Analyzing the relationships between car accidents and public transportation in NYC

Team 09: Yaseen Ahmed, Kelvin Borges, Pauline Mbae, Nicole Rojas, Briana Sosa

1 Table of Contents

2 Problem Overview

2.1 Context

America's metropolitan areas are often lauded as places of opportunity, amusement, and fortune, capturing the interests of millions of people worldwide as city dwellers or visitors. On top of their reputations, the sheer number of people traveling in and out of these areas put tremendous pressure on public officials and local governments to keep residents and visitors alike safe. From campaigns aimed at reducing texting while driving to signs asking drivers to buckle up, it is no secret that road safety is one of the top public safety concerns in American cities. Understanding what factors may contribute to accidents when they do occur must also be at the top of the priority list.

In addition to massive skyscrapers and iconic neighborhoods, major cities often boast public transportation systems that their visitors have never seen. In these areas, public transportation often serves as an alternative to driving. In no region is this truer than in New York, or as many call it, the "concrete jungle." Our focus in this project was examining the characteristics of accidents in NYC in terms of frequency, location, and proximity to alternative transportation means.

2.2 Specific Issue

There are millions of car crashes every year in the U.S. alone, with tens of thousands occurring daily. Newer technologies for intelligent driving are emerging, like lane assists in more recent models. The landmark $1 trillion Infrastructure bill passed in November 2021 requires future generations of cars to have DUI detection technologies to combat the menace of DUI driving in the United States. Some states have even enacted strategies like **Vision Zero,** whose primary goal is to eliminate all traffic fatalities and severe injuries while increasing safe, healthy, equitable mobility for all.

New York is one of the busiest cities worldwide regarding vehicles and human traffic, with about **66 million annual visitors** during the pre-pandemic years.

2.3 Our Solution / Approach

Our strategy was to find and visualize demographic data about public transportation in NYC, including subway ridership volumes, Citi bikes ridership, and accidents data for NYC. We explored the busiest subway stations and distance analysis of accidents within a 200m radius of the stations and accidents contributing factors. By identifying accidents prone subway stations, we can answer to what extent human traffic influences car accidents in NYC.

A few of our initial hypotheses included the following. First, there would be a relationship between subway station locations and where accidents are occuring, possibly more severe accidents would occur near public transportation hubs. Next, there would be a correlation between contributing factors and the number of persons involved, those who were injured or killed. Finally to build off the contributing factors, each borough would have similar characteristics in regards to statistics such as; contributing factors, total numbers of accidents

3 Project Overview

3.1 Impact

In the U.S. alone there are millions of car crashes every year with tens of thousands occurring daily. New York, Chicago, and Washington D.C are no strangers to such events and are even notorious for their traffic conditions. With drivers back on the road and more and more people packing trains and buses every day it is important that local agencies and the greater public understand the relationship between the two. Doing so will allow local officials to make informed policy decisions in their attempts to improve road safety. It also will provide a look at a public safety issue that the general public may not have considered before.

3.2 Technical Framework

- Data sourcing
- Data wrangling
- Exploratory Data Analysis
- Front-End Design & Data Visualization
- Communication & Stakeholder Engagement

3.3 Tools used

- Python (pandas, seaborn, matplotlib, numpy, scipy, statsmodel)
- Jupyter Notebook
- GitHub
- Google Colab
- Google Slides
- Zoom & Google Meet
- Excel
- Tableau

4 Data Analysis and Computation

4.1 Dataset details

We used three datasets for our project including: New York City Accidents Data, New York Subway Ridership Data and New York City-Bike Ridership Data.

Dataset	Source	Description	Strength & Weaknesses
2020 NYC Subway and Bus Ridership	NYC MTA	This dataset includes multiple files that describe ridership on NYC's subway and bus routes from 2015 to 2020 on an average daily basis. It provides information on every single route (train and bus) in the city as well as on station closures.	Strengths: This dataset includes information on both trains and buses allowing us to get a full picture of NYC's public transit. It also provides information on the change in ridership seen during the Covid-19 pandemic. The information on station closure will also provide an important factor to analyze against car accident prevalence. Weaknesses: The numbers provided are daily averages as opposed to exact ridership numbers. This information is also provided on a yearly basis as opposed to a daily basis as seen in the previous dataset.
Citibike System Data	Citibike	Provides downloadable files of NYC Citi Bike trip data. Includes information on day, trip start and end time, origin and final	Strengths: This information provides some insight into an increasingly more popular form of transportation that may be getting overlooked in the public transportation landscape. Information on exact location of

		station, gender, and customer type.	trip origin and end help with comparison to other datasets and mapping efforts. Weaknesses: Besides gender and user membership type (24-hr or 3-day pass, annual member, casual rider) not much is given on the rider. Data does not provide any information on what happens during the trip.Information is also sorted in files organized by month making data wrangling process arduous.
NYC Motor vehicle Accidents Data	NYC OpenData	Dataset allowed us to get a sense of where car accidents happen in NYC and provides information about location, date, severity, and weather conditions	Strengths: This dataset is inclusive of all accidents in the five boroughs of NYC. It allowed us to paint a big picture of one of the busiest cities in the world. Weaknesses: The "Unspecified" accident contributing factor is not enough data to determine the causes of accidents even though it is among the most popular contributing factors for accidents according to the data. The data is preliminary and subject to change when the MV-104AN forms are amended based on revised crash details. Police reports (MV104-AN) are required to be filled out for collisions where someone is injured or killed, or where there is at least $1000 worth of damage.

4.2 Data Wrangling and Manipulation

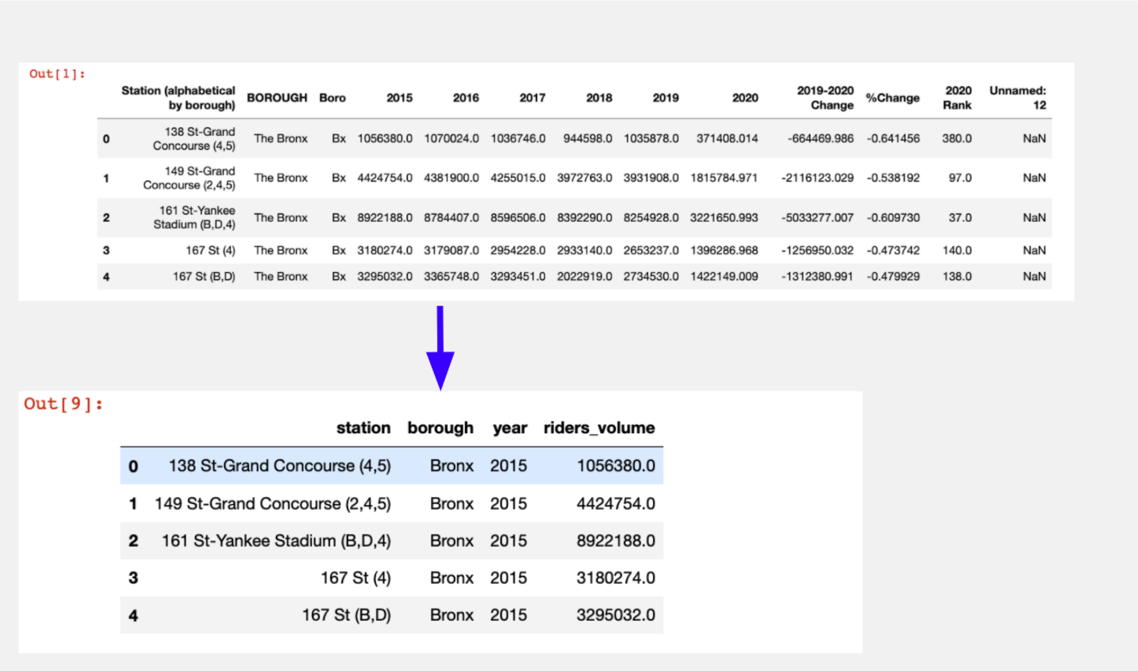
The team's primary goal during the cleanup and data wrangled journey was ensuring that we could merge data from the three datasets and draw meaningful insights from them. We hoped that at the end of the 15 weeks, we would find relationships between the distribution of accidents in NYC and access and proximity to public transportation.

Data cleaning Highlights: -

- Eliminating all data points with null values.
  - As the plan for our analysis was extremely location oriented, we removed any accident that did not provide us a zip code, latitude, and longitude. Then we proceeded to add columns that we would use later in our analysis. (i.e Total Injured, Total Killed, Severity)
- Renaming columns and categorical row values to ensure uniformity

Data Wrangling: -

- Creating new columns
- Extracting Categorical column values from strings
- Merging NYC Subway station spatial dataset with Motor Vehicle Accidents dataset
- Concatenating monthly ridership files into one large dataset
- Unpivoting data from to desirable formats, see example below



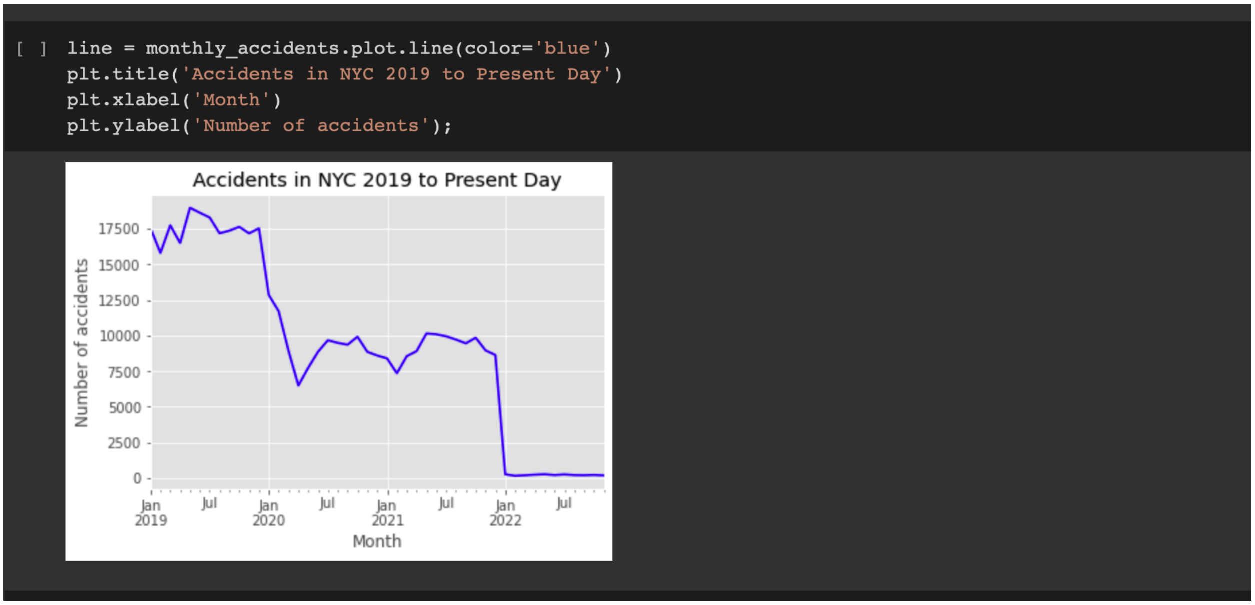
Out[1]:

	Station (alphabetical by borough)	BOROUGH	Boro	2015	2016	2017	2018	2019	2020	2019-2020 Change	%Change	2020 Rank	Unnamed: 12
0	138 St-Grand Concourse (4,5)	The Bronx	Bx	1056380.0	1070024.0	1036746.0	944598.0	1035878.0	371408.014	-664469.986	-0.641456	380.0	NaN
1	149 St-Grand Concourse (2,4,5)	The Bronx	Bx	4424754.0	4381900.0	4255015.0	3972763.0	3931908.0	1815784.971	-2116123.029	-0.538192	97.0	NaN
2	161 St-Yankee Stadium (B,D,4)	The Bronx	Bx	8922188.0	8784407.0	8596506.0	8392290.0	8254928.0	3221650.993	-5033277.007	-0.609730	37.0	NaN
3	167 St (4)	The Bronx	Bx	3180274.0	3179087.0	2954228.0	2933140.0	2653237.0	1396286.968	-1256950.032	-0.473742	140.0	NaN
4	167 St (B,D)	The Bronx	Bx	3295032.0	3365748.0	3293451.0	2022919.0	2734530.0	1422149.009	-1312380.991	-0.479929	138.0	NaN

Out[9]:

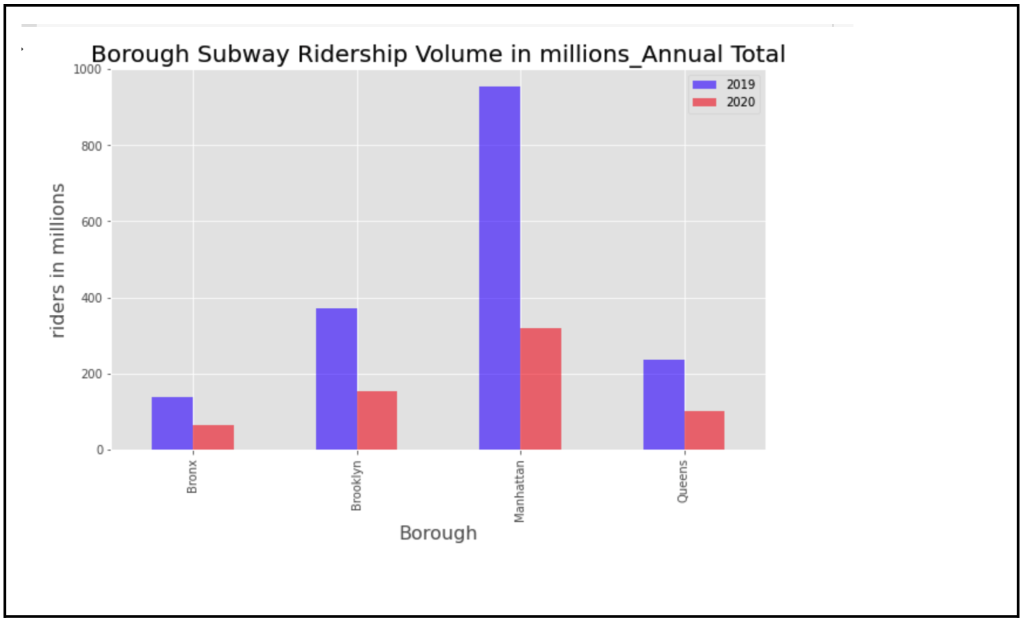
	station	borough	year	riders_volume
0	138 St-Grand Concourse (4,5)	Bronx	2015	1056380.0
1	149 St-Grand Concourse (2,4,5)	Bronx	2015	4424754.0
2	161 St-Yankee Stadium (B,D,4)	Bronx	2015	8922188.0
3	167 St (4)	Bronx	2015	3180274.0
4	167 St (B,D)	Bronx	2015	3295032.0

4.3 Exploratory Data Analysis and Results



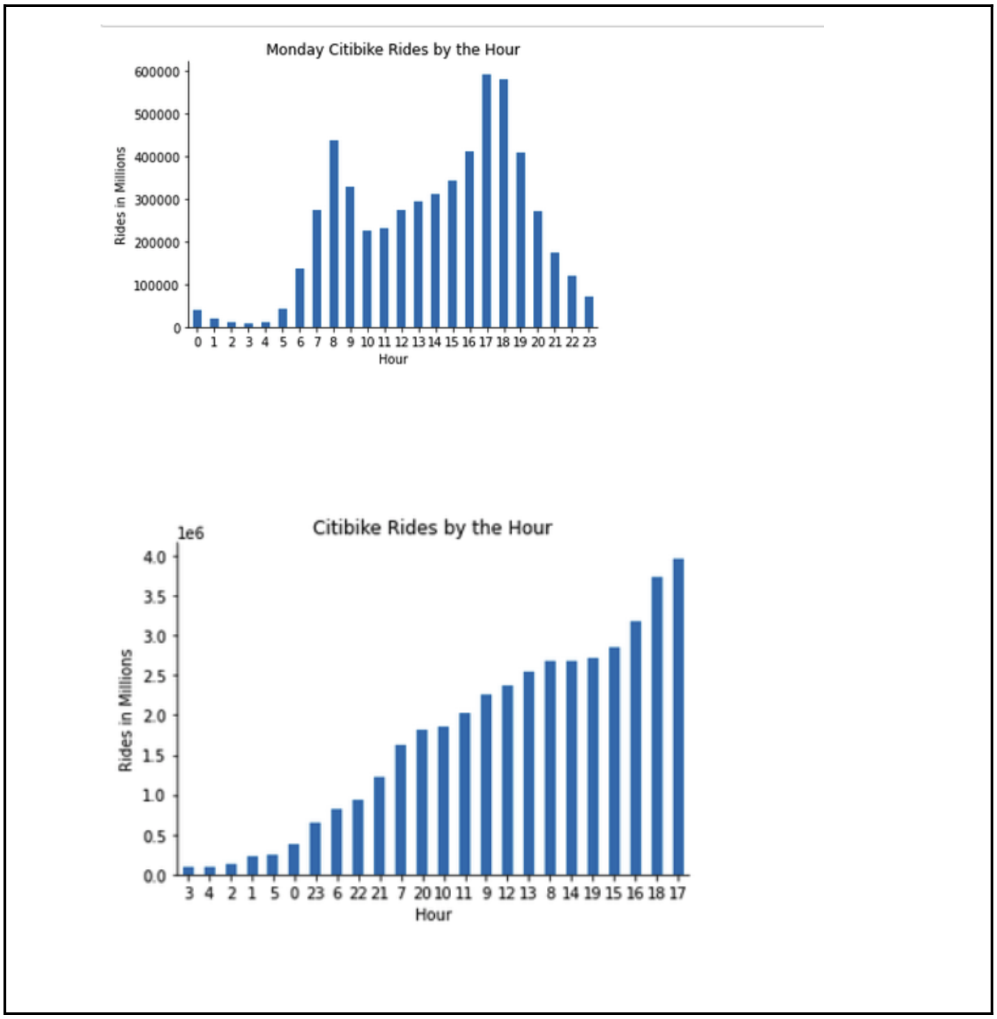
```
[ ] line = monthly_accidents.plot.line(color='blue')
    plt.title('Accidents in NYC 2019 to Present Day')
    plt.xlabel('Month')
    plt.ylabel('Number of accidents');
```

- We created these visuals during the Exploratory Data Analysis stage. Before the pandemic, NYC had about 17,000 accidents monthly, but the number reduced significantly during the COVID-19 lockdown.



- Like accidents, subway ridership volume dropped significantly during the pandemic. Pre-pandemic Manhattan had close to 1B annual riders, but the number dropped to 400m. Like accidents, subway ridership volume dropped significantly during the pandemic. Pre-pandemic Manhattan had close to 1B annual riders, but the number dropped to 400m. Subway Ridership volume was **slightly higher** during the weekday compared to the weekend. Manhattan had the most subway riders followed by Queens and Brooklyn.

- **D**irectly above we see a visual comparison between Citibikes ridership on Monday specifically and across all hours of the day regardless of day



- Directly above we see a visual comparison between Citibikes ridership on Monday specifically and across all hours of the day regardless of day of the week.

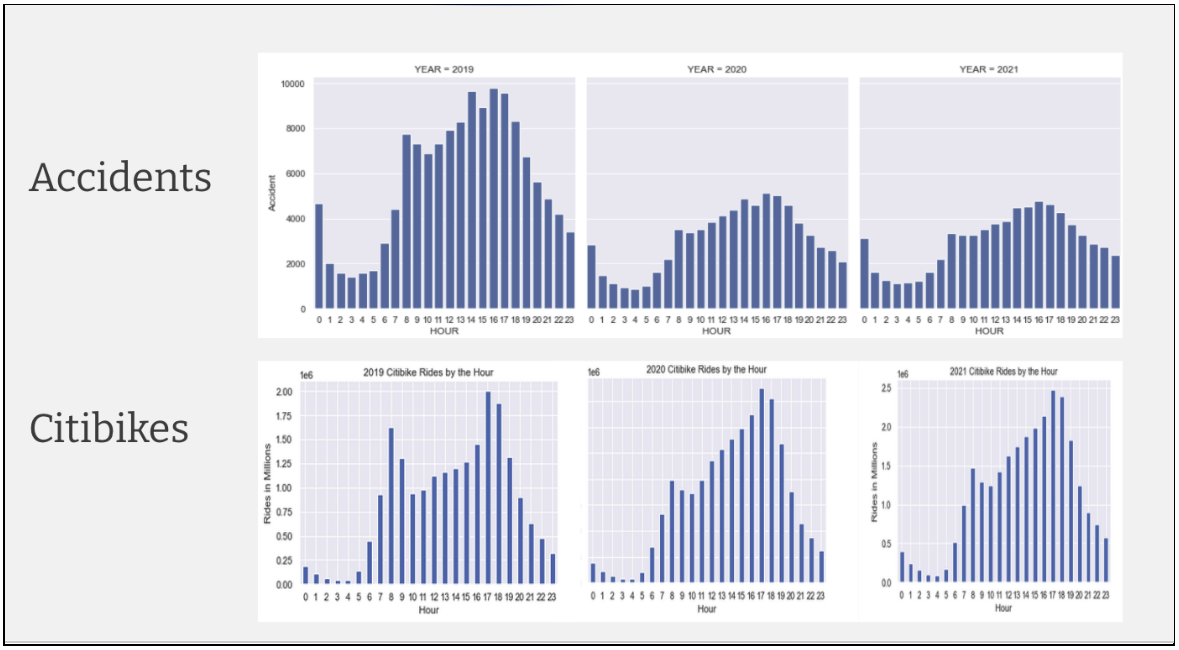
**The above visuals are just a few of the many visuals we created throughout our EDA process.

4.4 Project highlights

We mainly focused on three main things during our EDA and the bulk of our project.

1. Time-based trends across all the datasets.
2. Severity Score
3. Distance Analysis.

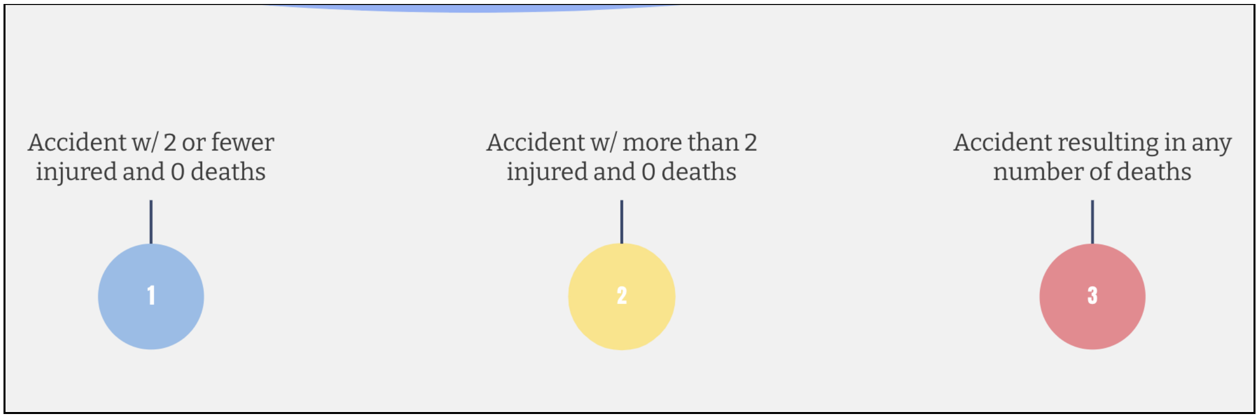
Accidents and Citi bike rides followed the **same time-based trends**; they tended to spike during the morning commute to work, and in the evening, the trends remained relatively consistent over the years.



Accidents

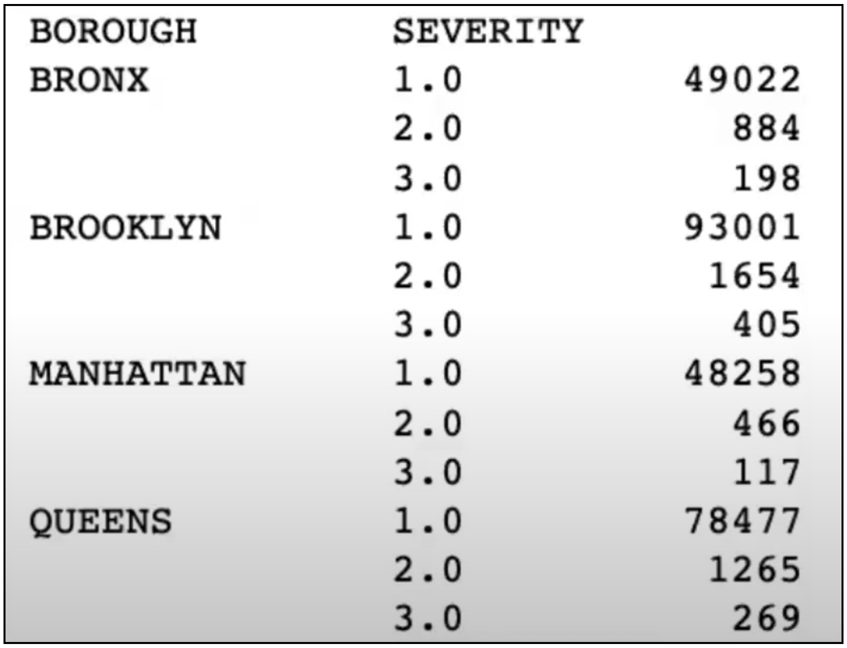
Citibikes

Severity score measured how much physical harm was sustained by motorists or passengers in the cars and pedestrians, as shown below.


Accident w/ 2 or fewer injured and 0 deaths
1
Accident w/ more than 2 injured and 0 deaths
2
Accident resulting in any number of deaths
3

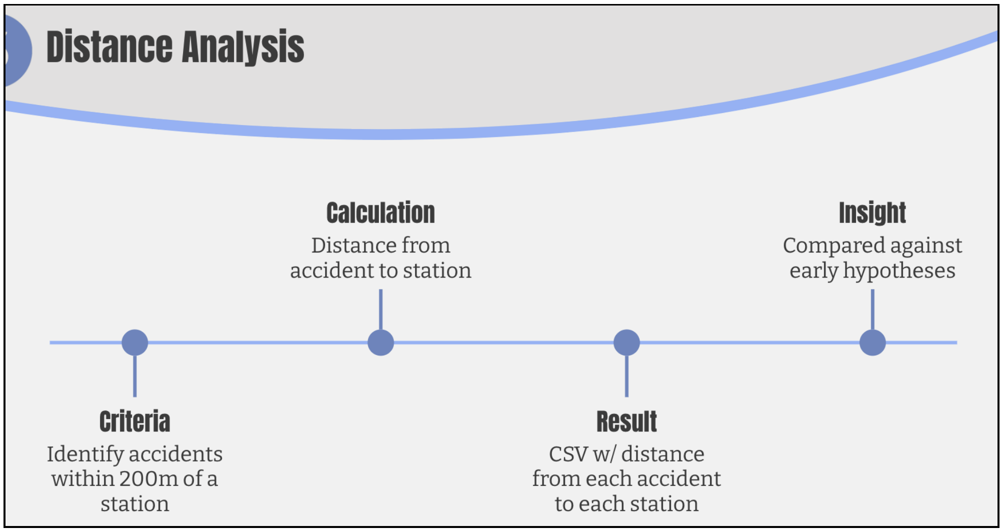
It was a rather exciting discovery that while the NYC boroughs differed in size and volume of motorists, train riders, and city bike users, every borough averaged a severity score of one , which meant that two or fewer getting hurt weren't fatal accidents.

Here is a summary of the number of accidents and severity scores across the boroughs: -



```
BOROUGH        SEVERITY
BRONX          1.0         49022
               2.0           884
               3.0           198
BROOKLYN       1.0         93001
               2.0          1654
               3.0           405
MANHATTAN      1.0         48258
               2.0           466
               3.0           117
QUEENS         1.0         78477
               2.0          1265
               3.0           269
```

We calculated how far the accidents were occurring from the nearest subway stations for the **distance analysis.**



Distance Analysis

Calculation
Distance from accident to station

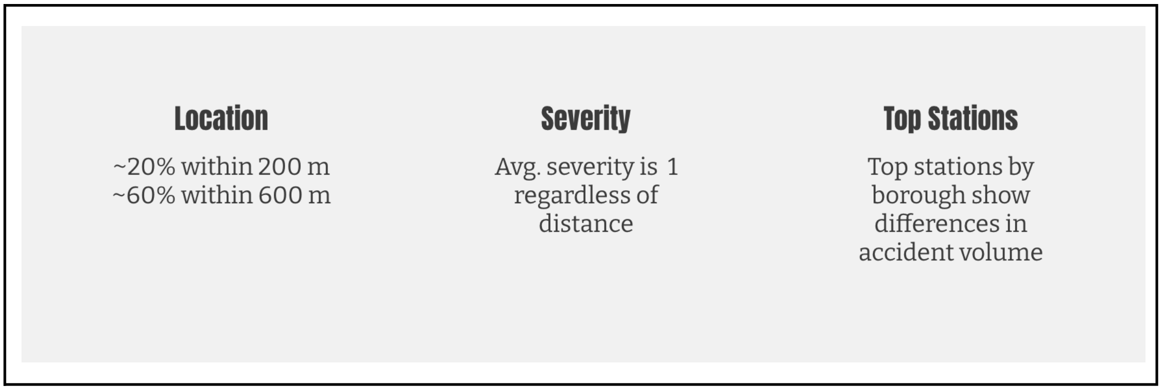
Insight
Compared against early hypotheses

Criteria
Identify accidents within 200m of a station

Result
CSV w/ distance from each accident to each station

Using spatial data in our datasets (latitudes and longitudes), the distance in meters from the location of an accident to the nearest subway stations and all subway stations in our datasets.

- About 20% of our accidents with an average severity score of 1 occurred within 200m from the subway station, approximately two building blocks.
- An estimated 60% of the accidents occurred within 600m from the subway stations.
- Even though the severity score and distance analysis were consistent across the boroughs, top stations by borough showed differences in the volume of accidents.



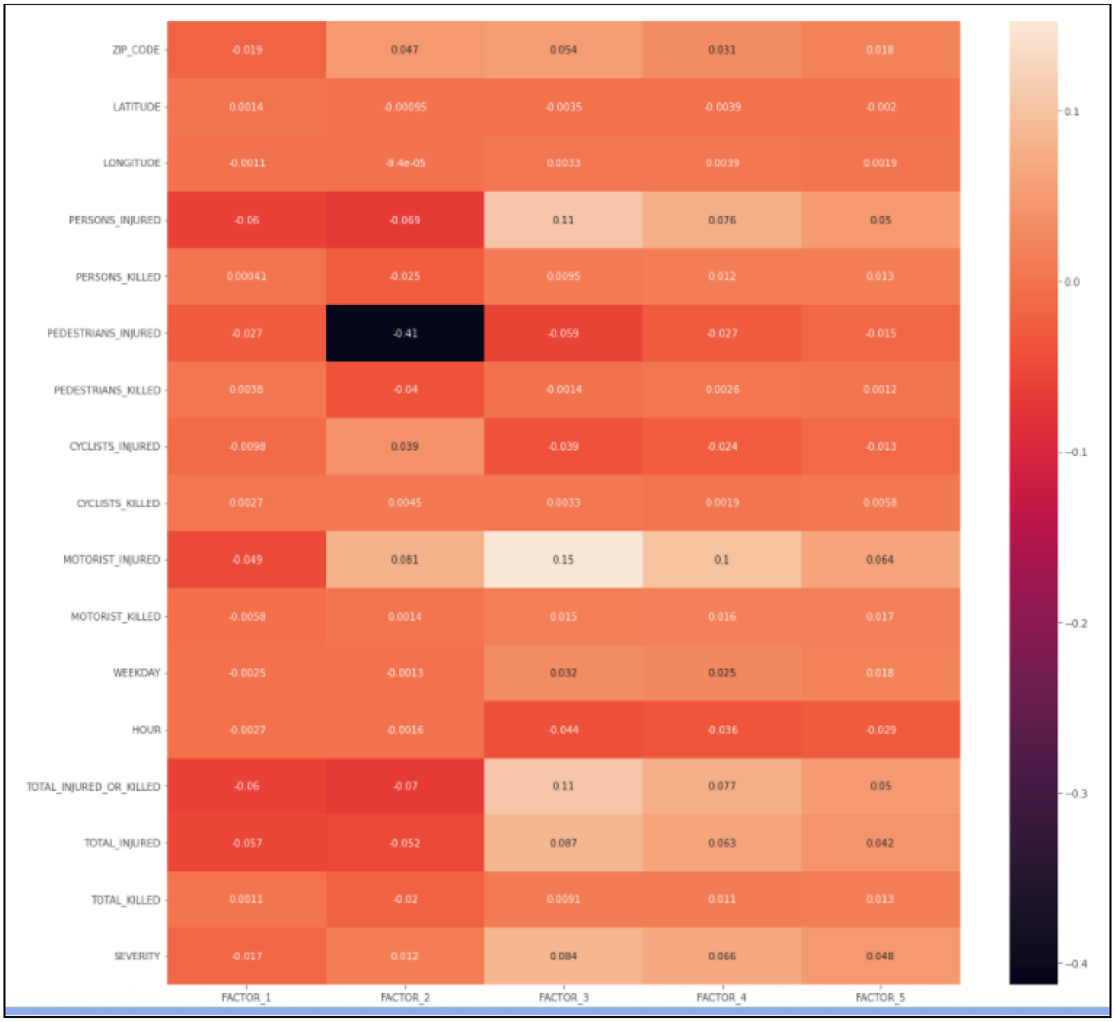
Location
~20% within 200 m
~60% within 600 m

Severity
Avg. severity is 1 regardless of distance

Top Stations
Top stations by borough show differences in accident volume

4.5 Statistical Analysis and Modeling

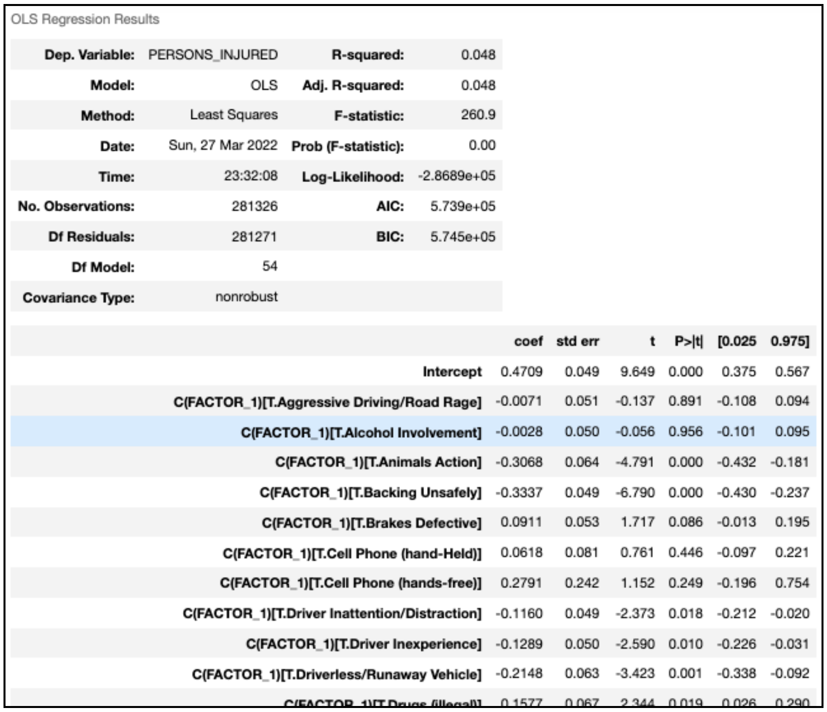
In our analysis of the Motor Vehicle Accidents datasets, we attempted to examine the Contributing Factors of the car accidents in relation to any other data point that was within the dataset provided. We began this analysis by first creating a correlation matrix to see if any relationships were either very strongly or negatively correlated. As you can see in the matrix below, most of the values remained below 0.1.



	FACTOR_1	FACTOR_2	FACTOR_3	FACTOR_4	FACTOR_5
ZIP_CODE	-0.019	0.047	0.054	0.031	0.018
LATITUDE	0.0014	-0.00095	-0.0035	-0.0039	-0.002
LONGITUDE	-0.0011	-9.4e-05	0.0033	0.0039	0.0019
PERSONS_INJURED	-0.06	-0.069	0.11	0.076	0.05
PERSONS_KILLED	0.00041	-0.025	0.0095	0.012	0.013
PEDESTRIANS_INJURED	-0.027	-0.41	-0.059	-0.027	-0.015
PEDESTRIANS_KILLED	0.0038	-0.04	-0.0014	0.0026	0.0012
CYCLISTS_INJURED	-0.0098	0.039	-0.039	-0.024	-0.013
CYCLISTS_KILLED	0.0027	0.0045	0.0033	0.0019	0.0058
MOTORIST_INJURED	-0.049	0.081	0.15	0.1	0.064
MOTORIST_KILLED	-0.0058	0.0014	0.015	0.016	0.017
WEEKDAY	-0.0025	-0.0013	0.032	0.025	0.018
HOUR	-0.0027	-0.0016	-0.044	-0.036	-0.029
TOTAL_INJURED_OR_KILLED	-0.06	-0.07	0.11	0.077	0.05
TOTAL_INJURED	-0.057	-0.052	0.087	0.063	0.042
TOTAL_KILLED	0.0011	-0.02	0.0091	0.011	0.013
SEVERITY	-0.017	0.012	0.084	0.066	0.048

This depicts the weak correlation in relation to this aspect of our analysis. To ensure this was accurate, we then followed up with creating linear regressions for various dependent variables. Our main test was again focused on the contributing factors. The highest r-squared value depicted in each of our tests remained below .005. As with the correlation matrix, we did not have the ability to argue there was a strong relationship between these factors and the rest of the dataset.

When asking the question “why?”, we felt the descriptors for contributing factors and the lack thereof for certain descriptions, was a strong reason a correlation was not found. Much of the data contained “Unspecified” as a category, and the total number of unique options was 55 for contributing factors. Beyond the chance that no correlation exists, this could have been the cause for such a low r-squared.



OLS Regression Results

Dep. Variable:	PERSONS_INJURED	**R-squared:**	0.048
Model:	OLS	**Adj. R-squared:**	0.048
Method:	Least Squares	**F-statistic:**	260.9
Date:	Sun, 27 Mar 2022	**Prob (F-statistic):**	0.00
Time:	23:32:08	**Log-Likelihood:**	-2.8689e+05
No. Observations:	281326	**AIC:**	5.739e+05
Df Residuals:	281271	**BIC:**	5.745e+05
Df Model:	54		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	0.4709	0.049	9.649	0.000	0.375	0.567
C(FACTOR_1)[T.Aggressive Driving/Road Rage]	-0.0071	0.051	-0.137	0.891	-0.108	0.094
C(FACTOR_1)[T.Alcohol Involvement]	-0.0028	0.050	-0.056	0.956	-0.101	0.095
C(FACTOR_1)[T.Animals Action]	-0.3068	0.064	-4.791	0.000	-0.432	-0.181
C(FACTOR_1)[T.Backing Unsafely]	-0.3337	0.049	-6.790	0.000	-0.430	-0.237
C(FACTOR_1)[T.Brakes Defective]	0.0911	0.053	1.717	0.086	-0.013	0.195
C(FACTOR_1)[T.Cell Phone (hand-Held)]	0.0618	0.081	0.761	0.446	-0.097	0.221
C(FACTOR_1)[T.Cell Phone (hands-free)]	0.2791	0.242	1.152	0.249	-0.196	0.754
C(FACTOR_1)[T.Driver Inattention/Distraction]	-0.1160	0.049	-2.373	0.018	-0.212	-0.020
C(FACTOR_1)[T.Driver Inexperience]	-0.1289	0.050	-2.590	0.010	-0.226	-0.031
C(FACTOR_1)[T.Driverless/Runaway Vehicle]	-0.2148	0.063	-3.423	0.001	-0.338	-0.092
C(FACTOR_1)[T.Drugs (illegal)]	0.1577	0.067	2.344	0.019	0.026	0.290

5 Dashboard Description

Our front-end dashboard provides visual representation of the insights and trends we found through our work. This includes graphs detailing the top 3 contributing factors to accidents and counts for the number of subway riders with adjustable filters for year and borough. Our dashboard is highlighted by a heatmap revealing the subway stations with the most accidents within 200 meters— also filterable by year and borough.

This dashboard can be used by drivers and public transportation users alike to get a better sense of where accidents are occurring most and where they should be the most cautious. It can also be used by local government agencies to identify potential areas of focus in NYC’s battle against road accidents.

6 Conclusions and Future Work

We did not find a strong relationship between station location & where accidents occur, their severity, or contributing factors. However, this gives leeway for further exploration, we have the ability to expand analysis on other contributing factors for accidents, delve into neighborhood data where accidents have occured and, explore the demographics of these respective areas. Another piece of data we can incorporate are other transportation options such as bus data. Finally, other metropolitan areas like Washington DC, Chicago, and Los Angeles can also be examined. Each of these cities contain large populations and robust public transportation systems.

Further Exploration

- Analyze other contributing factors, demographic information of neighborhoods with most accidents like proximity to landmarks and road design factors like busy intersections.
- Incorporate more transportation options available in NYC
- Research other metropolitan areas like Washington DC, Chicago and Los Angeles.

7 References / Sources

- Motor Vehicle Collisions - Crashes - Source: https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Crashes/h9gi-nx95
- Citi Bike System Data - Source: https://ride.citibikenyc.com/system-data
- MTA Subway and Bus Ridership 2020 - Source: https://new.mta.info/agency/new-york-city-transit/subway-bus-ridership-2020