NYC Traffic Accidents

Jared Goroski, Eric Lin, Brodie McCarthy, Sung Park

INTRODUCTION

Introduction

Problems

- Challenging Traffic Conditions: Heavy congestion and intricate intersections
- Traffic Accidents: Prone to accidents involving cars, pedestrians, and cyclists. Considered the worst city in the country to drive in on multiple metrics.

Questions

- Identify specific locations in New York City that exhibit a high frequency of collisions, indicating potential public safety concerns that may need infrastructural modifications?
- Can injury or death be predicted through machine learning?

DATA CLEANING

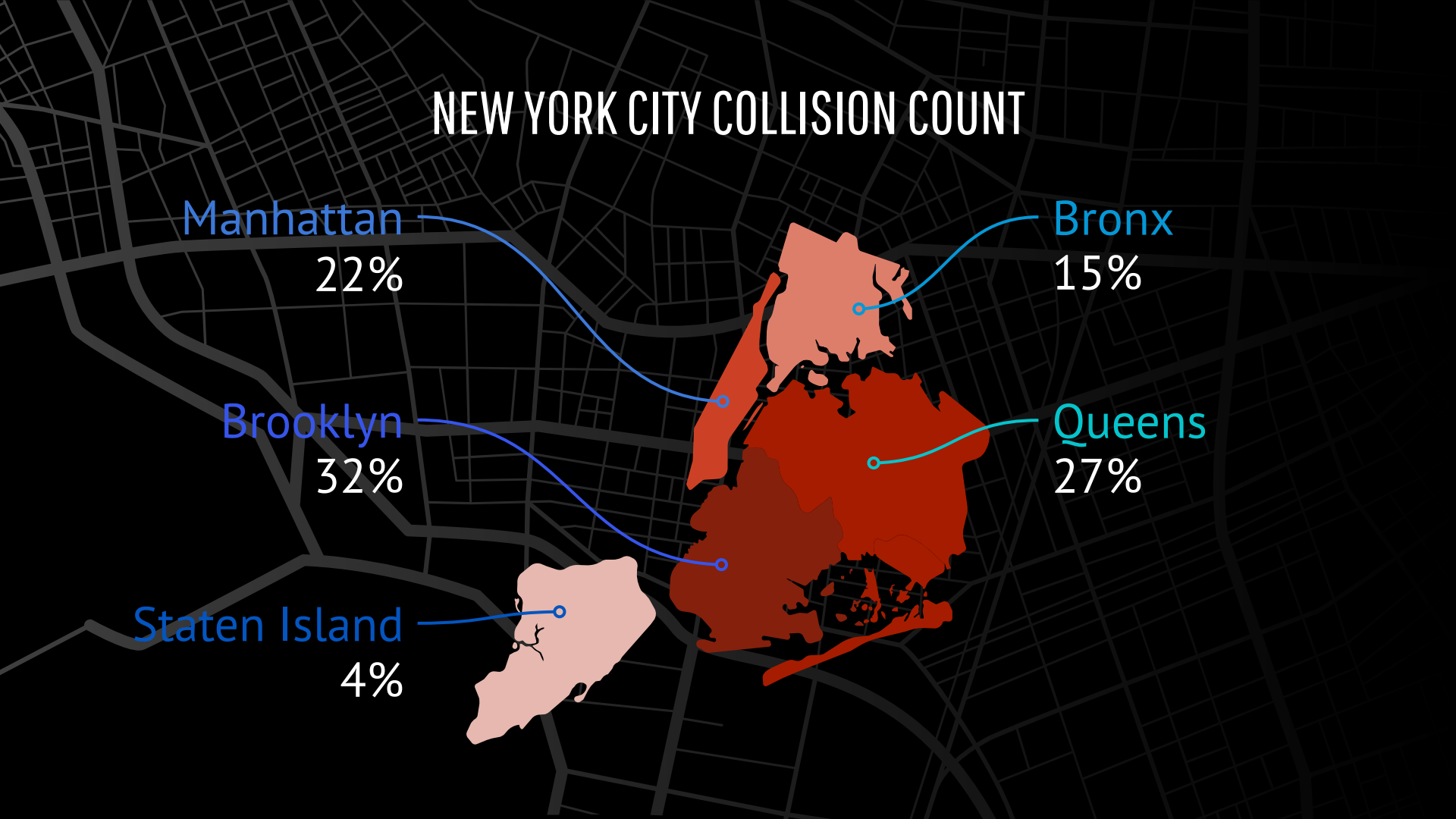
Data Cleaning

- NOT predicting killed, injured, or unspecified per person or per vehicle
- Predict if the accident had resulted in killed, injured, or unspecified per crash

Data	# Rows	# Columns
Crashes	2.08 M	29
Person	5.33 M	21
Vehicles	4.17 M	25
Master	600 K	45

ANALYSIS

NEW YORK CITY COLLISION COUNT



NEW YORK CITY COLLISION PER CAPITA

#1: Flushing
11362

#2: Long Island City
11101

#3: Bronx
10464

#4: Staten Island
10303

#5 Staten Island
10302

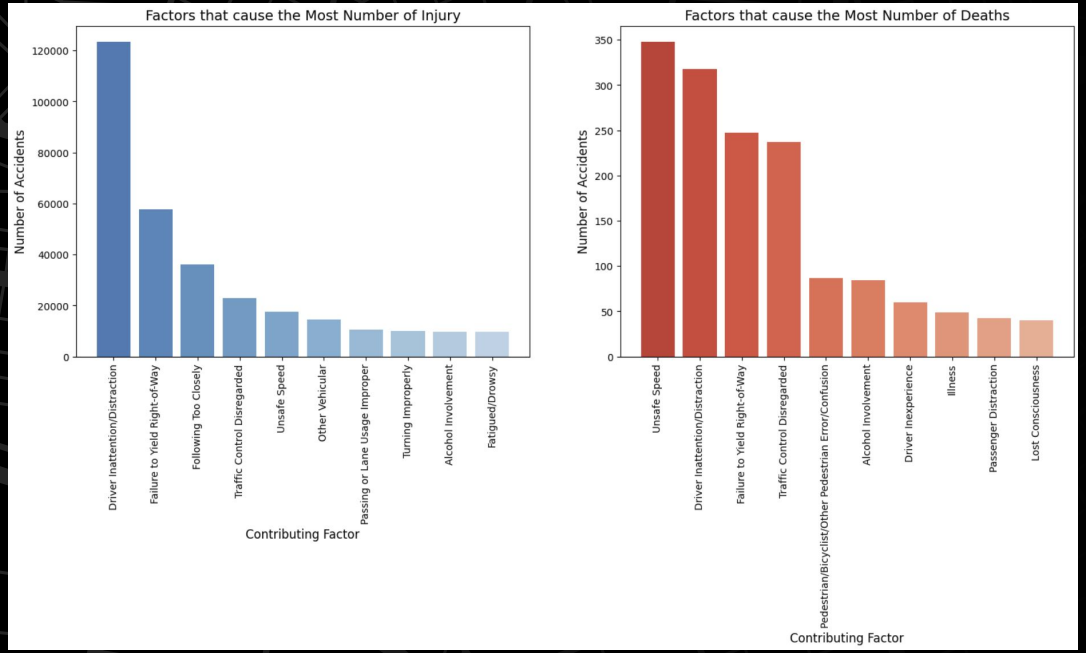
Accidents Factors Injury and Death Count

Injury Accident Factors

- Driver Distraction
- Failure to Yield Right-of-Way
- Following too Closely

Death Accident Factor

- Unsafe Speed
- Driver Distraction
- Failure to Yield Right-of-Way



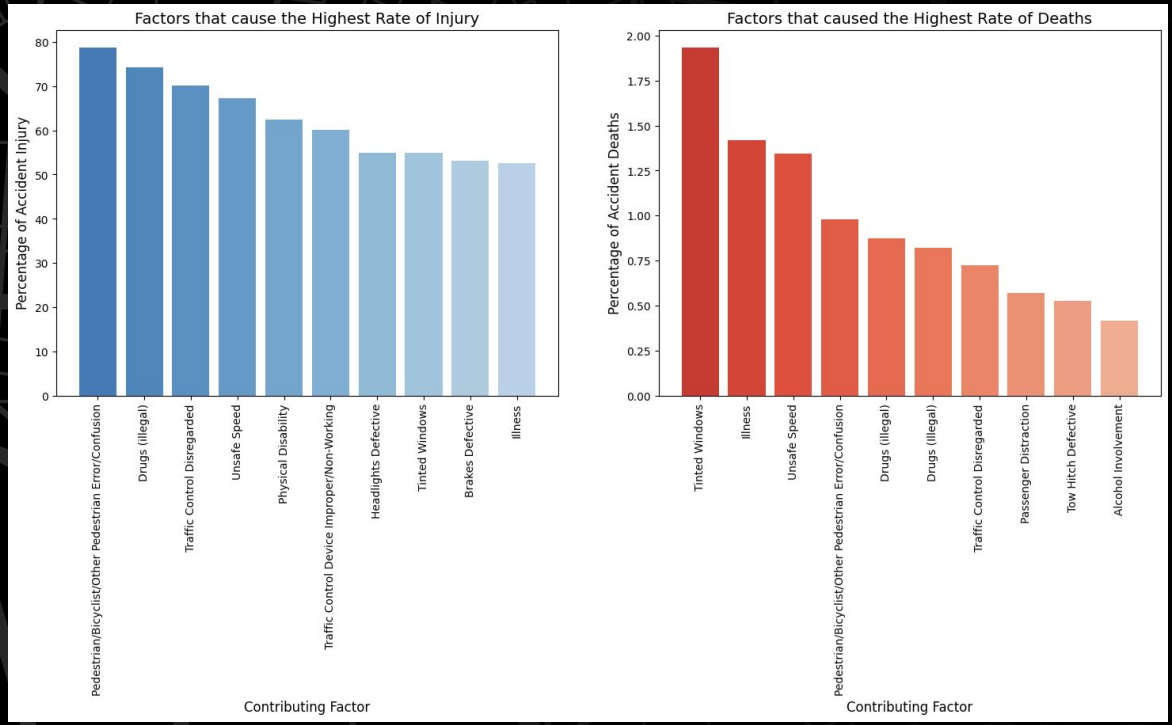
Accidents Factors Injury and Death Rate

Rate of Injury

- Pedestrians/Bicyclist
- Drugs
- Disregard Traffic Control

Rate of Death

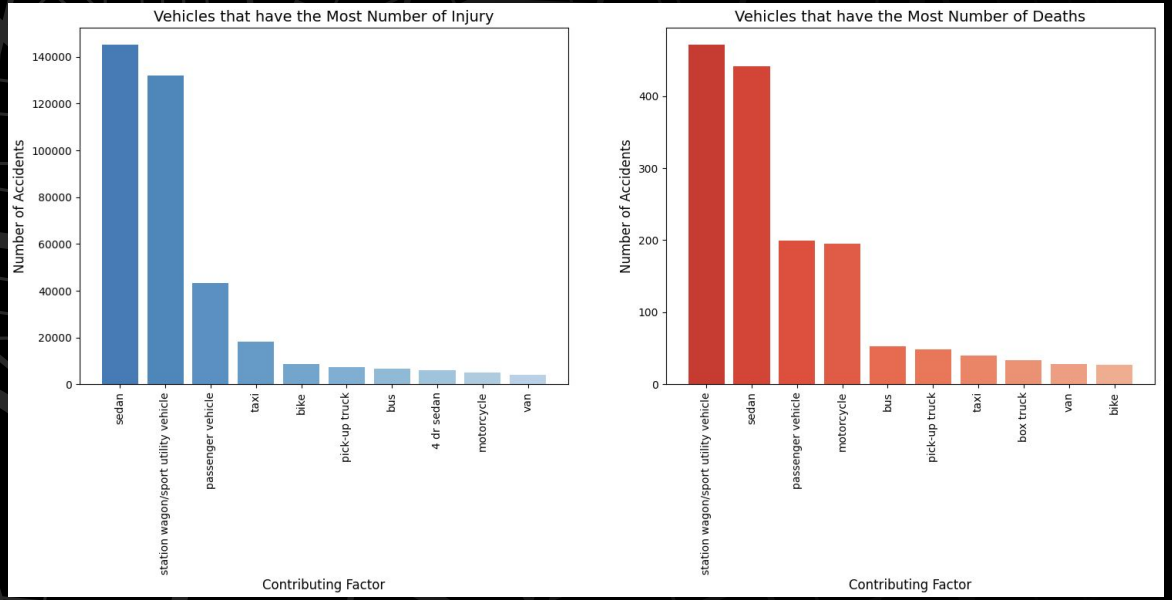
- Tinted Windows
- Illness
- Unsafe Speed



Vehicle Injury and Death Count

Most Accidents/Injuries/Deaths

- Sedans
- Station Wagons
- Passenger Vehicles



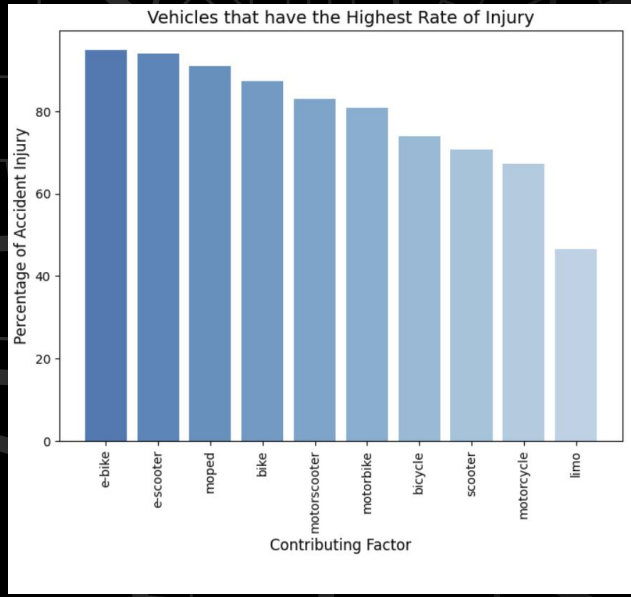
Vehicle Injury and Death Rate

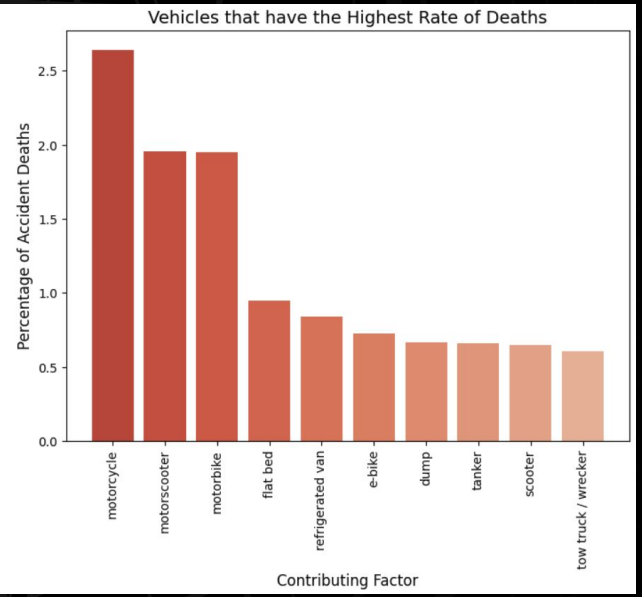
Rate of Injury

- Two Wheelers

Rate of Death

- Two Wheelers
- Large Vehicles





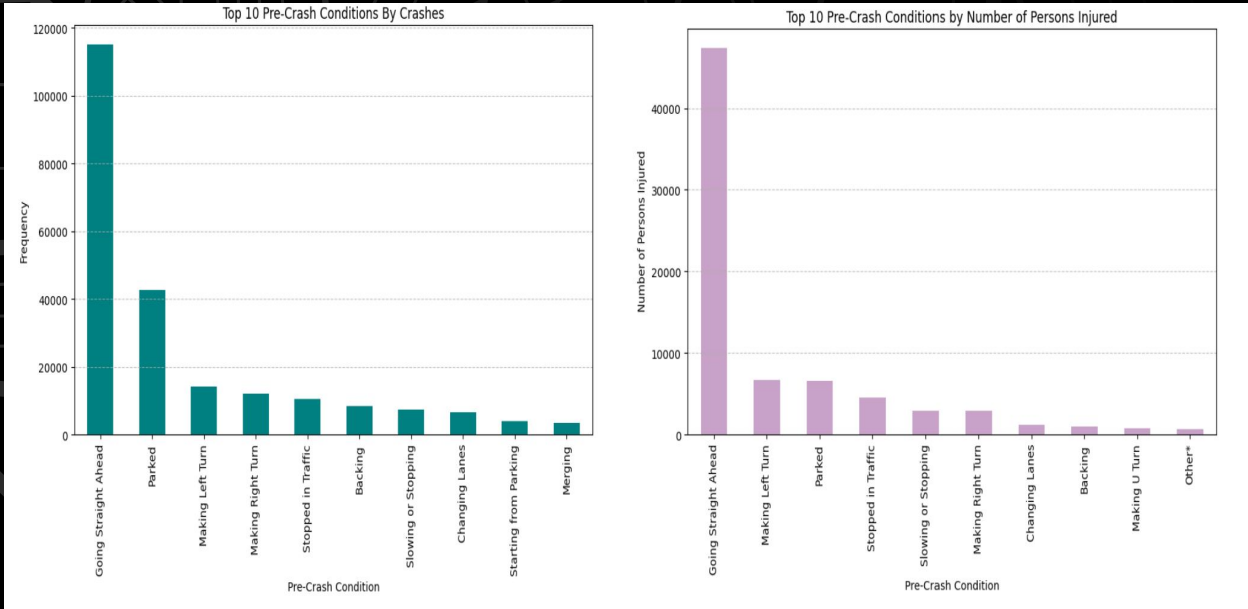
Collision Injuries By Pre-Crash Conditions

Most Crashes

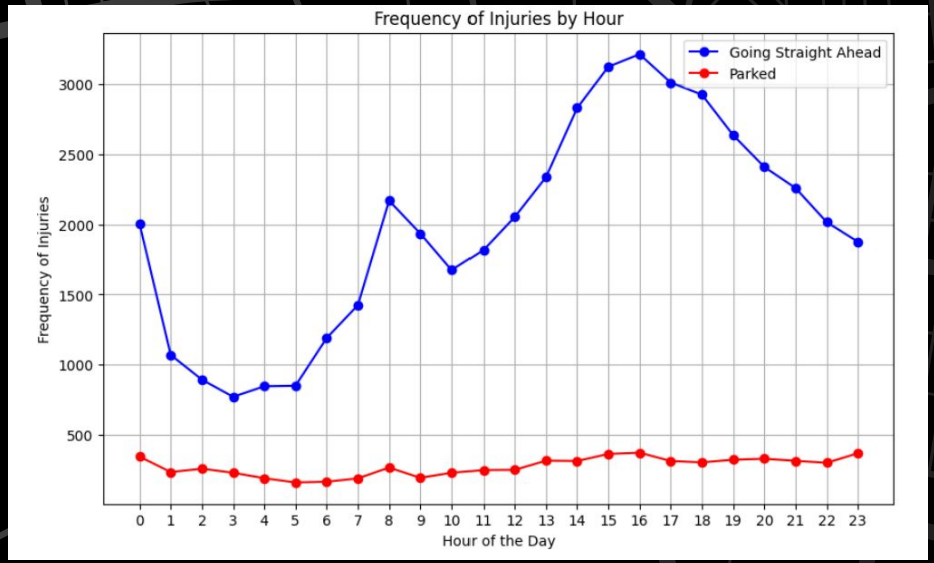
- Going Straight Ahead
- Parked
- Making Left Turn

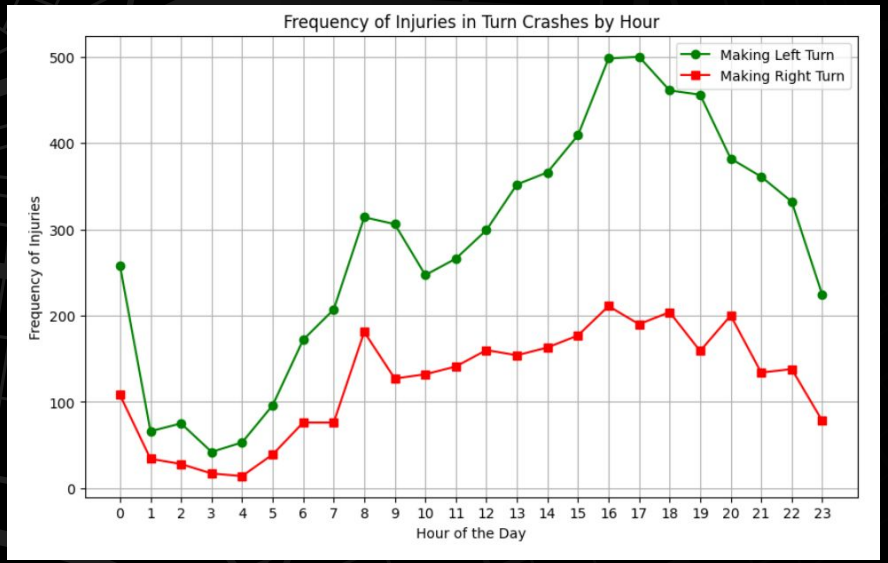
Most Injures

- Going Straight Ahead
- Making Left Turn
- Parked



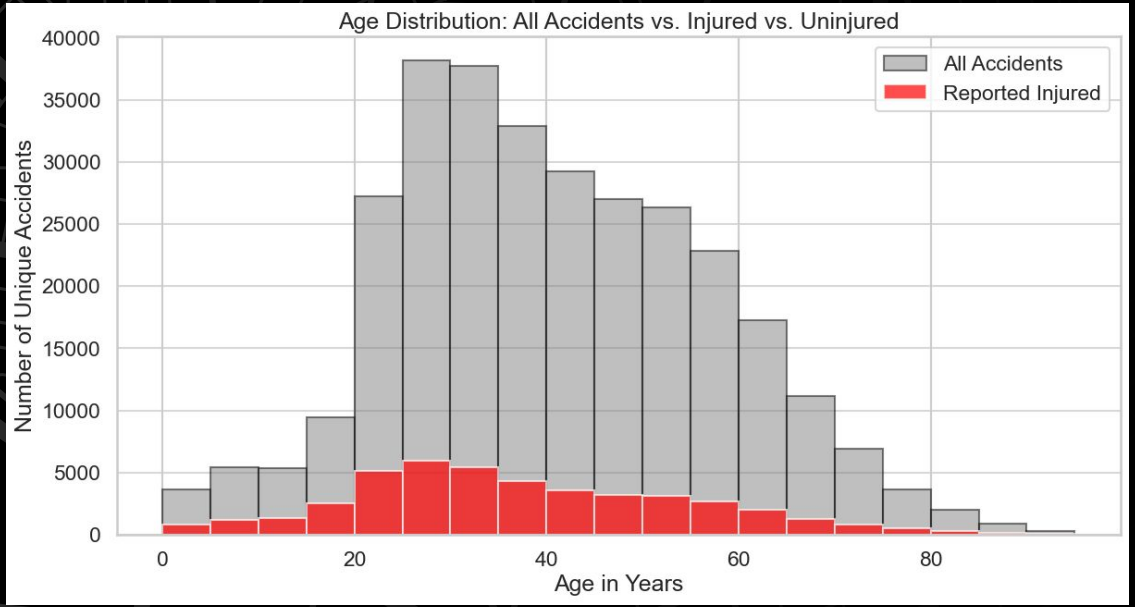
Pre-Crash Conditions By Hour





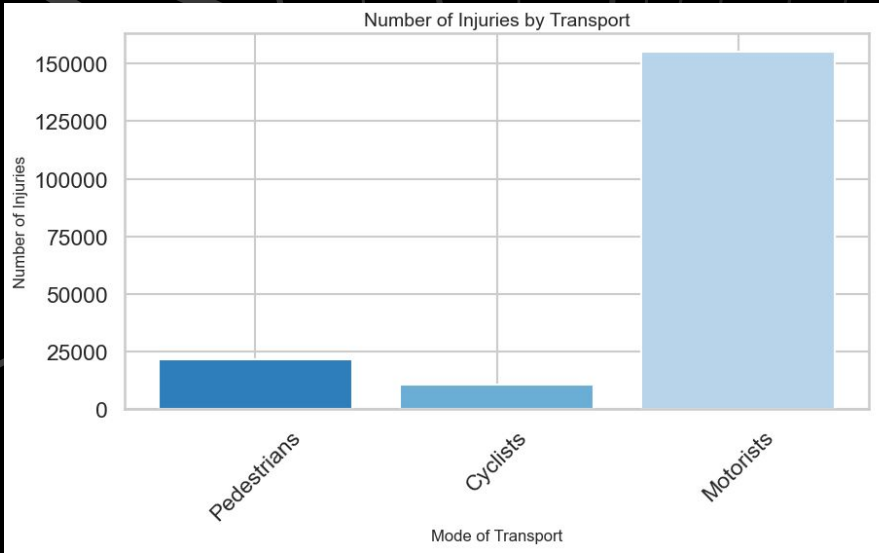
Accidents by Ages

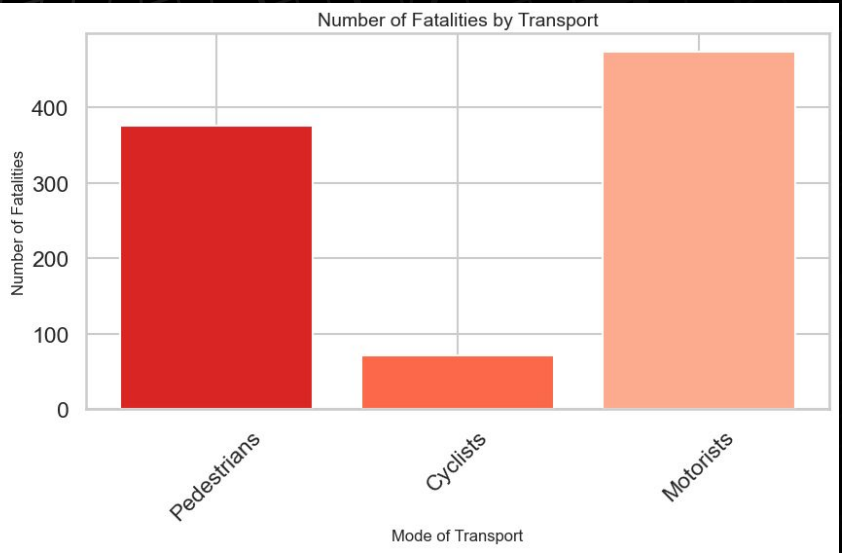
- Age distribution of ALL accidents resulting in injuries and unspecified
- Highest rates of accidents between 25-40
- Majority of accidents are minor involving no injuries



Injuries and Fatalities by Transport

- Motorists make up a majority of Injuries and fatalities
- Pedestrians had disproportionately Fatality struck compared to Cyclists and Motorists
- Risk hierarchy for injuries and fatalities: pedestrians > cyclists > motorists

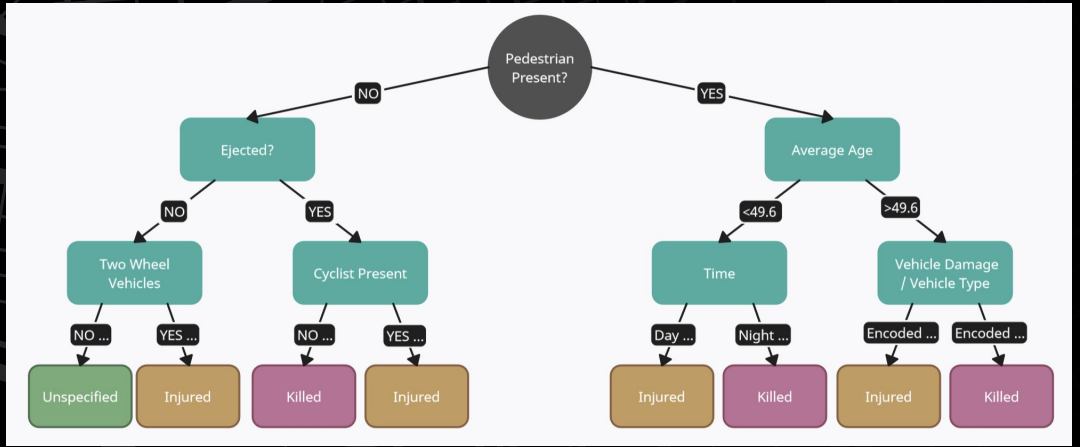




MACHINE LEARNING

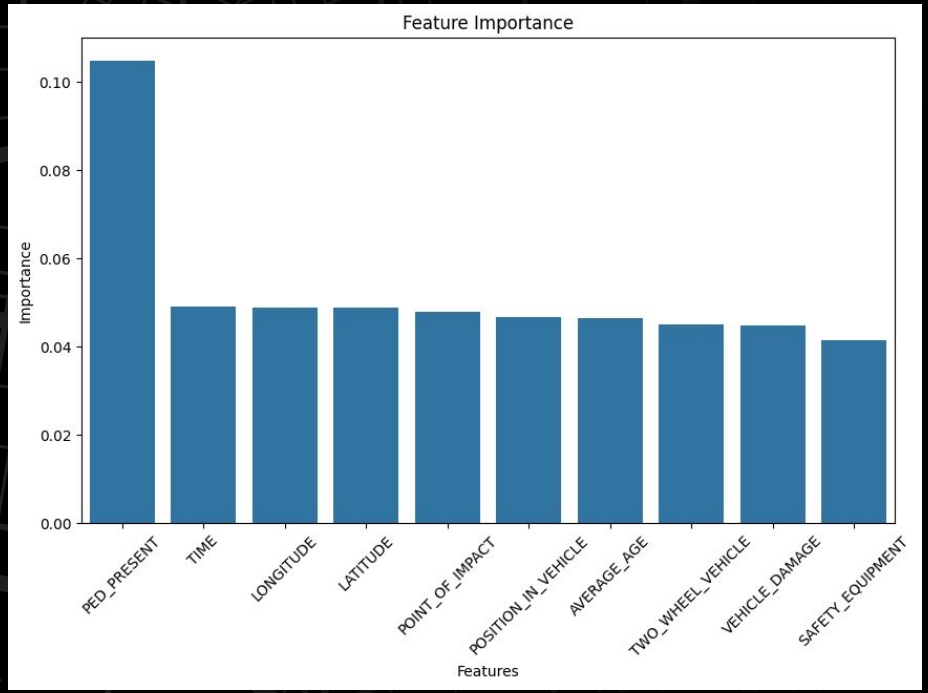
Decision Tree

- Accurate Decision Trees were too large to display easily
- Accuracy of 75.5%
  - Very low precision for Killed class and low recall for Injured
- Interesting Points:
  - Later time played a factor



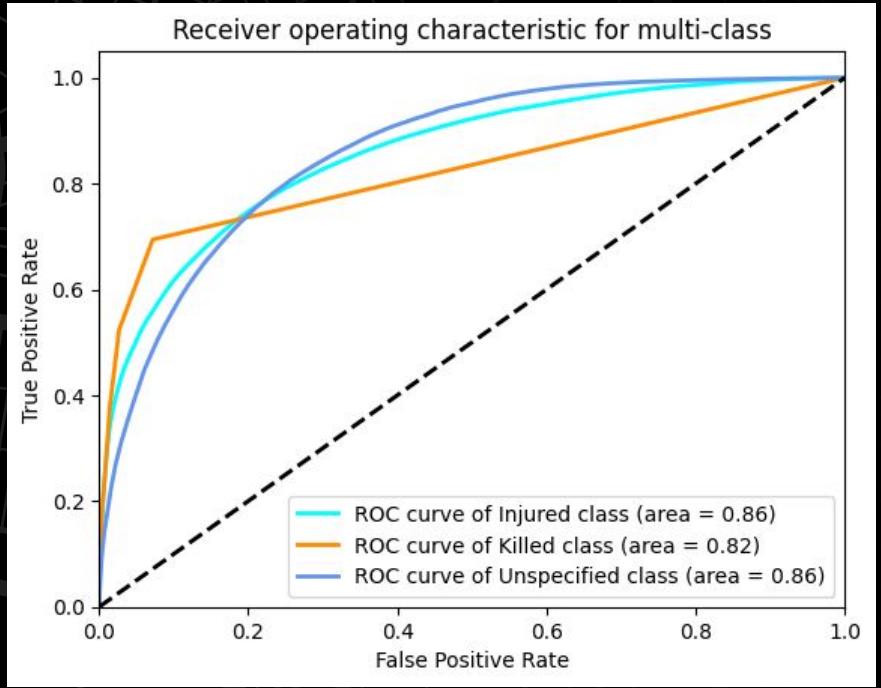
Random Forest Classifier

- Average Accuracy of 81.3% over 5 folds
  - High of 83%
  - Low of 73%
- Most important features aligned with Decision Tree
  - Found location to be important

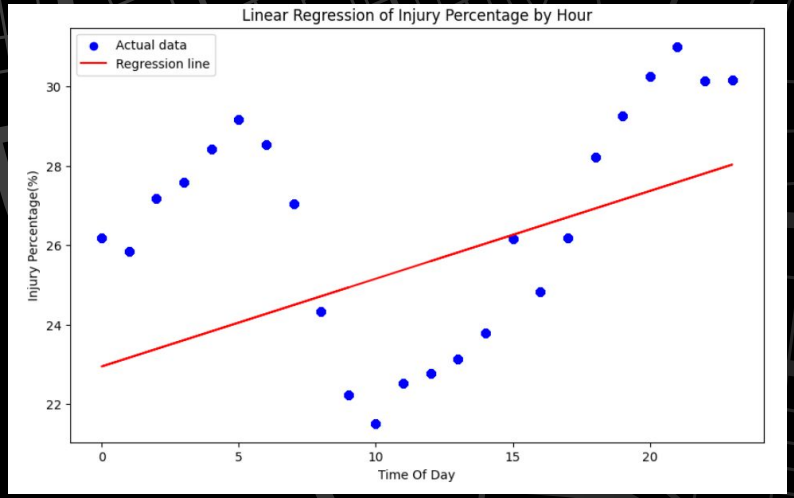


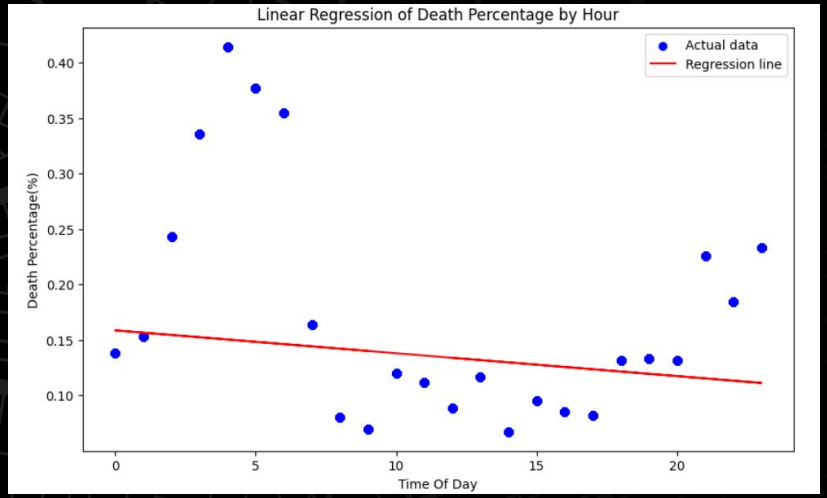
Random Forest Classifier

- Better at classifying Unspecified and Injured
- Followed Decision Tree with issues in sensitivity towards Killed class



Linear Regression





Used Variables For Injury Prediction	Correlation Coefficient	R-Squared	RMSE
Crash time by hour	0.453141623	0.2053373305	2.504416118
Average age	0.118258	0.01398487504	2.789695954
Number of vehicles	0.047524	0.00225855152	2.806235351
Number of people	0.035149	0.001235459845	2.80767375
Pre-Crash Conditions	0.07261363506	0.005272739997	2.80199331

	CRASH HOUR	# VEHICLES	AVG AGE	# PEOPLE
Injury Ratio	0.450	- 0.048	- 0.120	0.035
Death Ratio	- 0.160	- 0.071	- 0.076	- 0.025

Neural Network

Beginning Accuracy

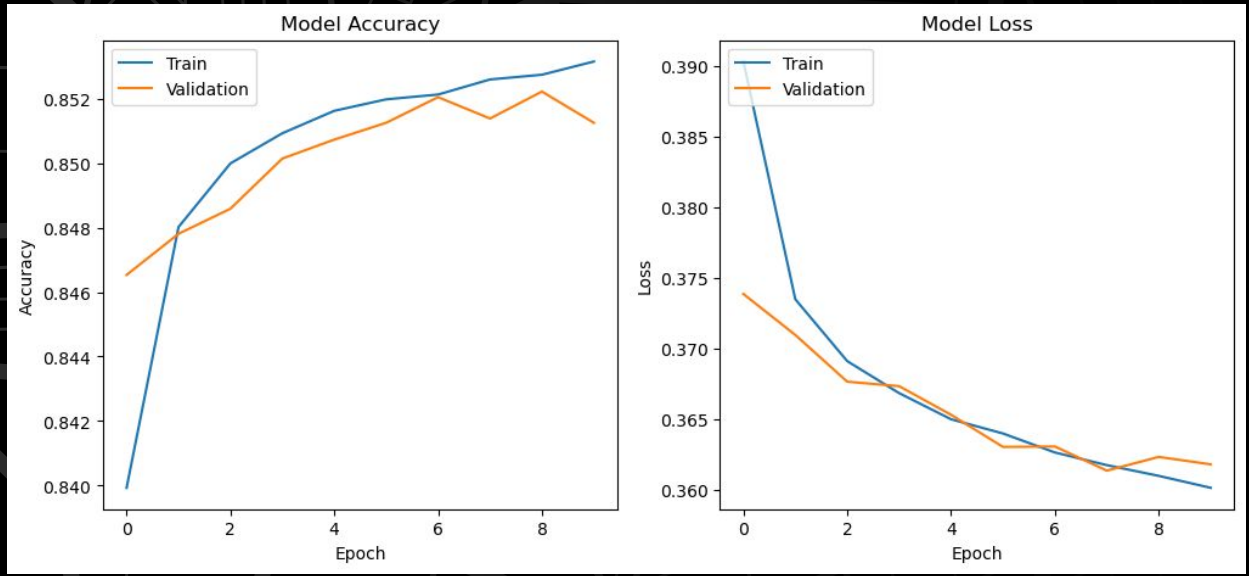
- 82.82%

Ending Accuracy

- 85.16%

Network

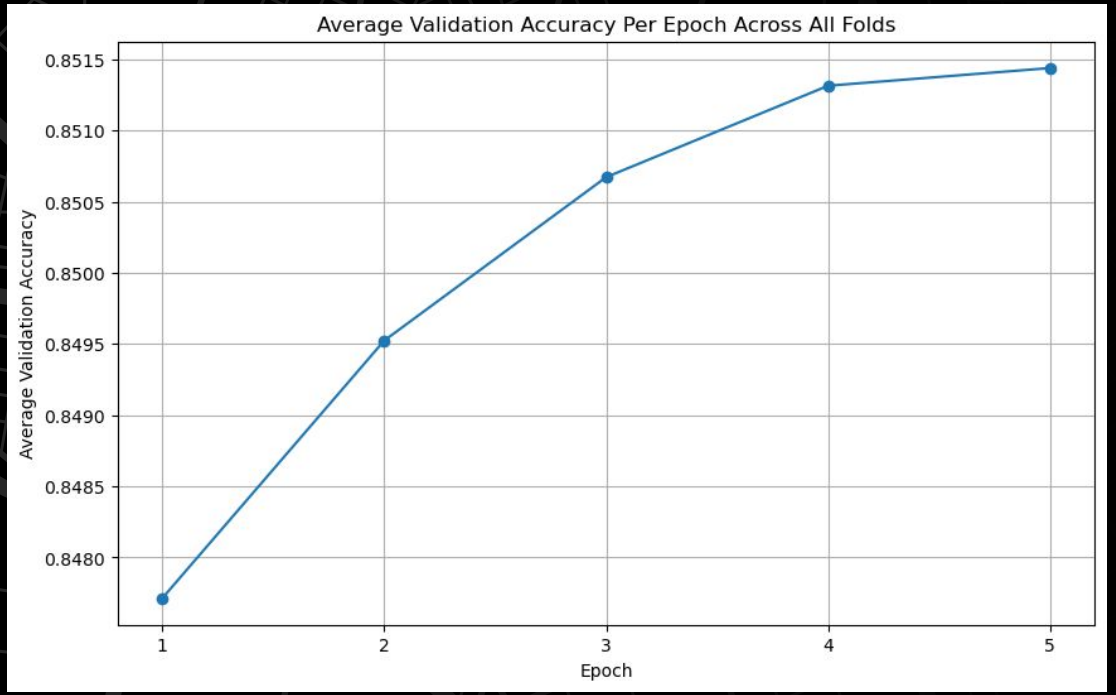
- 3 internal layers
- 10 epochs
- 80-20 train split



Neural Network

- Using k-fold with 5 splits
- Utilizes shuffle to ensure good representation

- Lowest score: 85.08%
- Highest score: 85.19%
- Average score: 85.14%



WHAT'S NEXT?

Future Work

Weather Data

- Precipitation
- Temperature
- Visibility

OpenWeatherMap

- 40 years of daily weather data for each New York City location/borough

Pothole Data

- Road Conditions
- Pothole Location vs Accident Location

NYC OpenData

- 311 Service Requests (2010 - Present)
- Potholes

Future Work

Vehicle Distribution

- Number of vehicles currently in motion
- Tunnels, bridges

Race Data

- Ethical Challenges
- Traffic accidents rates among different races

CONCLUSION

Conclusion

Through the data we have seen that pedestrians and bikers are the groups that are most likely to be injured or fatally struck compared to the parties that are in vehicles. There are several major factors that contribute to the high rate of accidents that are happening across the boroughs, some of the leading factors for both fatal and injury accidents are distracted drivers, failing to yield, and speeding. It seems that new policies to protect pedestrians and bikers by creating protected bike lanes or high visibility crosswalks. Majority of accidents are in Brooklyn and Queens which had the highest per capita.

To predict the injury level of parties in an accident, the neural network slightly outdid the random forest model, with 85% vs. 82% accuracy, in predicting traffic accident outcomes. The neural network deep learning worked well at identifying complex patterns, while the random forest ensemble method offered robustness and visual layout . Both models could be improved by adjusting their complexity and incorporating more data. Linear regression lagged behind, showing that more sophisticated models are needed for this task.

THANK YOU
Q & A