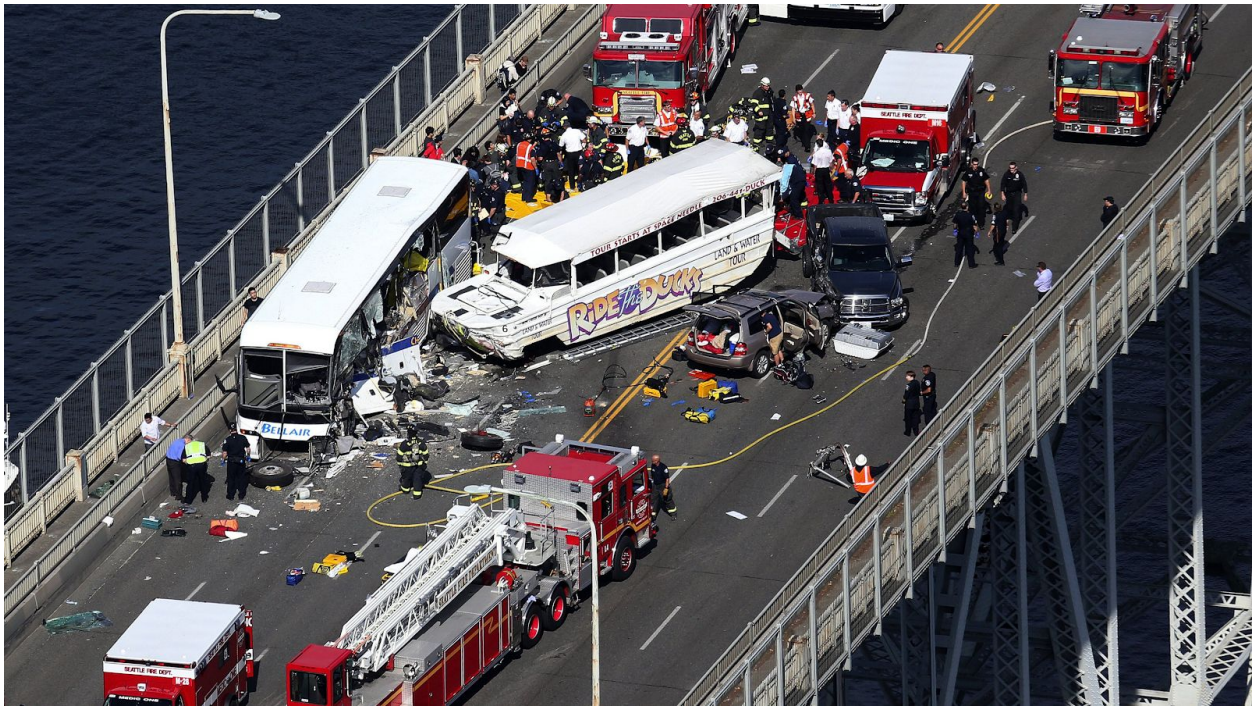
An Investigation on Road Accident And Severity in Seattle



HARSHIT JAIN

DATA SCIENCE AND PROFESSIONAL SPECIALISATION
IBM

Table Of Contents:

Introduction

Seattle is the largest city in the state of Washington and consists of roughly 710000 residents.The bestway to travel within the city is by car,and therefore evaluating the likelihood and severity of road accidents is of great importance. Not only for the safety of the civilians, but also for the economy. Reducing road accidents, improves the safety of the residents, reduces potential time off work for busy commuters and less demand for road repairs. This will allows for less demand for paid sick leave and for employees to continue working. The most common job type in the area, as of 2018, is a software developer which receives an average income of $91000.This Overtook Retail Salesperson 2016,which has average salary of $28000. This jump in average wage, shows the necessity for preserving and furthering the economy.

Business Problem

The Seattle government is going to prevent avoidable car accidents by employing methods that alert drivers, health system, and police to remind them to be more careful in critical situations.

In most cases, not paying enough attention during driving, abusing drugs and alcohol or driving at very high speed are the main causes of occurring accidents that can be prevented by enacting harsher regulations. Besides the aforementioned reasons, weather, visibility, or road conditions are the major uncontrollable factors that can be prevented by revealing hidden patterns in the data and announcing warning to the local government, police and drivers on the targeted roads.

The target audience of the project is local Seattle government, police, rescue groups, and last but not least, car insurance institutes. The model and its results are going to provide some

advice for the target audience to make insightful decisions for reducing the number of accidents and injuries for the city.

Data

The data was collected by 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 from 0 to 4.

Severity codes are as follows:

0: Little to no Probability	(Clear Conditions)
1:Very Low Probability	Chance or Property Damage
2: Low Probability	Chance of Injury

3: Mild Probability	Chance of Serious Injury
4: High Probability	Chance of Fatality

Attribute	Data type, length	Description
OBJECTID	ObjectID	ESRI unique identifier
SHAPE	Geometry	ESRI geometry field
INCKEY	Long	A unique key for the incident
COLDETKEY	Long	Secondary key for the incident
ADDRTYPE	Text, 12	Collision address type: • Alley • Block • Intersection
INTKEY	Double	Key that corresponds to the intersection associated with a collision
LOCATION	Text, 255	Description of the general location of the collision
EXCEPTRSNCODE	Text, 10	
EXCEPTRSNDESC	Text, 300	
SEVERITYCODE	Text	A code that corresponds to the severity of the collision: • 3—fatality • 2b —serious injury • 2—injury • 1—prop damage • 0—unknown
SEVERITYDESC	Text	A detailed description of the severity of the collision
COLLISIONTYPE	Text	COLLISION TYPE

PERSONCOUNT	Double	The total number of people involved in the collision
PEDCOUNT	Double	The number of pedestrians involved in the collision. This is entered by the state.
PEDCYLCOUNT	Double	The number of bicycles involved in the collision. This is entered by the state.
VEHCOUNT	Double	The number of vehicles involved in the collision. This is entered by the state
INJURIES	Double	The number of total injuries in the collision. This is entered by the state.
SERIOUSINJURIES	Double	The number of serious injuries in the collision. This is entered by the state.
FATALITI ES	Double	The number of fatalities in the collision. This is entered by the state

INCDATE	Date	The date of the incident
INCDTTM	Text	The date and time of the incident
JUNCTIONTYPE	Text	The date and time of the incident.
SDOT_COLCODE	Text	A code given to the collision by SDOT.
SDOT_COLDESC	Text	Text, 300 A description of the collision corresponding to the collision code.
INATTENTIONIND	Text	Whether or not collision was due to inattention. (Y/N)
UNDERINFL	Text	Whether or not a driver involved was under the influence of drugs or al

		Whether or not a driver involved was under the influence of drugs or alcohol.
WEATHER	Text	A description of the weather conditions during the time of the collision.
ROADCOND	Text	The condition of the road during the collision.
LIGHTCOND	Text	The light conditions during the collision.
PEDROWNOTGRNT	Text	Whether or not the pedestrian right of way was not granted. (Y/N)
SDOTCOLNUM	Text	A number given to the collision by SDOT.
SPEEDING	Text	Whether or not speeding was a factor in the collision. (Y/N)
ST_COLCODE	Text	A code provided by the state that describes the collision. For more information about these codes, please see the State Collision Code Dictionary .
SEGLANEKEY	Text	A key for the lane segment in which the collision occurred.
CROSSWALKKEY	Long	A key for the crosswalk at which the collision occurred.
HITPARKEDCAR	Long	Whether or not the collision involved hitting a parked car. (Y/N)

Furthermore, because of the existence of null values in some records, the data needs to be preprocessed before any further processing.

Data Cleaning

The first quality to check is to make sure there is at least one attribute to use as a universal identifier. This attribute needs to comprise of unique values of length comparable to the amount of rows in the dataframe. a Boolean check was performed on OBJECTID to verify this. The other unique attributes (INCKEY, COLDETKEY, REPORTNO, INTKEY) can be redacted from the dataframe as some have missing values or have different types of formats within the column.

Since this report will focus predominantly on the date and time of the collision, the date and time format will need to be altered. The dataframe itself has fairly inconsistent values for these columns.

The datetime (INCDTTM) attribute was split into two separate columns, and then put into d/m/Y and H/M/S string format. The year could then be easily extracted by taking the last four letters. The date (INCDATE) attribute was cleaned by using the pandas function strftime to split the date into day, month, year and day of year. A 00:00 time appears for cells with date and no time after applying string format to INCDTTM. This was created by first checking how many cells had incidents at midnight (this turned out to be zero), then 00:00 was replaced with an undefined value (NaN). For the purpose of consistent dataframes, the rows with NaN values were removed. This was

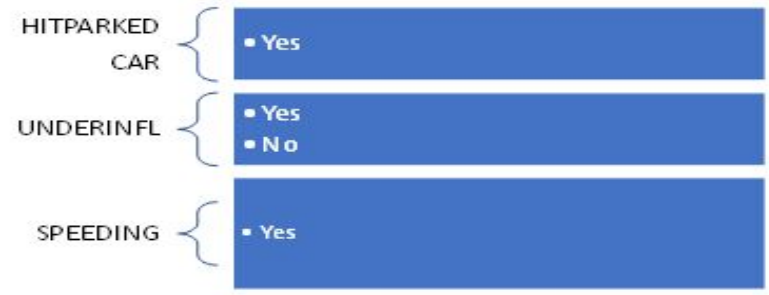
around 2% of the dataset. The hour was extracted from the time the same way the year was extracted from date.

One Hot Encoding

Before using classification methods to predict the outcome of severity, the attributes were organised into numerical form. The attributes being

WEATHER	A description of the weather conditions during the time of the collision.
ROADCOND	The condition of the road during the collision.
LIGHTCOND	The light conditions during the collision.
SPEEDING	Whether or not speeding was a factor in the collision. (Y/N)
UNDERINFL	Whether or not a driver involved was under the influence of drugs or al Whether or not a driver involved was under the influence of drugs or alcohol
HITPARKEDCAR	Whether or not the collision involved hitting a parked car. (Y/N)

Firstly, the a new dataframe is created merging day number and month number from the previous date dataframe. All rows with 'Unknown' or 'Other' values were removed to keep all data entries consistent. This resulted in 7% drop in data. Again, 'OBJECTID' was kept as a consistent attribute. Many of the attributes have multiple unique values, some of which are repeated. In order to employ the best possible outcome of severity but to also keep runtime fairly low, these values have been grouped as follows One hot encoding will take these categorical variables and transform them into numerical ones under binary variables.


HITPARKED CAR
• Yes
UNDERINFL
• Yes
• No
SPEEDING
• Yes


Dark
• Dark – No Street Lights
• Dark – Street Lights On
• Dark – Street Lights Off
• Dark – Unknown Lighting
Dawn
• Dawn
• Dusk
Daylight
• Daylight
Dark
• Dark – No Street Lights
• Dark – Street Lights On
• Dark – Street Lights Off
• Dark – Unknown Lighting
Dawn
• Dawn
• Dusk
Daylight
• Daylight
Wet
• Wet
• Standing Water
• Snow/Slush
• Oil
• Sand/Mud/Dirt
• Ice
Dry
• Dry

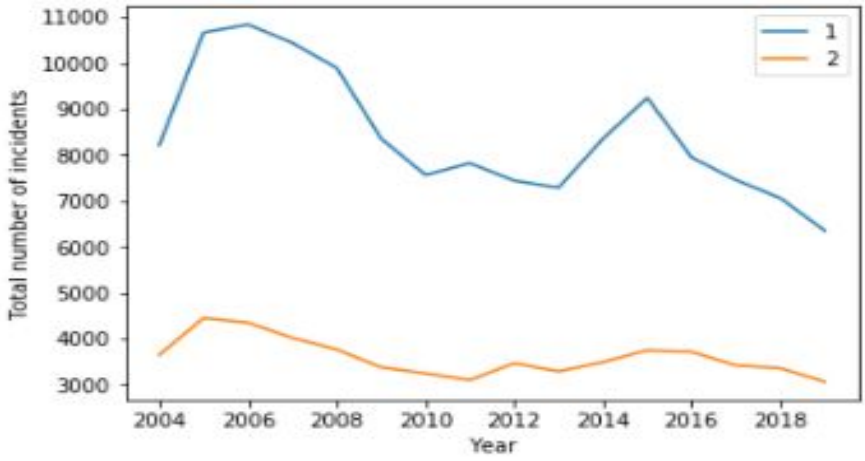
Data Analysis

The first comparison made is the severity against the new date attributes. The severity code obeys the following,

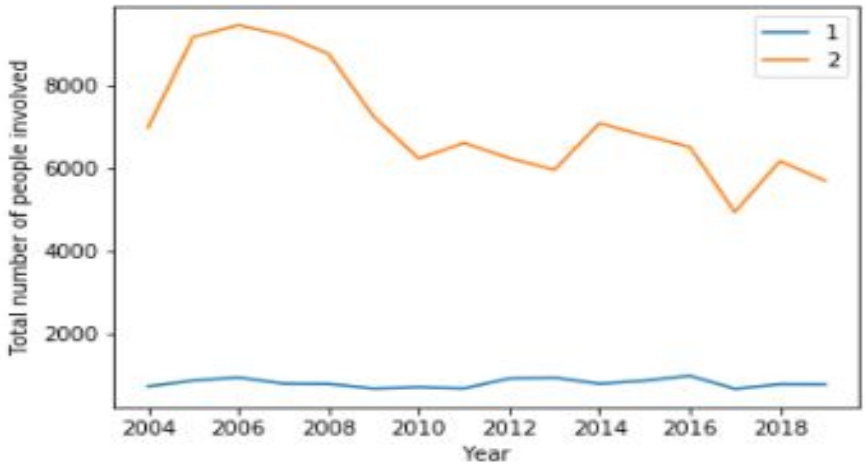
- 3 - Fatalilty
- 2b - Serious injury
- 2 - Injury
- 1 - Prop damage
- 0 - Unknown

The column was analysed to see the count of each severity. The only severity codes within the dateset are 1 and 2. This allowed for optimising the function of the code and improve the runtime.

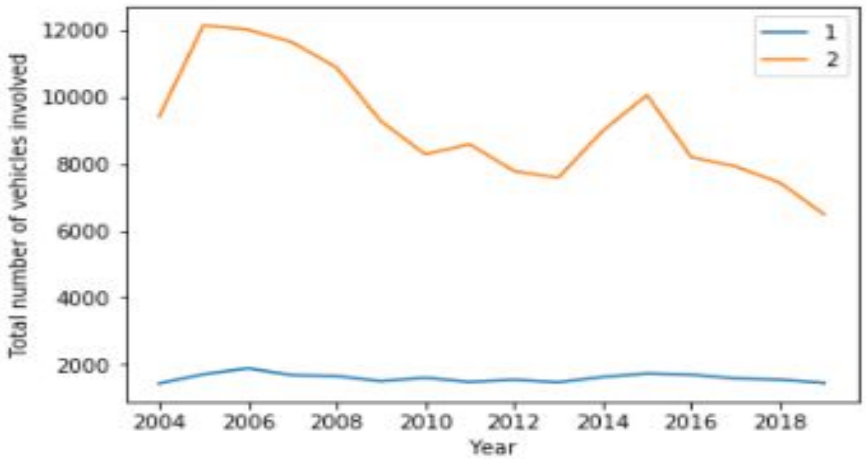
Total per Year



(a) Total incidents per year



(b) Total number of people involved in incidents per year



(c) Total number of vehicles involved in incidents per year

First, let's look at the total number of incidents across the years. This can be seen in 3.1a.Thisshowsapeakat2006butthencontinuestodecreasebeyondthispoint. The number of incidents have effectively halved across the 16 years (neglecting the peak at 2015). This figure also shows that incidents including injury are a lot less likely than those of property damage. Suggesting that injuries are very minimal in Seattle, and decreasing. Equally, the total amount of vehicles involved peaked at the same time (as expected), and is also decreasing. This will be important to note, as the weather conditions will affect the amount of vehicles damaged. An interesting observation is the amount of vehicles involved in incidents is more than that of people involved, suggesting there are a noticeable amount of stationary, empty vehicles involved in collisions. An example being a car driving at night in January on black ice, could swerve and hit multiple parked cars.

Total per Month

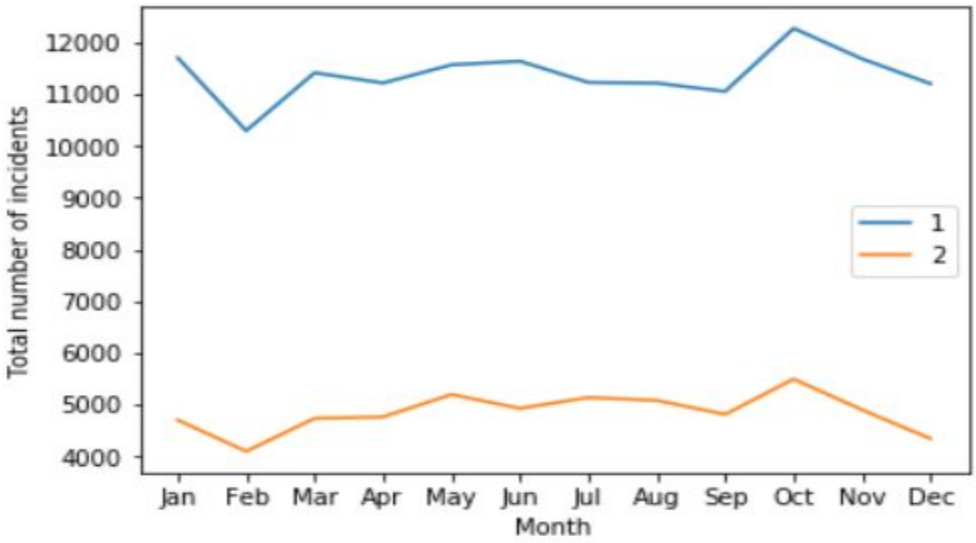


Fig. 3.2 Graph showing incidents across each month

Having analysed the total amounts across the year, it's important to also analyse incidents across the months. From 3.2 we can see that the amount of incidents reported are fairly consistent across the months. The drop in incidents occurs in February, potentially due to fewer drivers on the road. A spike in property damage
8
incidents occurs in October, which could be due to weather, similar to January. Otherwise, in the Spring/Summer/Autumn months, the number of incidents are consistent.

Total Per Day

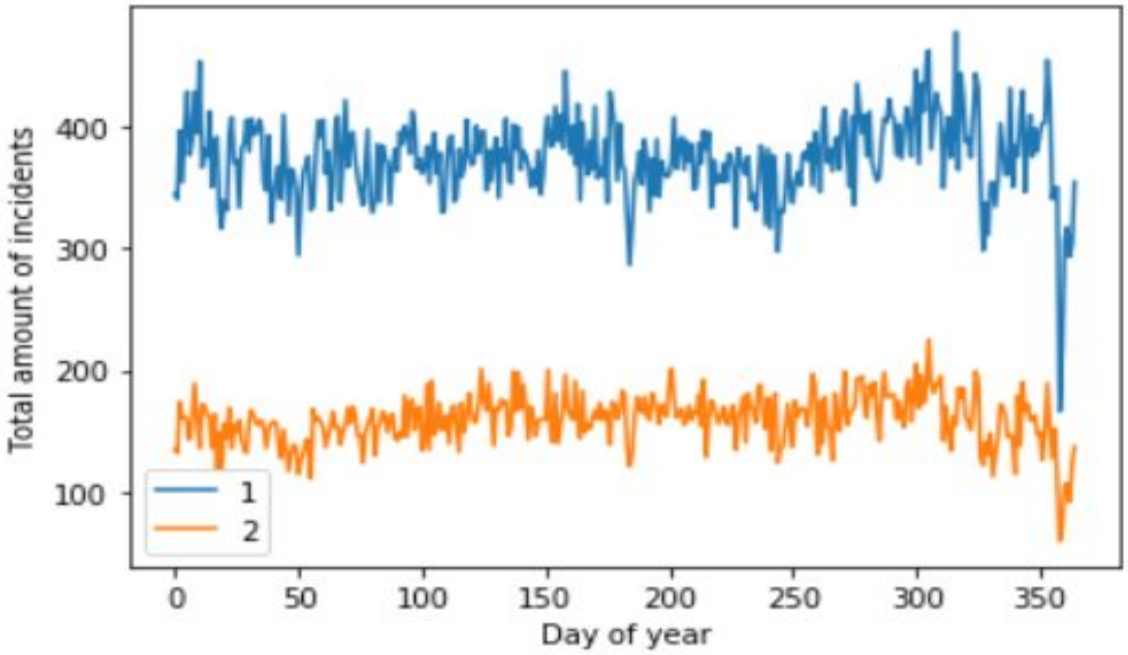
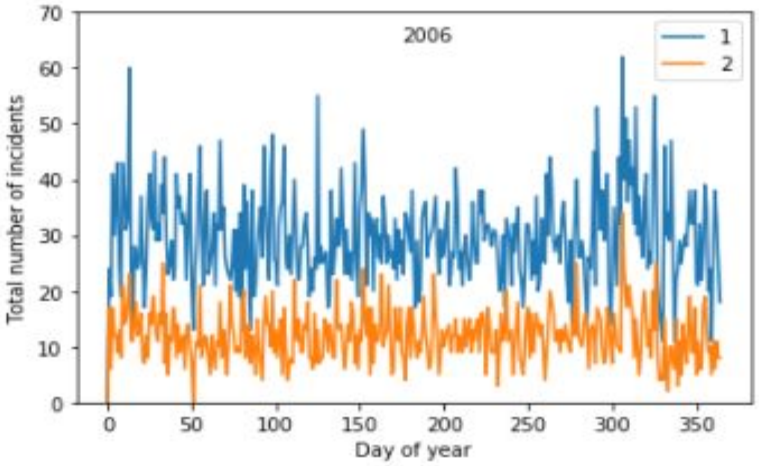


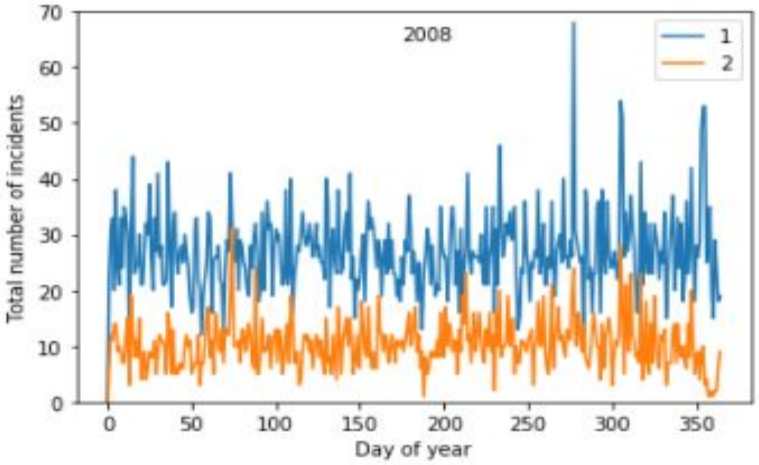
Fig. 3.3 Graph showing incidents across each day of the year

As can be seen in 3.3,property damage incidents occur twice as often as injuries. What is noticeable is that Christmas day has the least amount of incidents, which is expected as far fewer people travel on that day. It is noticeable also that days with a drop in incidents occur around public holidays [11]. What is noticeable across 3.4 is that although the number of incidents involving property damage reduces across the year, the amount of injuries stays fairly con- sistent. This could be due to people

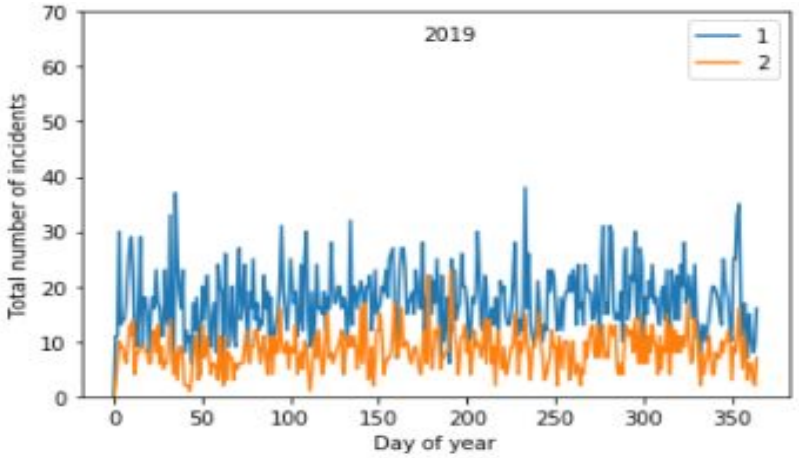
being more careful on the road to avoid cars being damaged. 2008 had the highest number of incidents per day as shown in 3.4b. What was seen in the data however that after this period, the incidents decreased dramatically. 3.4c has dropped to have more consistent amounts of incidents.



(a) Total incidents per day for year 2006 which had the highest number of incidents.



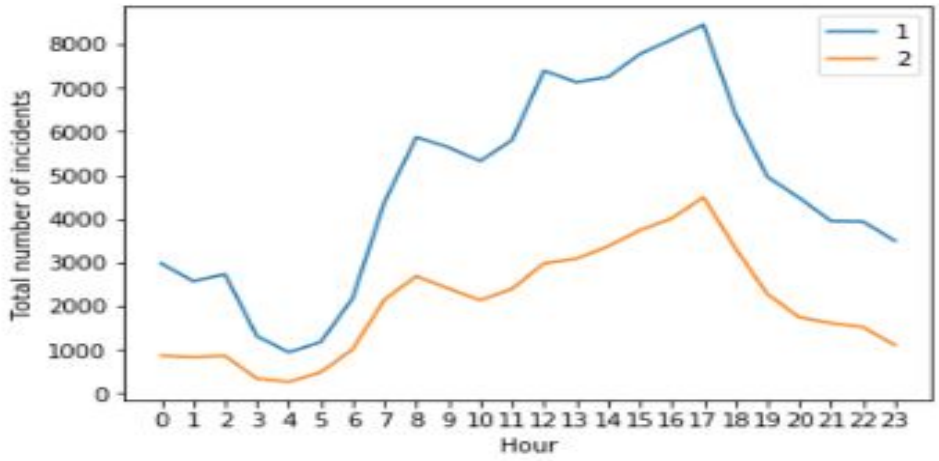
(b) Total incidents per day for year 2008 which had the day with the highest incidents.



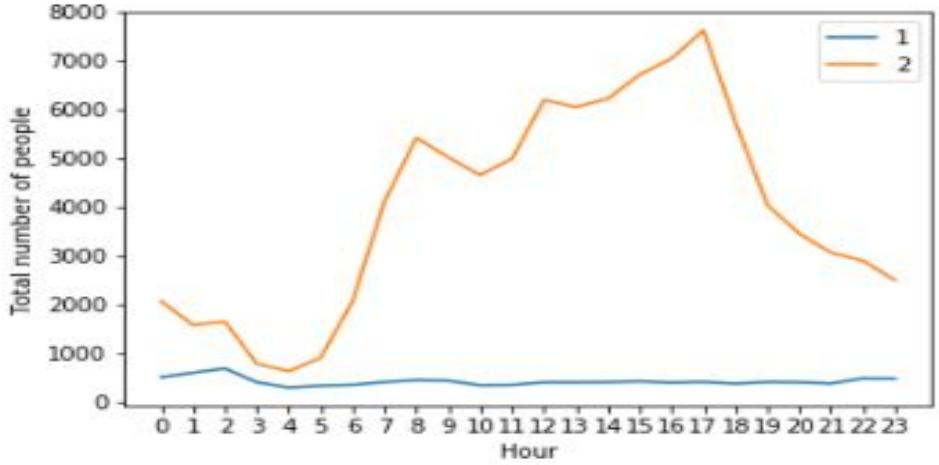
(c) Total incidents per day for year 2019 which had the lowest incidents.

Total Per Hour

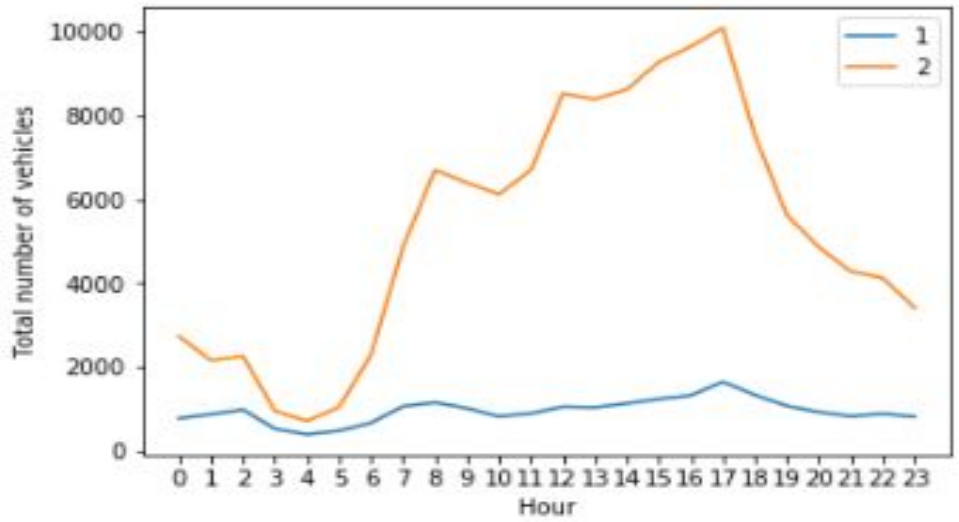
Finally, let's look at the total number of incidents across time. As can be seen in 3.5a,theamountofincidentsdropsbetween2amand 6am.This Expected As most people will be sleeping during these hours. The incidents peak at 5pm, which would be rush hour. This suggests that travelling later in the day could reduce the likelihood of an incident occurring. In 3.5b and 3.5c,property damage has no correlation to time. Otherwise, these graphs suggest that there is a correlation in number of incidents and number of people involved at peak times.



(a) Total incidents per hour.



(b) Total number of people involved in incidents per hour.



Predictive Modelling

The severity of an accident could take two values, either a 1 (less severe) or a 2 (more severe). In this sense it is a binary classification problem. Importantly, there was a large majority of data labelled with a 1 (75%).
As data contain outliers so by using Robust Scaler data can remove outliers and process and scale the data according to outliers.

All data used is labelled, so the data was split into a test/train model of ratio 1:5. The models I chose were k-Nearest Neighbours, SGD classifier, Logistic Regression and Random forest
The accuracy was measured through:
• Accuracy - Measures accuracy of predicted values and true values
• F1 - score - Measures accuracy based on true/false positives and true/false negatives

K-Nearest Neighbours

K-Nearest Neighbours classifies a new data point, based on the majority vote of the classes of it's nearest neighbours. The K value describes how many nearest neighbours should be taken into account in order to classify a new data point. The training process involves simply finding what value of K allows for the highest classification score on the training data. This model is

then used on the test data and the Accuracy and F1-Scores are calculated..One sees that the score significantly increases at even numbers of nearest neighbours. This highlights an issue with the way the experiment was carried out, which is that when K is an even number, there may not be a majority vote (the nearest neighbour's may be labelled with equal numbers of 1's and 2's). In this case, the function that was used to carry out this research automatically predicted that the new data point had a label of 1. This improved accuracy due to the fact that the majority of the data was labelled with a 1. The best accuracy was found at k=14.

Random Forest

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of overfitting to their training set. Random forests generally outperform decision trees, but their accuracy is lower than gradient boosted trees. However, data characteristics can affect their performance.

SGD Classifer

Stochastic Gradient Descent (SGD) is a simple yet efficient optimization algorithm used to find the values of parameters/coefficients of functions that minimize a cost

function. In other words, it is used for discriminative learning of linear classifiers under convex loss functions such as SVM and Logistic regression. It has been successfully applied to large-scale datasets because the update to the coefficients is performed for each training instance, rather than at the end of instances.

Logistic Regression

Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist. In regression analysis, logistic regression (or logit regression) is estimating the parameters of a logistic model (a form of binary regression).

Evaluation

Algorithm	Accuracy	f1_scores
SGDClassifier	0.715297	0.680553
Logistic Regression	0.724806	0.668367
KNN	0.719184	0.673489
Random Forest	0.727894	0.673298

Random forest has the highest accuracy in all models.

Conclusion

This report has explored the severity of incidents in Seattle. It has looked at the total number of incidents, people involved and vehicles involved as part of data analysis. Then it has used the weather and speeding, abusing drugs to see whether the type of incident can be predicted. The results weren't definitive but did suggest overfitting and therefore the models do need to be looked at again.

Future Improvements

As Data cannot contain various severity variables like 0,2b ,3 which give a prediction error so data need to be modified to get a higher accuracy and crosswalk key is used to get a better accuracy .