Applied Probability and Statistics
DEC 518Q-00A
Team A8: Yash Raj, Elias Melul, Wentao Yu, Yicen Song, Rosemary Ahn
Final Project

The Relationship between the Crime Rate and Weather in Boston

Introduction

Previous research has demonstrated the significant role weather plays on the occurrence of crime-related events. This study analyzes the weather and crime incidents in Boston, Massachusetts to find a possible correlation between weather and crime rate. The aim from gathering data and carrying out the results is to inform the city and especially the police department on the distinct locations and other important details to take note of at certain times of the year. This analysis on the relationship between the weather and the criminal incidents in Boston will add benefit to the Boston Police Department. Although, according to the Federal Bureau of Investigation's Uniform Crime Report, the crime rate has been steadily declining between 1999 and 2017, a study like this will help continue to decrease the crime rate by providing the police department with detailed information like the effects of temperature changes and the significance in knowing the specific locations. As a result, the Boston Police Department may be able to allocate their resources more strategically and effectively based on the time of the year.

Materials

The study used two datasets provided by Kaggle: the Boston Weather Report from January 2014 to April 2018, and the Boston Crime Report from January 2015 to October 2018. Since the study requires the use of both datasets to compare the two variables based on the time of the year, both datasets need to be matched by date. Therefore, the reports recorded between January 2015 to April 2018 will be used to carry out the analysis.

The Boston Weather Report contains data on contents such as the daily average temperature, humidity level, and precipitation count, and the Boston Crime Report is comprised of data on the time of the incidents, the offense types, and the location of the incidents, among others. In this study, we have extracted data on the criminal incidents to analyze it with the corresponding weather attributes on those days. Since the aim is to find the correlation between weather and crime rates in Boston, the target variables will be the location, date, crime rate, and weather (temperature).

Exploratory Data Analysis

Due to the nature of our data, we decided to explore the datasets in terms of the frequency of the number of daily crimes, categorical variables, such as the offense group and the day of the week, weather predictors and time series.

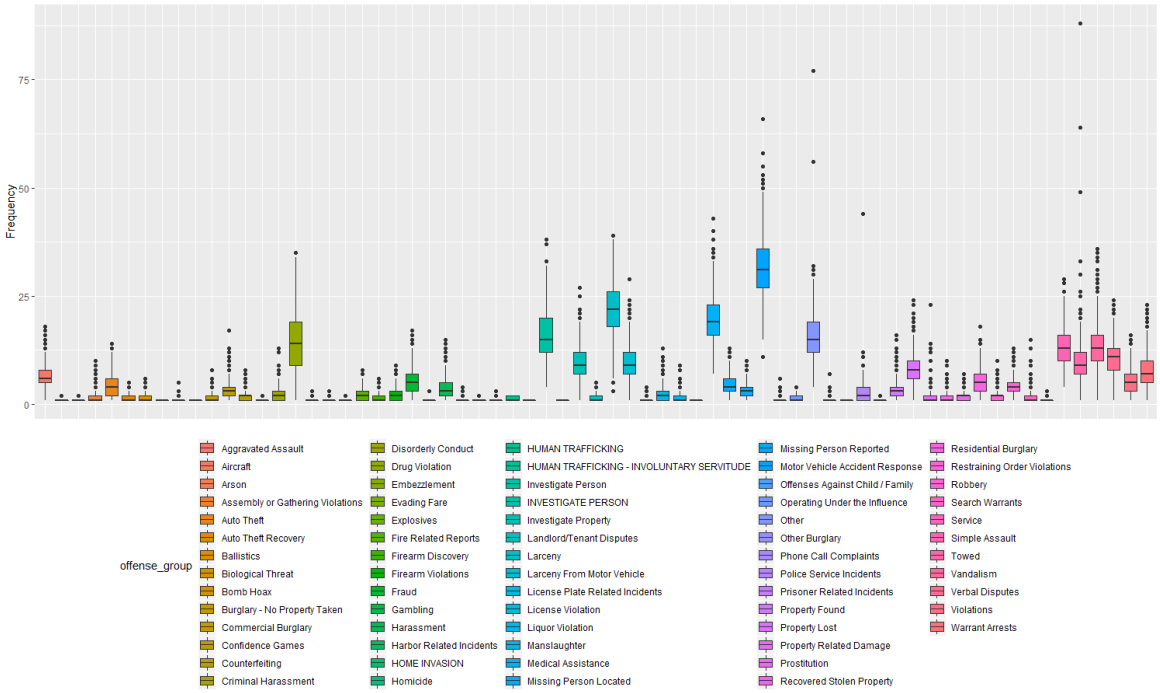


Figure 1. Boxplot on the Frequency for Each Offense Group

As seen on Figure 1, a boxplot was constructed to display the frequencies for each offense category listed in the dataset. There are 67 categories of crime. The outliers observed for all the crimes are seen to be higher in frequency, and the crimes that tend to be more common have a much larger frequency than the mean crime frequency by type. We can observe that most crimes can be represented by a small number of offenses. More specifically, the top 15 most frequent offense group represent 75% of the total crimes, as can be observed from the Pareto Chart in the Appendix.

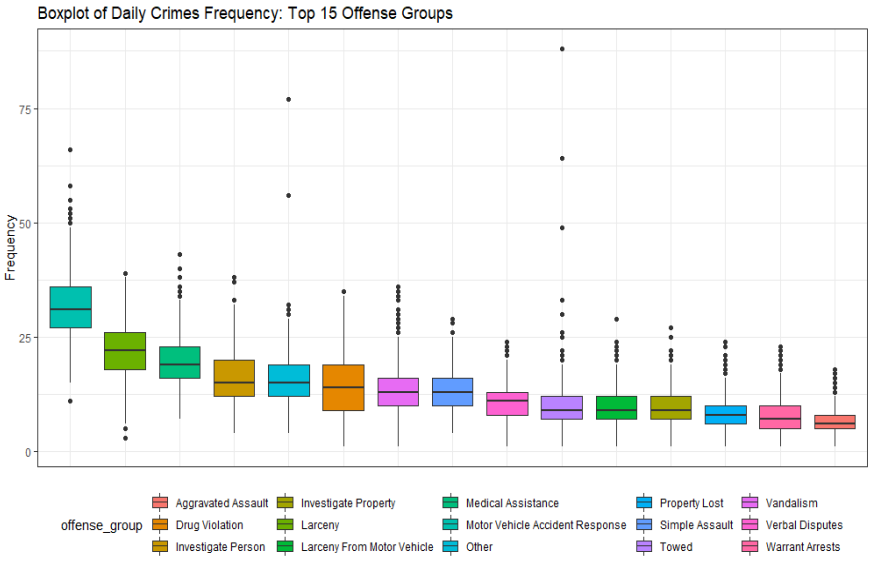


Figure 2. Boxplot on the Frequency for the Top 15 Offense Group

Figure 2 takes a closer look into the 15 highest offense groups and their daily crime frequency as displayed on Figure 1. As seen on this figure, Motor Vehicle Accident Response is by far the most common crime, followed by Larceny, Medical Assistance, Investigate Person and Other. Note the high-leverage outliers of Towed and Other.

Taking a closer look to the data, we can determine when the extraordinary outliers of the Towed and the Other offense groups occurred and whether there is time-wise trend and seasonality in the model.

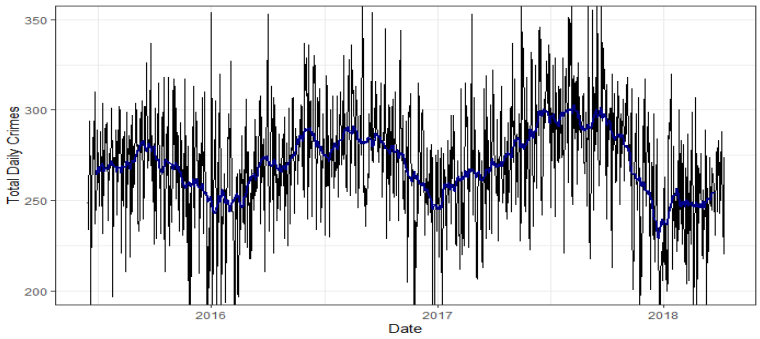


Figure 3. Time Series on the Frequency of Crimes Per Day

From the time-series represented in Figure 3 of total daily crimes (black) and a 30-day moving average (blue), we can observe a clear yearly seasonality with the trough during winter and the peak during summer. However, the variance of the total daily crimes is very high day to day. It is possible that there is another seasonality that accounts for such variance.

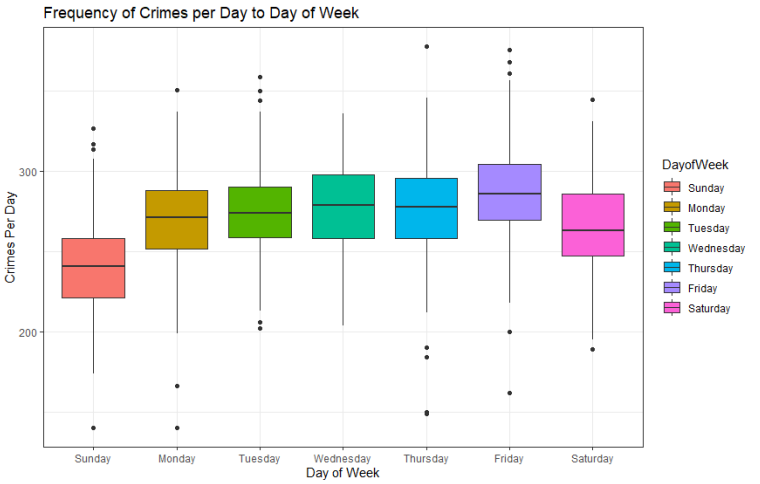


Figure 4. Boxplot on the Frequency of Each Offense Group

Crime during the weekends is significantly lower than that in weekdays. It increases steadily from Monday to Friday and drops suddenly in Saturday and even further on Sunday. Reasons can be criminals would like to rest in weekends or people are at home during weekends, so it is harder to commit crimes. The lowest crime per day is approximately 220 and the highest is approximately 305.

The range of crime per day is about 85. More specifically, Tuesday has smallest range and Wednesday has the largest ones. The ranges of weekends do not deviate lot from those of weekdays. Wednesday has no outliers and other days have 2-4 outliers (Figure 6). We investigated whether this pattern is followed for the same statistics but subdivided into offense groups, as shown in the Appendix, and concluded that this pattern is followed. That is, in weekends there are fewer crimes and in weekdays there are more. However, this trend is the least evident in the motor vehicle accident, which may be due to traffic accidents happen on the road instead of indoor. In addition, the ACF plot in the Appendix confirms that there is a 7-day seasonality.

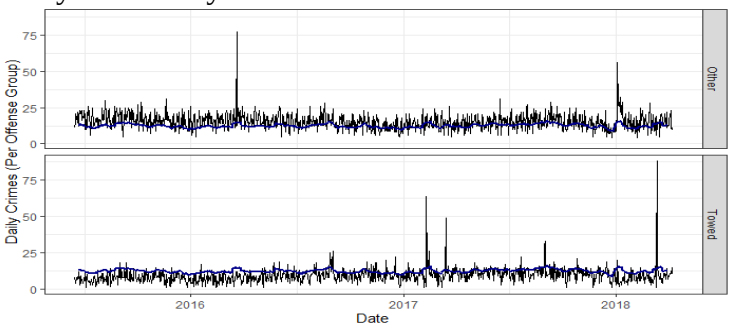


Figure 5. Daily Crimes time-series for 'Other' and 'Towed'

For the 'Other' offense group:
1) Freq: 77; March 20th 2016
   a) Red-Sox versus Mets
   b) St. Patrick's Day Parade
2) Freq: 56, Jan. 4th 2018
   a) Blizzard
3) Freq: 32, Jan. 8th 2018
   a) Blizzard

For the 'Towed' offense group:
1) Freq: 88; March 13th 2018
   a) Blizzard + Outages
2) Freq: 64, Feb. 9th 2017
   a) Blizzard + Outages
3) Freq: 49, March 14th 2017

As we can observe, blizzards present opportunities for different crimes to occur sporadically. The Boston Police Department (BPD) should focus efforts investigating what effects of blizzards cause different types of robberies.

When comparing weather parameters to the crime frequency, we first considered a correlation plot between frequency and the weather parameters.

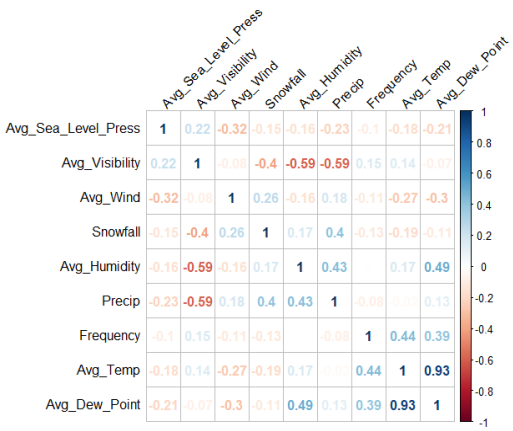


Figure 6. Correlation Plot

The correlation plot depicted in Figure 6 shows that the only parameters that correlate to the frequency of crimes are average temperature (F) and average dew point (F), both of which are significantly correlated between them. This may be an indicator for the linear regression model in a later section.

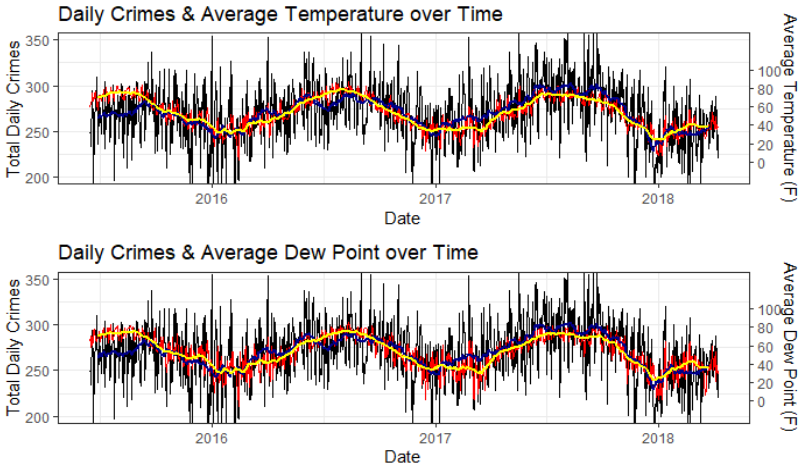


Figure 7. Timeseries of daily crimes over different weather parameters.

The correlation between frequency and average temperature and average dew point can also be observed in the timeseries plots in Figure 7, where the black represents daily crimes, the blue line represents the 30 day moving average of the daily crimes, and the red and yellow lines represent the same characteristics but for the weather parameter respectively.

Modeling

Numerous models and transformations were considered and constructed in addition to the models discussed in this section. However, we concluded that it wasn't appropriate to make any transformations due to the nature and relationships of the datasets being used.

Intuitive Linear Regression Model

A linear regression model was conducted, in which the estimation model can be illustrated as:

Frequency = Intercept + Avg_Temp + DoW + Events

where Avg_Temp is represented with integers, and DoW and Events are binary indicators of each day of the week and the events of precipitation, respectively. Based on our data exploration, we believe that that the best predictors for crime rate are (1) the average temperature, (2) the day of

the week, and (3) the event – Rain, None, Snow, or Both. Average Dew Point was excluded due to the extremely high correlation with average temperature. The model is summarized in the following table:

Variable	Coefficient	P.value	Significance	Variable	Coefficient	P.value	Significance
(Intercept)	187.20	<2e-16	***	**DoWFriday**	47.70	<2e-16	***
Avg_Temp	0.80	<2e-16	***	**DoWSaturday**	27.02	<2e-16	***
DoWMonday	3089	<2e-16	***	**EventNone**	12.57	0.00982	**
DoWTuesday	35.10	<2e-16	***	**EventRain**	8.44	0.09056	
DoWWednesday	35.17	<2e-16	***	**EventSnow**	7.27	0.21120	
DoWThursday	35.31	<2e-16	***				

Figure 8. A linear regression model, with Sunday and 'both events' as the reference groups

The intuitive model of Figure 8 delivers an adjusted R-squared of 0.3573 and AIC of 9682.215. All variables are significant at 99.9% except for EventRain and EventSnow. According to this model, with everything else constant, weekdays tend to have a higher average for the number of crimes per day with Sunday being the day with the least average amount of crimes by 27 crimes. Furthermore, crimes tend to increase by 12 crimes per day if doesn't rain and snow and increase by 0.8 crimes for every increase in the average temperature (F). Since our range for average temperature is 86 Fahrenheit, with a minimum of 2 Fahrenheit, the increase amount of the count of the daily crimes can be up to 69 daily crimes within the set of average temperatures recorded.

Backward Selection Model

A backward selection model was conducted, in which the estimation model can be illustrated as:

$$\text{Frequency} = \text{Avg\_Dew\_Point} + \text{Avg\_Humidity} + \text{Avg\_Visibility} + \text{DoW} \quad ,$$

where Avg_Dew_Point, Avg_Humidity, and Avg_Visibility are represented with integers, and DoW is a binary indicator of each day of the week. A summary of the model is described below:

Variable	Coefficient	P-value	Significance	Variable	Coefficient	P-value	Significance
(Intercept)	222.75586	2e-16	***	**DoWTuesday**	34.66787	2e-16	***
Avg_Dew_Point	0.88625	2e-16	***	**DoWWednesday**	35.18560	2e-16	***
Avg_Humidity	-0.46690	4.16e-08	***	**DoWThursday**	34.95001	2e-16	***
Avg_Visibility	1.35604	0.226	*	**DoWFriday**	47.19213	2e-16	***
DoWMonday	30.67225	2e-16	***	**DoWSaturday**	26.74568	2e-16	***

Figure 9. Backward Selection, from a model that fits "all" to selecting the output shown above

From running a backward selection model on our dataset, we found the weekdays to have a higher average crime rate than the weekends, with Sunday at the lowest and Saturday as the second lowest. We see that the average dew point and humidity are also highly statistically significant in predicting the crime rates in this model. With the daily crime rate as the independent variable, the model gets an adjusted R-squared value of 0.3663 and an AIC of 9666.663. With a resembling interpretation of the coefficients as the Intuitive model, we also have a level-log relationship between the average humidity (represented as a percentage) and the crime rate. This relationship

can be interpreted as, with all other factors unchanged, per every percent increment of the average humidity, the crime rate decreases by roughly half percent.

Forward and Stepwise Selection Model

A forward and stepwise selection model was conducted, in which the estimation model can be illustrated as:

Frequency = Avg_Temp + Avg_Visibility + DoW

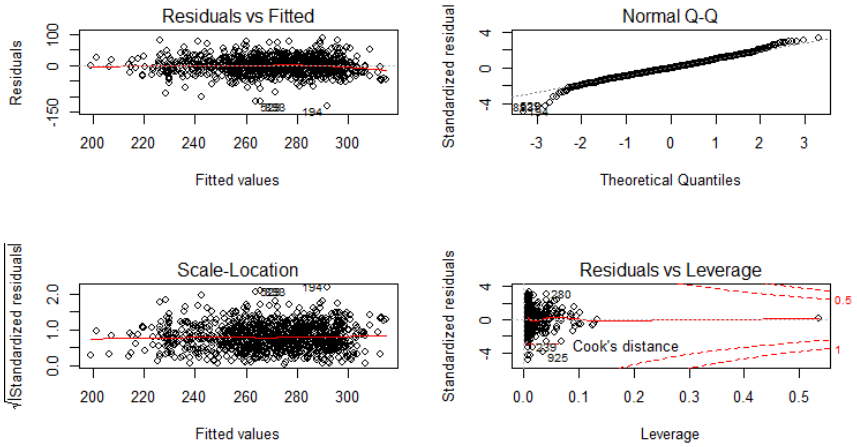
where Avg_Temp and Avg_Visibility are represented with integers, and DoW is a binary indicator of each day of the week. A summary of the model is described below:

Variable	Coefficient	P-value	Significance	Variable	Coefficient	P-value	Significance
(Intercept)	180.80115	<2e-16	***	DoWThursday	35.23966	<2e-16	***
Avg_Temp	0.79836	<2e-16	***	**DoWFriday**	47.51379	<2e-16	***
DoWMonday	31.05023	<2e-16	***	**DoWSaturday**	27.07369	<2e-16	***
DoWTuesday	35.09256	<2e-16	***	**Avg_Visibility**	1.84839	6.21e-05	***
DoWWednesday	35.16306	<2e-16	***				

Figure 10. Forward and Stepwise Selection, from a model that has no x's to selecting the output shown above

The forward selection model and stepwise selection model end up yielding the exact same model. As figure 10 shows, all the variables in the model are highly significant to the model. When looking at "day of the week", the model shows that, compared to Sundays, weekdays have a higher average crime rate holding all other variables constant. With the daily crime rate as the independent variable, we get the average temperature, the average visibility, and the day of the week with an adjusted R-squared value of 0.3616 and an AIC of 9673.246. When comparing the R-squared value and the AIC between the backward selection model and the forward and stepwise selection model, we can see that the values are very similar to each other.

Normality Assumption – Intuitive Model



1. From the residual versus fitted graph, we can observe that the number of data points above and below the red line are roughly the same, demonstrating that the mean of the residuals is 0. Furthermore, we observe a quasi-constant distribution of data, indicating linearity of the model.

Figure 11. Normality Assumption

2. From the normal quantile to quantile plot, we can see that the data passes the "fat pencil test" indicating normality of data except for the lower tail of data.
3. The residual versus leverage plot shows that there is only one high-leverage observation with a cook's distance (distance from centroid) below 0.5, hence the model doesn't seem to be significantly influenced by outliers.

Discussion and Evaluation

A correlation between weather and crime rate in Boston has been observed from the years 2015 to 2018. The frequency of crimes tends to move drastically with Boston's seasonal changes, where there is an overall higher count during the summer following a lower count in the winter. According to the intuitive linear regression model, a unit increase in the temperature (degree) results in an increase in the daily crime rate by a statistically significant value of approximately 187 records. This could possibly be explained by the inclination of people to be more active during warmer, endurable temperatures than during the winter of Boston, which is commonly known to be very cold. Furthermore, it has been observed that the crime rate is significantly higher during the weekdays than during the weekend. As depicted in both the forward and backward selection models, the crime rate recorded between the years 2015 and 2018 gradually increases from Monday to Friday and declines on Saturday and even further on Sunday, all of which have been established as statistically significant results.

Figure	**Model**	**R^2 Value**	**AIC**
8	Intuitive Linear Regression	0.3573	9682.215
9	Backward Selection	0.3663	9666.663
10	Forward, Stepwise Selection	0.3616	9673.246

From figure 12, we can see that the backward selection model has the largest R-square value and the smallest AIC value compared to the other two models. Accordingly, the backward selection model interprets the crime frequency relatively better but is not as intuitive as the intuitive linear regression model. We also notice that all the models provide a similar story with most important variables relating to the average temperature and day of week. And we should admit that the estimations of goodness of fit of the three models are very close. Therefore, combining the results of 3 models can be more effective in analyzing how various factors influence the crime frequency.

It is important to address that the results obtained from this study does not assure complete knowledge and prediction of the crime tendencies within specific locations of Boston during each season. This study only observes the past occurrences of crime in order to help raise awareness of possible criminal activity in certain areas of Boston during a certain period of the year. In addition, it is likely that not all crimes have been reported to the police department for multiple reasons such as the fear to report. Therefore, a dataset covering more years of records on crime and on temperature would increase the accuracy of our results.

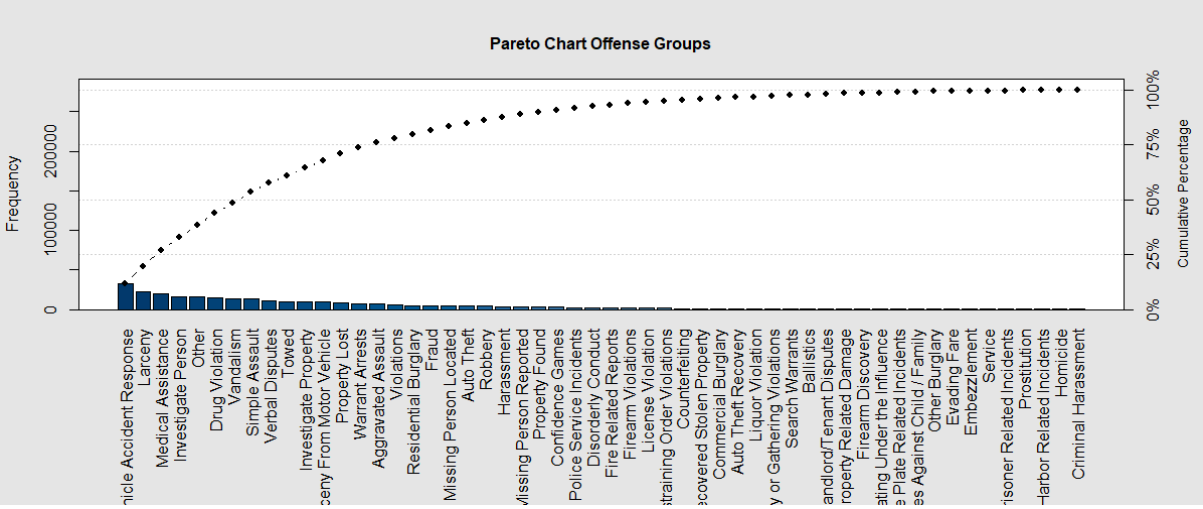
Conclusion

It's recommended to continue the investigation with a larger dataset – if possible – and separating the data into types of crimes (offense groups), since there may be some relationships that are only shown for specific crimes (example: traffic violations during high holidays) and could prove helpful in the pursuit of eradicating crime. A relevant example would be Motor Vehicle Accident Response, the most common type of accident, in which the BDP could use heatmaps, such as the one included in the Appendix, to locate the places with high accident rates and investigate how the city planning and traffic coordination department could alter things to help ease the situation.

The aim for this study is to help the general public as well as the Boston Police Department (BPD) in being prepared for possible criminal situations that could be avoided through effective preparation. The study establishes the likeliness of the daily crime rate to increase between Monday to Friday and to decrease from Saturday to Sunday and establishes the positive relationship between temperature and crime rate. Consequently, the models could help predict future values of crime to help the Boston Police Department allocate the resources.

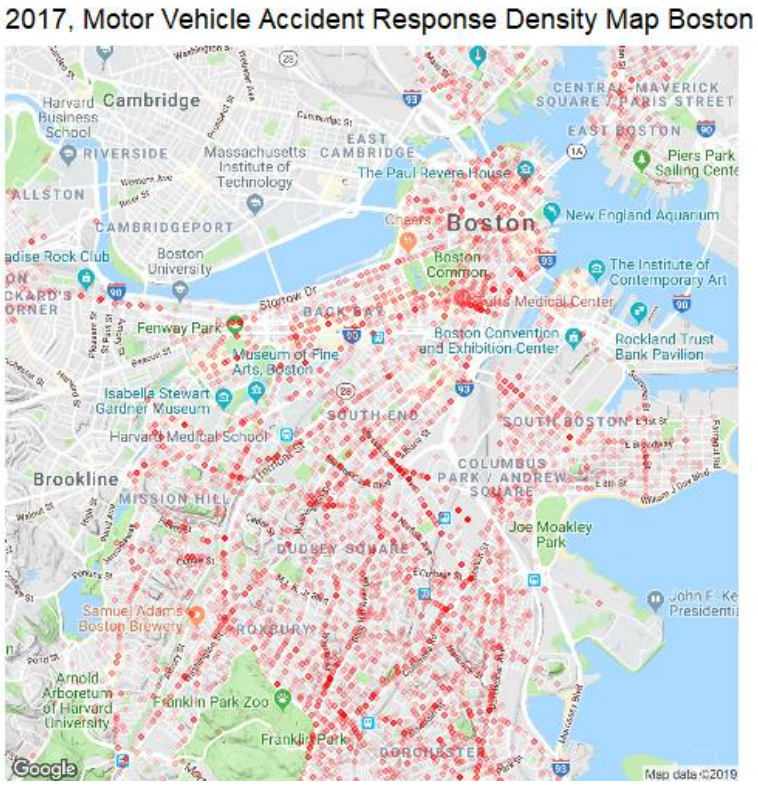
Appendix

Pareto Chart:



Pareto Chart: The Pareto Chart shows that around 75% of crimes are part of the top 15 most frequent offense groups. Note that representation above has been "cut" to hide infinitesimally small frequency offense groups.

Heat Map of 2017 Motor Vehicle Accident Response:



We can observe several locations and crossings that have a higher than average number of accidents. It would be keen to investigate more about the causes of the high accident rates in these specific locations.

Scatter Plot Facetted by Offense Groups:

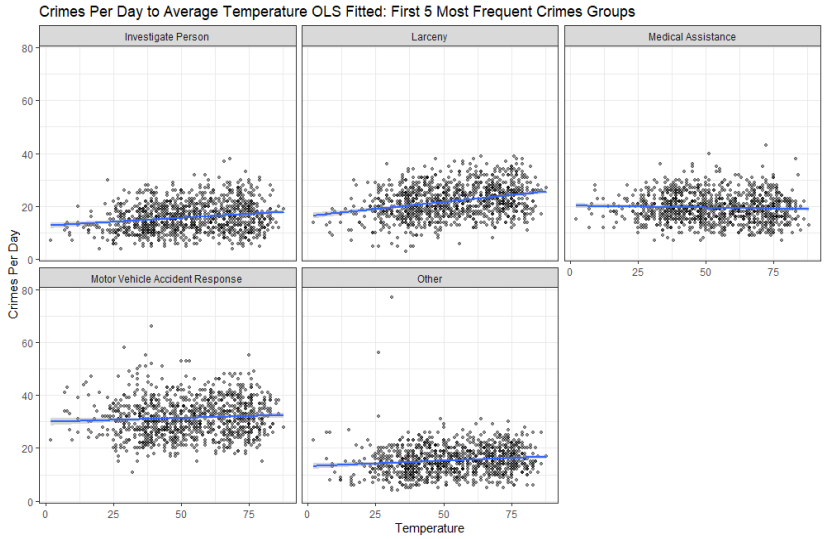
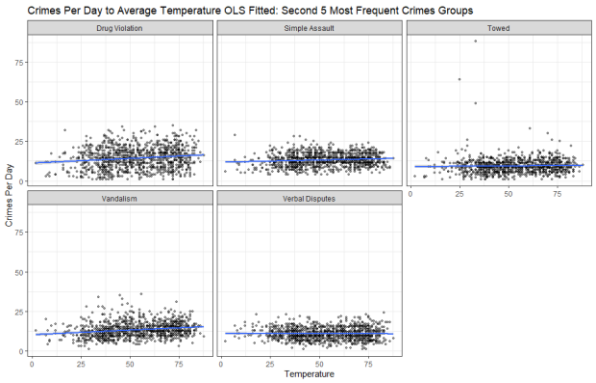
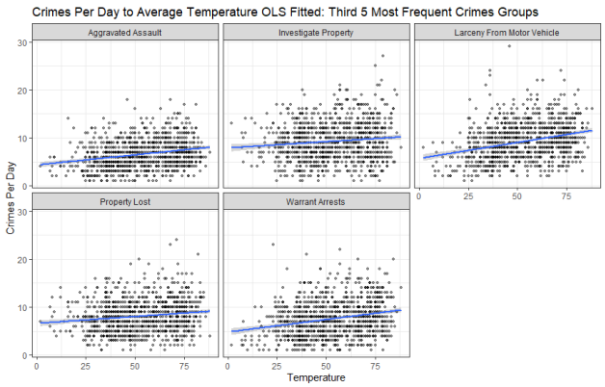


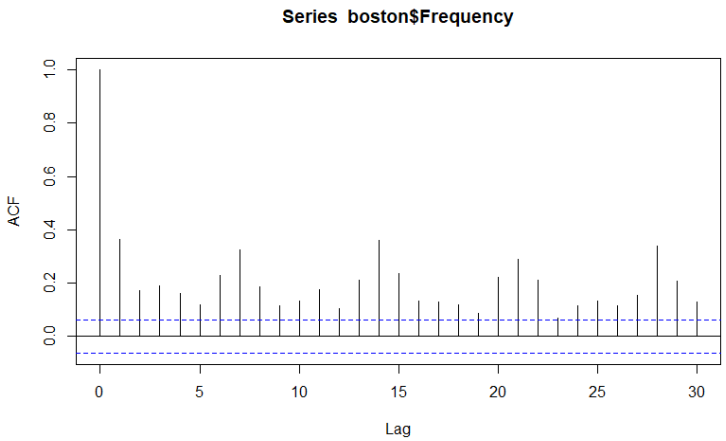
Figure 12. Scatterplots of the Top 5 Most Frequent Offense Groups

Figure 12 shows that given the data we have on Boston crimes and temperature, it is not reasonable to make any statement on the correlation of two variables. The only crime that seems to show a correlation is larceny. This could be because larceny is more common when people have more activities and people are less likely to go outside when the temperature is below 25. However, as showed in the graph above, the distribution of the data is heavily entered around the 50-75 temperature range which might result in bias in the OLS process. Additional scatterplots on the next 10 most frequent offense groups can be seen in the appendix below.



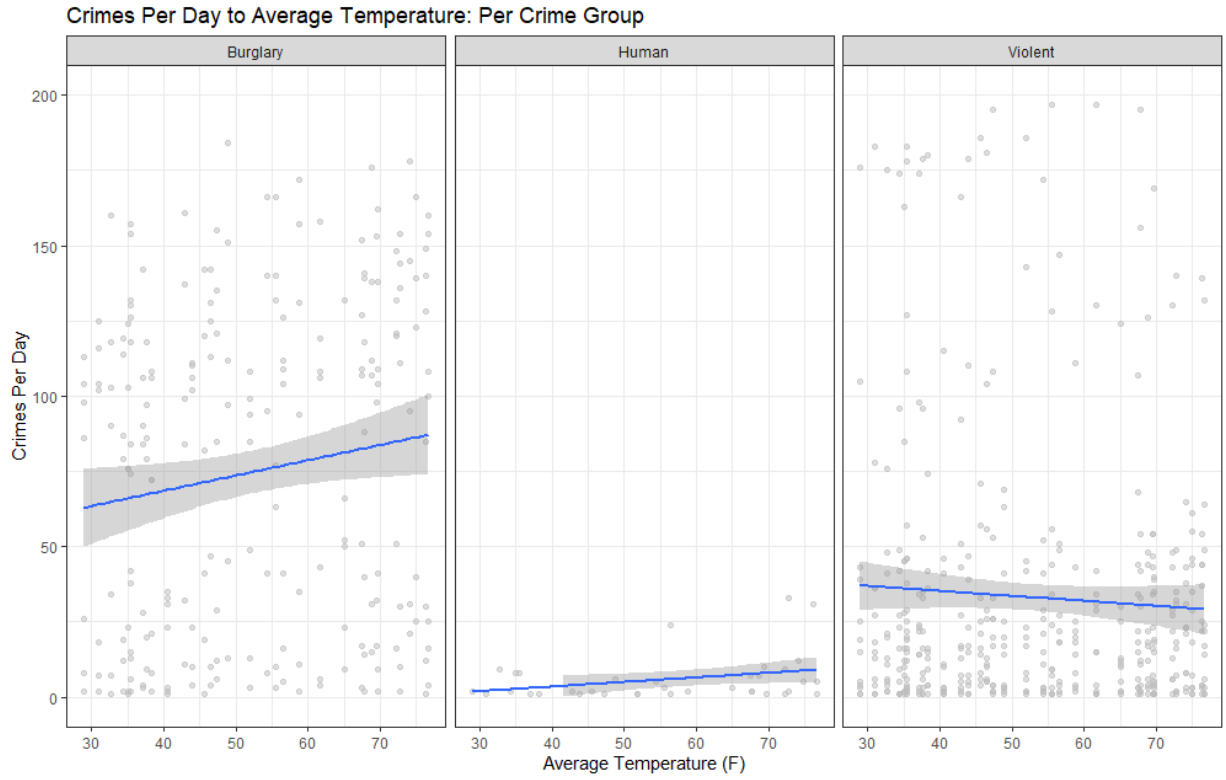


ACF Plot (Autocorrelation Function):



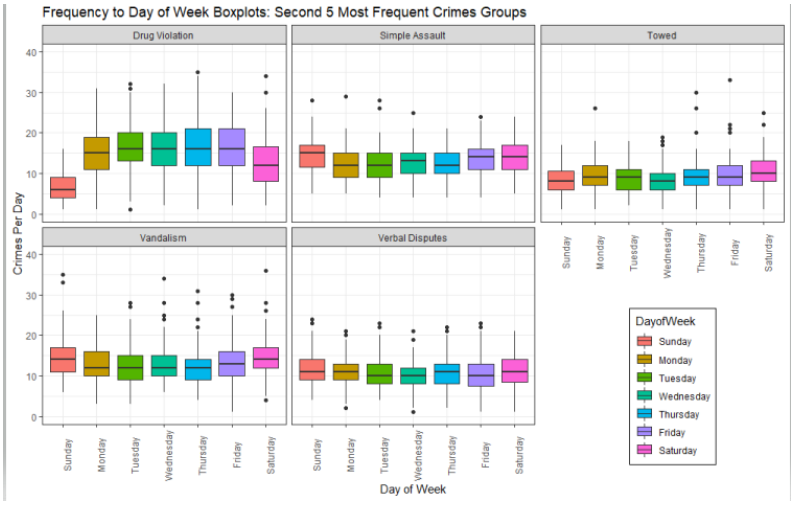
The ACF confirms once more that there is a clear weekly seasonality. It shows there's a correlation between points separated by 7 days (by 7 time-lags).

Relationship between the Frequency of Crime and Average Temperature for Each Crime Group:



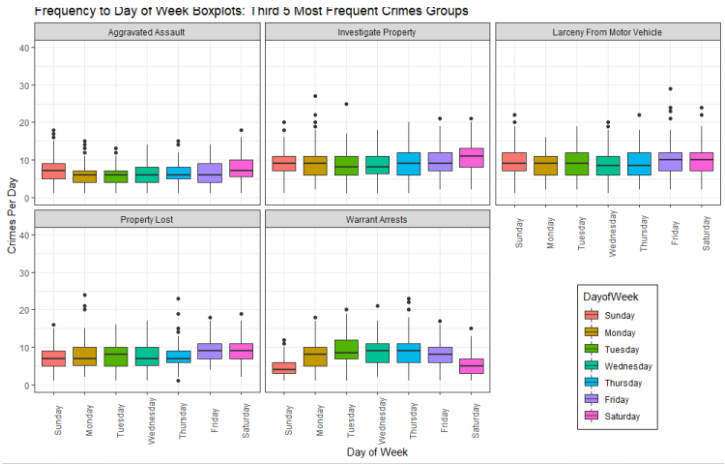
The figure above illustrates the relationship between the crimes per day with the average temperature in three different classifications of crime. Offense groups were further categories into three sections for this study, which are divided as: burlary, human, and violent. As seen on Figure 13, crimes in the human category were far less frequent than the other two groups. However, while, both the burglary and human scatterplots show that there is a positive linear relationship between the frequency of crimes and the average temperature, a negative relationship can be observed for that of the violent category.

Crimes Per Day versus Day of the Week Facetted by Offense Groups:



The trend of second 5 most frequent crimes from Sunday to Saturday in a week do not follow that of overall crimes. Drug violation and towed have less crimes on weekends and more on weekdays, which is similar to overall situation. Simple assault, vandalism and verbal disputes have less in weekdays and more on weekends. The range of crimes per day of second 5 most frequent crimes follow no clear rules

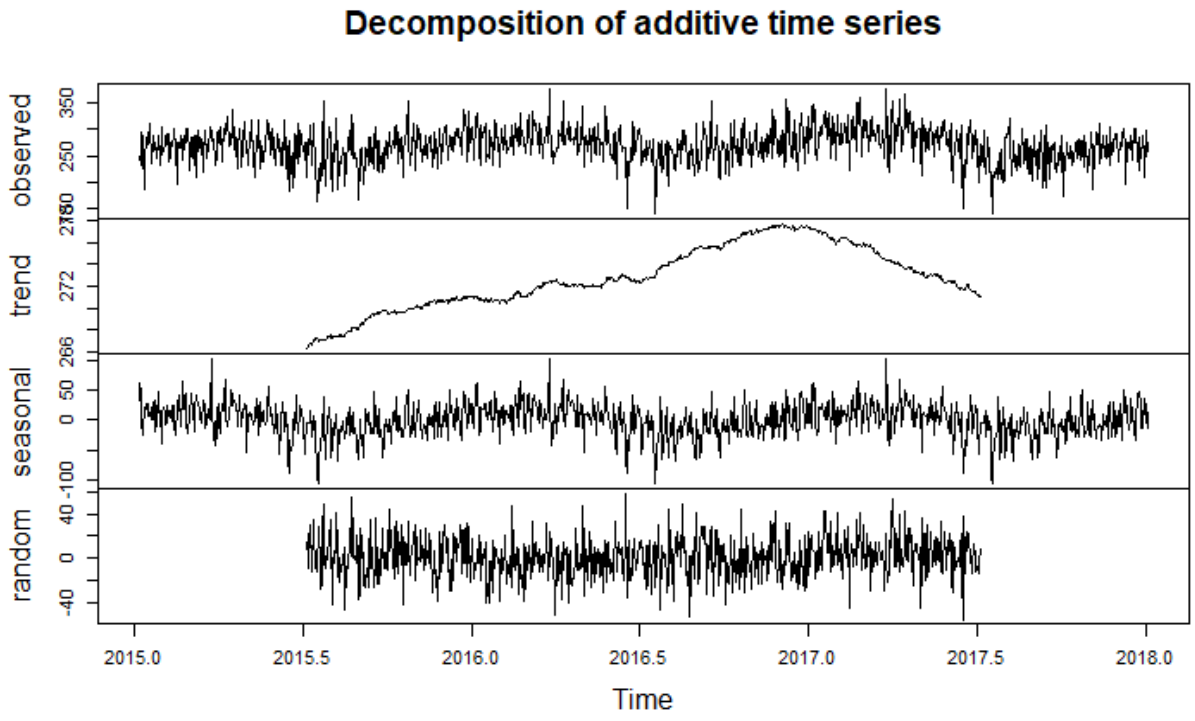
between weekends and weekdays. The range of drug violation is larger than other crimes. The second 5 most frequent crimes have more outliers than first 5 most frequent crimes.



The trend of third 5 most frequent crimes from Sunday to Saturday in a week do not follow that of overall crimes. Only warrant arrests have less crimes on weekends and more on weekdays, which is similar to overall situation. Aggravated assault, investigate property, larceny from motor vehicle and property lost have less in weekdays and more on weekends.

The range of crimes per day of third 5 most frequent crimes follow no clear rules between weekends and weekdays. In detail, larceny from motor vehicle has the largest average range in a week. The third 5 most frequent crimes have less outliers than second 5 most frequent crimes and approximately the same as first 5 crimes and the overall one.

Decomposition Plots:



Time series decomposition showing the trend, seasonality and randomness of our timeseries.