Airbnb Statistical Analysis

Table of Contents

Introduction and Motivation

Vacation homestay rentals have increased dramatically over the past few years, accommodating millions of guests every night (Wyman, Mothorpe and McLeod, 2020). The biggest player in the shared accommodations market is Airbnb®, with over 7 million listings worldwide (Dogru et al. 2021).

Vacation homestay rentals are here to stay. The fact that this is such a new service means research on it is lacking in comparison to hotels, resorts, etc. There have been analyses on the effects of COVID-19 on Airbnb (Gyódi, 2021), as well as price analyses (Zhang et al. 2017). These studies have not investigated all the variables we hypothesize will have an impact. Our goal is to fill a gap in the data and analyze price and COVID-19 under a new lens.

The insights gained from an analysis like this are invaluable. With vacation homestays being such a large market, gaining a deeper understanding of the inner working of the industry is important, especially from *all points of* view. Hosts, guests, and Airbnb would all benefit from at least one analysis in the following study.

Dataset

The dataset was sourced from Kaggle (Bhat, 2021), and was originally pulled through Airbnb's API (Airbnb, 2022). We have permission to use the dataset through CC0 1.0 public domain dedication (CC0 1.0, 2022), as it is open public domain data with no copyright. We will be looking at two data tables in total:

- One contains information 250,000+ Airbnb listings across 10 major cities. This table has 33 columns
- The other contains 5 million+ reviews on these listings from 2008-2021. This table has 4 columns

Both data tables are structured and tabular, containing different data types including:

- Categorical (e.g., property type, room type)
- Numerical (e.g., number of bedrooms, price, review score)
- Date (e.g., review date, host since data)
- Boolean (e.g., host is superhost, host identity verified)

The "listing ID" column is common to both tables and was used to merge them. In terms of the types of columns, w

- Listing information
  - o Listing ID, city, district, neighborhood, latitude, longitude, property type, room type, accommodates, bedrooms, amenities, names, price, min. nights, max. nights, instant bookable

- Host information
    - Host ID, host since, host location, host response time, host response rate, host acceptance rate, host is superhost, host total listings, host has profile picture, host is verified
- Review information
    - Review score, accuracy score, cleanliness score, check-in score, communication score, location , score, value score, review ID, review date, reviewer ID

Guiding Questions

GQ 1: Which metrics have the largest influence on listing price?
With this guiding question, we are attempting to investigate what columns in our data frame have the largest influence on price. On a survey we took of our DATA 602 class, we found that location was the principal factor people thought impacted price (Ahmad, Rasheed, and Yip, 2022). We tested this, among other metrics. This is a critical guiding question, as it helps us gain insight into one of the most crucial factors for hosts, guests, and Airbnb itself. Price is correlated with revenue among other things, and so an investigation of it is crucial.

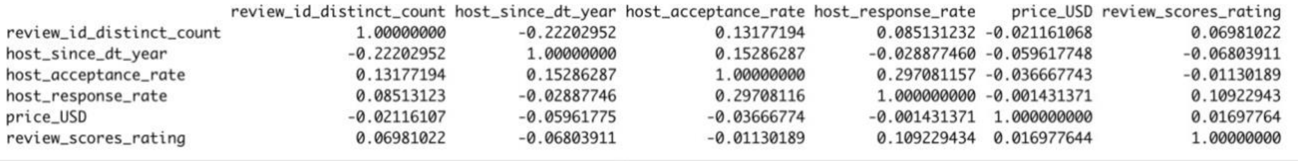
GQ 2: What impact has COVID-19 had on Airbnb?
This guiding question has evolved a little since our proposal. Instead of just looking at host information, we have decided to investigate the impact COVID-19 has had on Airbnb in general. This is extremely important as Airbnb has been hit quite hard by COVID-19 (Jang and Kim, 2022). Insights from this can be extrapolated to prepare for disasters and emergencies in the future.

Analysis

Guiding Question 1

Is there any strong correlation between available variables and listing price?

We correlated multiple available data variables with the price of listings and found the below results.



```
                        review_id_distinct_count host_since_dt_year host_acceptance_rate host_response_rate    price_USD review_scores_rating
review_id_distinct_count              1.00000000        -0.22202952           0.13177194        0.085131232 -0.021161068           0.06981022
host_since_dt_year                   -0.22202952         1.00000000           0.15286287       -0.028877460 -0.059617748          -0.06803911
host_acceptance_rate                  0.13177194         0.15286287           1.00000000        0.297081157 -0.036667743          -0.01130189
host_response_rate                    0.08513123        -0.02887746           0.29708116        1.000000000 -0.001431371           0.10922943
price_USD                            -0.02116107        -0.05961775          -0.03666774       -0.001431371  1.000000000           0.01697764
review_scores_rating                  0.06981022        -0.06803911          -0.01130189        0.109229434  0.016977644           1.00000000
```

Our analysis revealed existence of weak positive and negative correlations between listing price and other available variables. It is important to note that our analysis is limited, as we do not yet have methods to calculate correlation between more than 2 variables. It may not be specific metrics that influence price, but a *combination* of metrics.

Does host identity verification have an influence on listing price?

A recent study looking into trust perception in Airbnb found that people were more likely to book if the host was verified (Zhang, Yan, and Zhang, 2020). Does this also mean that hosts who were verified charge more than regular hosts? This is what we aimed to answer in our next analysis of price.

Recall that the two-sample t-distribution can be expressed as a confidence interval of the difference between two means as:

$$(\overline{X}_1 - \overline{X}_2) \pm t_{1-\frac{\alpha}{2},df}\sqrt{\frac{S_1^2}{n_1} - \frac{S_2^2}{n_2}}$$

$$df = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1}\left(\frac{S_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{S_2^2}{n_2}\right)^2}$$

Condition Checking for t-test

Before we can begin with our t-test, we must state our statistical hypotheses and check conditions. Our statistical hypotheses are stated below:

$$H_0: \mu_{verified\ host} \leq \mu_{non-verified\ host}$$

$$H_A: \mu_{verified\ host} > \mu_{non-verified\ host}$$

Our null hypothesis states that the mean price is the same between verified hosts and non-verified hosts. Now, for condition checking. We must check that both verified hosts and non-verified hosts have normally distributed prices. The normality plots can be seen below:

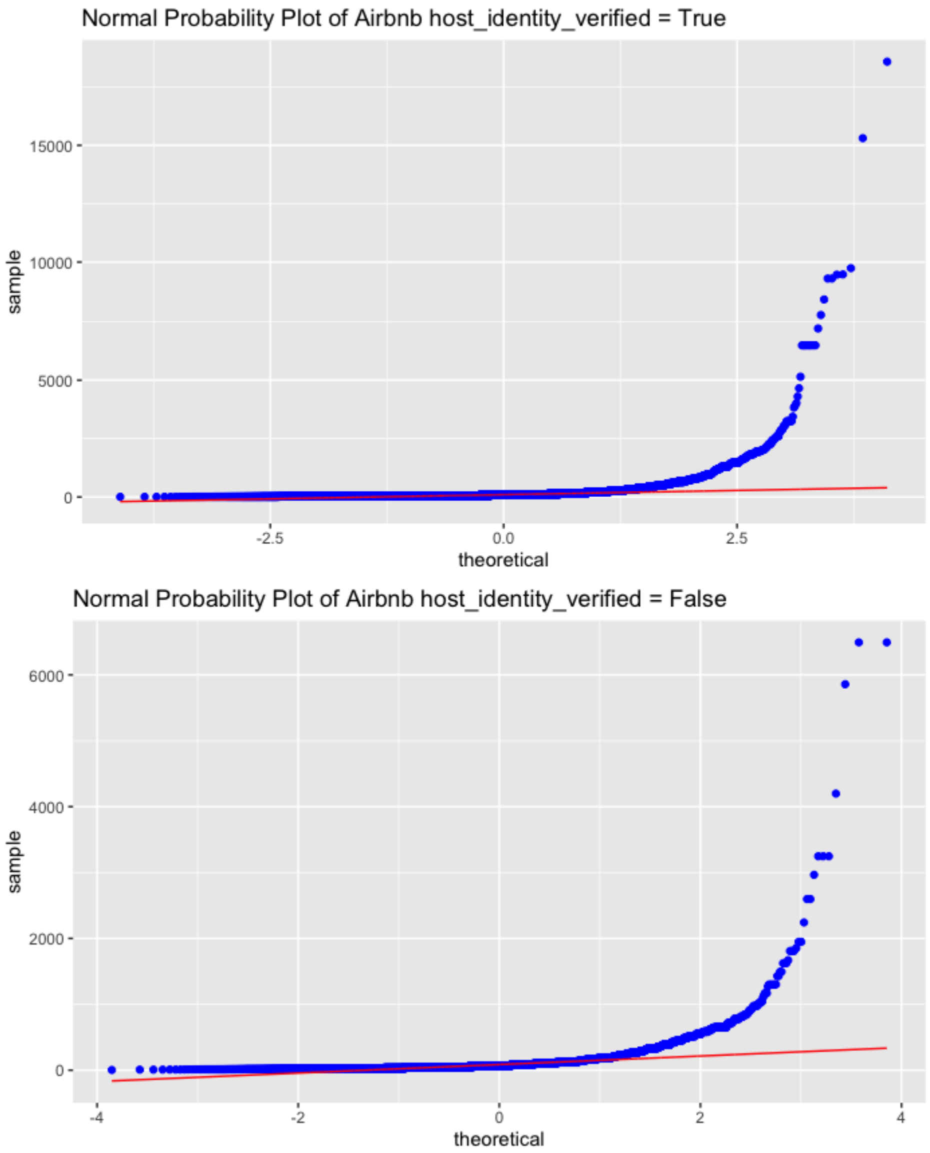
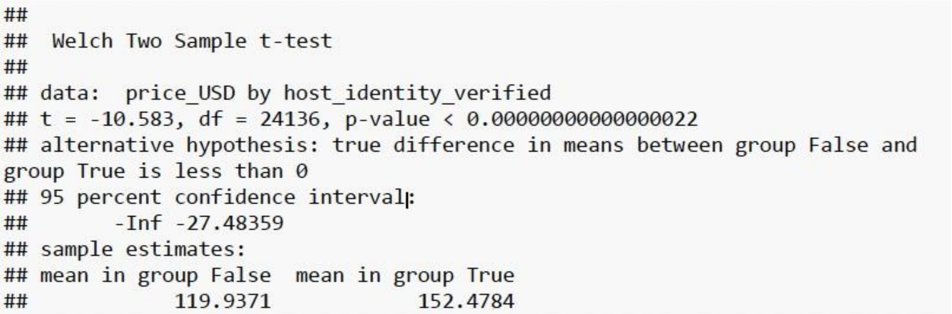


Figure 1a and 1b: Normality plots for prices for verified hosts Vs. non-verified hosts, respectively.

Data Analysis and Visualization

Both normality plots indicate that the data is normal, so we can continue with our analysis. In performing our Welch two-sample t-test, we obtain the following results:



```
##
##  Welch Two Sample t-test
##
## data:  price_USD by host_identity_verified
## t = -10.583, df = 24136, p-value < 0.00000000000000022
## alternative hypothesis: true difference in means between group False and
group True is less than 0
## 95 percent confidence interval:
##       -Inf -27.48359
## sample estimates:
## mean in group False  mean in group True
##            119.9371            152.4784
```

$$t = -10.583$$

$$p - value = 2.2 * 10^{-16}$$

$$27 \le \mu_{verified\ host} - \mu_{non-verified\ host} \le \infty$$

Inference

We get an extremely low p-value, indicating with high confidence that we can reject the null hypothesis, and say that there is a true difference in the mean price of verified hosts and unverified hosts. Our 95% confidence interval also does not contain any mean differences of zero, further backing our conclusion.

This shows a deeper insight. Previous studies have shown higher revenue from verified hosts, indicating that more people choose these hosts (Zhang, Yan, and Zhang, 2020). However, our findings show that these hosts charge more as well, which would also contribute to higher revenue. Therefore, it is always important to approach a statistical analysis from all angles before concluding.

We have also included a boxplot to further visualize this point:

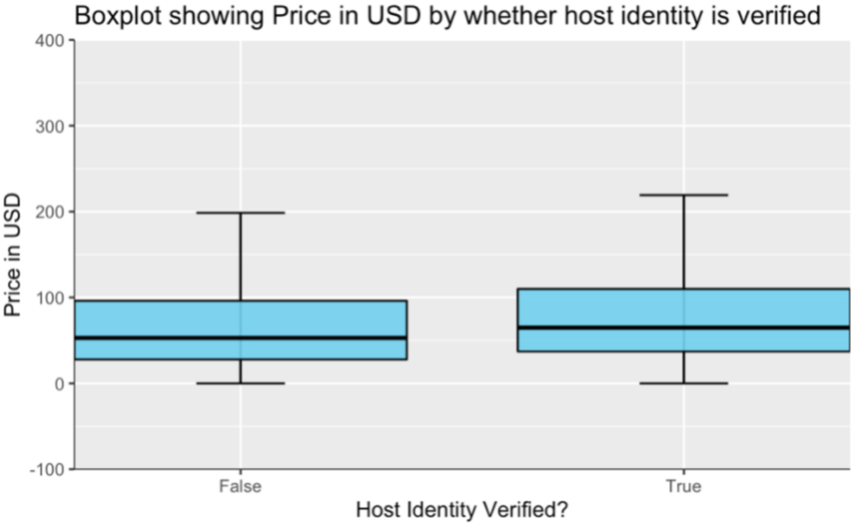


Figure 1c: Boxplot of price for hosts vs superhosts

We see that the median price for listings is slightly higher if the host identity is verified.

Does number of amenities offered have an influence on listing price?

Let us shift now from exploring price from the angle of verified hosts to amenities. We found variation between cities in terms of the average number of amenities offered, so for this analysis we only looked at Sydney, Australia.

Data Analysis and Visualization

The specific aspect of amenities we are analyzing is: Would the proportion of Airbnbs priced over $200 with over 7 amenities be greater than the proportion of Airbnbs prices *lower than* $200 with over 7 amenities? We decided to use a proportion test, below are our statistical hypotheses:

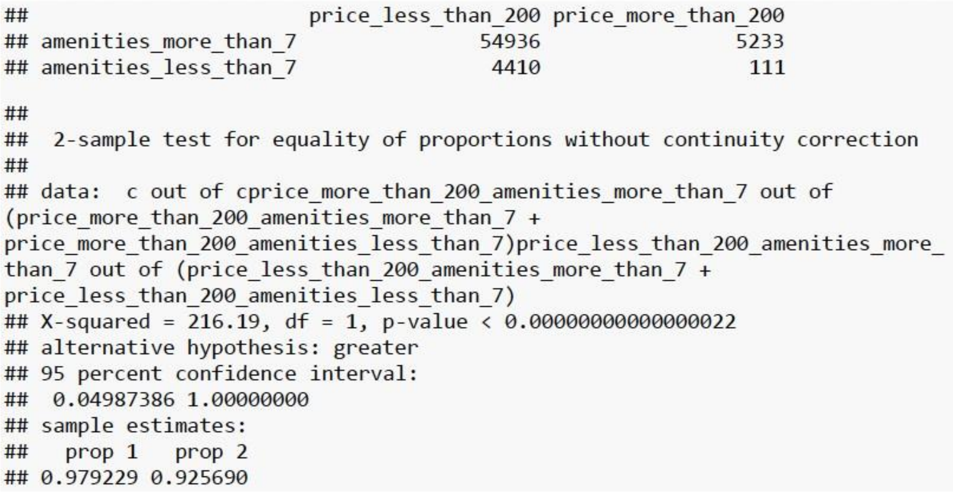
$$H_0: p(> 7amenities)_{price>200} \leq p(> \ 7amenities)_{price\leq 200}$$

$$H_A: p(> 7amenities)_{price>200} \ > p(> 7amenities)_{price\leq 200}$$

Our null hypothesis states that the proportions will not be different between the two price groups. We obtained the following table of proportions:

	Price > 200	Price ≤ 200
Amenities > 7	54936	5233
Amenities ≤ **7**	4410	111

Performing our proportion test, we obtain the following results:



```
##                       price_less_than_200 price_more_than_200
## amenities_more_than_7               54936                5233
## amenities_less_than_7                4410                 111

##
##  2-sample test for equality of proportions without continuity correction
##
## data:  c out of cprice_more_than_200_amenities_more_than_7 out of
(price_more_than_200_amenities_more_than_7 +
price_more_than_200_amenities_less_than_7)price_less_than_200_amenities_more_
than_7 out of (price_less_than_200_amenities_more_than_7 +
price_less_than_200_amenities_less_than_7)
## X-squared = 216.19, df = 1, p-value < 0.00000000000000022
## alternative hypothesis: greater
## 95 percent confidence interval:
##  0.04987386 1.00000000
## sample estimates:
##   prop 1   prop 2
## 0.979229 0.925690
```

$$p = 2 * 10^{-16}$$

$$0.04987 \leq p_1 - p_2 \leq 1.000$$

Inference

We see an exceedingly small p-value, and a confidence interval that does not include 0. Thus, we can reject our null hypothesis and say that there is a mean difference between the two proportions. This is what we would expect, as previous research has shown that hosts that offer higher quality amenities tend to charge more (Wang and Jeong, 2018). We took this a step further, showing that not only is this true for higher quality amenities, but also a higher number of amenities.

Does instant bookability have an influence on listing price?

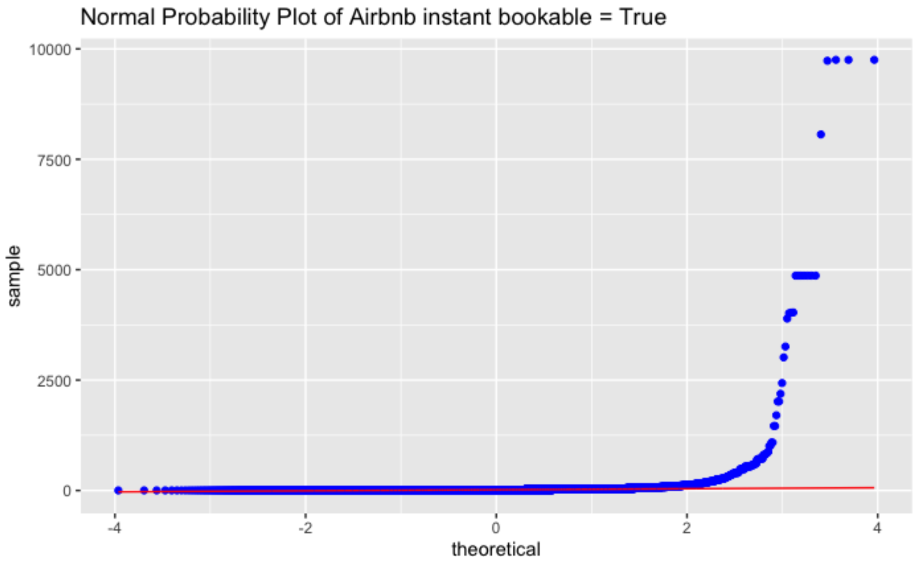
Let us move to another metric, whether an Airbnb is instantly bookable. There is a gap in the research, as this has not been previously compared to price. We wanted to compare prices between instantly bookable airbnbs vs non instantly bookable. We decided to use a two-sample t-test. Our statistical hypotheses are stated below:

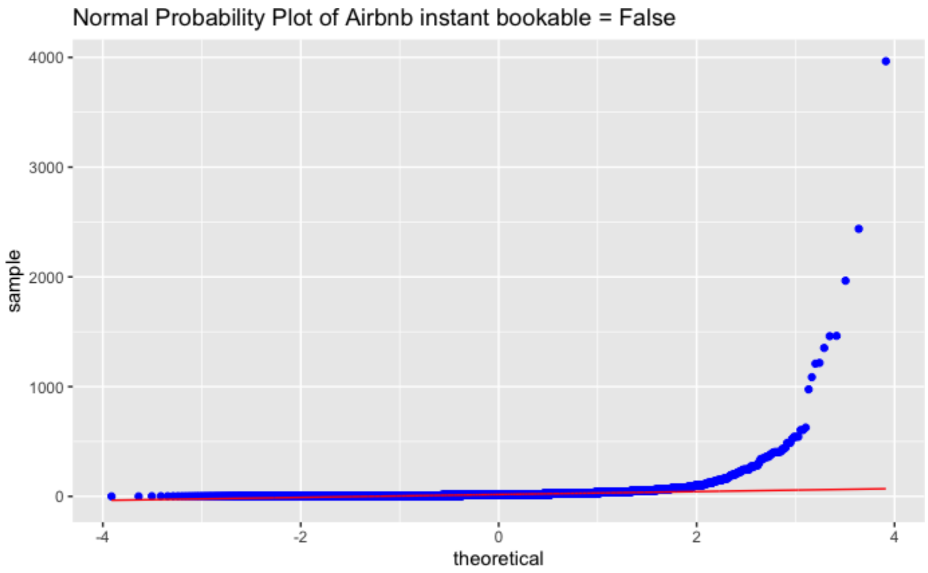
$$H_0\text{: } \mu_{instantbookable} = \mu_{non-instantbookable}$$

$$H_A\text{: } \mu_{instantbookable} \neq \mu_{non-instantbookable}$$

Condition Checking for t-test

To perform a t-test, we checked the condition of normality below:

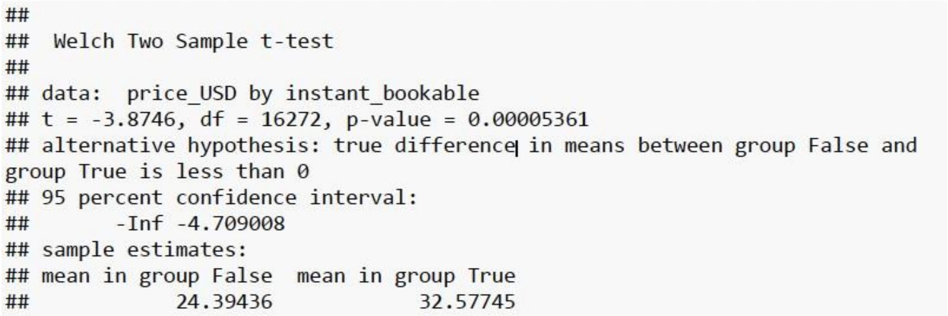




Figures 1d and 1e: Normal probability plots of prices for instant bookable and non-instant bookable hosts, respectively.

Data Analysis and Visualization

Based on these plots, we can accept the condition of normality, and continue with our test. Based on our Welch two-samples t-test, we obtained the following results:



```
##
##  Welch Two Sample t-test
##
## data:  price_USD by instant_bookable
## t = -3.8746, df = 16272, p-value = 0.00005361
## alternative hypothesis: true difference in means between group False and
group True is less than 0
## 95 percent confidence interval:
##       -Inf -4.709008
## sample estimates:
## mean in group False  mean in group True
##            24.39436            32.57745
```

$$t = -3.8746$$

$$p - value = 0.00005361$$

$$4.71 \leq \mu_1 - \mu_2 \leq \infty$$

Inference

We obtain a low p-value and a confidence interval that does not include a mean difference of zero. This means we can reject our null hypothesis and conclude that there is a mean difference in price between instant bookable vs non-instant bookable hosts.

This is an interesting insight. For the past few analyses, we can infer that if hosts feel like they offer more features (more amenities instant bookable, etc.), they can charge more.

Some population price visualizations

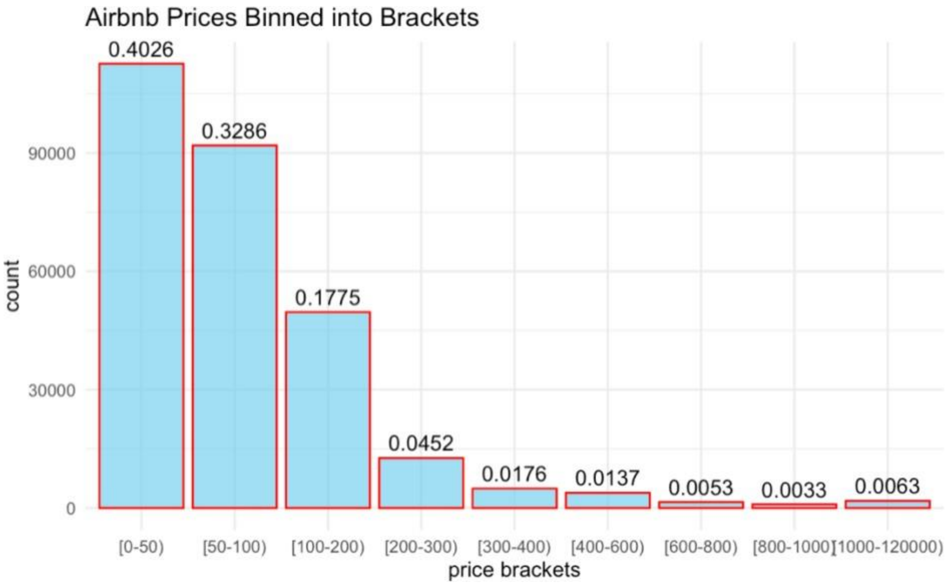


Figure 1f: Bar chart of Airbnb Prices binned into price brackets

Overall, we notice that the one of the biggest influences on listing price is market demand. More people are looking for economically priced Airbnb units. From our above analysis of more than 250,000 Airbnb listings, we observe that 40% of all the Airbnb listings were priced between 0-50 USD. Cumulatively, listings priced between 0-100 USD take up a share of 72%.

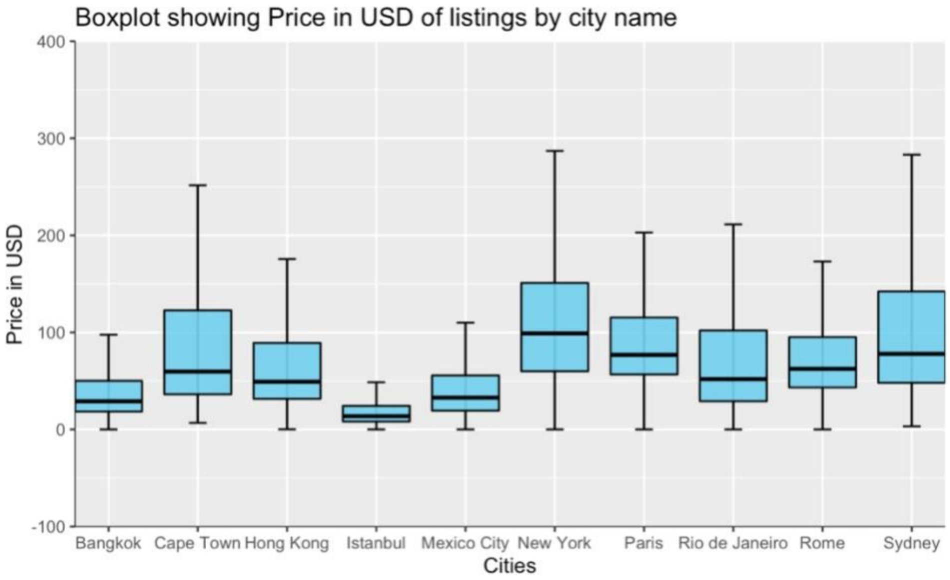


Figure 1g: Boxplots showing price distribution in each city

By this chart, we can clearly see that location influences listing price. As we can see, New York's listings have the highest median price followed by Paris and Sydney. Our initial correlation analysis revealed no single factor alone influences the listing price, but city by city analysis has made it clear that this is a significant factor. This result was also revealed from our class survey, where 57.6% of participants (19/33) considered location to the biggest factor that increases listing price the most (Ahmad, Rasheed, and Yip, 2022).

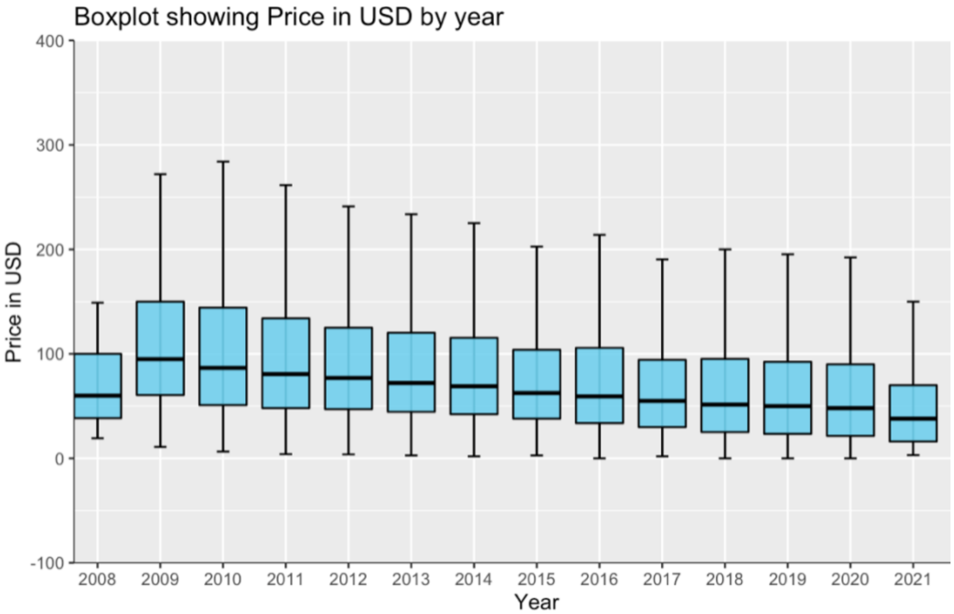


Figure 1h: Boxplots showing distribution of price by year

Passage of time is also affecting the price of listings as we can see that the median price of listings has been facing a downtrend since 2009 and started to plateau as we go forward to 2021. This was surprising to us, as inflation would lead us to believe that prices have increased. However, “supply” has also increased (there are a lot more listings now), which may play a role in the price decrease. Recent studies have found Airbnb price decreases with increased supply (Yang and Mao, 2019).

Guiding Question 2

Assumptions

Shifting to our second guiding question, we now want to see the impact that COVID-19 has had on people becoming hosts. We are making the following assumptions with our analysis of guiding question 2:

- We assume the COVID outbreak began on January 1st, 2020
- We assume the data from different cities is from different populations, therefore, each statistical test will only be done on one of the cities

What is the probability that a city would gain at least 3 new Airbnb host creations per day, before and after Covid?

Our first step in comparing pre and post COVID-19 data was to look at host account creations. We hypothesize that host account creations decreased post-covid. Specifically, we wanted to look at if the probability that there were at least 3 account new host account creations per day has changed since COVID-19. This is an important metric to study, as we can gain powerful insights by looking at host information as well as bookings. Lower host account creations can cause a butterfly effect lowering the entire Airbnb booking rate, so it is worth looking in to.

As we are looking into the number of times an event occurs over a time-interval, we can model host account creation times with a Poisson distribution. Recall that a Poisson distribution can be modeled as:

$$P(X = x) = \frac{e^{-\lambda}(\lambda)^{x}}{x!} \qquad \text{for } x = 0,1,2,3,\ldots$$

In our case, λ would be equal to the mean number of host account creations every day, and we are calculating:

$$P(X \geq 3)$$

Condition for Poisson Distribution

Before we can utilize the Poisson distribution model, we must ensure that our data meets a few conditions:

1. **All events are independent of each other**. This can be assumed as *most of the time* host account creations have nothing to do with one another.
2. **The rate of events through time is unchanged**. In our dataset, the mean amount of host creations does not vary greatly.

3. **Two events cannot occur simultaneously**. We checked this condition by looking for unique host account creation dates and found that there was a negligible amount that happened at the same time.

We calculated this probability for each city pre-covid and post-covid. The resulting probabilities can be visualized below in Figure 1:

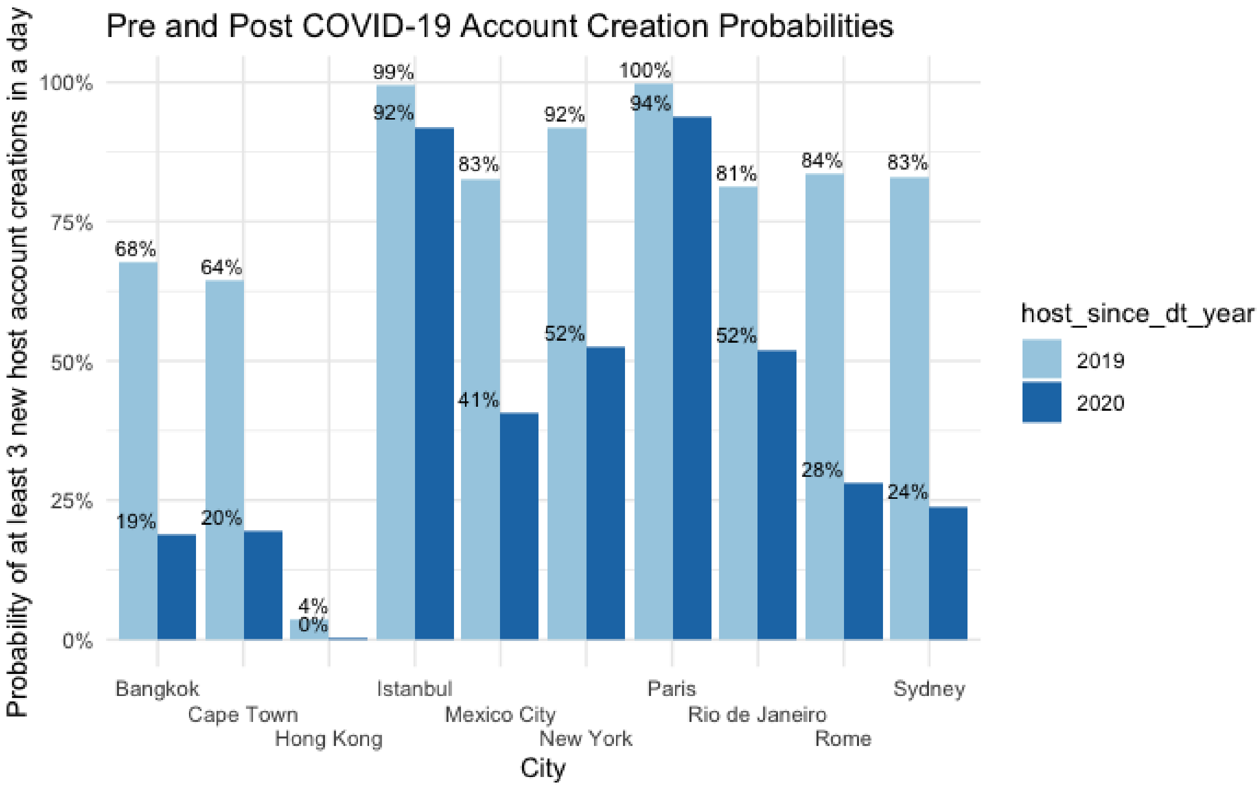


Figure 2a: Poisson Probabilities of at least 3 new host account creations per day, pre and post COVID-19

Inference

From figure 1, we can see a clear decrease in probability pre and post COVID-19. This agrees with our hypothesis that host account creations have decreased. Interestingly, this effect is much less pronounced in Paris and Istanbul. This was an interesting insight, as neither of these places lifted their covid restrictions earlier than most cities.

Looking deeper, a study found that, due to psychological fatigue from early case spikes, the covid attitude in Istanbul was unbothered near the start (Morgul et al., 2020). It is difficult to say why the same case exists for Paris, as they did not have the same early case spikes. Nonetheless, we can see that it is a clear outlier in the data, and most cities saw a decrease in the probability of at least three host account creations a day post covid.

At least 1 day has passed since the last Airbnb host creation, what is the probability that in total, at least 2 days will pass until the next Airbnb account creation?

Data Analysis and Visualization

Next, let us look at our Poisson distribution from a different angle. Let us look now at the probability that, given that at least one day as passed since the last host account creation, what is the probability that at least two more days will pass until the next host account creation. We are calculating:

$$P(X \geq 2 | x \geq \ 1)$$

Our assumptions still hold, and we have visualized the data in figure 2 below:

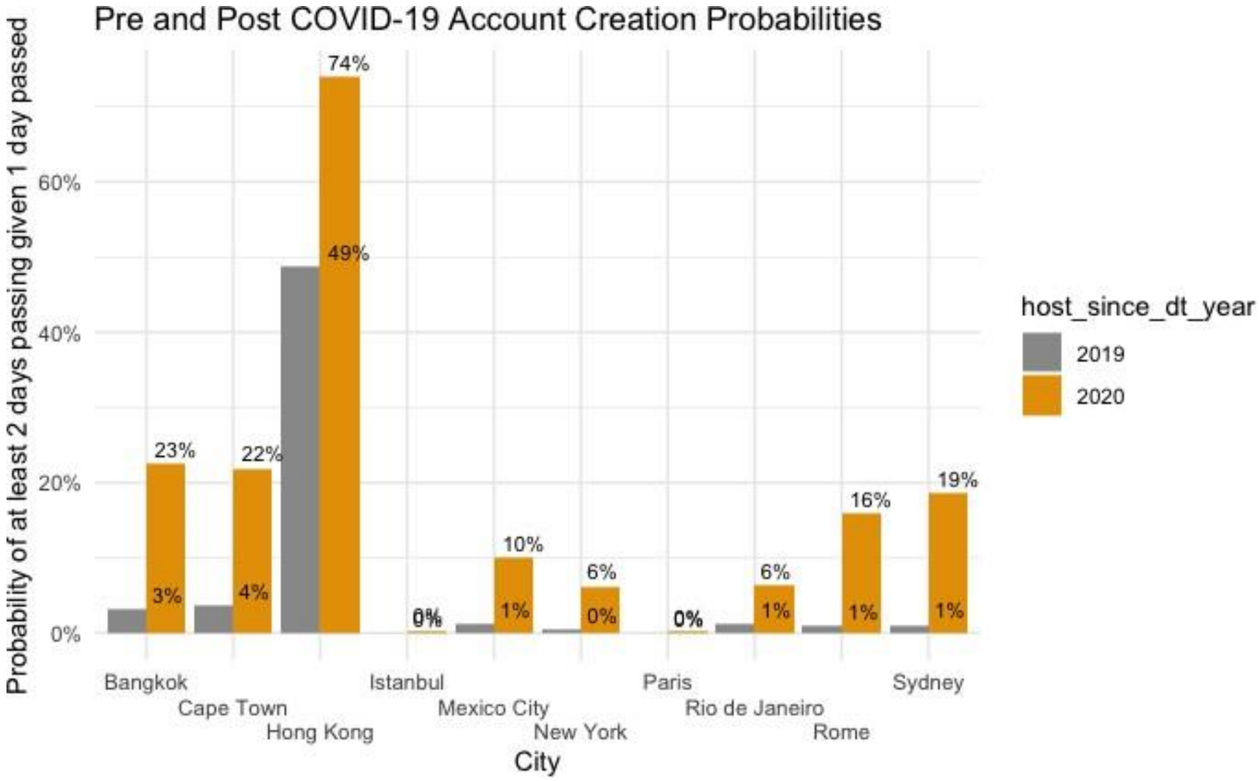


Figure 2b: Poisson Probabilities of at least 3 days passing between account creations given that at least one day has passed, pre and post COVID-19

Inference

This data complies with our conclusion from the last visualization. As we would suspect, the probability that at least 2 days will pass has increased, from which we can infer that less accounts are being created.

This effect is again not seen in Istanbul or Paris. This may again be related to covid attitudes or other external factors, as it looks like there was no change in account creations, and there was a near 0% change of 2 days between creations.

From these two visualizations, we can clearly infer that the number of accounts being created decreased after COVID-19. This same trend did not hold for Paris and Istanbul; however, these can be considered outliers in the data.

Has the mean days between host account creation changed since COVID-19

Data Analysis and Visualization

To further visualize this effect, we created a violin plot of the number of days between account creations before and after COVID-19. For this plot, we sampled n=30 data points from our dataset. We have included the visualization for Bangkok below:

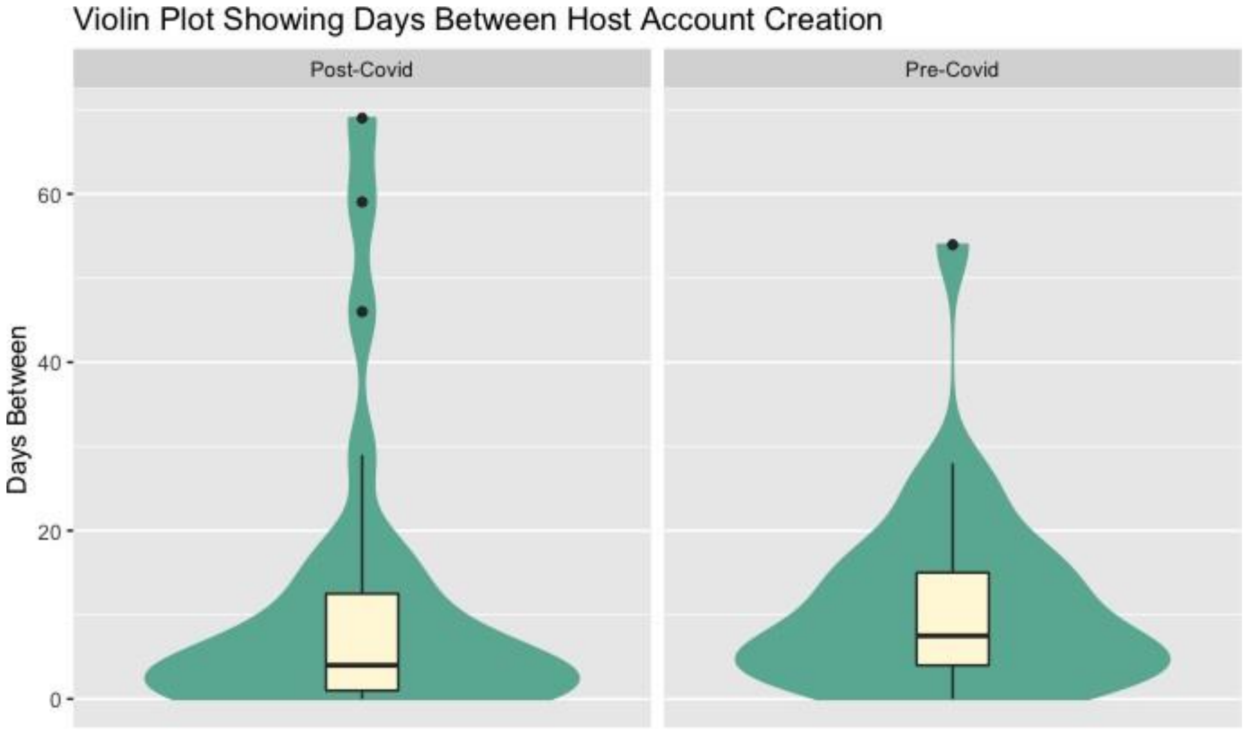


Figure 2c: Violin Plot showing days between host account creation for Bangkok

This violin plot further agrees with our other plots, showing a vertically tighter distribution pre-Covid.

These 3 visuals do help us infer that host account creations have decreased post COVID-19. However, no conclusions can be made until a statistical test is completed. We will start by stating our statistical hypotheses:

$$H_0: \mu_{pre-covid} = \mu_{post-covid}$$

$$H_A: \mu_{pre-covid} \neq \mu_{post-covid}$$

Condition Checking for t-test

Our null hypothesis states that there is no difference in the mean days between host account creation pre covid and post covid. Before we can complete our t-test, we must check the normality condition. Figures 4 and 5 below include normality plots for both the pre and post covid data:

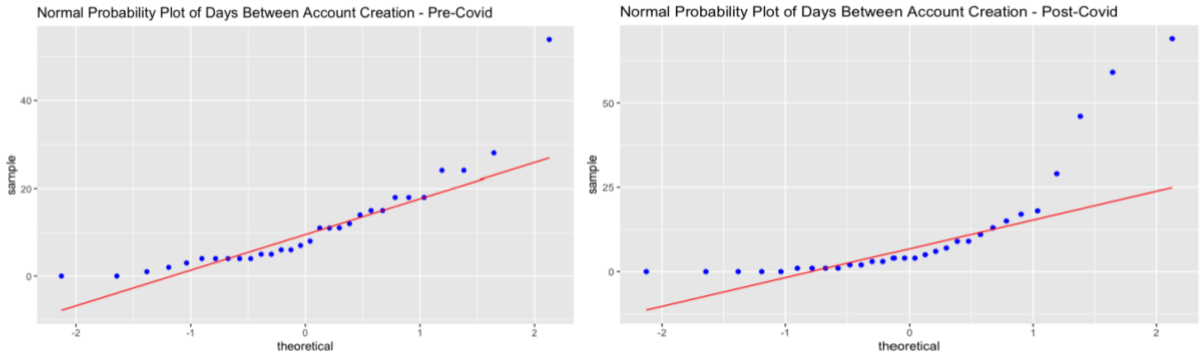
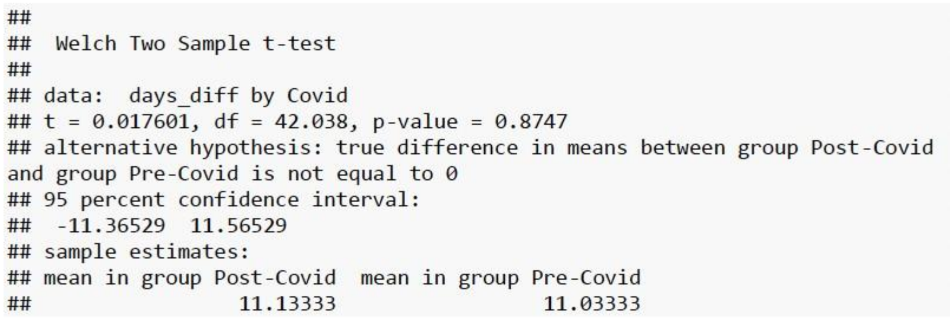


Figure 2d and 2e: Normal probability plots of days between host account creations, pre and post covid

Assessing our normality plots, both datasets are normal, and thus we can continue with our t-test. Recall that the two-sample t-distribution can be expressed as a confidence interval of the difference between two means as:

$$(\bar{X}_1 - \bar{X}_2) \pm t_{1-\frac{\alpha}{2},df}\sqrt{\frac{S_1^2}{n_1} - \frac{S_2^2}{n_2}}$$

Using this, we can model our distribution through a t-distribution. We performed a two-tailed test with a 95% confidence interval. We found the following results:



```
##
##  Welch Two Sample t-test
##
## data:  days_diff by Covid
## t = 0.017601, df = 42.038, p-value = 0.8747
## alternative hypothesis: true difference in means between group Post-Covid
and group Pre-Covid is not equal to 0
## 95 percent confidence interval:
##  -11.36529  11.56529
## sample estimates:
## mean in group Post-Covid  mean in group Pre-Covid
##                 11.13333                 11.03333
```

$$t = 0.017601$$
$$p - value = 0.979$$
$$p = 0.8747$$

$$-11.3652 \leq \mu \leq 11.5652$$

Inference

Our p-value is greater than 0.05, and our confidence interval includes zero, both indicating that we fail to reject the null hypothesis, and that there is no difference in mean number of days between host account creations pre and post covid.

This was a great learning experience for us. All our visuals leaned toward days between bookings increasing post COVID-19. However, our statistical analysis proved that this was very wrong. This is a good example of why visual analysis is often not enough, and it should be performed in tandem with statistical analysis to ensure that we are coming to the correct conclusion.

Do the data provide statistically significant evidence to conclude that there is an association between the Covid and the room type offered? Assume we only have a sample of size 300.

Data Analysis and Visualization

Let us now shift our gaze now toward the effects COVID-19 has had on the type of room being booked. Our data set had 4 different room types: Entire place, hotel room, private room, and shared room.

Studies have found that more people are choosing places to themselves over shared places post covid (Bagnera, S.M., Stewart, E. and Edition, S., 2020). We hypothesize that booking an entire place has increased post-covid. Our analysis considered data from Hong Kong. Since we are analyzing bivariate categorical data, a Chi-squared test of independence seemed appropriate.

Our data contained E_{ij} values lower than 5, we had to simulate p-values in our test. The chi squared statistic was calculated by the following:

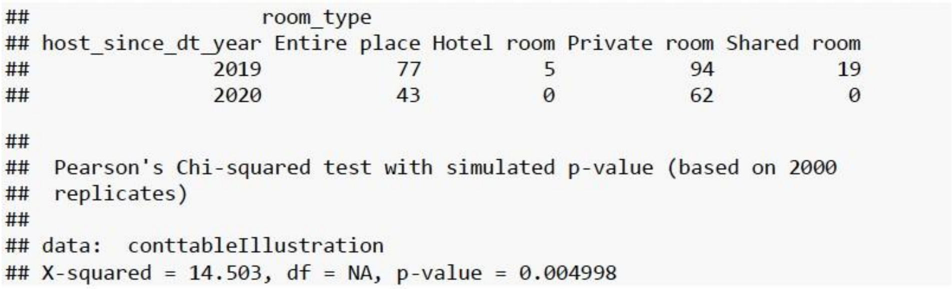
$$\chi^2_{Obs} = \sum_{i=i}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2_{df=(c-1)*(r-1)}$$

Before performing our chi-squared test, we must state our statistical hypotheses:

H_0: Covid-Status and Room Type Booked are Independent

H_A:Covid-Status and Room Type Booked are not Independent

All our conditions have been met (if we simulate p-values), and we received the following results:



```
##                  room_type
## host_since_dt_year Entire place Hotel room Private room Shared room
##               2019           77          5           94          19
##               2020           43          0           62           0

## 
##  Pearson's Chi-squared test with simulated p-value (based on 2000
##  replicates)
## 
## data:  conttableIllustration
## X-squared = 14.503, df = NA, p-value = 0.004998
```

$$\chi^2 = 16.854$$
$$\chi^2 = 14.503$$

$$p - value = 0.0004998$$

Inference

We get a p-value remarkably close to zero. Because of this, we can reject the null hypothesis, and say that covid-status and room type booked are not independent. This agrees with our initial hypothesis.

Does the host response rate differ before and after Covid?

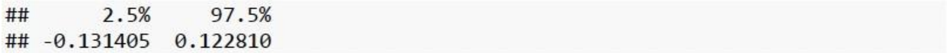
We next aimed to study if host behaviour has changed since covid. Specifically, we aimed to analyze what the change in host response rate has been pre and post covid. We hypothesize that there has been no change in host behaviour. We started by stating our statistical hypotheses.

$$H_0\text{: } \mu_{pre-covid} = \mu_{post-covid}$$

$$H_A\text{: } \mu_{pre-covid} \neq \mu_{post_{c}ovid}$$

Data Analysis and Visualization

To perform our statistical test, we decided to bootstrap the mean difference, with 1000 replicates. In each bootstrap, we would record the mean difference in that bootstrapped sample. We then found the 95% interval of this data and used it to accept/reject our null hypothesis. The 95% confidence interval was as follow:



```
##      2.5%     97.5% 
## -0.131405  0.122810
```

$$-0.1314 \leq \mu_{pre-covid} - \mu_{post-covid} \leq 0.1228$$

Inference

Since our confidence interval captures a mean difference of zero, we can say with confidence that we fail to reject our null hypothesis and there is no mean difference between host response rate pre and post covid. This agrees with our hypothesis. We have also included a figure of our bootstrap distribution:

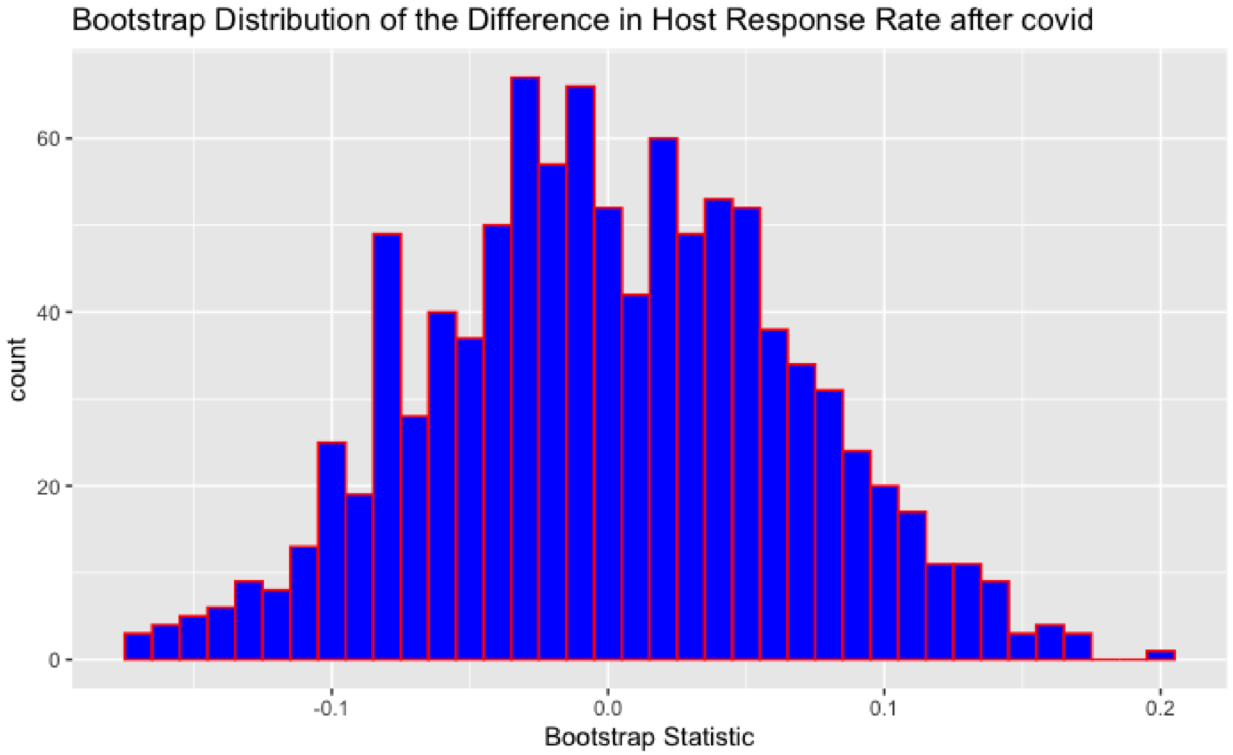


Figure 2f: Bootstrap distribution of difference in mean host response rate pre and post covid

You may be wondering why we even chose to test this metric, as we hypothesized there would be no difference. With a statistical analysis of the effects of COVID-19 on the Airbnb market, it is important to analyze the data from all angles (including host behaviour), to come to a strong conclusion.

Build a model to attempt to forecast the number of new Airbnb joiners which Mexico City Airbnb will have post-covid

Our last step in the analysis of guiding question 2 is to build a linear model to forecast new Airbnb host account creations in Mexico City. We will use this model to attempt to forecast how many new accounts were made in 2020 (the covid year). Our linear model will not contain the real data for the year 2020.

Condition Checking for LM model

The first step in constructing our linear model is to check conditions. The response variable must be normally distributed with a mean and standard deviation. To check this condition, we plotted a normality plot with the residual values of Airbnb host account creations. The plot can be seen below:

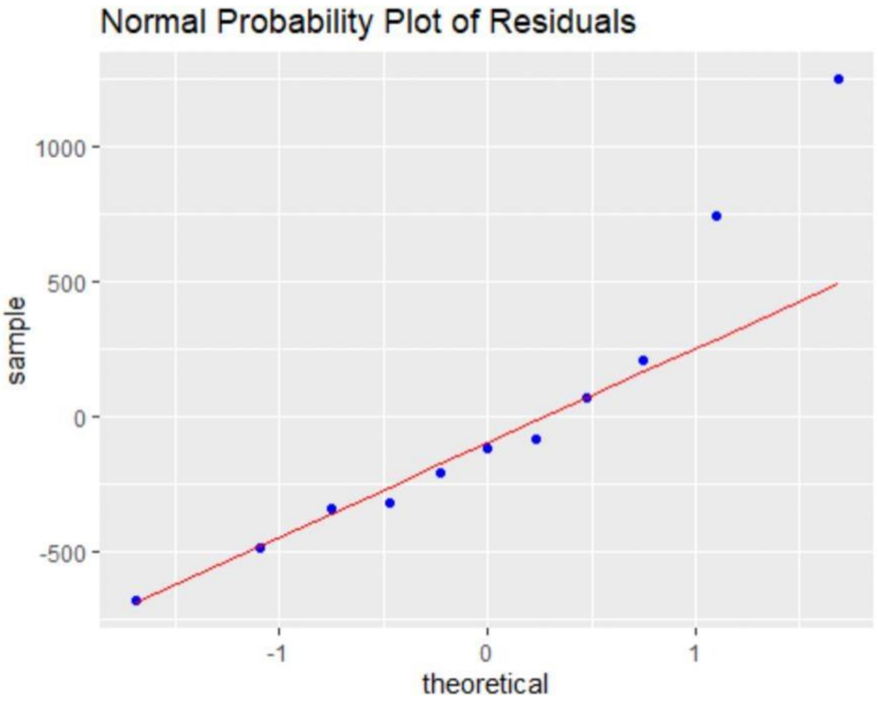


Figure 2g: Normality plot for host account creations in Mexico City pre-covid

The next condition that we need to check is that the data is homoscedastic. This means that for each distinct value of the predictor variable, and response variable has the same standard deviation. We can check this with a plot of residuals, shown below:

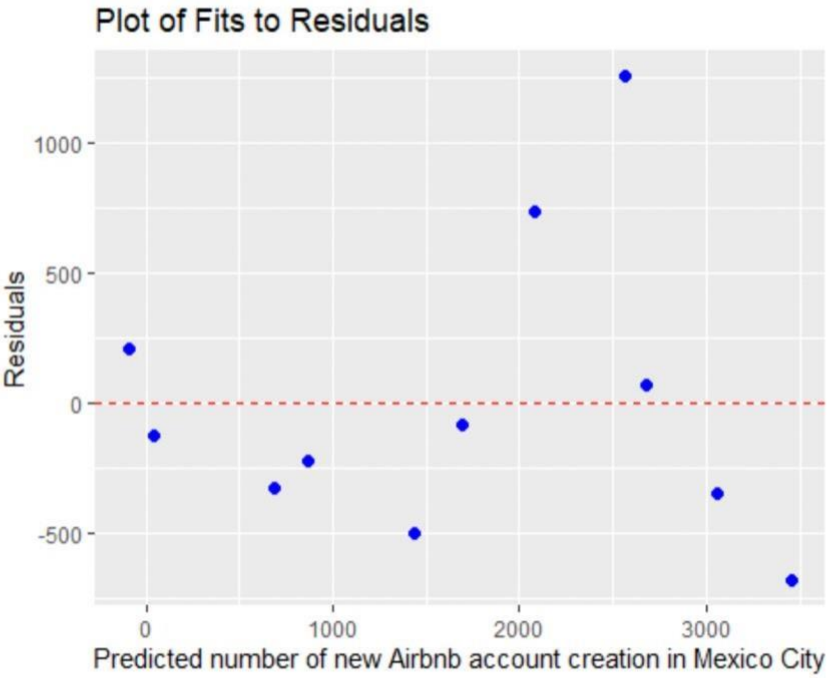
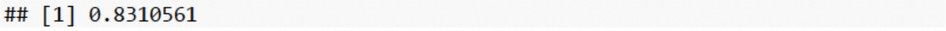


Figure 2h: Residual plot for host account creation in Mexico City pre-covid

From the above plot, we can see that the residuals are scattered relatively uniformly around the y=0 line, and so we can conclude that the data is homoscedastic.

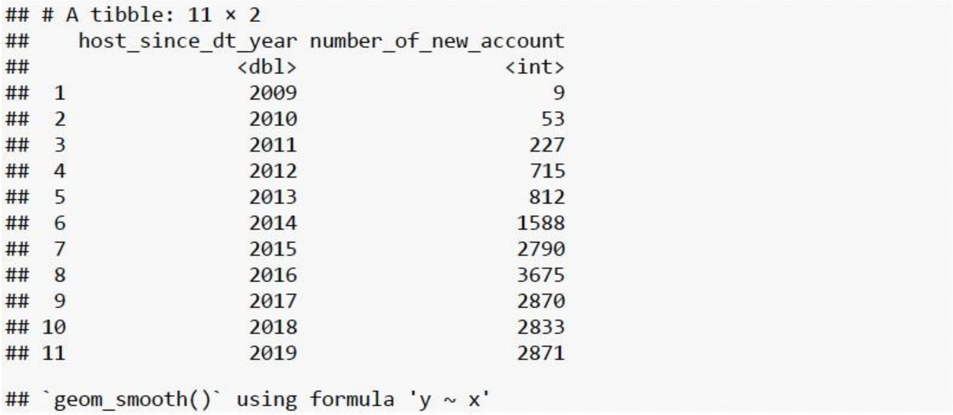
Just to ensure our data is correlated, we calculated a correlation coefficient and got the following value:



```
## [1] 0.8310561
```

$$R^2 = 0.83$$

The R-squared of the model is 0.83, which means 83% of the variance in the dependent variable (the Airbnb new host creations) is explained by an independent variable (Year) in a regression model. This is high enough to say that our data is quite strongly correlated.



```
## # A tibble: 11 × 2
##    host_since_dt_year number_of_new_account
##                 <dbl>                 <int>
##  1               2009                     9
##  2               2010                    53
##  3               2011                   227
##  4               2012                   715
##  5               2013                   812
##  6               2014                  1588
##  7               2015                  2790
##  8               2016                  3675
##  9               2017                  2870
## 10               2018                  2833
## 11               2019                  2871

## `geom_smooth()` using formula 'y ~ x'
```

Data Analysis and Visualization

Finally, we added a linear regression plot to get a better understanding of our linear model:

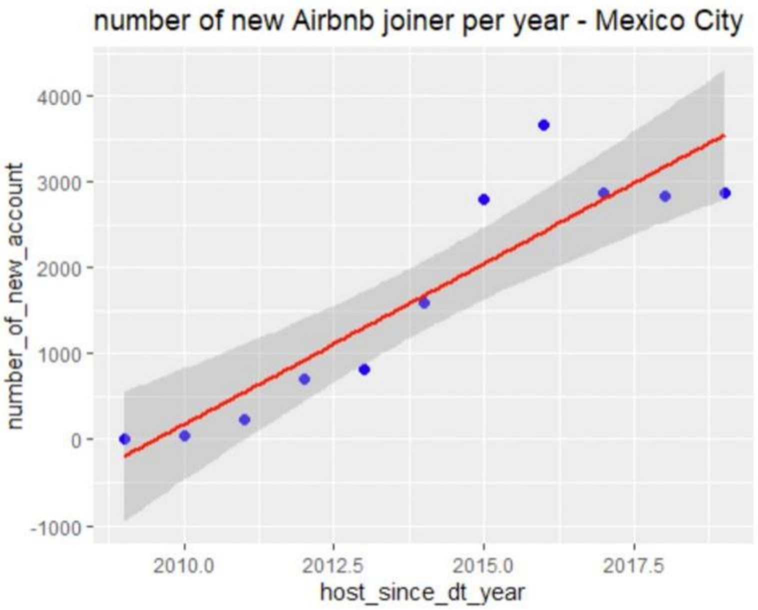
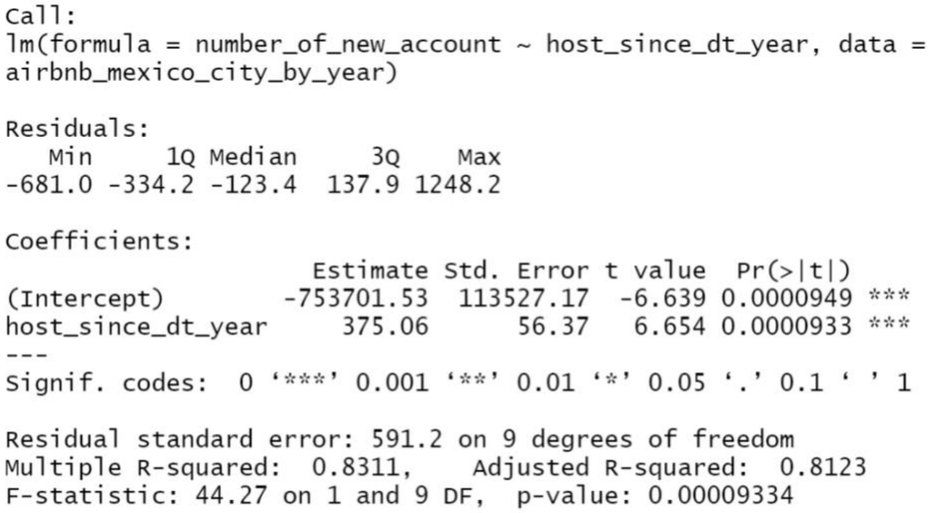


Figure 2i: Linear regression plot of host account creation in Mexico City pre-covid

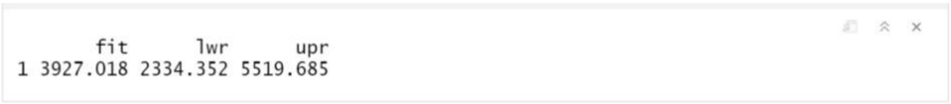
Now, it is time to construct the model:



```
Call:
lm(formula = number_of_new_account ~ host_since_dt_year, data =
airbnb_mexico_city_by_year)

Residuals:
   Min     1Q Median     3Q    Max
-681.0 -334.2 -123.4  137.9 1248.2

Coefficients:
                     Estimate Std. Error t value  Pr(>|t|)
(Intercept)        -753701.53  113527.17  -6.639 0.0000949 ***
host_since_dt_year     375.06      56.37   6.654 0.0000933 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 591.2 on 9 degrees of freedom
Multiple R-squared:  0.8311,    Adjusted R-squared:  0.8123
F-statistic: 44.27 on 1 and 9 DF,  p-value: 0.00009334
```

$$Airbnb\ new\ host\ total_i = -753701 + (375 * Year_i)$$

We have a model with a slope of 375 and an intercept of –753701, with standard errors of 56.37 and 113527.17, respectively. 375 can be considered the average number of new hosts that join per year. Let us forecast the Airbnb new host total for year 2020 (after Covid).



```
       fit      lwr      upr
1 3927.018 2334.352 5519.685
```

$$2334 \leq Airbnb\ new\ host\ total_{Year_i=2020} \leq 5520$$

Inference

Using the model to forecast 2020, Mexico City should have between 2334 and 5520 new host account creations. The actual number of host account creations in 2020 was 1480, which is outside the 95% confidence interval. Therefore, there is statistical evidence that Covid influences the number of Airbnb new host account creation.

Conclusions

In the analysis of our first guiding question (the price analysis), we first found that no one column strongly correlated with price, although a combination of columns may be. However, we found that a mean listing price difference exists in the following 3 categories:

- Host verified (more expensive) Vs. unverified (cheaper)
- Higher number of amenities (more expensive) Vs. lower number of amenities (cheaper)
- Instant bookable (more expensive) Vs. non-instant bookable (cheaper)

These are indicative of an interesting idea. Previous studies have shown that hosts who feel like they have more to offer tend to charge more than their unit is worth (Toader et al., 2022). With our analysis, we can see actual metrics of what hosts feel like they have to offer: verification, number of amenities, and instant bookability. In future research, it would be interesting to see other metrics that cause hosts to increase their listing prices, as well as rank these metrics on which one has the biggest influence on price.

Looking at the other visualisations for price analysis (figs. 1f, 1g, 1h), a few trends can be observed.

From figure 1f, we notice the obvious trend of number of bookings decreasing as price increases. What is interesting is that 72% of listings are priced under $100 USD. It was difficult to find the true average listing price in an article; however, we felt on average the listings we saw were more expensive than this.

From figure 1g, we can see that the median price does not vary much from city to city. Istanbul has the lowest median price, while New York has the highest, with Paris in second place. Paris and Istanbul are often outliers (or near outliers in this case) in our analysis. An interesting further study would be to investigate why they tend to differ from the norm so much, sometimes in opposite directions.

Finally, figure 1h shows us that median price is decreasing. As mentioned above, this is surprising when you think about inflation, but makes more sense when you think about supply and demand for Airbnb. The fact that price is decreasing indicates that supply is increasing faster than demand in this market.

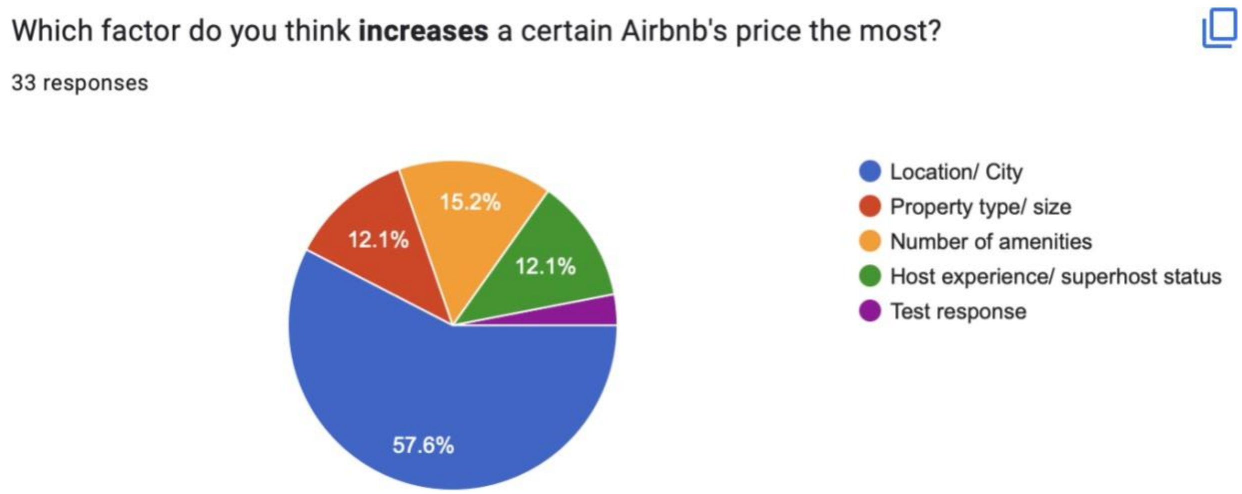
Moving to our second guiding question (the COVID analysis), we first analyzed host account creations. Figure 2a showed “the probability of at least 3 new host account creations” has decreased since COVID, while figure 2b showed “the probability of at least 2 days passing since an account creation given that at least one day has passed” has also decreased since COVID. From this, we inferred that host account creations have decreased since COVID. However, upon statistical testing, we found that we were incorrect (by quite a large margin). Therefore, data visualisation and statistical analysis must be done in tandem.

We also found that people were more likely to book an entire place to themselves vs a shared room post covid. This agreed with past research. A limit of this analysis is that we do not know if more people are choosing rooms like this, or if more rooms like this are being offered. We suspect it is a mix of both, as everyone (hosts and guests) has become more cautious of their health post covid.

The last step in our analysis was constructing a linear model to predict new host creations in Mexico City in 2020. We suspected our prediction would be a lot higher than the true value, as COVID is playing a role. Our 95% confidence interval did not capture the amount of host creations in Mexico City, and we are confident it deviates from the model.

Airbnb continues to be impacted by COVID-19, and our study is limited in the amount of post covid data we had. Further studies should be conducted to investigate trends that have formed in the covid years, as well as what is happening with certain outlier cities. This can be valuable information on these cities. The investigation we conducted not only gave us insight into the Airbnb industry, but also the cities we were analysing.

Appendix



Which factor do you think **increases** a certain Airbnb's price the most?

33 responses

Figure 1c: Data from our class survey used in some statistical analyses

References

Ahmad, I., Rasheed, I., Yip, C., 2022. Google Survey <https://docs.google.com/forms/u/0/d/1adnCI6hq2kiY8BSa_K7xmP2E6NVlGxiza0PNPHDL1A8/viewform?edit_requested=true>

Airbnb, 2022, *Airbnb API*

Bagnera, S.M., Stewart, E. and Edition, S., 2020. Navigating hotel operations in times of COVID-19. *Boston Hospitality Review*, *1*(7).

Bhat, M. 2021, *Airbnb Listings & Reviews*, electronic dataset, Kaggle, viewed 13 Sept. 2022, <https://www.kaggle.com/datasets/mysarahmadbhat/airbnb-listings-reviews>

CC0 1.0 Universal Public Domain Dedication, 2022, *Creative Commons*, <https://creativecommons.org/publicdomain/zero/1.0/>

Dogru, T., Mody, M., Line, N., Hanks, L., Suess, C. and Bonn, M., 2021. The Effect of Airbnb on Hotel Performance: Comparing Single- and Multi-Unit Host Listings in the United States. Cornell Hospitality Quarterly, 63(3), pp.297-312.

Gyódi, K., 2021. Airbnb and hotels during COVID-19: different strategies to survive. *International Journal of Culture, Tourism and Hospitality Research*.

Jang, S. and Kim, J., 2022. Remedying Airbnb COVID-19 disruption through tourism clusters and community resilience. Journal of Business Research, 139, pp.529-542.

Morgul, E. *et al.* (2020) “Covid-19 pandemic and psychological fatigue in Turkey,” *International Journal of Social Psychiatry*, 67(2), pp. 128–135. Available at: https://doi.org/10.1177/0020764020941889.

Toader, V., Negrușa, A.L., Bode, O.R. and Rus, R.V., 2022. Analysis of price determinants in the case of Airbnb listings. Economic Research-Ekonomska Istraživanja, 35(1), pp.2493-2509.

Wang, C.R. and Jeong, M., 2018. What makes you choose Airbnb again? An examination of users’ perceptions toward the website and their stay. *International Journal of Hospitality Management*, *74*, pp.162-170.

Wyman, D., Mothorpe, C. and McLeod, B., 2020. Airbnb and VRBO: the impact of short-term tourist rentals on residential property pricing. Current Issues in Tourism, pp.1-12.

Yang, Y. and Mao, Z., 2019. Welcome to my home! An empirical analysis of Airbnb supply in US cities. Journal of Travel Research, 58(8), pp.1274-1287.

Zhang, L., Yan, Q. and Zhang, L., 2020. A text analytics framework for understanding the relationships among host self-description, trust perception and purchase behavior on Airbnb. *Decision Support Systems*, *133*, p.113288.

Zhang, Z., Chen, R.J., Han, L.D. and Yang, L., 2017. Key factors affecting the price of Airbnb listings: A geographically weighted approach. *Sustainability*, *9*(9), p.1635.