IBM DATA SCIENCE FINAL PROJECT

ACCIDENT SEVERITY DATA ANALYSIS

INTRODUCTION:

This is a data analysis project based around the severity of reported accidents. In this project we shall find the degree of relation between the location of the accident, light condition and road condition with regards to the severity of the accident. Depending on the result we will be able to analyse the factors that cause severe accidents and hopefully allow the respective authorities to be able to take certain measures to reduce the impact of these factors.

DATA :

The dataset we will use for this analysis will be imported into our database by a url. On first inspection we shall inspect the various labels of the dataset. We have 37 columns with various details and factors with regard to the accident. We have to take into consideration the 4 columns that we will need for this analysis.

These are as follow:

- ADDRTYPE
- LIGHTCOND
- ROADCOND
- SEVERITYCODE

On further inspection of the dataset we find unbalanced data.We shall now balance the data using different data wrangling techniques.

```
In [24]: df.head(5)
Out[24]:
```

	SEVERITYCODE	X	Y	OBJECTID	INCKEY	COLDETKEY	REPORTNO	STATUS	ADDRTYPE	INTKEY	...	ROADCOND	LIGHTCOND	PED
0	2	-122.323148	47.703140	1	1307	1307	3502005	Matched	Intersection	37475.0	...	Wet	Daylight	
1	1	-122.347294	47.647172	2	52200	52200	2607959	Matched	Block	NaN	...	Wet	Dark - Street Lights On	
2	1	-122.334540	47.607871	3	26700	26700	1482393	Matched	Block	NaN	...	Dry	Daylight	
3	1	-122.334803	47.604803	4	1144	1144	3503937	Matched	Block	NaN	...	Dry	Daylight	
4	2	-122.306426	47.545739	5	17700	17700	1807429	Matched	Intersection	34387.0	...	Wet	Daylight	

5 rows × 38 columns

TYPE TO ENTER A CAPTION.

METHODOLOGY:

The current dataset has unbalanced labels and therefore will not yield an optimal result.

The first step requires us to perform data wrangling operations on the dataset to remove unwanted and unnecessary data.

We do this by first checking the different data types and also the missing value in the data.

In [15]:
```
missing_data = df.isnull()
missing_data.head(5)
```

Out[15]:

	SEVERITYCODE	X	Y	OBJECTID	INCKEY	COLDETKEY	REPORTNO	STATUS	ADDRTYPE	INTKEY	...	ROADCONI
0	False	False	False	False	False	False	False	False	False	False	...	Fals
1	False	False	False	False	False	False	False	False	False	True	...	Fals
2	False	False	False	False	False	False	False	False	False	True	...	Fals
3	False	False	False	False	False	False	False	False	False	True	...	Fals
4	False	False	False	False	False	False	False	False	False	False	...	Fals

5 rows × 38 columns

In [16]:
```
for column in missing_data.columns.values.tolist():
    print(column)
    print (missing_data[column].value_counts())
    print("")
```

```
SEVERITYCODE
False    194673
Name: SEVERITYCODE, dtype: int64

X
False    189339
True       5334
Name: X, dtype: int64

Y
False    189339
True       5334
Name: Y, dtype: int64

OBJECTID
False    194673
Name: OBJECTID, dtype: int64

INCKEY
```

Further more we shall now remove columns with redundant data.

```
In [18]: df.drop(["SEGLANEKEY","CROSSWALKKEY","HITPARKEDCAR","INATTENTIONIND","STATUS"], axis=1)
```

Out[18]:

INCKEY	COLDETKEY	REPORTNO	ADDRTYPE	INTKEY	LOCATION	...	SDOT_COLDESC	UNDERINFL	WEATHER	ROADCO
1307	1307	3502005	Intersection	37475.0	5TH AVE NE AND NE 103RD ST	...	MOTOR VEHICLE STRUCK MOTOR VEHICLE, FRONT END ...	N	Overcast	
52200	52200	2607959	Block	NaN	AURORA BR BETWEEN RAYE ST AND BRIDGE WAY N	...	MOTOR VEHICLE STRUCK MOTOR VEHICLE, LEFT SIDE ...	0	Raining	
26700	26700	1482393	Block	NaN	4TH AVE BETWEEN SENECA ST AND UNIVERSITY ST	...	MOTOR VEHICLE STRUCK MOTOR VEHICLE, REAR END	0	Overcast	
1144	1144	3503937	Block	NaN	2ND AVE BETWEEN MARION ST AND MADISON ST	...	MOTOR VEHICLE STRUCK MOTOR VEHICLE, FRONT END ...	N	Clear	
17700	17700	1807429	Intersection	34387.0	SWIFT AVE S AND SWIFT AV OFF RP	...	MOTOR VEHICLE STRUCK MOTOR VEHICLE, FRONT END ...	0	Raining	
...	...	...	...	...	...	...	...	...	...	

On further inspection of the data it is found that the required key columns ADDRTYPE, WEATHER, ROADCOND are categorical in nature and cannot be used in different analytical methods until they are converted to integer values.

The following numerical values are assigned to the different categorical values:

WEATHER

CLEAR	1
OVERCAST	2
RAINING	3
UNKNOWN	4
OTHER	5
SNOWING	6
FOG	7
SLEET	8
DIRT	9
SEVERE CROSSWINDS	10
PARTLY CLOUDY	11

ADDRTYPE

INTERSECTION	1
BLOCK	2
ALLEY	3

ROADCOND

DRY	1
WET	2
UNKNOWN	3
SNOW	4
ICE	5
OTHER	6
SAND	7
STANDING WATER	8
OIL	9

SEVERITY CODE has categorical values an hence needs a modification.

EXCEPTRSNDESC	Text, 300	
SEVERITYCODE	Text, 100	A code that corresponds to the severity of the collision: • **3**—fatality • **2b**—serious injury • **2**—injury • **1**—prop damage • **0**—unknown

It is now important to redefine the datatypes for the columns to allow analytical methods to be used.

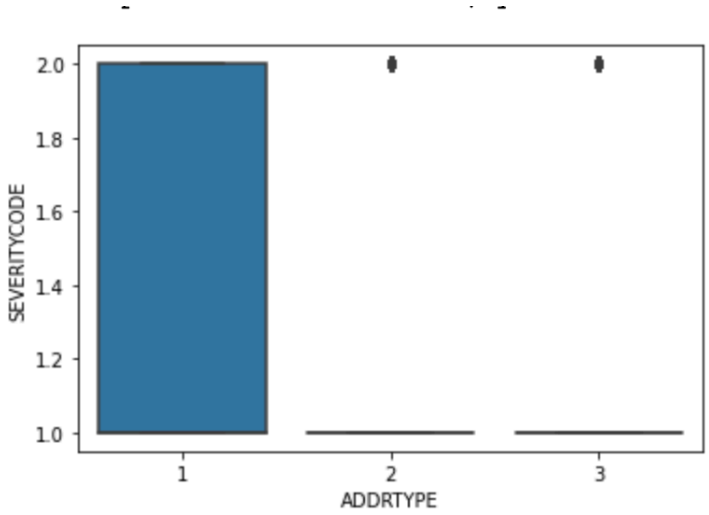
In [25]:
```
df[["SEVERITYCODE"]]=df[["SEVERITYCODE"]].astype("int")
df[["ADDRTYPE"]]=df[["ADDRTYPE"]].astype("int")
df[["WEATHER"]]=df[["WEATHER"]].astype("int")
df[["ROADCOND"]]=df[["ROADCOND"]].astype("int")
```

Now a new data frame is created based on the formatted dataset which can now be used for exploratory analysis.

Our exploratory analysis will be based on upon finding a correlation between the severity of the collision and the address type.

Pearsons correlation coefficient is used to measure the relationship.

First a boxplot is created to visualise the relationship



From the above graph we can see that the majority of the severity code 2 cases are found at address type 1(Intersection).

We shall now find the Pearson correlation Pearson Correlation
The Pearson Correlation measures the linear dependence between two variables X and Y.
The resulting coefficient is a value between -1 and 1 inclusive, where:
1: Total positive linear correlation.
0: No linear correlation, the two variables most likely do not affect each other.
-1: Total negative linear correlation.

```
The Pearson Correlation Coefficient is
-0.19976858449641707  with a P-value of P = 0.0
```

From the Pearson correlation coefficient we can deduce that the there is no implicit relationship i.e linear relationship between the address of the accident and the severity.

MACHINE LEARNING:

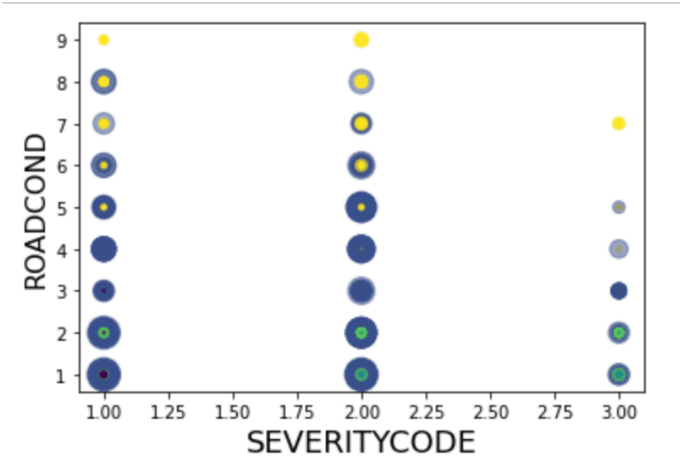
K MEANS CLUSTERING

***k*-means clustering** is a method of vector quantization, originally from signal processing, that aims to partition *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean (cluster centers or cluster centroid), serving as a prototype of the cluster

The aim is to find a clusters which can take into account

- SEVERITYCODE
- WEATHER
- ROADCOND

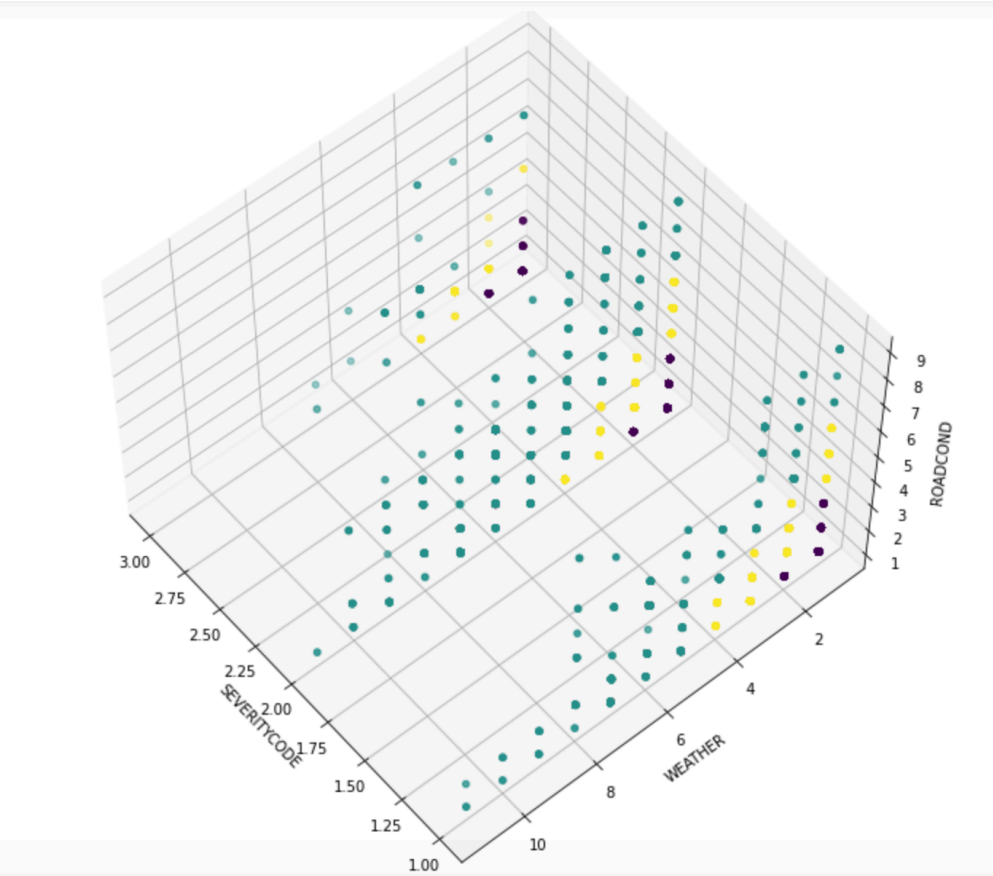
First we plot a scatterplot to view the relationship between SEVERITY OF THE COLLISION and the ROADCONDITION.



The above scatterplot shows us that most of the type 2 severity collisions take place in road conditions 1-4. These are:
1-DRY 2-WET 3-Unknown 4-SNOW

Now we apply K Means algorithm to create three clusters based around the three attributes

The following diagram depicts these clusters.



We now have 3 clusters based on the severity of the accident with respect to the road condtions and weather. From the above chart we can see that the majority of the code 2 type collisions take place with the road conditions 1-4 and weather conditions 1-7, whereas most of the type 1 collisions take place with road conditions 1-3 and weather conditions 1-6.

RESULTS:

The results of the data analysis can be summarised as follows

- Majority of the severity code 2-INJURY collisions take place at address type 1-INTERSECTIONS.
- There is no linear relationship between the severity of the collision and the address type.
- Majority of the severity code 2-INJURY collisions take place with the following road conditions: 1-DRY 2-WET 3-Unknown 4-SNOW
- Majority of the code 2 type collisions take place with the road conditions 1-4 and weather conditions 1-7.
- Most of the type 1 collisions take place with road conditions 1-3 and weather conditions 1-6.

DISCUSSION:

The analysis conducted can be useful for providing the road safety organisations an insight into the various factors that contribute to an accident or collision.

The relationship between these factors and the severity of the collision can play a vital role in preventing further such accidents or reducing the severity of the accident.

As the results mentioned the majority of the Injury causing collisions happened at Intersections. This finding can allow the road safety organisation to enforce stricter rules e.g lower speed limit so as to help reduce accidents in that type of location.

The next result showed us how the majority of injury causing collision were caused by certain road conditions. This can be used by the officials to improve road conditions which depict these characteristics.

And finally the clusters based on Severity, Weather and Road conditions can provide an insight into the relationship between these factors and lead to better decisions being made regarding road safety at locations which have these certain characteristics.

CONCLUSION:

The objective of this project is to gain valuable insights from the given dataset of accident/collisions in a particular location.

The relationships between the different attributes of the dataset can be used to make the necessary changes to ensure public safety and overall road safety is improved.

The concerned authorities can take up various initiatives to try and counter the different characteristics of the collisions and accidents that have taken place to ensure that such collisions can be reduced in the near future.