MTA Daily Ridership

John Cruz

2023-03-02

Introduction

The Metropolitan Transportation Authority (MTA), provides a daily ridership dataset containing systemwide ridership and traffic estimates for subways, buses, Long Island Rail Road, Metro-North Railroad, Access-A-Ride, and Bridges and Tunnels. The data starts March 1, 2020 (April 1, 2020 for LIRR and Metro-North) until current date, and provides a percentage comparison against a comparable pre-pandemic date.

Note : Will need to update the code from using CSV file to API connection to establish up to date data. This data was last pulled March 3, 2023.

The objective will be to compare daily ridership between different modes of transportation against estimated pre-pandemic levels.

Data Source: MTA Daily Ridership

Required Libraries

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.4.0     v purrr   1.0.1
## v tibble  3.1.8     v dplyr   1.1.0
## v tidyr   1.3.0     v stringr 1.5.0
## v readr   2.1.3     v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

Import file

The data is read from a CSV file downloaded from the (NY Open Data)[https://data.ny.gov/]. We can see that there are multiple columns for each unique mode of transportation.

```
mta_data <- read_csv('mta_daily_ridership.csv', show_col_types = FALSE)
mta_data$Date <- as.Date(mta_data$Date, format = '%m/%d/%Y')

knitr::kable(head(mta_data))
```

Date	Subways: Total Estimated Ridership	Subways: % of Comparable Pre-Pandemic Day	Buses: Total Estimated Ridership	Buses: % of Comparable Pre-Pandemic Day	LIRR: Total Estimated Ridership	LIRR: % of Comparable Pre-Pandemic Day	Metro-North: Total Estimated Ridership	Metro-North: % of Comparable Pre-Pandemic Day	Access-A-Ride: Total Scheduled Trips	Access-A-Ride: % of Comparable Pre-Pandemic Day	Bridges and Tunnels: Total Traffic	Bridges and Tunnels: % of Comparable Pre-Pandemic Day	Staten Island Railway: Total Estimated Ridership	Staten Island Railway: % of Comparable Pre-Pandemic Day
2023-03-02	3760245	0.68	1390337	0.62	195243	0.62	176678	0.64	28978	0.97	934427	1.01	7428	0.47
2023-03-01	3773706	0.68	1477880	0.66	192142	0.61	176661	0.64	29449	0.99	901530	0.98	7401	0.46
2023-02-28	3408751	0.63	1334393	0.62	177532	0.59	139326	0.52	26116	0.89	751643	0.85	6689	0.41
2023-02-27	3335204	0.61	1415210	0.66	180277	0.60	160346	0.60	26730	0.91	854625	0.97	6751	0.42
2023-02-26	1623164	0.74	677706	0.69	74226	0.95	67702	0.74	15891	0.94	807475	1.07	1095	0.39
2023-02-25	2041901	0.71	802244	0.63	86403	0.92	81742	0.62	15461	0.95	831613	1.00	1396	0.33

Transform Data into Long Format

The data is difficult to compare with its current wide format. The following code creates a loop that goes through each column, creating subset data frames that correspond to each unique transit type. The data is then merged back together into one main dataframe containing the date, transit type, and statistics.

```
# create matrix to uniquely identify columns
col_name_matrix <-
  colnames(mta_data) |>
  str_split(pattern = ":", simplify = TRUE)

# find distinct transit types
col_name <- unique(col_name_matrix[, 1])

# loop through data frame and create new long format data frame
merged_transit <- data.frame()
```

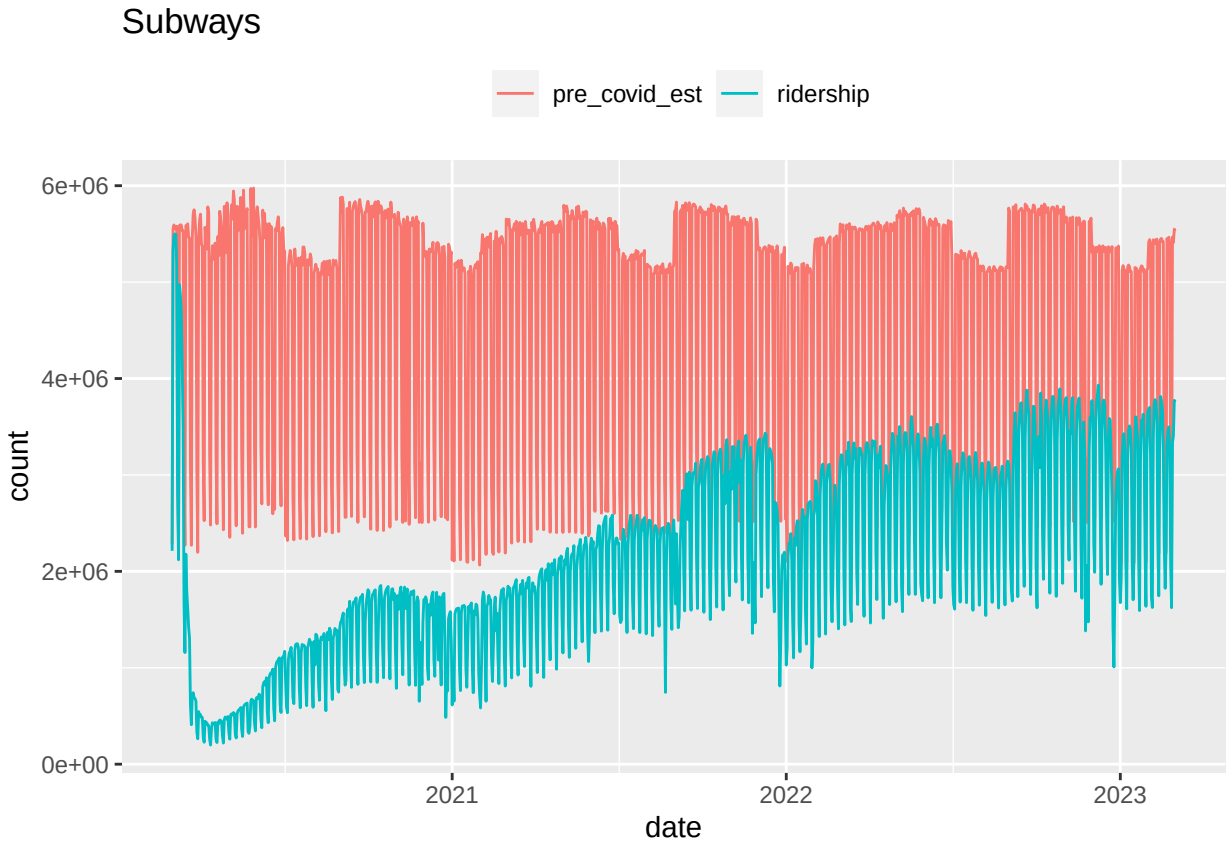
```
for (name in col_name){
  if (str_to_lower(name) == 'date'){
    next
  }

  temp_df <-
    mta_data |>
    select('Date', starts_with(name)) |>
    rename(date = 1, ridership = 2, pre_covid_pct = 3) |>
    mutate(transit_type = name, pre_covid_est = round(ridership / pre_covid_pct, 0)) |>
    relocate(transit_type, .after = 'date')

  merged_transit <-
    rbind(merged_transit, temp_df)

}

knitr::kable(head(merged_transit))
```

date	transit_type	ridership	pre_covid_pct	pre_covid_est
2023-03-02	Subways	3760245	0.68	5529772
2023-03-01	Subways	3773706	0.68	5549568
2023-02-28	Subways	3408751	0.63	5410716
2023-02-27	Subways	3335204	0.61	5467548
2023-02-26	Subways	1623164	0.74	2193465
2023-02-25	Subways	2041901	0.71	2875917

Comparing Transit Types to their Pre-Pandemic Estimated Ridership

First we see that to compare a transit type, such as the subways, there is seasonality to account for.

```
merged_transit_long <-
  merged_transit |>
  filter(transit_type == 'Subways')|>
  select(!pre_covid_pct) |>
  gather('category', 'count', -c('date', 'transit_type'))



merged_transit_long |>
  ggplot(aes(x = date, y = count, color = category)) +
  geom_line() +
  ggtitle('Subways') +
  theme(legend.position = 'top', legend.title = element_blank())
```



Trends

The following graphs shows us with a smoothing function how each transit type currently compares to their pre-pandemic levels. Because the data frame has numbers in a wide format for ridership and pre-pandemic percentage, I calculated the estimated pre-pandemic ridership and used the *gather()* function to transform the data into a longer format. A for loop was used to be able to do this while also printing out each unique transit type.

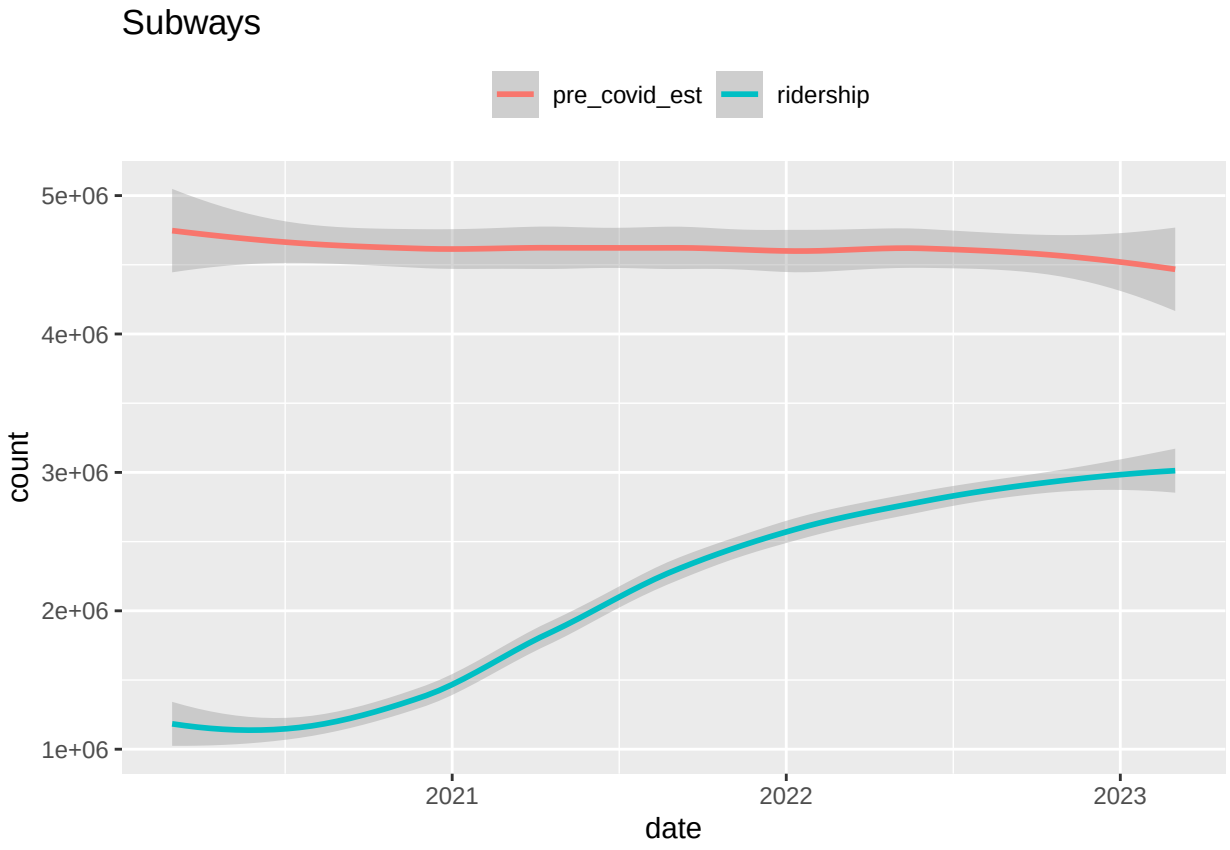
Note: I learned that *ggplot()* does not like to automatically print its plot while in a for loop. You have to explicitly call the *print()* function to ensure it prints out.

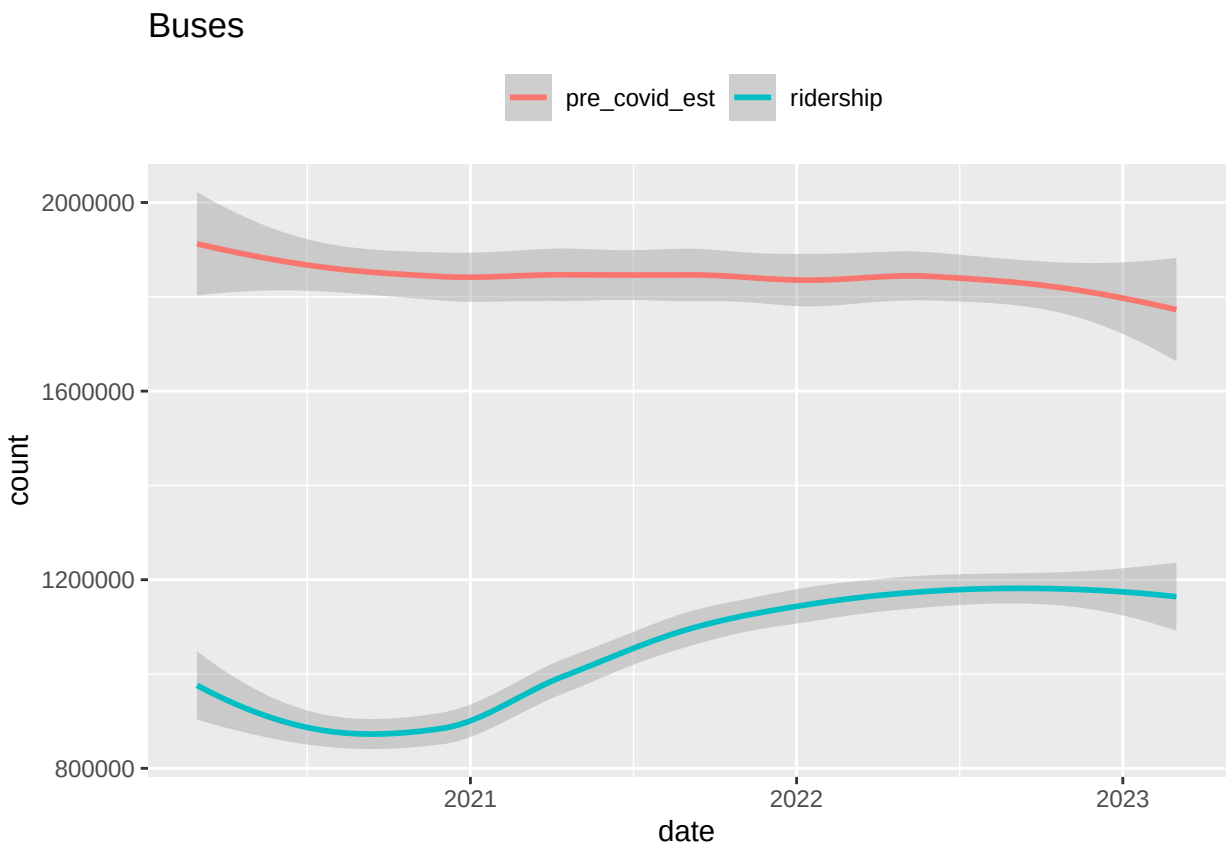
Loops and ggplot

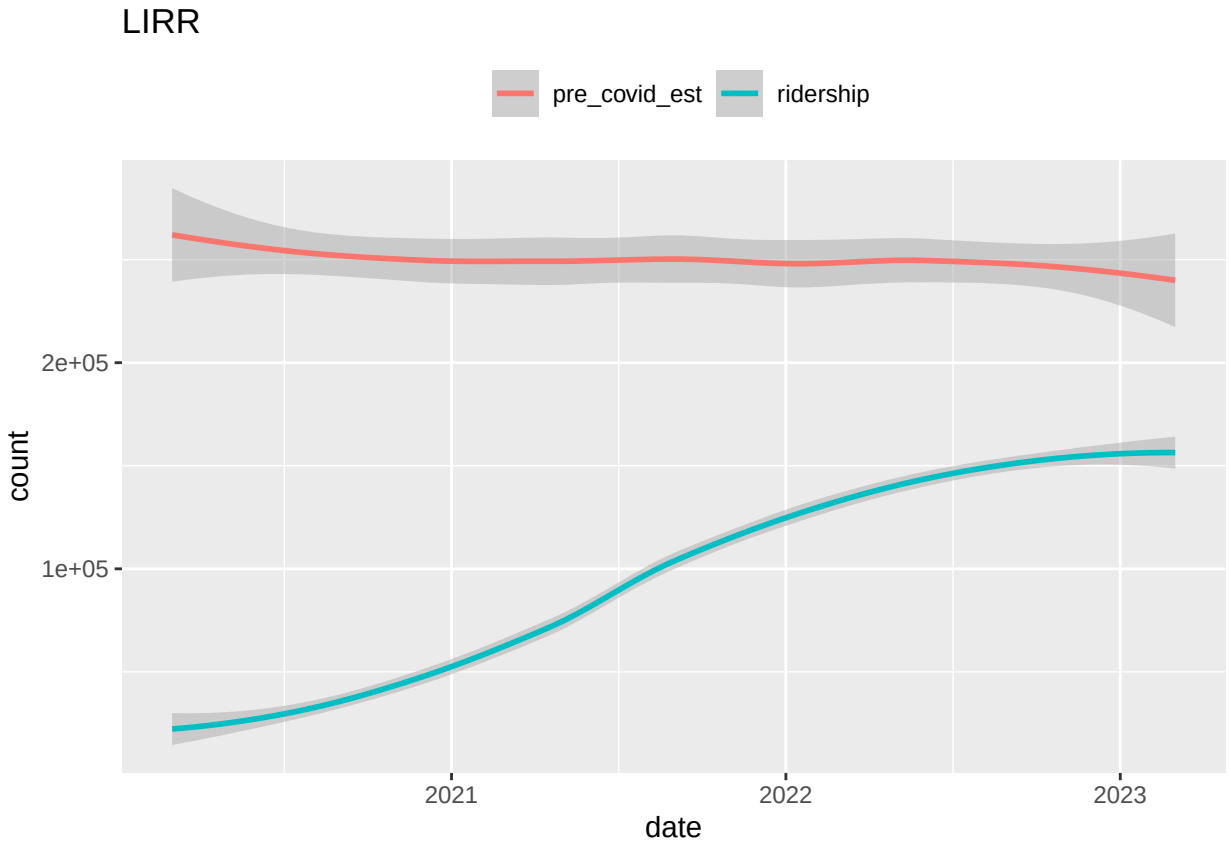
Surprisingly, both **Bridges and Tunnels** and **Access-A-Ride** have both returned close to pre-pandemic levels.

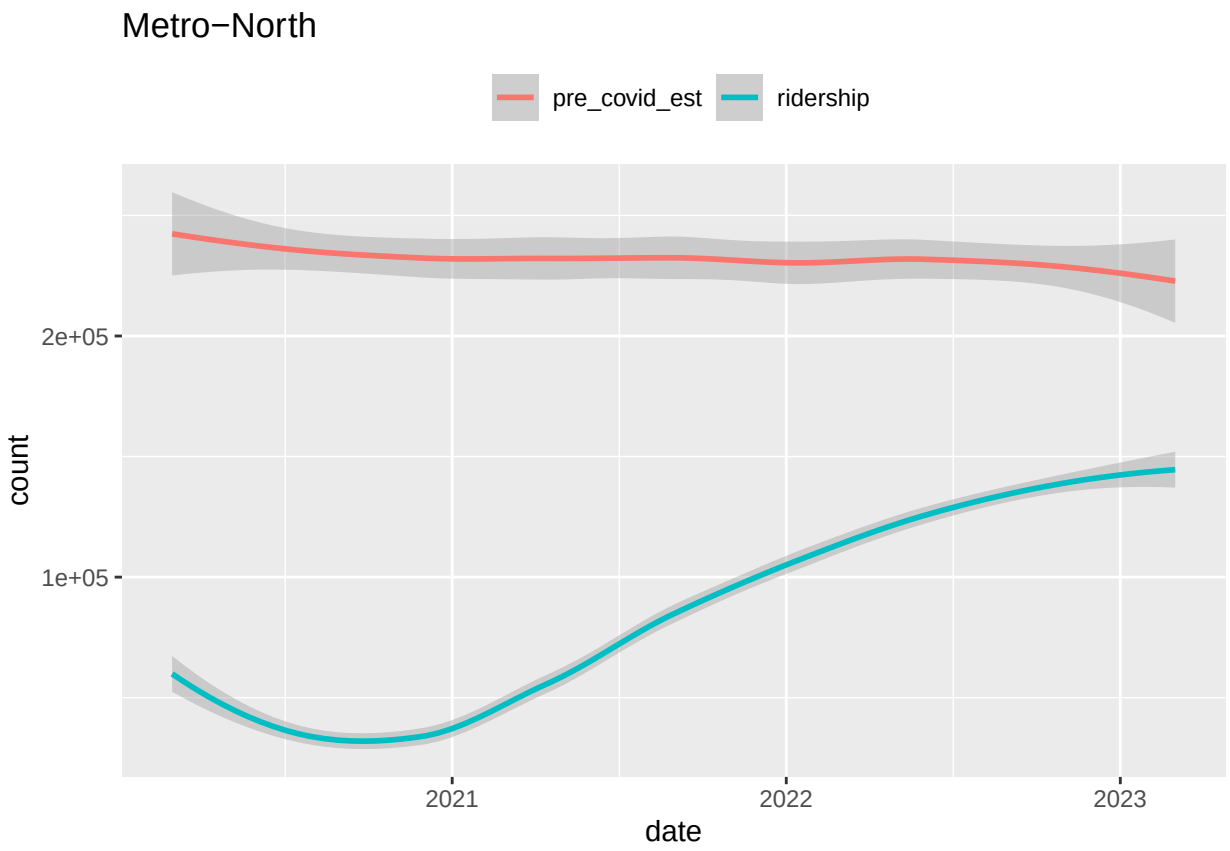
```
for (name in col_name){
  if (str_to_lower(name) == 'date'){
    next
  }

  merged_transit_long <-
    merged_transit |>
    filter(transit_type == name)|>
    select(!pre_covid_pct) |>
    gather('category', 'count', -c('date', 'transit_type'))
```

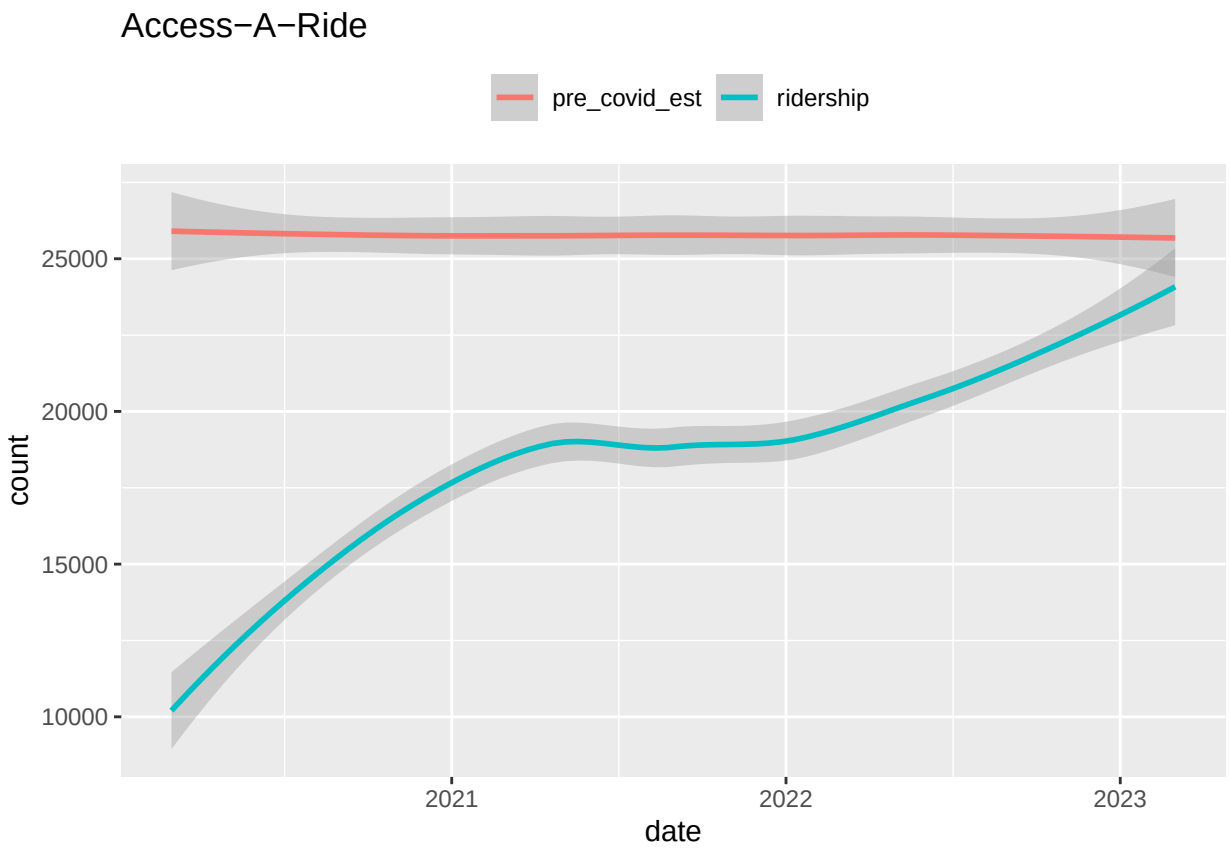
```
  print(merged_transit_long |>
    ggplot(aes(x = date, y = count, color = category)) +
    geom_smooth(method = 'loess') +
    ggtitle(name) +
    theme(legend.position = 'top', legend.title = element_blank()))

}
```

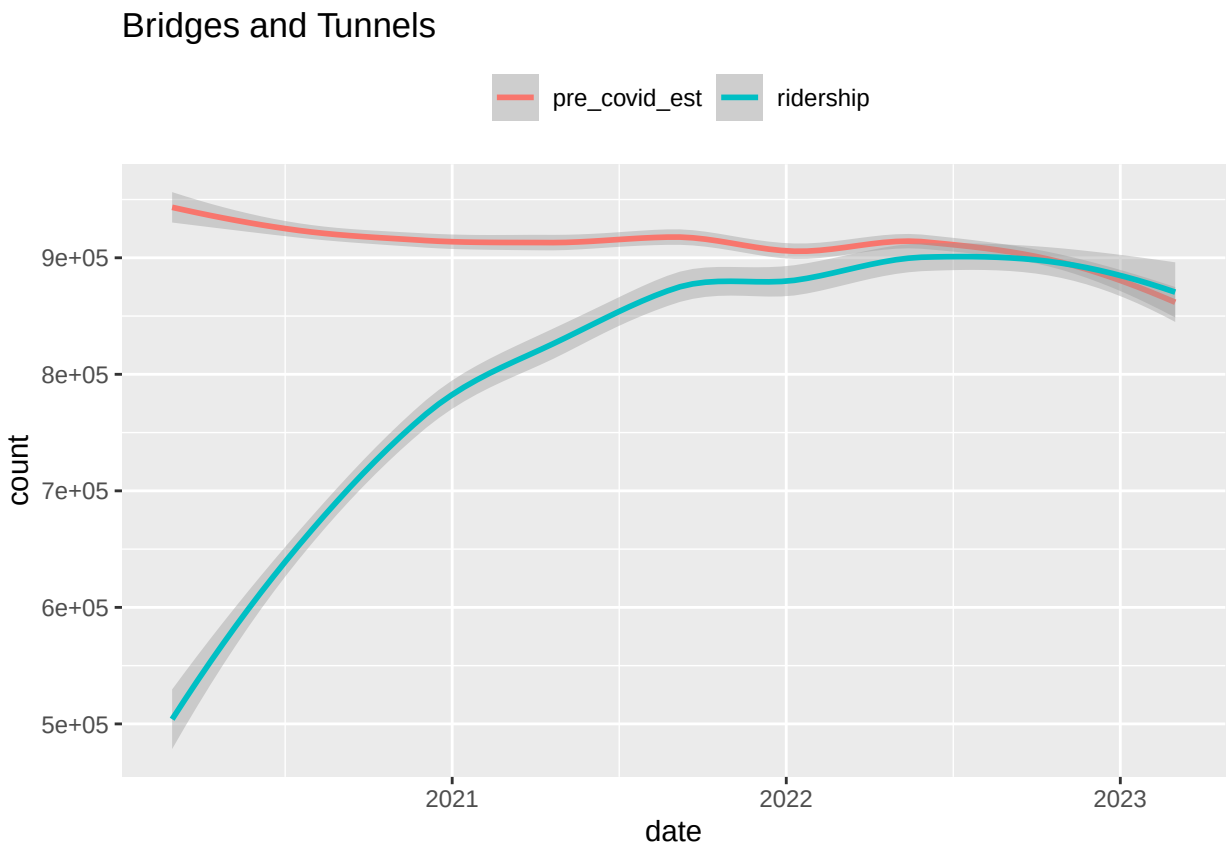


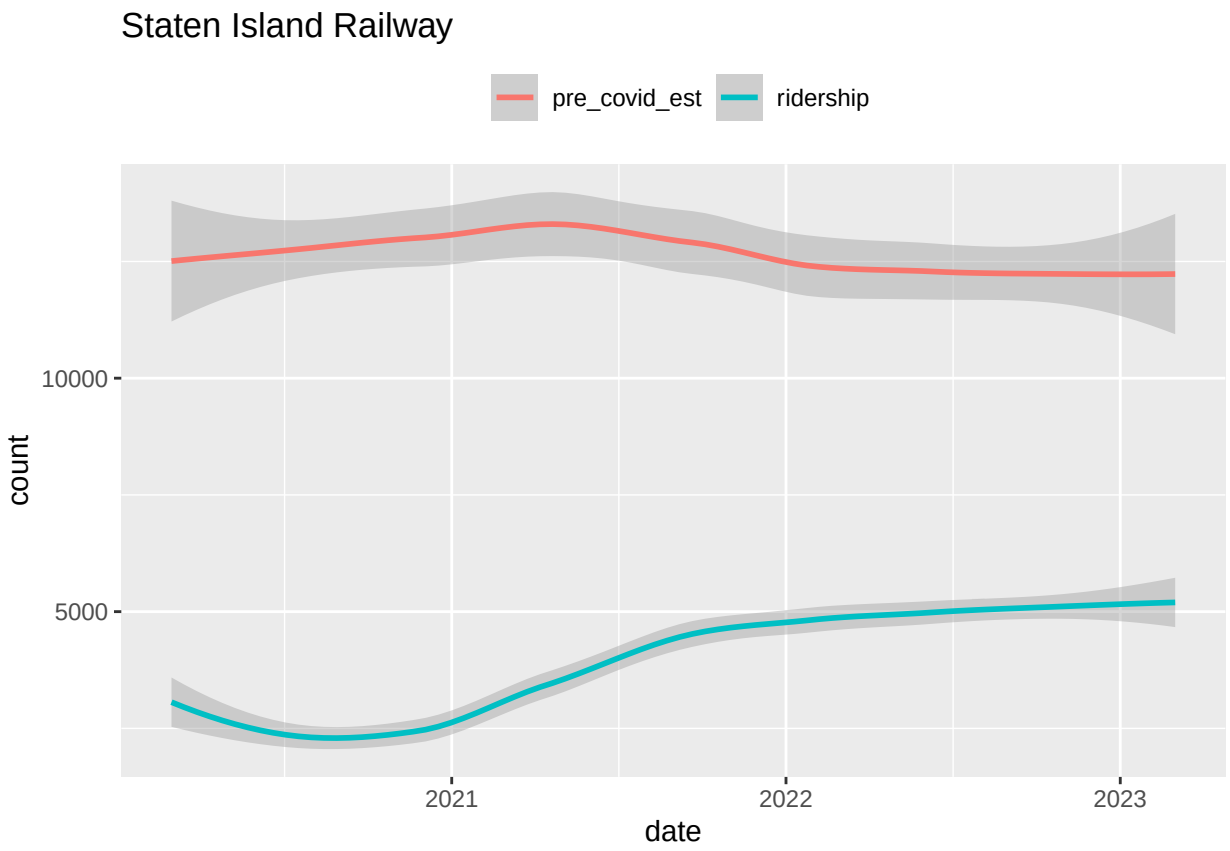

Buses
pre_covid_est
ridership
2000000
1600000
1200000
800000
count
2021
2022
2023
date


LIRR
pre_covid_est
ridership
2e+05
1e+05
count
2021
2022
2023
date


Metro−North
pre_covid_est
ridership
2e+05
1e+05
count
2021
2022
2023
date


Access−A−Ride
pre_covid_est
ridership
25000
20000
15000
10000
count
2021
2022
2023
date


Bridges and Tunnels
pre_covid_est
ridership
9e+05
8e+05
7e+05
6e+05
5e+05
count
2021
2022
2023
date



Reccommendations

Further research should be done on how people, who were fully remote, have returned back to the office. Are Uber and Lyft, providing more personal space, increased since the pandemic? Also, looking into transportation deserts within the city, did the numbers ever change from pre-pandemic numbers and can we account for tourism versus native citizens?