTCOND','ROADCOND','COLLISIONTYPE'],axis=1)
df1.head()
```

Out[46]:

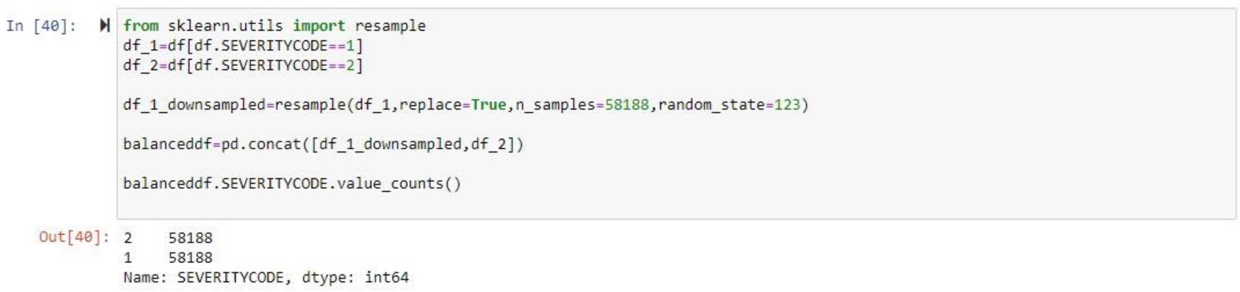
	SEVERITYCODE	WEATHER	LIGHTCOND	ROADCOND	COLLISIONTYPE
0	2	Overcast	Daylight	Wet	Angles
1	1	Raining	Dark - Street Lights On	Wet	Sideswipe
2	1	Overcast	Daylight	Dry	Parked Car
3	1	Clear	Daylight	Dry	Other
4	2	Raining	Daylight	Wet	Angles

Then I started analyzing the dependent variable – accident severity with value counts and found the accidents were skewed to less severe (1) at 70%.



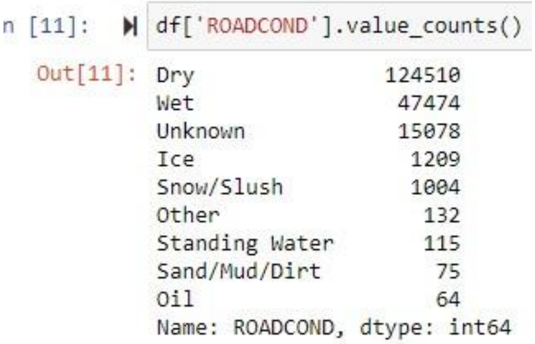
```
In [26]:  df['SEVERITYCODE'].value_counts()

Out[26]: 1    136485
         2     58188
         Name: SEVERITYCODE, dtype: int64
```

For further analysis (beyond the scope of this report), I downsampled the 1 severity to match the 2 cases.

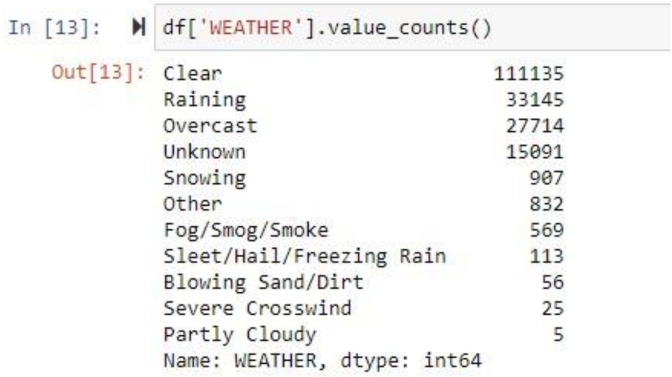


```
In [40]:  from sklearn.utils import resample
          df_1=df[df.SEVERITYCODE==1]
          df_2=df[df.SEVERITYCODE==2]

          df_1_downsampled=resample(df_1,replace=True,n_samples=58188,random_state=123)

          balanceddf=pd.concat([df_1_downsampled,df_2])

          balanceddf.SEVERITYCODE.value_counts()

Out[40]:  2    58188
          1    58188
          Name: SEVERITYCODE, dtype: int64
```

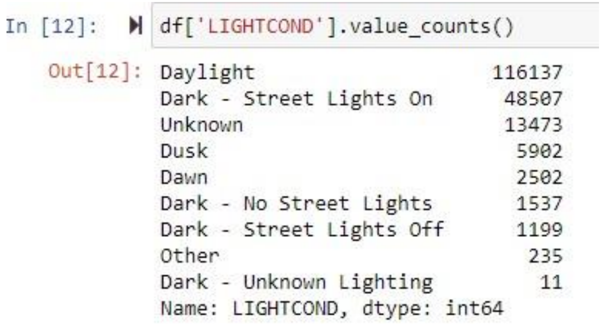
Then, I ran the value counts for the 4 independent variables:



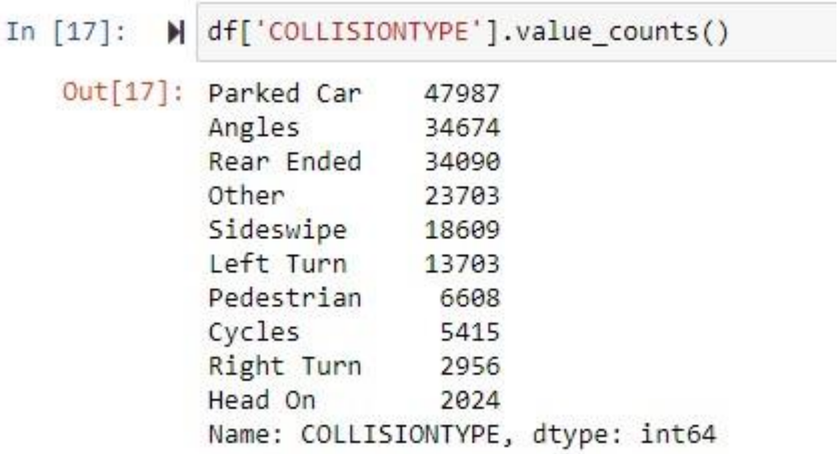
```
n [11]:   df['ROADCOND'].value_counts()

Out[11]:  Dry               124510
          Wet                47474
          Unknown            15078
          Ice                 1209
          Snow/Slush          1004
          Other                132
          Standing Water       115
          Sand/Mud/Dirt         75
          Oil                   64
          Name: ROADCOND, dtype: int64
```



```
In [13]:  df['WEATHER'].value_counts()

Out[13]: Clear                       111135
         Raining                      33145
         Overcast                     27714
         Unknown                      15091
         Snowing                        907
         Other                          832
         Fog/Smog/Smoke                 569
         Sleet/Hail/Freezing Rain       113
         Blowing Sand/Dirt               56
         Severe Crosswind                25
         Partly Cloudy                    5
         Name: WEATHER, dtype: int64
```



```
In [12]:  df['LIGHTCOND'].value_counts()

Out[12]: Daylight                    116137
         Dark - Street Lights On      48507
         Unknown                      13473
         Dusk                          5902
         Dawn                          2502
         Dark - No Street Lights       1537
         Dark - Street Lights Off      1199
         Other                          235
         Dark - Unknown Lighting         11
         Name: LIGHTCOND, dtype: int64
```



```
In [17]:  df['COLLISIONTYPE'].value_counts()

Out[17]: Parked Car     47987
         Angles         34674
         Rear Ended     34090
         Other          23703
         Sideswipe      18609
         Left Turn      13703
         Pedestrian      6608
         Cycles          5415
         Right Turn      2956
         Head On         2024
         Name: COLLISIONTYPE, dtype: int64
```

Results

The 1st 3 independent variable outcomes did not pass the sense test as with most accidents, conditions were:

- Dry
- Clear
- Daylight

Discussion

However, the 4th variable collision type was the telling variable: it showed the most common collisions were with parked cars.

> **Accidents with parked cars are minor and are usually out of carelessness rather than a major contributing factor. This explains why the severity was skewed to less severe.**

Conclusions

The dataset we were provided involves non-severe minor accidents that were likely out of carelessness or inexperience than some external condition.

For Seattle, they should run more data with severity skewed towards the more fatal accidents to see the contributing factors in order to assist city planning and promote driver awareness.

Submitted By: Abhiraam Khanna