Assignment 7: Project Report

Anant Rajeev

Introduction

Conducting this analysis would be extremely useful because while the COVID-19 pandemic for the past 2 years has been thoroughly documented, researched, and analyzed, there are different angles of supplemental research that would be extremely beneficial to look at, particularly with regards to criminal and police incidents. Performing this analysis will give key insight into further and more potential serious societal impacts of the COVID-19 pandemic in Davidson County, Tennessee in specific. For a once-in-a-generation pandemic, it is unclear how the added social restrictions, unorthodox government mandates, and alternative forms of business practices would have impacted Nashville's (the biggest metropolitan city in Davidson County) police incidents and crime rates. In this analysis, I plan to uncover some productive takeaways which would be helpful in a few key ways: assisting government officials to allocate resources properly within the police department in the future, prepare for crises better, and even model the policing patterns during a pandemic for future reference. One thing is clear: We know that the COVID-19 pandemic has affected different parts of the country in different ways, but the deep, common human-centered consequences are there for all to see. Policing incidents are just one way of logging the human impact of this pandemic and I hope to tell a cohesive story about that using the data I have aggregated. As part of my Assignment 5, I developed a concise research question to tee off this analysis: Did the COVID-19 pandemic have a material impact on the number of police incidents in Nashville, Tennessee at critical times in the pandemic?

Background/Related Work

There existed a lot of literature about COVID-19 pandemic case counts being linked in some capacity to increased crime rates across cities both domestically and internationally. For example, in the article "Crime in the Time of COVID" written by David S. Abrams suggests that there could be a relationship between the number of COVID cases and level of restrictions and the counts of criminal incidents across large metropolitan cities across the United States. The research explained that states across the nation have experienced larger increases in violent crime and similar decreases in property crime. This inspired me to see if there was an actual mathematical way to produce an analysis that addressed this issue.

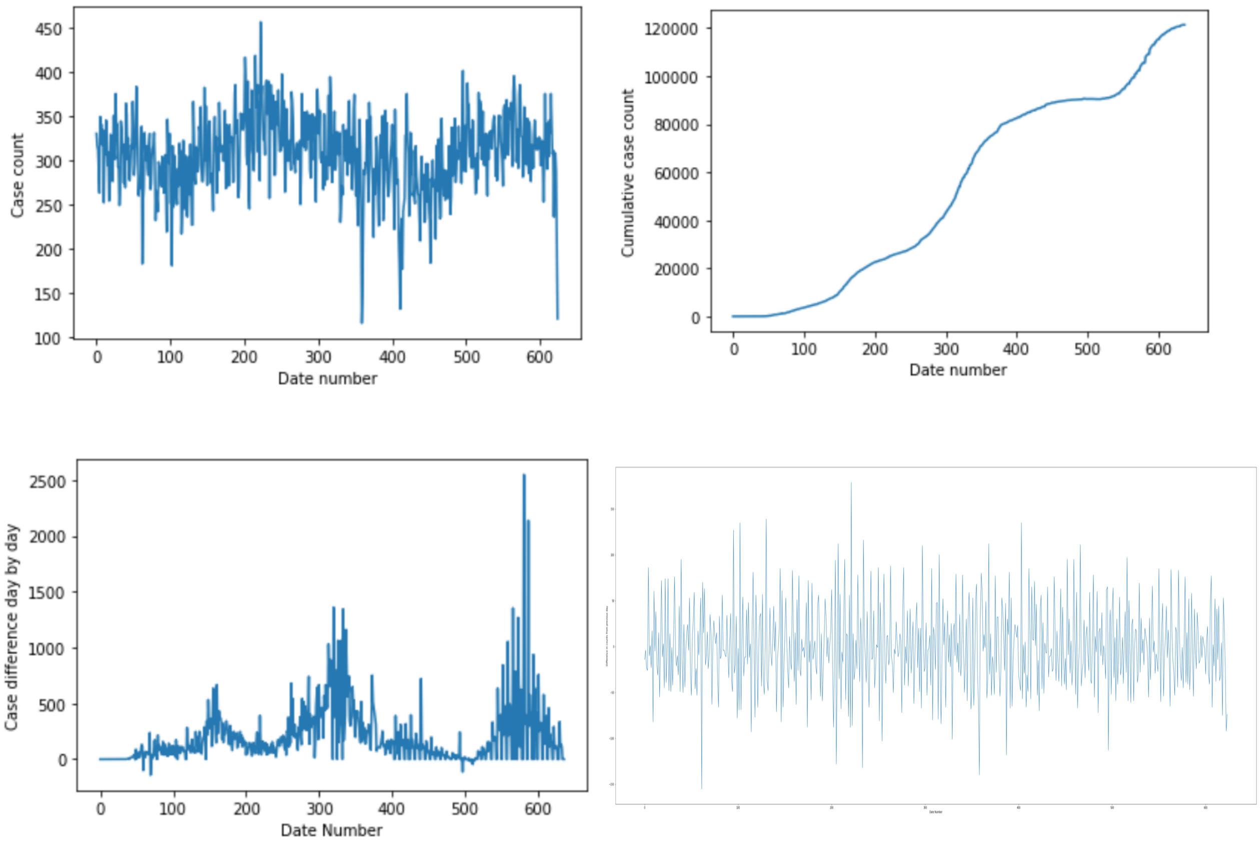
Another article entitled “Crime in California during the COVID-19 Pandemic” went through statistical research of how crime in California was impacted by the COVID-19 pandemic. A key outcome of this research was that “...between 2019 and 2020, property crime declined by 8% in California, while violent crime increased slightly by 0.8%.” This further cemented my inclination to partake in this analysis. It showed that the correlation and relationships we might see in the Nashville area might go either way, and the only way to find out was to go through my own statistical research.

This brought me to my hypotheses.

Null Hypothesis: Through each critical time period in the pandemic, there was not a significant difference in the means of counts of police incidents than in any other period of time during the pandemic.

Alternative Hypothesis: Through each critical time period in the pandemic, there was a significant difference in the means of the counts of police incidents than in any other period of time during the pandemic.

To do some initial exploratory research, I developed some plots in Python using the “matplotlib” library to assess any patterns and relationships right off the bat between Davidson County’s COVID-19 case counts and their count of police incidents.



Figures 1-4 (left to right & top to bottom) **Figure 1:** Count of police incidents by day in Davidson County, TN. **Figure 2:** Count of cumulative COVID cases over time in Davidson County, TN. **Figure 3:** Difference in COVID cases per day in Davidson County, TN. **Figure 4:** Difference in police incident cases per day in Davidson County, TN.

Just by looking at the data it is hard to discern patterns, but I hoped that through my analytical methods I would perform next, I would be able to break down some of the nuances in the plots and extract some relationships that might not be visible from just simple visualizations.

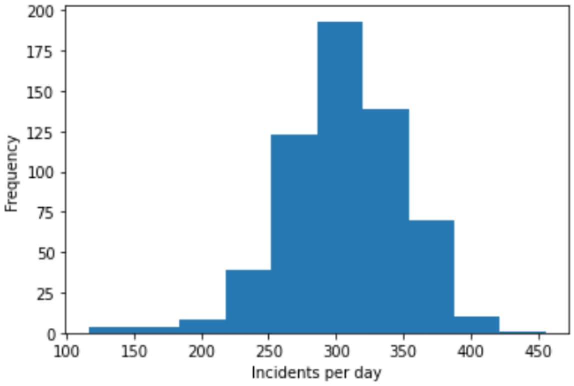
Methodology

Based on my literature research about this topic, I narrowed my tests down to be focused on 3 specific and critical time periods that related to the pandemic:

1) The COVID-19 mask mandate in Davidson County, TN from 7/3/2020 to 5/14/2021

2) The presidential election and subsequent lame duck period from 11/7/2020 to 1/20/21
3) The county’s vaccine availability from 3/8/2021 to 9/16/2021 (last day I have data for in the dataset)

To test whether or not there was a material difference in the number of police incidents as it related to the COVID-19 pandemic, I wanted to employ two methods: Two Sample t-testing to test the difference in population means and a linear regression to visualize any patterns and relationships between populations and see how well we can use a one-degree polynomial to model our police incident counts. I wanted to go about this in a two-pronged fashion because I felt there was ample space to understand the relationship between COVID-19 case counts and crime incidents in Davidson County. To try and capture the nuances that are sometimes lost with a simple statistical test like a two sample t-test, I wanted to build a very elementary linear model to assess and potentially discover similar patterns in case counts and police incident counts.



The assumptions that I had to validate in order to confirm the use of two sample t-testing were that the police incident counts in two different time periods were independent of one another, that the data points were continuous, and that the histogram of the data points in question followed a general normal distribution. All of these assumptions held up for both the distribution of data for the police incidents in Nashville, TN and the pandemic case counts in Davidson County, TN.

It is important to understand that there are ethical implications that exist while going about this analysis. One key human-centered issue to be considered is that the police incidents do reflect actual real-life incidents that involve people’s actual experiences. As I went through my analysis, I did my utmost to be cognizant of the fact that police incident counts should not solely be a matter of numbers but also be used in discussion and findings in the context that they were developed in. This meant addressing the research question as a human-centered issue as well as being a mathematical issue.

Findings

The findings did not point to anything concrete with regards to my initial hypotheses. But with any research, there are results and conclusions that are also drawn that answer questions that weren't initially poised at the inception.

Two sample t-testing with COVID-19 case counts in Davidson County, TN

p-value	**Population 1**	**Population 2**
0.02319	COVID vaccine start/end date	All other days
2.29e-31	Start/End of presidential election	All other days
1.40e-10	Start/End of mask mandate	All other days

With regards to case counts, the three time periods I analyzed had significant results. The difference in the case counts in each of these time periods as compared to all other days were statistically significantly different. What this indicated is that these time periods were good time periods to use with regards to testing whether or not police incident differences were also statistically significant.

Two-sample t-testing with police incident counts in Nashville, TN.

p-value	**Population 1**	**Population 2**
0.87558	COVID vaccine start/end date	All other days
0.08850	Start/End of presidential election	All other days
0.41732	Start/End of mask mandate	All other days

We found that with regards to crime data, there were no time periods in which the police incident counts were significantly different than the rest of the days. When computing the statistical tests however, the time period between the start and the end of the presidential election (11/7/2020 - 1/20/21) yielded closer to significantly different results than any other days in our dataset (0.08 compared to 0.9 and 0.4).

Discussion/Implications

What these statistical testing results tell us is that these three time periods during the pandemic are certainly critical to look at. Based on the analysis above and the results of my tests they do carry a level of significance. If we can find a way to find the data to expand the research past just police incidents during these critical time periods, we could expose some patterns we may never know existed.

Based on our findings above, we know there were in fact times during the pandemic that deviated from the typical pattern of case counts and those should be looked at more closely. We knew that COVID-19 case counts fluctuated throughout the pandemic, finding certain periods during the pandemic required some literature research as well as these tests to properly identify.

After training my linear regression machine learning model, I used the model to predict incident counts. In this plot below, the blue line and points refer to the predicted values and the orange line and points represent the true values that I was comparing against. To evaluate the performance of the regression model at a more granular and mathematical level I wanted to compute the root mean square error (RMSE),

- **Root Mean Squared Error:** 287.85 incidents
- **Mean difference between values in predicted/true set:** 101.16 incidents

I used the mean difference between values in the predicted and true set (101.16 incidents) as a threshold to further examine how accurate my predictions were. Of the 119 cases I used in my test set, 27 predicted values were less than 101.16 incidents away from the corresponding true values.

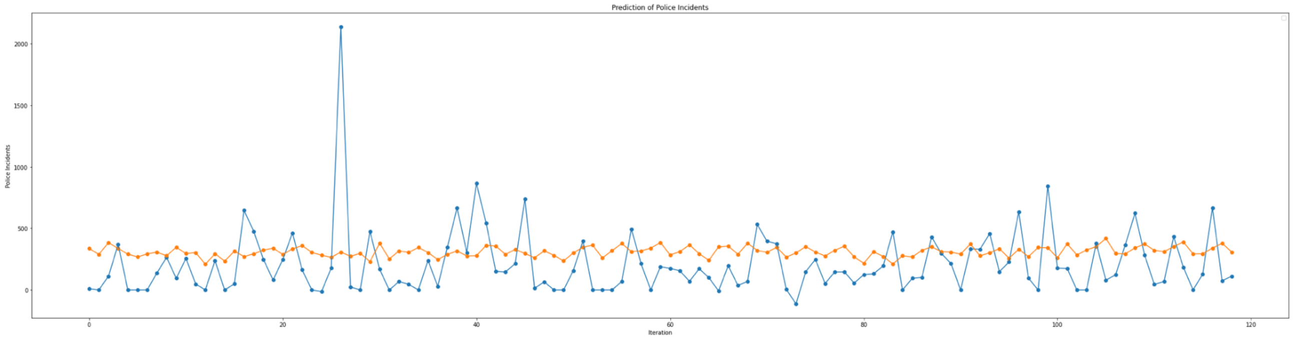


Figure 5: Plot of predicted values (blue) vs. true values (orange) of police incidents

Limitations

There were several limitations that started to give way as I went through this analysis. However, the biggest key limitation that came with going about producing this analysis was the lack of publicly available data at the incremental, daily level we wanted in the specific county of Davidson County, Tennessee. Even though I had a one key data set that addressed my hypothesis, I was hoping that there would be more publicly available data that I could use en route to accurately answer my research question. For example, this would come in the form of more granular data available for metropolitan police incident data for all of Davidson County, TN rather than just Nashville, TN. Although Nashville does contain a majority of the population of Davidson County, the police incident data does not reflect all of the incidents throughout the county. This is potentially a reason for us not finding any statistically significant results.

Because my linear regression model was only a one-degree polynomial, my predicted values had a high risk of being very turbulent and containing a lot of outliers. The model could not be trained enough and thus would likely render some inaccurate results. Going off of this, it was also difficult to assess the accuracy of the model itself. A linear regression model is only as good as the data that is used to train it, so I had to assess accuracy with a slightly more lenient lens than I would have with a stronger and more robust model.

Furthermore, with regards to the statistical assumptions, I had to make some decisions as to whether or not the COVID-19 case counts in Davidson County were actually independent and followed a normal distribution. Because pandemic case counts inherently grow by people having it spreading it to people who don't have it, assuming that the counts of COVID-19 cases per day are independent of one another is an issue worth addressing.

A potential limitation that I was concerned about at the outset was the licensing on the dataset I used. The Kaggle dataset I used for my analysis entitled "Metro Nashville Police Department Open Data" did not have any licensing information as part of the metadata on the website. However, after doing some research on Kaggle's licensing policies is that once a dataset is published to the site, the data is essentially public domain, similar to an MIT General Public license.

Conclusion

Going back to the initial research questions that prompted this analysis, did COVID-19 case counts have any statistically significant relationship with regards to the police incidents during those critical time periods and during the pandemic overall?

The short answer is no, but there are nuances and some interesting findings. With any concrete analysis that uses a few different techniques to accomplish it, we can extract some conclusions about our scenario from different angles. Because our p-values for assessing the differences in police incidents through those particular time periods were less than 0.05 (our significance threshold) we know that those time periods were different than others in relation to the pandemic. Whether or not criminal activity and police incidents were significantly different than normal in those time periods is unclear but significant differences in other socioeconomic activity during those time periods remains to be seen.

Going through an analysis like this, it opened my eyes to the socioeconomic implications of a once-in-a-century pandemic. While there are so many angles to look at the effects of a pandemic, police incidents are just one of them, and by even achieving a p-value of 0.08, we can deduce that police incidents during some of those time periods have differences that approach significance. By obtaining more broken down data about police incidents and more informative specifics of those incidents, we can better assess this relationship.

References

1) https://econofact.org/crime-in-the-time-of-covid
2) https://www.nature.com/articles/s41562-021-01139-z
3) https://www.capolicylab.org/wp-content/uploads/2021/09/Crime-in-California-During-the-Covid-19-Pandemic.pdf

Data Sources

1) https://www.kaggle.com/kennethhesse/metro-nashville-police-department-open-data?select=Metro_Nashville_Police_Department_Incidents.csv
2) https://www.kaggle.com/antgoldbloom/covid19-data-from-john-hopkins-university?select=RAW_us_confirmed_cases.csv