Steven Weil

Aaron Anderson

Professor Adjei

DSC 200

12/12/2024

Final Project Report

For our final project we decided to collect, clean, and output housing data from the Chicago, Illinois area. We knew that Chicago was one of the biggest cities in American and felt good about our ability to find sufficient data for it. Overall, we were able to find data for housing market trends, neighborhood demographic data, median housing and rental prices, rental listings, and relevant info from government housing policy reports. We used web scraping (Selenium), api access (requests), CSV files, excel files, and pdf files to gather all of the needed information. The data was found across the sites census.gov, apartments.com, and a pdf policy report from the United Stated Department of Housing and Urban Development. Once we had found all of the data, we began the process of extracting and cleaning it. The CSV and Excel files contain information about the neighborhood demographics sorted by zip code. However, census.gov had the zip codes in complicated columns. This was the first real struggle of the assignment. We eventually figured out we could use pandas and a mix of string splicing / Boolean series to sort out unnecessary columns and focus on the data we needed. Then, we transposed the data frames the data was stored in to ensure we could merge on the zip code column. Then, we accessed an API hosted by census.gov to access housing market trends. We had to sign up for an API key to access the data, but it was free and allowed us to page through the API by zip code using requests. Next, we used ethical web scraping to scrape housing and rental price data from Apartments.com. We also were able to scrape rental listings and descriptions from Apartments.com. It was here we ran into another struggle. Requests was unresponsive for this website, therefore we were required to use Selenium. Also, the price listings were often given as a price range, to handle this we averaged the lower and upper bounds together. We used the price data we found here to create calculated columns for the final csv file. The end result was two columns that contained the average price for houses and rentals per zip code.

Once we had all of this data and had thoroughly cleaned it, it was time to merge the data. To merge, we called all of the separate functions from main, and each returned a data frame. Then, we called the merge function and passed the data frames into it. It performed inner joins on the zip code column for each table and created the final merged data frame. Finally, the main function converted that to a CSV using pandas and output the final product.

The other deliverable was the PDF file. We were unable to find a government pdf file that had data at the granularity we needed. So, we found a pdf of a report from the US department of Housing and Urban Development. It is a comprehensive housing market analysis, and we were able to find a chart that deals with many housing market statistics over the past 30 years in downtown Chicago. We used slate3k to access and extract the data from the PDF. Then we used pandas to convert it to a data frame and clean the data before returning it back to main to be output.

Overall, this assignment was an excellent opportunity for us to put everything we learned this semester into practice. The skills we have obtained will be invaluable going forward in both data science and in life. We worked hard on this assignment and hope that it turned out well. Thank you for a great semester professor!