EECE5644
Introduction to Machine Learning
Final Project Milestone Report
Members: Srinishanth Rajarajan, James Packard, Ethan Neal

The MBTA Bus and Commuter Rail Ridership data sets contain trip data for bus routes and rail lines in the Greater Boston area. Our project aims to identify trends in ridership and trip duration through exploratory data analysis, which could be used to estimate commute times based on the factors in this dataset. For this milestone, we have prepared visualizations of trends in these data sets.

In the Commuter Rail data set, the time of day is predictably a large factor in determining the number of riders. The peak hours are at 8am and 6pm EST, so these values will likely become the cut-points in our regression tree analysis, where the first node determines whether the stop time is greater than the first peak of 8am. A bar chart of the average ridership is shown in Figure 1, in UTC.

Certain Commuter Rail lines are less utilized than others. Specifically, the Fairmount line has far fewer riders on average than the most popular line, Providence. We expect the line to be a factor deeper in the regression tree, selecting out the most or least popular lines. Figure 2 shows the average ridership across each of the Commuter Rail lines in a bar chart.

Before the final project report, we plan to build the regression tree to identify the most relevant measurements from the data sets, and determine a regression model using those measurements.

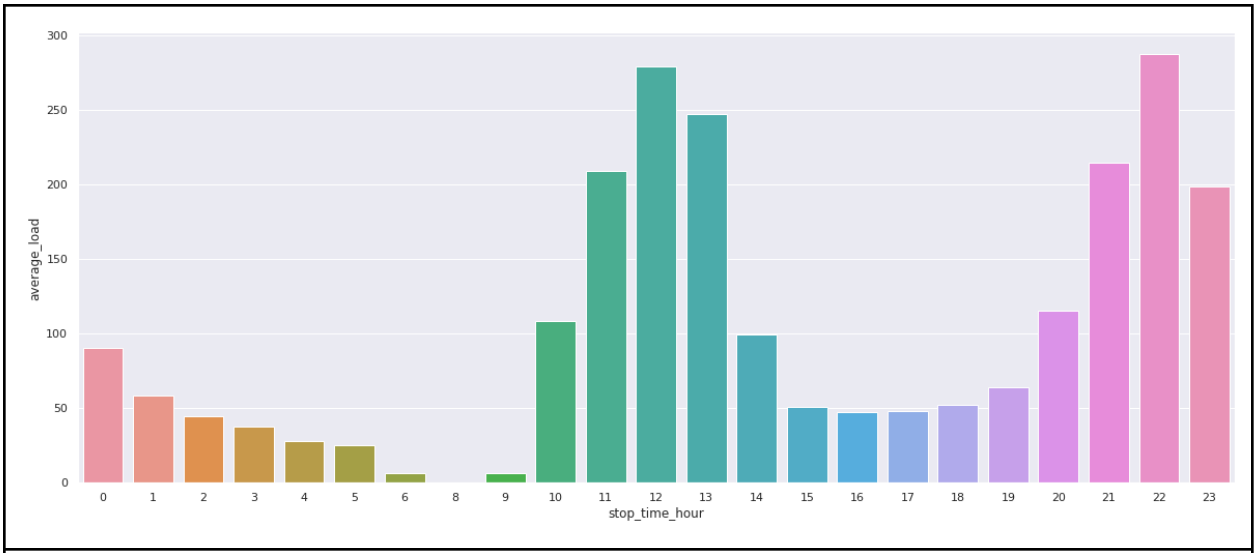

Figure 1: Average Commuter Rail Ridership by Time of Day (UTC)

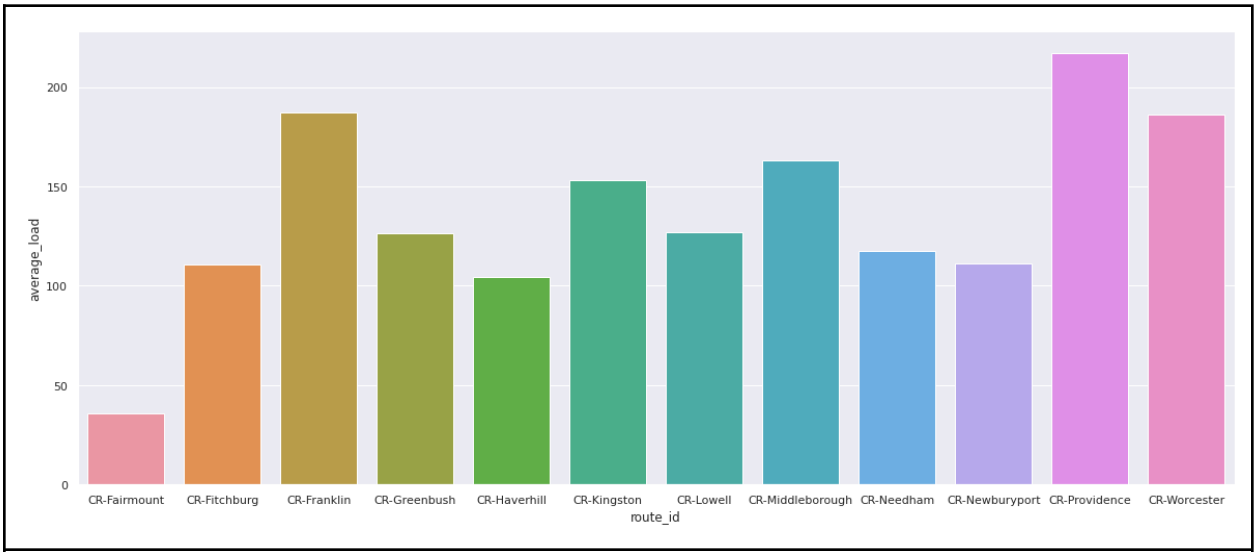

Figure 2: Average Commuter Rail Ridership by Line