# Fall 2020 CS 687 Capstone Project
# Progress Report

# COVID-19 Effects on the US Labor Market

Ruan Almeida
Advisor: Sion Yoon
MS in Computer Science
School of Technology & Computing (STC)
City University of Seattle (CityU)
almeidaruan@cityuniversity.edu, yoonhee@cityu.edu

## Abstract

The ongoing COVID-19 recession is already one of the major global economic crises that modern society has faced, causing abnormally high and rapid increases in unemployment rates in many countries. The lack of perspective about market heat has also caused a sharp drop in the hiring numbers, with the job opportunities dispute becoming more competitive. The situation can lead to an inability of the United States to provide state-funded unemployment insurance. The proposal for this project is to implement a monitoring application able to combine data from US labor market and healthcare datasets. The goal is to present a meaningful relationship between COVID-19 cases, unemployment and job opening rates, that can be used for statistical analysis and forecast. Data analytics approaches such as data wrangling and data preprocessing are described in the implementation methodology. The system is based on technology tools and development languages with support for data analysis such as Python (libraries Pandas and NumPy) and R (R Markdown package). The research demonstrates that after the initial impact caused by the COVID-19 pandemic lockdown, job opening numbers have gradually recovered their previous figures, but unemployment is still high. That indicates a shift in the workforce, with emerging professionals taking the jobs that used to belong to the previous workers. I expect that academics, data analysts, and other kinds of professionals can benefit from this project – both the application and the methodology – using it as a source for different analyses that can shed light on the impacts of the COVID-19 downturn in the US labor market.

**Keywords:** Data analytics, unemployment rate, hiring numbers, job openings, COVID-19, labor market, monitoring software, Big Data, R Markdown, data wrangling, data preprocessing.

## 1. INTRODUCTION

**Problem Statement**
The Coronavirus (COVID-19) changed the lives of almost every human being in 2020 and brought considerable impacts to the labor market worldwide. There are many statistical data available related to both COVID-19 and hiring/unemployment numbers during the pandemic period. However, none of them were able to combine that information in a way that allows us to best measure those impacts.

The proposal for this project is to implement a monitoring application able to combine data from unemployment/hiring numbers datasets and COVID-19 cases datasets, and to provide a view that can establish a meaningful relationship between different official data sources such as the U.S. Bureau of Labor Statistics and the World Health Organization (WHO). The main obstacle is to be able to create a methodology for data preprocessing and data cleaning that will ease retro alimentation of the system with updated information on a regular basis.

The goal is to present a dynamic system using data analytics procedures and techniques to be able to provide combined quantitative data that can be used as input for relevant information about the effects of the pandemic on the professionals in the US labor market.

**Motivation**
There has been a significant economic impact on people's lives around the world due to COVID-19. In three months of the pandemic, the number of unemployed in the USA increased by more than 14 million, going from 6.2 million in February this year to 20.5 million in May 2020 (Kochhar, 2020). Those numbers are worse than the numbers registered during the Great Recession period when unemployment was increased by around 9 million from 2008 to 2009 (U.S. Bureau of Labor Statistics, 2020). That information exposes how alarming the situation is and justifies the current demand for a monitoring application.

**Approach**
After the initial evaluation of some previous work related to this topic, it was possible to confirm relevant gaps where this project can focus in order to bring significant contributions. All available information is static and cannot be transformed to allow different analyses and perspectives. At the same time, the literature review brought important inputs for my project that can be used as parameters for the proposed system.

Regarding the application's development, data analytics approaches such as data wrangling and data preprocessing (which involves data cleaning, data integration, data reduction and data transformation) will be applied. Technology tools and languages with support for data analysis such as Python (libraries Pandas and NumPy) and R (R Markdown package) will be explored for the implementation of the system.

**Conclusions**
Data science and, more specifically, data analytics, are trend topic nowadays, both in academics and corporate environments. IT 4.0 and big data technology have wide opened the range where we can apply data analytics procedures and techniques. COVID-19 pandemics have caused impacts in so many different areas that it is difficult to narrow down the vast data into meaningful parameters. As an outcome for this project, I hope to deliver an application and a methodology that can be used by academics, data analysts, and other kinds of professionals as a relevant source for analyses that can shed light on the impacts of the COVID-19 downturn in the USA labor market.

## 2. BACKGROUND

To provide an overview of my topic, I performed a background analysis based on different scholarly sources, research, and peer-reviewed publications related to the COVID-19 pandemic and its effect on employment and labor. The central questions I target for my topic were:

- How COVID-19 is affecting different groups of employees?
- How COVID-19 is affecting different segments of employers?
- How COVID-19 can be related to the unemployment rate and hiring numbers?

The background analysis of each of those key aspects – COVID19 and its effects on employees, employers, unemployment rates and hiring numbers – provided me numbers to evaluate my assumptions. The current information about those aspects is very comprehensive when it comes to each one of them (as it will be demonstrated in section 3), but there is no direct link or clear correlation among them.

## 3. RELATED WORK

As a result of my literature review, I was able to establish a comprehensive look at previous writings around my topic prior to the arguments I intend to present in my study. I was also able to identify potential gaps I can explore to add value to my project.

According to Rakesh Kochhar, a senior researcher at Pew Research Center, in the first three months of the COVID-19 pandemic, unemployment rose higher than it did in two years of the Great Recession. In his research, he was able to estimate that the number of unemployed in the USA increased by more than 14 million, from 6.2 million in February to 20.5 million in May 2020. This number is significantly higher than the data related to the Great Recession period when unemployment numbers increased by around 9 million from the end of 2007 to the beginning of 2010 (U.S. Bureau of Labor Statistics, 2020). This data reflects the massive unemployment rate increase, which according to Kochhar's research, leaped this year from 4% in February to more than 14% in April. After performing an analysis of the US unemployment rate records of the past ten decades, I was able to observe that unemployment numbers went from the lowest values in post-World War II to the second highest in this era (U.S. Bureau of Labor Statistics, 2020). As shown in Figure 1, the unemployment rate starts to decrease gradually as the COVID cases curve starts flattening (after April), which causes a market heat. This effect echoes with different other references I found. That supports my initial assumption of a strong relationship between the pandemic and the US labor market.

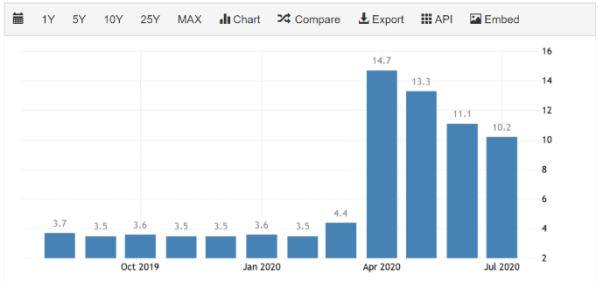


Figure1: US unemployment rate in the past months (U.S. Bureau of Labor Statistics, 2020)

Stepping further in my assumptions, I sliced and drilled down the related works I found to verify the effects of this unemployment increase during COVID-19 recession among different groups of workers. The economic crisis vastly increases unemployment and creates competition between different kinds of workers (Blustein et al., 2020). As shown in Figures 2 and 3, Kochhar's research indicates that women – especially Hispanic – experienced a sharper rise in the unemployment rate than men. Immigrants experienced higher impacts than US-born workers, as shown in Figure 4. This is consistent with the indications that the minorities were more impacted by the abrupt increase in unemployment (Fairlie et al., 2020).


Figure2: Unemployment rate – women vs. men (Kochhar, 2020).

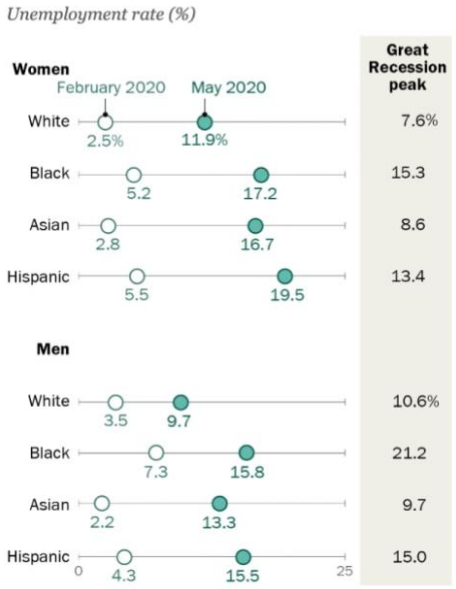


Figure3: Unemployment rate among racial and ethnic groups (Kochhar, 2020)

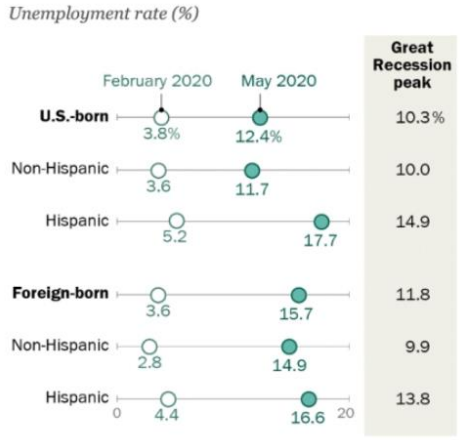


Figure4: Unemployment rate – US-born vs. Foreign-born (Kochhar, 2020)

COVID-19 downturn seems to have affected emergent workers more severely than workers more than 24 years old. As shown in Figure 5, one in every four young professionals between 16 and 24 years old are unemployed. Kochhar's analysis also indicates that the impacts are lower among professionals with higher levels of education, as shown in Figure 6.

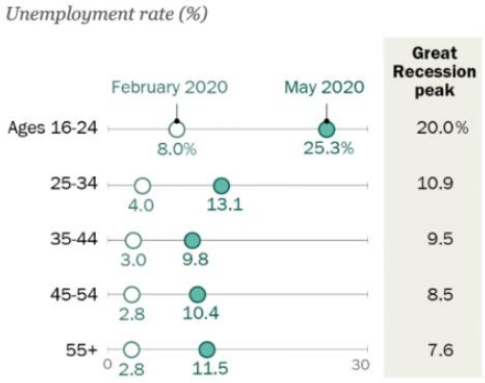


Figure5: Unemployment rate in different age groups (Kochhar, 2020)

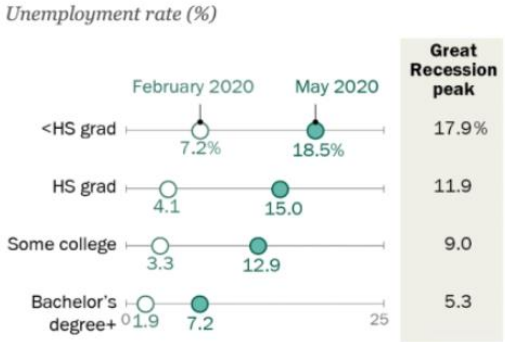


Figure6: Unemployment rate in different education levels (Kochhar, 2020)

Kochhar's research brought me excellent resource and data to evaluate the impacts of the pandemic in unemployment rates, but it does not directly associate that data with job opportunities and market expectations in terms of hiring. In order to close this gap, I needed other references that would provide me insight into the economic impact of COVID-19 downturn on businesses. I was able to obtain that information on a survey conducted by researchers Alexander W. Bartik (Department of Economics, University of Illinois), Marianne Bertrand (Booth School of Business, The University of Chicago), Zoe Cullen, Michael Luca, Christopher Stanton (Harvard Business School, Boston) and Edward L. Glaeser (Department of Economics, Harvard University). This team conducted a survey with more than 5.800 companies between March 28 and April this year as an attempt to shed light on the impacts of the pandemic on business. The results indicate both the financial fragility of many small businesses and the lack of expectations about a short-term improvement of the situation. Those combined factors make the risk of investment in new projects too high, preventing the heat of the market.

Bartik and his team's work suggests that the pandemic had already caused mass layoffs and small business closures a few weeks immediately after its inception. This effect was increased because there was not enough time for governmental support through the Coronavirus Aid, Relief, and Economic Security (CARES) Act. Three-quarters of the survey respondents revealed that they only had enough funding to last around two months. Small businesses with more than $10,000 in monthly expenses had only about two weeks of available money to maintain the operations in the current circumstances. The data indicates that 43% of businesses had temporarily closed, and active employment was reduced by 39% (Bartik et al., 2020). The research also shows that small businesses employ almost 50% of American workers, which makes them extremely relevant to my project.

In another related work, former Chief Economist - Department of Labor, Harry J. Holzer, attempted to compare US unemployment rate and COVID-19 outcomes with other OECD (Organization for Economic Co-operation and Development) nations. His study focused on the 25 richest OECD countries, and he used data from two reliable datasets: Trading Economics for the economic data and the Johns Hopkins University Coronavirus Resource Center for information on COVID-19 cases and mortality. The relevance of

his finding for my project can be explained: by comparing employment and health data for the USA with those of other OECD nations (as shown in Table 1), I will be able to establish the level to which USA federal response to the crisis has either mitigated or exacerbated the damage. Furthermore, his research demonstrates that Unemployment Insurance (UI) claims in US have risen by several million workers in the four weeks followed by the stroke of the pandemic, in April.

| | Others | U.S. |
|---|---|---|
| **Economic Outcomes** | | |
| Unemployment Rates (%) Jan.-Mar. 2020 | 0.32 | 0.8 |
| Unemployment Rates (%) Jan.-April 2020 | 1.44 | 11.1 |
| **Viral Cases and Deaths** | | |
| Cases/Thousand | 3.2 | 5.3 |
| Deaths/Thousand | 0.2 | 0.3 |
| New Cases/Million | 21.5 | 69 |
| New Deaths/Million | 2 | 3.7 |

Table1: Changes in employment and virus outcomes: US v. other OECD countries (Holzer, 2020)

The reviewed literature indicates that there are enough data available related to both US labor market and COVID-19 downturn that will enable my project to implement a dynamic system capable of performing different analytic views. Furthermore, I have concluded that my proposal can fill a gap in the current research to provide a more specific analysis related to job hiring numbers.

## 4. APPROACH

Selecting the appropriate research methodology is of great importance for defining the approach on an academic project. Nevertheless, when it comes to a data analytics project, it is also essential to determine the most suitable method of collecting and handling data.

To reach the goal of presenting a dynamic system able to provide combined data that can be used as input for relevant information about the effects of the pandemic in the US labor market, I focused my approach on quantitative methods. Quantitative research is based on a systematic empirical investigation of observable phenomena via mathematical, statistical, or computational techniques. It can be directly related to collecting sample surveys or other quantitative datasets to contribute to the body of substantive knowledge in the data analysis field (Treiman, 2014). It is important to consider data preparation as an iterative, multidisciplinary process that depends on setting applicable routines and solutions for the most common sources of error (Endel & Piringer, 2015). Because of that, it was essential to design a data wrangling and data preprocessing methodology to reach the best approach to handle the transformation of the raw data into relevant information. The challenge of reducing the gap between the user requirements and the designer is a key aspect of research (Butkiene & Butleris, 2018). This methodology will also allow me to better address user requirements with high quality and presentable information.

### Data Wrangling

A good data wrangler approach for this project will allow me to integrate information from different data sources, handle common transformation issues and solve data cleansing quality problems. To implement that approach, I designed a methodology that will consider six core activities:

i. Discovering: Before implementing the data cleaning methods, I will perform an analysis of the data sources in order to understand them better (especially their patterns and correlations). This step includes some aspects of Exploratory Data Analysis (EDA), a key part of the data science process.
ii. Structuring: Evaluate the need to reshape, order, or merge the raw data to be suitable to my resulting dataset.
iii. Cleaning: Fix missing data values (null values) and some data types that usually come in different formats (e.g., date) and can cause issues in my database
iv. Enriching: Evaluate the inclusion of additional data that brings the outcome closer to the original proposal.
v. Validating: Assurance of the quality and consistency of the resulting data. Make sure that the result is relevant to the proposed context.
vi. Publishing: Make sure that the wrangled data is going to be presented in an appropriate format and that it can be available for my target audience.

### Data Preprocessing

Together with the proposed data wrangling methodology, it is also necessary to establish an approach for data preprocessing, with techniques to help improve the quality of the data and make the data mining efforts easier. Basically, data preprocessing can be described as a set of data mining techniques used to transform raw data into efficient and useful formats. (Patil &

Hiremath, 2018) Different methods of data preprocessing have been implemented, and there are some important aspects to evaluate when preprocessing data. Han, Kamber & Pei defined those aspects in four groups of techniques: data cleaning, data integration, data reduction and data transformation (Han et al., 2018).

As mentioned in the data wrangling section, data cleaning is the process of handling and/or filling in missing data. It also involves reducing data noise, removing outliers, and solving inconsistencies. To handle those problems, the designed approach will include:

- Delete records with null attributes.
- Consider the average of values for equivalent attributes.
- Consider the median value for equivalent attributes.
- Fill missing attributes with the most frequent values in the dataset.

Data integration is, essentially, the process of merging data from different data sources. The most common issues data analysts find while performing data integration are associated with inconsistencies and redundancies. The first step to deal with those issues is to establish a data validation mechanism for the integration procedures. Adding up to this mitigation step, I will evaluate and consider including automated data integration tools and data transformation tools.

The data mining involved in this project will handle large volumes of data. To reduce data complexity and increase efficiency, the following data reduction techniques are being considered in this methodology:

- Attribute subset selection, to eliminate irrelevant or redundant attributes.
- Numerosity reduction, to replace the original data by smaller form of data representation (applying practices such as grouping, sampling, or using histograms).
- Prioritization of highly relevant attributes to the detriment of less relevant ones.

Finally, data transformation techniques will be performed to convert the original data into more suitable formats for the mining process. In a project of this nature, there might be issues related to unbalanced data. Also, different scales of data and data values out of range are common problems. To deal with those problems, the implementation will include the following techniques:

- Normalization, to scale data values over a specified range.
- Attribute construction, with new attributes being generated from the provided set of attributes to assist in the mining process.
- Concept hierarchy generation for nominal data, with attributes being converted to a higher level in the hierarchy.

Nowadays, when it comes to data analytics, it is crucial to think about how to implement a high-performance platform to efficiently analyze big data and how to design an appropriate mining algorithm to best extract useful data. (Tsai et al., 2015). With my implementation methodology, I expect to reach my objectives in terms of data wrangling and data preprocessing, that consists of:

- Establish a rational method for gathering data from different sources.
- Gather accurate and actionable data as input for my monitoring application.
- Avoid issues and time wasted in fixing problems caused by dirty (unreliable) data.

## 5. DATA COLLECTION

For this project, the main method of data collection is the stratification of data from databases and case studies. Since I am integrating big data, it would not be efficient to do questionnaires and checklists. The time constraint for this project also does not allow me to collect data from human subjects through interviews and surveys. The data that is being used as input for the system is coming from different sources and under different conditions. For the data analysis, it is necessary to separate the data that were coming from those different sources and with different conditions. In fact, the stratification procedure was established even before the data collection process, as a part of the data collection plan.

The data collection plan consists of three aspects:

- Identify what data is needed.
- Define how the data will be collected.
- Delineate how the data will be analyzed (data assessment).

**Data needs**

The fundamental inputs for the monitoring application are:

- Data about COVID-19: number of cases, segmented by location (US cities or states) and time (monthly figures).
- Data about unemployment: official monthly rates in US cities or states.

- Data about hiring: job opening numbers segmented by location (US cities or states) and time (monthly figures).

Monthly updates (at least) must be provided for all the needed data.

**Data collection**

There is a multitude of resources related to COVID-19 available on the Internet nowadays. However, these resources are spread all over different databases, and are often buried by a mass of information that does not concern this research (e.g., shocking headlines, number of deaths, names of infected celebrities, etc.). The same goes for data about unemployment and hiring numbers (e.g., unemployment rate by different sectors, companies showing off how they assisted health care agencies, etc.). To deal with this issue, I applied the data wrangling and data preprocessing methodology I designed, as explained in section 4 of this paper.

For the COVID-19 data collection, I was able to extract the data I needed from the Johns Hopkins University Center for Systems Science and Engineering (JHU CCSE), Coronavirus repository (Johns Hopkins University & Medicine, 2020). The raw data arranged by JHU CCSE pulls data from other several different data sources, such as the World Health Organization (WHO) and the Centers for Disease Control and Prevention (CDC). One other major advantage of using the JHU CCSE database is that I was able to easily implement an access routine to it in my R Markdown application through the *coronavirus* package (Krispin & Byrnes, 2020).

The decision of using R Markdown as the development language proved to be an excellent choice for this project. R Markdown provides not only an authoring framework for data science but also various packages that were extremely beneficial and reduced a lot of the coding efforts. Plus, R Markdown applications are based on metadata, and allows different chunks of code from different languages in the same code script. This flexibility came in handy since part of my data collection procedures were implemented in R and part in Python.

After installing the R Markdown *coronavirus* package in my VSCode environment, I had access to a vast and tidy dataset of the COVID-19 epidemic. It provided me the fields that I needed for my stratification procedure and filters. Also, it made my data wrangling and data preprocessing efforts easier. Some of the most relevant fields that I explored from the *coronavirus* dataset (such as date, province, country, type, and cases) are shown in Table 2.


Table2: Relevant fields from coronavirus dataset used in my application

The *coronavirus* pack comprehensive R archive network (CRAN) version is updated monthly. There is also the possibility of using the Github (Dev) version, which is updated on a daily basis. To keep the dashboard with the most recent information, I just need to invoke the *update_dataset()* function and refresh the application.

Similarly, for data about unemployment, I used R packages *rUnemploymentData* and *blsAPI*. The *blsAPI* is an R package that allows users to request and extract data from the U.S. Bureau of Labor Statistics (BLS) datasets through its application programming interface (API). The *blsAPI* gives public access to economic data from all BLS programs (U.S. Bureau of Labor Statistics, 2020). It enabled me to pull monthly unemployment and labor force estimates to my monitoring application.

For data about hiring, I used R Markdown to scrape HTML tables from the U.S. Bureau of Labor Statistics database. For Instance, The Economic News Release contains information about job openings levels and rates by region, seasonally adjusted, as shown in Table 3. The job openings rate is the number of job openings on the last business day of the month as a percent of total employment plus job openings.

| Industry and region | Rates[2] | | | | | |
|---|---|---|---|---|---|---|
| | Sept. 2019 | May 2020 | June 2020 | July 2020 | Aug. 2020 | Sept. 2020[P] |
| Total | 4.4 | 3.9 | 4.2 | 4.6 | 4.3 | 4.3 |
| INDUSTRY | | | | | | |
| Total private | 4.7 | 4.1 | 4.4 | 4.7 | 4.5 | 4.6 |
| Government | 3.1 | 2.9 | 3.0 | 3.7 | 3.2 | 3.1 |
| REGION[4] | | | | | | |
| Northeast | 4.1 | 3.7 | 4.3 | 4.5 | 4.0 | 4.1 |
| South | 4.6 | 4.0 | 4.2 | 4.6 | 4.6 | 4.6 |
| Midwest | 4.5 | 3.7 | 4.2 | 4.9 | 4.3 | 4.3 |
| West | 4.4 | 3.9 | 4.0 | 4.3 | 4.1 | 4.1 |

Table3: Simplified view of job openings rates by industry and region (U.S. Bureau of Labor Statistics, 2020)

**Data assessment**

The delineation about how the data will be analyzed starts with a data stratification procedure. Data stratification is the partition of data into smaller and well-defined strata (layers) based on predetermined criteria. For this project, I used a simple stratification procedure for the data analysis, based on:

i. Different colors for each source
ii. Different charts (such as control charts, bubble charts, histogram charts, scatter diagrams, etc.).

With this procedure, I was able to drill down the information and perform more specific analyses, as shown in Figures 7, 8 and 9.

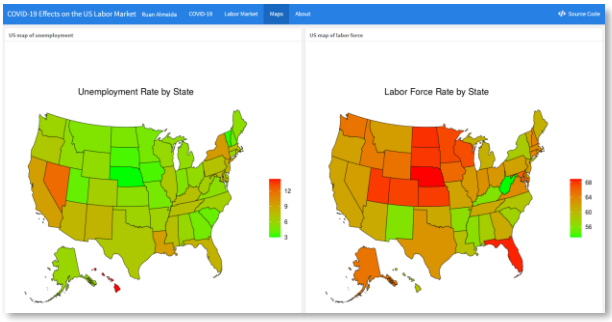


Figure7: US labor market figures by state (Almeida, 2020)


Figure8: US unemployment and job openings situation (Almeida, 2020)


Figure9: US COVID-19 situation (Almeida, 2020)

R Markdown supports dozens of static and dynamic/interactive output formats, making it easier to generate the different views that I needed to interpret the data.

The following packages were used for the dashboard interface and the visualizations:

- *flexdashboard*: This package enabled me to compile a group of related data visualizations as a dashboard and create interactive dashboards. (R Studio, 2020).
- *plotly*: This graphing library enabled me to produce interactive, publication-quality graphs, and generate the different kinds of charts I needed. (Plotly, 2020).
- *leaflet*: This package enabled me to create interactive maps. (R Studio, 2020).

The following packages were used for data manipulation:

- *dplyr*: A grammar of data manipulation that provided me a consistent set of verbs that helped to solve the most common data manipulation challenges. (Tidyverse, 2020).
- *tidyr*: This package provided me a standard way of storing data and manipulate columns and rows in a table. (Tidyverse, 2020).

Finally, to assure that the algorithm will have no performance issues, the designed data collection plan considers some data quality key performance indicators (KPI), based on:

- Validity.
- Reliability.
- Completeness.
- Precision.
- Integrity.

- Timeliness.

Adding up to that, as a general usability rule, the presentation of the data results was kept as simple as possible, with clear visualization to strengthen the methods.

## 6. DATA ANALYSIS

My analysis of the collected data was focused on evaluating the effects of COVID-19 downturn on the US labor market. More specifically, I assessed the following hypotheses:

i. The increase of COVID-19 cases is directly proportional to the US unemployment rates.
ii. The increase of COVID-19 cases is inversely proportional to the US job openings.

To perform this analysis, my strategy based on:

a. Identify the five US states that currently have the highest number of COVID cases.
b. Identify the five US states that currently have the lowest number of COVID cases.
c. Check the unemployment rates and job openings of those ten states in SEP-2019 and SEP-2020. Calculate the gap (proportional variation) between those months.
d. Compare the gaps of the states more affected by COVID with the gaps of the states less affected.

To implement this strategy using the collected data, it was necessary to use techniques such as grouping, filtering, summarizing, sorting, categorizing, and comparing. For instance, my COVID dataset has *counties* as an attribute. Because of that, I had multiple instances for the *state* attribute instead of a single instance. I had to group and summarize the data per state before sorting and comparing. Also, for this analysis, I only intended to consider contiguous states. So, I had to filter the dataset to exclude Hawaii and Alaska (I also removed Washington DC during this process). As a result, I generated the information I needed for steps a and b, as shown in Tables 4 and 5.

| | State | Region | COVID Cases |
|---|---|---|---|
| 0 | Texas | South | 1,082,625 |
| 1 | California | West | 1,026,918 |
| 2 | Florida | South | 875,088 |
| 3 | Illinois | Midwest | 564,086 |
| 4 | New York | Northeast | 561,308 |

Table4: Top 5 US states affected by COVID cases (Almeida, 2020)

| | State | Region | COVID Cases |
|---|---|---|---|
| 0 | Vermont | Northeast | 2,843 |
| 1 | Maine | Northeast | 8,791 |
| 2 | New Hampshire | Northeast | 14,311 |
| 3 | Wyoming | West | 21,881 |
| 4 | Delaware | South | 28,395 |

Table5: Bottom 5 US states affected by COVID cases (Almeida, 2020)

The next steps involved getting the hiring and the unemployment data, on specific dates, for each one of the ten states identified in the previous steps. For that, I used the categorizing and comparing techniques. It was necessary to perform a cross-tabulation analysis between tables from different datasets, as shown in Tables 6 and 7. Cross-tabulation analysis, also known as contingency table analysis, is a powerful analytical tool often used to analyze categorical (nominal measurement scale) data. In that case, I used it to compare the variables COVID cases, unemployment rates and job openings, all connected by the attribute "state". However, first I needed to solve one more issue. This one is related to data preprocessing: my data collection related to hiring information (job offers) does not have this information per state. The lowest level I was able to reach while drilling-down was "job openings per region". To solve this problem, I resorted to the designed methodology (section 4 of this paper). The methodology specific approach for data cleaning involves filling in missing values. In this case, what I did was to include (fill), for each *state* instance, the job numbers I had for its respective region as shown in Figure 10. The result, of course, is not precise, but it is reasonable enough for this analysis. In fact, there

is not too much variation on those numbers, as we can see in Table 7. In a deeper analysis, I could notice that the most noticeable impact related to hiring happened in MAR-2020, when the quarantine order spread over the country.

| | State | Region | Rate SEP-19 | Rate SEP-20 | Over-the-year change |
|---|---|---|---|---|---|
| 0 | Texas | South | 3.5 | 8.3 | 4.8 |
| 1 | California | West | 3.9 | 11.0 | 7.1 |
| 2 | Florida | South | 2.9 | 7.6 | 4.7 |
| 3 | Illinois | Midwest | 3.7 | 10.2 | 6.5 |
| 4 | New York | Northeast | 3.9 | 9.7 | 5.8 |
| 5 | Delaware | South | 3.9 | 8.2 | 4.3 |
| 6 | Wyoming | West | 3.7 | 6.1 | 2.4 |
| 7 | New Hampshire | Northeast | 2.6 | 6.0 | 3.4 |
| 8 | Maine | Northeast | 2.9 | 6.1 | 3.2 |
| 9 | Vermont | Northeast | 2.4 | 4.2 | 1.8 |

Table6: Unemployment rates gaps (Almeida, 2020)

| | State | Region | Rate SEP-19 | Rate SEP-20 | Over-the-year change |
|---|---|---|---|---|---|
| 0 | Texas | South | 4.6 | 4.6 | 0.0 |
| 1 | California | West | 4.4 | 4.1 | -0.3 |
| 2 | Florida | South | 4.6 | 4.6 | 0.0 |
| 3 | Illinois | Midwest | 4.5 | 4.2 | -0.3 |
| 4 | New York | Northeast | 4.1 | 4.1 | 0.0 |
| 5 | Delaware | South | 4.6 | 4.6 | 0.0 |
| 6 | Wyoming | West | 4.4 | 4.1 | -0.3 |
| 7 | New Hampshire | Northeast | 4.1 | 4.1 | 0.0 |
| 8 | Maine | Northeast | 4.1 | 4.1 | 0.0 |
| 9 | Vermont | Northeast | 4.1 | 4.1 | 0.0 |

Table7: Job openings gaps (Almeida, 2020)

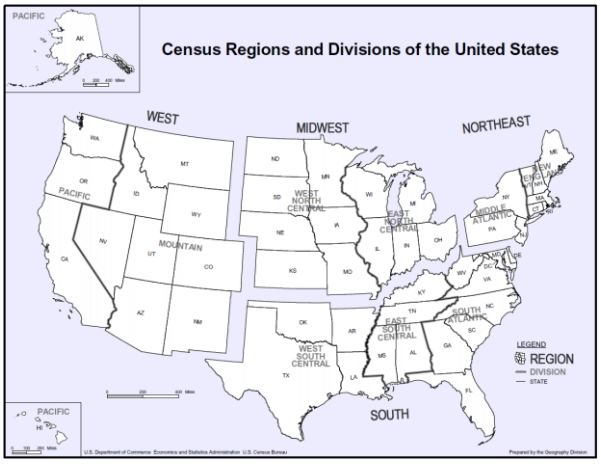


Figure10: Regions vs. states (US Census Bureau, 2020)

As one of the takeaways of this data analysis process, it became even more clear to me the value of a dynamic system that presents an update dashboard with both simple and straight to the point information about the COVID-19 effects on the US labor market.

## 7. FINDING

The monitoring application offered a timely insight into what US workers are facing, along with an indicator of labor market conditions. The analysis of the system's data demonstrated that the unemployment rate, as widely informed by the media, has increased accordingly with COVID-19 cases. According to the data presented in Table 6, the average gap of the top five US states affected by COVID-19 cases – Texas, California, Florida, Illinois, and New York – is 5.78. The same average gap of the bottom five US states affected by COVID-19 cases – Vermont, Maine, New Hampshire, Wyoming, and Delaware – is 3.02. That indicates a 2.76 higher average gap at the unemployment rate in the states that were more affected by COVID-19 cases.

Meanwhile, the system's data demonstrated that there were no significant gaps in terms of job openings, in the same period, for that group of states (see Table 7). As mentioned before, a deeper analysis revealed there was a massive impact in terms of hiring in MAR-2020 when the quarantine order spread over the country but after that the numbers started to gradually increase until SEP-2020 (as shown in Table 8).

| State | Rate SEP-19 | Rate MAY-20 | Rate JUN-20 | Rate JUL-20 | Rate AUG-20 | Rate SEP-20 |
|---|---|---|---|---|---|---|
| Texas | 4.6 | 4.0 | 4.2 | 4.6 | 4.6 | 4.6 |
| California | 4.4 | 3.9 | 4.0 | 4.3 | 4.1 | 4.1 |
| Florida | 4.6 | 4.0 | 4.2 | 4.6 | 4.6 | 4.6 |
| Illinois | 4.5 | 3.7 | 4.2 | 4.9 | 4.3 | 4.2 |
| New York | 4.1 | 3.7 | 4.3 | 4.5 | 4.0 | 4.1 |
| Delaware | 4.6 | 4.0 | 4.2 | 4.6 | 4.6 | 4.6 |
| Wyoming | 4.4 | 3.9 | 4.0 | 4.3 | 4.1 | 4.1 |
| New Hampshire | 4.1 | 3.7 | 4.3 | 4.5 | 4.0 | 4.1 |
| Maine | 4.1 | 3.7 | 4.3 | 4.5 | 4.0 | 4.1 |
| Vermont | 4.1 | 3.7 | 4.3 | 4.5 | 4.0 | 4.1 |

Table8: Job openings in 2020 by month (Almeida, 2020)

In order to provide additional findings, I dove deeper into those results, and combined them with findings from related work. I started with an extensive evaluation of the labor market situation in California, the state that presented the highest unemployment gap of all (7.1, as shown in Table 7). The rates topped out at 14.5% and 16.4% in the US and California, respectively, before starting to fall. These unprecedented unemployment rates are unlikely unrelated to the impact COVID-19 induced in the labor market. It is important to emphasize that unemployment rates in the US were at historic lows before the

pandemic (U.S. Bureau of Labor Statistics, 2020). This echoes with the fact that, throughout the decade prior to the pandemic, the US was experiencing a historical period of monthly job opportunities increase (U.S. Bureau of Labor Statistics, 2020). Starting from October 2010 US job growth averaged 196,000 a month, totaling 22.1 million before the virus-induced halt (U.S. Bureau of Labor Statistics, 2020). California's share of that job growth was 3.3 million (15% of the US total). Despite that, I could notice that after the COVID-19 stroke, losses in California have been relatively larger than the country overall (as indicated in Table 7). The swift and devastating magnitude of job losses during April and May of 2020 was also unprecedented. However, according to the data presented, hiring numbers seemed to recover faster than unemployment numbers. That led me to the following findings:

- The sudden implementation of public health precautions that included shuttering businesses and sending workers home as an attempt to control the spread of the novel coronavirus had a direct impact on the increase of the unemployment rates in the US. It also caused an immediate market reaction of drastically hiring-freeze at the beginning of the pandemic.
- Thousands of people lost their jobs, but only a small part of those was able to reintegrate the workforce, as the unemployment rate keeps high even though the job offerings are coming back to the same numbers of 2019. Combined with the findings of the related work described in section 3, this indicates a possible shift in the workforce, with emergent (younger) and more educated professionals taking the jobs that belonged to the previous workers.
- Related work also indicates that a good portion of the people who lost their jobs claimed and are using the state-funded unemployment insurance benefit before seeking jobs again. This contributes to the contrast between unemployment and hiring data curves.

## 8. CONCLUSION

The current moment is one of uncertainty and transformation. The impacts on the economy and on the organizations due to the COVID-19 pandemic is inevitable. As a result, Human Resource (HR) areas – responsible for hiring and firing employees – are also hit. Despite many statistical data available related to COVID-19, hiring and unemployment numbers, none of them provides a combined view of those aspects to best demonstrate the mentioned impacts. To narrow down the vast related data into meaningful information for this research, I proposed the implementation of a monitoring system using data analytics procedures and techniques. The goal was to provide combined quantitative data (extracted from different datasets) related to the unemployment rate, job opening rate and COVID-19 cases on a time basis. This metadata can be used as input for relevant evaluation about the effects of the pandemic on the professionals in the US labor market. The system allowed me to evaluate the assumptions that the increase of COVID-19 cases is directly proportional to the US unemployment rates and inversely proportional to the US job openings. Based on all the data, graphs and findings provided along this paper, it is confirmed that unemployment rates and hiring numbers (job openings) were indeed severely impacted by the COVID-19, with massive layoffs and severe hiring-freeze when the pandemic struck. However, while the number of job openings started to increase again, the unemployment rate continues high. That indicates that many people lost their jobs but only a small part of those was able to reintegrate the workforce. One of the explanations for this contrast between unemployment and hiring data could be extracted from related work, which indicates a possible shift in the workforce, with emerging (younger) and more educated professionals taking the jobs that used to belong to the previous workers. Adding up to that, it was also indicated that a good portion of the people who lost their jobs claimed and are using the state-funded unemployment insurance benefit before seeking jobs again.

It was also reasonable to conclude that we might face volatile trends in job growth for more time. Although a hasty rebound is not discarded if the virus is reduced to a controlled level (or with an eminent discovery of a vaccine), given the lack of an effective national public strategy to maintain the COVID-19 cases low, it is more likely in the coming months that the trend in jobs may exhibit a hectic and disturbed trajectory. That justifies once more the importance of a monitoring system able to track this data.

## 9. FUTURE WORK

The application and the methodology created as a result of this project can be used by academics, data analysts, and other kinds of professionals as

a relevant source for further analyses related to the COVID-19 downturn in the USA labor market. The continuation of my work concerns a deeper analysis of my findings, experiments, and tests that can reinforce the fundamental basis of my conclusions. Also, new insights about the behavior of hiring areas can emerge with the confirmation of a vaccine that can change the balance of the current quarantine measures. The monitoring application will allow me to keep tracking for further opportunities to foster my research with new information.

Finally, I also recommend the use of the system for extended research on future trends in virus transmission, customer confidence, and public policies. This can determine when and how fast affected businesses and their related payrolls might rebound.

## 10. REFERENCE


Almeida, Ruan (2020). COVID-19 Effects on the US Labor Market - Monitoring Application (with source code and demo). Retrieved from https://github.com/ruanmurta/Capstone.

Bartik, A., Bertrand M., Cullen Z. (2020). The impact of COVID-19 on small business outcomes and expectations. Retrieved from https://www.pnas.org/content/pnas/117/30/17656.full.pdf

Blustein, Duffy, Ferreira, Cohen-Scali (2020). Unemployment in the time of COVID-19: A research agenda. Retrieved from https://www.sciencedirect.com/science/article/pii/S0001879120300610

Butkiene, R., Butleris, R. (2018). The Approach for User Requirements Specification. Retrieved from https://www.researchgate.net/profile/Rimantas_Butleris/publication/267953604_The_Approach_for_User_Requirements_Specification/links/54d89d3e0cf25013d03e7077/The-Approach-for-User-Requirements-Specification.pdf

Endel, F., Piringer, H. (2015). Data Wrangling: Making data useful again. Retrieved from https://www.sciencedirect.com/science/article/pii/S2405896315001986

Fairlie, Couch & Xu (2020). The Impacts of COVID-19 on Minority Unemployment: First Evidence from April 2020 CPS Microdata. Retrieved from https://www.nber.org/papers/w27246

Han, J., Kamber, M. & Pei, J. (2018). Data Mining: Concepts and Techniques. Morgan Kaufmann (MK), 2018.

Holzer, H. (2020). The COVID-19 crisis: How do U.S. employment and health outcomes compare to other OECD countries? Retrieved from https://www.brookings.edu/research/the-covid-19-crisis-how-do-u-s-economic-and-health-outcomes-compare-to-other-oecd-countries/

Johns Hopkins University & Medicine (2020). Coronavirus Resource Center. Retrieved from https://coronavirus.jhu.edu/

Kochhar, R. (2020). Unemployment rose higher in three months of COVID-19 than it did in two years of the Great Recession. Retrieved from https://www.pewresearch.org/fact-tank/2020/06/11/unemployment-rose-higher-in-three-months-of-covid-19-than-it-did-in-two-years-of-the-great-recession/

Krispin, R., Byrnes, R. (2020). R Markdown Coronavirus package. Retrieved from https://ramikrispin.github.io/coronavirus/

Patil, M., Hiremath, B. (2018). A Systematic Study of Data Wrangling. Retrieved from http://www.mecs-press.org/ijitcs/ijitcs-v10-n1/IJITCS-V10-N1-4.pdf

Plotly (2020). Plotly R Open Source Graphing Library. Retrieved from https://plotly.com/r/

R Studio (2020). flexdashboard for R. Retrieved from https://rmarkdown.rstudio.com/flexdashboard/

R Studio (2020). Leaflet for R. Retrieved from https://rstudio.github.io/leaflet/

Tidyverse (2020). dplyr. Retrieved from https://dplyr.tidyverse.org/

Tidyverse (2020). tidyr. Retrieved from https://tidyr.tidyverse.org/

Treiman, D.J. (2014) Quantitative Data Analysis: Doing Social Research to Test Ideas. Wiley, 2014

Tsai, C., Lai, C., Chao, H., Vasilakos, A. (2015). Big data analytics: a survey. Retrieved from https://journalofbigdata.springeropen.com/track/pdf/10.1186/s40537-015-0030-3

U.S. Bureau of Labor Statistics (2020). Accessing the Public Data API with R. Retrieved from https://www.bls.gov/developers/api_r.htm

U.S. Bureau of Labor Statistics (2020). Job Openings and Labor Turnover Summary. Retrieved from https://www.bls.gov/news.release/pdf/jolts.pdf

US Census Bureau (2020) Census Bureau Regions and Divisions with State FIPS Codes. Retrieved from https://www2.census.gov/geo/pdfs/maps-data/maps/reference/us_regdiv.pdf