Lab 4 - Class Data Visualization

Name: Yiran Hu, netid: yiranhu3

Assignment Overview

- We'll be exploring our class survey data that we cleaned in Lab 1. This time, we'll focus on visualizations!
- Note that each row represents one student in our class, and each column is a variable/question from the survey.
- Don't use your own Lab 1 file for this assignment—use the cleaned data provided in the Canvas instructions.
- When finished, upload a pdf of your report to Gradescope.

Step 0

Pre-lab work: Complete the pre-lab tutorials for Lab 4 first: https://stat212-learnr.stat.illinois.edu/ (https://stat212-learnr.stat.illinois.edu/)

Load `tidyverse` package.

```
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────────────── tidy
verse 2.0.0 ──
## ✓ dplyr     1.1.2     ✓ readr     2.1.4
## ✓ forcats   1.0.0     ✓ stringr   1.5.0
## ✓ ggplot2   3.4.3     ✓ tibble    3.2.1
## ✓ lubridate 1.9.2     ✓ tidyr     1.3.0
## ✓ purrr     1.0.2
## ── Conflicts ────────────────────────────────────────────────────────────
────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to beco
me errors
```

Load the data

- Download the Class_F23.xlsx file to your computer. Save it to the same folder as this RMarkdown file.
- If you haven't already, you might need to install `readxl`. Remove the `#` symbol below to run it. Then be sure to remove that line once installed.
- Make sure the name of the file *matches* what you input inside `read_excel`. If it's `Class_F23_1` for example, be sure to adjust that!

```
#install.packages("readxl)
library(readxl)
Class_F23 = read_excel("Class_F23.xlsx")
```

View data. Run once below, but delete before knitting your markdown!

```
#View(Class_F23)
```

- Coding Tip: Remember that R is CaSe AnD sYmBoL_SeNsItIvE. As you code, type in your variable names exactly as they appear in the data frame. sleep /= Sleep. Grad Plans /= Grad_Plans

Question 1 (5pts)

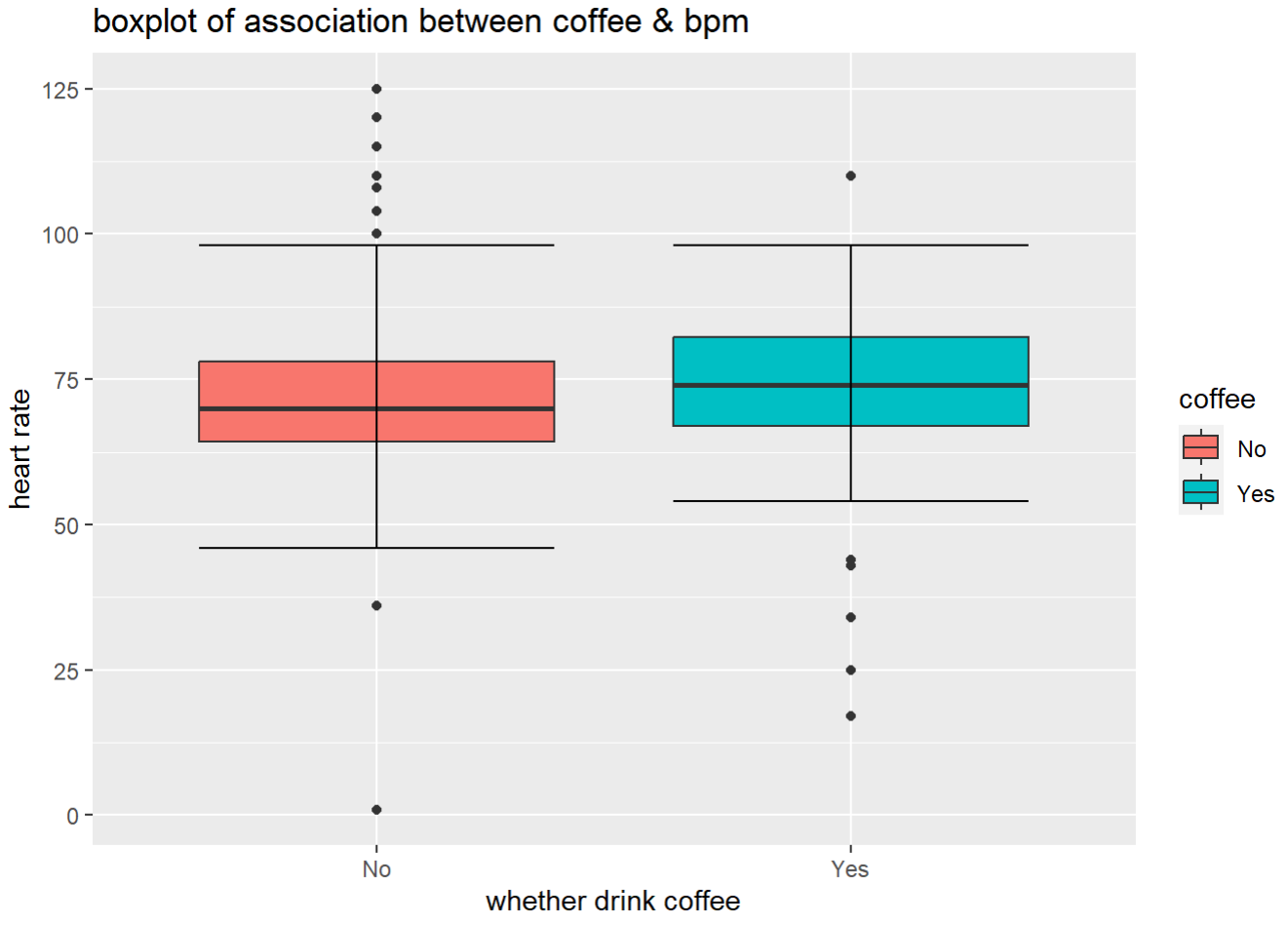
Is there any association between students' heart rates (measured in beats per minute) and whether they drank coffee in the last 24 hours? Create side by side boxplots to make the comparison

Include the image of the boxplots here.

- Add an appropriate title and appropriate axes labels
- Each box should be a different fill color
- Add whiskers (errorbars) to your boxplots

```
ggplot(data = Class_F23, aes(x = coffee, y = bpm, fill = coffee)) +
  geom_boxplot() +
  stat_boxplot(geom = "errorbar") +
  labs(x = "whether drink coffee", y = "heart rate", title = "boxplot of association between co
ffee & bpm")
```

```
## Warning: Removed 2 rows containing non-finite values (`stat_boxplot()`).
## Removed 2 rows containing non-finite values (`stat_boxplot()`).
```



Do you see any difference in BPM between these two groups of students? If so, which group seems to have higher BPM values and by how much?

There are not big differences between the two groups.

Is this the result you expected?

No, I thought people who drank coffee would have higher heart rate.

Question 2 (5pts)

Next, let's look at the values student reported as their expected salary in 20 years.

Report the numeric summary (min, Q1, Q2, mean, Q3, max) in salary expectation for the class.

```
summary(Class_F23$salary)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.      NA's
## 1.000e+00 9.775e+04 1.300e+05 2.750e+08 2.000e+05 1.000e+11         8
```

min: 1.000e+00, Q1: 9.775e+04, Q2: 1.300e+05, mean: 2.750e+08, Q3: 2.000e+05, max: 1.000e+11

Include an image of a density curve for this variable here

- Add an appropriate title
- Add a fill color (change the fill color from the default "white" option it currently has)
- **OPTIONAL:** If you're curious how to turn o ff scientific notation and report values in normal notation, you can run library(scales) and add the following line to your ggplot code: scale_x_continuous(labels = comma)

```
library(scales)
```

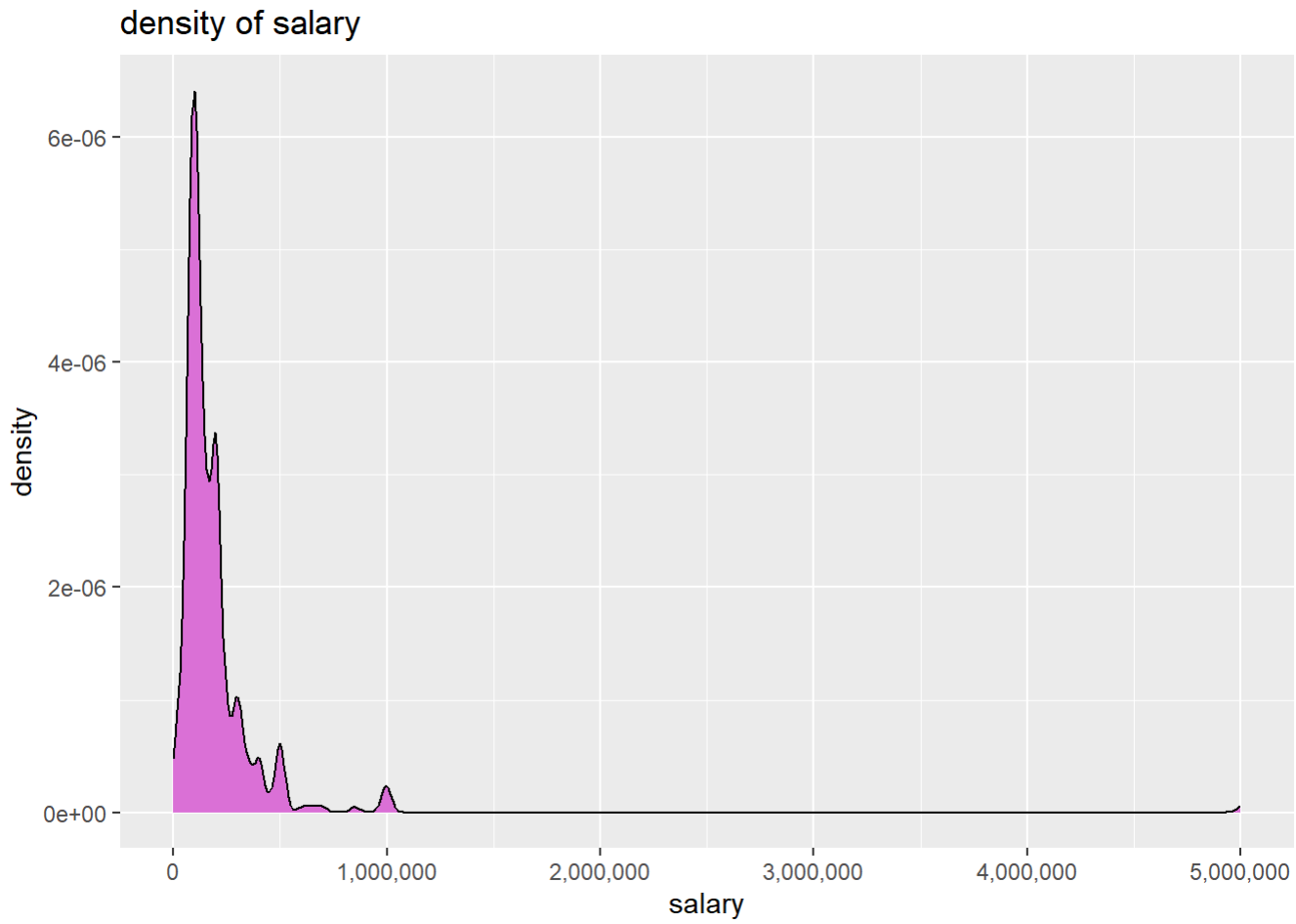
```
##
## 载入程辑包：'scales'
```

```
## The following object is masked from 'package:purrr':
##
##     discard
```

```
## The following object is masked from 'package:readr':
##
##     col_factor
```

```
ggplot(data = Class_F23, aes(x = salary)) +
  geom_density(fill="orchid")+
  scale_x_continuous(labels = comma,limits = c(0, 5000000))+
  labs(title="density of salary")
```

```
## Warning: Removed 11 rows containing non-finite values (`stat_density()`).
```



The middle 50% of students reported expected salary levels between what two values?

Between Q1 & Q3: from 9.775e+04 to 2.000e+05.

Why does the scale of this plot stretch so high? Are class responses scattered pretty evenly across this range, or more concentrated in one numeric range of this plot? Hint: sort the salary variable and scroll to the bottom

Because some students answered very large numbers like 1e+11. But the class responses are concentrated around 10^4 to 10^5.

Question 3 (5pts)

Are students' salary expectations associated with their plans after graduation?

To investigate this, we will make a summary table using a pipe that reports the mean, median, and standard deviation in projected salary based on students' graduation plans.

Hint: some people have no entry in the salary column (which creates a default response of "NA"). You'll need to program in a response to remove the NAs when telling R to calculate the statistics.

Include an image of your summary table

```
Class_F23 |>
  group_by(grad_plans) |>
  summarise(Mean = mean(salary,na.rm=TRUE),
            Median = median(salary,na.rm=TRUE),
            St_Dev = sd(salary,na.rm=TRUE))
```

grad_plans <chr>	Mean <dbl>	Median <dbl>	St_Dev <dbl>
A Job	1785853917.4	100000	1.336304e+10
Dental School	195882.4	200000	8.970847e+04
Graduate School	247174.0	100000	1.050271e+06
Medical School	314987.2	200000	8.583563e+05
PA School	124956.5	120000	4.763925e+04
Something else	132000.0	101000	8.504544e+04

6 rows

Based on the summary stats, does there seem to be any association between these two variables? How might you explain this result in context?

I think there is some relationships between the two variables. Because on the summary stats, students who plan “A job” or “Medical School” expects higher salaries. Salaries are related to the job, becoming a doctor or dentist may have more salary on average. But some jobs like PA may have low salary.

Which grad plan group has the highest standard deviation, and what do you think is contributing to that? Hint: sort the salary column and scroll to the bottom!

“A job” group. Because one student choose 1e+11 as expected salary, but other students’ salaries are about 1e+5 or 1e+4. That makes the standard deviation large.

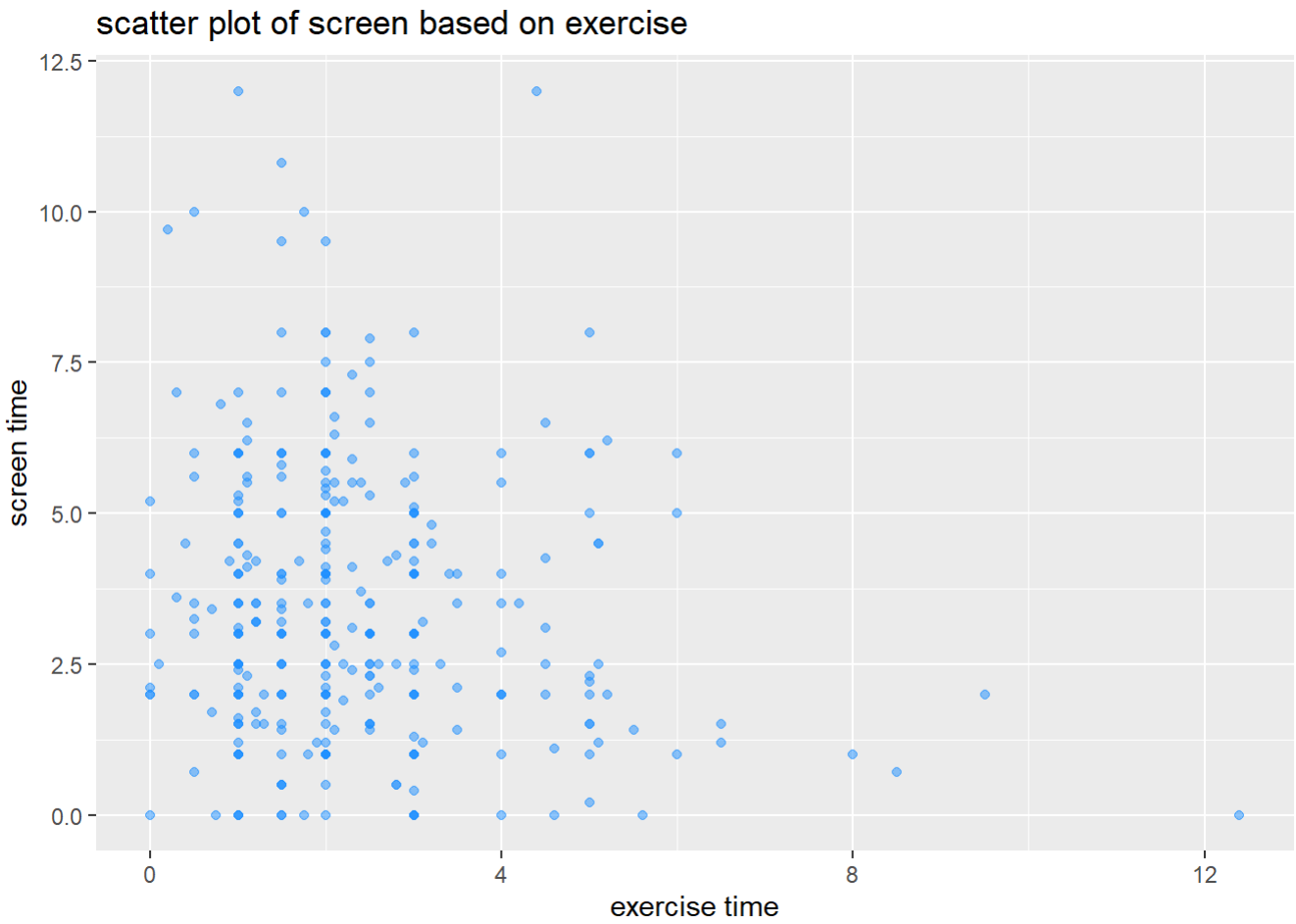
Question 4 (5pts)

Is there any association between how much time students reported using a screen other than their phone in the last 24 hours and how much time they spent exercising or actively moving?

Include an image of a scatterplot for these variables here, with exercise placed on the x-axis.

- Build your plot inside a pipe that only includes students whose daily activities (phone time, other screen time, exercise time, and sleep time) are less than or equal to 24 hours
- Add an appropriate title and axes labels

```
Class_F23 |>
  filter(phone+screen+exercise+sleep_total<=24) |>
    ggplot(aes(x = exercise, y = screen)) +
    geom_point(alpha = 0.5, color = "dodgerblue") +
    labs(x = "exercise time", y = "screen time", title="scatter plot of screen based on exercis
e")
```



Do you see any association between these variables? How might you explain this relationship?

I think the more time students spend on exercise, the less time they will spend on screen. Because students who like using screens may spend a lot of time on that, so they have less time to do exercise.

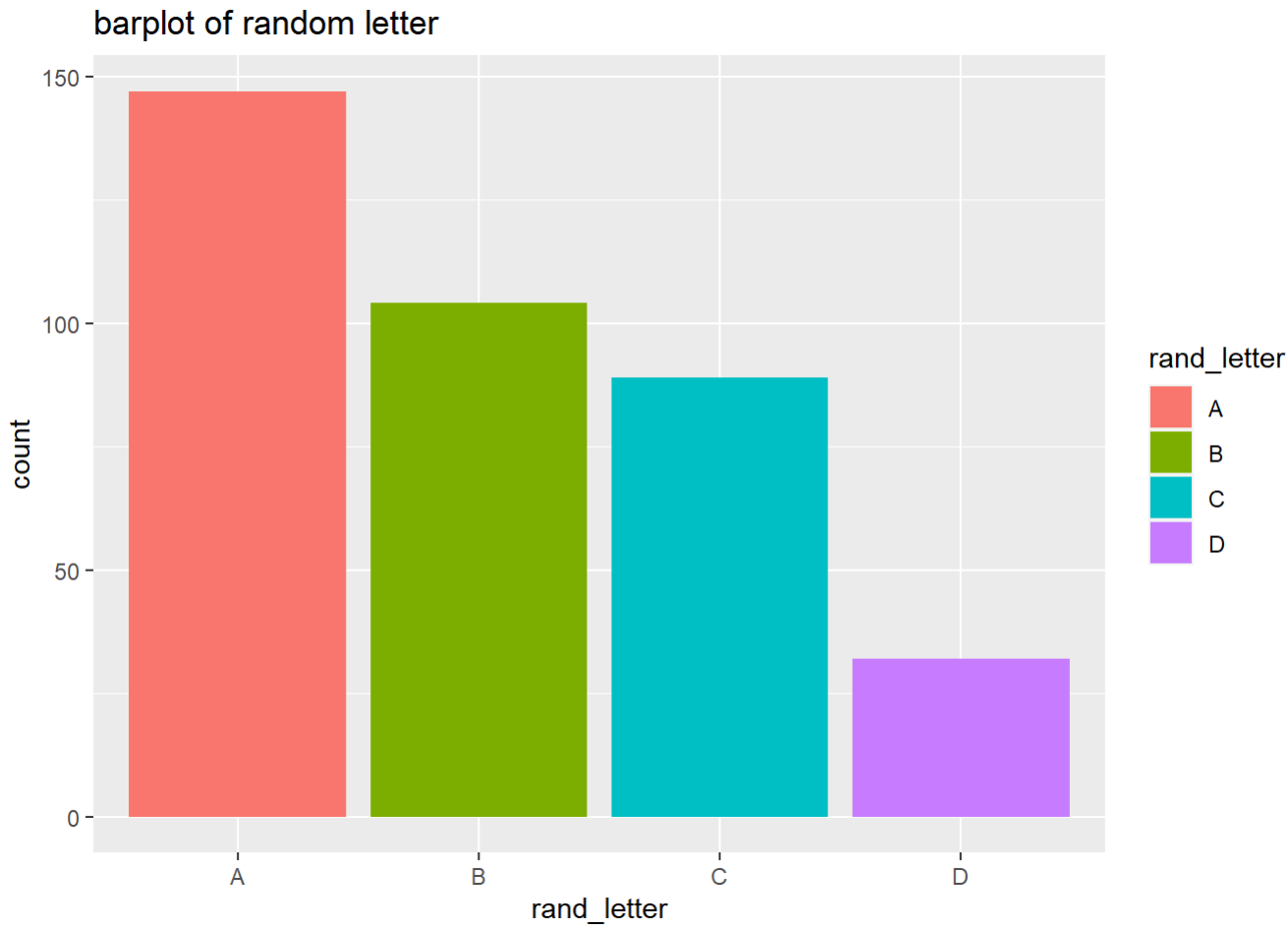
Question 5 (5pts)

When asked to choose a letter or number at random, how did the class do?

Create a univariate barplot showcases the results of the random letter question.

- Fill each bar with a different color
- Add an appropriate title

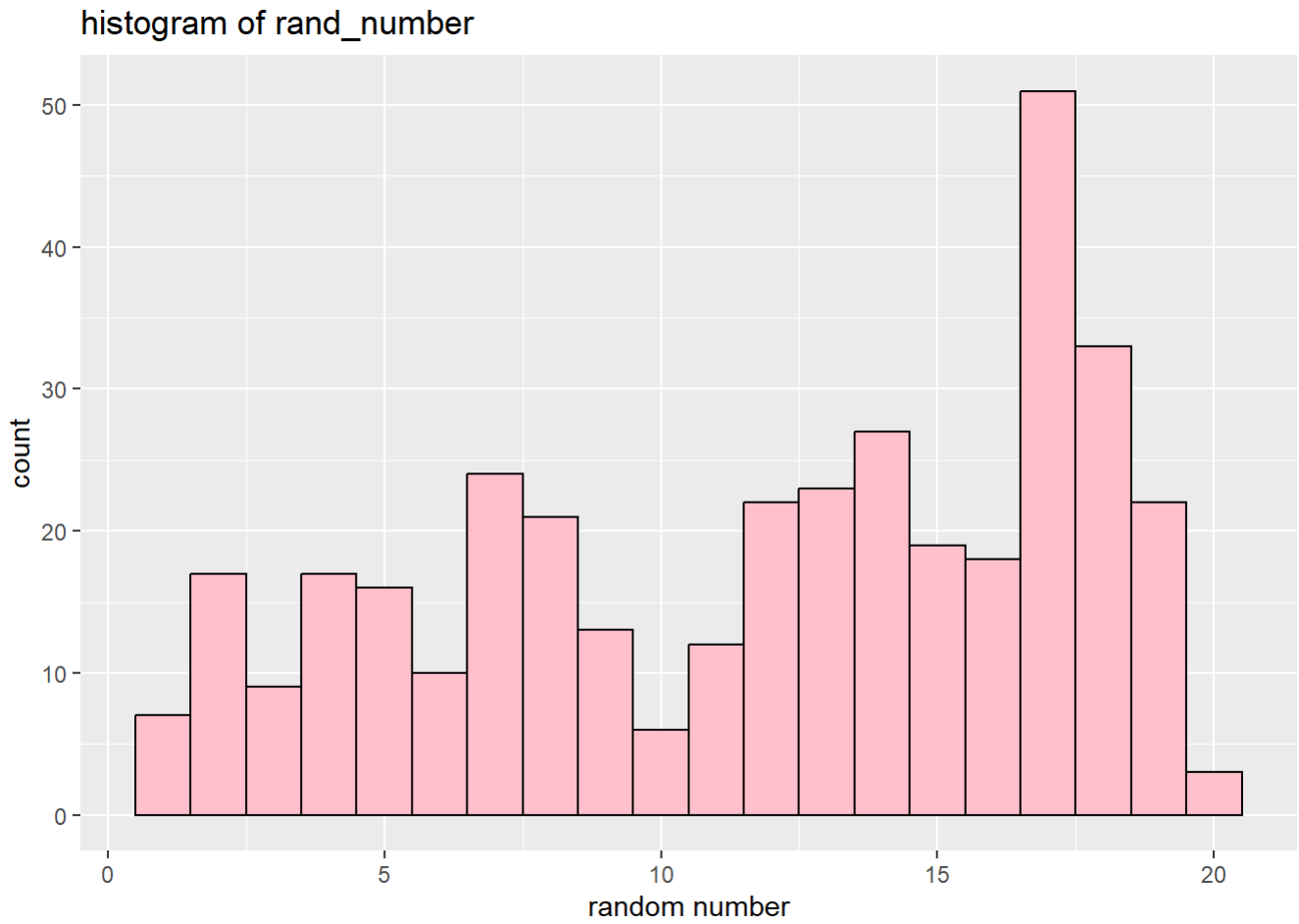
```
ggplot(data=Class_F23,aes(x=rand_letter,fill=rand_letter))+
  geom_bar()+
  labs(title="barplot of random letter")
```



Create a histogram that showcases the results of the random number question

- Filter out any numbers outside the range from 1 to 20
- Set your histogram to have 20 bins
- Choose distinct fill colors and border color for your histogram bins
- Add an appropriate title and x axis label

```
Class_F23 |>
  filter(rand_number>=1 & rand_number<=20) |>
  ggplot(aes(x=rand_number))+
  geom_histogram(bins=20,color="black",fill="pink") +
  labs(title="histogram of rand_number", x="random number")
```



Based on the results, how well do you think the class did at choosing at random?

When choosing a number or letter, they seem to choose specific letters or numbers more often, like letter A and number 17. It looks not very random.

Question 6 (5pts)

Let's explore the relationship of two categorical variables: academic level and whether or not a student owns a car. *Consider whether these are categorical or numeric variables and choose an appropriate visualization to represent them!*

Intermediate step: *Before* creating the graph, note that the academic level variable will list the categories alphabetically, rather than in order of seniority. Use the following template to complete a custom re-ordering of the levels. Identify your data frame name and variable name correctly and plug that into each slot. Then run this code to restructure the variable. Nothing will output—but you'll see in your graph that the order is correct! If you make a mistake and accidentally messed up something with the data, try re-importing the data again. *This part is only worth 1 point, so even if you fail to re-order the categories, just move on to the graph!*

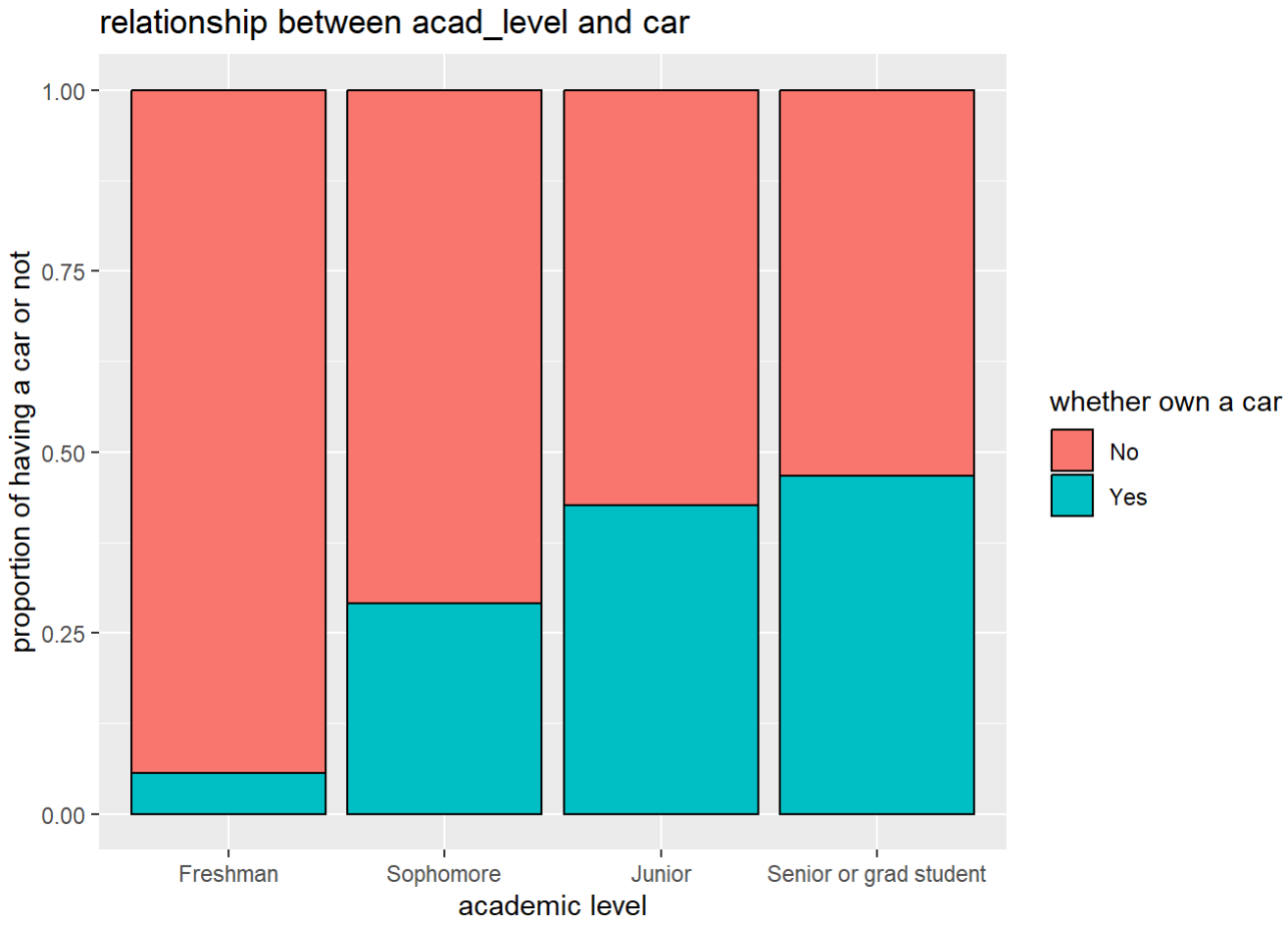
Remove the `#` sign at the beginning as well.

```
Class_F23$acad_level = factor(Class_F23$acad_level, levels = c("Freshman", "Sophomore", "Junior", "Senior or grad student"))
```

Include an image of your plot

- Add an appropriate title and an appropriate axis label for any axis a variable is assigned to
- You are welcome to add or adjust any other features if appropriate

```
ggplot(data = Class_F23, aes(x = acad_level, fill = car)) +
  geom_bar(color = "black", position = "fill") +
  labs(x = "academic level",
       y = "proportion of having a car or not",
       fill="whether own a car",
       title = "relationship between acad_level and car")
```



Does there appear to be any association between students' academic level and car ownership status? Briefly explain what you notice in your graph to make this conclusion.

Yes. As academic level increases, a larger proportion of students have a car.

Question 7 (5pts)

What's a multivariate question that you have about the class data?

Pose a question involving two variables in our class dataset.

Are there any relations between sleep_total and academic level?

Create an appropriate visualization and/or summary table that helps you address this question.

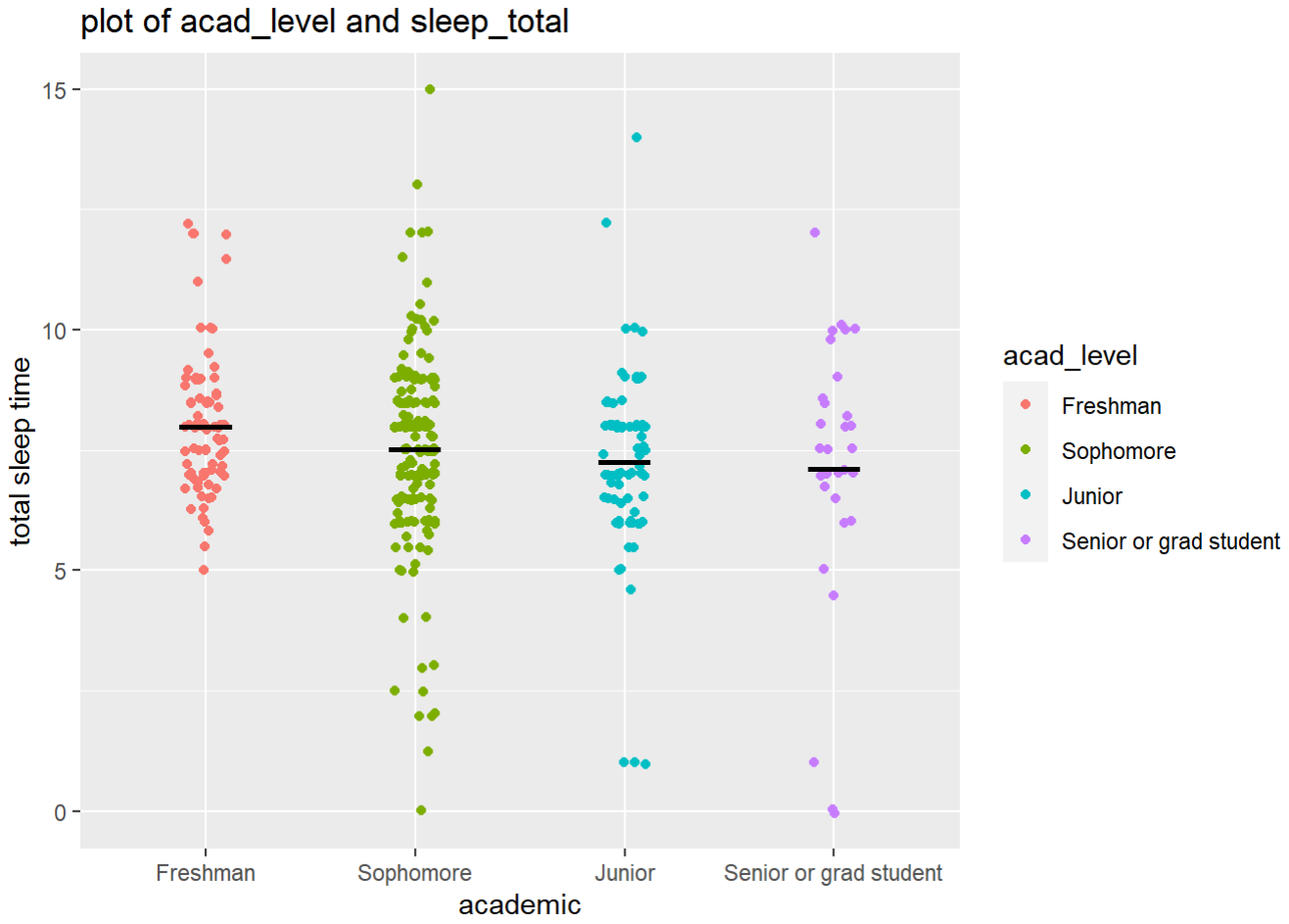
- Be proactive to filter out any outliers as needed
- Please format any visualizations (titles, axes labels, color as appropriate)
- If making a summary table, please add appropriate column headers

```
ggplot(data = Class_F23, aes(x = acad_level, y = sleep_total, color = acad_level)) +
  geom_jitter(width=0.1) +
  labs(x = "academic", y = "total sleep time", title = "plot of acad_level and sleep_total")+
  stat_summary(fun = mean, fun.min = mean, fun.max = mean, geom = "errorbar", color = "black",
width = 0.25, size = 1.0)
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## ℹ Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```
## Warning: Removed 2 rows containing non-finite values (`stat_summary()`).
```

```
## Warning: Removed 2 rows containing missing values (`geom_point()`).
```



Briefly answer your question based on your findings in the data

The distribution of sleep in 4 acad_level looks very similar. The mean of sleep_total slightly decrease as the grade increases.