Mini Project #6
David McCormick (DTM190000)
I am in a solo group.

Section 1

1. Part 1
   a. In order to build the model, we first create a correlation matrix in order to see what variables are correlated to PSA level. Cancer volume has the highest correlation with a value of .624.

	subject	psa	cancervol	weight	age	benpros	vesinv	capspen	gleason
subject	1.000	0.603	0.621	0.114	0.197	0.165	0.567	0.477	0.538
psa	0.603	1.000	0.624	0.026	0.017	-0.016	0.529	0.551	0.430
cancervol	0.621	0.624	1.000	0.005	0.039	-0.133	0.582	0.693	0.481
weight	0.114	0.026	0.005	1.000	0.164	0.322	-0.002	0.002	-0.024
age	0.197	0.017	0.039	0.164	1.000	0.366	0.118	0.100	0.226
benpros	0.165	-0.016	-0.133	0.322	0.366	1.000	-0.120	-0.083	0.027
vesinv	0.567	0.529	0.582	-0.002	0.118	-0.120	1.000	0.680	0.429
capspen	0.477	0.551	0.693	0.002	0.100	-0.083	0.680	1.000	0.462
gleason	0.538	0.430	0.481	-0.024	0.226	0.027	0.429	0.462	1.000

Afterwards, we plot each predictor variable against the PSA level and find their correlation values:

We begin creating the models, first by taking into account all of the predictors. Initially, only two of the predictors–cancer volume and seminal vesicle invasion–are statistically significant.

```
Response: psa
           Df Sum Sq Mean Sq F value    Pr(>F)
cancervol   1  62202   62202 63.8370 4.279e-12 ***
weight      1     85      85  0.0869  0.768857
benpros     1    638     638  0.6546  0.420613
vesinv      1   6861    6861  7.0415  0.009415 **
capspen     1    869     869  0.8921  0.347426
gleason     1   1321    1321  1.3557  0.247359
Residuals  90  87695     974
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

If we take the natural logarithm of PSA, then we get more significant predictors.

```
Response: log(psa)
           Df Sum Sq Mean Sq F value    Pr(>F)
cancervol   1 55.164  55.164 94.5479 1.073e-15 ***
weight      1  1.790   1.790  3.0682 0.0832452 .
benpros     1  6.219   6.219 10.6591 0.0015499 **
vesinv      1  7.308   7.308 12.5253 0.0006377 ***
capspen     1  0.141   0.141  0.2424 0.6237053
gleason     1  4.637   4.637  7.9467 0.0059236 **
Residuals  90 52.510   0.583
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Afterwards, we begin the backward elimination method with a significance level of .05 in order to get the most correlated predictor variables. Capsular

penetration is the first predictor to be eliminated.

```
Response: log(psa)
          Df Sum Sq Mean Sq F value    Pr(>F)
cancervol  1 55.164  55.164 94.8970 8.858e-16 ***
weight     1  1.790   1.790  3.0795  0.082650 .
benpros    1  6.219   6.219 10.6984  0.001515 **
vesinv     1  7.308   7.308 12.5715  0.000621 ***
gleason    1  4.390   4.390  7.5517  0.007229 **
Residuals 91 52.898   0.581
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Next, we remove weight as a predictor.

```
Response: log(psa)
          Df Sum Sq Mean Sq F value    Pr(>F)
cancervol  1 55.164  55.164 95.3440 7.145e-16 ***
benpros    1  7.803   7.803 13.4873 0.0004030 ***
vesinv     1  7.334   7.334 12.6758 0.0005886 ***
gleason    1  4.239   4.239  7.3264 0.0080997 **
Residuals 92 53.229   0.579
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Now, we can predict log(psa) with strong predictor variables. Our final model has an adjusted R^2 of .5653.

```
lm(formula = log(psa) ~ cancervol + benpros + vesinv + gleason)

Residuals:
     Min       1Q   Median       3Q      Max
-1.88531 -0.50276  0.09885  0.53687  1.56621

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.65013    0.80999  -0.803 0.424253
cancervol    0.06488    0.01285   5.051 2.22e-06 ***
benpros      0.09136    0.02606   3.506 0.000705 ***
vesinv       0.68421    0.23640   2.894 0.004746 **
gleason      0.33376    0.12331   2.707 0.008100 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7606 on 92 degrees of freedom
Multiple R-squared:  0.5834,    Adjusted R-squared:  0.5653
F-statistic: 32.21 on 4 and 92 DF,  p-value: < 2.2e-16
```

Our model is as follows:

$$log(psa) = -.65013 + .06488x1 + .09136x2 + .68421x3 + .33376x4$$

Next, we calculate the predicted value for cancer volume of 6.999, benign prostatic hyperplasia of 2.535, 0 for seminal vesicle invasion which represents an absence, and 6.876 for

gleason score. Our predicted value if 2.331, but we need to raise this to the power of e in order to get the PSA, instead of log(PSA). Our final value for PSA is 10.2835.

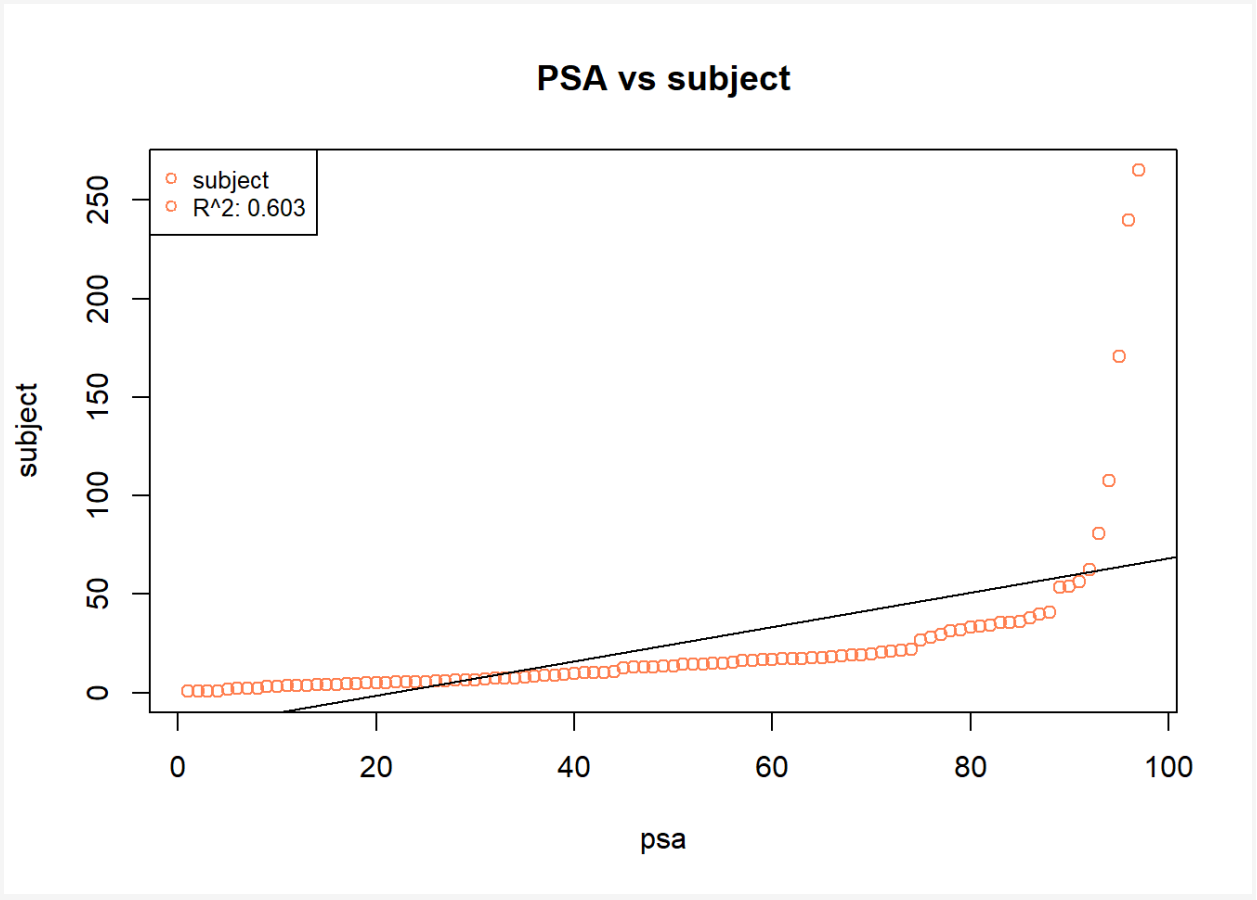
Section 2

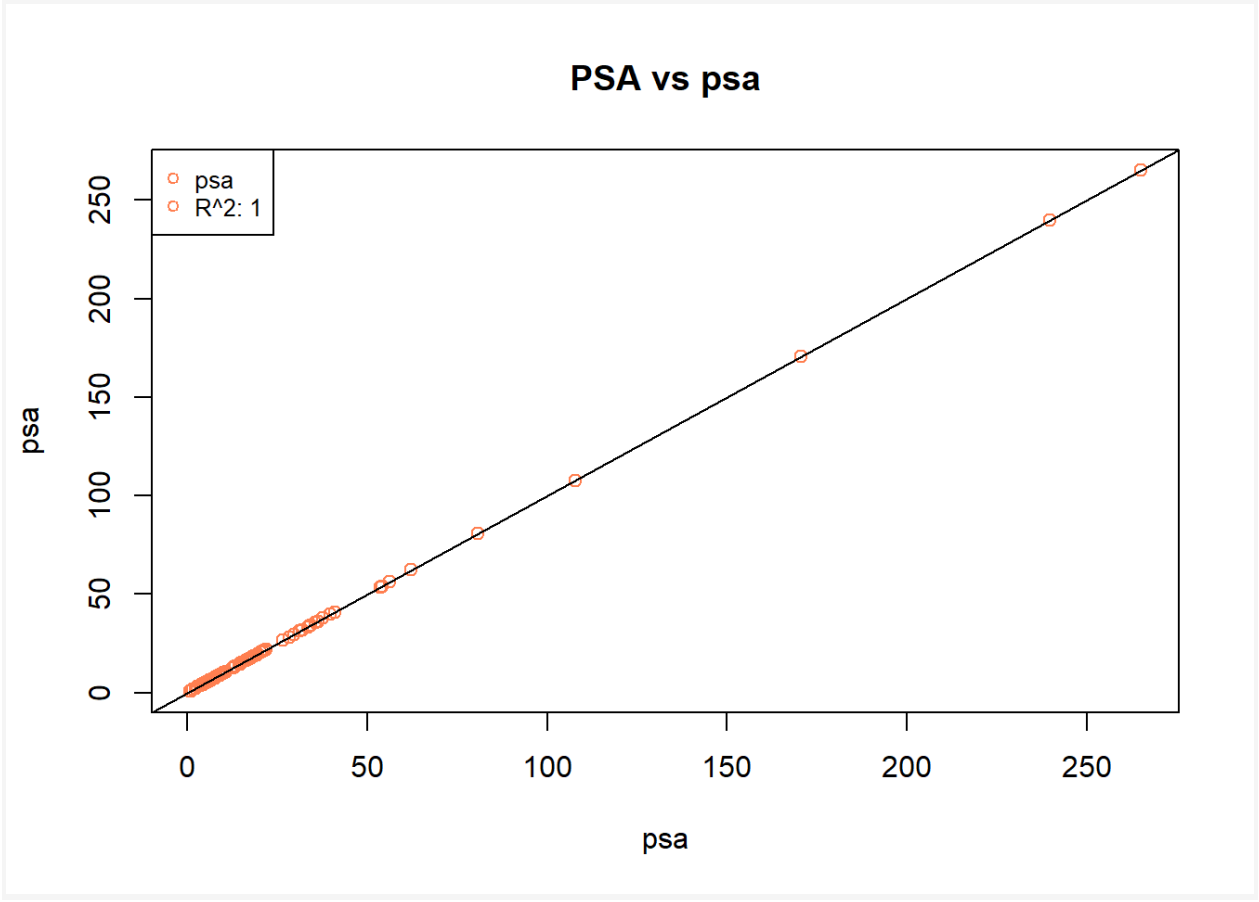
project6.R

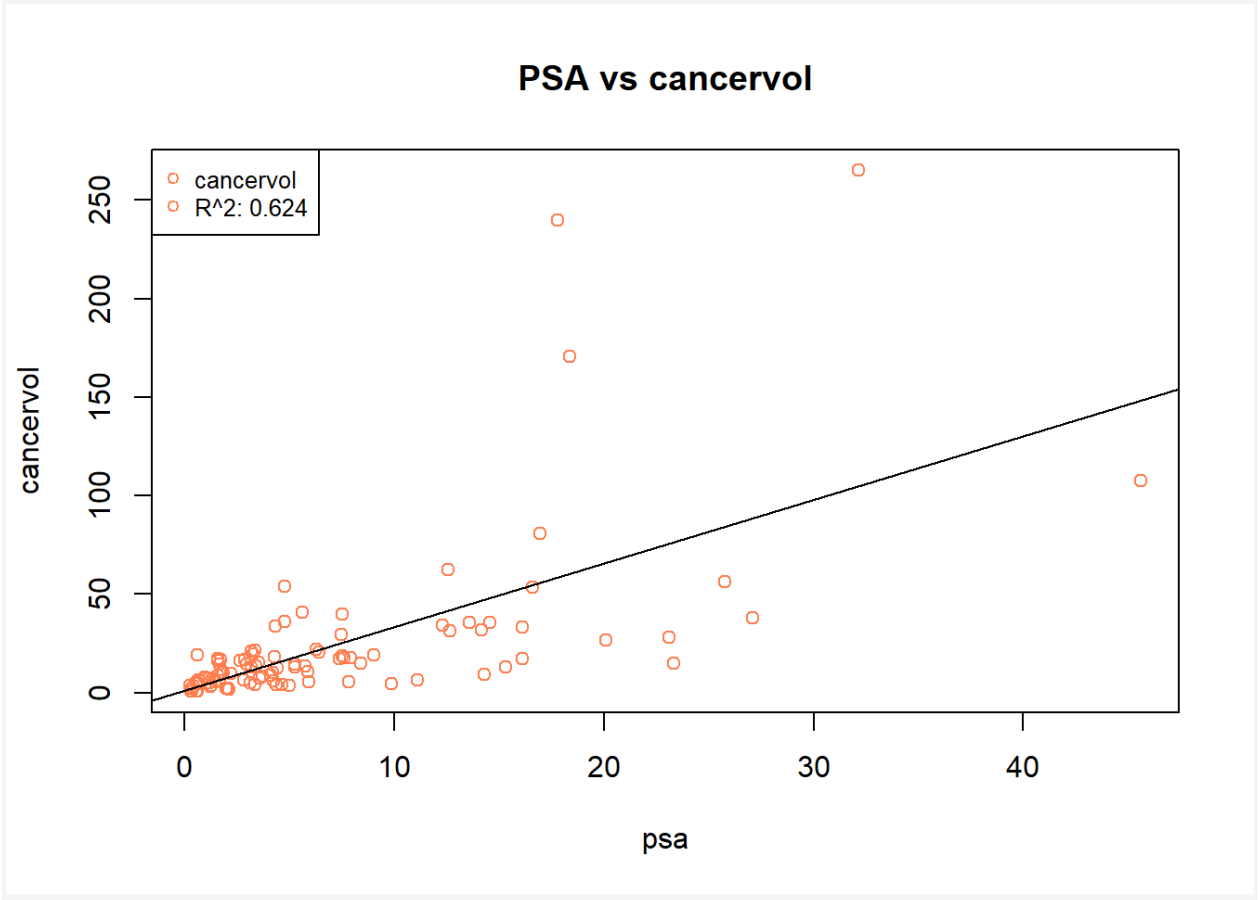
David

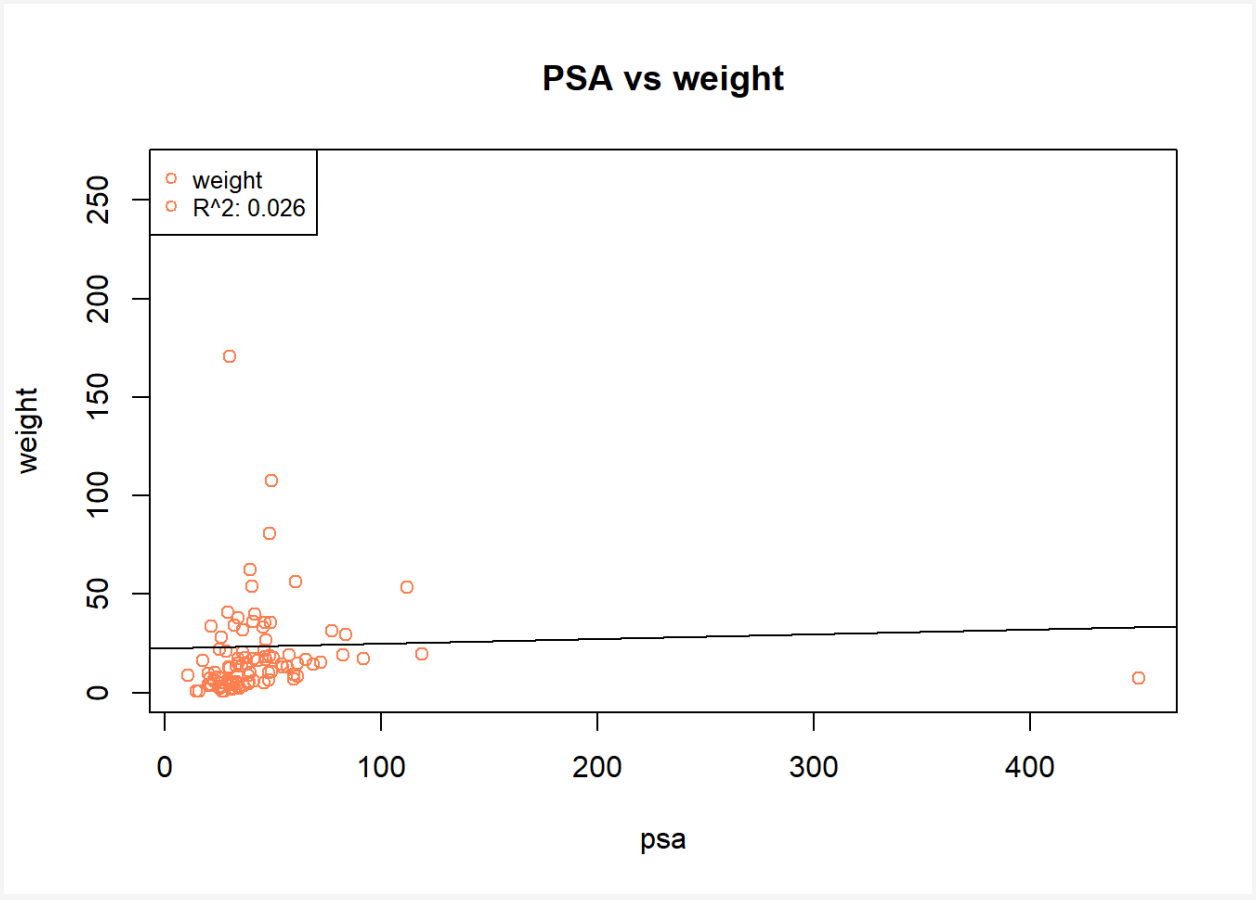
2021-12-09

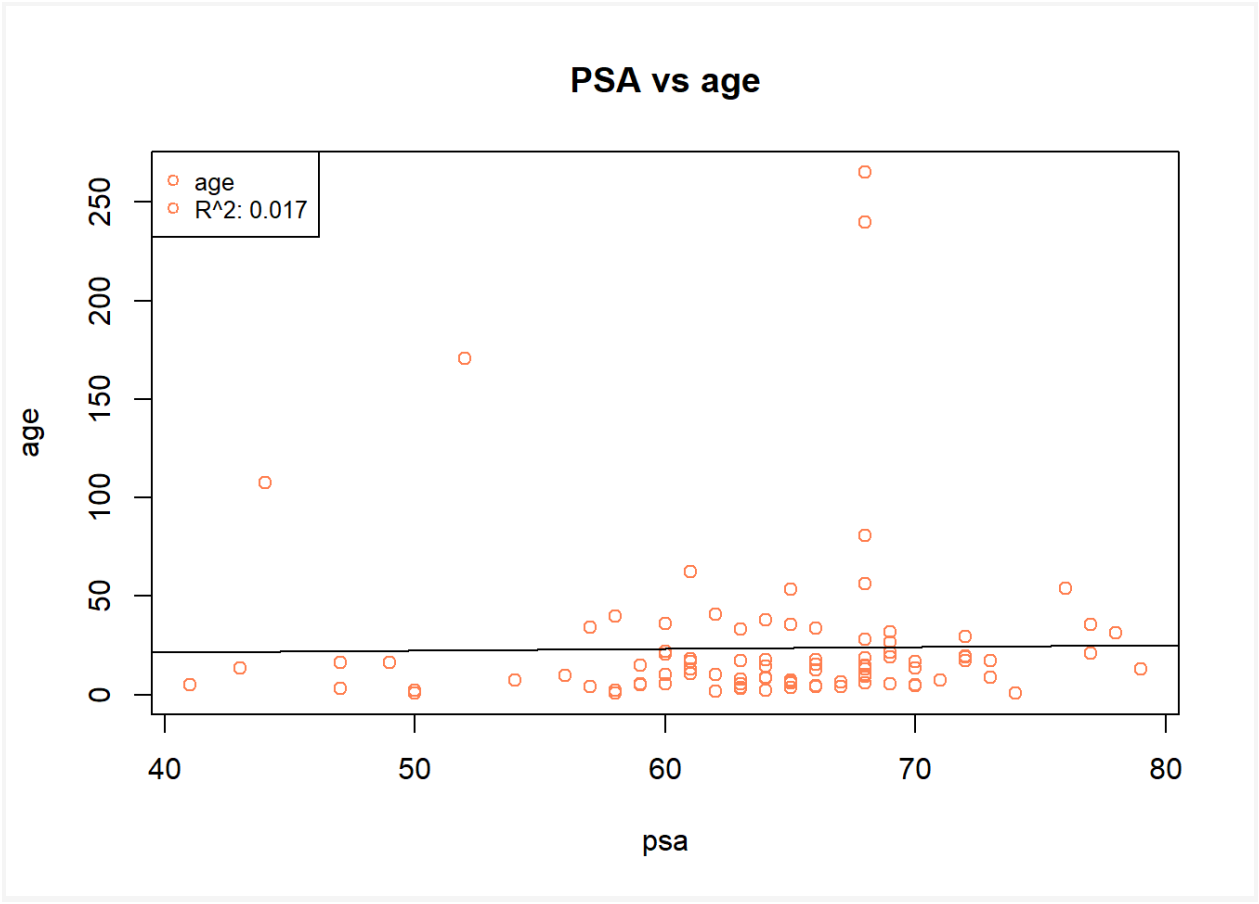
```
data = read.csv("C:/Users/David/Desktop/prostate_cancer.csv")

correlationMatrix = round(cor(data), 3)
print(correlationMatrix)

##           subject    psa cancervol weight   age benpros vesinv capspen gleason
## subject     1.000  0.603     0.621  0.114 0.197   0.165  0.567   0.477   0.538
## psa         0.603  1.000     0.624  0.026 0.017  -0.016  0.529   0.551   0.430
## cancervol   0.621  0.624     1.000  0.005 0.039  -0.133  0.582   0.693   0.481
## weight      0.114  0.026     0.005  1.000 0.164   0.322 -0.002   0.002  -0.024
## age         0.197  0.017     0.039  0.164 1.000   0.366  0.118   0.100   0.226
## benpros     0.165 -0.016    -0.133  0.322 0.366   1.000 -0.120  -0.083   0.027
## vesinv      0.567  0.529     0.582 -0.002 0.118  -0.120  1.000   0.680   0.429
## capspen     0.477  0.551     0.693  0.002 0.100  -0.083  0.680   1.000   0.462
## gleason     0.538  0.430     0.481 -0.024 0.226   0.027  0.429   0.462   1.000

subject = data$subject
psa = data$psa
cancervol = data$cancervol
weight = data$weight
age = data$age
benpros = data$benpros
vesinv = data$vesinv
capspen = data$capspen
gleason = data$gleason

for (iter in seq(1,length(colnames(data)))) {
  plot(data[,iter], psa, xlab="psa", ylab=colnames(data)[iter], main=paste("PSA vs", colnames(data)[iter]),
col="coral")
  #points(n_arr[(1):(4)], bResN[(4 * iter-3): (4 * iter)], col="coral")
  abline(lm(psa~data[,iter]))
  legend("topleft", legend=c((colnames(data)[iter]), paste("R^2:", round(cor(psa, data[,iter]),3))),
col="coral", pch=1:1,cex=0.8)
}
```

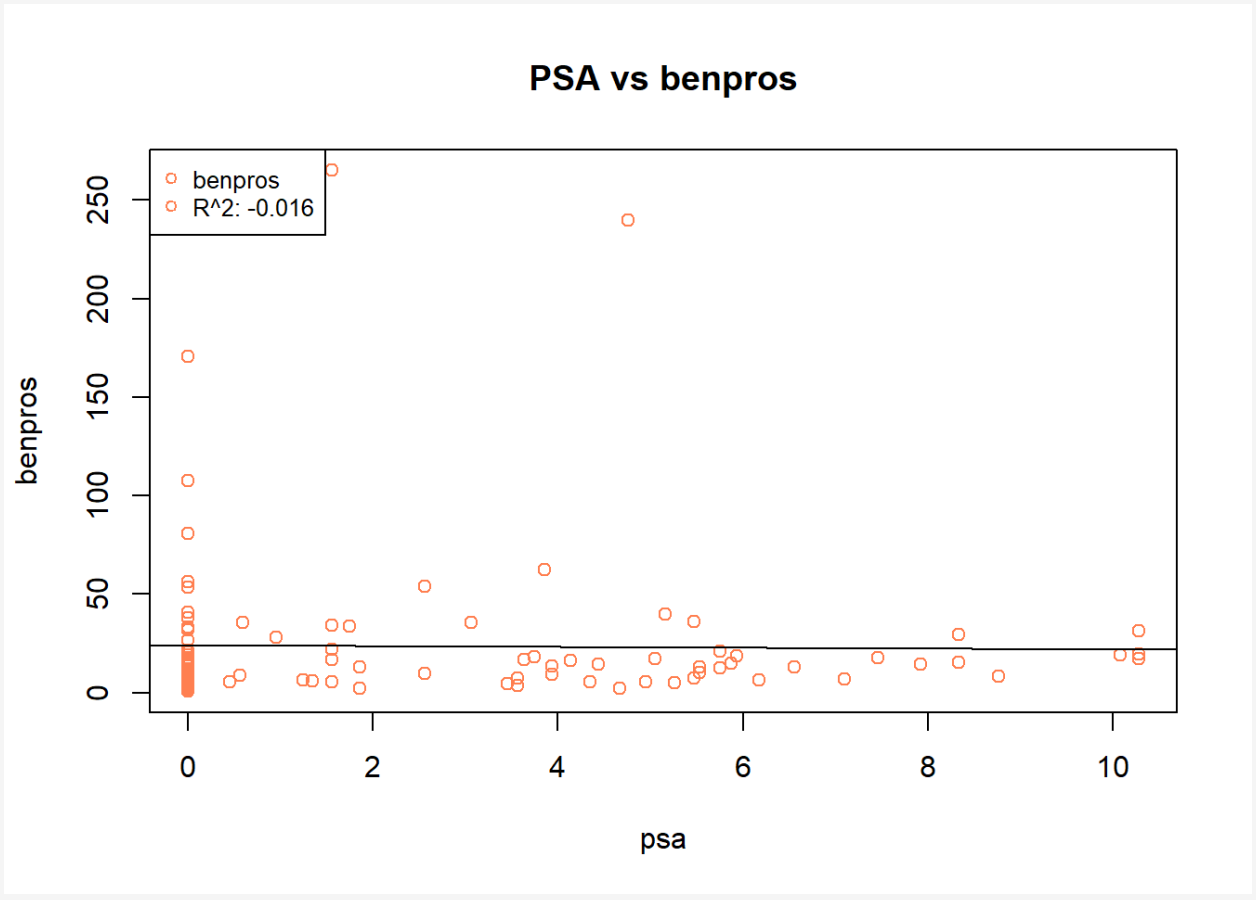

PSA vs subject
subject
R^2: 0.603
subject
0
50
100
150
200
250
0
20
40
60
80
100
psa

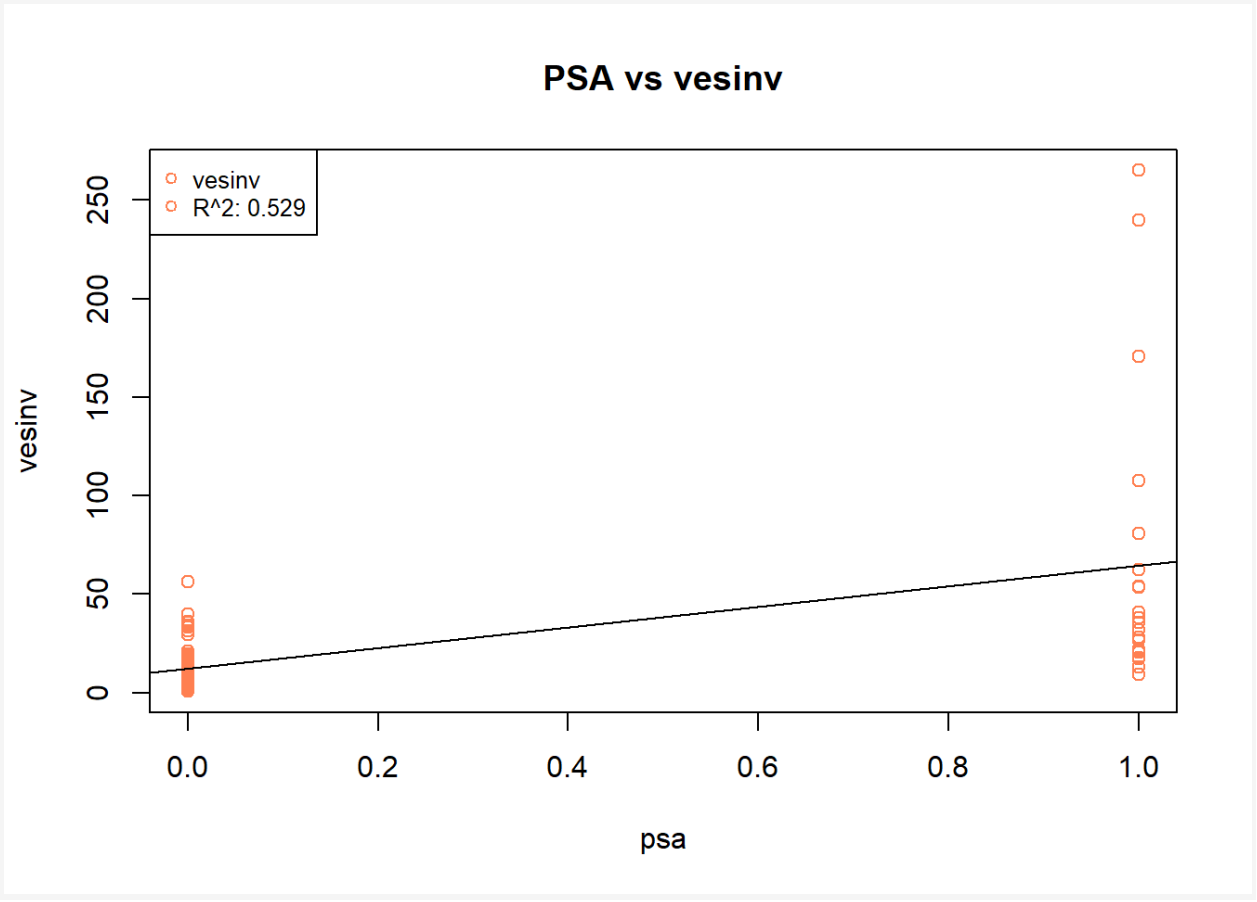

PSA vs psa
psa
R^2: 1
psa
0
50
100
150
200
250
0
50
100
150
200
250
psa

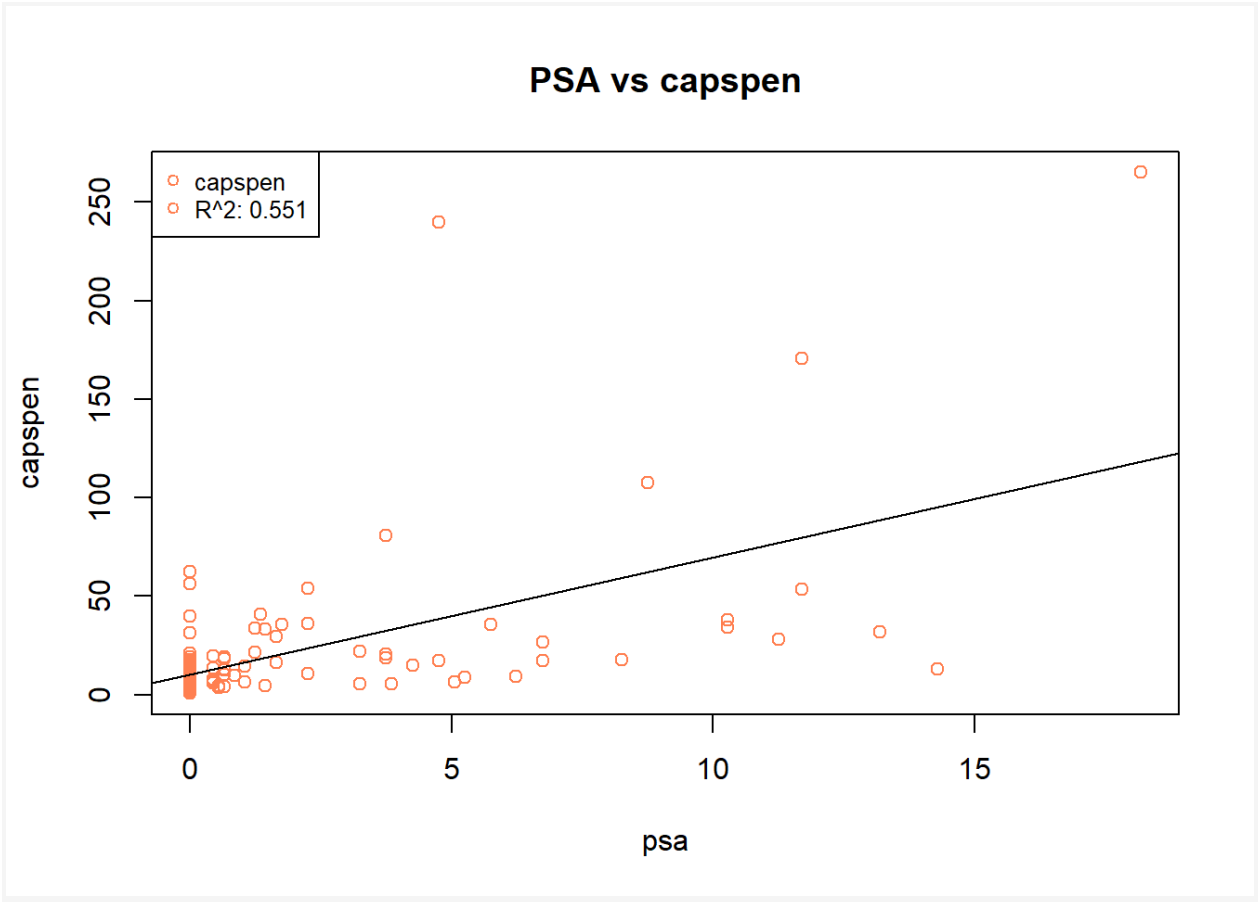

PSA vs cancervol
cancervol
R^2: 0.624
cancervol
0
50
100
150
200
250
0
10
20
30
40
psa

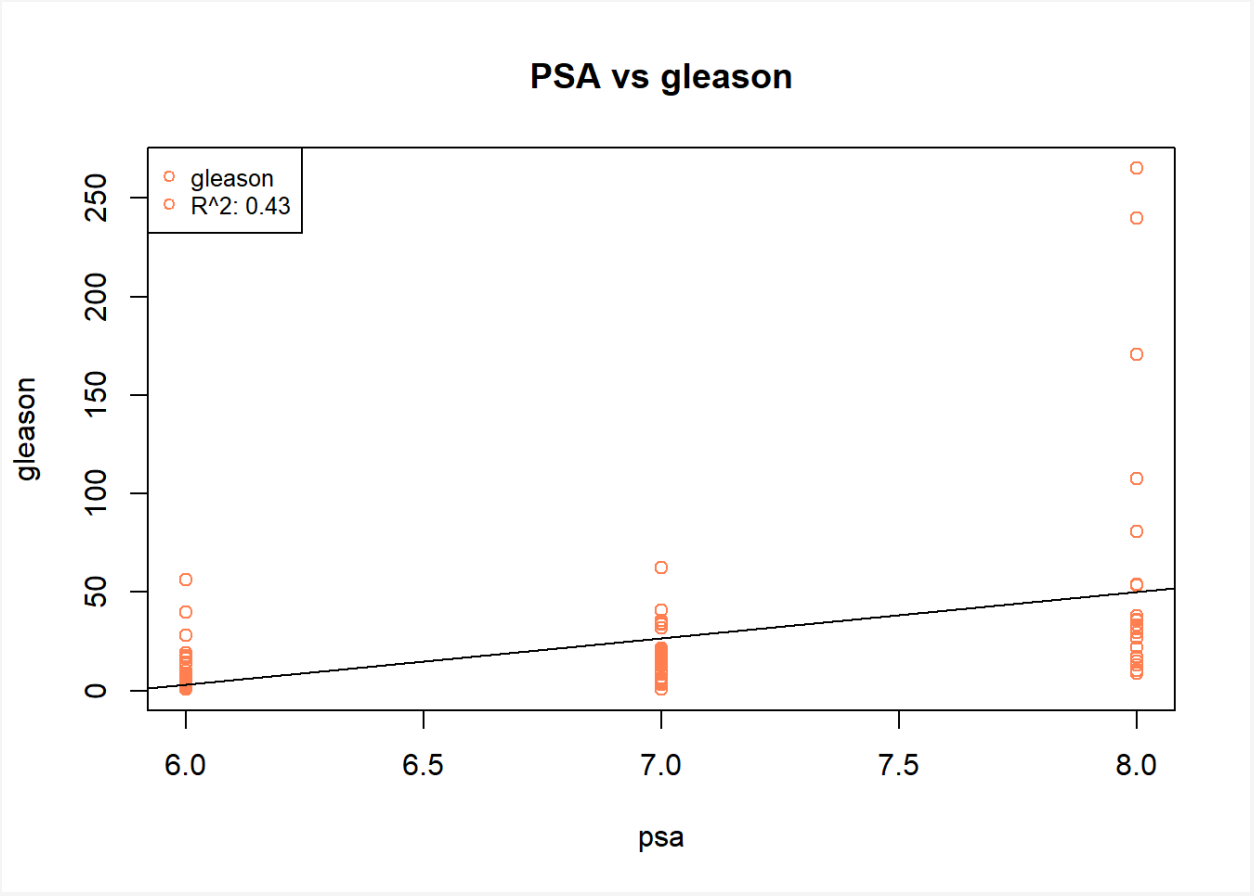

PSA vs weight
weight
R^2: 0.026
weight
0
50
100
150
200
250
0
100
200
300
400
psa


PSA vs age
age
R^2: 0.017
age
0
50
100
150
200
250
40
50
60
70
80
psa


PSA vs benpros
benpros
R^2: -0.016
benpros
0
50
100
150
200
250
0
2
4
6
8
10
psa


PSA vs vesinv
vesinv
R^2: 0.529
vesinv
0
50
100
150
200
250
0.0
0.2
0.4
0.6
0.8
1.0
psa


PSA vs capspen
capspen
R^2: 0.551
capspen
0
50
100
150
200
250
0
5
10
15
psa




```
model1 = lm(psa ~ cancervol + weight + benpros + vesinv + capspen + gleason)
anova(model1)
```

```
## Analysis of Variance Table
##
## Response: psa
##           Df Sum Sq Mean Sq F value    Pr(>F)
## cancervol  1  62202   62202 63.8370 4.279e-12 ***
## weight     1     85      85  0.0869  0.768857
## benpros    1    638     638  0.6546  0.420613
## vesinv     1   6861    6861  7.0415  0.009415 **
## capspen    1    869     869  0.8921  0.347426
## gleason    1   1321    1321  1.3557  0.247359
## Residuals 90  87695     974
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
model1AdjR = summary(model1)$adj.r.squared
```

```
model2 = lm(log(psa) ~ cancervol + weight + benpros + vesinv + capspen + gleason)
anova(model2)
```

```
## Analysis of Variance Table
```

```
##
## Response: log(psa)
##           Df Sum Sq Mean Sq F value    Pr(>F)
## cancervol  1 55.164  55.164 94.5479 1.073e-15 ***
## weight     1  1.790   1.790  3.0682 0.0832452 .
## benpros    1  6.219   6.219 10.6591 0.0015499 **
## vesinv     1  7.308   7.308 12.5253 0.0006377 ***
## capspen    1  0.141   0.141  0.2424 0.6237053
## gleason    1  4.637   4.637  7.9467 0.0059236 **
## Residuals 90 52.510   0.583
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
model2AdjR = summary(model2)$adj.r.squared
```

```
model3 = lm(log(psa) ~ cancervol + weight + benpros + vesinv + gleason)
anova(model3)
```

```
## Analysis of Variance Table
##
## Response: log(psa)
##           Df Sum Sq Mean Sq F value    Pr(>F)
## cancervol  1 55.164  55.164 94.8970 8.858e-16 ***
## weight     1  1.790   1.790  3.0795  0.082650 .
## benpros    1  6.219   6.219 10.6984  0.001515 **
## vesinv     1  7.308   7.308 12.5715  0.000621 ***
## gleason    1  4.390   4.390  7.5517  0.007229 **
## Residuals 91 52.898   0.581
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
model3AdjR = summary(model3)$adj.r.squared
```

```
model4 = lm(log(psa) ~ cancervol + benpros + vesinv + gleason)
anova(model4)
```

```
## Analysis of Variance Table
##
## Response: log(psa)
##           Df Sum Sq Mean Sq F value    Pr(>F)
## cancervol  1 55.164  55.164 95.3440 7.145e-16 ***
## benpros    1  7.803   7.803 13.4873 0.0004030 ***
## vesinv     1  7.334   7.334 12.6758 0.0005886 ***
## gleason    1  4.239   4.239  7.3264 0.0080997 **
## Residuals 92 53.229   0.579
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
model4AdjR = summary(model4)$adj.r.squared
```

```
finalModel = summary(model4)
print(finalModel)
```

```
##
## Call:
## lm(formula = log(psa) ~ cancervol + benpros + vesinv + gleason)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.88531 -0.50276  0.09885  0.53687  1.56621
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.65013    0.80999  -0.803 0.424253
## cancervol    0.06488    0.01285   5.051 2.22e-06 ***
## benpros      0.09136    0.02606   3.506 0.000705 ***
## vesinv       0.68421    0.23640   2.894 0.004746 **
## gleason      0.33376    0.12331   2.707 0.008100 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7606 on 92 degrees of freedom
## Multiple R-squared:  0.5834, Adjusted R-squared:  0.5653
## F-statistic: 32.21 on 4 and 92 DF,  p-value: < 2.2e-16
```

```
temp = finalModel$coefficients
intercept = temp[1]
betaCancervol = temp[2]
betaBenpros = temp[3]
betaVesinv = temp[4]
betaGleason = temp[5]

x1 = mean(cancervol)
x2 = mean(benpros)
x3 = 0
x4 = mean(gleason)

pred = intercept + betaCancervol*x1 + betaBenpros*x2 + betaVesinv*0 + betaGleason*x4
print(pred)
```

```
## [1] 2.330541
```