**Extract**

We extracted the data from several sources. We used the likelihood of automation data from an existing dataset (CSV format) on dataworld – It was a part of a study that was published in 2013. It contained data on occupation names, codes, and total employment numbers. We obtained data on occupation names, codes, total employment numbers, annual median salary and hourly median wage in the US for the years 2013 and 2018 from the Bureau of Labor Statistics (xlsx) – these files were to large to upload to GitHub, so the links to these datasets are available in the ReadMe.

**Transform**

We used Jupyter notebooks to clean and transform all of the data in this project. The automation likelihood data was accompanied by total employment numbers across states, which we were not utilizing so we removed. In the data from the bureau of labor statistics, we removed all columns except for the occupation names, codes, total employment numbers, annual median salary and hourly median wage. We created columns for the percent difference in total employment numbers in the automation likelihood data by using the columns for the different years in the BLS table. We also found the percent difference in annual median salary and hourly median wage through the same method, after adjusting for inflation using a calculator (also available in the ReadMe)

**Load**

We then loaded all the cleaned data to a PostgreSQL database as two tables, 1 for all the data from the BLS for the years 2013 and 2018, and one for the automation likelihood and the associated percent changes. We were able to successfully pull from the database to create visualizations in the Jupyter notebook.