Rui Lin

Kena Jackson

8/31/2020

Northwestern Data Science Bootcamp

**ETL Report Analysis**

**Extract:**  your original data sources and how the data was formatted (CSV, pgAdmin 4, etc).

The data sources we used were ‘Unemployment Rates June 2020’, ‘PPP Loans’ and “State Name”. These data sources were formatted as CSV files and then converted into dataframes via Pandas. Later we connect to pgAdmin 4 for further transformations

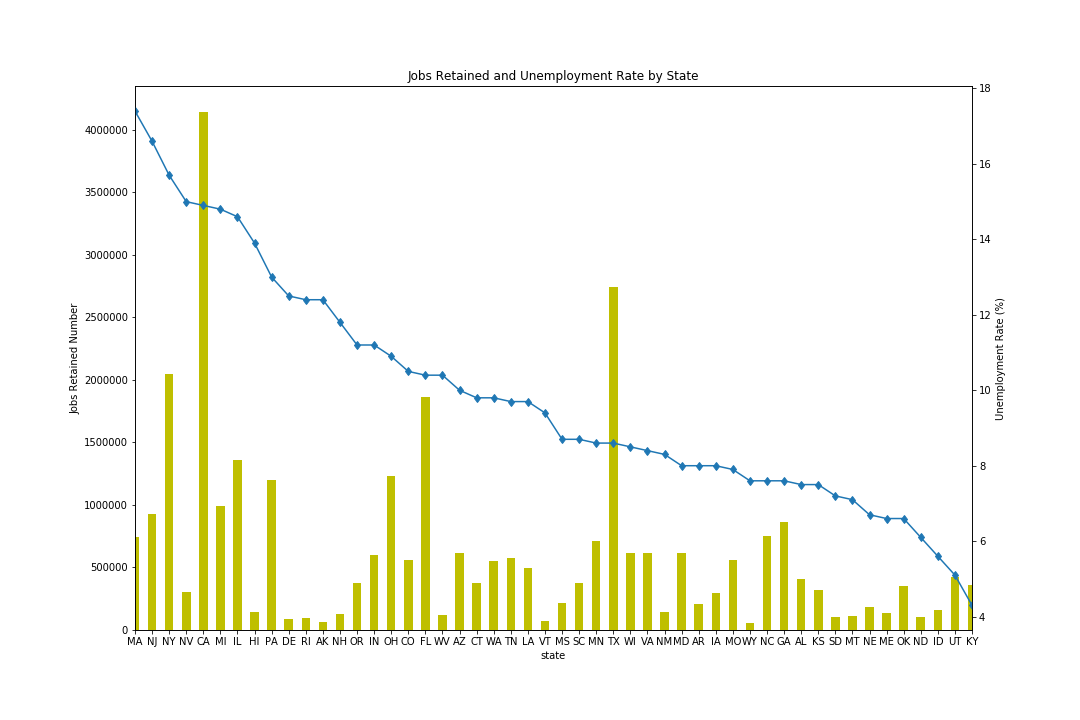
**Transform:**what data cleaning or transformation was required.

The PPP Loans data initially had over 600,000 rows. First we pulled only the columns we intended to upload to our main database. The columns we kept were, “State” and “Jobsretained”. We then proceeded to group the jobsretained column by state. This way we had one lump sum for each state and 50 rows which helped us join our other datasets. We cleaned up our unemployment rates table by removing the extraneous columns and leaving only the rate and the state columns. The state columns of these two datasets(PPP Loans, June Employment) were different. The employment rate’s table displayed full state name while the PPP Loans displayed the state abbreviations. To work through the discrepancy we added another table, ‘state name’, that had states as well as the abbreviations so we can properly join the tables in our database. We had to capitalize the state name in the ‘June Employment Rate’ table so we could have a linear merge. We also renamed the columns so when we ran our SQL query all naming conventions were cohesive.

After we cleaned all of our data and created our dataframes in Pandas, we proceeded to connect the data to our local db. From pgAdmin, we created a database and our tables/ We checked the engine to be sure that the tables have populated. We returned to pgAdmin and created the tables as well as the SQL query for our join. We did run into a few bugs along the way with joining. These were fixed once we adjusted the names of our columns in pandas and transposed those into SQL for more continuity. We joined the state abbreviation with the jobs retained and the unemployment rate. This way only three columns are visible in our final database.

**Load:**the final database, tables/collections, and why this was chosen.

After all of our data was cleaned and transformed, we converted the data into our main database. We did a read sql query to confirm the addition of the tables and all of the above changes. For our final dataframe, we set the state as the index and had an unemployment rate and jobsretained as our remaining columns. We also plotted the data as shown below so we had a visualization of the data we cleaned. We chose to explore this data because of the current pandemic our economy is constantly shifting. We explored the PPP Loan data as we wanted to see how many businesses were applying and approved for loans. And how many jobs were actually saved due to the loans. Is there a visible trend? Did the loans assist with job retention? We also wanted to see how this data compared to the current unemployment rate. If businesses in these states were able to save so many jobs, then surely the unemployment rates will not be as high? We found that why some states did have a large amount of jobs saved, the unemployment rates were still pretty high. Although it is not adequately reflected in our data we saw that some businesses were approved for $5-$10 million in loans and still did not retain any jobs which was interesting to see.



**Other Elements:** Surprises? Unanswered questions?

We did want to divulge deeper into the data analysis but was unable to do so due to the time restraint. It would have been interesting to see the amount of loans that were applied for in each state as an alternative comparison piece. It was also interesting to see that Florida had less jobs retained than New York. The population of New York is approximately 475,000 less than Florida so it was interesting to see this disparity.

**Sources:**

<https://www.bls.gov/news.release/laus.nr0.htm>

<https://www.kaggle.com/susuwatari/ppp-loan-data-paycheck-protection-program>

<https://www.50states.com/abbreviations.htm>