Statistical Methods for Data Science

Mini Project #6

Group Members: Arpita Kumane, Harshali Dube
Contribution: Both the team members collaborated and worked on the project together. We analyzed, discussed, and efficiently worked to submit the two questions.

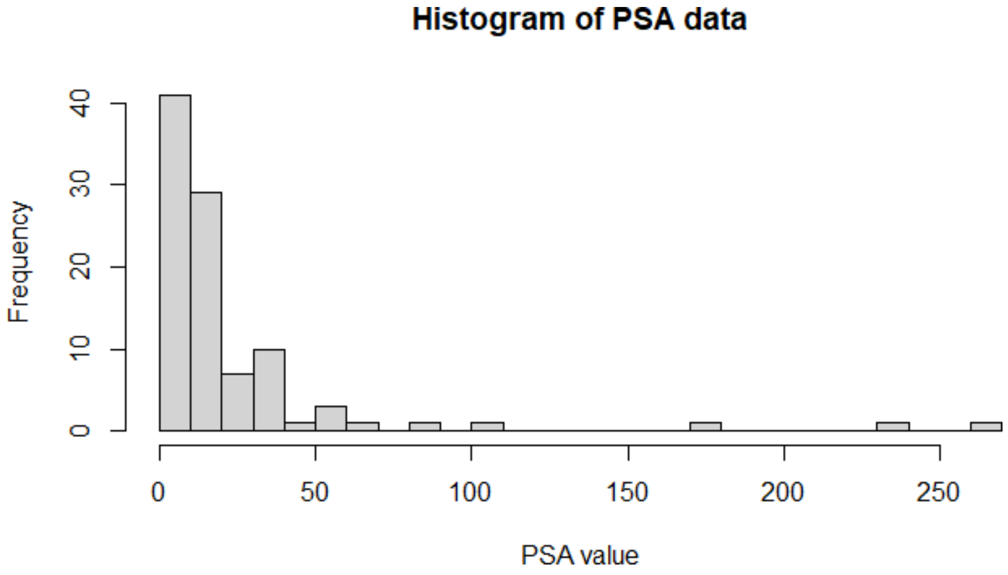
Question 1

Consider the prostate cancer dataset available on eLearning as prostate cancer.csv. It consists of data on 97 men with advanced prostate cancer. A description of the variables is given in Figure1. We would like to understand how PSA level is related to the other predictors in the dataset. Note that vesinv is a qualitative variable. You can treat gleason as a quantitative variable.
Build a “reasonably good” linear model for these data by taking PSA level as the response variable. Carefully justify all the choices you make in building the model. Be sure to verify the model assumptions. In case a transformation of response is necessary, try the natural log transformation. Use the final model to predict the PSA level for a patient whose quantitative predictors are at the sample means of the variables and qualitative predictors are at the most frequent category.

Sol:

To create the best linear model for the supplied data, we must undertake exploratory analysis on the data in order to have a thorough understanding of the data.

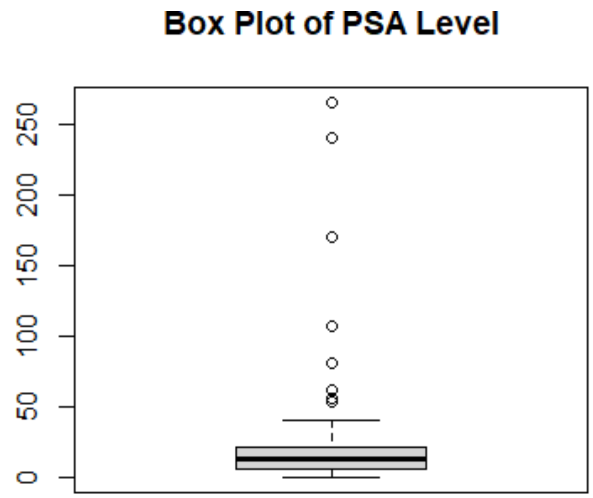
Analysis of PSA data

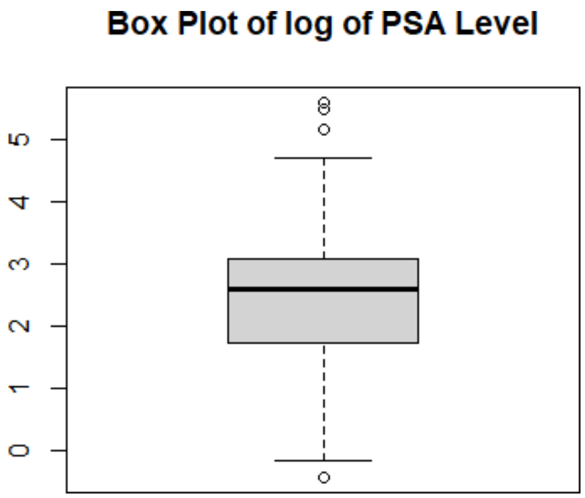


Above histogram provides us some key insights:

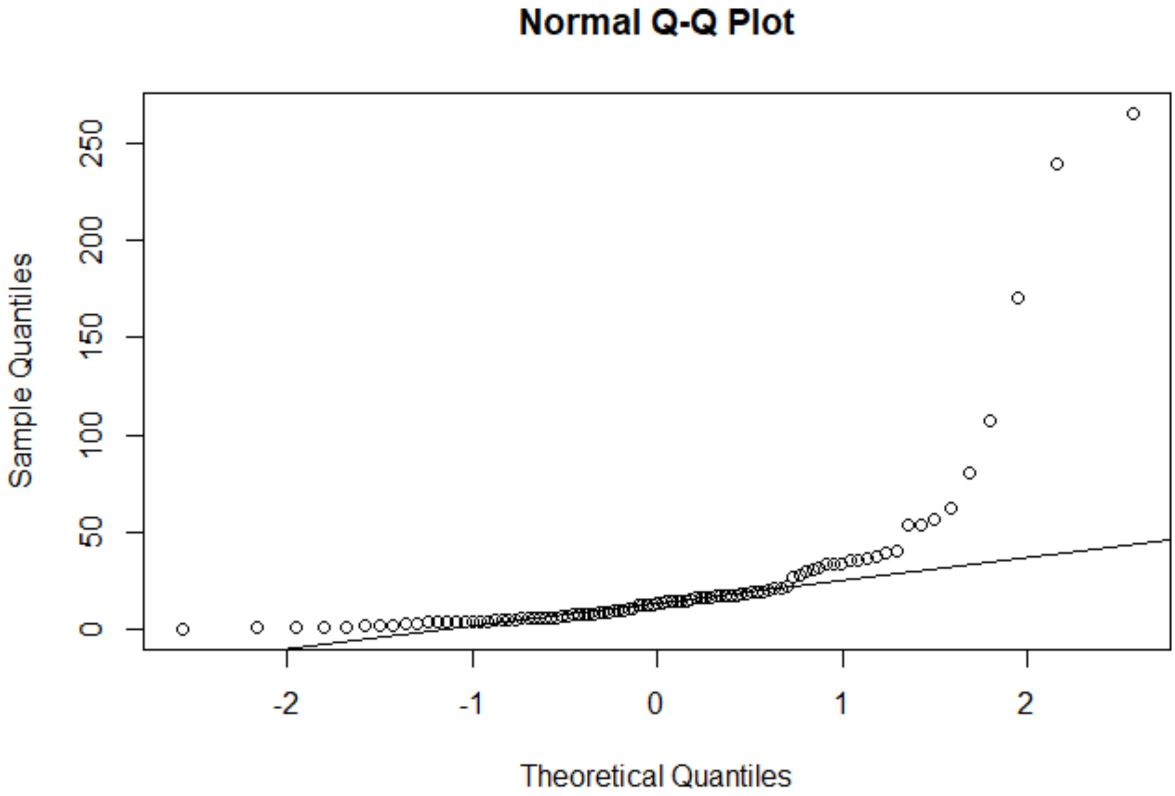
- Distribution of histogram looks like exponential
- Population is inversely proportional to PSA value i.e. as the PSA value increase number of people are less

- Majority of the people have less PSA value.



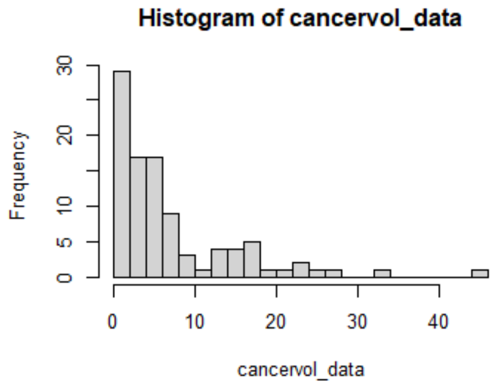


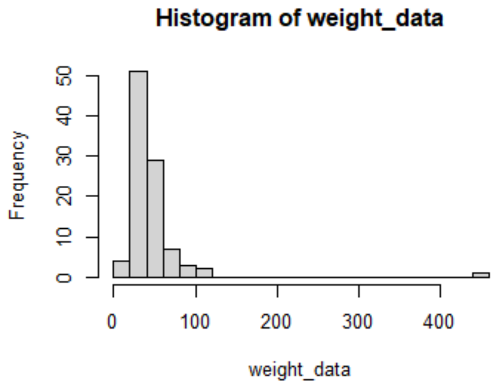
We were unable to learn much from the standard boxplot since the PSA values are so little, but we can see from the boxplot that there are a number of outliers in the data. Since the distribution of the data is symmetrical and there are relatively fewer outliers than expected, the boxplot generated by the natural log of the data will be used as the response variable.

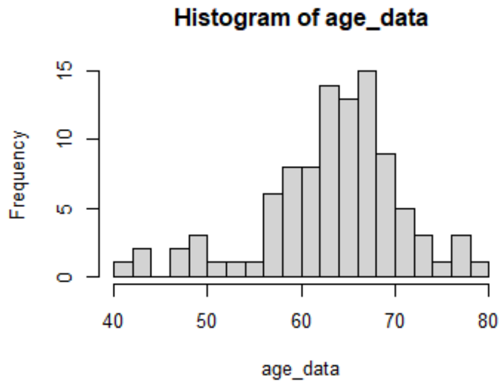


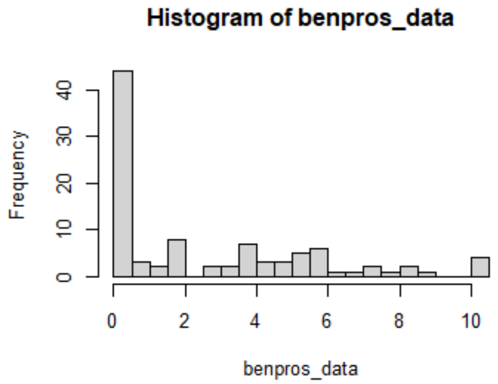
From the QQ plot we can conclude that PSA data does not follow a normal distribution because many points have a huge deviation from the QQ line.

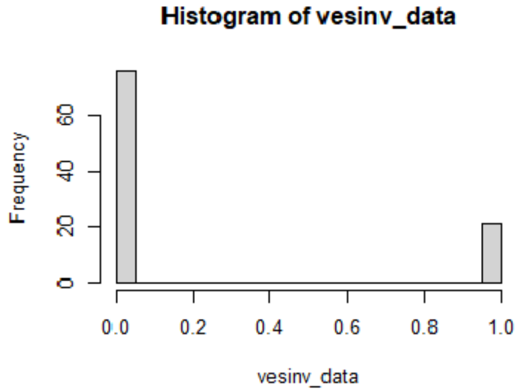
Analysis of other Quantitative data

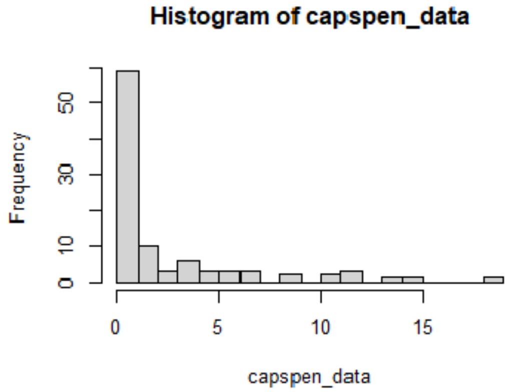

Histogram of cancervol_data

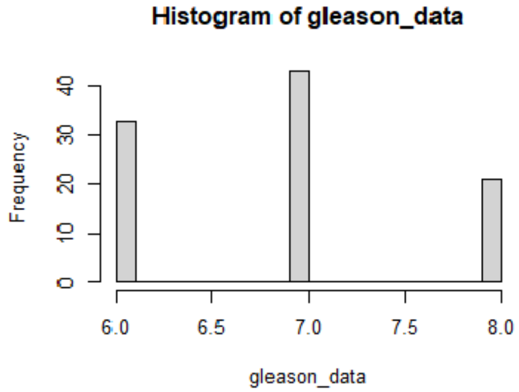

Histogram of weight_data


Histogram of age_data


Histogram of benpros_data


Histogram of vesinv_data

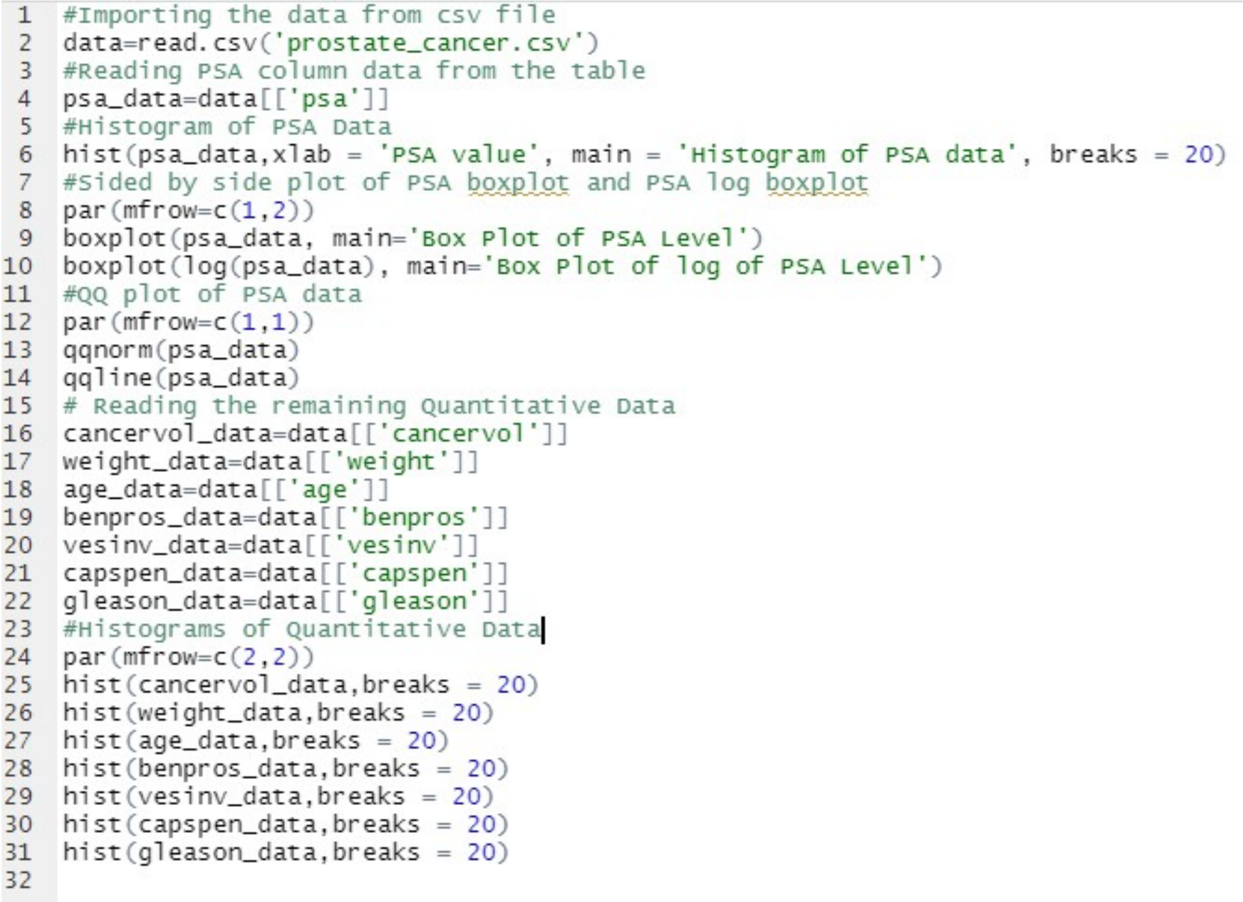

Histogram of capspen_data


Histogram of gleason_data

From the above histograms we can conclude the following:

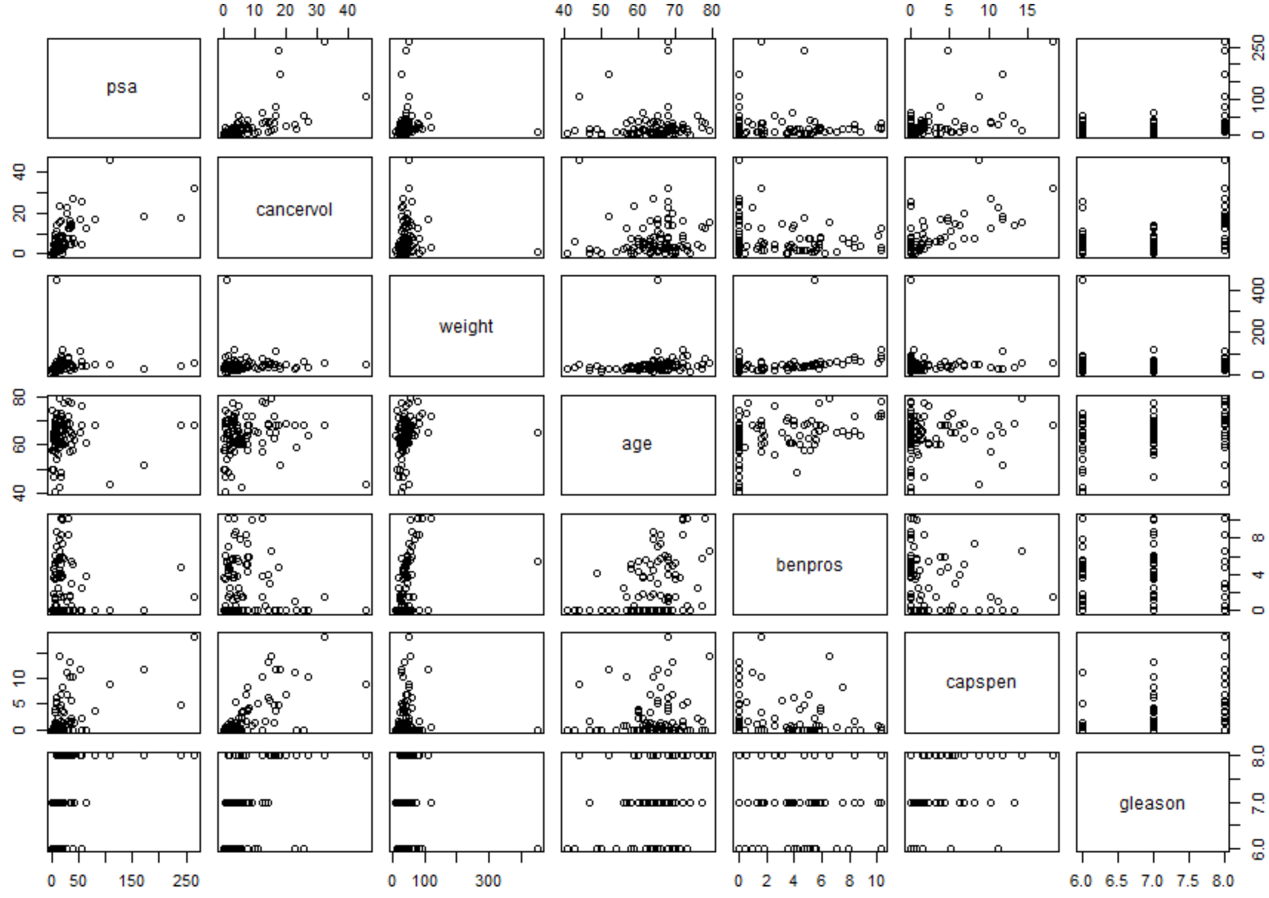
1. Cancervol: The histogram distribution resembles the histogram of PSA data. These variables might be related in a linear fashion.
2. Weight: The histogram for the weight data is less normal-looking and more gamma-like.
3. Age: The histogram of age data resembles a normal distribution.
4. Benpros: This histogram has an exponential distribution, same as the PSA and Cancervol data. A possible linear link between PSA and Cancervol data exists.
5. Vesinv: Similar to a Bernoulli variable, this variable has only two possible values: 0 or 1. In comparison to those who do not have Vesinv, fewer persons have Vesinv.
6. Caspspen: This histogram has an exponential distribution, just like PSA, Cancervol, and Benpros. A possible linear link between PSA, Cancervol, and Benpros data.
7. Gleason: This distribution has only three value 6, 7 and 8 in the histogram.

R Code:

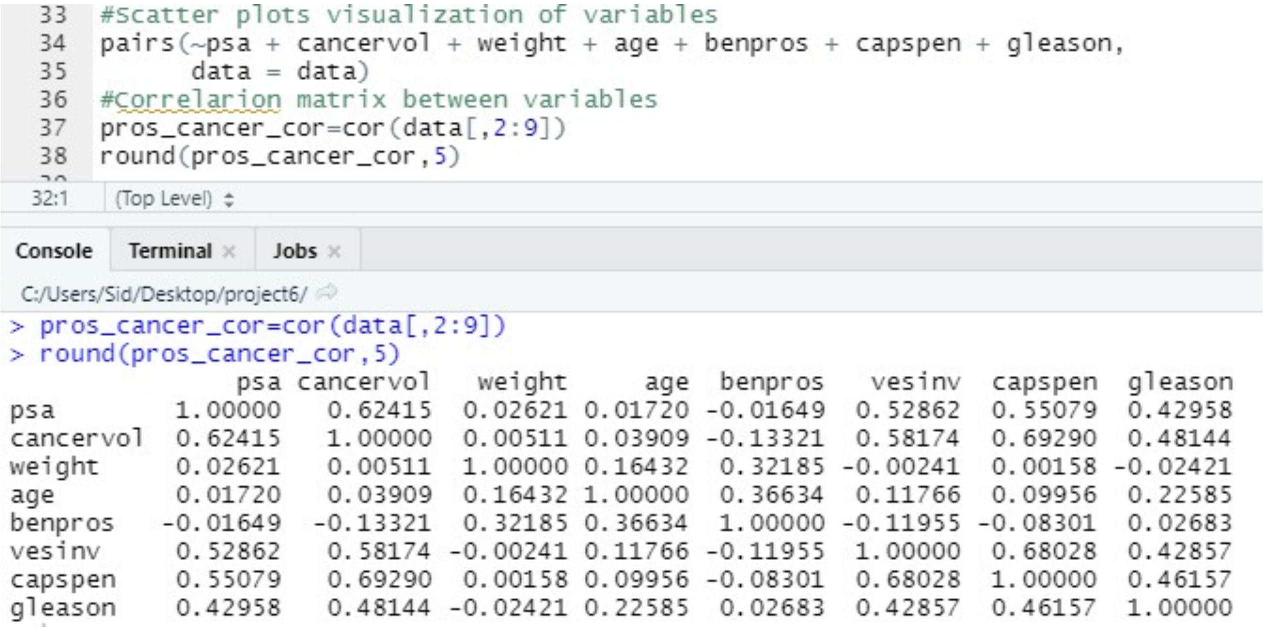


```r
#Importing the data from csv file
data=read.csv('prostate_cancer.csv')
#Reading PSA column data from the table
psa_data=data[['psa']]
#Histogram of PSA Data
hist(psa_data,xlab = 'PSA value', main = 'Histogram of PSA data', breaks = 20)
#Sided by side plot of PSA boxplot and PSA log boxplot
par(mfrow=c(1,2))
boxplot(psa_data, main='Box Plot of PSA Level')
boxplot(log(psa_data), main='Box Plot of log of PSA Level')
#QQ plot of PSA data
par(mfrow=c(1,1))
qqnorm(psa_data)
qqline(psa_data)
# Reading the remaining Quantitative Data
cancervol_data=data[['cancervol']]
weight_data=data[['weight']]
age_data=data[['age']]
benpros_data=data[['benpros']]
vesinv_data=data[['vesinv']]
capspen_data=data[['capspen']]
gleason_data=data[['gleason']]
#Histograms of Quantitative Data
par(mfrow=c(2,2))
hist(cancervol_data,breaks = 20)
hist(weight_data,breaks = 20)
hist(age_data,breaks = 20)
hist(benpros_data,breaks = 20)
hist(vesinv_data,breaks = 20)
hist(capspen_data,breaks = 20)
hist(gleason_data,breaks = 20)

```

To determine the correlation between the variables, we can use a scatter plot and a correlation matrix. These scatter plots provide us a clear picture of whether the variables have any linear relationships, which aids us in creating the best possible model.



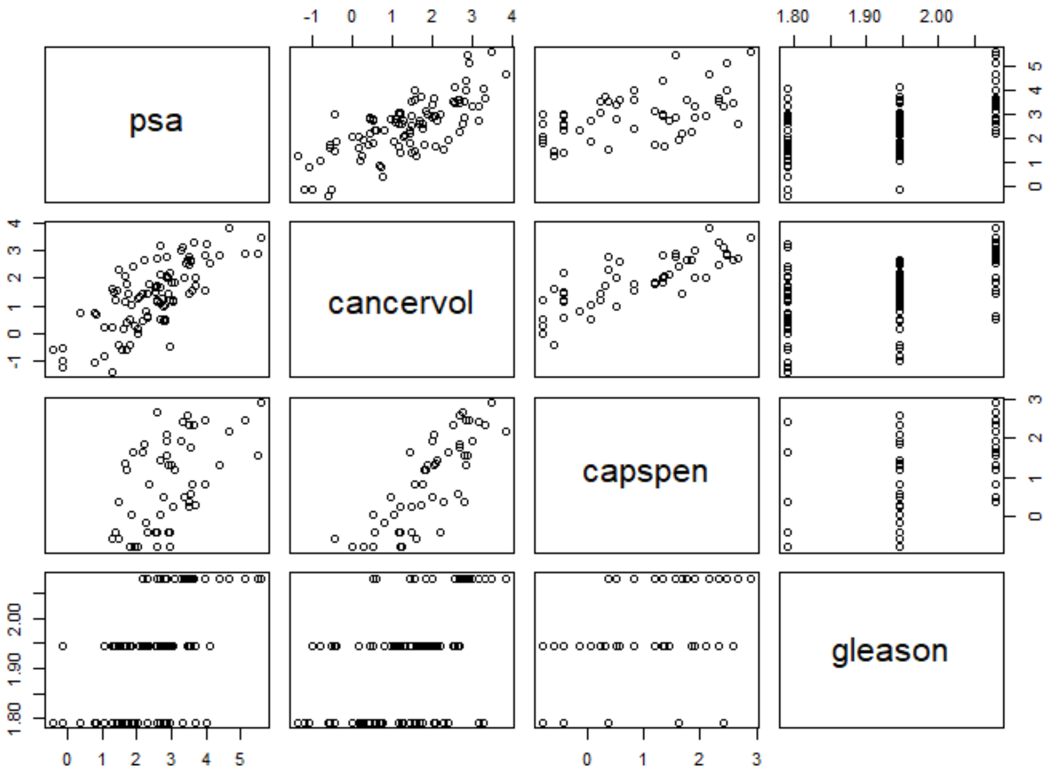
R Code:



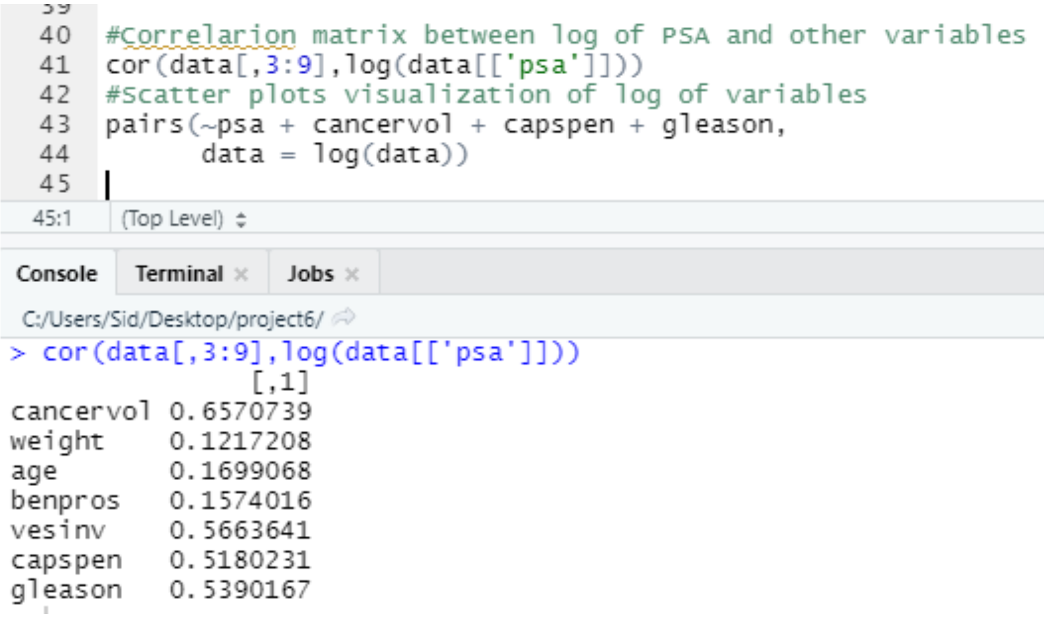
```
#Scatter plots visualization of variables
pairs(~psa + cancervol + weight + age + benpros + capspen + gleason,
      data = data)
#Correlarion matrix between variables
pros_cancer_cor=cor(data[,2:9])
round(pros_cancer_cor,5)
```

```
> pros_cancer_cor=cor(data[,2:9])
> round(pros_cancer_cor,5)
               psa cancervol   weight     age  benpros   vesinv  capspen  gleason
psa        1.00000   0.62415  0.02621 0.01720 -0.01649  0.52862  0.55079  0.42958
cancervol  0.62415   1.00000  0.00511 0.03909 -0.13321  0.58174  0.69290  0.48144
weight     0.02621   0.00511  1.00000 0.16432  0.32185 -0.00241  0.00158 -0.02421
age        0.01720   0.03909  0.16432 1.00000  0.36634  0.11766  0.09956  0.22585
benpros   -0.01649  -0.13321  0.32185 0.36634  1.00000 -0.11955 -0.08301  0.02683
vesinv     0.52862   0.58174 -0.00241 0.11766 -0.11955  1.00000  0.68028  0.42857
capspen    0.55079   0.69290  0.00158 0.09956 -0.08301  0.68028  1.00000  0.46157
gleason    0.42958   0.48144 -0.02421 0.22585  0.02683  0.42857  0.46157  1.00000
```

As predicted by the histogram analysis, we can see that there is a minor linearity between PSA, Cancervol, Caspen, and Gleason. Because the original numbers are so little, we can clearly see the picture from log converted data.



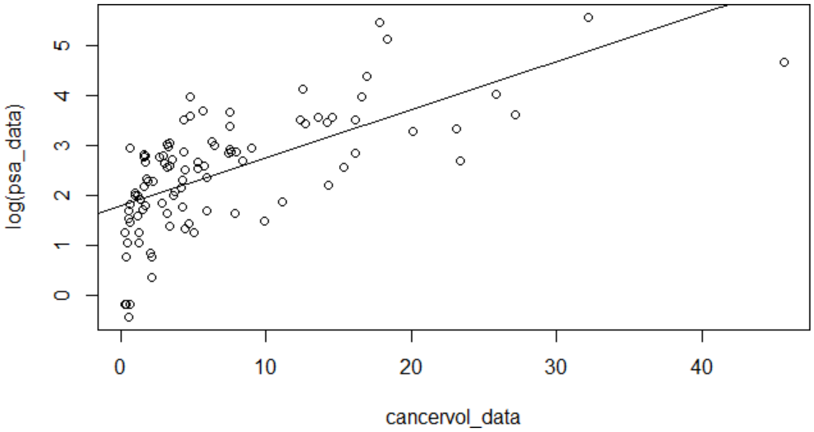
R Code:

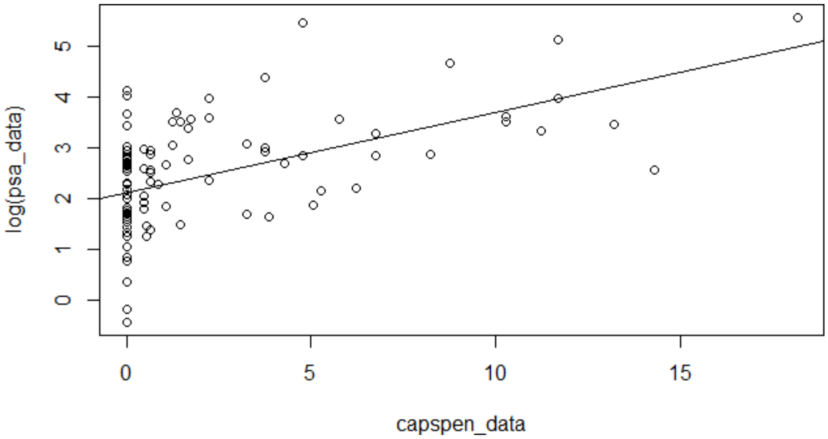

```
#Correlarion matrix between log of PSA and other variables
cor(data[,3:9],log(data[['psa']]))
#Scatter plots visualization of log of variables
pairs(~psa + cancervol + capspen + gleason,
      data = log(data))

```

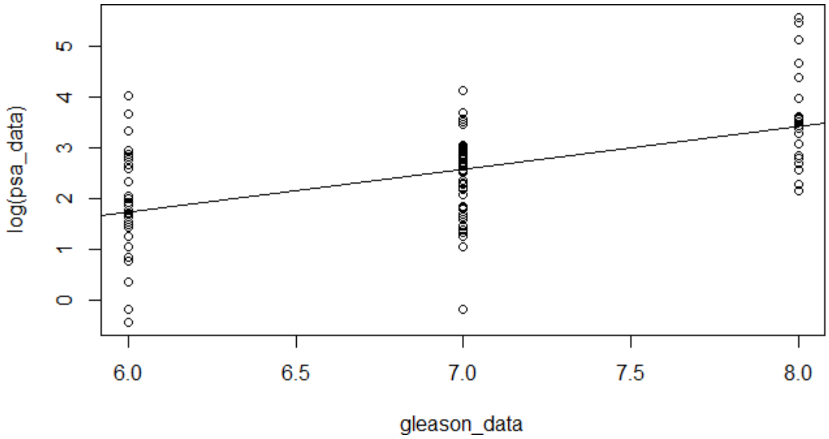
```
> cor(data[,3:9],log(data[['psa']]))
               [,1]
cancervol 0.6570739
weight    0.1217208
age       0.1699068
benpros   0.1574016
vesinv    0.5663641
capspen   0.5180231
gleason   0.5390167
```

Start by creating a straightforward linear model with the variables Cancervol, Capsen, and Gleason. Due to their low correlation values and lack of value to the model, additional variables like weight, age, and venpro can be eliminated.

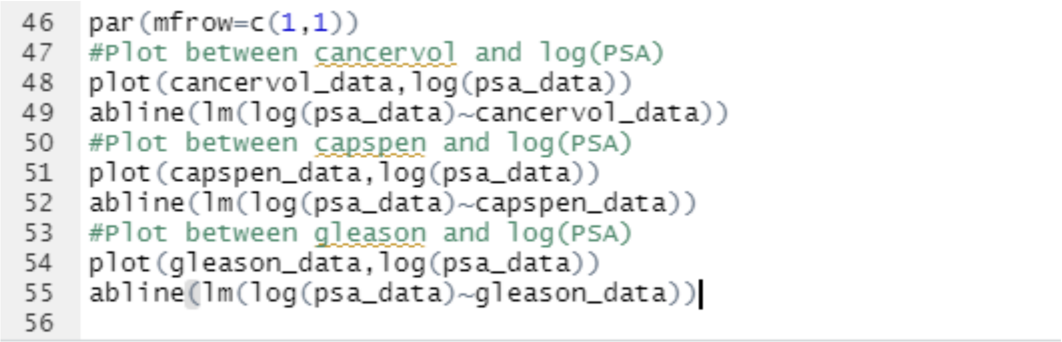
We will take a closer look at the log(psa) vs the variables which would be ideal for our linear model.





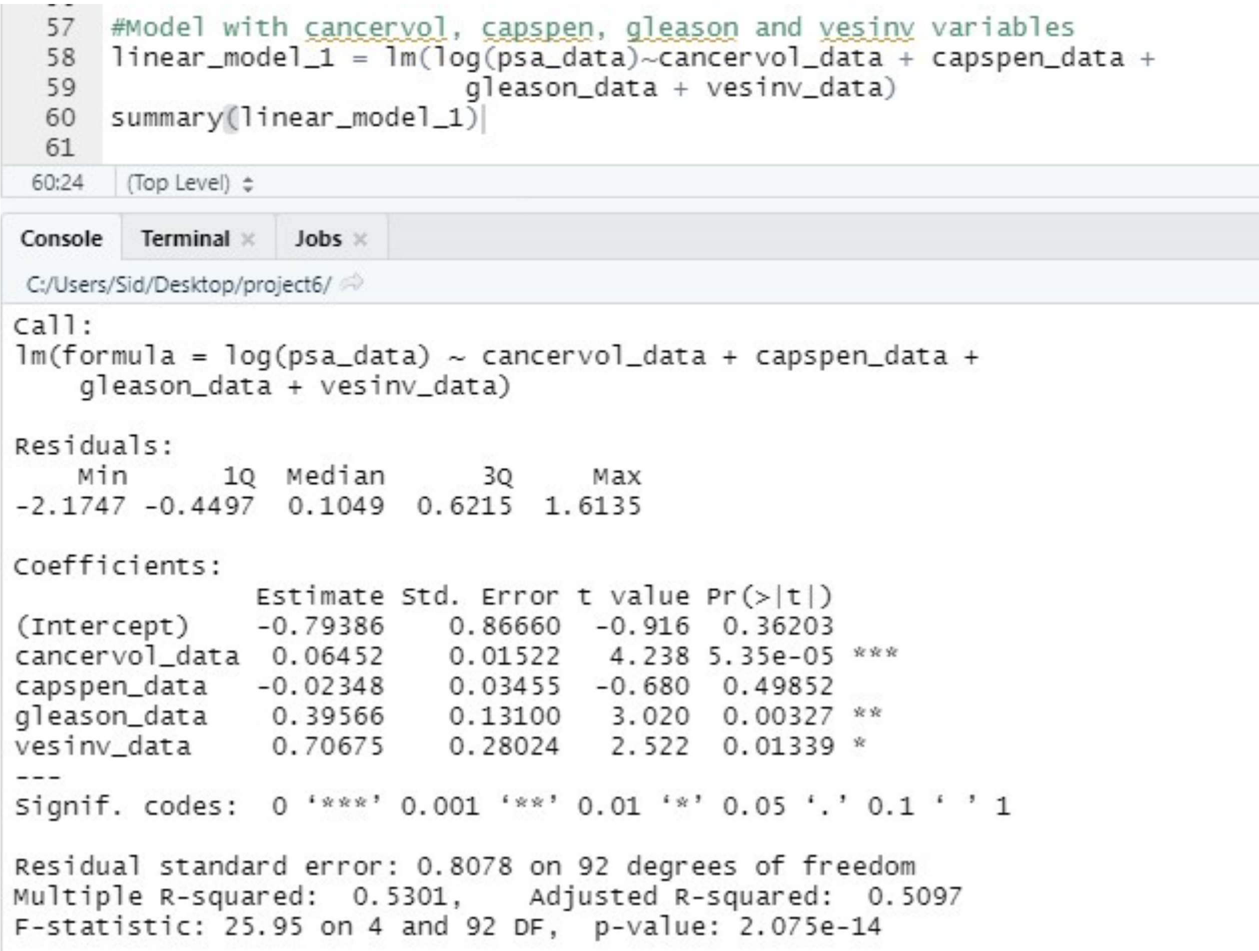


R Code:



```
par(mfrow=c(1,1))
#Plot between cancervol and log(PSA)
plot(cancervol_data,log(psa_data))
abline(lm(log(psa_data)~cancervol_data))
#Plot between capspen and log(PSA)
plot(capspen_data,log(psa_data))
abline(lm(log(psa_data)~capspen_data))
#Plot between gleason and log(PSA)
plot(gleason_data,log(psa_data))
abline(lm(log(psa_data)~gleason_data))
```

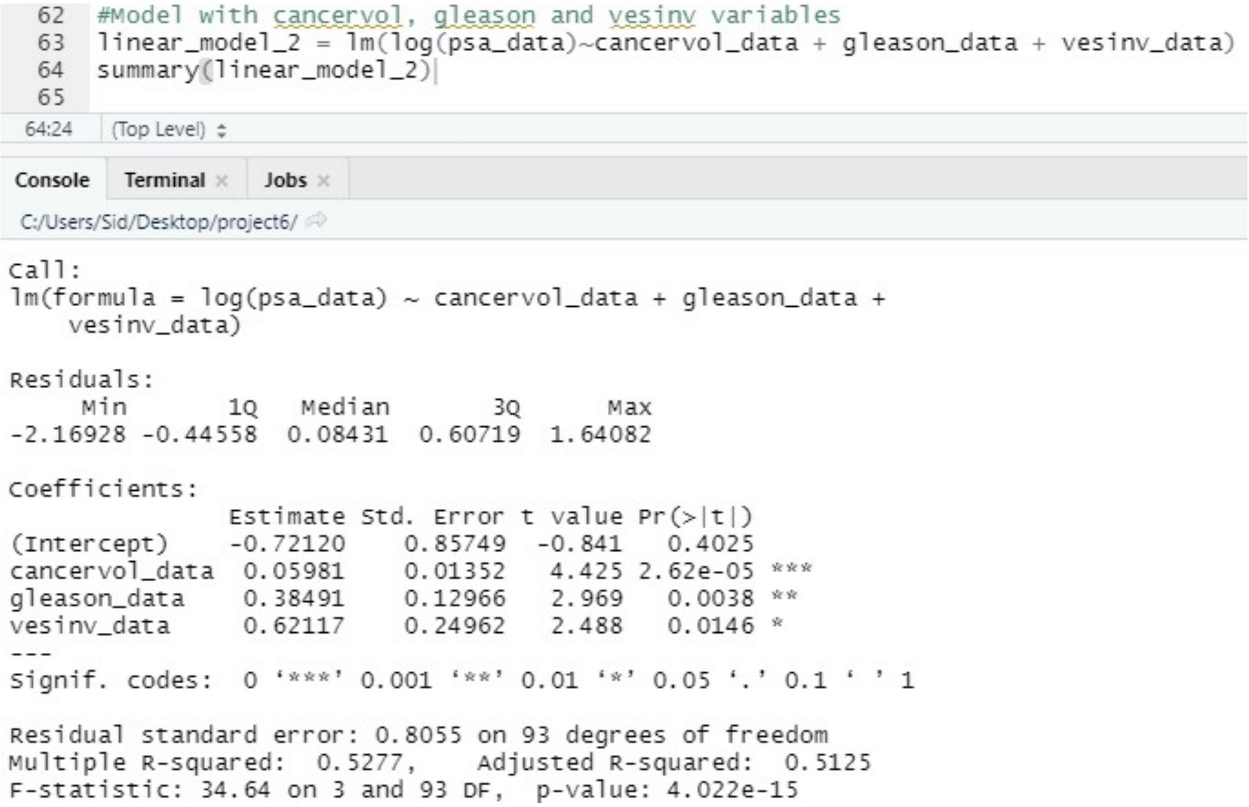
As anticipated, there is a strong positive trend between the variables and log(psa). This research allows us to construct a linear model using the quantitative predictors cancervol, capspen, and gleason. We additionally take into account the variable vesinv because of its strong connection and potential role in our model.



```
#Model with cancervol, capspen, gleason and vesinv variables
linear_model_1 = lm(log(psa_data)~cancervol_data + capspen_data +
                       gleason_data + vesinv_data)
summary(linear_model_1)
```

```
Call:
lm(formula = log(psa_data) ~ cancervol_data + capspen_data +
    gleason_data + vesinv_data)

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_data  0.06452    0.01522   4.238 5.35e-05 ***
capspen_data   -0.02348    0.03455  -0.680  0.49852
gleason_data    0.39566    0.13100   3.020  0.00327 **
vesinv_data     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
```

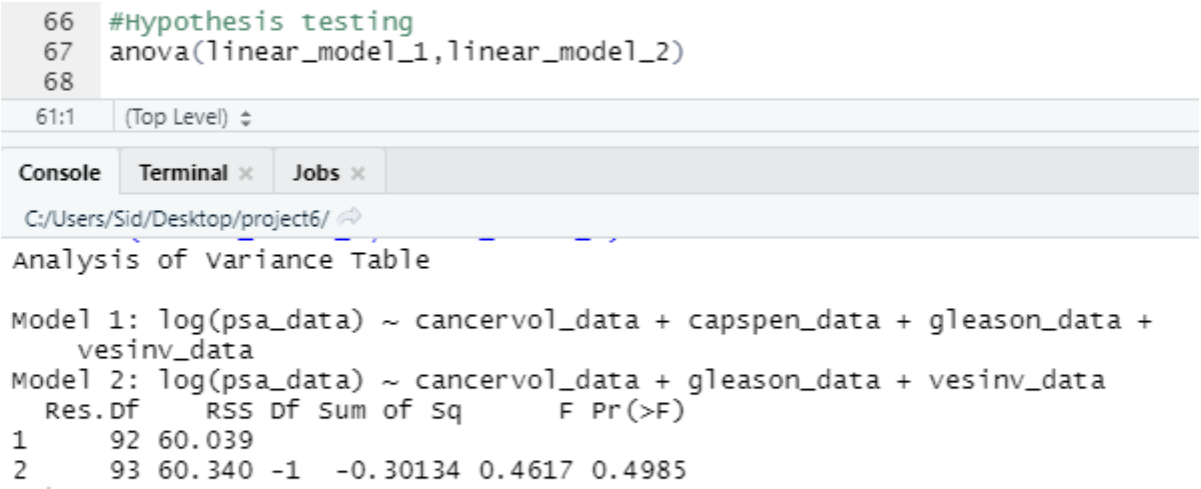
From the t-test of variable caspen we can conclude that it is not significant in our model so we can neglect and proceed with other variables.



```
#Model with cancervol, gleason and vesinv variables
linear_model_2 = lm(log(psa_data)~cancervol_data + gleason_data + vesinv_data)
summary(linear_model_2)
```

```
Call:
lm(formula = log(psa_data) ~ cancervol_data + gleason_data + 
    vesinv_data)

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_data  0.05981    0.01352   4.425 2.62e-05 ***
gleason_data    0.38491    0.12966   2.969   0.0038 ** 
vesinv_data     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
```

We perform hypothesis testing between model 1 and 2 in order to make an informed decision about the final model.



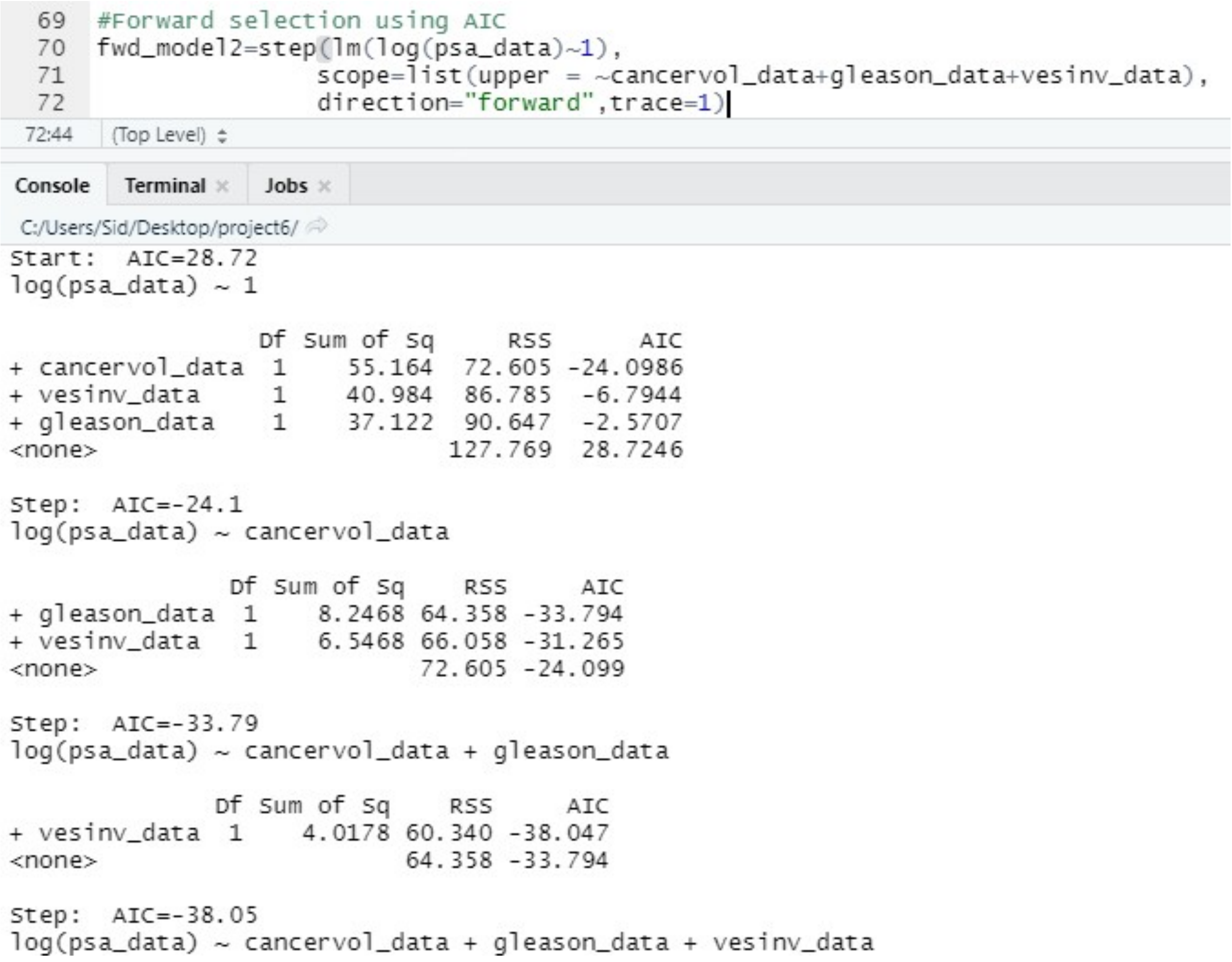
```
#Hypothesis testing
anova(linear_model_1,linear_model_2)
```

```
Analysis of Variance Table

Model 1: log(psa_data) ~ cancervol_data + capspen_data + gleason_data + 
    vesinv_data
Model 2: log(psa_data) ~ cancervol_data + gleason_data + vesinv_data
  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
```

Given that the null hypothesis is accepted and the partial F-static p value is high, we came to the conclusion that the simplified model is preferable. We tried further reducing model 2, but we were unable to discover compelling evidence to do so.

In order to compare model 2, we go on to model 2 and do an automatic stepwise selection based on AIC.

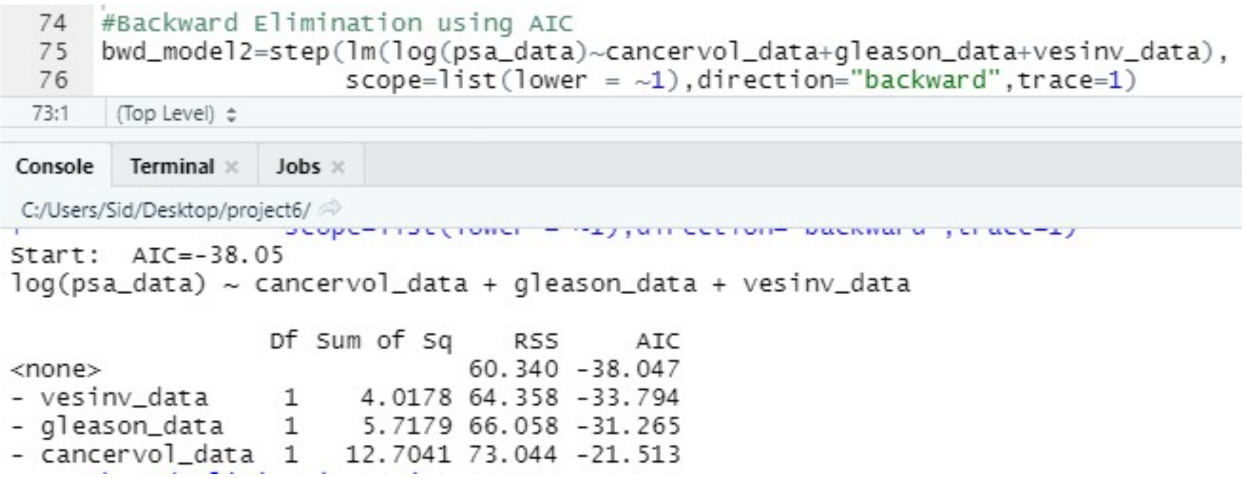
Forward selection using AIC:



```
#Forward selection using AIC
fwd_model2=step(lm(log(psa_data)~1),
                scope=list(upper = ~cancervol_data+gleason_data+vesinv_data),
                direction="forward",trace=1)
```

```
Start:  AIC=28.72
log(psa_data) ~ 1

                 Df Sum of Sq     RSS      AIC
+ cancervol_data  1    55.164  72.605 -24.0986
+ vesinv_data     1    40.984  86.785  -6.7944
+ gleason_data    1    37.122  90.647  -2.5707
<none>                        127.769  28.7246

Step:  AIC=-24.1
log(psa_data) ~ cancervol_data

               Df Sum of Sq    RSS     AIC
+ gleason_data  1    8.2468 64.358 -33.794
+ vesinv_data   1    6.5468 66.058 -31.265
<none>                      72.605 -24.099

Step:  AIC=-33.79
log(psa_data) ~ cancervol_data + gleason_data

              Df Sum of Sq    RSS     AIC
+ vesinv_data  1    4.0178 60.340 -38.047
<none>                     64.358 -33.794

Step:  AIC=-38.05
log(psa_data) ~ cancervol_data + gleason_data + vesinv_data
```

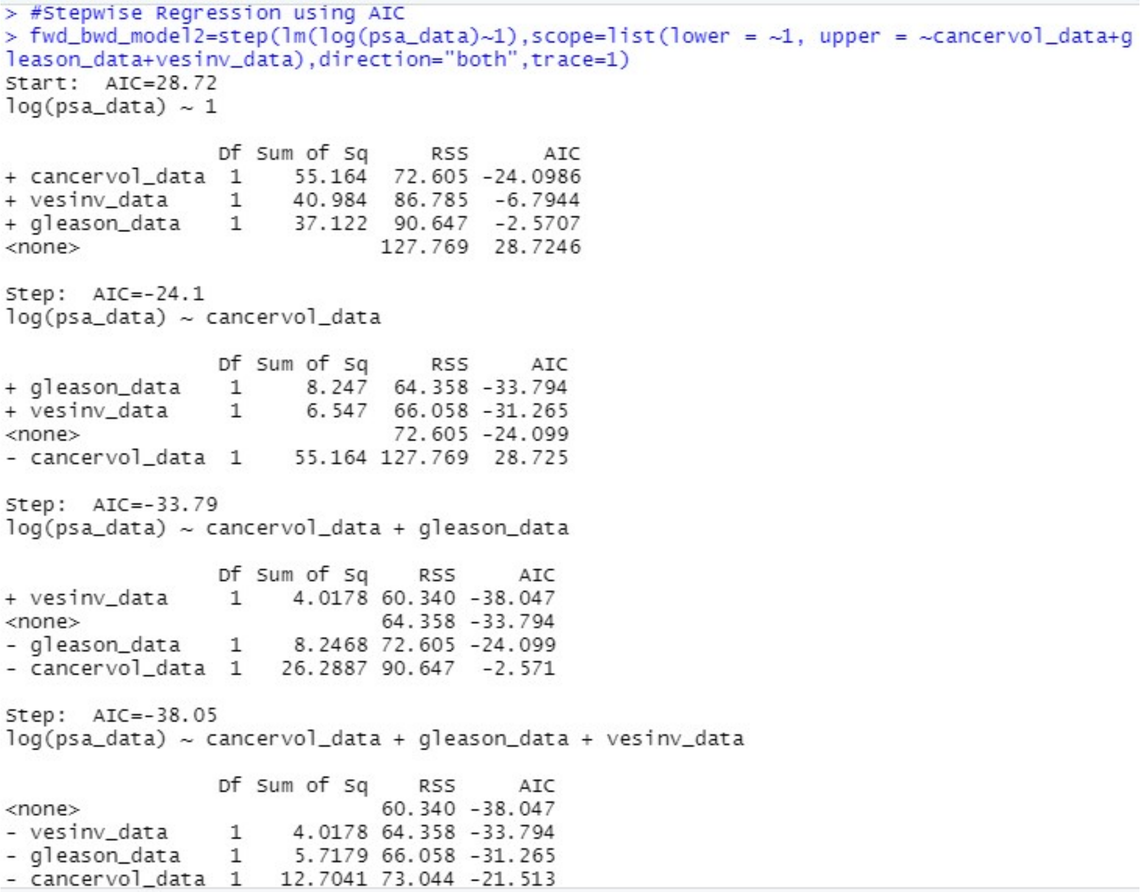
Backward Elimination using AIC:



```
#Backward Elimination using AIC
bwd_model2=step(lm(log(psa_data)~cancervol_data+gleason_data+vesinv_data),
                scope=list(lower = ~1),direction="backward",trace=1)
```

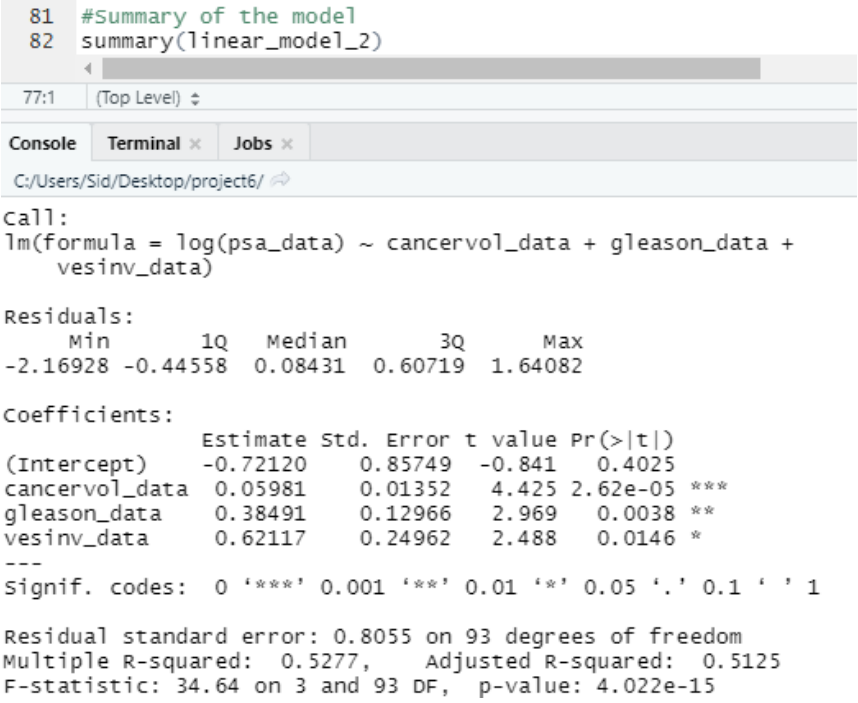
```
                scope=list(lower = ~1),direction="backward",trace=1)
Start:  AIC=-38.05
log(psa_data) ~ cancervol_data + gleason_data + vesinv_data

                 Df Sum of Sq    RSS     AIC
<none>                        60.340 -38.047
- vesinv_data     1    4.0178 64.358 -33.794
- gleason_data    1    5.7179 66.058 -31.265
- cancervol_data  1   12.7041 73.044 -21.513
```

Stepwise Regression using AIC:



```
> #Stepwise Regression using AIC
> fwd_bwd_model2=step(lm(log(psa_data)~1),scope=list(lower = ~1, upper = ~cancervol_data+g
leason_data+vesinv_data),direction="both",trace=1)
Start:  AIC=28.72
log(psa_data) ~ 1

                 Df Sum of Sq     RSS      AIC
+ cancervol_data  1    55.164  72.605 -24.0986
+ vesinv_data     1    40.984  86.785  -6.7944
+ gleason_data    1    37.122  90.647  -2.5707
<none>                        127.769  28.7246

Step:  AIC=-24.1
log(psa_data) ~ cancervol_data

                 Df Sum of Sq     RSS     AIC
+ gleason_data    1     8.247  64.358 -33.794
+ vesinv_data     1     6.547  66.058 -31.265
<none>                         72.605 -24.099
- cancervol_data  1    55.164 127.769  28.725

Step:  AIC=-33.79
log(psa_data) ~ cancervol_data + gleason_data

                 Df Sum of Sq    RSS     AIC
+ vesinv_data     1    4.0178 60.340 -38.047
<none>                        64.358 -33.794
- gleason_data    1    8.2468 72.605 -24.099
- cancervol_data  1   26.2887 90.647  -2.571

Step:  AIC=-38.05
log(psa_data) ~ cancervol_data + gleason_data + vesinv_data

                 Df Sum of Sq    RSS     AIC
<none>                        60.340 -38.047
- vesinv_data     1    4.0178 64.358 -33.794
- gleason_data    1    5.7179 66.058 -31.265
- cancervol_data  1   12.7041 73.044 -21.513
```

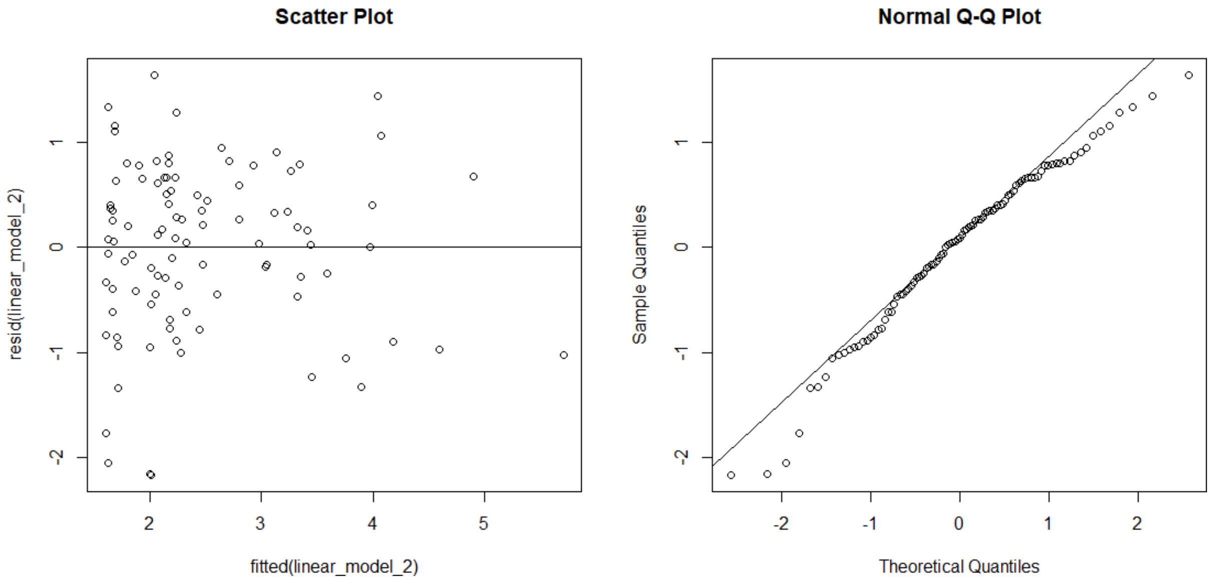
Summary:



```
#Summary of the model
summary(linear_model_2)
```

```
Call:
lm(formula = log(psa_data) ~ cancervol_data + gleason_data +
    vesinv_data)

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_data  0.05981    0.01352   4.425 2.62e-05 ***
gleason_data    0.38491    0.12966   2.969   0.0038 **
vesinv_data     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
```

Residual Plots:



We may verify that the normality assumption is true for residuals by looking at the QQ plot. We may infer that all of our model assumptions hold true, thus we take this to be our final model.