Data

To solve the objective, we need following data: -

1. List of all the accidents occurred and recorded in the Seattle.
2. Severity of every recorded accident.
3. Weather condition during the travel for every recorded accident.
4. Lighting condition on the Road for every recorded accident.
5. Road condition for every recorded accident.
6. Before the accident, whether the victim was under the influence of alcohol/substance.
7. Whether it was violation of speed limit.
8. Whether the driver was attentive.

Sources of Data

The data has been made available by the IBM online certification team. The collisions data is provided by Seattle Police Department and recorded by Traffic Records which is being updated on weekly basis since 2004.

Data Preparation and Cleaning

The first step to Data Analysis and Machine Learning is acquisition of data from the reliable sources and converting it into the required format. This format should be complete in every aspect and should help the Scientist make his/her case. It should narrate a story that the target audience should listen to with interest. The case should be made on the basis of data rather than instincts. The required dataframe should not contain any unwanted values or null values which can significantly affect the final model accuracy. The null values should either be dropped from the dataframe or appropriate values should be putted to make it more sensible.

After building a dataframe, the values are checked for null and appropriate corrections are made. There were 11 types of input in the WEATHER columns which were all converted to integers for further analysis as follows –

Serial Number	WEATHER original column values	WEATHER replaced numerical values
1.	Unknown	0
2.	Clear	1
3.	Other	2
4.	Partly Cloudy	3
5.	Overcast	4
6.	Severe Crosswinds	5
7.	Blowing Sand/Dirt	6
8.	Raining	7

9.	Fog/Smog/Smoke	8
10.	Sleet/Hail/Freezing Rain	9
11.	Snowing	10

Similarly, all the string inputs in ROADCOND and LIGHTCOND were replaced with numerical inputs for the sake of ease in analysis as given below –

Serial Number	ROADCOND original column values	ROADCOND replaced numerical values
1.	0	0
2.	Unknown	0
3.	Ice	1
4.	Sand/Mud/Dirt	2
5.	Oil	3
6.	Snow/Slush	4
7.	Standing Water	5
8.	Wet	6
9.	Other	7
10.	Dry	8

Serial Number	LIGHTCOND original column values	LIGHTCOND replaced numerical values
1.	0	0
2.	Unknown	0
3.	Dark – No Street Lights	1
4.	Dark – Street Lights Off	2
5.	Dark – Unknown Lighting	3
6.	Dark – Street Lights On	4
7.	Dusk	5
8.	Dawn	6
9.	Other	7
10.	Daylight	8

In the UNDERINFL column, the entries were in both string as well as binary format. They were all replaced with binary inputs as follows –

Serial Number	UNDERINFL original column values	UNDERINFL replaced column values
1.	Y	1
2.	N	0
3.	0	0
4.	1	1

Similarly, SPEEDING and INATTENTIONIND columns were also corrected and converted since only positives i.e. 'Y's were inputted in the dataframe as follows –

Serial Number	SPEEDING original column values	SPEEDING replaced column values
1.	Y	1
2.	Null	0

Serial Number	INATTENTIONIND original column values	INATTENTIONIND replaced column values
1.	Y	1
2.	Null	0

All the rows with null values in every column were dropped.

After cleaning and formatting, from the given dataframe "df", SEVERITYCODE, WEATHER, LIGHTCOND, ROADCOND, UNDERINFL, SPEEDING and INATTENTIONIND were extracted by dropping the remaining columns. Only these columns were selected from the large given dataframe because these were the only independent variable which could be used to predict the severity of a possible accident. In this renewed dataframe "df", the values were checked for null and appropriate corrections were made. Finally, the columns were renamed as given below –

Serial Number	Original column name	Renamed column name
1.	SEVERITYCODE	severity
2.	WEATHER	weather
3.	LIGHTCOND	light
4.	ROADCOND	road
5.	UNDERINFL	influence
6.	SPEEDING	speeding
7.	INATTENTIONIND	inattention

Ultimately, the dataframe "df" consists of 7 columns plus one index column. There are 194673 entries so the shape of the dataframe "df" is (7,194673).