ETL Project

**Team Members**: Matthew Barlow, Edmund Annobil, Natalie Cowart and Stanislav Mruz

**Background**: Homelessness is observed in every state in the US. Individuals are often found at risk of losing their homes due to various factors. Even in high income states whose salaries and wages are slightly higher than the average, homelessness is seen at alarming levels. Experts and researchers attribute this to many factors such as wages and salaries, lack of affordable housing, unemployment, taxes etc.

**Dataset Links**:

Icon

Description automatically generated1. <https://www.huduser.gov/portal/sites/default/files/xls/2007-2020-PIT-Estimates-by-state.xlsx>

2. <https://www.bls.gov/lau/lastrk20.htm>

**Extraction**

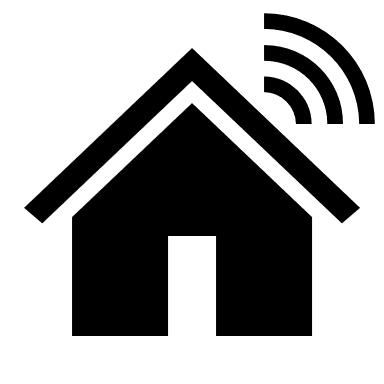
Icon

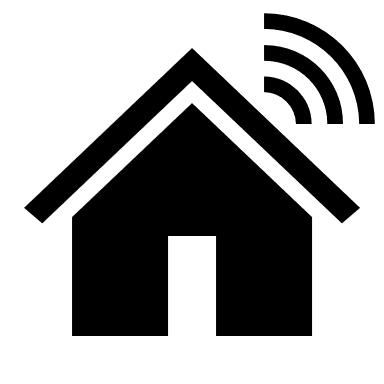
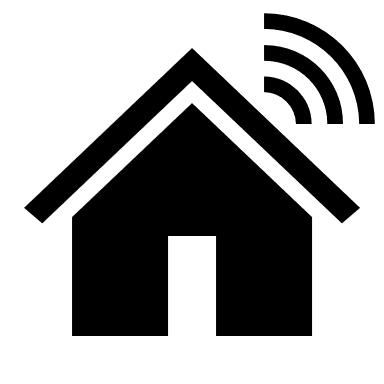
Description automatically generated Our datasets were retrieved from the Department of Housing and Urban Development and the Bureau of Labor Statistics. The focus of our data centered on unemployment and homelessness for all US States within the year 2020. We chose that year since it was a year that recorded unemployment rates at very high levels.

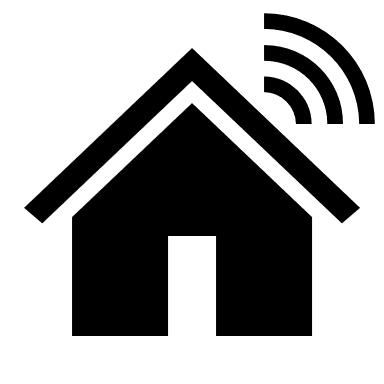
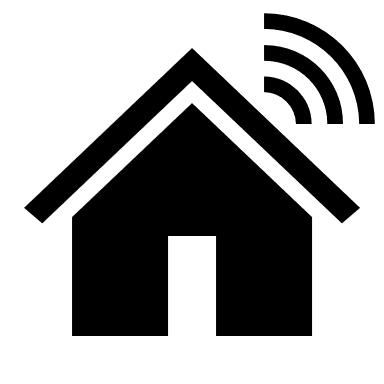
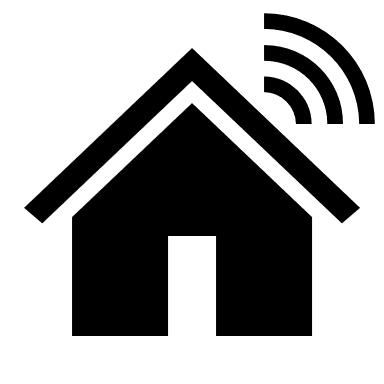
**Transformation**

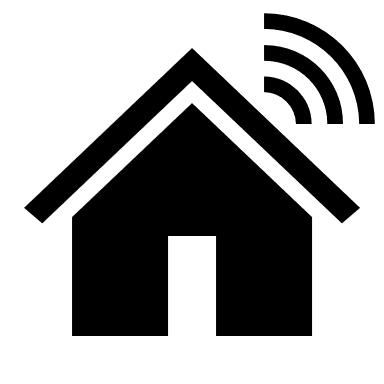
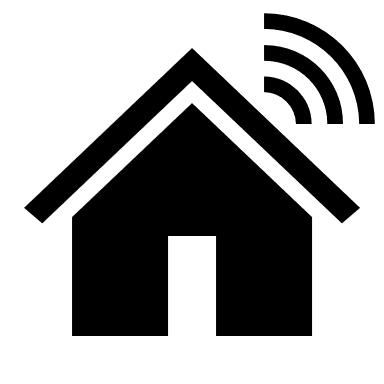
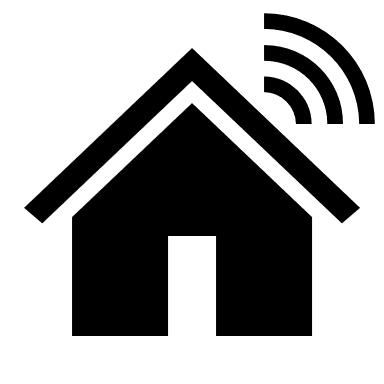
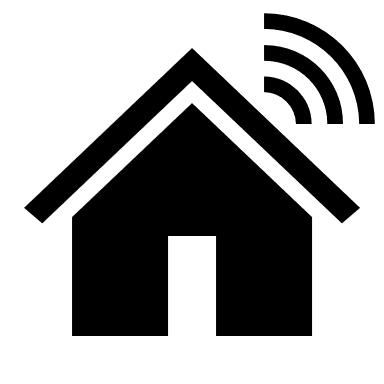
Icon

Description automatically generated After extracting the data into a Jupyter notebook, we inspected the data to ensure it contained the columns that we needed. In the unemployment data we discovered some rows had missing values and those were dropped. The unemployment data also contained certain regions in the US which were not states and those were deleted from the data as well. In the homelessness dataset, the state column was abbreviated which we needed to change into the full state name to merge with our unemployment data. We then renamed the columns that were relevant for our work and dropped the demographic data that was not. Figure 1 and Figure 2 below shows the unemployment dataset and homelessness dataset after these steps were performed.



**Imported the data set –** Imported through csv file and web scraping

**Cleaned each data set –** Dropped rows/NaN, fixed columns

**Checked/Fixed data types –** Assigned proper data type to each column

**Merged/Joined two data sets –** Merged csv was used to create the database

Figure 2: Homelessness Dataset after dropping and renaming columns

Figure 1: Unemployment Dataset after dropping and cleaning columns

Graphical user interface

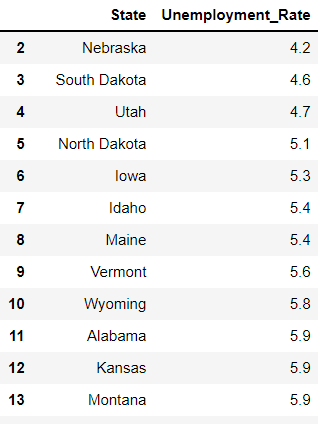
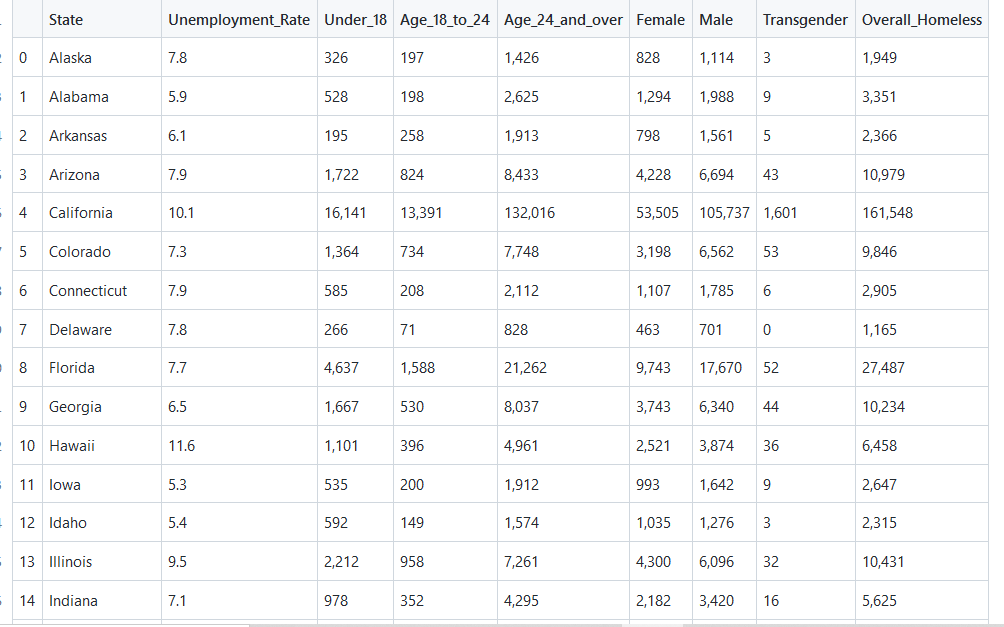
Description automatically generated

Figure 3: One single table, the imported index was dropped, and the remaining columns were reordered to a more logical format.



**Load**

Icon

Description automatically generated Lastly, the final step of the ETL process involved loading our final dataset into a database. An Entity Relationship Diagram (ERD) was created to help structure the design of the database and its objects. We created the database using the corresponding table(s) to match the columns in the Pandas DataFrame using PostgreSQL and then connected to the database using SQLAlchemy. This will allow any user to easily perform multiple queries in the database to retrieve information and/or perform analysis on that data. Figure 4 reflects the ERD diagram of the table(s) in the database and Figure 5 shows a sample query of the database to ensure the data was properly loaded.

Graphical user interface, application

Description automatically generated

Figure 4: ERD Diagram

Table

Description automatically generated

Figure 6: Sample Query using Pandas and SQLAlchemy.

**Conclusion:**

These datasets identify the unemployment rate, homelessness population and demographics by state. The final output will help us to recognize which state has the following.

* High/Low Unemployment Rate
* High/Low Homelessness
* Gender Index
* Age Index

These variables will help outline areas with high unemployment rate, homelessness, gender, and age brackets that need more assistance in any future aid or development plans. Also, this data could help companies identify which states need to create more jobs and contribute to a better economy.