Description of Problem and Data

Description of Problem

- 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 situatins.

Description of Data

- The issue we are trying to solve is taking the dependent variables
- A) car accidents in Seattle and
- B) the severity (severity code: 0-5 although 1&2 are only in this dataset) of such accidents with a higher number indicating a more serious accident

- Numerous categorical vehicles are provided that seem to be relevant independent variables.
- A) road conditions
- B) light conditions
- C) weather
- D) collision type

- Location is provided to target popular intersections.

Introduction

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.*

Description of Data (repeated)

- The issue we are trying to solve is taking the dependent variables
- A) car accidents in Seattle and
- B) the severity (severity code: 0-5 although 1&2 are only in this dataset) of such accidents with a higher number indicating a more serious accident

- Numerous categorical vehicles are provided that seem to be relevant independent variables.
- A) road conditions
- B) light conditions
- C) weather
- D) collision type

- Location is provided to target popular intersections.

Data and Methodology

Data

- We are provided a CSV file which needs to be cleansed. There are too many columns. I used **pandas** to load the csv file as a dataframe.

- After looking at the data, there are simply too many columns. I reduced the table to the dependent variable (accident severity) and a couple of 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

Severity Code: Dependent Variable

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

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

- Severity code is skewed to LESS negative outcomes.
- I balanced data for machine learning purposes; not needed for this exercise, using SKLEARN.

```
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
```

Independent Variables: Value Counts

- The first 3 independent variables did not equate to telling a story of being in an accident.

Accidents happened most when it was:

- Dry
- Clear
- Daylight
-

```
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 [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 [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
```

The Telling Independent Variable and Conclusions

```
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
```

- The collision type explains the dependent variable.
- The most accidents occurred with parked cars.
- **This explains why the majority of the data involved *less* severe accidents.**
- It also explains why other variables that normally contribute to accidents didn't contribute.

Accidents with parked cars are minor and are usually out of carelessness rather than a major contributing factor.