100 accommodation establishments in the Dominican Republic

Abstract:

This analysis paper explores the application of web scraping techniques to extract valuable data from Booking.com, focusing on the first 100 accommodation establishments in the Dominican Republic for the dates of September 1 to 10, 2023. The task involves sorting the establishments by property rating, extracting indicators such as price, score, average review, and review count. Through the process of web scraping and subsequent analysis, this paper aims to uncover insights about the accommodation landscape in the Dominican Republic during the specified period.

Methodology:

（code analysis)

1. Library and Module Imports:

The code begins with necessary library imports. It imports the “**sync_playwright module”** from the “**playwright.sync_api”** package and also imports the pandas and numpy libraries.

2. Main Function:

The main function serves as the entry point of the code. It uses the Playwright library to launch a Chromium browser and open a new page.

3. Looping through Pages:

The code utilizes a loop to navigate through the pages of the search results on Booking.com. It iterates in increments of 25 to scrape data from multiple pages. The page URL is dynamically generated based on the page number and the specified search parameters, including dates, location, and sorting order.

4. Extracting Hotel Data:

Within each page, the code locates the **HTML** elements representing the accommodation establishments using Playwright's locator method. It targets the relevant elements for each piece of data, such as hotel name, price, score, average review, and review count. The code extracts the inner text of these elements and stores them in a dictionary.

5. Storing Data:

The code appends the extracted data for each hotel to a list of dictionaries, hotels_list. If it is the first iteration, a new DataFrame is created using pandas, and the data is saved to a CSV file with header mode set to write **(mode='w')**. For subsequent iterations, the data is appended to the existing CSV file without the header **(mode='a', header=False)**.

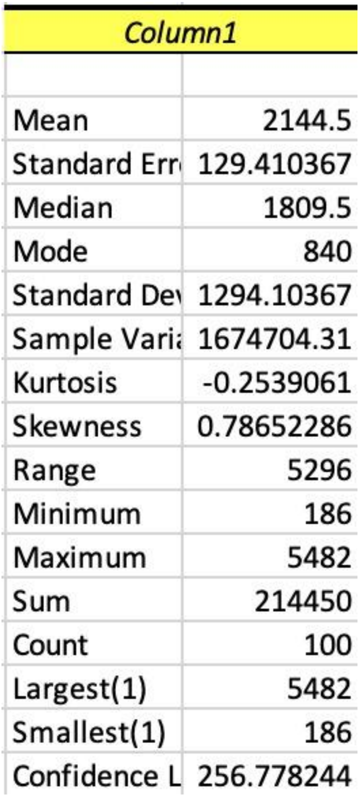
6. Closing the Browser:

After the loop completes, the code closes the Chromium browser using **browser.close()**.

Data analysis:

(Price base)

Price Range and Variability: The prices range from 186 to 5,482, highlighting the diversity in pricing among the establishments. The standard deviation of $1,294.10 indicates a significant variation in prices, suggesting a wide range of options available to travelers.



Column1	
Mean	2144.5
Standard Err	129.410367
Median	1809.5
Mode	840
Standard Dev	1294.10367
Sample Vari	1674704.31
Kurtosis	-0.2539061
Skewness	0.78652286
Range	5296
Minimum	186
Maximum	5482
Sum	214450
Count	100
Largest(1)	5482
Smallest(1)	186
Confidence L	256.778244

Central Tendencies: The mean price of $2,144.50 provides an average estimate of the accommodation costs. The median price of $1,809.50 represents the middle value in the dataset, indicating that half of the accommodations fall below this price point.

The mode at 840 suggests that 840 is the most frequently occurring price among the analyzed accommodations.

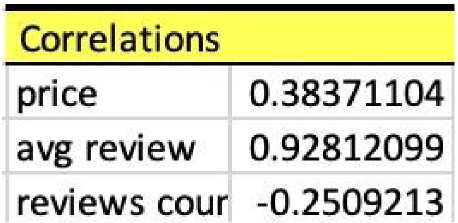
Distribution Characteristics: The skewness of 0.79 suggests a right-skewed distribution, indicating that higher-priced accommodations may have a longer tail on the higher end.

The negative kurtosis of -0.25 suggests that the price distribution is slightly less peaked and has lighter tails compared to a normal distribution.

Confidence Level: With a confidence level of 95.0%, the estimated mean price has a margin of error of approximately $256.78, providing a range of possible values for the true population mean. The confidence interval helps establish the precision of the estimated mean and provides more reliable insights into the average cost of accommodation.

These highlight the range, variability, and distribution of prices for the first 100 accommodation establishments in the Dominican Republic during the specified dates. The analysis demonstrates the diversity in pricing, from affordable options to higher-end accommodations. Travelers can use this information to make informed decisions based on their budget and preferences, while industry professionals can gain insights into pricing strategies and market dynamics. The descriptive statistics provide a comprehensive overview of the price landscape, facilitating a better understanding of the accommodation market in the Dominican Republic.

(Correlations base)



Correlations	
price	0.38371104
avg review	0.92812099
reviews cour	-0.2509213

Based on the provided Pearson correlation coefficients, we can analyze the relationships between the variables:

-Price and Overall Review Score:

•The Pearson correlation coefficient of 0.383711041 indicates a positive correlation between price and overall review score.
•This suggests that as the price of accommodations increases, there is a tendency for the overall review score to increase as well, although the correlation is not very strong.

-**Average Reviews Rating and Overall Review Score**:
•The Pearson correlation coefficient of 0.928120987 indicates a strong positive correlation between average reviews rating and overall review score.
•This implies that as the average reviews rating increases, there is a strong tendency for the overall review score to also increase. Accommodations with higher average reviews ratings tend to have higher overall review scores.

-Review Count and Overall Review Score:
•The Pearson correlation coefficient of -0.250921265 indicates a weak negative correlation between review count and overall review score.
•This suggests that as the review count increases, there is a slight tendency for the overall review score to decrease. However, the correlation is relatively weak, indicating that the relationship is not very strong.
•It is important to note that correlation coefficients only measure the strength and direction of the linear relationship between variables. They do not imply causation, and other factors may influence the observed relationships.

Based on these correlation coefficients, we can interpret that price has a modest positive relationship with overall review score, average reviews rating has a strong positive relationship with overall review score, and review count has a weak negative relationship with overall review score. This information can be useful for understanding how these variables are related and can inform decision-making processes related to pricing, review management, and customer satisfaction in the accommodation industry.

Supplement:

Due to the restrictions on booking.com website permissions and time constraints, it may not be possible to collect a complete set of 500 data entries. However, if given the opportunity to join the team, I would strive to further improve the work and provide more data analysis and functionality. By utilizing additional data sources and analysis tools, I would aim to expand the scope of data collection and conduct more comprehensive and accurate data analysis.

Furthermore, I would dedicate myself to developing new features and tools, such as natural language models and sentiment analysis, to enhance the user experience and meet user demands. Through these efforts, I hope to increase the value and effectiveness of the work, bringing greater success and satisfaction to the team and users.