ETL Project – Constructing a Database for Political Science Analysis Keyed by State

James Bryant II, Amee Yang, Greg Spahlinger

Introduction

In recent years data analytics related journalism has been growing in importance. The trend was started by Nate Silver and fivethirtyeight.com, who made a splash initially by using aggregate polling data to predict the outcome of the 2008 presidential election. Since then, they have branched into politics, sports and culture related journalism, with a mindset informed by relevant data. Because fivethirtyeight.com publishes its datasets, they are providing a resource that could empower others, who admire their approach to follow in their footsteps. In looking for data to archive for our ETL project, we decided that we would archive some of the data provided there.

While a large number of datasets are available at fivethirtyeight.com, they often look at very different phenomenon, so our first challenge was to find sets that share a reference for relational database formation. We decided to find sets that are organized by US state, so that we could enable research into the differing character of various US states. We initially found two suitable datasets on the site – one on urbanization in each state, and one on political elasticity – a measure developed to describe how apt a state is to follow national electoral trends towards one major political party or the other. We decided to bring this together with racial demographics by state, which came from data world. This makes good sense, as living arrangements (urbanization) and political preferences (elasticity) describe the people living in a particular state, and racial demographics can provide insight into the level of diversity in those communities. In principle our database can be expanded to include other such factors since US states are a convenient key for organizing data.

Methods

We obtained our data in the form of .csv files and one .xls file (from data world). The .xls file was then converted to a .csv using Microsoft excel. In looking at the states included, we realized quickly that we had several problems. The first problem was that our elasticity file identified us states by two letter abbreviation rather than full state name. As we could not think of a slick programmatic way to change one for the other using python, we edited the key manually in excel, so that the full name of the state was present. The other manual cleaning we did involved deleting unstructured notes out of the bottom of the demographics file.

The next problem we encountered was that each file had a different idea of what should be listed as a state – our demographics file listed US totals as well as DC as states, while the urbanization data file included DC and 5 US territories as states. Finally, the elasticity file contained DC as the only non-state US polity. We decided that we should trim entries that did not reference DC or a formal US state. In order to do this, we read all the files into Pandas DataFrames. The demographics .csv was cleaned by dropping the US total row manually, along with a ‘total’ column that appeared to total all of the demographic proportions to 1. In order to get rid of unwanted rows in the urbanization data we merged it with the elasticity data using an inner join. New DataFrames were then copied from the merged DataFrame. After cleaning the three cleaned DataFrames were written out to .csv files.

We modeled the Data using QuickDBD (app.quickdatabasediagrams.com). The model was relatively straightforward, as all tables contain the same primary key (state). We used the model to export SQL commands for table setup, then modified them to remove “NOT NULL” from our demographics table. We did this because our demographics dataset omits some values due to high margins of error, leading to many null values. We then set up our tables in PostgreSQL, and imported our cleaned .csv files.

Our final step with the database was to run test queries to make sure it works correctly. We first queried all of the tables independently using “SELECT \* FROM table” commands. After this we wrote a query where we selected one column from each table, using a double inner join to join all three tables. We initially had problems with this and had to rename one of our keys from “Location” to “state”, but after this the query worked, proving that any relations in the database should be joinable on state.

Conclusion and Future Directions

We have detailed our creation of a new database for political science data by state, including some of the metrics used by fivethirtyeight.com. Because of the structure of the database, in the future we should be able to add additional tables, which include other information about demographics such as socioeconomic data, or consumer preference data. Because of the convenient key system, this database could be very useful to researchers.