GW Data Analysis Bootcamp
Project 1

Siobhan Byrne
Bryan Johns
Katie Loosararian
Robert Takan

Variants of Time As a Factor of Victim-Based Crime in Baltimore

This exploratory data analysis looks at the relationship between the frequency of crime and time of day, to better inform decision makers on the efficient allocation of police resources. We find a strong statistical association between the time of day, day of the week, and month of the year upon the frequency of crimes.

Data Selection

We found the data set on Kaggle: https://www.kaggle.com/datasets/sohier/crime-in-baltimore

It's based upon the Open Baltimore data, released to the public by the Baltimore Police Department (BPD): https://www.baltimorepolice.org/crime-stats/open-data

The dataset represents the location and characteristics of major (Part 1) crime against persons, such as homicide, shooting, robbery, aggravated assault, etc., within the City of Baltimore. These Victim-Based crimes are collectively referred to as Part 1 crimes by the BPD.

The data set runs from the first of January, 2012 to the second of September, 2017.

We chose it, quite frankly, because it seemed interesting, and most of us have personal ties to Baltimore. We're also big fans of The Wire, so it seemed fitting.

Data Cleaning

We removed "CrimeCode" from the data. Looking at the Crime Codes, we discovered that the only data of interest was already included under other pandas series, especially crime "Description". We removed the police "Post" from the data, as we were not interested in it.

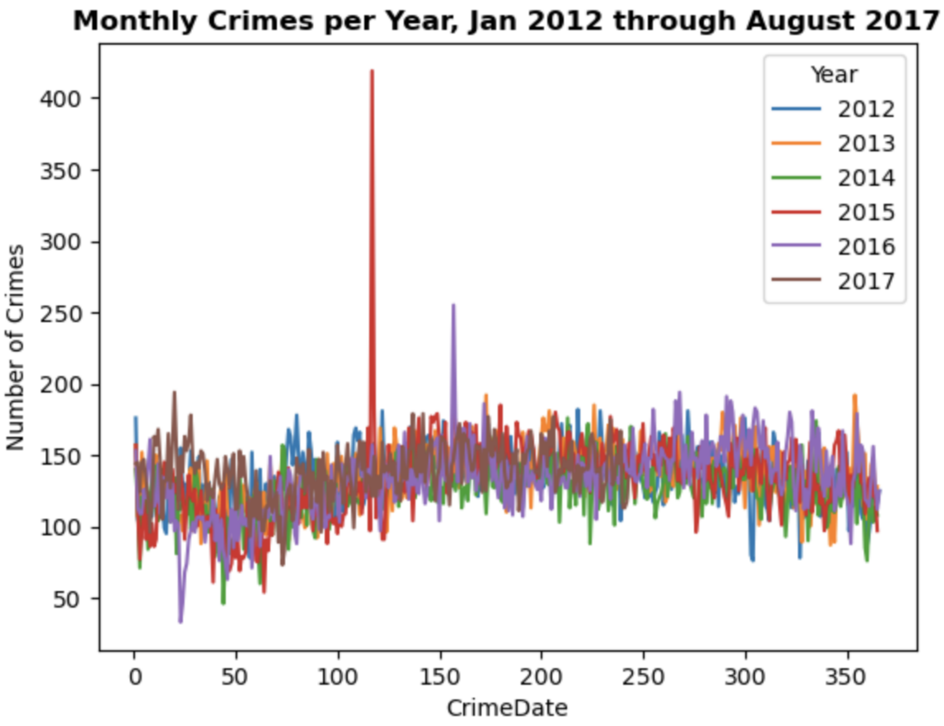
We combined the values “I” and “O” with “Inside” and “Outside” in the “Inside/Outside” series.

Time required some cleaning. We switched “CrimeDate” to a datetime64 data type. We had to switch 24:00:00 to 00:00:00 to convert it from a string to datetime64.. We added a series for the “Day of the Week”. We also added a series for “Time of Day”, broken down into “Late Night”, “Morning”, “Afternoon”, and “Evening”, based upon bins of midnight - 6am - noon - 6pm - midnight. We later conduct more granular analysis on an hour by hour level as well.

We did not drop any null values. We didn’t drop them to allow for flexibility of data analysis. For example, if you are checking the type of crime happening at a certain hour of the day, does it matter if the “Inside/Outside” category has a few null values? (It has 10,278 null values, roughly 3% of the total data)

The biggest example of this in action comes when analyzing data by year. Since the data set only runs through 2-Sep-2017, 2017 often has to be dropped while performing data analysis, to avoid skewing the results. This frequently happens in our data analysis. You will notice that at times the visuals will be marked through 2016, or 2017, as appropriate.

Our biggest outlier was the aftermath of the death of Freddie Gray in April of 2015.



Line chart clearly showing outlier data from the aftermath of Freddie Gray.

Data Exploration

We performed a fair amount of data exploration, starting with “initial_data_cleaning.ipynb”. We broke apart the data, looking at the counts for values in each column, to see what could be cleaned up and to get a sense of what was in the data. We even took a look inside the Open Baltimore data from 2012-2021. These early data explorations happened in tandem with the early data cleaning.

After getting a general sense of what was in the data, we created a wide array of plots and ran some rudimentary statistical analyses, generally looking around to see what we could find. All of this early exploratory data analysis can be found under “Resources/miscellaneous” in the repo root.

We started noticing some trends in how time impacts crime levels, and began to focus on that.

Crime and the Effect of Time of Day, Day of Week, Month of Year

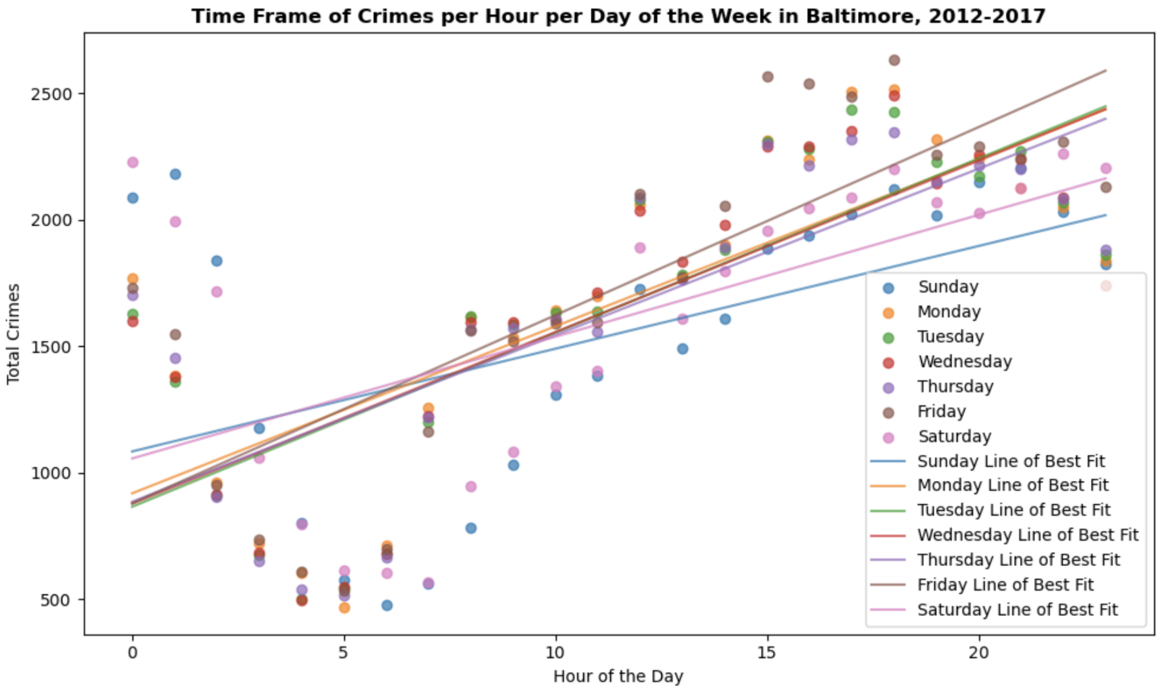
Overview

Our hypothesis was that time, day of week, and month would impact the frequency of crime in the city of Baltimore. The null hypothesis being that time, day of week, and month does not impact crime in Baltimore.Our analysis showed that time of day, day of the week, and month of the year all have significant impact on victim-based crime rates. Across all years of data, afternoon and evening crime is the most prevalent. 6:00pm was the hour of the day with the most reported crime incidents, with the most prevalent type of crime at that hour being larceny. Friday was the day of the week that saw the most amount of crime on average.

We used various methods of statistical analysis on our data (ANOVA, linear regression, and chi-square). All supported our hypothesis that time of day, day of the week, and month of the year have a significant impact on victim-based crime rates.

Does the Hour of Day Impact Crime Rates?

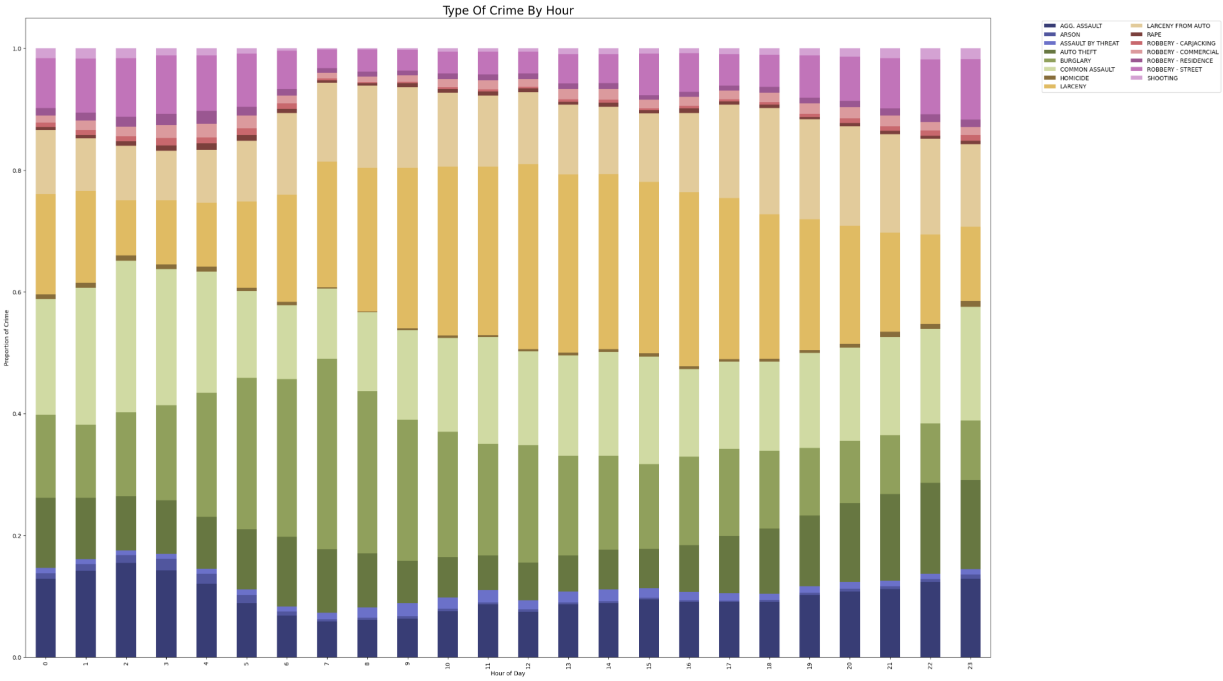
Our analysis showed that 6:00pm was the hour of the day with the most reported crime incidents, and that the bulk of crime happens in the afternoon and evening times of day.



Scatterplot of Crime by Hour and Day of Week

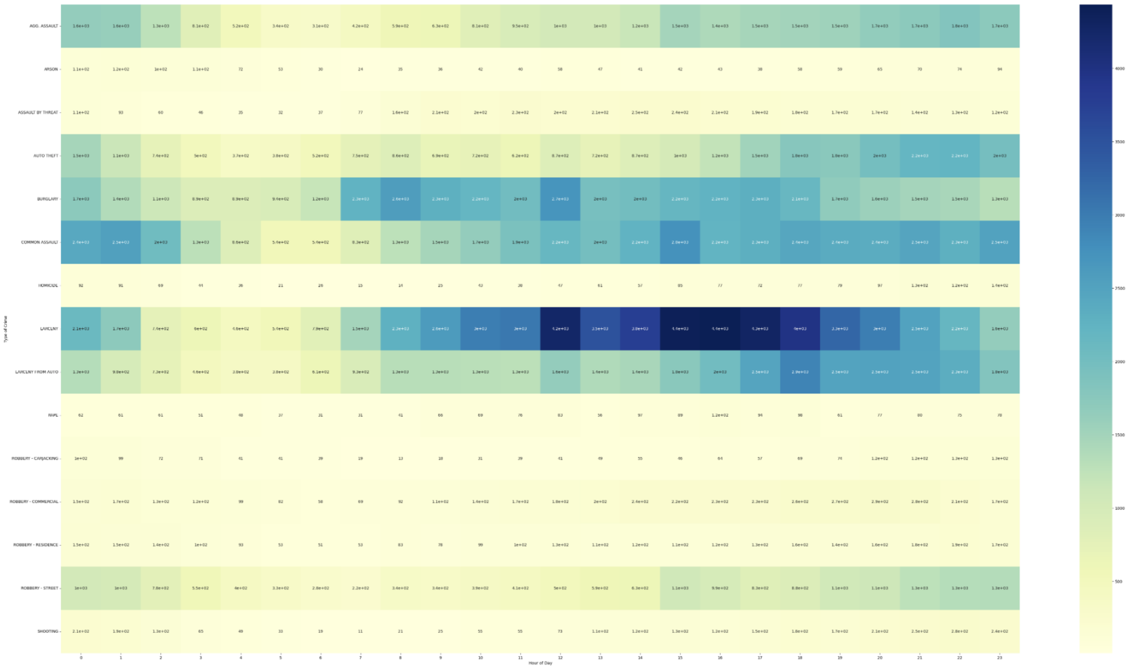
Linear regression was performed on crime by hour and day of the week. The pvalue is well below .05, signifying a statistical correlation between time of day and the number of victim-based crimes. The rvalue is also high, at least .5, and often close to .8, meaning the scatter points fit the line relatively well.

The linear regression line shows how the value of a scatter point on the y-axis can be predicted by the value on the x-axis. The r-value measures how closely the scatter points fit the line, on a scale of -1 to 1, with zero meaning no relationship between the scatter points and the slope of the line, and an absolute value of 1 meaning an exact fit between the line and the points. So, an r-value close to 1 or -1 means the line detects a significant relationship between the x & y axis. An r-value close to zero signifies the line has no bearing on the relationship between the x & y axis.

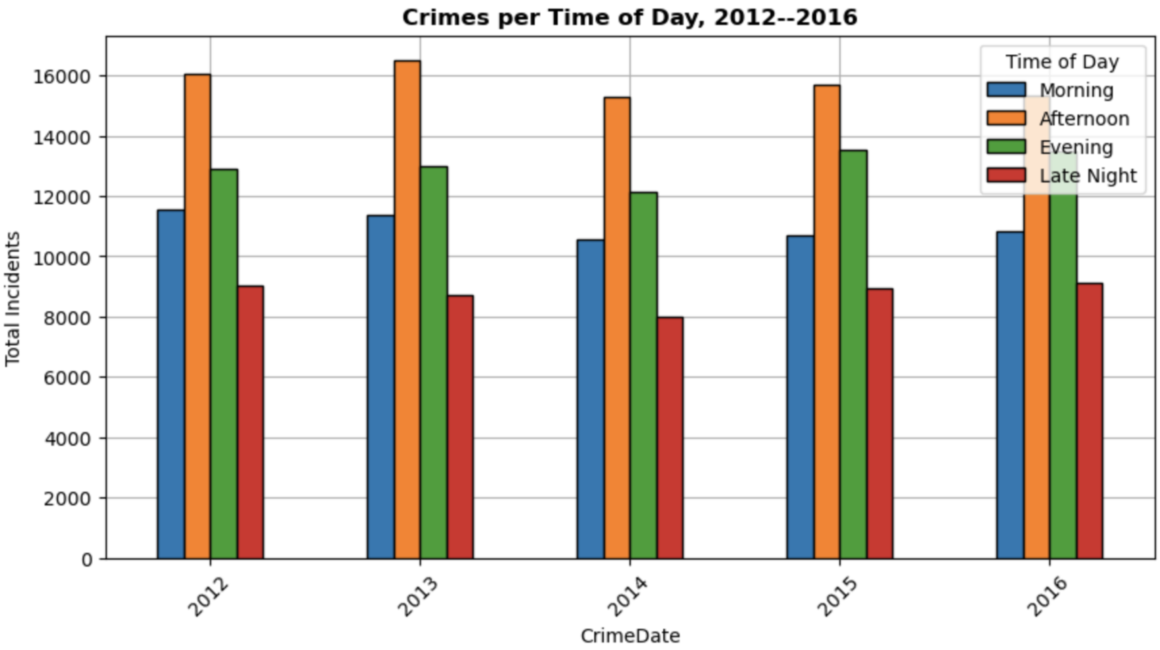


100% stacked chart of Type of Crime by Hour

We also performed chi-square on this data. The chi-square statistic is large, and the p-value zero, indicating there is a strong statistical association between the frequency of crimes and the two categorical variables of Type of Crime and Hour of the Day.



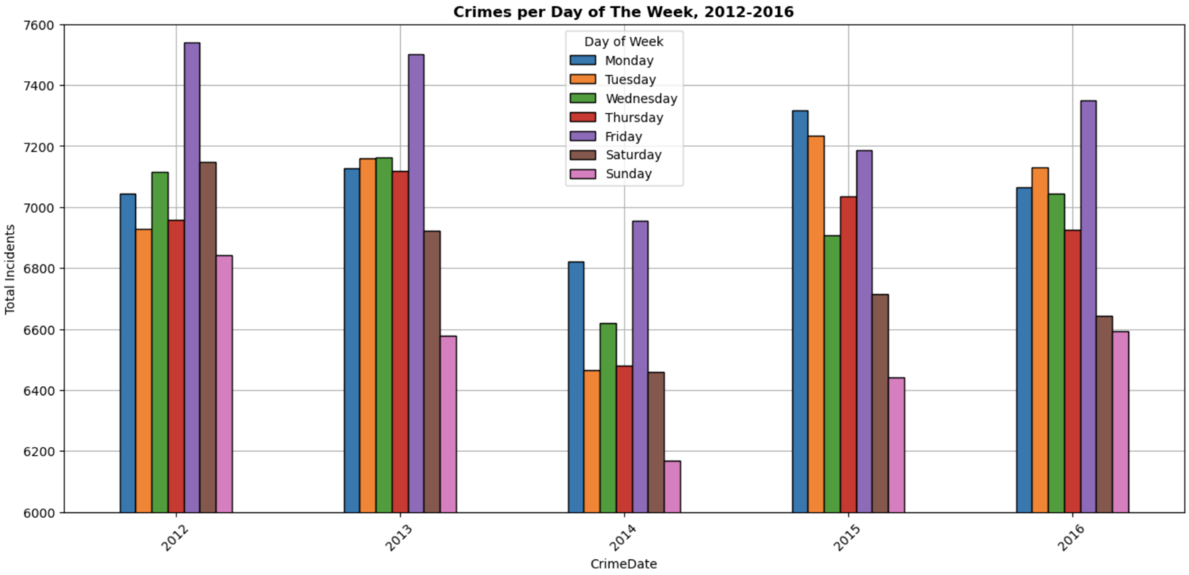
Heatmap of Type of Crime by Hour



Bar chart of Crime per Time of Day by Year

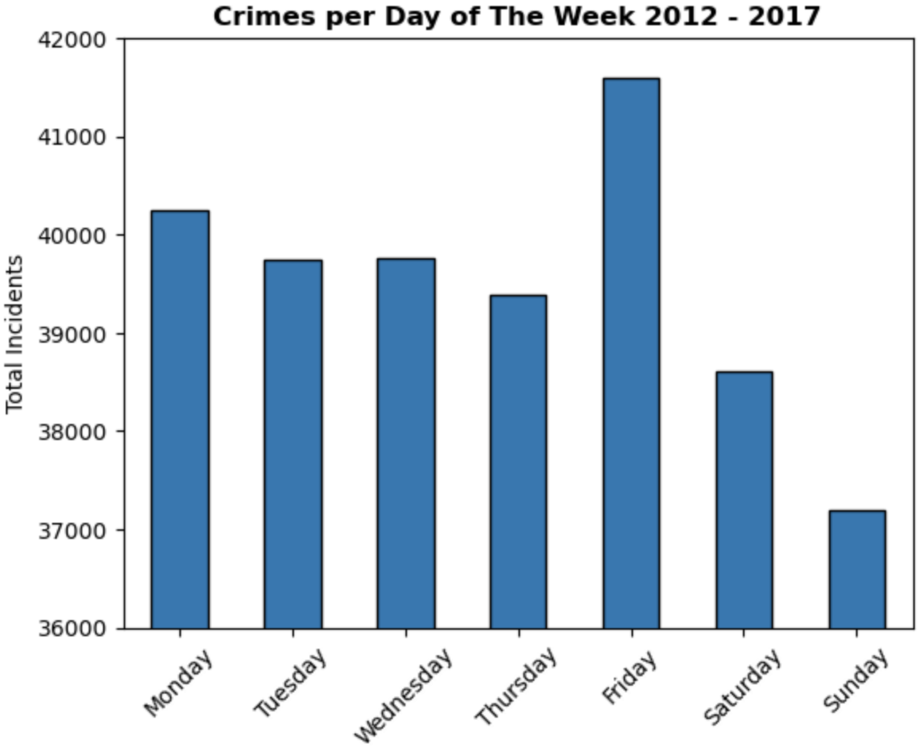
Does the Day of the Week Impact Crime Rates?

According to our analysis of the data, Friday was the day of the week with the highest amount of crime.



Bar chart of Crimes per Day of the Week by Year

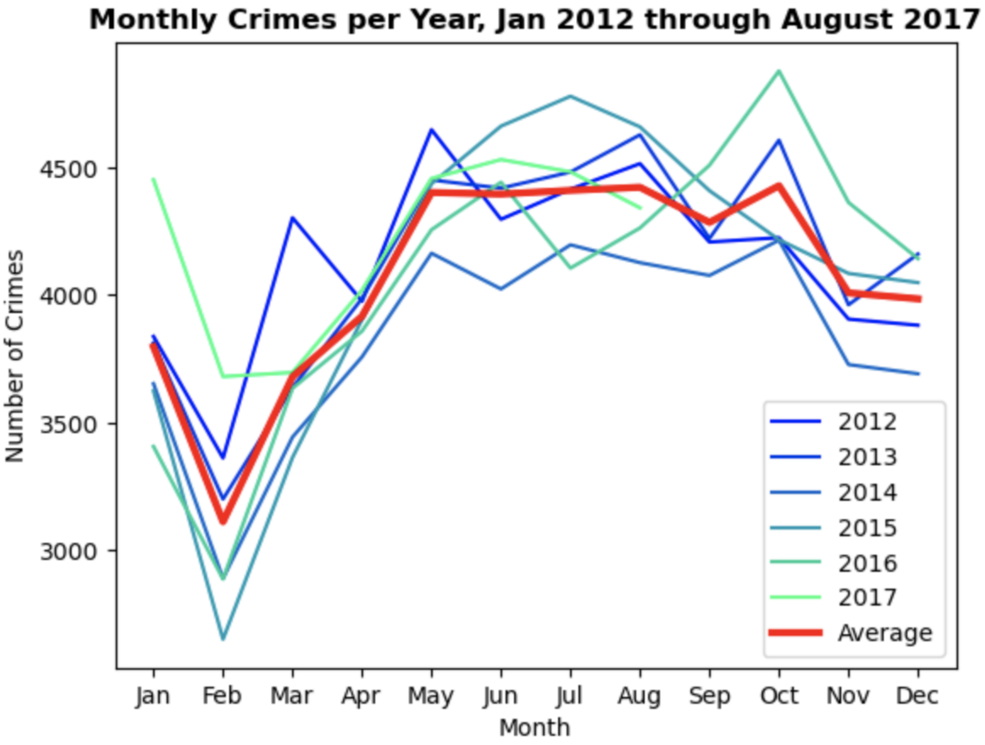
We performed chi-square analysis on this data. The chi-square statistics is quite large, and the p-value effectively zero, indicating there is a strong statistical association between the frequency of crimes and the two categorical variables of time of day and day of the week.



Bar chart of Crime per Day of Week

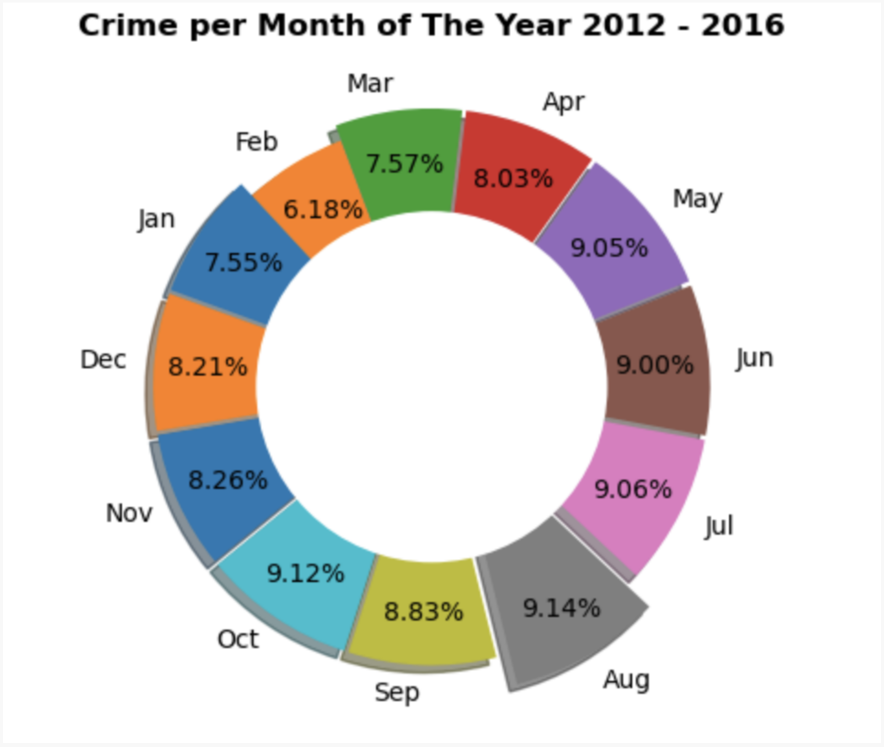
Does the Month of the Year Impact Crime Rates?

The data shows that summer months have more crime on average than fall or winter.



Line chart of Monthly Crimes per Year

We performed statistical analysis on this data using ANOVA. The resulting p-value of 0.001 supports our hypothesis that month of the year does impact the rate of crime. The p-value is very small, so we reject the null hypothesis.



Pie chart of Crime by Month

Conclusions

This data can potentially be useful from an allocation of resources perspective–deciding what days of the week, times of day, and months of the year might benefit from increased coverage by law enforcement. The data and analysis strongly supports that afternoon and evening crime is the most prevalent, Friday is the day of the week that sees the most crime, and the summer months see more crime than fall or winter. All of our hypotheses (hour of day affects crime rate, time of year affects crime rate, day of week affects crime rate) were supported by the various statistical analyses performed.

Further analysis could include a deeper dive into various types of crime day of week and month, and inside versus outside crime, as well as types of crime that are more prevalent in different neighborhoods and districts.

References

Assistance on the stacked bar charts came from (https://python-charts.com/part-whole/stacked-bar-chart-matplotlib/)