ISQS 3358 Written Report:

Business Intelligence – Group 2 Final Project Report

School Name: Texas Tech University

Instructor: David Lucus

Date: December 1st, 2020

Course: Business Intelligence

Group Names: Branson Hebel, Maria Tan de Castro,

Yanely Ramos, Chukwudi Agbakwu

Word Count: 3,101

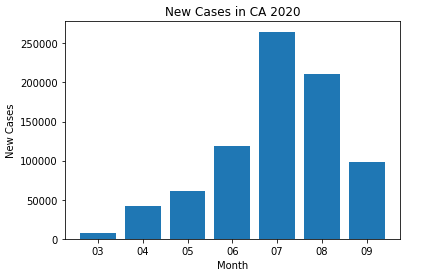
Analysis of Data

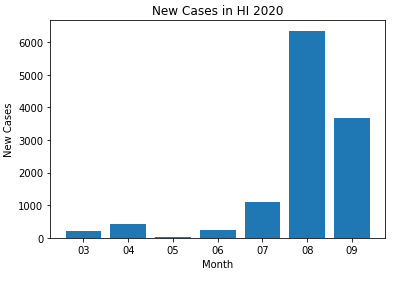
The data we collected for our project was to determine the effect that the Coronavirus Disease 2019 (COVID-19) had on unemployment rates in three highly populated states and a state that is surrounded by water and disconnected from the United States. These states include California, Texas, New York, and Hawaii. The unemployment rate of each state since COVID-19 and the number of positive cases in each state per month are what we derived our variables from.

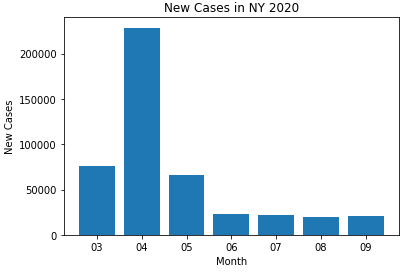
To achieve the analysis of whether COVID-19 had a drastic effect on the unemployment rates for the states mentioned above, we gathered data for the positive cases of COVID-19 in the United States by each state. Additionally, we gathered the data of the monthly unemployment rates of each state since they have reported positive cases. For each state, that dated back to March 2020. Additionally, since the unemployment rate is only measured by month-to-month basis, we had to convert our cases of COVID-19 from a daily amount to a monthly amount, which ultimately only gave us a limited amount of data to work with.

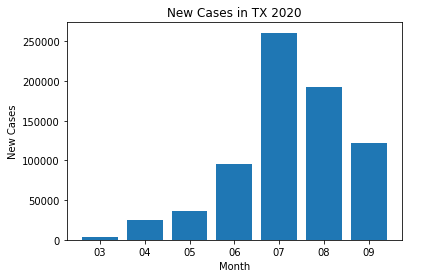
Our analysis determined that COVID-19 did influence the unemployment rate of each state. COVID-19 had the greatest impact on the unemployment rate from March 2020 to April 2020, which was caused by the stay-at-home order that was mandated in March 2020. This caused many companies to furlough many of their employees, which resulted in an increase in unemployment rates for April 2020. From the data analyzed, it was evident that the positive cases of COVID-19 initially drove the unemployment rates upwards, but as time passed and as each state decided to lift their stay-at-home order, unemployment rates for each state began to decrease. Overall, this data told us that the unemployment rate is not directly correlated with the amount of positive COVID-19 cases, but with when each state had their stay-at-home order mandated and lifted.

Although, positive COVID-19 cases did not directly cause the unemployment rate to fluctuate, they did have an impact on the decision of how long the stay-at-home order was in effect for. For instance, Hawaii originally had a stay-at-home order mandated from March 25, 2020 until April 30, 2020. However, Governor David Ige extended this order until May 31, 2020, which was caused by the severeness of the virus and the positive cases doubling from March to April. Furthermore, California began their stay-at-home order on March 19, 2020 and extended the order until May 31, 2020. As for California’s COVID-19 cases, they had an increase of 34,006 positive cases from March until the end of April and an additional 61,666 positive cases in April. California began soft openings for different counties within the state in April, which explains the dramatic increase of cases throughout the month. Additionally, the New York Governor mandated their stay-at-home order to begin March 22, 2020 and was set to expire on May 15, 2020. Governor Andrew M. Cuomo extended this order until June 13, 2020, due to no sign of the pandemic easing. From March to June, New York had 393,448 positive cases. Lastly, Texas Governor Greg Abbott, announced the stay-at-home order to begin on April 2, 2020 and be lifted on April 30, 2020. Unlike the other states, Governor Abbott did not extend the stay-at-home order, but many county leaders decided to extend the order in their county. Texas COVID-19 cases increased by 21,556 from March to April. The graphs below illustrate how each state’s new cases varied on a month-to-month basis.

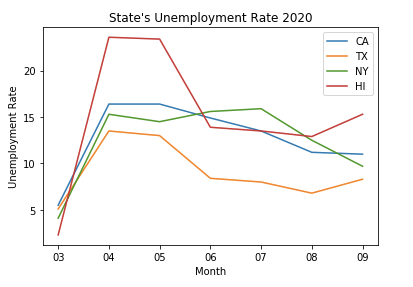








From our data, and the graph below it is shown that the unemployment rate of each state maintained a high unemployment rate during the durations of the stay-at-home orders.



The graph illustrates that COVID-19 had an initial impact on the unemployment rate, but a month after the stay-at-home order was lifted in each state, the unemployment rate slowly began decreasing for each state listed.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Unemployment Rate for Each State (%)** | | | | | |
| **Year** | **Month** | **CA** | **HI** | **NY** | **TX** |
| 2020 | 3 | 5.5 | 2.3 | 4.1 | 5.1 |
| 2020 | 4 | 16.4 | 23.6 | 15.3 | 13.5 |
| 2020 | 5 | 16.4 | 23.4 | 14.5 | 13.0 |
| 2020 | 6 | 14.9 | 13.9 | 15.6 | 8.4 |
| 2020 | 7 | 13.5 | 13.5 | 15.9 | 8.0 |
| 2020 | 8 | 11.2 | 12.9 | 12.5 | 6.8 |
| 2020 | 9 | 11.0 | 15.3 | 9.7 | 8.3 |

The table above is illustrating the unemployment rate for each state we had analyzed beginning when COVID-19 was first reported in each state. For our data, the first report of a positive case was in March 2020, which is why the unemployment rate for each state was at its lowest during the pandemic.

Additionally, travel restrictions seem to also be correlated to the unemployment rate increase and decrease trend differing from state to state. For starters, Hawaii is a tourist attraction and due to COVID-19 Hawaii has had an international and national travel ban, which deemed to be the cause of Hawaii having the largest unemployment rate increase between the chosen states during the pandemic. This travel ban also allowed Hawaii to keep COVID-19 cases relatively low compared to the heavily populated states. On October 15, 2020, Hawaii reopened to allow U.S travelers to enter the state with proof of a negative COVID-19 test result. This decision, “will help bolster the state’s labor force and improve the economy as more residents go back to work and more hotels, restaurants and other businesses reopen,” according to Eugene Tian, chief economist for the state Department of Business, Economic Development and Tourism. As for California and Texas, they have not strictly mandated travel requirements, other than encouraging travelers to self-quarantine for 14 days. This is a logical explanation for their unemployment rate to continuously decrease. New York on the other hand, began a travel restriction on June 24, 2020 that is still ongoing. This restriction includes a 14-day quarantine period if you were traveling from a state with a high infection rate. This ties to the New York unemployment rate increasing again in June and July. With additional data and time, we could have potentially analyzed the travel restrictions and stay-at-home-orders in our original data. These could have been additional variables that would show a more correlated effect that COVID-19 had on the unemployment rate.

Our data could be applied to business application by showing the effect that the increase of the unemployment rate directly has with GDP, since only essential workers could work during the duration of the stay-at-home orders, unless individuals could work from home. During the pandemic many small businesses were forced to closed along with larger businesses forcing to go bankrupt, which also affected the economy.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Variable Name** | **Type** | **Function** |
| 1 | us\_states\_covid19\_daily | .csv | Original csv where the Positiveincrease originated from. |
| 2 | CAUR | .csv | Original unemployment rate csv from California |
| 3 | HIURN | .csv | Original unemployment rate csv from Hawaii |
| 4 | NYUR | .csv | Original unemployment rate csv from New York |
| 5 | TXUR | .csv | Original unemployment rate csv from Texas |
| 6 | df\_merged | DataFrame | This is our main dataframe with everything combined including all 4 state's positive Increase and unemployment rate. |
| 7 | df\_merged\_ca | DataFrame | This dataframe is the same as df\_merged but just for the state of California. |
| 8 | df\_merged\_hi | DataFrame | This dataframe is the same as df\_merged but just for the state of Hawaii. |
| 9 | df\_merged\_ny | DataFrame | This dataframe is the same as df\_merged but just for the state of New York. |
| 10 | df\_merged\_tx | DataFrame | This dataframe is the same as df\_merged but just for the state of Texas. |
| 11 | df\_daily\_covid | DataFrame | df\_daily\_covid is a variable name for the .csv file "us\_states\_covid\_19" which included all 50 states Covid data which was filtered down to df\_merged. |
| 12 | positiveIncrease | DataFrame Column | positiveIncrease is a column from "us\_states\_covid19\_daily" which is merged into df\_merged. This has the positive new cases for each state by day. |
| 13 | df\_daily\_covid\_ag1 | DataFrame | This dataframe is a group by State, Month and Year. It also includes all 50 states. |
| 14 | California Graphs | Graphs | All California data is represented via *blue*. |
| 15 | New York Graphs | Graphs | All California data is represented via *green*. |
| 16 | Hawaii Graphs | Graphs | All California data is represented via the color *red*. |
| 17 | Texas Graph | Graphs | All California data is represented via *orange*. |

This table give clear instruction about all variables defined within the code. States the type of the variable and its end function for the code.

Data Cleaning

The quality of the dataset overall was good. The timeframe ranged from January 2020 to September 2020 which provided an exceptionally large amount of data that we were then able to use in finding the correlation between positive COVID-19 tests and unemployment rates. Since the amount of data was immense, we decided to restrict our analysis to major states except for Hawaii as an outlier in our references. Now when viewing the data, our data on the unemployment rates stops at month 09 (September). The reason for this is due to our data frame df\_daily\_covid also ending at that same month and it wouldn’t make sense to our analysis to continue to view unemployment past that month as our focus is to find the correlation between those rates and positive COVID-19 cases.

Most of the missing variables in our data set came from the Daily U.S. Covid-19 dataset. After pulling the Daily U.S. Covid-19 data, we dropped all unnecessary columns. We only kept the date, state, and the positiveIncrease column. The positiveIncrease column shows the number of positive cases by day. As we are focusing on the correlation of unemployment rates to the positive COVID-19 cases, we decided to omit the “extra” data that doesn’t fully illustrate our analysis. To further elaborate – there were variables that dealt with the number of deaths on a specific day, how many people were hospitalized in each state, if there were any tests pending, etc. and since those variables (for example) don’t have much of an effect on unemployment, it was in the best interest of our perusal to remove all unnecessary variables (also for the readability of those reading the code).

There were multiple rows of missing data under the different columns in the Daily U.S. COVID-19 dataset. Now, this data was a mix between being missing at random or missing not at random. None of the data missing was missing completely at random as all the data collected was intentional. Through the data that was either missing at random or missing not at random, we viewed the data that was critical to the thorough analysis of the correlation of unemployment to positive cases. However, once all the unnecessary columns were dropped, the column we did use did not contain any missing data.

On the positive column, there are missing data that are shown from in the beginning of the 2020 year, mostly during January to March. However, in the positiveIncrease column, any date that had missing data in the positive column was listed as 0. We assume this is due to there being no increase in cases as there is no data in those columns. This is probably because the data from that time period is missing due to no positive cases as the virus had not reached all the states in the United States. These variables are not related to anything specific. Testing was not as widely available, and people were not getting tested. Therefore, the states were not able to report any cases, resulting in no data for the dates of those months.

We did not use any method to clean this data due to there being no missing values. However, there were some negative values in our dataset. In the state of Hawaii for two dates in the month of May, there was a variable of –1 and –5 on the positiveIncrease column. We concluded that this is either due to Hawaii misreporting the data or there were possible false positive tests during those days.

Data Merging

Our process of merging the data involved finding the common element within the Daily U.S. COVID-19 dataset and each of the unemployment rate datasets (California, New York, Texas, Hawaii). The common element was the state name abbreviation (ex: CA, NY, TX, HI) and the month and year. Although the Daily U.S. COVID-19 dataset had data that was reported daily, we combined the daily positive increase case numbers together to accurately show the positive increase case numbers by each month. From there, we were able to match the month and year from the Daily U.S. COVID-19 dataset to the months on each of the unemployment rate datasets by their respective states. The elements that were not common were the positive increase cases and the unemployment rate, which were the two elements we put into the same table to compare the two datasets.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **state** | **year** | **month** | **positiveIncrease** | **CAUR** |
| **0** | CA | 2020 | 3 | 7429 | 5.5 |
| **1** | CA | 2020 | 4 | 41435 | 16.4 |
| **2** | CA | 2020 | 5 | 61666 | 16.4 |
| **3** | CA | 2020 | 6 | 118149 | 14.9 |
| **4** | CA | 2020 | 7 | 264856 | 13.5 |
| **5** | CA | 2020 | 8 | 210497 | 11.2 |
| **6** | CA | 2020 | 9 | 98223 | 11.0 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **state** | **year** | **month** | **positiveIncrease** | **HIURN** |
| **7** | HI | 2020 | 3 | 203 | 2.3 |
| **8** | HI | 2020 | 4 | 409 | 23.6 |
| **9** | HI | 2020 | 5 | 38 | 23.4 |
| **10** | HI | 2020 | 6 | 249 | 13.9 |
| **11** | HI | 2020 | 7 | 1089 | 13.5 |
| **12** | HI | 2020 | 8 | 6350 | 12.9 |
| **13** | HI | 2020 | 9 | 3679 | 15.3 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **state** | **year** | **month** | **positiveIncrease** | **NYUR** |
| **14** | NY | 2020 | 3 | 75789 | 4.1 |
| **15** | NY | 2020 | 4 | 228577 | 15.3 |
| **16** | NY | 2020 | 5 | 66398 | 14.5 |
| **17** | NY | 2020 | 6 | 22684 | 15.6 |
| **18** | NY | 2020 | 7 | 21560 | 15.9 |
| **19** | NY | 2020 | 8 | 19742 | 12.5 |
| **20** | NY | 2020 | 9 | 20870 | 9.7 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **state** | **year** | **month** | **positiveIncrease** | **TXUR** |
| **21** | TX | 2020 | 3 | 3265 | 5.1 |
| **22** | TX | 2020 | 4 | 24821 | 13.5 |
| **23** | TX | 2020 | 5 | 36200 | 13.0 |
| **24** | TX | 2020 | 6 | 95699 | 8.4 |
| **25** | TX | 2020 | 7 | 260960 | 8.0 |
| **26** | TX | 2020 | 8 | 192023 | 6.8 |
| **27** | TX | 2020 | 9 | 122163 | 8.3 |

The tables above are the combined tables we created to analyze the correlation between each state’s positive cases and unemployment rate by month starting from March 2020 to September 2020.

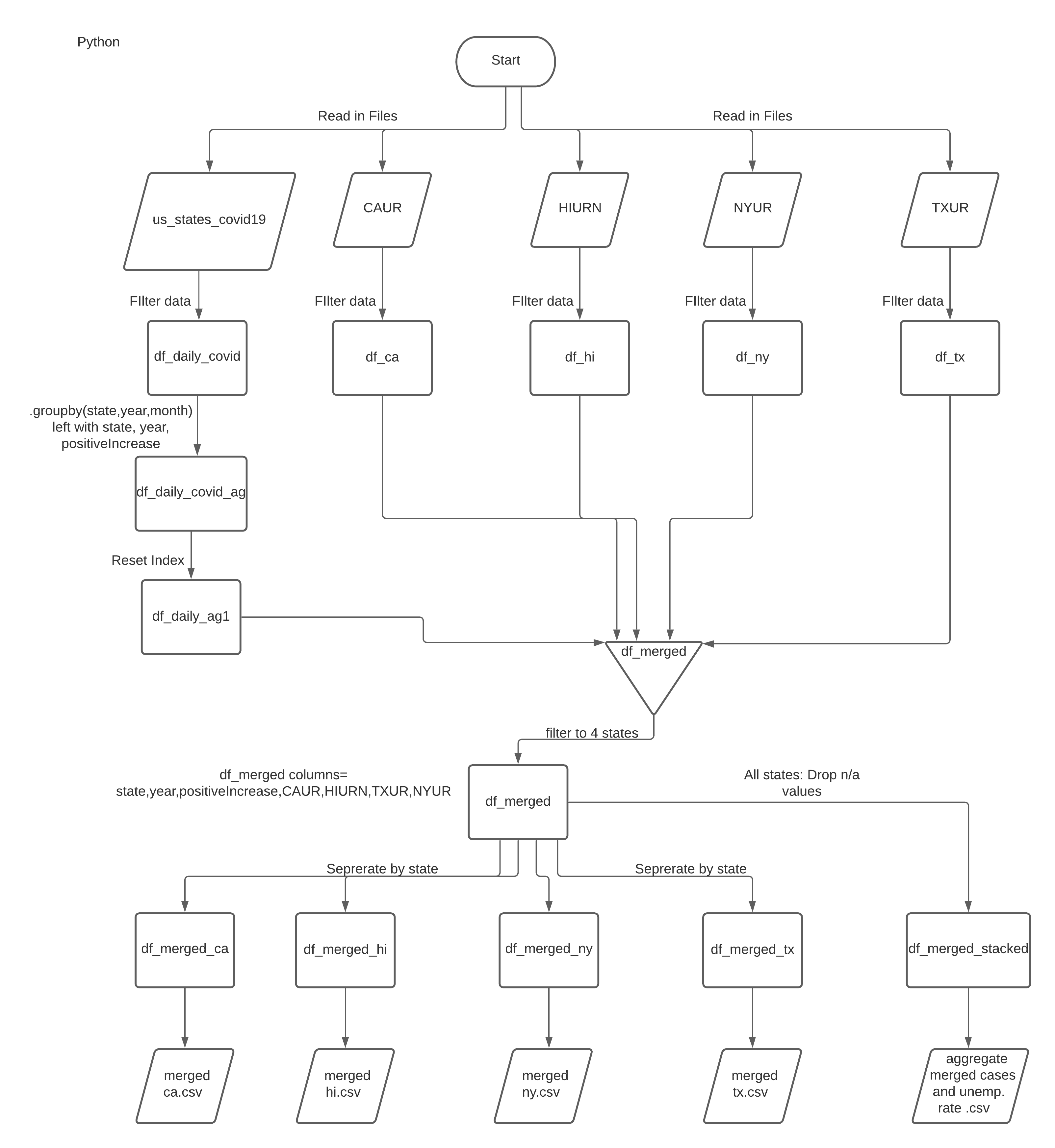
There were no issues with multilevel measurements in our final dataset. As stated before, the original Daily U.S. COVID-19 dataset positive case numbers were by day, so we compiled it to change it into months because the unemployment rate datasets were by month. We did this to make all the datasets be at the same granularity. By doing this, we eliminated any multilevel measurement issues in our final dataset.

When the separate datasets are combined, the variables that become valuable are the state’s positive increase case numbers and the unemployment rate. Individually, the positiveIncrease column would only show the number of positive increase of cases by month and state. Similarly, the unemployment rate data shows how much unemployment rate has changed by month and state. Both of their data show specific information, which can help answer some questions. Together, the datasets become even more valuable as they provide more information and can help answer even more questions.

With our data combined, there are multiple ways that the data has become more valuable since being merged. To begin, we can see the number of positive COVID-19 cases by month and how it relates to the unemployment rates. We can analyze what is the effect of COVID-19 and unemployment rates and answer more questions with this combined data than with the individual data. Additionally, we can also create graphs and tables with the combined data to see trends and give us answers into our questions when we analyze this data.

In this data, we can create aggregate-measurements with the positiveIncrease column where it shows the positive increase of cases by month. If we add the positive increase of cases per month to the month prior, we could create a running total of COVID-19 cases by the state. With this information, we could compare it to the unemployment rate to show us if the number of total running total of the positive cases would change the analysis that we had for the monthly positive increase cases.

Visualizations



This flow diagram is being used to illustrate all steps necessary to gather the data to visualizing the data. When referencing the language used for the entirety of the code, “Python” is labeled at the top to state the language is consistent throughout.

Instructions for code

To have our code run successfully, the user will have to:

1. First unzip the files in an appropriate place.
2. The user must then open spyder or any corresponding application that will allow you to use and run py files.
3. Once the application is open and the user opens our code file, the user will have to manipulate the path to correspond where the user originally placed the unzipped file that contains the csv files for the unemployment rate of each state and the Daily Covid count. This will allow python to be able to read and pull the data from those csv files.
4. Next the output path, which is labeled ‘path\_2’ will also have to be manipulated to be exported in the correct designated output folder within the folder where our other data is located.
5. After those path variables are manipulated the user will be able to run the code with ease.

After running the code, five files will be exported in the output folder, which contain our aggregate data, and the data for each individual state.

Citations

CBS News. (2020, May 15). Lockdown extended for most of coronavirus-battered New York. Retrieved December 02, 2020, from <https://www.cbsnews.com/news/new-york-stay-at-home-extended-coronavirus-lockdown/>

Dave Segal [dsegal@staradvertiser.com](mailto:dsegal@staradvertiser.com) Oct. 20, 2. (2020, October 21). Hawaii's unemployment rate hit nation-high 15% in September. Retrieved December 02, 2020, from <https://www.staradvertiser.com/2020/10/20/breaking-news/hawaiis-unemployment-rate-hit-nation-high-15-in-september/>

Governor's Office – News Release – Gov. Ige approves Mayor Caldwell's Stay at Home Order. (n.d.). Retrieved December 02, 2020, from <https://governor.hawaii.gov/newsroom/latest-news/governors-office-news-release-gov-ige-approves-mayor-caldwells-stay-at-home-order/>

Livengood, P. (2020, April 27). Texas 'stay home' order will expire April 30, Gov. Abbott says. Retrieved December 02, 2020, from <https://www.kvue.com/article/news/health/coronavirus/coronavirus-texas-covid-19-stay-home-order-to-expire/269-1d6de0a9-a9ad-48ae-b43b-6045d08824bb>

McKibbin, W. (2020, March 27). What are the possible economic effects of COVID-19 on the world economy? Warwick McKibbin's scenarios. Retrieved December 02, 2020, from <https://www.brookings.edu/blog/up-front/2020/03/06/what-are-the-possible-economic-effects-of-covid-19-on-the-world-economy-warwick-mckibbins-scenarios/>

Statewide Stay-at-Home Order: Effective March 25, 2020 through April 30, 2020. (2020, September 22). Retrieved December 02, 2020, from <https://hawaiicovid19.com/statewide-stay-at-home-order-effective-march-25-2020-through-april-30-2020/>

Svitek, P. (2020, March 31). Gov. Greg Abbott tells Texans to stay home except for essential activity in April. Retrieved December 02, 2020, from <https://www.texastribune.org/2020/03/31/greg-abbott-texas-executive-order-closures/>

Travel restrictions issued by states in response to the coronavirus (COVID-19) pandemic, 2020. (n.d.). Retrieved December 02, 2020, from <https://ballotpedia.org/Travel_restrictions_issued_by_states_in_response_to_the_coronavirus_(COVID-19)_pandemic,_2020>