MTA Turnstile Exploratory Data Analysis

A Fair Fare

Metis Bootcamp
Project 1
Brandon McNeil

The Fight Against Unfair Fare Hikes:

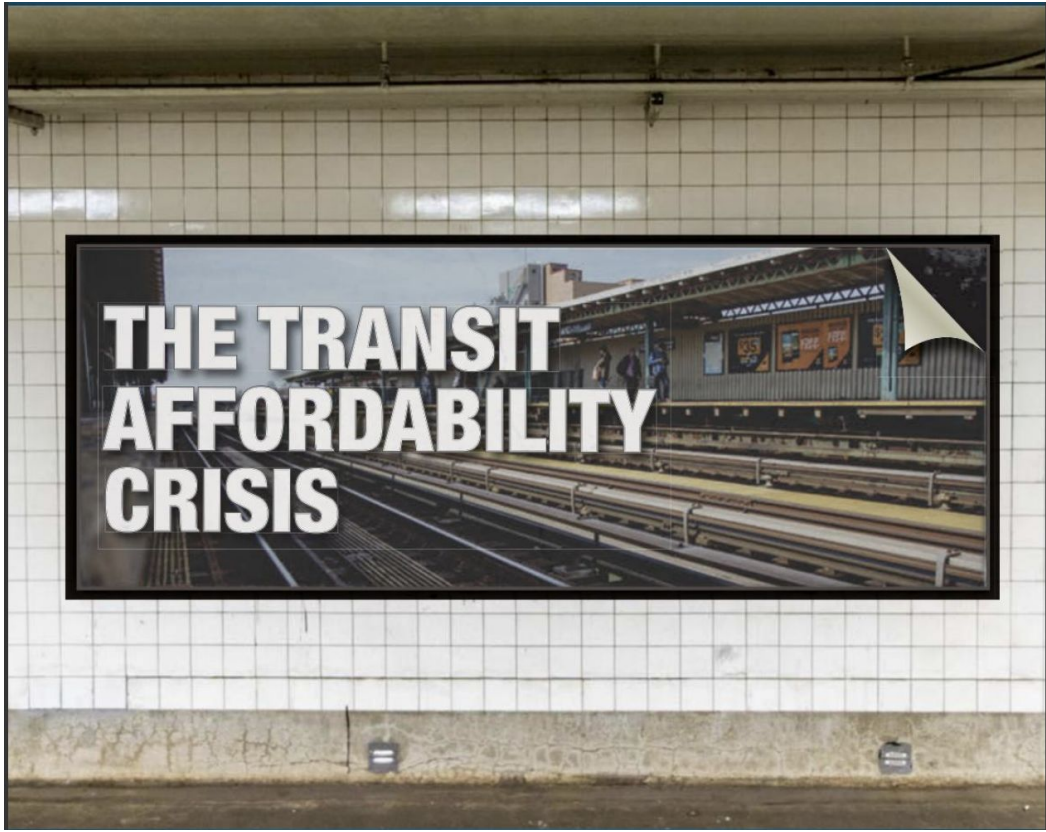
The Issue:

The MTA is planning on raising the Subway fare from the current $2.75 to $3.25 per ride.

Over the course of a year, that additional fare could end up costing the average commuter an additional $300 a year.

Our Analysis:

Examine what information we can extract from MTA turnstile data, as well as other supplementary datasets, to find which New Yorkers would be most affected by a proposed fare hike.



https://issuu.com/cssnyorg/docs/the_transit_affordability_crisis_fi

The Data!

Data Sources:

MTA Turnstile Data (Dec2018-Mar2019, & Dec2020-Mar2021)): http://web.mta.info/developers/turnstile.html

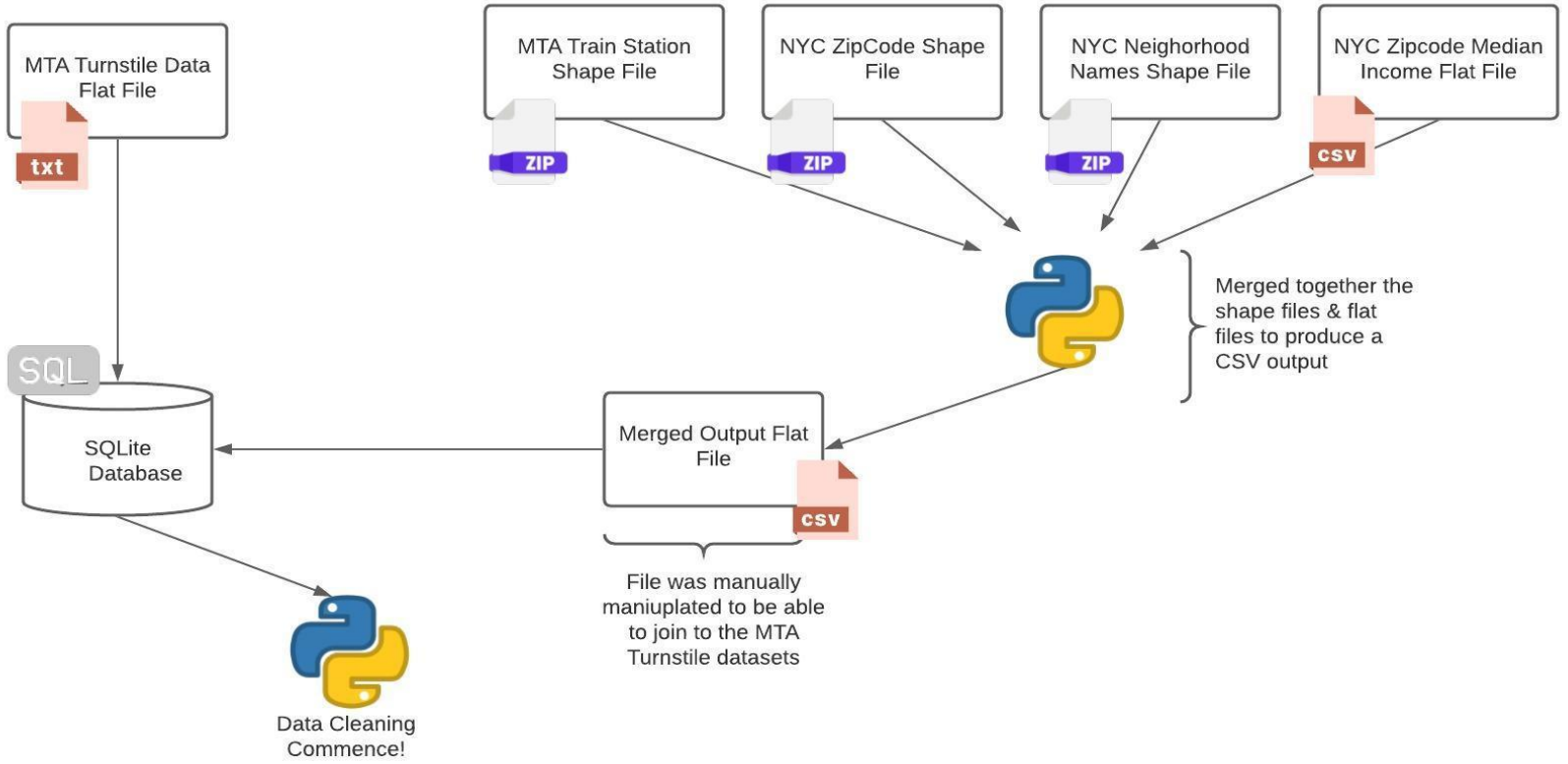
NYC Zip Code Boundary: https://data.cityofnewyork.us/Business/Zip-Code-Boundaries/i8iw-xf4uZipCode

NYC Neighborhood-Tabulation-Areas: https://data.cityofnewyork.us/City-Government/Neighborhood-Tabulation-Areas-NTA-/cpf4-rkhq

NYC Subway Station Map: https://data.cityofnewyork.us/Transportation/Subway-Stations/arq3-7z49

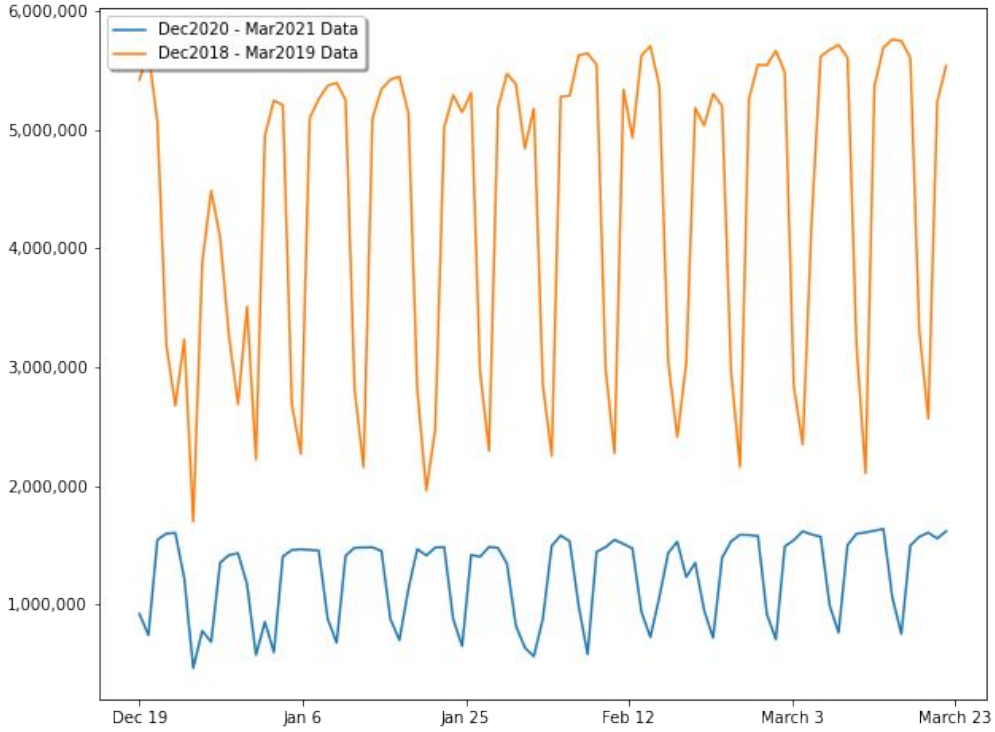
NYC Median Income by Zip Code: https://data.cccnewyork.org/data/map/66/median-incomes#66/39/6/107/62/a/a

Workflow Process:

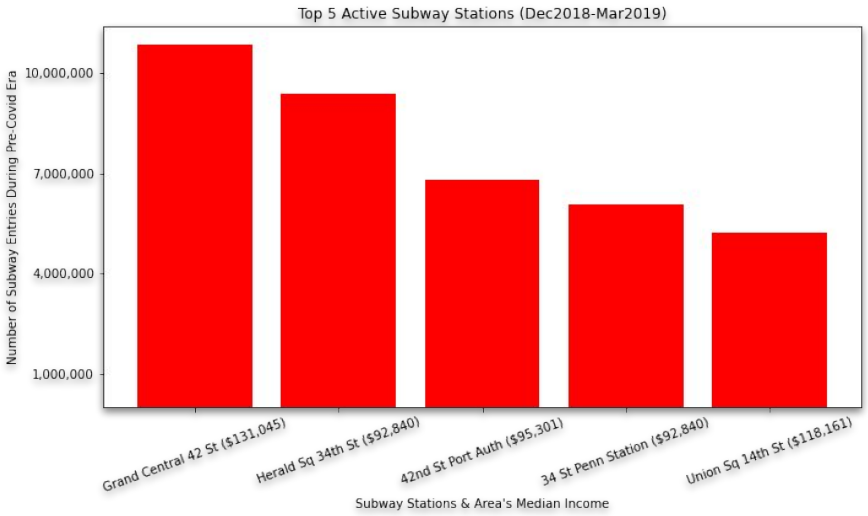


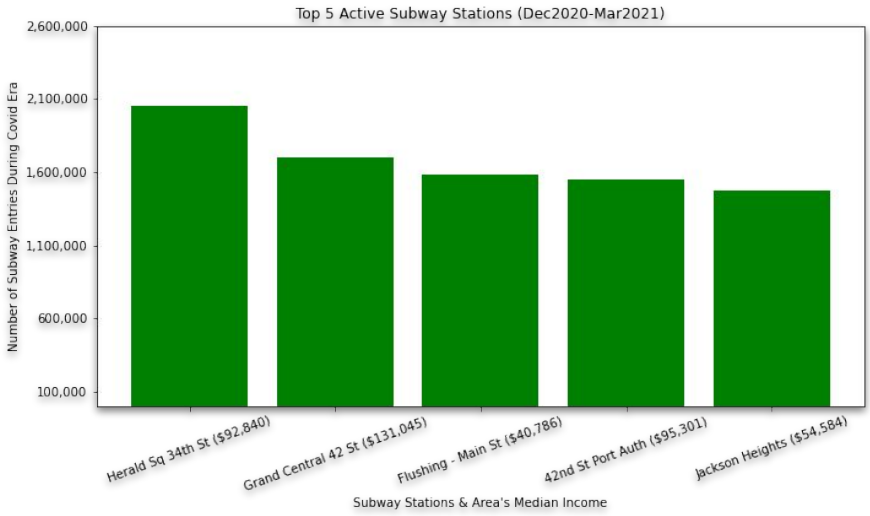
The Findings!

It is no surprise that current train usage has dropped dramatically!



Top 5 Active Subway Stations:





Pre-Covid Results:
Expected! Top 5 are all major hubs for transportation, tourism, and commerce.

During Covid Results:
Rise of 2 outer borough (not Manhattan) train stations in the Top 5.

Calculating Rider Usage % Compared to Pre-Pandemic Levels:

Method:

1. Aggregate Station's Entry Totals by Neighborhood Name (many Stations to 1 Neighborhood Name).

2. Divide Current Neighborhood Entries to Pre-Covid Station Entries.

Observations:

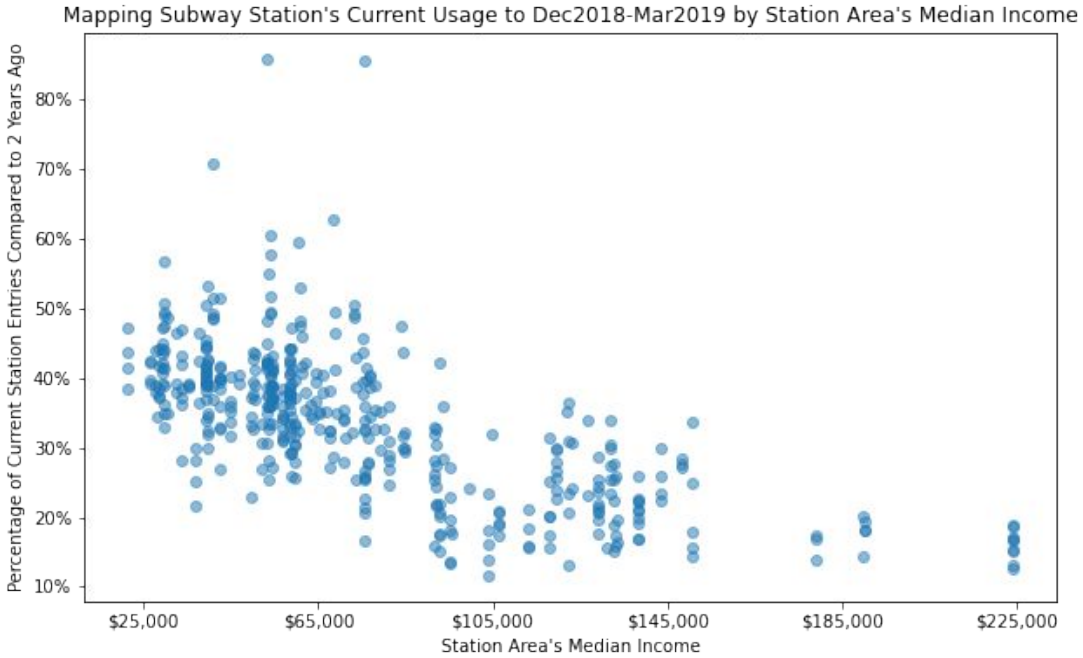
1. Outer-boroughs are experiencing higher "return to normal" percentages than the major transportation hubs.

Current Entries / PreCovid Entries = "Return to Normal" %

Neighorhood Name	Entries_Current	Entries_PreCovid	Return_to_Normal_Percentage	BoroName
Richmond Hill	233655.0	386563.0	0.60	Queens
Borough Park	454667.0	838559.0	0.54	Brooklyn
North Corona	1204759.0	2396372.0	0.50	Queens
Woodhaven	389651.0	789230.0	0.49	Queens
Jackson Heights	1139264.0	2379487.0	0.48	Queens
West Farms-Bronx River	228476.0	487069.0	0.47	Bronx
West Brighton	458720.0	1009759.0	0.45	Brooklyn
West Concourse	1491831.0	3338077.0	0.45	Bronx
Fordham South	356742.0	800243.0	0.45	Bronx
Bensonhurst East	619220.0	1374120.0	0.45	Brooklyn
Woodlawn-Wakefield	273432.0	617880.0	0.44	Bronx
Ocean Parkway South	167067.0	376796.0	0.44	Brooklyn
Mount Hope	628152.0	1438316.0	0.44	Bronx
Eastchester-Edenwald-Baychester	137523.0	321717.0	0.43	Bronx
Norwood	750124.0	1737511.0	0.43	Bronx

Top 15 Results of DataFrame

Plotting Rider Percentage and Station Area's Median Income:



Stations in lower median income areas are seeing a higher return to normal percentage than higher median income areas.

Commuters in these areas may be essential workers, or workers that do not have a work from home option available.

Further Research Required!

Conclusion:

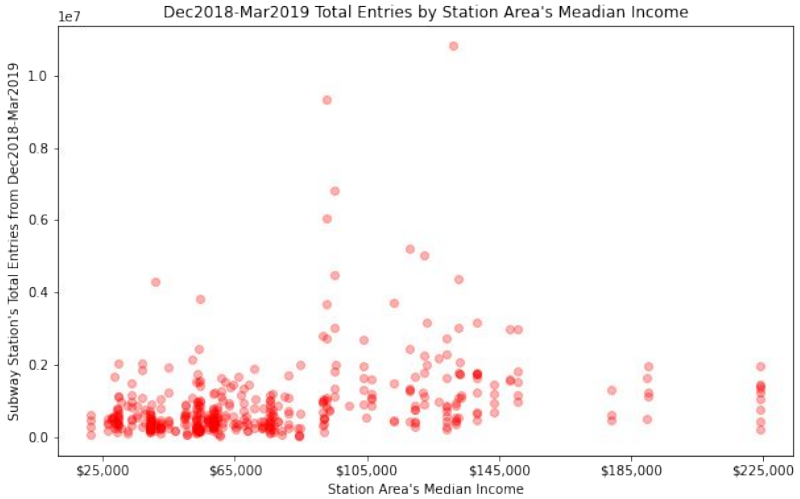
Major Train Hubs are not returning to normal ridership levels as quickly as the outer boroughs.

While stations like 34th St Penn Station & 42nd St Grand Central are large revenue sources for the MTA, the fare raise can disproportionately impact working class commuters who are returning to their regular train habits - but more work will be required!

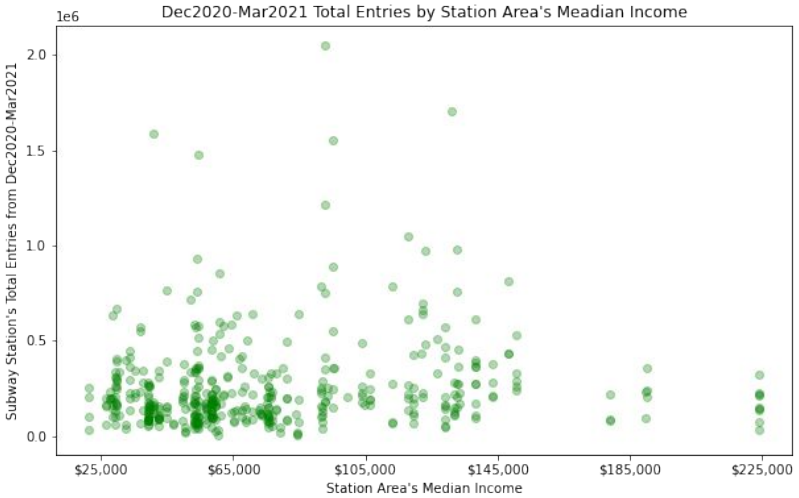
Any proposed fare hike will burden on working class folks who work are already paying their fair share of fares.

Appendix:

Pre-Covid Total Entries per Station by Station Area's Median Income



Covid Total Entries per Station by Station Area's Median Income



Geospatial Joining!

Using geopandas library, I was able to merge two dataset together not by the traditional unique key, but by a physical location using lat / long coordinates.

```
In [38]: joined = geopandas.sjoin(nycnta, mta_stations, op = 'contains', how='left')
         joined.info()

<class 'geopandas.geodataframe.GeoDataFrame'>
Int64Index: 543 entries, 0 to 194
Data columns (total 14 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   BoroCode     543 non-null    int64
 1   BoroName     543 non-null    object
 2   CountyFIPS   543 non-null    object
 3   NTACode      543 non-null    object
 4   NTAName      543 non-null    object
 5   Shape_Leng   543 non-null    float64
 6   Shape_Area   543 non-null    float64
 7   geometry     543 non-null    geometry
 8   index_right  473 non-null    float64
 9   OBJECTID     473 non-null    float64
 10  NAME         473 non-null    object
 11  URL          473 non-null    object
 12  LINE         473 non-null    object
 13  NOTES        473 non-null    object
dtypes: float64(4), geometry(1), int64(1), object(8)
memory usage: 63.6+ KB
```