BALTIMORE CRIME ANALYSIS

Intriguing insights in to crime in the City of Baltimore. Data analysis and modeling performed in Python.

KOLTEN SAFRON

SYNOPSIS

PROJECT BACKGROUND

For as long as I can remember the world of crime has always been something that was intriguing to me. I probably have my mom to thank for this, as she also shared this same interest with me. She was always there to watch crime documentaries or tell stories from her constant crime mystery books she reads. So I sat there and I figured what better way to put my interest of crime and passion of data together, then to do a crime analysis. Thankfully for me, the City of Baltimore posts crime data online so I was able to get a data a great data set for free. This consisted of each major crime that occurred in Baltimore from 2011 to present. Including such data as, crime date, crime time, neighborhood, district, crime code, crime description, weapon.

My goal with this project was to dive into Baltimore's Crime and see what trends I could discover. I wanted to go behind just looking at the data that was found in Baltimore's data set, so I tried to get creative and think what else could affect crime and how can I try and draw conclusions from it. To take it a step further I wanted to challenge myself and see if I could create a machine learning model that would predict the number of crimes that would occur in each neighborhood on a given day or given month.

In order to save everyone time, I prepared a quick synopsis at the start summarizing the key takeaways from my data analysis as well as a summary of the model. Please sit back and enjoy the read as we dive into this together.

KEY TAKEAWAYS

Below summarizes my key takeaways from my exploratory data analysis done over the period 2011 through 2021:

- One of the toughest parts about creating this model and the general data analysis, was dealing with the large amount of missing values and inconsistent data in the data sets used. Addressing this was probably the single largest part of this project as a lot of data cleanup was necessary to get it into a usable form.
- Data set included major crimes with 14 different descriptions (Ex. larceny, agg assault, robbery – residence, homicide), the top 3 crimes (larceny, common assault, burglary) made up 54% of the major crimes from the period.
- Of the total crimes 43% of them incurred inside where as 57% occurred outside. Excluding any auto related crimes, 54% of the crimes incurred inside whereas only 46% incurred outside.
- The city was broken down into 9 sections. The east side of town (northeast, east, southeast) consisted of the highest occurrences making up 38% of the total crime. With the northeast containing the highest occurrences of crime.
- Neighborhood showed to have a strong correlation with the number of crimes that occurred.
- Crime was at its highest during May through October, afterwards it would slow down an average of 16% for the winter and early spring. February was the slowest month from crime, down from the average for the year by 21% or down from the high crime rate months of May through October by 27%.
    - An additional takeaway from the monthly crime rates, is that the amount of crime that occurred inside stayed relatively consistent throughout the year (with the exception of February as it has such a significant drop in crime). So the significant increase in crime that occurs from May through October happens primarily outside.

- On a yearly level we can see crime has stayed relatively stable with the exception of a drop in 2016, jump up in 2017, and drop in 2020 & 2021. As expected the drop in crime from 2020 & 2021 would be a result of the COVID-19 pandemic, as people were sticking to themselves more. This is interesting though given the negative effect on the economy and unemployment, you would think that people would become desperate and crime would increase. Nothing in particular stands out for the drop in 2016, however for 2017 a likely contributor to the increase in crime may be the US Presidential Election. When reviewing the data we saw that January 2017 was by far the highest amount of crime that had occurred in January through all the years, being 17% higher than the next closest. The correlation here would be that January 2017 would have been when Trump officially took office and Baltimore (Maryland) has been historically a strong Democratic state.
- When looking at the crime as broken down by the day of the month, there isn't much to take away other than the 1st day of the month being the highest crime rate by a large margin. It was 8% higher than the average and 5% higher than the next closet.
- The time of the day, as done in hour increments, also has a large effect on the occurrence of crime. Crime is more common during the hours of noon to midnight. That half of the day makes up 63% of the crime that occurs. With the hours of 3AM to 6AM being the slowest making up only 7% of the total crime, yet it represents 17% of the total hours in the day.
    - Additionally, you can see that certain crimes tend to follow this overall trend of being slower after midnight through the morning, however some do not. Take Larceny which this crime occurs almost double from the hours of 11 to 17 compared to any other point of the day. Is the conclusion to draw here that Larceny is more likely to occur when people are gone to work?
- Crime is fairly consistent across the week with only small deviations. Friday has the highest crime with an increase of 5% over the average, whereas Sunday has the lowest with a decrease of 5% over the average.
- Looking at a large public event like a NFL football game, there was little to no correlation with an increase in crime. In the downtown area, where the football stadium was, there was barely any change in crime. This may be a result of increased police presence on game days knowing the potential for it to get out of hand.
- Baltimore City Data
    - Population – The population has been decreasing consistently. This should lead to a decrease in crime, however this is not perfectly apparent when comparing crime by year with population. When looking at the crime rate per capita we can see that even though the population decreased 2.3% combined in 2017 & 2018 the crime rate increased an average of 0.5%.
    - Poverty Population – One of the surprising findings was that the amount of the population below the poverty threshold has been consistently dropping since 2014 (23.2% of the population in 2014 vs 20.9% of the population in 2019), but there doesn't appear to be a correlation with the drop in crime.
    - Median Household Income – Incorporating the median household income in Baltimore vs US seemed to provide little correlation with the overall crimes. However, in the predictive model both of these data points had a strong enough weight assigned to them, that there is a correlation that exists. It is just not evident to the human eye.
    - Unemployment rate – There does appear to be a correlation with unemployment rate, as the unemployment took a significant drop from 2011 to 2015, dropping from 10.6% to 7.4%. We can then see that through those years, there was a corresponding decrease in the crimes that occurred.

PREDICTIVE MODEL

Two different machine learning models were created to attempt to predict the number of crimes that would occur. The first model was created to predict the number of crimes that would occur in each neighborhood on a given month (Ex. Crimes in Abcd Neighborhood in January 2018) whereas the second model would predict this on a daily level (Ex. Crimes in Abcd Neighborhood on January 8, 2018). Both models were created using the XGBoost machine learning algorithm (decision tree-based). The models incorporated much of the data and findings as discussed in the key takeaways above. The performance of these models was judged based off of their mean average error (the average error or difference that existed between the predicted crimes per day/month for each neighborhood vs the actual number of crimes that occurred on that day/month for each neighborhood).

1) Crimes per neighborhood on a given month
    - Mean average error of the model – 0.62
    - Average crimes per neighborhood that actually occurred – 13.3
    - Model accuracy – 95%

2) Crimes per neighborhood on a given day
    - Mean average error of the model – 0.43
    - Average crimes per neighborhood that actually occurred – 1.57
    - Accuracy – 73%

As you can see from the summary above, the model on a monthly level was much more accurate. This would be expected as we would not anticipate enough crimes to occur in each neighborhood that it would be easily predictable, as shown by only an average of 1.57 crimes occurring in a neighborhood each day in Baltimore.

LINKS MY TO WORK

Main repository for project - https://github.com/kds55/Baltimore-Crime-Analysis

Project code as completed in python - https://github.com/kds55/Baltimore-Crime-Analysis/blob/main/baltimore_crime_analysis_code.py

EXPLORATORY DATA ANALYSIS

CHARTS AND FURTHER DETAILS TO BE ADDED/UPDATED IN THE NEAR FUTURE

All charts for the time being can be viewed through the source code as linked above in my GitHub.