Analysis of crime in Denver, Colorado, 2018-2020

Ella Hidveghy, Charlie Davies and Rhiannon Wallis

Non-Technical Summary

Denver, Colorado, consistently ranks among the American cities with the highest crime rates. Thus, understanding the prevalence of various crimes across the city and examining the impact of different factors is crucial for developing effective crime prevention strategies.

Utilising the dataset provided by Denver's Open Data Catalog (https://www.denvergov.org/opendata/), our analysis revealed no significant variation in white-collar crimes across police districts in Denver, Colorado, throughout 2019. Additionally, the daily occurrence of white-collar crimes did not display any substantial fluctuations. However, it is essential to recognize the limitations of these conclusions due to the unique classification of Denver International Airport and the lack of data on police district 7.

We also investigated the influence of weather and time of year on burglary rates, using data from the National Centers for Environmental Information website. Our findings suggest that weather and seasonal factors do not reliably predict the number of burglaries in the city. Consequently, further research into environmental, geographical, and socio-economic determinants is necessary to better understand burglary patterns.

Interestingly, our analysis uncovered that the Covid-19 pandemic has had a differential impact on crime rates. Following the onset of the pandemic, car thefts surged by over 50%, while other crime rates dropped by a third. This highlights the non-uniform effect of Covid-19 on various types of crime. It is worth noting, however, that the limited availability of post-pandemic data may influence the robustness of these conclusions.

Introduction

Denver, Colorado is a largely populated city in the United States of America, with one of the highest crime rates in the country. Looking into what events or context affects particular crimes in cities is vital to help predictive policing, future government policy, and strengthening public perception of transparency of crime data.

Our report aims to investigate changes and effects on specific types of crimes in the city of Denver in 2018-2020, focusing on the following:

- The variability of white-collar crimes across the city's seven districts and whether the daily count varied across 2019.
- Whether weather data can be used to predict the daily number of burglaries.
- Whether the effects of the Covid-19 pandemic impacted Car Thefts and other crimes.

White-collar crimes are generally nonviolent and financially motivated crimes such as Fraud, Insider Trading, Embezzlement, Money Laundering or Intellectual Property Theft. As estimated by the FBI, these types of crimes cost the United States between $300 and $600 million annually (Wikipedia, 2023). Using statistical methods, we aim to explain and understand any difference in white-collar crime occurrences across the Denver districts, and look further into these crimes from 2019 data, studying whether they vary significantly over the year.

Previous studies have shown links between the weather and crime and its use in prediction. Baryshnikova, Davidson and Wesselbaum, 2021 highlighted a significant increase of violent crimes due in hotter daily temperatures, and Solomon et al., 2022a found that contextual factors such as historical burglaries, spatial information, weather features and socio-economic information play an important role in predicting burglaries using deep learning. Our report focuses on Denver's daily burglary counts and building prediction models on weather data elements. We will look specifically at the minimum and maximum daily temperatures and daily precipitation levels and evaluate their predictive abilities on data from 2018-2020.

The onset of COVID-19 led to widespread job losses and school closures, potentially increasing crime levels. Conversely, lockdowns and restricted mobility might have reduced crime rates. Studies found that stay-at-home orders from 16th March 2020 initially suggested a global crime decrease (Nivette et al., 2021). However, further research revealed that the impact varied across different crimes and cities (Nivette et al., 2021). We will explore the effect of COVID-19 on car thefts and other criminal activities using various statistical models.

Methodology

Our analysis was supported by three datasets.

The first data set was a list of all the crimes reported across Denver in 2019. It contains 62,817 observations of the following variables:

- Crime categories: aggravated assault, all other crimes, arson, auto theft, burglary, drug alcohol, larceny, murder, other crimes against persons, public disorder, robbery, sexual assault, theft from motor vehicle, and white-collar crime
- Geographical coordinates where the crime happened
- District the crime occurred
- Neighbourhood the crime occurred
- First possible month of the year the crimes occurred
- First possible day of the month the crimes occurred
- First possible hour of the day the crimes occurred
- Month of the year the crime was reported
- Day of the month the crime was reported
- Hour of the day the crime was reported

Before proceeding with our analysis, we removed missing values from the district variable.

The second dataset contained data of the daily weather in central Denver, Colorado, from 2018-2020. It contains 1096 observations of the following variables:

- Calendar date: from 1st January 2018 to 31st December 2020
- Maximum and minimum temperature reported that day (in Fahrenheit)
- The precipitation level recorded (in inches)

There are no missing values in this dataset, which we checked before analysis.

The final data set contained daily counts of crimes for 14 categories from 2018-2020 in Denver, Colorado. It contained 13,375 observations of the following variables:

- Crime categories: aggravated assault, all other crimes, arson, auto theft, burglary, drug alcohol, larceny, murder, other crimes against persons, public disorder, robbery, sexual assault, theft from motor vehicle, and white collar crime
- Daily count: from 1st January 2018 to 31st December 2020
- Year when the crime occurred.
- First possible month of the year the crimes occurred
- First possible day of the month the crimes occurred

For example, for crimes that occurred on 28th January 2019 would have the observation:
Year when the crime occurred = 2019, First possible month of the year the crimes occurred = 1, First possible day of the month the crimes occurred = 28. Before beginning analysis, we checked for missing data and found none.

In our first analysis, we investigated whether the proportion of crimes classified as white-collar varied across the seven police districts in 2019. Then, we explored whether the daily number of white-collar crimes varied across 2019

To begin, we used the data set that contained a list of all the crimes reported across Denver in 2019 to count the total number of crimes across all categories for each district as well as the number of crimes called white collar. Since the total number of crimes per district varies, we plotted a bar chart illustrating the proportion of crime in each district that was white-collar. To note, *Other Crime* refers to crime that is **not** white-collar.

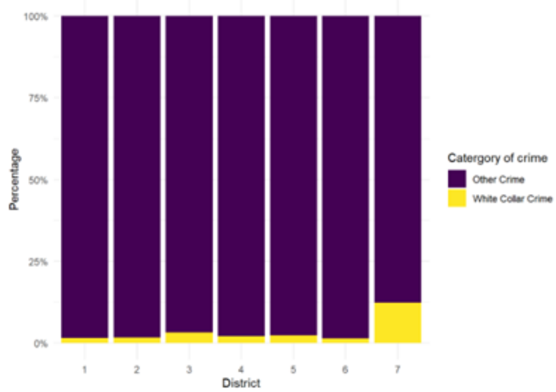

Figure 1 *: Bar Chart of the proportion of White-Collar crime*

From this exploratory analysis, districts 1 through 6 seem to have a consistent proportion of white-collar crime (≈2%) whereas district 7 has a much larger proportion of white-collar crime (≈12%). To further investigate this, we conducted an ANOVA test which analyses the difference between the means of two or more groups.

The p-value of this test is 0.151 which is larger than any reasonable significance level. Therefore, there is insufficient evidence to suggest that white-collar crime varied across districts in 2019.

To explore whether the daily number of white-collar crimes varied across 2019, we filtered the data set that includes the counts of crimes from 2018-2020 in Denver, resulting in data that was exclusively from 2019 data and was categorised as *White Collar*. There is little evidence to suggest that the daily number of white-collar crimes vary within each month. Therefore, we averaged the daily number of white-collar crimes for each month in 2019 to see whether there is a seasonal trend across the year.

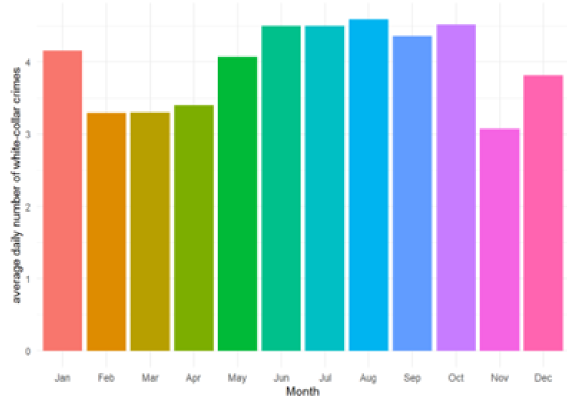

Figure 2 *: Bar Chart of Daily Average White-Collar Crimes per month in 2019*

From figure 2, it is unclear whether a trend between month of the year and average daily white-collar crimes exists. Therefore, we should include the variable in any Generalised Linear Model.

As our response variable is a count, a Poisson model with a logarithmic link is a good option. We defined daily count of white-collar crime (DW) as the response variable and Month of the Year (M) as a factor variable. Note: from our analysis previously, we do not include district number as a variable as there is no significant difference between them.

$$DW_i \sim Poisson(\theta_i)$$

$$Log(\theta_i) \sim \beta_0 + \alpha_i \tag{1}$$

$$\alpha_1 = 0,\ i = 1,\ldots,12$$

To check for overdispersion, a goodness of fit test is performed which results in a significant p-value, implying that our data is over dispersed. To compensate for this, a negative binomial model with logarithmic link is fit to the data.

$$DW_i \sim \mathrm{NegBinom}(\pi_i, \vartheta)$$

$$Log(\pi_i) \sim \beta_0 + \alpha_i \tag{2}$$

$$\alpha_1 = 0,\ i = 1,\ldots,12$$

This choice of change is further validated by the AIC of the negative binomial model (1486) being smaller than the poisson model (1487).

When plotting the model, we see that only 5 months are significant (June – October). To evaluate the significance of this variable to the model, we compared the AIC of the model with and without the month of the year variable. We found that the AIC does not differ and remains at 1486. When performing diagnostic tests on the 2 models, we found the residuals better fit the model with the month of the year variable in the GLM. Therefore, it is appropriate to keep the variable in the GLM.

In our second analysis, the variables of interest are the minimum and maximum temperature, precipitation and date, which are present in the second dataset, and the number of burglaries daily, which we extracted from the third dataset. This dataset provides the daily counts for all types of crimes over the same period (2018-2020), which we subsetted to select daily counts of burglaries. For analysis of the predictors, we restructured and combined the necessary data from these two datasets.

As one would expect, there is high correlation between daily minimum and maximum temperature. Since our goal is prediction, the multicollinearity between these variables won't violate assumptions of a GLM here as we're assuming that the correlation between these variables is the same in data the model is built on and the 'new' data. Thus we may include all important variables in any modelling for prediction.

We may also question whether the minimum and maximum temperature of the day is representative of the overall temperature that day, we get a better representation of the weather if the data included an average temperature over the whole day.

First, we investigated the response variable, *Burglary Count*. To investigate the predictive abilities of the weather we fit a Poisson model with logarithmic link function, as our response variable is a count. We chose to use a model with just the variables of interest, relating to temperature and precipitation. Defining the response, Daily Burglary Count, as BC and our predictors, Maximum Temperature, Minimum Temperature, and Precipitation as $MaxTemp$, $MinTemp$ and $Precip$ respectively, we fit the following model:

$$BC \sim Poisson(\theta)$$

$$\log(\theta) \sim \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 MaxTemp + \boldsymbol{\beta}_3 MinTemp + \boldsymbol{\beta}_4 Precip \qquad (3)$$

To further investigate prediction with these variables, we added the Month and Year to the model. We fit the following model, with variables defined as before, and Month of the Year defined as $\boldsymbol{\alpha}_i$: and Year defined as $Year_i$. We left Year as a numerical variable as we'd prefer to be able to forecast burglary count for future data (e.g. 2021-present) and using years as levels would eliminate this. Including the Year variable accounts for the underlying upwards trend in burglaries we noticed in explanatory analysis.

$$BC_i \sim Poisson(\theta_i)$$

$$\log(\theta_i) = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 MaxTemp_i + \boldsymbol{\beta}_2 MinTemp_i + \boldsymbol{\beta}_3 Precip_i + \boldsymbol{\beta}_4 Year_i + \boldsymbol{\alpha}_i \qquad (4)$$

$$i = 1,2,\ldots,12$$

$$\boldsymbol{\alpha}_1 = 0$$

To assess how well the models can predict, we split the data by generating a test set, a random sample of 10% of the dataset (110 observations), then trained the model on the remaining 90%

(986 observations) and calculated the predictive error of the test set.

For the predictive error we used Root Mean Squared Error (RMSE):

$$\text{RMSE:} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(\widehat{y}_i - y_i\right)^2} \qquad (5)$$

where y_i are the observed values, $\widehat{y}_i$ are the predicted values and m is the number of observations in the test set.

Our final aim was to analyse whether the Covid-19 pandemic, which began on 16th March 2020 led to a change in the number of crimes classified as *Car Thefts*. We then wanted to explore whether the conclusion changes when considering the crime category *All other crimes*. We began by analysing a subset of the data which contains daily counts of crime category *Car Thefts*. In the secondary analysis, we used a subset that contains the daily counts of the crime category *All Other Crimes*. As we were comparing crime rates before and after 16th March 2020, we separated the data into *Pre* and *Post* Covid by adding a variable to the data set defined as Covid Period with categories *Pre* or *Post*. *Pre* contains all the daily counts for Car Thefts from 1st January 2018 to 15th March 2020 and *Post* contains the daily counts from 16th March 2020 to 31st March 2020.

To determine the changes in Car Thefts that Covid-19 pandemic has caused we began by looking at exploratory analysis and then fit a linear model to the data. Because we have split our data into two important date categories, we were less interested in the exact date of the crimes. Therefore, the two variables we used were Covid Period and Daily Count. There have been suggestions that crime can be seasonal. According to research (Block, Gould and David Coldren, 1984), it has been discovered that almost all types of personal and household crimes are more likely to occur during warmer months and crimes against property are much higher in winter months. Therefore, we suggested that it might be ideal to include Month of the Year in our analysis. As we can see from figure 3, there is a difference in car thefts between the different months.

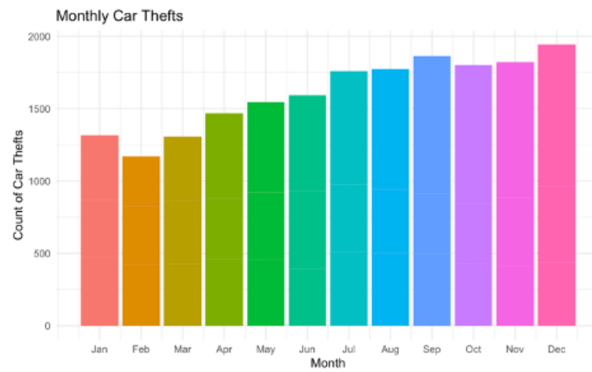

Figure 3 *: Bar Chart of Average Monthly Car Thefts in 2018-2020*

Therefore, in any modelling taking place, Month of the Year was included.
We explored different possible GLM's to decide which fits our model best. As our response variable is a count, a Poisson model with a logarithmic link is the ideal option. We fit the model with Daily Count as our response and Covid Period & Month of the Year as our explanatory variables. We defined Daily Count of Car Thefts as CT_{ij}, and Month of the Year as M_{ij}:

$$CT_{ij} \sim Poisson(\theta_{ij})$$

$$Log(\theta_{ij}) \sim \beta_o + \alpha_i + \beta_2 M_{ij} \qquad (6)$$

$$i = Pre, Post,\ \alpha_{pre} = 0$$
$$j = 1, \ldots, 12$$

We checked that adding Month of the Year to the model is appropriate by comparing the AIC with and without the variable and found that it is appropriate to include in the model. We checked for overdispersion by performing a goodness of fit test. We found a significant p value = 3.11e-15, implying that our data is over dispersed. We validated that the data is over-dispersed and that negative binomial was the more suitable model by

comparing AIC. Negative binomial had AIC=6544.6 and Poisson had AIC=6604.9.
We therefore fitted a Negative Binomial:

$$CT_{ij} \sim \text{NegBinom}(\pi_{ij}, \vartheta)$$

$$Log(\pi_{ij}) \sim \beta_o + \alpha_i + \beta_1 M_{ij} \qquad (7)$$

$$i = Pre, Post, \alpha_{pre} = 0$$
$$j = 1,\dots,12$$

Now we wanted to find out how our conclusion changes for the category *All Other Crimes*. As before, we used variable Daily Count as our response variable and Covid Period and Month of the Years as our explanatory variables.

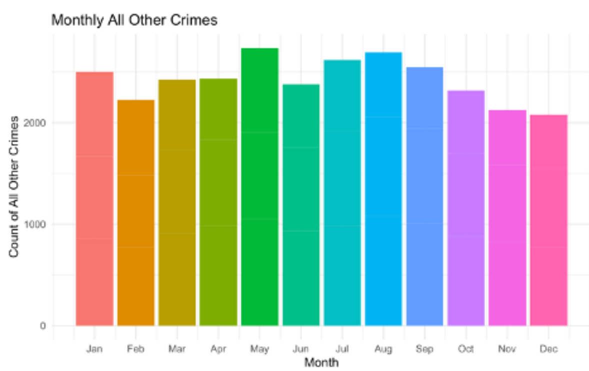

Figure 3 *: Bar Chart of Average Monthly All Other Crimes in 2018-2020*

We found that removing Month of the Year from the model reduced the AIC from 7887.6 to 7886.4, so we removed it from the model. We again fit a Poisson model with logarithmic link and through the same methods as above, found that the data is over-dispersed.
Writing AOC_i for *All Other Crimes*, we describe our GLM as:

$$AOC_i \sim \text{NegBinom}(\pi_i, \vartheta)$$

$$\text{Log}(\mu_i) \sim \beta_o + \alpha_i \qquad (8)$$

$$i = \text{Pre, Post}$$
$$\alpha_{pre} = 0$$

Results

In our first investigation, we found the following results. Between districts, we concluded (from the ANOVA test) that there was not a significant difference in the proportion of white-collar crime across the seven police districts. It is important to note that from the American Community Survey 2015-2019, no district 7 exists. Instead, Denver International Airport is categorised as district 5. This change in district classification could affect our results. Equally, there is a significant lack of data for district 7 compared to the other districts affecting the reliability of our conclusions.

To explore the variation of daily white-collar crime across 2019, we first plot a boxplot.

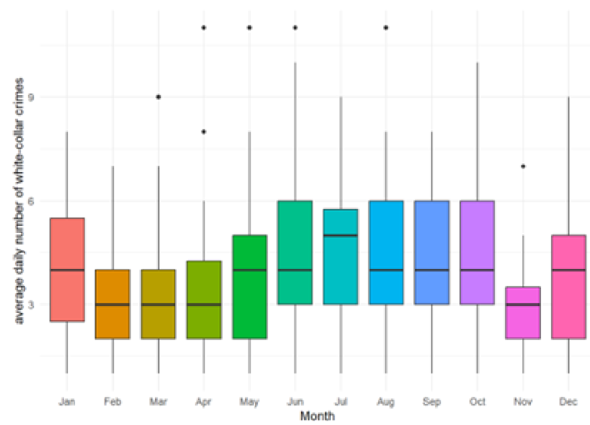

Figure 4 *: Box Plot of Daily Count of White Collar Crimes each month in 2019*

Figure 4 shows there is variation in the number of white-collar crimes across the months in 2019. Using model 2, when we set the reference level to January, we observed that the coefficient for the month November was significant (p=0.04).

Coefficient	Estimate
β_0	1.42583
α_{11}	-0.345040

Figure 5 *: Table of coefficients from model 2*

Looking at the coefficients from figure 5, we estimated that white-collar crimes decreased by 35% in November (CI [-0.011, -0.594]).
It is important to note that these results are only obtained when setting the reference level as January. If the reference level changes, the month November may no longer be significant. However, at least one α_i will be significantly different from the reference level.
In turn, although the coefficient for the month November is significant, in general it appears that the daily number of white-collar crimes does not vary month to month. Therefore, we conclude that the daily number of white-collar crimes does not change throughout 2019.

In our second analysis, we observed the following outcomes. In exploratory analysis, we investigated the daily number of burglaries over the time period 2018-2020.

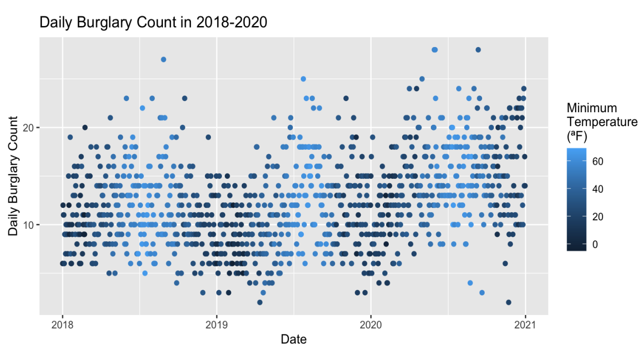

***Figure 6** : Scatter graph of Daily Burglary Count in 2018-2020*

Here we saw that the daily count of burglaries seems to be seasonal, with the highest counts generally occurring in the months with the highest minimum recorded temperature. We also observe a larger number of burglary counts in 2020, than in 2019 and 2018 which we verified with statistically significant Student's t-tests:

Test:	**p-value:**
H_0 ; $\mu_{2020} = \mu_{2018}$ vs H_1 : $\mu_{2020} \neq \mu_{2018}$	<2.2e-16
H_0 ; $\mu_{2020} = \mu_{2019}$ vs H_1: $\mu_{2020} \neq \mu_{2019}$	<2.2e-16

***Figure 7** : Hypothesis test of Burglary counts in 2020 vs 2019 and 2018*

Testing the prediction of the first fitted model, we obtained:

$$\text{RMSE} = 4.031038$$

Since the mean daily burglary count is 12.31569 and the number of daily burglaries ranges from 2 to 28, prediction error 4 implies very little accuracy.
Analysing the second fitted model, we found that including the month of the year as a predictor lowers the prediction error, however not but a large amount as the RMSE is 3.910902. This is still poor predictive ability. To check the reproducibility of this result, we repeated the test-train split and calculated the RMSE for each model on each rerun and took the mean prediction error over the repeats and we reached the same outcome; both models have similar RMSE and neither models have good prediction error. From these results, we can conclude that data of weather and time of year alone does not yield accurate predictions of the daily number of burglaries.

For our final analysis, We first plotted the following boxplot.

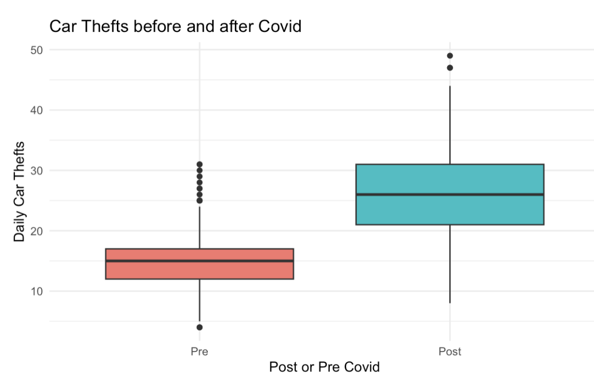

***Figure 8** : Box Plot of Car Thefts before and after Covid-19*

From the first interpretation, figure 7 indicates that the Covid-19 pandemic increased Car Thefts. Our model also found that Covid Period is a statistically significant factor at α=0.05, with p-value=<2e-16. We found that the presence of the Covid-19 pandemic led to an increase of Car Thefts. We found our coefficients to be:

Coefficient	Estimate
β_0	2.544064
α_{Post}	0.535138
β_1	0.022621

Figure 9 *: Table of coefficients from model 7*

Figure 9 shows that the Covid-19 pandemic increased Car Thefts by 53.5%, as indicated by the coefficient for Covid Period Post = 0.535138.
Our confidence interval for α_{Post} is (0.4992624, 0.57099536). Meaning that our confidence interval for percentage increase is (49.93%, 57.1%).
For our second analysis into the impact of Covid-19 on the category *All Other Crimes*, our initial exploratory analysis found the following.

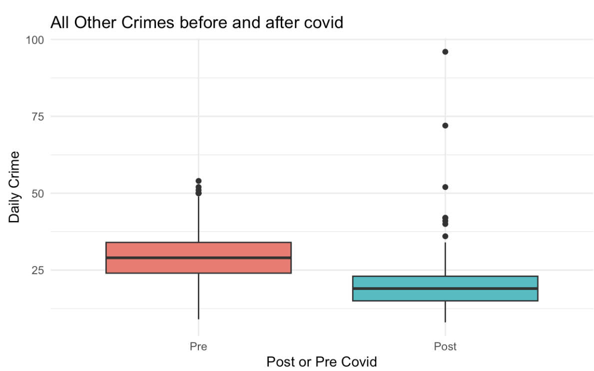

Figure 8 *: Box Plot of All Other Crimes before and after Covid-19*

This suggests a decrease in *All Other Crimes* due to the pandemic. In our model, like before we found that Covid Period is statistically significant at α=0.05 level, with p-value = <2e-16. Our coefficients were the following

Coefficient	Estimate
β_0	3.358821
α_{Post}	-0.345040

Figure 9 *: Table of coefficients from model 8*

This therefore indicates that Covid-19 led to a 34.50% decrease in *All Other Crimes*. Our confidence interval for α_{Post} is (-0.3846815, -0.3054248), resulting in the following confidence interval for the percentage, (30.54%, 38.47%).

Our assumptions for our negative binomial GLMs can be verified by plots of the residuals in the appendix. The four plots suggest that the assumptions of normality, independence, linearity, and homoscedasticity are met. We also plotted our variables Covid Period and Month against the residuals and found that our model fits the data nicely, confirming our results' reliability.

Conclusion

We first examined whether there were any notable differences in the proportion of white-collar crimes across the seven police districts in 2019. Our ANOVA test results indicated no significant difference in white-collar crime proportions among the districts, as evidenced by the large p-value of 0.151.
Subsequently, we investigated the variation of white-collar crimes throughout 2019. While our analysis revealed a significant decrease (35%) in white-collar crimes between January and November, the overall fluctuations observed in Figure 4 were not statistically significant. Consequently, we concluded that white-collar crimes did not exhibit significant variation across 2019.
It is important to note that our conclusions may be influenced by the scarcity of data for District 7 in comparison to the other districts. This data

limitation could affect the reliability of our findings, and future analyses should consider addressing this issue to obtain more robust conclusions.

In the second phase of our analysis, our objective was to forecast daily occurrences of burglaries based on daily temperature and precipitation data. Our findings suggest that relying solely on weather and time-of-year information does not lead to accurate predictions of daily burglary occurrences, as evidenced by the prediction errors for both models. Specifically, the first model had a prediction error of 4.031038, while the second model's error was 3.910902. This outcome is not unexpected, as it is improbable that burglary counts can be accurately predicted based on temperature and precipitation alone. For more precise predictions of burglary counts, a comprehensive approach should be taken that considers other factors beyond weather variables. These may include environmental aspects (e.g., urban design, lighting, and accessibility), geographical differences (e.g., variations in crime rates between urban and rural areas or across neighbourhoods), and socio-economic variables (e.g., income levels, unemployment rates, and education). By incorporating a broader range of factors, future models may yield more accurate predictions and better capture the complexities of burglary occurrence patterns.

In our final analysis, we aimed to determine the impact of COVID-19 on car thefts and all other crimes. Our exploratory findings revealed that the pandemic led to an increase in car thefts but a decrease in all other crimes. The GLM results showed a substantial 53.5% increase in car thefts during the pandemic, as indicated by the Covid Period coefficient α_{Post}=0.535138.

According to our GLM, the Covid-19 pandemic resulted in a 34.50% decrease in all other crimes, as represented by the coefficient α_{Post} = - 0.345040. The category "All Other Crimes" encompasses offences such as possession of stolen property, abduction, and money laundering. The complete list can be found in the provided dataset. Thus, our analysis demonstrates that COVID-19's impact on crime is not uniform across all categories.

The reliability of our conclusions could be improved by addressing certain limitations in our analysis. For instance, the dataset comprised 805 pre-Covid observations and only 291 post-Covid observations, reducing the statistical power of our results and calling into question the robustness of our conclusions. Furthermore, our model may not have accounted for other factors that could have influenced car theft rates, such as variations in crime counts across different districts. Investigating each district individually might yield a more accurate understanding of the impact of Covid-19 on crime.

Appendix

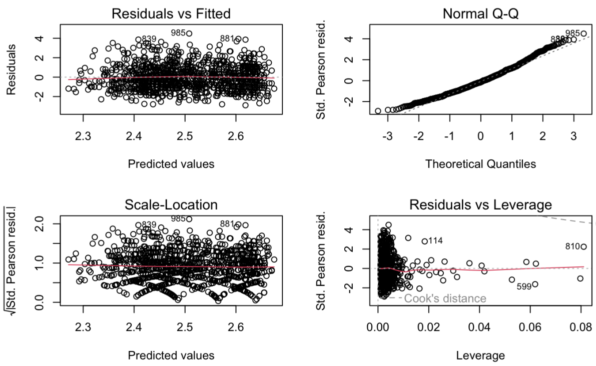

Figure 9: *Model Diagnostics for first prediction model 3*

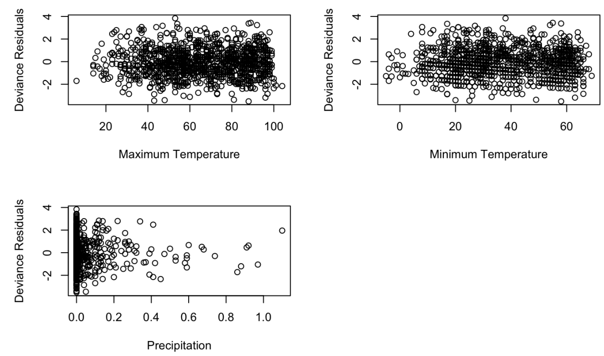

Figure 10 *: Variable vs deviance residual plots for the first prediction model 3*

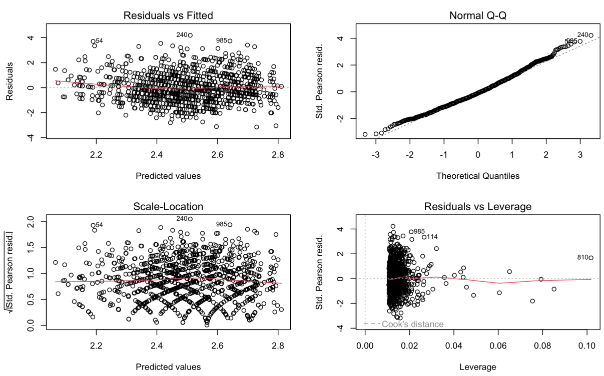

Figure 11: *Model Diagnostics for second prediction model 4*

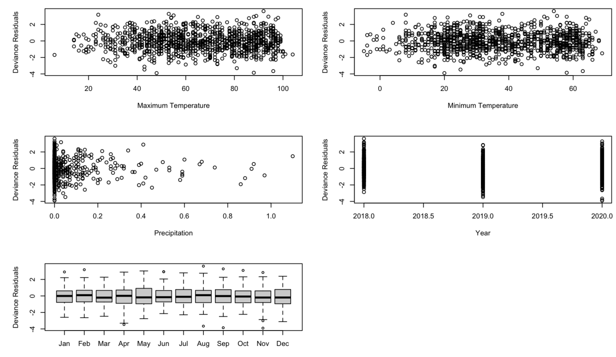

Figure 12 *: Variable vs deviance residual plots for the second prediction model 4*

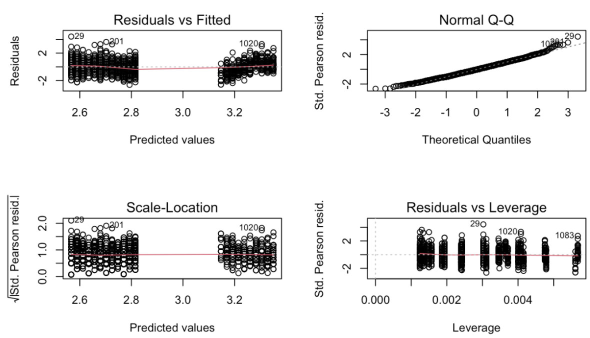

Figure 13 *: Model Diagnostics for model 7*

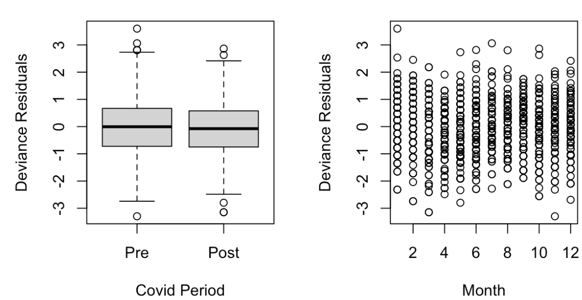

Figure 14 *: Variables vs deviance residuals plots for model 7*

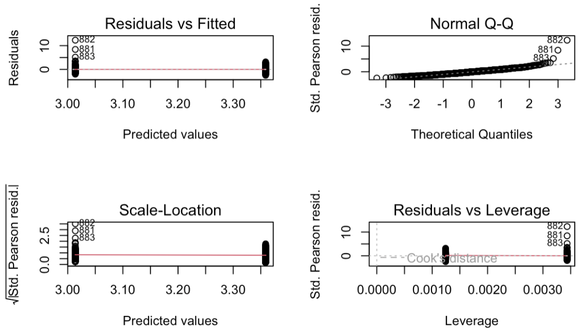

Figure 15 *: Model Diagnostics for model 8*

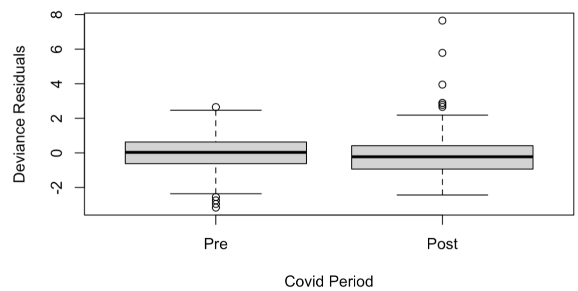

Figure 16 *: Variables vs deviance residuals plots for model 7*

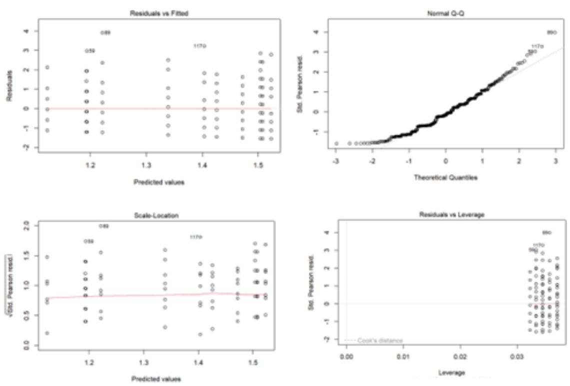

Figure 17 *: Model Diagnostics for model 2*

Bibliography

Block, C., Gould, W. and David Coldren, J. (1984). IS CRIME SEASONAL? [online] Available at: https://bjs.ojp.gov/content/pub/pdf/ics.pdf.

Nivette, A.E., Zahnow, R., Aguilar, R., Ahven, A., Amram, S., Ariel, B., Burbano, M.J.A., Astolfi, R., Baier, D., Bark, H.-M., Beijers, J.E.H., Bergman, M., Breetzke, G., Concha-Eastman, I.A., Curtis-Ham, S., Davenport, R., Díaz, C., Fleitas, D., Gerell, M. and Jang, K.-H., 2021. A global analysis of the impact of COVID-19 stay-at-home restrictions on crime [Online]. Nature Human Behaviour [Online]. Available from: https://doi.org/10.1038/s41562-021-01139-z.

Wikipedia, 2023. White-collar crime [Online]. Available from: https://en.wikipedia.org/w/index.php?title=White-collar_crime&oldid=1150545131 [Accessed 30th April May 2023].

Baryshnikova, N., Davidson, S. and Wesselbaum, D., 2021. Do you feel the heat around the corner? The effect of weather on crime [Online]. Empirical Economics [Online], Available from: https://doi.org/10.1007/s00181-021-02130-3.

Solomon, A., Kertis, M., Shapira, B. and Rokach, L., 2022. A deep learning framework for predicting burglaries based on multiple contextual factors [Online]. Expert Systems with Applications [Online]. Available from: https://doi.org/10.1016/j.eswa.2022.117042.