**ETL PROJECT**

TEAM “COVIDATA”

GROUP MEMBERS

Evans

Mohammed

Stevene

With the current situation we live in, our emotions and interest drove us to perform our ETL project on the current pandemic COVID-19 and UNEMPLOYMENT. Although there is lots of information on the national level our interest was centered on knowing how the pandemic has a correlation with unemployment across all states in the U.S.

Our original data sources for our csv files came from kaggle and BLS (Bureau of Labor Statistics). The data from kaggle included updated information on current infection/confirmed cases across the globe and the U.S states. The second source is one that is constantly being updated by the BLS (monthly) and includes information on monthly unemployment rate for the past two months (January – February). These files were used due to the connection that would be derived by joining on the state and finding the unemployment rate to confirmed cases with the intention to analyzed which states employment rates are being affected by the number of confirmed cases.

usa\_county\_wise.csv:

* + From the start, we realized that this file included the data necessary: Province\_State, Country\_Region, Date and Confirmed
  + With the dates formatted by mm/dd/yyyy in rows

Unemployment.csv:

* + No changes were necessary on this file as the columns were formatted properly and included all the information needed: State, Jan. and Feb., values for months being the unemployment rate.

The following steps were used to conduct our study:

* + We used Jupyter Notebook and imported pandas to facilitate loading our csv files
  + Pandas continued to be used to help us read our csv files and print out results using .head()
  + Not all the information in usa\_county\_wise.csv for confirmed cases was needed therefor we used the following function to select columns:

state\_infection =data\_df[["Province\_State","Country\_Region","Date","Confirmed"]]

* + Since we planned to work on specific periods of time, our next opportunity was in filtering County\_Region and Month (1= January 2= February) to select only the US region and

infectionByMonth= state\_infection.loc[(state\_infection["month"] == "1")|(state\_infection["month"] == "2"), :]

infectionByMonth=state\_infection.loc[(state\_infection["Country\_Region"] == "US"), :]

infectionByMonth.head()

infectionByState= infectionByMonth.groupby(["month","Province\_State"]).sum()#.dt.strftime("%m"))

infectionByState.reset\_index().head()

Our final steps were to write the data into a database for storage and export to postgres. We initiated a connection and created an engine to postgres. Followed by creating tables and concluding with using pandas to load csv converted dataframe into database.

rds\_connection\_string= ‘postgres:postgres@localhost:5432/etl\_pro\_db’

engine= create\_engine(F’postgresql://{rds\_connection\_string}’)

engine.table\_names()

infectionByMonth.to\_sql(name= ‘Infection, con=engine, if\_exist=’append’, index= False)

unemployment\_df.to\_sql(name= ‘unemployment’, con=engine, if\_exist= ‘append’, index=False)

Our approach for this analysis was to draw attention on the number of cases and rate in unemployment. Our thoughts were to find the trend between the increase in confirmed cases in US states and how each state’s unemployment rate changes. However, our functions created difficulty creating easily accessible dataframe which we could access to perform future analysis. In conclusion, our data that was transformed into a postgres database for future joins for those who are interested in analyzing how covid-19 has created drastic changes in unemployment rates in the U.S.