BTRY 4100 – Multivariate Analysis Final Project Plan

Andrew Chung, Ryan Gomez

March 24, 2025

1 Team Members

Andrew Chung (hc893), Ryan Gomez (rg673).

2 Problem Background

As the largest and busiest of its kind in the U.S. and the Western Hemisphere, the New York City subway system is unparalleled in its scale of service, comprising 36 lines (28 services), operating 472 stations, serving a city of 8 million, and boasting a daily ridership of 3.6 million passengers. Its vast scope of operations poses challenges in efforts to optimize transportation systems, mitigate delays, improve service efficiency/quality, and manage large-scale infrastructure; understanding passenger traffic and its relationship with systemic performance features is vital for sensible MTA policies that help maintain such an intricate transit system.

3 Research Question

Our final project seeks to identify and address **key predictors of station-wise ridership trends** in the **NYC Subway system during weekday peak hours**, specifically utilizing data collected from January-February 2025. In particular, with subway stations as data points, we seek to engineer a **regression model** to quatify the relationship between covariates that contain information on service-wise performance and station-specific metrics, potentially with the capacity to predict future trends.

4 Data Sources

All datasets (publicly available) were obtained from New York State's Open Data Portal, at `data.ny.gov`. (My data sets are up to date effective March 24, 2025, although my analysis will be confined to the months of January to February 2025 due to the lack of March data.) Seven (7) line-level datasets and an hourly ridership log of subway station complexes, compiled by the Metropolitan Transportation Authority (MTA), were rigorously pre-processed and integrated into (as of writing) two (2) data files: one containing service performance metric values (covariates) on a subway line basis, and the other a database of stations, respective lines served, and ridership numbers by the hour. I plan to bind the two datasets into a single data file for full analysis soon.

5 Head of the Dataset

See next page.

Final_Project_Head

Andrew Chung

2025-03-24

BTRY 4100 - Final Project Plan, Head of Datasets

Andrew Chung, hc893

Service Performance Metrics Dataset

```
line_data = read.csv("MTA_Subway_Line_Data_2025.csv")
head(line_data, 10)
```

```
##    line division num_passengers additional.platform.time additional.train.time
## 1     1        A       11835657                0.8519514             0.5591533
## 2     2        A        7686450                1.1443002             0.5599256
## 3     3        A        5991257                0.7963752             0.5513993
## 4     4        A        8645927                0.8821731             0.5303611
## 5     5        A        7388073                0.9987821             0.4962281
## 6     6        A       12414366                1.1352443             0.5931913
## 7     7        A       10437490                1.0755958             0.5421852
## 8    SG        A        1408022                0.3846406            -0.1867343
## 9     A        B        9342499                1.1466301             0.4169279
## 10    B        B        5887043                2.0705641             1.1342136
##    over_five_mins_perc wait.assessment service.delivered
## 1          0.086699270       0.7634315         0.9826627
## 2          0.137992605       0.6642381         0.9392052
## 3          0.102887618       0.7017032         0.9484935
## 4          0.120673805       0.6789162         0.9607604
## 5          0.124684875       0.6470091         0.9146056
## 6          0.115997280       0.7095128         0.9598140
## 7          0.096209730       0.6871326         0.9255027
## 8          0.001124943       0.9848287         0.9987327
## 9          0.136582320       0.6595975         0.9535108
## 10         0.231646440       0.5931124         0.8807690
##    terminal_on_time_performance Percent.Late infra_critical noninfra_critical
## 1                     0.8228496   0.03770504              4                 1
## 2                     0.7159579   0.04282938              4                 4
## 3                     0.8116235   0.03556113              3                 2
## 4                     0.8093127   0.03018528              8                 1
## 5                     0.7849674   0.03565719              2                 4
## 6                     0.8418102   0.03278275              1                11
## 7                     0.9200905   0.03970280              3                 2
```

```
## 8                    0.9990777   0.00000000             0                 0
## 9                    0.8179950   0.03325508             1                 1
## 10                   0.6384484   0.05679924             2                 5
##    infra_noncritical noninfra_noncritical
## 1               2282                 1049
## 2               2187                 1570
## 3               1279                  992
## 4               1645                 1259
## 5               1407                 1476
## 6               1834                 1775
## 7               1448                  627
## 8                  3                   16
## 9               2149                  718
## 10              2303                  894
```

Hourly Ridership Summarized Dataset

```
hourly_ridership = read.csv("MTA_Subway_Ridership_Summarized_2025.csv")
head(hourly_ridership, 10)
```

```
##      transit_timestamp         station_complex   borough ridership lines
## 1  2025-01-01 07:00:00               1 Av (L) Manhattan       104     L
## 2  2025-01-01 07:00:00             103 St (1) Manhattan        90     1
## 3  2025-01-01 07:00:00             103 St (6) Manhattan       138     6
## 4  2025-01-01 07:00:00           103 St (C,B) Manhattan        27   C,B
## 5  2025-01-01 07:00:00 103 St-Corona Plaza (7)    Queens       455     7
## 6  2025-01-01 07:00:00             104 St (A)    Queens        32     A
## 7  2025-01-01 07:00:00           104 St (J,Z)    Queens        46   J,Z
## 8  2025-01-01 07:00:00             110 St (6) Manhattan       111     6
## 9  2025-01-01 07:00:00             111 St (7)    Queens       208     7
## 10 2025-01-01 07:00:00             111 St (A)    Queens        38     A
```