Group Assignment

Data 630 9040

Due Date: August 12th, 2021

UMGC, Summer Semester 2021

Professor Firdu Bati

Team Members

Samuel Adebayo

Tegan Classens (Coordinator)

Melissa Hunfalvay

Patrick Mugwanja

Introduction

Objective

The dataset used for this project was the Hotel Reviews data provided by Datafiniti's Business Database (Kaggle, 2019). The dataset included information on hotels throughout the world including textual reviews by patrons who stayed at the hotel. The data was collected between 2002 and 2017.

The objective of the analysis was to use text mining and Natural Language Processing (NLP) methods to determine some insights into the reviews, as well as the sentiment (positive or negative) of patrons who stayed at the hotels, both on an overall level, but also at a specific province level for a few of the provinces in the dataset. More specifically, the purpose is to determine which hotels to recommend and which to avoid.

A second analysis performed on the data is cluster analysis. Clustering of data occurs via mathematical algorithms that allow the hotels to understand what items within the data are similar to one another, therefore forming a cluster. Clustering also allows the hotel owners to understand what differs between the clusters (Han, Kamber, Pai, 2011).

Both analyses are forms of unsupervised learning techniques that are designed to provide insights into the trends, patterns, and grouping of the textual content. Unsupervised learning does not have a ground truth or correct outcome, instead it finds insights and common trends to help the hotels know where to focus their attention in order to satisfy their customers.

Problem Domain

Hotels operate on thin margins, according to world's leading hotel industry experts (Mandelbaum, 2020). The profit margin is heavily influenced by occupancy rates. A 39.8% annual occupancy produces an 11.6% profit margin.

In order to keep profitable, hotel owners focus on occupancy rates. One of the most influential factors in occupancy rates is online customer reviews.

Customers choose hotels based on the hotel reputation which they find on social media on sites such as Yelp and Tripadvisor. These are highly ranked platforms that provide independent reviews that can "destroy or market" a hotel's reputation, according to Linchpin. This is big business, as a one-star increase in reviews has meant 5-9% increase in revenue (StayinTouch, n.d.)!

Extracting patron feedback on certain features of the hotel, to include health and safety precautions such as cleanliness in the time of COVID is important. Families traveling for summer vacations will likely be looking for clean, family-friendly hotels that provide safe and comfortable lodging. Poor customer reviews on cleanliness will likely decrease occupancy rates.

Reviews on amenities such as breakfast, internet, wellness facilities and services can be important to different types of travelers. For example, business travelers in today's world rely on fast speed and reliable internet. Customer reviews on these services for the business traveler will influence which hotel they choose.

Even though the data in this dataset was collected prior to COVID, it is relevant today. Post-COVID the primary challenge of the hotel industry is customer satisfaction (Linchpin, n.d.). According to Linchpin, a leading marketing strategy website, the primary challenges for hotels in fulfilling customer satisfaction include:

1. *Electronic check-in:* which includes a text-message conveying the customer's room number with electronic key card for access.
2. *International travelers:* have different demands that include hotels providing members that speak multiple languages fluently. This is big business, and

technology giant Google recently launched earbuds which translate 40 different languages that could assist hotels in managing this challenge.

3. *Personalized experience:* customers look for unique designs and locally sourced food and drink. In the era of Airbnb, this is becoming a very competitive challenge for hotels.
4. *Loyalty Programs:* that include discounts and specials are influential factors in customer satisfaction.

Understanding feedback from customers is critical to a hotel's business. The feedback can make the difference between being in-business or out-of-business, and between surviving and thriving as a hotel business. Therefore, the objective of this analysis is to find recurring themes and trends in the data that affect customer satisfaction levels, and ultimately influence occupancy rates and profit margins for the hotel industry.

Method Rationale

The methodology chosen was text classification and cluster analysis. These are statistical techniques and forms of unsupervised learning.

The rationale for *unsupervised learning* methodology includes:

a) The lack of a known “truth” or outcome from which the data has a correct answer

b) The need to explore the data for trends, patterns, groupings and insights (Ng, 2021)

c) The size of the data is large which lends itself to unsupervised learning

The rationale for using text classification methodology includes:

a) The need to extract features within the unstructured patron reviews that represent the hotels

b) The desire to interpret meaning from the reviews and extract concepts such as “I like…” or “I don’t like…” (positive or negative) in order to inform the recommendation to stay at the hotel or look elsewhere

The rationale for using *cluster analysis* includes:

a) The need to group data via the use of a mathematical algorithm (Machine Learning, n.d.)

b) The desire to find groupings within the data to inform the hotel industry.

c) The research question is looking to identify patterns in the data (Maimon & Rokach, 2010).

d) The nature of the problem is to inform the hotel industry of factors that may be influencing occupancy rates and ultimately profit margins, therefore, an analysis that provides groupings, patterns, and trends within the data will help inform the industry.

The hotel owners, hotel industry executives and customers who are patrons of the hotels all want to ask questions of a dataset like this to include:

1. What are the most common (frequent) occurrences, good or bad, being expressed by previous customers?

2. Are these insights of interest to me as a business traveler? Or for our family vacation? Or as an international traveler?

3. As a hotel owner, what are the things we are doing well? What are the services customers like that keep them returning?

4. By understanding these insights, how may the hotel owners mitigate concerns of the patrons? Changes can lead to increased ratings, increased occupancy and increased profit margin. Mitigating strategies may take the form of:

    a. Faster check-in

    b. Changes in reward programs

    c. Changes in menu items at the hotel restaurant

    d. Changes in cleaning routines or cleaning supervision

Analysis

Data

This dataset was a list of 1000 hotel reviews. The version of data used was version 5, collected between 2002 and 2017, and owned by Datafiniti (Kaggle, 2019). The data was collected from hotels located throughout the world.

The goal of the data was to examine the textual reviews and ratings of the hotels in order to inform members of the hotel industry, owners of the hotels, and patrons who visit the hotels. This information can be used to provide insights into features of importance for different groups of travelers. Furthermore, the feedback can inform hotels as to what is important to travelers, what the hotels are doing well, and what needs improvement.

The number of observations in this dataset was 35,912. Each observation was unique, in that each observation was from a different traveler. There were 19 variables in the data.

Location-related variables included the hotel's name, street address (address), city, country, latitude, longitude, postal code (postalCode), and province.

The variable (categories) revealed various types of hotels, for example: hotels, motels, lodging, and corporate lodging.

Variables related to customer reviews included the date of the review (reviews.date); the date the review was added to the dataset (reviews.dateAdded); the recommendation of the review (reviews.doRecommend); and the identification of the review (reviews.id). Information on the reviewer included: the reviewers' province (reviews.userProvince), username (reviews.username), and the reviewers' city (reviews.userCity). A rating, between zero (poor) and ten (excellent) was also included (reviews.rating). Textual components of the reviews included the reviewers' comments (reviews.text) and the title of the review (reviews.title).

Exploratory Analysis

First, the str command was ran in order to see the different data types of the variables:

```
> str(HotelReviews)
'data.frame':   35912 obs. of  19 variables:
 $ address             : Factor w/ 999 levels "1 Main St","1 Miracle Strip Pkwy Se",..: 973 973 973 973 973 973 973 973 973 973 ...
 $ categories          : Factor w/ 396 levels "Accommodation Reservations,Hotel & Motel Reservations,Hotels,Accommodations & Lodging,Motel
s",..: 121 121 121 121 121 121 121 121 121 121 ...
 $ city                : Factor w/ 761 levels "Abbeville","Aberdeen",..: 429 429 429 429 429 429 429 429 429 429 ...
 $ country             : Factor w/ 1 level "US": 1 1 1 1 1 1 1 1 1 1 ...
 $ latitude            : num  45.4 45.4 45.4 45.4 45.4 ...
 $ longitude           : num  12.4 12.4 12.4 12.4 12.4 ...
 $ name                : Factor w/ 879 levels "1785 Inn","1900 House",..: 449 449 449 449 449 449 449 449 449 449 ...
 $ postalCode          : Factor w/ 912 levels "","05156-9127",..: 186 186 186 186 186 186 186 186 186 186 ...
 $ province            : Factor w/ 287 levels "AK","AL","Andyville",..: 85 85 85 85 85 85 85 85 85 85 ...
 $ reviews.date        : Factor w/ 3010 levels "","2002-05-16T00:00:00Z",..: 1526 2224 1777 1562 2174 2226 1805 2286 2469 2290 ...
 $ reviews.dateAdded   : Factor w/ 1029 levels "2015-01-28T14:40:46Z",..: 529 529 529 529 529 529 529 529 529 529 ...
 $ reviews.doRecommend : logi  NA NA NA NA NA NA ...
 $ reviews.id          : logi  NA NA NA NA NA NA ...
 $ reviews.rating      : num  4 5 5 5 5 5 4 4 3 4 ...
 $ reviews.text        : Factor w/ 34399 levels "","- . 80,86 . , 0,33 , . )",..: 18360 18869 4360 32349 32349 31603 14239 17430 6000 13711
...
 $ reviews.title       : Factor w/ 21964 levels "","' Old but good '",..: 7437 8590 12442 7470 11085 20441 12409 12409 7028 17483 ...
 $ reviews.userCity    : Factor w/ 2898 levels "","12582","94503",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ reviews.username    : Factor w/ 15493 levels "","'Kim L","@AFRomero_",..: 12589 375 9780 7487 13707 375 375 375 4188 375 ...
 $ reviews.userProvince: Factor w/ 649 levels "","a","Afton mn",..: 1 1 1 1 1 1 1 1 1 1 ...
```

Figure 1: Structure of the Hotel Reviews Dataset

Of the 19 variables in the dataset, 14 were character data types, three were numeric data types, and two were logical data types. The two variables with logical data types - reviews.doRecommend and reviews.id - only contained null values, but that will be explored further later in the analysis.

```
> apply(newdata, 2, function(x) length(unique(x)))
             address           categories                 city              country             latitude
                 999                  396                  761                    1                  983
           longitude                 name           postalCode             province         reviews.date
                 984                  879                  912                  287                 3010
   reviews.dateAdded  reviews.doRecommend           reviews.id       reviews.rating         reviews.text
                1029                    1                    1                   44                34400
        reviews.title     reviews.userCity     reviews.username reviews.userProvince
                21964                 2898                15494                  649
```

Figure 2: Unique Values per Variable

The following command was run in R - apply(HotelReviews, 2, function(x) length(unique(x))) - in order to identify how many unique values each variable contained. As mentioned, the two variables with logical data types (reviews.doRecommend and reviews.id) only have one unique value, which is the null value. The country variable also has only one unique value, which is the US, or United States. The rest of the variables have a variety of unique values, ranging from a low of 44 unique values for the reviews.rating variable to a high of 34,400 unique values for the reviews.text variable, which represents the actual text of the hotel reviews.

Most of the variables are character data types. Therefore, the summary command does not provide a lot of information. The summary command does provide the length (35,912, which is the number of observations), and the class and mode, which are both “character” variables. For the two variables with logical data types, the mode is “logical” and the NAs (or Not Availables/Nulls) are 35,912, which represents the length of observations. For the three variables with numeric data types, the minimum, maximum, and median values are given, as well as the 1st and 3rd quartiles, and the amount of NAs. For the latitude and longitude variables, most of these values are less compelling, although the number of NAs is insightful, which informs us that there are 86 NA’s for both latitude and longitude.

For the reviews.rating variable, the minimum rating is 0, the mean is 3.776, the median is 4, the 3rd quartile is 5, and the maximum is 10. Therefore, it is obvious there are some outliers in this variable. Additionally, there are 862 null values in this variable.

As there were so many variables with character values (including several text-heavy variables), and many unique values, and many missing values within the dataset, there were limited appropriate visualizations that could be completed on the dataset during the exploratory phase of the analysis. However, several barplots were created to explore some of the variables.

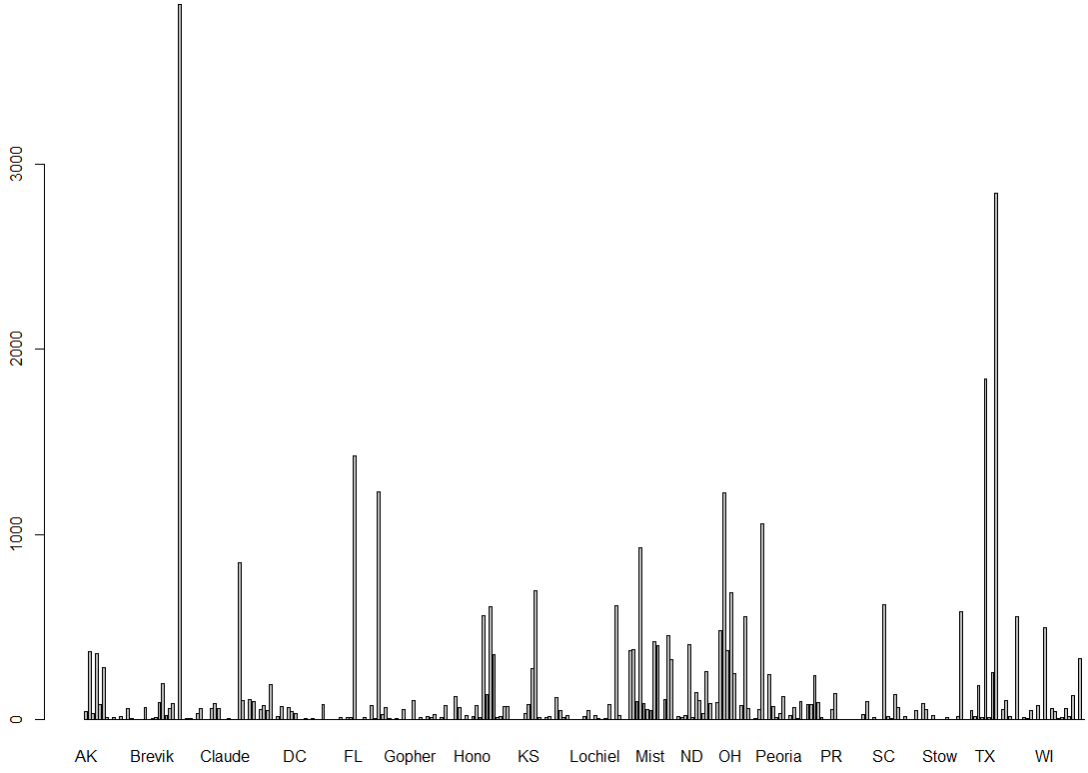

Figure 3: Bar Plot of the Province Variable

The first bar plot (*Figure 3*) illustrates the province variable, which is the province in which the hotel is located. Due to aspect constraints in the graphical interface, not all of the provinces are listed in the visualization. The unique count showed, there are 287 unique

provinces captured in the dataset. There is also a wide range in the distribution of the different provinces, with some provinces only showing a few times in the dataset, and some showing often.

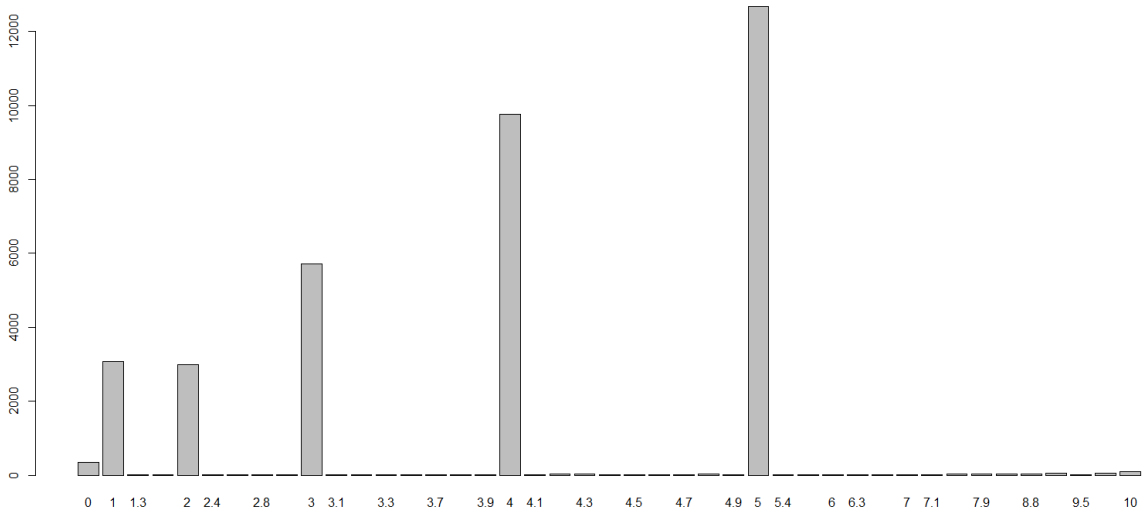

Figure 4: Bar Plot of the Reviews.Rating Variable

The second bar plot (*Figure 4*) illustrates the reviews.rating variable, which is the rating provided by the guests at the hotel. While the ratings range from 0 to 10, and include decimal scores, there are obviously certain whole numbers that are most prevalent in the dataset, specifically 1, 2, 3, 4, and 5. As a reminder, the unique values function indicated there are 44 unique review ratings within this variable in the dataset.

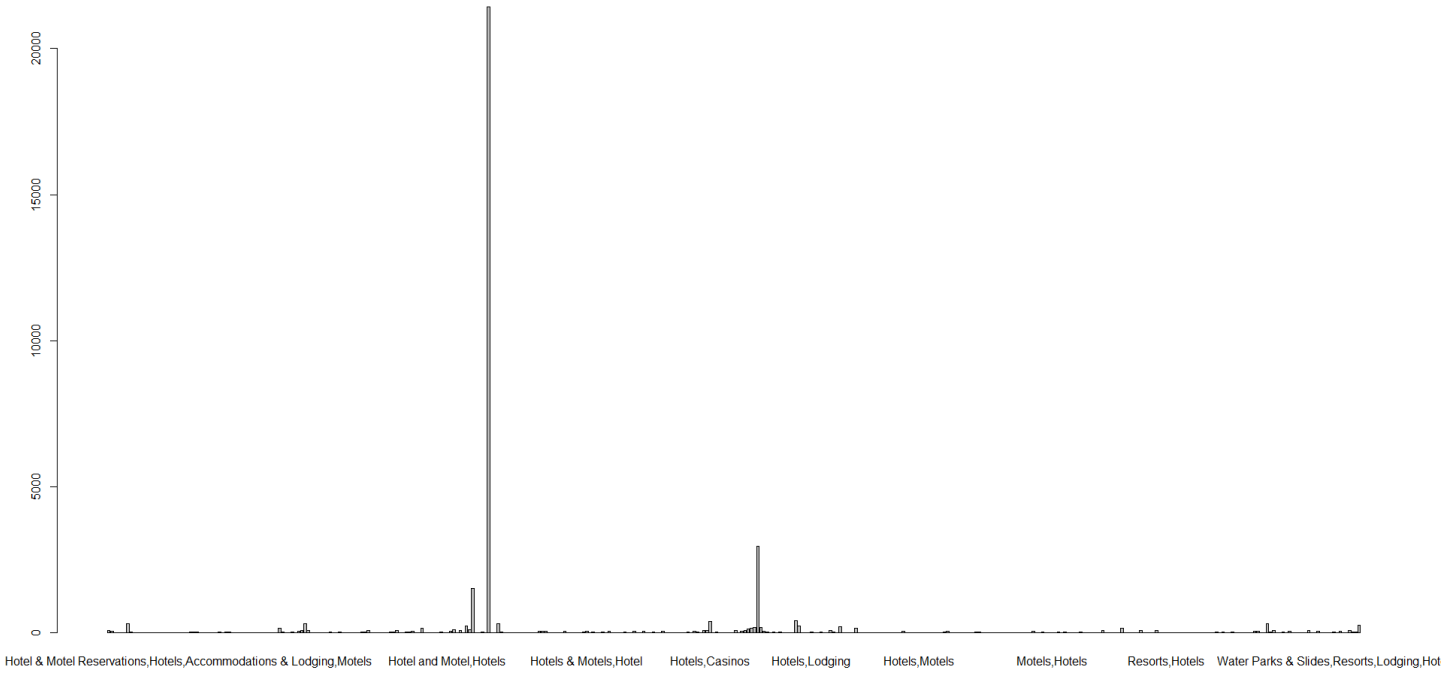

Figure 5: Bar Plot of the Categories Variable

The third bar plot (*Figure 5*) illustrates the categories variable. There are 396 unique values in this variable, but it is clear based on the visualization that there is an outlier in this variable. In taking a look at the data in R, it is not clear why the data is clumped together in this way with this variable. There seems to be a lot of redundant clusters in this variable.

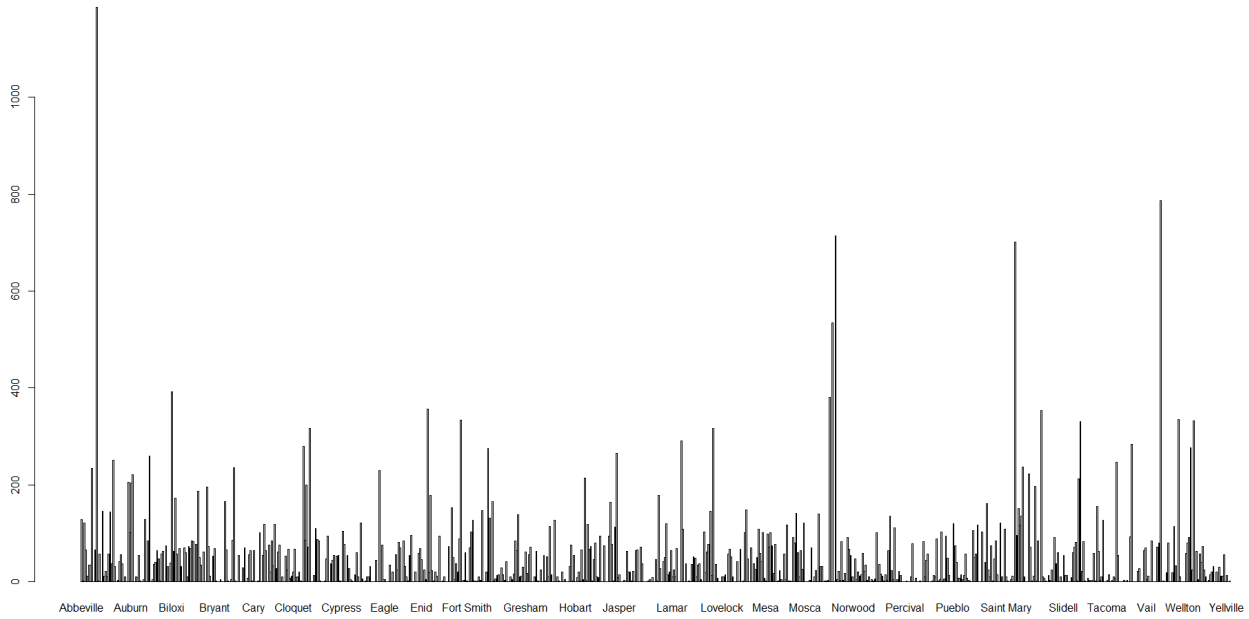

Figure 6: Bar Plot of the City Variable

The fourth bar plot (*Figure 6*) illustrates the city variable, which is the city in which the hotel is located. There are 761 unique cities located in the dataset, and although not all of their names are listed in the bar plot due to graphical limitations, all of their counts are represented. There is a wide range in their distributions, as well as some obvious outliers in this variable.

Preprocessing

The first step in preprocessing was identifying outliers. Since the methods to identify outliers work with numeric or integer data types, there were only a few variables to which they would be applied. Of those three, two were the latitude and longitude variables, where outlier detection would be irrelevant, so only the reviews.rating variable was analyzed:

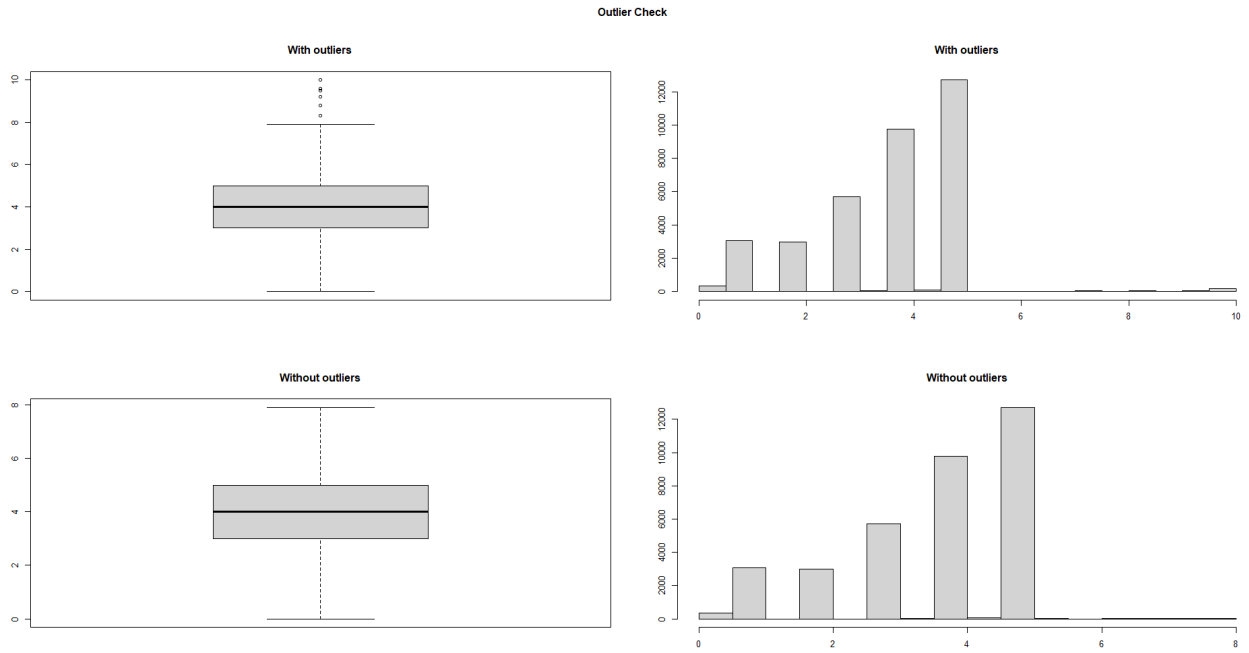

Figure 7: Outlier Analysis of Reviews.Rating Variable

On the R-Bloggers website, a method and complementary R function was found that was utilized in this analysis in order to identify and remove the outliers by replacing them with NAs; the full code is provided in the Appendix (Dhana, 2016). These NAs were also subsequently removed, but this will be explained further later in the preprocessing section. As can be seen in the "with outliers" plots, there were many values outside of the box plot, namely any value after eight. In the "without outliers" plots, these values have been removed, and are no longer visible in the plots.

Next, as part of preprocessing, a check was completed to look for null values in all of the variables.

```
> apply(newdata, 2, function (newdata) sum(is.na(newdata)))
             address           categories                 city              country             latitude
                   0                    0                    0                    0                   86
           longitude                 name           postalCode             province         reviews.date
                  86                    0                    0                    0                    0
   reviews.dateAdded  reviews.doRecommend           reviews.id       reviews.rating         reviews.text
                   0                35912                35912                 1115                    1
       reviews.title     reviews.userCity     reviews.username reviews.userProvince
                   0                    0                    1                    0
```

Figure 8: Null Values in All Variables

At this point, the following variables had null values: latitude, longitude, reviews.doRecommend, reviews.id, reviews.rating, reviews.text, and reviews.username. Since the reviews.doRecommend and the reviews.id variables contain only null values, they were removed completely from the dataset, as they would not have added anything of value to the analysis. Next, since the review date column might come in handy later on, but it contains both the date, as well as a time value, so that a sample value reads as "2013-09-22T00:00:00Z", the substr command was utilized so that a new column was created to capture only the first 10 characters from the column, so that the new column would capture only "2013-09-22" from the earlier example. Lastly, the na.omit command was ran to get rid of the last remaining NA values, and finally, the dataset was checked for null values to make sure there were none remaining:

```
> apply(cleaneddata, 2, function (cleaneddata) sum(is.na(cleaneddata)))
             address           categories                 city              country             latitude
                   0                    0                    0                    0                    0
           longitude                 name           postalCode             province         reviews.date
                   0                    0                    0                    0                    0
   reviews.dateAdded       reviews.rating         reviews.text       reviews.title     reviews.userCity
                   0                    0                    0                    0                    0
    reviews.username reviews.userProvince substring_reviewDate
                   0                    0                    0
```

Figure 9: Null Values in All Variables after PreProcessing

There were additional preprocessing techniques utilized to aid in the text mining methods, but since they were specific to that portion of the analysis, they are discussed in that portion of this paper.

Algorithm Intuition

The analysis went through six stages or processes in achieving the set objectives: 1. Reading the hotel reviews data; 2. Preprocessing; 3. Vector representation of text; 4. Model; 5. Output result; and 6. Evaluation.

In this paper, the text mining analysis focused on sentiment and clustering classification algorithms in gaining insights into hotel reviews data, and determining the sentiment of patrons, which includes which hotels need improvements and in what areas, based on the result of the analysis.

Sentiment Analysis helps organizations in measuring their strength, reshape delivery mechanisms, and identify public sentiment towards their products or services delivery; in other words, it is computation techniques of extracting subjective information from text.

Sentiment analysis can be broken into two approaches: 1. The rules-based approach, and 2. Automatic sentiment approach. The rules-based approach is the common type of sentiment analysis where text documents are analyzed without training the data or using any supervised machine learning algorithms. This method is based on the predefined rules on which text is labeled as positive, negative, or neutral. This approach is also known as the lexicon approach and generally follows this process:

1. Stemming - cleaning the text data
2. Tokenization - The method of breaking down the text into smaller pieces
3. Speech or text tagging
4. Parsing - segmenting the words based on polarity

5. Lexicon analysis - determining the sentiment bearing phrase and scoring them accordingly.

The second approach is the automatic sentiment machine learning analysis. In this approach, the model is built to dig into the text and present the results with little human intervention. It uses supervised classification approaches, where sentiment detection is framed as binary (positive /negative) and labeled data needed to train classifiers (Manasee, 2015). Examples include: Naive Bayesian classification, linear regression, and support vector machines.

The main idea behind K-means clustering is to define clusters that will minimize the within cluster variation, and the standard algorithm for achieving this is the Hartigan-Wong algorithm (1979), which defined total within cluster variation as the sum of squared distances between items and the corresponding centroid, given as -

$$W(C_k) = \sum_{x_i \in C_k} (x_i - \mu_k)^2$$

Figure 10: Hartigan-Wong K-Means Algorithm

- where xi is the data points assigned to the cluster. Ck, uk is the mean value of the points assigned to clusters Ck. Each instance is assigned to a cluster through minimization techniques in a way that variation between the sum of squared distance of the instance to their respective clusters is minimal.

The total within cluster variation is defined as:

$$tot.withiness = \sum_{k=1}^{k} W(C_k) = \sum_{k=1}^{k} \sum_{x_i \in C_k} (x_i - \mu_k)^2$$

Figure 11: K-Means Minimization function

This is the total within the cluster sum of square functions, and validates the goodness of the clustering algorithm. The idea here is that we want our value to be as small as possible.

Text Mining

Prior to conducting sentiment analysis on the data, additional processing needed to be conducted on the dataset that was specific to text mining. Specifically, the reviews.text variable was transformed into a text corpus first. Next, stopwords were removed from the reviews section. Stop words are words that are identified prior to analysis as being irrelevant to analysis, and are therefore removed (Welbers et al., 2017).

Next, a command was run in order to lower all of the text to lowercase so that the case of the words would all be the same. This was done so that, for example, the word “clean” and “Clean” would not be treated separately. This is considered a form of normalization, which is a large part of preprocessing for text mining.

Normalization ensures uniformity across the corpus so that the program the text mining is being performed on recognizes similar text as having the same meaning (Welbers et al., 2017). Next, a line of code was included to remove any emojis, if there were any in the text reviews, since that is popular in text communications on the internet nowadays, since those will be irrelevant in the analysis.

Next, numbers and punctuation were removed as well. Again, these will be irrelevant in the analysis, so keeping them in will only confuse matters and create noise in the results. Next, the words in the corpus were stemmed, meaning words are taken down to their most basic form or root (Shirani, n.d.). For example, words like talked, talking, and talks, would all be stemmed down to talk, so that they wouldn't all be treated separately in text mining analysis (El Marie, 2018a). Lastly, extraneous whitespace was stripped from the corpus.

Next, a Term Document Matrix was created, and the word frequencies were calculated, which was used to create a word cloud of the words used in the hotel reviews (Van den Rul, 2019):

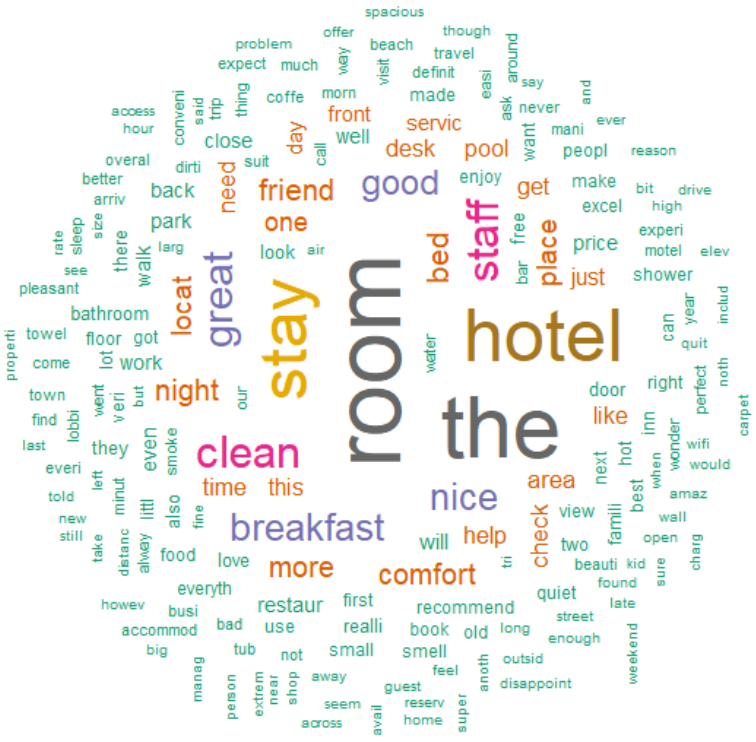

Figure 12: Word Cloud of Hotel Reviews Word Frequencies

As seen above, the word cloud identifies the most populous words used in the hotel reviews. The words used most frequently are sized accordingly, and are also colored to differentiate themselves, with the less frequently used words colored the same, and the more frequently used words colored differently to separate themselves.

Afterwards, a different Term Document Matrix was created, and called myTdm, in order to conduct additional text mining methods. By calling myTdm, it is possible to see some descriptive statistics, such as that the term document matrix is composed of 23,013 terms and 34,719 documents (which are the reviews from the original dataset; Shirani, n.d.). From here, it is possible to choose specific letters of the alphabet to see some of the terms associated with that letter. For example, the following shows the first six terms starting with the letter "s":

```
> idx<-which(dimnames(myTdm)$Terms=="s")
> inspect(myTdm[idx+(0:5),100:110])
<<TermDocumentMatrix (terms: 6, documents: 11)>>
Non-/sparse entries: 1/65
Sparsity           : 98%
Maximal term length: 7
Weighting          : term frequency (tf)
Sample             :
        Docs
Terms    100 101 102 103 104 105 106 107 108 109
  s        0   0   0   0   0   0   0   0   0   1
  safe     0   0   0   0   0   0   0   0   0   0
  satisfi  0   0   0   0   0   0   0   0   0   0
  sausag   0   0   0   0   0   0   0   0   0   0
  scrambl  0   0   0   0   0   0   0   0   0   0
  see      0   0   0   0   0   0   0   0   0   0
```

Figure 13: Term Document Matrix for Six Terms Beginning with Letter "S"

What is interesting about the above term document matrix for the letter "s" is that the word stemming that was conducted earlier is visible. For example, the word "scrambl" was possibly stemmed from longer words, such as: scramble, scrambled, scrambling, etc.

Next, word frequencies were also looked at by using the findFreqTerms command, and then plotted using a bar plot. First, the words in the term document matrix were filtered by a minimum frequency of 10. This still resulted in a lot of terms since the original dataset was so large. In order to have a manageable plot, the top ten terms of the frequent terms were captured and plotted in a bar plot, which is shared below:

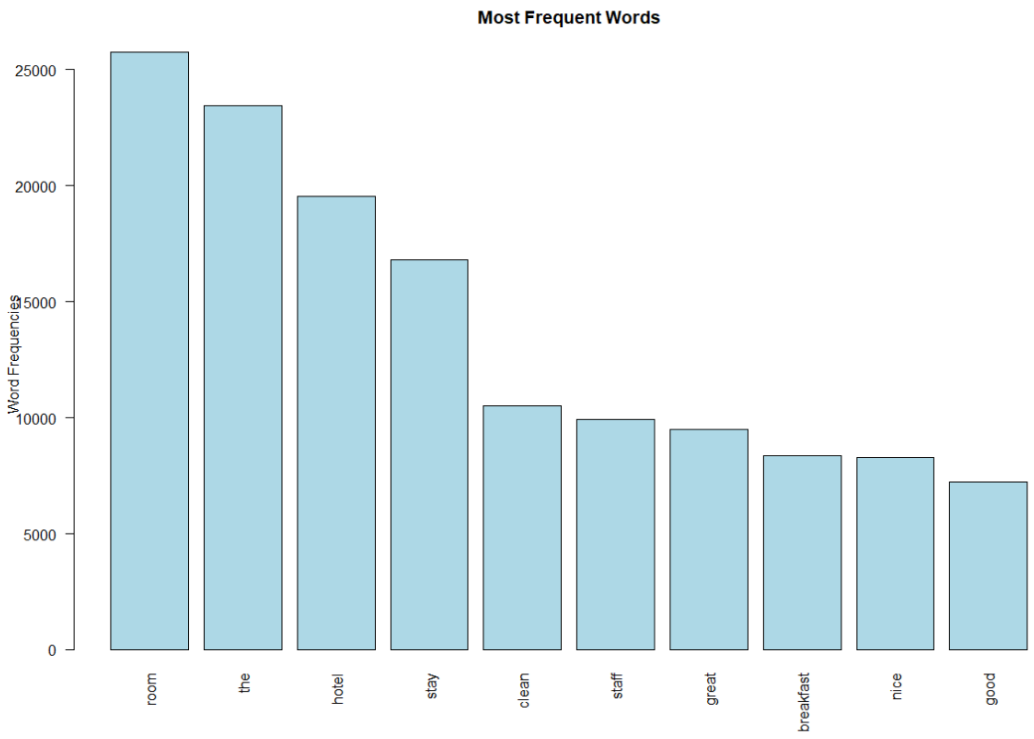

Figure 14: Barplot of Ten Most Frequent Words

Following the look at the most frequent distributions of terms, sentiment analysis was implemented on the text of the hotel reviews. As mentioned, sentiment analysis is ideal for analyzing hotel reviews because its purpose is to identify "...reactions, attitudes, context, and emotions." (El Marie, 2018b). Sentiment analysis is also referred to as word polarities, where

stereotypically positive words and stereotypically negative words are taken into account, while neutral words are ignored (El Marie, 2018b). First, the sentiment mining was conducted on the data using the syuzhet package in R. After all of the processing in R, the first sentiment analysis plot included a series of eight distinct emotions: anger, anticipation, disgust, fear, joy, sadness, surprise, and trust (El Marie, 2018b). This resulted in the following plot:

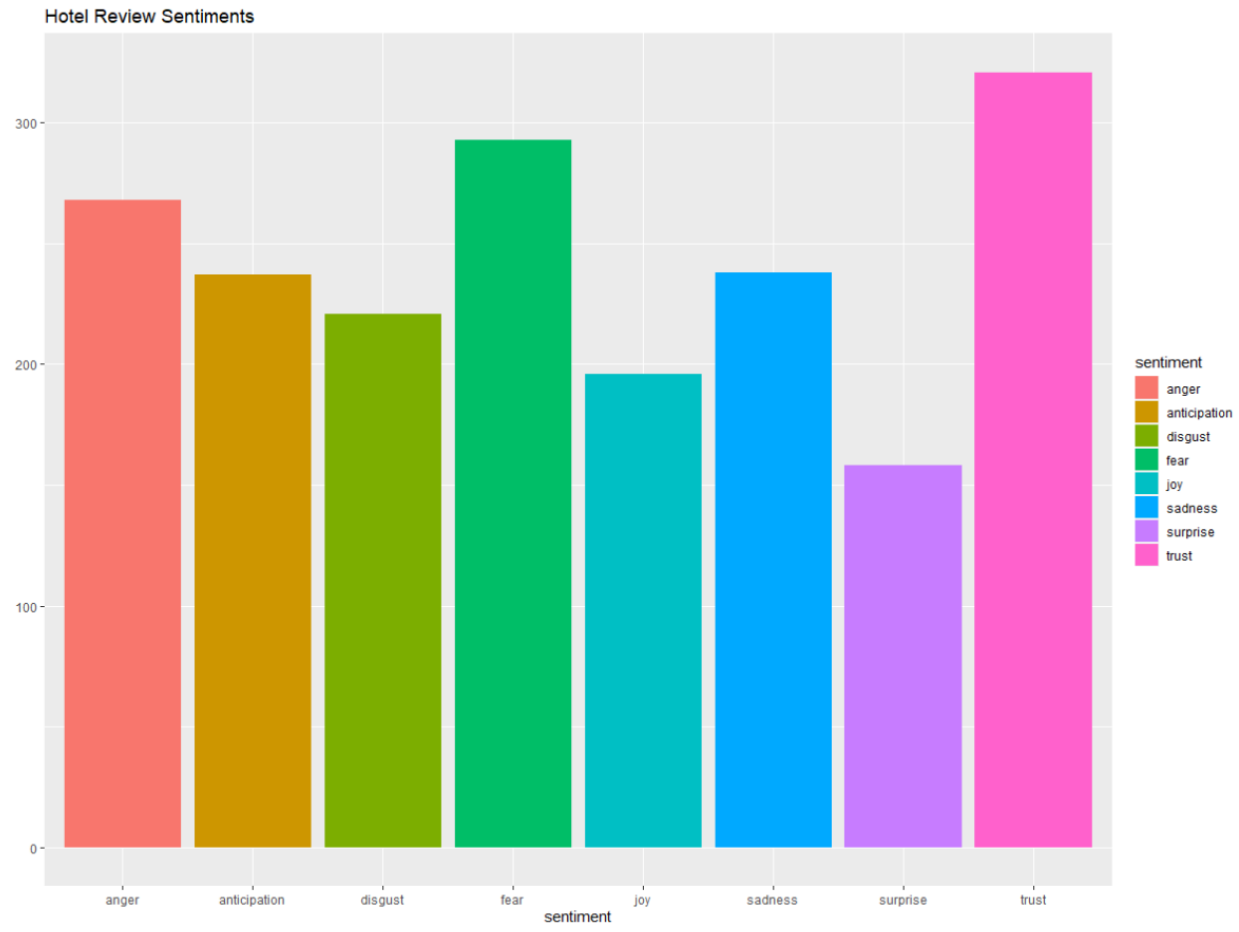

Figure 15: Overall Sentiment Analysis of Eight Distinct Emotions

The next sentiment analysis plot was an overall polarity plot of negative versus positive, and resulted in the following plot:

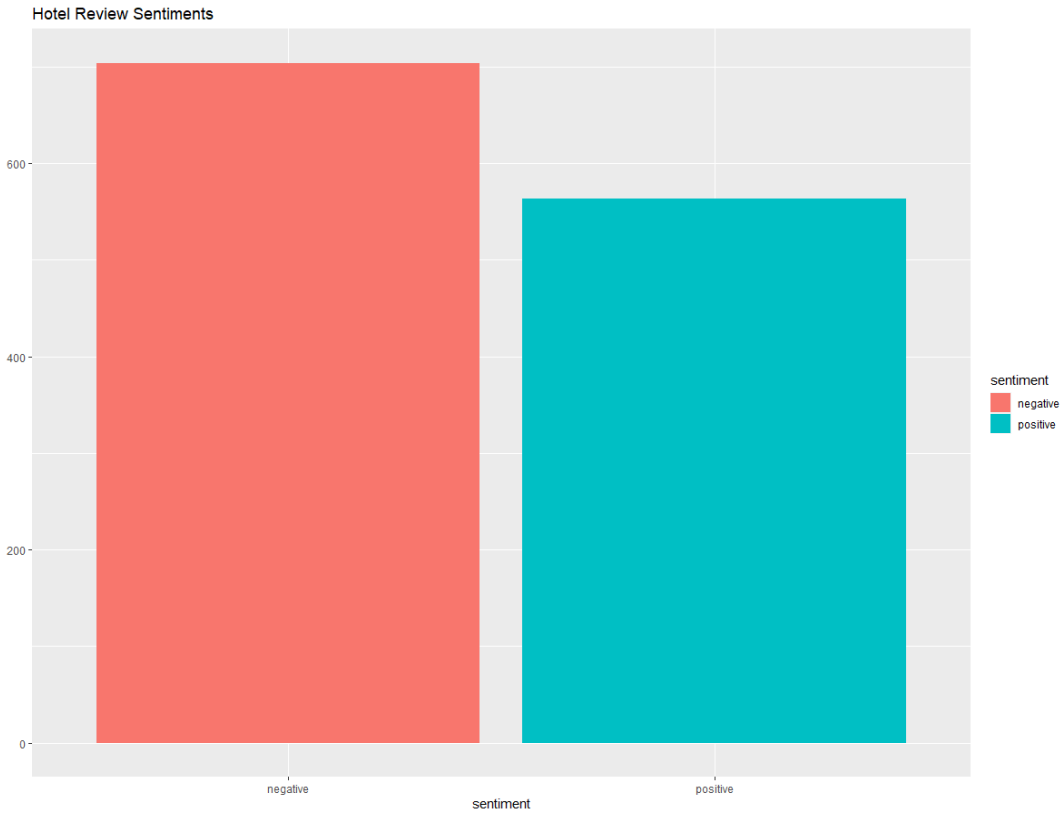

Figure 16: Overall Sentiment Analysis of Two Primary Polarities

In looking at the first sentiment analysis plot (*Figure 15*), the overarching sentiment is trust, which is a positive sentiment for hotels to receive. However, the second overarching sentiment is fear, which is a directly conflicting emotion to trust, so that is a surprising finding. The third most populous sentiment is anger, while the sentiment that appears the least is surprise. Looking at these findings overall is interesting, but they would be far more compelling for a hotel owner to dig into for their hotel property specifically, or perhaps for a city tourism manager to dig into for their city specifically.

In the second plot (*Figure 16*), the overall sentiment is negative, while a positive sentiment is second. This would mean that more reviews had overall negative sentiments. One thing to note about sentiment analysis is that it is not good at differentiating between slang or

sarcasm, so that can skew results. For example, if someone says that the hotel room was “sick”, which in slang terms can actually mean that it was really cool, this can be interpreted to be a negative review in sentiment analysis. Additionally, according to a review of sentiment computation methods with R packages, the default lexicon in the syuzhet package tends to skew towards dominant negative words since, out of 10,748 words, 7,161 words are negative, while 3,587 are positive (Naldi, 2019).

Cluster Analysis

Machine learning paradigms can broadly be classified as either supervised, unsupervised, or reinforcement learning. Clustering is an unsupervised learning methodology whose primary objective is to classify unlabeled data into their unique categories based on aspects such as the similarity of the different characteristics of the provided data. In this work, clustering was considered a necessary step in the general understanding of the given problem (Denny and Spirling, 2018). This section therefore outlines the significance, steps, and results of clustering analysis provided in the data.

Significance of Clustering Hotel Reviews

In the data provided, no clear relationship could be mapped between the text provided in the reviews and the different categories to which the reviews belonged to. Having a clear understanding of this relationship would be essential in developing subsequent models for the data that are aimed at understanding patterns followed in developing different reviews. For instance, by understanding the relationship between different reviews in different categories, it would be possible to establish how customer reviews on a specific category affected their

perception about the services offered in another category. Other similar insights can be drawn from such analysis, thus making the entire process of text mining beneficial in the long run.

Steps Followed

The text column of the data first had to be cleaned by removing all forms of URLs that may have been present in the reviews. This step was necessary since such URL patterns in text reviews have no meaningful impact as far as the semantics of the reviews are concerned. The removal of URLS implied that more whitespace in the data was created. This effect was therefore rectified by removing extra whitespace in the reviews without the URLS. Finally, the last step was to remove stopwords present in the text. This removal was based on the fact that stopwords comprise a large proportion of text data while at the same time having no important contribution to the structure of the data. Therefore, the removal helps in reducing the dimensionality of the data to be analyzed, subsequently increasing the general computational time.

Feature engineering was also performed on the cleaned data to make it appropriate for the clustering algorithms. Text data in its raw text format bears little information that can be analytically evaluated. However, through vectorization, different words in the provided text can be assigned different weights which would subsequently make the analytic process more effective and its subsequent results more insightful. To accomplish this objective, a document term matrix was created from the original text data. This matrix was then assigned weights using the term frequency inverse document frequency approach (Qaiser and Ali, 2018). Finally, the data's dimensions had to be reduced again to reduce the size due to R's inability to effectively

handle memory from the clustering to be performed. A cosine similarity approach was then used for calculation of distance metrics in the data.

Clustering requires that one passes the appropriate number of clusters into which the data has to be categorized into. To select the most appropriate number of clusters, various mechanisms can be used. The Elbow method was used in this case to determine the best number of clusters to fit the data into. This approach leverages on the within-cluster sum of square metrics to determine the optimum value of *k* (Kothary et al., 2018). A range of between 1 and 10 was passed as the possible range of the data, and thereafter the computations were allowed to run for 10 iterations for them to converge. This resulted in the following plot:

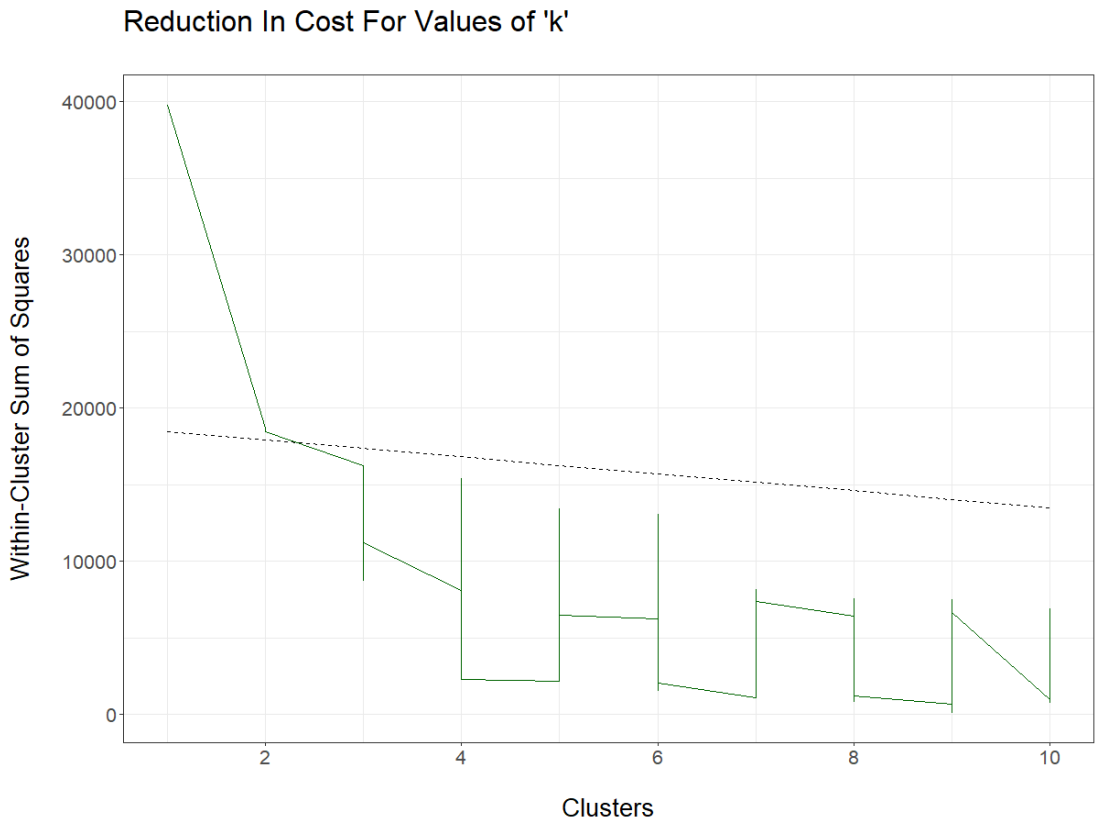

Figure 17: Plot of Reduction in Cost for Values of k

The results of the elbow method showed that 8 was the most optimum number of clusters for the text data. In the steps that followed, the data was clustered into 8 categories using the K-means clustering algorithm. The clustering was then followed by a visual representation of the distribution of the text clusters in a 2-dimensional plane. The image below shows the distribution of the data in the selected plane:

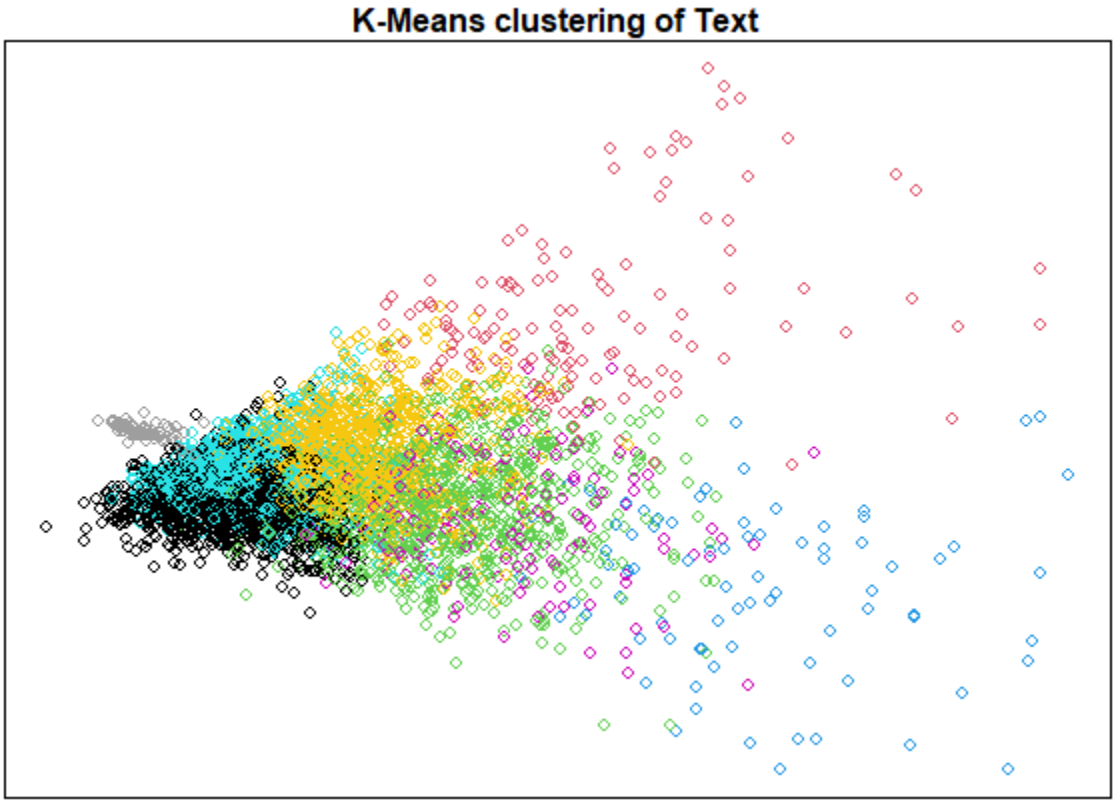

Figure 18: Plot of K-means Clustering of Hotel Reviews Text

From the findings above, it can be seen that there are instances where certain clusters are distinctly separate in the plane. However, most of the clusters are closely located and in some cases, it is difficult to clearly separate any two clusters. These results therefore hint at the possibility of similar word terms being used in writing the reviews for different categories

provided in the data. A clearer comprehension of the data would however be obtained by performing the t-distributed stochastic neighbor embedding process on the data (Wattenberg and Johnson, 2016). Unlike the 2-D visualization provided above, the t-SNE representation provides one with a higher-dimensional visualization of the data thus providing a clear representation of the relationships between different terms.

Result

Output:

At the beginning of this analysis, the hotel reviews data was explored for insights using barplots, and we were able to find that hotels located in the provinces of Florida, California, Virginia, Georgia, Ohio, New York, Pennsylvania, and Texas dominated the dataset. An initial insight from the text document shows that sentiment from hoteliers are not randomly distributed across the hotels as shown in the diagram below.

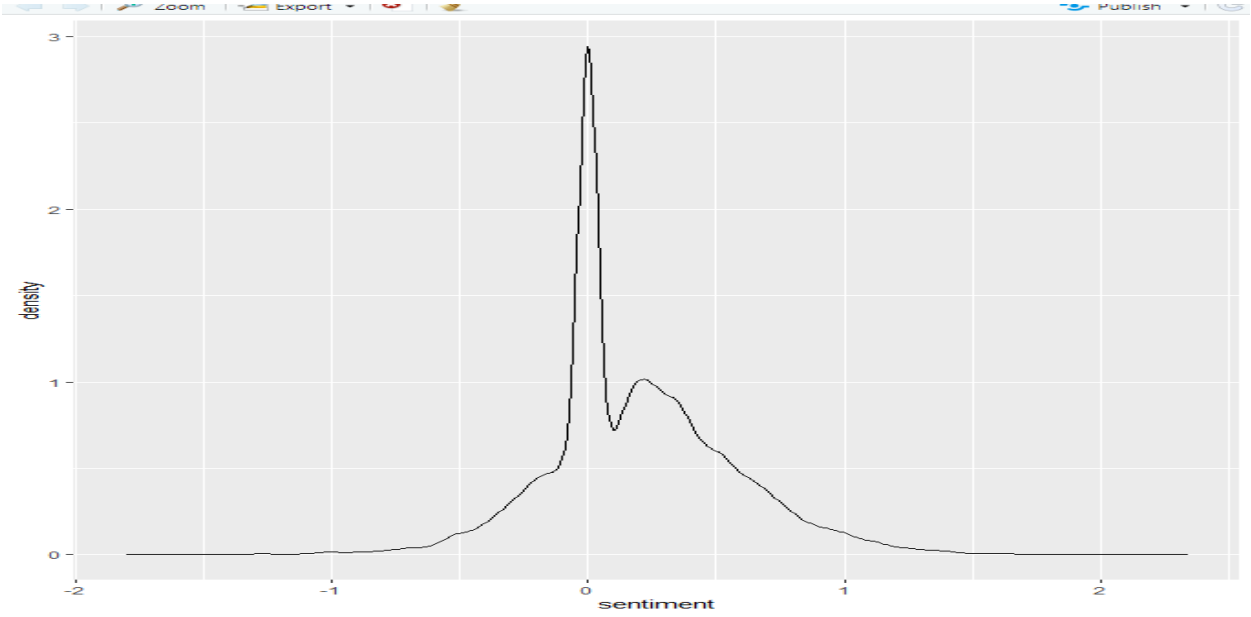

Figure 19: Hotel Reviews Sentiments Distribution Plot

Prior to looking into specific provinces, some overall text mining and NLP methods were incorporated into the dataset in order to get an overview of the reviews' content and sentiment.

Due to the volume of the dataset and computation complexity, further digging into the data using barplots and boxplots resulted in the plots being unreadable. Hence, the data was segmented into regions and three provinces with high dominated hotels were selected for the sentiment analysis, and they are California, Texas, and Florida.

The barplots for the California province suggested that Wine Valley lodge, Hotel Valencia Santana Row, Hotel Aurora SFO, Anaheim Marriott Suite, Best Western of Long Beach, Simpson House Inn, Residence Inn by Marriott Irvine John Wayne Airport, Best Valley Western Silicon Valley Inn, Cherokee Lodge Bed and Breakfast, Hawthorn Suites by Wyndham Livermore Wine Country, California all had impressive customers positive ratings compared to negative sentiment, as shown on the bar chart below.

The emotional review plot for California shows that customers were more concerned about trust, anticipation, fear, and sadness. The joy bar is seen relatively high too but not as much as trust, fear, and anticipation. These suggested that some hotels are still meeting customers' needs in terms of satisfaction (see *Figure 20*, for California sentiment analysis barplots).

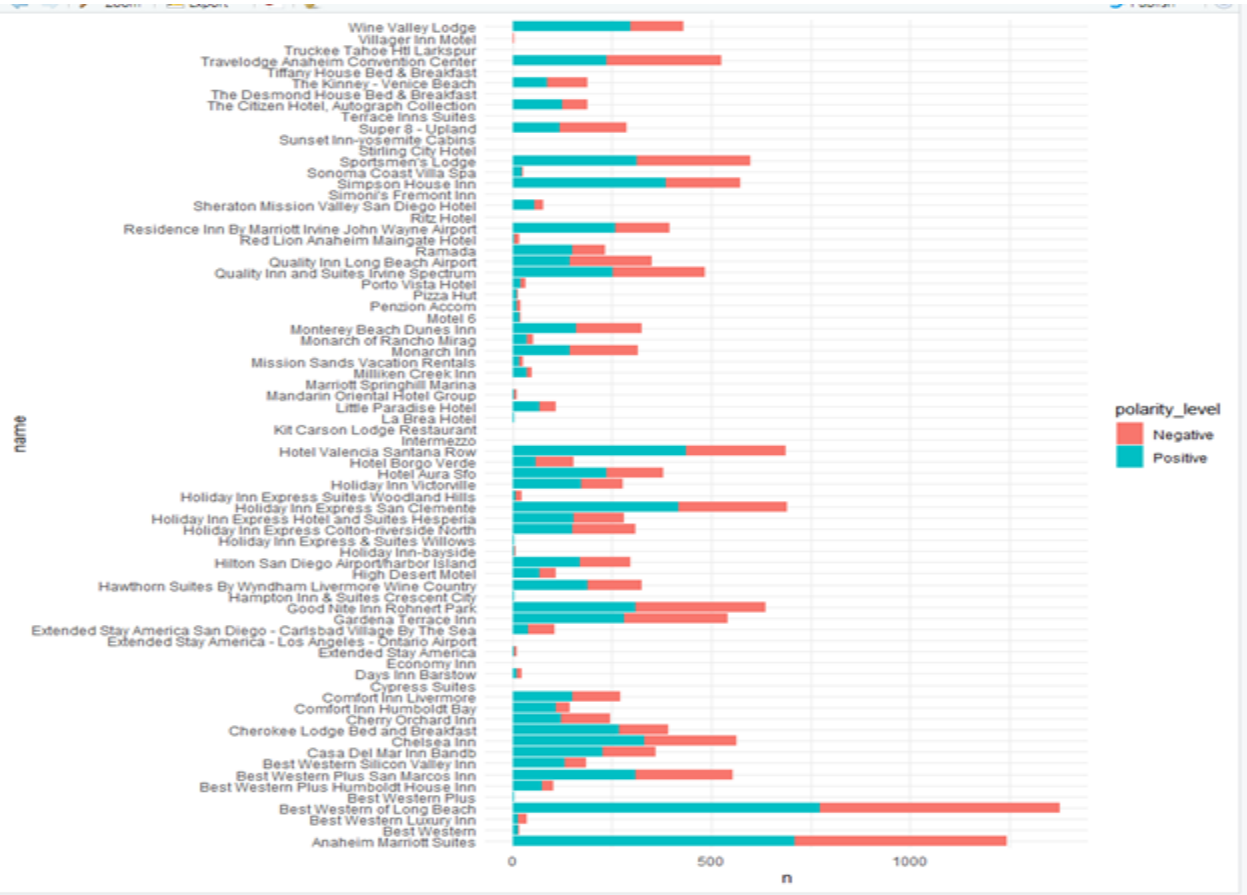

Figure 20: California Hotel Sentiment Barplot

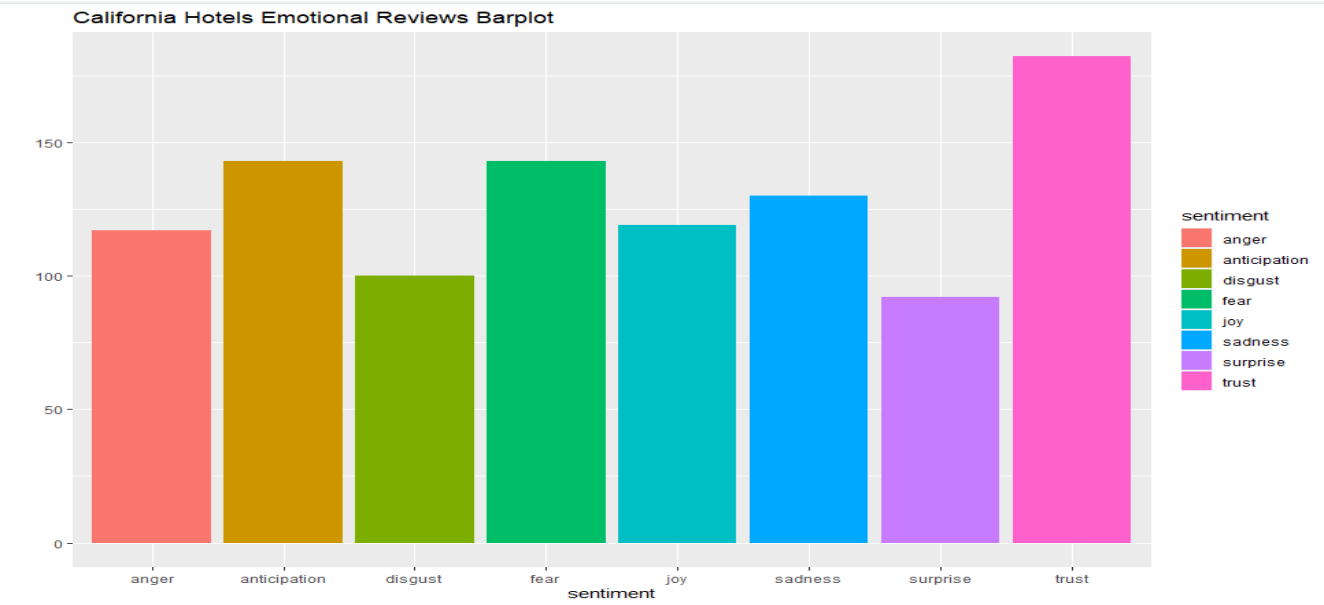

Figure 21: California Hotel Emotional Reviews Barplot

Florida is a tourist-driven economy; in 2017, Florida visitors spent $90 billion in hotels, transportation, etc. (UCF). Therefore, sentiment analysis for this province is crucial for this analysis.

The Florida hotel emotional and sentiment reviews bar plots suggested that only few hotels actually received a positive sentiment; for instance, Doubletree By Hilton West Palm Beach received more positive ratings than any other hotels in Florida (in our data), followed by Jacksonville Plaza Hotel and Suite, and Hampton Inn Sarasota-1-75 Bee Ridge, while Comfort Inn, Vista Hotel on Lake Tarpon, and Extended Stay America-Orlando Theme Parks-Vineland Rd are at marginal sentiment. In addition, the Florida emotion barplot suggested a similar pattern to that of California, in the sense that customers' trust and anticipation still take the lead. This simply means all customers look forward to receiving what was advertised, either on the company web page or on social media platforms (value for money).

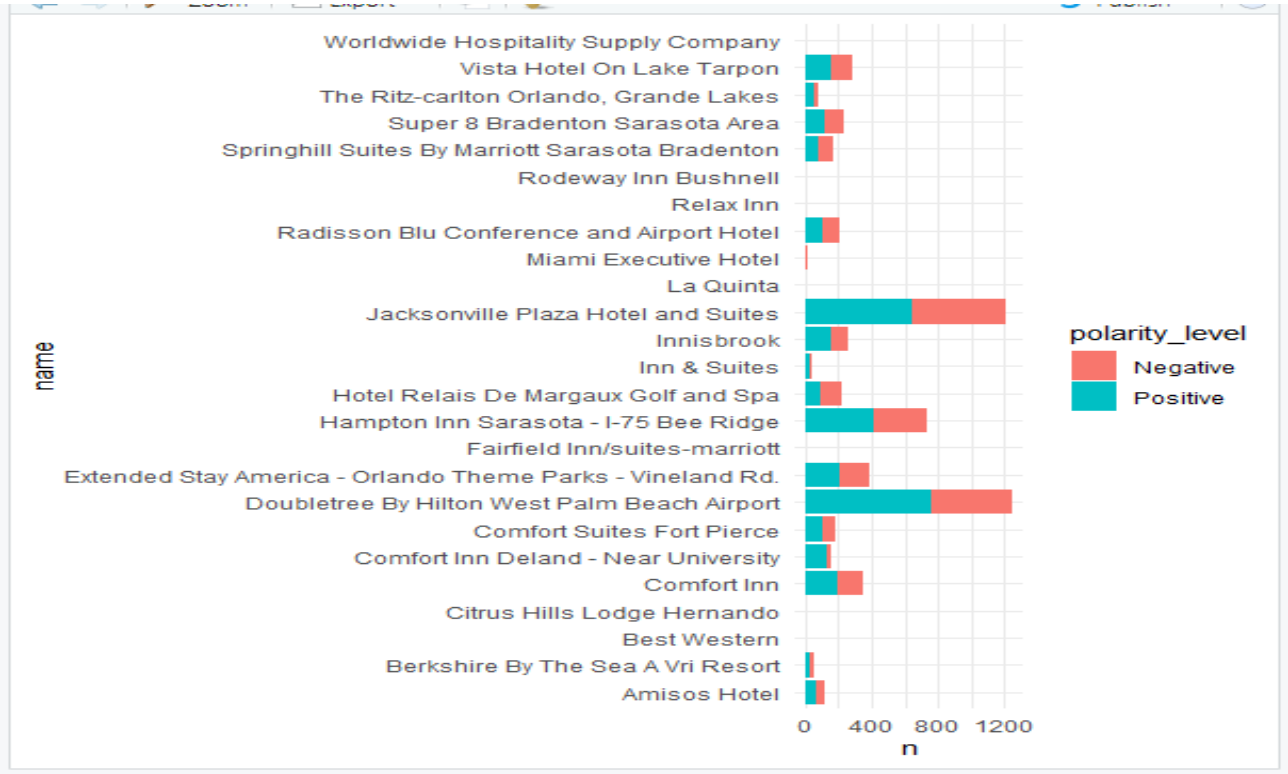

Figure 22: Florida Hotel Sentiment Barplot

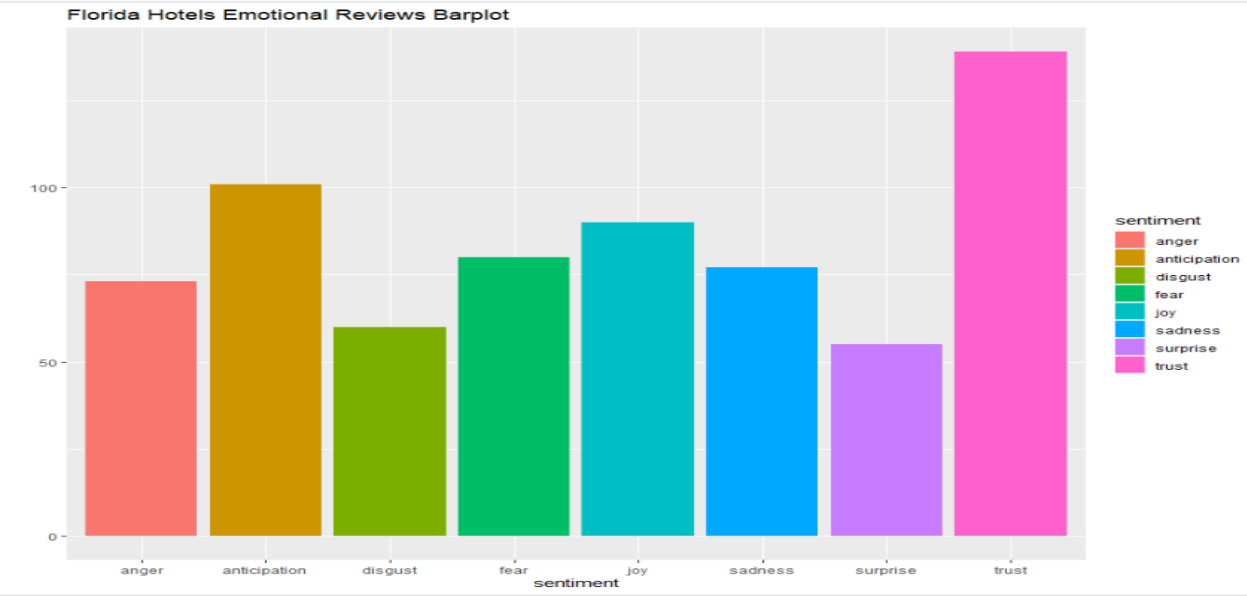

Figure 23: Florida Hotel Emotional Reviews Barplot

Texas hotels sentiment analysis showed a similar outcome compared to the previous provinces examined, meanwhile one hotel (Fiesta Inn and Suites) was observed to have a higher amount of customers' negative sentiment compared to its positive customers' sentiment.

Hampton Inn Abilene, Hampton Inn & Suites Fort Worth-West -1-130, and Excellence Riviera Cancun - Adult Only - All-inclusive dominated the customer ratings in the province (as shown on the plot). The emotional barplot actually gives more insight into Texas hotel reviews, we can see that trust and fear dominated the tone in the text review and determined the ratings of the hotels. Patrons clearly want to receive value for money spent and obtain satisfaction with hotel services; this finding aligns with the word cloud diagram in this analysis (found in the Appendix).

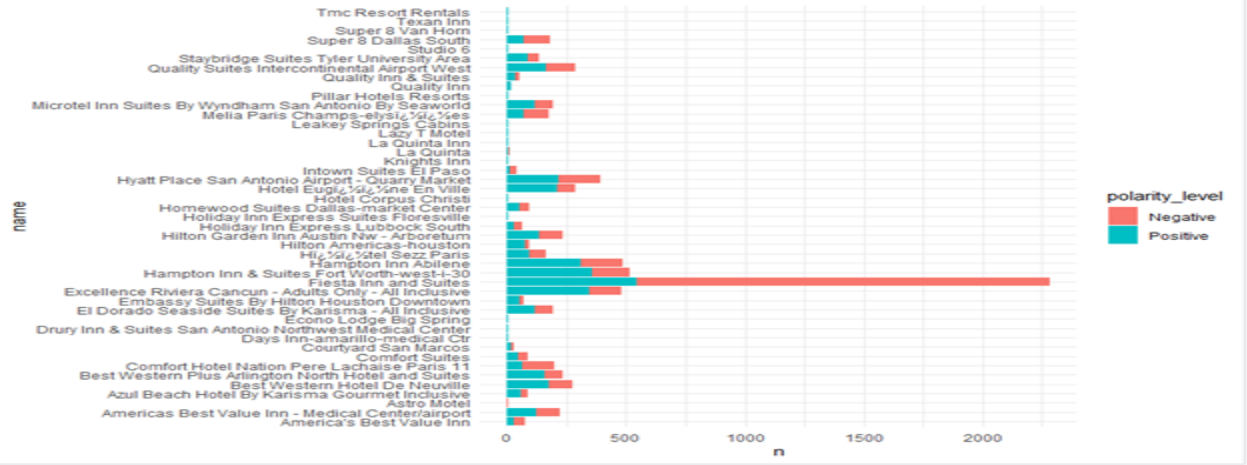

Figure 24: Texas Hotel Sentiment Barplot

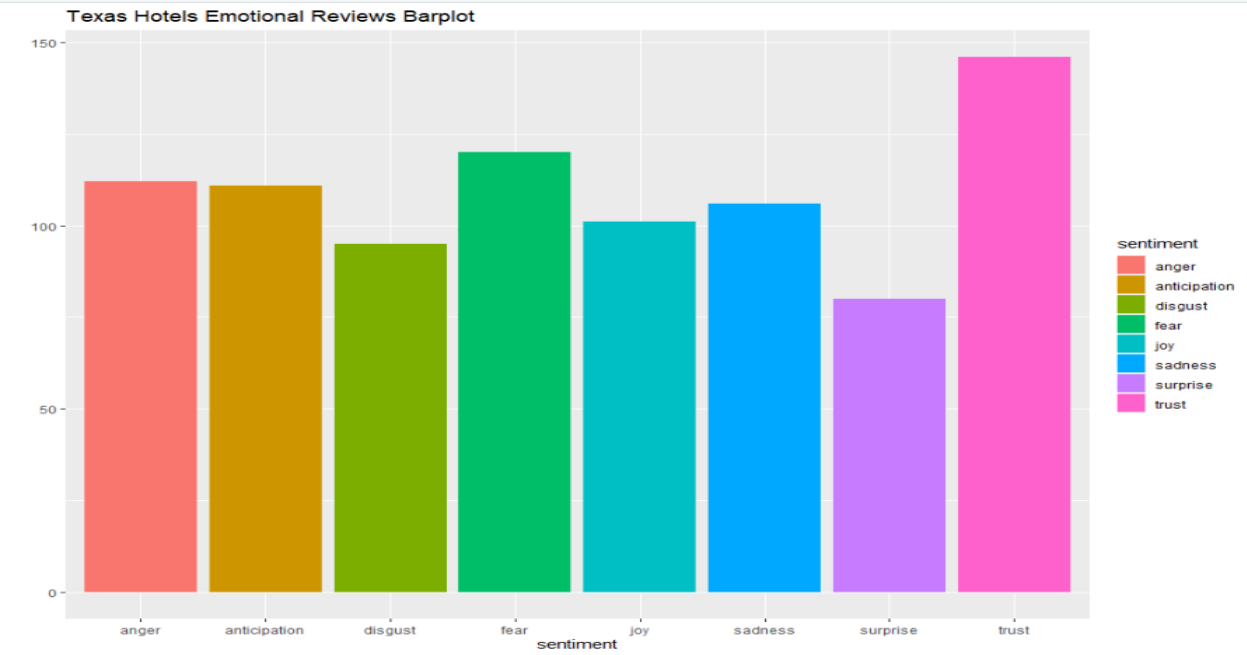

Figure 25: Texas Hotel Emotional Reviews Barplot

Properties

The K-means clustering analysis properties or parameters used in this paper are K-means iteration (iter) which explain how many loop or iteration steps the algorithm runs prior delivering an output. In our model, the iteration steps = 6.

The other properties are tot.withinss (the total within the cluster sum of squared), kmeans $centers (to display the attribute values at the cluster), kmeans $size (number of instances in each cluster), kmeans $totss (to display the total sum of squared errors), and K-means $betweenss (to display the between cluster sum of squares).

Evaluation

The K-means clustering analysis went through 6 iterations, this means the algorithm returned output after the 6th run of iteration (see *Figure xxx, below*) with eight (8) clusters, and the instances in each cluster are 118, 1082, 142, 1932, 232, 108, 186 and 1200, making a total of 5000 observations.

```
> kmeans_cluster$iter
[1] 6
```

Figure 26: K-Mean Clustering Model Iteration

```
> kmeans_cluster
K-means clustering with 8 clusters of sizes 118, 1082, 142, 1932, 232, 108, 186, 1200
```

Figure 27: K-means Clustering Model Cluster Sizes

The within cluster sum of squares by cluster are 848.4404, 7770.1122, 1865.8191, 11600.9837, 1879.0594, 1239.9793, 1625.3996, and 6956.2350. This shows the variability of the observations within the cluster,

The cluster has a total variance of 15% (i.e., the sum of squared distances divided by the total squared error) which is not that impressive because the higher the variance the better the cluster analysis. In other words, the 15% suggested that the points in clusters are not well separated.

```
Within cluster sum of squares by cluster:
[1]   848.4404  7770.1122  1865.8191 11600.9837  1879.0594  1239.9793  1625.3996  6956.2350
 (between_SS / total_SS =  15.0 %)
```

Figure 28: Within Cluster Sum of Squares by Cluster

In further evaluating the cluster analysis, a comparison of how the K-means method grouped the hotel reviews dataset into clusters with the actual reviews ratings, cross-table evaluation was built with the command below and it was found that 38% of the 5000 observations matches the actual reviews rating.

table(HotelReview$reviews.rating, kmeans_cluster$cluster)

```
> table(HotelReview$reviews.rating, kmeans_cluster$cluster)

      1   2   3   4   5   6   7   8
0     1   7  12  25   0   0   1  10
1     5  62  11 165  22   4  12 100
2     5  61   8 177  20  10   7 102
3     7 156  21 311  34  20  29 168
4    34 307  55 509  89  38  49 346
5    61 461  32 697  64  30  84 437
6     1   0   0   1   0   0   0   1
7     0   1   0   0   0   2   1   0
8     1   2   1   9   0   2   0   4
9     1   3   0   5   1   0   1   1
10    0   6   0   2   1   1   0   6
```

Figure 29: Model Clusters Evaluation Matrix

Conclusion

Summary

The objective of the analysis was to determine the sentiment (positive or negative) of patrons who stayed at the hotels. More specifically, the purpose is to determine which hotels to recommend and which to avoid. Through this process, it was anticipated that the insights and results generated would be used to advise the hotel managers on how to effectively implement their operations to the satisfaction of their clients. The sentiment analysis tasks found out that in general, most of the reviews, when viewed holistically, were negative, but when some of the provinces were viewed separately, the reviews were more positive with few of them being negative. In this regard, it can be argued that the hotel managers and the relevant shareholders

are generally performing well as far as customer satisfaction is concerned. It is important to realize that this is a generalized conclusion and that independent hotels may have varying comparisons of positive and negative sentiments. This work further explored the concept of clustering and how it could be applied to text data. To achieve the clustering, the K-means algorithm as implemented in R was used. To determine the optimum number of clusters in the data, the elbow plot method was used where the cluster with the least within-cluster sum of squares was selected. This was found as 8 in this work. The results of the clustering revealed that most clusters overlapped. In principle, this relationship is expected since in hotel reviews, there are no clear distinctions in terms of the topics being discussed in the reviews.

Limitations

There were several limitations in this project. They took two forms - the first limitations were in the dataset, and the second limitations were in the analysis.

Data Limitations

The first limitation in the data is that there was significant overlap between the different categories, as well as a lot of duplicative-sounding entries. This caused large spikes in the distributions, and also caused there to be far more categories than needed in this variable. Additionally, slight spelling differences in some of the category types (such as hotel and hotels, resort and resorts) caused entirely new category types (hotel, resort was treated differently than hotels, resorts, etc.), which caused even more redundant/duplicative categories, which compounded the issue.

A second limitation in the data was that there was no clear relationship between the text provided in the reviews and the different categories to which the reviews belonged to. Having a

clear understanding of this relationship would be essential in developing subsequent models for the data that are aimed at understanding patterns followed in developing different reviews.

Analysis Limitations

The first limitation of this analysis was the computer power needed to complete the work. Some of the group's computers seemed to handle the large amount of data and processing needed for the analysis without issue. Others had significant trouble. This was first identified during the exploratory phase of the analysis when, as mentioned above, due to aspect constraints in the graphical interface, not all of the provinces from the reviews were able to be listed in the visualization. This resulted in a reduction of the data to 5000 records for the clustering portion; it was left at the original amount for the text mining and NLP portions. Although still a substantial number of reviews, the limited data may have also limited further insights into the cluster models. K-means clustering by itself was found to be insufficient in fully separating the different word clusters in the provided data. This inability is due to the high dimensionality of the provided text.

A second limitation of the analysis was the general nature of the results in relation to the stated objective. More specifically, the objective was to determine which hotels to recommend and which to avoid. A limitation of the analysis is that specific hotels within such a large dataset are difficult to review in any great depth. The methods of text analysis and clustering are more global analytical processes. Therefore, specific guidance for an individual hotel manager or potential customers as it pertains to a hotel is limited. Instead the manager of a hotel should look for trends of what customers find positive or negative and use that information to review their specific hotel's level of customer satisfaction.

A third limitation of the analysis was that due to the volume of the dataset and computation complexity, further digging into the data using barplots and boxplots made the plots unreadable. Without clear visualizations, it becomes difficult to see groupings in the data, especially from a large dataset.

The fourth limitation of the analysis was that there were also some outlier clusters in the data with significantly few instances. It can be hypothesized that these clusters correspond to the three least common sentiments of the eight sentiments generated in the analysis.

Improvement Areas

Future studies may consider incorporating dimensionality reduction techniques such as t-Distributed Stochastic Neighbor Embedding in the clustering process (Wattenberg and Johnson, 2016). This approach is known for its effectiveness in visualizing high-dimensional data. It has found significant applications in data-intensive fields such as in the study of the human genome. In addition, a comprehensive sentiment analysis can be achieved if the dataset is segmented according to regions because each region has distinct characteristics that need to be explored rather than overshadowed by larger provinces.

t-distributed Stochastic Neighbor Embedding is a dimensionality reduction technique that has been found to be quite helpful when working with high-dimensional data. The technique has been successfully applied in studying bioinformatics, natural language processing and biomedical signal processing among many others. Future studies can therefore include t-SNE in the scope when the data provided is high-dimensional. To overcome the limitation of overlapping clusters, future data can be obtained from different contexts. The semantic meaning of the texts would subsequently be distinct enough, resulting in well-separated clusters. Well-

separated clusters would make it feasible to display common words in each cluster, thus giving more insights into the analysis being performed.

Another improvement area could be in the dataset. For instance, a more recent set of data during the COVID pandemic may be more relevant to world events today and assist hotel owners more effectively with the impact of COVID on travelers. Furthermore, sub-group ratings for specific areas of hotel services, such as a cleanliness rating, a service rating, or a food and beverage rating would be more specific, and provide valuable feedback to both hotel management and consumers looking for their next hotel.

Lastly, some improvement areas for the analysis could include adding different types of analysis, such as: extracting patron feedback on certain features of the hotel, such as health and safety precautions; for example, cleanliness in the time of COVID, which could be an important feature to extract and examine further. As a use-case, families traveling for summer vacations will likely be looking for clean, family-friendly hotels that provide safe and comfortable lodging.

Another different type of analysis to improve on would be to extract reviews on amenities such as breakfast, internet, wellness facilities, and other types of services, which can be important to different types of travelers. For example, business travelers in today's world rely on fast and reliable internet. Have past patrons to the hotel commented on the internet capabilities?

Lastly, building a classification model in order to classify a hotel with a "recommend" or "do not recommend" rating could be really compelling. More specifically the goal of the analysis would be to interpret meaning from the reviews, and extract concepts such as "I like…" or "I don't like…" (positive or negative sentiment) in order to inform the recommendation to stay at the hotel or look elsewhere (Linguamatics, n.d.).

References

Big Data University. (2017). *Machine Learning – Unsupervised Learning K-means Clustering Advantages & Disadvantages* [Video]. YouTube. https://www.youtube.com/watch?v=77psYm8bLsQ

Datafiniti. (2019). *Hotel Reviews. Dataset: Version 5.* Kaggle. https://www.kaggle.com/datafiniti/hotel-reviews

Denny, M. J., & Spirling, A. (2018). Text Preprocessing for Unsupervised Learning: Why It Matters, When It Misleads, and What to Do About It. *Political Analysi*s, 26(2), 168-189.

Dhana, K. (2016). *Identify, Describe, Plot, and Remove the Outliers from the Dataset.* R-Bloggers. https://www.r-bloggers.com/2016/04/identify-describe-plot-and-remove-the-outliers-from-the-dataset/

El Marie. (2018a). *Understanding and Writing your First Text Mining Script with R.* Dev. https://dev.to/lornamariak/understanding-and-writing-your-first-text-mining-script-withr-345k

El Marie. (2018b). *Exploring Sentiment Analysis as an Application of Text Mining.* Dev. https://dev.to/lornamariak/exploring-sentiment-analysis-o6j

Garcia, R.A. (2012). *What is text classification?* [Video]. YouTube. https://www.youtube.com/watch?v=vfnxRGmP_ss &feature=youtu.be

Han, J., Kamber, M., & Pei, J. (2011). *Data Mining Concepts and Technique, 3rd Ed.* Ch.3. Elsevier Science. New York, NY.

Kothari, P. K., Steinhardt, J., & Steurer, D. (2018, June). Robust Moment Estimation and Improved Clustering Via Sum of Squares. *In Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing* (pp. 1035-1046).

Linchpin. (2021). Hospitality Industry Challenges and Opportunities in 2021. Retrieved from: https://linchpinseo.com/common-challenges-facing-the-hospitality-industry/

Linguamatics. (2012). *Text Mining for Beginners.* [Video]. YouTube. https://www.youtube.com/watch?v=40QIW9Sr6Io

Maimon, O & Rokach, L. (2010). *Data Mining and Discovery Handbook.* Springer, New York, NY.

Mandelbaum, R. (2020). *Hotel Operators Adapt and Survive in 2020.* CBRE. https://www.cbrehotels.com/en/research/articles/hotel-operators-adapt-and-survive-in-2020

Merceron, A., Dierenfeld, H. (n.d.). Some Principles of Unsupervised Learning & Application in Education. beuth-hochschule.de

Naldi, M. (2019). A Review of Sentiment Computation Methods with R Packages. University of Rome Tor Vergata. Dpt. of Civil Engineering and Computer Science. *Via del Politecnico 1, 00133 Rome, Italy.* https://arxiv.org/pdf/1901.08319.pdf

Ng, A. (2021). *Introduction to Unsupervised Learning.* Retrieved from: https://www.coursera.org/lecture/machine-learning/unsupervised-learning-1VkCb

Qaiser, S., & Ali, R. (2018). Text mining: use of TF-IDF to examine the relevance of words to documents. *International Journal of Computer Applications*, *181*(1), 25-29.

Shirani, A. (n.d.). *An Introduction to Text Mining.* UMGC. https://learn.umgc.edu/d2l/le/content/583899/viewContent/21942392/View

StayNTouch Blog. (2018). *How Online Reviews Impact Hotel Revenue.* https://www.stayntouch.com/blog/how-online-reviews-impact-hotel-revenue/#:~:text=Consumers%20tend%20to%20shortlist%20hotels,%2D9%25%20increase%20in%20revene

UC. (2018). Business Analytics K-means Cluster Analysis R Programming Guide. uc-r.github.io

Van den Rul, C. (2019). *How to Generate Word Clouds in R.* Towards Data Science. https://towardsdatascience.com/create-a-word-cloud-with-r-bde3e7422e8a

Wattenberg, M., Viégas, F., & Johnson, I. (2016). How to Use t-SNE Effectively. *Distill*, 1(10), e2.

Welbers, K., Van Atteveldt, W., & Benoit, K. (2017). Text Analysis in R. *Communication Methods and Measures.* VOL. 11, NO. 4, 245–265. https://doi.org/10.1080/19312458.2017.1387238

Appendix

Outlier Function (Dhana, 2016):

```
outlierKD <- function(dt, var) {
    var_name <- eval(substitute(var),eval(dt))
    na1 <- sum(is.na(var_name))
    m1 <- mean(var_name, na.rm = T)
    par(mfrow=c(2, 2), oma=c(0,0,3,0))
    boxplot(var_name, main="With outliers")
    hist(var_name, main="With outliers", xlab=NA, ylab=NA)
    outlier <- boxplot.stats(var_name)$out
    mo <- mean(outlier)
    var_name <- ifelse(var_name %in% outlier, NA, var_name)
    boxplot(var_name, main="Without outliers")
    hist(var_name, main="Without outliers", xlab=NA, ylab=NA)
    title("Outlier Check", outer=TRUE)
    na2 <- sum(is.na(var_name))
    cat("Outliers identified:", na2 - na1, "n")
    cat("Propotion (%) of outliers:", round((na2 - na1) / sum(!is.na(var_name))*100, 1), "n")
    cat("Mean of the outliers:", round(mo, 2), "n")
    m2 <- mean(var_name, na.rm = T)
    cat("Mean without removing outliers:", round(m1, 2), "n")
    cat("Mean if we remove outliers:", round(m2, 2), "n")
    response <- readline(prompt="Do you want to remove outliers and to replace with NA? [yes/no]: ")
    if(response == "y" | response == "yes"){
        dt[as.character(substitute(var))] <- invisible(var_name)
        assign(as.character(as.list(match.call())$dt), dt, envir = .GlobalEnv)
        cat("Outliers successfully removed", "n")
        return(invisible(dt))
    } else{
        cat("Nothing changed", "n")
        return(invisible(var_name))
    }
}
```

Figure 30: California Hotels Wordcloud

Figure 31: Florida Hotels Wordcloud

Figure 32: Texas Hotels Wordcloud