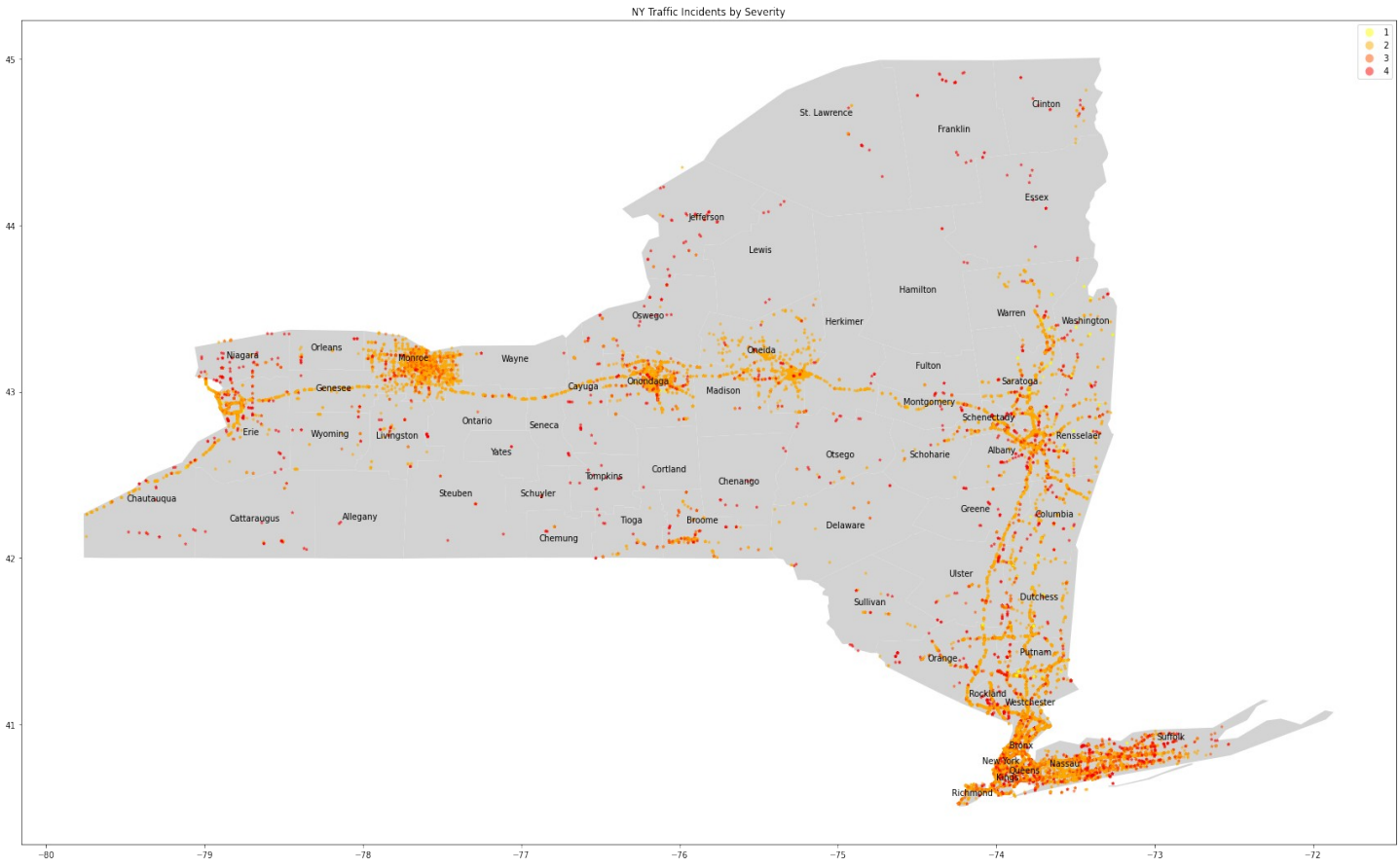
Analysis of New York Traffic Incidents 2019-2020

HOW DO KEY CONTRIBUTORS IMPACT THE DURATION OF TRAFFIC INCIDENTS?



Contents

Executive Summary

Reducing traffic incident occurrence is an important public safety challenge that is at the forefront of transportation industries, government policy and road safety services. Accident analysis and prediction is vital for reducing the frequency of accidents occurring, equally, understanding the most efficient strategies for reducing the hindering affect on traffic flow on major roads by decreasing the duration of an incident when it occurs. To address these challenges, we have conducted in-depth analysis on a dataset that contains traffic incident data and have found some key contributors that cause an accident or increase the likelihood that an incident will occur.

We found different variables that can affect the duration of the traffic incident. After analyzing the data and examining the data visualizations, we found evidence of the following:

- Highest duration of incidents occurs during weekends over weekdays, Autumn (fall) and Winter over the other seasons and night-time in general compared to the rest of the day.
- Higher duration also occurs on the main and auxiliary interstate roads compared to other roads

To have a better understanding, further analysis was conducted. From Levene's & Bartlett's tests it seems the variables were not homogenous and from the Welch test we can state that:

- the difference in traffic incident duration between weekday and weekends is statistically significant, and
- the difference in traffic incident duration between night and day is statistically significant.

We also investigated the effects of the time of year and type of route on the incident durations. The ANOVA results have indicated that:

- the traffic incident duration differs between the seasons with Autumn (fall) being the season with the highest occurrences of traffic incidents, and
- the traffic incident duration differs between the route types.

Introduction

Traffic incidents significantly impact people, businesses and essential services worldwide. These issues weigh heavily on the daily lives of all that are affected, and furthermore, have the potential to cripple the traffic flow of major cities and busy thoroughfares. Both Individuals and governments have an interest in identifying key contributors and relationships relating to the cause and effect of traffic incidents. Gaining an understanding of these key factors allow for behavioral changes relating to road safety within individuals by amplifying caution in high-risk scenarios. Equally, governments can implement prevention strategies, policy and infrastructure to minimize the risk of a traffic incident occurring. Additionally, governments can review current strategies for emergency response procedures to speed up the clean-up process that assist to normalize traffic flow.

Application Programming Interface (API), which is a software intermediary that allows two applications to communicate with one another, are used by governments to collect live data relating to traffic incidences. The dataset that has been analyzed in this report consists of accident data collected in the US State of New York, using multiple APIs that deliver streaming traffic incident data. Law enforcement agencies, the US and state Departments of Transportation, traffic sensors and traffic cameras within the road network are all entities that capture data that is broadcasted by these APIs.

This report analyses traffic incidences that occurred in the State of New York from February 2016 to December 2020, which is a small portion of a larger dataset that contains the traffic incidences of 49 states in the USA (Moosavi, Parthasarathy, Hossein Samavatian & Ramnath, 2019). This dataset was created and made publicly available to allow for large scale analysis into the key contributors that cause accidents to occur, equally, to gain insight into preventative measures, policy and overall road safety. The aim of this report was to focus on the key contributors that impact the duration of an incident and the severity of interruption to traffic flow, to then gain an understanding of what conditions are high-risk and thus conclude when drivers should avoid travel.

Initially, we conducted analysis of the dataset by loading the data frame into a Jupyter Notebook and using Python to explore incident statistics to identify patterns and relationships between them. Once we gained a thorough understanding of the dataset, we isolated variables of interest and conducted the following analyses:

1. Using T-tests and Analysis of Variance (ANOVA), we examined the differences in the mean duration (how long the incident went for) between multiple pairs in the dataset. We examined pairs such as the mean duration for weekdays and weekends, and day and night. We also ran these tests of different seasons and types of routes.
2. We used multiple linear regressions to explore and model the relationship between these variables in the dataset to determine the severity and duration of an incident in different conditions.
3. Developed a predictive model using linear regression and continuous numerical values to understand and identify the conditions that increase the duration of a traffic incident when it occurs. This also allows us to make a prediction of how long it may take to normalize traffic flow after an accident has taken place.

Dataset Description

The dataset that we have worked on contains information about traffic accidents that occurred in New York in 2019 and 2020. The dataset used for this report can be found at https://www.kaggle.com/sobhanmoosavi/us-accidents. The original file contains information about all 50 states in the US, but only the data for New York is used in this report.

The dataset initially had 39,537 records with 48 columns. After examining all of the variables, a few changes were made to the dataset, which is listed below in the data preprocessing section.

Every row of information refers to an accident that occurred in the state of New York, including a description of the incident and various location, weather and other factors that are also listed for analysis.

The variables in this dataset are described in Appendix A, along with their variable types.

Data Preprocessing

Missing Values

The columns that contained the majority of the missing data were 'Number', 'City', and 'Airport_Code'. None of these columns were considered important for the analysis in this report, so it was considered best to drop them and remove any remaining rows that contained missing values. This meant that 2758 rows were removed from the original data set.

New Variables

Some new variables were created to help with the data analysis.

Duration (minutes)

The duration of the incident was calculated by subtracting the ‘End Date Time’ from the ‘Start Date Time’ of the incident.

Adverse Weather Condition

A new variable was created in an attempt to capture the types of bad weather conditions that could affect traffic. These values were extracted from the ‘Weather_Condition’ column in the original data set.

Hour of Day

The hour of the day that the incident occurred, in 24 hour format.

Part of Day

The part of the day the incident occurred, broken into: ‘Early Morning’, ‘Morning’, ‘Late Morning’, ‘Afternoon’, ‘Evening’, and ‘Night’.

Day of Week

The name of the day of the week the traffic incident occurred on.

Part of Week

The part of the week the traffic incident occurred on, i.e. ‘Weekday’ or ‘Weekend’.

Time of Year

Seasons were used in an attempt to differentiate between the different times of the year i.e. ‘Spring’, ‘Summer’, ‘Autumn’, ‘Winter’. The ‘Start_Time’ variable was used to extract the date of the incident and assigned a season name based on the seasonal dates in North America.

Route Type

An attempt was made to extract main and auxiliary interstate roads from the 'Description' variable. Regular expressions were used to match patterns of I-78, I-90N, I-478 etc. Where 2 digit codes (e.g. I-78) are main interstates and 3 digit codes (e.g. I-478) are auxiliary interstate roads. The remaining records where these pattern matches were not found were labeled as 'Other'. An assumption was made that references to interstates using these codes meant that the traffic incident occurred on (or in the vicinity of) that interstate.

Trimming the Data

The new variable for 'Duration (mins)' was selected as the response (or dependent) variable for the analysis (this is discussed in more detail in the next section). On inspection of this new value it was discovered that there were very long durations, sometimes days or weeks in length, so it was decided to trim the data set of records where the duration seemed to be an extreme outlier. Normally, an observation more than three standard deviations away from the mean is used as a rule of thumb for extreme outliers. Using this rule would have meant trimming the data for observations where the duration was longer than around seven hours, however, it was decided to use a more conservative value of ten hours as the cutoff point for trimming the data. This resulted in a further 1,787 records being removed from the data set. Given that the data set has been trimmed of these outliers, any conclusions drawn from the analysis will need to take this into account.

Descriptives

Response Variable (Duration of Traffic Incident)

The new variable for 'Duration (mins)' was selected as the response (or dependent) variable for the analysis. It was considered the best way to measure the impact. Linear regression had been selected as the preferred method for building a predictive model since the goal was to try to find the way that the key contributors to the impact of traffic incidents related to each other. Using a continuous numerical value as opposed to a categorical variable for the response allowed for the use of linear regression.

The distribution of 'Duration (mins)' is provided in the histogram in Figure 1. The data is not normally distributed, this could prove problematic to some of the analysis methods.

	Count	**Mean**	**Std. Dev.**	**Min**	**25%**	**50%**	**75%**	**Max**
Duration (mins)	34992	104.82	92.01	5.5	30	79.11	137.13	599.77

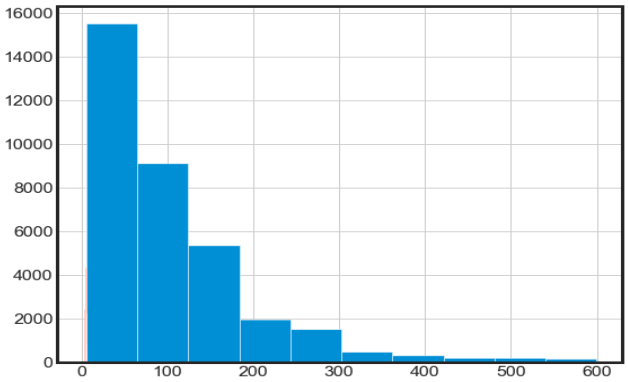


Figure 1.

Numerical Predictors

Detailed descriptive statistics for the numerical predictors have not been provided in the report as none were used in the final analysis, however a correlation matrix has been provided to show the relationships.

A correlation matrix is a table consisting of columns and rows that show the variables, where each cell has a correlation coefficient between the variables (all values are using the Pearson correlation coefficient). These are used to create simple linear relationships between the variables and can measure the strength and direction of the relationship.

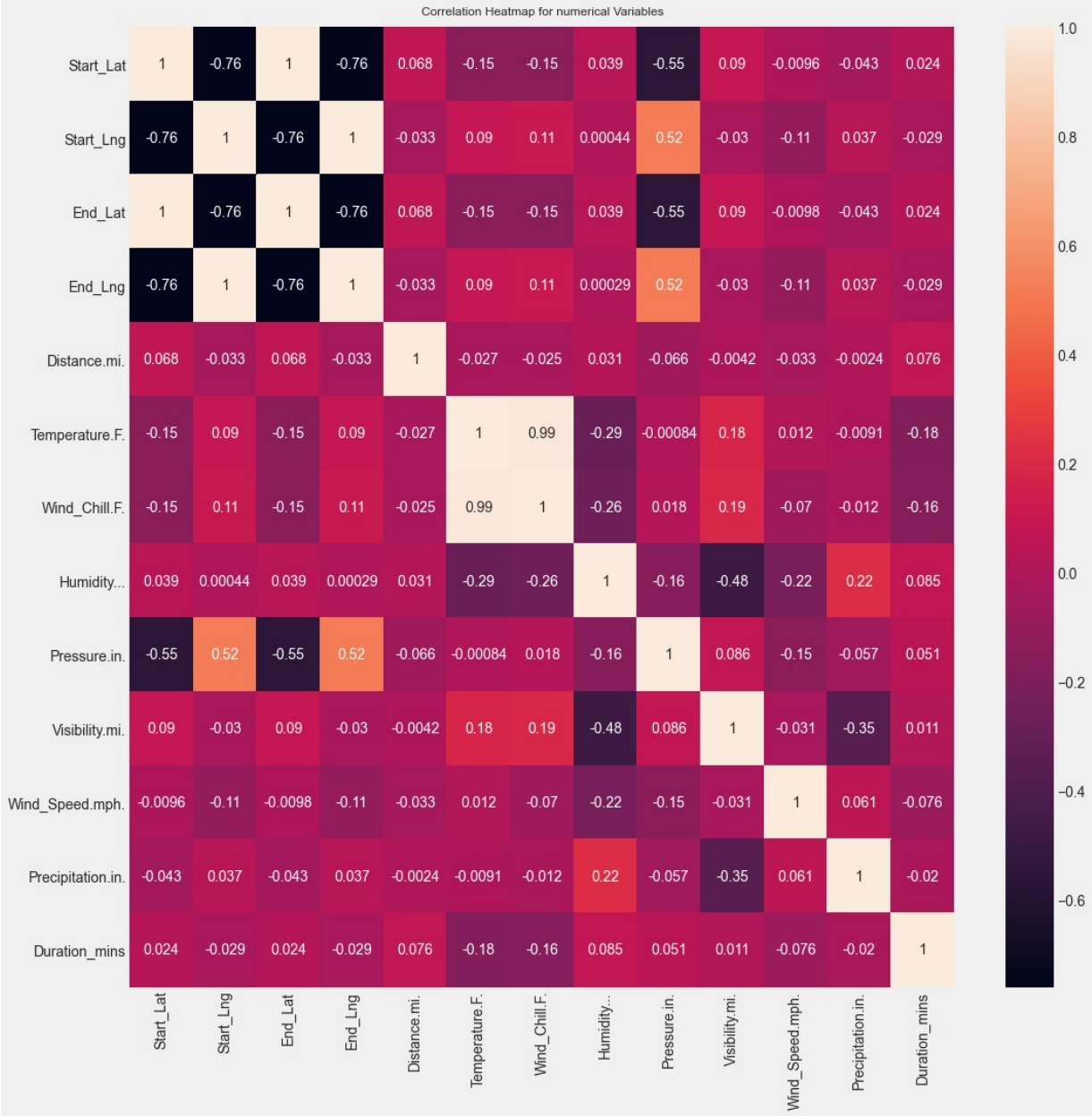


Figure 2.

In Figure 2, we can see a few relationships emerging, namely wind chill with temperature chill (they are highly correlated at 0.99), wind with humidity and temperature, start and end longitude with pressure. A few relationships also can be observed to very weak and thus not likely to related, start latitude and longitude with end latitude and longitude. Pressure also has a weak relationship with start and end latitude and humidity with visibility. However, we can observe that humidity is related to precipitation and visibility is related to temperature and wind chill. Some other notable relationships include temperature with precipitation and duration with distance, temperature and wind chill.

There is very little correlation between traffic incident duration and the other numerical variables. There is an exception to this with temperature and wind chill that have correlation of -

0.18 and -0.16 respectively, so there is some evidence that these variables could have some impact on the duration of a traffic incident.

Categorical Predictors

Adverse Weather Conditions

In Figure 3 we see a comparison of weather conditions that would be considered potentially bad for driving in. There doesn't seem to be much difference in the mean duration of a traffic incident between the different types of weather conditions, with the exception of storms, where duration tends to much lower. Given this plus the lack of correlation between the duration and the weather measurements (shown in the correlation matrix earlier), it appears that it would be more useful to focus on other areas (like the differences in time of day\week\year) in the analysis rather than weather conditions.

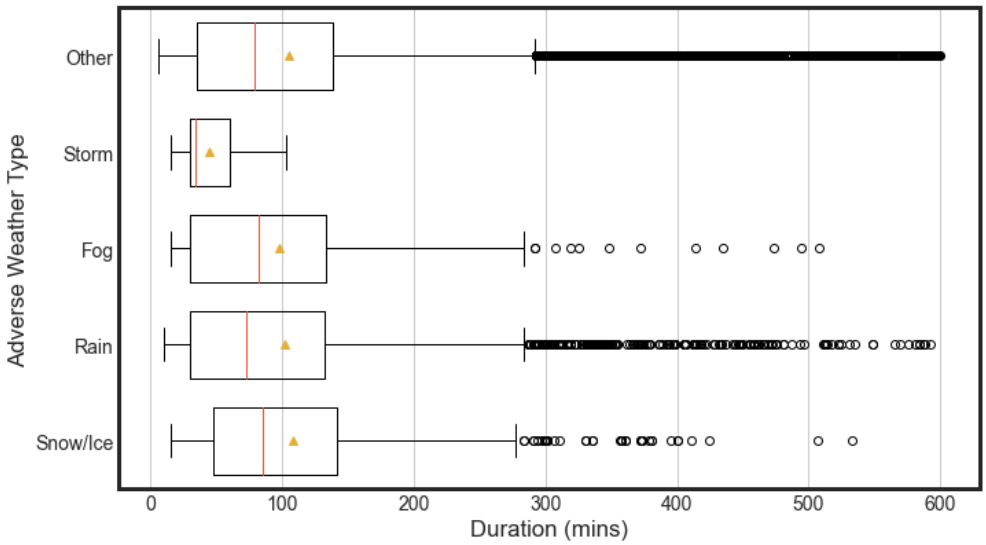


Figure 3.

Location of Incidents

In the tables below we can see the top ten in terms of number of traffic incidents by New York state counties, and by the route type i.e. on an interstate type route (where the code is provided), or other types of roads. ‘Other’ types of routes make up 73% of traffic incidents, but the mean duration of those incidents is much lower than the next two highest in the table, which are the I-87 and I-90 which contribute 11% of traffic incidents combined. We can visually see this in the boxplots in Figure 4, where it appears that the duration of traffic incidents are longer on main and auxiliary interstates than on other types of roads.

The counties of Monroe, Queens and Westchester combined comprise 35% of all traffic incidents.

ROUTE TYPE	MEAN DURATION	N	N %
OTHER	86.82	25590	73

COUNTY	MEAN DURATION	N	N %
MONROE	108.75	5376	15

I-87	155.71	2114	6
I-90	161.14	1720	5
I-95	185.91	1110	3
I-495	157.65	903	3
I-278	153.31	862	2
I-490	111.07	326	1
I-678	186.20	321	1
I-84	125.46	321	1
I-287	126.54	275	1

QUEENS	101.46	3464	10
WESTCHESTER	103.74	3382	10
BRONX	115.97	2628	7
NASSAU	78.69	2518	7
SUFFOLK	77.19	1952	6
ONONDAGA	80.96	1876	5
NEW YORK	127.75	1858	5
ERIE	111.75	1312	4
KINGS	109.99	1260	4

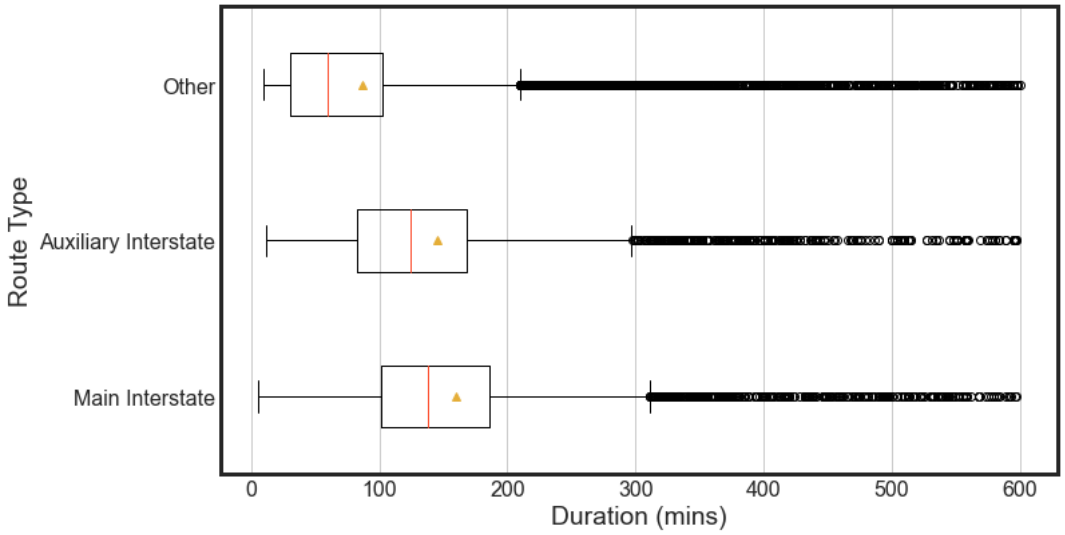


Figure 4.

Date & Time

The idea of focusing on the date\time characteristics of traffic incidents was explored in this section, including the time of day, time of week, and time of year.

Time of Day

Figure 5 portrays a histogram of incidents against one whole day broken down in 24 hours. Based on the plot we can observe low levels of incidents at 0400, 1000 and 2200 and the highest numbers are observed at around 1700. Figure 6 shows the boxplots representing the same data in another graphical form but for the traffic incident duration, where we can observe the mean, median, interquartile range and the outliers. Here we can see that the duration of the traffic incidents at night appear to be longer those during the day.

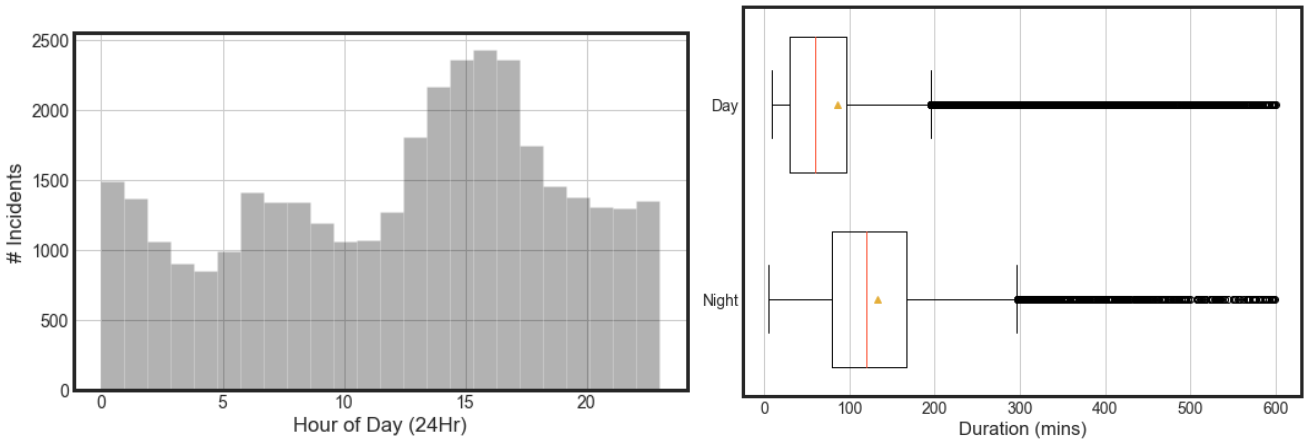


Figure 5.

Figure 6.

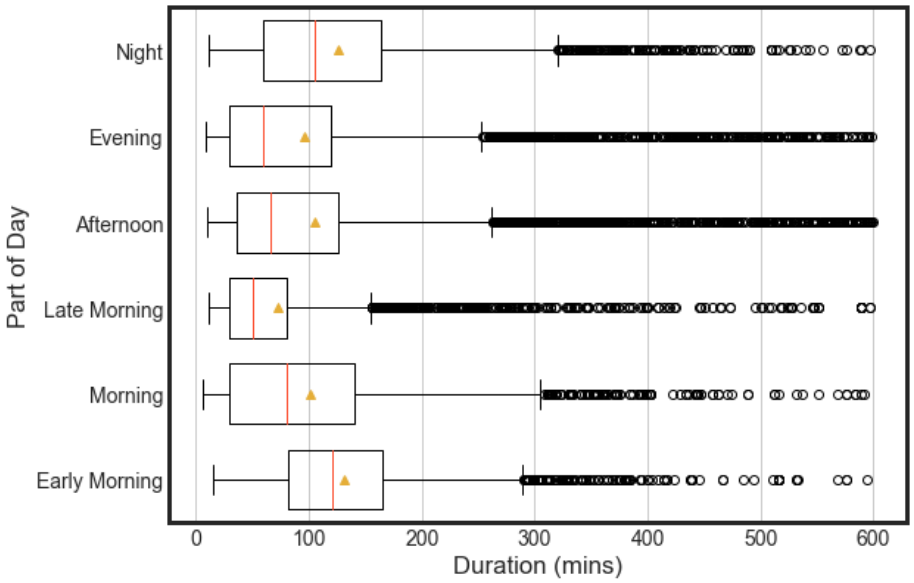


Figure 7.

In Figure 7, we see a breakdown of the incident durations across different parts of the day, where it appears that incidents late at night or early in the morning seem to have longer durations, whereas late morning incidents appear to have the shortest durations.

Time of Week

Based on Figure 8 the number of traffic incidents seem to be higher during the week than on the weekends. However, in Figure 9, the duration of the traffic incidents appears to be longer on the weekend than on a weekday (i.e. Monday – Friday).

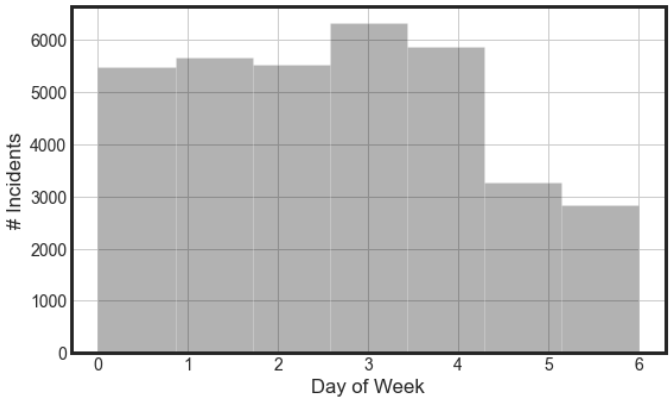


Figure 8.

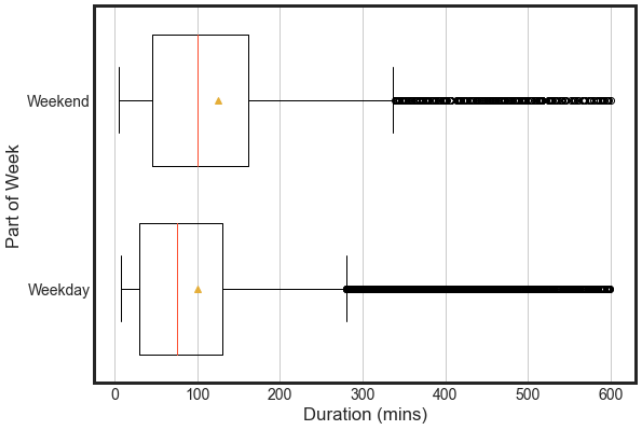


Figure 9.

Time of Year

The plot in Figure 10 shows the number of traffic incidents compared to the mean duration of traffic incidents over the time period January 2019 to December 2020. Here we see some interesting trends. There appears to be a large spike in the number of traffic incidents from March 2020 to July 2020, and then another large spike from October 2020 the end of December 2020. These spikes are quite different compared to the data in 2019. It's possible that there are some issues with the way the data has been collected over this period that might account for these differences. Another possibility is the anomalous nature of the year 2020, where lockdowns and quarantines were put into effect across the state of New York. Is the nature of closing down and reopening through lockdowns causing the spikes we are seeing in Figure 10? Or are there other factors at play here, such as data collection or changing traffic during holiday periods etc.? It's difficult to draw conclusions without more data.

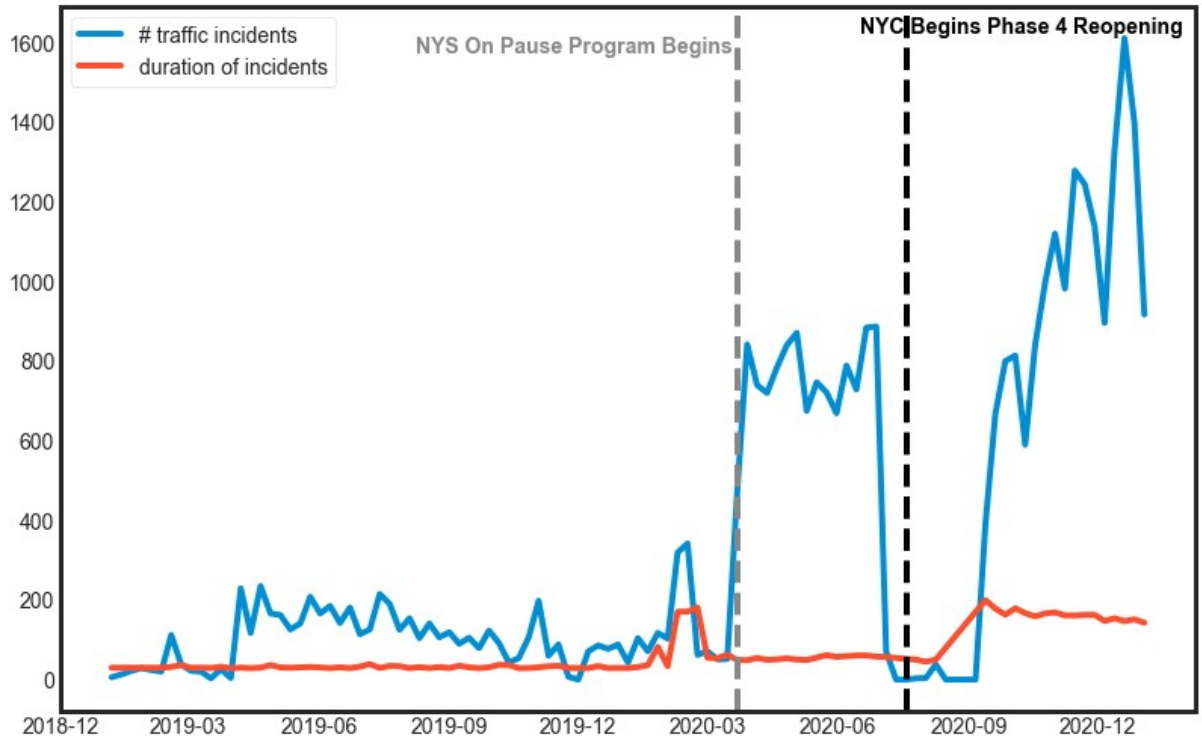


Figure 10.

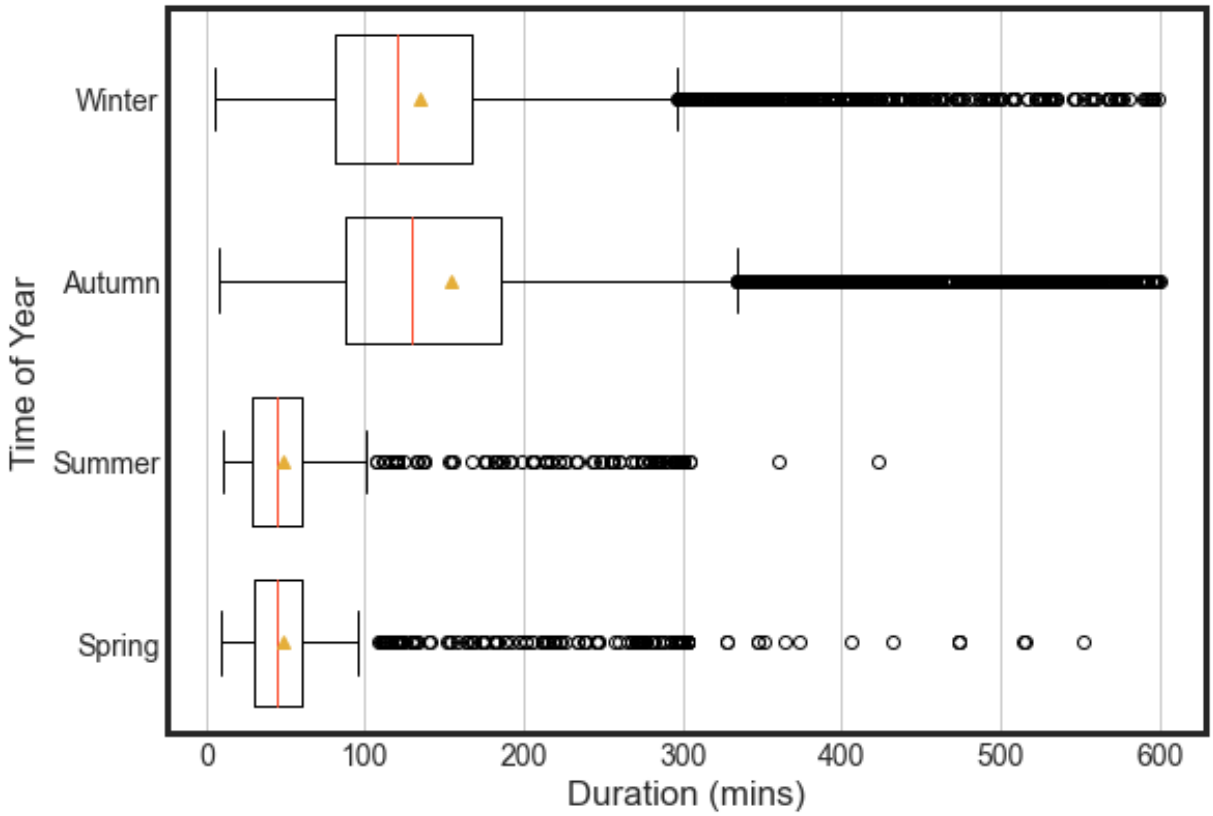


Figure 11.

In Figure 11, we have boxplots for each of the seasons (used to differentiate between different times of year). It appears that there is a difference between the duration of traffic incidents, where it seems to be longer in the colder months than in the warmer months. Is this due to the weather or other factors like, holidays or COVID-19 lockdowns?

Analysis & Results

In this section we cover the analysis that was performed to verify whether there is statistical significance in the differences highlighted in the previous section for the indicators related to date\time periods, and route types. T-tests and ANOVA have been performed to check the statistical significance. And then a linear regression was performed to try to gauge the relative impact of these predictors.

From the observations in Section 5 (descriptive statistics), further tests are required to have an understanding of the traffic incident durations between

- Time of day - day & night were used to represent this
- Time of week - weekdays & weekends were used to represent this
- Time of year – seasons were used to represent this
- Type of route - main and auxiliary interstates were used to represent this

Assumptions for tests:

For all tests the assumption of normality relied on the central limit theorem given the high number of observations (N = 34992), and independence of samples was assumed based on the quality of the data sources.

Is the difference in the mean traffic incident duration between weekdays and weekends statistically significant?

Perform Levene's & Bartlett's test for variance homogeneity:

$$H0: \sigma^2_{duration,weekday} = \sigma^2_{duration,weekend}$$
$$H1: \neg HO$$

The null hypothesis is that the variance of the traffic incident duration is equal between the groups where the incident was on a weekday and when it was on the weekend, and the alternative hypothesis states that it is not equal.

Both tests produced a very small p-value ($p < .001$), suggesting that it is very unlikely that the variances are equal in traffic incident durations between the groups where the incident occurred on a weekday and those on the weekend. Therefore, we can reject the null hypothesis and assume that the variances are not homogeneous.

Given the above, perform a Welch test to compare the means:

The null hypothesis is that the mean of the traffic incident duration is equal between the groups when the traffic incident occurred on a weekday and those on the weekend i.e.

$$H0: \mu_{duration,weekday} = \mu_{duration,weekend}$$
$$H1: \neg\ HO$$

The difference in duration between traffic incidents that occur on weekdays (M = 124.87, SD = 103.31) and those on weekends (M = 100.57, SD = 88.86) is significant (t = 17.11, p < .001).

Is the difference in the mean traffic incident duration between day and night statistically significant?

Perform Levene's & Bartlett's test for variance homogeneity:

$$H0: \sigma^2_{duration,\text{night}} = \sigma^2_{duration,day}$$
$$H1: \neg\ HO$$

The null hypothesis is that the variance of the traffic incident duration is equal between the groups where the incident was during the day and when it was during the night. The alternative hypothesis states that it is not equal.

Both tests produced a very small p-value (p<.001), suggesting that it is very unlikely that the variances are equal in traffic incident durations between the groups where the incident occurred during the day and those during the night. Therefore, we can reject the null hypothesis and assume that the variances are not homogeneous.

Given the above, perform a Welch test to compare the means:

The null hypothesis is that the mean of the traffic incident duration is equal between the groups when the traffic incident occurred during the day and those during the nigh. The alternative hypothesis states that it is not equal.

$$H0: \mu_{duration,night} = \mu_{duration,day}$$
$$H1: \neg\ HO$$

The difference in duration between traffic incidents that occur during the night (M = 133.11, SD = 87.10) and those during the day (M = 86.04, SD = 90.37) is significant (t = 48.76, p < .001).

Is the difference in the mean traffic incident duration between different time's of the year statistically significant?

The null hypothesis is that the mean of the traffic incident duration is equal between the groups where the incident was in spring, summer, autumn or winter i.e.

$$H0: \mu_{duration,spring} = \mu_{duration,summer} = \mu_{duration,autumn} = \mu_{duration,winter}$$
$$H1: \neg\ HO$$

There is as significant relationship between the duration of a traffic incident and the time of year (season) (F = 4619.63, p < .001).

Perform post hoc analysis on time of year:

Both Tukey HSD and Pairwise T-tests were performed but only the results of the Tukey comparison have been shown here as both tests agreed on which groups differed.

group 1	group2	mean diff	p-adj	lower	upper	reject
Autumn	Spring	-105.5013	0.001	-108.221	-102.7816	True
Autumn	Summer	-105.6015	0.001	-108.8922	-102.3107	True
Autumn	Winter	-19.5926	0.001	-22.4997	-16.6856	True
Spring	Summer	-0.1001	0.9	-3.5093	3.309	False
Spring	Winter	85.9087	0.001	82.8683	88.9491	True
Summer	Winter	86.0088	0.001	82.4484	89.5693	True

The post hoc analysis shows that for the duration of a traffic incident, there is a statistically significant difference ($p < .001$) between all of the seasons except for Spring and Summer, where the difference is not considered to be statistically significant.

Is the difference in the mean traffic incident duration between different route types statistically significant?

The null hypothesis is that the mean of the traffic incident duration is equal between the groups where the incident was occurred on a main interstate road, an auxiliary interstate road, or another type of road i.e.

$$H0: \mu_{duration,main\text{-}interstate} = \mu_{duration,auxiliary\text{-}interstate} = \mu_{duration,other}$$
$$H1: \neg HO$$

There is as significant relationship between the duration of a traffic incident and the type of road ($F = 2069.36$, $p < .001$).

Perform post hoc analysis on type of route:

Both Tukey HSD and Pairwise T-tests were performed but only the results of the Tukey comparison have been shown here as both tests agreed on which groups differed.

group1	group2	mean diff	p-adj	lower	upper	reject
Auxiliary Interstate	Main Interstate	14.7389	0.001	10.4611	19.0166	True
Auxiliary	Other	-	0.00	-	-	True

Interstate		58.2659	1	61.7925	54.7393	
Main Interstate	Other	-73.0048	0.001	-76.0236	-69.9861	True

The post hoc analysis shows that for the duration of a traffic incident, there is a statistically significant difference ($p < .001$) between all of the route types.

Predictive Modeling Using Linear Regression

Given the results we saw from the previous analysis, it was decided to use the following predictors for the linear regression:

- Dummy variables were created for 'Time of Year' (i.e. season), but only Autumn and Winter were used in the regression, since they seemed to have the bigger impact according to the boxplots in Figure 11.
- Dummy variables were created for 'Time of Week' i.e. 'Weekend' and 'Weekday'.
- Dummy variables were created for 'Time of Day' based on the variable 'Civil_Twilight' i.e. 'Day' and 'Night'.
- Dummy variables were created for 'Type of Route' i.e. 'Route_I_Main' (main interstate), 'Route_I_Aux' (auxiliary interstate), and 'Other'.

The response variable for 'Duration (mins)' was log transformed to help meet the assumptions for linear regression (see Appendix B for more details).

A multiple linear regression was calculated to predict the mean duration of traffic incidents in New York state for the years 2019 and 2020. A significant regression equation was found ($F(6, 34985) = 4295$, $p < .001$), with an R^2 of .424.

$$\text{Log(Duration)} = \beta_0 + \beta_1 \times \text{Autumn} + \beta_2 \times \text{Winter} + \beta_3 \times \text{Weekend} + \beta_4 \times \text{Night} + \beta_5 \times \text{Main-Interstate} + \beta_6 \times \text{Auxiliary-Interstate}$$

	Default			
Independent Variables	**B**	**SE**	**t-value**	**Sig.**
Dummy: Autumn	.904	.009	102.426	<.001***
Dummy: Winter	.770	.010	79.873	<.001***
Dummy: Weekend	.038	.009	4.476	<.001***
Dummy: Night	.077	.008	10.110	<.001***
Dummy: Route_I_Main	.299	.010	30.911	<.001***
Dummy: Route_I_Aux	.223	.011	20.519	<.001***
Constant	3.75	.005	743.169	<.001***
	N = 34992			
	R^2 = .424			

All of the predictors appear to be statistically significant, all with $p < .001$. Since the response was log transformed, the coefficients will be reported as a percentage increase in the response (see references on interpreting log transformation in linear regression for the information used to determine this).

The intercept is reported as 3.75, so if all of our predictors are not true, then the average duration of a traffic incident is approximately: $e^{3.75} = 42$ (minutes).

- If the traffic incident occurs **in autumn**, the mean duration (mins) of an incident is predicted to **increase** by approximately **147%** [142.78% 151.18%].
- If the traffic incident occurs **in winter**, the mean duration (mins) of an incident is predicted to **increase** by approximately **116%** [112.12% 120.12%].
- If the traffic incident occurs **on the weekend**, the mean duration (mins) of an incident is predicted to **increase** by approximately **4%** [2.22% 5.65%].
- If the traffic incident occurs **at night**, the mean duration (mins) of an incident is predicted to **increase** by approximately **8%** [6.4% 9.53%].
- If the traffic incident occurs **on a main interstate road**, the mean duration (mins) of an incident is predicted to **increase** by approximately **35%** [32.31% 37.44%].
- If the traffic incident occurs **on an auxiliary interstate road**, the mean duration (mins) of an incident is predicted to **increase** by approximately **25%** [22.38% 27.76%].

Discussion & Conclusions

This study used 2019-2020 New York Accidents dataset to explore the variables that might be associated with the traffic incident durations in New York state.

We found different variables that can affect the duration of the traffic incident. After analyzing the data and examining the data visualizations, we found evidence of the following:

- Highest duration of incidents occurs during weekends over weekdays, Autumn and Winter over the other seasons and night-time in general compared to the rest of the day.
- Higher duration also occurs on the main and auxiliary interstate roads compared to other roads

To have a better understanding, further analysis was conducted. From Levene's & Bartlett's tests it seems the variables were not homogenous and from the Welch test we can state that:

- the difference in traffic incident duration between weekday and weekends is statistically significant, and
- the difference in traffic incident duration between night and day is statistically significant.

We also investigated the effects of the time of year and type of route on the incident durations. The ANOVA results have indicated that:

- the traffic incident duration differs between the seasons, and
- the traffic incident duration differs between the route types.

A multiple linear regression was conducted to predict the mean duration of traffic incidents and provide a gauge for the impact of each predictor. From the results we can see that night and weekday incidents have the smallest impact on the mean duration of traffic incidents relative to the other predictors. The type of road has a much bigger impact, with incidents that have occurred on main interstate roads having duration times roughly 4-8 times larger relative to nights or weekends. However, the biggest impact by far, appears to be related to the time of year.

Given that this large difference in incident durations exists depending on the time of year, this raises the question of why? What is contributing to this pattern? It could be that the weather has more of an impact on incident duration in colder months (we did see some correlation with temperature in the correlation matrix). Or is it related to events like school and public holiday periods? Or is the year 2020 anomalous due to the effects of lockdowns throughout the year that may be skewing the data? According to Investopedia's Covid-19 Timeline (2021), New York was in lockdown from March 2020 and slowly reopened from July 2020. Could these lockdowns have created some abnormalities in the data? (see Figure 10).

Since the impact of lockdowns due to the pandemic is possibly affecting this dataset, it's difficult to draw any firm conclusions based on the data available. Data for other years (past or future) would need to be added for any further analysis to help investigate this pattern in the data.

It should be safe to assume that traffic incidents on interstate roads do have a large impact on the incident duration times. This is one area where it should be safe to advise the New York State Department of Transportation. It could help with how they organize and deploy their resources and perhaps invest in research in how to lessen the interstate rode incident duration times. If the time of year patterns revealed in the analysis turn out to be unaffected by the pandemic lockdowns in 2020, we would also advise that there should be further investigation into why there is such a large difference in traffic incident duration times depending on the time of year.

References

Moosavi, S., Parthasarathy, S., Hossein Samavatian, M., & Ramnath, R. (2019). A Countrywide Traffic Accident Dataset [Ebook]. Ohio: The Ohio State University. Retrieved from https://arxiv.org/abs/1906.05409

Kerr, A., (05 April, 2021). A Historical Timeline of COVID-19 in New York City. Investopedia. Retrieved 13 November, 2021, from https://www.investopedia.com/historical-timeline-of-covid-19-in-new-york-city-5071986

Clay Ford (2018). Interpreting Log Transformations in a Linear Model. University of Virginia Library. Retrieved from https://data.library.virginia.edu/interpreting-log-transformations-in-a-linear-model/

Appendix A

The variables and their description in the original dataset. This does not include variables added as part of this report, details of any added variables are provided in Data Preprocessing – New Variables – Section.

Variable	**Description**	**Variable Type**
ID	Unique Identifier for the incident	Numerical (continuous)
Severity	The severity of the crash denoted by a number	Categorical (nominal)
Start_Time	The start time of the incident	Numerical (continuous)
End_Time	The end time of the incident	Numerical (continuous)
Start_Lat	The latitude of the crash during the start	Numerical (discrete)
Start_Lng	The longitude of the crash during the start	Numerical (discrete)
End_Lat	The latitude of the crash during the end	Numerical (discrete)
End_Lng	The longitude of the crash during the end	Numerical (discrete)
Distance.mi.	Distance given in miles	Numerical (discrete)
Description	Description of the crash	
Number	Street Number where the crash occurred	Categorical (nominal)
Street	Street Name where the crash occurred	Categorical (nominal)
Side	The side of the car which crashed	Categorical (Binary)
City	City where the crash occurred	Categorical (nominal)
County	County where the crash occurred	Categorical (nominal)
State	State where the crash occurred	Categorical (nominal)

Zipcode	Zipcode where the crash occurred	Categorical (nominal)
Country	Country where the crash occurred	Categorical (nominal)
Timezone	Timezone, when the crash occurred	Categorical (nominal)
Airport_Code	Airport identifier code	Categorical (nominal)
Weather_Timestamp	The time of the accident	Numerical (continuous)
Temperature.F.	Temperature denoted in Fahrenheit	Numerical (discrete)
Wind_Chill.F.	Wind Chill denoted in Fahrenheit	Numerical (discrete)
Humidity...	Humidity in the air	Numerical (discrete)
Pressure.in.	The atmospheric pressure denoted in inches	Numerical (discrete)
Visibility.mi.	Visibility during that time denoted in miles	Numerical (discrete)
Wind_Direction	The direction of the wind at the time of the	Categorical (nominal)
Wind_Speed.mph.	The speed of the wind denoted in miles per hour	Numerical (discrete)
Precipitation.in.	The rate of precipitation or rainfall denoted in	Numerical (discrete)
Weather_Condition	Weather condition during the time of the	Categorical (nominal)
Amenity	Type of incident, denoted by true or false	Categorical (Binary)
Bump	Type of incident, denoted by true or false	Categorical (Binary)
Crossing	Type of incident, denoted by true or false	Categorical (Binary)
Give_Way	Type of incident, denoted by true or false	Categorical (Binary)
Junction	Type of incident, denoted by true or false	Categorical (Binary)
No_Exit	Type of incident, denoted by true or false	Categorical (Binary)
Railway	Type of incident, denoted by true or false	Categorical (Binary)
Roundabout	Type of incident, denoted by true or false	Categorical (Binary)
Station	Type of incident, denoted by true or false	Categorical (Binary)
Stop	Type of incident, denoted by true or false	Categorical (Binary)
Traffic_Calming	Type of incident, denoted by true or false	Categorical (Binary)
Traffic_Signal	Type of incident, denoted by true or false	Categorical (Binary)
Turning_Loop	Type of incident, denoted by true or false	Categorical (Binary)
Sunrise_Sunset	Astronomical features denoted by day or night	Categorical (Binary)
Civil_Twilight	Astronomical features denoted by day or night	Categorical (Binary)
Nautical_Twilight	Astronomical features denoted by day or night	Categorical (Binary)
Astronomical_Twilight	Astronomical features denoted by day or night	Categorical (Binary)

Appendix B

Check Assumptions for Linear Regression

A log transform of the response was required to help meet the assumption of normality, even so, the normal Q-Q plot in Figure 12 still looks a little problematic. The residuals vs fitted plot in Figure 12 also looks problematic, even after trimming the data there still appears to be an outlier problem that may be affecting linearity and homoscedasticity. But this model may be 'good enough' for the purposes of the analysis, which is to attempt to find the relative effects of different variables on the duration of traffic incidents. However, choosing a different analysis method may allow for a more robust interpretation of the results. In Appendix C a different approach using decision trees is shown which roughly reflects similar results achieved in the linear regression above.

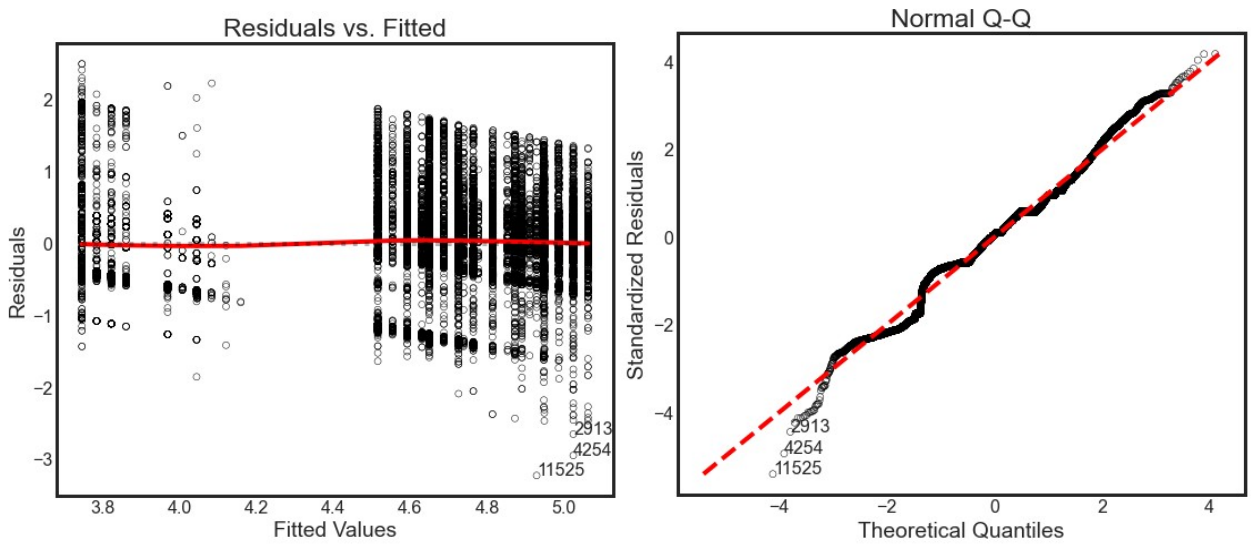


Figure 12

Appendix C

While not exactly the same, we can see quite a similar pattern to the impact on incident duration for certain predictors that we saw in the linear regression. This provides some extra confidence that while the assumptions for linear regression may have been a little problematic, it still seems to do a reasonable job of estimating the impact of the predictors on incident duration. So we can get a fairly good feel for what is having the bigger effect on the duration of traffic incidents.

Results of the decision tree are summarized below (and can also be seen in Figure 13):

				Mean Duration (mins)
Roo				**104.**
	Summer\Spring			**78.4**
		Summer\Spring		**49.2**
			Weekend	**46.5**
			Weekday	**49.5**
		Winter		**135.**
			Not a Main	**129.**
			Main Interstate	**151.**
	Autumn			**154.**
		Not a Main		**148.**
			Weekend	**143.**
			Weekday	**163.**
		Main Interstate		**170.**
			Night	**162.**
			Day	**182.**

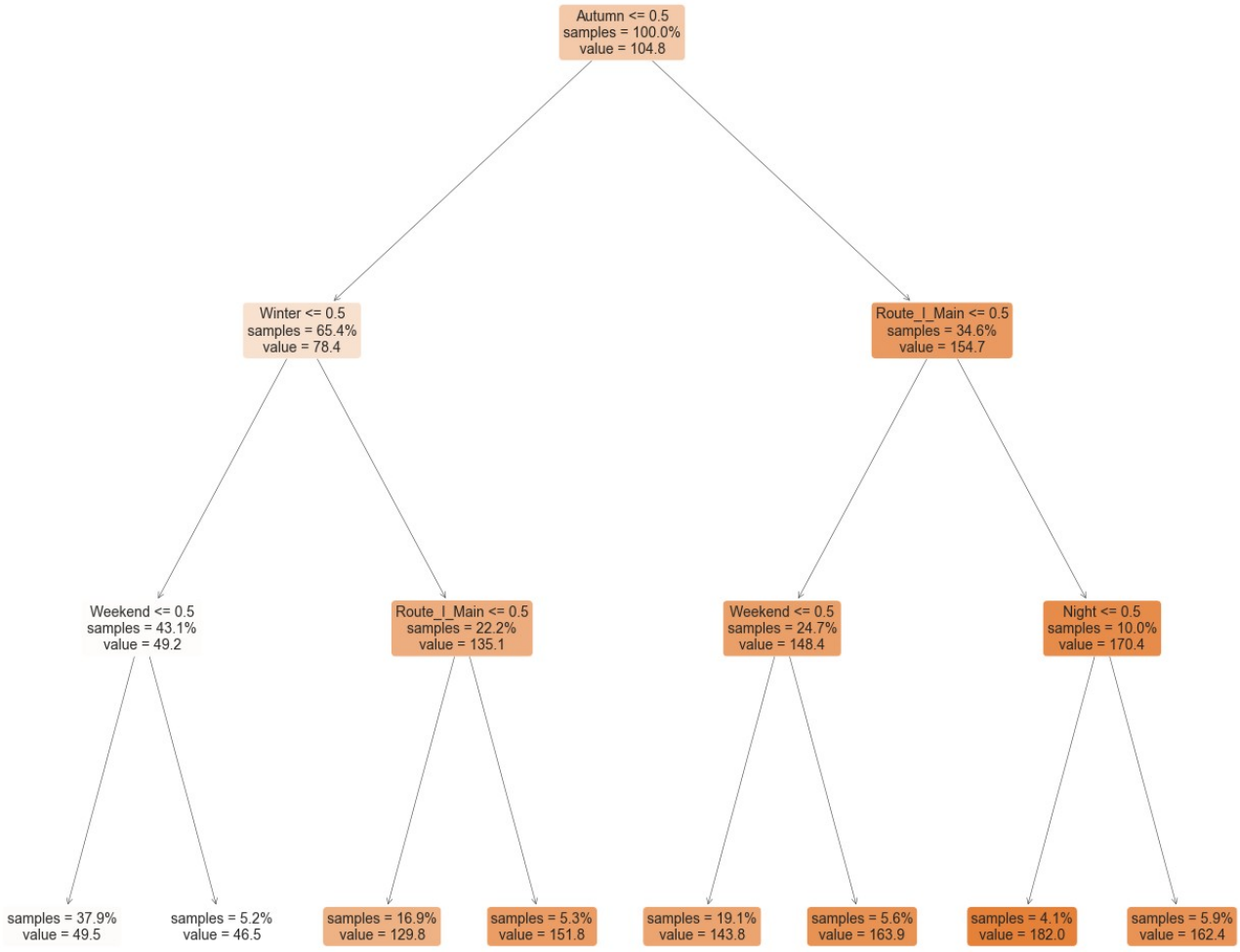

Autumn <= 0.5
samples = 100.0%
value = 104.8
Winter <= 0.5
samples = 65.4%
value = 78.4
Route_I_Main <= 0.5
samples = 34.6%
value = 154.7
Weekend <= 0.5
samples = 43.1%
value = 49.2
Route_I_Main <= 0.5
samples = 22.2%
value = 135.1
Weekend <= 0.5
samples = 24.7%
value = 148.4
Night <= 0.5
samples = 10.0%
value = 170.4
samples = 37.9%
value = 49.5
samples = 5.2%
value = 46.5
samples = 16.9%
value = 129.8
samples = 5.3%
value = 151.8
samples = 19.1%
value = 143.8
samples = 5.6%
value = 163.9
samples = 4.1%
value = 182.0
samples = 5.9%
value = 162.4

Figure 13