Course: DATA 512 A – Human-Centered Data Science
Assignment: Project part 2 - Extension plan
Name: Sravan Hande
Student Id: 2129885

Motivation/problem statement

As a part of this extended analysis on different aspects of COVID period, I would like to focus on the changes in employment conditions of the New Haven County population as the disease progress through multiple stages of infection, fatalities, masking policy changes, and vaccination. As an initial that it strikes for us that unemployment would be at its peaks during this pandemic, however, if we keenly look at it, certain sectors of employment were at their peak due to the societal requirement. On the other hand, with the spread of virus and increase in fatalities a bulk of changes occurred in the employment, a large chunk of workforce was moved to remote work in the interest of employees' safety. We can intuit the employment trend to have multiple peaks and valleys as there were several changes with respect to masking policies and vaccination policies. This analysis will help us approach these employment change in a human-centric manner to understand the changes in economic conditions of the population using the change in employment.

Research questions and/or hypotheses

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

1. The overall employment would have seen a significant fall during the covid period.
    a. H_0: Employment rate is same before and during COVID.
    b. H_a: Employment rate is lower during COVID.
2. Do certain industry sectors have higher correlation with the confirmed cases than others?
    a. H_0: Employment rate is same across different industry sectors.
    b. H_a: Employment rate is different in certain sectors during COVID.
3. Do we see a positive correlation between confirmed cases and unemployment rate, and a negative correlation between vaccination and unemployment rate?
4. How long did it take for the employment to regain its normal state after the start of vaccination, masking policy?
5. Do we see a significant change in employment with the masking policy changes?

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, and vaccination. The results might show us a lack of evidence to

support our hypothesis, in which case further analysis is performed by bucketing the data into different categories like pre/post masking policy, pre/post vaccination. This will enable us to obtain more meaningful correlations for each timespan buckets.

Data to be used

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
COVID-19 Vaccinations in the New Haven County	https://data.cdc.gov/Vaccinations/COVID-19-Vaccinations-in-the-United-States-County/8xkx-amqh/data

Unknowns and dependencies

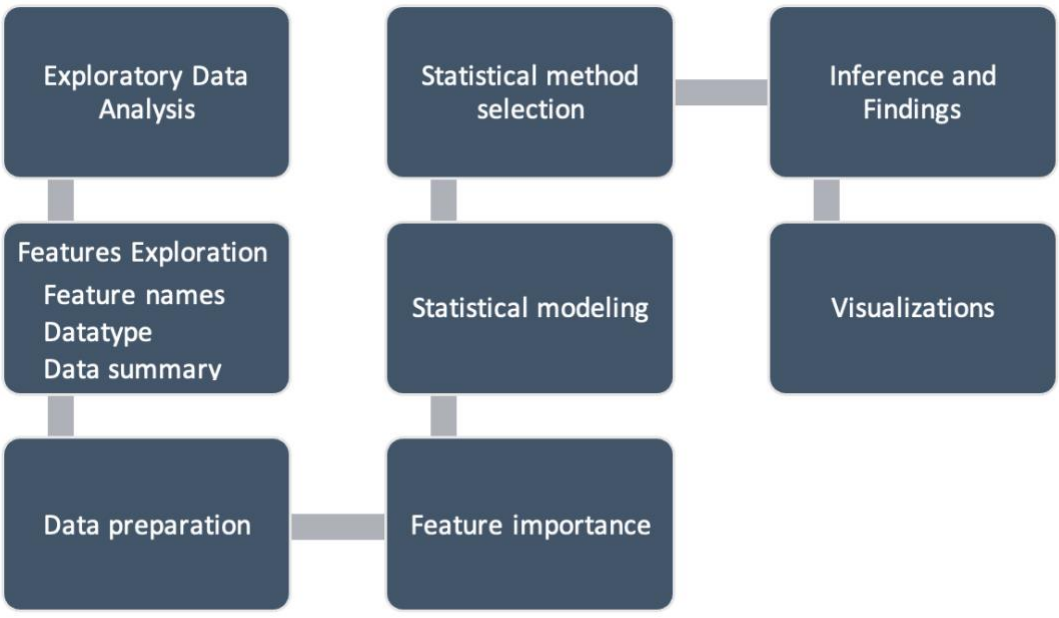
One of the key considerations in our analysis is that there are a number of factors that would have impacted the employment conditions in the New Haven County during the pandemic, however, we will be focusing only on certain factors related to confirmed cases, vaccination and masking data. There is possibility 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. Our analysis would not be considering these inaccuracies along with other factors that are beyond the scope of this project.

Methodology

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 several statistical approaches like correlation, linear regression, and statistical testing methods like ANOVA and T-test. Each of these methods have different aspect of explainability and we will be choosing the right method based on the data and question. We will further need to 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.

Timeline to completion

Date	**Task scheduled for completion**
Nov 10, 2022	Problem statement submission (Part 2) Data source curation, methodology and plan of execution Deliverable: Extension plan (Part 2 assignment submission)
Nov 16, 2022	Data curation, cleaning, and exploratory data analysis (EDA)

	Collect the data from the mentioned sources, extract, filter the data by county and perform exploratory data analysis to understand the general trends in employment based on vaccination, confirmed cases and death rate.
Nov 20, 2022	Data modeling and analysis Analyzing the change in employment in different sectors with the progression of daily cases, fatalities and vaccinations in the county
Nov 25, 2022	Inference and fine tuning Derive inferences from the modelling performed on employment data, see the possibility of finding different trends over time.
Dec 1, 2022	Visualization and result documentation Create plots to visualize interesting pattens that support our initial intuition and hypothesis
Dec 5, 2022	Presentation of findings (Part 3) Prepare a story using the finding from my analysis to deliver in the class on how covid cases and vaccination has affected employment in the New Haven County. Deliverable: Pecha Kucha presentation
Dec 12, 2022	Final report (Part 4) Submitting the final report with details about the analysis, inferences, caveats, and any possible gaps in the findings that can be addressed using further datasets. Deliverable: Final Report