Extension Plan

Aaron Jacobson

Motivation and Problem Statement

The prevalence of Covid19 has reshaped the way we live. On first glance, the data seems to suggest that non-pharmaceutical interventions like mask mandates have no effect on the spread of the disease. Some have used this to argue that masks aren't needed. I want to investigate whether there is a relationship between changes in human activity and the spread of the virus rather than changes in government policy. This research will address the issue of noncompliance with government mandates obscuring the apparent efficacy of the mandated actions. Additionally, I hope to analyze whether there is a measured change in people's behavior in response to government mandates. This will be a human-centered analysis because people's response to changing situations will never fully reflect government policies.

Research Question and Hypothesis

To address the motivations described above, the primary research question I hope to answer is:

> "Is there a relationship between changes in measurements of the level of human activity and changes in the speed at which Covid 19 spreads in New York?"

My hypothesis is that there will be a positive correlation between the level of human activity and the increase in Covid 19 cases in New York.

Data Required

To fulfill this research I will need data from a few sources. First, I will need the number of daily covid cases in New York. Additionally, I will use Apple's Mobility Trends data which can be accessed from https://covid19.apple.com/mobility. I plan to use Apple's data as a proxy for the level of human activity in New York as it measures the percent of activity relative to January 13, 2020. Apple's data represents the number of routing requests per day made through Apple's services. Apple's terms specify "You may use Mobility Trends Reports provided on the Site, including any updates thereto (collectively, the "Apple Data"), only for so long as reasonably necessary to coordinate a response to COVID-19 public health concerns (including the creation of public policy) while COVID-19 is defined as a pandemic by the World Health Organization." The terms also require that I not use the data to try to identify any specific users of Apple's services. Given that the data is aggregated to a very high level, cities are the smallest unit of analysis provided, there is not much of a concern of violating an individual person's privacy. There is, however, a concern that I may over-rely on Apple's data and assume it's more representative of the general population than it really is. This data will help me analyze whether

changes in the level of human activity in New York over the past two years are associated with changes in the spread of Covid 19.

I also want to use OpenTable's Seated Dinner dataset, which can be found here: https://www.opentable.com/state-of-industry. This dataset represents the % change in the daily number of seated dinners at participating restaurants in 2020 and 2021 compared to 2019 data for the same day. I cannot find any terms of use for OpenTable's data. The main ethical consideration with using this data set is that there might be a bias in terms of what kinds of restaurants participate in OpenTable's services. There is a potential that those restaurants had a significantly different experience during the pandemic compared to restaurants that don't participate in OpenTable. This data will help me analyze whether changes in the level of human activity, as measured by seated dinners at restaurants, are associated with changes in the spread of Covid 19.

Unknowns and Dependencies

The main concern I have is that I have no ability to verify how representative Apple's and OpenTable's data is for the US population as a whole. To properly answer my question I would use a better source of data that encompasses all people but that data source doesn't exist.

Methodology

I plan to primarily use OLS autoregression to statistically determine whether there is a relationship between changes in the spread of Covid 19 and changes in the level of human activity. I am choosing to use OLS with a lagged variable because I don't want to overfit my model and because OLS is capable of handling time series and polynomial data.

Project Timeline

Task	Planned Completion Date
Collect and Clean Data	Thursday Nov 18th
Analyze and Validate Analysis	Thursday Nov 25th
Produce Visualizations	Thursday Dec 2nd
Create Presentation	Tuesday Dec 7th
Create Final Report	Tuesday Dec 14th