

**NEW YORK UNIVERSITY**

**Text as Data(DS - GA 1015)**

**Spring Semester, 2024-25**

***The Impact of Viral & Polarizing Discourse about the Subway***

***On Subway Ridership***

**Group Members:**

**Lavya Midha (lm5465)**

**Shashank Pandey (sp8108)**

**Introduction**

The New York City subway system is one of the largest, most popular and complex public transit networks in the world. It serves thousands if not millions of commuters each day. It is very well known for its efficiency and connectivity, but is also notoriously famous for its delays and safety concerns on a daily basis. In the era of social media, there is a lot of conversation about how the public feels about the MTA.

This paper explores the relationship between social media discourse and overall ridership trends in New York City, aiming to understand if social media discourse, particularly the top posts about the MTA everyday tend to affect the subway ridership. In a sentence, the paper aims to answer the question, *“ Does public perception, as reflected on social media, influence how people move through the city?”*

We find that there is a consistent negative effect on ridership in the aftermath of high-volume and high-polarity discourse. We also find that this negative effect is seen across different areas in the city, and even across all five boroughs.

**Motivation**

Understanding social media discourse about a topic has become increasingly important in today’s era, where platforms like Twitter have become primary channels through which individuals share information and reactions to events in real time. The subway is the primary method of transportation for a majority of people living in the New York City Metropolitan Area; any delays or discourse around the subway widely affects their daily lives.

While getting information through survey statistics and other datasets may be useful, text data coming from social media discussions has become increasingly important because it gives us a real-time raw idea about public emotions, and the virality of said conversations, directly from the source. Therefore, looking at text data here seems to be the best approach to understand the relationship between public sentiment and subway ridership.

**Data Collection**

To answer the questions we posed in the motivation for our project, as defined above, we needed two datasets:

* A corpus of public discussion about the New York City subway that includes engagement metrics, enabling us to identify dates with both high-volume and highly charged commentary.
* A dataset recording daily subway ridership, ideally per-station, so we could analyze whether these high-impact dates had an effect on ridership, and, if they did, whether there were differing effects based on borough/location.

For the first dataset, we explored several possible data sources to gather a complete corpus for analysis. Since the New York City Subway is an extremely popular topic of discourse, we identified two possible types of text data for our project.

One possible source of text data would be **from social media**, including content from platforms like X/Twitter, which would include tweets and replies about the subway. The advantage of this approach is that we would be able to get reflections of real-time sentiment (from people who would be airing their opinion in the moment), along with several metrics that would indicate the popularity and engagement of these opinions (through likes, comments). The one caveat of this data collection approach was that there could be a lot of noise in this data, and hence, we would need to clean the data carefully.

The other source of text data we considered was from **traditional media**, including comments on platforms like The New York Times on their articles about the subway. The advantage of this approach is that we would be more likely to get well-formed opinions, which would be easier to analyze. However, this data could end up having self-selection biases (as not all people are inclined to write well-formed comments on these platforms).

Because of the above reasons, we decided to use data collected from social media. For uniformity and consistency, we decided to collect data only from Twitter.

Since Twitter’s API did not allow us to gather the data without running into rate limit complications, we decided to use Selenium to scrape the platform. We adapted an open-source script to fetch the ‘top 100’ tweets of each day, as ranked by Twitter’s internal recommendation system. We decided to scrape the top 100 tweets for each day, so that we would have a decently representative dataset (≈36,000 tweets), and we would be able to compare the levels of activity across all days.

For reproducibility, we are attaching the exact structure of the keyword search we automated on Twitter to gather the top 100 tweets of each day of the year:

**"MTA" OR "NYC subway" OR "New York train" OR "New York subway" OR "NYC train"**

For the second dataset, we used New York State’s ‘MTA Subway Hourly Ridership’ dataset, publicly available on the NY Open Data website. We filtered this data to only include tweets from December 2023 to January 2025 (as we wanted a 14-day margin of ridership data to possibly analyze changes in ridership in the first few/final days of 2024). We gathered relevant columns from this dataset, including timestamps, station name, borough, ridership, and geolocation data, and proceeded to pre-processing.

**Data Cleaning and Preprocessing**

To ensure accurate analysis, we first consolidated all collected tweets into a CSV file containing tweet text along with associated metrics such as likes, comments, retweets, and analytics. The text data then went through a series of preprocessing steps for better interpretation for the main analysis. These included converting all text to lower case, removing hashtags, hyperlinks and tags, removing nltk stopwords and removing punctuations.

Additionally, we identified and excluded tweets from accounts unrelated to the Metropolitan Transportation Authority (MTA) that contained keywords such as “MTA” or “NYC Subway,” which could introduce noise. Irrelevant tweets included terms like *[insert examples, e.g., “MTA bus in Canada,” “NYC Subway sandwich”]* which were removed for better precision. The resulting cleaned dataset was then stored separately for sentiment analysis.

Ridership data, in contrast, had been filtered for relevance during initial collection and therefore required very little preprocessing. We aggregated the hourly ridership data to represent daily ridership.

**Methodology**

*Sentiment Analysis of Tweets*

Our first step after preprocessing was sentiment analysis of the tweets to later see document social media sentiment about the New York City subway. For our sentiment analysis two types of frameworks were used: one which followed a rule-based (classical) approach, while another followed a transformer-based approach.

The first framework used was regular VADER scores, which scored each tweet on a scale of -1 and 1. VADER is a rule-based sentiment analysis tool that is specifically designed for analyzing social media texts. It uses a dictionary of words and rules to determine the sentiment of a piece of text, using a valence score for each word to determine its positivity or negativity and aggregating those scores to get an overall sentiment.

For the second framework, we used a transformer-based natural language model created specifically for social media analytics. We utilized a RoBERTa-based sequence classification model, trained by Cardiff NLP, which has been optimized for short-form social media text. The model was set up, and tokenizers were loaded using the *transformers* library. In addition, to optimize performance, Google Colab's T4-GPU environment was used.

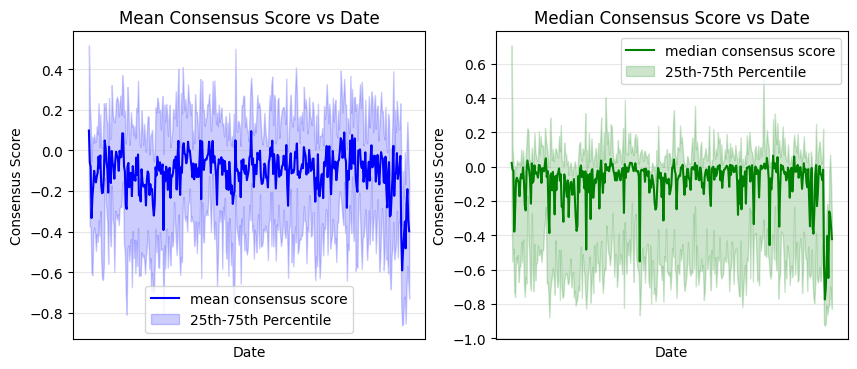
Next, we defined a custom function, *predict\_sentiment,* to tokenize each tweet, run it through the model, and extract the predicted sentiment. The model returns logits for three sentiment classes: *negative, neutral, and positive*, which are converted to probability scores using a softmax function. The function was then iteratively applied to all preprocessed tweets to output the probability scores for each sentiment class.

However, for consistency, we wanted our scores to be in the same range of -1 and 1 (as the VADER scores). For that, we subtracted the negativity score from the positivity score for each tweet (net\_score = positive\_score - negative\_score) and then normalized the score to get the final sentiment scores for this model. We went with this approach as we want the overall score to reflect the polarity of opinion, like the VADER scores.

Finally, to get an overall score, we used the average of our VADER and BERT scores for all tweets. This average allowed us to advantage consensus opinions, and penalize disagreements between the models (so only **unambiguously** strong tweets would have high scores).

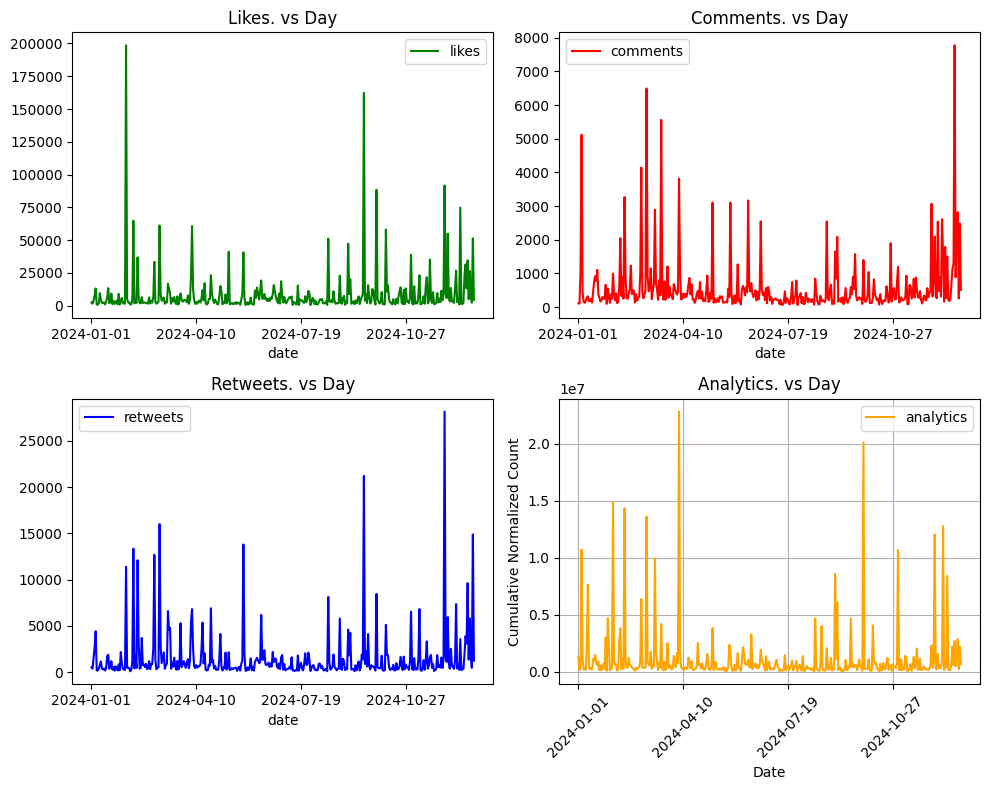
*Identifying Dates with High-Polarity & High-Volume Commentary:*

Having calculated the sentiment scores for each tweet, we now needed to get a measure of the ‘sentiment’ and the ‘polarity’ of the discussion, for each date. We considered using either the mean daily sentiment ( the mean of the combined sentiment scores for all tweets in a day) or the median daily sentiment, and plotted both below. We decided to use the mean consensus score because it is a more accurate metric for dates with strongly polarizing opinions (the median is more likely to blindly represent the majority opinion).



From the above plots, we can see that there are a few dates where we can see that there are noticeable spikes in the consensus score - we identify the top 5 %ile of such dates (for both positive and negative outliers), and store them. These dates are identified as our **‘high polarity +ve/-ve dates’**. We can also see that the consensus score tends to skew negative across most dates, which backs up our initial motivation of study.

Having identified our high-polarity dates, we now move to identifying dates of high activity. For this, we look at the averages of each of the activity metrics (number of likes, comments, retweets, and analytics/views) per day. The plot of these metrics is seen below.



We identify **‘high activity dates’**, as the top 10% of dates in each metric (to ensure that we gather dates that have all kinds of engagement).

Now that we have both high-polarity and high-activity dates, we take the intersection of these dates as the dates we should study. We find that this subset contains 5 dates with highly positive polarity, and 12 dates with highly negative polarity - these disproportionate results are explained by the simple fact that, generally, high activity on a social media platform about something like the subway is likely to happen when something goes wrong (or) something bad happens.

*Identifying what happened on dates where discourse was observed*

After identifying the dates associated with spikes in both positive and negative discourse, we aimed to uncover the underlying causes driving these conversations. To do this, we leveraged YAKE (Yet Another Keyword Extractor), which we hypertuned according to our preferences to improve keyword relevance. This allowed us to efficiently generate the top 20 most frequent bigrams for each day, offering insight into the dominant topics driving the discourse. With these keywords in mind it was easy to look at tweets to understand what happened on the day of discourse.

*Analyzing Impact on Ridership*

Having identified the dates we want to study, we now find the average ridership in the 14 days before and after the date we study. We take a wide window of 14 days, in an attempt to prevent any system events (like short-term repair work, higher ridership on the weekends) from impacting our averages.

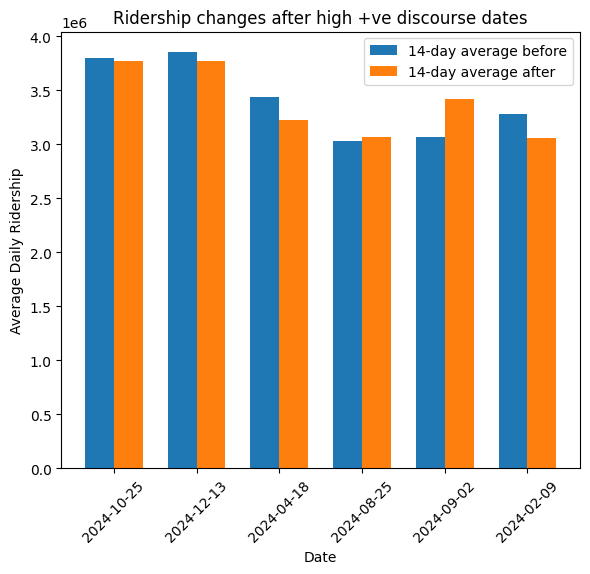
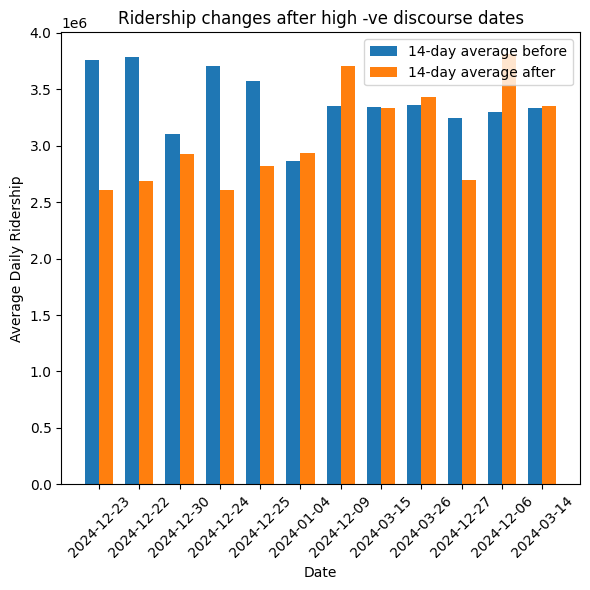
Since we are also interested in the geographically stratified impact on ridership, we also calculate the changes in average ridership for each station in the network, and for each of the city’s five boroughs. This can allow us to see the extent of variations in ridership.

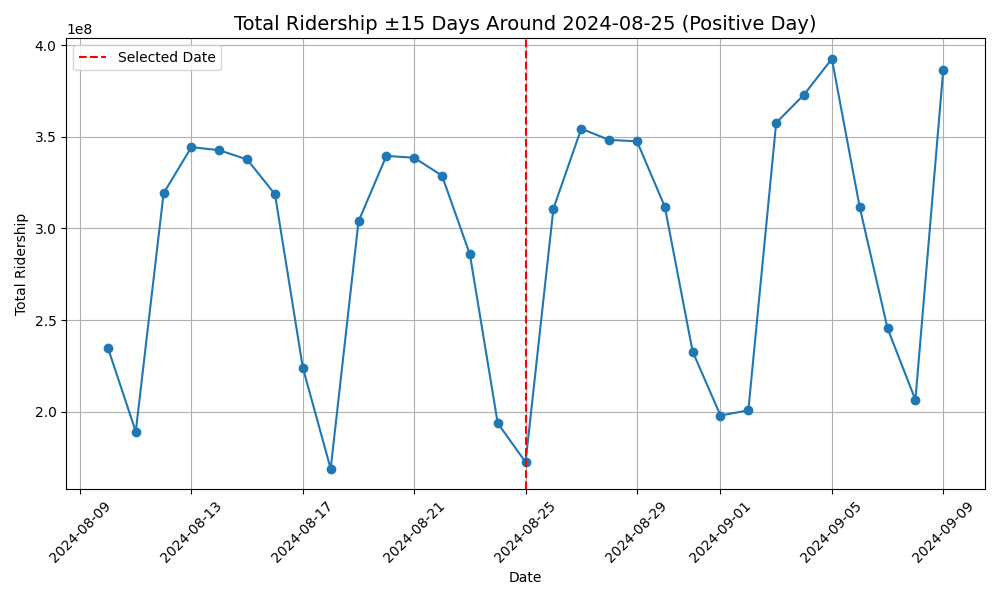
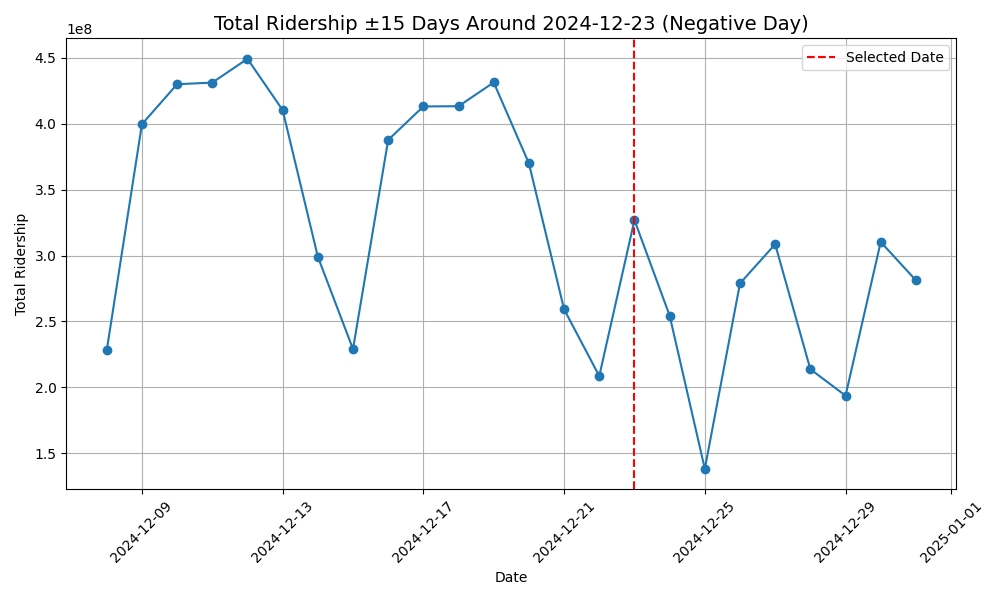
**Results**

Firstly, in addition to our keyword extraction (which extracted the topics of discourse), we wanted to look up whether there were **actual** incidents on the subway on or around the high-polarity dates. We found incidents on nearly all the dates, including:

* 22 December 2024: A woman was burned alive at the Coney Island subway station (coverage of this event dominated discourse until the end of the year)
* 4 January 2024: The 1 train derailed after colliding with a vandalized train, 26 injured.
* 9 December 2024: A jury found Daniel Neely not guilty of causing the death of Jordan Penny on an F train in May 2023, in a controversial verdict.

From the above, we are fairly confident that we have been able to identify dates that had high levels of heated discourse. Next, to study the impact of positive and negative dates on ridership, we plot ridership changes around these dates:

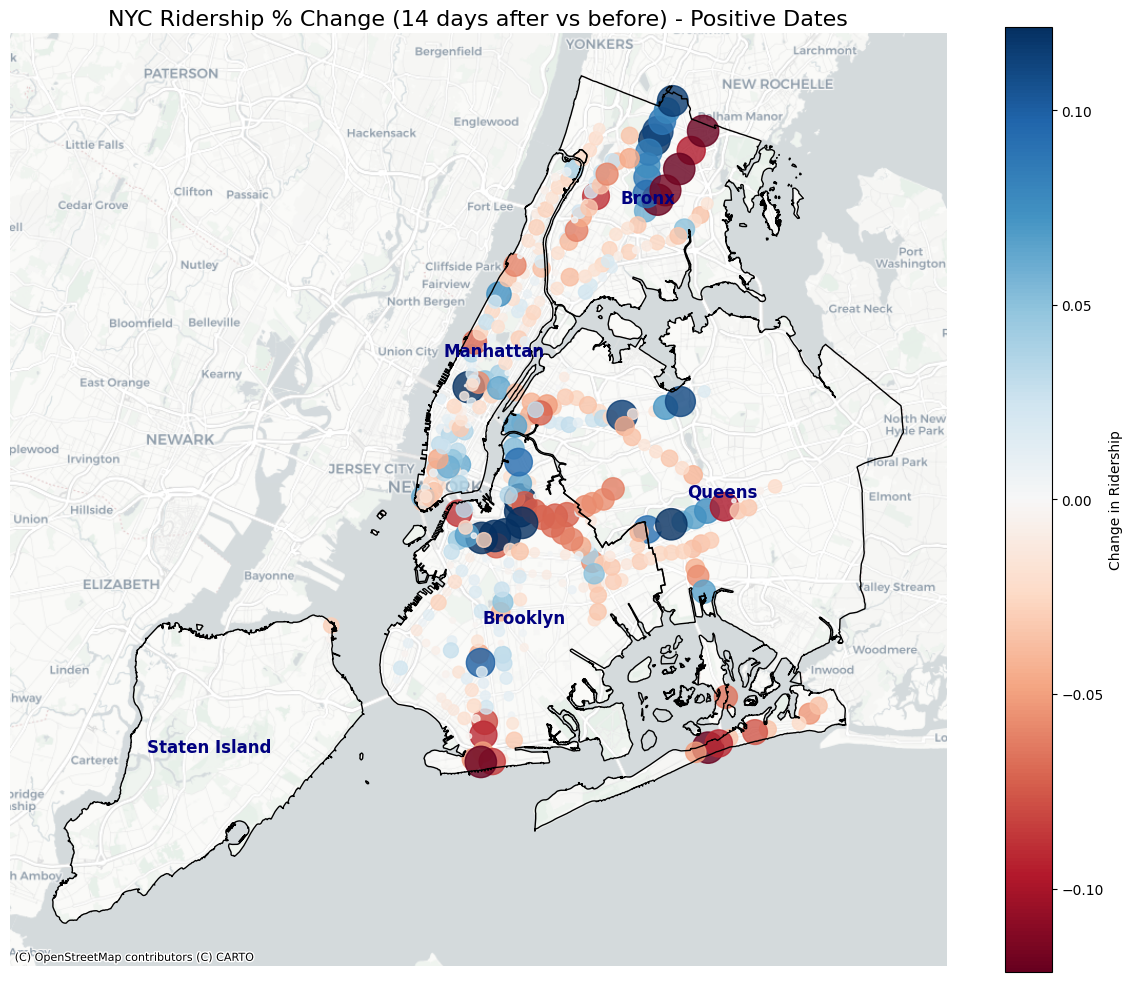
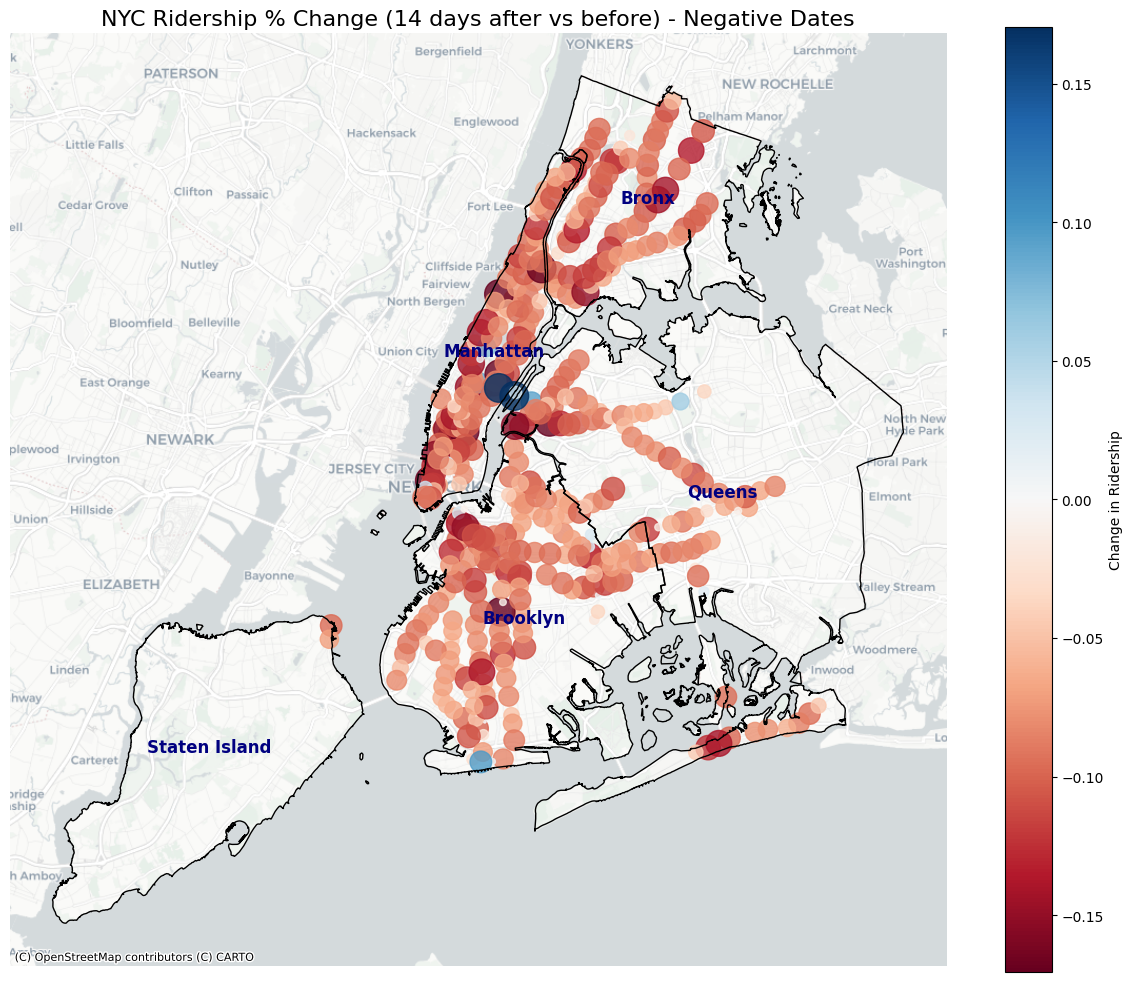
 

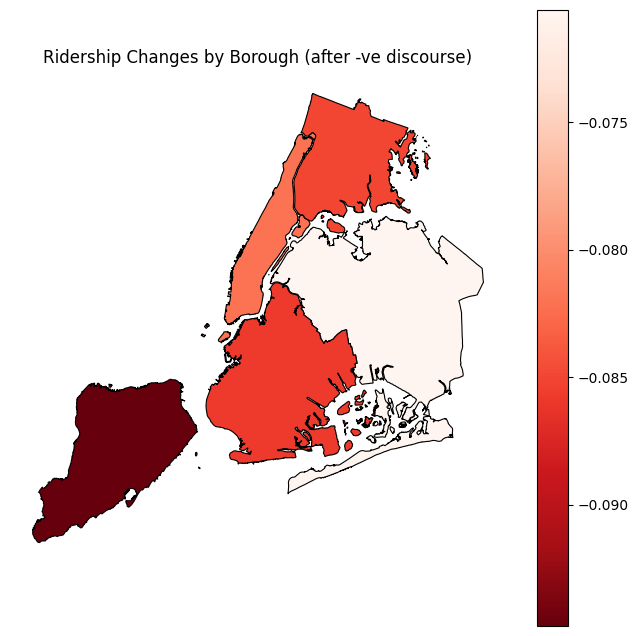
 

We see that, in the aftermath of high +ve discourse dates, there is a mixed impact on ridership (some positive, some negative). There is no strong trend that allows us to predict a clear impact on ridership. However, when we see the aftermath of high -ve discourse dates, we see a clear decline of ridership after most dates, albeit to varying intensities. This allows us to draw our first conclusion: **Viral and negative discourse does indeed negatively affect subway ridership**.

Another aim of our study, as we had mentioned in the motivation, was to see whether there were different impacts on ridership across different parts of the city. We know that the city has high levels of socio-economic inequality (indeed, the city contains both the most billionaires on Earth, and also contains the U.S’s poorest congressional district). These factors could impact the ability of people to change their ridership patterns.

To see this stratified impact, we visualize the change in ridership (per subway station) across all five boroughs, around both positive and negative high-discourse dates.

**** ****

For positive events (which are also relatively rarer), we see mixed impacts on ridership. 

We see, however, that there does not appear to be any geographic impact on ridership in the aftermath of negative events. While we see a negative impact on ridership (as the bar chart earlier had already shown us), this effect is, rather remarkably, seen across almost all subway stations. This leads us to our second conclusion: **the negative impact of viral & negative discourse on ridership is consistently seen across the city.**

After getting the above results, we attempt to see how the impact of ridership varied across boroughs, to find where the negative impact was relatively stronger (and whether such variation existed across boroughs). We generated the following plot.

From the map, we can see that Staten Island sees the highest decrease in ridership (almost 9%), while other boroughs see slightly lower decreases. This difference may be explained by the relative unpopularity of the Staten Island Railway, which is not integrated with the rest of the subway system and serves fewer riders overall.However, we still see a relatively similar decrease in ridership across all five boroughs, supporting our earlier conclusion.

**Conclusion & Takeaways**

From our results, we can arrive at two key conclusions:

* Strongly negative and popular commentary on social media does correlate to a decline in ridership, indicating that, to some extent, people do seek out alternate modes of transportation when the subway suffers from incidents.
* While there is a negative impact on ridership, this impact is seen across all parts of town. From the well-off West Village to the transit deserts in Queens, it appears that all neighbourhoods have a similar negative response to these incidents.

A clear takeaway from our results is that it highlights just how important the appearance of public perception is on subway ridership. When isolated incidents happen that make the subway appear to be unsafe or dangerous, there is a noticeable decrease in ridership. The importance of maintaining this ‘public perception’ is also highlighted by the second conclusion we arrived at above, i.e., that ridership in all regions of the city (rich districts, less well-off regions) suffers when events that cause negative discourse happen. This is likely to disproportionately hurt the less well-off, as they will find it more difficult to arrange alternative transportation.

**Future Steps**

Our research has left several trails for future work to follow. One obvious possibility could be using texts from traditional media (as discussed earlier), to get a more ‘cleaner’ corpus to analyze. Integration could also be done with open-source socio-economic NY State data, which would allow for more granular analysis of whether higher-income residents respond differently to these incidents. Another future diagonal area of study could examine the impact of these incidents on the ridership in the **specific station** where the incident took place.

**References**

1. Das, A., Prajapati, A. K., Zhang, P., Srinath, M., & Ranjbari, A. (2023). *Leveraging Twitter data for sentiment analysis of transit user feedback: An NLP framework*. arXiv. <https://doi.org/10.48550/arXiv.2310.07086>
2. New York State Metropolitan Transportation Authority. (n.d.). *MTA Subway Hourly Ridership: 2020–2024*. State of New York. <https://data.ny.gov/Transportation/MTA-Subway-Hourly-Ridership-2020-2024/wujg-7c2s>
3. Chen, X., Wang, Y., & Di, X. (2023). *Whose attitudes toward transit are most affected by rising subway crimes in New York City?* <https://doi.org/10.1109/ITSC57777.2023.10421971>
4. GitHub User (2023-25). *Selenium-Twitter-Scraper*, GitHub <https://github.com/godkingjay/selenium-twitter-scraper>