Packages

We will use the following packages in today's lab. You may need to install the package `skimr` using the `install.packages()` command in your console.

```
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   2.1.2      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)
```

Data

Today's data is about Airbnb listings in Santa Cruz, CA. The data can be obtained from http://insideairbnb.com/; it was originally scraped from airbnb.com.

You can see a visualization of some of the data used in today's lab at http://insideairbnb.com/santa-cruz-county/.

First, you need to download the file listings.csv from Santa Cruz in this website and save it on your computer

```
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.
```

We will use the following variables in this lab:

- `price`: Cost per night (in U.S. dollars)
- `neighbourhood`: Santa Cruz county City (Santa Cruz, Capitola, Scotts Valley, etc.)
- `room_type`:
  - *Entire home/apt* (guests have entire place to themselves)
  - *Private room* (Guests have private room to sleep, all other rooms shared)

  - *Shared room* (Guests sleep in room shared with others)

- `number_of_reviews`: Total number of reviews for the listing
- `reviews_per_month`: The number of reviews per month the listing has over the lifetime of the listing.
- `minimum_nights`: Number of nights required to book rental

Exercises

Data wrangling & EDA

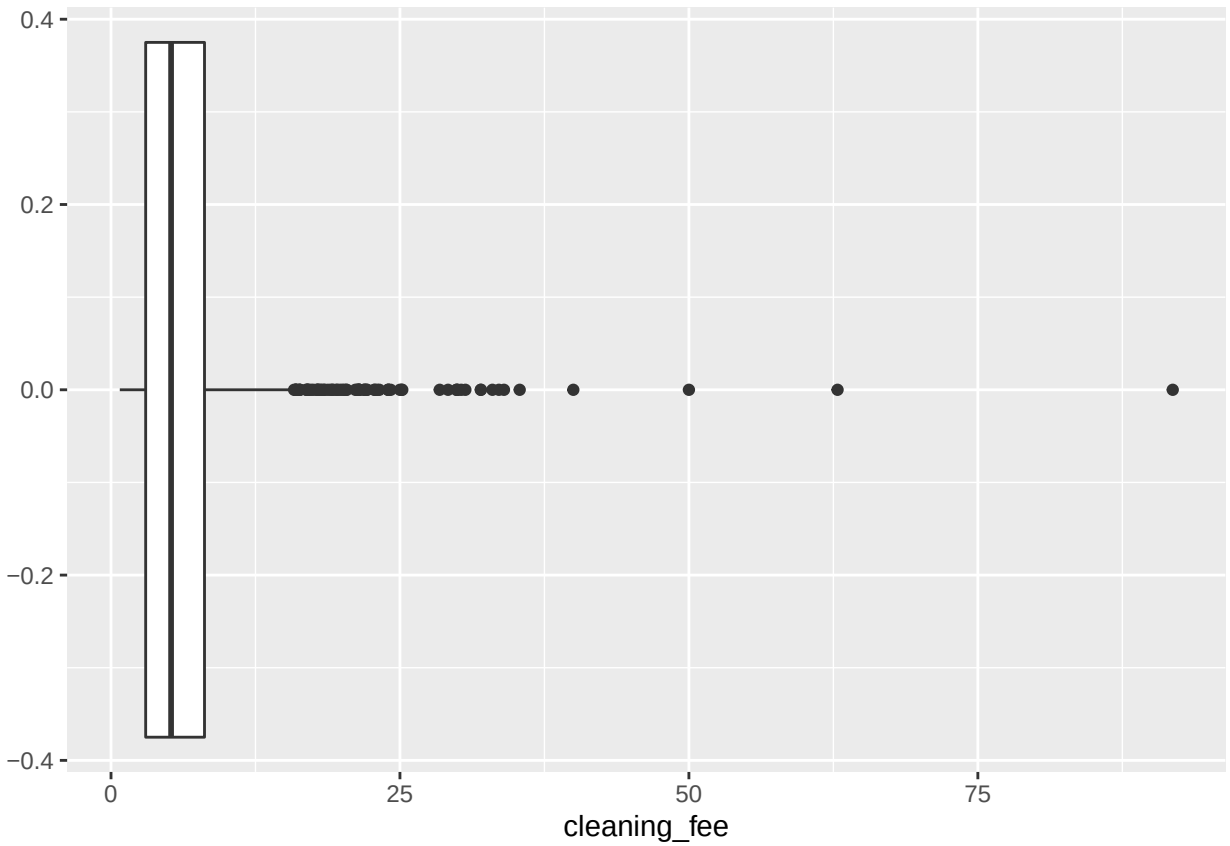
1. Some Airbnb rentals have cleaning fees, and we want to include the cleaning fee when we calculate the total rental cost. Create a variable call `cleaning_fee` calculated as the 2% of the price per night.

```
airbnb$cleaning_fee <- airbnb$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`.

ANSWER: The distribution of cleaning fee is right-skewed and have many outliers.

```
ggplot(airbnb, aes(x = cleaning_fee)) + geom_boxplot()
```



```
summary(airbnb$cleaning_fee)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.760   3.000   5.200   6.495   8.090  91.860
```

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.
    - 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.

ANSWER: The 3 most common neighborhoods are Unincorporated Areas, City of Santa Cruz, and City of Capitola. They make up 97.42% of the observations in the data.

```
airbnb %>%
  count(neighbourhood, sort = TRUE) %>%
  mutate(percentage = n / nrow(airbnb) * 100)
```

```
## # A tibble: 5 x 3
##   neighbourhood             n percentage
##   <chr>                 <int>      <dbl>
## 1 Unincorporated Areas    920     56.5
## 2 City of Santa Cruz      413     25.4
## 3 City of Capitola        252     15.5
## 4 City of Scotts Valley    29      1.78
## 5 City of Watsonville      13      0.799
```

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.

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

airbnb %>%
  count(neigh_simp, sort=TRUE)
```

```
## # A tibble: 4 x 2
##   neigh_simp               n
##   <fct>                <int>
## 1 Unincorporated Areas   920
## 2 City of Santa Cruz     413
## 3 City of Capitola       252
## 4 Other                   42
```

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.

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‘.

ANSWER: 1 - 3 and 30 are the most common value for minimum nights. “30” stands out in the top 4 because this value is much larger than others. Airbnb listings with 30 nights minimum requirement are usually for long-term renting.

```
airbnb %>%
  count(minimum_nights, sort=TRUE)
```

```
## # A tibble: 21 x 2
##    minimum_nights     n
##             <dbl> <int>
##  1              2   594
##  2              1   456
##  3              3   233
##  4             30   141
##  5              4    52
##  6              5    30
##  7              7    28
##  8             31    27
##  9             28    17
## 10              6    13
## # ... with 11 more rows
```

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

You will use this filtered dataset for the remainder of the lab.

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.*

```
airbnb$price_3_nights <- airbnb$cleaning_fee + 3 * airbnb$price
```

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.

```
linear_reg <- lm(price_3_nights
                 ~ neigh_simp
                 + number_of_reviews
                 + reviews_per_month, data = airbnb)
tidy(linear_reg) %>%
  kable(digits = 3)
```

term	estimate	std.error	statistic	p.value
(Intercept)	1488.215	59.830	24.874	0.000
neigh_simpCity of Santa Cruz	-236.340	68.253	-3.463	0.001
neigh_simpUnincorporated Areas	-298.390	60.142	-4.961	0.000
neigh_simpOther	-601.972	160.791	-3.744	0.000
number_of_reviews	-0.335	0.169	-1.984	0.047
reviews_per_month	-85.949	11.006	-7.809	0.000

```
confint(linear_reg, level=0.95)
```

```
##                                    2.5 %        97.5 %
## (Intercept)                   1370.8328888  1.605597e+03
## neigh_simpCity of Santa Cruz   -370.2464950 -1.024327e+02
## neigh_simpUnincorporated Areas -416.3839317 -1.803961e+02
## neigh_simpOther                -917.4322807 -2.865108e+02
## number_of_reviews                -0.6671755 -3.795131e-03
## reviews_per_month              -107.5410785 -6.435624e+01
```

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

ANSWER: Holding other variables constant, increase in one number of reviews, there will be an decrease of 0.33 dollars in prices for three nights for listings in the City of Capitola, and we are 95% confident that the true value is between -0.66 to -0.003.

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

ANSWER: Holding other variables constant, mean prices for three nights for listings in the City of Santa Cruz is 236.34 dollars cheaper than City of Capitola, and we are confident that 95% of true value of such difference is between -370.24 to -102.43.

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

ANSWER: When there's no reviews for the listing, the mean prices for three nights in the City of Capitola is 1488.21. This interpretation is meaningful but not helpful because it's possible that some new listings don't have reviews.

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.

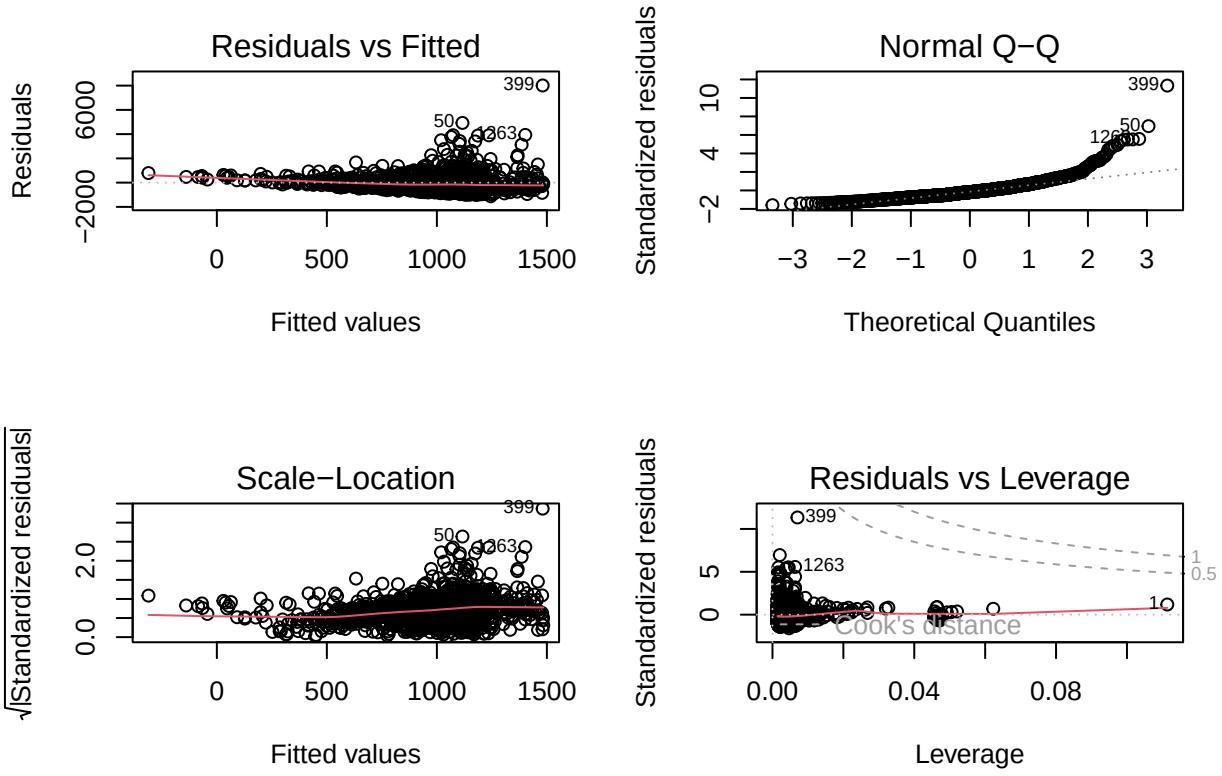
```
newdata <- data.frame(neigh_simp = "Other", number_of_reviews = 10, reviews_per_month = 5.14)
predict(linear_reg, newdata, interval = "confidence", level = 0.95)
```

```
##        fit      lwr      upr
## 1 441.1127 137.9465 744.2789
```

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

I'm not confident on interpreting the inferential results of your model. By looking at the plots below, we can identify some outliers. By the residual plots, the pattern and variance of residuals aren't random and constant respectively.

```
par(mfrow = c(2, 2))
plot(linear_reg)
```



You're done and ready to submit your work! Knit, commit, and push all remaining changes. You can use the commit message "Done with Lab 05!", and make sure you have pushed all the files to GitHub (your Git pane in RStudio should be empty) and that all documents are updated in your repo on GitHub. Then submit the assignment on Gradescope following the instructions below.

Submitting the Assignment

As before, what the instructor is going to check is your repo. Make sure to produce a pdf, and include it in your repo with the name Lab05.pdf. Also, include a folder called raw_data where the original data should be stored, and another folder called mod_data where the final version of your data table should be stored. Finally, a Readme.md should be created with a short description of this lab and the data.

```
write.csv(airbnb, "mod_data/airbnb.csv")
```

Acknowledgement

The data from this lab is from insideairbnb.com