NYC Car Accident Predictor

Davida Rosenstrauch
February 18, 2021

Agenda

- Why it matters
- The Process
- The Data
- The Pandemic Factor
- Exploratory Data Analysis
- The Results
- Recommendations, Next Steps

Why It Matters

- Car accidents are one of the leading causes of death and injury in the United States.
- Predictions can help:
  - Police departments, EMT services, and hospitals appropriately allocate resources
  - Individuals assess risk when going out on the road

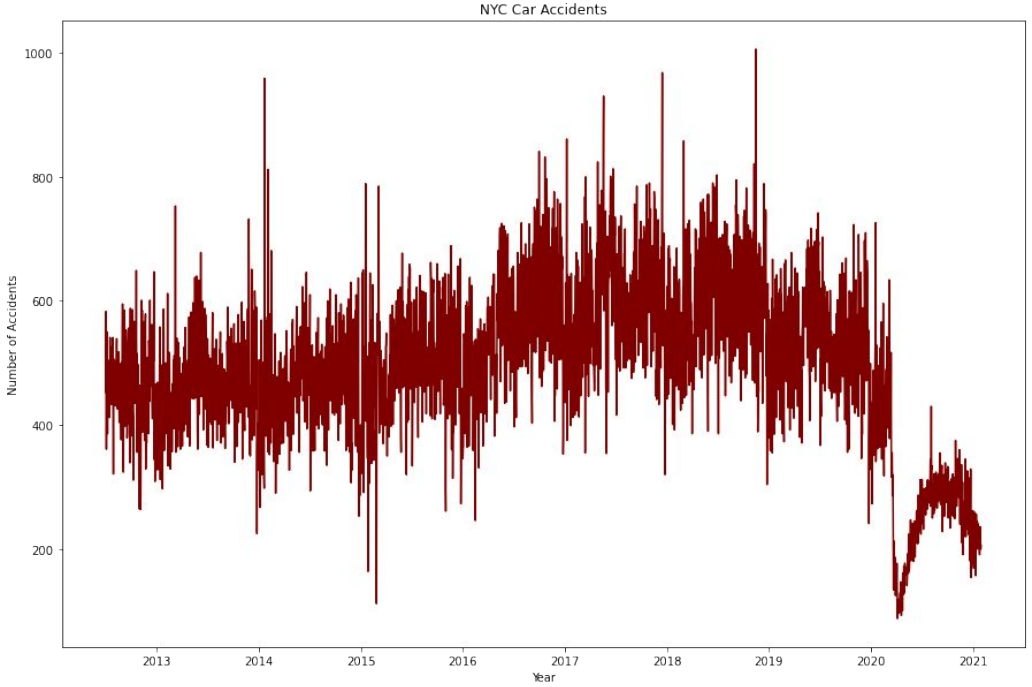
The Process

1. Data collection, cleaning, and analysis
2. Time series models on each borough, in order of accident frequency
    - ARIMA-style time series models:
        - ARIMA (baseline)
        - ARIMAX
        - SARIMA
        - SARIMAX
    - Facebook Prophet model
3. Apply predictions of best-performing models
4. Deploy Streamlit dashboard for public use

The Data

- Primary data: NYC Open Data Motor Vehicle Collisions dataset
  - 1.75 million crashes in New York City between July 1, 2012 and January 29, 2021
- Supplemental data: Nominatim reverse geocoding API to identify crash boroughs and zip codes based on GPS coordinates identified in primary dataset

The Pandemic Factor

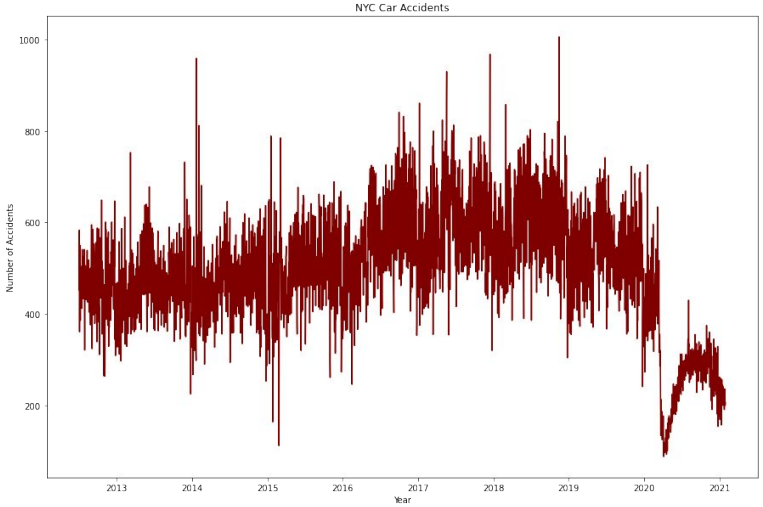


The Pandemic Factor

- COVID projections
- Transportation projections
- Workforce projections

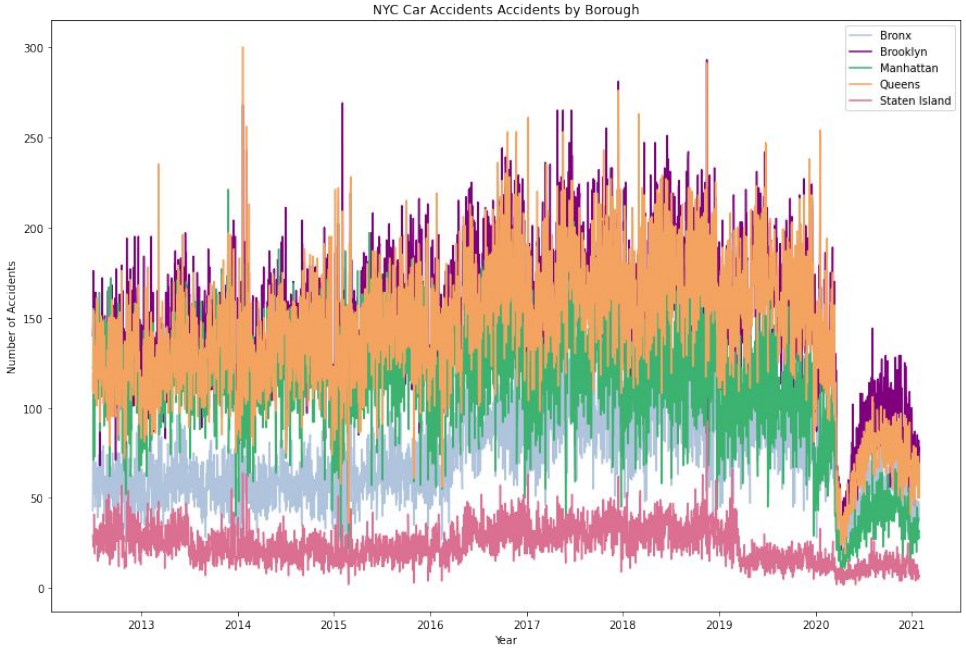
- Sources:
  - https://www.cdc.gov/coronavirus/2019-ncov/downloads/covid-data/Consolidated-Forecasts-Incident-Cumulative-Deaths-2021-02-01.pdf
  - https://covid19.healthdata.org/united-states-of-america?view=total-deaths&tab=trend
  - https://analytics-tools.shinyapps.io/covid19simulator04/
  - https://covid19-projections.com/path-to-herd-immunity/
  - https://www.theatlantic.com/ideas/archive/2020/12/the-2021-post-pandemic-prediction-palooza/617332/
  - https://www.govtech.com/analytics/Has-COVID-19-Forever-Changed-Rush-Hour-Traffic-Patterns.html

The Pandemic Factor

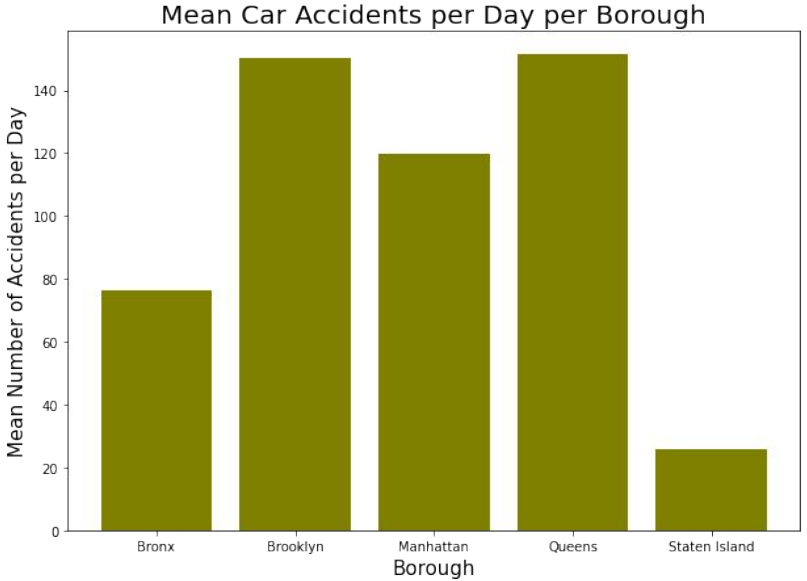


- Model on data until March 13, 2020
- Starting July 1, 2021, predict 25% of predictions that would have come from pre-COVID data
- Connect projections linearly between end of current data and July

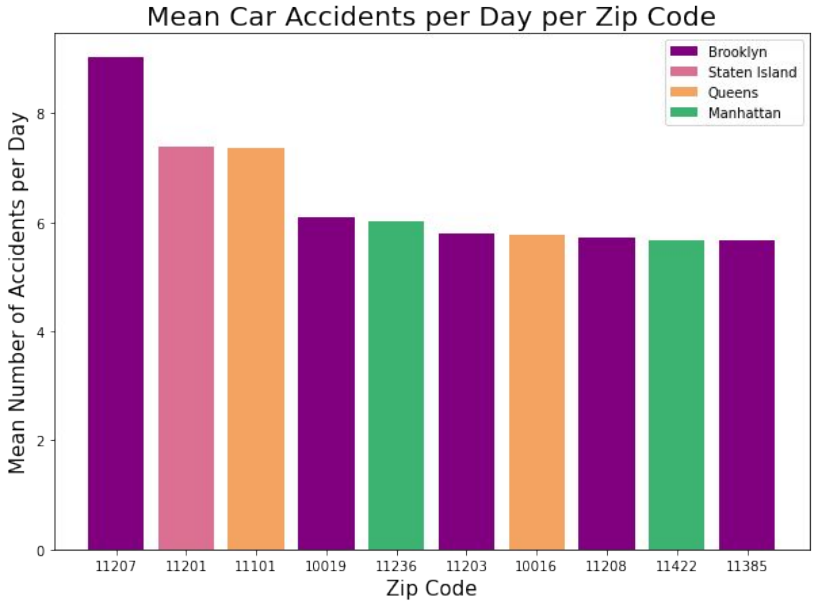
Exploratory Data Analysis:
Where Do Accidents Occur?



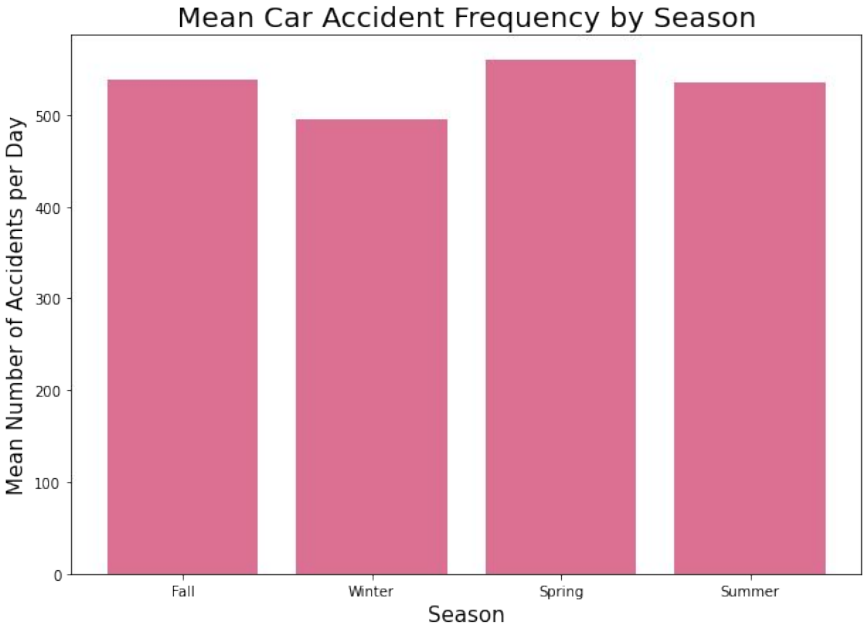
Exploratory Data Analysis:
Where Do Accidents Occur?



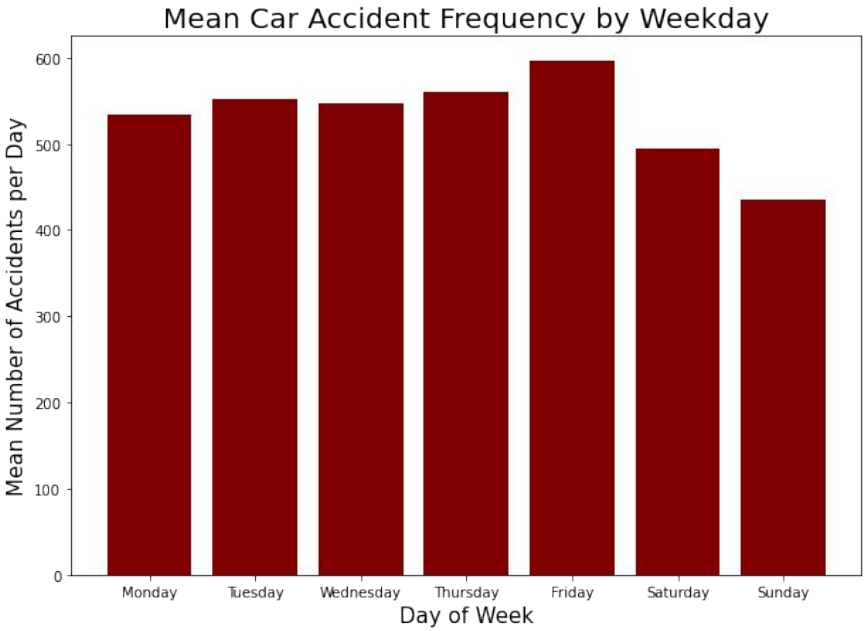
Exploratory Data Analysis:
Where Do Accidents Occur?



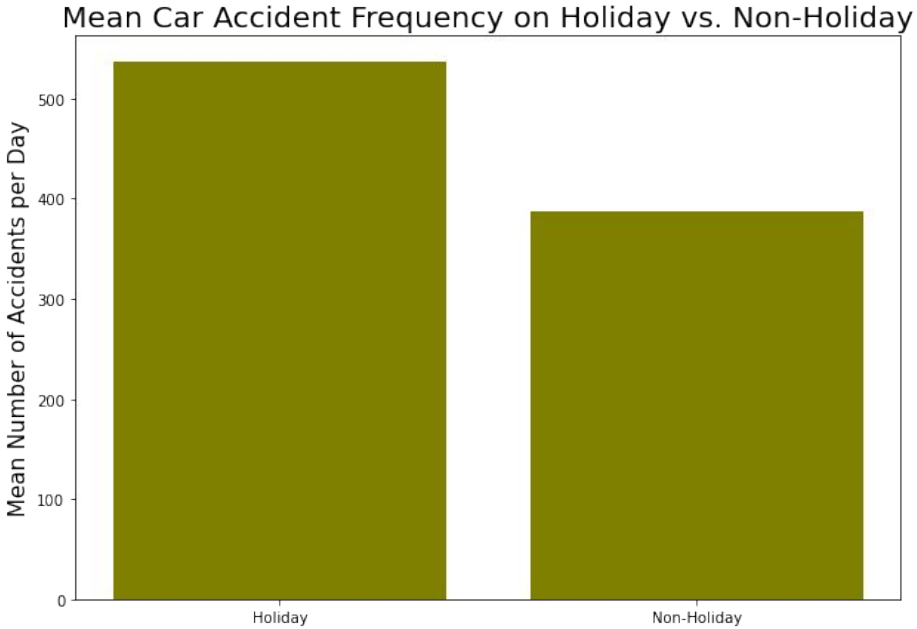
Exploratory Data Analysis:
When Do Accidents Occur?



Exploratory Data Analysis:
When Do Accidents Occur?

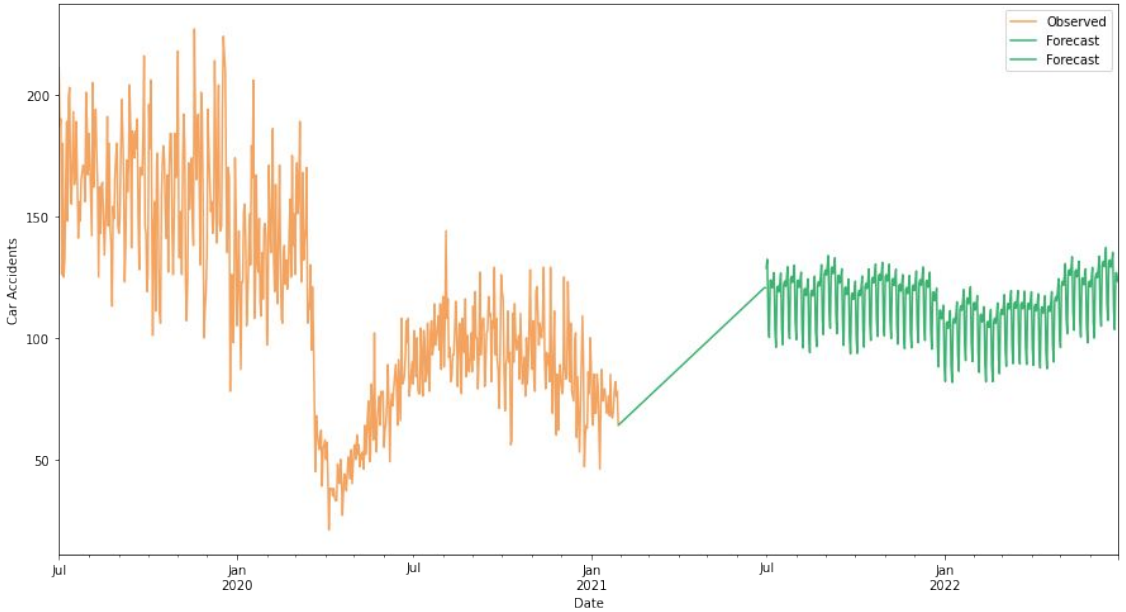


Exploratory Data Analysis:
When Do Accidents Occur?



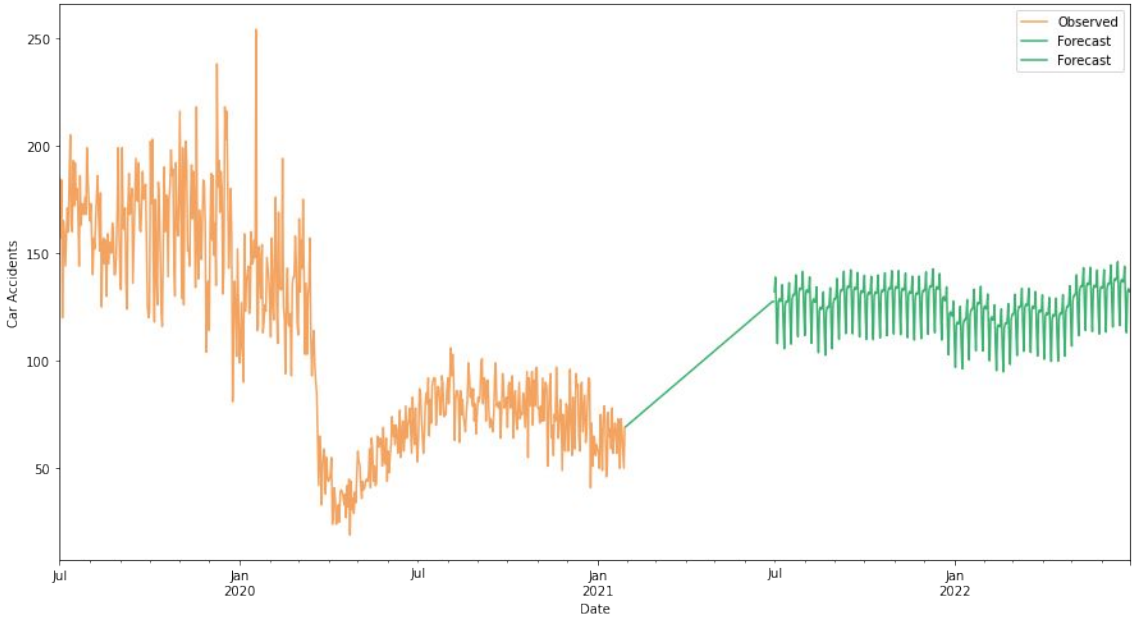
Results: Brooklyn

- Facebook Prophet model
- RMSE: 24.44
- Predictions wrong by an average of 9% of total accident range



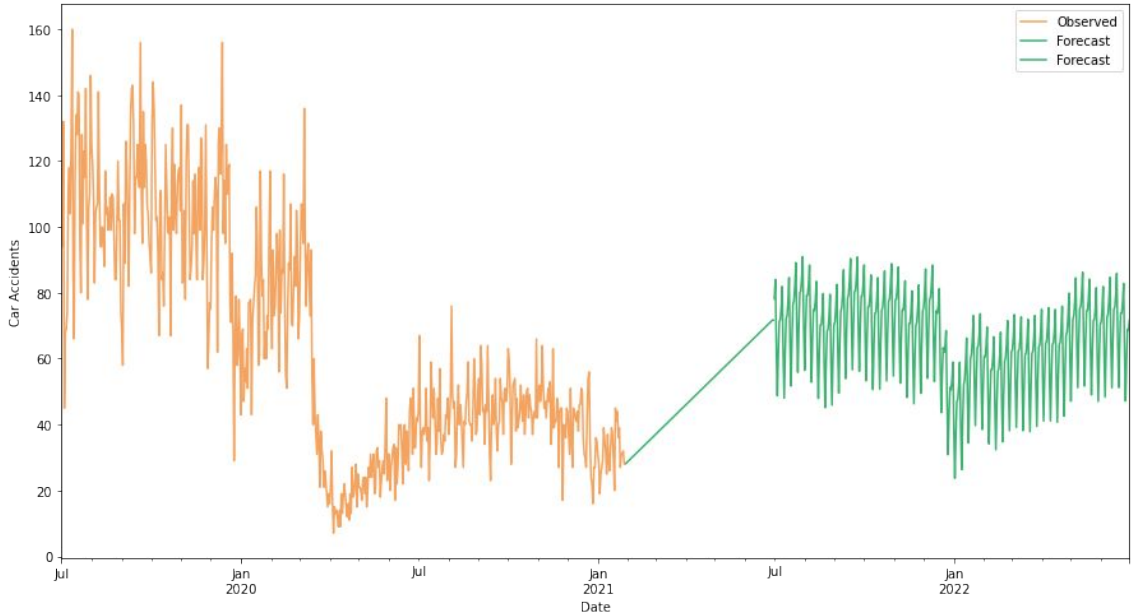
Results: Queens

- Facebook Prophet model
- RMSE: 26.62
- Predictions wrong by an average of 10% of total accident range

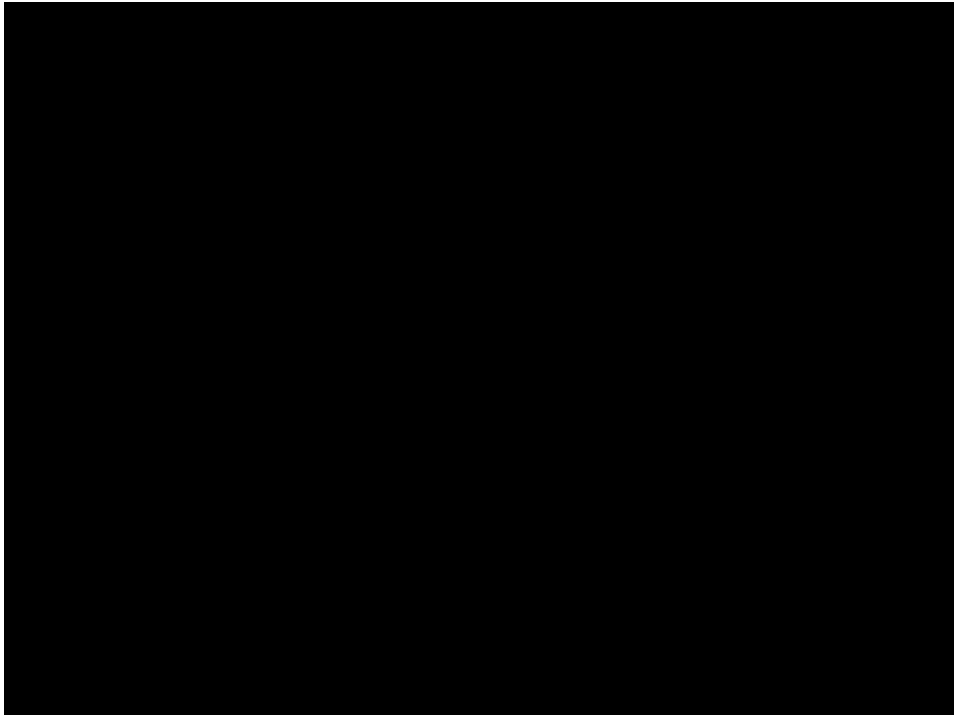


Results: Manhattan

- Facebook Prophet model
- RMSE: 19.05
- Predictions wrong by an average of 9% of total accident range



Results: Streamlit Dashboard



Recommendations

- For institutions:
  - Increase crash-related resources in time periods with most accidents, for example during the spring.
  - Save resources in time period with fewer accidents, for example on holidays and weekends.
  - Increase crash-related resources in zip codes with most accidents.
- For individuals:
  - If your timing is flexible, avoid hours of peak accident frequency.

What's Next?

- Generate predictions for remaining boroughs
- Try other modeling types like neural networks
- Predict number of accidents by zip code
- Predict number of injuries per borough and zip code
- Incorporate weather APIs and web-scraping
- Adjust model as pandemic-related updates occur

Thank you!

https://github.com/Davida1014/NYC-Car-Accident-Predictor/

http://192.168.0.11:8501/

https://www.linkedin.com/in/davida-rosenstrauch-0b1b88a6/