Data visualization - Milestone 1

Seasonal trends in hotel reviews

Hilda Abigél Horváth, Adrien Schurger-Foy, Julian Schnitzler

April 8, 2022

Dataset

We use a dataset from Kaggle called 515K Hotel Reviews Data in Europe[1], since we wanted to work with a dataset that contains both location and date data. This dataset contains hotel costumer reviews for European luxury hotels originated from booking.com.

In the dataset there is data about the reviews such as the number of words in the positive review or the reviewer's score for the hotel, as well as data about the hotel, for example its location, name . Moreover, the reviewers' nationality can also be found in the dataset. The more detailed description of the data in the dataset can be found below in the Table 1.

Name	Description	Example
Hotel_Address	Address of the hotel	Gravesandestraat 55 Oost 1092 AA Amsterdam Netherlands
Additional_Number_of_Scoring	Number of scores without reviews for the hotel	194
Review_Date	The date of the review	7/24/2017
Average_Score	The average score for the hotel at the review date	7.7
Hotel_Name	The name of the hotel	Hotel Arena
Reviewer_Nationality	The country the reviewer comes from	Poland
Negative_Review	The negative review of the reviewer	Backyard of the hotel is total mess shouldn t happen in hotel with 4 stars
Review_Total_Negative_Word_Counts'	The number of words in the negative review	17
Total_Number_of_Reviews'	The total number reviews the hotel has	1403
Positive_Review	The positive review of the reviewer	Only the park outside of the hotel was beautiful
Review_Total_Positive_Word_Counts	The number of words in the positive review	20
Total_Number_of_Reviews_Reviewer_Has_Given	Total number of reviews reviewer has given	1
Reviewer_Score	Reviewers' score for the hotel	6.7
Tags	Reviewers' tags for the hotel	' Leisure trip ', ' Group ', ' Duplex Double Room ', ' Stayed 1 night '
days_since_review	Duration between the review and scrape date	10 days
lat	Latitude for the hotel location	52.3605759
lng	Longitude for the hotel location	4.9159683

Table 1: Data featured in the dataset

The dataset is very clean and does not require much preprocessing. One main part of preprocessing is to deal with missing values. When a user decided to not write a positive review, the value of this cell is simply 'No positive'. Similar for negative reviews, with 'No negative'. We filter those values out and replace them by None, to make sure that we do not confuse our visualizations with these non-user generated words. But, we do not need to change the Review_Total_Positive_Word_Counts and Review_Total_Negative_Word_Counts, since those are already 0 for the respective rows. Also, for some of the hotels the latitude and longitude values are missing. By carefully analyzing when that happens, we assume that the preprocessing removed some non-unicode characters from some of the addresses, and that then lead to missing values in the geocoordinates since probably some lookup failed. Since it only occured for about 20 of the over 1000 hotels, we decided to fix those addresses manually and search for the respective coordinates by entering the adresses into latlong.net.

Even though the authors of the dataset report that they dealt with preprocessing, we found out that there are still upper- and lowercase words in the reviews. We lowercased all and removed stopwords, and saved the new Reviews in the columns Positive_Review_clean and Negative_Review_clean.

[1] https://www.kaggle.com/datasets/jiashenliu/515k-hotel-reviews-data-in-europe

	Additional_Number_of_Scoring	Average_Score	Review_Total_Negative_Word_Counts	Total_Number_of_Reviews
count	515738.000000	515738.000000	515738.000000	515738.000000
mean	498.081836	8.397487	18.539450	2743.743944
std	500.538467	0.548048	29.690831	2317.464868
min	1.000000	5.200000	0.000000	43.000000
25%	169.000000	8.100000	2.000000	1161.000000
50%	341.000000	8.400000	9.000000	2134.000000
75%	660.000000	8.800000	23.000000	3613.000000
max	2682.000000	9.800000	408.000000	16670.000000

	Review_Total_Positive_Word_Counts	Total_Number_of_Reviews_Reviewer_Has_Given	Reviewer_Score
count	515738.000000	515738.000000	515738.000000
mean	17.776458	7.166001	8.395077
std	21.804185	11.040228	1.637856
min	0.000000	1.000000	2.500000
25%	5.000000	1.000000	7.500000
50%	11.000000	3.000000	8.800000
75%	22.000000	8.000000	9.600000
max	395.000000	355.000000	10.000000

Table 2: Overview of relevant continuous features

Problematic

With our visualization, we would like to see how reviews about hotels change over time and location. By that, we aim to give people planning their vacation an easy to grasp way to find hotels that receive a lot of negative feedback in certain periods, or more general find cities and regions where hotels receive remarkable positive or negative feedback in a certain time period. It can also allow travellers to quickly check average reviews in a certain area, and more importantly identify hotels with significantly poor reviews in some time. From the number of reviews in a certain period,

Moreover, using clouds of words of the most frequent words in the positive and negative reviews (excluding stopwords), users can quickly identify points of criticism and commendation among a specific hotel, and evaluate whether or not those points are of concern or are so severe that they might want to consider another hotel.

We can also show for each hotel the nationalities of the reviewers. Then, we could draw an arrow from each country to this hotel, that appears whenever the user hovers over or clicks on the dot of the hotel on the map. Assuming that the likelihood of reviewing the hotel is not influenced by the country that the reviewer is from, that could help users to roughly estimate the distribution of countries that the guests are from. Even if that assumption does not hold, it nicely shows which places the most reviewers of a specific hotel come from.

Exploratory Data Analysis

This dataset contains 515,000 customer reviews and ratings of 1493 luxury hotels from 6 metropolises across Europe, from August 2015 to August 2017. (Table 3) For an overview of the distribution of the numerical features in the dataset, we refer to Fig. 1.

We first have a look at the distribution of the number of reviews (that appear in the dataset, i.e. not the number reported in Total_Number_of_Reviews) per hotel in the dataset. The hotels with the lowest number of reviews have less than 10, while the highest reach over 2000 reviews. (Fig. 2).

It seems that most positive reviews feature the staff or the location of the hotel. Negative reviews often mention the room. (Fig. 3)

There is a peak in the number of reviews in July and August 2016, reaching until over 25000 reviews in August 2016. The other months seem to be fairly evenly distributed on a level of slightly below 20000 reviews per month. (Fig. 4)

Column	no. of unique values
Hotel_Address	1493
Additional_Number_of_Scoring	480
Review_Date	731
Average_Score	34
Hotel_Name	1492
Reviewer_Nationality	227
Negative_Review	330010
Review_Total_Negative_Word_Counts	402
Total_Number_of_Reviews	1142
Positive_Review	412600
Review_Total_Positive_Word_Counts	365
Total_Number_of_Reviews_Reviewer_Has_Given	198
Reviewer_Score	37
Tags	55242
days_since_review	731
lat	1489
lng	1489
Positive_Review_clean	383363
Negative_Review_clean	310002

Table 3: Number of unique values per column

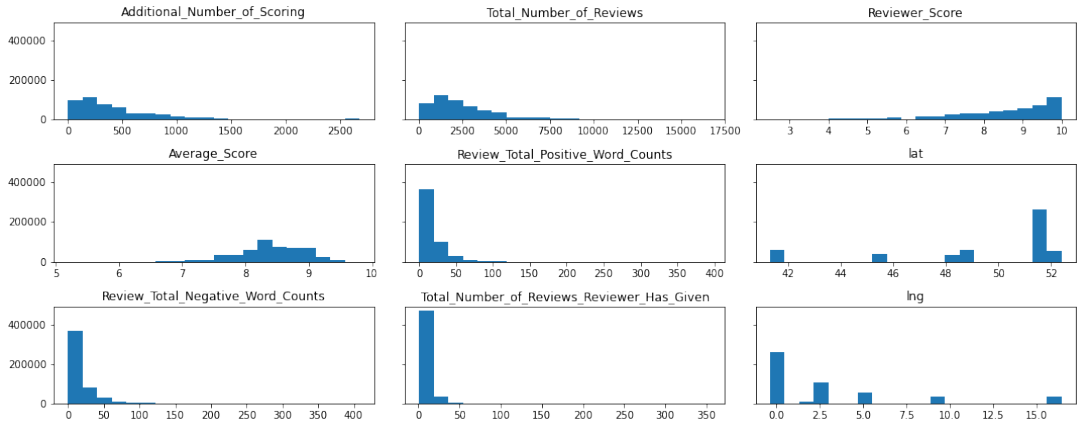

Figure 1: Histograms of numerical features

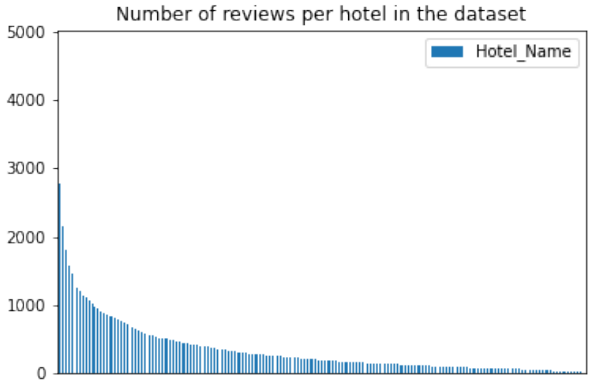

Figure 2: Number of reviews in the dataset for each hotel

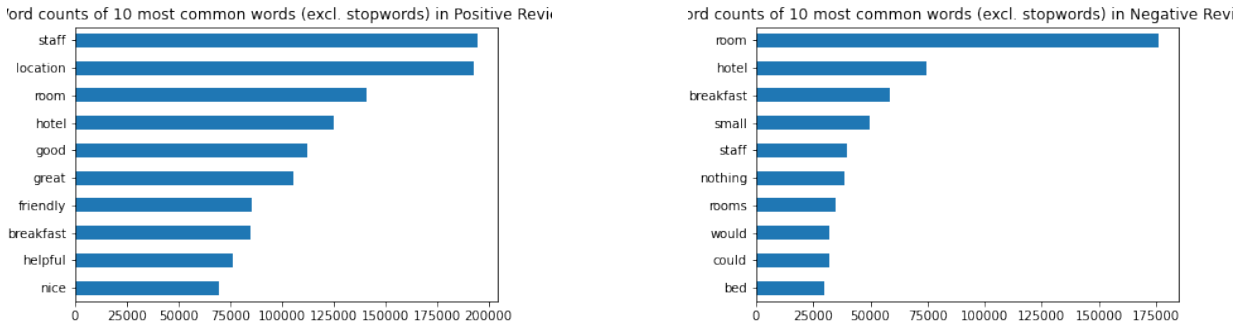

Figure 3: Most frequent words in positive and negative reviews

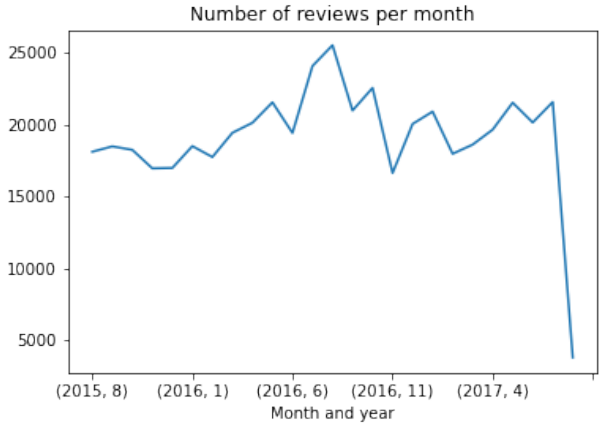

Figure 4: Number of reviews per month

Related work

Directly on Kaggle, it is possible to find previous work on this dataset in several shared notebooks. This work focuses mainly on sentiment analysis and its relationship with user ratings, since positive and negative points are mentioned in separate fields. One of the notebooks that got the most upvotes provides a quick visualization of the dataset, however it does not show nicely which hotel got which reviews, and it also leaves the time component out. Also, their preprocessing, especially stopword removal, seems to little. [2]

We also found many research papers that use this dataset. For example the authors of this paper use the dataset to train a LSTM-based model to classify hotel reviews into positive and negative. In another paper based on this dataset, the authors build a recommender system based on the cosine similarity between the tf-idf vectors of several hotel reviews. Our main source of inspiration are interactive maps containing information about the location of a hotel/apartment, together with additional metadata like the prices, or the score, e.g. on Airbnb (Fig. 5).

We also found a lot of wordclouds online, which inspired us to do something similar with the text of the reviews. For example, the wordcloud library (`https://github.com/amueller/word_cloud`) allows for creating those easily in python.

[2] https://www.kaggle.com/code/jiashenliu/quick-visualization-of-data

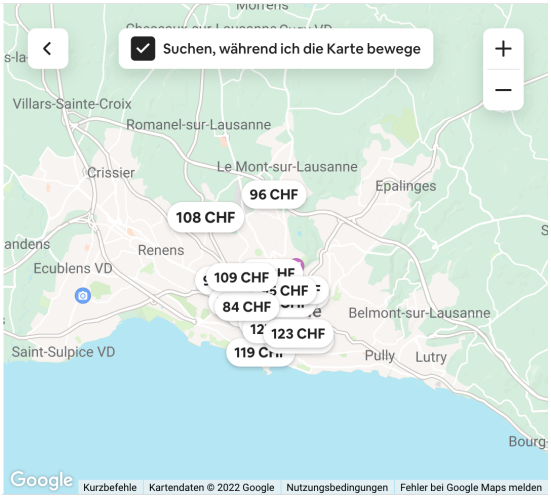

Figure 5: Airbnb Map to choose the location and price

Our approach differs in that we seek to identify seasonal and location-based trends in user ratings over time, not just presenting a snapshot of the current situation, like for example airbnb does in showing the current prices.
None of us ever worked with this dataset before.