ETL Project: Income, Housing, and Restaurants

Extract

- I extracted the data from Kaggle.com and Zillow, links are below
    - https://www.zillow.com/research/data/
        - ZHVI all Homes; Geography: Zip code
    - https://www.kaggle.com/goldenoakresearch/us-household-income-stats-geo-locations
    - https://www.kaggle.com/PromptCloudHQ/restaurants-on-yellowpagescom
        - This 3rd one was not part of my original proposal but I added it to make my project more robust and challenge myself more
- The files were all CSV, I had hoped to work with some other types but the data I needed was all in CSV format
- I used pd.read_csv to extract the data into a pandas dataframe

Transform

- As I extracted the data, I was able to keep it clean by adding the code below to keep leading zeros in the data frames, by initially loading them all converted as a string
    - converters={'Zip_Code': lambda x: str(x)}
    - **I after I figured out this code, I was able to eliminate a few steps that I had spent a lot of time on before I figured it out (highlighted in grey below). This made my code a lot more straightforward and cleaner. It was a huge Ah-Ha moment. (if you want to see the mess I had before, I saved a copy, you can just ask me on slack)**
- I first transformed by taking only the columns that I needed from the data sets
- Then I renamed the columns that needed a clearer or better formatted name
- I dropped duplicates in from the income dataset, to eliminate the duplicated data
- Then from the income and housing data sets I removed the 3-digit zip codes the were input errors, it only dropped a small handful of rows
    - Side note about zip codes: all three data sets had the same error, where if there was a 0(s) at the beginning of the zip code it did not appear on the table value. Luckily it effected all 3 data sets so the joins worked out fine
- Then for then for all three datasets I needed to convert the Zip_Code column to a string for purposes that will make sense as you keep reading
- Then I set the index to the Zip_Code so there was a good index to work off of and so I could later merge the indexes
- Then with the Restaurants (eats_df) I had to remove .0 from the end of the Zip_Code values
    - I could only figure out how to do this by converting the Zip_Code to a string
    - This is why I converted all of the Zip_Code columns to a string
- Then I joined the income and housing data sets together
- I then cleaned those by removing duplicate columns and changing the names of some columns
- Then I took that merged dataset and joined it with the restaurant dataset
- I then did similarly and cleaned up the columns by removing some and changing names

Load

- First, I created a connection to postgres
- Then I loaded each of the individual tables, and the joined tables, making a total of 5 tables that were loaded
    - I felt it was important to have all the tables since they could all be used for different types of analysis and they were all clean on their own
- Then I checked the tables in python to see if they were all there
- I also checked in PGadmin and everything looked as expected

This project was enjoyable to do, I learned a lot and it helped solidify some stuff I already knew. I was intimidated at first to work alone, but I showed myself I could come up with good results all on my own.