Week 3 Assignment

Nicole Cataland

2025-01-13

Step 1 - Import Data

NYPD Shooting Incident Data

Our first step will be importing the NYPD Shooting Incident Data. This dataset has been compiled by the Office of Management Analysis and Planning each quarter since 2006. Each record in the dataset represents a shooting incident and includes details, such as location, time of day, and suspect/victim demographics. The dataset is made publicly available on the data.gov website, where we will directly source it in CSV format using the tidyverse library. We will be using the ggplot2 library later to visualize our data, so let's go ahead and import that now too to be tidy.

```
library(ggplot2)
library(tidyverse)
data_url <- "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD"
incident_data_raw <- read_csv(data_url)
incident_data_raw
```

```
## # A tibble: 28,562 x 21
##    INCIDENT_KEY OCCUR_DATE OCCUR_TIME BORO     LOC_OF_OCCUR_DESC PRECINCT
##           <dbl> <chr>      <time>     <chr>    <chr>                <dbl>
##  1    231974218 08/09/2021 01:06      BRONX    <NA>                    40
##  2    177934247 04/07/2018 19:48      BROOKLYN <NA>                    79
##  3    255028563 12/02/2022 22:57      BRONX    OUTSIDE                 47
##  4     25384540 11/19/2006 01:50      BROOKLYN <NA>                    66
##  5     72616285 05/09/2010 01:58      BRONX    <NA>                    46
##  6     85875439 07/22/2012 21:35      BRONX    <NA>                    42
##  7     79780323 07/12/2011 22:26      BROOKLYN <NA>                    71
##  8     85744504 07/14/2012 23:45      BROOKLYN <NA>                    69
##  9    142324890 04/21/2015 15:36      BROOKLYN <NA>                    75
## 10    152868707 05/07/2016 15:23      BROOKLYN <NA>                    69
## # ... with 28,552 more rows, and 15 more variables: JURISDICTION_CODE <dbl>,
## #   LOC_CLASSFCTN_DESC <chr>, LOCATION_DESC <chr>,
## #   STATISTICAL_MURDER_FLAG <lgl>, PERP_AGE_GROUP <chr>, PERP_SEX <chr>,
## #   PERP_RACE <chr>, VIC_AGE_GROUP <chr>, VIC_SEX <chr>, VIC_RACE <chr>,
## #   X_COORD_CD <dbl>, Y_COORD_CD <dbl>, Latitude <dbl>, Longitude <dbl>,
## #   Lon_Lat <chr>
```

Step 2 - Clean Data

Now that we have our data imported, we need to tidy it up.

After reviewing the data footnotes provided by the NYPD, we find that our dataset may have duplicate INCIDENT_KEYs in the event a shooting incident had multiple victims. Investigating the proportion of shootings that have multiple victims will serve as the starting point for our analysis.

First, we will begin by making sure all columns are in the appropriate format and selecting only the columns that are relevant to our analysis.

```
incident_data <- select(incident_data_raw, INCIDENT_KEY, OCCUR_DATE, OCCUR_TIME, BORO, STATISTICAL_MURDE
incident_data$OCCUR_DATE <- as.Date(incident_data$OCCUR_DATE, format = "%m/%d/%Y")
incident_data$PERP_SEX[incident_data$PERP_SEX == "(null)"] <- NA
incident_data$PERP_SEX <- as.factor(incident_data$PERP_SEX)
incident_data$VIC_SEX <- as.factor(incident_data$VIC_SEX)
incident_data$BORO <- as.factor(incident_data$BORO)
summary(incident_data)
```

```
##   INCIDENT_KEY         OCCUR_DATE          OCCUR_TIME
##  Min.   :  9953245   Min.   :2006-01-01   Length:28562
##  1st Qu.: 65439914   1st Qu.:2009-09-04   Class1:hms
##  Median : 92711254   Median :2013-09-20   Class2:difftime
##  Mean   :127405824   Mean   :2014-06-07   Mode  :numeric
##  3rd Qu.:203131993   3rd Qu.:2019-09-29
##  Max.   :279758069   Max.   :2023-12-29
##             BORO       STATISTICAL_MURDER_FLAG PERP_SEX     VIC_SEX
##  BRONX        : 8376   Mode :logical           F   :  444   F: 2760
##  BROOKLYN     :11346   FALSE:23036             M   :16168   M:25790
##  MANHATTAN    : 3762   TRUE :5526              U   : 1499   U:   12
##  QUEENS       : 4271                           NA's:10451
##  STATEN ISLAND:  807
##
```

Ah, much better. Now to the fun part!

Step 3 - Analysis

There are a few ways we can approach determining how many shooting incidents had multiple victims. We will start with the simplest to give us an idea: count the total number of unique INCIDENT_KEYs and compare that to the total number of records in our dataset. To see this as a percentage requires only some basic arithemetic:

Percentage Multiple Victims Incidents = ((Total INCIDENT_KEYs - Unique INCIDENT_KEYs)/Unique INCIDENT_KEYs) x 100

```
incident_records_count <- nrow(incident_data)
unique_incidents_count <- length(unique(incident_data$INCIDENT_KEY))
multi_victim_incidents <- incident_records_count - unique_incidents_count
percentage_multi_victim <- (multi_victim_incidents / unique_incidents_count) * 100

print(paste0("Total incident records: ", incident_records_count))
```

```
## [1] "Total incident records: 28562"
```

```
print(paste0("Total unique incidents: ", unique_incidents_count))
```

```
## [1] "Total unique incidents: 22394"
```

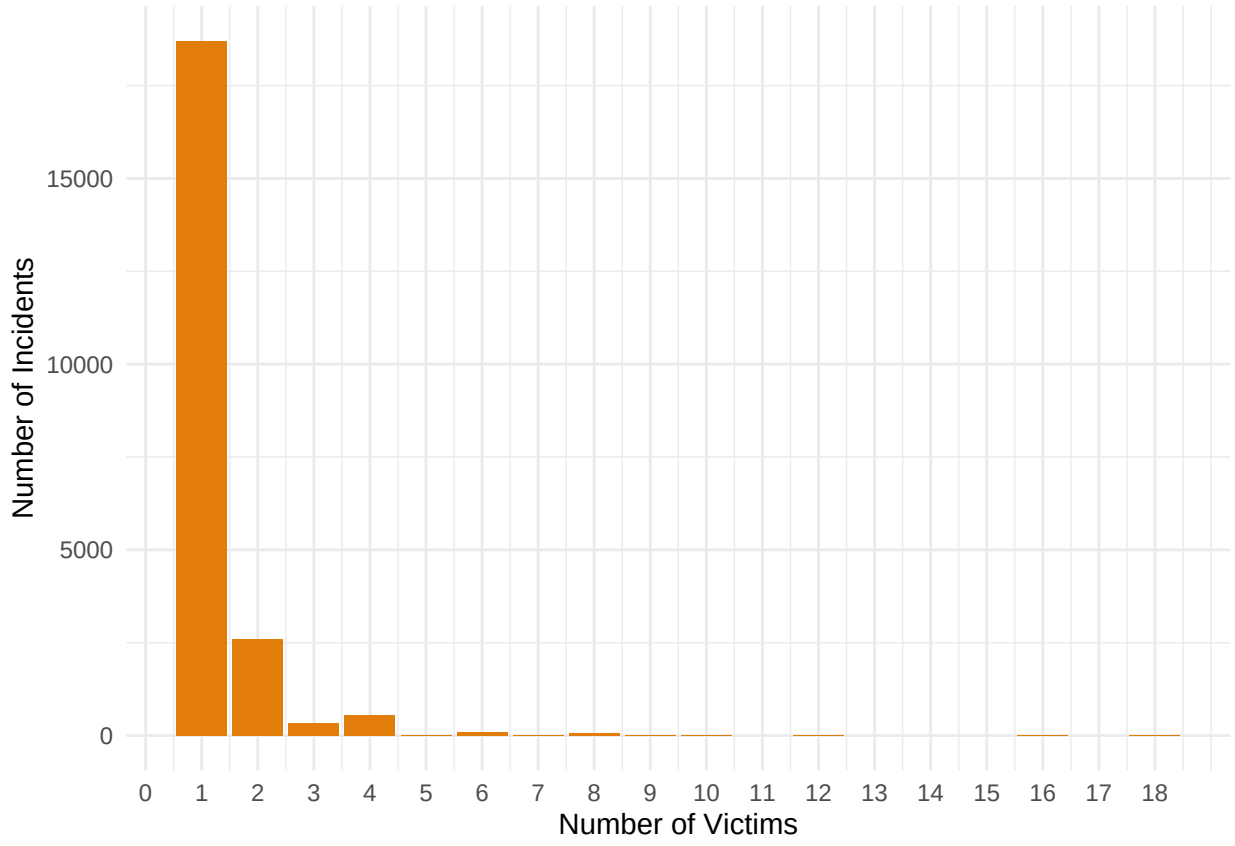
```
print(paste0("Percent multiple victim incidents: ", percentage_multi_victim, "%"))
```

```
## [1] "Percent multiple victim incidents: 27.543091899616%"
```

OK, so not an insignificant number of shooting incidents must have had multiple victims. Let's investigate this further. We will begin by grouping and counting records with the same INCIDENT_KEY to get a better look at the distribution of victim counts for shooting incidents in NYC. Also, we will need to create a new dataframe where each unique incident is represented by a single record to visualize our data.

```
incident_data <- incident_data %>% group_by(INCIDENT_KEY) %>% mutate(VICTIM_COUNT = n()) %>% ungroup()
unique_incidents <- incident_data %>% distinct(INCIDENT_KEY, .keep_all = TRUE)

ggplot(unique_incidents, aes(x = VICTIM_COUNT)) +
  geom_bar(fill = "#E27C0B") +
  scale_x_continuous(breaks=seq(0,max(unique_incidents$VICTIM_COUNT),by=1)) +
  #scale_y_log10() +
  labs(x = "Number of Victims", y = "Number of Incidents" ) +
  theme_minimal()
```



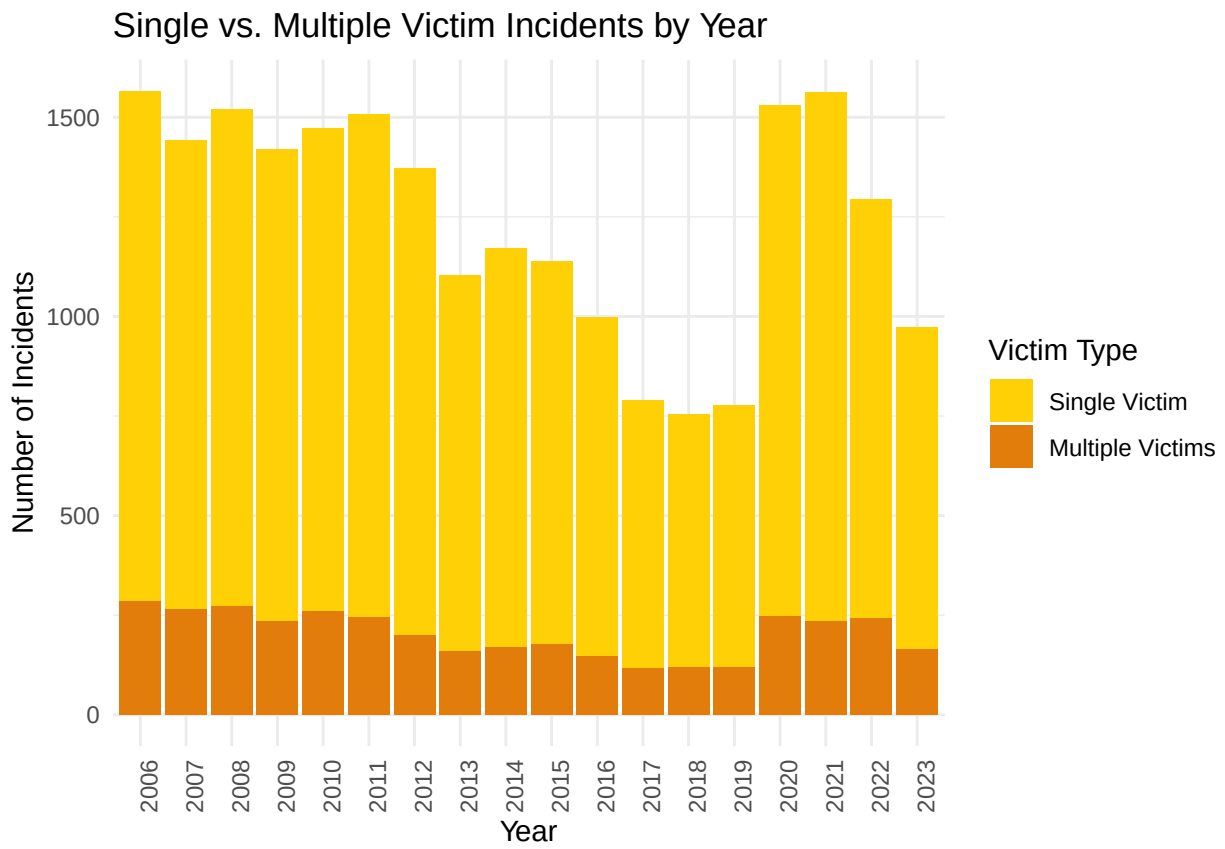
Well... Now we see it. But it doesn't tell us much.

Except for that it will probably be more valuable to us to think about the data in more binary terms of "single" vs "multiple" victim incidents moving forward. Let's go ahead and add this as a new column factor in our dataset, then break down the data by year to see if any patterns emerge. (Moving forward, we will refer to this new factor column as "incident type" to describe if a shooting incident had a single victim or multiple victims.)

```
unique_incidents <- unique_incidents %>% mutate(INCIDENT_TYPE = if_else(VICTIM_COUNT == 1, "Single Victi
unique_incidents$INCIDENT_TYPE <- as.factor(unique_incidents$INCIDENT_TYPE)

incidents_by_year <- unique_incidents %>%
  mutate(YEAR = as.numeric(format(OCCUR_DATE, "%Y"))) %>%
  group_by(YEAR, INCIDENT_TYPE) %>%
  summarize(INCIDENT_COUNT = n(), .groups = "drop")


ggplot(incidents_by_year, aes(x = factor(YEAR), y = INCIDENT_COUNT,
  fill = factor(INCIDENT_TYPE, levels = c("Single Victim","Multiple Victims")))) +
  geom_col(position = "stack") +
  labs(
    title = "Single vs. Multiple Victim Incidents by Year",
    x = "Year",
    y = "Number of Incidents",
    fill = "Victim Type"
  ) +
  theme_minimal() +
  scale_fill_manual(values = c("#FED005", "#E27C0B")) +
  theme(
    axis.text.x = element_text(angle = 90, hjust = 1)
  )
```



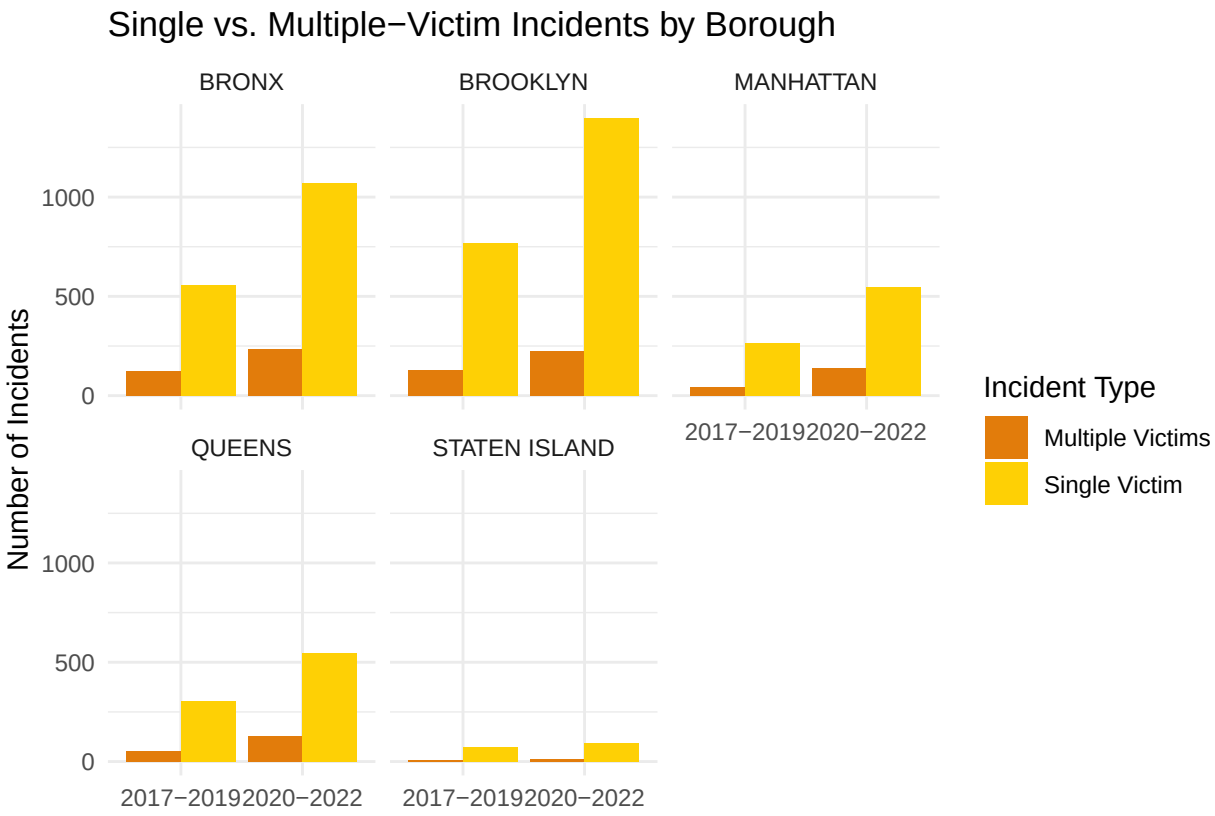
Now that's more interesting. Shooting incidents as a whole appeared to be trending downward before spiking back up in 2020. Let's add another column factor to our dataset to zero in our attention on the 3-year periods preceding (2017-2019) and following (2020-2022) this spike. Let's also further break the data down by borough to add a new dimension to help us understand the abrupt increase in incidents and if it was accompanied by changes in the proportion of multi-victim incidents.

```
incidents_by_year_range <- unique_incidents %>%
  mutate(YEAR = as.numeric(format(OCCUR_DATE, "%Y"))) %>%
  mutate(YEAR_RANGE = case_when(
    YEAR >= 2017 & YEAR <= 2019 ~ "2017-2019",
    YEAR >= 2020 & YEAR <= 2022 ~ "2020-2022",
    TRUE ~ NA_character_ ) ) %>%
  filter(!is.na(YEAR_RANGE))

incidents_by_year_range <- incidents_by_year_range %>%
  group_by(BORO, YEAR_RANGE, INCIDENT_TYPE) %>%
  summarize(INCIDENT_COUNT = n(), .groups = "drop")

ggplot(incidents_by_year_range, aes(x = YEAR_RANGE, y = INCIDENT_COUNT, fill = INCIDENT_TYPE)) +
  geom_col(position = "dodge") +
  facet_wrap(~ BORO) +
  scale_fill_manual(values = c("#E27C0B", "#FED005")) +
  labs(
    title = "Single vs. Multiple-Victim Incidents by Borough",
    x = "",
    y = "Number of Incidents",
```

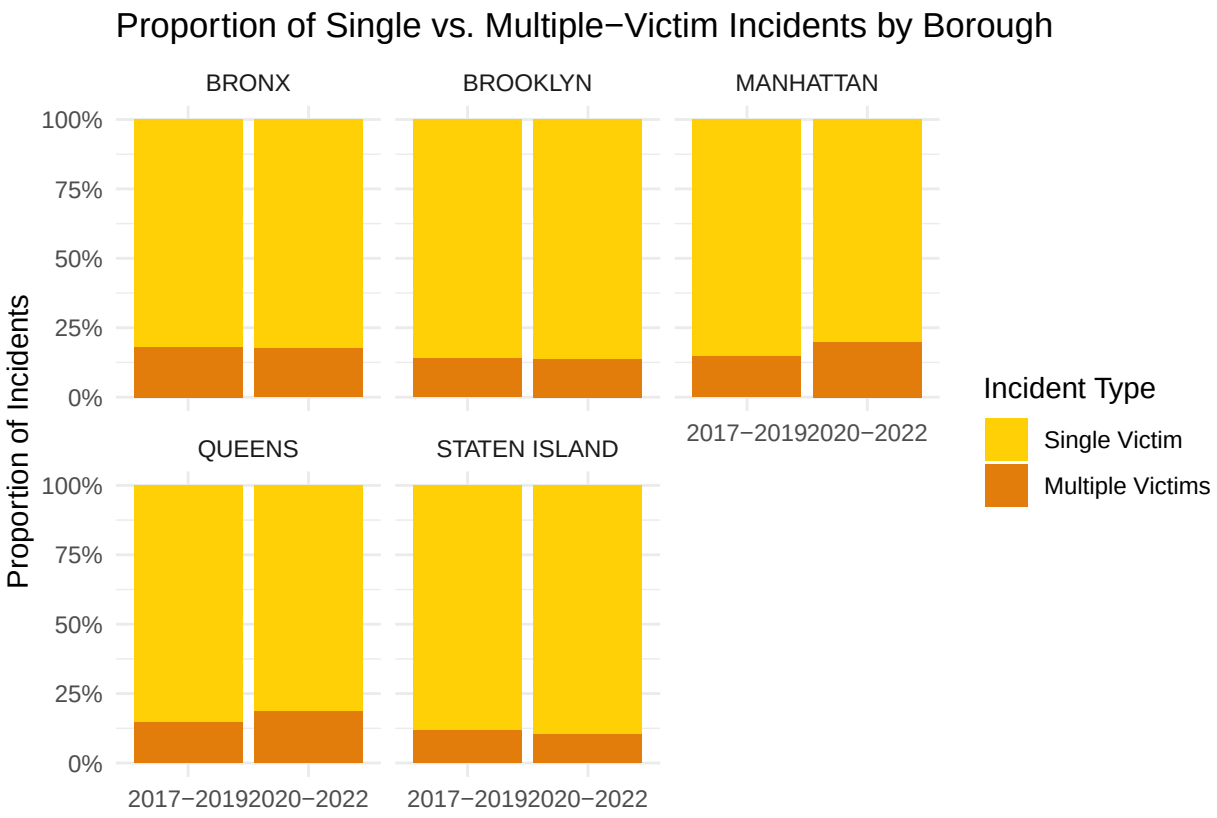
```
    fill = "Incident Type"
  ) +
  theme_minimal()
```



We can see above that the majority of shooting incidents in NYC during our selected year ranges took place in Brooklyn and the Bronx. However, this picture may be not be a fully accurate depiction due to bias in our dataset which could result from some boroughs having more consistent and reliable reporting standards than others. This graph also doesn't much help us to visually appreciate how the proportion of multi-victim incidents may or may not have changed.

To get a clearer view, let's visualize the same data again using spine plots so that all of our borough data is set to the same scale.

```
ggplot(incidents_by_year_range, aes(x = YEAR_RANGE, y = INCIDENT_COUNT, fill = factor(INCIDENT_TYPE, le
  geom_col(position = "fill") +
  facet_wrap(~ BORO) +
  scale_y_continuous(labels = function(x) paste0(round(100 * x), "%")) +
  labs(
    title = "Proportion of Single vs. Multiple-Victim Incidents by Borough",
    x = "",
    y = "Proportion of Incidents",
    fill = "Incident Type"
  ) +
  theme_minimal() +
  scale_fill_manual(values = c("#FED005", "#E27C0B"))
```



How about that. Viewing the same dataset in the spine plot format (where all data bars are scaled to 1) reveals some curious observations that otherwise would have been lost on us. The proportion of multiple victim shooting incidents actually increased in Manhattan and Queens, whereas the proportion in Bronx and Brooklyn largely stayed the same or even slightly decreased.

Step 4 - Modeling

Before we get too far ahead of ourselves, we need to determine if there is any statistical significance to our suspicion there's a relationship between NYC borough and the incident type during our periods of observation.

Let's first return to our earlier complete dataset of unique incidents to determine if there's a relationship between these two factors across all years of available data. Since our variables of interest are categorical, we will use a chi-squared test to evaluate their independence. To perform this test, we must do a little restructuring of our data to get it into a frequency table format but after that, R does all of the heavy lifting. After calling R's built in chisq.test() function, it only remains for us to interpret the results.

```
frequency_all_incidents <- table(unique_incidents$BORO, unique_incidents$INCIDENT_TYPE)
frequency_all_incidents
```

```
##
##                Multiple Victims Single Victim
##   BRONX                    1154          5181
##   BROOKLYN                 1351          7793
##   MANHATTAN                 535          2374
```

```
##   QUEENS                       555          2805
##   STATEN ISLAND                 92           554
```

```
chisq_all_incidents <- chisq.test(frequency_all_incidents)
chisq_all_incidents
```

```
##
##  Pearson's Chi-squared test
##
## data:  frequency_all_incidents
## X-squared = 43.298, df = 4, p-value = 8.974e-09
```

The value we are most concerned with here is the p-value, which tells us if our factors are related or if patterns emerging between the two resulted from something akin to coincidence. Our p-value comes in well below that cut-off point (.05) at 0.000000008974, strongly suggesting our hunch was right and there is a relationship between borough and incident type.

Now that's established, let's repeat the same test for the datasets we were working with earlier to better understand the periods preceding and following the incident spike in 2020.

```
#2017-2019
incidents_2017_2019 <- subset(unique_incidents, OCCUR_DATE >= as.Date("2017-01-01") & OCCUR_DATE <= as.
frequency_2017_2019_incidents <- table(incidents_2017_2019$BORO, incidents_2017_2019$INCIDENT_TYPE)
frequency_2017_2019_incidents
```

```
##
##                Multiple Victims Single Victim
##   BRONX                     122           554
##   BROOKLYN                  127           768
##   MANHATTAN                  45           262
##   QUEENS                     53           304
##   STATEN ISLAND              10            75
```

```
chisq_2017_2019_incidents <- chisq.test(frequency_2017_2019_incidents)
print('Chi-Square Test Results for 2017-2019:')
```

```
## [1] "Chi-Square Test Results for 2017-2019:"
```

```
chisq_2017_2019_incidents
```

```
##
##  Pearson's Chi-squared test
##
## data:  frequency_2017_2019_incidents
## X-squared = 5.7218, df = 4, p-value = 0.2209
```

```
#2020-2023
incidents_2020_2023 <- subset(unique_incidents, OCCUR_DATE >= as.Date("2020-01-01") & OCCUR_DATE <= as.
frequency_2020_2023_incidents <- table(incidents_2020_2023$BORO, incidents_2020_2023$INCIDENT_TYPE)
frequency_2020_2023_incidents
```

```
##
##                 Multiple Victims Single Victim
##   BRONX                      295          1326
##   BROOKLYN                   274          1684
##   MANHATTAN                  161           685
##   QUEENS                     151           652
##   STATEN ISLAND               13           120
```

```
chisq_2020_2023_incidents <- chisq.test(frequency_2020_2023_incidents)
print('Chi-Square Test Results for 2020-2023:')
```

```
## [1] "Chi-Square Test Results for 2020-2023:"
```

```
chisq_2020_2023_incidents
```

```
##
##  Pearson's Chi-squared test
##
## data:  frequency_2020_2023_incidents
## X-squared = 23.395, df = 4, p-value = 0.0001056
```

The plot thickens! There appears to be no statistically significant relationship between borough and incident type in the years preceeding the spike, given our p value of .2209 is greater than .05. Yet, we see strong evidence for a relationship between these two factors in the years 2020-2023 with a p-value of 0.0001056.

Step 5 - Conclusion

Our cursory analysis into multiple victim shooting incidents in NYC yielded more questions than answers, but provded valuable direction for future analysis. We sought to better understand the spike in shooting incidents in the year 2020 by focusing on the preceeding and following 3-year periods. This lead us to observe that changes in the prevalence of multiple victim incidents were not uniform across all New York City boroughs, which inspired later analysis where we found the relationship between borough and incident type has changed–and gotten stronger–over time. These observations invite further investigation into how different features or strategies in each NYC borough may enable or mitigate multiple victim incidents. Our analysis also benefits such future efforts by allowing them to dial in on policy or policing changes that occurred in 2020 which could help explain our findings.

Though steps were taken to mitigate bias by considering scaled data and statistically supporting our hypothesis using chi-square tests, our analysis was still subject to personal bias due to the arbitrary selection of year ranges. Furthermore, the decision to classify the records into single victim or multiple victim incidents may have introduced undue influence on the interpretation of results. Other forms of bias (such as underreporting in certain boroughs or data entry errors) may have also impacted the quality of our data, thus skewing our results. All considered, a much more rigorous analysis would need to be performed to substantiate any of our claims.