CS601C Fall Project One

2024-10-17

Peter Treacy

College

A

```r
setwd("~/Desktop/CompStatistics")

data <- read.csv("~/Desktop/CompStatistics/college.csv")
```

B

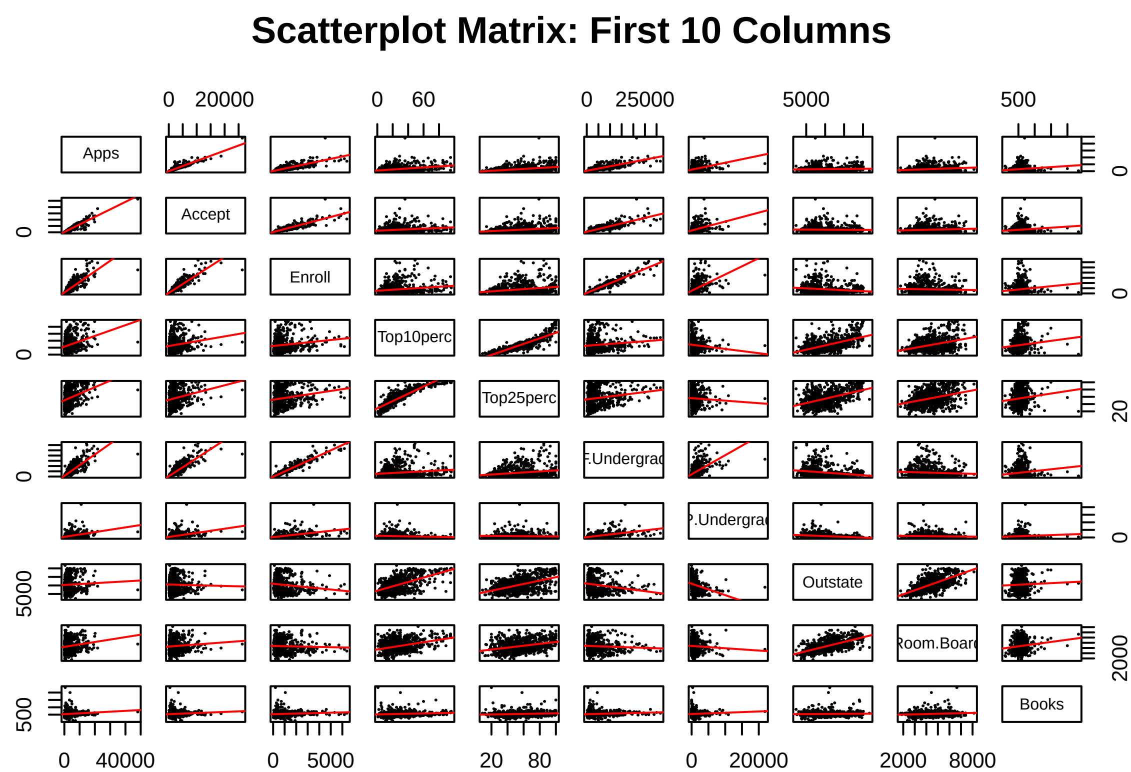
```
##                              X Private Apps Accept Enroll Top10perc Top25perc
## 1 Abilene Christian University     Yes 1660   1232    721        23        52
## 2           Adelphi University     Yes 2186   1924    512        16        29
## 3               Adrian College     Yes 1428   1097    336        22        50
## 4          Agnes Scott College     Yes  417    349    137        60        89
## 5    Alaska Pacific University     Yes  193    146     55        16        44
## 6            Albertson College     Yes  587    479    158        38        62
##   F.Undergrad P.Undergrad Outstate Room.Board Books Personal PhD Terminal
## 1        2885         537     7440       3300   450     2200  70       78
## 2        2683        1227    12280       6450   750     1500  29       30
## 3        1036          99    11250       3750   400     1165  53       66
## 4         510          63    12960       5450   450      875  92       97
## 5         249         869     7560       4120   800     1500  76       72
## 6         678          41    13500       3335   500      675  67       73
##   S.F.Ratio perc.alumni Expend Grad.Rate
## 1      18.1          12   7041        60
## 2      12.2          16  10527        56
## 3      12.9          30   8735        54
## 4       7.7          37  19016        59
## 5      11.9           2  10922        15
## 6       9.4          11   9727        55
```

C

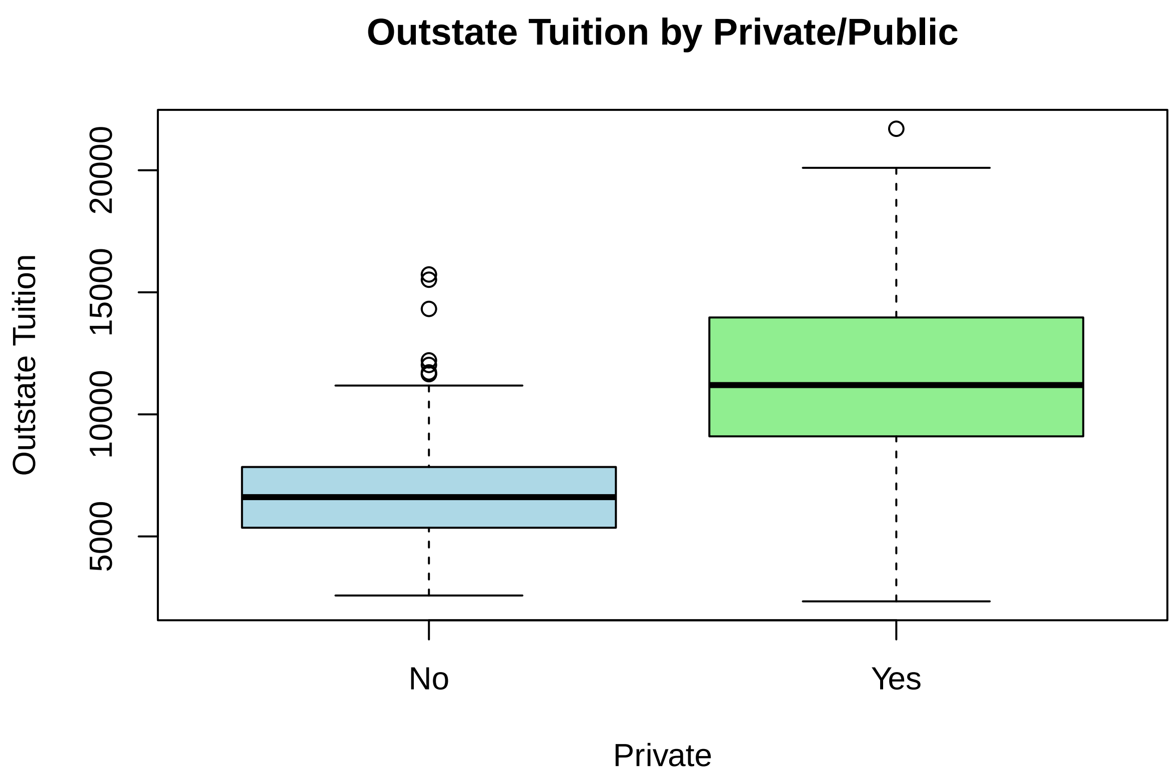
```r
#i
summary(data)
```

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

```r
#ii

data <- read.csv("~/Desktop/CompStatistics/college.csv")

num_data <- data[, sapply(data, is.numeric)]

selectnum_data <- num_data[, 1:10]

panel.custom <- function(x, y) {
  points(x, y, pch = 1, col = "black", cex = .1)
  abline(lm(y ~ x), col = "red")
}
par(mar = c(5, 5, 4, 2))

pairs(selectnum_data,
      panel = panel.custom,
      main = "Scatterplot Matrix: First 10 Columns"
)
```




```r
#iii Private
data$Private <- as.factor(data$Private)

# Boxplot
boxplot(Outstate ~ Private, data = data,
        main = "Outstate Tuition by Private/Public",
        xlab = "Private",
        ylab = "Outstate Tuition",
        col = c("lightblue", "lightgreen"))
```




```r
#iv

# Elite = 'Top10perc' column
data$Elite <- ifelse(data$Top10perc > 50, "Elite", "Non-Elite")

# Elite to a categorical variable
data$Elite <- as.factor(data$Elite)

# Out Total
table(data$Elite)
```

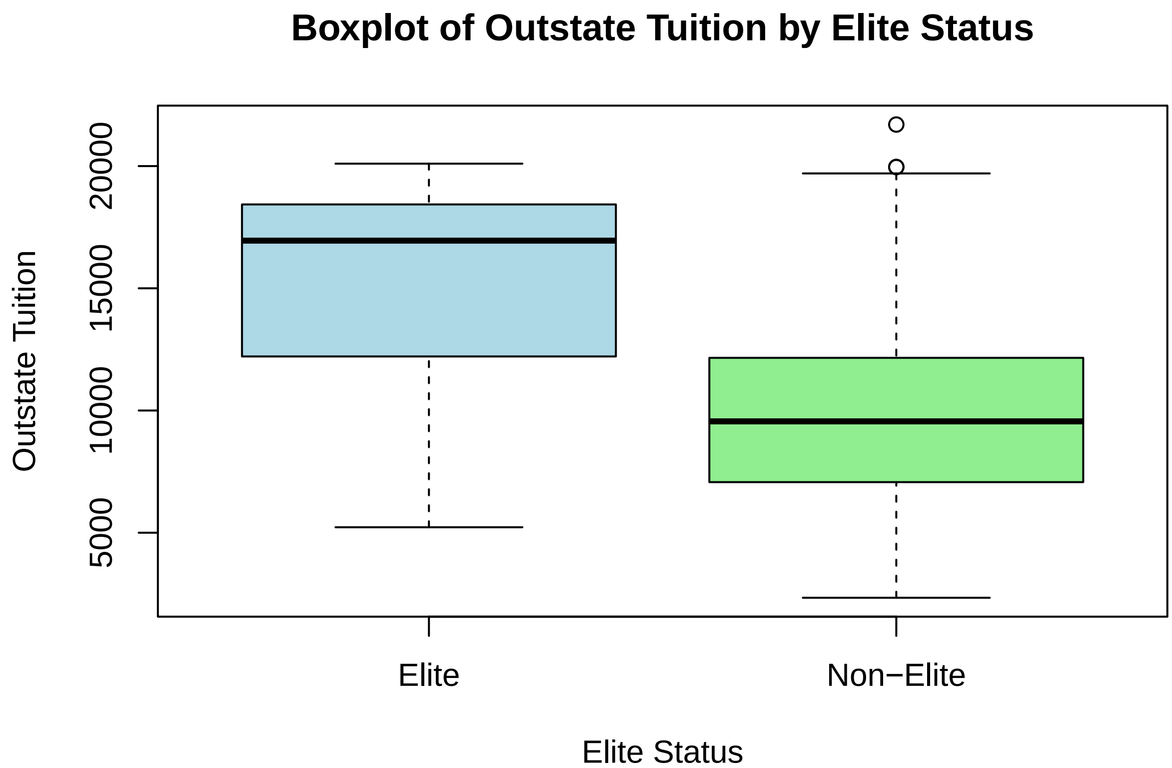
```
##
##     Elite Non-Elite
##        78       699
```

```r
#v
summary(data$Elite)
```

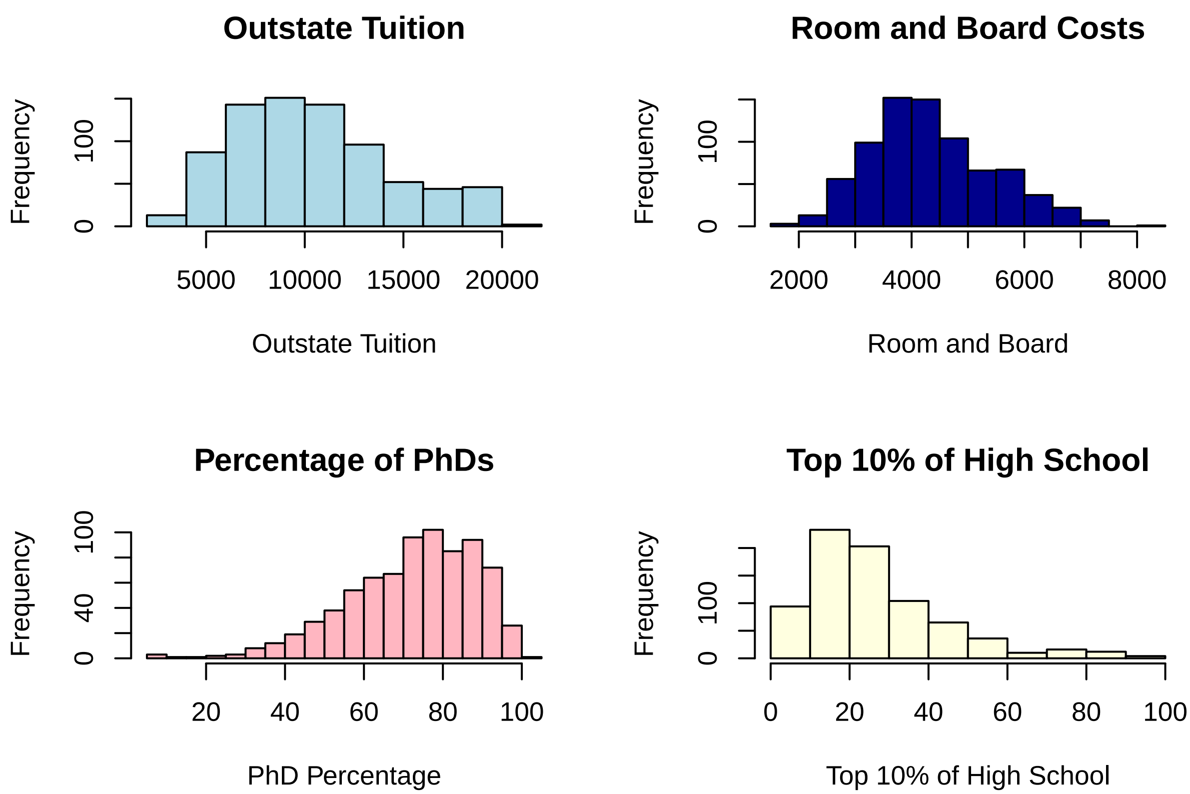
```
##     Elite Non-Elite
##        78       699
```

```r
boxplot(Outstate ~ Elite, data = data,
        main = "Boxplot of Outstate Tuition by Elite Status",
        xlab = "Elite Status",
```

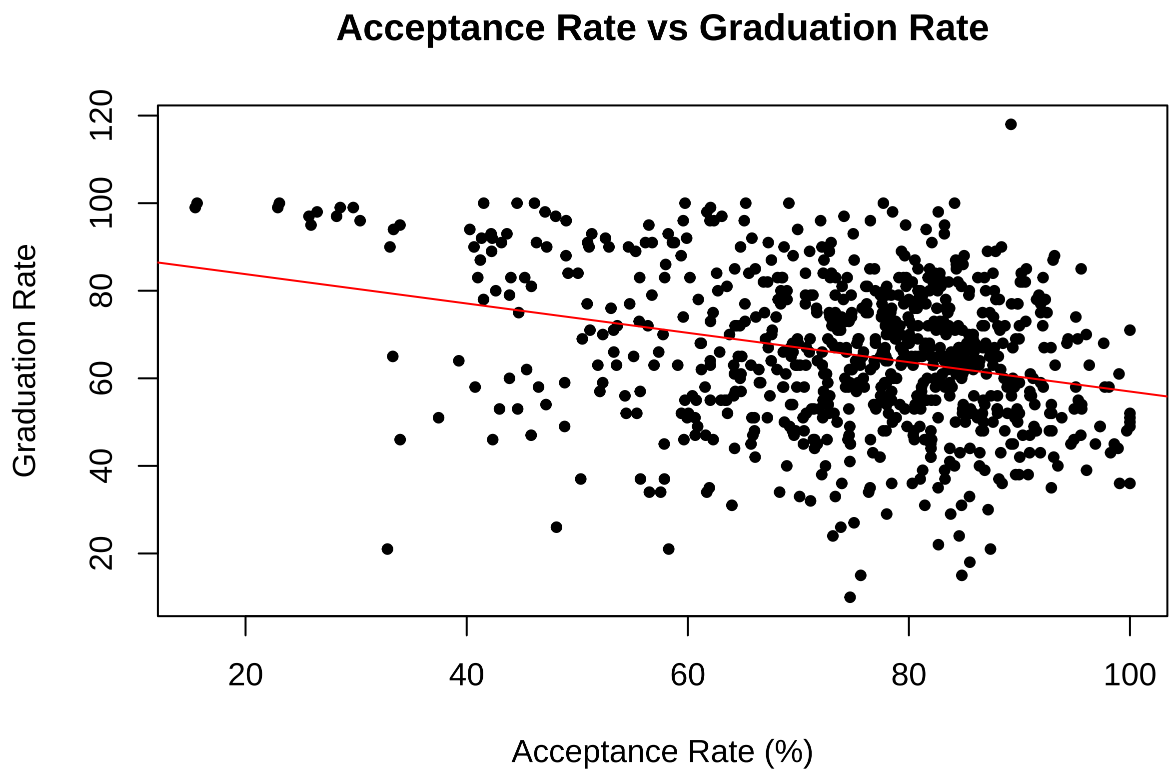
```r
        ylab = "Outstate Tuition",
        col = c("lightblue", "lightgreen"))
```




```r
#vi
data <- read.csv("~/Desktop/CompStatistics/college.csv", stringsAsFactors = TRUE)

par(mfrow=c(2,2))

hist(data$Outstate, breaks = 10,
     main = "Outstate Tuition", xlab = "Outstate Tuition",
     col = "lightblue")

hist(data$Room.Board, breaks = 15,
     main = "Room and Board Costs", xlab = "Room and Board",
     col = "darkblue")

hist(data$PhD, breaks = 20,
     main = "Percentage of PhDs", xlab = "PhD Percentage",
     col = "lightpink")

hist(data$Top10perc, breaks = 8,
     main = "Top 10% of High School", xlab = "Top 10% of High School",
     col = "lightyellow")
```




```r
#vii

data <- read.csv("college.csv")

cdata <- data[!is.na(data$Accept) & !is.na(data$Apps) & !is.na(data$Grad.Rate), ]

cdata$Accept_Rate <- (cdata$Accept / cdata$Apps) * 100

model <- lm(Grad.Rate ~ Accept_Rate, data = cdata)

plot(cdata$Accept_Rate, cdata$Grad.Rate,
     main = "Acceptance Rate vs Graduation Rate",
     xlab = "Acceptance Rate (%)",
     ylab = "Graduation Rate",
     pch = 20, col = "black")
abline(model, col = "red")
```



Conclusion:

You could conclude that colleges with higher acceptance rates tend to have higher graduation rates. This might suggest that more inclusive colleges do well at retaining and graduating students.

Auto

```
setwd("~/Desktop/CompStatistics")

auto_data <- read.csv("~/Desktop/CompStatistics/auto.csv")

str(auto_data)
```

```
## 'data.frame':    392 obs. of  9 variables:
##  $ mpg         : num  18 15 18 16 17 15 14 14 14 15 ...
##  $ cylinders   : int  8 8 8 8 8 8 8 8 8 8 ...
##  $ displacement: num  307 350 318 304 302 429 454 440 455 390 ...
##  $ horsepower  : int  130 165 150 150 140 198 220 215 225 190 ...
##  $ weight      : int  3504 3693 3436 3433 3449 4341 4354 4312 4425 3850 ...
##  $ acceleration: num  12 11.5 11 12 10.5 10 9 8.5 10 8.5 ...
##  $ year        : int  70 70 70 70 70 70 70 70 70 70 ...
##  $ origin      : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ name        : chr  "chevrolet chevelle malibu" "buick skylark 320" "plymouth satellite" "amc rebel
```

2.A

Quantitative: mpg, cylinders, displacement, horsepower, weight, acceleration, year.

Qualitative: origin, name.

2.B

```r
range(auto_data$mpg)
```

```
## [1]  9.0 46.6
```

```r
range(auto_data$cylinders)
```

```
## [1] 3 8
```

```r
range(auto_data$displacement)
```

```
## [1]  68 455
```

```r
range(auto_data$horsepower)
```

```
## [1]  46 230
```

```r
range(auto_data$weight)
```

```
## [1] 1613 5140
```

```r
range(auto_data$acceleration)
```

```
## [1]  8.0 24.8
```

```r
range(auto_data$year)
```

```
## [1] 70 82
```

2.C

```r
auto_data <- read.csv("~/Desktop/CompStatistics/auto.csv")

#mpg
mean(auto_data$mpg)
```

```
## [1] 23.44592
```

```r
sd(auto_data$mpg)
```

```
## [1] 7.805007
```

```r
#cylinders
mean(auto_data$cylinders)
```

```
## [1] 5.471939
```

```r
sd(auto_data$cylinders)
```

```
## [1] 1.705783
```

```r
#displacement
mean(auto_data$displacement)
```

```
## [1] 194.412
```

```r
sd(auto_data$displacement)
```

```
## [1] 104.644
```

```r
#horsepower
mean(auto_data$horsepower)
```

```
## [1] 104.4694
```

```r
sd(auto_data$horsepower)
```

```
## [1] 38.49116
```

```r
#weight
mean(auto_data$weight)
```

```
## [1] 2977.584
```

```r
sd(auto_data$weight)
```

```
## [1] 849.4026
```

```r
#acceleration
mean(auto_data$acceleration)
```

```
## [1] 15.54133
```

```
sd(auto_data$acceleration)
```

```
## [1] 2.758864
```

```
#year
mean(auto_data$year)
```

```
## [1] 75.97959
```

```
sd(auto_data$year)
```

```
## [1] 3.683737
```

2.D

```
auto_data <- read.csv("~/Desktop/CompStatistics/auto.csv")

#range, mean standard deviation for the original dataset
roriginal <- sapply(auto_data[, sapply(auto_data, is.numeric)], range)
moriginal <- sapply(auto_data[, sapply(auto_data, is.numeric)], mean)
sdoriginal <- sapply(auto_data[, sapply(auto_data, is.numeric)], sd)

print(roriginal)
```

```
##       mpg cylinders displacement horsepower weight acceleration year origin
## [1,]  9.0         3           68         46   1613          8.0   70      1
## [2,] 46.6         8          455        230   5140         24.8   82      3
```

```
print(moriginal)
```

```
##          mpg    cylinders displacement   horsepower       weight acceleration
##    23.445918     5.471939   194.411990   104.469388  2977.584184    15.541327
##         year       origin
##    75.979592     1.576531
```

```
print(sdoriginal)
```

```
##          mpg    cylinders displacement   horsepower       weight acceleration
##    7.8050075    1.7057832  104.6440039   38.4911599  849.4025600    2.7588641
##         year       origin
##    3.6837365    0.8055182
```

```
#remove 10 rows
rows_to_remove <- sample(1:nrow(auto_data), 10)
auto_data_subset <- auto_data[-rows_to_remove, ]

#range, mean, and standard deviation for the subset
```

```r
range_subset <- sapply(auto_data_subset[, sapply(auto_data_subset, is.numeric)], range)
mean_subset <- sapply(auto_data_subset[, sapply(auto_data_subset, is.numeric)], mean)
sd_subset <- sapply(auto_data_subset[, sapply(auto_data_subset, is.numeric)], sd)

print(range_subset)
```

```
##       mpg cylinders displacement horsepower weight acceleration year origin
## [1,]  9.0         3           68         46   1613          8.0   70      1
## [2,] 46.6         8          455        230   5140         24.8   82      3
```

```r
print(mean_subset)
```

```
##          mpg    cylinders displacement   horsepower       weight acceleration
##    23.496335     5.473822   194.053665   104.392670  2976.356021    15.554974
##         year       origin
##    75.986911     1.578534
```

```r
print(sd_subset)
```

```
##          mpg    cylinders displacement   horsepower       weight acceleration
##    7.8352683    1.7031993  104.6630658   38.6670660  851.2296734    2.7574788
##         year       origin
##    3.6825953    0.8054017
```

```r
#mean (percentage change)
mean_comparison <- 100 * (moriginal - mean_subset) / moriginal
mean_comparison
```

```
##          mpg    cylinders displacement   horsepower       weight acceleration
## -0.215034064 -0.034415846  0.184312127  0.073435482  0.041246952 -0.087812912
##         year       origin
## -0.009633058 -0.127077721
```

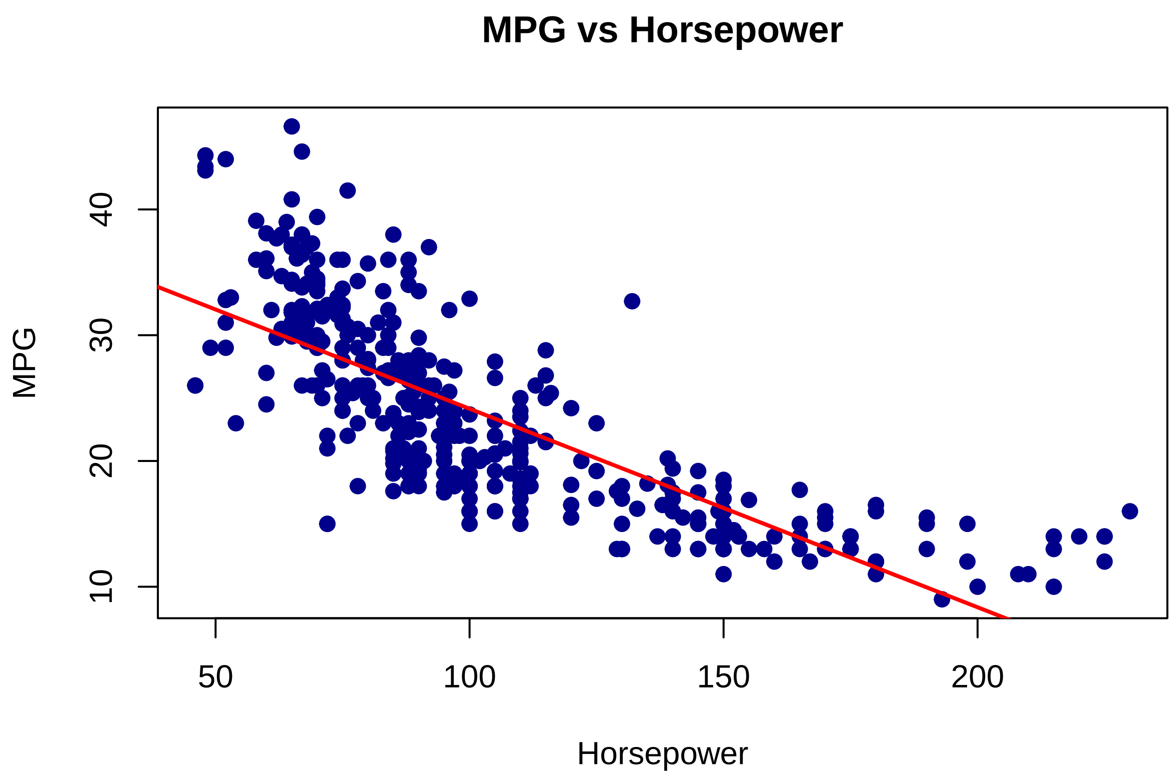
```r
#standard deviation (percentage change)
sd_comparison <- 100 * (sdoriginal - sd_subset) / sdoriginal
sd_comparison
```

```
##          mpg    cylinders displacement   horsepower       weight acceleration
##  -0.38771087   0.15148398  -0.01821597  -0.45700382  -0.21510571   0.05021487
##         year       origin
##   0.03098106   0.01446356
```

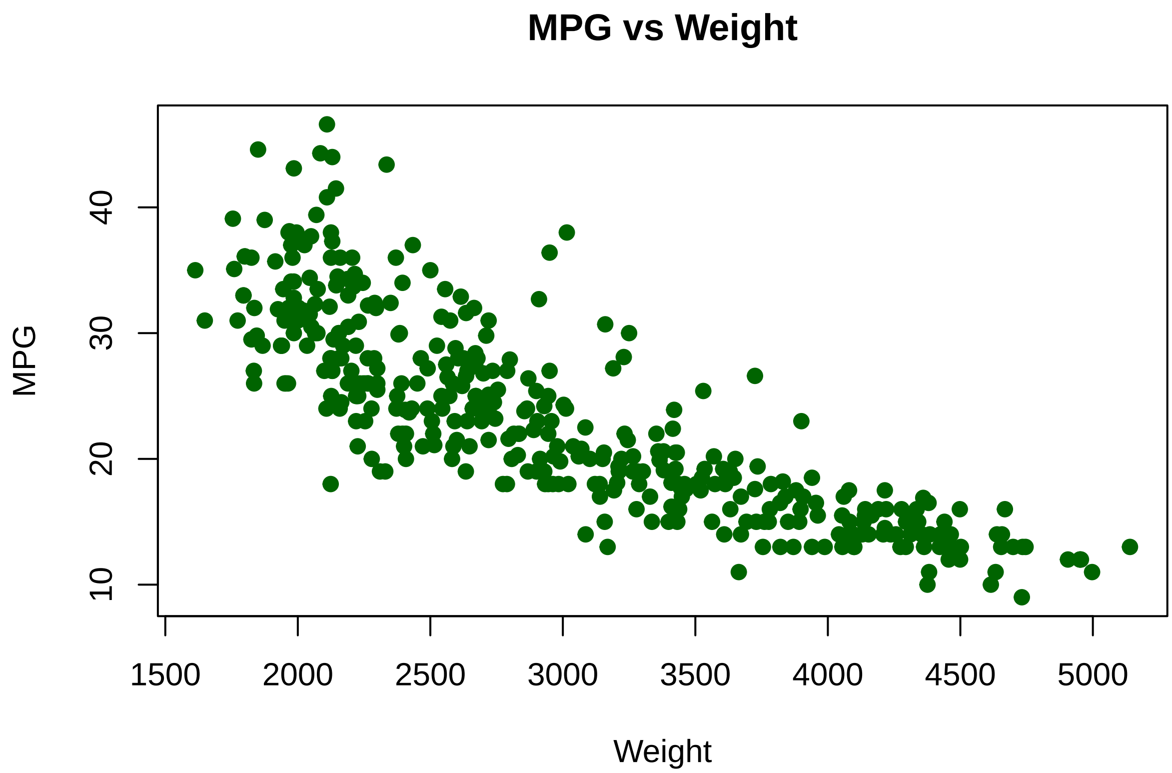
Conclusion The changes in _original to _subset were fairly minor and did not have a subsational impact on the outputs. This is probably due to how large the original dataset is.

#2.E

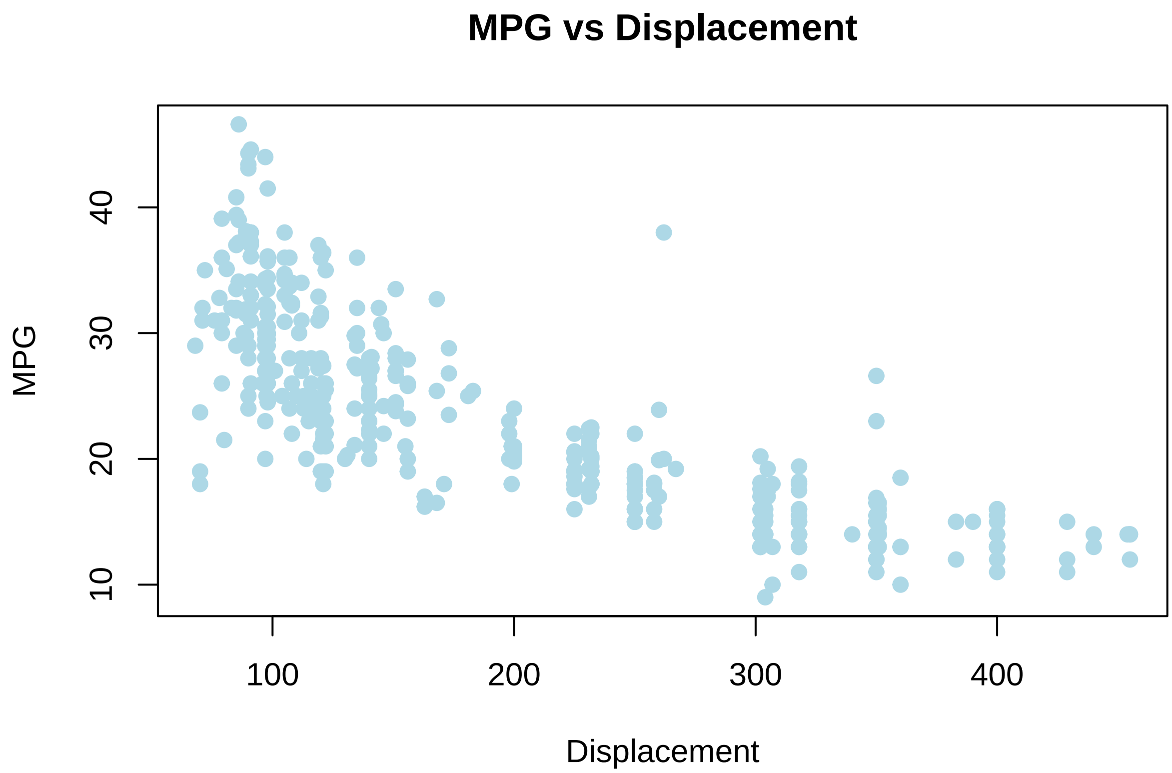
```r
auto_data <- read.csv("~/Desktop/CompStatistics/auto.csv")

library(ggplot2)

lm_mpg_hp <- lm(mpg ~ horsepower, data = auto_data)

# Scatterplot of mpg vs horsepower
plot(auto_data$horsepower, auto_data$mpg,
     main = "MPG vs Horsepower",
     xlab = "Horsepower", ylab = "MPG", pch = 19, col = "darkblue")
abline(lm_mpg_hp, col = "red", lwd = 2)
```




```r
plot(auto_data$weight, auto_data$mpg,
     main = "MPG vs Weight",
     xlab = "Weight", ylab = "MPG", pch = 19, col = "darkgreen")
```




```r
plot(auto_data$displacement, auto_data$mpg,
     main = "MPG vs Displacement",
     xlab = "Displacement", ylab = "MPG", pch = 19, col = "lightblue")
```



These plots show negative relationships between mpg and horsepower, weight, and displacement because as these variables increase the mpg tends to decrease. Horsepower is a strong predictor of mpg, as we see with the negative slope in the scatterplot. Displacement's plot shows us that larger engines(cars) typically consume more fuel, leading to lower mpg.

Boston

3.A

```
setwd("~/Desktop/CompStatistics")

boston_data <- read.csv("~/Desktop/CompStatistics/boston.csv")
# number of row
nrow(boston_data)
```

```
## [1] 506
```

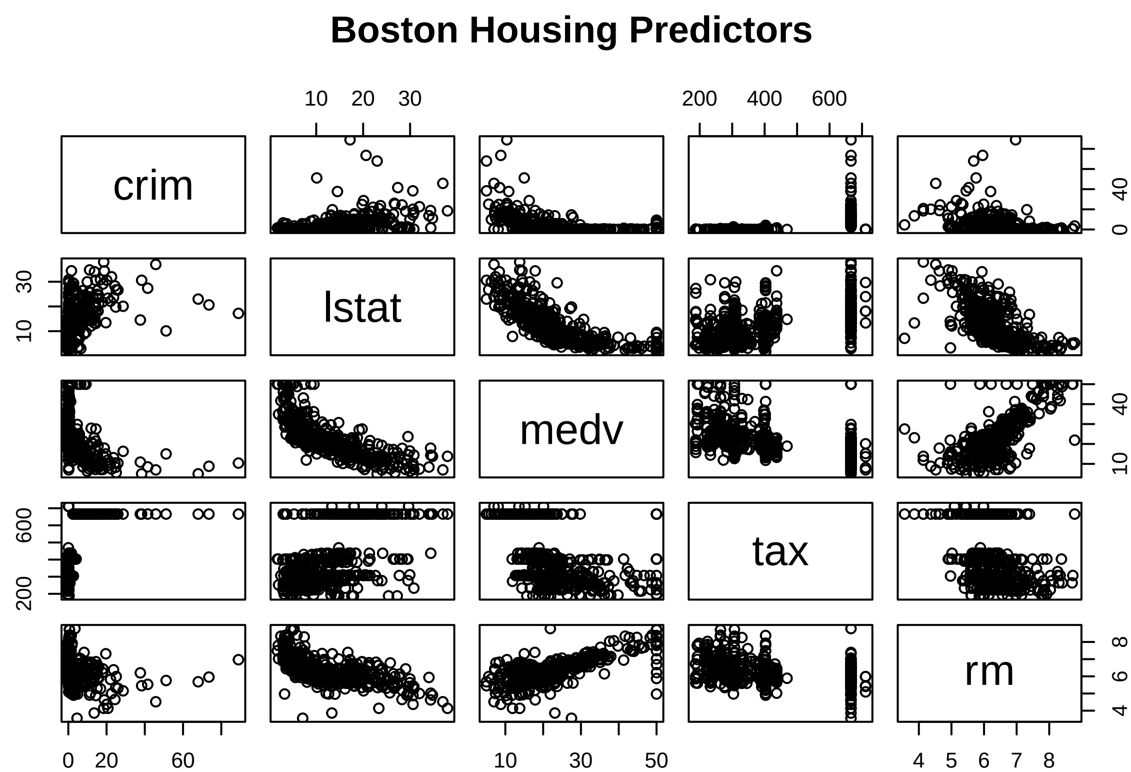
```
# number of columns
ncol(boston_data)
```

```
## [1] 14
```

Each row represents a single observation, a house or property, in a specific suburb of Boston. Each column in the dataset typically represents a different attribute (feature) of the house or the neighborhood such as crime per capita, or proportion of residential land zoned for lots over 25,000 sq. ft

3.B

```r
boston_data <- read.csv("~/Desktop/CompStatistics/boston.csv")

pairs(boston_data[, c("crim", "lstat", "medv", "tax", "rm")],
      main = "Boston Housing Predictors")
```



Conclusion

Higher crime rates and lower-status populations are associated with lower home values.

Crime rate (crim) vs. median home value (medv), and possibly crim vs. rm (number of rooms), are likely to show strong negative relationships. Higher crime rates generally correspond to lower home values and smaller homes.

Number of rooms (rm) vs. median home value (medv) typically shows a strong positive correlation, as larger homes are associated with higher home values.

Relationships between property tax (tax) and other variables like crim, rm, and medv may show weak correlations, as taxes can vary independently of home size or crime rates in different areas.

3.C

```r
boston_data <- read.csv("~/Desktop/CompStatistics/boston.csv")

#crime rate vs lower status population
crimlstat <- lm(crim ~ lstat, data = boston_data)

summary(crimlstat)
```

```
##
## Call:
## lm(formula = crim ~ lstat, data = boston_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.925  -2.822  -0.664   1.079  82.862
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.33054    0.69376  -4.801 2.09e-06 ***
## lstat        0.54880    0.04776  11.491  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.664 on 504 degrees of freedom
## Multiple R-squared:  0.2076, Adjusted R-squared:  0.206
## F-statistic:   132 on 1 and 504 DF,  p-value: < 2.2e-16
```

Conclusion

The relationship between crim and lstat is likely to be significant predictors, indicating that crime rates increase as the percentage of lower-status population increases. The scatterplot and regression model provide a clearer view of this.

3.D

```r
summary(boston_data$crim)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##  0.00632  0.08204  0.25651  3.61352  3.67708 88.97620
```

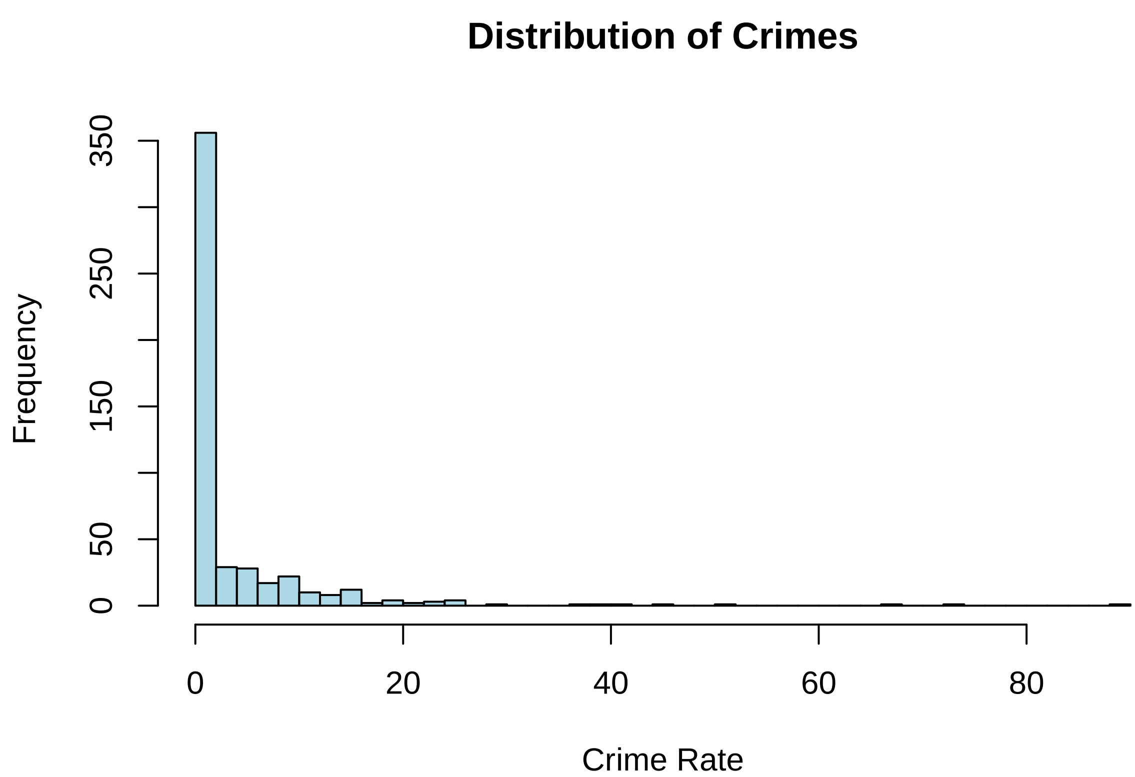
```r
summary(boston_data$tax)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   187.0   279.0   330.0   408.2   666.0   711.0
```

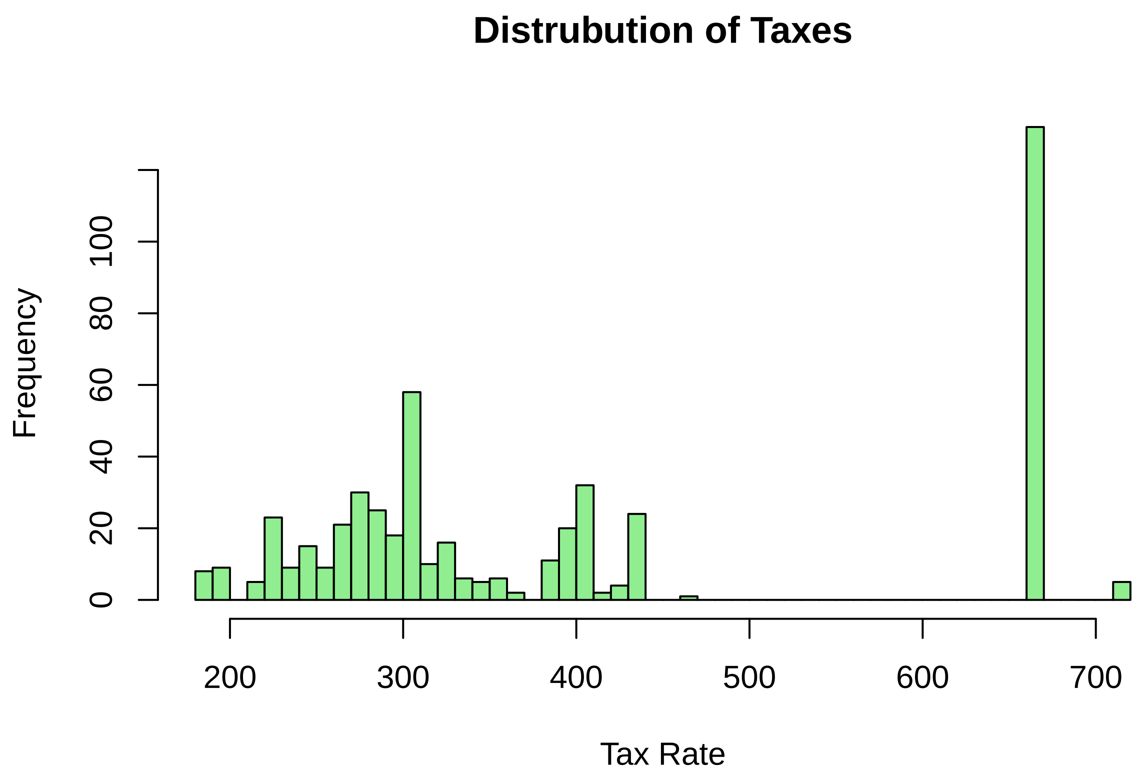
```r
# suburbs with crime rates in the top 10%
high_crime_threshold <- quantile(boston_data$crim, 0.9)
# 90th percentile
high_crime_suburbs <- boston_data[boston_data$crim > high_crime_threshold, ]

# suburbs with tax rates in the top 10%
high_tax_threshold <- quantile(boston_data$tax, 0.9)
# 90th percentile
high_tax_suburbs <- boston_data[boston_data$tax > high_tax_threshold, ]

# suburbs with high tax rates
high_tax_suburbs[, c("tax", "medv", "crim")]
```

```
##     tax medv    crim
## 489 711 15.2 0.15086
## 490 711  7.0 0.18337
## 491 711  8.1 0.20746
## 492 711 13.6 0.10574
## 493 711 20.1 0.11132
```

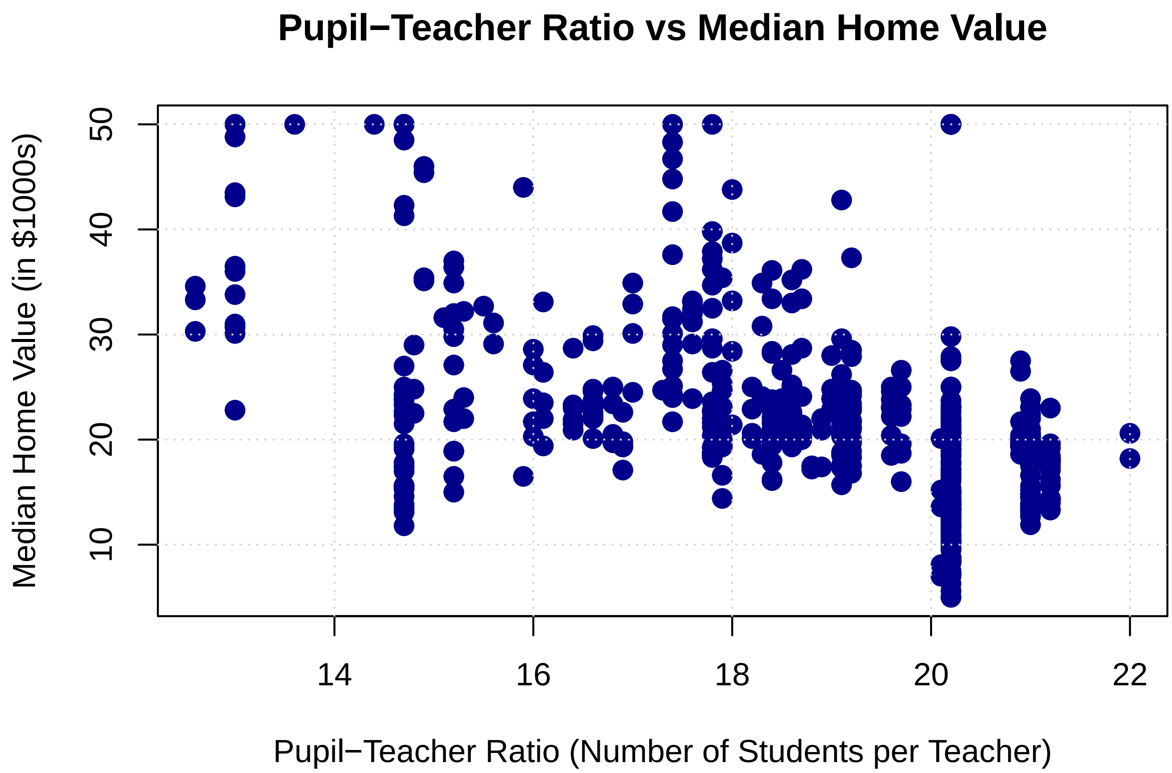
```r
hist(boston_data$crim, breaks = 50, main = "Distribution of Crimes",
     xlab = "Crime Rate", col = "lightblue")
```



```r
hist(boston_data$tax, breaks = 50, main = "Distrubution of Taxes",
     xlab = "Tax Rate", col = "lightgreen")
```




```r
# scatterplot for Pupil-Teacher Ratio vs Median Home Value
plot(boston_data$ptratio, boston_data$medv,
     main = "Pupil-Teacher Ratio vs Median Home Value",
     xlab = "Pupil-Teacher Ratio (Number of Students per Teacher)",
     ylab = "Median Home Value (in $1000s)",
     pch = 19, col = "darkblue", cex = 1.3)
grid()
```



Conclusion

The range of predictors such as crime rate (crim), tax rate (tax), and pupil-teacher ratio (ptratio) highlight disparities across suburbs.

Areas with high crime rates are likely less desirable, leading to lower property values and a lower ptratio, while some suburbs maintain very low crime levels, suggesting safer environments.

Tax rates vary widely, reflecting differences in local policies and possibly the quality of public services or schools. Likewise, the range in pupil-teacher ratios suggests disparities in educational resources, where some suburbs benefit from lower ratios and better educational quality, while others face higher ratios, potentially indicating overcrowded schools.

This variability shows the contrasting living conditions across the Boston area, with some suburbs offering safer environments, better educational opportunities, and potentially higher taxes for improved services.

3.E

```
#3.E
# number of suburbs along the Charles River

num_suburbs_bound_river <- sum(boston_data$chas == 1)

num_suburbs_bound_river
```

```
## [1] 35
```

3.F

```r
boston_data <- read.csv("~/Desktop/CompStatistics/boston.csv")

# median pupil-teacher ratio
median_ptratio <- median(boston_data$ptratio)

median_ptratio
```

```
## [1] 19.05
```

3.G

```r
# lowest median value of suburban homes
min_medv_row <- which.min(boston_data$medv)

lowest_medv_suburb <- boston_data[min_medv_row, ]
lowest_medv_suburb
```

```
##       X    crim zn indus chas   nox    rm age    dis rad tax ptratio lstat medv
## 399 399 38.3518  0  18.1    0 0.693 5.453 100 1.4896  24 666    20.2 30.59    5
```

Conclusion

The suburb with the lowest median home value (medv) (399) likely suffers from a combination of factors that make it less desirable for homebuyers.

A high crime rate (crim) is the most important negative predictor when homebuyers are viewing a suburban town. The ptratio (pupil-teacher) suggests that the educational oppurtunies in the suburb are limited, which could further qdd to it's negative appeal (children, families, etc)

The lstat (high percentage of lower-status residents) is 30.59, which is far away from the lowest on the Boston Housing list, which is 1.73. This indicates socio-economic challenges, which often correlate with lower demand for housing and lower property values. Fewer rooms per dwelling (rm) 5.453 is the median for this Boston suburb town vs the highest listed at 8.780 which is suggests that the homes are smaller, which is another reason why property values are low.

This shows how a combination of socio-economic, safety, and educational factors can influence housing markets and 399 would not be viewed as desirable.

3.H

```r
# number of suburbs with more than 7 rooms
suburb7_rooms <- sum(boston_data$rm > 7)

# the number of suburbs with more than 8 rooms
suburb8_rooms <- sum(boston_data$rm > 8)

# Display the results
suburb7_rooms
```

```
## [1] 64
```

```
suburb8_rooms
```

```
## [1] 13
```

Conclusion

Suburbs with an average of more than 8 rooms per dwelling are typically found in wealthier areas. These neighborhoods often feature higher property values, lower crime rates, and access to better school resources, offering a distinct contrast to other regions in the dataset. This analysis sheds light on the characteristics that set these affluent suburbs apart.