Project introduction:

The data set used consists of arrest data from the New York Police Department (NYPD) for the period of January 1st to September 30th, 2024. This data was downloaded from NYC Open Data (https://tinyurl.com/mw4ftb8u). Each entry in the dataset contains details about a specific arrest, including information about time of arrest, location, and suspect description including race and gender.

By analyzing this data, we aim to gain insights into crime patterns and law enforcement activities and explore potential disparities in policing across different boroughs and demographic groups. This study seeks to determine whether arrests are proportional to borough populations, whether federal holidays see a higher number of arrests, and whether there is an association between race and gender in Queens. The findings of this analysis may be able to provide patterns and detailed information about law enforcement practices and contribute to discussions on policing in New York City.

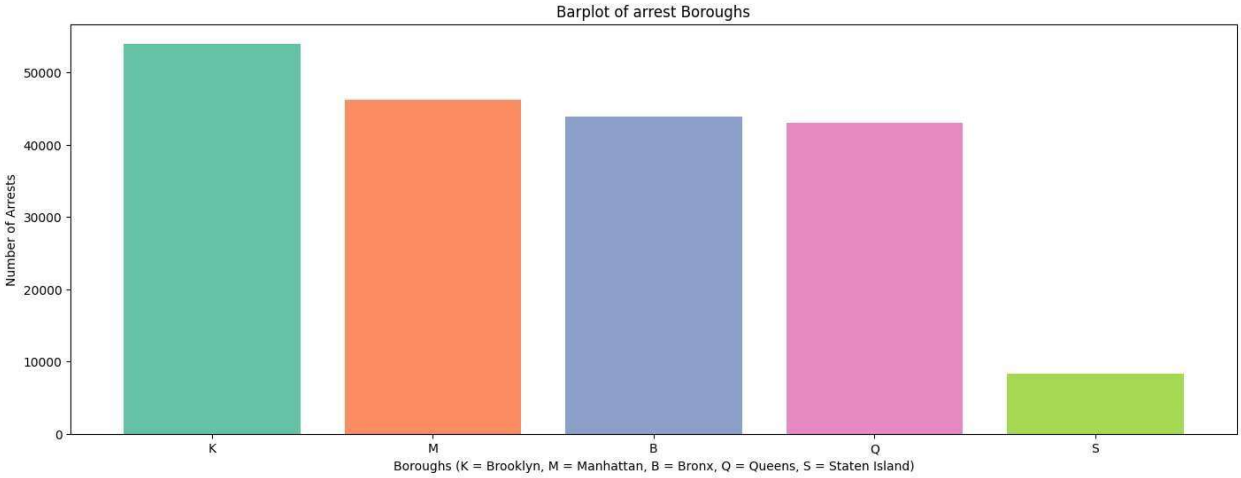
Research Questions and Hypotheses:

1. Are arrests in Bronx and Brooklyn proportional to their populations?
    a. H_0: Arrests in the Bronx and Brooklyn are proportional to their populations
    b. H_1: Arrests in the Bronx and Brooklyn are not proportional to their populations
2. Are there more arrests on Federal holidays?
    a. H_0: There is no difference in the number of arrests on federal holidays and non-federal holidays
    b. H_1: There is a difference in the number of arrests on federal holidays and non-federal holidays
3. Is there an association between gender and race (Black vs. Asian) in the population of Queens?
    a. H_0: Sex and race (Black/Asian) are independent. The proportion of males and females is the same across the racial groups.
    b. H_1: Sex and race (Black/Asian) are dependent. The proportion of males and females differs across the racial groups.

Analysis with visualizations tables and explanatory texts

1. Are arrests in Bronx and Brooklyn proportional to their populations?
    a. H_0: Arrests in the Bronx and Brooklyn are proportional to their populations
    b. H_1: Arrests in the Bronx and Brooklyn are not proportional to their populations

Visualizing the data:



Borough	Number of Arrests
Brooklyn	53994
Manhattan	46235
The Bronx	43899
Queens	42978
Staten Island	8341
Total	195447
Average	39089.4

To visualize arrest data between the boroughs of New York City a bar plot was chosen. Each borough having its distinct arrest count shown can help easily identify the higher crime rate zones. It can be seen by this graphic that Brooklyn has more arrests than the Bronx. The original question asked if there was more crime in the Bronx than Brooklyn, which is disproved.

The bar plot shows the distribution of data across all the different boroughs. The total amount of arrest was 195447. The lowest arrest count is in Staten Island at 8341 arrests, and the highest happens to be in Brooklyn at 53994 arrests. The average amount of arrest per borough is approximately 39089. The first quartile of the mount of arrests was 42978, and the third quartile was 46235. Using that the IQR can be found to be 3257. The lower bound (Q1 – (1.5*IQR)) is 38092.5, and the upper bound (Q1 – (1.5*IQR)) is 41349.5. This shows that the arrests for both Staten Island and Brooklyn would be considered outliers.

One expectation that was set was that the Bronx has a significantly higher poverty rate than Brooklyn, which leads to a higher arrest rate. This expectation led me to believe that the number of arrests would potentially be higher the Bronx than in Brooklyn. However, the graphic has changed my expectation. It seems that the poverty rate does not impact the amounts of arrests happening.

While the bar plot plainly answered the question of which borough has more arrests it does not account for any other differences between them. To further investigate the relationship between the Bronx and Brooklyn, the population was factored in to determine whether arrests in these boroughs are proportional to their populations.

Borough	Population
Brooklyn	2679620
The Bronx	1419250

Using the population data in tandem with the arrest data, a statistical test was run.

Test run:

- Chi-squared test Goodness of fit
    - Tests on whether the observed distribution a single categorical variable (Number of arrests) matches an expected distribution

Assumptions:

1. Data is categorical:
    a. Chi-squared is for categorical data
    b. The Boroughs are categories, and the number of arrests and population are counts
2. Observations must be independent
    a. Each arrest should be independent of others. An arrest in Brooklyn should not influence an arrest in the Bronx
    b. Each arrest is recorded as a separate event, and an individual can only be arrested once per crime at a single location. However, external factors such as policing strategies, and related case crime rates may introduce some level of dependence between boroughs.
        i. Because only arrest count information is available the results should be interpreted cautiously, i.e. Not making policy decisions based on results
3. Counts should be sufficiently large
    a. The expected frequency for each category should be at least 5 to ensure that the Chi-squared approximation is valid.
    b. Expected arrests were much larger than 5 (63997 for Brooklyn and 33895 for Bronx)
4. The data is a Random Sample
    a. The arrest data covered the entirety of both boroughs so as not to be biased in collection.

5. No More Than 20% of Expected Frequencies Should Be Less Than 5
    a. There are only two categories (Brooklyn and Bronx) both are above 5

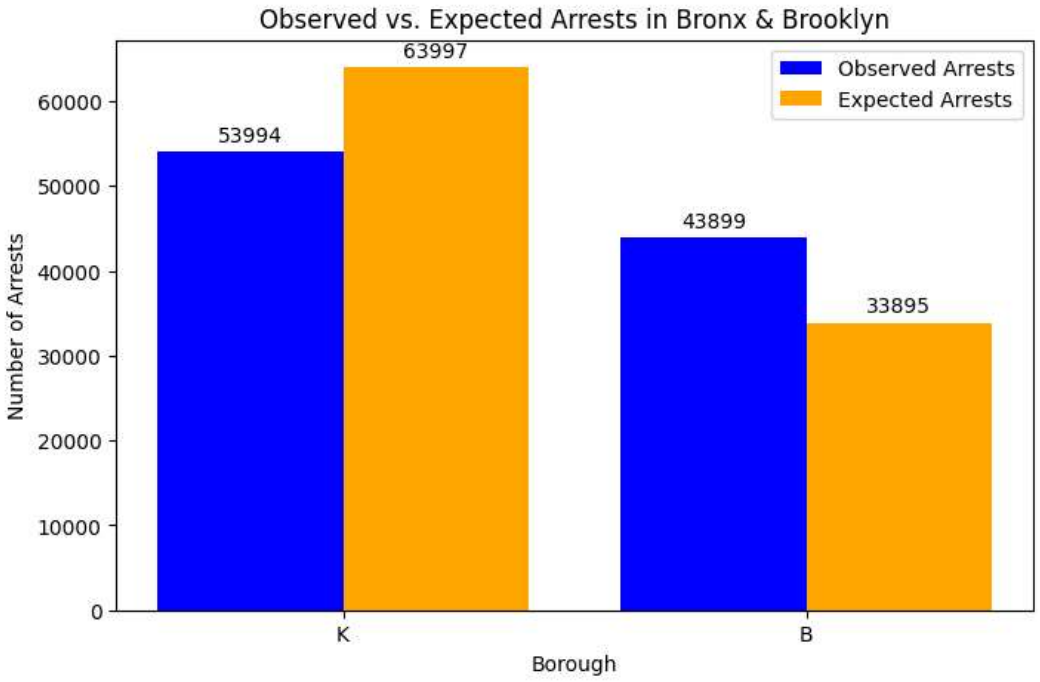
Alpha = .05

Chi-Square Statistic: 4515.636750396534

P-value: 0.0

P < alpha so **reject the null hypothesis** (H_1): Arrests are NOT proportional to population.

Visualization:



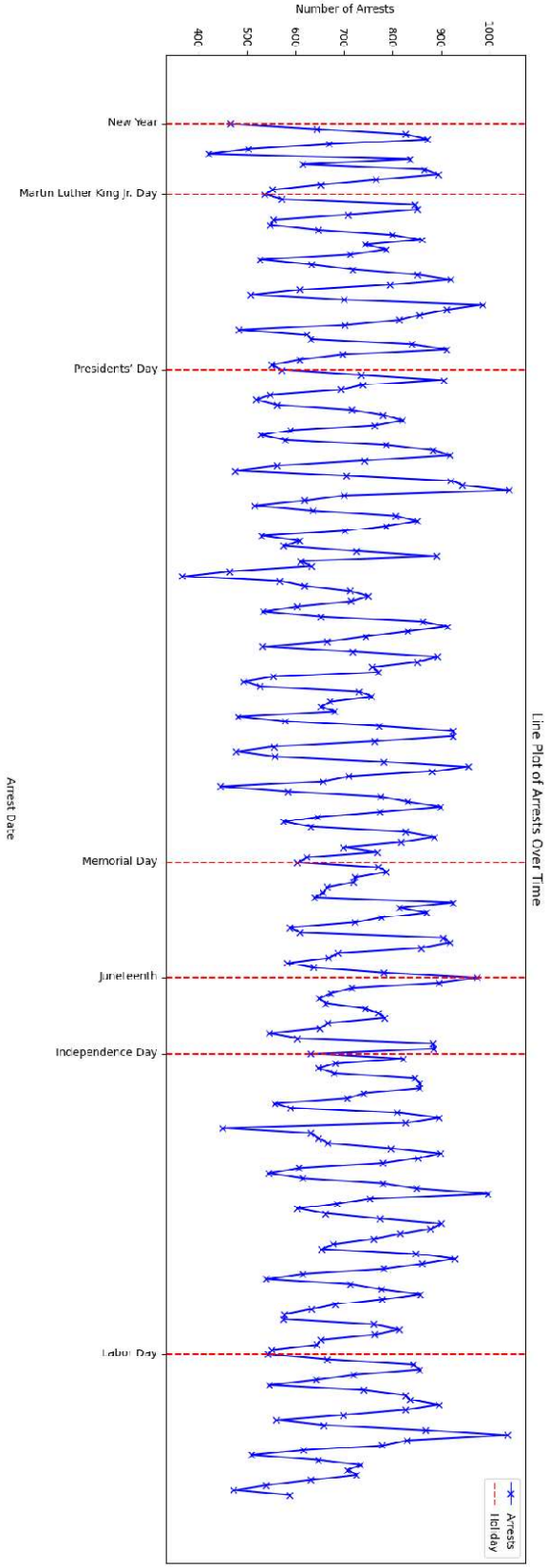
Conclusions:

From the graphic we can see that in Brooklyn 53994 arrests were observed, which is lower than the expected amount of 63997. The Bronx had 43899 arrests, which is higher than the expected amount of 33895. Brooklyn had both higher observed and expected arrest amounts.

As we reject the null hypothesis (H_0: Arrests in the Bronx and Brooklyn are proportional to their populations) other factors beyond the population size influence the arrest numbers influence arrest amounts. These could be potential disparities in enforcement practices, policy, and/or socioeconomic factors.

2. Are there more arrests on Federal holidays?
    a. H_0: There is no difference in the number of arrests on federal holidays and non-federal holidays
    b. H_1: There is a difference in the number of arrests on federal holidays and non-federal holidays

Visualization:



Statistic	Daily arrest amount	Date(s)
Min	366	2024-03-31
Max	1040	2024-03-14
Median	712.5	
Average	713.3102189781022	
Standard deviation	132.53054981820873	
Mode 1	630	2024-05-20, 2024-07-04, 2024-07-20, 2024-09-27
Mode 2	651	2024-01-13, 2024-04-08, 2024-04-26, 2024-08-30
Mode 3	855	2024-02-08, 2024-07-10, 2024-07-11, 2024-09-05
Q1	614	
Q3	824.75	

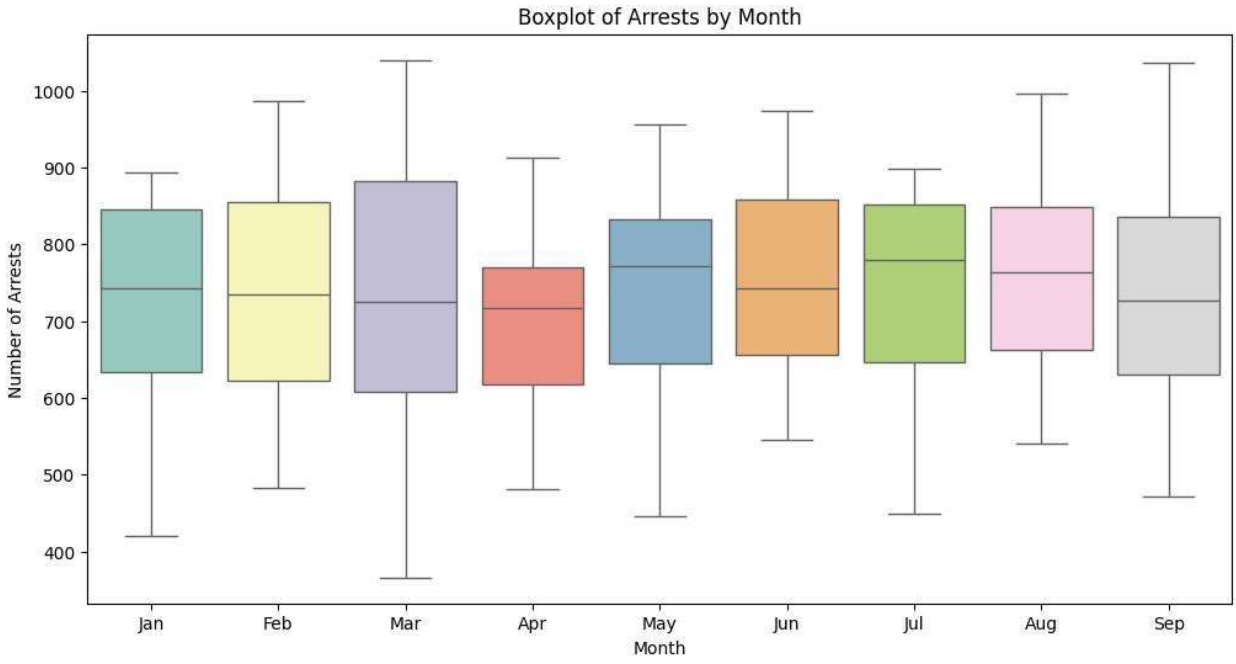
The graphic above is a line plot with each date's arrest amount for the entire city of New York. Federal holidays are marked out and labeled on the x axis. The original question posed (are there more arrests on Federal holidays?) is disproved here except for Juneteenth having a spike in arrests. It is also seen on the graphic that there are relative minimums happening at most of the federal holidays except for presidents' day and Juneteenth.

The highest daily arrest in the city was made on 2024-03-14 at 1040 arrest and the lowest amount was 366 on 2024-03-31. The IQR is 210.75. The upper bound is 1140.875 and the lower bound is 297.875. Using that information there are no outliers in the arrests over time. The standard deviation is 132.53054981820873. There happens to be three modes 630, 651, and 855 daily arrests. The first mode occurs once on the federal holiday of July 4th.

The expectation that was made before creating the graphic was that on federal holidays people do not have work so they would have more time for illicit and illegal activity. This led me to believe that there would generally be spikes on the holidays in arrest amounts. The graphic has led me to change my expectation. The one exception here is the spike on Juneteenth. Juneteenth is the commemoration of the end of slavery in the US. Another expectation that could be an explanation is that African Americans are disproportionately arrested compared to other races. Juneteenth is a celebration of African American freedom so

because of festivities perhaps more arrests were made. I would like to investigate further into seeing the races that were arrested at this time.

Presidents' day and Juneteenth happen to not occur at relative minimums in the data set. It is worth noting though presidents' days daily arrest occurs the day after a relative minimum. Juneteenth happens at a relative maximum. A question that I would like to investigate further is about the celebrations that occur on these holidays that may lead to this phenomenon.



The above graphic is a boxplot of monthly arrest. It helps establish a further understanding of the line plot of daily arrest. It shows the median amount of arrest as well as the upper and lower quantiles. In addition to that the monthly minimums and maximums are shown on the whiskers of the plots.

Some interesting things in this plot are that the median amount of arrest that occurred for the first four months are very close to each other. The 3rd quartiles for January, February, June, July,

August and September are all very close to one another. There appears to be no outliers for any of the months. The month with the lowest minimum appears to be March. March also happened to have the largest maximum number of arrests.

Some expectations I had here were that the months with holidays would have a higher number of arrests, and possibly outliers that occurred. This expectation was proven to be incorrect, in the fact that there are no outliers that occurred in this plot.

I would like to further inquire about other factors that may impact the sample monthly statics like how the number of non-federal holidays in a month could have impacted results.

To determine whether there are more arrests on federal holidays and to test the null hypothesis (H_0: There is no difference in the number of arrests on federal holidays and non-federal holidays), a statistical test was conducted.

Test run:

- Two sample t-test
    - The two-sample t-test allows you to assess if there is a statistically significant difference between groups.

Results:

T-Statistic: -1.5537383022050255

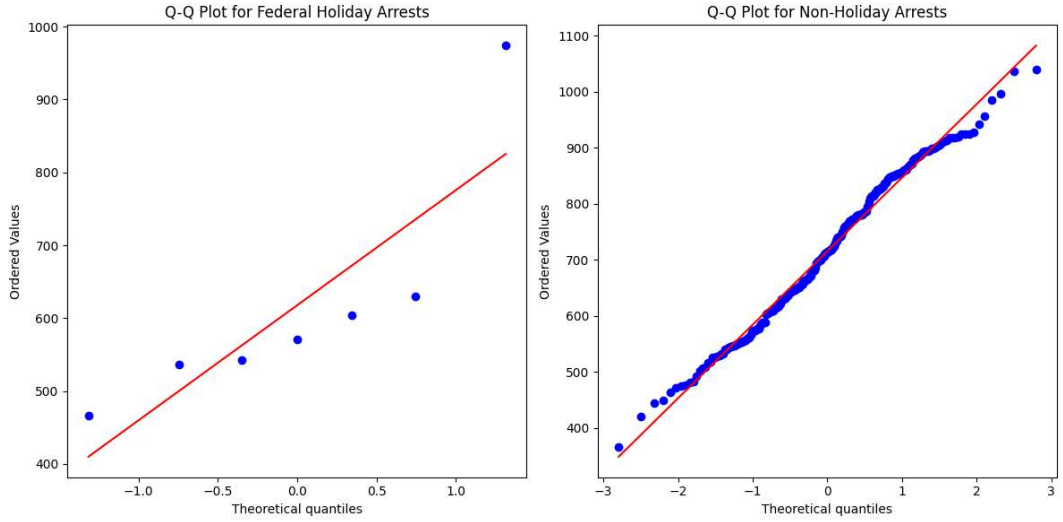
P-Value: 0.16966995710225058

Alpha: .05

P > alpha so we **fail to reject the null hypothesis** (H_0): No significant difference in the number of arrests on holidays vs. non-holidays.

Assumptions:

1. The samples from Federal holidays and non-holidays are independent of each other. Arrests on consecutive days did not influence each other. The holidays are completely different days, and there were no repeated days
2. Both groups (holiday and non-holiday arrest counts) are approximately normally distributed:



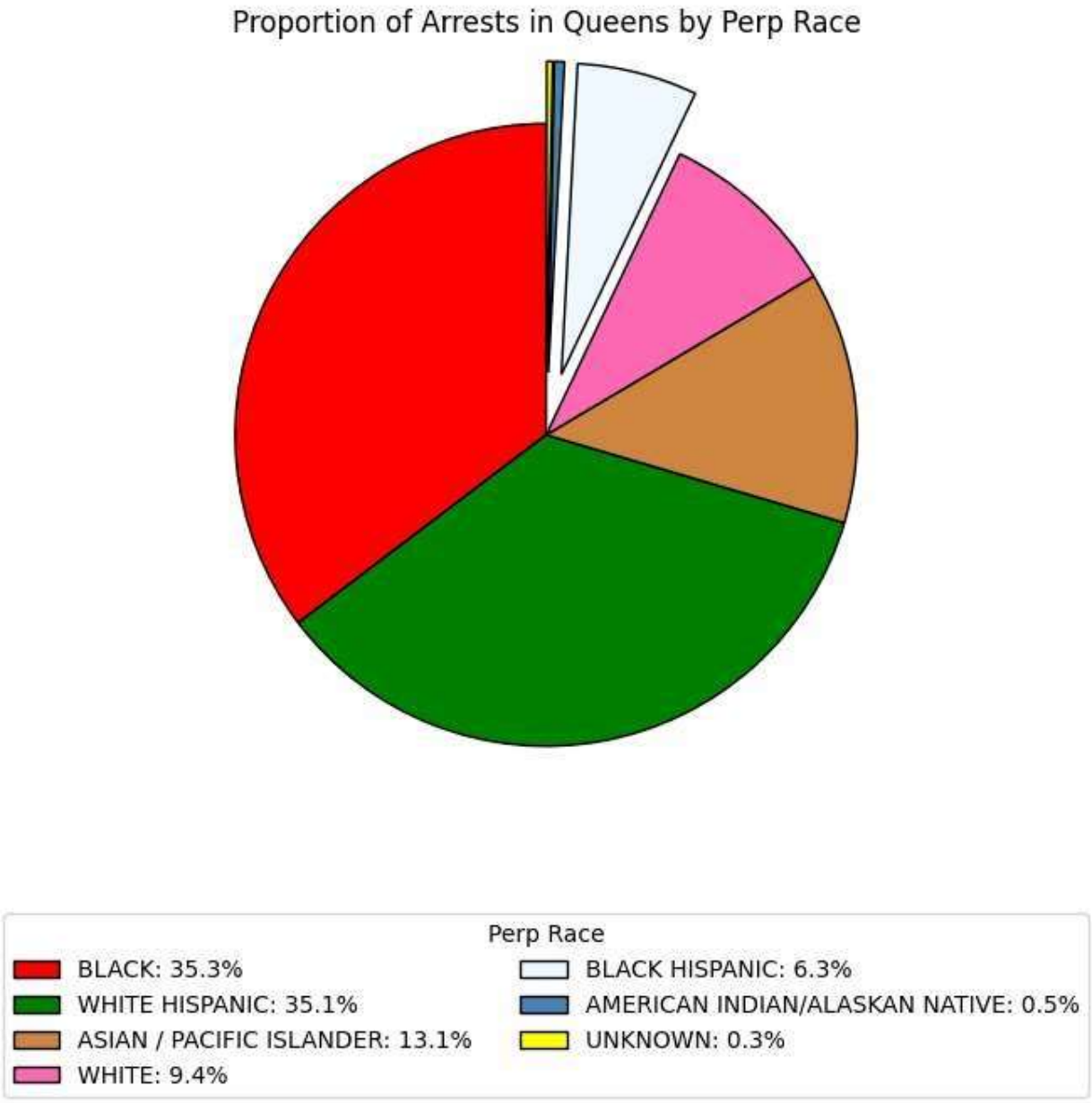
Note: The QQ plot for federal holidays shows that the data does not really follow a normal distribution. This means that the results of this test are probably not very accurate – policy decisions shouldn't be made with the current test results.

Conclusions:

As we fail to reject the null hypothesis, We conclude that there is no statistical evidence to suggest that federal holidays impact the number of arrests.

3. Is there an association between gender and race (Black vs. Asian) in the arrests of Queens?
    a. H_0: Sex and race (Black/Asian) are independent. The proportion of males and females is the same across the racial groups.
    b. H_1: Sex and race (Black/Asian) are dependent. The proportion of males and females differs across the racial groups.

Visualizations:

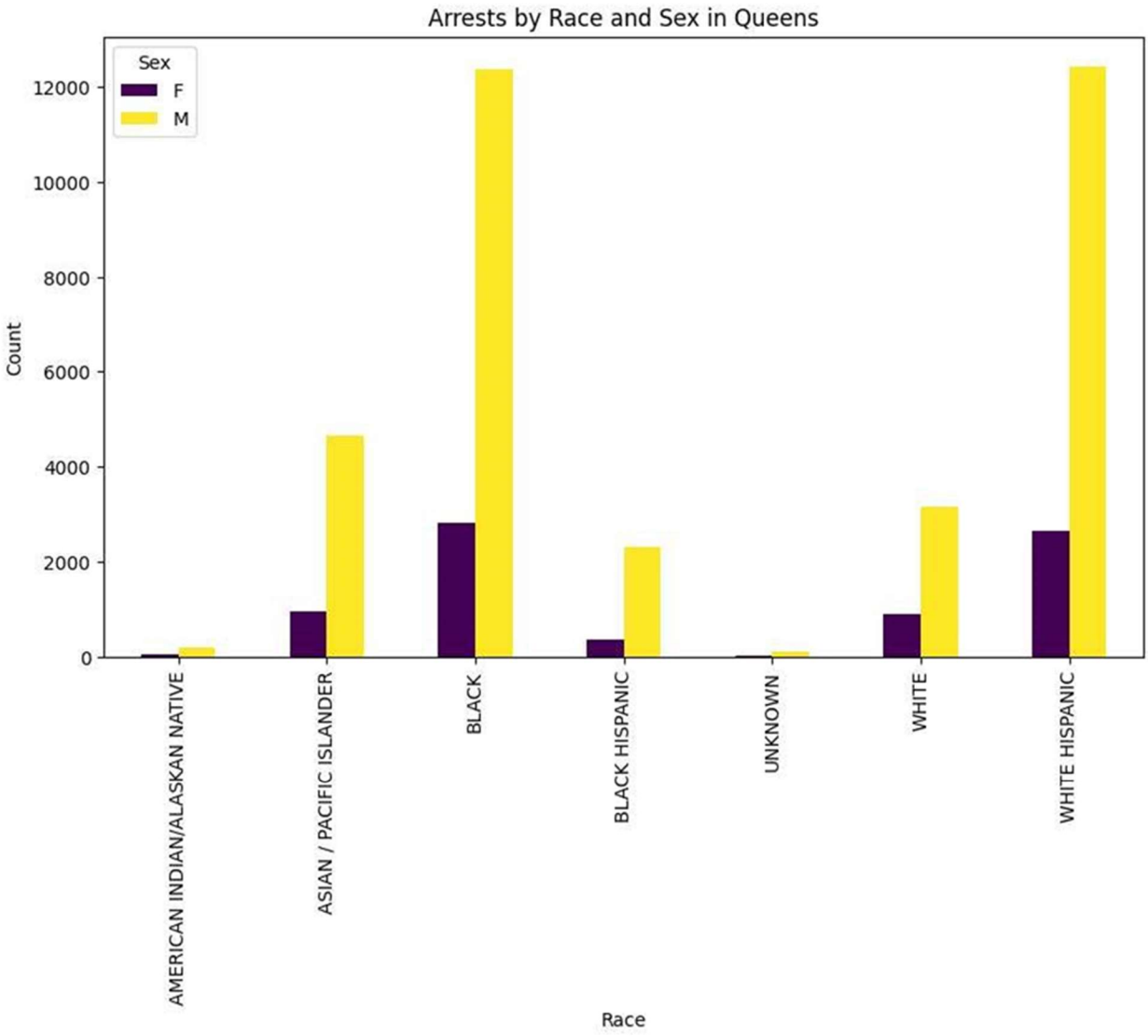


	Amount	Race
Max	15181	Black
Min	133	Unknown
Average	6139.714285714285	
Q1	1455.5	
Median	4041.0	
Q3	10356.0	
Upper Bound	23706.75	

The above chart shows the proportions of races arrested in Queens borough. The graphic affirms the posed question (Are there more Black people being arrested than Asian people in Queens?). It is seen on the chart that in Queens in 2024 from January of September 35.3% of arrests made were of Black perpetrators, whereas only 13.1% peoples of Asian/Pacific Island descent. One note that should be made is that the unknown category is not missing information, it is that the perpetrators race could not be determined.

The max amount of people arrested in Queens by race were Black at 15181 people. The minimum amount was 133 people of unknown race. The average arrests by race were 6139.71. There are no outliers in this set. The upper bound is 23706.75. The lower bound is negative, that cannot occur the as the lowest theoretical arrest that can be made are 0. There are no outliers in this sample.

One expectation I had is that there would be a slightly higher proportion of peoples of Asian descent represented on this chart. Queens is home to Flushing which has a very large community of people from Asian. My expectation was that because there was a large community of Asian people living in Queens, there would have been a higher representation on this graphic.



Perp Race	Male count	Female count
AMERICAN INDIAN/ALASKAN NATIVE	183	41
ASIAN / PACIFIC ISLANDER	4653	965
BLACK	12361	2820
BLACK HISPANIC	2326	361

UNKNOWN	117	16
WHITE	3155	886
WHITE HISPANIC	12432	2662

The above graphic further breaks down the race category to further understand the question posed (Are there more Black people being arrested than Asian people in Queens?). It shows a grouped bar graph of each race broken down by sex. This helps breakdown and further understand the question. One thing that would be interesting to investigate here is the total amount of males and females broken down by these races. Something that was unexpected here was how close the amount of white Hispanic women was to the number of black women. From the previous graphic, pie chart, it was seen that the proportion was very similar for two races, it is interesting to note that it seemed to hold true through the sexes as well.

The lowest amount of people arrested were females of unknown races at 16 people. The highest was that of black men. The lowest amount for men was that of people of unknown race at 117 and for females the unknown race as well. I am curious what entails someone being classified as unknown race as there are many races such as middle eastern, or Jewish descent not being included as options.

One expectation I had for this graphic is that there would be a higher arrest amount for men vs. women. This expectation was proven to be true as for each race there are more men that were arrested than there was female. Something that would require further study is the types of crimes being committed broken down by sex as well.

After visualizing both the number of arrests in Queens by race and by gender, the question was expanded to include both factors: "Is there an association between gender and race (Black vs. Asian) in the arrests of Queens?" the revised question provides a deeper insight into the data. To explore this further, a statistical test was conducted.

Test run:

- Chi-squared test of independence
    - Tests whether two categorical variables are independent of each other

Assumptions:

1. Independence of Observations
    a. Each arrest falls into one category; a person is either male or female, Black or Asian

b. The information of one arrest does not influence another
    i. Though arrests of multiple perpetrators in a singular case might be present, this information is not present in the data. Therefore, independence **may be** violated. However, since only arrest count information is available the results of the test should be interpreted cautiously, i.e. not make any policy decisions based on the data.

2. Counts should be sufficiently large
    a. The expected frequency for each category should be at least 5 to ensure that the Chi-squared approximation is valid.
    b. Expected arrest were much larger than 5:

	Black	Asian/Pacific Islander
Male	12418.36	4595.64
Female	2762.64	1022.36

3. Random Sampling
    a. The data is taken from all over Queens borough from the 1st of January to the 30th of September 2024. Not one time, or precinct meaning the sample is random for all of Queens
4. Large sample size:
    a. The samples are all larger than 5, meaning they are sufficiently large.

Results:

Chi-Square Statistic: 5.2970480959256605

P-value: 0.02136159656502185

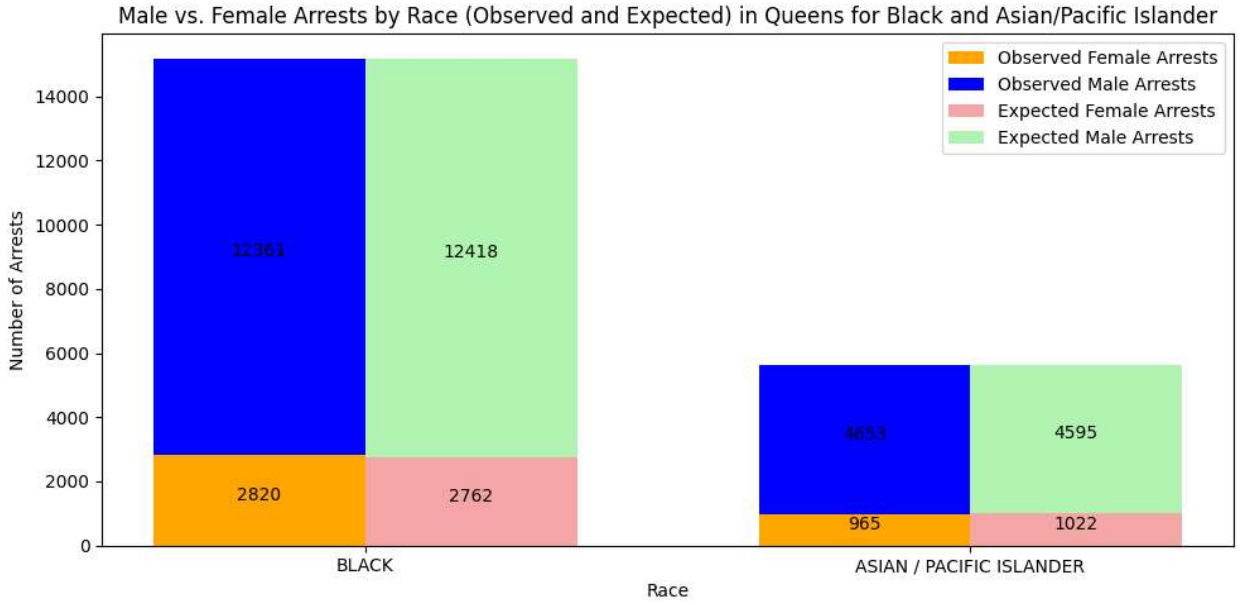
Degrees of Freedom: 1

Alpha: .05

Conclusion:

P is less than alpha so:

Reject the null hypothesis! (H_1): Sex and race (Black/Asian) are dependent. The proportion of males and females differs across the racial groups. The below plot shows that observed arrests for black females is larger than the expected amounts (2820 compared to 2762 expected). The black males observed number of arrests are less than the expected amounts (12361 compared 12418 expected). The Asian/Pacific Islander females have a lower number of observed arrests than the expected amount (965 compared to 1022 expected). Asian/Pacific Islander males have a higher number of observed arrests than the expected amount (4653 compared to 4595)


Male vs. Female Arrests by Race (Observed and Expected) in Queens for Black and Asian/Pacific Islander
Observed Female Arrests
Observed Male Arrests
Expected Female Arrests
Expected Male Arrests
Number of Arrests
0
2000
4000
6000
8000
10000
12000
14000
12361
12418
2820
2762
4653
4595
965
1022
BLACK
ASIAN / PACIFIC ISLANDER
Race

Meta-data and Variables for dataset

- The dataset has 195,447 entries in the file
- Each entry has 19 attributes:
    1. **ARREST_KEY**: Randomly generated persistent ID for each arrest
        - This is text
    2. **ARREST_DATE**: Exact date of arrest for the reported event
        - This is a floating Timestamp
    3. **PD_CD**: Three-digit internal classification code (more granular than Key Code)
        - This is a number
    4. **PD_DESC**: Description of internal classification corresponding with PD code (more granular than Offense Description)
        - This is a text field
    5. **KY_CD**: Three-digit internal classification code (more general category than PD code)
        - This is a number
    6. **OFNS_DESC**: Description of internal classification corresponding with KY code (more general category than PD description)
        - This is a text field
    7. **LAW_CODE**: Law code charges corresponding to the NYS Penal Law, VTL, and other various local laws
        - This is a text field
    8. **LAW_CAT_CD**: Level of offense: felony, misdemeanor, or violation
        - This is a text field
    9. **ARREST_BORO**: Borough of arrest
        - B: Bronx, S: Staten Island, K: Brooklyn, M: Manhattan, Q: Queens
        - This is a text field
    10. **ARREST_PRECINCT**: Precinct where the arrest occurred
        - This is a number
    11. **JURISDICTION_CODE**: Jurisdiction responsible for the arrest.
        - 0: Patrol, 1: Transit, 2: Housing (NYPD), Codes 3 and above represent non-NYPD jurisdictions
        - This is a number
    12. **AGE_GROUP**: Perpetrator's age within a category
        - This is a text field
    13. **PERP_SEX**: Perpetrator's sex description
        - This is a text field
    14. **PERP_RACE**: Perpetrator's race description
        - This is a text field
    15. **X_COORD_CD**: Midblock X-coordinate for New York State Plane Coordinate System, Long Island Zone, NAD 83, units in feet (FIPS 3104)
        - This is a number.

16. **Y_COORD_CD**: Midblock Y-coordinate for New York State Plane Coordinate System, Long Island Zone, NAD 83, units feet (FIPS 3104)
    - This is a number
17. **Latitude**: Latitude coordinate for Global Coordinate System, WGS 1984, decimal degrees (EPSG 4326)
    - This is a number
18. **Longitude**: Longitude coordinate for Global Coordinate System, WGS 1984, decimal degrees (EPSG 4326)
    - This is a number
19. **New Georeferenced Column**: combined Latitude and Longitude
    - This is a point

Appendix:

- Git hub link:
    - https://github.com/emirbeg2017/505project
- Link to dataset:
    - https://tinyurl.com/mw4ftb8u