Tom Ashkenazi

ETL Project Report

Extract:

* The goal of my project was to extract listings from Zillow’s website to analyze the real estate market in Santa Clara county, California. I used Splinter and BeautifulSoup to gather and parse the HTML code from the Zillow page. For each listing I extracted the address, list price, number of beds, number of baths, and square ft. I then used Pandas to append the listings scraped from Zillow into a list, and then turned that list into a dataframe.

Transform

* Once I had the listings in a Dataframe I was able to transform and clean the data. I changed the datatypes of the list price and square ft columns from string to integer and float (respectively), allowing me to calculate price per square ft. I created another column displaying the listings’ price per square ft.
* I used the .split() function to split the address column into separate “street address”, “city”, and “zip code” columns, so that I could analyze listing data by geographic area. When I searched the zip code and city columns for unique values, I noticed that for some of the listings, the city or zip code field was either empty or filled with the wrong information (like having the name of the housing development instead of city). I manually searched these listings and imputed the correct city name and zip code.

Load

* I loaded my cleaned dataframe as a table in my local sql database. Then I confirmed that my data had been correctly added by querying the table. I used sql alchemy to count the number of listings in each city and in each zip code. I noticed that there were zip codes and cities with only a few listings. I then used the read\_sql\_query function to display cities and zip codes both by highest average price and highest average price per square ft, filtering out the cities and zip codes with only a few listings to avoid outliers.

Notes / Things to Improve

* The initial goal of my project was to analyze the broad real estate market in Santa Clara County, but I ended up researching only the higher-end real estate market in this area. When I scraped Zillow’s site I sorted my search by descending listing prices. Later I realized that Zillow caps its search at 20 pages, which meant that my search only included listings worth $1.4 million and above. To study the broader real estate market in this area I should have scraped separate Zillow searches for each zip code or city.