Lab 6 – Practice with `dplyr` and `ggplot2`

2024-10-12

```
knitr::opts_chunk$set(echo = TRUE,
                      fig.align = 'center',
                      fig.width = 4,
                      fig.height = 3,
                      message = FALSE,
                      warning = FALSE)
```

```
library(ggplot2)
library(dplyr)

## Less uggo plots
theme_set(theme_bw())
```

Question 1 This question uses the penguins dataset.

```
penguins <- read.csv("https://collinn.github.io/data/penguins.csv")
```

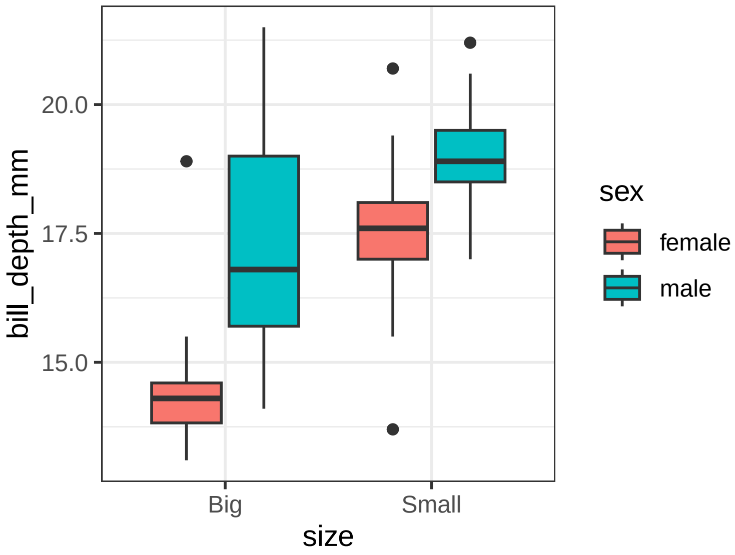
1. Using the appropriate `dplyr` functions, first filter the penguin dataset to only include observations in which the variable `sex` takes the values "male" and "female" (a handful of observations are simply missing entries). Then, create a new variable in this dataset called `size` that takes the value "Small" if the penguin's body mass is less than the median and the value "Big" if the body mass is greater than or equal to the median
2. Create a boxplot illustrating the relationship between dill depth, size, and sex. Do small or large penguins tend to have deeper bills? Which groups appears to have the greatest amount of variability?
3. The code below will take the variable `size` and re-code it as a factor variable. The utility here comes in the `levels` argument: doing this before creating a plot allows you to modify the order in which the values in a plot

```
penguins <- mutate(penguins, size = factor(size, levels = c("Small", "Big")))
```

Recreate the plot and compare it to what you saw in (2.). What has changed? Note that this will work for all future categorical variables when creating ggplots.

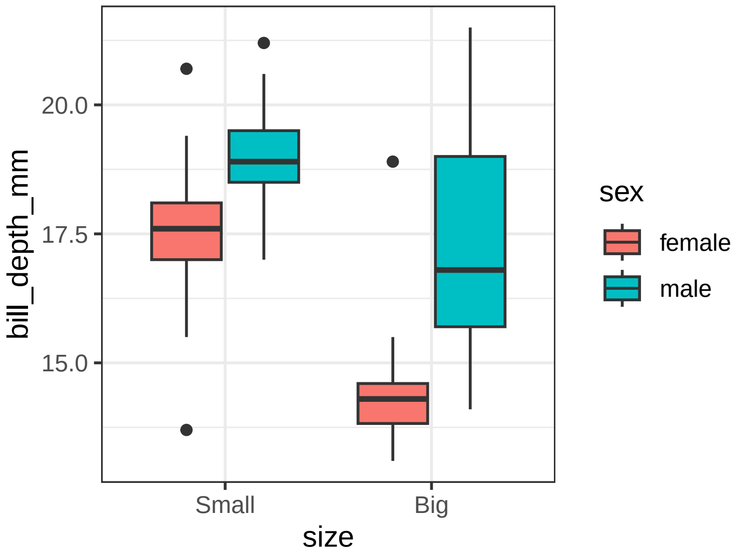
```
## Part 1
penguins <- penguins %>%
  filter(sex %in% c("male", "female")) %>%
  mutate(size = ifelse(body_mass_g < median(body_mass_g), "Small", "Big"))

## Part 2
(pengplot1 <- ggplot(penguins, aes(size, bill_depth_mm, fill = sex)) +
  geom_boxplot())
```



```
## Part 3 -- the order of the categorical variables has changed
penguins <- mutate(penguins, size = factor(size, levels = c("Small", "Big")))

(pengplot2 <- ggplot(penguins, aes(size, bill_depth_mm, fill = sex)) +
  geom_boxplot())
```

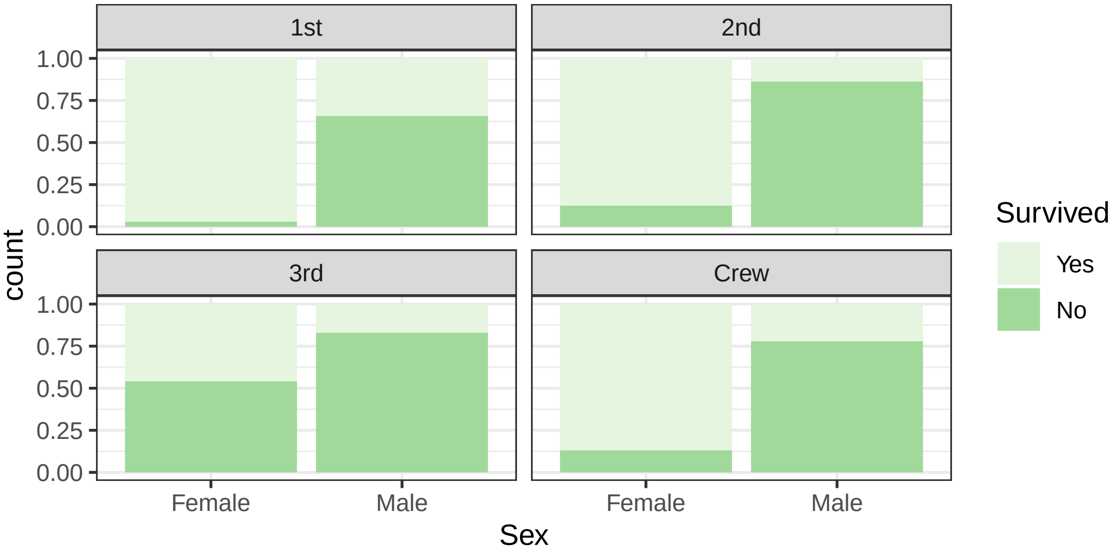


Question 2 This question will use the titanic dataset.

```
data(Titanic)
titanic <- as.data.frame(Titanic)
titanic <- titanic[rep(1:nrow(titanic), times = titanic$Freq), ]
titanic$Freq <- NULL
```

Recreate the plot below by *refactoring* the levels as we did in Question 1. Pay special note to the order of values presented for Sex, Class, and Survived.

```
titanic <- mutate(titanic, Survived = factor(Survived, levels = c("Yes", "No")),
                  Sex = factor(Sex, levels = c("Female", "Male")),
                  Class = factor(Class, levels = rev(c("Crew", "3rd", "2nd", "1st"))))
ggplot(titanic, aes(Sex, fill = Survived)) +
  geom_bar(position = "fill") +
  scale_fill_brewer(palette = "Greens") +
  facet_wrap(~Class)
```



Question 3 This question uses the college dataset.

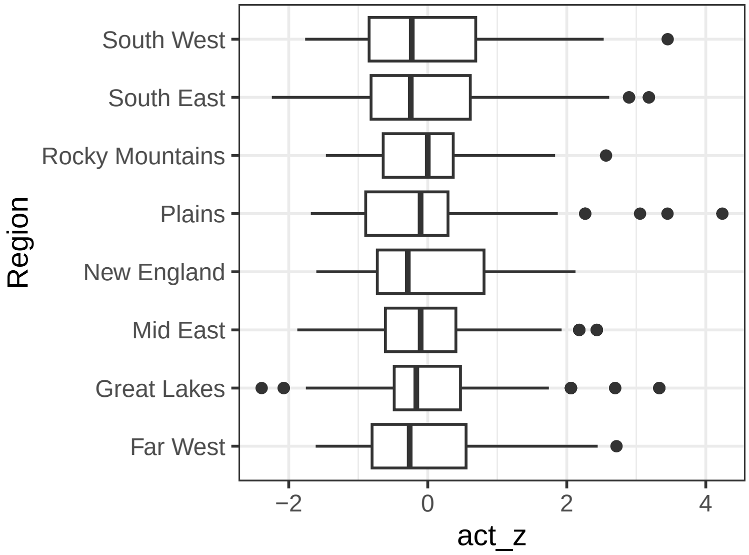
```
## College data
college <- read.csv("https://collinn.github.io/data/college2019.csv")
```

Here, we are interested in answering the question: "which region has the largest outlier *relative to the region* for median ACT score"

1. First, standardize the `ACT_median` variable *by region* and then create the appropriate boxplot showing the relationship between region and standardized median ACT score (Hint: how do you find standardized values of a variable?). Which region appears to have the greatest outlier?
2. Next, create a summary of the college dataset that shows the maximum and minimum median ACT values for each region. Which region has the largest median ACT? Which has the smallest?

```
## Part 1: Which region has the most positive outliers?
cc <- college %>%
  group_by(Region) %>%
  mutate(meanAct = mean(ACT_median),
         sdAct= sd(ACT_median),
         act_z = (ACT_median - meanAct) / sdAct)

## Should be plains
ggplot(cc, aes(y = Region, act_z))+
  geom_boxplot()
```



```
## Part 2
college %>% group_by(Region) %>%
  summarize(maxact = max(ACT_median),
            minact = min(ACT_median))
```

```
## # A tibble: 8 x 3
##   Region          maxact minact
##   <chr>            <int>  <int>
## 1 Far West            34     18
## 2 Great Lakes         34     16
## 3 Mid East            34     17
## 4 New England         35     18
## 5 Plains              34     19
## 6 Rocky Mountains     31     20
## 7 South East          34     15
## 8 South West          34     17
```

3. Compare your answers from (2) with what you found in (1). Does having the largest/smallest ACT values necessarily result in having outliers? Explain what is happening.

Question 4 This question uses the tips dataset

```
tips <- read.csv("https://collinn.github.io/data/tips.csv")
```

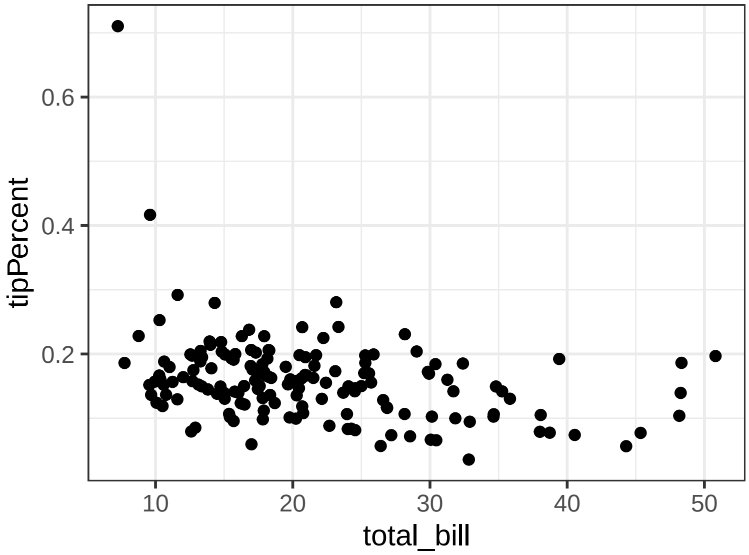
1. Begin by using the appropriate `dplyr` functions to filter the dataset to only include bills that occurred on either Saturday or Sunday and in which the size of the party was greater than 1.
2. Create a new variable, `tipPercent` that finds what percentage of the total bill was offered as a tip
3. Construct a summary that includes the average tipping percentage of an individual by sex and smoker status. Along with this summary, also include the number of individuals making up each group (`n()`)
4. Create the appropriate plot to illustrate the relationship between total bill (x-axis) and tip percent (y-axis). What type of relationship do you see? Would Pearson's correlation or Spearman's correlation be more appropriate to describe this?

```
## Part 1 and 2
tt <- tips %>% filter(size > 1, day %in% c("Sat", "Sun")) %>%
  mutate(tipPercent = tip / total_bill)

## Part 3 - Who seems to be a better tipper?
group_by(tt, sex, smoker) %>% summarize(mean(tipPercent), N = n())
```

```
## # A tibble: 4 x 4
## # Groups:   sex [2]
##   sex    smoker `mean(tipPercent)`     N
##   <chr>  <chr>               <dbl> <int>
## 1 Female No                  0.158    26
## 2 Female Yes                 0.171    18
## 3 Male   No                  0.160    75
## 4 Male   Yes                 0.152    42
```

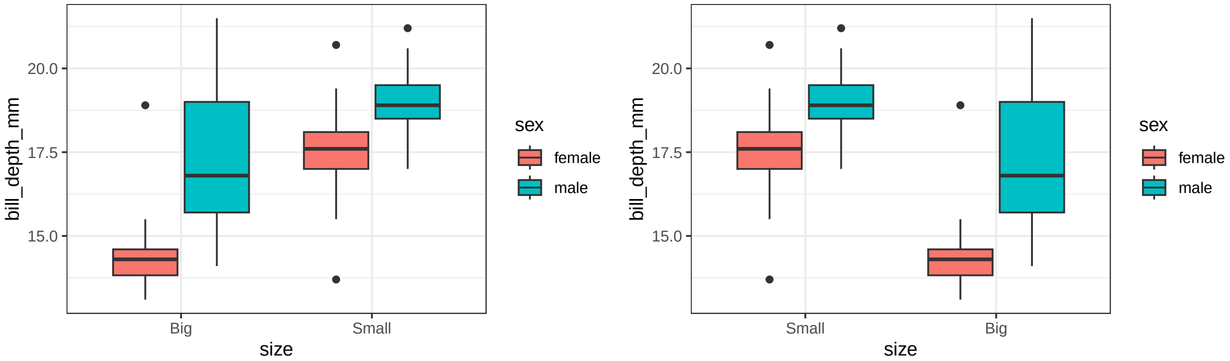
```
## Part 4 - Relationship between total bill and percent tip? (Spearman better)
ggplot(tt, aes(total_bill, tipPercent)) +
  geom_point()
```



Extending ggplot

Question 5 Using `grid.arrange()`, reproduce the two plots you created in Q1 to show side-by-side how using `factor()` and `levels` can allow you to rearrange your plots.

```
library(gridExtra)
grid.arrange(pengplot1, pengplot2, nrow = 1)
```



Question 6 Recreate the following plot that shows the mean and median net tuition for each of the regions in the college dataset

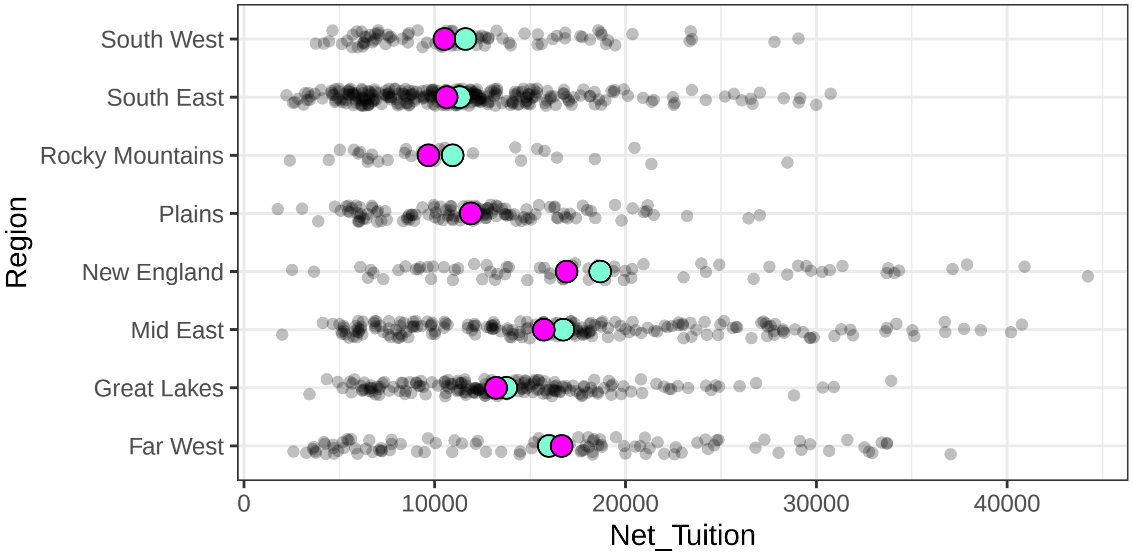
```
college <- read.csv("https://collinn.github.io/data/college2019.csv")
```

Note that:

- Jitter function has `height = 0.15` and `alpha = 0.25`
- The colors are "**a**quaremarine" for **a**verage and "**m**agenta" for **m**edian
- The size of the colored points is `size = 3.5`

```
cc <- group_by(college, Region) %>%
  summarize(meanDebt = mean(Net_Tuition),
            medianDebt = median(Net_Tuition))

ggplot(college, aes(y  = Region, Net_Tuition)) +
  geom_jitter(height = 0.15, alpha = 0.25) +
  geom_point(data = cc, aes(y = Region, x = meanDebt), fill = 'aquamarine', size = 3.5,
             shape = 21, color = 'black') +
  geom_point(data = cc, aes(y = Region, x = medianDebt), fill = 'magenta', size = 3.5,
             shape = 21, color = 'black')
```


Region
South West
South East
Rocky Mountains
Plains
New England
Mid East
Great Lakes
Far West
0
10000
20000
30000
40000
Net_Tuition

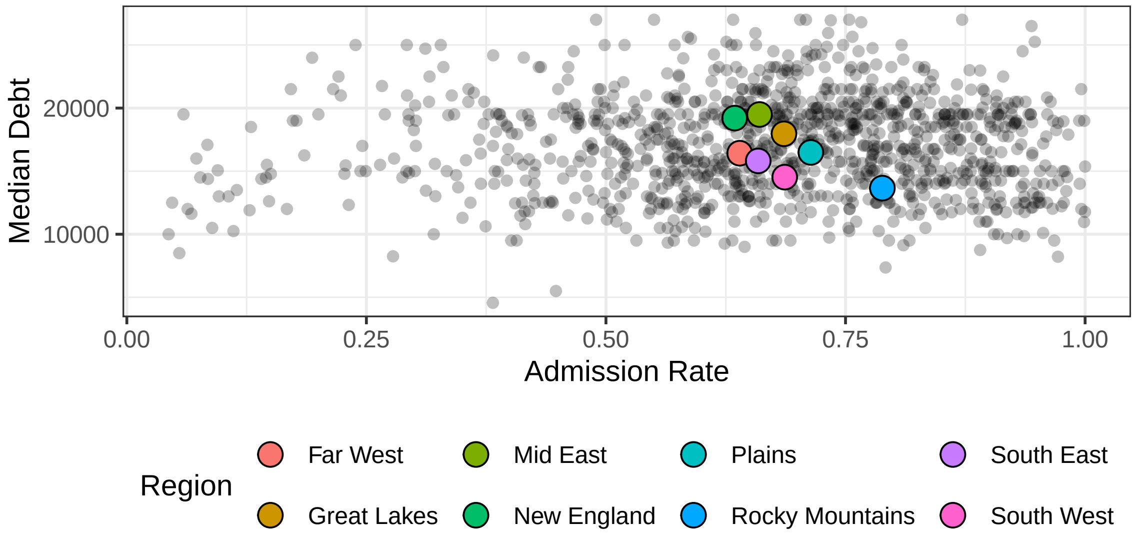
Question 7 Recreate this from the ecological correlation lecture, where the individual points are the schools and the colored dots get their x and y values from the regions' average admission rate and median debt, respectively

Note that:

- `alpha = 0.25` for the black dots
- `size = 4` for the colored dots
- You can add `theme(legend.position = "bottom")` to move the legend

```
cc <- group_by(college, Region) %>%
  summarize(adm_rate = mean(Adm_Rate),
            dm = mean(Debt_median))

ggplot(college, aes(Adm_Rate, Debt_median)) +
  geom_point(alpha = 0.25) +
  geom_point(aes(adm_rate, dm, fill = Region), data = cc, pch = 21, color = "black",
             size = 4) +
  labs(x = "Admission Rate", y = "Median Debt") +
  theme(legend.position = "bottom")
```


Median Debt
Admission Rate
Region
● Far West
● Great Lakes
● Mid East
● New England
● Plains
● Rocky Mountains
● South East
● South West