Rishi Hebbar
Prathyusha Challa

Project 2: Extract, Transform, Load

Project Statement:

For this project, our goal was to create a table showing select county-level data along with employment data . In addition we created a separate table showing ethnic breakdown in Illinois, also by county.

Methodology:

Extract:
The first dataset we worked with was the 2017 census county-level data (acs_county_data). This set included counties from all 50 states along with columns for child poverty, commuter information, eligible voters and more. The second data set was found on data.illinois.gov. This detailed employment numbers by county, total labor force, as well as the unemployment rate. The extraction process was done by simply downloading the csv files and reading them into dataframes using the pandas library.

Transform:
In order to widdle the datasets down to the information we need, columns in both the Illinois unemployment dataset and the census dataset were removed. The census data set included counties from all 50 states. Using the .loc function, we were able to select only the rows with Illinois in the state column. Once cleaned, the datasets were merged on the county column. The county column was set as an index and primary key. Column names were changed for readability and aesthetics. For the ethnicity data, we pulled select columns from the 2017 census data along with the county column to give us a break down of ethnicity.

Load:
Once the data was cleaned and merged, we loaded the data into MySQL using sqlalchemy. Our decision to go with a relational database was partly due to our familiarity as well as it's usefulness. If we did want to scale up and add more tables detailing county-level data, relational databases are very helpful and flexible.

Challenges:

We ran into issues while adding the Primary key on the column after creating the Table.

1. Problem: As we created the column using text and it gave us an error "Blob/Text column was used without any specification of Data" .

Solution : To Fix this error we converted the TEXT data to VARCHAR and specified the length

2. Problem: We also noticed that the data was duplicated in the table

   Solution: we created as we ran the code multiple times and to fix this we dropped the duplicate data added to the table and altered the Column name and added a Primary Key constraint.

Data Sources:

Census Data:
https://www.kaggle.com/muonneutrino/us-census-demographic-data#acs2017_county_data.csv
Illinois Gov Data:
https://data.illinois.gov/dataset/733unemployment_rate_by_county_apr17