Exploring Data Science: Understanding, Predicting, & Visualizing Crime in Syracuse

A Capstone Project Submitted in Partial Fulfillment of the
Requirements of the Renée Crown University Honors Program at
Syracuse University

Ryan H. French

Candidate for Bachelor of Science in Information Management & Technology
and Renée Crown University Honors
Spring 2019

Honors Capstone Project in Your Major

Capstone Project Advisor: Deborah L Nosky, Assistant Professor of Practice

Capstone Project Reader: Avinash Kadaji, Adjunct

Honors Director: Dr. Danielle Taana Smith

Abstract

With the advent of the open data portal for the city of Syracuse came an opportunity previously impossible; anyone could download, mine, and visualize information about Syracuse direct from the source. Over the course of this project, I will be performing these processes on a selection of crime data from 2017 in order to better understand the patterns of crime in Syracuse, where they occur, and if it can be predicted whether or not a crime will lead to an arrest.

This project will begin with an overview of the data, how it was obtained, and the meanings of the variables within. Next, I will begin the process of cleaning and feature engineering so as to optimize the data set for analysis. Following that, I will visualize the different aspects of the data and point out interesting insights I have drawn as a result. Finally, I will attempt to predict whether or not an arrest will occur based off of other variables present within the data.

Executive Summary

Over the course of this project, I endeavored to analyze and visualize occurrences of crime in Syracuse throughout 2017 so as to aid in the understanding of its citizens and, potentially, learn how to prevent more crimes from occurring in the future. In order to do so, I utilized a wide variety of Data Science techniques and processes on the data available to me in order to generate insights.

I began with Data Cleaning, the process of removing imperfections from the data. This involved fixing a number of empty values in various rows as well as reformatting some data into more readily accessible formats.

From there, I turned to Feature Engineering; re-tooling the data so as to be easier to analyze. This involved creating separate columns for date and month, grouping the time crimes were reported into hourly bins, and removing unnecessary columns.

Next, I began the process of data mining, visualization, and predictive modeling, leading to some interesting insights and results. I have included a few of these below:

Large retail locations correlate with an increase in crime; Destiny USA exhibits the most crime reports in Syracuse by far at 579 in 2017 and, of the three areas with the next most crime reports, 2 of the 3 have large Tops stores. Increasing police and/or security presence around Destiny USA and these other locations may have an impact on the amount of larcenies occurring in the Syracuse area.

Arrests occur for only about 1/4 of the crimes reported; larceny is the offense which most frequently leads to an arrest.

The vast majority of crimes reported involve robbery or larceny codes (76.4%); it may be worth investigating how many households have home security systems and

what measures they are taking to prevent such crimes as well as providing additional information sessions and/or material on preventative measures for citizens to employ.

Random Forest insight analysis indicates that LarcenyCode and CODE DEFINED are the most significant variables in determining whether or not an arrest will be made.

Machine Learning algorithms appear to be able to predict arrests with a respectable degree of accuracy (greater than 70%). Running Support Vector Machines on a dataset containing 50% arrests and 50% non-arrests predict arrests with an accuracy of 74.0%. Running Random Forest on a dataset containing 50% arrests and 50% non-arrests predicts arrests with an accuracy of 71.3%

It is my hope that, after reading this report, one will have a better understanding of the most common crimes in Syracuse and where they occur as well as some insights as to how they can be potentially reduced.

Project Background

Initial Research

As an undergraduate student in the School of Information Studies accepted to begin my studies as a Master's student in Applied Data Science next Fall, I was interested in pursuing a project that would allow me to manipulate data, better explore and refine my Data Science skills, and produce some interesting results that might be useful to the Syracuse community.

At the time, I had recently listened to an episode of the popular internet podcast Reply All which featured an extensive story on Jack Maple, the man credited with first deploying predictive analytics in the realm of crime and the creator of the New York Police Department's predictive crime system, COMPSTAT. The first of its kind, COMPSTAT demonstrated immense success and went on to be widely adopted in other metropolitan cities in one form or another, beginning the modern age of crime analytics (Vogt and Goldman).

My interest in crime analytics piqued, I began researching what kind of crime data was available on the Syracuse area and stumbled across the Syracuse Open Data repository. Open data repositories are essentially collections of data made available to the public, in this case by a government entity (the city of Syracuse). In this way, the data can serve a variety of purposes. The average citizen is provided an opportunity to become more involved in (and aware of) the domain of the data and its relevancy to their local community. Data driven individuals are provided information to conduct experiments upon and potentially generate valuable insights as a result. Finally, governments are seen as more transparent, community minded, and can potentially

benefit from the insights generated by independent researchers without needing to utilize their own valuable resources.

Some examples of well-developed open data projects are those present in Chicago, Illinois and Los Angeles, California, vast repositories containing a wide variety of easily navigable data available to the public. Topics available online range from trash container locations, to mapped landmarks, to report cards for public schools, and quite a few in-between.

However, simply because the data has been made available by the relevant organizations from these locations does not guarantee its integrity. There is plenty of room for error between a crime being observed, an officer entering it into the database on their end, and the database consequently being uploaded to the internet for public use. Additionally, as discussed in the podcast previously mentioned, frequently the implementation of crime analysis and prediction can lead to vicious feedback loops, creating unattainable goals for reducing crime (Vogt and Goldman).

For instance, if a police force is doing an effective job of catching offenders and reducing crime, it makes logical sense that the amount of gross crime occurring would decrease. However, due to the goals set by management algorithms, police are expected to be catching increasingly more criminals even as crime in general decreases, a difficult objective.

Another problem that frequently arises from such situations is the under-reporting of major crime in an attempt to demonstrate its decline. For instance, crimes involving acts of prostitution may instead be reported as lesser infractions such as loitering in order to prevent more prostitution being reported in a given area, potentially implying that law

enforcement is not doing an adequate job. In this way, the manipulation of crime reporting presents an additional barrier to authenticity in the data.

With this in mind, I will be treating the data utilized in this project as if I can guarantee its integrity, however I also invite readers to keep a healthy degree of skepticism in mind as they read. In this case, the data utilized consists of crime reports from Syracuse city police officers which had been reported over the course of 2017. In order to protect the identity of those involved, the addresses for each report were generalized to the block at which the crime occurred. The data was initially made available in April of 2018 by Chief Data Officer of Syracuse, Sam Edelstein and consists of 5598 rows, one for each crime reported. Having found a data set, I set out to understand it, visualize it, and see what interesting insights might reside within.

Data Utilized

Initially, I had intended to utilize the most current data from 2018 as well as that from 2017 in order to provide the most contemporary reporting and to be able to train my predictive models on the most data possible. However, currently only the first 6 months of data are available for 2018 (January to June) which raised concerns for me as to skewing my results by including this data. As such, I have chosen to simply utilize the data from 2017.

The 2017 crime data implemented in this analysis can be downloaded from:

http://data.syrgov.net/datasets/0583c4cbea2d4edf9f13e8dcbe21eefa_0

In order to effectively manipulate, analyze, and visualize this information, I pulled down the available .csv file from the web address above.

Contained in the data are a variety of attributes associated with what are called “Part I Crimes”, the 7 serious and commonly occurring crimes which the FBI utilizes to track crime around the United States. These attributes include information about the crime such as the address it occurred at, the type of crime, and the time it was reported. For the 5629 rows in the data set, this information is represented for each instance. The time frame for this data is the year of 2017 with a few outliers (points that fall outside of this date range) from years previous.

This data was posted by the Chief Data Officer of Syracuse, Sam Edelstein and contains information related to violent crimes committed in the city. The definitions for the crimes included can be found at: https://github.com/CityofSyracuse/OpenDataDictionaries/blob/master/PartICrimeSelected.pdf

Explanation of Attributes

Each entry in the crime data set consists of 10 attributes which I will define below:

Attribute	**Definition**
ADDRESS	The address at which the crime occurred, scaled to the block level for anonymization purposes.
Arrest	Whether or not an arrest occurred.
Attempt	Whether a crime was completed or merely attempted.
CODE DEFINED	The code associated with the type of crime (LARCENY, ROBBERY, AGRIVATED ASSAULT, etc).
DATE	The date on which the crime was reported.

DRNUMB	The unique ID for each instance of crime reported.
FID	The unique ID for each row in the data from the repository.
LarcenyCode	The location at which the Larceny took place (From Mailbox, From Building, From Motor Vehicle).
TIMEEND	The time at which responding to the crime ended.
TIMESTART	The time at which the crime was first reported.

Data Challenges

In my effort to understand the data, a number of challenges arose which I will now enumerate. For instance, simply understanding the data available to me initially provided some problems as not all of the fields present were readily intelligible. While a file with definitions for each attribute in the data was provided in the notes for the data, some were incomplete and required additional information which was provided by Sam Edelstein.

Although there is little way to know for sure, the integrity of the data is also questionable when pulled from an open source such as this. As there is little way to conference with those who initially entered these reports (or even figure out who they were due to the anonymization of the data) or to confirm the validity of the data (aside from checking for missing values and assuredly mis entered information), I operated under the assumption that all of the information provided was correct.

Finally, the inability to easily access those involved with the data creation presented additional boundaries in regard to domain knowledge and checking assumptions or

expectations. That being said, being able to reach out to Sam Edelstein with questions and to obtain his feedback on this overall endeavor was invaluable.

Data Exploration

The first step of my analysis was to explore the data that I had available to me to better understand the information present, the way it was formatted, and the kind of questions (and answers) that I might be able to synthesize from it. In order to do so, I utilized R and R Studio.

R Studio is a workspace for the open source programming language known as R which is commonly used for statistical and data modeling practices. Using an R script, I imported the data into R Studio to allow me to better view it, utilizing the View() function to view the information in spreadsheet format.

Taking note of the of the information present, I began to do some initial visualizations using the histogram() and barplot() functions built into R on the various columns of the dataset to better view the distributions within each. I also inspected the structure of the data itself using the str() function which provides insight into the layout of the information and the types of variables used (characters, integers, factors, etc).

Once I felt that I had adequately explored the data enough to get an idea of its shape and what interesting insights I might be able to pull from within it, I moved onto the Data Cleaning phase.

Data Cleaning

Raw data is often unclean; it contains outliers, mis-entered information, and missing values. The data cleaning process primarily involves reformatting data to be easier to work with, removing outliers and errors, and performing discretization.

I began by removing time outliers in the data; 30 rows that were from years previous to 2017 (one being from 1994!). Next, I converted the Arrests and Attempt attributes to binary; they had previously only had flags where said incidents occurred. In this way, I removed the large number of empty entries present in the data that could affect my subsequent graphing. In similar fashion, for crimes without larceny codes, I added the value of ‘Not Larceny’ in their ‘LarcenyCode’ column so as not to have empty values.

From there, I discretized the times in both TIMESTART and TIMEEND into hourly bins for a total of 24 separate bins. Essentially, the discretization process involves grouping attributes with a wide variety of unique values into overarching bins (in this case, by hour) in order to better perform analysis upon them. For example, times such as ‘12:05’ and ‘12:45’ would both become grouped under the bin of ‘12’. In this way, it becomes easier to visualize trends in occurrences of crime over the course of a day.

Similarly, I discretized the initial TIME column (which contained values such as ‘2017-10-07T04:00:00.000Z’) into separate, more legible columns for both month and day by slicing the strings present.

Finally, I added my new columns (‘weekDay’, ‘DAY’ and ‘MONTH’) to my data set while dropping columns that I did not want or need (‘TIME’, ‘DRNUMB’, and ‘FID’). Having finished cleaning the data, I was now prepared to begin my analysis and visualization.

Analysis & Visualization

My analysis began with referring back to my previously created rough barplots and histograms of the relevant variables. From there, I began the process of improving the visualization from both a legibility and aesthetic perspective using the package ggplot2 for R. ggplot2 is a powerful plotting and visualization package commonly used to augment R's default graphics by adding new elements to visualizations or allowing existing ones to be better customized. Each of these visualizations is included below along with my analysis for your viewing convenience.

Mapping

To begin my analysis, I was interested in visualizing the distribution of crime in Syracuse. However, this initially proved a bit more complicated a task than I had envisioned; while many crime databases consist of street address, city, and zip code (and sometimes latitude and longitude), the Syracuse crime data contained only the street address. Some examples include '400 WHITTIER AVE' and '200 ELSNER ST'. However, in order to plot these addresses on a map, I needed to reverse geocode them which involves passing the address to an API which will then return the latitude and longitude for each.

Due to the limited information available for each address, my first step was to append the string ', Syracuse, NY' onto each. This way, the API would be more likely to correctly determine the location. Next, I passed my newly created address book through the Data Science Toolkit API, which responded with latitudes and longitudes which I saved. Finally, I utilized the popular mapping library called Leaflet in order to plot the gathered addresses on a map:

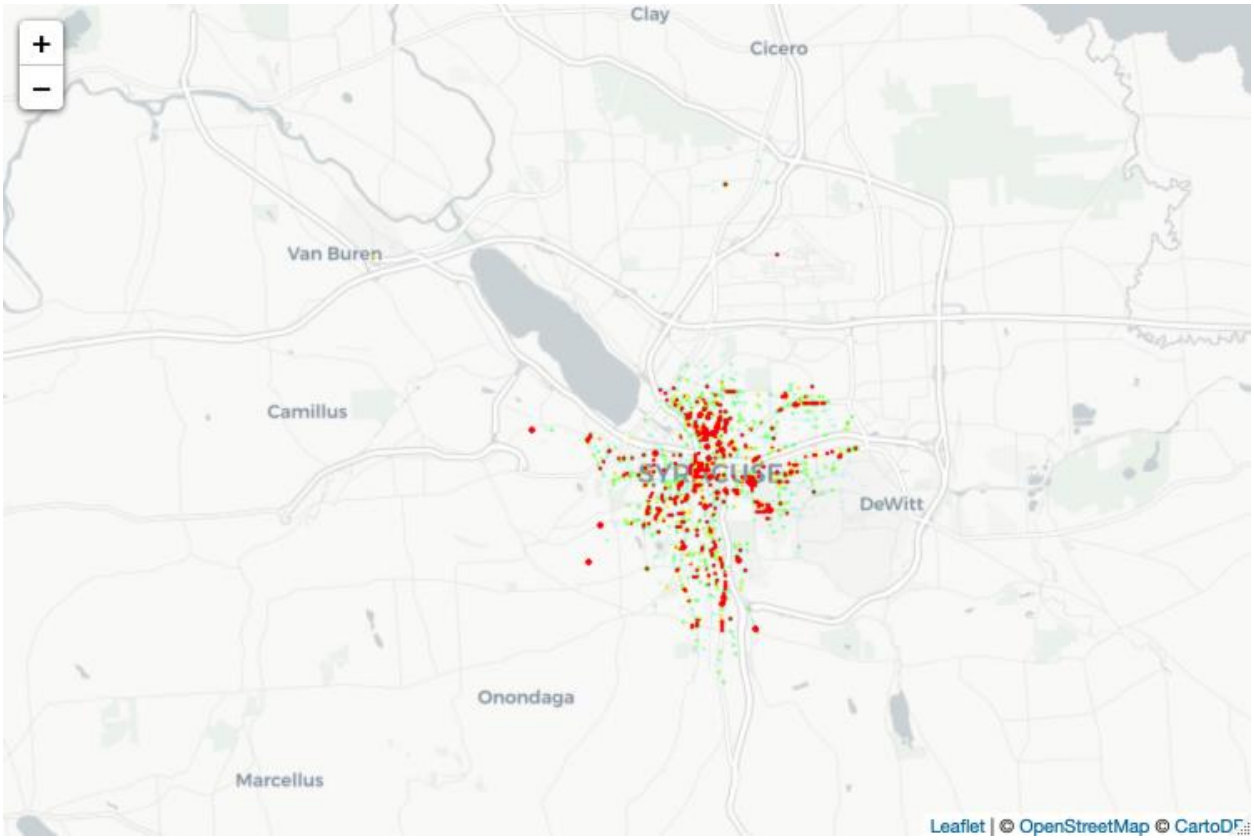

Figure 1: Heat Map of Crime in Syracuse

By looking at the hotspots (conveniently shaded in red) compared to the areas with less crime density (shaded green), it becomes easy to visualize the distribution of crimes in the greater Syracuse area.

Date Time Analysis

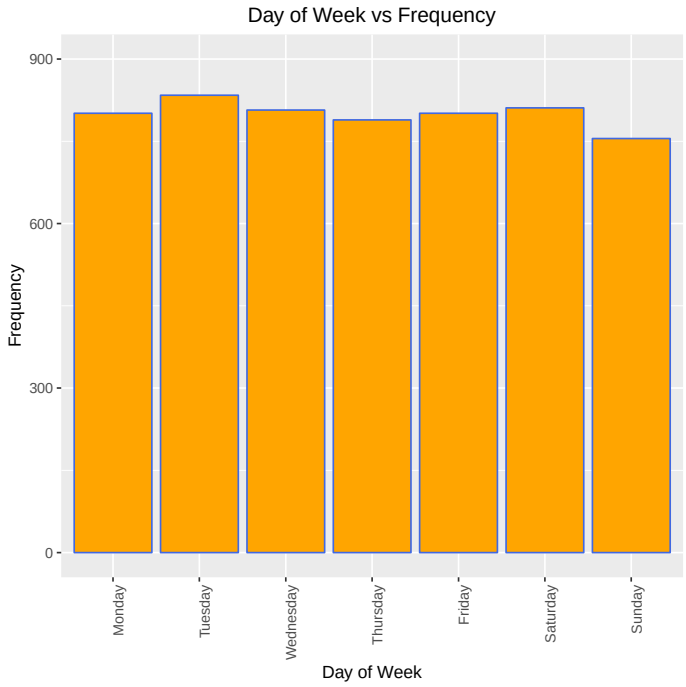

Figure 2: Week Day Frequency of Crimes

I began my time-based analysis by aggregating the number of crime occurrences by day of the week in order to look at overall crime frequency for each day. Overall, the trend is relatively consistent with most days hovering somewhere around the 800 mark for the year. However, it is worth noting that the count for Sunday is 755, well over one standard deviation from the median of 799.71. In this way, it appears as though Sundays truly do provide some rest for the wicked.

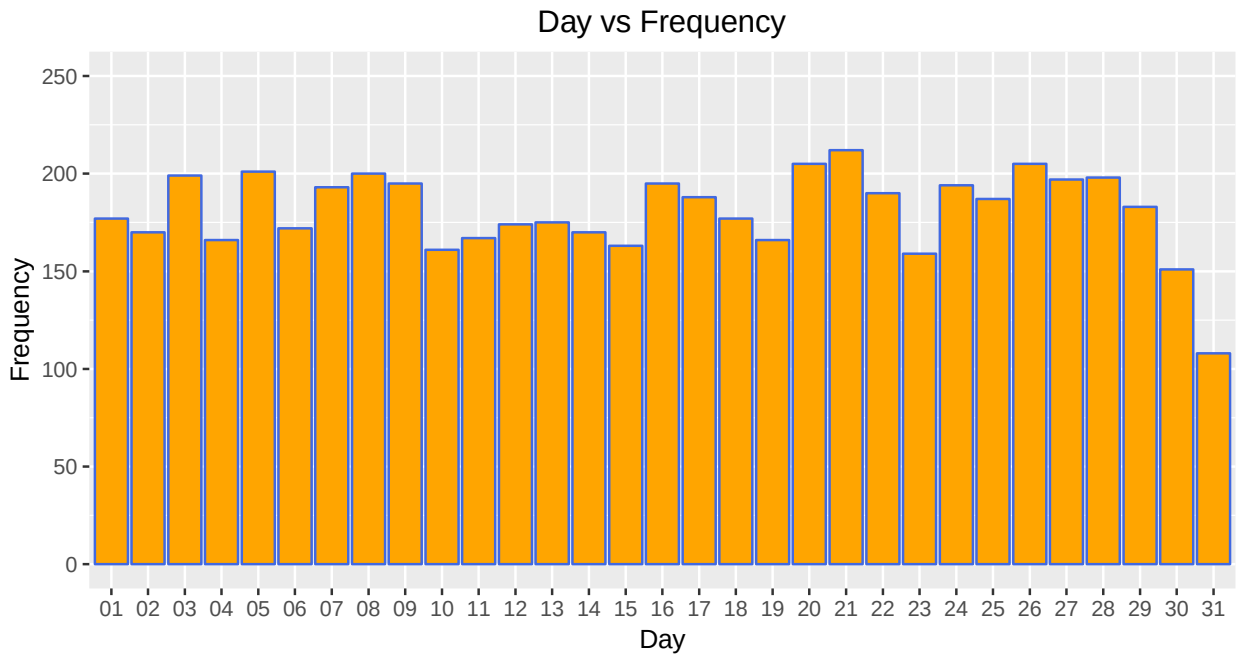

Figure 3: Daily Frequency of Crimes

My initial expectation was to find a relatively consistent trend along days of the month, which appears to be the case keeping in mind that this data is only from a singular year. There is a difference of 53 between the day with the most crime (212 on the 21st) and the day with the least (159 on the 23rd), not including the 30th or the 31st. However, if we were to have access to more data, perhaps the last 10 years, I would expect to find any inconsistencies here continue along the trend of leveling out.

One interesting detail to note is the decline in crime from the 29th to the 31st, likely due to the fact that only some months have that many days. It is also for this reason that I left these days out of the previous measure of days with the least crime.

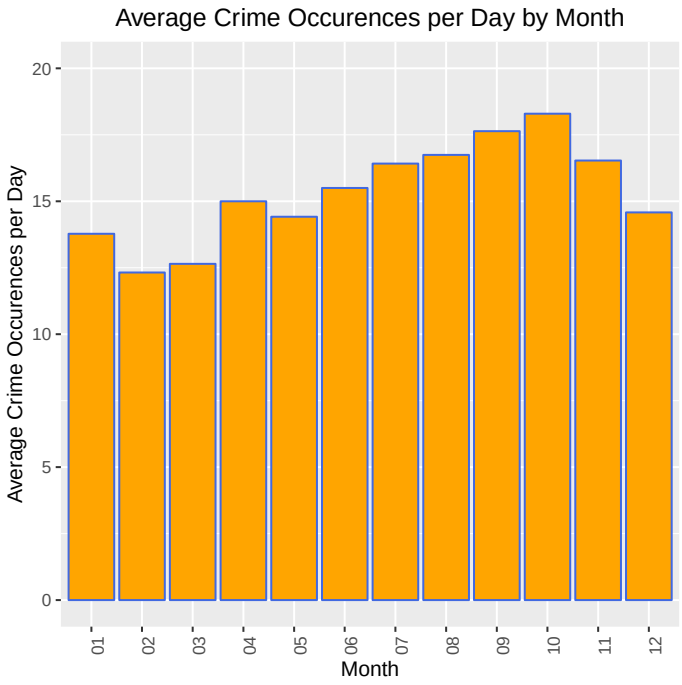

Figure 4: Monthly Frequency of Crimes

Many of the same presumptions can be brought into our analysis of the monthly frequency; we immediately notice that February, the month with the least number of days, has the least amount of crime occurrences. However, the overall number seem to follow a trend, being the lowest in February (12.3) and steadily rising each month to their peak in October (18.3) before lowering once again.

One possible theory is that crime may at least be correlated to warm weather; instances of crime increasing from the beginning of Spring, through the Summer, until they fall back off as the weather plummets towards the end of the Fall. However, to reiterate, this trend may simply be the result of variability within the available data.

According to Sam Edelstein, the reduction in crime from January to February is the result of students at Syracuse returning from break and discovering that their houses or belongings had broken into or damaged. While this is interesting in and of itself, it also

presents a potential problem in the crime data as a whole: when crimes are reported is not necessarily when they occurred.

Address Analysis

After initially plotting the number of crimes per each address, the graph was nearly unintelligible with the sheer number of addresses. In order to streamline my analysis, I elected to restrict my data to addresses reporting over 15 instances of crimes so as to focus on high volume locations.

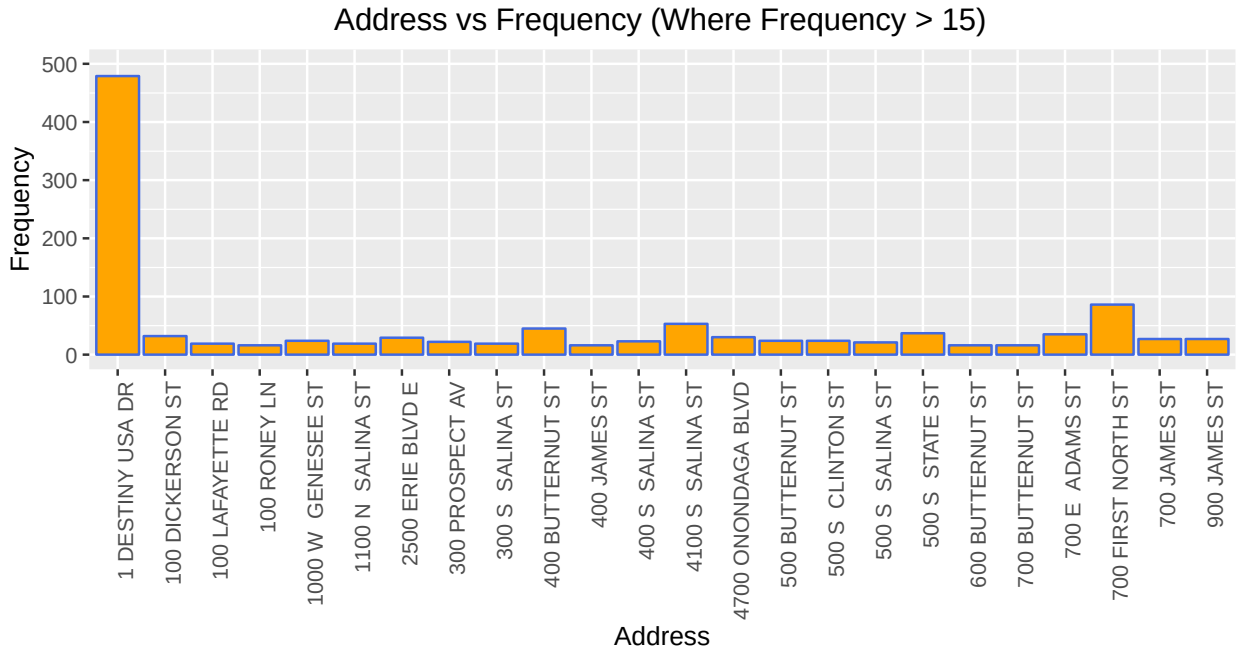

Figure 5: Crime Frequency by Address

Viewing my initial plot of crime frequency by address (which has been anonymized to the block level), one location clearly stands out. With a count of 479 crime reports, 1 Destiny USA Dr (better known as Destiny USA Mall) is hundreds of reports ahead of the nearest competitor, 700 First North St. at 86. When considering why this might be

however, it seems to make sense due to the highly trafficked nature of the mall. To better understand this, I plotted the codes defined for crime occurring at Destiny.

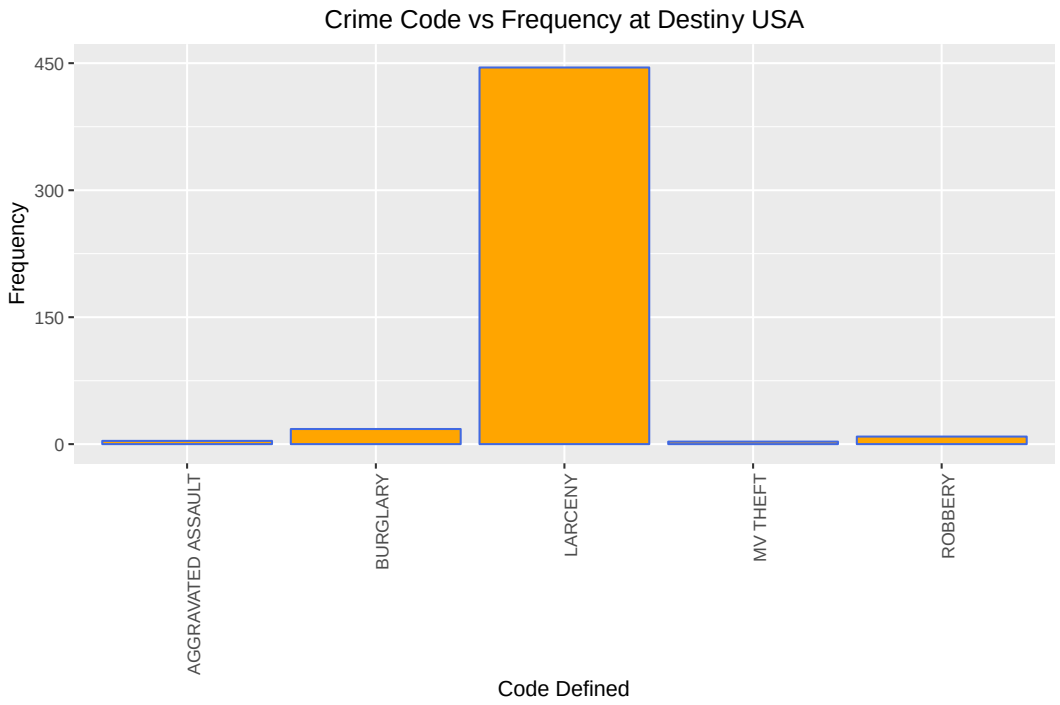

Figure 6: Crime Frequency at Destiny

As the 6th largest mall in the United States, Destiny attracts plenty of customers from the region and beyond meaning a large, highly concentrated population within its walls. Additionally, there are plenty of stores with plenty of customers walking in and out frequently, making larceny potentially easier than other locations. To this effect, the vast majority of reported crimes at Destiny were forms of larceny, 445 larceny reports to 34 of all other types. However, of the 479 crimes that occurred at Destiny, only 226 led to arrests. For this reason, due to the density of crime and the ability to condense resources into a reasonably small geographic area, it appears that investing more resources in the Destiny USA area may cut the amount of larcenies in Syracuse dramatically.

Having examined Destiny a bit closer, I then removed it from our previous plot as it is an outlier, if a legitimate one, that makes the rest of the data a bit difficult to interpret. The new plot is significantly more legible.

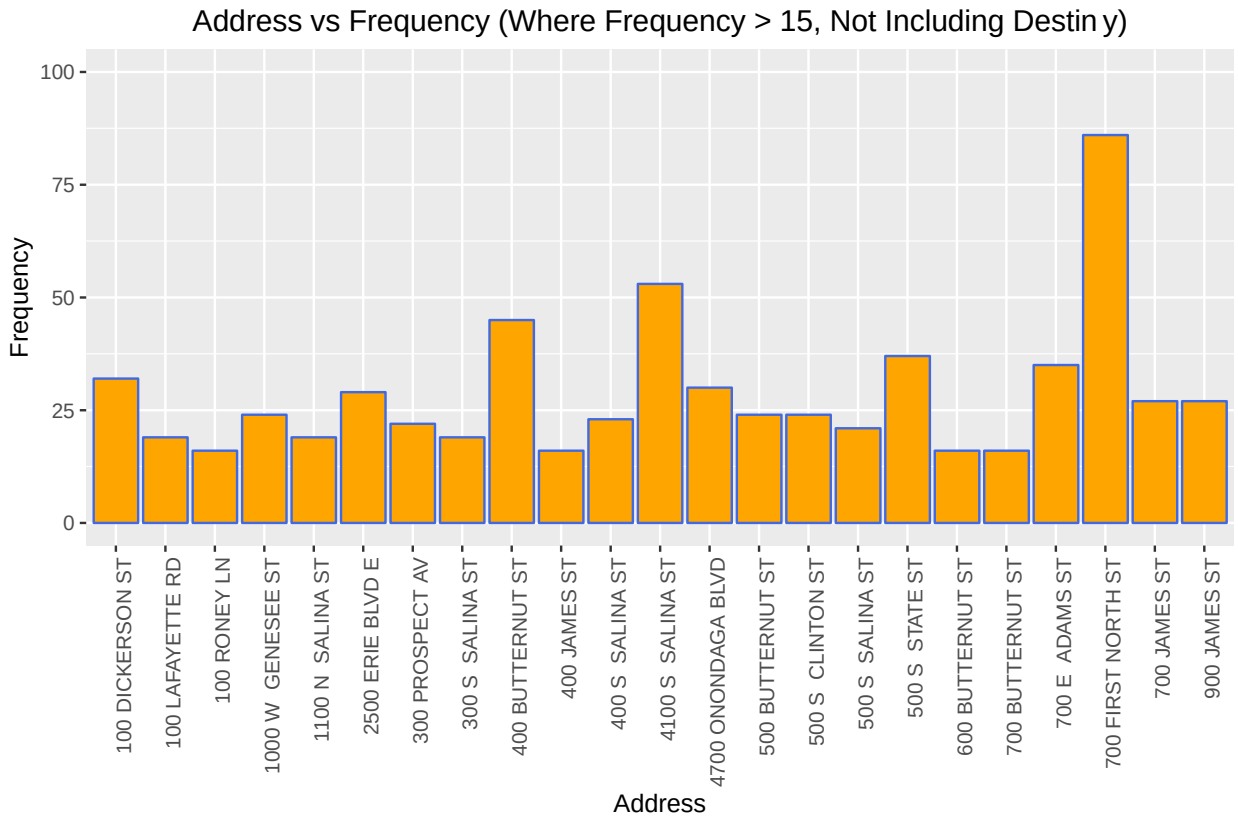

Figure 7: Crime Frequency by Address (Refined)

From here, it remains apparent that 700 First North St is a hot spot, along with 400 Butternut St and 4100 S Salina St. Looking at Google Maps, it appears that 700 First North St is located on the Northside and encompasses a block of residential housing across the street from a large Tops; a possible candidate for the cause of the crimes. Interestingly, 4100 S Salina St includes a small strip mall along with another similarly sized Tops, 400 Butternut St also being located by a large drug store. Although anecdotal, this pattern of high crime blocks encompassing large grocery and drug stores is interesting to say the least.

Code Defined Analysis

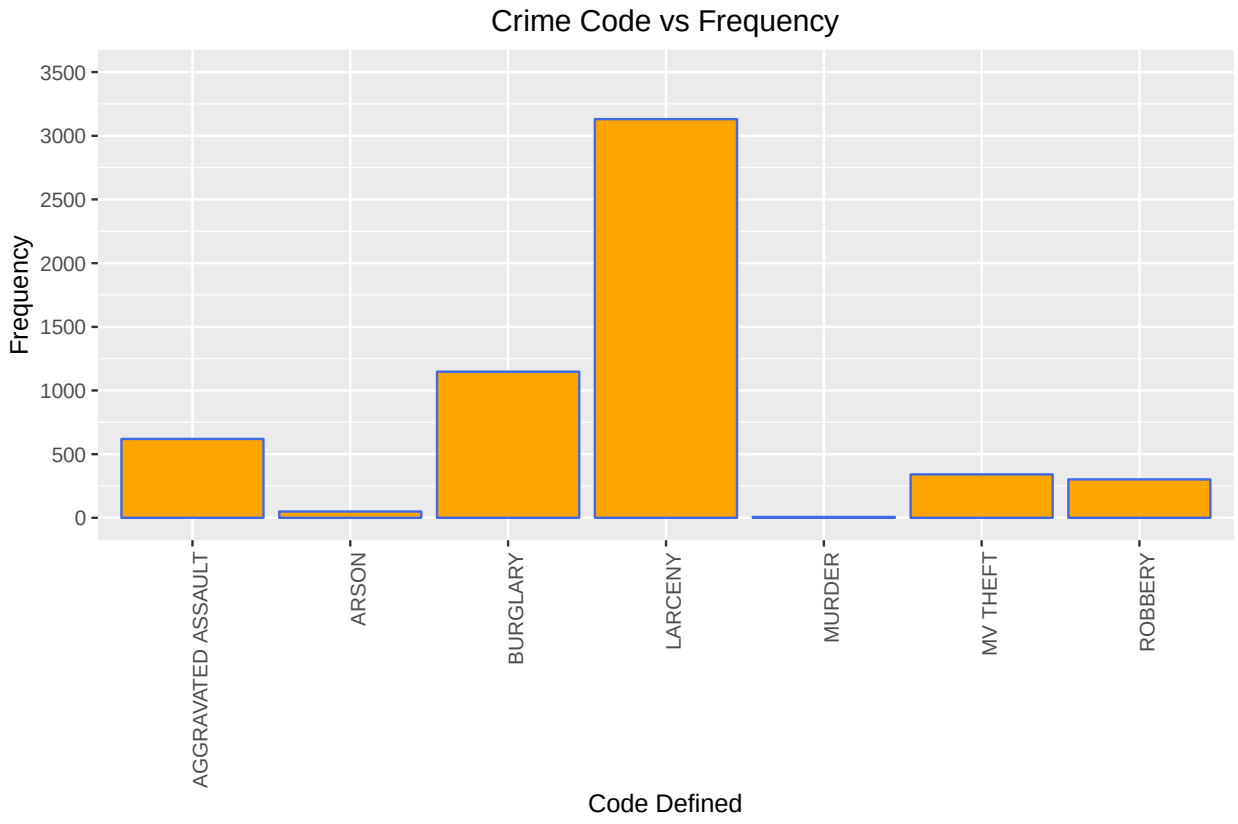

Figure 8: Crime Frequency by Code Defined

By examining the 'CODE_DEFINED' column, I hoped to get a better understanding of the types of crime most frequently occurring in Syracuse. Unsurprisingly, larceny remained the most common form of crime at 3131 counts, followed by burglary at 1148 and then aggravated assault at 620. Of the crimes reported, 4620 or 82.5% are nonviolent (burglary, larceny, motor vehicle theft) while only 978 or 21.1% were violent. However, of the violent crime, there were 7 murders and 49 instances of arson, a rather unsettling figure.

	MONTH	DAY	TIMESTART	TIMEND	ADDRESS	CODE_DEFINED	Attempt	Arrest	LarcenyCode

2639	07	05	0	0	2100 E FAYETTE ST	MURDER	0	1	Not Larceny
2977	08	26	0	0	300 N SALINA ST	MURDER	0	1	Not Larceny
3023	05	27	15	15	400 SHONNARD ST	MURDER	0	1	Not Larceny
3048	12	12	1	1	300 MERRIMAN AV	MURDER	0	0	Not Larceny
3099	09	25	21	21	1100 AVERY AV	MURDER	0	1	Not Larceny
3100	09	25	21	21	1100 AVERY AV	MURDER	1	1	Not Larceny
4595	04	03	22	22	500 GIFFORD ST	MURDER	0	0	Not Larceny

Figure 9: Murder Related Data

Of the 7 murders, row numbers 3099 and 3100 immediately stand out; this was most likely a double homicide considering both crimes occurred on the same day, at the same time, and at the same location. Of the 7, only 5 arrests have been made, although aside from that each occurrence is rather unique as far as the data goes.

Additionally, for larceny related crimes (which consist of larceny, burglary, motor vehicle theft, aggravated assault, arson) there is usually an accompanying code designating the type of larceny conducted. The reason for the caveat of usually is that, for some crimes such as aggravated assault or arson, the larceny element is not inherent and therefore there are many occurrences where no such value is entered. In order to better visualize

the frequency of each code, I have removed all entries without larceny codes from the data.

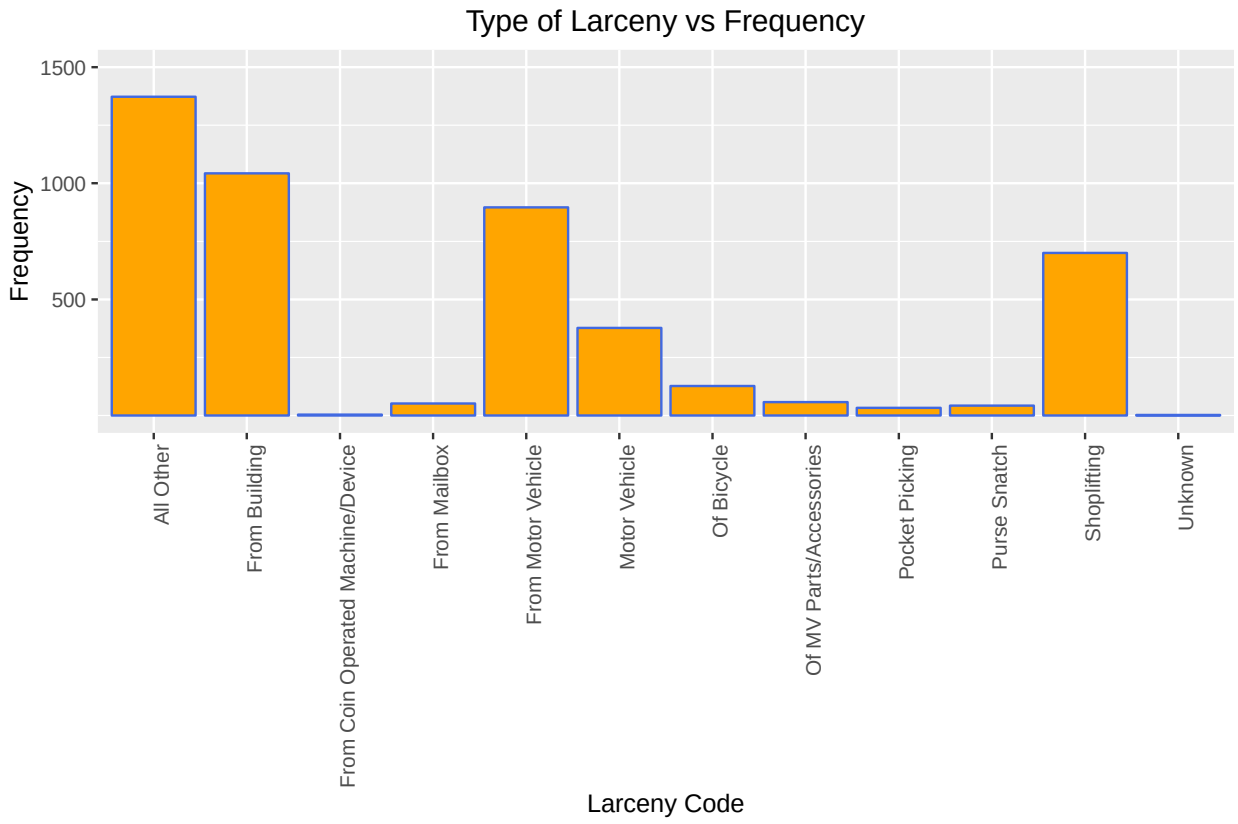

Figure 9: Crime Frequency by Larceny Code

Aside from the catch-all category of ‘All Other’, ‘From Building’ appears to be the most common category at 1043 occurrences followed closely by ‘From Motor Vehicle’ (896) and ‘Shoplifting’ at 700. Aside from the other catch-all category of ‘Unknown’, the next lowest category is ‘From Coin Operated Machine/Device’, likely a relic from the golden age of traditional parking meters. It appears that even criminals are aware of newer parking technology!

Arrests Analysis

Finally, I looked at the ‘Arrests’ indicator, a marker of whether or not an arrest was made in conjunction to the crime reported.

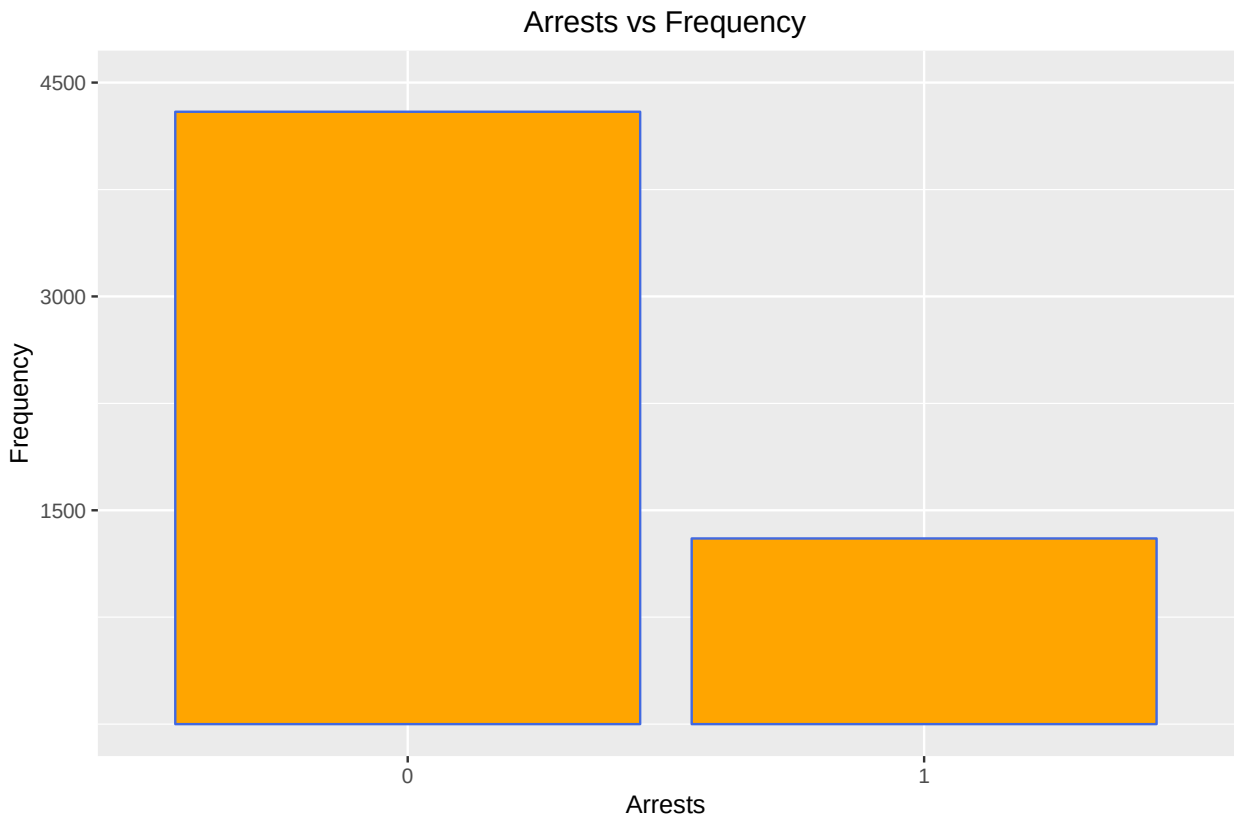

Figure 10: Arrests

Of the 5598 total crimes, only 1302 (or 23.2%) actually resulted in an arrest. By subsetting the data to only crimes which resulted in an arrest, I looked for trends as to who is most frequently caught.

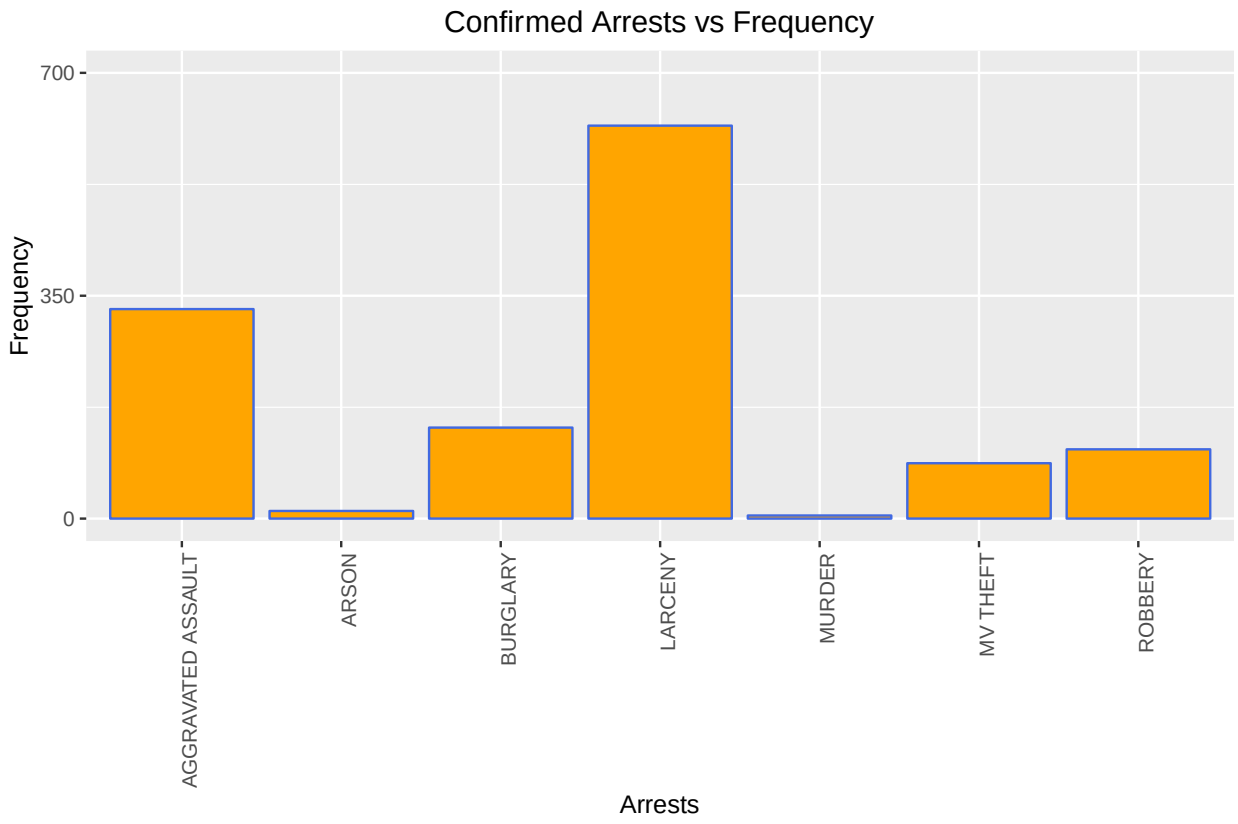

Figure 11: Confirmed Arrests by Code Defined

Viewing the subset of the 1302 confirmed arrests, the trends in arrests by code defined generally mirror the initial plot of Crime Frequency by Code Defined (Figure 6), with larceny leading by a wide margin, trailed by aggravated assault and burglary, then motor vehicle theft and robbery, and finally arson and murder. However, in this case, catching those committing aggravated assault appears to be more common than catching burglars.

Predicting Arrests

Data Preparation

Having generated some interesting insights and plotted most of the information available, I then moved on to predictive modeling for arrests. Predictive modeling utilizes machine learning algorithms in order to analyze data and attempt to make predictions about new, unseen data. In order to create my models and test them for accuracy, I created both a training data set and a testing data set from my initial data set, used for training the model and then testing it respectively.

In order to create these sets, I created a cut point 2/3 of the way through the data set and then randomized the rows present in either portion to ensure that similar lumps of time were not grouped together. By setting a seed (an arbitrary number) for this randomization, the process can be ensured to be random but consistently so; the results of running the code will always be the same. This is important for the sake of reproducibility, that anyone with access to my code would be able to recreate the same results themselves.

Support Vector Machines (SVM) Modeling

For my first model, I elected to utilize the Support Vector Machines or SVM modeling algorithm. SVM is a classification algorithm usually used for binary classification, provided a set of data it will plot it on an X Y axis and attempt to draw a line to separate the points on the plane into a series of two groups; effectively classifying them by which side of the line they fall on. For the sake of predicting arrests, utilizing this model makes sense as we have a binary variable (either an arrest or not) that we would like to classify based on the other available variables from our data.

Essentially, the model predicts ‘Arrest’ vs the rest of the data from the trainData variable, utilizing the kernel rbfdot. The cost indicates how harshly the classification line is drawn and the cross parameter divides the data into 10 groups within the trainData for testing to decrease the chances of an anomalous result. Finally the prob.model parameter indicates that the model’s output is a probability model of whether an arrest occurs or not.

Having generated the model, I then ran it against the test data I had previously generated to see how accurate it was. The result was an accuracy of 79.5%.

Support Vector Machines Model			
Accuracy: 79.5%		**Prediction**	
		No Arrest	Arrest
testData	No Arrest	1286	130
	Arrest	251	199

However, looking at the confusion matrix pictured above, this accuracy is not entirely honest. The purpose of a confusion matrix is to demonstrate how many instances the classifier determined correctly, as well as how many false positives and negatives were selected. In this case, although the model predicted non arrests well (1286 right to 251 wrong), it did not do so well with successful arrests (199 right to 130 wrong). In order to correct this, I subsetted the sample data once again to include an equal distribution of successful and non-arrests before putting it through the same 2/3 training and testing analysis. The results are as follows:

50/50 Support Vector Machines Model

Accuracy: 74.0%		**Prediction**	
		No Arrest	Arrest
testData	No Arrest	354	107
	Arrest	118	289

As you can see, although the accuracy of the model has fallen a bit, the classification itself appears to be a bit more robust as far as false positives are concerned. However, the best way to test these models would be to run both of them on an entirely new data set, such as another year of crime, and see which performed better.

Random Forest Model

In my goal of predicting arrests, the other algorithm that I utilized was the Random Forest algorithm. Random Forest functions are a form of a more basic algorithm known as Decision Tree in which the data set is systematically split based on decision points, starting with broad similarities and differences and becoming more specific as the branches of the tree extend outward. However, due to variance in the data, this end result might be more or less accurate depending on the luck of the draw, which is where Random Forest comes in; it performs the Decision Tree classification many times and then takes average values from the results to generate one, consistent, tree.

Due to the way that Decision Trees are formulated, they are unable to accept columns with more than 53 unique factors, meaning that addresses had to be removed so as to allow the model to function correctly. After noticing its adverse effect on the accuracy of the models, I also removed the column containing days of the week.

Having recognized the problem of the skewed data from the SVM model, I have again created two different decision trees; one with the skewed distribution of the database itself, and another with a 50/50 split, in hopes of eventually testing them against a more complete database. The results of these runs are as follows:

Random Forest Model			
Accuracy: 78.6%		**Prediction**	
		No Arrest	Arrest
testData	No Arrest	1232	184
	Arrest	215	325

50/50 Random Forest Model			
Accuracy: 71.3%		**Prediction**	
		No Arrest	Arrest
testData	No Arrest	283	178
	Arrest	71	336

The most notable difference between the Random Forest and SVM models is the amount of predictions of successful arrests, whether this is correct or not. In this way, it is apparent that, although both modeling formats are being utilized to predict the same thing, their results can be quite different. Overall, the Random Forest models both appear to be less accurate overall compared to their SVM counterparts.

However, despite their slight decrease in accuracy, Random Forest models are also much more interpretable compared to the black box nature of most SVM models. For

instance, below is a multi-way importance plot meant to indicate the most important variables in creating the Random Forest.

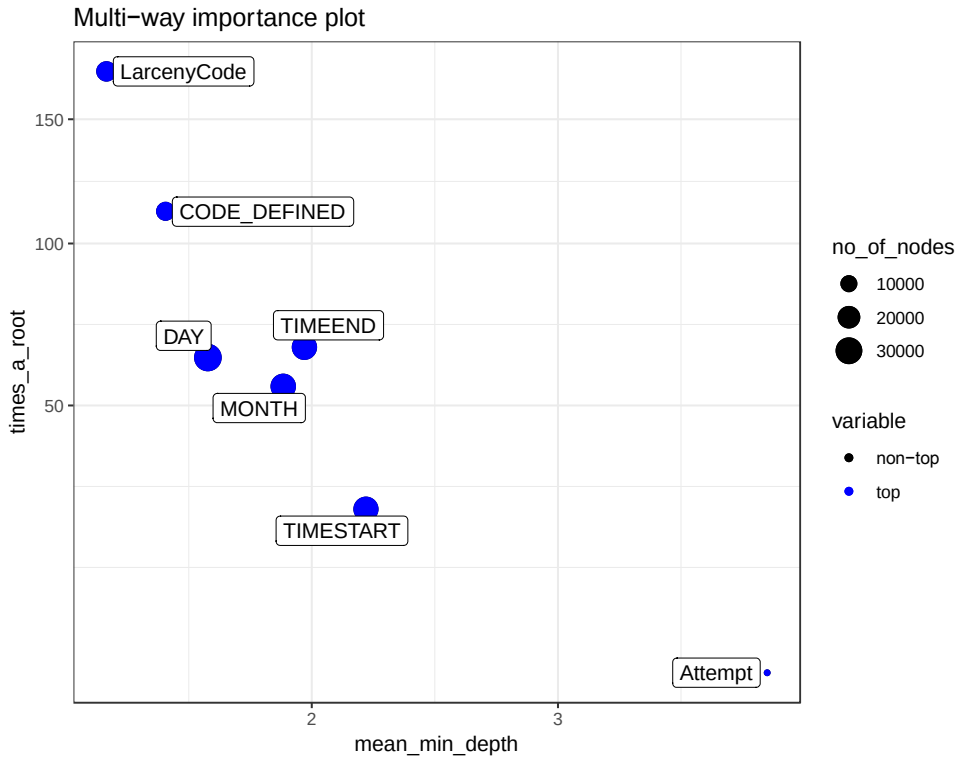

Figure 12: Random Forest Multi-Way Importance Plot

The x-axis, labeled mean_min_depth indicates the average minimal depth of the variable at each of the individual Decision Trees utilized in the creation of the overarching forest. The y-axis, labeled times_a_root indicates the number of times that a variable was the splitting factor between two branches of the decision tree. With these definitions in mind, it is easy to see that LarcenyCode and CODE_DEFINED are the most useful variables in determining whether an arrest was made or not.

However, the fact that TIMEEND is among the four most significant variables is rather interesting, indicating that the time at which an initial investigation concludes may be more important in determining if an arrest is made than the time that the crime actually

occurs. Finally, it is worth noting that the Attempt variable is of little to no use in this decision making process.

Insights

The following are some insights and potential topics for further investigation that resulted from my analysis. To further expound upon these would involve discussion with law enforcement personnel and those with more specialized domain knowledge, as such these topics have not been fully realized.

Increase police and/or security presence around Destiny USA, it may dramatically impact larcenies occurring in the Syracuse area. Destiny has a dramatically higher number of crime reports than any other address at 479, the vast majority of which are larcenies (445). Increasing police presence or implementing other measures such as secuty cameras in this area could lead to a reduction in larcenies.

Investigate the correlation between large department stores (such as Tops) and why crime is particularly high at 700 First North St, 400 Butternut St, and 4100 S Salina St. Increasing police presence here may also prove worthwhile. Understanding this trend and the underlying factors may help not only reduce crime in these areas, but also identify other potential hotspots for activity.

As the vast majority of crimes reported involve robbery or larceny codes (76.4%), it may be worth investigating how many households have home security systems and what measures they are taking to prevent such crimes as well as providing additional information sessions and/or material on preventative measures for

citizens to employ. In this way, the number of robberies and larcenies may be reduced without the need for more traditional policing.

Conclusion

Over the course of this project I have had the opportunity to not only better understand the distribution and nuances of crime in Syracuse, but to practice my Data Science skills as well. Dealing with raw data such as this was a learning experience in both how to properly clean data, and how to engineer new information, such as the appended addresses used to create the heat map.

By performing data mining, it was also possible to uncover new information in crime trends such as the fact that the vast majority of the crimes at Destiny USA are larceny, that larcenies are the crime that most commonly results in an arrest, and that 2/3 blocks with the most crime have a Tops (not including Destiny USA). The process of framing the questions to discern this information and its potential usefulness in predicting future crimes as a result.

Finally, through the use of predictive modeling, it was possible to predict whether or not a crime would result in an arrest with between a 70% and 80%. While interesting and potentially useful, I would enjoy the ability to test these models against a larger test data set, a goal that will hopefully be realized as more crime information on Syracuse is inevitably gathered.

Bibliography

Edelstein, Sam. “Part I Crime Definitions.” Github, 4 Dec. 2017, github.com/CityofSyracuse/OpenDataDictionaries/blob/master/PartICrimeSelected.pdf.

Edelstein, Sam. “Weekly Part 1 Crime Offenses 2018.” Syracuse Open Data, 5 June 2018, http://data.syrgov.net/datasets/0583c4cbea2d4edf9f13e8dcbe21eefa_0

Vogt, PJ, and Alex Goldman. “Reply All: The Crime Machine, Part I.” *Reply All*, episode 127, Gimlet Media, 11 Oct. 2018.