Week 3 Reading Guide: Adding in Categorical Variables

## IMS – Chapter 4, Section 6

In Figure 4.8, which data display do you prefer?

What is a strength of the stacked boxplots?

What is a weakness of the stacked boxplots?

What is a strength of stacked histograms?

What is a weakness of stacked histograms?

What is a **ridge plot**?

## Chapter 3 – Data Wrangling

## The Pipe (%>%)

How would you write the following pipe operation in **one** line?

x %>%
 f() %>%
 g() %>%
 h()

|  |  |
| --- | --- |
|  | Think about howx %>%  f() can be written as f(x). |

## filter()

What does the filter() function do?

How would you describe the following filter() statement **in words**?

filter(origin == "JFK" | dest == "SEA",
 month >= 10)

What does the %in% do in the following code?

filter(dest %in% c("SEA", "SFO", "PDX", "BTV", "BDL")

## summarize()

What does the summarize() function do?

You will write lines of code that look like:

 summarize(mean = mean(temp, na.rm = TRUE),
 std\_dev = sd(temp, na.rm = TRUE))

What does the name (e.g., mean, std\_dev) on the right hand side of the = sign correspond to?

What does the code (e.g., mean(temp, na.rm = TRUE)) on the left hand side of the = sign correspond to?

## group\_by()

What does the group\_by() function do?

When you group\_by() a variable before summarize() what happens?

What happens when you group by more than one variable in group\_by()?

## mutate()

What does the mutate() function do?

Suppose we wanted to create a new variable in the penguins dataset that took the original body\_mass\_g variable and converted it to kilograms. How would that code look?

## arrange()

What does the arrange() function do?

What order are the rows organized by default when you use arrange()?

What would you do if you wanted the rows to be organized in descending order?