**ETL project: Mark Clayton and Jenny Bui-**

For our project, we decided to tackle US Income rates and US Presidential political party representation. All of the information was separated based on State.

The extraction process begins with the datasets that we gathered. Our sources were from Kaggle:

<https://www.kaggle.com/jayrav13/us-representation-by-zip-code>

<https://www.kaggle.com/goldenoakresearch/us-household-income-stats-geo-locations?select=kaggle_income.csv>

The datasets included a CSV and JSON file in which we took to Python/Pandas to transform in our next step. First step for the data cleaning process was to load the json file into the Jupyter Notebook. The transformation that was required for this included trimming the file into a smaller and more manageable file compared to the original set of over 500 thousand rows. We also had to figure out a way to load the large JSON file through using an additional add-on installation on Jupyter Notebook. Nonetheless, we attempted to un-nest the JSON data to only have the president name, political party and zip code. Next, we created a new notebook and cleaned up the CSV file. Several rows were dropped and the ones remaining included State, Zip Code, Mean, Median, and St. Deviation. The only common data between the two data files was the zip code.

The final database includes a merge between the two dataframes based on zip code. The tables were then loaded onto PostgresSQL using an engine connection. This was chosen because we wanted to see if there is a direct relationship between presidential political party and income based on zip code. There were some limitations to our data. For example, we only had one variable that could be tested since there is no direct relationship between income and presidential party. In the end we discovered that most presidents with the zip code 11542 lived in NY and the state had a mean income value of 11680. In summary, even with some limitations, we were able to utilize the ETL process.