ANALYSIS OF AIRBNB DATA

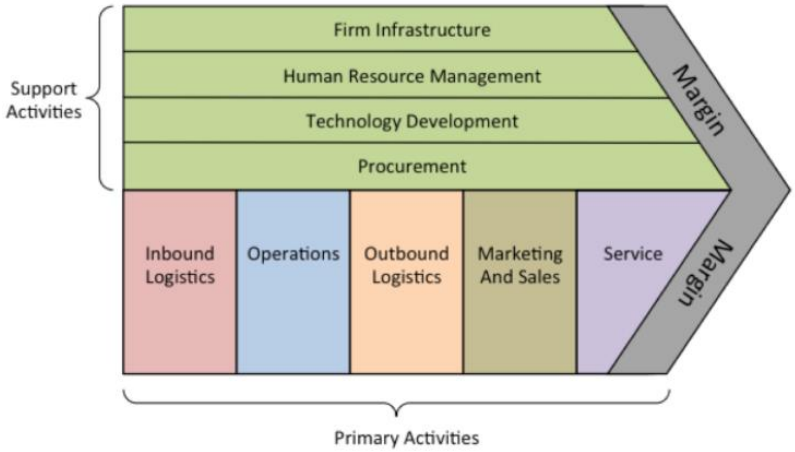
-By Debolina Sasmal

Introduction:

Airbnb is an American vacation rental online marketplace company that allows house owners to rent out their properties to guests looking for a place to stay. In this report, I have used a dataset that includes the data on various Airbnb listings in NYC to understand how different attributes such as location, house type, availability etc. can be used to predict the price of a new listing as well as to predict what kind of listings get the most reviews. This data analysis can help new renters decide on aspects that will give them the optimum profit. The Value chain analysis and Sentiment analysis helps us understand the Airbnb business activities and how users are reacting to Airbnb's efforts to keep the accommodations sanitized during this current pandemic.

Airbnb Value Chain Analysis:

The value chain analysis of Airbnb can help identify business activities that can create a competitive advantage for Airbnb as well as for the renters. The online platform business model is the actual source of value creation for Airbnb because it cuts down the cost to a great extent due to the absence of logistics.



Primary activities:

- Inbound Logistics – Inbound logistics involve receiving and storing of raw materials, but since the Airbnb platform operates completely online, addition of new renters comes at no extra cost. The inbound logistics that takes place online consists of listing of accommodations and experiences.
- Operations – Operations involve the transformation of raw materials into finished products. For Airbnb, it involves activities such as handling of keys, cleaning, sanitizing, and providing amenities (towels, soap, shampoo etc.) within properties., but these have to be done by the renters. So, it is a complete profit for Airbnb but the renters have to bear this minimal initial cost.
- Outbound Logistics – Outbound logistics involves warehousing and distribution if the finished products. Airbnb doesn't have to do anything in this.
- Marketing and Sales – Marketing and sales activities include application of marketing strategy. For Airbnb, marketing promotions includes brand partnerships, social media, celebrity endorsements, word-of-mouth marketing, TV ads etc.

- Service – Service involves customer care support. Excellent customer service is a good competitive advantage for Airbnb. Their customer support services include agents and support team who are available through mail, phone and chat to answer customer queries and resolve their complains. This service is available 24/7 and in various languages since Airbnb is available in many countries. Airbnb goes to extreme lengths to please their customers such as the following instances mentioned on Business Insider (Jim Edwards, Aug 14 2013, "Check Out The Extreme Lengths Airbnb Will Go To In Order To Please Customers", https://www.businessinsider.com/insane-lengths-airbnb-will-go-to-in-order-to-please-customers-2013-8)
    - When an Airbnb host cancelled the cottage Shay Fan had booked for her bridesmaids on the eve her wedding, Airbnb pitched in a $1,000 credit to help her find a new place to stay.
    - Adam Jackson discovered that one of his guests had stolen a pizza cutter from his kitchen, so he tweeted about it and Airbnb overnighted a new one to him with this cute apology note.
    - Jessica Dowdy tweeted that she had booked with Airbnb in Seattle to see a Pearl Jam gig, and the company sent her a T-shirt.

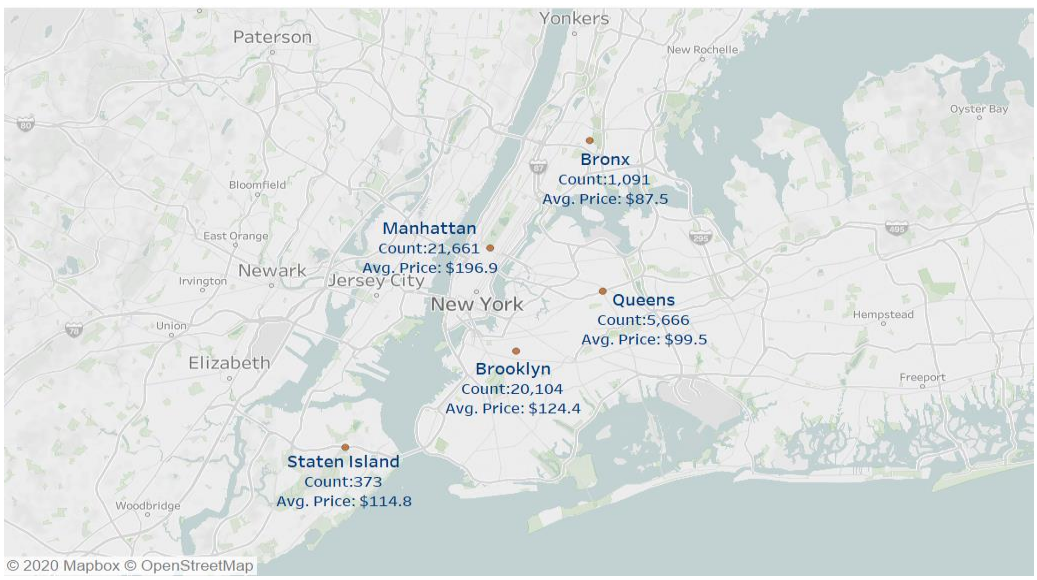
Support Activities:

- Firm infrastructure – Invest in real estates that hosts offer to the customers.
- Human resource management – Employ high skilled professionals in IT, engineering, and management areas. Conduct skill development programs for its employees so that they are up to date with the current technologies since Airbnb uses the online business platform.
- Technology Development – Research and Development, Update to the latest technology for marketing and analysis purposes since technology and analysis is the most important part for the companies dependent on online business platforms. Optimization of machine learning to predict better ways to run the business.
- Procurement – Procurement of renters who are offering to put up their houses for rent. For the renters, procurement would be buying real estates.

Exploring the Dataset:

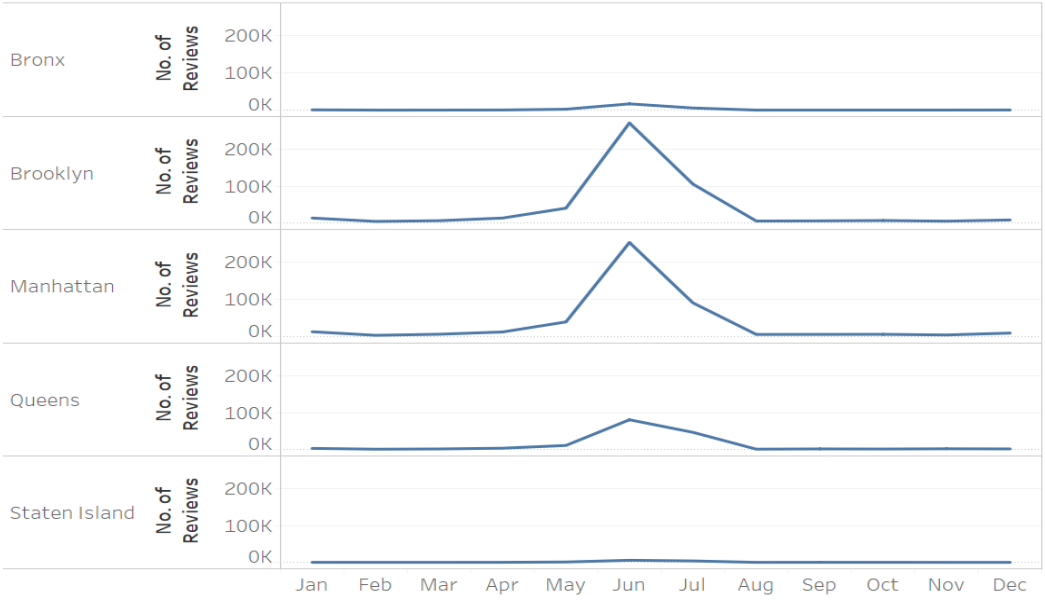
Dataset used: New York City Airbnb Open Data, https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data

Analysis of each Borough:



According to the dataset, there are five neighborhood groups, that is, Boroughs in New York viz. Brooklyn, Manhattan, Queens, Bronx, and Staten Island. The above figure shows a Map view of the same. Using Tableau, I have depicted the Average price and the number of operating Airbnb in each borough. As expected, the prime locations Manhattan and Brooklyn have the most listings and average price.

Monthly Analysis of Airbnb usage:



The dataset contains the number of reviews each Airbnb host has received from the year 2011 to 2019. I have used Tableau to create a visualization that shows how many reviews the Airbnb in each Borough gets every month in a year. The "number of reviews" helps us guess indirectly how busy the Airbnb have been, because more occupancy means more reviews. From the above illustration, we can conclude that the occupancy peaks during the summers, from May to August, probably because of the summer break. Also, the temperature is pleasant during these months if one wants to visit New York. (Santorini Dave, October 26 2019, "The Best Time to Visit New York City", https://santorinidave.com/best-time-to-visit-nyc#:~:text=The%20best%20time%20to%20visit%20New%20York%20City%20is%20from,in%20New%20York%20is%20September.)

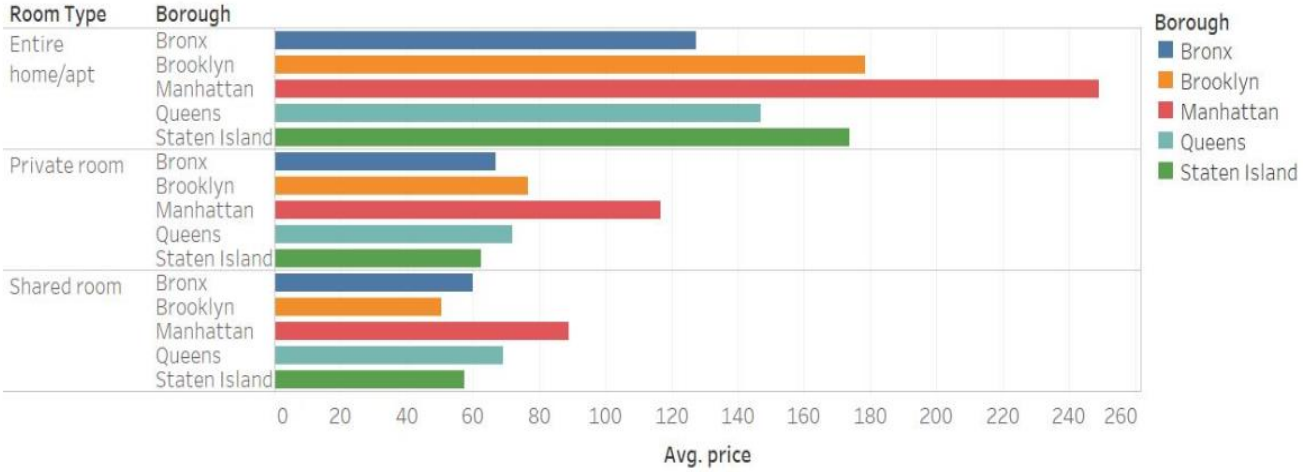
Hosts with the most number of listings:

```
# A tibble: 11,452 x 2
   vars              n
   <chr>         <int>
 1 Michael         417
 2 David           403
 3 Sonder (NYC)    327
 4 John            294
 5 Alex            279
 6 Blueground      232
 7 Sarah           227
 8 Daniel          226
 9 Jessica         205
10 Maria           204
```

```
count1 <- count(NYC_airbnb, vars = host_name)
count1 %>% arrange(desc(n))
view(count1)
```

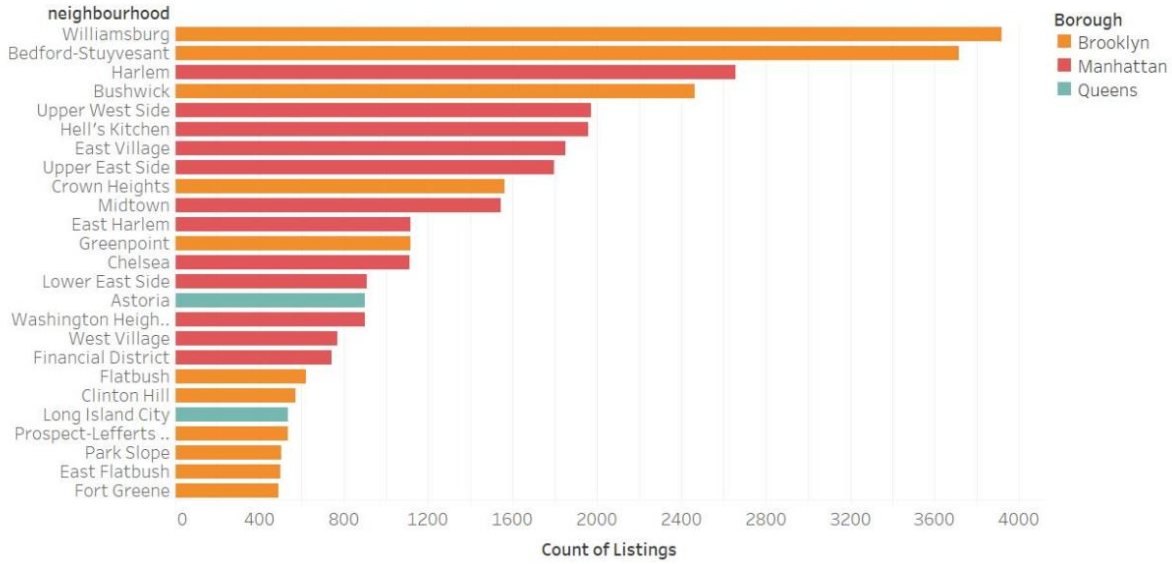
The above table depicts the top 10 hosts with the highest number of listings. It is interesting to see that a host can have multiple listings. From this, we can conclude that it is a good business for those who have a lot of holdings in the real-estate.

Room Type Analysis:



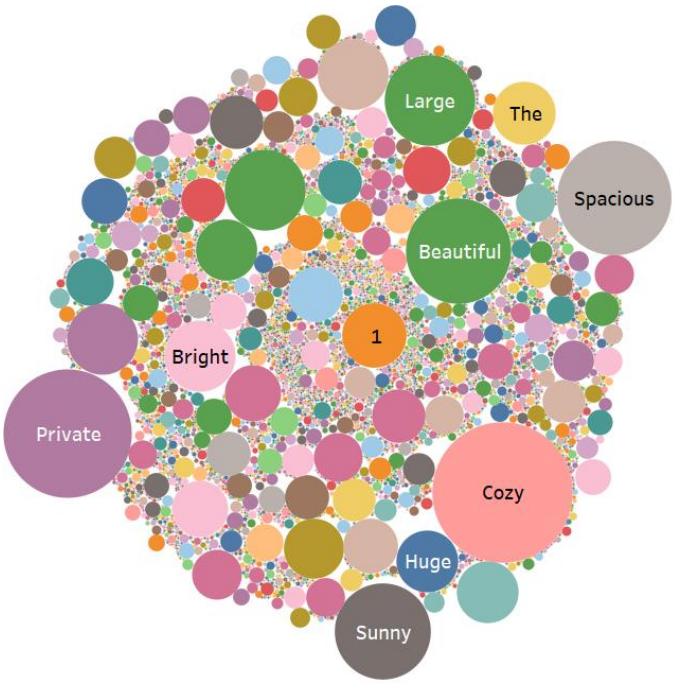
From the above illustration, we can see that people prefer to rent an entire home/apartment and shared rooms are the least preferred ones. The demand for entire room/apartment is the most in Manhattan.

Analysis of Neighborhoods in each Borough:



The above illustration shows the top 25 neighborhoods with the most listings. We can see that most of the neighborhoods are from either Brooklyn or Manhattan. Bronx and Staten Island are not even a part of the list.

Analyzing the most used words for listing:



As per the above wordcloud, the most used words by the hosts while listing their houses on Airbnb are words like Private, Cozy, Huge, Sunny, Bright, Beautiful, Spacious, Large. Simple but mind-catching words are used by the hosts that can aptly describe the houses and grab attention as well.

Regression Analysis for Rating and Price Predictions:

I have used Regression Analysis and created two models. One model predicts the price of the listings and the other model predicts the rating of the listings depending upon various factors.

The dataset has been cleansed and been divided into 70% training set and 30% test set as following:

```
airbnb_train <- NYC_airbnb %>% sample_frac(.7) %>% filter(price > 0)
airbnb_test <- NYC_airbnb %>% sample_frac(.3) %>% filter(price > 0)
```

Initially, I was a R-squared value of 0.03-0.04 for my models, so I removed the outliners present in price:

```
Price_out <- airbnb_train %>% filter(price < quantile(airbnb_train$price, 0.9)
                                     & price > quantile(airbnb_train$price, 0.1))
```

<u>Price Prediction:</u>

```
model1 <- lm(price ~ room_type + neighbourhood_group + latitude + longitude
                   + number_of_reviews + availability_365
                   + reviews_per_month + calculated_host_listings_count +
                minimum_nights, data = Price_out)
summary(model1)
```

Output:

```
Coefficients:
                                  Estimate Std. Error  t value Pr(>|t|)
(Intercept)                     -1.512e+04  7.531e+02  -20.077  < 2e-16 ***
room_typePrivate room           -6.130e+01  4.970e-01 -123.329  < 2e-16 ***
room_typeShared room            -7.396e+01  2.272e+00  -32.554  < 2e-16 ***
neighbourhood_groupBrooklyn     -7.168e+00  2.167e+00   -3.308 0.000942 ***
neighbourhood_groupManhattan     1.630e+01  1.990e+00    8.190 2.71e-16 ***
neighbourhood_groupQueens        4.308e+00  2.095e+00    2.056 0.039790 *
neighbourhood_groupStaten Island -7.240e+01  3.988e+00  -18.156  < 2e-16 ***
latitude                        -7.210e+01  7.309e+00   -9.864  < 2e-16 ***
longitude                       -2.461e+02  8.470e+00  -29.055  < 2e-16 ***
availability_365                 3.764e-02  2.008e-03   18.747  < 2e-16 ***
reviews_per_month               -4.256e-01  1.490e-01   -2.856 0.004290 **
calculated_host_listings_count   1.212e-01  9.907e-03   12.231  < 2e-16 ***
minimum_nights                  -1.750e-01  1.334e-02  -13.117  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 40.63 on 28362 degrees of freedom
Multiple R-squared:  0.4445,    Adjusted R-squared:  0.4442
F-statistic:  1891 on 12 and 28362 DF,  p-value: < 2.2e-16
```

Explanation:

The p-value of this model is 2.2×10^{-16}, which is quite low for any value of α. So, this model is significant at any standard level of significance. Also, the Multiple R-squared value is 0.4445, that is, 44.45% of the variation in the price can be predicted correctly using this model.

Rating Prediction:

```
model2 <- lm(number_of_reviews ~ room_type  +
                    + availability_365 + calculated_host_listings_count +
                    latitude + longitude + reviews_per_month +
               neighbourhood_group + price +
               minimum_nights, data = Price_out)
summary(model2)
```

Output:

```
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                     -7.132e+03  7.126e+02 -10.008  < 2e-16 ***
room_typePrivate room           -2.862e+00  5.790e-01  -4.943 7.72e-07 ***
room_typeShared room            -1.472e+01  2.178e+00  -6.761 1.39e-11 ***
availability_365                 5.369e-02  1.901e-03  28.248  < 2e-16 ***
calculated_host_listings_count  -1.535e-01  9.331e-03 -16.447  < 2e-16 ***
latitude                         4.481e+01  6.885e+00   6.508 7.71e-11 ***
longitude                       -7.173e+01  8.073e+00  -8.885  < 2e-16 ***
reviews_per_month                1.571e+01  1.407e-01 111.663  < 2e-16 ***
neighbourhood_groupBrooklyn      1.312e+01  2.046e+00   6.411 1.46e-10 ***
neighbourhood_groupManhattan     8.055e+00  1.883e+00   4.278 1.89e-05 ***
neighbourhood_groupQueens        7.994e+00  1.981e+00   4.036 5.46e-05 ***
neighbourhood_groupStaten Island -5.137e+00  3.797e+00  -1.353    0.176
price                           -2.188e-02  5.586e-03  -3.917 8.99e-05 ***
minimum_nights                  -5.520e-02  1.270e-02  -4.346 1.39e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 40.24 on 31513 degrees of freedom
Multiple R-squared:  0.3282,    Adjusted R-squared:  0.328
F-statistic:  1184 on 13 and 31513 DF,  p-value: < 2.2e-16
```

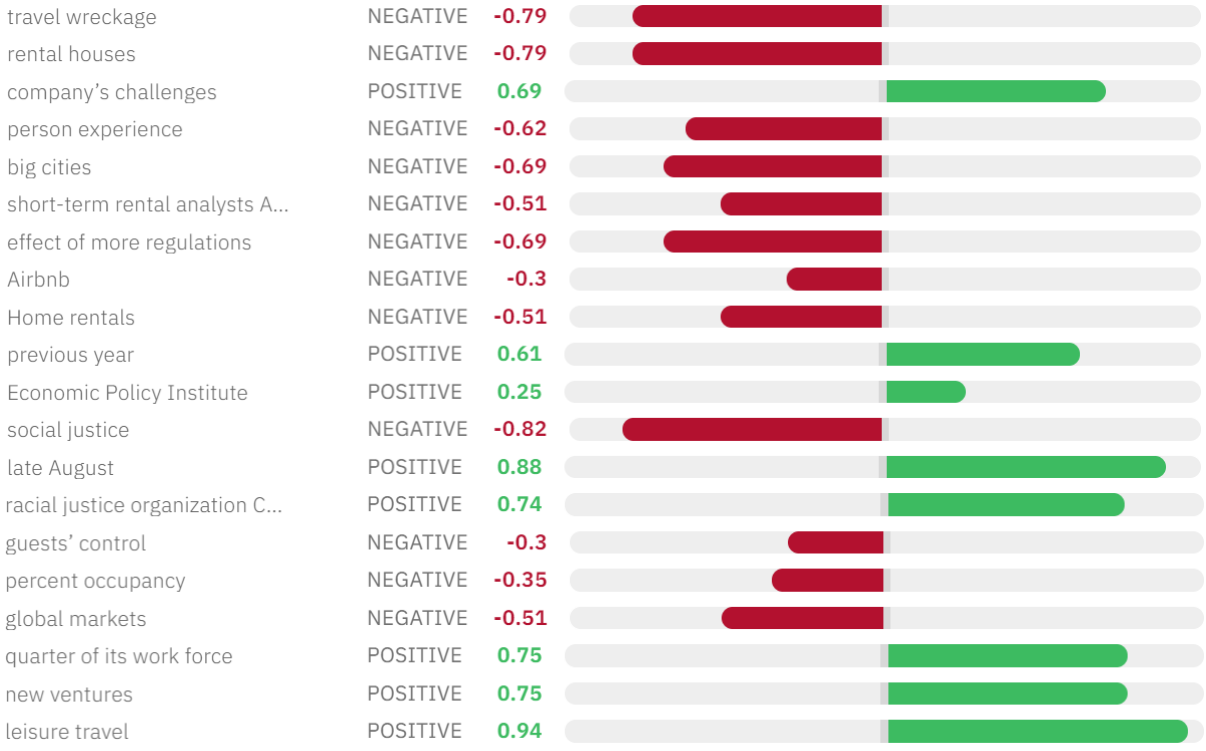
Explanation:

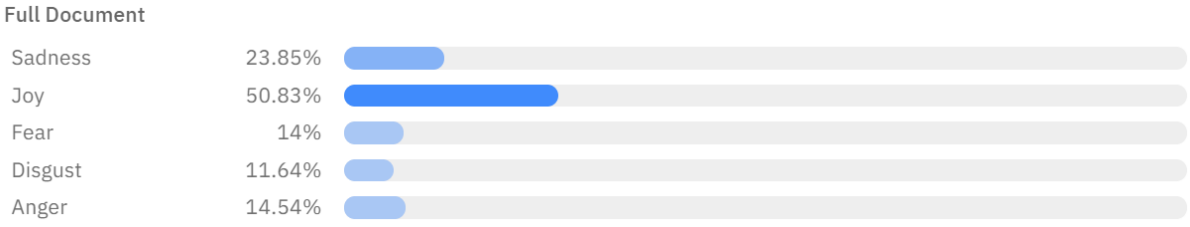
The p-value of this model is 2.2×10^{-16}, which is quite low for any value of α. So, this model is significant at any standard level of significance. Also, the Multiple R-squared value is 0.3282, that is, 32.82% of the variation in the price can be predicted correctly using this model.

Current performance of Airbnb during the pandemic:

Article used: Elaine Glusac, Nov. 16, 2020, "The Future of Airbnb", https://www.nytimes.com/2020/09/24/travel/airbnb-pandemic.html

To assess the current performance of Airbnb, that is to check whether it has taken a blow due to the current covid-19 pandemic situation, I have performed a sentiment analysis on an article called "The Future of Airbnb"





As mentioned in the article, "Home-sharing's challenges aren't only about social distancing and hygiene. Overtourism, racial bias, fee transparency and controlling the party crowd are also in the mix.". According to the sentiment analysis of the article, the most positive keyword is "leisure travel" and the most negative keyword is "social justice". The positivity of the leisure travel is self-explanatory. Social justice has been marked negative because of its usage pertaining to the George Floyd incident this summer.

The emotion of the overall article is "Joy". That is because people are still using Airbnb, and even more than before. As mentioned in the article, "Now the pandemic is changing how people want to work, travel and live. Remote school and work unbind families from their

homes. People are living differently and people want to live anywhere. Cities have taken the blow of the pandemic, but the less crowded districts haven't. People are booking Airbnb at such isolated places as a getaway because this monotonous life that has resulted from covid.

Overall Conclusion:

From the value chain analysis, dataset analysis, and sentiment analysis, it can be concluded that for those, who are owners of the real estate, or even for the people who are thinking about renting a part of their house, Airbnb is a profitable option. The initial cost of set up is quite low because of the absence of inbound/outbound logistics. There is a slight difference in the conclusion drawn from the dataset and from the sentiment analysis. According to the dataset, which is a pre-pandemic one, it is easier to rent out a place in the big boroughs such as Brooklyn and Manhattan, whereas according to the sentiment analysis, people are looking forward to book Airbnb in the less populated areas now (preferably not cities).

References:

Jim Edwards, Aug 14 2013, “Check Out The Extreme Lengths Airbnb Will Go To In Order To Please Customers”, https://www.businessinsider.com/insane-lengths-airbnb-will-go-to-in-order-to-please-customers-2013-8

New York City Airbnb Open Data, https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data

Santorini Dave, October 26 2019, “The Best Time to Visit New York City”, https://santorinidave.com/best-time-to-visit-nyc#:~:text=The%20best%20time%20to%20visit%20New%20York%20City%20is%20from,in%20New%20York%20is%20September

Elaine Glusac, Nov. 16, 2020, “The Future of Airbnb”, https://www.nytimes.com/2020/09/24/travel/airbnb-pandemic.html

Stackoverflow, https://stackoverflow.com/

John Dudovskiy, September 16, 2019, “Airbnb Value Chain Analysis”, https://research-methodology.net/airbnb-value-chain-analysis/

Samar Singla, “How Airbnb Works | Insights into Business & Revenue Model”, https://jungleworks.com/airbnb-business-model-revenue-insights/#:~:text=Booking%20is%20made%20through%20Airbnb,host%20after%20deducting%20their%20commission.

Wanderlust, Nov 13 2018, “Airbnb: Utilizing Machine Learning to Optimize Travel”, https://digital.hbs.edu/platform-rctom/submission/airbnb-utilizing-machine-learning-to-optimize-travel/