DSC423 – Data Analysis and Regression

Milestone 04 – Data Analysis

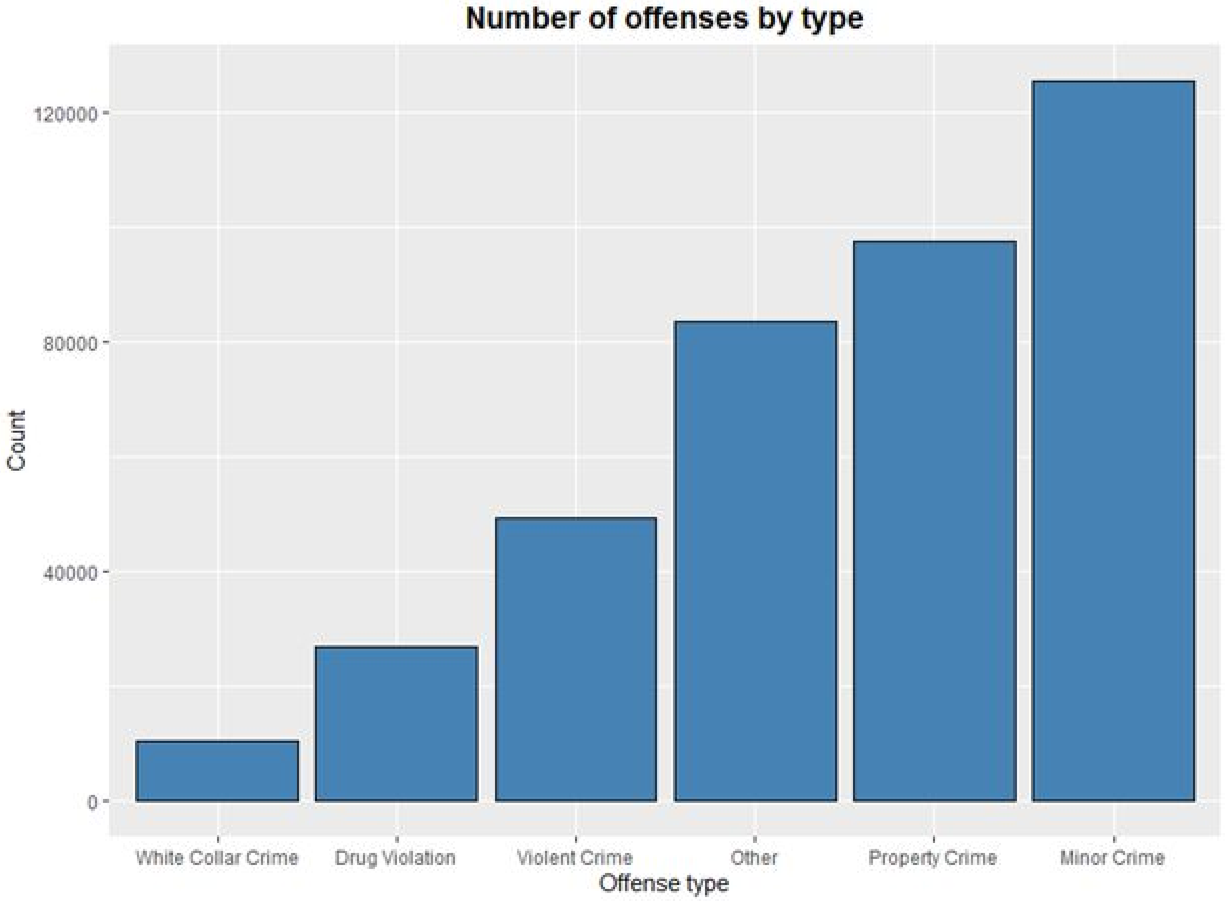
By:

Group Dogwood Rose Doppelgangers

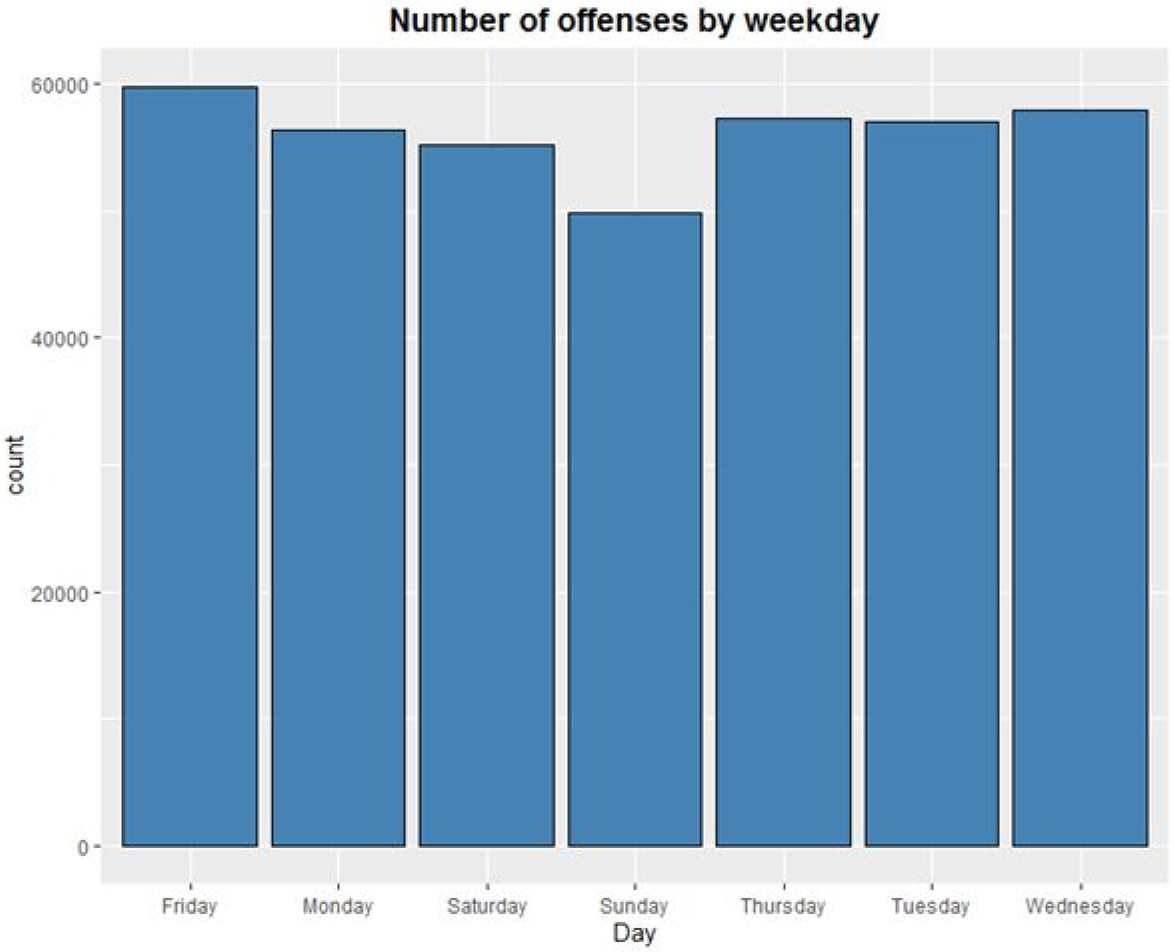
Jesse Franks, Pete Gillespie, Reid Case, Robert Cordrey, Salik Hussaini, M. Omar Khan

For the purposes of our analysis, we elected to combine several different techniques to better understand and visualize the data we have collected thus far. Our primary objective for this stage was to provide some context to the data collected and theorize actionable experiments based upon these insights. To better understand what possible explanations exists for the crimes being committed, it was our goal to ensure we considered several possible variables and their relationship to the frequency and nature of criminal behavior. While crime is unpredictable by its very nature and is subject to the whims and misanthropic minds of the masses, we endeavor through the initial analysis to help us paint in broad brush strokes and develop something of a picture surrounding criminal activity.

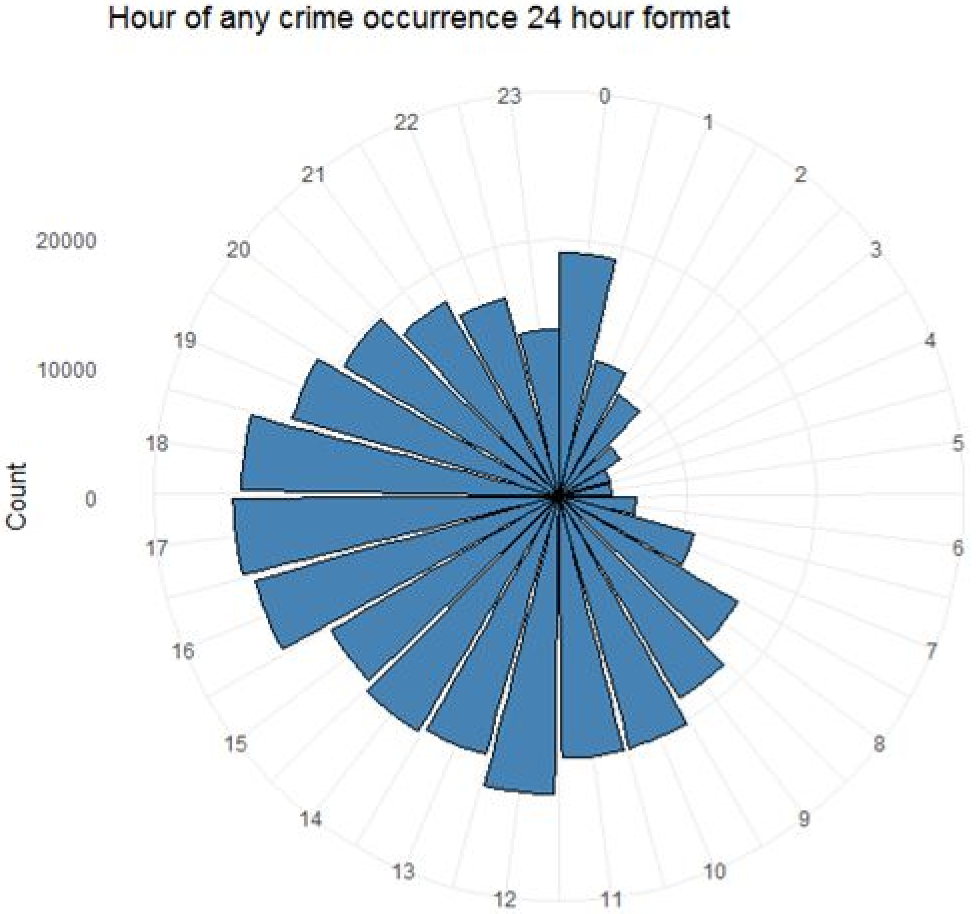
BAR CHARTS:



We began our observations by dividing our dataset to show the distribution of the categorical variable as represented by the “OFFENSE_CODE_GROUP” field into a bar chart. By categorizing as such, we were able to determine an approximate range for frequency of crimes depending on their nature. As indicated in the bar graph it is easily determined that “minor crimes” tended to be the most prolific of crimes reported, representing 31.9% of the crime reported for the period under review. This result may not be all that surprising as it represents an amalgamation of petty crimes, those with the least amount of effort to commit, and perhaps are the simplest to report. The chart goes on to show some interesting pieces of information, namely that property crime is the second largest group of crime committed at 24.8% and finally that white collar crime attributes to only 2.6% of criminal activity. Based on the chart above we believe that the majority of datapoints in our analysis will be constituted of minor crimes. One actionable experiment would be to compare time of day as an explanatory variable using linear regression on each category of criminal behavior to see if there exists a relationship. We expect that crimes committed later in the day (after dark) would mainly be constituted of violent crimes, or drug violations, while white collar crime would almost exclusively occur during business hours (9:00 AM to 5:00 PM).



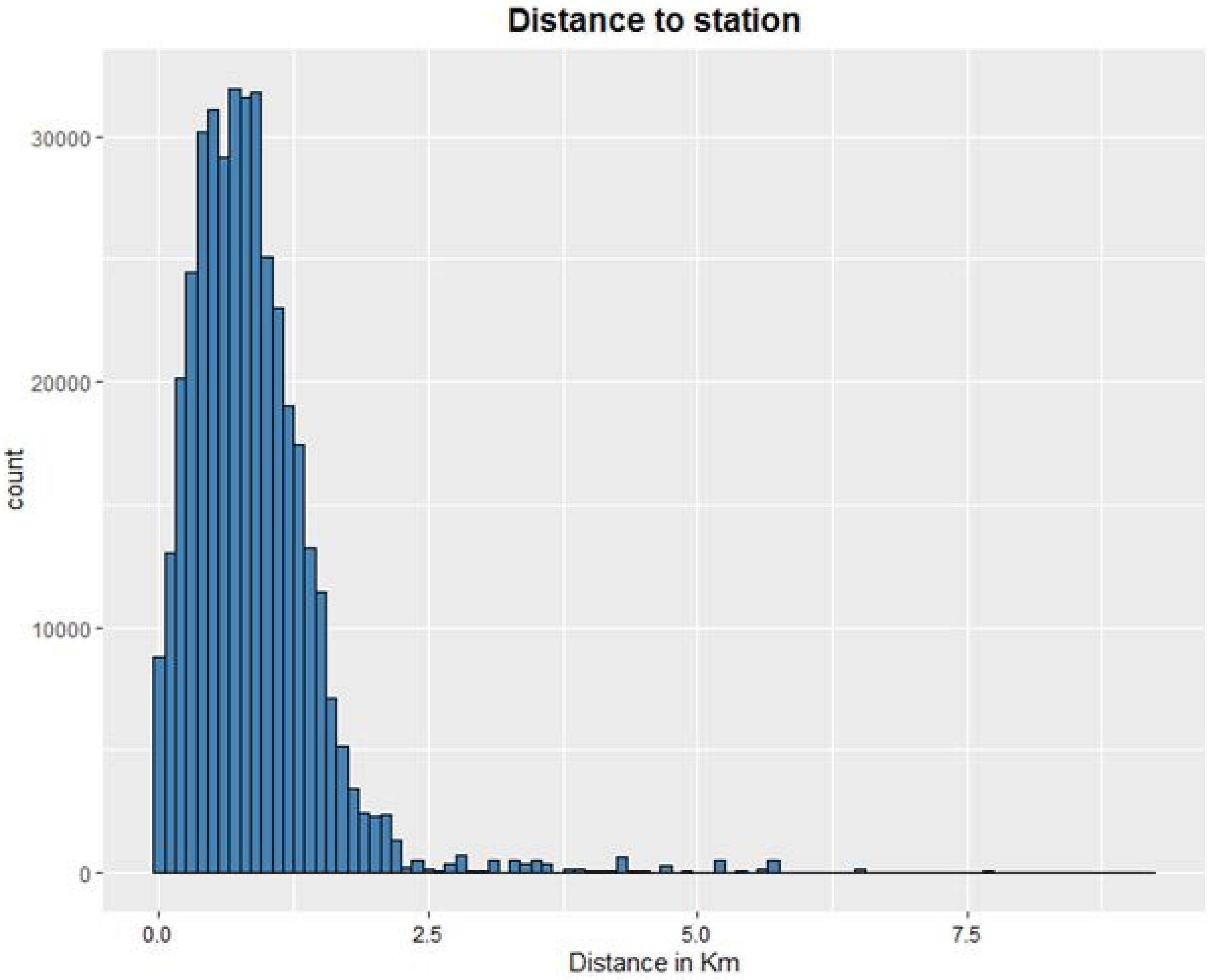
With this chart we elected to focus on the frequency of criminal activity based on the day of the week. By doing so we can determine whether a possible relationship exists between the day of the week and the frequency of crimes. We can clearly see that Fridays represent the most prolific day of the week for a crime to occur. Thereafter, all the other days of the week appear to be generally in line with each other, except for Sunday. Based on this data we believe that crimes in the Boston Metropolitan Area are statistically less likely to occur on a Sunday. Additionally, this bar chart suggests several possible insights; for instance, criminal activity may peak on Fridays as it’s the last day in the work week. We believe that criminal activity committed by those individuals that are otherwise preoccupied during the week becomes a viable opportunity for those who do not have to report in for work on Saturday. Sunday may be a “day off” for criminal behavior. Again, one possible insight is because of the requirement to go into work on Monday. Another possible insight is that since fewer people are out and about on a Sunday, fewer opportunities for crime exist. Based on these insights, we will use linear regression we will see if there is a clear relationship between the day of the week and the likelihood of a crime being committed, and we expect there to be a strong positive correlation based on the data from the bar chart above.



Although the presentation may be unorthodox, this bar chart indicates the time of day (in the 24 hour format) and the number of crimes. By treating the 24 hour day as a categorical variable we are able to see the frequency of crimes based on the daily cycle. Based on the chart above it appears that the “witching hours” of 15:30 – 18:30 represent the majority of criminal behavior; however there are a number of curious cases, such as midnight (0:00) and noon (12:00). Also based on this chart, it is clear that the early morning hours are far less likely to have a crime committed during that time. Specifically, it is clear that from 1:00 – 7:00 there is a precipitous drop in criminal activity. One insight as to the driving factor behind this is because no one is awake at those hours and therefore the opportunity to commit crimes is less prevalent.

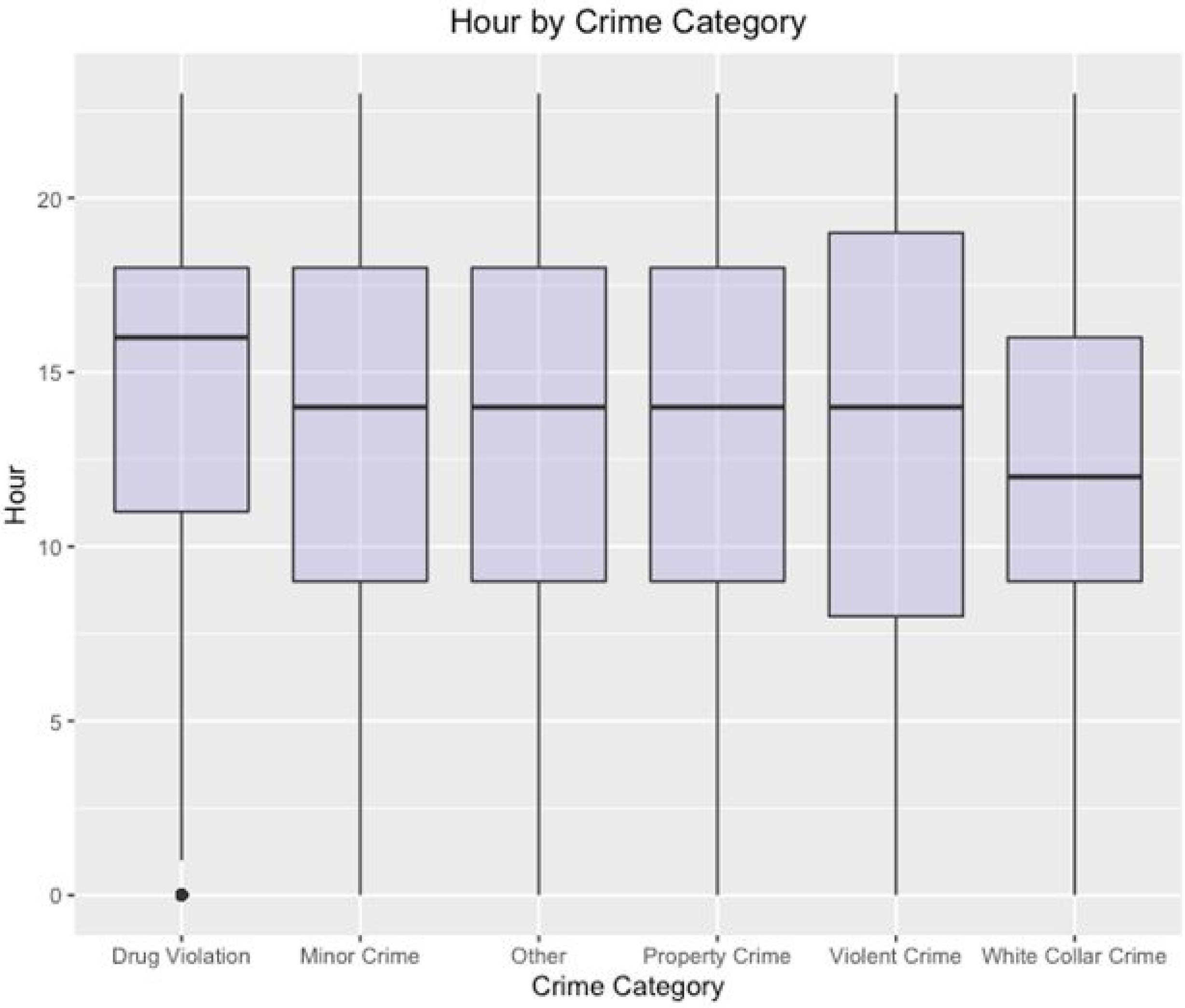
On the other end of the spectrum the 15:30 – 18:30 virtually lines up with post work “rush hours” for commuters. As such, the opportunity is rife for criminal behavior. As to noon and midnight, these outliers may be harder to explain, however one possibility is that data entry of crimes reported on certain days may have followed on standard formatting if the exact time a crime took place is not known. For example, a clerk who is filing the report may not know the exact hour a crime took place and as such reported it as being 0:00 or 12:00 out of convenience. Based on these insights we again plan to use linear regression to see if there is a direct relationship between time of day and criminal activity. We anticipate that the two are strongly correlated based on the above.

HISTOGRAMS:

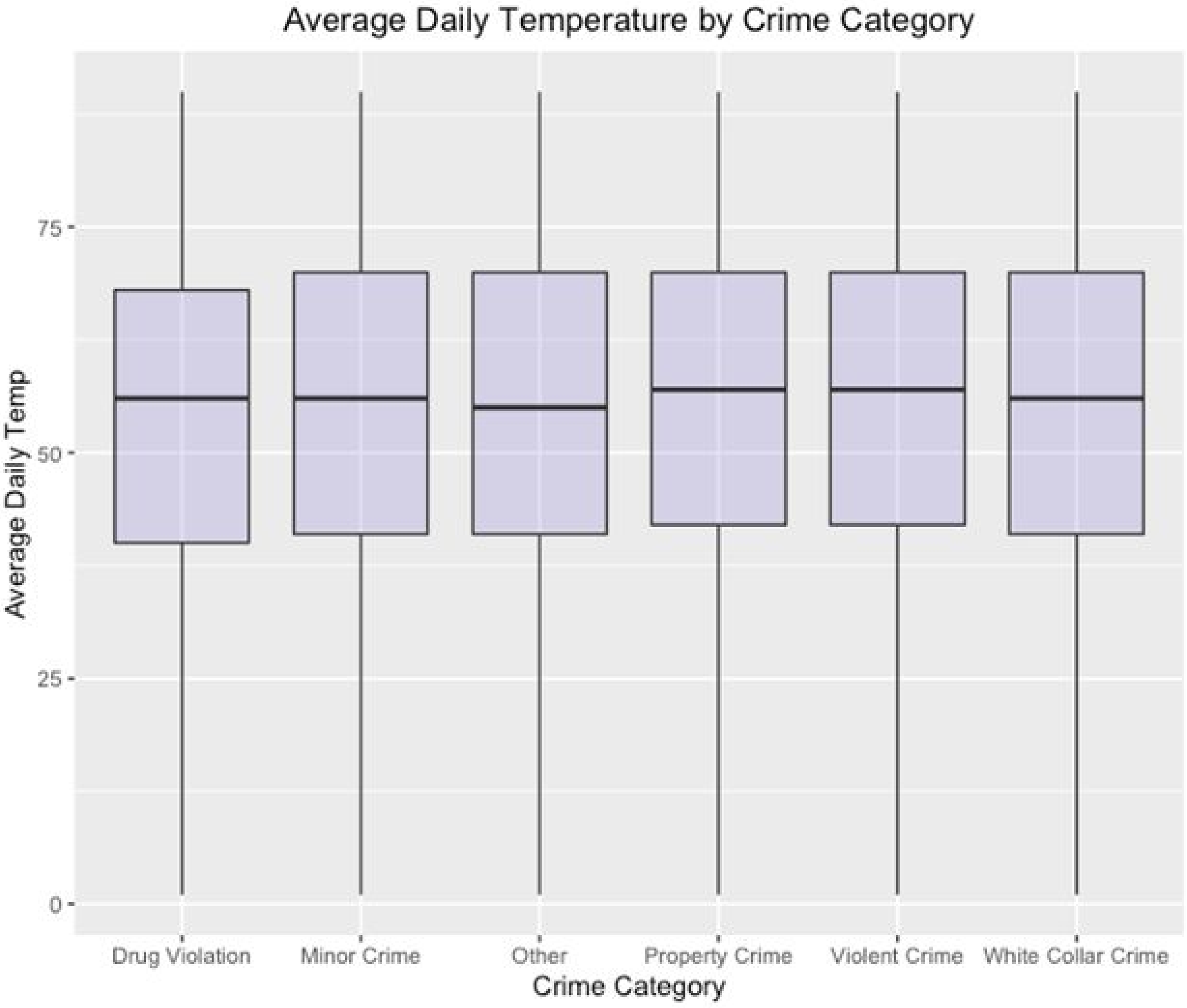


In the above histogram we wanted to determine what the typical distance to the nearest police station was in comparison to how frequent crime was being committed. We used the latitude and longitude data to help us in determining what the distance “as the crow flies” from the spot of the incident to the nearest police station. As we can see the data above is right skewed, with most crimes being reported within a 2.5-kilometer radius to the station. Insights from this histogram were varied; however, we believe that the proximity of police stations has a direct impact on the ability to report crimes and thus gain representation within our dataset. As such, criminal activity occurring a great distance away from a police station is far less likely to be reported, although such activity could be occurring. One perhaps superfluous insight could be that the creation of police stations somehow increases the crime rate within its immediate vicinity. However, a more logical explanation would be that police stations simply have an effective radius of about 2.5 kilometers. Based on this insight we intend to use linear regression to determine what the correlation is between the distance to a station and a crime being reported. We anticipate that our data will indicate that there is a strong correlation between the distance and the likelihood of a crime being reported. This may lead us to the conclusion that the Boston Police Department would need to place their stations no more than 5 kilometers apart to ensure adequate coverage of the metropolitan area.

BOX PLOTS:

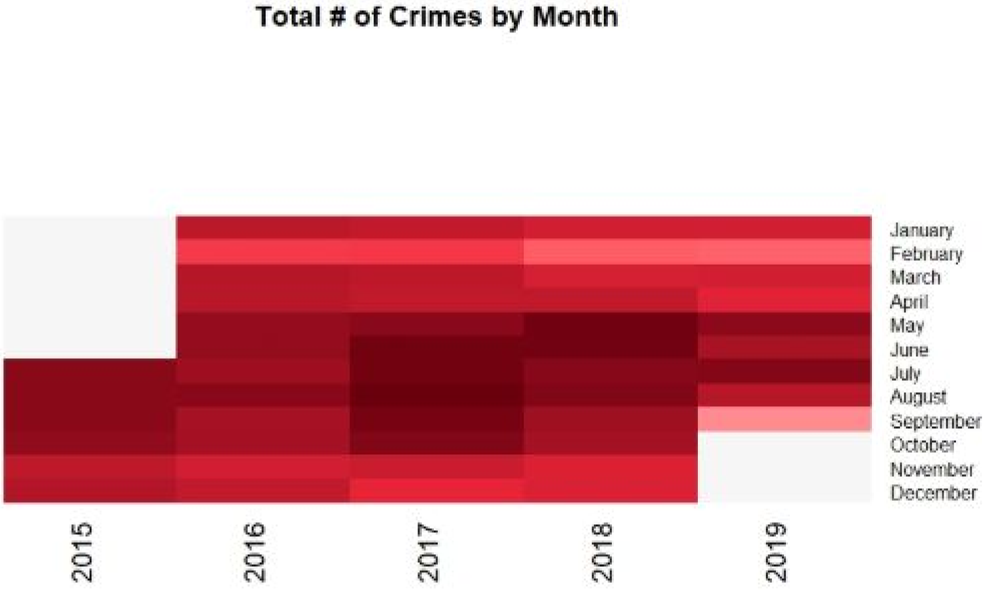


Here we elected to examine a box plot for each category of crime as noted earlier. This depicts similar data to the bar chart we saw regarding the time of day and count of criminal activity; however in this box plot we have included specific information on the nature of the criminal category. This is useful to us because we can quickly identify possible correlations between the type of criminal activity and the time of day in which such activity is said to occur. One interesting insight is that generally speaking criminal activity occurs within the same band of time, which confirms our insight from the prior bar chart; one unique case however is the white collar crime category. We can see this is limited to almost exclusively the traditional business hours of 9:00 – 17:00 (9:00 AM – 5:00 PM). This makes complete sense as when else could a white collar crime be committed. Another interesting insight is violent crime has the largest box covering virtually the entire waking day. Based on these insights we believe that conducting a linear regression for each category of criminal behavior against the time of day will help us to determine if a correlation exists. That is to see if time of day is an explanatory variable for the type of criminal behavior.



Similar to the prior box plot, we discussed the time of day as an explanation for the category of crime committed, we examined the average daily temperature to see if there was an explanation between temperature and the category of crime committed. Based on this box plot, we are able to see that generally speaking, criminal activity in the Boston Metropolitan Area is most likely committed above 30 degrees and below 70 degrees Fahrenheit. This is a generally broad range however it is interesting to note that virtually all categories of crime seem to occur generally within the same band. Based on this analysis, we believe that during the colder winter months, crime is significantly less likely to occur. As such, based on these insights we believe that a linear regression comparing temperature and crime for each criminal category will show a strong positive correlation. We anticipate that the R-squared between each of the categories will be relatively similar as based on this box plot there is no major variance.

HEAT MAP:



This heat map describes a relationship between the time of year and the occurrence of criminal activity. It is easy to note that generally the darkest portions of this graph shows up during the warmer months for Boston. This can be further validated based on our prior analysis on the average daily temperature and the occurrence of a crime. Based on this we believe that the time of year greatly influences the likelihood of criminal activity. Generally, December through February show the lowest likelihood of criminal activity, while May through August have the highest likelihood. It is likewise fascinating to note that we inadvertently were able to draw comparisons between the years for criminal activity. 2016 appeared to have fewer total crimes compared to 2017 and 2018. While the data is still out on 2019, it appears that the winter months behaved as expected, with May and July showing peak criminal behavior. Based on these insights we believe that a linear regression model between the time of year and criminal activity will show us a distinct relationship. Time of year as an explanatory variable will determine the rate of criminal activity.

In conclusion, it is clear that there are any number of ways to compare the data on criminal behavior. We have chosen to examine a few of the many variables in order to gain better insights into what can explain criminal behavior. With these insights in hand, we anticipate that we will be able to formulate multiple lines of inquiry into this indelibly complex problem. It is worth mentioning again that while we endeavor to predict something as irascible as the criminal workings of human nature, armed with these insights we are prepared for the challenge.