# MATH560 HW 1

Jared Andreatta

2025-03-13

## Problem 1

a. In this problem, we should use a flexible method. With the presence of a large sample size of data, flexible methods are better able to capture the patterns in the data.

b. For this, it would be better to use an inflexible method. Flexible methods are prone to overfitting on small sample sizes where n«p.

c. Flexible methods are better for nonlinear data. These methods are better able to capture complex nonlinear patterns without imposing a strict structure on them.

d. Inflexible methods are better for data with large variance. The restrictions of the model would prevent overfitting on noise better than flexible models could.

## Problem 2

a. n = observations for top 500 firms, p = profit, # employees, CEO salary, industry

This is a regression problem, as the response variable is continuous. We are more concerned with inference since we want to know what drives these factors.

b. n = 20 observations of similar products, p = 1/0 success/failure, price, marketing budget, competitor price, 10 other vars

This is a classification problem, since we are trying to predict a binary, or qualitative, response. We are converned with prediction, as we want to predict whether our product would be a success.

c. n = 52 observations of weekly data in 2012, p = % change in USD/EURO, % change in US market, % change in British market, % change in German market.

This is a regression problem, since we are concerned with predicting a continuous value. We are concerned with prediction, since we want to predict the % change.

# Problem 5

The biggest advantage that flexible methods have over inflexible methods is that they are able to capture more complex patterns in data, thus they are often able to offer better predictive accuracy over inflexible methods, since inflexible methods impose restrictions on the structure of the model. The disadvantage of flexible methods is the lack of interpretability; these types of methods can be somewhat of a “black box”, since they are not as interpretable.

In studies where we are concerned with predictive accuracy with a large sample size, flexible methods may be preferred. In a study where we are concerned with inference with smaller sample sizes, parametric methods might be preferred.

# Problem 6

Parametric approaches are more strict: it uses the data to estimate parameters according to a specific structure of the model. Nonparametric approaches do not make any assumptions about the functional form of the function. Instead, they seek to estimate a smooth function that is the best fit of the given data points.

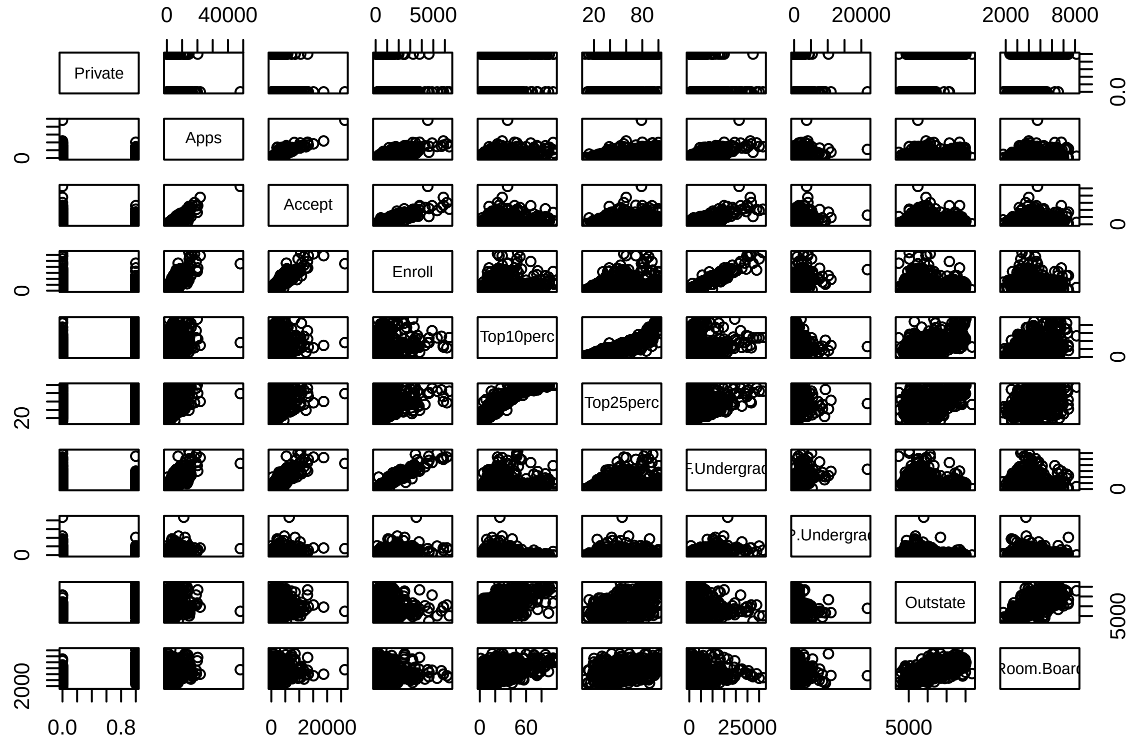
Parametric approaches have a few advantages over nonparametric methods. First, they perform better on small sample sizes. Nonparametric methods are prone to overfitting on the noise of the data when the sample size is small and the estimate of the variance is poor, whereas parametric methods can generalize better off of a small sample size. The second advantage is for the sake of inference. The black-box nature of nonparametric methods can be quite difficult to inference off of. However, since parametric models estimate a set of parameters for a model, it is easy for the analyst to statistically inference which variables have effect on other variables.

# Problem 8

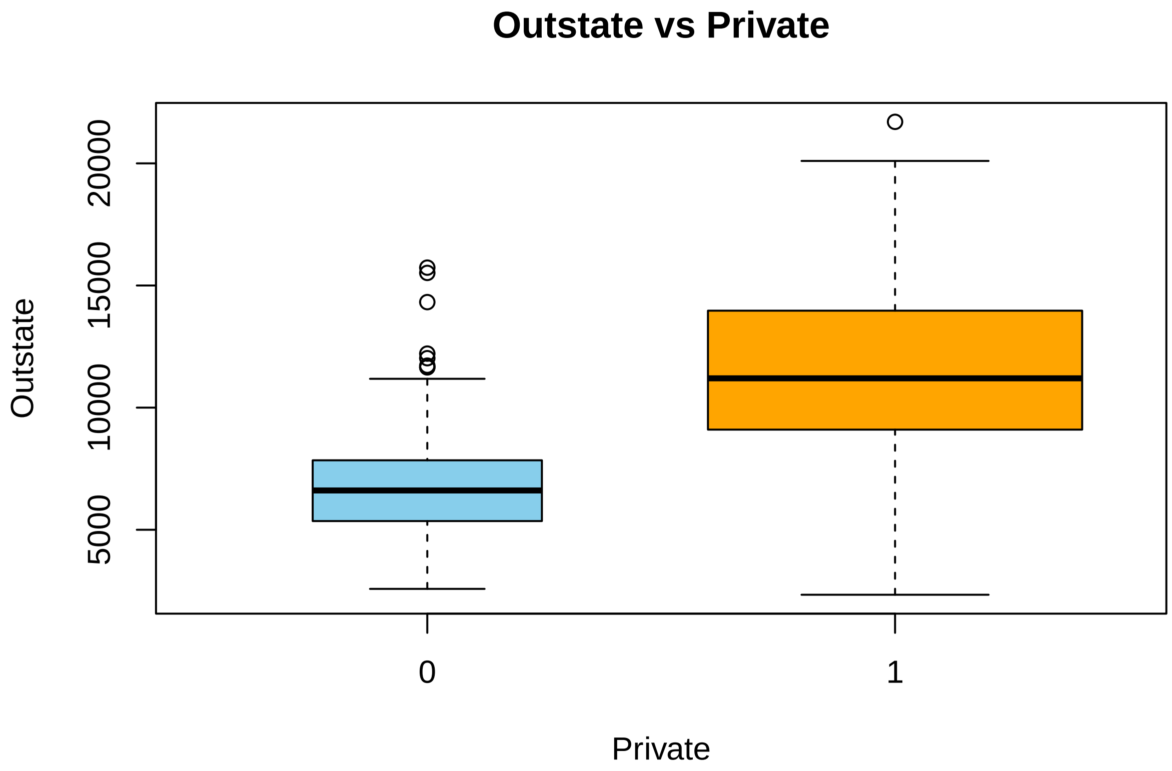
```r
# a
college <- read.csv("College.csv")

# b
rownames(college) <- college[, 1]
college <- college[, -1]
View(college)

# c
attach(college)

# i
summary(college)
```

```
##    Private               Apps          Accept          Enroll
##  Length:777         Min.   :   81   Min.   :   72   Min.   :  35
##  Class :character   1st Qu.:  776   1st Qu.:  604   1st Qu.: 242
##  Mode  :character   Median : 1558   Median : 1110   Median : 434
##                     Mean   : 3002   Mean   : 2019   Mean   : 780
##                     3rd Qu.: 3624   3rd Qu.: 2424   3rd Qu.: 902
##                     Max.   :48094   Max.   :26330   Max.   :6392
##    Top10perc       Top25perc      F.Undergrad     P.Undergrad
##  Min.   : 1.00   Min.   :  9.0   Min.   :  139   Min.   :    1.0
##  1st Qu.:15.00   1st Qu.: 41.0   1st Qu.:  992   1st Qu.:   95.0
##  Median :23.00   Median : 54.0   Median : 1707   Median :  353.0
##  Mean   :27.56   Mean   : 55.8   Mean   : 3700   Mean   :  855.3
##  3rd Qu.:35.00   3rd Qu.: 69.0   3rd Qu.: 4005   3rd Qu.:  967.0
##  Max.   :96.00   Max.   :100.0   Max.   :31643   Max.   :21836.0
##     Outstate       Room.Board        Books           Personal
##  Min.   : 2340   Min.   :1780   Min.   :  96.0   Min.   : 250
##  1st Qu.: 7320   1st Qu.:3597   1st Qu.: 470.0   1st Qu.: 850
##  Median : 9990   Median :4200   Median : 500.0   Median :1200
##  Mean   :10441   Mean   :4358   Mean   : 549.4   Mean   :1341
##  3rd Qu.:12925   3rd Qu.:5050   3rd Qu.: 600.0   3rd Qu.:1700
##  Max.   :21700   Max.   :8124   Max.   :2340.0   Max.   :6800
##       PhD            Terminal       S.F.Ratio      perc.alumni
##  Min.   :  8.00   Min.   : 24.0   Min.   : 2.50   Min.   : 0.00
##  1st Qu.: 62.00   1st Qu.: 71.0   1st Qu.:11.50   1st Qu.:13.00
##  Median : 75.00   Median : 82.0   Median :13.60   Median :21.00
##  Mean   : 72.66   Mean   : 79.7   Mean   :14.09   Mean   :22.74
##  3rd Qu.: 85.00   3rd Qu.: 92.0   3rd Qu.:16.50   3rd Qu.:31.00
##  Max.   :103.00   Max.   :100.0   Max.   :39.80   Max.   :64.00
##      Expend        Grad.Rate
##  Min.   : 3186   Min.   : 10.00
##  1st Qu.: 6751   1st Qu.: 53.00
##  Median : 8377   Median : 65.00
##  Mean   : 9660   Mean   : 65.46
##  3rd Qu.:10830   3rd Qu.: 78.00
##  Max.   :56233   Max.   :118.00
```

```r
# ii
college$Private <- as.numeric(college$Private=="Yes")
pairs(college[,1:10])
```



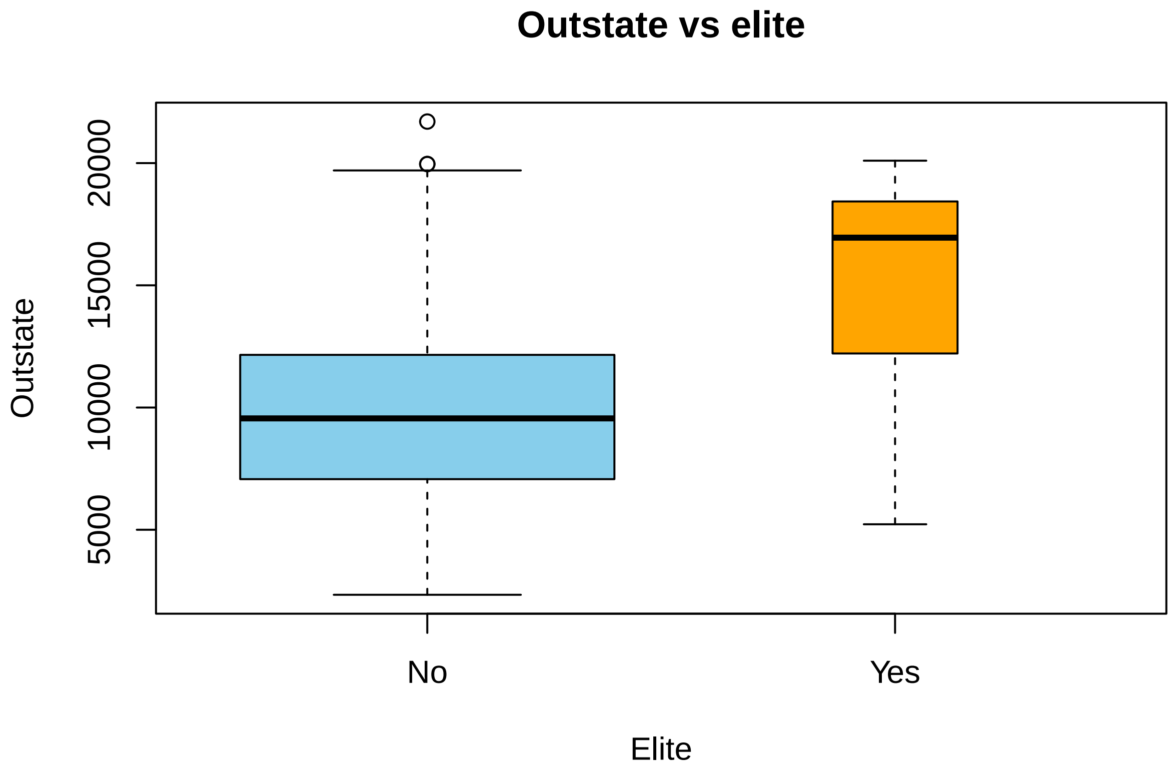
```r
# iii
boxplot(Outstate~Private, data=college,
        varwidth=TRUE,
        col=c("skyblue","orange"),
        main="Outstate vs Private")
```



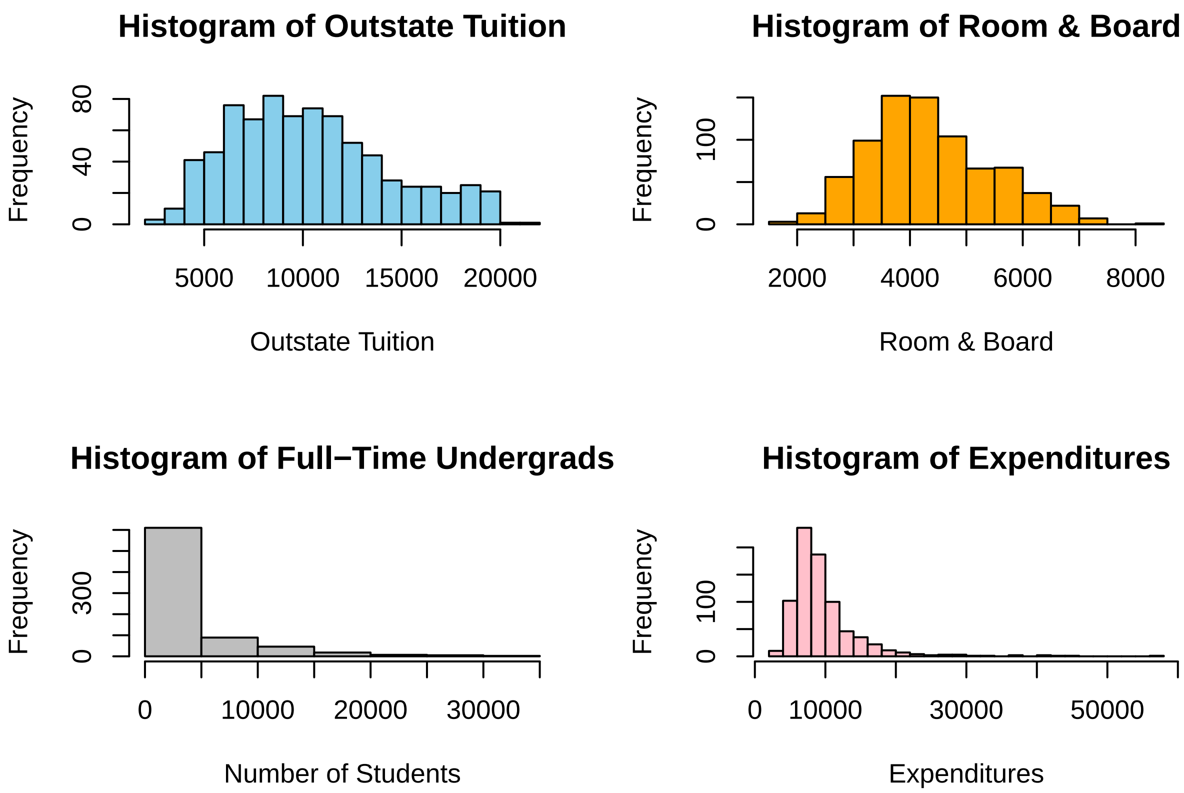
```
# iv

Elite <- rep("No", nrow(college))
Elite[college$Top10perc > 50] <- "Yes"
Elite <- as.factor(Elite)
college <- data.frame(college, Elite)

summary(college) # 78 Universities
```

```
##     Private            Apps           Accept          Enroll
##  Min.   :0.0000   Min.   :   81   Min.   :   72   Min.   :  35
##  1st Qu.:0.0000   1st Qu.:  776   1st Qu.:  604   1st Qu.: 242
##  Median :1.0000   Median : 1558   Median : 1110   Median : 434
##  Mean   :0.7272   Mean   : 3002   Mean   : 2019   Mean   : 780
##  3rd Qu.:1.0000   3rd Qu.: 3624   3rd Qu.: 2424   3rd Qu.: 902
##  Max.   :1.0000   Max.   :48094   Max.   :26330   Max.   :6392
##    Top10perc       Top25perc      F.Undergrad     P.Undergrad
##  Min.   : 1.00   Min.   :  9.0   Min.   :  139   Min.   :    1.0
##  1st Qu.:15.00   1st Qu.: 41.0   1st Qu.:  992   1st Qu.:   95.0
##  Median :23.00   Median : 54.0   Median : 1707   Median :  353.0
##  Mean   :27.56   Mean   : 55.8   Mean   : 3700   Mean   :  855.3
##  3rd Qu.:35.00   3rd Qu.: 69.0   3rd Qu.: 4005   3rd Qu.:  967.0
##  Max.   :96.00   Max.   :100.0   Max.   :31643   Max.   :21836.0
##     Outstate       Room.Board       Books           Personal
##  Min.   : 2340   Min.   :1780   Min.   :  96.0   Min.   : 250
##  1st Qu.: 7320   1st Qu.:3597   1st Qu.: 470.0   1st Qu.: 850
##  Median : 9990   Median :4200   Median : 500.0   Median :1200
##  Mean   :10441   Mean   :4358   Mean   : 549.4   Mean   :1341
##  3rd Qu.:12925   3rd Qu.:5050   3rd Qu.: 600.0   3rd Qu.:1700
##  Max.   :21700   Max.   :8124   Max.   :2340.0   Max.   :6800
##       PhD             Terminal       S.F.Ratio      perc.alumni
##  Min.   :  8.00   Min.   : 24.0   Min.   : 2.50   Min.   : 0.00
##  1st Qu.: 62.00   1st Qu.: 71.0   1st Qu.:11.50   1st Qu.:13.00
##  Median : 75.00   Median : 82.0   Median :13.60   Median :21.00
##  Mean   : 72.66   Mean   : 79.7   Mean   :14.09   Mean   :22.74
##  3rd Qu.: 85.00   3rd Qu.: 92.0   3rd Qu.:16.50   3rd Qu.:31.00
##  Max.   :103.00   Max.   :100.0   Max.   :39.80   Max.   :64.00
##      Expend          Grad.Rate      Elite
##  Min.   : 3186   Min.   : 10.00   No :699
##  1st Qu.: 6751   1st Qu.: 53.00   Yes: 78
##  Median : 8377   Median : 65.00
##  Mean   : 9660   Mean   : 65.46
##  3rd Qu.:10830   3rd Qu.: 78.00
##  Max.   :56233   Max.   :118.00
```

```
boxplot(Outstate~Elite, data=college,
        varwidth=TRUE,
        col=c("skyblue","orange"),
        main="Outstate vs elite")
```



```r
# v
# Divide the plotting window into 2 rows and 2 columns
par(mfrow = c(2, 2))

# Outstate tuition
hist(college$Outstate, breaks = 15,
     col = "skyblue",
     main = "Histogram of Outstate Tuition",
     xlab = "Outstate Tuition")

# R&B costs
hist(college$Room.Board, breaks = 20,
     col = "orange",
     main = "Histogram of Room & Board",
     xlab = "Room & Board")

# FT Undergrads
hist(college$F.Undergrad, breaks = 10,
     col = "gray",
     main = "Histogram of Full-Time Undergrads",
     xlab = "Number of Students")

# Expenditure
hist(college$Expend, breaks = 25,
     col = "pink",
     main = "Histogram of Expenditures",
     xlab = "Expenditures")
```



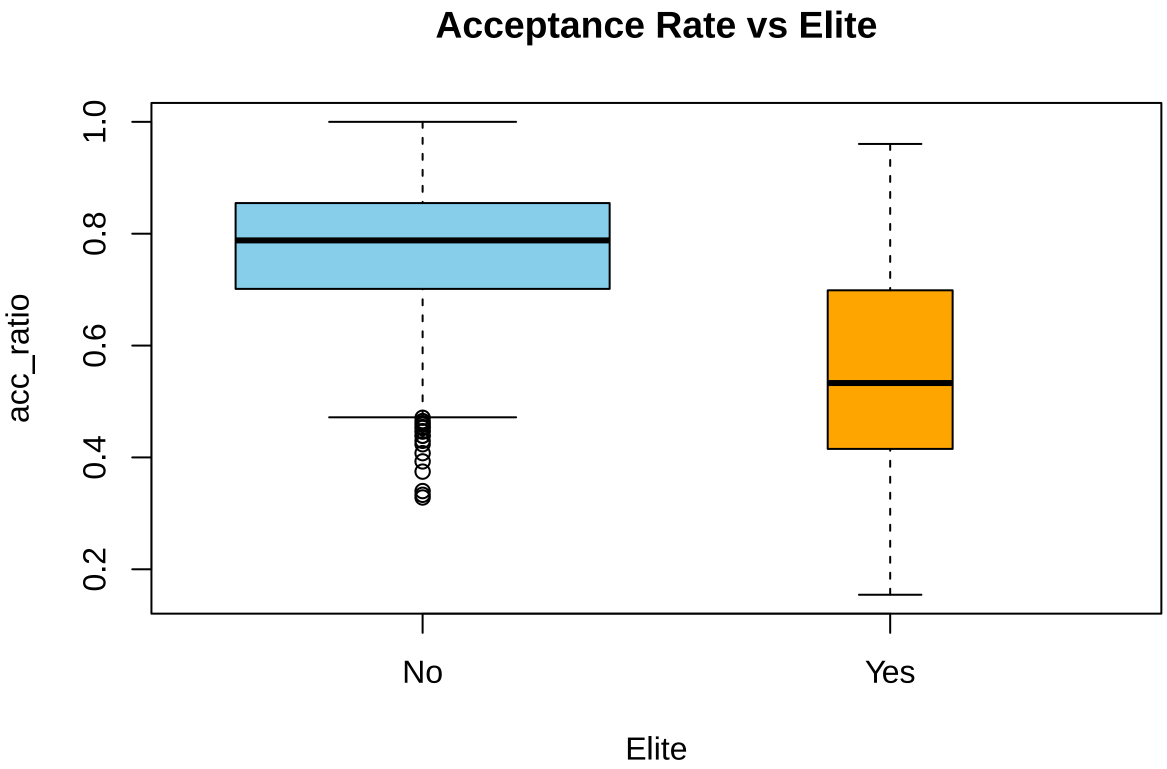
## Problem 8 part vi

Here, I will do some extra data analysis and visualization.

**Acceptance rate by prestige.**

I made another variable to quantify acceptance rate for each college by dividing number of applicants accepted divided by number of applicants. As we can see, "elite" schools clearly have a lower acceptance rate with a mean of ~50%. It also has much more variability: it can get lower than 20% acceptance rate, but it can also have higher than an 80% acceptance rate, whereas the middle 50% of the non-elite schools seem to be closely centered around the mean, although there is a greater presence of outliers that have low acceptance rates.

```r
acc_ratio <- Accept/Apps # Acceptance rate

boxplot(acc_ratio~Elite, data=college,
        varwidth=TRUE,
        col=c("skyblue","orange"),
        main="Acceptance Rate vs Elite")
```



### Distribution of costs

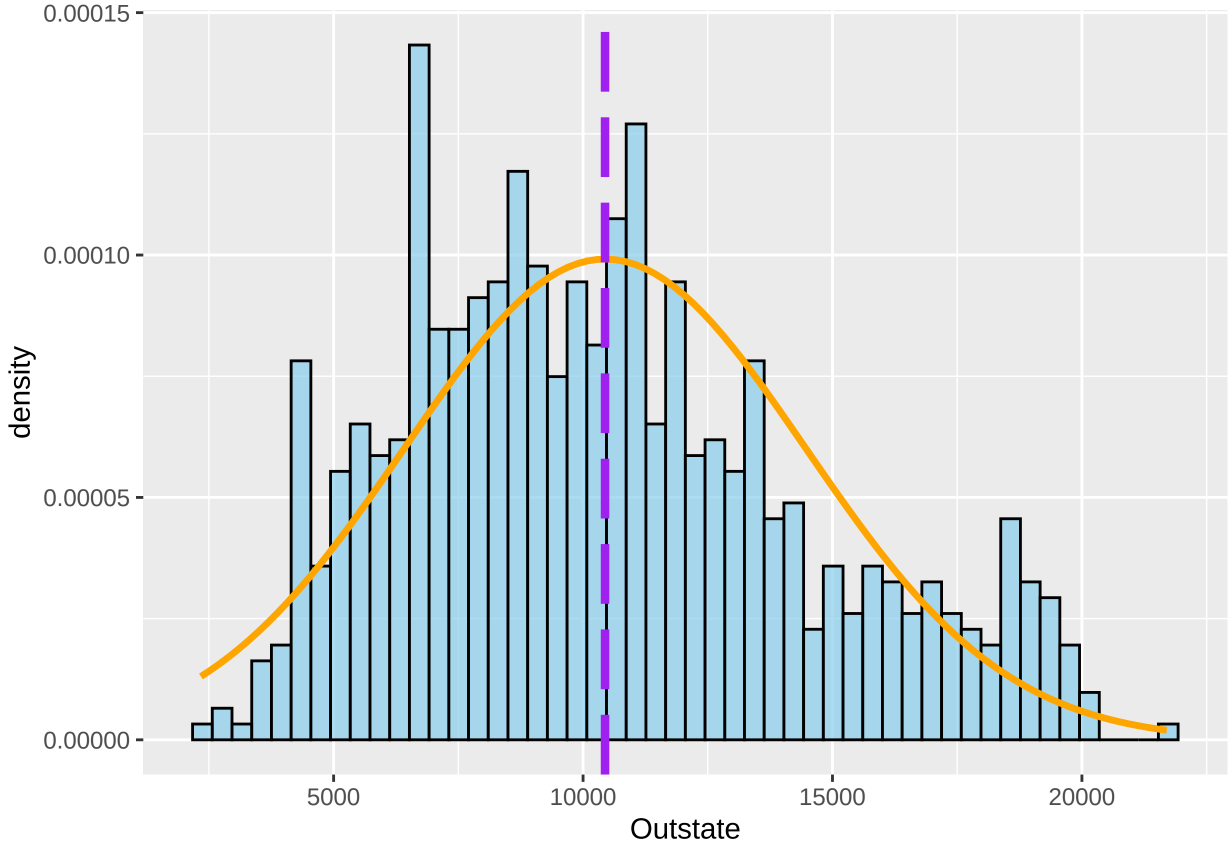
First, I plotted the distribution of cost-related variables against the assumed normal distribution. At first sight, there seems to be a degree of negative skew for each variable, with “fat tails” on the right. I also plotted for the total costs, which is the cost of everything aggregated together, which followed a similar distribution shape as the other variables. This suggests that most of the colleges have lower costs, but there are a handful of expensive colleges that skew the mean significantly in terms of costs.

```
library(ggplot2)
library(ggExtra)
```

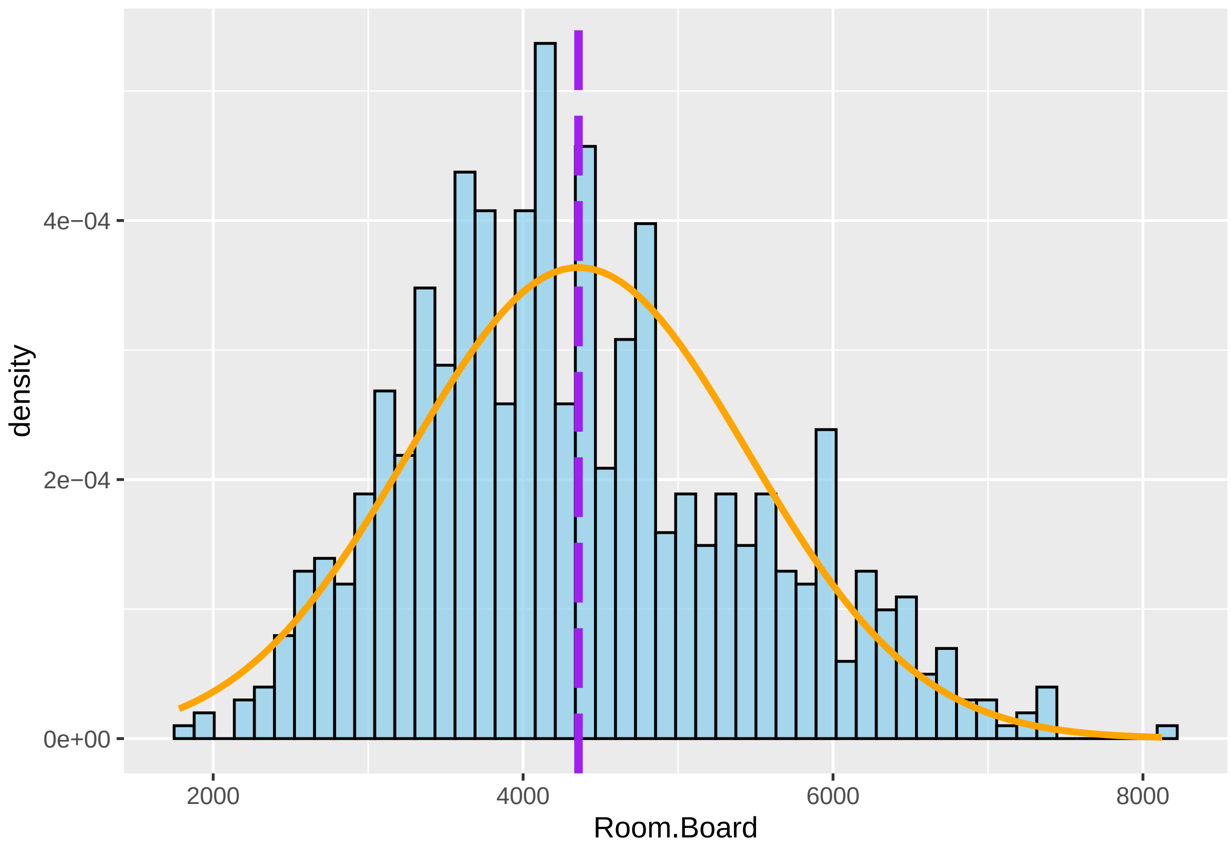
```
## Warning: package 'ggExtra' was built under R version 4.4.3
```

```
par(mfrow = c(2, 2))

# Outstate tuition
ggplot(college, aes(x = Outstate)) +
  geom_histogram(aes(y = ..density..), bins = 50, fill = "skyblue", color = "black", alpha = 0.7) +
  stat_function(fun = dnorm,
                args = list(mean = mean(Outstate, na.rm = TRUE),
                            sd = sd(Outstate, na.rm = TRUE)),
                color = "orange", linewidth = 1.2) +
  geom_vline(xintercept = mean(Outstate), color="purple", linetype="longdash", linewidth=1.5)
```

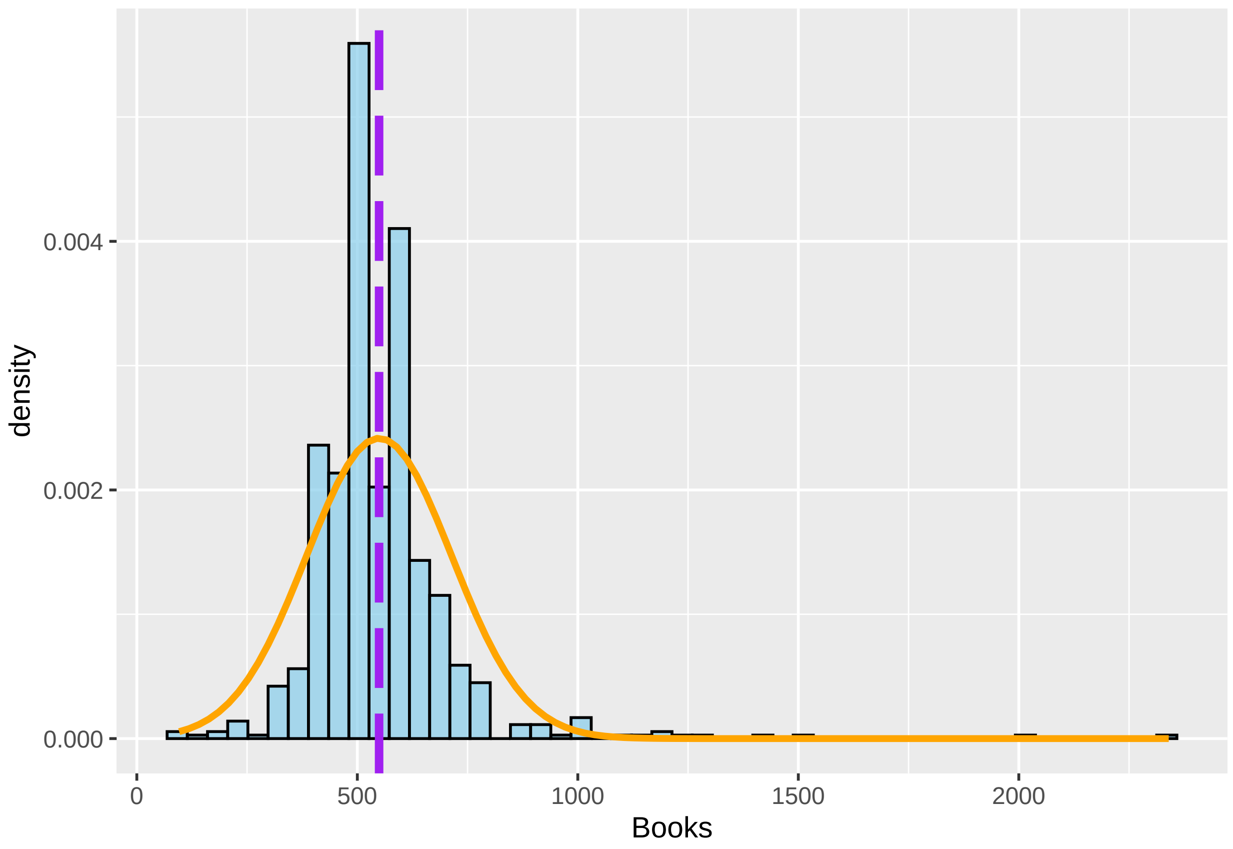
```
## Warning: The dot-dot notation (`..density..`) was deprecated in ggplot2 3.4.0.
## i Please use `after_stat(density)` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



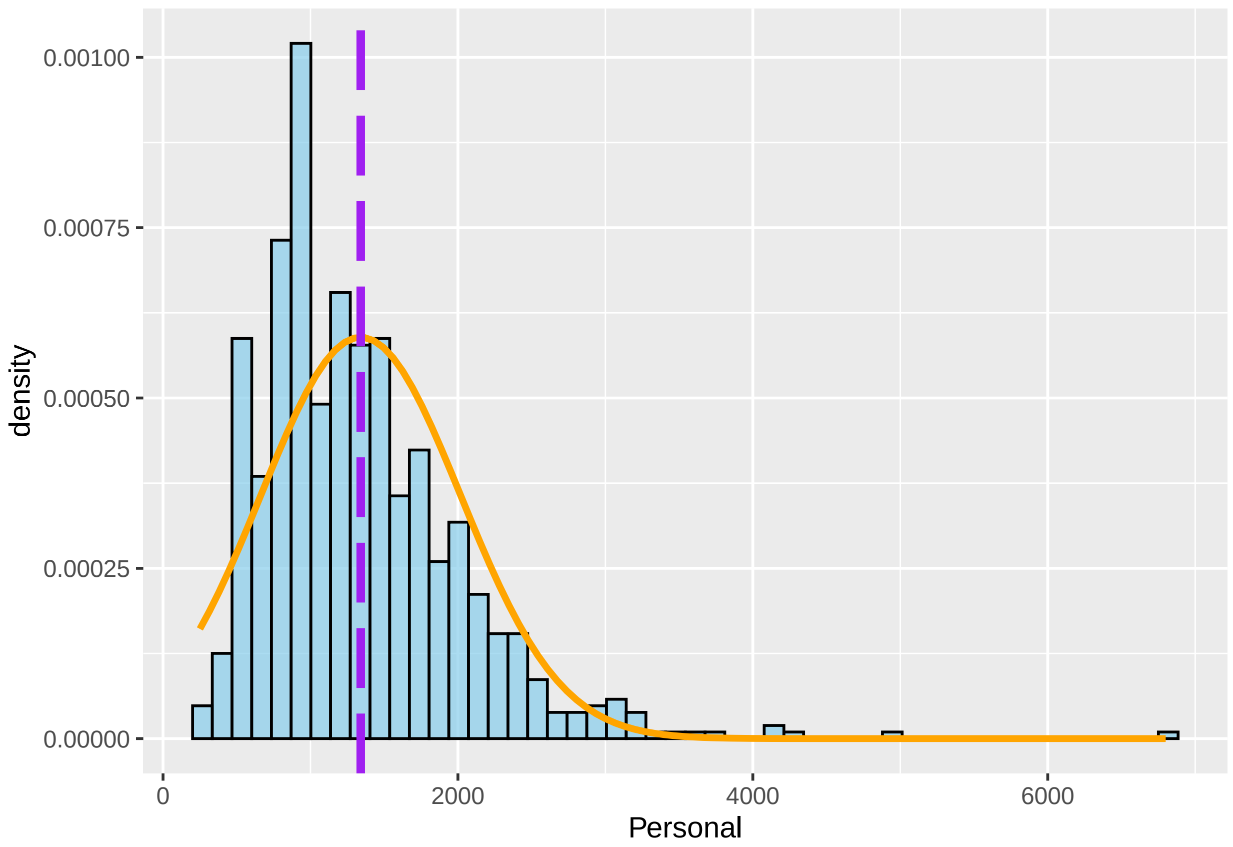
```r
# R&B costs
ggplot(college, aes(x = Room.Board)) +
  geom_histogram(aes(y = ..density..), bins = 50, fill = "skyblue", color = "black", alpha = 0.7) +
  stat_function(fun = dnorm,
                args = list(mean = mean(Room.Board, na.rm = TRUE),
                            sd = sd(Room.Board, na.rm = TRUE)),
                color = "orange", linewidth = 1.2)  +
  geom_vline(xintercept = mean(Room.Board), color="purple", linetype="longdash", linewidth=1.5)
```



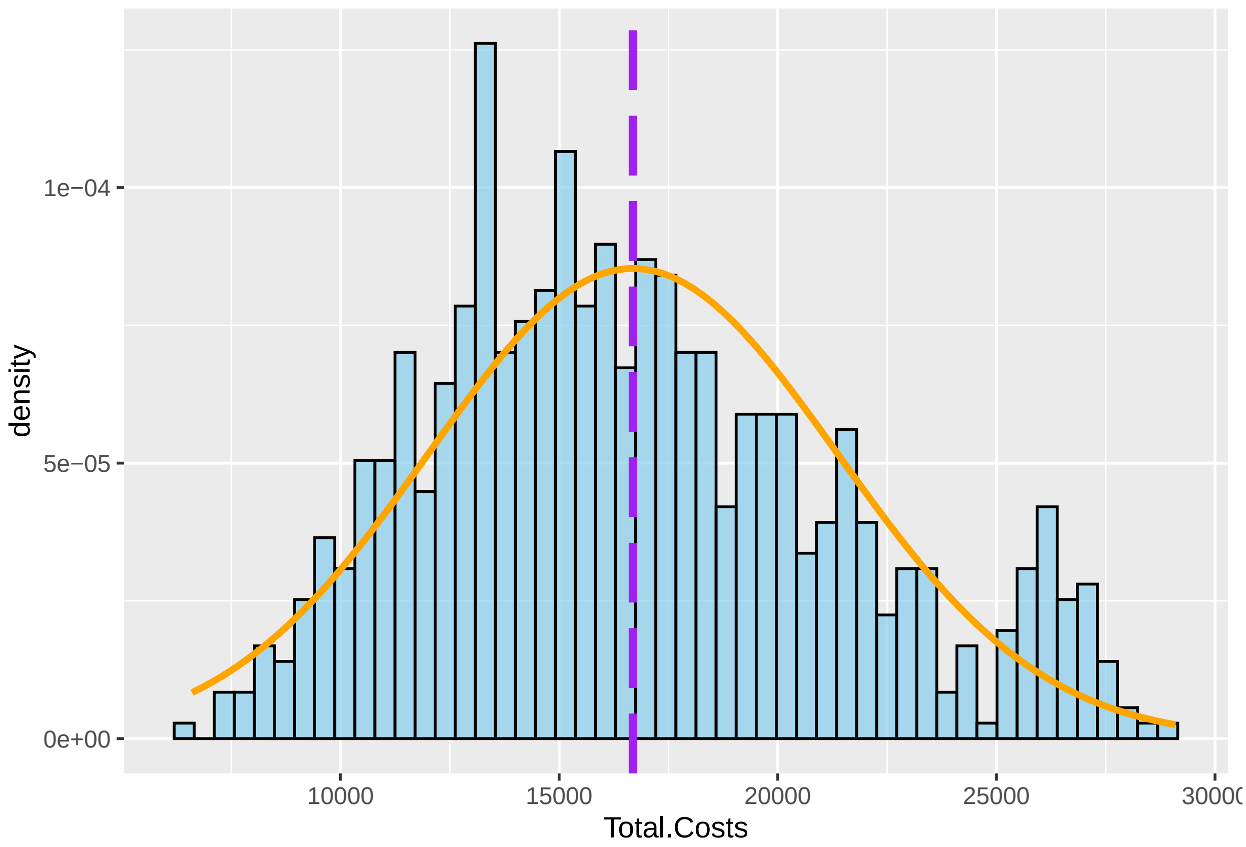
```r
# Book Costs
ggplot(college, aes(x = Books)) +
  geom_histogram(aes(y = ..density..), bins = 50, fill = "skyblue", color = "black", alpha = 0.7) +
  stat_function(fun = dnorm,
                args = list(mean = mean(Books, na.rm = TRUE),
                            sd = sd(Books, na.rm = TRUE)),
                color = "orange", linewidth = 1.2)  +
  geom_vline(xintercept = mean(Books), color="purple", linetype="longdash", linewidth=1.5)
```



```r
# Personal spending
ggplot(college, aes(x = Personal)) +
  geom_histogram(aes(y = ..density..), bins = 50, fill = "skyblue", color = "black", alpha = 0.7) +
  stat_function(fun = dnorm,
                args = list(mean = mean(Personal, na.rm = TRUE),
                            sd = sd(Personal, na.rm = TRUE)),
                color = "orange", linewidth = 1.2)  +
  geom_vline(xintercept = mean(Personal), color="purple", linetype="longdash", linewidth=1.5)
```



```
# Total Costs
Total.Costs <- Outstate+Room.Board+Books+Personal
ggplot(college, aes(x = Total.Costs)) +
  geom_histogram(aes(y = ..density..), bins = 50, fill = "skyblue", color = "black", alpha = 0.7) +
  stat_function(fun = dnorm,
                args = list(mean = mean(Total.Costs, na.rm = TRUE),
                            sd = sd(Total.Costs, na.rm = TRUE)),
                color = "orange", linewidth = 1.2)  +
  geom_vline(xintercept = mean(Total.Costs), color="purple", linetype="longdash", linewidth=1.5)
```


1e−04
5e−05
0e+00
density
10000
15000
20000
25000
30000
Total.Costs