Denise Dodd

Trends and Predictions in St. Louis Crime

White Paper

Business Problem

St. Louis is frequently rated among the most dangerous cities in America (Schneider, 2023). This study aims to analyze various aspects of crime in St. Louis, including the types of crime occurring across its districts, the timing of crime incidents, and predicting future crime rates. With this information, the St. Louis police department can better allocate their resources, enhance community education, and engage in preventative measures to curve future crime.

Background/History

The St. Louis police department provides monthly datasets titled SLMPD Crime Files (St. Louis Metropolitan Police Department, 2021). The most recent files span the year 2020. I have concatenated each of the monthly 2020 files into one dataframe spanning the entire year of 2020. The primary data points that I will be utilizing from this dataframe are: Month Reported, Date Occurred, District, and Crime. I will be using splitting and mapping to transform these columns into additional variables.

Data Explanation (Data Prep/Data Dictionary/etc)

As a standard practice, my first steps of data cleaning typically involve reviewing for null values and checking for duplicates. However, in "Crime Data Frequently Asked Questions" (St. Louis Metropolitan Police Department, 2008), the Police Department mentions that each of these anomalies should be anticipated in the dataset. There are many times where every data point is not available for a Complaint which will result in null values. Additionally, if there is more than one incident of the same crime occurring under the same complaint, it will appear to be a duplicate, but it will not be.

The "Crime" column of the dataset contains a six-digit Uniform Crime Reporting (UCR) number. However, leading zeros have been dropped. I added leading zeros to entries in this column less than 5 characters and then confirmed that all resulting entries in this column are 6 characters. After this, I extracted the first two characters from the "Crime" column and placed them in a "UCR" column. I used the FBI's UCR Handbook (Federal Bureau of Investigation, 2004) to map the first two characters of a UCR number to an overall category that a crime falls into (i.e. larceny, assault, fraud, etc.) This resulted in a new column titled "Category". This will be important as different crimes require different expertise and prevention techniques.

Methods

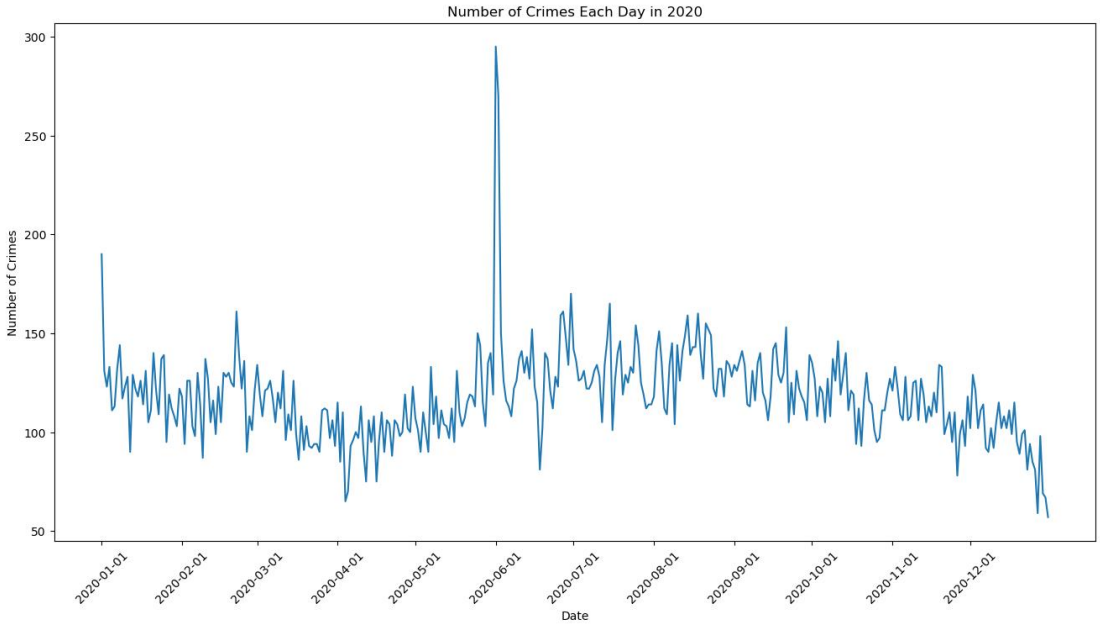
Many of the methods pertaining to the trends in when and where crime occurred were reviewed by grouping data either by a time element (month, hour etc), category of crime, or the district where the crime occurred and counting the number of criminal incidents for each group. These groupings were then displayed using various visualization techniques such as line graphs, horizontal bar charts, stacked bar charts and heat maps. When the visualizations revealed anomalies in the data, I researched by either filtering the dataframes or using a violin graph to review distribution. When reviewing the reporting time, I extracted the month and year elements from "MonthReported" and "DateOccured." I calculated the difference and placed the resulting figure in a "TimeLag" column. I then found the min, max, and average of this column overall and by category.

Finally, to predict future crime rates, I split the data into training data (everything in the dataset that happened prior to 2020-09-01) and testing data (everything from 2020-09-01 to the end of 2020). The "Date" column was set as the index for both data sets and notated as a daily frequency. I then

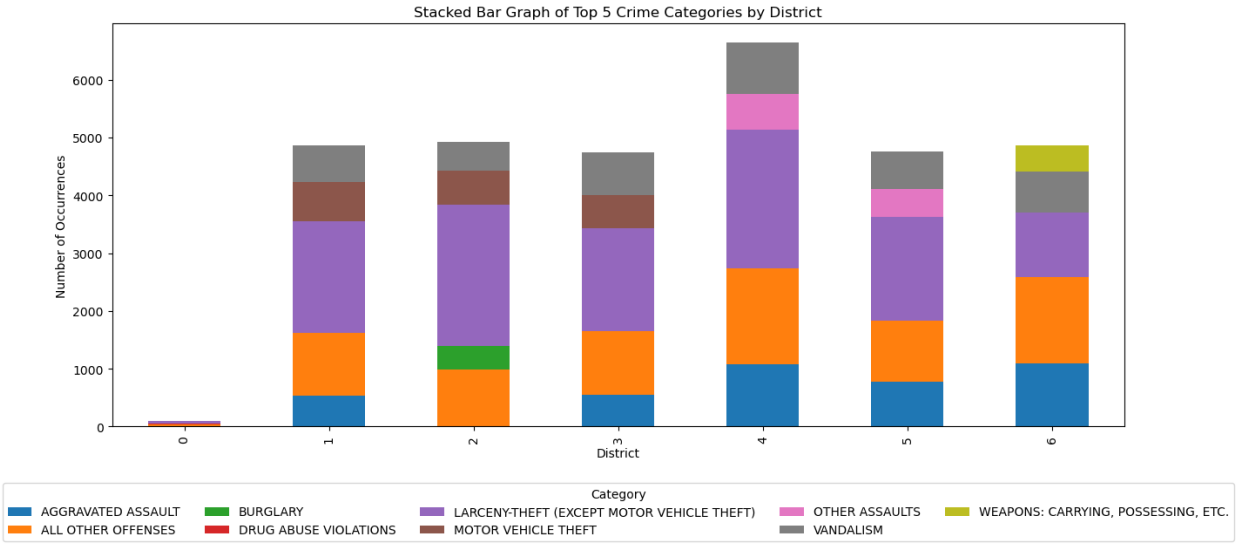
used a loop to train, fit, generate predictions, and calculate RMSE for Holt-Winters, SARIMA, and ARIMA time-series prediction models. Each model had parameters specifying daily lag in the data and making daily predictions. The RMSE score determined that on average, the predictions generated by the Holt-Winters model varied the least from the data in the test set.

Analysis

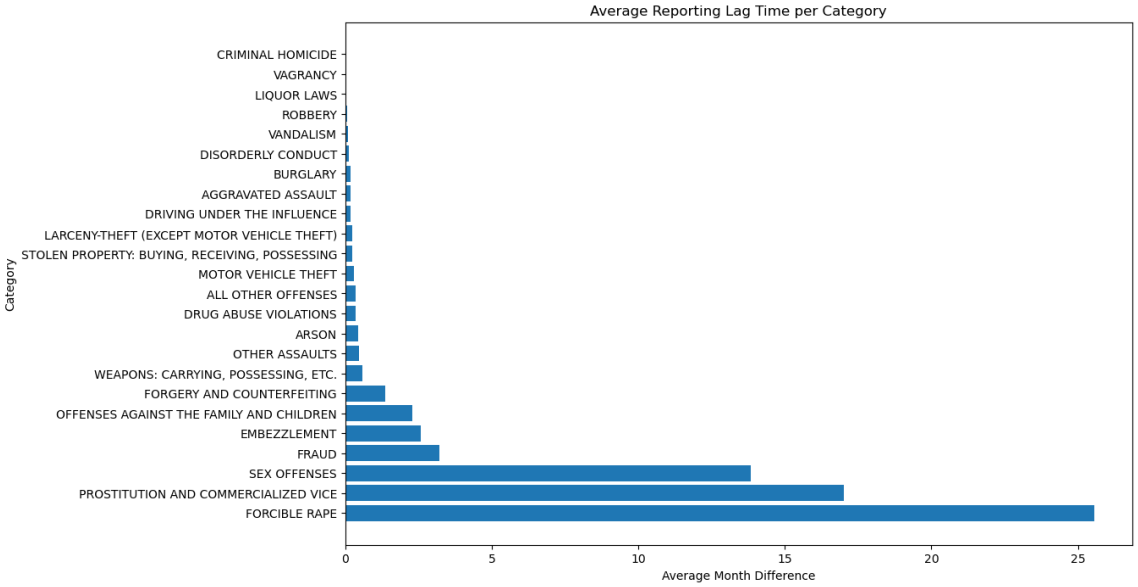
The line graph, which detailed the daily volume of crime, shows a slight increase in crimes during the summer months but mostly there is a steady ebb and flow of crime. There is also a noticeable spike in crime near the beginning of June. By filtering, I narrowed down the cause of this spike to 3 complaints with 24-43 incidents per complaint. All incidents were Aggravated Assault charges. These multi-crime complaints are causing large jumps in the crime volume, but they are valid data points and will remain in the dataset.



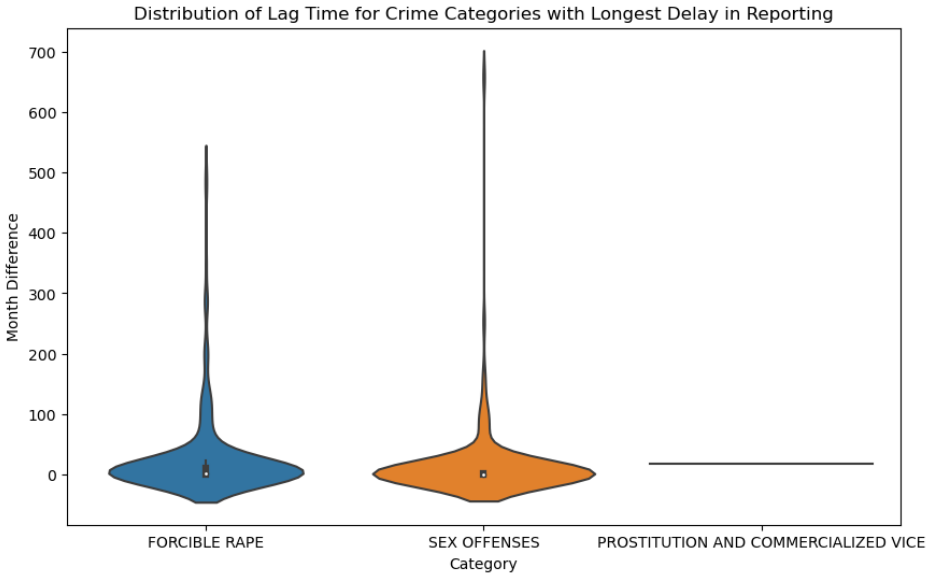
After reviewing when crime occurs, I studied where crime occurs. The below stacked bar graph highlights the top 5 categories of crime in each district. It is apparent that District 0 has nominal crime, raising a question about its population. The viewer can also see the tall purple sections in each district which represent the volume of Larceny. A final interesting observation is that District 6 is the only district where Weapons charges are in the top 5 crime categories and District 2 is the only district with Burglary charges in the top 5 crime categories.



I reviewed how reporting times were dispersed among the categories and found that the three categories with by far the longest reporting lag time are all some form of sexual offense.



I isolated the three categories with the largest reporting delays and made violin plots of these crimes to determine the distribution of this data and if there are outliers causing the data to skew towards a longer lag time. The item that stands out the most is the flat line for Prostitution and Commercialized Vice. I investigated this further and learned there were only two incidents of this crime in the dataset and those incidents had a large lag time. A generalization on the reporting time for Prostitution and Commercialized Vice would require additional data. Both Forcible Rape and Sex Offenses have outliers on the high end of the reporting lag time (such as the 658 month reporting time found previously), but the bulk of the reporting times seem to range from 0-50 months and tapering off as the reporting time approaches 100 months. Overall, the large reporting times for these categories are due to either not enough data or outliers. But even without the outliers, the average reporting time is greater than the 4 month reporting time of the crime with the next highest reporting time.



After reviewing the trends in current crime, I wanted to make predictions for future crime. As noted above, I calculated the RMSE for three time-series predictive models and determined that the Holt-Winters model, with an RMSE of 14.74 was the best fit for my data. The RMSE of 14.74 indicates that on average, the predicted volume of crime will be off by roughly 15 crimes per day. The dataset has a daily mean of 117.92 crimes per day. Therefore, on average the predictions will be about 80% accurate. I then trained and fit a Holt-Winters model outside of the loop so the police department can generate future predictions. For this example, I opted to make predictions for the next 6 months by forecasting the next 182 entries proceeding the training set. I compared the result to the trends in the previously generated line graph and found the results to be reasonable.

Conclusion

The results of this study show that crime happens most often during the summer months and during the afternoon/evening hours. Additionally, more aggressive crimes such as larceny, assault, and rape are happening in higher volumes than less aggressive crimes such as embezzlement, counterfeiting, and liquor laws. This is contributing the “Most Dangerous City in America” title that St. Louis often holds. If the city wants to separate itself from this title, it will need to focus on combating these dangerous crimes. They can begin their efforts by focusing on District 4 which has the highest volume of crime overall as well as some of the highest volume in the “dangerous” crime categories. The study also shows that sex crimes such as rape and sexual offense have a delay in reporting time that is far greater than the reporting delay in other crime categories. The police department will want to research options for shortening the reporting time of sex crimes.

Assumptions

Because the original data source recorded crimes that were reported in 2020, this study assumes that all crimes that occurred in 2020 were also reported in 2020. It is known that this is not accurate, specifically in the case of sex crimes that are often reported many months or years after the incident occurred.

Additionally, when reviewing the above visualizations, it can be easy to assume that the population across the districts is equal and equate crime rate with volume. Due to the lack of a population data point, this is a study strictly on the volume of crime.

Limitations

Many of the limitations in this project are a result of missing data points. It would be helpful to know the population of each district so I could join it with this dataframe and a true crime rate could be calculated as opposed to the volume of crime per district. This would add context to the above findings.

Additionally, when predicting the data, I split the data into testing and training sets with Sept 1st being the split date. There is slight seasonality to this data so I would have preferred to train on a full year and test on a different year's full data set. Having two full years of data that can be divided into training and testing sets would have likely lowered the RMSE.

Challenges

One of the reasons I was drawn to this dataset is because I had previously only done one project with time series data, and I was interested in enhancing these skills. As such, one of my challenges was ensuring that data, specifically the data related to dates and time, was formatted correctly so I could extract the necessary month, day, and time elements. I then had to familiarize myself with Holt-Winters, ARIMA, and SARIMA predictive models and their various hypertuning options. After a steep learning curve, I achieved the best RMSE I could and opted to move forward with Holt-Winters. I believe Holt-Winters resulted in the best RMSE because it has a built-in capability to smooth data and account for anomalies such as the spike in crime in early June. Becoming more educated in time series data and their associated predictive methods proved to be a rewarding challenge.

Future Uses/Additional Applications

In addition to providing details to the police department regarding how they might want to allocate their personnel and resources to combat various categories of crime, this study can also be used as a benchmark that can be referred to in future years. This study utilized datasets containing crimes that were reported in 2020. A study of future crimes can be compared to this to determine if the volume of crime has decreased, if the lag time between a crime reporting and a crime occurring has decreased, or if there is a category of crime that has seen a substantial increase and needs to be addressed. The police department can also determine if they want to focus on a certain category of crime, a certain district, or a combination of crime category and district and make predictions on that subset of data. Having a benchmark such as this will help the police department know if any mitigation efforts have been successful or if they need to change their approach.

Recommendations

Based on the above findings, I have four overarching recommendations for the police department.

1) Increase staffing and patrolling during the summer months, the evening/nighttime hours, and in District 4. These time periods and places have the largest volume of crime compared to their counterparts.

2) Larceny is by far the category of crime with the greatest number of occurrences both city wide and in most districts. The police department should make a focused effort to lower larceny crimes and educate the public on preventative measures that can be taken.
3) The crime categories with the greatest reporting delays are all various categories of sex crimes. To shorten that lag time, offer varying platforms and methods for someone to report a sex crime so a victim or advocate can report in a manner that is most comfortable to them.
4) Compare the predicted volume of crime to the actual volume of crime that occurred during the predicted period. This will inform the department if crime decreased and their mitigation methods worked, or if they need to alter their approach to deterring crime.

If the police department addresses these topics, I anticipate that a future study will show a decrease in the volume of crime in times and places with high criminal activity, a decrease in larceny, and a decrease in the lag time when reporting a sex crime.

Implementation Plan

To implement the findings and recommendations in the above study, the heads of the police department and city officials will need to collaborate and determine how they want to prioritize the main findings and recommendations. They will also need to develop a timeline and budget for how they will address each recommendation. It would be a good practice to engage the community and media partners to keep citizens informed of the goals of the department, how to prevent crime, how to protect themselves from crimes, and how to report crime. Predictions were run for the 6 months (182 days) following the last day of the training set. The police department can determine if this is a meaningful time frame to them or if they would like to adjust the number of predictions made to better suit their needs.

Ethical Assessment

The largest ethical concern I have with this study is that I do not want to “victim blame.” It is ideal if a crime is reported immediately so evidence is more readily available. However, if a victim is not comfortable coming forward until months or years later, it is important that the victim is not shamed. This will likely result in future victims opting to not report at all, which would be more detrimental than a delay in reporting. Additionally, part of the recommendations is to have a public education effort advising citizens how to prevent crime. It is ideal if valuables are locked in a safe, but if valuables are not secured and someone is robbed the victim is not responsible for the robbery.

Another ethical concern is the logic of using 2020 as the dataset for this study. Due to covid and low enlistment in the police force that year (Police Executive Research Forum, 2022), viewers should be aware that the 2020 data might be an anomaly compared to the crime volumes and trends of the surrounding years.

CITATIONS

Federal Bureau of Investigation. (2004). *Uniform Crime Reporting Handbook*. https://ucr.fbi.gov/additional-ucr-publications/ucr_handbook.pdf

Police Executive Research Forum. (2022, March). *PERF Survey Shows Steady Staffing Decrease Over the Past Two Years*. https://www.policeforum.org/workforcemarch2022

Schneider, J. (2023, November 25). New crime ranking lists St. Louis as third “most dangerous” US city. FOX 2. https://fox2now.com/news/missouri/new-crime-ranking-lists-st-louis-as-third-most-dangerous-us-city/

St. Louis Metropolitan Police Department. (2008). Crime Data Frequently Asked Questions. https://www.slmpd.org/Crime/CrimeDataFrequentlyAskedQuestions.pdf

St. Louis Metropolitan Police Department. (2021). SLMPD Downloadable Crime Files. https://www.slmpd.org/Crimereports.shtml

APPENDIX

Appendix A – FBI Uniformed Crime Reporting

Title: " Uniform Crime Reporting Statistics: Their Proper Use "

Publisher: Federal Bureau of Investigation (FBI)

Published: May 2017

URL: https://ucr.fbi.gov/ucr-statistics-their-proper-use

Uniform Crime Reporting Statistics: Their Proper Use

Since 1930, participating local, county, state, tribal, and federal law enforcement agencies have voluntarily provided the nation with a reliable set of crime statistics through the Uniform Crime Reporting (UCR) Program. The FBI, which administers the program, periodically releases the crime statistics to the public.

Usefulness of UCR Data:
UCR crime statistics are used in many ways and serve many purposes. They provide law enforcement with data for use in budget formulation, planning, resource allocation, assessment of police operations, etc., to help address the crime problem at various levels. Chambers of commerce and tourism agencies examine these data to see how they impact the particular geographic jurisdictions they represent. Criminal justice researchers study the nature, cause, and movement of crime over time. Legislators draft anti-crime measures using the research findings and recommendations of law enforcement administrators, planners, as well as public and private entities concerned with crime problems. The news media use the crime statistics provided by the UCR Program to inform the public about the state of crime as it compares to the national level.

Pitfalls of Ranking
UCR data are sometimes used to compile rankings of individual jurisdictions and institutions of higher learning. These incomplete analyses have often created misleading perceptions which adversely affect geographic entities and their residents. For this reason, the FBI has a longstanding policy against ranking participating law enforcement agencies on the basis of crime data alone. Despite repeated warnings against these practices, some data users continue to challenge and misunderstand this position. Data users should not rank locales because there are many factors that cause the nature and type of crime to vary from place to place. UCR statistics include only jurisdictional population figures along with reported crime, clearance, or arrest data. Rankings ignore the uniqueness of each locale. Some factors that are known to affect the volume and type of crime occurring from place to place are:

- Population density and degree of urbanization.
- Variations in composition of the population, particularly youth concentration.
- Stability of the population with respect to residents; mobility, commuting patterns, and transient factors.

- Economic conditions, including median income, poverty level, and job availability.
- Modes of transportation and highway systems.
- Cultural factors and educational, recreational, and religious characteristics.
- Family conditions with respect to divorce and family cohesiveness.
- Climate.
- Effective strength of law enforcement agencies.
- Administrative and investigative emphases on law enforcement.
- Policies of other components of the criminal justice system (i.e., prosecutorial, judicial, correctional, and probational).
- Citizens' attitudes toward crime.
- Crime reporting practices of the citizenry.

Ranking agencies based solely on UCR data has serious implications. For example, if a user wants to measure the effectiveness of a law enforcement agency, these measurements are not available. As a substitute, a user might list UCR clearance rates, rank them by agency, and attempt to infer the effectiveness of individual law enforcement agencies. This inference is flawed because all the other measures of police effectiveness were ignored. The nature of the offenses that were cleared must be considered as those cleared may not have been the most serious, like murder or rape. The agency's clearances may or may not result in conviction, the ultimate goal. The agency may make many arrests for Part II offenses, like drug abuse violations, which demonstrate police activity but are not considered in the clearance rate. The agency's available resources are also critical to successful operation, so its rate of officers to population and budget should be considered. The UCR clearance rate was simply not designed to provide a complete assessment of law enforcement effectiveness. In order to obtain a valid picture of an agency's effectiveness, data users must consider an agency's emphases and resources; and its crime, clearance, and arrest rates; along with other appropriate factors.

Because of concern regarding the proper use of UCR data, the FBI has the following policies:

- The FBI does not analyze, interpret, or publish crime statistics based solely on a single dimension interagency ranking.
- The FBI does not provide agency-based crime statistics to data users in a ranked format.
- When providing/using agency-oriented statistics, the FBI cautions and, in fact, strongly discourages, data users against using rankings to evaluate locales or the effectiveness of their law enforcement agencies.

Promoting Responsible Crime Analysis

For more information about the UCR Program, visit https://ucr.fbi.gov. For web assistance, please contact the FBI's Crime Statistics Management Unit at (304) 625-4830.

10 Questions

Question #1: Are there any trends in the time of year that crime occurs?

Question #2: What caused the spike in crime at the beginning of June?

Question #3: Are there any trends in the time of day that crime occurs?

Question #4: Which categories of crime are most/least prevalent in St. Louis?

Question #5: How is crime distributed among the districts?

Question #6: What are the top crimes in each district?

Question #7: What is the delay between when a crime occurs and when it is reported?

Question #:8 How is the reporting delay dispersed among the different categories of crime?

Question #:9 Why are the top three average lag times so much greater than the reporting lag time of other categories?

Question #10: Can future volumes of crime be predicted?