1 Introduction

The project aims to understand the relationship between traffic collisions and population distribution in New York City (NYC). By analyzing motor vehicle collision data alongside demographic information, the goal is to identify boroughs with the highest collision risks and fatalities. These insights can help city planners and policymakers design effective interventions to improve urban safety.

Three datasets from NYC Open Data were used: motor vehicle collisions data, population by community districts (2010), and a street-to-borough mapping dataset. The raw datasets contained inconsistencies, missing information, and mismatched identifiers, making them unsuitable for direct analysis. An automated data pipeline was developed to clean, transform, and integrate these datasets. This report describes the data sources, challenges, and improvements made through the pipeline and highlights key findings.

2 Questions

How can we analyze the relationship between traffic collisions and population data in NYC to identify areas with the highest collision risks and fatalities?

3 Data Sources

3.1 Descriptions of Data Sources

- **Datasource 1: Motor Vehicles Collisions - Crashes**

  This dataset contains detailed information on traffic collisions in NYC, including the date, time, location, contributing factors, and the number of fatalities or injuries. The data spans several years and includes over one million records. However, it lacks borough identifiers for some incidents and has inconsistencies in location fields such as on_street_name and off_street_name.

- **Datasource 2: NYC Population by Community Districts (2010 Census)**

  This dataset provides population counts for NYC's community districts, segmented by boroughs. The _2010_population column was the focus of this analysis, as it represents a consistent demographic snapshot. Challenges included isolating borough-level data and normalizing identifiers for integration with the collisions dataset.

- **Datasource 3: Street-to-Borough Mapping**

  This dataset maps street names to boroughs using full_stree and borocode. The borocode column uses numerical codes (e.g., 1 for Manhattan, 2 for Bronx), requiring translation into borough names for compatibility with the other datasets.

  The datasets were chosen for their relevance to the project goals, and their combined use allowed for a multi-dimensional analysis of collisions relative to population density.

3.2 Data Pipeline

The automated data pipeline was designed to address the challenges posed by the raw datasets and produce clean, integrated data for analysis. The steps for this step is as follows:

1. Data Download
   The raw datasets were downloaded directly from NYC Open Data using automated scripts to ensure reproducibility.
2. Data Cleaning
   - Missing values in critical columns (e.g., borough) were imputed by cross-referencing street names (on_street_name, off_street_name) with the Street-to-Borough mapping.
   - The _2010_population column was isolated, and invalid values were removed from the population dataset.
   - The borocode column in the Street-to-Borough dataset was translated into borough names using a mapping:
     1 → Manhattan, 2 → Bronx, 3 → Brooklyn, 4 → Queens, 5 → Staten Island.
3. Data Integration
   - The cleaned collisions dataset was enriched with borough information using the Street-to-Borough mapping.
   - Borough-level population data was merged to compute normalized collision rates (e.g., fatalities per 100,000 residents).
4. Data Storage
   The cleaned datasets were saved in SQLite databases for ease of access and analysis. Summary tables and visualizations were generated as additional outputs.

4 Results and Limitations

The automated pipeline significantly improved the quality and usability of the data. Below are the key outcomes:

4.1 Improvements in Data Quality

- **Borough Completeness:**

  In the collisions dataset, 15% of incidents initially lacked borough identifiers. Using the Street-to-Borough mapping, this was reduced to just 2%, significantly improving completeness.

- **Population Consistency:**

  Population data was filtered to focus on 2010 figures, ensuring temporal alignment with the collision data.

4.2 Limitations

- **Street Mapping Gaps:**

  Some streets in the collisions dataset could not be matched to boroughs, leading to 2% of records remaining unclassified.

- **Static Population Data:**
  Using 2010 census data does not account for demographic changes, potentially affecting collision rate accuracy.

References

NYC Open Data. (n.d.). Motor Vehicle Collisions - Crashes. Retrieved from https://data.cityofnewyork.us/

NYC Open Data. (n.d.). NYC Population by Community Districts (2010 Census). Retrieved from https://data.cityofnewyork.us/

NYC Open Data. (n.d.). Street-to-Borough Mapping. Retrieved from https://data.cityofnewyork.us/