MA415 Final Project Deliverable #1

Xiao Lu, Yange Bian, Luwen Mai, Yuyang Li

Background and Motivation

Online reviews are becoming increasingly important, especially for hotels. When booking accomodations, travelers are paying more attention on online reviews about the hotels. This special relationship between hotels and their online reviews motivates our study. Since cultural differences may affect travelers' opinions on what the important features of hotels are, we would like to look at whether the travelers' nationalities have any impacts on their preferences in different hotel features. The dataset we use contains information on online reviews for hotels in Europe, which is one of the most popular travel destinations around the world. By using data from one specific continent, Europe, we are able to generate a focused, coherent and detail-oriented analysis.

Mini Abstract

This project deliverable performs an preliminary analysis of online reviews for hotels in Europe. We analyzed and visualized the multiple relationships between the top 20 travelers' nationalities and their online reviews. Based on their nationalities, we examined the travelers' positive and negative reviews both qualitatively and quantitatively, and compared the average review scores of the hotels from last year with that of this year. In addition, we also briefly analyzed the relationship between the countries where hotels locate and the corresponding review scores based on the travelers' nationalities.

Table of Content

Data Description

The dataset is originally adapted from *kaggle.com*, and the data were collected from and owned by *Booking.com*. Some of the variables are continuous, such as Review_Total_Negative_Word_Counts, Review_Total_Positive_Word_Counts and Reviewer_Score, While other variables are categorical, such as Reviewer_Nationality, Negative_Review and Positive_Review.

The reference link of the dataset is as follows: https://www.kaggle.com/jiashenliu/515k-hotel-reviews-data-in-europe. The raw dataset contains 515,000 hotel reviews and scoring of 1493 luxury hotels across Europe. Below are the variable descriptions of all 17 variables in the dataset:

- Hotel_Address: Address of hotel.
- Review_Date: Date when reviewer posted the corresponding review.
- Average_Score: Average Score of the hotel, calculated based on the latest comment in the last year.
- Hotel_Name: Name of Hotel
- Reviewer_Nationality: Nationality of Reviewer
- Negative_Review: Negative Review the reviewer gave to the hotel. If the reviewer does not give the negative review, then it should be: 'No Negative'
- Review_Total_Negative_Word_Counts: Total number of words in the negative review.
- Positive_Review: Positive Review the reviewer gave to the hotel. If the reviewer does not give the negative review, then it should be: 'No Positive'
- Review_Total_Positive_Word_Counts: Total number of words in the positive review.
- Reviewer_Score: Score the reviewer has given to the hotel, based on his/her experience
- Total_Number_of_Reviews_Reviewer_Has_Given: Number of Reviews the reviewers has given in the past.
- Total_Number_of_Reviews: Total number of valid reviews the hotel has.
- Tags: Tags reviewer gave the hotel.
- days_since_review: Duration between the review date and scrape date.
- Additional_Number_of_Scoring: There are also some guests who just made a scoring on the service rather than a review. This number indicates how many valid scores without review in there.
- lat: Latitude of the hotel
- lng: longitude of the hotel
- Hotel_Country: the country in which the hotel is located. (Created by selecting the last word of the Hotel_Address)

Research Question

Do travelers from different countries have different preferences in hotel features?

How does the occurrence frequencies of keywords depend on travelers' nationalities? [1]

Response Variable: Reviewer_Score

Categorical Covariates: Reviewer_Nationality, Hotel_Country, Positive_Review, Negative_Review,

Continuous Covariates: Average_Score, Review_Total_Positive_Word_Counts, Review_Total_Negative_Word_Counts, frequencies of keywords*

Data Import & Cleaning

Fortunately, there is no significant problem when importing data. There is no missing, unique or unusual values. The only thing worth mentioning is that we originally defined column type of 'Review_Date' as col_date(). However, R studio informs us that the actual type of this column is col_character().

[1] *The frequencies of keywords will be created and analyzed in project deliverable 2.*

Variation of Single Variables

- Top 20 countries in terms of the number of travelers

Reviewer_Nationality <chr>	num_tourist_by_nation <int>
United Kingdom	245246
United States of America	35437
Australia	21686
Ireland	14827
United Arab Emirates	10235
Saudi Arabia	8951
Netherlands	8772
Switzerland	8678
Germany	7941
Canada	7894
France	7296
Israel	6610
Italy	6114
Belgium	6031
Turkey	5444
Kuwait	4920
Spain	4737
Romania	4552
Russia	3900
South Africa	3821

From the above table, we see that United Kingdom has the highest number of tourists. Most of the top 20 countries where travelers come from are developed countries. This result makes sense since people from developed countries generally have higher disposable income, and that enables them to travel around the world.

- The ranking, in descending order, of average review scores from the Top 20 tourist sources of countries.

Reviewer_Nationality <chr>	Top_AvSC <dbl>
United States of America	8.761474
Australia	8.635575
Canada	8.624518
Israel	8.608940
United Kingdom	8.470613
Ireland	8.458005
South Africa	8.443367
Romania	8.272174
Russia	8.268000
Germany	8.228333
Belgium	8.219113
Switzerland	8.218881
Spain	8.180088
France	8.163260
Netherlands	8.136842
Italy	8.106769
Turkey	8.029091
United Arab Emirates	8.008398
Saudi Arabia	7.949787
Kuwait	7.730705

The United States of America has the highest average review scores of around 8.76. So far, we don't see a geographical or socio-economic pattern associated with average review scores.

- The ranking, in descending order, of last year's average review scores from the Top 20 tourist sources of countries, calculated based on the latest comment in the last year.

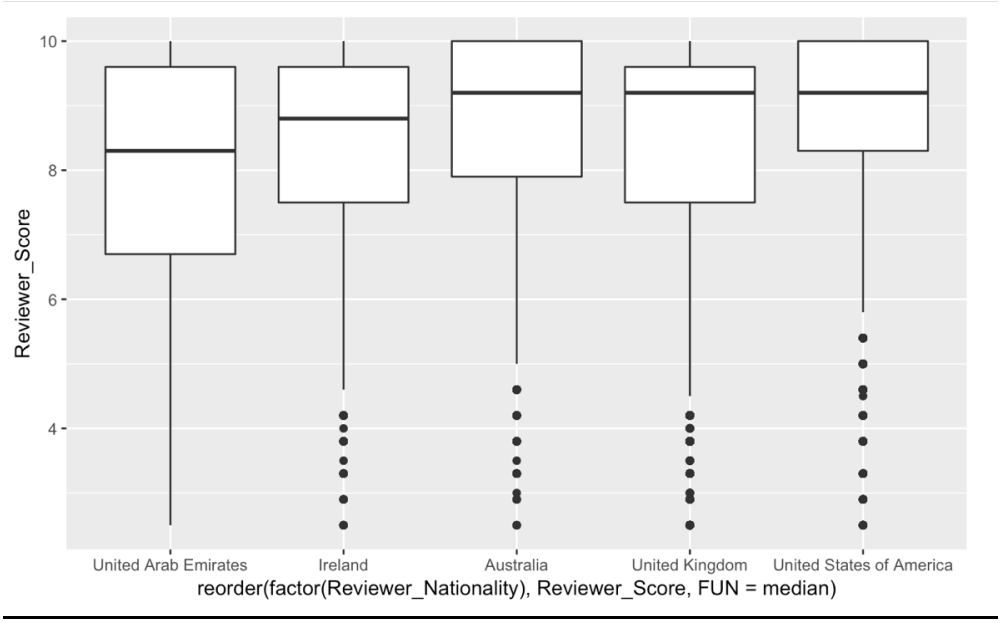
Reviewer_Nationality <chr>	AvSc <dbl>
United States of America	8.536783
Israel	8.529139
Switzerland	8.490210
Australia	8.475310
Belgium	8.465529
Canada	8.451791
South Africa	8.444388
Russia	8.431000
United Arab Emirates	8.415234
Saudi Arabia	8.406170

Reviewer_Nationality <chr>	AvSc <dbl>
Spain	8.387611
Netherlands	8.385088
United Kingdom	8.377530
Germany	8.360556
France	8.357182
Turkey	8.351273
Romania	8.347391
Italy	8.344923
Kuwait	8.339004
Ireland	8.333858

The United States of America has the highest average last year's review score of 8.53, while Ireland has the lowest score of 8.33 among the top 20 countries. Compared with the previous table about the average review scores from the Top 20 tourist sources of countries, Ireland has descended on the list.

Covariation between Multiple Variables

- Boxplot for the 'Reviewer_Score' for Top 5 sources of tourists countries, ordered by median value



The boxplot shows that Australia has the highest median value of Reviewer_score while United Arab Emirates has the lowest median value among the top 5 countries. Comparing to the other four sources of tourists countries, United Arab Emirates has no outliers shown on the graph, which means the feedback scores are pretty coherent.

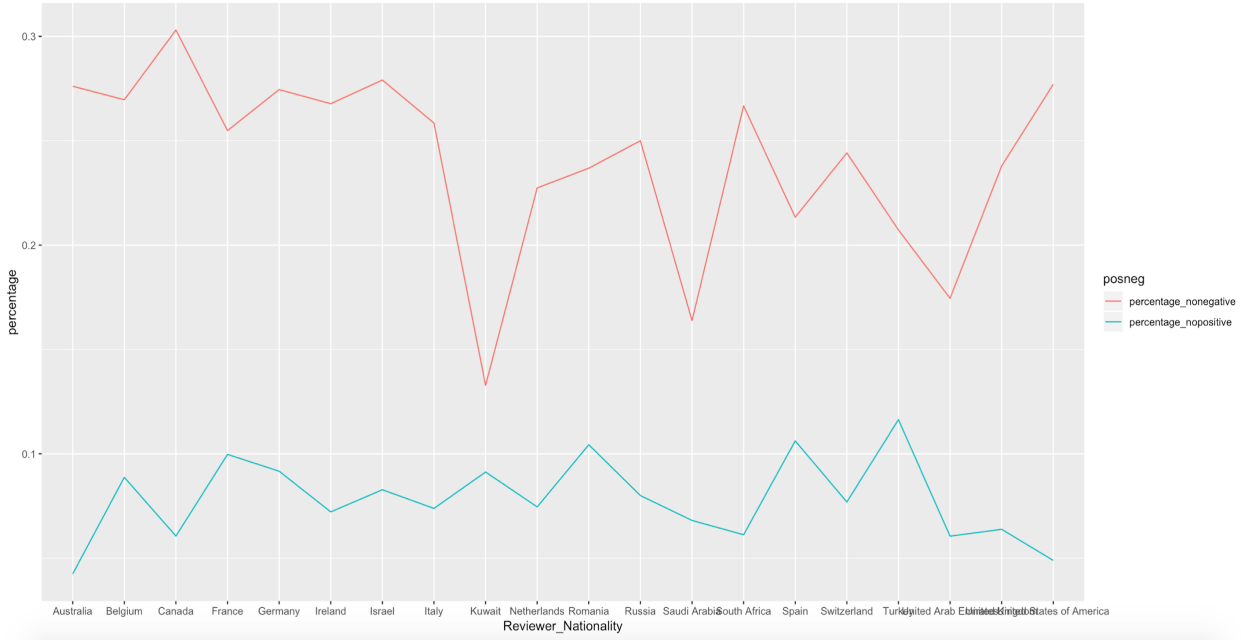
- Percentages for "No Positive"/ "No Negative" Reviews for Top 20 tourist countries

Reviewer_Nationality <chr>	percentage_nopositive <dbl>	percentage_nonegative <dbl>
Australia	0.04247788	0.2761062
Belgium	0.08873720	0.2696246
Canada	0.06060606	0.3030303
France	0.09972299	0.2548476
Germany	0.09166667	0.2745098
Ireland	0.07217848	0.2677165
Israel	0.08278146	0.2790698
Italy	0.07384615	0.2584615
Kuwait	0.09128631	0.1327801
Netherlands	0.07456140	0.2273731

Reviewer_Nationality <chr>	percentage_nopositive <dbl>	percentage_nonegative <dbl>
Romania	0.10434783	0.2368421
Russia	0.08000000	0.2500000
Saudi Arabia	0.06808511	0.1638298
South Africa	0.06122449	0.2666667
Spain	0.10619469	0.2133333
Switzerland	0.07692308	0.2441315
Turkey	0.11636364	0.2072727
United Arab Emirates	0.06054688	0.1745098
United Kingdom	0.06382627	0.2378946
United States of America	0.04893138	0.2770270

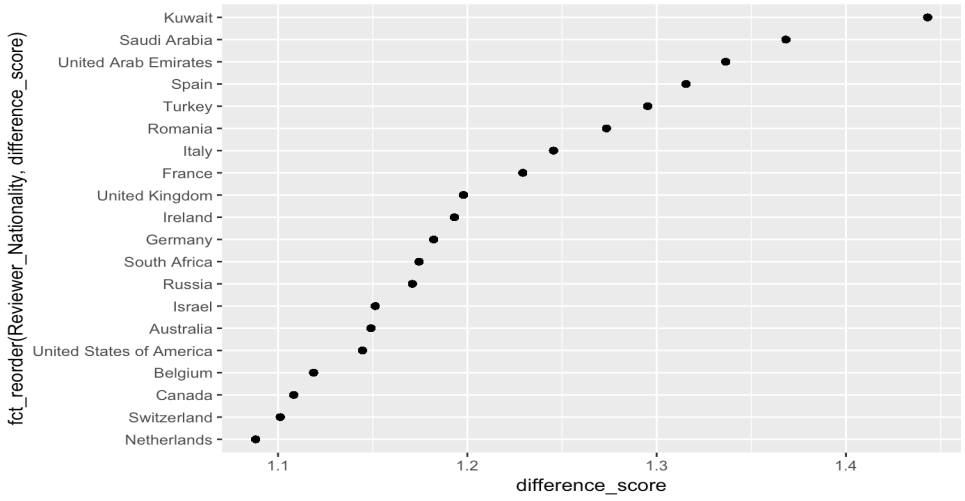
From the table, Turkey has the highest rate of people leaving completely negative reviews, while Canada has the highest rate of people leaving completely positive reviews. By comparing rates of all-positive and all-negative reviews from each country, we see that there are significant differences in terms of travelers' habits of leaving reviews. However, people are more likely to leave all-positive reviews than all-negative reviews in general.

- Percentages for "No Positive"/ "No Negative" Reviews for Top 20 tourist countries



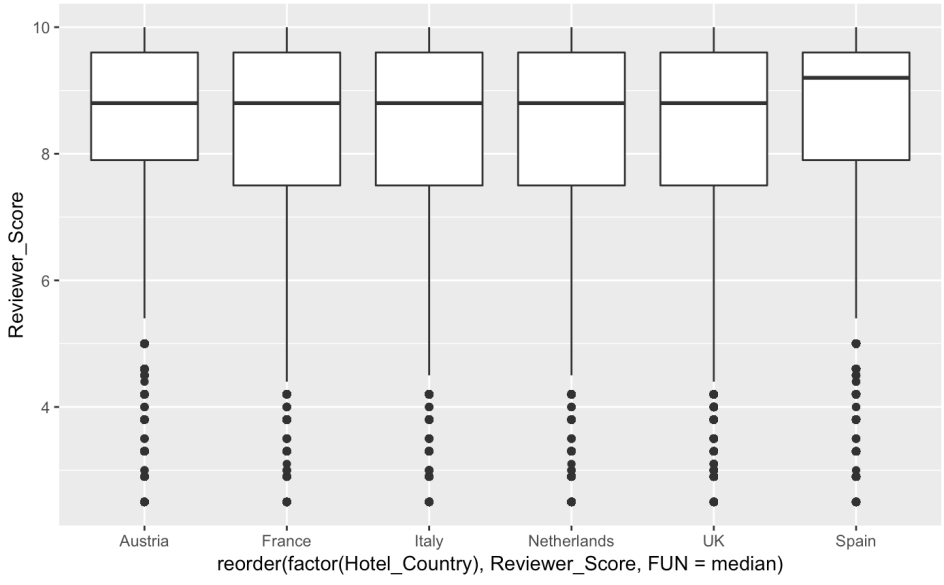
The line graph shows that people are more likely to leave all-positive reviews than all-negative reviews, which is consistent with the observation from the bullet point above. However, we do see that the two trend lines are closer to each other for some countries, such as Kuwait and Saudi Arabia. In other cases, two trend lines are farther away from each other, such as USA, Australia and Canada.

- Scatterplot for the absolute difference between last year's and this year's average review scores among the Top 20 source-country of tourists.



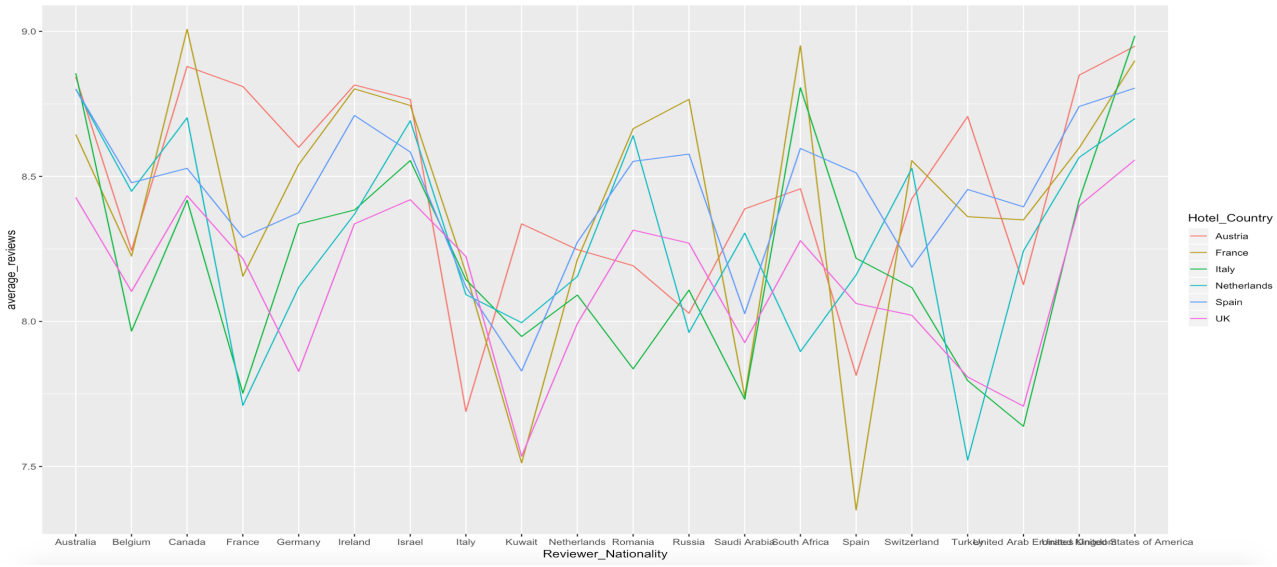
From the above graph, Kuwait has the largest score difference of 1.45, while Netherlands has the lowest score difference of around 1.08.

- Boxplot of review scores for hotels in 6 countries, ordered by median value



From the above graph, Spanish hotels has the highest median review score, while other five countries have similar median review scores. We are curious about the reason behind it and willing to figure it out later.

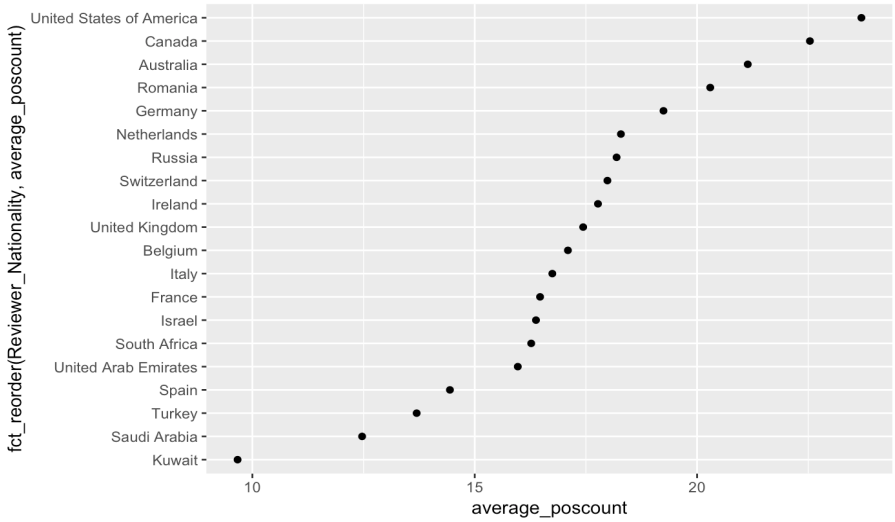
- Line graph for the mean review score for Top 20 tourist countries, ordered by the countries of hotels



From the above graph, we can see that travelers from France and Kuwait tend to give lower scores to hotels, regardless of their geographic areas. On the other hand, travelers from

the USA, Canada and Australia tend to give higher scores to hotels around the world. This is consistent with our findings in previous graphs.

- Scatterplot for the mean Total Positive Word Count for Top 20 tourist countries, ordered by the overall median



From this scatterplot, we can see that travelers from the USA, Canada and Australia are the most generous when giving positive reviews, while travelers from Kuwait, Saudi Arabia and Turkey

are less interested in giving long positive reviews. Interestingly, travelers from countries that are likely to give higher review scores are also more likely to provide longer positive reviews, and vice versa.

Discussion

Based on the tables and graphs presented above, we first notice that travelers are mostly from developed countries, where people have more disposable income on average. We then conclude that travelers from different countries have different habits of leaving reviews - some are more likely to leave all-positive reviews, some are more likely to leave all-negative ones, and others have more balanced ratio of all-positive and all-negative reviews. We also see that even though hotels from different countries have similar median scores, the distribution of scores could be quite different. Finally, we seem to discover an association between average review score and word counts on positive reviews. Travelers from countries that are likely to give higher review scores are also more likely to provide longer positive reviews, such as the USA and Australia. The opposite is also valid. Kuwait and Saudi Arabian travelers have higher rates of giving lower hotel review scores, and they are also more likely to give shorter positive reviews. This pattern could be correlated with different communication patterns among countries.

However, we have not yet got the opportunity to investigate what composed of these reviews, such as the intensity of positive/negative words, and keywords that represent features of hotels important to travelers. We need more such information so that we will be able to look at the usefulness of our empirical model. This model will focus on enabling hotels to provide customized features to serve travelers from different countries.

References

Xie, K.L., Zhang, Z., Zhang, Z.Q. (2014). The business value of online consumer reviews and management response to hotel performance. *International Journal of Hospitality Management, 43*, 1–12.

Sparks, B.A., So, K.K.F, Bradley, G.L. (2016). Responding to negative online reviews: The effects of hotel responses on customer inferences of trust and concern. *Tourism Management, 53*, 73–85.

Zhang, Y. & Vasquez, C. (2014). Hotels' responses to online reviews: Managing consumer dissatisfaction. *Discourse, Context and Media, 6*, 54–64.

Xie, K.L., So, K.K.F, Wang, W. (2017). Joint effects of management responses and online reviews on hotel financial performance: A data-analytics approach. *International Journal of Hospitality Management, 62*, 101–110.

Sparks, B.A. & Victoria, B. (2011). The impact of online reviews on hotel booking intentions and perception of trust. *Tourism Management, 32*, 1310–1323.