Running head:

Jian Ruan, Zach Yu, Carrie Hashimoto, Sophia Lyu (Undergraduate), Patrick Yee (Graduate)

February 21, 2023

1 Motivation and Overview of Primary Data.

The aim of this project is to analyze contributing factors for motor vehicle accidents and related injuries in the New York Metropolitan area. A better statistical understanding of these factors could help guide more intelligent decisions in the interest of public safety by preemptively identifying conditions under which accidents are more likely. These analyses would be of interest to officials in both the public (legislative, infrastructual) and private (insurance, engineering) sectors.

The primary dataset for this project is a list of motor vehicle accidents in New York City between 1 Januray 2013 and 31 December 2022 (MVCC). The data are collected by the New York Police Department and maintained by the City of New York [1] (1.97m rows $\times$ 29 cols) and contains date/time, vehicle type, location, and injury data for each raw observation. Since the MVCC set is complete population data (eg, contains every known car accident over the specified 10-year interval and is not a sample), the main focus of this study will be on regression and causality, not statistical inference.

2 Number and severity of accidents as a function of location.

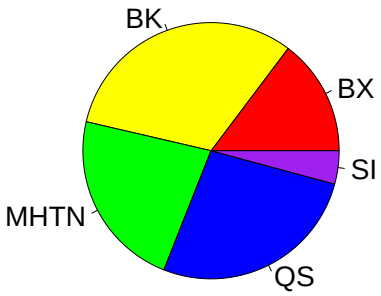

Figure 1: Number of accidents/borough, raw

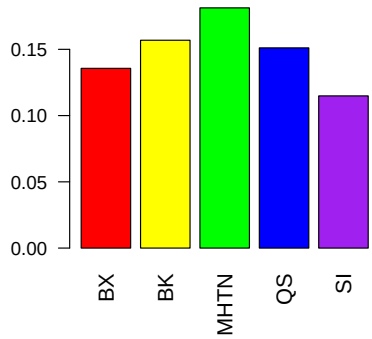

Figure 2: Number of accidents/borough, per capita

As expected, the raw number of accidents agrees with time-averaged population per borough. [2] When controlling for population, however, the number of accidents per capita do not follow a uniform distribution, with Manhattan being the most dangerous and Staten Island being the least dangerous. These observations can be verified quantitatively:

```
        Chi-squared test for given probabilities

data:  observed.counts
X-squared = 16619, df = 4, p-value < 2.2e-16
```

3 Incorporating weather data.

It was hypothesized that weather influences the likelihood of accidents, as certain humidity levels can affect visibility and precipitation/high wind speeds can create unsafe road conditions. To investigate the correlation between accident risk and weather, the MVCC was cross referenced against weather data collected by JFK international airport from 2013 to 2022 (NCDC). [3] These data (3652 rows $\times$ 21 cols) contain precipitation, wind, temperature and other weather data for each day in the interval of interest. For each day, accident counts were aggregated from the primary data and regressed against rainfall for the corresponding entry in NCDC.

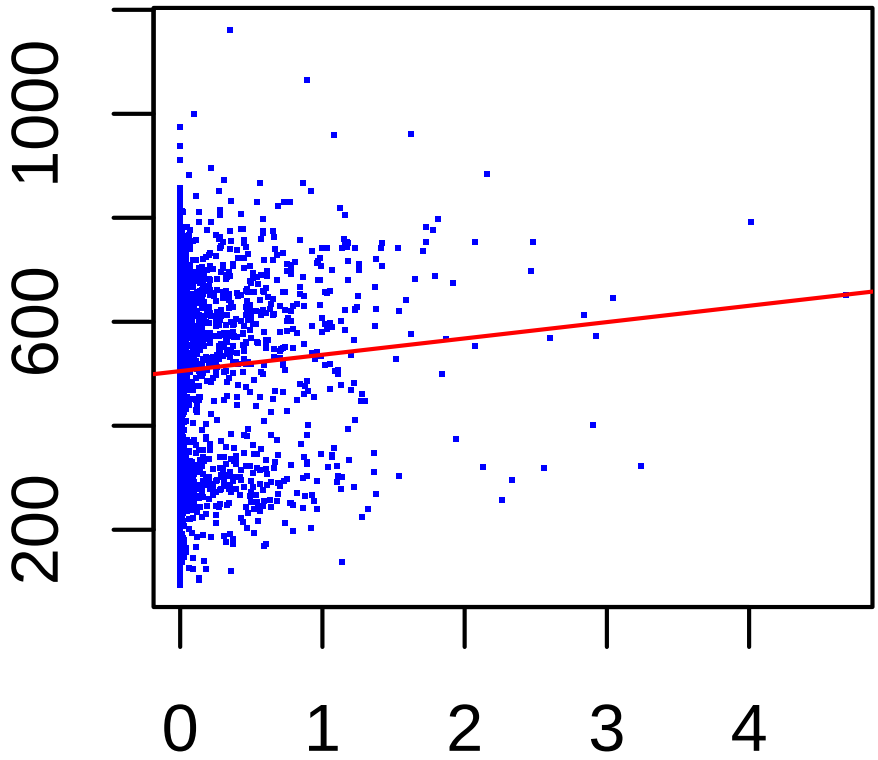

Figure 3: Precipitation (inches) versus number of crashes/day

The coefficient of this model is 31.4, which is significant ($p = .0002$).

References

[1] (NYPD), Police Department. "Motor Vehicle Collisions - Crashes: NYC Open Data." Motor Vehicle Collisions - Crashes | NYC Open Data, 17 Feb. 2023, https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Crashes/h9gi-nx95.

[2] "America." New York City Boroughs (USA): Boroughs - Population Statistics, Charts and Map, http://www.citypopulation.de/en/usa/newyorkcity/.

[3] National Centers for Environmental Information (NCEI). "Climate Data Online Search." Search | Climate Data Online (CDO) | National Climatic Data Center (NCDC), https://www.ncdc.noaa.gov/cdo-web/search.