Phoenix Crime Analysis

Capstone Project Final Report

Lyttia McManus

Springboard: Intro to Data Science

Fall 2018

Introduction

Phoenix, AZ is the 5th largest city in the country with a population of over 1,600,000 residents. Despite summer temps reaching the 120s and a poorly ranked public education system, people are increasingly adopting Phoenix as their new home. 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

At this point, 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. Families than can afford it, often move to the suburbs usually to provide a safer environment for their children.

As my first attempt at analyzing data in R and data science in general, 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 this city. Examining the conditions in which these crimes occur can lead us to understand its contributing factors, and what really 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 safer and keep 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. This type of prediction could 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. It could also 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 (like 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:

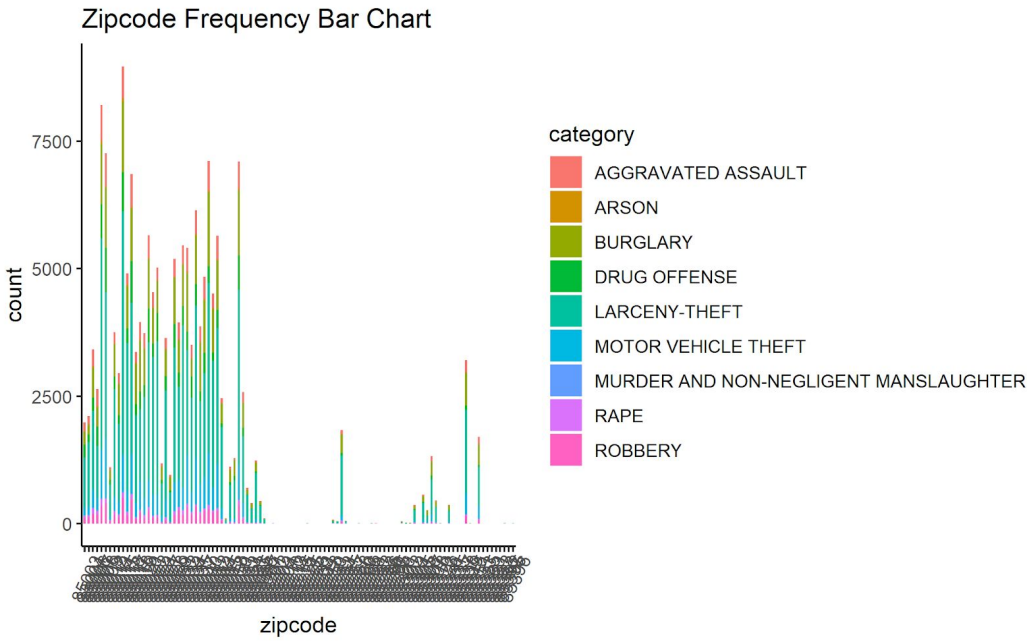
- Crime incident number
- Earliest and latest date and time the crime occurred*
- Uniform Crime Reporting (UCR) crime type:
    - Homicides, rapes, robberies, aggravated assaults, burglaries, thefts, motor vehicle thefts, arsons, and drug offenses*
- Hundred block address (precise address confidential)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

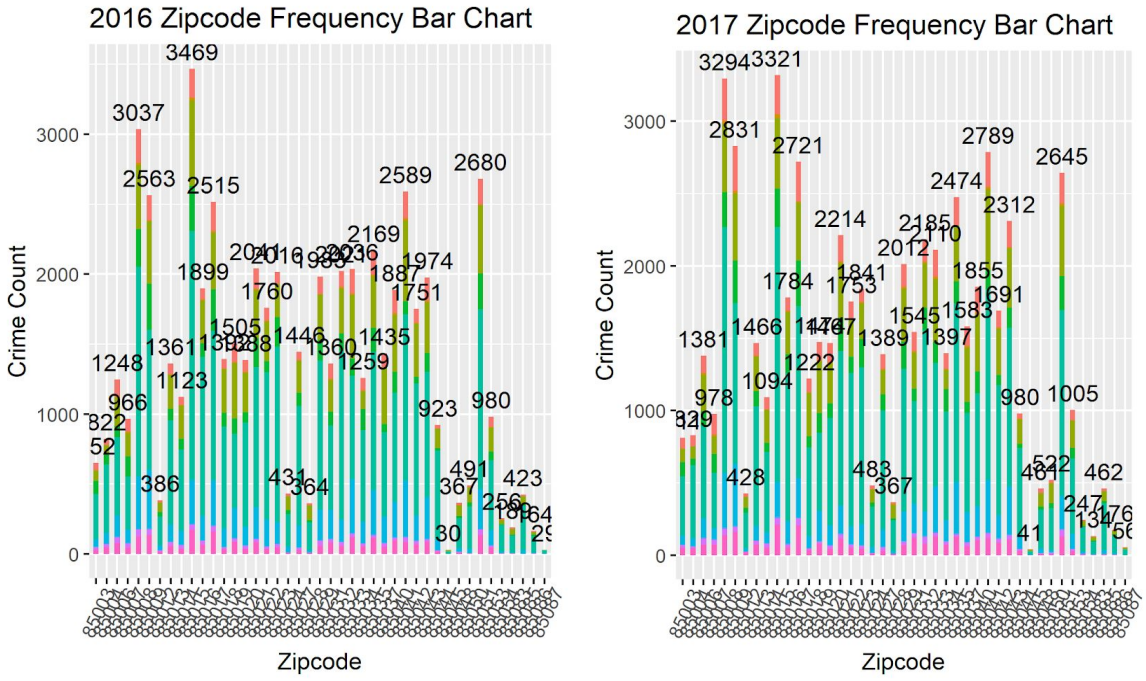
- Median & Average list price*
- Luxury list price
- Median & Average days on market
- Total active listings*
- New listings
- Price increases and price reductions

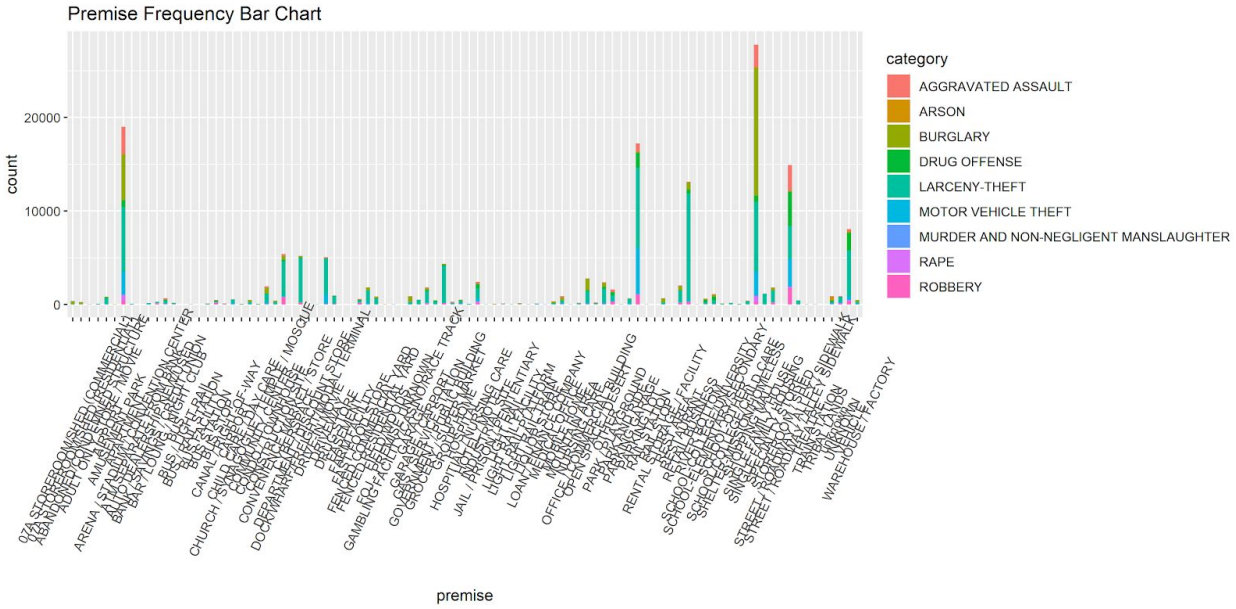
Cleaning and Wrangling

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.

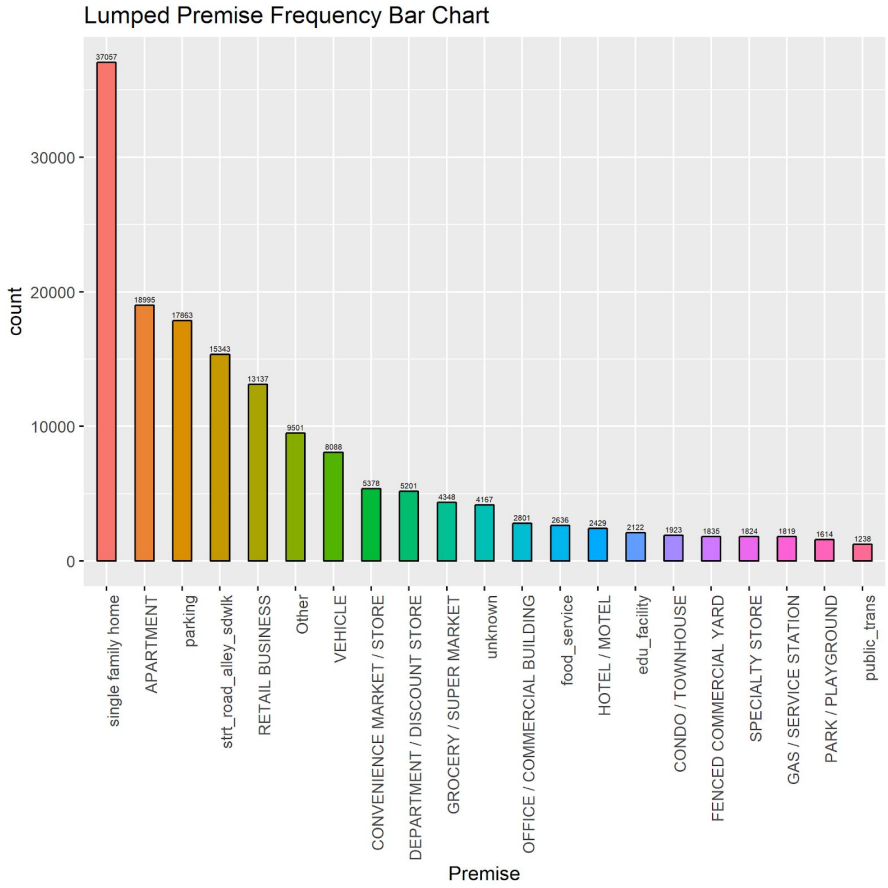


Zip-codes were filtered from the original data to more accurately represent a true crime count, since many of these zip-codes are located outside of official Phoenix boundaries, and likely have many additional crimes reported to their respective police departments. Phoenix zip-codes start with 850. This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





There are originally 95 unique premise types, so smaller premise categories are added with dummy binary variables.

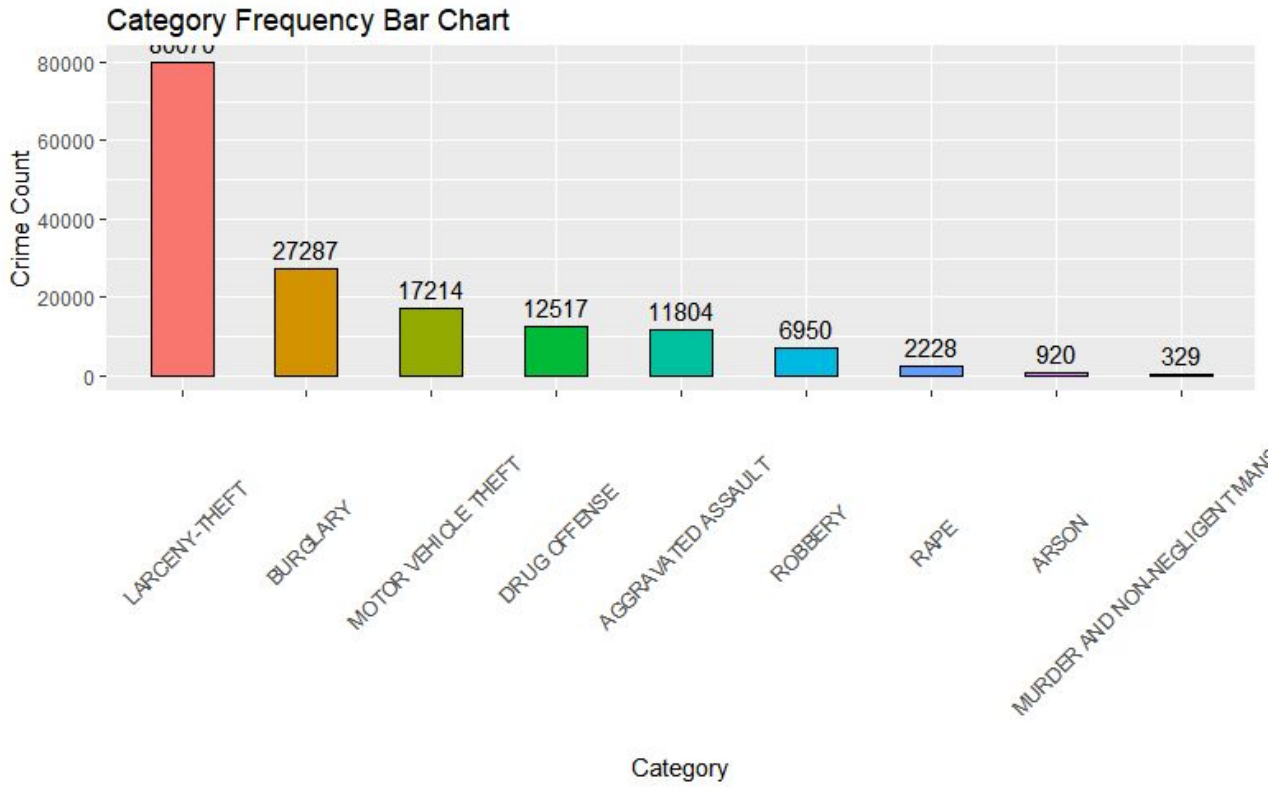


The graph above shows the top 20 premise types reported to PPD. The remaining premise types were combined into "Other", which accounts for 5% of the total data. The most common premise type is the single family home, which accounts for 23% of the data.

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" to produce the "crimes3" data frame. 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.

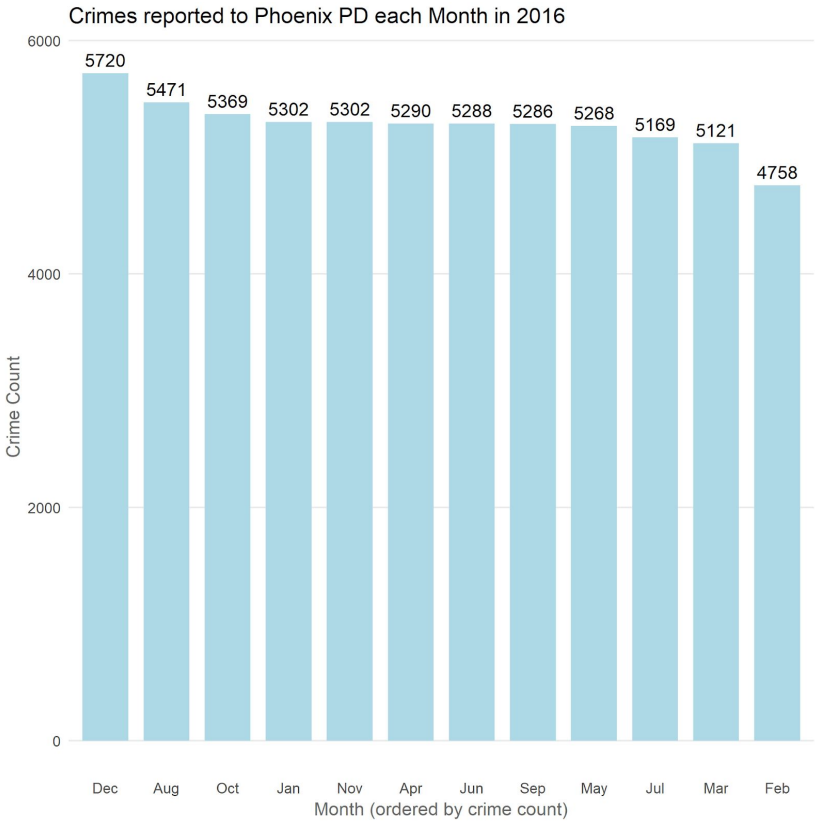
Preliminary Exploration and Initial Findings

Of the nine categories of crime, larceny-theft was the most common crime category reported to PPD with __ out of __ total crimes, coming in at _% of the crimes reported.

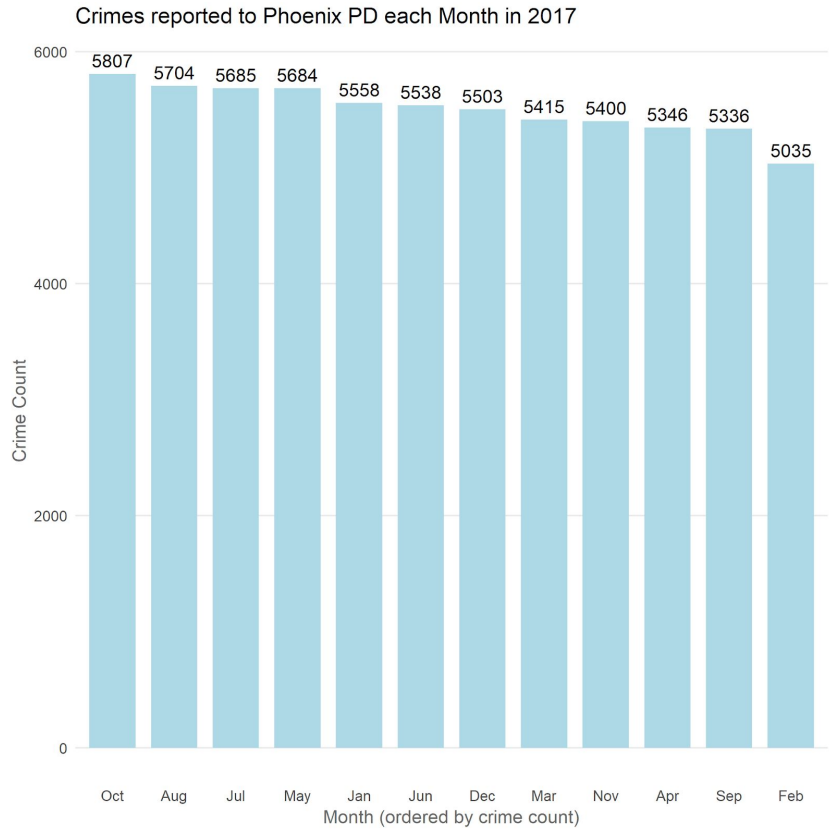


[1] It is not possible to get up-to-date property value estimates on the addresses from the Phoenix Open Crime Data because addresses are displayed to the hundred block for confidentiality.

Bottom limits of about 5000 crimes to upper limits of 6000 crimes reported monthly:

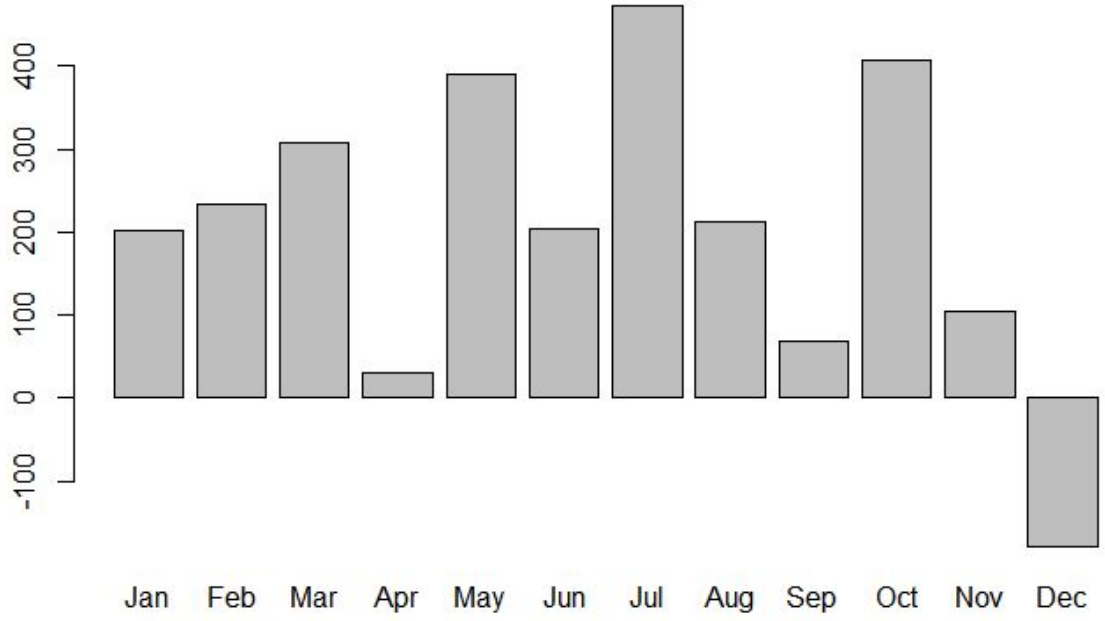


Range in 2016- 962



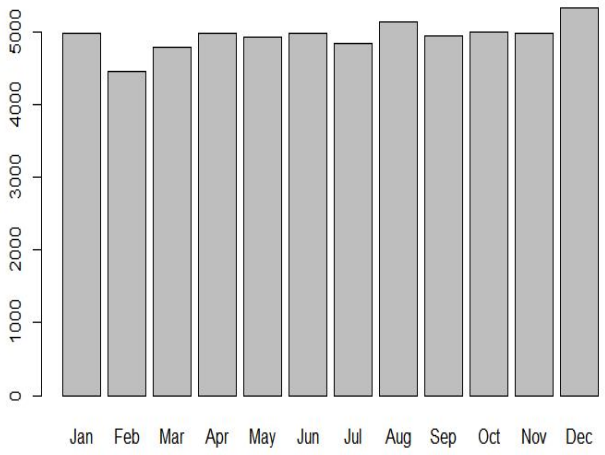
Range in 2017- 772

Monthly crime change from 2016-2017

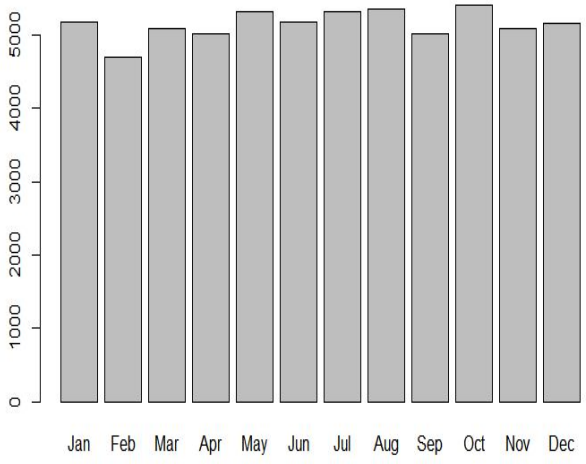


There were 2,450 more crimes in 2017 than in 2016, a 4.1% increase. Crimes increased each month in 2017 except for December in which crimes decreased. December went from being the highest crime month in 2016 to #7 in 2017.

2016

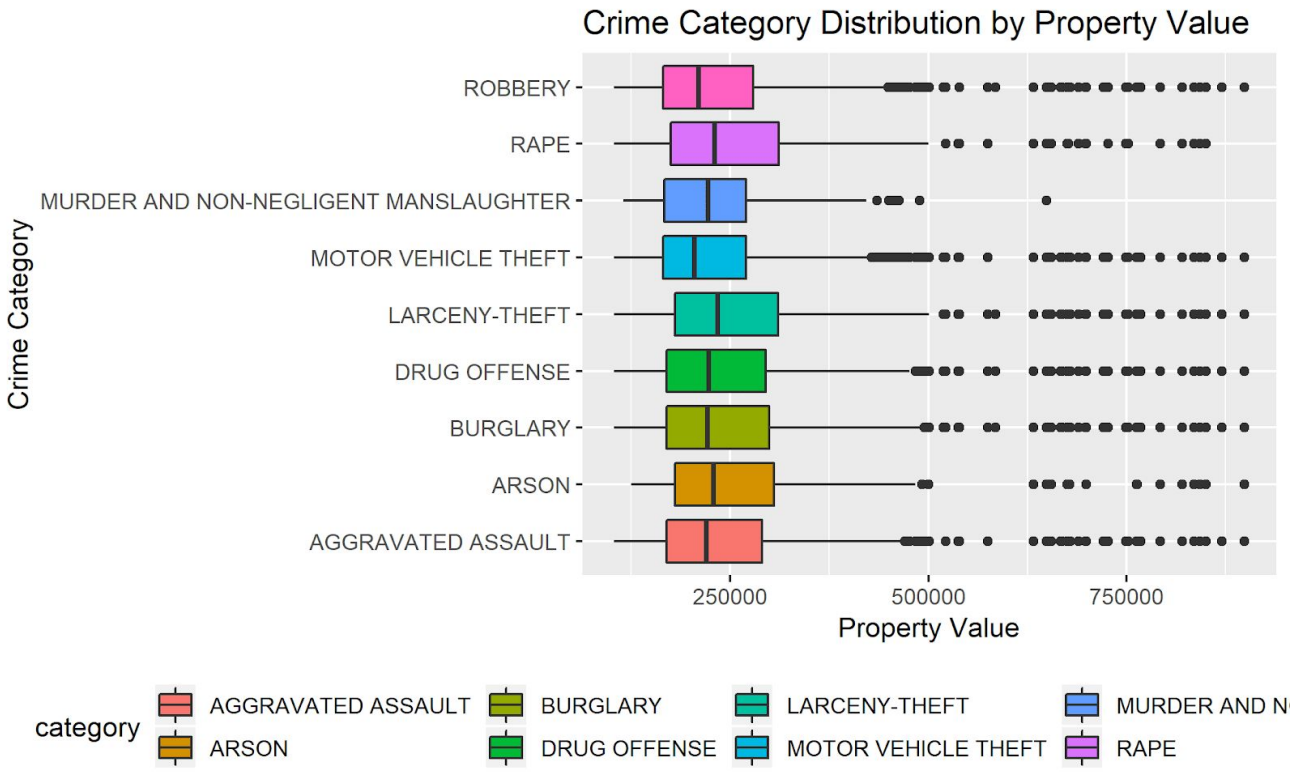


2017

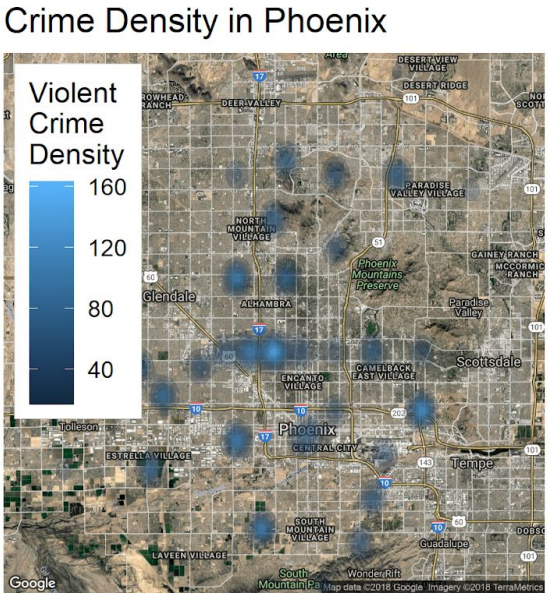


I do not see any seasonal trends. Daily temperature data would be more accurate measure of whether there was a crime association with daily temperature or certain weather conditions.

Graphing trends in property value:



Heat density map:



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

Ordered 2016 count-

Ordered 2017 count-

Crimes by Time of Day:

Approach

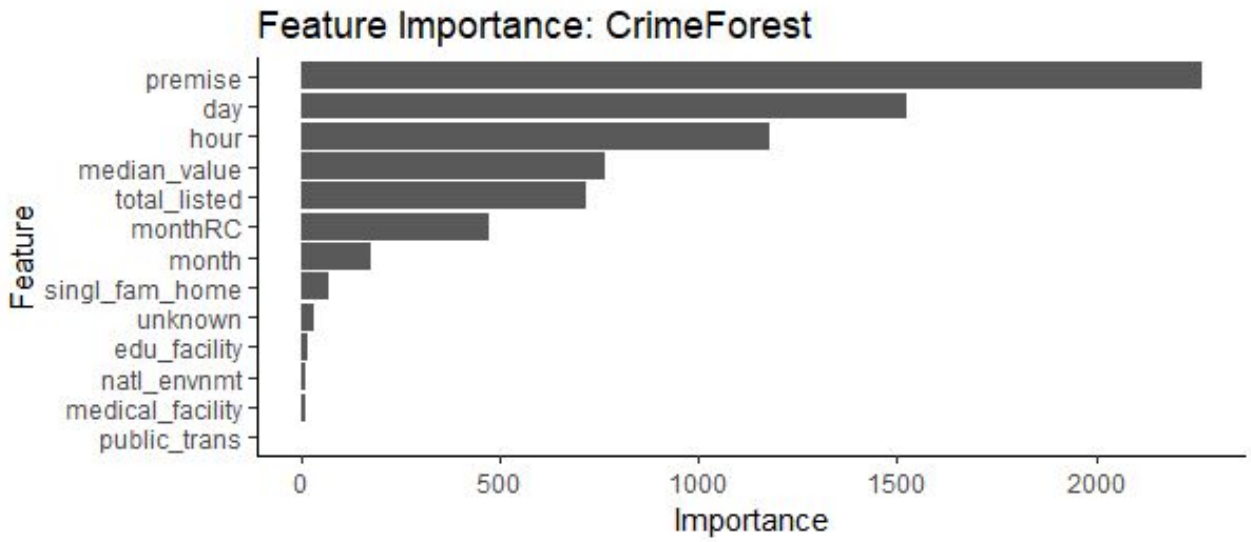
Since I wanted to fit a model that can find a pattern in categorical variables that may or may not be linear and may have a degree of multicollinearity, a decision tree algorithm works best. The Random Forest algorithm is a predictive modeling tool that can find patterns in data by randomly creating 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.

The intent is to use a Random Forest model to create a model that can predict whether a crime will be violent or nonviolent based on various indicators. The CART model did not produce a comprehensive tree. The data will be split into training and test data randomly and evaluated using cross- validation.

Some predictors I will be engineering and testing for use in my Random Forest model include:

- Time of Day (Hour), Part of Day (Morning, Afternoon, Evening, Night)*
- Season (Winter, Spring, Summer, Fall), Month*
- Premise Type
- Zip-code, Median Property Value, Value categories(Lower, Middle, Upper)*
- Part of Town (Northwest, Southwest, Northeast, Southeast, Central, North_central, South_central)
- Extras: Weather data (temperature (F) & conditions)*- could be correlated with season

*variables listed together have a high chance of multicollinearity



Model Evaluation

I used a confusion matrix to evaluate the random forest model:

PredictCV

	0	1
0	41126	0
1	6669	0

Proposed Use

This type of prediction could be used along with crime hotspot maps and other visualizations, to provide:

- decision support to police departments
- inform policy makers
- It could also 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 (like Zillow).

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 may not be able to create a very telling hotspot map with this data. 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. My 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.

Next Steps

As this was my first data project, I had a lot to learn from the beginning and did not make the best choices. At this point, I realize that my project needs more data. Unfortunately, my plans for obtaining weather data failed since there were problems with the call through rwunderground's API. This is possibly due to the fact that it is no longer free.
It is very clear that I need more information to analyze before I can make more pointed predictions. True crime analysis usually 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. While detailed data collection should be in place, it is understandably not open to the public.

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)