Using Data Science Pipeline Principles to tackle "Did the Coronavirus pandemic impact crime within the UK?"

Leihl Zambrano, Foundation Computer Science, University of East Anglia

Abstract

In response to the coronavirus (COVID-19) pandemic, crime data was collected throughout the UK and published to the public. Data Science Pipeline is the process and tools to extract data to present in an understandable format. By utilising local crime data, specifically Suffolk crime data between years 2019 and 2020 (Pre-COVID and POST-Covid), by dissecting the crime data, this document will tackle whether the coronavirus has impacted crime on a local and national level.

Introduction

- When conducting this research, data is sourced from https://data.police.uk/data/, this is open-source data that is released to the public by the UK police.
  Some notes for coronavirus and measuring crime:
  Not all incidents will result in a crime being identified or recorded. If it is concluded that an offence has been committed, either immediately when reported, or after being recorded as an incident, the incident will be recorded as a crime and will be identified as domestic abuse-related
- Data from **2019** and **2020** Files from police.data was selected
- Utilising Python, with libraries "**numpy**", "**matplotlib**", "**panda**" and "**remoteZip**", these are tools to create visualisations from the data.
- Python is more flexible as it can run multiple data files from police.data to create an environment where all files are merged. With python automating same actions to different files.
- This would reduce the repetition actions and speed up calculations within datasets.
- Whereas excel, data will have to be manually edited which would not be feasible as I am working with multiple datasets

Methods

- During the data collection phase, I decided to download documents through python using the **RemoteZip** library. Through collecting the datasets this method, it allowed me to directly download datasets from police.data, without downloading the entire archive. I could have approached data collecting through downloading all archives from police.data, however, it requires the download of excess datasets that are not in use.
- The months **April-September** was collected from 2019 and 2020 to compare Pre-COVID and Post-COVID crime rates. Based on restrictions, the April-September time frame is most appropriate as it highlights **key times**, such as restrictions coming into effect from 23 March 2020 and were gradually eased from 13 May 2020.
- The **Suffolk** dataset will be utilized for this research, as that is the area that I am familiar with during the occurrence of the pandemic.
- This data is sourced from a public organization, the data may have minimal **cleansing** done to it. Therefore, before approaching the data I checked whether the data has been cleaned, I utilised **isNull**. With both datasets, I searched for cells that contained no values.
- I decided not to cleanse the data of empty cells, as both 2019 and 2020 showed similar crime trends, presenting a positive increase of crime followed by a decline. However, if I had an opportunity in the future to develop this work, I would remove the empty cells so that data can show more accurate results between datasets.
- Graphing the data through a line graph as, it allows the data to compare **changes over the same time**. It provides the visual representation of the difference between 2019 and 2020 cumulative frequencies during the months.
- To compare the crime rates of Suffolk, I created a total crime bar graph, directly comparing the total crime committed during the April-September timespan. Further calculation was done to figure whether the crime has been impacted. This is done by using % change calculation

Results

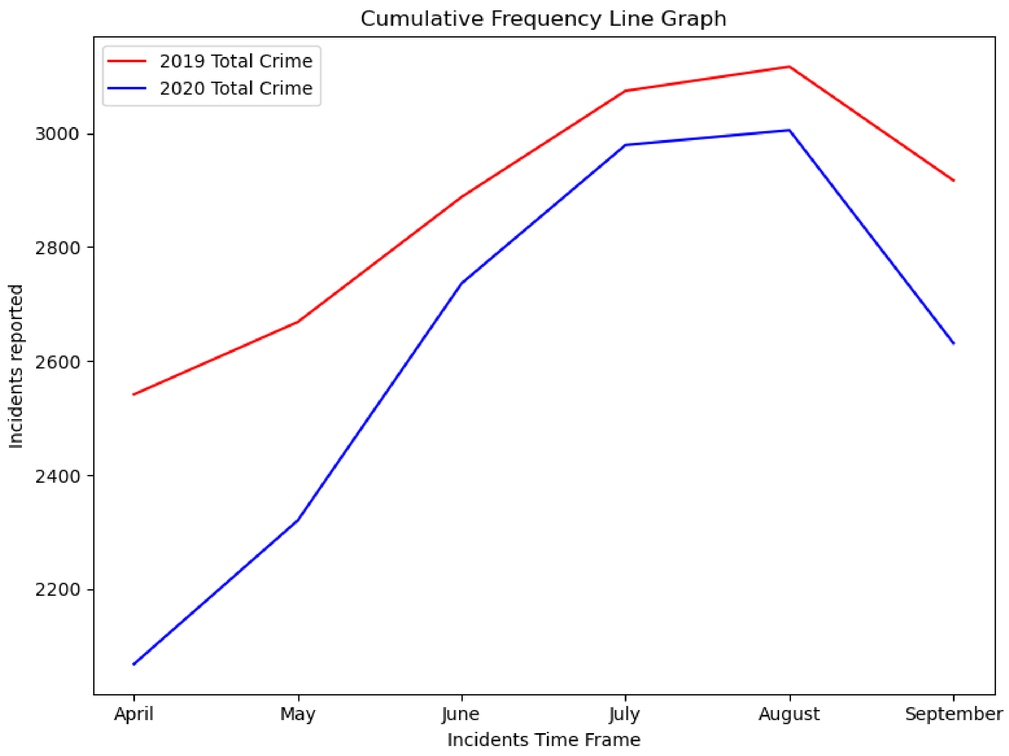

Figure 1: Total Crime Line Graph from April to September. Monthly totals for number of offences will be affected by the number of days in a month as well as the number of weekends and bank holidays.

Both years (2019 and 2020) could suggest that an increase of crime rates during April-August, until September where it shows a decline in crime.

Conclusion

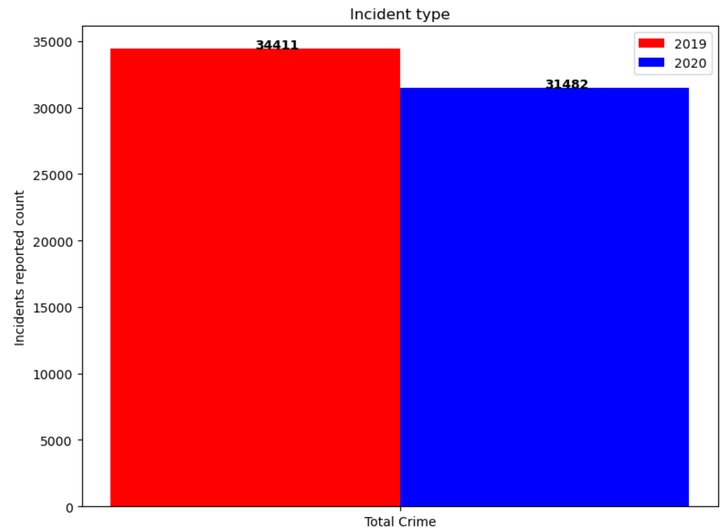

Figure 2: Total Crime Reported from April to September. During the 2019 timespan, it yielded 34411 crime incidents. Compared to 2020, where it yielded 31482 crime incidents.

Through these results, it implies that there is a difference in crime rates between both 2019 and 2020. Within this timeframe there is a 8.5% decrease in crime.

Through the use of data science pipeline, the Suffolk crime dataset shows interesting data collected from 2019-2020. With the use of data science pipeline, further research upon this topic would allow more concise results to answer whether or not the pandemic had an effect. Due to limiting myself to look at one specific dataset, it may be that other areas around the UK may yield similar or vastly different results. Nevertheless, with the research done, the data implies that there was a slight decline in overall crime in 2020.

References

DOMO (no date) *How to Use the Data Science Pipeline*. Available at: https://www.domo.com/glossary/what-is-the-data-science-pipeline#:~:text=The%20data%20science%20pipeline%20refers,insights%20based%20on%20real%20data. (Accessed: 10th May 2023).

Police Data (2020 and 2019) Crime Data Download Available at: https://data.police.uk/data/ (Accessed: 4th May 2023).