PROJECT DOCUMENTATION

Table of Contents

BACKGROUND

Overview

Mappa is a company that aims to redefine the experience of finding a new home by using the power of data and technology. The goal of this project is to create a master list of listed properties across the Greater London area (United Kingdom) in one dataset. The purpose of this dataset is to provide house hunters with a comprehensive and up-to-date list of available properties, enabling them to make informed decisions and secure the best value for their money.

Project Objectives

The main objective of this project is to extract and blend property datasets across the Greater London area in one dataset. This involves scraping data from various sources, cleaning and processing the data, and then combining it into a single, comprehensive dataset. The specific objectives of the project are:

1. To identify and scrape property data from various sources across the Greater London area.
2. To clean and pre-process the scraped data to ensure consistency and accuracy.
3. To combine the scraped data into a single, comprehensive dataset.
4. To make the final dataset available for use by house hunters.

Project Approach

To achieve the project objectives, we will be using web scraping techniques to extract property data from various sources, including property listing websites and real estate agents' websites. We will then use data cleaning and pre-processing techniques to ensure that the data is consistent and accurate. Once the data has been cleaned and pre-processed, we will merge the data into a single dataset using data blending techniques. Finally, we will make the dataset available for use by house hunters.

Project Deliverables

The main deliverable of this project is the comprehensive dataset of listed properties across the Greater London area. This dataset will be made available in a format that can be easily accessed and used by house hunters. In addition, we will also provide documentation on the data sources used, the data cleaning and pre-processing techniques used, and the data blending techniques used to create the final dataset.

METHODOLOGY

For this project, we followed a three-stage approach, which included:

1. Data Gathering
2. Data cleaning, and
3. Data merging.

Data Gathering

The scraping process began by utilizing the Beautifulsoup package for data gathering. However, the websites being scraped heavily depend on JavaScript, requiring the implementation of alternative strategies. To address this, the web scraping process was carried out using the Selenium library in Python. Selenium is an effective tool for extracting information from dynamic websites that heavily rely on JavaScript to generate content. By utilizing Selenium, interactions with the website, searches, and extraction of the desired data were automated.

The first step was to download and Install Selenium drivers or Chrome (the driver is also available for other browsers). The link to download the driver is here.

For each of the three websites (Rightmove, On the Market, and Zoopla), the web scraping process was divided into two stages: Properties for sale (Sales) and properties for rent (Rent). In addition, a web scraping function was created for each stage for all websites, making it a total of six functions.

Taking the web scraping process for Right Move as a case study. The two functions for sales and rent are rightMove_sales and rightMove_rent, respectively. Both functions scrape property data from Rightmove for a given list of postcodes and transaction types and return the data as a pandas DataFrame. Here are the four arguments the function takes:

1. Postcodes -- a list of postcodes for which to scrape property data
2. Trans_type -- the transaction type of the properties to be scrapped ('sales' or 'rent')
3. Website -- the name of the website being scraped (in this case, 'Rightmove')
4. df -- an empty panda DataFrame to store the scraped data

Inside the function, the ChromeDriver executable is set as the WebDriver, and the ChromeDriver is launched with the specified service. Then, the function navigates to the website and interacts with the user interface to initiate a search for properties.

Using the Selenium WebDriver, the function locates and interacts with different elements on the webpage, such as canceling pop-up windows, maximizing the window, selecting the transaction type, entering the postcode in the search bar, and clicking the search button.

To ensure all relevant data is obtained, the function goes through multiple pages of search results. First, it scrapes property details such as an address, property type, number of bedrooms and bathrooms, price, description, listing date, agent details, property URL, and website source. These details are stored in individual lists.

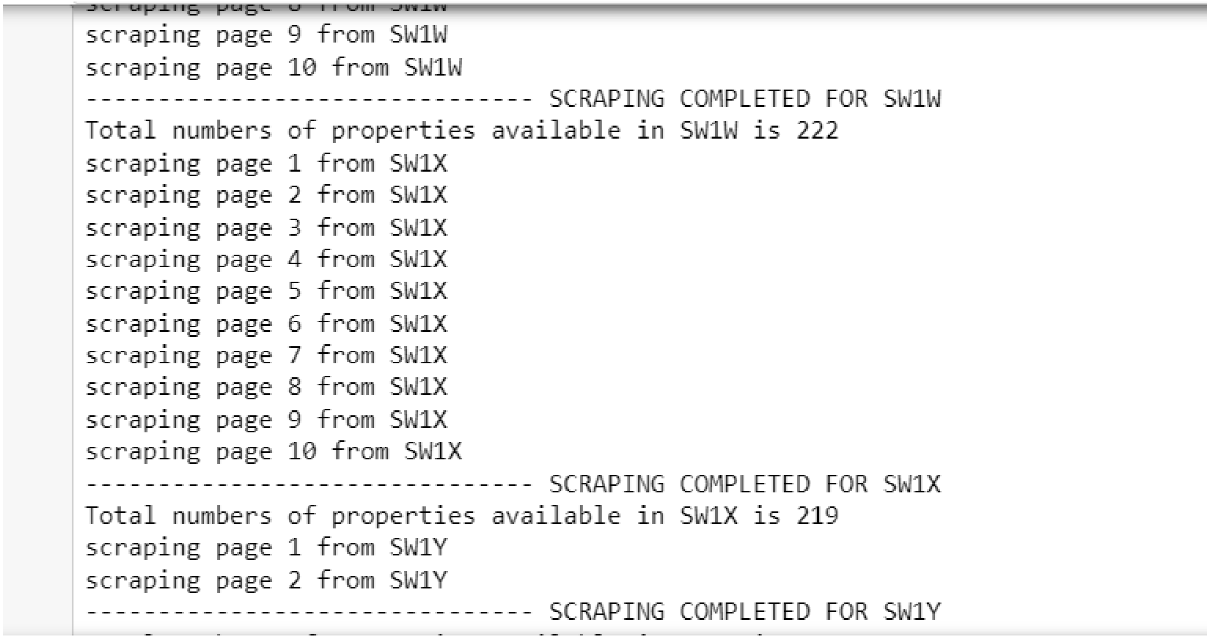
After scraping the data from each page, the function scrolls down to load more properties and clicks on any pop-up windows that may appear. It then finds the next button to navigate to the next page of results, repeating this process until all pages have been scrapped.

This same technique is used for the other two websites. Finally, the scraped data for all postcodes are stored in a Pandas DataFrame. The scraped data has 14 different features, namely:

1. Unique Id: Consist of the postal code, the transaction type, the scrapping serial number, and the website name. E.g., For BR1S00001OM, BR1 - Postal code, S - Sales, 00001 - Serial number, OM - On the Market Website.
2. Location
3. Transaction type
4. Property type
5. Address
6. Number of bedrooms
7. Number of Bathrooms
8. Price
9. Description
10. Listing date
11. Agent name
12. Listing source
13. Property URL
14. Current date: The date the web scraping was performed (Applicable to only On the mar website)

The web scraping function was enhanced with informative print statements to monitor the progress for each website page and postal code. This feature allows for easy tracking and ensures that in case of network issues or interruptions, the scraping process can be resumed from the point where it left off. Figure 2.1 provides a visual representation of these print statements for reference. The Datasets obtained from the scrapping were then saved into a Comma-Separated-Value (CSV) file.

N.B.: When web scraping the Zoopla website, the installation of an undetected chrome driver was necessary to bypass the website security due to its protection against third-party applications using Cloudware.



```
scraping page 8 from SW1W
scraping page 9 from SW1W
scraping page 10 from SW1W
------------------------------ SCRAPING COMPLETED FOR SW1W
Total numbers of properties available in SW1W is 222
scraping page 1 from SW1X
scraping page 2 from SW1X
scraping page 3 from SW1X
scraping page 4 from SW1X
scraping page 5 from SW1X
scraping page 6 from SW1X
scraping page 7 from SW1X
scraping page 8 from SW1X
scraping page 9 from SW1X
scraping page 10 from SW1X
------------------------------ SCRAPING COMPLETED FOR SW1X
Total numbers of properties available in SW1X is 219
scraping page 1 from SW1Y
scraping page 2 from SW1Y
------------------------------ SCRAPING COMPLETED FOR SW1Y
```

Figure 2.1: A visual representation of these print statements for reference

Next, we will move on to the data cleaning stage to ensure the scraped data is consistent and accurate for further analysis.

Data Cleaning

It is commonly acknowledged that data is often characterized by its unclean and disorderly nature, necessitating data wrangling and cleaning. In this project, the sales and rent datasets obtained from each website were subjected to separate cleaning processes. These cleaning procedures involved three distinct stages:

a) Data Assessment: In this stage, the data was meticulously examined to gain an initial understanding. The shape of the data was investigated, along with an analysis of value distributions within the dataset. Additionally, the data types of each column were examined to ensure consistency.

b) Dealing with Null Values: This stage focused on addressing null values within the dataset. Given the unique nature of this project, null values were replaced with the term 'Not Listed' rather than being dropped. This replacement signifies that the corresponding feature was not listed on the website, providing clear communication of missing values.

Note: It is important to highlight that the majority of columns, including the number of bathrooms and bedrooms, have an object data type. This is because, instead of dropping null values, they were replaced with the string 'Not listed'.

c) Data Cleaning: The data cleaning process encompassed various techniques to address any remaining inconsistencies and untidiness. This involved modifying the column structure to ensure uniformity across datasets and removing irrelevant words or characters from values.

These data cleaning procedures were applied to all six datasets, ensuring that the data was prepared and standardized for the subsequent merging process.

Data Merging

The merging of datasets was performed to generate a comprehensive master dataset. Initially, the sales and rent data from the three websites were merged individually. Subsequently, the sales data was merged with the rent data to form the final dataset.

To ensure an accurate merging process between the sales data from different websites, a systematic approach was adopted to eliminate duplicates. The identification of duplicates was based on specific columns, namely 'Location', 'Transaction_Type', 'Property_Type', 'Address', 'Bedrooms', 'Bathrooms', 'Price', 'Description', and 'Agent'. Notably, the 'unique_id', 'listing_source', and 'source_url' columns were excluded from the duplicate check, as these attributes are website specific.

It is important to note that the 'Price' column in the sales and rent datasets had different interpretations. In the sales dataset, it represented the property price, whereas in the rent dataset, it denoted the monthly rental price. To address this discrepancy, an additional column was incorporated into both datasets. In the sales dataset, the new column 'Price per month' was added, and its values were populated with 'Not listed'. Similarly, in the rent dataset, the new column 'Price' was introduced, and its values were filled with 'Not listed'. This adjustment ensured a consistent data structure across the merged datasets.

Upon merging the sales and rent datasets, the final master dataset was obtained, incorporating all relevant information from the three websites.

CONCLUSION

Result

The result of this project is the creation of a comprehensive master dataset, which consists of the merged sales and rent data from three different websites: Rightmove, On the Market, and Zoopla. The master dataset comprises 175,395 rows and 14 columns, providing a rich collection of property information for analysis and further exploration.

Limitations and Challenges

The scope of this project was primarily focused on data scraping and cleaning processes. Due to time constraints and project requirements, further in-depth analysis and exploration of the dataset were not conducted as part of this project. However, the master dataset serves as a valuable foundation for future analysis and insights.

Throughout the project, several challenges were encountered, primarily during the web scraping phase. The process of web scraping was time-consuming and relied on the speed and stability of the network connection. Variations in website structures, dynamic content, and anti-scraping mechanisms posed additional challenges during data collection. Nonetheless, with careful handling and the utilization of tools such as Selenium, these challenges were successfully overcome.

Despite the limitations and challenges, the project has achieved its main objectives of gathering, cleaning, and merging property data from multiple sources. The resulting master dataset provides a valuable resource for further analysis and decision-making in the real estate domain.

In conclusion, this project has laid the foundation for harnessing the power of data to enhance the house-hunting experience. The master dataset serves as a comprehensive source of property information, empowering users with valuable insights and knowledge to make informed decisions in their search for a new home.