lab05

REPO: https://github.com/theeho/lab05

```
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)
```

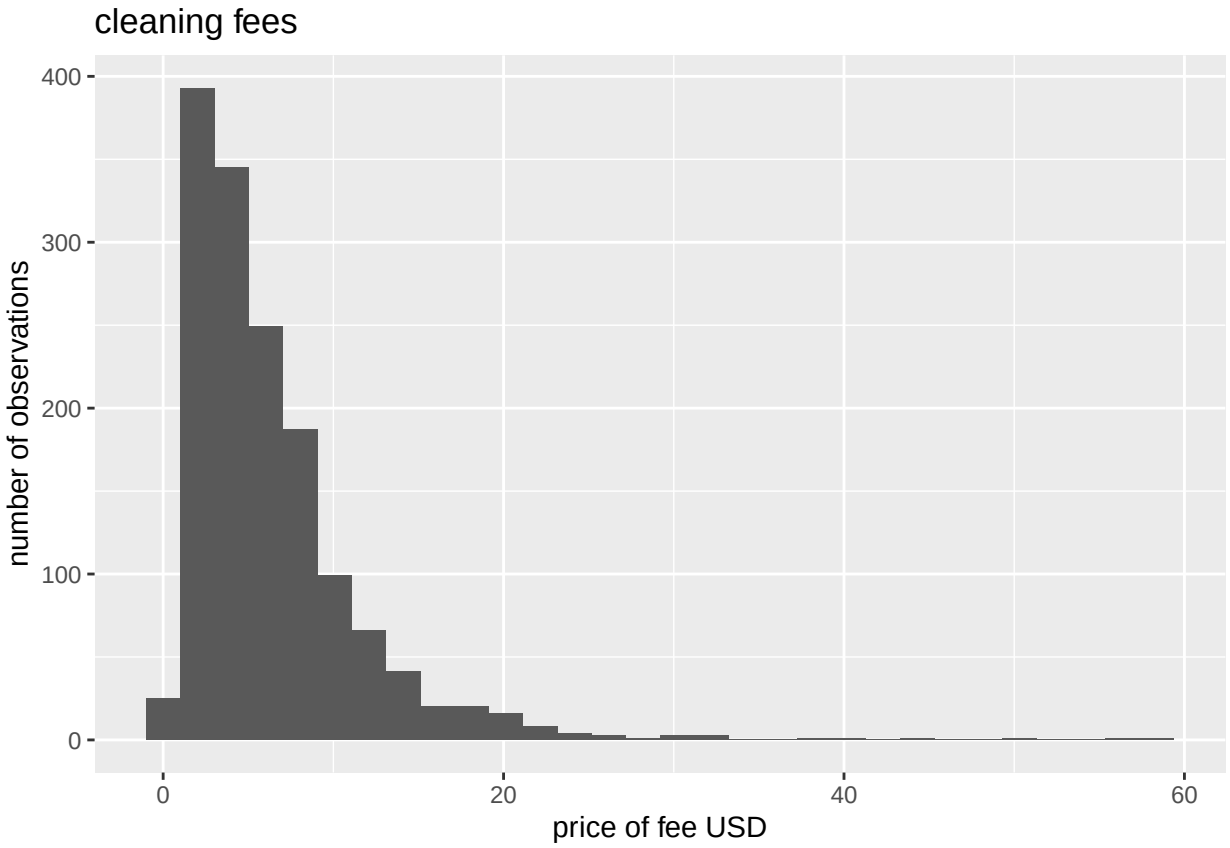
```
airbnb <- read_csv("listings.csv")
```

```
## 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.
```

Exercise 1/2

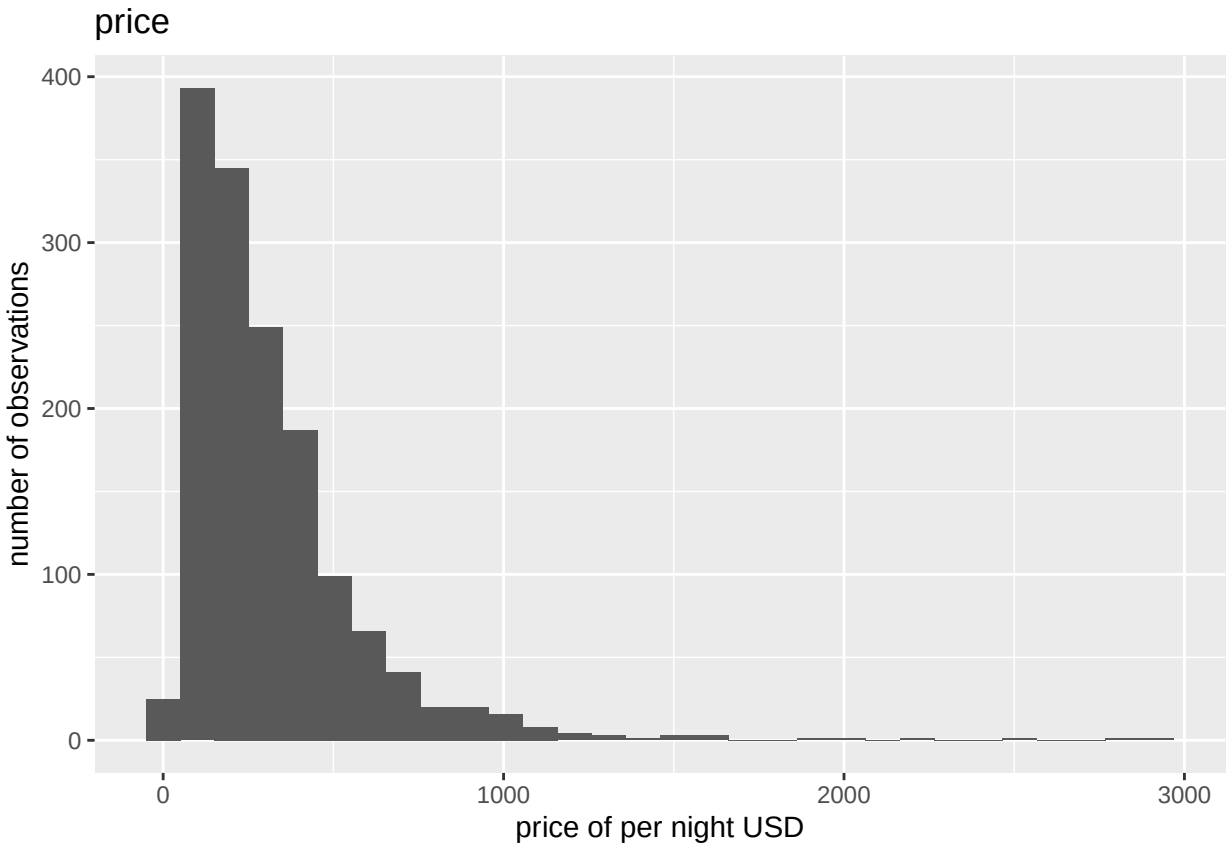
```
cleaning_fee <- transform(airbnb, fee = price * .02)
cplot<- ggplot(cleaning_fee, aes(x=fee)) + geom_histogram() + labs(title = "cleaning fees", x = "price
pplot<- ggplot(cleaning_fee, aes(x=price)) + geom_histogram() + labs(title = "price", x = "price of per
cplot
```

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




```
pplot
```

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



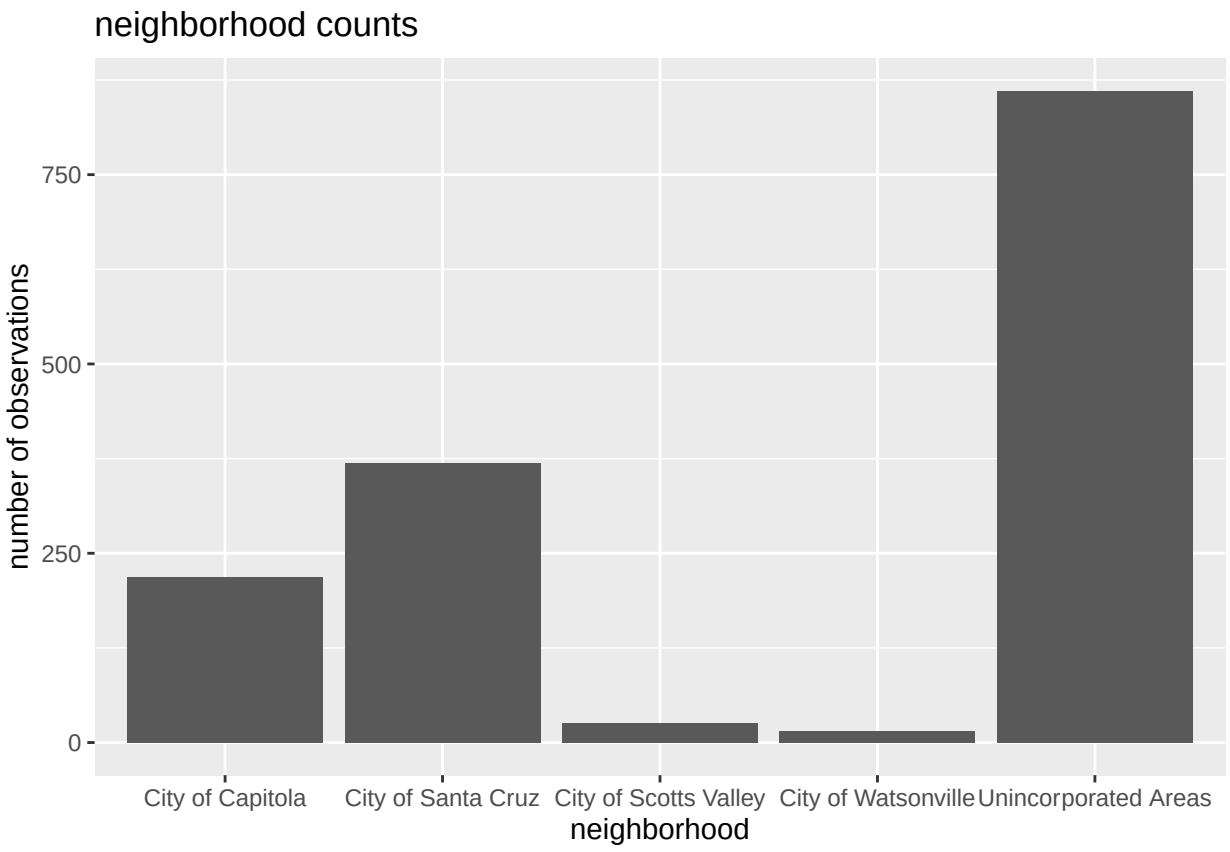

```
summary(cleaning_fee)
```

```
##        id                 name              host_id            host_name
##  Min.   :    8357   Length:1489        Min.   :     3177   Length:1489
##  1st Qu.:13436121   Class :character   1st Qu.: 15867956   Class :character
##  Median :28966253   Mode  :character   Median : 45862270   Mode  :character
##  Mean   :28997705                      Mean   : 90615670
##  3rd Qu.:46470409                      3rd Qu.:118394874
##  Max.   :54036247                      Max.   :436679905
##
##  neighbourhood_group neighbourhood         latitude       longitude
##  Mode:logical        Length:1489        Min.   :36.85   Min.   :-122.3
##  NA's:1489           Class :character   1st Qu.:36.96   1st Qu.:-122.0
##                      Mode  :character   Median :36.97   Median :-122.0
##                                         Mean   :36.99   Mean   :-122.0
##                                         3rd Qu.:36.99   3rd Qu.:-121.9
##                                         Max.   :37.19   Max.   :-121.7
##
##   room_type             price        minimum_nights  number_of_reviews
##  Length:1489        Min.   :  31.0   Min.   : 1.00   Min.   :   0.00
##  Class :character   1st Qu.: 144.0   1st Qu.: 1.00   1st Qu.:   8.00
##  Mode  :character   Median : 250.0   Median : 2.00   Median :  35.00
##                     Mean   : 318.9   Mean   : 4.86   Mean   :  84.18
##                     3rd Qu.: 403.0   3rd Qu.: 3.00   3rd Qu.: 105.00
##                     Max.   :2950.0   Max.   :90.00   Max.   :1623.00
```

```
##
##   last_review         reviews_per_month calculated_host_listings_count
##  Min.   :2014-03-24   Min.   : 0.010    Min.   : 1.000
##  1st Qu.:2021-09-24   1st Qu.: 0.650    1st Qu.: 1.000
##  Median :2021-11-27   Median : 1.640    Median : 1.000
##  Mean   :2021-08-17   Mean   : 2.298    Mean   : 7.021
##  3rd Qu.:2021-12-19   3rd Qu.: 3.305    3rd Qu.: 5.000
##  Max.   :2021-12-29   Max.   :13.540    Max.   :43.000
##  NA's   :114          NA's   :114
##  availability_365 number_of_reviews_ltm license             fee
##  Min.   :  0.0    Min.   :  0.00        Mode:logical   Min.   : 0.620
##  1st Qu.: 76.0    1st Qu.:  2.00        NA's:1489      1st Qu.: 2.880
##  Median :175.0    Median : 12.00                       Median : 5.000
##  Mean   :185.1    Mean   : 22.35                       Mean   : 6.378
##  3rd Qu.:312.0    3rd Qu.: 34.00                       3rd Qu.: 8.060
##  Max.   :365.0    Max.   :159.00                       Max.   :59.000
##
```

Exercise 3

```
np <- ggplot(airbnb, aes(x=neighbourhood)) + geom_bar() + labs(title = "neighborhood counts", x = "neigh
np
```




```
c <- airbnb %>% count(neighbourhood)
c
```

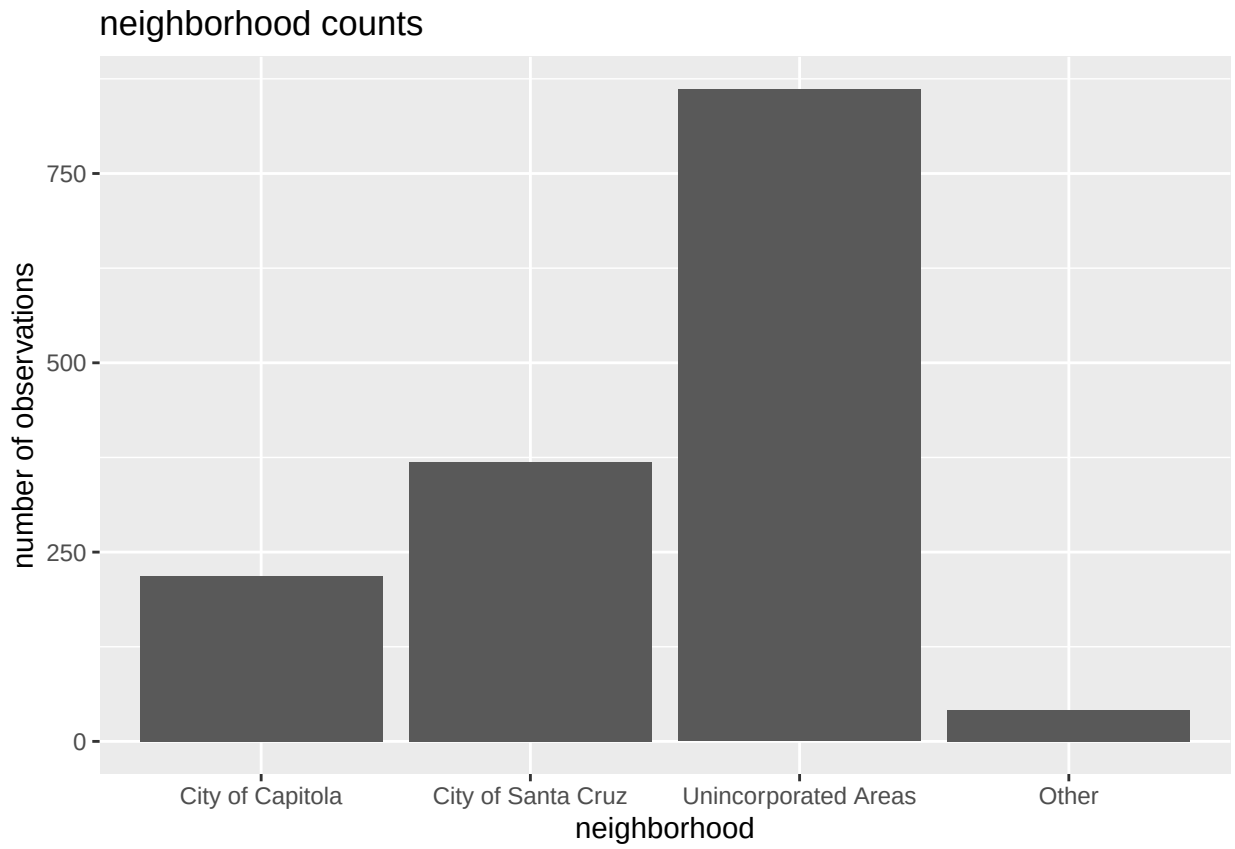
```
## # A tibble: 5 x 2
##   neighbourhood             n
##   <chr>                 <int>
## 1 City of Capitola        218
## 2 City of Santa Cruz      369
## 3 City of Scotts Valley    26
## 4 City of Watsonville      15
## 5 Unincorporated Areas    861
```

There are 5 different categories of neighbourhood.

Capitola, unicorperated areas, and Santa Cruz are the most common. They make up 97% of the population.

Exercise 4:

```
neigh_simp_df <- airbnb %>%
  mutate(neighbourhood = fct_lump(neighbourhood, n = 3, other_level = "Other"))

np_s <- ggplot(neigh_simp_df, aes(x=neighbourhood)) + geom_bar() + labs(title = "neighborhood counts", 
np_s
```




```
cs <- neigh_simp_df %>% count(neighbourhood)
cs
```

```
## # A tibble: 4 x 2
##   neighbourhood             n
```

```
##   <fct>                <int>
## 1 City of Capitola       218
## 2 City of Santa Cruz     369
## 3 Unincorporated Areas   861
## 4 Other                   41
```

Exercise 5:

```
cm <- neigh_simp_df %>% count(minimum_nights)
cm
```

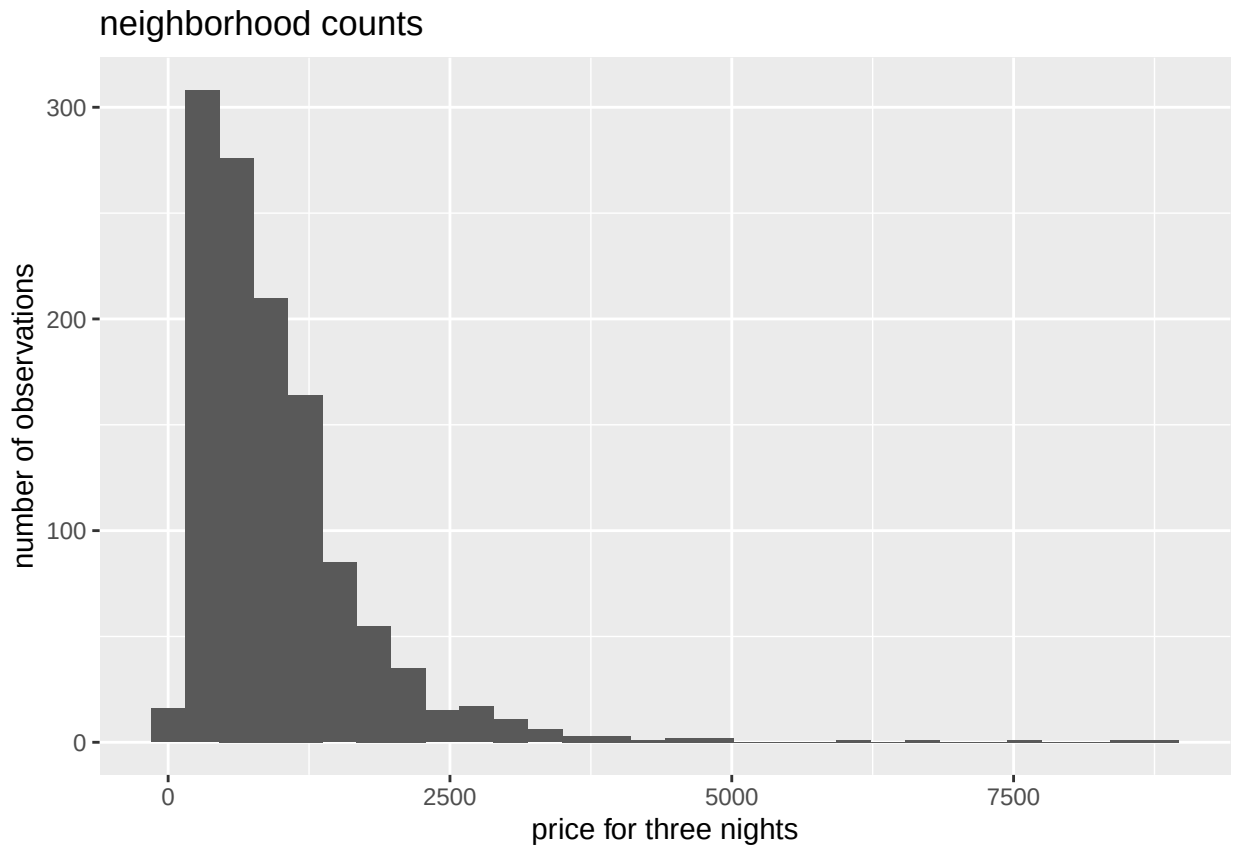
```
## # A tibble: 21 x 2
##    minimum_nights     n
##             <dbl> <int>
##  1              1   420
##  2              2   571
##  3              3   223
##  4              4    56
##  5              5    32
##  6              6    10
##  7              7    30
##  8              8     1
##  9             10     3
## 10             14     7
## # ... with 11 more rows
```

The 4 most common minimum nights are 1,2,3 and 4 nights. 1 minimum night stands out because it a minimum of 1 night implies that there can be less than 1 night. The intended purpose of this value is likely listings without a minimum.

Exercise 6

```
fm <- neigh_simp_df %>% filter(minimum_nights <= 3)
three_n <- transform(fm, price_three_nights = .02 * price + (price * 3))
np_tn <- ggplot(three_n, aes(x=price_three_nights)) + geom_histogram() + labs(title = "neighborhood cou
np_tn
```

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




```
summary(three_n)
```

```
##        id                 name              host_id           host_name
##  Min.   :    8357   Length:1214        Min.   :     3177   Length:1214
##  1st Qu.:13690600   Class :character   1st Qu.: 15958711   Class :character
##  Median :29047859   Mode  :character   Median : 47316700   Mode  :character
##  Mean   :29203488                      Mean   : 93836810
##  3rd Qu.:46688456                      3rd Qu.:126617361
##  Max.   :53932000                      Max.   :435944193
##
##  neighbourhood_group              neighbourhood    latitude
##  Mode:logical        City of Capitola    :178   Min.   :36.85
##  NA's:1214           City of Santa Cruz  :260   1st Qu.:36.96
##                      Unincorporated Areas:749   Median :36.97
##                      Other               : 27   Mean   :36.99
##                                                 3rd Qu.:37.01
##                                                 Max.   :37.19
##
##    longitude       room_type             price        minimum_nights
##  Min.   :-122.3   Length:1214        Min.   :  31.0   Min.   :1.000
##  1st Qu.:-122.0   Class :character   1st Qu.: 146.2   1st Qu.:1.000
##  Median :-122.0   Mode  :character   Median : 255.5   Median :2.000
##  Mean   :-122.0                      Mean   : 322.9   Mean   :1.838
##  3rd Qu.:-121.9                      3rd Qu.: 409.0   3rd Qu.:2.000
##  Max.   :-121.7                      Max.   :2950.0   Max.   :3.000
```

```
##
##  number_of_reviews  last_review         reviews_per_month
##  Min.   :   0.00   Min.   :2015-08-30   Min.   : 0.030
##  1st Qu.:  14.00   1st Qu.:2021-10-17   1st Qu.: 0.975
##  Median :  46.50   Median :2021-11-28   Median : 2.070
##  Mean   :  96.88   Mean   :2021-09-20   Mean   : 2.623
##  3rd Qu.: 125.50   3rd Qu.:2021-12-20   3rd Qu.: 3.725
##  Max.   :1623.00   Max.   :2021-12-29   Max.   :13.540
##                    NA's   :71           NA's   :71
##  calculated_host_listings_count availability_365 number_of_reviews_ltm
##  Min.   : 1.000                 Min.   :  0.0    Min.   :  0.00
##  1st Qu.: 1.000                 1st Qu.: 80.0    1st Qu.:  5.00
##  Median : 2.000                 Median :177.0    Median : 18.00
##  Mean   : 7.405                 Mean   :191.0    Mean   : 26.32
##  3rd Qu.: 6.000                 3rd Qu.:321.8    3rd Qu.: 39.00
##  Max.   :43.000                 Max.   :365.0    Max.   :159.00
##
##  license        price_three_nights
##  Mode:logical   Min.   :  93.62
##  NA's:1214      1st Qu.: 441.68
##                 Median : 771.61
##                 Mean   : 975.30
##                 3rd Qu.:1235.18
##                 Max.   :8909.00
##
```

Exercise 7:

```
m1 <- lm(price_three_nights~neighbourhood + number_of_reviews + reviews_per_month, data = three_n)
tidy(m1, conf.int = 95) %>% 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
neighbourhoodCity of Santa Cruz	-208.001	75.923	-2.740	0.006	-356.966	-59.036
neighbourhoodUnincorporated Areas	-312.632	65.758	-4.754	0.000	-441.652	-183.613
neighbourhoodOther	-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: Every 1 increase in number_of_reviews results in a .437 USD decrease in price. We are 95% confident the true paramater is between -.834 and -0.04

Exercise 9: If the neighbourhood of the listing is in Santa Cruz, then our price of the listing for 3 nights would be 208.001 USD less than the price if the listing was in Scotts Valley. We are 95% confident the true paramater is between -356.966 -59.036.

Exercise 10: The intercept has meaningful interpretation. It estimates the price for 3 nights for a listing in Scotts Valley with no reviews. This is a valid case.

Exercise 11: Estimated price for 3 nights = 1457.38 + 10(-0.437) + 5.14(-85.171) = 1015.23106 USD

Exercise 12:

I think there are a few concerning things about our asumptions. It seems that the number of reviews, and the reviews per month would be dependent on eachother. A higher number of reviews per month will always result in a higher number of total reviews. Also, the p-value for number_of_reviews is much higher than the others. I think it can be assumed that all observations would be independent of eachother given that the price of one listing will not directly affect the price of another. Based on these assumptions I would be cautious when using this model.