Data Science Capstone Project

Introduction

The road transportation provides fast means of transportation. However, it is also subjected to the various kind of accidents. There are many factors and sources of an accident. Data on accident can provide the insights into these factors and sources and can help building better infrastructure or planning travelling schedule based on situational conditions.

The report would focus on infrastructural factors into accidents and help better infrastructure building and retrofitting of existing infrastructure along with keeping in mind the environmental & situational factors. Safety feature embedded on the time of building infrastructure have potential to reduce number of accidents and decreasing the severity of accidents. This report would help the architects and builder of road infrastructure to build safe road.

Data understanding

For the analysis, data should have infrastructural, environmental and situation factors for an accident.

I. Infrastructural factors would include type of junctions and type of turn. For example, existence of bland turn may increase the probability of accidents.

II. Environmental factors include raining and lighting condition at the time of accident. For example, heavy rainfall may increase the probability of accidents, specially, if the road has slop and it is slippery.

III. Situational factors involve factors like vehicle types and drivers' condition. For example, speedy and driver under influence of drug or alcohol have high probability of meeting an accident.

Methodology

Data cleaning:

The datasets have 38 columns and many of these columns were not useful in training of the model. Therefore, a number of columns were deleted from the dataset.

```
df_clear = df.drop(columns=['SDOT_COLDESC', 'ST_COLDESC','SEGLANEKEY',
'CROSSWALKKEY', 'SDOTCOLNUM', 'SEVERITYCODE.1', 'SEVERITYDESC',
'LOCATION','STATUS', 'INCKEY', 'COLDETKEY', 'X', 'Y', 'INTKEY',
'EXCEPTRSNCODE', 'OBJECTID', 'REPORTNO','EXCEPTRSNDESC' ,'SPEEDING',
'PEDROWNOTGRNT', 'INATTENTIONIND', 'SDOT_COLCODE', 'ST_COLCODE'])
```

Datasets have one target category of accident namely the severity of accident, which is represent by a code having either value of 1 or value of 2. However, there were imbalance in the number of rows having severity category of 1 or 2. The severity category 1 have 136485 number of rows and the severity category 2 has 58188 number of rows. The imbalance in the target category would induce the biases in the trained model.

```
df_clear['SEVERITYCODE'].value_counts()
Out[7]:
1    136485
2     58188
Name: SEVERITYCODE, dtype: int64
```

Therefore, the random rows of data where severity code is '1' were deleted to balance the number of rows from both the severity category.

```
df_clear.drop(df_clear.query('SEVERITYCODE ==1') .sample(frac=.57).
index , inplace=True)
```

The column “HITPARKEDCAR” which give the information of Whether or not the collision involved hitting a parked car have the values of ‘0’, ‘1’, ‘N’ and ‘Y’. Since the machine learning algorithm can only be trained on numerical values, the value of ‘N’ and ‘Y’ were replaced with the value of ‘0’ and ‘1’, respectively.

```
df_clear.replace({'HITPARKEDCAR': {'N': '0', 'Y': '1'}}, inplace=True)
```

Similarly, the column ‘UNDERINFL’ which give the information of Whether or not a driver involved was under the influence of drugs or alcohol have the values of ‘N’ and ‘Y’, which were replaced with the value of ‘0’ and ‘1’, respectively.

```
df_clear.replace({'UNDERINFL': {'N': '0', 'Y': '1'}}, inplace=True)
```

Data extraction

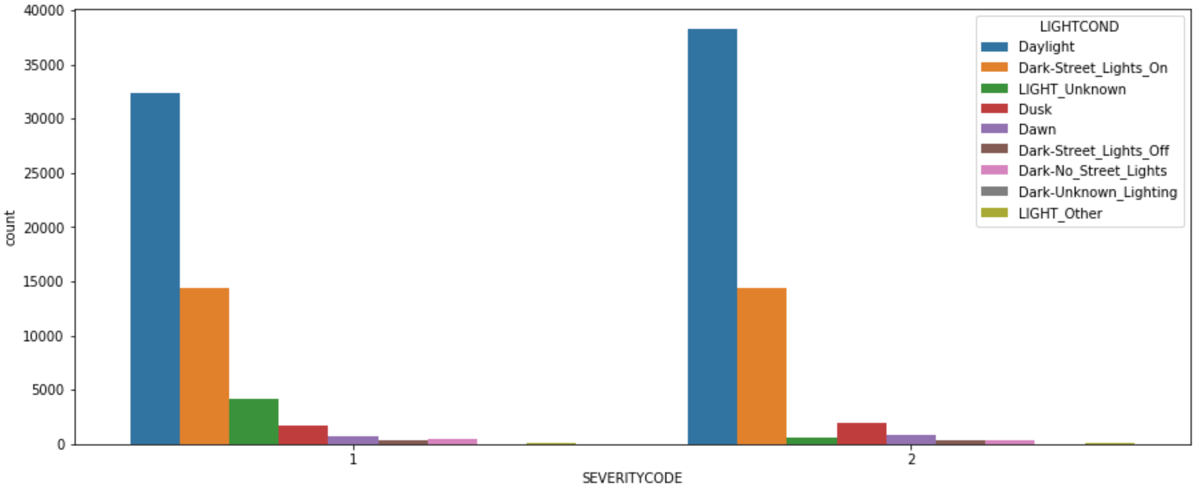
The dataset has the datetime of the accident. From datatime of the accident, the day of week and the hour of accident can be extract. This extracted data can be used in analysing the pattern of accident on the basis of day of week and the hour of accident.

```
df_clear['DAYOFWEEK'] = df_clear['INCDATE'].dt.dayofweek

df_clear['TIMEINHOURS'] = df_clear['INCDTTM'].dt.hour
```

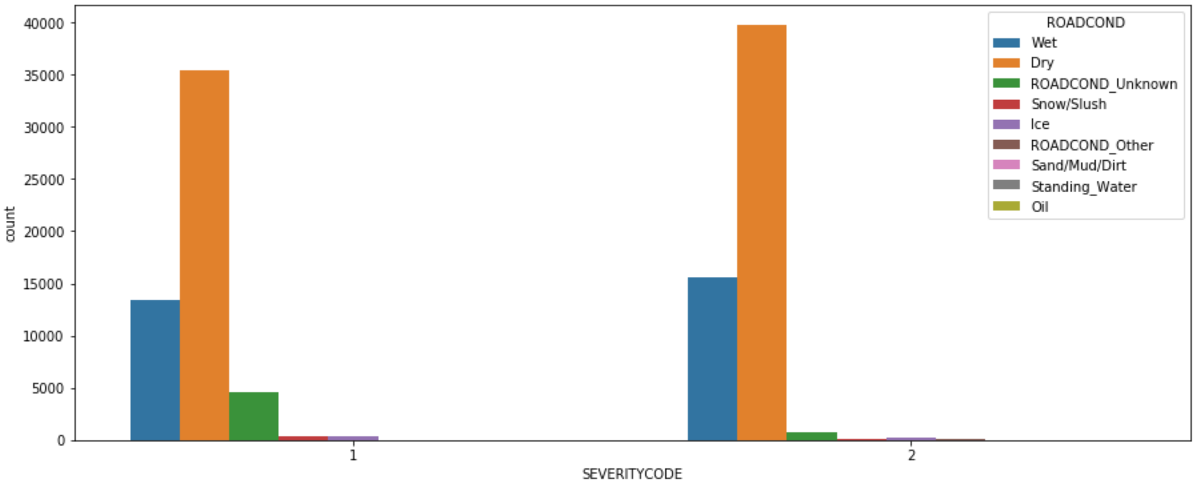
Exploratory Data Analysis

1. Relationship between Lighting condition and severity of accident:



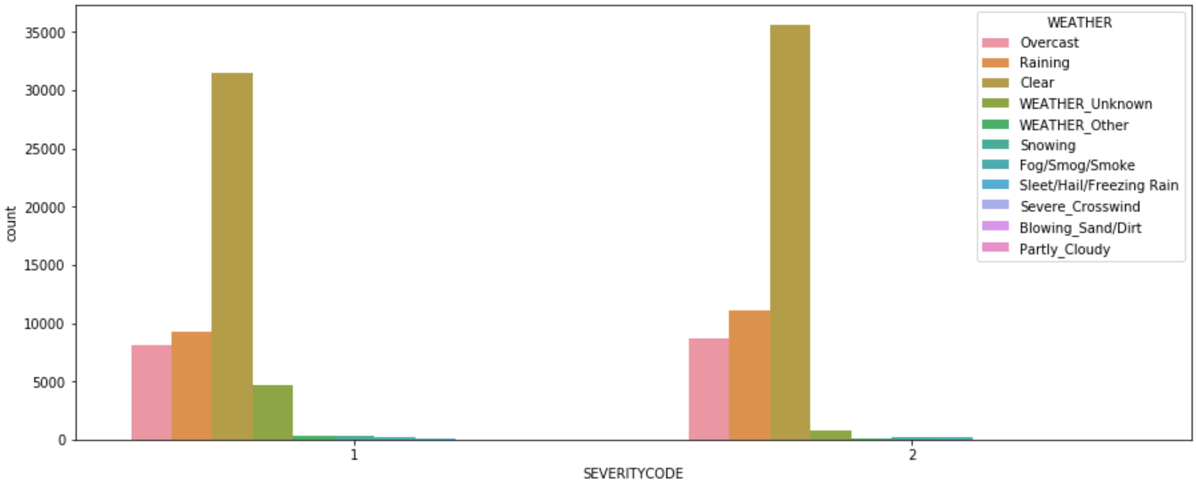
From the plot, it is clear that in daylight condition the severity category-2 accident are higher than the severity category-1 accident. This relationship is contractor to the common understanding that daylight would have more of severity category-1 accidents.

2. Relationship between Road condition and severity of accident:



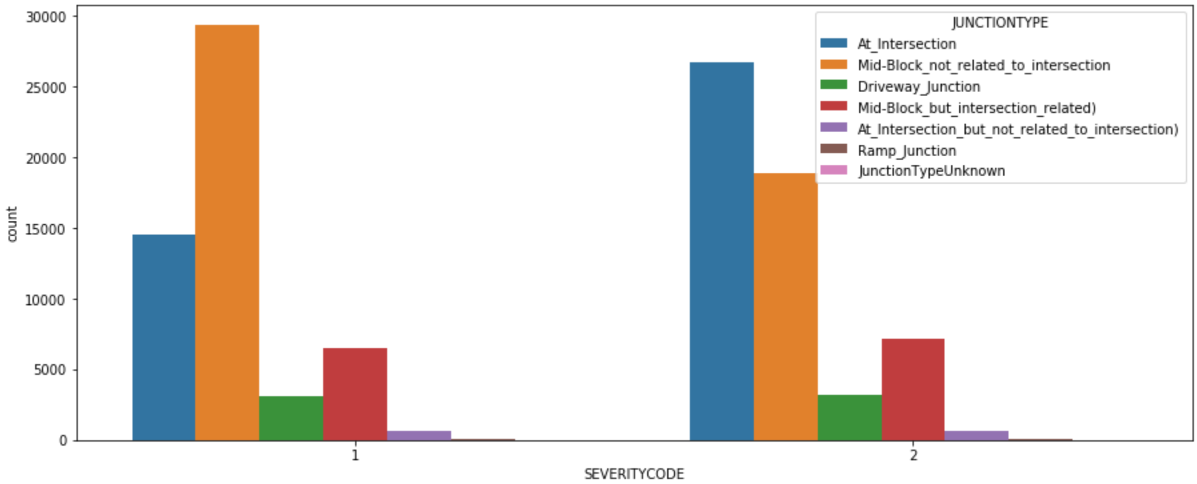
From the plot, it is clear that road condition doesn’t have much influence in degerming the type of severity of the accident. The wet road conditions almost have similar number of accidents for both the severity. Further, Dry road conditions contribute to a greater number of accidents as against the wet road conditions. It may be due the fact that dry condition is more prevalent than wet road condition. Since we don't have time proportion of dry to wet condition, we can't make a prediction that wet condition have more probability than dry condition.

3. Relationship between weather condition and severity of accident:



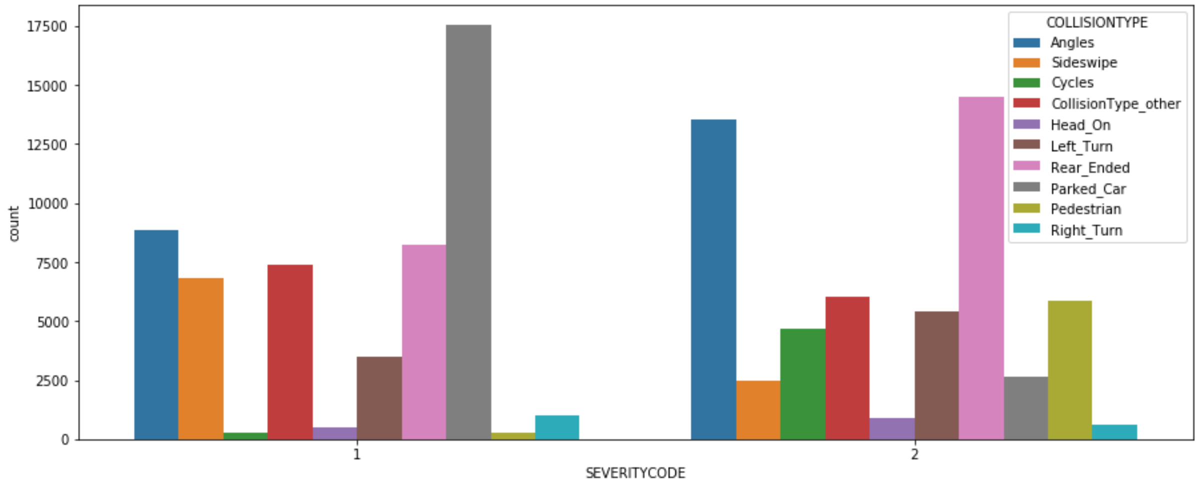
From the plot, it is clear that weather condition doesn't have much influence in degerming the type of severity of the accident. The clear weather has marginally a greater number of Severity-2 accidents the Severity-1 accidents. Further, raining conditions doesn't seem to have influence on determining the severity of accidents as both the category of severity of accident have almost similar values.

4. Relationship between Junction type and severity of accident:



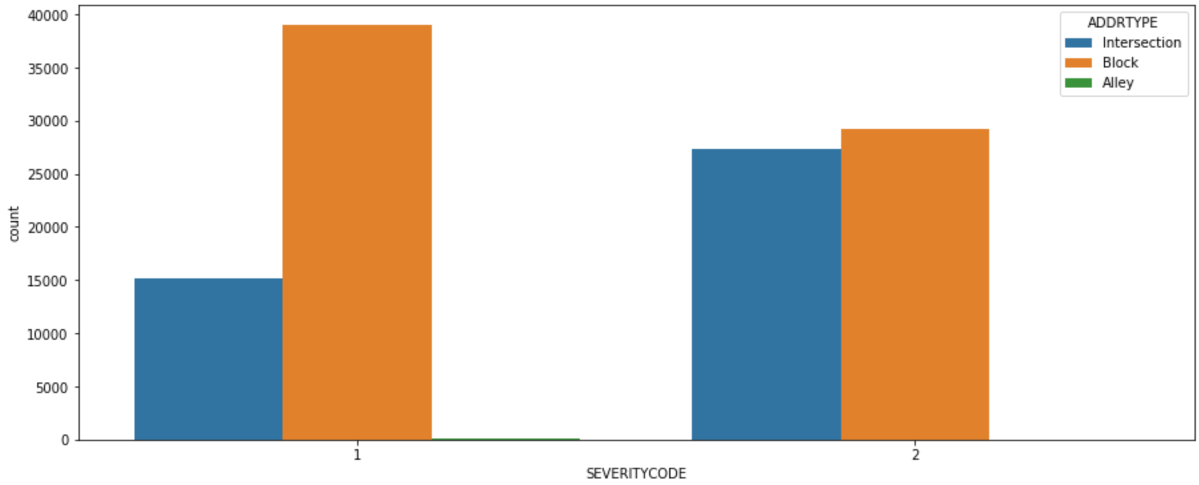
From the plot, it is clear that junction type has clear relationship with the severity of accidents. At intersection junction, there are more severity-2 accident than severity-1 accidents. In contrast, mid-block junction have a greater number ofe severity-1 accident than severity-2 accidents.

5. Relationship between Collision type and severity of accident:



From the plot, it is clear that collision type has some corelative relationship with type of severity of accident. As most of parked car collision have severity-1 accidents and the greater number of angles collision have severity-2 type accidents.

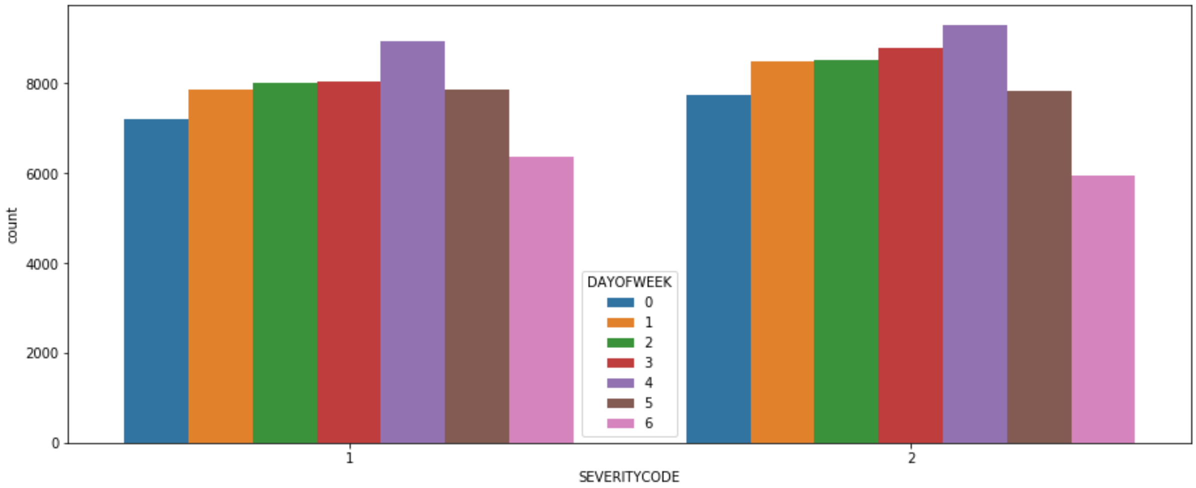
6. Relationship between address type and severity of accident:



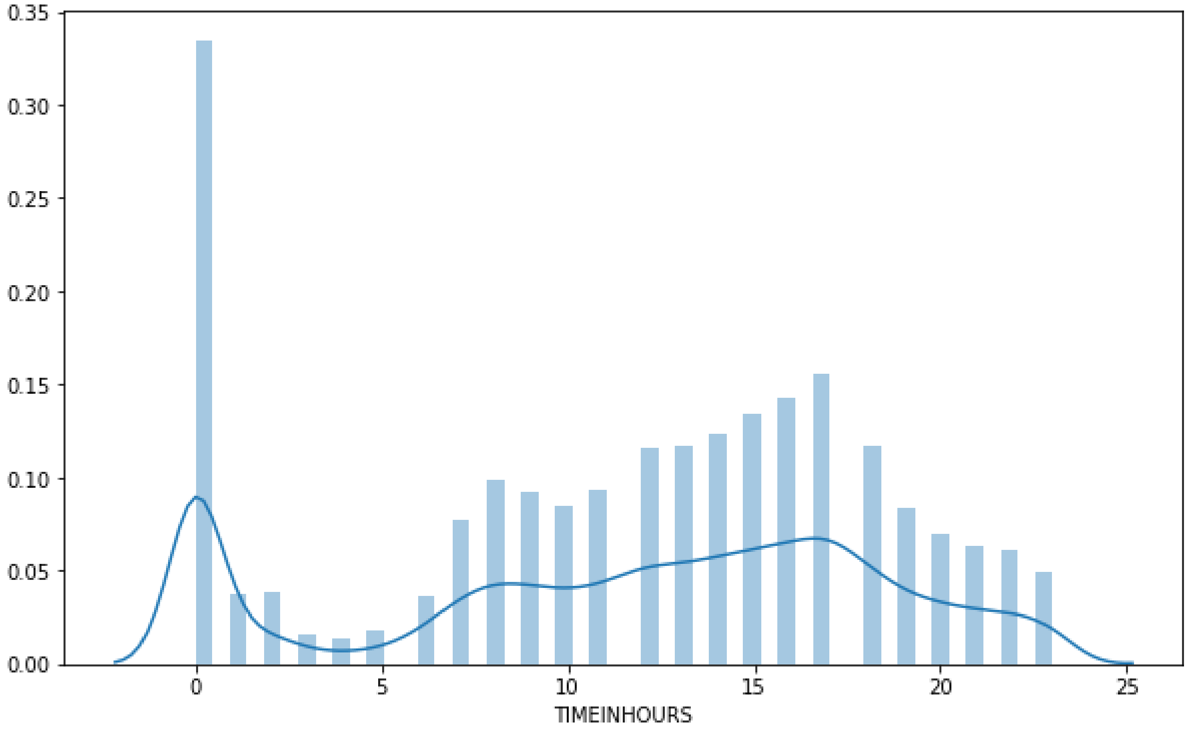
From the plot, it is clear that address type ‘Block’ has a greater number of severity-1 type accidents compared to the severity-2 type accidents. In contrast, address type ‘intersection’ type has a greater number of severity-2 type of accidents compared to the severity-1 type of accidents.

7. Relationship between day of the week of the accident and severity of accident:

From the plot, it seems that day of the week don't have a relationship with the severity of the accidents. However, it is clear from the plot that the numbers of accidents are highest on Fridays and the lowest on Sundays.



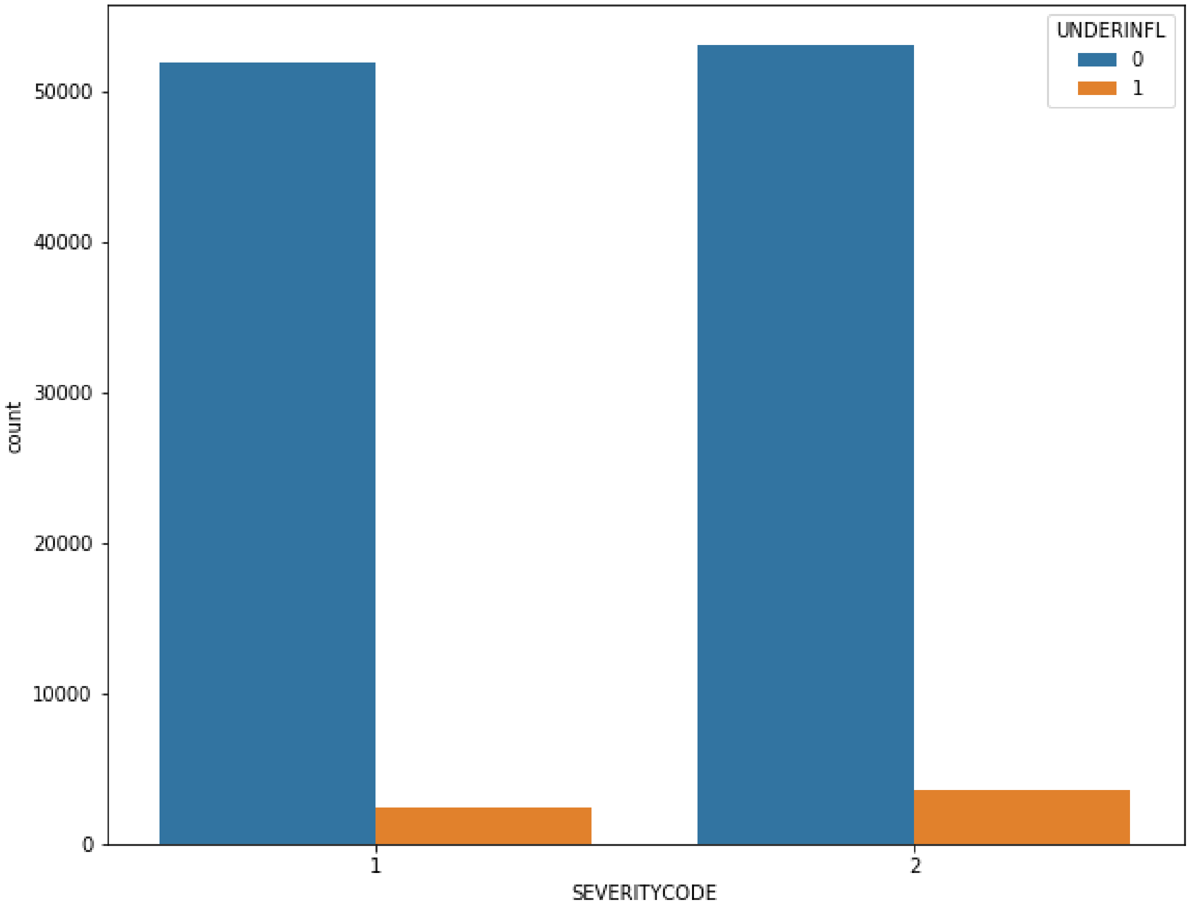
8. Relationship between hour of the accident and severity of accident:



From the plot above, it is clear that the highest numbers of accidents happened between 5PM and 6PM and lowest of accidents happened between 1AM to 5AM. The highest frequency on 12AM hours is due the fact the time of the accident is missing and 00:00 is taken as a default value.

9. Relationship between driver under the influence of drug or alcohol and severity of accident:

From the plot below, it seems that there is no correlation between driver under the influence of drug or alcohol and the severity of the accidents. However, it is clear from the plot that a greater number of accidents happen by the driver who were not under the influence of drug or alcohol.



Dealing with categorical features- creating dummy variables

The machine learning model algorithms can only understand numerical values for training and prediction of result. Therefore, the categorial features are needs to be converted into values, which machine learning algorithm could understand. The way to do is by creating dummy variable of the distinctive values of categorial feature and assigning 1 or 0 for these dummy variables.

```
addressType = pd.get_dummies(df_clear['ADDRTYPE'], drop_first=True)
df_clear.drop(['ADDRTYPE'], axis=1, inplace =True)
df_clear = pd.concat([df_clear, addressType], axis=1)
Collision = pd.get_dummies(df_clear['COLLISIONTYPE'], drop_first=True)
df_clear.drop(['COLLISIONTYPE'], axis=1, inplace =True)

JUNCTION = pd.get_dummies(df_clear['JUNCTIONTYPE'], drop_first=True)
df_clear.drop(['JUNCTIONTYPE'], axis=1, inplace =True)

WEATHER = pd.get_dummies(df_clear['WEATHER'], drop_first=True)
df_clear.drop(['WEATHER'], axis=1, inplace =True)

ROAD = pd.get_dummies(df_clear['ROADCOND'], drop_first=True)
df_clear.drop(['ROADCOND'], axis=1, inplace =True)
```

```
LIGHT = pd.get_dummies(df_clear['LIGHTCOND'], drop_first=True)
df_clear.drop(['LIGHTCOND'], axis=1, inplace =True)

df_clear = pd.concat([df_clear, Collision, JUNCTION, WEATHER, ROAD,
LIGHT], axis=1)
```

Result Section

Predictive Modelling

Here the target variable is the severity of accident which have two possible values of ‘1’ or ‘2’. The model is trained on the feature of dataset to predict the severity of the accident. Since the target variable is a categorical variable, classification models were applied on the dataset. The classification model like KNeighbors Classifier uses the Euclidean distances, therefore, features of datasets are standardised to remove biases of an individual feature on the model.

```
from sklearn import preprocessing
X = preprocessing.StandardScaler().fit(X).transform(X.astype(float))
```

The dataset is divided into training set and test set so that models the performance can be assessed on the out of sample dataset.

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split( X, y, test_size =
0.3, random_state=7)
print ('Train set:', X_train.shape,  y_train.shape)
print ('Test set:', X_test.shape,  y_test.shape)
```

Result

The KNeighbors Classifier, Logistic Regression, Support Vector Classifier algorithms were used to train and test the model. The performance of these model is provided below:

	KNeighbors Classifier	Logistic Regression	Support Vector Classifier
Train set Accuracy	0.74906608269998	0.70358108978487	0.71201854952982
Test set Accuracy	0.68260647	0.70463767244747	0.70788374259865

Fig: Performance of classification models.

The SVM model is the best performing model though the difference with the other model is very small.

Discussion

The accident dataset is analysed with the help of visualisation and modelling of data. The relationships between the severity of the accidents with other factors is analysed. There emerged a clear understanding that the address type of the site of the accident have strong correlation with the severity of the accident. Similarly, junction type of the site of the accident have strong correlation with the severity of the accident.

And the highest number of accidents happened on the Fridays and lowest on Sundays. This could be explained on the basis of the facts that the on Sundays, most of us don’t have to go to work and traffic are mostly at its lowest on the Sundays. On the other hand, on Fridays, most of us attend work and go to attend various pressure seeking gathering to celebrate the next two days’ off.

Conclusion

Our analysis starts with the finding the correlation between accidents and its severity with the infrastructural, environmental and situational factors of accidents. With the exploratory data analysis with visualisation of correlation, our understanding of accidents and its various factors have improved. Our analysis shows that environmental factors like weather, road condition and lighting condition have not had any significant effect on severity of accidents.

In contrast, there is strong correlation between infrastructural factors and severity and frequency of accidents as more sever accidents happened on the intersection as compared to, on the blocks. Further, situational factors can be seen in two ways. Firstly, individual situational factors like driver under influence of drug and alcohol have least effect on determining the frequency and severity of accidents. Secondly, cumulative situational factors like hours of the day and day of the week have greater effect on the accidents.