JBI010 Assignment 2

Abdullah Alhariri (2184796), Rafi Azmani (2158256)

1 Introduction

When considering booking accommodation, there is a sea of options. Luckily, many people leave reviews after their stay. We created a tool to analyse these reviews and the hotels providing us with a clear image which hotels are great and which not. With the help of a dataset consisting of 515 738 reviews and 1 493 we can deliver strong and concrete advice and information with enough argumentation.

2 AI Statement

AI hasn't been used in any shape or form for this assignment.

3 Summary statistics

We have done exploratory data analysis on reviews, of which we where able to gather important information about the reviews posted. For every field (attribute) of review, we have gathered negative words, positive words, scores and days since the post of the review. On the data we have gathered, we performed exploratory data analysis and got the following results.
Negative words:

- Mean: 18.54
- Median: 9
- Standard deviation 29.69

Positive words:

- Mean: 17.78
- Median: 11
- Standard deviation 21.8

Score:

- Mean: 8.4
- Median: 8.8
- Standard deviation 1.64

Days since the review:

- Mean: 354.44
- Median: 353.0
- Standard deviation 208.93

4 Own analysis

Doing our own analysis we found out that Austrian hotels are doing the best in terms of scores, which greatly explains why there's none Austrian hotels in the bottom 10 hotels, and 4 hotels in the top 30; all of that while just having 157 hotels in total. From the pie chart below (Figure 1) we see that Austria doesn't have the lowest amount of hotels, thus being credible enough for it delivering the best hotel review scores.

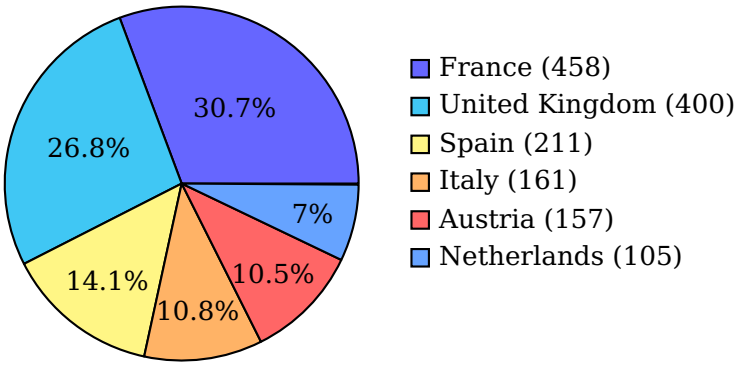

Figure 1: Pie chart of the contribution of each country. Number of hotels inside brackets.

Analysing this more we see that there's in total 38 655 reviews on Austrian hotels with over 169 nationalities. On the pie chart (Figure 2) we see that UK and USA are the major reviewers of Austrian hotels.

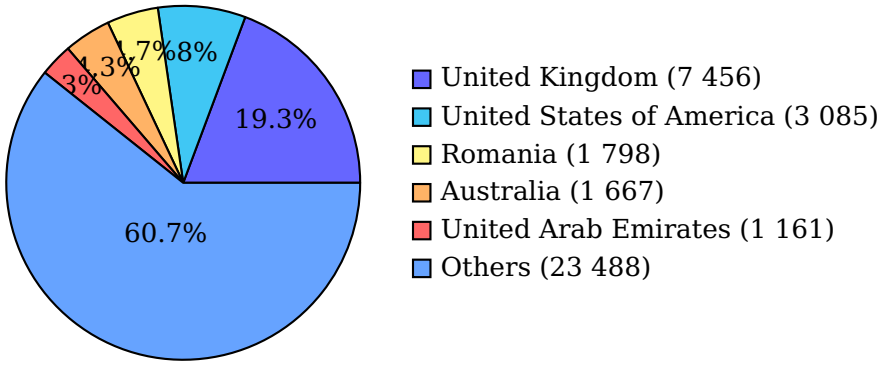

Figure 2: Pie chart of the review nationalities gathered on Austrian hotels. Number of reviews inside brackets.

Further looking at those two nationalities we see that they generally give a higher scores than the average score. The general average score is 8.24 and 8.49, 8.79 for UK and USA respectively, this could be the reason why Austria scores better than other countries.

5 Score improvement

We decided to create two different analysis to strengthen arguments and thus booking.com can have more (concrete) feedback's. For the first analysis we asked ourselves: "What is the average score of each country?", and for that we found the following answer:

- Austria: 8.6
- France: 8.5
- Spain: 8.5
- United Kingdom: 8.5
- Netherlands: 8.4
- Italy: 8.3

To answer this question we had to create a function with hotels as the input and expecting the function to return the mean score of each country based on the reviews posted. For the next analyse we asked ourselves: "What nationalities give the highest scores?", and surprisingly there are a couple of countries that have always given a 10, they are:

- Crimea
- Equatorial Guinea
- Comoros

- Svalbard Jan Mayen

The way we did answer this question is very similar to the first one, we made a function that takes in hotels and determines the nationalities that gives the highest scores by searching in the reviews of each hotel.

6 The main finding and conclusions

We found out that there are in total 1 492 hotels and 515 738 reviews in the dataset, of which 30 hotels scoring 10 and thus being in the top, on the other hand there are 4 hotels scoring below score 6. On top of that we found out that 16 out of 30 top hotels and 7 out of 10 bottom hotels have been decreasing, this gives the information that 53% of the top hotels and 70% of the bottom hotels haven been decreasing.
Based on our previous findings we can conclude that Austrian hotels are doing the best in terms of scores, which explains why there's none Austrian hotels in the bottom 10 and 4 hotels in the top 30 while just having 157 hotels in total; see the pie chart (Figure 1)

7 Your advice

Based on the analyzing of Austria review nationalities, we can say that attracting nationalities which generally give score higher than the mean (8.24) can result into boosting your hotel score above others; try to attract tourist from e.g. Puerto Rico, El Salvador, Congo, Australia, United Kingdom, United States of America and Canada because generally they all give on average a score higher than 9.1. On the other hand it may be a good idea to communicate or repel the tourist that give below average score of the nationalities e.g. Montserrat, American Samoa and Northern Mariana Islands.

8 Implementation

The work done on this assignment has been only in hotel_types.py in the correct boxes as instructed. The design choices have been taken from PEP 8 as being the conventions; for doc strings we decided to continue using the reST convention.
We have created 4 functions as instructed, therefor we are documenting each input and output:

- average_score_per_country
  - Input: Booking
  - Output: {'Netherlands': 7.7, 'United Kingdom': 8.47} i.e. dict[str, float]
- highest_nationalities_average_score
  - Input: Booking
  - Output: highest nationalities in a list as a string

- check_improvements
  - Input: Booking
  - Output: {'Hotel Arena': (7.904.., 7.767.., False), ...}
- get_top_and_bottom
  - Input: hotels: list[Hotel] (list of hotels), sorting_order: bool (i.g. False)
  - Output: A list containing both hotel name and score as a tuple. i.e. list[tuple[str, float]]