**Name:** HOANG VO

**MyUH ID:** 1671058

**Math 4322**

*Homework #1*

Conceptual

**Exercise #1:**

1. This scenario is a **Classification** problem because the result of having cancer is a categorical variable with Yes/No answer.

We are most interested in **Inference** because we want to know how potential of having cancer is affected as suggested factors change.

-sample size: n = 200 patients

-# of variables: p = 17 variables = measuring 12 variables + 5 survey variables

1. This scenario is a **Regression** problem because the final year GPA of a student is a quantitative variable with [0.0, 4.0].

We are most interested in **Inference** because we want to know how the final year GPA of a student is affected by changes of analyzed factors.

-sample size: n = 1500 UH students

-# of variables: p = 9 variables

1. This scenario is a **Classification** problem because the question emphasizes that the outcome result is about Who wins, A or B, a Categorical variable, not about the exact score.

We are most interested in **Prediction** because we want to know the result of Y in Y=f(X)

-Sample size: n = 300 games

-# of variables: p = 8 variables

**Exercise #2:**

A very flexible approach for regression or classification:

-advantages:

+ it may give a better fit for non-linear models

+ it decreases the bias

-disadvantages:

+ it requires estimating a greater number of parameters (variables)

+it follows the noise too closely (overfit)

+ it increases the variance

Apply:

1. The sample size is large (>20), and # of variables is large.

I prefer a very flexible approach because it has better result on large sample size while the # of variables is large.

1. The sample size is very large (>1000) but the # of variables is small (<10).

I prefer the less flexible approach because with the very large sample size, we can have good variance, good fit for non-linear models, and less bias. The extremely large sample size covers well the small # of variables.

1. The sample size is large(>10) but the # of variables is small (<10).

I suggest the less flexible approach because the # of variables is not large enough. Moreover, the variables are mostly categorical with bias choices, not quantitative variables that could provide more accurate and non-bias result.

**Exercise #3:**

-A **Parametric** statistical learning approach is using assumptions about the model. We make an assumption about the function form or shape of relationship f in Y=f(X)+E.

-A **Non-parametric** statistical learning approach has **NO** assumptions about the model. The method seeks an estimate of f that get as close to the data points as possible without being too rough or wiggly.

-Advantages and disadvantages of Parametric approach to regression and classification.

+Advantages:

> the simplifying of modeling f to a few parameters

> Not as many observations are required compared to a non-parametric approach.

+Disadvantages:

> There is potential to inaccurately estimate f if the form of f assumed is wrong

> Overfit the observations if more flexible models are used.

For example: In exercise 1, c part, if the n is small (<10 maybe), then it would require a lot of assumptions over weathers, stadium sizes, and so on to form the estimate which could be wrong. Since the values of variables are mostly of classification basing on assumptions, the chance of wrong results are high.

**Exercise #4:**

a)

> College <- read.csv("C:/Users/hoang/Downloads/math 4322/College.csv")

> View(College)

> names(College)

[1] "X" "Private" "Apps" "Accept" "Enroll"

[6] "Top10perc" "Top25perc" "F.Undergrad" "P.Undergrad" "Outstate"

[11] "Room.Board" "Books" "Personal" "PhD" "Terminal"

[16] "S.F.Ratio" "perc.alumni" "Expend" "Grad.Rate"

> College

X Private Apps Accept Enroll

1 Abilene Christian University Yes 1660 1232 721

2 Adelphi University Yes 2186 1924 512

3 Adrian College Yes 1428 1097 336

4 Agnes Scott College Yes 417 349 137

5 Alaska Pacific University Yes 193 146 55

6 Albertson College Yes 587 479 158

7 Albertus Magnus College Yes 353 340 103

8 Albion College Yes 1899 1720 489

9 Albright College Yes 1038 839 227

10 Alderson-Broaddus College Yes 582 498 172

11 Alfred University Yes 1732 1425 472

12 Allegheny College Yes 2652 1900 484

13 Allentown Coll. of St. Francis de Sales Yes 1179 780 290

14 Alma College Yes 1267 1080 385

15 Alverno College Yes 494 313 157

16 American International College Yes 1420 1093 220

17 Amherst College Yes 4302 992 418

18 Anderson University Yes 1216 908 423

19 Andrews University Yes 1130 704 322

20 Angelo State University No 3540 2001 1016

21 Antioch University Yes 713 661 252

22 Appalachian State University No 7313 4664 1910

23 Aquinas College Yes 619 516 219

24 Arizona State University Main campus No 12809 10308 3761

25 Arkansas College (Lyon College) Yes 708 334 166

26 Arkansas Tech University No 1734 1729 951

27 Assumption College Yes 2135 1700 491

28 Auburn University-Main Campus No 7548 6791 3070

29 Augsburg College Yes 662 513 257

30 Augustana College IL Yes 1879 1658 497

31 Augustana College Yes 761 725 306

32 Austin College Yes 948 798 295

33 Averett College Yes 627 556 172

34 Baker University Yes 602 483 206

35 Baldwin-Wallace College Yes 1690 1366 662

36 Barat College Yes 261 192 111

37 Bard College Yes 1910 838 285

38 Barnard College Yes 2496 1402 531

39 Barry University Yes 990 784 279

40 Baylor University Yes 6075 5349 2367

41 Beaver College Yes 1163 850 348

42 Bellarmine College Yes 807 707 308

43 Belmont Abbey College Yes 632 494 129

44 Belmont University Yes 1220 974 481

45 Beloit College Yes 1320 923 284

46 Bemidji State University No 1208 877 546

47 Benedictine College Yes 632 620 222

48 Bennington College Yes 519 327 114

49 Bentley College Yes 3466 2330 640

50 Berry College Yes 1858 1221 480

51 Bethany College Yes 878 816 200

52 Bethel College KS Yes 202 184 122

Top10perc Top25perc F.Undergrad P.Undergrad Outstate Room.Board Books

1 23 52 2885 537 7440 3300 450

2 16 29 2683 1227 12280 6450 750

3 22 50 1036 99 11250 3750 400

4 60 89 510 63 12960 5450 450

5 16 44 249 869 7560 4120 800

6 38 62 678 41 13500 3335 500

7 17 45 416 230 13290 5720 500

8 37 68 1594 32 13868 4826 450

9 30 63 973 306 15595 4400 300

10 21 44 799 78 10468 3380 660

11 37 75 1830 110 16548 5406 500

12 44 77 1707 44 17080 4440 400

13 38 64 1130 638 9690 4785 600

14 44 73 1306 28 12572 4552 400

15 23 46 1317 1235 8352 3640 650

16 9 22 1018 287 8700 4780 450

17 83 96 1593 5 19760 5300 660

18 19 40 1819 281 10100 3520 550

19 14 23 1586 326 9996 3090 900

20 24 54 4190 1512 5130 3592 500

21 25 44 712 23 15476 3336 400

22 20 63 9940 1035 6806 2540 96

23 20 51 1251 767 11208 4124 350

24 24 49 22593 7585 7434 4850 700

25 46 74 530 182 8644 3922 500

26 12 52 3602 939 3460 2650 450

27 23 59 1708 689 12000 5920 500

28 25 57 16262 1716 6300 3933 600

29 12 30 2074 726 11902 4372 540

30 36 69 1950 38 13353 4173 540

31 21 58 1337 300 10990 3244 600

32 42 74 1120 15 11280 4342 400

33 16 40 777 538 9925 4135 750

34 21 47 958 466 8620 4100 400

35 30 61 2718 1460 10995 4410 1000

36 15 36 453 266 9690 4300 500

37 50 85 1004 15 19264 6206 750

38 53 95 2121 69 17926 8124 600

39 18 45 1811 3144 11290 5360 600

40 34 66 9919 484 6450 3920 600

41 23 56 878 519 12850 5400 400

42 39 63 1198 605 8840 2950 750

43 17 36 709 131 9000 4850 300

44 28 67 1964 623 7800 3664 650

45 26 54 1085 81 16304 3616 355

46 12 36 3796 824 4425 2700 660

47 14 24 702 501 9550 3850 350

48 25 53 457 2 21700 4100 600

49 20 60 3095 1533 13800 5510 630

50 37 68 1620 49 8050 3940 350

51 16 41 706 62 8740 3363 550

52 19 42 537 101 8540 3580 500

Personal PhD Terminal S.F.Ratio perc.alumni Expend Grad.Rate

1 2200 70 78 18.1 12 7041 60

2 1500 29 30 12.2 16 10527 56

3 1165 53 66 12.9 30 8735 54

4 875 92 97 7.7 37 19016 59

5 1500 76 72 11.9 2 10922 15

6 675 67 73 9.4 11 9727 55

7 1500 90 93 11.5 26 8861 63

8 850 89 100 13.7 37 11487 73

9 500 79 84 11.3 23 11644 80

10 1800 40 41 11.5 15 8991 52

11 600 82 88 11.3 31 10932 73

12 600 73 91 9.9 41 11711 76

13 1000 60 84 13.3 21 7940 74

14 400 79 87 15.3 32 9305 68

15 2449 36 69 11.1 26 8127 55

16 1400 78 84 14.7 19 7355 69

17 1598 93 98 8.4 63 21424 100

18 1100 48 61 12.1 14 7994 59

19 1320 62 66 11.5 18 10908 46

20 2000 60 62 23.1 5 4010 34

21 1100 69 82 11.3 35 42926 48

22 2000 83 96 18.3 14 5854 70

23 1615 55 65 12.7 25 6584 65

24 2100 88 93 18.9 5 4602 48

25 800 79 88 12.6 24 14579 54

26 1000 57 60 19.6 5 4739 48

27 500 93 93 13.8 30 7100 88

28 1908 85 91 16.7 18 6642 69

29 950 65 65 12.8 31 7836 58

30 821 78 83 12.7 40 9220 71

31 1021 66 70 10.4 30 6871 69

32 1150 81 95 13.0 33 11361 71

33 1350 59 67 22.4 11 6523 48

34 2250 58 68 11.0 21 6136 65

35 1000 68 74 17.6 20 8086 85

36 500 57 77 9.7 35 9337 71

37 750 98 98 10.4 30 13894 79

38 850 83 93 10.3 33 12580 91

39 1800 76 78 12.6 11 9084 72

40 1346 71 76 18.5 38 7503 72

41 800 78 89 12.2 30 8954 73

42 1290 74 82 13.1 31 6668 84

43 2480 78 85 13.2 10 7550 52

44 900 61 61 11.1 19 7614 49

45 715 87 95 11.1 26 12957 69

46 1800 57 62 19.6 16 3752 46

47 250 64 84 14.1 18 5922 58

48 500 35 59 10.1 33 16364 55

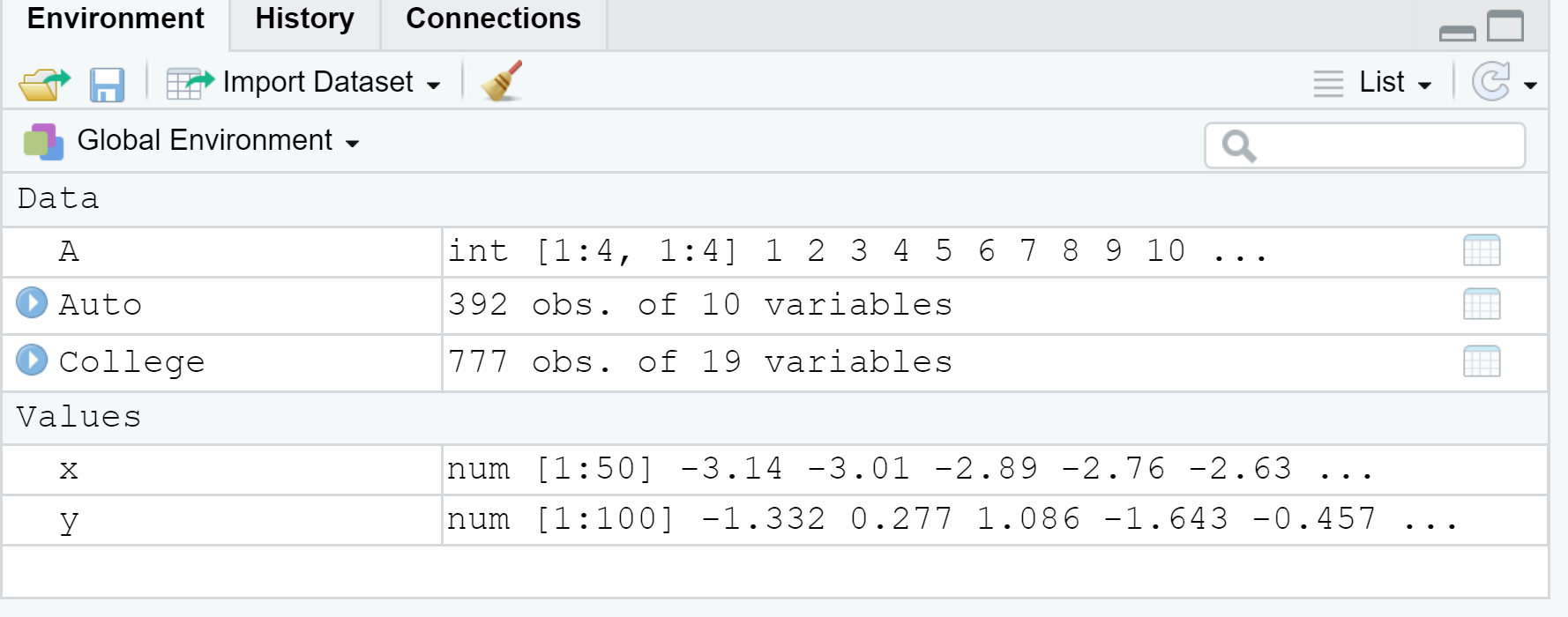
49 850 87 87 17.5 20 10941 82

50 2375 80 80 16.3 17 10511 63

51 1700 62 68 11.6 29 7718 48

52 1400 61 80 8.8 32 8324 56

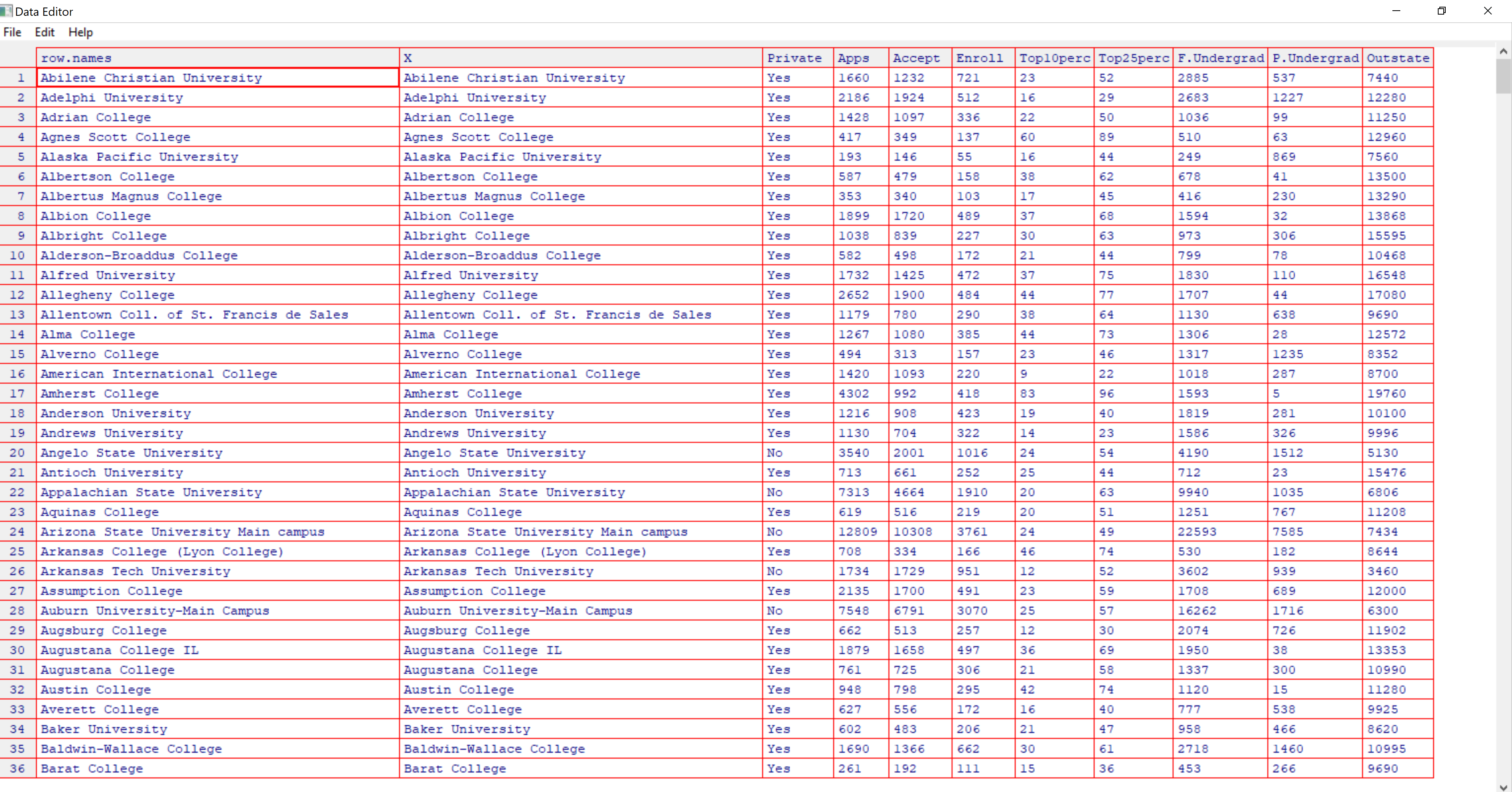
[ reached 'max' / getOption("max.print") -- omitted 725 rows ]



b)

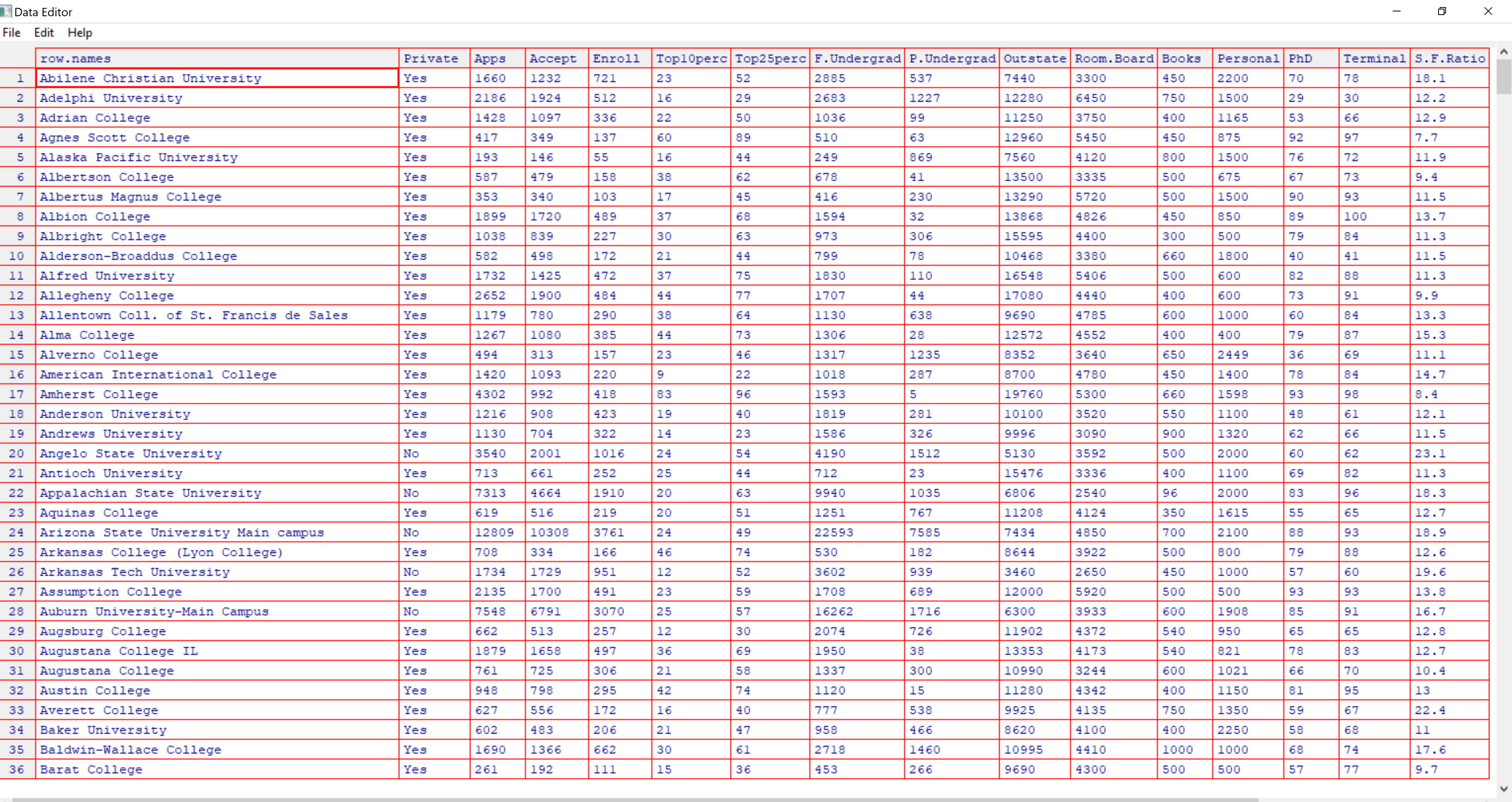
> rownames(College)= College[ ,1]

> fix(College)



> College = College[, -1]

> fix(College)



C)

c).i:

> 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

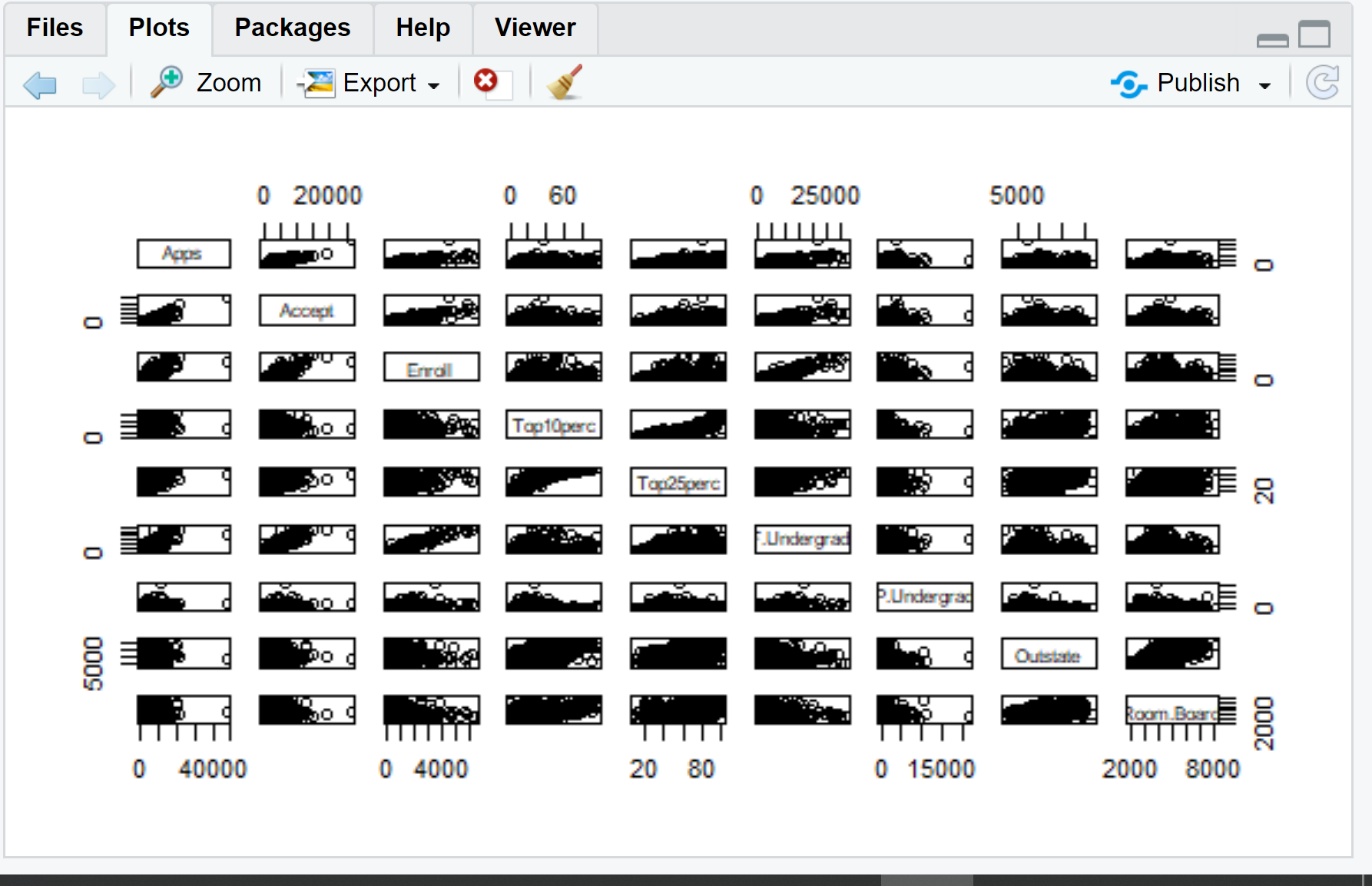
Mean : 65.46

3rd Qu.: 78.00

Max. :118.00

c).ii:

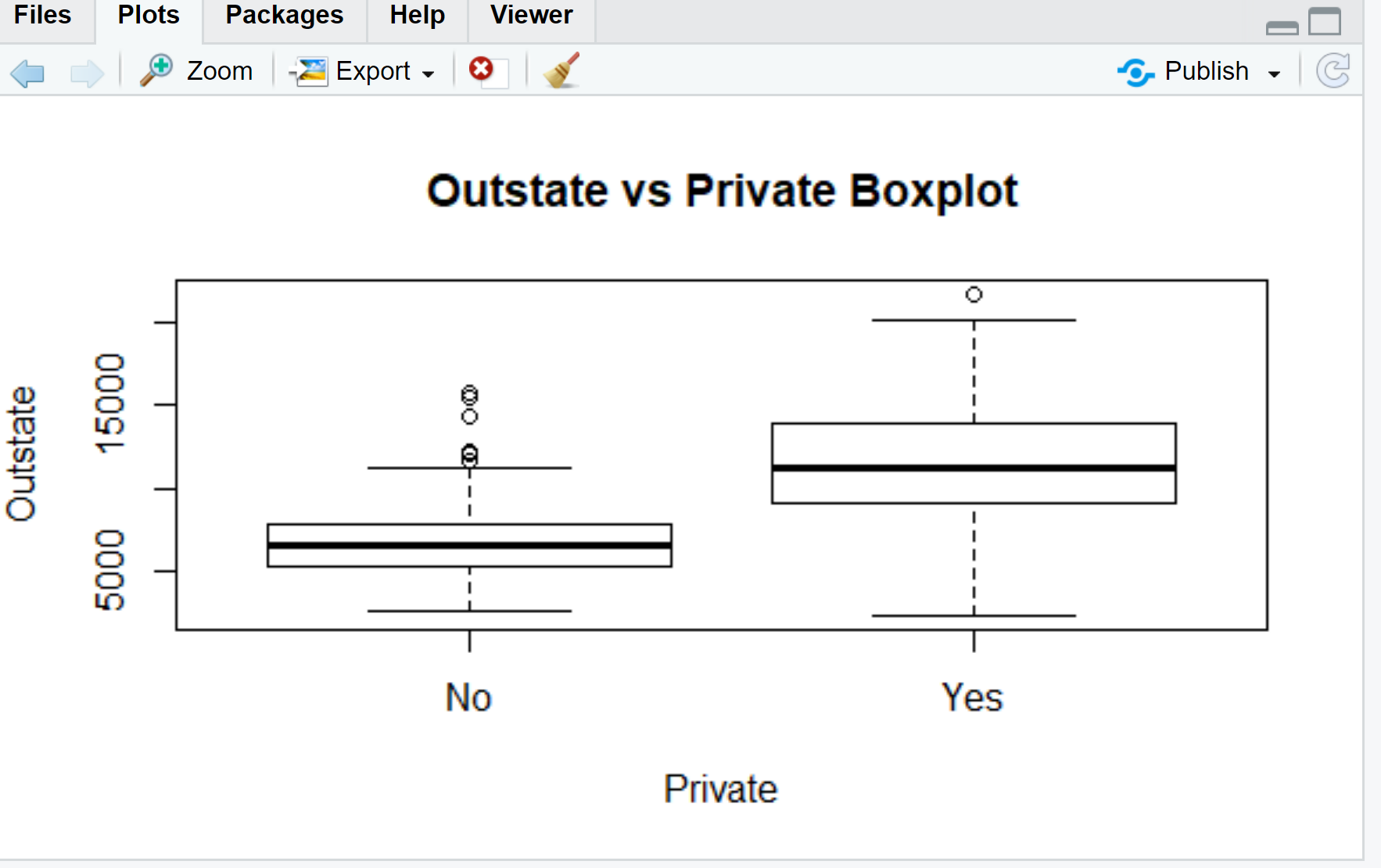
pairs(College[, 2:10])



c).iii:

> boxplot(Outstate~Private, data=College, ylab="Outstate", xlab="Private")

> title("Outstate vs Private Boxplot")



c)iv:

> summary(Elite)

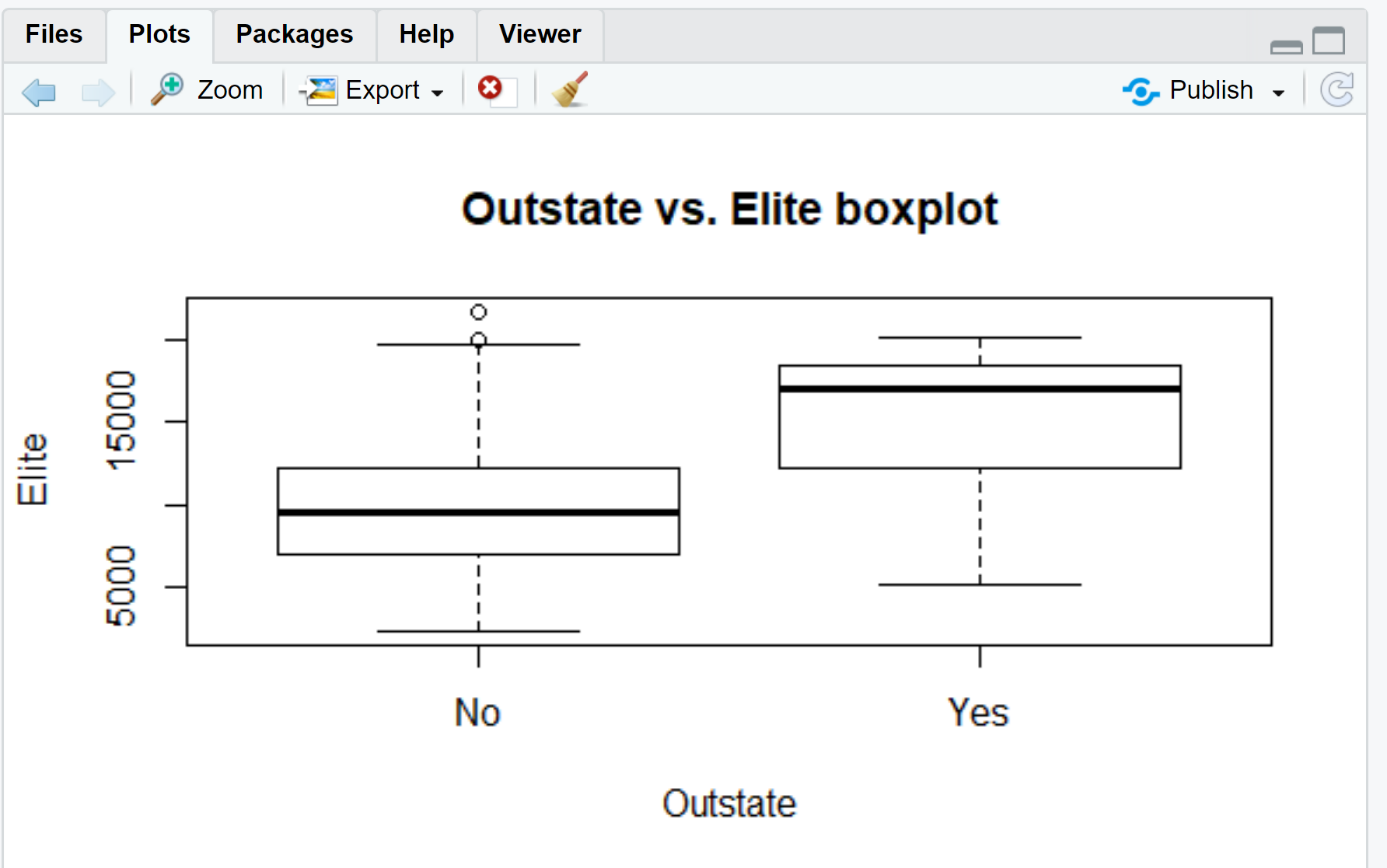
No Yes

699 78

Totally 78 Elite schools

> boxplot(Outstate~Elite, data = College, xlab="Outstate", ylab="Elite")

> title("Outstate vs. Elite boxplot")



c)v:

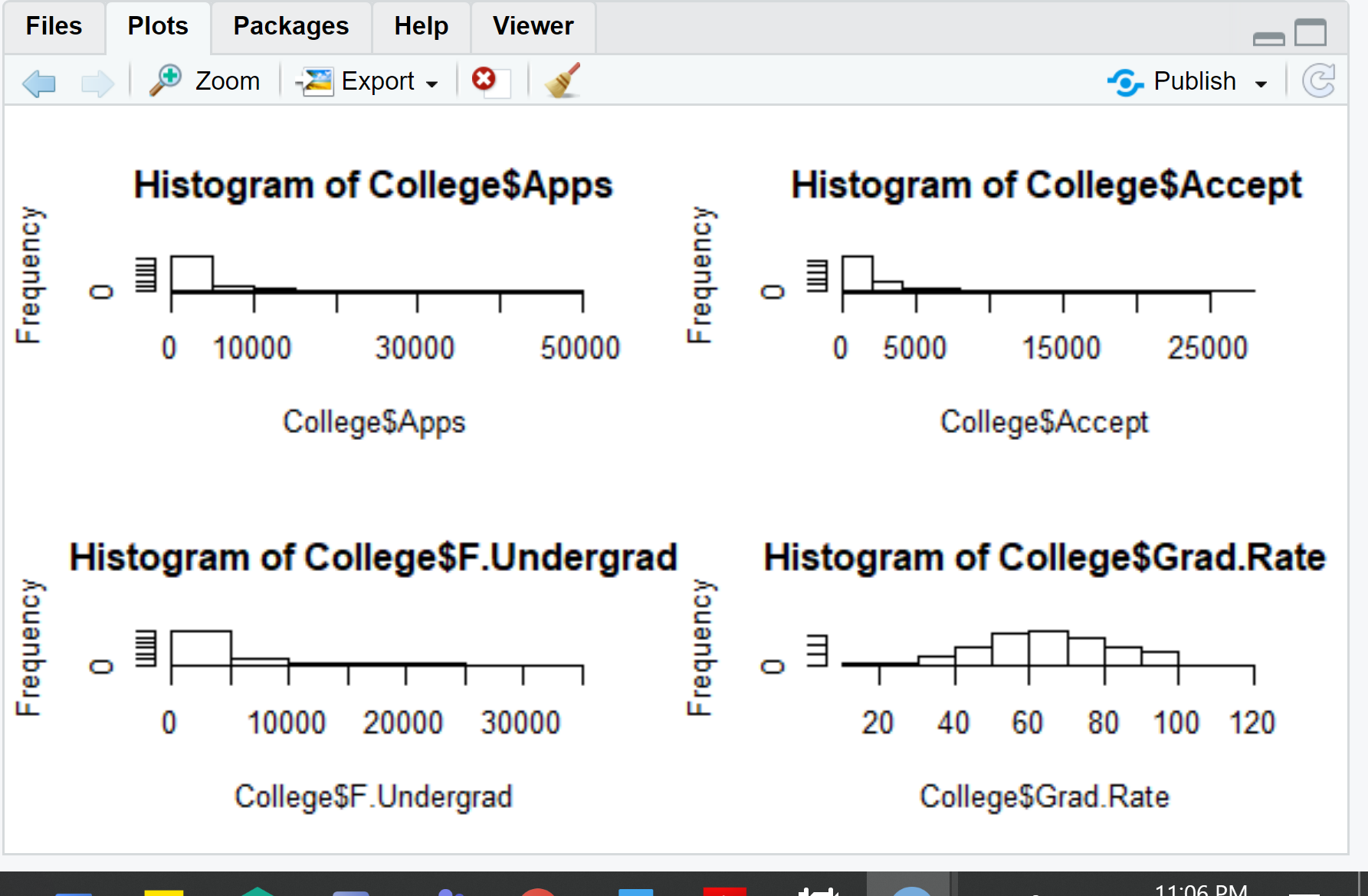
> par(mfrow= c(2,2))

> hist(College$Apps)

> hist(College$Accept)

> hist(College$F.Undergrad)

> hist(College$Grad.Rate)



**Exercise #5:**

a)

> library(MASS)

> Boston

> ?Boston

The Boston data frame has **506 rows** and **14 columns**.

b)Quantitative

c)

> names(Boston)

[1] "crim" "zn" "indus" "chas" "nox" "rm" "age"

[8] "dis" "rad" "tax" "ptratio" "black" "lstat" "medv"

> range(Boston$crim)

[1] 0.00632 88.97620

> range(Boston$zn)

[1] 0 100

> range(Boston$indus)

[1] 0.46 27.74

> range(Boston$chas)

[1] 0 1

> range(Boston$nox)

[1] 0.385 0.871

> range(Boston$rm)

[1] 3.561 8.780

> range(Boston$age)

[1] 2.9 100.0

> range(Boston$dis)

[1] 1.1296 12.1265

> range(Boston$rad)

[1] 1 24

> range(Boston$tax)

[1] 187 711

> range(Boston$ptratio)

[1] 12.6 22.0

> range(Boston$black)

[1] 0.32 396.90

> range(Boston$lstat)

[1] 1.73 37.97

> range(Boston$medv)

[1] 5 50

|  |  |  |
| --- | --- | --- |
| Predictors | Min | max |
| crim | 0.00632 | 88.97620 |
| zn | 0 | 100 |
| indus | 0.46 | 27.74 |
| chas | 0 | 1 |
| nox | 0.385 | 0.871 |
| rm | 3.561 | 8.780 |
| age | 2.9 | 100 |
| dis | 1.1296 | 12.1265 |
| rad | 1 | 24 |
| tax | 187 | 711 |
| ptratio | 12.6 | 22 |
| black | 0.32 | 396.90 |
| lstat | 1.73 | 37.97 |
| medv | 5 | 50 |

d)

> summary(Boston)

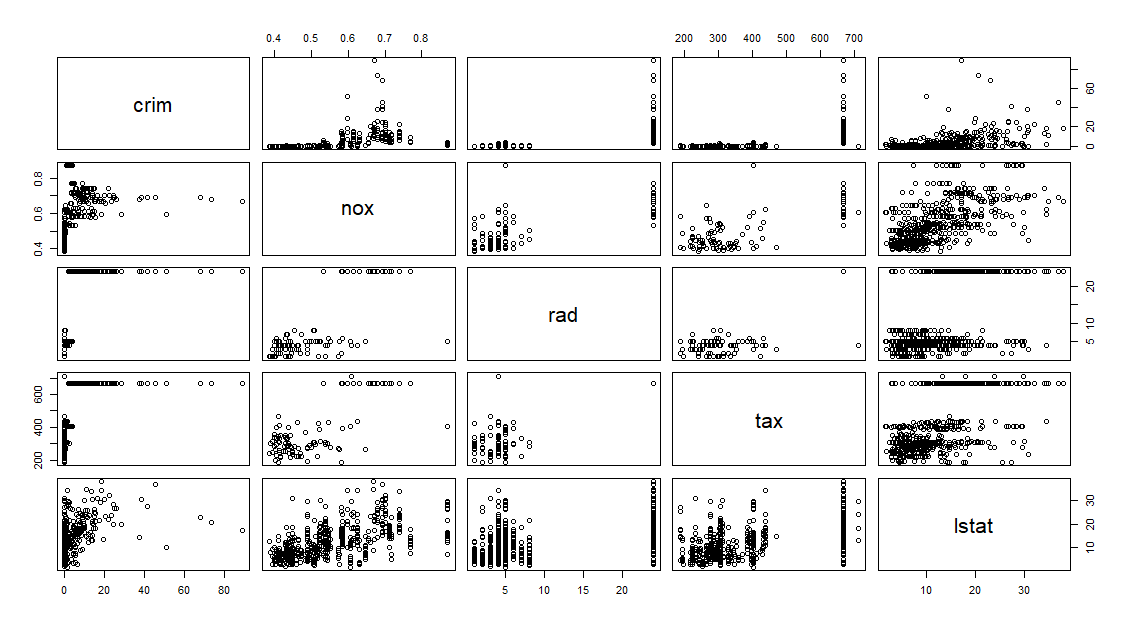
|  |  |  |
| --- | --- | --- |
| Predictors | Mean | St. Dev. |
| crim | 3.613 | 8.602 |
| zn | 11.36 | 23.322 |
| indus | 11.14 | 6.86 |
| chas | 0.069 | 0.254 |
| nox | 0.555 | 0.116 |
| rm | 6.285 | 0.703 |
| age | 68.57 | 28.149 |
| dis | 3.795 | 2.106 |
| rad | 9.549 | 8.707 |
| tax | 408.2 | 168.537 |
| ptratio | 18.46 | 2.165 |
| black | 356.67 | 91.295 |
| lstat | 12.65 | 7.141 |
| medv | 22.53 | 9.197 |

e)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Predictors | Min | Max | Mean | St. Dev. |
| crim | 0.00632 | 88.9762 | 4.190 | 9.211 |
| zn | 0 | 95 | 10.816 | 22.917 |
| indus | 0.45 | 27.74 | 11.864 | 7.111 |
| chas | 0 | 1 | 0.081 | 0.274 |
| nox | 0.385 | 0.871 | 0.5695 | 0.118 |
| rm | 3.561 | 8.78 | 6.316 | 0.741 |
| age | 6.8 | 100 | 71.187 | 27.069 |
| dis | 1.1296 | 12.1265 | 3.503 | 2.075 |
| rad | 1 | 24 | 10.474 | 9.124 |
| tax | 187 | 711 | 427.928 | 174.47 |
| ptratio | 12.6 | 22 | 18.357 | 2.222 |
| black | 0.32 | 396.9 | 352.06 | 97.489 |
| lstat | 1.73 | 37.97 | 12.863 | 7.376 |
| medv | 5 | 50 | 22.833 | 9.737 |

f)

pairs(Boston[,c(1,5,9,10,13)])



Of the four predictors with the strongest correlations with per capita crime rate, accessibililty to radial highways (rad) and full-value property-tax rate (tax) have moderately strong positive relationships with crime rate (crim).

g)

> Boston.hicrim <- subset(Boston, crim > 1)

> Boston.hitax <- subset(Boston, tax > 500)

> Boston.hiptratio <- subset(Boston, ptratio > 20)

-There were approximately 66% with lower than 1% crime rate per capita. The range for crim is (0.006, 88.976). The suburbs that appear to have high crime rates (> 1%) generally have:

+ No residential land zoned for lots over 25,000 sq ft

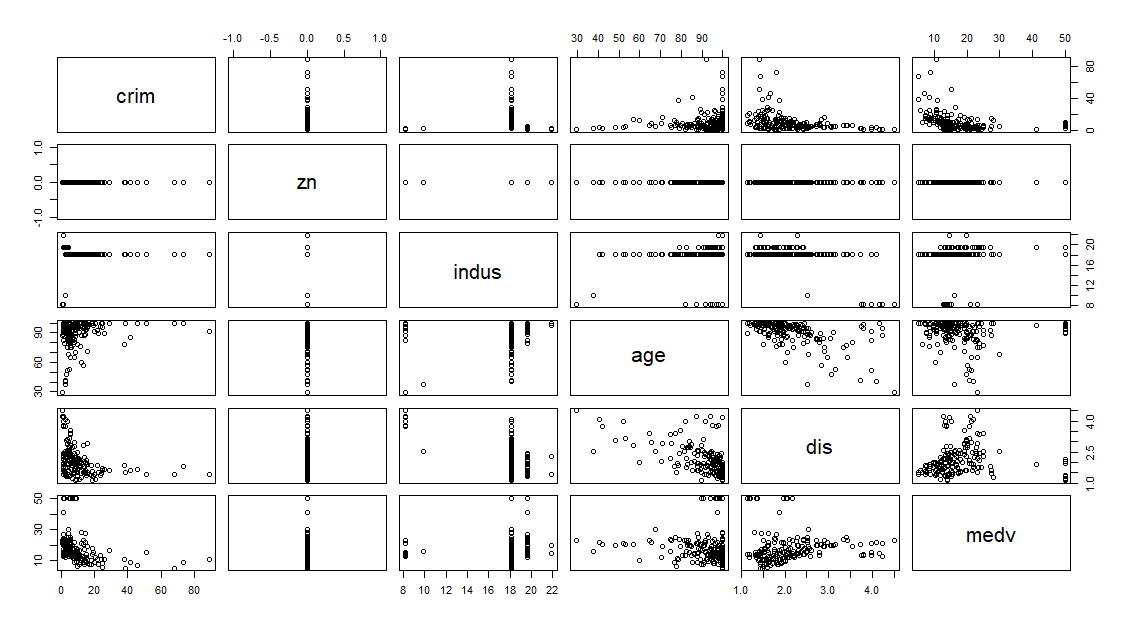
+ Less than 18% of non-retail business acres per town

+ A higher proportion of older housing

+ Tend to be closer to employment centers

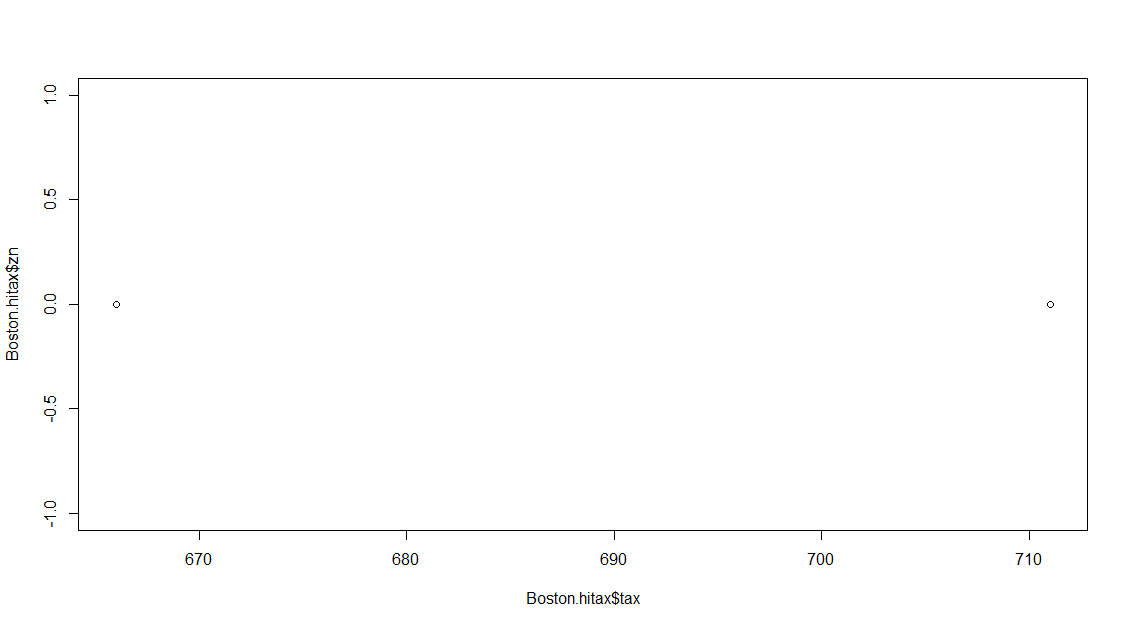
+ Lower median home values

> pairs(Boston.hicrim[,c(1,2,3,7,8,14)])



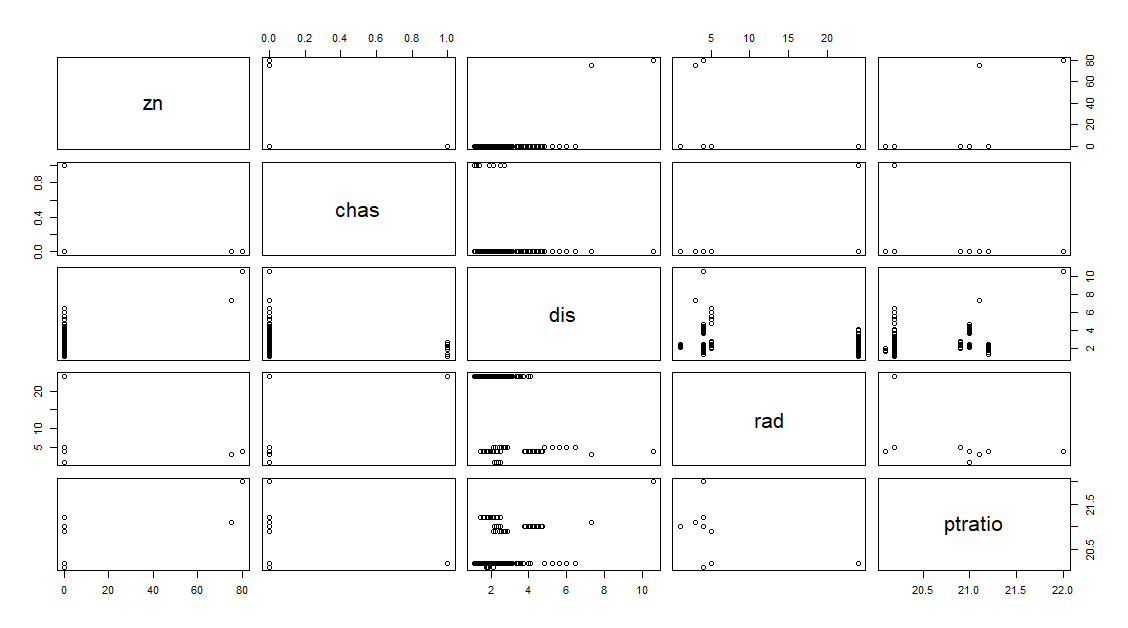
- The range of tax rates appear to form two clusters – 187 to 469 and 666 to 711. No residential land zoned for lots over 25,000 sq ft belong in the higher tax group.

> plot(Boston.hitax$tax, Boston.hitax$zn)



* The range for ptratio is (12.6, 22.0). Of the observations, there are approximately 40% having a pupil-teacher ration of 20-22.

pairs(Boston.hiptratio[,c(2,4,8,9,11)])



h)

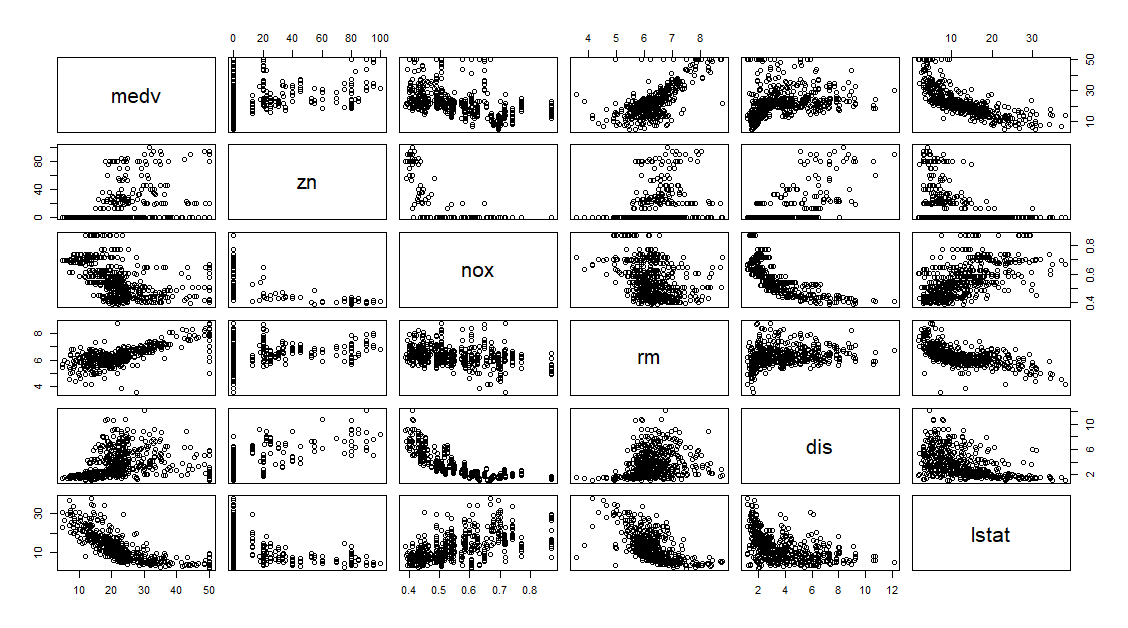
> length(Boston$chas[Boston$chas==1])

[1] 35

There are 35 suburbs in this data set that bound Charles River.

i)

pairs(Boston[c(14, 2, 5,6, 8, 13)])



We can depend on zn, nox, rm, dis, and lstat to predict medv.

+linear: medv ~ zn, medv ~ nox, medv ~ rm

+exponential: medv ~ dis, medv ~ lstat

**Exercise #6:**

a)

> names(Auto)

[1] "X" "mpg" "cylinders" "displacement"

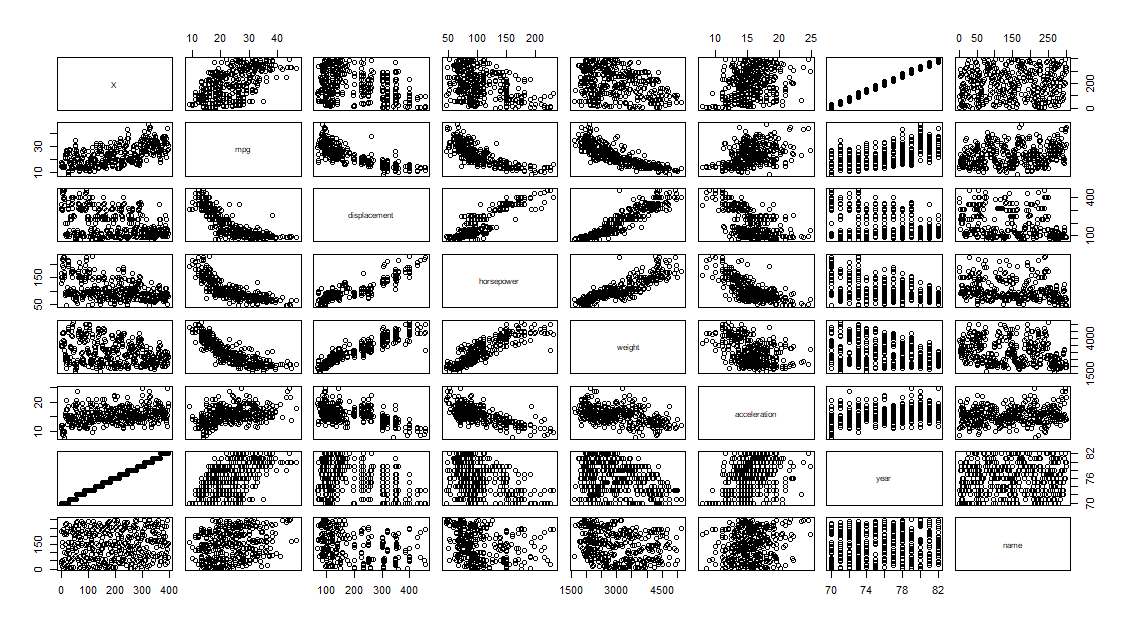
[5] "horsepower" "weight" "acceleration" "year"

[9] "origin" "name"

There are 392 rows and 10 columns.

b)

> pairs( Auto[c(1,2,4,5,6,7,8,10)])



Strongest variable relationships:

-mpg ~ displacement: it is almost negative linear. The increase of mpg leads to decrease of displacement.

- Horsepower ~ weight: it is a clear positive linear graph. The increase of the weight relates to the increase of the horsepower.

- weight ~ displacement: it is a clear positive linear graph. The increase of the weight relates to the increase of the horsepower.

- Horsepower ~ displacement ~ weight: these three factors have positive linear graphs. They increase together.

There is a negative relationship between mpg and displacement, horsepower, and weight which indicates a loss in fuel efficiency as vehicles grow larger and more powerful. However, there is a positive relationship between mpg and year which indicates an overall improvement in efficiency as time passes.

c)

Predictor associates with Horsepower:

- Horsepower ~ weight: it is a clear positive linear graph. The increase of the weight relates to the increase of the horsepower.

- Horsepower ~ displacement ~ weight: these three factors have positive linear graphs. They increase together.

d)

> range(Auto$horsepower)

[1] 46 230

> range(Auto$displacement)

[1] 68 455

> range(Auto$weight)

[1] 1613 5140

group4 <- subset(Auto, range(Auto$horsepower) == max(range(Auto$horsepower)))

group5 <- subset(Auto, range(Auto$weight) == max(range(Auto$weight)))

group6 <- subset(Auto, range(Auto$displacement)== max(range(Auto$displacement)))

> group5 <- subset(Auto, range(Auto$weight) == max(range(Auto$weight)))

> summary(group5$origin)

Min. 1st Qu. Median Mean 3rd Qu. Max.

1.000 1.000 1.000 1.617 2.000 3.000

> group6 <- subset(Auto, range(Auto$displacement)== max(range(Auto$displacement)))

> summary(group6$origin)

Min. 1st Qu. Median Mean 3rd Qu. Max.

1.000 1.000 1.000 1.617 2.000 3.000

**The origins are 1,2,3**

e)

group7 <- subset (Auto, year > 79)

> nrow(group7)

[1] 85

There are **85 models** produced at 1980 or later.