Lab 05

Rohit Kulkarni

02/09/2022

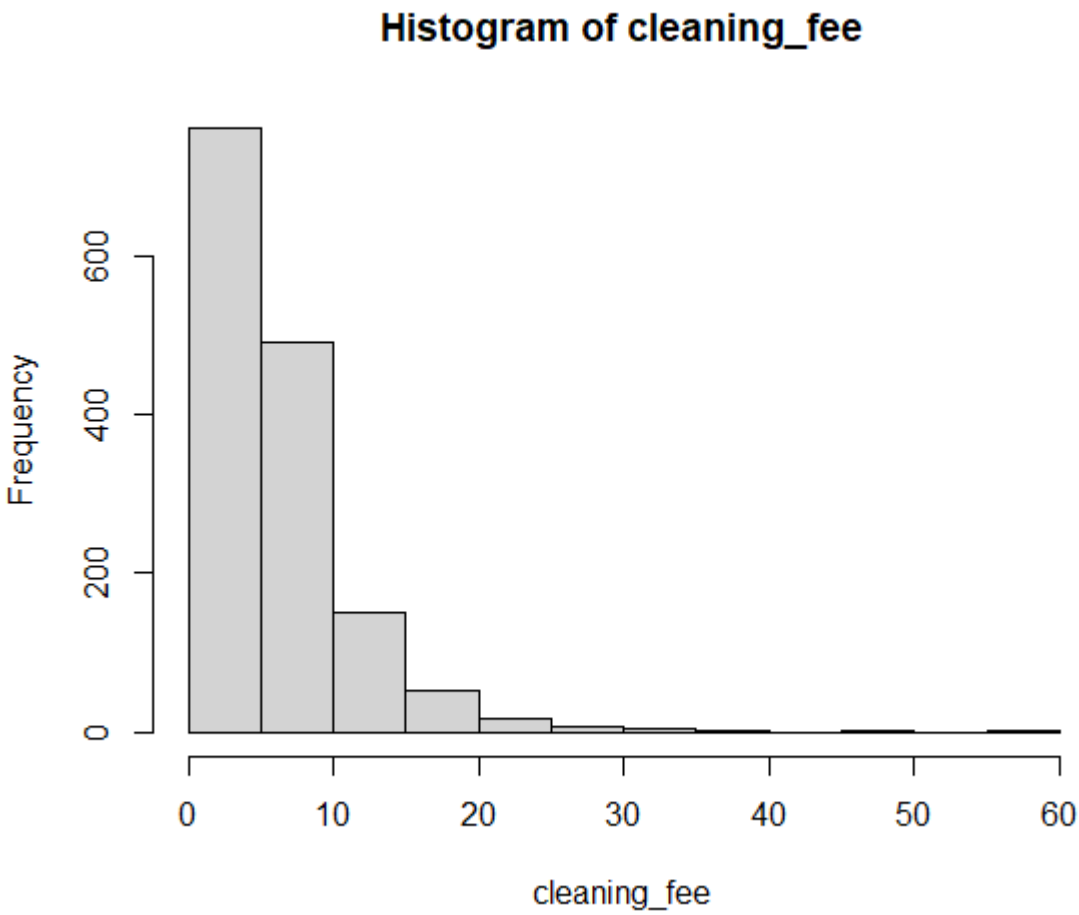
##Data wrangling & EDA 1. Some Airbnb rentals have cleaning fees, and we want to include the cleaning fee when we calculate the total rental cost. Create a variable call cleaning_fee calculated as the 2% of the price per night.

```
cleaning_fee <- airbnb$price * 0.02
```

2. Visualize the distribution of cleaning_fee and display the appropriate summary statistics. Use the graph and summary statistics to describe the distribution of cleaning_fee.

```
hist(cleaning_fee)
summary(cleaning_fee)
```

Summary Statistics: Min. 1st Qu. Median Mean 3rd Qu. Max. 0.620 2.880 5.000 6.378 8.060 59.000



Histogram:
The cleaning fee distribution is skewed to the right, with a median of 5 dollars per night. There are outliers as well, for example the max which is a cleaning fee of 59 dollars per night.

3. Next, let's examine the neighbourhood.

- How many different categories of neighbourhood are in the dataset? Show code and output to support your answer.

```
unique(airbnb$neighbourhood)
```

There are 5 categories in neighborhood: "Unincorporated Areas", "City of Santa Cruz", "City of Capitola", "City of Scotts Valley", "City of Watsonville"

- Which 3 neighborhoods are most common in the data? These 3 property types make up what percent of the observations in the data? Show code and output to support your answer.

```
table(airbnb$neighbourhood)
```

Output: City of Capitola City of Santa Cruz City of Scotts Valley City of Watsonville 218 369 26 15 Unincorporated Areas 861

The three property types that are most common are Unincorporated Areas and City of Santa Cruz, and City of Capitola. By proportion, these categories make up 57.8% for Unincorporated Areas, 24.7% for City of Santa Cruz, and 14.6% for City of Capitola. Altogether, these three categories make up 97.2% of the data.

4. Since an overwhelming majority of the observations in the data are one of the top 3 cities, we would like to create a simplified version of the neighbourhood variable that has 4 categories.

Create a new variable called neigh_simp that has 4 categories: the three from the previous question and "Other" for all other places. Be sure to save the new variable in the data frame.

```
neigh_simp <- fct_recode(airbnb$neighbourhood, "Other" = "City of Scotts Valley",
                      "Other" = "City of Watsonville")

airbnb$neighbourhood <- neigh_simp
```

5. What are the 4 most common values for the variable minimum_nights? Which value in the top 4 stands out? What is the likely intended purpose for Airbnb listings with this seemingly unusual value for minimum_nights? Show code and output to support your answer.

```
count(airbnb, minimum_nights)
```

This code shows the count for each number of minimum nights in the Airbnb. This is the output:

minimum_nights n 1 1 420 2 2 571 3 3 223 4 4 56 5 5 32 6 6 10 7 7 30 8 8 1 9 10 3 10 14 7

The four most common values for minimum_nights are, in order: 2 nights, 1 night, 3 nights, and 30 nights. The most likely explanation is that some airbnb's are intended for a model more similar to paying for a month of rent. These airbnb's most likely accomodate customers for longer periods of time.

Airbnb is most commonly used for travel purposes, i.e. as an alternative to traditional hotels, so we only want to include Airbnb listings in our regression analysis that are intended for travel purposes. Filter airbnb so that it only includes observations with minimum_nights <= 3‘.

```
airbnb <- filter(airbnb, minimum_nights <= 3)
```

##Regression

6. For the response variable, we will use the total cost to stay at an Airbnb location for 3 nights. Create a new variable called price_3_nights that uses price and cleaning_fee to calculate the total cost to stay at the Airbnb property for 3 nights. Note that the cleaning fee is only applied one time per stay.

Be sure price is in the correct format before calculating the new variable.

```
price_3_nights <- airbnb$price + airbnb$cleaning_fee
airbnb$price_3_nights <- price_3_nights
```

7. Fit a regression model with the response variable from the previous question and the following predictor variables: neigh_simp, number_of_reviews, and reviews_per_month. Display the model with the inferential statistics and confidence intervals for each coefficient.

```
model <- lm(price_3_nights ~ neighbourhood + number_of_reviews + reviews_per_month,
            data=airbnb)
summary(model)
```

This is the summary of the model printed: Coefficients: Estimate Std. Error t value Pr(>|t|)
(Intercept) 498.30724 21.99962 22.651 < 2e-16 ***neighbourhoodCity of Santa Cruz -70.25199 25.64290 -2.740 0.00625*** *neighbourhoodOther -226.81481 53.96435 -4.203 2.84e-05* **neighbourhoodUnincorporated Areas -105.59104 22.20950 -4.754 2.25e-06** *number_of_reviews -0.14747 0.06834 -2.158 0.03113*
reviews_per_month -28.76622 4.24340 -6.779 1.94e-11 ***

The formula for the confidence interval would be the estimate, plus or minus the standard error mulitplied by the t-value.

neighbourhood (City of Santa Cruz): [-140.5135, 0.0095] neighbourhood (Other): [-453.6269, -0.0026] neighbourhood (Unincorporated Areas): [-211.1750, -0.0070] number_of_reviews: [-0.2949, 0.0000] reviews_per_month: [-57.5322, -0.0002]

8. Interpret the coefficient of number_of_reviews and its 95% confidence interval in the context of the data.

In the context of this problem, the number of reviews is either negatively correlated or not correlated at all with the price of the Airbnb.

9. Interpret the coefficient of neigh_simpCity of Santa Cruz and its 95% confidence interval in the context of the data.

In the context of this problem,whether or not the Airbnb is in Santa Cruz is negatively correlated with the price (ie Airbnb's in Santa Cruz are cheaper).

10. Interpret the intercept in the context of the data. Does the intercept have a meaningful interpretation? Briefly explain why or why not.

This intercept does not have a meaningful interpretation in this context because there are no Airbnb's that have a negative cost.

11. Suppose your family is planning to visit Santa Cruz over Spring Break, and you want to stay in an Airbnb. You find an Airbnb that is in Scotts Vallye, has 10 reviews, and 5.14 reviews per month. Use the model to predict the total cost to stay at this Airbnb for 3 nights. Include the appropriate 95% interval with your prediction.

```
predict(model, data.frame(neighbourhood=c("City of Santa Cruz"), number_of_reviews=c(10),
                          reviews_per_month=c(5.14)))
```

With this model, I can predict that the cost of the Airbnb will be $278.72 per night. By slightly tweaking this code, we can also get the 95% interval for the prediction (although it does not show, the predict() method automatically assumes 95% interval):

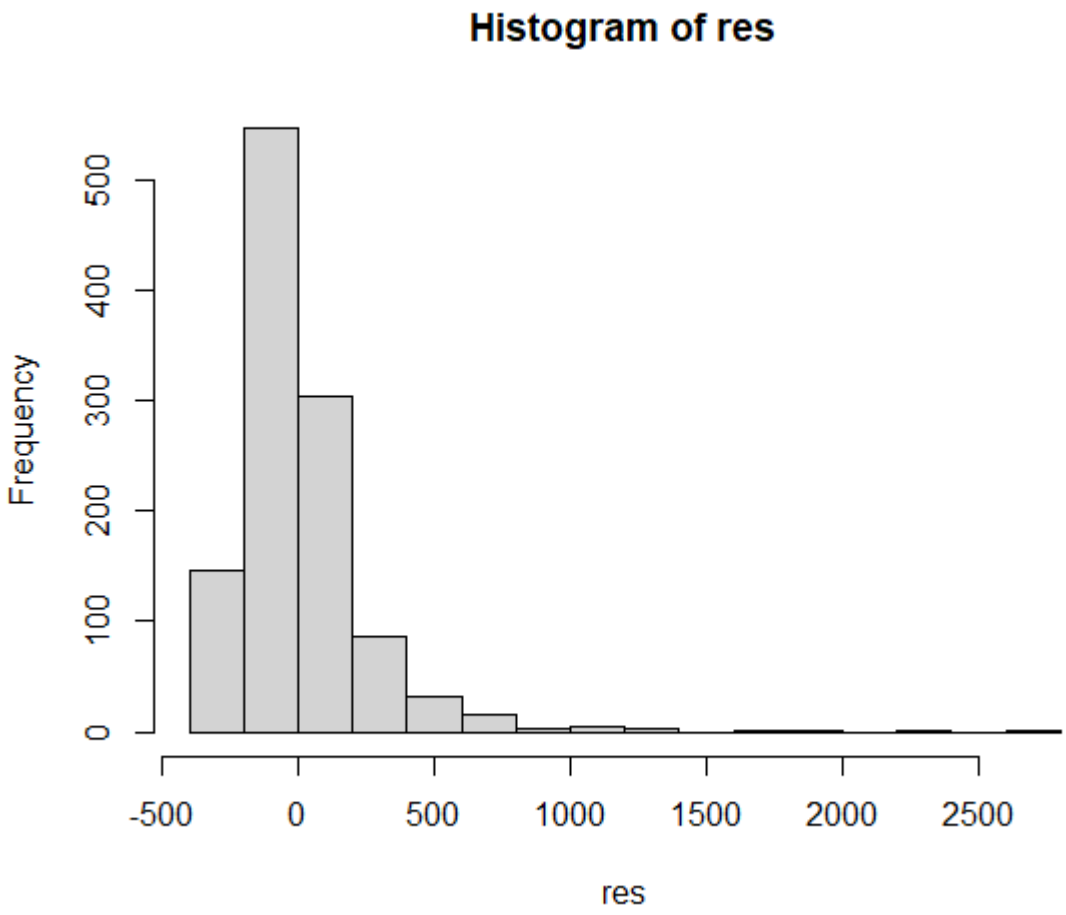
```
predict(model, data.frame(neighbourhood=c("City of Santa Cruz"), number_of_reviews=c(10),
                          reviews_per_month=c(5.14)), interval="confidence")
```

12. Now check the assumptions for your regression model. Should you be confident on interpreting the inferential results of your model?

One assumption we make is the Normality assumption, which assumes that the distribution of y values is Normal at every X value. To check this, we can plot the residuals on a histogram and check if the spread is normal:

```
res <- resid(model)

hist(res)
```



Output:

We can clearly see that the spread of the residuals is skewed to the right heavily and not a Normal distribution. Therefore, we should not be confident in interpreting inferential results of the model I created.