Homework Assignment #1

Math 437 - Modern Data Analysis

Due February 10, 2023

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 **35 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.

An "input variable" is also known as predictors, independent variables, or features and is usually denoted by the letter "X," sometimes with a subscript. An "output variable" is also known as the "response" or "dependent variable" and is usually denoted by the letter "Y."

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.

"Reducible error" can be improved by changing the accuracy using the most appropriate statistical learning technique to measure f. "Irreducible error" cannot be changed no matter how accurate the measure for f is, hence it being called 'irreducible.' For an example in which irreducible error is greater than zero, a basketball player's free-throw accuracy might vary, depending on their distance from the free-throw line or the pressure they are feeling from the game.

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?

Types of questions that can be answered with *inference* include finding the predictor values and examining the relationship between the predictor and response variables. Types of questions that can be responded using *predictions* are questions that are looking for actual values based on our data. It is possible to use the same model for both inference and prediction. We should know that some models are better for interpreting *inference* than *prediction* and vice versa.

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

Quantitative prediction questions, taking on numerical values, are generally answered using *regression* methods and qualitative prediction questions, utilizing categories, are answered using *classification* methods.

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

Nonparametric methods seek an estimate of f to get as close to the actual point as possible. This implies that these methods can be used on a wider range of shapes of f.

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.

Mean squared error is an example of a loss function for regression problems. *Training error rate* is an example of a loss function for classification problems.

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

We fit the model on a *training set* to build the model. The rest of the data is used to evaluate how good the *training set* is.

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?

As a model becomes more complex, the *bias* of the model will decrease and the *variance* of the model will increase The *bias* increases, because more simple models have higher chances of error, which is what *bias* is. Complex models will be more *variable* because smaller changes in the data lead to greater changes in f_hat.

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

Overfitting means that the data follows the errors closely. In terms of bias-variance, a model that *overfits* the data has higher variance and lower bias and vice versa.

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

A *Bayes classifier* is the conditional probability that is used to produce the lowest possible error rate called the *Bayes error rule*. The conditional probability for *Bayes classifier* is P(Y = j | X = x0), where j = 1, 2 depending on which response variable it is referring to and x0 is the predictor value. . . ..

(I think this answers the question, but I didn't want to delete your answer) A Bayes classifier assigns each observation to the most likely class, given its predictor values. The Bayes classifier will always choose the class for which the conditional probability is largest.

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).

Conceptual Problem 2 (3 pts)

Textbook Exercise 2.4.4

a Classification might be useful in everyday life when I am classifying between a man and a woman, whether the weather is going to be cold or not cold, and classifying . . . The response variables for comparing male and female would be hair, clothing, and mannerisms to name a few. Based on these variables, I believe the

goal of this application is prediction because we are using the response variable to classify whether or not the person is male or female. The response variables for classifying the weather is the temperature, if the sun is out and if there are clouds. Using these response variables, we predict whether it will be cold, rainy, warm, or hot. #####

b Regression can be used in everyday life when comparing sales for a company, predicting the stock price, and when we go shopping for clothes. Comparing sales for a company can be something as simple as using the sales from the previous week to predict the following week or more complex in using sales from the previous year to predict sales for this year. The same thing can be used when determining whether or not stock prices could go up or go down in value between today and tomorrow based on previous days. When you go shopping, you can predict how expensive something is based on response variables such as the company you are shopping from, the material and quality, as well as the size of the clothing item.

c Some real-life applications of cluster analysis include streaming services where they can collect data to see which areas have lower usage users and focus more on advertisements in that area, health insurance where companies could determine their monthly premiums based on the number of doctor visits a year, household size, and average age in the household along with other variables, and earthquake studies where researchers can cluster different areas based on whether or not they are on fault lines

Conceptual Problem 3 (3 pts)

Textbook Exercise 13.7.2

(a) Bernoulli with probability alpha

(b) Binomial with m trials and probability alpha

(c) sqrt(m*alpha*(1-alpha))

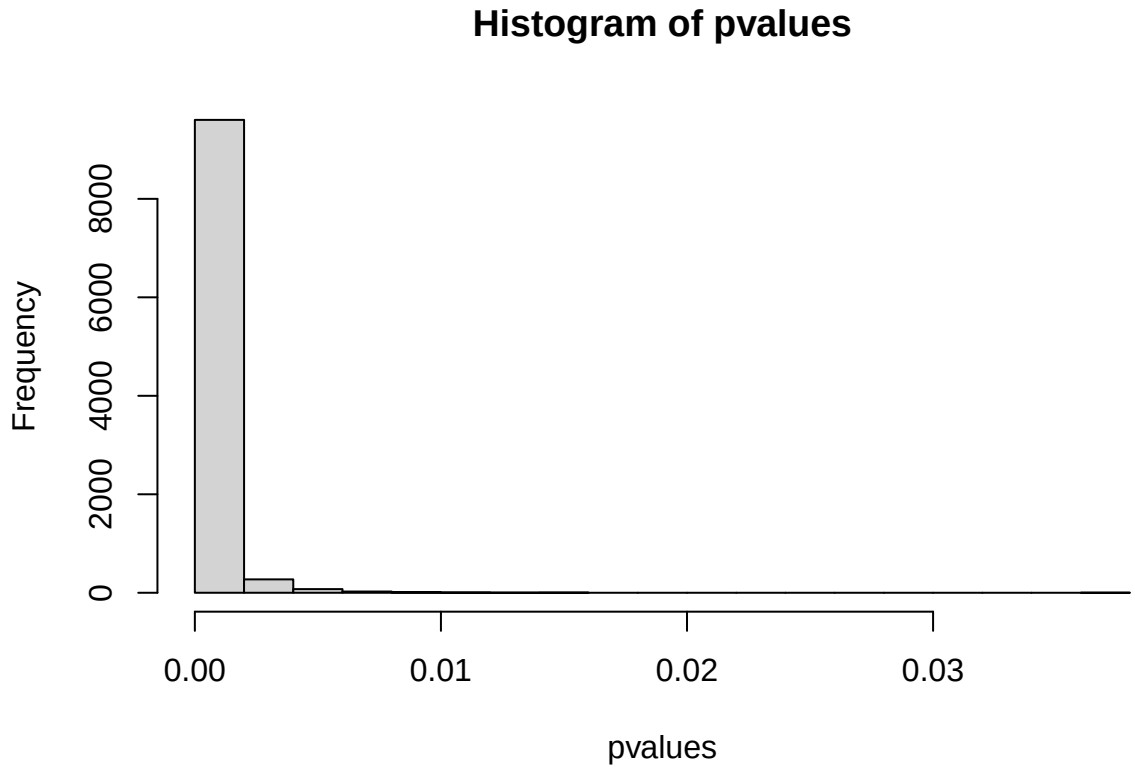
Simulation Problems

Simulation Problem 1 (Code: 1.5 pts; Explanation: 3.5 pts)

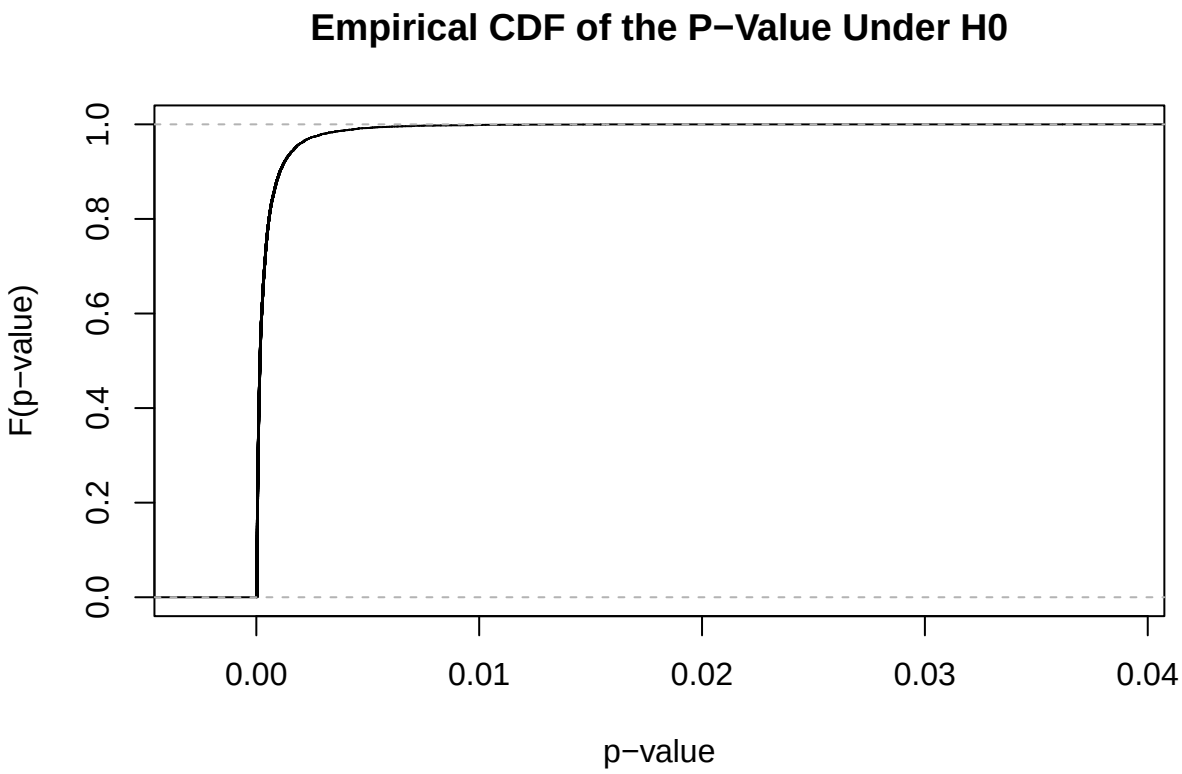
From the Parametric vs. Nonparametric Tests: Two-Sample Tests activity, copy to this homework your simulation code/results from the *Assumptions Violated, Ha True* section of each test as well as the results tables for all simulations (in the Class Results section). Write a couple of paragraphs explaining the difference between parametric and nonparametric methods and describing under what conditions we might prefer to use a classic nonparametric method (Mann-Whitney) Instead of the corresponding parametric method (two-sample t-test).

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

nG <- 50
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*.9, mean = 0, sd = sqrt(0.19)), rnorm(nG*.1, mean = 3, sd = sqrt(0.19)))
  y <- c(rnorm(nG*.9, mean = 0.8, sd = sqrt(0.19)), rnorm(nG*.1, mean = 3.8, sd = sqrt(0.19)))

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

}
```

```
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] 1
```

Parametric methods are utilized when a variable is assumed to be normally distributed and there are no outliers. Parametric methods are used with data that sufficiently fits a distribution. They can also estimate the value of a point when there are no data. Nonparametric methods seek an estimate of f to get as close to the actual point as possible. This implies that these methods can be used on a wider range of shapes of f. These methods are used when outliers in the data cannot be removed or when a distribution cannot be applied to the dataset. Based on the Parametric vs. Nonparametric activity we did in class, increasing power led to a lower probability of making a type 2 error.

Applied Problems

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

Textbook Exercise 2.4.8 with the following changes:

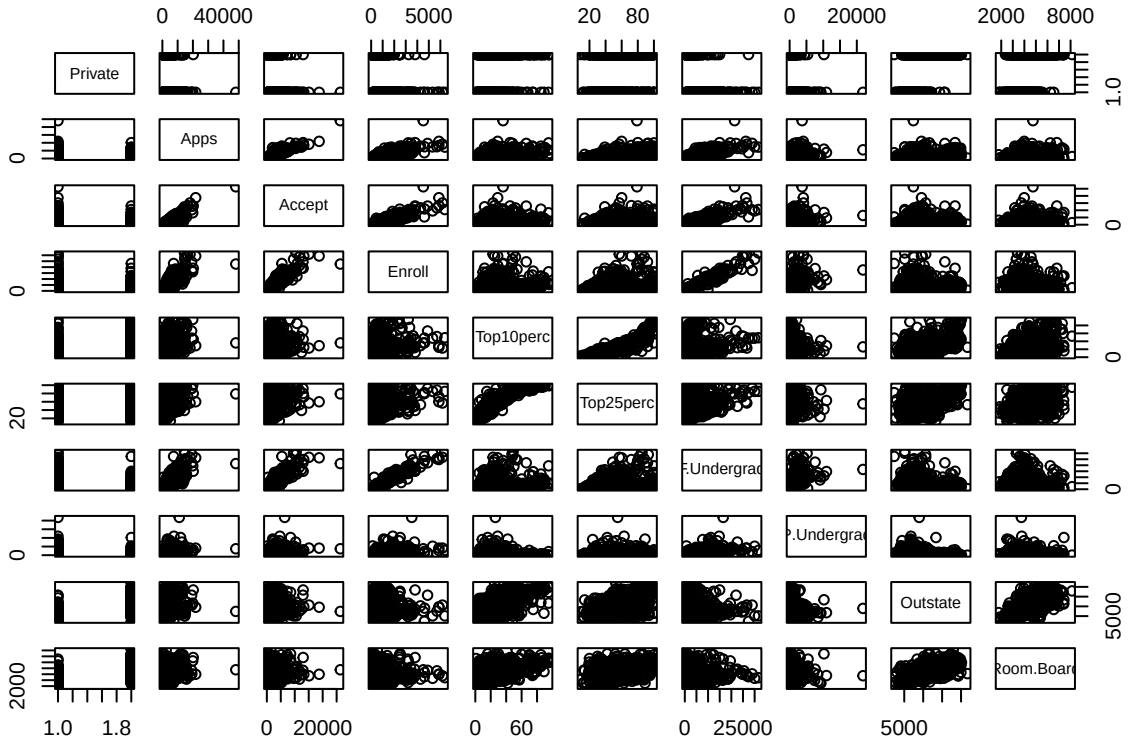
- Use the `College` dataset already in the `ISLR2` package instead of doing parts (a) and (b).
- Replace the four lines of code in part (c.iv) with a single line that accomplishes the same thing, using the `mutate` and either `if_else` or `case_when` functions from the `dplyr` package.
- As part of your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

```
coll <- ISLR2::College
```

```
# c.i
summary(coll)
```

```
##  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
```

```
# c.ii
A <- coll[ , 1:10]
pairs(A)
```

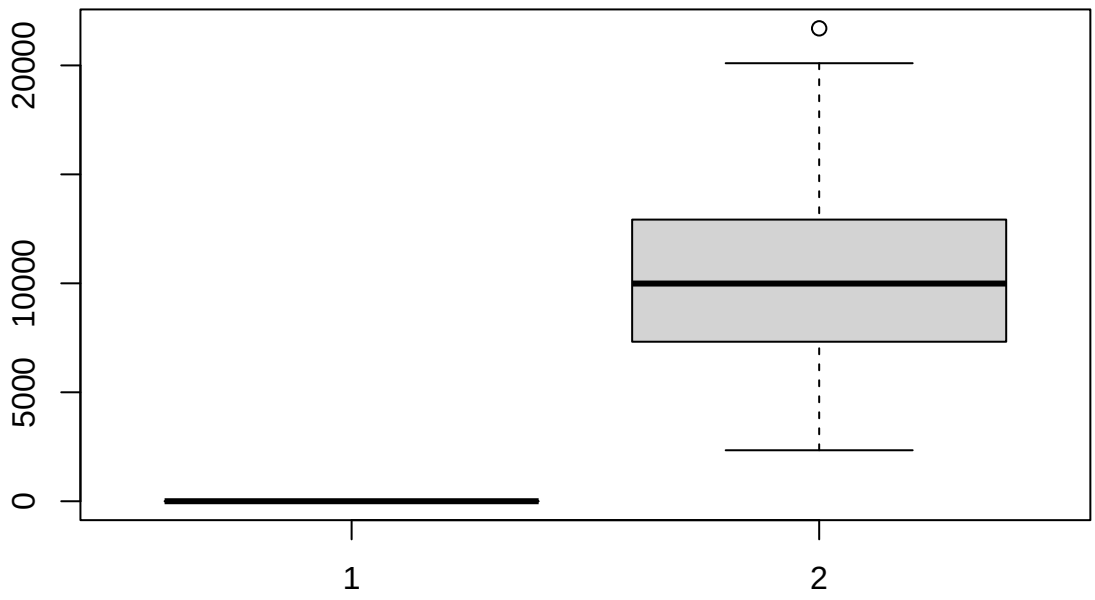



```
# c.iii
coll$Outstate
```

```
##   [1]  7440 12280 11250 12960  7560 13500 13290 13868 15595 10468 16548 17080
##  [13]  9690 12572  8352  8700 19760 10100  9996  5130 15476  6806 11208  7434
##  [25]  8644  3460 12000  6300 11902 13353 10990 11280  9925  8620 10995  9690
##  [37] 19264 17926 11290  6450 12850  8840  9000  7800 16304  4425  9550 21700
##  [49] 13800  8050  8740  8540  6200  5188 11660  6500  7844  7150  9900 18420
##  [61] 19030  7452 14080 10870 19380  9592  4371 10260 10265  2340 19528 18165
##  [73] 18550 13306 13130 10518  8900 12950  7380  7706 10230  7550  6060 10750
##  [85] 13050  8400 19292 17900 12200  8150 13125 15700  7656  9270 13712  9384
##  [97] 14340  7344 11400  8950 11230 10938  5962  4620  7242  8300 11850 16624
## [109] 13500 10335  8730  9300  7860  4412 17000 17500 10740 15960  8116  7168
## [121] 13925  9888 18930 19510 10860  6120  9800 11790 12600 11180 12247 12224
## [133] 10900  9990 11138  8300 11844 18000 11720 16240 17142  8412  8294 10425
## [145] 18624 10500  6900 10800  9216 18740 12050 15248 10628  8000  6230  8920
## [157]  9130 12292 19545 17295 10850  4528 16900 14300 18700  4486  6700  9570
## [169]  8310  9800  9000 13420 18432  8730 18590 15036  7248  5800  4950 11190
## [181]  5962  5710  9650  8770 15360 14190 14990 11800  9100  7800  8578 17600
## [193]  5401 10485 10955  6297 15000  6806  9400  5120 13900  6597  8025  6680
## [205]  8390 14235  6198  5840  9650 10390 13320  5500  9900 13440 10970  8180
## [217]  9476 12500 10800 17450  8100 18300  6489  6744  9150 19964  6120 13000
## [229] 12200  9420 15588  8958  9100  6108 11750  8330 10310 15688  5224 13404
## [241] 14125 11000 19700 13252 13218  7161  8200  6300  5504 17480 18485 17230
## [253]  9376  8800 11090 14067 19029 11600 13470 13960 12275  9990  8080  9950
## [265]  7260  7800 10500  8050 14550  7799 14360 10000  8840  6892  9766  9210
## [277] 10690  7550 14424  7994  7620  3946  6398 11700 18800  7656  9414 14850
```

```
## [289]  6995  8400  7870  8000 19240  9600 10910  8664 15747  8842 12600 18730
## [301]  6987 16880  9400  4752  5170  4938 17163 11040 13850 18700 10100 11700
## [313]  8840 15800 10560  5950  4818  9200 13380  4400  7352  7920 11200  5150
## [325]  5925  3957 12990 13592 11100 11500 13240 13900 12450  7320 15909  9620
## [337]  9858 10440 12370 14700  4300  9400 13850 10700 11610  5094 11200  6490
## [349] 11510 10200 11390 11200  9250 11040 20100  4486  7680  6930  7950 11985
## [361]  9813  6720 12500  5016 10300  8856 10658  8127  6840  7844  8200 11910
## [373] 11320 11505  5580  9866  4386  3840  8550 13000 12480  6073  5552  3648
## [385]  8438  4426 14990  7050 10520  4515 19300  6844  8950 10500  9900 12850
## [397]  7470 12474 12250  7400 16975  4738 13240  9090 10850  8832  5376 17748
## [409] 10194 10320  5542  6806  8400  8242 11718  5834 12580  4856 13380  6746
## [421]  7799  3735  9840  9900 16404 14134  9990  9114 19670 16560 12900 15990
## [433]  7629 16732  5390  6400  5336 12888  6530  8530 11000 13312 11925 14210
## [445]  6360 10645 18200  2580  8640 11690 10500  5640  6000 17688 10178  9700
## [457] 16200  4290 11859 19900 14400  9556 11020 10100 12030  6684  4449 13840
## [469] 13970 19960 12700 17475 15200 13250 15200  9870 13425  9490  8734 12520
## [481] 16425 10950  4356  7410  7411  7410 11070 10450 12950  4259  8670 10880
## [493] 12247 11200  9985 12750 12200 11690 12730 10800 10300 13030 14350  9408
## [505] 10850 10860 10575 10475  5130  8236  8384 13584 19300  8325  8955 17238
## [517] 12669 12825 12000 11240  7844  7210 10800  9240 16160 11250  8990 18710
## [529] 18820  3811  4680  3738  9520  5472 12772  7070  4740  4285  7536  7200
## [541] 11850  8400  7000  8600 10456 16150 10570 18720 11550 13332  6800  8678
## [553] 12140  5000  8650 13900 12315 16900  3040 12170  6550  6550  6550  6550
## [565]  6550  6550  6550  6550  6550  6550  6550  6550  6840  6550 16130 14500
## [577] 15150  7850  5666 10965  7070  5130  4860  8490  7850  7860  6400  7070
## [589] 11172  7600 10900  5391  9456 18810 11412 11010 12240 19040  7700  6735
## [601]  7800 18732  6874  4440  5028 11648 12024  6618  9500 18930  8907 11656
## [613] 10760 11380 10220 15192 11130 10430 11800  7090  5697 14220  4460  7560
## [625]  7230 11120  6994 13540  6540  6810  6600  6600  8594  8723  8566  6919
## [637] 16500 15732  8828  9843  8949  4916  9057  9057  7246  6150  4440  5595
## [649] 11450 11180  5972  8400  7248  8677  7558  5634  6634  4104  7731  6197
## [661] 16850  5173 10602 17020 10786 12040 16230 10330 14500 17840 13600 13226
## [673]  3687 11584  5800  8074  6760 17230  7100  4973  4652 11712  8550  5764
## [685]  4422  5130  4104 12520 16320 15350 11750  6857 15516 12212  8199  6172
## [697]  6704  7032  6950  6900  9096  8786  5988  8840 14900  9600  4286 11800
## [709] 17865 18920 15925 10217  5587 10260  7384 10900  9140  4450 12925 13500
## [721] 13850  8670 10000 11600 16260 13750 15276  8200 18350  2700  8840  5590
## [733]  9160 18345 14900  9850  9890 19130  7844  4470 14200  6390 14510  6940
## [745]  8994  5918  8124  5542 10720 12065  8820 14320 11480 18460 10500 16670
## [757] 16249 12660 12350 11150 14800 10060 10535 19629 11428  7820  4200  6400
## [769]  9100 15948 12680 15884  6797 11520  6900 19840  4990
```

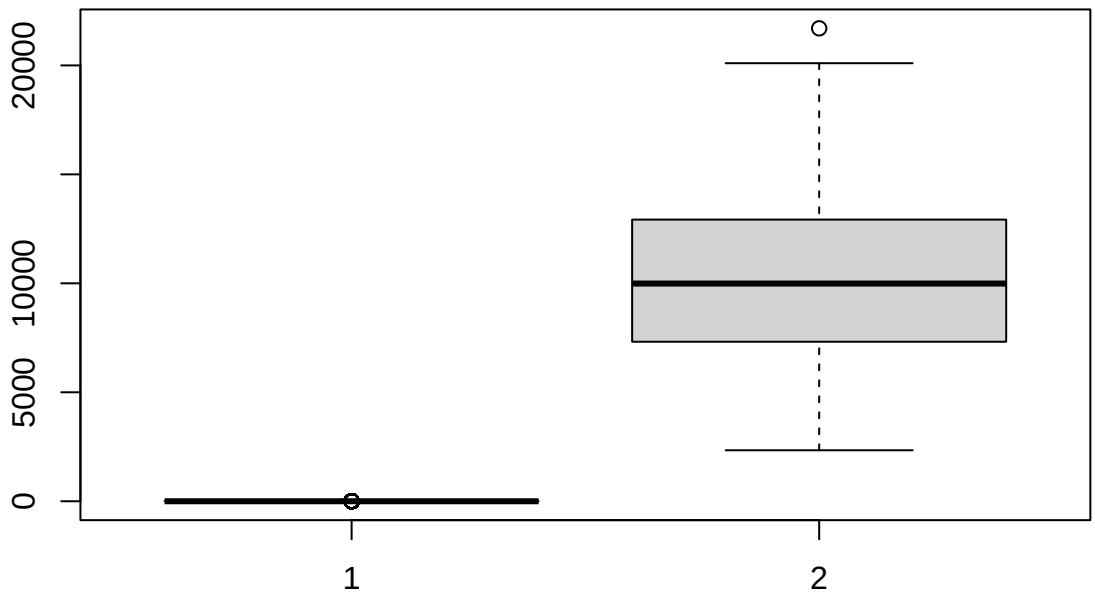
```
coll$Private <- as.factor(coll$Private)
boxplot(coll$Private, coll$Outstate)
```




```
#idk if the boxplot is correct
# c.iv
Elite <- rep ( " No " , nrow ( coll ) )
Elite [ coll $ Top10perc > 50] <- " Yes "
Elite <- as.factor ( Elite )
coll <- data.frame ( coll , Elite )
summary(Elite)
```

```
##   No   Yes
##   699    78
```

```
boxplot(coll$Elite,coll$Outstate)
```



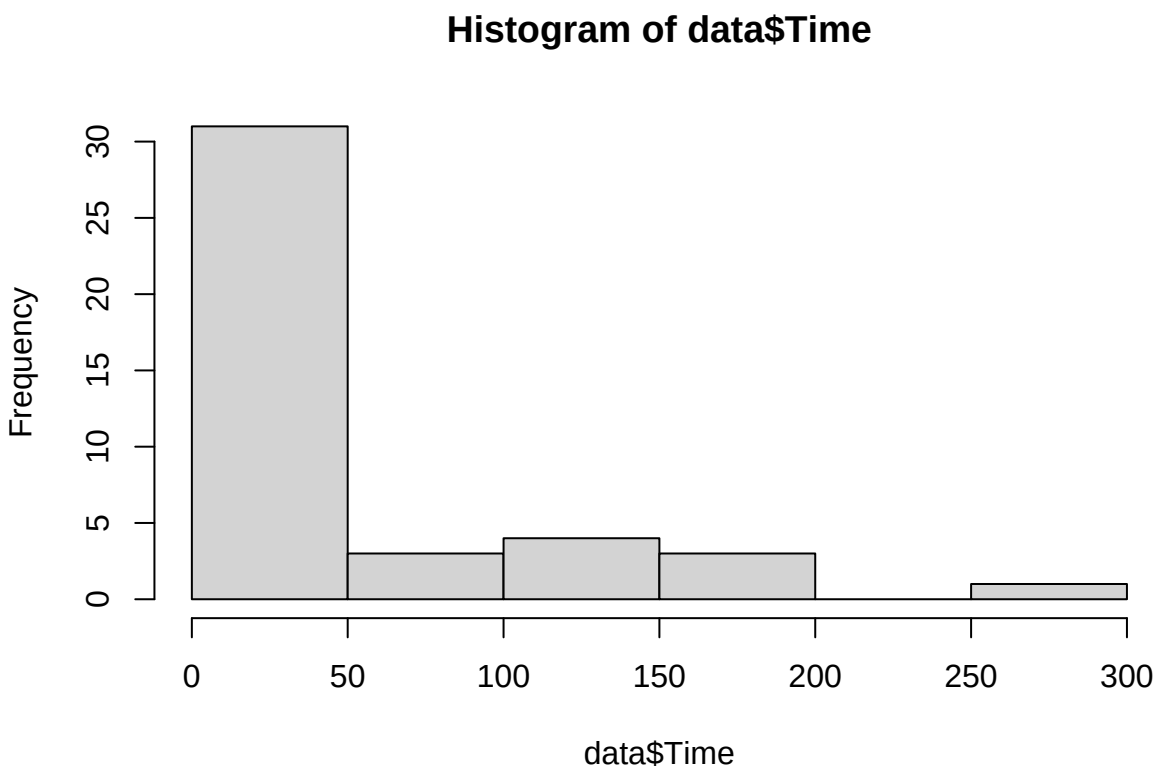

```
#I still don't think the boxplot is correct
# c.v


# c.vi
```

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.

```
data = read.csv("violence.csv", h = T)
View(data)
hist(data$Time)
```



In order to do a two sample t-test, the data must be normally distributed, but this dataset has an outlier and skewed right, as seen by the boxplot.

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.