ETL Project Technical Report

Team Members: Bernardo Abecia, Ivan Choi, Meng-Yin Lee

Date: 11/24/2020

1. Project Objective Summary:

The objective of this project is to prepare data using ETL program and obtain information from the recently concluded 2020 Gubernatorial race. Our process started from data sets that were sourced from Kaggle in the form of CSV’s, transforming the dataset and upload the data into SQL database for further analysis.

1. Extract/Data Utilized

Our team decided to get CSV files from Kaggle.com in two separate locations.

We picked up CSV’s of governors county; governors county candidates; governors state from (<https://www.kaggle.com/unanimad/us-election-2020>), and for demographics to identify the gender counts who voted for the candidates (<https://www.kaggle.com/etsc9287/2020-general-election-polls>).

There were about 5,000+ lines of data to be extracted, cleaned and load to database.

1. Transformation

Once the data files were extracted, the data was transformed using Python/Jupyter Notebooks. Multiple columns were eliminated to streamline the process. Some were renamed to focus on the state, state code, county, votes, and candidates. Next, we installed “pip install psycopg2” to connect the python data to SQL database.

1. Load

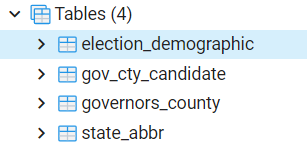
All the data were uploaded into a relational database SQL/Postgres program which composed of 4 tables.

1. Other Steps

The next step in the process will be to further analyze the cleaned dataset using Postgres/SQL ETL Project Database for any information with regards to who won the gubernatorial race per state, number of votes cast, gender info and etc.

***ETL Data Process Flow***

***ETLproject\_db/postgres@PostgreSQL***



Kaggle

governors\_county.csv

governors\_county.jpynb

Kaggle

demo\_county\_statistics.csv

Kaggle

governors\_county\_cancsdidate.csv

<https://www.ssa.gov/international/coc-docs/states.html>

state\_abbreviation.jpynb

election\_demographic.jpynb

Governor\_County\_Candidate.jpynb