ITEC 620 - Group Project Proposal

Kogod School of Business

Group 2: Binh Minh An Nguyen, Ho-Ah Kim, Ahmed Malik, Steve Rodriguez

12/10/2021

Table of Contents

Executive Summary

According to the International Association of Safe International Travel, approximately 38,000 people dies annually due to traffic accidents in the US (Q, A., 2021). Our objective is to draw actionable conclusions based on trends and causes of traffic accidents to better inform the public domain on potential significant causes and trends of traffic accidents and what can be done to mitigate the high number of accidents in the United States.

Our dataset, retrieved from Kaggle (Moosavi, 2021), contains approximately 46 distinct columns attributing to one dependent variable - the severity of accidents and observations totaling approximately 1.04 million. The high number of data points was crucial in selecting the data set due to its robustness and exploration potential.

The methodology used to analyze the data consisted of both descriptive and predictive models that allowed for valuable insight into the frequency and significant factors contributing to the cause of the accidents. Predictive models allowed for the analysis leading to the probability of certain variables being identified as the major causes of severe accidents.

The most significant findings when applying descriptive and predictive methods are as follows:

Most accidents (40,000 - 50,000) with a severity level of two, which is also the highest severity frequency in the data set, occur between 3:00 PM and 6:00 PM. For accidents that occurred on highways, the driving factor causing all levels of severity was mainly caused by weather factors. For the accidents that occurred on city roads, all levels of severity were found to be caused primarily by road condition factors.

Through Logistic Regression (LR) analysis, we found 41 statistically significant variables associated with the probability of a severe accident. However, for predictions, we recommend the Classification Tree model as it generates a significantly smaller error rate (3.08%) in comparison with that by LR (10.6%).

Lastly, we use Holt Winter Model and predicted approximately 43,000 accidents for the first week of 2021.

Introduction

The problem of traffic accidents in the United States affects thousands of drivers every year. Not only cause significant traffic jams, but also death, and injuries which makes more than 20,000 people die on American roadways (Zipper, 2021). Besides, the other impacts follow up with traffic jams, death, and injuries, which are long-lasting emotional, physical, and sometimes financial burdens.

In the European Union, they dropped the accident rate by thirty-six percent between 2010 and 2020 and traffic deaths by making regulations and adjusting the restriction (Zipper, 2021). Reducing accidents is one of the crucial tasks for the United States. Therefore, the government's data as a reference for making regulations and restrictions will be a valuable source. In this project, by analyzing the accidents dataset, we would like to provide the government's data to provide solutions to reduce the accidents.

Our preliminary analysis of the dataset indicates leading variables related to the accidents' severity, including the location of the accident, weather, time that accident occurred, and road conditions. By implementing descriptive and predictive methods on this dataset, we will find the correlation among the variables and determine the relationships between what events, conditions, and causing effects impact more on the accident.

We will use two descriptive methods to explore and overview our dataset. These two techniques are: (1) Measure of Frequency to identify the pattern of accident frequency by hours and to check which states have the highest rates of accidents over 2016-2020, (2) Association Rules to understand the likelihood of weather, locations, lighting factors that could cause different levels of severity.

For predictive methods, we will examine Logistic Regression and Classification Tree on the prediction of accident severities. Next, we will analyze the potential seasonality of historic data to predict the next 7-day car accident frequency.

At the end of the project, we will provide actionable insights and suggestions to help US Government reduce the accident rates and accident severities.

Data ETL

Step 1 – A primary data wrangling was done in Excel before we imported the dataset to R. These transformations include inserting a zero value on blank cells under the **wind_speed** column if the **wind_direction** is *calm*. This is based on other existing combination values under these 2 columns. The total rows that had been revised is approximately 600,000 rows.

Step 2 - Import the dataset to R using *read_csv()* function under the *tidyverse* package and assign it into a variable named accidents. Since our dataset has multiple logical variables, read_csv() functions will be more efficient and productive.

```
accidents <- read_csv("~/Documents/Kogod - AU/Fall 2021/ITEC 620/Group Project/US_
Accident_calm_0.csv")

## Rows: 1048575 Columns: 47
```

In this section, we will primarily clean up our data; while moving on to the **Analysis** section, we will conduct secondary data wrangling for each analysis technique, if necessary.

Step 2 - In the following code chunk, we are going to:

1. Select only 26 relevant attributes from the original dataset
2. Rename some of the columns/variables
3. Re-format the date and time columns, and extract the respective year, month, week, date from the **Start_Date** column

Since this dataset involves more than 1 million distinct observations - by **Accident_ID**, duplicate removal is not required. Instead, the new dataset will be saved into a variable named *accident_c*.

```
accident_c <- accidents %>%
  select(
    Severity, Start_Time, End_Time, `Distance(mi)`, Description, Number, Street, City, State,
Timezone, Weather_Timestamp, `Temperature(F)`, `Wind_Chill(F)`, `Humidity(%)`,
`Visibility(mi)`, Wind_Direction, `Wind_Speed(mph)`, `Precipitation(in)`,Weather_Condition,
Amenity, Crossing, Give_Way, Junction, Railway, Station, Stop, Traffic_Signal,
Nautical_Twilight
  ) %>%
  rename(
    distance = `Distance(mi)`, temperature = `Temperature(F)`, wind_chill = `Wind_Chill(F)`,
humidity = `Humidity(%)`,visibility = `Visibility(mi)`, wind_speed = `Wind_Speed(mph)`,
```

```
precipitation = `Precipitation(in)`
  ) %>%
  # Find out if the accidents happened on highway or on city road
  mutate(is.highway = ifelse(is.na(Number), 1, 0)) %>%
  select(-c('Description', 'Number', 'Street')) %>%
  # Uniform column name to lower letter for better productivity
  select_all(tolower) %>%
  mutate(start_time_2 = parse_datetime(start_time, "%m/%d/%Y %H:%M")) %>%
  # Skip the missing values in start_time
  filter(!is.na(start_time_2)) %>%
  # Extract year, quarter, month, hour from the start frame
  mutate(
    year = year(start_time_2),
    quarter = quarter(start_time_2),
    month = month(start_time_2),
    hour = hour(start_time_2)
  )
```

There were 34,680 rows dropped from the first wrangling process because of the improper date-time values under the **Start_Time** column in our original data.

Step 3 - We identified cells containing multiple values under the column **weather_condition**. Such values observed are conflicting with the single-value criterion of a data frame. To address such issues, we use *separate_rows()* function to split the chunks and use *pivot_wider()* under the *tidyverse* package to set them up as new column names since these are weather conditions that potentially affect the outcome of an accident. The new dataset will then be assigned to a variable named *accident_pivot*.

```
accident_pivot <- accident_c %>%
  mutate(id = seq(1:n())) %>%
  filter(!is.na(weather_condition)) %>%
  mutate(observations = 1) %>%
  separate_rows(weather_condition, sep = " / ", convert = TRUE) %>%
  pivot_wider(names_from = weather_condition, values_from = observations, values_fill = 0)
```

The new dataset *accident_pivot* contains 103 variables. Now, we will clean up the column names for our productivity.

```
names(accident_pivot) <- str_replace_all(names(accident_pivot), " ", "_")

accident_pivot <- accident_pivot %>% select_all(tolower)
```

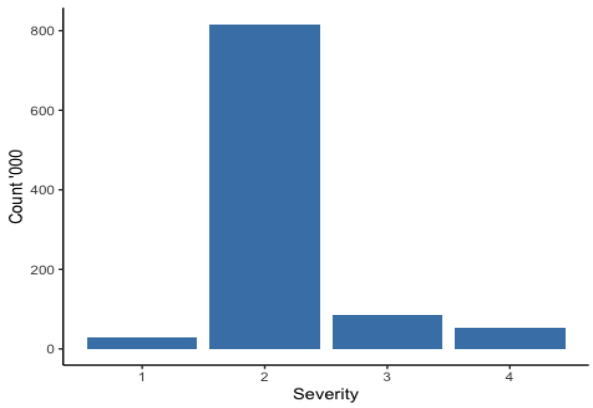
Analysis

Descriptive Data

Measure of Frequency

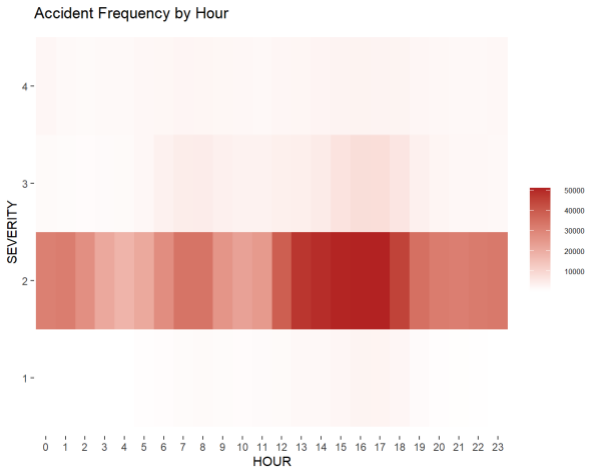
1. Count the number of each severity

We developed R code to generate the number of data points by each severity *[Appendix 2]*. Because of the majority of data points (just over 800K) in the data set having a severity level of two (slightly moderate), we experienced some skewness and accepted the fact.



2. Accident Rate by Hours

We will firstly group the occurrences of accidents overall by 24 hours and severity and use a heatmap to display how frequent accidents happen during the day. Given the overwhelming amount of data points related to Level 2 severity the heat map below accurately represents the time frame in which approximately 40,000 - 50,000 accidents occur, which is between 3:00 PM - 6:00 PM.



It is also logical that the bulk of accidents occur when millions of drivers are on the road due to rush hour. Refer to *[Appendix 1]* for the code to create this heatmap.

Association Rules

We are interested in learning which particular types of factors causing the most likelihood of each accident severity. Thus, from the initial dependent variable **severity**, we pick up and split each severity into new columns: **severity_1, severity_2, severity_3, and severity_4**. Similar process from the weather_condition will apply here. The new data with 107 columns is assigned to a variable named *accident_binary*.

```
accident_binary <- accident_pivot %>%
  # Create dummy variables for each accident severity
  mutate(
    observations = 1,
    is.day = ifelse(nautical_twilight == 'Day', TRUE, FALSE)) %>%
  pivot_wider(names_from = severity, values_from = observations, values_fill = 0) %>%
  rename(severity_1 = `1`, severity_2 = `2`, severity_3 = `3`, severity_4 = `4`) %>%
  select(-c('start_time', 'end_time', 'distance', 'city', 'state', 'timezone',
'weather_timestamp', 'temperature', 'wind_chill','wind_speed', 'id', 'humidity', 'visibility',
'wind_direction', 'precipitation', 'start_time_2', 'year', 'month', 'quarter', 'hour',
'nautical_twilight')
  ) %>%
  mutate_if(is.double, ~. > 0.5)
```

Furthermore, we segmented the analysis to two separate scenarios which are highway and city since the dynamics of driving are different from each other. In addition, our dataset skewed more to highway accidents with approximately as twice as the city road accidents. Thus, splitting the dataset by road types will help us to gain more valuable insights on the cause of accidents.

1. Association Rules for Highway accident severities

Step 1 - For highway category, location conditions such as *amenity*, *crossing*, *stop*, and *railway* become irrelevant. Thus, we removed these variables prior to running the association rules for the severity of highway accidents. The association rules will then be sorted by the lift ratio in descending order.

```
accident_hway <- accident_binary %>%
  filter(is.highway == TRUE) %>%
  select(-amenity, -crossing, -stop, -railway, -is.highway)

hway_rules <- apriori(accident_hway, parameter = list(supp=0.0001, conf=0.0001))
hway_sorted <- sort(hway_rules, by = "lift")
```

Step 2 - Subset the association rules for each severity at the rhs and select those rules with the lift ratio of at least 2 by using the *subset()* function. *[Appendix 3]*

Refer to *[Appendix 3]*, for **Severity 1**, the highest rule is *{traffic_signal, fair, is.day}* with the highest lift ratio of *8.871098*. The variables of *traffic signal, daytime, cloudy*, *windy*, and *rain* are the most frequently shown among the rules as causing effects for Severity 1.

We filtered the association rules with a lift ratio being equal or more than 2, to find out the association rules for **Severity 2**, it seems that there were no rules returned. Thus, for this section only, we will decrease the criterion of lift ratio for this severity.

```
hway_rules_2 <- subset(hway_sorted, subset = rhs %in% c("severity_2") & lift >= 1)
```

More than 500 combinations of conditions were generated; however, none of them reached the likelihood or a lift ratio of 1.5, since Severity 2 is dominant among all severities *[Appendix 4]*. Therefore, any conditions can cause the likelihood of this type of accident. In other words, none of them hold extreme effects. Overall, weather conditions that reduce visibility, day time, and highway junctions are the most common factors causing the likelihood of Severity 2, the most effective rule here is {*junction, smoke, is.day*} with a lift ratio of *1.201788*.

For **Severity 3**, the highest lift ratio of *2.982320* belongs to *{t-storm, is.day} [Appendix 5]*. The variables, which are most frequently shown up, are *storm, thunder*, and *rain*.

For the **Severity 4**, the most attentive rule is *{junction, light_snow}* with a lift ratio of *2.108521 [Appendix 6]*. Overall, *giveaway, junction, snow, overcast,* and *cloudy* have the highest occurrence.

In summary, for highway accidents, severities, especially Severity 3, are significantly impacted by the weather condition.

2. Association Rules for City Road accident severities

For city road accidents, all possible independent variables would be used to run the Association Rules.

```
accident_city <- accident_binary %>%
  filter(is.highway == FALSE) %>%
   select(-is.highway)

city_rules <- apriori(accident_city, parameter = list(supp=0.0001, conf=0.0001))
city_sorted <- sort(city_rules, by = "lift")
```

Similarly, we subset the association rules for each type of severity with the lift ratio being at least 2 to support our analysis.

Refer to *[Appendix 7]*, the highest lift ratio of *10.221148* for **Severity 1** belongs to *{crossing, traffic_signal, fair, is.day}***.** The most critical factors noted include *crossing, traffic_signal, amenity, station,* and *daytime*.

Similar to the association rules for Highway - Severity 2 and given the skewness of this dataset, we reduced the threshold for the lift ratio for **Severity 2** in the condition of city roads as well. Refer to *[Appendix 8]*, none of the rules' lift ratios are more than 1.5. In particular, the most effective rule is *{stop, haze}* with a lift ratio of *1.185398*. The popular variables are *stop, crossing, railway, station,* and *haze*.

For **Severity 3**, *{station, traffic_signal, overcast}* has the highest lift ratio of *7.937146 [Appendix 9]*. The frequent variables are *station, traffic_signal, overcast,* and *cloudy*.

Severity 4 has the highest lift ratio of *5.128429*, which belongs to *{stop, overcast} [Appendix 10]*. Out of all, *stop, overcast, cloudy,* and *clear* are the frequent variables.

In summary, City Road accident severities are more likely related to the road conditions. Typically, for the **Severity 3**, accidents usually occur at crowded areas, such as *station, traffic_signal, and amenities*.

3. Conclusions

While The weather factors more likely cause highway accident severities, City Road accident severities are affected by a combination of both weather and location conditions. However, the standard distinctive variable that causes the likelihood of both highway accidents and city road accidents is daytime. Given such surprising findings, we can primarily conclude that weather and road conditions are the only causes along with human factors.

Predictive Data

To predict the accident severity, we built and compared the results between Classification Tree and Logistic Regression models, while excluding k-NN Classification. since k-NN is a simple "lazy" machine learning method that is only suitable for small-medium sized datasets. For big datasets such as US Accidents, k-NN Classification method requires a considerable amount of processing power resulting in computing restrictions.

Logistic Regression

Before performing the logistic regression, we converted the logical variables into binary values and dropped missing values.

```
accident_glm <- accident_pivot %>%
  mutate_if(is.logical, as.integer) %>%
  mutate(is.day = ifelse(nautical_twilight == 'Day', 1, 0)) %>%
  filter(!is.na(weather_timestamp), !is.na(humidity), !is.na(visibility), !is.na(wind_speed),
!is.na(precipitation)) %>%
  filter(!is.na(is.day)) %>%
  filter(!is.na(weather_timestamp)) %>%
  filter(wind_direction != "North", !is.na(wind_speed)) %>%
  select(-start_time, -end_time, -distance, -city, -state, -timezone, -weather_timestamp, -
wind_chill, -wind_direction, -precipitation, -nautical_twilight, -start_time_2, -year, -
quarter, -month, -hour, -id, -'n/a_precipitation') %>%
  rename(heavy_storm = 'heavy_t-storm', storm = 't-storm') %>%
  drop_na()
```

Logistic Regression Model

Step 1 - We partitioned the *accident_glm* data with 60% of data points into a training set while the rest became the validating set. After partitioning, we built a logistic regression model on the training set with eighty-five independent variables, indicating location, day/night, and weather conditions. *[Appendix 11]*

Step 2 - We observed that there are ten *NA* results in our *summary()* output. These results are because the associated variables contain only 0 *(FALSE)* values. Based on the associated *p-values*, there are 41 statistically significant variables, including: *temperature, humidity, visibility, wind_speed, amenity, crossing, give_way, junction, railway, station, traffic_signal, is.highway, windy, is.day, drizzle_and_fog, light_sleet, light_rain_shower, widespread_dust, heavy_storm, storm, light_rain_with_thunder, thunder,*

thunderstorm, heavy_thunderstorm_and_rain, light_freezing_rain, smoke, light_rain_showers, light_thunderstorm_and_rain, thunderstorms_and_rain, drizzle, rain, haze, light_drizzle, light_freezing_drizzle, clear, scattered_clouds, light_snow, snow, overcast, light_rain, precipitation. Thus, we rebuilt the logistic regression model with these forty-one significant variables. *[Appendix 12]*

The new model shows three coefficients whose p-values are considerably significant but very near 0.05. As we have many variables, there is a problem with *The Curse of Dimensionality*. In other words, the larger the number of variables is, the demand for data points increases exponentially. This concept applies to our scenario. It is hard to distinguish p-values near 0.05 or determine whether those variables are significant. More observations from future data can be included to validate the accuracy of this model.

Step 3 - Thus, we can move on with the predictions by applying our logistic regression model on the validating set and calculate the overall accuracy:

```
sum(accident.glm.classifications == accidents.test.results) / length(accidents.test.results)
```

```
## [1] 0.8938174
```

Step 4 - Given the logistic regression model, we can accurately predict up to *89.40%* classification of Severity of accident on our test set, which is equivalent to an error rate of 10.6%. Additionally, we are interested in knowing how in particular, our model predicts the non-Severity (0) and Severity (1) and how accurate the model predicts severed accidents by generating the confusion matrix table given the built logistic regression model:

```
table(accident.glm.classifications, accidents.test.results)
```

```
##                             accidents.test.results
## accident.glm.classifications  FALSE   TRUE
##                            0 263882  31281
##                            1     69     14
```

Precision Rate: The percentage of correct prediction for severe accidents is:

14/(14 + 69) = 0.1687 or 16.87%

Classification trees

The second method is Classification tree to predict the binary accident severity. *[Appendix 13]*

Step 1 - Similar to Logistic Regression, we split data into a training set (60%) and a validating set (40%).

Step 2 - Computing the most optimal value of mindev and the best error rate for our classification tree. Result involves an error rate of *3.08%* with a mindev of *0.00053*.

Classification tree analysis

In our case, *TRUE* represents the more severe accidents. To see whether accidents are severe, we rounded the values at each end of the node. If value is less than 0.5, converted to 0 *(Not Severe)*, otherwise converted to 1 *(Severe)*. The first combination of independent variables (starting from the left) in the Classification Tree represents severe accidents given the conditions of *crossing*, *day*, *traffic_signal*, *humidity* < *20.5* and *temperature* > 93.5, *fair* weather. Firstly, this combination might represent accidents that occurred in the city, where a traffic signal was available. Secondly, the temperatures might indicate that this data is coming from states such as California, Texas and Florida which have similar climatic conditions. This is also confirmed because frequency of accidents occurring from these states in our data set is high. A deeper state-wide analysis could bring much more clarity on how these independent variables affect the severity and to what extent.

Model	Error Rate
Classification Tree	**3.08%**
Logistic Regression	10.6%

To analyze the classification accuracy of this model, we ran a similar R script to assess how many dependent variables in the validation set match the predictions produced from the classification model. As a result, the error rate produced by this model is **3.08%**, which is significantly lower than the logistic regression model (**10.6%)**. Thus, Classification Tree is recommended to predict whether the accident will be severe as comparatively it has a lower error rate.

Time-Series Analysis

Given the incomplete data point during 2018 and 2019, for this Time-Series analysis, we decided to use the daily data points grouped by weeks throughout the year 2020 as this year has the most completed data over time. Here, our group:

1. Filtered only those observations that happened in 2020
2. Used the *as.Date()* function on start_time_2 column to extract only date

Used the *cut.Date()* function to allocate our date values into the respective weeks, starting from week one by Sunday.

```
accident_ts_dt <- accident_c %>%
  filter(year == 2020) %>%
  mutate(
    year = year(start_time_2),
    month = month(start_time_2),
    date = as.Date(start_time_2)
  ) %>%
  mutate(week = cut.Date(date, breaks = "1 week", labels = FALSE))
```

Before running the Time-Series model, we firstly visualized our dataset. There were some data errors during **2020-07-01** to **2020-09-09** with the records of accidents much lower than normal (count of 400 or below). *[Appendix 14]*

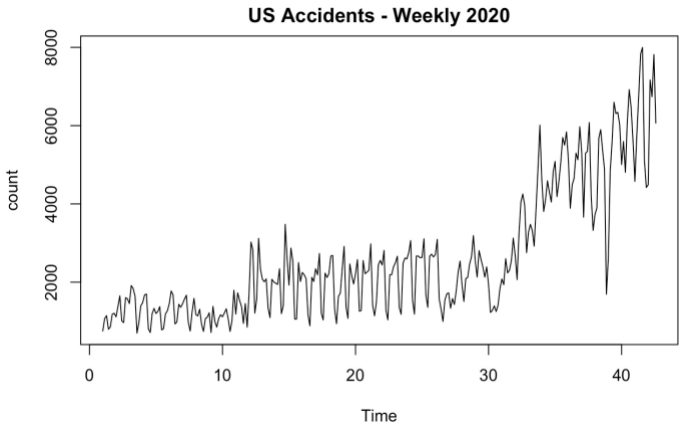
Fortunately, these data errors fit right within a period of nine weeks, which would not create much impact on our time-series analysis. To remove these data errors, we set a criterion of *count > 410* and connect the data from June-2020 to September 10th, 2020, onward.

```
# Make data possible for time-series analysis
accident_ts <- accident_ts_dt %>%
  group_by(year, month, day) %>%
  summarize(count = n()) %>%
  ungroup() %>%
  group_by(week) %>%
  summarize(count = sum(count)) %>%
  arrange(day) %>%
  select(count > 410) %>%
  ts(start = 1, freq=7)
```

```
## `summarise()` has grouped output by 'year', 'month'. You can override using the `.groups`
argument.
```

```
plot(accident_ts, main="US Accidents - Weekly 2020")
```

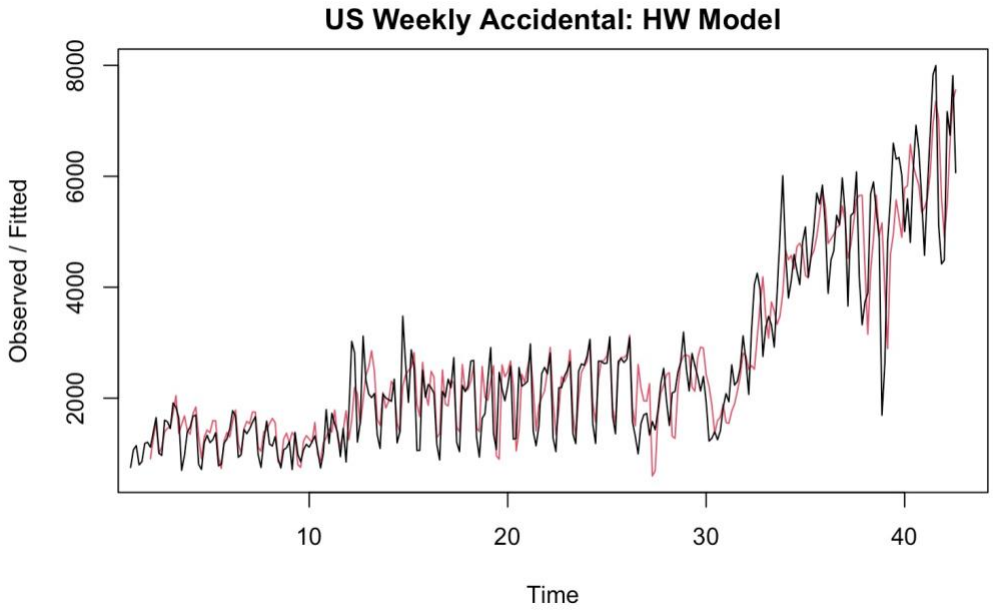
We could recognize an overall increasing trend and some sense of seasonality briefly. For example, it seems like the number of accidents went up during the weekdays toward Saturday. In addition to our Measure of Frequency section analysis, accidents mostly happen during peak hours because more people go down on roads. On the other hand, we see a pattern of accident number drops on Sunday, and perhaps, people prefer to stay at home for energy recharge. To ascertain which model better suited our time series, we computed the Mean Square Error (MSE) for the Single Exponential Smoothing model, Double Exponential Smoothing model, and Holt Winter model. Finally, we compared the results as another support for our potential model.



	MSE
Hwmodel	405,422.50
SESmodel	529,864.50
DESmodel	565,698.60

Refer to *[Appendix 15]*, on the right side; the three mean square errors were computed and displayed in ascending order.

As expected, the smallest MSE belongs to the Holt Winter model. Combining two arguments - (1) Sense of seasonality and (2) Returning the smallest MSE, Holt Winter works out the best for our time-series prediction case.



By applying the built Holt Winter model, we predicted the number of accidents during the first week of 2021 below:

```
predict(accident.HWmodel)

## Time Series:
## Start = c(46, 4)
## End = c(46, 4)
## Frequency = 7
##           fit
## [1,] 5822.509
## [2,] 5222.521
## [3,] 5121.899
## [4,] 6423.801
## [5,] 6552.980
## [6,] 7354.945
## [7,] 6866.619
```

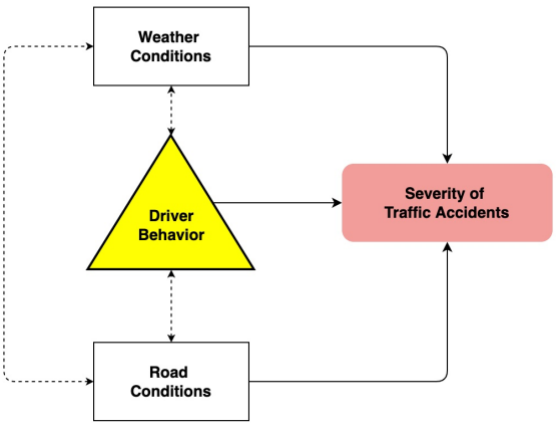
The prediction results seem to fit with what we observed earlier that the number of accidents is higher during weekdays toward the next Saturday.

Conclusions

Based on our analysis, we can conclude that:

1. Most of the accidents frequently happened during the peak hours, when more people took part in the traffic flows.
2. The dominant factor that is most likely causing accidents, applicable to all types of severity and the location is during the daytime. This finding also suggests that the drivers' behaviors also caused car accident severity.
3. We recommend the Classification Tree method to predict the accident severity, and this model returned a much better overall accuracy rate in comparison to that by the Logistic Regression model. Nevertheless, the Logistic Regression model is applicable to identify and assure the relationship between accident severity and the significant variables.
4. By using daily data points, we would predict the next seven-day of traffic accidents by using the Holt Winter model.

All analysis is to assist the US Traffic Department, Body, and local authority to have a more comprehensive view of accident severity and its root cause. Based on such analysis, we develop a cause-result diagram for traffic accidents on the right side.



Furthermore, the purpose of our analysis is to support the National Safety Transportation Board, and local authorities to assess and improve their current Road Safety Programs.

Thus, we propose the authority to:

- Review and revise the existing traffic code, provisions, law enforcement, and educate the traffic participants. For example, the California Department of Motor Vehicles can tighten the

requirements on taking driving licenses; promote safe driving skills and traffic guidance among teen drivers or raise drivers' awareness when they drive across states

- Immediately raise drivers' awareness regarding sudden weather changes, traffic flows, or road conditions ahead
- Frequently check the road quality and surrounding conditions for clearer sight

APPENDIX

Appendix 1 - R code to generate the Accident Frequency by Hour

```
accident_hour <- accident_pivot %>%
  group_by(hour, severity) %>%
  summarize(total_accidents = n())

## `summarise()` has grouped output by 'hour'. You can override using the `.groups` argument.

accident_hour %>%
  ggplot(mapping = aes(x = factor(hour), y = factor(severity))) +
  geom_tile(mapping = aes(fill = total_accidents)) +
  scale_fill_continuous(low = "white", high = "firebrick", name = NULL) +
  ggtitle("Accident Frequency by Hour") +
  xlab("HOUR") +
  ylab("SEVERITY") + theme(axis.ticks = element_line(linetype = "dashed"),
    axis.text = element_text(family = "Times",
        colour = "gray24"), axis.text.x = element_text(family = "Times"),
    axis.text.y = element_text(family = "Times"),
    panel.background = element_rect(fill = NA),
    legend.position = "left") +labs(fill = "Frequency") + theme(legend.text =
element_text(size = 6),
    legend.title = element_text(size = 24),
    legend.position = "bottom", legend.direction = "horizontal")  + theme(legend.position =
"right", legend.direction = "vertical")
```

Appendix 2 - R code to generate the bar chart displaying data count by severity

```
accident_pivot %>%
  group_by(severity) %>%
  summarize(count = n()) %>%
  ggplot(aes(x = severity, y = count/1000)) +
  geom_col(fill = "steelblue") +
  labs(x = "Severity", y = "Count '000") +
  theme_classic()
```

Appendix 3 – Association Rules for Highway – Severity 1

```
hway_rules_1 <- subset(hway_sorted, subset = rhs %in% c("severity_1") & lift >= 2)

inspect(hway_rules_1)
```

Description: df [20 × 8]

	lhs <chr>		rhs <chr>	support <dbl>	confidence <dbl>	coverage <dbl>	lift <dbl>	count <int>
[1]	{traffic_signal,fair,is.day}	=>	{severity_1}	0.0041752753	0.21183460	0.019710073	8.871098	2710
[2]	{traffic_signal,windy,is.day}	=>	{severity_1}	0.0001818017	0.14640199	0.001241798	6.130946	118
[3]	{traffic_signal,fair}	=>	{severity_1}	0.0045203903	0.14317783	0.031571860	5.995926	2934
[4]	{traffic_signal,windy}	=>	{severity_1}	0.0001818017	0.12102564	0.001502175	5.068248	118
[5]	{traffic_signal,is.day}	=>	{severity_1}	0.0068391317	0.11589171	0.059013125	4.853252	4439
[6]	{traffic_signal,mostly_cloudy,is.day}	=>	{severity_1}	0.0010615368	0.10138317	0.010470543	4.245671	689
[7]	{traffic_signal,cloudy,is.day}	=>	{severity_1}	0.0005808409	0.08935767	0.006500179	3.742074	377
[8]	{traffic_signal}	=>	{severity_1}	0.0073367752	0.08862173	0.082787543	3.711254	4762
[9]	{traffic_signal,mostly_cloudy}	=>	{severity_1}	0.0011185424	0.08590699	0.013020388	3.597568	726
[10]	{traffic_signal,partly_cloudy,is.day}	=>	{severity_1}	0.0005777595	0.08348175	0.006920788	3.496005	375

1–10 of 20 rows Previous 1 2 Next

	lhs <chr>		rhs <chr>	support <dbl>	confidence <dbl>	coverage <dbl>	lift <dbl>	count <int>
[11]	{traffic_signal,light_rain,is.day}	=>	{severity_1}	0.0002249410	0.07862143	0.002861065	3.292467	146
[12]	{station,traffic_signal,is.day}	=>	{severity_1}	0.0001818017	0.07861426	0.002312579	3.292167	118
[13]	{traffic_signal,partly_cloudy}	=>	{severity_1}	0.0006116547	0.07091819	0.008624794	2.969875	397
[14]	{station,fair,is.day}	=>	{severity_1}	0.0001186333	0.06666667	0.001779499	2.791832	77
[15]	{traffic_signal,cloudy}	=>	{severity_1}	0.0006286023	0.06020363	0.010441270	2.521176	408
[16]	{traffic_signal,light_rain}	=>	{severity_1}	0.0002326445	0.06008754	0.003871759	2.516315	151
[17]	{station,traffic_signal}	=>	{severity_1}	0.0001987493	0.05764075	0.003448069	2.413849	129
[18]	{fair,windy,is.day}	=>	{severity_1}	0.0001972086	0.05541126	0.003558998	2.320484	128
[19]	{mostly_cloudy,windy,is.day}	=>	{severity_1}	0.0001170926	0.05463695	0.002143103	2.288058	76
[20]	{fair,is.day}	=>	{severity_1}	0.0098157486	0.05385142	0.182274647	2.255162	6371

11–20 of 20 rows Previous 1 2 Next

Appendix 4 – Association Rules for Highway – Severity 2

```
hway_rules_2 <- subset(hway_sorted, subset = rhs %in% c("severity_2") & lift >= 1)
```

```
inspect(hway_rules_2)
```

	lhs <chr>		rhs <chr>	support <dbl>	confidence <dbl>	coverage <dbl>	lift <dbl>	count <int>
[1]	{junction,smoke,is.day}	=>	{severity_2}	0.0001864237	0.9918033	0.0001879644	1.201788	121
[2]	{junction,smoke}	=>	{severity_2}	0.0003343302	0.9774775	0.0003420336	1.184429	217
[3]	{haze,windy}	=>	{severity_2}	0.0001694761	0.9649123	0.0001756389	1.169203	110
[4]	{haze,windy,is.day}	=>	{severity_2}	0.0001247961	0.9642857	0.0001294181	1.168444	81
[5]	{blowing_snow}	=>	{severity_2}	0.0001633134	0.9464286	0.0001725575	1.146806	106
[6]	{smoke}	=>	{severity_2}	0.0020722307	0.9392458	0.0022062709	1.138103	1345
[7]	{blowing_snow,is.day}	=>	{severity_2}	0.0001155519	0.9375000	0.0001232554	1.135987	75
[8]	{smoke,is.day}	=>	{severity_2}	0.0012094432	0.9257075	0.0013065068	1.121698	785
[9]	{haze}	=>	{severity_2}	0.0104951938	0.9109388	0.0115212947	1.103803	6812
[10]	{snow,windy}	=>	{severity_2}	0.0002465107	0.9090909	0.0002711618	1.101564	160

1–10 of 57 rows Previous 1 2 3 4 5 6 Next

	lhs <chr>		rhs <chr>	support <dbl>	confidence <dbl>	coverage <dbl>	lift <dbl>	count <int>
[11]	{patches_of_fog}	=>	{severity_2}	0.0003682254	0.9087452	0.0004052020	1.101145	239
[12]	{wintry_mix}	=>	{severity_2}	0.0010276416	0.9087193	0.0011308679	1.101113	667
[13]	{n/a_precipitation}	=>	{severity_2}	0.0003250860	0.9055794	0.0003589812	1.097309	211
[14]	{junction,wintry_mix}	=>	{severity_2}	0.0001032264	0.9054054	0.0001140112	1.097098	67
[15]	{light_snow,windy}	=>	{severity_2}	0.0008843572	0.9053628	0.0009767987	1.097046	574
[16]	{wintry_mix,is.day}	=>	{severity_2}	0.0004252310	0.9049180	0.0004699111	1.096507	276
[17]	{blowing_dust}	=>	{severity_2}	0.0001124705	0.9012346	0.0001247961	1.092044	73
[18]	{blowing_dust,is.day}	=>	{severity_2}	0.0001093891	0.8987342	0.0001217147	1.089014	71
[19]	{junction,traffic_signal,cloudy}	=>	{severity_2}	0.0001633134	0.8983051	0.0001818017	1.088494	106
[20]	{fog}	=>	{severity_2}	0.0126305929	0.8971329	0.0140788434	1.087074	8198

11–20 of 57 rows Previous 1 2 3 4 5 6 Next

	lhs <chr>		rhs <chr>	support <dbl>	confidence <dbl>	coverage <dbl>	lift <dbl>	count <int>
[21]	{junction,haze}	=>	{severity_2}	0.0019720857	0.8957313	0.0022016488	1.085375	1280
[22]	{haze,is.day}	=>	{severity_2}	0.0073321532	0.8927031	0.0082134290	1.081706	4759
[23]	{patches_of_fog,is.day}	=>	{severity_2}	0.0002557549	0.8924731	0.0002865687	1.081427	166
[24]	{n/a_precipitation,is.day}	=>	{severity_2}	0.0001494471	0.8899083	0.0001679354	1.078320	97
[25]	{light_snow,windy,is.day}	=>	{severity_2}	0.0003759288	0.8840580	0.0004252310	1.071231	244
[26]	{junction,traffic_signal,cloudy,is.day}	=>	{severity_2}	0.0001032264	0.8815789	0.0001170926	1.068227	67
[27]	{traffic_signal,fog}	=>	{severity_2}	0.0007010149	0.8800774	0.0007965378	1.066407	455
[28]	{fair}	=>	{severity_2}	0.2949346670	0.8784698	0.3357368128	1.064459	191430
[29]	{junction,haze,is.day}	=>	{severity_2}	0.0013897042	0.8765792	0.0015853721	1.062169	902
[30]	{station,fair}	=>	{severity_2}	0.0026315019	0.8670051	0.0030351632	1.050567	1708

	lhs <chr>		rhs <chr>	support <dbl>	confidence <dbl>	coverage <dbl>	lift <dbl>	count <int>
[31]	{junction,fog}	=>	{severity_2}	0.0021369398	0.8657928	0.0024681886	1.049098	1387
[32]	{shallow_fog}	=>	{severity_2}	0.0002865687	0.8651163	0.0003312488	1.048279	186
[33]	{mist,is.day}	=>	{severity_2}	0.0001833423	0.8623188	0.0002126155	1.044889	119
[34]	{traffic_signal,haze}	=>	{severity_2}	0.0004760738	0.8607242	0.0005531084	1.042957	309
[35]	{junction,traffic_signal,fair}	=>	{severity_2}	0.0003666847	0.8592058	0.0004267717	1.041117	238
[36]	{cloudy}	=>	{severity_2}	0.1089839290	0.8590321	0.1268682816	1.040906	70737
[37]	{traffic_signal,heavy_rain}	=>	{severity_2}	0.0002603769	0.8578680	0.0003035163	1.039496	169
[38]	{fair,windy}	=>	{severity_2}	0.0040782117	0.8577447	0.0047545755	1.039346	2647
[39]	{junction,fair,windy}	=>	{severity_2}	0.0005931664	0.8574610	0.0006917707	1.039003	385
[40]	{junction,fair}	=>	{severity_2}	0.0450868103	0.8556975	0.0526901252	1.036866	29264

	lhs <chr>		rhs <chr>	support <dbl>	confidence <dbl>	coverage <dbl>	lift <dbl>	count <int>
[41]	{fog,is.day}	=>	{severity_2}	0.0052999804	0.8498024	0.0062367212	1.029722	3440
[42]	{give_way,traffic_signal,fair}	=>	{severity_2}	0.0003774695	0.8477509	0.0004452600	1.027237	245
[43]	{traffic_signal,haze,is.day}	=>	{severity_2}	0.0003497371	0.8407407	0.0004159868	1.018742	227
[44]	{heavy_snow}	=>	{severity_2}	0.0005500270	0.8400000	0.0006547941	1.017845	357
[45]	{station,clear,is.day}	=>	{severity_2}	0.0002742432	0.8396226	0.0003266267	1.017388	178
[46]	{give_way,traffic_signal,cloudy}	=>	{severity_2}	0.0001525285	0.8389831	0.0001818017	1.016612	99
[47]	{junction,traffic_signal,fair,is.day}	=>	{severity_2}	0.0002002900	0.8387097	0.0002388073	1.016281	130
[48]	{heavy_rain}	=>	{severity_2}	0.0049194295	0.8358639	0.0058854434	1.012833	3193
[49]	{junction,traffic_signal}	=>	{severity_2}	0.0011354900	0.8327684	0.0013635124	1.009082	737
[50]	{mist}	=>	{severity_2}	0.0003589812	0.8321429	0.0004313938	1.008324	233

	lhs <chr>		rhs <chr>	support <dbl>	confidence <dbl>	coverage <dbl>	lift <dbl>	count <int>
[51]	{junction,fog,is.day}	=>	{severity_2}	0.0009583104	0.8315508	0.0011524376	1.007607	622
[52]	{junction,light_snow,windy}	=>	{severity_2}	0.0001063077	0.8313253	0.0001278774	1.007333	69
[53]	{light_snow}	=>	{severity_2}	0.0125689652	0.8288966	0.0151634905	1.004391	8158
[54]	{rain}	=>	{severity_2}	0.0127923656	0.8276515	0.0154562220	1.002882	8303
[55]	{traffic_signal,scattered_clouds,is.day}	=>	{severity_2}	0.0016824356	0.8266465	0.0020352541	1.001664	1092
[56]	{fair,windy,is.day}	=>	{severity_2}	0.0029411810	0.8264069	0.0035589985	1.001374	1909
[57]	{junction,traffic_signal,is.day}	=>	{severity_2}	0.0007672646	0.8258706	0.0009290373	1.000724	498

Appendix 5 – Association Rules for Highway – Severity 3

```
hway_rules_3 <- subset(hway_sorted, subset = rhs %in% c("severity_3") & lift >= 2)

inspect(hway_rules_3)
```

	lhs <chr>		rhs <chr>	support <dbl>	confidence <dbl>	coverage <dbl>	lift <dbl>	count <int>
[1]	{t-storm,is.day}	=>	{severity_3}	0.0003374115	0.3164740	0.0010661589	2.982320	219
[2]	{heavy_t-storm,is.day}	=>	{severity_3}	0.0001617727	0.3106509	0.0005207539	2.927446	105
[3]	{light_rain_with_thunder,is.day}	=>	{severity_3}	0.0002434293	0.3098039	0.0007857529	2.919464	158
[4]	{light_rain_with_thunder}	=>	{severity_3}	0.0002757839	0.3070326	0.0008982234	2.893348	179
[5]	{thunder,is.day}	=>	{severity_3}	0.0001833423	0.2960199	0.0006193582	2.789569	119
[6]	{light_thunderstorms_and_rain}	=>	{severity_3}	0.0002495921	0.2903226	0.0008597061	2.735880	162
[7]	{t-storm}	=>	{severity_3}	0.0003651440	0.2883212	0.0012664488	2.717020	237
[8]	{thunder}	=>	{severity_3}	0.0001910458	0.2755556	0.0006933114	2.596722	124
[9]	{heavy_t-storm}	=>	{severity_3}	0.0001787203	0.2735849	0.0006532534	2.578151	116
[10]	{light_thunderstorms_and_rain,is.day}	=>	{severity_3}	0.0001848830	0.2575107	0.0007179625	2.426675	120

	lhs <chr>		rhs <chr>	support <dbl>	confidence <dbl>	coverage <dbl>	lift <dbl>	count <int>
[11]	{thunderstorm}	=>	{severity_3}	0.0001617727	0.2343750	0.0006902300	2.208653	105
[12]	{station,traffic_signal,mostly_cloudy,is.day}	=>	{severity_3}	0.0001109298	0.2236025	0.0004961028	2.107138	72
[13]	{thunderstorm,is.day}	=>	{severity_3}	0.0001386623	0.2205882	0.0006286023	2.078732	90
[14]	{thunder_in_the_vicinity,is.day}	=>	{severity_3}	0.0001556099	0.2200436	0.0007071776	2.073600	101

Appendix 6 – Association Rules for Highway – Severity 4

```
hway_rules_4 <- subset(hway_sorted, subset = rhs %in% c("severity_4") & lift >= 1.5)

inspect(hway_rules_4)
```

	lhs <chr>		rhs <chr>	support <dbl>	confidence <dbl>	coverage <dbl>	lift <dbl>	count <int>
[1]	{junction,light_snow}	=>	{severity_4}	0.0002326445	0.09431605	0.002466648	2.108521	151
[2]	{traffic_signal,clear}	=>	{severity_4}	0.0005731374	0.09360845	0.006122710	2.092702	372
[3]	{junction,light_snow,is.day}	=>	{severity_4}	0.0001355809	0.09034908	0.001500634	2.019835	88
[4]	{overcast}	=>	{severity_4}	0.0038486486	0.08740684	0.044031436	1.954059	2498
[5]	{snow,is.day}	=>	{severity_4}	0.0001247961	0.08653846	0.001442088	1.934646	81
[6]	{traffic_signal,overcast}	=>	{severity_4}	0.0002249410	0.08639053	0.002603769	1.931339	146
[7]	{give_way}	=>	{severity_4}	0.0002033713	0.08466966	0.002401939	1.892867	132
[8]	{clear}	=>	{severity_4}	0.0083875272	0.08278211	0.101320527	1.850669	5444
[9]	{traffic_signal,scattered_clouds}	=>	{severity_4}	0.0001648540	0.07604833	0.002167754	1.700129	107
[10]	{snow}	=>	{severity_4}	0.0001571506	0.07544379	0.002083016	1.686614	102

	lhs <chr>		rhs <chr>	support <dbl>	confidence <dbl>	coverage <dbl>	lift <dbl>	count <int>
[11]	{give_way,is.day}	=>	{severity_4}	0.0001232554	0.07373272	0.001671651	1.648362	80
[12]	{traffic_signal,clear,is.day}	=>	{severity_4}	0.0003805509	0.07209574	0.005278411	1.611766	247
[13]	{scattered_clouds}	=>	{severity_4}	0.0020645273	0.07143619	0.028900300	1.597021	1340
[14]	{junction,overcast}	=>	{severity_4}	0.0007503170	0.06851435	0.010951239	1.531700	487
[15]	{traffic_signal,scattered_clouds,is.day}	=>	{severity_4}	0.0001386623	0.06813020	0.002035254	1.523112	90
[16]	{light_snow,is.day}	=>	{severity_4}	0.0006085733	0.06739464	0.009029996	1.506668	395

Appendix 7 – Association Rules for City Road – Severity 1 (The first 90 rows)

```
city_rules_1 <- subset(city_sorted, subset = rhs %in% c("severity_1") & lift >= 2)

inspect(city_rules_1)
```

	lhs		rhs	support	confidence	coverage	lift
	<chr>	<chr>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>
[1]	{crossing,traffic_signal,fair,is.day}	=>	{severity_1}	0.0092442305	0.3741111	0.0247098521	10.221148
[2]	{amenity,crossing,station,traffic_signal,fair,is.day}	=>	{severity_1}	0.0001221048	0.3445378	0.0003544019	9.413171
[3]	{amenity,crossing,railway,station,traffic_signal}	=>	{severity_1}	0.0001131703	0.3333333	0.0003395110	9.107052
[4]	{amenity,crossing,station,fair,is.day}	=>	{severity_1}	0.0001250830	0.2978723	0.0004199216	8.138217
[5]	{crossing,fair,is.day}	=>	{severity_1}	0.0097832788	0.2902200	0.0337098729	7.929145
[6]	{amenity,railway,station,traffic_signal}	=>	{severity_1}	0.0001131703	0.2835821	0.0003990744	7.747790
[7]	{crossing,traffic_signal,fair}	=>	{severity_1}	0.0102568073	0.2805474	0.0365599788	7.664879
[8]	{amenity,crossing,railway,station}	=>	{severity_1}	0.0001131703	0.2773723	0.0004080089	7.578131
[9]	{crossing,station,traffic_signal,light_rain,is.day}	=>	{severity_1}	0.0001369957	0.2705882	0.0005062884	7.392783
[10]	{crossing,station,traffic_signal,cloudy,is.day}	=>	{severity_1}	0.0002412315	0.2621359	0.0009202536	7.161856

	lhs		rhs	support	confidence	coverage	lift
	<chr>	<chr>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>
[11]	{amenity,crossing,station,traffic_signal,fair}	=>	{severity_1}	0.0001429520	0.2608696	0.0005479827	7.127258
[12]	{amenity,crossing,station,traffic_signal,is.day}	=>	{severity_1}	0.0003335547	0.2598608	0.0012835900	7.099697
[13]	{crossing,station,traffic_signal,fair,is.day}	=>	{severity_1}	0.0004794849	0.2580128	0.0018583762	7.049208
[14]	{amenity,crossing,railway,traffic_signal,is.day}	=>	{severity_1}	0.0001310394	0.2573099	0.0005092666	7.030005
[15]	{amenity,railway,traffic_signal,is.day}	=>	{severity_1}	0.0001310394	0.2365591	0.0005539391	6.463069
[16]	{crossing,railway,station,traffic_signal,is.day}	=>	{severity_1}	0.0001786900	0.2362205	0.0007564544	6.453816
[17]	{amenity,crossing,railway,is.day}	=>	{severity_1}	0.0001399739	0.2361809	0.0005926552	6.452735
[18]	{amenity,railway,station}	=>	{severity_1}	0.0001131703	0.2360248	0.0004794849	6.448471
[19]	{amenity,crossing,railway,traffic_signal}	=>	{severity_1}	0.0001518865	0.2339450	0.0006492404	6.391646
[20]	{crossing,station,traffic_signal,light_rain}	=>	{severity_1}	0.0001399739	0.2315271	0.0006045679	6.325588

	lhs		rhs	support	confidence	coverage	lift
	<chr>	<chr>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>
[21]	{amenity,crossing,station,is.day}	=>	{severity_1}	0.0003395110	0.2280000	0.0014890835	6.229223
[22]	{crossing,traffic_signal,is.day}	=>	{severity_1}	0.0143220054	0.2256052	0.0634826090	6.163794
[23]	{crossing,railway,station,is.day}	=>	{severity_1}	0.0002322970	0.2247839	0.0010334240	6.141355
[24]	{amenity,railway,is.day}	=>	{severity_1}	0.0001459302	0.2227273	0.0006551968	6.085166
[25]	{amenity,crossing,station,fair}	=>	{severity_1}	0.0001489084	0.2192982	0.0006790221	5.991481
[26]	{crossing,railway,traffic_signal,fair,is.day}	=>	{severity_1}	0.0002710132	0.2177033	0.0012448738	5.947907
[27]	{amenity,station,traffic_signal,fair,is.day}	=>	{severity_1}	0.0001310394	0.2167488	0.0006045679	5.921827
[28]	{crossing,traffic_signal,fair,windy,is.day}	=>	{severity_1}	0.0001131703	0.2159091	0.0005241574	5.898886
[29]	{amenity,crossing,station,traffic_signal}	=>	{severity_1}	0.0003603582	0.2145390	0.0016796862	5.861454
[30]	{traffic_signal,fair,is.day}	=>	{severity_1}	0.0122492011	0.2138178	0.0572880215	5.841750

	lhs		rhs	support	confidence	coverage	lift
	<chr>	<chr>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>
[31]	{crossing,station,cloudy,is.day}	=>	{severity_1}	0.0002978167	0.2136752	0.0013937822	5.837854
[32]	{railway,traffic_signal,fair,is.day}	=>	{severity_1}	0.0002799477	0.2136364	0.0013103935	5.836792
[33]	{crossing,station,light_rain,is.day}	=>	{severity_1}	0.0001489084	0.2109705	0.0007058256	5.763957
[34]	{amenity,railway,traffic_signal}	=>	{severity_1}	0.0001518865	0.2090164	0.0007266728	5.710569
[35]	{amenity,crossing,traffic_signal,fair,is.day}	=>	{severity_1}	0.0002591005	0.2032710	0.0012746555	5.553599
[36]	{crossing,fair}	=>	{severity_1}	0.0108494626	0.2024451	0.0535921162	5.531035
[37]	{crossing,railway,station,traffic_signal}	=>	{severity_1}	0.0001935809	0.2018634	0.0009589698	5.515140
[38]	{amenity,crossing,traffic_signal,mostly_cloudy,is.day}	=>	{severity_1}	0.0001876245	0.2006369	0.0009351445	5.481633
[39]	{crossing,station,traffic_signal,cloudy}	=>	{severity_1}	0.0002501660	0.1990521	0.0012567865	5.438334
[40]	{crossing,station,traffic_signal,is.day}	=>	{severity_1}	0.0012091358	0.1947242	0.0062094783	5.320091

	lhs <chr>		rhs <chr>	support <dbl>	confidence <dbl>	coverage <dbl>	lift <dbl>
[41]	{amenity,crossing,traffic_signal,cloudy,is.day}	=>	{severity_1}	0.0001250830	0.1935484	0.0006462623	5.287966
[42]	{crossing,traffic_signal,fair,windy}	=>	{severity_1}	0.0001131703	0.1928934	0.0005866989	5.270071
[43]	{amenity,crossing,railway}	=>	{severity_1}	0.0001608210	0.1928571	0.0008338868	5.269080
[44]	{railway,station,traffic_signal,is.day}	=>	{severity_1}	0.0001846464	0.1925466	0.0009589698	5.260595
[45]	{crossing,station,fair,is.day}	=>	{severity_1}	0.0005450046	0.1826347	0.0029841234	4.989792
[46]	{crossing,traffic_signal}	=>	{severity_1}	0.0156919622	0.1822427	0.0861047660	4.979080
[47]	{amenity,crossing,station}	=>	{severity_1}	0.0003692927	0.1815520	0.0020340881	4.960210
[48]	{crossing,railway,station}	=>	{severity_1}	0.0002471879	0.1796537	0.0013759132	4.908346
[49]	{crossing,traffic_signal,partly_cloudy,is.day}	=>	{severity_1}	0.0012925245	0.1791908	0.0072131206	4.895698
[50]	{amenity,crossing,traffic_signal,mostly_cloudy}	=>	{severity_1}	0.0002025154	0.1789474	0.0011317035	4.889049

	lhs <chr>		rhs <chr>	support <dbl>	confidence <dbl>	coverage <dbl>	lift <dbl>
[51]	{amenity,railway}	=>	{severity_1}	0.0001667774	0.1777778	0.0009381226	4.857094
[52]	{crossing,station,light_rain}	=>	{severity_1}	0.0001518865	0.1758621	0.0008636684	4.804755
[53]	{crossing,is.day}	=>	{severity_1}	0.0155817700	0.1748020	0.0891395182	4.775794
[54]	{railway,station,is.day}	=>	{severity_1}	0.0002412315	0.1734475	0.0013908040	4.738787
[55]	{crossing,station,traffic_signal,fair}	=>	{severity_1}	0.0005092666	0.1703187	0.0029900797	4.653304
[56]	{station,traffic_signal,cloudy,is.day}	=>	{severity_1}	0.0002918604	0.1695502	0.0017213806	4.632307
[57]	{amenity,crossing,traffic_signal,is.day}	=>	{severity_1}	0.0006909348	0.1681159	0.0041098705	4.593122
[58]	{station,traffic_signal,light_rain,is.day}	=>	{severity_1}	0.0001518865	0.1677632	0.0009053628	4.583483
[59]	{crossing,traffic_signal,mostly_cloudy,is.day}	=>	{severity_1}	0.0020668479	0.1661082	0.0124427820	4.538268
[60]	{crossing,railway,traffic_signal,is.day}	=>	{severity_1}	0.0006135024	0.1638823	0.0037435560	4.477453

	lhs <chr>		rhs <chr>	support <dbl>	confidence <dbl>	coverage <dbl>	lift <dbl>
[61]	{amenity,crossing,cloudy,is.day}	=>	{severity_1}	0.0001310394	0.1635688	0.0008011269	4.468888
[62]	{crossing,railway,partly_cloudy,is.day}	=>	{severity_1}	0.0001310394	0.1617647	0.0008100614	4.419599
[63]	{railway,station,traffic_signal}	=>	{severity_1}	0.0002025154	0.1615202	0.0012538083	4.412918
[64]	{amenity,crossing,fair,is.day}	=>	{severity_1}	0.0002739914	0.1611208	0.0017005334	4.402008
[65]	{amenity,crossing,mostly_cloudy,is.day}	=>	{severity_1}	0.0002025154	0.1607565	0.0012597647	4.392053
[66]	{crossing,station,traffic_signal,mostly_cloudy,is.day}	=>	{severity_1}	0.0002412315	0.1607143	0.0015009962	4.390900
[67]	{crossing,traffic_signal,partly_cloudy}	=>	{severity_1}	0.0013878258	0.1597532	0.0086873133	4.364641
[68]	{crossing,fair,windy,is.day}	=>	{severity_1}	0.0001221048	0.1595331	0.0007653889	4.358628
[69]	{crossing,traffic_signal,cloudy,is.day}	=>	{severity_1}	0.0011376598	0.1587038	0.0071684481	4.335971
[70]	{railway,traffic_signal,is.day}	=>	{severity_1}	0.0006313714	0.1585639	0.0039818094	4.332150

	lhs <chr>		rhs <chr>	support <dbl>	confidence <dbl>	coverage <dbl>	lift <dbl>
[71]	{crossing,railway,traffic_signal,fair}	=>	{severity_1}	0.0003216420	0.1583578	0.0020311099	4.326517
[72]	{station,traffic_signal,fair,is.day}	=>	{severity_1}	0.0006105242	0.1580571	0.0038626827	4.318301
[73]	{crossing,railway,traffic_signal,mostly_cloudy,is.day}	=>	{severity_1}	0.0001340175	0.1578947	0.0008487776	4.313867
[74]	{crossing,station,cloudy}	=>	{severity_1}	0.0003067512	0.1574924	0.0019477213	4.302873
[75]	{amenity,station,traffic_signal,is.day}	=>	{severity_1}	0.0003544019	0.1567852	0.0022604288	4.283554
[76]	{amenity,station,traffic_signal,fair}	=>	{severity_1}	0.0001518865	0.1559633	0.0009738606	4.261098
[77]	{railway,traffic_signal,fair}	=>	{severity_1}	0.0003305765	0.1526823	0.0021651275	4.171456
[78]	{amenity,crossing,traffic_signal,fair}	=>	{severity_1}	0.0002918604	0.1519380	0.0019209178	4.151121
[79]	{crossing,station,traffic_signal}	=>	{severity_1}	0.0012597647	0.1501597	0.0083894966	4.102538
[80]	{traffic_signal,fair}	=>	{severity_1}	0.0134940749	0.1500132	0.0899525578	4.098535

	lhs <chr>		rhs <chr>	support <dbl>	confidence <dbl>	coverage <dbl>	lift <dbl>
[81]	{traffic_signal,fair,windy,is.day}	=>	{severity_1}	0.0001935809	0.1497696	0.0012925245	4.091878
[82]	{crossing,traffic_signal,windy,is.day}	=>	{severity_1}	0.0002322970	0.1491396	0.0015575814	4.074666
[83]	{railway,traffic_signal,mostly_cloudy,is.day}	=>	{severity_1}	0.0001340175	0.1490066	0.0008994065	4.071033
[84]	{amenity,station,fair,is.day}	=>	{severity_1}	0.0001429520	0.1481481	0.0009649261	4.047579
[85]	{crossing,station,is.day}	=>	{severity_1}	0.0014265420	0.1480222	0.0096373486	4.044139
[86]	{crossing,traffic_signal,mostly_cloudy}	=>	{severity_1}	0.0022276690	0.1473311	0.0151201541	4.025256
[87]	{crossing,railway,traffic_signal,mostly_cloudy}	=>	{severity_1}	0.0001429520	0.1472393	0.0009708825	4.022747

Appendix 8 – Association Rules for City Road – Severity 2 (the first 50 rows)

```
city_rules_2 <- subset(city_sorted, subset = rhs %in% c("severity_2") & lift > 1)

inspect(city_rules_2)
```

	lhs <chr>		rhs <chr>	support <dbl>	confidence <dbl>	coverage <dbl>	lift <dbl>	count <int>
[1]	{stop,haze}	=>	{severity_2}	0.0003901399	0.9924242	0.0003931181	1.185398	131
[2]	{stop,haze,is.day}	=>	{severity_2}	0.0002263407	0.9870130	0.0002293189	1.178935	76
[3]	{railway,haze}	=>	{severity_2}	0.0001995372	0.9852941	0.0002025154	1.176882	67
[4]	{crossing,railway,haze}	=>	{severity_2}	0.0001697555	0.9827586	0.0001727337	1.173853	57
[5]	{snow,windy}	=>	{severity_2}	0.0001429520	0.9795918	0.0001459302	1.170071	48
[6]	{railway,haze,is.day}	=>	{severity_2}	0.0001310394	0.9777778	0.0001340175	1.167904	44
[7]	{crossing,junction,railway}	=>	{severity_2}	0.0001250830	0.9767442	0.0001280612	1.166669	42
[8]	{haze,windy}	=>	{severity_2}	0.0002263407	0.9743590	0.0002322970	1.163820	76
[9]	{crossing,railway,haze,is.day}	=>	{severity_2}	0.0001131703	0.9743590	0.0001161485	1.163820	38
[10]	{crossing,station,stop,fair}	=>	{severity_2}	0.0003395110	0.9743590	0.0003484455	1.163820	114

	lhs <chr>		rhs <chr>	support <dbl>	confidence <dbl>	coverage <dbl>	lift <dbl>	count <int>
[1]	{stop,haze}	=>	{severity_2}	0.0003901399	0.9924242	0.0003931181	1.185398	131
[2]	{stop,haze,is.day}	=>	{severity_2}	0.0002263407	0.9870130	0.0002293189	1.178935	76
[3]	{railway,haze}	=>	{severity_2}	0.0001995372	0.9852941	0.0002025154	1.176882	67
[4]	{crossing,railway,haze}	=>	{severity_2}	0.0001697555	0.9827586	0.0001727337	1.173853	57
[5]	{snow,windy}	=>	{severity_2}	0.0001429520	0.9795918	0.0001459302	1.170071	48
[6]	{railway,haze,is.day}	=>	{severity_2}	0.0001310394	0.9777778	0.0001340175	1.167904	44
[7]	{crossing,junction,railway}	=>	{severity_2}	0.0001250830	0.9767442	0.0001280612	1.166669	42
[8]	{haze,windy}	=>	{severity_2}	0.0002263407	0.9743590	0.0002322970	1.163820	76
[9]	{crossing,railway,haze,is.day}	=>	{severity_2}	0.0001131703	0.9743590	0.0001161485	1.163820	38
[10]	{crossing,station,stop,fair}	=>	{severity_2}	0.0003395110	0.9743590	0.0003484455	1.163820	114

	lhs <chr>		rhs <chr>	support <dbl>	confidence <dbl>	coverage <dbl>	lift <dbl>	count <int>
[1]	{stop,haze}	=>	{severity_2}	0.0003901399	0.9924242	0.0003931181	1.185398	131
[2]	{stop,haze,is.day}	=>	{severity_2}	0.0002263407	0.9870130	0.0002293189	1.178935	76
[3]	{railway,haze}	=>	{severity_2}	0.0001995372	0.9852941	0.0002025154	1.176882	67
[4]	{crossing,railway,haze}	=>	{severity_2}	0.0001697555	0.9827586	0.0001727337	1.173853	57
[5]	{snow,windy}	=>	{severity_2}	0.0001429520	0.9795918	0.0001459302	1.170071	48
[6]	{railway,haze,is.day}	=>	{severity_2}	0.0001310394	0.9777778	0.0001340175	1.167904	44
[7]	{crossing,junction,railway}	=>	{severity_2}	0.0001250830	0.9767442	0.0001280612	1.166669	42
[8]	{haze,windy}	=>	{severity_2}	0.0002263407	0.9743590	0.0002322970	1.163820	76
[9]	{crossing,railway,haze,is.day}	=>	{severity_2}	0.0001131703	0.9743590	0.0001161485	1.163820	38
[10]	{crossing,station,stop,fair}	=>	{severity_2}	0.0003395110	0.9743590	0.0003484455	1.163820	114

	lhs <chr>		rhs <chr>	support <dbl>	confidence <dbl>	coverage <dbl>	lift <dbl>	count <int>
[31]	{junction,railway,is.day}	=>	{severity_2}	0.0001042358	0.9459459	0.0001101922	1.129882	35
[32]	{station,stop,fair,is.day}	=>	{severity_2}	0.0002978167	0.9433962	0.0003156857	1.126837	100
[33]	{light_snow,windy}	=>	{severity_2}	0.0009857733	0.9430199	0.0010453366	1.126387	331
[34]	{n/a_precipitation}	=>	{severity_2}	0.0003275984	0.9401709	0.0003484455	1.122984	110
[35]	{patches_of_fog,is.day}	=>	{severity_2}	0.0002263407	0.9382716	0.0002412315	1.120716	76
[36]	{station,fair,windy,is.day}	=>	{severity_2}	0.0002025154	0.9315068	0.0002174062	1.112636	68
[37]	{station,cloudy,windy}	=>	{severity_2}	0.0001191267	0.9302326	0.0001280612	1.111114	40
[38]	{wintry_mix}	=>	{severity_2}	0.0010691620	0.9300518	0.0011495725	1.110898	359
[39]	{haze}	=>	{severity_2}	0.0091310602	0.9274047	0.0098458203	1.107736	3066
[40]	{light_snow,windy,is.day}	=>	{severity_2}	0.0004169434	0.9271523	0.0004497032	1.107434	140

	lhs <chr>		rhs <chr>	support <dbl>	confidence <dbl>	coverage <dbl>	lift <dbl>	count <int>
[41]	{crossing,give_way,stop}	=>	{severity_2}	0.0001131703	0.9268293	0.0001221048	1.107049	38
[42]	{station,cloudy,windy,is.day}	=>	{severity_2}	0.0001131703	0.9268293	0.0001221048	1.107049	38
[43]	{station,heavy_rain}	=>	{severity_2}	0.0001489084	0.9259259	0.0001608210	1.105970	50
[44]	{crossing,station,stop}	=>	{severity_2}	0.0005658517	0.9223301	0.0006135024	1.101675	190
[45]	{station,fair,windy}	=>	{severity_2}	0.0002471879	0.9222222	0.0002680350	1.101546	83
[46]	{crossing,haze}	=>	{severity_2}	0.0006641313	0.9214876	0.0007207164	1.100668	223
[47]	{crossing,give_way,cloudy}	=>	{severity_2}	0.0001042358	0.9210526	0.0001131703	1.100149	35

Appendix 9 – Association Rules for City Road – Severity 3 (the first 50 rows)

```
city_rules_3 <- subset(city_sorted, subset = rhs %in% c("severity_3") & lift >= 2)

inspect(city_rules_3)
```

	lhs		rhs	support	confidence	coverage	lift
	<chr>	<chr>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>
[1]	{station,traffic_signal,overcast}	=>	{severity_3}	0.0001906027	0.4183007	0.0004556596	7.937146
[2]	{station,traffic_signal,scattered_clouds}	=>	{severity_3}	0.0001637992	0.3846154	0.0004258779	7.297977
[3]	{station,traffic_signal,scattered_clouds,is.day}	=>	{severity_3}	0.0001340175	0.3571429	0.0003752490	6.776693
[4]	{amenity,traffic_signal,overcast}	=>	{severity_3}	0.0001042358	0.3431373	0.0003037730	6.510940
[5]	{station,traffic_signal,overcast,is.day}	=>	{severity_3}	0.0001191267	0.3333333	0.0003573800	6.324913
[6]	{station,overcast}	=>	{severity_3}	0.0002829259	0.3074434	0.0009202536	5.833658
[7]	{station,scattered_clouds}	=>	{severity_3}	0.0002203844	0.2771536	0.0007951706	5.258917
[8]	{station,scattered_clouds,is.day}	=>	{severity_3}	0.0001846464	0.2530612	0.0007296509	4.801771
[9]	{amenity,overcast}	=>	{severity_3}	0.0001340175	0.2472527	0.0005420264	4.691557
[10]	{station,overcast,is.day}	=>	{severity_3}	0.0001786900	0.2419355	0.0007385854	4.590663

	lhs		rhs	support	confidence	coverage	lift
	<chr>	<chr>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>
[11]	{traffic_signal,cloudy,windy,is.day}	=>	{severity_3}	0.0002114499	0.2390572	0.0008845156	4.536049
[12]	{amenity,overcast,is.day}	=>	{severity_3}	0.0001161485	0.2378049	0.0004884194	4.512286
[13]	{junction,traffic_signal,is.day}	=>	{severity_3}	0.0001250830	0.2346369	0.0005330919	4.452174
[14]	{traffic_signal,cloudy,windy}	=>	{severity_3}	0.0002203844	0.2208955	0.0009976860	4.191435
[15]	{amenity,station,traffic_signal,mostly_cloudy}	=>	{severity_3}	0.0001340175	0.2112676	0.0006343496	4.008748
[16]	{amenity,station,traffic_signal,mostly_cloudy,is.day}	=>	{severity_3}	0.0001131703	0.2065217	0.0005479827	3.918696
[17]	{stop,overcast}	=>	{severity_3}	0.0001221048	0.2060302	0.0005926552	3.909369
[18]	{junction,traffic_signal}	=>	{severity_3}	0.0001489084	0.2057613	0.0007236946	3.904268
[19]	{traffic_signal,light_snow,is.day}	=>	{severity_3}	0.0001727337	0.2056738	0.0008398431	3.902606
[20]	{railway,station,traffic_signal,is.day}	=>	{severity_3}	0.0001935809	0.2018634	0.0009589698	3.830305

	lhs		rhs	support	confidence	coverage	lift
	<chr>	<chr>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>
[21]	{railway,station,traffic_signal}	=>	{severity_3}	0.0002501660	0.1995249	0.0012538083	3.785934
[22]	{junction,cloudy,is.day}	=>	{severity_3}	0.0001280612	0.1990741	0.0006432841	3.777379
[23]	{junction,clear}	=>	{severity_3}	0.0002263407	0.1968912	0.0011495725	3.735959
[24]	{junction,mostly_cloudy,is.day}	=>	{severity_3}	0.0002144280	0.1951220	0.0010989436	3.702388
[25]	{traffic_signal,overcast}	=>	{severity_3}	0.0011138345	0.1881288	0.0059205961	3.569695
[26]	{station,traffic_signal,mostly_cloudy,is.day}	=>	{severity_3}	0.0005688299	0.1838306	0.0030943156	3.488138
[27]	{station,traffic_signal,mostly_cloudy}	=>	{severity_3}	0.0006820003	0.1817460	0.0037524905	3.448584
[28]	{junction,is.day}	=>	{severity_3}	0.0010721401	0.1813602	0.0059116616	3.441263
[29]	{junction,overcast}	=>	{severity_3}	0.0001131703	0.1792453	0.0006313714	3.401133
[30]	{junction,fair,is.day}	=>	{severity_3}	0.0002174062	0.1767554	0.0012299830	3.353889

	lhs		rhs	support	confidence	coverage	lift
	<chr>	<chr>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>
[31]	{junction,partly_cloudy}	=>	{severity_3}	0.0001429520	0.1758242	0.0008130396	3.336218
[32]	{amenity,station,mostly_cloudy,is.day}	=>	{severity_3}	0.0001310394	0.1746032	0.0007504981	3.313050
[33]	{traffic_signal,overcast,is.day}	=>	{severity_3}	0.0008845156	0.1734813	0.0050986220	3.291763
[34]	{junction,cloudy}	=>	{severity_3}	0.0001518865	0.1722973	0.0008815374	3.269296
[35]	{junction,clear,is.day}	=>	{severity_3}	0.0001518865	0.1717172	0.0008845156	3.258289
[36]	{traffic_signal,windy,is.day}	=>	{severity_3}	0.0007356073	0.1711712	0.0042974951	3.247928
[37]	{amenity,station,mostly_cloudy}	=>	{severity_3}	0.0001518865	0.1683168	0.0009023846	3.193768
[38]	{junction,mostly_cloudy}	=>	{severity_3}	0.0002382534	0.1670146	0.0014265420	3.169059
[39]	{amenity,station,traffic_signal,is.day}	=>	{severity_3}	0.0003752490	0.1660079	0.0022604288	3.149957
[40]	{junction,partly_cloudy,is.day}	=>	{severity_3}	0.0001042358	0.1650943	0.0006313714	3.132622

	lhs		rhs	support	confidence	coverage	lift
	<chr>	<chr>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>
[41]	{railway,station,is.day}	=>	{severity_3}	0.0002293189	0.1648822	0.0013908040	3.128597
[42]	{station,traffic_signal,light_rain}	=>	{severity_3}	0.0001816682	0.1622340	0.0011197908	3.078349
[43]	{station,traffic_signal,light_rain,is.day}	=>	{severity_3}	0.0001459302	0.1611842	0.0009053628	3.058428
[44]	{junction}	=>	{severity_3}	0.0013431533	0.1600426	0.0083924748	3.036766
[45]	{overcast}	=>	{severity_3}	0.0041158269	0.1595290	0.0257998612	3.027022
[46]	{amenity,railway,is.day}	=>	{severity_3}	0.0001042358	0.1590909	0.0006551968	3.018709
[47]	{crossing,traffic_signal,overcast}	=>	{severity_3}	0.0002620787	0.1588448	0.0016499045	3.014038
[48]	{traffic_signal,light_snow}	=>	{severity_3}	0.0002293189	0.1577869	0.0014533455	2.993965
[49]	{traffic_signal,windy}	=>	{severity_3}	0.0007743234	0.1564380	0.0049497136	2.968371
[50]	{crossing,station,light_rain,is.day}	=>	{severity_3}	0.0001101922	0.1561181	0.0007058256	2.962301

Appendix 10 – Association Rules for City Road – Severity 4

```
city_rules_4 <- subset(city_sorted, subset = rhs %in% c("severity_4") & lift >= 2)

inspect(city_rules_4)
```

	lhs		rhs	support	confidence	coverage	lift	count
	<chr>	<chr>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<int>
[1]	{stop,overcast}	=>	{severity_4}	0.0002233625	0.3768844	0.0005926552	5.128429	75
[2]	{stop,overcast,is.day}	=>	{severity_4}	0.0001846464	0.3690476	0.0005003321	5.021791	62
[3]	{stop,scattered_clouds}	=>	{severity_4}	0.0001101922	0.3274336	0.0003365329	4.455531	37
[4]	{stop,clear}	=>	{severity_4}	0.0005033102	0.3243762	0.0015516250	4.413927	169
[5]	{smoke,is.day}	=>	{severity_4}	0.0005866989	0.3182553	0.0018434854	4.330637	197
[6]	{stop,clear,is.day}	=>	{severity_4}	0.0003871617	0.3016241	0.0012835900	4.104330	130
[7]	{amenity,traffic_signal,clear}	=>	{severity_4}	0.0001399739	0.2901235	0.0004824631	3.947835	47
[8]	{clear}	=>	{severity_4}	0.0166568884	0.2819337	0.0590808781	3.836393	5593
[9]	{overcast}	=>	{severity_4}	0.0071714263	0.2779638	0.0257998612	3.782373	2408
[10]	{amenity,clear}	=>	{severity_4}	0.0002501660	0.2616822	0.0009559916	3.560823	84

	lhs		rhs	support	confidence	coverage	lift	count
	<chr>	<chr>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<int>
[11]	{crossing,railway,clear}	=>	{severity_4}	0.0001935809	0.2610442	0.0007415636	3.552141	65
[12]	{smoke}	=>	{severity_4}	0.0007653889	0.2603850	0.0029394509	3.543171	257
[13]	{railway,clear}	=>	{severity_4}	0.0001995372	0.2567050	0.0007773016	3.493096	67
[14]	{clear,is.day}	=>	{severity_4}	0.0115255065	0.2382125	0.0483833020	3.241460	3870
[15]	{overcast,is.day}	=>	{severity_4}	0.0049556700	0.2349619	0.0210913791	3.197228	1664
[16]	{scattered_clouds}	=>	{severity_4}	0.0034606301	0.2268202	0.0152571498	3.086441	1162
[17]	{amenity,clear,is.day}	=>	{severity_4}	0.0001697555	0.2209302	0.0007683671	3.006293	57
[18]	{crossing,railway,clear,is.day}	=>	{severity_4}	0.0001250830	0.2110553	0.0005926552	2.871920	42
[19]	{railway,clear,is.day}	=>	{severity_4}	0.0001280612	0.2087379	0.0006135024	2.840386	43
[20]	{scattered_clouds,is.day}	=>	{severity_4}	0.0027518264	0.2004338	0.0137293501	2.727390	924

	lhs		rhs	support	confidence	coverage	lift	count
	<chr>	<chr>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<int>
[21]	{amenity,overcast}	=>	{severity_4}	0.0001012577	0.1868132	0.0005420264	2.542048	34
[22]	{station,clear}	=>	{severity_4}	0.0003305765	0.1846922	0.0017898784	2.513186	111
[23]	{crossing,clear}	=>	{severity_4}	0.0010721401	0.1755241	0.0061082206	2.388433	360
[24]	{crossing,overcast}	=>	{severity_4}	0.0004139652	0.1713933	0.0024152935	2.332223	139
[25]	{drizzle}	=>	{severity_4}	0.0001072140	0.1706161	0.0006283932	2.321647	36
[26]	{station,traffic_signal,clear}	=>	{severity_4}	0.0001280612	0.1692913	0.0007564544	2.303620	43
[27]	{crossing,traffic_signal,clear}	=>	{severity_4}	0.0006700876	0.1592357	0.0042081501	2.166789	225
[28]	{traffic_signal,clear}	=>	{severity_4}	0.0021353458	0.1587688	0.0134494024	2.160436	717
[29]	{junction,clear}	=>	{severity_4}	0.0001816682	0.1580311	0.0011495725	2.150397	61
[30]	{traffic_signal,overcast}	=>	{severity_4}	0.0009053628	0.1529175	0.0059205961	2.080815	304

	lhs		rhs	support	confidence	coverage	lift	count
	<chr>	<chr>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<int>
[31]	{crossing,clear,is.day}	=>	{severity_4}	0.0007594326	0.1472286	0.0051581853	2.003404	255

Appendix 11 – Initial Logistic Regression Model *accident_severity*

```
accident_glm <- as.data.frame(accident_glm)
# To make our result reproductive
set.seed(12345)
# To create training set
training <- sample(1:nrow(accident_glm), 0.6*nrow(accident_glm))
ycol <- match('severity',colnames(accident_glm))

accidents.training <- accident_glm[training,-ycol]
accidents.training.results <- accident_glm[training,ycol]

# To create a validating set
accidents.test <- accident_glm[-training,-ycol]
accidents.test.results <- accident_glm[-training,ycol]
# Convert to TRUE/FALSE values
accident_glm$severity <- accident_glm$severity > 2
accidents.training.results <- accidents.training.results > 2
accidents.test.results <- accidents.test.results > 2

# Generate LR model with all variables
accident_severity <- glm(severity ~.,
                         family = binomial(link="logit"),
                         data=accident_glm[training,])
summary(accident_severity)
```

```
Call:
glm(formula = severity ~ ., family = binomial(link = "logit"),
    data = accident_glm[training, ])

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7166  -0.5245  -0.4189  -0.3043   2.9022

Coefficients: (10 not defined because of singularities)
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                 -3.667e+00  2.691e-01 -13.626  < 2e-16 ***
temperature                  5.260e-03  3.429e-04  15.340  < 2e-16 ***
humidity                    -5.135e-04  2.945e-04  -1.744 0.081235 .
visibility                   1.603e-02  2.298e-03   6.977 3.02e-12 ***
wind_speed                   3.065e-02  1.027e-03  29.855  < 2e-16 ***
precipitation                6.282e-02  3.023e-02   2.078 0.037731 *
amenity                      2.423e-01  4.932e-02   4.913 8.97e-07 ***
crossing                    -4.726e-01  2.363e-02 -20.002  < 2e-16 ***
give_way                     3.723e-01  9.068e-02   4.105 4.04e-05 ***
junction                     4.022e-01  1.492e-02  26.953  < 2e-16 ***
railway                      3.677e-01  5.198e-02   7.074 1.50e-12 ***
station                     -8.237e-02  3.432e-02  -2.400 0.016387 *
stop                        -5.384e-02  4.948e-02  -1.088 0.276561
traffic_signal               3.636e-01  1.571e-02  23.137  < 2e-16 ***
is.highway                   2.318e-01  1.149e-02  20.172  < 2e-16 ***
light_rain                   5.319e-01  2.670e-01   1.992 0.046381 *
overcast                     1.292e+00  2.697e-01   4.791 1.66e-06 ***
mostly_cloudy                1.331e-01  2.671e-01   0.498 0.618281
snow                         6.851e-01  2.822e-01   2.427 0.015206 *
light_snow                   7.571e-01  2.683e-01   2.822 0.004769 **
cloudy                       2.335e-01  2.669e-01   0.875 0.381600
scattered_clouds             1.044e+00  2.967e-01   3.518 0.000435 ***
clear                        1.099e+00  2.979e-01   3.690 0.000224 ***
partly_cloudy                1.373e-01  2.673e-01   0.514 0.607444
light_freezing_drizzle       1.521e+00  4.052e-01   3.753 0.000174 ***
light_drizzle                8.071e-01  2.747e-01   2.938 0.003304 **
haze                        -6.684e-01  2.754e-01  -2.428 0.015203 *
rain                         4.626e-01  2.688e-01   1.721 0.085230 .
heavy_rain                   3.190e-01  2.727e-01   1.170 0.242042
fair                        -1.259e-01  2.670e-01  -0.472 0.637144
drizzle                      7.100e-01  3.098e-01   2.292 0.021928 *
fog                          1.894e-01  2.703e-01   0.701 0.483394
thunderstorms_and_rain       1.184e+00  3.213e-01   3.684 0.000229 ***
patches_of_fog              -2.742e-01  4.236e-01  -0.647 0.517401
light_thunderstorms_and_rain 1.427e+00  2.939e-01   4.857 1.19e-06 ***
mist                         5.803e-02  4.244e-01   0.137 0.891238
rain_showers                 1.356e+01  1.970e+02   0.069 0.945098
light_rain_showers           1.924e+00  9.544e-01   2.016 0.043819 *
heavy_drizzle                9.067e-01  6.149e-01   1.474 0.140356
smoke                        5.479e-01  2.793e-01   1.962 0.049774 *
light_freezing_fog           1.462e+01  1.132e+02   0.129 0.897208
light_freezing_rain          1.070e+00  3.272e-01   3.271 0.001072 **
blowing_snow                -6.831e-01  5.788e-01  -1.180 0.237908
heavy_thunderstorms_and_rain 1.315e+00  3.204e-01   4.105 4.04e-05 ***
heavy_snow                  -1.634e-04  3.376e-01   0.000 0.999614
snow_grains                         NA         NA      NA       NA
squalls                     -9.509e+00  6.503e+01  -0.146 0.883744
light_fog                           NA         NA      NA       NA
shallow_fog                 -8.620e-01  5.267e-01  -1.637 0.101712
thunderstorm                 8.346e-01  3.756e-01   2.222 0.026297 *
light_ice_pellets            1.143e+00  7.644e-01   1.495 0.134860
thunder                      6.078e-01  2.884e-01   2.107 0.035101 *
thunder_in_the_vicinity      4.661e-01  2.902e-01   1.606 0.108218
windy                       -3.200e-01  3.429e-02  -9.332  < 2e-16 ***
light_rain_with_thunder      1.164e+00  2.858e-01   4.072 4.67e-05 ***
heavy_thunderstorms_and_snow 1.329e+00  1.275e+00   1.042 0.297462
light_snow_showers                  NA         NA      NA       NA
ice_pellets                 -9.479e+00  1.132e+02  -0.084 0.933238
light_thunderstorms_and_snow        NA         NA      NA       NA
storm                        1.172e+00  2.807e-01   4.174 2.99e-05 ***
wintry_mix                   1.858e-01  2.998e-01   0.620 0.535411
heavy_storm                  9.086e-01  2.975e-01   3.054 0.002258 **
sand                        -1.013e+01  1.970e+02  -0.051 0.958970
widespread_dust              2.216e+00  8.104e-01   2.734 0.006254 **
blowing_dust                -6.943e-01  5.409e-01  -1.284 0.199308
volcanic_ash                        NA         NA      NA       NA
freezing_rain                5.620e-01  1.115e+00   0.504 0.614145
small_hail                  -9.263e+00  1.393e+02  -0.067 0.946971
heavy_ice_pellets                   NA         NA      NA       NA
dust_whirls                         NA         NA      NA       NA
showers_in_the_vicinity      1.042e-01  4.137e-01   0.252 0.801204
funnel_cloud                        NA         NA      NA       NA
light_rain_shower            1.072e+00  5.916e-01   1.812 0.069935 .
partial_fog                  1.369e+00  8.595e-01   1.593 0.111115
sleet                       -8.645e+00  7.425e+01  -0.116 0.907303
snow_and_sleet              -8.654e+00  4.501e+01  -0.192 0.847538
light_sleet                  1.252e+00  7.291e-01   1.717 0.085971 .
freezing_drizzle                    NA         NA      NA       NA
drizzle_and_fog              1.768e+00  6.022e-01   2.935 0.003332 **
light_snow_and_sleet         1.874e+00  1.275e+00   1.470 0.141609
rain_shower                  1.608e+00  1.441e+00   1.116 0.264542
dust_whirls_nearby                  NA         NA      NA       NA
heavy_rain_shower            1.938e+00  1.441e+00   1.344 0.178812
thunder_and_hail            -8.872e+00  1.970e+02  -0.045 0.964075
drifting_snow               -8.824e+00  1.970e+02  -0.045 0.964268
light_snow_shower            1.462e+01  1.970e+02   0.074 0.940820
is.day                       7.679e-01  1.271e-02  60.438  < 2e-16 ***
```

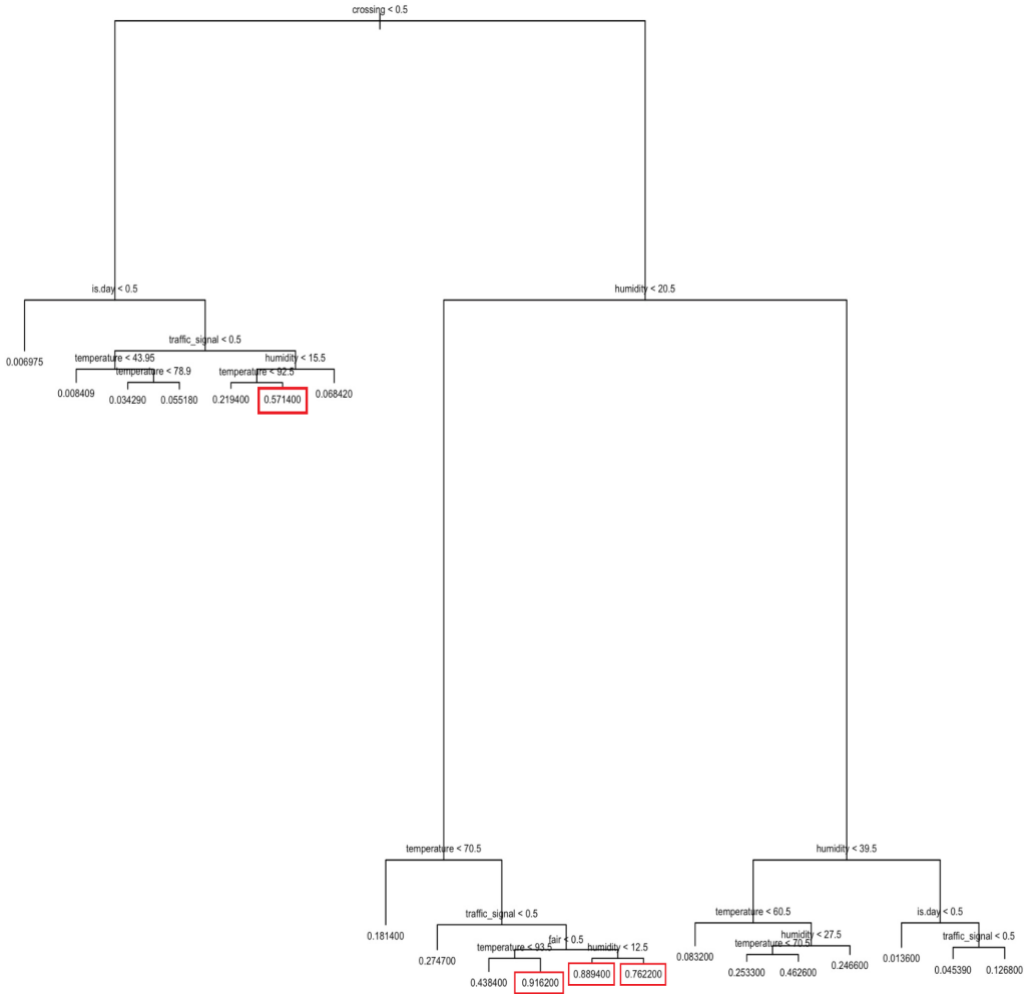
Appendix 12 - Logistic Regression final model - Coefficients

```
accident_severity_1 <- glm(severity ~
temperature+humidity+visibility+wind_speed+amenity+crossing+give_way+junction+railway+station+
traffic_signal+is.highway+windy+is.day+drizzle_and_fog+light_sleet+light_rain_shower+widespread_dust+heavy_storm+storm+light_rain_with_thunder+thunder+thunderstorm+heavy_thunderstorms_and_
rain+light_freezing_rain+smoke+light_rain_showers+light_thunderstorms_and_rain+thunderstorms_and_rain+drizzle+rain+haze+light_drizzle+light_freezing_drizzle+clear+scattered_clouds+light_snow+snow+overcast+light_rain+precipitation,
                        family = binomial(link="logit"),
                        data=accident_glm[training,])

summary(accident_severity_1)
```

```
Coefficients:
                               Estimate Std. Error z value Pr(>|z|)
(Intercept)                  -3.8666354  0.0407071 -94.987  < 2e-16 ***
temperature                   0.0053684  0.0003357  15.992  < 2e-16 ***
humidity                      0.0024086  0.0002693   8.943  < 2e-16 ***
visibility                    0.0136048  0.0020964   6.489 8.62e-11 ***
wind_speed                    0.0348437  0.0010107  34.476  < 2e-16 ***
amenity                       0.2542626  0.0492637   5.161 2.45e-07 ***
crossing                     -0.4674296  0.0235766 -19.826  < 2e-16 ***
give_way                      0.3666613  0.0903818   4.057 4.97e-05 ***
junction                      0.4024689  0.0149095  26.994  < 2e-16 ***
railway                       0.3624607  0.0519287   6.980 2.95e-12 ***
station                      -0.0767513  0.0342729  -2.239 0.025129 *
traffic_signal                0.3700375  0.0156483  23.647  < 2e-16 ***
is.highway                    0.2378313  0.0114255  20.816  < 2e-16 ***
windy                        -0.3487579  0.0342551 -10.181  < 2e-16 ***
is.day                        0.8110198  0.0126142  64.294  < 2e-16 ***
drizzle_and_fog               1.6366127  0.5420449   3.019 0.002533 **
light_sleet                   1.1735066  0.6774543   1.732 0.083233 .
light_rain_shower             0.9531814  0.5284854   1.804 0.071293 .
widespread_dust               2.1511585  0.7652916   2.811 0.004940 **
heavy_storm                   0.7480270  0.1329445   5.627 1.84e-08 ***
storm                         1.0404342  0.0885619  11.748  < 2e-16 ***
light_rain_with_thunder       1.0395246  0.1034641  10.047  < 2e-16 ***
thunder                       0.5219892  0.1119398   4.663 3.11e-06 ***
thunderstorm                  0.7311942  0.2646646   2.763 0.005732 **
heavy_thunderstorms_and_rain  1.1501888  0.1787204   6.436 1.23e-10 ***
light_freezing_rain           0.9514066  0.1911052   4.978 6.41e-07 ***
smoke                         0.5599317  0.0827269   6.768 1.30e-11 ***
light_rain_showers            1.8190724  0.9171018   1.984 0.047311 *
light_thunderstorms_and_rain  1.3075270  0.1239484  10.549  < 2e-16 ***
thunderstorms_and_rain        1.0460657  0.1798513   5.816 6.02e-09 ***
drizzle                       0.5915891  0.1584905   3.733 0.000189 ***
rain                          0.3226807  0.0371179   8.693  < 2e-16 ***
haze                         -0.7072600  0.0692715 -10.210  < 2e-16 ***
light_drizzle                 0.6754530  0.0679259   9.944  < 2e-16 ***
light_freezing_drizzle        1.3983749  0.3060095   4.570 4.88e-06 ***
clear                         1.0341283  0.1325690   7.801 6.16e-15 ***
scattered_clouds              0.9579445  0.1299460   7.372 1.68e-13 ***
light_snow                    0.6321549  0.0369800  17.094  < 2e-16 ***
snow                          0.5270698  0.0955523   5.516 3.47e-08 ***
overcast                      1.1766081  0.0416804  28.229  < 2e-16 ***
light_rain                    0.4104191  0.0189047  21.710  < 2e-16 ***
precipitation                 0.0728129  0.0300628   2.422 0.015434 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 297546  on 442866  degrees of freedom
Residual deviance: 283043  on 442825  degrees of freedom
```

Appendix 13 - Classification Tree

```
# To make sure there was no missing value in our dataset
accident_tree <- accident_glm  %>%
  mutate(severity = ifelse(severity == TRUE, 1, 0)) %>%
  drop_na()

accident_tree <- as.data.frame(accident_tree)

# To reproduce the results
set.seed(12345)

# Randomly partitioning 60% of data points into a training set to build the tree model
tree_training <- sample(1:nrow(accident_tree), 0.6*nrow(accident_tree))
# To identify the dependent variable y-value
ycol <- match('severity', colnames(accident_tree))

# Generate variables that contain the training set
accident_tree_training <- accident_tree[tree_training, -ycol]
accident_tree_training_results <- accident_tree[tree_training, ycol] > 0.5
# Generate the variables that contain the validating set
accident_tree_test <- accident_tree[-tree_training, -ycol]
accident_tree_test_results <- accident_tree[-tree_training, ycol] > 0.5

# Find the most optimum mindev values and the most minimum error rate
best.mindev <- -1
error.rate <- -1
best.error.rate <- 99999999
for (i in seq(from=0.00004, to=0.05, by=0.0005)) {
  accident.tree <- tree(severity ~ ., data=accident_tree[tree_training,], mindev=i)
  accident.tree.proportions <- predict(accident.tree,accident_tree[-tree_training,])
  accident.tree.classifications <- round(accident.tree.proportions,0)
  error.rate <- 1- (sum(accident.tree.classifications == accident_tree_test_results) /
nrow(accident_tree[-tree_training,]))
  if (error.rate < best.error.rate) {
    best.mindev <- i
    best.error.rate <- error.rate
  }
}
print(paste("The optimal value of mindev is",best.mindev,"with an overall error rate
of",best.error.rate))

## [1] "The optimal value of mindev is 0.00053 with an overall error rate of
0.0308162745264589"

DC.best.tree <- tree(severity ~ ., data=accident_tree[training,], mindev=best.mindev)
plot(DC.best.tree)
text(DC.best.tree, cex=0.6)
```


crossing < 0.5
is.day < 0.5
humidity < 20.5
0.006975
traffic_signal < 0.5
temperature < 43.95
humidity < 15.5
temperature < 78.9
temperature < 92.5
0.008409
0.034290
0.055180
0.219400
0.571400
0.068420
temperature < 70.5
humidity < 39.5
0.181400
traffic_signal < 0.5
temperature < 60.5
is.day < 0.5
fair < 0.5
humidity < 27.5
traffic_signal < 0.5
0.274700
temperature < 93.5
humidity < 12.5
temperature < 70.5
0.083200
0.013600
0.438400
0.916200
0.889400
0.762200
0.253300
0.462600
0.246600
0.045390
0.126800

Appendix 14 - Data Errors

year	month	date	count
2020	7	2020-07-01	4
2020	7	2020-07-21	37
2020	7	2020-07-22	5
2020	7	2020-07-27	13
2020	7	2020-07-28	1
2020	7	2020-07-29	4
2020	7	2020-07-30	9
2020	7	2020-07-31	105
2020	8	2020-08-01	1
2020	8	2020-08-02	2
2020	8	2020-08-03	403
2020	8	2020-08-05	1
2020	8	2020-08-06	17
2020	8	2020-08-12	1
2020	8	2020-08-15	1
2020	8	2020-08-16	3
2020	8	2020-08-18	14
2020	8	2020-08-19	17
2020	8	2020-08-25	1
2020	8	2020-08-28	2
2020	9	2020-09-04	1
2020	9	2020-09-05	1
2020	9	2020-09-08	227
2020	9	2020-09-09	144

Appendix 15 - Calculate Mean Square Error for Time-Series predictions

a. The MSE of the **Holt Winter** model is:

```
accident.HWmodel$SSE / nrow(accident.HWmodel$fitted)
```

```
## [1] 405422.5
```

b. The MSE of the **Double Exponential Smoothing** model:

```
accident.DESmodel <- HoltWinters(accident_ts, gamma = FALSE)

accident.DESmodel$SSE / nrow(accident.DESmodel$fitted)
```

```
## [1] 565698.6
```

c. The MSE of the S**ingle Exponential Smoothing** model:

```
accident.SESmodel <- HoltWinters(accident_ts, beta = FALSE, gamma = FALSE)

accident.SESmodel$SSE / nrow(accident.SESmodel$fitted)
```

```
## [1] 529864.5
```

REFERENCE

1. Q, A. (2021, August 25). *Road Safety Facts*. Association for Safe International Road Travel. Retrieved (2021, October 29) from https://www.asirt.org/safe-travel/road-safety-facts/
2. Moosavi, Sobhan. (2021, September 10). *US Accidents (updated)*. Kaggle. Retrieved November 17, 2021, from https://www.kaggle.com/sobhanmoosavi/us-accidents
3. Zipper, D. (2021, November 26). *The Deadly Myth That Human Error Causes Most Car Crashes*. The Atlantic. Retrieved December 10, 2021, from https://www.theatlantic.com/ideas/archive/2021/11/deadly-myth-human-error-causes-most-car-crashes/620808/