Analysis of Airbnb Listings During the Coronavirus Pandemic

Prepared by:
Dazzelle Bagtas
Kylie Wise
Logan Hylton
Braeden Obar

May 6, 2021

Contents

List of tables

List of figures

Executive summary

This project investigates what characteristics are important in producing a high reservation rate for an Airbnb listing. This information is important considering that the Covid-19 pandemic has hurt the hospitality industry, and determining which factors are important in helping listings get reserved may provide guidance as to how Airbnb can assist hosts to continue receiving reservations and sustain their business. The data used for this analysis are listing information for Airbnbs in New York City from January 2020 to March 2021, which was obtained from Insideairbnb.com. The data includes information about listings such as their physical characteristics, corresponding host information, prices, and the days which they were reserved within a year. Through descriptive analysis, it was revealed that the number of reservations booked within a month increased during 2020. Additionally, the number of bookings within the first three months of 2021 was less than the first three months of 2020. Through predictive analysis, specifically random forest modeling, it was revealed that the most important variables that play a factor in customers' decision to reserve a listing are the total number of listings a host has, the number of people it accommodates, the minimum nights required to stay, whether the host is a superhost, and price. Our recommendations for Airbnb to help their hosts are to decrease host costs and provide real estate assistance.

1 Introduction

To say the Coronavirus pandemic has altered the way the world works would be an understatement. Covid-19 has fundamentally changed the way many industries function, thrive, and survive. When the Coronavirus outbreak hit the United States, many businesses struggled as lockdowns and travel restrictions greatly diminished the number of consumers buying goods and services.

The lodging industry has been especially affected by the pandemic. For Airbnb specifically, about 64% of guests had canceled or planned to cancel their Airbnb reservations three months after the Coronavirus entered the United States (Lane 2020). In response, Airbnb has implemented new policies and practices such as new cleaning protocols, limited guest capacities, and masking and social distancing guidelines (Airbnb 2020). A question remains, however, if these new policies are enough to sustain the business after the losses it has experienced. According to an article published by McKinsey & Company, the accommodation and food industry, which includes the lodging industry, may not fully recover until 2025 (Dua et al. 2020). Assessing more specifically how the pandemic has affected the lodging industry and specific businesses within it will be important to show how well they have adjusted and allow business leaders to better understand what more must be done to sustain their businesses and gain insight on what the future of the lodging industry might look like.

1.1 Business problem

In order to examine an aspect of how the pandemic has affected the lodging industry, the following question will be explored: What are the attributes of listings that have closed or are not receiving reservations during the pandemic? Factors such as price and location that may possibly affect the likelihood of these listings not getting reserved or having to close altogether will be observed. By determining these specific factors, Airbnb can target problem areas and design solutions to combat them. This is important because when listings are not booked, revenue is lost. The problem will be examined within the context of New York City because it saw one of the highest totals of Covid-19 cases in the United States and, thus, will ensure that any effects found are likely attributable to the presence of Covid-19 (USA Facts 2021).

1.2 Intended audience

The main entity that would benefit from the solution for the business problem is Airbnb itself. By determining certain attributes and factors that may be contributing to the low performance of certain listings, Airbnb may allocate more time and resources to help these areas and hosts. This could look like providing more resources to help hosts abide by safety protocols, guidance in attracting more guests, or cutting their operating costs in the meantime. Doing so may incur additional costs for Airbnb, but it also has the potential to increase Airbnb's overall revenue and its recovery from the initial shock caused by the pandemic. Airbnb hosts will also benefit from this solution. Receiving help from Airbnb in maintaining and advertising their listings may help increase the number of reservations made and their personal revenue.

Additionally, the solution may help guide hosts' decision on whether to continue business with Airbnb at all if they determine that they would not be able to make up for losses and continued operating costs even with Airbnb's assistance. Furthermore, other businesses within the lodging industry could benefit from this analysis. For example, Tripping.com and Vrbo, which are lodging companies similar to Airbnb, may take note of what kinds of listings continue to struggle in booking guests and design their own solutions to help those hosts.

2 Data

After cleaning and preparing the data from the data set that contained listing information for each Airbnb listing in New York City, the data was narrowed down to 21 columns that contain information about each listing such as their unique ID number, physical listing characteristics, characteristics about the hosts, location, and amenities.

The second data set used for analysis contains columns with information of Listing IDs, Dates, Availability, Prices, and Maximum and Minimum Nights. This data set shows whether a listing is booked or available for each date of the year and for what price.

2.1 Data collection

Our data sets come from Insideairbnb.com, which verifies, cleanses, analyzes, and aggregates publicly available information about various cities' Airbnb's listings directly from the Airbnb website. Inside Airbnb describes themselves as an independent, non-commercial set of tools and data that allows people to explore how Airbnb is being used in cities around the world

(Inside Airbnb 2020). They are not associated with or endorsed by Airbnb, but all their information comes directly from the Airbnb website, which helps support that the data are accurate and consistent with the true listings. Additionally, the data is updated monthly, which ensures that the data is up to date. While some hosts do not always keep their listings calendar updated, after cross-checking a few dates from our data with current listing on the Airbnb website, the data appears to remain consistent. All these factors ensure that the data is of high quality.

Our data sets provide detailed information about each listing and provide both qualitative and quantitative information that contains insight about the listings ranging from physical characteristics, price, details about the host, and days reserved out of the year. These characteristics make it appropriate and useful to answer the question of determining what the common characteristics are among listings that are struggling to get reservations or have closed down altogether.

2.2 Data preparation

Inside Airbnb has already prepared and cleaned much of the data; and consequently, we found that there was not as much for us to do in terms of cleaning. The first step of preparing the data consists of removing categories that are not useful for analysis within the "listings" datasets. The original data set has 104 variables and is reduced to 21 variables after removing the unhelpful columns. Many columns that are removed are ones that included links to profile pictures, personal name information, and redundant, less helpful location identifiers such as latitude and longitude points, since these do not provide any useful information. Next, the data is then checked to ensure there are no duplicate listings. After doing so, value types are then converted into their appropriate form such as numeric, currency, or percentage. Additionally, certain column names will be renamed in order to increase the clarity of the data. Cells with blank or missing values are not removed because they may give insight as to what helps boost a listing's number of reservations among customers. We additionally created a response variable column which indicated whether a listing had a high, moderate, or low reservation rate. A listing had a high reservation rate if it was reserved for 75% or more of the year, moderate it if was reserved between 50%-74% of the year, and low otherwise.These data preparation steps ensure that our data is even more useful in determining what the common characteristics are among listings that are struggling to get reservations or have closed out altogether and to further confirm that the data is of high quality.

3 Descriptive analytics

The following descriptive analytics visualizations provide insights that will be helpful in determining common characteristics among Airbnb listings in New York City that have closed or are not receiving reservations during the Covid-19 pandemic. Exploring this data will give us insights into patterns and trends within our data, which will inform how to best approach our business question through statistical analysis and assess how the pandemic has altered the Airbnb New York City landscape. In these visualizations, we will explore different metrics from listing price and listing price variation to booking counts and year over year trends.

The following graph shows the difference in the number of nights a listing was booked or reserved between 2020 and 2021 for the first three months of each year.

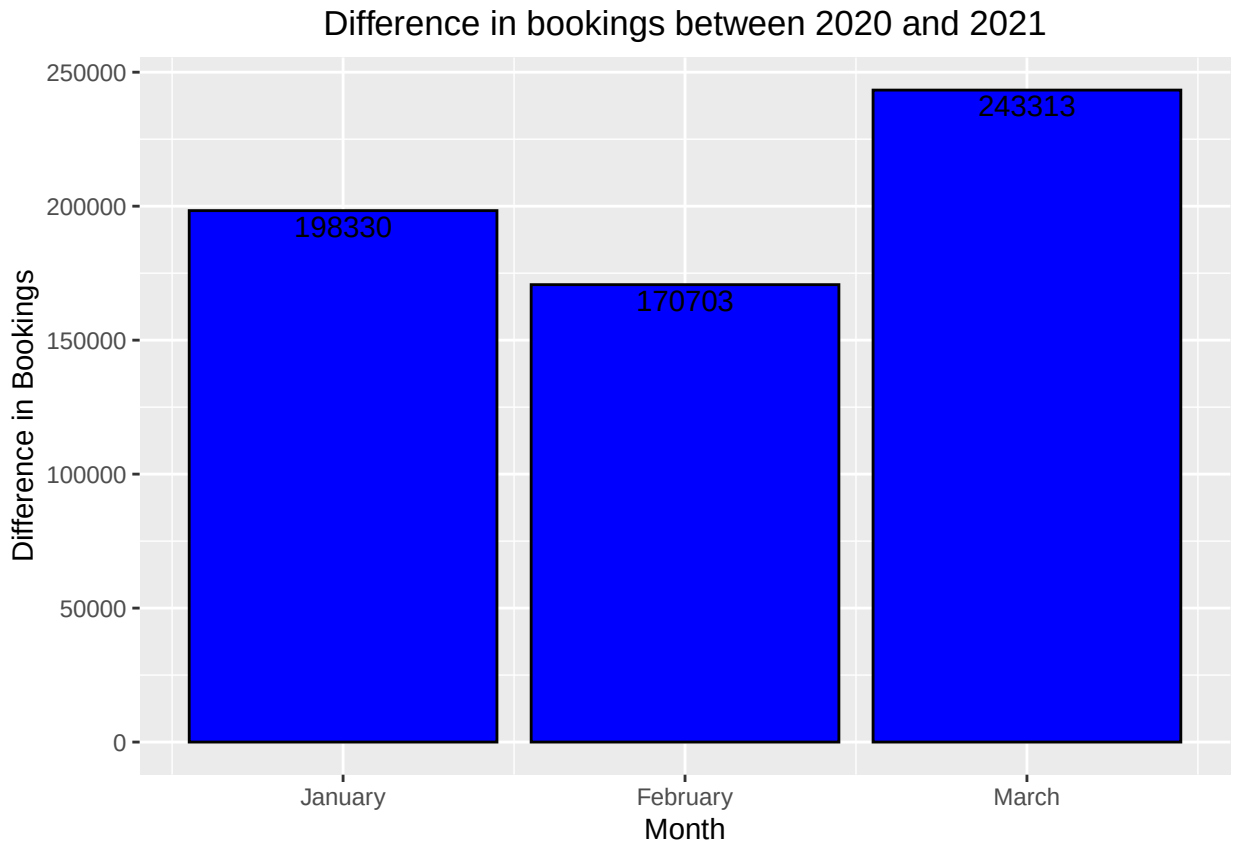


Figure 1: Difference in Number of Bookings from Jan-March, 2020 vs. 2021

The graph reveals that the greatest difference in number of bookings is in the month of March. This means that March 2021 had 243,313 less bookings that March 2020.

The following plot shows the number of nights that a listing was booked for each month in 2020.

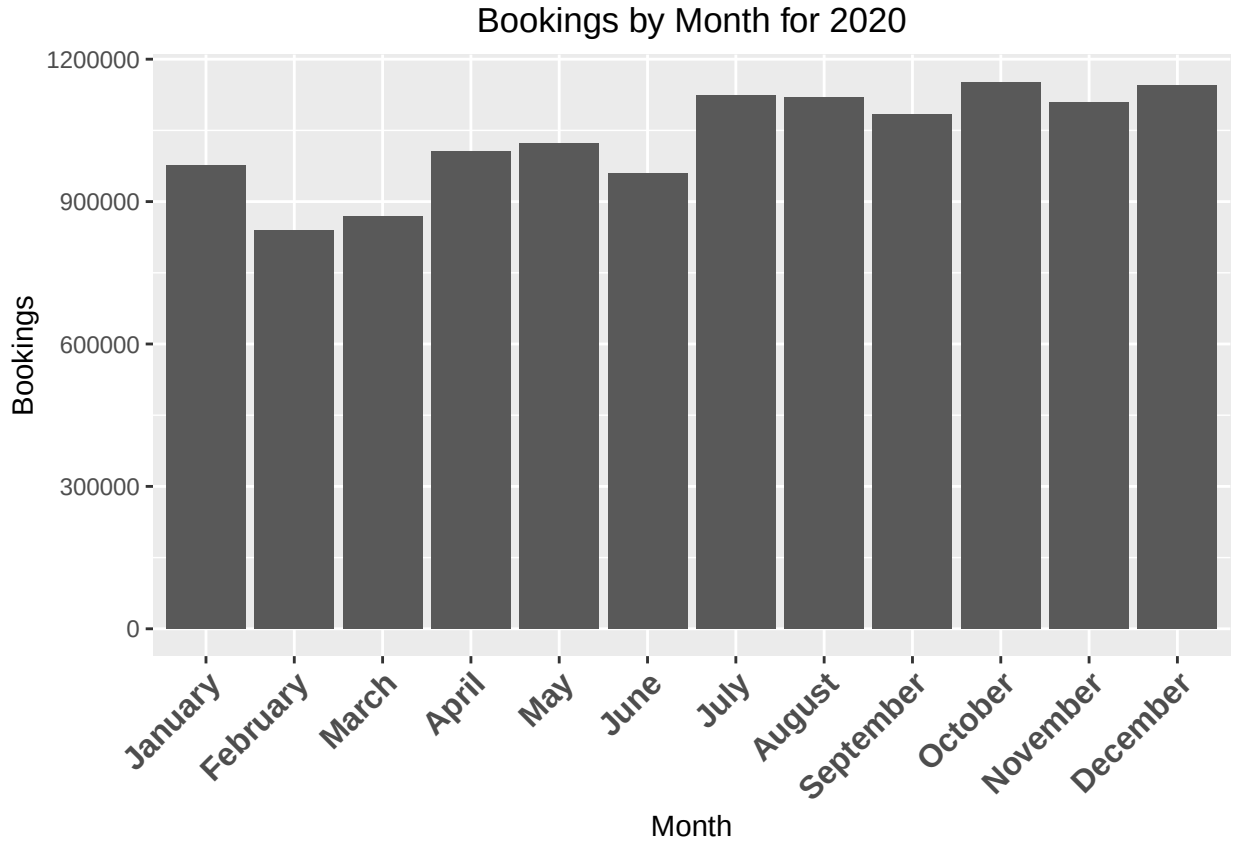


Figure 2: Bookings Per Month, 2020

The plot shows an increasing trend in the number of bookings for each month in the year 2020.

The following bar graphs show the top 5 most frequently mentioned adjectives used in the summaries for the 2020 and 2021 listings datasets with their appropriate counts of times mentioned.

We can see similar trends for the most used adjectives in both 2020 and 2021, with top words in both being spacious, available, comfortable. In 2020, there were more mentions of beauty and renovation whereas in 2021 there were more mentions of being equipped with certain amenities or features and being furnished.

The following boxplots show the distribution of listing prices by neighborhood in 2020.

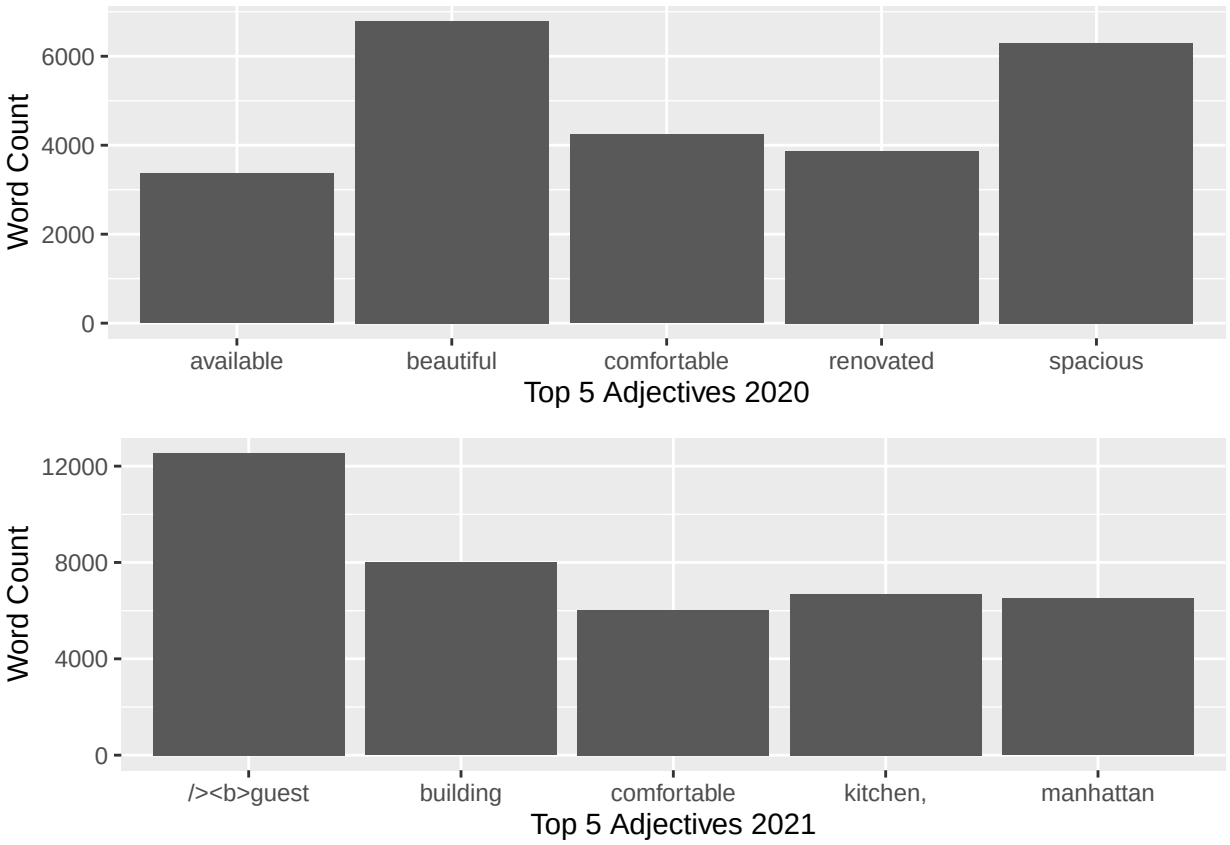


Figure 3: Top 5 Used Words In Descriptions, 2020 vs. 2021

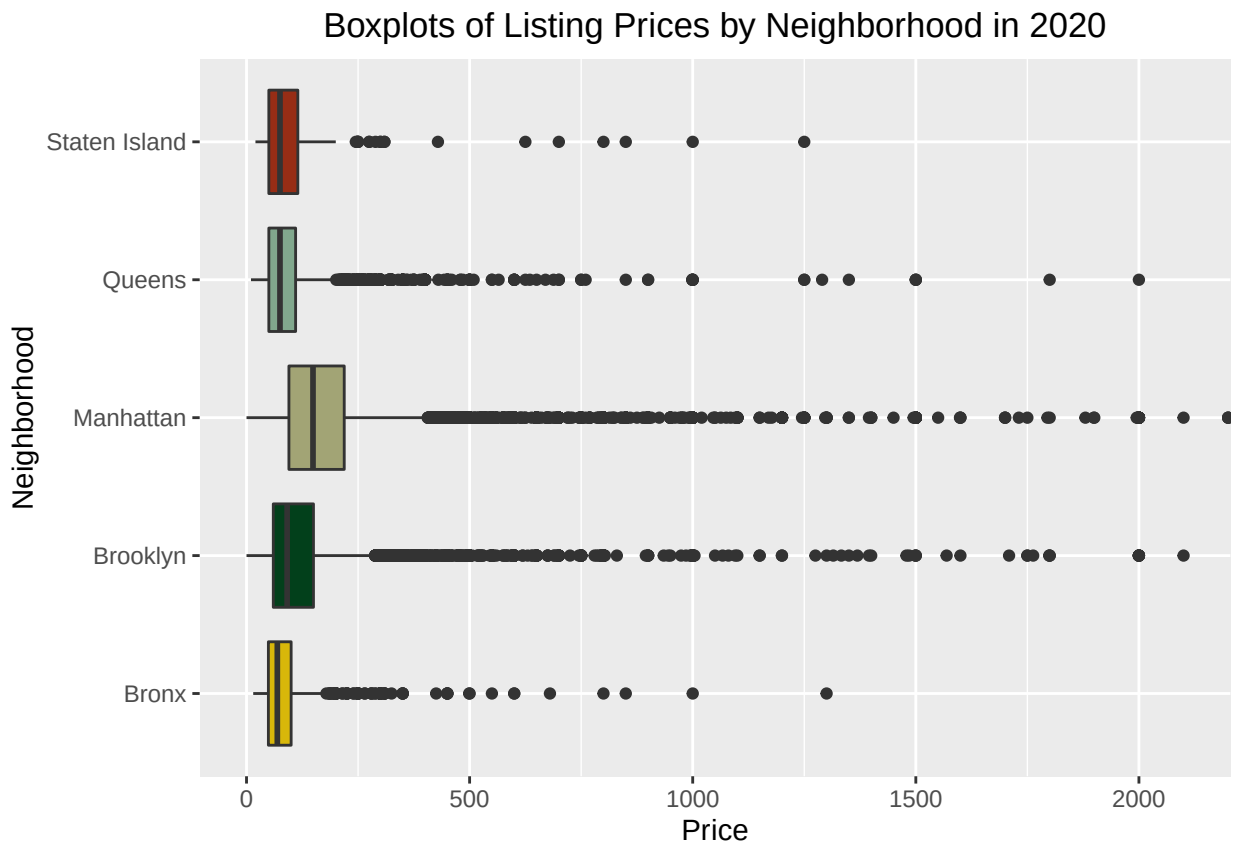


Figure 4: Boxplots of Listing Prices by Neighborhood, 2020

The boxplots reveal that most listings fall somewhere within the $60 to $250 range but there are many outliers. Manhattan had the highest average listing price while Staten Island, Queens, and the Bronx are closely tied for having the lowest average listing price.

3.1 Insight summary

From the chart displaying difference in bookings between 2020 and 2021, it is revealed that the first three months of 2021 had fewer overall bookings for New York City Airbnb's than the first three months of 2020. This is reasonable as cases in the first few months of 2021 spiked in New York City, thus it is to be expected that less people felt comfortable renting in New York during this time.

In the next visualization of the month over month for the entire year of 2020, the number of total bookings gradually increased. This is surprising because one would predict that as the pandemic progressed and worsened, the number of total bookings would decrease, however the trend displays the opposite. This implies that there may be factors concerning the listings that have encouraged customers to reserve Airbnbs.

As for pricing, from the boxplots, we see more clearly how the distribution and variation of Airbnb listing price differs for each neighborhood. Manhattan has the highest average listing prices while the lowest average prices are around the same range in the Bronx, Queens, and Staten Island. This understanding may be particularly interesting as we continue into our predictive analyses because we may want to split our data by neighborhood to see if some neighborhoods have different trends than others.

Additionally, we investigated how the listing descriptions changed between 2020 and 2021. It is revealed that in 2020, there were more mentions of beauty and renovation whereas in 2021 there were more mentions of being equipped with certain amenities or features and being furnished. This shift in wording may reflect a change in customers' values concerning listings; therefore, it may be important to evaluate its relationship to the number of bookings each listing receives.

4 Predictive analytics

For our predictive analytics section, we chose to use random forest modeling and text analysis. Through random forest modeling, our goal is to determine which various listing variables are important in determining the number of days a listing is reserved in a year in New York

City. Our original data set contained over 20 different variables and descriptors that could be factors in booking an Airbnb. The resulting random forest model will tell us which quantitative and qualitative variables are most important/significant when reserving a property. We also performed a bag-of-words text analysis. Each listing in our data set had a paragraph description about the property. By creating a word analysis, we can find which words or phrases are most common. These results would signal which listing features are most important and effective when attracting guests to stay at a specific listing.

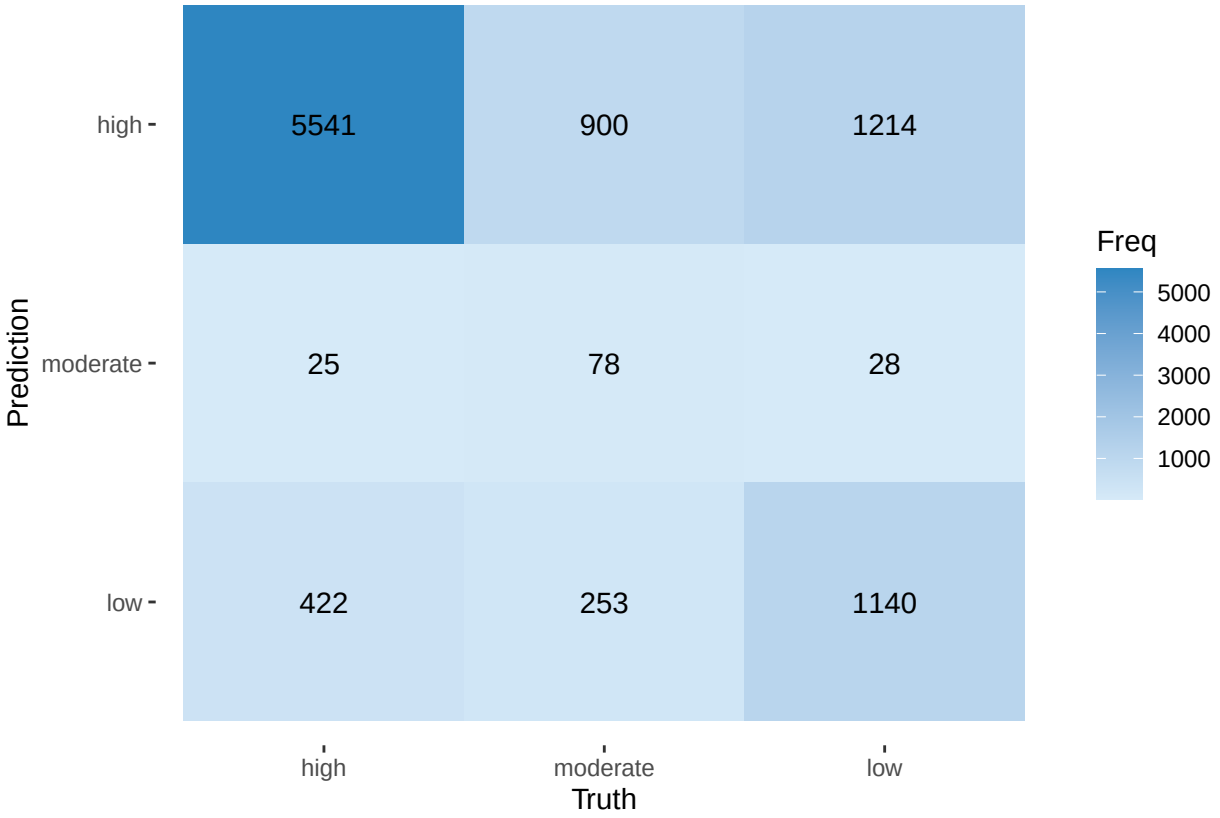
4.1 Process

For our random forest model, we first divided our Listings data set into three smaller data sets: one to construct a model to perform the analysis, one to estimate how well our model was constructed, and another to apply our model to data and validate the quality of our model. We then created the initial decision tree which included all our variables. After that, we iterated the model by removing variables that did not appear as important in prediction by considering how much the model accuracy decreased if they were removed.

As for the text analysis portion of our predictive analysis, we used bag-of-words unigram and bigram modeling. Bag-of-words modeling is concerned with occurrences of words or combination of words in a particular dataset. For our analysis, we were interested in looking at the description of Airbnb listings that were successfully rented in order to compare the most used words in 2020 listings descriptions versus 2021 listings descriptions. We first began by splitting up each word for each description and creating a list of these words. After compiling this list, we ensured that it was simplified, meaning it only contained the words we were interested in. To do this, we removed stop words, which are commonly used words such as "the," "I," "which," etc., punctuation, and symbols from this list. We also removed words that had a low weight according to term frequency and inverse document frequency. This removed words that were common, but common across all descriptions, so that they did not give us uniquely interesting information about each year. We then took this simplified list and did a count of each word, to see which words were used the most in descriptions for each year. We followed a similar process for word pairs in order to create a bigram model that counts pairs of words that occur consecutively. We then put each of these side by side in a table to more easily see if and where the differences occurred.

4.2 Assessments

The two important assessments to consider for random forest analysis are the confusion matrix and the prediction accuracy of the model used to create the decision tree. From our confusion matrix we can see that our model correctly predicted 5,635 listings that were categorized as high, 106 listings as moderate, and 1,094 listings as low. It incorrectly predicted 2,766 listings; however, there were not an overwhelming amount of true low values that were were incorrectly predicted as high or true high values incorrectly predicted as low, which assures us that the model we have chosen predicts listings' reservation rates well.



The second important assessment of our decision tree model is the accuracy of our model. Our model has an accuracy of 71.2% which is fairly high and helps validate that our decision tree model is sufficient to use. (Shish 2016)

	Accuracy
Final Random Forest Model	0.712

Random forest models make no assumptions about the distribution of the data. Thus, it is appropriate for our data to be used in such model and we did to investigate the data any farther. There also are no assumptions required to complete a successful bag-of-words modeling.

For the bag-of-words modeling, we assessed which stop words to exclude from our model. We first used common English stop words; however, it was also important to add additional stop words particular to our analysis. In order to do so we went in and removed words that had a low weight according to term frequency and inverse document frequency. This removed words that were common across all descriptions, as they did not give us uniquely interesting information about each year. For example, we decided to remove words such as bedroom, apartment, apt, etc. because we saw that these words appeared in most of the descriptions, thus they do not provide unique information about differences that exist from description to description. Unfortunately, even after removing stop words, we found many of the same words appeared in both the 2020 and 2021 datasets. This is not ideal because it makes it more difficult to answer our business question regarding how listings have changed from 2020 to 2021.

4.3 Results

As a result of creating, pruning, and removing variables from our random forest model, we found that the most important variables in determining whether a listing has a high, moderate, or low reservation rate in a given year is the total number of listings a host has, the number of people it accommodates, the minimum nights required to stay, whether the host is a superhost, and price.

Variable	Mean_Decrease_Accuracy
Host Total Listing Count	203.59
Accommodates	107.57
Minimum Nights	103.74
Host is Superhost	103.00
Price	97.96

The results from our bag-of-words text analysis are shown below. Some interesting results are that the top word in 2021 descriptions is space, which was mentioned more than 40,000 times whereas space is much less frequently mentioned in 2020. We can also see that words like

restaurants and subway are mentioned frequently in 2020, however not included in the top 10 most frequently mentioned words in 2021. We will explore these results more thoroughly in our insight summary.

Table 3: Top Description Words 2020 & 2021

2020 Words	Frequency	2021 Words	Frequency
kitchen	40924	space	42048
walk	28588	kitchen	28018
restaurants	26404	access	25562
manhattan	26275	bathroom	17550
bathroom	25251	guest	17482
park	24833	building	14663
train	24296	manhattan	14605
space	24267	floor	13828
subway	22434	walk	13622
neighborhood	21982	park	13054

4.4 Insight summary

From the random forest, we find that the total number of listings a host has, the number of people it accommodates, the minimum nights required to stay, whether the host is a superhost, and price are important variables in determining whether a listing has a high, moderate, or low reservation rate. From this list of variables, this reveals that customers may be looking for hosts who have expertise and experience in running and maintaining Airbnb rentals when considering which listing to rent since they are considering a host's total number of listings. This is reasonable, especially in the Covid pandemic era, where everyone is taking additional safety precautions.

As for our unigram model, we saw that "space" was mentioned the most frequently in 2021 rented listings. This gives us some more insight into how the needs and wants of renters has changed as a result of the pandemic. It is reasonable that after almost a year of socially distancing, as a result of the pandemic, people prefer more spacious environments. It is interesting to see Airbnb listings descriptions reflect this cultural change and cater to the new preferences of renters. Pre-pandemic, people felt comfortable going to restaurants and taking the subway; however, now many people view these activities as unsafe. Thus, it is

not surprising to see Airbnb listings no longer mentioning as frequently restaurants and the subway. Ultimately, this text analysis gave us a better idea of how listings have adjusted because of the pandemic to better suit the preferences of renters. It also may give hosts a better idea of how to better market their listing to the changing needs of the market.

5 Conclusions

Ultimately, through this analysis we were able to investigate which factors are important in resulting in high reservation rates for an Airbnb listing. We were able to find that the number of listings that a host has, the minimum number of nights for the booking, the number of guests it can accommodate, the minimum nights required to stay, and the price are all important factors. Given that the number of listings that a host has was the most important, this may signal that guests are looking for hosts with experience and expertise in running Airbnb listings as they are looking for safety and a reliable place to stay, specifically within the pandmeic. We also determined that certain favorable Airbnb description features have changed as a result of the pandemic. For example, we revealed an increase in listings using the word "space", which may indicate that hosts should include this within their descriptions to attract customers in the meantime. This information is valuable to Airbnb stakeholders and Airbnb itself in how they can change their business to focus more on marketing what the consumers value the most when selecting a listing. The Covid-19 pandemic has undoubtedly hurt the hospitality and lodging industry, and determining which factors are important in helping listings get reserved may provide guidance as to how Airbnb can assist hosts to continue receiving reservations and sustain their business.

6 Recommendations

1. **Reduce Host Costs**
   Given that price appears to be an important factor in whether customers choose to reserve a listing or not, one recommendation is that Airbnb help subsidize hosts' costs of running Airbnbs to possibly provide leeway for hosts to decrease their prices and attract more customers. For example, Airbnb may decrease the rates of their hosts' service fees or possibly provide stipends so that hosts may better afford home upgrades or cleaning supplies. A drawback of this is the cost that Airbnb would have to take on. This would be extra expenses or reduced revenue on their part; however, if the decrease in host costs increases their reservation rates enough,

the revenue they receive may be enough to also offset the increased expenses for Airbnb.

2. **Provide Real Estate Assistance**
   Additionally, given that the total number of listings that a host has is also important in factor in whether a customer reserves a listing or not, it may be useful for Airbnb to provide additional resources of how hosts may acquire new properties to lease. This may include providing databases of properties that are available and guiding hosts through the process of acquiring them. The drawback is that this would require additional skills within Airbnb to create such a resource. Additionally, the success of such a resource would vary depending on location and how many properties are available in general.

References

Airbnb. 2020. "What Are the Health and Safety Requirements for Airbnb Stays?" Accessed February 12, 2021. https://www.airbnb.com/help/article/2839/what-are-the-health-and-safety-requirements-for-airbnb-stays.

CNN. 2021. "Travel to New York City During Covid-19: What You Need to Know Before You Go." Accessed March 6, 2021. https://www.cnn.com/travel/article/new-york-city-travel-covid-19/index.html.

Dua, André, Deepa Mahajan, Lucienne Oyer, and Sree Ramaswamy. 2020. "US Small-Business Recovery After the Covid-19 Crisis." Accessed February 12, 2021. https://www.mckinsey.com/industries/public-and-social-sector/our-insights/us-small-business-recovery-after-the-covid-19-crisis.

Inside Airbnb. 2020. "Get the Data." Accessed February 12, 2021. http://insideairbnb.com/get-the-data.html.

———. 2021. "About Inside Airbnb." Accessed February 12, 2021. http://insideairbnb.com/about.html#disclaimers.

Kerr, Alexandra. 2020. "A Historical Timeline of Covid-19 in New York City." October 6, 2020. https://www.investopedia.com/historical-timeline-of-covid-19-in-new-york-city-5071986.

Lane, Lea. 2020. "How Bad Are Covid-19 Pandemic Effects on Airbnb Guests, Hosts?" June 9, 2020. https://www.forbes.com/sites/lealane/2020/06/09/how-bad-are-covid-19-pandemic-effects-on-airbnb-guests-hosts/?sh=4fc310887432.

NYC Health. 2021. "COVID-19: Data." Accessed February 12, 2021. https://www1.nyc.gov/site/doh/covid/covid-19-data.page.

Shish. 2016. "R How to Visualize Confusion Matrix Using the Caret Package." September 13, 2016. https://stackoverflow.com/questions/23891140/r-how-to-visualize-confusion-%20matrix-using-the-caret-package.

USA Facts. 2021. "US Coronavirus Cases and Deaths." Accessed February 12, 2021. https://usafacts.org/visualizations/coronavirus-covid-19-spread-map/?utm_source=usnews&utm_medium=partnership&utm_campaign=2020&utm_content=healthiestcommunitiescovid.

Appendix A: Data

Table 4: Listings Dataset Split into 4 Tables

id	name	host id	host since	host location
2595	Skylit Midtown Castle	2845	9/9/08	New York, New York, United States
3831	Cozy Entire Floor of Brownstone	4869	12/7/08	New York, New York, United States
5099	Large Cozy 1 BR Apartment In Midtown East	7322	2/2/09	New York, New York, United States
5121	BlissArtsSpace!	7356	2/3/09	New York, New York, United States
5178	Large Furnished Room Near B'way	8967	3/3/09	New York, New York, United States

id	is superhost	host total listings	host response rate	host verifications	has profile pic
2595	FALSE	6	63	9	TRUE
3831	FALSE	1	100	4	TRUE
5099	FALSE	1	NA	5	TRUE
5121	FALSE	1	100	8	TRUE
5178	FALSE	1	94	4	TRUE

id	neighborhood	neighborhood group	property type	room type	accommodates
2595	Midtown	Manhattan	Apartment	Entire home/apt	1
3831	Clinton Hill	Brooklyn	Guest suite	Entire home/apt	3
5099	Murray Hill	Manhattan	Apartment	Entire home/apt	2
5121	Bedford-Stuyvesant	Brooklyn	Apartment	Private room	2
5178	Hell's Kitchen	Manhattan	Apartment	Private room	2

Table 7: Calendar Dataset

id	bedrooms	beds	price	minimum nights	maximum nights	year	reservation rate
2595	0	1	225	7	1125	2020	high
3831	1	4	89	1	730	2020	moderate
5099	1	1	200	3	21	2020	high
5121	1	1	60	29	730	2020	low
5178	1	1	79	2	14	2020	low

listing id	date	available	price	minimum nights	maximum nights
17693	2020-01-04	TRUE	75	2	28
2595	2020-01-04	TRUE	175	7	1125
2595	2020-01-05	TRUE	175	7	1125
2595	2020-01-06	TRUE	175	7	1125
2595	2020-01-07	TRUE	175	7	1125

Appendix B: Data preparation details

R

We only want to work with a select number of columns, so we selected the relavent columns and ignored the unwanted columns.

```
listings.2020.cleansed <- listings.2020 %>%
  select(id, name, host_id, host_since, host_location, host_is_superhost,
         host_total_listings_count, host_response_rate, host_verifications,
         host_has_profile_pic, neighbourhood_cleansed,
         neighbourhood_group_cleansed, property_type, room_type, accommodates,
         bedrooms, beds, price, minimum_nights, maximum_nights)

listings.2021.cleansed <- listings.2021 %>%
  select(id, name, host_id, host_since, host_location, host_is_superhost,
         host_total_listings_count, host_response_rate, host_verifications,
         host_has_profile_pic, neighbourhood_cleansed,
         neighbourhood_group_cleansed, property_type, room_type, accommodates,
         bedrooms, beds, price, minimum_nights, maximum_nights)
```

Then we added a new column named "year", so the 2020 data could be combined with the 2021 data using the rbind function.

```
listings.2020.cleansed <- listings.2020.cleansed %>%
  mutate(year=2020)
listings.2021.cleansed <- listings.2021.cleansed %>%
  mutate(year=2021)

listings.cleansed <- rbind(listings.2020.cleansed, listings.2021.cleansed)
```

We then converted Host_is_superhost and host_has_profile_pic from character vectors to logical vectors using the mutate function.

```
listings.cleansed <- listings.cleansed %>%
  mutate(host_is_superhost=host_is_superhost == "t") %>%
  mutate(host_has_profile_pic=host_has_profile_pic == "t")
```

The price variable in the listings dataset started as a character string with dollar symbols and commas for larger numbers. Using the mutate function, we converted the price variable to a numeric vector.

```
listings.cleansed <- listings.cleansed %>%
  mutate(price=gsub("\\$", "", price)) %>%
  mutate(price=as.numeric(gsub(",", "", price)))
```

Using the sapply function with a custom function, we converted the host_verifications column to be the number of verifications instead of listing the verifications.

```
listings.cleansed$host_verifications <-
  sapply(listings.cleansed$host_verifications,
         function(x) {length(unlist(strsplit(x, ", ")))})
```

Using the rename function, we renamed neighbourhood variables to remove "cleansed" at the end of the variable name.

```
listings.cleansed <- listings.cleansed %>%
  rename(neighbourhood=neighbourhood_cleansed,
         neighbourhood_group=neighbourhood_group_cleansed)
```

Then, we converted host_response_rate variable into a numeric vector using the gsub function to remove the % symbol and then mutate.

```
listings.cleansed <- listings.cleansed %>%
  mutate(host_response_rate=as.numeric(gsub("%", "", host_response_rate)))
```

The price variable in the calendar dataset started as a character string with dollar symbols and commas for larger numbers. We used the mutate function to convert the price variable to a numeric vector.

```
calendar.2020.cleansed <- calendar.2020 %>%
  mutate(price=gsub("\\$", "", price),
         adjusted_price=gsub("\\$", "", adjusted_price)) %>%
  mutate(price=as.numeric(gsub(",", "", price)),
```

```
        adjusted_price=as.numeric(gsub(",", "", adjusted_price))) %>%
  mutate(available=available == "t")

calendar.2021.cleansed <- calendar.2021 %>%
  mutate(price=gsub("\\$", "", price),
         adjusted_price=gsub("\\$", "", adjusted_price)) %>%
  mutate(price=as.numeric(gsub(",", "", price)),
         adjusted_price=as.numeric(gsub(",", "", adjusted_price))) %>%
  mutate(available=available == "t")
```

Lastly, we created a reservation rate variable by determining the proportion of days a listing was reserved from the calendar data set by grouping the data by Id and summing the days when it was not available and dividing that count by 365. We then created the thresholds for "high", "moderate", and "low" values and assigned them accordingly.

```
booked<-calendar.2020.cleansed[calendar.2020.cleansed$available==FALSE,]

tfcounts<-booked%>%
  group_by(id)%>%
  count()

fulldata <- merge(listings.2020.cleansed, tfcounts, by="id")%>%
  rename(days_reserved=n)%>%
  drop_na()

fulldata <- fulldata%>%
  mutate(value = if_else(days_reserved/365 >= 0.75, "high",
                         if_else(days_reserved/365 >=0.50, "moderate", "low")))%>%
  mutate(value = factor(value, levels = c("high","moderate","low")))%>%
   drop_na()
```

Excel

The first step in preparing the data is removing columns that are not useful for answering the business problem using the "Hide" function.

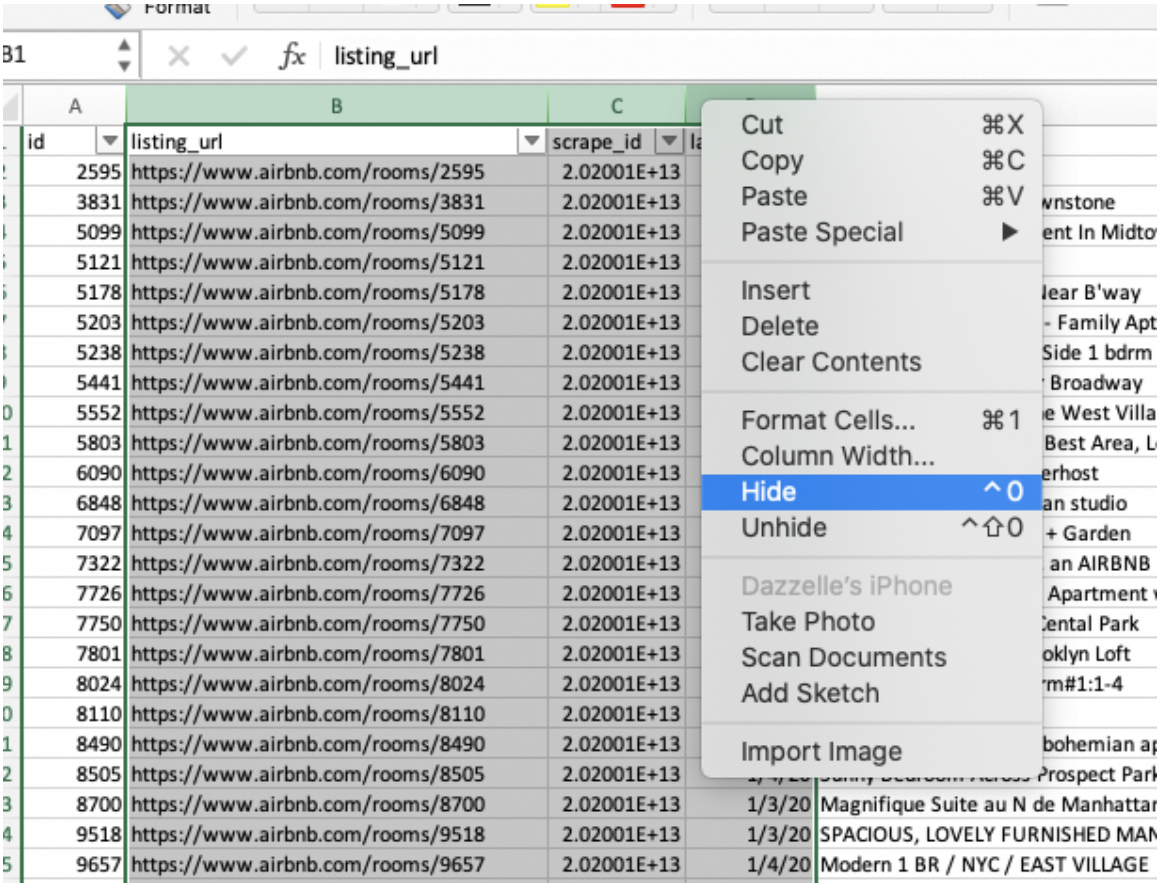


Figure 5: Hiding Unwanted Columns

It is then important to check that there are no duplicate values. In this dataset, each listing is given a unique ID number. We checked for duplicate values using the "Conditional Formatting" function to highlight any duplicate values in red. Next, in order to see the duplicate values, we use the "Sort & Filter" function. This should reveal a drop-down menu, where one will now select "Custom Sort". The data will be sorted by the "id" column, sorted on "Cell Color" with the red cell color on top. However, in this instance, under the "Color/Icon" column, there is no option to sort for the red color which indicates that there are no duplicate values.

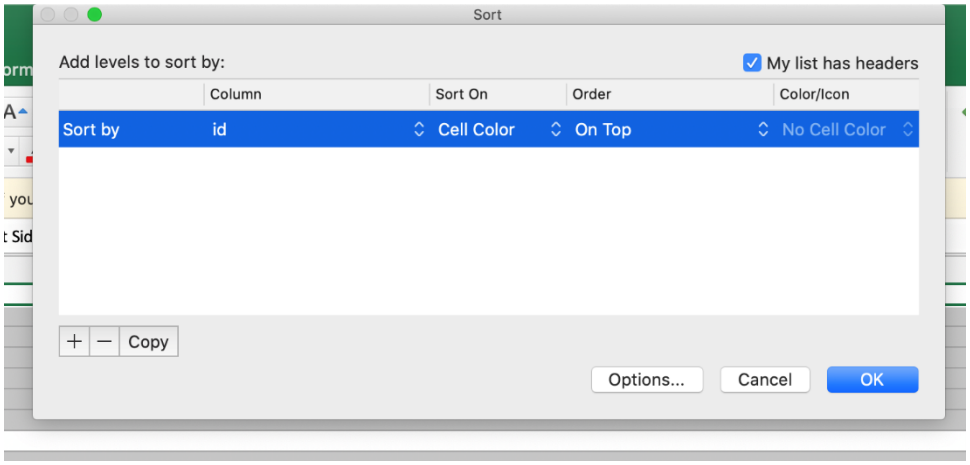


Figure 6: Sort by ID on Cell Color

Afterwards, certain columns were converted into appropriate number formats. For example, the "date" column was converted into Short Date, the "host_response_rate" column was converted into Percentage, and "price" was converted into Currency. Any columns containing numerical values were converted into Number. This was done by highlighting the specific columns one at a time, clicking the drop-down menu located to the left of "Conditional Formatting", and choosing the appropriate number format.

We then inserted a new column titled "year" in each data set in order to be able to combine the data for both 2020 and 2021 into one data set while still maintaining the year in which the data was collected. We also added a new column titled "Host Verifications" which gives a count of the number of verifications that hosts required of their customers rather using the full list. This was done using the "Text to Columns" feature to separate the verifications list by commas. We then used the formula "COUNTA" to sum the number of verifications into a new column.

The "calendar-2020" dataset was already prepared well by Insideairbnb.com (Inside Airbnb 2020). In order to further prepare it, the "listing_id" column was converted to Num-

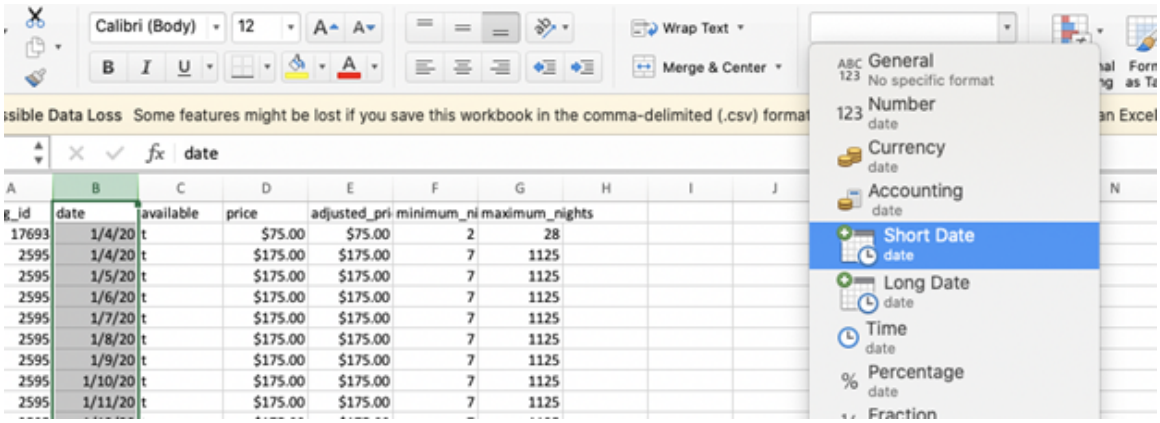


Figure 7: Convert Date to Short Date Type

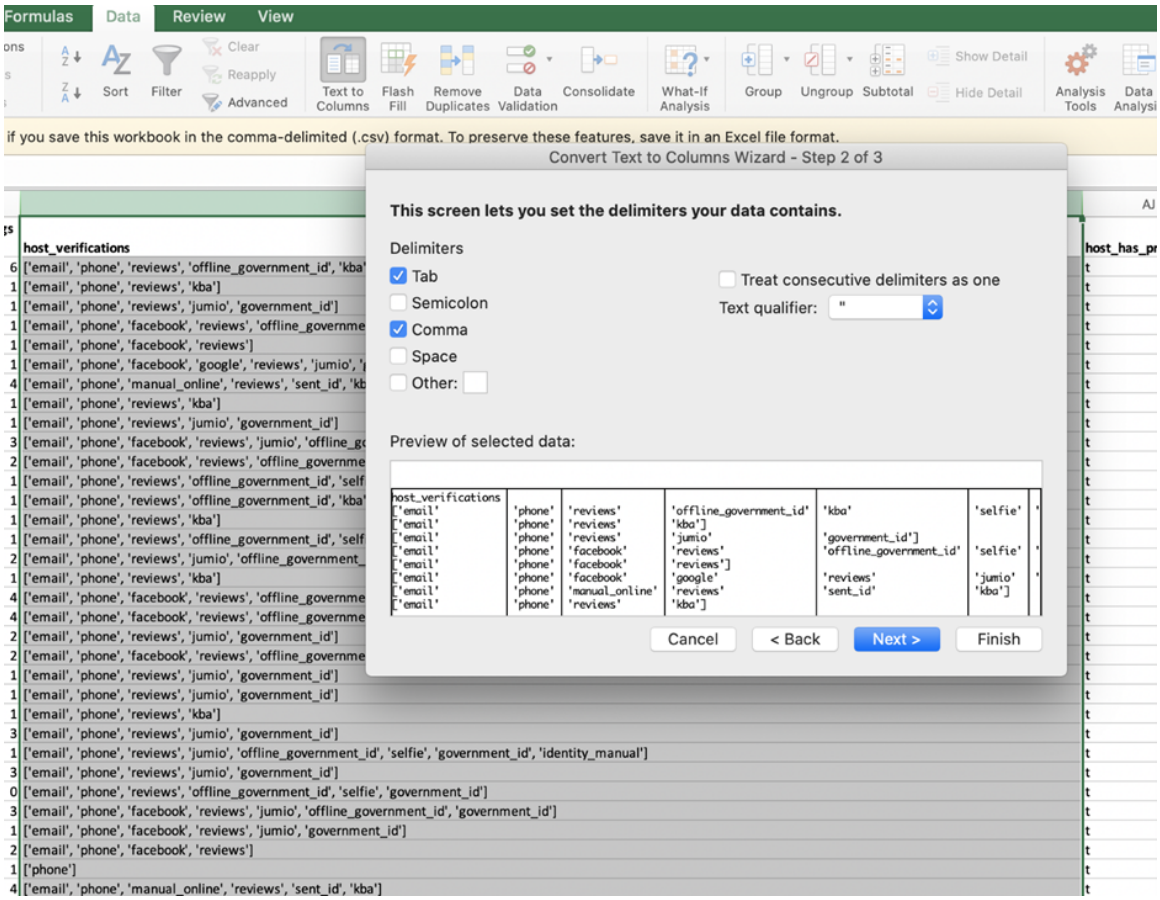


Figure 8: Separate Column with Text to Columns

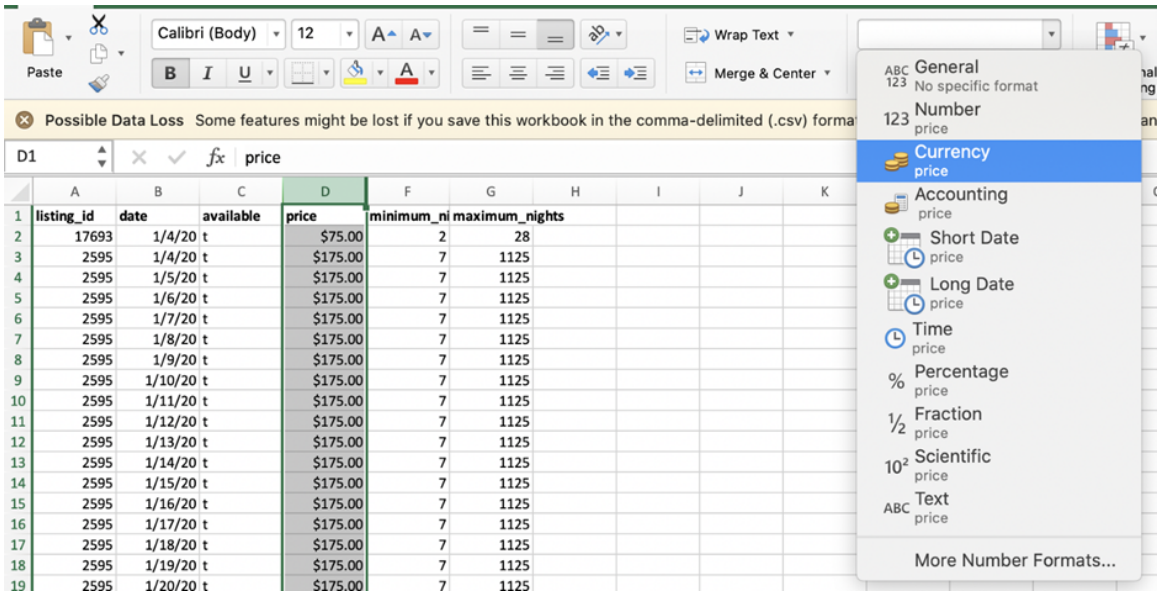
id	name	host_id	host_since	host_location	host_response_rate	host_is_superhost	host_total_listings_count	host_has_profile_pic	neighbourhood
2595	Skylit Midtown Castle	2845	9/9/08	New York, New York, United States	63.00%	f	6	t	Midtown
3831	Cozy Entire Floor of Brownstone	4869	12/7/08	New York, New York, United States	100.00%	f	1	t	Brooklyn
5121	BlissArtsSpace!	7356	2/3/09	New York, New York, United States	100.00%	f	1	t	Bedford-Stuyvesant
5178	Large Furnished Room Near B'way	8967	3/3/09	New York, New York, United States	94.00%	f	1	t	Manhattan
5238	Cute & Cozy Lower East Side 1 bdrm	7549	2/7/09	New York, New York, United States	100.00%	t	4	t	Lower East Side
5441	Central Manhattan/near Broadway	7989	2/16/09	New York, New York, United States	100.00%	t	1	t	Manhattan
5803	Lovely Room 1, Garden, Best Area, Legal rental	9744	3/10/09	New York, New York, United States	90.00%	t	3	t	Park Slope
7097	Perfect for Your Parents + Garden	17571	5/17/09	New York, New York, United States	100.00%	t	1	t	Brooklyn
7322	Chelsea Perfect by Doti, an AIRBNB Super Host!	18946	5/27/09	New York, New York, United States	100.00%	t	1	t	Manhattan
7726	Hip Historic Brownstone Apartment with Backyard	20950	6/10/09	New York, New York, United States	100.00%	t	1	t	Brooklyn
7750	Huge 2 BR Upper East Cental Park	17985	5/20/09	New York, New York, United States	50.00%	f	2	t	Manhattan
7801	Sweet and Spacious Brooklyn Loft	21207	6/12/09	New York, New York, United States	100.00%	f	1	t	Williamsburg
8024	CBG CtyBGd HelpsHaiti rm#1:1-4	22486	6/20/09	Brooklyn, New York, United States	93.00%	f	4	t	Park Slope
8110	CBG Helps Haiti Rm #2	22486	6/20/09	Brooklyn, New York, United States	93.00%	f	4	t	Brooklyn
8490	MAISON DES SIRENES1,bohemian apartment	25183	7/10/09	New York, New York, United States	100.00%	t	2	t	Brooklyn
8505	Sunny Bedroom Across Prospect Park	25326	7/12/09	New York, New York, United States	100.00%	t	2	t	Brooklyn
9518	SPACIOUS, LOVELY FURNISHED MANHATTAN BEDROOM	31374	8/12/09	New York, New York, United States	100.00%	f	1	t	Inwood
9657	Modern 1 BR / NYC / EAST VILLAGE	21904	6/16/09	New York, New York, United States	100.00%	f	1	t	East Village
9668	front room/double bed	32294	8/18/09	New York, New York, United States	90.00%	f	3	t	Harlem

Figure 9: Prepared Listings Data Split for Readability

neighbourhood_group_cleansed	property_type	room_type	accommodates	bathrooms	bedrooms	beds	price	minimum_nights	maximum_nights	Year	Host Verifications
Manhattan	Apartment	Entire home/apt	1	1	0	1	$225.00	7	1125	2020	9
Brooklyn	Guest suite	Entire home/apt	3	1	1	4	$89.00	1	730	2020	4
Brooklyn	Apartment	Private room	2		1	1	$60.00	29	730	2020	8
Manhattan	Apartment	Private room	2	1	1	1	$79.00	2	14	2020	4
Manhattan	Apartment	Entire home/apt	3	1	1	2	$150.00	1	1125	2020	6
Manhattan	Apartment	Private room	2	1	1	1	$99.00	2	7	2020	4
Brooklyn	Townhouse	Private room	2	1.5	1	0	$89.00	4	14	2020	7
Brooklyn	Apartment	Entire home/apt	4	1	1	2	$199.00	1	30	2020	8
Manhattan	Apartment	Private room	3	1	1	1	$120.00	1	180	2020	4
Brooklyn	Townhouse	Entire home/apt	4	1	1	1	$99.00	3	730	2020	8
Manhattan	Apartment	Entire home/apt	4	1	2	2	$190.00	7	365	2020	6
Brooklyn	Loft	Entire home/apt	4	1		2	$299.00	3	90	2020	4
Brooklyn	Bed and breakfast	Private room	4	3.5	1	2	$115.00	1	120	2020	8
Brooklyn	Bed and breakfast	Private room	3	2.5	4	8	$32.00	2	730	2020	8
Brooklyn	Loft	Entire home/apt	5	1	1	4	$120.00	2	365	2020	5
Brooklyn	Condominium	Private room	2	1	1	1	$60.00	1	20	2020	9
Manhattan	Apartment	Private room	2	1	1	1	$44.00	3	30	2020	5
Manhattan	Apartment	Entire home/apt	3	1	1	1	$175.00	7	60	2020	4
Manhattan	Apartment	Private room	2	1	1	1	$50.00	3	365	2020	5
Manhattan	Apartment	Private room	2	1	1	1	$52.00	2	730	2020	8

Figure 10: Prepared Listings Data Split for Readability

ber, the "price" column was converted into Currency, and "minimum_nights" and "maximum_nights" were converted into Number format. This was done by highlighting the specific columns one at a time, clicking the drop-down menu located to the left of "Conditional Formatting", and choosing the appropriate number format.



	A	B	C	D	F	G
1	listing_id	date	available	price	minimum_ni	maximum_nights
2	17693	1/4/20	t	$75.00	2	28
3	2595	1/4/20	t	$175.00	7	1125
4	2595	1/5/20	t	$175.00	7	1125
5	2595	1/6/20	t	$175.00	7	1125
6	2595	1/7/20	t	$175.00	7	1125
7	2595	1/8/20	t	$175.00	7	1125
8	2595	1/9/20	t	$175.00	7	1125
9	2595	1/10/20	t	$175.00	7	1125
10	2595	1/11/20	t	$175.00	7	1125
11	2595	1/12/20	t	$175.00	7	1125
12	2595	1/13/20	t	$175.00	7	1125
13	2595	1/14/20	t	$175.00	7	1125
14	2595	1/15/20	t	$175.00	7	1125
15	2595	1/16/20	t	$175.00	7	1125
16	2595	1/17/20	t	$175.00	7	1125
17	2595	1/18/20	t	$175.00	7	1125
18	2595	1/19/20	t	$175.00	7	1125
19	2595	1/20/20	t	$175.00	7	1125

Figure 11: Convert Price to Currency Type

	listing_id	date	available	price	minimum_nights	maximum_nights
2	17693	1/4/20	t	$75.00	2	28
3	2595	1/4/20	t	$175.00	7	1125
4	2595	1/5/20	t	$175.00	7	1125
5	2595	1/6/20	t	$175.00	7	1125
6	2595	1/7/20	t	$175.00	7	1125
7	2595	1/8/20	t	$175.00	7	1125
8	2595	1/9/20	t	$175.00	7	1125
9	2595	1/10/20	t	$175.00	7	1125
10	2595	1/11/20	t	$175.00	7	1125
11	2595	1/12/20	t	$175.00	7	1125
12	2595	1/13/20	t	$175.00	7	1125
13	2595	1/14/20	t	$175.00	7	1125
14	2595	1/15/20	t	$175.00	7	1125
15	2595	1/16/20	t	$175.00	7	1125
16	2595	1/17/20	t	$175.00	7	1125

Figure 12: Prepared Calendar Data

Appendix C: Analytics details

Descriptive analytics

6.0.1 R

I. Difference in Bookings Graph

First we convert the data string to a date format in order to extract the month for each booking. We also subset the data to be only the first 3 months out of each year.

```
bookings.2020 <- calendar.20203 %>%
  filter(available==F) %>%
  mutate(month=format(as.Date(date, "%Y-%m-%d"), "%m")) %>%
  filter(as.numeric(month) < 4) %>%
  group_by(month) %>%
  summarise(n())

bookings.2020 <- bookings.2020 %>%
  mutate(year=2020)
```

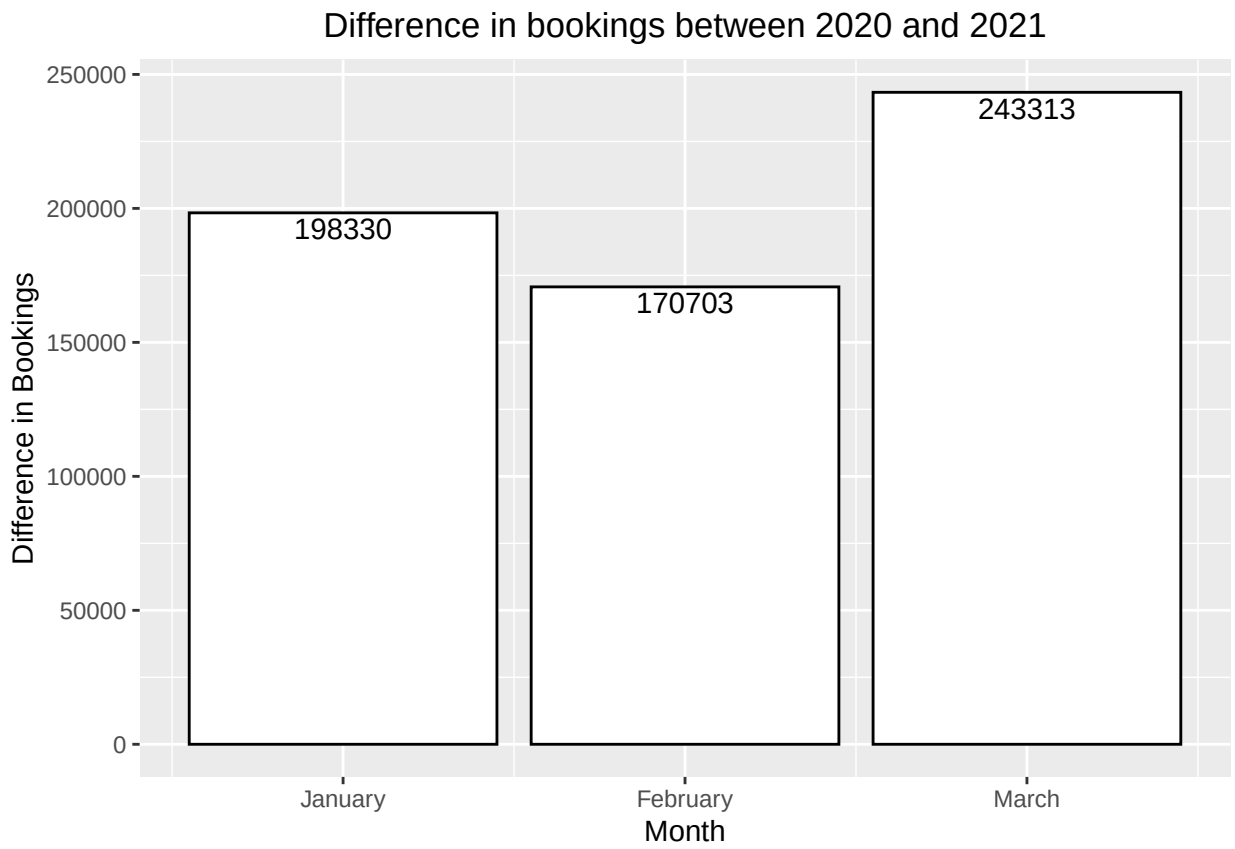
```
bookings.2021 <- calendar.20213 %>%
  filter(available==F) %>%
  mutate(month=format(as.Date(date, "%Y-%m-%d"), "%m")) %>%
  filter(as.numeric(month) < 4) %>%
  group_by(month) %>%
  summarise(n())

bookings.2021 <- bookings.2021 %>%
  mutate(year=2021)
```

We then combined the two years of bookings data and renamed the columns. Using this combined dataset, we created a barplot with the month on the x axis and the number of bookings on the y axis and the color of the bar plot is based on the year. We also changed the labels to be the name of the month instead of the number.

```
bookings <- rbind(bookings.2020, bookings.2021)
colnames(bookings) <- c("month", "bookings", "year")
bookings.diff <- bookings[bookings$year==2020,c("bookings")] -
  bookings[bookings$year==2021, c("bookings")]
bookings.diff <- cbind(bookings.diff, c(1, 2, 3))
colnames(bookings.diff) <- c("bookings", "month")

ggplot(bookings.diff, aes(x=month, y=bookings)) +
  geom_bar(stat = "identity", fill="white", color="black") +
  labs(x="Month", y="Difference in Bookings", title="Difference in bookings between 2020
  theme(plot.title = element_text(hjust = 0.5)) +
  geom_text(aes(label=bookings), vjust=1.25) +
  scale_x_continuous(labels=c("January", "February", "March"),
                     breaks = c(1, 2, 3))
```



II. Bookings Trend in 2020 Graph

First, we converted the data string to a date format in order to extract the month for each booking in 2020.

```
bookings.full.2020 <- calendar.20203 %>%
  filter(available==F) %>%
  mutate(month=format(as.Date(date, "%Y-%m-%d"), "%m")) %>%
  group_by(month) %>%
  arrange(as.numeric(month)) %>%
  summarise(n())

colnames(bookings.full.2020) <- c("month", "bookings")
```

We then created another barplot with month on the x axis and bookings on the y axis for each month in 2020. We also changed the labels and text on the plot, so that the names of each month would be displayed instead of the numeric month.

```
ggplot(bookings.full.2020, aes(x=month, y=bookings)) +
  geom_bar(stat = "identity") +
  labs(title="Bookings by Month for 2020", x="Month", y="Bookings") +
  theme(plot.title = element_text(hjust = 0.5)) +
  scale_x_discrete(labels=c("January", "February", "March", "April", "May",
                            "June", "July", "August", "September", "October",
                            "November", "December")) +
  theme(axis.text.x=element_text(angle=45, vjust = 1, hjust = 1,
                                 face="bold", size=12))
```

III. Most Used Adjectives in Descriptions Graph We first read in the data that took the summaries from the description category in each data set, and then took the count of how frequently each word was mentioned.

```
summary_word_count20 <- read.csv("summary_word_count.csv")
summary_word_count21 <- read.csv("words.csv")
```

We then filtered the data so that we were only looking at longer words, to exclude stop words such as and, the, that, etc. Words that don't actually provide any context

```
summary20new <- summary_word_count20[nchar(summary_word_count20$X) > 7 ,]
summary21new <- summary_word_count21[nchar(summary_word_count21$X) > 7 ,]
```

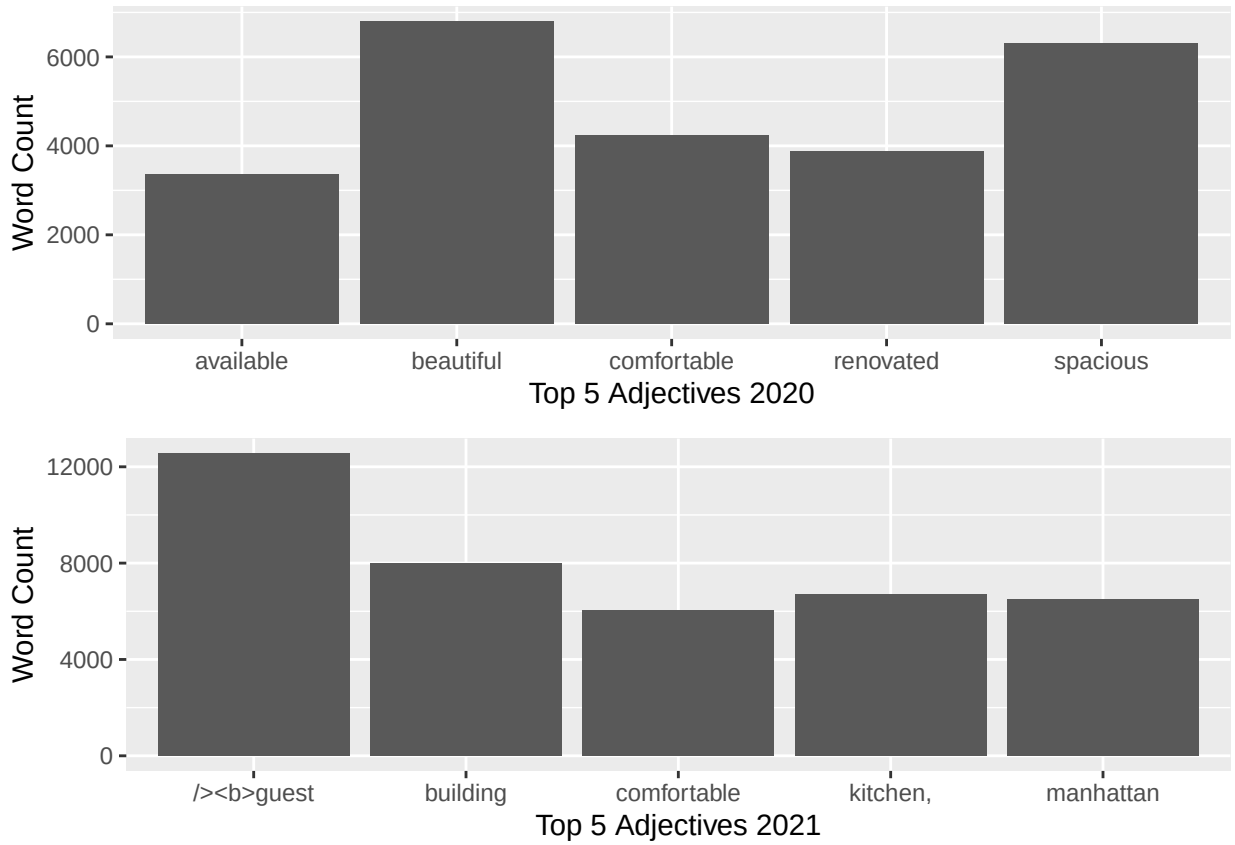
Next, we went through and picked the adjectives then plotted them as a bar chart.

```
p<-ggplot(data=summary21new[c(14, 7, 12,13,4),], aes(x=X, y=X0)) +
  geom_bar(stat ="identity") + xlab("Top 5 Adjectives 2021") +
  ylab("Word Count")

d <- ggplot(data=summary20new[c(18,4,13,2,12),], aes(x=X, y=X0)) +
  geom_bar(stat ="identity") + xlab("Top 5 Adjectives 2020") +
  ylab("Word Count")
```

Lastly, we put the two graphs together.

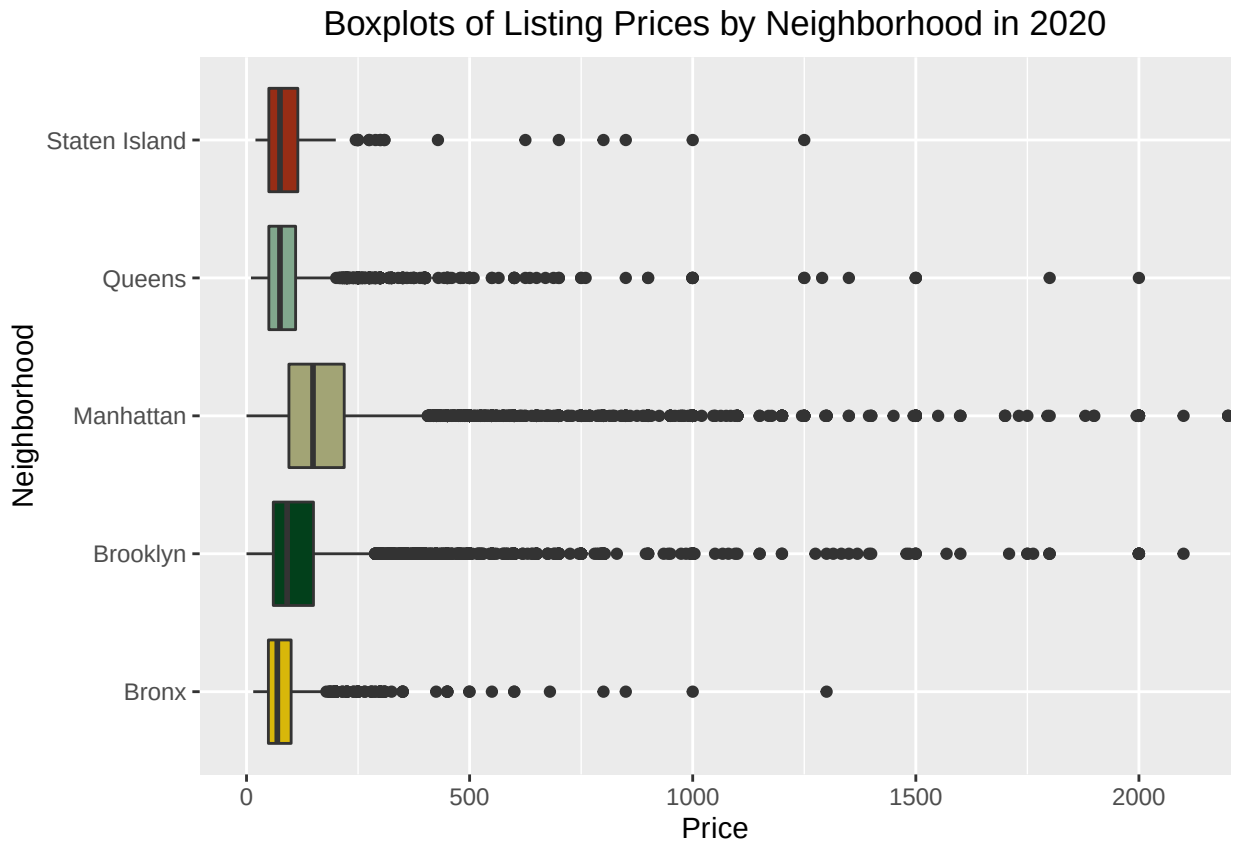
```
grid.arrange(d, p, nrow = 2)
```



III. Boxplot of prices by neighborhood.

We created a boxplot of price distributions grouped by neighborhood using the listings.2020.cleansed data with "Price" on the x-axis and "Neighborhood" on the y-axis.

```
library(wesanderson)
ggplot(listings3[listings3$year==2020,],
       aes(x=price,y=neighbourhood_group,
       fill=neighbourhood_group))+
  geom_boxplot()+
  coord_cartesian(xlim = c(0, 2100)) +
  ggtitle("Boxplots of Listing Prices by Neighborhood in 2020")+
  theme(plot.title = element_text(hjust = 0.5))+
  xlab("Price")+ylab("Neighborhood") +
  scale_fill_manual(values=wes_palette(n=5, name="Cavalcanti1")) +
  theme(legend.position="none")
```



Predictive analytics

I. Random Forest Analysis

First we subsetted the data to remove any listing identification columns and then split up the data into training, validation, and testing sets.

```
fulldata<-fulldata%>%
  select(host_is_superhost, host_total_listings_count,
         host_verifications, host_has_profile_pic,
         room_type, accommodates, bedrooms, beds, price,
         minimum_nights, maximum_nights, value)

listings.3div <- fulldata %>%
  initial_split(prop = 0.6)

listings.3div2 <- listings.3div %>%
  testing() %>%
  initial_split(prop = 0.5)

listings.train <- training(listings.3div)
listings.validate <- training(listings.3div2)
listings.test <- testing(listings.3div2)
```

Next, we created the initial random forest using all variables with the training set.

```
RF.mod1 <- randomForest(value~., data=listings.train,mtry = 2, importance = TRUE)
```

The second random forest model was created by looking at the important variables from model 1, and removing those that did not result in large decreases in accuracy if removed.

```
RF.mod2 <- randomForest(value~. - host_has_profile_pic - beds, data=listings.train,mtry
```

The third random forest model was created by removing an additional variable that did not result in large decreases in accuracy if removed .

```
RF.mod3 <- randomForest(value~. - host_has_profile_pic -bedrooms - beds, data=listings.t
```

We then added the validation data to all the models and added predicted values to the models and compared their confusion matrices and accuracy rates.

```
RF.add1 <- listings.validate%>%
  add_predictions(RF.mod1) %>%
  rename(pred_value = pred) %>%
  mutate(method = "RF.mod1")

RF.add2 <- listings.validate %>%
  add_predictions(RF.mod2) %>%
  rename(pred_value = pred) %>%
  mutate(method = "RF.mod2")

RF.add3 <- listings.validate %>%
  add_predictions(RF.mod3) %>%
  rename(pred_value = pred) %>%
  mutate(method = "RF.mod3")

RF.add <- RF.add1 %>%
  bind_rows(RF.add2) %>%
  bind_rows(RF.add3) %>%
  group_by(method)
```

```
RF.add1 %>%
  conf_mat(truth = value, estimate = pred_value)
```

```
##           Truth
## Prediction high moderate  low
##   high     5603      957 1366
##   moderate    4        9    3
##   low       357      295 1007
```

```
RF.add2 %>%
  conf_mat(truth = value, estimate = pred_value)
```

```
##           Truth
```

```
## Prediction high moderate  low
##   high     5600      948 1364
##   moderate   13       21    9
##   low       351      292 1003
```

```
RF.add3 %>%
  conf_mat(truth = value, estimate = pred_value)
```

```
##           Truth
## Prediction high moderate  low
##   high     5573      946 1358
##   moderate   15       19   18
##   low       376      296 1000
```

```
RF.add %>%
  metrics(truth = value, estimate = pred_value) %>%
  filter(.metric == "accuracy")
```

```
## # A tibble: 3 x 4
##   method  .metric  .estimator .estimate
##   <chr>   <chr>    <chr>          <dbl>
## 1 RF.mod1 accuracy multiclass     0.689
## 2 RF.mod2 accuracy multiclass     0.690
## 3 RF.mod3 accuracy multiclass     0.687
```

We then added the testing data to all the models and added predicted values to the models and compared their confusion matrices and accuracy rates to confirm how well our models predict reseration rates.

```
RF.add1test <- listings.test %>%
  add_predictions(RF.mod1) %>%
  rename(pred_value = pred) %>%
  mutate(method = "RF.mod1")
```

```
RF.add2test <- listings.test %>%
  add_predictions(RF.mod2) %>%
```

```
  rename(pred_value = pred) %>%
  mutate(method = "RF.mod2")

RF.add3test <- listings.test %>%
  add_predictions(RF.mod3) %>%
  rename(pred_value = pred) %>%
  mutate(method = "RF.mod3")

RF.addtest <- RF.add1test %>%
  bind_rows(RF.add2test) %>%
  bind_rows(RF.add3test) %>%
  group_by(method)

RF.add1test %>%
  conf_mat(truth = value, estimate = pred_value)

##           Truth
## Prediction high moderate  low
##   high     5605      968 1391
##   moderate    8       21   11
##   low       375      242  980

RF.add2test %>%
  conf_mat(truth = value, estimate = pred_value)

##           Truth
## Prediction high moderate  low
##   high     5586      942 1361
##   moderate   21       40   16
##   low       381      249 1005

RF.add3test %>%
  conf_mat(truth = value, estimate = pred_value)

##           Truth
## Prediction high moderate  low
```

```
##   high     5570      945 1354
##   moderate   20       36   25
##   low       398      250 1003
```

```
RF.addtest %>%
  metrics(truth = value, estimate = pred_value) %>%
  filter(.metric == "accuracy")
```

```
## # A tibble: 3 x 4
##   method  .metric  .estimator .estimate
##   <chr>   <chr>    <chr>          <dbl>
## 1 RF.mod1 accuracy multiclass     0.688
## 2 RF.mod2 accuracy multiclass     0.691
## 3 RF.mod3 accuracy multiclass     0.688
```

II. Text Analysis

First, we tokenize the name columns so each word of the names is on a new line for 2020 and remove punctuation and then remove stop words for each year.

```
listings.names2020 <- data2020 %>%
  unnest_tokens(word, description) %>%
  mutate(word = str_extract(word, "[0-9a-z']+"))
```

```
listings.names2021 <- data2021 %>%
  unnest_tokens(word, description) %>%
  mutate(word = str_extract(word, "[0-9a-z']+"))
```

```
check2020 <- listings.names2020$word
check2021 <- listings.names2021$word
```

```
# remove stop words 2020
listings.names2020 <- listings.names2020 %>%
  anti_join(stop_words, by = "word")  %>%
  mutate(word.stem = wordStem(word))
```

```
# remove stop words 2021
listings.names2021 <- listings.names2021 %>%
  anti_join(stop_words, by = "word")  %>%
  mutate(word.stem = wordStem(word))
```

Then, we determined the count of words across all descriptions.

```
word.count2020 <- listings.names2020 %>%
  group_by(word) %>%
  summarize(word.count = n()) %>%
  arrange(desc(word.count))

word.count2021 <- listings.names2021 %>%
  group_by(word) %>%
  summarize(word.count = n()) %>%
  arrange(desc(word.count))
```

Then, we added in additional stop words specific to this analysis.

```
new_stop_words <- tibble(word = c(NA, "bedroom","private","apartment","cozy","br", "bed"
                                  "living", "1","2","3","apt","studio","located"))
```

Then we removed new stop words.

```
listings.names2020 <- listings.names2020 %>%
  anti_join(new_stop_words, by = "word")

listings.names2021 <- listings.names2021 %>%
  anti_join(new_stop_words, by = "word")
```

Then we performed a new word count with all of the new stop words gone.

```
new.word.count2020 <- listings.names2020 %>%
  group_by(word) %>%
  summarize(word.count = n()) %>%
  arrange(desc(word.count))
```

```
new.word.count2021 <- listings.names2021 %>%
  group_by(word) %>%
  summarize(word.count = n()) %>%
  arrange(desc(word.count))
```

We then took the top 10 words used in descriptions from 2020 and 2021 and combined them into one dataframe.

```
# display of top 10 words in name 2020
display2020 <- new.word.count2020 %>%
  top_n(10, word.count)

# display of top 10 words in name 2021
display2021 <- new.word.count2021 %>%
  top_n(10, word.count)

colnames(display2020) <- c("2020 Words","Frequency")
colnames(display2021) <- c("2021 Words","Frequency")

cbind(display2020, display2021) %>% kable(caption = "Top Description Words 2020 & 2021")
```

Table 9: Top Description Words 2020 & 2021

2020 Words	Frequency	2021 Words	Frequency
kitchen	40924	space	42048
walk	28588	kitchen	28018
restaurants	26404	access	25562
manhattan	26275	bathroom	17550
bathroom	25251	guest	17482
park	24833	building	14663
train	24296	manhattan	14605
space	24267	floor	13828
subway	22434	walk	13622
neighborhood	21982	park	13054

Appendix D: Comment incorporation

6.1 Deliverable 1

Comment 1: Try to be a bit more specific about your business problem. Airbnb is talked about very little, and it is unclear what aspect of the lodging market you are focusing on.

Answer 1: We changed our introduction paragraph to explain why we chose Airbnb (the largest short-term renter in NYC) and what variables in listings we are analyzing.

Comment 2: It could be good to include something more focused on looking at how they will recover rather than just how COVID has impacted them since you included how it could be until 2025.

Answer 2: In our final deliverable, we made recommendations as to what Airbnb could do to recover fast enough. These recommendations came from our predictive analysis and are focused on keep renter costs down in order to decrease the average price per night.

Comment 3: Make it clear whether you are interested in the number of listings or in the profitability of the companies.

Answer 3: In our business problem section, we changed our question to "What are the attributes of listings that have closed or are not receiving reservations during the pandemic?" We chose to focus on listings because profitability numbers for Airbnb were more difficult to find when filtering by city.

Comment 4: You should work as a group to dig deeper into a problem that, when solved, will provide significant returns to the intended audience. As stated, it is too broad to have these sorts of measured effects. Think about how you can benefit Airbnb, if they remain your intended audience, in a way that is more than looking at the historical data they have on hand. Perhaps analyzing different attributes of listing or hosts that closed or re-listed may be one path to go down.

Answer 4: Our initial dataset included over 120 variables. We narrowed it down to 21 that we thought would be most effective and influential for analyzing listing data. We focused on different attributes of listings and why one listing is booked more often than others.

Comment 5: With Airbnb comes an important audience that you did not go into too much detail about which are the active and potential Airbnb-ers. They could also benefit from analysis like yours to gain knowledge about predictive trends in the industry and could be good to include in a few sentences.

Answer 5: We decided not to use customers as our audience because we would have to use different data about customer demographics. Our dataset only included information about the listings and the Airbnb hosts themselves, so our audience will be Airbnb and hosts.

Comment 6: You can be more specific on the costs and benefits to Airbnb. Consider how this affects their resource allocation decisions, or, perhaps, consider how this could affect the decisions of individual Airbnb hosts. Also, I would recommend either more deeply evaluating how other lodging businesses will benefit or narrowing your scope to not include them.

Answer 6: We chose to just focus on Airbnb instead of other lodging businesses (like hotels) because Airbnb had the largest rental market share in NYC for short-term rentals. We also narrowed our main audience down to Airbnb and their hosts. In our final deliverable, we made some recommendations on how Airbnb could better allocate resources to their hosts in order to support them after the pandemic. These recommendations came from our text and predictive analysis.

Comment 7: As a reader, I would be asking about a specific metric of how badly Airbnb was hurt at this start of the pandemic. (i.e., how much business did they lose then). And then this could be compared to an analysis of more recent losses/gains.

Answer 7: Because we decided to focus on listings rather than profitability, there was not one singular metric that could measure how badly Airbnb was hurt. In the descriptive analytics section, we did include a bar plot showing the average price of rentals in various boroughs in New York City. We also looked at the number of listings booked and available before and after the pandemic and created charts for those.

Comment 8: If the point is to analyze how Airbnb is affected by COVID-19 then you will also need to collect data on the trend of cases in the city as well so that you can compare them rather than just looking at pre and post COVID. Also, if the data is not associated with Airbnb, it may be difficult to make a claim for Airbnb based on the analysis.

Answer 8: In the data collection section, we explained how our data is from a site called inairbnb.com, which aggregates information directly from Airbnb itself. While the site is not endorsed by Airbnb, it uses an API connected to Airbnb's website and combines the data by city. We chose to not collect data on the trend of cases in the city and just included information from news articles about when lockdowns were lifted.

Comment 9: I understand what is needed and I think it is good that you include a discussion of the potential issue of data accuracy. I would ask if you were looking at month by month data or yearly data? If yearly, would you compare it to 2019? How would this work considering the pandemic has spanned multiple years and since we have only just started

2021?

Answer 9: We added in our data section that we were looking at both monthly and yearly data. Most of our plots compared data from the first 3 months of 2020 to the first 3 months of 2021. We thought this was the most accurate way to do this because these months represented the before and aftereffects of COVID in the city.

Comment 10: You've listed some of the variables that are available. What are the variables that you need? Are they the same?

Answer 10: In our required data section, we listed some of the data that we will use rather than random columns like we had before. We explained why we chose certain variables and why we removed others. These decisions are also included in our data section in deliverable 2.

Comment 11: A full blank line is needed between paragraphs. Each reference entry should have an author, even if it is the site that has posted the information.

Answer 11: We made all format and reference changes in our final deliverable.

Comment 12: Some of this information may be better suited to the intended audience section. Consider splitting into two paragraphs.

Answer 12: We rewrote our intended audience and business problem paragraph in order to make them less repetitive and more specific.

Comment 13: I think intended audience section was stated clearly and it provided a great explanation on who would actually benefit from this business solution. You definitely left the room on the particular audience, which I especially liked, just because you do not know the results yet, so I believe that that was a very clever idea.

Answer 13: We ended up focusing on Airbnb and Airbnb hosts as our intended audiences because our information and results best fit these groups.

Comment 14: Required Data section was truly insightful, and I understood what type of data you will be looking into. It is great that you already found the dataset, so you could be more precise in this section. I honestly do not see a place for improvement here, just because I believe the information given is covering everything we needed to know at this stage of the project.

Answer 14: In our final deliverable, we narrow down to what 21 variables we are focusing on rather than including all 120.

6.2 Deliverable 2

Comment 1: I understand the description provided and it is complete in describing the columns. Having the column titles and description as a list or table so it will be easy for anyone to go back and quickly check the descriptions. It would also be a little clearer that there were 2 datasets if you showed the listing data right after its description rather than both of them back-to-back. Other than that good info in this section. There should be a way to get rid of that error message at the top of your paper. I don't think a full list of the column names is necessary when you are showing your data. I am not sure I would agree the ID column is the most important column, and you probably don't need to conjecture on the importance of columns here anyway. In general, this section's text can be made much more concise / less wordy.

Answer 1: We made this section much more concise and removed the list of column names to make it much less wordy. We also hid the error message that was appearing at the top of the page. We also removed the conjecture about the importance of the columns.

Comment 2 Summary of a few comments: I understand the description provided and it is very thorough with how it explains every column. It also shows example tables of the dataset, for a visual element to help the reader understand. One question I have afterwards is whether or not the ID is the most important variable as you say, because although it connects the sets and identifies listings, this seems like a more logistical variable and less of one that gives critical information that you will analyze. But I understand that without this variable, you can't really analyze your data so it does make sense of why it is important logistically. I also understand the table and do not have suggestions regarding that.

Answer 2: We removed the section about the ID being the most important column.

Comment 3 Summary of a few comments: This section is really clear and describes data collection methods appropriately. Consider adding in any challenges when collecting data or elaborating on the reliability of the data source since issues regarding that were listed. You have a good description of how you collected the data and why the data is "good data." I think it was great that you noted how often the data is updated. Regarding considerations, I think it would be useful to do a little more analysis on whether the data (which is not actually from Airbnb) is accurate. To elaborate on this, you say that you still think that the data is accurate, but I think if you went into the verification process in a little more detail that would be useful.

Answer 3: We talked more about the reliability of the data source and how we cross checked some of the observations to ensure that the listings were consistent. We also added that the

data source gets their data directly from Airbnb.

Comment 4: Instead of saying "These characteristics make it appropriate to answer the proposed business problem", you can be more explicit about what your problem is and how the characteristics are valuable to it.

Answer 4: We changed our data collection section to include more information about how the characteristics of the dataset fit into our business problem.

Comment 5 Summary of a few comments: I believe the data collection part clearly showed where the data is from. I think the part where you talk about usefulness of the data is really informative because the data is updated monthly.

Answer 5: This comment suggested no improvements.

Comment 6 Summary of a few comments: This is a really clear, non-technical description and I understand the process that was undergone. Consider adding in any challenges when preparing data and perhaps look into any extreme values to make sure 0's won't mess up further analysis. It seems like you did not do much cleaning here, hopefully that's because you've determined it isn't necessary. Also, consider adding any additional considerations that you have yet to mention above. Looking below, there are more cleaning steps you could mention here (at least at a high-level).

Answer 6: We changed this section to address more of the actual cleaning steps that were used to prepare the data and also mentioned some of the challenges and more justifications for why we took the steps we did.

Comment 7: Data preparation part clearly shows the way you manipulated the data. I believe that by removing duplicated data and missing values you ensured the data is ready for analysis. I like the way you showed manipulation with Excel, i believe showing screenshots made it easier to follow.

Answer 7: Comments suggested no improvements.

Comment 8: I understand the process and it is well commented throughout which makes it easy to follow. My only question is why you chose to display the number of N/A values in the table when you said previously you were not removing rows with blank values in preparation. Perhaps another output that can be compared to excel would be helpful as well.

Answer 8: We ended up removing the section to display the number of N/A values in the table.

Comment 9: Try using more active voice to be more direct—you changed the variables instead of "they were changed".

Answer 9: We changed our comments to include more active voice.

Comment 10: I understand how it worked, although a few in-code comments would be useful. It was useful that you wrapped the code so it didn't go off the page. I think it would also be useful to explain what data is being piped into which functions at times. There are also warnings that the code shows that I think should be explained. But overall, I think the code was clear and the explanations were good.

Answer 10: We changed this section to include more comments in the code to better help the reader understand what our R code was doing. We also changed it so that the warnings would be hidden.

Comment 11: This is a clear and easy-to-follow section, ample screenshots. Since this is supposed to be understandable by someone who has Excel knowledge, I am not sure screenshots for some simpler actions like value type changes may not require screenshots. Could make it a little cleaner, but what you have is plenty helpful.

Answer 11: We changed some of our screenshots to not include some of the more trivial things.

Comment 12: Creating a new paragraph at "Then," was not necessary. Consider breaking up the text with the screenshots so that the reader can follow along, and at least mention the figure number that corresponds to the step you are explaining. Answer 12: We changed this section to not start the new paragraph and formatted the screenshots to be in the text instead of at the bottom.

Comment 13: I liked the explanation and screenshots of the specific process that was used. This made me able to understand it completely. Sometimes you say that you changed "certain columns" and I think it would not hurt to be specific instead. After your last set of images, there is no explanation, and I think an explanation would be useful there.

Answer 13: We changed it to clearly state which columns were changed and moved the images to inside of the text, since this was a formatting issue and not a lack of explanation issue.

Comment 14: 1. As this paper stands alone currently, there is no indication from this section that your data is for Airbnb or the location. Even combined, this section should make it clear that the data are appropriate and match the business problem. 2. Several tenses don't seem to match. 3. The data set names do not need to be used until the details sections. 4. As you find which variables are not useful in your analysis, they can be removed here. 5. Column titles usually have some capitalization.

Answer 14: We changed the tenses to match and be more grammatically correct. We also removed the dataset names from this section. We also capitalized the column titles.

Comment 15: One typo in this paragraph ("useful appropriate").

Answer 15: We fixed this type to just be useful.

Comment 16: 1. Thorough job. 2. Consider replacing "that we wanted" with "relevant". 3. Missing output.

Answer 16: We replaced "that we wanted" with "relevant" and added in the output.

Comment 17: 1. Typo in Figure 2 name. 2. Some screenshots are hard to read.

Answer 17: We fixed the typos in the figure names and tried to make the screenshots easier to read.

6.3 Deliverable 3

Comment 1: You can provide a bit more explanation/introduction here as it is not accompanied by past deliverables. 1st graphic: this doesn't seem to provide insight beyond common knowledge. 2nd: How does this relate to the first graphic? Was there a large drop off between Dec and Jan? 3rd: line plots generally imply continuous data, consider using a different way to compare boroughs. Also, discussion of insights should be left for the next section.

Answer 1: We elaborated on our introduction to include more information about our business problem. For the first graphic, we ultimately changed it to be completely different as we realized this initial one we had was not as insightful. We then changed our line graph into a bar chart.

Comment 2: I think that some of the visualizations could be refined with color and themes a little more. In addition, some of the visualizations' text is small and is hard to see trends in some places. Small refinements should fix this pretty easily. Answer 2: We added color and attempted to make the graphs appear larger.

Comment 3: Many of these insights are not beyond common knowledge. Also, the 5th graphic is mentioned but no real insights are explained. I think a little more elaboration would be useful on the individual patterns, but not so much that it makes it not concise.

Answer 3: We removed our fifth graphic as a result since we ourselves could not determine any further insights.

Comment 4: I believe Insight summary is very clear and not complicated. I like how you provided the summary of everything you believed was important for your future analysis. So

overall I think you did a good job. The coding part was well commented and easy to follow. I do not have any suggestions as I believe this part was very well made.

Answer 4: Since no recommendation was made, we did not make any corresponding changes.

Comment 5: More complete description of Excel process should be included.

Answer 5: Since we removed this visualization, this part was also removed.

Comment 6: I understand the process in R and it is very efficiently done. I wonder if it is completely necessary to show each of the visualizations twice or not. In the last graph (the two-line graph), one of the labels in the key has an underscore, but this is only a small issue. Besides this, I don't see any problems.

Answer 6: We improved the overall appearance of visualizations by adding titles and axis labels.

Comment 7: The first paragraph could be significantly shortened.

Answer 7: We shortened this paragraph to make it more succinct but still informational.

Comment 8: If the y-axis scale is difference in bookings, consider making the axis label more descriptive.

Answer 8: We changed this axis title to "Difference in Bookings".

Comment 9: With so many values on the y-axis tick marks, becomes hard to quickly process the value. Consider reformatting these values.

Answer 9: We reformatted the values.

Comment 10: All visualizations should be captioned.

Answer 10 : We captioned our visualizations

Comment 11: How does knowing that the trend by month in 2020 increases impact your business problem?

Answer 11: It reassures that there may be a factor that is contributing to the increase in reservations.

Comment 12: This section should be non-technical (ie. bar graph, box plots).

Answer 12: We adjusted our visualizations.

Comment 13: It's almost impossible to read the words in the third visualization. The font needs to be larger. A title is needed.

Answer 13: We added a title and attempted to make the visualization appear larger.

Comment 14: There are points that are half on the box plot. Either cut it off at a value that does not slice a point in half or show the entire boxplot.

Answer 14: We changed the limits of the x axis so that none of the points were cut off.

Comment 15: The font used in the fourth visualization is also too small.

Answer 15: We attempted to make the visualization bigger overall.

Comment 16: Boxplots don't show averages unless you specifically add it. You should not draw an insight about a value that's not shown.

Answer 16: We removed our comment about averages and instead commented on other features of the boxplot.

Comment 17: The last visualization should have a title.

Answer 17: This visualization was removed.

Comment 18: How does the last visualization uniquely informs your business problem? Is there more to it than the expected result of decreased demand?

Answer 18: This visualization was ultimately removed.

Comment 19: The insight summary section should be non-technical and more succinct. It should be a summary instead of a detailed explanation. Consider showing potential relationships between number of bookings and price here.

Answer 19: We refined our summary section to make it more succinct and non-technical.

Comment 20: When introducing the second visualization, either make clear that you are moving to the next visualization or change the word first to something like next.

Answer 20 : We added a header that made it clear we were talking about the next visualization.

Comment 21: Note that the visualizations here should not be captioned. Make sure that all explanations are in full sentences. The size of the word visualizations here is much better. The words in this visualization and the one show in the body are different.

Answer 21: We removed captions and made sure the explanations were done in full sentences.

Comment 22: A full line is needed between paragraphs. Some of the reference list items need to also list the publication.

Answer 22: We added full lines in between paragraphs.

6.4 Deliverable 4

Comment 1: You can be a bit more descriptive (provide a bit more context) in description of your goals. I'm not sure I understand your connection between words in listing descriptions being most prevalent and being "most important and effective".

Answer 1: We added more context and information about our business question, because without the other deliverables it was not exactly clear what our context was. We clarified that we are only looking at listing descriptions from successfully rented properties, which a lot of people commented on. Thus, the words that are mentioned the most frequently are representative of features that are probably most desirable to renters.

Comment 2 & 3: Both predictive analytics methods make sense for their project and are explained well. There are no questions left unanswered. I have no recommendations on improvements as this explanation covered all the bases. The descriptions were very well written. I liked how it wasn't too long, but still provided all the information needed. I was also able to understand what the predictive analytics would be used for in terms of answering the business problem.

Answer 2: & 3 We did not make changes as a result of this comment. Insinuated no changes were to be made for improvement.

Comment 4: Calling it a "giant list" isn't necessary (unprofessional).

Answer 4: We removed unprofessional language and adjusted terms such as giant.

Comment 5: The one improvement I would recommend would be to include the significance level that was chosen to assess the linear regression model. Other than that, the description made sense, and left no questions unanswered.

Answer 5: We ended up removing the linear regression model, as it did not fit the assumptions well enough to draw meaningful conclusions. Thus, this comment is void.

Comment 6 & 7: Based upon the visualizations provided, linear regression does not seem like the best method to analyze this data as many of the data is discrete aka. non-continuous. This makes analyzing it accurately difficult. Also, consider the other assumptions of linear regression.

Answer 6 & 7: From this extremely helpful feedback we decided to remove our linear regression model and instead do a decision tree, which didn't rely on the same assumptions that are required of linear regression. We were trying to force our data to fit a linear regression model, when it was not correctly suited for our data.

Comment 8: Improve visualizations by adding captions or titles to the graphs. This will provide more context as to what the graph is trying to display.

Answer 8: We added captions and titles to all our graphs, in order to provide more context and improve readability and professionalism.

Comment 9: We mentioned that, "Most travelers are often business-related and are traveling alone and for short times." It was confusing because people weren't sure if this was supported by our data or conjecture.

Answer 9: We removed this comment because it was unnecessary part of our analysis and caused confusion rather than adding anything meaningful.

Comment 10: It's certainly possible that certain words are negated, which is lost with single-word analysis.

Answer 10: To solve this problem, we added an additional bigram analysis to see if negated words were mentioned and so that we could get another look at the texts in order to pick up on trends we might not have seen in the unigram model.

Comment 11: Suggested moving some of the information into other deliverables.

Answer 11: We moved our response variable creation discussion into the data section and simplified this section to include less about the overall business question which is mentioned more thoroughly in earlier sections.

Comment 12: Since the majority of 2020 was during the pandemic, there may be more similarities in the reviews over these two years, unless you're looking at Jan-Mar for both 2020 and 2021.

Answer 12: We took this into account using term document frequency which was a comment mentioned earlier. Should take care of reoccurring words in 2020 then 2021.

Comment 13: You don't need to teach the reader what the terms like confusion matrix mean. This section shouldn't mention R or its packages. Consider rounding the values in table 2. 4.

Answer 13: We removed the parts where we go into an extensive explanation of what a confusion matrix is, however kept a succinct explanation because in a professional setting many will be unfamiliar with this term and concept. Also removed mention of R packages, as it would be unprofessional and possibly confusing to the reader. Rounded the values in the table.

Comment 14: Consider using term and document frequency instead of bag of words to filter out those commonly used words.

Answer 14: We used term and inverse document frequency in order to assess which additional stop words to exclude for our analysis. Also added a brief explanation of what exactly this means for our analysis.

Comment 15: It's hard to compare the words using the graphics. Consider putting this information into a table with a column for 2020 and one for 2021 to more easily compare the words.

Answer 15: We reformatted the way we displayed this so that instead of being side by side bar graphs it is instead a table that has the top ten words and columns for 2020 and 2021. Improved readability and ability to draw insights.

Comment 16: The original and pruned trees aren't different because there was nothing to prune. Did you consider random forests?

Answer 16: We changed our decision tree to a random forest model.