**Group 1 Documentation**

**NAMES:**

* Thomas Westerkamp
* Melissa Mongrella
* Domenic Padulo
* Lauren Stredler

**Instructions**:

Submit a final report that describes the following:

* Extract: your original data sources and how the data was formatted (CSV, etc.)
* Transform: what data cleaning or transformation was required
* Load: the final database, tables/collections and why this was chosen
* Upload to Github and submit

**Extract:**

* Data sources:
  + Our group used two CSV files to compare the U.S. minimum wage to the census population by state in 2019.
  + The two data sources were pulled from Kaggle.
    - 2019 Census Population Data by State: <https://www.kaggle.com/peretzcohen/2019-census-us-population-data-by-state>
    - US Minimum Wage by State from 1968 to 2020: https://www.kaggle.com/lislejoem/us-minimum-wage-by-state-from-1968-to-2017

**Transform:**

* We renamed the:
  + The Minimum Wage CSV file by removing the spaces
  + The Population data CSV file by shortening to make the name easier to import into Pandas
* The CSV files were imported using Pandas
* We created a filtered data frame using specific columns
  + The following columns were dropped because they did not provide additional information that was useful:
    - 2019 Census Population Data by State: Lat Coordinate of Capital, Long Coordinate of Capital
    - US Minimum Wage by State from 1968 to 2020: State.Minimum.Wage.2020.Dollars, Federal.Minimum.Wage, Federal.Minimum.Wage.2020.Dollars, Effective.Minimum.Wage, Effective.Minimum.Wage.2020.Dollars, CPI Average, Department.Of.Labor.Uncleaned.Data
      * In addition, we only kept data from the year 2019
  + The following columns were renamed:
    - 2019 Census Population Data by State: STATE was renamed to state; POPESTIMATE2019 was renamed to population\_estimate
    - US Minimum Wage by State from 1968 to 2020: State.Minimum.Wage was renamed to state\_minimum\_wage and Year variable was changed to \_year in order for Postgres to not automatically assign it a function and to differentiate it as the name of one of our columns.
* We converted our State variables in both data frames from objects to strings before we merged the two dataframes. Once they were merged, the State variable reverted back to an object. We also did this to the Year variable.

**Load**:

* We made the database connection to Postgres and loaded the merged dataframe into the database. Our team was most comfortable with the Postgres platform and chose to use it. We also preferred that the visuals provided in Postgres are similar to that of Jupyter Lab and could therefore make easier comparisons.

**Tables**:

* Matching tables were created in Postgres and joined on the State variable