1. **Motivation/problem statement:**

I’m interested in seeing how human mobility has changed over the course of the COVID-19 pandemic because this could give more context into the upticks in cases during certain periods of time as we have seen in the common analysis. By studying where people have been going, we may be able to pinpoint areas which have high transmissibility of the virus. It is also interesting to see the percentage change of destination visitations as restrictions loosen and I am interested to see which destinations people started going to most after the mask mandate has lifted. This is a human-centered problem since we are looking at how people have changed their traveling habits during the quarantine period, mask mandate, and the period where there are no restrictions, and making inferences based on the timing of the events. By studying human mobility, this would help give more insight to people’s perception of risk and this can be helpful in behavioral economics if a similar situation like COVID-19 were to arise again.

1. **Research questions and/or hypotheses:**

There are several research questions that I would like to investigate that will be built upon each other. The first question that comes to mind is how has each phase of the pandemic affected how much people go to areas tagged as retail or recreation? After looking into this initial question, I would like to investigate how retail or recreation visitations compare to parks and residences. I would also like to do an analysis to compare all the destinations if time allows. A hypothesis that comes to mind is that people stayed within their residences the most at the start of the pandemic with occasional visits to grocery stores and pharmacies, during the warmer months, people go to parks more in 2020 than in 2021 and minimize traveling to retail or recreational destinations because of the mask mandate. After the mask mandate has lifted, more people utilize transit stations and go to workplaces/recreation/retail more often as well as leaving their residences more often. Some more research questions and hypotheses are listed below:

**Research Question:**

Is there a difference in mean percentage change for location X during June 2020 vs. June 2021?

**Null Hypothesis:** There is no difference in mean percentage change for location X during June 2020 vs. June 2021.

**Alternate Hypothesis:** There is a significant difference D in mean percentage change for location X during June 2020 vs. June 2021.

**Research Question:** How much effect did the COVID-19 pandemic have on location X’s visitations?

**Null Hypothesis:** COVID-19 had no effect on location X’s visitations.

**Alternate Hypothesis:** COVID-19 yielded a -10% on location X’s visitations.

1. **Data to be used**:

The additional data being used is the [Google COVID-19 Community Mobility Reports](https://www.google.com/covid19/mobility/). The data charted movement trends over time by geography, across different categories of places such as retail and recreation, groceries and pharmacies, parks, transit stations, workplaces, and residential. We will be looking specifically at Cuyahoga County, Ohio, but the data is available on a global and regional scale. The data reported has been continuously updated up until October 15, 2022. This dataset is intended to help remediate the impact of COVID-19. It shouldn’t be used for medical diagnostic, prognostic, or treatment purposes. It also isn’t intended to be used for guidance on personal travel plans. Location accuracy and the understanding of categorized places varies from region to region, so it is not recommended to use this data to compare changes between countries, or between regions with different characteristics (e.g. rural versus urban areas). Changes for each day are compared to a baseline value for that day of the week:

The baseline is the median value, for the corresponding day of the week, during the 5-week period Jan 3–Feb 6, 2020 so the percentage change would be based off of the mobility during this time period. Data is collected from users who have opted in to location history through their Google accounts. No personally identifiable information, like an individual’s location, contacts or movement, is made available at any point. Insights in these reports are created with aggregated, anonymized sets of data from users who have turned on the location history setting, which is off by default. People who have location history turned on can choose to turn it off at any time from their Google account and can always delete location history data directly from their timeline. Google also added artificial noise into the dataset to further anonymize the data. These [privacy-preserving protections](https://policies.google.com/privacy?hl=en) also ensure that the absolute number of visits isn’t shared.

This data will help me answer where people have been going during different phases of the pandemic and how have the different phases influenced the types of locations people are frequenting more often than before. This expands upon the common analysis by providing context of where the virus may have been more transmissible during peaks and also providing insight about people’s activity while navigating an ongoing pandemic.

1. **Unknowns and dependencies**:

As time passes and we move further away from the baseline period, populations might vary due to relocation or new regional and remote working options. Google’s understanding of categorized places might also change. For example, the same value today and in April 2020 might not indicate the same behavior or adherence—it might be that Google has updated information about shops and restaurants in the region or that fewer people live there now. These differences could shift the values up or down over long time periods, so we recommend using some caution when analyzing data from longer time intervals (6+ months).

As mentioned in the data section, the data collected is only from users who have opted-in to sharing their location history through their Google accounts which means that the data represents a small fraction of the region population. We will still maintain the assumption that the users who have opted in are still representative of the population for the purposes of this project. Seasonality is also another unknown since we are dealing with time series data. It may be hard to extrapolate insights for parks for example since parks may have a dramatically lower percentage change during colder months. It may be hard to differentiate seasonality changes from COVID-19 phase changes.

1. **Methodology**:

**Gathering Data and Preprocessing:** The additional data is a zip file containing CSVs per country by date i.e. a separate csv for US 2020, 2021, and 2022. I plan to extract just the data pertaining to the US and joining each year’s data to form one CSV and then joining that CSV with the mask mandate and daily cases CSVs. This would result in one final CSV containing confirmed cases, location percentage changes, and the respective dates.

**Visualizations:**  I plan to create several visualizations to depict the year over year change of COVID-19 confirmed cases and linking the mobility changes to within the same time period. Specifically I want to look into the impact of the mask mandate on mobility destination changes, the case counts vs. mobility destination changes, and mobility destination changes in general.

**Analysis:**  I plan to conduct t-tests for multiple locations across different time periods to see if there is a significant change in mobility before, during, and after the mask mandate. A t-test would be appropriate since we have a large sample and the standard deviations are unknown. Alongside t-tests, I’m interested in creating a regression model to see if we can predict COVID-19 confirmed cases by location mobility or predict location mobility of a specific location based on the COVID-19 phase time frame.

1. **Timeline to completion**:

November 10 - Submit Extension Plan

November 17 - Complete exploratory data analysis of the community mobility dataset and create initial visualizations

November 24 - Build a model and conduct statistical tests to confirm hypotheses and answer research questions

December 1- create clean documentation and push all necessary files to the project repository

December 5 - Project Presentation

December 12 - Submit Final Report and Project Repository

**Sources:**

T-test: <https://www.scribbr.com/statistics/t-test/>

Regression: <https://towardsdatascience.com/understanding-regression-using-covid-19-dataset-detailed-analysis-be7e319e3a50>

Time series analysis: <https://www.tableau.com/learn/articles/time-series-analysis>