GR5063 Data Visualization Proposal

Team Members: Jiaqing Ge, Ruowang Li, Yejie Yu, Shangqing Li, Zhoujun Zheng

Emails: jg4185@columbia.edu, rl3096@columbia.edu, yy2835@columbia.edu, sl4633@columbia.edu, zz2687@columbia.edu

Title: Group Y - Hotel Reviews - How to be a guest favorite hotel with top rated

Abstract: We are curious about what factors make a hotel popular and receive positive reviews from guests. And also whether hotels receiving positive reviews have similar characteristics. In order to discover these, we will explore the following questions step by step to find possible answers:

- Whether the location of a hotel would influence the sentiment of reviews and the score given by the reviewers? If yes, how do the review sentiment and the score vary from different locations?
- Whether the sentiment of reviews and the score given by the reviewers are related to the type of trip, the number of days staying, and the number of guests who stayed? We want to investigate if some attributes of the trip and/or the guests are related to the sentiment of reviews and the score given.
- Are there any relationships between the reviewer's nationality and the scores they give? We would explore guests from which country are more generous in giving a high score.
- If a hotel is doing business in more than one location, we want to investigate whether the hotels under the same name but in different locations receive similar reviews and scores or not.
- What are the words that appear frequently in the positive reviews and the negative reviews separately?

Techniques: ggplot2, ggmap, interaction, clustering, text analysis, and network analysis

Data Description: The dataset we plan to use is named 515K Hotel Reviews Data in Europe. This dataset can be found at https://www.kaggle.com/jiashenliu/515k-hotel-reviews-data-in-europe.

- The data was scraped from Booking.com. This dataset is originally owned by Booking.com, but it is publicly available to everyone already.
- This dataset contains 515,000 customer reviews and scoring of 1,493 luxury hotels across Europe.
- This dataset provides the geographical location of hotels by specifying their latitude and longitude.
- It also includes details about customers and reservations such as nationality, type of trip, type of room reserved, the number of days staying, the number of people traveled with, and so on.

Visualizations:

- Country Map (UK based): We would explore the geographical distribution of the hotels with positive or negative reviews and different scoring. We may add crime data to further investigate possible relationships between review sentiments and location with the crime rate.
- Bar chart: The dataset provides information about the guests and the trip. We can divide them into specific groups. For example, for the guests who stay for business purposes and those who stay for leisure purposes, we can see whether the reviews and the scores given are similar or different between these two groups.
- Line chart: Plotting the scores against the review date can reveal if there is any seasonal pattern of the scoring given to the hotels.

- Histogram: For each hotel, a histogram can show the distribution of the scoring.
- Boxplot: We can compare the distribution of the scoring given to different hotels and find out which hotels are top-rated.
- Scatter plot: We can explore the relationship between the scoring of the hotels and the continuous attributes that we are interested in, such as the number of days staying and the number of guests who stayed.
- Word Cloud: Based on the positive and negative reviews, we can cluster hotels. And for each cluster, a word cloud could be developed to find what words are frequently appeared in the reviews. In this case, we can understand in what aspects the hotel is doing well and in what aspects improvement is needed.