CROSS-SECTIONAL AND TIME SERIES ANALYSIS OF CRIME RATE ACROSS NORTH CAROLINA

by

Isabella Johnson

A thesis submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Master of Science in
Mathematics

Charlotte

2022

Approved by:

_____
Dr. Yinghao Pan

_____
Dr. Yanqing Sun

_____
Dr. Qingning Zhou

ABSTRACT

Isabella Johnson. CROSS-SECTIONAL AND TIME SERIES ANALYSIS OF CRIME RATE ACROSS NORTH CAROLINA (Under the direction of Dr. Yinghao Pan)

Unemployment rate reached an astronomical high in 2020 due to the Covid-19 pandemic. The media began reporting on increasing crime rates due to the unemployment rate rising. The Center for Disease Control (CDC) reported that over 81,000 drug overdose deaths occurred in the United States in the 12 months ending in May 2020, the highest number of overdose deaths ever recorded in a 12-month period. This leads to an interesting question of whether there is a relationship between unemployment rate, drug use, and crime rate. This study will explore that relationship across 26 North Carolina cities and counties. This will be done in a two-part analysis: a cross sectional analysis and a time series analysis.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

Covid-19 shook the world in a way no one had seen before. Unemployment rates hit an all-time high since The Great Depression [1]. Traditionally with high unemployment rates there comes a higher crime rate, as citizens do not have the means to provide income, so they resort to crime. Oftentimes, crime is related to drug usage within the media. Not only are those two related, but mental health and drug use are highly correlated [2]. The Center for Disease Control (CDC) reports that over 81,000 drug overdose deaths occurred in the United States in the 12 months ending in May 2020, the highest number of overdose deaths ever recorded in a 12-month period. This leads to the question; during this time of high unemployment, will the crime rate and drug use also be on the rise? For the purpose of this study, we will focus on violent crime. Violent crime is defined by the National Institute of Justice as any crime a victim is harmed or threatened by violence; these include but are not limited to rape, sexual assault, robbery, assault, and murder [3]. Additionally, we will focus only on opioid overdose drug use. It is estimated Synthetic opioids appear to be the primary driver of the increases in overdose deaths, increasing 38.4 % from the 12-month period leading up to June 2019, compared with the 12-month period leading up to May 2020[2]. We will focus on North Carolina for this analysis.

To see if there is a change, we must also analyze data from the time before the Covid-19 virus shut down the United States on March 26, 2020[4]. Data will span from January 2020-December 2021. One consideration is that North Carolina is a large state with much diversity in its population density, socioeconomic status, and available jobs. Due to this, we wanted to ensure that all parts of the state were represented to limit any bias. A list of 26 counties and cities were chosen and are listed in Table 1. More information about these counties will be discussed in the

data description section. In addition to the variables of crime rate, unemployment, and opioid drug overdose, other variables were selected based on their correlation with crime rates [5]. This analysis has 2 parts: a cross-sectional analysis and a time series analysis. The cross-sectional analysis focuses on March 2020 and looks at all 26 counties at that time. The time series analysis spans January 2020- December 2021 and only looks at Charlotte, NC. Additional information regarding each of these will be discussed in the Main results section.

Table 1: List of 26 Counties and Cities used in the analysis.

Charlotte	Alexander	Currituck	Jackson	Macon
Onslow	Swain	Wilmington	Winston	Hickory
Durham	Burlington	Ashe	Concord	Cherokee
Davidson	Yadkin	Wilson	Scotland	Surry
Fayetteville	Greensboro	New Bern	Rocky Mount	Asheville
Raleigh				

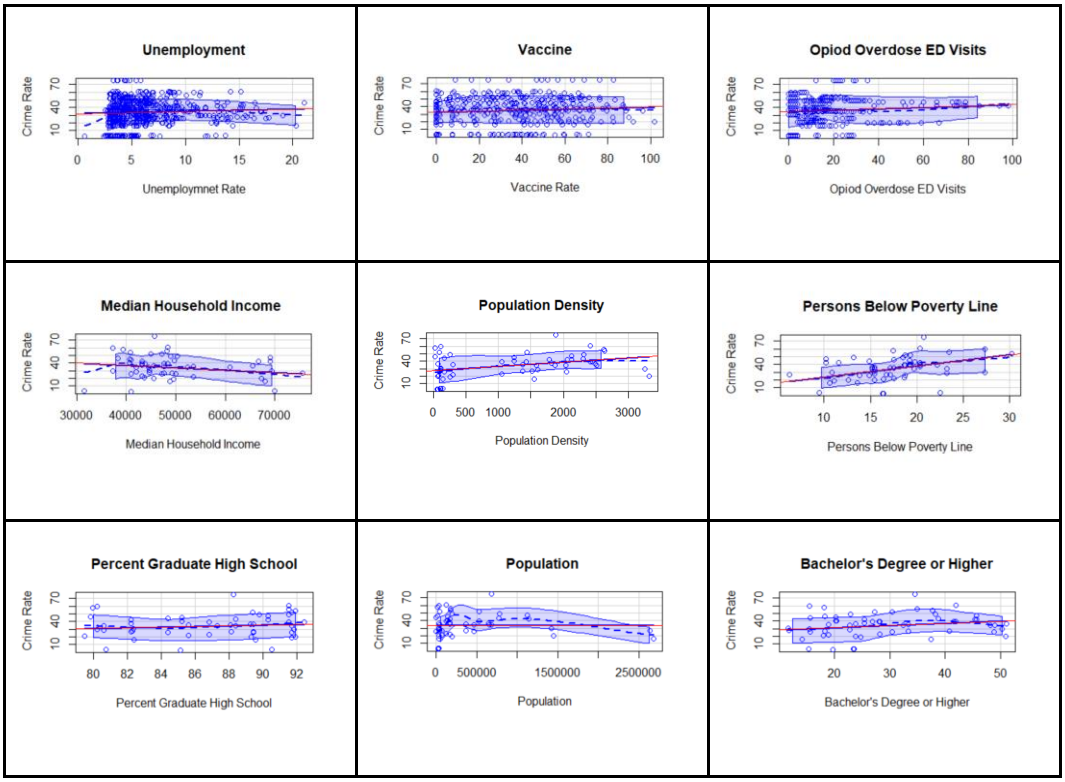
DATA DESCRIPTION

The dependent variable for these analyses is the crime rate. At first, we used a count of the number of violent crimes in each county but then decided to change to something more continuous. We changed it to the crime rate per 1000 people which is calculated by dividing the number of reported violent crimes by the total population; the result is then multiplied by 1000 to give us our crime rate variable [6]. The crime rate is reported yearly so for all of 2020 we have the same value for all months; the same is true for 2021.

In total there are 12 independent variables: date, place, unemployment rate, vaccine rate, opioid overdose, population, Median household income, people per square mile, persons below the poverty line, percent graduated high school, and percent graduated college. The date is the same as mentioned above monthly from January 2020 to December 2021. The place is also mentioned above and listed in Table 1 which are the 26 selected counties and cities across North Carolina. Graphs of each independent variable plotted against the crime rate are in Figure 1. Unemployment rates are our first independent variable. This variable is calculated monthly and has a range of .6% (Ashe County June 2020) to 21.1% (Swain County May 2020). The unemployment rate is obtained by dividing the number of unemployed persons by the number of persons in the labor force (employed or unemployed) and multiplying that figure by 100. The vaccine rate is again a monthly variable and ranges from 0-101.85(Raleigh December 2021). It is obtained by taking the total population that has received one dose of a covid-19 vaccine of any type divided by the population of the county and multiplying that figure by 100[7]. It should be noted that from January 2020-December 2020 vaccine rate is 0 due to no Covid-19 vaccine being approved for use by the Food and Drug Administration. Additionally, the vaccine rate can be

above 100 because it measures the number of vaccines administered in the county/city so someone could receive their vaccine in that county but then live in another county/city. Opioid Overdose Emergency Department Visits is the next variable and is again a monthly variable and has a range of 0-98(Charlotte August 2021) [8]. This variable is calculated by taking the total count of all the visits made to the emergency departments in the county because of opioid overdose for any reason.

Figure 1: Graphs of each variable against crime rate.



The following variables were chosen because of an article published by the Federal Bureau of Investigation which evaluates factors that impact the crime rate in an area [9]. The first is the population, which is the total number of individuals living in that county and is calculated yearly. There is a large range in these values 2046-2680820 people again because we made sure to include all parts of North Carolina in this analysis and there are many small towns and big cities

in North Carolina. Median household income is the next variable, with a range of $31563-$75578, this value is calculated yearly as well. This is calculated by taking the median in the data set, which can be determined by placing all the numbers in value order and finding the middle number in the data set. If there are two middle numbers, then take the average of the two middle numbers to obtain your median income. The median was chosen as opposed to the mean since it is less sensitive to outliers which are common in larger cities in our data set. The threshold for poverty in North Carolina for 2021 is $26,500, which leads to our next variable, the percent of the population living below the poverty line [10]. Which is calculated by taking the percent of the population in the county living below the poverty line. The range for these values is 6.3%-30.23%. It should be noted that none of the median income values fall below this line.

Population density (people per square mile) is the next variable of interest. This is calculated by dividing the total population or number of housing units within a geographic entity by the land area of that entity measured in square miles. Population density will tend to increase in cities and decrease in rural areas. The graph of population density(x-value) against crime rate(y-value) is shown in Figure 2. It should be noted there is a lot of variability in the crime rate when the population density is smaller. The last two variables deal with education level. The first is the percentage of people in the county/city that graduated high school. It has a range of 79.48%-92.43% and is calculated yearly. The second is the percentage of the population with a bachelor's degree or higher. The range for this variable is 11.65%-51.1%.

Figure 2: Graph of population density against crime rate.

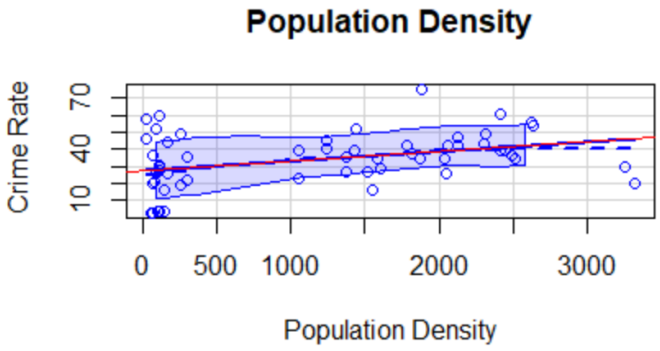
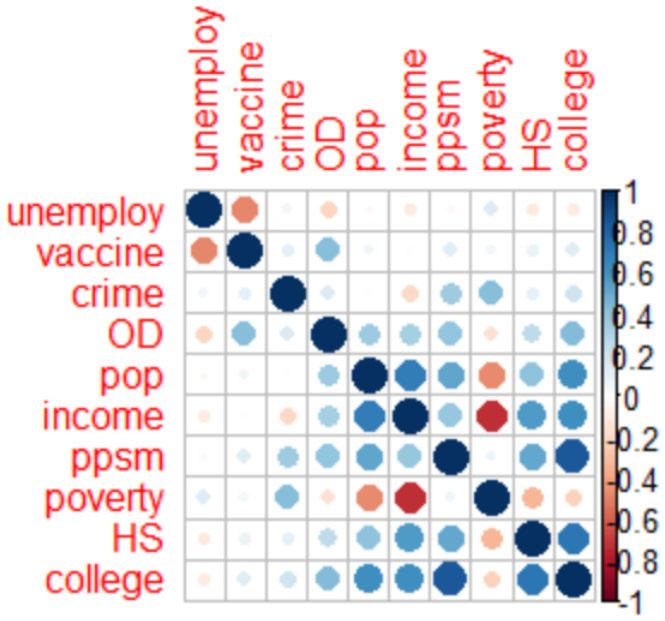


Figure 3 is a graph of the correlation between each variable. Blue represents a positive correlation with the darkest shade being closest to 1 and red represents a negative correlation with the darkest shade being closest to -1. The variables with the strongest negative correlation are Vaccine rate versus unemployment (value of -0.488), and poverty rate versus median household income (value of -0.7311) The variables with a strong positive correlation are Opioid overdose versus vaccine rate (value of 0.429), median household income versus population (value of 0.695) and percent of the population with a bachelor's degree or higher versus population density (value of 0.849).

Figure 3: Correlation graph between each variable.



MAIN RESULTS

As mentioned previously this analysis has two parts: cross-sectional and time series. The cross-sectional part focuses only on March 2020. This month was chosen because that is when the majority of the shutdowns began. The variables for this analysis were crime rate (dependent), unemployment rate, population, median household income, population density, poverty rate, percent of the population with a bachelor's degree or higher, percent of people in the county/city that graduated high school. Vaccine rates and opioid overdose were excluded in this analysis because data were not available for this period. The model that was fitted was a linear regression model using the lm command in R Studio. The fitted model is $y=53.28 + 0.4329*\text{unemployment} + 8.072*10^{-7}*\text{population} - 3.828*10^{-4}*\text{income} + 8.623*10^{-3}*\text{ppsm} + 0.5373*\text{poverty} - 0.3298*\text{hs}+0.188*\text{bachelors}$. The summary output of the model is seen in Figure 4. The model was then evaluated for each of the linear regression assumptions: heteroskedasticity, residuals vs. fitted values, normality, and linearity. Heteroskedasticity was evaluated using the Breusch-Pagan test where the p-value was 0.7578 thus, we fail to reject the null hypothesis and conclude that Heteroskedasticity is not present. Figure 5 shows the residuals plotted against the fitted values. There does not appear to be any relationship between the two, therefore it passes this assumption. A QQ plot and the Shapiro Wilke's test were used to test for normality. The plot is shown in Figure 6. The p-value for the Shapiro Wilke's test is 0.6454 hence we fail to reject the null hypothesis and say that the normality assumption is met. The last test is for normality Figures 7.1 & 7.2 show a pairwise plot of each variable against crime rate and they all have a linear relationship so therefore the model is linear. Looking at the summary of the model in Figure 4 we can see that only population density (ppsm) is the only variable that is significant and is significant at the 0.1 level. No other variables appear to be significant. Upon this finding, I

decided to take the natural log of crime rate and run the analysis above again to see if any more variables would be significant. Upon that analysis, it returned the same result. Additionally, the analysis was run for several other randomly chosen months in our interval, they also produced the same result that only population density was significant.

Figure 4: Summary output of the cross-sectional analysis.

```
Residuals:
     Min       1Q   Median       3Q      Max
-21.2325  -4.4941   0.3736   4.3425  18.3038

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       5.328e+01  7.027e+01   0.758    0.458
data_20$unemploy  4.329e-01  3.650e+00   0.119    0.907
data_20$pop       8.072e-07  6.200e-06   0.130    0.898
data_20$income   -3.828e-04  3.976e-04  -0.963    0.348
data_20$ppsm      8.623e-03  4.490e-03   1.920    0.070 .
data_20$poverty   5.373e-01  1.020e+00   0.527    0.604
data_20$hs       -3.298e-01  7.738e-01  -0.426    0.675
data_20$bach      1.883e-01  4.557e-01   0.413    0.684
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.895 on 19 degrees of freedom
Multiple R-squared:  0.6037,    Adjusted R-squared:  0.4577
F-statistic: 4.135 on 7 and 19 DF,  p-value: 0.006396
```

Figure 5: Residual versus Fitted values for the cross-sectional analysis.

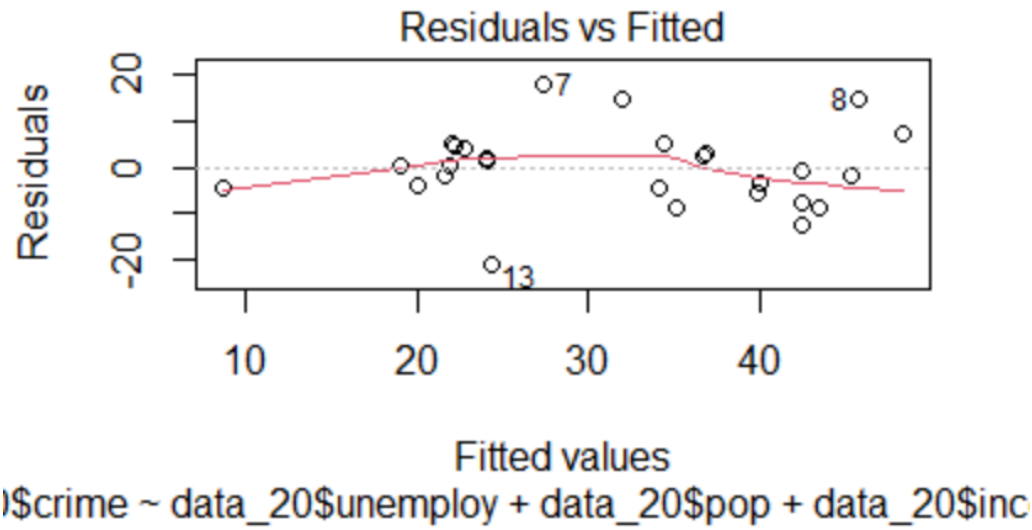


Figure 6: QQ-plot for the cross-sectional model.

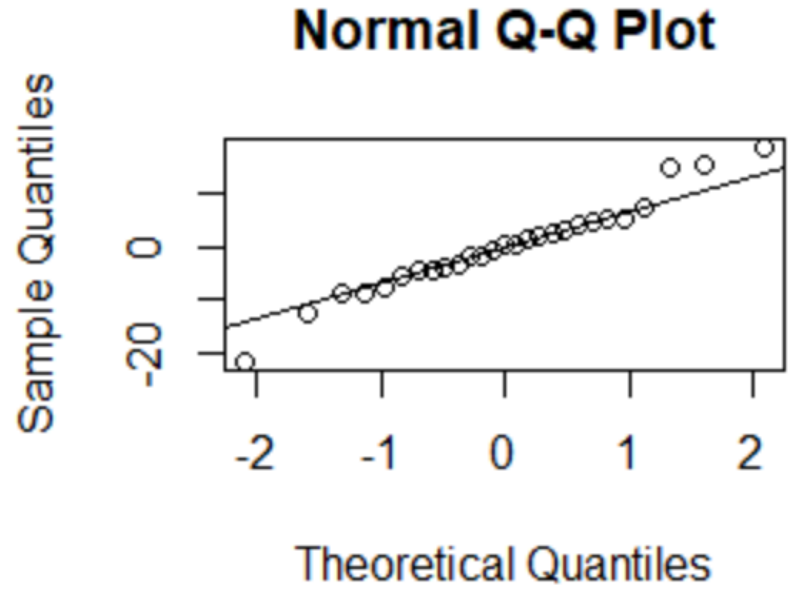


Figure 7.1: Pairwise plot of variables in the cross-sectional analysis.

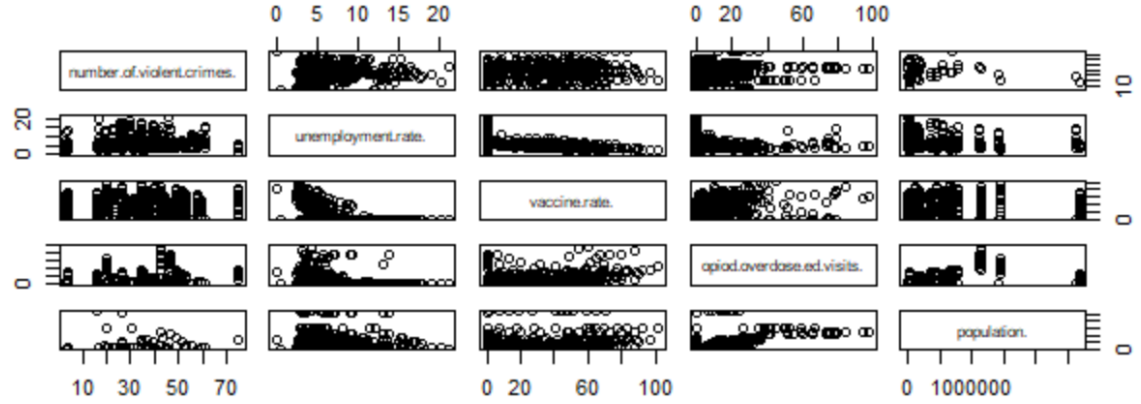
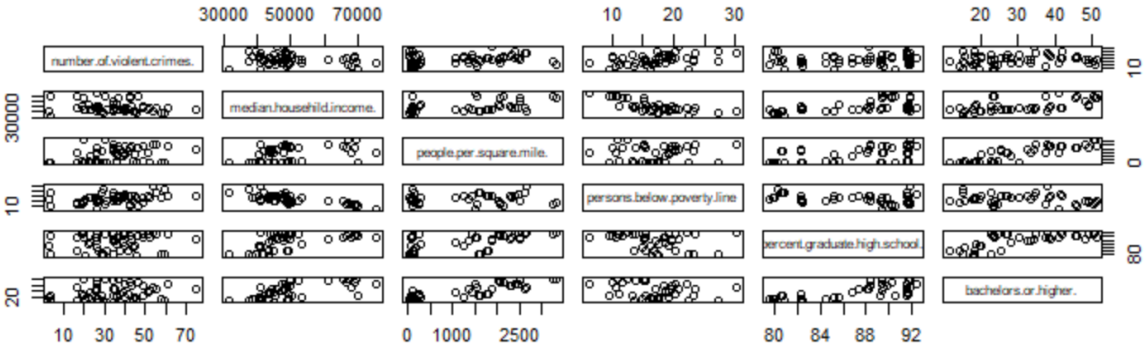


Figure 7.2: Continuation of Pairwise plot.



The second part of the analysis is a time series. The time series analysis spans January 2020 through December 2021 and it only included the city of Charlotte. Due to the crime rate being a yearly variable and therefore being the same value for all of 2020 and then a different value for all of 2021, the decision was made to fit the model for Opioid overdose and then for the Unemployment rate. Again, R Studio was used to fit these models, the command auto. Arima was used to fit the models. This command uses specific formulas to derive either an ARIMA

(Autoregressive Integrated Moving Average) or ARMA (Autoregressive Moving Average) model. For the ARIMA model Auto-Regressive (AR) terms refer to the lags of the differenced series, Moving Average (MA) terms refer to the lags of errors and I is the number of differences used to make the time series stationary. AR the time series is regressed with its previous values i.e., $y_{(t-1)}$, $y_{(t-2)}$, etc. The order of the lag is denoted as p. I integration/difference the time series uses differencing to make it stationary. The order of the difference is denoted as d. MA the time series is regressed with residuals of the past observations i.e., error $\varepsilon_{(t-1)}$, error $\varepsilon_{(t-2)}$, etc. The order of the error lag is denoted as q. For ARIMA models with differencing, the differenced series follows a zero-mean ARMA model. The formula for the general ARIMA model is $X_t = a_1 X_{t-1} + \ldots + a_p X_{t-p} + e_t + b_1 e_{t-1} + \ldots + b_q e_{t-q}$, where we Let X_t be the d^{th} difference of the time series Y_t and it is time. The ARMA model is equivalent to an ARIMA model of the same MA and AR orders with no differencing. It is also assumed to be stationary. The formula for the general ARMA model is $(X_t - m) = a_1 (X_{t-1} - m) + \ldots + a_p (X_{t-p} - m) + e_t + b_1 e_{t-1} + \ldots + b_q e_{t-q}$, m is the mean of the model, t is time. For the ARIMA model m= 0 since there is differencing so that is how we derive the formula.

The first time series model used the opioid overdose data. Before fitting the model, we must first evaluate the data for two assumptions of an ARIMA model. Data should be stationary, by stationary it means that the properties of the series don't depend on the time. Data should also be univariate because ARIMA works on a single variable. Auto-regression is all about regression with the past values. The opioid data is both stationary and univariate, so we may precede. A plot of the data against time is shown in Figure 8. Running the Auto. ARIMA command in R we can see the following output in Figure 9. The value on the right is the AIC (Akaike Information

Criterion), which is an estimator of the out-of-sample prediction error and thereby relative quality of statistical models for a given set of data. Given a collection of models for the data, AIC estimates the quality of each model, relative to each of the other models. Thus, AIC provides a means for model selection. The model with the least AIC (197.3595) was chosen as the best model, which is an ARIMA (1,0,0) with a non-zero mean. Figure 10 shows a summary of the coefficients of the fitted model. Plugging these coefficients into the formula above for an ARIMA model we get $(Y_t - 67.93) = e_t + 0.6323 * (Y_{t-1} - 67.93)$. This is also known as an AR (1) model. A forecast was then for the model and can be seen in Figure 11. The line in blue is the forecast values. The gray shaded area is the 95% prediction interval. The forecast is for 6 months ahead. There does not appear to be any trend in the forecast but appears pretty linear if it is reaching an equilibrium.

Figure 8: Time series plot of Opioid Overdose ER Visits

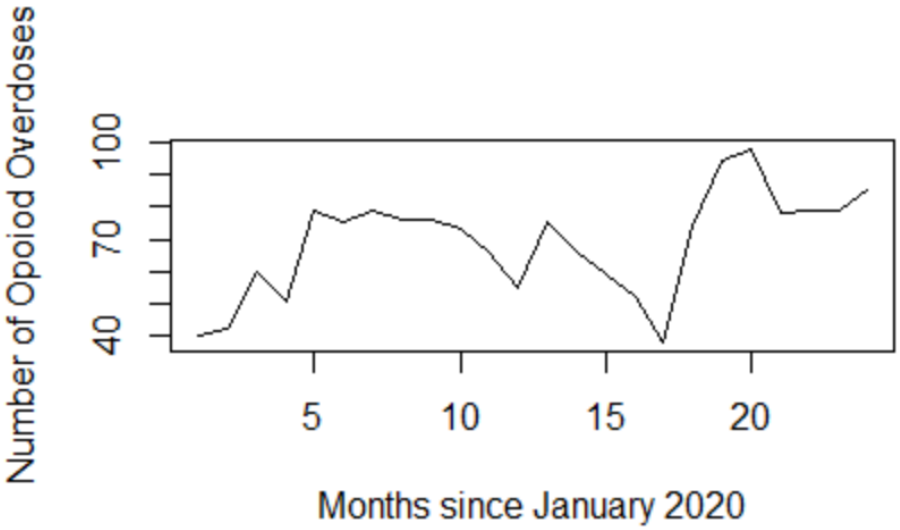


Figure 9: R output from AUTO.ARIMA command for Opioid overdose ER visits

Figure 9: R output from AUTO.ARIMA command for Opioid overdose ER visits

```
ARIMA(2,0,2) with non-zero mean : Inf
ARIMA(0,0,0) with non-zero mean : 204.9986
ARIMA(1,0,0) with non-zero mean : 197.3595
ARIMA(0,0,1) with non-zero mean : 199.7701
ARIMA(0,0,0) with zero mean     : 274.5532
ARIMA(2,0,0) with non-zero mean : 199.8552
ARIMA(1,0,1) with non-zero mean : 200.02
ARIMA(2,0,1) with non-zero mean : 203.4818
ARIMA(1,0,0) with zero mean     : Inf

Best model: ARIMA(1,0,0) with non-zero mean
```

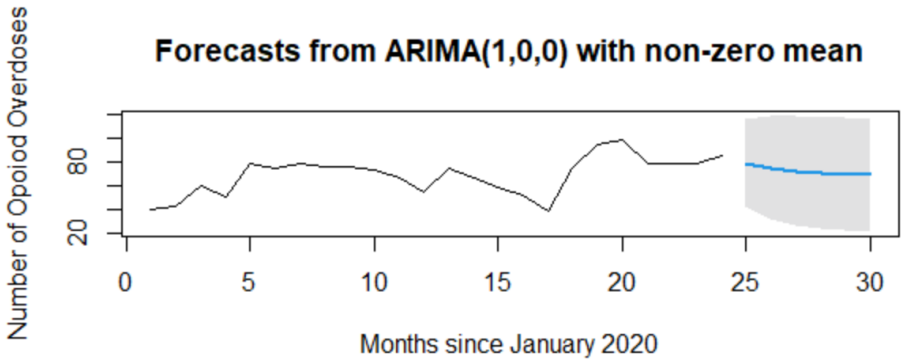
Figure 10: Coefficients for Opioid overdose Er visits ARIMA model

```
ARIMA(1,0,0) with non-zero mean

Coefficients:
         ar1     mean
      0.6323  67.9306
s.e.  0.1701   6.5499

sigma^2 = 172.6:  log likelihood = -95.08
AIC=196.16   AICc=197.36   BIC=199.69
```

Figure 11: 6-step ahead forecast for Opioid overdose ED visits



The second time series model uses the unemployment rate. The data was again evaluated under the two assumptions mentioned above and it showed to be both univariate and stationary. Figure 12 is a plot of the Unemployment rate over the time period for this analysis. Auto. ARIMA was also used in this model fitting and the output from this can be seen in Figure 13. The AIC was again used to choose the best model and the model with an AIC of 105.2032 was chosen. That is an ARIMA (0,1,0) model. This means the coefficients for the AR and the MA part are both equal to 0. Therefore $X_t = e_t$ which implies $Y_t=Y_{t-1}+e_t$ because X_t is the first difference of Y_t. Again, a forecast was done for the model and can be seen in Figure 14. The line in blue is the forecast values. The gray shaded area is the 95% prediction interval. The forecast is for 6 months ahead. The forecast is consistent with the analysis that the model is white noise.

Figure 12: Time series plot for unemployment rate

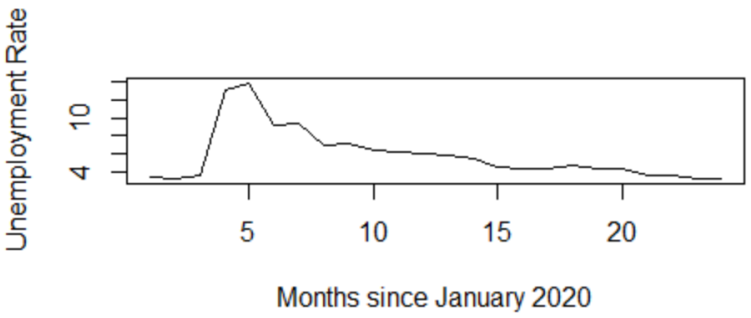
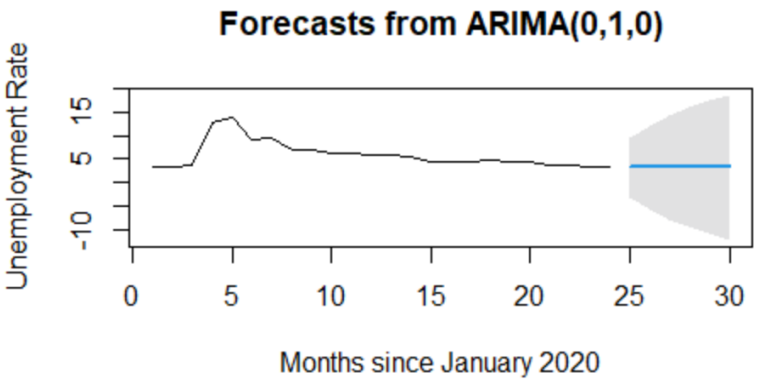


Figure 13: R output from AUTO.ARIMA command for unemployment rate

```
ARIMA(2,1,2) with drift         : Inf
ARIMA(0,1,0) with drift         : 107.6124
ARIMA(1,1,0) with drift         : 110.1991
ARIMA(0,1,1) with drift         : 110.1164
ARIMA(0,1,0)                    : 105.2032
ARIMA(1,1,1) with drift         : Inf

Best model: ARIMA(0,1,0)
```

Figure 14:6-step ahead forecast for unemployment rate



DISCUSSION AND CONCLUSION

There are several recommendations for future work on this topic. The first would be to expand the date range of the analysis to January 2019 through December 2023. This would allow for a greater time period before the onset of Covid-19 and for a greater recovery from its economic effects. An additional recommendation would be to select variables and data sets that are more accessible on a monthly basis. Many of the variables in these analyses were only available yearly which does not give as many data points which can give a deeper insight into events that may trigger statistically significant events. The last recommendation would be to use different cities across the United States instead of just North Carolina. This is because each state handled the Covid-19 shutdowns differently and that may affect the analysis. Applying these suggestions to similar analyses may lead to more interesting results.

In the beginning of this thesis the question was posed: during this time of high unemployment, will the crime rate and drug use also be on the rise? Through both the cross sectional and time series analysis, I believe we can say no to both. The cross-sectional model proved that only population density was significant between January 2020-December 2021. There is a spike in the initial onset of Covid-19 (March 2020), but after that initial spike the level trends back to normal. The time series analysis also supports this claim. In both the time series plots for Opioid overdose ED visits and unemployment they have spikes at the initial onset but then reach an equilibrium. Therefore, no assumption can be made using these analyses that there was an increase in crime rate based on the 12 selected factors during this time period. In conclusion, there seems to be no correlation between drug use and crime rate as suggested in the media.

REFERENCES

1. C. Thorbecke. US unemployment rate skyrockets to 14.7%, the worst since the Great Depression https://abcnews.go.com/Business/us-economy-lost-205-million-jobs- april-unemployment/story?id=70558779
2. Center for Disease Control. Overdose Deaths Accelerating During COVID-19 https://www.cdc.gov/media/releases/2020/p1218-overdose-deaths-covid-19.html
3. D.A. Stork.Violent Crime https://nij.ojp.gov/topics/crimes/violent-crime
4. R. Cooper. Governor Cooper Announces Statewide Stay at Home Order Until April 29 https://www.ncdhhs.gov/news/press-releases/2020/03/27/governor-cooper-announces-statewide-stay-home-order-until-april-29
5. Federal Bureau of Investigation. FBI Urges Vigilance During COVID-19 Pandemic. https://www.fbi.gov/coronavirus
6. American Violence. https://www.americanviolence.org/cities/raleigh?compChartType=differenceChart&compare=none&crimeType=300&customCompareInterval&customTimespan
7. American Violence. Interval&metric=total&precision=monthly&selectedCensusTractsIds&selectedCitiesIds&sortColumn=name&sortPage=0&sortReversed=false&timespan=last12Months
8. NY Databases.com. North Carolina COVID-19 Vaccine Tracker. Https://data.democratandchronicle.com/covid-19-vaccine-tracker/north-carolina/mecklenburg-county/37119/

9. North Carolina Department of Health and Human Services. Opioid and Substance Use Action Plan Data Dashboard. https://www.ncdhhs.gov/opioid-and-substance-use-action -plan-data-dashboard
10. Federal Bureau of Investigation. Variables Affecting Crime. https://ucr.fbi.gov/hate -crime/2011/resources/variables-affecting crime#:~:text=Modes%20of%20transportation %20and%20highway,to%20divorce%20and%20family%20cohesiveness
11. A. M. Costello. CMCS Informational Bulletin. https://www.medicaid.gov/federal -policy-guidance/downloads/cib031821.pdf
12. B. Sessoms. Fayetteville crime down but violent crimes up, much higher than NC overall. https://carolinapublicpress.org/52023/fayetteville-crime-down-but-violent- crimes-up-much-higher-than-nc-overall/
13. Bureau of Labor Statistics. https://data.bls.gov/pdq/SurveyOutputServlet
14. Bureau of Labor Statistics.Labor force data by county, not seasonally adjusted, April 2021-May 2022.https://www.bls.gov/web/metro/laucntycur14.txt
15. World Population Review. North Carolina Population 2022 https://worldpopulationreview. com/us-cities/charlotte-nc-population
16. North Carolina State Bureau of Investigation. SBI Statistics. https://www.ncsbi.gov /Services/SBI-Statistics
17. AreaVibes. NC Crime. https://www.areavibes.com/asheville-nc/crime/
18. Neighborhood Scout. CRIME RATES. https://www.neighborhoodscout.com/ nc/new-bern/crime