```
#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))

#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)


#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

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)

#summary of the model
summary(fit4)


#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)
#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)
#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)
#Evidence against vesinv is also not strong enough
#Hence we accept fit5 as a preliminary model
summary(fit5)

#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")

ackward elimination based on AIC
fit9.backward <- step(lm(y~cancervol+capspen+gleason+vesinv, data = MyData2),scope
= list(lower = ~1), direction = "backward")

#Both forward and backward
fit10.both <- step(lm(y ~ 1, data = MyData2),scope = list(lower = ~1, upper =
~cancervol+capspen+gleason+vesinv),direction = "both")


#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)


#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.
```