Course: DATA 512 A – Human-Centered Data Science **Name:** Sravan Hande
Assignment: Individual Analysis Project Report **Student Id:** 2129885

Introduction

In this analysis, the population of New Haven County is examined in relation to changes in job conditions as the disease develops through various stages of infection, fatalities and masked policy changes. Although it initially seems likely that unemployment would be at its highest during this epidemic, if we examine it closely, we can see that several industries had boomed in employment because of societal demands. A significant portion of the workforce was relocated to remote work in the interest of employee safety, however, as a result of the spread of the virus and rise in mortality. Given the numerous changes to immunization and masking laws, it makes sense that the employment trend has multiple peaks and dips. With the use of this analysis, we may approach these job shifts from a human-centric perspective and better comprehend how they affect the population's changing economic circumstances. In this analysis we primarily focus on the different aspects of employment – labor force, jobs availability, employed population, unemployed population as a whole and for individual employment sectors. With the help of statistical testing techniques, this analysis aims to validate the significance of COVIDs impact on different employment sectors.

Background/Related Work

An Exploration of Machine Learning Models to Forecast the Unemployment Rate
https://pesquisa.bvsalud.org/global-literature-on-novel-coronavirus-2019-ncov/resource/en/covidwho-1109412

This paper explores univariate machine learning techniques to forecast the South African unemployment rate. Six traditional statistical models are compared with seven machine learning models. The multi-layer perceptron achieves the lowest error rate, whilst the ridge regression model achieved the highest R − squared. These are closely followed by ARIMA, LASSO, and the elastic net, showing that machine learning models can forecast the South African unemployment rate with higher accuracy than traditional statistical methods. A univariate time-series analysis was undertaken to forecast the South African unemployment rate, this is similar to what was used by several researchers. The model's performance was measured based on their mean absolute percentage error (MAPE): a common performance measure in time-series analysis. The following subsection will provide the descriptive statistics of unemployment as well as the results of the analysis

COVID-19 pandemic and unemployment rate prediction for developing countries
https://journals.plos.org/plosone/article/file?id=10.1371/journal.pone.0275422&type=printable

This research paper's primary goal is to examine the effect of the Covid-19 pandemic on the unemployment rate in selected countries of Asia through an advanced hybrid modeling approach, using unemployment data of seven developing countries of Asia: Iran, Sri Lanka; Bangladesh; Pakistan; Indonesia; China; and India, and compare the results with conventional modeling approaches. Finding shows that the hybrid ARIMA-ARNN model outperformed its competitors for Asia developing economies. In addition, the best-fitted model was utilized to predict the five years ahead unemployment rate. According to the findings, unemployment will rise significantly in developing economies in the next years, and this will have a particularly severe impact on the region's economies that aren't yet developed.

Predicting Impact of COVID-19 on the Global Economy Based on Hybrid Model
https://www.atlantis-press.com/article/125971867.pdf

In this research a mixed prediction model (AdaBoost, Linear Regression, Decision Tree) is constructed to predict the GDP per capita of different countries. Many algorithms have been developed or applied to economic development prediction, such as K-Nearest Neighbor (KNN), random forest, AdaBoost and so on. In order to obtain more accurate prediction results, this paper uses a hybrid algorithm (a new algorithm obtained by integrating linear regression, random forest and AdaBoost algorithms) to predict GDP per Capita, which is a good indicator to measure a country's economy and living standards. The GDP prediction under the epidemic situation is a highly nonlinear problem, which will be affected by many different dimensions, such as the country, region, population density, government or local policies and so on. In this article, in order to improve the prediction accuracy, three algorithms (linear regression, random forest and AdaBoost algorithms) are combined together to get a new model and the final experimental results also show that the prediction performance of the hybrid model has been greatly improved. In this paper, the proposed model is simpler and more convenient to analyze the multi factor model, but because it is a linear model, it is difficult to accurately analyze and predict the nonlinear patterns.

In all these articles and research, the authors focused on predicting the unemployment rate or predicting the economy on a larger scale, however, there is not enough research that focuses on the impact of COVID on the employment as a whole and the different sectors within. In the current analysis, the primary focus is to understand how employment, labor force, job availability has changed over time across different sectors due to the pandemic. More specifically, this article focusses on addressing below research questions.

As a part of this analysis, a few key questions/hypotheses I would like to address are,

1. How is the overall employment impacted during and after the pandemic, did the employment status gain its previous state as before covid?
2. Do certain industry sectors have higher impact with the confirmed cases than others?
3. Are there any specific employment sectors that are impacted the more and are there any sectors which does not see a major change during this period
4. Do we see a significant change in employment with the masking policy changes?
5. How did the policies like masking mandate effect the overall employment (labor force, jobs availability, employed population) and different sectors?

Though there are several other factors that would have affected the employment conditions within the county, the idea here is to address the above questions with the help of data limited to confirmed cases, masking policy changes, employment. Along with visual analysis of these explorations, the idea is to use the appropriate statistical testing techniques to see if there is a significant impact of the pandemic on employment domain.

Methodology

The methodology is primarily designed with People centered considerations. Though multiple analysis techniques are performed as part of the project, only interpretable results are selected for the final analysis and discussion. This will enable audience from wide domains to easily perceive the analysis and results which will help for a broader reach of the work. The implementation of this work is done with reproducibility embedded using interactive jupyter notebooks and inline documentations that describes each step performed in the analysis.

We will be using the standard data science methodology to address the above hypothesis and questions. The initial step in the process is to perform exploratory analysis to better understand our data. This will enable us to access the data quality and perform data cleaning operations to eliminate any inconsistencies. Each of the data sources have the information available in different metrics, for example the employment data has attributes to compare at aggregated level and for individual sectors, we have employment, unemployment and the workforce available at each data point. Similarly, the vaccination data has 72 columns each indicating a different aspect or a comparative metric at each data point. We will access each of these available features and pick only those that will help us accurately address our questions. Once we arrive at the final set of features for each of these data sources, we will combine these sources based on the timestamp. This will enable us to obtain a single view for employment statistics, fatalities, confirmed cases, vaccinations and masking policy on a given date. We will use this merged data as our master data in our further analysis to address our hypothesis and questions.

To address our intuitions in this analysis, we will be utilizing statistical testing methods Welch's T-test and Causal Impact Analysis. Each of these methods have different aspect of explainability and we will be choosing the right method based on the data and question. We further break the data into different timeframes to better fit our models accounting for positive and negative correlations. As most of our analysis is based on temporal data, we will be using several time series plots to better judge the model and infer the finding from our modeling. Though the tests can be performed to analyze the statistical significance of these hypothesis, for questions regarding employment changes across different sectors visualizations would be of a great use to represent the findings.

In the stages of data collection and prior to performing the data preprocessing the data is analyzed for ethical considerations and violations. The three data sets being used in this analysis are aggregated at a county and department level, which will mask all the possible personally identifiable information of the residents of that county. Though the data describes the county as a whole, we will not be operating with any specific resident's data of this county. This will ensure the privacy of these residents and keep this work more distributable.

To proceed with the data cleaning operation, firstly the 3 data sets were ingested - COVID-19 fatality data from John Hopkins University, New Haven Employment data and COVID-19 Vaccinations in the New Haven County. The analysis involves performing several transformation and translation, few of which are – Converting the columnar data to row and merging year, month columns to obtain a time series data, converting features to appropriate data types for better filtering and visual representation, and dropping nan values. Ingesting sector wise employment data into a single dataset and cleaning the final dataset by dropping invalid values. From the new havens mask mandate data, we assume the missing values as No, meaning people were not mandated to wear masks. Once the data cleaning is done, we proceed to visualize the data.

On a high level we will be using these below data plots for visual analysis,

- We want to examine the trends in COVID instances between 2020 and now.
- We would like to know the trends of employment, number of jobs, unemployment, labor force over the span of 5 years 2018 to 2022 and see if there's a noticeable change in the trends during the pandemic time and other significant dates like masking policy changes
- In order to determine whether or whether there is an impact of covid on the job market, we also want to evaluate trends on jobs in the job market over the course of these five years, from 2018 to 2022.
- We wish to examine whether employment variation in new havens with masking mandatory rules has changed.
- If there is a variation, we also wish to examine the employment situation by sector in New Haven under the mask mandate policy.

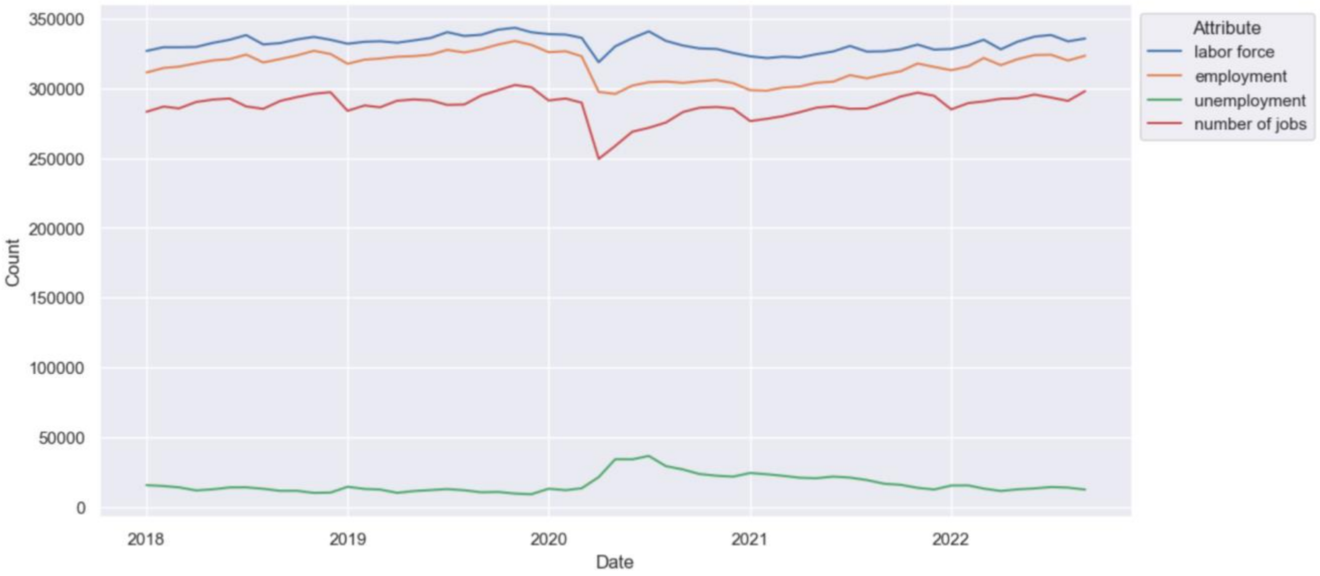
While choosing the statistical modelling techniques, our primary goal is to use a model or technique that would guarantee fairness and transparency for our analysis. This will enforce us to perform analysis using explainable statistical algorithms instead of opaque algorithms. Using such explainable models will be easier to demonstrate the process with clarity and infer the request in a more meaningful manner. The methods that would match such criteria and also work well with our data are statistical testing models, which proven to be success over validating significance of a change.

We then proceed with testing the significance of these findings. To validate if there is significant effect of pandemic on the employment, we wanted to perform hypothesis testing using Welch's t test. Hypothesis testing is a method for determining how well one can extrapolate observed results in a study sample to the larger population from which the sample was drawn, hypothesis testing is a process used to assess the strength of the evidence from the sample and provides a framework for making decisions related to the population. Welch’s T Test is a nonparametric univariate test called **t**hat checks whether there is a statistically significant difference between the means of two unrelated groups. When the equality of variances is violated, it serves as a substitute for the independent t-test. The proposition being examined here is: The null hypothesis (H0) states that u1(employment mean before COVID) = u2(employment mean from start of COVID), which means that the means of samples 1 and 2 are equal. Alternative hypothesis (HA): The mean of sample 1 is not equal to the mean of sample 2 is expressed as u1 u2. We can reject the null hypothesis if the p-value is less than the threshold being tested, which is often 0.05, however, as we are performing a social experiment, we will be using the standard 0.1 as our significance threshold.

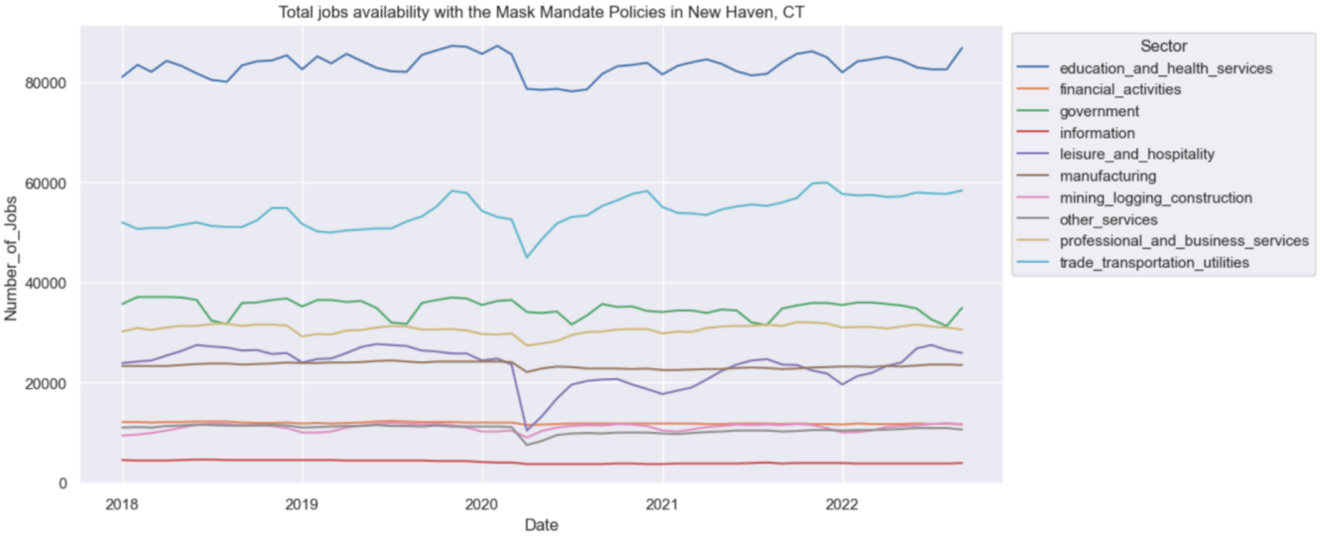
As we are using time series data, the independence assumption of the Welch’s T test cannot be fully met, as the values will be autocorrelated in the time series data, due to which this test often provides overoptimistic results. Hence, we are using another testing method, Causal Impact Analysis (CIA) to validate the significance of interruption by COVID. In CIA, we need two time series that have similar trend over time prior to the intervention. In our case the intervention we are trying to validate is that cause by COVID. For this purpose, we split the employment data by sectors, and we notice that the government sector does not have a significant change due to COVID and has a similar trend to that of total employment prior to the pandemic. This will serve as a control group time series. Now, we can use the control group time series along with our original (to be validated) time series data sets and perform CIA.

Findings

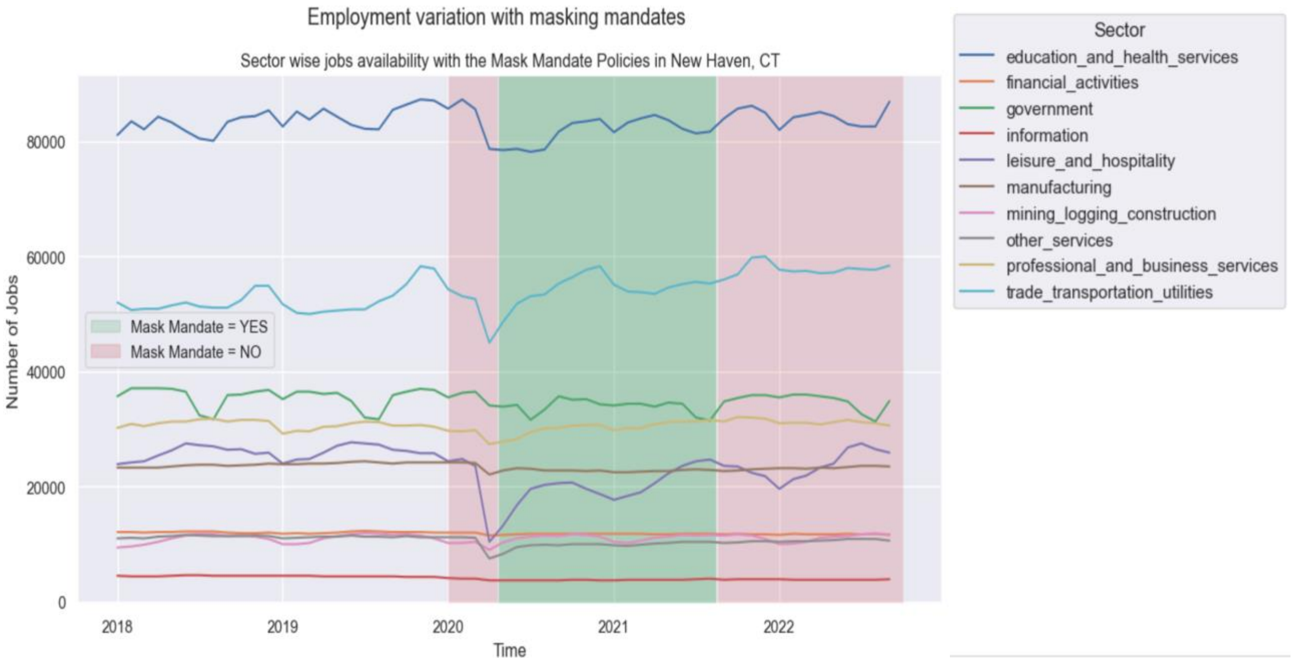
From the initial visualizations we can see that there has been a noticeable fluctuation during the pandemic period. However, these are different for each of the employment factors we are analyzing. From the below visual, we see that both the labor force and employment have fallen during early COVID period. Though the fall in labor force is lesser and sharp, we can see that the employed population took a little longer to regain its previous state. This briefly shows us how the labor force and employed population changed over different interventions like different waves of COVID, mask mandates, vaccination etc.



The next analysis is on the different employment sectors jobs availability. While few of the sectors does not show much fluctuation during the pandemic, the sectors like leisure_and_hospitality, trade_transportation_utilities, education_and_health_services face a huge dip. This scenario can be explained by the nature of each of these sectors work environment. It is surprising to see that education and health services has also fallen a big time, which might be due to the risk aspect.



The next analysis is with respect to the mask mandating policies. We see that the polity seem to have a significant positive trending effect on the job availability. We see that the situation has again faced a dip when the mask mandates were removed, which shows us a considerable positive correlation. Though there is a second dip, which might be due to the second wave of the pandemic, the mask mandates seem to have improved the employment condition overall. We can also notice that by the end of the year 2022, the employment situation with respect to jobs availability is almost back to the place where it was prior to pandemic.



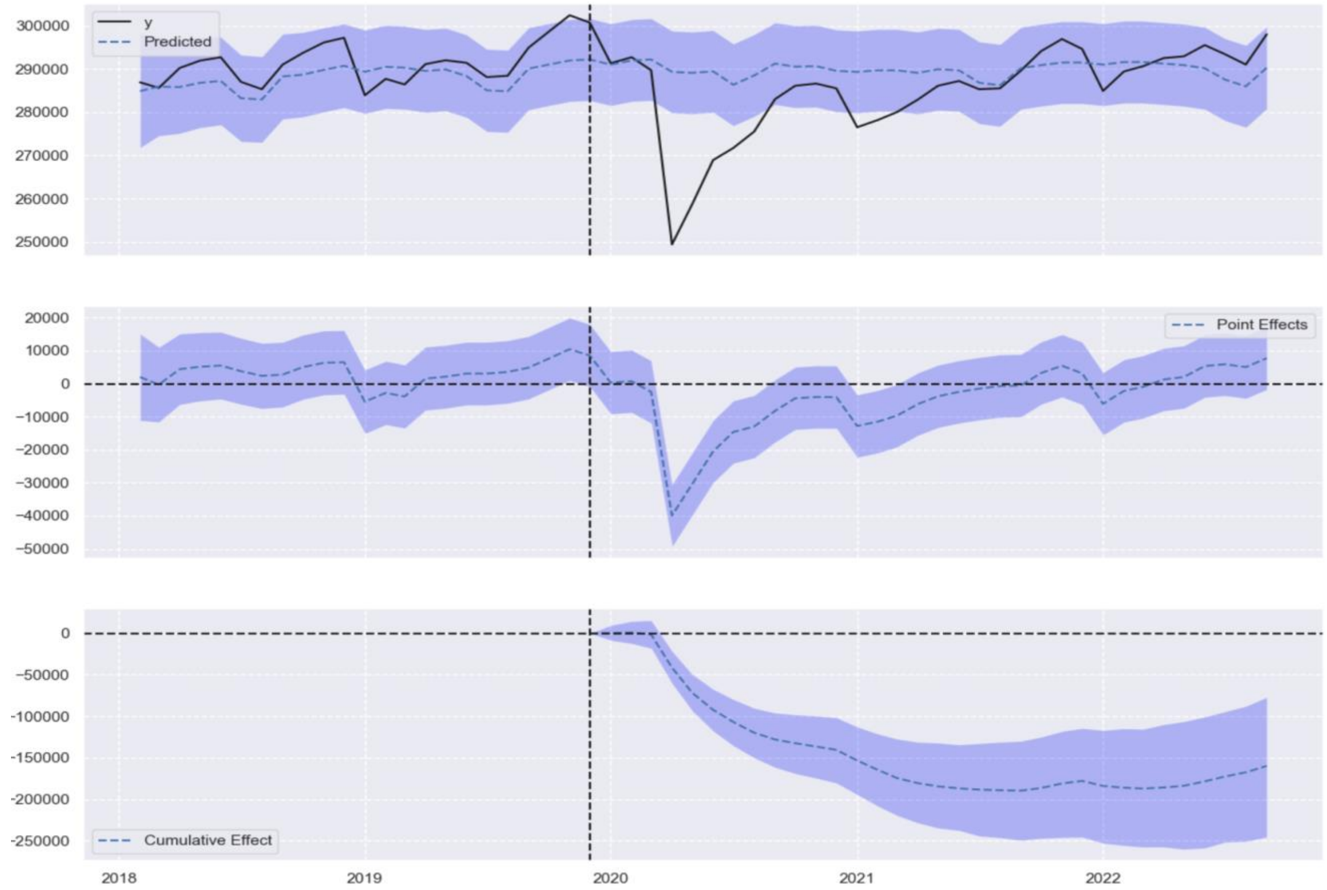
The Welch's T test results are as below,

Welch's t-test= 2.8544

p-value = 0.0063

Welch-Satterthwaite Degrees of Freedom= 49.0716

The p-value is much lower than our significance threshold values of 0.01, which shows that there is a significant impact on the employment. As we discussed in the methodology, these results cannot be reliable due to the autocorrelated nature of the time series data. We will be proceeding with the Causal Impact Analysis to further validate our hypothesis.



From the above plot we can see that, during the post-intervention period, the response variable had an average value of approx. 284975.76. By contrast, in the absence of an intervention, we would have expected an average response of 289820.3. The 95% interval of this counterfactual prediction is [287301.36, 292429.06]. Subtracting this prediction from the observed response yields an estimate of the causal effect the intervention had on the response variable. This effect is -4844.54 with a 95% interval of [-7453.31, -2325.6]. Summing up the individual data points during the post-intervention period (which can only sometimes be meaningfully interpreted), the response variable had an overall value of 9404200.0. By contrast, had the intervention not taken place, we would have expected a sum of 9564069.79. The 95% interval of this prediction is [9480944.94, 9650159.07].

The above results are given in terms of absolute numbers. In relative terms, the response variable showed a decrease of -1.67%. The 95% interval of this percentage is [-2.57%, -0.8%]. This means that the negative effect observed during the intervention period is statistically significant. The probability of obtaining this effect by chance is very small (Bayesian one-sided tail-area probability p = 0.0). This means the causal effect can be considered statistically significant.

Discussion/Implications

In this analysis we clearly see that COVID has disrupted the employment situation in the New Haven County. All the four views of employment, labor force, employed population, unemployed population, jobs availability show clear fall during the initial stage of COVID. However, we see that each of these have a unique trend during the past two years of pandemic. It is interesting to see how the number of jobs has fallen much more than the labor force and employed population. From a human centered view, this explains how the employers reacted to the COVID situation by freezing all their hiring activities to achieve stability during the tough time. The labor force is another interesting trend that explains people's reaction towards the pandemic, in the initial stage, we can see a sharp dip, which later pick up in nor more than a quarter. Unlike labor force, the employed population and number of jobs took a very long time to get back to their pre pandemic state. These findings will help us estimate and expect the changes in employment in case of any future outbreaks.

From the visual analysis, we can see that education and health services are the highest employed sectors in the county, both of which were severely impacted during the pandemic. This seems deviated from general expectation that health services would have seen a huge rise in employment during the pandemic period. Despite the medical emergency situation, the sector seen a drop in number of jobs available through the pandemic. The fact that both education and health services both combined to a single sector should be highlighted in this discussion, as it is possible that the decline is solely due to the fall in number of jobs in education alone. The next highest impacted is the transportation sector. Considering that there was a state of emergency and lock down in place, this finding is totally in line with our intuition about the impacted sectors. This analysis helps us be prepared to fortify the weaker sectors for any future pandemics.

The statistical significance analysis performed using the Welch's t test and the Causal Impact Analysis clearly validate our hypothesis about the impact on employment and different employment sectors. In the post intervention graph we can see how the cumulative effect is slowly reducing and getting back to the neutral axis. This indicates that after more than two years, the employment situation in the New Haven County is slowly regaining its position as of pre pandemic time.

Limitations

One of the key considerations in our analysis is that there are a number of factors that would have impacted the employment conditions in the New Haven County during the pandemic, however, we will be focusing only on certain factors related to confirmed cases and masking data. There is possibility of gaps and inconsistencies within the data sources we have considered. The confirmed cases data contains only the reported cases, however, there would be a significant chunk of population who would not have reported their results or would not have tested for the infection at all. This would have an impact on our analysis as during the peak and much later of the infections, the reported confirmed cases might not be accurate due to the wide availability of at home test kits and vaccinations.

The employment information available on the US Bureau of labor statistics is a publicly available dataset that is collected and updated monthly. It is very much possible for the data to be inaccurate during the pandemic time for many reasons, we cannot expect every organization to report their updated employment statistics in the havoc of pandemic and similarly we could expect errors in curating the data in such chaotic situations. For the welch t-test the significance is very high, however we cannot rely on this due to the auto correlation assumption that is not met completely due to the underlying time series data. Further, we have considered the Causal Impact Analysis approach to test the significance of COVID intervention for which we have considered government sector as a control group. Though the assumptions of similar trend prior to intervention met for these time series plots, we cannot validate the actual impact on government sector due to COVID. Our analysis would not be considering these inaccuracies along with other factors that are beyond the scope of this project.

Conclusion

In this analysis I intended to analyze the impact of COVID on the employment situation of New Haven County from multiple views of work force. The analysis is designed with human centric decision principles in consideration and aimed to produce a simple and explainable analysis to better understand the employment state. Different visual analysis techniques are applied to tailor the hypothesis for our analysis, which are then validated for significance using the statistical techniques Welch’s t test and Causal Impact Analysis. Though Welch’s test provided evidence to reject the null hypothesis, we utilized CIA to strengthen our evidence as the assumptions were weak in Welch’s t test.

References

[1] Lai, H., Khan, Y.A., Thaljaoui, A. et al. RETRACTED ARTICLE: COVID-19 pandemic and unemployment rate: A hybrid unemployment rate prediction approach for developed and developing countries. Soft Comput (2021). https://doi.org/10.1007/s00500-021-05871-6

[2] Jincheng Zuo1*, Fenglan Ma1, Shaun Chen2. Predicting Impact of COVID-19 on the Global Economy Based on Hybrid Model. ICFIED 2022. https://www.atlantis-press.com/article/125971867.pdf

[3] Mulaudzi, R.; Ajoodha, R.. An Exploration of Machine Learning Models to Forecast the Unemployment Rate of South Africa: A Univariate Approach, *IMITEC 2020, https://pesquisa.bvsalud.org/global-literature-on-novel-coronavirus-2019-ncov/resource/en/covidwho-1109412*

Data Sources

This dataset taken from the **COVID-19 Vaccinations in the United States** represents county-level data on overall US COVID-19 vaccination administration. The information is representative of all vaccine partners, including federal entity facilities, long-term care institutions, retail pharmacies, jurisdictional partner clinics, and dialysis centers. This dataset has 1.8M rows and 72 columns.

COVID-19 data from John Hopkins University data covers:

- confirmed cases and fatalities by US county.
- confirmed cases and deaths at the international level
- some metadata that is present in the unprocessed JHU data

New Haven, economy at glance basically provides the data about the employment statistics. National, regional, state, and urban area BLS Glance pages. The information on these pages was put together using information from various BLS surveys and initiatives. Every time one of the source programs provides new statistics, the Economy at Glance pages are updated with the most recent information. On average, this happens 7-9 times per month. The seven programs produced at glance These programs are Current Population Survey, Current Employment Statistics, Consumer Price Index, Producer Price Indexes, International Price Indexes, Employment Cost Index, and Major Sector Productivity. Along with the data used in the part 1 of the project – Confirmed cases and the masking policy data, we will be using the below additional sources that are available as public datasets. All the data sets are aggregated at county level and does not expose personal information, hence, do not pose any significant ethical concerns.

Datasets	Links
COVID-19 fatality data from John Hopkins University	https://www.kaggle.com/datasets/antgoldbloom/covid19-data-from-john-hopkins-university?select=RAW_us_deaths.csv
New Haven Employment data	https://www.bls.gov/eag/eag.ct_newhaven_mn.htm
The CDC dataset of masking mandates by county	https://data.cdc.gov/Policy-Surveillance/U-S-State-and-Territorial-Public-Mask-Mandates-Fro/62d6-pm5i