ETL Project

April 17, 2001

By Amol Bhatia, Jennifer Duffy, Robin Evans

**Objective**

Our objective was to Extract, Transform, and Load data related to the United States stock market and United States presidential administrations so that analysts can examine how the market has reacted to executive orders, cabinet appointments, and ???? during three different administrations. Our time period included: the first and second terms of Barack Obama (January 20, 2009 – January 20, 2017), the only term of Donald Trump (January 20, 2017 – January 20, 2021), and the beginning of the first term of Joe Biden (January 20 , 2021 – April 17, 2021).

**Extraction**

We used three different datasets from the platforms listed below. These were the most recent ones we could find. The sources for our data are as follows:

* The Yahoo Finance API
* Executive Orders from the Federal Register of the National Archives (<https://www.federalregister.gov/presidential-documents/executive-orders>)
* U.S. Cabinet Appointments from data.world (<https://data.world/government/us-cabinet-appointments>)

Prior to choosing these three datasets, we looked at a number of other sources, including 34 different “president” datasets on data.world; however, we rejected them because the information was not recent enough or the datasets were linked to elections, which is not in our scope. The Yahoo Finance API is the most up-to-date as it provides live market data with less than 0.1-second lag.

**Transformation**

Our first steps in cleaning up the datasets involved figuring out which variables were not relevant. For the ??? Dataset (Figure 1), we dropped the ??? and renamed the ?? column to ??.

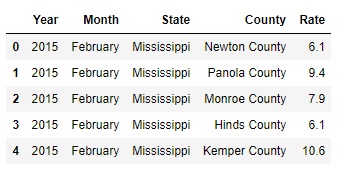
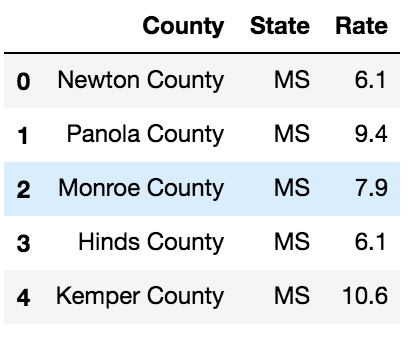
SHOW SCREENSHOT OR DF EXAMPLE BELOW??

**Figure 1:** Dropped unnecessary columns

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

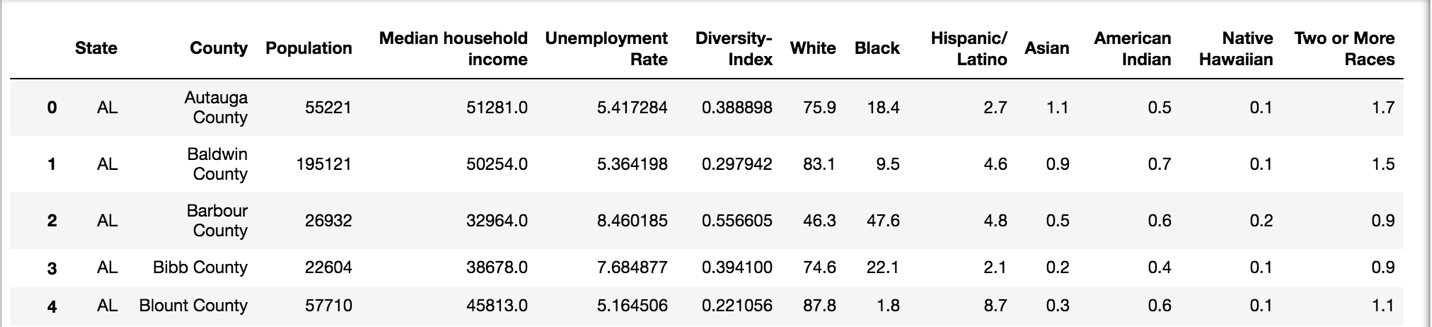
**Figure 2:** edit this caption

Give a description here of other transforming steps………

**Figure 3:** replace table and edit caption **Figure 4:** replace table and edit caption

After cleaning the datasets and workable, we started merging. To start, the ??? and ?? datasets were merged on ??, using an inner join. The ?? dataset was then merged on that by ???, using an inner join as well. With all the three datasets combined (Figure 5) into one universal table, the ??? 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**

Edit this text……..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 ???

ADD SAMPE QUERY HERE

**Figure 6:** Sample Query

**Summary**

Edit this section………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.