Practice1S24

Sofia Guttmann

2024-02-12

Practice1 - Due Thursday, 2/8 by midnight to Gradescope

Reminder: Practice assignments may be completed working with other individuals.

For academic integrity, the second page of this assignment provides a place where you will list who you worked with (and for which problems) as well as outside resources that you used (meaning resources besides our textbook, course materials, and default R help for functions/packages.) The problems begin after that. (Next week, this page will be the cover page and the workflow review will be removed.)

Reading

The associated reading for the week is Chapter 2, Chapter 3, and Section 8.2.

Git Workflow Review

1. Before editing this file, verify you are working on the copy saved in *your* repo for the course (check the filepath and the project name in the top right corner).
2. Before editing this file, make an initial commit of the file to your repo to add your copy of the problem set.
3. Change your name at the top of the file and get started!
4. You should *save, knit, and commit* the file each time you've finished a question, if not more often.
5. You should occasionally *push* the updated version of the file back onto GitHub.

6. When you think you are done with the assignment, save the pdf as "*YourFirstInitialYourLastName_thisfilename*.pdf" before committing and pushing (this is generally good practice but also helps me in those times where I need to download all student homework files). For example, I would save this file as AWagaman_Practice1.pdf.

Practicing Academic Integrity

If you worked with others or used resources outside of provided course material (anything besides our textbook, course materials in the repo, labs, R help menu) to complete this assignment, please acknowledge them below using a bulleted list.

I acknowledge the following individuals with whom I worked on this assignment:

Name(s) and corresponding problem(s)

- Sofia Guttmann

I used the following sources to help complete this assignment:

Source(s) and corresponding problem(s)

-

1 - Enhancing a plot

The *mpg* data set is available in R (use the help file to learn more about it). We are interested in examining the relationship between the variables *hwy* and *cty* across the variable *drv*.

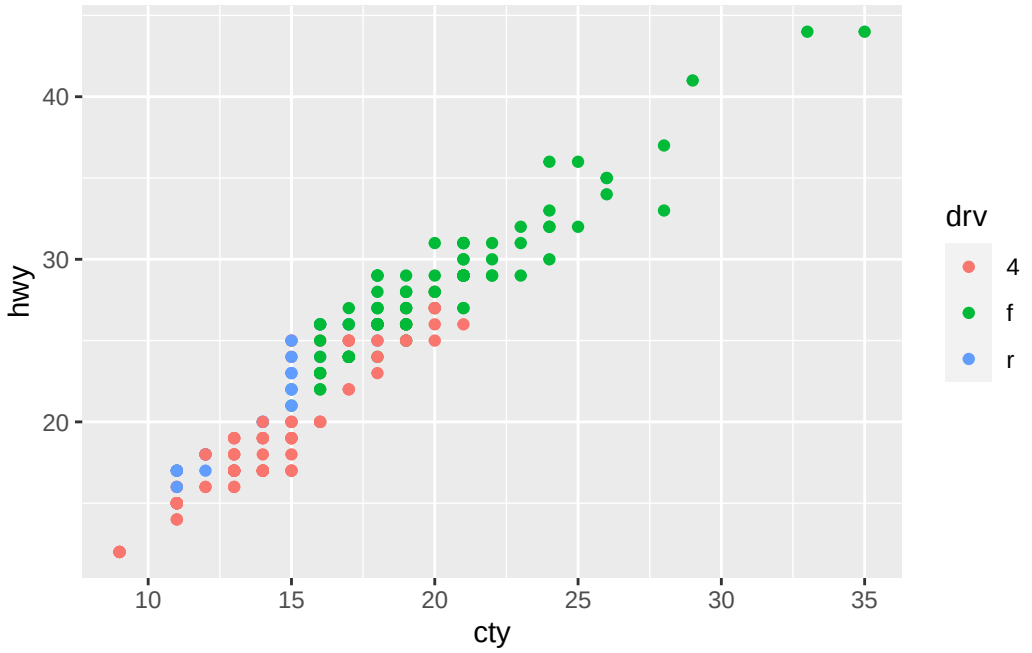
A preliminary plot is provided which needs enhanced. Add new code chunks to make new plots as described below.

```
# preliminary plot, eval: false means plot will not show in pdf; set yours to true
# or remove option in chunks below

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

part a - Add *drv* to the preliminary plot using color or size.

Solution:

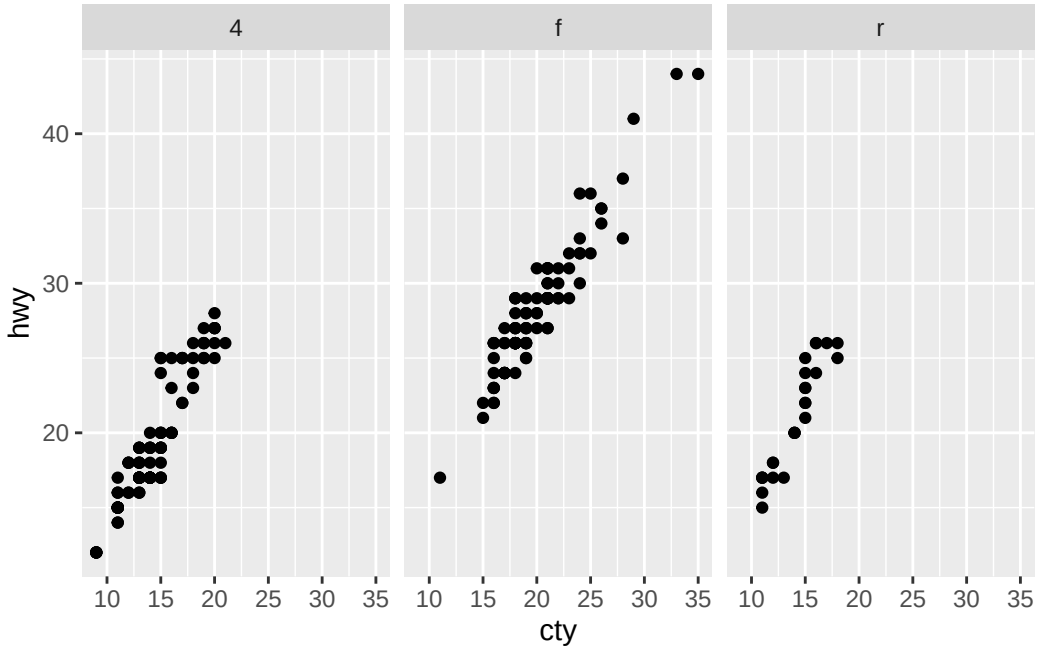
```
library(ggplot2)

g <- ggplot(data = mpg, mapping = aes(x = cty, y = hwy, color = drv)) +
  geom_point()
g
```



part b - Use facets to incorporate *drv* to the preliminary plot instead of using either color or size.

Solution:

```
library(ggplot2)

g <- ggplot(data = mpg, mapping = aes(x = cty, y = hwy)) +
  geom_point() +
  facet_wrap(~ drv)
g
```



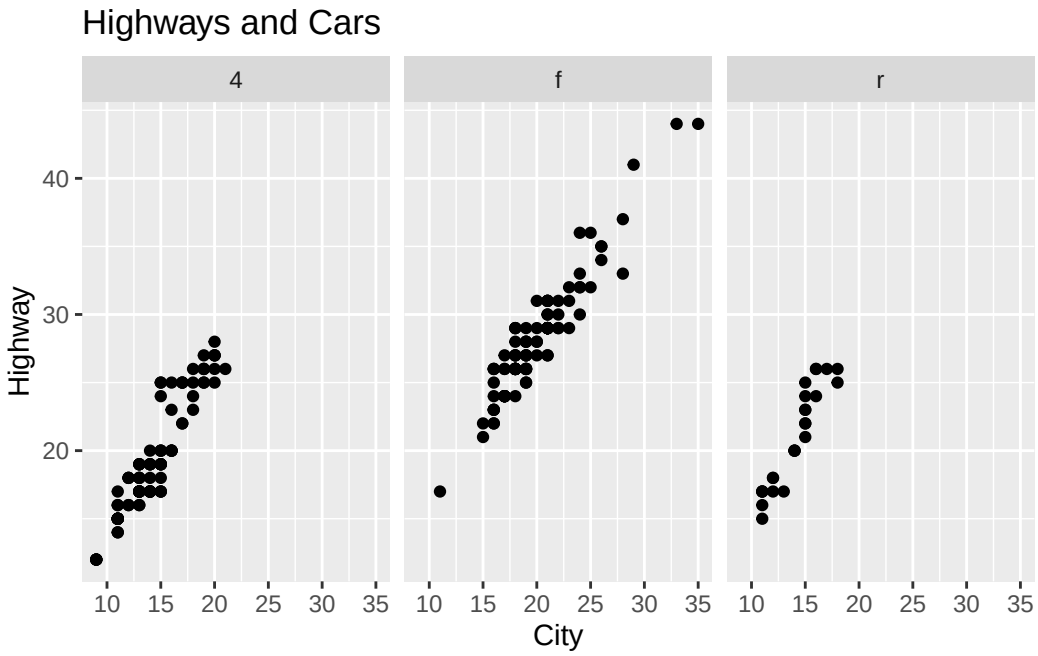
part c - Which graphic do you prefer - the one from part a or part b - for exploring the relationship between these 3 variables? Justify your response.

Solution: I prefer part b because I can clearly distinguish the differences in cty and relate it to the number of highways.

part d - Improve your preferred plot by adding a title and making more appropriate axis labels.

Solution:

```
library(ggplot2)

g <- ggplot(data = mpg, mapping = aes(x = cty, y = hwy)) +
  geom_point() +
  facet_wrap(~ drv) +
  labs(title = "Highways and Cars",
       x = "City",
       y = "Highway")
g
```



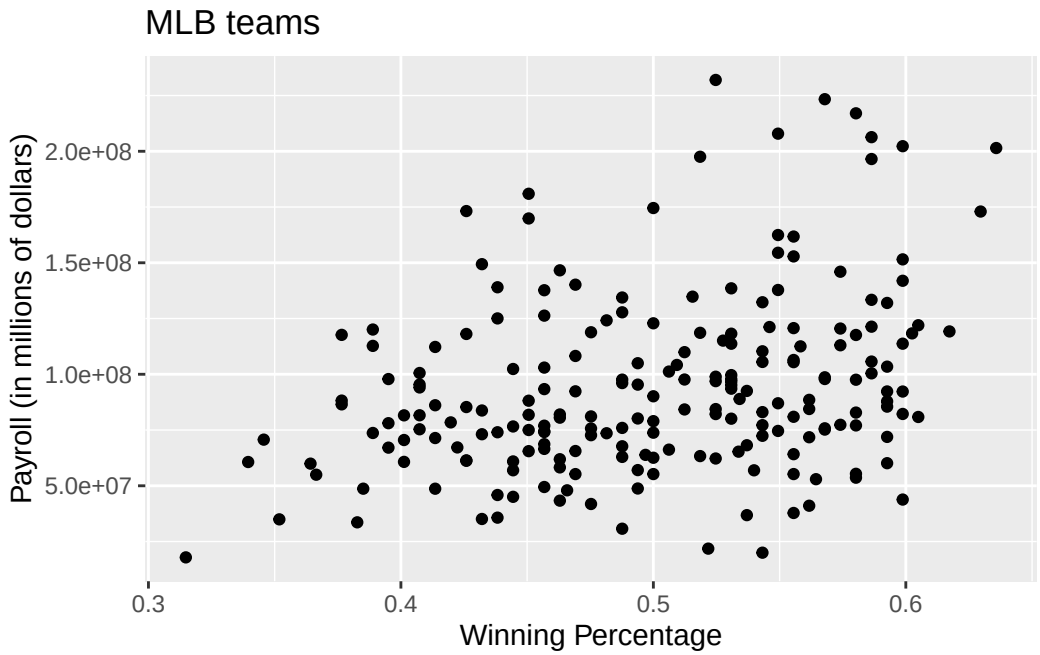
2 - Baseball (Based on MDSR 3.5)

We want to explore the relationship between winning percentage and payroll in context using the *MLB_teams* data in the *mdsr* package.

> part a - Create an informative data graphic that illustrates the relationship between these 2 variables. Be sure your graphic has appropriate labels and a title (i.e. it has context).

Solution:

```
library(ggplot2)

g <- ggplot(data = MLB_teams, mapping = aes(x = WPct, y = payroll)) +
  geom_point() +
  labs(title = "MLB teams",
       x = "Winning Percentage",
       y = "Payroll (in millions of dollars)")

g
```

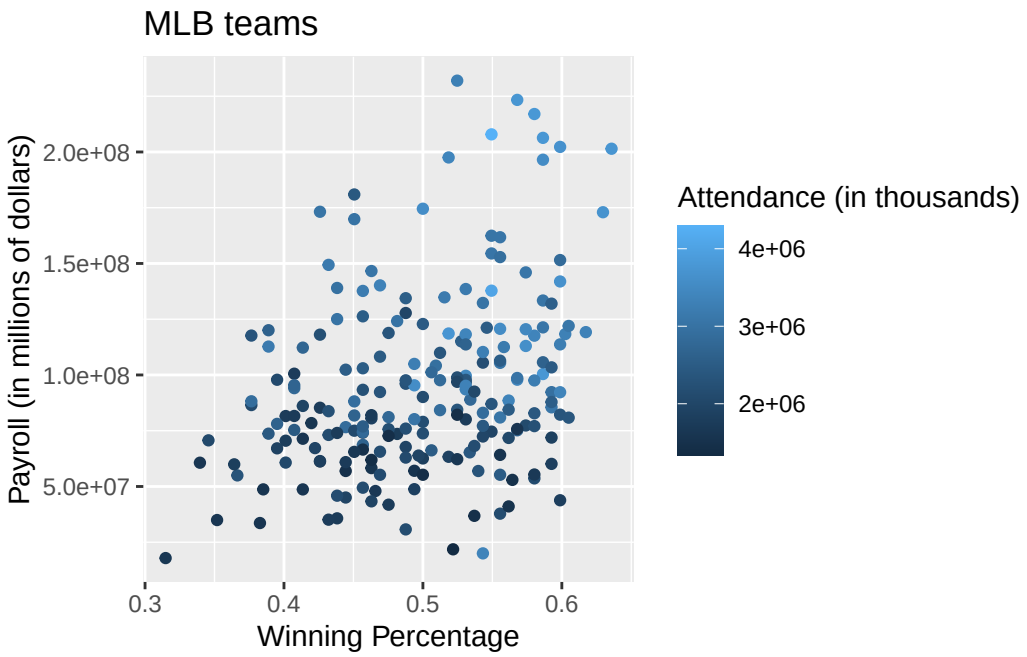


> part b - Now, add a third variable to your plot, making sure to update titles, labels, etc. as needed.

Use the help file to choose a variable that you think will be interesting to add. (You may play around with this, of course.)

Solution:

```
library(ggplot2)

g <- ggplot(data = MLB_teams, mapping = aes(x = WPct, y = payroll, color = attendance)) +
  geom_point() +
  labs(title = "MLB teams",
       x = "Winning Percentage",
       y = "Payroll (in millions of dollars)",
       color = "Attendance (in thousands)")

g
```



part c - What story does your graph from part b tell?

Solution: The graph shows that payroll increases generally when winning percentage is above .5.

3 - Storms (Based on MDSR 3.8)

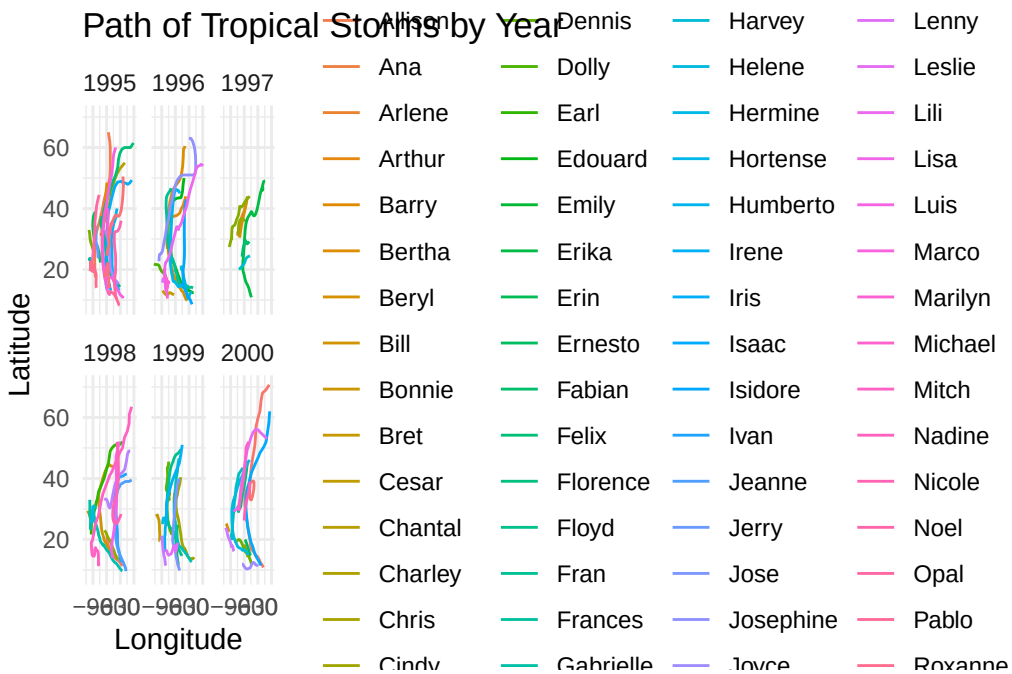
MDSR 3.8 reads: "Using data from the *nasaweather* package, use the *geom_path()* function to plot the path of each tropical storm in the *storms* data table . Use color to distinguish the storms from one another, and use faceting to plot each year in its own panel."

part a - Complete MDSR 3.8

Hint: latitude should be your y-axis and longitude should be your x-axis.

Solution:

```
library(nasaweather)
library(ggplot2)
data("storms")

storms_df <- as.data.frame(storms)
ggplot(storms_df, aes(x = long, y = lat, group = name, color = name)) +
  geom_path() +
  facet_wrap(~year) +
  labs(x = "Longitude", y = "Latitude", title = "Path of Tropical Storms by Year") +
  theme_minimal()
```



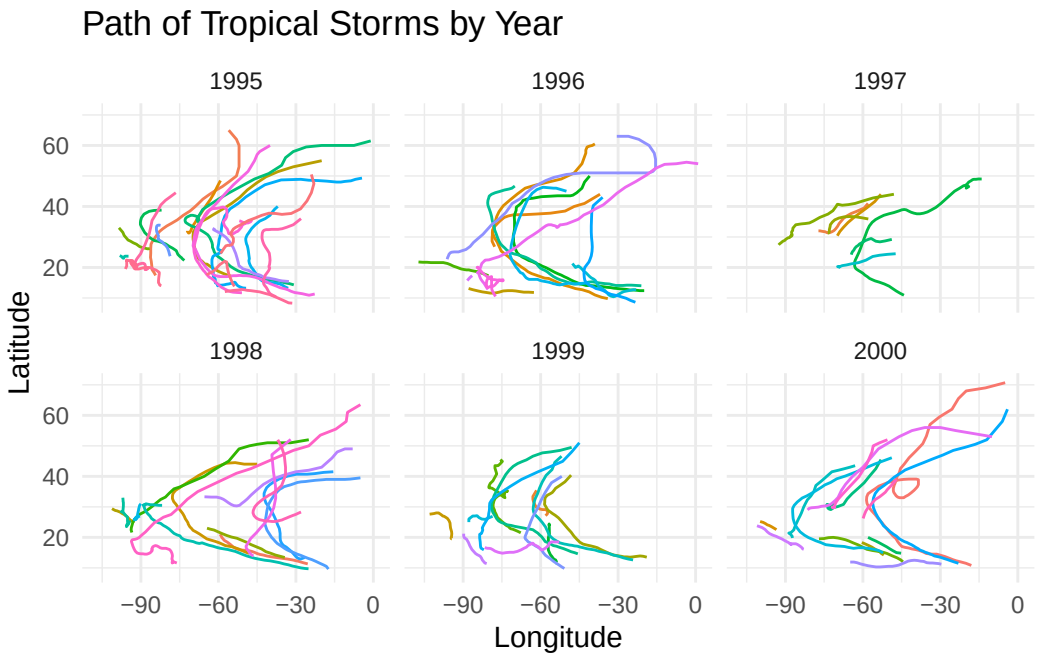
part b - How useful do you find the legend of storm names and colors? If your overall goal was to just look for common paths, would you need the names? Is the plot in part a what you would consider accessible? What issues do you see with the plot? (We will not address all issues.)

Solution:I find the legend and the colors unhelpful. The difference in shades is not distinguishable enough, but the longitude and latitude are distinguishable.

part c - Remove the legend of storm names/colors by adding *scale_color_discrete(guide = "none")*. Be sure your final graph has appropriate labels and a title. Use your plot to discuss what year has the most "variability" in storm paths (in your opinion).

Solution: I think year 2000 has the most variability- especially in latitude.

```
library(nasaweather)
library(ggplot2)

data("storms")

storms_df <- as.data.frame(storms)

ggplot(storms_df, aes(x = long, y = lat, group = name, color = name)) +
  geom_path() +
  facet_wrap(~year) +
  labs(x = "Longitude", y = "Latitude", title = "Path of Tropical Storms by Year") +
  scale_color_discrete(guide = "none") +
  theme_minimal()
```


Path of Tropical Storms by Year
1995
1996
1997
1998
1999
2000
Latitude
Longitude
60
40
20
−90
−60
−30
0

4 - Metabolic Rate

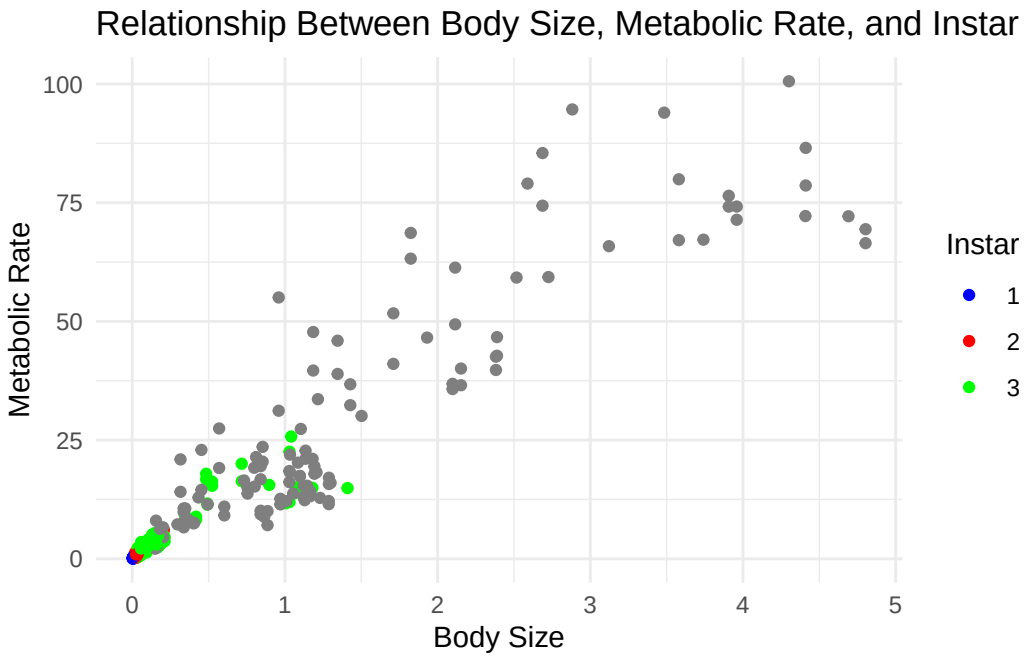
The *MetabolicRate* data set from the *Stat2Data* package contains measurements of body size and metabolic rates for Manduca Sexta caterpillars. We want to model the relationship between these variables using a regression (with body size as the explanatory variable). However, there is some concern that the relationship between the original variables is not linear. (You can investigate this yourself.)

Create an informative data graphic that shows the relationship between these two variables - *BodySize* and *Mrate*. Use appropriate changes to scale to identify a relationship that regression modeling seems appropriate for (you do not need to add the regression line). Then, add the variable *Instar* to the plot in an appropriate way. Be sure your final plot has a title, appropriate labels, and a legend (as appropriate). Finally, describe what your graphic reveals in a few sentences.

Hint: The help menu for the data set will describe what the variables are and may help you identify potential scales to use. However, do not change the variables used in the plot - use the variables above, and make necessary modifications in other ways in the code.

Solution: The graph displays that as Body Size increases, so does Metabolic Rate.

```
library(ggplot2)
ggplot(MetabolicRate, aes(x = BodySize, y = Mrate, color = factor(Instar))) +
  geom_point() +
  labs(title = "Relationship Between Body Size, Metabolic Rate, and Instar",
       x = "Body Size", y = "Metabolic Rate", color = "Instar") +
  scale_color_manual(values = c("1" = "blue", "2" = "red", "3" = "green")) +
  theme_minimal()
```



Going forward, every plot created for our assignments should have clear context in terms of good labels and a title. Be sure to get in the practice of using labs().