Stat628: Data Science Practicum
Module 3: Executive Summary

Group 15: Lanxi Zhang, Yifan Ren, Yuhan Zheng

Introduction:

In 2023, the world will end the COVID-19 pandemic, economies worldwide will recover, and the tourism industry will prosper. We obtained data from Yelp for various industries. Among them, we are interested in the entire hotel industry in Santa Barbara County and want to explore what is related to the prosperity of the hotel industry and provide some suggestions for hotel owners for their reference.

Data Merging and Data Pre-processing:

To further investigate what factors result in the success or failure of the Santa Barbara County hotel industry, it is crucial to integrate the datasets to see if we can reveal any important findings. We subset California state out and cities belonging to Santa Barbara County for both JSON and CSV files and ignore missing county FIPS. After merging JSON files first, we reformatted the date to one common format for matching, for any matched dates, we added the needed features from CSV to the merged data frame from JSON.

1. data type conversion:

First, we replace NA or empty strings in all columns of the dataset with zeros, convert logical strings to numeric values, e.g., map specific strings representing true conditions (e.g., "True", "yes_corkage", etc.) to 1, and strings representing false conditions (e.g., "False", "none", etc.) to 0, and speak of mapping multiple categorical strings to multiple numeric values.

2. Scale:

We find for example that the Trip of numbers of variable ranges in the millions and is not on the same scale as the rest of the data. To make the parametric model more stable and accurate, we normalize the data to a range between 0 and 1 based on the minimum and maximum values of the columns.

Exploratory Data Analysis:

We first explore the average star rating and text length in Santa Barbara County, the average star rating is around 3.60 and the average text length is around 695 characters. We also draw the histogram to describe the distribution of star ratings (**Fig.1**). We notice that a few attributes are missing in some of the hotels, so we want to find attributes shared by some hotels and count their frequency (**Fig.2**). According to the plot, we find the most common attributes in

AcceptCreditCards, which means using credit cards offers much more convenience for customers when staying in hotels.

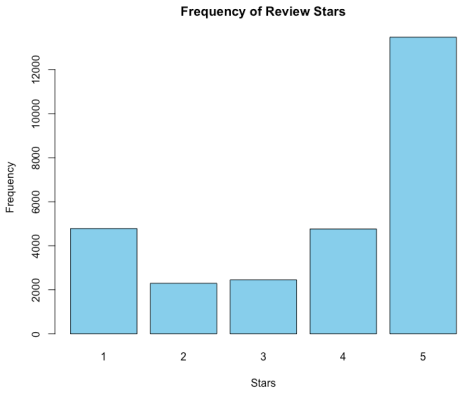

Fig.1 Barplot of Rating Stars

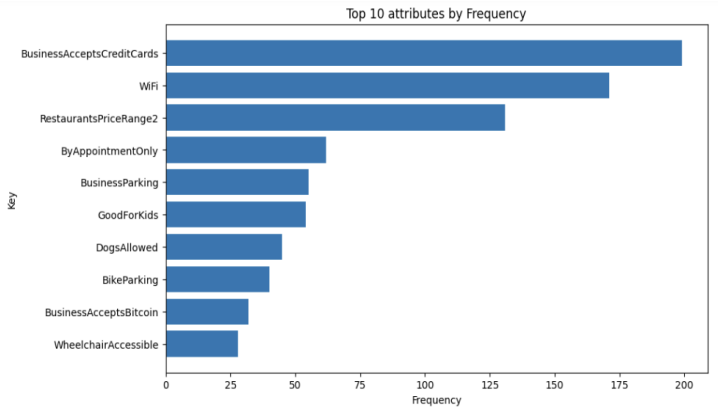

Fig.2 Top 10 Attributes Frequency

After that, we are also curious about the relationship between rating stars, review count, and text characteristics (i.e. text length, average word length). We calculate correlation by the Pearson correlation method (**Fig.3**). The result shows that star reviews have a negative correlation with comment length, which implies that customers are likely to laud when they feel satisfaction while using many words to complain about the bad things about the hotel. Because of that, we draw a scatter plot to depict this phenomenon (**Fig.4**). From this figure, we truly find that hotels with 1 star has longer comment than hotels with 5 stars. This phenomenon is consistent with the previous assumption.

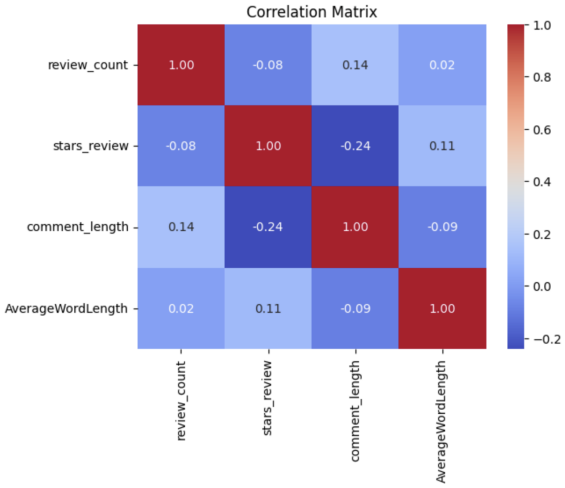

Fig.3 Correlation Matrix

Fig.4 Scatter Plot

Model:

We decided to explore the impact of different attributes and demographic variables of the hotel on the business star rating through a linear regression model. Based on the AIC criterion used

to optimize the model by adding or deleting variables using the stepwise regression method. And then we have the regression model,

$$\begin{aligned} Starsbusines = {} & 3.78945 + 0.34490 \times WheelchairAccessible + 0.40396 \times BikeParking \\ & + 0.14645 \times ByAppointmentOnly - 0.23009 \times DogsAllowed \\ & + 0.36834 \times GoodForKids + 0.07981 \times BusinessAcceptsCreditCards \\ & - 0.14095 \times WiFi - 0.05202 \times RestaurantsPriceRange2 \\ & + 0.63085 \times RestaurantsReservations + 1.02001 \times RestaurantsDelivery \\ & - 2.11767 \times RestaurantsAttire + 0.18537 \times RestaurantsGoodForGroups \\ & - 1.15036 \times RestaurantsTableService + 0.07984 \times Number.of.Trips \end{aligned}$$

The coefficients for each attribute indicate how much that attribute is expected to contribute to stars_business (business star rating). A positive coefficient indicates that the star rating is expected to increase as the attribute increases, while a negative coefficient does the opposite. For example, a coefficient of 1.02001 for RestaurantsDelivery indicates that stars_business is expected to increase by approximately 1.02 stars if the attribute increases by one unit. Conversely, DogsAllowed has a coefficient of -0.23009, which means that for every unit increase in this attribute, the star rating is expected to decrease by about 0.23 stars.

Sentimental Analysis:

We then conduct the sentimental analysis. First, we do the data cleaning for the text part, which includes the review data in the hotel industry from Yelp in Santa Barbara County. We remove the numbers, special characters, stop words, proper nouns, and single words, then change all the reviews to lowercase for convenience to do the following steps. We calculate the sentiment score for each review, if the sentiment score is greater than 0, we label this review as 'positive', if the sentiment score is 0, we label this review as 'neutral', and if the sentiment score is less than 0, we label this review as 'negative'. We draw the relationship between the review stars (1-5) with sentimental labels (**Fig.5**).

Furthermore, we extract the nouns from positive reviews and negative reviews and find that there are some important characters in both kinds of reviews (i.e. room, location, staff), we make the word frequency plot to reflect this feature directly (**Fig.6**). To provide some suggestions for hotel owners, we examine the former words and the latter words around these key characteristics. The positive reviews may contain 'clean' for the room, 'friendly' for the staff, and 'great' for the location, etc.

From **Fig.5,** the hotel with 5-star ratings has the most positive ratings, and the one with 1-star ratings has relatively more negative ratings than the other 4 categories. Based on our analysis, it is clear that the positive rating has positively influenced the hotel rating stars.

From **Fig.6**, we can see that both of the top 20 negative ratings and top 20 positive ratings have something in common, for instance, the two most frequently occurring words are room and hotel. We can infer that customers are concerned about the same characteristics in the hotel

industry. If different hotels provide Different quality of services, customers are more likely to compare the same characteristics in different hotels and give different ratings.

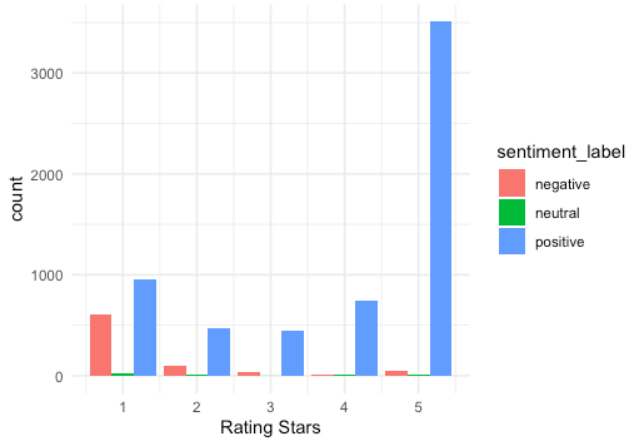

Fig.5 Barplot of rating stars and sentimental label

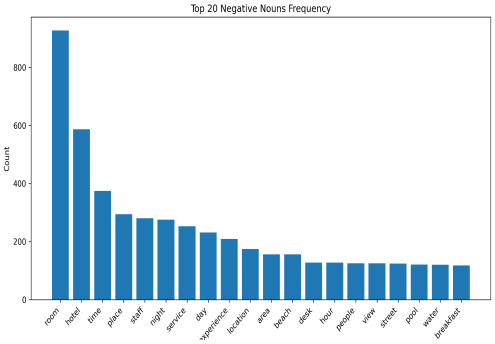

Fig.6 (a) Top 20 Negative Nouns

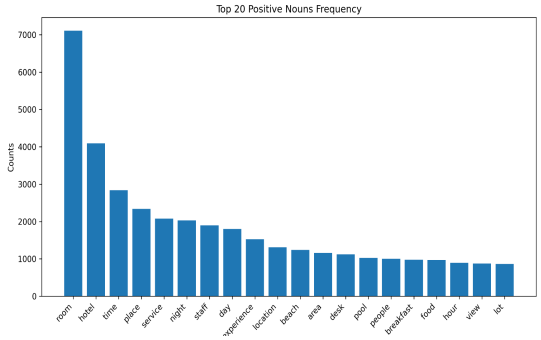

Fig.6 (b) Top 20 Positive Nouns

Fig.6 Top Nouns Frequency

Conclusion:

Based on the former analysis, we can conclude that star ratings have a positive effect on WheelchairAccessible, BikeParking, ByAppointmentOnly, GoodForKids, BusinessAcceptsCreditCards, RestaurantsPriceRange2, RestaurantsReservations, RestaurantsDelivery, RestaurantsGoodForGroups, and Number.of.Trips. If people want to start up a new hotel or hotel owners would like to improve their service quality, they can consider these aspects to enhance their star level. At the same time, the hotel owners and people who want to set up a new hotel might care more about the cleanliness of the room, the attitude of the staff, and the location of the hotel. By improving these areas, hotels are more likely to receive higher ratings and thus be successful in the hospitality industry.

Contribution:

Lanxi Zhang: data merging, regression part, shiny part, summary part.
Yifan Ren: data pre-processing part, shiny part, summary part.
Yuhan Zheng: sentimental analysis part, EDA part, summary part.