

Scraping Booking.com Hotels: Uncovering Insights

422 Final Project

Master of Science in Business Analytics

University of California, Davis



Priya Iddalgi

Yifan Zhu

Kangjian(James) Gao

Table of Contents

<i>Summary.....</i>	<i>2</i>
<i>Background Information & Business Problems.....</i>	<i>3</i>
<i>Web-Scraping Routine(s) & Database Design Choice</i>	<i>3</i>
<i>Data Cleaning & Potential Implementation</i>	<i>6</i>
<i>Conclusion.....</i>	<i>7</i>
<i>Appendix.....</i>	<i>7</i>

Summary

In this project, the team is targeting one central business problem, that is how hotel industry in the U.S should react to the booming post pandemic market demand by properly planning immediate action items. The primary data source for this project is collected through web-scraping Booking.com. Booking.com is one of the world's largest online travel agencies that offer a wide range of accommodation options for travelers, including hotels, apartments, villas, and hostels, in over 220 countries worldwide. During the web-scraping process, the team deployed the Selenium and Python Beautiful Soup to download and parse web pages. The final dataset include 14 columns and 1024 rows is then proceeded to MongoDB. The reason we choose MongoDB over SQL DB is mainly because MongoDB has a more flexible data structure and more scalable for future use. MongoDB do not require a fixed Schema, which make it easier for us to first scrape the desired data then proceed to define the exact data type for each feature. Also, consider the on-going data collection, MongoDB is more scalable than SQL DB. To make the scraped data ready for business use, the team also applied data cleaning include removing duplicate rows and fill nulls with median. Finally, the team suggested how the business could leverage the three action items to solve the business problems defined before.

Background Information & Business Problems

The COVID-19 pandemic has inflicted significant harm on various industries, but as we enter the post-pandemic era, many are experiencing a tremendous rebound - one of these being the hotel industry. According to Statista, there are 132,228 hotels and motels in the US as of 2022, which is 8.3% more than in 2021. As more and more countries remove their COVID policies, we expect this trend to continue go strong. Therefore, it is important for hotel business owners to catch this opportunity and properly plan for immediate action items to continue stand out in this increasingly competitive market.

To address this problem, we came up with three crucial actions items. Firstly, we want to understand the formulation of the overall rating score for each hotel listed in the Booking.com search page. More specifically, we want to find factors that potentially impact the overall rating score, and how businesses can leverage themselves through a better understanding of the composition of the overall rating score. Secondly, we want to investigate potential competitors for each hotel in all three cities listed above. We believe it is important for business owners to always know their potential competitors, especially where they are doing well, and we can adjust based on that. Finally, we want to come up with a list of criteria that can better help hotel business owners decide the price. The primary data source for this project is collected through web-scraping Booking.com. Booking.com is one of the world's largest online travel agencies that offer a wide range of accommodation options for travelers, including hotels, apartments, villas, and hostels, in over 220 countries worldwide. The platform has more than 28 million listings and attracts millions of visitors daily. This makes Booking.com a valuable source to collect latest data in the hotel industry.

Web-Scraping Routine(s) & Database Design Choice

We used web-scraping to collect all the data for this project. To make the extracted data less sensitive to bias, we randomly picked a date time range between May-30-2023 to May-31-2023 and scraped the lowest price available in each hotel for a typical one-person accommodation. We web-scraped a total of thirteen features for this project (price, miles to center, address, facilities rating, free WiFi rating, label, location

rating, name, number of reviews, overall rating, staff rating, city, room type) as showed in *Figure. 1* and *Figure. 2* in the appendix. Our approach to scraping all these data is to first download all eighteen pages of hotel searching results using Selenium. The reason we choose Selenium over Python Beautiful Soup is that we want to automate the downloading process, so there is no need to manually copy all eighteen search page URLs. To keep updating the browser landing page, we decomposed the searching page URL to increase the offset value by 25 each time using a for loop, which suggests the next page. The Selenium then controls the cursor to click on the next page button and finish downloading the HTML file.

After having all eighteen HTML files downloaded to the local hard drive, we then start extracting miles to center, room type, and price fields from each hotel information box using Python Beautiful Soup (*Figure. 1*). We also parse and stored the linked URL in order to extract all the other fields located in the hotel detailed page (*Figure. 2*). All the field information was then stored in MongoDB. The reason we choose MongoDB over SQL DB is mainly due to a more flexible data structure and scalability. In MongoDB, data is stored in collections, which contain documents. A document is a JSON-like data structure that can contain fields with values of different data types. This allows for a highly flexible data model, where you don't have to define a fixed schema beforehand. Furthermore, if you want to update or insert new data into an existing collection using Python, you can simply do that by defining a start document ID along with a filter query specifying that ID as we did in this project. In contrast, SQL requires a fixed schema. Also, if you want to add a new field or change a data type, you need to modify the schema and update all existing records accordingly, which can be time-consuming and inefficient. Another advantage of MongoDB is scalability. MongoDB can automatically distribute data across multiple nodes in a cluster, allowing it to handle large volumes of data and traffic. In our case, we collected hotel information from only three cities in the U.S.: New York, Chicago, and Los Angeles. However, we do see the potential of data incrementation for future business use. Therefore, we want to build an initial set-up that is scalable in the future, which makes us choose MongoDB over SQL DB.

We believe all thirteen features extracted above are meaningful in answering the business problem we defined above, and there are four major reasons. Firstly, pricing strategy. By analyzing the hotel prices and room types, businesses can develop a competitive pricing strategy to attract more customers. For instance, if the data shows that hotels with a certain room type are in high demand, businesses can set a higher price for these rooms. On the other hand, if the data shows that there is a low demand for a particular room type, businesses can offer discounts to attract more customers. By analyzing the pricing trends of their competitors, businesses can also adjust their prices to remain competitive and maintain profitability.

Secondly, Marketing and Advertising. The miles-to-center data can be used to create marketing campaigns that highlight the hotel's proximity to popular tourist destinations, such as museums, theme parks, or shopping centers. Businesses can also use the data to target specific customer segments that prioritize the hotel's location. For instance, a hotel located close to a business district can target business travelers by promoting its proximity to the central business district. By tailoring their marketing messages to the right customer segments, businesses can increase their brand awareness and attract more customers.

Thirdly, quality improvement. By examining the facilities rating, free WiFi rating, overall rating, and staff rating, hotel management can identify areas of improvement to enhance the overall customer experience. For instance, if the facilities rating for a hotel is low, management can investigate the reasons for the low rating and take steps to improve the hotel's amenities. Similarly, if the staff rating is low, management can provide training to staff members to improve their customer service skills. By addressing the areas of improvement identified through the data, businesses can improve the overall customer experience and increase customer loyalty.

Finally, competitive analysis. By analyzing the data on the number of reviews and the overall rating, hotels can benchmark themselves against their competitors and identify areas where they can outperform them. For instance, if a competitor has a higher overall rating than the business, management can examine the reasons for the higher rating and take steps to improve their hotel's ratings. Similarly, if a competitor has a

higher number of reviews, businesses can examine their marketing strategy and identify ways to increase their hotel's visibility and attract more reviews. By staying on top of their competition, businesses can remain relevant in the market and attract more customers.

Data Cleaning & Potential Implementation

Data cleaning is an essential step in the data analysis process. It involves identifying and correcting errors, inconsistencies, and incomplete information in the dataset to make it usable for analysis. In this section, we will go through the different steps taken to clean our hotel booking dataset scraped. Furthermore, we will talk about how the cleaned dataset can help provide value to businesses. After importing the dataset from the csv file, we check the high-level data to determine the number of columns and rows and data types of each column. We then check the statistical summary of the dataset to get a better understanding of the numerical features' distribution. In order to ensure we have unique hotels in our dataset, we removed duplicated based on the address column, keeping only the first (lowest priced) occurrence for each address. This step resulted in 641 unique entries compared to 729 entries earlier. Additionally, we replaced the missing values with the median (for numerical columns) and mode (for categorical columns) to ensure data completeness. Another step taken in data cleaning was using regular expressions and lambda functions to extract room type information, and provide clear labels for 'Entire Studio', 'Private Suite' or 'Basic Room'. To maintain data integrity, we standardized naming conventions. Finally, we saved the cleaned dataset to a CSV file to be used for analysis to solve our business problem.

Data cleaning is an essential step in data analysis that can help businesses derive useful insights from their datasets. By having the cleaned dataset, we can now think about how to achieve the three action items we identified before. Since we already have the overall rating as one of the features, the company can perform linear regression to address the first action item. Since we have features that covering different aspects of the hotel performance, the business can perform clustering method for the second action item. Finally, for the last action item, the company could potentially run a tree model. In conclusion, the

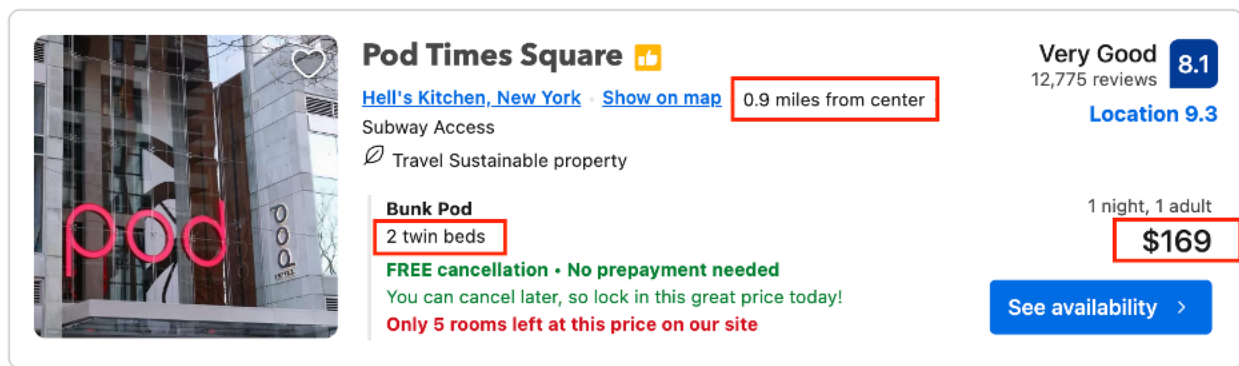
businesses can use this cleaned dataset to create targeted marketing campaigns and improve their advertising efforts to attract more customers. It can also provide insights into market trends, competitive analysis, and strategic planning. The cleaned data will make it easy to analyze this data and provide fair statistical and machine learning results.

Conclusion

In this project, we applied web scraping skills to parse hotel information from Booking.com for three major cities in the US and storing it in MongoDB, which allows for a highly flexible data model and scalability. By having this cleaned data ready for business use, we believe all three action items can be properly conducted and the business value can be realized.

Appendix

Figure. 1



Pod Times Square

[Hell's Kitchen, New York](#) - [Show on map](#) 0.9 miles from center

Subway Access
Travel Sustainable property

Bunk Pod
2 twin beds

FREE cancellation • No prepayment needed
You can cancel later, so lock in this great price today!
Only 5 rooms left at this price on our site

Very Good 8.1
12,775 reviews
Location 9.3

1 night, 1 adult
\$169

[See availability >](#)

Figure. 2



8.1 Very Good 12,910 reviews [Read all reviews](#)

Categories: [Show details](#)

Staff	8.4	Facilities	8.0	Cleanliness	8.2
Comfort	8.2	Value for money	7.9	Location	9.3
Free WiFi	8.6				