From Stats to Safety: Analyzing Traffic Accidents in NYC

Kailai Wang, Kaiwen Tong, Yuchi Liu, Xuyan Zhao

Explorer Transportation Data Science Project

COLUMBIA UNIVERSITY IN THE CITY OF NEW YORK

NORTHEAST BIG DATA INNOVATION HUB

National Student Data Corps

Background and Research Purpose

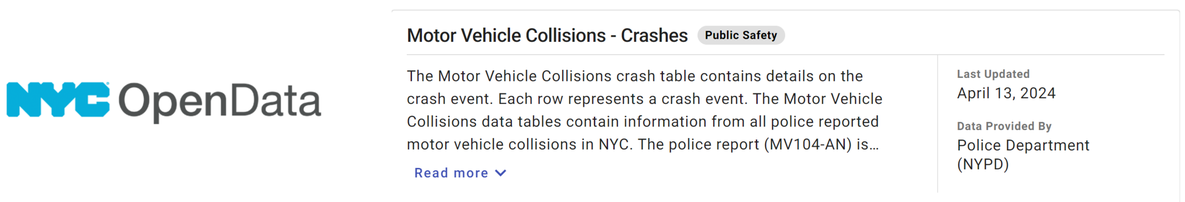
Background: By analyzing a dataset of traffic crashes in New York City for the year 2021, this project aims to shed light on the core factors and trends of traffic crashes in the city. New York City is one of the most traffic-congested cities in the United States and the world[1]. New York's status as a highly busy metropolis with a high volume of traffic flow and pedestrian activity on a daily basis makes it particularly important to understand the context of traffic crashes.

Purpose: Through comprehensive analysis of a detailed dataset containing over 45,000 accidents,

- Reveal when and where accidents occur, their common causes, and the impact of different types of traffic participants.
- Explore areas with high accident rates to help relevant authorities formulate more effective road safety policies and preventive measures.
- Expect to contribute data-supported insights for reducing the occurrence of traffic accidents and improving public safety.
- This report will cover aspects ranging from the time distribution of accidents to the types of victims, providing valuable information for urban planners, policy makers, and the public.

Data Preparation

Data Source: NYC OpenData Motor Vehicle Collisions - Crashes dataset



According to NYC Open Data, "each row represents a crash event. The Motor Vehicle Collisions data tables contain information from all police reported motor vehicle collisions in NYC. Given the sheer volume of data, this project focuses on traffic accident records for the year 2021.

Data Selection: This project focuses on traffic accident records for the year 2021.

Data Cleaning and Preprocessing

- Perform data cleansing, including removing duplicate records, dealing with missing values, etc.
- Ensure the quality of the data and the accuracy of the analysis.

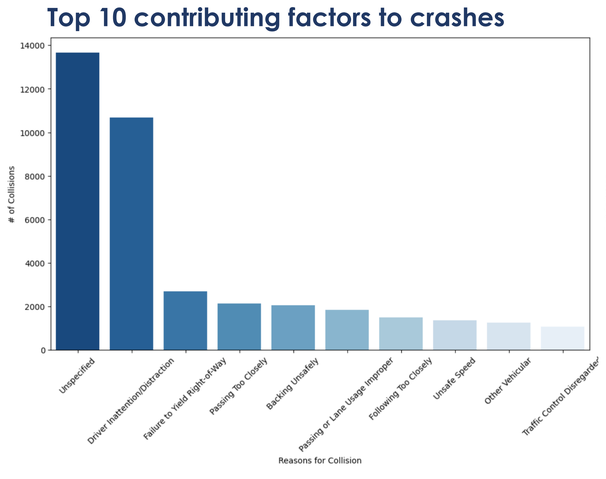
	CRASH DATE	CRASH TIME	BOROUGH	ZIP CODE	LATITUDE	LONGITUDE	LOCATION	NUMBER OF PERSONS INJURED	NUMBER OF PERSONS KILLED	NUMBER OF PEDESTRIANS INJURED	...	CONTRIBUTING FACTOR VEHICLE 1	CONTRIBUTING FACTOR VEHICLE 2	CONTRIBUTING FACTOR VEHICLE 3	CONTRIBUTING FACTOR VEHICLE 4	CONTRIBUTING FACTOR VEHICLE 5	VEHICLE TYPE CODE 1	VEHICLE TYPE CODE 2	VEHICLE TYPE CODE 3	VEHICLE TYPE CODE 4	VEHICLE TYPE CODE 5
0	1/1/2021	19:30:00	BRONX	10463	40.882700	-73.89273	(40.8827, -73.89273)	0	0	0	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1	1/1/2021	7:40:00	BROOKLYN	11218	40.637910	-73.97864	(40.63791, -73.97864)	0	0	0	...	Unspecified	Unspecified	NaN	NaN	NaN	Station Wagon/Sport Utility Vehicle	Taxi	NaN	NaN	NaN
2	1/1/2021	4:51:00	BROOKLYN	11212	40.660090	-73.90055	(40.66009, -73.90055)	0	0	0	...	Other Vehicular	Other Vehicular	NaN	NaN	NaN	Sedan	Station Wagon/Sport Utility Vehicle	NaN	NaN	NaN
3	1/1/2021	16:14:00	BROOKLYN	11237	40.705807	-73.93176	(40.705807, -73.93176)	0	0	0	...	Passing Too Closely	Unspecified	NaN	NaN	NaN	Sedan	NaN	NaN	NaN	NaN

Dataset contains 45,152 traffic accident records with 25 fields covering the following information:

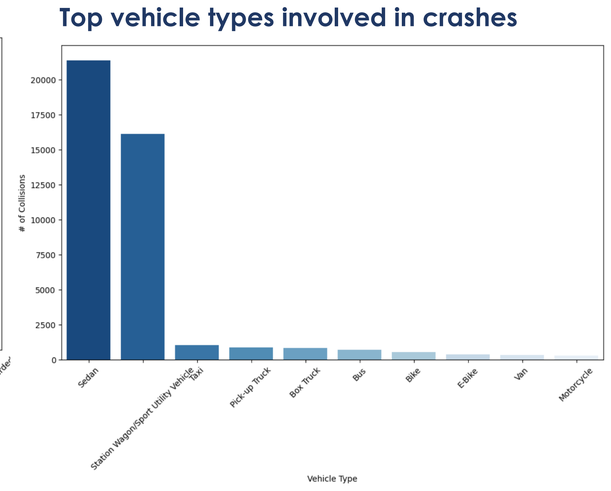
- CRASH DATE & CRASH TIME - The date and time of the accident.
- BOROUGH & ZIP CODE - The borough and zip code where the accident occurred.
- LATITUDE & LONGITUDE - Latitude and longitude of the incident location.
- LOCATION - A detailed description of the location.
- NUMBER OF PERSONS INJURED & KILLED - A count of the number of people injured and killed.
- NUMBER OF PEDESTRIANS/CYCLISTS/MOTORISTS INJURED & KILLED - Injury and fatality statistics for pedestrians, cyclists and motorists.
- CONTRIBUTING FACTOR VEHICLE 1-5 - Causes of accidents for up to five vehicles.
- VEHICLE TYPE CODE 1-5 - Type descriptions for up to five vehicles.

Data Visualization

Top 10 contributing factors to crashes



Top vehicle types involved in crashes



Understanding these top causes helps in designing targeted traffic safety measures and educational campaigns to reduce accidents, enhance road safety, and ultimately save lives.

Time Series Analysis: The main goal of this milestone is to dive deeper into Time Series Analysis in order to better understand our data's trends over time.

- Hourly Distribution of Traffic Accidents

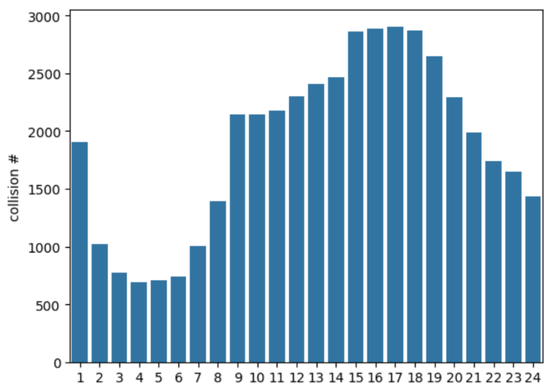
This bar chart illustrates the cumulative number of accidents that occurred during each hour of the day throughout 2021.

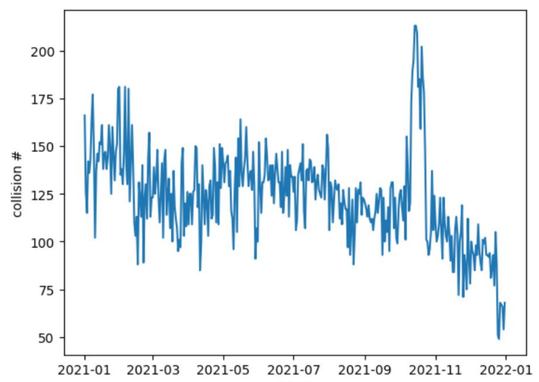
- Seasonal Variation in Traffic Accidents in 2021

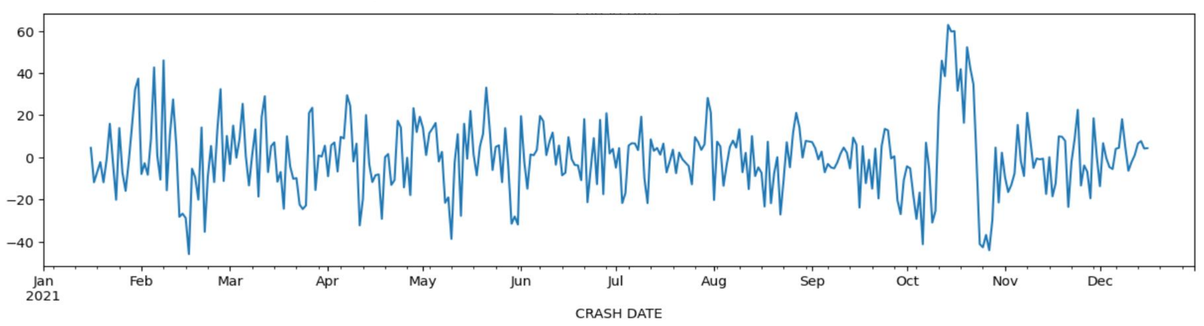
This time series graph shows the seasonal trends in traffic accidents over the year 2021.

- Residual Components of Accident Data

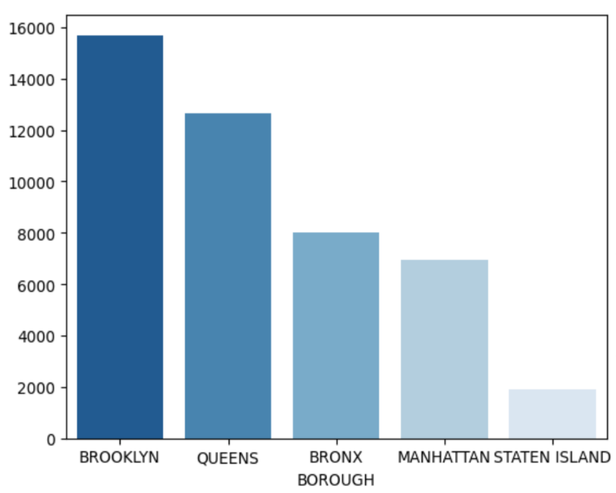
This plot displays the residual components after extracting the trend and seasonal factors from the accident data

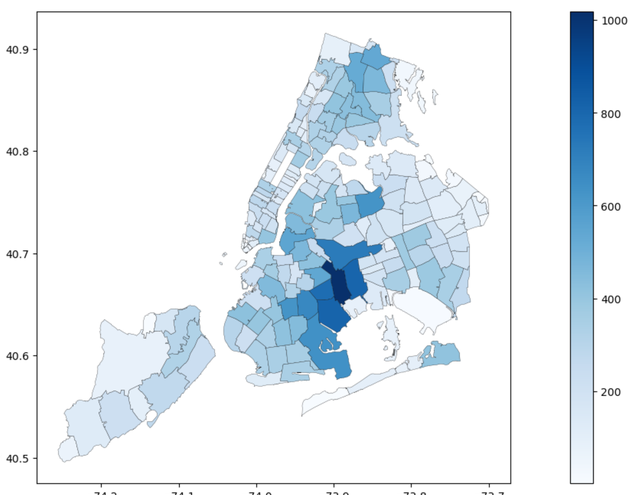






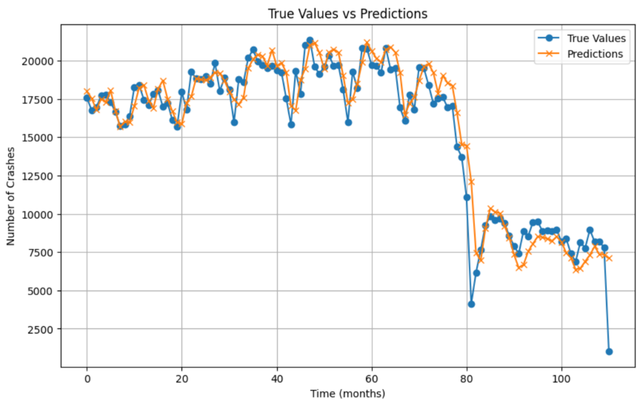
Geospatial Analysis: explore geospatial aspects of the and geospatial visualizations.





This heat map provides a visual way to see how often traffic accidents occur in different areas of New York City. Darker areas indicate areas with a higher number of accidents, while lighter areas indicate fewer accidents. With heat maps, decision makers can better pinpoint areas that need special attention and improvement, such as through additional traffic signals, signs, or other safety measures.

Neural Network Prediction (Based on LSTM): build a time series prediction model based on a Long Short-Term Memory (LSTM) neural network



- **Model Structure:** LSTM neural network for time series prediction.
- **Data Utilization:** The LSTM model utilizes the past 12 months of car crash data as features to forecast the number of car crashes for the upcoming month, leveraging historical patterns and trends.
- **Training and Testing:** First 100 months for training; rest for testing.
- **Performance Objective:** Minimize mean square error between predictions and actual data.

Conclusion

A comprehensive analysis of 2021 New York City crash data reveals major patterns in crash causes, types of vehicles involved, regional distribution, and temporal distribution. Driver inattention, failure to yield to pedestrians, failure to obey traffic controls, and unsafe travel speeds were the major contributing factors to crashes, underscoring the urgency of improving traffic safety awareness and regulating driving behavior. Sedans and sport utility vehicles had higher accident rates, reflecting the popularity of these models. The frequency of accidents in Brooklyn and Queens calls for optimization of traffic management strategies, and the high accident rate between 3 p.m. and 7 p.m., as well as the seasonal peak in November, should be the focus of future safety interventions.

Future research should delve into the unspecified causes of accidents in the data records and categorize these unspecified accidents in order to more accurately target interventions. Examining the relationship between infrastructure, traffic volumes, and accident types in different areas can help design safety strategies that are better suited to specific neighborhoods. At the same time, future trends in crash frequency can be determined based on neuronal prediction models to inform subsequent rule enforcement.By integrating these analyses, we will not only be able to reduce the number of crashes, but also provide data-driven insights for roadway safety planning in New York City.

References and Data Sources

[1] Shaaban, Khaled, and Mohamed Ibrahim. "Analysis and identification of contributing factors of traffic crashes in New York City." Transportation research procedia 55 (2021): 1696-1703.

NYC OpenData Motor Vehicle Collisions - Crashes dataset

https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Crashes/h9gi-nx95

Contribution Explanation

Kailai Wang, Yuchi Liu, Xuyan Zhao: Data Preparation and Visualization,

Kaiwen Tong: Background information

Yuchi Liu, Xuyan Zhao, Kailai Wang, Kaiwen Tong: Conclusion