lab 5

Aisha Lakshman

2/18/2022

Packages

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.1.1     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(stringr)
library(knitr)
library(skimr)
library(broom)
```

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

Data wrangling & EDA

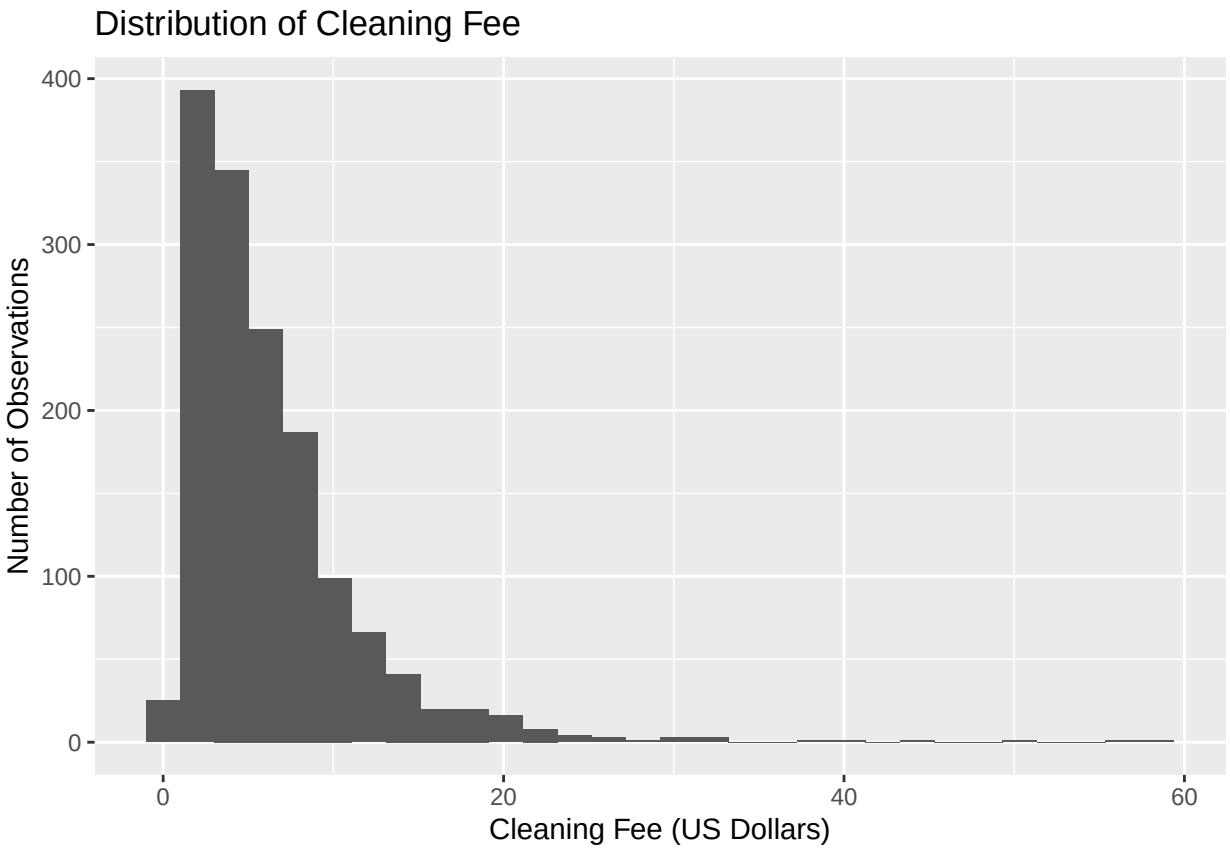
Exercise 1

```
airbnb <- airbnb %>% mutate(cleaning_fee = price * 0.02)
```

Exercise 2

```
ggplot(data = airbnb, aes(x = cleaning_fee)) +
  geom_histogram() +
  labs(x = "Cleaning Fee (US Dollars)", y = "Number of Observations", title = "Distribution of Cleaning
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
airbnb %>% summarise(mean = mean(cleaning_fee),
                     median = median(cleaning_fee),
                     std_dev = sd(cleaning_fee),
                     iqr = IQR(cleaning_fee))
```

```
##       mean median  std_dev  iqr
## 1 6.377582      5 5.394347 5.18
```

The histogram for cleaning fee indicates a skewed right distribution. According to the summary statistics, the mean value is 5 and the median value is approximately 6.4. The difference in these values are a result of outliers, as seen in the histogram. ## Exercise 3

```
count(airbnb, neighbourhood)
```

```
##           neighbourhood   n
## 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
```

There are 5 different categories of "neighborhood": City of Capitola, City of Santa Cruz, City of Scotts Valley, City of Watsonville, and Unincorporated Areas. The three most common neighborhoods in the

data are Unincorporated Areas, City of Santa Cruz, and City of Capitola. There are 1489 observations for "neighborhood", and these three neighbourhoods make up approximately 97.25 % of the data ((861 + 369 + 218)/1489 = 0.97246). ## Exercise 4

```
airbnb <- airbnb %>% mutate(neigh_simp = fct_lump_n(neighbourhood, n = 3))

count(airbnb, neigh_simp)
```

```
##             neigh_simp   n
## 1     City of Capitola 218
## 2   City of Santa Cruz 369
## 3 Unincorporated Areas 861
## 4                Other  41
```

Exercise 5

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

```
##    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
## 11             15   2
## 12             20   2
## 13             21   2
## 14             25   1
## 15             28  14
## 16             29   3
## 17             30  89
## 18             31  15
## 19             45   3
## 20             60   4
## 21             90   1
```

The four most common values for minimum nights are 1, 2, 3, and 30. The value of 30 minimum nights stands out the most, and seems slightly unusual without context. However, I think this value can be explained by airbnb listings suited for long-term residents, rather than tourists who generally stay 1-3 nights.

```
airbnb <- airbnb %>% filter(minimum_nights <= 3)
count(airbnb, minimum_nights)
```

```
##   minimum_nights   n
## 1              1 420
## 2              2 571
## 3              3 223
```

Regression

Exercise 6

```
airbnb <- airbnb %>% mutate(price_3_nights = 3 * price + cleaning_fee)
```

Exercise 7

```
model <- lm(price_3_nights ~ neigh_simp + number_of_reviews + reviews_per_month, data = airbnb)
tidy(model, conf.int = TRUE) %>%
  kable(format = "markdown", digits = 3)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	1475.380	65.136	22.651	0.000	1347.580	1603.181
neigh_simpCity of Santa Cruz	-208.001	75.923	-2.740	0.006	-356.966	-59.036
neigh_simpUnincorporated Areas	-312.632	65.758	-4.754	0.000	-441.652	-183.613
neigh_simpOther	-671.550	159.777	-4.203	0.000	-985.040	-358.059
number_of_reviews	-0.437	0.202	-2.158	0.031	-0.834	-0.040
reviews_per_month	-85.171	12.564	-6.779	0.000	-109.821	-60.520

Exercise 8

According to our model, the coefficient of number_of_reviews = -0.437 and the confidence interval for number_of_reviews = (-0.834,-0.040). The coefficient tells us that the price of a three night stay decreases by $0.437 for each review. The confidence interval tells us that if we repeat our sample data, 95% of the time the coefficient for number_of_reviews will fall in the range of (-0.834,-0.040). ## Exercise 9 According to our model, the coefficient of neigh_simpCity of Santa Cruz = -208.001 and the confidence interval for neigh_simpCity of Santa Cruz = (-356.966,-59.036). The coefficient tells us that the price for a three night stay in the City of Santa Cruz is on average $208 less than a three night stay in the City of Capitola. The confidence interval tells us that if we repeat our sample data, 95% of the time the coefficient for neigh_simpCity will fall in the range of (-356.966,-59.036). ## Exercise 10 The intercept in our model represents inferential statistics and the confidence interval for the City of Capitola. The intercept is meaningful because it acts as our "baseline", as the other inferential statistics and confidence intervals are dependent on the intercept. ## Exercise 11

```
1475.380 - 671.550 - 0.437 * 10 - 85.171 * 5.14
```

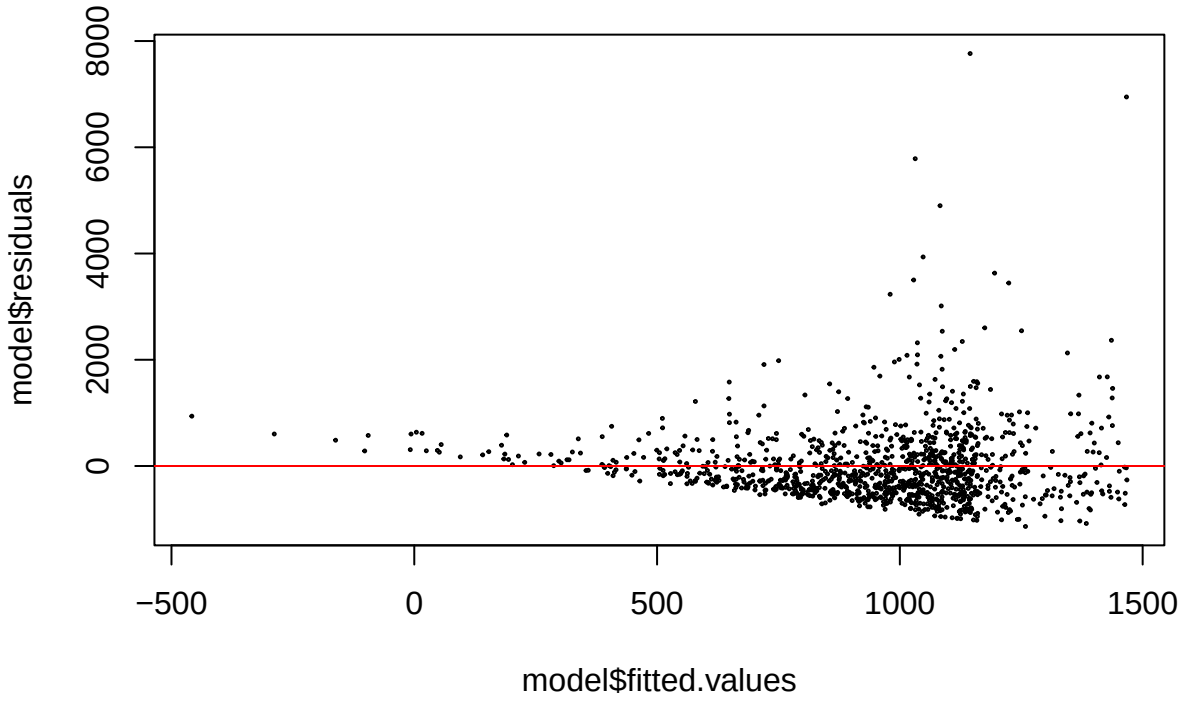
```
## [1] 361.6811
```

```
predic_data <- data.frame(neigh_simp = "Other", number_of_reviews = 10, reviews_per_month = 5.14)
predict(model, predic_data, interval = "confidence")
```

```
##        fit      lwr      upr
## 1 361.6874 59.63618 663.7387
```

Using the model, I predict a three night stay at a Scotts Valley airbnb with 10 reviews and 5.14 reviews per month would cost $361.6811. The 95% confidence interval is (59.63618, 663.7387). ## Exercise 12 To check the linearity assumption, we can inspect a plot of the residuals vs predictors for our model. Based on the plot below, I would say the linearity assumption is not satisfied because there is a clear fan-shaped pattern. To check the constant variance assumption, we can also inspect the plot below. I would say the constant variance assumption is not satisfied because there the points do not create a consistent shape from left to right.

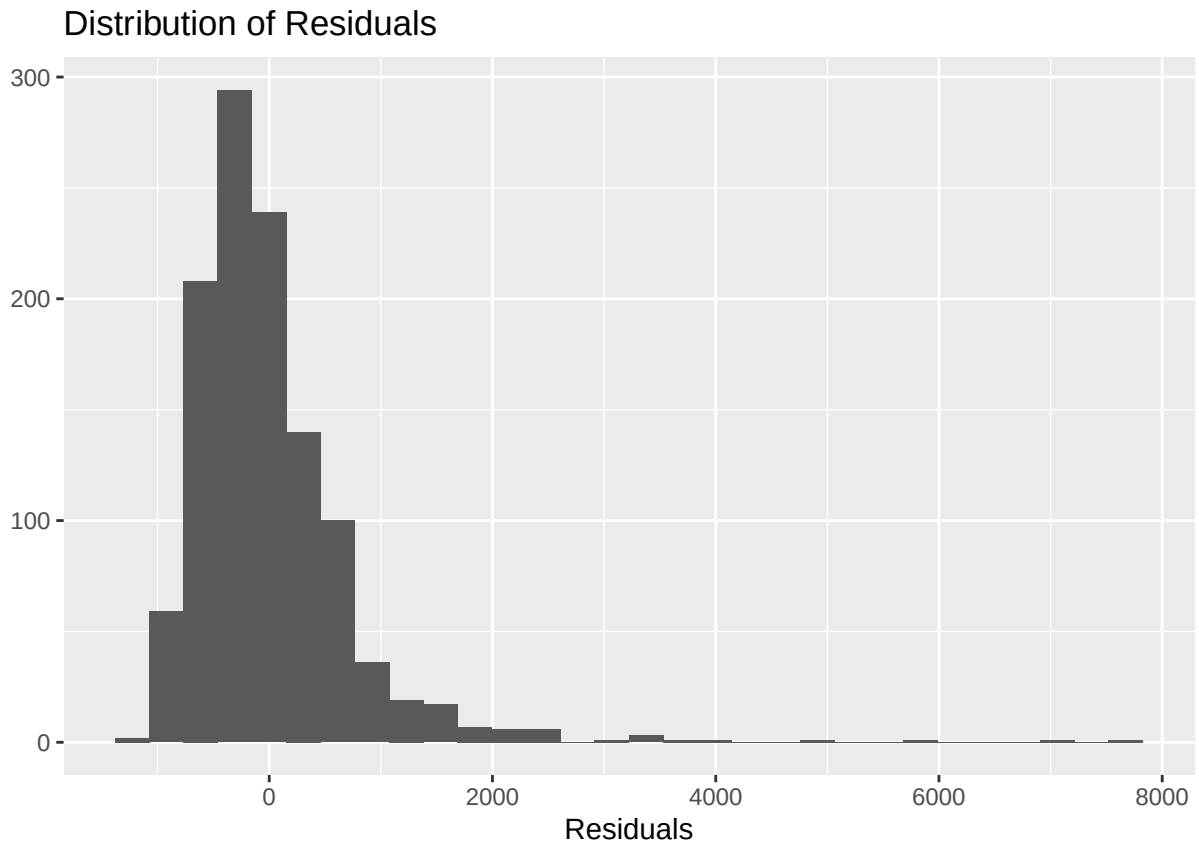
```
#residuals vs predicted
plot(model$fitted.values, model$residuals, cex=0.2)
abline(h = 0, col = "red")
```



To check the normality assumption for our model, we can create a histogram of the model residuals. Based on the plot below, I would say the normality assumption is not satisfied because the histogram does not follow a normal distribution.

```
ggplot(data = model, aes(x = model$residuals)) +
  geom_histogram() +
  labs(x = "Residuals", y = "", title = "Distribution of Residuals")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Based on these assumption checks, I would not be too confident in the interpretation based on the inferential results of this model.