Data Analytics Mini Project 5

M.Tech- Group 9.2

Name of the Student: Vijay Deshpande
Student ID: 201761003
Role: Question 1

Name of the Student: Jayant Mahawar
Student ID: 201761004
Role: Question 2 and 3

Exercise 1:
Take PSA level is as the response variable. Make scatterplots of PSA level with other variables. Based on these, choose one quantitative variable that you think may be used effectively to predict PSA level. Highlight any potential outliers on the scatterplot of this variable with PSA level.

Solution:
Scatter plot of response variable with other variable are shown below.

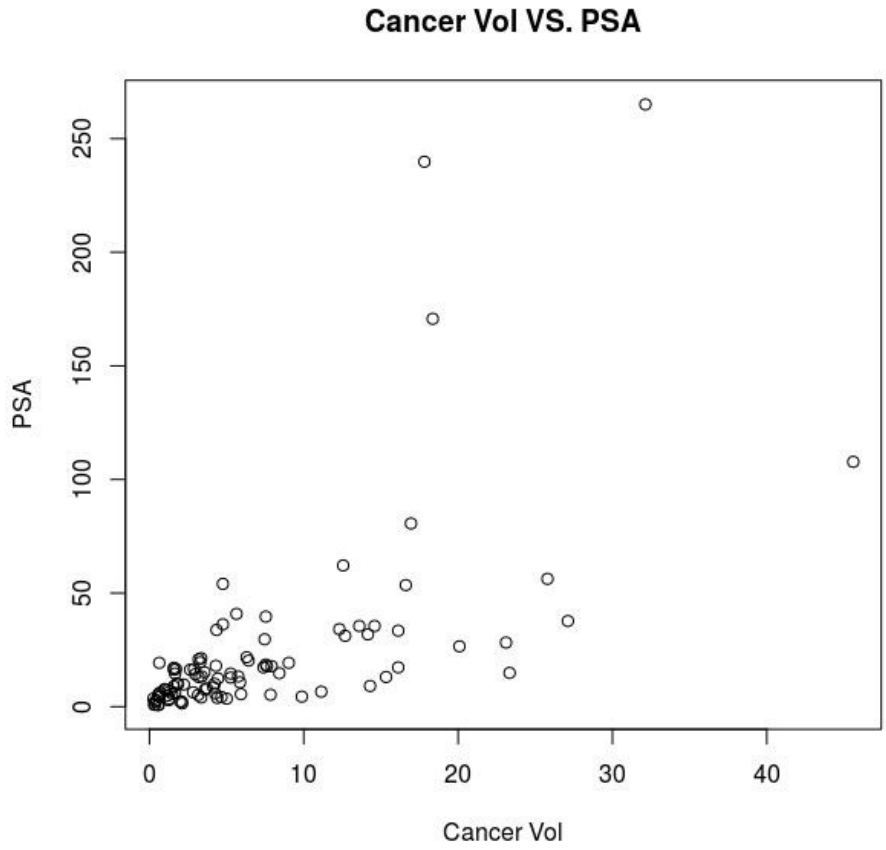

Fig1: Cancer Vol vs. PSA

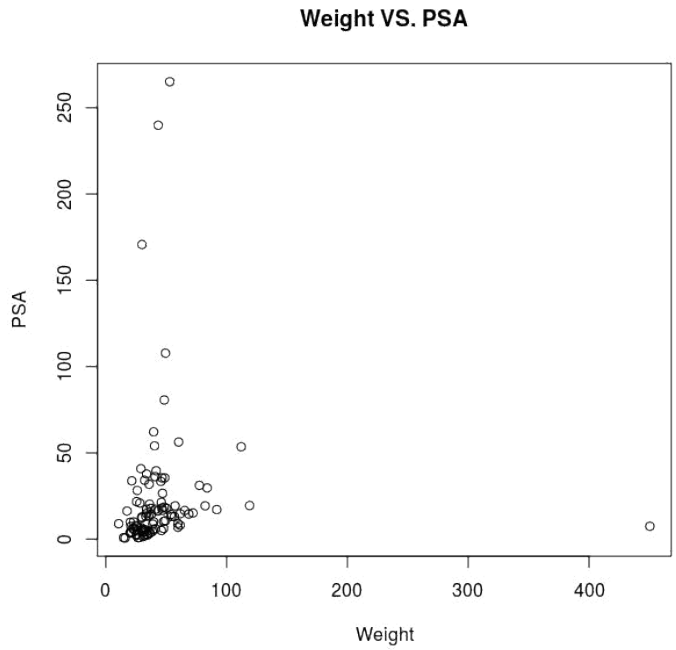

Fig 2:Weight vs. PSA

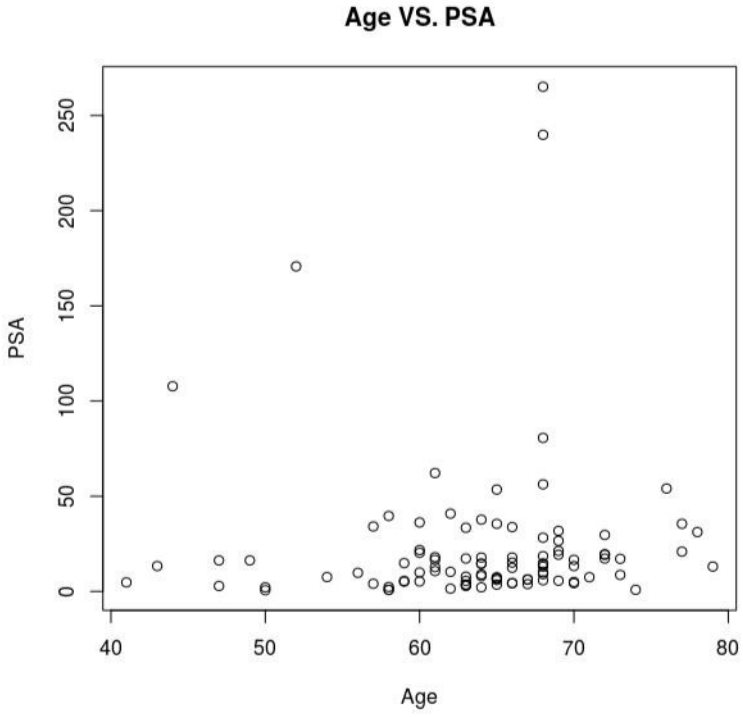

Fig 3:Age vs. PSA

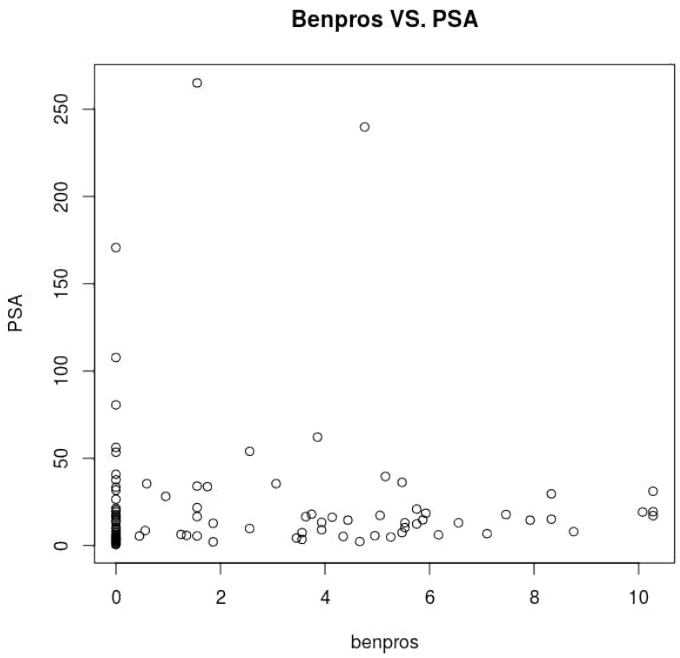

Fig 4:Benpros vs. PSA

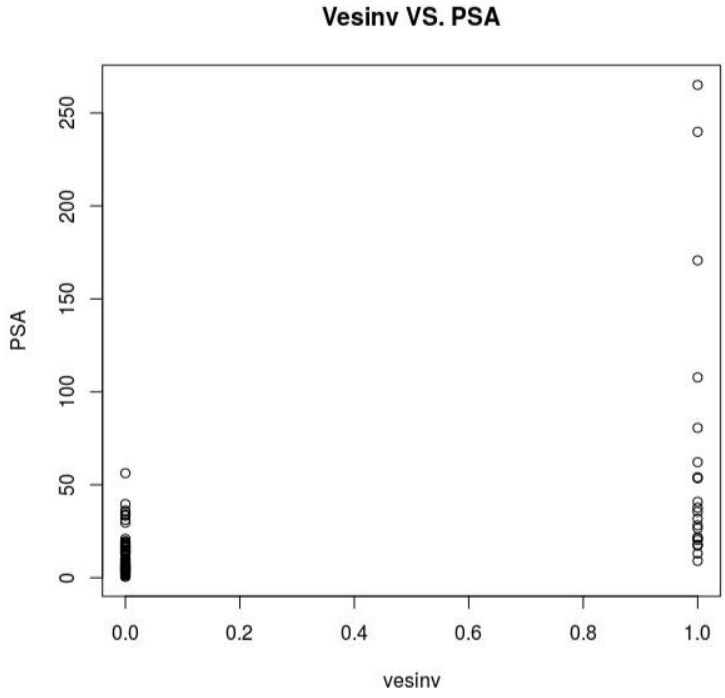

Fig 5:Vensiv vs. PSA

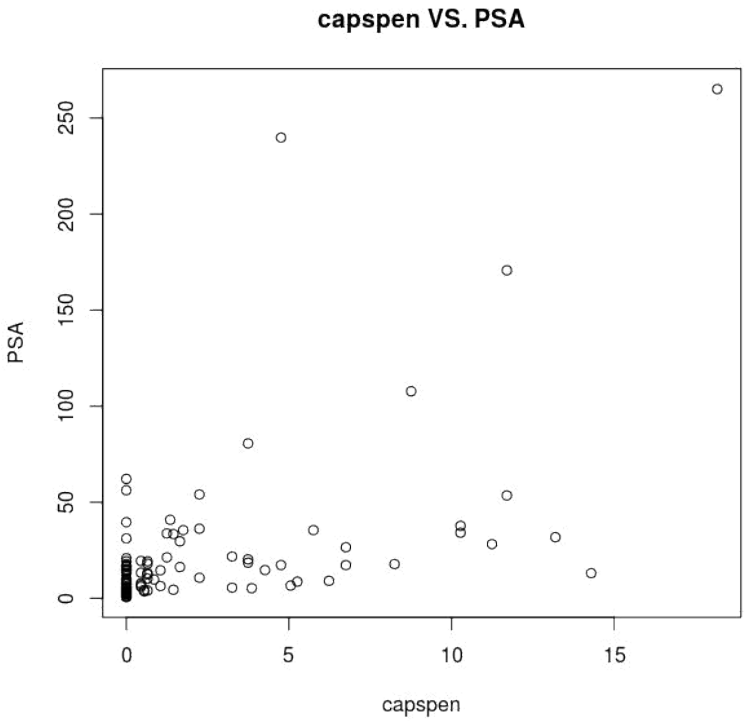

Fig 6:Capsepen vs. PSA

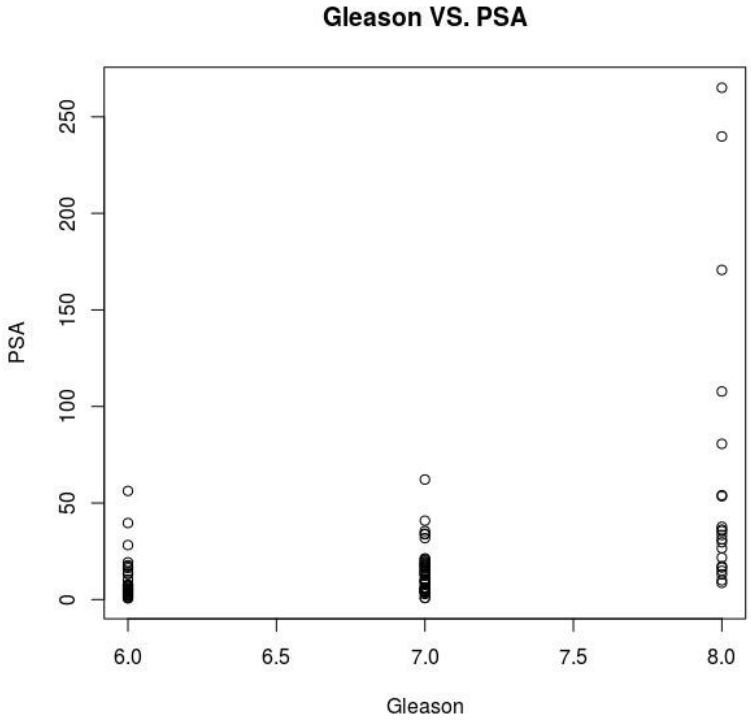

Fig 7:Gleason vs. PSA

Here correlation between PSA and other variables are shown below. As per correlation value, cancervol is highly correlated with PSA. We consider cancervol as a quantitative variable. Fig1 shows a linear increase in PSA level with cancervol.

Correlation Value:

	psa	cancervol	weight	age	benpros	vesinv
psa	1.00000	0.624150588	0.026213430	0.01719938	-0.01648649	0.528618785

	capspen	gleason
psa	0.528618785	0.550792517

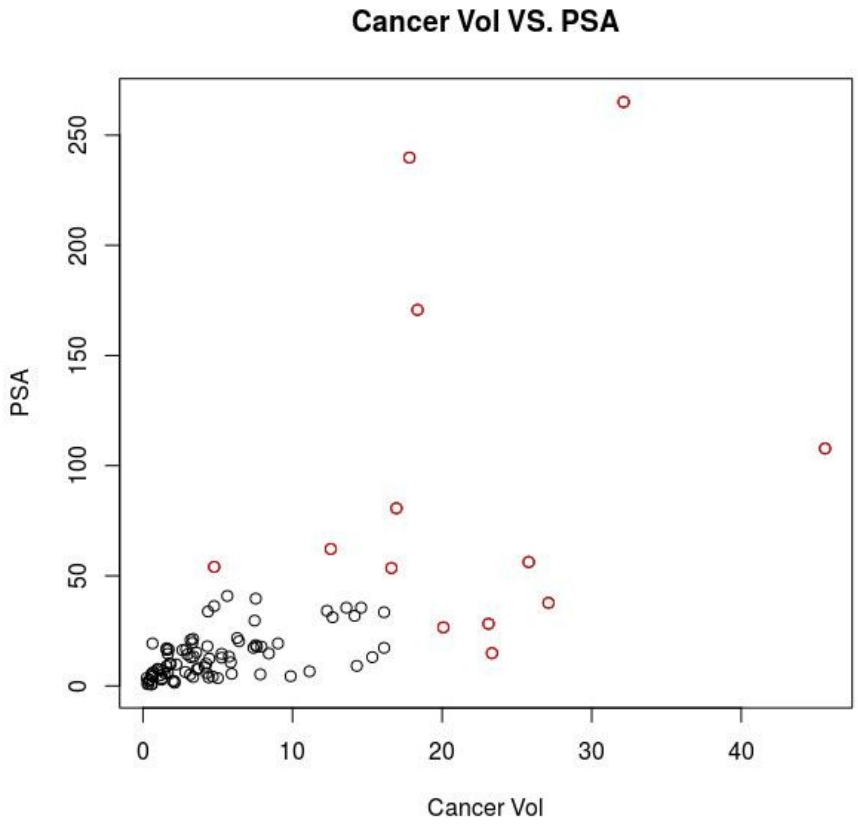

Fig 8: Outliers in Cancer Vol vs. PSA

In given data PSA level has a outlier data which are greater then 44.85 and for cancervol data value greater then 18.5393 is outlier. Here Red colors points shows outliers in data.

R-code:

```
#read data from File
cancer=read.csv("/home/vijay/prostate_cancer.csv")
cancer

# Make a scatterplot

#Cancer Vol vs. PSA
x=cancer$cancervol
y=cancer$psa
plot(x, y, xlab="Cancer Vol", ylab="PSA",main="Cancer Vol VS. PSA")

#Weight vs. PSA
x=cancer$weight
y=cancer$psa
plot(x, y, xlab="Weight", ylab="PSA",main="Weight VS. PSA")

#Age vs. PSA
x=cancer$age
y=cancer$psa
plot(x, y, xlab="Age", ylab="PSA",main="Age VS. PSA")

#Benpros Vs. PSA
x=cancer$benpros
y=cancer$psa
plot(x, y, xlab="benpros", ylab="PSA",main="Benpros VS. PSA")

#Vesinv Vs. PSA
x=cancer$vesinv
y=cancer$psa
plot(x, y, xlab="vesinv", ylab="PSA",main="Vesinv VS. PSA")

#Capspen Vs. PSA
x=cancer$capspen
y=cancer$psa
plot(x, y, xlab="capspen", ylab="PSA",main="capspen VS. PSA")

#Gleason Vs. PSA
x=cancer$gleason
y=cancer$psa
plot(x, y, xlab="gleason", ylab="PSA",main="gleason VS. PSA")

#Correlation beetween variables.
cor(cancer[,c(2,3,4,5,6,7,8,9)])

#outliers
out=quantile(cancer$psa,0.75)+1.5*IQR(cancer$psa)
out1=quantile(cancer$cancervol,0.75)+1.5*IQR(cancer$cancervol)
ca=subset(cancer,cancer$psa>out | cancer$cancervol>out1)
ca
x=cancer$cancervol
y=cancer$psa
plot(x, y, xlab="Cancer Vol", ylab="PSA",main="Cancer Vol VS. PSA")
points(ca$cancervol,ca$psa,col="RED")
```

Exercise 2 :

Fit a simple linear regression model and carry out regression diagnostics. The analysis should include an assessment of the degree to which the key regression assumptions are satisfied. If an assumption is not met, attempt to remedy the situation. Comment on the fit of the final model using appropriate tests and statistics.

Solution:

Fig 9 shows regression model with PSA and Cancer Vol.

Regression assumptions are:

1. Errors have Mean Zero and constant variance (Residual Plot Test)
2. Errors are Normally distributed(Normal QQ Plot)
3. Errors are independent (Time Series Plot)

Fig 10 shows residual plot of assumed model. This figure shows that mean is almost zero but variance is different which is not constant. Fig 11 shows normality assumption which is not hold for this model. Data is not normally distributed. This model is not going to work as Adjusted R-squared: 0.3831 value is also less then 0.4. So,This model is not good fit

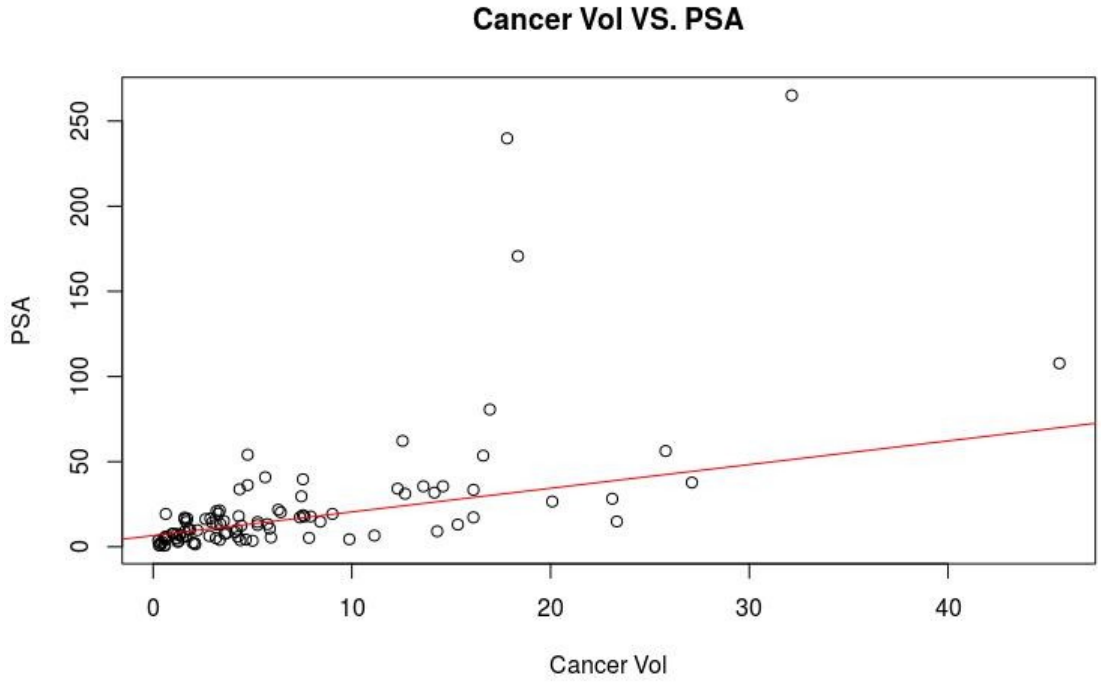

Fig 9: Regression Model (PSA ~ Cancer vol)

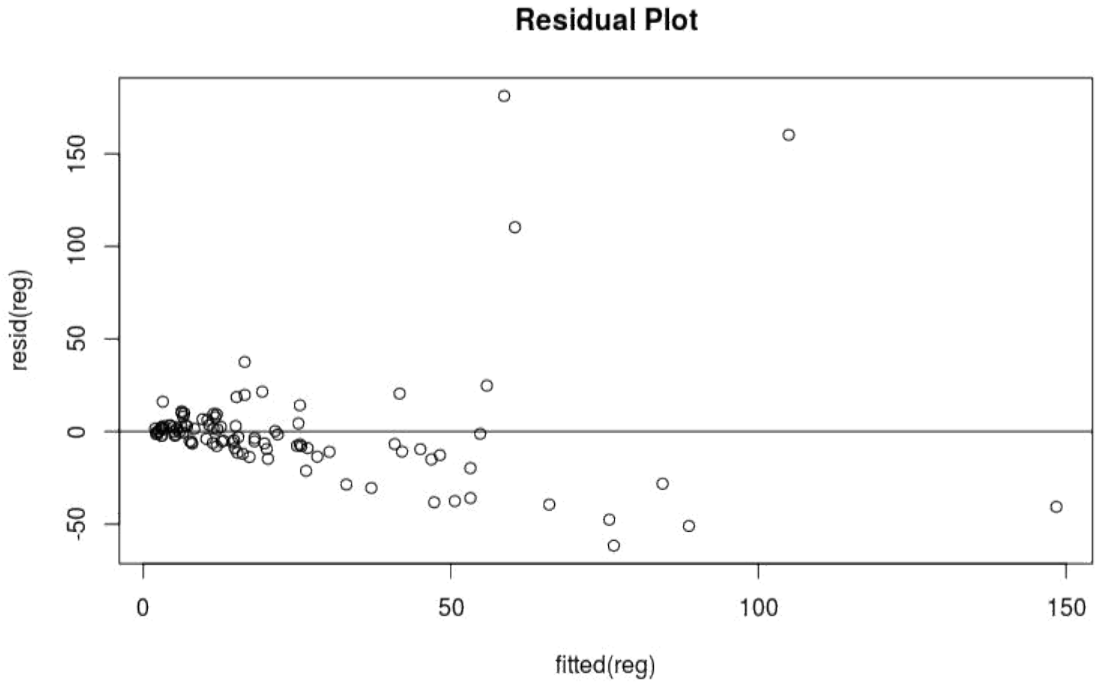

Fig 10: Residual Plot

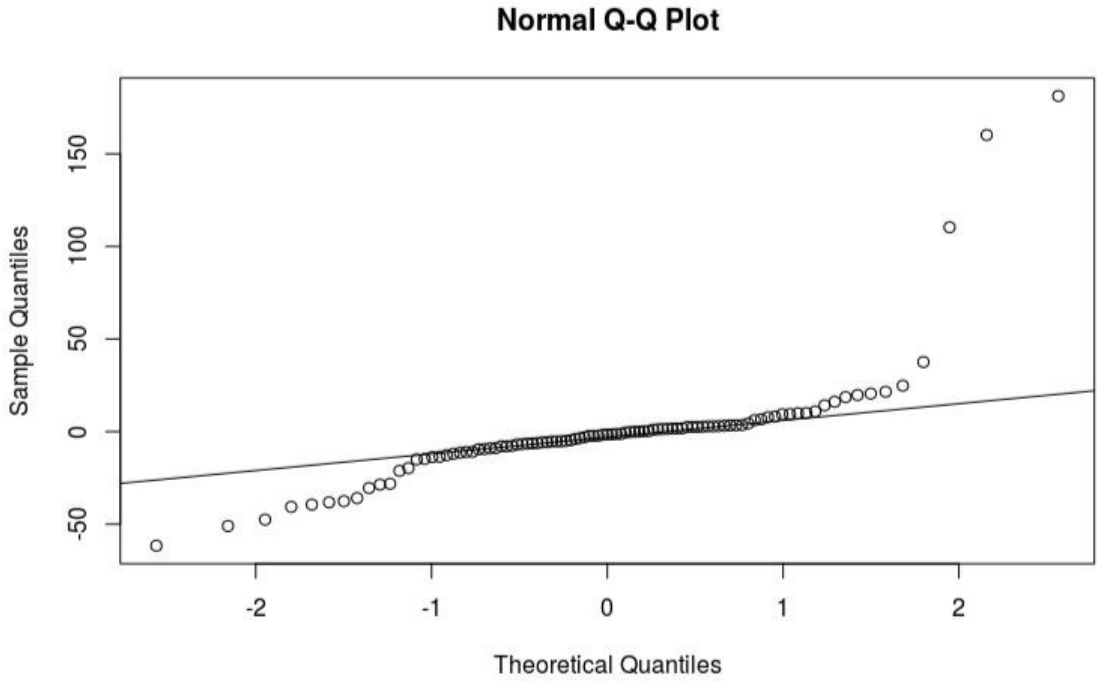

Fig 11: Q-Q Plot

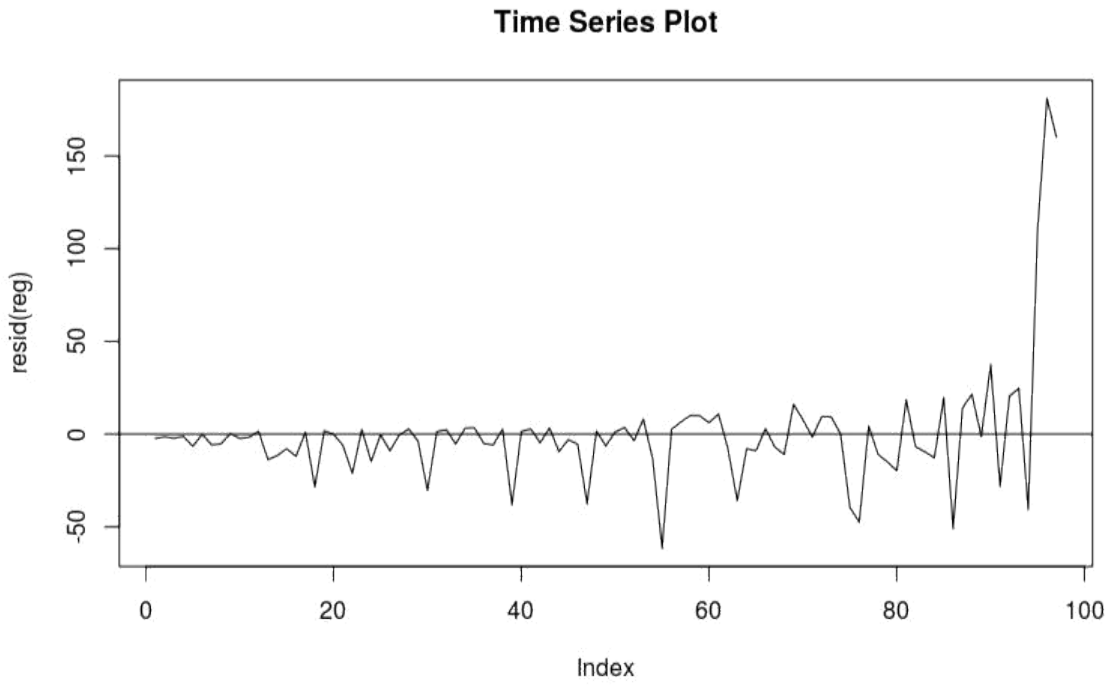

Fig 11: Time Series Plot

Fig 12 shows fitted regression line with Log(PSA) ~ Cancer Vol . Fig 13,Fig 14 and Fig 15 shows that all regression assumption holds true. Even Adjusted R-squared: 0.4258 value which shows that this model is good fit. Fig 13: shows Errors have Mean Zero constant variance. Fig 14: shows Error are Normally distributed as most of the points are on the line. Fig 15: shows Error are independent as there is no trend in plot.

Compare the p-value (2.688e-13)for the F-test to significance level 0.05. If the p-value is less than the significance level,sample data provide sufficient evidence to conclude that regression model fits the data better than the model with no independent variables.

Hence the independent variables in your model improve the fit!

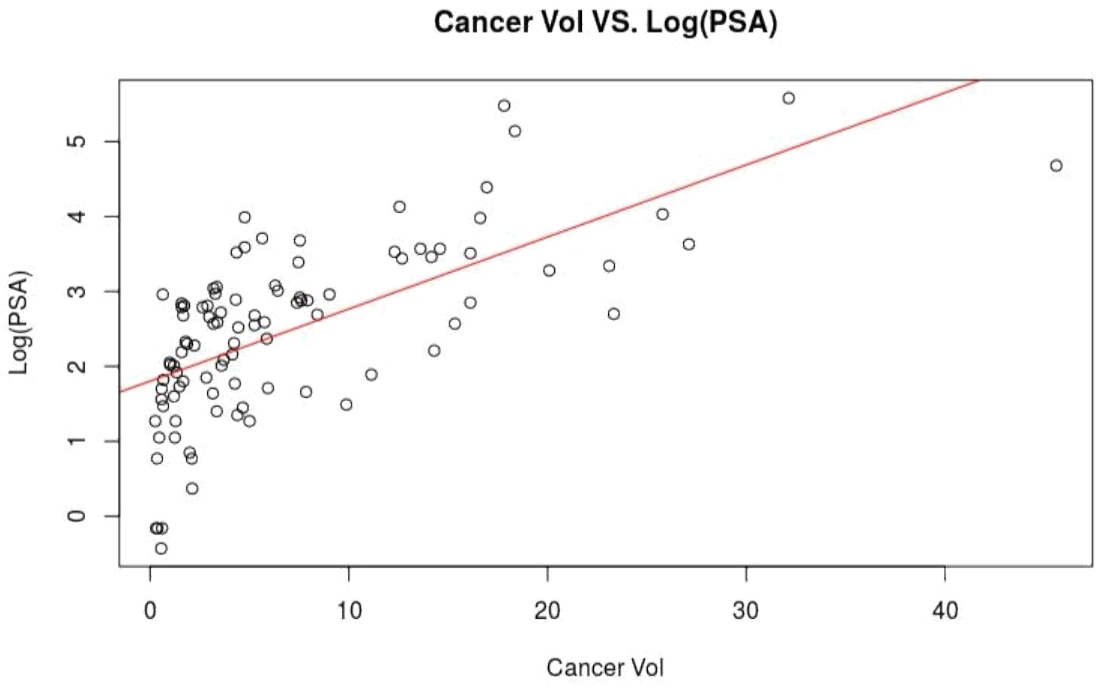

Fig 12: Regression Model (Log(PSA) ~ Cancer vol)

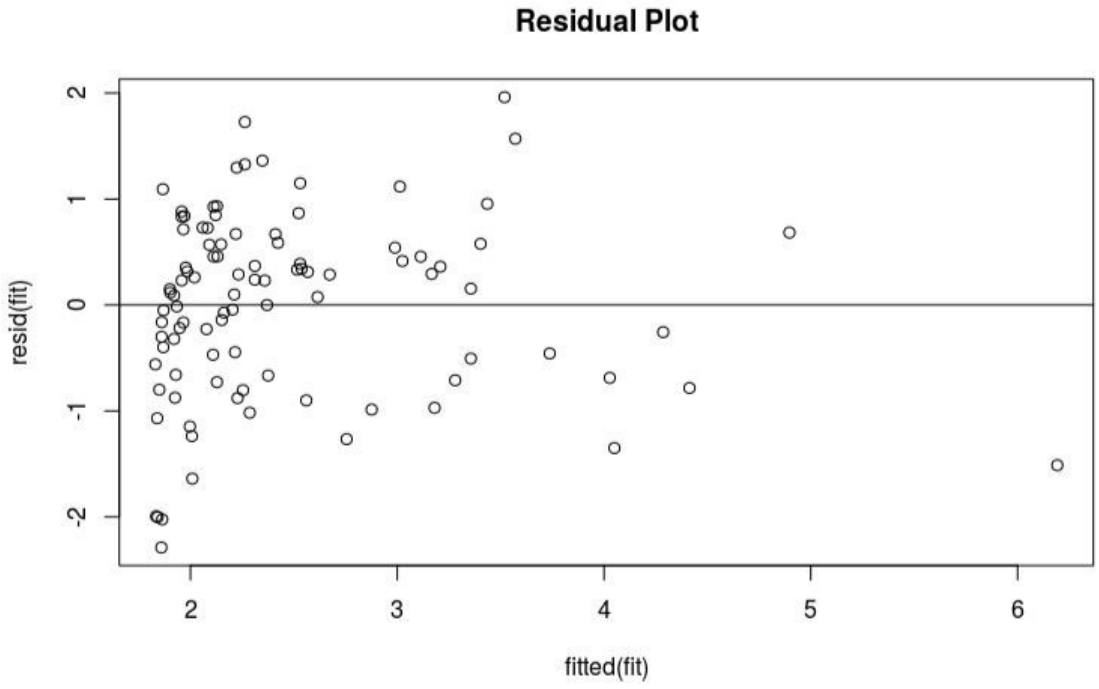

Fig 13: Residual Plot

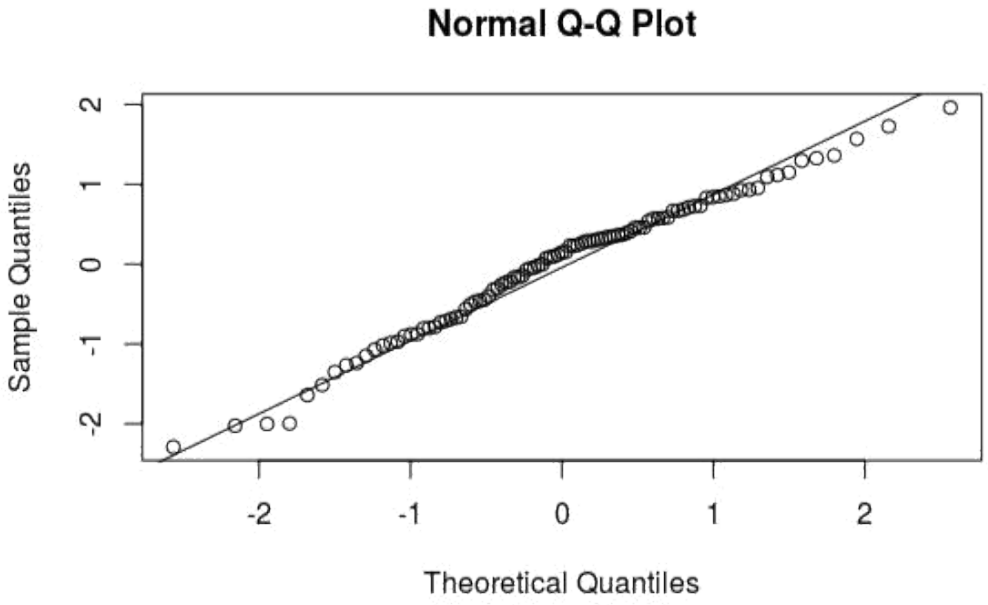

Fig 14: Normal Q-Q Plot

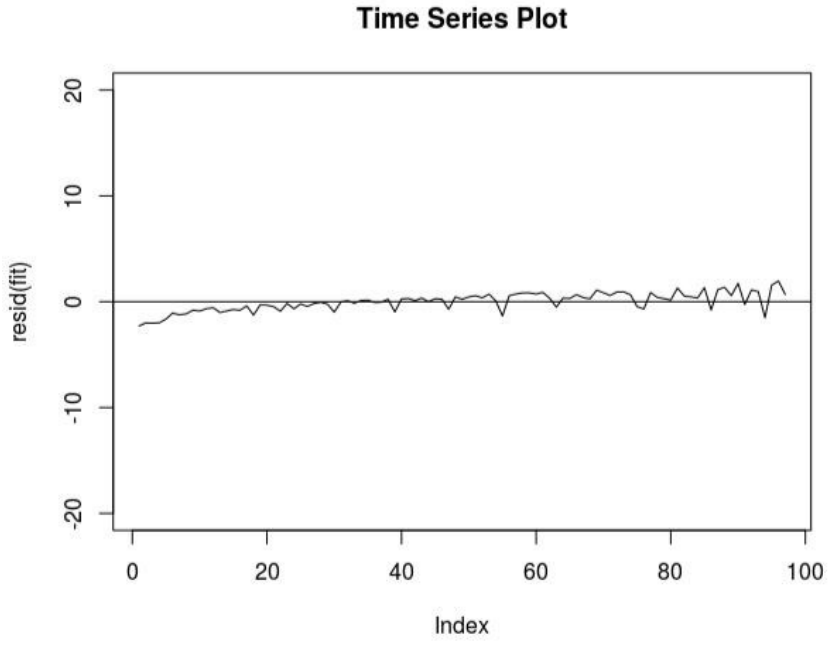

Fig 15 : Time Series Plot

RCODE:

#reading data

```
cancer=read.csv("/home/vijay/prostate_cancer.csv")
```

#Fitting Regression

```
x=cancer$cancervol
y=cancer$psa
plot(x, y, xlab="Cancer Vol", ylab="PSA",main="Cancer Vol VS. PSA")
reg=lm(cancer$psa~cancer$cancervol)
abline(regc,col="RED")
```

#Checking assumption

```
mean(residuals(reg))
```

[1]4.237733e-16

#Residual Plot(Test of Mean Zero constant variance)

```
plot(fitted(reg),resid(reg),main="Residual Plot")
abline(h=0)
```

#Normal QQ Plot(Test of Normality Assumption)

```
qqnorm(resid(reg))
qqline(resid(reg))
```

#Time Series Plot Test of independent assumptions

```
plot(resid(reg),type="l",main="Time Series Plot")
abline(h=0)
```

#summary data

```
summary(reg)
```

Call:
lm(formula = cancer$psa ~ cancer$cancervol)

Residuals:

Min	1Q	Median	3Q	Max
-61.619	-9.023	-1.586	3.151	181.183

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	1.1249	4.3596	0.258	0.797	
cancer$cancervol	3.2299	0.4148	7.786	8.47e-12	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 32.03 on 95 degrees of freedom

Multiple R-squared: 0.3896, Adjusted R-squared: 0.3831

F-statistic: 60.63 on 1 and 95 DF, p-value: 8.468e-12

#anova test

```
anova(reg)
```

Analysis of Variance Table

Response: cancer$psa

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
cancer$cancervol	1	62202	62202	60.627	8.468e-12	***

```
Residuals                 95     97469          1026
---
Signif. codes:          0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’
```

```
#log of PSA data
data=cancer
data$lpsa <- log(data$psa)

#linear Model for Logarithmic data
fit = lm(lpsa ~ cancervol, data = data)
x=data$cancervol
y=data$lpsa
plot(x, y, xlab="Cancer Vol", ylab="Log(PSA)",main="Cancer Vol VS. Log(PSA)")
abline(fit,col="RED")

#Residual Plot(Test of Mean Zero constant variance)
plot(fitted(fit),resid(fit),main="Residual Plot")
abline(h=0)

#Normal QQ Plot(Test of Normality Assumption)
qqnorm(resid(fit))
qqline(resid(fit))

#Time Series Plot Test of independent assumptions
plot(resid(fit),type="l",main="Time Series Plot",ylim=range(-20:20))
abline(h=0)
```

#Test and Statastics

```
anova(fit)
```

Analysis of Variance Table

```
Response: lpsa
           Df Sum Sq Mean Sq F value    Pr(>F)
cancervol  1 55.164  55.164  72.179 2.688e-13 ***
Residuals 95 72.605   0.764
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
summary(fit)
```

```
Call:
lm(formula = lpsa ~ cancervol, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-2.2886 -0.6590  0.1493  0.5769  1.9610

Coefficients:
```

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	1.80549	0.11899	15.174	< 2e-16	***
cancervol	0.09619	0.01132	8.496	2.69e-13	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8742 on 95 degrees of freedom
Multiple R-squared: 0.4317, Adjusted R-squared: 0.4258
F-statistic: 72.18 on 1 and 95 DF, p-value: 2.688e-13

Exercise 3:

Use the final model to predict the PSA level for a patient whose predictor variable value is at the median of the variable.

Solution:

Median value of cancervol variable data, x = 4.2631.
Response variable (PSA level) at median value x will be y= 9.166368

R-CODE:

```
#Median of cancervol
x=median(data$cancervol)

#Finding Log(y) using slope and intersection
logy=fit$coefficients[1]+fit$coefficients[2]*x

#Predicting value of y
y=exp(logy)
```