36315 Final Project

Jade Huang, Peile Li, Ruiwen Liu, Jiashen Wang

May 2, 2023

Contents

Introduction

Crime poses a major challenge in numerous metropolitan areas worldwide, and New York City is not exempt. Understanding the distribution of crime in the city, as well as the characteristics of the victims and suspects, is crucial for law enforcement officials and policymakers. In this data analysis report, we will explore the crime rates in New York City and investigate the factors that contribute to crime in the city.

Our report will focus on three primary research questions. Firstly, we will examine the distribution of crimes across the city and identify which borough is the most dangerous and which is the safest. We will also analyze the most common types of offenses in each borough. Secondly, we will investigate how the number of crimes has changed from 2018 to 2021, which can help identify any patterns or trends in crime rates. Finally, we will look at the characteristics of the groups of victims and suspects, including who is most vulnerable and if there is a relationship between the characteristics of suspects and victims.

We use the dataset from Kaggle to answer the proposed questions. By analyzing the data, we hope to help people better understand crime patterns and trends in NYC, and to provide insights that can be used in creating effective strategies to reduce crime in New York City.

Data Description

Our dataset contains the crimes committed from 2004-2021 in New York. In our analysis, we only used data from 2018 to 2021 and select only female and male for the victim and suspect sex. After the data processing, our dataset has 775,028 rows, each represents a criminal complaint in NYC and includes information about the type of crime, the location and time of enforcement. There are 12 variables. Below are the descriptions of the variables:

1. `CMPLNT_FR_DT`: Exact date of occurrence for the reported event (or starting date of occurrence, if CMPLNT_TO_DT exists).

2. `LAW_CAT_CD`: Level of offense: felony, misdemeanor, violation.
3. `BROR_NM`: The name of the borough in which the incident occurred.
4. `PREM_TYP_DESC`: Specific description of premises; grocery store, residence, street, etc.
5. `SUSP_AGE_GROUP`: TSuspect's Age Group.
6. `SUSP_RACE`: Suspect's Race Description
7. `SUSP_SEX`: Suspect's Sex Description.
8. `Latitude`: Midblock Latitude coordinate for Global Coordinate System, WGS 1984, decimal degrees (EPSG 4326).
9. `Longitude`: Midblock Longitude coordinate for Global Coordinate System, WGS 1984, decimal degrees (EPSG 4326).
10. `VIC_AGE_GROUP`: Victim's Age Group.
11. `VIC_RACE`: Victim's Race Description.
12. `VIC_SEX`: Victim's Sex Description (F=Female, M=Male).

Research Questions

In this project, we want to answer the following three research questions:

1. **How are crimes distributed across New York?**

- In what areas of NYC do crimes happen the most/least frequently?
- What is the most common level of offense in each borough?

2. **How had the number of crimes in NYC changed from 2018 to 2021?**

- What is the general trend of the number of crimes during this time period?
- How did crimes by borough / level of offence change during this time period?

3. **What are the characteristics of the groups of victims and suspects?**

- What groups of people are the most vulnerable?
- Is there a relationship between the characteristics of suspects and victims?

Visualization and Statistical Analysis

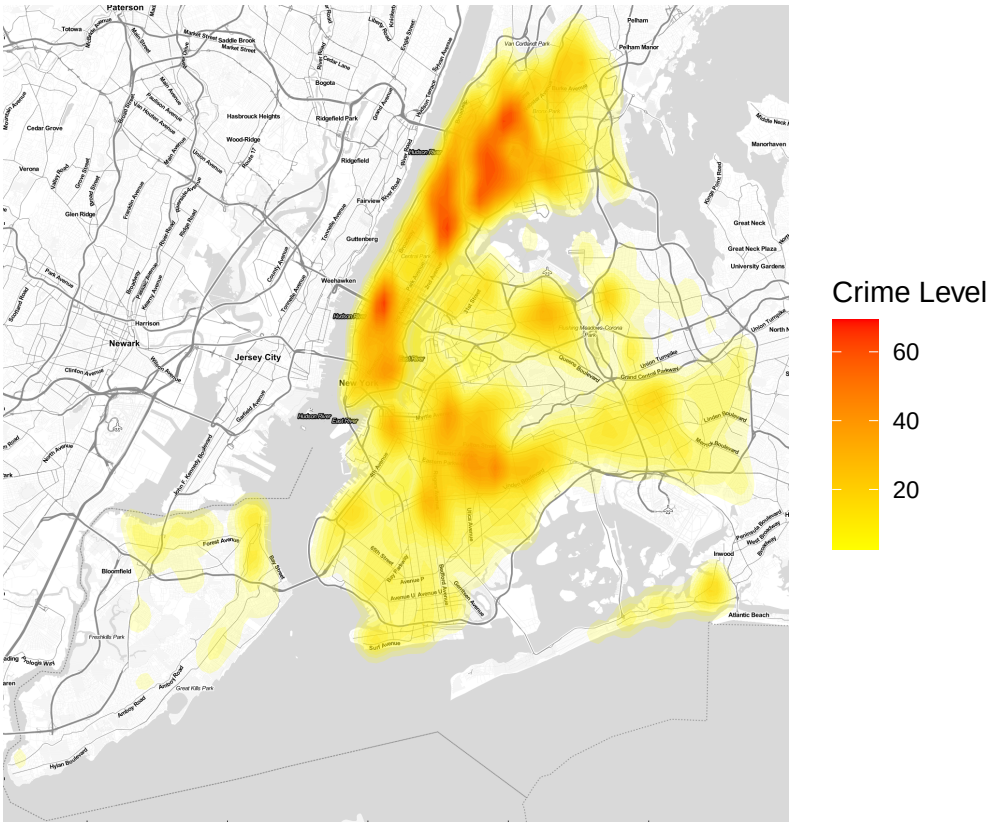
Research Question 1: How are crimes distributed across New York City?

First of all, we want to learn about where crimes happen the most/least frequently. The variables we mainly analyzed here are `Latitude`, `Longitude`, `BROR_NM`, and `PREM_TYP_DESC`,

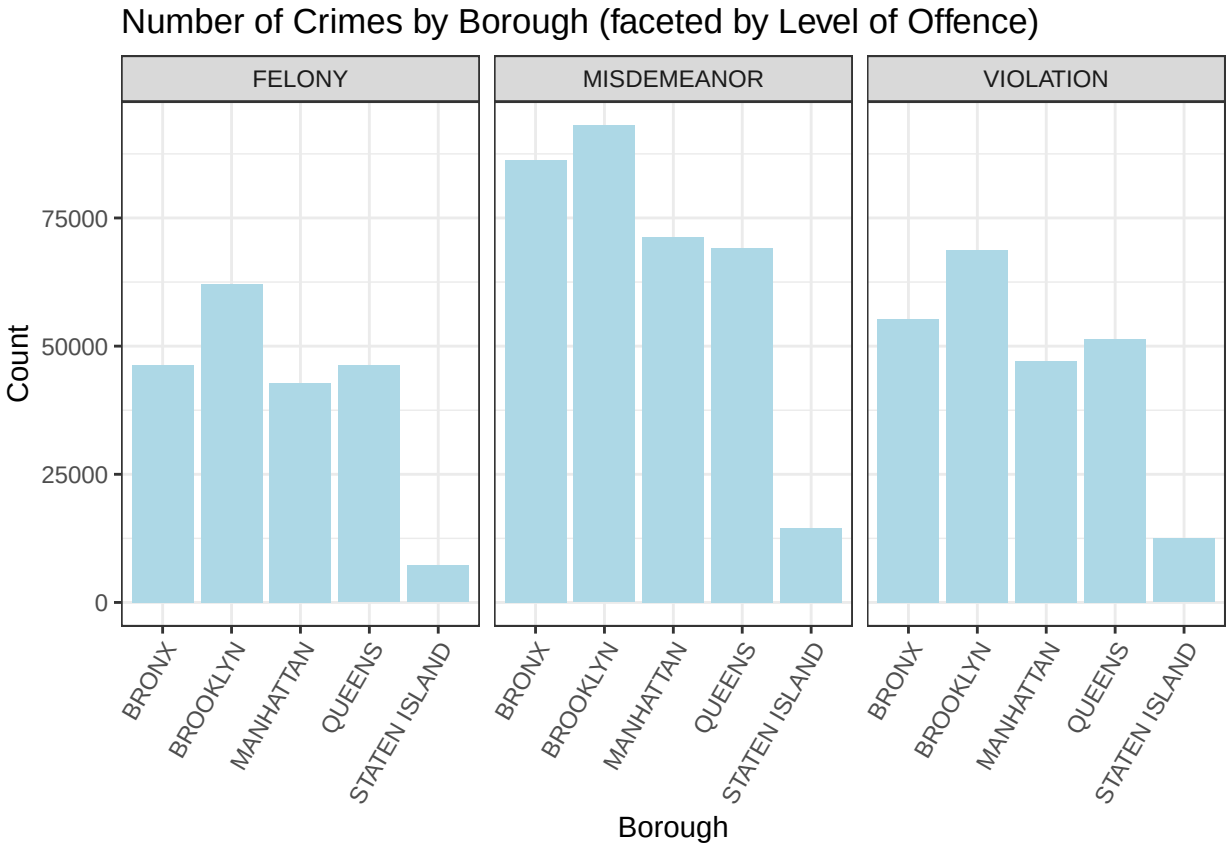
Word Cloud of Crime Location Keywords

Our first plot is a word cloud showing the specific location description keywords of crimes. The variable being analyzed in the word cloud is 'PREM_TYP_DESC'. In the word cloud above, we can see that the most frequent key words are "residence", "house", and "apt". The less frequent key words are "street", "store", "public", and "residence house". From this word cloud, we learned that the places where crimes happen most frequently are in residences, houses and apartments, while it is relatively less common to see crimes in restaurants, gyms, playgrounds, or parks. In general, in-door and private spaces are more susceptible to crimes compared to public facilities in NYC.

Crime Density Map of NYC



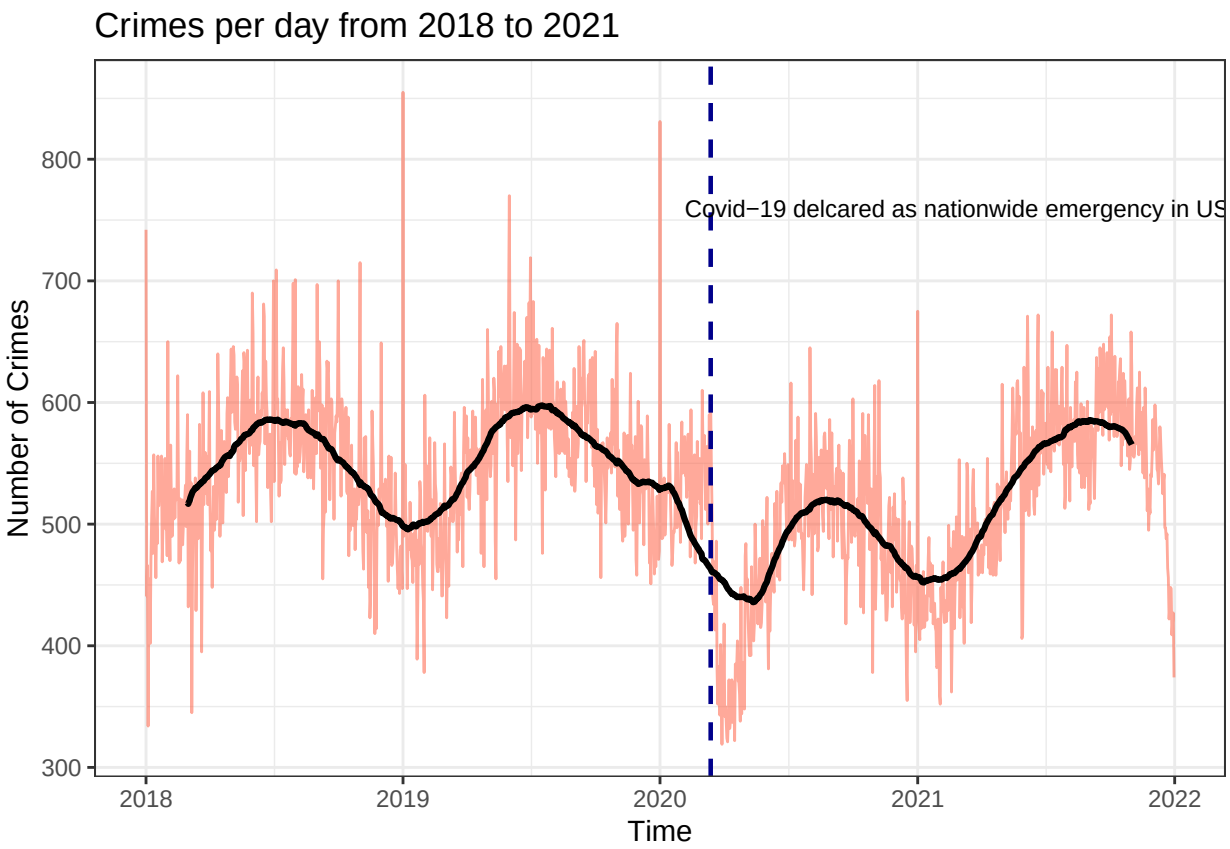
In the density map, we can see the frequency of crime across New York City. The areas in which crimes take place most frequently are Midtown/Downtown Manhattan, Uptown Manhattan, Bronx, and Brooklyn. These areas also happen to be densely populated. The darkest spot on the heat map is Midtown Manhattan, which is famous for the Times Square and millions of tourists every year. The relatively safer areas are Queens, Staten Island, and East New York City. In general, there are more crimes in the north of NYC than in the south. One interesting observation here is that Manhattan is covered in dark red and orange except the central area, which corresponds to Central Park East/West. One possible explanation is that a park itself is safer than populated areas, as identified from the word cloud above. Another plausible reason is that the many of residences around the Central Park are upscale and the residents are also relatively well-off.



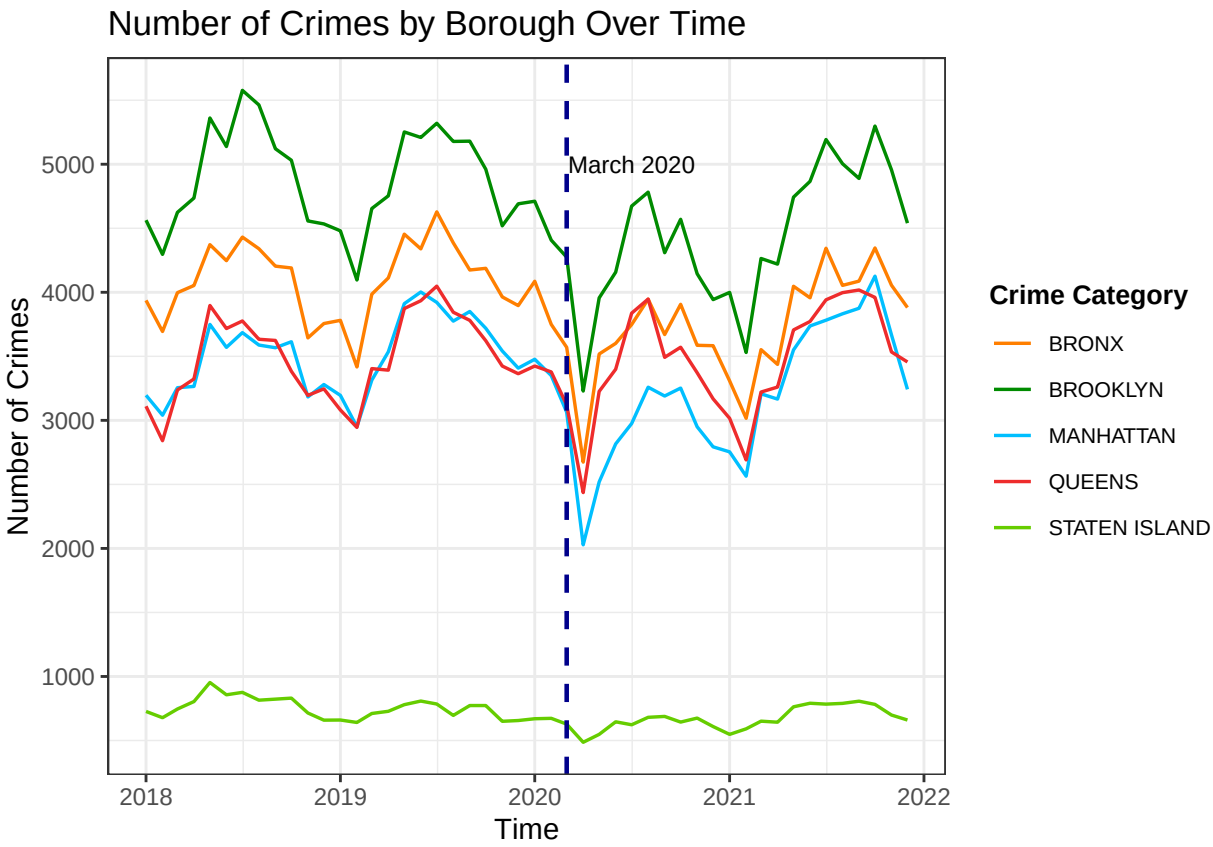
In the bar plot above, we can see the distribution of each type of crime across different boroughs. Brooklyn has the highest number of crimes in general, followed by Manhattan. Bronx and Queens have similar numbers of crime. Queens has a higher number of felonies, while Bronx has higher numbers of misdemeanors and violations. Staten Island, which is scarcely populated and far away from the downtown area, has the least numbers of crimes in every level. Across all boroughs, there are more incidents of misdemeanors, followed by felonies and then violations.

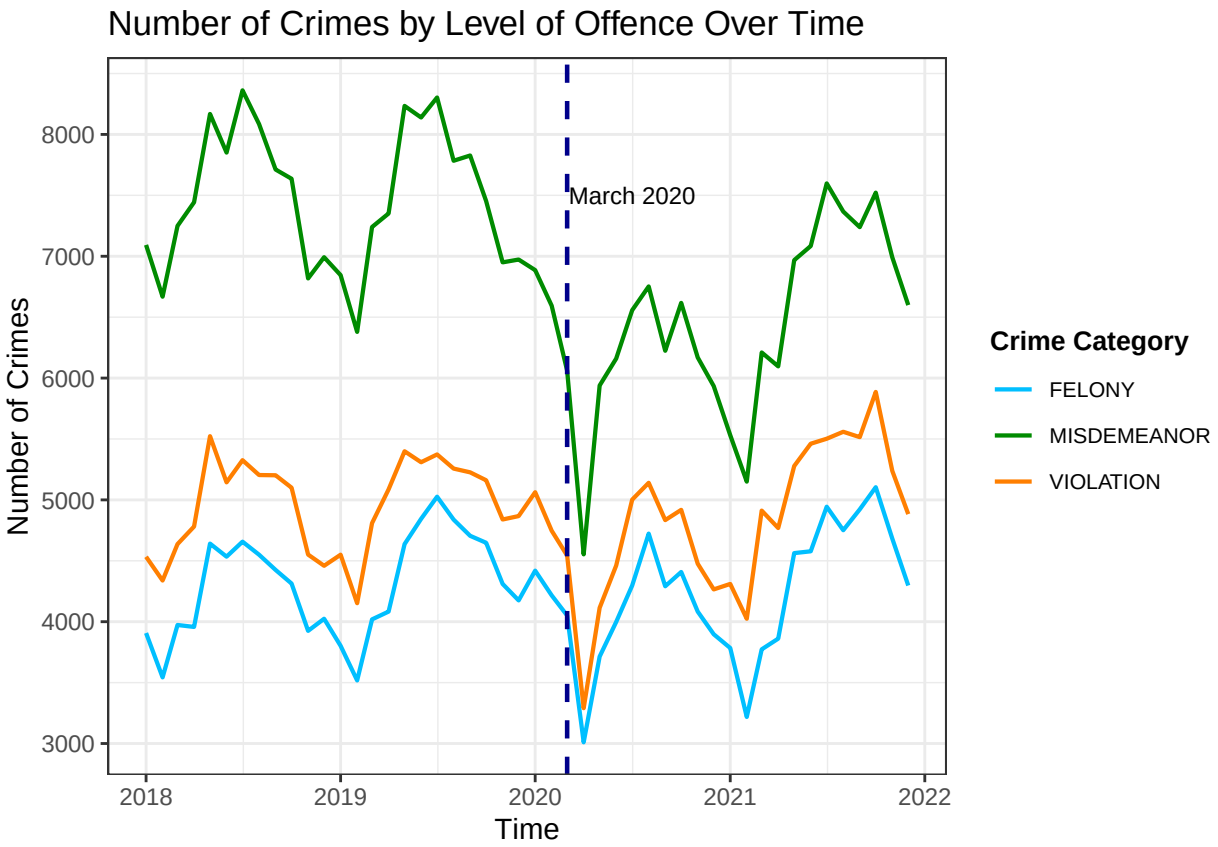
Research Question 2: How had the number of crimes in NYC changed from 2018 to 2021?

After looking at the location distribution of the crimes, we wanted to investigate the trend of crimes in NYC over time. In this sections, we focused on the variable `CMPLNT_FR_DT`. Three new dataframes were created. The first is grouped by `CMPLNT_FR_DT`, which contains one observation for each day from 2018 to 2021. It is being used to plot time series data. A new variable `yearmon` is created for the second and third dataframe. It encompasses the month and year information of `CMPLNT_FR_DT`. The second dataframe is grouped by `yearmon` and `BROR_NM`, while the third is grouped by `yearmon` and `LAW_CAT_CD`.



This plot shows the moving average of the number of crimes per day from 2018 to 2021, with the time series marked in orange in the background. The trends of the moving average curves in 2018 and 2019 are highly similar. Both years peak in the middle of the year and decrease toward the end of the year. The dashed line marks the day "2020/03/13" when the Trump Administration declared COVID-19 as a nationwide emergency in the US. There has been a significant drop in the number of crimes after entering 2020, especially following the first several months of the COVID-19 pandemic. The overall level of crime throughout 2020 is notably lower than in the previous two years. The total number of crimes will increase again in 2021, matching the levels of pre-pandemic times.


Number of Crimes by Borough Over Time
March 2020
5000
4000
3000
2000
1000
Number of Crimes
2018
2019
2020
2021
2022
Time
Crime Category
BRONX
BROOKLYN
MANHATTAN
QUEENS
STATEN ISLAND



These two plots are plotted using the second and third dataframe in this section respectively. They show the number of crimes by borough and by level of offense in NYC from 2018 to 2021. Both plots contain a dashed line that marks March 2020, the month when COVID-19 started in the US.

In the first plot, all boroughs demonstrate similar trends over time: periodic in the years 2018 and 2019 followed by a steep drop after the start of the pandemic. Brooklyn always had the most number of crimes, while Staten Island had the least. The trends of Queens and Manhattan are almost identical except in 2020 when Manhattan had a notably lower number of crimes. One plausible explanation is that the total number of people in Manhattan dropped as companies could not operate in person.

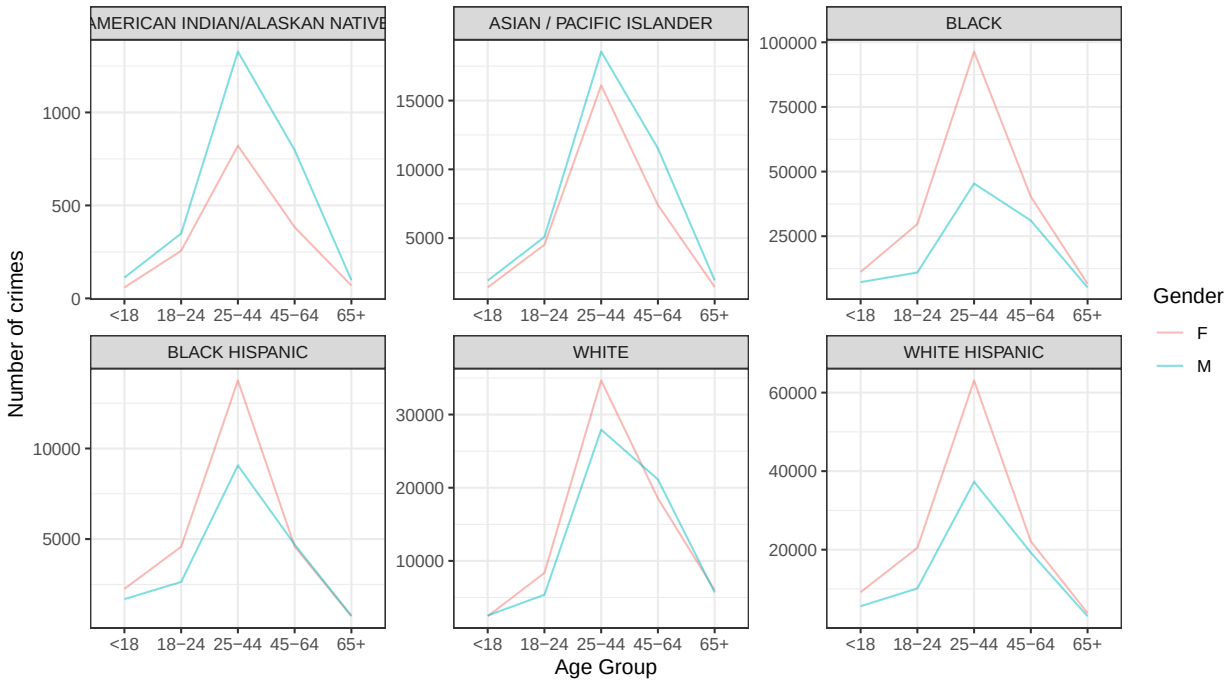
The second plot is similar to the first one in the sense that all curves follow similar trends over time. However, the three crime categories differ in their post-pandemic trends. The number of felony crimes bounced back to pre-pandemic levels in 2021. The number of violations spiked to a level that exceeds 2018 or 2019 in the latter half of 2021. Misdemeanors saw the most significant decrease among the three during COVID-19. It failed to match its pre-pandemic levels in 2021.

Research Question 3: What are the characteristics of the groups of victims and suspects?

After looking at the location and change over time, we wanted to find out what the victims and suspects are like in NYC. In this section, in order to explore the relationship between the characteristics of suspects and victims, we create a subset that contains the following variables:

1. `<18'`: 1: If the suspect's age is smaller than 18, 0: Otherwise.
2. `18-24'`: 1: If the suspect's age is between 18 and 24 (inclusive), 0: Otherwise.
3. `25-44'`: 1: If the suspect's age is between 25 and 44 (inclusive), 0: Otherwise.
4. `45-64'`: 1: If the suspect's age is between 45 and 64 (inclusive), 0: Otherwise.

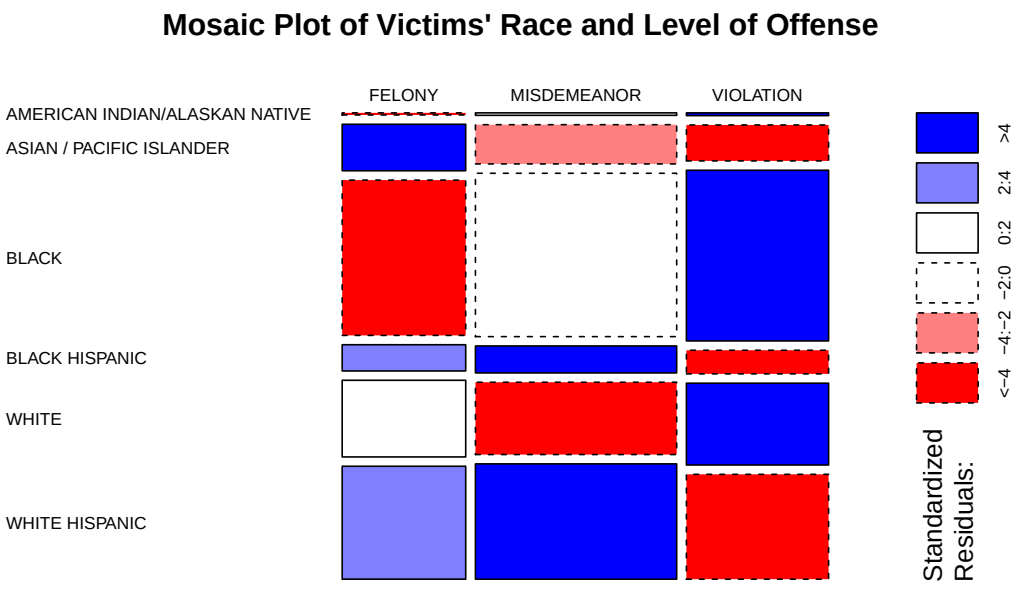
5. `65+'`: 1: If the suspect's age is greater or equal to 65, 0: Otherwise.
6. `Male'`: 1: If the suspect is Male, 0: Otherwise.
7. `Female'`: 1: If the suspect is Female, 0: Otherwise.
8. `Native'`: 1: If the suspect is native American, 0: Otherwise.
9. `Asian'`: 1: If the suspect is Asian, 0: Otherwise.
10. `Black'`: 1: If the suspect is Black, 0: Otherwise.
11. `BlackHispanic'`: 1: If the suspect is Black Hispanic, 0: Otherwise.
12. `White'`: 1: If the suspect is White, 0: Otherwise.
13. `WhiteHispanic'`: 1: If the suspect is White Hispanic, 0: Otherwise.
14. `<18'`: 1: If the victim's age is smaller than 18, 0: Otherwise.
15. `18-24`: 1: If the victim's age is between 18 and 24 (inclusive), 0: Otherwise.
16. `25-44`: 1: If the victim's age is between 25 and 44 (inclusive), 0: Otherwise.
17. `45-64`: 1: If the victim's age is between 45 and 64 (inclusive), 0: Otherwise.
18. `65+`: 1: If the victim's age is greater or equal to 65, 0: Otherwise.
19. `Male`: 1: If the victim is Male, 0: Otherwise.
20. `Female`: 1: If the victim is Female, 0: Otherwise.
21. `Native`: 1: If the victim is native American, 0: Otherwise.
22. `Asian`: 1: If the victim is Asian, 0: Otherwise.
23. `Black`: 1: If the victim is Black, 0: Otherwise.
24. `BlackHispanic`: 1: If the victim is Black Hispanic, 0: Otherwise.
25. `White`: 1: If the victim is White, 0: Otherwise.
26. `WhiteHispanic`: 1: If the victim is White Hispanic, 0: Otherwise.



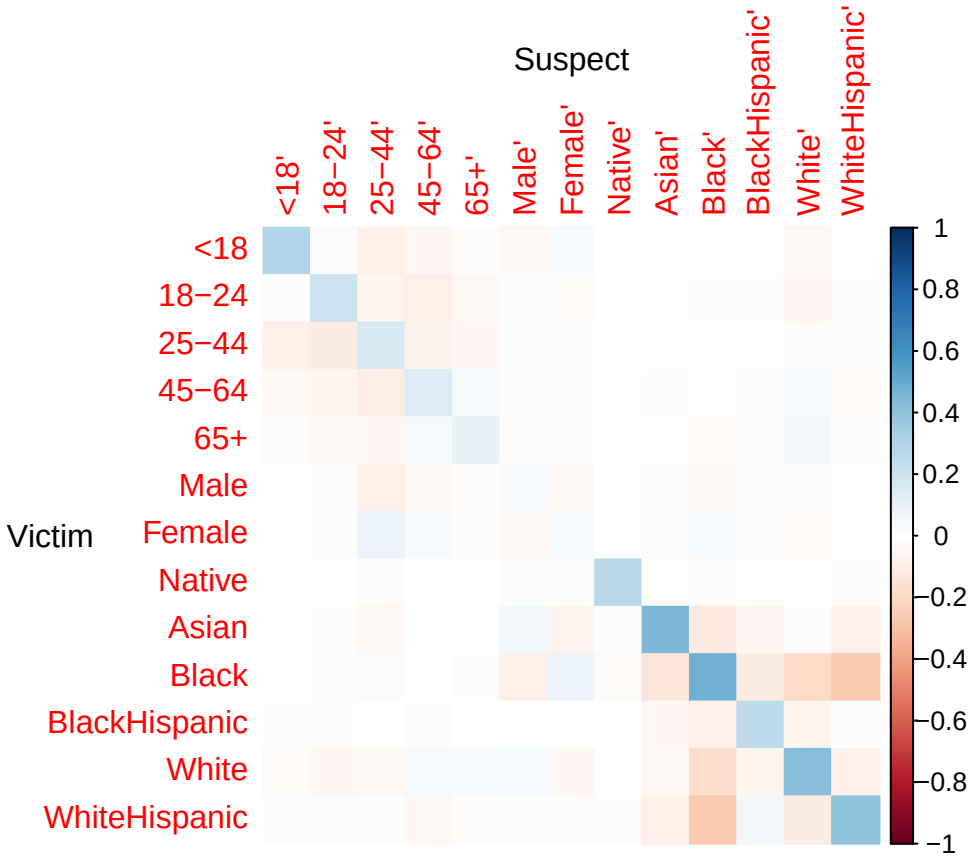
We began with an EDA plot that shows the distribution of victims' age groups and sex faceted by race. Among victims with a known race, the numbers of females and males are about the same among Black Hispanic and White. For American Indian and Asian, male victims are more than female victims. For Black and White Hispanic, female victims are more common. Among all races, 25-44 is the most common age group of victims. This plot is informative for our third research question which focuses on the relationship between the offense and the age, race or gender of the victim. This EDA helps us explore the joint distributions of victims' age, race, and gender.

##

```
##  Pearson's Chi-squared test
##
## data:  crime$LAW_CAT_CD and crime$VIC_RACE
## X-squared = 2018.8, df = 12, p-value < 2.2e-16
```



To further investigate the relationship between the victims' characteristics, we produced a mosaic plot and used a Pearson's Chi-squared test. The mosaic plot shows the marginal distribution of the level of offense and the distribution of victims' race conditioned on the level of offense. We can observe that misdemeanor is the most common level of offense, followed by violation. Felony is the least common type. Among felony-level crimes, the number of Asian, Black Hispanic, and White Hispanic victims are significantly higher than expected, and the number of Native and Black victims is lower than expected. Among misdemeanor-level crimes, the number of Black Hispanic and White Hispanic victims is significantly higher than expected, whereas the number of Asian and White victims is lower than expected. Among violation-level crimes, the number of Native, Black, and White victims are significantly higher than expected, while the number of Asian, Black Hispanic, and White Hispanics is lower than expected. In the corresponding Pearson's Chi-squared test, the p value is $< 2.2e\text{-}16$. Both the mosaic plot and the Pearson's test suggest that the level of offense and the race of the victims are not independent of each other.



The correlation plot using data from the subset shows the correlation between the characteristics of suspects and victims, including their age group, gender, and race. Blue squares indicate positive correlations between the two characteristics, and red squares indicate negative correlations. We can observe some interesting results that suspects tend to target the victims in their same age, gender, and race group. One possible explanation could be that suspects know more victims in their same age, gender, and race group, and that most crimes are committed by acquaintances. This phenomenon is most common among the races Asian and Black, and least common among same genders.

Conclusion

From our analysis on the distribution of crimes in New York City, we found that the number of crimes and level of offense varied by boroughs and specific locations, highlighting the need for targeted interventions to reduce crime rates in specific neighborhoods. Additionally, analyzing the number of crimes from 2018 to 2021, we found that COVID-19 pandemic had a significant impact on the number of crimes in NYC, with a steep drop in 2020 followed by a rebound in 2021. Besides, from the analysis of the characteristics of victims and suspects, we found that the most vulnerable groups are those between 25-44 years old. Also,there was an interesting pattern that suspects tend to target victims in their same age, gender, and race group, suggesting that most crimes are committed by acquaintances. Overall, our analysis provides insights into the changing nature of crime in New York City over the past few years.

Future Research

In our analysis, we did not take a look into the different types of crimes. Thus, we can investigate the impact of COVID-19 on specific types of crimes and explore the possible reasons behind the observed changes in pre and post COVID-19 period. Also, we may need to examine more demographic characteristics of

suspects and victims in the future, such as income, education, and prior criminal records, as well as the characteristics of boroughs, such as population density, law enforcement policies, and economic conditions, to better understand how these factors contribute to the differences in crime rates. Finally, we could conduct more advanced statistical techniques to identify the key predictors of crime and to develop predictive models for crime rates in different parts of the city.