***ETL Project Proposal***

**Team BIM: Bernardo, Ivan, Meng-Yin**

**Date:** 11/19/2020

**ETL goal:**

1. To read data from Kaggle ‘US Election 2020’ selected datasets
2. Perform data cleansing, data column selection, data column rename, and load into a database
3. Plan to use join to combine the data into a reportable dataset for analysis

**Proposal:** US Election 2020

**Data Source**: Kaggle

**Data sets:**

1. governors\_county.csv -- b

[https://www.kaggole.com/unanimad/us-election-2020?select=governors\_county.csv](https://www.kaggle.com/unanimad/us-election-2020?select=governors_county.csv)

1. governors\_county\_candidate.csv - i

<https://www.kaggle.com/unanimad/us-election-2020?select=governors_county_candidate.csv>

1. demo\_county\_statistics.csv - m

<https://www.kaggle.com/etsc9287/2020-general-election-polls>

1. 2 letters state abbreviation - m

**url = '**[**https://www.ssa.gov/international/coc-docs/states.html**](https://www.ssa.gov/international/coc-docs/states.html)**'**

**Data Extract/Transform/Load:** Panda(read\_csv)/dataframe/to\_sql

* etl - use panda read\_csv
* column rename to be the same as database table/column
* data cleansing - remove ‘county’ from county column content
* create dataframe with the cleansed data
* create database and tables
* load dataframe to the database

**Target Database**: Plan to use PostgreSQL as our target database