**Bio 5202 Problem Set #1**

Due by Friday Mar 31 8:00 pm Central Time, by email to Dr Powers

**Instructions**

Please provide all of your answers within this document directly underneath the question. For questions that ask you to provide figures, you can use insert or copy/paste. This is an "Open Everything" assignment and it is fine to work together, but you must provide your own answers generated on your own machine.

# General comments

# A. Complete code means the following:

-Include loading of packages.

-Including steps to read data if there are any files to read in.

-Include code that assigns name to objects called in the code.

-The code you provide should be functional code. It should work on someone else’s

machine if the machine has the necessary packages installed.

# B. Avoid exporting ggplot objects that have not yet been assigned names. Doing so will export the most recent plot created, which may or may not be the plot you intended to export.

# C. Avoid machine-specific code that loads data from a directory that no one else has on their machine. For example…

scholar <- read\_csv("C:\\Users\\baylorbear\\OneDrive\\Desktop\\scholar.csv")

Instead, in general, keep all necessary files in the same directory and set that as your working directory…

scholar <- read\_csv("scholar.csv")

Also keep in mind that read\_csv() can also read data from subfolders without changing your working directory…

read\_csv("data/scholar.csv")

and it can read data from the web…

scholar <- read\_csv('https://github.com/st3powers/bio5202/raw/main/scholar.csv')

# D. Avoid assigning object names that are already function names. For example…

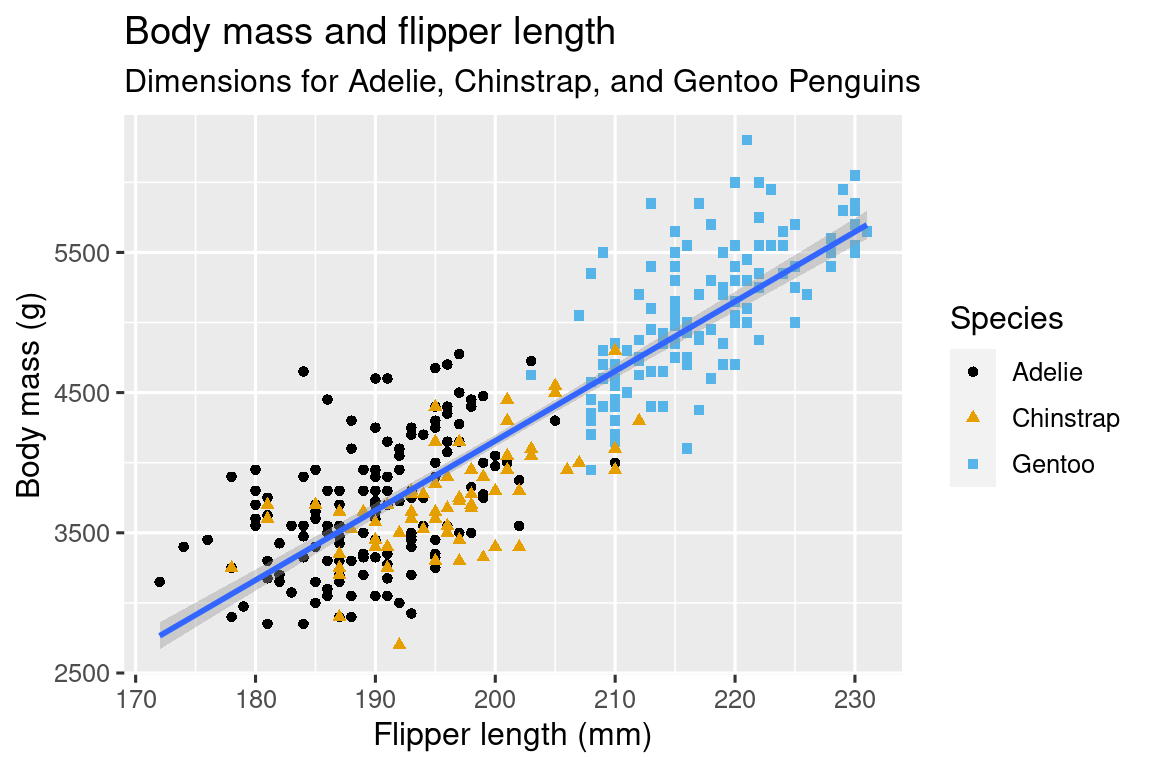
function <- function(ls(“package:dplyr”))

# E. There were some peculiar uses of the na.omit() function.

Instead of this… scholar %>% na.omit(scholar)

Just use this… na.omit(scholar)

**1a.** Provide a complete sequence of R code that successfully generates the below ggplot figure.



library(tidyverse)

library(palmerpenguins)

library(ggthemes)

ggplot(penguins, aes(x=flipper\_length\_mm, y=body\_mass\_g)) +

geom\_point(aes(color=species, shape=species)) +

geom\_smooth(method=lm) +

labs(title="Body mass and flipper length", subtitle = "Dimensions for Adelie, Chinstrap and Gentoo Penguins", x="Flipper length (mm)", y="Body mass (g)",color= "Species", shape="Species") + scale\_color\_colorblind()

**1b.** Revise the figure by removing the gray background, and provide both the revised code and the plot image immediately below.

library(tidyverse)

library(palmerpenguins)

library(ggthemes)

ggplot(penguins, aes(x=flipper\_length\_mm, y=body\_mass\_g)) +

geom\_point(aes(color=species, shape=species)) +

geom\_smooth(method=lm) +

labs(title="Body mass and flipper length", subtitle = "Dimensions for Adelie, Chinstrap and Gentoo Penguins", x="Flipper length (mm)", y="Body mass (g)",color= "Species", shape="Species") +

scale\_color\_colorblind() + theme\_bw()

**1c.** Revise the figure title (Body mass and flipper length) so that your name is included in the title, and provide both the revised code and the plot image immediately below.

library(tidyverse)

library(palmerpenguins)

library(ggthemes)

ggplot(penguins, aes(x=flipper\_length\_mm, y=body\_mass\_g)) +

geom\_point(aes(color=species, shape=species)) +

geom\_smooth(method=lm) +

labs(title="Body mass and flipper length- R Guru ", subtitle = "Dimensions for Adelie, Chinstrap and Gentoo Penguins", x="Flipper length (mm)", y="Body mass (g)",color= "Species", shape="Species") +

scale\_color\_colorblind()+theme\_bw()

# Comments on #1

# The below method is not ideal, because it completely removes the axis lines.

ggplot(penguins, aes(x=flipper\_length\_mm, y=body\_mass\_g)) +

geom\_point(aes(color=species, shape=species)) +

geom\_smooth(method=lm) +

labs(title="Body mass and flipper length- R Guru ", subtitle = "Dimensions for Adelie, Chinstrap and Gentoo Penguins", x="Flipper length (mm)", y="Body mass (g)",color= "Species", shape="Species") +

scale\_color\_colorblind()+

theme(panel.background = element\_blank())

**2a.** Use R code to generate an alphabetized list of function names for every function contained in the dplyr package. Then use R to restrict the list of function names to only first *n* elements, where *n* is the number of characters in your last name. For example, if your last name has 7 characters, restrict the function list to only the first 7 elements. Provide the complete code sequence used to generate this restricted list, as well as the actual names of those functions.

library(dplyr)

funs\_dplyr <- ls("package:dplyr")

head(funs\_dplyr,6)

#[1] "%>%" "across"

#[3] "add\_count" "add\_count\_"

#[5] "add\_row" "add\_rownames"

**2b.** Technically speaking, the R object you created in 2a may or may not be a "list" object by R standards. What kind of object is it? Provide the answer along with code that generates the answer.

library(dplyr)

funs\_dplyr <- ls("package:dplyr")

head(funs\_dplyr,6) %>% class()

# [1] "character"

# or, same thing without the pipe

class(head(funs\_dplyr,6))

**3a.** Provide a complete sequence of R code that successfully generates the below ggplot figure. Don't worry about considerations of the width or height of the figure yet.

Chart, scatter chart, bubble chart

Description automatically generated

library(tidyverse)

ggplot(mpg, aes(x = displ, y = hwy)) +

  geom\_point(aes(color = drv)) +

  geom\_smooth(se = FALSE, aes(linetype = drv))

Chart, scatter chart, bubble chart

Description automatically generated

library(tidyverse)

ggplot(mpg, aes(x=displ, y=hwy)) +

  geom\_point(aes(color=drv)) +

  geom\_smooth(se = FALSE)

Chart, scatter chart

Description automatically generated

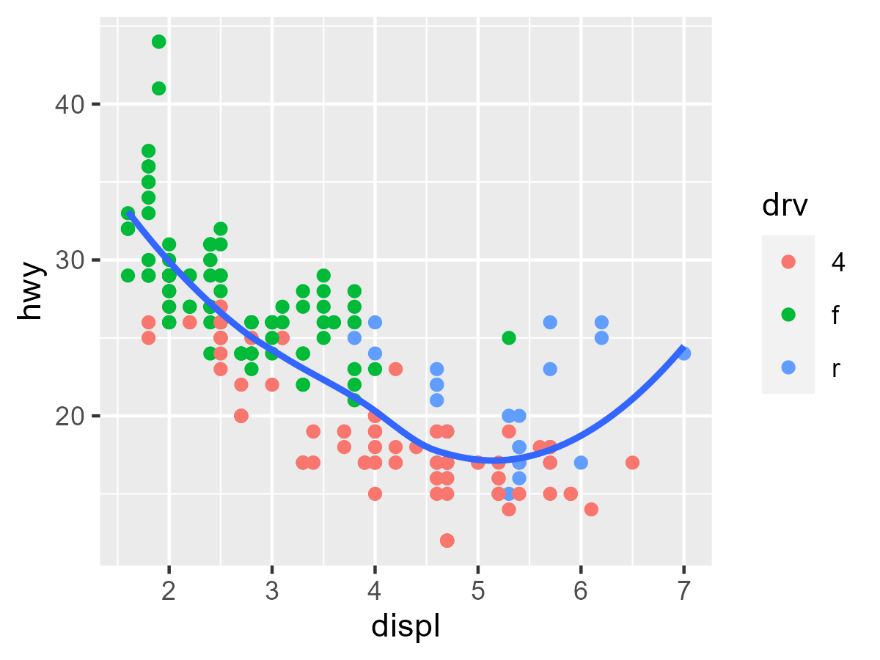
library(tidyverse)

ggplot(mpg, aes(x=displ, y=hwy)) +

  geom\_point(color = "black") +

  geom\_smooth(aes(group = drv), se = FALSE)

**3b.** Export the ggplot object as a png file with width of 4 inches and height of 3 inches. Provide the plot image immediately below.



library(tidyverse)

theplot <- ggplot(mpg, aes(x=displ, y=hwy)) +

  geom\_point(aes(color=drv)) +

  geom\_smooth(se=FALSE)

ggsave("theplot.png",theplot,width=4,height=3,units="in")

# or more specifically...

ggsave(filename="theplot.png",plot=theplot,width=4,height=3,units="in")

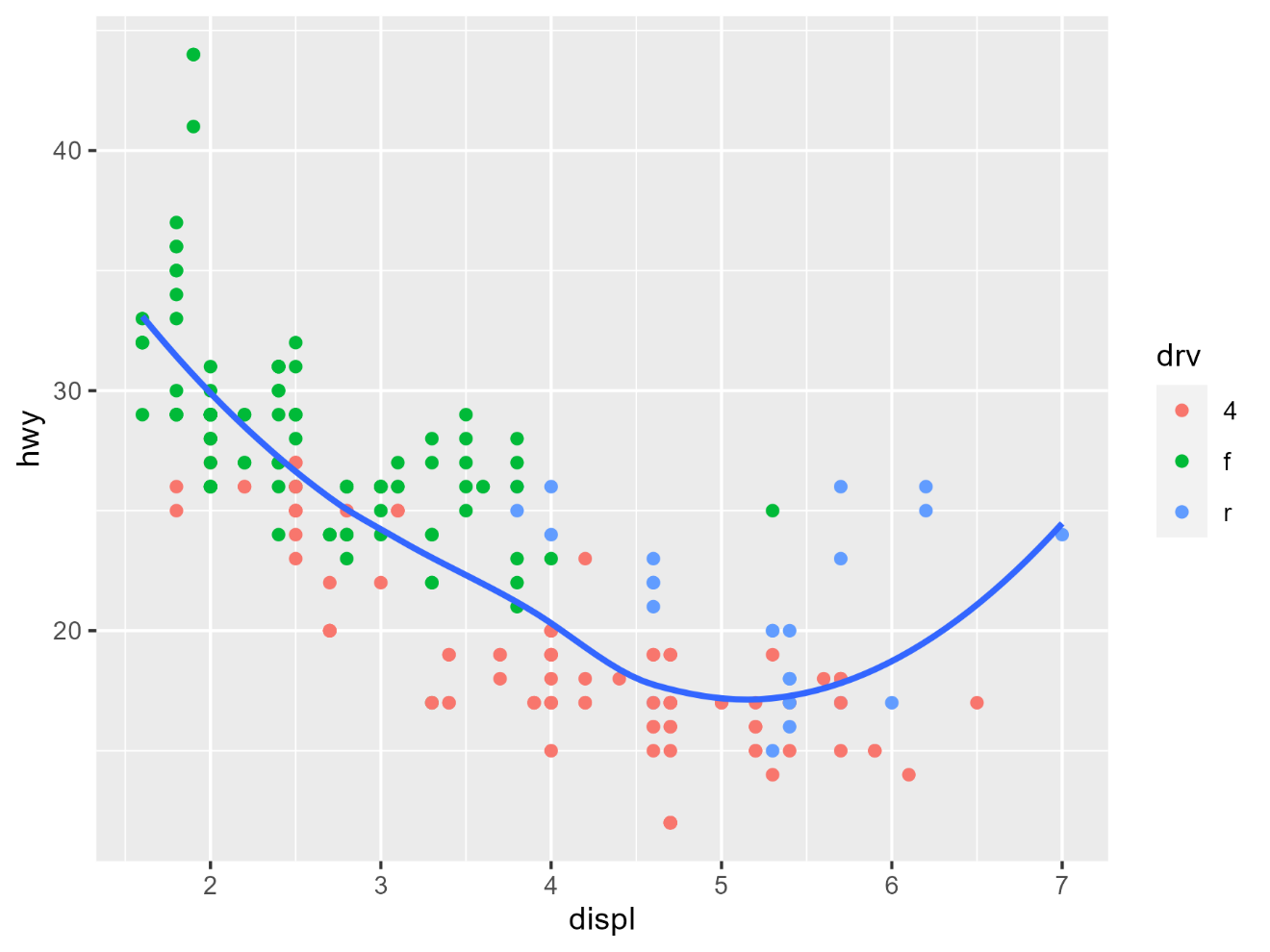
# or alternatively…

png(filename="theplot.png", width=4,height=3,units="in",res=300)

theplot

dev.off()

**3c.** Export the ggplot object as a png file with width of 6 inches and height of 4.5 inches. Provide the plot image immediately below.



library(tidyverse)

theplot <- ggplot(mpg, aes(x=displ, y=hwy)) +

  geom\_point(aes(color=drv)) +

  geom\_smooth(se=FALSE)

ggsave("theplot.png",theplot,width=6,height=4.5,units="in")

**4a.** Find the "scholar.csv" file in the class github repository and load the file into R. This dataset contains four productivity indicators compiled from Google Scholar for 18 Baylor Biology primary research faculty (anonymized using the "person\_number" column). These indicators are the h-index, and the number of times their works were cited in each of the last three years (2020, 2021, and 2022). Ignoring NA values (not all individuals had a Google Scholar profile), calculate the total number of citations earned by these faculty in 2021, in other words the sum across all individuals in 2021, providing the complete code sequence used to generate your answer.

\*Note that the instructor's answer will double count cited articles coauthored by two more of these individuals, but for this question you can ignore that issue.

library(tidyverse)

scholar <- read\_csv('https://github.com/st3powers/bio5202/raw/main/scholar.csv')

scholar |> filter(indicator=="2021") |> summarize(sum(value,na.rm=TRUE))

These faculty were cited 4422 times in 2021 according to google scholar.

# A tibble: 1 × 1

  `sum(value, na.rm = TRUE)`

                       <dbl>

1                       4422

**4b.** Consider the sum of citations across years 2020, 2021, and 2022 for each individual.

Create a 2 column x 4 row table that shows which person\_numbers had the top four (highest) citation rates over this three year period, and the number of citations for each individual over this three year period. Provide the i) complete code sequence used to generate the table, and ii) the table itself (a screenshot of the table viewed in Rstudio will suffice, as will an inserted or copy/pasted table that was exported from Rstudio).

library(tidyverse)

scholar |> filter(indicator %in% c("2020","2021", "2022")) |>

  group\_by(person\_number) |>

  summarize(sums=sum(value,na.rm=TRUE)) |>

  arrange(desc(sums)) |>

  head(4)

Text

Description automatically generated with medium confidence

**5a.** Find the "gamerstate.csv" file in the class github repository and load the file into R. This dataset contains survey results from recreational video game players from different states of the USA. It reports, for each state, the average hours per week that video games were played by survey respondents. According to the survey, in which three states did gamers spend the most hours per week playing video games? Provide the complete code sequence used to generate a table that contains the state names, and the number of hours per week played, for these top three states.

library(tidyverse)

gamerstate <- read\_csv('https://github.com/st3powers/bio5202/raw/main/gamerstate.csv')

gamerstate |> arrange(desc(hours\_per\_week)) |> head(3)

# A tibble: 3 × 2

state hours\_per\_week

<chr> <dbl>

1 MI 17

2 OH 17

3 TX 17

**5b.** Create two of the following three visualizations for these gaming survey data: i) histogram,

ii) horizontal barplot with state abbreviations on the y axis and hours per week on the x axis, iii) scatter plot with state abbreviations on the y axis and hours per week on the x axis. Provide the complete sequence of R code for generating these figures, as well as the figure itself.

library(tidyverse)

gamerstate <- read\_csv('https://github.com/st3powers/bio5202/raw/main/gamerstate.csv')

# histogram

gamerstate %>% ggplot(aes(x=hours\_per\_week)) +

  geom\_histogram()+

  theme\_bw()

# From Wickham and Grolemund Chapter 10- Layers

# "Consider a basic bar chart, drawn with geom\_bar() or geom\_col()."

# barplot

gamerstate %>% ggplot(aes(x=state,y=hours\_per\_week)) +

  geom\_col()+

  coord\_flip()+

  theme\_bw()

# or...

gamerstate %>% ggplot(aes(y=state,x=hours\_per\_week)) +

  geom\_col()+

  theme\_bw()

# scatterplot

gamerstate %>% ggplot(aes(x=state,y=hours\_per\_week)) +

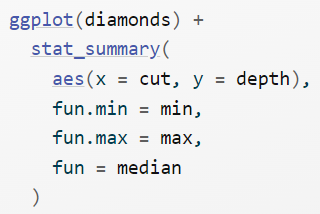
  geom\_point()+

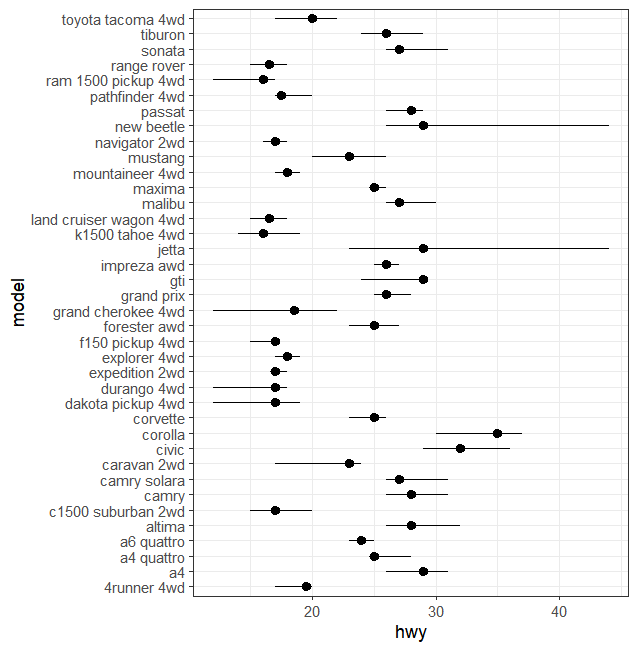
  theme\_bw()

**5c.** Consider the number of hours per week that the Texas gamer respondents played video games (17 hours per week). What percent of the current Baylor undergraduate population would you predict typically plays video games for 17 hours per week or higher, in an average week? What percent of Baylor Biology graduate students?

This question was intended to give freebie points. There are no incorrect answers for this question.

**6a.** Consider the 'diamonds' dataset contained in the tidyverse package, and the below code. Then use the 'mpg' dataset, also available in the tidyverse package, to generate the plot underneath. In your response please provide the code that successfully generates the plot.





library(tidyverse)

ggplot(mpg) +

  stat\_summary(aes(x=hwy, y=model),

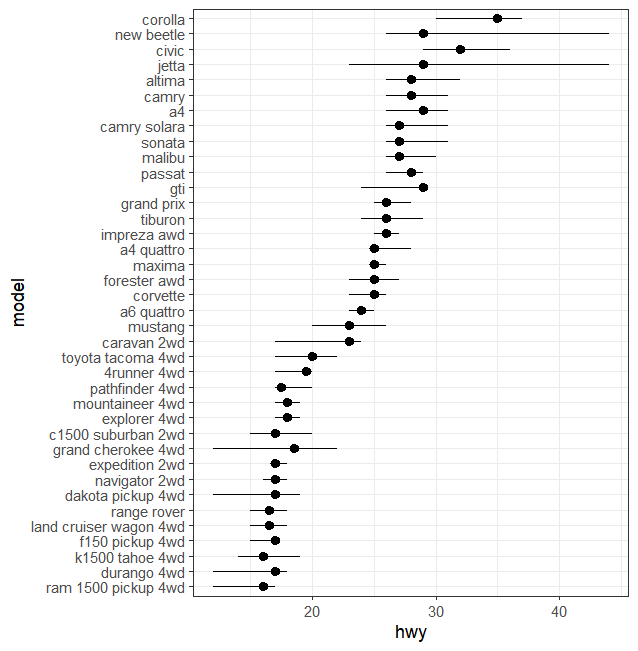
               fun.min= min,

               fun.max=max,

               fun =median) +

  theme\_bw()

**6b.** Revise the above plot so it is sorted by the hwy values, as shown below. Provide the code that successfully generated the plot.



library(tidyverse)

ggplot(mpg) +

  stat\_summary(aes(x=hwy, y=reorder(model,hwy)),

               fun.min= min,

               fun.max=max,

               fun =median) +

  theme\_bw()