Homework 1: Due Wednesday, February 8 at 11:59pm

For this assignment, you will be analyzing the **`College`** data frame included in the **`ISLR`** package, which contains 777 observations on 18 features of U.S. colleges from the 1995 issue of U.S. News and World Report. The goal of this assignment is to become more familiar with R and data visualization, and therefore you will need to use resources such as the R **`help()`** function and **`ggplot2`** cheat sheet to learn how to create different types of plots. All analyses must be performed in R using **`tidyverse`** and other packages discussed in class. Provide your responses (including R code pasted in text format) in the designated spaces in this Word document, and then save it as a pdf and upload it to Canvas.

List column eatures:

```
> colnames(College)
 [1] "Private"     "Apps"        "Accept"      "Enroll"      "Top10perc"   "Top25perc"   "F.Undergrad" "P.Undergrad"
 [9] "Outstate"    "Room.Board"  "Books"       "Personal"    "PhD"         "Terminal"    "S.F.Ratio"   "perc.alumni"
[17] "Expend"      "Grad.Rate"
```

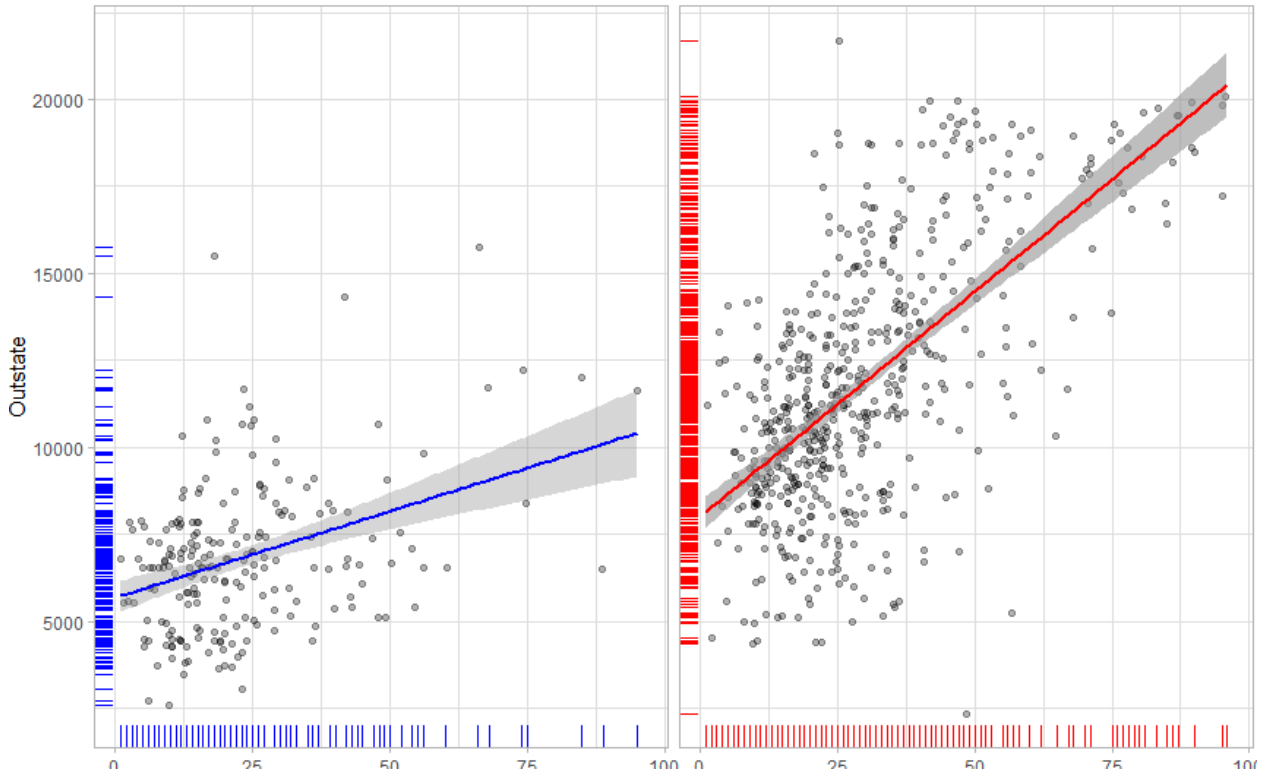
1. **[15%]** Generate jittered scatterplots of out-of-state tuition as a function of the percentage of new students who are from the top 10% of their high school classes that are faceted by public colleges (left facet) and private colleges (right facet). Overlay these scatterplots with straight lines (not smoothed lines, hint: **`help(geom_smooth)`**) containing 95% confidence bands, with lines colored differently for each facet. Add rug plots to both facets, using the same colors as for the lines, and set the background of the plotting region to white rather than the default color of gray.

   Provide code below:

```
#1
# Create jittered  scatterplot
ggplot(College, aes(x = Top10perc, y = Outstate)) +
  geom_jitter(alpha = 0.3) +
  facet_grid(. ~ Private) +
  geom_smooth(method = "lm", se = TRUE, color = "blue") +
  geom_smooth(method = "lm", se = TRUE, color = "red", formula = y ~
x, data = subset(College, Private == "Yes")) +
  theme_light()+
  theme(strip.background = element_blank())+
  geom_rug(color = "blue")+
  geom_rug(data = subset(College, Private == "Yes"), color = "red")
```

```
# Create jittered  scatterplot
ggplot(College, aes(x = Top10perc, y = Outstate)) +
  geom_jitter(alpha = 0.3) +
  facet_grid(. ~ Private) +
  geom_smooth(method = "lm", se = TRUE, color = "blue") +
  geom_smooth(method = "lm", se = TRUE, color = "red", formula = y ~ x, data = subset(College, Private == "Yes")) +
  theme_light()+
  theme(strip.background = element_blank())+
  geom_rug(color = "blue")+
  geom_rug(data = subset(College, Private == "Yes"), color = "red")
```

   Provide figure below:



2. **[15%]** Compute two correlation coefficients between the percentage of new students from the top 10% of their high school classes and out-of-state tuition – one for public colleges, and one for private colleges. What do both correlation coefficients say about the general relationship between these features? Is this relationship stronger for public or private colleges? Provide an explanation for this difference that is based on your examination of the plot from question 1 (hint: compare points, confidence intervals, and/or rug plots).

Provide code and console output below:

```
#2
# Correlation coefficient for public colleges
cor(subset(College, Private == "No")$Top10perc, subset(College, Private == "No")$Outstate)
# Correlation coefficient for private colleges
cor(subset(College, Private == "Yes")$Top10perc, subset(College, Private == "Yes")$Outstate)
```

```
# Correlation coefficient for public colleges
cor(subset(College, Private == "No")$Top10perc, subset(College, Private == "No")$Outstate)

# Correlation coefficient for private colleges
cor(subset(College, Private == "Yes")$Top10perc, subset(College, Private == "Yes")$Outstate)
```

Provide answer below:

Correlation coefficient for public colleges

```
[1] 0.3748554
```

```
Correlation coefficient for private colleges
[1] 0.6222538
```

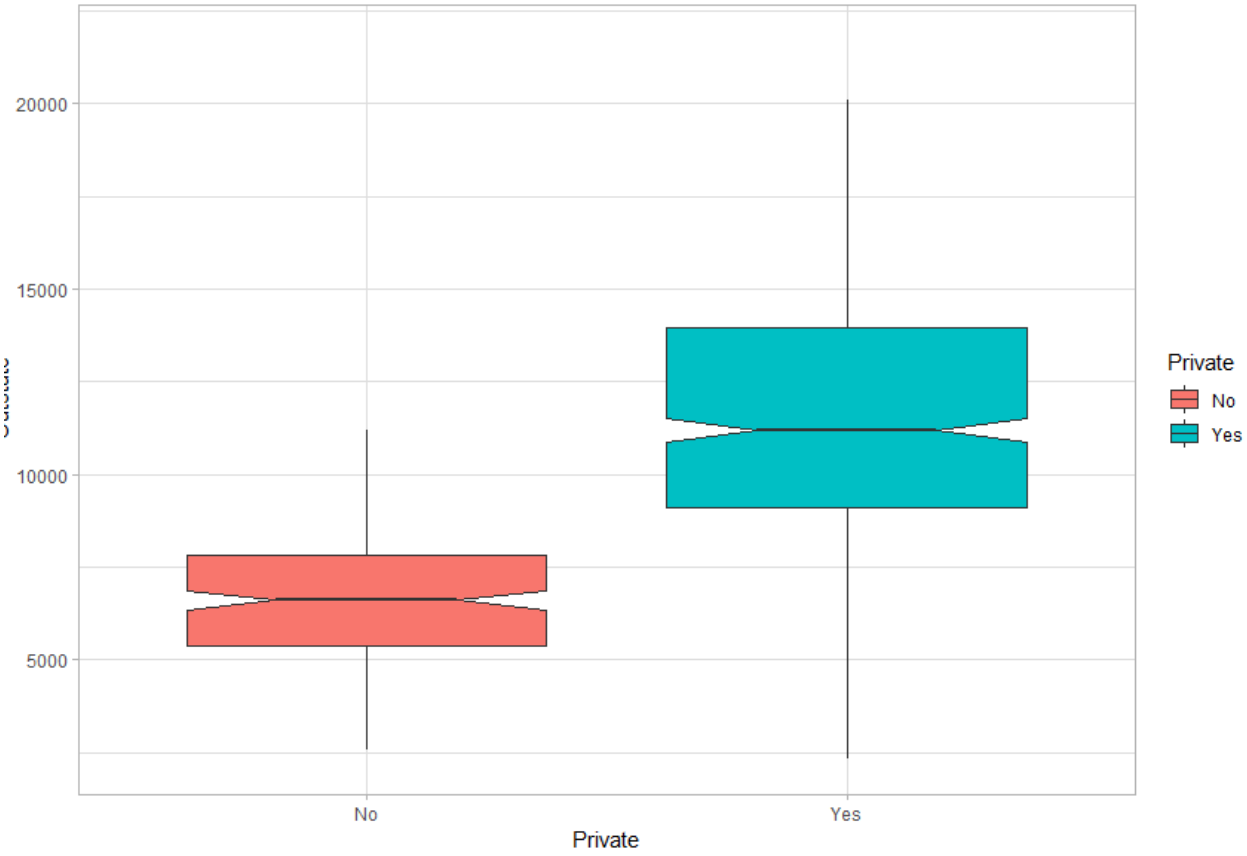
3. **[15%]** Generate a pair of box plots showing the distributions of out-of-state tuition for public and private colleges, using a different color to fill each box plot. Include notches and remove outliers (hint: **`help(geom_boxplot)`**) for easier comparison of the distributions between the two groups, and set the background of the plotting region to white rather than the default color of gray. Non-overlapping notches of box plots tell us that the distributions of two groups are significantly different from one another. Based on this information, are the distributions of out-of-state tuition significantly different between public and private colleges? What can you conclude about the difference between out-of-state tuition at public and private colleges?

Provide code below:

```
#3
# Box plots showing the distributions of out-of-state tuition for
public and private colleges,
ggplot(College, aes(x = Private, y = Outstate, fill = Private)) +
  geom_boxplot(outlier.shape = NA, notch = TRUE) +
  theme_light()+
  theme(strip.background = element_blank())
```

```
ggplot(College, aes(x = Private, y = Outstate, fill = Private)) +
  geom_boxplot(outlier.shape = NA, notch = TRUE) +
  theme_light()+
  theme(strip.background = element_blank())
```

Provide figure below:



Provide answer below:
The data shows that private colleges tend to have higher out-of-state tuition than public colleges

4. **[15%]** Generate a pair of violin plots showing the distributions of out-of-state tuition for public and private colleges, using a different color to fill each violin plot. Set the background of the plotting region to white rather than the default color of gray. What is the difference in information displayed by box plots and violin plots? Based only on these violin plots, what can you say about the difference between out-of-state tuition at public and private colleges?

Provide code below:

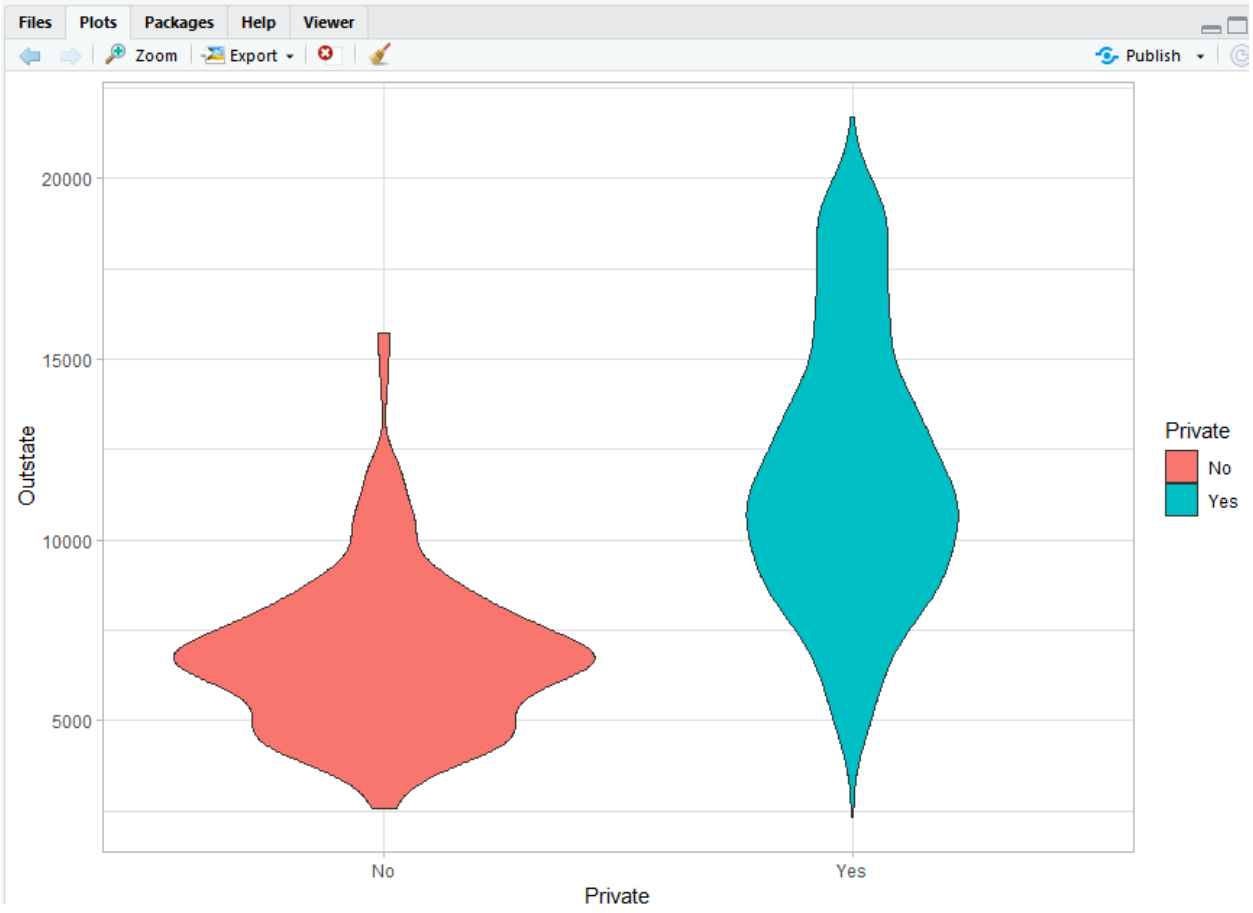
```
#4
# Pair of violin plots showing the distributions of out-of-state
tuition for public and private colleges.

ggplot(College, aes(x = Private, y = Outstate, fill = Private)) +
  geom_violin(trim = TRUE) +
  theme_light() +
  theme(strip.background = element_blank())
```

```r
#4
# Pair of violin plots showing the distributions of out-of-state tuition for public and private colleges.

ggplot(College, aes(x = Private, y = Outstate, fill = Private)) +
  geom_violin(trim = TRUE) +
  theme_light() +
  theme(strip.background = element_blank())
```

Provide figure below:



Provide answer below:

Based on the violin plots, we can see that the distribution of out-of-state tuition at private colleges has a longer tail on the right side, indicating that there are more private colleges with higher out-of-state tuition than public colleges.

5. **[15%]** Generate a #, with points colored in gray. Overlay these strip plots with box plots, using a transparency level of 0.5 and a different color to fill each box plot. Include notches, color outliers in red, and set the background of the plotting region to white rather than the default color of gray. Based on these plots, are book costs generally higher at public or private colleges? Is the college with the highest book costs public or private?

Provide code below:

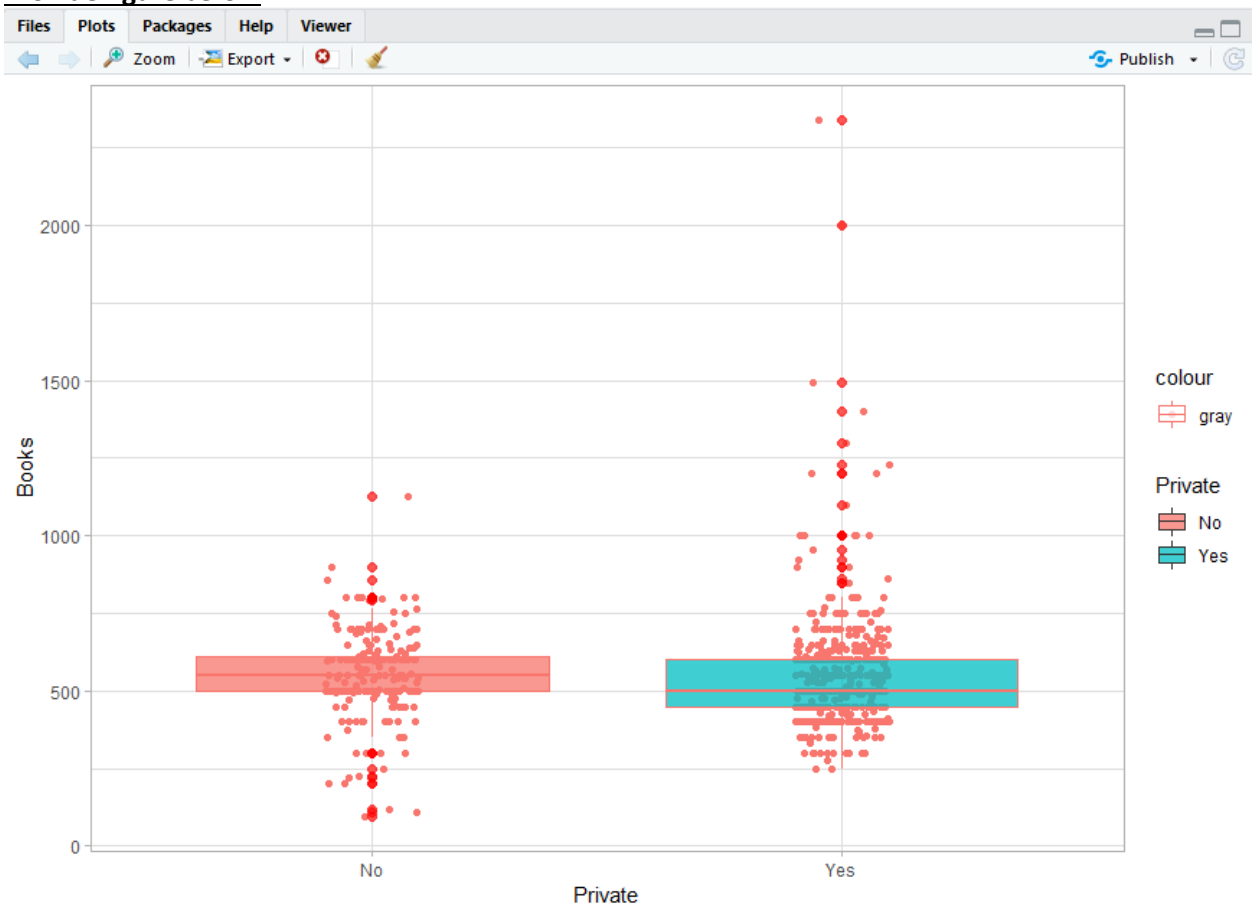
```
#5
# Pair of strip plots showing the distributions of book costs for
public and private colleges.

ggplot(College, aes(x = Private, y = Books, color = "gray")) +
  geom_jitter(width = 0.1) +
  geom_boxplot(alpha = 0.5, aes(fill = Private)) +
  geom_boxplot(alpha = 0.5, aes(fill = Private), outlier.color =
"red", outlier.size = 2) +
  theme_light() +
  theme(strip.background = element_blank())
```

```
#5
# Pair of strip plots showing the distributions of book costs for public and private colleges.

ggplot(College, aes(x = Private, y = Books, color = "gray")) +
  geom_jitter(width = 0.1) +
  geom_boxplot(alpha = 0.5, aes(fill = Private)) +
  geom_boxplot(alpha = 0.5, aes(fill = Private), outlier.color = "red", outlier.size = 2) +
  theme_light() +
  theme(strip.background = element_blank())
```

Provide figure below:



Provide answer below:
We can see that the book costs are generally higher at private colleges than public colleges.

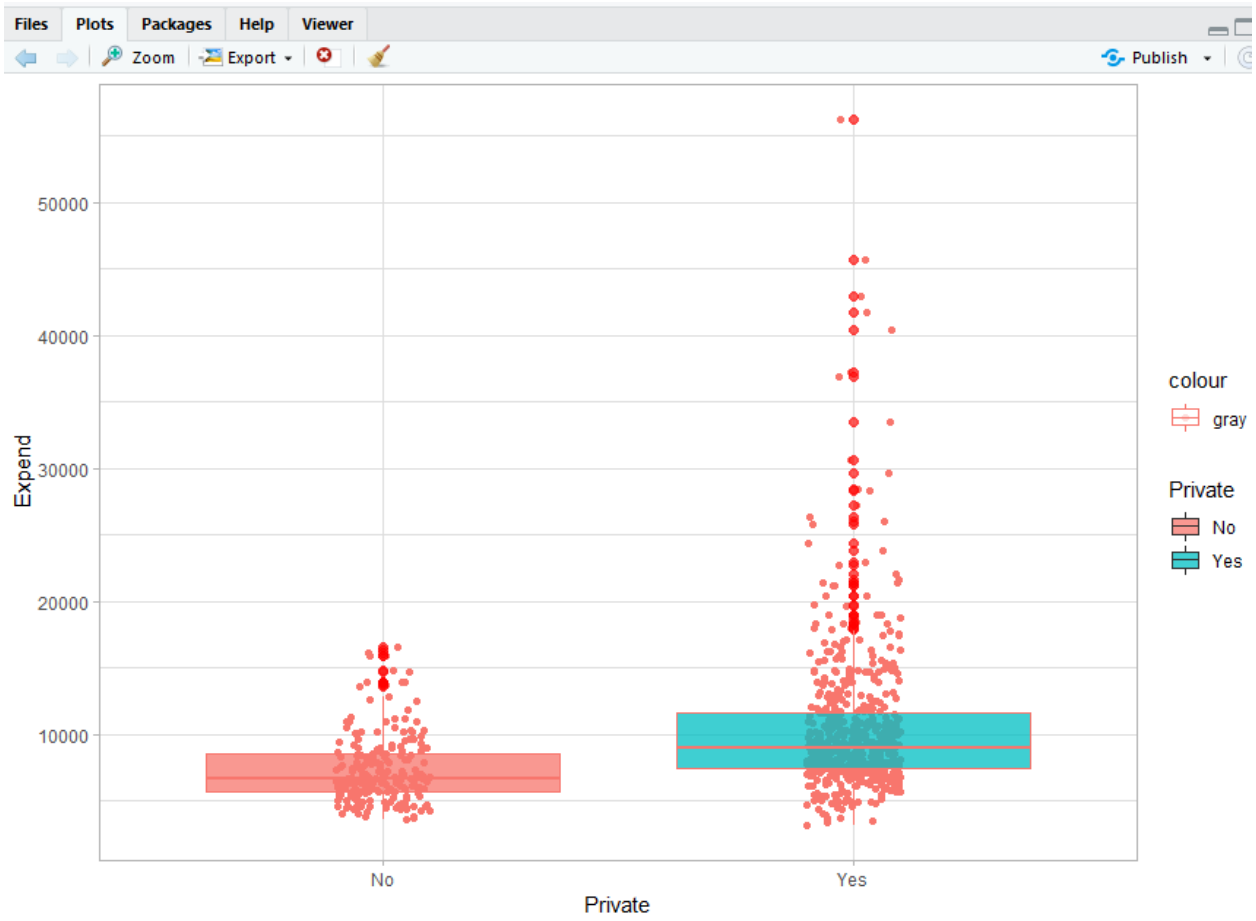
6. **[15%]** Generate a pair of jittered strip plots showing the distributions of total expenses (out-of-state tuition, room and board, books, and personal spending) for public and private colleges, with points colored in gray. Overlay these strip plots with box plots, using a transparency level of 0.5 and a different color to fill each box plot. Include notches, color outliers in red, and set the background of the plotting region to white rather than the default color of gray. Based on these plots, are total expenses generally higher at public or private colleges? Are there more outliers in total expenses for public or private colleges?

Provide code below:

```
#6
# pair of jittered strip plots showing the distributions of total
expenses
#  (out-of-state tuition, room and board, books, and personal
spending) for public and private colleges
ggplot(College, aes(x = Private, y = Expend, color = "gray")) +
  geom_jitter(width = 0.1) +
  geom_boxplot(alpha = 0.5, aes(fill = Private)) +
  geom_boxplot(alpha = 0.5, aes(fill = Private), outlier.color =
"red", outlier.size = 2) +
  theme_light() +
  theme(strip.background = element_blank())
```

```
#6
# pair of jittered strip plots showing the distributions of total expenses
# (out-of-state tuition, room and board, books, and personal spending) for public and private colleges
ggplot(College, aes(x = Private, y = Expend, color = "gray")) +
  geom_jitter(width = 0.1) +
  geom_boxplot(alpha = 0.5, aes(fill = Private)) +
  geom_boxplot(alpha = 0.5, aes(fill = Private), outlier.color = "red", outlier.size = 2) +
  theme_light() +
  theme(strip.background = element_blank())
```

Provide figure below:



Provide answer below:

we can see that the total expenses are generally higher at private colleges than public colleges.

7. **[10%]** As discussed in class, an important part of being a data scientist is communicating your findings. Therefore, summarize the dataset and your findings from question 1-6 in a short paragraph. Do not simply copy your answers from above, but rather assume that you need to briefly explain your analysis to an audience with no prior knowledge of the dataset or data science.

Provide answer below:

The first inference shows plots for both public and private colleges. The scatter plot shows the out-of-state tuition on the y-axis and the percent of new students from the top 10% of their high school classes on the x-axis. For each facet, there is a straight-line layer with 95% confidence bands, denoted by a different color. Rug plots with the same color as lines are also available. It can be concluded that private colleges attract more students than public colleges.

The correlation coefficient is a statistical measure that describes the strength and direction of a linear relationship between two variables. In our case, both correlation coefficients indicate that out-of-state tuition for public and private colleges is positively correlated with the percentage of new students

from the top 10% of their high school classes. The correlation coefficient for public colleges is stronger than the private colleges. It is possible though that we can contribute the analysis that private colleges that may tend to have more selective admissions processes, and therefore a higher percentage of top 10% students may be more strongly related to tuition costs at private colleges than at public colleges?

The third analysis shows the distributions of out-of-state tuition for public and private colleges. There is a measure of uncertainty around the median represented by the notches in the box plot. There is a significant difference between public and private colleges in the distribution of out-of-state tuition.

In the fourth inference, we show the out-of-state tuition distributions for public and private colleges using violin plots. Violin plots indicate that the distribution of out-of-state tuition at private colleges has a longer tail on the right side, suggesting that there are more private colleges with higher out-of-state tuition than public colleges. In addition, it shows that out-of-state tuition at private colleges is more evenly distributed than that at public colleges. As a result, out-of-state tuition at private colleges is generally higher than that at public colleges.

In the fifth plot, gray points indicate distributions of book costs between public and private colleges. In general, private colleges charge higher book prices than public universities. As can be seen, private colleges' boxplots have higher medians (the line inside the box) and the boxes are generally taller than those of public colleges. As can be seen from the boxplot, the college with the highest book costs is a private college.

Strip plots showing total expenses (out-of-state tuition, room and board, books, and personal spending) for public and private colleges are used for the final inference. Private colleges have generally higher expenses than public colleges, as shown by these plots. Private colleges' box plots have higher medians (the line inside the box) and are generally taller than public colleges'. Compared to public colleges, private colleges have a higher number of outliers.

Finally, we can conclude that private schools are more sought after, yet they are more expensive and more costly for students in the top 90 percentile.