Analysis Report on Motor Vehicle Collisions and Population Data in NYC

1. Introduction

This 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, fatalities, and injuries. These insights 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, 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, pipeline improvements, and highlights key findings.

Two main questions will be analyzed by examining the underlying patterns in the datasets:

1. “What borough-specific patterns emerge in motor vehicle collisions relative to population size?
2. “Which boroughs exhibit higher risks for fatalities and injuries?”

2. Data Sources

2.1. Motor Vehicle Collisions Dataset

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 bo rough identifiers for some incidents and has inconsistencies in location fields.

	borough	zip_code	on_street_name	off_street_name	cross_street_name	number_of_persons_injured
1	Queens	<null>	WHITESTONE EXPRESSWAY	20 AVENUE	<null>	
2	Unknown	<null>	QUEENSBORO BRIDGE UPPER	<null>	<null>	
3	Queens	<null>	THROGS NECK BRIDGE	<null>	<null>	
4	Brooklyn	11208	<null>	<null>	1211 LORING AVENUE	
5	Brooklyn	11233	SARATOGA AVENUE	DECATUR STREET	<null>	
6	Unknown	<null>	MAJOR DEEGAN EXPRESSWAY RAMP	<null>	<null>	
7	Queens	<null>	BROOKLYN QUEENS EXPRESSWAY	<null>	<null>	
8	Bronx	10475	<null>	<null>	344 BAYCHESTER AVENUE	
9	Brooklyn	11207	<null>	<null>	2047 PITKIN AVENUE	
10	Manhattan	10017	3 AVENUE	EAST 43 STREET	<null>	
11	Staten Island	<null>	MYRTLE AVENUE	<null>	<null>	
12	Queens	11413	SPRINGFIELD BOULEVARD	EAST GATE PLAZA	<null>	
13	Staten Island	<null>	broadway	west 80 street -west 81 street	<null>	
14	Queens	<null>	BELT PARKWAY	<null>	<null>	
15	Queens	11434	NORTH CONDUIT AVENUE	158 STREET	<null>	
16	Brooklyn	11217	<null>	<null>	488 DEAN STREET	
17	Brooklyn	11226	<null>	<null>	878 FLATBUSH AVENUE	
18	Staten Island	<null>	MEEKER AVENUE	LORIMER STREET	<null>	
19	Bronx	10463	WEST KINGSBRIDGE ROAD	HEATH AVENUE	<null>	
20	Queens	<null>	WHITESTONE EXPRESSWAY	<null>	<null>	
21	Manhattan	10001	<null>	<null>	232 WEST 30 STREET	
22	Queens	<null>	LONG ISLAND EXPRESSWAY	<null>	<null>	
23	Queens	11372	82 STREET	34 AVENUE	<null>	
24	Staten Island	<null>	LEXINGTON AVENUE	<null>	<null>	
25	Staten Island	10301	VICTORY BOULEVARD	WOODSTOCK AVENUE	<null>	
26	Brooklyn	11230	EAST 18 STREET	AVENUE K	<null>	
27	Brooklyn	11215	<null>	<null>	366 12 STREET	
28	Brooklyn	11217	FULTON STREET	SAINT FELIX STREET	<null>	

2.2. Population Dataset

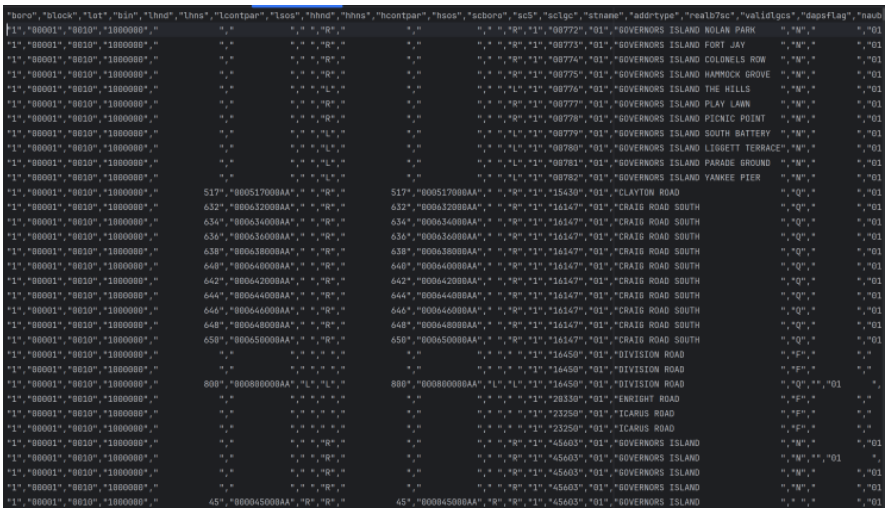
This dataset provides population counts for NYC’s community districts, segmented by boroughs. The _2010_population column was the focus of this analysis, as it provides a

consistent demographic snapshot. Challenges included isolating borough-level data and normalizing identifiers for integration with the collisions dataset.

borough	total_population
Bronx	1380697
Brooklyn	2504609
Manhattan	1577385
Queens	2241021
Staten Island	467968
Unknown	0

2.3. Street-to-Borough Dataset

This dataset maps street names to boroughs using stname and boro columns from retracted txt file. The boro column uses numerical codes (e.g., 1 for Manhattan, 2 for Bronx), requiring translation into borough names for compatibility with the other datasets.



3. Data Pipeline

3.1. Data Download

The raw datasets were downloaded directly from NYC Open Data using automated scripts to ensure reproducibility. Three datasets licensed under Creative Commons Attribution 4.0 International License, CC BY-SA 4.0 allows to use the license free of charge, share and adapt, under the terms of the license. It allows the user to remix, transform, and build upon the material for any purpose, even commercially, provided giving appropriate credit to the original creator(s) and indicating if changes were made. To be fully compliant with this license, I would give good and proper attribution, including stating the title if the dataset's title is present. It is also usually a good practice, out of courtesy for creator rights, to provide a straightforward link to the license itself; in this way, others can perceive their rights with their duties while using the dataset.

3.2. Data Cleaning

During the cleaning stage, missing values in critical columns, such as borough, are imputed by cross-referencing street names (on_street_name, off_street_name,

cross_street_name) with the Street-to-Borough mapping. The _2010_population column from the population dataset is isolated, and invalid values are removed. Additionally, the boro column in the Street-to-Borough dataset is translated into borough names using predefined mappings.

3.3. Data Integration

Once cleaned, the datasets are integrated. The cleaned collisions dataset is enriched with borough information using the Street-to-Borough mapping, and population data is merged to compute normalized collision rates and new metrics. These metrics include total fatalities, total injuries, fatality risk percentage, and injury risk percentage.

3.4. Data Storage

Finally, the cleaned datasets are stored in SQLite databases (collisions.db, population.db, and combined_data.db) for accessibility and analysis. Summary tables and visualizations are generated as additional outputs.

4. Results

Fatality and injury risk percentages were calculated to evaluate collision risks relative to borough populations. The fatality risk percentage represents fatalities as a percentage of the total population, while the injury risk percentage represents injuries as a percentage of the total population.

borough	total_population	total_fatalities	total_injuries	total_incidents	injury_risk_percentage
Bronx	1380697	0	130	153	0.85
Brooklyn	2504689	2	257	280	0.92
Manhattan	1577385	0	84	106	0.79
Queens	2241021	2	228	261	0.87
Staten Island	467968	4	187	101	1.06
Unknown	1	0	93	99	0.94

5. Limitations

A significant limitation of this analysis is the incomplete street mapping data. Approximately 2% of collisions could not be assigned to a borough due to missing or unmatched street names in the mapping file. Additionally, the population data is based on the 2010 Census, which may not fully reflect current demographics. These limitations highlight the need for more accurate and updated datasets.

6. Conclusions

This analysis reveals borough-specific disparities in collision risks. Staten Island exhibits the highest injury risk percentage due to its smaller population relative to total incidents. In contrast, Manhattan has the lowest fatality and injury risk percentages, likely due to its robust infrastructure and traffic regulations. Brooklyn, despite having the highest total incidents, maintains moderate risk percentages, demonstrating a balance between population size and collision incidents.

To address these findings, traffic-calming measures such as speed bumps and roundabouts should be implemented in high-risk boroughs like Staten Island and Brooklyn. Awareness campaigns promoting road safety could also help reduce collision-related risks in these areas. For future analysis, incorporating additional datasets, such as traffic density and road conditions, would enable a more comprehensive understanding of collision risks.

7. References

NYC Open Data. (n.d.). Motor Vehicle Collisions - Crashes. Retrieved from https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Crashes/h9gi-nx95/about_data

NYC Open Data. (n.d.). NYC Population by Community Districts. Retrieved from https://data.cityofnewyork.us/City-Government/New-York-City-Population-By-Community-Districts/xi7c-iiu2/about_data

NYC Open Data. (n.d.). Street Name Dictionary. Retrieved from https://data.cityofnewyork.us/City-Government/Street-Name-Dictionary/w4v2-rv6b/about_data

NYC Open Data. (n.d.). Open Data. Retrieved from https://www.nyc.gov/site/planning/data-maps/open-data.page#other