ETL Project

Extract:

The original data sources were extracted from redfin.com, homefinder.com, and trulia.com via BeautifulSoup to get html parser and set up browser to scrape data using splinter. The following was scraped from listed sites: price of listed properties, number of bedrooms, number of bathrooms, sq. footage, and zip code of listed properties.

Transform:

We transformed the data by removing “$” and “,” from the price and converting it to a numeric value. This conversion was repeated for “number of beds”, “number of baths”, and “area sq. ft.” values by removing text and converting the remaining value to float. This process was repeated with varying extracted values from data scraped from listed sites.

Once data was converted to consistent formats between data sets, Null and duplicate values were removed from data. This was done by dropping all addresses that were duplicated or had a value of na. The cleaned data was then combined into a single dataframe.csv file and ready to load into MongoDB.

Load:

To load the data into the database, a staging table was created in pgadmin. The data was loaded from the Final_DataFrame csv file using sql copy.