Group 2 Text Analysis of Hotel Reviews

Theodore Fitch, Michael Goddard, Betapho Hannah, and Jeffrey Morey

University of Maryland Global Campus

DATA 630

Dr. Ami Gates

August 9, 2021

Introduction

With the increasing competition in the hospitality industry, tourism infrastructures have increasingly become digital - enhancing the interconnections amongst suppliers, firms, and customers. The popularity of the Internet in business operations has led to social media content, such as electronic word-of-mouth (eWOM) messages that provide shared information regarding a firm's products and services (Podnar & Javernik, 2012). As such, online customer reviews –an example of the eWOM – have emerged as the most relevant source of information for customer decision making (Filieri & McLeay, 2013) and are perceived to be more effective in influencing customer behavior than the traditional marketing information offered by product and service providers or the third-party websites (Yang & Mai, 2010).

From 2019 to 2020, the U.S Travel Association noted a spending decline of 42% ($500 billion), with international travel along with business travel experiencing the worst declines. (Singh & Wang, 2021). International travel spending dropped by 76% (as opposed to 34% of domestic travel), whereas business travel spending fell by 70% (as opposed to 27% for leisure travel) (Singh & Wang, 2021). Since the WHO declared the COVID-19 a global pandemic in March 2020, hotels globally have experienced sudden declines in occupancy. In 2019, hotel occupancy averaged 66% (Lardieri, 2021). However, this value reduced to a historic low of 24.5% by April 2020, with hotel revenue falling by about half to $84.6 billion in 2020 (Lardieri, 2021).

Although the market projections demonstrate positive prospects for the hotel industry once the pandemic is over, Airbnb and other similar platforms are transforming the market structure of the hospitality industry. The advent of COVID-19 at the beginning of 2020 accelerated the structural shift from traditional hotel accommodation to Airbnb accommodation options (Lardieri, 2021).

Airbnb offered short-term rental accommodation alternatives that complied with social distancing procedures, which most traditional hotel accommodations could not match. In turn, this attracted more guests. In this case, guests can book Airbnb accommodation, check-in, and enjoy their stay, devoid of contacting people outside their party.

Due to digitalization in the tourism industry, it is typical for online consumers to deal with massive amounts of online content, search engines, varying devices, and new strategic approaches for making a purchasing decision. Therefore, online reviews have become the most trusted source for customers when making e-commerce purchasing decisions.

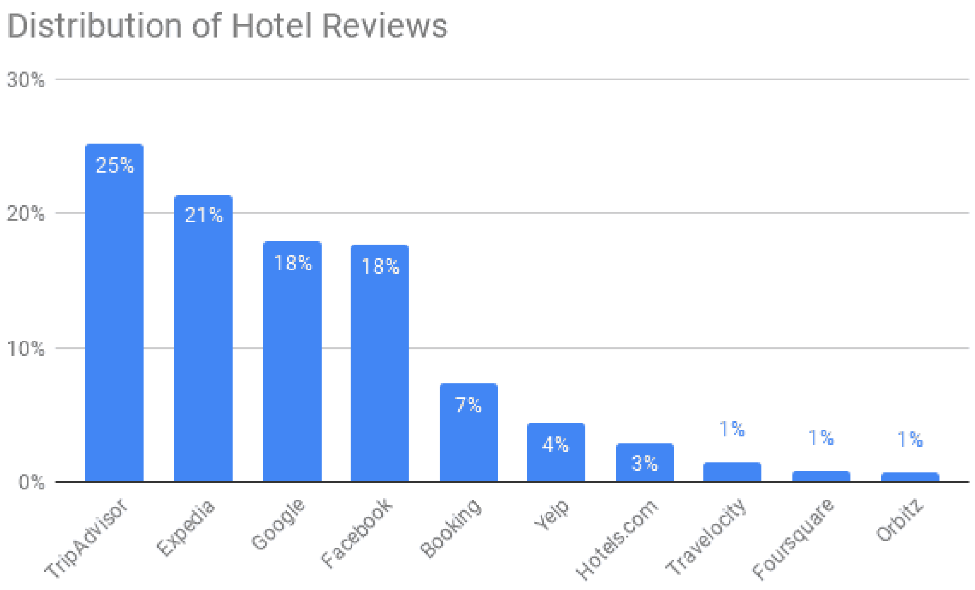
A Nielsen report in 2012 established that consumers' reviews are the second most trusted source of brand recognition after a recommendation from friends and family members (Consumer Trust in Online, Social, and Mobile Advertising Grows, 2012). Hotels are aware of this trend and are thus leveraging the power of the Internet to create platforms that gather and convey individual customer reviews on their services to attract a large customer base. A ComScore study in 2007 investigated the effect of consumer-generated ratings on the price customers were willing to pay for an item to be delivered offline (Online Consumer-Generated Reviews Have Significant Impact on Offline Purchase Behavior, 2007). According to ComScore, the study "revealed that consumers were willing to pay at least 20 percent more for services receiving an "Excellent," or 5-star, rating than for the same service receiving a "Good" or 4-star, rating" (Online Consumer-Generated Reviews Have Significant Impact on Offline Purchase Behavior, 2007).

Furthermore, amenities offered by hotels in the United States attract customers. These amenities include a fitness center, pool, continental breakfast, conference rooms, and a 5-star Michelin restaurant. Reviews vary depending on the level at which these amenities satisfy guests.

From statistics in Datafiniti's Business Database, several hotels have many reviews, while other hotels across the country have just one review (Kaggle, 2021). The data reveals that online reviews about (a) physical attributes, (b) food and drinks, (c) staff service, (d) hotel location, and (e) number of online reviews affect hotel booking considerations. Upon this rationale, the proposed study seeks to ascertain the influence of online ratings and reviews on hotel booking considerations. The study's findings would enable hotel managers to implement innovative services that enhance hotel reviews, increasing hotel bookings and profitability.

Lastly, hotels must monitor review sites, especially for negative reviews that could detract from hotel bookings. Figure 1 below provides one with a sense of the variety of sites where hotel reviews are stored. This potential for negative reviews is a reason to develop text mining and analysis tools to assess the impact these reviews can have on hotel operations.

Figure 1



Note: TripAdvisor, Expedia, Google, and Facebook account for nearly 75% of the hotel reviews submitted by guests, so one should ensure text mining and analysis tools can work with the reviews stored on these sites (The Hotel Review Sites You Should Monitor, 2020).

Analysis

Data Information and Processing

The Hotel Reviews dataset is available for download from the Kaggle website and provided by Datafiniti (Kaggle, 2021). The dataset consists of 35,912 observations and 19 variables. See Appendix A, Table 1. One fundamental assumption of this paper is the need to perform text mining to analyze hotel reviews, so one does not use a response variable to predict outcomes with supervised learning algorithms.

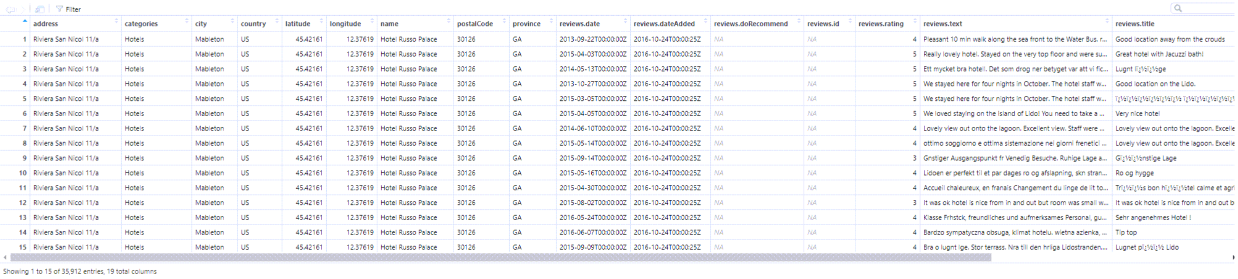
The fields for this dataset consist of hotel and review information. Each observation is an individual review. Most of the fields are factors with a high number of levels. All have too many levels to be used with specific classification algorithms. Twelve of the columns contain nulls that one will need to address. There is a more significant percentage of nulls in the reviewer information, but our analysis's vital information is the sentiment and the hotel information. In addition to the missing information, there is also clearly erroneous data in the dataset. For example, the province attribute has multiple length values greater than two, while most are abbreviations for U.S. states. Inconsistency in data exists with formats like some zip codes with five digits and others with ten digits.

Our group will approach Exploratory Data Analysis (EDA) in two phases. The first phase will be to understand the structure of the data. In this case, there is only one table that needs to be understood. Once the information is understood, the second phase is to understand the data distributions of the columns to decide how to address the conversion to a machine learning rectangular dataset compatible with the selected algorithms.

First, one can describe the data *qualitatively* using several R methods. Use the read.csv command to import the HotelReviews.csv file into a variable named "reviews," then use the View command to display the data.frame.

Figure 2

View Command Output for Original Dataset



	address	categories	city	country	latitude	longitude	name	postalCode	province	reviews.date	reviews.dateAdded	reviews.doRecommend	reviews.id	reviews.rating	reviews.text	reviews.title
1	Riviera San Nicol 11/a	Hotels	Mableton	US	45.42161	12.37619	Hotel Russo Palace	30126	GA	2013-09-22T00:00:00Z	2016-10-24T00:00:25Z	NA	NA	4	Pleasant 10 min walk along the sea front to the Water Bus. r...	Good location away from the crouds
2	Riviera San Nicol 11/a	Hotels	Mableton	US	45.42161	12.37619	Hotel Russo Palace	30126	GA	2015-04-03T00:00:00Z	2016-10-24T00:00:25Z	NA	NA	5	Really lovely hotel. Stayed on the very top floor and were su...	Great hotel with Jacuzzi bath!
3	Riviera San Nicol 11/a	Hotels	Mableton	US	45.42161	12.37619	Hotel Russo Palace	30126	GA	2014-05-13T00:00:00Z	2016-10-24T00:00:25Z	NA	NA	5	Ett mycket bra hotell. Det som drog ner betyget var att vi fic...	Lugnt ï¿½ï¿½ge
4	Riviera San Nicol 11/a	Hotels	Mableton	US	45.42161	12.37619	Hotel Russo Palace	30126	GA	2013-10-27T00:00:00Z	2016-10-24T00:00:25Z	NA	NA	5	We stayed here for four nights in October. The hotel staff w...	Good location on the Lido.
5	Riviera San Nicol 11/a	Hotels	Mableton	US	45.42161	12.37619	Hotel Russo Palace	30126	GA	2015-03-05T00:00:00Z	2016-10-24T00:00:25Z	NA	NA	5	We stayed here for four nights in October. The hotel staff w...	ï¿½ï¿½ï¿½ï¿½ï¿½ï¿½ ï¿½ï¿½ï¿½ï¿½ï¿½
6	Riviera San Nicol 11/a	Hotels	Mableton	US	45.42161	12.37619	Hotel Russo Palace	30126	GA	2015-04-05T00:00:00Z	2016-10-24T00:00:25Z	NA	NA	5	We loved staying on the island of Lido! You need to take a ...	Very nice hotel
7	Riviera San Nicol 11/a	Hotels	Mableton	US	45.42161	12.37619	Hotel Russo Palace	30126	GA	2014-06-10T00:00:00Z	2016-10-24T00:00:25Z	NA	NA	4	Lovely view out onto the lagoon. Excellent view. Staff were ...	Lovely view out onto the lagoon. Excelle
8	Riviera San Nicol 11/a	Hotels	Mableton	US	45.42161	12.37619	Hotel Russo Palace	30126	GA	2015-05-14T00:00:00Z	2016-10-24T00:00:25Z	NA	NA	4	ottimo soggiorno e ottima sistemazione nei giorni frenetici ...	Lovely view out onto the lagoon. Excelle
9	Riviera San Nicol 11/a	Hotels	Mableton	US	45.42161	12.37619	Hotel Russo Palace	30126	GA	2015-09-14T00:00:00Z	2016-10-24T00:00:25Z	NA	NA	3	Gnstiger Ausgangspunkt fr Venedig Besuche. Ruhige Lage a...	Gï¿½ï¿½nstige Lage
10	Riviera San Nicol 11/a	Hotels	Mableton	US	45.42161	12.37619	Hotel Russo Palace	30126	GA	2015-05-16T00:00:00Z	2016-10-24T00:00:25Z	NA	NA	4	Lidoen er perfekt til et par dages ro og afslapning, skn stran...	Ro og hygge
11	Riviera San Nicol 11/a	Hotels	Mableton	US	45.42161	12.37619	Hotel Russo Palace	30126	GA	2015-04-30T00:00:00Z	2016-10-24T00:00:25Z	NA	NA	4	Accueil chaleureux, en franais Changement du linge de lit to...	Trï¿½ï¿½s bon hï¿½ï¿½tel calme et agri
12	Riviera San Nicol 11/a	Hotels	Mableton	US	45.42161	12.37619	Hotel Russo Palace	30126	GA	2015-08-02T00:00:00Z	2016-10-24T00:00:25Z	NA	NA	3	It was ok hotel is nice from in and out but room was small w...	It was ok hotel is nice from in and out b
13	Riviera San Nicol 11/a	Hotels	Mableton	US	45.42161	12.37619	Hotel Russo Palace	30126	GA	2016-05-24T00:00:00Z	2016-10-24T00:00:25Z	NA	NA	4	Klasse Frhstck, freundliches und aufmerksames Personal, gu...	Sehr angenehmes Hotel !
14	Riviera San Nicol 11/a	Hotels	Mableton	US	45.42161	12.37619	Hotel Russo Palace	30126	GA	2016-06-07T00:00:00Z	2016-10-24T00:00:25Z	NA	NA	4	Bardzo sympatyczna obsuga, klimat hotelu. wietna azienka, ...	Tip top
15	Riviera San Nicol 11/a	Hotels	Mableton	US	45.42161	12.37619	Hotel Russo Palace	30126	GA	2015-09-09T00:00:00Z	2016-10-24T00:00:25Z	NA	NA	4	Bra o lugnt lge. Stor terrass. Nra till den hrliga Lidostranden....	Lugnet pï¿½ï¿½ Lido

Showing 1 to 15 of 35,912 entries, 19 total columns

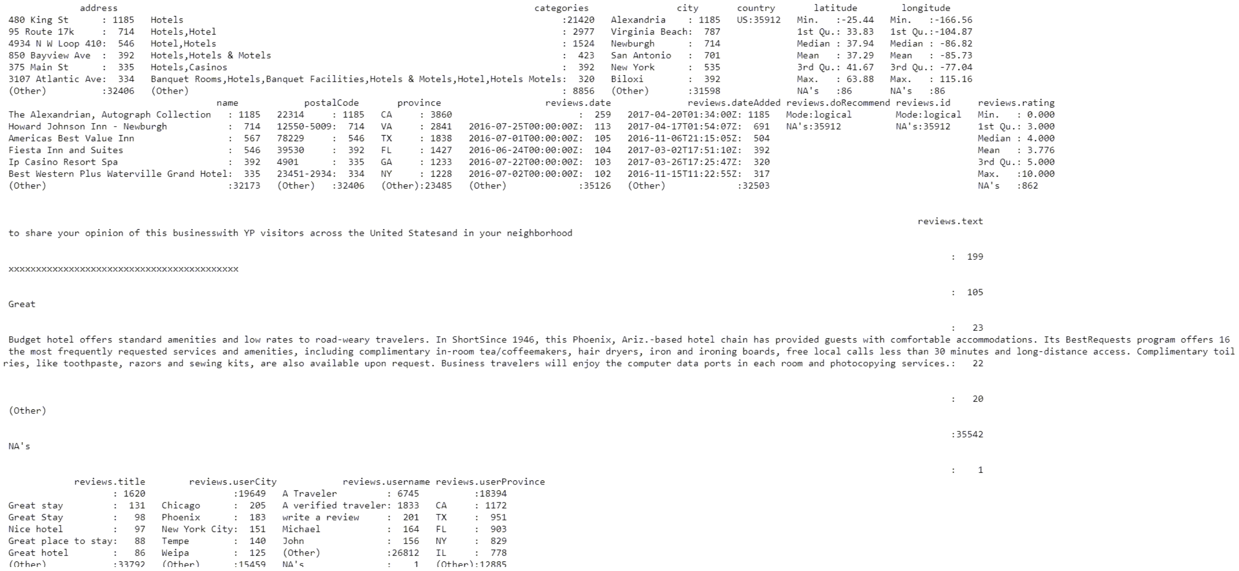
Note: The first fifteen rows of the original reviews data.frame object, for brevity purposes. Some column names are not displayed due to limitations while in document Portrait mode.

One key takeaway from executing the View command is that several columns may prove valuable for the analysis. In contrast, other columns will add no value to the analysis, such as the hotel address, id, and username of the hotel reviewer. Some reviews contain gibberish in the review title, so one must either remove the gibberish observations in the review title column or drop the column name from the analysis altogether. Some of the reviews contain foreign language, so one must either translate reviews with foreign language into English or remove the observations from the analysis. As mentioned earlier, the reviews contain several columns that can add value to the analysis, such as the latitude, longitude, city, province, and categories. Our team will elaborate on the use of such columns in the analysis section of this paper.

Second, one can review descriptive statistics for the dataset to understand the column names, sample values for the column names, how many times a value appears in the dataset, and vital statistical measures such as mean, median, mode, and quartiles.

Figure 3

Summary Command Output for Original Dataset



```
               address                                                                 categories              city          country       latitude        longitude     
 480 King St      : 1185   Hotels                                                           :21420   Alexandria    : 1185   US:35912   Min.   :-25.44   Min.   :-166.56  
 95 Route 17k     :  714   Hotels,Hotel                                                     : 2977   Virginia Beach:  787              1st Qu.: 33.83   1st Qu.:-104.87  
 4934 N W Loop 410:  546   Hotel,Hotels                                                     : 1524   Newburgh      :  714              Median : 37.94   Median : -86.82  
 850 Bayview Ave  :  392   Hotels,Hotels & Motels                                           :  423   San Antonio   :  701              Mean   : 37.29   Mean   : -85.73  
 375 Main St      :  335   Hotels,Casinos                                                   :  392   New York      :  535              3rd Qu.: 41.67   3rd Qu.: -77.04  
 3107 Atlantic Ave:  334   Banquet Rooms,Hotels,Banquet Facilities,Hotels & Motels,Hotel,Hotels Motels:  320   Biloxi        :  392              Max.   : 63.88   Max.   : 115.16  
 (Other)          :32406   (Other)                                                          : 8856   (Other)       :31598              NA's   :86       NA's   :86       
                                        name           postalCode        province                     reviews.date                  reviews.dateAdded reviews.doRecommend reviews.id      reviews.rating  
 The Alexandrian, Autograph Collection   : 1185   22314      : 1185   CA      : 3860                        :  259   2017-04-20T01:34:00Z: 1185   Mode:logical        Mode:logical    Min.   : 0.000  
 Howard Johnson Inn - Newburgh           :  714   12550-5009:  714   VA      : 2841   2016-07-25T00:00:00Z:  113   2017-04-17T01:54:07Z:  691   NA's:35912          NA's:35912      1st Qu.: 3.000  
 Americas Best Value Inn                 :  567   78229      :  546   TX      : 1838   2016-07-01T00:00:00Z:  105   2016-11-06T21:15:05Z:  504                                       Median : 4.000  
 Fiesta Inn and Suites                   :  546   39530      :  392   FL      : 1427   2016-06-24T00:00:00Z:  104   2017-03-02T17:51:10Z:  392                                       Mean   : 3.776  
 Ip Casino Resort Spa                    :  392   4901       :  335   GA      : 1233   2016-07-22T00:00:00Z:  103   2017-03-26T17:25:47Z:  320                                       3rd Qu.: 5.000  
 Best Western Plus Waterville Grand Hotel:  335   23451-2934:  334   NY      : 1228   2016-07-02T00:00:00Z:  102   2016-11-15T11:22:55Z:  317                                       Max.   :10.000  
 (Other)                                 :32173   (Other)   :32406   (Other):23485   (Other)             :35126   (Other)             :32503                                       NA's   :862     
                                                                                                                                                                                                  
                                                                                                                                                                                      reviews.text
 to share your opinion of this businesswith YP visitors across the United Statesand in your neighborhood                                                                                              
                                                                                                                                                                                      :  199  
 xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx                                                                                                                                                           
                                                                                                                                                                                      :  105  
 Great                                                                                                                                                                                                
                                                                                                                                                                                      :   23  
 Budget hotel offers standard amenities and low rates to road-weary travelers. In ShortSince 1946, this Phoenix, Ariz.-based hotel chain has provided guests with comfortable accommodations. Its BestRequests program offers 16
 the most frequently requested services and amenities, including complimentary in-room tea/coffeemakers, hair dryers, iron and ironing boards, free local calls less than 30 minutes and long-distance access. Complimentary toil
ries, like toothpaste, razors and sewing kits, are also available upon request. Business travelers will enjoy the computer data ports in each room and photocopying services.:   22  
                                                                                                                                                                                                     
                                                                                                                                                                                      :   20  
 (Other)                                                                                                                                                                                              
                                                                                                                                                                                      :35542  
 NA's                                                                                                                                                                                                 
                                                                                                                                                                                      :    1  
            reviews.title          reviews.userCity              reviews.username reviews.userProvince
                   : 1620                   :19649   A Traveler          : 6745           :18394      
 Great stay         :  131   Chicago        :  205   A verified traveler: 1833   CA     : 1172      
 Great Stay         :   98   Phoenix        :  183   write a review     :  201   TX     :  951      
 Nice hotel         :   97   New York City:  151   Michael            :  164   FL     :  903      
 Great place to stay:   88   Tempe          :  140   John               :  156   NY     :  829      
 Great hotel        :   86   Weipa          :  125   (Other)            :26812   IL     :  778      
 (Other)            :33792   (Other)        :15459   NA's               :    1   (Other):12885      
```

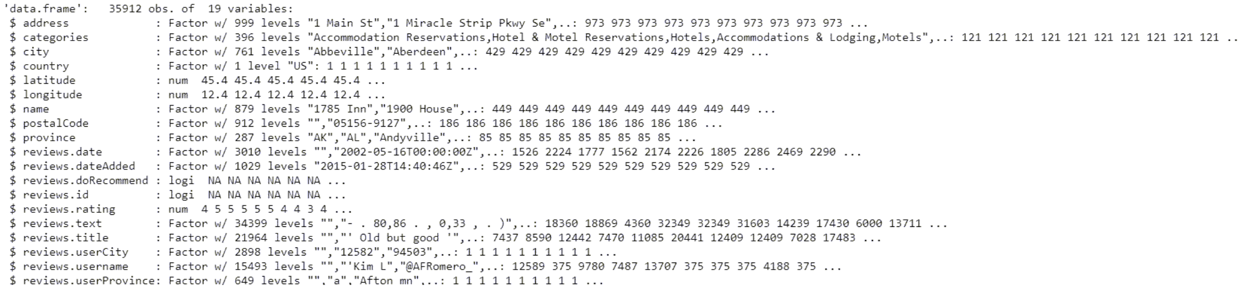
Note: The summary command output for the original reviews data.frame object.

One key takeaway from executing the Summary command is that the Categories column contains various values that describe the same category. For example, one observation contains the text "Hotels," and another observation contains the text "Hotels, Hotels," and another observation contains the text "Hotels, Hotels & Motels." If one performs an analysis by grouping hotel reviews based on category, one may find the effort challenging. One will need to normalize the category column to derive value during the analysis. Also, it is evident from the output that some columns contain numeric values while others contain text, so one will need to consider what normalization tasks one will need to perform on these columns.

Third, one can review the dataset structure to understand each variable's data types and the number of categories or levels. See Figure 4 on the next page for output from the structure command.

Figure 4

Structure Command Output for Original Dataset



```
'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,Motels",..: 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 ...
```

Note: The structure command output for the original reviews data.frame object.

One key takeaway from executing the structure command is that several variables contain many levels. For example, the *city* variable has 761 levels. One can deduce from so many levels that there is a legitimate, wide variety of cities or multiple observations where the city names are similar but not normalized. Another takeaway is that some of the observations contain null values, which is evident when viewing the reviews.id and reviews.doRecommend variables. One will need to decide how to handle the missing values during data cleaning activities. With the dataset structure, variables, and values reviewed, one can focus on data cleaning activities.

Data Cleaning and Feature Engineering

First, one can remove variables that add no value to the analysis. For example, the hotel address adds no value to the analysis. The country variable was a constant, thus providing no differentiation. Since the analysis will not include hotel address or country information, one can remove the address and country variables from the dataset. One can remove variables that identify information about the reviewer, which are the reviews.userCity and reviews.username variables. The reviews.id is a unique identifier that one can also remove. One should remove the reviews.doRecommend variable since it contains many missing values.

Second, one should normalize the review ratings since some of the reviews contained values from one to five, while others contained values from one to ten. Additionally, some of the review ratings contained decimal values. Using a custom function that divides the review rating by two, one can accomplish this normalization if the rating is greater than five and rounds the decimal values.

Third, a dataset review during data preprocessing activities revealed that the hotel province values were unreliable. To overcome this issue, one can leverage the latitude and longitude to map each set of coordinates to a state in the United States. This feature engineering effort revealed that a few coordinates mapped to hotels in overseas locations. For example, a hotel in Georgia had latitude/longitude coordinates in Italy. One can remove these observations without any significant impact on the overall analysis.

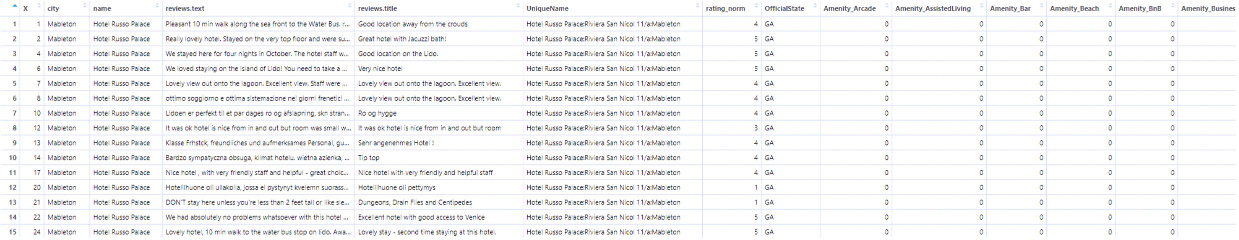
Fourth, numerous observations contained gibberish or foreign text. One can use a custom function that leveraged the grepl method in R to find all reviews that contain gibberish or foreign text and then remove those observations without any significant impact on the overall analysis.

Fifth, one might add value to the analysis by identifying and analyzing reviews based on amenities, such as a conference room, pool, fitness center, and other amenities that attract guests and influence the overall review. Unfortunately, the Hotel Reviews dataset does not contain an "amenities" variable. The dataset contains a "categories" column, which is a candidate for such analysis. However, the categories column contains 396 levels and values that are similar but not the same. For example, one review may contain a category value of "Conference Center," while another review may contain a category of "Conference." To address this issue, one can generalize the categories and then use the generalized categories to engineer a feature that enables hotel review analysis based on amenity.

Sixth, one completes additional cleanup to remove unnecessary columns. For example, one can remove the categories column since one has engineered an amenity feature. After completing data preprocessing activities, one should review the modified dataset to ensure it is ready for analysis. Figure 5 is an abbreviated view of the modified dataset, while Figure 6 depicts the structure.

Figure 5

View Command Output for Original Dataset

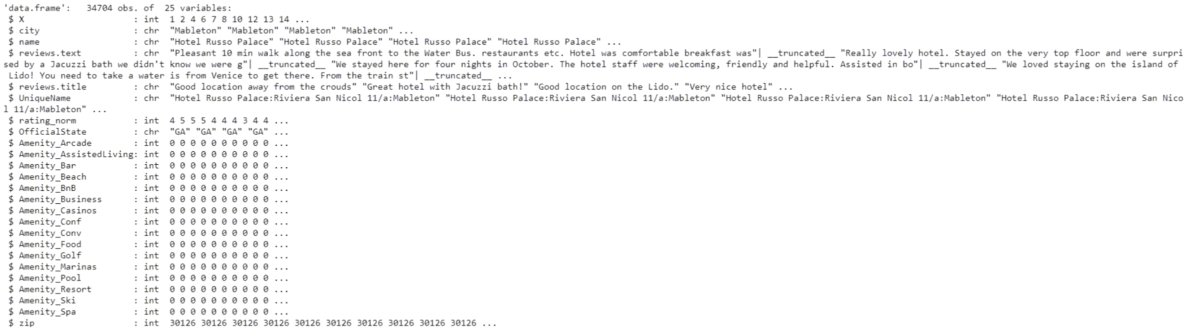


	X	city	name	reviews.text	reviews.title	UniqueName	rating_norm	OfficialState	Amenity_Arcade	Amenity_AssistedLiving	Amenity_Bar	Amenity_Beach	Amenity_BnB	Amenity_Busines
1	1	Mableton	Hotel Russo Palace	Pleasant 10 min walk along the sea front to the Water Bus. r...	Good location away from the crouds	Hotel Russo Palace:Riviera San Nicol 11/a:Mableton	4	GA	0	0	0	0	0	
2	2	Mableton	Hotel Russo Palace	Really lovely hotel. Stayed on the very top floor and were su...	Great hotel with Jacuzzi bath!	Hotel Russo Palace:Riviera San Nicol 11/a:Mableton	5	GA	0	0	0	0	0	
3	4	Mableton	Hotel Russo Palace	We stayed here for four nights in October. The hotel staff w...	Good location on the Lido.	Hotel Russo Palace:Riviera San Nicol 11/a:Mableton	5	GA	0	0	0	0	0	
4	6	Mableton	Hotel Russo Palace	We loved staying on the island of Lido! You need to take a ...	Very nice hotel	Hotel Russo Palace:Riviera San Nicol 11/a:Mableton	5	GA	0	0	0	0	0	
5	7	Mableton	Hotel Russo Palace	Lovely view out onto the lagoon. Excellent view. Staff were ...	Lovely view out onto the lagoon. Excellent view.	Hotel Russo Palace:Riviera San Nicol 11/a:Mableton	4	GA	0	0	0	0	0	
6	8	Mableton	Hotel Russo Palace	ottimo soggiorno e ottima sistemazione nei giorni frenetici ...	Lovely view out onto the lagoon. Excellent view.	Hotel Russo Palace:Riviera San Nicol 11/a:Mableton	4	GA	0	0	0	0	0	
7	10	Mableton	Hotel Russo Palace	Lidoen er perfekt til et par dages ro og afslapning, skn stran...	Ro og hygge	Hotel Russo Palace:Riviera San Nicol 11/a:Mableton	4	GA	0	0	0	0	0	
8	12	Mableton	Hotel Russo Palace	It was ok hotel is nice from in and out but room was small w...	It was ok hotel is nice from in and out but room	Hotel Russo Palace:Riviera San Nicol 11/a:Mableton	3	GA	0	0	0	0	0	
9	13	Mableton	Hotel Russo Palace	Klasse Frhstck, freundliches und aufmerksames Personal, gu...	Sehr angenehmes Hotel !	Hotel Russo Palace:Riviera San Nicol 11/a:Mableton	4	GA	0	0	0	0	0	
10	14	Mableton	Hotel Russo Palace	Bardzo sympatyczna obsuga, klimat hotelu. wietna azienka, ...	Tip top	Hotel Russo Palace:Riviera San Nicol 11/a:Mableton	4	GA	0	0	0	0	0	
11	17	Mableton	Hotel Russo Palace	Nice hotel , with very friendly staff and helpful - great choic...	Nice hotel with very friendly and helpful staff	Hotel Russo Palace:Riviera San Nicol 11/a:Mableton	4	GA	0	0	0	0	0	
12	20	Mableton	Hotel Russo Palace	Hotellihuone oli ullakolla, jossa ei pystynyt kvelemn suorass...	Hotellihuone oli pettymys	Hotel Russo Palace:Riviera San Nicol 11/a:Mableton	1	GA	0	0	0	0	0	
13	21	Mableton	Hotel Russo Palace	DON'T stay here unless you're less than 2 feet tall or like sle...	Dungeons, Drain Flies and Centipedes	Hotel Russo Palace:Riviera San Nicol 11/a:Mableton	1	GA	0	0	0	0	0	
14	22	Mableton	Hotel Russo Palace	We had absolutely no problems whatsoever with this hotel ...	Excellent hotel with good access to Venice	Hotel Russo Palace:Riviera San Nicol 11/a:Mableton	5	GA	0	0	0	0	0	
15	24	Mableton	Hotel Russo Palace	Lovely hotel, 10 min walk to the water bus stop on lido. Awa...	Lovely stay - second time staying at this hotel.	Hotel Russo Palace:Riviera San Nicol 11/a:Mableton	5	GA	0	0	0	0	0	

Note: The View command output for the modified reviews data.frame object. One has normalized the rating and established an official state feature based on latitude and longitude. One has also added columns to indicate if a hotel has specific amenities.

Figure 6

Structure Command Output for Modified Dataset



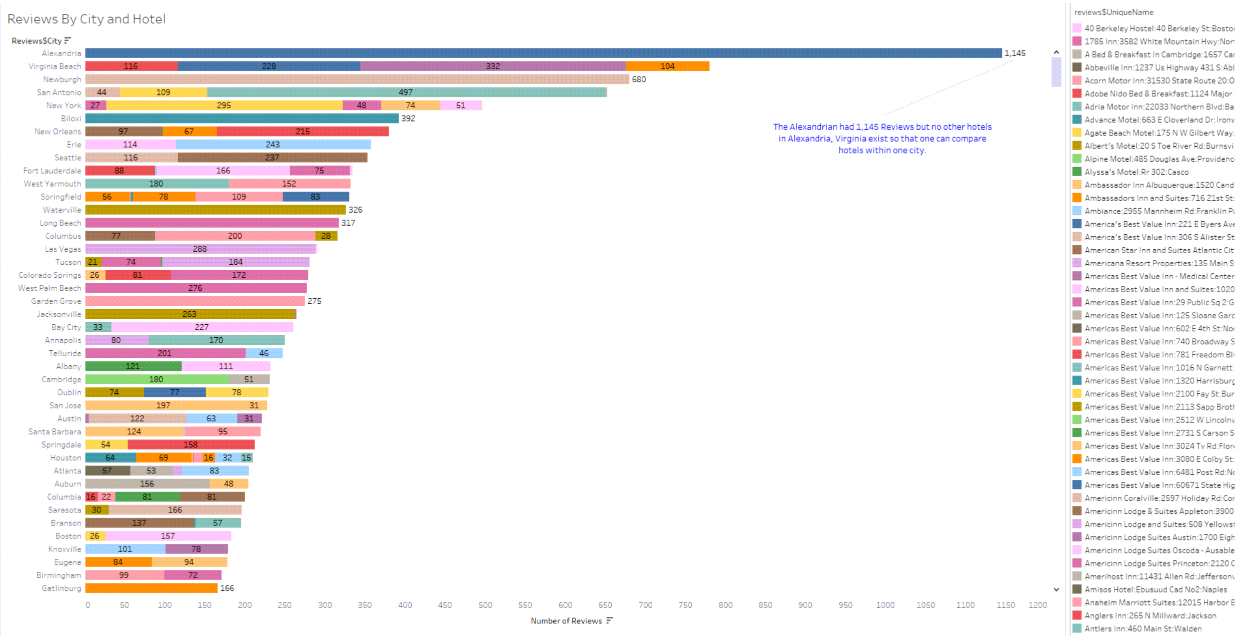
```
'data.frame':	34704 obs. of  25 variables:
 $ X                     : int  1 2 4 6 7 8 10 12 13 14 ...
 $ city                  : chr  "Mableton" "Mableton" "Mableton" "Mableton" ...
 $ name                  : chr  "Hotel Russo Palace" "Hotel Russo Palace" "Hotel Russo Palace" "Hotel Russo Palace" ...
 $ reviews.text          : chr  "Pleasant 10 min walk along the sea front to the Water Bus. restaurants etc. Hotel was comfortable breakfast was"| __truncated__ "Really lovely hotel. Stayed on the very top floor and were surpri
sed by a Jacuzzi bath we didn't know we were g"| __truncated__ "We stayed here for four nights in October. The hotel staff were welcoming, friendly and helpful. Assisted in bo"| __truncated__ "We loved staying on the island of
 Lido! You need to take a water is from Venice to get there. From the train st"| __truncated__ ...
 $ reviews.title         : chr  "Good location away from the crouds" "Great hotel with Jacuzzi bath!" "Good location on the Lido." "Very nice hotel" ...
 $ UniqueName            : chr  "Hotel Russo Palace:Riviera San Nicol 11/a:Mableton" "Hotel Russo Palace:Riviera San Nicol 11/a:Mableton" "Hotel Russo Palace:Riviera San Nicol 11/a:Mableton" "Hotel Russo Palace:Riviera San Nico
l 11/a:Mableton" ...
 $ rating_norm           : int  4 5 5 5 4 4 4 3 4 4 ...
 $ OfficialState         : chr  "GA" "GA" "GA" "GA" ...
 $ Amenity_Arcade        : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Amenity_AssistedLiving: int  0 0 0 0 0 0 0 0 0 0 ...
 $ Amenity_Bar           : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Amenity_Beach         : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Amenity_BnB           : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Amenity_Business      : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Amenity_Casinos       : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Amenity_Conf          : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Amenity_Conv          : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Amenity_Food          : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Amenity_Golf          : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Amenity_Marinas       : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Amenity_Pool          : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Amenity_Resort        : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Amenity_Ski           : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Amenity_Spa           : int  0 0 0 0 0 0 0 0 0 0 ...
 $ zip                   : int  30126 30126 30126 30126 30126 30126 30126 30126 30126 30126 ...
```

Note: The structure command output for the modified reviews data.frame object.

Lastly, further exploratory data analysis revealed some interesting insights into the data that would serve one well before selecting a text mining methodology and algorithm. For example, it is essential to determine if there are enough observations in the dataset to support the comparison of hotels within specific cities. To understand which hotels have the most reviews, one can count the number of reviews by the city and hotel name. It is worth noting that a high frequency of reviews does not correlate to hotel popularity. Figure 7 below illustrates that although certain cities have more hotels with reviews, not all cities have multiple hotels. For example, The Alexandrian hotel in Alexandria, Virginia, was the most frequently reviewed hotel with 1,145 reviews. However, this hotel is the only hotel in the city with reviews. Therefore, it is impossible to compare hotels with each other in this city.

Figure 7

Reviews by City and Hotel



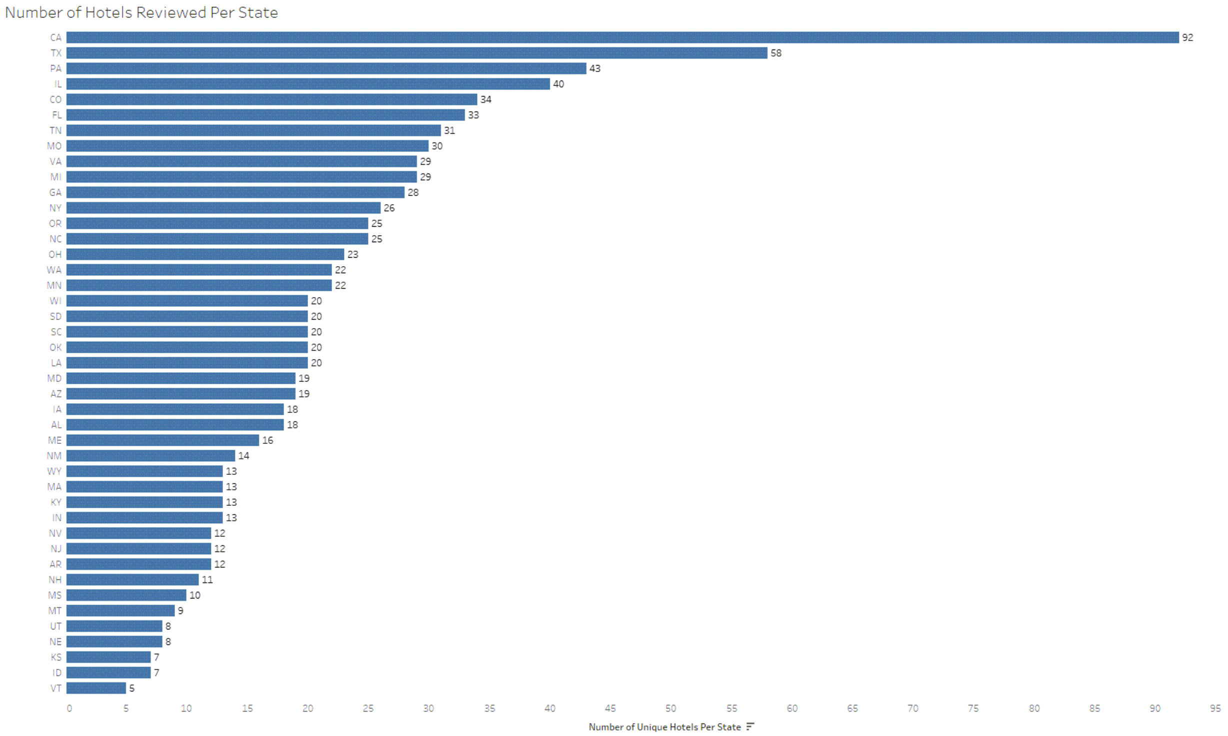
The Alexandrian had 1,145 Reviews but no other hotels in Alexandria, Virginia exist so that one can compare hotels within one city.

Note: One can use the count method in R to assess how many reviews are present by city and unique hotel name. There are not enough hotels in each city to effectively compare hotels within a city.

Further inspection of Figure 7 reveals that other cities, such as Biloxi, Long Beach, West Palm Beach, and Jacksonville, have only one hotel reviewed. This realization guided our team to compare hotels by the city but only for cities with at least four hotels. Additionally, when one counts the number of hotels reviewed per state, one can see that certain states have many hotels. One could use the information in Figure 8 later to choose reviews selectively for further analysis and comparison.

Figure 8

Unique Hotels, By Official State

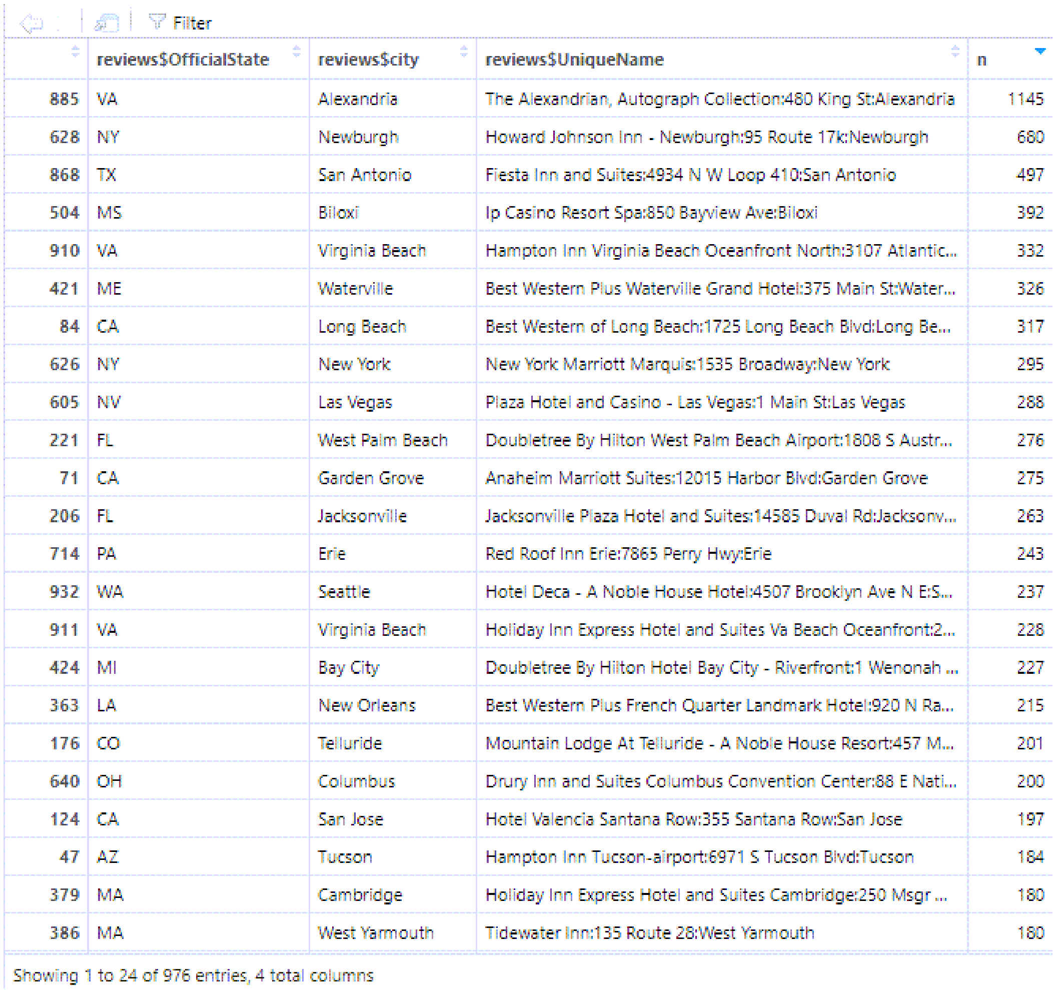


Note: One can use the count method in R to assess the number of unique hotels in each state. California tops the list with 92 unique hotels, while Vermont has the fewest unique hotels at 5.

One can execute the count method from the dplyr package in R to obtain a count for variables of interest from the modified dataset that includes features engineered and discussed in the data cleaning and feature engineering section. Figure 9 below provides a truncated view (for brevity) of the hotel review counts, grouped by official state.

Figure 9

Hotel Review Counts, By Official State, With City and Unique Hotel Name



	reviews$OfficialState	reviews$city	reviews$UniqueName	n
885	VA	Alexandria	The Alexandrian, Autograph Collection:480 King St:Alexandria	1145
628	NY	Newburgh	Howard Johnson Inn - Newburgh:95 Route 17k:Newburgh	680
868	TX	San Antonio	Fiesta Inn and Suites:4934 N W Loop 410:San Antonio	497
504	MS	Biloxi	Ip Casino Resort Spa:850 Bayview Ave:Biloxi	392
910	VA	Virginia Beach	Hampton Inn Virginia Beach Oceanfront North:3107 Atlantic...	332
421	ME	Waterville	Best Western Plus Waterville Grand Hotel:375 Main St:Water...	326
84	CA	Long Beach	Best Western of Long Beach:1725 Long Beach Blvd:Long Be...	317
626	NY	New York	New York Marriott Marquis:1535 Broadway:New York	295
605	NV	Las Vegas	Plaza Hotel and Casino - Las Vegas:1 Main St:Las Vegas	288
221	FL	West Palm Beach	Doubletree By Hilton West Palm Beach Airport:1808 S Austr...	276
71	CA	Garden Grove	Anaheim Marriott Suites:12015 Harbor Blvd:Garden Grove	275
206	FL	Jacksonville	Jacksonville Plaza Hotel and Suites:14585 Duval Rd:Jacksonv...	263
714	PA	Erie	Red Roof Inn Erie:7865 Perry Hwy:Erie	243
932	WA	Seattle	Hotel Deca - A Noble House Hotel:4507 Brooklyn Ave N E:S...	237
911	VA	Virginia Beach	Holiday Inn Express Hotel and Suites Va Beach Oceanfront:2...	228
424	MI	Bay City	Doubletree By Hilton Hotel Bay City - Riverfront:1 Wenonah ...	227
363	LA	New Orleans	Best Western Plus French Quarter Landmark Hotel:920 N Ra...	215
176	CO	Telluride	Mountain Lodge At Telluride - A Noble House Resort:457 M...	201
640	OH	Columbus	Drury Inn and Suites Columbus Convention Center:88 E Nati...	200
124	CA	San Jose	Hotel Valencia Santana Row:355 Santana Row:San Jose	197
47	AZ	Tucson	Hampton Inn Tucson-airport:6971 S Tucson Blvd:Tucson	184
379	MA	Cambridge	Holiday Inn Express Hotel and Suites Cambridge:250 Msgr ...	180
386	MA	West Yarmouth	Tidewater Inn:135 Route 28:West Yarmouth	180

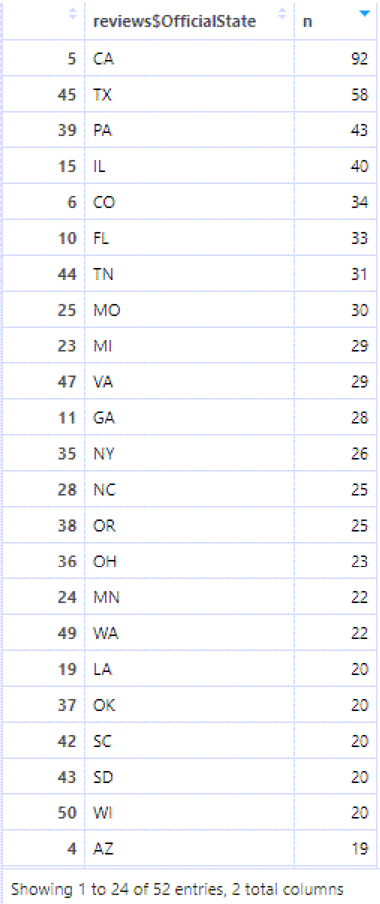
Showing 1 to 24 of 976 entries, 4 total columns

Note: This figure illustrates which hotel had the most reviews: The Alexandrian in Alexandria, VA. The Howard Johnson Inn in Newburgh, NY, had the second-highest number of reviews.

Alternatively, one can group the unique hotels within a state and obtain a count to determine if there are enough unique hotels to compare to hotels in other states. For example, a hotel chain with hotels in two states might compare hotel reviews to understand what hotel amenities and services resonate with guests for hotels in different geographical areas. Like the method described above, one can count the number of unique hotels within a state. Figure 10 below provides a truncated view (for brevity) of the unique hotels grouped by the official state.

Figure 10

Unique Hotels, By Official State



	reviews$OfficialState	n
5	CA	92
45	TX	58
39	PA	43
15	IL	40
6	CO	34
10	FL	33
44	TN	31
25	MO	30
23	MI	29
47	VA	29
11	GA	28
35	NY	26
28	NC	25
38	OR	25
36	OH	23
24	MN	22
49	WA	22
19	LA	20
37	OK	20
42	SC	20
43	SD	20
50	WI	20
4	AZ	19

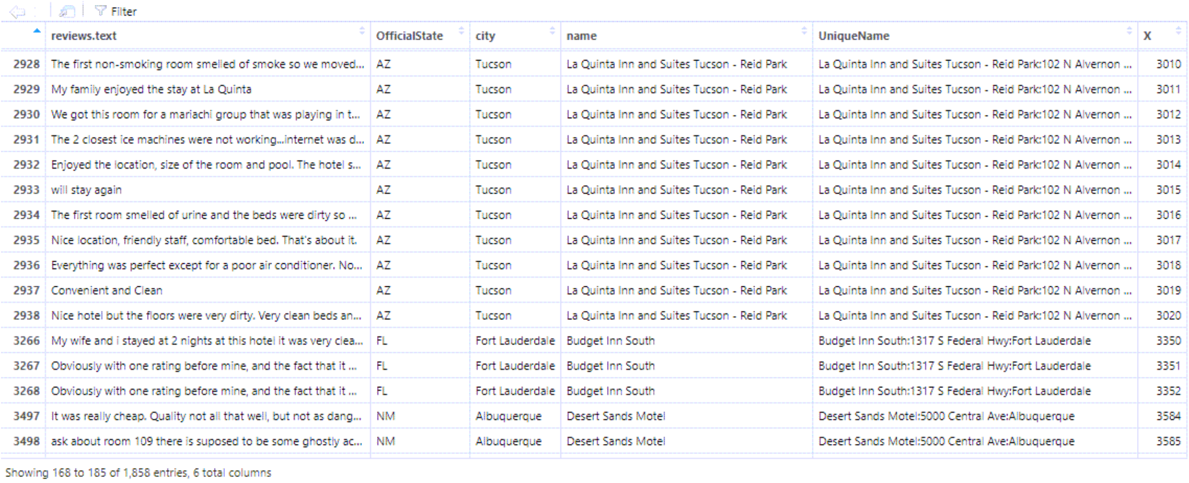
Showing 1 to 24 of 52 entries, 2 total columns

Note: This figure illustrates that some states have more unique hotels with reviews than other states. For example, California (C.A.) has 92 unique hotels, and Texas has 58 unique hotels. These two states would be good candidates for further hotel review analysis and comparison.

Equipped with this knowledge, one can now filter states and cities containing enough hotel review observations to produce results worth consideration for drawing conclusions and making recommendations to hotel management. To this end, one can use the *which* method in R to distill the cleaned dataset into a manageable dataframe for text analysis. Our team selected the top ten cities and states based on the earlier counts. Figure 11 below provides a truncated view (for brevity) of the top ten city reviews dataframe.

Figure 11

Original Top Ten City Reviews, Truncated View



	reviews.text	OfficialState	city	name	UniqueName	X
2928	The first non-smoking room smelled of smoke so we moved...	AZ	Tucson	La Quinta Inn and Suites Tucson - Reid Park	La Quinta Inn and Suites Tucson - Reid Park:102 N Alvernon ...	3010
2929	My family enjoyed the stay at La Quinta	AZ	Tucson	La Quinta Inn and Suites Tucson - Reid Park	La Quinta Inn and Suites Tucson - Reid Park:102 N Alvernon ...	3011
2930	We got this room for a mariachi group that was playing in t...	AZ	Tucson	La Quinta Inn and Suites Tucson - Reid Park	La Quinta Inn and Suites Tucson - Reid Park:102 N Alvernon ...	3012
2931	The 2 closest ice machines were not working...internet was d...	AZ	Tucson	La Quinta Inn and Suites Tucson - Reid Park	La Quinta Inn and Suites Tucson - Reid Park:102 N Alvernon ...	3013
2932	Enjoyed the location, size of the room and pool. The hotel s...	AZ	Tucson	La Quinta Inn and Suites Tucson - Reid Park	La Quinta Inn and Suites Tucson - Reid Park:102 N Alvernon ...	3014
2933	will stay again	AZ	Tucson	La Quinta Inn and Suites Tucson - Reid Park	La Quinta Inn and Suites Tucson - Reid Park:102 N Alvernon ...	3015
2934	The first room smelled of urine and the beds were dirty so ...	AZ	Tucson	La Quinta Inn and Suites Tucson - Reid Park	La Quinta Inn and Suites Tucson - Reid Park:102 N Alvernon ...	3016
2935	Nice location, friendly staff, comfortable bed. That's about it.	AZ	Tucson	La Quinta Inn and Suites Tucson - Reid Park	La Quinta Inn and Suites Tucson - Reid Park:102 N Alvernon ...	3017
2936	Everything was perfect except for a poor air conditioner. No...	AZ	Tucson	La Quinta Inn and Suites Tucson - Reid Park	La Quinta Inn and Suites Tucson - Reid Park:102 N Alvernon ...	3018
2937	Convenient and Clean	AZ	Tucson	La Quinta Inn and Suites Tucson - Reid Park	La Quinta Inn and Suites Tucson - Reid Park:102 N Alvernon ...	3019
2938	Nice hotel but the floors were very dirty. Very clean beds an...	AZ	Tucson	La Quinta Inn and Suites Tucson - Reid Park	La Quinta Inn and Suites Tucson - Reid Park:102 N Alvernon ...	3020
3266	My wife and i stayed at 2 nights at this hotel it was very clea...	FL	Fort Lauderdale	Budget Inn South	Budget Inn South:1317 S Federal Hwy:Fort Lauderdale	3350
3267	Obviously with one rating before mine, and the fact that it ...	FL	Fort Lauderdale	Budget Inn South	Budget Inn South:1317 S Federal Hwy:Fort Lauderdale	3351
3268	Obviously with one rating before mine, and the fact that it ...	FL	Fort Lauderdale	Budget Inn South	Budget Inn South:1317 S Federal Hwy:Fort Lauderdale	3352
3497	It was really cheap. Quality not all that well, but not as dang...	NM	Albuquerque	Desert Sands Motel	Desert Sands Motel:5000 Central Ave:Albuquerque	3584
3498	ask about room 109 there is suposed to be some ghostly ac...	NM	Albuquerque	Desert Sands Motel	Desert Sands Motel:5000 Central Ave:Albuquerque	3585

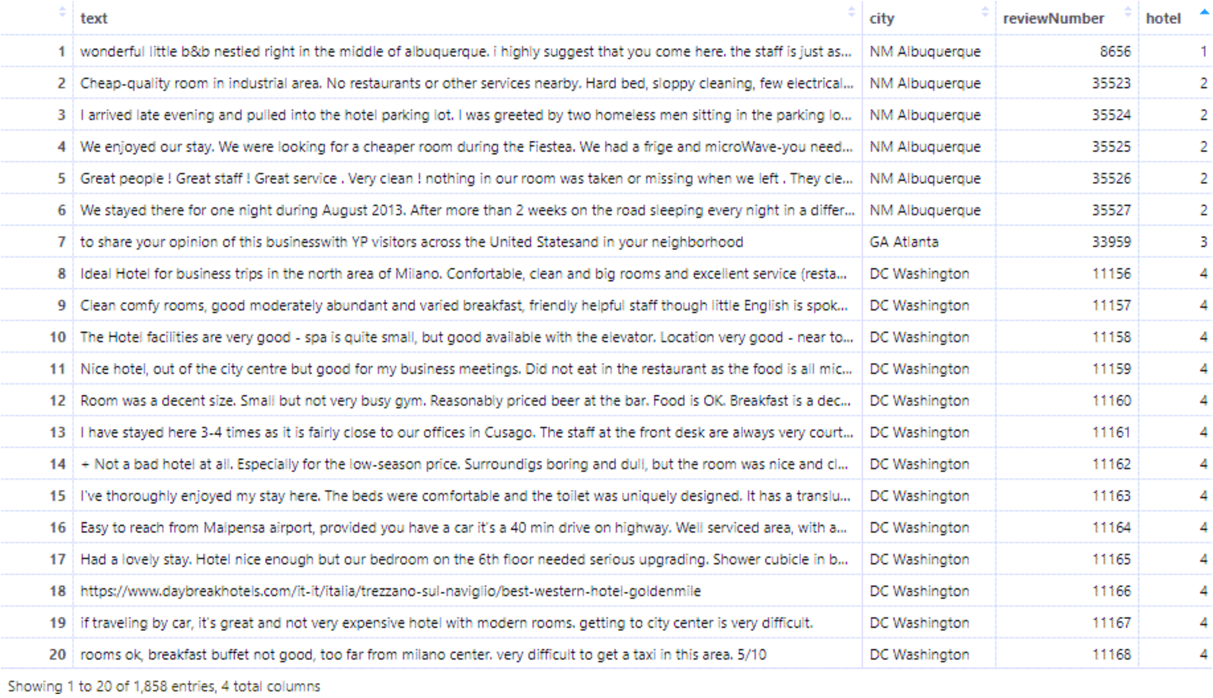
Showing 168 to 185 of 1,858 entries, 6 total columns

Note: This figure illustrates that reviews are available for multiple states, such as Florida (FL), Arizona (AZ), and New Mexico (NM). There are 1,858 entries in the top ten city reviews dataframe, which is sufficient to complete a targeted analysis of hotel reviews.

Although the top ten city reviews dataframe provides a distilled dataset to complete a targeted analysis of hotel reviews, one must complete some additional data transformation. For example, each hotel has a unique name, but there is no unique identifier to differentiate one hotel from another easily. The column labeled "X" represents the review number, but it is not very descriptive. The city and state columns are separate, so creating groups based on a city/state combination are difficult. Therefore, one must execute another sequence of commands to transform the top ten city reviews into a more usable format. See Figure 12 below.

Figure 12

Transformed Top Ten City Reviews, Truncated View



	text	city	reviewNumber	hotel
1	wonderful little b&b nestled right in the middle of albuquerque. i highly suggest that you come here. the staff is just as...	NM Albuquerque	8656	1
2	Cheap-quality room in industrial area. No restaurants or other services nearby. Hard bed, sloppy cleaning, few electrical...	NM Albuquerque	35523	2
3	I arrived late evening and pulled into the hotel parking lot. I was greeted by two homeless men sitting in the parking lo...	NM Albuquerque	35524	2
4	We enjoyed our stay. We were looking for a cheaper room during the Fiestea. We had a frige and microWave-you need...	NM Albuquerque	35525	2
5	Great people ! Great staff ! Great service . Very clean ! nothing in our room was taken or missing when we left . They cle...	NM Albuquerque	35526	2
6	We stayed there for one night during August 2013. After more than 2 weeks on the road sleeping every night in a differ...	NM Albuquerque	35527	2
7	to share your opinion of this businesswith YP visitors across the United Statesand in your neighborhood	GA Atlanta	33959	3
8	Ideal Hotel for business trips in the north area of Milano. Confortable, clean and big rooms and excellent service (resta...	DC Washington	11156	4
9	Clean comfy rooms, good moderately abundant and varied breakfast, friendly helpful staff though little English is spok...	DC Washington	11157	4
10	The Hotel facilities are very good - spa is quite small, but good available with the elevator. Location very good - near to...	DC Washington	11158	4
11	Nice hotel, out of the city centre but good for my business meetings. Did not eat in the restaurant as the food is all mic...	DC Washington	11159	4
12	Room was a decent size. Small but not very busy gym. Reasonably priced beer at the bar. Food is OK. Breakfast is a dec...	DC Washington	11160	4
13	I have stayed here 3-4 times as it is fairly close to our offices in Cusago. The staff at the front desk are always very court...	DC Washington	11161	4
14	+ Not a bad hotel at all. Especially for the low-season price. Surroundigs boring and dull, but the room was nice and cl...	DC Washington	11162	4
15	I've thoroughly enjoyed my stay here. The beds were comfortable and the toilet was uniquely designed. It has a translu...	DC Washington	11163	4
16	Easy to reach from Malpensa airport, provided you have a car it's a 40 min drive on highway. Well serviced area, with a...	DC Washington	11164	4
17	Had a lovely stay. Hotel nice enough but our bedroom on the 6th floor needed serious upgrading. Shower cubicle in b...	DC Washington	11165	4
18	https://www.daybreakhotels.com/it-it/italia/trezzano-sul-naviglio/best-western-hotel-goldenmile	DC Washington	11166	4
19	if traveling by car, it's great and not very expensive hotel with modern rooms. getting to city center is very difficult.	DC Washington	11167	4
20	rooms ok, breakfast buffet not good, too far from milano center. very difficult to get a taxi in this area. 5/10	DC Washington	11168	4

Showing 1 to 20 of 1,858 entries, 4 total columns

Note: This figure illustrates that reviews now have a unique review number and hotel identifier. One has also transformed the individual city and state columns into one column representing the city and state to group reviews by city/state combination.

Analysis and Model Methods

Once one has finished initial exploratory data analysis and data preprocessing activities, one can discuss the text mining algorithm and model methods used to perform text mining and analysis. As mentioned earlier, this paper aims to ascertain the influence of online ratings and reviews on hotel bookings and identify areas where a hotel could improve amenities and services. One can work towards this goal by exploring the reviews before selecting a text mining methodology and algorithm.

First, one can perform the text analysis using tidy text mining principles that extend and are compatible with earlier R text mining library methods, such as using the *tm* method to explore a corpus and build a document-term matrix. Instead, one utilizes a tidy text format with a specific data structure consisting of columns for each variable, rows for each observation, and a table for each type of observational unit (The Tidy Text Format, n.d.). Tidy text mining enables one to store a single word, an n-gram, a sentence, or a paragraph as a token (The Tidy Text Format, n.d.)

Second, one can execute additional methods to complete additional text mining tasks, such as removal of stop words, filtering, counting word frequencies, and sentiment analysis – to name a few. Furthermore, one can use the output from these tasks to produce plots with word frequency counts, word clouds, and positive and negative sentiment analysis insights.

To convert text into tokens, one uses the unnest_tokens method with word and text as input parameters. One uses the word parameter to specify that each word in a line of text (in this case, one hotel review) corresponds to a token. The unnest_tokens method produces a "tibble," - a modern class of dataframe in R that is available in packages such as the dplyr and tibble package (The Tidy Text Format, n.d.). One can easily manipulate the data using tidy tools after tokenizing the hotel reviews into a one-token-per-row format.

For example, one can remove stop words from three lexicons in the tidytext package along with a custom-built set of stop words appropriate to the hotel reviews dataset. Our team elected to use a custom set of stop words in conjunction with the three lexicons available in the tidytext package for this analysis. A review of the output described in the next paragraph revealed a need to implement custom stop words to eliminate words like hotel and text like "0", "1", and "2."

Third, one can use the count method included with the dplyr package to obtain a sorted list of most frequently used words for the narrowed list of cities and states selected for the analysis. After one produces the sorted list, one can utilize the ggplot package to display a bar chart of words and counts for each word.

Fourth, one can compare the average review rating between cities to assess how cities ranked relative to each other. For example, one can compare the average review rating for Atlanta, Georgia hotels to hotels in other cities. Not only could hotels benefit from this information, but event planners could also use this information to select venues close to hotels in cities with more robust reviews. Guests could also use this information to avoid booking hotel stays in cities with poor ratings. Furthermore, a hotel chain experiencing positive reviews in one city might elect to implement features or enhance services in another city to increase bookings and profitability. A hotel chain could also identify cities where hotels are not as widely reviewed and initiate a marketing campaign or a system to reward guests for submitting hotel reviews. One can also compare overall sentiment scores between cities to determine which cities have the most positive sentiment and which ones have a less positive or negative sentiment.

Fifth, one can explore the most powerful words in the corpus using TF-IDF scores. TF-IDF stands for Term Frequency-Inverse Document Frequency and is formed by multiplying two values: TF*IDF. (Scott, 2019). Term frequency is simply the number of times a word appears divided by the number of total words in a document. DF is the number of documents in the total corpus of words in which a word will appear. So, a document could be a chapter where the corpus is the book. A document would be a city in this analysis, whereas the corpus is the top ten cities. IDF is that number inversed (1/DF) (Scott, 2019). This application of TF-IDF shows the power of a word and its uniqueness to a particular city.

Sixth, no text analysis would be complete without understanding the opinions or sentiment of hotel guests who have provided an online review. One of the more common approaches to programmatically detect sentiment is to consider the text as a combination of individual words and the sentiment content of the whole text. For human readers, one can infer if a hotel review is positive or negative by reading and understanding the emotional intent of words within the review or other emotional nuances such as anger or disappointment (Sentiment Analysis with Tidy Data, n.d.). Therefore, one goal of sentiment analysis is to select and implement an algorithm capable of analyzing individual words and the combination of words in a passage of text to determine if a hotel review is positive or negative or full of a specific emotion.

To this end, three general-purpose lexicons are available for sentiment analysis using the tidytext package: AFINN from Finn Arup Nielsen, bing from Bing Liu and collaborators, and nrc from Saif Mohammad and Peter Turney (Sentiment Analysis with Tidy Data, n.d.). The lexicons contain many English words with assigned scores for positive/negative sentiment, emotions like joy, anger, disgust, or a scaled sentiment score ranging between -5 (more negative) to 5 (more positive) (Sentiment Analysis with Tidy Data, n.d.). One can leverage the get_sentiments method in the tidytext package while specifying the lexicon to determine sentiment scores.

For example, to obtain scaled sentiment scores using the AFINN lexicon, one can execute the *get_sentiments("afinn")* command against a specific tibble. The sentiment lexicons were constructed and validated by either the authors or using crowdsourcing mechanisms, such as

Amazon Mechanical Turk, so one must carefully assess the applicability of the lexicons against the subject domain (Sentiment Analysis with Tidy Data, n.d.).

After executing the get_sentiments method using the lexicons, one can plot the results using the popular ggplot library. One can also compare sentiment lexicons by binding the three lexicons together and displaying the most common positive and negative words on one plot.

Lastly, one can explore relationships between words by examining adjacent words or words that appear closely in the same hotel review. One can use the unnest_tokens function to tokenize reviews based on words and based on sequences of consecutive words, known as n-grams (Relationships Between Words: n-grams and Correlations, n.d.). One can explore words that appear in pairs, known as bigrams, which could be helpful for text analysis of hotel reviews.

For example, if bigrams like "bad service," "dirty pool," or "rude staff" appear frequently, one could conclude that these are areas of improvement for the hotel. One may also find it helpful to visualize words by arranging them in a network, where one represents each word as a node (or vertex) on a graph with lines (or edges) connecting each node. One can represent the strength of the relationship using line weight, with darker lines reflecting stronger relationships between words.

To summarize, the text mining and analysis steps are as follows:

1. Complete all data information and preprocessing activities
2. Identify features that will add value to text mining and analysis effort
3. Add engineered features into the dataset for further analysis
4. Identify and select an algorithm to granularly analyze the text
5. Use the selected algorithm to break the text into tokens and n-grams
6. Leverage stop word lexicons to exclude words with little or no value

7. Utilize filters to display categories or entities of interest
8. Plot word frequency counts to understand which words appear most often
9. Explore relationships between words using n-grams and network graphs
10. Complete sentiment analysis to understand the emotion behind the words

This text mining and analysis approach assumes that the text to be analyzed is large enough to break into tokens. The process for determining stop words is iterative in that one may have to plot word frequency first, assess which words add no value to the analysis, and then customize the stop words. This approach also assumes that one may need to narrow down the scope of the analysis by filtering.

Results

First, one can use several methods to complete text mining tasks, such as removing stop words, filtering, counting word frequencies, and sentiment analysis. One initial task was to generate a word frequency plot using hotel reviews for the top ten cities. However, to generate this plot, one must first break the text for each hotel review into individual words or tokens. One can execute the tidy *unnest_tokens* method and parameters to specify which text to tokenize and what type of tokens to create (words, sentences, or n-grams). For example, to tokenize the top ten city reviews, one can execute the following command:

```
tidy_top10CityReviews <- top10CityReviews %>%  unnest_tokens(word, text)
```

Second, one can preview the results of this modern dataframe, called a tibble. A tibble consists of multiple rows and columns. Each row contains a body of text that could be a word, n-gram, or sentence, depending on which options one specifies as parameters when executing the unnest_tokens method. See Table 2 on the next page for the top ten city reviews tibble.

Table 2

Tibble From Top Ten City Reviews, Truncated

```
# A tibble: 78,508 x 4

   city       reviewNumber hotel word
   <fct>             <int> <int> <chr>
 1 TX Houston           82     7 if
 2 TX Houston           82     7 you
 3 TX Houston           82     7 are
 4 TX Houston           82     7 being
 5 TX Houston           82     7 treated
 6 TX Houston           82     7 at
 7 TX Houston           82     7 any
 8 TX Houston           82     7 texas
 9 TX Houston           82     7 medical
10 TX Houston           82     7 center
# ... with 78,498 more rows
```

Note: This tibble has 78,508 rows of four columns. The last column is the word, or token, to be analyzed further.

The tibble has 78,508 rows of four columns and lists the city/state combination for each row, the review number, the unique hotel identifier, and each word extracted from the reviews.

Third, one can use the bind_rows method to create a custom tibble of stop words which one can use to remove words of no interest. After the initial generation of a word frequency plot, our team discovered that certain words were of no interest in the analysis. For example, it makes sense that the word "hotel" would frequently appear in a tibble of hotel reviews but using that word in text analysis adds no value. One can eliminate that word by adding it to the stop words list. One can then display the custom stop words tibble that one will use later to ensure one does not include specific stop words in the analysis. See Table 3 below.

Table 3

Tibble of Custom Stop Words

```
# A tibble: 1,160 x 2
   word  lexicon
   <chr> <chr>
 1 hotel custom
 2 0     custom
 3 1     custom
 4 2     custom
 5 3     custom
 6 4     custom
 7 5     custom
 8 6     custom
 9 7     custom
10 8     custom
# ... with 1,150 more rows
```

Note: This tibble has 1,160 rows of two columns. The word column represents the stop word.

Fourth, one can execute the tidy *anti_join* method to remove the custom stop words from the top ten city reviews tibble that one generated earlier. The anti_join method also removes the stop words included by default with the three tidytext lexicons. The three lexicons are extensive, with many frequently used words in the English language such as "the," "a," "and" "he," and "she" included. Once one executes the anti_join method, one can pipe the top ten city reviews tibble (which has had the stop words removed) to a count method with word and sort = TRUE as input parameters. See Table 4 on the next page.

Table 4

Tibble of Word Frequency Counts from Hotel Reviews of Top Ten Cities

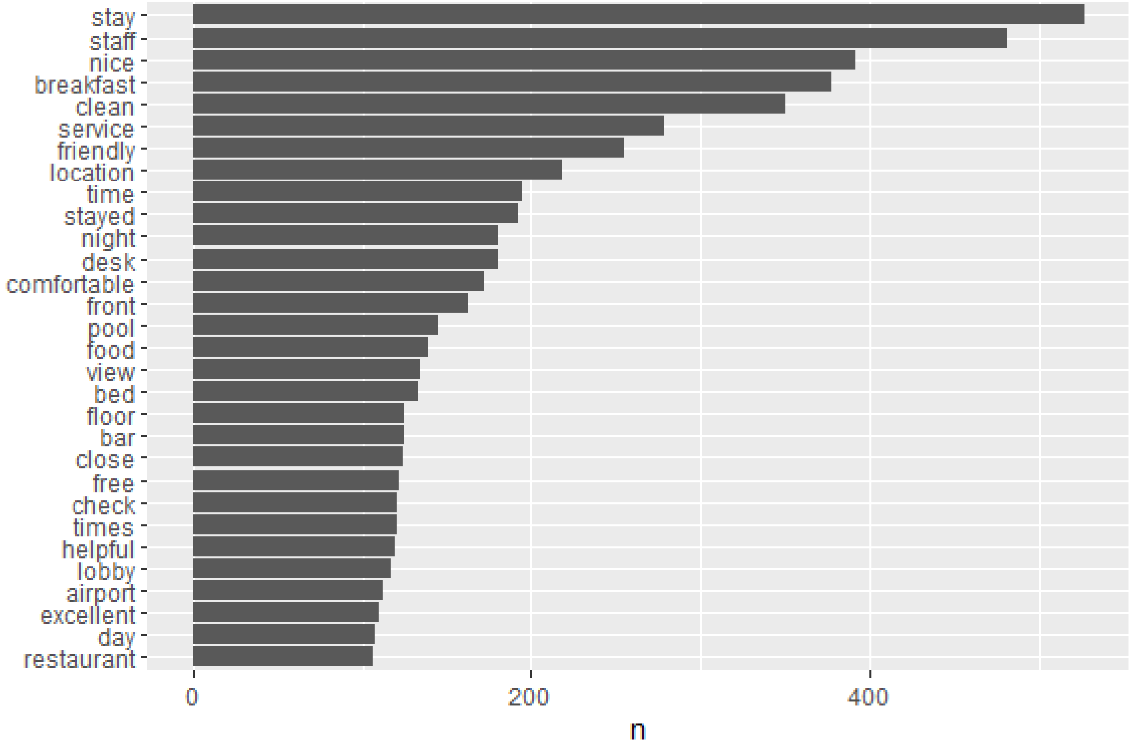
```
# A tibble: 5,419 x 2
   word        n
   <chr>     <int>
 1 stay        527
 2 staff       481
 3 nice        392
 4 breakfast   377
 5 clean       350
 6 service     278
 7 friendly    255
 8 location    218
 9 time        195
10 stayed      193
# ... with 5,409 more rows
```

Note: This tibble has 5,419 rows of two columns. The word column represents the word in the review, while the "n" column represents the number of times the word appeared.

Fifth, one can pipe the top ten city reviews tibble to several methods such as count, filter, mutate, and ggplot to produce a word frequency plot as indicated in Figure 13 on the next page. The word frequency plot reveals that the word "stay" appears the most frequently, followed by the words "staff" and "nice." Words such as "excellent," "day," and "restaurant" appear less frequently, but may still be words the hotel wants to be aware are appearing in online reviews.

Figure 13

Word Frequency Plot for Hotel Reviews from Top 10 Cities



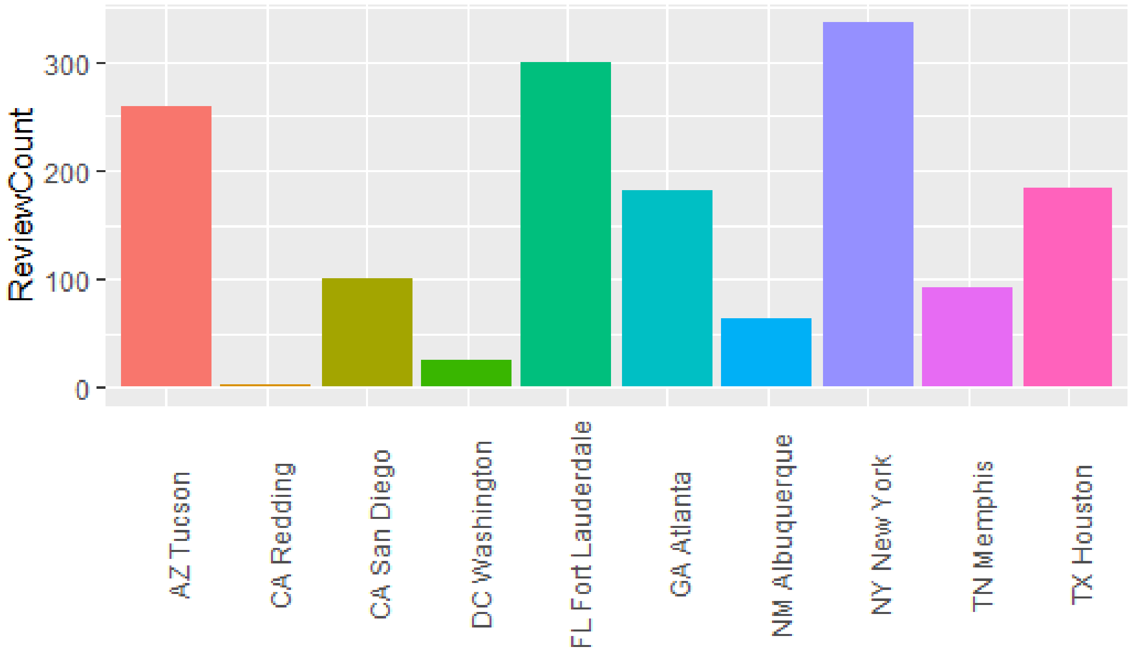
Note: This word frequency plot helps one visualize which words appear most frequently in the top ten cities tibble. One can use this plot to understand what amenities and hotel features guests write about the most.

Noteworthy, some words on the word frequency plot are related to customer service and others related to amenities. For example, breakfast, pool, bar, and restaurant are related to amenities, while staff, friendly, and helpful are related to customer service. One could use this information to deduce that these amenities and customer service traits are essential to guests and expend effort to advertise and improve these areas. Also, it is essential to note that this word frequency plot does not consider surrounding words for context. This lack of word context could lead one to conclude that guests enjoy the pool area when the reviews may reflect those guests did not like the "dirty pool." Lack of context in this word frequency plot is even more reason to consider performing a bi-gram analysis to understand word pair frequency.

Sixth, one can generate a plot to depict the number of hotel reviews for the top ten cities selected for the text analysis. Figure 14 below indicates that New York, NY had the highest number of reviews while Redding, CA had the lowest. Also, Fort Lauderdale, FL, and Tucson, AZ had many reviews, while Washington, DC and Albuquerque, NM had very few reviews.

Figure 14

Hotel Reviews Count for Top Ten Cities

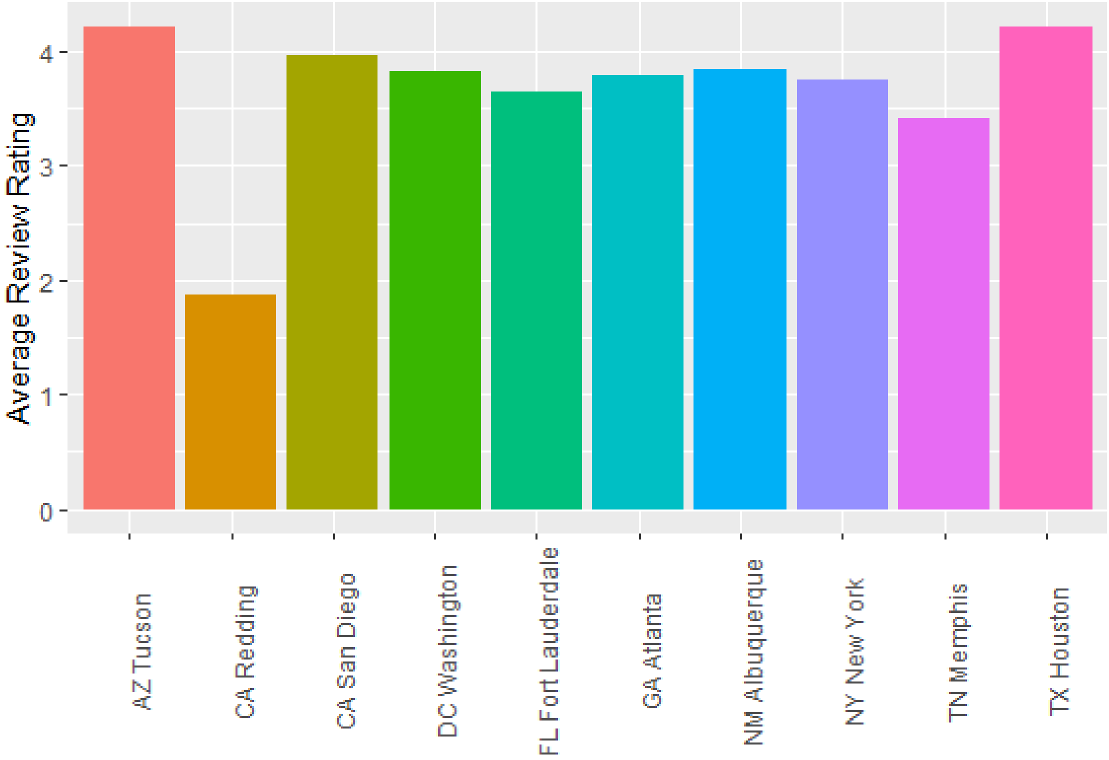


Note: This chart illustrates that New York, NY had the highest number of reviews while Redding, CA, has the lowest.

Seventh, one can generate a plot to depict the average review rating for hotels in each of the top ten cities. See Figure 15 on the next page. All the cities except for Redding, CA, had ratings near 3.5 and 4.0. Further inspection of the Redding, CA reviews revealed only eight reviews. Six of the reviews were for "Pizza Hut," with a review rating of 3, while three of the hotels in the same city had a review rating of zero. With such little data and low ratings for the three hotels, it is no surprise that the average review rating is low for this city.

Figure 15

Average Review Rating for Top Ten Cities

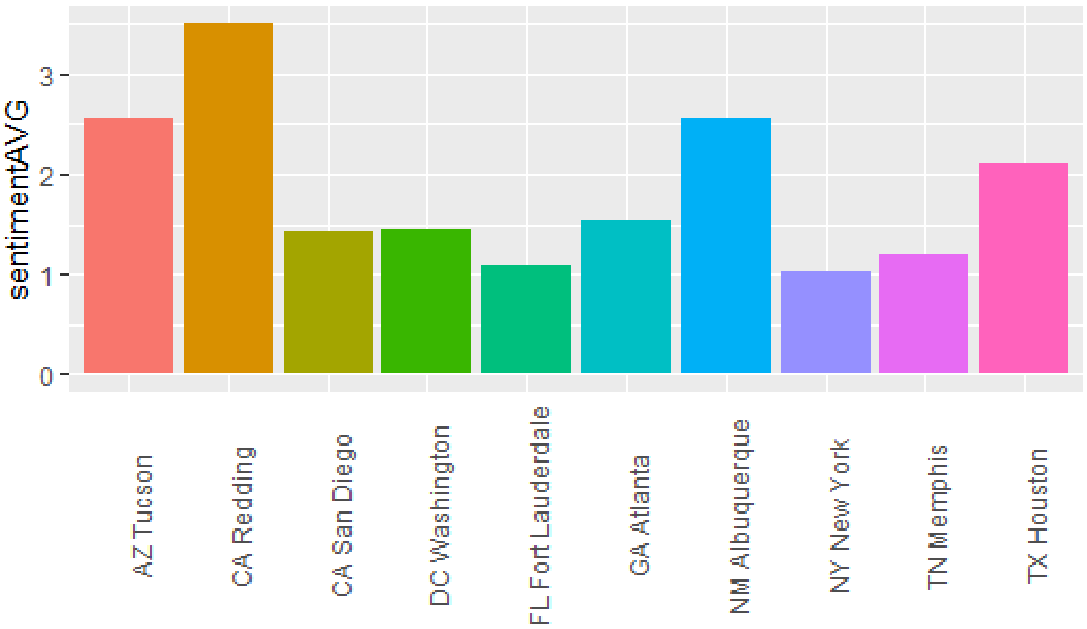


Note: This chart illustrates that most hotels had an average review rating near 3.5 and 4.0, except for Redding, CA. Further inspection of the reviews revealed that there were only eight reviews for hotels in the city, and three of the reviews had a rating of zero.

Eighth, one can generate a plot to depict the sentiment average review rating for hotels in each of the top ten cities. From Figure 16 on the next page, one can deduce that there was a lot of variability in the sentiment average. Interestingly, even though Redding, CA had a lower-than-average hotel rating, it had an extremely high sentiment average. However, further inspection of the Redding, CA reviews revealed that it appears someone from management inserted positive comments into the reviews. Two of the reviews contained the following text: "Great Pizza and Great people!! Come check out the new lobby!!!" Another review sounded like a write-up for a TV commercial. One can infer those guests did not submit these reviews and that the use of certain positive words unfairly influenced the average sentiment score for this city

Figure 16

Sentiment Average for Top Ten Cities

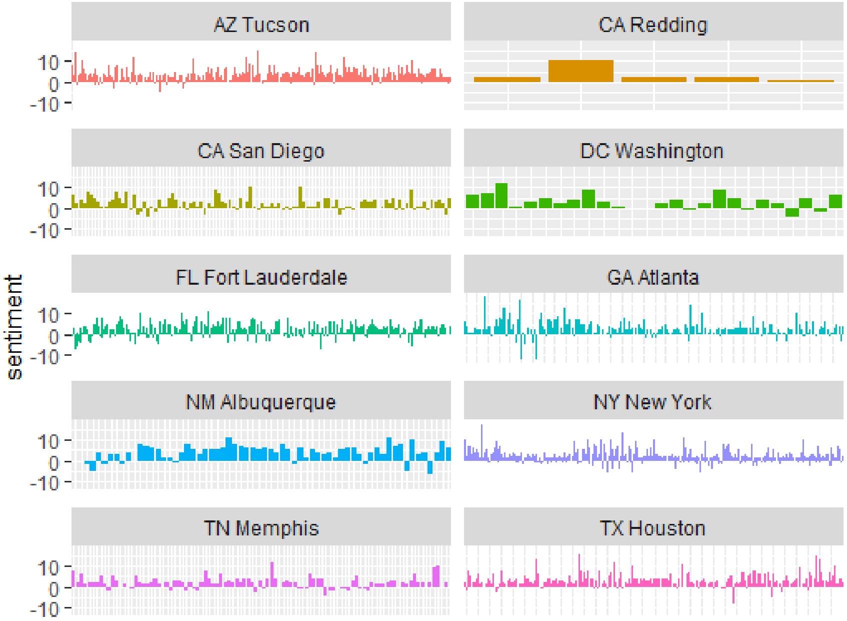


Note: This chart depicts the sentiment average for the top ten cities. Although Redding, CA had extremely high sentiment, further inspection revealed management likely submitted some reviews. The use of positive words in reviews by management can unfairly influence the sentiment average.

Ninth, one can generate a plot to compare the sentiment scores for the top ten cities. One can use the Bing analysis previously discussed to generate a sentiment score, with positive and negative words in a review evaluated to capture the feeling of a review. In Figure 17 on the next page, one can see that Redding stands out with no negative sentiment scores (and explained earlier). New York has the most variability, with sentiment scores above and below zero. One can also see that negative sentiment scores are significant, but they are all lower than many of the highest positive sentiment scores. One can deduce that customers who have negative experiences will usually write shorter reviews and not use as much negative language in their reviews. While businesses need to address negative reviews, one can also argue that businesses need to encourage pleased customers to write positive reviews.

Figure 17

Sentiment Score Comparison Matrix for Top Ten Cities

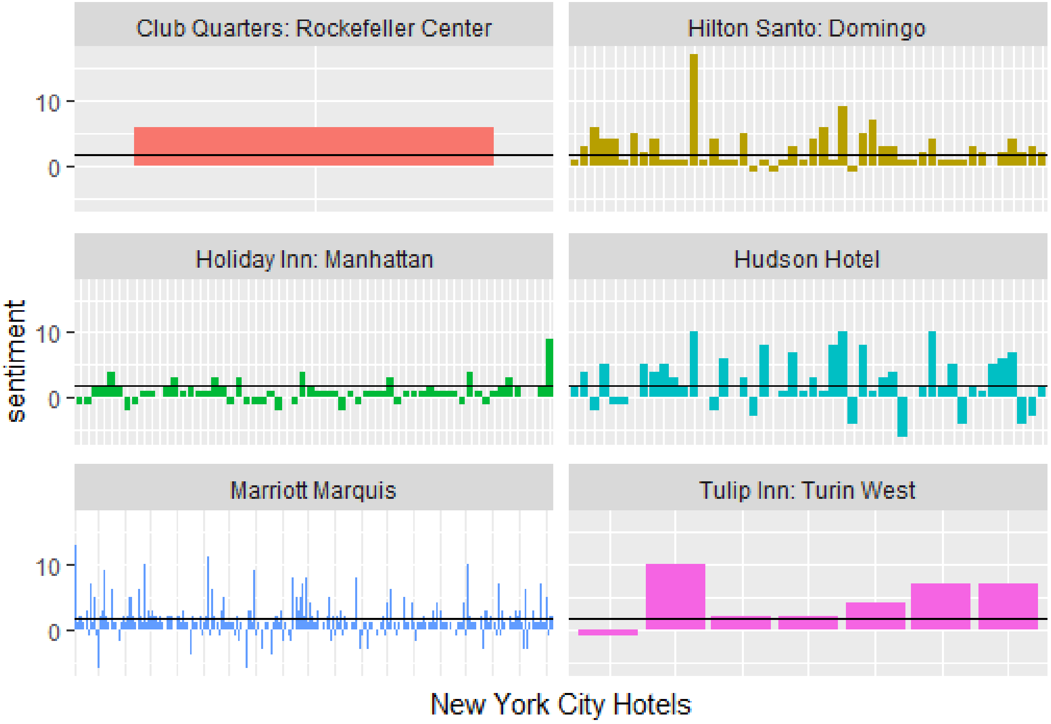


Note: This chart illustrates the sentiment score of each review for the top 10 cities. This score is the overall feeling of a review. There is some variability in all cities with primarily positive scores. Even the most significant negative scores are smaller than the most significant positive scores.

Tenth, one can generate a plot to compare the sentiment scores for hotels in New York City, which is a city that travelers visit often. In Figure 18 on the next page, one can see that there is quite a bit of variability in the number of reviews for the eight hotels analyzed in New York City, with the lowest of 1 and the highest of nearly 300. A similar pattern exists in Figure 17, where the lowest negative sentiment reviews are smaller than the highest positive sentiment reviews. This pattern shows that customers with positive experiences should be encouraged to write about their experiences at a more granular level.

Figure 18

Sentiment Score Comparison Matrix for New York City Hotels



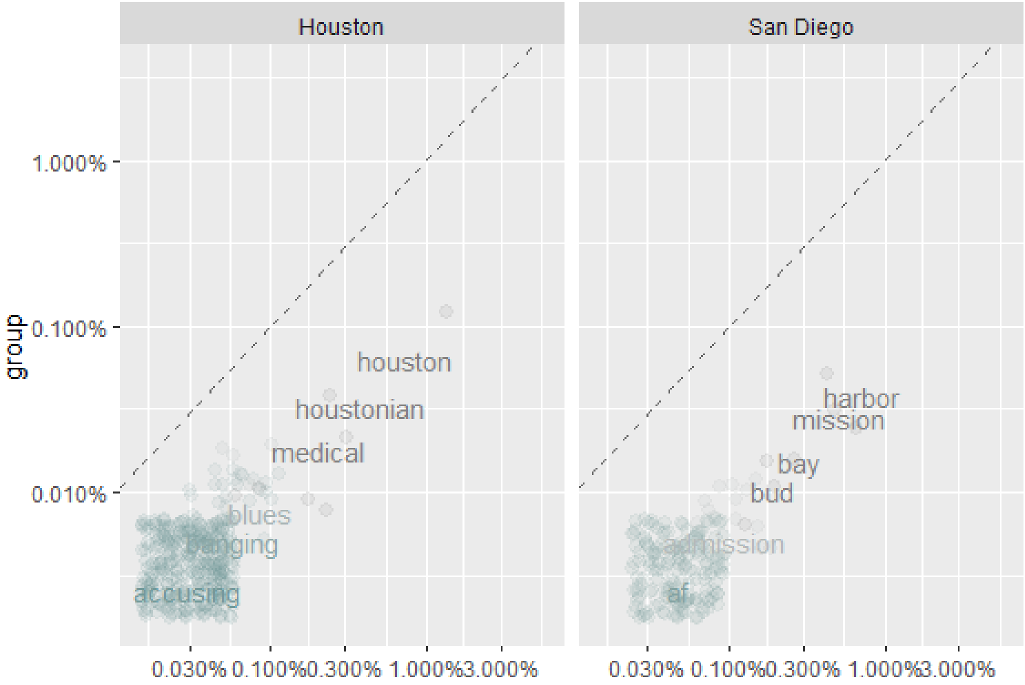
Note: This chart illustrates the sentiment scores in reviews for eight hotels in New York City. The black line is the average sentiment score for all reviews in New York City (1.78). It holds a similar pattern to Figure 17.

For most hotels, most reviews are above average, with several reviews being very high above average. On the other hand, the negative sentiment reviews can tell a hotel how to improve. The Hilton Santo Domingo has had only a few rare mildly negative reviews, meaning they should focus on rectifying minor errors when they happen to leave customers happy. In contrast, the Marriot Marquis has many mildly and strongly negative reviews meaning there are systemic issues that management needs to address. There is excellent value in both qualitative reviews (so new customers have ideas of what to expect more thoroughly) and quantitative reviews (so a hotel is boosted to the top-scoring hotels in a region).

Eleventh, one can generate a plot to compare the frequency of words in one city versus the frequency in the top 10 cities. In Figure 19 below, words further to the right are used more frequently just for that city versus the total corpus of the top ten cities.

Figure 19

Word Frequency Plot Comparison – Houston versus San Diego



Note: This chart illustrates the frequency of a word appearing just in the reviews for a particular city (Houston and San Diego) versus the overall reviews for the top 10 cities. Words appearing farther to the right are more unique to the cities shown.

On the left, "houston", "houstonian", "medical", and "blues" stand out. The word "medical" may indicate that hotels having close access to the many medical centers in Houston is a primary draw for many customers. Hotels can accommodate their customers in this way by ensuring they have maps or resources regarding the medical centers in their immediate vicinity.

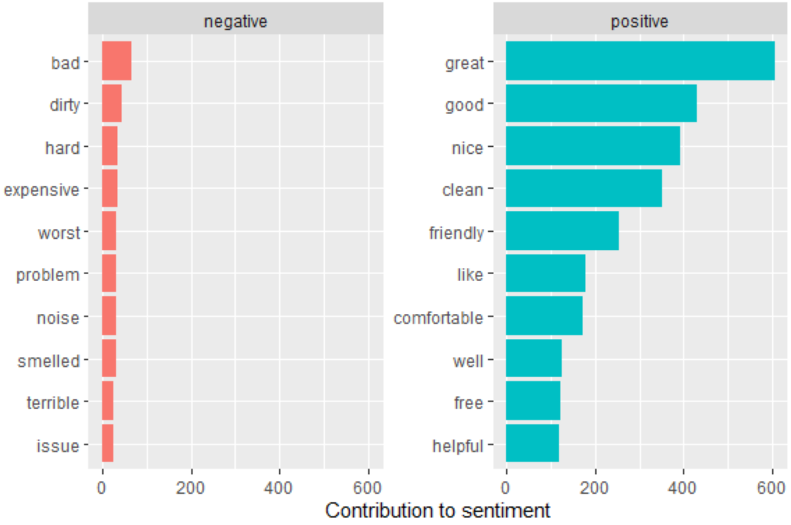
On the other hand, San Diego's frequent words include "harbor," "bay," "mission," and "bud." The first two words refer to the beautiful scenery and local beaches. Many customers may also have business at the San Diego Bay since significant shipping happens through that port. San Diego also has many historical "missions" that are of tourist value. Hotels usually have plenty of maps and brochures regarding these sorts of things; however, in this digital age, businesses must also market that they are near some of these areas of interest.

Twelfth, one can generate a plot to illustrate how different words contribute to negative and positive sentiment. One can readily deduce from Figure 20 on the next page that negative words have a more negligible effect on negative sentiment than positive words have on positive sentiment. One can also derive value from this plot by understanding trigger words that lead to negative reviews. For example, a hotel could use the word "smelled" to guide a management decision to install an air circulation system that infuses pleasant-smelling fragrance into the air, much like the casino industry. A hotel could use the word "clean" to continue emphasizing good housekeeping practices to staff members.

Thirteenth, one can generate a word cloud to depict the top 100 most frequently used words in the hotel reviews of the top ten cities. Figure 21 on the next page illustrates those words like staff, clean, breakfast, and nice appear more frequently, while words like town, wifi, and price appear less frequently. A hotel could use the word cloud to understand what hotel amenities and service characteristics guests mention in reviews. For example, a hotel could use the word staff to guide hiring practices and training literature for front desk receptionists, housekeepers, restaurant servers, and chefs.

Figure 20

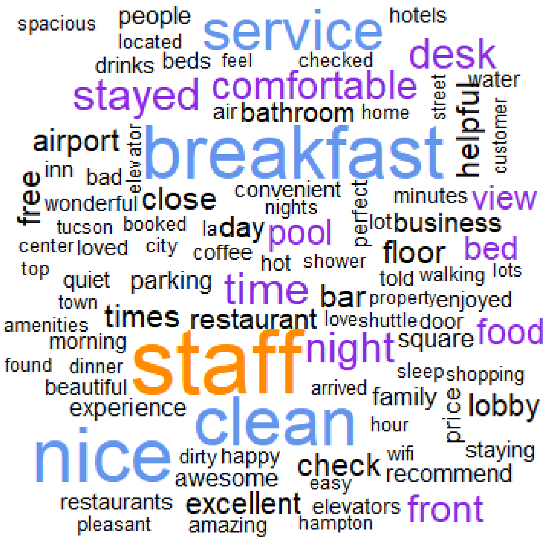
Word Contribution to Sentiment – Negative and Positive



Note: This plot depicts words that contribute to positive and negative sentiment. A hotel could use this information to guide hotel management policies and procedures.

Figure 21

Word Cloud for Hotel Reviews from Top Ten Cities

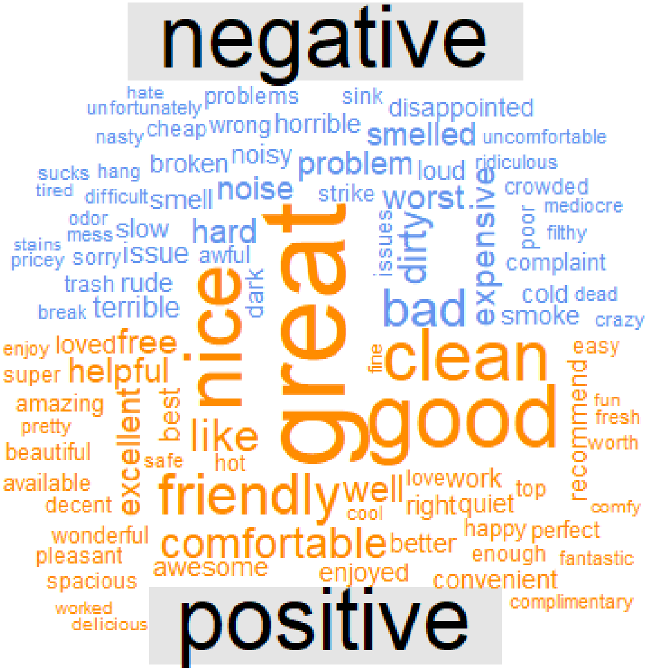


Note: This word cloud depicts some of the words in the hotel reviews of the top ten cities. A hotel could use this information to guide hiring practices, market amenities, improve breakfast offerings, justify pool upgrades, and make other decisions impacting hotel operations.

Fourteenth, one can generate a sentiment word cloud to depict some of the words that contribute to positive and negative sentiment. As depicted in Figure 23 below, this word cloud would benefit hotels that need to address negative reviews quickly. Suppose a hotel could generate this word cloud monthly. In that case, the hotel could use this information to spot negative trends quickly and address them by responding to the negative review online. Quick response time is crucial, as other potential hotel guests may see negative reviews before the hotel response and then decide not to book a stay at the hotel. A sentiment word cloud is a powerful tool that the hotel could leverage to improve hotel bookings and profitability.

Figure 23

Negative and Positive Sentiment Word Cloud

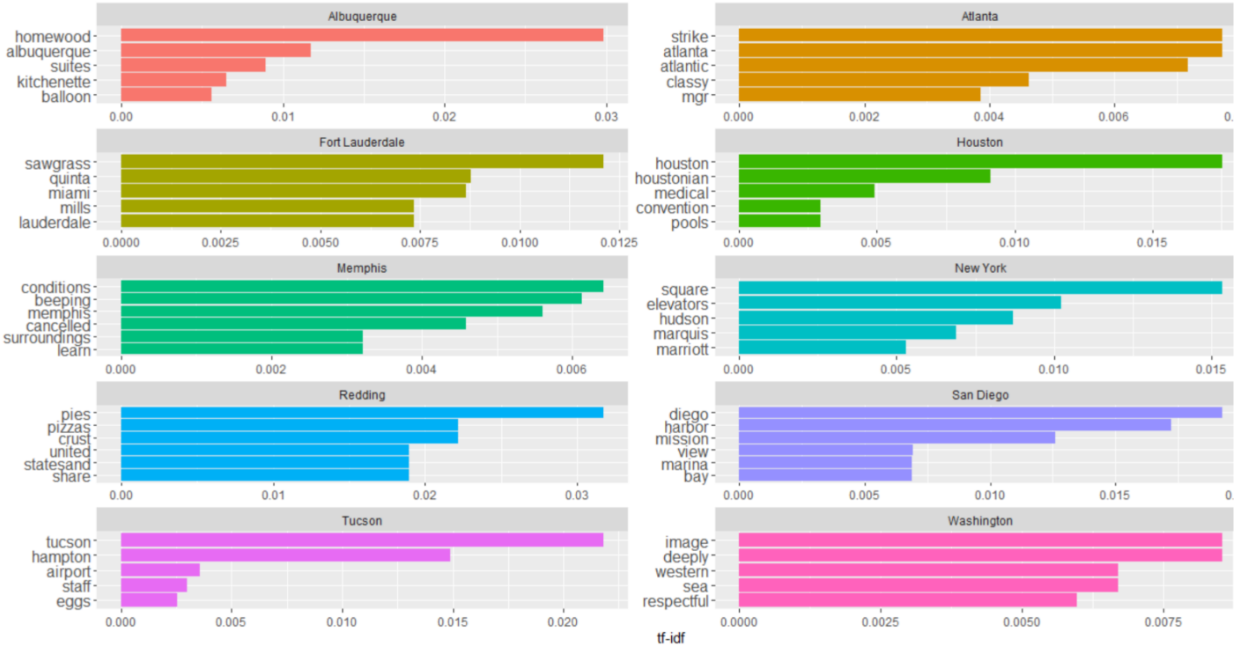


Note: This sentiment word cloud depicts some of the words that contribute to positive and negative sentiment. This word cloud would be beneficial to hotels that need to address negative reviews quickly.

Fifteenth, one can generate a plot showing the top 5 words with the highest TF-IDF scores for each top 10 city. Figure 24 below reveals that many of the words are simply rehashing the hotel names (Sawgrass, Homewood, Marriott, Maquis, etc.) or unique city name variations (Diego, Tucson, Memphis, etc.). However, the items of value are the other words. "Balloon" for Albuquerque likely is referring to the Albuquerque balloon festival. New York highlights "Square" and "Elevators"; the former likely refers to be close by to some of the famous squares. The latter could be positive or negative; elevators are necessary for a very "tall" city like New York. Speedy, reliable elevators could make or break a customer's experience.

Figure 24

TF-IDF Comparison of the Top 5 Words for Each of the Top 10 Cities

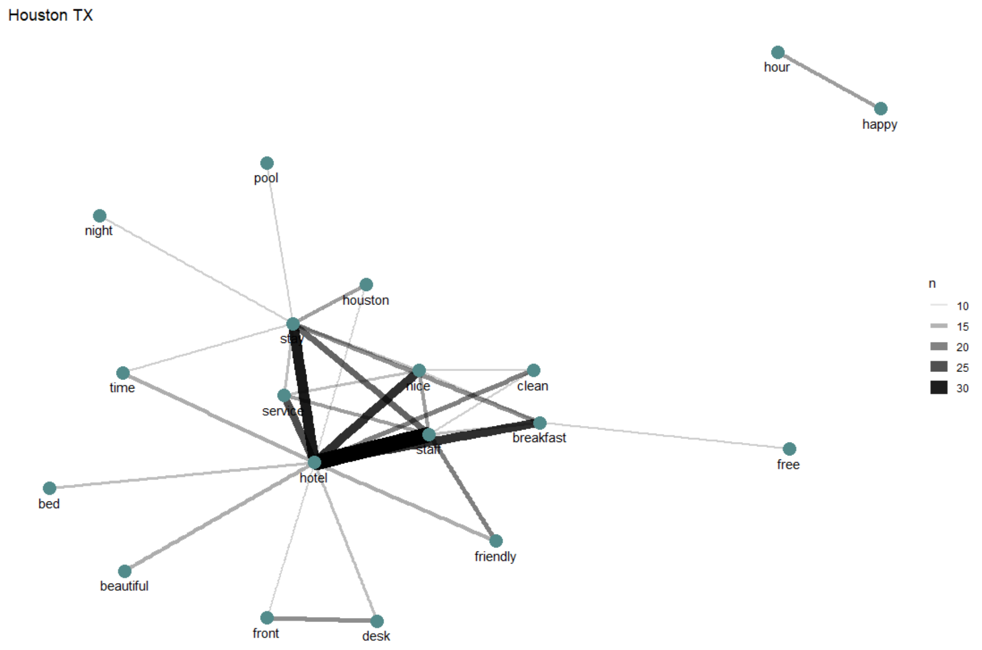


Note: This chart illustrates the top 5 most potent words as determined by the TF-IDF score for the top 10 cities. Many unique words are specific hotel names (Sawgrass, Homewood, Marriott, Maquis, etc.) or unique city name variations (Diego, Tucson, Memphis, etc.).

Lastly, one can generate a spider diagram displaying the most frequently co-occurring words. This diagram provides more context to a single word. Figure 25 on the next page has many common co-occurring words; there is a prominent rectangle at the center between "hotel," "staff," "nice," and "service." However, unexpectedly, "happy" and "hour" occur on the margin unconnected to any other words shown. Also, "free" and "breakfast" indicate an enticing offer that most customers look for in any hotel where they would stay. Finally, "beautiful," "clean," and "bed" all co-occur with "hotel," indicating that customers strongly care about the cleanliness and aesthetics of a hotel.

Figure 25

Spider Diagram Showing Most Frequently Co-Occurring Words



Note: This chart illustrates the strongest co-occurrences between words in reviews about Houston hotels. The thicker lines indicate stronger co-occurrences. One expects many of the words to occur together. "Free" and "breakfast," as well as "happy" and "hour" are essential words to which one should pay attention.

Conclusions

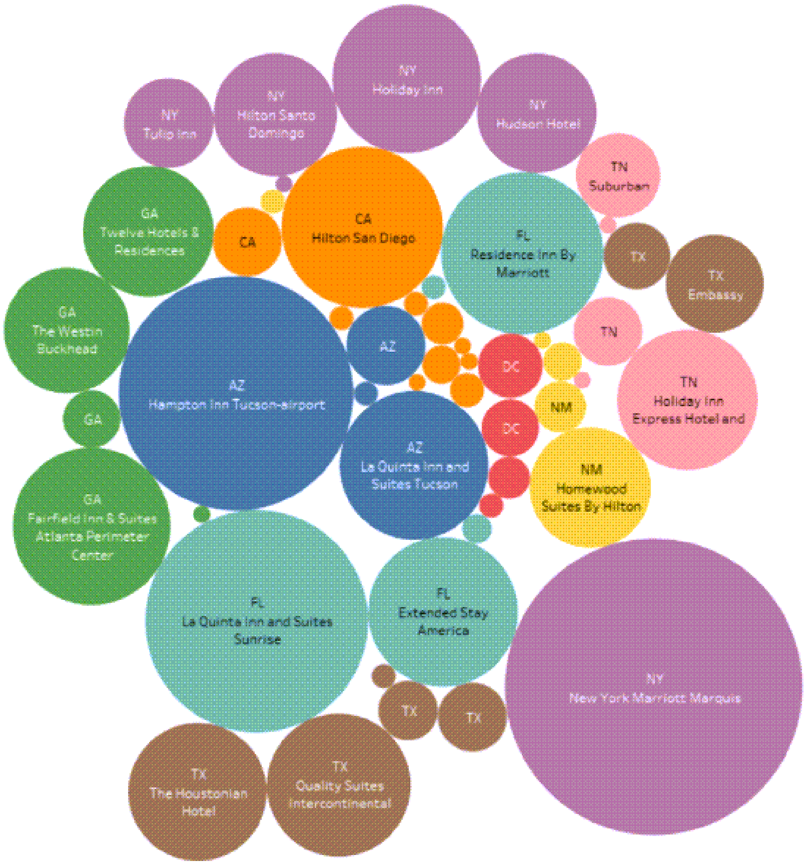
In conclusion, this project sought to ascertain the influence of online ratings and reviews on hotel booking considerations. The study established how online reviews about (a) physical attributes, (b) food and drinks, (c) staff service, (d) hotel location, and (e) number of online reviews affect hotel booking considerations. Our team analyzed the hotel reviews dataset to establish the influence of online ratings and reviews on hotel booking considerations. The dataset was analyzed using Exploratory Data Analysis, conducted in two phases. The first phase entailed understanding the data structure. In contrast, the second phase involved understanding the data distributions of the columns to decide how to address the conversion to a machine learning rectangular dataset compatible with the selected algorithms.

Our team performed data cleaning to remove variables from the dataset that added no value to the analysis, such as hotel address or country information. Data cleaning entailed removing all identifying information that revealed the reviewer's identities, including identifying variables, such as usernames and reviews.id. Our team normalized the rating and established an official state feature based on latitude and longitude. Afterward, we also added columns to indicate if a hotel has specific amenities.

After completing the initial exploratory data analysis and processing activities, our team discussed the text-mining algorithm and model methods. Guided by the research objective, which was to ascertain the influence of online ratings and reviews on hotel bookings and identify areas where a hotel could improve amenities and services, the team explored the reviews further using text mining methodology and algorithms. See Figure 26 on the next page for a visual representation of the number of hotel reviews by hotel and state that were part of the original dataset considered for the analysis.

Figure 26

Number of Hotel Reviews by Hotel and State for Top Ten Cities



Note: The size of each bubble represents the number of reviews for each hotel for the top ten cities, with the largest bubble being the hotel with the highest number of reviews and the smallest bubble being the hotel with the least number of reviews.

The study's findings revealed high positive scores of hotel attributes that included service, rooms, location, food and drinks, and the number of online reviews in that order. The highest positive score among the four attributes (service) supports the notion that customers put a high value on customer service. If they are satisfied, they will leave a positive comment concerning the overall hotel business dynamics, the quality of the room, food, drink, and location are critical aspects that can influence positive reviews.

While exploring the TF-IDF scores, which entailed probing the number of times a word appeared in the data set for the top 10 cities, the word that appeared the most in the reviews was "stay," while the word that appeared the least in the review was "restaurant." In this chart, the top ten words that appeared the most in the reviews included stay, staff, nice, breakfast, clean, service, friendly, location, time, and stayed night. These top ten words resonate with the vital hotel attributes explored in this study: physical attributes, food and drinks, staff service, and hotel location. Therefore, hotel managers should improve on these attributes to attract new clients and increase their customer base.

One also noted that local attractions play a part in the hotel's positive reviews. For instance, the medical center around Houston or the missions and the bay of San Diego influenced positive reviews among hotels in the vicinity. Thus, a hotel's physical location significantly impacts the reviews it is likely to receive from its customers. The positive reviews regarding the hotel's location increase in the event major attractions such as amusement parks, natural wonders, metropolitan areas, or museums are nearby. The findings suggest that in the absence of local attractions nearby, the management should enhance other attributes such as customer service and other hotel amenities like room appearance.

However, although the study findings reveal invaluable information regarding the impact of online reviews on hotel booking considerations, the study had a few limitations requiring improvement from future researchers. There is a need to perform analysis using amenities features to enhance the reliability of the data. Amenities entail residential or commercial property characteristics that are regarded as beneficial by potential users or tenants. Perming analysis using amenities feature would increase the depth and scope of data analysis, thus enhancing data triangulation, an essential aspect of study validity.

Moreover, some reviews contain foreign languages, which presents a language barrier among people who do not understand those languages. Given this, it would be advisable to translate the foreign languages to English and other commonly used international languages to allow a large pull of potential customers to understand the reviews and make informed decisions. One can translate the reviews into other languages, including French, Chinese, and Germany, among other international languages.

Finally, it is recommended for future researchers to pinpoint reviews from a specific hotel chain and examine it against the study variables. Different hotel chains face different problems and therefore need to formulate solutions unique to their situation. Exploring a specific hotel chain would lead to recommendations that are relevant to that particular hotel chain. It is also vital for future studies to combine TD-IDF with sentiment analysis to enhance data triangulation, thus improving the reliability of the findings. Also, future studies on the same topic should analyze customer characteristics and explore more current and longitudinal data stretching extended periods.

Appendix A

Table 1

Dataset Attributes, Data Type, Category, and Notes

#	Attribute	Data Type	Category	Notes
1	address	factor	hotel	999 levels
2	categories	factor	hotel	396 levels
3	city	factor	hotel	761 levels
4	country	factor	hotel	Constant only U.S. for the entire file
5	latitude	num	hotel	Paired with longitude, 982 distinct values. 86 missing values.
6	longitude	num	hotel	Paired with latitude, 983 distinct values. 86 missing values.
7	name	factor	hotel	879 levels
8	postalCode	factor	hotel	912 levels. 55 missing values.
9	province	factor	hotel	287 levels – there are only 50 states
10	reviews.date	int	review	3010 levels. 259 missing values.
11	reviews.dateAdded	factor	review	1029 levels
12	reviews.doRecommend	logi	review	All Nulls
13	reviews.id	logi	review	All Nulls
14	reviews.rating	num	review	292 levels. 43 distinct values. 862 missing values.
15	reviews.text	factor	review	The main text for analysis. 22 missing.
16	reviews.title	factor	review	4 levels. 1,622 missing rows.
17	reviews.userCity	factor	review	2898 levels. 19,649 missing rows.
18	reviews.username	factor	review	15493 levels. 43 missing rows.
19	reviews.userProvince	factor	review	649 levels. 18,394 missing rows.

References

Consumer Trust in Online, Social, and Mobile Advertising Grows (2012, April 11). *Nielsen.com*. Retrieved from Newswire | Consumer Trust in Online, Social and Mobile Advertising Grows | Nielsen – Nielsen

Filieri, R. & McLeay, F. (2013, March 25). *E-WOM and Accommodation: An Analysis of the Factors That Influence Travelers' Adoption of Information from Online Reviews*. Retrieved from E-WOM and Accommodation: An Analysis of the Factors That Influence Travelers' Adoption of Information from Online Reviews - Raffaele Filieri, Fraser McLeay, 2014 (sagepub.com)

Kaggle. (2021). *Hotel Reviews*. Retrieved from https://wwww.kaggle.com/datafiniti/hotel-reviews

Lardieri, A. (2021). *Coronavirus Pandemic Sets Hotel Industry Back 10 Years, Report Finds*. Retrieved from https://www.usnews.com/news/national-news/articles/2021-01-27/coronavirus-pandemic-sets-hotel-industry-back-10-years-report-finds

Online Consumer-Generated Reviews Have a Significant Impact on Offline Purchase Behavior. (2007, November 29). *ComScore*. Retrieved from https://www.comscore.com/Insights/Press-Releases/2007/11/Online-Consumer-Reviews-Impact-Offline-Purchasing-Behavior

Podnar, K., & P. Javernik. 2012. *The Effect of Word of Mouth on Consumers' Attitudes Toward Products and their Purchase Probability*. Journal of Promotion Management 18 (2): 145–168.

Relationships Between Words: n-grams and Correlations. (n.d.) *TidyTextMining.com.* Retrieved from https://www.tidytextmining.com/ngrams.html

Scott, W. (2019). *TF-IDF from Scratch in Python on Real-World Dataset. Towards Data Science.* Retrieved from: https://towardsdatascience.com/tf-idf-for-document-ranking-from-scratch-in-python-on-real-world-dataset-796d339a4089

Sentiment Analysis with Tidy Data (n.d.). *TidyTextMining.com.* Retrieved from https://www.tidytextmining.com/sentiment.html

Singh, J. & Wang, B. (2021). *Impact of COVID-19 on the Hospitality Industry and Implication for Operations and Asset Management.* Retrieved from https://www.bu.edu/bhr/2021/05/31/impact-of-covid-19-on-the-hospitality-industry-and-implication-for-operations-and-asset-management/

The Hotel Review Sites You Should Monitor (2020, August 19). *ReviewTrackers.com.* Retrieved from https://www.reviewtrackers.com/blog/hotel-review-sites/

The Tidy Text Format (n.d.). *TidyTextMining.com*. Retrieved from https://www.tidytextmining.com/tidytext.html

Yang, J. & Mai, E. (2010). *Experiential Goods with Network Externalities Effects: An Empirical Study of Online Rating System*. Volume 63, Issues 9–10, 2010, Pages 1050-1057, ISSN 0148-2963. Retrieved from https://doi.org/10.1016/j.jbusres.2009.04.029