Effect of socioeconomic, demographic, political factors and mental health on crime

Rishabh Kandoi
Data Science
University of Rochester
Rochester, United States
rkandoi@ur.rochester.edu

Richa Yadav
Data Science
University of Rochester
Rochester, United States
ryadav3@ur.rochester.edu

***Abstract*— We investigate if a social media setting, public safety workforce, mental health, and political affiliations might give a socio-behavior indication for a crime prediction problem. Traditional crime prediction methods rely on historical crime statistics and demographic information of the region of interest. The proposed idea is that publicly available data from Twitter may contain predictive elements that, unlike the availability of previous crime data for places, can also correlate with changes in crime rates. We analyze the correlation for crime trend prediction with the goal of using Twitter content, public safety workforce%, mental health, and political affiliations. We considered 10 cities from all over the United States and analyzed their correlations with crime statistics for the previous decade. We also report changepoints in the overall crime trends in the past decade and try to correlate that with actual political events or law enforcement that could have caused the change. Overall, the research provides a deep insight into the correlation between our proposed variables, and crime trends.**

Keywords—Twitter data analytics, text mining, sentiment analysis, social trend prediction, change point analysis, crime prediction

I. Introduction

Crime, defined as an illegal act punishable by the governing body, affects not only the victim and their family but also leaves a scar on society for generations. Unfortunately, criminal actions have been prevalent across the globe since the dawn of mankind. Many researchers and organizations have tried to understand criminal psychology and develop crime prediction tools using conventional location-based or demographics-based information [1][2]. Fast forward to today, when each netizen has a stage to speak his/her mind on Internet gives an additional perspective to the above problem. One of the most widely used social media platforms, Twitter, offers social data that is naturally occurring; people seem to be more eager to freely communicate their thoughts, interests, and actions on Twitter. As a result, Twitter offers content that reflects the social behavior of its users. The extraction of behavioral patterns from Twitter has been the subject of numerous studies, including personality recognition, language variances, and crowd behavior to track lifestyles.
In this study, we collect Twitter data based on specific crime-related keywords like theft, violence, etc. for 10 selected cities. The objective of this study is as follows:
a. Does negative sentiment expressed over social media correlate with a high crime rate in that location?
b. If yes, what crime type are people more vocal about on Internet Vs. the actual number of cases reported of that crime in the city?
c. What is the trend pattern of the above 2 scenarios across multiple cities?
d. Do mental health, political affiliations, and the public safety workforce also correlate with criminal activities?
We have also performed a changepoint analysis on the temporal data of crime records to catch any subtle changes in the mean or variance. It determines the number of changes and estimates the time of each change.

II. Data Collection and Transformation

A. *Selection of cities*

We wanted to compare our findings from multiple cities all over the United States of America that represent the demographic and economic expanse that is present in the country. Hence, we considered cities based on their population (as per the 2021 Census data) and their geographical distribution.
The chart below shows the population distribution of the 10 selected cities, as clearly visible, densely populous cities like NYC and sparsely populated cities like New Orleans are both considered in our study.

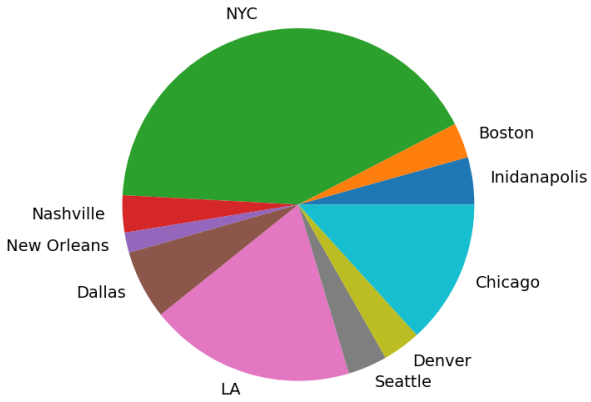

Fig 1.: Population distribution of selected 10 cities.

We have also considered the geographical territory to ensure coverage of most prime regions of the country as follows:

Fig 2.: Geographical distribution of the cities

Using the above approach, we have selected the cities – Boston, New York City, Indianapolis, Nashville, New Orleans, Dallas, Los Angeles, Las Vegas, Denver, Seattle, and Chicago.

B. Twitter Data Collection

Twitter Data is collected using the Tweepy library in python. It is an open-source Python package that gives a very convenient way to access the Twitter API with Python. Tweepy allows location-based tweet extraction using geographical coordinates and hence, we entered the coordinates of the selected 10 cities. To further filter our Twitter data extraction, we used the following keywords –

"crime, theft, violence, gun violence, abscond, harassment, sexual abuse, assault, bribe, burglar, offense, unlawful, forgery, illegal, gambling, violation, homicide, imprisonment, trafficking, extortion, manslaughter, criminal, perjury, robbery, terrorism, terrorist"

Based on the keywords above, detailed information like the tweet ID, user ID, username, tweeted on, actual tweet, number of followers, etc. was collected.

C. Census Data Collection

We used the yearly census data from 2011 to 2021 reported on https://data.census.gov/. The following variables were collected at a city level for the years mentioned above:

a. Gender Ratio - This is the percentage of females in the entire population. Directly collected from https://data.census.gov/table?q=DP05 at a city level.

$$\frac{Number\ of\ females}{Total\ Population}$$

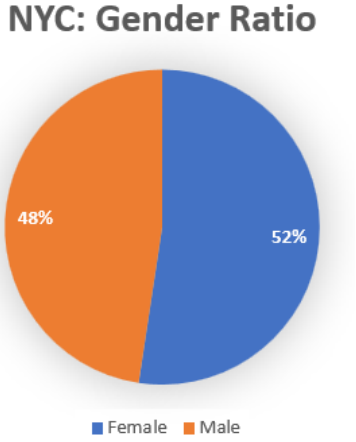

Fig 3.: Average gender ratio in NYC.
(Most of the selected cities follow the same distribution)

b. Median Income Rate - This column signifies the median income of all household income reported each year. Directly collected from https://data.census.gov/table?q=employment&tid=ACSDP1Y2021.DP03&moe=false.

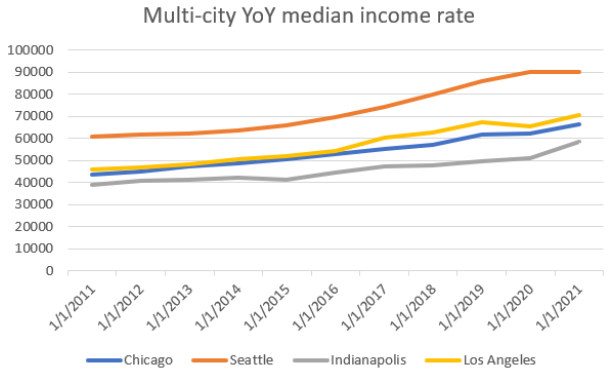

Fig 4.: Median income rate comparison for multiple cities

c. Public Safety Workforce% - This column is the number of people in public administration divided by the total population. Directly collected from https://data.census.gov/table?q=employment&tid=ACSDP1Y2021.DP03&moe=false

$$\frac{\#people\ in\ public\ safety\ workforce}{Total\ Population}$$

d. Unemployment Rate - This column is available at a year-month level. It basically means the #unemployed people over the population. Data collected from (for NYC only) https://www.bls.gov/regions/new-york-new-jersey/data/xg-tables/ro2xglausnyc.htm

$$\frac{\#Unemployed\ People}{Total\ Population}$$

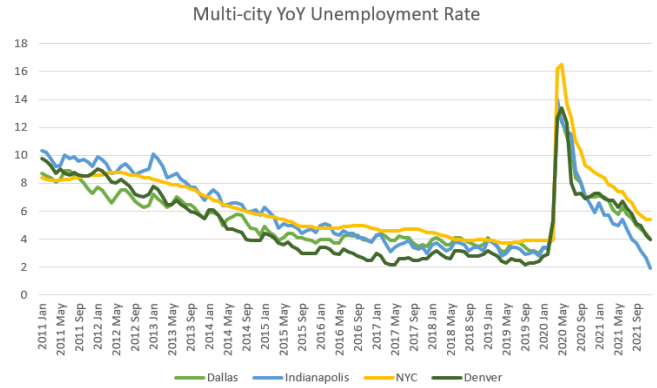

Fig 5.: Unemployment Rate for multiple cities (Sudden increase in Unemployment rate around 2nd Quarter of 2020 due to COVID-19)

D. Presidential election vote%

Presidential votes were collected for that year's Democratic candidate for each of the election years of 2008,2012,2016 and 2020. Data was collected from nytimes.com.

$$\frac{\#votes\ for\ democratic\ candidate}{Total\ Votes}$$

E. *Mental health data*

This data is collected from Google Trends at each city level for the word "depression" from 2011 to 2021. The idea is to capture the popularity of the term "depression" in the given city over the years. The year-month with the highest search count has a value of 100, all the other year-months are relatively ranked on the basis of this. Data collected from *https://trends.google.com/trends/*

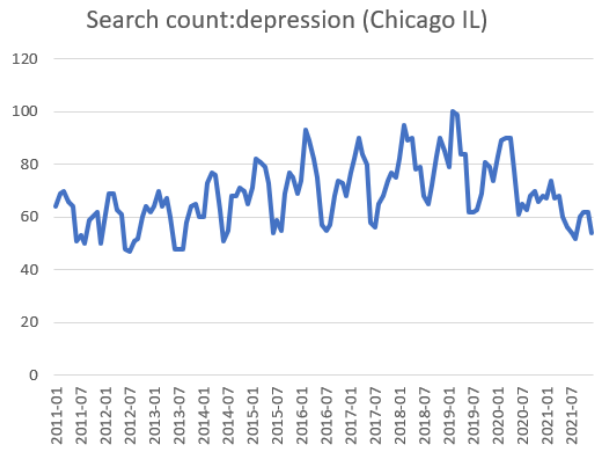

Fig 6.: Google trend for 'depression' for Chicago city

F. *Crime Count data*

This data is collected from the municipal open data portal created by several cities. Dozens of US cities have fairly comprehensive open data portals, with information on varied types of government activity like police or property records. Data collected from *https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-Present/ijzp-q8t2/data* (for Chicago only).

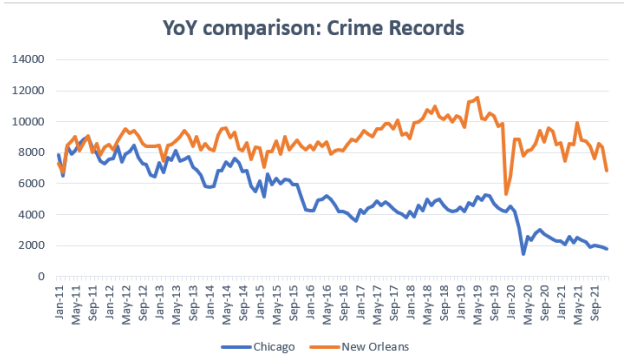

Fig 7.: Crime trend for Chicago and New Orleans
(The crime trend in Chicago is decreasing YoY while the trend for New Orleans is rather increasing over time. This corroborates the article by WSJ.com which calls New Orleans the No.1 city in the USA for homicides [4])

III. Modelling Techniques

The main idea of this study is to collect enough evidence that additional variables do give a better understanding of the crime trend. Since most of the variables in the dataset are of numerical type, we chose to perform a regression analysis to derive a conclusion. The further section discusses all our methods and techniques.

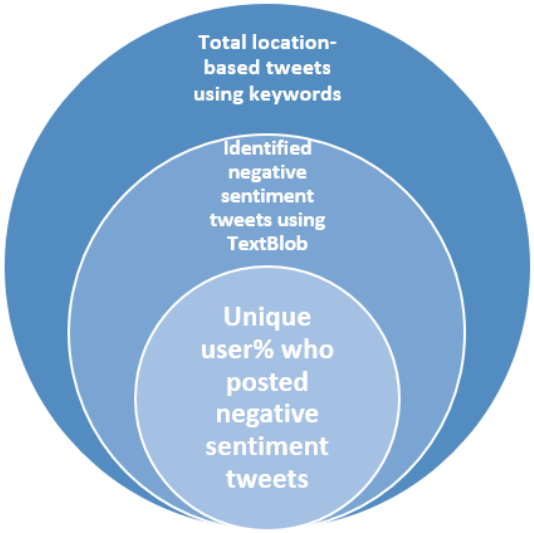

Fig 8.: Calculation of Unique User% with negative tweets

A. *Sentiment Analysis for collected tweets*

As discussed in the Data Collection section, keyword-specific, location-based tweets were collected for the selected 10 cities for the years 2011 – 2021. To get the truly negative tweets, we used TextBlob's sentiment analysis library. The function returns a named tuple of the form Sentiment (polarity, subjectivity). The polarity score is a float within the range [-1.0, 1.0]. Any tweet that had a polarity of less than 0 was flagged as a negative tweet. Based on the set of negative tweets filtered by TextBlob and the original pool of tweets, unique users were identified. To summarize,

1) Collected keyword-specific, location-based tweets for each year-month.

2) Identified the number of unique users in the above pool for each year-month (total_unique_users).

3) To ensure that before sentiment analysis, a tweet is "neutral" certain keywords like 'crime', 'criminal', 'violence', etc were omitted from the tweet. For example, consider a tweet – "the show criminal minds is crazy". It would be incorrect if this tweet gets flagged as negative because of the word criminal. Hence, it's best to omit such words before performing sentiment analysis.

4) Performed sentiment analysis and filtered truly negative tweets using TextBlob with polarity < 0.

5) Based on the above-filtered tweets, aggregated the number of unique users. (negative_unique_users).

6) For each year-month,

$$PercentNegativeUsers = \frac{negative\_unique\_users}{total\_unique\_users}$$

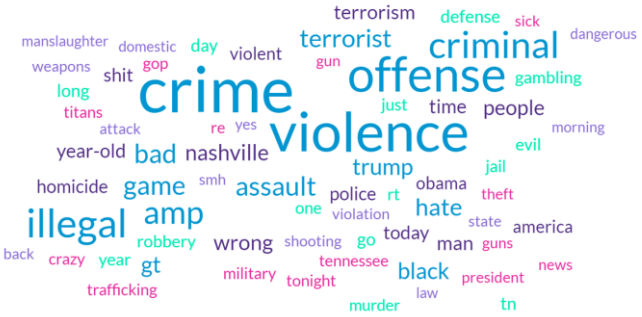

Fig 9.: Nashville Word Cloud – negative sentiments captured after sentiment analysis

B. *Regression Techniques*

The power of regression analysis is that it is simple, easy to interpret, and yet a very robust statistical method to find relationships between numerical variables. For our study, we implemented multiple linear regression and polynomial regression techniques.

1) *Multiple Linear Regression*

	City-specific data.org portals	Census Data				Sentiment analysis of tweets	Nytimes.com	City-level Google Trend Data
Year-Month	Population normalized crime%	Female% in population	Unemployment rate	Median Income rate	Public safety workforce%	Unique User% with negative tweets	%Presidential votes for Democrats	Search Count for Depression
11-Jan	0.02265	0.515	8.9	37,325	0.058	0.3478	0.79	54
11-Feb	0.02098	0.515	7.8	37,325	0.058	0.2813	0.79	69
11-Mar	0.02641	0.515	7.8	37,325	0.058	0.3019	0.79	53
11-Apr	0.02728	0.515	7.4	37,325	0.058	0.2342	0.79	69
11-May	0.02814	0.515	7.9	37,325	0.058	0.2823	0.79	44
11-Jun	0.02530	0.515	9.2	37,325	0.058	0.2581	0.79	63
11-Jul	0.02693	0.515	9.2	37,325	0.058	0.0000	0.79	50
11-Aug	0.02832	0.515	8.8	37,325	0.058	0.0000	0.79	57

Fig 10.: Example of the input dataset

Multiple linear regression is a regression analysis with more than one independent variable, meaning that we try to predict a value based on two or more variables. For our analysis, the target variable is the population-normalized crime%. The independent variables are:

- Female% in population [1]
- Unemployment rate [1]
- Median Income rate [1]
- Public safety workforce% *(New variable)*
- Unique User% with negative tweets *(New variable)*
- %Presidential votes for Democrats *(New variable)*
- Search Count for Depression *(New variable)*

As the data from the last 10 years, there is a lot of variability in the data due to many factors like elections, covid-19, etc. To ensure that our regression analysis doesn't suffer from a class imbalance problem which simply means unequal distribution of input training data, we used a 10-fold cross-validation technique to compute the average error rates.

2) *Polynomial Regression*

As a specific case of multiple linear regression, polynomial regression is a type of linear regression that estimates the correlation as an nth-degree polynomial. As many of our critical variables don't show a particularly linear relationship, we had to try polynomial regression.

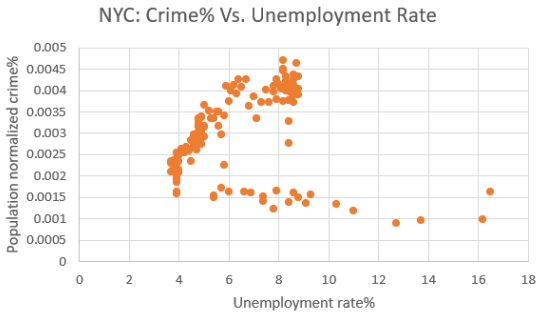

Fig 11.: NYC Crime% Vs. the unemployment rate shows a non-linear data relationship

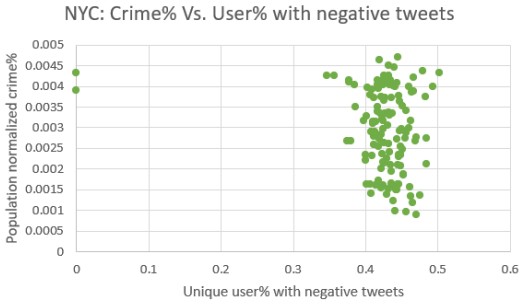

Fig 12.: NYC Crime% Vs. Unique user% show a non-linear data relationship.

The key distinction in the application of a Polynomial regression is that the user needs to specify the degree of polynomial to be considered.

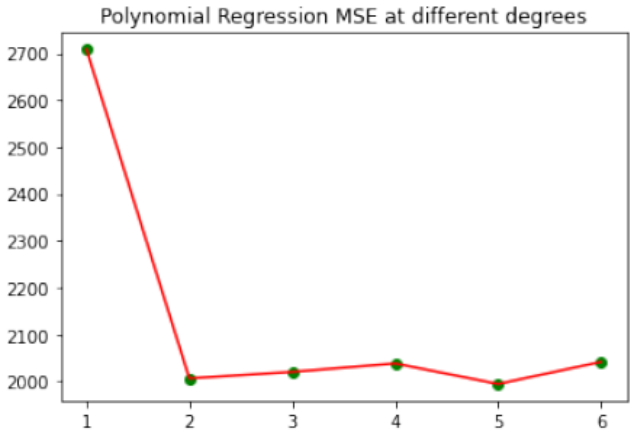

Fig 13.: Mean Squared Error for different degrees of polynomial

For most of the cities, we found that a polynomial degree of 2 gave the best fit and the least error.

```
Polynomial Model (degree=2) summary:
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.903
Model:                            OLS   Adj. R-squared:                  0.897
Method:                 Least Squares   F-statistic:                     143.7
Date:                Fri, 25 Nov 2022   Prob (F-statistic):           1.23e-58
Time:                        12:32:58   Log-Likelihood:                -1223.0
No. Observations:                 132   AIC:                             2464.
Df Residuals:                     123   BIC:                             2490.
Df Model:                           8
Covariance Type:            nonrobust
==============================================================================
                                 coef    std err          t      P>|t|
[0.025      0.975]
------------------------------------------------------------------------------
const                          3.47e+06   1.97e+06      1.761      0.081
-4.3e+05    7.37e+06
Gender_ratio                  -3.875e+06   2.24e+06     -1.733      0.086
-8.3e+06    5.51e+05
UnemploymentRate               1.712e+05   1.28e+05      1.336      0.184
-8.25e+04    4.25e+05
MedianIncomeRate               -88.4198     35.878     -2.464      0.015
-159.438     -17.402
Gender_ratio^2                -5.072e+06   2.92e+06     -1.738      0.085
-1.08e+07    7.03e+05
Gender_ratio UnemploymentRate -3.158e+05   2.44e+05     -1.292      0.199
-7.99e+05    1.68e+05
Gender_ratio MedianIncomeRate  168.1235     68.286      2.462      0.015
32.956      303.290
UnemploymentRate^2             82.3844     38.491      2.140      0.034
6.195      158.574
UnemploymentRate MedianIncomeRate  -0.1256      0.041     -3.045      0.003
-0.207      -0.044
MedianIncomeRate^2            8.477e-07   1.45e-05      0.058      0.953
-2.78e-05    2.95e-05
==============================================================================
```

Fig 14.: NYC polynomial regression output

As seen from the output of polynomial regression on NYC, we were not able to explain the meaning of terms like those highlighted above. For example, the term *Gender_ratio * MedianIncomeRate* has a very low p-value and can be considered an important interaction term in deciding the regression equation; but there is no logical way to explain this terminology. Hence, despite giving impeccable results due to a lack of interpretation, we had to discard this approach.

C. *Changepoint analysis*

A change-point analysis is run on a set of time-ordered data to see if any changes have taken place. It counts the changes and calculates how long each modification will take. Additionally, it offers confidence intervals for the timing of each shift as well as levels of confidence for each change. We intend to perform a changepoint analysis on our crime trend time series to understand subtle changes and try to correlate the same with certain political or social events that might have caused the sudden change in the trend. We used the Pruned Exact Linear Time (PELT) algorithm [5] for changepoint detection for this use case.

IV. Results and Inferences

In this section, we want to talk about our findings and the conclusions that we gathered.

A. *Multiple Linear Regression*

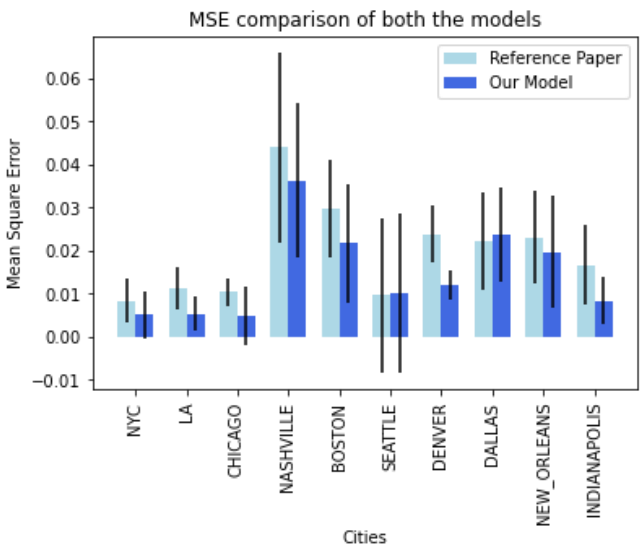

Fig 15.: MSE report linear regression

The above chart reports the average Mean Squared Error (MSE) of the linear regression for old variables Vs. the old + additional variables. The vertical bars represent the average MSE value while the vertical black lines represent the standard deviation of MSE across the 10-fold cross-validation sets. We can see that our proposed model with additional attributes performs better than the reference paper in all cities.

City	Female%	Unemployment rate	Median Income rate	Public safety workforce%	Unique User% with negative tweets	%Presidential votes for Democrats	Search Count for Depression
Boston	-5.01E+04	-169.9182	-0.0713	-1.21E+05	-1261.5069	2.93E+04	-19.6378
New York City	2.16E+05	-632.3858	-1.0918	-2.84E+05	-7403.9871	1.99E+04	38.5746
Indianapolis	7.78E+04	-131.8578	-0.1373	5.71E+04	673.3547	7603.6165	-29.1779
Nashville	-3.53E+05	-16.8131	-0.0308	3.78E+04	3165.423	-7553.8812	-32.3356
New Orleans	2.98E+05	-27.2478	-0.1851	3.16E+05	186.616	-1.06E+04	0.1052
Dallas	-1267.7038	-8.0127	-0.0313	8784.844	-27.3213	650.9958	0.3175
Los Angeles	-1.81E+05	-99.6388	-0.3244	3474.7768	-194.7272	-4722.3648	4.4887
Denver	8.01E+05	232.6884	0.4928	1.39E+05	-1567.8939	-1.42E+05	-36.2239
Seattle	-3.78E+05	116.675	0.041	2.39E+05	-986.65	-1.35E+04	-0.3506
Chicago	-1.63E+05	-58.3957	-0.2474	-1.79E+05	1570.1766	4.80E+04	18.4053

Fig 16.: Correlation coefficients after multiple linear regression

The table above shows the values of all correlation coefficients across all the 10 selected cities using multiple linear regression. The cells with red font indicate that for that city, the variable has a significant p-value (less than 0.05).

B. *City-specific Inferences*

1) *Effect of political affiliations on crime in Chicago and Denver*

Post-training the regression model for the city of Chicago, we found a clear positive relationship between the votes for the Democratic presidential candidate and crime. This relationship is also clearly visible in the graph below.

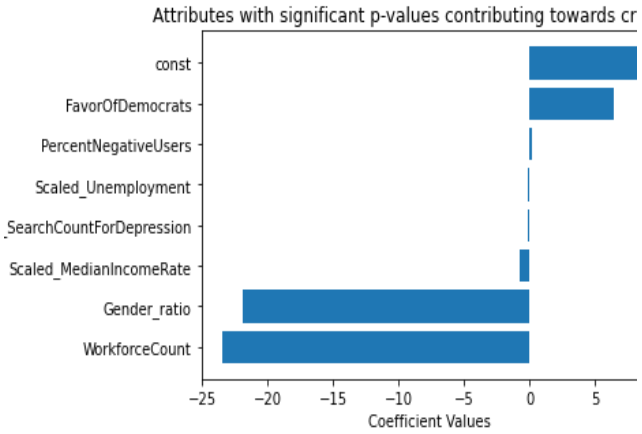

Fig 17.: Correlation coefficients for Chicago (FavorofDemocrats shows a positive coefficient)

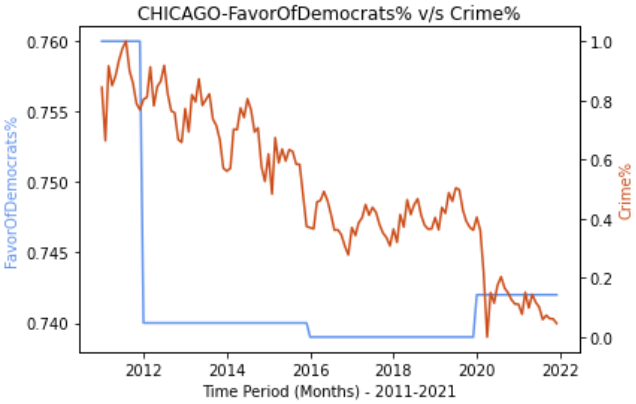

Fig 18.: Relationship between Crime and political affiliations - Chicago

Given this relationship, an opposite relation is seen in a smaller city like Denver. The below chart shows a negative correlation coefficient for the favor of the democrats variable. A negative correlation would in fact mean the crime rate is more when there is more population in favor of Republicans. This is evident from the p-value graph as well as the time-series graph depicted here.

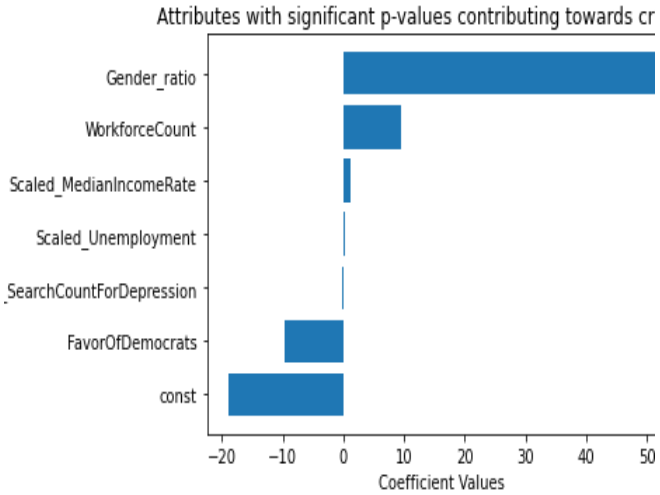

Fig 19.: Correlation coefficients for Denver

(FavorofDemocrats shows a negative coefficient)

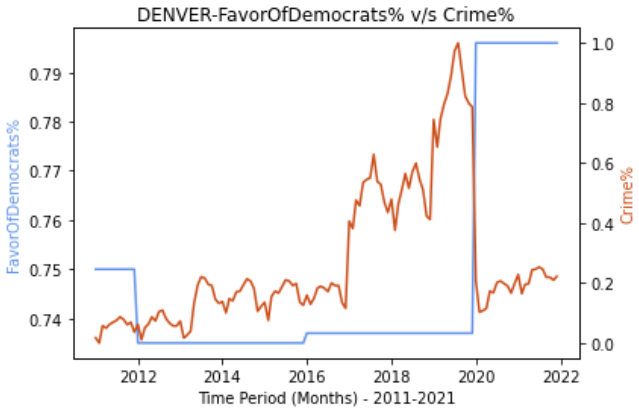

Fig 20.: Relationship between Crime and political affiliations - Denver

This information also correlates with the fact that generally larger cities like Chicago tend to be Democratic in nature while smaller cities tend to be more Republican.

2) Effect of public safety workforce on the crime rate in New Orleans

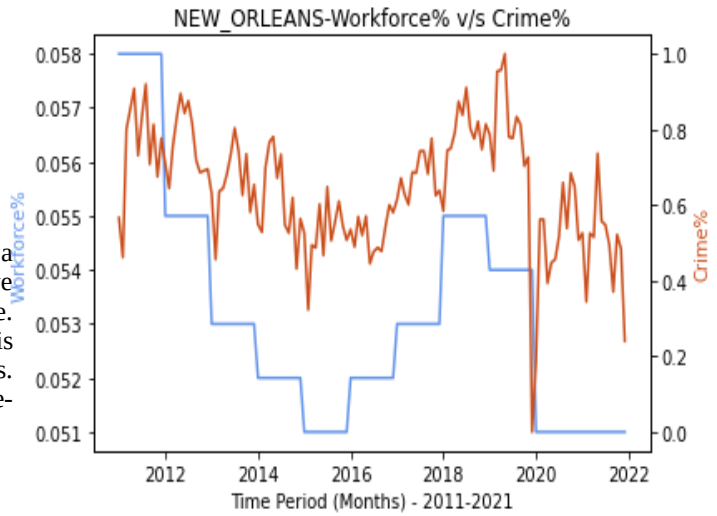

Fig 21.: Relationship between Crime and public safety administration – New Orleans

Regarding the percentage of the population employed in public safety, the regression findings revealed that, for the most part in all places, growing the workforce was associated with a decline in crime rates, which is the ideal outcome. However, as we all know, the perfect world does not exist. There are anomalies, such as the city of New Orleans, which exhibits the exact opposite tendency from what is normal. There are just too many additional variables contributing to such absurd criminal activity in this city in particular, which consistently lists as one of the highest crimes committed in all US cities [4].

3) Effect of negative tweet sentiment

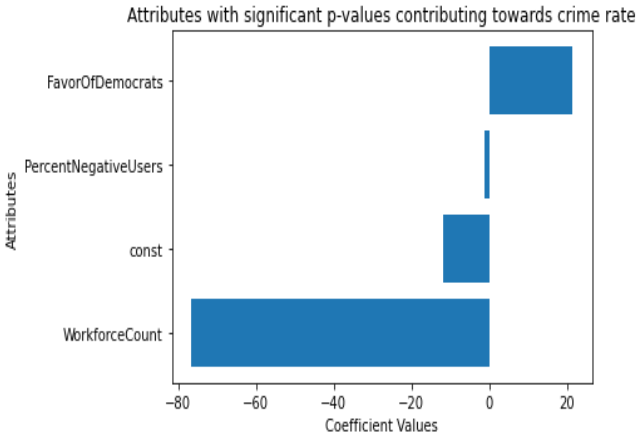

Fig 22.: Relationship between Crime and Negative tweet sentiment – NYC

(Lesser people tweeting negative sentiment)

Finally, we wanted to discuss some inferences on the Twitter negative sentiment variable and how the number of negative tweets posted by users is affected by the crime rate. We discovered that in New York, despite an increase in crime rate, fewer people are posting tweets with negative sentiments about the same, clearly visible from the negative correlation coefficient.

While it is displaying a positive response from users in Denver as shown in the chart below, more people are tweeting about an increase in crime rates or the opposite is also true.

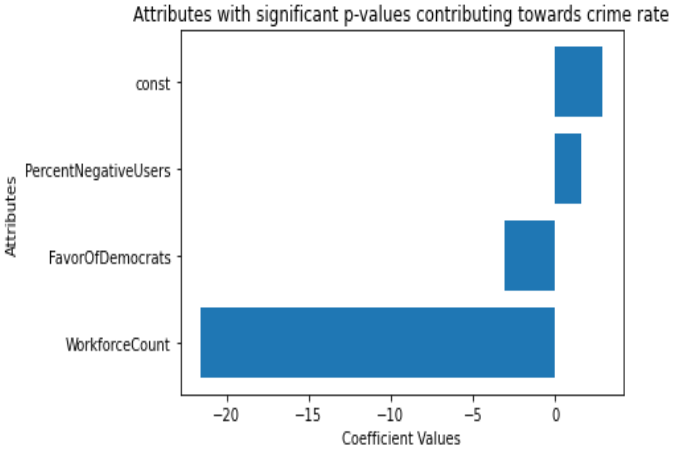

Fig 23.: Relationship between Crime and Negative tweet sentiment – Denver

(More tweeting about the increase in the crime rate or vice versa)

C. *Changepoint analysis insights*

Using the PELT changepoint analysis approach [5], we discovered the following interesting insight for the city of Boston.

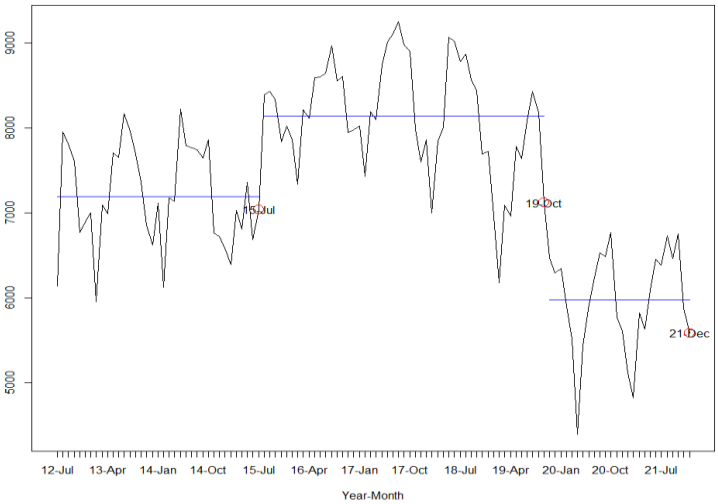

Fig 24.: Changepoint analysis for Boston Crime Trend

We are able to justify the 2019-Oct changepoint as the government of Massachusetts passed a historic criminal justice reform bill in April 2019 that significantly reduced the severity of numerous low-level offenses and raised the minimum age of criminal responsibility from 7 to 12, resulting in a drop in crime. This is also clear from the graph below, which demonstrates that crime rates actually decline when more people support the Democrats who passed this legislation.

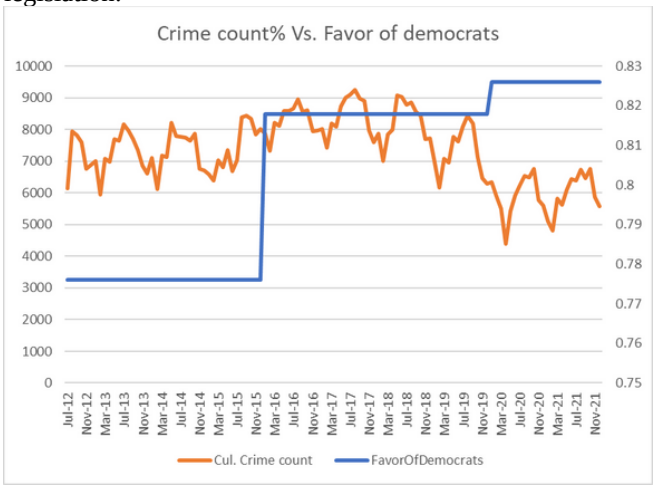

Fig 25.: the relationship between Crime and political affiliations for Boston

V. Conclusion

Prediction of crime rate is a critical requirement, extensive research has been conducted in the past, but still leaving it as an open-ended problem. Although the traditional approach[1] did well by establishing the dependency of Female% in the population, Unemployment rate, and Median Income rate towards crime rate prediction, we saw that our model with Public safety workforce%, Unique User% with negative tweets, %Presidential votes for Democrats, Search Count for Depression as additional attributes performed far better in terms of MSE, and gave some interesting insights specific to multiple cities as already discussed. There is yet deeper

analysis possible, like building a model to accommodate data across cities. We can also increase the granularity of the crime rate to distinguish the effect of social media response for violent crimes or non-violent cases and find out the type of crime most talked about.

REFERENCES

[1] Ajimotokin, S., Haskins, A. and Wade, Z., 2015. The effects of unemployment on crime rates in the US.

[2] Towers, S., Chen, S., Malik, A. and Ebert, D., 2018. Factors influencing temporal patterns in crime in a large American city: A predictive analytics perspective

[3] South, S.J. and Messner, S.F., 2000. Crime and demography: Multiple linkages, reciprocal relations. Annual Review of Sociology, pp.83-106.

[4] https://www.wsj.com/articles/new-orleans-murder-rate-crime-11663338008

[5] Killick, R., Fearnhead, P. and Eckley, I.A.∗, 2012. Optimal detection of changepoints with a linear computational cost