ETL Project

By Akshat Jain, Vishal Gaurh and Satyam Sharma

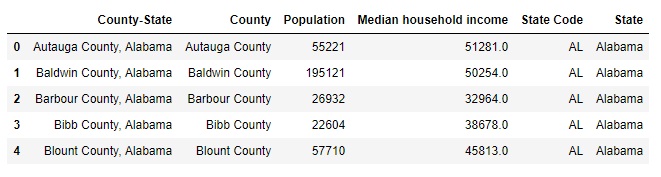
**Extraction**

We used 3 datasets from the public platform Kaggle and Data World. All of our data was based on county through all the States ranging over various years from 2010 to 2015. These were the most recent ones we could find. The sources for our dataset are as follows

* Diversity Index from Kaggle.
* Unemployment from Kaggle.
* Median Income by county from Data World.

**Transformation**

Our first steps in cleaning up the datasets involved figuring out which variables were not relevant. For the Median Income Dataset (Figure 1), we dropped the columns County-State & State and renamed the State Code column to State.



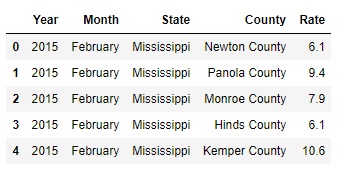
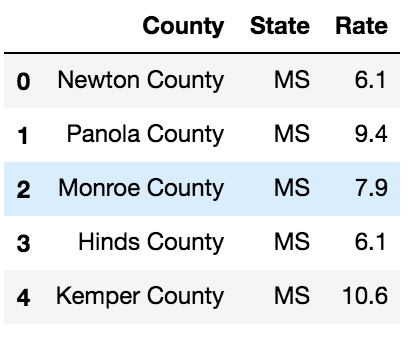
**Figure 1:** Dropped unnecessary columns

We also split certain columns such as Location from Diversity table into County & State to make it easier to merge and group with the other 2 datasets (Figure 2).



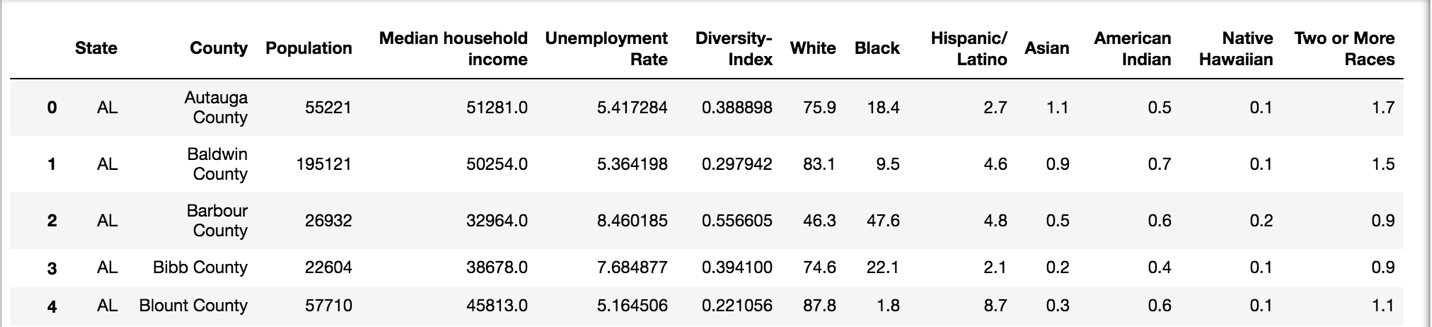
**Figure 2:** Split Location column to County & State, and dropped it

For the unemployment rate dataset, State column was in full name (Figure 3), we needed to abbreviate them to merge with the rest. To do this, a dictionary of states and their abbreviations were used with a loc loop to iterate through the dataset (Figure 4). Finally, we grouped this dataset by State and County and averaged for each County.

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

After cleaning the datasets and workable, we started merging. To start, the Unemployment dataset and Median Income dataset were merged on State and County, using an inner join. The Diversity dataset was then merged on that by State and County again, using an inner join as well. With all the three datasets combined (Figure 5) into one universal table, the State and County index was reset turning them back into columns, and columns were reordered to a more logical format.

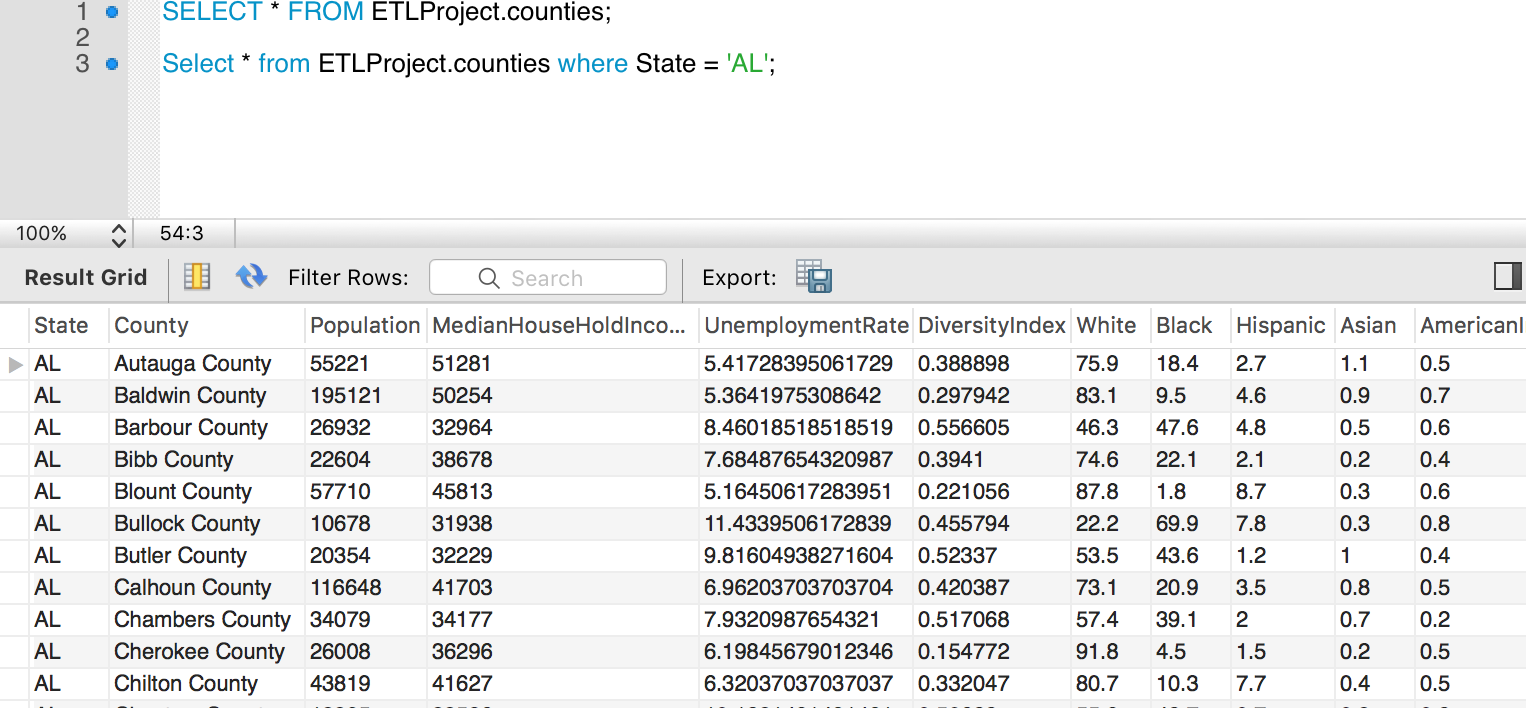


**Figure 5:** Final Output after merging all the datasets

**Load**

The last step was to transfer our final output into a DataBase. We created a database and respective tables to match the columns from the final Panda’s Data Frame using MYSQL and then connected to the database using SQLAlchemy and loaded the result. Here we were able to perform multiple queries to suit a desired criterion.

Sample Query (Figure 6) using MySQL that returns all the variables for the State of Alabama (AL).



**Figure 6:** Sample Query

**Summary**

We used these datasets so we could identify the diversity ratio, median income and unemployment rates per county for each state. The final output will help us to recognize which county, state that has the following.

* Most or Least Population.
* Median Household Income
* Unemployment Rate
* Diversity Index
* Race Index

These indices can be used to identify if any development aid/work are required for areas with high unemployment rate, low median income or with high population.

If the population is high and if the schools/colleges/educational institutions are less, these indices can be used to build more educational institutions, which in turn create more employment opportunities. It is a cascading effect of more the population, more educational institutions, more employment, more household income.