



# ANALYZING NYPD SHOOTING INCIDENTS (2006–2024)

FINAL PROJECT  
DTSA 5301: DATA SCIENCE AS A FIELD  
JINNAJATE ACHALAPONG



# PROJECT GOALS & QUESTION

## Project Goals:


- Demonstrate the end-to-end data science process
- Use real-world NYC data
- Predict risk factors for shootings

## Key Question:

- How can past data help us decide where and when to deploy police resources most effectively?


# DATA OVERVIEW







**City of New York**

There is no description for this organization

 Topics

 This is a Non-Federal dataset covered by different Terms of Use than Data.gov. [See Terms](#)

## NYPD Shooting Incident Data (Historic)

 Metadata Updated: April 19, 2025

List of every shooting incident that occurred in NYC going back to 2006 through the end of the previous calendar year.

This is a breakdown of every shooting incident that occurred in NYC going back to 2006 through the end of the previous calendar year. This data is manually extracted every quarter and reviewed by the Office of Management Analysis and Planning before being posted on the NYPD website. Each record represents a shooting incident in NYC and includes information about the event, the location and time of occurrence. In addition, information related to suspect and victim demographics is also included. This data can be used by the public to explore the nature of shooting/criminal activity. Please refer to the attached data footnotes for additional information about this dataset.

## Where Our Data Comes From:

- **Source:** NYPD Open Data Portal
- **Period:** 2006–2024
- ~30,000 records
- **Features:** Date, time, location, demographics (victim & perpetrator)

# DATA CLEANING & PREPARATION

## Step 2: Tidy and Transform Data

### Remove Unnecessary Columns

The following columns are not needed for this assignment:

PRECINCT, JURISDICTION\_CODE, LOCATION\_DESC, X\_COORD\_CD, Y\_COORD\_CD, Lon\_Lat

```
nypd_shooting <- nypd_shooting %>%  
  select(-c(PRECINCT, JURISDICTION_CODE, LOCATION_DESC, X_COORD_CD, Y_COORD_CD, Lon_Lat)) %>%  
  mutate(OCCUR_DATE = mdy(OCCUR_DATE),  
         OCCUR_TIME = hms(OCCUR_TIME),  
         Shootings = 1,  
         OCCUR_YEAR = year(OCCUR_DATE),  
         OCCUR_MONTH = month(OCCUR_DATE, label = TRUE, abbr = TRUE),  
         OCCUR_WDAY = weekdays(OCCUR_DATE),  
         OCCUR_HOUR = hour(hms(OCCUR_TIME)))
```

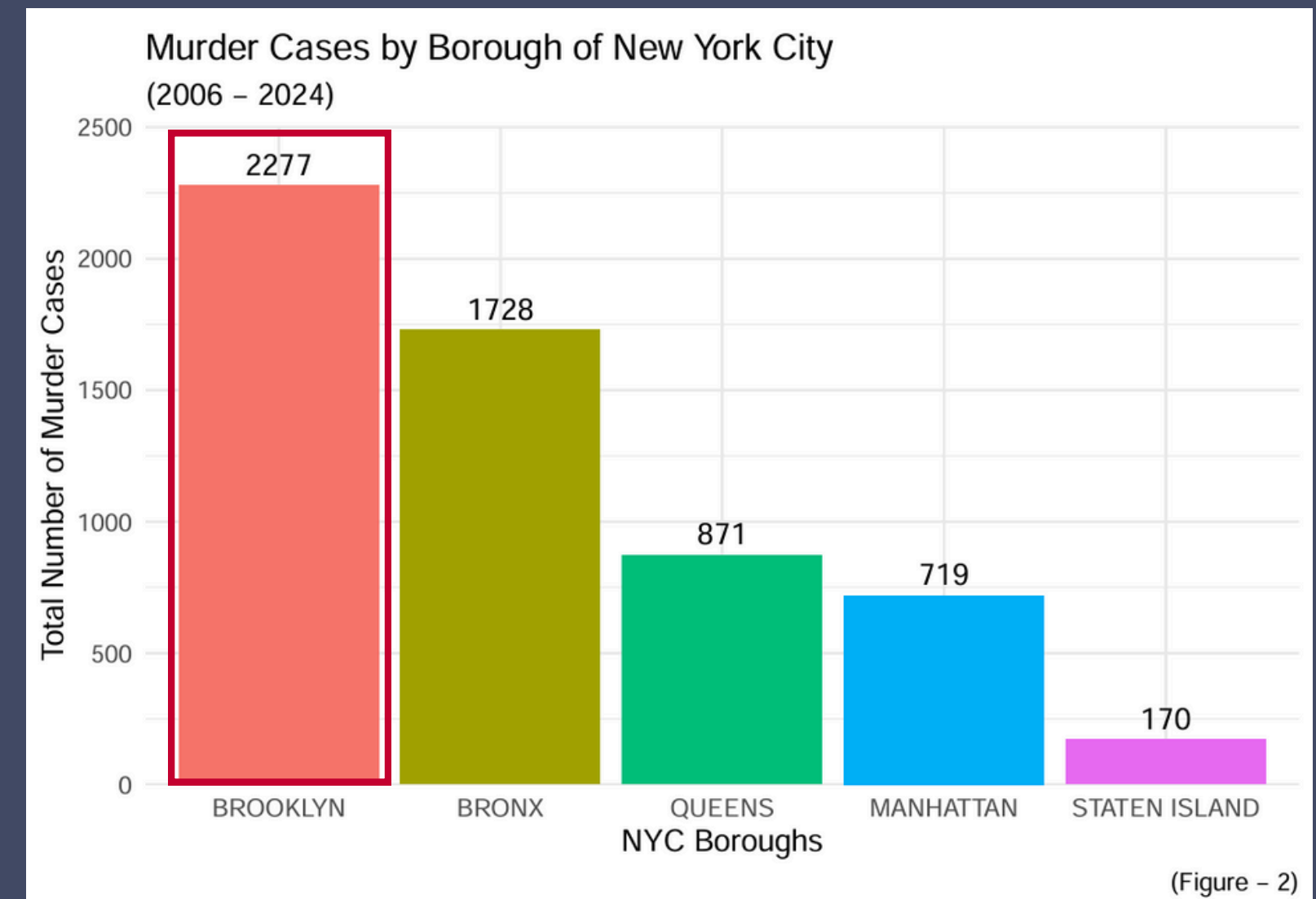
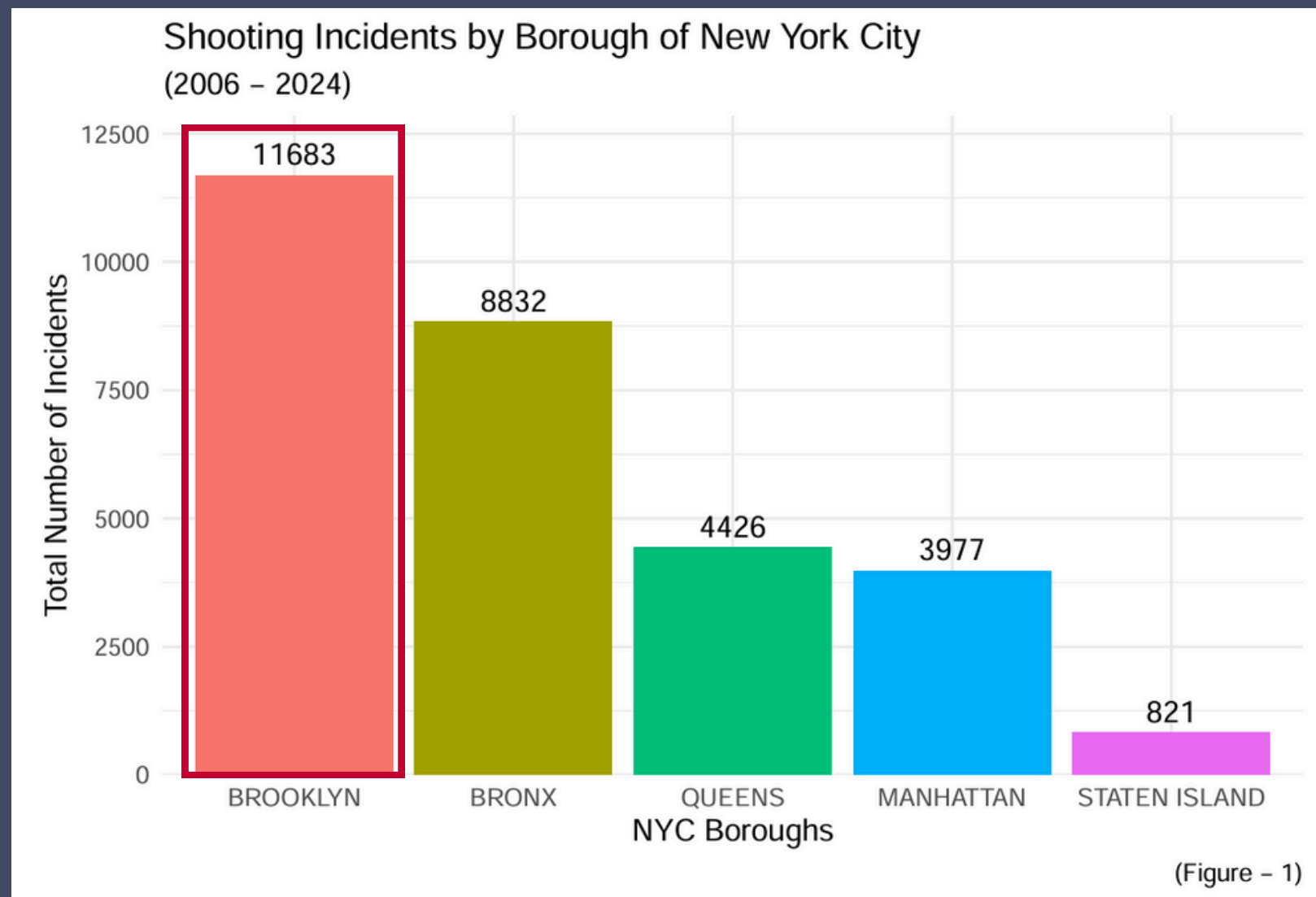
## Making Raw Data Useful:

- Remove unnecessary columns
  - PRECINCT, JURISDICTION\_CODE, LOCATION\_DESC, X\_COORD\_CD, Y\_COORD\_CD, Lon\_Lat
- Replace missing values and remove extreme values in the data
- Converted time/date formats
- Descriptive statistics

# GEOGRAPHIC HOTSPOTS OF VIOLENCE

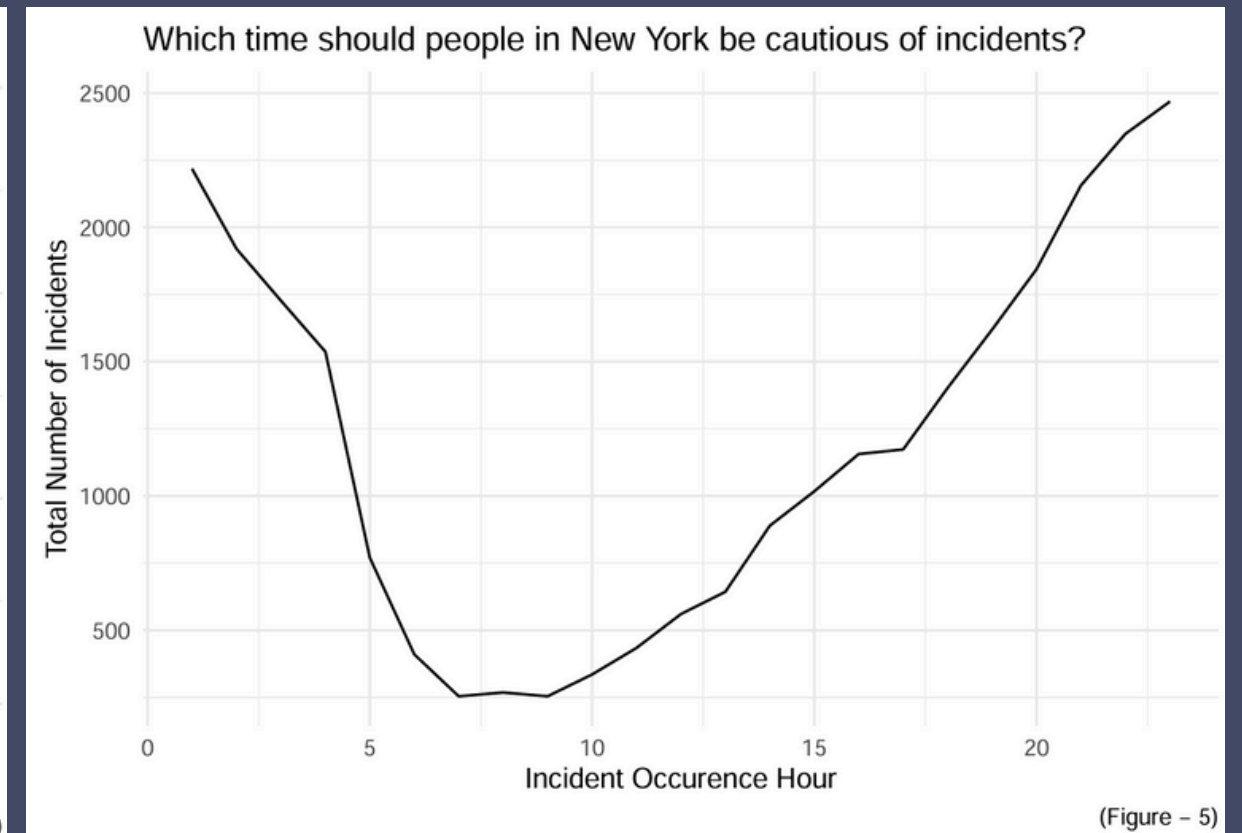
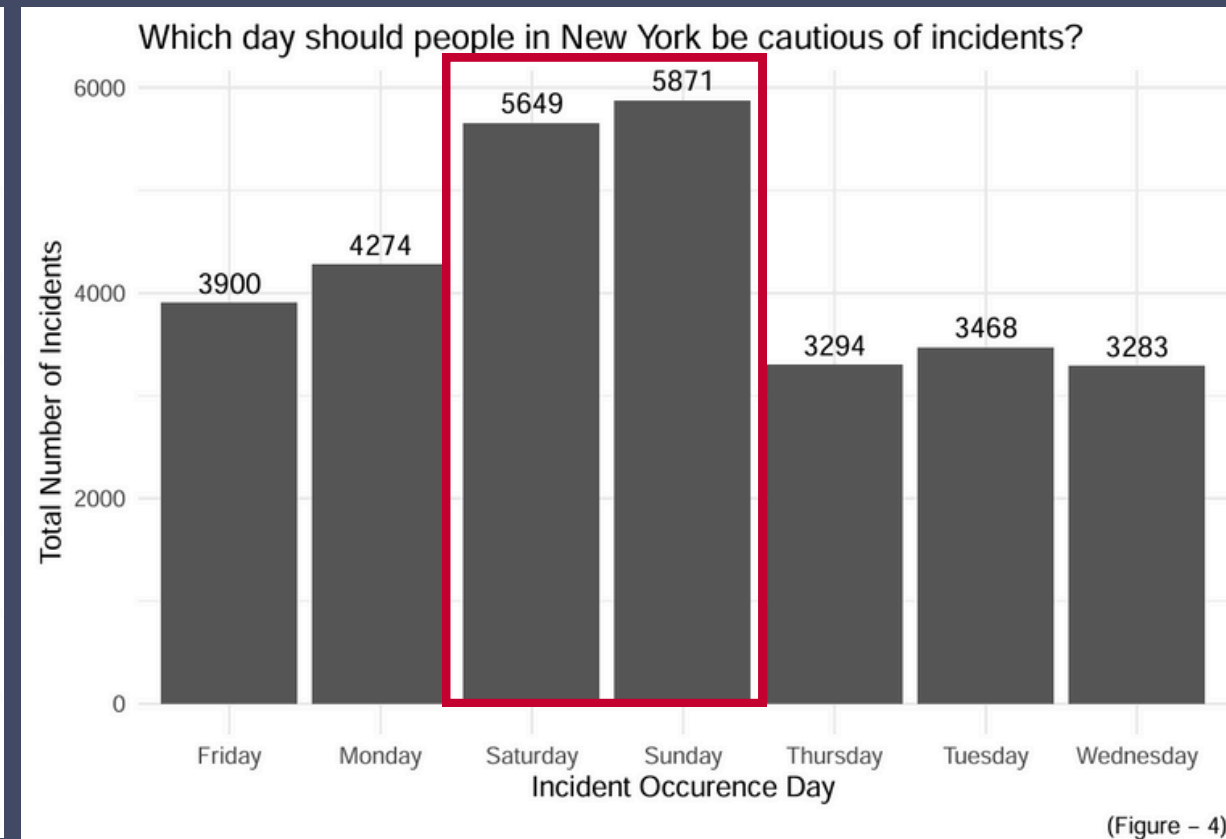
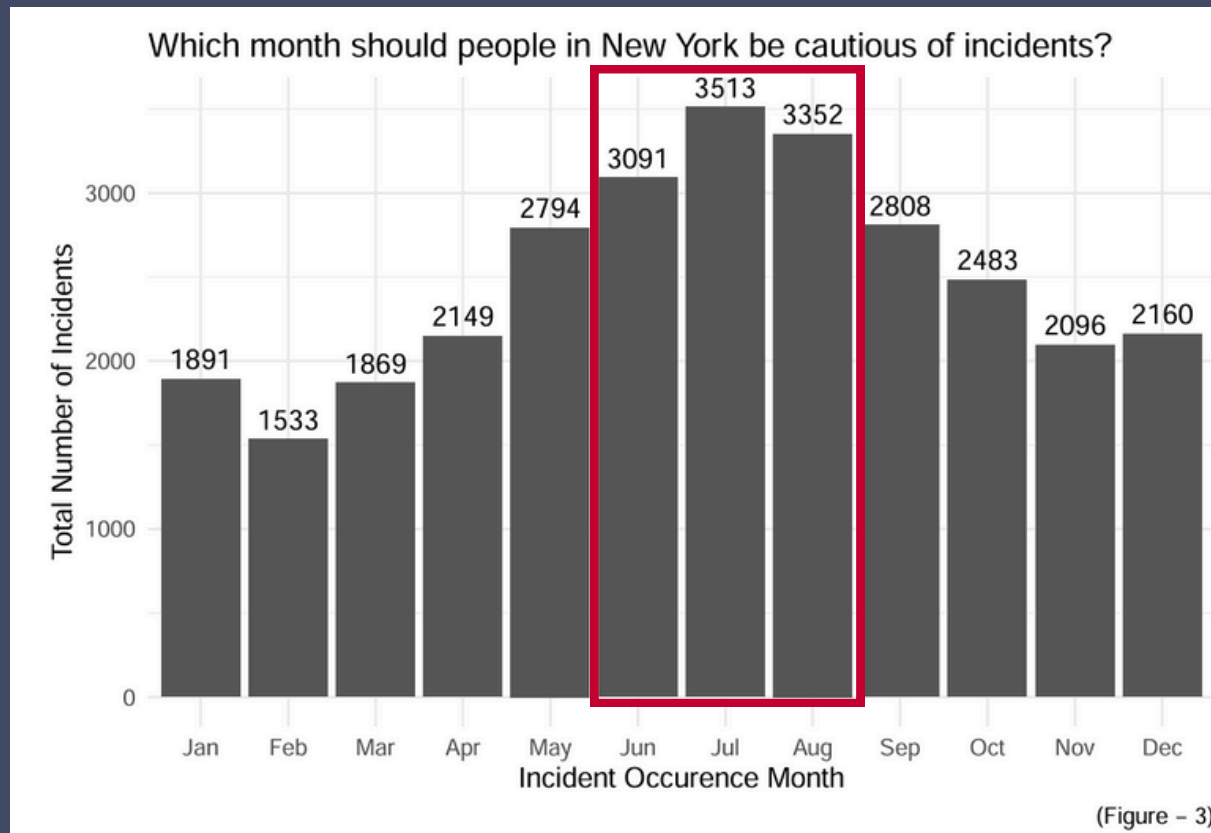
## Where Shootings Happen Most

- **Brooklyn** recorded the highest number of shooting incidents
- The pattern is similar when looking specifically at murder cases



# WHEN CRIME HAPPENS

- **Summer** months—particularly June, July, and August
- **Weekends** tend to have more criminal activity
- **Evenings** and **nighttime** are the riskiest hours.



# THE PROFILE OF PERPETRATORS AND VICTIMS

## Who's Involved?

- A significant number of incidents involve individuals aged **18–24** and **25–44**
- **Black** and **White Hispanic** individuals appear most frequently in incident records.
- The vast majority of incidents involve **male** individuals

```
table(nypd_shooting$PERP_AGE_GROUP, nypd_shooting$VIC_AGE_GROUP)
```

##							
##		<18	1022	18-24	25-44	45-64	65+ UNKNOWN
##		812	0	3568	4342	573	44 5
##	(null)	156	0	457	859	135	21 0
##	<18	566	0	669	455	90	23 2
##	18-24	825	1	2903	2483	355	49 14
##	25-44	284	0	1622	3773	571	52 40
##	45-64	22	0	90	419	221	18 5
##	65+	0	0	2	27	25	13 0
##	Unknown	416	0	1364	1202	148	16 2

```
table(nypd_shooting$PERP_SEX, nypd_shooting$VIC_SEX)
```

##			
##		F	M Unknown
##		693	8614 3
##	(null)	176	1452 0
##	F	80	380 1
##	M	1830	15003 7
##	Unknown	112	1387 1



# PREDICTING MURDER CASES

- Use logistic regression to estimate the probability that a shooting incident
- The **victim's age group** and the **perpetrator's race** were statistically significant predictors of whether the victim survived

```
# Logistics Regression
glm_model <- glm(STATISTICAL_MURDER_FLAG ~ PERP_RACE + PERP_SEX + PERP_AGE_GROUP + VIC_RACE + VIC_SEX +
summary(glm_model)

##
## Call:
## glm(formula = STATISTICAL_MURDER_FLAG ~ PERP_RACE + PERP_SEX +
##     PERP_AGE_GROUP + VIC_RACE + VIC_SEX + VIC_AGE_GROUP + OCCUR_HOUR +
##     OCCUR_WDAY + OCCUR_MONTH + Latitude + Longitude + BORO, family = binomial,
##     data = nypd_shooting)
##
## Coefficients: (3 not defined because of singularities)
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  67.926327 144.547278   0.470  0.638409

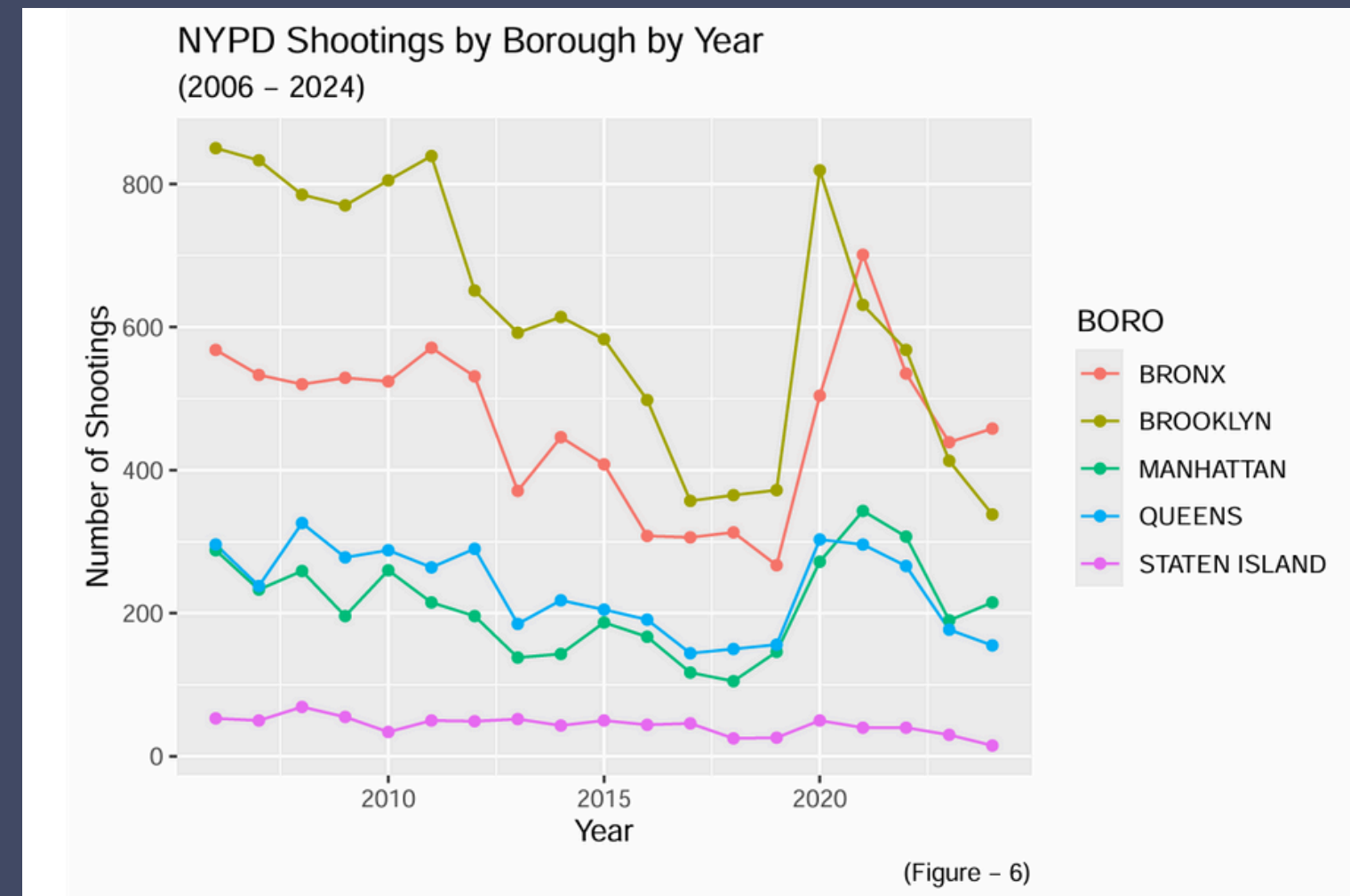
## PERP_RACE(null) ***
## PERP_RACEAMERICAN INDIAN/ALASKAN NATIVE
## PERP_RACEASIAN / PACIFIC ISLANDER ***
## PERP_RACEBLACK ***
## PERP_RACEBLACK HISPANIC ***
## PERP_RACEUnknown ***
## PERP_RACEWHITE ***
## PERP_RACEWHITE HISPANIC ***
## PERP_SEX(null)
## PERP_SEXF ***
## PERP_SEXM ***

## VIC_AGE_GROUP1022
## VIC_AGE_GROUP18-24 ***
## VIC_AGE_GROUP25-44 ***
## VIC_AGE_GROUP45-64 ***
## VIC_AGE_GROUP65+ ***
## VIC_AGE_GROUPUNKNOWN

## PERP_SEX(null)
## PERP_SEXF ***
## PERP_SEXM ***
```

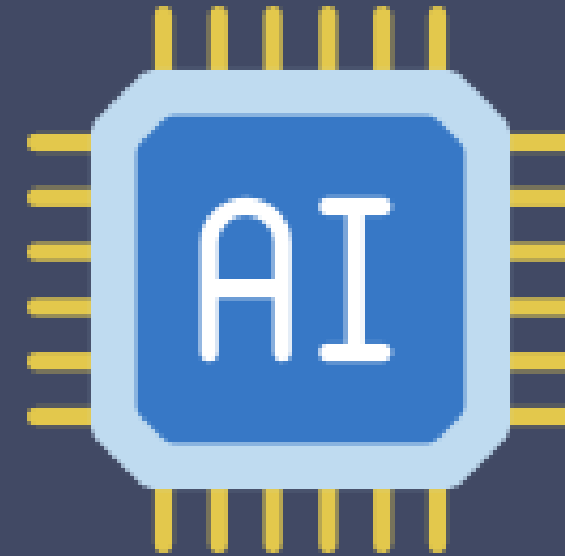


# WHAT I LEARNED



- **My assumptions were challenged** — I initially expected the Bronx to lead in incidents
- Shows the risk of relying on media-driven **stereotypes**
- Supports CNN reporting: NYC shootings rose 73% in May 2021 vs. May 2020

# FINAL THOUGHTS



- Data helps fight crime more effectively
- Summer & weekends are high-risk
- Young men are most involved
- Predictive modeling has real-world value



A wide-angle photograph of a busy city street, likely Times Square in New York City. The scene is filled with tall buildings and numerous large, brightly lit billboards. In the foreground, two NYPD patrol cars are visible on the street. One car is in the left lane, and another is further ahead in the right lane. The street is marked with white arrows and lines. On the right side of the street, two people are sitting on a low wall or barrier. The overall atmosphere is one of a bustling urban environment.

**THANK YOU FOR LISTENING**