Lab 5

Ravi Gupta

2/18/2022

```
airbnb <- read_csv("listings.csv")
```

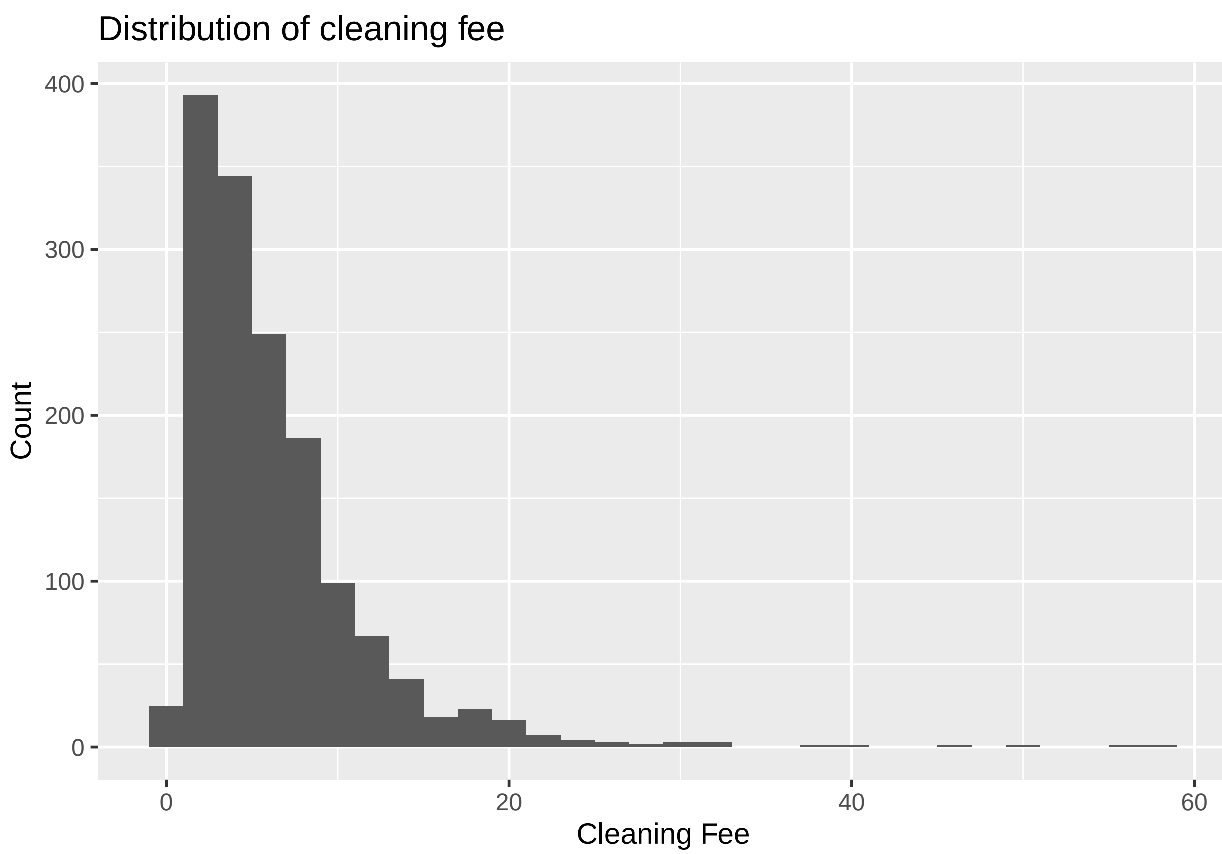
```
## Rows: 1489 Columns: 18

## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr   (4): name, host_name, neighbourhood, room_type
## dbl  (11): id, host_id, latitude, longitude, price, minimum_nights, number_o...
## lgl   (2): neighbourhood_group, license
## date  (1): last_review

##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

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.

```
cleanairbnb <- airbnb %>%
  mutate(cleaning_fee = .02 * (price))
```

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. The distribution is skewed right.

```
ggplot(data = cleanairbnb, aes(x = cleaning_fee)) + geom_histogram(binwidth = 2) + labs(x = "Cleaning Fe
       y = "Count",
       title = "Distribution of cleaning fee")
```



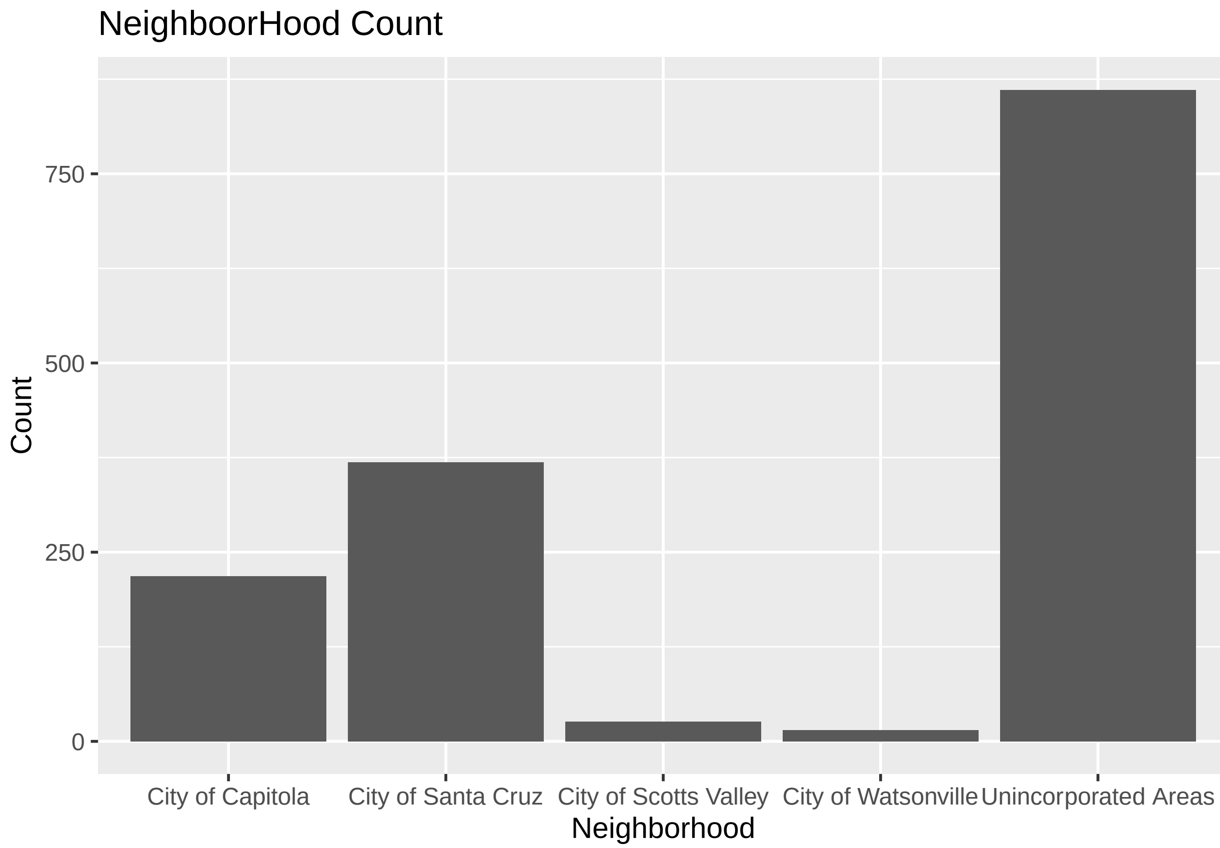

```
cleanairbnb %>%
  summarise(min = min(cleaning_fee),
            q1 = quantile(cleaning_fee)[2],
            median = median(cleaning_fee),
            q3 = quantile(cleaning_fee)[4],
            max = max(cleaning_fee),
            iqr = IQR(cleaning_fee),
            mean = mean(cleaning_fee),
            std_dev = sd(cleaning_fee)
            )
```

```
## # A tibble: 1 x 8
##     min    q1 median    q3   max   iqr  mean std_dev
##   <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl>   <dbl>
## 1  0.62  2.88      5  8.06    59  5.18  6.38    5.39
```

Next, let's examine the neighbourhood.

How many different categories of neighbourhood are in the dataset? Show code and output to support your answer. 22 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. City of Capitola, City of Santa Cruz, Unincorporated Areas 97.2%

```
ggplot(data = cleanairbnb, aes(x = neighbourhood)) + geom_bar() + labs(x = "Neighborhood",
     y = "Count",
    title = "NeighboorHood Count")
```




```
n_distinct(cleanairbnb$neighbourhood)
```

```
## [1] 5
```

```
cleanairbnb %>%
distinct(neighbourhood, id) %>%
group_by(neighbourhood) %>%
summarize("count" = n())
```

```
## # A tibble: 5 x 2
##   neighbourhood         count
##   <chr>                 <int>
## 1 City of Capitola        218
## 2 City of Santa Cruz      369
## 3 City of Scotts Valley    26
## 4 City of Watsonville      15
## 5 Unincorporated Areas    861
```

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.

```r
fourcleanairbnb <- cleanairbnb %>%
   mutate(neigh_simp = ifelse(neighbourhood == "City of Capitola", "City of Capitola",ifelse(neighbourho
```

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. 1,2,3,4 Some people allow people to stay by the month until they want to leave.

```r
n_distinct(cleanairbnb$minimum_nights)
```

```
## [1] 21
```

```r
cleanairbnb %>%
distinct(minimum_nights, id) %>%
group_by(minimum_nights) %>%
summarize("count" = n())
```

```
## # A tibble: 21 x 2
##    minimum_nights count
##             <dbl> <int>
##  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
## # ... with 11 more rows
```

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.

```r
pricefourcleanairbnb <- fourcleanairbnb %>%
  mutate(price_3_nights = cleaning_fee + (price * 3))
```

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.

```r
model <- lm(price_3_nights ~ neigh_simp + number_of_reviews + reviews_per_month , data = pricefourclean
summary(model)
```

```
##
## Call:
## lm(formula = price_3_nights ~ neigh_simp + number_of_reviews +
##     reviews_per_month, data = pricefourcleanairbnb)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1149.9  -454.9  -144.6   266.3  7801.4
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   1383.7515    59.6938  23.181  < 2e-16 ***
## neigh_simpCity of Santa Cruz   -241.5932    69.2460  -3.489   0.0005 ***
## neigh_simpOther               -691.2533   140.3976  -4.924 9.53e-07 ***
## neigh_simpUnincorporated Areas -260.2090    61.8741  -4.205 2.77e-05 ***
## number_of_reviews               -0.4522     0.2052  -2.204   0.0277 *
## reviews_per_month              -71.1334    12.3056  -5.781 9.21e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 762.2 on 1369 degrees of freedom
##   (114 observations deleted due to missingness)
## Multiple R-squared:  0.08445,    Adjusted R-squared:  0.0811
## F-statistic: 25.25 on 5 and 1369 DF,  p-value: < 2.2e-16
```

```
summary(model)$coefficient
```

```
##                                  Estimate  Std. Error   t value      Pr(>|t|)
## (Intercept)                   1383.751521  59.6937751 23.180834 1.503531e-100
## neigh_simpCity of Santa Cruz   -241.593239  69.2459847 -3.488913  5.003671e-04
## neigh_simpOther               -691.253292 140.3975934 -4.923541  9.532187e-07
## neigh_simpUnincorporated Areas -260.209008  61.8740937 -4.205460  2.774745e-05
## number_of_reviews               -0.452178   0.2051745 -2.203870  2.769941e-02
## reviews_per_month              -71.133445  12.3055651 -5.780592  9.206809e-09
```

```
confint(model)
```

```
##                                      2.5 %        97.5 %
## (Intercept)                   1266.6503413 1500.85270059
## neigh_simpCity of Santa Cruz   -377.4329724 -105.75350571
## neigh_simpOther               -966.6710181 -415.83556601
## neigh_simpUnincorporated Areas -381.5873152 -138.83070121
## number_of_reviews               -0.8546684   -0.04968754
## reviews_per_month              -95.2732517  -46.99363860
```

Interpret the coefficient of number_of_reviews and its 95% confidence interval in the context of the data. We are 95% confident that the true population mean is between -0.854 and -0.05. Interpret the coefficient of neigh_simpCity of Santa Cruz and its 95% confidence interval in the context of the data. We are 95% confident that the true population mean is between -377.43 and -105.75. Interpret the intercept in the context of the data. Does the intercept have a meaningful interpretation? Briefly explain why or why not. No it does not because that is when our predictor variables are all 0. 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. 322.35

```r
new_obs = data.frame(neigh_simp = 'Other', number_of_reviews = 10, reviews_per_month = 5.14)
predict(model, new_obs)
```

```
##        1
## 322.3505
```

Now check the assumptions for your regression model. Should you be confident on interpreting the inferential results of your model? The assumptions are not satisfied because there isn't a linear relationship between the variables.

```r
ggplot(data = pricefourcleanairbnb, aes(x = number_of_reviews, y = price_3_nights)) + geom_point() + lal
     y = "Price",
    title = "Number of Reviews x Price")
```