Gurpinder Singh

STAT 108

11/9/2022

Load all the followin library

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.4
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.3      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

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

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

```
## Rows: 1627 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.
```

```
glimpse(airbnb)
```

```
## Rows: 1,627
## Columns: 18
## $ id                            <dbl> 8357, 24548, 31721, 43785, 54948, 57031~
## $ name                          <chr> "The Mushroom Dome Retreat & LAND of Pa~
## $ host_id                       <dbl> 24281, 99532, 136376, 191477, 258675, 4~
```

```
## $ host_name                       <chr> "Kitty And Michael", "Kerstin", "Annie"~
## $ neighbourhood_group             <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ neighbourhood                   <chr> "Unincorporated Areas", "City of Santa ~
## $ latitude                        <dbl> 37.00939, 36.97191, 36.95849, 36.97774,~
## $ longitude                       <dbl> -121.8855, -121.9973, -121.9721, -122.0~
## $ room_type                       <chr> "Entire home/apt", "Private room", "Ent~
## $ price                           <dbl> 159, 100, 239, 99, 335, 155, 98, 499, 1~
## $ minimum_nights                  <dbl> 2, 1, 4, 2, 2, 10, 3, 3, 1, 1, 3, 2, 2,~
## $ number_of_reviews               <dbl> 1724, 526, 293, 529, 123, 1, 490, 16, 5~
## $ last_review                     <date> 2022-09-25, 2022-09-05, 2022-09-26, 20~
## $ reviews_per_month               <dbl> 10.73, 3.48, 2.45, 3.60, 0.85, 0.02, 3.~
## $ calculated_host_listings_count <dbl> 2, 1, 2, 1, 2, 2, 1, 1, 2, 2, 1, 1, 2, ~
## $ availability_365                <dbl> 100, 0, 160, 348, 0, 26, 88, 0, 294, 10~
## $ number_of_reviews_ltm           <dbl> 125, 18, 48, 43, 5, 0, 61, 4, 80, 74, 3~
## $ license                         <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
```

```
spec(airbnb)
```

```
## cols(
##   id = col_double(),
##   name = col_character(),
##   host_id = col_double(),
##   host_name = col_character(),
##   neighbourhood_group = col_logical(),
##   neighbourhood = col_character(),
##   latitude = col_double(),
##   longitude = col_double(),
##   room_type = col_character(),
##   price = col_double(),
##   minimum_nights = col_double(),
##   number_of_reviews = col_double(),
##   last_review = col_date(format = ""),
##   reviews_per_month = col_double(),
##   calculated_host_listings_count = col_double(),
##   availability_365 = col_double(),
##   number_of_reviews_ltm = col_double(),
##   license = col_logical()
## )
```

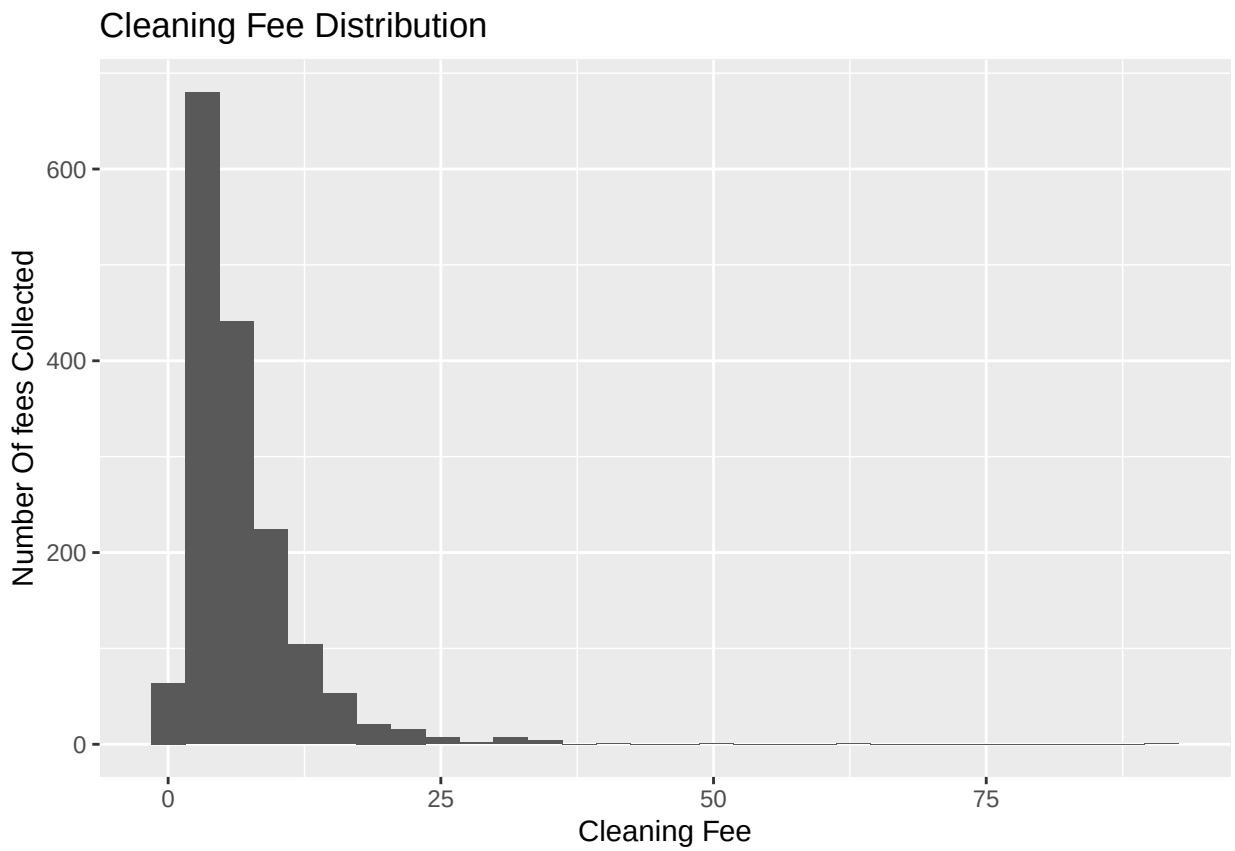
Exercise 1:

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

Exercise 2

```
ggplot(data = airbnb, aes(x =cleaning_fee)) +
  geom_histogram() +
  labs(x ="Cleaning Fee",
       y = "Number Of fees Collected",
       title = "Cleaning Fee Distribution")
```

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




```
airbnb %>%
  summarise(max = max(cleaning_fee),
            min = min(cleaning_fee),
            mean = mean(cleaning_fee),
            med = median(cleaning_fee),
            sd = sd(cleaning_fee),
            iqr = IQR(cleaning_fee))
```

```
## # A tibble: 1 x 6
##     max   min  mean   med    sd   iqr
##   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1  91.9  0.76  6.49   5.2  5.60  5.09
```

The following distribution shows a distribution with a right tail. Most values are between 0 to 25 with a mean of 6.5

Exercise 3

```
airbnb %>%
  group_by(neighbourhood) %>%
  summarise(obsvationCount = n())
```

```
## # A tibble: 5 x 2
##   neighbourhood         obsvationCount
##   <chr>                          <int>
```

```
## 1 City of Capitola                252
## 2 City of Santa Cruz              413
## 3 City of Scotts Valley            29
## 4 City of Watsonville              13
## 5 Unincorporated Areas            920
```

The following shows 5 different neighborhoods and more over it shows Unincorporated Areas, City of Santa, City of Capitola are the most observation they make up 97.48002459 percent of the observations

Exercise 4

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

Exercise 5

```
combinedData %>%
  group_by(minimum_nights) %>%
  summarise(obsvationCount = n())
```

```
## # A tibble: 21 x 2
##    minimum_nights obsvationCount
##             <dbl>          <int>
##  1              1            456
##  2              2            594
##  3              3            233
##  4              4             52
##  5              5             30
##  6              6             13
##  7              7             28
##  8             10              6
##  9             12              1
## 10             14              6
## # ... with 11 more rows
```

```
combinedData <- combinedData %>%
  filter(minimum_nights<= 3)
combinedData %>%
  group_by(minimum_nights) %>%
  summarise(obsvationCount = n())
```

```
## # A tibble: 3 x 2
##   minimum_nights obsvationCount
##            <dbl>          <int>
## 1              1            456
## 2              2            594
## 3              3            233
```

The four most common values for the variable minimmum nights is 1,2,3,30 The value 30 stands out as there is a drastic jump in observations at 30 days which is one month exactly. The likely purpose is to take into

account that the price will be different if they set minimum nights to 1 vs 14 since they want to give better deal for 14.

Exercise 6:

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

Exercise 7:

```
reg_model <- lm(price_3_nights ~ neigh_simp +number_of_reviews+ reviews_per_month, data = combinedData)
  tidy(reg_model, conf.int = TRUE) %>% # output model
  kable(digits = 3) # format model output
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	1488.215	59.830	24.874	0.000	1370.833	1605.597
neigh_simpCity of Santa Cruz	-236.340	68.253	-3.463	0.001	-370.246	-102.433
neigh_simpUnincorporated Areas	-298.390	60.142	-4.961	0.000	-416.384	-180.396
neigh_simpOther	-601.972	160.791	-3.744	0.000	-917.432	-286.511
number_of_reviews	-0.335	0.169	-1.984	0.047	-0.667	-0.004
reviews_per_month	-85.949	11.006	-7.809	0.000	-107.541	-64.356

```
coef(reg_model)
```

```
##                   (Intercept)   neigh_simpCity of Santa Cruz
##                  1488.2151911                   -236.3396034
## neigh_simpUnincorporated Areas                neigh_simpOther
##                   -298.3900021                   -601.9715299
##              number_of_reviews               reviews_per_month
##                     -0.3354853                    -85.9486577
```

Exercise 8: The coeffienient which indicates the number the price for three night stay at an airbnb will decrease or increase per review given on the airbnb rating. In this case it is .335 The confidence interval indicates that the magnitude decrease/increase of price per review will lie between -.667 and .004 with 95 percent confidence

Exercise 9: The coefficent of neigh_simp city of santa cruz is -236.340. This indicates that in comparrison to capitola the only one not shown in the tab, the price of an airbnb for three nights will be cheaper by -236.340.

Exercise 10: intercept indicates given neigh_simp= capitola number_of_reviews, and reviews_per_month are both zero the price for the airbnb will be predicted to be 1488.215.

Excersize 11: Estimate is 1488.215+10*(-0.335)+5.14*(-85.949)-601.972 = 441.11514. So it is estimated it will cost 441.12 dollars