SEVERITY PREDICTION MODEL

INTRODUCTION: BUSINESS PROBLEM

Road accidents are one of the major problems leading to loss of lives and properties in many countries. They further have an impact on social and economic conditions in terms of major costs like healthcare. But accidents prediction on roadways is quite a difficult task as they are caused by wide variety of reasons like road conditions, road designs, user behaviours, traffic rules, vehicle conditions, weather conditions etc. Gathering data about each and every attribute with an extra factor of severity attached to it like property damage, injury, serious injury, fatality etc. can help us build some hypothesis and build statistical models of prediction around them.

We can test the hypothesis and build perfect prediction models or even gain new insights regarding what are the main reasons apart from already assumed reasons. The total number of accidents happening and the severity level of accidents have been and can be applied as one of the many possible indicators to measure the efficiency of road network system services.

Interest:

The resulting models can help the respective government road transport agencies to identify and rectify the accident factors by answering the question

"What are all the factors which increases the probability of severity in accidents?"

DATA

A. DATA UNDERSTANDING

The Proposed model tries to predict the severity of accidents from the dataset of all accidents reported within the Seattle-area provided by the SPD from the year 2004 to present(Data / Metadata). The dataset includes several attributes relating to accident from location & time to junction type to weather conditions to severity.

From the data, removing irrelevant and redundant attributes before feeding the data into the model is very important as it increases both efficiency and reliability of model outputs. Accordingly, we will remove attributes from the dataset mentioned above like

"OBJECT ID, SHAPE, INCKEY, COLDETKEY, INTKEY, LOCATION, EXCEPTRSNCODE, EXCEPTRSNDESC, SEVERITY DESC, INJURIES, SERIOUSINJURIES, FATALITIES, INCDATE, SDOT_COLCODE, SDOT_COLCODE, PEDROWNOTGRNT, SDOTCOLNUM, ST_COLCODE, ST_COLDESC, SEGLANEKEY, CROSSWALKKEY and STATUS" from the data.

We will choose the parameter Severity which we will use to determine the hazardous conditions leading to accidents based on the outputs provided by other independent attributes provided in the data as mentioned below:

1. Severity Code (Property damage, injury, fatality)
2. Location data (Coordinates, address type etc)
3. Collision Type (Pedestrians, Vehicles, cycles, junction etc.)
4. Persons Involved (Pedestrians, Cyclists etc)
5. Vehicle Count

6. Date and Time
7. Junction Type
8. Inattention, Under Influence and Speeding
9. Weather and Road Conditions

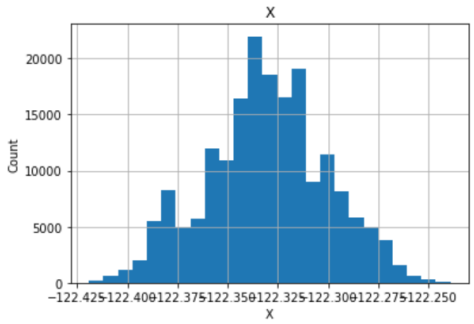
Eliminating irrelevant attributes and creating a new dataframe with relevant attributes:

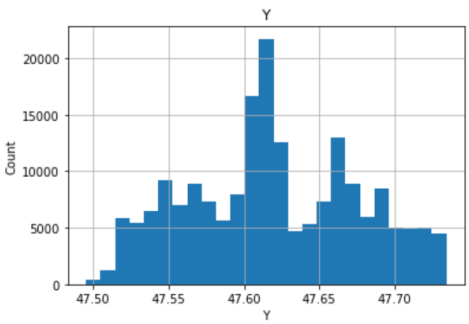
ADDRTYPE	COLLISIONTYPE	PERSONCOUNT	PEDCOUNT	PEDCYLCOUNT	VEHCOUNT	INCDTTM	JUNCTIONTYPE	INATTENTIONIND	UNDERINFL	WEATHER	ROADCOND	LIGHTCOND
Intersection	Angles	2	0	0	2	3/27/2013 2:54:00 PM	At Intersection (intersection related)	NaN	N	Overcast	Wet	Daylight
Block	Sideswipe	2	0	0	2	12/20/2006 6:55:00 PM	Mid-Block (not related to intersection)	NaN	0	Raining	Wet	Dark - Street Lights On
Block	Parked Car	4	0	0	3	11/18/2004 10:20:00 AM	Mid-Block (not related to intersection)	NaN	0	Overcast	Dry	Daylight
Block	Other	3	0	0	3	3/29/2013 9:26:00 AM	Mid-Block (not related to intersection)	NaN	N	Clear	Dry	Daylight
Intersection	Angles	2	0	0	2	1/28/2004 8:04:00 AM	At Intersection (intersection related)	NaN	0	Raining	Wet	Daylight

B. DATA PREPARATION

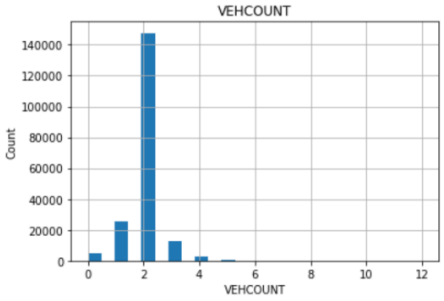
From the variables chosen, we will segregate both numerical and categorical variables to see patterns and count of data relevant to an accident in both buckets. After analysing the patterns individually for each attribute, we will select attributes which might not be needed as inputs as they might skew the output or will have no influence on the output.

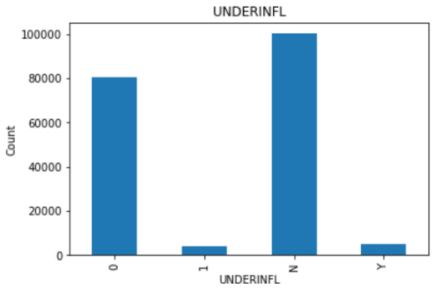
Chosen Attributes individual Evaluation:

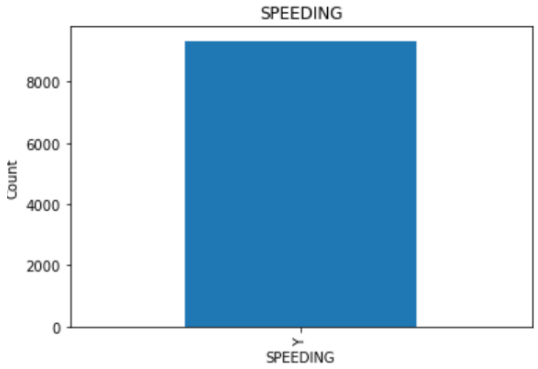


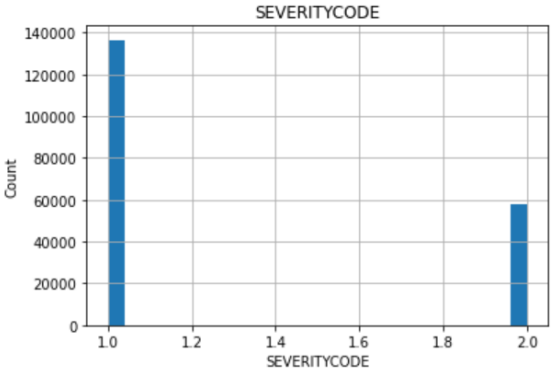


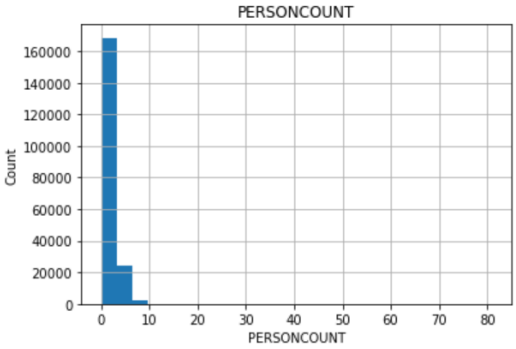
Location data provided Latitude, and Longitudinal coordinates denoted as X and Y, respectively, as shown in the above graphs. Above graphs also display a normal distribution providing insight that accidents happen more at the centre of and the frequency and number decrease as we move away from the city. We can also safely assume this due to more traffic and population density distribution at the centre than outskirts. So, it is safe to delete these two attributes as inputs.

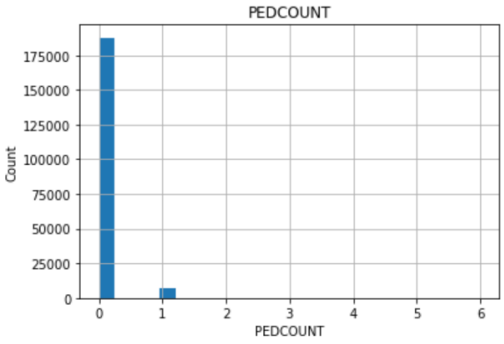


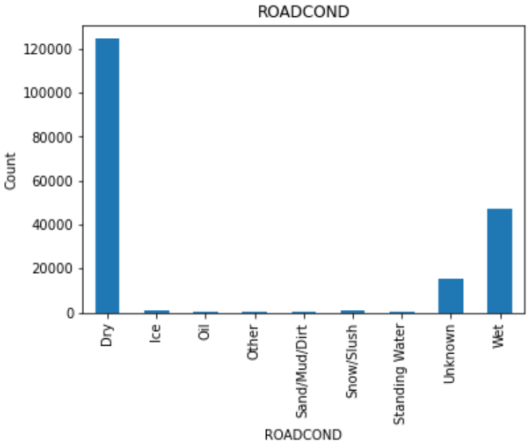


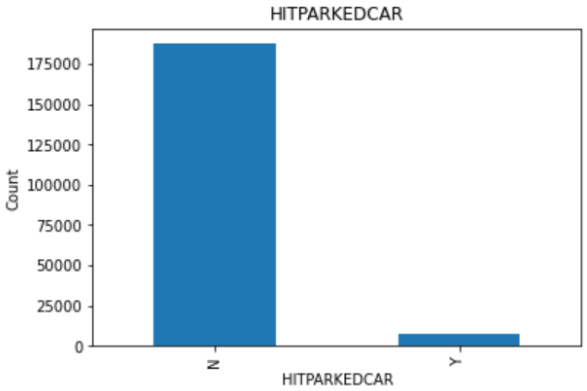

SPEEDING
Count
SPEEDING
Y

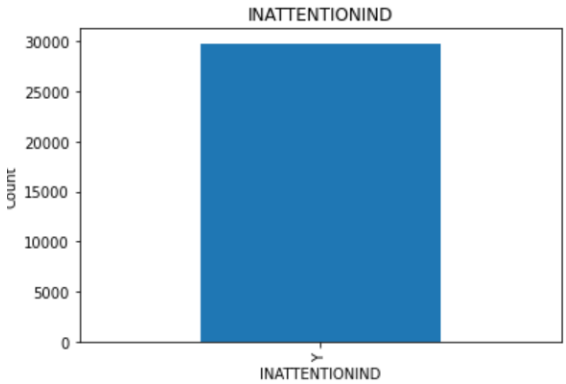

SEVERITYCODE
Count
SEVERITYCODE

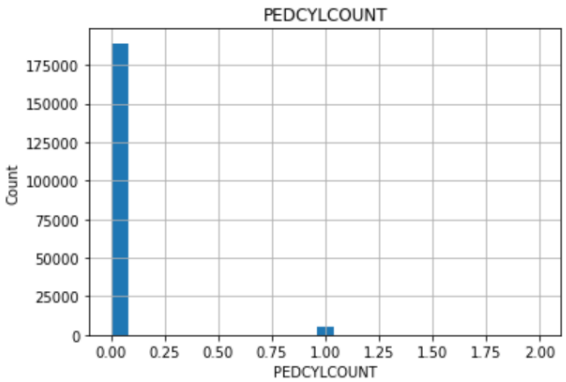

PERSONCOUNT
Count
PERSONCOUNT

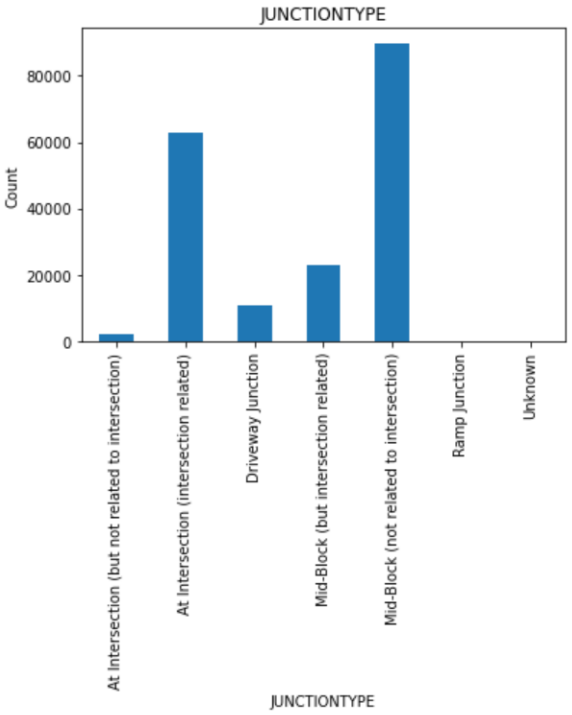

PEDCOUNT
Count
PEDCOUNT

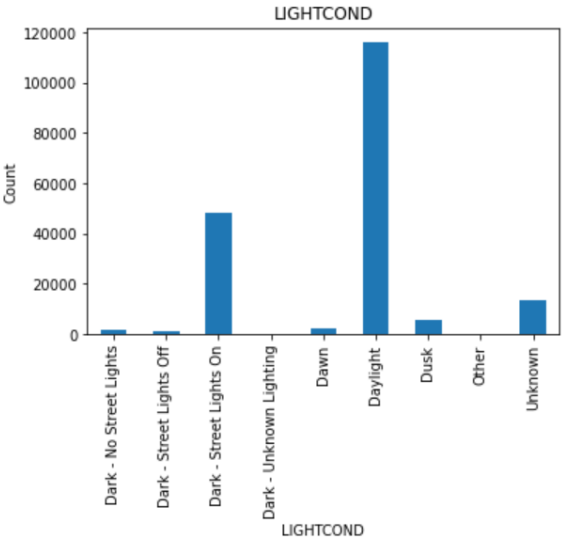

ROADCOND
Count
Dry
Ice
Oil
Other
Sand/Mud/Dirt
Snow/Slush
Standing Water
Unknown
Wet
ROADCOND

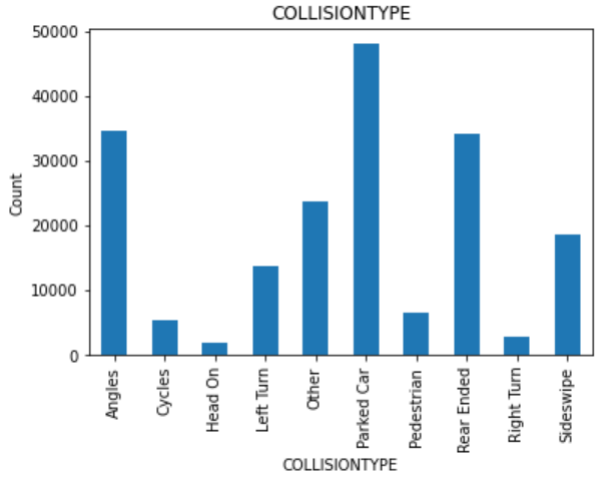

HITPARKEDCAR
Count
N
Y
HITPARKEDCAR

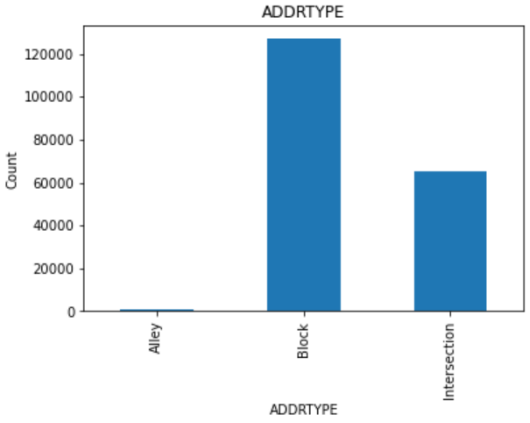

INATTENTIONIND
Count
Y
INATTENTIONIND

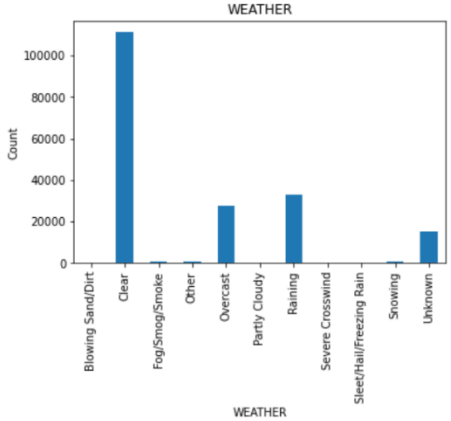

PEDCYLCOUNT
Count
PEDCYLCOUNT


JUNCTIONTYPE
Count
0
20000
40000
60000
80000
At Intersection (but not related to intersection)
At Intersection (intersection related)
Driveway Junction
Mid-Block (but intersection related)
Mid-Block (not related to intersection)
Ramp Junction
Unknown
JUNCTIONTYPE


LIGHTCOND
Count
0
20000
40000
60000
80000
100000
120000
Dark - No Street Lights
Dark - Street Lights Off
Dark - Street Lights On
Dark - Unknown Lighting
Dawn
Daylight
Dusk
Other
Unknown
LIGHTCOND


COLLISIONTYPE
Count
0
10000
20000
30000
40000
50000
Angles
Cycles
Head On
Left Turn
Other
Parked Car
Pedestrian
Rear Ended
Right Turn
Sideswipe
COLLISIONTYPE

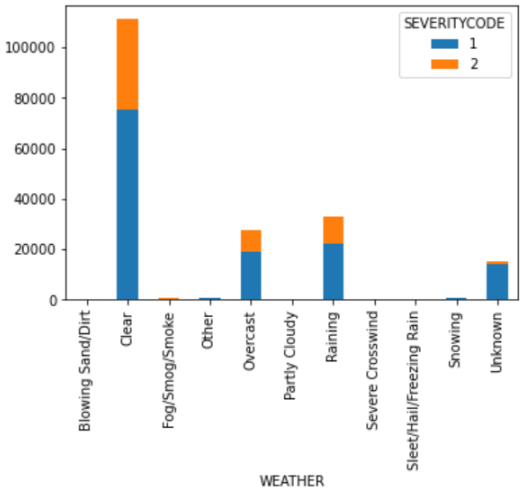

ADDRTYPE
Count
0
20000
40000
60000
80000
100000
120000
Alley
Block
Intersection
ADDRTYPE

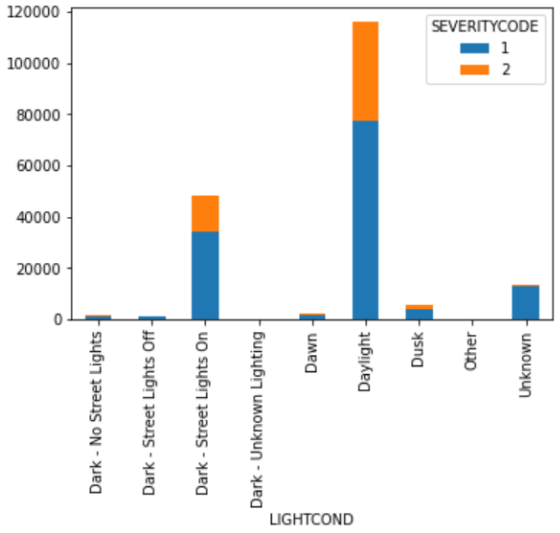

WEATHER
Count
0
20000
40000
60000
80000
100000
Blowing Sand/Dirt
Clear
Fog/Smog/Smoke
Other
Overcast
Partly Cloudy
Raining
Severe Crosswind
Sleet/Hail/Freezing Rain
Snowing
Unknown
WEATHER

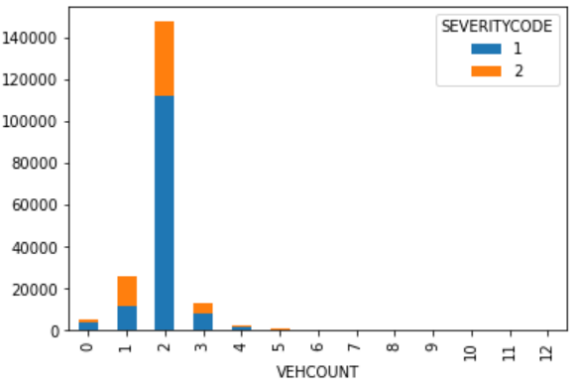
Observations and deletions: From the above graphs, it is clear that Pedestrian Count and Pedestrian Cycle Count has zero values in more than 90% of the entries, hence deleting those attributes from the inputs list of the model.

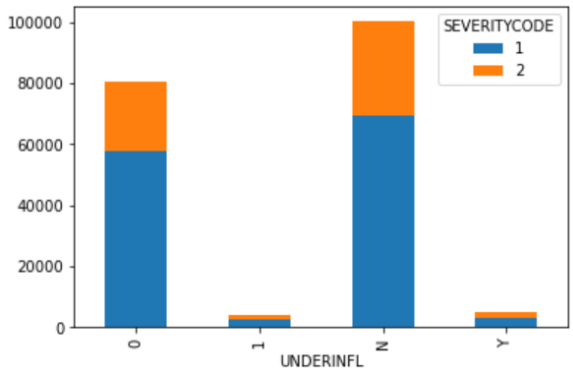
Independent Variables vs Dependent Variables:

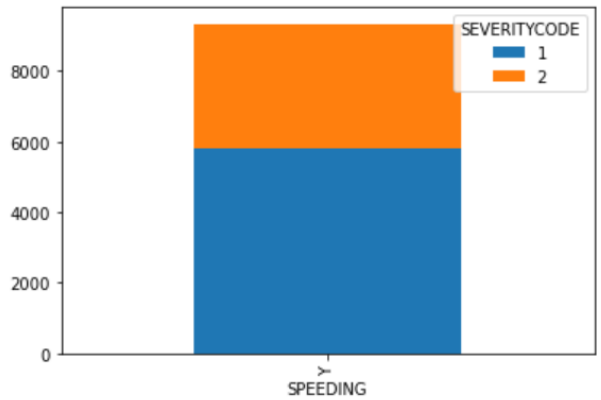
Evaluating each chosen variable against severity to understand if all the values of independent variables are evenly distributed with dependent variable values and identify any significatn outliers.

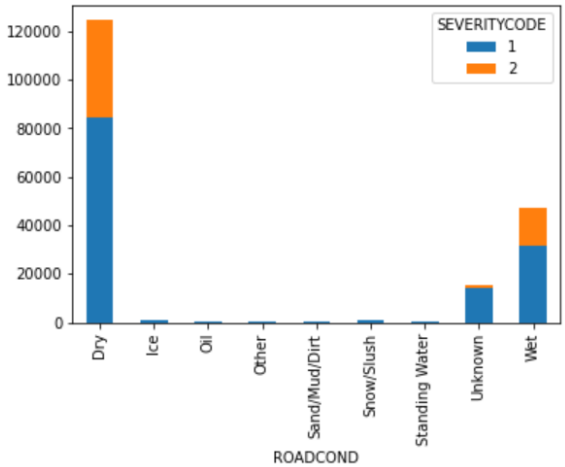


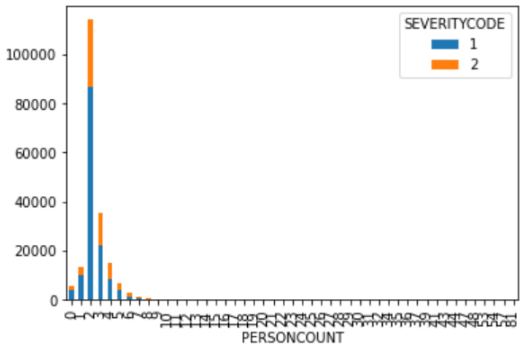


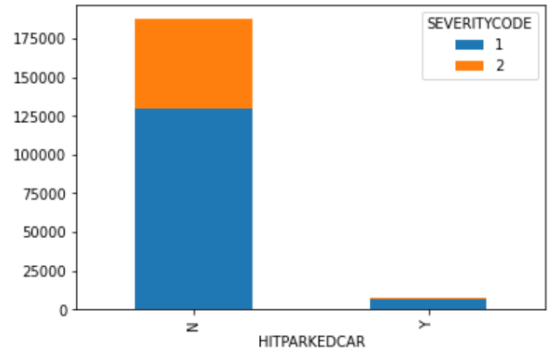


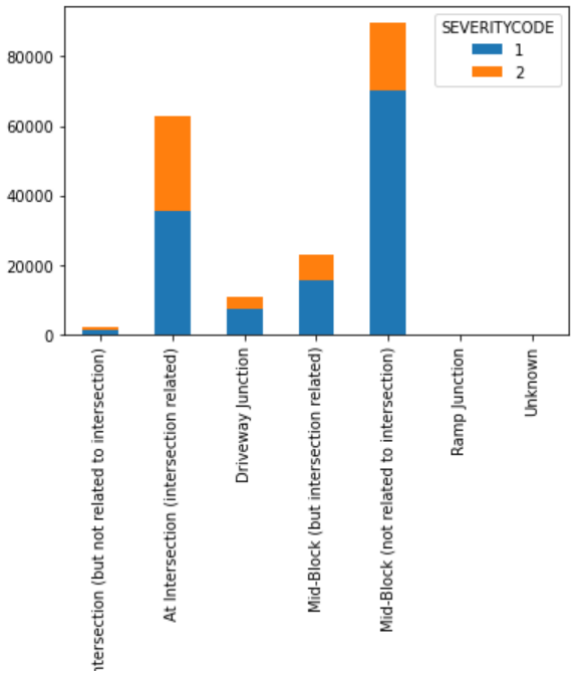


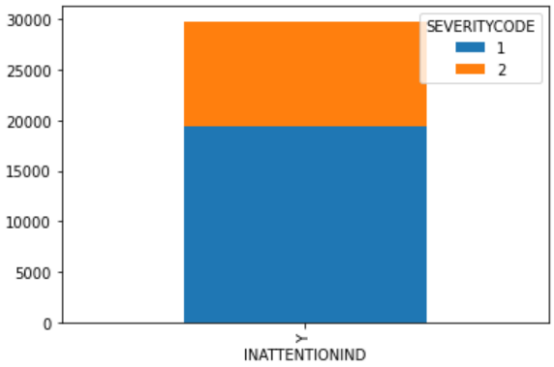


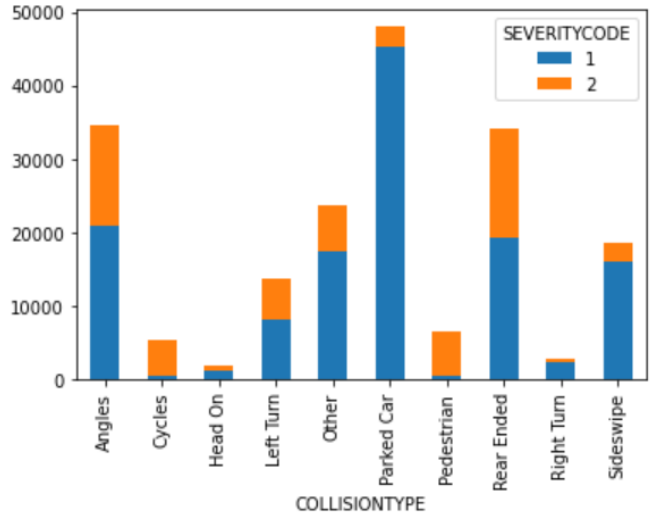




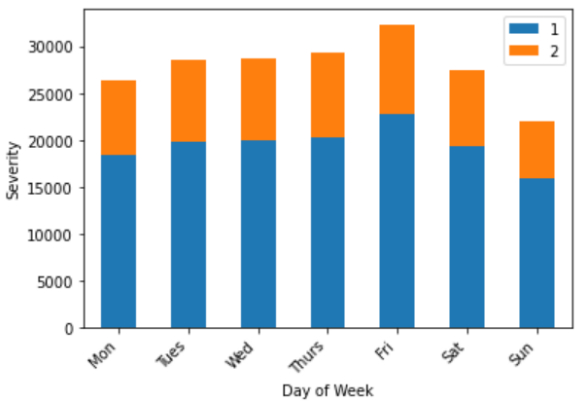


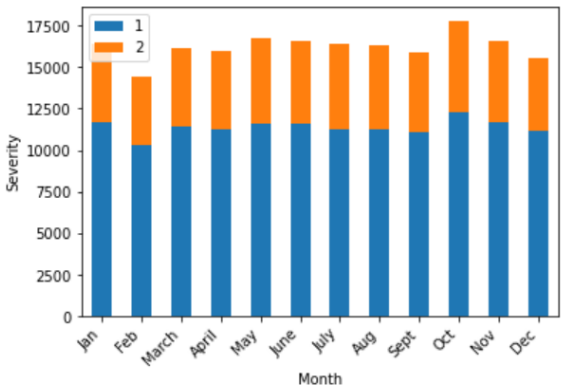






There aren't any major outliers in all the attributes chosen and data of severity levels 1 and 2 are distributed in enough manner to perform the analysis and further analysis is needed if time is a factor in increasing severity like a particular day of the week or a particular month has recorded highest number of accidents significantly. But it is clear from the below figures that there aren't any such noticeable insights.





Hence as all chosen variables have not many outliers and we can use these as independent variables to build our model. Then all the chosen attributes having missing values are identified and filled with relevant data accordingly to ensure smooth operations and avoid any skewed output. For categorical data, I replaced the empty values with NaN or missing. For numerical data, I filled with the median of the data as mean considers outliers and all categorical values are converted into numerical values to perform modelling. All duplicate values like 0,1 in Y, N columns are converted into 0,1 respectively (For example Under influence column had both Y,N and 0,1 entries which were converted into 0,1 respectively). Below find the snapshot of the final dataset for model consideration.

	ADDRTYPE	COLLISIONTYPE	PERSONCOUNT	VEHCOUNT	JUNCTIONTYPE	INATTENTIONIND	WEATHER	ROADCOND	LIGHTCOND	HITPARKEDCAR	SPEEDING	UNDERINFLnew
0	3	5	2.0	2.00000	7	0	7	8	8	1	0	0
1	2	1	2.0	2.00000	1	0	10	8	5	1	0	0
2	2	0	4.0	3.00000	1	0	7	6	8	1	0	0
3	2	4	3.0	3.00000	1	0	8	6	8	1	0	0
4	3	5	2.0	2.00000	7	0	10	8	8	1	0	0
5	3	5	2.0	2.00000	7	0	8	6	8	1	0	0
6	3	5	2.0	2.00000	7	0	10	8	8	1	0	0
7	3	9	3.0	1.00000	7	0	8	6	8	1	0	0
8	2	0	2.0	2.00000	1	0	8	6	8	1	0	0
9	3	5	2.0	2.00000	7	0	8	6	8	1	0	0
10	1	4	2.0	2.00000	4	0	7	6	8	1	0	0

PREDICTIVE MODELLING AND RESULTS

I have used both classification and regression models such as Logistic Regression, Decision Tree and Random Forest in order to identify which will provide better accuracy and precision. Different models were also used in order to test the efficiency of the independent variables chosen to predict severity. I have skipped KNN due to large entries of data and multiple independent variables.

LOGISTIC REGRESSION RESULTS

```
Jaccard: 0.7181841585857719
F1 Score: 0.8359804215363101
Accuracy:  0.7469628868627199
LogLoss: 0.4952020776023226
              precision    recall  f1-score   support

         1.0       0.77      0.90      0.83     27312
         2.0       0.62      0.38      0.47     11623

    accuracy                           0.75     38935
   macro avg       0.70      0.64      0.65     38935
weighted avg       0.73      0.75      0.73     38935
```

DECISION TREE RESULTS

```
Accuracy:  0.7493001155772441
Jaccard: 0.734741018533616
F1 Score: 0.8470901543040652
              precision    recall  f1-score   support

         1.0       0.77      0.92      0.84     27312
         2.0       0.65      0.34      0.44     11623

    accuracy                           0.75     38935
   macro avg       0.71      0.63      0.64     38935
weighted avg       0.73      0.75      0.72     38935
```

RANDOM FOREST CLASSIFIER RESULTS

```
              precision    recall  f1-score   support

         1.0       0.77      0.92      0.84     27312
         2.0       0.65      0.35      0.45     11623

    accuracy                           0.75     38935
   macro avg       0.71      0.63      0.65     38935
weighted avg       0.73      0.75      0.72     38935
```

From the results from above models, it is clear that Decision Tree and Random Forest models have similar results and we can use any of these two predict severity of results and chosen factors are indeed the ones influencing the severity of accidents and government and road welfare agencies should look after these in developing better infrastructure and better measures during different weather conditions to reduce the number of accidents and also the severity.

CONCLUSION

In this analysis, I analysed the relationship between the ,severity of accidents and the attributes that contribute to them. Identified several factors like Inattention, Under Influence and Parked Cars are one of the major contributors of accidents and measures are needed to control them. I built both classification and regression models to cross-check the ability of data attributes I have chosen to predict the severity and almost the same results in multiple models do reflect.

RECOMMENDATIONS

As main factors like Inattention, Under Influence and Parked Cars did come out as some significant contributors and road-safety measures needs to be improved to educate the users about inattention and strict measures to control alcohol/drugs influence.