# Homework Assignment #1

Nicholas Noel & Lissette Villa

Due February 10, 2023

```
library(tidyverse)
library(ISLR2)
```

## Instructions

You should submit either two or three files:

1. You should write your solutions to the Simulation and Applied Problems in this R Markdown file and submit the (.Rmd) file.
2. You should knit the final solution file to pdf and submit the pdf. If you are having trouble getting code chunks to run, add `eval = FALSE` to the chunks that do not run. If you are having trouble getting R Studio to play nice with your LaTeX distribution, I will begrudgingly accept an HTML file instead.
3. Solutions to the Key Terms and Conceptual Problems can be submitted in a separate Word or pdf file or included in the same files as your solutions to the Simulation and Applied Problems.

This homework assignment is worth a total of **40 points**.

## Key Terms (5 pts)

Read Chapter 2 of Introduction to Statistical Learning, Second Edition. Based on your reading, answer the following questions.

1. What is the difference between an *input variable* and an *output variable* in a model? Provide synonyms for each term.

Answer: An input variable is typically our X, or predictors, there can be multiple, whereas the output variable is denoted as Y and is the response variable. The reading uses the idea of sales being Y, our output, and variables such as tv and radio budget being X, or input variables.

2. What is the difference between *reducible error* and *irreducible error*? Give an example (other than those given in the book) of a situation in which the irreducible error is greater than zero.

Answer: Reducible error comes from inaccuracy in the estimate that can be reduced by using the appropriate statistical technique and irreducible error comes during the data collection process because of influential but unmeasured variables and variability in the collected data. Irreducible error cannot be eliminated. An example would be in a study about teenager growth patterns in America, their nutrition may not be recorded exactly which would be a variable influencing their growth and therefore would introduce irreducible error.

3. Generally, what types of questions are answered using *inference* and what types are answered using *prediction*? Is it possible to use the same model for both inference and prediction?

Answer: In general, questions that would be answered in inference would include trying to understand the deeper relationship between our predictors, x, and our response, y, whereas in prediction we are less concerned with the relationship but instead how the predictors, x, lead to the outcome, y. The reading uses the example as inference asks the relationship between x and y and prediction asks what values of x leads to y. Also, it is

possible to use the same model for both inference and prediction, just depending on how one interprets as well as the complexity of the model used.

4. Generally, what types of prediction questions are answered using *regression* methods and what types are answered using *classification* methods?

Answer: The types of prediction questions answered using regression methods are those that have a quantitative response and prediction questions with qualitative responses are generally answered using classification methods. Although some methods can be used for either response type.

5. What are the major advantages of using a *nonparametric* method over a *parametric* method? What are the disadvantages?

Answer: The major advantages of nonparametric methods is the fact that the problem is reduced to simply finding the parameters instead of the function itself, this is great when we are able to assume the function has a form we can understand such as a linear form but when the form of the data is more complex, our linear model could become a super complicated 13th Taylor polynomial that breaks when one more piece of data is added(over fitting). When the form of the data is complex a nonparametric approach is generally better to use.

6. In prediction, we typically aim to minimize a *loss function* that more-or-less represents the total error in our predictions. Give one example each for regression and classification problems of a measure of model (in)accuracy.

Answer: A measure of model accuracy for regression problems is mean squared error and for classification problems the most common measure used is error rate. In both cases, the MSE and error rate should be computed using test data rather than data from the training set.

7. Why do we only fit the model on a *training set*? What do we do with the rest of the data?

Answer: We only fit the model of a training set as it has the majority of the data. This allows us to test the rest of the data into the model to determine how well the model is doing using real collected data.

8. Generally, as a model becomes more complex, what happens to the *bias* of the model and why? What happens to the *variance* of the model and why?

Answer: As a model becomes more complex, or flexible, the bias of the model will decrease as it will more closely follow the training data which often has non-linear relationships. However, as the model becomes more complex, it will have higher variance as it follows the training data because its flexibility makes the model sensitive to changes in the training data that would not have a major effect on more simplified, rigid models.

9. What is meant by the term *overfitting*? Explain this in terms of the bias-variance trade-off.

Answer: Overfitting is when too much power is given to the training data set that inevitably leads to large variance when cross validating. The best way to see this is when the training dataset changes considerably as a result from new or different data values being used when modeling. This can also be seen in the bias when working with real life examples of the data in which our error rate becomes increasingly large implying that the data is not following a trend but instead our model is solely basing its' values on that of the training data (Very clear when cross validating to determine whether or not we are overfitting).

10. Briefly explain how a *Bayes classifier* works.

Answer: Bayes classifier works by assigning an observation to a class based on the probability that an observation belongs to that class given an observed predictor variable. The threshold that determines to which class an observation is assigned is called the Bayes decision boundary.

# Conceptual Problems

## Conceptual Problem 1 (4 pts)

Write me a brief (2-3 paragraphs) summary of what you learned in the P-Values and Power in-class activity about how the distribution of p-values (over very many tests) is affected by the validity/violation of test assumptions and the power of the test. Did anything surprise you or clarify a concept for you? Support your writing with a few graphs you produced in class (it is easiest to copy and re-run the relevant code chunks).

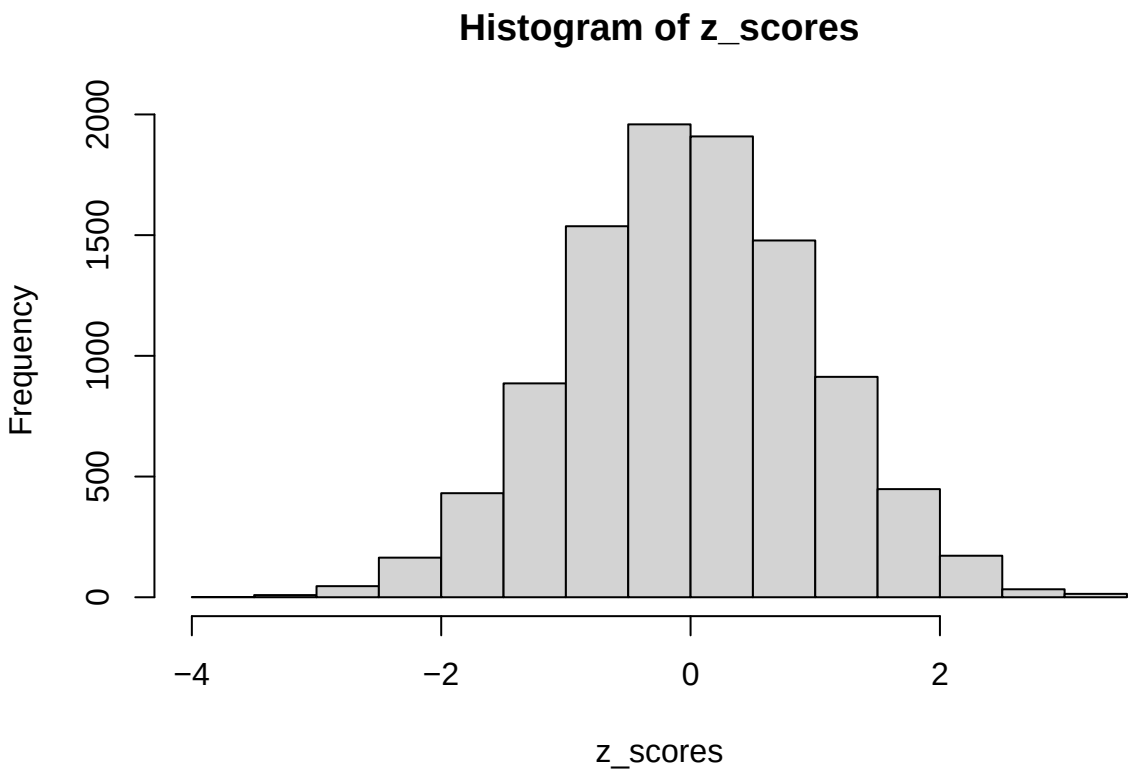
The understanding of p-values as a probability is part of the fundamentals of understanding statistics. I enjoyed the idea of how p-values are a function of z-scores and population sizes. The basis of p-values are the inputs, which produce a value between 0 and 1. When all of the assumptions of a hypothesis test are met, the distribution of that probability is evenly, uniformly, distributed. This is the foundation of meeting all of the assumptions but the curve occurs when we break assumptions.

When assumptions are broken we were still seeing a uniformly distribution of p-values but only when our n, sample size, was large. The reason we were seeing the the same uniform distribution is because a large sample size meant that whatever we were testing, i.e. a two sample t-test on two populations in which the distribution was binomial was the difference between the values were still normal! This is how the central limit theorem covered us.The general rule that we decided on is so long as the z-score is roughly normally distributed, we can trust the p-value since it would be uniformly random so long as our sample size is large, our z-scores will be normally distributed despite the population parameters not being normally distributed.

This idea does not hold for when our n is small since the differences between the two populations is much less likely to be normal. Power was incorporated into this idea as the activity taught us what power really was. One of the key summaries that made sense was understanding "if H_a is correct and we have 80% power, 80% of p-values will be below 0.05." This intuitively makes sense and clarified how the ability of a test to reject H_0 when H_a is true, or also, to prevent the likelihood of having a type II error occur.

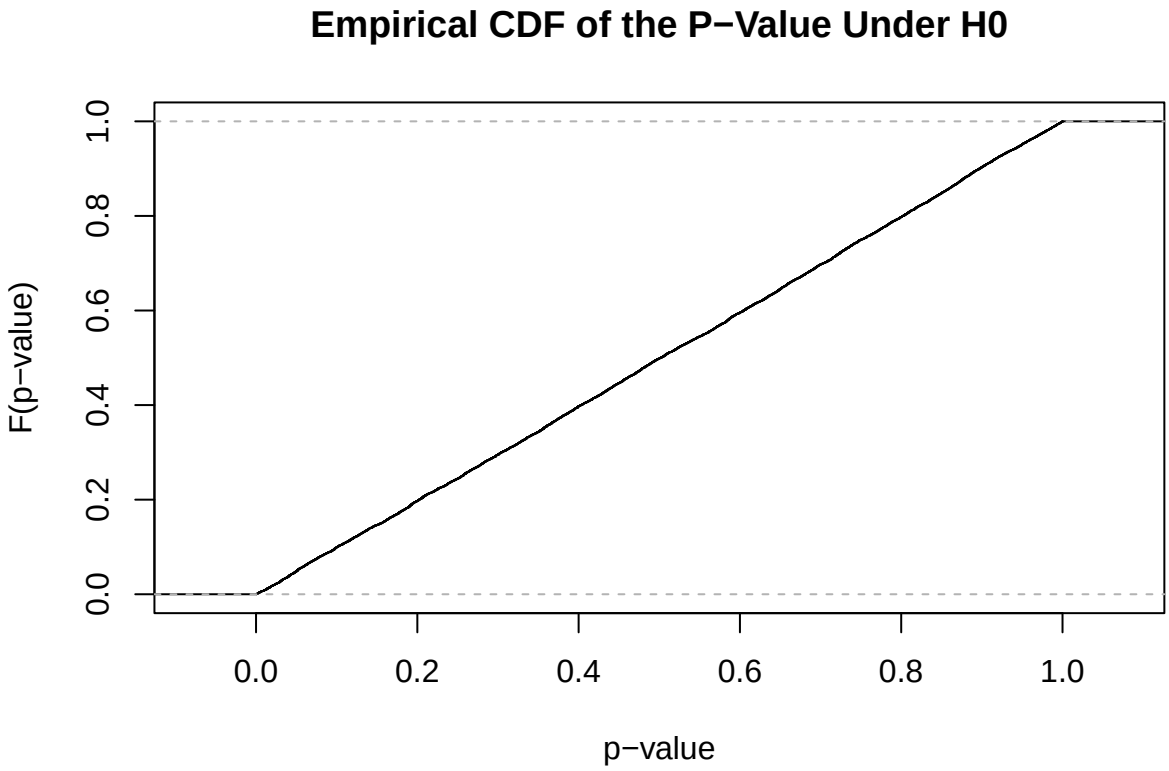
```
set.seed(23)
n <- 100
mu <- 70
sigma <- 10

xbar <- numeric(10000) # empty vector to store sample means in

for(i in 1:10000){
  sim_values <- rnorm(n, mean = mu, sd = sigma)
  xbar[i] <- mean(sim_values)
}
```

```
z_scores <- (xbar - mu)/(sigma/sqrt(n))

hist(z_scores)
```




```r
p_lower <- pnorm(z_scores, mean = 0, sd = 1, lower.tail = TRUE)
p_upper <- pnorm(z_scores, mean = 0, sd = 1, lower.tail = FALSE)
two_sided_p_value <- 2*pmin(p_lower, p_upper)
# pmin is a pairwise minimum, comparing the values at each index in the vectors
# Here it will return p_lower when z < 0 and p_upper when z > 0

hist(two_sided_p_value)
```


```
plot(ecdf(two_sided_p_value),
     xlab = "p-value",
     ylab = "F(p-value)",
     main = "Empirical CDF of the P-Value Under H0")
```




## Conceptual Problem 2 (3 pts)

Textbook Exercise 2.4.4

4. You will now think of some real-life applications for statistical learning.

(a) Describe three real-life applications in which classification might be useful. Describe the response, as well as the predictors. Is the goal of each application inference or prediction? Explain your answer. Answer: Classification might be useful in:

1. Predicting a persons risk for developing a cardiovascular disease which would be the response and some predictors might be genetics, BMI, and the presence of comorbidities such as abdominal obesity.
2. Determining the risk level of a customer buying auto insurance. The response would be the likelihood of having a claim. The predictors would be their age, driving record, average daily miles driven, car make and model, etc. The goal of this application is prediction of the frequency of a claim which would thereby help the insurer determine the buyers coverage and payments.
3. Inferring the relationship between a persons diet, physical activity level, and intellectual stimulation and the types of disease or ailments they develop later in life. The response variable would be some measure of health, a simple one would be if they have a disease or not. Some predictors would be the types of foods they eat, how often they exercise and for how long, their level of education, and their job type.

(b) Describe three real-life applications in which regression might be useful. Describe the response, as well as the predictors. Is the goal of each application inference or prediction? Explain your answer. Answer: Regression might be useful for:

1. Predicting a persons future income. The response would be their yearly salary. The predictors would be their level of education, field of work, location of residence, and the location of their workplace.

2. Predicting the final score of a basketball game. The response would be the number of points scored. Some predictors would be the team playing, their previous scores, the presence of high scoring players, and if it's a home or away game.
3. Making inferences related to the number of hot dogs someone buys. The response would be number of hot dogs purchased in a month. Some predictors would be the person's yearly income, their nationality, their current location of residence, if they have a Costco membership, the month or season, if they attend baseball games or not, etc.

(c) Describe three real-life applications in which cluster analysis might be useful. Answer: Cluster analysis might be useful for:

1. Grouping loan applicants into high and low risk categories for defaulting on a loan.
2. Grouping potential customers into appropriate groups to effectively advertise to them based on factors such as age, gender, and income.
3. Determining geographic areas that would utilize public transit based on factors such as population density and the average income of a resident.

### Conceptual Problem 3 (3 pts)

Textbook Exercise 13.7.2

## Simulation Problems

### Simulation Problem 1 (Code: 4 pts; Explanation: 6 pts)

From the Parametric vs. Nonparametric Tests: Two-Sample Tests activity, copy to this homework your simulation code/results from the *Assumptions Violated, Ha True* section of each test as well as the results tables for all simulations (in the Class Results section).

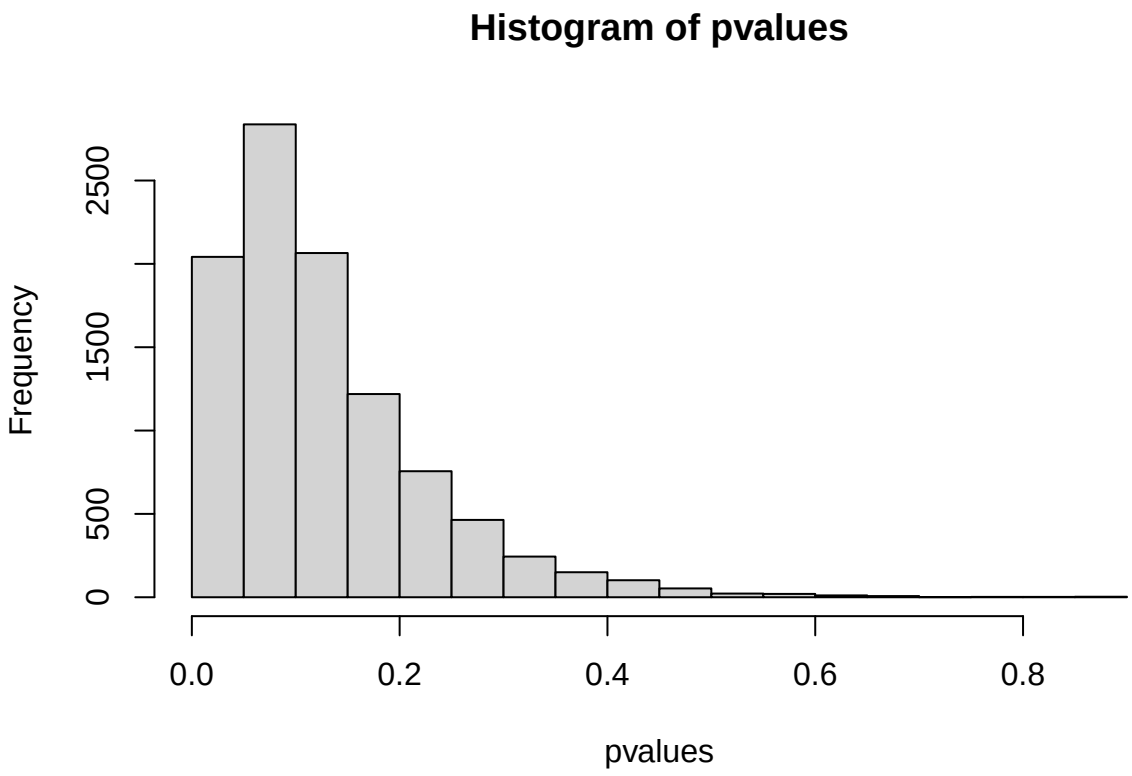
```
pvalues <- numeric(length = 10000)
```

---

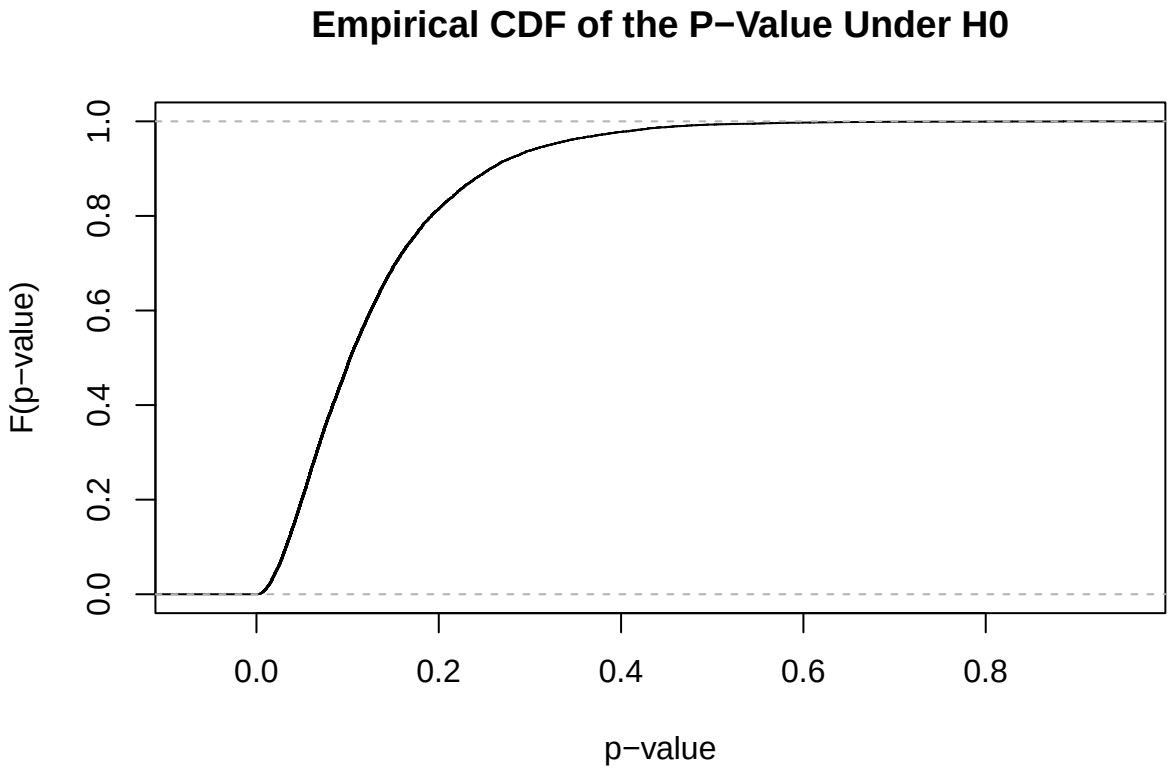
t.test section:

```
nG <- 10
d <- 0.8
for (i in 1:length(pvalues)){
  set.seed(i)  # notice that the seed changes every time inside the for loop
  # you could also set a single seed outside the for loop

  # Create the vectors x and y
  x <- c(rnorm(nG*0.9, mean = 0, sd = sqrt(0.19)),
         rnorm(nG*0.1, mean = 3, sd = sqrt(0.19))
)
  y <- c(rnorm(nG*0.9, mean = d, sd = sqrt(0.19)),
         rnorm(nG*0.1, mean = 3 + d, sd = sqrt(0.19))
)

  # Perform the t-test and get the p-value
  pvalues[i] <- t.test(x, y, alternative = "t")$p.value

}
```

```
# histogram of the p-values under H0
hist(pvalues)
```




```
plot(ecdf(pvalues),
     xlab = "p-value",
     ylab = "F(p-value)",
     main = "Empirical CDF of the P-Value Under H0")
```



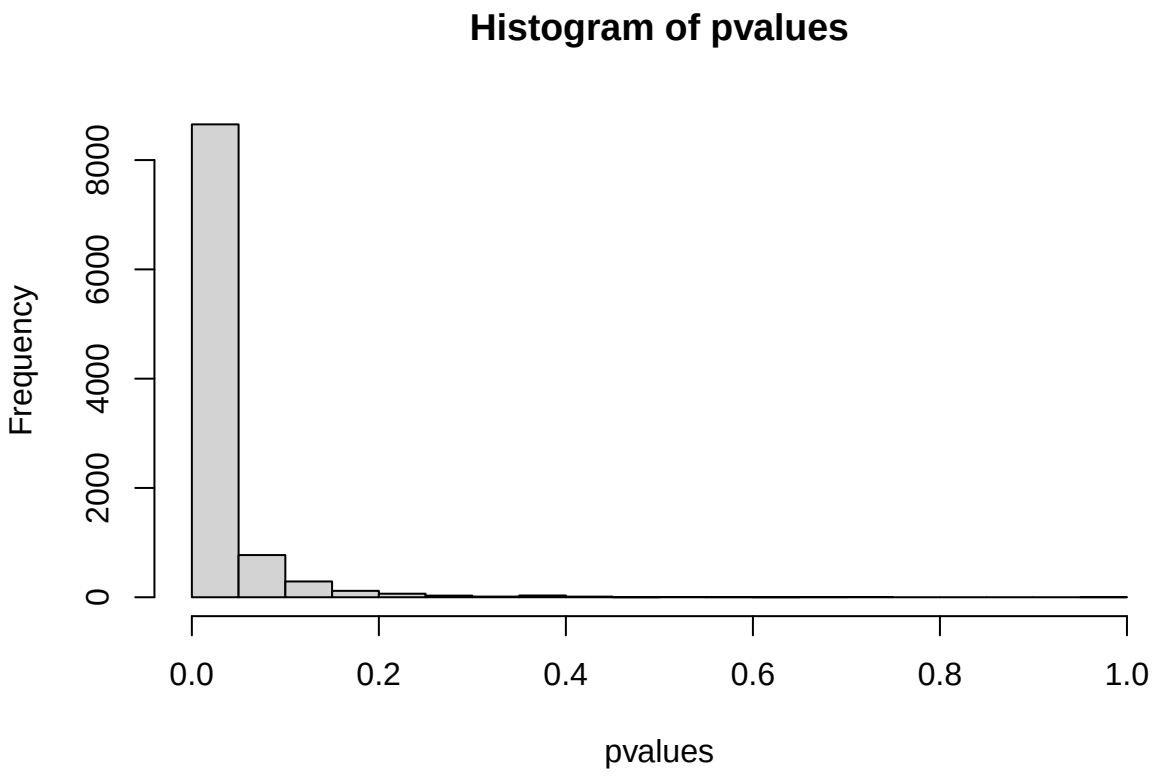

```
mean(pvalues <= 0.05)
```

```
## [1] 0.2042
```

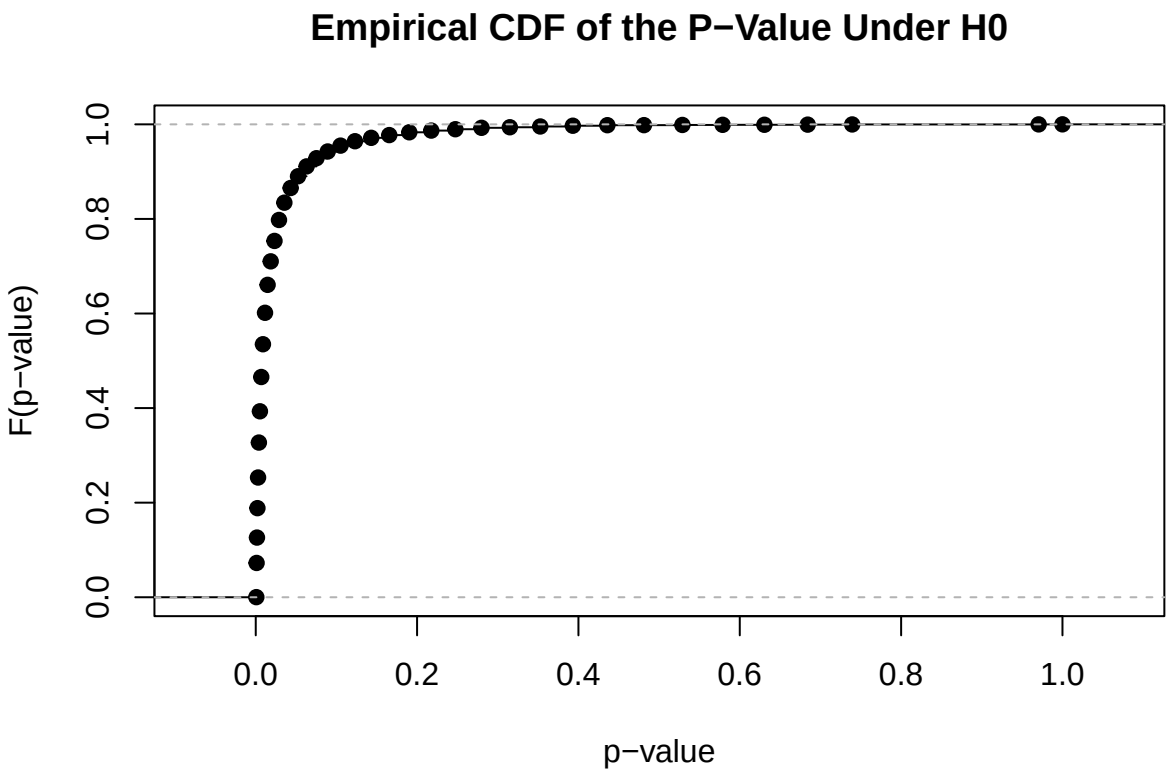
Mann-Whitney Section:

```
nG <- 10
d <- 0.8
for (i in 1:length(pvalues)){
  set.seed(i)  # notice that the seed changes every time inside the for loop
  # you could also set a single seed outside the for loop

  # Create the vectors x and y
  x <- c(rnorm(nG*0.9, mean = 0, sd = sqrt(0.19)),
         rnorm(nG*0.1, mean = 3, sd = sqrt(0.19))
)
  y <- c(rnorm(nG*0.9, mean = d, sd = sqrt(0.19)),
         rnorm(nG*0.1, mean = 3 + d, sd = sqrt(0.19))
)

  # Perform the t-test and get the p-value
  pvalues[i] <- wilcox.test(x, y, alternative = "t")$p.value

}
```

```
# histogram of the p-values under H0
hist(pvalues)
```




```
plot(ecdf(pvalues),
     xlab = "p-value",
     ylab = "F(p-value)",
     main = "Empirical CDF of the P-Value Under H0")
```




```
mean(pvalues <= 0.05)
```

```
## [1] 0.8654
```

| $n_G$ | t-Test Assumptions | Type I Error Rate | Power at $d = 0.2$ | Power at $d = 0.5$ | Power at $d = 0.8$ |
|---|---|---|---|---|---|
| 10 | Met | 0.0507 | 6.95% | 18.18% | 38.3% |
| 25 | Met | 0.05 | 10.32% | 40% | 79% |
| 50 | Met | 0.0504 | 15.92% | 68.12% | 98% |
| 10 | Violated | 0 | 0.02% | 16.34% | 20.4% |
| 25 | Violated | 0 | 0.06% | 89% | 93% |
| 50 | Violated | 0 | 1.23% | 86.55% | 100% |

Write a couple of paragraphs explaining the difference between parametric and nonparametric methods and explain why classic nonparametric methods (Mann-Whitney and Kruskal-Wallis) are a better choice than the corresponding parametric methods (two-sample t-test and one-way ANOVA) when the assumptions of the parametric method are clearly violated.

Answer: The primary reason why we would use a non-parametric model is when we violate our assumptions. It was clear that violated assumptions destroyed the power and validity of the tests that we were performing but when using non-parametric methods those violations were not of concern. This ties into the idea of not having to assume a family of functions that the data would follow but instead we tested based the shape of the data as it was.

Overall, the relationship between parametric and non-parametric is best understood when trying to understand the shape that a distribution follows. For example, the relationship between age and height in children is linear so fitting a linear model makes sense. However, if we were fitting the model of a molecules position in

space at a certain point in time a linear model may no longer follow the spread of the data and thus instead of doing a thirteenth degree Taylor polynomial that severely over fits the data a non-parametric solution allows us to answer prediction based problems.

Clearly violating assumption in a t-test also decreases power causing the likelihood of making a type two error to increase. Especially with small data sets the violation of assumptions is particular dangerous as the central limit theorem no longer protects a lot of the necessary conditions (as seen in the applied problem 2). Note that nonparametric methods are based solely on the data within the data set and thus the collection of good data is much more important as the thousands of simulations are based on it, it is also important in traditional parametric methods but catching it becomes increasingly important if the error is significant.

# Applied Problems

## Applied Problem 1 (Code: 6 pts; Explanation: 3 pts)

Textbook Exercise 2.4.8 with the following changes:

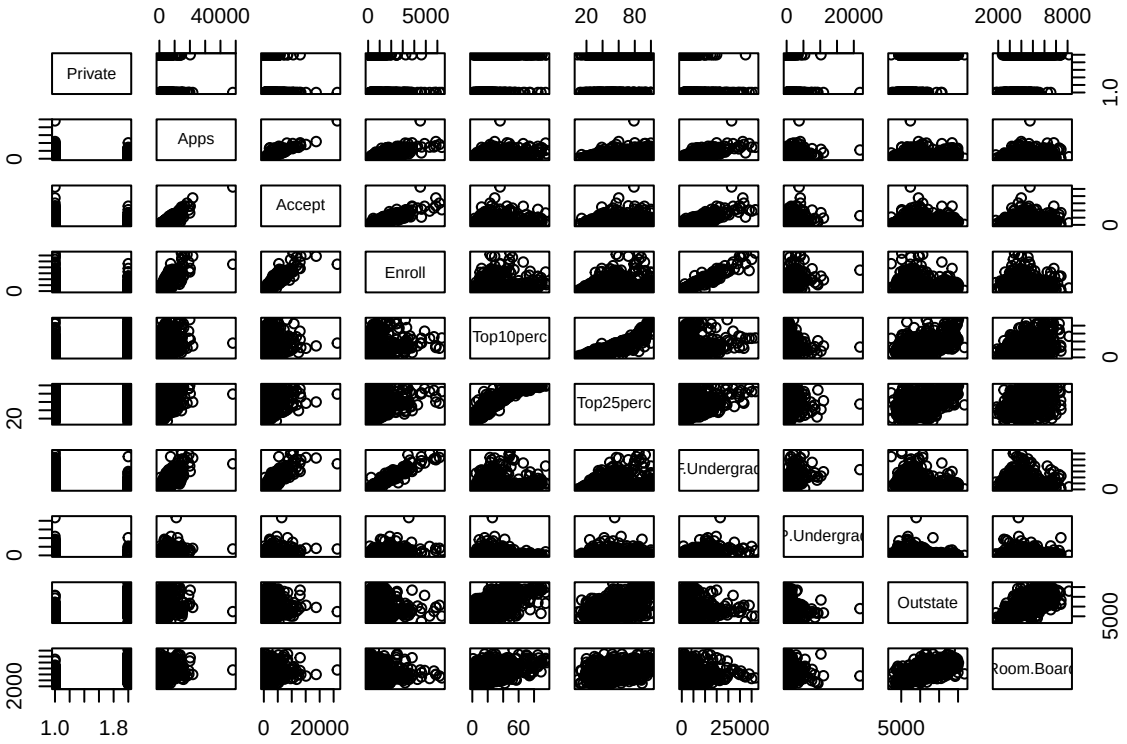
- Use the `College` dataset already in the `ISLR2` package instead of doing parts (a) and (b).
- Replace the four lines of code in part (c.iv) with a single line that accomplishes the same thing, using the `mutate` and either `if_else` or `case_when` functions from the `dplyr` package.
- As part of your brief summary in part (c.vi), identify at least one data point that cannot possibly have been recorded correctly, and explain why.

```
# row.names (College) <- College[, 1]
# View(College)
# Elite <- rep ("No", nrow (college))
# Elite[college$Top10perc > 50] <- "Yes "
# Elite <- as.factor (Elite)
# college <- data.frame (college , Elite)
summary(College)
```

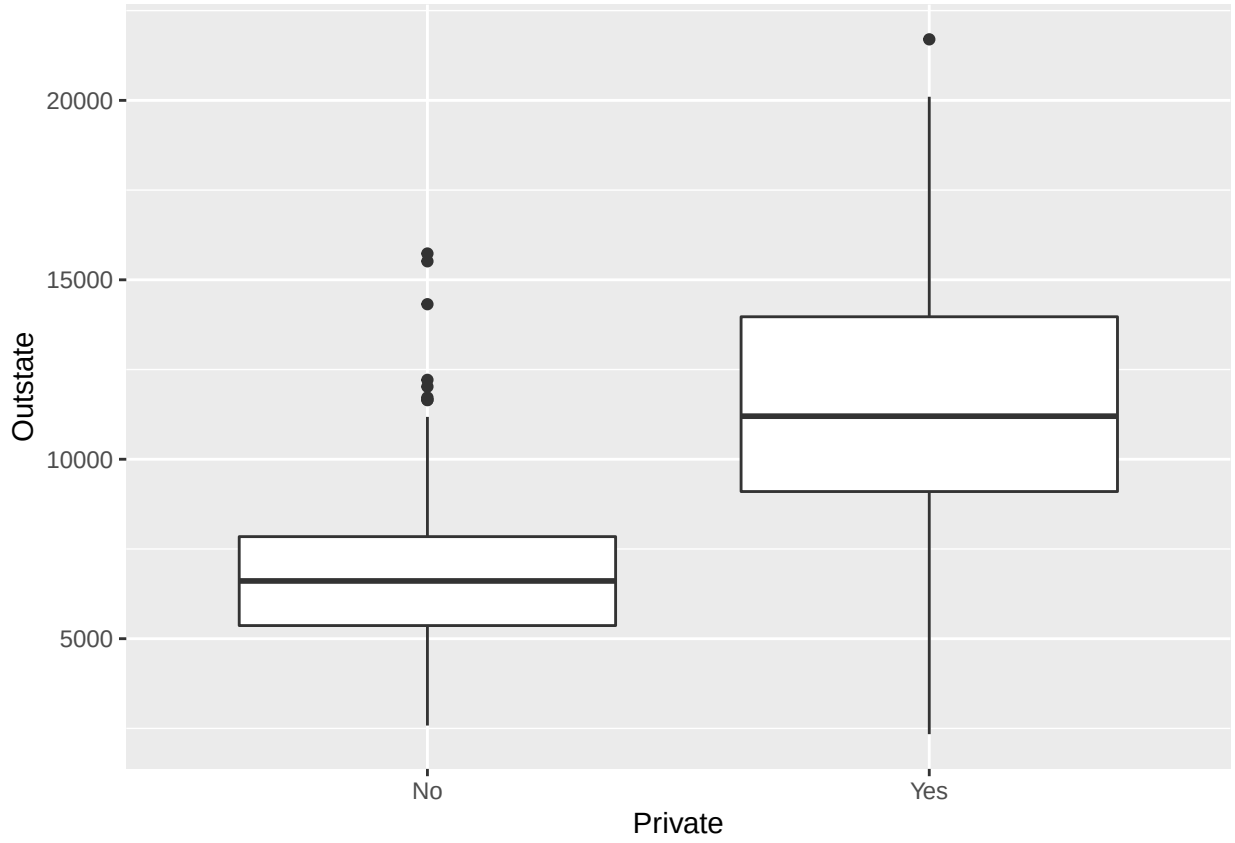
```
##  Private        Apps           Accept          Enroll       Top10perc
##  No :212   Min.   :   81   Min.   :   72   Min.   :  35   Min.   : 1.00
##  Yes:565   1st Qu.:  776   1st Qu.:  604   1st Qu.: 242   1st Qu.:15.00
##            Median : 1558   Median : 1110   Median : 434   Median :23.00
##            Mean   : 3002   Mean   : 2019   Mean   : 780   Mean   :27.56
##            3rd Qu.: 3624   3rd Qu.: 2424   3rd Qu.: 902   3rd Qu.:35.00
##            Max.   :48094   Max.   :26330   Max.   :6392   Max.   :96.00
##    Top25perc      F.Undergrad     P.Undergrad         Outstate
##  Min.   :  9.0   Min.   :  139   Min.   :    1.0   Min.   : 2340
##  1st Qu.: 41.0   1st Qu.:  992   1st Qu.:   95.0   1st Qu.: 7320
##  Median : 54.0   Median : 1707   Median :  353.0   Median : 9990
##  Mean   : 55.8   Mean   : 3700   Mean   :  855.3   Mean   :10441
##  3rd Qu.: 69.0   3rd Qu.: 4005   3rd Qu.:  967.0   3rd Qu.:12925
##  Max.   :100.0   Max.   :31643   Max.   :21836.0   Max.   :21700
##    Room.Board       Books           Personal         PhD
##  Min.   :1780   Min.   :  96.0   Min.   : 250   Min.   :  8.00
##  1st Qu.:3597   1st Qu.: 470.0   1st Qu.: 850   1st Qu.: 62.00
##  Median :4200   Median : 500.0   Median :1200   Median : 75.00
##  Mean   :4358   Mean   : 549.4   Mean   :1341   Mean   : 72.66
##  3rd Qu.:5050   3rd Qu.: 600.0   3rd Qu.:1700   3rd Qu.: 85.00
##  Max.   :8124   Max.   :2340.0   Max.   :6800   Max.   :103.00
##     Terminal       S.F.Ratio      perc.alumni        Expend
##  Min.   : 24.0   Min.   : 2.50   Min.   : 0.00   Min.   : 3186
##  1st Qu.: 71.0   1st Qu.:11.50   1st Qu.:13.00   1st Qu.: 6751
```

```
##  Median : 82.0   Median :13.60   Median :21.00   Median : 8377
##  Mean   : 79.7   Mean   :14.09   Mean   :22.74   Mean   : 9660
##  3rd Qu.: 92.0   3rd Qu.:16.50   3rd Qu.:31.00   3rd Qu.:10830
##  Max.   :100.0   Max.   :39.80   Max.   :64.00   Max.   :56233
##    Grad.Rate
##  Min.   : 10.00
##  1st Qu.: 53.00
##  Median : 65.00
##  Mean   : 65.46
##  3rd Qu.: 78.00
##  Max.   :118.00
```

```
pairs(College[,1:10])
```




```
ggplot(data = College,
       mapping = aes(x = Private,
                     y = Outstate)) +
  geom_boxplot()
```



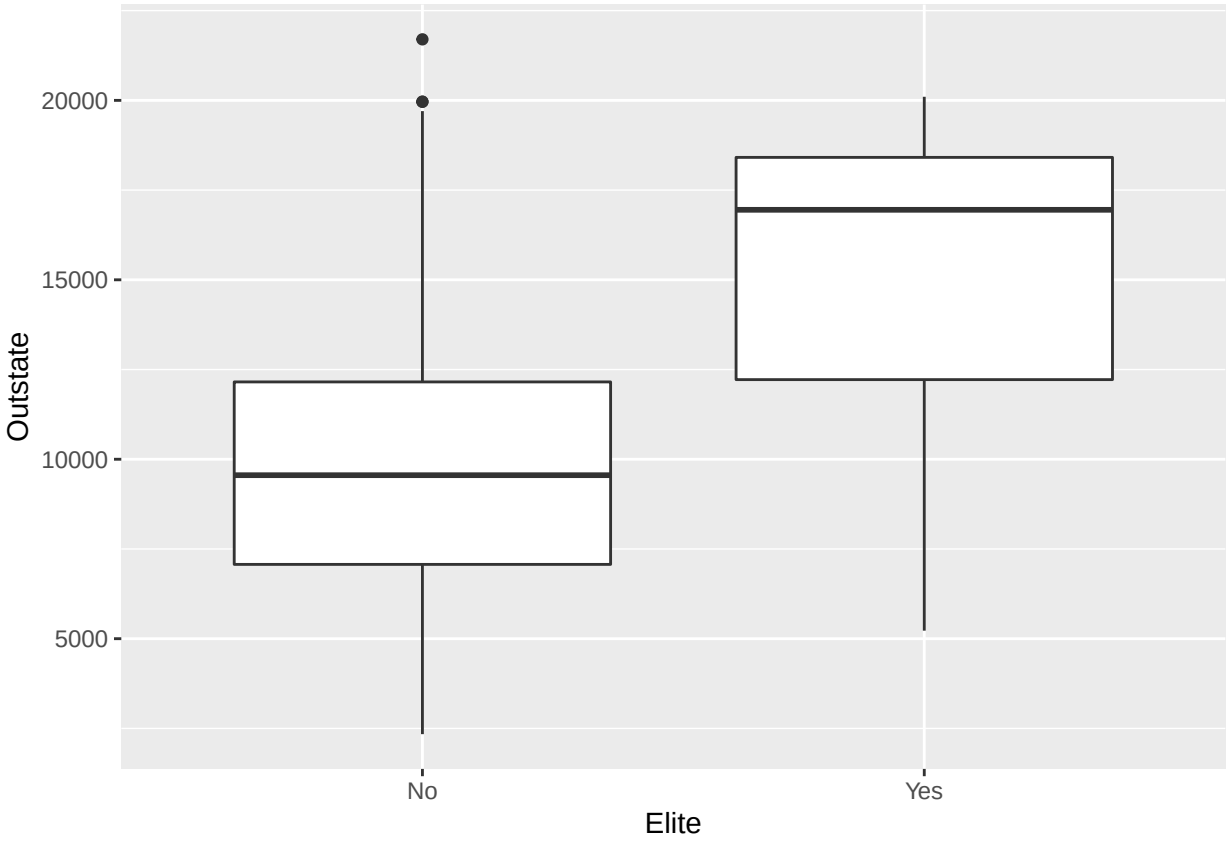

```
College %>%
  filter(PhD >100)
```

```
##                                  Private Apps Accept Enroll Top10perc
## Texas A&M University at Galveston      No  529    481    243        22
##                                  Top25perc F.Undergrad P.Undergrad Outstate
## Texas A&M University at Galveston        47        1206         134     4860
##                                  Room.Board Books Personal PhD Terminal
## Texas A&M University at Galveston       3122   600      650 103       88
##                                  S.F.Ratio perc.alumni Expend Grad.Rate
## Texas A&M University at Galveston      17.4          16   6415        43
```

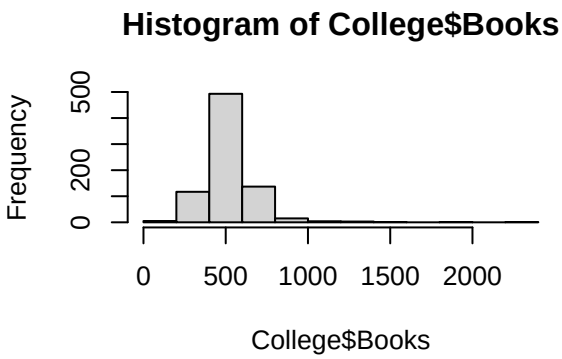
```
College <- College %>% mutate(Elite = case_when(Top10perc > 50 ~"Yes",Top10perc<= 50 ~ "No"))

summary(College$Elite)
```

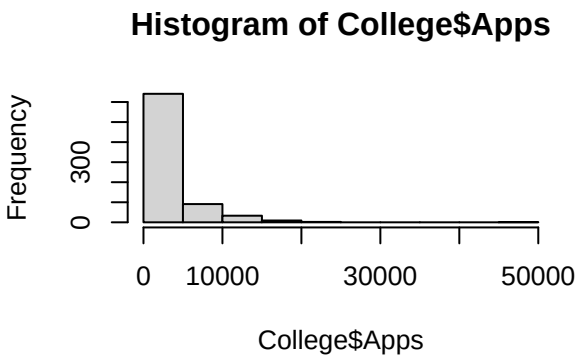
```
##    Length     Class      Mode
##       777 character character
```

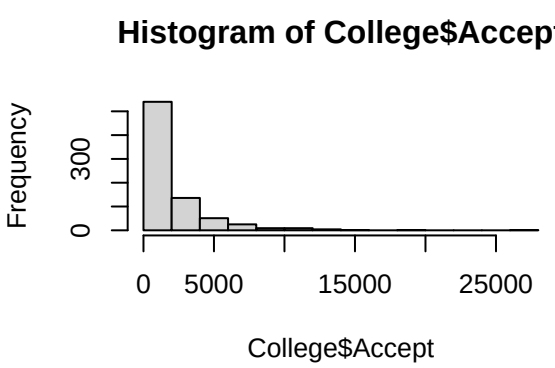
```
ggplot(data = College,
       mapping = aes(x = Elite,
                     y = Outstate)) +
  geom_boxplot()
```

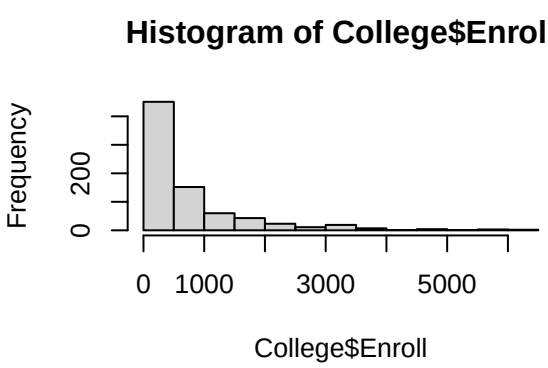



```
par(mfrow=c(2,2))
hist(College$Books)
hist(College$Apps)
hist(College$Accept)
hist(College$Enroll)
```



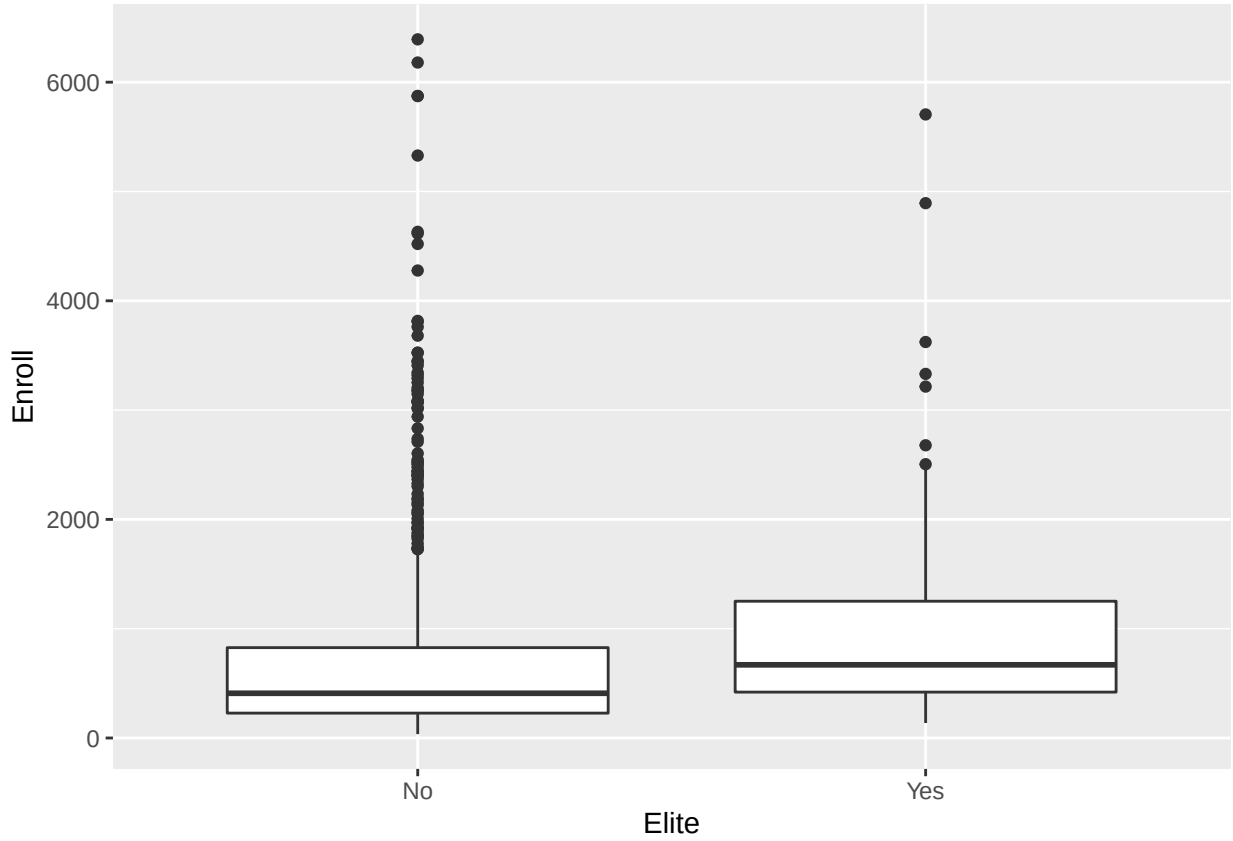




```
summary(College$Elite)
```

```
##    Length     Class      Mode
##       777 character character
```

```
ggplot(data = College,
       mapping = aes(x = Elite,
                     y = Enroll)) +
  geom_boxplot()
```




Texas A&M University at Galveston could not have been properly collected as the percentage of faculty with PH.D.'s is over 100% which is clearly an error in the collection of data processes.

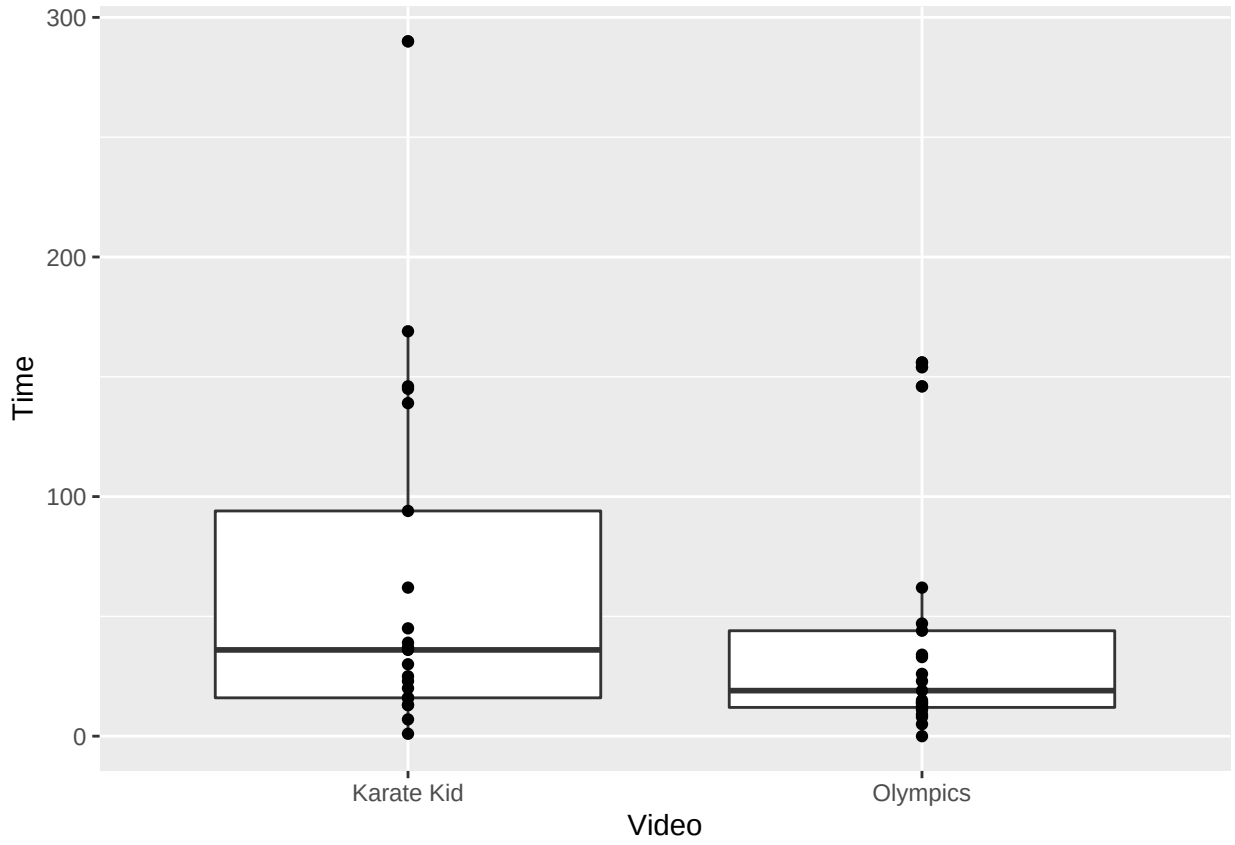
It is clear with just a little bit of data exploration that colleges tend to be in the same area regarding most numerical variables but seemingly skewed right. One reason this is occuring can be thanked to the idea that "elite" colleges cost far greater than the "average" ones resulting in this inbalance of distribution.

## Applied Problem 2 (Code: 1 pt; Explanation: 2 pts)

Molitor (1989) hypothesized that children who watched violent film and television were more tolerant of violent "real-life" behavior. A sample of 42 children were randomly assigned to watch footage from either the 1984 Summer Olympics (non-violent) or the movie *The Karate Kid* (violent). They were then told to watch (by video monitor) two younger children in the next room and get the research assistant if they "got into trouble" (the monitor actually showed a pre-recorded video of the children getting progressively more violent).

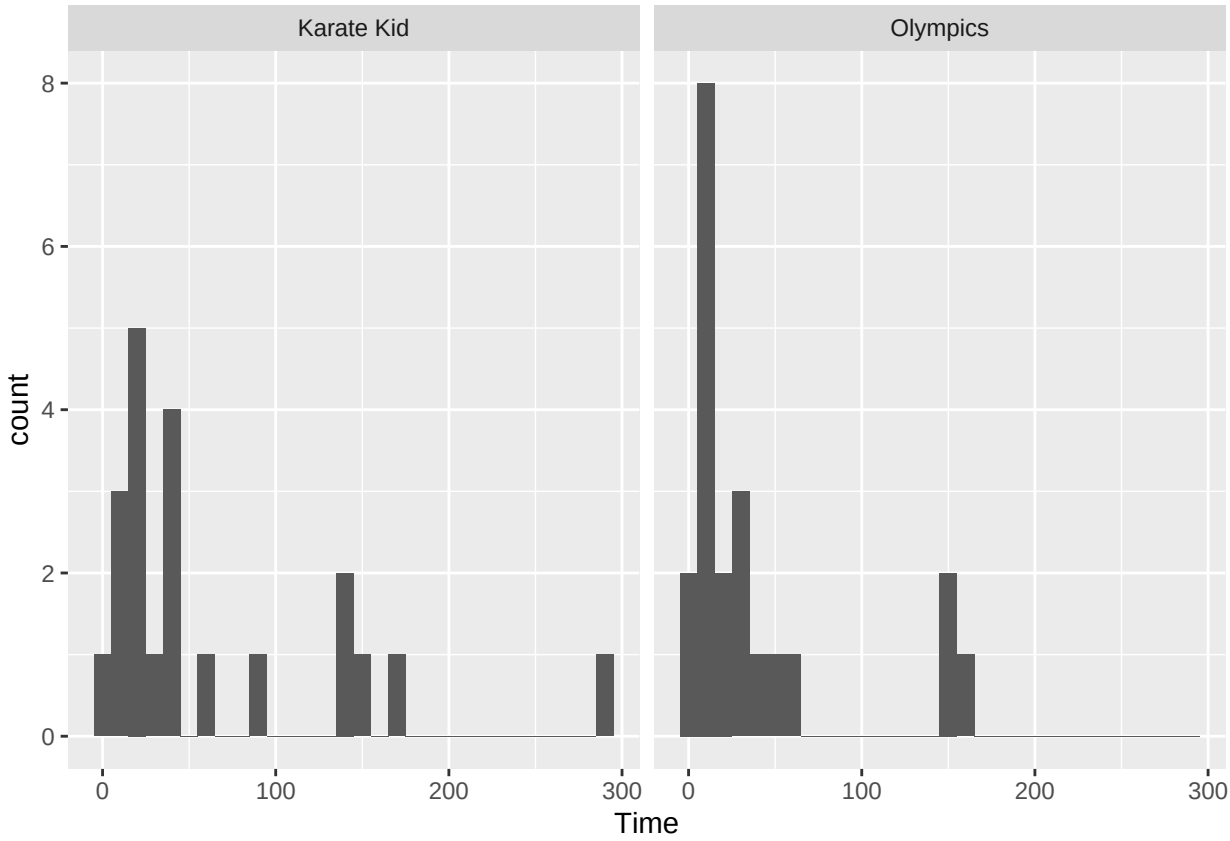
The file *violence.csv* contains the time (in seconds) that each child stayed in the room. Longer stays are assumed to indicate more tolerance of violent behavior. Produce an appropriate graph showing the sample data and, based on your graph, explain why a two-sample t-test might not be the best idea.

```
violent_data <- read.csv("violence.csv")

ggplot(data = violent_data ,aes(x = Video, y= Time)) +
  geom_boxplot()+
  geom_point()
```




```
ggplot(data = violent_data,
       mapping = aes(x = Time)) +
  geom_histogram() +
  facet_wrap(~Video)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
violent_data %>%
  filter(Video == "Olympics") %>%
  summary()
```

```
##     Video               Time
##  Length:21          Min.   :  0.00
##  Class :character   1st Qu.: 12.00
##  Mode  :character   Median : 19.00
##                     Mean   : 40.24
##                     3rd Qu.: 44.00
##                     Max.   :156.00
```

```
violent_data %>%
  filter(Video == "Karate Kid") %>%
  summary()
```

```
##     Video               Time
##  Length:21          Min.   :  1.00
##  Class :character   1st Qu.: 16.00
##  Mode  :character   Median : 36.00
##                     Mean   : 65.05
##                     3rd Qu.: 94.00
##                     Max.   :290.00
```

The reason why a t-test may not be okay for us is the fact that our sample size is very small. The issue is that with larger data sets we are covered by the central limit theorem that tells us even if the distribution is not normal the difference between y1 and y2 will be. However this is not the case with smaller data sets thus we have failed to meet the requirements necessary to perform the test.

## Applied Problem 3 (Code: 1 pt; Explanation: 2 pts)

Use the permutation test function you wrote in Lab 2 to determine whether the research hypothesis in the previous question was supported. Be sure to follow all steps of hypothesis testing, up to and including writing a conclusion that answers the research question in context.

```
permutation_t_test <- function(formula, data, alternative = "t",
                               permutations = 10000, seed = 9034){
 set.seed(seed)

 permutation_df <- model.frame(formula = formula, data = data)

 t_obs <- t.test(formula = formula, data = data,
                 alternative = alternative, var.equal = TRUE)$stat

 t_perm <- numeric(permutations)


 for (i in 1:length(t_perm)){

  permutation_df[[1]] <-sample(permutation_df[[1]])


  t_perm[i] <- t.test(formula, permutation_df, var.equal = TRUE)$statistic}

 T_all <- c(t_obs, t_perm)

  p_left <- sum(T_all <= t_obs)/(length(t_perm) + 1)

  p_right <- sum(T_all >= t_obs)/(length(t_perm) + 1)

  p_value <- dplyr::case_when(alternative == "g" ~ p_right,
                              alternative == "l" ~ p_left,
                              alternative == "t" ~ 2*min(p_left, p_right),
                              TRUE ~ NaN )

  results <- list(obs = t_obs,
                  sim = t_perm,
                  p_val = p_value)

  return(results)
}
```

Step 1: $H_0$: There is no difference in the violence tolerance of children who watched violent film or television and those who did not. $H_a$: There is a difference in the violence tolerance of children who watched violent film or television and those who did not.

Step 2: Our significance level will be $\alpha = 0.05$

Step 3: Testing

```
violent_output <- permutation_t_test(Time ~ Video, violent_data)
violent_output$p_val
```

```
## [1] 0.210179
```

Step 4: Conclusion Fail to reject $H_0$ at an $\alpha = 0.05$ level of significance. Their is insufficient evidence to

conclude that watching violent film or television affects a child's violence tolerance.