ETL PROJECT

SHALYN LAVOIE | MORIAH TAYLOR

22 MAY 2021

E

Our first source of data comes from Kaggle (https://www.kaggle.com/muonneutrino/us-census-demographic-data?select=acs2017_county_data.csv). It contains census data from the year 2017. The data source is from the GitHub repository for #TidyTuesday, which is an ongoing R community project (https://github.com/rfordatascience/tidytuesday/blob/master/data/2021/2021-05-11/broadband.csv). This dataset includes information about broadband speeds and usage. The reason we picked these two sources to work with is because they both are organized by county, which made them easy to merge.

T

There were a few cleaning tasks we had to perform using the pandas package. With the **census** data, we first had to select only the columns we were interested in, because there were over 50 columns in the original dataset. We select *County*, *County ID*, *State*, *County*, *TotalPop*, *White*, *Income*, *IncomePerCap*, and *Employed* because we thought if someone were to use this dataset, those would be interesting factors to juxtapose against broadband access. With the **broadband** data, we also had to select which columns we were interested in. We only selected *County ID* and *Broadband Usage*, leaving out the state and county name columns because we already got that information from the census dataset. Next, we merged these datasets on County ID and dropped all NA values. Finally, we put the data into 3rd Normal Form. This meant breaking the merged data into three datasets: (1) **county_names**, containing *County ID* and *County*; (2) **states**, containing *County ID* and *State*; and (3) **counties**, containing *County ID, TotalPop, White, Income, IncomePerCap,* and *Employed*. *County ID* acts as the primary key for all of these datasets.

L

Since we put our data into 3rd Normal Form, we chose to use a relational database - postgreSQL. We uploaded our data from Jupyter Notebook into pgAdmin using the sqlalchemy package.