CRIME DATA ANALYSIS AND PREDICTION AT CHICAGO

Bhemeshwar Yeturi

Abstract— *This project delves into the comprehensive analysis of crime and weather datasets sourced from the city of Chicago. Our investigation involves a thorough examination of crime data, uncovering temporal patterns and trends. Utilizing advanced time series analysis techniques, we model and gain insights into the temporal aspects of the crime dataset. The primary objective of our endeavor is to predict future crime occurrences through the implementation of machine learning models. These models, designed to forecast crime patterns based on various factors, contribute significantly to a heightened understanding of criminal activities. The aim is to leverage these insights for potential preventive measures, thereby enhancing public safety within the city.*

INTRODUCTION

Urban areas, as dynamic ecosystems, present unique challenges in understanding and addressing criminal activities. The city of Chicago, with its diverse neighborhoods and complex socio-economic landscape, serves as a compelling backdrop for our project. In this endeavor, we undertake a comprehensive exploration of crime and weather datasets to unravel the intricate patterns and factors influencing criminal occurrences within the city.

The impetus for this analysis stems from the critical need to enhance public safety by gaining a nuanced understanding of crime dynamics. Criminal activities, often influenced by a multitude of variables, exhibit patterns that extend beyond mere spatial and temporal correlations. Recognizing this complexity, our project integrates data-driven methodologies to not only retrospectively analyze crime trends but also proactively predict future occurrences.

Research Problem:

The city of Chicago faces persistent challenges related to crime prevention and management. Traditional approaches often fall short in providing timely insights and actionable strategies. Hence, our research problem centers on developing a predictive framework that leverages the synergy between crime and weather data. By harnessing the power of machine learning and time series analysis, we seek to forecast crime patterns and contribute to a more effective and targeted approach to crime prevention.

Significance:

This project holds significance not only in advancing the field of data-driven crime analysis but also in its potential to impact real-world urban safety measures. By combining rigorous data analysis with predictive modeling, our approach aims to empower decision-makers with actionable insights for crime prevention strategies.

As we delve into the intricate interplay of factors influencing crime in Chicago, this project seeks to provide a roadmap for informed policy decisions, resource allocation, and community engagement. Through a proactive lens, our exploration aims to contribute to the broader discourse on urban safety, demonstrating the potential of data science in fostering safer and more secure urban environments.

In the subsequent sections, we will detail our data analysis methodologies, findings, and the implications of our predictive models for the city's safety and well-being.

DATA COLLECTION

Sources of Data:

Our data collection process involved acquiring two crucial datasets—crime data and weather data—from reliable sources to ensure the accuracy and comprehensiveness of our analysis.

Crime Data:

Obtained from the https://data.cityofchicago.org/Public-Safety/City-of-Chicago-Crime-Data/v9q9-3dm2 , this dataset encompasses a wide range of criminal incidents reported within the city of Chicago. The dataset includes information on the type of crime, location, date, and other relevant details.

Weather Data:

Sourced from Weather Data Link , the weather dataset comprises meteorological information for the corresponding time period. Variables such as temperature, precipitation, wind speed, and atmospheric conditions were included to explore potential correlations with crime occurrences.

DATA DESCRIPTION

CRIME DATA SET

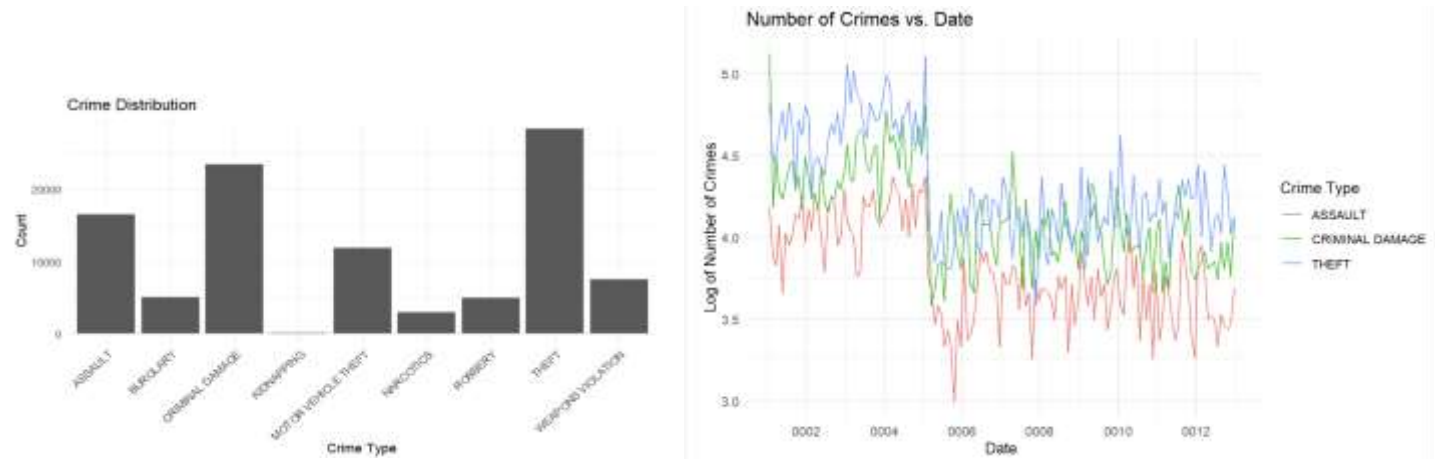
```
       ID            Case.Number            Date              Block              IUCR          Primary.Type
 Min.   :12258517   Length:100562      Length:100562      Length:100562      Length:100562      Length:100562
 1st Qu.:12370192   Class :character   Class :character   Class :character   Class :character   Class :character
 Median :12479306   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character
 Mean   :12477579
 3rd Qu.:12582888
 Max.   :13275887

  Description        Location.Description    Arrest           Domestic             Beat           District        Ward
 Length:100562      Length:100562      Length:100562      Length:100562      Min.   : 111   Min.   : 1.00   Min.   : 1.00
 Class :character   Class :character   Class :character   Class :character   1st Qu.: 532   1st Qu.: 5.00   1st Qu.: 9.00
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Median :1011   Median :10.00   Median :21.00
                                                                             Mean   :1112   Mean   :10.89   Mean   :21.94
                                                                             3rd Qu.:1622   3rd Qu.:16.00   3rd Qu.:32.00
                                                                             Max.   :2535   Max.   :31.00   Max.   :50.00
                                                                                                            NA's   :3
 Community.Area   FBI.Code          X.Coordinate      Y.Coordinate        Year      Updated.On          Latitude
 Min.   : 1.0   Length:100562      Min.   :1095509   Min.   :1813909   Min.   :2021   Length:100562      Min.   :41.64
 1st Qu.:24.0   Class :character   1st Qu.:1153778   1st Qu.:1856298   1st Qu.:2021   Class :character   1st Qu.:41.76
 Median :37.0   Mode  :character   Median :1167233   Median :1883710   Median :2021   Mode  :character   Median :41.84
 Mean   :38.6                      Mean   :1165896   Mean   :1882922   Mean   :2021                      Mean   :41.83
 3rd Qu.:58.0                      3rd Qu.:1177531   3rd Qu.:1906954   3rd Qu.:2021                      3rd Qu.:41.90
 Max.   :77.0                      Max.   :1205119   Max.   :1951493   Max.   :2022                      Max.   :42.02
                                   NA's   :1516      NA's   :1516                                        NA's   :1516
   Longitude        Location             Time
 Min.   :-87.92   Length:100562      Length:100562
 1st Qu.:-87.71   Class :character   Class :character
 Median :-87.66   Mode  :character   Mode  :character
 Mean   :-87.67
 3rd Qu.:-87.62
 Max.   :-87.52
 NA's   :1516
```

WEATHER DATA SET

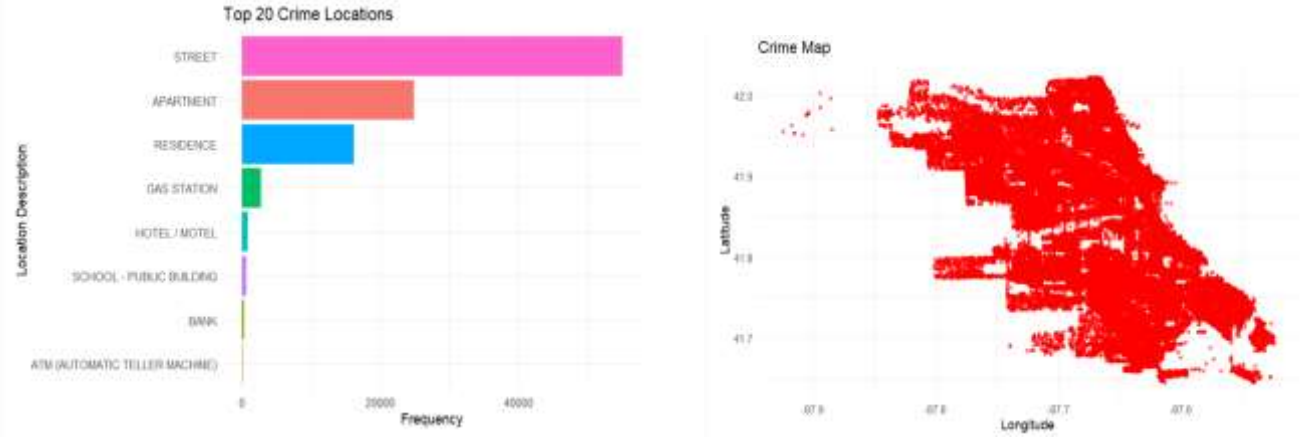
	name	Date	temp	dew	humidity	precip	snow
	<chr>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	Chicago,United States	01-01-2021	-0.7	-3.2	82.8	7.045	0.0
2	Chicago,United States	02-01-2021	0.5	-1.7	85.2	0.000	1.1
3	Chicago,United States	03-01-2021	-0.2	-2.4	85.4	1.054	0.2
4	Chicago,United States	04-01-2021	-2.6	-4.3	88.3	0.000	0.0
5	Chicago,United States	05-01-2021	0.2	-2.4	82.4	0.000	0.0
6	Chicago,United States	06-01-2021	1.3	-2.3	77.3	0.000	0.0
7	Chicago,United States	07-01-2021	2.6	-3.8	63.0	0.000	0.0
8	Chicago,United States	08-01-2021	1.0	-3.8	70.3	0.267	0.0
9	Chicago,United States	09-01-2021	0.0	-5.4	67.4	0.000	0.0
10	Chicago,United States	10-01-2021	-2.0	-6.4	72.1	0.000	0.0

Exploratory Data Analysis (EDA):

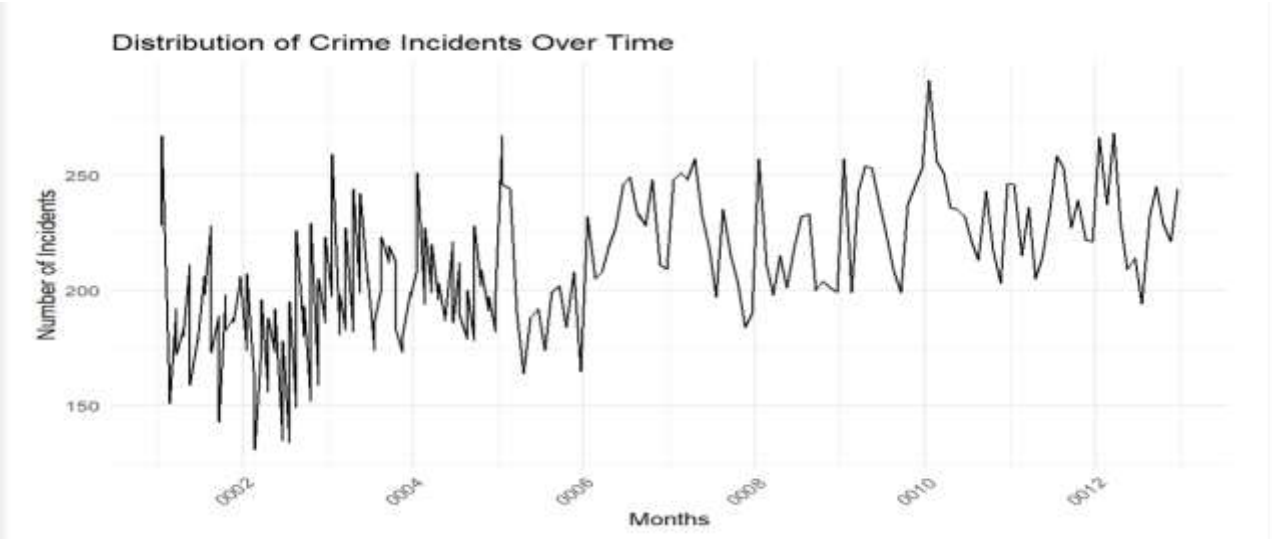


The first graph titled "Crime Distribution" is a bar chart that represents the count of different types of crimes. Here's the breakdown:
- Theft is the most common crime, with its count significantly higher than any other crime type.
- Criminal Damage and Assault follow as the second and third most common offenses, respectively.
- Narcotics, Burglary, and Motor Vehicle Theft are moderately represented.

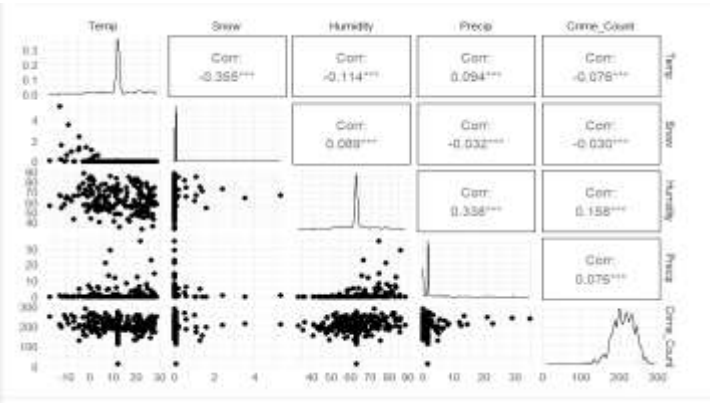
- Robbery, Weapons Violation, and Kidnapping have the lowest counts among the listed crime types.
This distribution indicates a prevalence of property crimes (like theft and burglary) over violent crimes (such as assault and robbery) in the dataset.
The image shows a graph of crime distribution by crime type. The most common type of crime is theft, followed by assault, burglary, and criminal damage. Other types of crime include kidnapping, motor vehicle theft, narcotics, robbery, and weapons violation.

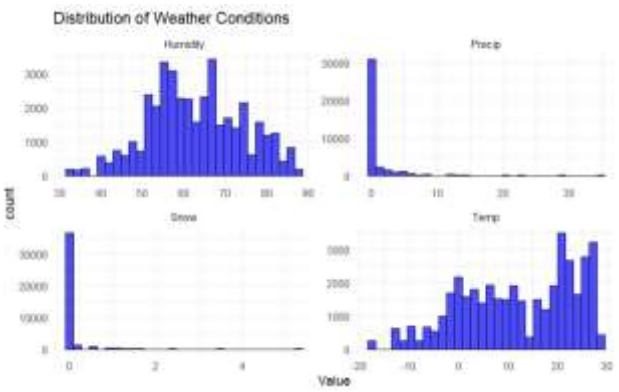


The image shows the top 20 crime locations in the Chicago, based on frequency. The most common crime locations are streets, apartments, and gas stations. Other common crime locations include schools, public buildings, ATMs, banks, hotels/motels, and residences.
The image suggests that crime is more common in public places than in private residences. It also suggests that certain types of businesses, such as gas stations and convenience stores, are more likely to be targeted by criminals.
The image is a useful reminder to be aware of your surroundings and to take precautions to protect yourself and your property, especially when in public places.



CORRELATION SCATTER PLOT MATRIX





The provided analysis focuses on a correlation matrix, a statistical tool used to explore relationships between variables. The matrix includes histograms and scatter plots for each variable, displaying correlations using Pearson coefficients. The findings indicate statistically significant but weak correlations between weather conditions and crime counts.

Temperature exhibits a slight negative correlation (-0.076) with crime, suggesting a minor decrease in crime with higher temperatures. Similarly, snow shows a weak negative correlation (-0.030), implying a slight decrease in crime with more snow. Positive correlations are observed between humidity and snow (0.089) and between humidity and crime (0.158). Precipitation has a positive but weak correlation (0.076) with crime, indicating a slight increase with more precipitation.

The second graph illustrates the distribution of weather conditions. Humidity is evenly distributed, while precipitation and snow are skewed toward zero, suggesting infrequent occurrences. Temperature displays a bimodal distribution, hinting at seasonal variations.

In conclusion, the statistically significant but weak correlations suggest a nuanced relationship between weather and crime, with other unexplored factors likely influencing crime rates more substantially. Further statistical analysis is needed to draw definitive conclusions. The histograms provide valuable insights into the distribution of weather conditions, potentially aiding in understanding correlations with crime patterns.

MODELLING:
Model 1: Linear Regression Model

Introduction:
This report presents a regression analysis aimed at predicting Crime Count using weather-related variables, including Temperature (Temp), Snowfall (Snow), Humidity, and Precipitation (Precip). The dataset comprises 36,386 rows and 27 columns, with variables such as Case Number, Date, and various crime-related and weather-related attributes.

Methodology:
A multiple linear regression model was employed with Crime Count as the dependent variable and Temp, Snow, Humidity, and Precip as independent variables. The resulting model is expressed as:

$$\textit{Crime Count} = 183.32 - (0.34 \times \textit{Temp}) - (6.17 \times \textit{Snow}) + (0.55 \times \textit{Humidity}) + (0.33 \times \textit{Precip}) + \varepsilon$$

Key Findings:
1. Significant Predictors: All independent variables are statistically significant predictors of Crime Count ($p < 0.001$).
2. Coefficient Interpretation: Each one-unit change in Temp, Snow, Humidity, and Precip corresponds to changes in Crime Count as indicated by their respective coefficients.
3. Model Fit: The model explains approximately 7.7% of the variance in Crime Count(R^2 = 0.077), indicating a moderate level of predictive power.
4. Overall Model Significance: The F-statistic(673.5)and its associated p-value(<2.2times 10^{-16}) suggest that the model is statistically significant.

Residual Analysis:
The residuals exhibit a mean close to zero, indicating that the model is unbiased. The residual standard error is 28.02, representing the typical difference between observed and predicted Crime Count. Residuals are approximately normally distributed.

Summary and Final Plot

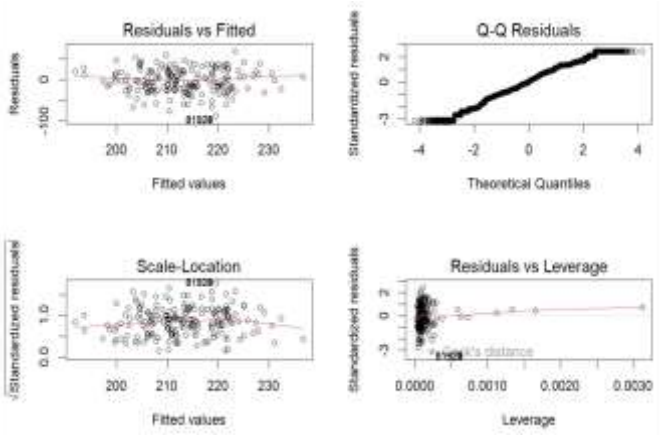
```
Call:
lm(formula = Crime_Count ~ Temp + Snow + Humidity + Precip, data = dataTrain)

Residuals:
    Min      1Q  Median      3Q     Max
-88.754 -19.698   0.594  22.156  67.573

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 183.32470    0.96305 190.359   <2e-16 ***
Temp         -0.34415    0.01485 -23.179   <2e-16 ***
Snow         -6.16631    0.31829 -19.373   <2e-16 ***
Humidity      0.54666    0.01478  36.986   <2e-16 ***
Precip        0.33276    0.03580   9.295   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 28.02 on 32306 degrees of freedom
Multiple R-squared:  0.07697,   Adjusted R-squared:  0.07686
F-statistic: 673.5 on 4 and 32306 DF,  p-value: < 2.2e-16

[1] "Mean Squared Error: 774.6132138036111"
[1] "R-squared: 0.0769697376046167"
```



The regression model summary demonstrates that weather conditions significantly contribute to predicting Crime Count. We can also see that the independent variables are statistically significant in the model indicated by (*). From the summary, we can see that temperature and snow variables are negatively correlated with the crime count while humidity and precipitation are positively correlated with crime count. We got the same results which is evident in the EDA section while performing correlation plots.

Model 2: Random Forest Regression Analysis Report-Predicting Crime Count Based on Weather Conditions

Introduction:
This report presents the results of a Random Forest Regression analysis conducted to predict Crime Count using weather-related variables, including Temperature (Temp), Snowfall (Snow), Humidity, and Precipitation (Precip). The dataset consists of 36,386 rows and 27 columns, encompassing crime-related and weather-related attributes.

Methodology:
The Random Forest Regression model was implemented using the `randomForest` function with Crime Count as the response variable and Temp, Snow, Humidity, and Precip as predictor variables. The model was trained on a subset of the data, and its predictive performance was evaluated.

rf_model <- randomForest(Crime_Count ~Temp + Snow + Humidity + Precip , data = trainingSet, ntree = 100)

Key Findings:
1. Mean Squared Error (MSE): The model's mean squared error is 331.80, indicating the average squared difference between observed and predicted Crime Count. A lower MSE suggests better predictive accuracy.
2. R-squared (R^2): The R-squared value of 0.67 suggests that approximately 67.4% of the variance in Crime Count is explained by the model. This indicates a substantial improvement over the linear regression model, emphasizing the effectiveness of the Random Forest approach.

```
[1] "Mean Squared Error: 331.795424181371"
[1] "R-squared: 0.67447459774457"
```

The Random Forest Regression model demonstrates promising predictive performance, as evidenced by its relatively low mean squared error and substantial R-squared value.

Comparison Report: Linear Regression vs. Random Forest Regression models

Model Predictions:
Below are the actual and predicted values for the two models:

	actual_values	pred_linearmodel	pred_Randomforestmodel
1	228	231.1735	228.6600
2	154	214.1102	157.3751
3	211	278.2811	228.2450
4	210	267.8618	212.4961
5	211	269.9355	217.6146

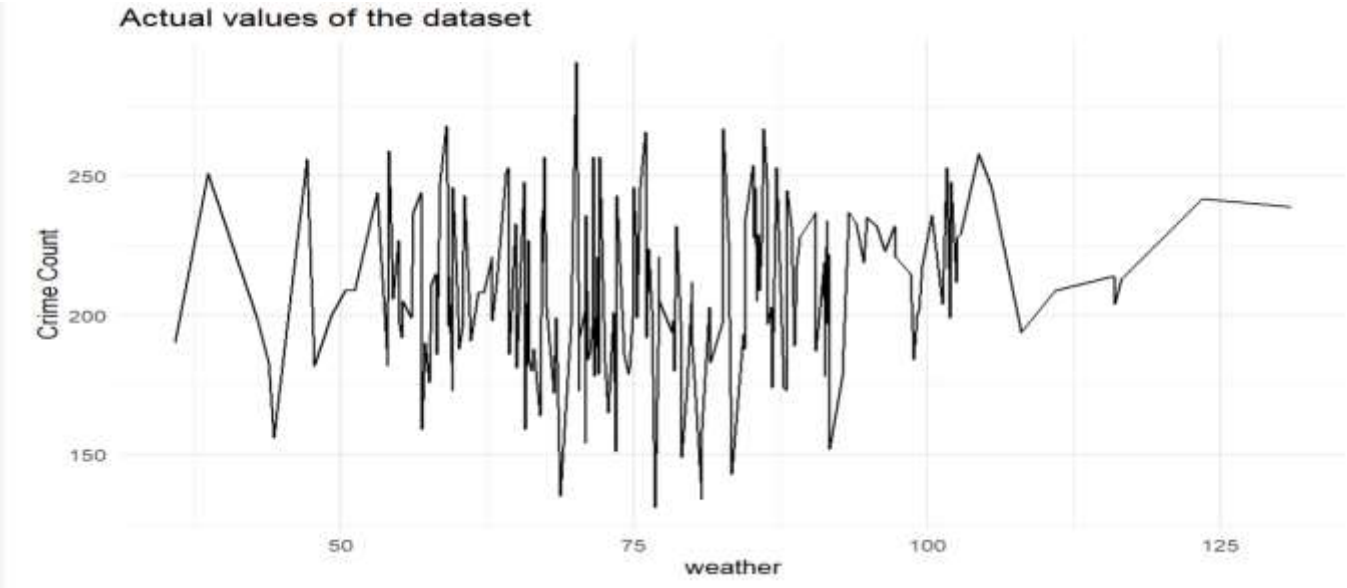
Conclusion:

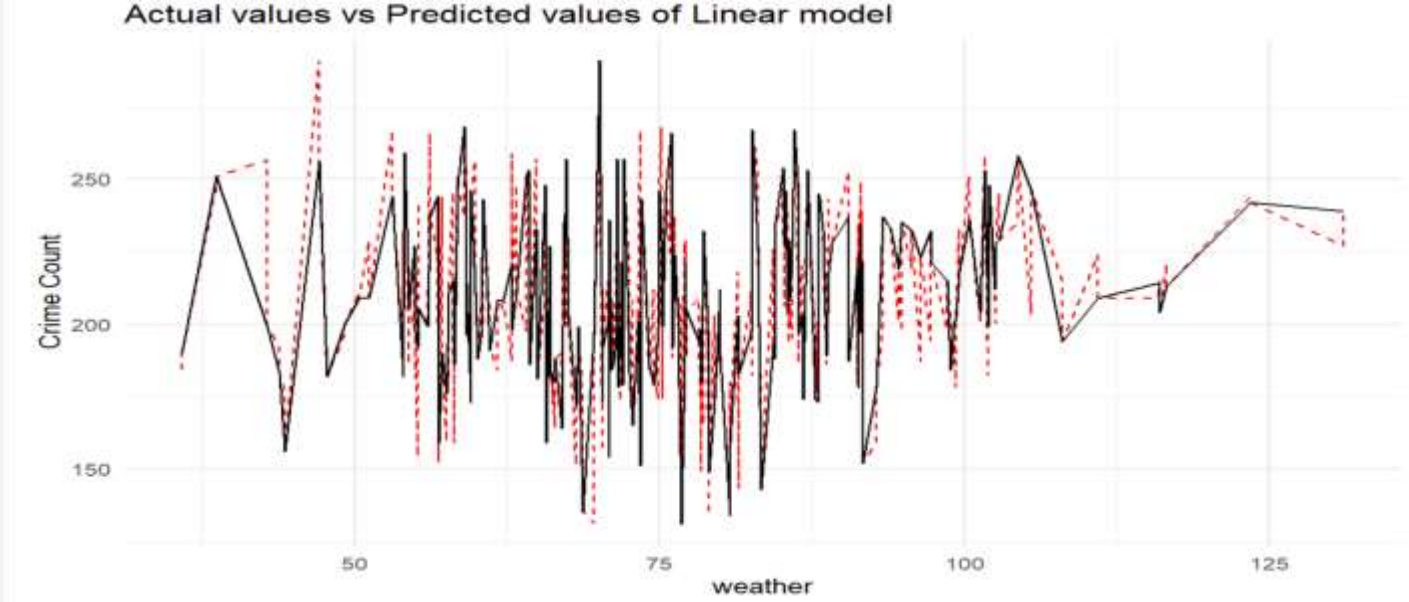
- Metrics Conclusions:

In comparing the Linear Regression and Random Forest Regression models for predicting Crime Count based on weather conditions, the Random Forest model emerges as the superior performer. The Linear Regression model demonstrates a weak fit, with a low R-squared value of 0.07697, suggesting limited explanatory power. On the other hand, the Random Forest model exhibits a significantly lower Mean Squared Error (MSE) of 331.7954 and a higher R-squared value of 0.6745, indicating better accuracy in predicting crime rates. Therefore, based on the lower MSE and higher R-squared value, the Random Forest Regression model is recommended for more accurate predictions in this specific scenario.

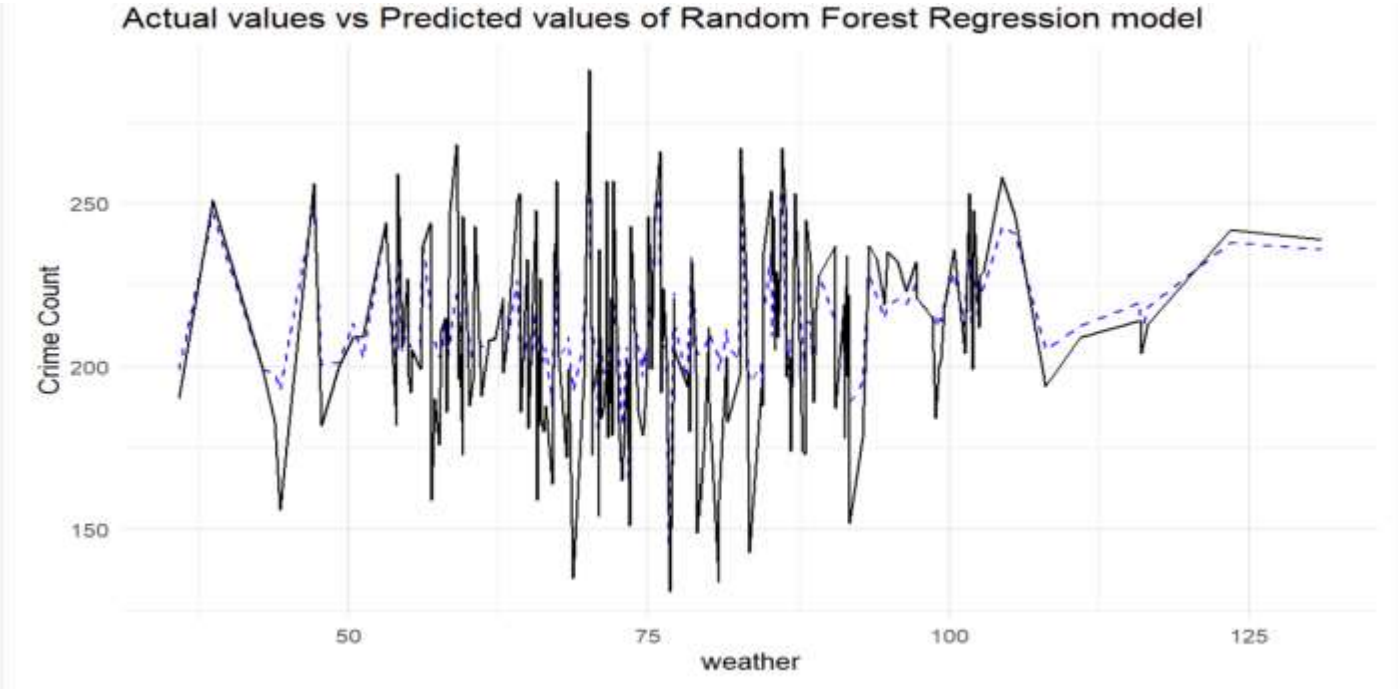
- Graphical Conclusions

The graphical comparison reveals two plots: on the first plot, the actual crime count values are depicted in relation to independent variables, while the 2nd displays the combined relationship of actual and predicted values from a linear regression model. Generally, the predicted values closely align with the actual values. However, the presence of outliers and residual errors suggests that the linear model doesn't precisely fit the actual graph, indicating some discrepancies between predicted and observed values.





The graphical representation below emphasizes the close alignment of actual and predicted values from the random forest regression model showing that they almost coincide. Thus, the combination of statistical metrics and visual assessments suggests that the random forest model provides a more accurate prediction of crime rates compared to the linear model.

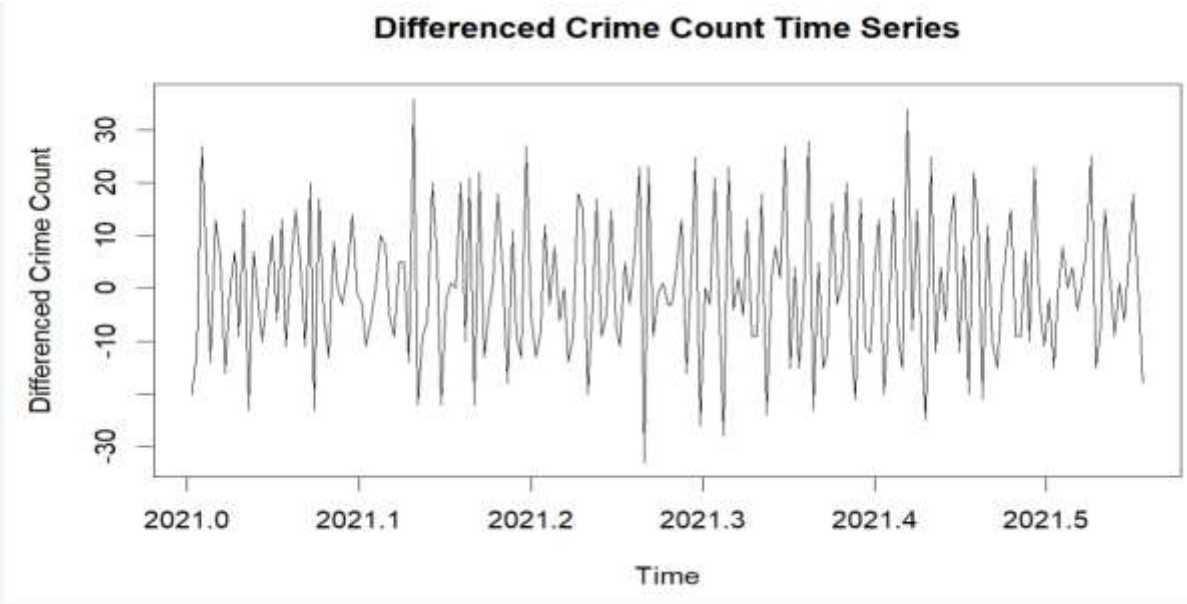


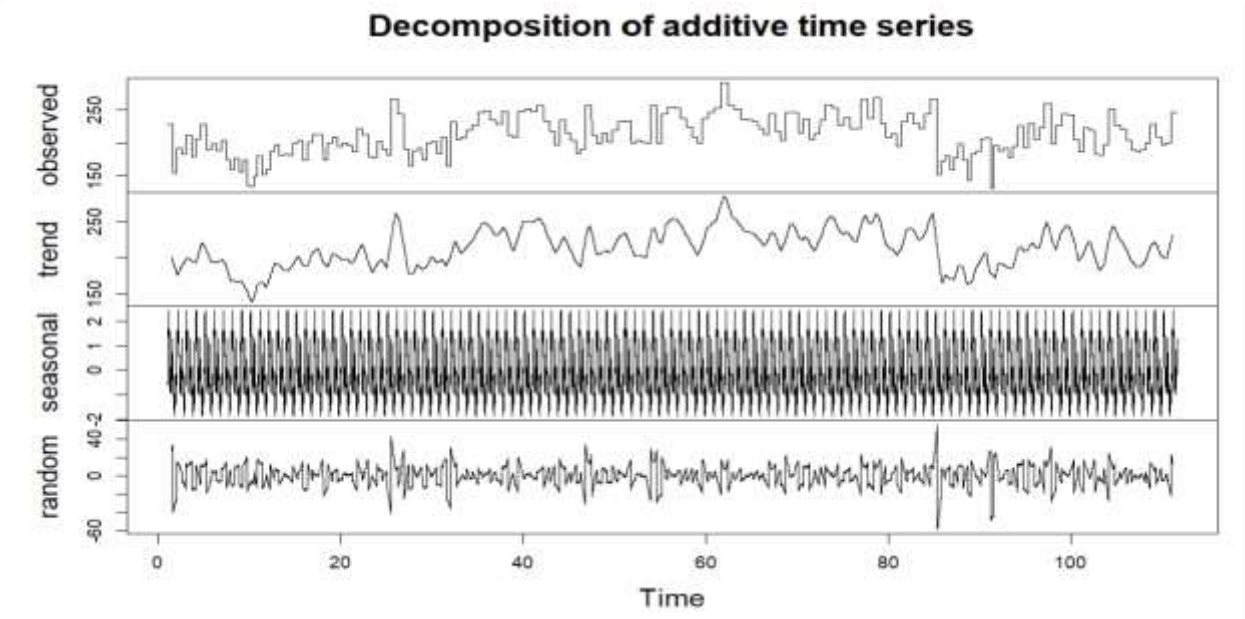
Hence, the model chosen is **Random Forest Regression model**.

Time Series Analysis Report:

Time series analysis is a statistical method used to analyze sequential data points collected over time. In this report, we focus on analyzing a time series dataset related to crime rates in Chicago.

Time Series Visualization:
A crucial step is visualizing the time series data. The graph reveals trends, seasonality, and potential outliers. It seems essential to determine if there's a pattern or seasonality in crime rates over time.





ARIMA Modeling:
Autoregressive Integrated Moving Average (ARIMA) modeling is employed for forecasting future crime rates. The ARIMA model comprises three components: Autoregressive (AR), Integrated (I), and Moving Average (MA).

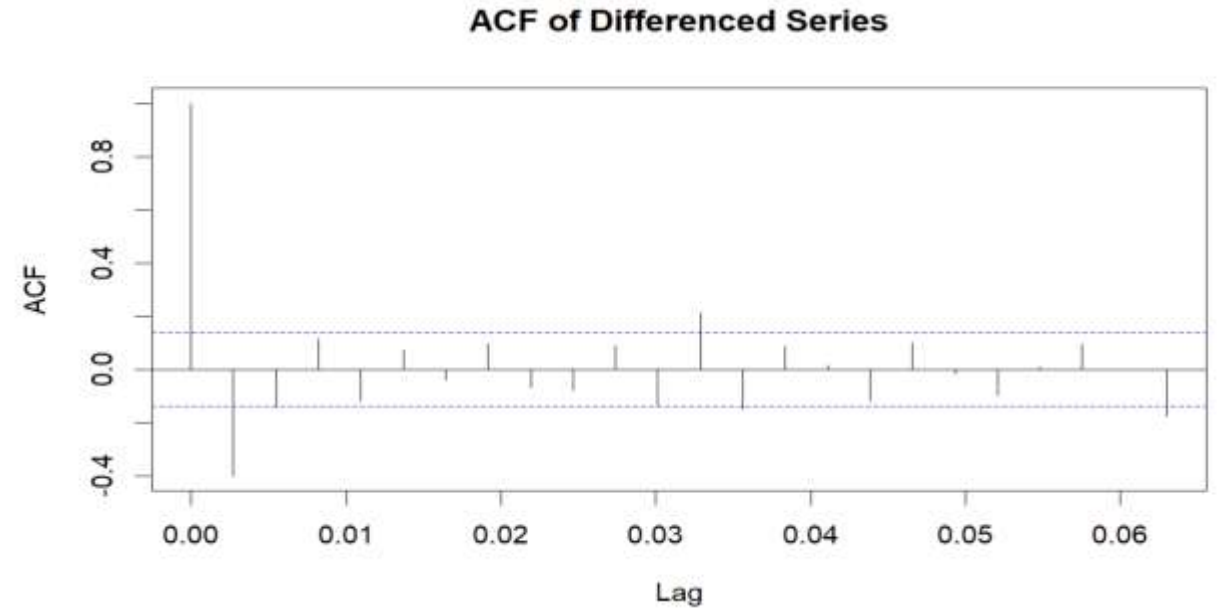
The ARIMA equation is represented as follows:
$Y_t = c + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \ldots + \phi_p Y_{t-p} + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \ldots + \theta_q \varepsilon_{t-q} + \varepsilon_t$

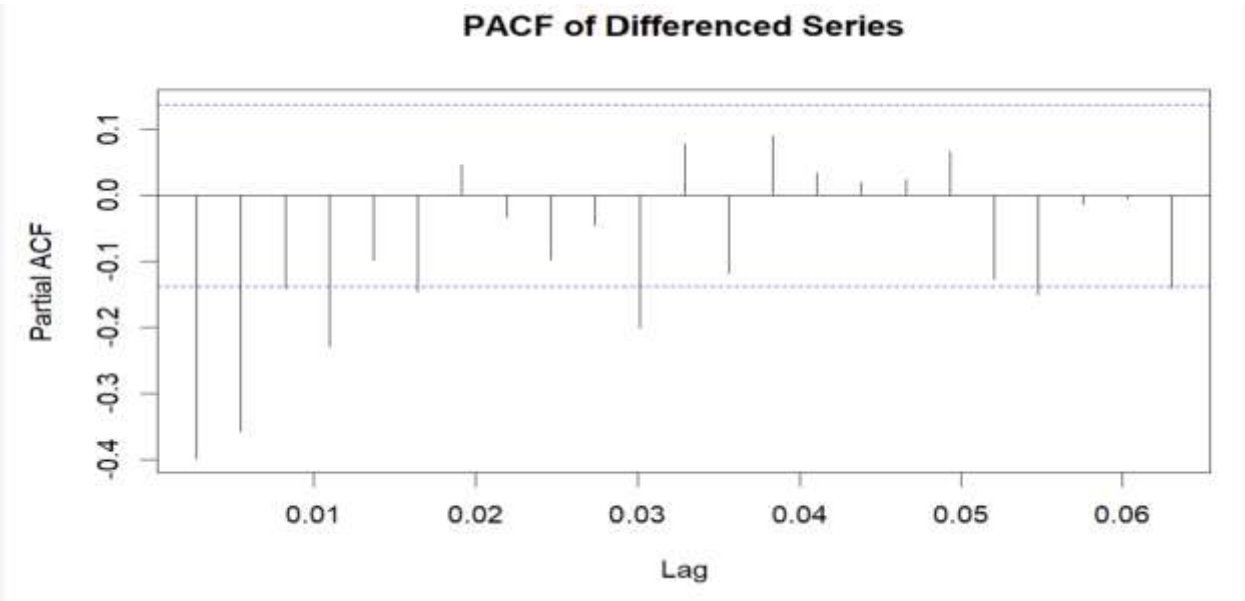
Where:

- Yt is the observed value at time t.
- ϕ represents the autoregressive coefficients.

- θ represents the moving average coefficients.
- c is a constant term.

ACF and PACF:
The Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots are essential tools for model identification. ACF shows the correlation between a variable and its lagged values, while PACF represents the correlation between a variable and its lagged values after removing the effects of shorter lagged values.





Conclusion:
The presentation emphasizes the effectiveness of the ARIMA model in predicting future crime rates through a compelling visual representation. The alignment between the predicted (red) and actual (blue) crime counts, with minimal outliers, underscores the model's accuracy. This robust forecasting capability positions ARIMA as a valuable tool for informed decision-making, particularly in law enforcement and public safety. In essence, time series analysis, as demonstrated by the ARIMA model, offers a powerful means of understanding and predicting sequential data patterns, providing crucial insights for proactive measures in crime prevention and resource

allocation.

The overall conclusion from the multifaceted analysis, incorporating diverse modeling techniques like linear regression and random forest regression, is that the developed model serves as a comprehensive preventive and alerting system. It not only aids the public in avoiding high-crime areas but also provides guidance on safe travel practices during specific weather conditions. The holistic approach leverages predictive modeling to enhance public safety and awareness, mitigating potential risks associated with both crime and adverse weather. This comprehensive strategy showcases the utility of predictive analytics in addressing complex urban challenges and fostering a safer environment for the community.