Difference in Differences Analysis of Traffic Congestion Pricing in Manhattan

Ian Wissel (Model specifications, assumptions), Brayden Venning (literature review), Vaughn Mitchell (data management, running Stata commands)

Inspiration/Context

- On Jan 5th, 2025, New York City implemented congestion pricing in the city's Central Business District in Manhattan (60th street and below).
- Drivers entering this zone must pay a $9 toll.
- The intended effects of the congestion pricing is to reduce traffic, lower CO2 emissions, and increase the support and revenue for the public transit system.

Research Question/Hypothesis

- Research Question: How has NYC traffic congestion pricing affected subway ridership?
- Hypothesis: We believe that the congestion pricing policy will affect subway ridership differently between stations inside and outside the zone.
- Ho:B3=0   Ha:B3>0

Literature Review

- The Short-Run Effects of Congestion Pricing in New York City
  - Looks at vehicle's speed on the roads for the congested pricing areas and non-congestion pricing area, as well as the estimated vehicles emissions rate. Study found increased speeds in both CBD and non-CBD zones, as well as decreased CO2 emissions. Also found equitable distribution of benefits between high- and low-income neighborhoods.

- The London Congestion Change
  - Paper looked at some of the effects of congestion pricing in London. Study found that bus Transportaion into central London (where there is congestion pricing) increased by over 50% more than was expected. Also found decrease in traffic delays and little opposition from the public.

- Effective and Equitable Congestion Pricing: New York and Beyond
  - Written before the congestion pricing arguing that the pricing won't solve the problem of traffic since the pricing will mostly impact drivers of personal vehicles instead of taxis and ride-hailing vehicles. Since most of the NYC traffic doesn’t come from personal vehicles, the congestion pricing would be ineffective unless a fairer system of pricing was implemented.

GitHub Documentation

- All elements of our project can be found at https://github.com/vaughnmitchell13/econ308-proj
    - Data sources, notebooks for preprocessing and postprocessing steps
        - Data-Preprocessing.ipynb, Data-Postprocessing.ipynb, Visualization.ipynb
    - Maps, visualizations, and steps for replication in ArcGIS Pro

README

Subway Ridership and NYC's Congestion Relief Zone - Preliminary Steps

👾 Data

- https://mega.nz/folder/S8o3iDQY#LB5H6vnRPKadVrwmf51e6g includes CSV files from the sources below.
- Source for 2025
- Source for 2024
- Note: This is *raw* data, and it includes information about the station (ID and conventional name), transit date and time, lat/long coordinates, and ridership and transfer numbers. These datasets were used in Data-Preprocessing.ipynb, which documents how this data was managed and prepared for further cleaning.
- The CSV files containing "_w_dummy" were created in ArcGIS Pro in order to attribute whether or not the station was inside the congestion zone. Further steps are provided below.

🗺️ ArcGIS Pro Documentation

- An NYC roads layer from census.gov's TIGER/Line collection was loaded to set context of NYC and serve as a foundation for the map.
- The projection for this data was NAD83(2011) / New York Long Island (ft).
- Both `Manhattan_Subway_Ridership_2024.csv` and `Manhattan_Subway_Ridership_2025.csv` were loaded into ArcGIS Pro using the `XY Table to Point` tool.
- A new layer called `Traffic Congestion Zone` was drawn to reflect Manhattan's current congestion relief area.
- A new field called `inside_zone` was added to both attribute tables, using the `Select by Location` tool to fill in 1's or 0's if the station was inside the zone.
- Both tables were exported for final revisions in Data-Postprocessing.ipynb; the final datasets can be found here.
- The map below, although not exactly our intention with a difference-in-differences analysis, is still valuable to see the change in average ridership from 2024 to 2025.

Data Sources

- *Data.gov* - MTA Subway Hourly Ridership: Beginning 2025
- *Data.gov* - MTA Subway Hourly Ridership: 2024
- *U.S. Census Bureau, Department of Commerce* - ShapeFile for NYC Roads (for ArcGIS Pro)

Variables

- station_ridership: *daily* ridership per station in Manhattan
- inside_zone: dummy for if the station was inside the congestion relief zone
- in_2025: dummy for if the data was collected in 2025 or 2024
- inter: inside_zone * in_2025, for the DID treatment effect

Note: For consistency, data for 2024 was 1-1 to 4-3 (same as 2025)

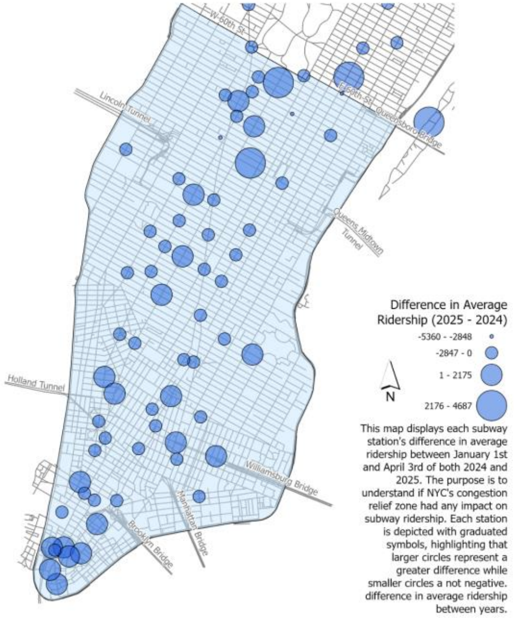
Data Management Map

Original Data

transit_timestamp	transit_mode	station_complex_id	station_complex	borough	payment_method	fare_class_category	ridership	transfers	latitude	longitude	Georeference
01/03/2025 12:00:00 AM	tram	TRAM2	RI Tramway (Roosevelt)	Manhattan	omny	OMNY - Full Fare	12	7	40.75734	-73.95412	POINT (-73.95412 40.75734)
01/03/2025 06:00:00 AM	tram	TRAM1	RI Tramway (Manhattan)	Manhattan	omny	OMNY - Seniors & Disability	1	1	40.761337	-73.96416	POINT (-73.96416 40.761337)
01/03/2025 07:00:00 AM	tram	TRAM2	RI Tramway (Roosevelt)	Manhattan	metrocard	Metrocard - Unlimited 30-Day	10	0	40.75734	-73.95412	POINT (-73.95412 40.75734)
01/03/2025 08:00:00 AM	tram	TRAM1	RI Tramway (Manhattan)	Manhattan	metrocard	Metrocard - Unlimited 30-Day	5	0	40.761337	-73.96416	POINT (-73.96416 40.761337)
01/03/2025 10:00:00 AM	tram	TRAM1	RI Tramway (Manhattan)	Manhattan	omny	OMNY - Students	1	0	40.761337	-73.96416	POINT (-73.96416 40.761337)
01/03/2025 10:00:00 AM	tram	TRAM2	RI Tramway (Roosevelt)	Manhattan	metrocard	Metrocard - Unlimited 7-Day	41	0	40.75734	-73.95412	POINT (-73.95412 40.75734)
01/03/2025 11:00:00 AM	tram	TRAM1	RI Tramway (Manhattan)	Manhattan	metrocard	Metrocard - Unlimited 30-Day	3	0	40.761337	-73.96416	POINT (-73.96416 40.761337)
01/03/2025 12:00:00 PM	tram	TRAM2	RI Tramway (Roosevelt)	Manhattan	metrocard	Metrocard - Full Fare	31	5	40.75734	-73.95412	POINT (-73.95412 40.75734)
01/03/2025 01:00:00 PM	tram	TRAM1	RI Tramway (Manhattan)	Manhattan	metrocard	Metrocard - Seniors & Disability	4	1	40.761337	-73.96416	POINT (-73.96416 40.761337)
01/03/2025 02:00:00 PM	tram	TRAM2	RI Tramway (Roosevelt)	Manhattan	metrocard	Metrocard - Unlimited 30-Day	9	0	40.75734	-73.95412	POINT (-73.95412 40.75734)
01/03/2025 03:00:00 PM	tram	TRAM1	RI Tramway (Manhattan)	Manhattan	metrocard	Metrocard - Unlimited 30-Day	9	0	40.761337	-73.96416	POINT (-73.96416 40.761337)
01/03/2025 03:00:00 PM	tram	TRAM1	RI Tramway (Manhattan)	Manhattan	metrocard	Metrocard - Unlimited 7-Day	86	0	40.761337	-73.96416	POINT (-73.96416 40.761337)
01/03/2025 04:00:00 PM	tram	TRAM1	RI Tramway (Manhattan)	Manhattan	metrocard	Metrocard - Fair Fare	6	1	40.761337	-73.96416	POINT (-73.96416 40.761337)
01/03/2025 04:00:00 PM	tram	TRAM1	RI Tramway (Manhattan)	Manhattan	metrocard	Metrocard - Unlimited 30-Day	11	0	40.761337	-73.96416	POINT (-73.96416 40.761337)
01/03/2025 06:00:00 PM	tram	TRAM1	RI Tramway (Manhattan)	Manhattan	metrocard	Metrocard - Seniors & Disability	6	2	40.761337	-73.96416	POINT (-73.96416 40.761337)
01/03/2025 09:00:00 PM	tram	TRAM1	RI Tramway (Manhattan)	Manhattan	metrocard	Metrocard - Full Fare	17	5	40.761337	-73.96416	POINT (-73.96416 40.761337)
01/03/2025 09:00:00 PM	tram	TRAM2	RI Tramway (Roosevelt)	Manhattan	metrocard	Metrocard - Fair Fare	1	0	40.75734	-73.95412	POINT (-73.95412 40.75734)
01/03/2025 10:00:00 PM	tram	TRAM1	RI Tramway (Manhattan)	Manhattan	metrocard	Metrocard - Fair Fare	3	2	40.761337	-73.96416	POINT (-73.96416 40.761337)
01/03/2025 11:00:00 PM	tram	TRAM2	RI Tramway (Roosevelt)	Manhattan	metrocard	Metrocard - Unlimited 30-Day	1	0	40.75734	-73.95412	POINT (-73.95412 40.75734)

Remove irrelevant columns in Data-Preprocessing.ipynb

Load into ArcGIS Pro, generating a Congestion Relief Zone feature layer so we could add our inside_zone dummy.



Export tables for Data-Postprocessing.ipynb, aggregating ridership by day and station and adding the in_2025 dummy.

Vertically combine the 2024 and 2025 datasets (each station has an entry for every day from 1-1-2024/5 to 4-3-2024/5

Final Data

manhattan_subways

	transit_date	station_complex_id	station_ridership	inside_zone	in_2025	day
0	1/1/2024	157	2171	0	0	0
1	1/1/2024	614	29653	1	0	0
2	1/1/2024	438	2811	0	0	0
3	1/1/2024	223	2965	0	0	0
4	1/1/2024	399	4407	0	0	0
5	1/1/2024	155	2076	0	0	0

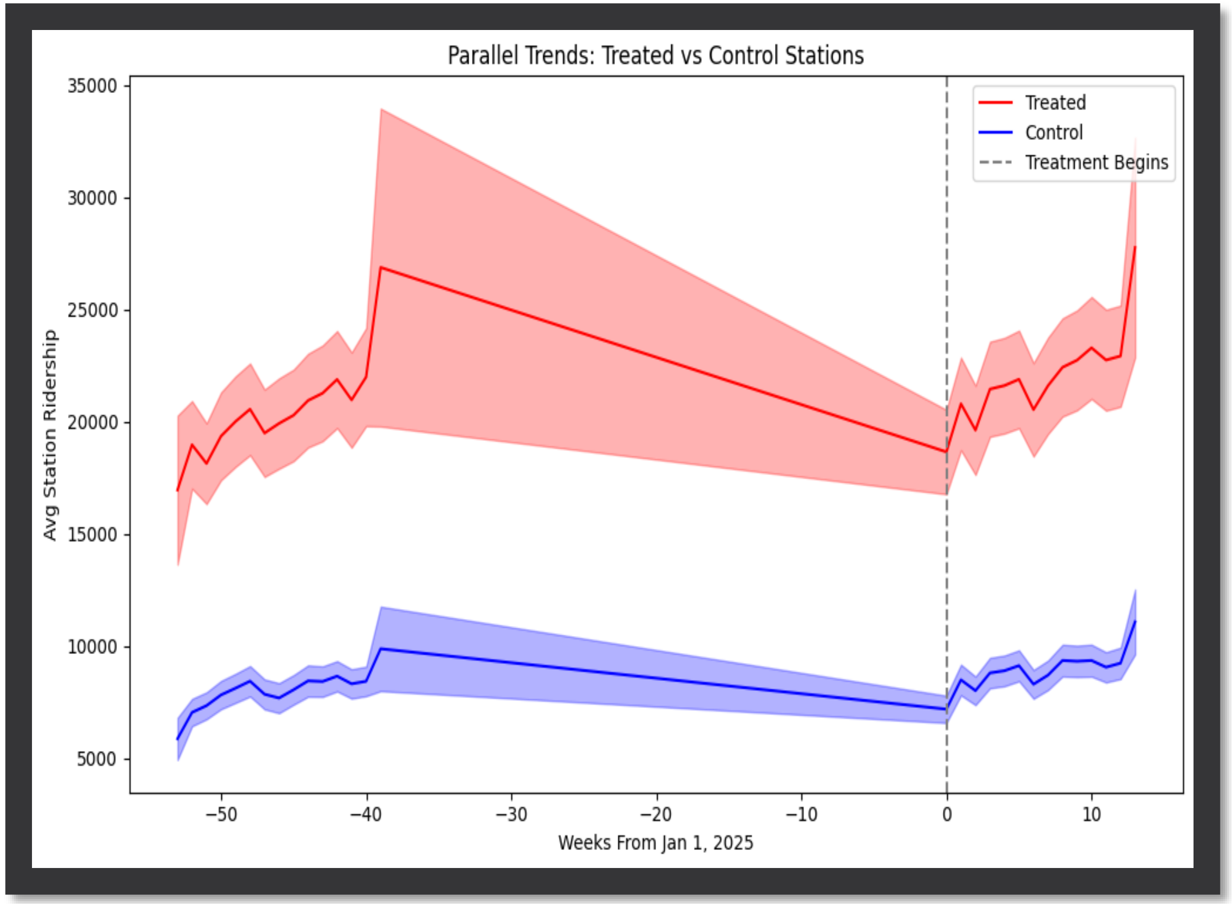
Theory and Model Specification

- Difference in Differences Approach
  - Equation: *Ridership* = B0 + B1 (*Treatment*) + B2 (*Post*) + B3 (*Treatment * Post*) + ε
    - *Ridership*: Number of riders at a given subway station on a given day
    - *Treatment*: Dummy variable — 1 if the station is inside the congestion pricing zone; 0 otherwise
    - *Post*: Dummy variable — 1 if the date is after policy enactment (January 5, 2025); 0 otherwise
    - *Treatment * Post:* Interaction term; equals 1 only for treated stations after the policy was enacted
    - ε: Error term, capturing unobserved influences on Ridership.
  - Our model Estimates whether congestion pricing caused a differential ridership change in treated vs. control stations.
    - Testing significance of interaction term to identify policy effect

Key Assumptions

- Parallel Trends:
  - Pre-treatment trends in ridership are parallel between treated and control stations
- No spillover effects:
  - Ridership in control areas isn't indirectly affected by the policy (e.g., shifting to nearby stations outside the zone)

- No simultaneous policies



Plotting pre-treatment ridership trends for treated and control stations verifies our parallel trends assumption

- Note: Estimation uses OLS, but causal interpretation depends on the assumptions above

Challenges

- Not Addressed
  - Potential spillover effects
    - Individuals may have switched their subway station/ transportation methods in anticipation of the policy, and this cannot be easily observed and isolated.
  - Confounding Policies
    - Although there is no significant other policy changes, potential confounding policies or shocks unrelated to congestion pricing (e.g., changes in subway service, new station openings) could bias our estimates if they differentially affected treated and control areas.
- Addressed
  - Time-invariant differences between stations
    - Our third model includes station fixed effects to control for time-invariant differences between stations, such as size, location, and baseline ridership levels

Model 1 – Simple DID

- Bad model
  - No station controls
  - No time controls
  - Pooled OLS

```
. reg station_ridership inside_zone in_2025 inter

      Source |       SS           df       MS      Number of obs   =    22,614
-------------+----------------------------------   F(3, 22610)     =   1022.52
       Model |  8.9682e+11         3  2.9894e+11   Prob > F        =    0.0000
    Residual |  6.6102e+12    22,610   292356479   R-squared       =    0.1195
-------------+----------------------------------   Adj R-squared   =    0.1193
       Total |  7.5070e+12    22,613   331977289   Root MSE        =     17098

------------------------------------------------------------------------------
station_ri~p | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
 inside_zone |   12255.84   321.2814    38.15   0.000      11626.1    12885.57
     in_2025 |   785.2321   331.3586     2.37   0.018     135.7464    1434.718
       inter |   616.9946   455.5815     1.35   0.176    -275.9765    1509.966
       _cons |   8033.623    233.634    34.39   0.000     7575.685    8491.562
------------------------------------------------------------------------------
```

Model 2 – DID w/ Standard Error Clusters

- Improved
  - Cluster S.E. by station ("busy stations are likely to be busy tomorrow"), this assumes that errors are independent with respect to the station
  - Still doesn't include fixed effects

```
. reg station_ridership inside_zone in_2025 inter, vce(cluster station_complex_id)

Linear regression                               Number of obs     =     22,614
                                                F(3, 120)         =      20.66
                                                Prob > F          =     0.0000
                                                R-squared         =     0.1195
                                                Root MSE          =      17098

                        (Std. err. adjusted for 121 clusters in station_complex_id)
------------------------------------------------------------------------------
             |               Robust
station_ri~p | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
 inside_zone |   12255.84   2702.254     4.54   0.000     6905.561    17606.11
     in_2025 |   785.2321   166.0119     4.73   0.000       456.54    1113.924
       inter |   616.9946   281.7883     2.19   0.030     59.07326    1174.916
       _cons |   8033.623    799.675    10.05   0.000     6450.323    9616.924
------------------------------------------------------------------------------
```

Model 3 – Fixed Effects

- Best
  - Fixed effects to control for time invariant characteristics of stations (e.g. holidays, planned construction)
  - Also controls for entity invariant characteristics (e.g. stations near Times Square versus residential areas, stations close to schools/businesses)
  - Average Treatment Effect: 641
    - The congestion pricing policy led to a predicted increase in 641 riders per day per station inside the zone

```
. xtreg station_ridership in_2025##inside_zone i.date_num, fe vce(cluster station_complex_id)
note: 1.inside_zone omitted because of collinearity.
note: 23834.date_num omitted because of collinearity.

Fixed-effects (within) regression               Number of obs     =     22,614
Group variable: station_co~d                    Number of groups  =        121

R-squared:                                      Obs per group:
     Within  = 0.3634                                         min =        183
     Between = 0.1267                                         avg =      186.9
     Overall = 0.0483                                         max =        187

                                                F(120, 120)       =          .
corr(u_i, Xb) = 0.0147                          Prob > F          =          .

                          (Std. err. adjusted for 121 clusters in station_complex_id)
------------------------------------------------------------------------------------
                   |               Robust
 station_ridership | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------------+----------------------------------------------------------------
         1.in_2025 |    11196.6   1109.322    10.09   0.000      9000.22    13392.98
     1.inside_zone |          0  (omitted)
                   |
in_2025#inside_zone|
               1 1 |   640.5835   287.6305     2.23   0.028      71.0952    1210.072
```

Conclusion

- When using stronger models that better control for variables like the differences in stations, we find statically significant evidence that the congestion pricing does have a positive effect on public transit ridership inside the congestion pricing zone.
- Based on evidence from both our own data analysis, backed up by a review of other literature, the congestion pricing policy in large cities seems to have had its intended effects on improving support for the public transit system as well as improving the roads themselves.
- Future research could see us directly comparing congestion pricing in NYC to congestion pricing of other cities or compare to other methods of traffic control (for example parking fees in San Francisco).