ETL Project Proposal:

Our proposal is to take two datasets from Kaggle, “How much do data scientists earn in 2017-2020,” and “HR Analytics: Job Change of Data Scientists” and extract the data via Jupyter Notebook and set up our tables with pgAdmin/postgres. Luis has been tasked with creating the tables in Postgres, using the csv file headers and see which columns he finds most relevant.

Extraction:
The dataset was pulled from Kaggle, as mentioned before, and we will be using Jupyter Notebook is extract those dataset .csv files and put them into pandas DataFrames.

Transform:
Simultaneously, Jimmy and Sarah will begin the process of taking the .csv files, which have already been uploaded into our group Github, and transform the data, both by getting rid of null values and dropping columns that seem less relevant to finding the commonalities of both datasets. If time allots, the three of us can also work on analyzing the data and seeing what conclusions we can draw from the data.

Load:
Lastly, we will all participate in loading the datasets from Pandas DataFrames into Postgres, making sure the data has loaded correctly, and joining the tables created into a new nice, clean table/dataset.