DTSA 5301 - DATA SCIENCE AS A FIELD

University of Colorado Boulder

NYPD SHOOTING INCIDENT DATA PRESENTATION

PRESENTED BY FILSAN MUSA

TABLE OF CONTENT

DATA DESCRIPTION

THIS DATASET COVERS **ALL REPORTED SHOOTINGS IN NYC** FROM **2006 THROUGH 2024**. EACH ENTRY SHOWS **WHERE AND WHEN THE SHOOTING HAPPENED**, AND **DEMOGRAPHIC INFO** ABOUT THE PEOPLE INVOLVED LIKE **AGE, SEX, AND RACE**.

- ROWS: 29,747
- COLUMNS: 21

NOTE: THE DATA IS EXTRACTED QUARTERLY AND CHECKED BY THE NYPD BEFORE IT'S SHARED PUBLICLY.

OBJECTIVE

OBJECTIVE

- IDENTIFY AND ANALYZE SPATIAL PATTERNS IN THE GEOGRAPHIC DISTRIBUTION OF SHOOTING INCIDENTS ACROSS NEW YORK CITY
- TO EXAMINE TEMPORAL TRENDS IN SHOOTING-RELATED INCIDENTS ACCROSS BOROUGHS

METHODOLOGY

METHODOLOGY

1

IMPORT DATA

Load necessary dependencies and import dataset

2

DATA EXPLORATION

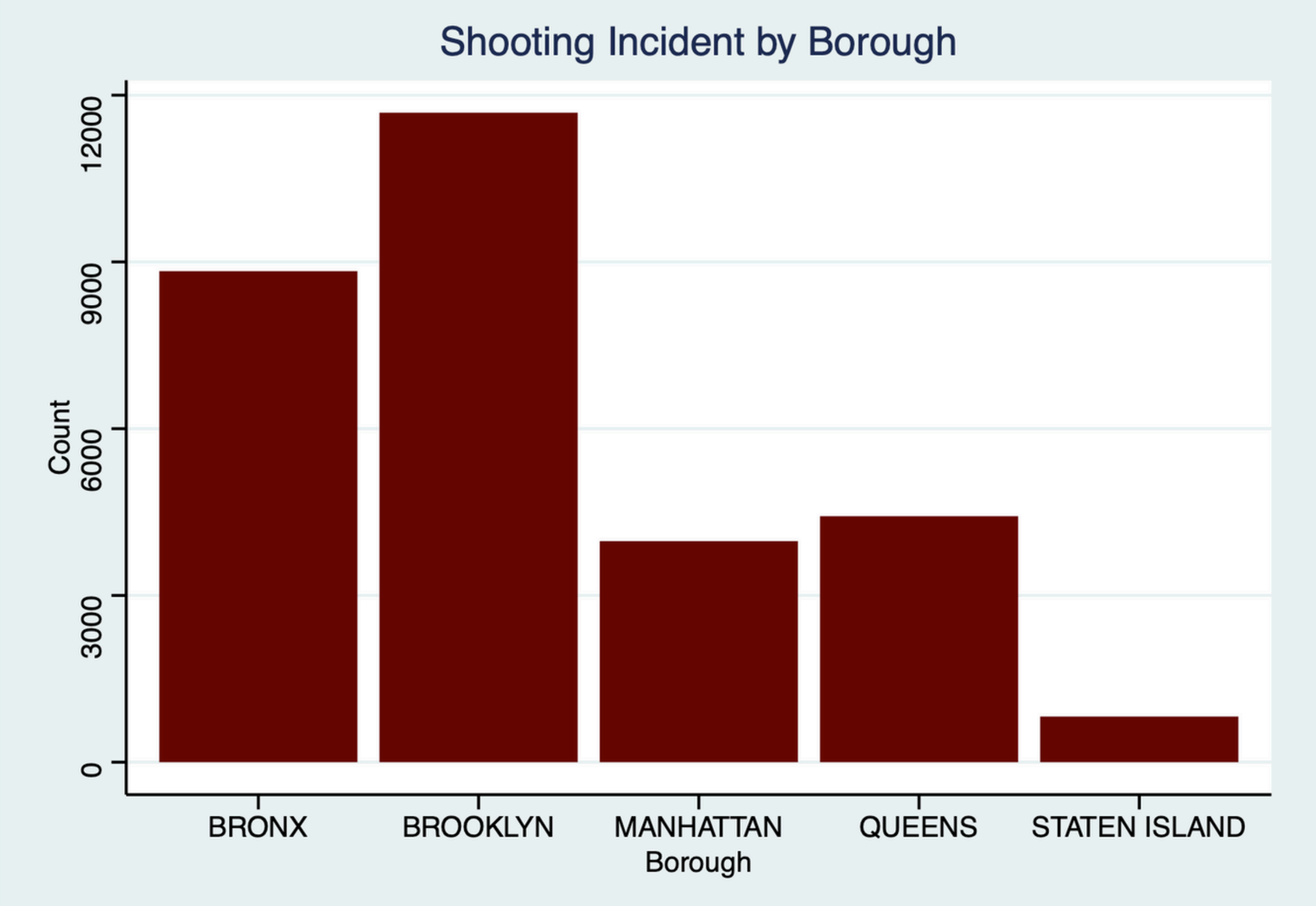
Learning about the data, and exploring relationships between variables

3

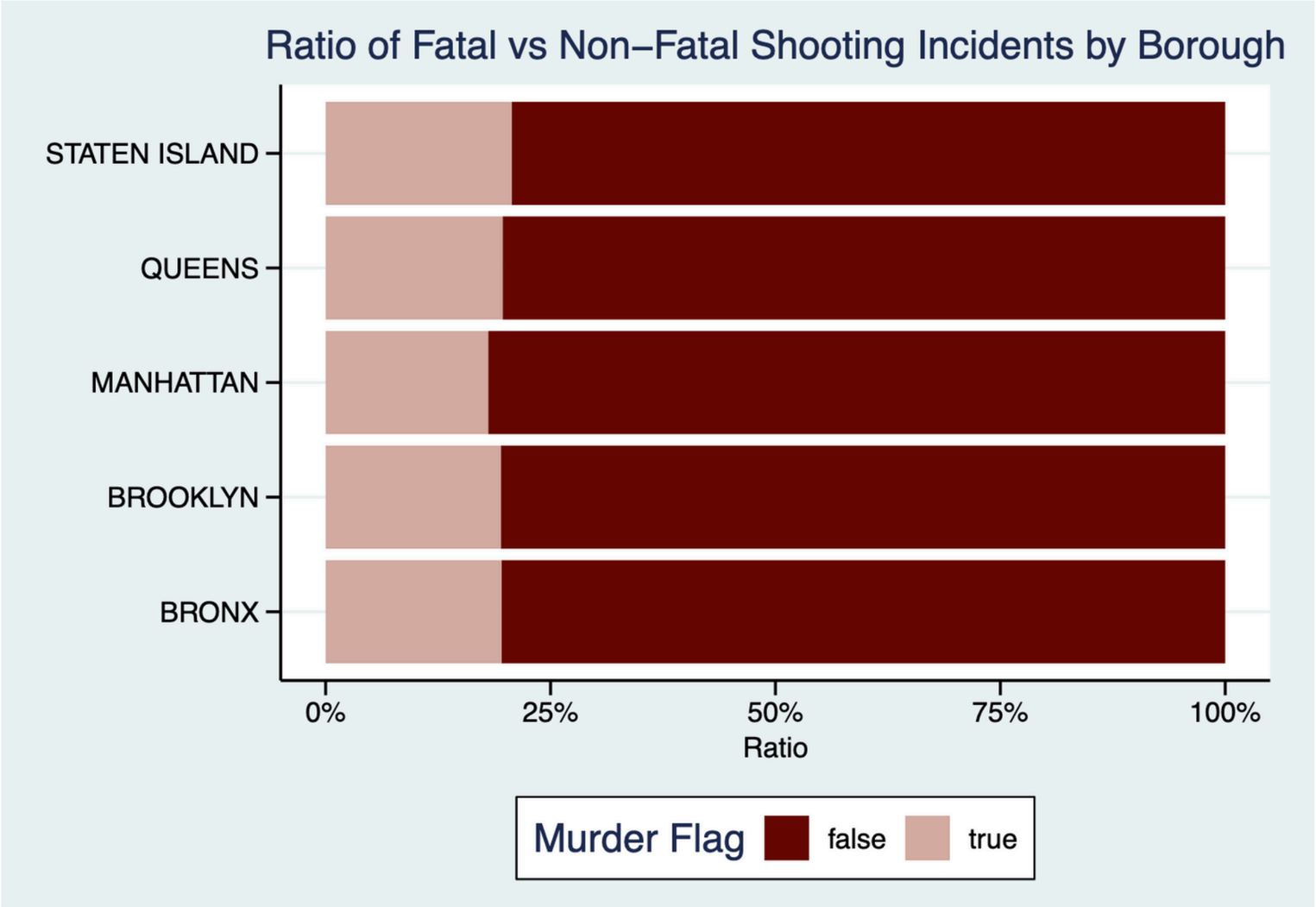
DATA CLEANING

Dropping Variables, Handling missing data, Fixing Data Types, Renaming Variables, Feature Engineering

DATA EXPLORATION



DATA EXPLORATION (CONT.)



DATA CLEANING

FEATURE ENGINEERING

CREATED MURDER_FLAG, OCCUR_DATETIME; FIXED DATA TYPES

DROP VARIABLES

DROPPED IRRELEVANT/REDUNDANT VARIABLES, AND VARIABLES W/ SUBSTANTIAL MISSING DATA

CHECK FOR DUPLICATES

THERE WERE NO DUPLICATES

HANDLE MISSING DATA

USED MEAN VALUES TO REPLACE MISSING VALUES IN COORDINATE COLUMNS

RENAME VARIABLES

RENAMED VARIABLES TO CONFORM TO VARIABLE NAMING CONVENTION (SCREAMING_SNAKE_CASE)

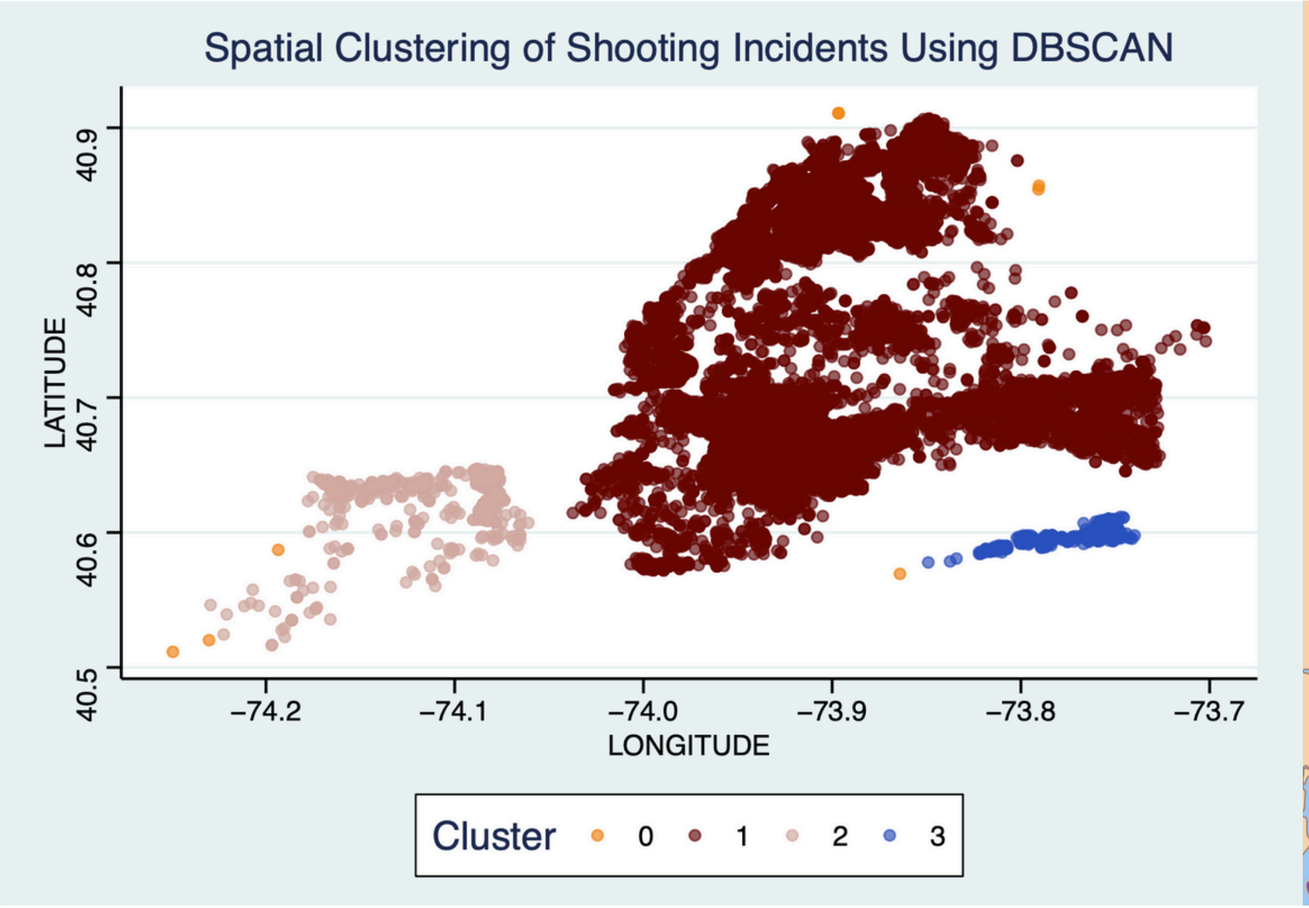
MODEL EVALUATION

QUESTIONS ON INTEREST

SPATIAL CLUSTERING MODEL:

- WHAT DOES THE CLUSTERING REVEAL ABOUT THE GEOGRAPHIC DISTRIBUTION OF INCIDENTS?
- IS THERE A DISCERNIBLE PATTERN?

SPATIAL ANALYSIS



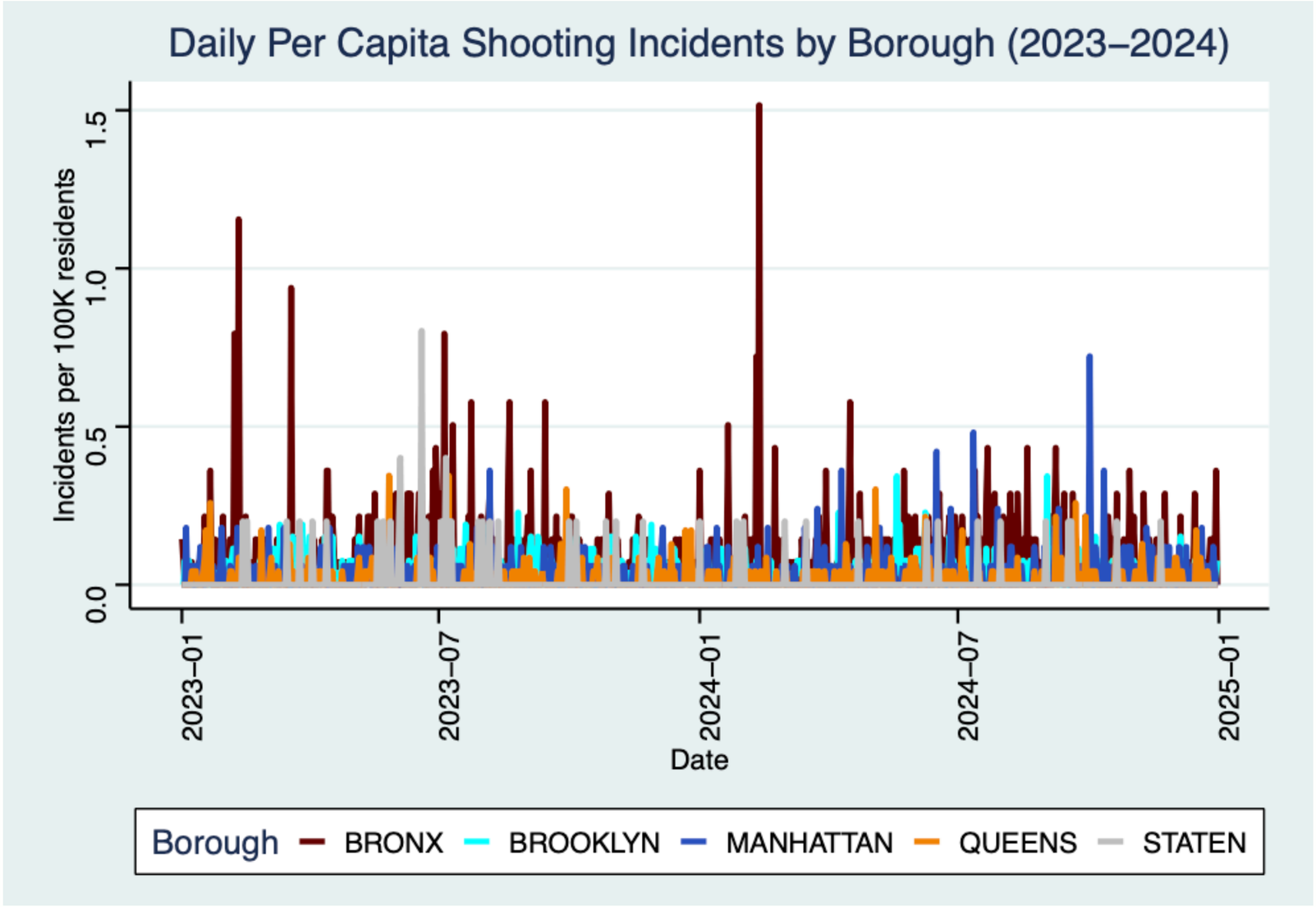
1. Manhattan (New York County) 2. Brooklyn (Kings County) 3. Queens (Queens County) 4. The Bronx (Bronx County) 5. Staten Island

QUESTIONS ON INTEREST

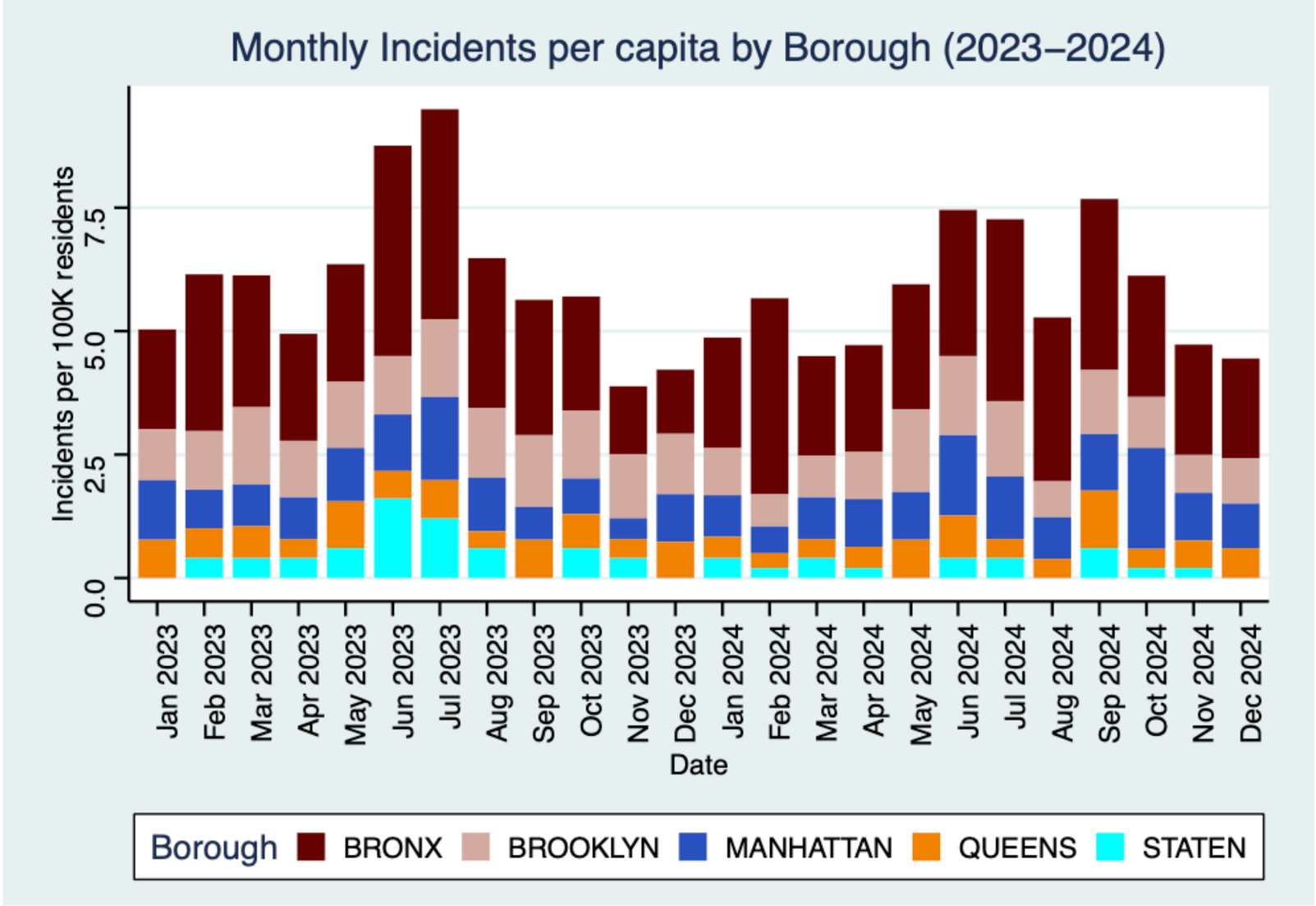
TIME SERIES MODEL:

- DO THE BOROUGHS SHOW NOTICEABLE TRENDS IN THE OCCURRENCES OF SHOOTING RELATED INCIDENTS?
- IS THERE EVIDENCE OF SEASONALITY OR RECURRING PATTERNS?

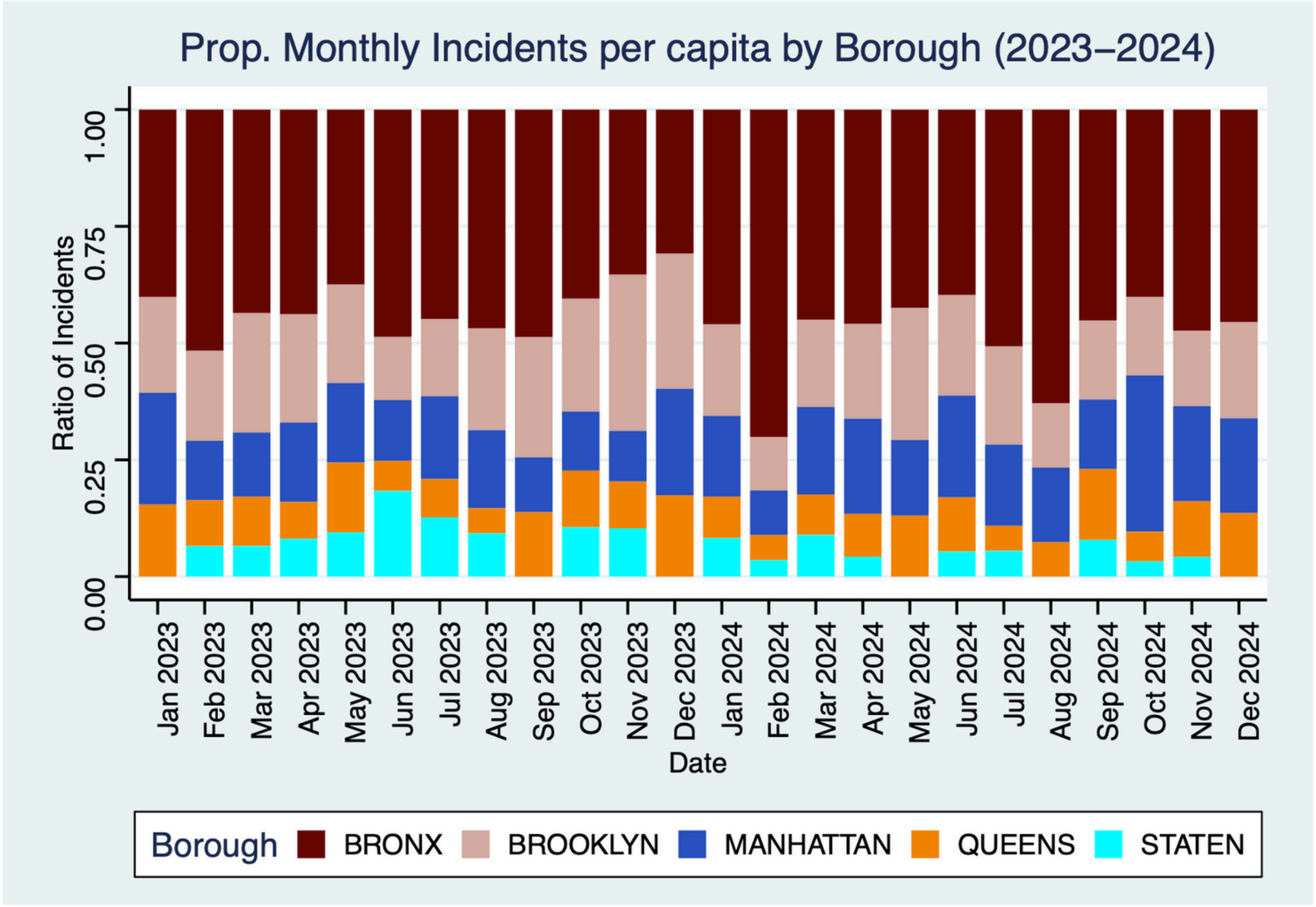
TIME SERIES ANALYSIS



TIME SERIES ANALYSIS (CONT.)



TIME SERIES ANALYSIS (CONT.)



LIMITATIONS & CHALLENGES

ASSUMPTIONS

- ## EXCLUSION OF DATA

  Excluded 2020–2022 due to COVID-related disruptions in crime patterns

- ## USED POP. ESTIMATES

  Used static 2024 population estimates, assuming minimal year-over-year change.

CHALLENGES

- ## INDENTIFYING MISSING VALUES

  Difficulty finding missing values in character and factor variables.

- ## HANDLING MISSING VALUES

  Do we impute, drop rows in NAs, or exclude the variable all togethe etc.r?

CONCLUSION

KEY FINDINGS

DBSCAN MODEL:

- SHOOTINGS WERE MOST CONCENTRATED IN THE BRONX, BROOKLYN, MANHATTAN, AND QUEENS.

TIME SERIES MODEL:

- THE BRONX HAD THE HIGHEST PER CAPITA RATE
- MORE REPORTED SHOOTING IN THE SUMMER, AND LESS IN THE WINTER

SUGGESTION:

- TARGETED INTERVENTIONS ARE NEEDED IN THE BRONX

END!