IBM Capstone Project – Predicting car accident severity in Seattle city.

Introduction:

According to the World Health Organisation, approximately 1.35 million people die each year because of road traffic accidents. Children and young adults in the age group of 5-29 years of age are the most effected and the most vulnerable among the road users are pedestrians, cyclists, and motorcyclists. Some of the reasons contributing to this problem are bad weather conditions like heavy rains, snow & bright sun, people who do not follow road rules, inexperienced drivers, and bad junctions etc. As a result of road accidents there can be huge economic loss to individuals and to the nation. Non-fatal injuries can cause lifelong disabilities and sometimes we can even lose our loved ones. With advancement in technology and vast availability of data, it is incredibly important to develop a model that can predict the accident and its severity so that we can save valuable lives and help the economy. In this project, I will be attempting to develop a model that can predict the accidents and its severity in the American city, Seattle.

The target audience who would be interested in this problem are the drivers themselves who can opt to get alerts on the probability of being in an accident, health and emergency response teams, the police and the insurance companies.

Data:

The data required for this project has been downloaded from the coursera IBM data science course, which was obtained from the Seattle Police Department and Accident Traffic Records Department from 2004 to present.

The data consists of 37 independent variables and 194,673 rows. The dependent variable, “SEVERITYCODE”, contains numbers that correspond to different levels of severity caused by an accident represented by 1 and 2.

Data dictionary:

SEVERITYCODE	Text, 100	Severity codes are as follows: 1: Very Low Probability — Chance or Property Damage 2: High Probability — Chance of injury and fatality.
ADDRTYPE	Text, 12	Collision address type: • Alley • Block • Intersection
COLLISIONTYPE	Text, 300	Collision type
VEHCOUNT	Double	The number of vehicles involved in the collision.

		This is entered by the state.
JUNCTIONTYPE	Text, 300	Category of junction at which collision took place
INATTENTIONIND	Text, 1	Whether or not collision was due to inattention. (Y/N)
UNDERINFL	Text, 10	Whether or not a driver involved was under the influence of drugs or alcohol.
WEATHER	Text, 300	A description of the weather conditions during the time of the collision.
ROADCOND	Text, 300	The condition of the road during the collision.
LIGHTCOND	Text, 300	The light conditions during the collision.
SPEEDING	Text, 1	Whether or not speeding was a factor in the collision. (Y/N)
HITPARKEDCAR	Text, 1	Whether or not the collision involved hitting a parked car. (Y/N)
Hour	Double	Derived from date and time column, represents time of the day.

Methodology:

- Data preparation: In this phase, all the necessary libraries are imported and the .csv dataset is imported and has been converted to a pandas dataframe for further evaluation. A Jupyter notebook has been used for this project.
- Exploratory data analysis: In this phase, descriptive statistics of the data was obtained to get a quick feel of the data and to see what we are dealing with.
  - Bar graphs and maps were used to explore the relationship between different variables and the target variable.
  - Variables that provide no meaningful information or with no information were dropped.
  - Highly correlated variable columns were deleted as they provide the same information as the target variable.
  - There are some columns with Boolean values like 0 and 1, they have been converted to No and Yes respectively to match the date.
  - The date and time column were also casted as date and time stamps.
  - Rows with missing values were dropped.
  - Dummy variables were created for the variables with categorical values.
  - Finally, we can make sure that the data has no missing values, no categorical values and the data is nice and clean and ready for analysis.
  - All the data except the target variable are converted to string type, as all the variables are now categorical type, mutual information feature selection technique is applied to choose the features for training and developing the model.

Modelling:

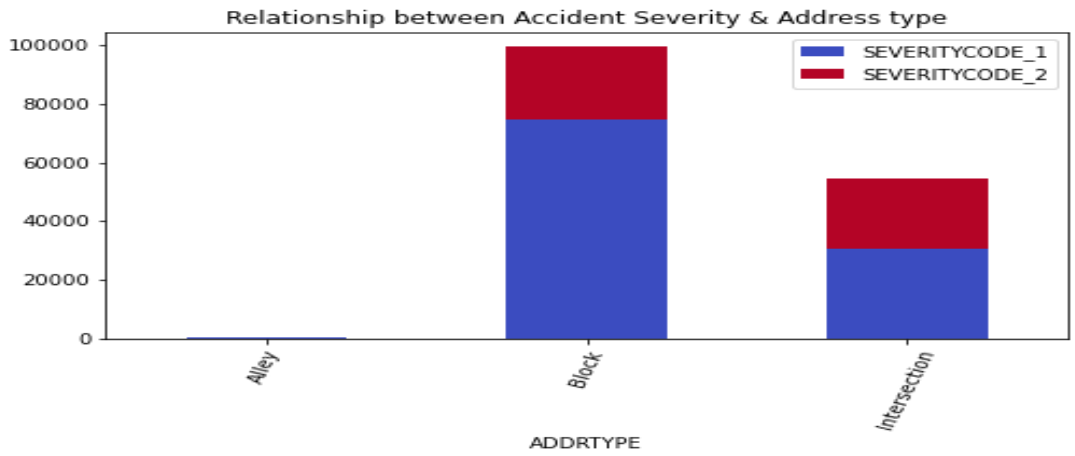
In this phase, the data was divided into X and y matrices, X being the independent variable and y as the target or the dependant variable.

- Data was scaled using the standard scalar function.
- The dataset was split into test set and training set, 80% of the data was used to train the model and it has been evaluated on 20% of the test data.
- The data was trained on three different models: Logistic regression, Knn classifier, and Decision Tree classifier.
- Training set results and test set results were compared to evaluate the model performance.

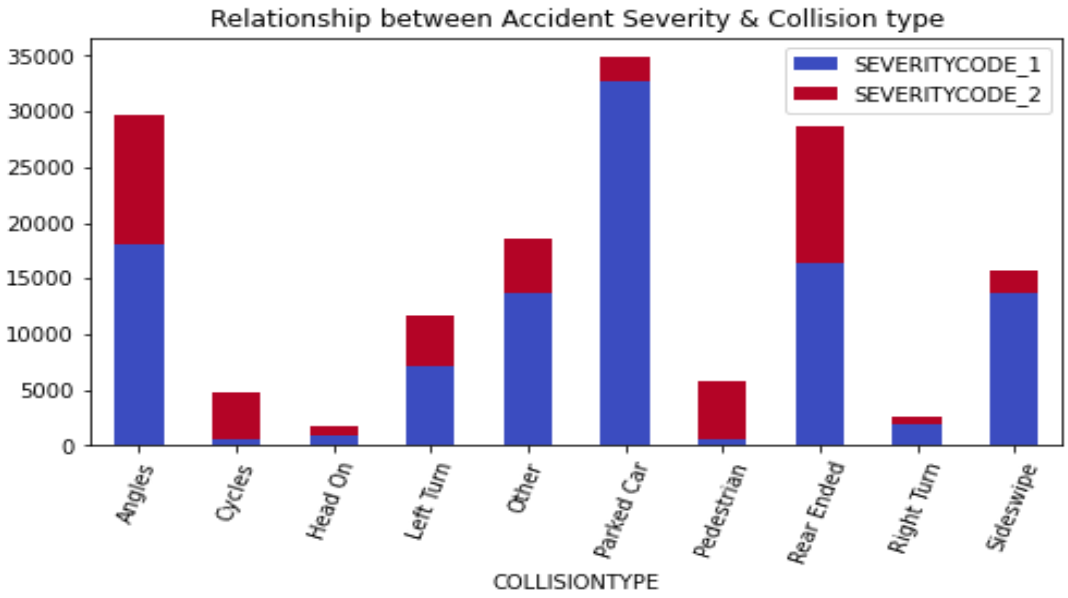
- Evaluation: In this phase, four different metrics were used to evaluate the models i.e., Jaccard index, F1-score, R2-score and accuracy.

Results:

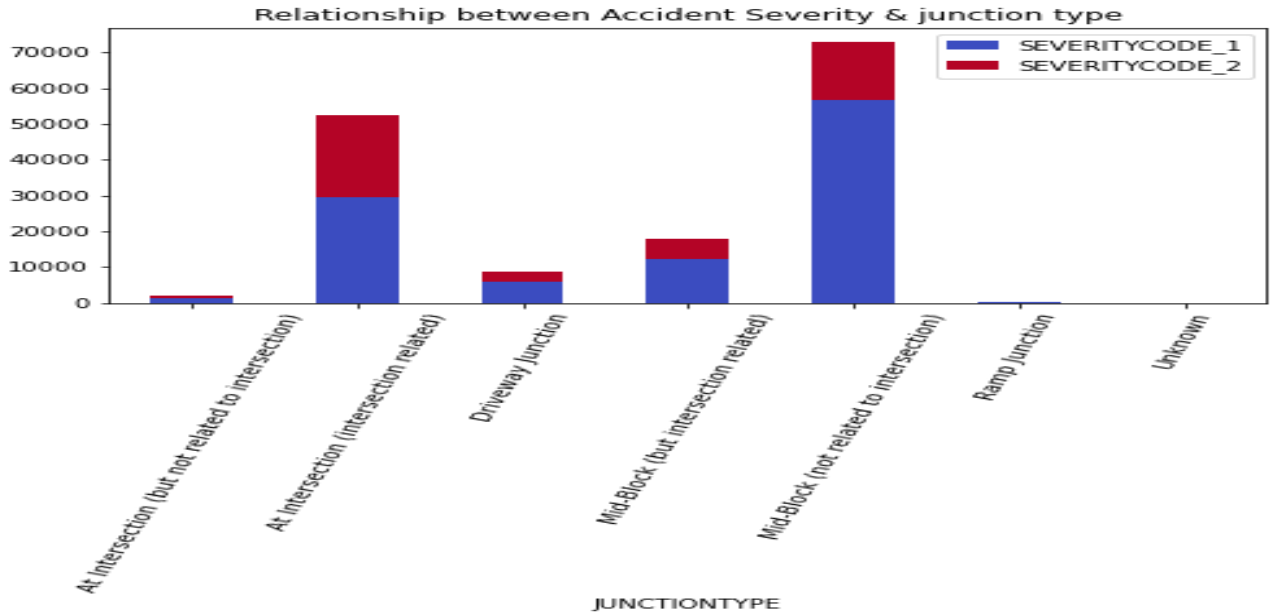
Exploring the relationship between address type and severitycode: From the graph, although it appears that there are more accidents happening near blocks, the accident severity seems to be high near the intersections. Traffic monitoring authorities can use this data to enforce more traffic regulations near blocks and intersections.



Exploring the relationship between Collision type and severitycode: From the collision graph, we can see that majority of the collisions happen with a parked car, but the severity of the accidents is high when collision happens from rear-end or angles. Car manufacturing companies can use this information to manufacture car models with more safety features like auto-park, parking sensors, lane departure warnings, auto-braking etc.

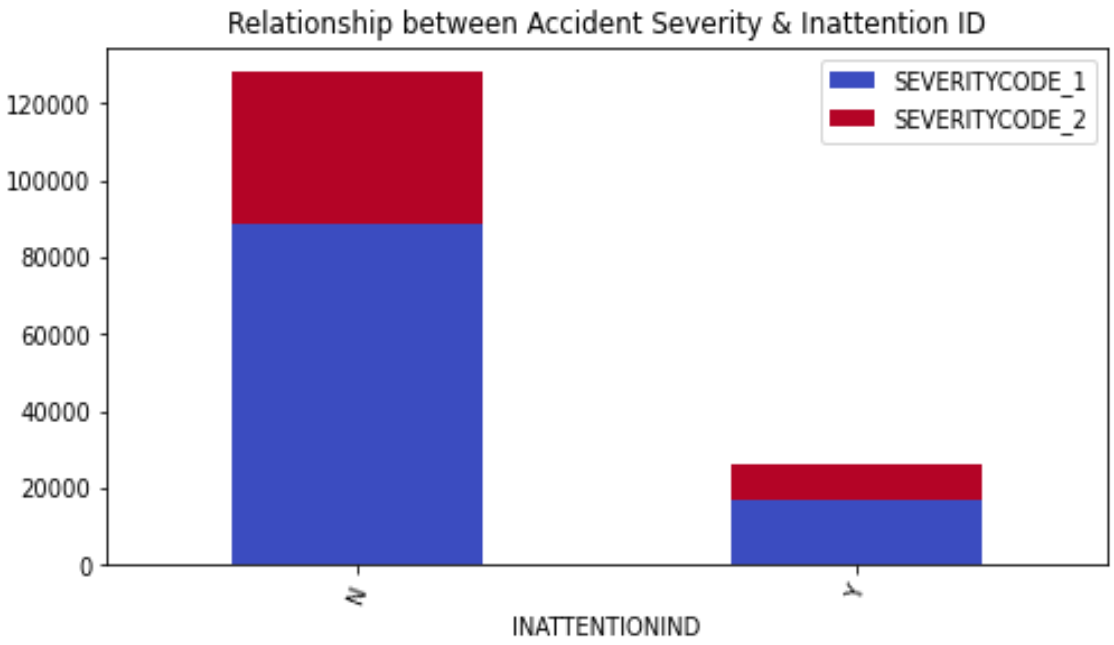


Exploring the relationship between junction type and severitycode: From the graph, we can note that majority of the accidents at intersection and mid-block, with high severity being at the intersections. Like discussed above in the Address type graphs, more traffic enforcements should be in place at intersections to avoid accidents and its severity.

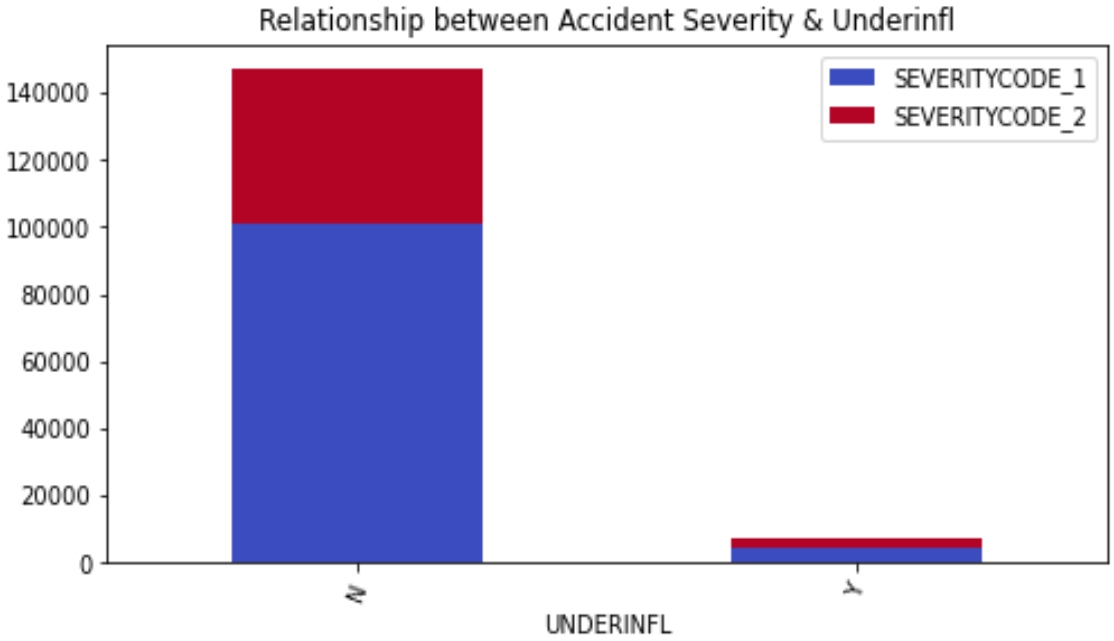


Exploring the relationship between driver's attention to road and severitycode of the accident: It is interesting to note that not many accidents happen due to driver's inattention.

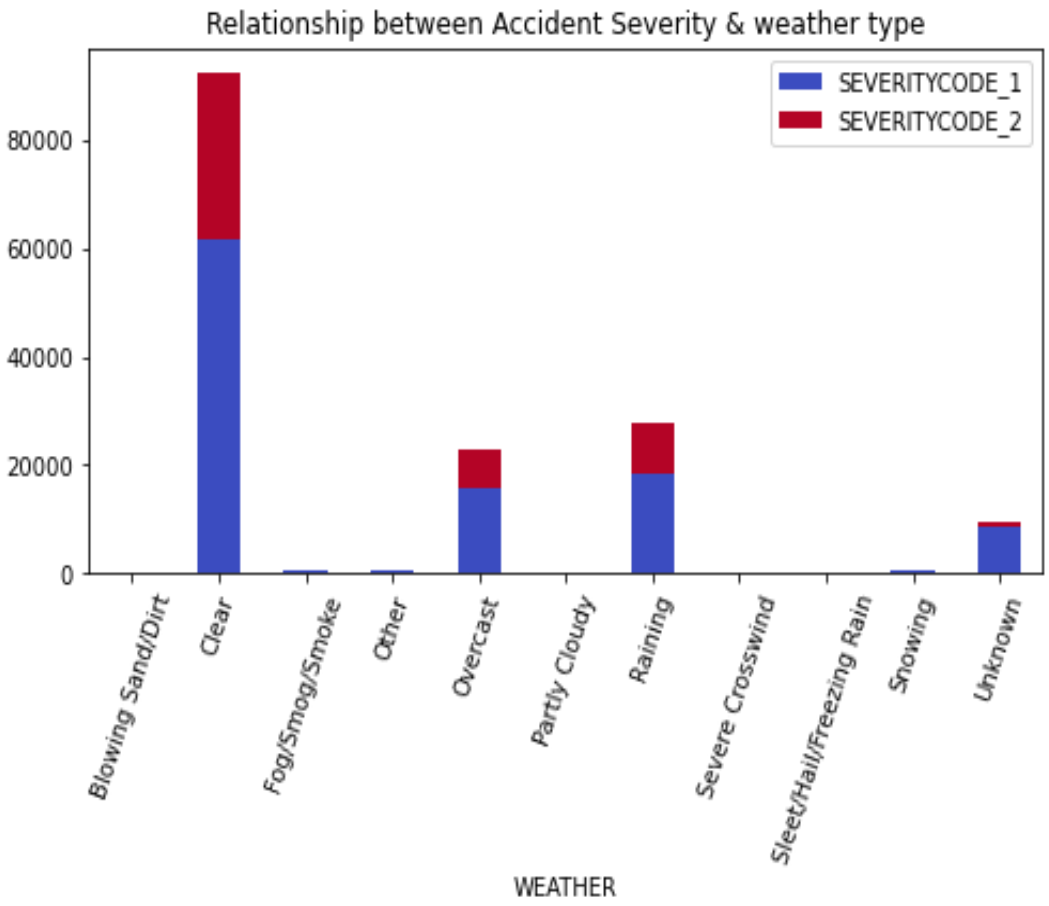
And, also with accidents happening due to drivers' inattention the severity looks low.



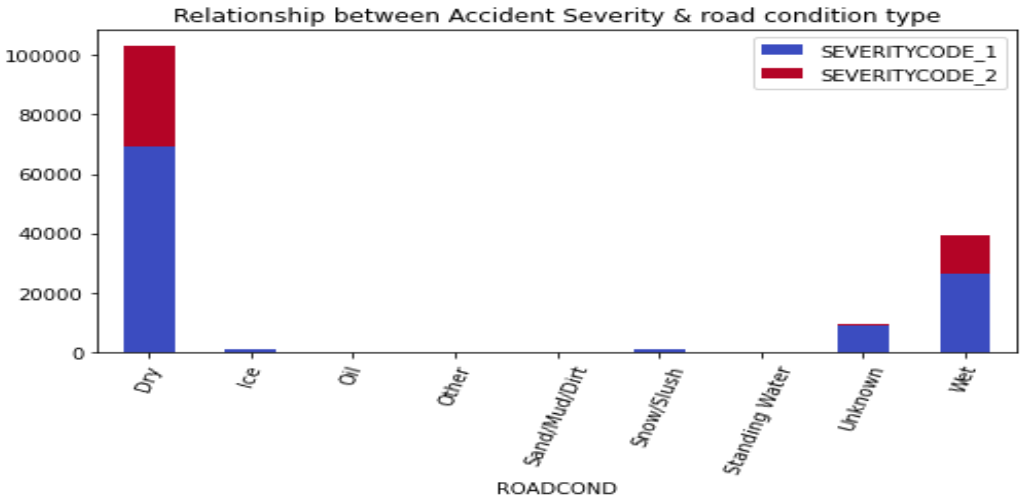
Exploring the relationship between influence of alcohol and severitycode: Like driver's inattention, the accidents happening under influence also appears to be low with low severity.



Exploring the relationship between weather condition and severitycode: From the bar graph, surprisingly majority of the accidents happen on a clear day and with high severity. Rainy days and overcast days also have some accidents.

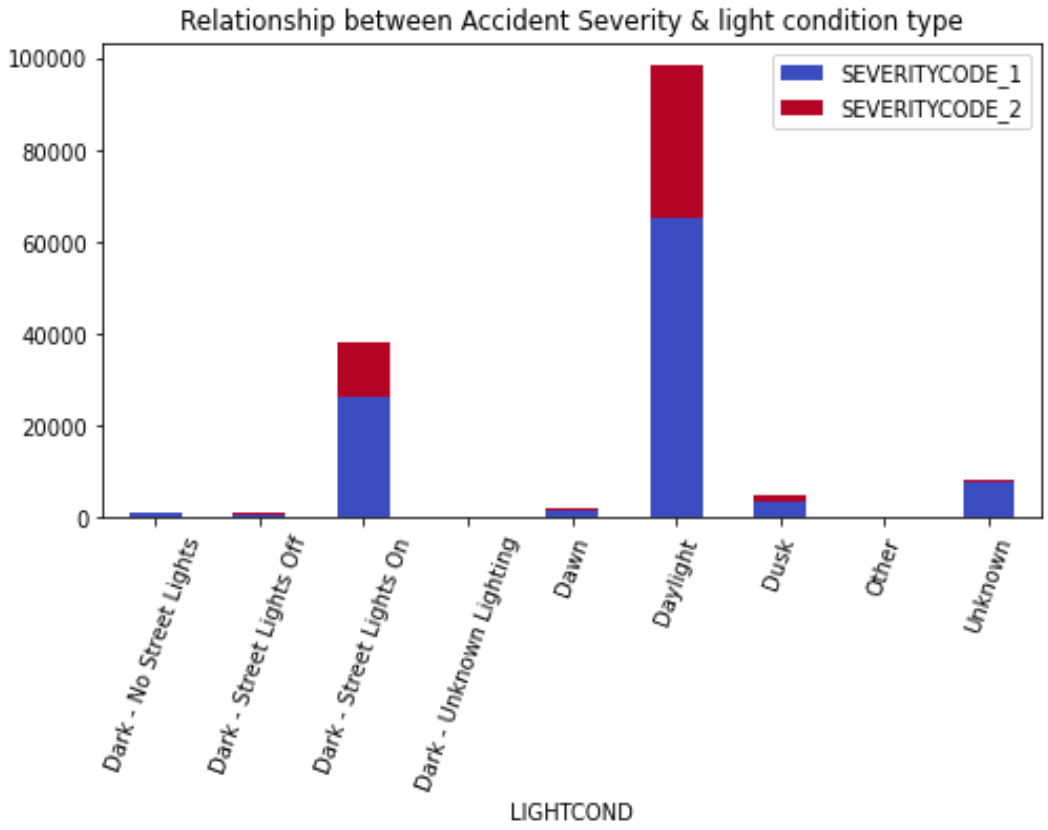


Exploring the relationship between road condition and severitycode: Just like the weather, more accidents and with high severity happen on a dry day. From the weather and road conditions graphs, we can see that light is one of the contributing factors for the accidents. Drivers should consider wearing sunglasses or other appropriate safety equipment when the weather is clear, dry, and sunny to avoid being in an accident.

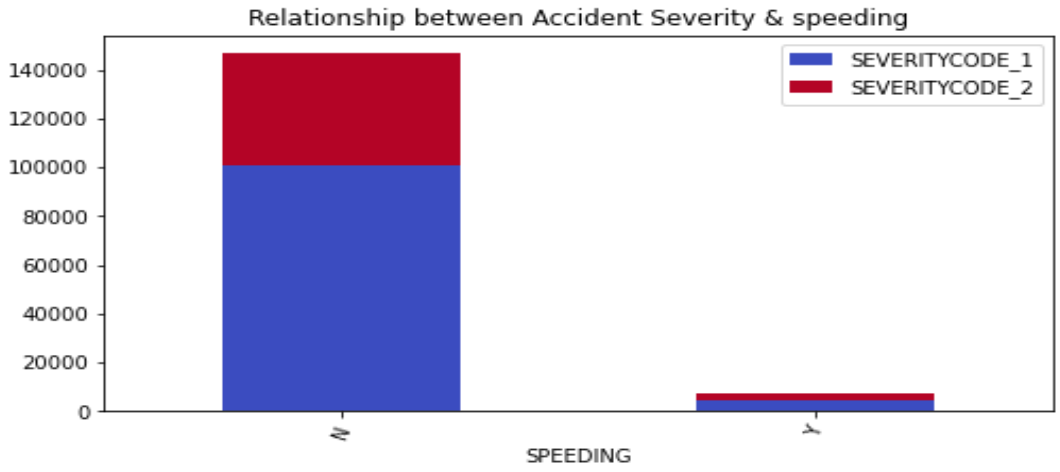


Exploring the relationship between light condition and severitycode: This graph gives us another reason to believe that light is contributing to the accidents, as we can see more

accidents happen in the daylight and the severity is also high. Second high variable, dark-street lights on also adds to the fact that light is the particularly important cause for the accidents.

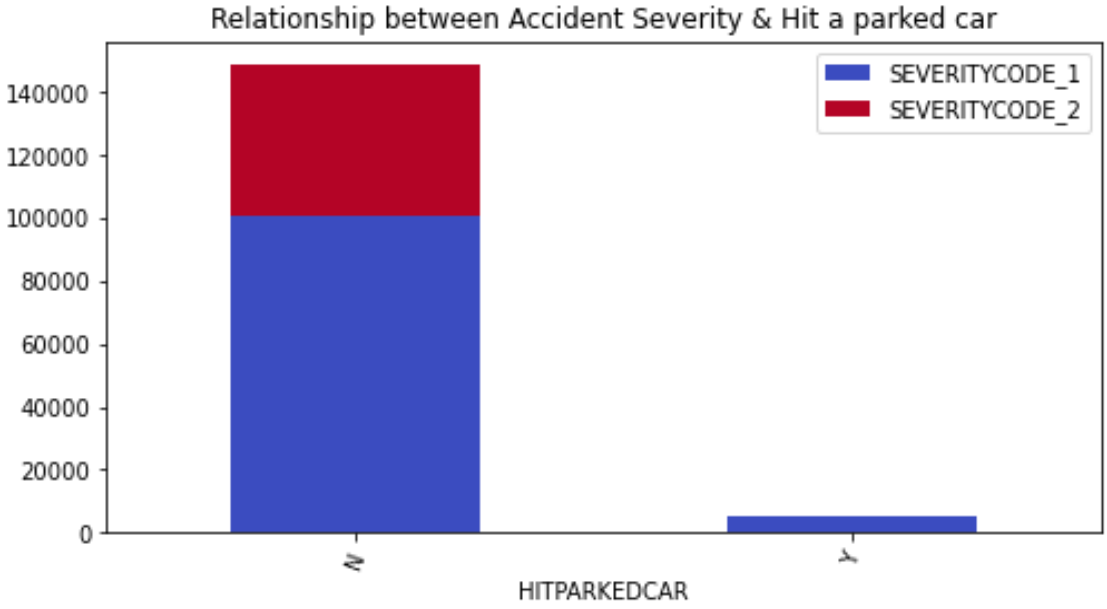


Exploring the relationship between speeding and severitycode: Not many accidents have occurred due to speeding, and the severity is also low.

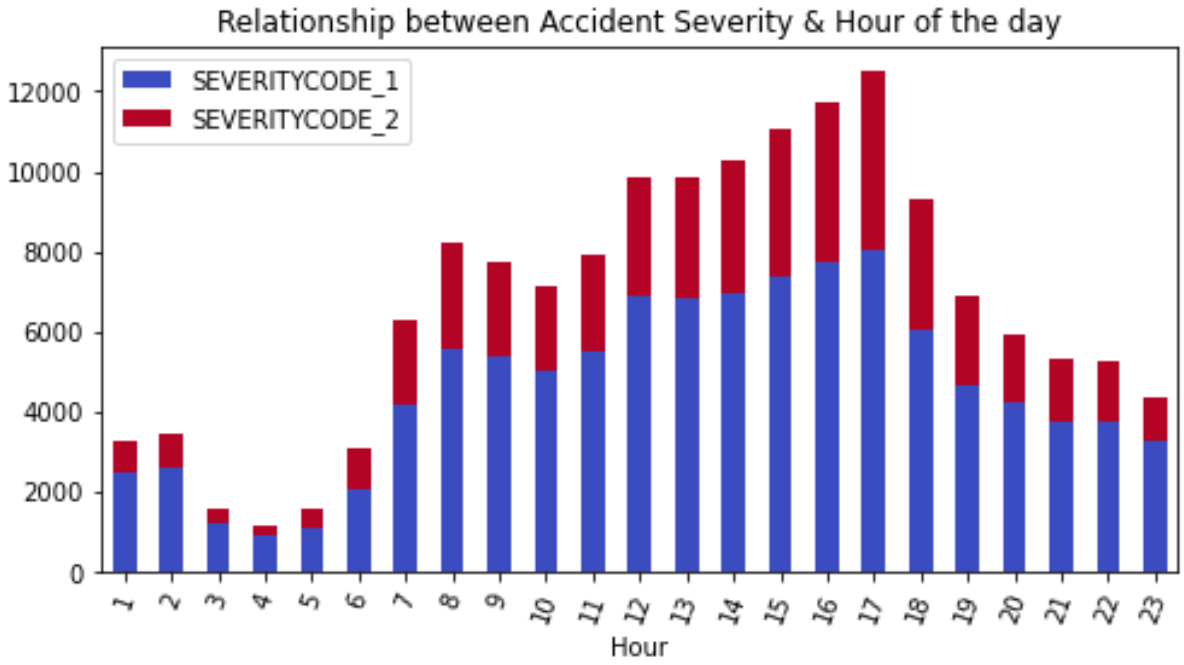


Exploring the relationship between whether the car is parked or not and severitycode: Not many accidents have occurred when the car is parked, and the severity is also low. Majority accidents happened when the cars are in motion, so another reason to suggest cars should

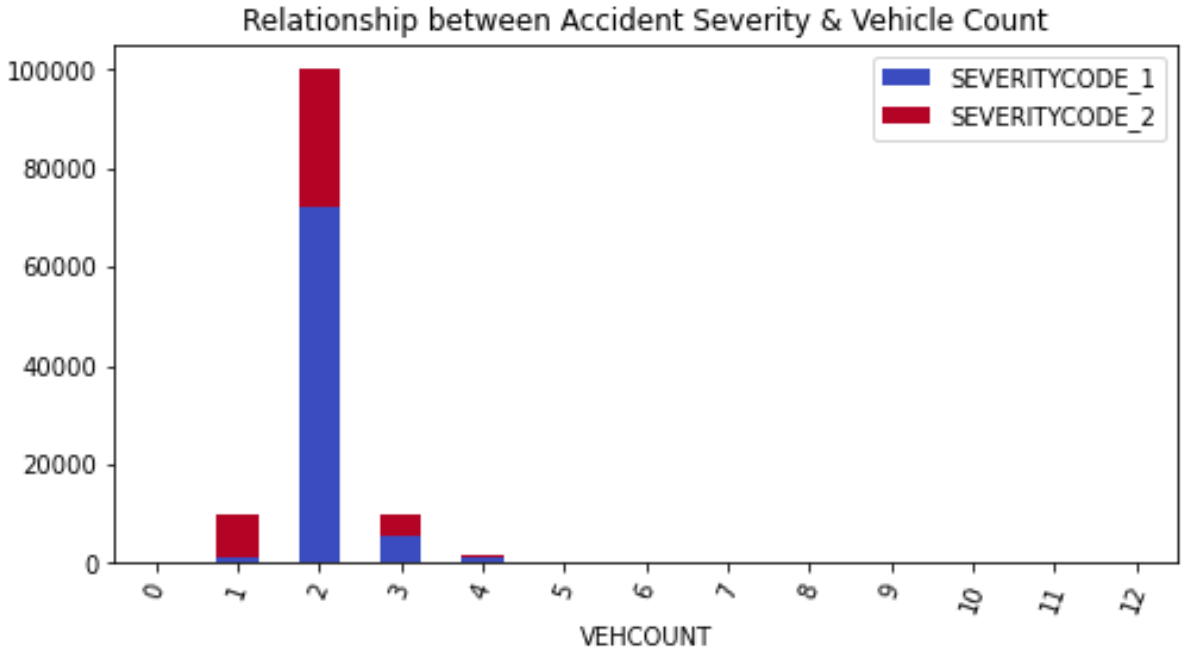
be designed with safety as number one priority and strict enforcement will traffic rules will help in reducing the accidents.



Exploring the relationship between hour of the day and severitycode: The graph tells us that, accidents happen between 7AM to 7PM and the peaks can be observed during the rush hour 2PM to 5PM.



Relationship between number of vehicles involved and severitycode: It appears that the accident severity is high when two to three vehicles are involved in collision.



Discussion:

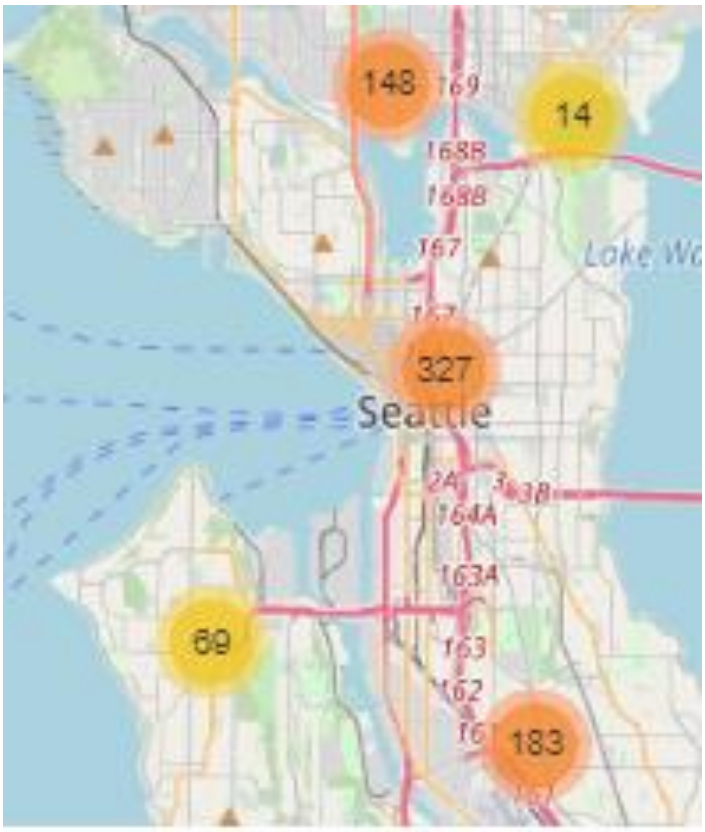
- From the above graphs most accidents happen during the daytime with clear light conditions and when the weather is dry. Also, accident severity graph peak starts from 7AM and 7PM, supports the argument that light plays a significant role in accidents and it's severity.

- Results from the model evaluation are as follows:

```
Confusion matrix of Logistic Regression:
LogisticRegression()
Accuracy score of logistic regression is  0.7246654699738904
Confusion matrix of K-NN classifier:
[[13035  2992]
 [ 4600  3885]]
Accuracy score of K-NN classifier is  0.6902741514360313
Confusion matrix of Decision Tree classifier:
[[13035  2992]
 [ 4600  3885]]
Accuracy score of Decision Tree classifier is:  0.6902741514360313
```

With respect to accuracy score from above, we can conclude that logistic regression model will be a good model to predict the accident severity for our data.

- To conclude, let us look at the map created using the X and Y co-ordinates from the dataset, we can see that a lot of accidents happen around the centre of the city.



- This project is an attempt to show that the occurrence of an accident and its severity can be predicted based on number of factors such as environmental factors, driver behaviour, traffic conditions and location. The results from the model predictions prove that the information from the data can be used for the prediction of future accidents with a given set of conditions.

Conclusion:

Comparing the results from various models, logistic regression model is a better performer in predicting the accident severity. This work also proves that historic data can used to train the models and predict future events.

Future work:

- The trained models from this work can be deployed onto various traffic monitoring applications used by police, health and emergency response teams and also can be used to develop apps to alter drivers on road and weather conditions to predict accident severity with a given set of conditions. Also, car manufacturing companies can use these data to develop car models with more safety options.