

5205 Group 1's Research on Association between Crime Rates and Seasons

Introduction

New York City is a densely populated metropolis. Crime is an essential factor affecting residents' quality of life and social stability there. On this basis, effective police management and prediction will help significantly improve police efficiency in dealing with crimes, reduce the occurrence of crimes, ensure public safety, and maintain social stability. This study aims to provide data support for the police by analyzing the association between crime rates and seasons, optimizing the police force allocation strategy, and helping prevent and reduce the occurrence of crime in advance.

Literature Review

Proactive Policing and Crime Trends

Hanna's analysis (2023) underscores the importance of proactive policing in mitigating urban crime, highlighting a decline in arrests over the past decade and its potential contribution to increased criminal activity. He advocates for a balanced approach, combining fair enforcement practices with targeted efforts to combat specific offenses such as criminal mischief and felony drugs. This perspective emphasizes the crucial role of law enforcement strategies in maintaining public safety.

Challenging Political Narratives

Farley's article (2023) delves into criticisms of Manhattan District Attorney Alvin Bragg's crime policies, particularly amidst claims of a rise in violent crime. However, statistical data from the first quarter of 2023 challenges these assertions, illustrating the nuanced nature of urban crime dynamics. Farley's analysis challenges simplistic political narratives, highlighting the need for a comprehensive understanding of the factors influencing crime rates in Manhattan.

The Relationship Between Seasons and Crime

Research into the relationship between seasons and crime rates in New York City is consistent with broader findings that weather and seasons can significantly affect crime patterns. Studies have shown that warmer temperatures and summer months are often associated with an increase in violent crime, such as assault and homicide (Corcoran & Zahnow, 2022). This association may be due to a combination of more social interaction during warmer months and the physiological effects of heat on human behavior. Studying the relationship between seasonal change and crime rate can help allocate police resources more effectively and reduce the occurrence of crime.

Importance of study

Understanding crime rates, especially in a city like New York, is critical to informed decision-making and effective law enforcement strategies. The literature review delves into the complex interplay of factors contributing to this phenomenon, including the effects of seasonal changes, types of crime, and social unrest. By examining the relationship between these factors and crime rates, these studies provide valuable insights that can inform evidence-based interventions to mitigate urban violence. In addition, the literature calls into question simplistic narratives surrounding crime trends and law enforcement policies, emphasizing the need for a nuanced approach to address these challenges. By studying the relationship between seasonal changes and the impact of crime rates, we can gain insight into the broader societal dynamics that influence criminal behavior and help develop more effective, data-driven public safety

policies. At the same time, it can promote cooperation with communities to solve crime problems, thereby improving public safety and social well-being.

Research Questions

Research Question 1: Is there an association between the total number of crimes and the seasons in New York City?

H0: There is no association between the total number of crimes and the seasons in New York City.

H1: There is an association between the total number of crimes and the seasons in New York City.

The alternative hypothesis is informed by the possibility that different seasons might affect the frequency of crimes due to varying social activities and environmental conditions.

Research Question 2: Is there an association between the number of assault crimes and the seasons in New York City?

H0: There is no association between the number of assault crimes and the seasons in New York City.

H1: There is an association between the number of assault crimes and the seasons in New York City.

One of the key reasons for examining this question is that we find assault to be the most common crime through our analysis. Thus, we believe that studying assault rates can also help the police department effectively deploy police officers to maintain safety in New York City. Additionally, the alternative hypothesis considers that certain conditions in different seasons, such as increased interactions during warmer months, might lead to higher incidents of assault.

Overall, we plan to conduct ANOVA tests to determine whether to reject the null hypotheses of these two questions. Specifically, we set the confidence level at 95%. In other words, if the p-value is less than 0.05, we will have enough evidence to reject the null hypothesis.

Datasets

This study obtains two raw datasets from the New York City Police Department (2024a; 2024b). They contain a vast range of information on crime events in New York City from 2017 to 2022, and in 2023 respectively. They capture over 5 million records with 19 columns. To be specific, the datasets provide the arrest event identifier, date, location, jurisdiction, coordinate, classification, relevant law code and severity. Also, they detail the age, gender and race of the perpetrator. Among them, 9 columns are 'plain text', while 8 columns are 'number'. Besides, 'ARREST_DATE' and 'Lon_Lat' are 'date' and 'point', respectively. Overall, based on these datasets, this study could have some insight into the crime situation in New York City in the past 7 years.

Overall, these two datasets have relatively favorable suitability. First, they come from official sources, which ensures the authenticity and accuracy of the data to a certain extent. In addition, it covers arrest information from 2017 to 2023, which can facilitate us to build time series with a long time span to analyze the seasonal fluctuation of crime rates. Moreover, we also find that these two datasets have fewer missing values in them despite the large number of records. This completeness not only reduces the workload of cleaning data, but also helps to reduce the possibility of biased results due to inappropriate data tidying. Furthermore, these datasets also include detailed information, such as specific types of crime, which are necessary for in-depth analyses. In conclusion, these two datasets provide a solid foundation for our study.

Techniques

In our study, we chose time series (ARIMA model and ETS model) and ANOVA tests as the main analysis techniques based on data characteristics and research needs. First, as the dataset covers daily crime records from 2018 to 2023, time series analysis seems to be the ideal tool. Specifically, it can effectively summarize such data and explore trends and seasonal patterns. Second, the purpose of our study is to explore the association between crime rates and seasons. Concretely, we divide the year into four seasons and compare whether there is a significant difference in the number of crimes between the seasons. Through the ANOVA test, we can exactly assess the difference in the mean value of crime rates across seasons, examining whether seasonal factors have a significant impact on the number of crimes. For instance, if the p-value of the ANOVA test is smaller than 0.05, it has evidence to support that there is an association between the crime rate and the seasons at a 95% confidence level.

Data Analysis

Data Tidying

The raw datasets have some issues. For example, the datasets store the six-year data in two CSV files. In addition, the crime severity 'LAW_CAT_CD' has some missing values. Moreover, some columns are weakly related to the research questions. Besides, some columns, such as 'LAW_CODE', also contain internal codes that are hard to understand. Furthermore, some columns, such as 'ARREST_BORO', use acronyms, which is not intuitive enough. Hence, in order to improve research efficiency, this study utilizes R functions to clean the raw dataset.

First, this study utilizes the 'rbind' to merge these two files. Then, it simplifies the whole raw dataset. Specifically, it uses the 'select' to drop columns that are relatively irrelevant or tough to decipher. As a result, the dataset only keeps the following columns from the raw dataset:

- **ARREST_KEY:** A unique identifier for each arrest record.
- **ARREST_DATE:** The date on which the arrest was made.
- **OFNS_DESC:** Offense description, which describes the nature of the crime.
- **LAW_CAT_CD:** Law category code classifying the offense as felony, misdemeanor, etc.
- **ARREST_BORO:** The borough in which the arrest was made.
- **AGE_GROUP:** The age group of the perpetrator.
- **PERP_SEX:** The sex of the perpetrator.
- **PERP_RACE:** The race of the perpetrator.
- **X_COORD_CD and Y_COORD_CD:** Coordinates for the arrest location.
- **Latitude and Longitude:** Geographical coordinates for the arrest location.
- **Lon_Lat:** A new or calculated field that combines geographic data.

Then, this study uses the 'mdy' to convert the 'ARREST_DATE' into the 'date' format. Following that, it extracts the month information from the 'ARREST_DATE' and stores it in the new column 'ARREST_MONTH'. Through this step, this study could directly use the monthly data to explore the changes in crime rates over a year.

After that, by leveraging the 'group_by' and 'summarise', the total number of monthly crime events is calculated. Then, this data is stored in the new column 'MONTHLY_CRIME_COUNT' through 'left_join'. In this way, the research could compare the changes in the number of crimes of different months.

Moreover, in order to make the dataset easier to read and understand, this research also transforms the values in some columns through ‘mutate’. For instance, for ‘LAW_CAT_CD’, it converts the ‘F’, ‘M’ and ‘V’ into ‘felony’, ‘misdemeanor’ and ‘violation’ respectively. Similarly, the ‘B’, ‘S’, ‘K’, ‘M’ and ‘Q’ in ‘ARREST_BORO’ are converted into ‘Bronx’, ‘Staten Island’, ‘Brooklyn’, ‘Manhattan’ and ‘Queens’ respectively. Also, it replaces the ‘F’ and ‘M’ with ‘female’ and ‘male’ respectively.

In addition, this study also stresses the abnormal values and missing values in the ‘LAW_CAT_CD’. To be specific, it classifies the null value as ‘not provided’. Also, some abnormal values, such as ‘9’ and ‘I’, are defined as ‘other’.

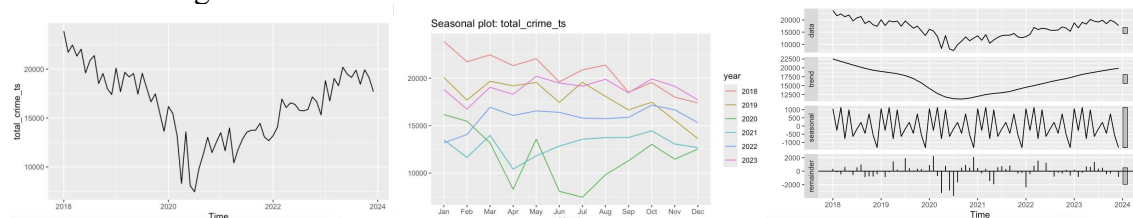
Furthermore, this study also uses ‘tolower’ to convert all values ‘PERP_RACE’ and ‘OFNS_DESC’ into lower letters. This could help to ensure consistency when comparing and searching data.

Finally, this study utilizes ‘range’ and ‘filter’ to sort the information in chronological order and filter data from 2018 onwards.

By taking these above measures, this study could obtain a cleaned dataset, including over 1 million records with 16 columns. Compared with the raw dataset, the cleaned dataset is cleaner and more organized. The higher quality, readability and usability of its data could offer a solid foundation for further research.

Research Question 1

For the first research question, we first create a time series about the total number of crimes in New York City. Then, we plot multiple graphs for this time series and observe whether there is some kind of pattern in the number of crimes over time. Through the graphs, we initially determine that there may be a trend and seasonality in the number of crimes. For example, from the beginning of 2018 to the end of 2023, the crime rate decreases and then increases. In addition, the crime rate is higher in the summer while it seems to be lower in the winter.



While long-term trends can show the overall direction of movement of crime rates over time, understanding the seasonality of crime is realistically more conducive to the government's ability to optimize the allocation of police resources. For instance, if crime is higher during a certain season, the police department can increase the number of patrols or police deployments. Therefore, the following study will focus on the seasonality of crime rates.

To better investigate whether there is an association between the number of crimes and time. We first build different models and run the Ljung-Box test to test whether the models sufficiently capture the main time-dependent patterns in the data. First, we assume that the data is stationary to build the ARIMA model. Next, we create the ETS: AAA model by assuming that the data are additive in terms of error, trend, and seasonality. Finally, we establish the ETS auto model.

```

> # AMRIMA Model
> arima_model_total <- auto.arima(total_crime_ts)
> arima_model_total
Series: total_crime_ts
ARIMA(2,1,0)(1,0,0)[12]

Coefficients:
    ar1      ar2    sar1
 -0.4715 -0.1835  0.2396
s.e.    0.1265  0.1188  0.1353

sigma^2 = 2652003; log likelihood = -624.76
AIC=1257.52  AICc=1258.13  BIC=1266.57

> #ES AAA Model
> ets_aaa_total = ets(total_crime_ts,model = 'AAA')
> ets_aaa_total
ETS(A,A,A)

Call:
ets(y = total_crime_ts, model = "AAA")

Smoothing parameters:
alpha = 0.7166
beta = 0.0255
gamma = 1e-04

Initial states:
l = 22453.5286
b = -131.0976
s = -985.5331 -230.3583 831.28 -332.1091 129.0769 -255.0669
-562.6673 936.8756 -932.9989 1172.132 -275.7949 505.1645

sigma: 1497.306

AIC  AICc  BIC
1376.670 1388.004 1415.374

```

Based on the results, we find that the ARIMA model performs optimally. It has the smallest AIC, AICc and BIC. However, the three indices of these three models are relatively close to each other, which makes it difficult to say decisively that the ARIMA model must be optimal. Therefore, we further compare the Ljung-Box test results of the three models with 1st-order and 14th-order lags.

```

> checkresiduals(arima_model_total)

Ljung-Box test

data: Residuals from ARIMA(2,1,0)(1,0,0)[12]
Q* = 15.77, df = 11, p-value = 0.1499
Model df: 3. Total lags used: 14

> Box.test(arima_model_total$residuals, type = "Ljung-Box")

Box-Ljung test

data: arima_model_total$residuals
X-squared = 0.034957, df = 1, p-value = 0.8517

> checkresiduals(ets_aaa_total)

Ljung-Box test

data: Residuals from ETS(A,A,A)
Q* = 23.928, df = 14, p-value = 0.04675
Model df: 0. Total lags used: 14

> Box.test(ets_aaa_total$residuals, type = "Ljung-Box")

Box-Ljung test

data: ets_aaa_total$residuals
X-squared = 0.015714, df = 1, p-value = 0.9002

> checkresiduals(ets_model_total)

Ljung-Box test

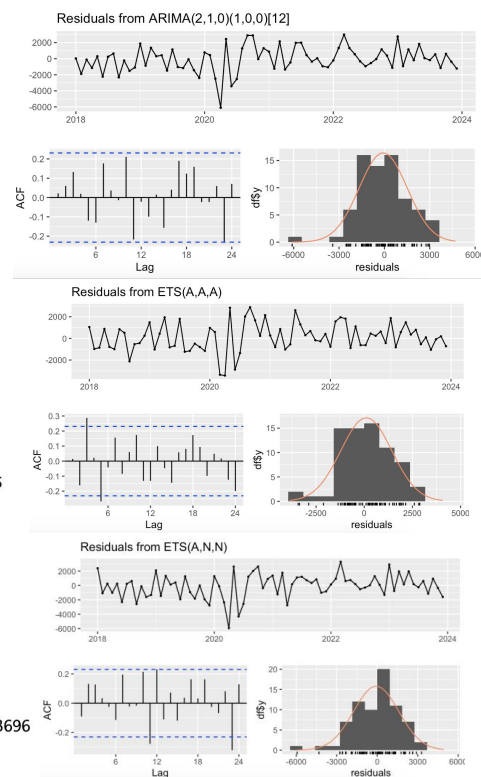
data: Residuals from ETS(A,N,N)
Q* = 24.765, df = 14, p-value = 0.03696
Model df: 0. Total lags used: 14

> Box.test(ets_model_total$residuals, type = "Ljung-Box")

Box-Ljung test

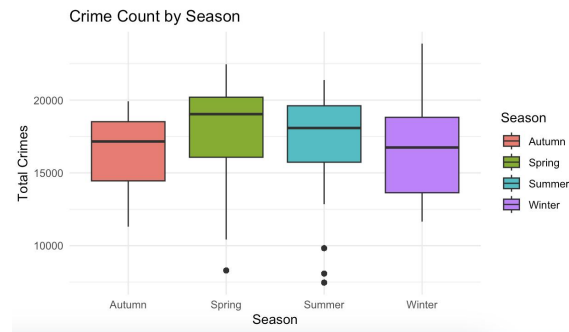
data: ets_model_total$residuals
X-squared = 0.62503, df = 1, p-value = 0.4292

```



Based on the results, we find that the p-values of the ARIMA model are all greater than 0.05. This indicates that the residuals are white noise at the 95% confidence level. This means that the ARIMA model adequately captures the time-dependent patterns. However, both ETS models have a p-value greater than 0.05 for the 1st-order lag and less than 0.05 for the 14th-order lag. This may imply that there is some sort of autocorrelation in the data over longer lags, which may be an indication of seasonality.

In addition, we plot box plots. Based on the picture, we can note that the crime rate seems to be higher in the spring and summer than in the fall and winter.



Hence, to further investigate the relationship between the number of crimes and seasonality, we further impose a statistical test for analysis. Specifically, we divide the 12 months of the year into four seasons and then perform an ANOVA test.

```
> anova_result <- aov(MONTHLY_CRIME_COUNT ~ Season, data = cleaned_dataset)
> summary(anova_result)
```

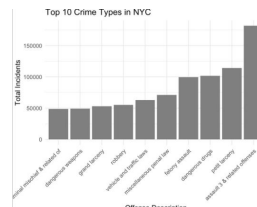
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Season	3	2.372e+11	7.906e+10	7092	<2e-16 ***
Residuals	1173952	1.309e+13	1.115e+07		

 Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

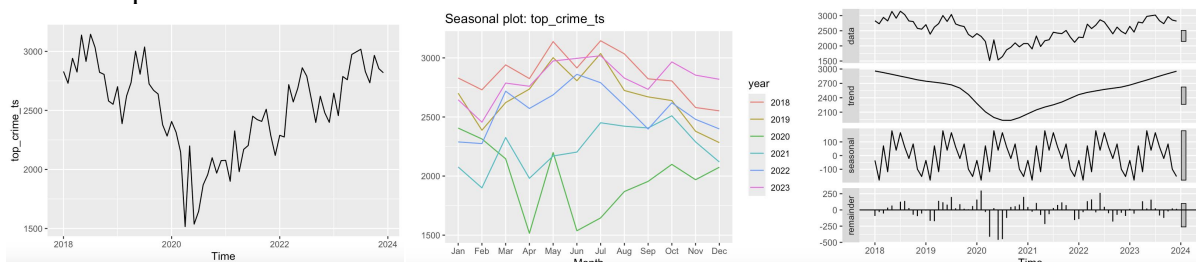
According to the ANOVA test result, the p-value is smaller than 0.05. Thus, there is enough evidence to reject the null hypothesis at the 95% confidence level. This indicates that there is an association between the total number of crimes and seasons at the 95% confidence level.

Research Question 2

We also plan to study the most high-frequency crime types in New York City and examine their number in relation to time. According to the code, assault is the most common type of crime in New York City.



Therefore, we create a time series about the number of assaults. Moreover, to visualize the change in the number of assaults, we plot multiple graphs. Based on the graphs, we find that the change in the number of assaults in New York City tends to coincide with the change in the total number of crimes. Specifically, It also shows a decrease and then an increase from the beginning of 2018 to the end of 2023. Additionally, numbers are higher in the summer whereas they appear to be lower in the winter. Similar to the first question study, we also believe that seasonality is more conducive to the police department's ability to target its police force. Therefore, in the following study, the focus will be on the seasonality of the amount of assaults. Also, it is important to note that the flow of research on this question is generally similar to that of the first question.



To further investigate whether there is an association between the number of assaults and the time. We plan to first build different models and run the Ljung-Box test to examine whether the models adequately capture the main time-dependent patterns in the data. First, we assume that the data is stationary to build the ARIMA model. Second, we create an ETS: AAA model by assuming that the data are additive in terms of error, trend, and seasonality. Finally, we also established an ETS auto model.

```
> ets_aaa_top <- ets(top_crime_ts, model = 'AAA')
> ets_aaa_top
ETS(A,A,A)

Call:
ets(y = top_crime_ts, model = "AAA")

Smoothing parameters:
alpha = 0.8028
beta = 1e-04
gamma = 0.001

Initial states:
l = 2973.9898
b = -7.4689
s = -124.3811 -74.6752 83.1911 -7.0169 88.9422 156.5102
26.3445 157.1525 -130.2471 74.1056 -168.5679 -81.438

sigma: 179.1143

AIC AICc BIC
1070.901 1082.234 1109.604

sigma^2 = 38350: log likelihood = -474.99
AIC=955.99 AICc=956.35 BIC=962.78
```

```
> ets_model_top_crime <- ets(top_crime_ts)
> ets_model_top_crime
ETS(A,N,A)

Call:
ets(y = top_crime_ts)

Smoothing parameters:
alpha = 0.8211
gamma = 1e-04

Initial states:
l = 2772.4533
s = -148.7577 -108.1422 82.3699 -2.7379 71.8235 166.9825
33.3375 174.6664 -119.6308 64.0678 -180.9864 -32.9927

sigma: 172.839

AIC AICc BIC
1064.292 1072.863 1098.442
```

```
> arima_model_top_crime <- auto.arima(top_crime_ts)
> arima_model_top_crime
Series: top_crime_ts
ARIMA(1,1,0)(1,0,0)[12]

Coefficients:
ar1 sar1
-0.3627 0.2848
s.e. 0.1225 0.1270
```

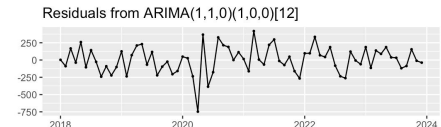
Based on the results, we find that ARIMA's model has the best performance. It has the smallest AIC, AICc and BIC. However, the three indicators of these three models are relatively close to each other. It is hard to say decisively that ARIMA's model is definitely optimal. Thus, we need to further compare the Ljung-Box test results with the 1st-order and 14th-order lags of the three models.

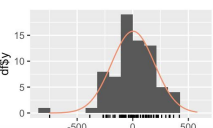
```
> checkresiduals(arima_model_top_crime)

Ljung-Box test

data: Residuals from ARIMA(1,1,0)(1,0,0)[12]
Q* = 17.839, df = 12, p-value = 0.1207

Model df: 2. Total lags used: 14
```





```
> Box.test(arima_model_top_crime$residuals, type = "Ljung-Box")

Box-Ljung test

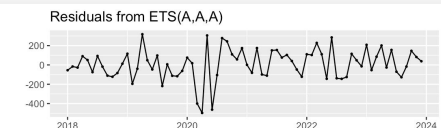
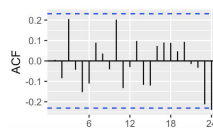
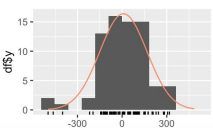
data: arima_model_top_crime$residuals
X-squared = 0.021554, df = 1, p-value = 0.8833
```

```
> checkresiduals(ets_aaa_top)

Ljung-Box test

data: Residuals from ETS(A,A,A)
Q* = 14.988, df = 14, p-value = 0.379

Model df: 0. Total lags used: 14
```

```
> Box.test(ets_aaa_top$residuals, type = "Ljung-Box")

Box-Ljung test

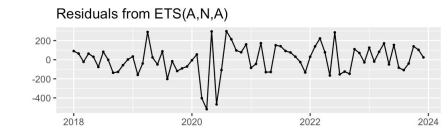
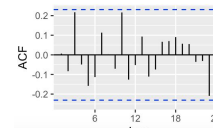
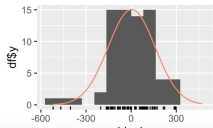
data: ets_aaa_top$residuals
X-squared = 0.002854, df = 1, p-value = 0.9639
```

```
> checkresiduals(ets_model_top_crime)

Ljung-Box test

data: Residuals from ETS(A,N,A)
Q* = 16.436, df = 14, p-value = 0.2875

Model df: 0. Total lags used: 14
```

```
> Box.test(ets_model_top_crime$residuals, type = "Ljung-Box")

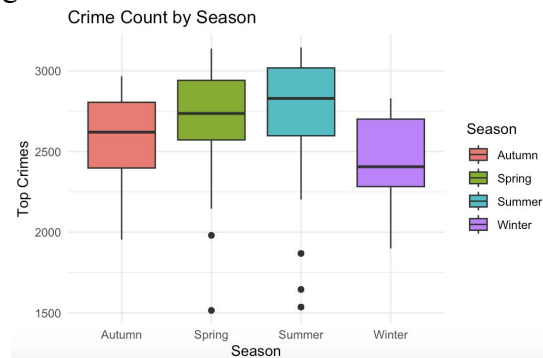
Box-Ljung test

data: ets_model_top_crime$residuals
X-squared = 0.003221, df = 1, p-value = 0.9547
```

Based on the results, we find that the p-value for all three models is greater than 0.05. This indicates that the residuals are white noise at the 95% confidence level. This suggests that all three model models adequately capture the time-dependent patterns. However, the premise

assumptions of all three models are different. For instance, ARIMA assumes that the data is stationary, while both ETS models assume that seasonality is additive. To some extent, this implies that the number of assaults may have seasonality.

Next, we draw the box plot. Based on the picture, we can see that the number of assaults seems to be higher in spring and summer than in fall and winter.



Hence, to investigate the relationship between the number of assaults and seasonality, we conduct further analysis by implementing the statistical test. To be specific, we now divide the 12 months of the year into four seasons and then perform an ANOVA test.

```
> summary(anova_result)
          Df    Sum Sq   Mean Sq F value Pr(>F)
Season      3  1.250e+10  4.167e+09   38880 <2e-16 ***
Residuals 1173952  1.258e+11  1.072e+05
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

According to the ANOVA test result, the p-value is less than 0.05, representing enough evidence to reject the null hypothesis at a 95% confidence interval. In other words, there is an association between the number of assaults and seasons.

Results

In our analysis of New York City crime data from 2018 to 2023, we validated two research questions with ANOVA tests that showed a p-value of less than 0.05 for both. This statistical result supports our decision to reject the null hypothesis at a 95% confidence level, showing a significant association between the total number of crimes and assaults in New York City and the season. This finding confirms that crime rates vary significantly from season to season. Especially in spring and summer, crime rates show an upward trend compared to autumn and winter. The trend is consistent with previous research showing that warmer months may lead to higher crime rates due to increased social interaction.

Future recommendations and policy initiatives

Longitudinal Analysis

Continuously monitoring crime trends over time can provide a deeper understanding of seasonal patterns and their underlying drivers. Future studies could conduct longitudinal analyses to track changes in crime rates and assess the effectiveness of intervention measures.

Community Partnerships

Strengthening partnerships between law enforcement agencies and local communities is essential for fostering trust and collaboration in crime prevention efforts. Initiatives such as community policing and neighborhood watch programs can enhance community resilience and promote active participation in crime reduction initiatives.

Predictive Analytics

Embracing predictive analytics and machine learning algorithms can enhance the predictive capabilities of law enforcement agencies. By leveraging historical crime data and environmental factors, predictive models can proactively forecast future crime hotspots and guide resource allocation strategies.

Policy Evaluation

Evaluating the impact of existing policies and interventions is critical for refining strategies and maximizing resource efficiency. Future research should focus on assessing the effectiveness of different policing strategies in reducing crime rates and enhancing public safety outcomes.

limitation and improvements

Limitations of seasonal studies

This study focused primarily on the relationship between seasonal factors and crime but may have overlooked other important factors. For example, the study showed that between the beginning of 2018 and the end of 2023, the number of crimes first decreases and then increases, which suggests that we should further explore long-term trends in the number of crimes in future studies.

Omitted variables

This study did not consider other variables affecting crime rates, such as weather conditions. Weather factors such as temperature, humidity, and precipitation may directly or indirectly impact criminal behavior. By introducing these weather variables, future studies could more accurately assess and explain the relationship between seasons and crime rates.

Effects of geography

In addition, our dataset, while covering a broad area of New York City, needs to explore in detail how different geographic locations affect patterns of crime occurrence. Factors such as socio-economic conditions, police deployment, and resident structure in different areas can significantly impact crime rates. Therefore, including geographical factors in the analytical framework will contribute to a deeper understanding of the relationship between criminal activities and environmental factors.

Conclusion

In conclusion, this study will use clustering and time series to answer these research questions. Through the analysis, police can identify age groups, regions, and periods with higher crime rates. On this basis, data on interventions can be built. It is not only expected to improve the efficiency of police allocation but also to help create a stable and safe social environment. The results and methods applied in this study can serve as a reference for other cities facing similar challenges, emphasizing the importance of data-driven policy making in public safety and crime prevention.

References

Corcoran, J., & Zahnow, R. (2022). Weather and crime: A systematic review of the empirical literature - crime science. BioMed Central.

<https://crimesciencejournal.biomedcentral.com/articles/10.1186/s40163-022-00179-8>

Farley, R. (2023). The facts on Manhattan crime. FactCheck.org.

<https://www.factcheck.org/2023/04/thefacts-on-manhattan-crime/>

Hanna, W. (2023). Stemming New York's crime surge. City Journal. (2023).

<https://www.cityjournal.org/article/stemming-new-yorks-crime-surge>

NYC Open Data. (2024a). NYPD arrests data (historic). https://data.cityofnewyork.us/Public-Safety/NYPD-Arrests-Data-Historic-/8h9b-rp9u/about_data

NYC Open Data. (2024b). NYPD arrest data (year to date). https://data.cityofnewyork.us/Public-Safety/NYPD-Arrest-Data-Year-to-Date-/uip8-fykc/about_data