ETL PROJECT

JUSTIN POT, NATALIE AVILES, NIKOLE YEUNG, PHOEBE WANG

COVID DEATHS IN RELATION TO INCOME LEVELS ACROSS COUNTIES IN CALIFORNIA

EXTRACT

We used data.world to find our first CSV file "John Hopkins COVID-19 Case Tracker" and The United State Census Bureau website to find our second CSV file "Median Household Income by County". We imported our CSV files into Jupyter Notebook to format the data and then loaded the files into our database that we created on PgAdmin.

TRANSFORM

To begin, we imported the COVID-19 cases CSV into Jupiter Notebook. We filtered the dataframe to the specific columns we needed and titled it "new_df". We then filtered the state column to California and the last accumulates report date. We then identifies NA values in our data frame and dropped them because they were outside of California or unassigned values. We finished cleaning the data by renaming the columns and setting any float field values to integers.

Next, imported the Household Income CSV into Jupyter Notebook. We filtered the dataframe to the specific columns we needed and titled it "income_df", we reset the index and named it "id". We then performed data cleaning where we removed the word "county" from the "County" column values so that we could set it as a Key value.We also removed dollar signs and commas from the income column values so that we could import it into our database as an integer.

LOAD

We created a database in PgAdmin named "COVID19_db" and created two tables to store our data.

We created a PostGres db connection to load our panda transformed data into PgAdmin. We chose postgres because it is a relational database and our data is structured data.Once the connection was established, we confirmed the connection by pulling the table names in Jupyter Notebook. We wrote a final query in Jupyter Notebook to read the tables we created.

We created two seperate tables in PgAdmin because COVID cases and deaths are changing everyday so the dataset can be updated daily but the income information would remain the same throughout the year.