ETL Project – Minimum Wage Analysis

Ryan Conway, Joe Fioro , Hammad Ali

**Extraction**

We used 2 datasets from the public platform Kaggle. All of our data was based on county through all the States ranging over various years from 1990 to 2016. These were the most recent ones we could find. The sources for our dataset are as follows

* National Minimum wage data from Kaggle.
* National Unemployment data from Kaggle.

**Transformation**

Before we could begin our process of transformation of the datasets, we had to remove certain elements that prevented Panadas from properly reading the csv files. Extra commas and parentheses caused errors with the utf-8 encoding. By using a “with/open” statement we were able to successfully remove the erroneous characters.

Our next step in cleaning up the datasets, involved dropping unneeded columns in each of the csv files. Simultaneously, we made sure the columns had a lowercase title to prevent an error in PGadmin.

The next transformative step was to drop data in on data set that was out of range (Years) of the other.

The final step in transformation involved “rolling up” the data to match the granularity of the highest level data set. In order to achieve this we did a panda’s groupby statement before importing the data frames to PGadmin. We aggregated the mean of the desired data by the State and Year columns. This cleanly rolled up the “month and county data into the same granularity of the other data set.

**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.

**Summary**

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

* Minimum and Maximum Unemployment Rate
* Highest and Lowest Minimum wage

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.