MA4710 Project

Cody Rorick

2024-05-24

Data description

Since 2008, guests and hosts have used Airbnb to expand on traveling possibilities and present more unique, personalized way of experiencing the world. This dataset describes the listing activity and metrics in NYC, NY for 2019.

The numerical predictors in this dataset are:

- Host Total number of listings (does the host own other properties on airbnb)
- Total number of reviews
- Price ($)
- Minimum nights per stay
- Number of days available in the year

The categorical predictors are:

- Host Total number of listings
- Room type (Entire home/apt, Private room, etc..)

The response variable of this data set is:

- Frequency of reviews the listing gets (reviews per month)

I am expecting to see a positive correlation between the total number of reviews and the frequency of reviews. The other variables I am less sure about but curious to dive into, as any one of these could have an impact on the frequency a listing gets reviewed. Airbnb reviews are a great way to establish legitimacy and bring attention to your posting, so having the ability to bring in more reviews per month is an important metric to rental property owners.

Regression equation

```
Review Frequency per Year = 8.9581 + -0.1191 * min_night_stay + -0.0004 * price
 + 0.0193 * host_listings + 0.0127 * availability + 0.2218 * tot_reviews
```

Are coefficients non negligible?

A table of the t-statistics, standard errors, and p-values in each of the tests that the null hypothesis coefficients are zero and alternative hypothesis that they are not equal to zero

```
intercept = 8.9581 t = 9.6471 p = 0.0000
min_night_stay = -0.1191 t = -4.0794 p = 0.0000
price = -0.0004 t = -0.0881 p = 0.4649
host_listings = 0.0193 t = 0.9726 p = 0.1655
availability = 0.0127 t = 3.2634 p = 0.0006
tot_reviews = 0.2218 t = 23.6864 p = 0.0000
```

Regression coefficient significance interpretation

All coefficients used except host_listings were shown to be significant at a 0.05-significance level in predicting frequency of reviews.

- intercept = 8.9581 (reviews/year), implies if all other factors are zero, baseline frequency starts at 8.9581 reviews per year. This number isn't meaningful to interpret
- price = -0.0004 (reviews/year*$), implies that for every dollar increase in Airbnb price, the frequency of reviews you get goes down by 0.0004 per year because the coefficient is negative
- min_night_stay = -0.1191 (reviews/year*days), implies that every additional required night for a minimum stay, the frequency of reviews goes down 0.1191 per year because the coefficient is negative
- host_listings = 0.0193 (reviews/year*listings), implies that every additional property a host owns, the frequency of reviews goes up 0.0193 per year
- availability = 0.0127 (reviews/year*days), implies that every additional available booking day a property has, the frequency of reviews goes up by 0.0127 per year
- tot_reviews = 0.2218 (reviews/year*reviews), implies that every additional review a property has, the frequency of reviews goes up by 0.2218 per year

F statistic, P-value, Sigma, and Correlation of Overall Regression

```
F statistic = 128.1324
```

```
P value = 0.0000
```

```
Sigma = 15.2679
```

```
R squared value = 0.3919
```

```
adjusted R squared value = 0.3889
```

Interpretation of sigma and R squared

```
The R squared value insinuates that 39.1923 percent of the variation in
review frequency can be attributed to the predictors included in this model
```

```
A sigma of 15.2679 reviews per year is an indicator of the amount of variation
that occurs around a predicted review frequency
```

P value associated with other hypothesis tests

The anova table below indicates if host number of host listings coefficients is equal to zero controlled for the other predictors. The p value indicates that we fail to reject the null hypothesis that it is equal to 0.

```
Analysis of Variance Table

Model 1: reviews_per_year ~ minimum_nights + price + availability_365 +
    number_of_reviews
Model 2: reviews_per_year ~ minimum_nights + price + calculated_host_listings_count +
    availability_365 + number_of_reviews
  Res.Df    RSS Df Sum of Sq     F Pr(>F)
1    995 231929
2    994 231709  1    220.53 0.946  0.331
```

The anova table below indicates whether or not the yearly availability coefficient could be equal to 0.015 reviews/year * days. The P value indicates that we fail to reject the null hypothesis that the coefficient is equal to this value.

```
Analysis of Variance Table

Model 1: reviews_per_year ~ minimum_nights + price + calculated_host_listings_count +
    number_of_reviews + offset(I(0.015 * availability_365))
Model 2: reviews_per_year ~ minimum_nights + price + calculated_host_listings_count +
    availability_365 + number_of_reviews
  Res.Df    RSS Df Sum of Sq     F Pr(>F)
1    995 231788
2    994 231709  1      79.5 0.341 0.5594
```

The anova table below indicates if yearly availability and minimum night stay coefficients give orthogonal information. The P value indicates that we should reject the null hypothesis that minimum night stay and yearly availability give the same information in the regression model.

```
Analysis of Variance Table

Model 1: reviews_per_year ~ price + calculated_host_listings_count + number_of_reviews +
    I(availability_365 + minimum_nights)
Model 2: reviews_per_year ~ minimum_nights + price + calculated_host_listings_count +
    availability_365 + number_of_reviews
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    995 236328
2    994 231709  1    4619.1 19.815 9.495e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```