Navigating and Data Collection of Airbnb Listings: A Selenium WebDriver Approach

Introduction

This Python script employs the Selenium package to navigate and scrape data from Airbnb's website. The process is automated through a web driver that simulates a user browsing through property listings on Airbnb. Specifically, the script is designed to:

Construct the URL for the Airbnb search page based on a specified location and date range.

Open the URL in a browser window controlled by Selenium.

Sequentially navigate through the first three pages of the search results.

For each listing found, extract the URL, nightly price, number of beds, and user rating, handling instances where information may be missing.

Store the gathered information in a Pandas DataFrame and Export the DataFrame to a CSV file for subsequent use or analysis.

In [2]:

```
#pip install selenium
```

```
Collecting selenium
  Obtaining dependency information for selenium from https://files.pythonhosted.org/p
ackages/b4/f9/e9ac5e4c5d84b07c7d117d67b2c84be221bcb9e62ff31fd0a1bbc06099c0/selenium-
4.19.0-py3-none-any.whl.metadata
  Downloading selenium-4.19.0-py3-none-any.whl.metadata (6.9 kB)
Requirement already satisfied: urllib3[socks]<3,>=1.26 in c:\users\shubh\anaconda3\li
b\site-packages (from selenium) (1.26.16)
Collecting trio~=0.17 (from selenium)
  Obtaining dependency information for trio~=0.17 from https://files.pythonhosted.or
g/packages/17/c9/f86f89f14d52f9f2f652ce24cb2f60141a51d087db1563f3fba94ba07346/trio-0.
25.0-py3-none-any.whl.metadata
  Downloading trio-0.25.0-py3-none-any.whl.metadata (8.7 kB)
Collecting trio-websocket~=0.9 (from selenium)
  Obtaining dependency information for trio-websocket~=0.9 from https://files.pythonh
osted.org/packages/48/be/a9ae5f50cad5b6f85bd2574c2c923730098530096e170c1ce7452394d7a
a/trio_websocket-0.11.1-py3-none-any.whl.metadata
  Downloading trio_websocket-0.11.1-py3-none-any.whl.metadata (4.7 kB)
Requirement already satisfied: certifi>=2021.10.8 in c:\users\shubh\anaconda3\lib\sit
e-packages (from selenium) (2023.7.22)
Collecting typing_extensions>=4.9.0 (from selenium)
  Obtaining dependency information for typing_extensions>=4.9.0 from https://files.py
thonhosted.org/packages/f9/de/dc04a3ea60b22624b51c703a84bbe0184abcd1d0b9bc8074b5d6b7a
b90bb/typing_extensions-4.10.0-py3-none-any.whl.metadata
  Downloading typing_extensions-4.10.0-py3-none-any.whl.metadata (3.0 kB)
Collecting attrs>=23.2.0 (from trio~=0.17->selenium)
  Obtaining dependency information for attrs>=23.2.0 from https://files.pythonhosted.
org/packages/e0/44/827b2a91a5816512fcaf3cc4ebc465ccd5d598c45cefa6703fcf4a79018f/attrs
-23.2.0-py3-none-any.whl.metadata
  Downloading attrs-23.2.0-py3-none-any.whl.metadata (9.5 kB)
Requirement already satisfied: sortedcontainers in c:\users\shubh\anaconda3\lib\site-
packages (from trio~=0.17->selenium) (2.4.0)
Requirement already satisfied: idna in c:\users\shubh\anaconda3\lib\site-packages (fr
om trio~=0.17->selenium) (3.4)
Collecting outcome (from trio~=0.17->selenium)
  Obtaining dependency information for outcome from https://files.pythonhosted.org/pa
ckages/55/8b/5ab7257531a5d830fc8000c476e63c935488d74609b50f9384a643ec0a62/outcome-1.
3.0.post0-py2.py3-none-any.whl.metadata
  Downloading outcome-1.3.0.post0-py2.py3-none-any.whl.metadata (2.6 kB)
Collecting sniffio>=1.3.0 (from trio~=0.17->selenium)
  Obtaining dependency information for sniffio>=1.3.0 from https://files.pythonhoste
d.org/packages/e9/44/75a9c9421471a6c4805dbf2356f7c181a29c1879239abab1ea2cc8f38b40/sni
ffio-1.3.1-py3-none-any.whl.metadata
  Downloading sniffio-1.3.1-py3-none-any.whl.metadata (3.9 kB)
Requirement already satisfied: cffi>=1.14 in c:\users\shubh\anaconda3\lib\site-packag
es (from trio~=0.17->selenium) (1.15.1)
Collecting wsproto>=0.14 (from trio-websocket~=0.9->selenium)
  Obtaining dependency information for wsproto>=0.14 from https://files.pythonhosted.
org/packages/78/58/e860788190eba3bcce367f74d29c4675466ce8dddfba85f7827588416f01/wspro
to-1.2.0-py3-none-any.whl.metadata
  Downloading wsproto-1.2.0-py3-none-any.whl.metadata (5.6 kB)
Requirement already satisfied: PySocks!=1.5.7,<2.0,>=1.5.6 in c:\users\shubh\anaconda
3\lib\site-packages (from urllib3[socks]<3,>=1.26->selenium) (1.7.1)
Requirement already satisfied: pycparser in c:\users\shubh\anaconda3\lib\site-package
s (from cffi>=1.14->trio~=0.17->selenium) (2.21)
Collecting h11<1,>=0.9.0 (from wsproto>=0.14->trio-websocket~=0.9->selenium)
  Obtaining dependency information for h11<1,>=0.9.0 from https://files.pythonhosted.
org/packages/95/04/ff642e65ad6b90db43e668d70ffb6736436c7ce41fcc549f4e9472234127/h11-
0.14.0-py3-none-any.whl.metadata
  Downloading h11-0.14.0-py3-none-any.whl.metadata (8.2 kB)
Downloading selenium-4.19.0-py3-none-any.whl (10.5 MB)
```

```
   ---------------------------------------- 0.0/10.5 MB ? eta -:--:--
   ---------------------------------------- 0.0/10.5 MB ? eta -:--:--
    --------------------------------------- 0.2/10.5 MB 1.8 MB/s eta 0:00:06
   - -------------------------------------- 0.5/10.5 MB 3.3 MB/s eta 0:00:04
   -- ------------------------------------- 0.6/10.5 MB 3.7 MB/s eta 0:00:03
   --- ------------------------------------ 1.0/10.5 MB 4.1 MB/s eta 0:00:03
   ---- ----------------------------------- 1.2/10.5 MB 4.4 MB/s eta 0:00:03
   ----- ---------------------------------- 1.5/10.5 MB 4.7 MB/s eta 0:00:02
   ------ --------------------------------- 1.7/10.5 MB 4.8 MB/s eta 0:00:02
   ------- -------------------------------- 2.0/10.5 MB 4.9 MB/s eta 0:00:02
   -------- ------------------------------- 2.3/10.5 MB 4.8 MB/s eta 0:00:02
   --------- ------------------------------ 2.5/10.5 MB 4.8 MB/s eta 0:00:02
   ---------- ----------------------------- 2.7/10.5 MB 4.9 MB/s eta 0:00:02
   ----------- ---------------------------- 2.9/10.5 MB 4.8 MB/s eta 0:00:02
   ------------ --------------------------- 3.3/10.5 MB 4.9 MB/s eta 0:00:02
   ------------ --------------------------- 3.4/10.5 MB 4.8 MB/s eta 0:00:02
   ------------- -------------------------- 3.7/10.5 MB 4.9 MB/s eta 0:00:02
   -------------- ------------------------- 3.8/10.5 MB 4.8 MB/s eta 0:00:02
   --------------- ------------------------ 4.2/10.5 MB 4.9 MB/s eta 0:00:02
   ---------------- ----------------------- 4.4/10.5 MB 4.9 MB/s eta 0:00:02
   ------------------ --------------------- 4.8/10.5 MB 5.1 MB/s eta 0:00:02
   ------------------- -------------------- 5.1/10.5 MB 5.2 MB/s eta 0:00:02
   -------------------- ------------------- 5.5/10.5 MB 5.3 MB/s eta 0:00:01
   --------------------- ------------------ 5.7/10.5 MB 5.3 MB/s eta 0:00:01
   ---------------------- ----------------- 5.9/10.5 MB 5.2 MB/s eta 0:00:01
   ----------------------- ---------------- 6.2/10.5 MB 5.2 MB/s eta 0:00:01
   ------------------------ --------------- 6.3/10.5 MB 5.2 MB/s eta 0:00:01
   ------------------------- -------------- 6.7/10.5 MB 5.3 MB/s eta 0:00:01
   ------------------------- -------------- 6.8/10.5 MB 5.2 MB/s eta 0:00:01
   --------------------------- ------------ 7.1/10.5 MB 5.2 MB/s eta 0:00:01
   --------------------------- ------------ 7.2/10.5 MB 5.1 MB/s eta 0:00:01
   ---------------------------- ----------- 7.6/10.5 MB 5.2 MB/s eta 0:00:01
   ----------------------------- ---------- 7.8/10.5 MB 5.1 MB/s eta 0:00:01
   ------------------------------ --------- 8.1/10.5 MB 5.2 MB/s eta 0:00:01
   ------------------------------- -------- 8.3/10.5 MB 5.2 MB/s eta 0:00:01
   -------------------------------- ------- 8.6/10.5 MB 5.2 MB/s eta 0:00:01
   --------------------------------- ------ 8.8/10.5 MB 5.2 MB/s eta 0:00:01
   ---------------------------------- ----- 9.1/10.5 MB 5.2 MB/s eta 0:00:01
   ----------------------------------- ---- 9.3/10.5 MB 5.2 MB/s eta 0:00:01
   ------------------------------------ --- 9.6/10.5 MB 5.2 MB/s eta 0:00:01
   ------------------------------------- -- 10.0/10.5 MB 5.3 MB/s eta 0:00:01
   ---------------------------------------  10.3/10.5 MB 5.3 MB/s eta 0:00:01
   ---------------------------------------  10.5/10.5 MB 5.5 MB/s eta 0:00:01
   ---------------------------------------- 10.5/10.5 MB 5.4 MB/s eta 0:00:00
Downloading trio-0.25.0-py3-none-any.whl (467 kB)
   ---------------------------------------- 0.0/467.2 kB ? eta -:--:--
   ---------------------------- ----------- 327.7/467.2 kB 9.9 MB/s eta 0:00:01
   ---------------------------------------- 467.2/467.2 kB 5.9 MB/s eta 0:00:00
Downloading trio_websocket-0.11.1-py3-none-any.whl (17 kB)
Downloading typing_extensions-4.10.0-py3-none-any.whl (33 kB)
Downloading attrs-23.2.0-py3-none-any.whl (60 kB)
   ---------------------------------------- 0.0/60.8 kB ? eta -:--:--
   ---------------------------------------- 60.8/60.8 kB ? eta 0:00:00
Downloading sniffio-1.3.1-py3-none-any.whl (10 kB)
Downloading wsproto-1.2.0-py3-none-any.whl (24 kB)
Downloading outcome-1.3.0.post0-py2.py3-none-any.whl (10 kB)
Downloading h11-0.14.0-py3-none-any.whl (58 kB)
   ---------------------------------------- 0.0/58.3 kB ? eta -:--:--
   ---------------------------------------- 58.3/58.3 kB ? eta 0:00:00
Installing collected packages: typing_extensions, sniffio, h11, attrs, wsproto, outco
```

```
me, trio, trio-websocket, selenium
  Attempting uninstall: typing_extensions
    Found existing installation: typing_extensions 4.7.1
    Uninstalling typing_extensions-4.7.1:
      Successfully uninstalled typing_extensions-4.7.1
  Attempting uninstall: sniffio
    Found existing installation: sniffio 1.2.0
    Uninstalling sniffio-1.2.0:
      Successfully uninstalled sniffio-1.2.0
  Attempting uninstall: attrs
    Found existing installation: attrs 22.1.0
    Uninstalling attrs-22.1.0:
      Successfully uninstalled attrs-22.1.0
Successfully installed attrs-23.2.0 h11-0.14.0 outcome-1.3.0.post0 selenium-4.19.0 sn
iffio-1.3.1 trio-0.25.0 trio-websocket-0.11.1 typing_extensions-4.10.0 wsproto-1.2.0
Note: you may need to restart the kernel to use updated packages.
```

In [ ]:

```
#import necessary libraries and modules
#selenium and its components for web scraping.
#time and random for managing delays in page loading and clicks to simulate human beha
#pandas for data manipulation and saving the scraped data into a CSV file
```

In [1]:

```
from selenium import webdriver
from selenium.webdriver.common.by import By
from selenium.webdriver.chrome.service import Service
from selenium.webdriver.support.ui import WebDriverWait
from selenium.webdriver.support import expected_conditions as EC
from selenium.common.exceptions import TimeoutException, NoSuchElementException
import time
import random
import pandas as pd
```

In [2]:

```
#Setting Up the WebDriver

serv_obj = Service('C:/Users/shubh/Downloads/chromedriver-win64/chromedriver-win64/chr
driver = webdriver.Chrome(service=serv_obj)
```

In [3]:

```
#Defining Search Parameters for the Airbnb location and the check-in and check-out dat

location = 'williamsburg-va'
checkin = '2024-03-18'
checkout = '2024-03-24'
```

In [4]:

```
# Building the Website URL to scrape
url_to_scrape = "https://www.airbnb.com/s/"+location+"/homes?tab_id=home_tab&date_pick

#The WebDriver opens the constructed URL in a browser window
driver.get(url_to_scrape)

#An empty list is initialized to store the scraped listing details
data = []

# Wait for the website to load with a timeout of 15 seconds
wait = WebDriverWait(driver, 15)  # Initialize once and use throughout
```

In [5]:

```
#Page Scraping Loop is set to scrape data from up to 3 pages of listings

page = 1
```

```
while page <= 3:
    try:
        #a try-except block is used to handle exceptions gracefully, such as timeouts
        listings_xpath_base = '//*[@id="site-content"]/div/div[2]/div/div/div/div/div[
        #//*[@id="site-content"]/div/div[2]/div/div/div/div/div[1]/div[1]/div/div[2]/d
        num_listings = len(driver.find_elements(By.XPATH, listings_xpath_base))

        for i in range(1, num_listings + 1):#Iterates over each listing found on the p
            try:
                listing_xpath = f'{listings_xpath_base}[{i}]'
                listing_url_xpath = f'{listing_xpath}/div/div[2]/div/div/div/div/a'
                price_xpath = f'{listing_xpath}/div/div[2]/div/div/div/div/div/div[2]/
                beds_xpath = f'{listing_xpath}/div/div[2]/div/div/div/div/div/div[2]/d
                rating_xpath = f'{listing_xpath}/div/div[2]/div/div/div/div/div/div[2]

                listing_details = {
                    'URL': driver.find_element(By.XPATH, listing_url_xpath).get_attrib
                    'Price': driver.find_element(By.XPATH, price_xpath).text,
                    'Beds': driver.find_element(By.XPATH, beds_xpath).text,
                    'Rating': driver.find_element(By.XPATH, rating_xpath).text
                }
                #constructs XPaths for various details (URL, price, beds, rating),
                #retrieves the information using Selenium, and stores it in a dictiona

                data.append(listing_details)
            except NoSuchElementException:
                print(f"Missing information for listing {i}, skipping.")
                continue

    except TimeoutException:
        print("Timeout while waiting for listings to load.")
        break

    if page < 3:
        try:
            #Finds the "Next" button, clicks it to go to the next page of listings and

            next_button = wait.until(EC.element_to_be_clickable((By.XPATH, '//a[contai
            driver.get(next_button.get_attribute('href'))
            page += 1
            time.sleep(random.randint(2, 15))  # Ensure the next page is fully loaded
        except NoSuchElementException:
            print("Reached the end of the pages.")
            break
        except Exception as e:
            print(f"An error occurred while navigating: {e}")
            break
    else:
        break
```

```
Missing information for listing 1, skipping.
Missing information for listing 16, skipping.
Missing information for listing 5, skipping.
Missing information for listing 11, skipping.
Missing information for listing 14, skipping.
Missing information for listing 15, skipping.
Missing information for listing 17, skipping.
```

In [6]:

```
#Finally, the browser is closed,
#the scraped data is converted into a pandas DataFrame

driver.quit()

df = pd.DataFrame(data)

print(df)
```

```
                                                  URL  \
0   https://www.airbnb.com/rooms/10711339744456620...   
1   https://www.airbnb.com/rooms/69072429954963487...   
2   https://www.airbnb.com/rooms/11015873820985475...   
3   https://www.airbnb.com/rooms/50155260?adults=1...   
4   https://www.airbnb.com/rooms/43614314?adults=1...   
5   https://www.airbnb.com/rooms/50027632?adults=1...   
6   https://www.airbnb.com/rooms/70212613522070648...   
7   https://www.airbnb.com/rooms/9595397?adults=1&...   
8   https://www.airbnb.com/rooms/11079587108399276...   
9   https://www.airbnb.com/rooms/20031636?adults=1...   
10  https://www.airbnb.com/rooms/41371157?adults=1...   
11  https://www.airbnb.com/rooms/63355636240451489...   
12  https://www.airbnb.com/rooms/50976430?adults=1...   
13  https://www.airbnb.com/rooms/14788657?adults=1...   
14  https://www.airbnb.com/rooms/15490423?adults=1...   
15  https://www.airbnb.com/rooms/72101225582397146...   
16  https://www.airbnb.com/rooms/71657074593237941...   
17  https://www.airbnb.com/rooms/49775865?adults=1...   
18  https://www.airbnb.com/rooms/74184877051672271...   
19  https://www.airbnb.com/rooms/53058322?adults=1...   
20  https://www.airbnb.com/rooms/10922808752520542...   
21  https://www.airbnb.com/rooms/34676321?adults=1...   
22  https://www.airbnb.com/rooms/80526980208395315...   
23  https://www.airbnb.com/rooms/47402400?adults=1...   
24  https://www.airbnb.com/rooms/14251373?adults=1...   
25  https://www.airbnb.com/rooms/20052973?adults=1...   
26  https://www.airbnb.com/rooms/33660829?adults=1...   
27  https://www.airbnb.com/rooms/83611198650835150...   
28  https://www.airbnb.com/rooms/36137997?adults=1...   
29  https://www.airbnb.com/rooms/45712169?adults=1...   
30  https://www.airbnb.com/rooms/45502133?adults=1...   
31  https://www.airbnb.com/rooms/82740254397010029...   
32  https://www.airbnb.com/rooms/52060654?adults=1...   
33  https://www.airbnb.com/rooms/70162207386414136...   
34  https://www.airbnb.com/rooms/70162207386414136...   
35  https://www.airbnb.com/rooms/39192659?adults=1...   
36  https://www.airbnb.com/rooms/71746746582977687...   
37  https://www.airbnb.com/rooms/77085030233667035...   
38  https://www.airbnb.com/rooms/59343109624631169...   
39  https://www.airbnb.com/rooms/73569383950491620...   
40  https://www.airbnb.com/rooms/17104856?adults=1...   
41  https://www.airbnb.com/rooms/49945692?adults=1...   
42  https://www.airbnb.com/rooms/13192779?adults=1...   
43  https://www.airbnb.com/rooms/86880023123489472...   
44  https://www.airbnb.com/rooms/53613567?adults=1...   
45  https://www.airbnb.com/rooms/85290474925597170...   
46  https://www.airbnb.com/rooms/25326413?adults=1...   

                                Price                 Beds  \
0                     $61 per night               2 beds   
1                    $174 per night               4 beds   
2                     $69 per night               3 beds   
3                    $164 per night               6 beds   
4   $101 per night, originally $160          1 queen bed   
5                    $125 per night               4 beds   
6                     $93 per night                1 bed   
7   $138 per night, originally $179               2 beds   
8                     $89 per night               3 beds   
9                     $98 per night               3 beds   
```

```
10                    $85 per night          1 queen bed
11  $123 per night, originally $147        2 double beds
12                   $175 per night               4 beds
13                   $122 per night         2 queen beds
14  $122 per night, originally $140               2 beds
15                    $75 per night               2 beds
16                    $85 per night                1 bed
17                    $79 per night  2 small double beds
18                    $92 per night        2 double beds
19                   $119 per night               3 beds
20                    $78 per night               3 beds
21                   $162 per night               3 beds
22                   $211 per night               4 beds
23                   $189 per night           1 king bed
24                   $188 per night               4 beds
25                    $98 per night               3 beds
26                   $135 per night               4 beds
27                   $142 per night           1 king bed
28                   $289 per night               8 beds
29   $81 per night, originally $123           1 king bed
30                   $113 per night               4 beds
31                   $160 per night               2 beds
32                   $180 per night               5 beds
33                   $111 per night               3 beds
34                   $111 per night               3 beds
35                   $111 per night               3 beds
36                    $73 per night                1 bed
37                    $85 per night           1 king bed
38                   $143 per night               3 beds
39                    $91 per night                1 bed
40                   $157 per night               3 beds
41                   $105 per night          1 queen bed
42                    $85 per night               3 beds
43                   $194 per night               3 beds
44                    $81 per night               2 beds
45                   $118 per night           1 king bed
46                   $145 per night               3 beds

                                       Rating
0      5.0 out of 5 average rating, 5 reviews
1    4.96 out of 5 average rating, 74 reviews
2                           New place to stay
3   4.99 out of 5 average rating, 199 reviews
4   4.83 out of 5 average rating, 357 reviews
5   4.86 out of 5 average rating, 164 reviews
6     4.88 out of 5 average rating, 8 reviews
7   4.81 out of 5 average rating, 747 reviews
8                           New place to stay
9   4.91 out of 5 average rating, 232 reviews
10  4.94 out of 5 average rating, 173 reviews
11   4.83 out of 5 average rating, 64 reviews
12  4.83 out of 5 average rating, 173 reviews
13  4.97 out of 5 average rating, 173 reviews
14  4.97 out of 5 average rating, 265 reviews
15    4.6 out of 5 average rating, 10 reviews
16    4.78 out of 5 average rating, 9 reviews
17   4.96 out of 5 average rating, 96 reviews
18   4.96 out of 5 average rating, 49 reviews
19  4.93 out of 5 average rating, 122 reviews
20                          New place to stay
```

```
21  4.98 out of 5 average rating, 295 reviews
22   4.93 out of 5 average rating, 43 reviews
23  4.94 out of 5 average rating, 103 reviews
24  4.96 out of 5 average rating, 311 reviews
25   4.92 out of 5 average rating, 48 reviews
26  4.99 out of 5 average rating, 247 reviews
27   4.98 out of 5 average rating, 64 reviews
28   4.69 out of 5 average rating, 81 reviews
29   4.88 out of 5 average rating, 40 reviews
30   4.73 out of 5 average rating, 59 reviews
31   4.93 out of 5 average rating, 44 reviews
32   4.97 out of 5 average rating, 78 reviews
33   4.97 out of 5 average rating, 33 reviews
34   4.97 out of 5 average rating, 33 reviews
35  4.89 out of 5 average rating, 105 reviews
36     5.0 out of 5 average rating, 8 reviews
37    5.0 out of 5 average rating, 25 reviews
38   4.96 out of 5 average rating, 72 reviews
39     5.0 out of 5 average rating, 7 reviews
40   4.96 out of 5 average rating, 68 reviews
41  4.97 out of 5 average rating, 200 reviews
42   4.95 out of 5 average rating, 40 reviews
43   4.95 out of 5 average rating, 21 reviews
44  4.97 out of 5 average rating, 143 reviews
45    4.67 out of 5 average rating, 3 reviews
46   4.9 out of 5 average rating, 131 reviews
```

In [7]:

```python
#Converted to CSV
df.to_csv('airbnb_listings.csv', index=False)
```

Observations from Output:

Several listings are missing critical information, causing the script to skip them. This could be due to inconsistencies in the page structure or dynamic content that hasn't loaded by the time of extraction.

Despite some skipped listings, the script successfully captures and stores the available data for the majority of the listings.

The output DataFrame presents a variety of information, including price discounts and listings without reviews, indicating "New place to stay."

The dataset compiled from the scraped data seems to include duplicates, as indicated by repeated URLs, which suggests that the script may benefit from additional logic to detect and handle such cases.

The ratings are detailed, providing both an average score and the number of reviews, offering insight into the listing's popularity and guest satisfaction.

In [ ]: