**Darren Pardo**

**Rutgers Data Science Bootcamp**

**Excel Homework**

**July 25, 2019**

**ETL Project**

**Extraction**

We used two datasets from the public platform Data World. Our data is based on the automation in the workforce and Occupational Employment Statistics. Below are our data set sources.

Occupational Employment Statistics summary from Data World:

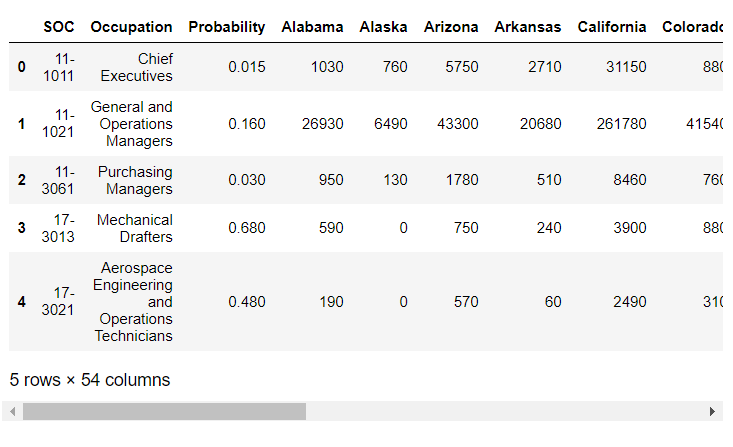
<https://data.world/data-ny-gov/gkgz-nw24/workspace/file?filename=occupational-employment-statistics-1.csv>

Occupations by State and Likelihood of Automation summary:

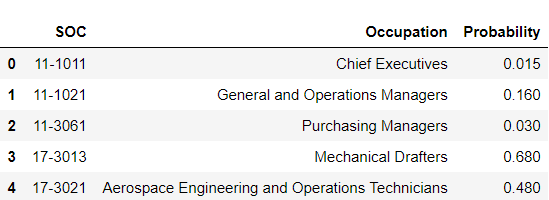
<https://data.world/wnedds/occupations-by-state-and-likelihood-of-automation/workspace/file?filename=raw_state_automation_data.csv>

**Transformation**

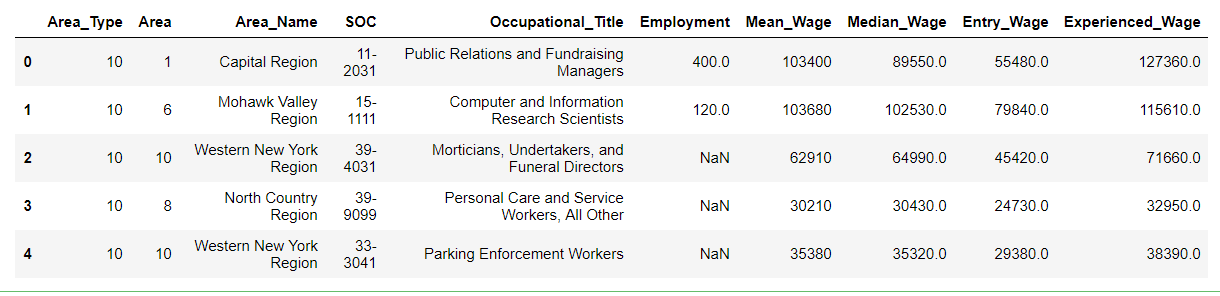
The first steps in cleaning up the datasets involved figuring out which columns were not useful. The amounts of employees for all of the states were removed.



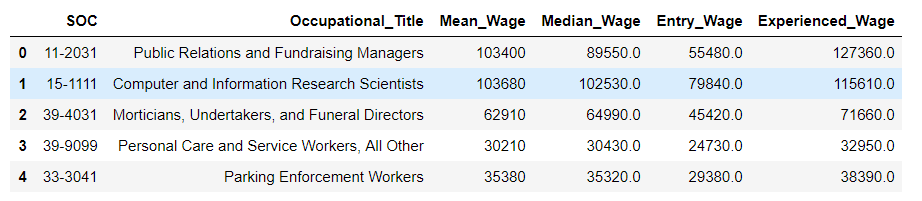
Which brings us to our new data with our three selected columns.



The same process was completed for the Occupational Employment Statistics Data.



New data with our selected columns.



**Load**

The last step was to load dataframes into database. We created a database and tables to match the columns from the Panda’s Data Frame using Postgressql.

CREATE TABLE automation\_probability (

SOC VARCHAR(10) primary key,

Occupation VARCHAR(250),

Probability FLOAT

)

CREATE TABLE occupational\_employment\_statistics (

SOC VARCHAR(10) primary key,

Occupational\_Title VARCHAR(250),

Mean\_Wage MONEY,

Median\_Wage MONEY,

Entry\_Wage MONEY,

Experienced\_Wage MONEY

)

SELECT automation\_probability.SOC, automation\_probability.Probability, automation\_probability.Occupation, occupational\_employment\_statistics.Mean\_Wage, occupational\_employment\_statistics.Median\_Wage, occupational\_employment\_statistics.Entry\_Wage, occupational\_employment\_statistics.Experienced\_Wage

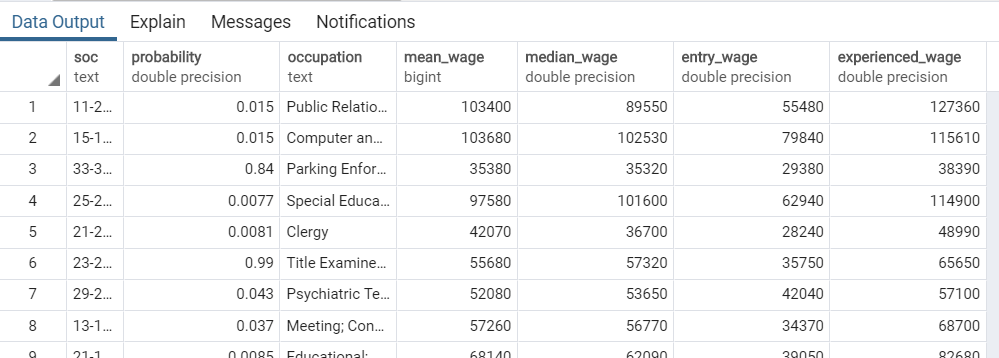
FROM automation\_probability

LEFT JOIN occupational\_employment\_statistics

ON occupational\_employment\_statistics.SOC = automation\_probability.SOC;

SELECT \* FROM automation\_probability

SELECT \* FROM occupational\_employment\_statistics



**Summary**

We used these datasets so we could identify the probability of automation for occupations and the current wages of these occupations. The final output will show us the disparity of wages and the likelihood of the job becoming automated.