Homework 2

Daniel Tshiani

2025-05-25

1

- for TV Ho: TV does not effect the number of units sold Ha: TV effects the number of units sold the P-value is < 0.01 which means we reject the Null. we conclude that a one unit increase in TVs increases the number of units sold by about 0.046 units.
- for radio Ho: radio does not effect the number of units sold Ha: radio effects the number of units sold the P-value is < 0.01 which means we reject the Null. we conclude that a one unit increase in radios increases the number of units sold by about 0.189 units.
- for newspaper Ho: newspaper does not effect the number of units sold Ha: newspaper effects the number of units sold the P-value is > 0.01 which means we fail to reject the Null. we conclude that does not have a significant effect of the number of units sold.

2

StartingSalary = beta0 + beta1 · GPA + beta2 · IQ + beta3 · Gender + beta4 · (GPA × IQ) + beta5 · (GPA × Gender)

Estimated model: StartingSalary = 50 + 20 · GPA + 0.07 · IQ + 35 · Gender + 0.01 · (GPA × IQ) - 10 · (GPA × Gender)

a

- For a fixed value of IQ and GPA, males earn more on average than females provided that the GPA is high enough

b

```r
50 + 20*4 + 0.07*110 + 35*1 + 0.01*(4 * 110) - 10*(4 * 1)
```

```
## [1] 137.1
```

c

False, that logic is referring to how big or small the effect of the interaction term is if there is an effect. However, to determine if there is little evidence of the interaction term we would need to look at its P-value.

3

a

the cubic model would have a lower RSS. It has more flexibility to fit the data, so its training RSS will be less than or equal to that of the linear model.

b

When it comes to testing data, I would expect the cubic model to have a higher RSS. I think the extra flexiblility in the training part would lead to overfitting for a cubic model, especially when the true relationship is linear. the cubic model would resemble the training data too closely, meanwhile the linear model would be able to capture the overall signal.

c

I would still expect the training RSS for the cubic model to be lower than or equal to the training RSS for the linear model. I think adding higher polynomial might not be efficient, however it wouldnt increase RSS for training data, especailly if we know the relationship is not linear. It would give us a similar RSS or lower.

d

I dont think we have enough information to tell. We would probably need to know more about the shape of the data and how far from linear it is.

4

a

```r
load("../data/College.rda")
```

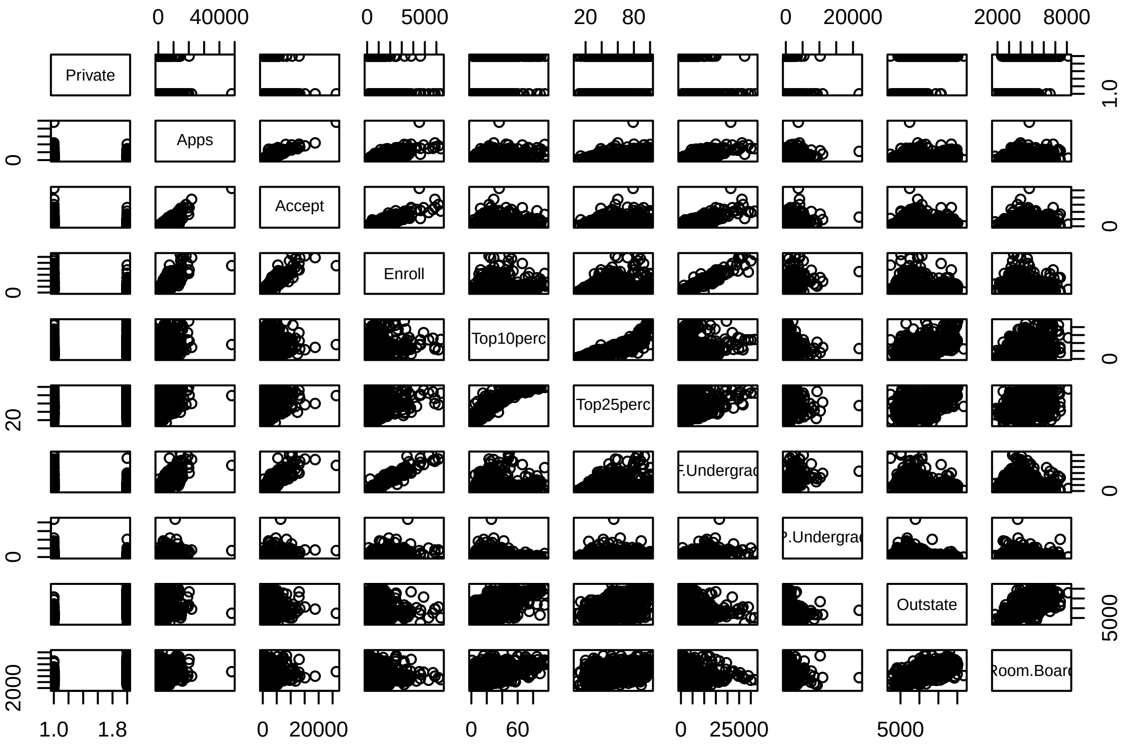
b

```r
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  
```

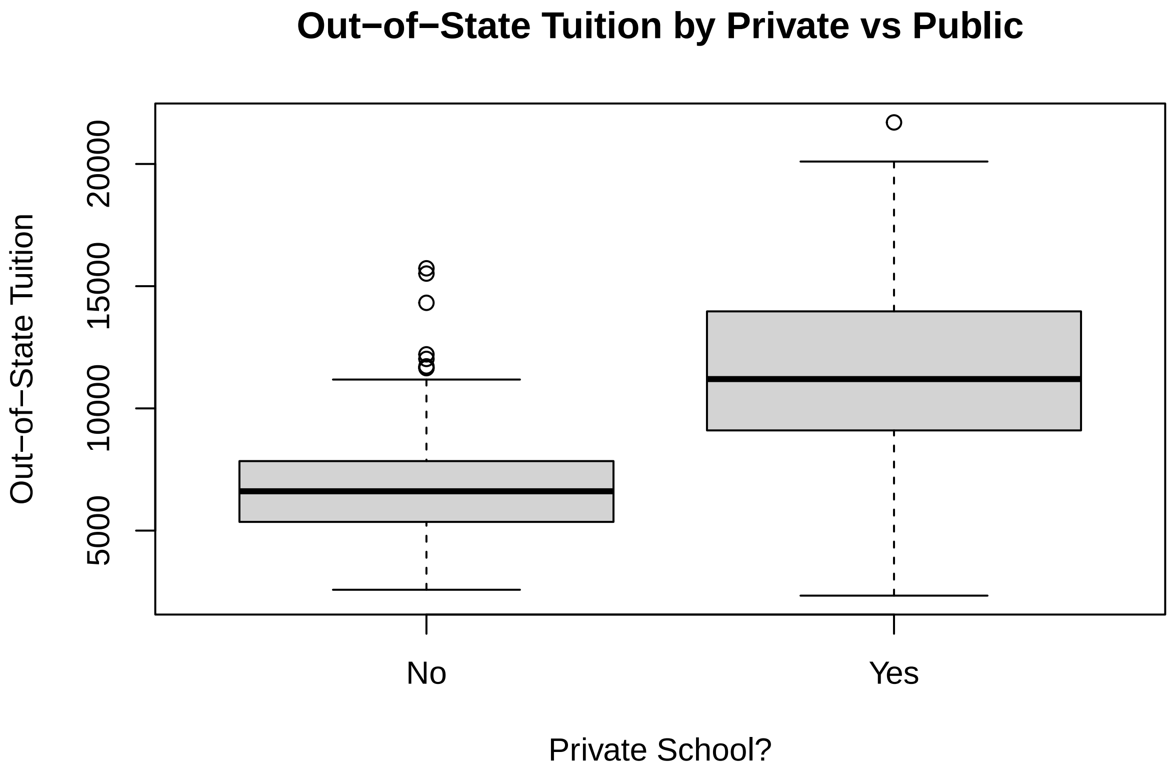
c

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



d

```r
plot(Outstate ~ Private, data = College,
     main = "Out-of-State Tuition by Private vs Public",
     ylab = "Out-of-State Tuition",
     xlab = "Private School?")
```



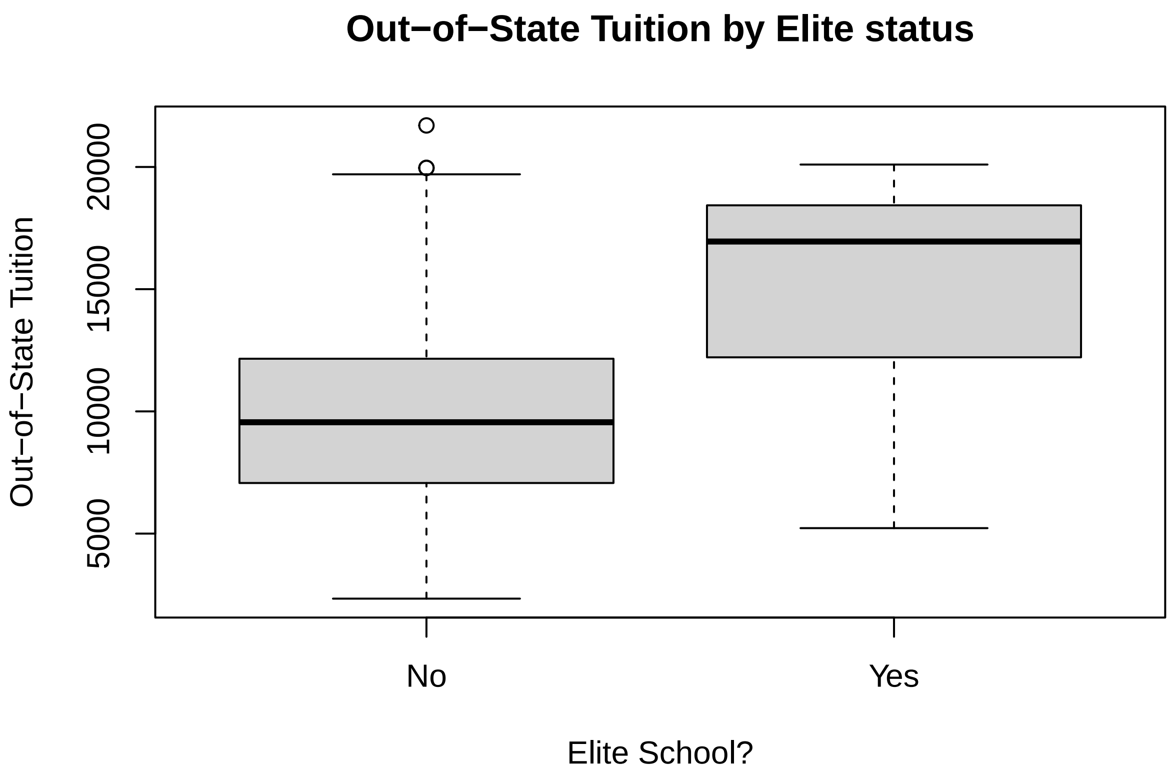
e

```
Elite = rep ("No",nrow(College))
Elite [College$ Top10perc > 50] = "Yes"
Elite = as.factor(Elite)
College = data.frame(College,Elite)
```

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

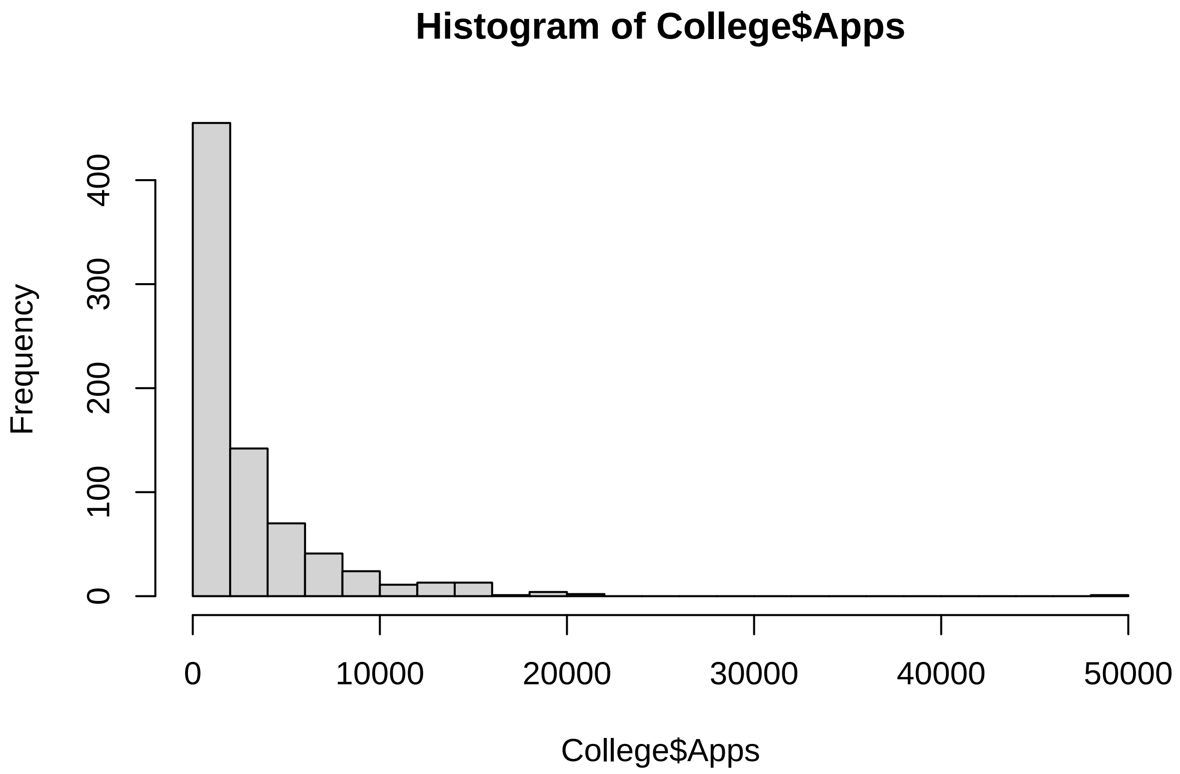
```
##  No Yes
## 699  78
```

```
plot(Outstate ~ Elite, data = College,
     main = "Out-of-State Tuition by Elite status",
     ylab = "Out-of-State Tuition",
     xlab = "Elite School?")
```

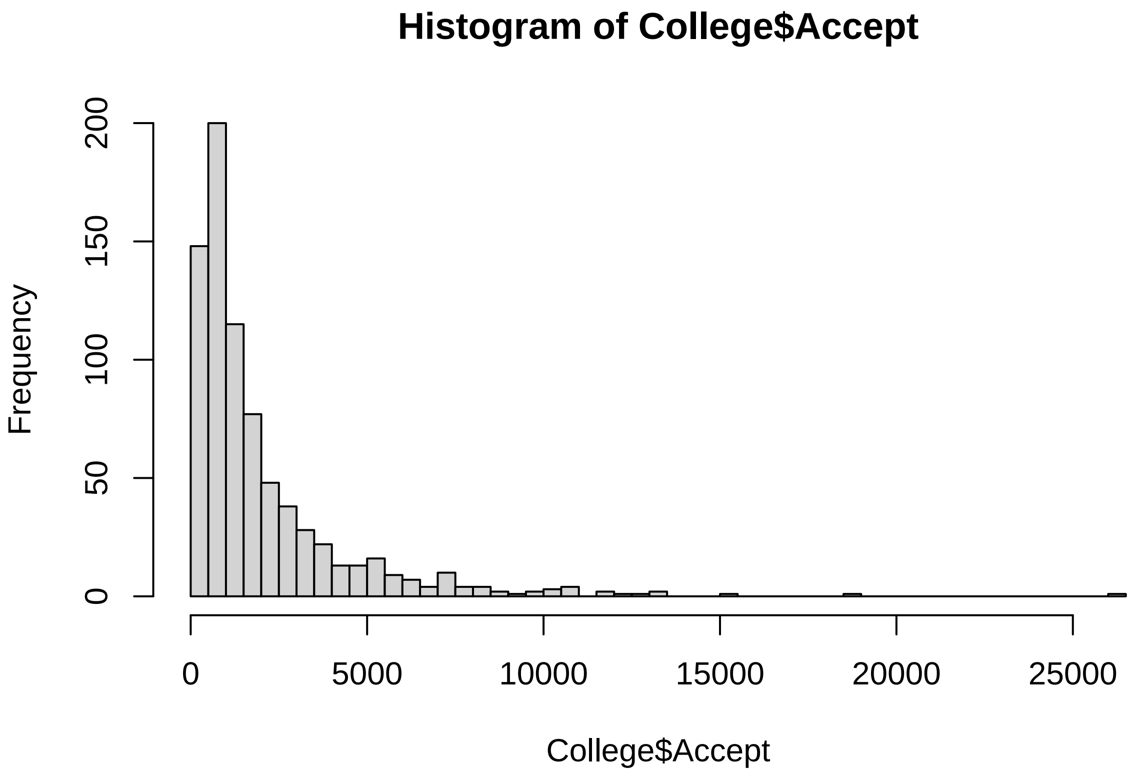


f

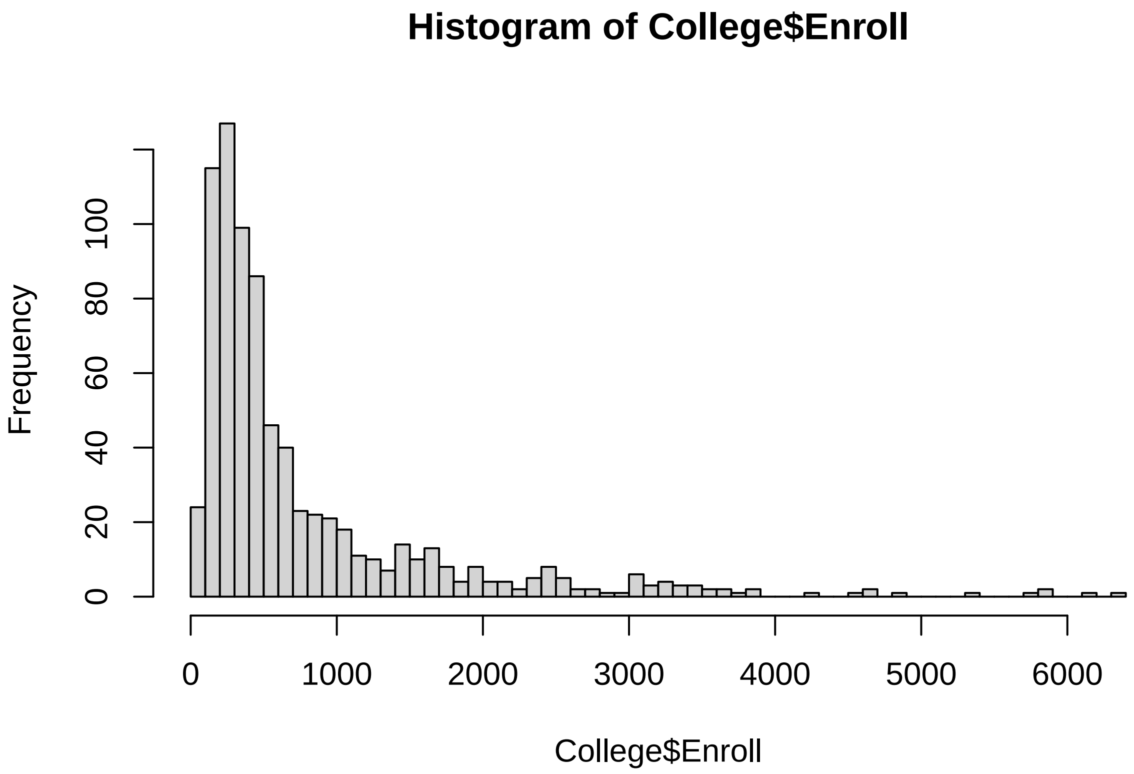
```r
hist(College$Apps, breaks = 20)
```



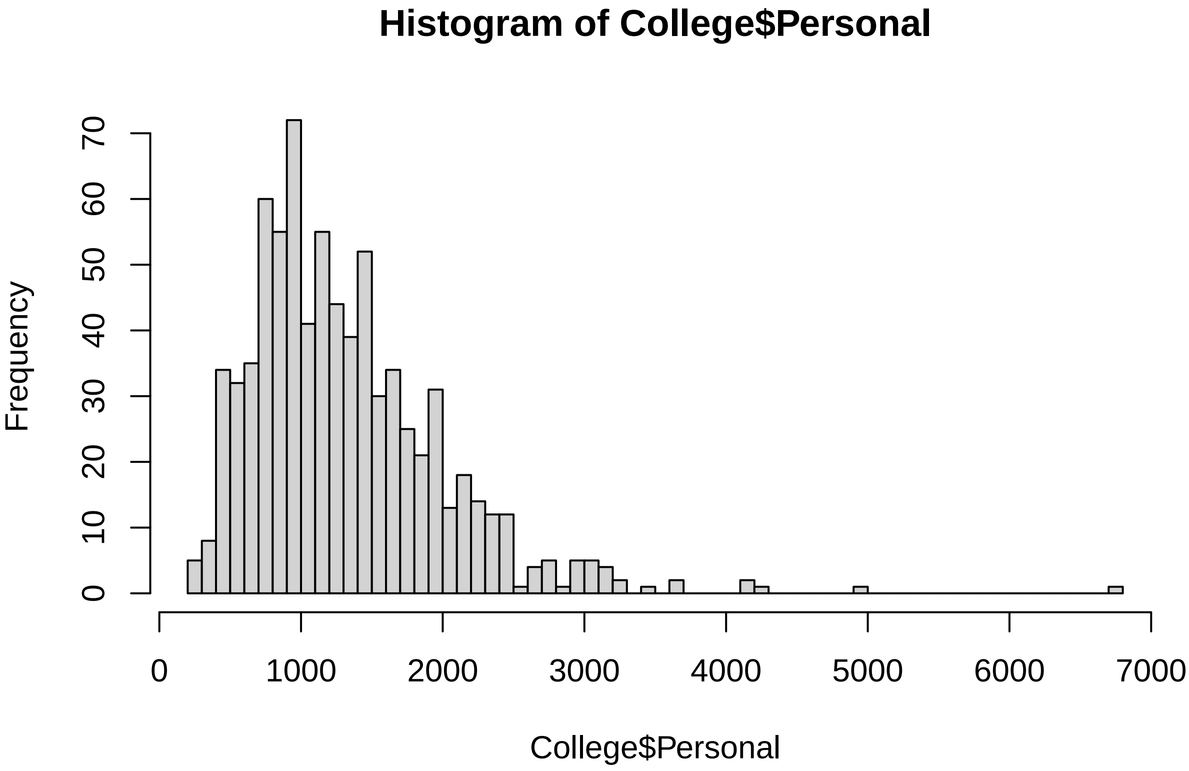
```r
hist(College$Accept, breaks = 40)
```



```r
hist(College$Enroll, breaks = 60)
```



```r
hist(College$Personal, breaks = 80)
```



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

g

```r
lm(data = College, Enroll ~Grad.Rate + PhD + Terminal + Expend)
```

```
##
## Call:
## lm(formula = Enroll ~ Grad.Rate + PhD + Terminal + Expend, data = College)
##
## Coefficients:
## (Intercept)    Grad.Rate          PhD     Terminal       Expend
##  -513.94748     -6.50790     16.54640      8.01388     -0.01253
```

5

when we calculate the line, we're trying to minimize how far off the predictions are from the actual data. The math behind it works out so that the best-fitting line naturally ends up passing through that center point of the data, which is the point (mean of x, mean of y). it's kind of like the line is balancing the data, and the center of balance is right at the average point. So no matter what the data looks like (as long as it's a simple linear regression), the line will always go through it.