ETL Project

Darya Roshan

**Extraction**

Using 3 different datasets as csv files from Data World and [Kaggle](https://www.kaggle.com/). The datasets are about diversity, unemployment rate and Median income in different counties and states.

**Transformation**

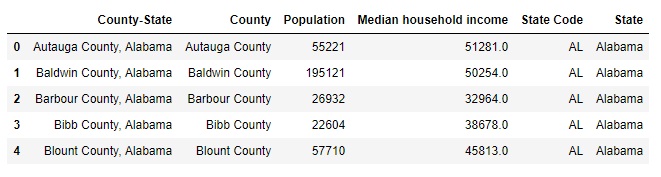
After calling the csv files the initial step was to clean them up by dropping unneeded and irrelevant columns and data along with renaming columns to relevant names.

As example, after reading the diversity csv file (Figure 1), the column Location was split into two columns of County and State for easier access, merge and grouping.



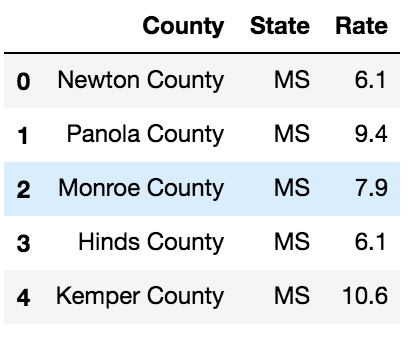
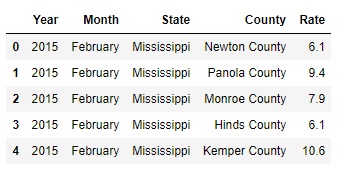
**Figure 1:** Split Location column to County & State, and dropped it

Additionally, in the Median Income Dataset shown table below (Figure 1), columns County-State was dropped, and the State Code column was renamed to State.



**Figure 2:** Dropped unnecessary columns and renamed

In other to be able to have a successful merge, in the unemployment rate table I had to change the full state names (Figure 3) to it’s abbreviation (Figure 4) by using a dictionary of states and their abbreviations were used with a loc loop to iterate through the dataset.

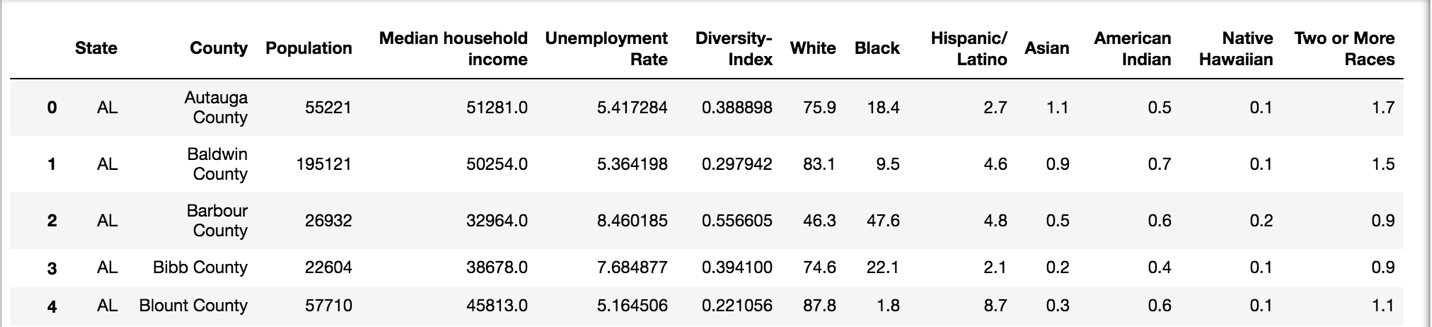


**Figure 3:** State without Abbreviation **Figure 4:** State with Abbreviation

Once that was done, I could group the dataset by State and County.

After the all datasets was cleaned it was time to start merging.

By using inner join, I connected the Unemployment and Median Income datasets on State and County and continued the same with the Diversity dataset. Finally, with the last merge all the three datasets combined into one universal table (Figure 5), the State and County index was reset turning them back into columns, and columns were reordered to a more logical format.



**Figure 5:** Final Output after merging all the datasets

**Load**

Once the table was ready (after the clean ups and merges) I needed to transfer the final output into a DB. For that I created a database in MYSQL and created the needed table. Then connected the final table with the data from panda to the database and table in MYSQL.

**Summary**

These datasets helped me identify the diversity ratio, median income and unemployment rates of each county for each state and recognize the population, diversity, race, median income and unemployment rate of each county and state.

The final output can be used in many aspects of developments nationwide. Overall information about the areas population, income, unemployment rate race and so on can help with development programs and opportunities depending on their need.