THE HOTEL INDUSTRY: COVID-19 IMPACTS

Aspen Schmidt, 11704154

Xitong Hu, 20061192

Trinh Phan, 20231984

Maryam Taherirani, 20165407

12.4.2020

TABLE OF CONTENTS

EXECUTIVE SUMMARY

The travel industry has been significantly affected by the COVID-19 pandemic. As safety and cleanliness have become an even more important consideration for traveling individuals, expectations and demands for the hotel industry have also increased. Through conducting descriptive statistical analysis and sentiment analysis on recent customer reviews, consumers' key considerations for hotel selection during the pandemic will be revealed. The descriptive statistical analysis will provide key insights found within the collected data. The sentiment analysis will uncover the primary feelings consumers are exhibiting and classify the various topics found within the data. This information combined can assist hotels in developing a strategic plan to increase and maintain their occupancy rates. The COVID-19 pandemic has caused travelers and hotel guests to be more conscious and concerned about where they stay in order to remain safe. By determining the most important expectations and demands of hotel guests, along with their current feelings towards their hotel experiences, hotel companies can improve their current processes and standards. By doing so, hotels will additionally improve their reviews and increase occupancy. If hotels are consistently unable to meet the identified guest expectations and guest feelings are negative, they may fail to remain in business. Therefore, it is very important for the hotel companies to understand what steps need to be implemented in order to remain viable and increases customer satisfaction.

PROJECT SCHEDULE

Below is the first GANTT Chart Schedule from Deliverable 1 which encompasses all of the project's activities and tasks for each deliverable. Each deliverable's task is represented by a different color which includes: Deliverable 1 in green, Deliverable 2 in blue, and the Final Presentation in pink. By each of the tasks is the name of who is responsible for completing it. Finally, the yellow diamonds represent the milestones of when each of the three deliverables are due.

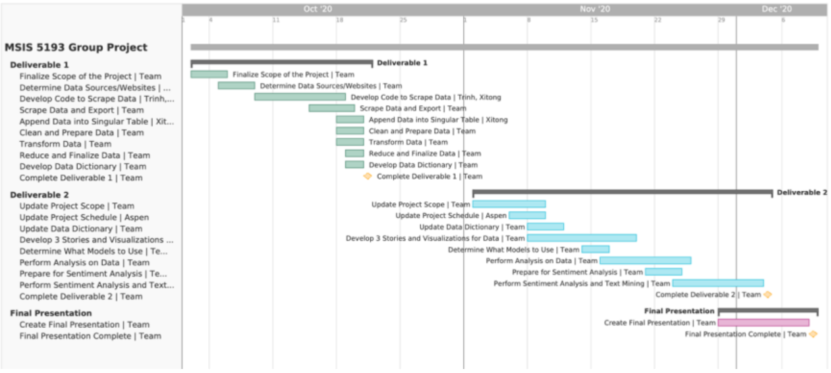


FIGURE 1. DELIVERABLE 1 INITIAL GANNT CHART SCHEDULE.

Below is the updated GANTT Chart Schedule. This schedule additionally includes the percentages of the remaining time needed to complete each of the tasks. Some of the changes to the schedule were shifting all of the Deliverable 2 tasks to be in a more condensed completion timeframe and adding various tasks. Since the Visualization section became optional, it was moved to the bottom of the Deliverable 2 schedule section and was noted as an optional task. Overall, the duration of each task remained the same, however it became easier to work on the various tasks simultaneously as opposed to being spread out. Because the schedule was created with plenty of flexible time to make adjustments as needed, it worked very well for the team.

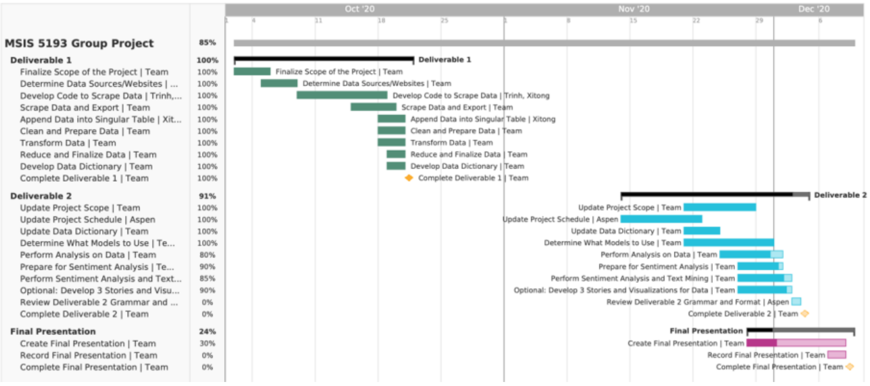


FIGURE 2. UPDATED GANTT CHART SCHEDULE.

STATEMENT OF SCOPE

STATEMENT

The purpose of this project is to uncover primary hotel guest expectations and the overall feelings guests have towards their hotel experience during the COVID-19 pandemic. Because the pandemic has increased concern for health safety, it is essential for hotels to understand guests' concerns and feelings in order to ensure they maintain their levels of quality. The results of this project will provide insights about consumer feelings regarding their recent stay(s) at a Chicago hotel within the previous few months. The city Chicago was selected as it is a centrally located, large city that would provide valuable hotel guest reviews. Specifically, this project is based on data scraped from hotels.com and contains basic hotel information, along with guest reviews at various hotels. Descriptive statistical analysis and sentiment analysis are used to gain the insights about the general consumer feelings and uncover concerns or areas of improvement hotels need to further explore. Additionally, to further develop the project, visualizations are also created explore the data in a different way and deepen the overall story.

OBJECTIVES

This project will identify the primary guest expectations hotels should meet during the COVID-19 pandemic. In order for hotels to remain in business due to the reduction of travel because of the pandemic, their current operation and cleaning processes may need to be adjusted or amplified. This project focuses on the city of Chicago for the sample hotel locations and the generalized population for this project is people, specifically those that have stayed in hotels during the COVID-19 pandemic. The project's objectives include the following:

- Determine the key considerations guests have when selecting a hotel

- Uncover basic statistics regarding the target variable of ‘price’ and explore any significant factors impacting the target variable
- Using sentiment analysis, determine the overall guests’ feelings about their stay found within their reviews
- Analyze the reviews through text analysis to determine the specific topics guests are concerned about and discussing when choosing and reviewing hotels
- Create and define identifiable standards that hotels should maintain during the COVID-19 pandemic; additionally, these standards could be implemented in other hotel chains in other cities

VARIABLES

This project has two datasets. One contains the main hotel features information. Variables in this dataset include: the hotel’s name, price, how many stars the hotel has, the total reviews, the rating, the latitude and longitude location of the hotel, check-in and check-out dates, the number of rooms, and the hotel’s URL. The variable ‘price’ will be used as the target variable for this dataset and variables such as stars, rating, and total reviews will be used as predictors. It is aimed to find whether there are any significant factors affecting the price.

The second dataset contains the hotel’s reviews. Variables in this dataset include: the hotel’s name, the hotel’s URL, the check-in date, the rating, and the review text. This dataset will be used for text analytics and topic modeling to determine the common themes in reviews.

DATA PREPARATION

After reviewing the data that was previously collected by web scraping for Deliverable 1, it was determined by the team that the data was sufficient to complete the analysis. The original datasets contained enough information to create visualizations, conduct descriptive statistical analysis, and complete text mining and sentiment analysis. As mentioned in the scope of the project, the purpose was to ultimately uncover the primary hotel guests' expectations and feelings they had towards their hotel stay and experience. By using the datasets already obtained, it provided enough information to conduct the desired analysis and reveal various insights contained within the data. Because of this, additional data was not collected. However, a brief summary about the data preparation steps from Deliverable 1, along with updates to the steps needed to conduct the analysis' will be discussed further in this section of the report.

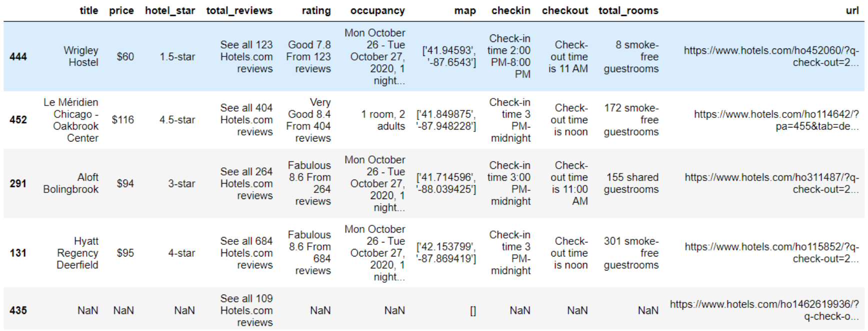
DATA ACCESS

The data was collected from https://www.hotels.com/ and contains data from over 400 hotels in Chicago, Illinois. The Selenium package and XPath Selector were used in Python to access and extract the data. First, a list of URLs for 466 hotels was extracted from the main page of hotels in Chicago. Then, the Selenium package set the driver through Chrome to access the individual websites and extract the required data. This data contains main features of the listed hotels such as: name, price, rating, and reviews, etc. By looping through the list, all of the data was collected and converted into a data frame for further analysis. Next, after opening the page for one hotel, the hotel features were extracted from this main page. However, in order to get the whole review list, the 'see all reviews' button must be clicked to check the review details for this hotel. After clicking the 'see all reviews' button and opening the review page, it shows 50 reviews per page. The reviews in this block are ordered by the date from new to old. The review features scraped for this study include: the rating from the customer, check-in date, and the review content.

DATA CONSOLIDATION

To consolidate the data, a CSV writer was opened, and the file was saved on the directory. Then the web crawler directly wrote data into each row for each hotel. The scraped data was converted into the following two tables: hotel_listings and hotel_reviews.

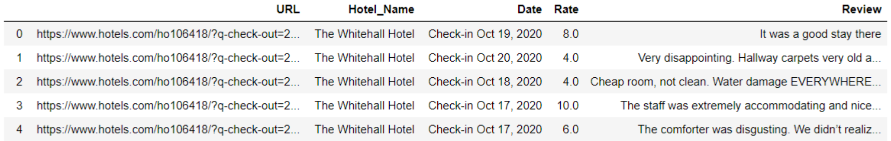
The table below shows the main features that are being taken into consideration. There are a total of eleven variables which include: title (hotel name), price, hotel_star, total_reviews, rating, occupancy, map (latitude and longitude), checkin, checkout, total_rooms, and url.



	title	price	hotel_star	total_reviews	rating	occupancy	map	checkin	checkout	total_rooms	url
444	Wrigley Hostel	$60	1.5-star	See all 123 Hotels.com reviews	Good 7.8 From 123 reviews	Mon October 26 - Tue October 27, 2020, 1 night...	['41.94593', '-87.6543']	Check-in time 2:00 PM-8:00 PM	Check-out time is 11 AM	8 smoke-free guestrooms	https://www.hotels.com/ho452060/?q-check-out=2...
452	Le Méridien Chicago - Oakbrook Center	$116	4.5-star	See all 404 Hotels.com reviews	Very Good 8.4 From 404 reviews	1 room, 2 adults	['41.849875', '-87.948228']	Check-in time 3 PM-midnight	Check-out time is noon	172 smoke-free guestrooms	https://www.hotels.com/ho114642/?pa=455&tab=de...
291	Aloft Bolingbrook	$94	3-star	See all 264 Hotels.com reviews	Fabulous 8.6 From 264 reviews	Mon October 26 - Tue October 27, 2020, 1 night...	['41.714596', '-88.039425']	Check-in time 3:00 PM-midnight	Check-out time is 11:00 AM	155 shared guestrooms	https://www.hotels.com/ho311487/?q-check-out=2...
131	Hyatt Regency Deerfield	$95	4-star	See all 684 Hotels.com reviews	Fabulous 8.6 From 684 reviews	Mon October 26 - Tue October 27, 2020, 1 night...	['42.153799', '-87.869419']	Check-in time 3 PM-midnight	Check-out time is noon	301 smoke-free guestrooms	https://www.hotels.com/ho115852/?q-check-out=2...
435	NaN	NaN	NaN	See all 109 Hotels.com reviews	NaN	NaN	[]	NaN	NaN	NaN	https://www.hotels.com/ho1462619936/?q-check-o...

TABLE 1. HOTEL_LISTINGS DATA EXAMPLE

The second table below shows the review features collected from the review page. Each record is one review. In addition to the date, rate, and review contents for the hotel, the URL and hotel name for each review have been added.



	URL	Hotel_Name	Date	Rate	Review
0	https://www.hotels.com/ho106418/?q-check-out=2...	The Whitehall Hotel	Check-in Oct 19, 2020	8.0	It was a good stay there
1	https://www.hotels.com/ho106418/?q-check-out=2...	The Whitehall Hotel	Check-in Oct 20, 2020	4.0	Very disappointing. Hallway carpets very old a...
2	https://www.hotels.com/ho106418/?q-check-out=2...	The Whitehall Hotel	Check-in Oct 18, 2020	4.0	Cheap room, not clean. Water damage EVERYWHERE...
3	https://www.hotels.com/ho106418/?q-check-out=2...	The Whitehall Hotel	Check-in Oct 17, 2020	10.0	The staff was extremely accommodating and nice...
4	https://www.hotels.com/ho106418/?q-check-out=2...	The Whitehall Hotel	Check-in Oct 17, 2020	6.0	The comforter was disgusting. We didn't realiz...

TABLE 2. HOTEL_REVIEWS DATA EXAMPLE

DATA CLEANING

Data cleaning is a critical step for any data science project. After extracting data from the website, the hotel_listings table indicates that all data has the type of string; and therefore, the desired data must be extracted or converted into the proper format. To clean the data, different techniques including replacing values and regex for finding and extracting the desired data were used. The complete list of data cleaning steps that were taken is below:

- Removal of the $ sign in price variable
- Extract only the value of hotel star
- Extract number of reviews from string in total_reviews column
- Collect check-in and check-out time, removing unnecessary strings
- Extract the numeric value in total_rooms column as number of rooms in that hotel
- Adjust numeric columns into integer or float types
- Remove missing values in rows; because XPath is changed for some URLs, the results return null

As a result, the final dataset includes 419 observations and 14 columns.

For the hotel_reviews table, any review without text review content (133/9,621) is dropped from the dataset. The current values in the date column are in string format and not in the necessary date format. For example: "Check-in Oct 19, 2020". Therefore, the date information is extracted from the review table and changed into the correct date format. Additionally, the rating data from the source table is in float pattern and string format and was changed to integer format. The final review dataset includes 9,488 reviews and 5 columns.

Additionally, more cleaning steps had to be taken in order to successfully conduct sentiment analysis. The first step in the positive and negative sentiment analysis and the emotion analysis was to prepare the data for plotting and analysis. Before any plots were created, data cleaning was conducted. This included the following steps:

1. Removing stop words
2. Removing numerical values

3. Removing punctuations
4. Changing all words to lower cases
5. Removing uninformative frequent words
6. Stemming the words

DATA TRANSFORMATION

In order to conduct linear regression, data transformation had to be done. Linear regression has four assumptions which include: normality, homoscedasticity, linearity between predictor and target variables, and independent residuals. In the Descriptive Statistics and Analysis section of the report, linearity will be examined in scatter plots, and normality will be examined in the distribution plots. Because continuous variables are typically not normally distributed, log transformation will be applied to the predictor variables with the goal of achieving a better distribution.

DATA REDUCTION

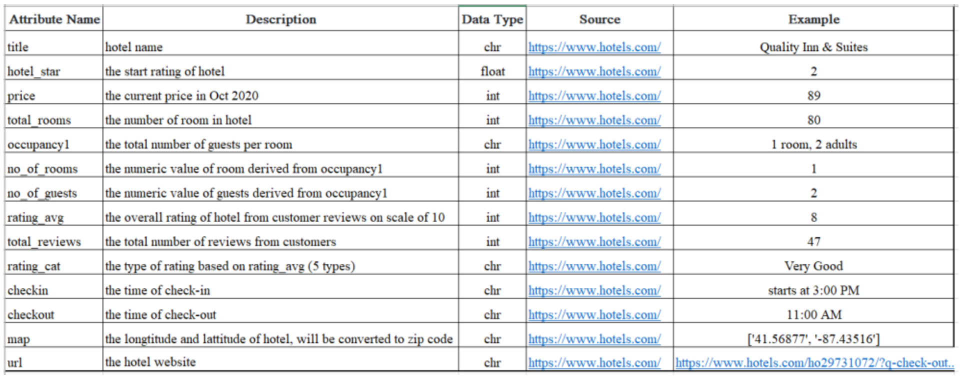
Some of the hotels on hotels.com have thousands of reviews dating back to 2015 or even earlier. Since the primary timeline of this project focuses on the COVID-19 pandemic timeframe, data for the most recent 100 reviews for each hotel was extracted. Additionally, by having approximately 10,000 reviews, topic extraction and topic modeling can be conducted. Therefore, the reviews for the first 100 hotel URLs were scraped. Because some hotels did not have 100 reviews and some reviews did not include review texts, the final dataset had a total of 9,488 reviews to be used for text analytics.

Additionally, in the Text Mining and Sentiment Analysis section of the report, stemming will be applied to the text in order to reduce the terms in the dataset and allow for topic modeling.

DATA DICTIONARY

Below are the two data dictionaries for the two datasets created for this research project.

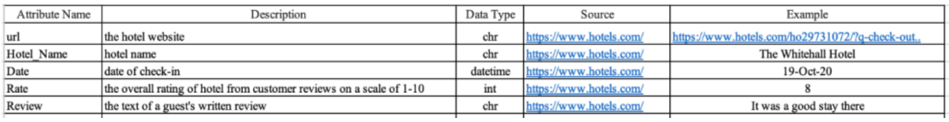
HOTEL MAIN FEATURES



Attribute Name	Description	Data Type	Source	Example
title	hotel name	chr	https://www.hotels.com/	Quality Inn & Suites
hotel_star	the start rating of hotel	float	https://www.hotels.com/	2
price	the current price in Oct 2020	int	https://www.hotels.com/	89
total_rooms	the number of room in hotel	int	https://www.hotels.com/	80
occupancy1	the total number of guests per room	chr	https://www.hotels.com/	1 room, 2 adults
no_of_rooms	the numeric value of room derived from occupancy1	int	https://www.hotels.com/	1
no_of_guests	the numeric value of guests derived from occupancy1	int	https://www.hotels.com/	2
rating_avg	the overall rating of hotel from customer reviews on scale of 10	int	https://www.hotels.com/	8
total_reviews	the total number of reviews from customers	int	https://www.hotels.com/	47
rating_cat	the type of rating based on rating_avg (5 types)	chr	https://www.hotels.com/	Very Good
checkin	the time of check-in	chr	https://www.hotels.com/	starts at 3:00 PM
checkout	the time of check-out	chr	https://www.hotels.com/	11:00 AM
map	the longtitude and lattitude of hotel, will be converted to zip code	chr	https://www.hotels.com/	['41.56877', '-87.43516']
url	the hotel website	chr	https://www.hotels.com/	https://www.hotels.com/ho29731072/?q-check-out..

TABLE 3. HOTEL MAIN FEATURES DATASET EXAMPLE

HOTEL REVIEWS



Attribute Name	Description	Data Type	Source	Example
url	the hotel website	chr	https://www.hotels.com/	https://www.hotels.com/ho29731072/?q-check-out..
Hotel_Name	hotel name	chr	https://www.hotels.com/	The Whitehall Hotel
Date	date of check-in	datetime	https://www.hotels.com/	19-Oct-20
Rate	the overall rating of hotel from customer reviews on a scale of 1-10	int	https://www.hotels.com/	8
Review	the text of a guest's written review	chr	https://www.hotels.com/	It was a good stay there

TABLE 4. HOTEL REVIEWS DATASET EXAMPLE

VISUALIZATIONS

EXPLORATION OF CONTINUOUS VARIABLES

To gain a better understanding of the continuous variables within the dataset, various visualizations were created for each of the variables. By doing this, the frequency, skewness, normality, and outliers of each of the variables were shown and more easily interpreted. This reveals issues that may be found within the data and has provided a deeper dive into the data.

TARGET VARIABLE: HOTEL PRICE

Below are various plots exploring the data's target variable: Hotel Price. The first plot shows the frequency of the distribution of hotel price per night. For example, there are 200 hotels priced around $100 per night.

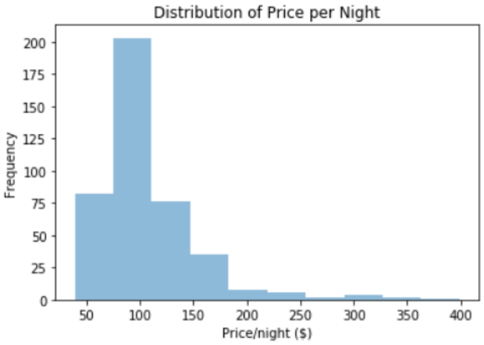


Figure 3. Frequency of Hotel Price

The second plot shows the skewness for this variable. The data is clearly heavily right skewed.

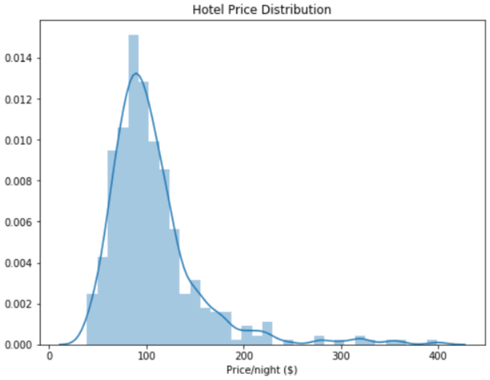


FIGURE 4. SKEWNESS OF HOTEL PRICE

The third plot shows the normality distribution. Because the data is not well aligned with the red line on the plot, it reveals that the distribution is not normal.

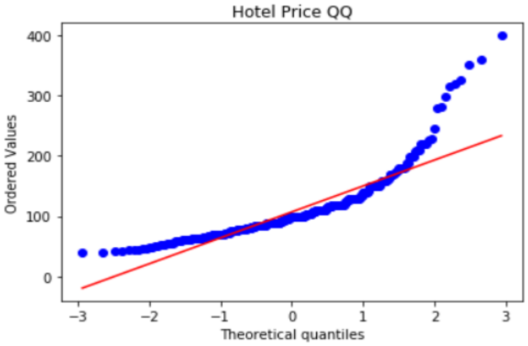


FIGURE 5. NORMALITY OF HOTEL PRICE

Finally, the fourth plot is a boxplot revealing any outliers found within this variable. All of the dots on the plot represent outliers. Because there are several dots, this means there are multiple outliers found in this variable.

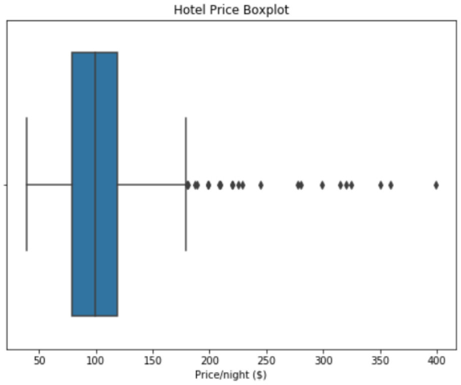


FIGURE 6. HOTEL PRICE OUTLIERS

VARIABLE: HOTEL RATING (STARS)

Below are various plots exploring the data's variable: Hotel Rating. The first plot shows the frequency of the distribution of hotel ratings, or how many stars a hotel has. For example, there are over 100 hotels with a 3.0 hotel star rating.

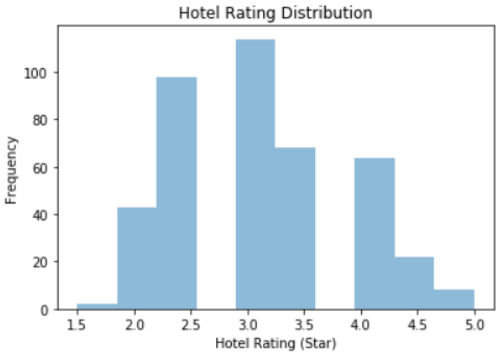


FIGURE 7. FREQUENCY OF HOTEL RATING

The second plot shows the skewness for this variable. The data fairly evenly distributed across the hotel ratings.

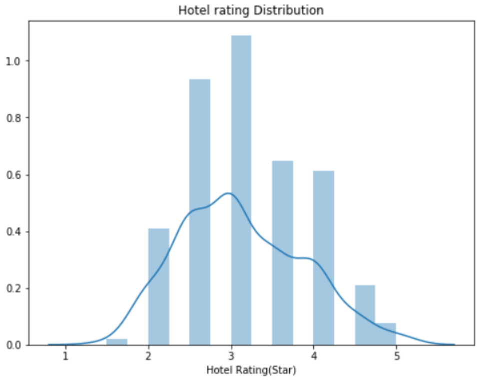


FIGURE 8. SKEWNESS OF HOTEL RATING

The third plot shows the normality distribution. Because the data is not well aligned with the red line on the plot, it reveals that the distribution is not normal.

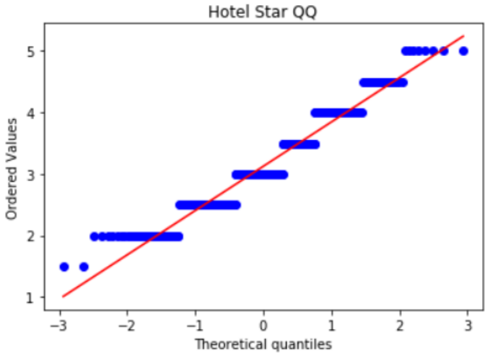


FIGURE 9. NORMALITY OF HOTEL RATING

The fourth plot is a boxplot revealing any outliers found within this variable. Because there are no dots on this plot, there are no outliers found for this variable.

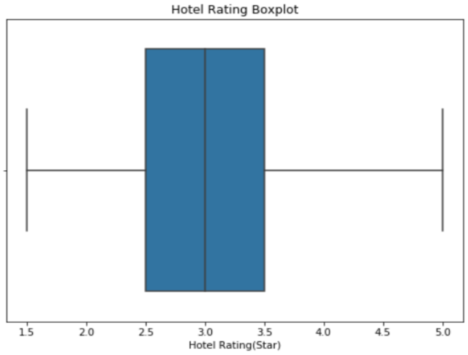


FIGURE 10. HOTEL RATING OUTLIERS

VARIABLE: NUMBER OF ROOMS PER HOTEL

Below are various plots exploring the data's variable: Number of Rooms. The first plot shows the frequency of the distribution of how many rooms a hotel has. For example, there are almost 300 hotels that have between 0-250 rooms.

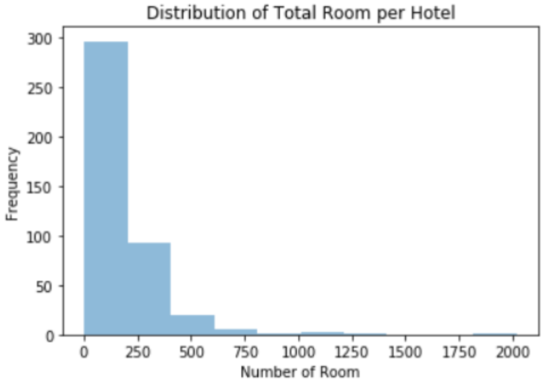


FIGURE 11. FREQUENCY OF NUMBER OF ROOMS

The second plot shows the skewness for this variable. The data is clearly heavily right skewed.

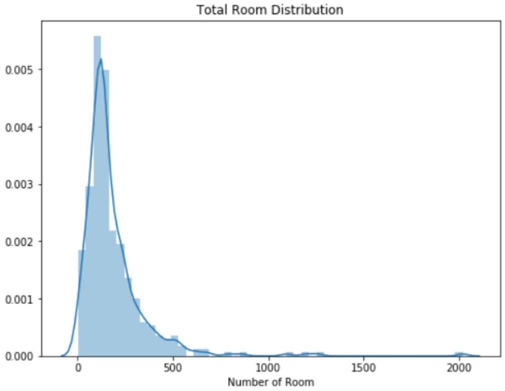


FIGURE 12. SKEWNESS OF NUMBER OF ROOMS

The third plot shows the normality distribution. Because the data is not well aligned with the red line on the plot, it reveals that the distribution is not normal.

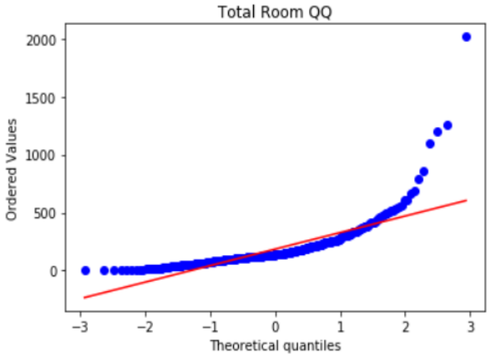


FIGURE 13. NORMALITY OF NUMBER OF ROOMS

Finally, the fourth plot is a boxplot revealing any outliers found within this variable. All of the dots on the plot represent outliers. Because there are several dots, this means there are multiple outliers found in this variable.

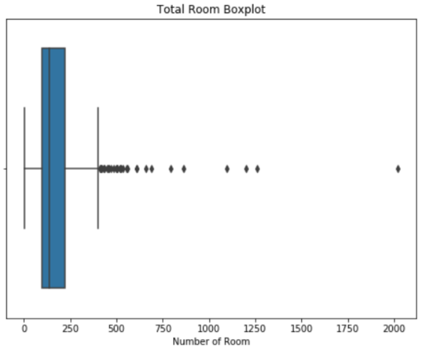


FIGURE 14. NUMBER OF ROOMS OUTLIERS

VARIABLE: CUSTOMER RATING

Below are various plots exploring the data's variable: Customer Rating. The first plot shows the frequency of the distribution of customer ratings. For example, there are almost 120 hotels that have an 8.5 customer rating.

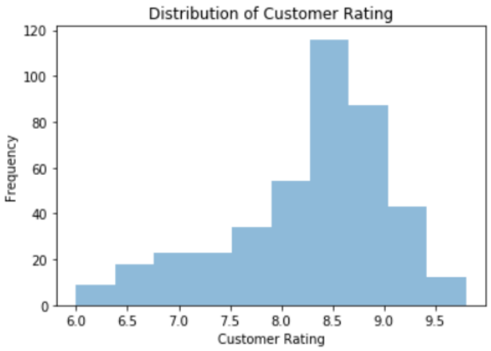


FIGURE 15. FREQUENCY OF CUSTOMER RATINGS

The second plot shows the skewness for this variable. While that data is somewhat evenly distributed, it still has heavier left skewness.

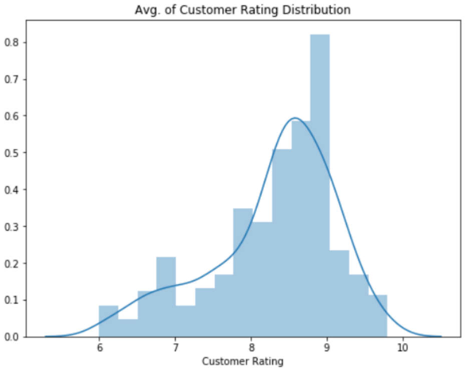


FIGURE 16. SKEWNESS OF CUSTOMER RATINGS

The third plot shows the normality distribution. Because the data is not well aligned with the red line on the plot, it reveals that the distribution is not normal.

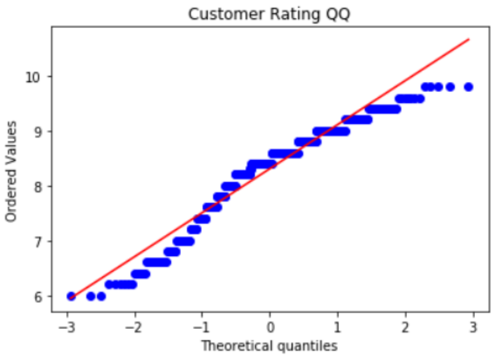


FIGURE 17. NORMALITY OF CUSTOMER RATINGS

Finally, the fourth plot is a boxplot revealing any outliers found within this variable. The two dots on the plot represent outliers. This variable does not have many outliers; however, it still has a couple.

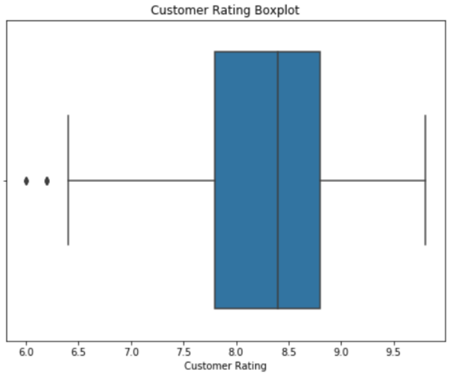


FIGURE 18. CUSTOMER RATINGS OUTLIERS

VARIABLE: TOTAL CUSTOMER REVIEWS PER HOTEL

Below are various plots exploring the data’s variable: Total Customer Reviews. The first plot shows the frequency of the distribution of customer ratings. For example, there are approximately 120 hotels that have approximately 500 reviews.

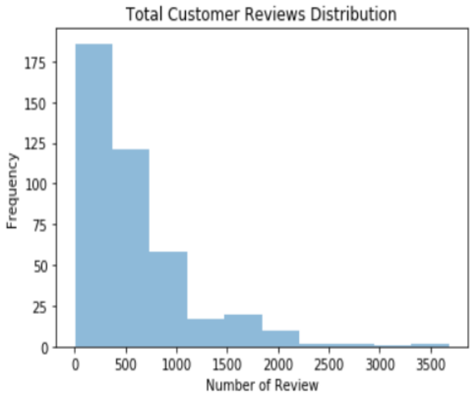


FIGURE 19. FREQUENCY OF NUMBER OF REVIEWS

The second plot shows the skewness for this variable. The data is clearly heavily right skewed.

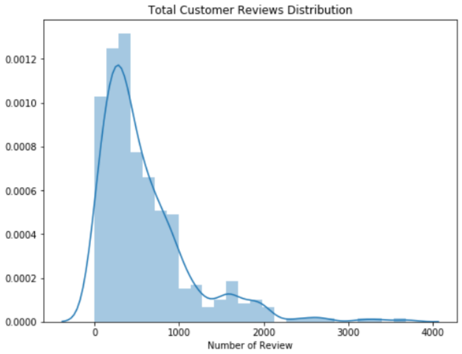


FIGURE 20. SKEWNESS OF NUMBER OF REVIEWS

The third plot shows the normality distribution. Because the data is not well aligned with the red line on the plot, it reveals that the distribution is not normal.

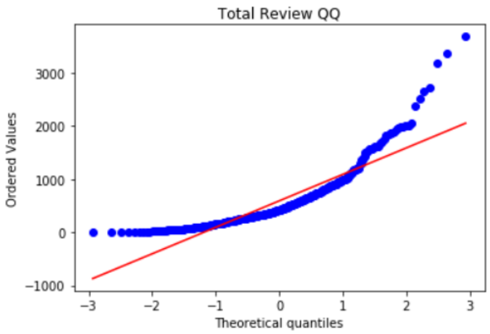


FIGURE 21. NORMALITY OF NUMBER OF REVIEWS

Finally, the fourth plot is a boxplot revealing any outliers found within this variable. All of the dots on the plot represent outliers. Because there are several dots, this means there are multiple outliers found in this variable.

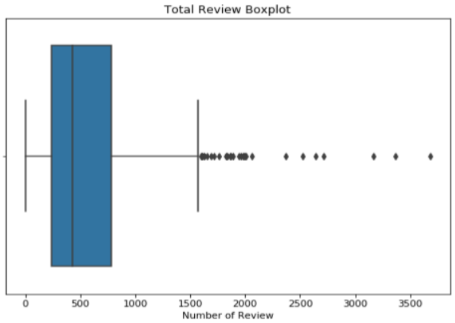
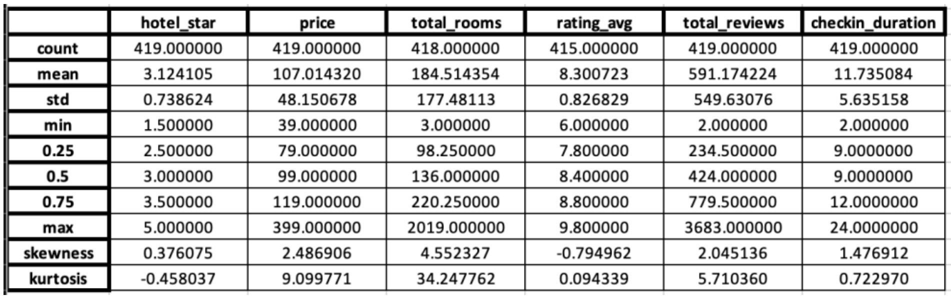


FIGURE 22. NUMBER OF REVIEWS OUTLIERS

DESCRIPTIVE STATISTICS AND ANALYSIS

SUMMARY STATISTICS FOR NUMERIC VALUES

The describe() function was used to collect the summary statistics for numeric values, and they are as follows:



	hotel_star	price	total_rooms	rating_avg	total_reviews	checkin_duration
count	419.000000	419.000000	418.000000	415.000000	419.000000	419.000000
mean	3.124105	107.014320	184.514354	8.300723	591.174224	11.735084
std	0.738624	48.150678	177.48113	0.826829	549.63076	5.635158
min	1.500000	39.000000	3.000000	6.000000	2.000000	2.000000
0.25	2.500000	79.000000	98.250000	7.800000	234.500000	9.0000000
0.5	3.000000	99.000000	136.000000	8.400000	424.000000	9.0000000
0.75	3.500000	119.000000	220.250000	8.800000	779.500000	12.0000000
max	5.000000	399.000000	2019.000000	9.800000	3683.000000	24.0000000
skewness	0.376075	2.486906	4.552327	-0.794962	2.045136	1.476912
kurtosis	-0.458037	9.099771	34.247762	0.094339	5.710360	0.722970

TABLE 5. NUMERIC VALUES SUMMARY STATISTICS

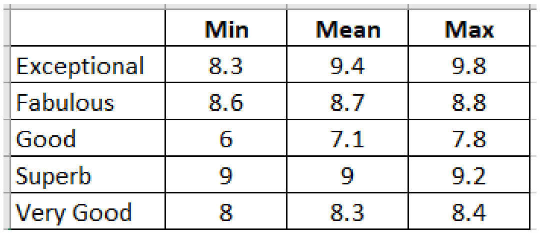
The statistics above align well with all of the visualizations discussed in the previous section.

CATEGORICAL VARIABLES EXPLORATION

For this section, the relationship between price and some of the categorical variables is explored.

PRICE VERSUS RATING STATISTICS

Each of the rating categories are determined by customer ratings. The statistics of the customer ratings and rating categories are in the following table:



	Min	Mean	Max
Exceptional	8.3	9.4	9.8
Fabulous	8.6	8.7	8.8
Good	6	7.1	7.8
Superb	9	9	9.2
Very Good	8	8.3	8.4

TABLE 6. PRICE VS CATEGORY STATISTICS

PRICE VERSUS HOTEL RATING BOX PLOT

The box plot is used to show the overall response pattern for the Price versus Hotel Rating variables. The given plot provides useful information about the distribution of price among the hotel rating types. It is revealed that the Exceptional rating group is more dispersed, while the Good rating group is comparatively condensed. This suggests that the price is quite different in the Exceptional rating group but not in Good rating group. In conclusion, there are obvious differences between price and hotel rating types.

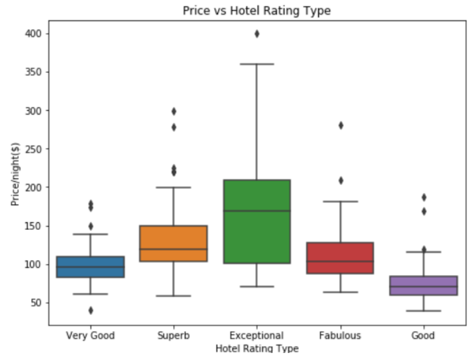


FIGURE 23. PRICE VERSUS HOTEL RATING BOX PLOT

DISTRIBUTION OF TIME

CHECK-IN TIME

Each hotel has a different check-in policy. The given bar chart below shows the varying flexible check-in times at the hotels. As seen below, most of the hotels allow customers to check-in from 3:00 PM to midnight. It was surprising to see that some hotels do not allow check-in until 5:00 PM.

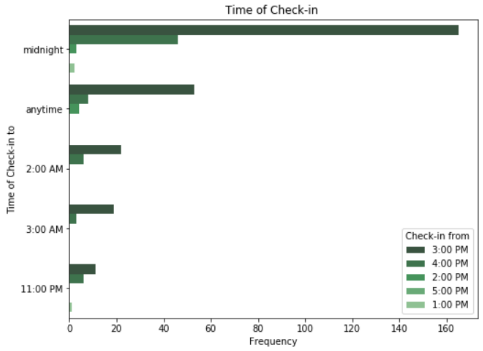


FIGURE 24. DISTRIBUTION OF CHECK-IN TIME

CHECK-OUT TIME

There is a primarily high frequency of hotels that have check-out times between 11:00 AM and 12:00 PM. However, at a small number of hotels, guests must check-out by 10:00 AM.

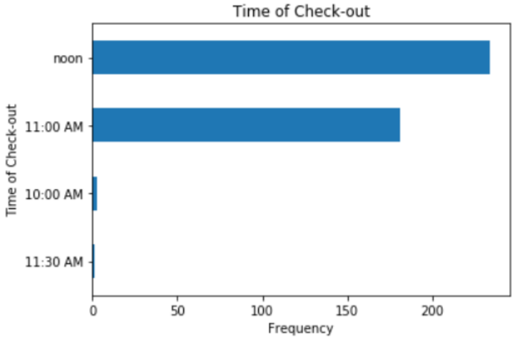


FIGURE 25. DISTRIBUTION OF CHECK-OUT TIME

MULTIVARIATE ANALYSIS

Next, multivariate analysis was used to develop the following charts that provide more information about the relationship between the target variable and the predictor variables.

PRICE VERSUS HOTEL RATING (STARS)

The scatter plot below shows the linear relationship between the Price Per Night and Hotel Rating variables. As expected, hotels with higher ratings above 4 stars are considered as luxury hotels and have significantly higher prices than lower rated hotels.

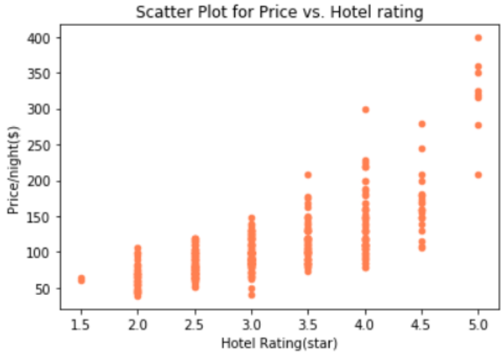


FIGURE 26. PRICE VERSUS HOTEL RATING (STARS) SCATTER PLOT

PRICE VERSUS NUMBER OF ROOMS

The number of rooms in a hotel represents the capacity or size of that hotel. It is assumed that the hotels with higher capacity rates will have higher revenue margins. It is also assumed that the hotel's capacity rate also affects the price of a hotel room. The scatter plot below reveals that as the number of hotel rooms increase, the price per night also increases. A linear regression plot is drawn to show the trend found in the data. However, in the plots below, there are some outliers. For example, there are four hotels which have high capacity rates but also have low prices – such as cost is less than $200 per night and there are over 1,000 rooms.

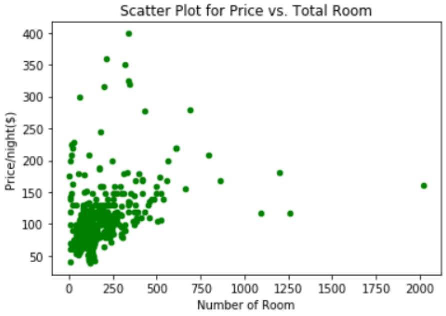


FIGURE 27. PRICE VS NUMBER OF ROOMS

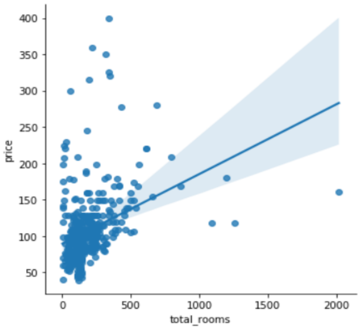


FIGURE 28. PRICE VS NUMBER OF ROOMS TREND

PRICE VERSUS CUSTOMER RATING

Customer ratings is one of the most important factors for evaluating any business' performance. Highly rated hotels provide guests with comparable services that strive to match guest expectations. The plot below clearly shows there is a strong positive relationship between the variables price and customer rating.

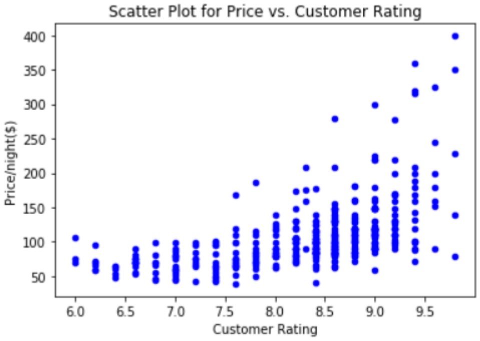


FIGURE 29. PRICE VS CUSTOMER RATING

PRICE VERSUS TOTAL CUSTOMER REVIEWS PER HOTEL

Higher numbers of total customer reviews per hotel is related to the hotel's popularity. If a hotel is more popular, it is more likely to have more reviews. Based on supply and demand principles, more well-known and popular hotels may have higher prices than other hotels, even though they both provide similar level of quality services and experiences. The plot below shows a light, positive relationship between price and total customer reviews per hotel. There are some outliers at the bottom right side of the plot.

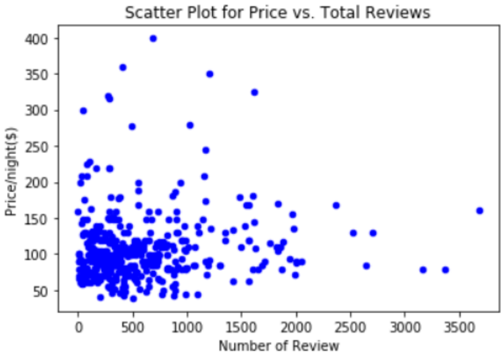


FIGURE 30. PRICE VERSUS TOTAL CUSTOMER REVIEWS PER HOTEL

PRICE LINEAR REGRESSION

Although there are limited features, linear regression is still applied to predict price using continuous variables. First, the data was split into training and validation (80:20) datasets. The split resulted in the training dataset containing 335 hotels, and the validation dataset containing 84 hotels.

The heatmap was determined to be the best way to obtain an overview of price and its corresponding relationships. According to the matrix, the variables listed below are most correlated with the target variable Price.

- The variables "hotel_star" and "rating_avg" are strongly correlated with "Price" as their correlations are 0.71 and 0.53 respectively.
- The "total_rooms" variable is moderately correlated with a correlation of 0.33.
- The "total_reviews" variable is weakly correlated with a correlation of 0.11.
- Even though the longitude and latitude variables show moderate correlation with Price, and location may impact a hotel's pricing, the variables longitude and latitude are converted to a

map and not for model building. Therefore, they are not selected as variables for the regression model.

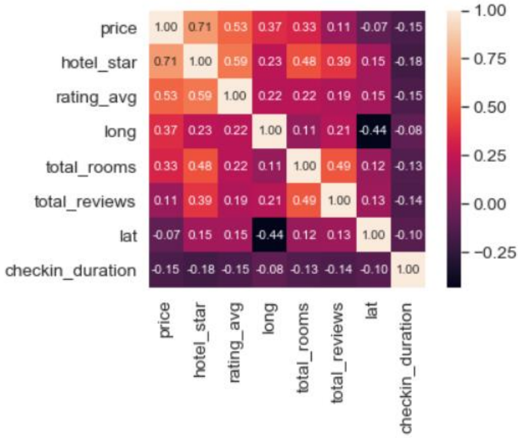


FIGURE 31. CORRELATION MATRIX

OUTLIERS

Because regression modeling is sensitive to outliers, they can significantly influence the model. Therefore, before further analysis can be conducted, the outliers must be examined and removed if deemed necessary.

First, the below scatter plot shows some of the outliers found within the data. These outliers include a hotel having a high total number of rooms and low prices or having a small total number of rooms but high prices.

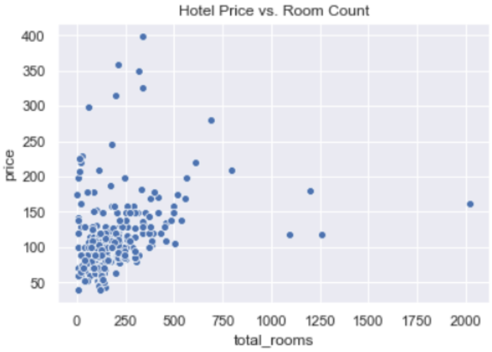


FIGURE 32. HOTEL PRICE AND TOTAL NUMBER OF ROOMS WITH OUTLIERS

After examining these outliers, it was determined that they needed to be removed from the dataset. The observations with the outliers were deleted from the training dataset, and the plot with these updates can be seen below. After these adjustments were made, the distribution was improved.

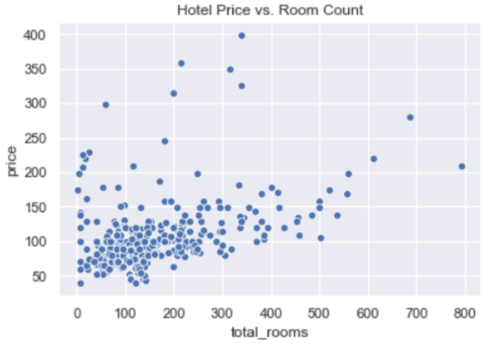


FIGURE 33. HOTEL PRICE AND TOTAL NUMBER OF ROOMS WITHOUT OUTLIERS

Next price and the total number of reviews variables were explored by creating a scatter plot. The first plot below shows some outliers such as hotels having over 2,000 reviews and prices less than $150.

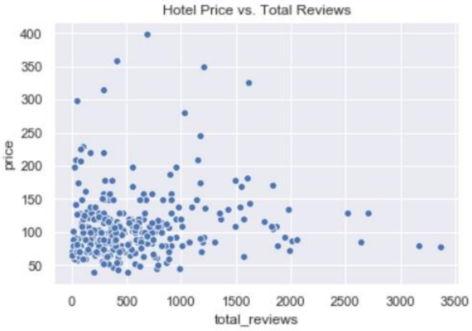


FIGURE 34. HOTEL PRICE AND TOTAL NUMBER OF REVIEWS WITH OUTLIERS

Again, after examining these outliers, it was determined that they needed to be removed from the dataset. The observations with the outliers were deleted from the training dataset, and the plot with these updates can be seen below. After these adjustments were made, the distribution was improved.

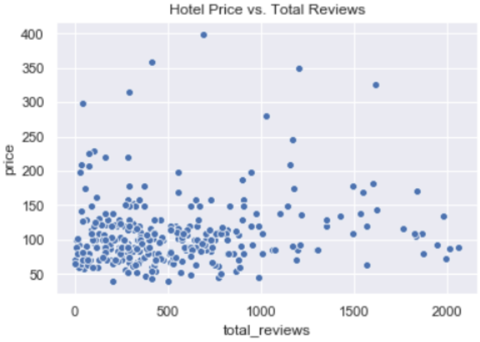


FIGURE 35. HOTEL PRICE AND TOTAL NUMBER OF REVIEWS WITHOUT OUTLIERS

DATA TRANSFORMATION

Linear regression has four assumptions which include: normality, homoscedasticity, linearity between predictor and target variables, and independent residuals. Linearity was examined in the above scatter plots, and normality was examined in the above distribution plots. Because continuous variables are typically not normally distributed, log transformation is applied to the predictor variables with the goal of achieving a better distribution. For example, the left plots below show the original distribution which have positive skewness. However, after price was transformed, the distribution of log values for price is normal which can be seen in the right plots below.

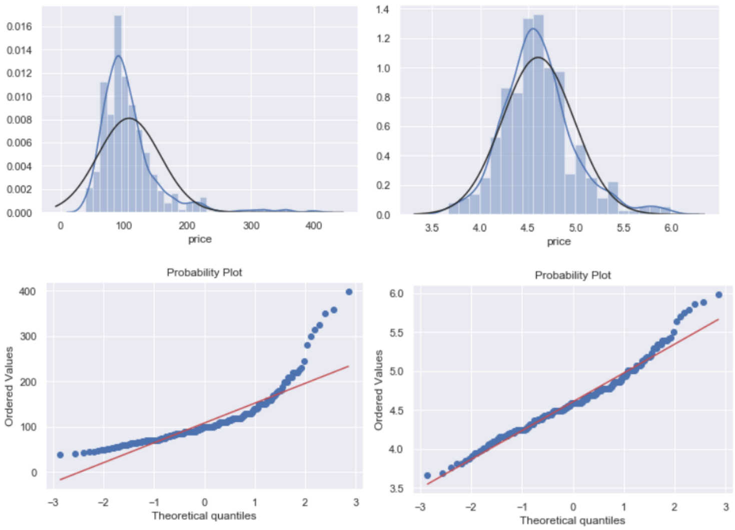
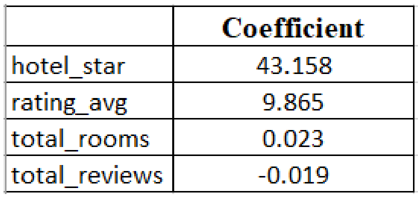


FIGURE 36. COMPARISON OF ORIGINAL AND TRANSFORMED TARGET VARIABLE: PRICE

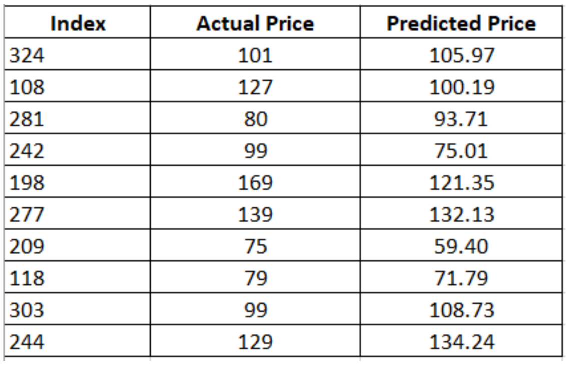
MODELING

The linear regression is fitted to predict the hotel price. As a result, the model got $R^2 = 55\%$ in the training dataset and $R^2 = 58\%$ in the validation dataset. The model explains 58% variance in the predicted price. Because of limited features, the model accuracy rate is not optimal. For future scope, it would be beneficial to add more records and features in order to improve the model's results. These additions could include things such as: location, type of listing (hotel, resort, villa, etc.), and the facilities and services provided by the hotel.



	Coefficient
hotel_star	43.158
rating_avg	9.865
total_rooms	0.023
total_reviews	-0.019

TABLE 7. COEFFICIENTS FROM THE MODEL



Index	Actual Price	Predicted Price
324	101	105.97
108	127	100.19
281	80	93.71
242	99	75.01
198	169	121.35
277	139	132.13
209	75	59.40
118	79	71.79
303	99	108.73
244	129	134.24

TABLE 8. SAMPLE OF PREDICTED VALUES FROM THE MODEL

TEXT MINING AND SENTIMENT ANALYSIS

Formatted: Heading 1

In this project, the most recent 100 (at maximum) reviews for 100 hotels in Chicago from Hotels.com was collected to create the datasets. There are a total of 9,488 records in the dataset, with hotel name, URL of the hotel, date the review was posted, rating from the review, and the review's text provided by the customers. In this section, the goal is to preform text analytics on reviews to understand the emotion of customers based on the reviews, find the possible factors that influence customers' emotions, and build an emotion classification model to classify the emotions of the customers based on their reviews.

The first step was tokenizing the review text and cleaning the unnecessary words/tokens. The stop words, numerical values, frequent words (such as "hotel", "chicago", "stay"), blank lines, tabs, and spaces from the review word tokens were removed since these words do not reflect the specific topics or sentiment of the customers. In addition, the words were stemmed to reduce the terms. After performing the text cleaning, the top 25 most frequent words are shown below:

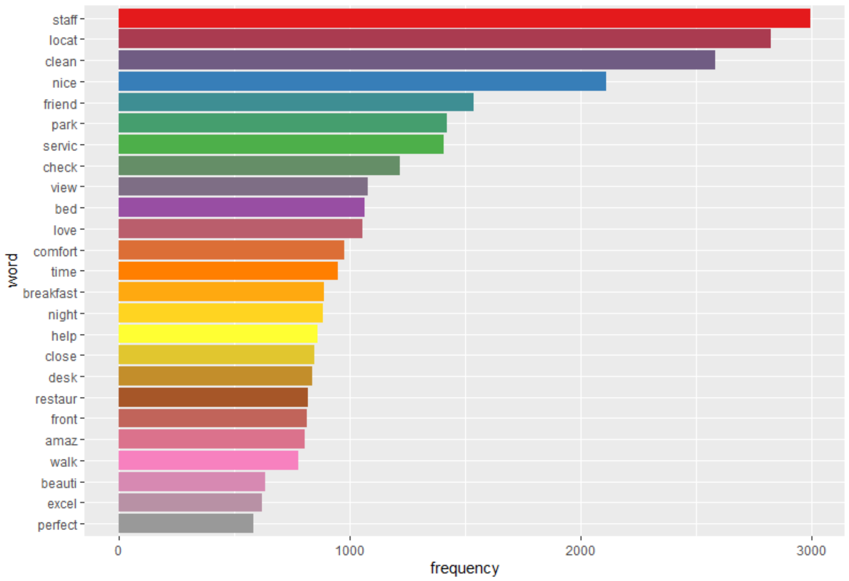


FIGURE 37. TOP 25 FREQUENT WORDS IN REVIEWS

In the figure above, the most frequent word is "staff" which occurred almost 3,000 times. This means that almost every three or four reviews contain the word "staff". Therefore, staff plays an important role in the customers' feelings about their hotel experience. Guests tend to have deep impressions related to the interaction with hotel staff and the services provided by the staff. Other words such as "locat" and "clean" suggest customers care about the location and the sanitary condition of hotel. Also, from this top word frequency chart, other import aspects guests evaluate the hotel on such as: parking, breakfast, and restaurant options are discussed. In the top frequent words, several words represent the sentiment of the customer including "nice", "love", "comfort", "excel", "perfect", etc. There do not appear to be any words related to negative sentiment. Overall, the reviews from the guests about their hotel experiences reveal that they are primarily satisfied and positive.

~~Before the analysis was conducted, data exploration was conducted. First, a plot with the most frequent words after changing all words to lower cases, removing stop words, numerical values, spaces, new lines, and tabs was created. Below are the results:~~

DATA CLEANING

As mentioned in the Data Preparation section of the report, the first step in the positive and negative sentiment analysis was to prepare the data for plotting and analysis. Before any plots were created, data cleaning was conducted. This included the following steps:

1. Removing stop words
2. Removing numerical values
3. Removing punctuations
4. Changing all words to lower cases
5. Removing uninformative frequent words
6. Stemming the words

DATA EXPLORATION

Based on the data exploration, it is revealed that the number of reviews has been significantly decreased in 2020. This may be due to the spread of the COVID-19 pandemic. Since the pandemic has negatively impacted the travel industry and reduced the number of trips being taking, it has consequently led to fewer reviews being written. The number of reviews revealed by date has been visualized in the following graph below:

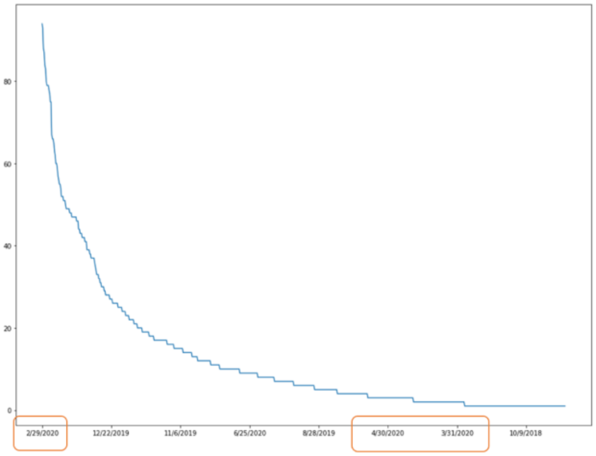


FIGURE 38. NUMBER OF REVIEWS BY DATE

Based on domain research, it was determined to select reviews with the rating above 8 (8 and 10) as Positive Reviews, and the ones with lower ratings (2, 4, and 6) as Negative Reviews. The distribution of Positive and Negative Reviews has been depicted in the following plots:

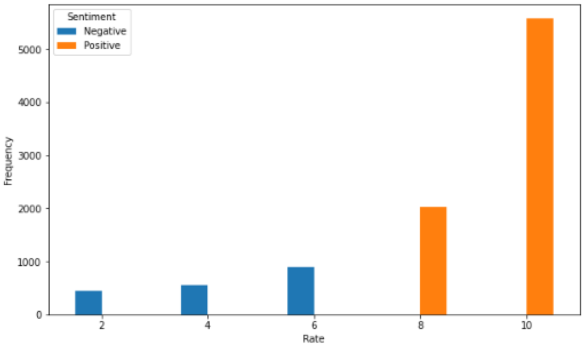


FIGURE 39. FREQUENCY OF POSITIVE AND NEGATIVE REVIEWS BY INDIVIDUAL RATES

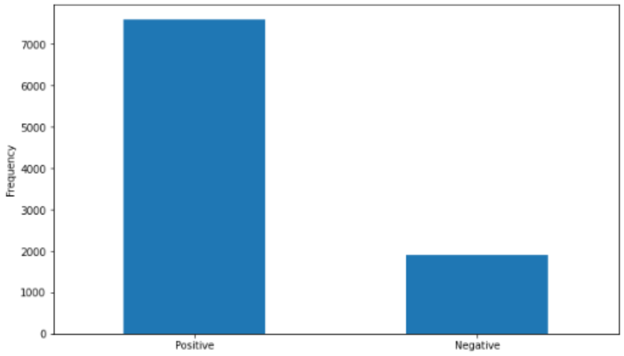


FIGURE 40. BAR GRAPH OF FREQUENCY OF POSITIVE AND NEGATIVE REVIEWS

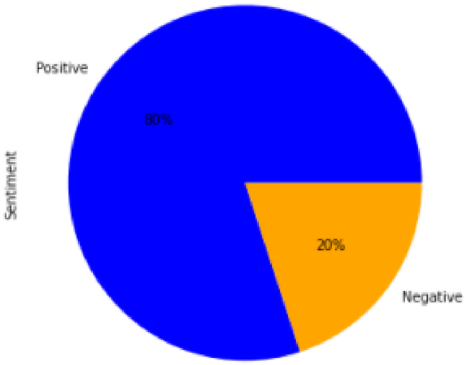


FIGURE 41. PIE CHART OF FREQUENCY OF POSITIVE AND NEGATIVE REVIEWS

Next, hotels were explored based on their sentiment. Looking at the figure below shows that nearly all of the hotels have substantially more positive reviews than negative ones.

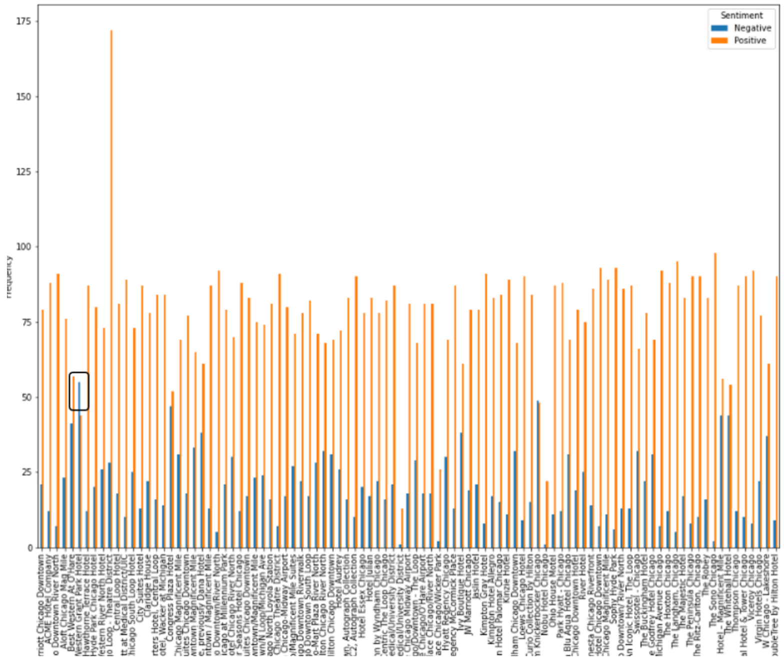
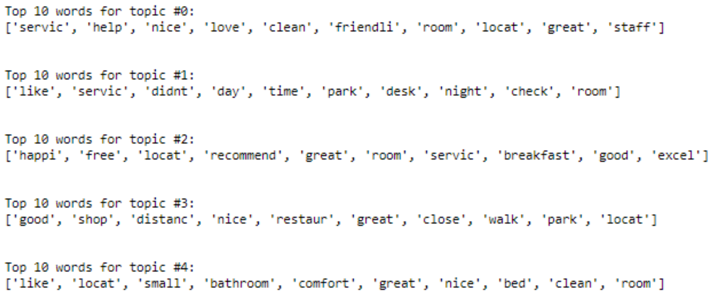


FIGURE 42. NEGATIVE AND POSITIVE SENTIMENT FOR INDIVIDUAL HOTELS

The graph above shows that nearly all hotels have more positive reviews than negative, except for a few exceptions (as shown above). Because looking at individual hotels is not within the scope of this study, future studies can conduct analysis about these few exceptional hotels.

TOPIC MODELING USING LDA

Five topics were derived from data using LDA Topic Modeling. These topics are:



```
Top 10 words for topic #0:
['servic', 'help', 'nice', 'love', 'clean', 'friendli', 'room', 'locat', 'great', 'staff']


Top 10 words for topic #1:
['like', 'servic', 'didnt', 'day', 'time', 'park', 'desk', 'night', 'check', 'room']


Top 10 words for topic #2:
['happi', 'free', 'locat', 'recommend', 'great', 'room', 'servic', 'breakfast', 'good', 'excel']


Top 10 words for topic #3:
['good', 'shop', 'distanc', 'nice', 'restaur', 'great', 'close', 'walk', 'park', 'locat']


Top 10 words for topic #4:
['like', 'locat', 'small', 'bathroom', 'comfort', 'great', 'nice', 'bed', 'clean', 'room']
```

FIGURE 43. LDA TOPICS

Topic 1 (#0): This topic is explaining the great services and the great staff who are nice, friendly, and helpful. It also mentions about the good location of their room and hotel and the cleanliness of the rooms. Guests seem to love their stays in the hotels.

Topic 2 (#1): This topic includes reviews about day and night check-in times. It also includes comments about the front desks, services, and parking, and some services which are available during the daytime, but are unavailable during the nighttime.

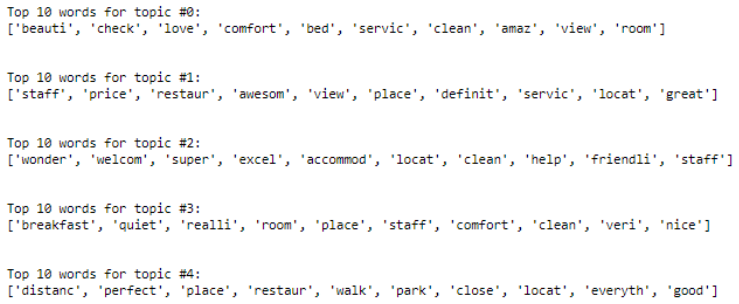
Topic 3 (#2): These reviews are about good breakfasts and services; they also include comments about excellent free services. It seems that people who commented here are willing to recommend the hotel.

Topic 4 (#3): This topic is about desired locations of the hotels, preferably in close distance to shopping centers and restaurants; it is also mentioned that walking-distance is preferred.

Topic 5 (#4): This topic incorporates comments about the comfortable facilities inside the room and its cleanliness. This topic includes great reviews about the beds and bathrooms of the rooms.

TOPIC MODELING USING NMF

Five topics were derived from data using NMF Topic Modeling. These topics are:



```
Top 10 words for topic #0:
['beauti', 'check', 'love', 'comfort', 'bed', 'servic', 'clean', 'amaz', 'view', 'room']


Top 10 words for topic #1:
['staff', 'price', 'restaur', 'awesom', 'view', 'place', 'definit', 'servic', 'locat', 'great']


Top 10 words for topic #2:
['wonder', 'welcom', 'super', 'excel', 'accommod', 'locat', 'clean', 'help', 'friendli', 'staff']


Top 10 words for topic #3:
['breakfast', 'quiet', 'realli', 'room', 'place', 'staff', 'comfort', 'clean', 'veri', 'nice']


Top 10 words for topic #4:
['distanc', 'perfect', 'place', 'restaur', 'walk', 'park', 'close', 'locat', 'everyth', 'good']
```

FIGURE 44. NMF TOPICS

Topic 1 (#0): This topic is explaining the beauty of the room, its cleanliness, and its comfortable beds.

Topic 2 (#1): This topic includes reviews about the awesome staff and services. It also has a large proportion of reviews about the location and room's view. Additionally, some reviews are about the price of the stay.

Topic 3 (#2): These reviews are mainly focused on staff and their behavior. It has reviews about staff who are friendly, helpful, and welcoming. It also includes guest experiences when expectations were exceeded, and special accommodations were met.

Topic 4 (#3): This topic includes comments about the breakfast, quiet atmosphere of the hotel, comfortableness, cleanliness of the rooms, and a very nice staff.

Topic 5 (#4): This topic incorporates comments about the hotels' locations. Additionally, reviews in this topic discuss being close to everything such as restaurants and shopping options. Reviews also include information about the hotels being in walking distance and guests not having to drive everywhere and pay for expensive parking.

LDA AND NMF RESULTS

The following graph shows that there are not many similarities between two topics derived using LDA and NFM. However, there are significant differences between two topics.

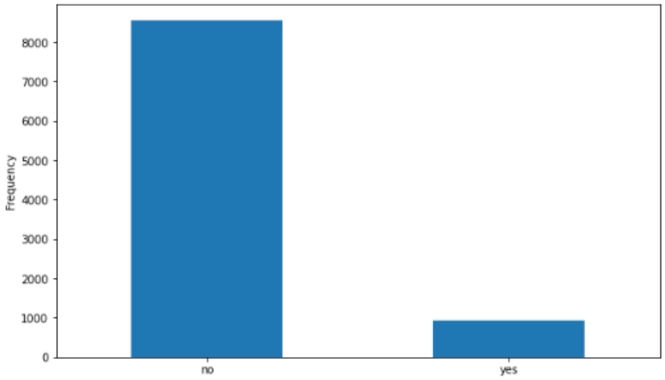


FIGURE 45. FREQUENCY OF DIFFERENCES BETWEEN LDA AND NFM

The following figure shows the distribution of reviews in topics using LDA:

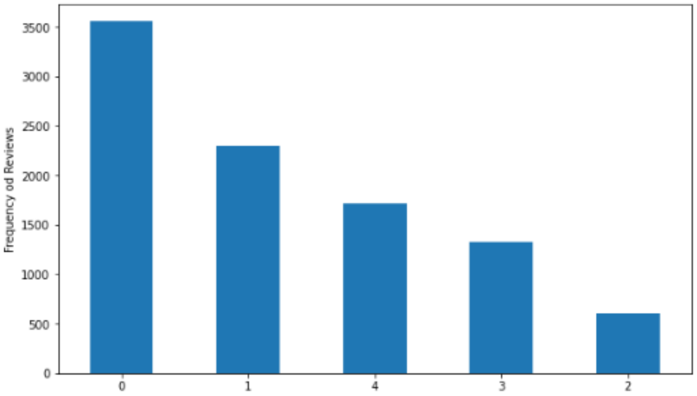


FIGURE 46. DISTRIBUTION OF REVIEWS FROM LDA TOPICS

The following figure shows the distribution of reviews in topics using NFM:

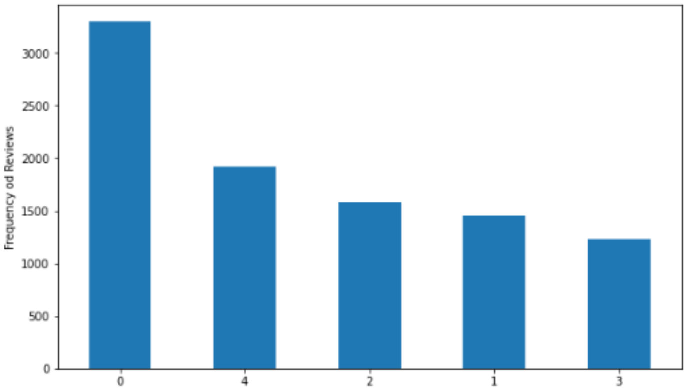
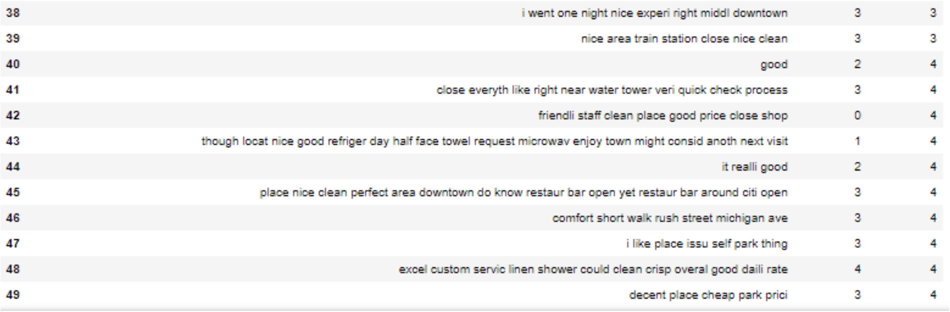


FIGURE 47. DISTRIBUTION OF REVIEWS FROM NMF TOPICS

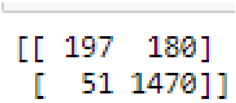
Topics then were assigned to each review, and a new data frame including topics was created. An example can be seen below:



38	i went one night nice experi right middl downtown	3	3
39	nice area train station close nice clean	3	3
40	good	2	4
41	close everyth like right near water tower veri quick check process	3	4
42	friendli staff clean place good price close shop	0	4
43	though locat nice good refriger day half face towel request microwav enjoy town might consid anoth next visit	1	4
44	it realli good	2	4
45	place nice clean perfect area downtown do know restaur bar open yet restaur bar around citi open	3	4
46	comfort short walk rush street michigan ave	3	4
47	i like place issu self park thing	3	4
48	excel custom servic linen shower could clean crisp overal good daili rate	4	4
49	decent place cheap park prici	3	4

FIGURE 48. EXAMPLE OF THE DATA FRAME INCLUDING THE TOPICS

Next, modeling is then done on sentiment, and the confusion matrix is achieved. The confusion matrix can be seen below:



```
[[ 197  180]
 [  51 1470]]
```

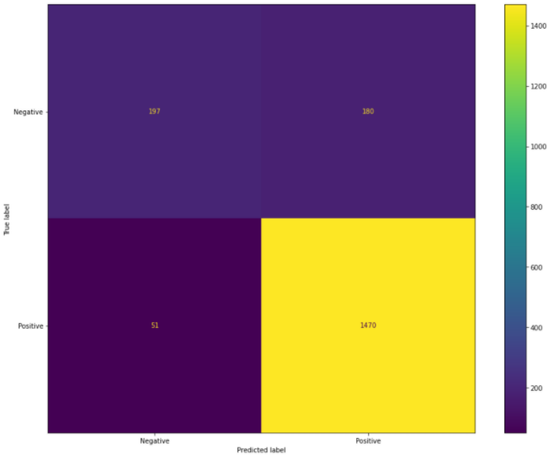
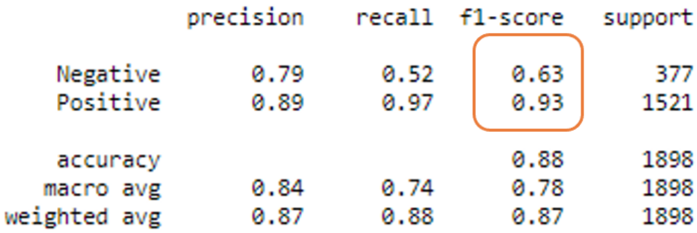


FIGURE 49. POSITIVE AND NEGATIVE SENTIMENT CONFUSION MATRIX

Classification report is then created and can be seen below:

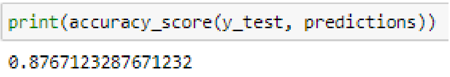


```
              precision    recall  f1-score   support

    Negative       0.79      0.52      0.63       377
    Positive       0.89      0.97      0.93      1521

    accuracy                           0.88      1898
   macro avg       0.84      0.74      0.78      1898
weighted avg       0.87      0.88      0.87      1898
```

FIGURE 50. POSITIVE AND NEGATIVE SENTIMENT CLASSIFICATION REPORT

The classification report shows perfect performance in predicting positive sentiments while poor performance when it comes to predicting negative sentiment.

The total accuracy is then achieved and can be seen below:



```
print(accuracy_score(y_test, predictions))
0.8767123287671232
```

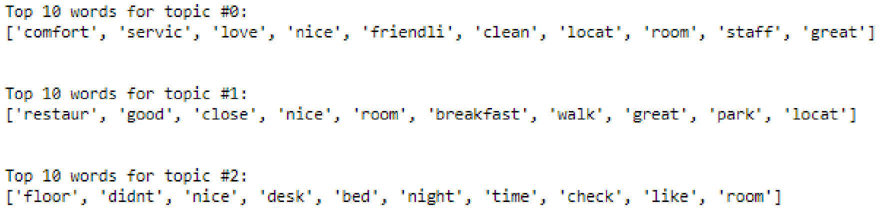
FIGURE 51. POSITIVE AND NEGATIVE SENTIMENT MODEL ACCURACY RATE

The accuracy rate is not very bad and is acceptable.

TOPIC MODELING USING LDA AND NMF FOR POSITIVE SENTIMENTS

All of the previous steps have been repeated on positive and negative sentiments separately to derive both positive and negative topics.

For Positive Sentiments, Rate >= 8. Three topics were derived from data. These topics for LDA are as follows:



```
Top 10 words for topic #0:
['comfort', 'servic', 'love', 'nice', 'friendli', 'clean', 'locat', 'room', 'staff', 'great']


Top 10 words for topic #1:
['restaur', 'good', 'close', 'nice', 'room', 'breakfast', 'walk', 'great', 'park', 'locat']


Top 10 words for topic #2:
['floor', 'didnt', 'nice', 'desk', 'bed', 'night', 'time', 'check', 'like', 'room']
```

FIGURE 52. POSITIVE LDA TOPICS

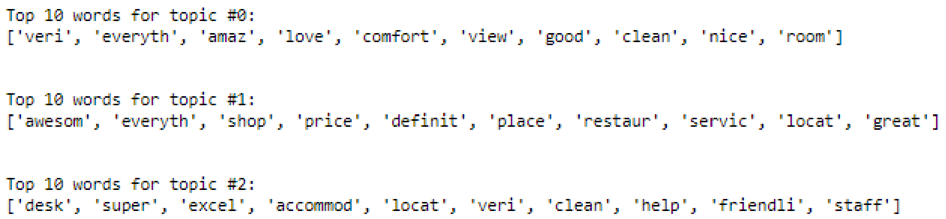
Topic 1 (#0): This topic is explaining the great services and the great staff who are nice, friendly, and helpful. It also mentions about the good location of their room and hotel and the cleanliness of the rooms. Guests seem to love their stays in the hotels.

Topic 2 (#1): This topic includes reviews about great breakfasts and parking sites. It also mentions reviews about the good location of the hotels which are locating close to restaurants and are accessible by walking.

Topic 3 (#2): This topic includes reviews about day and night check-in times. It also includes comments about the nice staff at the front desk.

After reviewing these three topics, it has been determined that they are nearly the same as the topics obtained in the LDA topic modeling using all of the data.

Next NMF topic modeling was used, and three topics were derived from data. These topics for NMF are as follows:



```
Top 10 words for topic #0:
['veri', 'everyth', 'amaz', 'love', 'comfort', 'view', 'good', 'clean', 'nice', 'room']


Top 10 words for topic #1:
['awesom', 'everyth', 'shop', 'price', 'definit', 'place', 'restaur', 'servic', 'locat', 'great']


Top 10 words for topic #2:
['desk', 'super', 'excel', 'accommod', 'locat', 'veri', 'clean', 'help', 'friendli', 'staff']
```

FIGURE 53. POSITIVE NMF TOPICS

Topic 1 (#0): This topic is explaining the beauty of the room, the room's cleanliness, and amazing views. It also includes the overall total satisfaction.

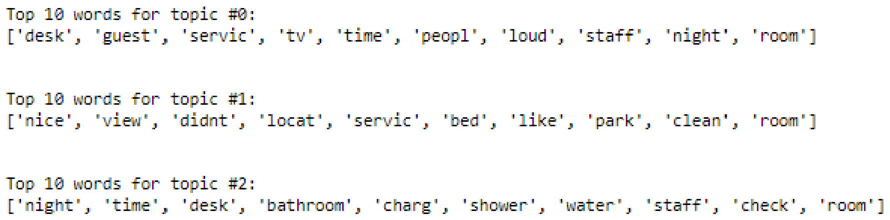
Topic 2 (#1): This topic includes reviews about the great location of the hotels which are close to shopping centers and restaurants.

Topic 3 (#2): These reviews are mainly focused on staff and their behaviors. It has reviews about staff who are friendly and helpful. It also includes guest experiences when expectations were exceeded, and special accommodations were met.

After reviewing these three topics, it has been determined that they are nearly the same as the topics obtained in the NMF topic modeling using all of the data.

TOPIC MODELING USING LDA AND NMF FOR NEGATIVE SENTIMENTS

For Negative Sentiments, Rate <= 6. Three topics were derived from data. These topics for LDA are as follows:



```
Top 10 words for topic #0:
['desk', 'guest', 'servic', 'tv', 'time', 'peopl', 'loud', 'staff', 'night', 'room']


Top 10 words for topic #1:
['nice', 'view', 'didnt', 'locat', 'servic', 'bed', 'like', 'park', 'clean', 'room']


Top 10 words for topic #2:
['night', 'time', 'desk', 'bathroom', 'charg', 'shower', 'water', 'staff', 'check', 'room']
```

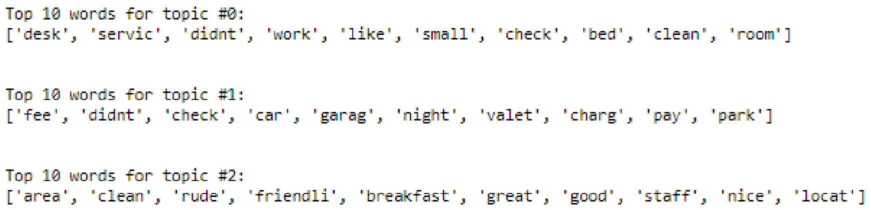
FIGURE 54. NEGATIVE LDA TOPICS

Topic 1 (#0): This topic includes reviews complaining about the noisy atmosphere, loud sounds, and TV's on high volume during nighttime which is annoying for guests trying to sleep.

Topic 2 (#1): This topic includes reviews about guests complaining about paying for the hotel parking which in some cases expensive. It also includes some comments about uncomfortable beds and being unclean.

Topic 3 (#2): This topic includes problems with bathroom amenities, and staff who had delays in solving issues.

Next NMF topic modeling was used, and three topics were derived from data. These topics for NMF are as follows:



```
Top 10 words for topic #0:
['desk', 'servic', 'didnt', 'work', 'like', 'small', 'check', 'bed', 'clean', 'room']


Top 10 words for topic #1:
['fee', 'didnt', 'check', 'car', 'garag', 'night', 'valet', 'charg', 'pay', 'park']


Top 10 words for topic #2:
['area', 'clean', 'rude', 'friendli', 'breakfast', 'great', 'good', 'staff', 'nice', 'locat']
```

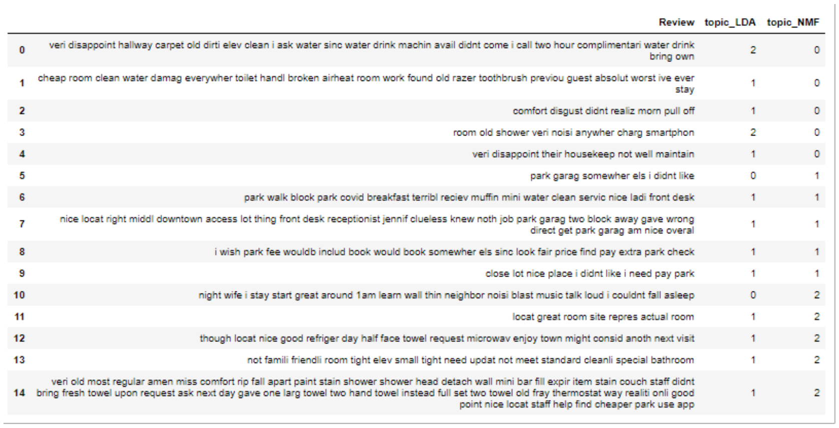
FIGURE 55. NEGATIVE NMF TOPICS

Topic 1 (#0): This topic includes issues about room cleanliness, being messy, and very small, tight rooms.

Topic 2 (#1): This topic includes comments about being charged for hotel parking or having to park in a different parking garage location with expensive rates, especially during the night.

Topic 3 (#2): These reviews are mainly focused on staff and their behavior. Some comments here evaluate staff as rude and unfriendly.

Topics then were assigned to each review, and a new data frame including topics was created. An example of Negative Reviews and their related topics has been shown here:



	Review	topic_LDA	topic_NMF
0	veri disappoint hallway carpet old dirti elev clean i ask water sinc water drink machin avail didnt come i call two hour complimentari water drink bring own	2	0
1	cheap room clean water damag everywher toilet handl broken airheat room work found old razer toothbrush previou guest absolut worst ive ever stay	1	0
2	comfort disgust didnt realiz morn pull off	1	0
3	room old shower veri noisi anywher charg smartphon	2	0
4	veri disappoint their housekeep not well maintain	1	0
5	park garag somewher els i didnt like	0	1
6	park walk block park covid breakfast terribl reciev muffin mini water clean servic nice ladi front desk	1	1
7	nice locat right middl downtown access lot thing front desk receptionist jennif clueless knew noth job park garag two block away gave wrong direct get park garag am nice overal	1	1
8	i wish park fee wouldb includ book would book somewher els sinc look fair price find pay extra park check	1	1
9	close lot nice place i didnt like i need pay park	1	1
10	night wife i stay start great around 1am learn wall thin neighbor noisi blast music talk loud i couldnt fall asleep	0	2
11	locat great room site repres actual room	1	2
12	though locat nice good refriger day half face towel request microwav enjoy town might consid anoth next visit	1	2
13	not famili friendli room tight elev small tight need updat not meet standard cleanli special bathroom	1	2
14	veri old most regular amen miss comfort rip fall apart paint stain shower shower head detach wall mini bar fill expir item stain couch staff didnt bring fresh towel upon request ask next day gave one larg towel two hand towel instead full set two towel old fray thermostat way realiti onli good point nice locat staff help find cheaper park use app	1	2

FIGURE 56. EXAMPLE OF THE DATA FRAME INCLUDING THE TOPICS FOR NEGATIVE SENTIMENT

~~DATA EXPLORATION FOR~~ EMOTIONS SENTIMENT ANALYSIS

~~In this project, the most recent 100 (at maximum) reviews for 100 hotels in Chicago from Hotels.com was collected to create the datasets. There are a total of 9,621 records in the dataset, with hotel name, URL of the hotel, date the review was posted, rating from the review, and the review's text provided by the customers. In this section, the goal is to preform text analytics on reviews to understand the emotion of customers based on the reviews, find the possible factors that influence customers' emotions, and build an emotion classification model to classify the emotions of the customers based on their reviews.~~

JOY AND SADNESS SENTIMENT

EMOTION ANALYSIS

The NRC lexicon is an unigrams emotion lexicon. It has the dictionary of the words and their corresponding sentiment/emotion. After inner-joining the lexicon with the word tokens, there are a total of 13,297 words assigned with the joy or sadness emotion. 11,072 words were assigned with the joy emotion, and 2,225 words were assigned with the sadness emotion.

In the joy emotion word cloud seen below, high frequency words include: “clean”, “friend”, “comfort”, “food”, “enjoy”, and “love”. Therefore, it can be concluded that when the hotel is clean and comfortable, customers always feel joy and happiness. Other high frequency words assigned to the joy emotion are words such as “safe” and “spa” which most likely indicates some aspects of hotel services that will make customers enjoy their experience.

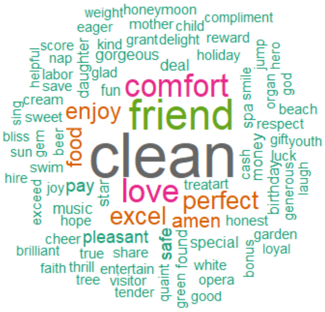


FIGURE 57. JOY EMOTION WORD CLOUD

The words classified with the sadness emotion include: “disappoint”, “quiet”, and “bad”. These words are the highest frequency words in this emotion group and can be seen in the word cloud below. In addition, words such as “broken”, “late”, “trash”, and “cancel” are also relatively high frequency words. This suggests that hotels need to provide the best services and keep the environment clean to satisfy the hotel guests.

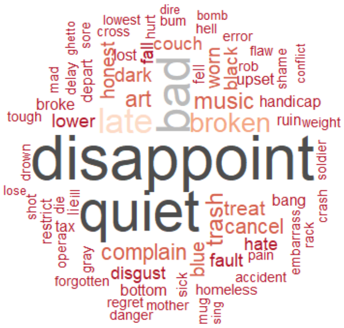


FIGURE 58. SADNESS EMOTION WORD CLOUD

Additionally, contentment was used to show the difference between joy and sadness, and the number of times specific words occurred in the reviews which represents the degree of emotion found in a specific word. The contentment diagram shown below reveals the top ten and bottom 10 contentment words found in the reviews. Similar to the words cloud, words such as “clean”, “comfort”, “friend”, and “love” have high contentment while words such as “disappoint”, “quiet”, and “bad” have low contentment scores.

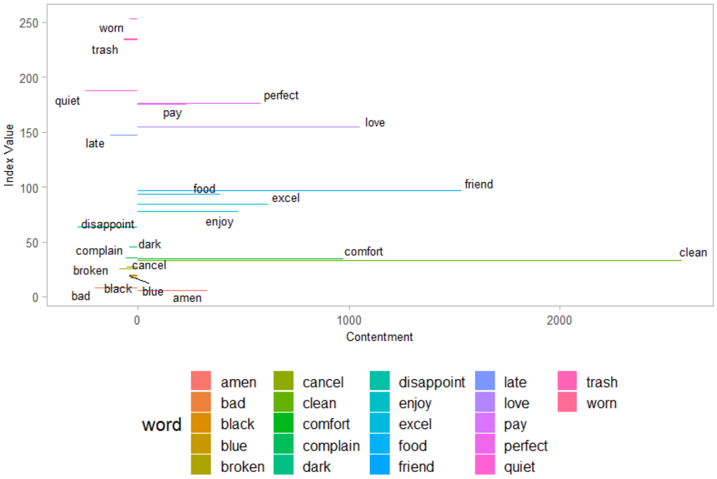


FIGURE 59. TOP AND BOTTOM CONTENTMENT WORDS OF JOY AND SADNESS EMOTIONS

After checking the contentment of each words, the emotion score of each review was calculated. The emotion score is the total counts of joy words in a review minus the total counts of sadness words in that review. If a review has a positive emotion score, it was classified as a joy emotion, else it was classified as a sadness emotion. There are a total 6,473 out of the ~~9,621~~ 9488 records containing the words that overlap with the joy and sadness words in NRC lexicon. Among the 6,473 reviews in the figure below, 5,406 reviews are assigned as the joy emotion, and 1,067 reviews are assigned as the sadness emotion.

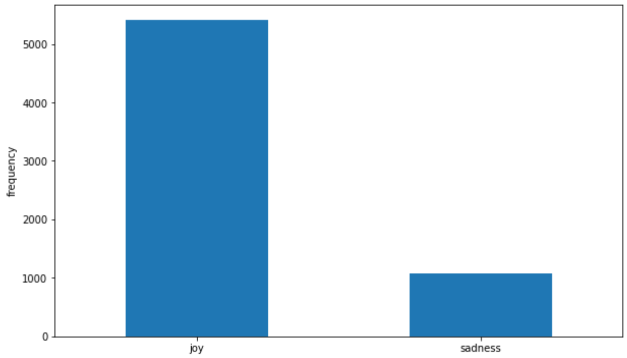


FIGURE 60. THE TOTAL VOLUME OF JOY REVIEWS AND SADNESS REVIEWS

From the hotel reviews, the top three hotels with the most joy classified review are the Silversmith Hotel Chicago Downtown, The Sono Chicago, and Satypineapple. These hotels have 70 or more joy classified reviews out of the most recent 100 reviews. This information can be seen in the graph below:

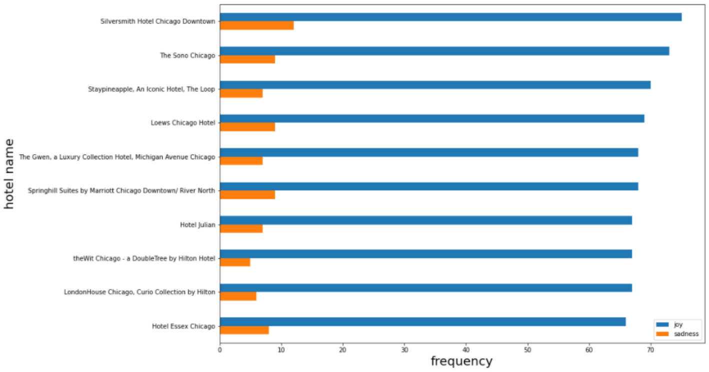


FIGURE 61. TOP 10 HOTELS WITH THE HIGHEST JOY EMOTION REVIEWS

The top three hotels with the most sadness classified reviews are the Best Western A O'Hare, The Whitehall Hotel, and The Buckingham Hotel. They have 20 or more sadness classified reviews out of the most recent 100 reviews. This information can be seen in the graph below:

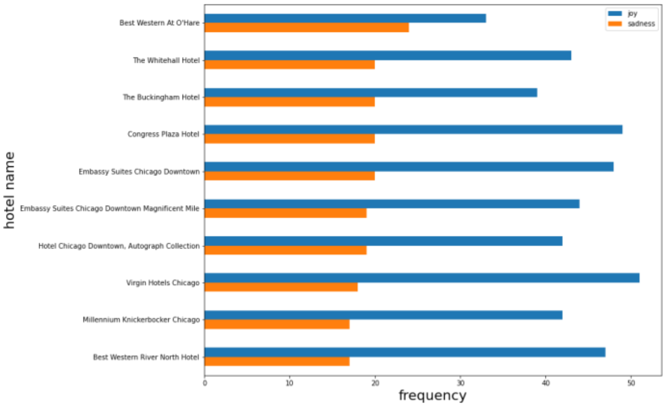
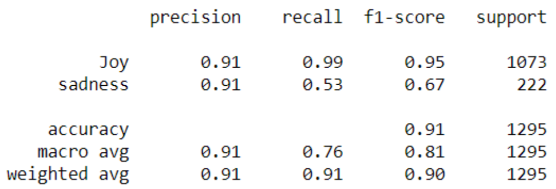


FIGURE 62. TOP 10 HOTELS WITH THE HIGHEST SADNESS EMOTION REVIEWS

CLASSIFICATION MODEL

Text from the guests' reviews was input as features to predict the emotion of a review. Text cleaning was conducted on the reviews' text. This included changing the text into lower case, removing the following: stop words, numerical values, punctuation, and common words (such as hotel, stay, and Chicago), and stemming the words. A TF-IDF Matrix for the review text was built with 2,500 maximum features and 7 absolute counts as minimum document frequency and 0.8 proportion of the document as maximum document frequency, with 7 minimum to 2,500 maximum features and 0.8 maximum document frequency. The The original data was split into 80% for training and 20% for validation, and the random forest classifier model was used to build the classification model. The overall accuracy of the model was 0.91 on the validation data. In the classification report below, precisions of joy and sadness are 0.91. The F1-Score for joy is 0.95 and the F1-Score for sadness is 0.67.



```
              precision    recall  f1-score   support

         Joy       0.91      0.99      0.95      1073
     sadness       0.91      0.53      0.67       222

    accuracy                           0.91      1295
   macro avg       0.91      0.76      0.81      1295
weighted avg       0.91      0.91      0.90      1295
```

FIGURE 63. MODEL CLASSIFICATION REPORT FOR JOY AND SADNESS EMOTIONS

In the confusion matrix below, more than half of the true sadness reviews have been classified as joy in the model which might be due to the imbalance problem of the data. In the future scope of the project, balancing the data and trying different models can be done to improve the model's overall performance.

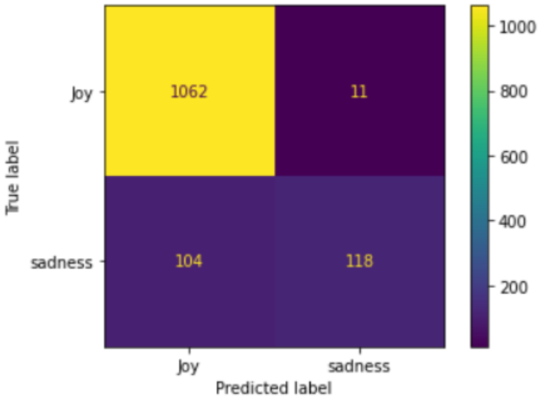


FIGURE 64. MODEL CONFUSION MATRIX FOR JOY AND SADNESS EMOTIONS

ANTICIPATION AND ANGER SENTIMENT

EMOTION ANALYSIS

In this section, the major terms in the reviews that related to customer satisfaction about their experience is analyzed. After inner-joining the anticipation and anger sentiment words listed in NRC lexicon with current words in reviews, there are 9,442 words in reviews assigned with these emotions. There are 7,162 words that are assigned with the anticipation emotion and 2,280 words that are assigned with the anger emotion. The following word clouds below show the top frequent words classified in the anticipation emotion and the anger emotion.

The most frequent words in anticipation emotion are as follows: “comfort”, “time”, “enjoy”, “perfect”, “excel”, and “pleasant”, etc. Therefore, when people feel happy or comfortable in the hotel, they are satisfied. This can be seen in the word cloud below:

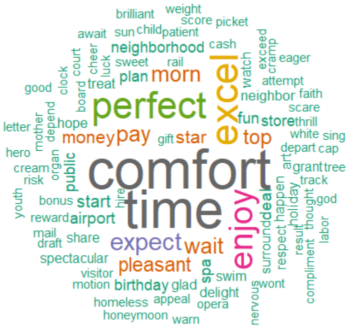


FIGURE 65. ANTICIPATION EMOTION WORD CLOUD

The most frequent words classified in the anger emotion word cloud are as follows: “disappoint”, “shock”, “bark”, “brutal”, “upset”, “bummer”, and “disrespect”, etc. These words indicate the customers probably went through an argument or were not satisfied with the hotel services and they feel angry about it. This can be seen in the word cloud below:

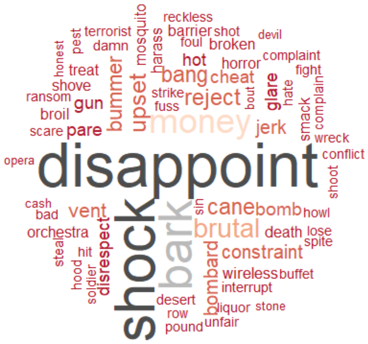


FIGURE 66. ANGER EMOTION WORD CLOUD

Additionally, contentment was used to show the difference between anticipation and anger, and the number of times specific words occurred in the reviews which represents the degree of emotion found in a specific word. If a word has a high contentment score, it means that word likely makes the customer feel anticipation, and if a word has a high negative contentment score, that word is associated with the anger emotion of the customer. Similarly, as shown in the word clouds above, the words “comfort”, “time”, “excel”, “perfect”, and “enjoy” are the top words in the anticipation emotion sentiment. This suggests that when customers feel the hotel environment is comfortable, they are highly satisfied. The words with lowest contentment scores are “disappoint”, “smell”, “bad”, “hot”, and “complaint”, etc. From these words, it can be assumed that when customers are disappointed with the hotel’s conditions or services, the customer will be dissatisfied and angry with their experience. Additionally, the smell most likely indicates that customers care about the air quality and smell in the hotel and their room. If the hotel has bad smell or smoke smell, they will most likely be unhappy. This information can be seen in the graph below:

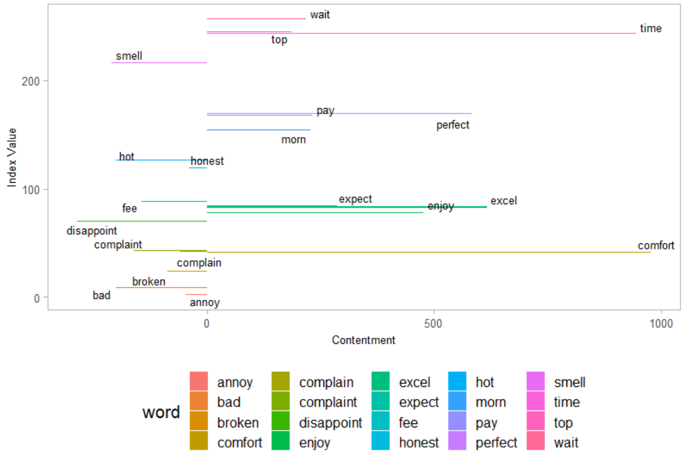


FIGURE 67. CONTENTMENT DIAGRAM OF ANTICIPATION AND ANGER EMOTIONS

The counts of the words of the anticipation emotion and the anger emotion for each review were then checked. The emotion score is used to represent the difference of the count of the anticipation emotion word and the count of anger emotion word in a review. If a review has a positive score, it means this review is prone to the anticipation emotion and is classified as anticipation. On the other hand, the review could be classified as the anger emotion. There are a total 4,994 records assigned with the anticipation or anger emotions. Specifically, 3,785 records are labeled as anticipation, and 1,209 records are labeled as anger.

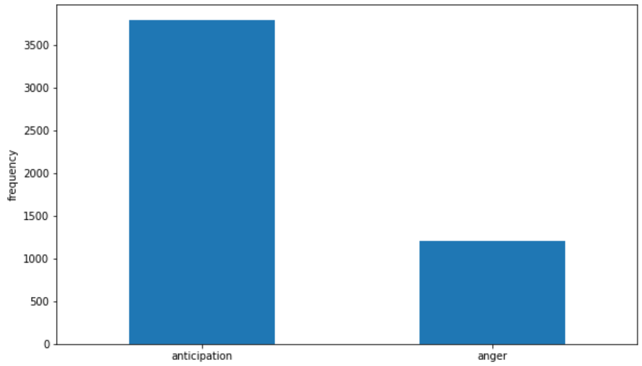


FIGURE 68. NUMBER OF ANTICIPATION REVIEWS AND ANGER REVIEWS IN OUR DATASET

After checking the anticipation and anger reviews distribution by hotel, it was revealed that the Langham, Chicago, Found Hotel Chicago River North, and The Sono Chicago has the highest number of anticipation emotion reviews. This can be seen in the plot below:

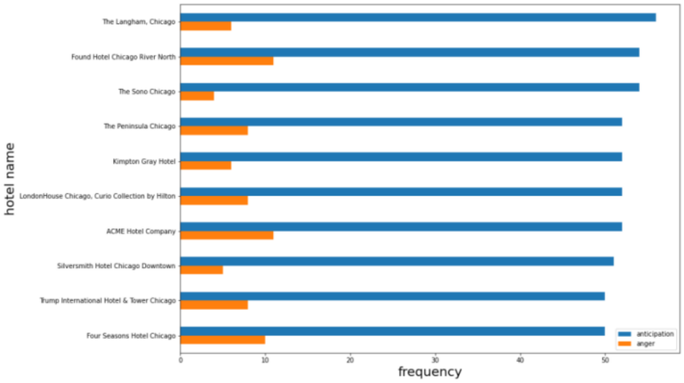


FIGURE 69. TOP HOTELS WITH THE HIGHEST AMOUNT OF ANTICIPATION EMOTION REVIEWS

The Whitehall Hotel, Hotel Audrey, and Embassy Suites Chicago Downtown Magnificent Mile are the top three hotels receiving the highest number of complaints and highest number of anger emotion reviews. This can be seen in the plot below:

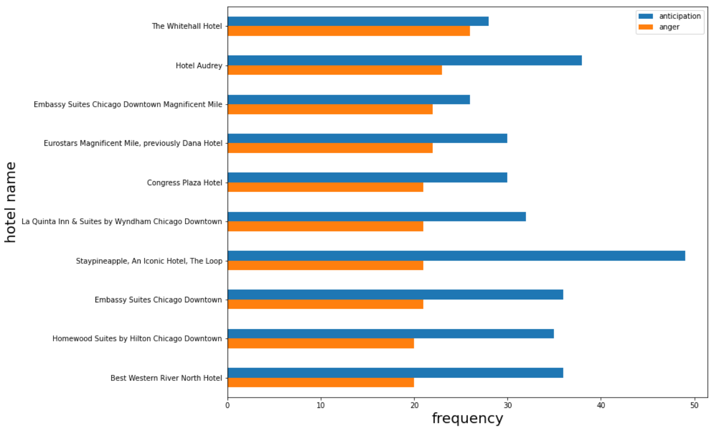
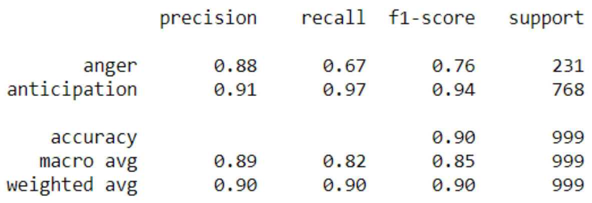


FIGURE 70. TOP HOTELS WITH THE HIGHEST AMOUNT OF ANGER EMOTION REVIEWS

CLASSIFICATION MODEL

Text from the guests' reviews was input as features to predict the emotion of a review. After text cleaning was conducted, A TF-IDF Matrix for the review text was built with ~~7 minimum to~~ 2,500 maximum features and 7 absolute counts as minimum document frequency and 0.8 proportion of the document as maximum document frequency. The original data was split into 80% for training and 20% for validation, and the random forest classifier model was used to build the classification model. The overall accuracy of the model was 0.90 on the validation data. In the classification report below, precisions of anger and anticipation are 0.91. The F1-Score for anger is 0.76 and the F1-Score for anticipation is 0.94.



```
              precision    recall  f1-score   support

       anger       0.88      0.67      0.76       231
anticipation       0.91      0.97      0.94       768

    accuracy                           0.90       999
   macro avg       0.89      0.82      0.85       999
weighted avg       0.90      0.90      0.90       999
```

FIGURE 71. MODEL CLASSIFICATION REPORT FOR ANTICIPATION AND ANGER

This model has a good performance in classifying the anger and anticipation emotions. In the model confusion matrix below, it is revealed that the model classifies the anticipation emotion very well. However, for the anger emotion, some of the true emotion reviews are classified as anticipation by the model. This model had better classification results than the joy and sadness model did. One of the reasons for this may be that the word list for the anticipation emotion is very different from the sadness word list. Therefore, the classification model has a better performance when classifying anticipation and anger emotions.

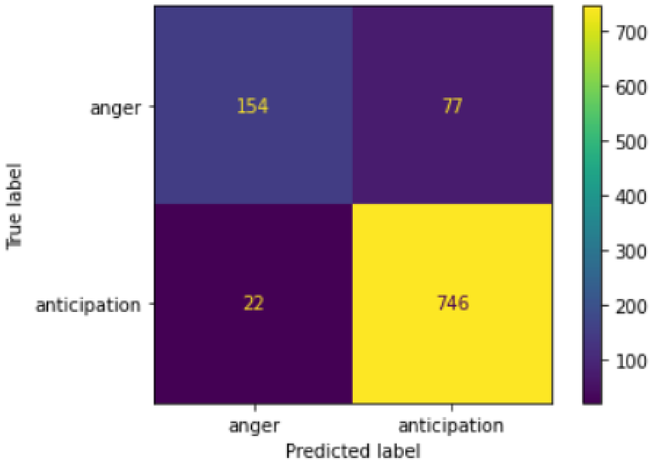


FIGURE 72. MODEL CONFUSION MATRIX FOR THE ANTICIPATION AND ANGER EMOTIONS

CONCLUSION

In conclusion, the datasets were able to provide insights about what guests expect when staying at a hotel. Things such as cleanliness, staff attitudes, price, and location are all influential factors in how a guest feels about their hotel experience. For future scope of the project, it would be beneficial to collect additional data from other major cities along with Chicago and see if there are any key differences among the cities. Additionally, it would be beneficial to add more records and features in order to improve the linear regression model's results. These additions could include things such as: specific location, type of listing (hotel, resort, villa, etc.), and the facilities and services provided by the hotel. For the classification model results, it would be beneficial to balance the data and try different models, aside from the random forest classifier model that was used, to improve the model's overall performance.