Project 2 - Extract, Transform and Load
Clara Bucar and Ariel Jones

The Project:
This project consists of researching and downloading two separate datasets from different sources, transforming as required for merging and loading the clean merged file into a SQL Database. The process is commonly known as ETL: extract, transform and load.
The new database generated from the merge between the Census 2015 with the Zillow Rent Prices for the same time period allows the analysis of rent prices per gender breakdown, race breakdown, income, type of employment, type of commute and other information contained in the Census. This new merged data can help paint the picture of gentrification, historically identified by increasing rent prices and changes in tenant income overtime, more often than not accompanied by a change in ethnicity prevalence.

Proposed datasets and sources:
Census 2015:
https://www.kaggle.com/datasets/muonneutrino/us-census-demographic-data
Data extracted from the DP03 and DP05 tables of the 2015 American Community Survey 5-year estimates, collected by the US Census Bureau as a product of the US Federal Government. The full datasets can be found at the American Factfinder website. This source includes two data files listed below. For the purposes of this Project, only the County Data dataset will be considered.

1. acs2015censustract_data.csv: Data for each census tract in the US, including DC and Puerto Rico.
2. acs2015countydata.csv: Data for each county or county equivalent in the US, including DC and Puerto Rico.

Zillow Rent Prices:
https://www.kaggle.com/datasets/zillow/rent-index?resource=download&select=price.csv
Zillow operates an industry-leading economics and analytics bureau which produces extensive data on rent and real estate prices throughout the US. Zillow's database covers more than one hundred million American homes containing details of square footage, quantity of bedrooms and restrooms, tax assessments, previous sales, etc. This database uses the median home rent for US counties, it covers multifamily, single family, condos and cooperative homes, it also includes homes not presently available for rent. This database includes rent median values from the years 2010-2017.

*Acknowledgement: The rent index data was calculated from Zillow's proprietary Rent Zestimates and published on its website.

Project Steps:
Part 1 - Extracting the data and loading into Jupyter Notebook

After downloading the two separate datasets (Census and Zillow, referenced above) in csv format from Kaggle, loading these into Jupyter Notebook in order to start the cleaning process and next merging the data now in dataframe format.

Part 2 - Transforming the data

Multiple steps were taken to clean each dataframe explained in detail in the Jupyter Notebook file contained in this repository. Each dataframe had to be modified, cleaned and/or renamed in order to achieve one column with matching results to be used as the key for the merge. For this project, the County and the State columns in each dataframe was utilized, but had to be previously adjusted. Using the State columns was crucial to correctly identify Counties with similar names but located in different States. The Zillow dataset was used as the left in a left join. 449 rows with null values generated by a mismatch in the County naming between dataframes were dropped. Examples of the mismatch issues included counties containing the word 'City' as opposed to not containing that info (Baltimore City x Baltimore), counties containing the word 'Saint' as opposed to 'St'. The team made the decision to eliminate the 449 rows with county naming issues from a sample of 13131 counties.

Part 3: Loading the merged data

Using Python to connect AWS RDS MySQL database and using the Create Engine from the SQLAlchemy library to interpret the database API