Predicting Accidents in New York City

Analog: Dev Ramesh, Matteo Lunghi, Richard Hill, Tom Liu

BROWN

Introduction & Goal

Traffic accidents seriously injure approximately 3,000 people and kill approximately 200 people in New York City each year.

We aim to predict the number of accidents in each borough of NYC during any period of time given weather conditions, traffic speeds, location, and time of day.

We hope our results can be used by government programs, like Vision Zero NYC, to improve traffic safety.

Dataset Information

Raw data is collected from years 2018 and 2019 from three sources:

1. Traffic Accidents (NYC OpenData) - ~450 thousand points

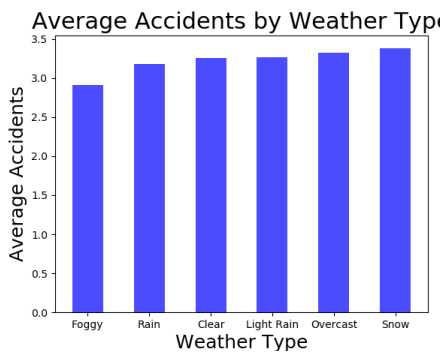
2. Traffic Speed (NYC OpenData) - ~26 million points

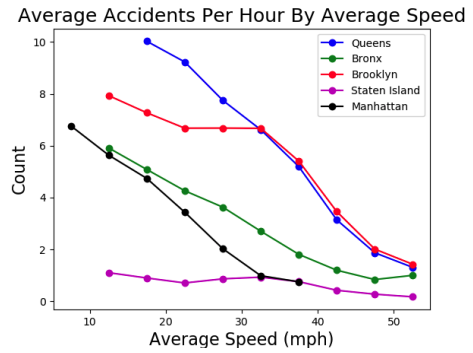
3. Historical Weather (DarkSky API) - ~88 thousand points

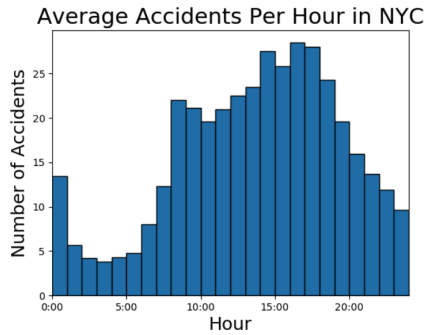
The table is cleaned and joined on location (borough) and time (date and hour) such that there is an entry for each hour for each borough in the two year span. Results in ~88 thousand points.

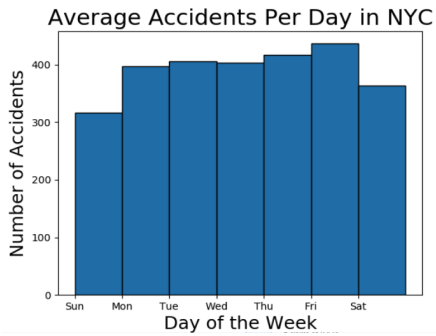
Acknowledgements

We would like to thank Professor Pavlick and all Data Science TAs for their support and guidance through this project.









Sample Entry

precipintensity (inches)	*windspeed (mph)*	*visibility (miles)*	*avgspeed (mph)*	*weather*	*borough*	*hour*	*day*	**accidents**	**prediction**
0.042	5.690	3.901	35.137	Rain	Bronx	2300	weekday	**2**	**1.928**

Methodology

1. City specific

2. 80/20 train-test random split

3. Feature-scaled continuous features

4. One-hot-encoded boolean features

Multiple Linear Regression Model

1. Features: **precipitation intensity, wind speed, visibility, average traffic speed**, *rain, cloudy, snow, Bronx, Brooklyn, Manhattan, Queens, hour 0, ..., hour 22, weekday*

2. Performance Metric: Comparing MSE and R-Squared to baseline model (predicts using average number of accidents across NYC per hour over two years)

	R-Squared	Train MSE	Test MSE
Baseline	0.000	10.5515	10.7523
Model 1	0.129	9.1931	9.3560
Model 2	0.411	6.2167	6.3054
Model 3	0.553	4.720	4.817

Results

1. Model substantially outperforms the baseline model. Model's MSE 50% lower than baseline model.

2. Weather conditions had a neligable impact on the model's MSE and R-Squared (near zero for both). Only visibility (p-value 0.04) and cloudiness (p-value of 0.00) were significant. The condition of snow (p-value 0.616) was insignificant.

3. Splitting by borough greatly improves performance. One-hot encoding borough features increased R-squared value from 0.286 to 0.553, and dropped test and train MSE from ~7.5 to 4.817.

Limits & Significance

1. The low impact of weather data challenges our biases and gives us a direction to explore confounding variables. We formed hypotheses relating weather and average speed to traffic volume, however we were unable to explore them since our dataset lacks this information

2. Since we did not always get a specific location for accidents, we could only compare accidents in a borough to summary statistics of the speed.

3. The model's R-Squared value of 0.553 implies that our data can predict the number of accidents with moderate success. This suggests that our model can roughly predict the number of accidents in a region per hour, which can help local governments coordinate resources needed to be prepared for accidents.