JT Ohlandt

Professor Li

B365-34395

12/14/2023

B365 Final Report

Overview

The goal of this project was to create a web scraper of Booking.com. This could be used in many applications including trip planning, competition research, and market analysis. First, I created a data scraper to scrape through a page of hotels. Next, I improved the data scraper by adding the ability to pull multiple pages and handle unrated hotels. Then, I made a data cleaner. Finally, I performed Linear Analysis on a sample dataset. The entire project was completed by me. I used Beautiful Soup to parse through the HTML, Pandas for storing the data frame and linear analysis and Chromedriver and Selenium to handle turning pages.

Part 1

The purpose of part 1 was to scrape a page of hotels for useful data. Utilizing Beautiful Soup's findAll method, I added all the property cards and iterated through them with a for loop. I added the attributes name, location, price, and rating for each hotel on the page. This was done by using the find method for each element, stripping the element and appending it to a list. I added this to a Panda's data frame and csv file called hotels.

Part 2

To grab a larger dataset my scraper would need to be able to scrape through multiple pages. This proved a challenge, as there was a JavaScript next button. To fix this, first I broke the program into 2 different functions get_hotel_data which operated like part 1 to scrape the data for a page. Next, I had the get_hotel_total method cycle through the pages. In each page it would call the get_hotel_data for each page url, updating that the list hotel with that pages data. The get_hotel_total method utilized selenium and ChromeDriver to calculate the total number of pages and simulate the page cycling.

In addition, an error occurred when a hotel lacked ratings as the scraper could not find the html segment. To fix this I added an if statement to return N/A when it was unable to get the rating.

Part 3

I created the DataCleaner file to clean the data. This allowed the data to be used in my future Linear Regression. First, I cleaned the ratings. I got rid of the hotels that had N/A ratings. Next, I converted the prices to floats. To achieve this I first had to remove the "$" and "," in some of the prices. Finally, I removed outliers. I used a IQR system to do this. Other methods could be used depending on the application of the data. I created a new csv file called "HotelsCleaned" for the new cleaned data.

Part 4

I chose an arbitrary date of January 31st 2024 and chose hotels with the NYC location for a large dataset. I then used matplotlib to run a linear regression between ratings and price. I then calculated the R squared value and the root mean squared error. I found R squared value was 0.234. That's not the best R squared value and the model could probably be improved. The RMSE value was $52. This meant the mean amount the regression line was off by was $52.