Part 2 - An Extension Plan

Socioeconomic Impact of COVID-19 (Maricopa County, AZ)
Rohit Lokwani
November 9th, 2022

1. Motivation/Problem statement

COVID-19 has had a massive impact on the world in the last couple of years. The rapid spread of COVID-19 had endangered lives, disrupted livelihoods, and had an impact on global trade, the economy, and businesses. The world economy had started to experience significant disruptions and was and is still moving toward a severe recession and an unprecedented economic crisis. All nations have experienced challenges as a result of COVID-19, but many nations have had to deal with a more difficult situation because of their large populations, subpar health services, high rates of poverty, poor socioeconomic conditions, and inadequate social protection systems.

The UN's Framework for the Immediate Socio-Economic Response to the COVID-19 Crisis warned that "The COVID-19 pandemic is far more than a health crisis: it is affecting societies and economies at their core. While the impact of the pandemic will vary from country to country, it will most likely increase poverty and inequalities on a global scale." (COVID-19 Socioeconomic Impact). To educate and customize government and partner responses to the COVID-19 issue and ensure that no one is left behind in this effort, it is essential to assess the implications of the crisis on communities, economies, and vulnerable people. These factors can directly affect the ability to earn, the representation of race or caste in society, education, etc. Hence, how did the pandemic impact such factors interest us the most? For this analysis, we target the socioeconomic factors primarily but focus mainly on economic ones. We try to do a fine-grained analysis of the socioeconomic implications of COVID-19 in Maricopa County, Arizona in the United States. We understand the implications at the individual level and industry level.

2. Research questions and/or hypotheses

Building on the analysis we conducted in part 1 of the project, where we tried to answer our first question, how did the (national-level) masking policies affect the COVID-19 infection rate? We plan to answer one main research question

"What was the impact of COVID-19 on socioeconomic factors at the individual level and industrial level in Maricopa County in Arizona?"

We hypothesize that the answers to this question would entail better measures for epidemics or such scenarios in the future. To deepen our analysis on this, we divide it into multiple subquestions:

1. What was the influence of the pandemic on the unemployment rate and Civil Labor Force in the county?
2. How did the socioeconomic factors (Education, Median Household Income, and Gross Domestic Product) change during the pandemic?
3. How did the COVID-19 cases, masking, and vaccination policies affect the economic indexes of the county in different industries?

4. As per research (Stanford), the real estate industry was one of the most impacted ones, we validate that and try to study trends in various factors (Active Listing, Total Listing, Pricing changes) over the course of the pandemic.

An additional/miscellaneous question, we would like to answer is that at any point where the COVID-19 cases could be accurately forecasted and in turn their impact on economic indexes.

3. Additional Datasets to be used

Following are the datasets that we plan to use for all of the questions mentioned above.

1. **Covid-19 data:**

   We plan to use the COVID-19 dataset as provided in the Common Analysis part.

   a. John Hopkins University COVID-19 data (License: Attribution 4.0 International (CC BY 4.0))
   b. Masking mandates by county (NCHS: Can be used for Statistical reporting and analyses)
   c. The New York Times mask compliance survey data (Copyright 2021 by The New York Times Company, used for non-commercial purposes)

2. **Federal Reserve Economic Data, FRED Monthly Data:**

   The dataset is licensed under FRED® Services General License and is allowed to be used none other than for statistical analysis purposes. This dataset has data points starting from 1990 to 2022 for unemployment rates, Civilian Labor Force participation, and housing parameters(Active Listing, Pricing). It is a two-dimensional dataset with timestamps and the respective measures in either case. We plan to understand the effect of the pandemic (COVID-19 cases) on these three measures. It helps in answering questions 1 and 4.

   The links to the dataset are as follows:

   a. Unemployment Rates in Maricopa County
   b. Civilian Labor Force (Total employed)
   c. Housing impact: We study the different aspects of the housing sector (Active listing, total listing, and pricing)

3. **CDC's Agency for Toxic Substances and Disease Registry Data:**

   This dataset keeps a track of the social vulnerability of counties given the diseases or abuse of toxic substances. Social Vulnerability Index (SVI) indicates the relative vulnerability of every U.S. Census tract. Census tracts are subdivisions of counties for which the Census collects statistical data. SVI ranks the tracts on 16 social factors, including unemployment, racial and ethnic minority status, and disability. We have the index values for all of these themes.

   The National Center for Health Statistics (NCHS), and Centers for Disease Control and Prevention (CDC), conduct statistical and epidemiological activities under the authority granted by the Public Health Service Act. NCHS survey data are protected by Federal confidentiality laws including Section 308(d) Public Health Service Act and the Confidential Information Protection

and Statistical Efficiency Act or CIPSEA. These confidentiality laws state the data collected by NCHS may be used only for statistical reporting and analysis.

Since it is the annual patterned data we use it to answer our 2nd question about the state of the socioeconomic variables before the pandemic and after the pandemic. Median household income and Gross Domestic Product, Education, and Poverty Estimates in different groups. The dataset can be found in United States Counties.

4. **Argonne National Laboratory Data for Different Sector Information**

This dataset contains the indexes for different industrial sectors across the country. We use it to research question 3. Since this dataset is available from January 2020-April 2022. We will be able to study the impact from almost the start of the pandemic. The data is spread out monthly for each of the counties as the index column and different sectors as the subindex. This dataset is again allowed to be used only for Statistical Analysis purposes and is licensed under DEAR 970.5204. The link for the dataset is US Counties Economic Information.

4. Unknowns and dependencies

In all the datasets mentioned above, the datasets are quite clean in general. The only concern is understanding the features accurately, which might require researching socioeconomic terminologies in order to avoid drawing incorrect conclusions. Another concern is here, we are directly correlating the variables like economic impact in industrial sectors and the effect of COVID-19, but it's important to note that these variables might be impacted by various other factors like International relations amongst others. For example, the CDC's Agency for Toxic Substances and Disease Registry Data measure the impacts of drug consumption and diseases on the economic indexes, but we are just correlating the impact with the disease part. Also, since the data is procured and licensed by Government affiliated organizations and the Personal Identifiable Information (PII) is anonymized, we do not see any ethical challenges with it.

5. Methodology

This section describes the end-to-end methods involved right from gathering, the tools, the analytical or statistical methods we foresee, and the presentation of findings.

5.1 Analysis Tools

Jupyter Notebook will be used to script the analysis steps using Python 3. While Python data analysis libraries such as NumPy, Pandas, and Sklearn will be used for data preprocessing, integration, wrangling, and executing correlations, Python's plotting libraries such as Matplotlib and Seaborn will be used to plot the results.

5.2 Data Gathering and Preprocessing

In all the datasets mentioned above, the datasets are quite clean in general. All of them are structured files stored as CSVs or Excel spreadsheets. Since, for answering all the questions, our primary data would

be the timestamp and the related features from our sparse dataset. By the looks of it, the data looks clean and quite well formatted, hence no preprocessing is required at this point. In order to create a consolidated dataset that is ready for analysis, the integration stage entails matching entries from all the processed datasets using state and county variables as unique keys. State and county names are included in each dataset. In order to map records from other datasets, these two fields serve as the primary keys. Some of the data is structured as monthly or in other cases as annual intervals. We will be adjusting our COVID-19 case data to these intervals in order to answer our questions. To be specific, the data will be adjusted(rolling average) to monthly levels for answering questions 1,3,4 and annually for question 2. For the miscellaneous question, we plan on using the daily cases for forecasting them.

5.3 Analysis Methods

To answer our research questions, we will mainly be using T-tests, correlation, and time-lagged cross-correlations to compare two sets of interesting features. We also plan on using heatmaps for visualizing multivariate data and plot bar charts and other basic plots for Exploratory Data Analysis. We plan on using Linear Regression for answering our miscellaneous questions.

5.3.1 Correlation Tests

Correlation analysis is a statistical method that gives insights into the existence of connections between quantitative variables and provides a metric to infer the strength of such relationships (Link). We hypothesize housing stability index correlates with the socio-economic country. We try to prove this hypothesis using Pearson's correlation coefficient (Linear Relationship) or Spearman's Correlation Coefficient (Less stringent, looks for monotonicity) to determine that. We also plan to determine the relationship between different industrial sectors and present these as a heatmap.

5.3.2 Time-Lagged Cross-correlation

This is a method to determine if there's a correlation between two-time series with a lag of time interval. This usually is helpful when you hypothesize that the patterns in one time series occur after another instead of at the same time interval. We will use this technique to determine the impact of COVID-19 on unemployment and Civil Labor Force participation as it is expected to occur with a time difference.

5.3.3 Linear Regression

In statistics, linear regression is a linear approach for modeling the relationship between a scalar response and one or more explanatory variables (also known as dependent and independent variables). The case of one explanatory variable is called simple linear regression; for more than one, the process is called multiple linear regression. We plan to use simple linear regression primarily as a method of forecasting and answering our miscellaneous question of, if at any point would we have been able to predict the COVID-19 cases after a day or a month. This would essentially improve the predictive power of our analysis and help in measures curtailing the spread of future epidemics.

5.3.5 Student's T-test

In statistics, a student's t-test is a method of testing hypotheses about the mean of a small sample drawn from a normally distributed population when the population standard deviation is unknown. We can compare two sample group means using two-sample T-tests. We plan to take the annual Median household income and Gross Domestic Product and use a t-test to figure out if there exists a statistically

significant difference between the period prior to the pandemic and one after that, supporting question 2 research.

5.3.6 Change Point Detection

In statistical analysis, change detection or change point detection tries to identify times when the probability distribution of a stochastic process or time series changes. In general, the problem concerns both detecting whether or not a change has occurred, or whether several changes might have occurred, and identifying the times of any such changes. We used the ruptures library in Python for doing this and then verified against the masking policy dates. We plan to use this technique again to answer our research question 3 for determining if the change points due to masking or vaccination policies had an impact on the economic indexes of the various industrial sectors.

5.3.6 Data Visualizations

We plan to do an exploratory data analysis of entire data sets to understand the data better. We also plan on using this analysis without any specific statistical methods and drawing conclusions for research question 2 from visualizing graphs like line graphs and bar plots.

5.4 Presenting findings

We intend to use heatmaps to communicate the correlation between different industrial sectors during the pandemic with Covid-19 cases being one of the dimensions. The change points for question 3 will be presented using highlighted vertical lines on time-series graphs as done in part 1 of the common analysis. The metrics for students' T-test, linear regression, and correlation analysis will be statistical values or coefficients with time-lagged cross-correlation shown as time series (questions 1,2,4).

6. Timeline to Completion

Following is the timeline of milestones or significant tasks for this project. Note that the following tasks will be done sequentially, hence the start date of the task following the prior one would be right after the latter's completion.

Task	Planned Completion Date	Progress
Part 1: Common Analysis	11/3/2022	Completed
Research, brainstorming and Data Collection	11/7/2022	Completed
Part 2: Extended Plan	11/10/2022	Completed
Exploratory Data Analysis	11/15/2022	In progress
Analysis for answering research questions	11/24/2022	-
Validating/Testing the Analysis	11/27/2022	-
Visualization and Documentation	11/30/2022	-
Part 3: Presentation	12/04/2022	-
Part 4: Submitting the report	12/09/2022	-

7. References

1. COVID-19 Socioeconomic Impact
2. John Hopkins University COVID-19 data
3. Masking mandates by county
4. Mask compliance survey
5. Unemployment Rates in Maricopa County
6. Civilian Labor Force (Total employed)
7. Housing impact
8. CDC's Agency for Toxic Substances and Disease Registry Data
9. US Counties Economic Information
10. Wikipedia.org
11. Britannica