Uncovering Hidden Patterns in NYPD Shooting Incidents

by Leveraging Analytics to Inform Public Safety

from Usman Khan

Project Overview & Objectives

Explore patterns in NYPD shooting incidents across boroughs and time

Predict fatal outcomes based on demographics using logistic regression

Identify data biases and ethical challenges in public safety analytics

Deliver insights to guide data-driven safety policies

Data Cleaning:

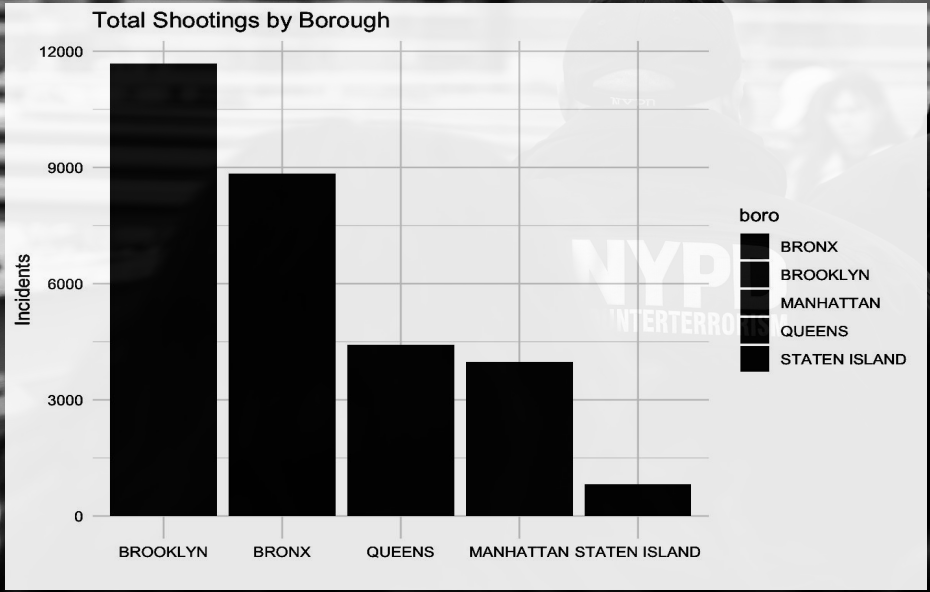
- Standardized column names for readability
- Converted date fields and extracted year/month
- Created binary flag for fatal outcomes
- Addressed missing demographic entries thoughtfully

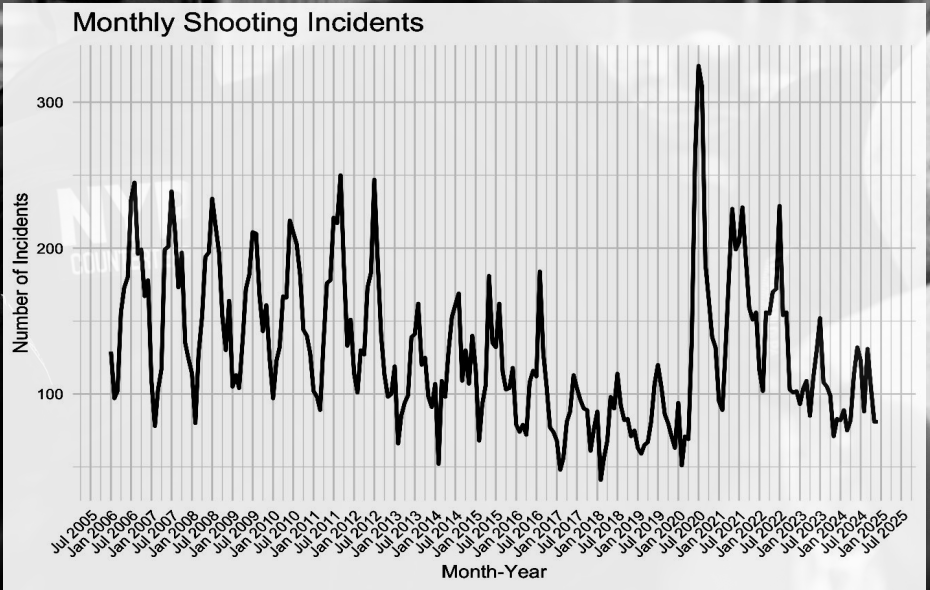
```
library(tidyverse)
library(lubridate)
library(janitor)

shootings <- read_csv("nypd-shooting-incident-data.csv")
%>%
  clean_names() %>%
  mutate(
    occur_date = mdy(occur_date),
    year = year(occur_date),
    month = month(occur_date, label = TRUE),
    is_fatal = ifelse(statistical_murder_flag == "Y", 1, 0)
)
```

Data Source:

- The dataset is from the NYPD Shooting Incident Database, containing detailed records of shooting incidents across New York City from 2006
- It includes location, date, suspect/victim demographics, and fatality outcomes.

Exploratory Data Analysis (EDA)





- Brooklyn and Bronx report the highest number of incidents
- Staten Island reports the least shootings

- Summer months show noticeable spikes in shootings
- Trends fluctuate year-over-year but spikes are consistent

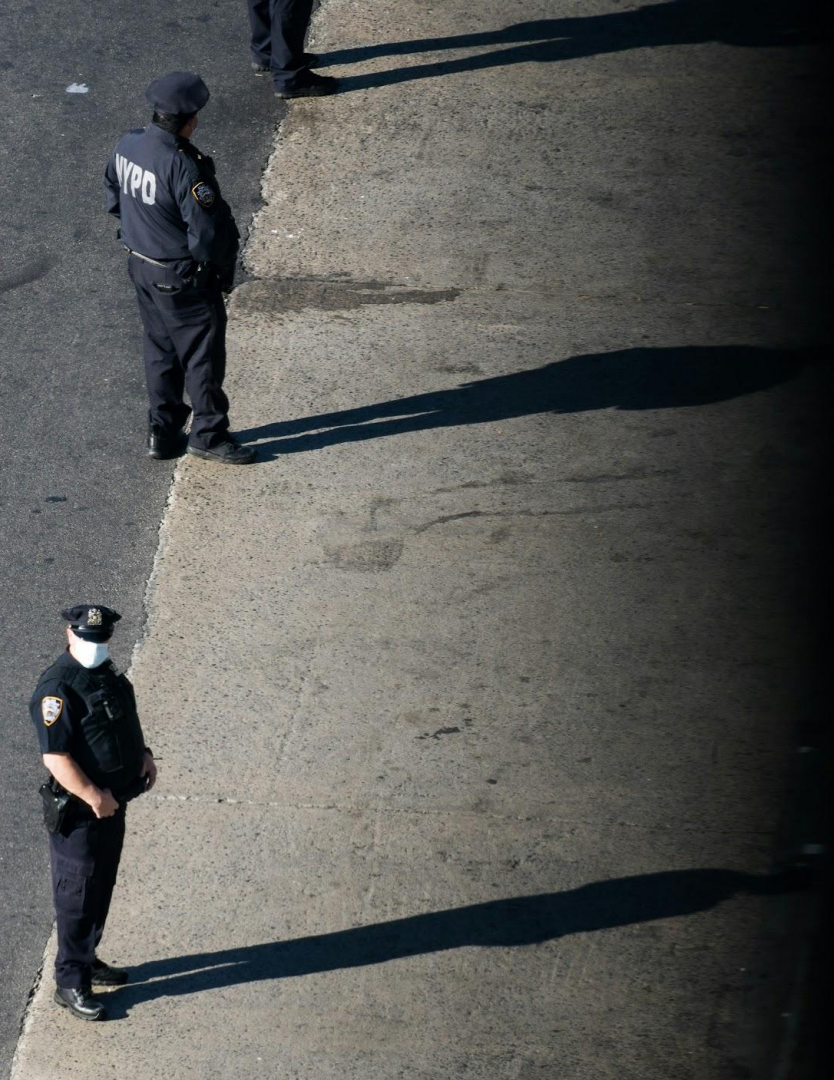
Predicting Fatality Outcomes

Approach

- Target: Predict whether a shooting incident resulted in a fatality (binary outcome: ***fatal / non-fatal***)
- Inputs: Borough, perpetrator demographics, victim demographics
- Method: Logistic regression for probability estimation
- Goal: Identify key factors associated with fatal outcomes

Findings

- Vast majority (99%) of shootings had a predicted fatality probability under 1%.
- Some variation by borough and perpetrator age, but overall fatalities were rare events.
- Logistic model was limited by missing and broad demographic data.



Bias Observed

- Reporting Bias: Inaccurate reporting or recording can easily lead to underrepresented or skewed data
- Demographic Bias: Age and sex categories are broad and potentially vague, which may oversimplify individual characteristics
- Policing Bias: Heavily policed boroughs may show artificially higher incident counts due to increased surveillance and reporting

Limitations Observed

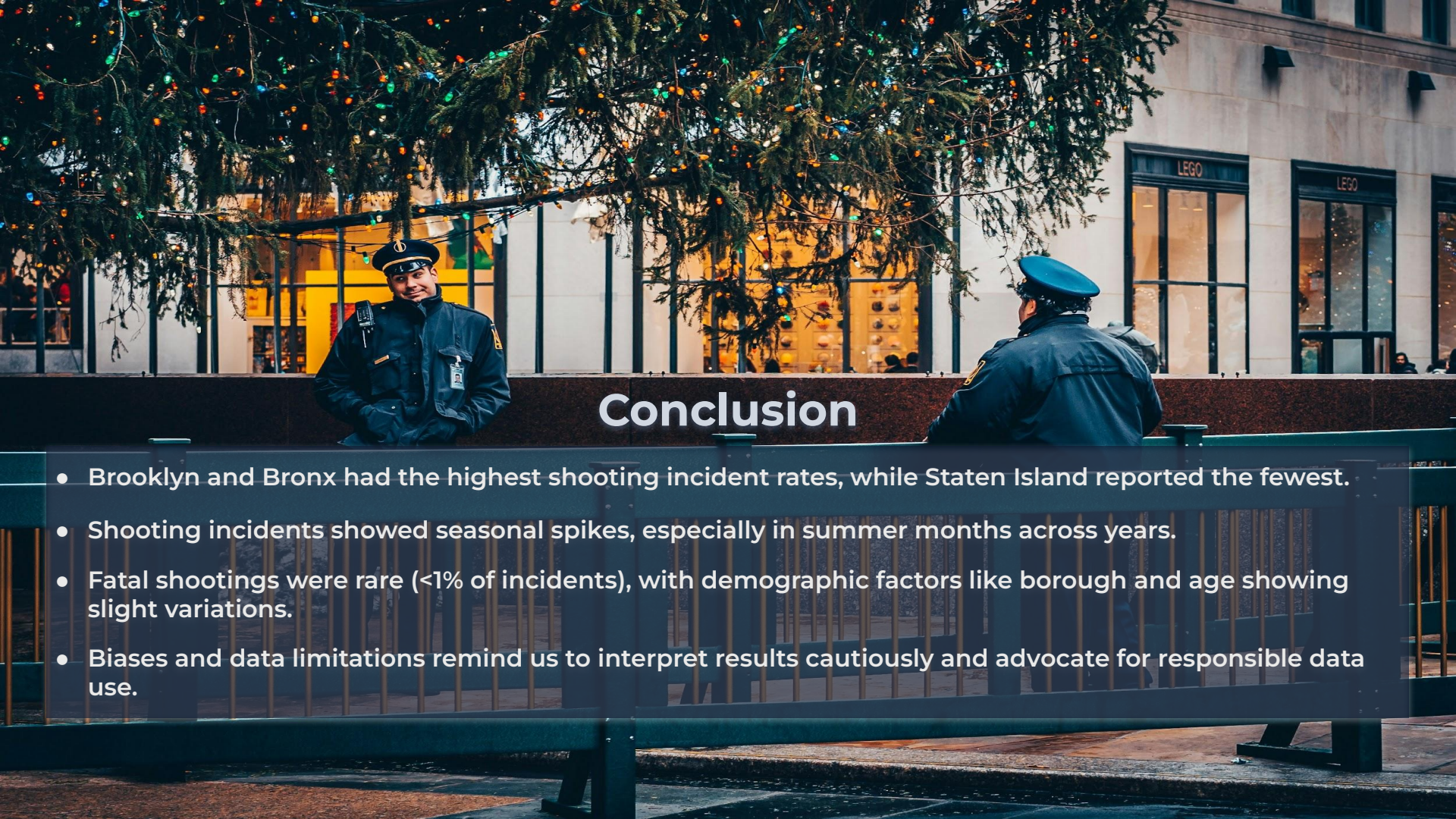
- Data Completeness: Missing demographic information (e.g., unknown suspect data) can reduce the model's reliability
- Model Assumptions: Logistic regression assumes independence between features and a linear log-odds relationship, which may be false
- *Rare Event Challenge:* Fatal shootings were rare (<1%), making model predictions naturally low-confidence.

Ethical Reflections: Data, Modeling & Responsibility

Ethical Concern	Key Takeaways
Bias Reinforcement	Predictive models based on incomplete or biased data risk reinforcing systemic disparities.
Demographic Sensitivity	Broad age/sex categories can misrepresent real people; granular, respectful representation matters.
Model Misuse	Misinterpreting low-fatality predictions could downplay serious systemic issues or misguide policy decisions.
Transparency & Reproducibility	Clear documentation and reproducibility are critical to maintain trust and accountability.

Ensuring Reproducibility & Transparency

- Full analysis conducted in RMarkdown for transparency and traceability
- All code is included and re-runs on the original dataset
- Packages used: tidyverse, janitor, and other CRAN-available libraries
- No manual data edits — fully scripted pipeline
- No randomness involved, ensuring consistent results across reruns
- Designed for repeatable, auditable, and trustworthy research



Conclusion

- Brooklyn and Bronx had the highest shooting incident rates, while Staten Island reported the fewest.
- Shooting incidents showed seasonal spikes, especially in summer months across years.
- Fatal shootings were rare (<1% of incidents), with demographic factors like borough and age showing slight variations.
- Biases and data limitations remind us to interpret results cautiously and advocate for responsible data use.

THANK YOU!

BY USMAN KHAN
APRIL 28, 2025