Final Memo Starting Point -> Revision 12/4

Questions:

1. How were daily covid cases affected by the initial state of emergency in NYC in March 2020?
2. How were daily covid cases affected when a mask mandate and late night subway closures for cleaning was put in place in April 2020?
3. Was there a noticeable increase of cases during the end of May 2020 when George Floyd protests started?
4. How were daily covid cases affected when the first phase of reopening started happening in June 2020?
5. How were daily covid cases affected when vaccines started becoming available December 2020?
6. How were daily covid cases affected when NYC was fully reopened in June 2021?
7. How were daily covid cases affected when vaccine requirements, mandates, and in person learning was back in full capacity starting August and September 2021, up to the end of vaccine requirements or mandates in private sector in Nov 2022 and for city workers in February 2023?

Additional Information Required:

1. **More details of all of these measures implemented in NYC during the timeframe I'm looking at** -> this can come in the form of many different news articles throughout the years that I have found
2. Compliance rates with lockdown measures across different boroughs
3. **Mobility data to gauge the extent of movement and potential non-compliance** -> There is 3 mobility surveys that were done in 2020 that could be examined further (1 of them being the one you sent, but there's one in May and July 2020
4. **Demographic data for NYC to understand if certain groups were more affected (though main covid dataset has information borough to borough)** -> a census can probably help with this, but it would be harder to break down neighborhood by neighborhood inside a whole borough, and huge generalizations on demographics by borough may have to be made
5. **Subway usages throughout this timeframe** -> I have found this through a data set that has been tracking MTA Daily Ridership since the beginning of March 2020 for everything except LIRR and Metro North (that started April 1st), right before the pandemic, and is still currently being updated
6. Traffic data throughout NYC (in the dataset I speak about right above for daily ridership, it introduces two columns about traffic in bridges and tunnels, and that made me think it may be a smart idea to try and find a dataset that has tracked traffic info throughout NYC)

COVID-19 Dataset Description:

1. **Data Collection:** The data was collected daily, reporting confirmed and probable cases, hospitalizations, and deaths due to COVID-19 -> legitimacy about data collection for any Covid related dataset has to be questioned with the

understanding that some sicknesses/deaths recorded as covid were in fact not actually Covid

2. **Observations/Rows:** The dataset comprises 1158 daily entries
3. **Timeframe:** March 2020 to May 2023
4. **Geography:** The data covers New York City, with breakdowns available by borough
5. **Variables/Columns:** There are 67 columns. Here's a summary of its structure:
    - `date_of_interest`: The date of the data record.
    - `CASE_COUNT`: The count of confirmed cases.
    - `PROBABLE_CASE_COUNT`: The count of probable cases.
    - `HOSPITALIZED_COUNT`: The count of hospitalizations.
    - `DEATH_COUNT`: The count of confirmed deaths.
    - `PROBABLE_DEATH_COUNT`: The count of probable deaths.
    - `CASE_COUNT_7DAY_AVG`: The 7-day average of confirmed cases.
    - Various other columns that provide the same data broken down by borough (Bronx (BX), Brooklyn (BK), etc.) and include 7-day averages for cases, hospitalizations, and deaths.
    - `INCOMPLETE`: A column that may indicate incomplete data for certain records.

Most important columns to answer questions:

- `CASE_COUNT`: To assess the number of confirmed COVID-19 cases reported each day
- `date_of_interest`: To keep track of month and year of cases
- `PROBABLE_CASE_COUNT`: To include probable cases in the analysis for broader understanding of Covid
- `ALL_CASE_COUNT_7DAY_AVG`: To smooth out daily fluctuations and identify broader trends
- The columns above, but for each of the boroughs: To look to see if there is trends in certain boroughs over other

MTA Daily Ridership Dataset Description:

1. **Data Collection:** This dataset provides daily ridership figures for the Metropolitan Transportation Authority (MTA) in New York City-> consider potential fluctuations in data accuracy due to events like service interruptions, weather conditions, or changes in reporting methodology OUTSIDE of just Covid
2. **Observations/Rows:** The dataset comprises 1352 daily entries
3. **Timeframe:** March 2020 to November 2023
4. **Geography:** This dataset specifically focuses on New York City, providing insights into transit patterns across various boroughs and transit lines
5. **Variables/Columns:** There are 15 columns. Here's a summary of its structure:
    - `Date:` The date of travel (MM/DD/YYYY).
    - `Subways: Total Estimated Ridership`: The daily total estimated subway ridership

- Subways: % of Comparable Pre-Pandemic Day: The daily ridership estimate as a percentage of subway ridership on an equivalent day prior to Covid 19
- Same type of columns for Buses, LIRR, Metro North, Access-A-Ride, and Staten Island Railway
- Bridges and Tunnels: Total Traffic: The daily total bridges and tunnels traffic. Blank values mean it wasn't available or applicable that day
- Bridges and Tunnels: % of Comparable Pre-Pandemic Day: The daily total traffic as a percentage of total traffic on an equivalent day prior to Covid 19

Most important columns to answer questions:

- For this data, all the columns are important, as everything is related to forms of transportation and general traffic in going across boroughs, with a percentage against pre-covid numbers. I will probably focus on the main two forms of public transportation: **subways and buses.** It's also interesting to see traffic in bridges and tunnels, and I wonder if there is data that can break down something like this further to help further see how compliant people were.

Putting this together in R:

- Loaded the 2 datasets
- Downloaded some libraries to do correlation analysis (dplyr, ggplot2, lubridate)
- Convert dates to Date format in datasets
- Combined Case_Count and Probable_Case_Count into one column
- Identified all columns that contained the word "Ridership" in them, replacing all NA values with zeroes to avoid NA value when adding all these up
- Summing up total ridership for each day and making a column for that
- Merging both datasets on the date
- Calculating correlation, where I got a very weak positive correlation of 0.03
- Plotted the correlation with a ggplot, showing absolutely no linear pattern
- Converted ridership data to long format
- Merged the Covid data and the Ridership long data by the data
- Made a color mapping to set colors to each of the 5 transportation types
- Plot with facets for each transportation type, showing ridership trend over time and COVID-19 case counts
- Plot with facets for each transportation type by only looking at data and ridership_count for each transportation type on a given day
- Create a time series plot with log transformed total cases and total ridership irrespective of type of transportation

Interpretation and Conclusions:

- The comprehensive analysis of public transportation ridership and daily COVID-19 cases in NYC presents a counterintuitive narrative to common perceptions about the pandemic's spread. Despite initial assumptions that densely packed transportation might significantly contribute to COVID-19 case surges, the data does not support this hypothesis. Instead, the slight decrease in ridership during major COVID-19 spikes indicates that factors other than transportation usage are at play.

- My argument, supported by the very weak positive correlation observed, suggests that the primary driver of COVID-19 case fluctuations was not public transportation but rather large gatherings and close-knit group interactions. This perspective is bolstered by the observation that significant case spikes were not aligned with increases in ridership. In fact, during these critical periods, there was a noticeable, albeit small, reduction in transportation use, which contradicts the notion that transportation was a central vector for the virus's spread
- In the two most notable case surges, the data reveals a dissociation between public transportation ridership, which continued a gradual recovery towards pre-pandemic numbers, and COVID-19 cases, which exhibited significant volatility with spikes and downturns. This pattern aligns with the interpretation that while public transit plays a role in urban mobility, its impact on pandemic dynamics is overshadowed by other social and environmental factors
- As we progress into 2023, the continuous downtrend in COVID-19 cases amidst recovering ridership levels further solidifies the argument. It underscores the multifaceted nature of pandemic spread mechanisms, where direct correlations with single variables such as public transportation usage may not capture the full complexity of transmission pathways
- Going forward, the analysis would benefit from a closer examination of community spread dynamics, potential superspreader events, and variations in policy adherence across different populations. Diving deeper into the demographic and geographic data may shed light on the nuances of how and where the virus spread most effectively. Moreover, incorporating mobility surveys and traffic data could refine our understanding of the interplay between movement patterns and case trends
- This investigation underscores the importance of considering a broad spectrum of societal behaviors in pandemic modeling and response strategies. It also calls for careful scrutiny of data quality and a rigorous approach to interpreting trends and correlations within the context of complex real-world phenomena