Part 2 - Extension Plan

Motivation/problem statement:

The COVID-19 pandemic disrupted human life in many ways. Starting from the sudden rise in use of masks, to only be able to meet people virtually, to not being able to attend long planned festivities with families and friends. Out of these, the one aspect that stood out to me was how COVID-19 disrupted people's travel plans. Being from Bangalore, a city infamous for its prolonged traffic jams throughout the year, the shift to empty roads was upending and even depressing to me personally, to say the least. I was missing the dreaded. However, I should be the last person to complain, since the first time I contracted COVID was because I went travelling during the uprise of the Delta variant in India in April 2021. All said and done though, stringent travel bans were possibly the most impactful policies to make a difference in decreasing the spread of COVID-19, from a scientific or practical perspective. With this motivation, I intend to explore the different ways travel bans had an impact on the spread of COVID-19, focusing specifically on the Cook County in Illinois. This problem statement is especially human centered as it reflects how human travel behavior can influence rise in COVID-19 cases, rise in hospitalized patients, deaths and more importantly, inform public policy during future pandemics or epidemics.

Research questions and/or hypotheses:

The specific research questions and their hypothesis, I'd like to explore are:

- **Question:** How did inter-regional travel habits between communities impact the growth of COVID cases?
    - **Hypothesis:** Every 10% increase in the number of people travelling between communities increased COVID cases by 6% increase in overall cases.

Data to be used:

As of now, I am planning to combine three datasets. I will require:

1. The RAW_us_confirmed_cases.csv file from the Kaggle repository of John Hopkins University COVID-19 data. This data is updated daily. There are no ethical considerations I feel we have to consider with this dataset, as these are the actual cases that were reported publicly. Licensed under Attribution 4.0 International (CC BY 4.0). Overall, this dataset will contain the daily COVID case count, which will act as our dependent variable in the analysis.

2. The Trips by Distance dataset to understand daily travel behaviors of people in Cook County, Illinois. The dataset summarizes how many people are staying at home during the COVID-19 pandemic and how far people are traveling when they don't stay home. It consists of 22 columns out of which I felt 16 relevant for our analysis. This dataset is licensed under U.S. Government Works. I couldn't find any information on Terms of Use, however.
   One ethical consideration I feel we should account for is that these data are experimental, which means they are created using new data sources or methodologies that benefit data

users in the absence of other relevant products. Since we can't be sure about the accuracy of the numbers in the data, we should be careful about the fidelity of the interpretations from any analytical studies on this data, in the pursuit to inform travel policy considerations during future.

Overall, this dataset will aid in answering the first two hypothesis questions – general influence of travel habits of people and effect of inter-regional travel on the pandemic.

Unknowns and dependencies:

There's just one factor I think can potentially impact our ability to answer our hypothesis:

- Vaccination might be having an interaction effect with travel habits. We might have to regress this data with vaccination data source for Cook County to understand how both these features interacted with each other over time.

Methodology:

Data Gathering and Processing

Two datasets will be combined for this analysis. Daily case data will be gathered from the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University, representing daily cumulative case counts by county. This data will be joined together at the county level, I am anticipating performing additional calculations to create the final dataset. New Cases will be calculated as the difference in total case count form day to day

Analysis

1. **Correlation analysis**: I would like to start with comparing how does each of my covariates vary with the total no. of cases over the period from April up to the period of time I have covariate information for.

   Some of the steps I plan to take are:
   - Plot a multi-line time series chart to visually understand maybe how a change in human travel behavior caused a change in number of COVID cases.
   - Utilize automated checkpoint detection methods to identify major inflection points during the period of our analysis.
   - Check if there are any trends over the entire period or during a specific period of time
   - Check for 3-day, weekly or monthly seasonality patterns using ACF and PACF plots for inference

2. **Regression Analysis**: For all the hypothesis, I am planning to regress the independent variables describing local travel & region-to-region travel habits with the dependent

variable, new number of COVID cases each day. This step is key, because the magnitude of the coefficients of our model will help us determine how the increase/decrease in the value of the independent variable results in an increase/decrease in the target variable.

3. **Counterfactual Analysis**: For all the time series regression models I will create, I intend to also see what would happen if perturbing the value of features would change our model's predictions. This can help us understand other hypothetical scenarios than the "reality" and how the number of COVID cases would covary at that particular point of time.

Presentation
Based on the **PechaKucha** presentation format, I intend to follow the "talk less, show more". I will try to do that by trying to present my results using as many visualizations as possible.

Timeline to completion

Data Gathering and Processing - November 15th 2022
Correlation Analysis - November 16th 2022
Regression Analysis- November 18th 2022
Counterfactual analysis – November 20th 2022
Visualization of results - November 24th 2022
Documentation - December 2nd 2022