Us Accident Data Case Study

Abbas Hussain

CS-989 Big Data Fundamentals

Contents

List of Figures

Chapter-1

Introduction

An accident is defined as an unplanned, undesirable event that happens unexpectedly and unintentionally, without anyone's fault or negligence. Accidents are unintended but foreseeable consequences leading to losses or harm. An accident can fall into many different categories however the category we are focusing on is road accidents. Road accidents can be referred to any collision involving a motor vehicle occurring on a roadway, public thoroughfare, private road open to public traffic or any publicly accessible private parking lot.

Automobile transportation has now become a part of everyone’s life. Given the astonishingly high rates of catastrophic accidents and deaths, auto-mobile improvement is unavoidable. According to the US National Safety Council, there are 8 million accidents every year, making it the greatest cause of mortality (National Safety Council, 2014). Every day, 12 people are killed in car accidents, and 2.7 million more are wounded. These astronomical figures may lead one to conclude that safety is not a top focus, yet they do not.

Accidents happen all the time, but no one expects them so suddenly. It's why they're called accidents. Everyone has an opinion on how accidents occur, but no one thinks about it until after the fact. Accidents and their consequences are a major problem in society today. To solve this matter, we can observe and analyse accident factors. By accident factors, I mean the external factors that lead to accidents. My goal with this report is to categorize what contributes to accidents, in order to provide insight into how accidents happen and gather the information that can be shown graphically to help break down the aspects that lead to these unfortunate situations. We can then use these insights to potentially prevent future accidents by addressing potential risk factors.

Chapter-2

Dataset

Data used is freely available on the Kaggle website. This dataset is nation-wide vehicle accident data, that spans 49 states in the United States. The accident data collected spans from over 1.5 million rows from February of 2016 to December of 2020 and is based on a variety of open-source data made available through various government portals.

The data folder for this repository contains a CSV, which contains details of each accident. Some columns and lines of the file contains, the severity of the accident, where the accident took place, its start time, Temperature, city/state and a few more. This dataset can answer and focus on specific questions about car accidents in the USA, including when, where how many, which states are most prone to the most amount of crashes, and how it affects the vehicle insurance industry in the upcoming years to come.

By combining wealthy coincidence facts with interactive visualizations, customers can get a glimpse of what passed off at injuries throughout the country. Car coinciding styles and info are discovered within the exploration, which allows it to be useful for researchers, policymakers, media or every other business to make choices or plans based totally on evidence from the facts. I have done a direct visualization to help understand the accident seasonal pattern, accident accumulation process, month performance and what regions are more vulnerable to accidents. I have further executed supervised and unsupervised learning to find the severity of an accident.

Chapter-3

Analysis of Dataset

Exploratory data analysis is the process of getting graphical approaches and basic summary statistics to explore and describe a data collection and derive findings. Patterns, trends, and outliers are communicated significantly more rapidly using graphical representations of data than with tables of numbers and text. Users can discover flaws and problems that require attention at a glance and take necessary action via visualisation. Rather than formal hypothesis testing, the purpose is to get an understanding of the facts by visual study. The goal of this activity is to learn about the data's structure and substance so that hypotheses may be developed for subsequent testing.

The dataset consists of 1516064 rows and 47 columns. The first part of this analysis was getting an overview of the data and to see what different aspects which contribute directly to an accident. Some of the aspects that I feel is directly related to an accident is being discussed below.

Time of Accident

We used the pandas datetime function to convert the original string ‘Start_Time’ column to timestamp. Next, the Hour of the day was extracted using ‘functions.hour’ and aggregated to count the number of accidents that happened during each hour.

Starting with a general analysis to determine the time of most accidents occurring, the data were grouped according to the various timestamp of the accident and presented in a histogram showing the number of accidents that occurred at each time of day.

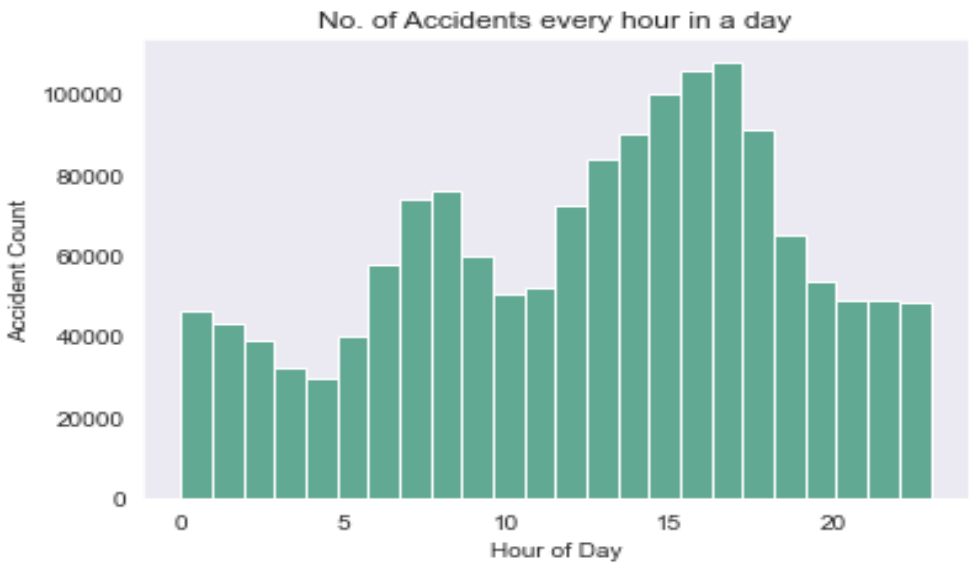

Fig 2.1 Distribution of accidents throughout the day

A higher percentage of accidents occur between 7 AM to 9 AM and 3 PM to 5 PM probably because people tend to be in a hurry to get to work and return from work. The safest time to travel can be depicted to be between 8 pm to 5 am as fewer accidents tend to happen during that time.

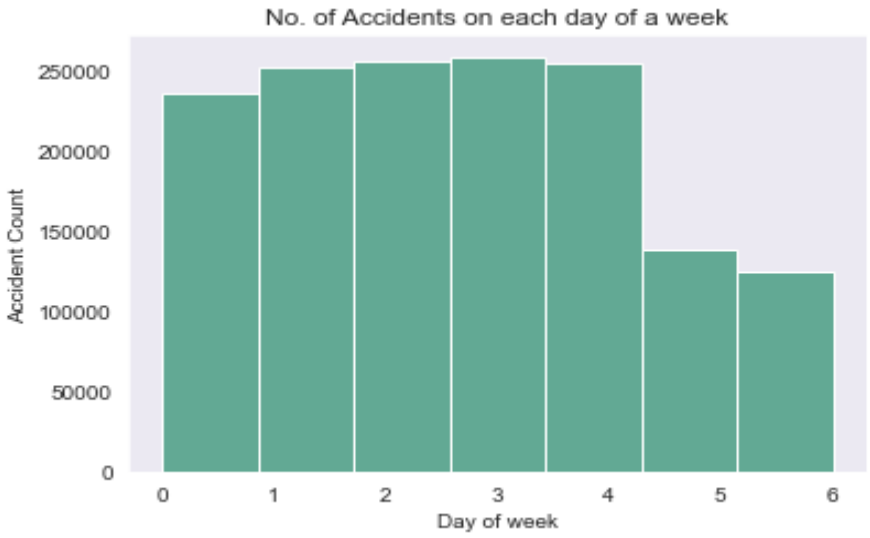

Fig 2.2 Distribution of accidents throughout the week

We used ‘functions. dayofweek’ to extract the ‘day’ on which each accident happened, aggregated to get the count for the number of accidents for each day. When looking at the data pattern in Fig 2.2 across the week, it's clear that there are a lot more accidents on weekdays, especially between Monday to Friday in the United States. On the contrary, on weekends, such as Saturday and Sunday, there are fewer accidents.

Further on, Figures 2.4 and 2.5 show how the number of accidents dispersed across weekends and weekdays, respectively, During weekends the graph reverses exceeds the count of an accident of that on a weekday. On weekends, the number of accidents is much lower, especially during peak hours, whereas on weekdays, the number of accidents is over three times higher.

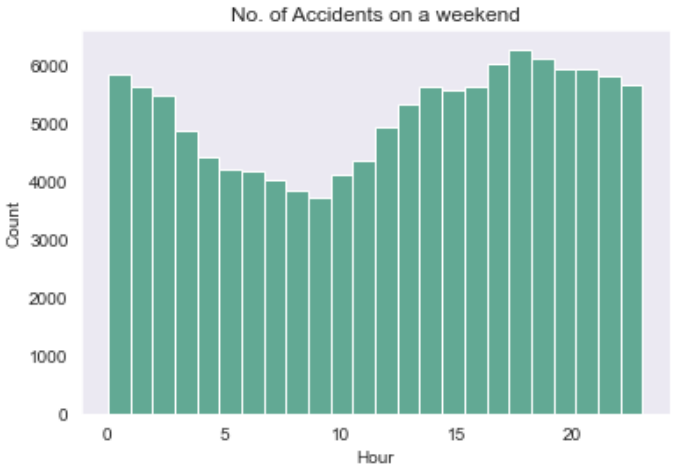

Fig 2.3 Distribution of accidents on a weekend

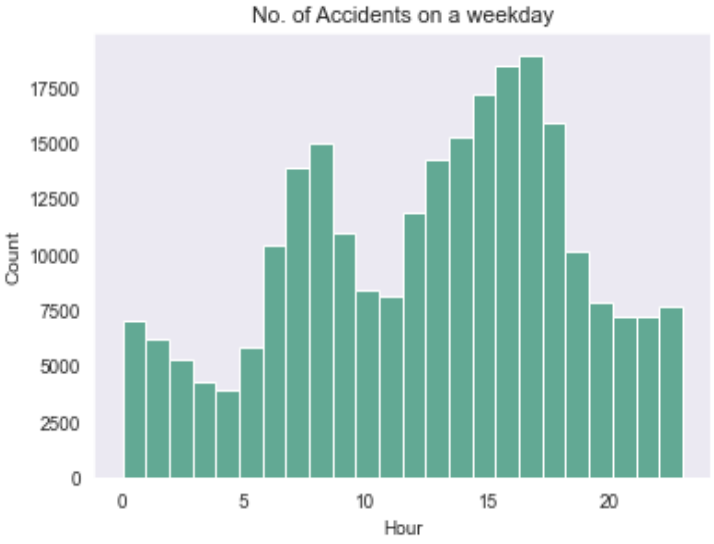

Fig 2.4 Distribution of accidents on a weekday

Accidents rise progressively from low points in January and February to a high point in October and December. In comparison to the second half of the year, the first half of the year had a lower number of accidents. Monthly accident rates have risen gradually from February's lows, culminating in the fourth quarter of the year.

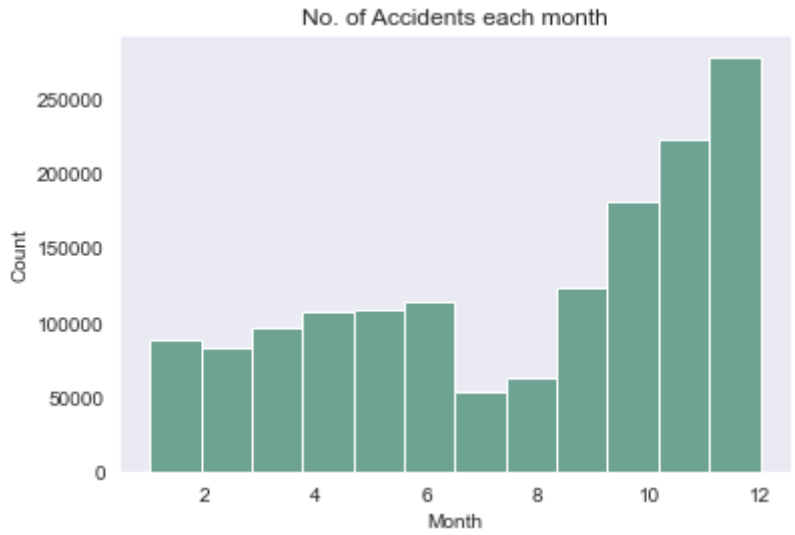

Fig 2.5 Distribution of accidents over each month

Weather impact on accidents

The temperature column contains the temperature during each time an accident happens. Since most accidents tend to happen during the end of the year as depicted in Fig2.6, the cold temperature might play a big role in that.

On plotting a graph, it shows that the accidents tend to happen in cold temperatures rather than warm temperatures. I have encoded all the temperatures less than 70 degrees Fahrenheit as cold climate.

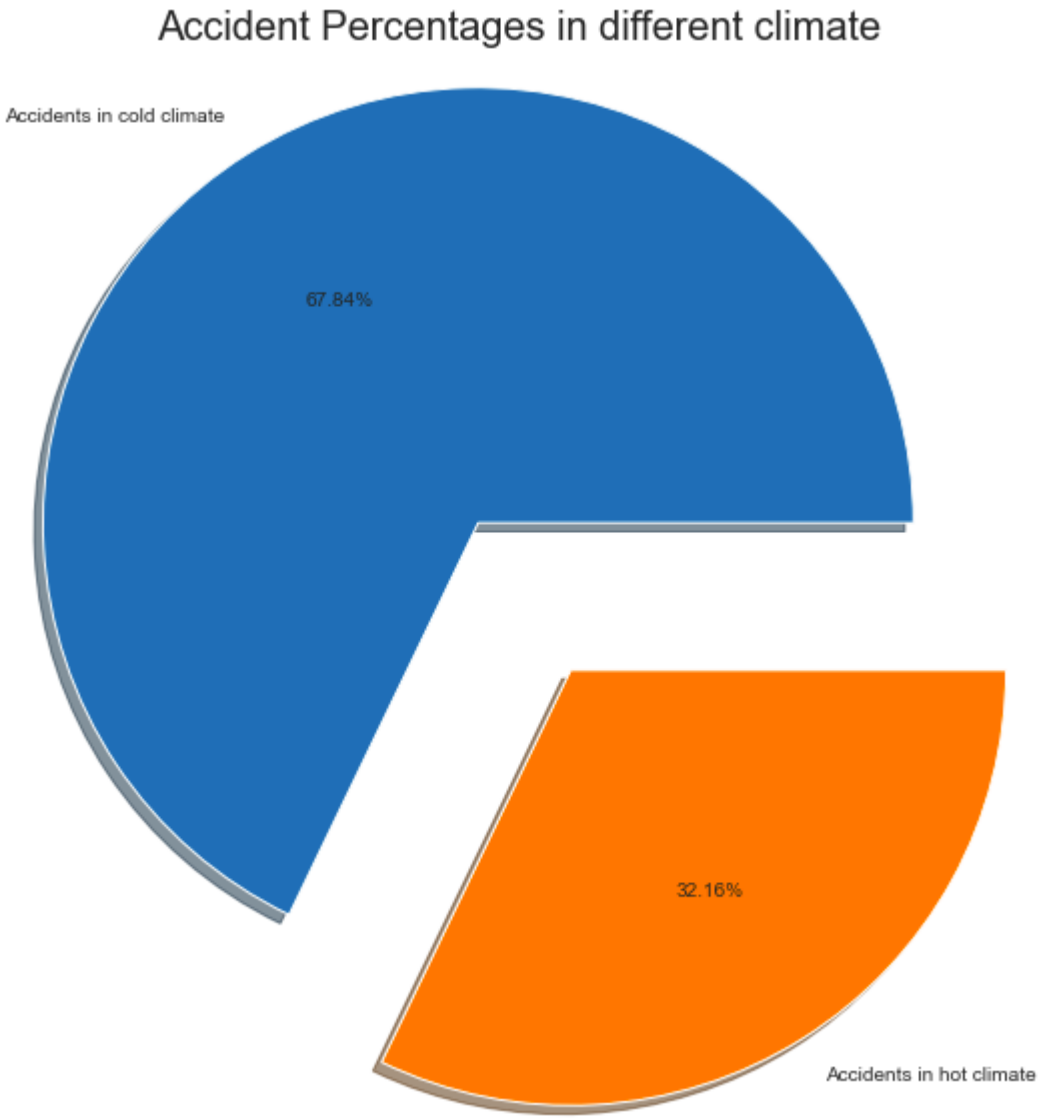

Fig 2.6 Distribution of accidents in different climate

The dataset contains 128 unique entries for the column 'Weather Condition,' as well as null values. We selected the top 12 weather conditions and averaged them to obtain the total number of accidents for each weather condition after filtering out the null values. In general, we assume that terrible weather will result in more accidents, with that we can also observe that when the weather is

clear, there are more accidents. This might be due to the fact that when the weather is terrible, people drive very carefully. The weather for the majority of the accidents was clear, followed by overcast and mostly cloudy, according to the map. In contrast to clear skies, overcast and mostly cloudy skies are tolerable causes for accidents, implying that meteorological conditions do not play a significant influence.

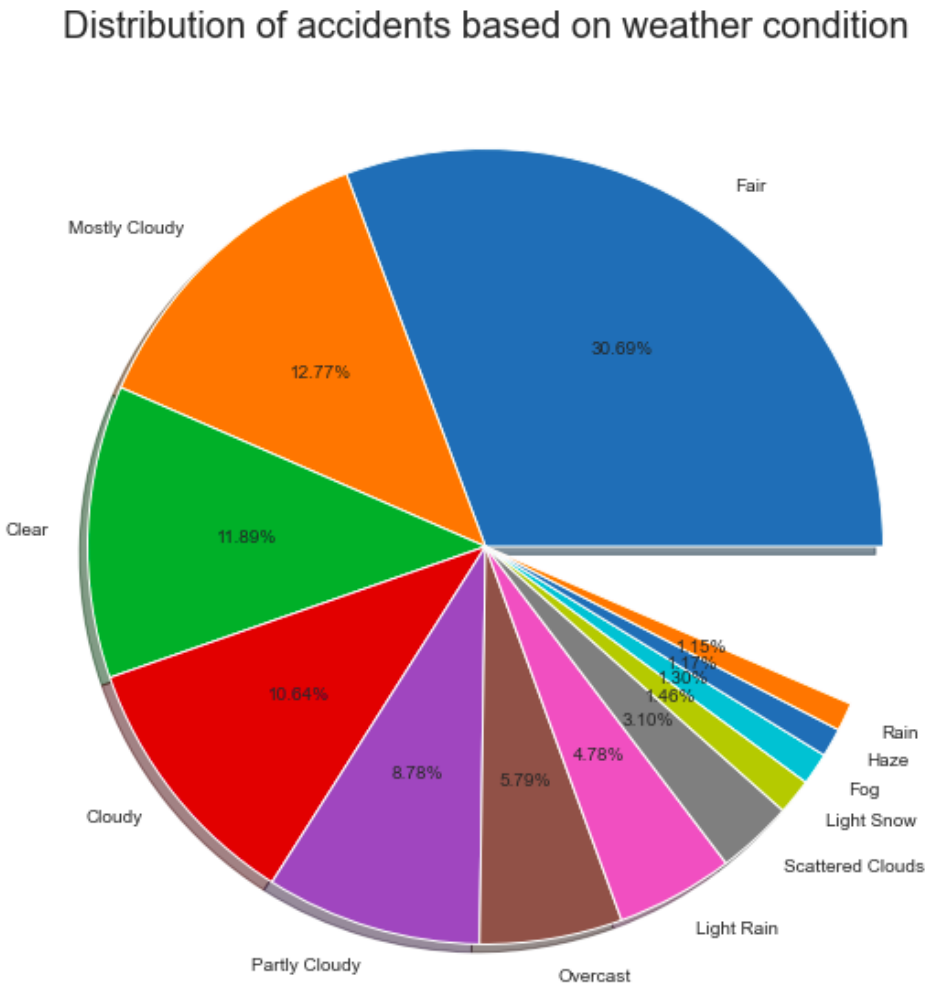

Fig 2.7 Distribution of accidents over weather condition

Fig 2.8 Heatmap correlation of accidents over weather condition

Location impact on accidents

The dataset contains 1516064 rows hence 1516064 accidents occurred in a given timeframe in the US. Fig 2.7 shows the distribution of accidents over the top twenty states with the highest number of accidents. Results showed that California had the highest number of accidents followed by Florida and Oregon amongst all the states in the country.

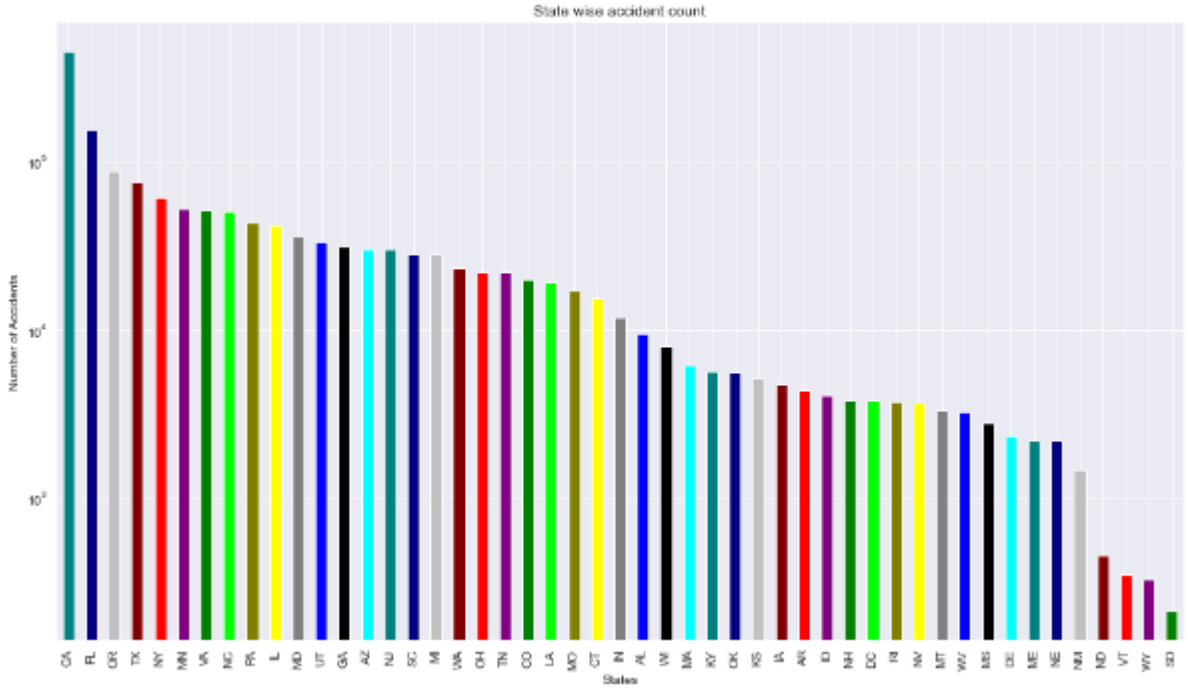

Fig 2.9 Distribution of accidents over states

To further our understanding the accidents time over every city were done and Los Angeles had the most accident followed by Miami. One of the most important factors to note is that New York doesn't come as the top accident-prone city in the USA. On further analysis of the dataset, it was found that New York City data was not present in the dataset. There is an exponential decrease in the number of accidents per city. About 4 per cent of the cities record more than 1000 accidents per year. Over 1300 cities have reported just 1 accident throughout the year.

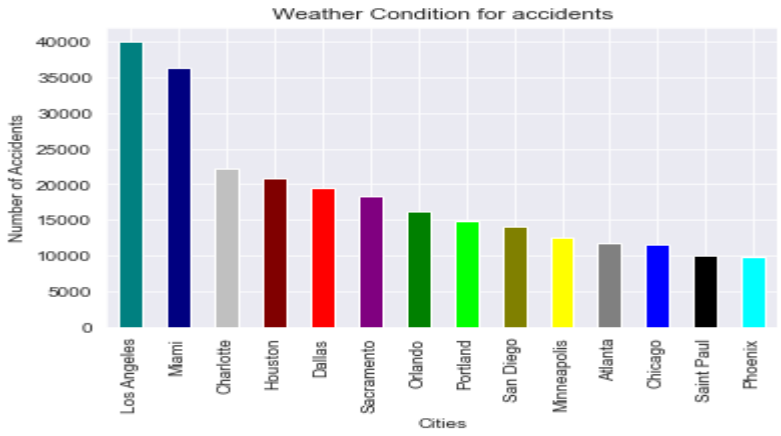

Fig 2.10 Distribution of accidents over cities

Accidents' proximity to traffic objects is also included in the dataset. The most common type of accident is Traffic Signal, followed by Junction, Crossing, Station, and Stop. As a result, states and their traffic cops should concentrate on their efforts in these locations to prevent accidents. If the issues are related to infrastructure issues, these must be addressed immediately.

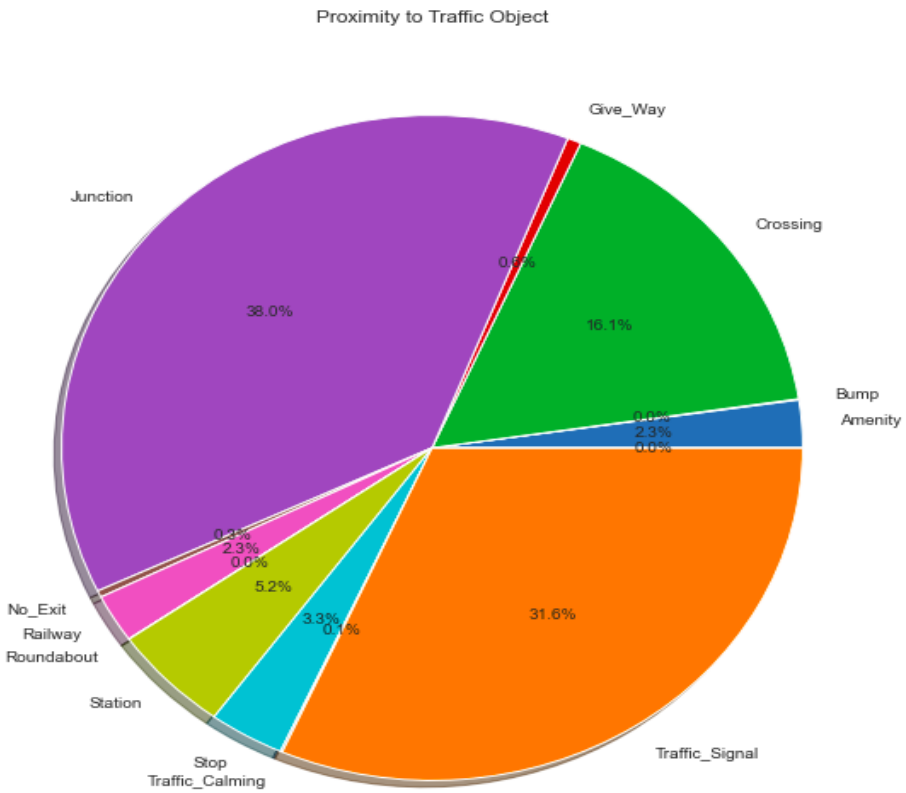

Fig 2.11 Distribution of accidents in proximity to Traffic objects

Severity of Accidents

The Severity is the most intriguing aspect of the dataset. The severity of the accident is determined by its impact. The graph below gives us details of the severity of each accident. Around 80% of the accidents are of average severity. It is followed by high severity which is around 10%. A very low percentage is formed by the very low severity(1) which is 1.86% altogether. As we can see the majority of the accidents had a severity of 2 (average) or 3 (above average), which is regrettable. There aren't many accidents that aren't fatal (0 and 1). Severity 3&4 have a significant influence. Accidents of severity 2 are more common than those of severity 3 and 4. Incidents of severity 3&4 are more likely to result in fatalities, while accidents with severity 2 are more likely to result in injuries.

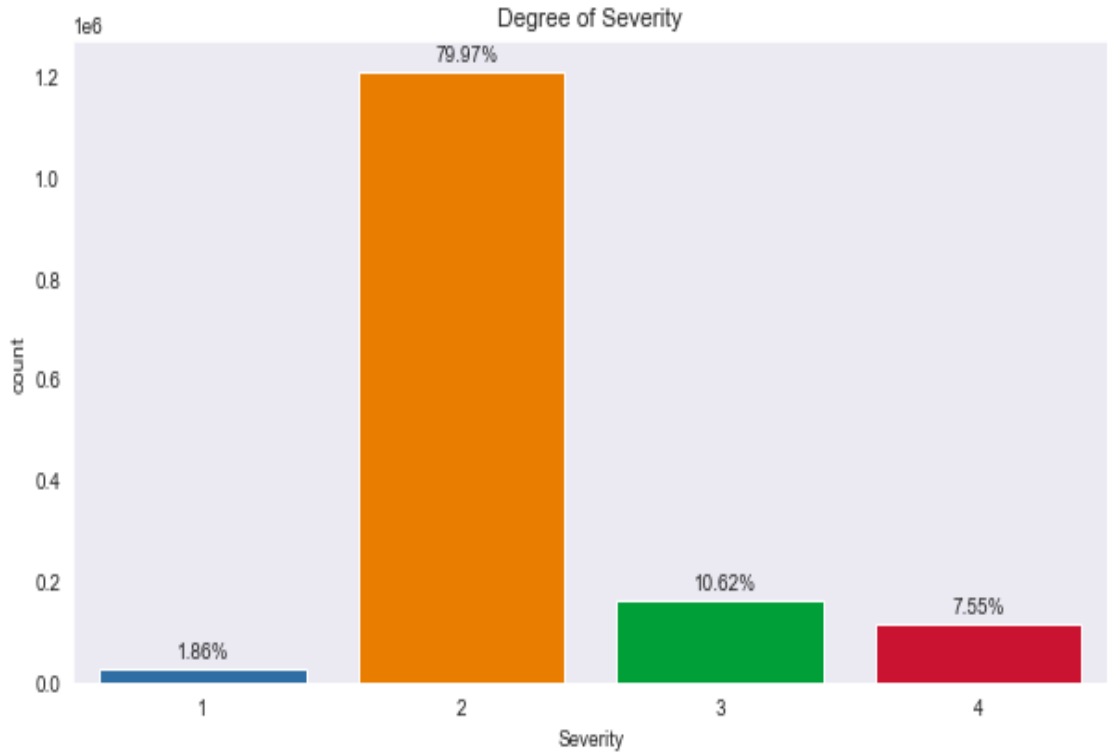

Fig 2.12 Severity of each accident

I have plotted the severity for each state individually. From the below plot, we can conclude that the accidents with the most severity occurred in Wyoming followed by Delaware and Colorado. The highest accident-prone states such as California and Florida which are pretty low with the overall severity of accidents.

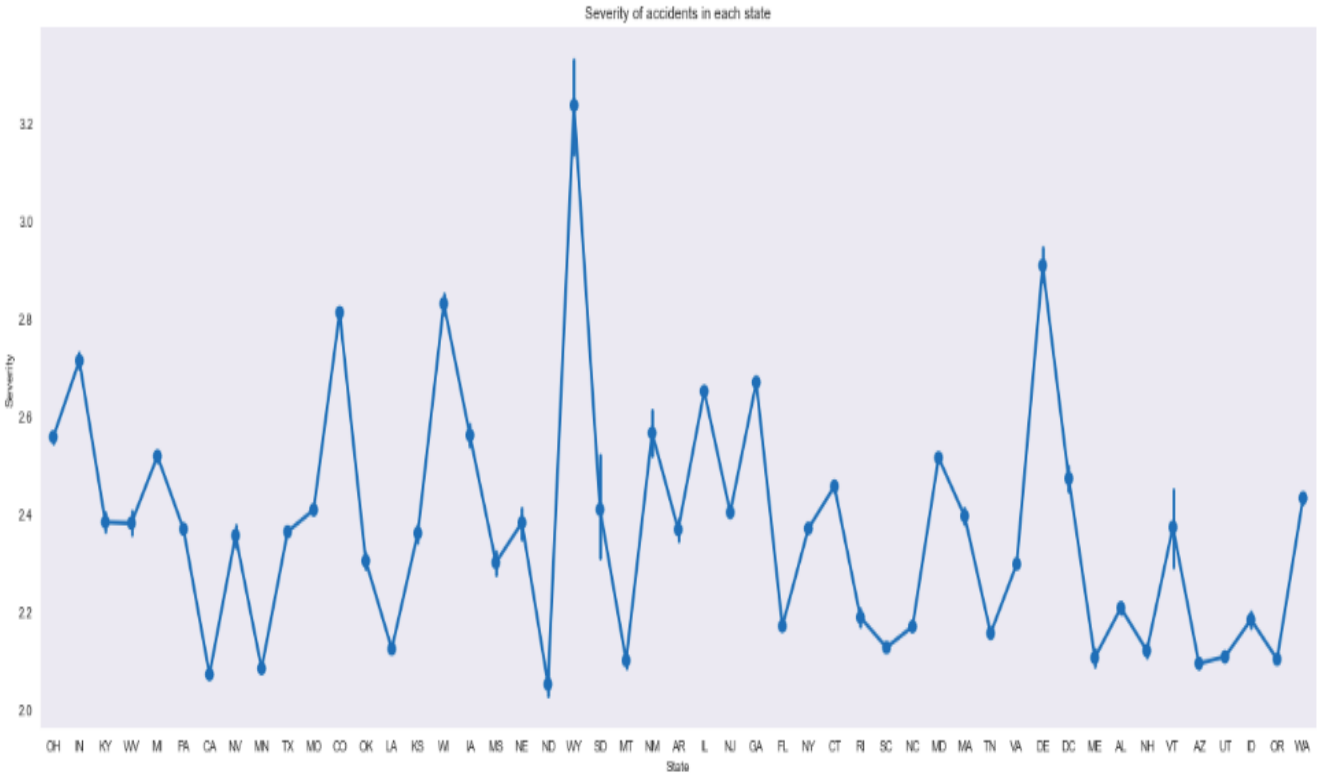

Fig 2.13 Severity of accidents in each state

Correlation

The correlation and multivariate regression both quantify how strongly the features are linked to the outcome (fatal accidents). When we compare the regression coefficients to the correlation coefficients, we notice that they differ slightly. The reason for this is that multiple regression calculates a feature's relationship with an outcome based on its association with all other characteristics, which is not taken into account when calculating correlation coefficients.

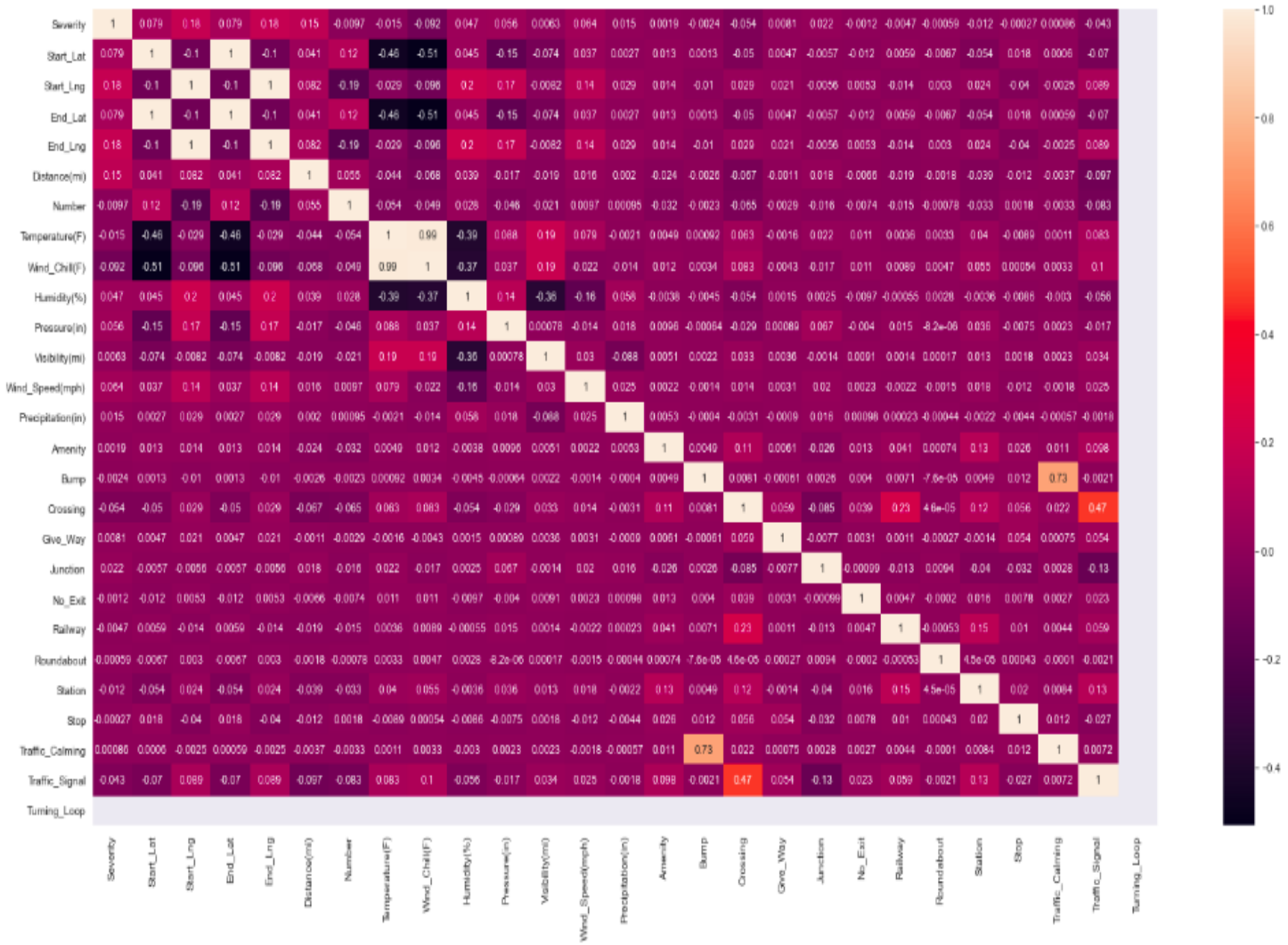

Fig 2.14 Heatmap of all variables

From the correlation table, we can see that the highest amount severity is most strongly correlated with Location. However, we can observe that other parameters such as crossings and traffic signals are positively correlated. As a result, we'd like to compute the target's (Severity) association with each feature while accounting for the effect of the other features.

```
Severity           1.000000
Start_Lat          0.078723
Start_Lng          0.179492
End_Lat            0.078729
End_Lng            0.179495
Distance(mi)       0.152869
Wind_Chill(F)      0.091947
Humidity(%)        0.047240
Pressure(in)       0.055531
Wind_Speed(mph)    0.063576
Crossing           0.053573
Junction           0.021885
Traffic_Signal     0.042802
Name: Severity, dtype: float64
```

Fig 2.15 features of correlation

Chapter-4

Unsupervised Learning

Unsupervised learning discovers patterns in a dataset without relying on known or labelled outcomes as a guide. Unsupervised machine learning methods can't be applied to a regression or classification problem right away since the values for the output data are unknown. Because training the algorithm is challenging, unsupervised learning may be used to find the data's underlying structure.

Clustering is one of the methods for unsupervised learning. Clustering helps you to divide a dataset into groups based on their similarity automatically. To perform clustering as a feature are supposed to be on a similar scale, and PCA is used to visualize data in reduced dimensional space. Figure 3.1 shows the scatter plot of the first four principal components.

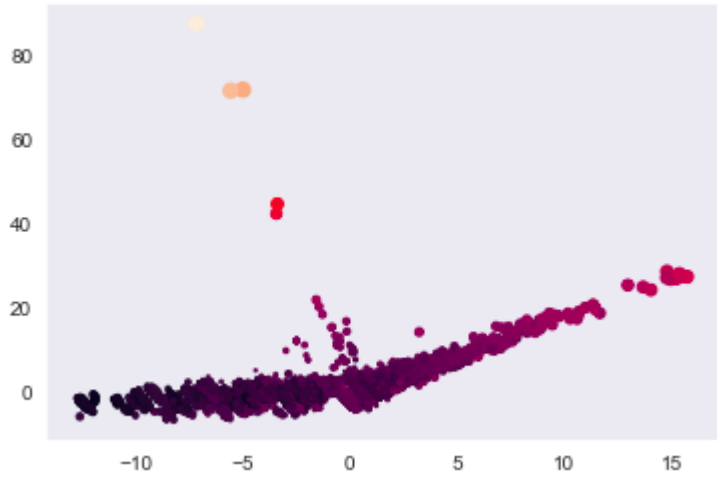

Fig 3.1 PCA principal component

The number of groups in which the accident should cluster was not obvious from the PCA scatter plot. So, I used KMeans clustering to help us locate an acceptable amount of groups by making an elbow plot and looking for the "elbow," which indicates that adding more clusters won't provide much explanatory power. Figure 3.2 shows an elbow plot made using k-mean. To my surprise, the elbow pot was more of a straight line rather than a curved plot making it impossible to recognise the cluster.

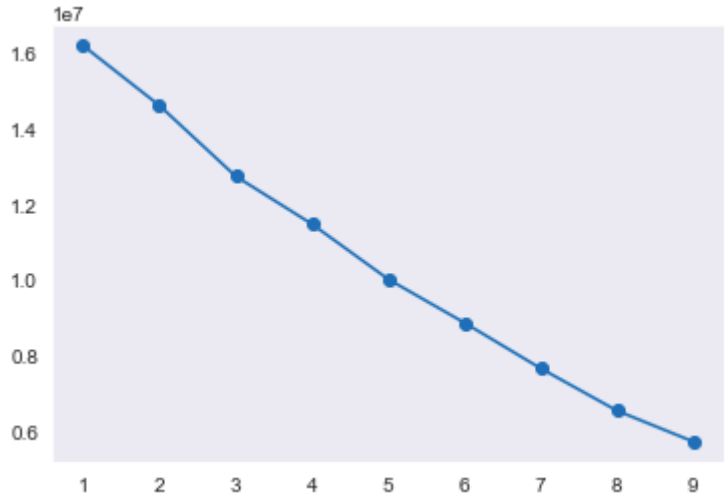

Fig 3.2 Elbow plot

I attempted to divide the data into four clusters based on the severity of the accidents using K-means clustering, however instead of treating each group as an individual, the algorithm overstated the similarities between them. As a result, cluster analysis proved an unsuitable method for my data.

Chapter-5

Supervised Learning

Supervised learning occurs when the data in our training set is labelled. In supervised learning, both the training and validation data are labelled before being fed to the model. Each piece of data given to the model during training is a pair consisting of the input item or sample and the associated label or output value, allowing the model to learn how to map specific inputs to specific outputs depending on what it learns from the labelled training data. In our situation, data refers to all numbers other than Severity, and categorization refers to the severity. The purpose of this research is to understand what factors influence the severity of an accident and how they affect it. Regression techniques are a collection of observable coefficients, not black boxes. As a result, regression approaches are responsible for the vast bulk of the analysis. The amount of data put aside for testing is 25% of the overall amount.

Linear Regression

Many factors influence the severity of an accident. A regression study utilising linear regression was undertaken to gain a better knowledge of the components that have the greatest impact during an accident. This dataset is vast, and numerous factors determine the severity of an accident, making it difficult to analyse. It also comprises a large number of completely useless, minor, and inconsequential features, which contribute significantly less to predictive modelling than the crucial variables. To find the critical function variable selection method, RFE, is used. Figure 4.1 shows the ranking made by RFE of each input feature. Recursive Feature Elimination works by deleting attributes iteratively and developing a model on the ones that remain. It employs model accuracy to determine which attributes are most important in predicting the target attribute.

```
Features sorted by their rank:
[(1, 'Amenity'), (2, 'Wind_Speed(mph)'), (3, 'Junction'), (4, 'Start_Lng'), (5, 'Crossing'), (6, 'Station'), (7,
'Severity'), (8, 'Stop'), (9, 'Start_Lat'), (10, 'Railway'), (11, 'Distance(mi)')]
```

Fig 4.1 Features by rank

Due to more than one independent variable, multiple linear regression was performed and Ordinary Least Squares was used to estimate the values of the coefficients. To perform linear regression the data is first divided into train and test sets. The training and test datasets must then be created. Scikit-Learn train test split() function is used for this. This function is given X and Y data, with the test size set to 0.25 and the random state set to 1. This returns four datasets as a result of this query. The main training data set, X train, contains 75% of the data, while y_train contains the target variable for that set of data. The X train data will be used to train our model, while the y_train data will be used to verify its performance and make any necessary improvements. Because the datasets differ in size, it is possible that our model might get confused, so the StandardScaler() function is used to scale the model to an even scale. The linear regression model is then created as a final stage. Figure 4.2 compares the actual y values and the predicted values by the model.

	Actual	Predicted
0	2	2.120414
1	2	2.089777
2	2	2.113969
3	2	2.086390
4	2	2.460954
...	...	...
367990	2	2.056560
367991	2	2.049743
367992	2	2.138908
367993	2	2.083350
367994	2	2.483777

367995 rows × 2 columns

Fig 4.2 Comparison of predicted and actual value

Logistic Regression

As the values are more categorical dependent rather than the continuous dependent variable, the linear regression model was unable to perform with high accuracy, and logistic regression was employed instead. Logistic regression is a method for modelling the probability of a discrete result given an input variable. Since the multi-class condition needs to be satisfied on the set of inputs, the x_train is transformed using the scaler.transform() function. Figure 2.19 shows the scale transformed function of the training set.

	0	1	2	3	4	5	6	7	8	9	10
0	-7.037856	5.373666	0.061667	-1.145515	-1.099053	-1.308452	-1.551419	-1.09955	-1.155852	-1.120573	-1.484624
1	-7.214456	5.281044	-0.176451	-1.644162	-1.099053	-1.308452	-1.551419	-1.09955	-1.155852	-1.120573	-1.484624
2	-7.224836	5.283574	-0.488979	-1.245245	-1.099053	-1.308452	-1.551419	-1.09955	-1.155852	-1.120573	-1.484624
3	-6.836446	5.353873	-0.582852	-0.979300	-1.099053	-1.308452	-1.551419	-1.09955	-1.155852	-1.120573	-1.484624
4	-7.515510	5.391865	-0.546219	-1.145515	-1.099053	-1.308452	-1.551419	-1.09955	-1.155852	-1.120573	-1.484624

Fig 4.3 Training set scale transformation

A confusion matrix is a metric that is used to assess a classification model's performance. In a confusion matrix, the number of correct and incorrect guesses is totaled for each class. Figure 4.4 shows how the actual target values compare to those predicted by the logistic regression model, giving us a clear view of how effectively our classification model is operating and what kinds of errors it produces.

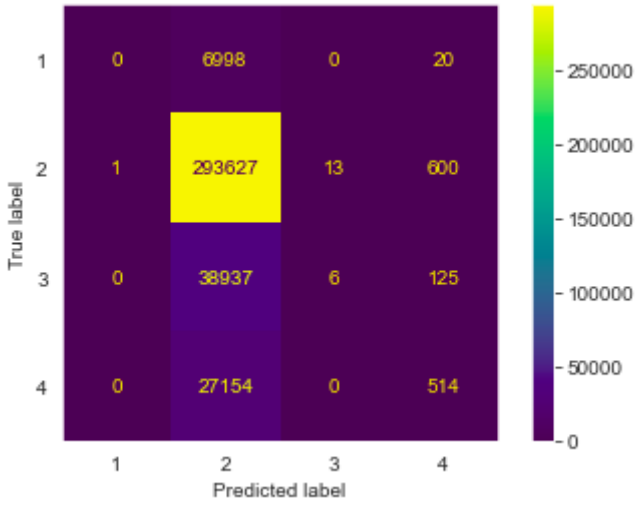

Fig 4.4 Confusion Matrix

Chapter-6

Reflection

If I were to restart the work, I might not have chosen this data, I would have chosen data that was more closely related to my knowledge base or a topic that was more appropriate for my expertise. I should have picked a data set that was easier to respond to using the skills I had learned in class. Because of its magnitude and the enormous number of variables that were linked to one another, working with this dataset was frequently intimidating. Focusing on a single aspect and going into further detail would have been preferable.

Because there are so many variables, the data didn't group themselves well to clustering. I put a lot of variables since I didn't know which ones had what impact, thinking that the important ones would stand out, which did not take place. I attempted to partition the data into four groups based on the severity of the accidents, but the algorithm overstated the similarities between them instead of treating each group as an individual. As a consequence, cluster analysis was shown to be an ineffective way for analysing my data.

This was a large dataset, and the severity of each accident is determined by a variety of factors, making it tough to analyse. It also include a slew of absolutely worthless, unimportant, and insignificant information that contribute far less to predictive modelling than the critical factors. Data being more categorical dependent rather than the continuous dependent variable made regression model unable to perform with high accuracy.

Chapter-7

Conclusion

Traffic accidents are a major public safety concern, and significant study has gone into analysing and predicting these uncommon occurrences. The research assisted us in determining the factors that cause accidents. A range of insights on the location, timing, weather, and points of interest of an accident may be gleaned from this dataset. One of the most crucial aspects of the data set to note is the absence of New York data, which, if added, would result in a few minor adjustments to the conclusions I reached. The severity of the accident, and how numerous factors might influence severity during an accident, was the most interesting component of the dataset for me. The research also aids us in determining the optimum month, day, and hour to travel. It can also assist us in predicting where each state has the most accident-prone locations are.

Finally, this research suggests infrastructural, policy, administrative, and human behaviour reforms that might help minimise US accidents.

Appendix A

Software used

Python : version 3.8.8

Jupyter Notebook :version 6.4.5

Dataset : https://www.kaggle.com/sobhanmoosavi/us-accidents

Bibliography

Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, and Rajiv Ramnath. "A countrywide traffic accident dataset." arXiv preprint arXiv:1906.05409 (2019).

Jain, Ayushi, Garima Ahuja, and Deepti Mehrotra. "Data mining approach to analyse the road accidents in India."*2016 5th International Conference on Reliability, Infocom Technologies and Optimisation (Trends and Future Directions)(ICRITO)*. IEEE, 2016.

Marom, Nadav David, Lior Rokach, and Armin Shmilovici. "Using the confusion matrix for improving ensemble classifiers."2010 IEEE 26-th Convention of Electrical and Electronics Engineers in Israel. IEEE, 2010.

Tukey, J. W. (1977). Exploratory data analysis (Vol. 2, pp. 131-160).