2.Data:

2.1 Data source:

Collision data had been fetched from Seattle Department of Transportation Open Data Program in CSV format.
Source:
https://s3.us.cloud-object-storage.appdomain.cloud/cf-coursesdata/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv

2.2 Data understanding:

Our dataset contains all the information about a total number of 194673 car accidents. It contains the location, time ,date ,weather conditions, road conditions etc. The dataset has 37 features and The dataset has total observations of 194673 with variation in number of observations for every feature.

The dataset also has a lot of empty columns which could have been beneficial had the data been present there. These columns included pedestrian granted way or not, segment lane key, crosswalk key and hit parked car.

2.3 Feature Selecting:

As we are using this dataset to predict the severity of an accident not all the features will

be useful for us. So, I will only pick up the features that will help me in this project.

The features that I selected are:

1.SEVERITYCODE:

a code that corresponds to the severity of collision and this column will be the main one as we are predicting the severity.

It has two values:

- 1:prop damage
- 2:injury

2.LIGHTCOND:

The conditions of the light during the collision.

Its values are:

(daylight / dark- street lights on / dark - no street lights / dark- street lights off /dark - unknown lighting / dusk / dawn /other / unknown)

3.WEATHER:

A description of weather conditions during the time of the collision.

Its values are:

(overcast - raining - clear - snowing - severe crosswind - partly cloudy - fog/ smog/

smoke - steel/ hail / freezing rain - blowing sand/ dirt - other - unknown)

4.ROADCOND:
The conditions of the road during the collision. Its values are:
(wet - dry- ice - snow/slush - sand/mud/dirt - standing water - oil - other - unknown)

2.4 Data cleaning:

First of all, we had to drop the null values from the dataset, which results in about a 10% decrease in the overall dataset. We were able to do this as we realise null values exist for all 3 feature columns normally for each row.

	SEVERITYCODE	WEATHER	LIGHTCOND	ROADCOND
15	1	NaN	NaN	NaN
36	1	NaN	NaN	NaN
53	2	NaN	NaN	NaN
60	1	NaN	NaN	NaN
75	1	NaN	NaN	NaN

So, Number of entries after dropping NA values will be 189337.

Secondly, we will have to down sample our dataset as we have 3 times more rows with a severity code of 1 compared to 2.

```
SEVERITYCODE
1    132285
2     57052
```

We will down sample the rows with severity code of 1 to be the same as severity code of 2 for our use case.

```
2    57052
1    57052
```

This an example of our data from the first 5 columns of our dataset:

	SEVERITYCODE	WEATHER	LIGHTCOND	ROADCOND
0	2	Overcast	Daylight	Wet
1	1	Raining	Dark - Street Lights On	Wet
2	1	Overcast	Daylight	Dry
3	1	Clear	Daylight	Dry
4	2	Raining	Daylight	Wet

The downsampled dataset:

	SEVERITYCODE	WEATHER	LIGHTCOND	ROADCOND
82770	1	Clear	Daylight	Dry
122946	1	Clear	Daylight	Dry
102968	1	Clear	Daylight	Dry
13906	1	Overcast	Daylight	Dry
123301	1	Raining	Dark - Street Lights On	Wet