A BETTER CITY: COMMUTER RAIL ANALYSIS

FINAL PRESENTATION

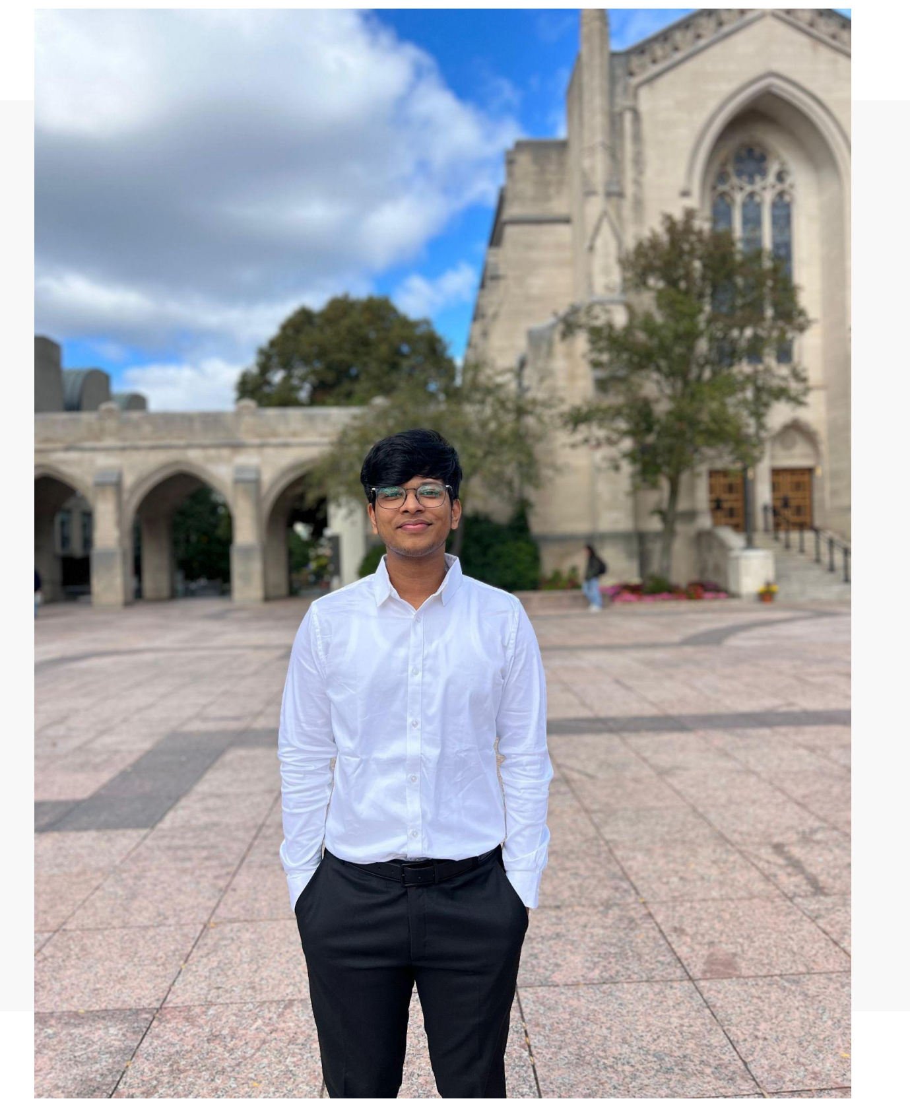
DEC. 5

By:
Bhuvan Gowda (bhuvansg@bu.edu)
Neel Gangrade (neelg@bu.edu)
Ningxin Guan (nxguan@bu.edu)
Sumanth Kamath (sumanth@bu.edu)
Yuanchen Yin (yinthyg@bu.edu)

Team Members

Bhuvan Gowda
team member
Data Science
1st Year Master

Yuanchen Yin
team member
Data Science
1st Year Master



Neel Gangrade
Team Member
Data Science
1st Year Master

Sumanth
Team Member
Data Science
1st Year Masters

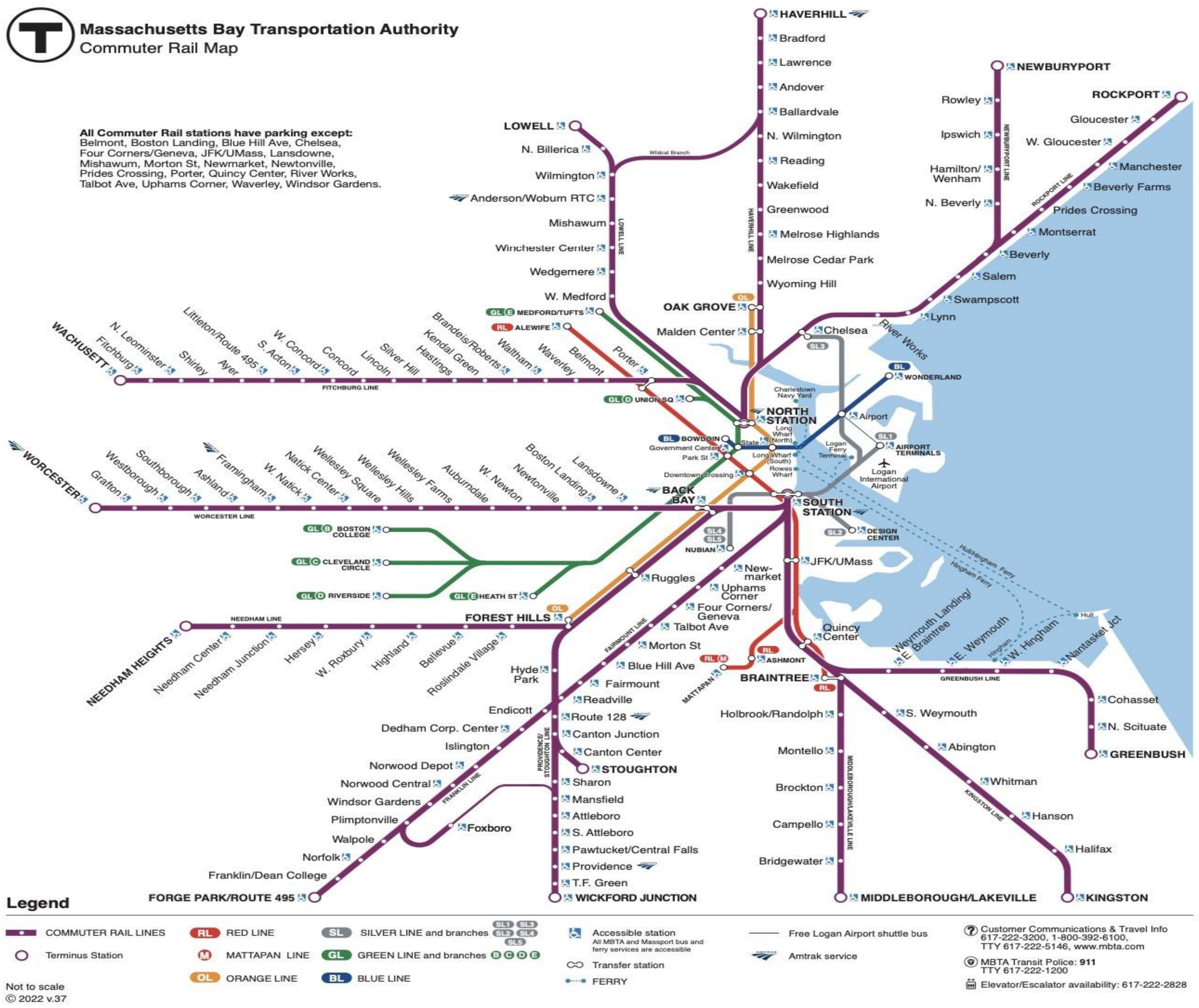
Ningxin Guan
team member
Data Science
1st Year Master

INTRODUCTION

- **Current Challenges of Boston Commuter Rail:**
  - Low Service Frequency
  - Low On-Time Performance

- **Project Goal:** Create data repository and analysis framework for MBTA schedules
- Support policy-making and decarbonization strategies by providing insights into transit schedules, timing, and operational changes over time.

Massachusetts Bay Transportation Authority



CLIENT INFORMATION

A Better City represents a multi-sector group of nearly 130 business leaders united around a common goal: To enhance the Greater Boston region's economic health, competitiveness, equitable growth, sustainability, and quality of life for all communities.

Goal: Developing solutions and influencing policy in three critical areas:

1. transportation and infrastructure
2. land use and development, and
3. energy and the environment.

DATA

Extracted the latest updated data from the MBTA archive_gtfs for all seasons and years.
Url: cdn.mbta.com/archive/archived_feeds.txt

Main tables:

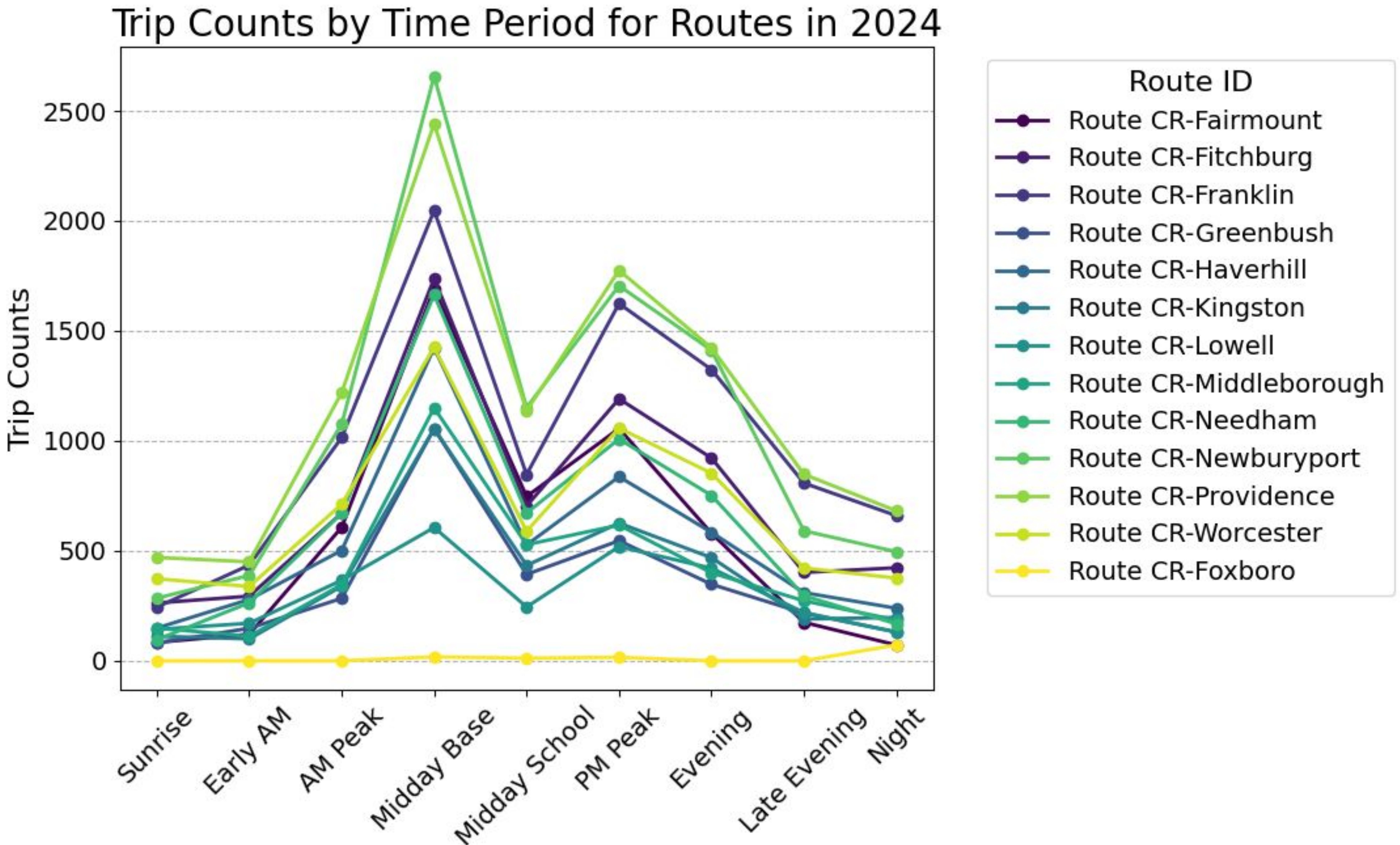
- **routes.txt**
- **lines.txt**
- **directions.txt**
- **calendar.txt**
- **trips.txt**
- **stops.txt**
- **stop_times.txt**

Cleaning Process:

- focus on **commuter rail** information: filtering rows where the *route_desc* column equals "*commuter_rail*".
- Remove null values.
- Remove duplicate values.

QUESTION 1

Breakdown in Number of Trips by day of the week

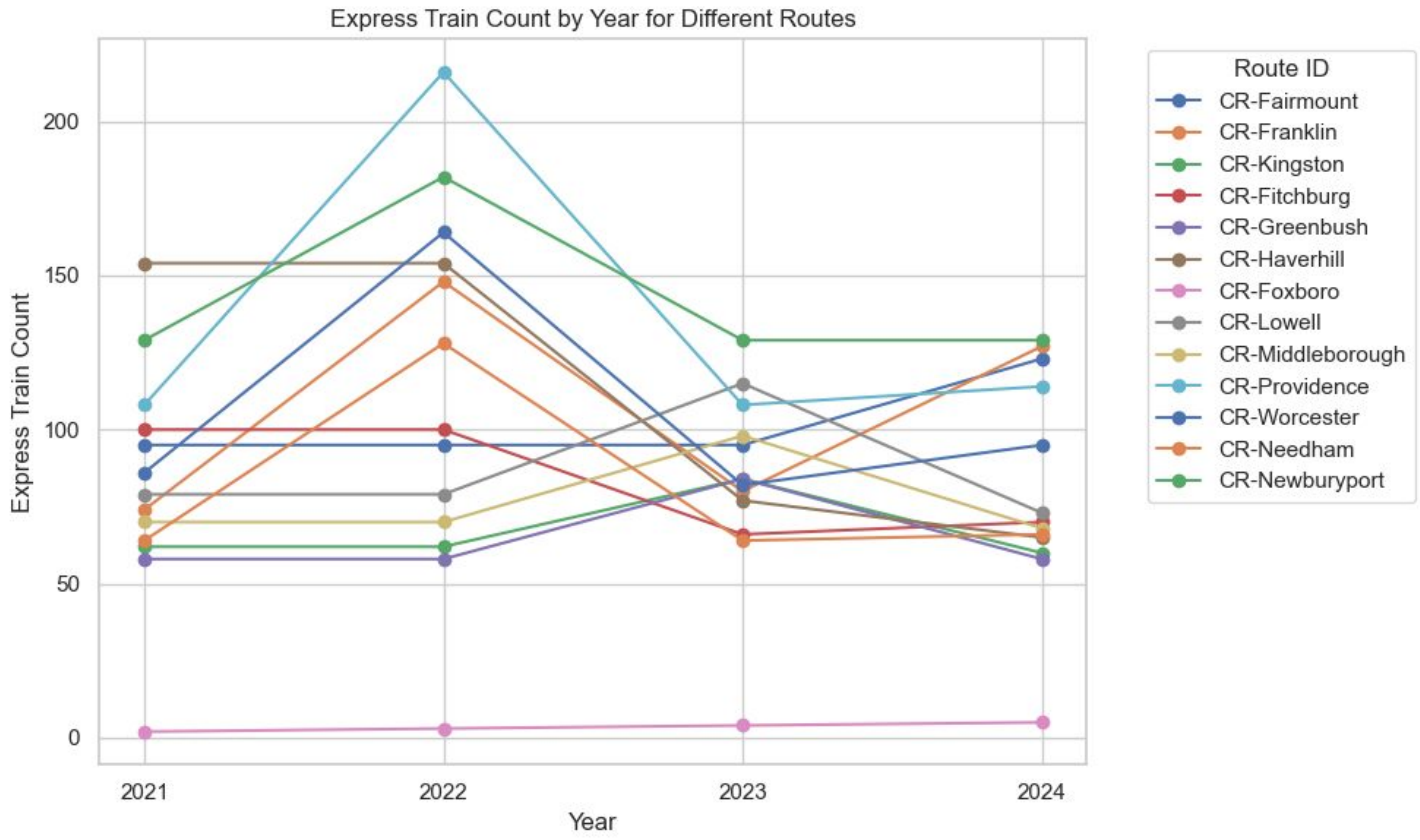


Breakdown in Number of Trips by time of day

x-axis: time period
y-axis: trip count

QUESTION 2

Number of Express Train on each line

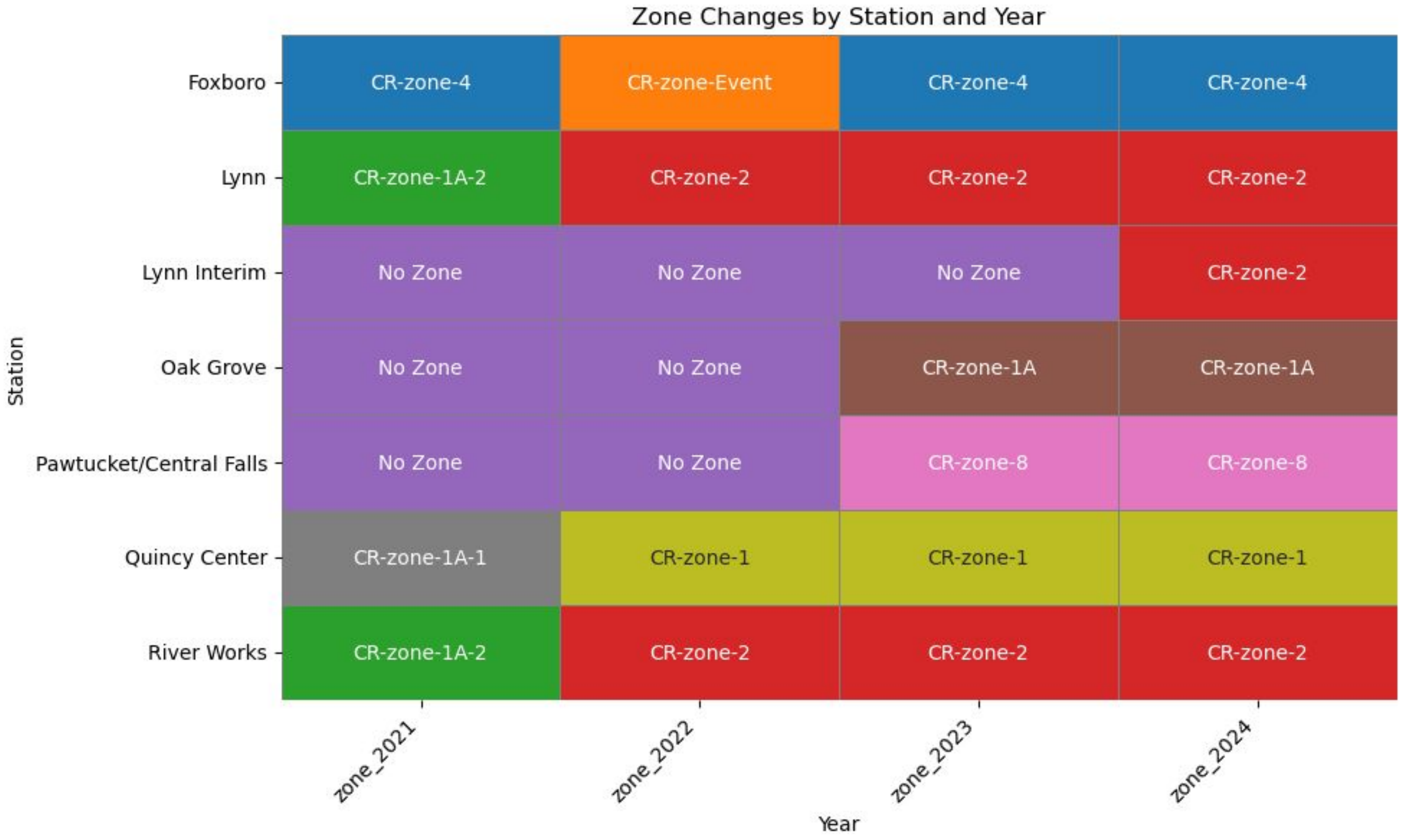


Express Train: number of stops for a trip is less than the maximum number of stops for that route

- No data for before 2021

QUESTION 3

Do we see incidents of a station changing fare zones?



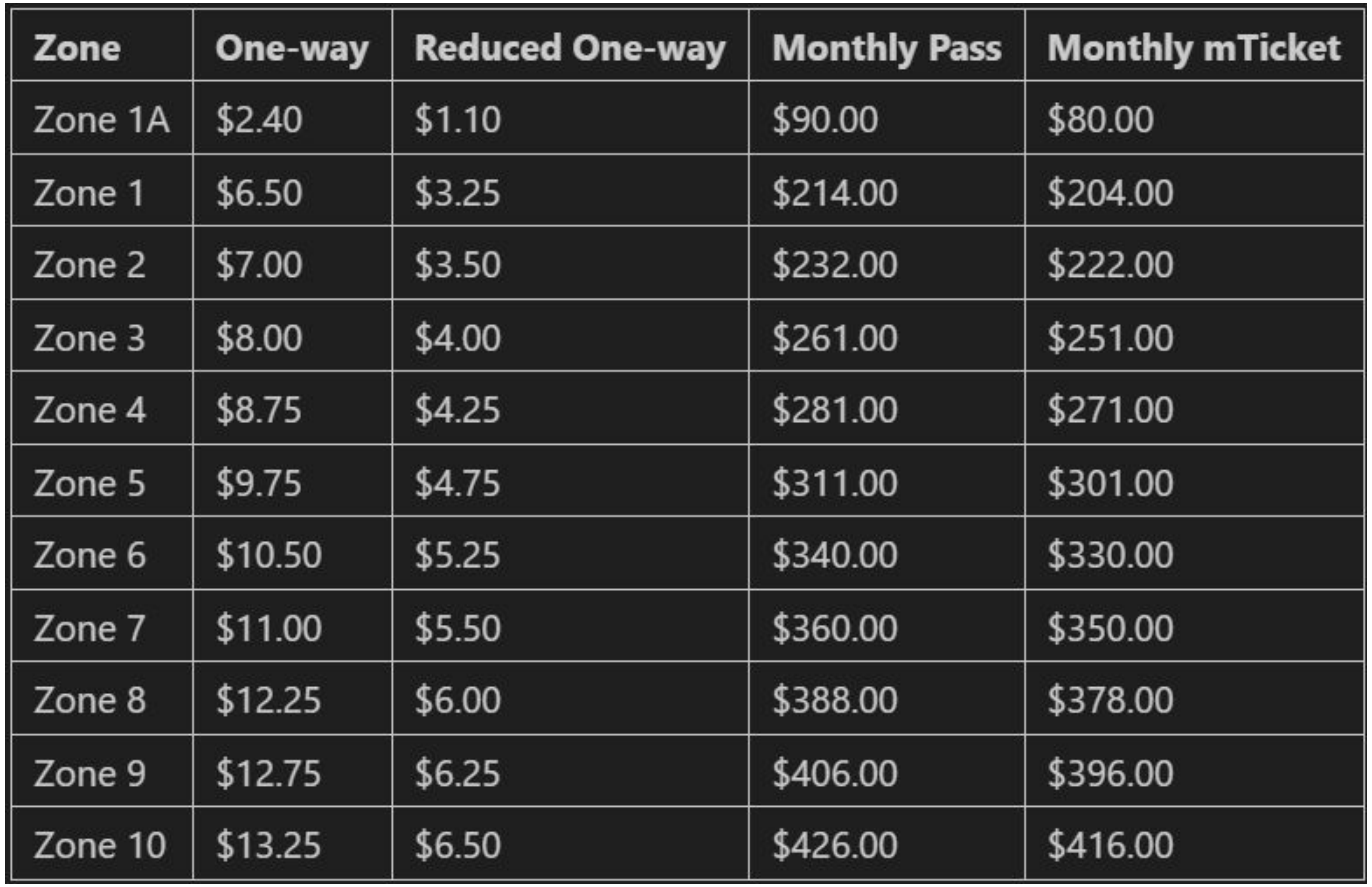
Zone Changes by Station and Year

Station	zone_2021	zone_2022	zone_2023	zone_2024
Foxboro	CR-zone-4	CR-zone-Event	CR-zone-4	CR-zone-4
Lynn	CR-zone-1A-2	CR-zone-2	CR-zone-2	CR-zone-2
Lynn Interim	No Zone	No Zone	No Zone	CR-zone-2
Oak Grove	No Zone	No Zone	CR-zone-1A	CR-zone-1A
Pawtucket/Central Falls	No Zone	No Zone	CR-zone-8	CR-zone-8
Quincy Center	CR-zone-1A-1	CR-zone-1	CR-zone-1	CR-zone-1
River Works	CR-zone-1A-2	CR-zone-2	CR-zone-2	CR-zone-2

The CR-zone-Event is a special fare zone which is only introduced at the times of major events in Boston, for example the Taylor Swift concert at the Gillette Stadium.

QUESTION 3.1

Do we see incidents of a station changing fare zones?

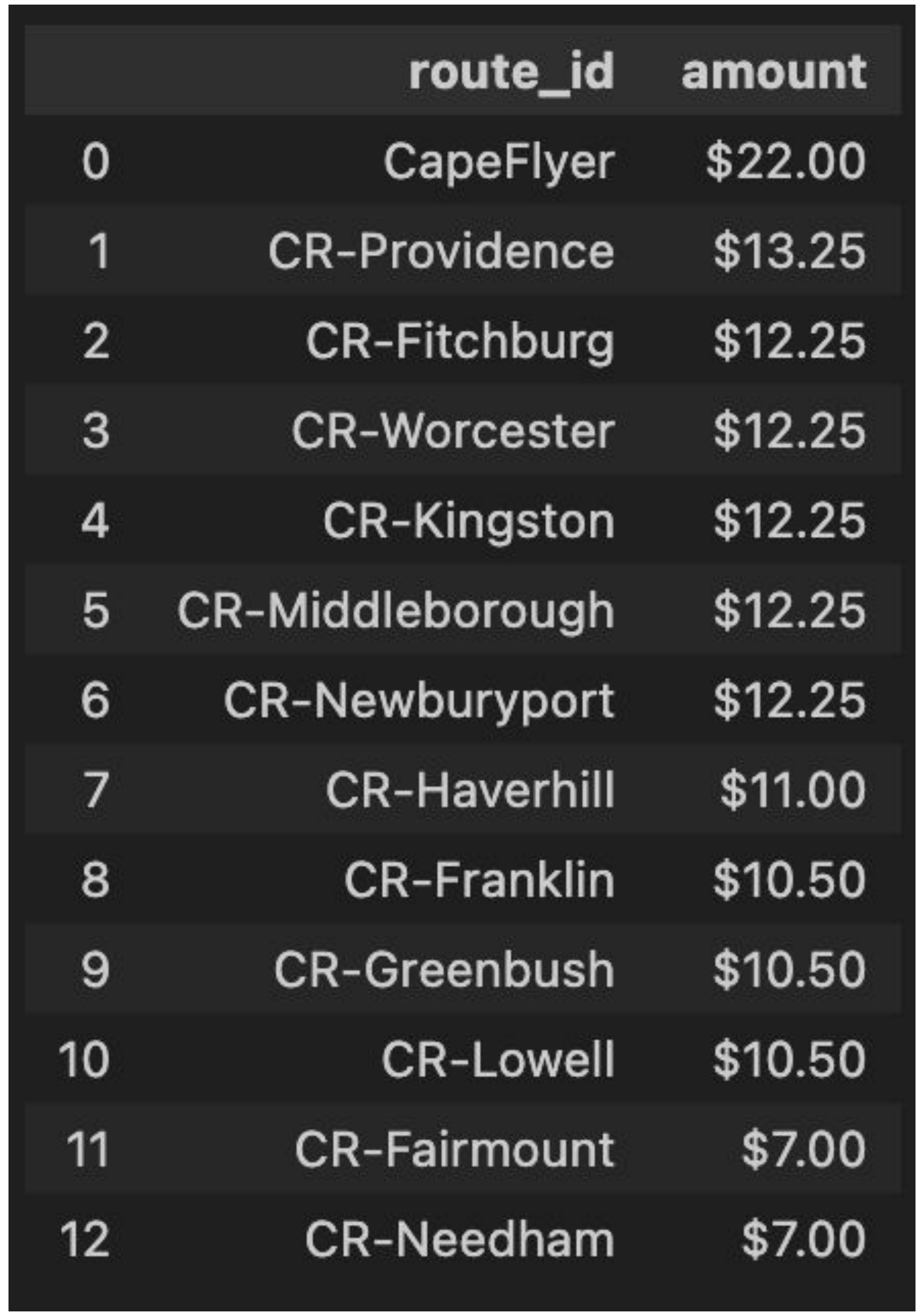


Zone	One-way	Reduced One-way	Monthly Pass	Monthly mTicket
Zone 1A	$2.40	$1.10	$90.00	$80.00
Zone 1	$6.50	$3.25	$214.00	$204.00
Zone 2	$7.00	$3.50	$232.00	$222.00
Zone 3	$8.00	$4.00	$261.00	$251.00
Zone 4	$8.75	$4.25	$281.00	$271.00
Zone 5	$9.75	$4.75	$311.00	$301.00
Zone 6	$10.50	$5.25	$340.00	$330.00
Zone 7	$11.00	$5.50	$360.00	$350.00
Zone 8	$12.25	$6.00	$388.00	$378.00
Zone 9	$12.75	$6.25	$406.00	$396.00
Zone 10	$13.25	$6.50	$426.00	$416.00

If your trip begins or ends at a station located in Zone 1A, your fare is based on the location of the other station you are traveling to or from.

QUESTION 4

Can we compare the changes in fare cost over time?



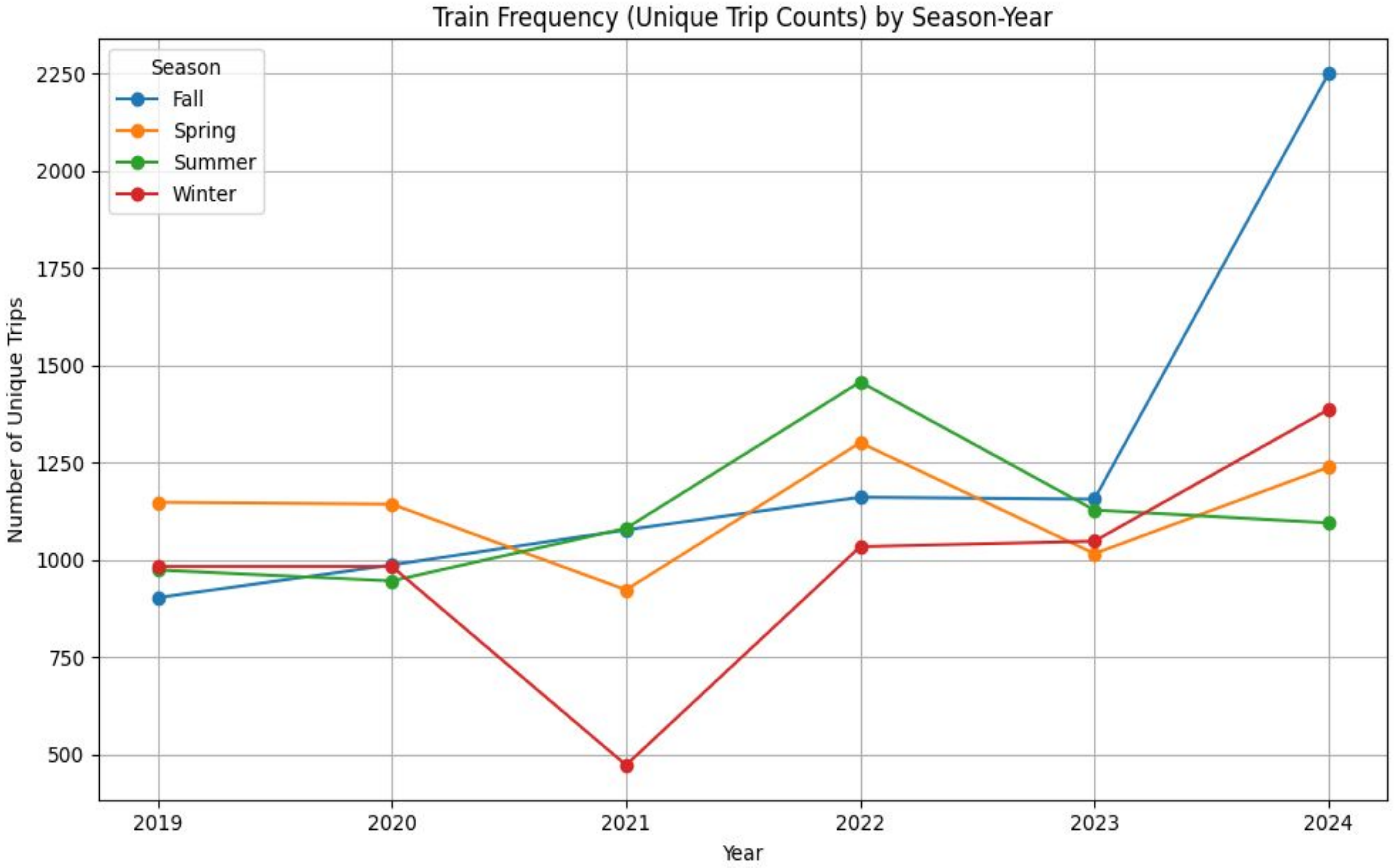
	route_id	amount
0	CapeFlyer	$22.00
1	CR-Providence	$13.25
2	CR-Fitchburg	$12.25
3	CR-Worcester	$12.25
4	CR-Kingston	$12.25
5	CR-Middleborough	$12.25
6	CR-Newburyport	$12.25
7	CR-Haverhill	$11.00
8	CR-Franklin	$10.50
9	CR-Greenbush	$10.50
10	CR-Lowell	$10.50
11	CR-Fairmount	$7.00
12	CR-Needham	$7.00

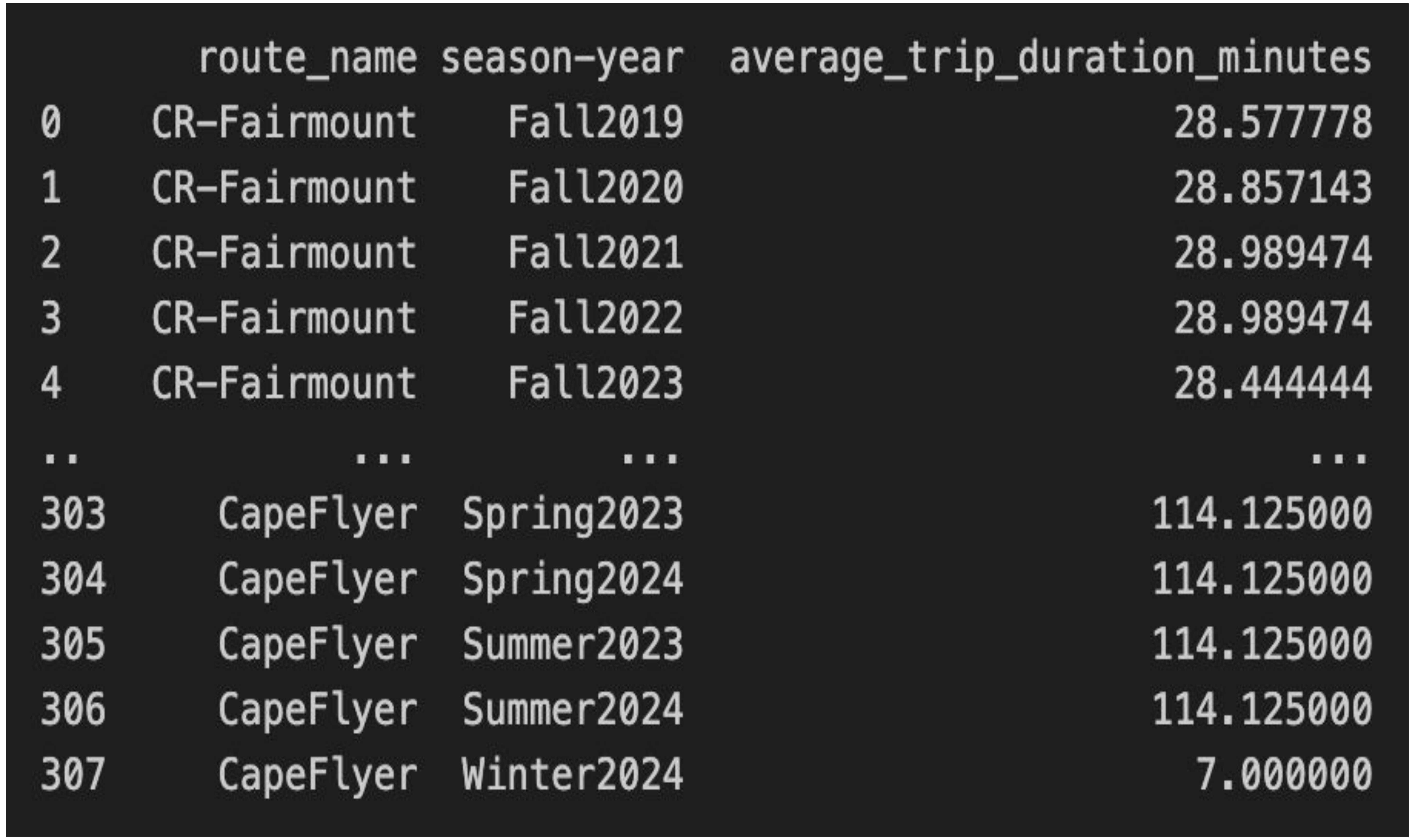
Since we have data for fares for 2 years (i.e, 2023 and 2024), we do not see any change in fares over the time.

NOTE - CapeFLYER is a seasonal train route that runs between South Station and Hyannis from Memorial Day to Labor Day.

QUESTION 5/7

How have schedules shifted over time? How do travel times vary across different schedules?

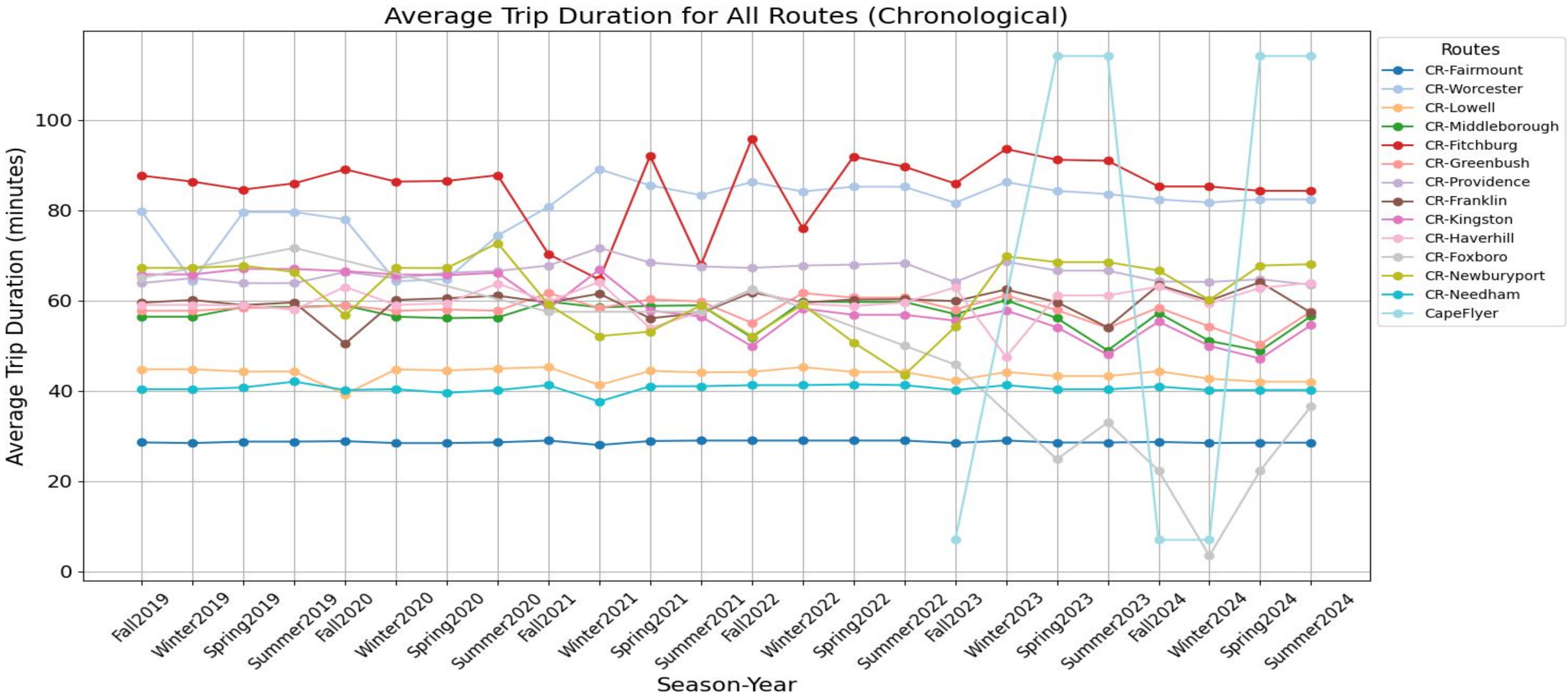




```
       route_name season-year  average_trip_duration_minutes
0    CR-Fairmount    Fall2019                      28.577778
1    CR-Fairmount    Fall2020                      28.857143
2    CR-Fairmount    Fall2021                      28.989474
3    CR-Fairmount    Fall2022                      28.989474
4    CR-Fairmount    Fall2023                      28.444444
..            ...         ...                            ...
303     CapeFlyer  Spring2023                     114.125000
304     CapeFlyer  Spring2024                     114.125000
305     CapeFlyer  Summer2023                     114.125000
306     CapeFlyer  Summer2024                     114.125000
307     CapeFlyer  Winter2024                       7.000000
```

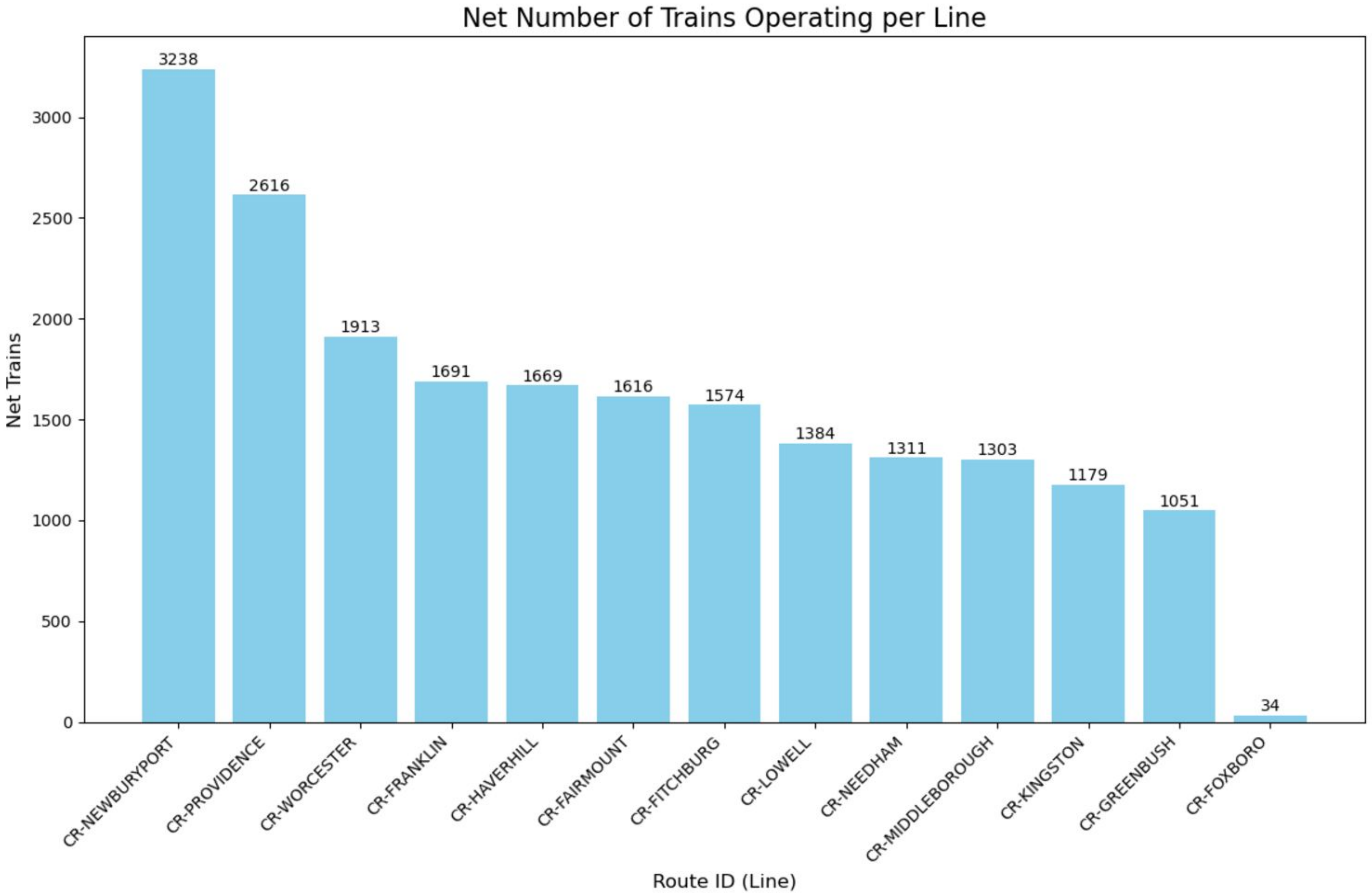
QUESTION 5/7

How have schedules shifted over time? How do travel times vary across different schedules?



QUESTION 6

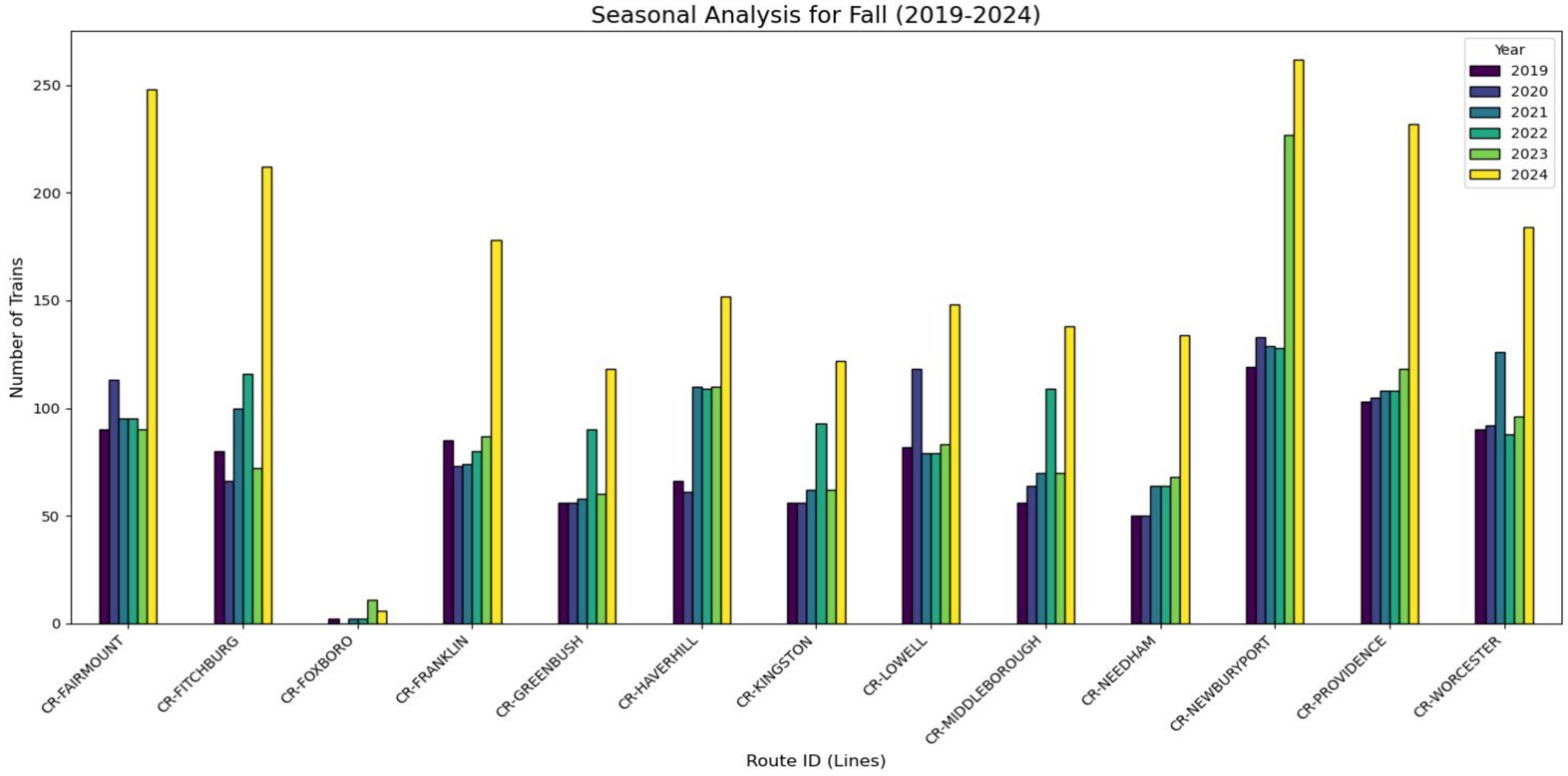
What is the net number of trains operating per line?



The CR-Newburyport line has the most trains, with 3,238 trips. The CR-Providence line follows, operating 2,616 trains. The CR-Worcester line has 1,913 trains, while the CR-Franklin line has 1,691. These numbers highlight their importance in the network.

In contrast, the CR-Foxboro line has only 34 trips. This is because it is a special line that operates exclusively during events.

QUESTION 6.1



Most lines show a steady increase in train numbers over the years. The highest values are consistently in 2024.

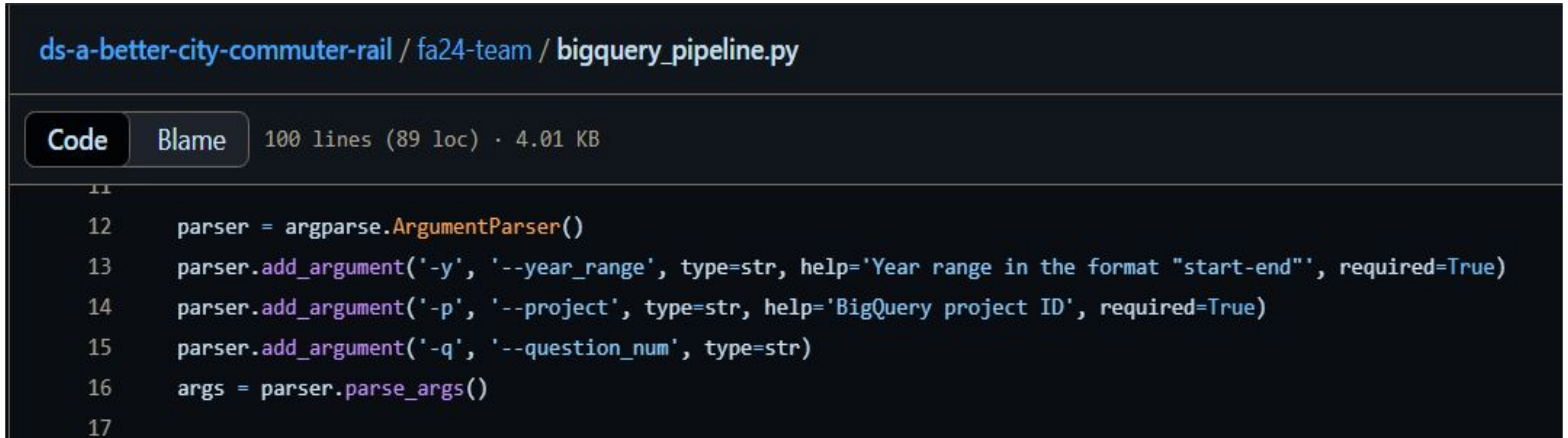
The CR-Newburyport and CR-Providence lines have the most growth, reflecting increased demand or service expansion.

The CR-Foxboro line remains low, as it only operates during events.

The seasonal graph shows an increase in train numbers during fall for most lines, especially in 2024. Declines in certain years, like 2020 and 2021. These changes likely result from evolving demand and adjustments in service under the influence of COVID-19.

Pipeline to Upload Raw Data to BigQuery

- The *bigquery_pipeline.py* script can be used to upload all the raw data for each season year from the MBTA archives to BigQuery.
- *python bigquery_pipeline.py --year_range 2019-2024 --project ds-better-city-commuter*
- *--year_range* is used to specify the range of years for which you want the data to be uploaded. *--project* specifies the project ID of your project on Google Cloud Platform.



ds-a-better-city-commuter-rail / fa24-team / bigquery_pipeline.py

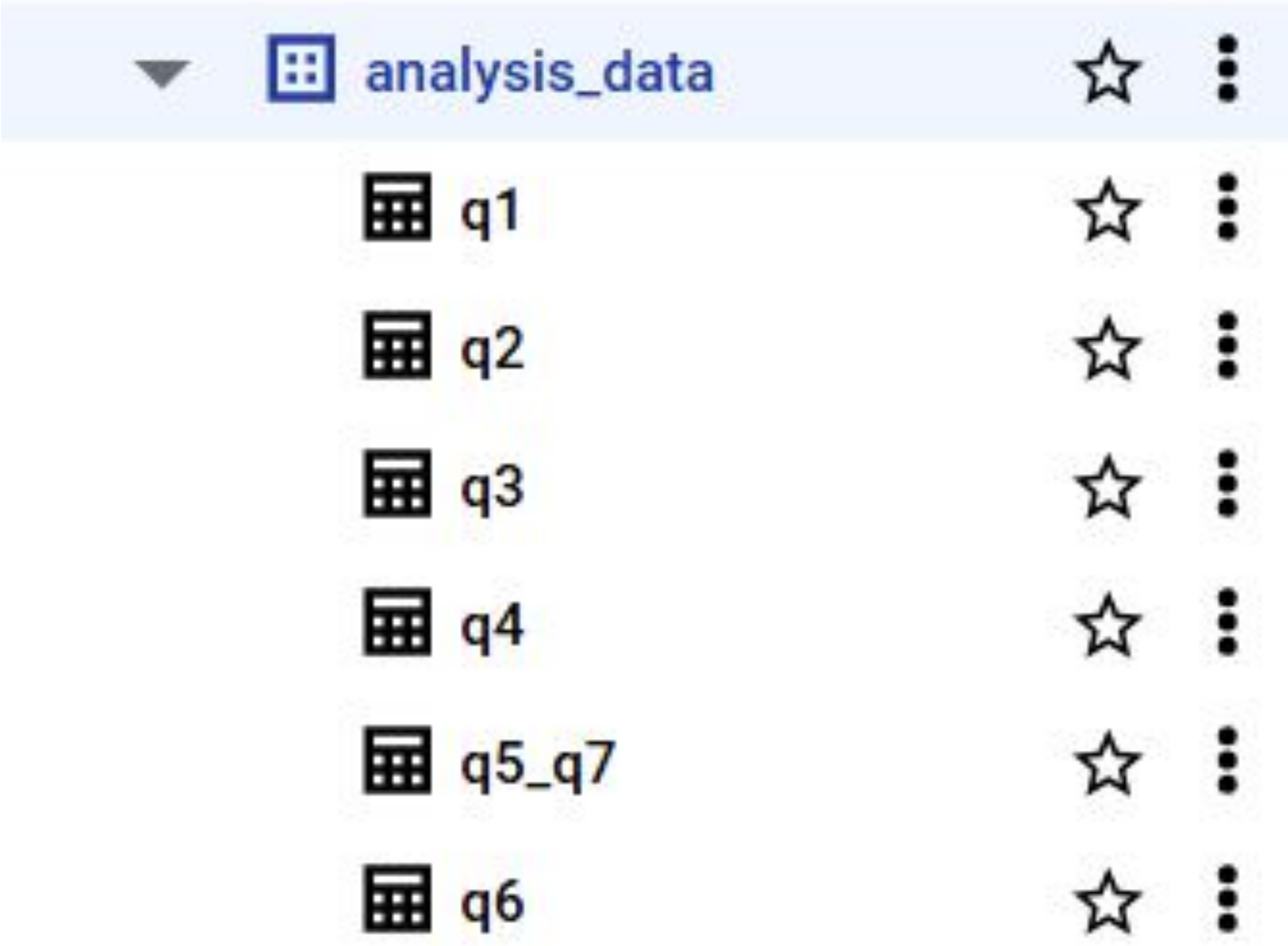
100 lines (89 loc) · 4.01 KB

```
parser = argparse.ArgumentParser()
parser.add_argument('-y', '--year_range', type=str, help='Year range in the format "start-end"', required=True)
parser.add_argument('-p', '--project', type=str, help='BigQuery project ID', required=True)
parser.add_argument('-q', '--question_num', type=str)
args = parser.parse_args()
```

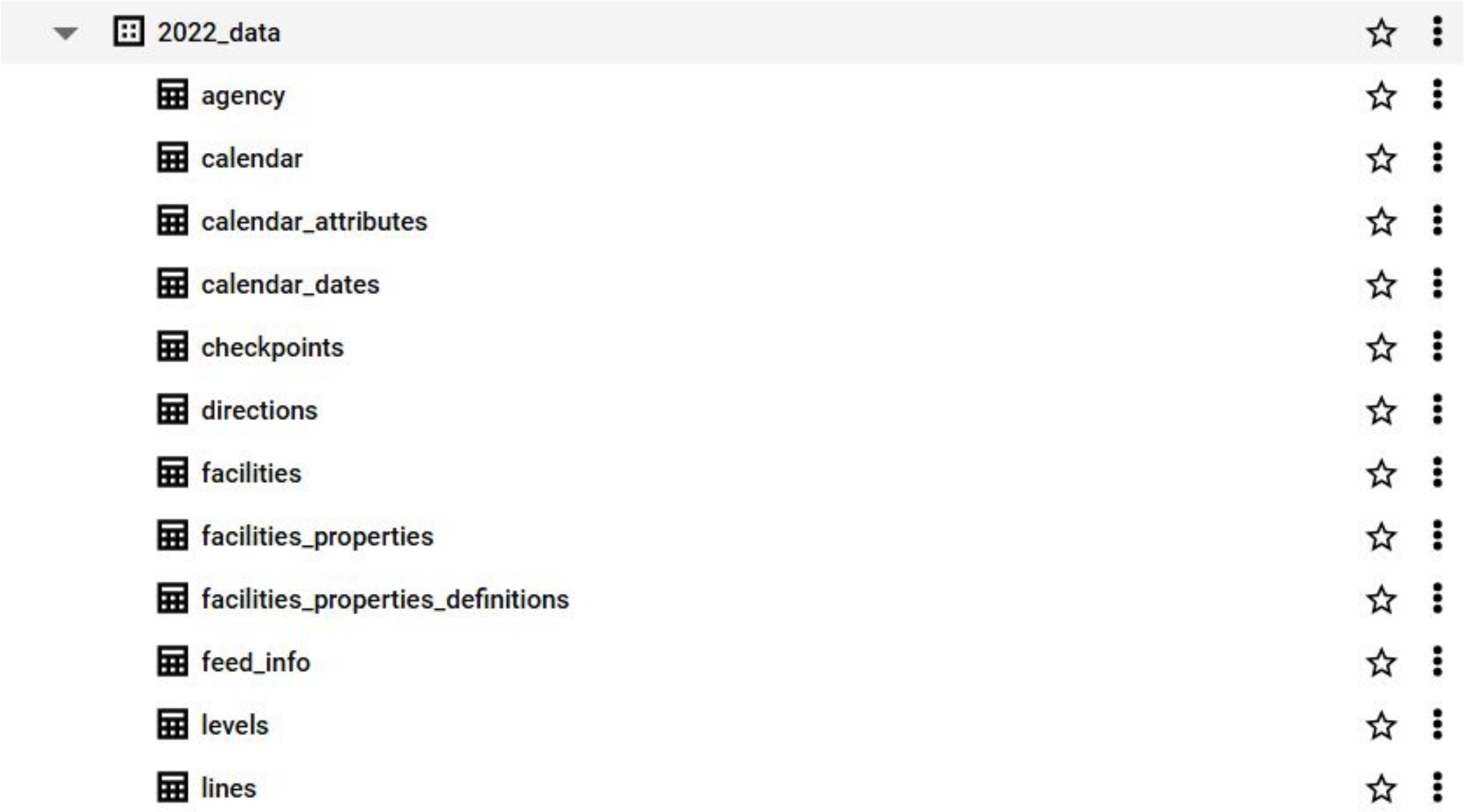
Pipeline to Upload the Answers for Key Questions on BigQuery

- The ***bigquery_cleaned_pipeline.py*** script is used to upload the dataset corresponding to a specific base question to BigQuery under the analysis_data Dataset.
- *python3 bigquery_cleaned_pipeline.py --project ds-better-city-commuter --question_num q4*
- The possible values of the argument --question_num are 'q1', 'q2', 'q3', 'q4', 'q5', 'q6', 'q7', 'q8'.
- Each question mentioned above has a table inside the *analysis_data* dataset.

Data on BigQuery



analysis_data
- q1
- q2
- q3
- q4
- q5_q7
- q6



2022_data
- agency
- calendar
- calendar_attributes
- calendar_dates
- checkpoints
- directions
- facilities
- facilities_properties
- facilities_properties_definitions
- feed_info
- levels
- lines

CONCLUSION & FUTURE STEP

Conclusion:

- Developing scripts to automatically fetch data.
- Visualization of MBTA schedule data.
- BigQuery pipeline for non-technical client to access data easily.
- Detailed tutorial and accompanying transition document detailing how others may self-serve data for future analyses.

Next Step:

- Further analyze the relationship between trip counts and ridership demand to support decarbonization

Thank you!

SPARK!