DS4200 Assignment 6

Anjali Tanna

Github Repository

---

1 - Dataset

Dataset downloaded.

---

2 - Data Cleaning and Exploration

I cleaned and explored the dataset using Jupyter Notebooks. The code showing the data cleanup can be found in the `.ipynb` file in my Github repository.

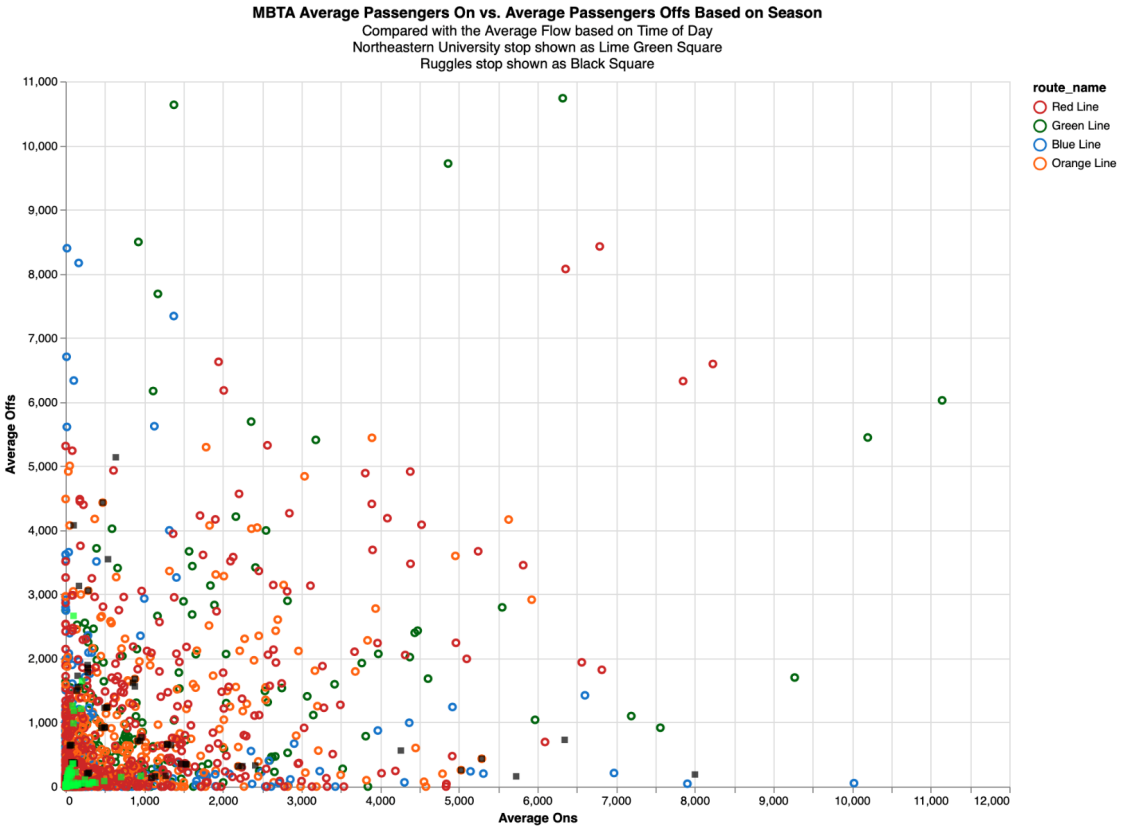
---

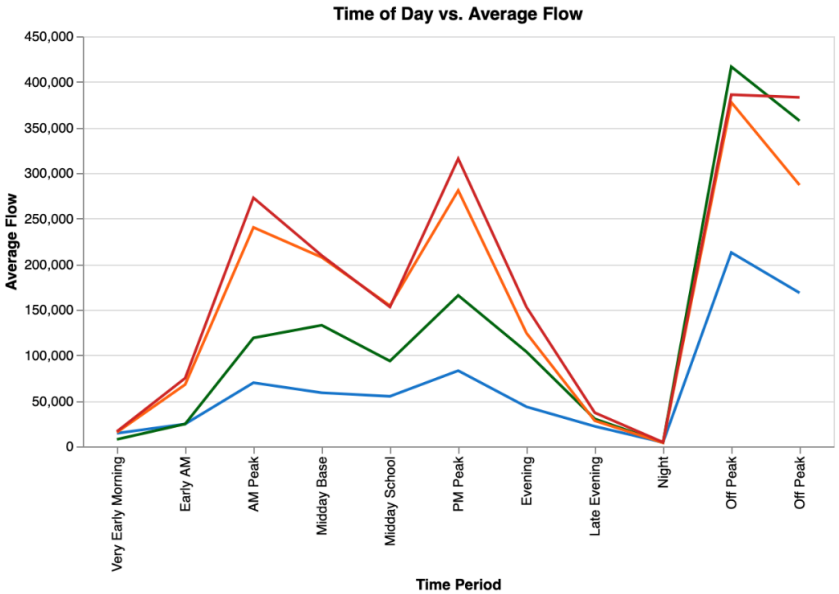
3 - Jupyter Notebook

The Jupyter Notebook for my Altair visualizations can be accessed through my Github repository here:

Jupyter Notebook file

4 - Scatter Plot and Line Chart (captions in part 7)

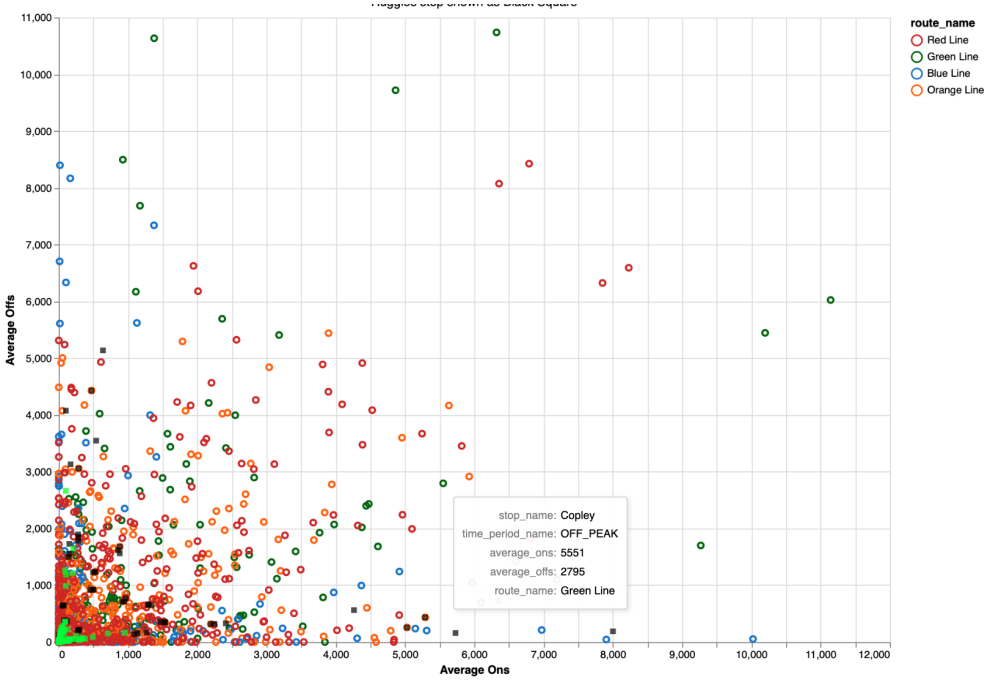


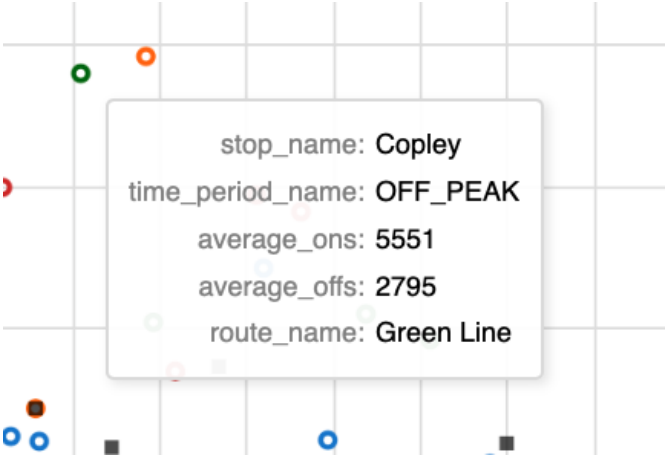


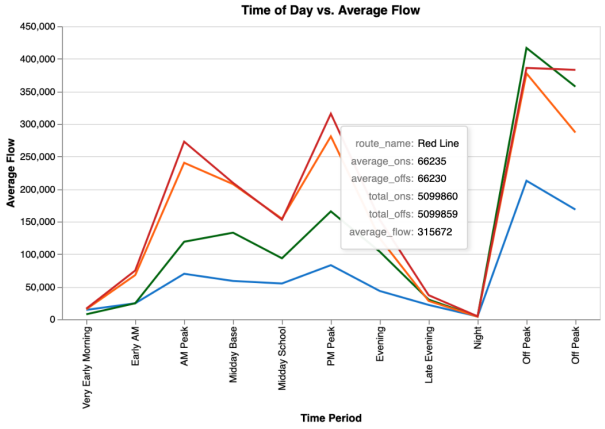
Season Fall 2019

5 - Details on Demand

On the scatter plot, there is a details-on-demand feature that shows the 'stop_name', 'time_period_name', 'average_ons', 'average_offs',` and `'route_name'`. On the line chart, there is a details-on-demand feature that shows the `'route_name', 'average_ons', 'average_offs', 'total_ons', 'total_offs', and `'average_flow'.







6 - Pop-Out Effect

The pop-out effect allows users to easily locate the Northeastern University and the Ruggles stops. The lime green squares denote the Northeastern University stop and the black squares denote the Ruggles stop, as mentioned in the title. Unfortunately, I could not layer this legend with the bigger scatter plot due to interactivity restrictions, but this description is listed in the title. Looking at the scatter plot, it is easy to tell that the Ruggles stop has higher averages at most points in the day compared with the Northeastern stop. This makes sense as those who use the Northeastern stop are primarily Northeastern students, but the Ruggles stop is generally used by the greater community.

7 - Explanations

Scatter Plot

The scatter plot represents the average passengers on compared with the average passengers off for all stops on all routes during all times of day. This scatter plot gives a good visual on all stops and time of day in the data, to see which route is the most popular and when. There is a details-on-demand feature that shows the 'stop_name', 'time_period_name', 'average_ons', 'average_offs', and 'route_name'. The color represents which line it is, either the Red, Green, Orange, or Blue lines. Using the dropdown menu at the bottom, you can select which season you want to look at (Fall 2017, Fall 2018, or Fall 2019). This graph is also linked to the line chart below, so when you select a point, all points in that route are shown, the rest are grayed out, and the corresponding line in the line chart is also shown with other lines grayed out. The lime green squares denote the Northeastern University stop and the black squares denote the Ruggles stop, as mentioned in the title. Looking at the scatter plot, many of the points at the higher ends of the x and y-axis are from the Red line and Green line. The Park Street Green line has the highest 'average_ons' during the 'OFF_PEAK' time period. It is interesting to see that the Blue line is the least popular but it makes sense as I have never been on it. I expected more Orange line points to be higher on the graph, so it is interesting to see that the Green and Red lines are more popular. It is also interesting to see how the plot changes as you select a different season.

Line Chart

This line chart shows the time period compared with the average flow of all the routes. This graph is a good indication of which lines have the most passengers during certain parts of the day. The line chart also responds to the dropdown menu, so you can see how the `'average_flow'` has changed throughout the years. It is clear that the Green line during the 'OFF_PEAK' has the highest 'average_flow'. However, the Red line has higher 'average_flow' during most time periods in the beginning of the day. Similar to the scatter plot, if you select a line on the line chart, only that route will appear and the same route will be reflected on the scatter plot. There is a details-on-demand feature that shows the 'route_name', 'average_ons', 'average_offs', 'total_ons', 'total_offs', and 'average_flow'. The color represents the line (Red, Green, Orange, or Blue).

Pop-Out Effect

The pop-out effect I chose to use utilized shape and color. I changed the shape of the points for Northeastern University and the Ruggles stops to squares, so it was easily seen against the outlined circles that were the other stops. I took it one step further and changed the color of these points so it was even easier to see them. Since there was a lot of data being plotted, these two effects made it easy for users to locate these stops on the plot.