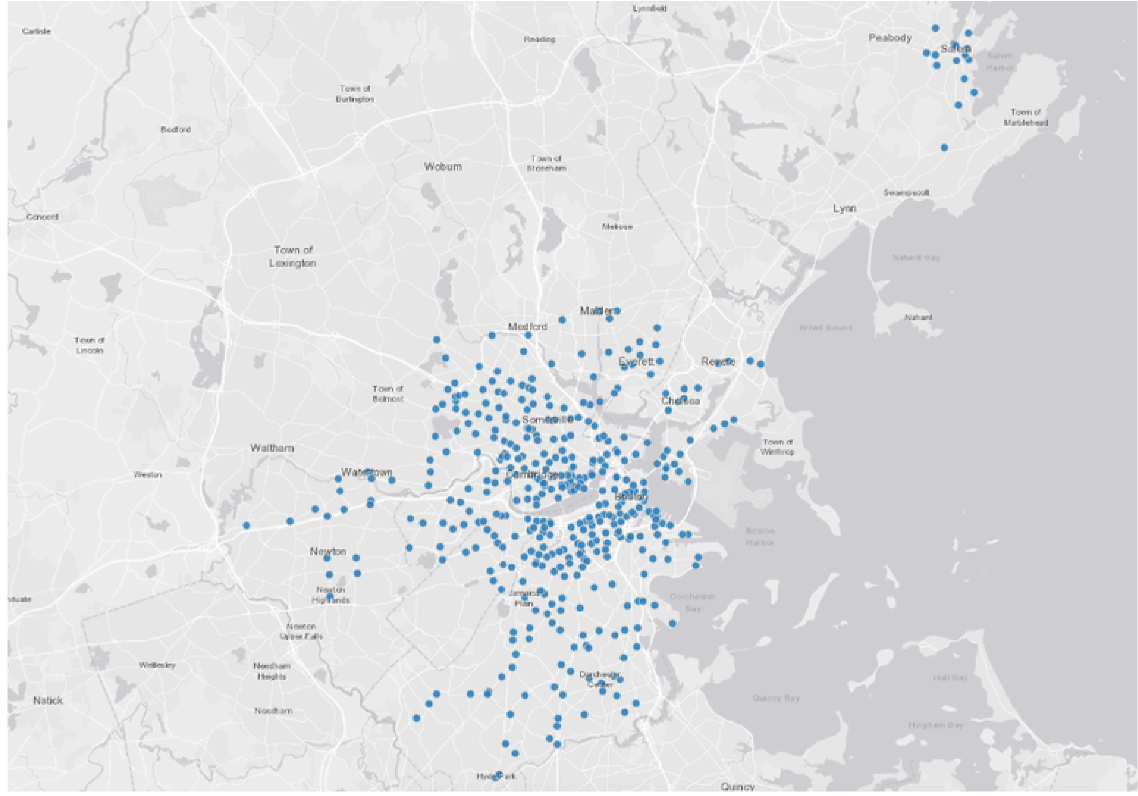
Predicting Bluebikes Usage and Demand in Greater Boston

Jessica Cannon, Ayu Izumo

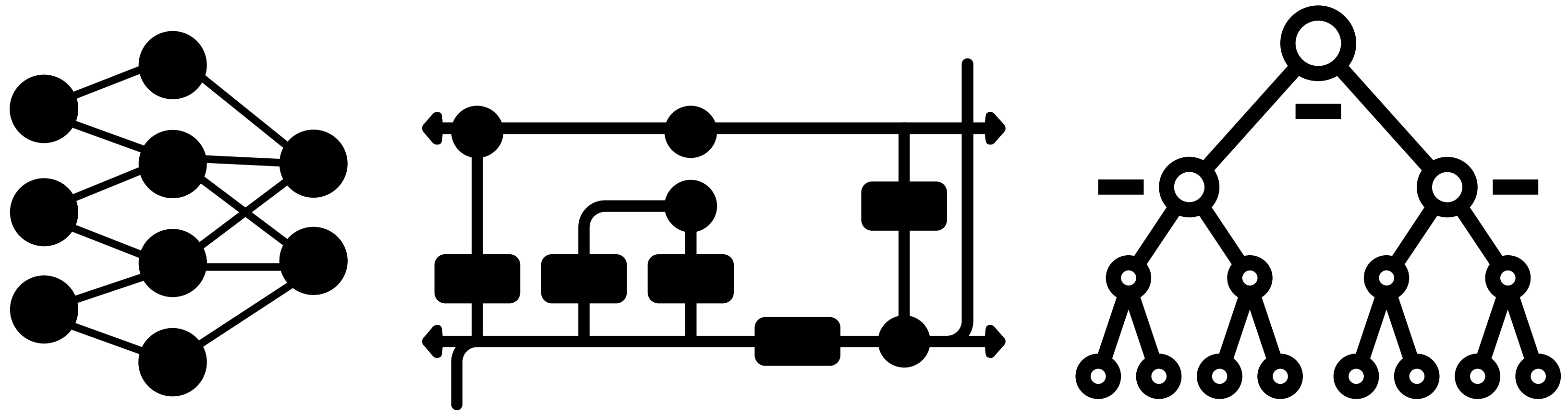
Problem: What we looked into

Which machine learning model, if any, would most effectively predict demand across Boston's Bluebikes system?

- **We wanted to find out:**
  - Which models best predict system usage across 13 different districts in Greater Boston?
  - What features best predict future demand?
- **Doing this allows for:**
  - Optimized bike distribution
  - Reduction of operational costs
  - Improve the overall user experience.
  - etc.

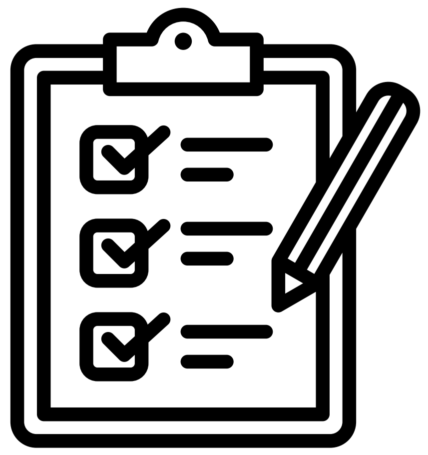


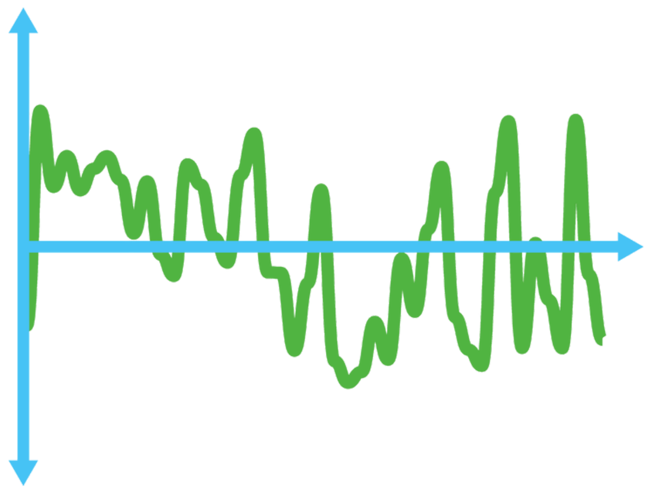
Methods: What we did first



- **Data Preprocessing, Cleaning and Integration**
  - Combined multiple monthly Bluebikes trip datasets from 2019-2023 to create a comprehensive dataset spanning each of those years
- **Exploratory Data Analysis**
  - Analyzed demand variations across time and regions.
- **Model Selection**
  - Chose three different models for initial training and comparison: **Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM), and XGBoost.**

Methods: What we did next





- **Model Evaluation using Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE)**
  - Selected the best model based on these metrics.
- **Model Training and Fine-tuning**
  - Trained the selected model using the combined dataset.
  - Fine-tuned hyperparameters to improve performance.
  - Experimented with different forms of data augmentation to optimize results.
- **Prediction Visualization**
  - Visualized the predicted values against the actual values for model comparison.

Challenge: Differences in column names

The datasets from March 2023 onwards had different column names compared to those of previous months.

Differences in the first four columns, for example:

tripduration	starttime	stoptime	start station id
180	2022-12-01 00:02:44.9630	2022-12-01 00:05:45.2260	115
295	2022-12-01 00:03:11.3990	2022-12-01 00:08:06.9320	32
737	2022-12-01 00:03:51.2520	2022-12-01 00:16:09.2200	200
887	2022-12-01 00:04:26.0750	2022-12-01 00:19:14.0430	74
307	2022-12-01 00:04:34.5180	2022-12-01 00:09:41.6680	515
895	2022-12-01 00:04:37.1280	2022-12-01 00:19:33.0940	74

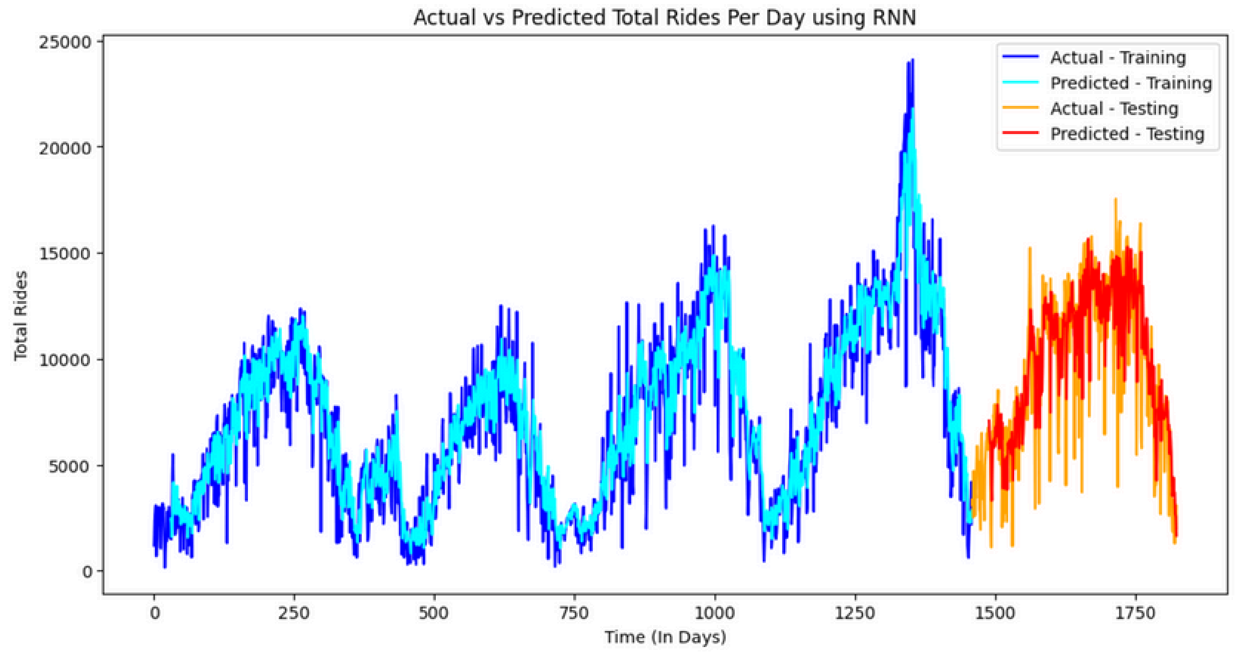
Before March ‘23

ride_id	rideable_type	started_at	ended_at
46B2D1F48BA690A6	docked_bike	2023-06-02 22:19:25	2023-06-02 22:22:01
D29E7DB5DF2DC595	docked_bike	2023-06-27 12:16:52	2023-06-27 12:38:28
DE1C7C6C734C79CC	docked_bike	2023-06-23 19:02:32	2023-06-23 19:18:11
7E74FB2FE8DDAB02	docked_bike	2023-06-07 18:15:48	2023-06-07 18:27:35
4F0FF8181AA60DAF	docked_bike	2023-06-10 15:51:14	2023-06-10 16:45:30
6F5BFB3BD760EB8E	docked_bike	2023-06-22 21:47:52	2023-06-22 21:55:36

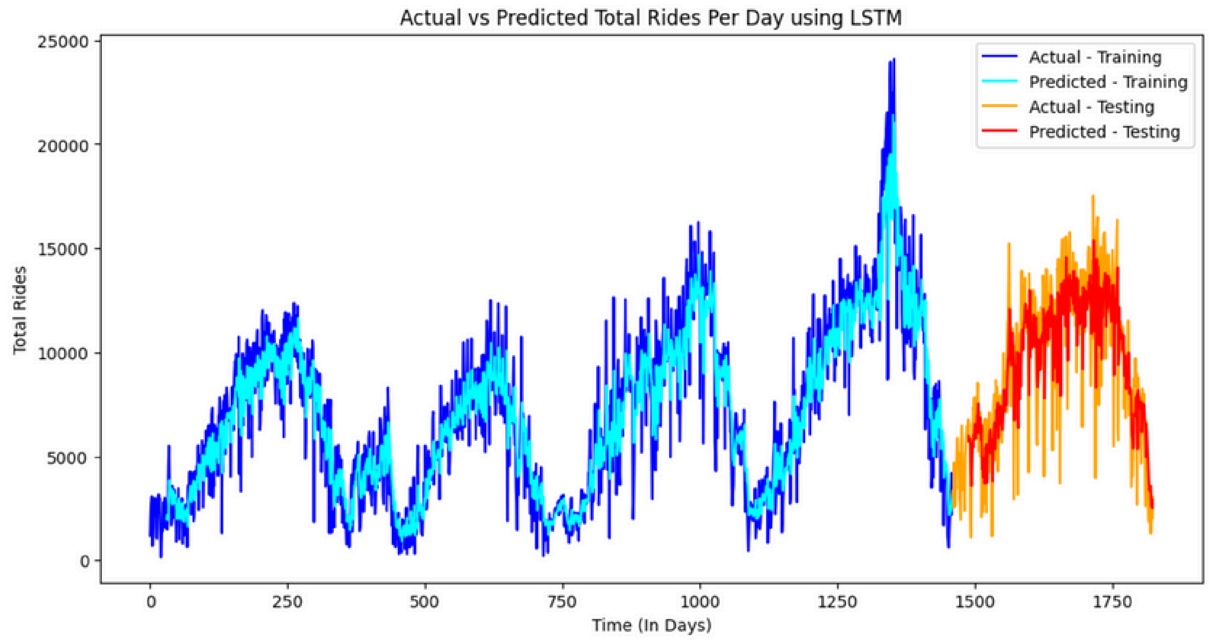
Starting March ‘23

➔ **To work around this, we conditioned on one of the new dataset column names** (‘ride_id’) to rename and rearrange the new datasets’ columns and concatenate them with the previous data.

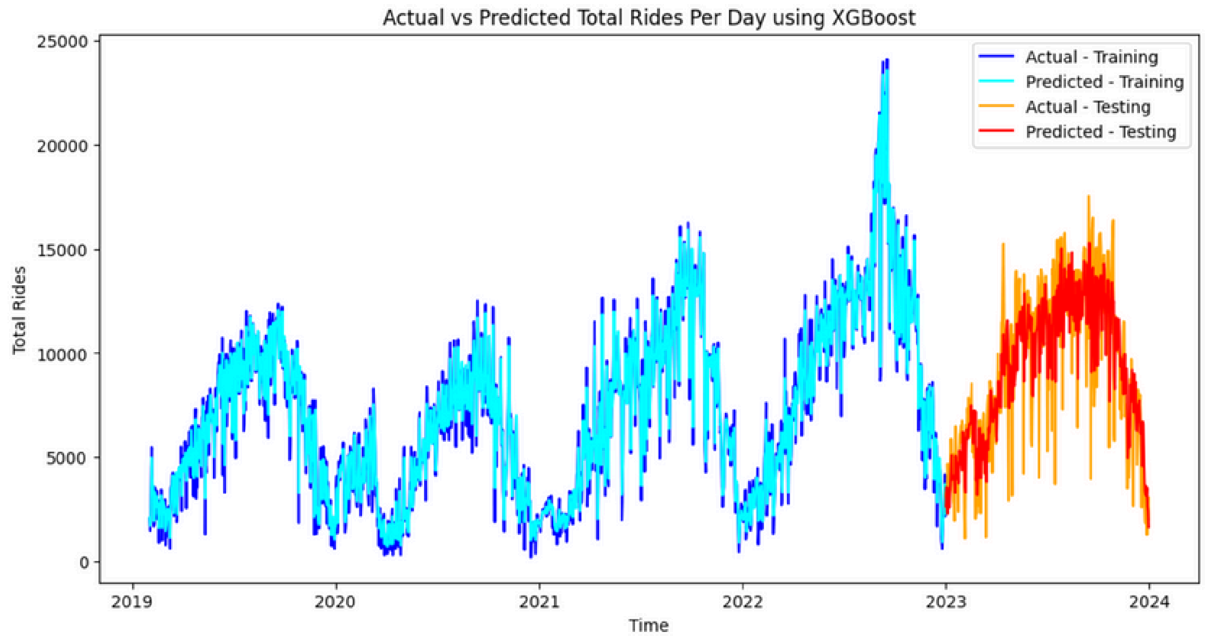
Results: Which model performed best?



RNN: RMSE = 2436.61, MAE = 1901.09



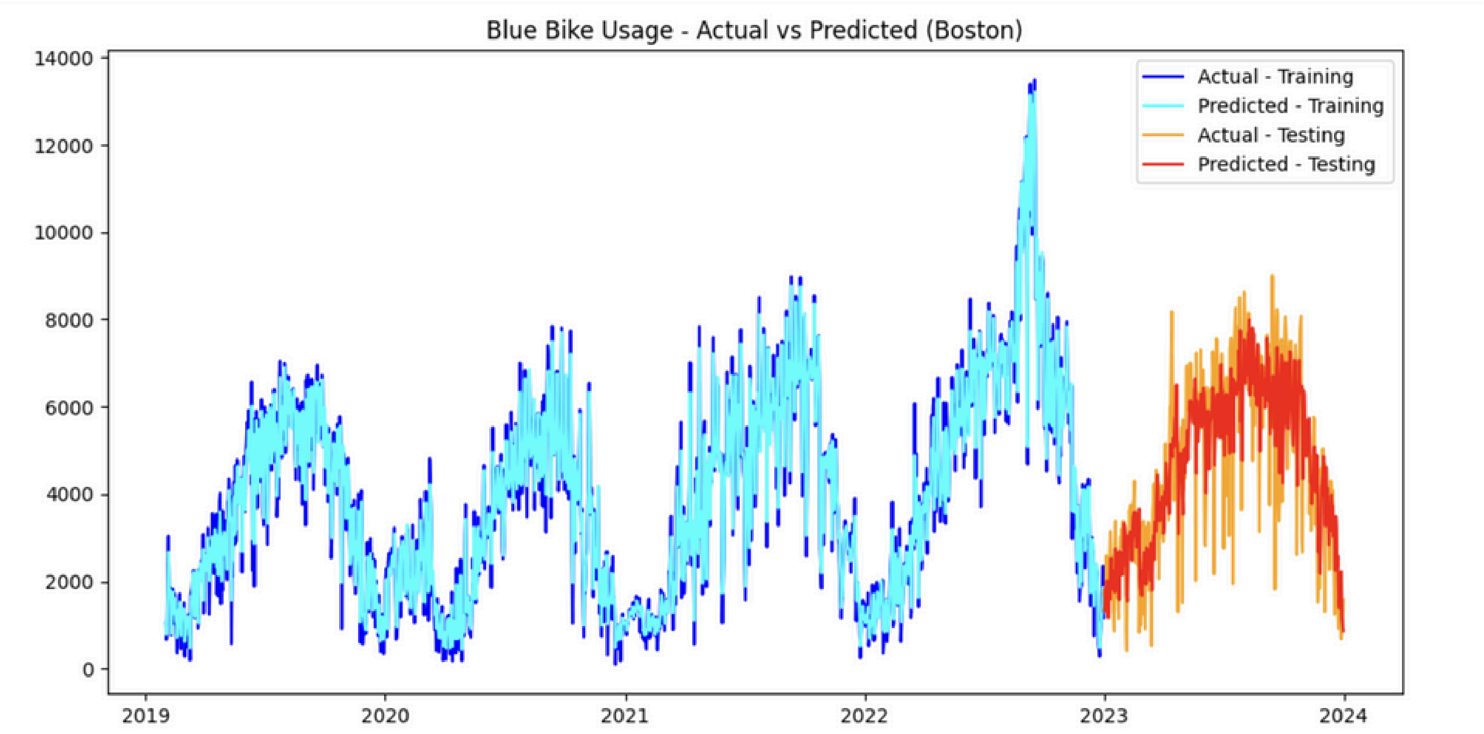
LSTM: RMSE = 2432.99, MAE = 1854.50



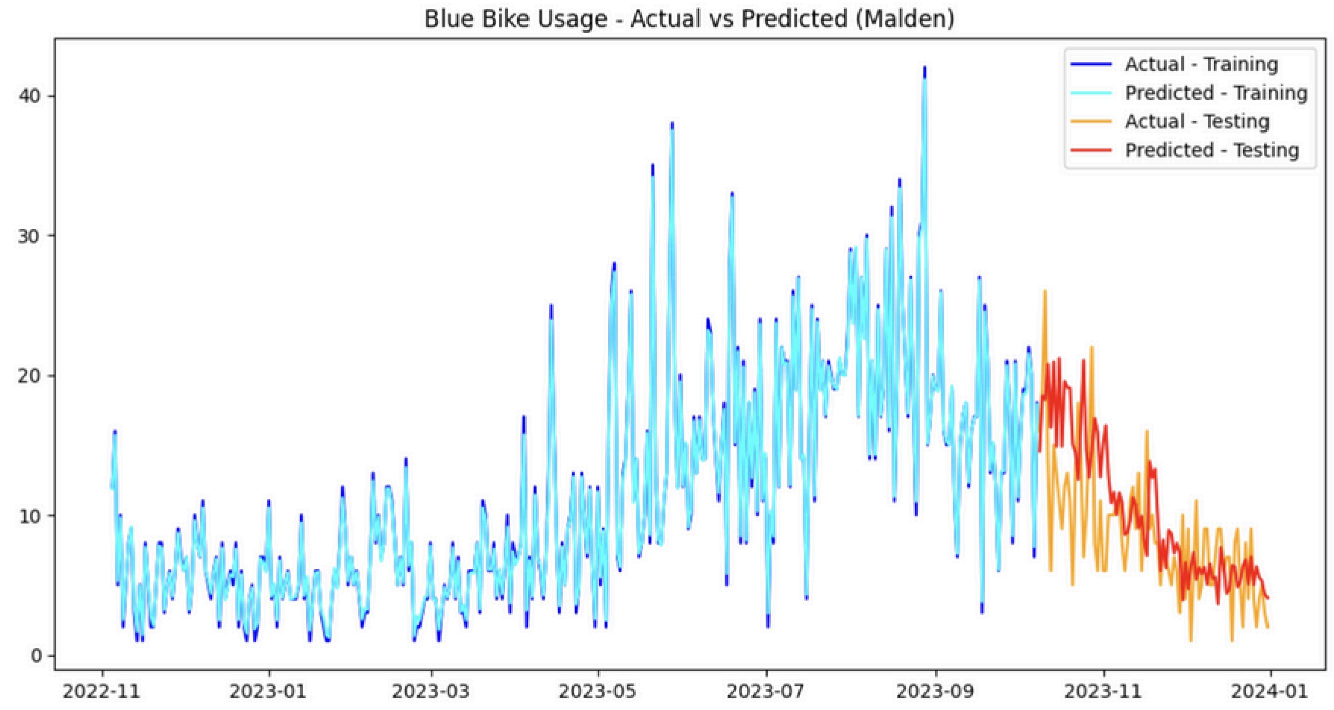
XGBoost: RMSE = 2303.17, MAE =1735.34

- XGBoost > LSTM > RNN ➜ LSTM outperformed RNN as expected, but surprising that XGBoost performed best because LSTM is generally thought to be better for large time series forecasting tasks.

Results: XGBoost and District Demand



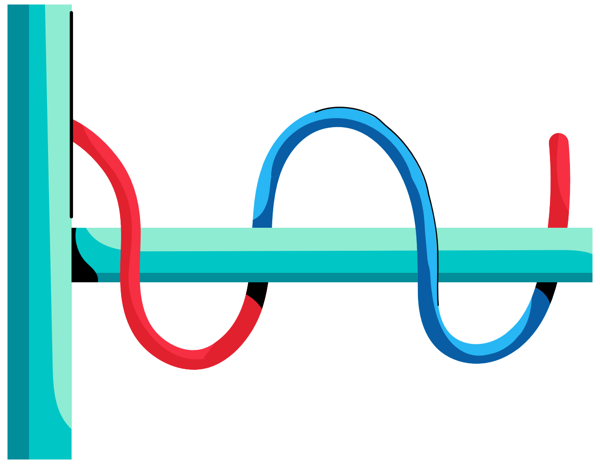
Boston, starting from January ‘19



Malden, starting from November ‘22

- XGBoost **did well in predicting demand in all districts**, but generally **did better predicting on districts with more available trip data, like Boston, with smaller Test RMSE/Train RMSE ratios.**
  - Test RMSE/Train RMSE ratio of 4.95 for Boston (data starting from 2019), 13.43 for Malden (data starting from 2022).
- Across all districts, we noticed that the usage tended to spike in the summer and fall months, and saw a drastic decline during the winter

Conclusions: Patterns and Seasons



- **XGBoost did best overall** in predicting trip data.
- Likely due to **patterns seen in demand over seasons** in a given year, making seasons a likely predictor for demand.
- **LSTMs and RNNs did better in making more consistent predictions** - less prone to overfitting than XGBoost, due to lower test RMSE/train RMSE ratios.
- XGBoost seems to do better with data that includes patterns, whereas LSTMs are better at handling long-term data that lacks immediate patterns or structures.

Fun Fact: The earliest recorded Bluebikes trip

- The earliest recorded Bluebikes trip took place in **July of 2011**, back when the system (then called Hubway) first became operational.
- This trip started at the **Boston Public Library** and totaled approximately 998.45 meters, or 0.6 miles.
- Since then, more than **23 million trips** have been taken by Bluebikes riders across the regions of Greater Boston

Thank you!