STATISTICAL METHODS FOR DATA SCIENCE

MINI PROJECT 6

Names of group members: 1. Arya Shah (Net Id: AAS190007)
2. Preethi Kesavan (Net Id: PXK190001)

Contribution of each group member:

Arya Shah:

Q1) Worked on summary statistics for body temperature (a) and the conclusion part in (c)

Q2) Worked on bootstrap interval estimation in (a) and (b).Also worked on the conclusion part in (c) and (d).

Preethi Kesavan:

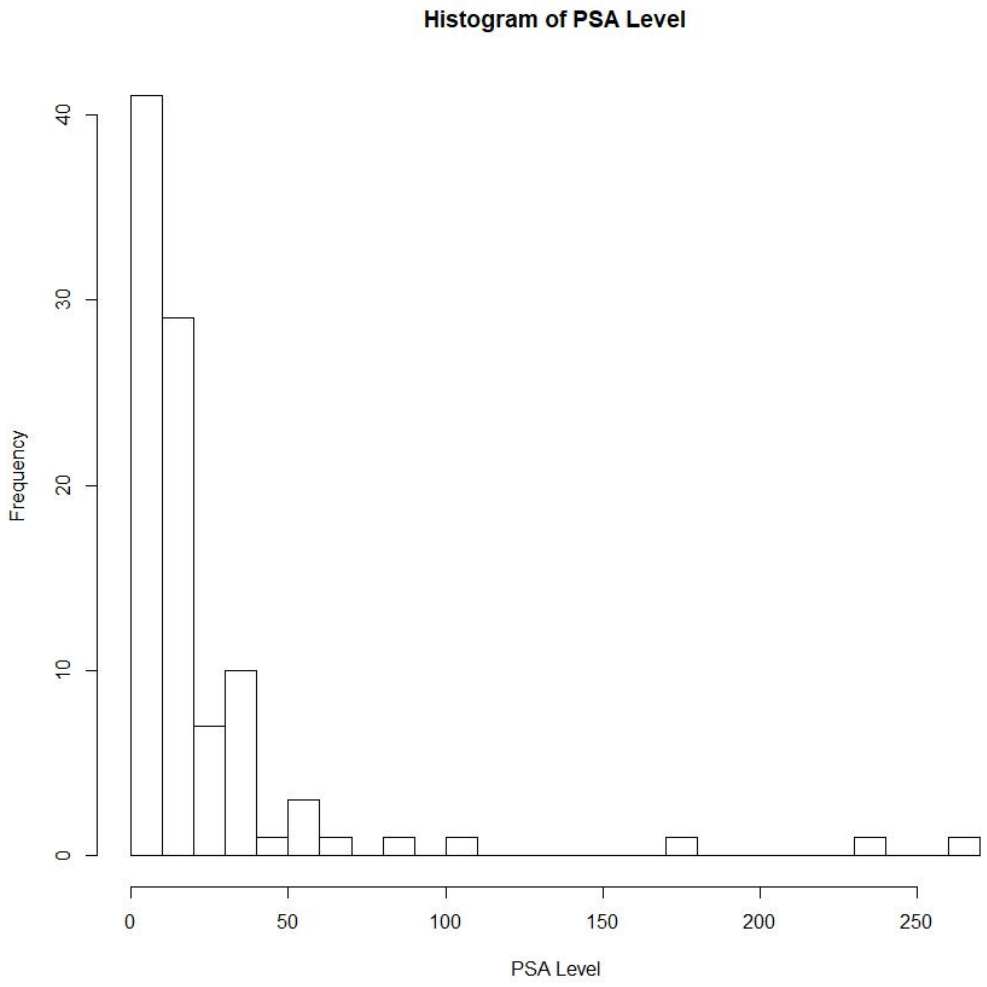
Q1) Worked on summary statistics for heart rate (b). Also worked on the conclusion part in (c)

Q2) Worked on the z-interval interval estimation in (a) and (b). Also worked on the conclusion part in (c) and (d).

Q1)
We need to make a “reasonably good” linear model for the data by taking PSA as the response variable.

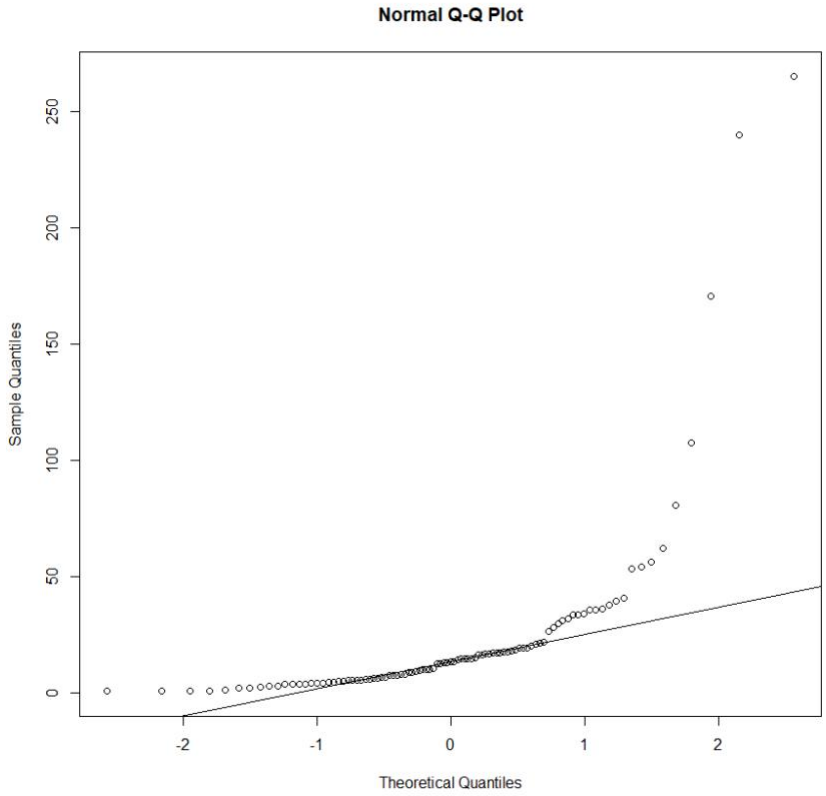
Exploratory Analysis of the response variable (PSA Level):

Histogram of PSA:



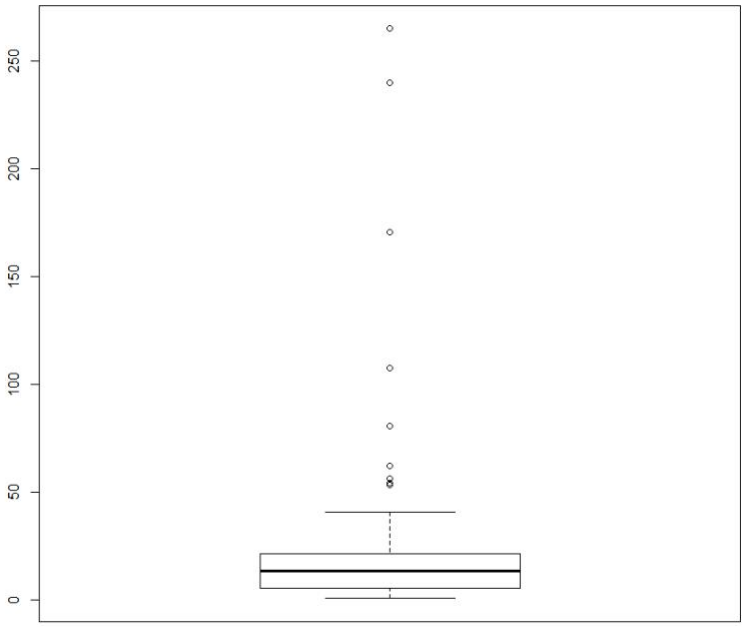
The Histogram shows that the distribution appears to be like an Exponential distribution, and very different to a normal distribution.
Most people have very low PSA levels, and the number of people drastically reduces as PSA level increases.

Normal Q-Q Plot of PSA Level:



The Normal Q-Q Plot above also shows that this data deviates from the Normal Q-Q Line and does not approximate the normal distribution.

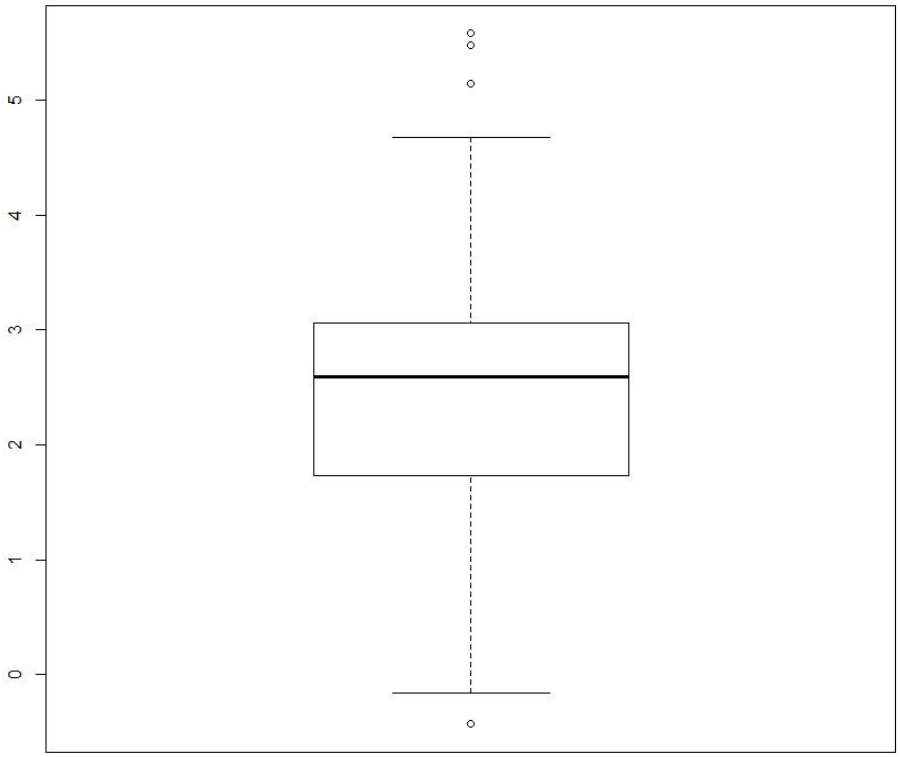
Boxplot of response variable(psa):



There are many outliers in the psa data, which can be seen from the boxplot above. Therefore, a transformation is required.

Let us use a log transformation on the response variable(psa) and again take a look at its distribution through a boxplot.

Boxplot of log(psa):

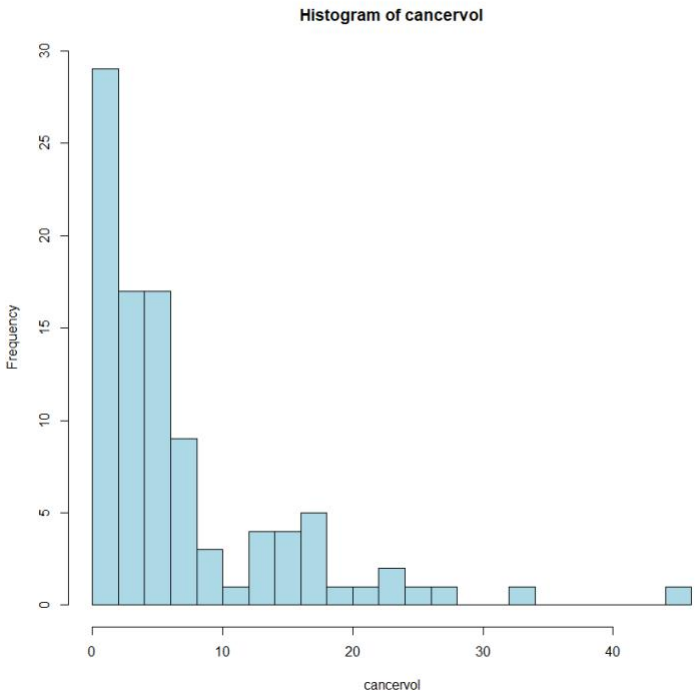


The number of outliers have been reduced and the distribution has now become more symmetric, so we now used the transformed response (log(psa)) as our response variable.

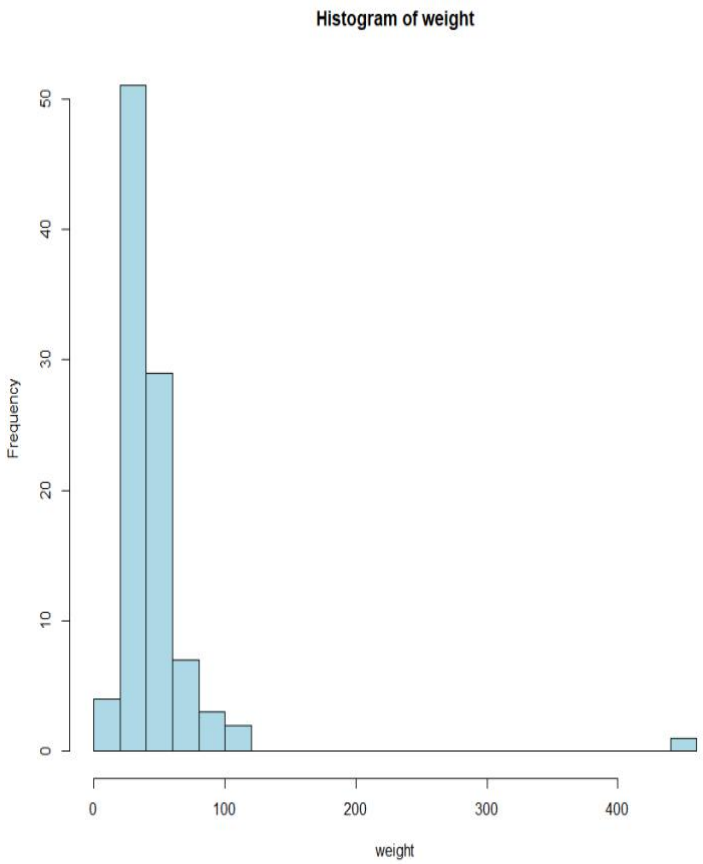
Exploratory Analysis of all possible predictor variables:

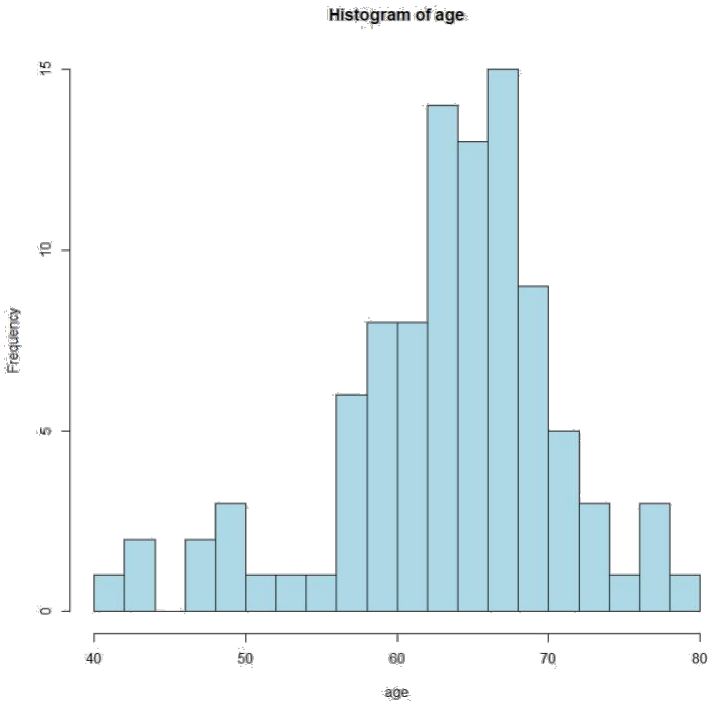
Let us now perform an exploratory analysis on the predictor variables which can be possibly introduced into our linear model and observe their distributions.

Histograms of predictor variables:

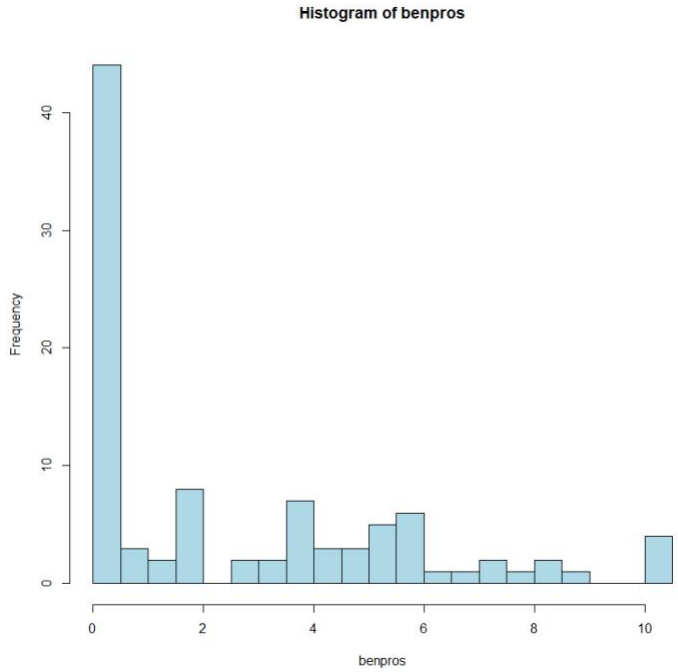


This predictor variable very closely approximates the distribution observed in our response variable (PSA level). This hints at a possible linear relationship between cancervol and psa.

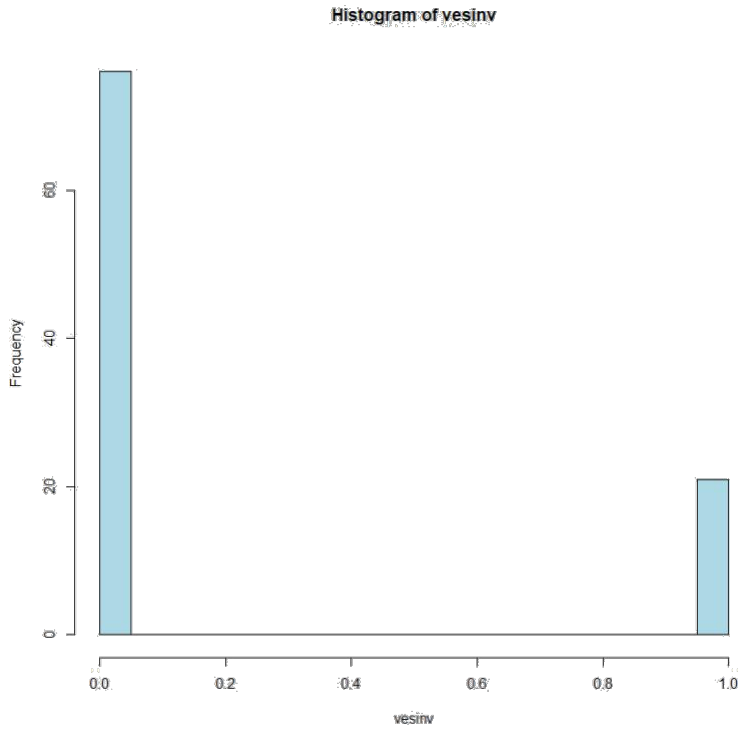

Histogram of weight
Frequency
0
10
20
30
40
50
0
100
200
300
400
weight


Histogram of age
Frequency
0
5
10
15
40
50
60
70
80
age

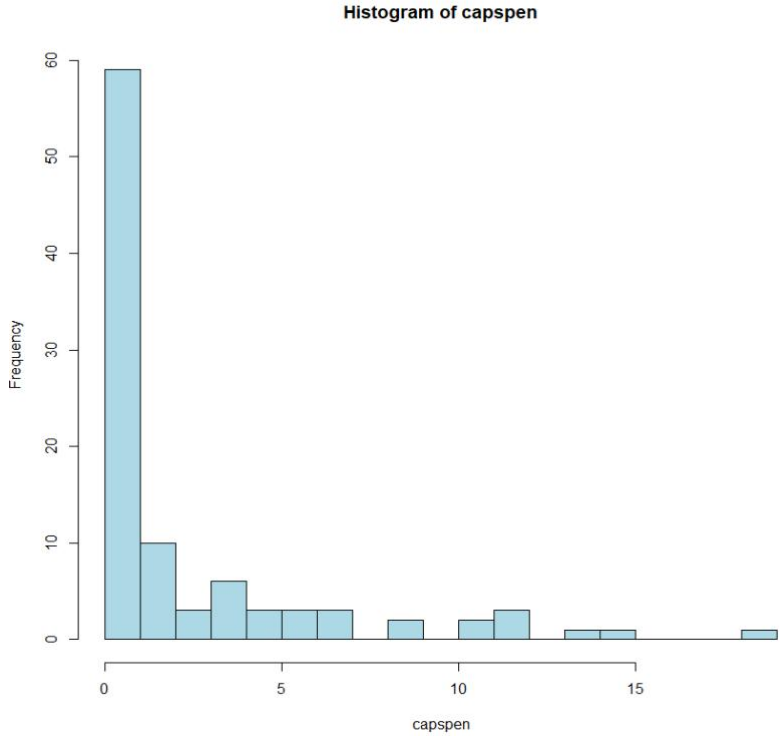
The distribution of ages of the sample approximates a normal distribution which is expected for a random sample from a human population.



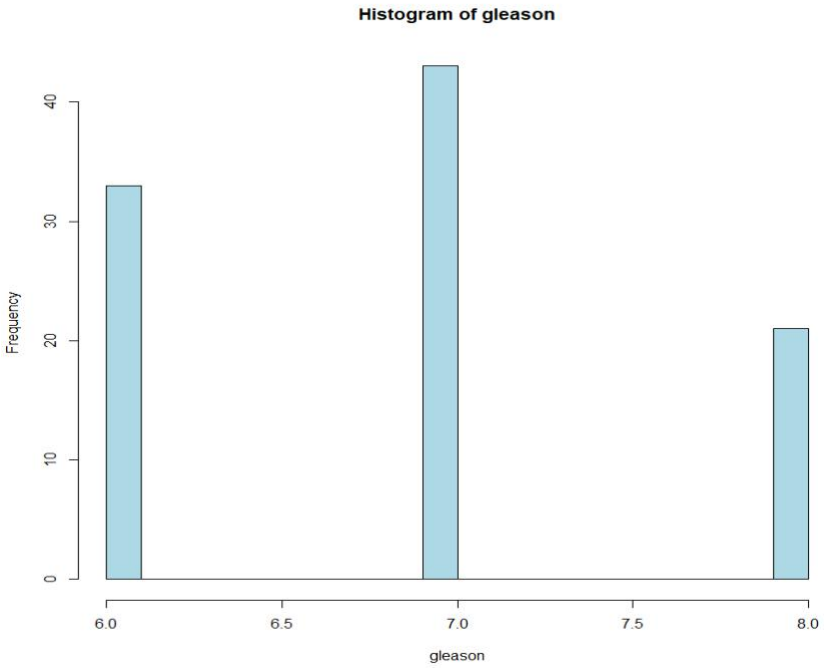
The above distribution also appears exponential; however, it appears less similar to psa than cancervol.



The variable vesinv is a categorical/qualitative variable and assumes only two possible values (0 or 1). There appears to be many more people without Seminal Vesicle Invasion (value 0) than those with it (value 1).

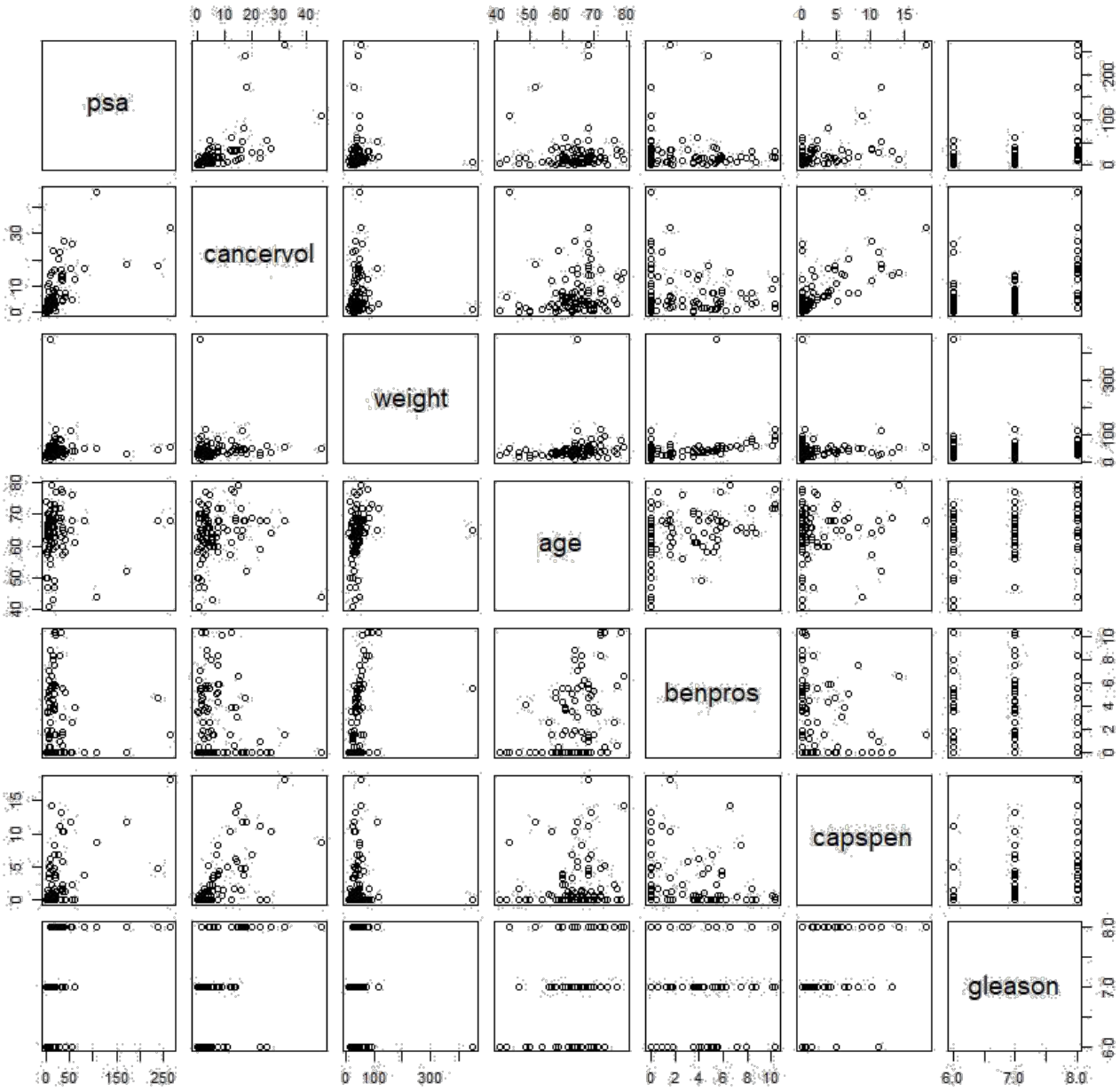


The distribution of the variable capspen is also very similar to those of psa and cancervol, and hints at a possible correlation between them.



The variable gleason has a distribution of only three values (6,7 and 8). Since the values have numerical meaning attached to them, let us treat this as a quantitative variable.

Now lets us look at the scatter plots and correlations between the variables to get a better understanding of the linear trends which may exist between these variables, before we start making our linear model.
Possible trends that exists can be visualized from the scatterplots presented below.



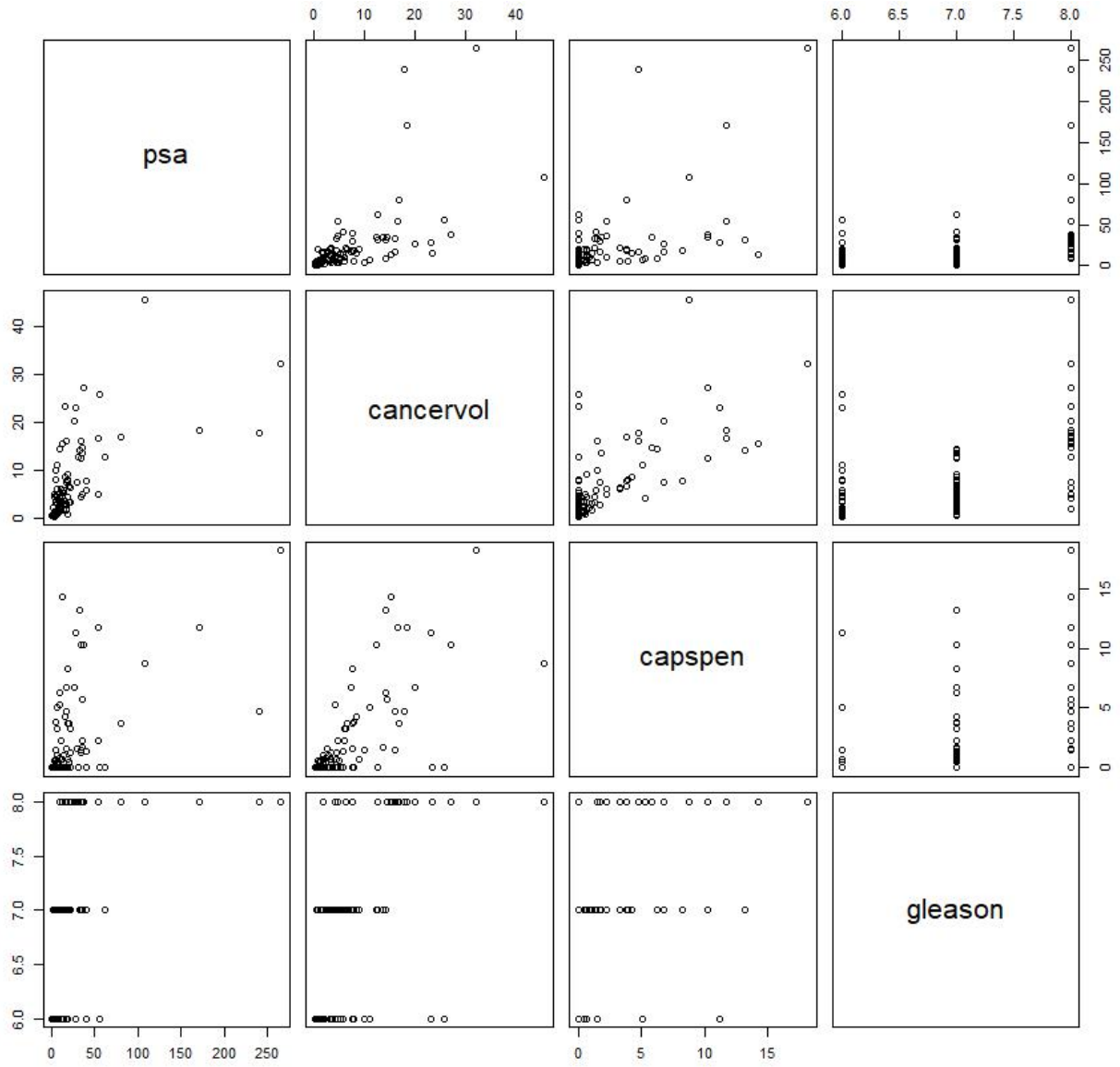
	psa	cancervol	weight	age	benpros	capspen	gleason
psa	1.000	0.624	0.026	0.017	-0.016	0.551	0.430
cancervol	0.624	1.000	0.005	0.039	-0.133	0.693	0.481
weight	0.026	0.005	1.000	0.164	0.322	0.002	-0.024
age	0.017	0.039	0.164	1.000	0.366	0.100	0.226
benpros	-0.016	-0.133	0.322	0.366	1.000	-0.083	0.027
vesinv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

From the above data we can conclude that there exists linear trend between response variable (PSA level) along with the following quantitative predictors: cancervol, capspen and gleason.

However, one more important thing to notice is the high correlation between the predictor variables themselves: the following predictoers show a high level of correlation between themselves, so we should try avoiding overfitting of data, which can be caused by this.

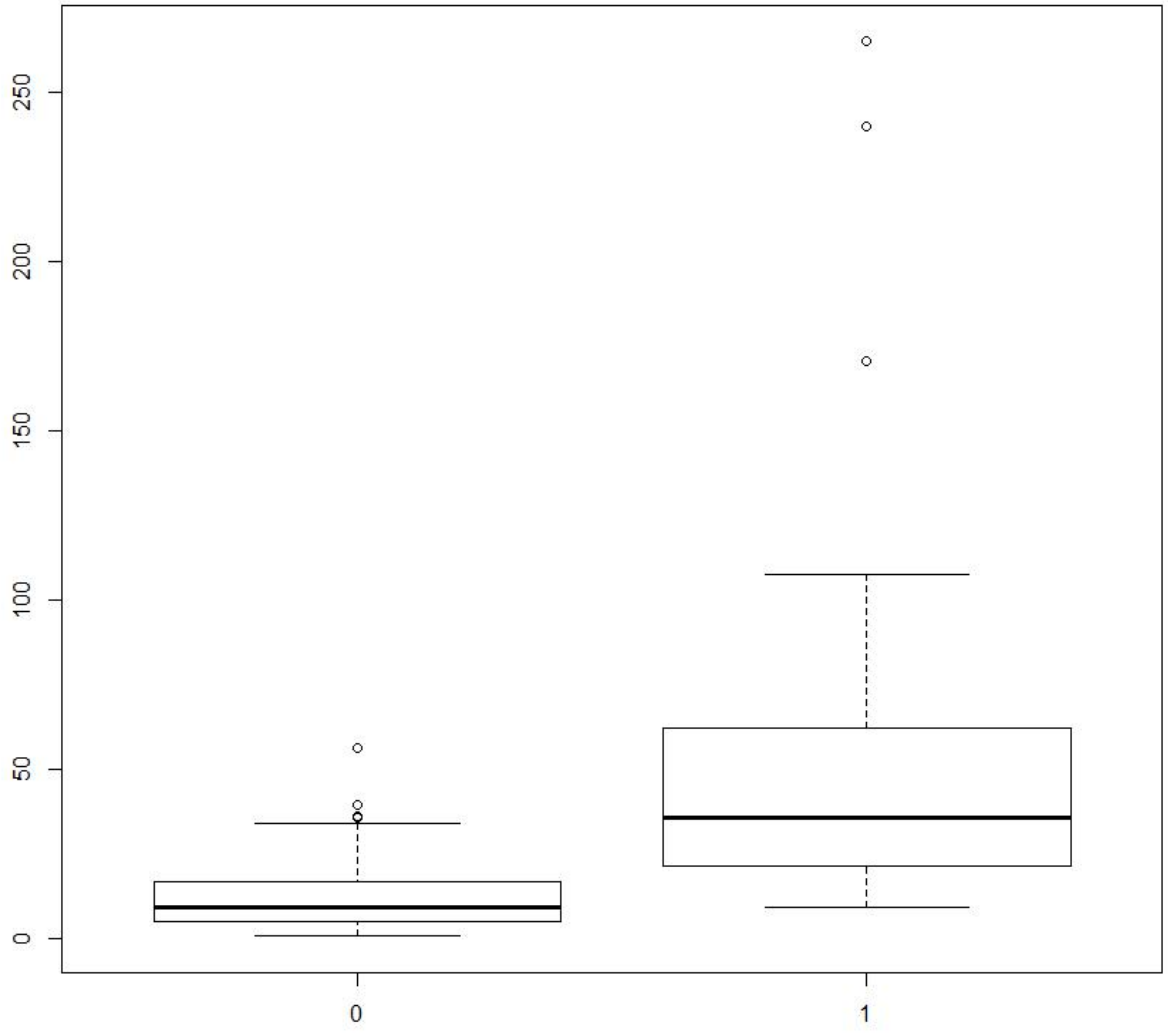
cancervol, capspen and gleason are quantitative predictor variables which show a significant degree of correlation amongst themselves, but also share significant degree of correlation to the response variable(psa). These variables seem to be the most important predictor variables for the respone(psa).

Let us have a closer look at the trend between these variables:



Now let us perform an exploratory analysis of the Qualitative variable : vesinv

Boxplot of relationship between different levels of the categorical variable (vesinv) and psa:



This boxplot suggests that the psa values vary significantly over the levels of vesinv:

Now let us start setting up a preliminary linear model using the quantitative variables cancervol, capspen and gleason.

We ignore the other variables, namely: weight, age and benpros since our exploratory analysis shows no statistically significant evidence of merit to include them into our linear model.

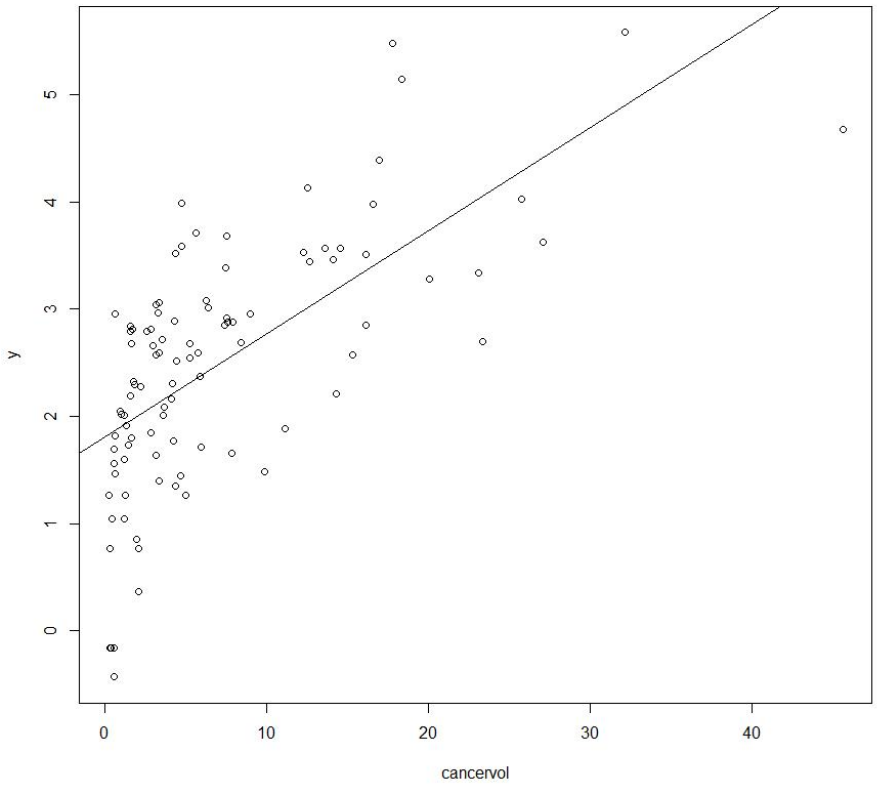
We have also decided that it is in our interest to transform the response variable (psa) using the log (natural logarithm) transformation.

Let us again observe the relationships between the transformed response (log psa) and each of the predictor variables that we have decided to include after our preliminary analysis:

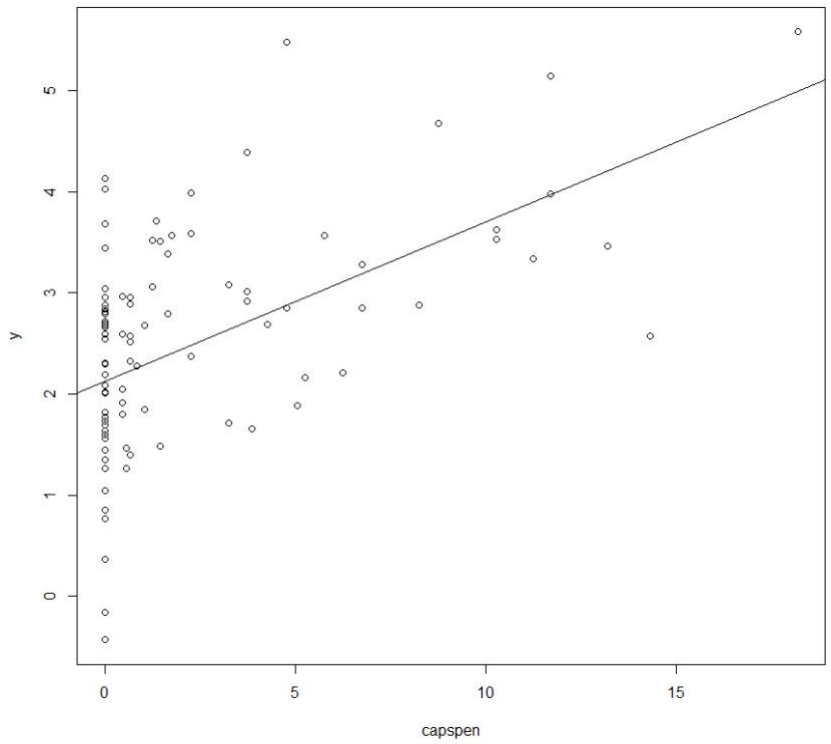
The y along the y-axis denotes the response variable-(log psa) :

Quantitative predictors:

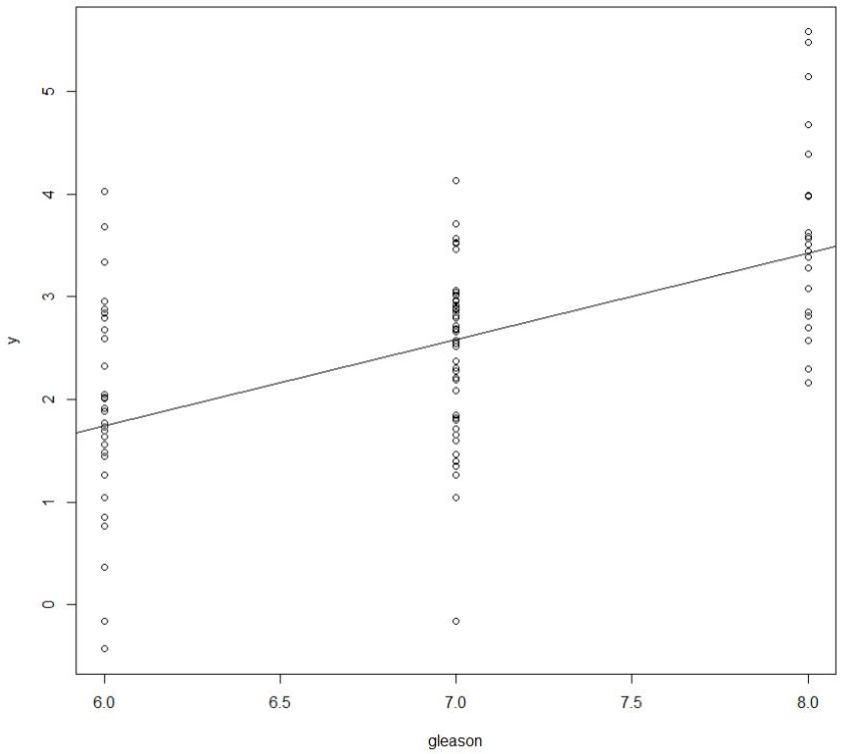
Cancervol and y:



Capspen and y:



Gleason and y:



Again let us look at all the correlations between our transformed response variable log(psa)

And other variables, just to be safe that the same correlations that existed before are still as strong as before:

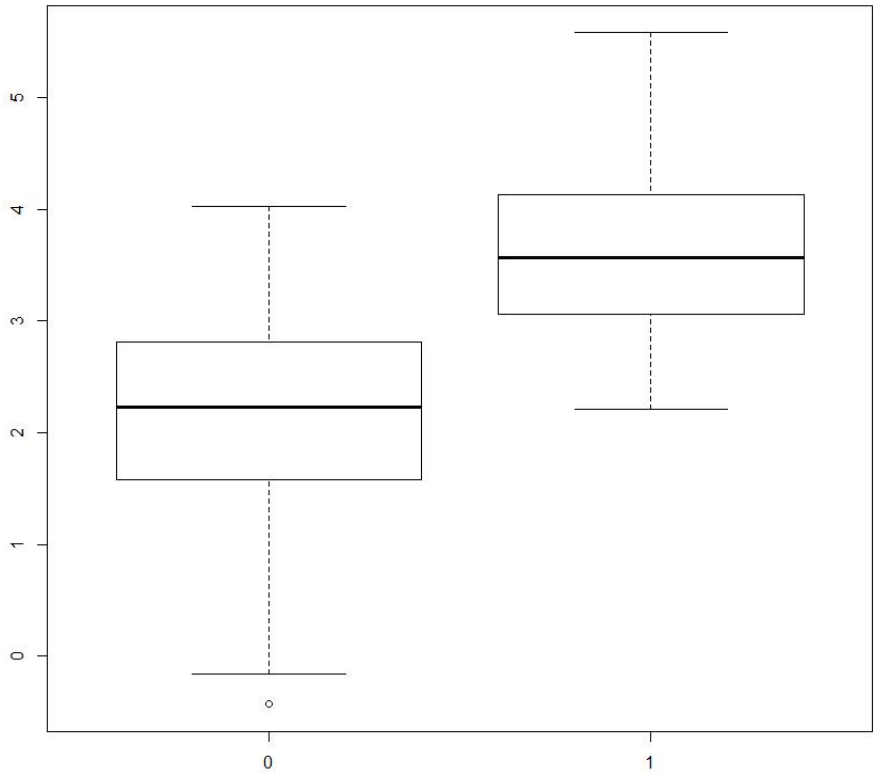
```
            psa cancervol weight age benpros capspen gleason
psa       1.000 0.657 0.122 0.170 0.157 0.518 0.539
```

Some of the correlations have slightly changed, but for the most part, our preliminary analysis of the best possible predictors still holds.

Qualitative predictor:

To make sure the relationship still holds with the newly transformed response we perform the same boxplot again:

Boxplot of relationship between different levels of the categorical variable (vesinv) and log psa:



The difference between the two levels is still significant even after the transformation.

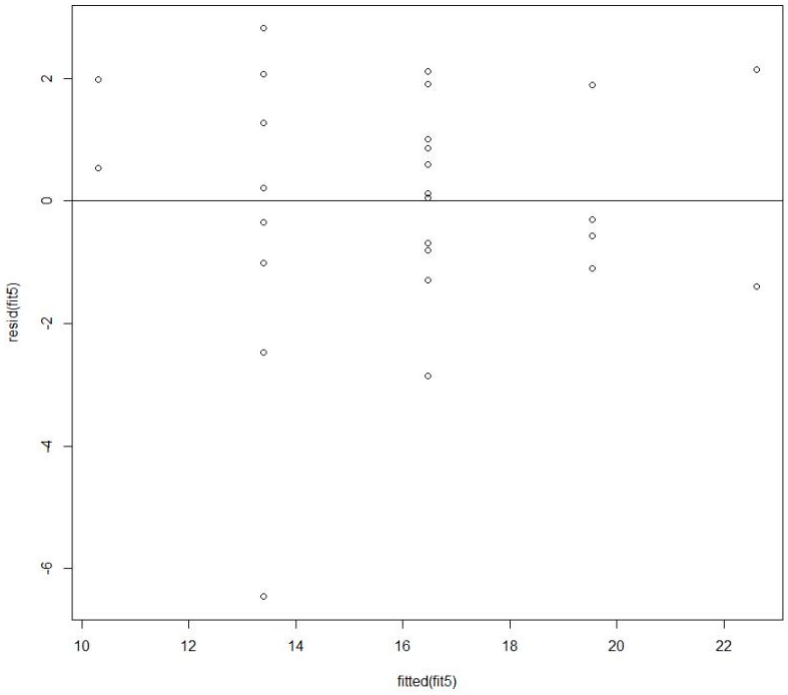
Now let us start building a linear model with the quantitative predictors cancervol, capspen and gleason since we observe a significant positive trend between each of these variables and our response variable (log psa), and the categorical predictor vesinv since our exploratory analysis found the variable to be significant.

The first model (fit4) has the variables described above. On performing a summary analysis through R, we see that a t-test of the variable capspen, provided evidence against its significance in our model. Perhaps the carriable capspen is not required as a predictor for this model. Now, we perform a partial F test for two Nested models (fit4 with all the variables; fit5 which is a nested model which does not have variable capspen). The p value for the partial F-statistic is high (0.4985), so we accept the null hypothesis and conclude that the reduced model is just as good. Hence, we can ignore the variable capspen. So let us continue testing reduced models against the newer regression model fit5.

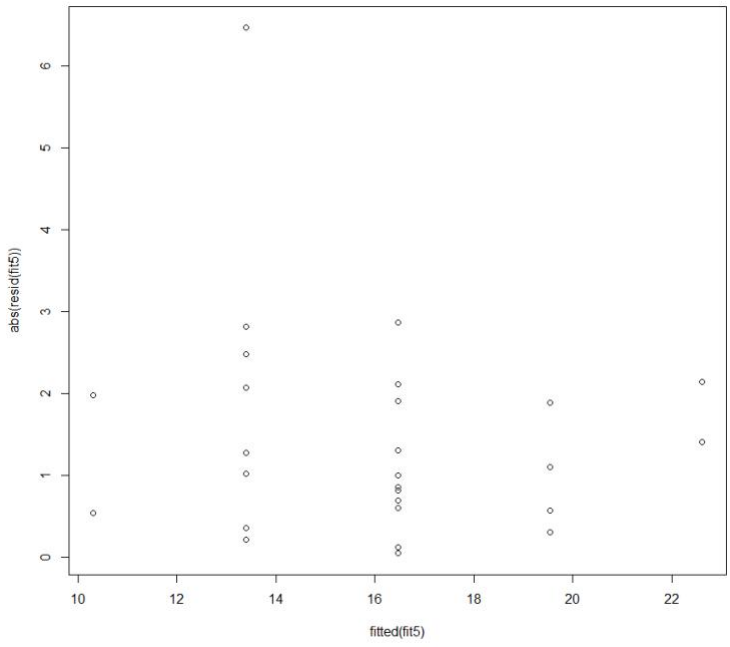
Similarly, we try to further reduce the model by reducing variables (quantitative variable gleason and then categorical variable vesinv). There however is evidence that we cannot reduce these variables (low p-value from the partial F-tests from tests on nested models), so we keep them in our model. We don't have any non-significant predictors, so let us use this(fit5) as our preliminary model.

Performing one last summary analysis shows us that all the predictor variables are significant, so now we take this model(y~cancervol+gleason+vesinv) as our final model and perform diagnostics:

Residual Plot:

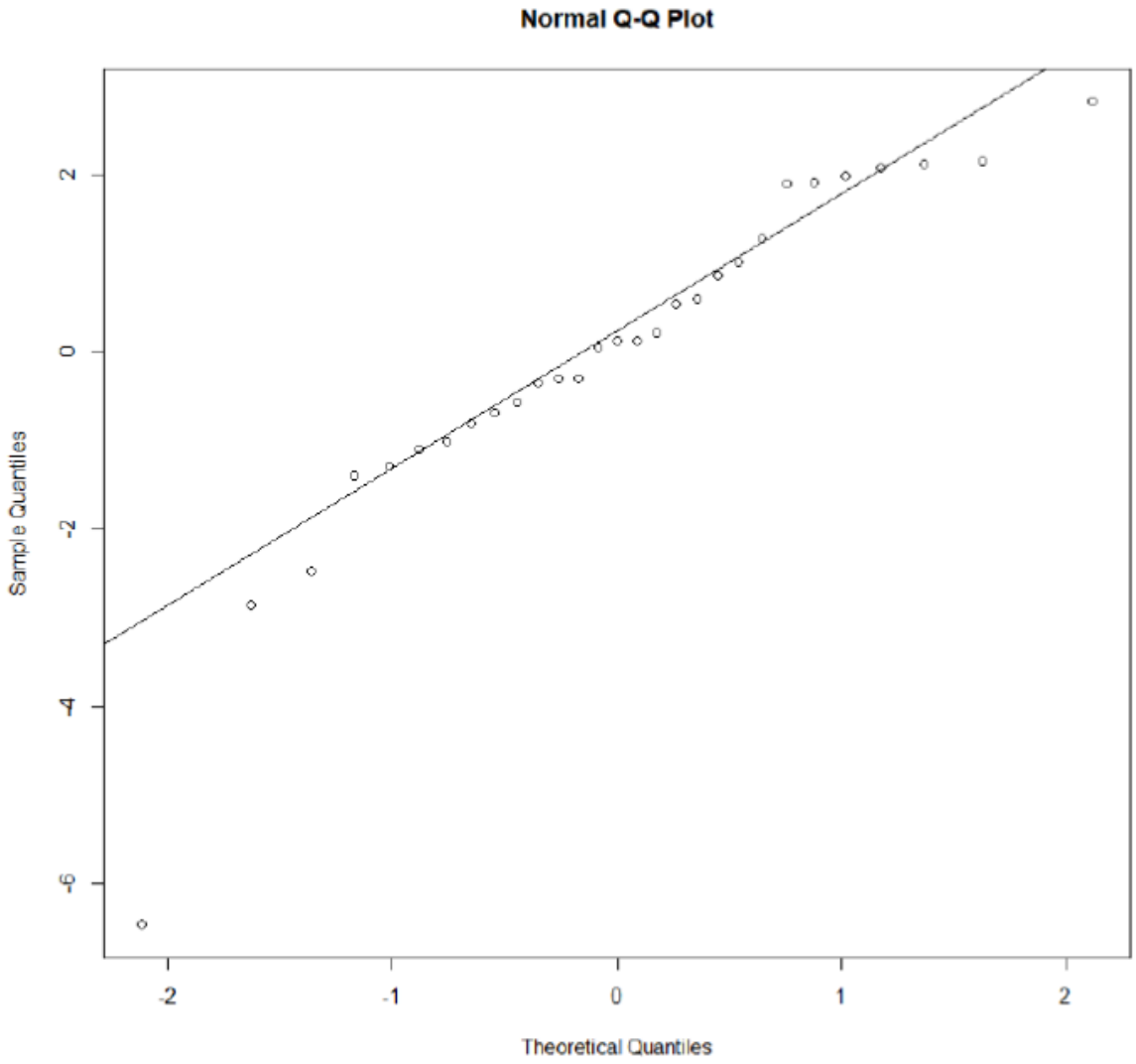


Plot of Absolute Residuals:



The above plots indicate no particular trend in the residuals.

Normal Q-Q Plot of Residuals:



The normality assumption of the Residuals also holds.

All of the model assumptions hold, so the model we have selected has passed the diagnostics. We can take this as our final model.

Q1)

Source Code:

```
#Setting working directory for easy access

setwd(‘Desktop/MiniProject 6')

MyData<-read.csv('prostate_cancer.csv')


#vesinv is a categorical variable (R treats factors as categorical variables)

MyData$vesinv=factor(MyData$vesinv)


#psa is the response

#we would like to see which of these variables could be used as accurate
predictors for the response variable (psa).


#Let us assign the variable names to their respective data(columns)

psa=MyData[,2]

cancervol=MyData[,3]

weight=MyData[,4]

age=MyData[,5]

benpros=MyData[,6]

vesinv=MyData[,7]

capspen=MyData[,8]

gleason=MyData[,9]


#EXPLORATORY ANALYSIS OF RESPONSE (PSA LEVEL)

#Histogram

hist(psa, xlab="PSA Level",main= "Histogram of PSA Level",breaks=20)

#Q-Q Plots

qqnorm(psa)

qqline(psa)

#Boxplot of psa level indicates many

outliers boxplot(psa)

#Looking at the distribution of the response variable(psa) after log transformation
is applied.

#Boxplot of transformed response (log(psa))

boxplot(log(psa))

We notice that the number of outliers has reduced, and the distribution becomes
more symmetrical.
```

```
#QUANTITATIVE VARIABLES EXPLORATORY ANALYSIS


#Single for-loop for histograms of each of the variables
for (j in 1:9) {
  hist(MyData[,j], xlab=colnames(MyData)[j],
       main=paste("Histogram of",colnames(MyData[j])),
       col="lightblue",breaks=20)


#scatterplots and correlations between all variables:
#using pairs for all scatterplots to get an overview of all existing trends
pairs(~psa + cancervol + weight + age + benpros + capspen + gleason, data = MyData)
#log PSA
pairs(~psa + cancervol + capspen + gleason, data = MyData)


#Getting all the correlations between each pair of variables
prostate.cor = cor(MyData[,2:9]) round(prostate.cor,3)


             psa cancervol weight   age benpros capspen gleason
psa        1.000     0.624  0.026 0.017  -0.016   0.551  0.430
cancervol  0.624     1.000  0.005 0.039  -0.133   0.693  0.481
weight     0.026     0.005  1.000 0.164   0.322   0.002 -0.024
age        0.017     0.039  0.164 1.000   0.366   0.100  0.226
benpros   -0.016    -0.133  0.322 0.366   1.000  -0.083  0.027
capspen    0.551     0.693  0.002 0.100  -0.083   1.000  0.462
gleason    0.430     0.481 -0.024 0.226   0.027   0.462  1.000


#We are most interested in the first line which is correlation between PSA and
other elements, however we also look at correlations between other variables to
avoid overfitting


#PSA has stronger correlations with quantitative variables cancervol, capspen, and
gleason


#log transformation of PSA with other
variabless cor(MyData, log(psa))


#QUALITATIVE VARIABLE EXPLORATORY ANALYSIS : vesinv


#Boxplots
```

```
#The boxplot shows a strong difference between the psa level based on the
two categories

boxplot(psa~vesinv)


#We have decided to use log(psa) as the new transformed response

#We have decided to exclude the following variables as predictors: weight, age and
benpros based on the previous analysis

#Now let us look at the relation between the response and each predictor one by one

#Since we are now transforming our response to log psa

#Quantitative

y=log(psa)

#cancervol and response(y)

plot(cancervol,y)'

fit1 = lm(y ~ cancervol, data = MyData)

abline(fit1)


#capspen and response(y)

plot(capspen,y)

fit2 = lm(y ~ capspen, data = MyData)

abline(fit2)


#gleason and response(y)

plot(gleason,y)

fit3 = lm(y ~ gleason, data = MyData)

abline(fit3)


#Checking correlations once again with newly transformed response log(psa), out of
#curiosity to make sure no adverse changes has occurred

#Lets make a new cop of the variable MyData and transform the response (psa) to
log(psa) in that copy

MyData2=MyDataMyData$psa=log(psa)
MyData2=MyData
MyData$psa=log(psa)
prostate.cor = cor(MyData[c(2,3,4,5,6,8,9)])
round(prostate.cor,3)
            psa cancervol weight   age benpros capspen gleason
psa       1.000     0.657  0.122 0.170   0.157   0.518   0.539
cancervol 0.657     1.000  0.005 0.039  -0.133   0.693   0.481
weight    0.122     0.005  1.000 0.164   0.322   0.002  -0.024
age       0.170     0.039  0.164 1.000   0.366   0.100   0.226
benpros   0.157    -0.133  0.322 0.366   1.000  -0.083   0.027
capspen   0.518     0.693  0.002 0.100  -0.083   1.000   0.462
gleason   0.539     0.481 -0.024 0.226   0.027   0.462   1.000
```

```
#qualitative:
boxplot(y~vesinv)

#Building first with quantitative variables and qualitative
variable #First we use all three variables: cancervol, capspen, and
gleason fit4=lm(y~cancervol+capspen+gleason+vesinv) fit4


Call:
lm(formula = y ~ cancervol + capspen + gleason + vesinv)

Coefficients:
(Intercept)     cancervol      capspen      gleason      vesinv1
   -0.79386       0.06452     -0.02348      0.39566      0.70675

#summary of the model
summary(fit4)

Call:
lm(formula = y ~ cancervol + capspen + gleason + vesinv)

Residuals:
    Min      1Q  Median      3Q     Max
-2.1747 -0.4497  0.1049  0.6215  1.6135

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.79386    0.86660  -0.916  0.36203
cancervol    0.06452    0.01522   4.238 5.35e-05 ***
capspen     -0.02348    0.03455  -0.680  0.49852
gleason      0.39566    0.13100   3.020  0.00327 **
vesinv1      0.70675    0.28024   2.522  0.01339 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8078 on 92 degrees of freedom
Multiple R-squared:  0.5301,      Adjusted R-squared:  0.5097
F-statistic: 25.95 on 4 and 92 DF,  p-value: 2.075e-14

#Based on the summary, it seems very clear that capspen is not required for the
model #Let us continue the tests with nested models

#We know that these three variables have significant correlation with each
other #so we need to check whether all of these are necessary #Let us reduce
the model ,removing capspen
fit5=lm(y~cancervol+gleason+vesinv)
#removing both capspen and gleason
fit6=lm(y~cancervol+vesinv)

#Now first performing partial F test to check the significance of capspen (fit4, fit5)
anova(fit4,fit5)
Analysis of Variance Table

Model  1: y ~ cancervol + capspen + gleason + vesinv
Model  2: y ~ cancervol + gleason + vesinv
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     92 60.039
2     93 60.340 -1 -0.30134 0.4617 0.4985
#Clearly capspen is not needed and is redundant

#Now let us check if gleason is needed performing partial F test to check the
significance
#of capspen (fit5, fit6)
anova(fit5,fit6)
Analysis of Variance Table

Model  1: y ~ cancervol + gleason + vesinv
Model  2: y ~ cancervol + vesinv
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1     93 60.340
2     94 66.058 -1   -5.7179 8.8127 0.003804 **
---
```

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
#It appears that gleason is an important predictor and no
statistically #significant evidence against it

#Just for the sake of curiosity, let us test whether the categorical variable vesinv
#can be ignored
fit7=lm(y~cancervol+gleason)
anova(fit5,fit7)
Analysis of Variance Table

Model  1: y ~ cancervol + gleason + vesinv
Model  2: y ~ cancervol + gleason
  Res.Df     RSS Df Sum of Sq      F  Pr(>F)
1     93 60.340
2     94 64.358 -1   -4.0178 6.1925 0.01461 *
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
#Evidence against vesinv is also not strong enough
#Hence we accept fit5 as a preliminary
model summary(fit5)

Call:
lm(formula = y ~ cancervol + gleason + vesinv)

Residuals:
     Min       1Q   Median       3Q      Max
-2.16928 -0.44558  0.08431  0.60719  1.64082

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.72120    0.85749  -0.841   0.4025
cancervol    0.05981    0.01352   4.425 2.62e-05 ***
gleason      0.38491    0.12966   2.969   0.0038 **
vesinv1      0.62117    0.24962   2.488   0.0146 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8055 on 93 degrees of freedom
Multiple R-squared:  0.5277,      Adjusted R-squared:  0.5125
F-statistic:  34.64 on 3 and 93 DF,  p-value: 4.022e-15


#Let us check how our fit5 compares with the automatic stepwise model
selection procedures based on AIC
# Forward selection based on AIC

fit8.forward <- step(lm(y ~ 1, data = MyData2),
+                      scope = list(upper = ~cancervol+capspen+gleason+vesinv),
+                      direction = "forward")
Start:   AIC=28.72
y ~ 1

             Df Sum of Sq      RSS      AIC
+ cancervol   1    55.164  72.605 -24.0986
+ vesinv      1    40.984  86.785  -6.7944
+ gleason     1    37.122  90.647  -2.5707
+ capspen     1    34.286  93.482   0.4169
<none>                    127.769  28.7246

Step:  AIC=-24.1
y ~ cancervol

           Df Sum of Sq    RSS      AIC
+ gleason   1    8.2468 64.358  -33.794
+ vesinv    1    6.5468 66.058  -31.265
<none>                 72.605  -24.099
+ capspen   1    0.9673 71.638  -23.400

Step:   AIC=-33.79
y ~ cancervol + gleason

           Df Sum of Sq    RSS      AIC
```

```
+ vesinv    1    4.0178 60.340 -38.047
<none>                  64.358 -33.794
+ capspen   1    0.1685 64.190 -32.048

Step:  AIC=-38.05
y ~ cancervol + gleason + vesinv

              Df Sum of Sq  RSS    AIC
<none>                    60.340 -38.047
+ capspen 1 0.30134 60.039 -36.532

#Backward elimination based on AIC

fit9.backward <- step(lm(y~cancervol+capspen+gleason+vesinv, data = MyData2),
+                         scope = list(lower = ~1), direction = "backward")
Start:  AIC=-36.53
y ~ cancervol + capspen + gleason + vesinv

            Df Sum of Sq     RSS     AIC
- capspen    1    0.3013  60.340  -38.047
<none>                    60.039  -36.532
- vesinv     1    4.1507  64.190  -32.048
- gleason    1    5.9535  65.993  -29.361
- cancervol  1   11.7209  71.760  -21.234

Step:  AIC=-38.05
y ~ cancervol + gleason + vesinv

            Df Sum of Sq     RSS     AIC
<none>                    60.340  -38.047
- vesinv     1    4.0178  64.358  -33.794
- gleason    1    5.7179  66.058  -31.265
- cancervol 1 12.7041 73.044 -21.513

#Both forward and backward

fit10.both <- step(lm(y ~ 1, data = MyData2),
+                    scope = list(lower = ~1, upper =
~cancervol+capspen+gleason+vesinv),
+                    direction = "both")
Start:  AIC=28.72
y ~ 1

            Df Sum of Sq     RSS      AIC
+ cancervol  1    55.164  72.605 -24.0986
+ vesinv     1    40.984  86.785  -6.7944
+ gleason    1    37.122  90.647  -2.5707
+ capspen    1    34.286  93.482   0.4169
<none>                   127.769  28.7246

Step:  AIC=-24.1
y ~ cancervol

            Df Sum of Sq     RSS     AIC
+ gleason    1     8.247  64.358 -33.794
+ vesinv     1     6.547  66.058 -31.265
<none>                    72.605 -24.099
+ capspen    1     0.967  71.638 -23.400
- cancervol  1    55.164 127.769  28.725

Step:  AIC=-33.79
y ~ cancervol + gleason

            Df Sum of Sq     RSS     AIC
+ vesinv     1    4.0178  60.340  -38.047
<none>                    64.358  -33.794
+ capspen    1    0.1685  64.190  -32.048
- gleason    1    8.2468  72.605  -24.099
- cancervol  1   26.2887  90.647  -2.571

Step:  AIC=-38.05
y ~ cancervol + gleason + vesinv
```

```
            Df Sum of Sq    RSS     AIC
<none>                    60.340 -38.047
+ capspen    1    0.3013 60.039 -36.532
- vesinv     1    4.0178 64.358 -33.794
- gleason    1    5.7179 66.058 -31.265
- cancervol  1   12.7041 73.044 -21.513

#Our preliminary model is the same as those produced by
#automatic stepwise model selection procedures based on AIC
#Hence we accept our model and perform the diagnostics #The
model selected is: cancervol+gleason+vesinv
#fit5(preliminary model), fit8.forward(Forward selection based on
AIC), #fit9.backward(Backward elimination based on AIC)
#and fit10.both(forward/backward) all follow this same model

summary(fit5)

Call:
lm(formula = y ~ cancervol + gleason + vesinv)

Residuals:
     Min       1Q   Median       3Q      Max
-2.16928 -0.44558  0.08431  0.60719  1.64082

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.72120    0.85749  -0.841   0.4025
cancervol    0.05981    0.01352   4.425 2.62e-05 ***
gleason      0.38491    0.12966   2.969   0.0038 **
vesinv1      0.62117    0.24962   2.488   0.0146 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8055 on 93 degrees of freedom
Multiple R-squared:  0.5277,      Adjusted R-squared:  0.5125
F-statistic: 34.64 on 3 and 93 DF,  p-value: 4.022e-15

#the summary tells us that our regression variables are all significant

# residual plot
plot(fitted(fit5),
resid(fit5)) abline(h = 0)
#No trend in the residuals

# plot of absolute residuals

plot(fitted(fit5), abs(resid(fit5)))
#Still no trend

# normal QQ plot
qqnorm(resid(fit5))
qqline(resid(fit5))
#The residuals approximate a normal
distribution #All assumptions hold
# This preliminary model passes the diagnostics. So we can take this as our final
model.
```