Phoenix Crime Analysis

Capstone Project Final Report

Lyttia McManus

Springboard: Intro to Data Science

2018

Introduction

Phoenix, AZ is the 5th largest city in the country with a population of over 1,600,000 residents. People are increasingly adopting Phoenix as their new home, despite summer temps in the 120s and a poorly ranked public education system. In 2017, Phoenix showed the 2nd highest population growth among large cities with populations of 50,000 or more, adding 24,036 people in 2017 (US Census Bureau).

The Problem

With this population, the Phoenix Police Department is dangerously understaffed. Phoenix PD employs just over 3,000 officers, about 1,500 short of the national average, that are needed to police a city the size of Phoenix (According to Ken Crane, president of the Phoenix Law Enforcement Association). Although all hands are on deck and calls are prioritized, there has been increased call response wait times for many years. Living in Phoenix, new residents quickly learn that near city center, crime has become practically unavoidable.

As my first attempt at analyzing data in R, the goal of this project was to better understand and locate the areas of Greater Phoenix with higher and lower crime rates, as well as determine when, where, and which types of crime occur most often in the city. Examining the conditions in which these crimes occur can lead us to understand its contributing factors as well as what doesn't seem to be making a difference.

Goal # 1:

Strategically deploy officers and patrol areas of Phoenix by understanding crime trends:

- Locate the areas of higher and lower crime rates.
- Determine when, where, and which types of crime occur most often.

Goal # 2:

Examine the conditions in which these crimes occur in order to determine:

- Contributing factors.
- What doesn't seem to be making a difference.

We can then use this information to create a plan to make the city a safe place for everyone and families from moving away.

In addition, we can try to predict whether a crime is likely to be violent or nonviolent using what is known as a "random forest". This algorithm is a predictive modeling tool that can find patterns in data by randomly creating numerous decision trees and determining the importance of each indicator to the outcome. The algorithm produced by this process can then be used to predict whether a crime is likely to be violent or not, depending on the conditions presented and help to prioritize crimes being reported.

This type of prediction could also be used (along with crime hotspot maps and other visualizations as mentioned above), to provide decision support to police departments and inform city policy makers. Additionally, it could be used in a program that assesses and assigns a community's crime rating to provide decision support to home buyers, and could be listed on websites that assist in home buying, such as Zillow.

Important fields and information

The crime data comes from the City of Phoenix Open Data website found here: Phoenix Crime Data. It comes in the form of a "csv file which is updated daily by 11am and includes incidents from November 1st, 2015 forward through 7 days prior to today's posting date. The file can be updated to include new crimes on a daily basis. At the time of original download (7/2/2018 8:00pm), the data included 169,818 crimes and recorded data for 7 variables:

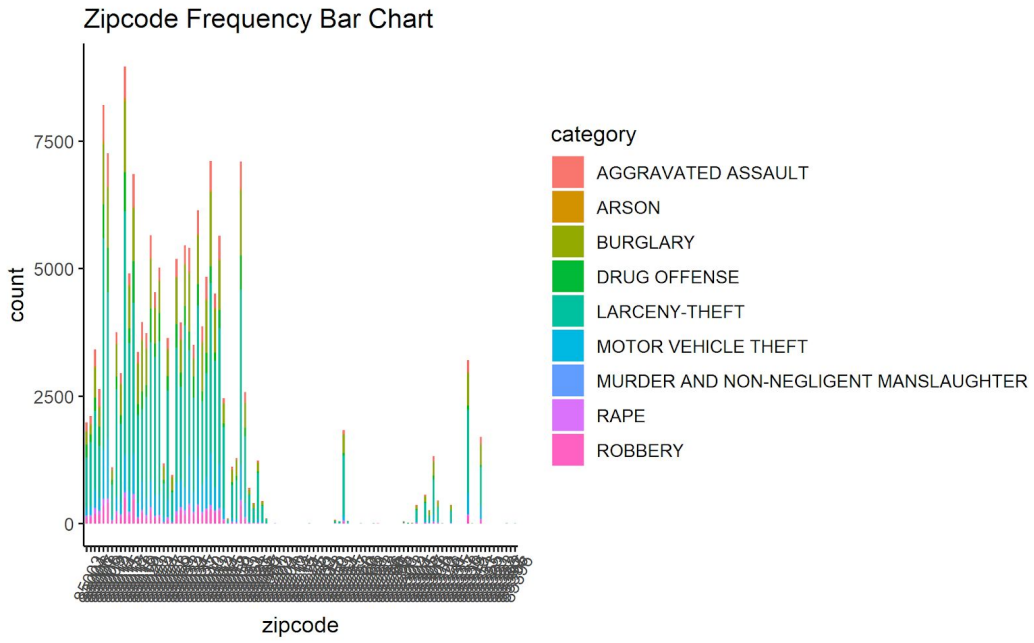
- Crime incident number
- Earliest and latest date and time the crime occurred*
- Uniform Crime Reporting (UCR) crime type:
    - Homicides, rapes, robberies, aggravated assaults, burglaries, thefts, motor vehicle thefts, arsons, and drug offenses*
- Hundred block address (precise address confidential)
- Zip code*
- Premise* (type of location)

Median house price by zip code (as a measure of prosperity) will come from realtor.com residential listings database found here: https://www.realtor.com/research/data. The data is "based on the most comprehensive and accurate database of MLS-listed for-sale homes in the industry". The data includes 1,094,091 observations of 34 variables. I am only using data from Phoenix zip codes- 2,993 entries- and 2 variables. Data was provided by zip code from 5-1-12 to 5-1-2018. Original variables include:

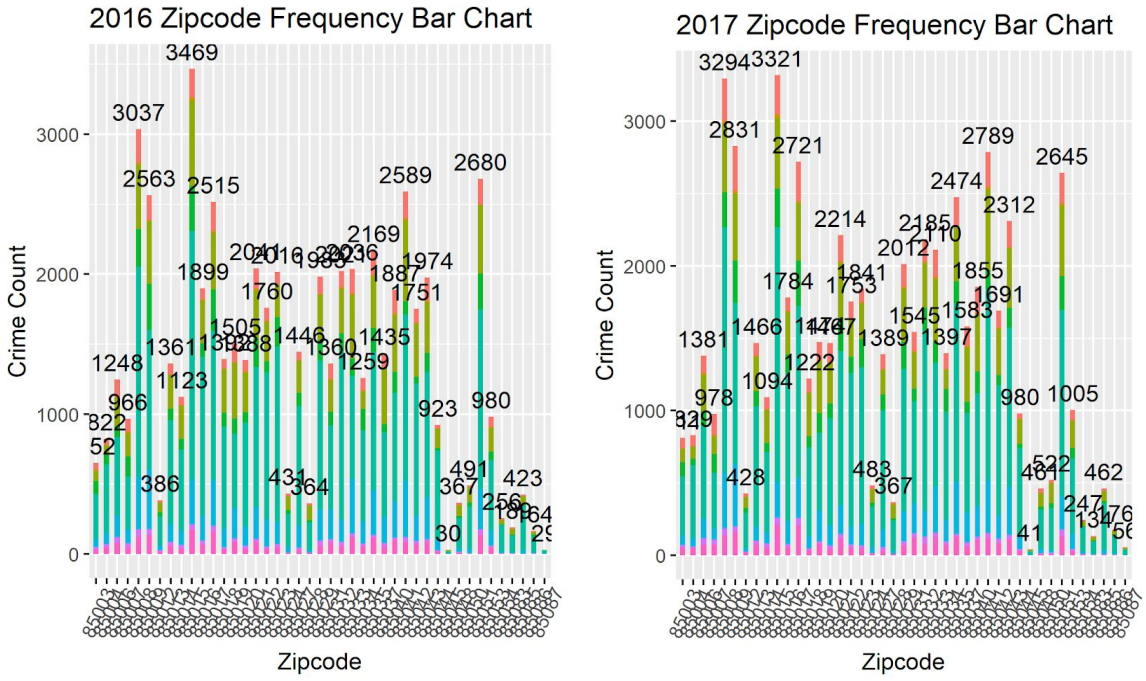
- Median & Average list price*
- Luxury list price
- Median & Average days on market
- Total active listings*
- New listings
- Price increases and price reductions

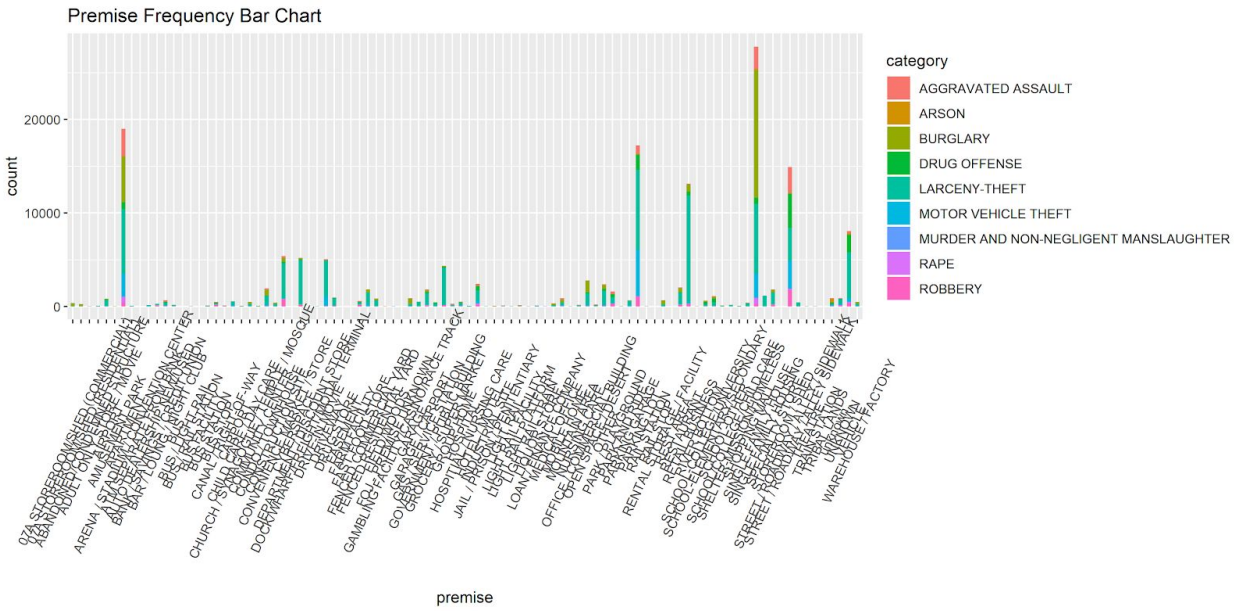
Cleaning and Wrangling

The data was wrangled using R scripts to clean, filter, and join the datasets as well as engineer other variables for use, and create some initial visualizations. Nearly all of the crime data remains intact, joined to columns containing the median listing price (property value), and number of surrounding listings (a measure of transiency) at the time and location of the crime. Many of the original variables contain capital letters and blank spaces which can be troublesome when writing code in R. The "OCCURRED ON" column is separated into multiple columns ("month", "day", "year", "hour", "minute") to make joining data frames, as well as more specific analysis, possible. The "OCCURRED TO" column is removed as it is irrelevant to the specific conditions leading to the development of each crime type. If any dates are not reported, or NA, the dates are filled by the order (incident number) in which the crime is reported. Dummy binary variables are added for crime categories to use in further analysis.

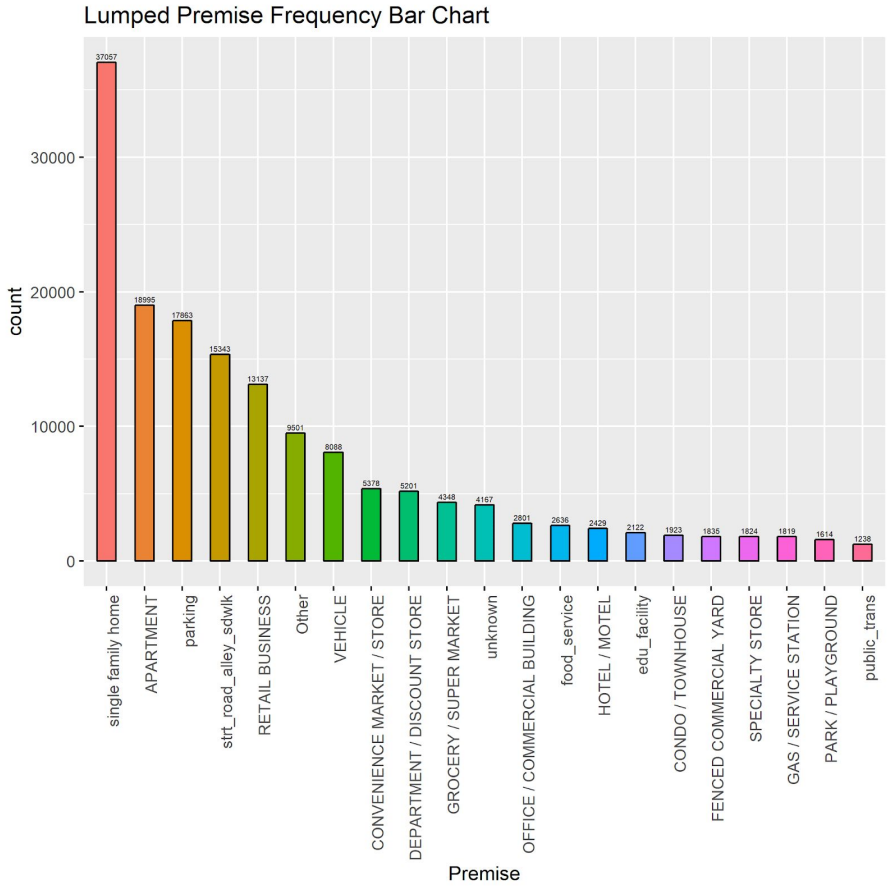


The graph above shows the crime count within each of the original 102 zip codes. Zip-codes were filtered from the original data to more accurately represent a true crime count, since many of these zip-codes are located outside of official Phoenix boundaries, and likely have many additional crimes reported to their respective police departments. Phoenix zip-codes start with 850. This left me with 43 zip-codes as shown below:





The graph above shows a count of the 95 unique premise types. These types are combined into premise categories, and are added with dummy binary variables. The lumped and ordered premise frequency chart is shown below:



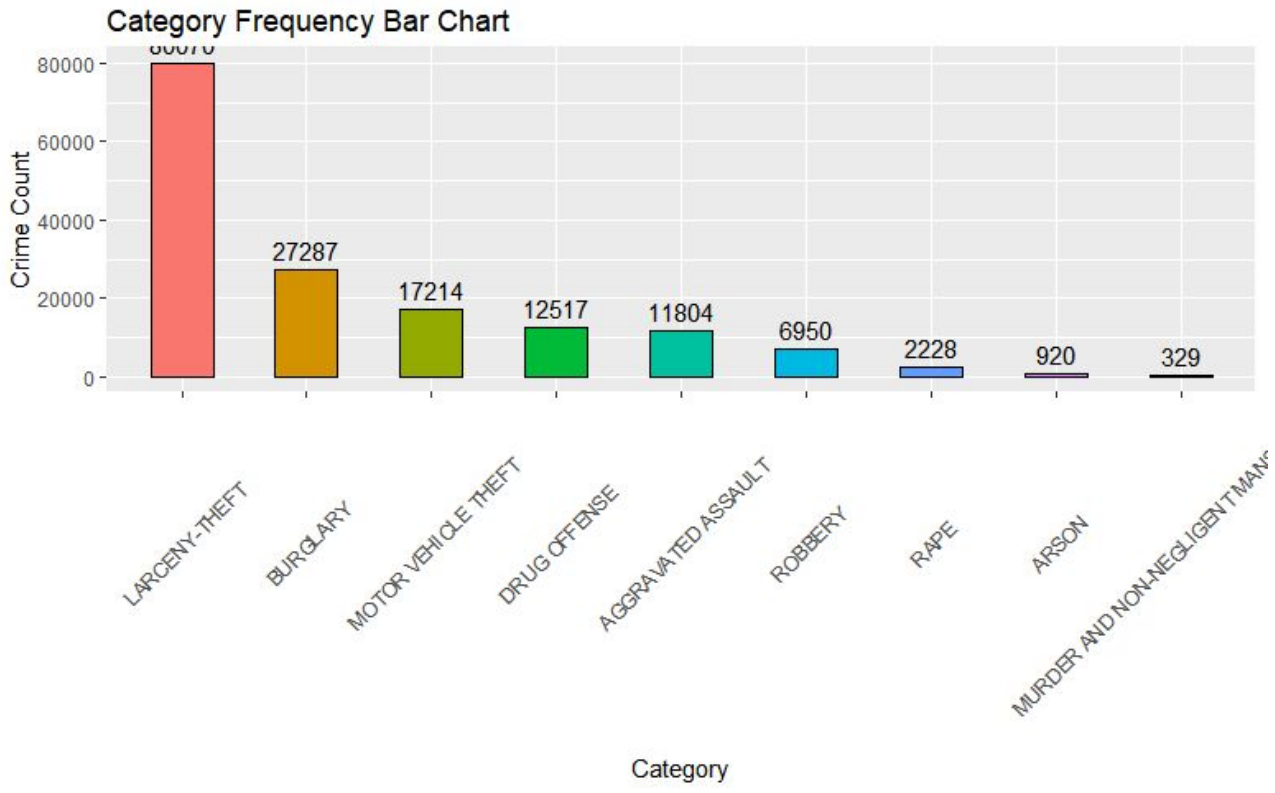
This graph consists of the top 20 premise types reported to PPD. The remaining premise types were combined into "Other", which accounts for 5% of the total data. The most common premise type is the single family home, which accounts for 23% of the data.

Many inconsequential variables are included in the Realtor.com data frame as it pertains to this project. Unnecessary variables are removed using dplyr. Property data joins the crimes data by "zip code", "month", and "year". This matches the crime to the median property value, accurate to the zip code, month, and year of the crime.[1] While exploring data for missing values, zip codes missing property values emerge. The NA zip codes are removed since an address or clues to a location are non-existent. The 85034 zip codes contain 3,511 crimes, so it is worthwhile to use historic property values from Zillow to fill these values.

Preliminary Exploration and Initial Findings

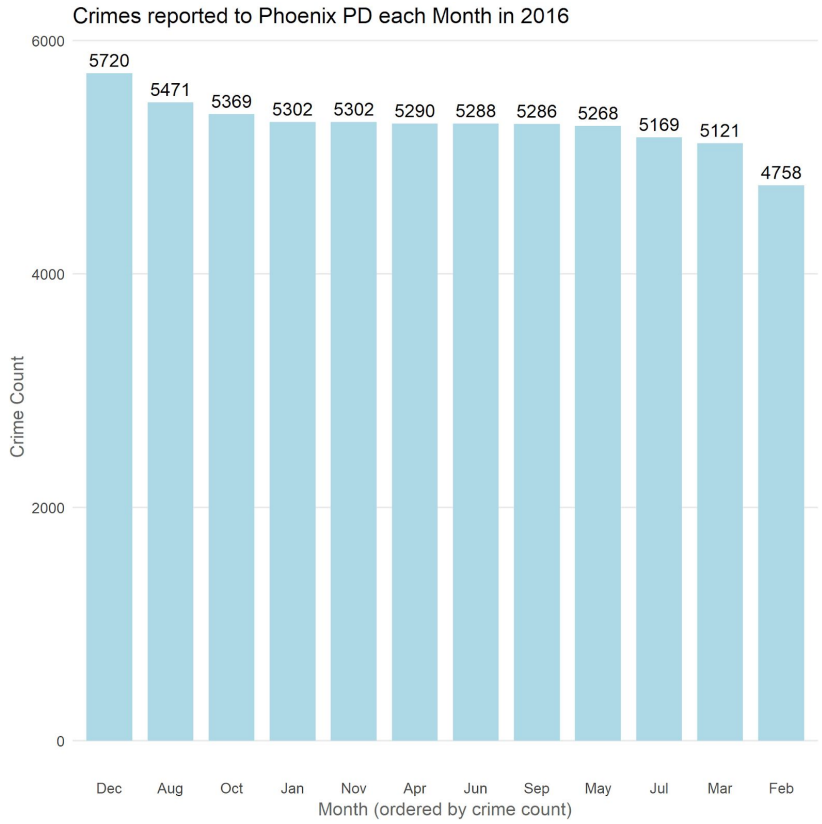
The section to follow will summarize some of the most important findings.

Of the nine categories of crime, larceny-theft was the most common crime category reported to PPD with 80,070 out of 159,319 total crimes, coming in at 50% of the crimes reported.

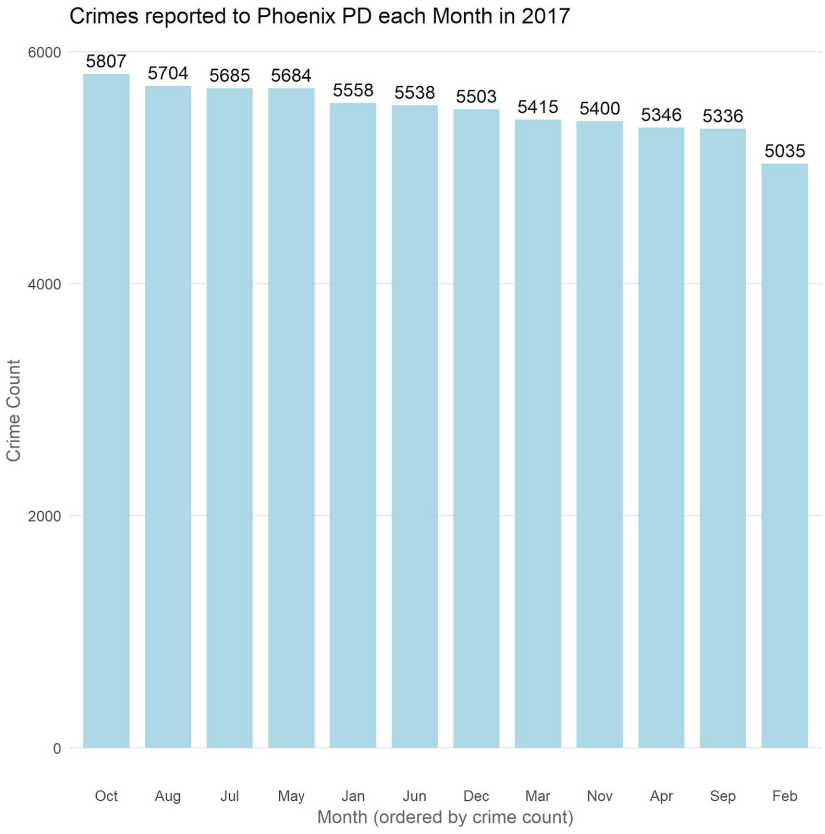


The next charts show 2016 and 2017 separately. We see bottom limits of about 5000 crimes to upper limits of 6000 crimes reported monthly and a range that dropped from 962 in 2016 to 772 in 2017, meaning that the crime count changed less from month to month in 2017.

[1] It is not possible to get up-to-date property value estimates on the addresses from the Phoenix Open Crime Data because addresses are displayed to the hundred block for confidentiality.

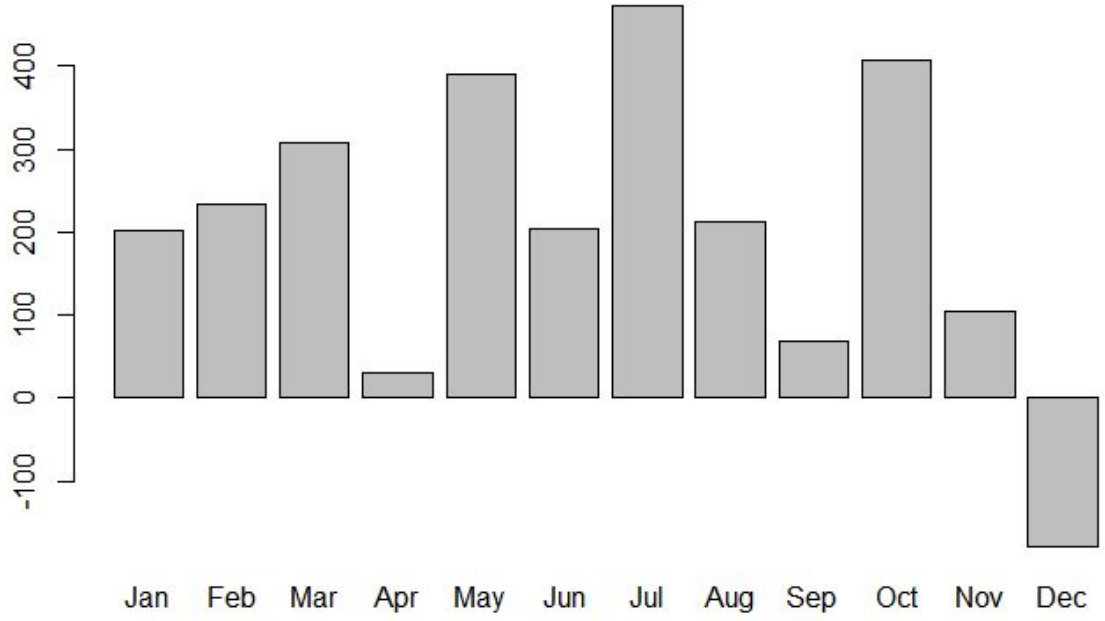


Range in 2016- 962



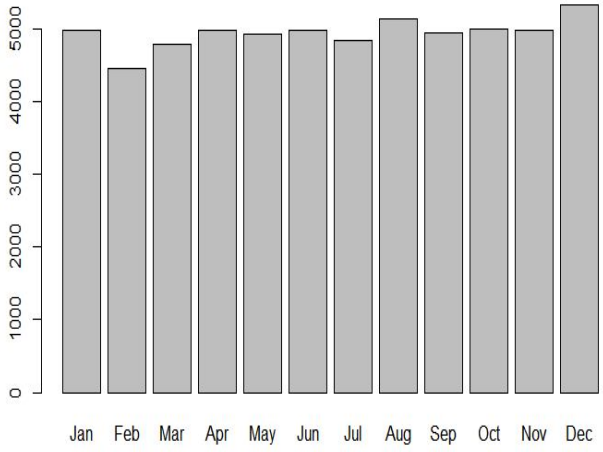
Range in 2017- 772

Monthly crime change from 2016-2017

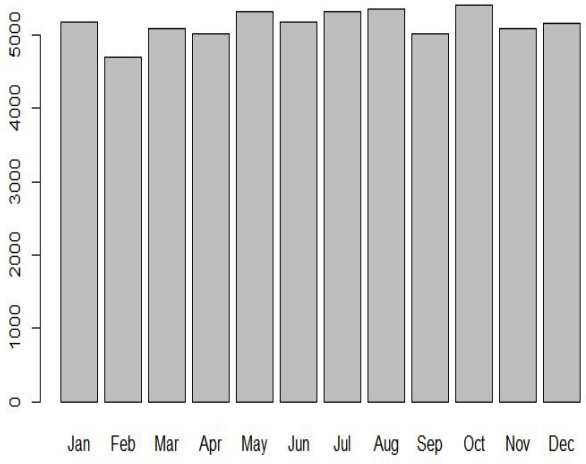


There were 2,450 more crimes in 2017 than in 2016, a 4.1% increase. Crimes increased each month in 2017 except for December in which crimes decreased. December went from being the highest crime month in 2016 to #7 in 2017.

Crime count each month in 2016

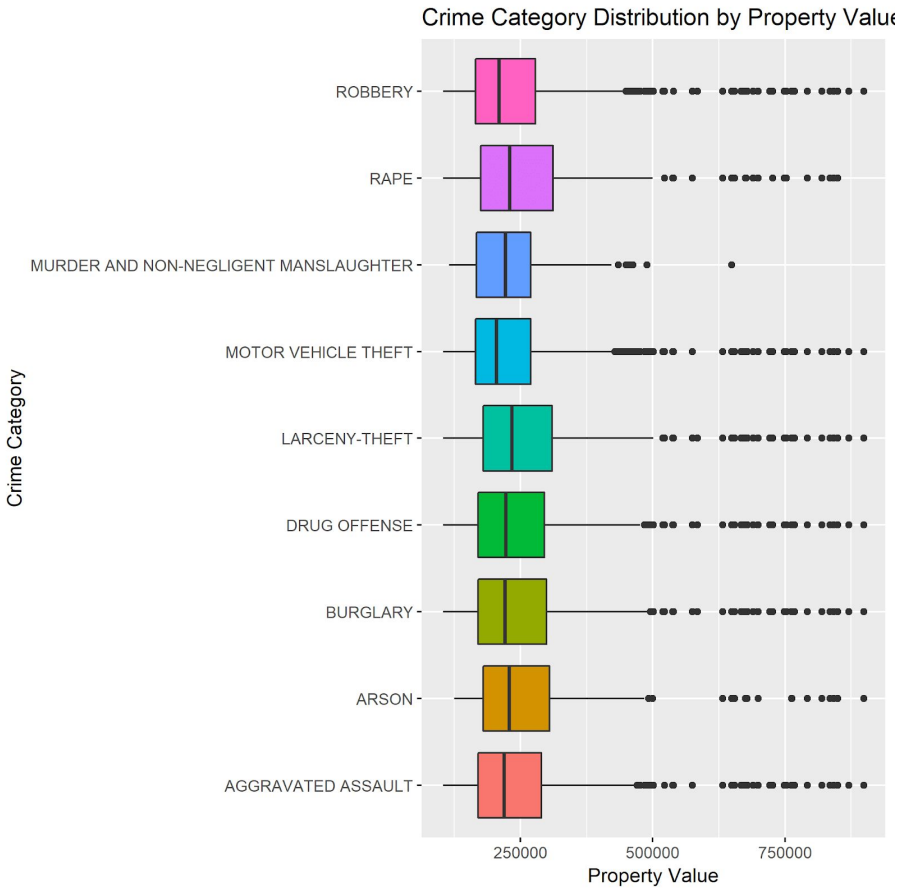


Crime count each month in 2017

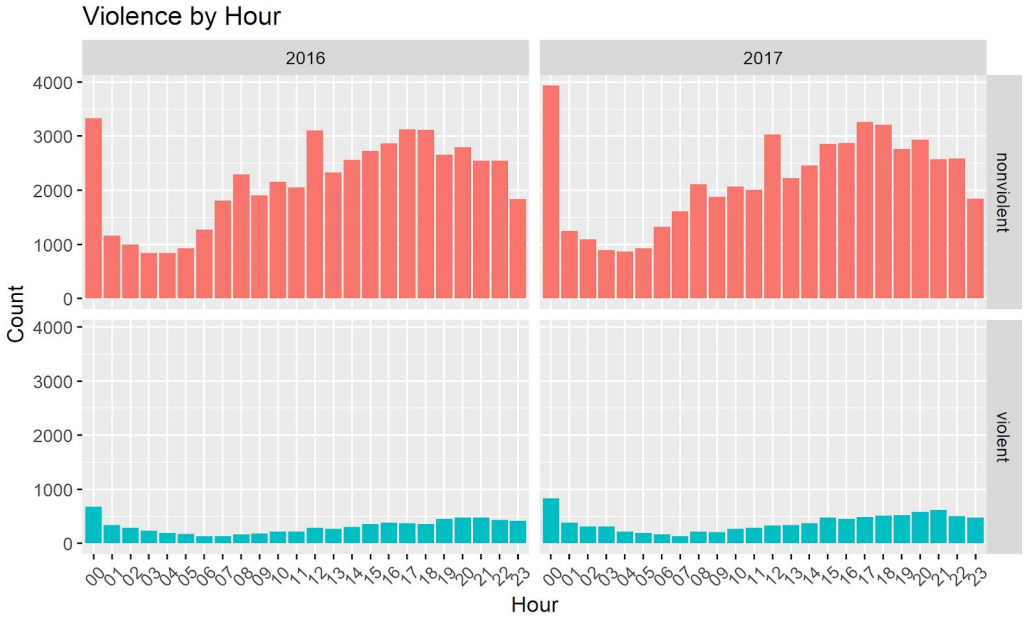


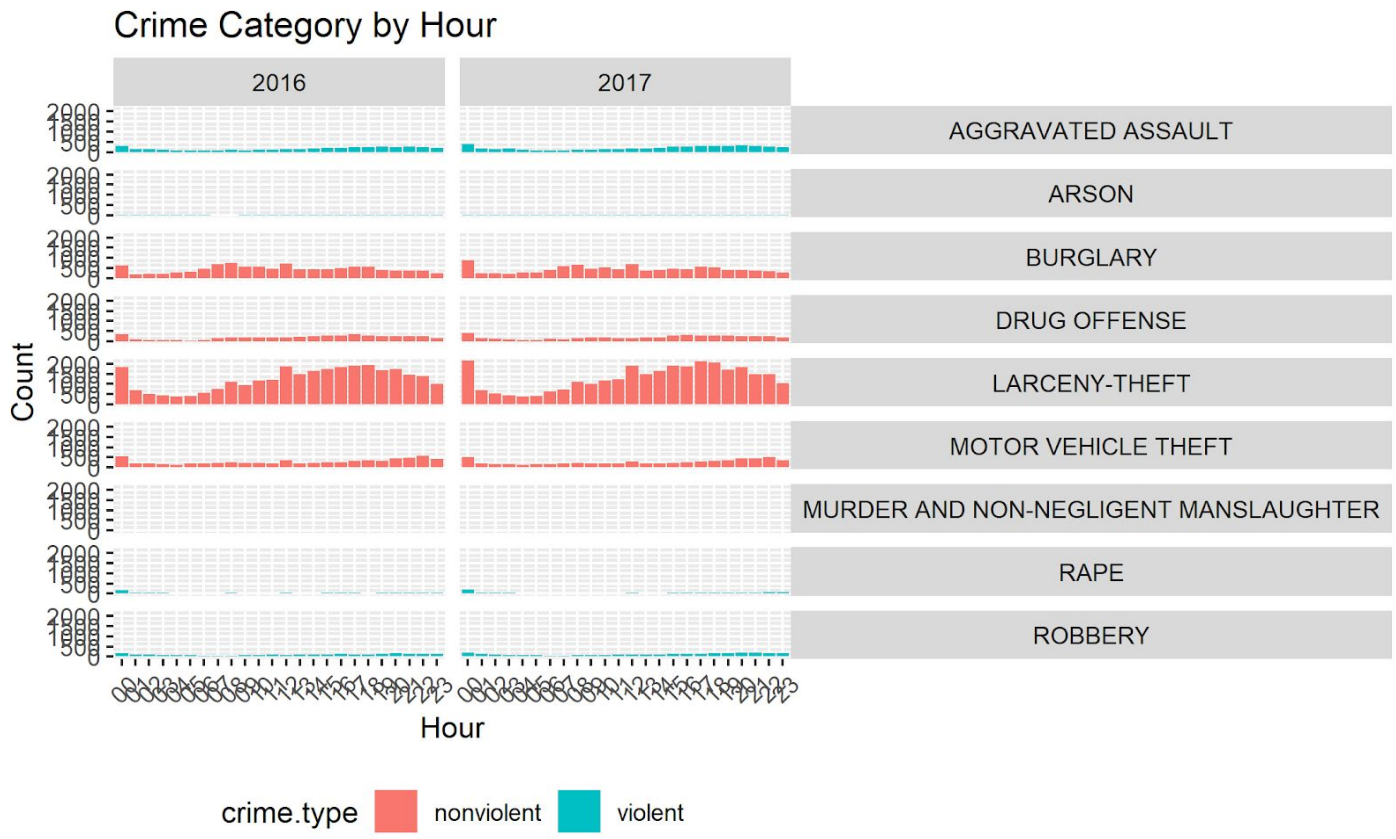
I do not see any seasonal trends in the above charts. Daily temperature data would be more accurate measure of whether there was a crime association with daily temperature or certain weather conditions.

The following box plots are another representation of the crime category distribution across property value in Phoenix. Crimes among all categories peak in areas of property valued at just under 250,000. This is also the most common value for single family housing in Phoenix, which is the location with the highest crime count with almost one quarter of all crimes occurring at a home (as shown by the chart on page 4).

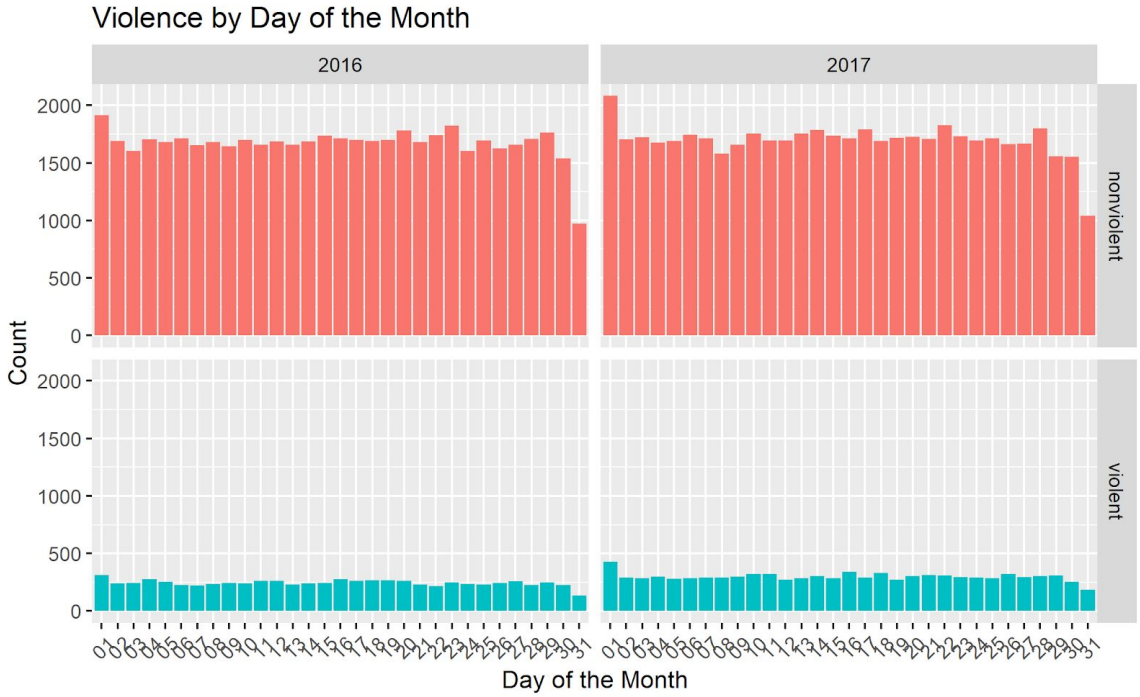


There doesn't seem to be a trend by day of the month, but this may need to be investigated further since the Feature Importance Graph shows the Day as a very important feature. There does seem to be a trend in hourly crime with both violent and non violent crimes peaking at midnight. Non violent crimes have a smaller peak again during the day around 5-6 pm and violent crimes reach a smaller peak around 9pm.

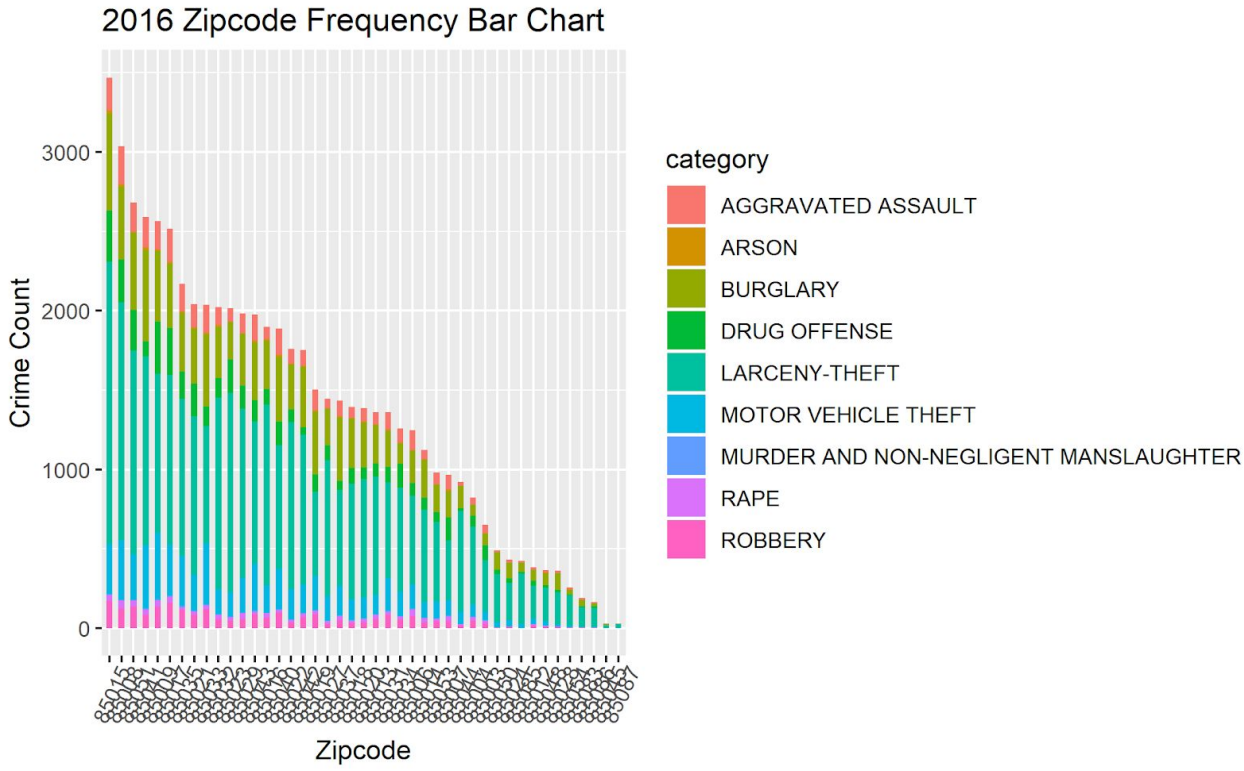




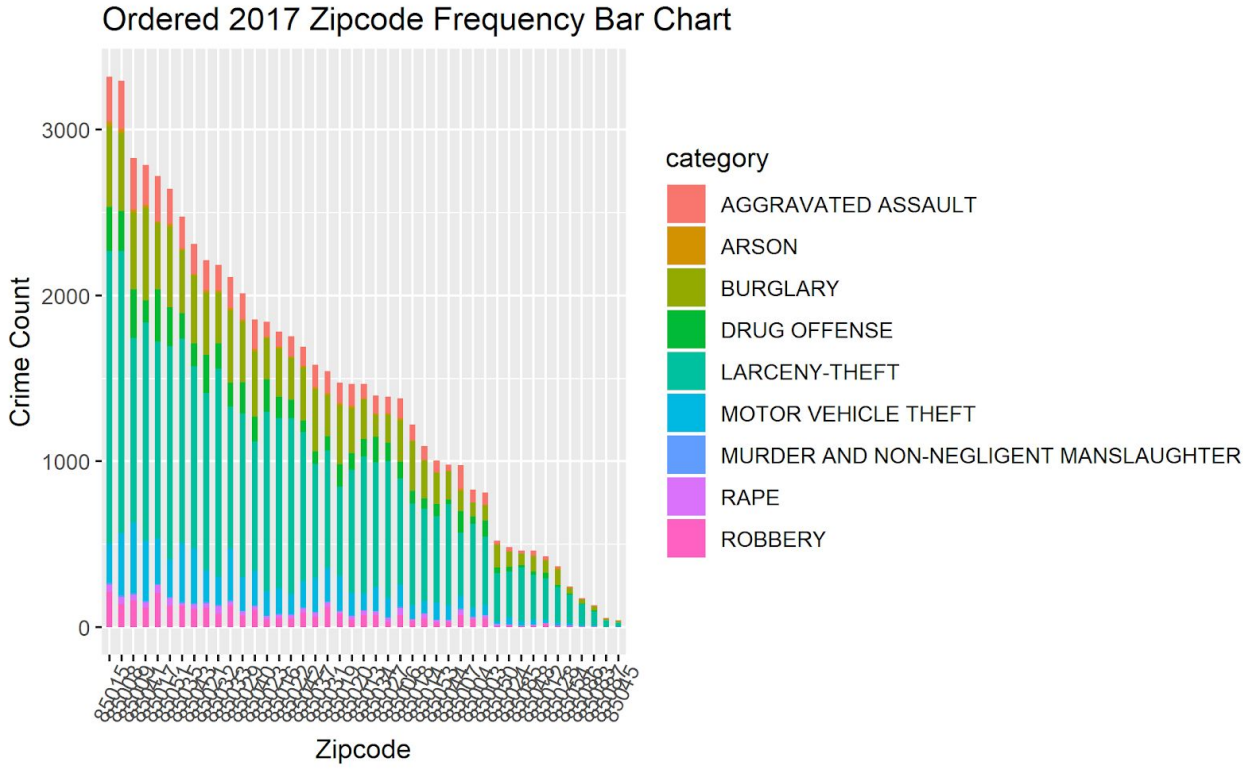
Apart from burglary, which seems to be pretty consistent and dips from 3-5 am, all crime types seem to be more concentrated in the afternoon and evening. There doesn't seem to be any pattern by day of the month, except that the 1st of the month seems to have an unusually high number of crimes.



85015 has the highest total crime count with 8,965 total crimes. Ordered 2016 count-



Ordered 2017 count-



Approach

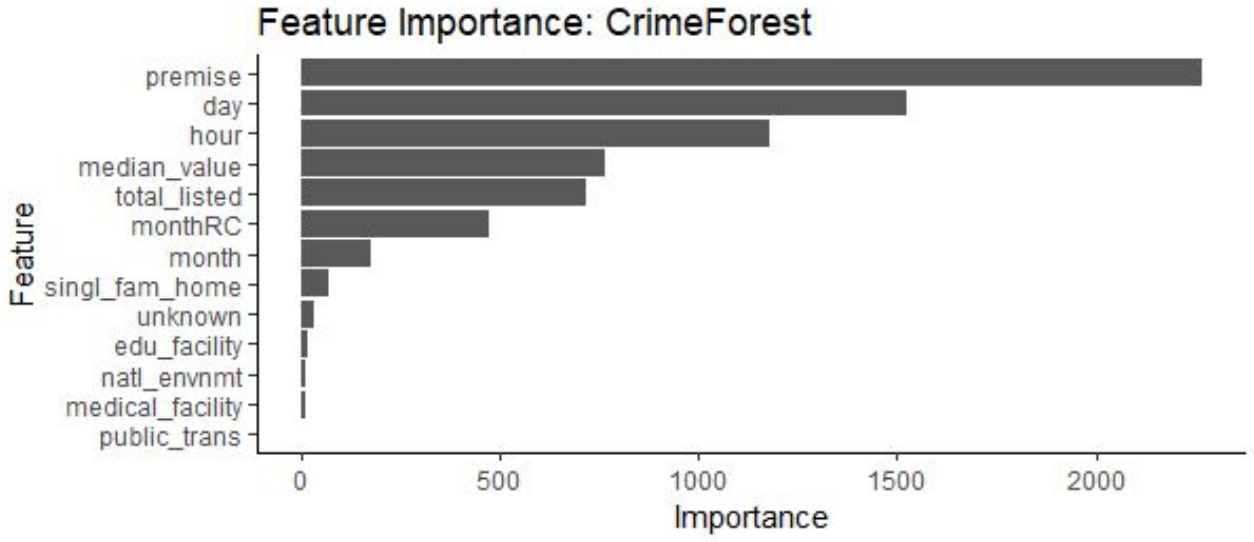
My variables are nonlinear and mainly consist of categorical data. Some algorithms that can be used to fit this problem would be CART, Random Forest, and Support Vector Machines. I set the random seed before testing each model to make sure each algorithm is run using exactly the same data splits. This ensures that my results from each model are comparable.

Upon testing I found the SVM to be inefficient on my machine with my data set. Training the CART model did not take very long, but also did not produce a tree past a root and simply predicted the most common outcome. That leaves the random forest algorithm. Random forest is a predictive modeling tool that can find patterns in data by randomly creating numerous decision trees and determining the importance of each indicator to the outcome. The algorithm produced by this process can then be used to predict whether a crime is likely to be violent or not, depending on the values of the indicators.

Model Evaluation & Fine-tuning

The first random forest model created did not perform any better than the CART model or simply predicting the most common outcome. Calculated from the confusion matrix, the first random forest model had an accuracy equal to (41125+1)/(41125+1+6665+4), or 86%. I needed to experiment with the parameters to get a better model.

One of the advantages of using the Random Forest model is that I can easily access the feature importance. Through looking at the feature importance, I can decide which features I may want to drop, because they don't contribute enough to the prediction process. This is important, because the more features you have, the more likely your model will suffer from overfitting. Looking at the Feature Importance, I am going to remove the premise categories and month variables from the feature selection and retest.



One disadvantage of using the random forest model, is that it has a weakness to class imbalances. Since 86% of my data consists of non-violent crimes, it could be said that violent crimes are rare. Without parameter tuning, the random forest is going to predict the most

common outcome to achieve what looks like a high prediction accuracy. This does not help us predict when we may be looking at what will turn out to be a violent crime, which is the purpose of this model. To handle this class imbalance I can use the ROSE or SMOTE functions to create synthetic sampling using random over sampling examples. After tuning the parameters, we can evaluate the model again to see our improvement.

When tested, I found that I could reduce violent crime type prediction errors the most using SMOTE in the trcontrol parameter in randomForest. I stopped with a 31.7% violent crime class error rate and 36.2% non-violent class error rate. This is a significant improvement in predicting violent crimes as opposed to the first random forest which simply predicted a non-violent crime to achieve a high accuracy score. The initial models had violent crime class error rates of and around 98.7%. Rather than allowing one class error rate to be significantly higher than another I chose the model that kept the error rates as close and as low as possible, while more accurately predicting the violent crime type. I chose a sample size of 15562 of both classes, which is the number of violent crimes in the training data, so as to level the violent and non-violent crime representation in the data.

Predictions

Now I want to get an idea of the accuracy of the model on our validation set. I ran the random forest model directly on the validation set and summarized the results in a confusion matrix. The accuracy for this model is 65%. It was a small validation dataset, but this result is able to better predict violent crimes than my initial model had. I feel that in order to create a more accurate model, more data is needed. In the following sections, I outline some of the data that would give more opportunity to create a better model.

Limitations

These variables provide a very small picture of the setting in which crimes occur and offer a limited view on the overall development of crime in Phoenix. Unfortunately, this dataset does not contain latitude or longitude coordinates, and the street addresses are rounded to the nearest hundred block for confidentiality. I was not able to create a very telling hotspot map with this data (see Appendix). It also does not contain all possible crime types, and only reports crimes that fall under one of the 9 types listed above. The data includes crimes that were reported to Phoenix PD, including those that took place in zip-codes from outside of Phoenix that are often covered by other police departments. Because the crime data from these zip-codes do not include crimes reported to other police departments, the data may unintentionally mislead users to conclude lower crime rates within those zip-codes. The model would also benefit from additional datasets that can provide information to explore other variables such as quality of educational opportunities offered in the area and demographic data including age, ethnicity, and income.

Recommendations to Phoenix Police Department

Phoenix PD could assign more officers to the top 5 crime zip codes following crime analysis reports each year or month, focusing on areas of higher violent crime. According to yearly data, in 2017 this would have been: 85015, 85008, 85009, 85041, 85017. Since burglary is the most common crime type, connecting with home cameras and communicating with residents about strategies to prevent burglary could be beneficial. It could be helpful to officers to share the

most current analysis visualizations and statistics during weekly meetings so they may be more informed about the types of crimes and changing conditions occurring in their areas. Strategically plan officer training to address the characteristics that can anticipate each crime type, while noting that the data shows that any and all crimes can occur anywhere at any time.

Next Steps

More data indicators are necessary to create a more insightful model. The original plan to obtain weather data failed since there were problems with the call through rwunderground's API. This is likely due to the fact that wunderground's API keys are no longer free. After an extensive amount of time was used trying to collect this data in small chunks, resulting in numerous errors in R, the process was deemed inefficient and discontinued.

True crime analysis often incorporates information about suspects and apprehended criminals as well as precise locations. This information can help in further predicting repeat offenses and stopping crime before it happens. A dataset containing information on suspects, victims, coordinates, and area/location characteristics would be enlightening. In this case, this type of detailed information surrounding the crimes in Phoenix is understandably not open to the public, but hopefully exists within the department. Piecing this information together from multiple sources and joining to my current data frame by zip code, would essentially assign each distinct zip code to one indicator value, which is inaccurate. From time to time, I plan to continue to search for open source data to add to better represent this problem and improve the model. This report will be updated when improvements are found.

Sources

Crime Data:[City of Phoenix Open Data]
(https://phoenixopendata.com/dataset/crime-data/resource/0ce3411a-2fc6-4302-a33f-167f68608a20?view_id=644b88ef-16b3-497d-9413-2ba3eedfd3c1)

Zip code lat/lon coordinates:[Free Zip Code Database](http://federalgovernmentzipcodes.us/)

Property Data:[Realtor.com](https://www.realtor.com/research/data)

Phoenix Area Base Map:[Google Maps Static
API](https://developers.google.com/maps/documentation/maps-static/intro)

GGMaps:D. Kahle and H. Wickham. ggmap: Spatial Visualization with ggplot2.
The R Journal, 5(1), 144-161. URL
http://journal.r-project.org/archive/2013-1/kahle-wickham.pdf

Weather Data (weather conditions):[rwunderground API
Rpackage](https://github.com/ALShum/rwunderground)

Appendix

Heat density map