Lab 5 - Class Data Visualization

Step 0

- 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

Load `tidyverse` package.

```
#library(rmarkdown)

cat("\014")
```

```
shell("cls")

#remove.packages("rlang")
#install.packages("devtools")
#install.packages("rlang")
#install.packages("tidyverse", dependencies = TRUE)
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse
2.0.0 ──
## ✔ dplyr     1.1.0     ✔ readr     2.1.4
## ✔ forcats   1.0.0     ✔ stringr   1.5.0
## ✔ ggplot2   3.4.1     ✔ tibble    3.1.8
## ✔ lubridate 1.9.2     ✔ tidyr     1.3.0
## ✔ purrr     1.0.1
## ── Conflicts ──────────────────────────────────────────
tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the ]8;;http://conflicted.r-lib.org/conflicted package]8;; to force
all conflicts to become errors
```

```
library(dplyr)
library(ggplot2)
#tinytex::install_tinytex()
```

Load the data

- Download the Class_S23.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_S23_1` for example, be sure to adjust that!

```
#install.packages("readxl)
#library(readxl)
#Class_S23 = read_excel("Class_S23.xlsx")
```

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

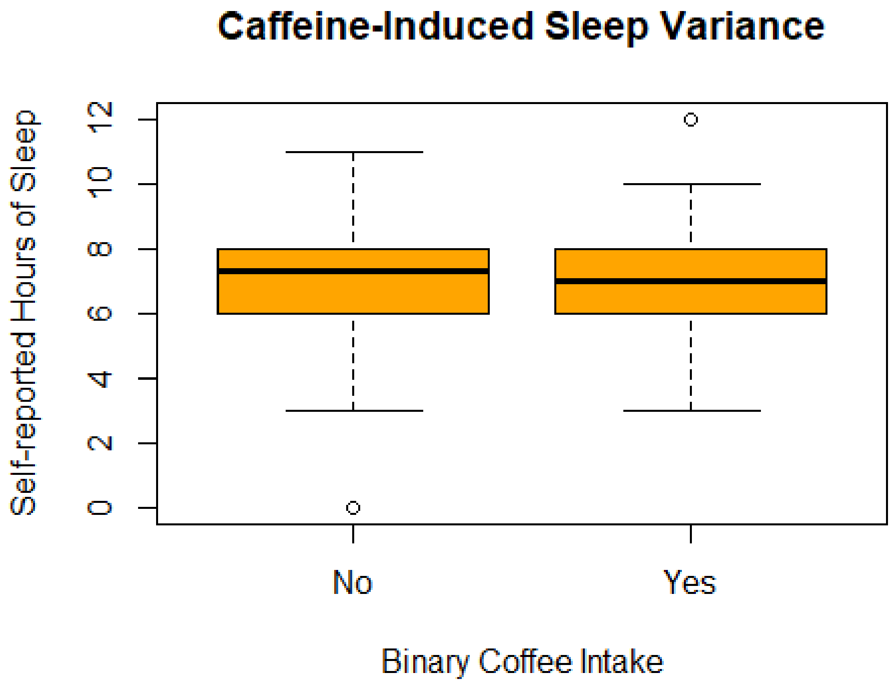
```
library(readxl)
Class_S23 <-
read_excel("C:/Users/varen/Desktop/Illini+Uni/SP23/STAT_212/STAT_212-
Lab_05/Class_S23.xlsx")
#View(Class_S23)
```

Question 1

Are students who reported having coffee in the last 24 hours reporting different amounts of sleep on average than the non-coffee drinkers? Create side by side boxplots to compare these two groups.

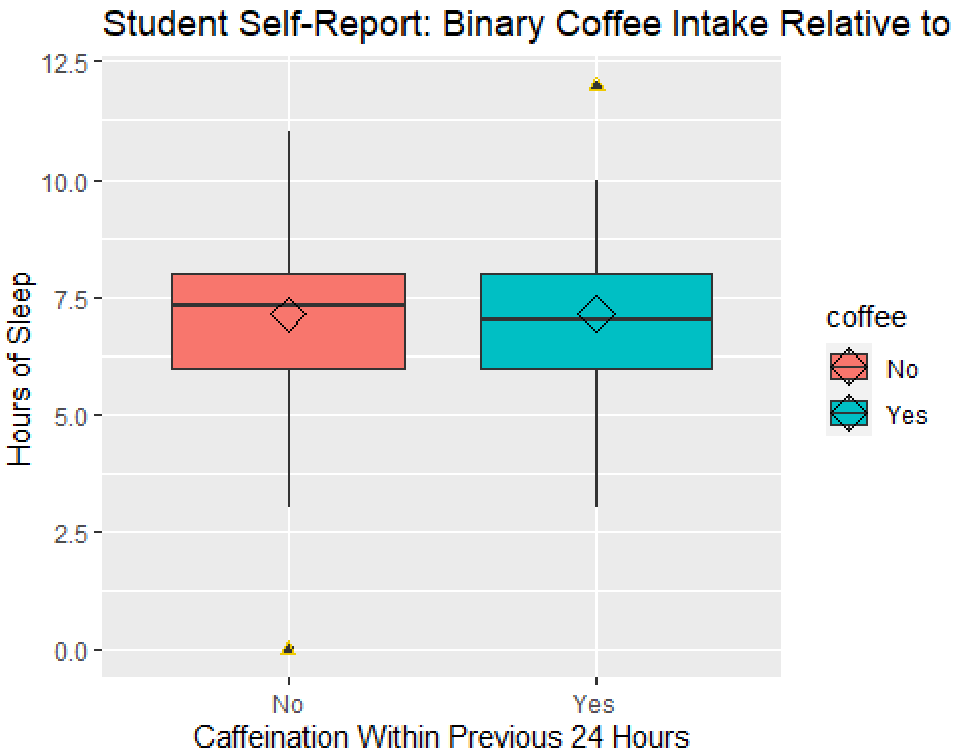
```
cat("\014")
```

```
shell("cls")
class_Data <- Class_S23

#Method 1: boxplot function
  boxplot(Class_S23$sleep ~ Class_S23$coffee, col="orange", main="Caffeine-
Induced Sleep Variance", ylab="Self-reported Hours of Sleep", xlab="Binary
Coffee Intake", horizontal=FALSE)
```



```
#Method 2: ggplot function
  class_Data$coffee <-factor(class_Data$coffee) # converts coffee into a
categorical variable
  coffee_sleep.bp <<-ggplot(data=class_Data, aes(y=sleep, x=coffee,
fill=coffee)) #creates boxplot
  coffee_sleep.bp <- coffee_sleep.bp + geom_boxplot() #Adds color
  coffee_sleep.bp <- coffee_sleep.bp + ggtitle("Student Self-Report: Binary
Coffee Intake Relative to Overnight Rest") # Adds a title
  coffee_sleep.bp <- coffee_sleep.bp + ylab("Hours of Sleep") +
xlab("Caffeination Within Previous 24 Hours") # Adds Labels
  coffee_sleep.bp <- coffee_sleep.bp + geom_boxplot(varwidth = TRUE) #boxes
are drawn with widths proportional to the square-roots of the number of
observations in the groups
  coffee_sleep.bp <- coffee_sleep.bp + geom_boxplot(notch = TRUE) # make a
notched box plot
  coffee_sleep.bp <- coffee_sleep.bp + geom_boxplot(outlier.colour = "gold2",
outlier.shape = 2)
  #coffee_sleep.bp <- coffee_sleep.bp + geom_jitter(width = 0.2) #adds a
small amount of random variation to the location of each point, and is a
```

```
useful way of handling overplotting caused by discreteness in smaller
datasets.
  #coffee_sleep.bp <- coffee_sleep.bp +guides(fill=FALSE) # removes legend
  coffee_sleep.bp <- coffee_sleep.bp + stat_summary(fun.y=mean, geom="point",
shape=5, size=4) # add a diamond at the mean

## Warning: The `fun.y` argument of `stat_summary()` is deprecated as of
ggplot2 3.3.0.
## ℹ Please use the `fun` argument instead.

  coffee_sleep.bp # Displays the Boxplot
```



```
  #coffee_sleep.bp <- coffee_sleep.bp + coord_flip() # rotates the boxplot

#Method 3:
```

Include the image of the boxplots here. (sharing your code is optional)

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

Briefly address these questions (suggested: 30-50 words):

```
print("The students who reported having coffee in the last 24 hours are
reporting slightly different amounts of sleep on average than the non-coffee
drinkers. They have a higher average and median amount of sleep, but nothing
outstanding.")
```

```
## [1] "The students who reported having coffee in the last 24 hours are
reporting slightly different amounts of sleep on average than the non-coffee
drinkers. They have a higher average and median amount of sleep, but nothing
outstanding."
```

- Do you think coffee drinking explains any variability in students' reported sleep?

```
print("Since there is no overwhelming difference between the two boxplots, I
do not think that drinking coffee explains any variability in students'
reported sleep.")
```

```
## [1] "Since there is no overwhelming difference between the two boxplots, I
do not think that drinking coffee explains any variability in students'
reported sleep."
```

- Is this the result you expected?

```
print("This is the result I expected since people usually drink coffee in the
morning, not at night. Therefore most of the obvious effects of coffee wear
off in a couple of hours in the early morning.")
```

```
## [1] "This is the result I expected since people usually drink coffee in
the morning, not at night. Therefore most of the obvious effects of coffee
wear off in a couple of hours in the early morning."
```

Question 2

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

Report the numeric summary in salary expectation for the class.

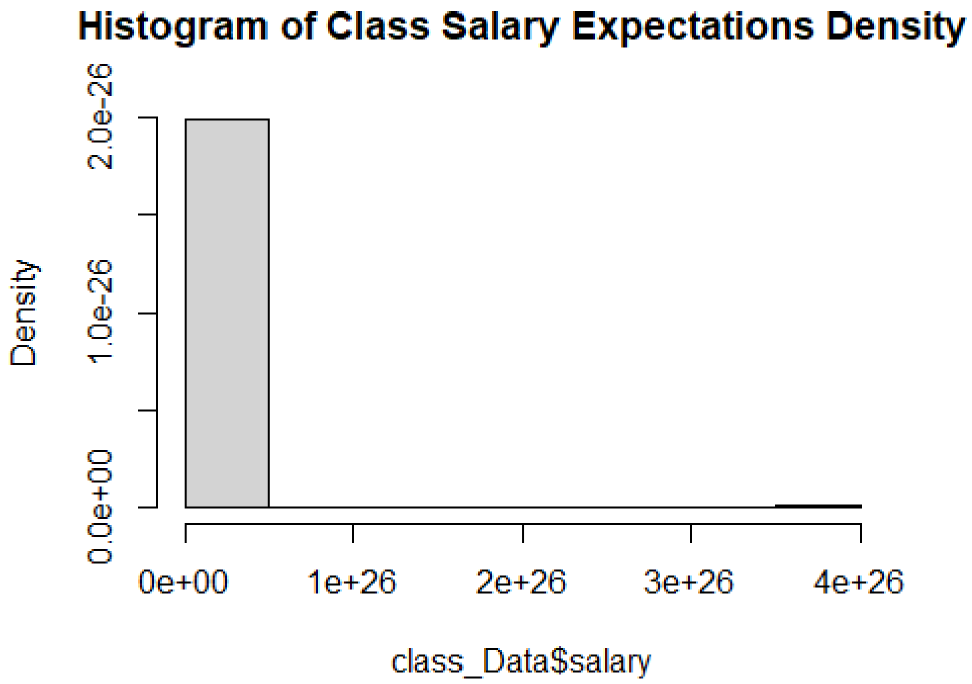
```
cat("\014")
```

```
shell("cls")

print("The numeric summary of class salary expectations is as follows:")

## [1] "The numeric summary of class salary expectations is as follows:"

summary(class_Data$salary)

##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max. 
## 2.000e+02 9.000e+04 1.325e+05 1.227e+24 2.500e+05 4.000e+26
```

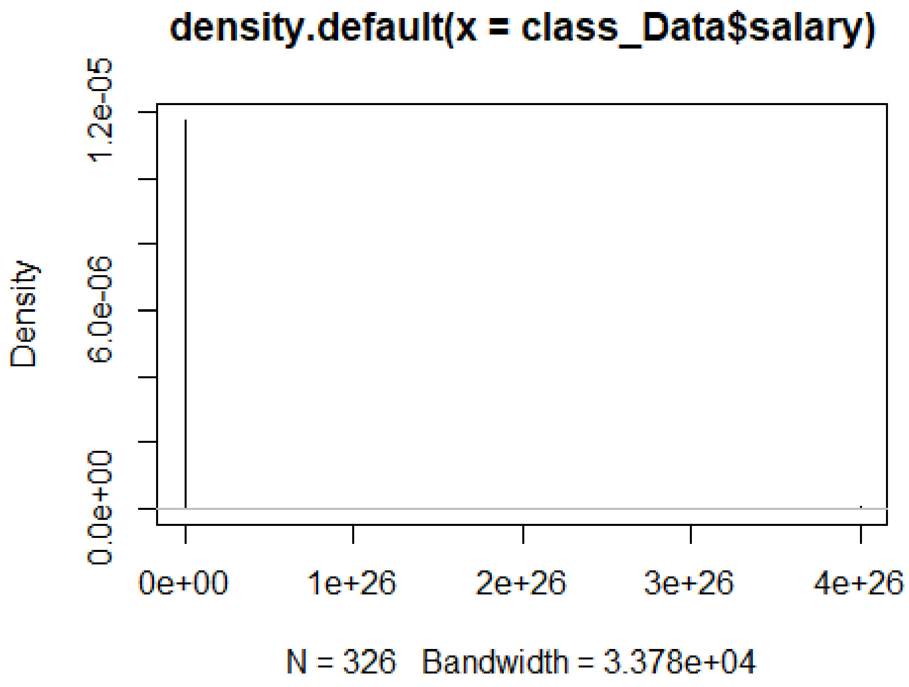
Include an image of a histogram for this variable here (sharing your code is optional)

```
cat("\014")
```

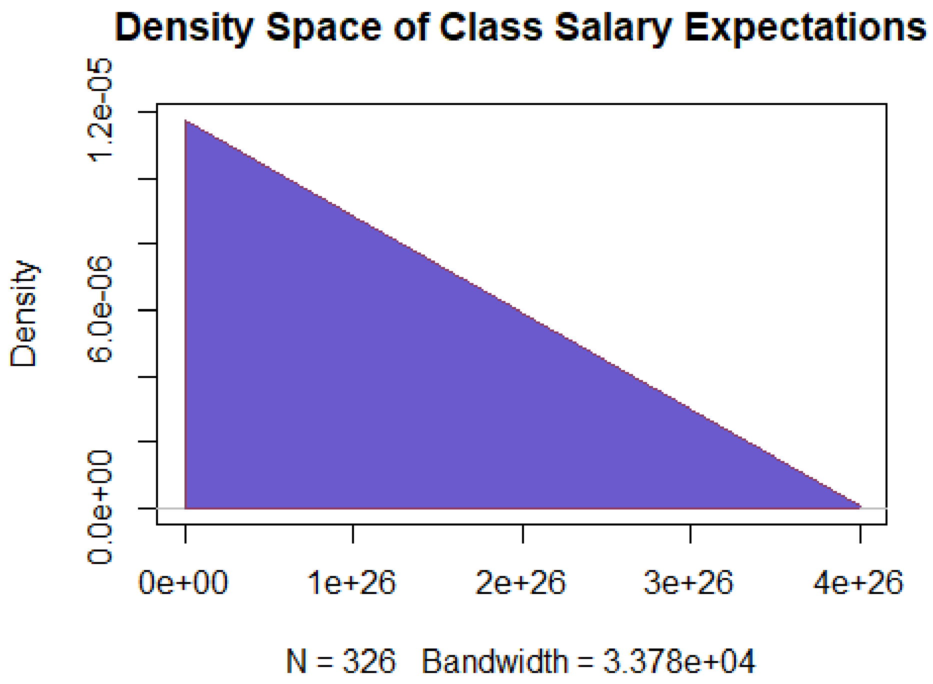
```r
shell("cls")

#hist(class_Data$salary)

#ggplot(data = class_Data, aes(x = salary)) + geom_histogram() +
scale_color_brewer(palette = "Set2") + theme_classic() + labs(title = "Class
Salary Expectations", x = "Expected Salary", y = "Frequency") +
theme(plot.title = element_text(size = 14, hjust = 0.5, face = "bold"),
axis.title = element_text(size = 12, face = "bold"), axis.text =
element_text(size = 11, color = "hotpink4"))

hist(class_Data$salary, freq = FALSE, main = "Histogram of Class Salary
Expectations Density")
```



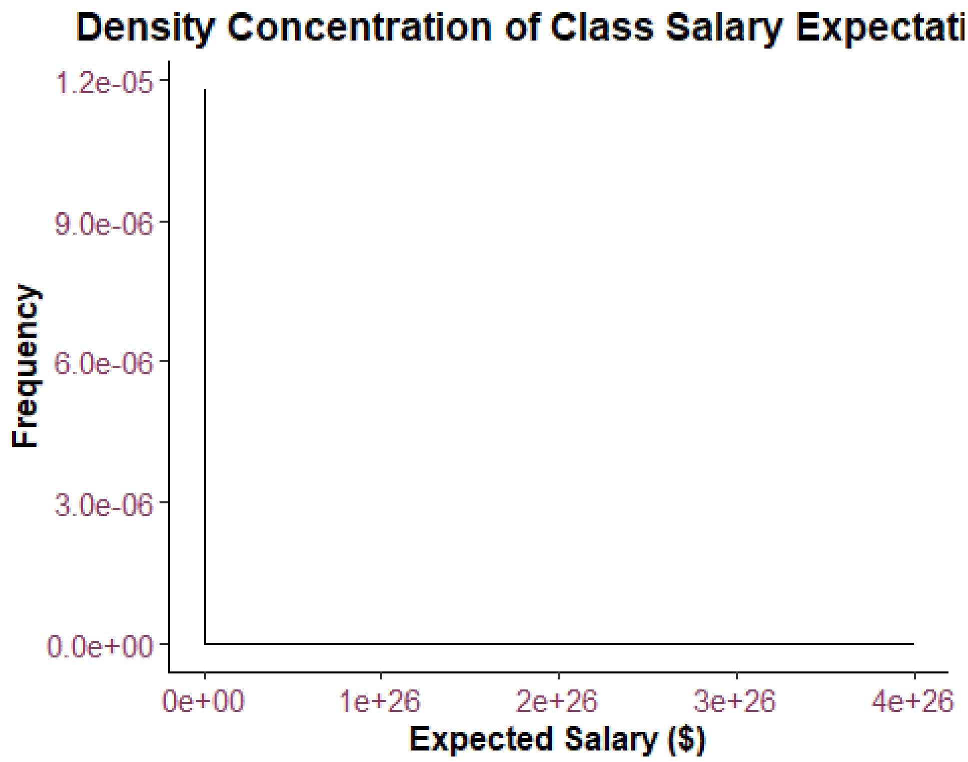
```r
# Kernel Density Plot
plot(density(class_Data$salary))
```



```
# Filled Density Plot
plot(density(class_Data$salary), main="Density Space of Class Salary
Expectations")
polygon(density(class_Data$salary), col="slateblue", border="hotpink4")
```



```
#ggplot(data = class_Data, aes(x = salary)) + geom_density() +
geom_vline(aes(xintercept=mean(salary)), color="slateblue",
linetype="dashed", size=1)

ggplot(data = class_Data, aes(x = salary)) + geom_density() +
scale_color_brewer(palette = "Set2") + theme_classic() + labs(title =
"Density Concentration of Class Salary Expectations", x = "Expected Salary
($)", y = "Frequency") + theme(plot.title = element_text(size = 14, hjust =
0.5, face = "bold"), axis.title = element_text(size = 12, face = "bold"),
axis.text = element_text(size = 11, color = "hotpink4"))
```



- Add an appropriate title
- Add a fill color (change the fill color from the default “white” option it currently has)
- Use a plot theme
- All other features optional!

Briefly address these questions

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

```
print("The middle 50% of students reported expected salary levels between
9.000e+04 and 2.500e+05 according to the numeric summary")

## [1] "The middle 50% of students reported expected salary levels between
9.000e+04 and 2.500e+05 according to the numeric summary"
```

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

```
print("The scale of the plot stretches so high because most of the responses
are concenctrated around 'low' salary values in comparison to the maximum.
They are scattered across different ranges, but they appear to be linearly
group due to the function of the density plot showing a continous outline of
the salary range; there is one ridiculously large salary value - more than
100 trillion times larger than a billion (already an extremely high number) -
which alters the shape of the graph dramatically")
```

```
## [1] "The scale of the plot stretches so high because most of the responses are concenctrated around 'low' salary values in comparison to the maximum. They are scattered across different ranges, but they appear to be linearly group due to the function of the density plot showing a continous outline of the salary range; there is one ridiculously large salary value - more than 100 trillion times larger than a billion (already an extremely high number) - which alters the shape of the graph dramatically"
```

Question 3

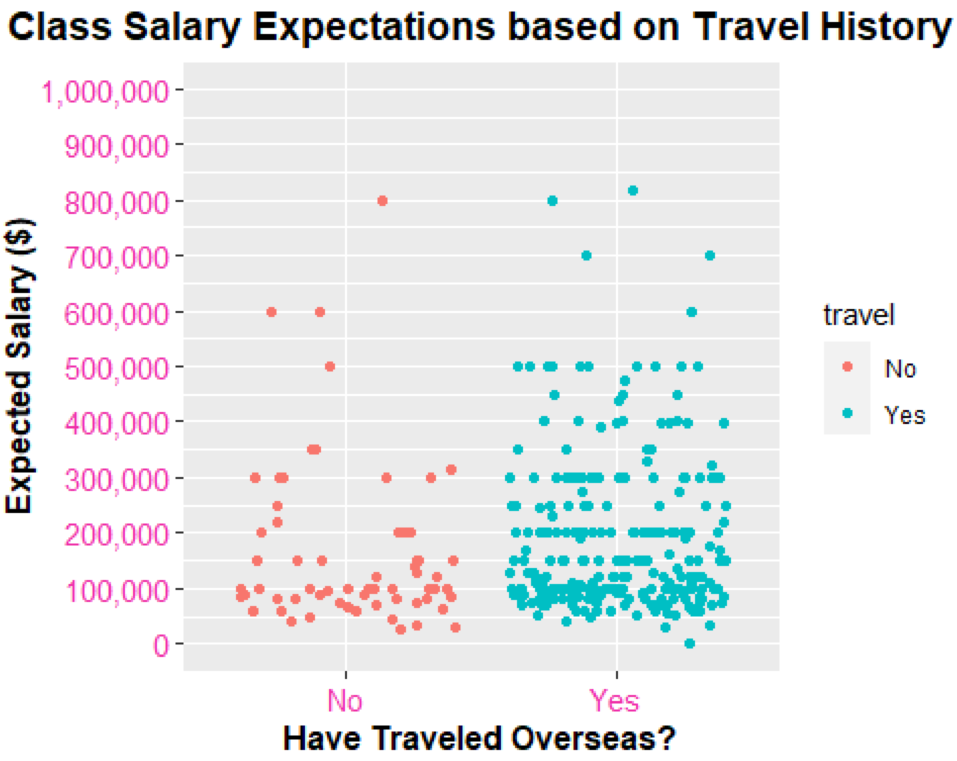
Let's try comparing students' expected salaries based on whether or not they have traveled overseas before.

Include an image of a jitter plot for these variables here, with salary played on the y-axis (sharing your code is optional).

- Color the points based on which travel group they are in (you can use the default colors or choose custom colors)
- jitter your points at a width of 0.05
- Use the limits argument to set the y axis to only span from 0 to 1 million dollars (this will leave out the 4 highest values and make the consensus data much easier to visualize!)
- Set the y axis breaks to be in increments of 100 thousand dollars
- Add an appropriate title and axes labels
- All other features optional!
- OPTIONAL: If you're curious how to turn off scientific notation and report comma form, try librarying the scales package and add labels = comma to your scale function. https://www.geeksforgeeks.org/change-formatting-of-numbers-of-ggplot2-plot-axis-in-r/

```
cat("\014")
```

```
shell("cls")

library(scales)

##
## Attaching package: 'scales'

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

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

plot_3 <- ggplot(data = class_Data, aes(x = travel, y = salary)) +
geom_jitter(aes(colour = travel)) + labs(title = "Class Salary Expectations
based on Travel History", x = "Have Traveled Overseas?", y = "Expected Salary
($)") + theme(plot.title = element_text(size = 14, hjust = 0.5, face =
"bold"), axis.title = element_text(size = 12, face = "bold"), axis.text =
element_text(size = 11, color = "maroon2")) + coord_cartesian(ylim=c(0,
1000000)) + scale_y_continuous(label=comma, breaks=seq(0,10000000,100000))

plot_3 + guides(fill = guide_legend(title = "Has the student travelled
overseas before?", title.position = "left"))
```



Question 4

Using a dplyr pipe, create a summary table that calculates the mean and median salary by travel. Add a filter option to only include salary levels below 1 million dollars (we will set a cut-off there so that our mean values aren't too susceptible to crazy high values). When you are done, you should have 4 values in a table style output, showing the mean and median salary of the "no" responses, and those for the "yes" responses.

Include (either directly copied, or screenshot) your summary table

Copy (or screenshot) the code you used to create that table

Briefly address these questions

- Do you think travel status explains any variability in students' projected salaries?

- What ideas or explanations do you have for any association or lack of association you see in the data?

```
cat("\014")
```

```
shell("cls")
class_Data2 <- class_Data
#View(class_Data2)

salary_travel_filter_table <-
class_Data2 %>%
  select(travel, salary) %>%
  filter(salary <= 1000000) %>%
  group_by(travel) %>%
  arrange(desc(salary)) %>%
  summarize(mean_salary = mean(salary), median_salary = median(salary)) #%>%
  #ungroup()
str(salary_travel_filter_table)
```

```
## tibble [2 × 3] (S3: tbl_df/tbl/data.frame)
##  $ travel       : chr [1:2] "No" "Yes"
##  $ mean_salary  : num [1:2] 165352 189917
##  $ median_salary: num [1:2] 100000 150000
```

```
#View(salary_travel_filter_table)

print("I think that travel status explains some increase in students’ self-
reported projected salaries. From the last graph, there seems to be an
increased density/ closer grouping of responses of those who report to have
travelled overseas. Also, these same responses tend to be of higher salary
values than those who have reported to not have travelled overseas.")
```

```
## [1] "I think that travel status explains some increase in students’ self-
reported projected salaries. From the last graph, there seems to be an
increased density/ closer grouping of responses of those who report to have
travelled overseas. Also, these same responses tend to be of higher salary
values than those who have reported to not have travelled overseas."
```

```
#aggregate(class_Data2$travel, class_Data2$salary, by=list(class_Data2$wage),
FUN=mean)
#aggregate(cbind(travel, salary) ~ academ_level + grad_plans, data =
class_Data2, FUN=mean)

#library(doBy)
#aggregate(cbind(travel, salary) ~ academ_level + grad_plans, data =
class_Data2, summary)
#summaryBy() function from the {doBy} package for 'travel' and 'salary'
variables by 'academ_level' and 'grad_plans'

#aggregate(class_Data2$travel, by=list(class_Data2$travel), FUN=summary)
#class_travel = subset(x = class_Data2, subset = travel >= 1)
#class_count = ifelse(class_Data2$travel >= 1, "1", "0")
#class_Data2 %>% summarise(mean(travel))
```

Question 5

Are students' minimum wage expectations associated with academic level?

Intermediate step: First, 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 pipe output that the order is correct! If you make a mistake and accidentally messed up something with the data, try re-importing the data again.

Include (either directly copied, or screenshot) your summary table

Copy (or screenshot) the code you used to create that table

Briefly address these questions:

Based on the summary stats, does there seem to be any association between academic level and minimum wage expectations for fast food jobs? How might you explain this result in context?

Why might the senior/grad students have such a high standard deviation compared to other groups? Hint: sort the wage column and scroll to the bottom!

```
cat("\014")
```

```
str(class_Data2)
```

```
## tibble [326 × 17] (S3: tbl_df/tbl/data.frame)
##  $ dist        : num [1:326] 120 139 6621 151 82 ...
##  $ bones       : num [1:326] 0 0 0 0 0 2 1 0 0 0 ...
##  $ wage        : num [1:326] 9 10 10 10 11 11 11 11 11 11 ...
##  $ sleep       : num [1:326] 5.5 8.5 6.5 5 9.5 5.2 9 0 6 8 ...
##  $ bpm         : num [1:326] 73 78 70 63 75 97 56 101 81 72 ...
##  $ shower      : num [1:326] 20 30 20 15 10 20 20 40 20 20 ...
##  $ salary      : num [1:326] 300000 200000 80000 120000 250000 75000
300000 70000 110000 80000 ...
##  $ travel      : chr [1:326] "Yes" "Yes" "Yes" "Yes" ...
##  $ academ_level: chr [1:326] "Freshman" "Freshman" "Freshman" "Sophomore"
...
##  $ academ_year : num [1:326] 1 1 1 2 1 1 1 3 1 1 ...
##  $ car         : chr [1:326] "Yes" "No" "No" "Yes" ...
##  $ grad_plans  : chr [1:326] "Medical School" "Graduate School" "Medical
School" "Medical School" ...
##  $ coffee      : Factor w/ 2 levels "No","Yes": 1 2 1 2 1 1 1 1 1 1 ...
##  $ residence   : chr [1:326] "Dorm Room" "Dorm Room" "Dorm Room" "Some
other residence" ...
##  $ section     : chr [1:326] "STAT 212 at 9am" "STAT 212 at 10am" "STAT
212 at 9am" "STAT 212 at 9am" ...
##  $ day         : chr [1:326] "Wednesday" "Saturday" "Wednesday"
"Wednesday" ...
##  $ letter      : chr [1:326] "A" "B" "A" "C" ...
```

```
#View(class_Data2)
#Class_S23$academ_level
shell("cls")

#Class_S23$academ_level = factor(Class_S23$academ_level, levels =
c("Freshman", "Sophomore", "Junior", "Senior or grad student"))

class_Data2$academ_level = factor(class_Data2$academ_level, levels =
c("Freshman", "Sophomore", "Junior", "Senior or grad student"))

wage_level_filter_table <-
class_Data2 %>%
  select(wage, academ_level, academ_year) %>%
  #filter(salary <= 1000000) %>%
  arrange(desc(academ_year), academ_level) %>%
  group_by(academ_year, academ_level) %>%
  summarize(mean_wage = mean(wage), median_wage = median(wage)) #%>%
```

```
## `summarise()` has grouped output by 'academ_year'. You can override using
the
## `.groups` argument.
```

```
  #cols <- c("academ_level","mean_wage", "median_wage") %>%
  #select(DF, !!cols)
  #ungroup()
head(wage_level_filter_table)
```

```
## # A tibble: 4 × 4
## # Groups:   academ_year [4]
##   academ_year academ_level           mean_wage median_wage
##         <dbl> <fct>                      <dbl>       <dbl>
## 1           1 Freshman                    14.0        14
## 2           2 Sophomore                   14.2        14
## 3           3 Junior                      15.8        15
## 4           4 Senior or grad student      21.3        15.2
```

```
#View(wage_level_filter_table)
print("Based on the summary stats,there seems to be an increasing
relationship between academic level and minimum wage expectations for fast
food jobs? This could be explained due to increased experience and
responsibility which - as far as explicitly this self-reported data is
concerned - are abstract variables not measured")
```

```
## [1] "Based on the summary stats,there seems to be an increasing
relationship between academic level and minimum wage expectations for fast
food jobs? This could be explained due to increased experience and
responsibility which - as far as explicitly this self-reported data is
concerned - are abstract variables not measured"
```

```
print("The senior/grad students might have such a high standard deviation
compared to other groupsdue to one of the seniors reporting $100.0/hr as an
expected hourly wage at a local fast food restaurant - this is generally
considered an incredibly ridiculous amount - attributing to the large
dispersion of wage values")
```

```
## [1] "The senior/grad students might have such a high standard deviation
compared to other groupsdue to one of the seniors reporting $100.0/hr as an
expected hourly wage at a local fast food restaurant - this is generally
considered an incredibly ridiculous amount - attributing to the large
dispersion of wage values"
```

Question 6

When asked to choose a letter at random, how did the class do? Create a univariate barplot showcases the results of the random letter question.

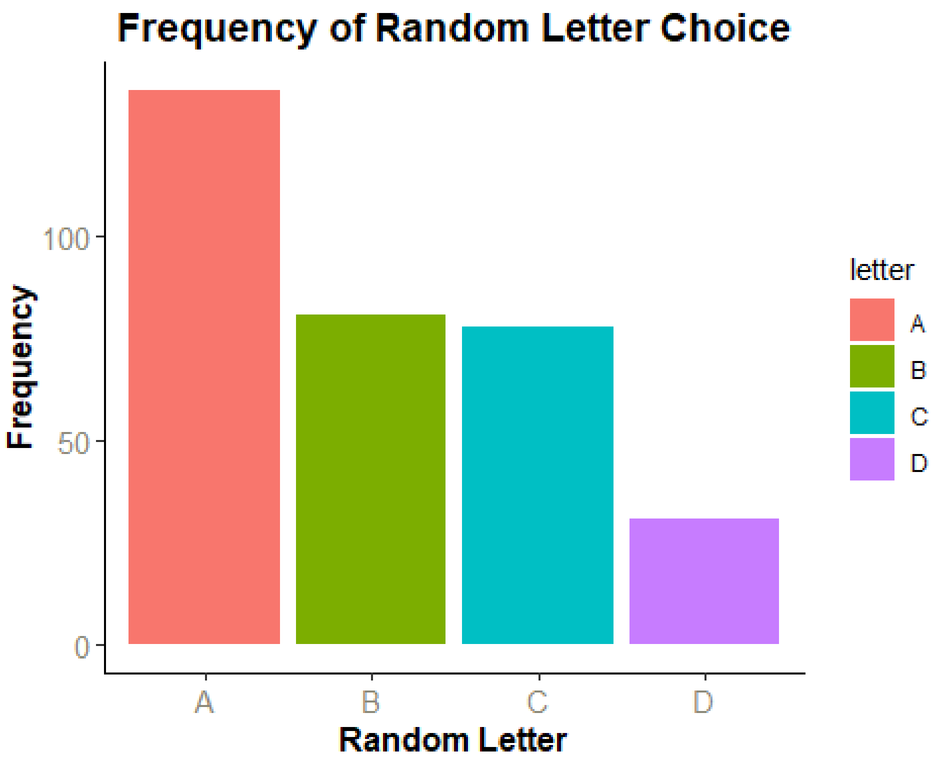
Include an image of your barplot (sharing your code is optional)

- Fill each bar a different color
- Use a color palette (or custom colors) for this plot
- Use a plot theme of your choice
- Add an appropriate title

Briefly comment on the plot. What do you notice? Is this what you expected? I would ask you if you have any explanation ideas, but I honestly have none! I've replicated this two semesters now and still don't know why this is what results. So if you have one, tell me!

```
cat("\014")
```

```
shell("cls")
ggplot(data = class_Data2, aes(x = letter, fill=letter)) + geom_bar() +
scale_color_brewer(palette = "Set2") + theme_classic() + labs(title =
"Frequency of Random Letter Choice", x = "Random Letter", y = "Frequency") +
theme(plot.title = element_text(size = 14, hjust = 0.5, face = "bold"),
axis.title = element_text(size = 12, face = "bold"), axis.text =
element_text(size = 11, color = "cornsilk4"))
```



```
#ggplot(data = class_Data2, aes(x = letter)) + geom_bar(stat = "identity",
color="blue", fill="red")
print("I do not notice a clear pattern, but if I had to conject I would say
that A is listed as the first letter, D is the last, and both B and C are
middle choices, and oftentimes people choose based on priority or in order of
options shown first to last.")
```

```
## [1] "I do not notice a clear pattern, but if I had to conject I would say
that A is listed as the first letter, D is the last, and both B and C are
middle choices, and oftentimes people choose based on priority or in order of
options shown first to last."
```

Question 7

Finally, let's explore the relationship of two categorical variables: academic level and whether or not a student owns a car. Create the appropriate graph to represent these two variables.

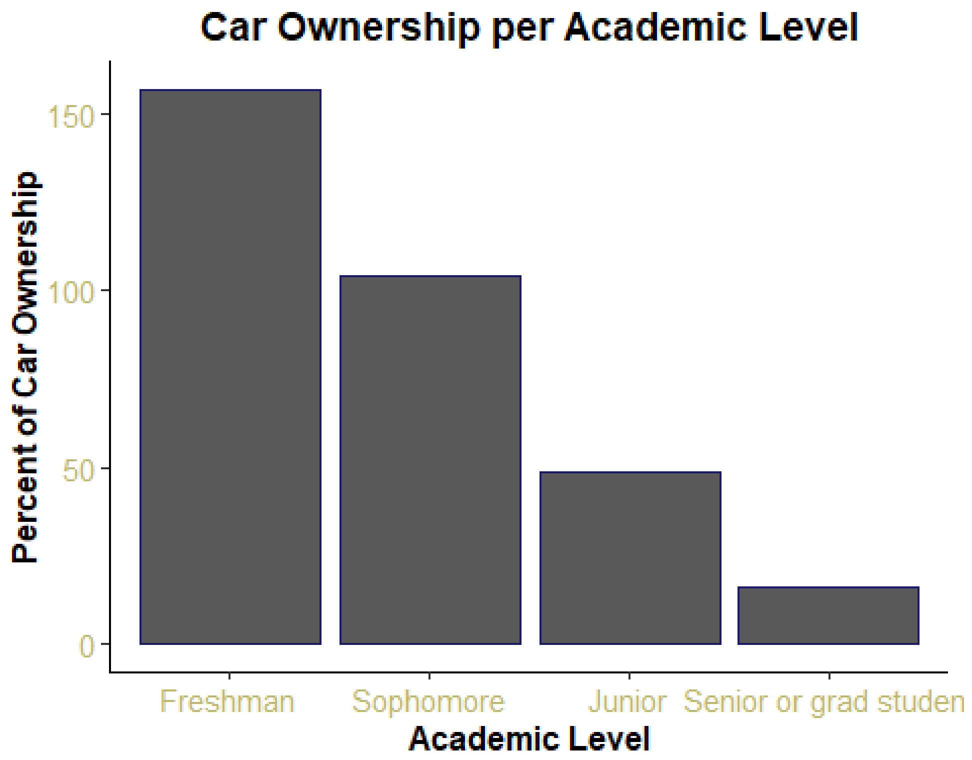
Include an image of your plot

- Use a color palette (or custom colors) for this plot.
- Add an appropriate title and an appropriate axis label for any axis a variable is assigned to
- Use a plot theme of your choice
- Use the theme function to center and bold the plot title

Briefly address this question: 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.

```
cat("\014")
```

```
shell("cls")
#ggplot(class_Data2, aes(x = academ_level, y = car)) +
geom_point(aes(color=academ_year)) + geom_smooth(method=lm) +
labs(stat="identity", title = "academic level and whether or not a student
owns a car", x = "Academic Level", y = "Car Ownership")

#class_Data2 %>%
#  ggplot() +
#  aes(x = academ_level, y = car) +
#  geom_histogram(colour='darkkhaki', aes(fill=academ_year)) +
#  #geom_smooth(method = 'loess', color='mediumvioletred') +
#  labs(stat="identity", title = "academic level and whether or not a student
owns a car", x = "Academic Level", y = "Car Ownership")

#ggplot(data=class_Data2, aes(x = academ_level)) +
geom_histogram(color='black', fill='mediumorchid4', bins=80, binwidth = 120)
+ labs(stat="identity", title = "academic level and whether or not a student
owns a car", x = "Academic Level", y = "Car Ownership")

#ggplot(data=class_Data2, aes(x=academ_level)) +
geom_bar(colour='midnightblue', aes(fill=car)) + labs(stat="identity", title
= "Car Ownership per Academic Level", x = "Academic Level", y = "Percent of
Car Ownership") +theme_classic() + theme(plot.title = element_text(size = 14,
hjust = 0.5, face = "bold"), axis.title = element_text(size = 12, face =
"bold"), axis.text = element_text(size = 11, color = "darkkhaki")) +
scale_color_brewer(palette = "Set2")
ggplot(data=class_Data2, aes(x=academ_level)) + scale_color_brewer(palette =
"Set2") + theme_classic() + geom_bar(colour='midnightblue') +
labs(stat="identity", title = "Car Ownership per Academic Level", x =
"Academic Level", y = "Percent of Car Ownership") + theme_classic() +
theme(plot.title = element_text(size = 14, hjust = 0.5, face = "bold"),
axis.title = element_text(size = 12, face = "bold"), axis.text =
element_text(size = 11, color = "darkkhaki"))
```



```
print("There does not appear to be any association between students’ academic
level and car ownership status since there is no consistent trend as academic
year/level increases.**")
```

```
## [1] "There does not appear to be any association between students’
academic level and car ownership status since there is no consistent trend as
academic year/level increases.**"
```