HW1: Math158

Joshua Jansen-Montoya

2022-09-06

Problem 1.1:

The dataset teengamb concerns a study of teenage gambling in Britain. Make a numerical and graphical summary of the data, commenting on any features that you find interesting. Limit the output you present to a quantity that a busy reader would find sufficient to get a basic understanding of the data.

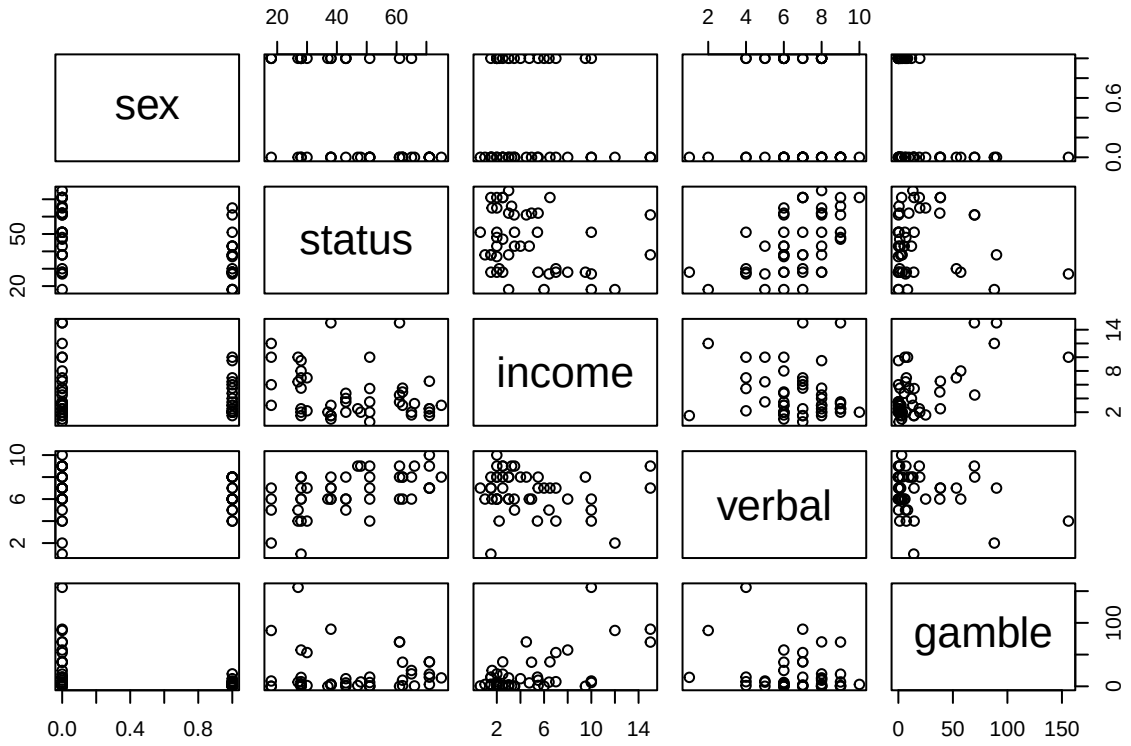
Answer 1.1:

We can complete the numerical summary with the following R-code.

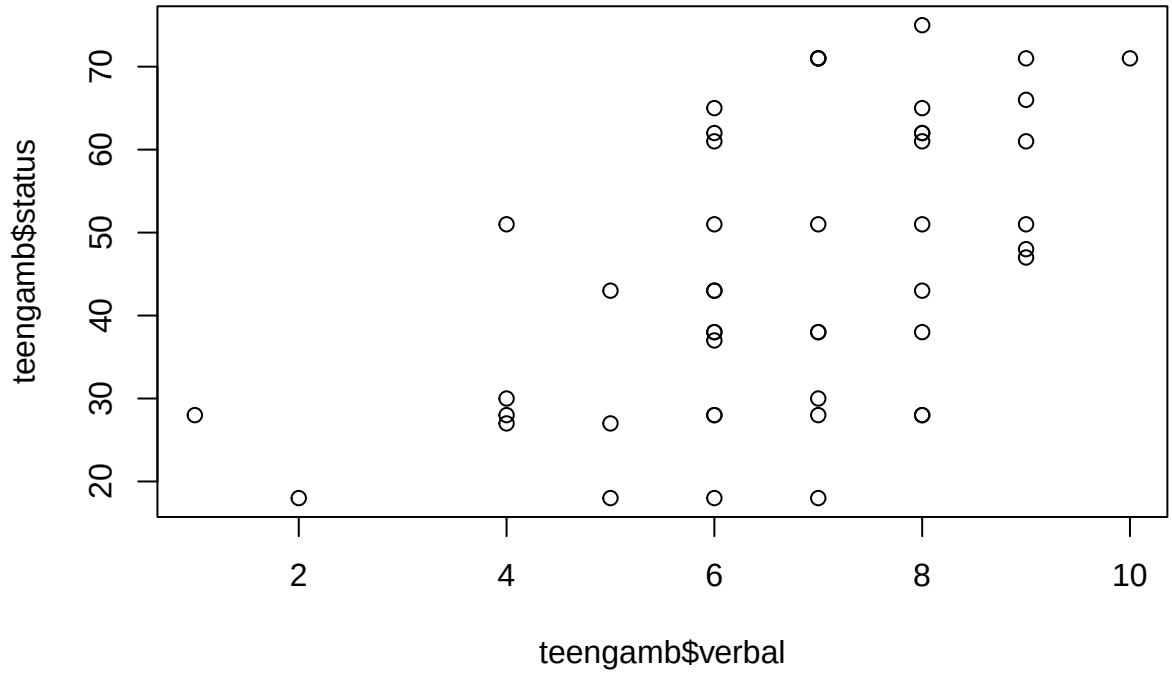
```
library(faraway)
summary(teengamb)
```

```
##       sex            status          income           verbal
##  Min.   :0.0000   Min.   :18.00   Min.   : 0.600   Min.   : 1.00
##  1st Qu.:0.0000   1st Qu.:28.00   1st Qu.: 2.000   1st Qu.: 6.00
##  Median :0.0000   Median :43.00   Median : 3.250   Median : 7.00
##  Mean   :0.4043   Mean   :45.23   Mean   : 4.642   Mean   : 6.66
##  3rd Qu.:1.0000   3rd Qu.:61.50   3rd Qu.: 6.210   3rd Qu.: 8.00
##  Max.   :1.0000   Max.   :75.00   Max.   :15.000   Max.   :10.00
##      gamble
##  Min.   :  0.0
##  1st Qu.:  1.1
##  Median :  6.0
##  Mean   : 19.3
##  3rd Qu.: 19.4
##  Max.   :156.0
```

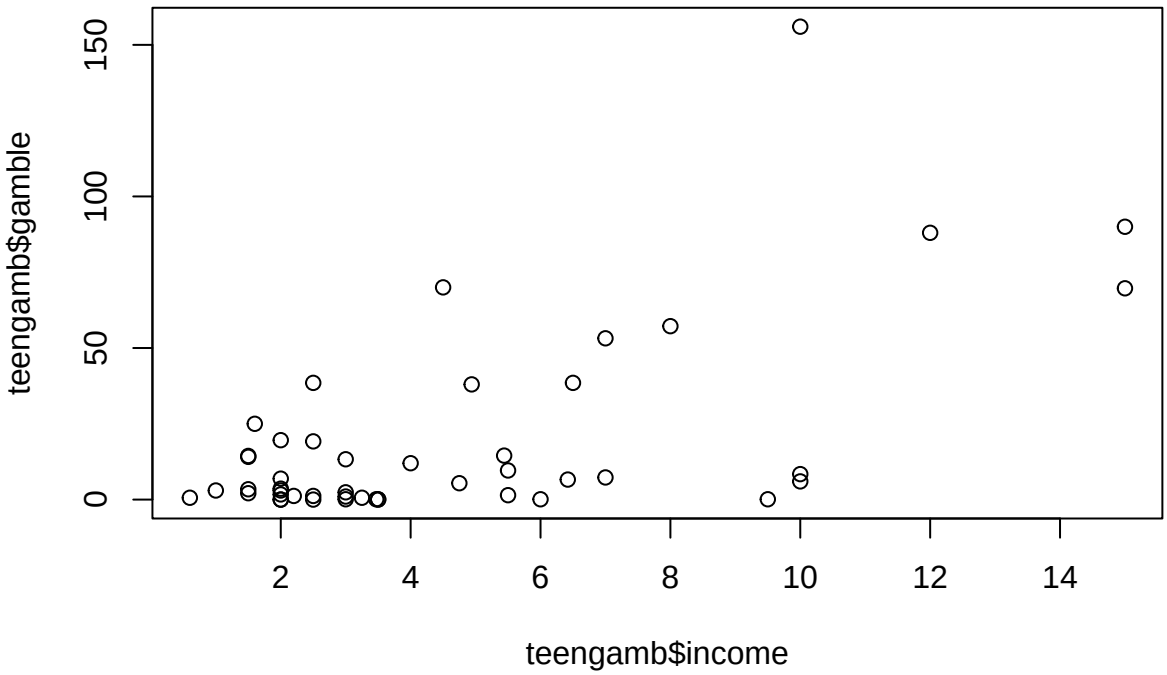
```
pairs(teengamb)
```



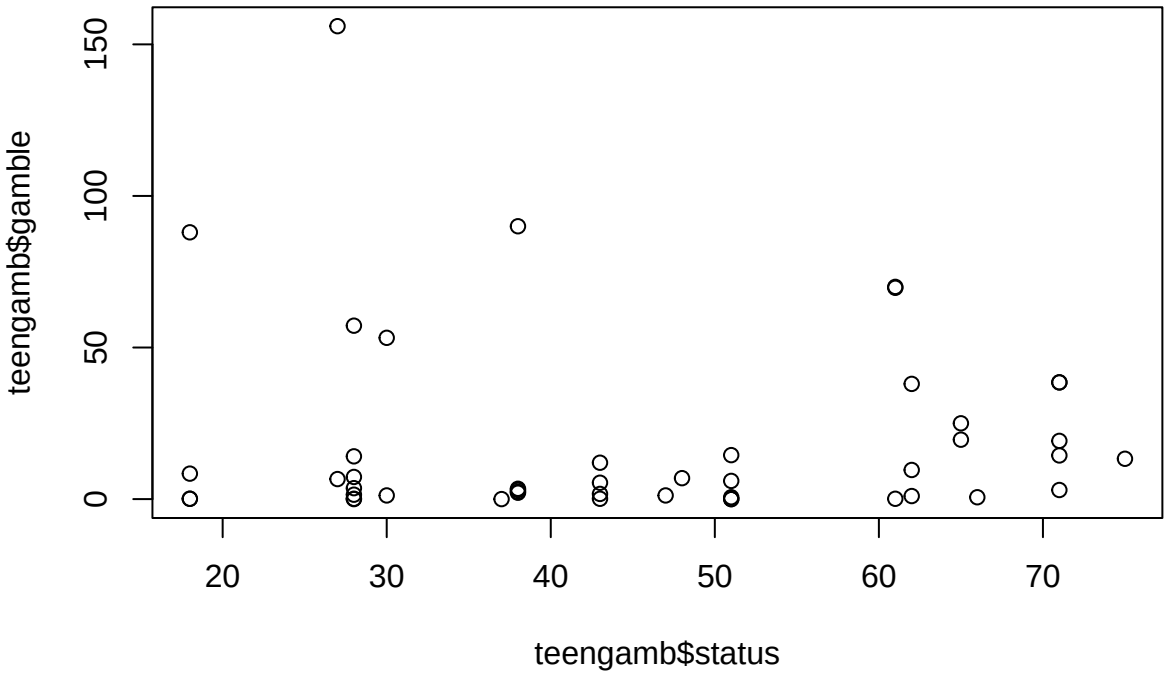

```
plot(teengamb$verbal, teengamb$status)
```




```
plot(teengamb$income, teengamb$gamble)
```




```
plot(teengamb$status, teengamb$gamble)
```



Now, looking at our graphs, we can note that there seems to be a positive linear relationship between the individuals socioeconomic status score and their verbal verbal word score. Similarly, there appears to be a positive relationship between the income that an individual has and their expenditure on gambling in pounds per year as indicated by the plot. However it is interesting to note that there does not seem to be a relationship between the socioeconomic status of an individual and their expenditure on gambling in pounds per year. We can note that in our study, our values for the pounds per week income of individuals ranged from (what seems to be) 600 pounds to 15,000 pounds per week.

Problem 1.3

The dataset prostate is from a study on 97 men with prostate cancer who were due to receive a radical prostatectomy. Make a numerical and graphical summary of the data as in the first question.

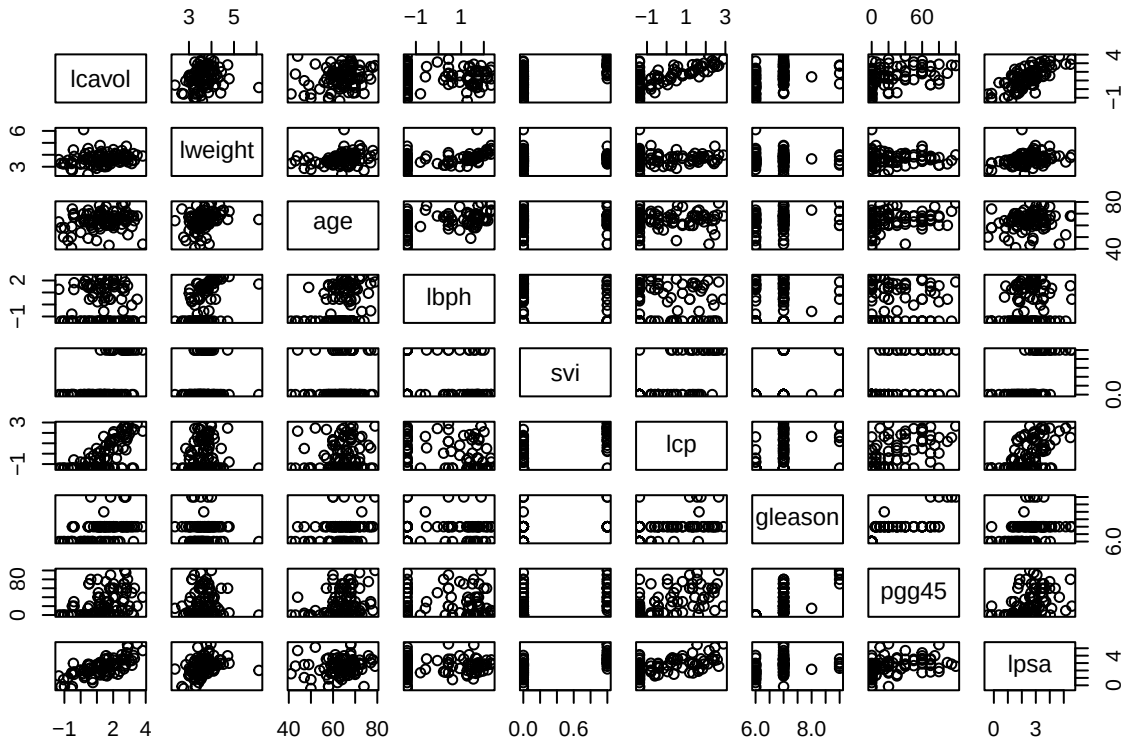
Answer 1.3

Once again, we can construct a numerical and graphical summary of the data using the following R-code.

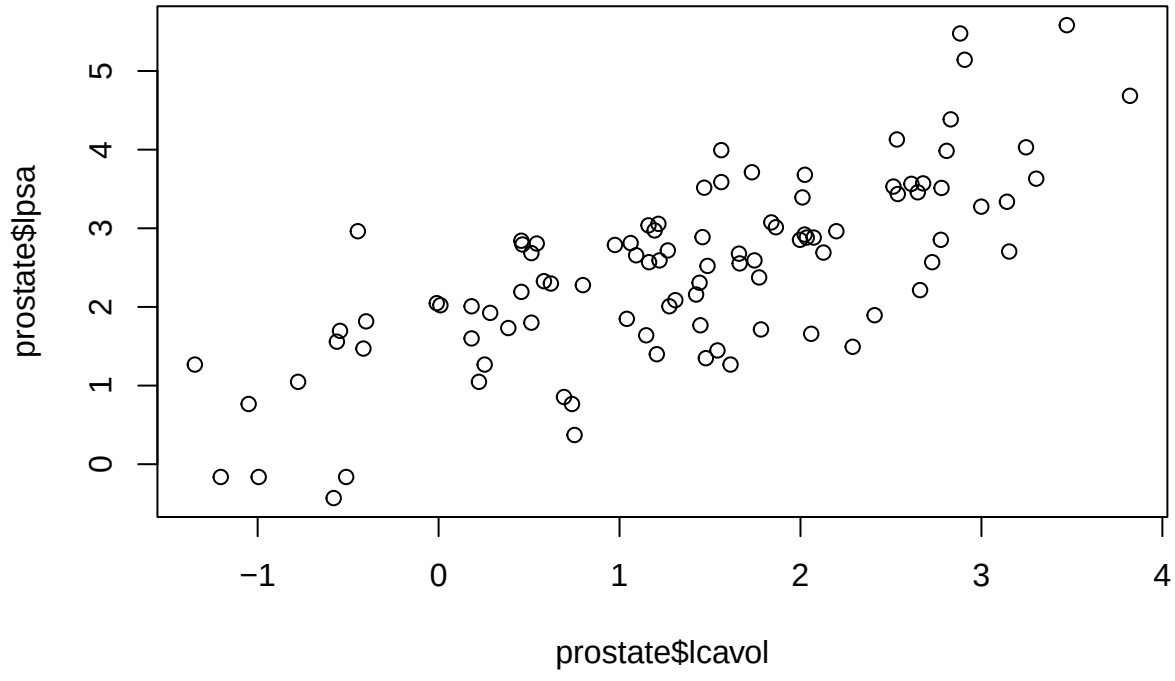
```
library(faraway)
summary(prostate)
```

```
##      lcavol           lweight           age             lbph        
##  Min.   :-1.3471   Min.   :2.375   Min.   :41.00   Min.   :-1.3863  
##  1st Qu.: 0.5128   1st Qu.:3.376   1st Qu.:60.00   1st Qu.:-1.3863  
##  Median : 1.4469   Median :3.623   Median :65.00   Median : 0.3001  
##  Mean   : 1.3500   Mean   :3.653   Mean   :63.87   Mean   : 0.1004  
##  3rd Qu.: 2.1270   3rd Qu.:3.878   3rd Qu.:68.00   3rd Qu.: 1.5581  
##  Max.   : 3.8210   Max.   :6.108   Max.   :79.00   Max.   : 2.3263  
##       svi              lcp             gleason          pgg45       
##  Min.   :0.0000   Min.   :-1.3863   Min.   :6.000   Min.   :  0.00  
##  1st Qu.:0.0000   1st Qu.:-1.3863   1st Qu.:6.000   1st Qu.:  0.00  
##  Median :0.0000   Median :-0.7985   Median :7.000   Median : 15.00  
##  Mean   :0.2165   Mean   :-0.1794   Mean   :6.753   Mean   : 24.38  
##  3rd Qu.:0.0000   3rd Qu.: 1.1786   3rd Qu.:7.000   3rd Qu.: 40.00  
##  Max.   :1.0000   Max.   : 2.9042   Max.   :9.000   Max.   :100.00  
##       lpsa        
##  Min.   :-0.4308  
##  1st Qu.: 1.7317  
##  Median : 2.5915  
##  Mean   : 2.4784  
##  3rd Qu.: 3.0564  
##  Max.   : 5.5829
```

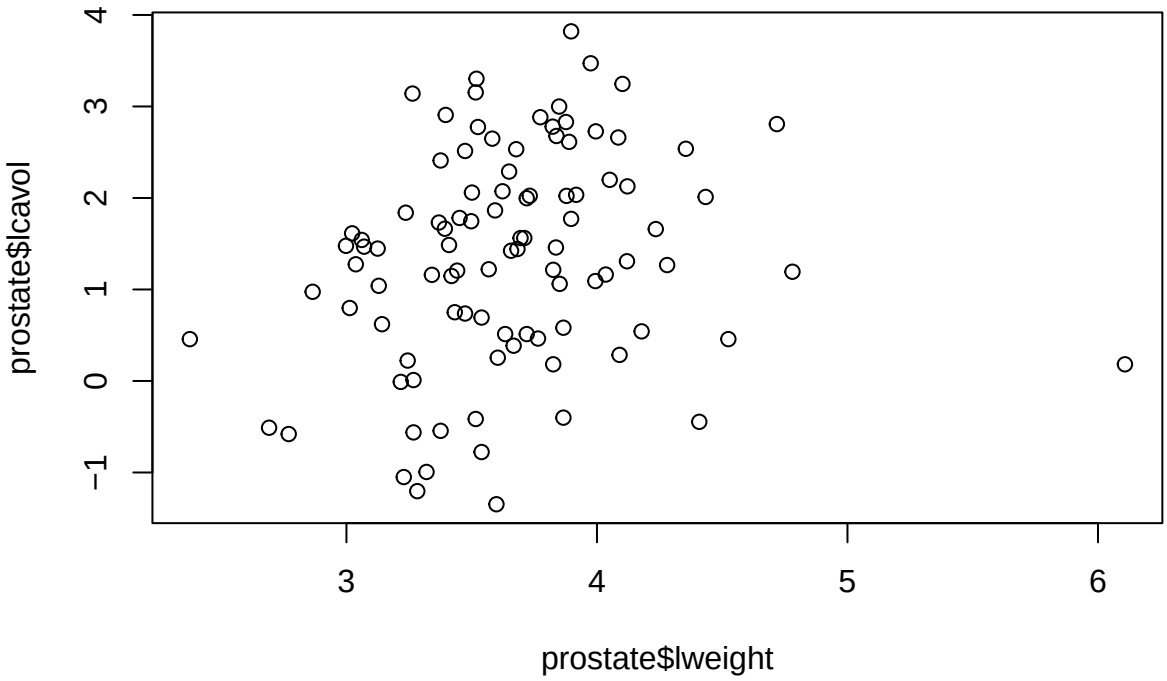
```
pairs(prostate)
```




```
plot(prostate$lcavol, prostate$lpsa)
```




```
plot(prostate$lweight, prostate$lcavol)
```



Evaluating our summaries, we can see that there appears to be a positive linear relationship between the log of prostate specific antigens and the log of the cancer volume, indicating some possible relationship between the amount of these prostate specific antigens and the volume of cancer in the prostate. There could also be a relationship between the log of the weight of an individual and the log of the cancer volume, though it is notable that there exists a point of high influence that is well outside of the spread of the rest of values.

Problem 1.4

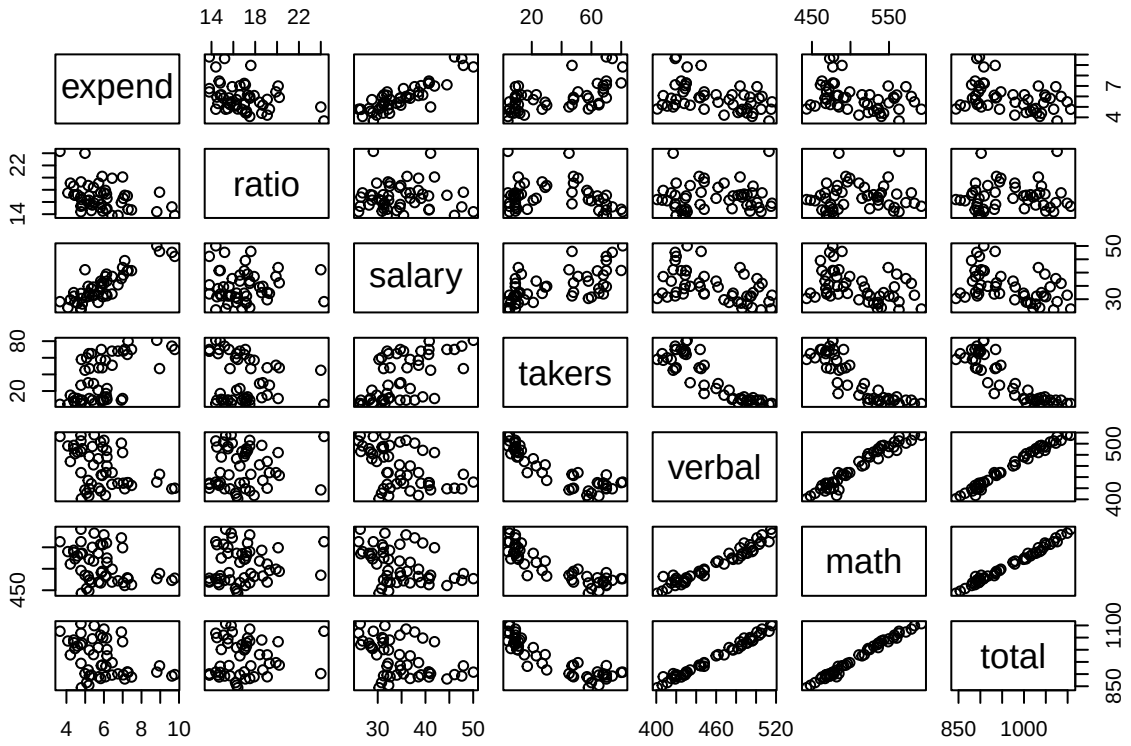
The dataset sat comes from a study entitled "Getting What You Pay For: The Debate Over Equity in Public School Expenditures." Make a numerical and graphical summary of the data as in the first question.

Answer 1.4

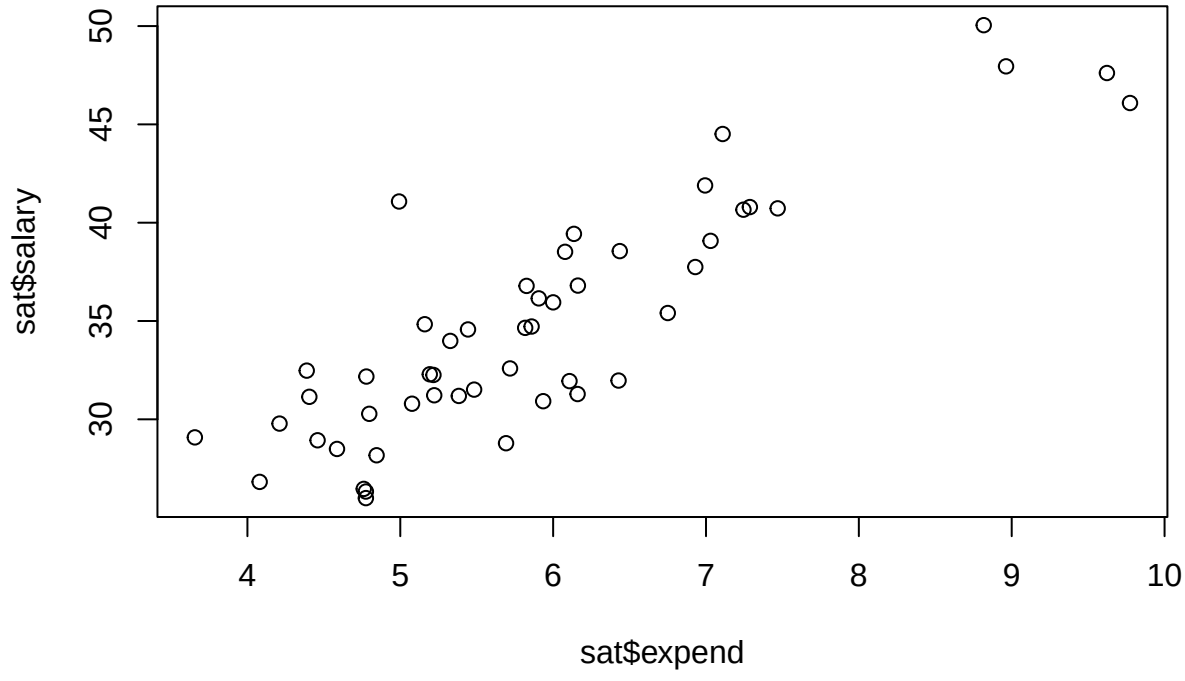
```
library(faraway)
summary(sat)
```

```
##      expend          ratio           salary          takers
##  Min.   :3.656   Min.   :13.80   Min.   :25.99   Min.   : 4.00
##  1st Qu.:4.882   1st Qu.:15.22   1st Qu.:30.98   1st Qu.: 9.00
##  Median :5.768   Median :16.60   Median :33.29   Median :28.00
##  Mean   :5.905   Mean   :16.86   Mean   :34.83   Mean   :35.24
##  3rd Qu.:6.434   3rd Qu.:17.57   3rd Qu.:38.55   3rd Qu.:63.00
##  Max.   :9.774   Max.   :24.30   Max.   :50.05   Max.   :81.00
##      verbal           math           total
##  Min.   :401.0   Min.   :443.0   Min.   : 844.0
##  1st Qu.:427.2   1st Qu.:474.8   1st Qu.: 897.2
##  Median :448.0   Median :497.5   Median : 945.5
##  Mean   :457.1   Mean   :508.8   Mean   : 965.9
##  3rd Qu.:490.2   3rd Qu.:539.5   3rd Qu.:1032.0
##  Max.   :516.0   Max.   :592.0   Max.   :1107.0
```

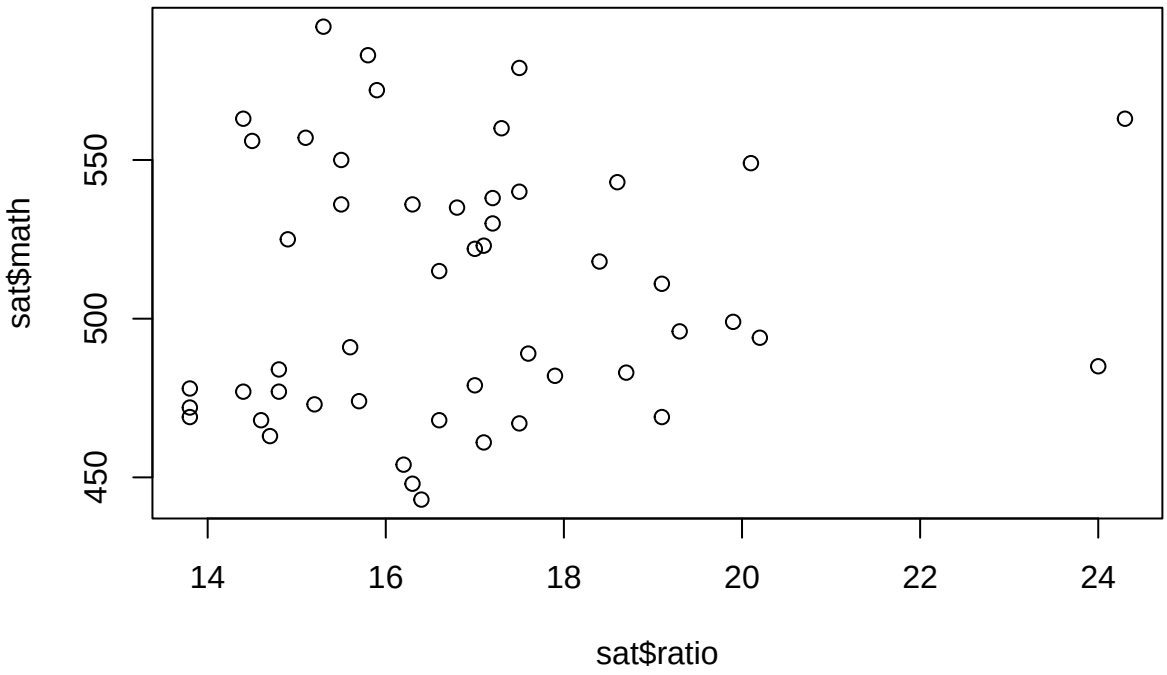
```
pairs(sat)
```




```
plot(sat$expend, sat$salary)
```




```
plot(sat$ratio, sat$math)
```



Evaluating our data, we can note that there is a positive linear relationship between the expendature per student in daily average attendance and the estimated salary of the teachers, which makes sense as if the school is spending more per student, part of that is spent on the teachers. Similarly, it is noteworthy that the relation between the average verbal SAT scores and the average math SAT scores is almost perfectly linear, which makes sense as if an individual scores well in one category, they could likely score well in the other. Surprisingly, there does not seem to be a significant relationship between the average math SAT scores and the average pupil/teacher ratio in public elementary and secondary schools, which I would expect to be related, as a hallmark of quality instruction seems to be a reduced student to teacher ratio and higher SAT scores.

Problem 2.1

The dataset teengamb concerns a study of teenage gambling in Britain. Fit a regression model with the expenditure on gambling as the response and the sex, status, income and verbal score as predictors. Present the output. 1. What percentage of variation in the response is explained by these predictors? 2. Which observation has the largest (positive) residual? Give the case number. 3. Compute the mean and median of the residuals. 4. Compute the correlation of the residuals with the fitted values. 5. Compute the correlation of the residuals with the income. 6. For all other predictors held constant, what would be the difference in predicted expenditure on gambling for a male compared to a female?

Answer 2.1

First, we can construct our linear regression model with the following R-code,

```
    library(faraway)
    reg <- lm(gamble ~ sex + status + income + verbal, data = teengamb)
    summary(reg)
```

```
##
## Call:
## lm(formula = gamble ~ sex + status + income + verbal, data = teengamb)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -51.082 -11.320  -1.451   9.452  94.252
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  22.55565   17.19680   1.312   0.1968
## sex         -22.11833    8.21111  -2.694   0.0101 *
## status        0.05223    0.28111   0.186   0.8535
## income        4.96198    1.02539   4.839 1.79e-05 ***
## verbal       -2.95949    2.17215  -1.362   0.1803
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.69 on 42 degrees of freedom
## Multiple R-squared:  0.5267, Adjusted R-squared:  0.4816
## F-statistic: 11.69 on 4 and 42 DF,  p-value: 1.815e-06
```

```
cor(fitted(reg), residuals(reg))
```

```
## [1] -1.070659e-16
```

1) We can see from our summary that the R^2 indicates that 53.7% of the variation is explained by the predicators.
2) The case that has the largest positive residual is row 24 in our data frame.
3) Using R, we can see that the mean of the residual is -3.065293e-17 and that the median is -1.451392.
4) Using R, we can calculate that the correlation of the residuals with the fitted values is -1.070659e-16.
5) Using R, we can calculate that the correlation of the residuals with the income is 0.857142.
6) For all other predictors held constant, the difference in predicted expenditure on gambling for a female compared to a male would be 22.12 dollars less.

Problem 2.4

The dataset prostate comes from a study on 97 men with prostate cancer who were due to receive a radical prostatectomy. Fit a model with lpsa as the response and lcavol as the predictor. Record the residual standard error and the R^2. Now add lweight, svi, lbph, age, lcp, pgg45 and gleason to the model one at a time. For each model record the residual standard error and the R^2. Plot the trends in these two statistics.

Answer 2.4

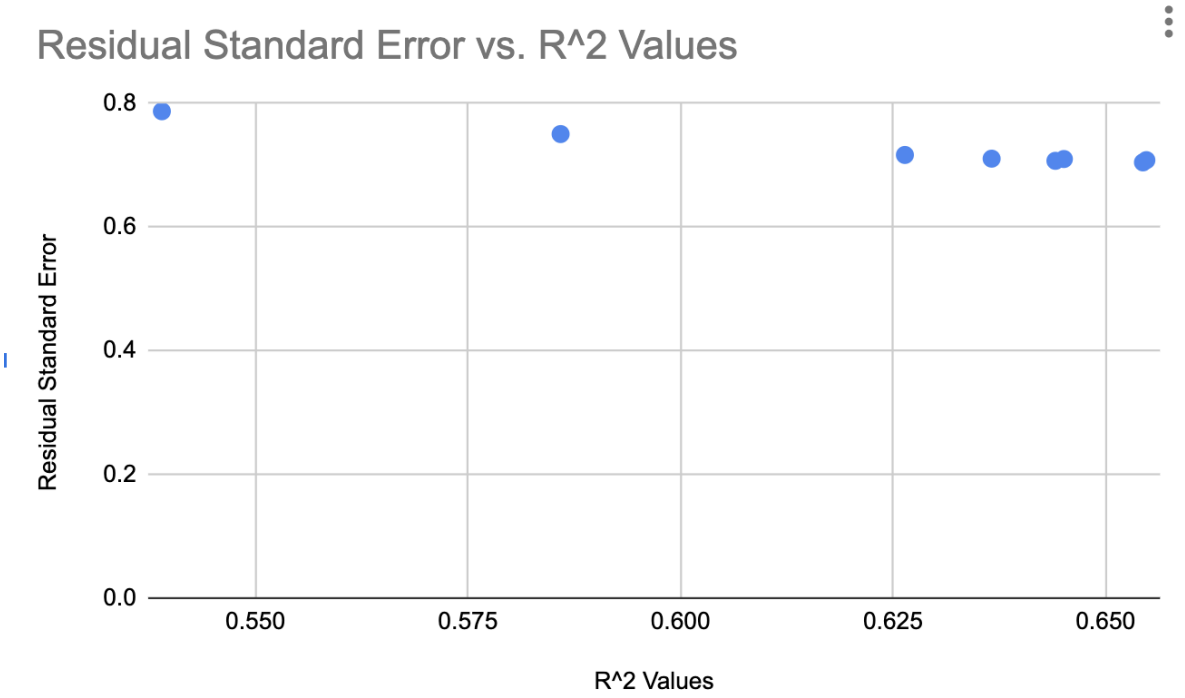
For the sake of ease, we will first constuct a list where we record the different residual values and R^2 values for each additional and then attatch said plot at the end.

```
library(faraway)
#summary(prostate)
reg <- lm(lpsa ~ lcavol + lweight + svi + lbph + age + lcp + pgg45+ gleason, data = prostate)
    summary(reg)
```

```
##
## Call:
## lm(formula = lpsa ~ lcavol + lweight + svi + lbph + age + lcp +
##     pgg45 + gleason, data = prostate)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7331 -0.3713 -0.0170  0.4141  1.6381
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)  0.669337   1.296387   0.516  0.60693
## lcavol       0.587022   0.087920   6.677 2.11e-09 ***
## lweight      0.454467   0.170012   2.673  0.00896 **
## svi          0.766157   0.244309   3.136  0.00233 **
## lbph         0.107054   0.058449   1.832  0.07040 .
## age         -0.019637   0.011173  -1.758  0.08229 .
## lcp         -0.105474   0.091013  -1.159  0.24964
## pgg45        0.004525   0.004421   1.024  0.30886
## gleason      0.045142   0.157465   0.287  0.77503
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7084 on 88 degrees of freedom
## Multiple R-squared:  0.6548, Adjusted R-squared:  0.6234
## F-statistic: 20.86 on 8 and 88 DF,  p-value: < 2.2e-16
```

1) lcavol: $R^2 = 0.5394$, residual standard error: 0.7875
2) lcavol + lweight: $R^2 = 0.5859$, residual standard error: 0.7506
3) lcavol + lweight + svi: $R^2 = 0.6264$, residual standard error: 0.7168
4) lcavol + lweight + svi + lbph: $R^2 = 0.6366$, residual standard error: 0.7108
5) lcavol + lweight + svi + lbph + age: $R^2 = 0.6441$, residual standard error: 0.7073
6) lcavol + lweight + svi + lbph + age + lcp:$R^2 = 0.6451$, residual standard error: 0.7102
7) lcavol + lweight + svi + lbph + age + lcp + pgg45:$R^2 = 0.6544$, residual standard error: 0.7048 lcavol + lweight + svi + lbph + age + lcp + pgg45 + gleason:$R^2 = 0.6548$, residual standard error: 0.7084 Now, we can construct the following plot of the two statistics as follows,



Looking at our plot, we can see that as our R^2 increases, there is a reduction in the Residual Standard Error.