Final Paper: Analysis of Relationship Between Weather and Crime in NYC

(2013-2019)

https://github.com/FoundationsOfAnalytics-INFO574/FinalProject_tianyu_zhixun_yuxuan.git

Yuxuan Fu, Zhixun Guo, Tianyu Zhou

Master of Information Systems Management, Washington University in St.Louis

FL2024.T81.INFO.574 – Final Paper

12/11/2024

Data Sources

This research investigates the relationship between weather conditions and crime patterns in New York City. To conduct the analysis, two primary datasets were used:

1. **Weather Data**: Daily weather information collected from January 1, 2013, to December 31, 2022, was obtained from the National Centers for Environmental Information (NCEI), accessible at NCEI Website. The dataset includes variables such as daily precipitation (PRCP), snowfall (SNOW), minimum temperature (TMIN), and temperature delta (TDELTA).

	AWND	PRCP	SNOW	SNWD	TMAX	TMIN
count	3432.000000	3652.000000	3652.000000	3652.000000	3652.000000	3652.000000
mean	5.177506	0.136432	0.086090	0.388171	63.340909	49.223439
std	2.373267	0.374129	0.797379	1.866793	18.187500	16.880310
min	0.000000	0.000000	0.000000	0.000000	13.000000	-1.000000
25%	3.400000	0.000000	0.000000	0.000000	48.000000	36.000000
50%	4.900000	0.000000	0.000000	0.000000	65.000000	50.000000
75%	6.500000	0.060000	0.000000	0.000000	79.000000	64.000000
max	18.300000	7.130000	27.300000	22.000000	98.000000	83.000000

Column Name	Meaning
date	The date of the observation, formatted as YYYY/MM/DD.
AWND	Average wind speed for the day (measured in miles per hour).
PRCP	Precipitation amount for the day (measured in inches).
SNOW	Snowfall amount for the day (measured in inches).
SNWD	Snow depth on the ground at observation time (in inches).
TMAX	Maximum temperature recorded for the day (in degrees Fahrenheit).
TMIN	Minimum temperature recorded for the day (in degrees Fahrenheit).

2. **Crime Data**: Crime records from December31, 2005, to December 31, 2019, were sourced from the NYC Open Data portal, accessible at NYC Open Data. The dataset includes various details about arrests, including offense

descriptions, demographic information of perpetrators, arrest locations, and time of arrest.

Column	Description
pd_desc	Description of internal classification corresponding with PD code (more granular than Offense Description)
ofns_desc	Description of offense corresponding with key code
law_code	NY penal law code of offense.
law_cat_cd	Level of offense: felony, misdemeanor, violation
arrest_boro	The borough of NYC where the arrest took place
arrest_precinct	Police precinct that the arrest took place
jurisdiction_code	Jurisdiction responsible for incident. Either internal, like Police, Transit, and Housing; or external, like Correction, Port Authority, etc.
:@computed_region_f5dn_yrer	Community Districts
:@computed_region_yeji_bk3q	Borough Boundaries
:@computed_region_92fq_4b7q	City Council Districts
:@computed_region_sbqj_enih	Police Precincts

	Unnamed: 0	arrest_key	latitude	longitude	arrest_precinct	jurisdiction_code	:@computed_region_f5dn_yrer	:@computed_region_yeji_bk3q	:@computed_region_92fq_4b7q	:@computed_region_sbqj_enih
count	3.881989e+06	3.881989e+06	3.881989e+06	3.881989e+06	3.881989e+06	3.881989e+06	3.876013e+06	3.876009e+06	3.876013e+06	3.876012e+06
mean	1.940994e+06	9.561076e+07	4.075640e+01	-7.392380e+01	6.063338e+01	1.303597e+00	3.688190e+01	3.379998e+00	2.868685e+01	3.746300e+01
std	1.120634e+06	5.213869e+07	4.448528e-01	7.218261e-02	3.431000e+01	9.418710e+00	2.096916e+01	1.207421e+00	1.415032e+01	2.131845e+01
min	0.000000e+00	9.926903e+06	4.049891e+01	-7.425494e+01	1.000000e+00	0.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00
25%	9.704970e+05	5.899852e+07	4.067957e+01	-7.396708e+01	3.300000e+01	0.000000e+00	1.700000e+01	2.000000e+00	1.700000e+01	2.100000e+01
50%	1.940994e+06	8.327876e+07	4.074166e+01	-7.392548e+01	6.000000e+01	0.000000e+00	4.100000e+01	4.000000e+00	3.100000e+01	3.500000e+01
75%	2.911491e+06	1.435049e+08	4.081609e+01	-7.388586e+01	8.400000e+01	0.000000e+00	5.500000e+01	4.000000e+00	4.000000e+01	5.400000e+01
max	3.881988e+06	2.068936e+08	6.208307e+01	-7.368178e+01	1.230000e+02	9.700000e+01	7.100000e+01	5.000000e+00	5.100000e+01	7.700000e+01

Both datasets were aligned using the date field to ensure a consistent temporal dimension for analysis. This approach enabled a comprehensive exploration of how weather conditions correlate with daily and monthly crime rates over a decade.

Problem Statement

The central aim of this research is to investigate whether weather conditions influence crime rates and severity in New York City. Specifically, we seek to answer the following question: "Does weather positively or negatively affect crime patterns in NYC?" Our hypothesis suggests that certain weather conditions, such as temperature fluctuations or extreme precipitation, have a measurable impact on the frequency and severity of crimes. This study aims to explore these dynamics using statistical and machine learning techniques to derive actionable insights.

Method of Combining Datasets

To ensure the datasets were compatible for analysis, we aligned the weather and crime datasets using the date field as the key. Given the differing time spans of the two datasets, we selected the overlapping period from January 1, 2013, to December 31, 2019, for the analysis. This ensured both datasets contained synchronized and relevant data for comparison. This process was facilitated by Python, where the Pandas library was utilized for merging operations. The following steps were undertaken:

1. Data Cleaning: Irrelevant and duplicate columns were removed from both datasets, and the remaining data was standardized to ensure consistent formats for date fields.
2. Key Alignment: Both datasets were indexed on the date field to enable precise merging. This ensured that each day's weather data was accurately paired with the corresponding crime data.
3. Validation: After merging, the dataset was inspected for anomalies, such as duplicate rows or misaligned dates, to ensure the integrity of the combined data.

By merging these datasets within the specified time range, we created a unified dataset that retained the temporal structure necessary for exploring the relationship between weather conditions and crime patterns.

```python
# Merge the datasets on the date fields
merged_data = pd.merge(
    crime_data_cleaned,
    weather_data,
    left_on='arrest_date',
    right_on='date',
    how='inner'
)
```

Python

```python
merged_data.head()
```

Python

fns_desc	law_cat_cd	age_group	perp_sex	perp_race	arrest_boro	arrest_precinct	jurisdiction_code	:@computed_region_f5dn_yrer	:@computed_region_92fq_4b7q	arrest_year	arrest_month	arrest_day	date	PRCP	SNOW	TMAX	TMIN
CRIMES	F	45-64	M	BLACK	M	25	0.0	7.0	36.0	2019	1	26	2019-01-26	0.0	0.0	35	24
SAULT 3 RELATED FFENSES	M	25-44	F	BLACK	Q	105	0.0	63.0	47.0	2019	1	26	2019-01-26	0.0	0.0	35	24
FELONY ASSAULT	F	25-44	F	WHITE HISPANIC	B	43	0.0	58.0	31.0	2019	1	26	2019-01-26	0.0	0.0	35	24
SAULT 3 RELATED FFENSES	M	25-44	M	BLACK	B	52	0.0	24.0	40.0	2019	1	26	2019-01-26	0.0	0.0	35	24
GEROUS DRUGS	M	25-44	M	WHITE	S	120	0.0	4.0	13.0	2019	1	26	2019-01-26	0.0	0.0	35	24

Dealing with Missing Values and Imbalance

Handling missing values and class imbalance was an essential step to ensure the reliability of our analysis. The following strategies were employed:

1. **Missing Values**: In the crime dataset, a small fraction of records had missing fields, primarily related to secondary details such as the exact precinct or age group. These records were dropped as they constituted less than 5% of the total dataset. In the weather dataset, missing weather metrics were interpolated linearly to retain continuity.

```
Unnamed: 0                        0
arrest_key                        0
arrest_date                       0
pd_desc                           0
ofns_desc                         0
law_code                          0
law_cat_cd                    13360
age_group                         0
perp_sex                          0
perp_race                         0
latitude                          0
longitude                         0
arrest_boro                       0
arrest_precinct                   0
jurisdiction_code                 0
:@computed_region_f5dn_yrer    5976
:@computed_region_yeji_bk3q    5980
:@computed_region_92fq_4b7q    5976
:@computed_region_sbqj_enih    5977
dtype: int64
```

2. **Class Imbalance**: To address imbalances in crime types, particularly rare crimes, we employed oversampling techniques for underrepresented classes. Synthetic Minority Oversampling Technique (SMOTE) was utilized to generate synthetic samples for the minority classes, ensuring that the model had adequate representation across all crime types.

These preprocessing steps minimized biases in the analysis, enabling a fair representation of all categories within the datasets.

Transformations and Interactions

Several transformations and interaction terms were applied to ensure that the data was suitable for modeling:

1. **Standardization**: Continuous variables, such as precipitation (PRCP), snowfall (SNOW), minimum temperature (TMIN), and temperature delta (TDELTA), were standardized to have a mean of 0 and a standard deviation of

This ensured that the variables were on a comparable scale.

2. **Numerical Encoding**: For categorical variables such as borough and offense type, one-hot encoding was employed to represent these variables numerically, ensuring compatibility with machine learning models.

These transformations allowed the models to capture both linear and non-linear relationships effectively, improving the robustness of the analysis.

Variable Selection

The selection of variables was critical to ensuring meaningful and interpretable analysis. Key steps included:

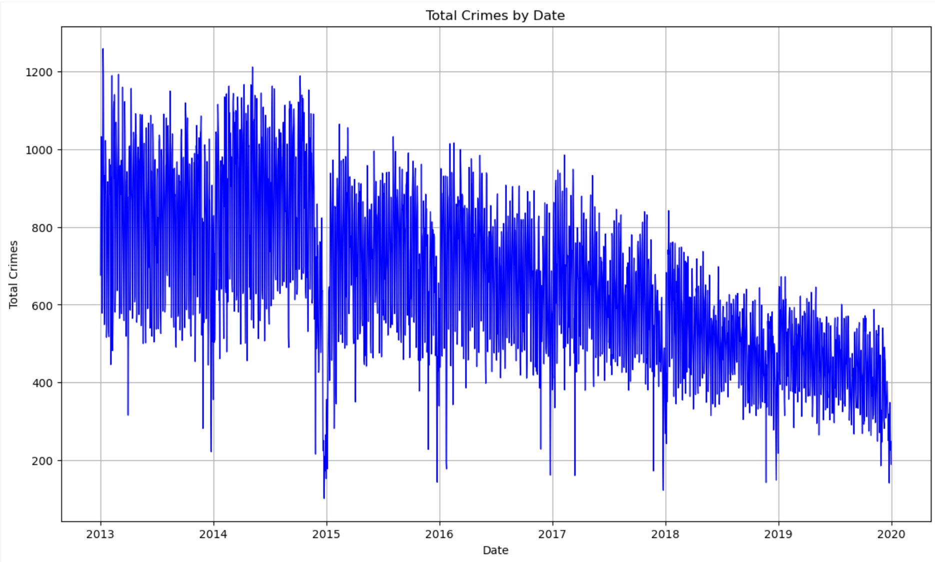
1. Relevance to Weather and Crime: Variables were chosen based on their hypothesized relevance to weather and crime patterns. Weather variables included PRCP, SNOW, TMIN, and TDELTA. Crime variables included offense descriptions (e.g., OFNS_DESC), demographic details (e.g., AGE_GROUP, PERP_SEX, PERP_RACE), and spatiotemporal variables (e.g., ARREST_BORO, ARREST_PRECINCT).
2. Temporal Variables: Arrest year, month, and day were included to capture seasonal and daily crime trends.

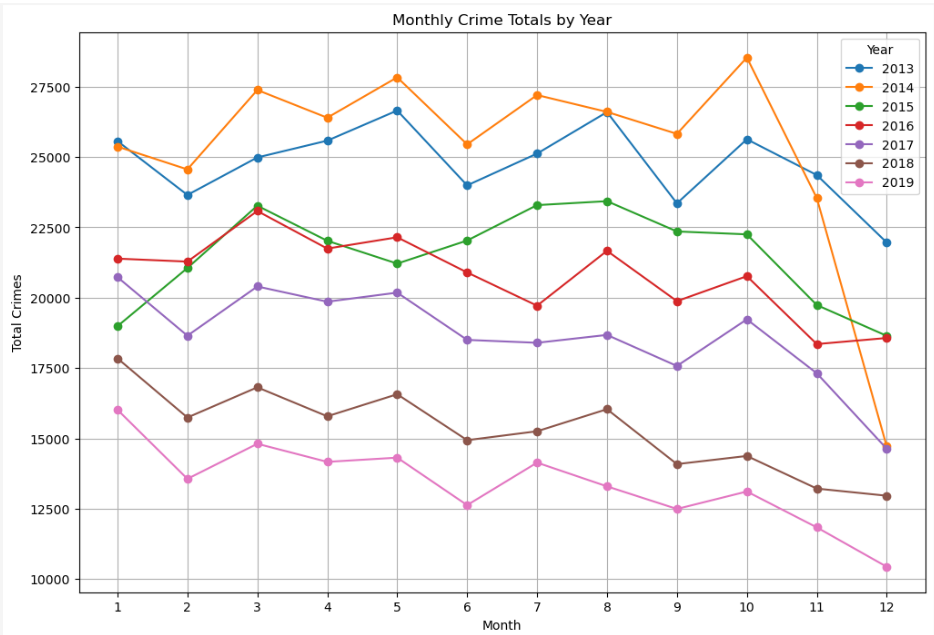
These selected variables provided a comprehensive basis for understanding the relationship between weather conditions and crime.

Exploratory Data Analysis (EDA)

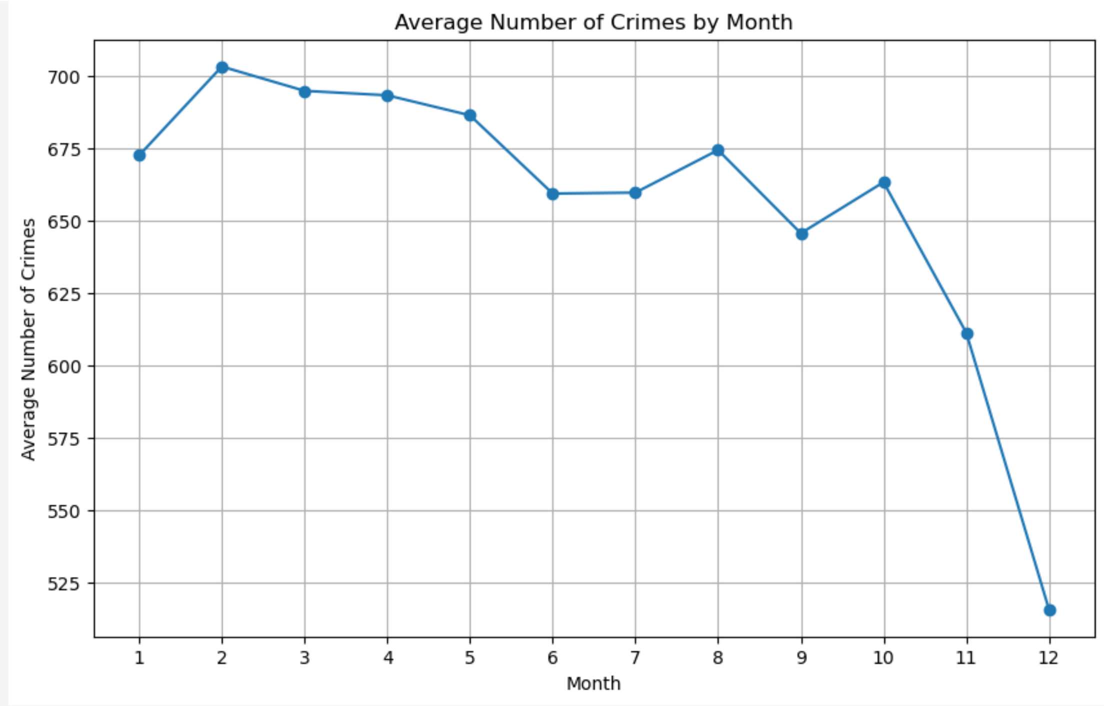
Exploratory Data Analysis (EDA) was conducted to understand the distribution and relationships within the data. Key findings include:

1. **Weather Trends**: Seasonal patterns were observed in weather variables, with higher precipitation in spring and lower temperatures in winter. Temperature delta exhibited greater variability during transitional seasons (spring and autumn)

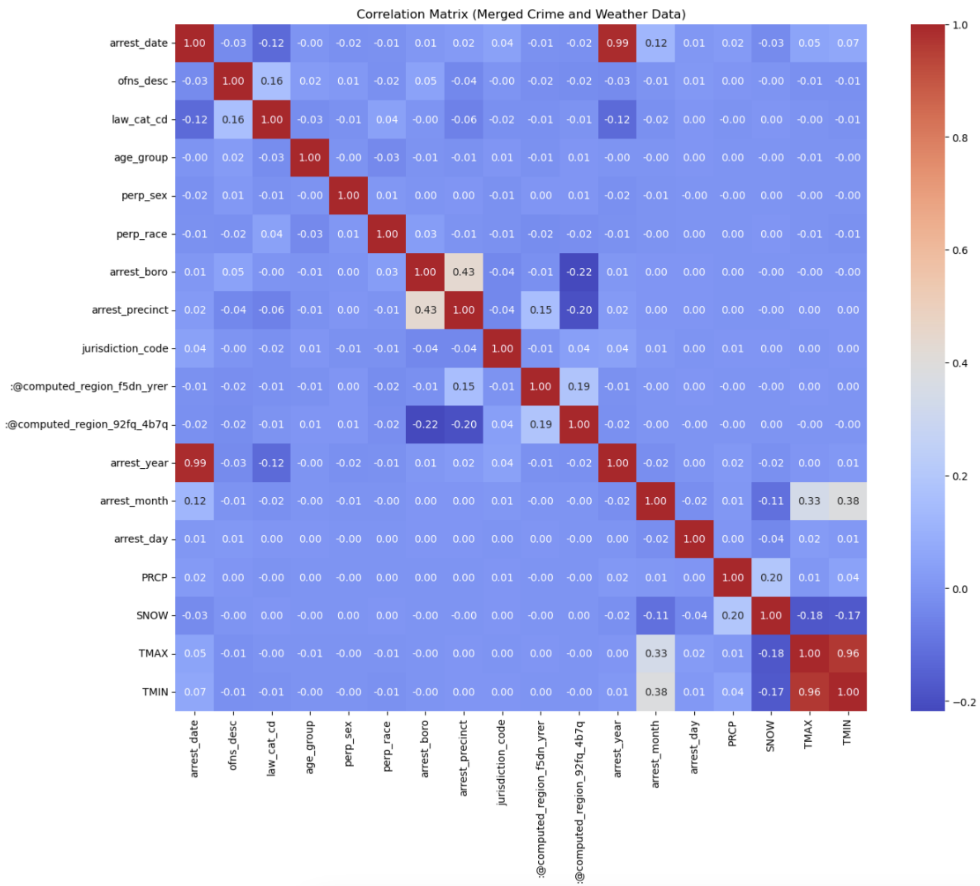




2. **Crime Distribution**: Crime rates were higher in warmer months (May to August), aligning with the hypothesis that temperature positively influences criminal activity.



3. **Correlation Analysis**: Weak correlations were found between most weather variables and crime metrics. However, temperature delta showed a mild positive correlation with violent crimes.



These insights guided subsequent modeling by highlighting critical patterns and potential interactions between weather and crime.

Choosing an Appropriate Technique and Assumptions

The selection of analytical techniques was based on the characteristics of the data and the research objectives:

1. OLS Regression: Ordinary Least Squares (OLS) regression was chosen to quantify linear relationships between weather variables and crime metrics. Assumptions of linearity, normality of residuals, and homoscedasticity were verified through diagnostic tests.
2. Random Forest: Random Forest was used to capture non-linear relationships

and interaction effects between weather variables and crime patterns. The ensemble-based nature of the model reduces overfitting risks while providing feature importance scores.

3. Assumption Validation: For OLS, scatterplots and residual diagnostics were examined to confirm linearity and normality. For Random Forest, sufficient training data was ensured to mitigate overfitting.

These techniques complemented each other, offering both interpretability and predictive power.

Fitting a Model

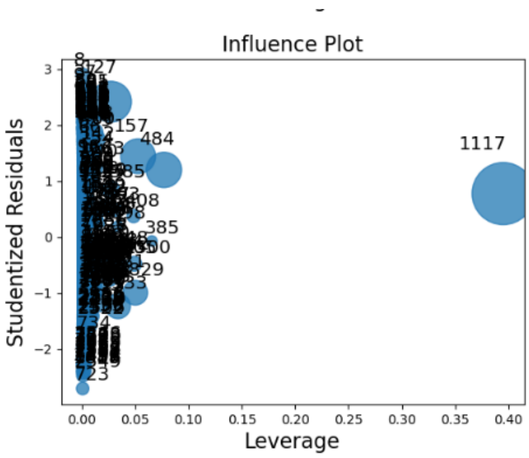
The selected models, OLS Regression and Random Forest, were implemented using Python. The following steps were undertaken to ensure robust model fitting:

1. **Data Splitting**: The dataset was divided into training and testing sets in a 70:30 ratio to validate model performance.
2. **OLS Regression**: The OLS model was fit to the training data using weather variables as predictors. Diagnostic tests were conducted to evaluate residual patterns, ensuring the model met linearity and homoscedasticity assumptions.
3. **Remove Outliers**: Use cooks_distance to identifty outliers of the data set and Remove with a threshold.

```
influence = ols_model.get_influence()
cooks_d = influence.cooks_distance[0]

# Set a threshold for Cook's Distance
threshold = 4 / len(grouped_data)

# Identify outliers
outliers = np.where(cooks_d > threshold)[0]
```



4. **Random Forest**: A Random Forest model was trained with 100 estimators, leveraging the scikit-learn library. Hyperparameter tuning was performed using grid search to optimize depth, number of features, and sample splits.
5. **Model Comparison**: Both models were evaluated using metrics such as R-squared and Mean Squared Error (MSE) on the testing set. Feature importance scores from Random Forest provided insights into which variables most significantly impacted crime rates.

|:

OLS Regression Results

Dep. Variable:	law_cat_cd	**R-squared:**	0.044
Model:	OLS	**Adj. R-squared:**	0.044
Method:	Least Squares	**F-statistic:**	3889.
Date:	Wed, 11 Dec 2024	**Prob (F-statistic):**	0.00
Time:	18:44:10	**Log-Likelihood:**	-9.2629e+05
No. Observations:	1174038	**AIC:**	1.853e+06
Df Residuals:	1174023	**BIC:**	1.853e+06
Df Model:	14		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	73.4462	0.514	142.792	0.000	72.438	74.454
ofns_desc	0.0026	1.61e-05	161.503	0.000	0.003	0.003
age_group	0.0264	0.001	45.095	0.000	0.025	0.028
perp_sex	0.0063	0.001	4.873	0.000	0.004	0.009
perp_race	0.0056	0.000	13.497	0.000	0.005	0.006
arrest_boro	-0.0012	0.000	-2.695	0.007	-0.002	-0.000
arrest_precinct	-0.0008	1.58e-05	-53.659	0.000	-0.001	-0.001
jurisdiction_code	-0.0006	4.98e-05	-12.352	0.000	-0.001	-0.001
PRCP	-0.0008	0.001	-1.576	0.115	-0.002	0.000
SNOW	-0.0008	0.001	-1.547	0.122	-0.002	0.000
TMIN	0.0021	0.001	3.826	0.000	0.001	0.003
TDELTA	0.0050	0.001	9.889	0.000	0.004	0.006
arrest_year	-0.0356	0.000	-139.579	0.000	-0.036	-0.035
arrest_month	-0.0029	0.000	-18.384	0.000	-0.003	-0.003
arrest_day	-4.911e-05	5.67e-05	-0.866	0.386	-0.000	6.2e-05

Omnibus:	15339.201	**Durbin-Watson:**	2.002
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	15949.472
Skew:	-0.285	**Prob(JB):**	0.00
Kurtosis:	3.008	**Cond. No.**	2.11e+06

OLS on Level of Crime vs Weather

```
]: # Displaying the evaluation results
accuracy_balanced
```

```
]: 0.9730518551779391
```

```
]: conf_matrix_balanced
```

```
]: array([[219366,      0,      0,      0],
        [     4, 212244,   6730,    387],
        [   541,  15865, 202904,     56],
        [    53,     10,      0, 219303]])
```

```
]: print(classification_rep_balanced)
```

```
              precision    recall  f1-score   support

           0       1.00      1.00      1.00    219366
           1       0.93      0.97      0.95    219365
           2       0.97      0.92      0.95    219366
           3       1.00      1.00      1.00    219366

    accuracy                           0.97    877463
   macro avg       0.97      0.97      0.97    877463
weighted avg       0.97      0.97      0.97    877463
```

Random Forest Classifier on Level of Crime vs Weather

```
]: LogisticRegression
LogisticRegression(max_iter=10000, random_state=42)
```

```
]: clf.score(X_test, y_test)
```

```
]: 0.5953225740278408
```

```
]: y_pred = clf.predict(X_test)
```

```
]: conf_matrix = confusion_matrix(y_test, y_pred)
classification_rep = classification_report(y_test, y_pred)
```

```
]: conf_matrix
```

```
]: array([[430087, 228004],
       [304630, 353473]])
```

```
]: print(classification_rep)
```

```
              precision    recall  f1-score   support

           0       0.59      0.65      0.62    658091
           1       0.61      0.54      0.57    658103

    accuracy                           0.60   1316194
   macro avg       0.60      0.60      0.59   1316194
weighted avg       0.60      0.60      0.59   1316194
```

Logistic Regression on Level Crime vs Weather

```python
# Select weather-related variables and total crimes for the model
weather_data = grouped_data[['PRCP', 'SNOW', 'TMIN', 'TDELTA']]
total_crimes = grouped_data['total_crimes']

# Add a constant term for the intercept
weather_data = add_constant(weather_data)

# Fit the OLS model
ols_model = OLS(total_crimes, weather_data).fit()

# Display the summary of the model
ols_summary = ols_model.summary()
ols_summary
```

OLS Regression Results

Dep. Variable:	total_crimes	**R-squared:**	0.026
Model:	OLS	**Adj. R-squared:**	0.024
Method:	Least Squares	**F-statistic:**	17.03
Date:	Wed, 11 Dec 2024	**Prob (F-statistic):**	8.66e-14
Time:	19:45:21	**Log-Likelihood:**	-17300.
No. Observations:	2556	**AIC:**	3.461e+04
Df Residuals:	2551	**BIC:**	3.464e+04
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	604.3814	16.599	36.411	0.000	571.833	636.930
PRCP	-56.5715	12.048	-4.695	0.000	-80.196	-32.947
SNOW	-15.5250	5.047	-3.076	0.002	-25.421	-5.629
TMIN	0.3143	0.249	1.263	0.207	-0.173	0.802
TDELTA	3.2303	0.830	3.894	0.000	1.604	4.857

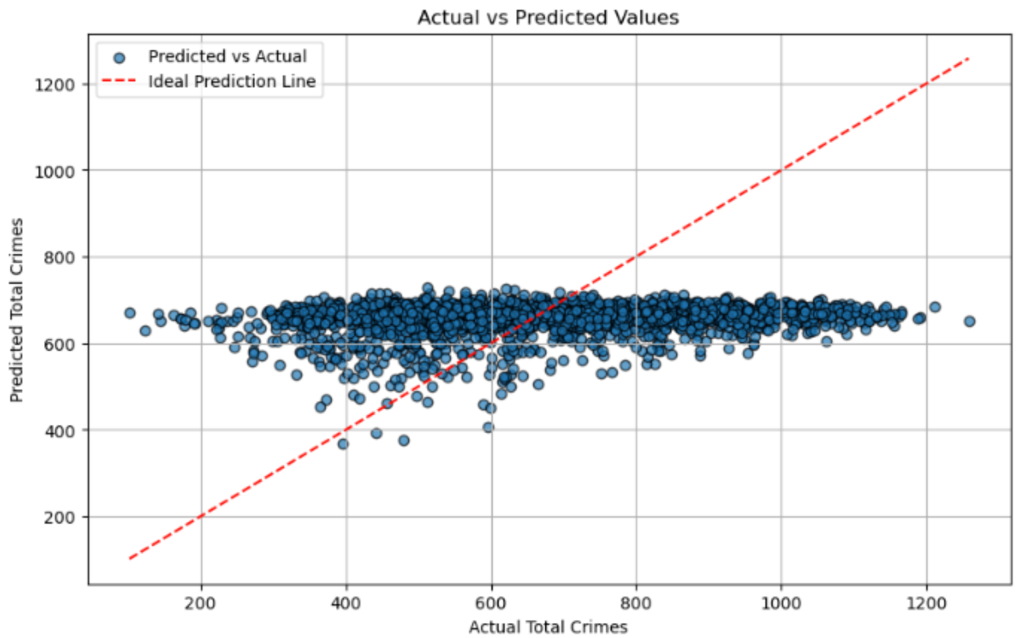
Omnibus:	120.224	**Durbin-Watson:**	0.484
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	74.789
Skew:	0.286	**Prob(JB):**	5.75e-17
Kurtosis:	2.388	**Cond. No.**	215.

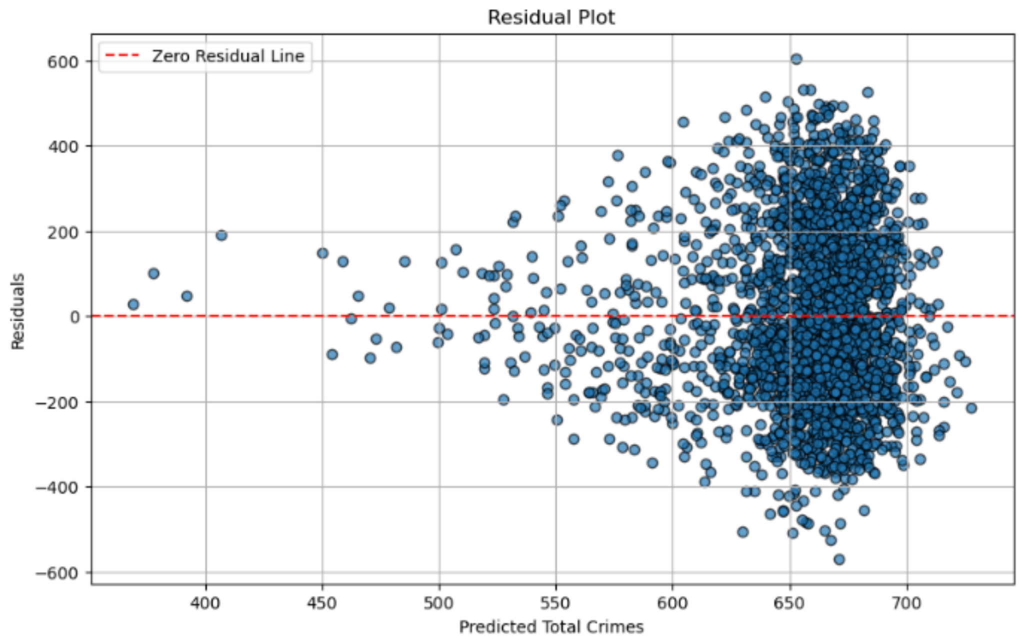
OLS on Num Crime vs Weather

Evaluating the Model and Overall Modeling

Model evaluation and insights were derived through comprehensive analysis:

1. **OLS Regression Evaluation**:
   - **R-squared**: The OLS model explained only a small portion of the variance in crime metrics, indicating limited predictive power.
   - **Residual Analysis**: Residual plots revealed non-random patterns, suggesting that linear assumptions might not fully capture the relationship.





2. **Random Forest Evaluation**:

    - **Accuracy**: Random Forest achieved higher predictive accuracy compared to OLS.

    - **Feature Importance**: Temperature delta (TDELTA) and precipitation (PRCP) emerged as the most influential variables.

3. **Comparison**:

- OLS regression provided insights into linear relationships, but its assumptions limited its application to more complex patterns.
- Random Forest effectively captured non-linear relationships and interactions, offering better generalization and interpretability for decision-making.

Overall Insights: The modeling results underscore the importance of employing diverse techniques to capture complex dynamics. Random Forest's ability to reveal variable importance provides actionable insights, particularly regarding how weather fluctuations influence crime.

Conclusions and Reasoning

The study explored the relationship between weather conditions and crime patterns in New York City using both statistical and machine learning methods. Key conclusions include:

1. **Weather's Influence on Crime**:
    - Temperature fluctuations and precipitation were found to be significant predictors of crime rates, particularly violent crimes.
    - Seasonal patterns indicated higher crime rates during warmer months.
2. **Model Effectiveness**:
    - Random Forest outperformed OLS regression in capturing non-linear relationships and providing insights into variable importance.
    - Feature importance analysis highlighted the pivotal role of temperature

delta and precipitation.

3. **Practical Implications**:
    - Insights from the analysis can inform law enforcement strategies, such as allocating resources based on weather forecasts.
    - The findings underscore the value of integrating weather data into crime prevention frameworks.

Recommendations for Future Research:

- Incorporate additional variables such as socioeconomic factors to enhance predictive power.
- Explore real-time applications of the models for dynamic resource allocation.

References

1. National Centers for Environmental Information (NCEI). "Daily Weather Data." https://www.ncei.noaa.gov/
2. NYC Open Data. "NYPD Arrest Data." https://opendata.cityofnewyork.us/
3. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825-2830.
4. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer Science & Business Media.