WEBSCRAPING PROJECT DESCRIPTION
HOTELS.NG

Magdalena Kowalewska 412860 | Anjola Olatunbosun 448005

Introduction

For our project we decided to scrape https://hotels.ng/. It is a Nigerian hotel booking website, resembling a worldwide known booking.com website. It was much easier to scrap in comparison to booking.com as there were less webscraping restrictions. This is also one of the reasons we decided to change our original website of choice, booking.com.

We focused on Lagos, Nigeria and scraped hotel names, standard room prices and overall rating. We scraped the details of the first 100 hotels from this website. Each page contained 10 links to different hotels webpages, so we scraped from the first 10 pages to get 100 hotels. After further extensions such as filtering using inputs in selenium and filtering of the .csv file, such a tool could be used to find best stay options for travelers. The purpose is to analyze the relationship that exists between the price of a hotel and the rating for each hotel. We visualized our data on a line graph to understand better the analysis of this data.

Scrapers

The scrapers are gathering the above information in various ways. For scrapy we created two spiders, the first one is extracting direct links to the hotels pages and then the second one is searching for the wanted information.

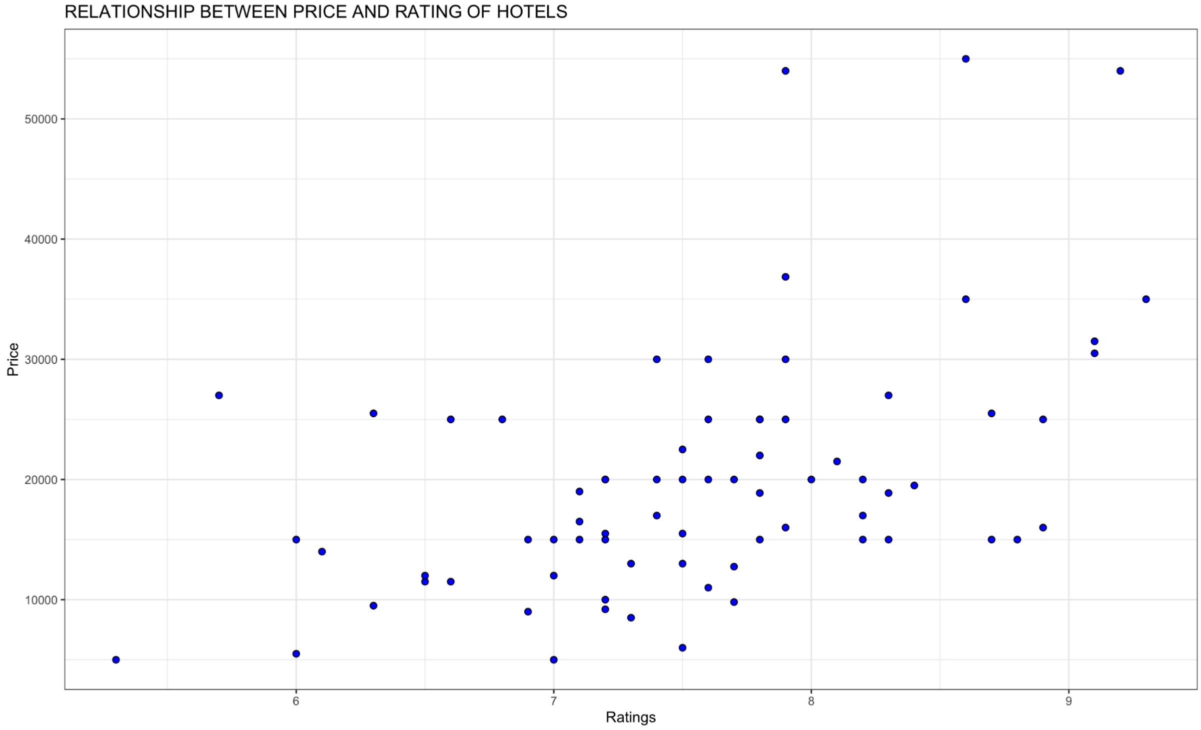
For Beautiful soup, we pulled our data out of an HTML file and created a parse tree from our page source code that we used to extract our data in a more readable manner, then exported to csv.

For Selenium we loaded our website, then we wrote a code to input our destination city-Lagos and check-in / check-out dates. After loading this input on our page, we scraped our data and exported it to csv.

Data Analysis

We will use the information from the extracted data for our analysis. For this, we will use the prices and the ratings. We will perform an analysis to check the correlation between the two variables, price and ratings. This is to see if there is a relationship between the prices of rooms and the customers ratings of these hotels.

NOTE:*Some of the hotels do not have ratings or prices because there is no availability for the assigned check in and check out, thus, we will get a few missing values. nevertheless our data visualization will only contain observations without missing values.*



To tidy up the dataset, we will remove all observations with NA's and outliers. We performed the Spearman correlation and we got the output below:

```
          price      rating
price  1.0000000 0.5093203
rating 0.5093203 1.0000000
```

Spearman coefficients measure the linear correlation between two variables X and Y. It has a value between +1 and −1. A value of +1 is a total positive linear correlation, 0 is no linear correlation, and −1 is a total negative linear correlation. From our output, we depicted a correlation of approximately 0.51 which indicates the variables, price and ratings can be considered **moderately correlated**. The interpretation of this basically means that an increase in the price of a hotel room does not imply a corresponding increase in the ratings of the hotel.