Homework-2: *Your Name Here*

All of these homework problems are from the *R for Data Science* book. The section numbers (e.g., "3.2.4 Exercises") refer to sections in this book. Although the questions are based on those in the book, some questions ask for additional details or analysis.

When solving these problems, you are allowed to use any method from the book or class, even if that method wasn't yet covered when the exercise was presented in the book.

Write answers that are as complete as possible. If a graph is helpful for formalizing the solution, provide the graph. If a table is helpful, provide a table. In the text part of the answer, outline the progression in your thinking as you perform the analysis.

As discussed in the syllabus, if you look up answers online, or get answers from your fellow students, or from Generative AI, these sources need to be cited.

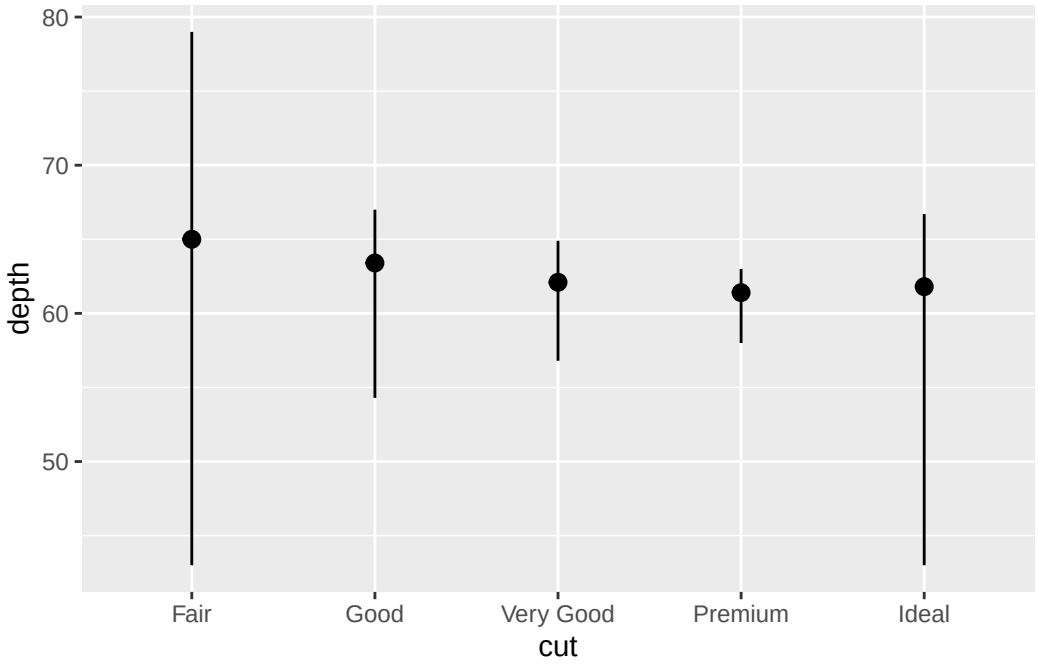
Note that you should type your answers in RStudio, by typing into the file **Homework-2.sa.Rmd**.

3.7.1 Exercises

(1) 3.7.1 Exercise 1, (10 pts)

Consider the following plot:

```
ggplot(data = diamonds) +
  stat_summary(
    mapping = aes(x = cut, y = depth),
    fun.min = min,
    fun.max = max,
    fun = median
    )
```



What is the default geom associated with `stat_summary()`?

How could you rewrite the plot code so that it drew the same graph, but used the default geom instead of `stat_summary()`?

(Note: Ed made use of `summarize()`, but other solutions may also exist.)

(2) 3.7.1 Exercise 2, (5 pts)

What does `geom_col()` do, and how is it different from `geom_bar()`?

(3) 3.7.1 Exercise 4, (5 pts)

What variables does `stat_smooth()` compute? What parameters control its behavior?

(4) 3.7.1 Exercise 5, (10 pts)

In our proportion bar chart, we must set `group = 1`. Why? In other words, what is wrong with these two plots?

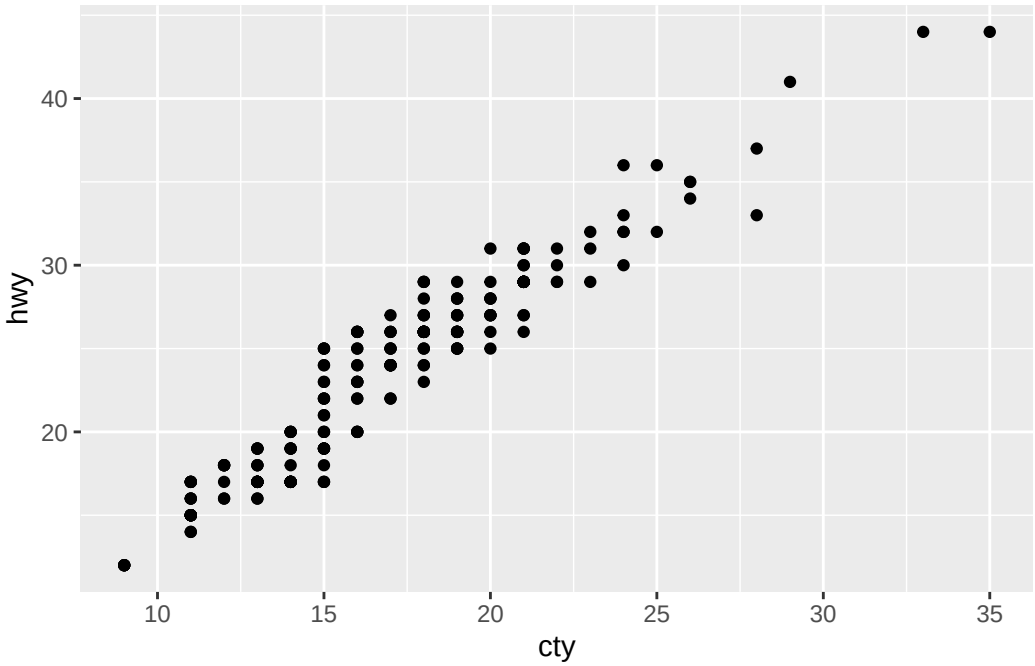
```
ggplot(data = diamonds) +
  geom_bar(mapping = aes(x = cut, y = after_stat(prop)))
ggplot(data = diamonds) +
  geom_bar(mapping = aes(x = cut, fill = color, y = after_stat(prop)))
```

3.8.1 Exercises

(5) 3.8.1 Exercise 1, (10 pts)

Given the context of this chapter, what is the problem with this plot? How could you improve it?

```
ggplot(data = mpg, mapping = aes(x = cty, y = hwy)) +
  geom_point()
```



(6) 3.8.1 Exercise 3, (10 pts)

Compare and contrast `geom_jitter()` with `geom_count()`. Demonstrate the kind of graph that `geom_count()` creates.

(7) 3.8.1 Exercise 4, (10 pts)

What is the default position adjustment for `geom_boxplot()`? Create a visualization of the `mpg` dataset that demonstrates it.

3.9.1 Exercises

(8) 3.9.1 Exercise 1, (10 pts)

Turn a stacked bar chart into a pie chart using `coord_polar()`.

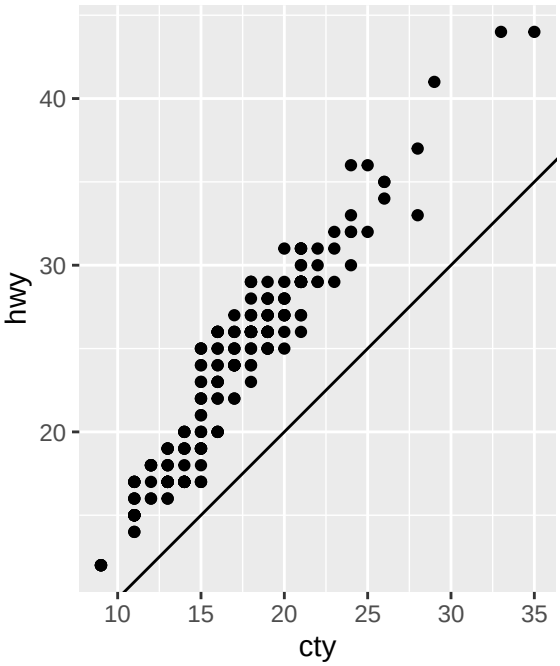
(9) 3.9.1 Exercise 3, (5 pts)

What is the difference between `coord_quickmap()` and `coord_map()`?

(10) 3.9.1 Exercise 4, (10 pts)

What does the plot below tell you about the relationship between city and highway mpg? Why is `coord_fixed()` important? What does `geom_abline()` do?

```
ggplot(data = mpg, mapping = aes(x = cty, y = hwy)) +
  geom_point() +
  geom_abline() +
  coord_fixed()
```



5.2.4 Exercises

(11) 5.2.4 Exercise 1, 70 pts (10 each))

For the `nycflights13::flights` dataset, find all flights that:

1. Had an arrival delay of two or more hours
2. Flew to Houston (`IAH` or `HOU`)
3. Were operated by United, American, or Delta
4. Departed in summer (July, August, and September)
5. Arrived more than two hours late, but didn't leave late
6. Were delayed by at least an hour, but made up over 30 minutes in flight
7. Departed between midnight and 6am (inclusive)

(12) 5.2.4 Exercise 4, (5 pts)

Why is `NA ^ 0` not missing? Why is `NA | TRUE` not missing? Why is `FALSE & NA` not missing? Can you figure out the general rule? (`NA * 0` is a tricky counterexample!)

5.3.1 Exercises

(13) 5.3.1 Exercise 1, (10 pts)

How could you use `arrange()` to sort all missing values to the start? (Hint: use `is.na()`).

(14) 5.3.1 Exercise 4, (10 pts)

Which flights traveled the longest distance? Which traveled the shortest?

5.4.1 Exercises

(15) 5.4.1 Exercise 1, (10 pts)

Brainstorm as many ways as possible to select `dep_time`, `dep_delay`, `arr_time`, and `arr_delay` from `flights`.

(16) 5.4.1 Exercise 3, (10 pts)

What does the `any_of()` function do? Why might it be helpful in conjunction with this vector?

```
vars <- c("year", "month", "day", "dep_delay", "arr_delay")
```

5.5.2 Exercises

(17) 5.5.2 Exercise 1, (10 pts)

Currently `dep_time` and `sched_dep_time` are convenient to look at, but hard to compute with because they're not really continuous numbers. Convert them to the more computationally convenient representation of number of minutes since midnight.

(18) 5.5.2 Exercise 2, (10 pts)

Compare `air_time` with `arr_time - dep_time`. What do you expect to see? What do you see? What do you need to do to fix it?