Mindpetal

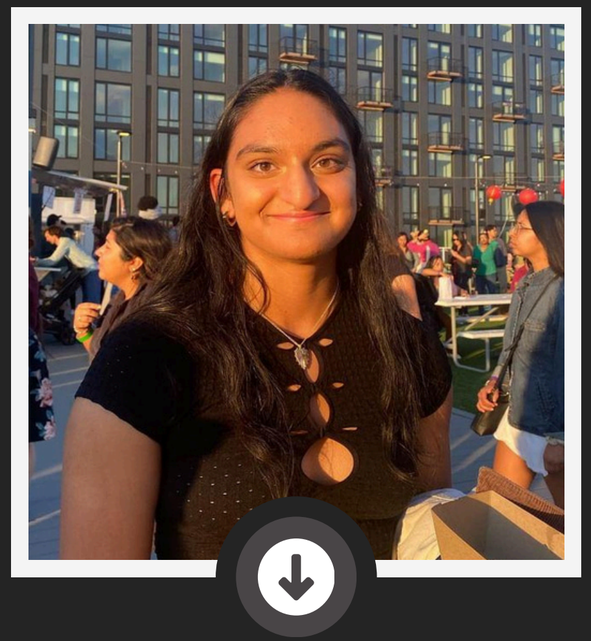
INFO CHALLENGE 2025

DC WMATA Metro Ridership Analysis

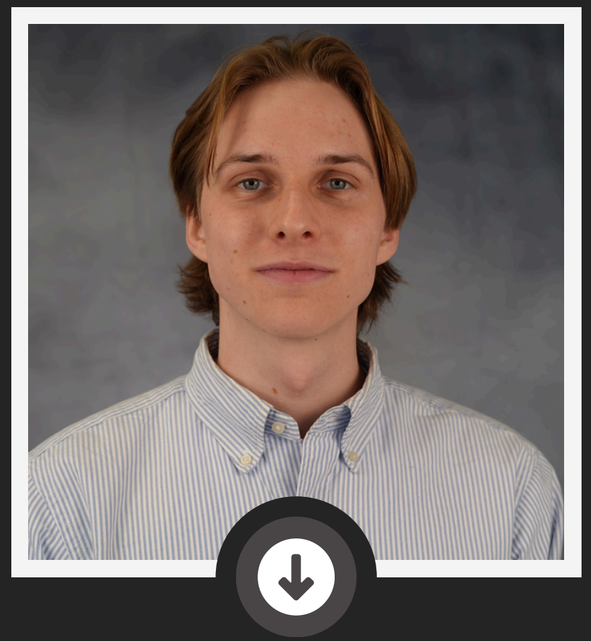
Team #: IC25069

Team Members: Maya Patel, Illia Polishchuck, Adrien Rozario

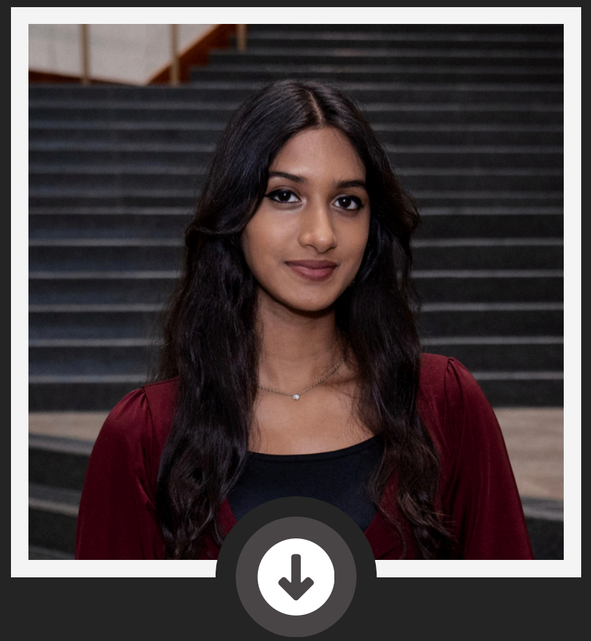
Our Team



Maya Patel



Illia Polishchuk



Adrien Rozario

Problem Statement

The Washington Metropolitan Transit Authority (WMATA) experiences fluctuating ridership patterns throughout the year, influenced by seasonal trends, station locations, and metro line usage.

However, without a data-driven approach to understanding these patterns, Metro may struggle with resource allocation, congestion management, and pricing optimization.

Additionally, businesses near metro stations may lack insights into potential customer foot traffic.

Proposal

This project aims to analyze ridership trends across different time periods and factors to:

- provide meaningful insights for Metro’s operational planning
- assist businesses seeking to optimize marketing and service strategies

Research Questions

1. Overall Ridership Trends (Big Picture for Metro & Businesses)

- What season has the highest overall ridership? (Metro can prepare for peak demand; businesses can plan marketing strategies.)
- What months have the highest overall ridership? (Reinforces seasonal trends.)

2. Metro Line Analysis (Broad Perspective → Business Relevance)

- Which metro lines handle the most ridership overall? (Metro can prioritize improvements; high-traffic lines are valuable for ads.)
- How does ridership fluctuate across metro lines by month? (Identifies seasonal trends for each line.)

Research Questions

3. Station-Specific Ridership (Metro Service & Business Insights Together)

- What are the top 10 busiest stations each month of the year? (Frequent usage → Metro needs more service; businesses near these stations can thrive.)
- What are the middle 10 busiest stations each month of the year? (Provides mid-range perspective on ridership.)
- What are the least 10 busiest stations each month of the year? (Metro may adjust resources; are there untapped business opportunities?)

Research Questions

4. Metro Line & Station Type Analysis (Broad Perspective → Business Relevance)

- How is ridership distributed across metro stations and lines over time? (Helps Metro assess usage trends and service needs.)
- How many stations of each type (residential, commercial, transfer) exist on each metro line? (Reveals structural distribution of different station types.)
- Which station types (residential, commercial, transfer) handle the most traffic? (Tells businesses where customer foot traffic is strongest.)

5. Business-Driven Optimization (Closing with Data-Backed Opportunities)

- Can Metro adjust pricing based on congestion? (Helps optimize revenue and ridership.)

Data Collection & Processing

We collected data from secondary sources with publicly available raw data, such as WMATA Metrorail Ridership Summary and WMATA Corridor Data Maps. In order to process the data, we normalized it to have the necessary tables and relationships.

The columns with relevant information were merged into data frames to be used in visualizations.

Datasets were extracted from these sites and manipulated with Python scripts to optimize our usage.

The dataset on the geographical data of each metro station (metro_stations.csv) had station names altered to match those of the other datasets to allow merge on station name.

Data Collection & Processing

Datasets:

- **All_Months.csv:** dataset on Average daily entries by station by month for all days of the week in 2024
- **Annual_Station_Boarding_Compiled.csv:** compiled average annual daily entries for all days, weekdays, and weekends in respective order
- **entries_exits_transformed_data.csv:** entries and exits per each day and station in 2024 (numeric only)
- **everyday_2024_w_metro_stations.csv:** entries and exits per each day and station in 2024 merged with metro_stations.csv
- **merged_df.csv :** merged_df is a result of merging All_Months.csv and metro_stations.csv on Station Name
- **metro_stations.csv:** data regarding location and specs of each metrorail station

Exploratory Data Analysis

Overall Ridership Trends

- Histogram of Average Daily Ridership by Season in 2024
- Histogram of Average Total Daily Entries by Month and Season

Metro Line Analysis

- Histogram of Lines and Ridership (Entries + Exits)
- Line Plot of Total Ridership Per Metro Line by Month

Station-Specific Ridership

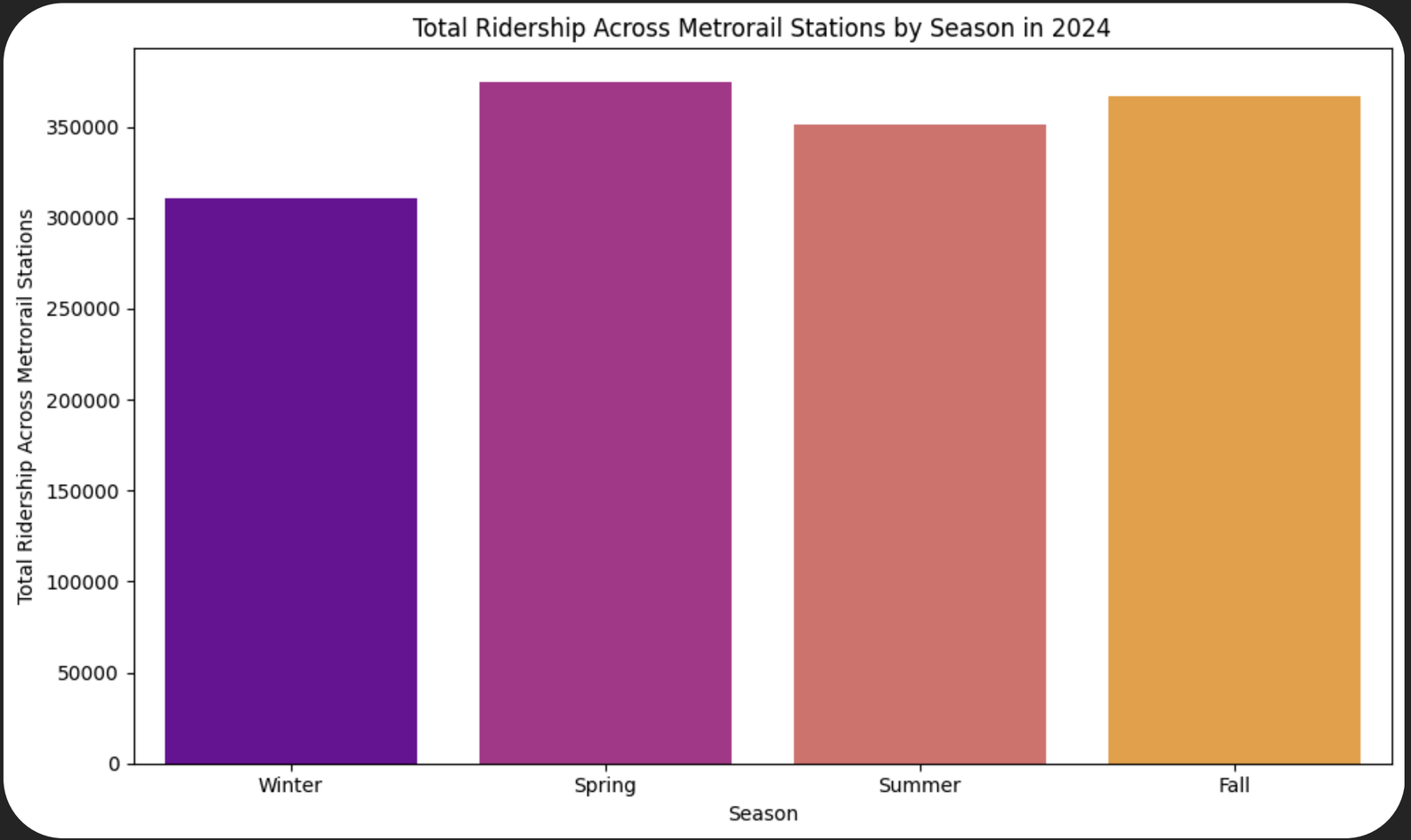
- Line Graph of Average Daily Entries for the Top 10 Busiest Metro Stations by Month in 2024
- Line Graph of Average Daily Entries for the Middle 10 Most Used Metro Stations by Month in 2024
- Line Graph of Average Daily Entries for the 10 Least Busy Metro Stations by Month in 2024

Metro Line & Station Type Analysis

- Heatmap of Monthly Total Daily Entries by Station by Line
- Grouped Bar Chart of Number of Each Station Type by Line
- Grouped Bar Chart and Pie Chart of Ridership by Station Type

Overall Ridership Trends:

What season has the highest overall ridership?

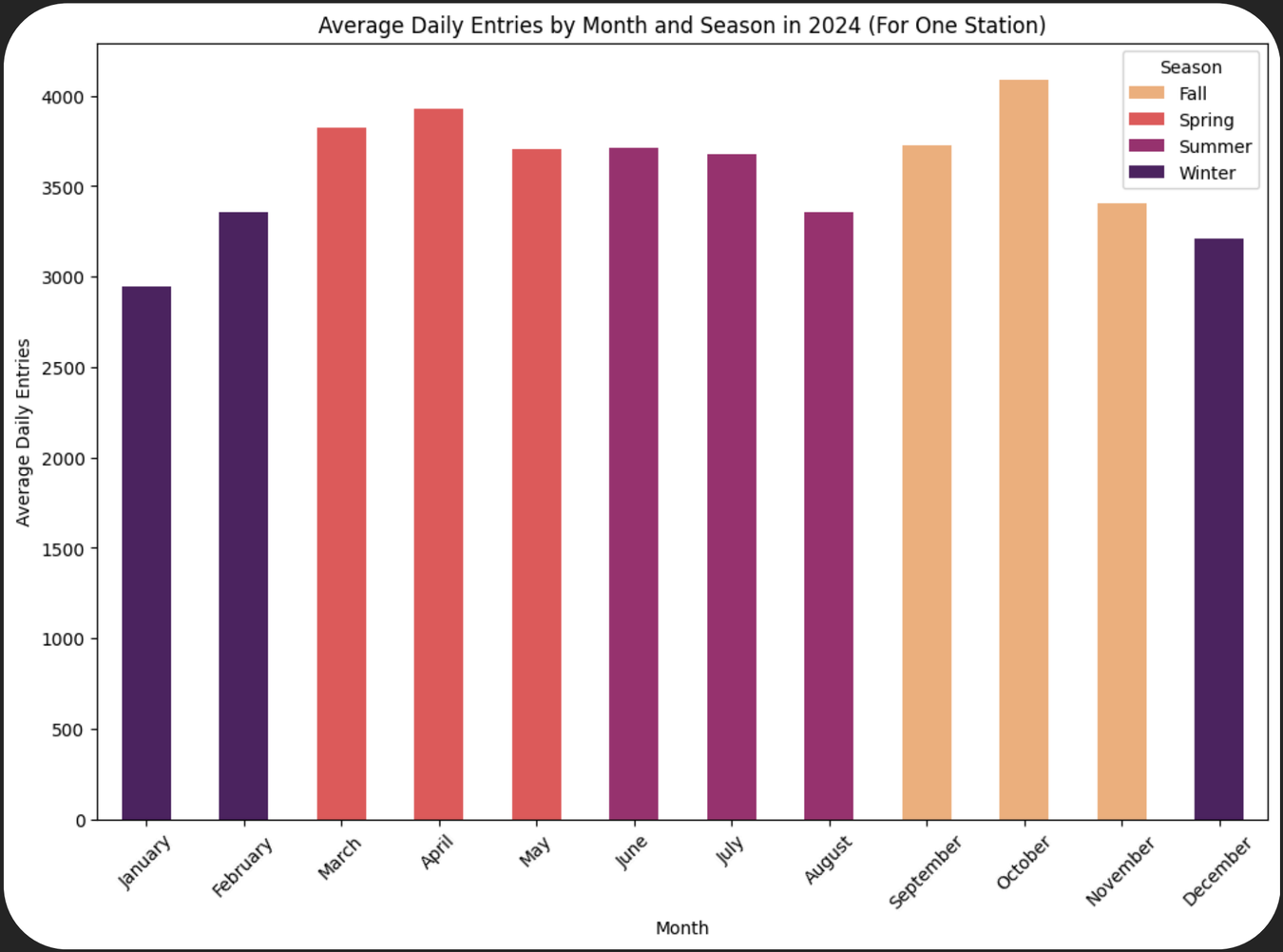


We started our analysis with the graph showcasing total ridership among all the metro stations throughout the four seasons of 2024.

We can see that although ridership fluctuates, peaking in spring and being the lowest in winter, this graph does not provide enough data to make any concrete conclusions regarding ridership patterns, so we decided to look into total ridership every month of the year.

Overall Ridership Trends:

What months have the highest overall ridership?



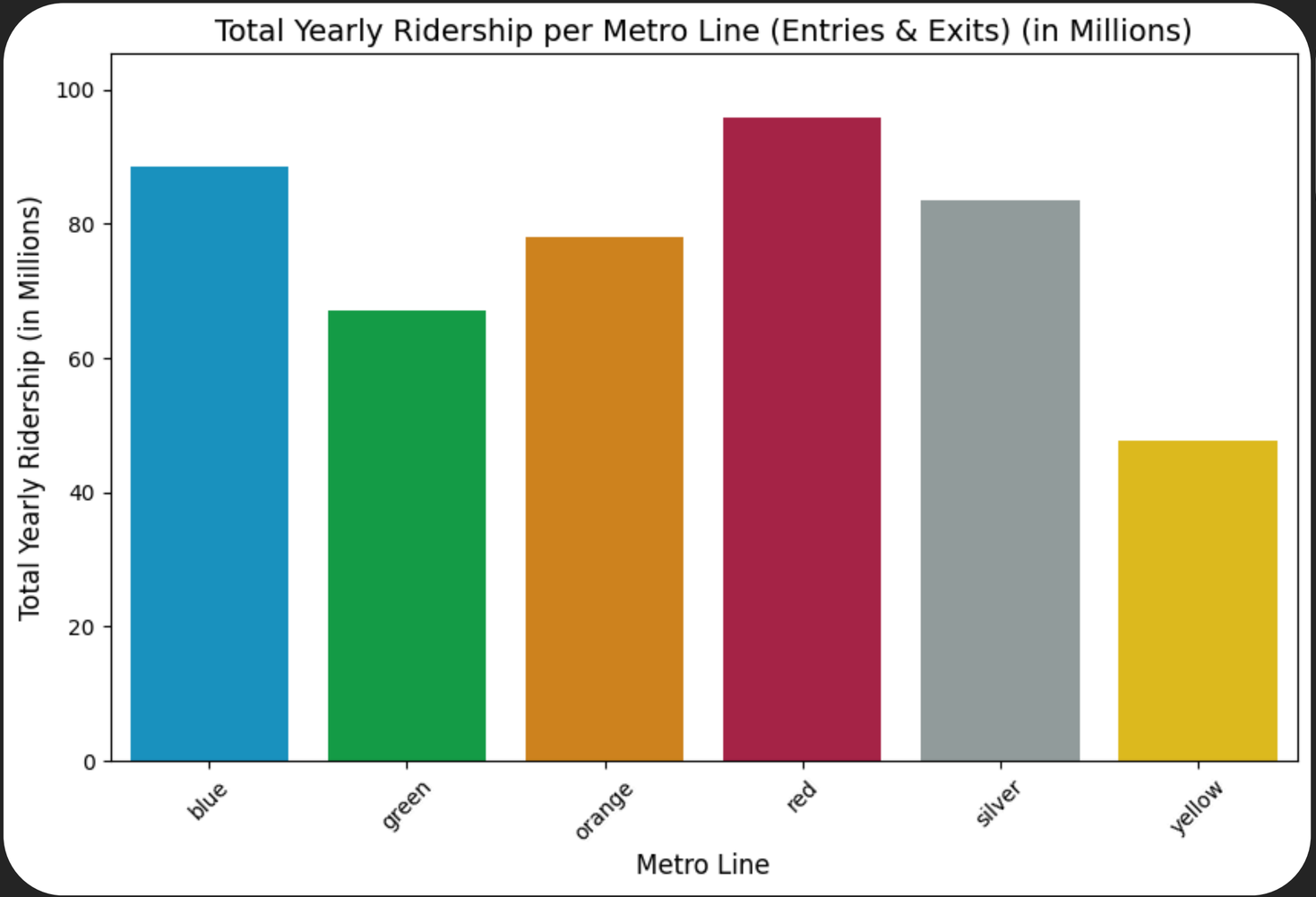
Overall trends for months correspond to average ridership to their respective seasons.

The spikes in April and October show that some months have peaks, but that the distribution of any season is not significantly impacted.

This prompted us to explore if the metro lines followed the same pattern, and that no particular line skews any month or season.

Metro Line Analysis:

Which metro lines handle the most ridership overall?



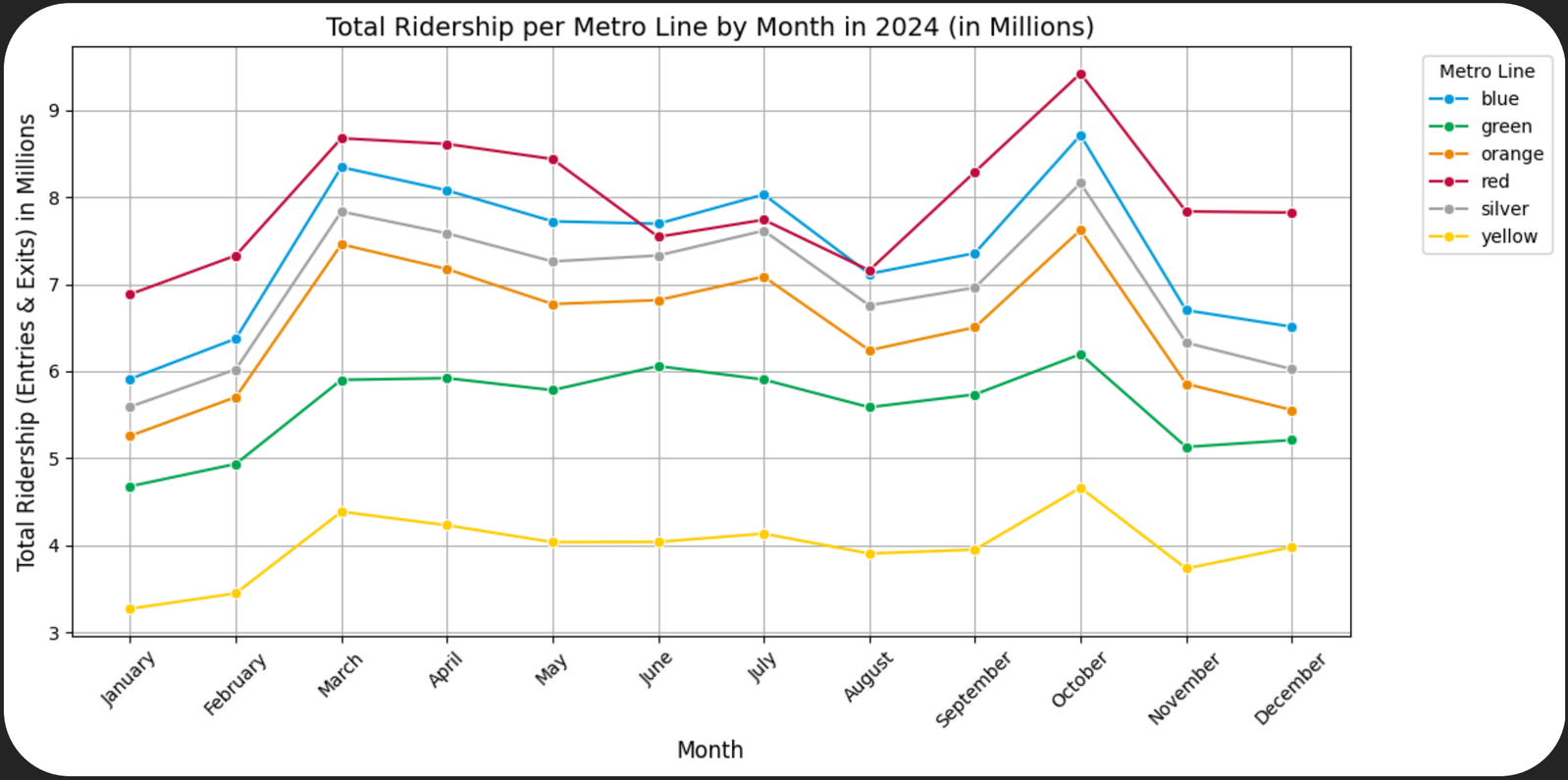
	Metro Line	Total Yearly Ridership (in Millions)
0	blue	88.558285
1	green	67.031008
2	orange	78.045014
3	red	95.759772
4	silver	83.468273
5	yellow	47.767306

This graph showcases the total yearly ridership in millions, with the red line coming in first with 95.76 million yearly riders, and the yellow line coming in last with 47.77 yearly riders.

This helped understand that some lines are more frequented than others and set ground for the upcoming line graphs, analyzing ridership per line per month.

Metro Line Analysis:

How does ridership fluctuate across metro lines by month?

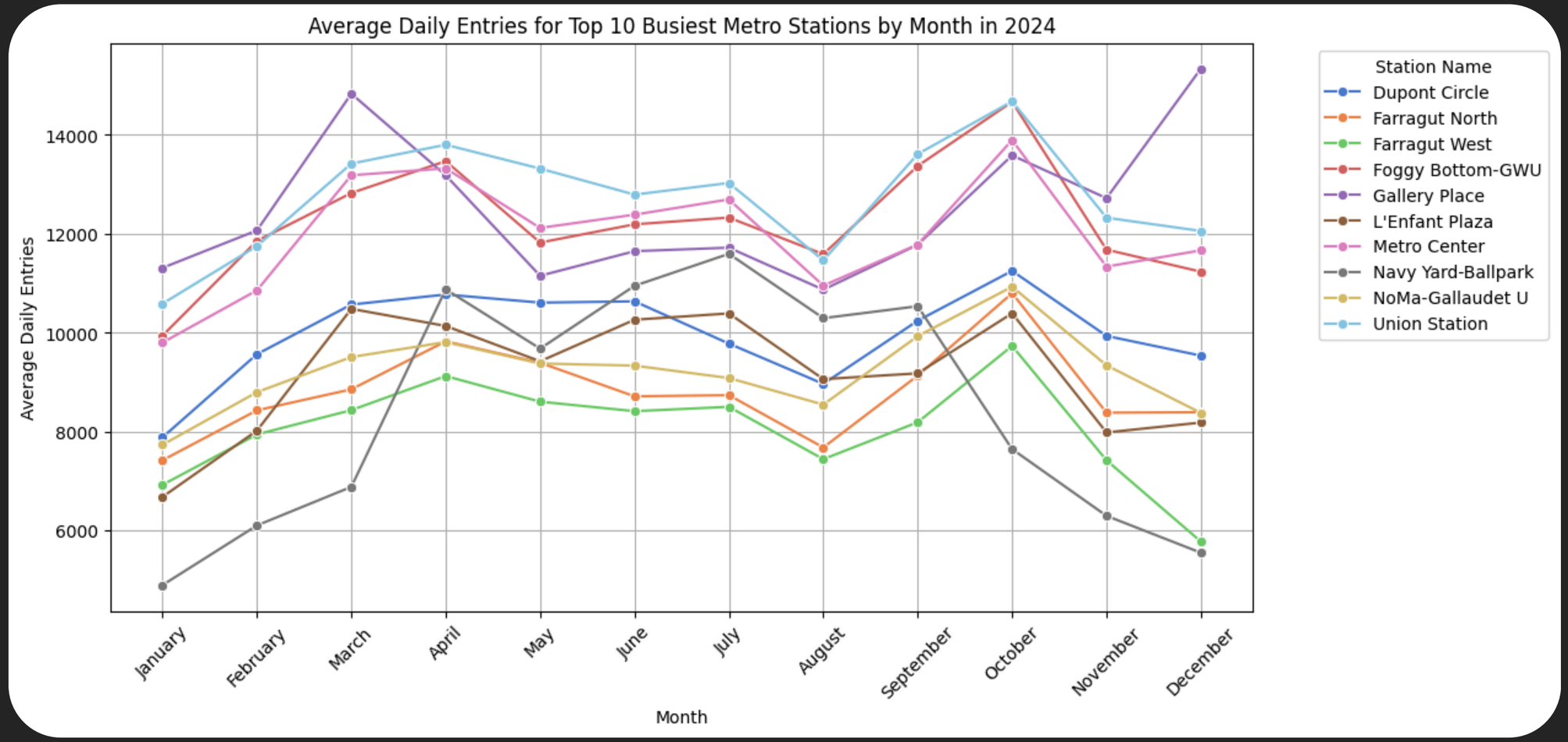


This graph depicts the total yearly ridership in millions, with the red line coming in first with 95.76 million yearly riders, and the yellow line coming in last with 47.77 yearly riders.

This helped understand that some lines are more frequented than others and set ground for the upcoming line graphs, analyzing ridership per line per month.

Station-Specific Ridership:

What are the top 10 busiest stations each month of the year?



In January and February, the average daily entries are at the lowest levels for all of the 10 listed stations, which can be attributed to the cold weather that discourages many people from stepping out and riding the metro.

In March, the average daily entries for Gallery Place suddenly reaches its first peak, which can be attributed to the warming weather and cherry blossom season. There are also many attractions nearby such as the National Portrait Gallery, Smithsonian American Art Museum, Capital One Arena, Shakespeare Theater, Kogo Courtyard, and more.

In April, the average daily entries for Union Station reaches its first peak, which can be attributed to the its many attractions nearby that people can enjoy in the warming weather, such as touring government buildings, visiting the Library of Congress, enjoying the blossomg trees at the National Japanese American Memorial. Some other popular destinations nearby are the National Air and Space Museum, the National Gallery of Art, and the U.S. Botanic Garden.

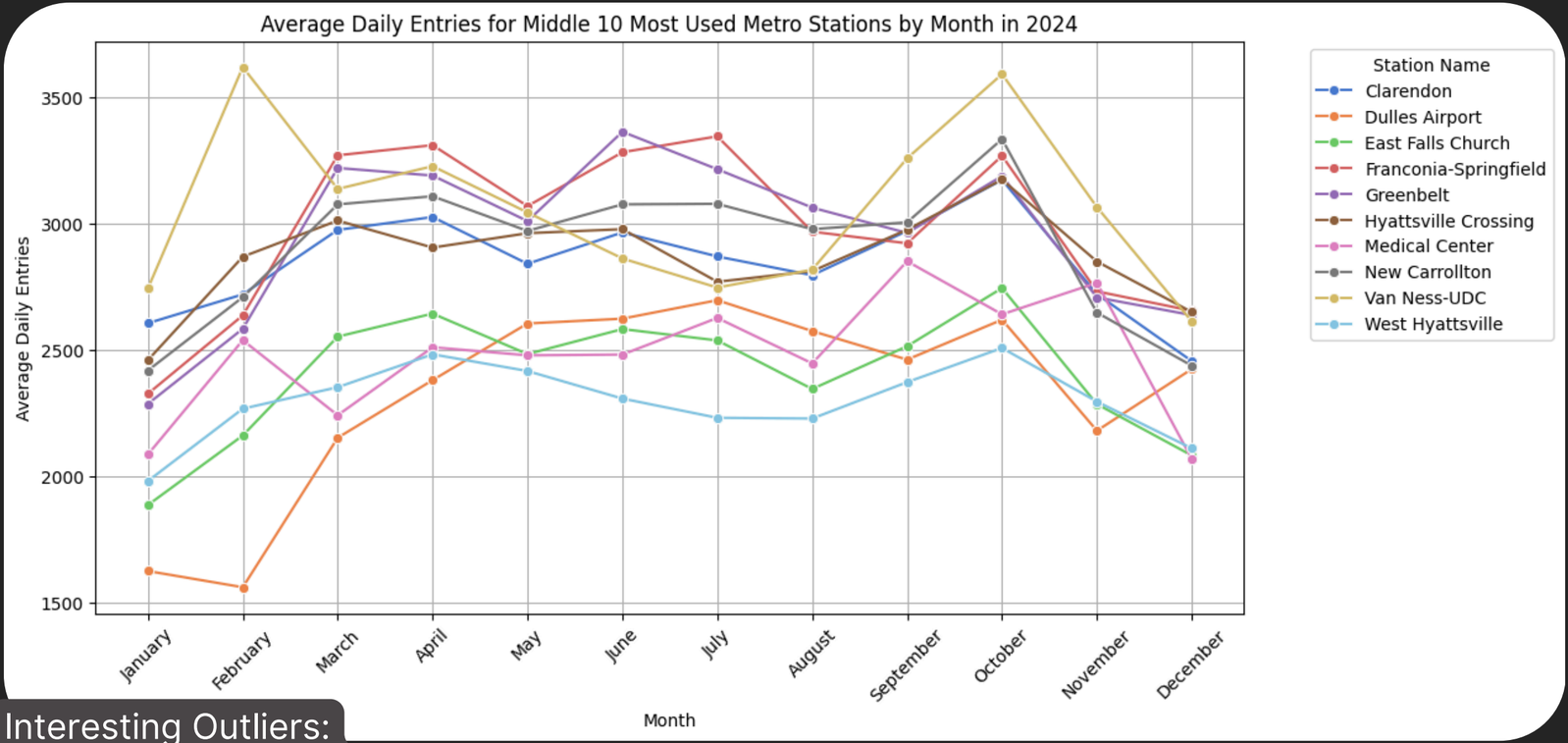
From May to August, the average daily entries are at the lowest levels for 9 out of 10 of the listed stations, which can be attributed to traveling out of state or out of the country by car or plane for vacation during the hot summer months.

In September, the average daily entries for 9 out of all of the 10 listed stationsdrastically increase and reach their peak in October. This may be attributed to popular events such as the Adams Morgan Day Festival, the DC State Fair, and the DC Jazz Fest in September, as well as the Oktoberfest, the Snallygaster, the Capital City Africa Cup, and the Fall Wine Festival in October. It is also the beginning of the Fall semester, so many commutes for college-going students begin this season, along with Halloween festivites.

In November and December, the average daily entres for all of the 10 listed stations decrease except for Gallery Place, which can be attributed to dropping temperatures discourage most metro riders, except for those interested in the Christmas festivities near Gallery Place.

Station-Specific Ridership:

What are the middle 10 busiest stations each month of the year?



The ten middle-most frequented Metro stations follow the same overall ridership pattern as the top ten busiest stations, but the peak is less apparent in March, and the summer plateau is less noticeable.

The October spike is still very visible, followed by the steep decrease in ridership into December.

Like the previous graph, the station trends tend to move with each other, but there are some discrepancies in February.

Interesting Outliers:

- Peak in Van Ness-UDC station in February

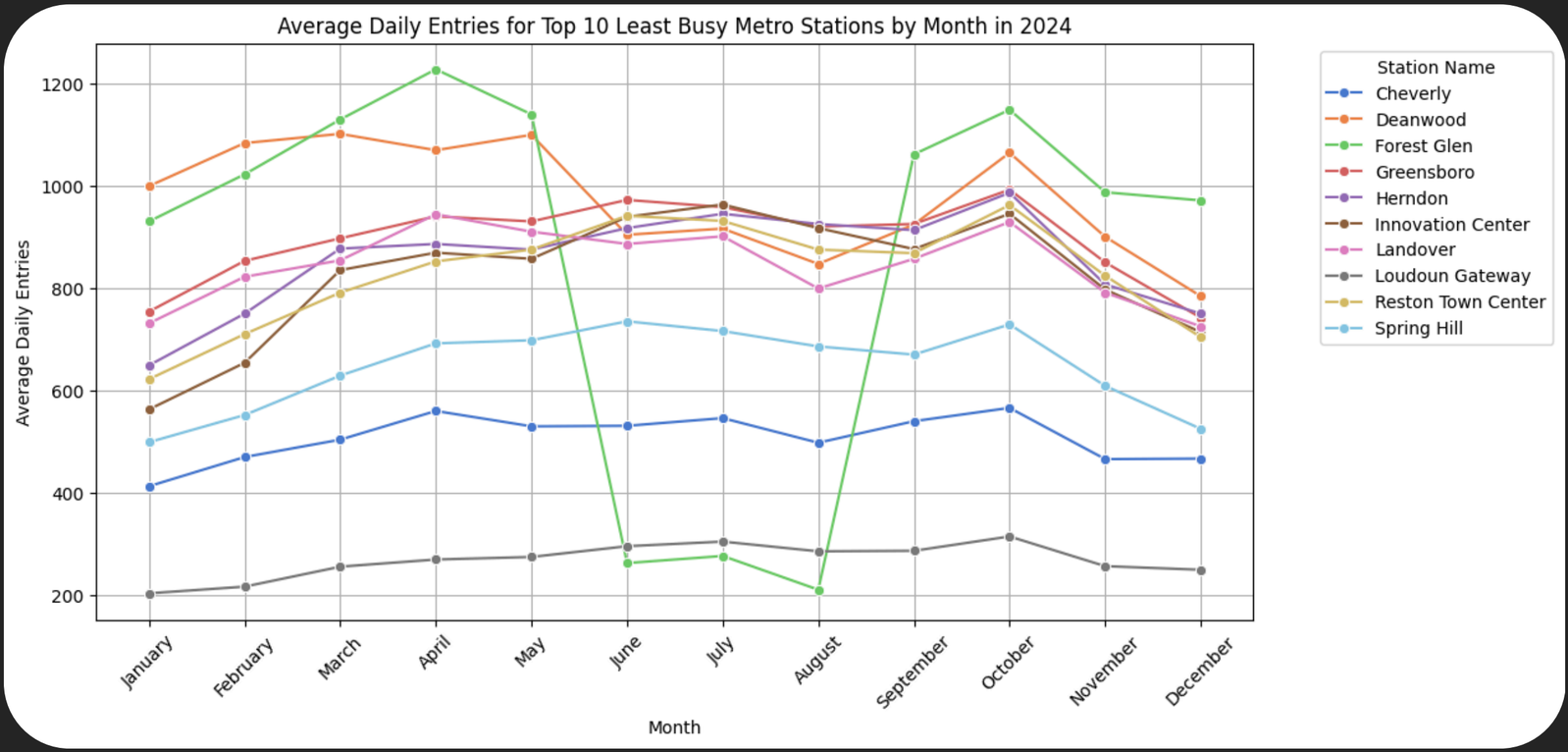
This could be attributed to UDC's Annual Holiday Concert.

- Dip in Dulles Airport in February

Travel around DC tends to be at its lowest during the winter season, and tourists likely hold off on traveling in February as the following month has the Cherry Blossoms as a major tourism event.

Station-Specific Ridership:

What are the least 10 busiest stations each month of the year?

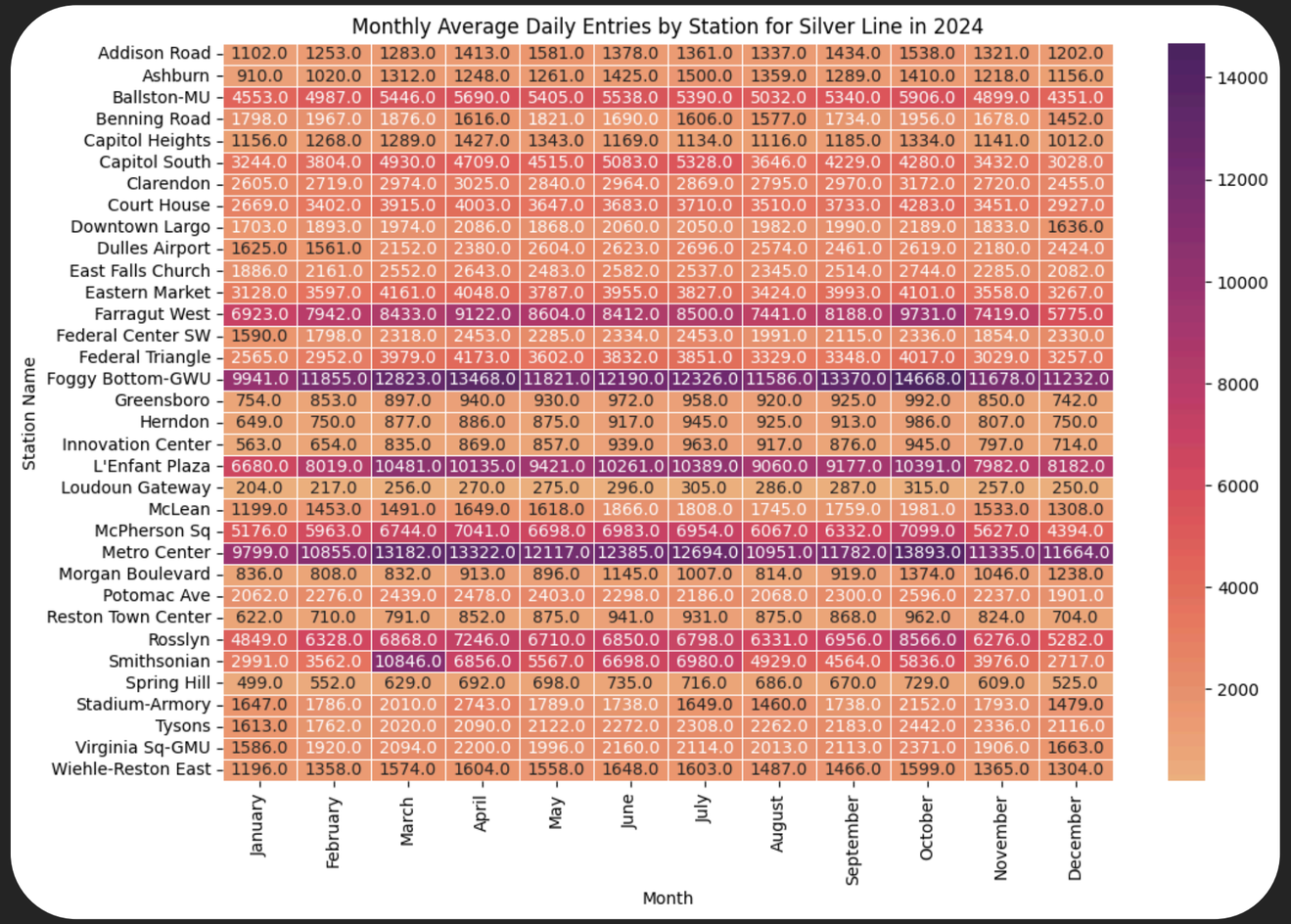


While the ten least frequented stations tend to follow the same trends as each other over time, the average daily entries appear to constantly plateau with small peaks in October and overall lower ridership in Winter months.

Note that the Y-axis scale is significantly lower in comparison to the previous two graphs, so discrepencies in ridership patterns are more sensitive in these lines.

Metro Line & Station Type Analysis:

How is ridership distributed across metro stations and lines over time?



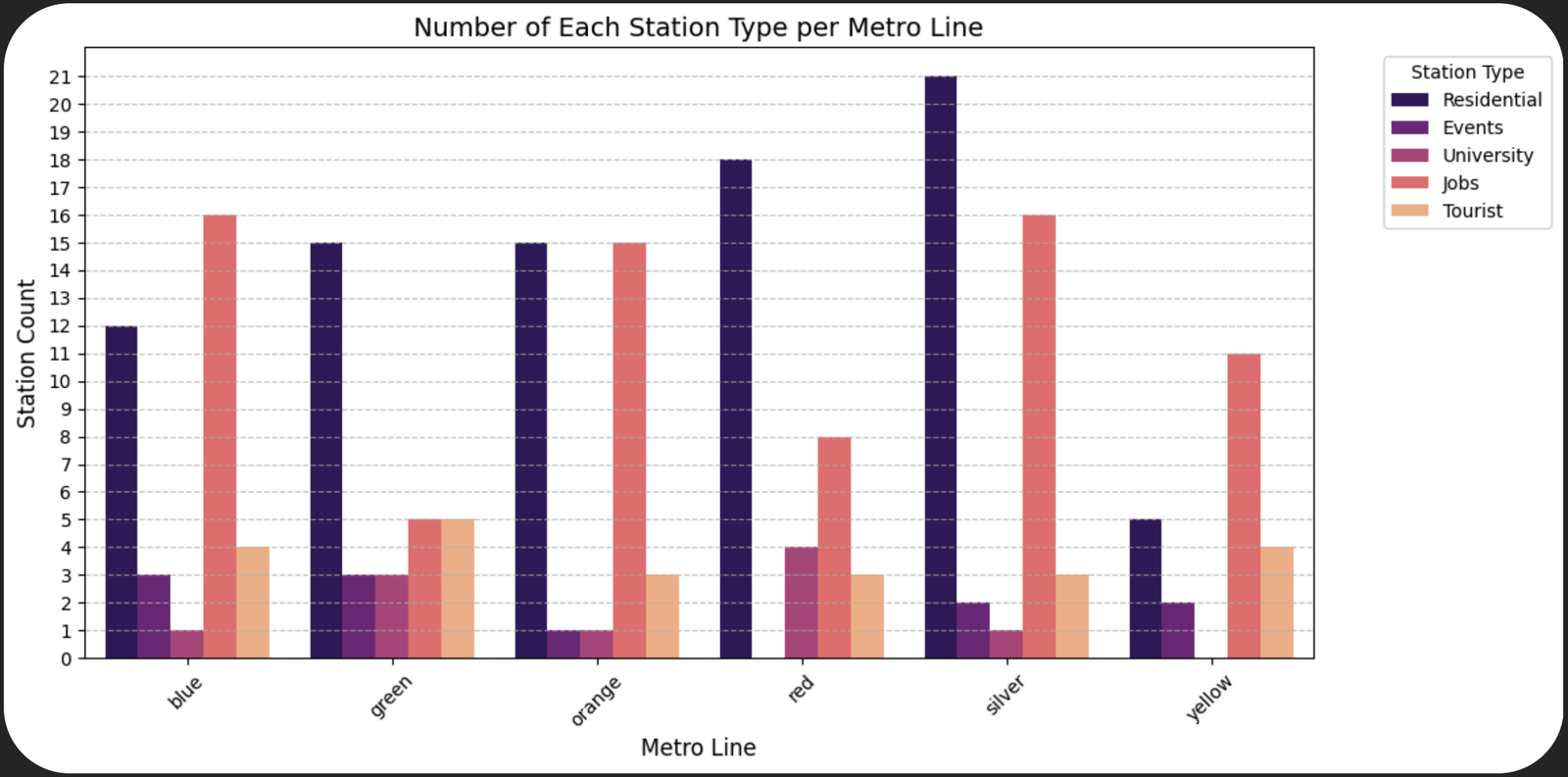
The heatmap measures the daily entries by stations across months in a certain line in order to show which months certain stations are most utilized in comparison to other stations on their line.

It also allows for easy identification of outliers in each station in which certain events or line closures might impact the overall ridership of a certain line during a particular month.

Outliers are significant as the previous visualizations show how overall metro trends throughout the months are generally not unique to a certain station. Such outliers prompted our idea to explore dynamic pricing of advertising in different stations in relation to these spikes by comparing the trends of specific stations to themselves.

Metro Line & Station Type Analysis:

How many stations of each type exist on each metro line?

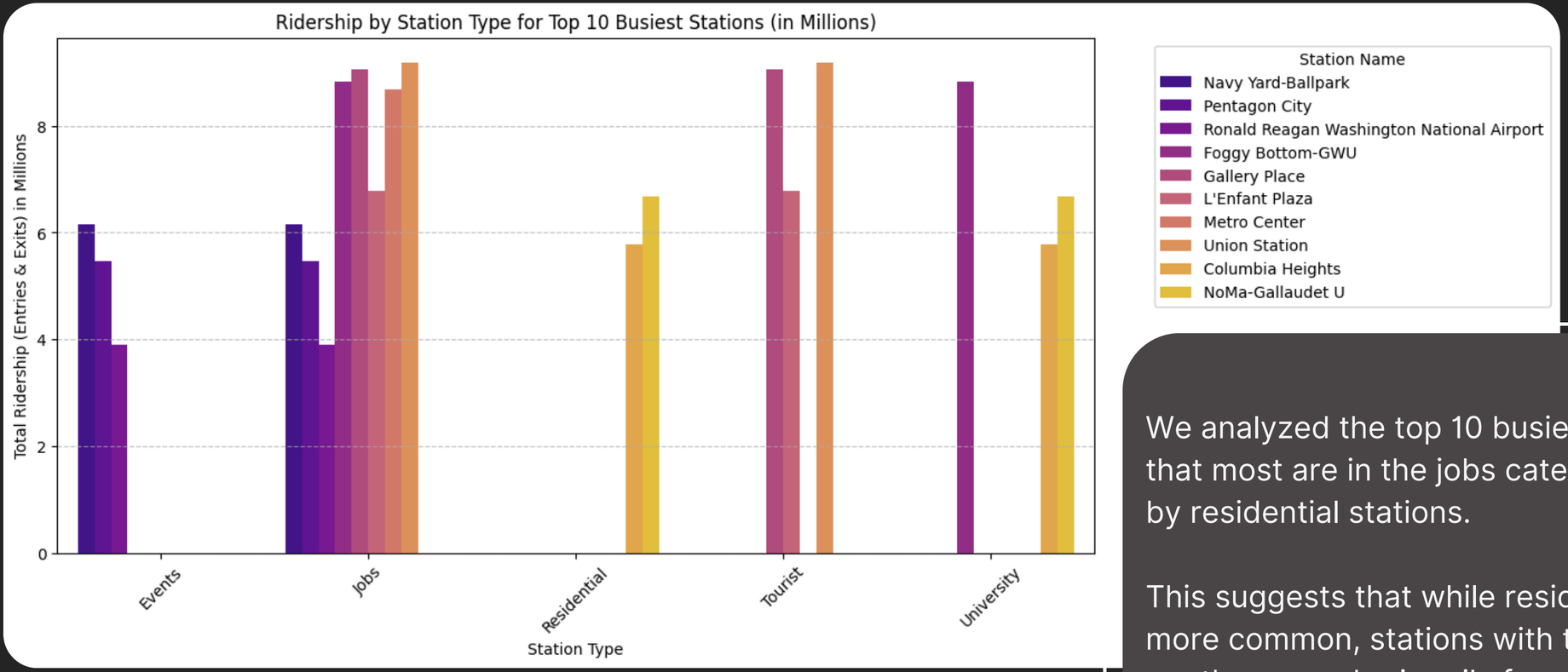


As each station in our analysis could be categorized as residential, events, university, jobs, tourist, with some stations fitting into two categories, we were trying to answer the question, which station types are most prominent.

As a result of our analysis, we can conclude that most stations could be categorized as residential, with the jobs category coming in second.

Metro Line & Station Type Analysis:

Which station types handle the most traffic?

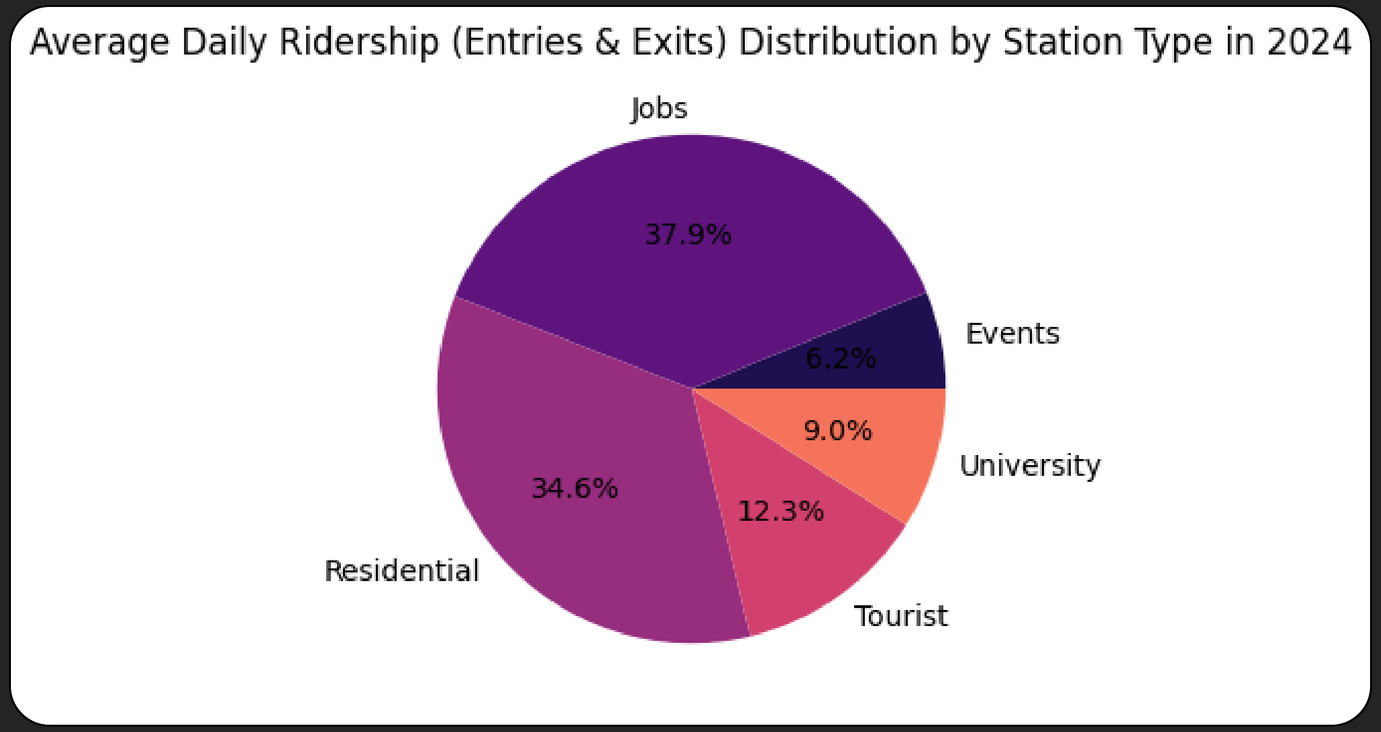


We analyzed the top 10 busiest stations and found that most are in the jobs category, followed closely by residential stations.

This suggests that while residential stations are more common, stations with the highest ridership are those used primarily for work-related commutes, and some stations belong to multiple categories.

Metro Line & Station Type Analysis:

Which station types handle the most traffic?



The pie chart also highlights the distribution of entries and exits between all station types, showcasing the almost equal distribution of residential and jobs categories, with the other three categories taking up a smaller percentage.

It can be assumed that the people who use residential metro stations are the same ones who later use jobs metro stations, making the distribution almost equal, accounting for people who perhaps work in the residential area or live in the “jobs” area.

Technological Tools

Google Sheets

Visual Studio Code

Jupyter Notebook

Python

Github

Future Implementations

- **Normalization of everyday_2024_w_metro_stations.csv**
  - Would allow us to create an SQL database for our data, and build queries and views to broaden the questions we can answer
  - Connect to Tableau to make dynamic data stories
  - Would allow us to utilize the X, Y, and Address data we have on stations with mapping sheets

- **Dynamic Pricing Script**
  - This script provides good ground for future research and implementation of dynamic advertisement pricing within metro lines.
  - Since we had limited data for this project, this script can be expanded by adding the data from the past years, not limited to 2024, to further improve the accuracy of the script.

- **Data Accuracy**
  - In the future, we would research the smaller lines in more depth to ensure proper assignment of station type.
  - In the future, we would also like to standardize the dataset columns (with data types, and pivoting vs unpivoting) across csv's
  - Utilizing geographical APIs on the events and frequented areas around the stations would allow us to understand the trends with more data-based backing and would allow us to make better predictions on the trends of 2025 if future DC event data is available.

INFO CHALLENGE 2025

Thank you

Team #:
IC25069

Team Members:
Maya Patel, Illia Polishchuck, Adrien Rozario