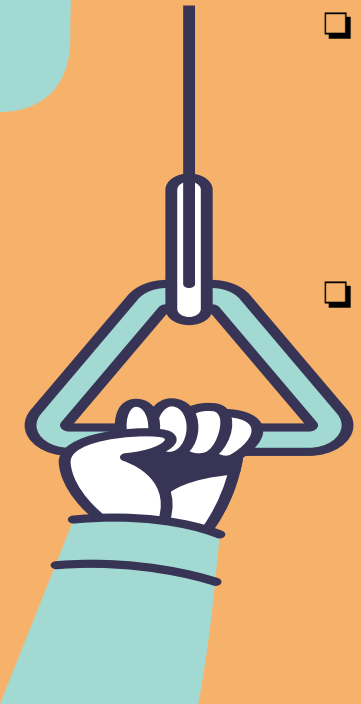
Going Places with Data:
A Study of the MBTA

Tim Demling, William Hanvey, Dan McCusker-Alvarez, Emma Sommers, Nick Perrotta, and Daniel Xu

Motivation

- Boston has the oldest public transit system in the nation
- How well the MBTA works has the potential to affect the entire city from citizens who use the T as their primary method of transit to businesses who need their employees to arrive on time
  - 34% of Boston commuters take the MBTA
- Northeastern was historically a commuter school and many students still use the orange or green line to get to campus



Research Question:

What correlations exist between MBTA service and outside factors?

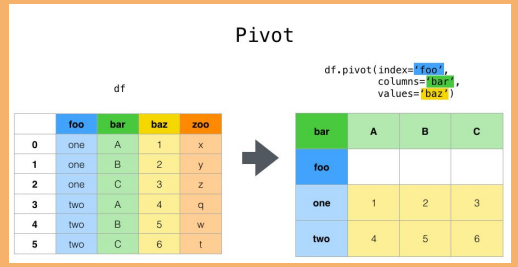
- Is there a relationship between potential delays and weather factors, day of the week, month of the year, and stops?
- Boston's demographics vary widely from neighborhood to neighborhood
- Some of these neighborhoods may be more reliant on public transportation
- Is there a relationship between better service and the demographic makeup of that area?

Data Sources

- ❏ MBTA 2020 Orange Line data
- ❏ Other MBTA informational data
- ❏ Weather Underground
- ❏ Boston.gov demographic data (2015-2019)
- ❏ Boston.gov geojson files

Preparing Data: Delays

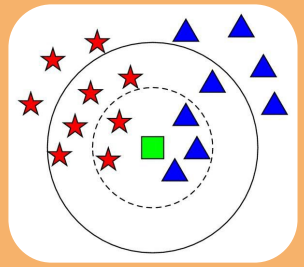
- EDA and Cleaning: fixing issues with NaN, direction ID
- Merged with other data to give each row next + previous stop
- Calculated travel time between stations for each train
- Pivot tables for the time between stations based on date, station, and direction using **median** as the aggfunc
  - Joined weather and precipitation data based on date



- Flagged data with mean > (5% * median) as ‘slow days’
  - Heavy positive skew -> variability in travel times

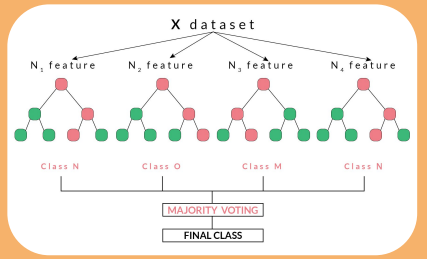
Two ML Algorithms Used:

K-NN



KNN for estimating slow day flag - 99.5% accuracy on test data

Random Forest



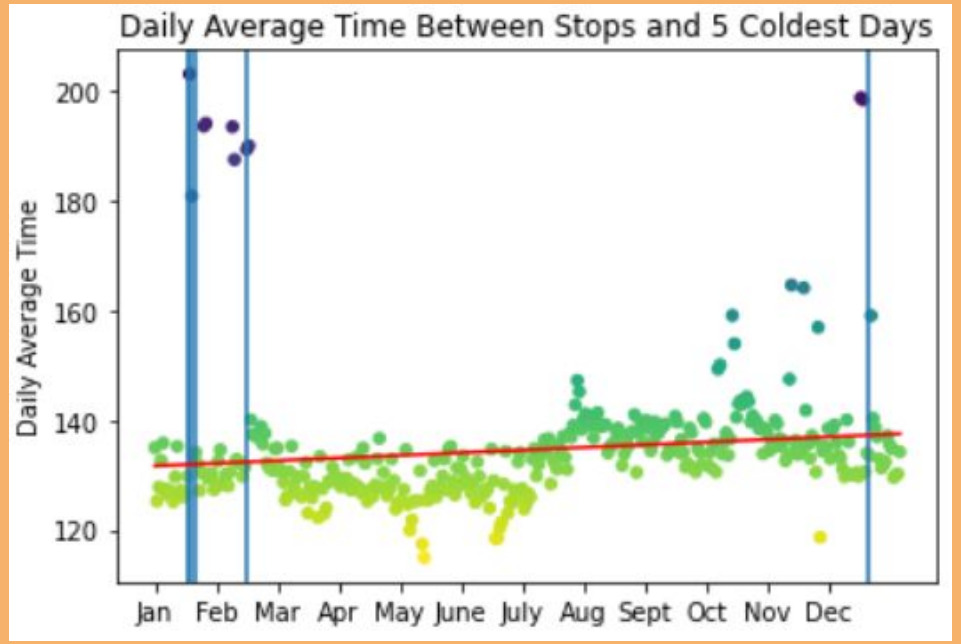
RF for estimating time between stops- 6.9 mean absolute error on test data

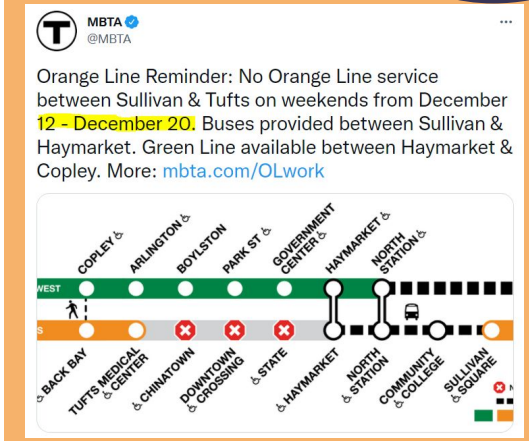
MODEL DEMO

Visualizations: Highest Delay Times

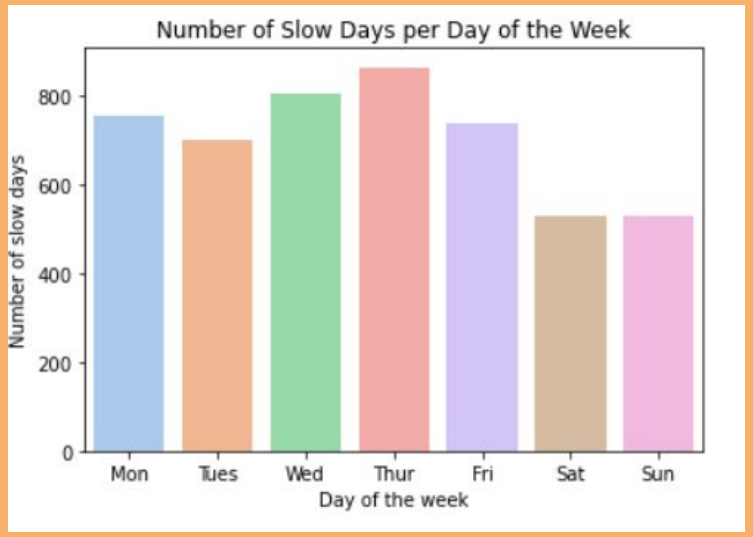
Largest daily average delays: changes in service affect data
1 point: 1 day

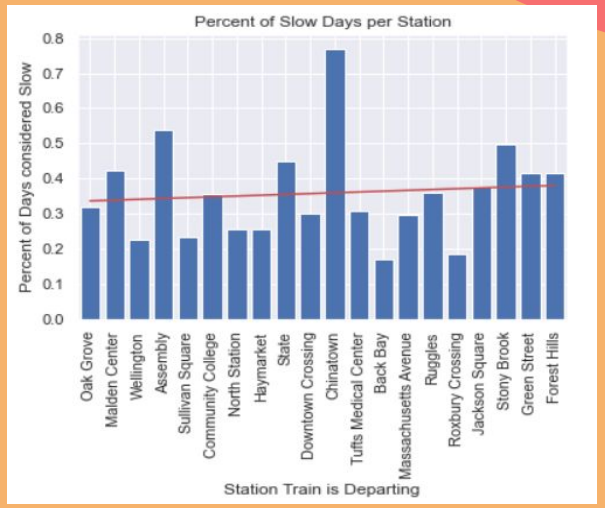
service_date_day	delay_time
2020-01-18	202.952381
2020-01-25	193.452381
2020-01-26	193.976190
2020-02-08	193.357143
2020-02-15	189.238095
2020-02-16	189.904762
2020-12-12	198.619048
2020-12-13	198.261905

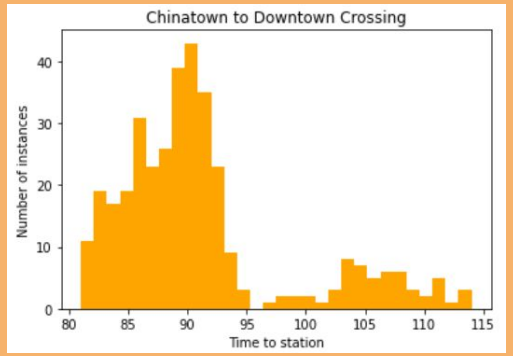


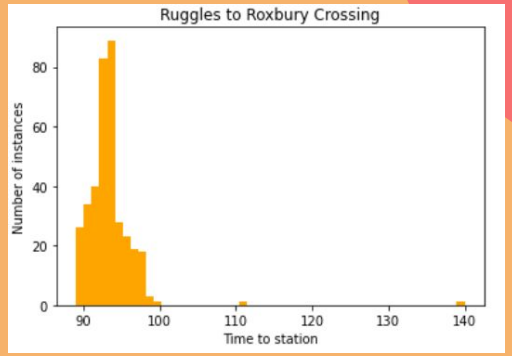


Visualizations: Slow Days & Slow Stations

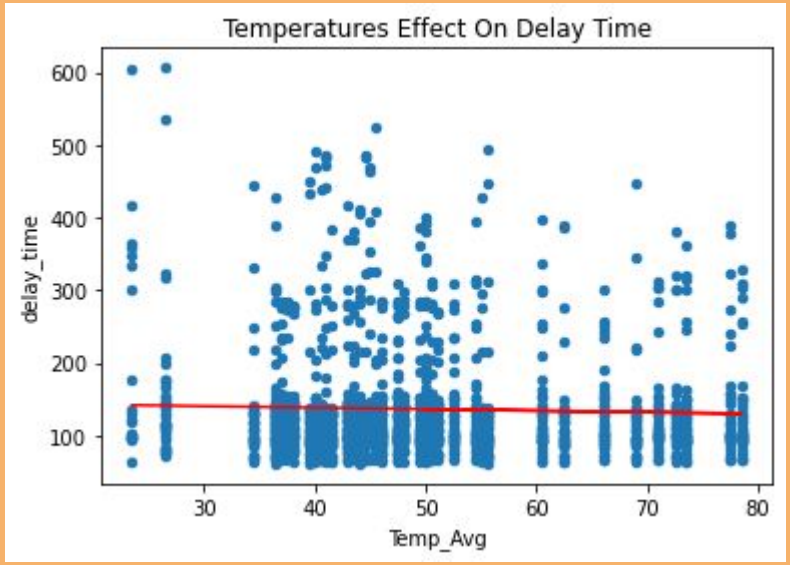


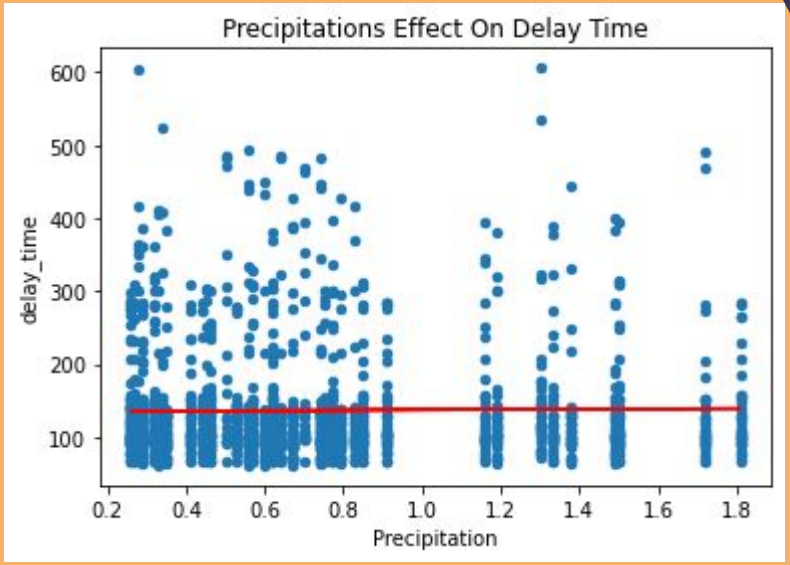




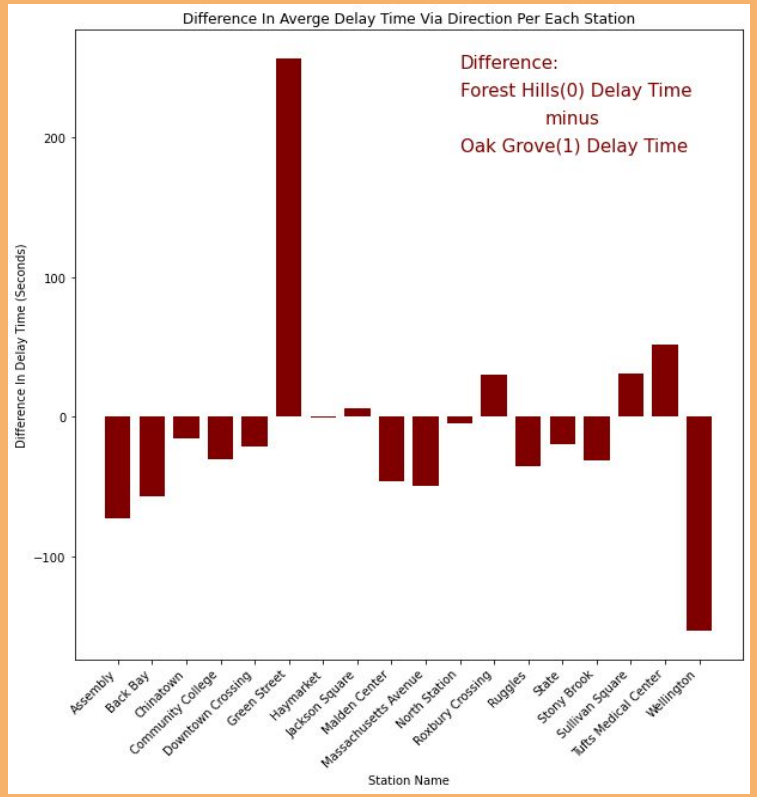


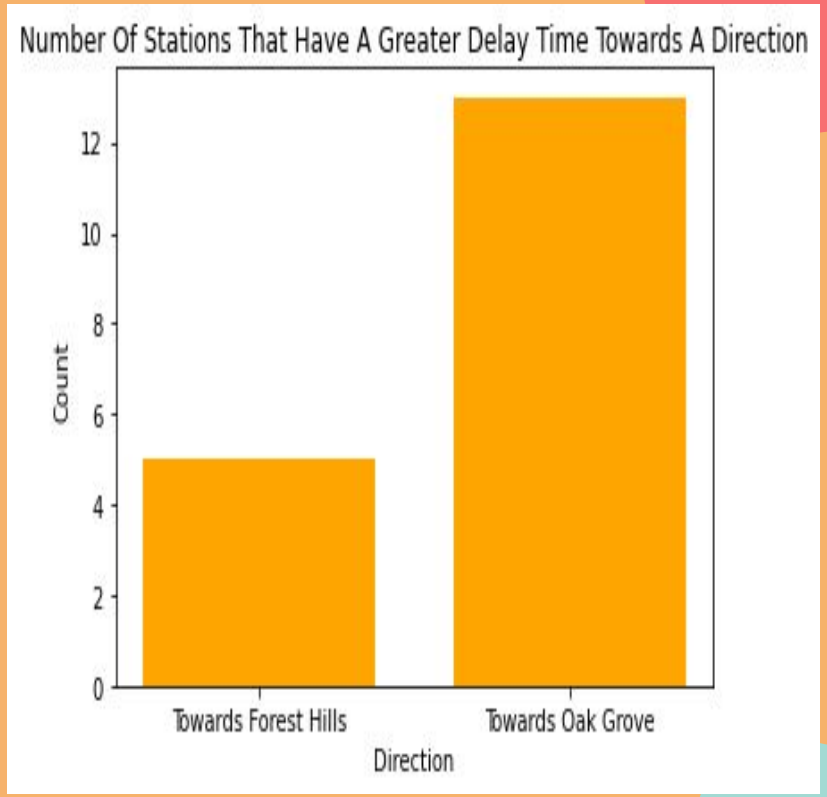
Visualizations: Looking At Weather & Delay Times





Visualizations: Looking At Delay Times At Each Direction

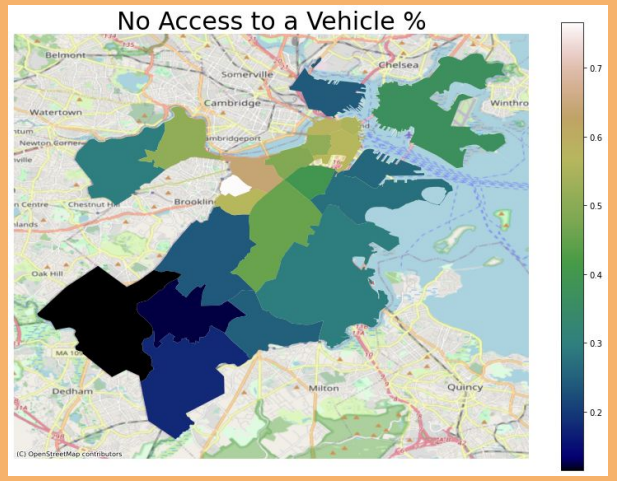




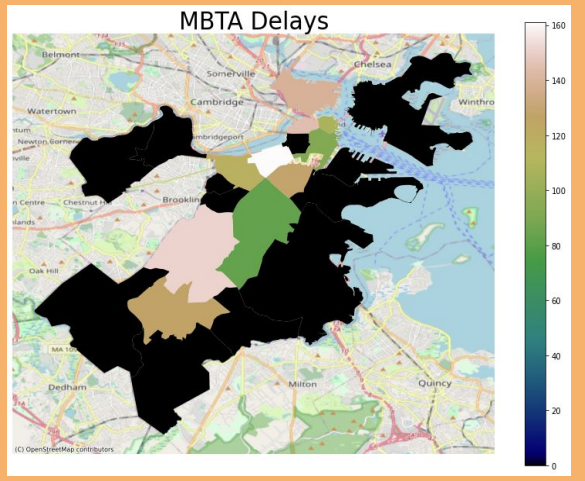
Moving Forward

Boston's Demographics

- Take each demographic statistic from dataset (ie. People without access to a vehicle), create an individual visualization for it
  - Result is a heatmap-esque plot depicting the neighborhood differences in that specific statistic
- Find the average delay time each neighborhood experiences
- Which statistic is most strongly associated with those delay times?
  - Which demographic groups are the most advantaged/disadvantaged?



GeoPandas



Room for Exploration

- Incorporate a wider range of data
  - How were delays pre-Covid?
  - Other MBTA lines?
  - How do other cities fare?

- Fill in the gaps of our demographic delay time visualization
  - Include all neighborhoods
  - More comprehensive and accurate visualization with different MBTA lines

- Relevance to current MBTA affairs
  - How can the MBTA be improved?
  - How can we ensure certain groups aren't as disadvantaged?
  - *What* are the causes of these delays?

Thank you for Listening!