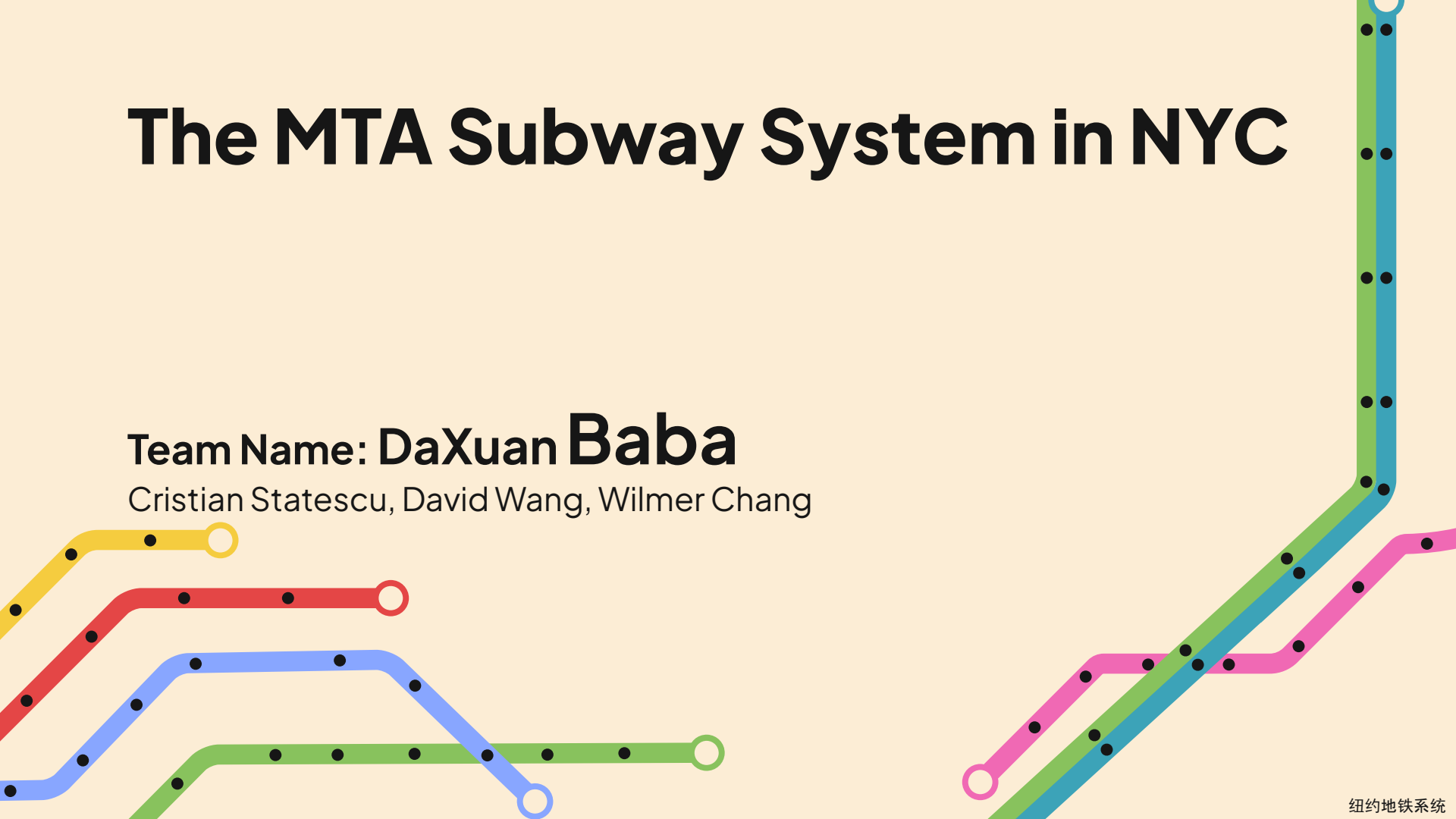
The MTA Subway System in NYC

Team Name: DaXuan Baba

Cristian Statescu, David Wang, Wilmer Chang

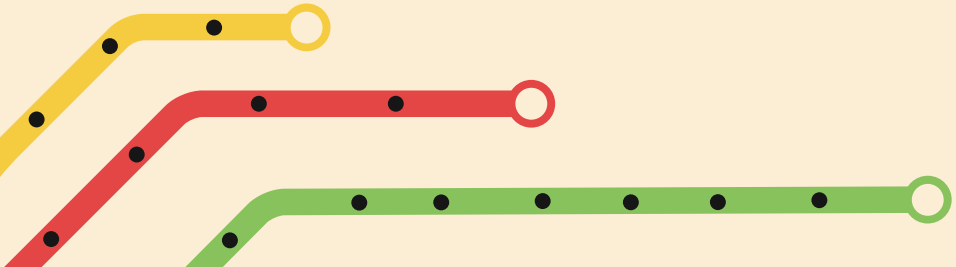


纽约地铁系统

Course Instructor

Xuan Wang

Table of contents



01

About the Project

Project Motivation

- Team members, along with millions of New Yorkers, uses the subway as the main source of transportation.
- Interests in a better future for public transit in New York City.

Problem Statement

The MTA subway system provides service to over 8 million New Yorkers with its 472 stations, each in a different state of maintenance and has a different capacity. **The problem is that the safety and comfort of commuters in NYC has been declining.** With subway crime rising up 13% from 2023 and, specifically, assaults rising up 11% on the transit system alone, the MTA is in dire need of safety improvements.

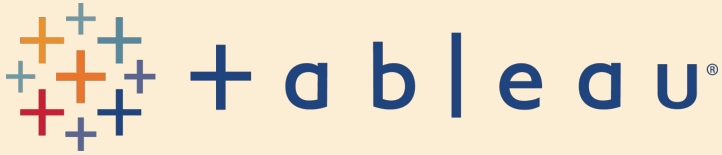
What is our data?

MTA Open Data
- MTA Subway Hourly Ridership Beginning February 2022
- MTA Turnstile Usage Data

NYPD
- Subway Transit Districts
- Subway Fare Evasion Summon and Arrests

General Research
- Information from various other city, state, government and news/specific transit websites

What tools did we use?



Observable

02
Data Exploration

MTA Subway Hourly Ridership (Beginning February 2022)

Data Label	Data Type	Data Description
transit_timestamp	DATE	Timestamp payment took place in local time. All transactions here are rounded **down** to the nearest hour. For example, a swipe that took place at 1:37pm will be reported as having taken place at 1pm.
transit_mode	TEXT	Distinguishes between the subway, Staten Island Railway, and the Roosevelt Island Tram
station_complex_id	ALPHANUMERIC	A unique identifier for station complexes
station_complex	TEXT	The subway complex where an entry swipe or tap took place. Large subway complexes, such as Times Square and Fulton Center, may contain multiple subway lines. The subway complex name includes the routes that stop at the complex in parenthesis, such as Zerega Av (6).
borough	TEXT	Represents one of the boroughs of New York City serviced by the subway system (Bronx, Brooklyn, Manhattan, Queens).
payment_method	TEXT	Specifies whether the payment method used to enter was from OMNY or MetroCard.
fare_class_category	TEXT	The class of fare payment used for the trip. The consolidated categories are: • MetroCard – Fair Fare • MetroCard – Full Fare • MetroCard – Other • MetroCard – Senior & Disability

- Aggregated all timestamps by date and summated ridership and transfer by each date
- Obtained all data points from 2023.

```
df = pd.read_csv('data/MTA_Subway_Hourly_Ridership__Beginning_February_2022_3.20.2024.csv', low_memory=False)
df['transit_timestamp'] = pd.to_datetime(df['transit_timestamp'])
df_ridership = df.groupby([df['transit_timestamp'].dt.date,'station_complex_id','station_complex','borough','Georeference']
                          ).agg({'ridership': 'sum', 'transfers': 'sum'}).reset_index()
df_2023_ridership = df_ridership[df_ridership['transit_timestamp'].dt.year == 2023]
df_2023_ridership.to_csv('data/ridership_daily_station_2023', index=False)
```

- Created a dictionary dataset that noted station_complex_id with their associated station, borough, and geo coordinates

pandas

NYPD Transit Districts

- Added transit bureau district number to each station complex in our station complex dictionary

station_complex_id	station_complex	borough	Georeference	transit_bureau
1	Astoria-Ditmars Blvd (N,W)	Queens	POINT (-73.91203308105469 40.7750358581543)	20
10	49 St (N,R,W)	Manhattan	POINT (-73.98413848876953 40.759899139404 3)	1
100	Hewes St (M,J)	Brooklyn	POINT (-73.95343017578125 40.706871032714844)	33
101	Marcy Av (M,J,Z)	Brooklyn	POINT (-73.95775604248047 40.70835876464844)	2
103	Bowery (J,Z)	Manhattan	POINT (-73.99391174316406 40.720279693603516)	2
107	Broad St (J,Z)	Manhattan	POINT (-74.01105499267578 40.70647430419922)	2

NYPD Subway Fare Evasion Summon and Arrests

Data Label	Data Type	Data Description
Quarter	DATE	Denotes quarter and year that the data was collected
Transit District	INTEGER	Denotes the transit district in which the data was collected
Gender	TEXT, INTEGER	Count based on Gender
Race	TEXT, INTEGER	Count based on Race
Age	TEXT,INTEGER	Count based on Age

- Data pull from NYPD summaries.
- Summons and arrests based count on transit district from 2018 to 2023
- Using excel, created a dataset that is usable for visualizations.
- Original data was in Excel Pivot Tables

Quarter	Quarter Start Date	Quarter End Date	Year	Transit Distict	Female	Male	Unknown	American	Asian/Pac	Black	Hispanic	Unknown	White	10 - 17	18 - 24	25 - 40	41 - 59	60+	Unknown	Grand Total
1	1/1/2018	3/31/2018	2018	1	507	1204	1	12	124	572	472	90	442	79	792	627	191	20	3	1,712
1	1/1/2018	3/31/2018	2018	2	376	666	4	6	134	294	221	92	299	41	403	433	146	20	3	1,046
1	1/1/2018	3/31/2018	2018	3	138	336	2	1	24	178	182	11	80	55	193	152	69	6	1	476
1	1/1/2018	3/31/2018	2018	4	353	906	3	11	93	423	322	55	358	49	516	517	168	11	1	1,262
1	1/1/2018	3/31/2018	2018	11	121	507	3	2	13	258	319	16	23	52	275	222	80	2	0	631

Turnstile Usage Data – Problems

Columns in this Dataset

Column Name	Description	Type
C/A	Control Area name/Booth name. This is the internal identifica...	Plain Text
Unit	Remote unit ID of station	Plain Text
SCP	Subunit/Channel/position represents a specific address for a ...	Plain Text
Station	Name assigned to the subway station by operations planning....	Plain Text
Line Name	Train lines stopping at this location. Can contain up to 20 sing...	Plain Text
Division	Represents the Line originally the station belonged to BMT, IR...	Plain Text
Date	Represents the date of the audit data	Date & Time
Time	Represents the time of the reported data (HH:MM:SS). The no...	Plain Text
Description	Represent the "REGULAR" scheduled audit event (Normally oc...	Plain Text
Entries	The cumulative ENTRY register value for a device. This regist...	Number
Exits	The cumulative EXITS register value for a device. This register...	Number

Turnstile Usage Data and MTA Subway Hourly Ridership didn't have matching columns.

- No column denoting entire station complex.
  - To include station complex name and the trains serviced
- No total ridership and transfer count.

Turnstile Usage Data

MTA turnstiles data released before 2022 and change into usable ridership data format

```
#Create special name to use crosswalk table
# creating a new column "unique_ID" based on the station, line name and division
turnstiles["unique_ID"] = turnstiles[["Station", "Line Name", "Division"]].apply("-".join, axis=1)
```

	MTA Station ID	unique_ID
0	H007	1 AV-L-BMT
1	N037	103 ST-BC-IND
2	R170	103 ST-1-IRT
3	R252	103 ST-6-IRT
4	R529	103 ST-CORONA-7-IRT
5	J034	104 ST-JZ-BMT
6	N137	104 ST-A-IND
7	R254	110 ST-6-IRT
8	J035	111 ST-J-BMT
9	N139	111 ST-A-IND
10	R530	111 ST-7-IRT
11	N030	116 ST-BC-IND
12	R256	116 ST-6-IRT
13	R302	116 ST-23-IRT
14	R173	116 ST-COLUMBIA-1-IRT
15	J037	121 ST-JZ-BMT
16	N026	125 ST-ACBD-IND
17	R174	125 ST-1-IRT
18	R258	125 ST-456-IRT
19	R304	125 ST-23-IRT
20	N024	135 ST-BC-IND
21	R306	135 ST-23-IRT
22	R176	137 ST CITY COL-1-IRT

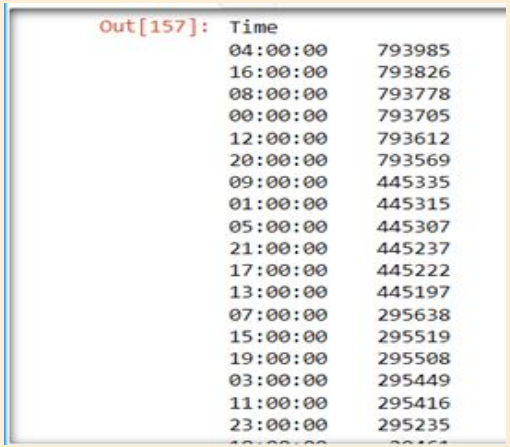
- Create unique ID & Use Crosswalk table to get from old format/ name to up to date names
- Get additional Georeference and geographic info data

	Missing Dates
0	2019-02-23
1	2019-02-24
2	2019-02-25
3	2019-02-26
4	2019-02-27
5	2019-02-28
6	2019-03-01
7	2019-06-15
8	2019-06-16
9	2019-06-17
10	2019-06-18
11	2019-06-19
12	2019-06-20
13	2019-06-21
14	2019-08-03
15	2019-08-04
16	2019-08-05

- There are missing dates
- Remove rows with null values / missing values since just very small percentage

Turnstile Usage Data

MTA turnstiles data released before 2022 and change into usable ridership data format



Out[157]:

Time	
04:00:00	793985
16:00:00	793826
08:00:00	793778
00:00:00	793705
12:00:00	793612
20:00:00	793569
09:00:00	445335
01:00:00	445315
05:00:00	445307
21:00:00	445237
17:00:00	445222
13:00:00	445197
07:00:00	295638
15:00:00	295519
19:00:00	295508
03:00:00	295449
11:00:00	295416
23:00:00	295235

- Outliners:
  - negative aggregated ridership number
  - Extreme nonsense hourly ridership number
- unusual system data collecting time

	station_complex_id	Date	ridership	net_exits	station_complex	Georeference	borough
0	A002	2019-01-01	12897	14333	Lexington Av (N,R,W)/59 St (4,5,6)	POINT (-73.967255 40.76266)	M
1	A002	2019-01-02	37302	37973	Lexington Av (N,R,W)/59 St (4,5,6)	POINT (-73.967255 40.76266)	M
2	A002	2019-01-03	40823	40826	Lexington Av (N,R,W)/59 St (4,5,6)	POINT (-73.967255 40.76266)	M
3	A002	2019-01-04	40207	41009	Lexington Av (N,R,W)/59 St (4,5,6)	POINT (-73.967255 40.76266)	M
4	A002	2019-01-05	17778	19431	Lexington Av (N,R,W)/59 St (4,5,6)	POINT (-73.967255 40.76266)	M
5	A002	2019-01-06	14638	16490	Lexington Av (N,R,W)/59 St (4,5,6)	POINT (-73.967255 40.76266)	M
6	A002	2019-01-07	39281	40364	Lexington Av (N,R,W)/59 St (4,5,6)	POINT (-73.967255 40.76266)	M
7	A002	2019-01-08	40687	41082	Lexington Av (N,R,W)/59 St (4,5,6)	POINT (-73.967255 40.76266)	M
8	A002	2019-01-09	41877	42428	Lexington Av (N,R,W)/59 St (4,5,6)	POINT (-73.967255 40.76266)	M
9	A002	2019-01-10	41695	43215	Lexington Av (N,R,W)/59 St (4,5,6)	POINT (-73.967255 40.76266)	M
10	A002	2019-01-11	40534	41233	Lexington Av (N,R,W)/59 St (4,5,6)	POINT (-73.967255 40.76266)	M

- Aggregate and sum up daily ridership of each turnstiles of each station entry at each station Complex
- Calculate the turnstiles rides (ridership count for each ride is not specified with constraints to avoid outliers

Research

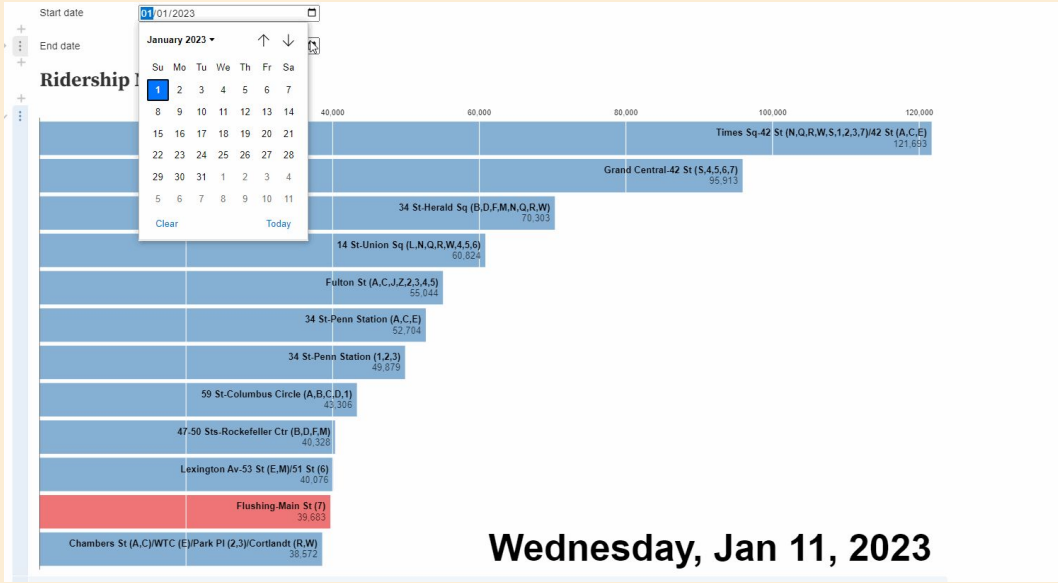
- Looked into different Metro Systems
- Researched various metrics of different Metro systems
  - Length of track per Metro System
  - Annual Budget for 2024
  - Average Cost per mile
    - Done by collecting average distance traveled per customer if the travel cost is NOT by distance (i.e. NYC MTA)
  - Annual Budgets for MTA in past 7 years

03 Visualizations and Discoveries

Visualizing Most Used Subway Stations

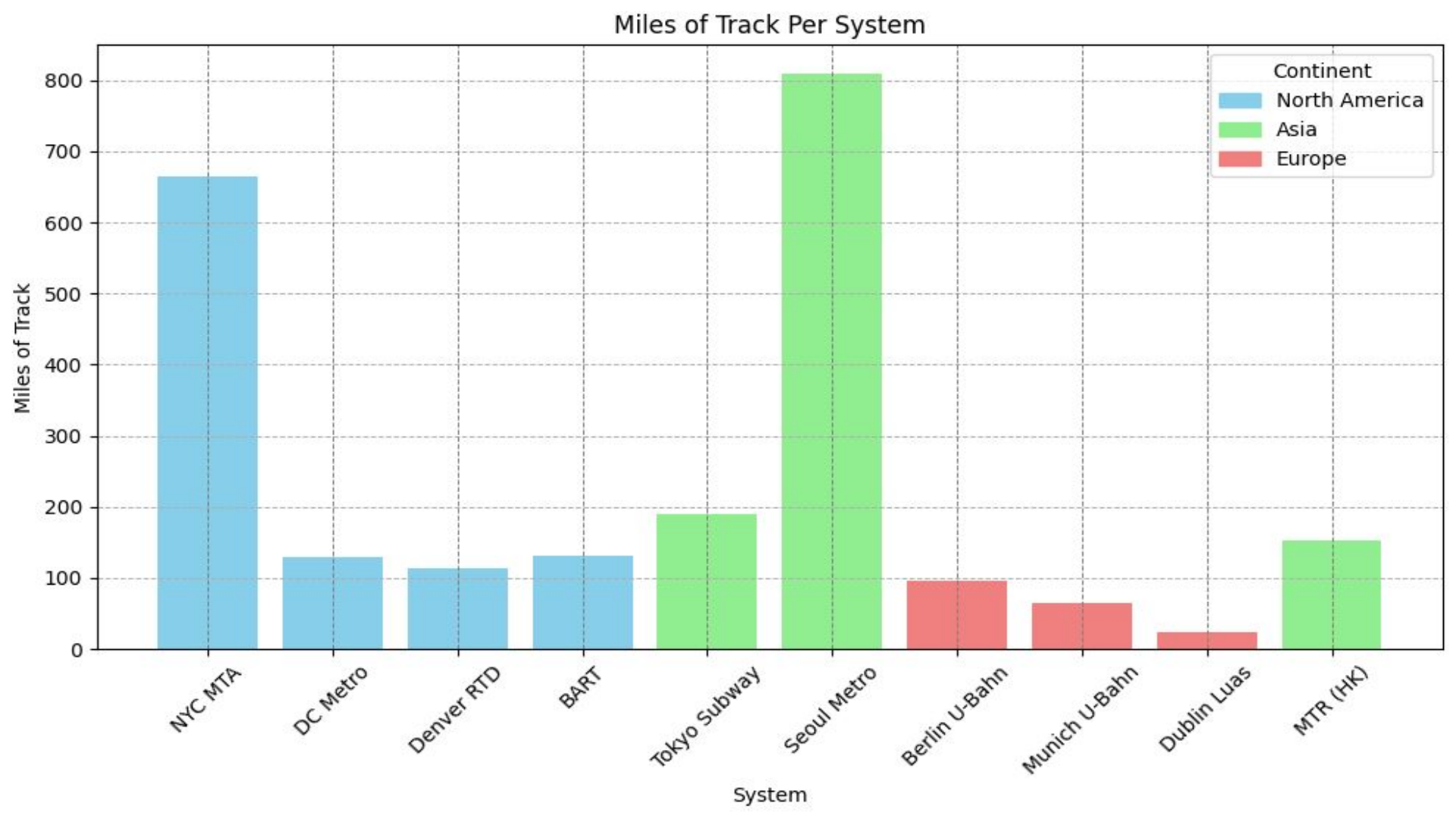
Fun Facts:

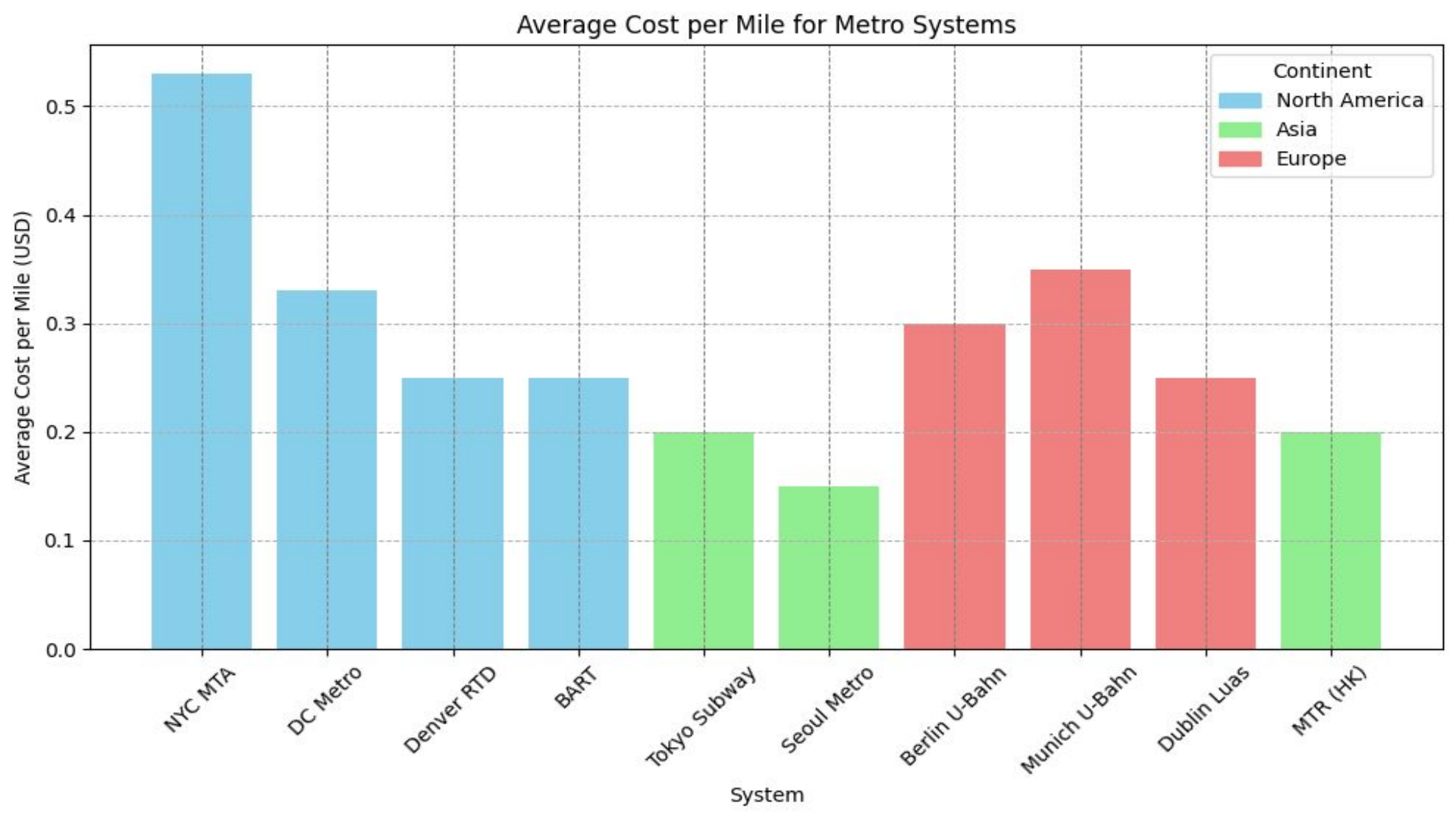
- Times Sq 42nd St complex always has the most ridership across the MTA system.
- On weekends: Bedford Ave has the most ridership in Brooklyn
  - Located in Williamsburg and has been a hotspot in recent years for weekend activities
- MTA Stations by CCNY: 125th St (A,B,C,D) is most used.

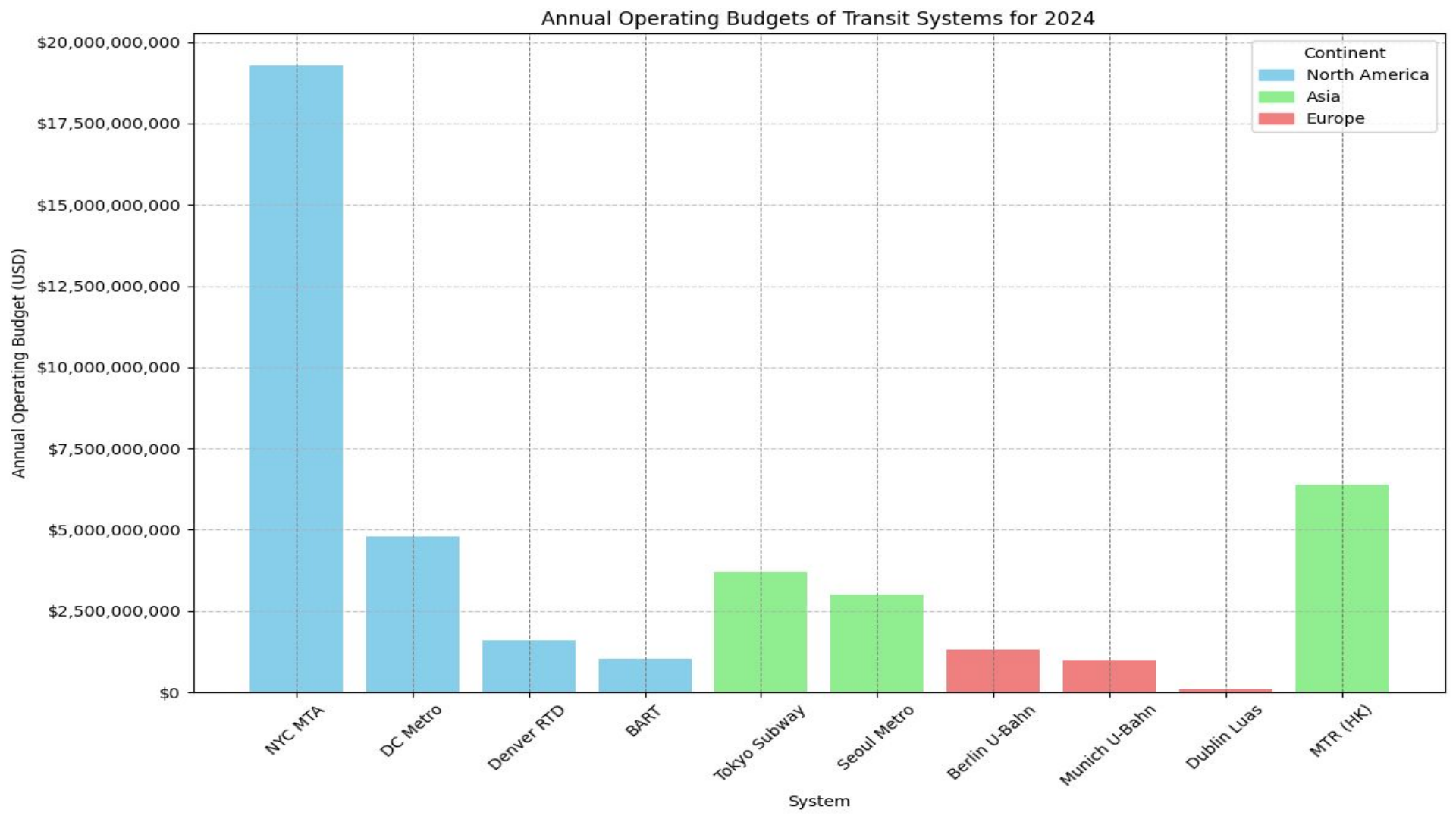


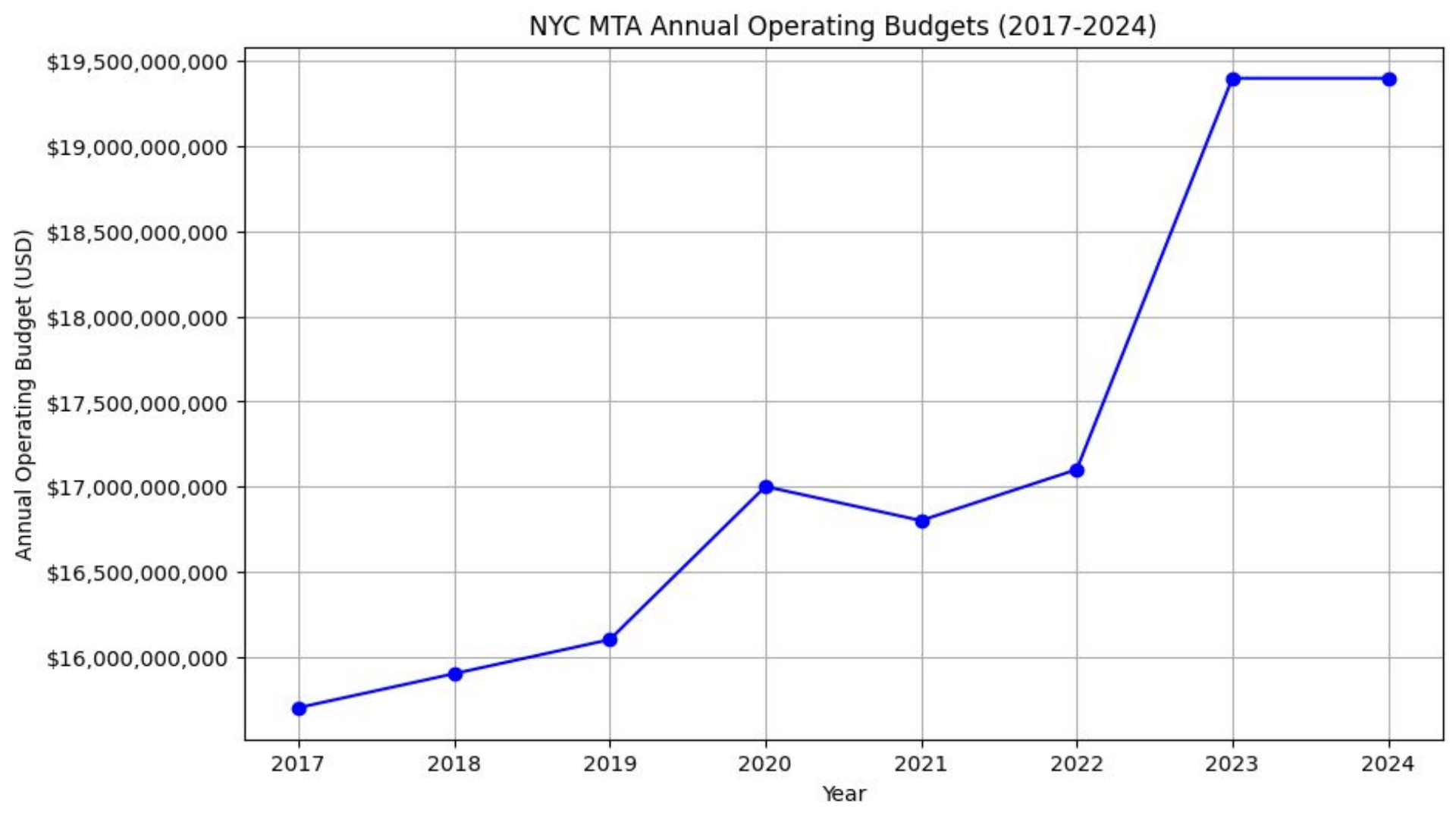
[Link to Observable Notebook]

Observable

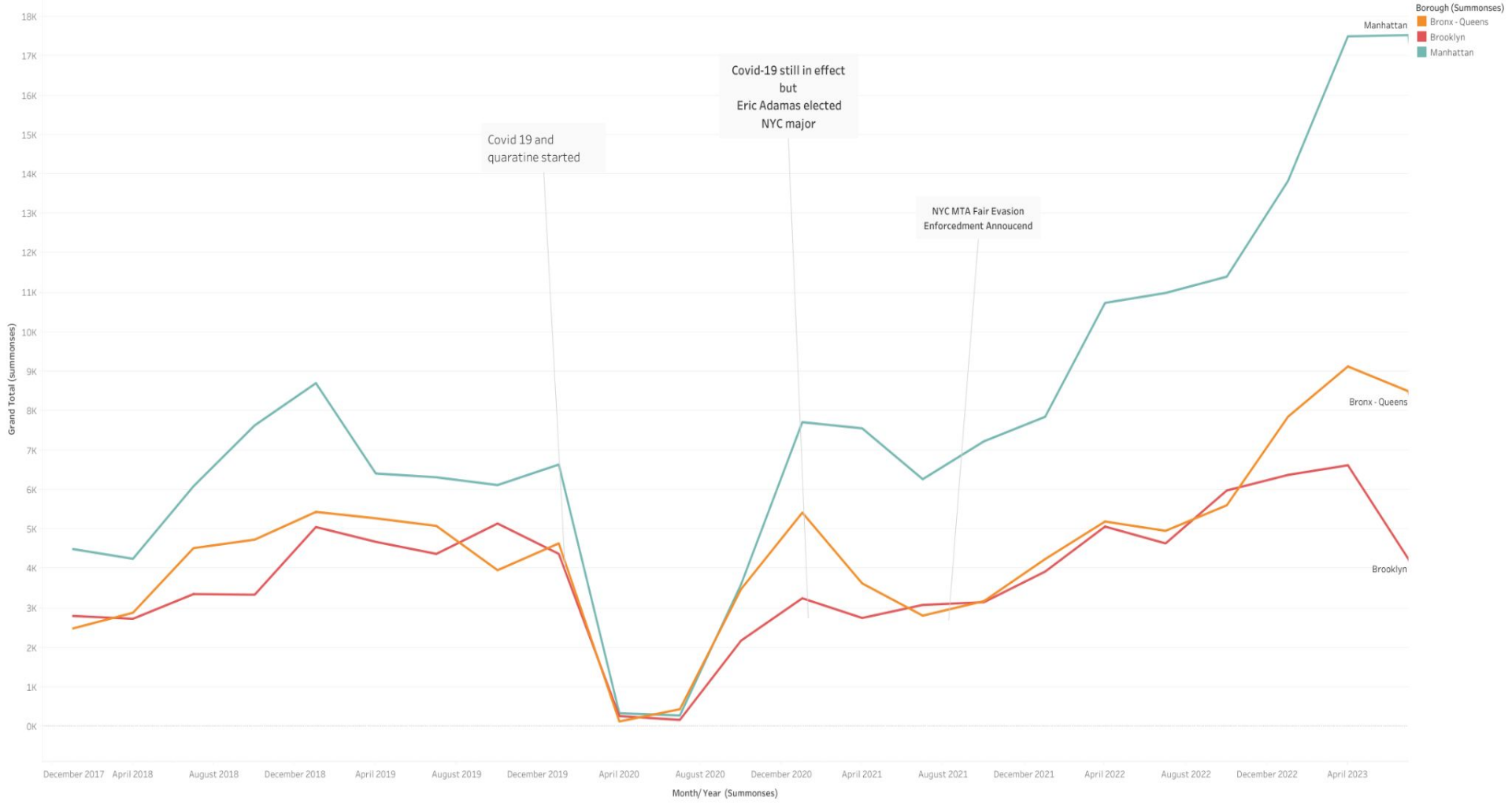

Miles of Track Per System
Continent
North America
Asia
Europe
Miles of Track
0
100
200
300
400
500
600
700
800
NYC MTA
DC Metro
Denver RTD
BART
Tokyo Subway
Seoul Metro
Berlin U-Bahn
Munich U-Bahn
Dublin Luas
MTR (HK)
System

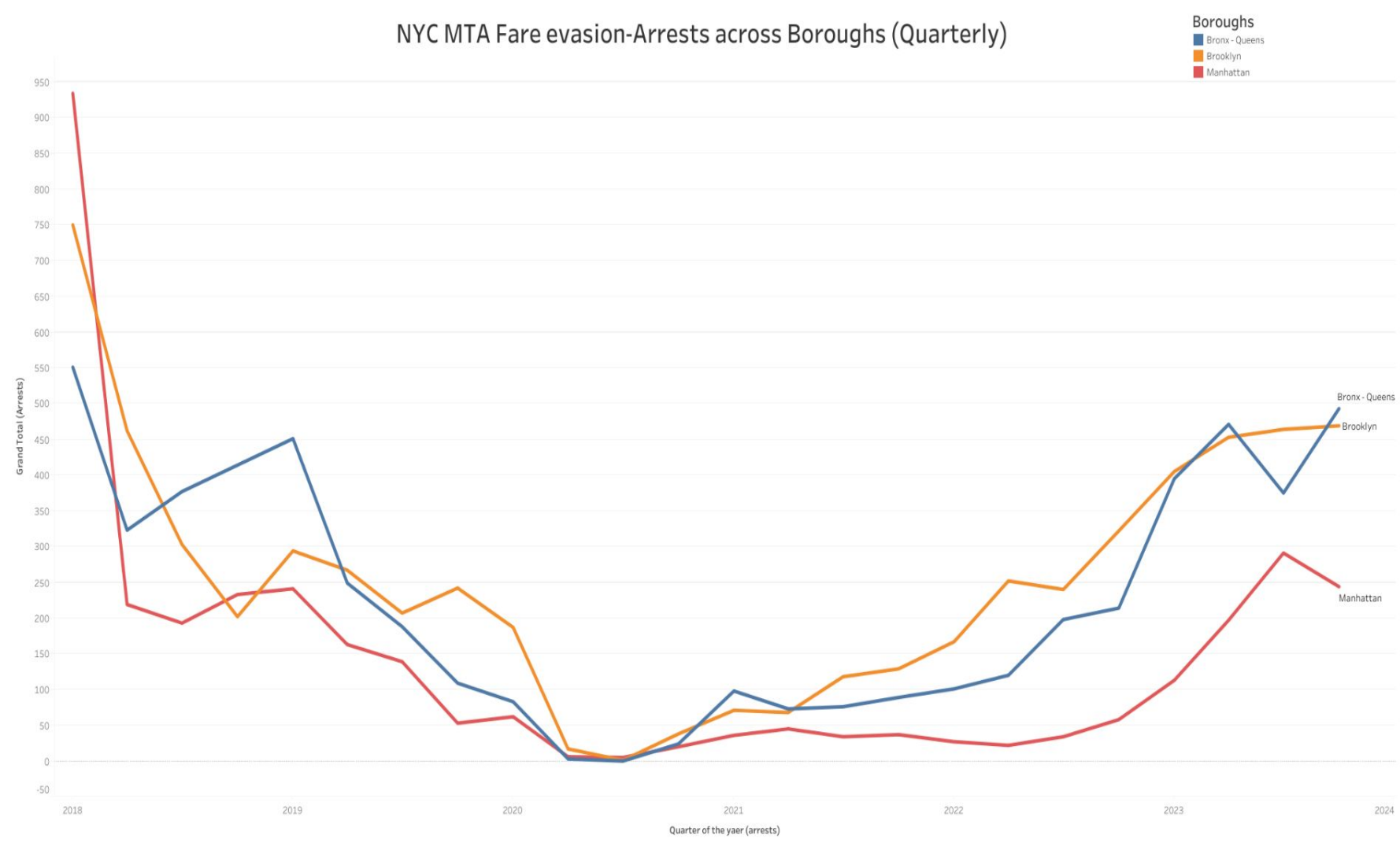

Average Cost per Mile for Metro Systems
Average Cost per Mile (USD)
0.0
0.1
0.2
0.3
0.4
0.5
Continent
North America
Asia
Europe
NYC MTA
DC Metro
Denver RTD
BART
Tokyo Subway
Seoul Metro
Berlin U-Bahn
Munich U-Bahn
Dublin Luas
MTR (HK)
System


Annual Operating Budgets of Transit Systems for 2024
Annual Operating Budget (USD)
$20,000,000,000
$17,500,000,000
$15,000,000,000
$12,500,000,000
$10,000,000,000
$7,500,000,000
$5,000,000,000
$2,500,000,000
$0
Continent
North America
Asia
Europe
NYC MTA
DC Metro
Denver RTD
BART
Tokyo Subway
Seoul Metro
Berlin U-Bahn
Munich U-Bahn
Dublin Luas
MTR (HK)
System


NYC MTA Annual Operating Budgets (2017-2024)
Annual Operating Budget (USD)
$19,500,000,000
$19,000,000,000
$18,500,000,000
$18,000,000,000
$17,500,000,000
$17,000,000,000
$16,500,000,000
$16,000,000,000
2017
2018
2019
2020
2021
2022
2023
2024
Year

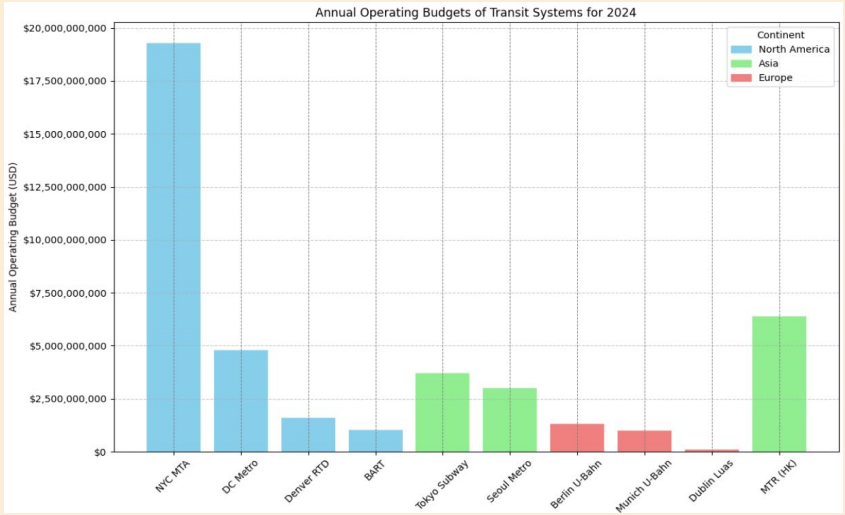
NYC MTA Fare evasion-Summonses across Boroughs (Quarterly)

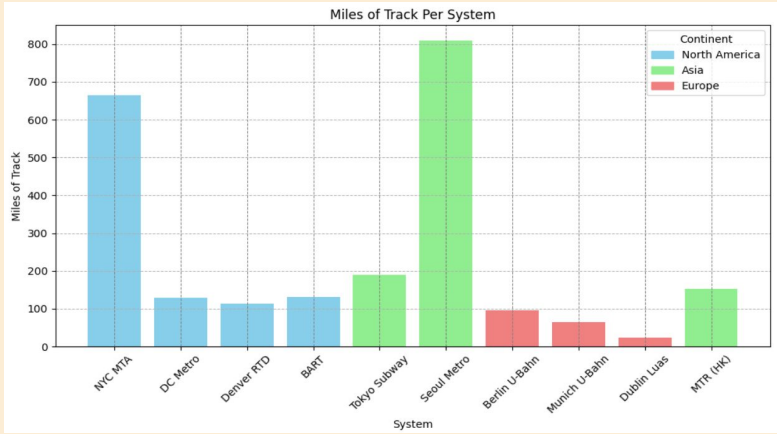



NYC MTA Fare evasion-Arrests across Boroughs (Quarterly)
Boroughs
Bronx - Queens
Brooklyn
Manhattan
Grand Total (Arrests)
950
900
850
800
750
700
650
600
550
500
450
400
350
300
250
200
150
100
50
0
-50
2018
2019
2020
2021
2022
2023
2024
Quarter of the yaer (arrests)
Bronx - Queens
Brooklyn
Manhattan

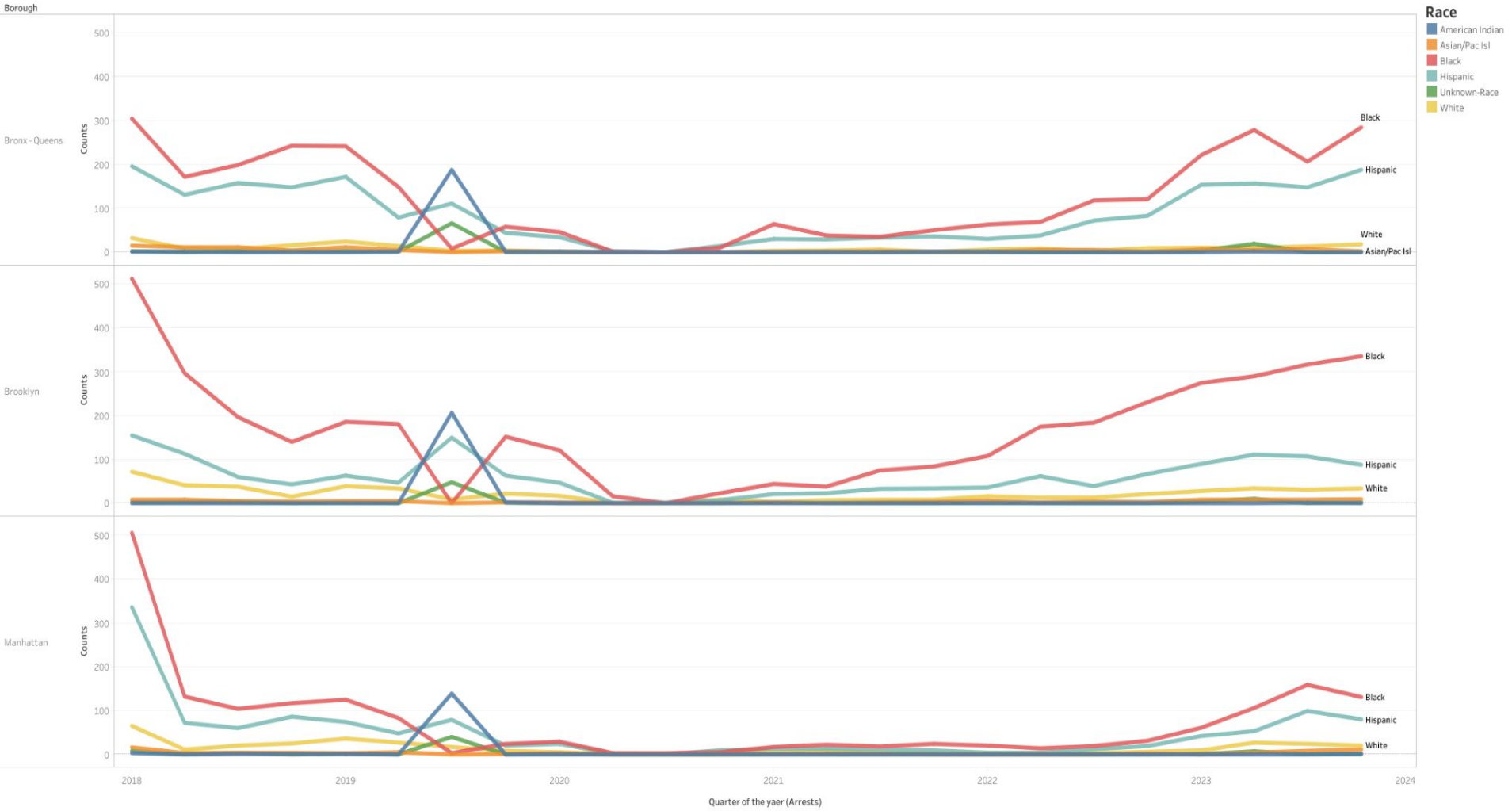
Discussion

- MTA has the biggest budget out of all systems we investigated, yet it is one of the most unclean, slow, untimely stations.
- Similar length of track to Seoul Metro, and with more money, yet still incredibly poor service.

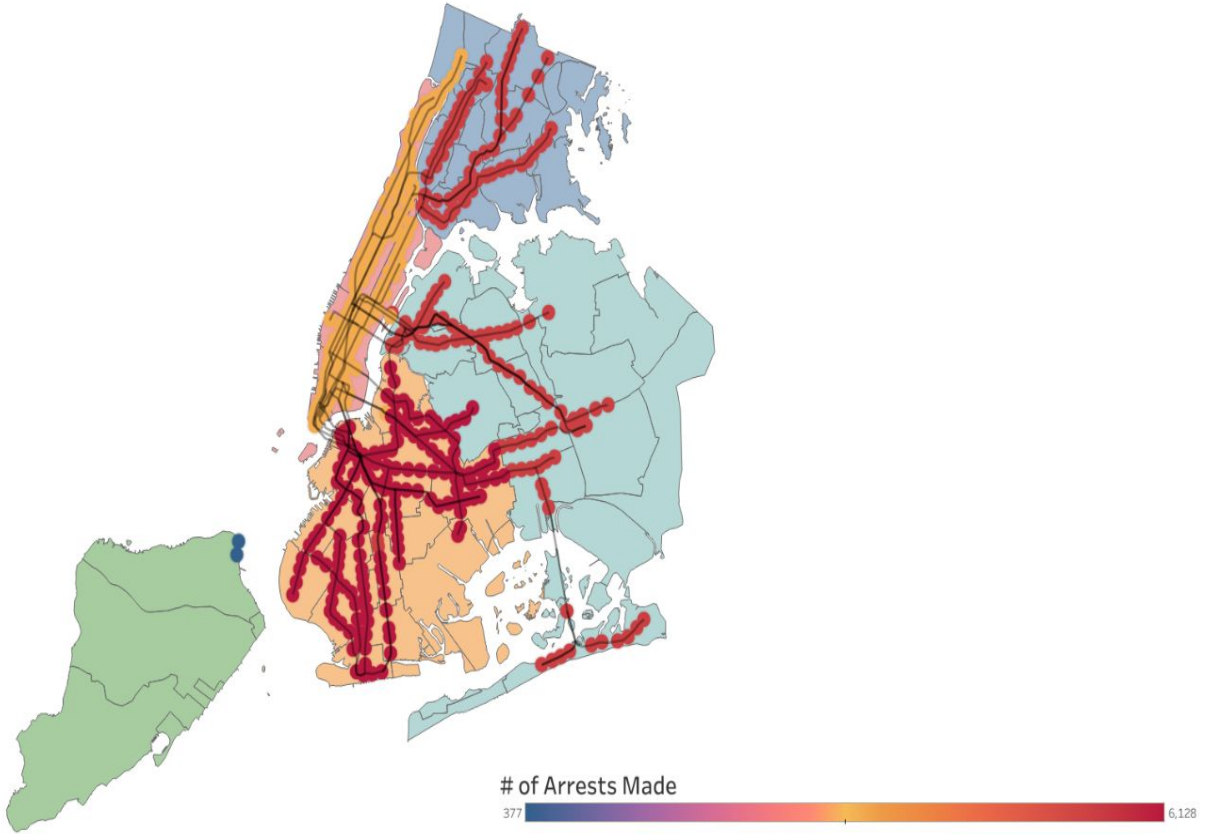




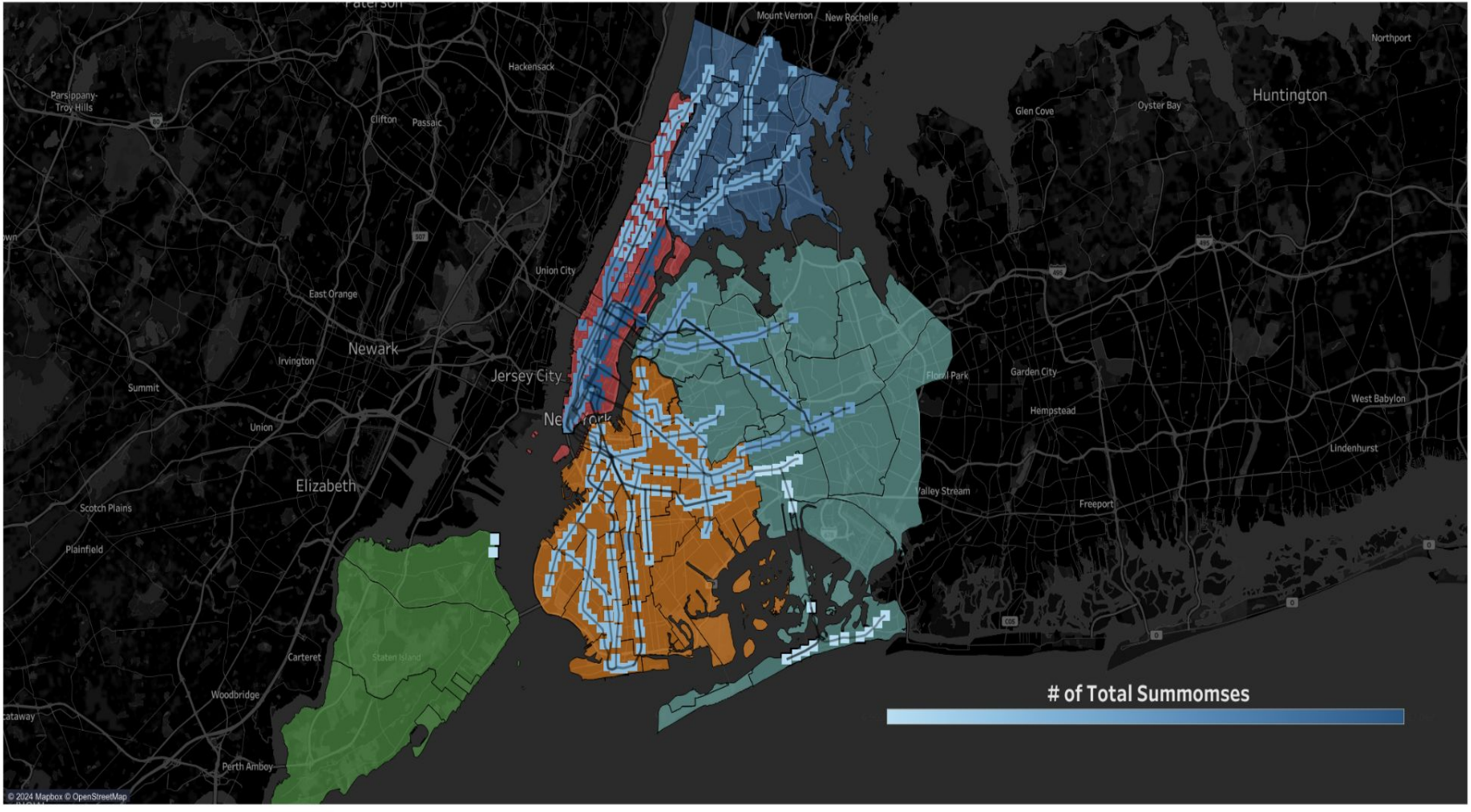
NYC MTA Fare evasion-Arrests Made Across Different Races 2018-2023(Quarterly)



NYC MTA Fare Evasion Total Arrests Made on each MTA Subway Lines



NYC MTA Fare Evasion Total Summonses Made on each MTA Subway lines



Observation

- Summonses Issued and Arrests Made in different borough are not consistent with which borough have larger crime numbers than the others:
  - Most summonses made are mainly in Manhattan and more in Transit Districts located in the East of Manhattan, followed by Queens/Bronx, and least in Brooklyn out of all boroughs.
  - More arrests were made in Brooklyn and then Bronx / Queens, and least in Manhattan Districts.

04

Final Thoughts

Possible Solutions

- Use money to create better infrastructure
- Security issues have not improved much from security guards - instead it has RISEN. Use less money on security, more on bettering homeless individuals, individuals with drug addictions, and more.

Reasoning

- Closing/opening doors to train when train arrives prevents pollution as well as shovings/suicide.
- Assistance for drug addiction, homelessness, housing can help cause a domino effect of positive changes.

Future Work with this Project

- Creating a shapefile to present the Transit District jurisdiction efficiently.
- Create a comprehensive dashboard to allow user to explore, visualize and follow the research.
- Further research what other countries do and think of concrete plans to better allocate MTA's budget