10-718 Assignment 1

Varun Rawal (Andrew ID : vrawal)

Data Collection and ETL

Question

Assignment1: ETL

Due: September 18th 6pm eastern

The goal of this assignment to learn how to get data from the web and load it into a database so it can be used in your project. We will use ACS data here from the US Census Bureau since you will probably use it for projects that have any spatial/location component in the US.

Tasks:

1. Get Data: Pick a US state and (write a script to) download the most recent ACS data (at the block group level) for every block group in that state.

There are various download sites and APIs available for this task. Feel free to pick one and justify why. There will be many variables available in the US. You don't have to get them all. Select and justify10-20 that you will find useful for your project (such as age, gender, race, income, education status, etc.). 2. Transform/Prep: Once you've downloaded the data, get it ready (with code) so that it can be loaded into a postgres database table.

3. Load: the data you've downloaded into a postgres database table.

You'll need to create the table schema Copy the data into the table You might find the following resources useful as you do this:

`https://dssg.github.io/hitchhikers-guide/curriculum/1_getting_and_keeping_data/csv-to-db/` (Links to an external site.) `https://github.com/dssg/ohio` (Links to an external site.) What you need to submit:

1. Code to do all three things. Try to keep it modular and generalizable so you can do it for a different geography, granularity (state, city, zipcode, etc.), and add additional variables. Push the code on github in a repository on your personal github account and submit the link on canvas *or* upload the code files on canvas in the assignment submission

2. Location of where the database table is in your project/team database. Submit it as part of the report on canvas

3. A short (1 page) write-up describing what you did and why. Submit it as a pdf on canvas.

Writeup

Following are the steps followed to fetch the data and save it as a csv :

```
python ACS_censusdata.py
```

This would run the script and save the data into a csv.

Then, to import the data from the csv into the PSQL database, we run the SQL script as follows :

```
psql -h mlpolicylab.db.dssg.io -U vrawal -d turnout1_database -f acs_script.sql
```

API for Census Data and why.

I used the following API available in PyPi :

`https://pypi.org/project/CensusData/`

I chose this API because of the convenience it allows. It is designed to provide easy access to the U.S. Census Bureau's API (`https://www.census.gov/developers/`) in Python. It supports pulling data from the American Community Survey (ACS) and the Census Summary Files.

It makes life easy by handling all the details of interacting with the Census API for us, so that we can focus on working with the data. It provides a class for representing Census geographies. It also provides functions for gaining further information about specific variables and tables and for searching for variables.

Selection & Justfifcation of variables

I selected the following 17 variables that I found useful for our project i.e. prediction of Voter turnout :

"Unemployed", "Income", "WhiteRace", "BlackRace", "IndianAlaskanRace", "AsianRace", "Population", "MaleGender", "FemaleGender", "HighSchoolEducation", "CollegeLess1yrEducation", "CollegeMore1YREducation", "AssociateDegreeEducation", "BachelorDegreeEducation", "MasterDegreeEducation", "ProfessionalDegreeEducation", "DoctorateDegreeEducation"

This is because I wanted to select the variables that were highly related to the following factors : **age, gender, race, income, education status** because these are the ones which play an important role and contribute to deciding whether the citizen is aware and responsible enough to vote or not.

Other Submission Details

1. All the code files are pushed to the github repository as instructed and also uploaded on canvas. `https://github.com/rawalvarun/10_718_Assignments`

2. Location of where the database table is in your project/team database : **acs_2018_ass1_schema** $\cdot acs\_2018\_ass1\_table$