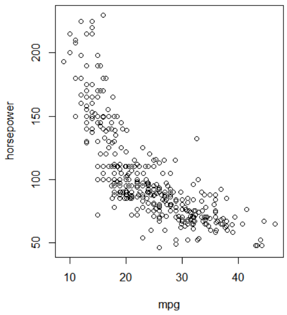
Q1

Question 1 (2+3+2+2+2+2+3+4 = 20)

This Question uses the data set Q1.txt. The data relates to gas millage, horsepower and other information on cars.

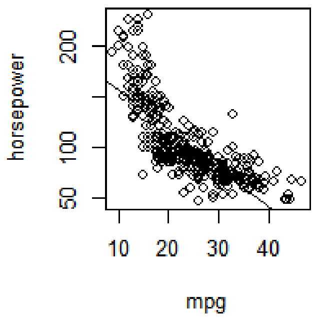
a. Plot the scatter plot of gas millage against horsepower and describe.



The data has the shape of a quadratic model rather than a linear model.

b. Perform a linear regression analysis of gas millage in terms of horsepower test the significance and plot the fitted line within the scatterplot.

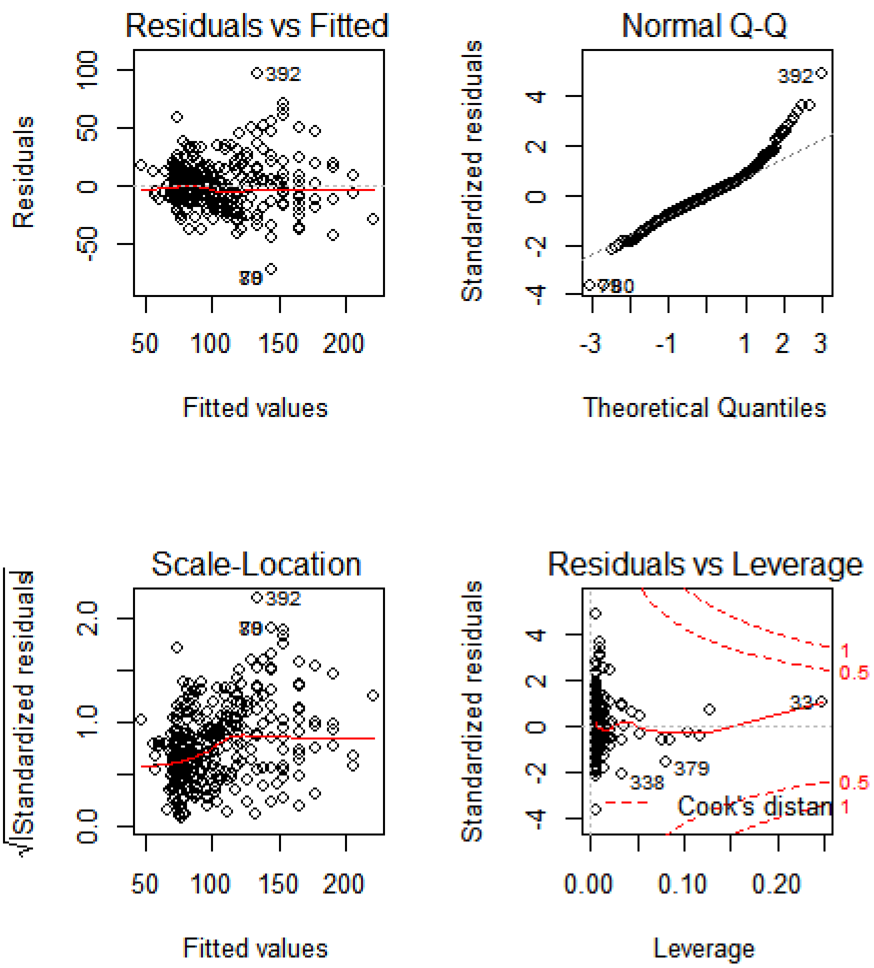
```
> fit1=lm(mydata$horsepower~mydata$mpg)
> summary(fit1)

Call:
lm(formula = mydata$horsepower ~ mydata$mpg)

Residuals:
    Min      1Q  Median      3Q     Max 
-64.892 -15.716  -2.094  13.108  96.947 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 194.4756     3.8732   50.21   <2e-16 ***
mydata$mpg   -3.8389     0.1568  -24.49   <2e-16 ***
---
Signif. codes:  
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 24.19 on 390 degrees of freedom
Multiple R-squared:  0.6059,	Adjusted R-squared:  0.6049 
F-statistic: 599.7 on 1 and 390 DF,  p-value: < 2.2e-16
```



c. Perform a polynomial regression of order 3 to model gas millage in terms of horsepower and select the best fit model and justify.

```
fit2 <- lm(mydata$horsepower ~ mydata$mpg + I(mydata$mpg^2) +
I(mydata$mpg^3), data = mydata)
> summary (fit2)

Call:
lm(formula = mydata$horsepower ~ mydata$mpg + I(mydata$mpg^2) +
    I(mydata$mpg^3), data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max
-71.935 -11.251  -1.338   9.324  95.459

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     429.581461  23.823018  18.032  < 2e-16
mydata$mpg      -30.131244   3.006538 -10.022  < 2e-16
I(mydata$mpg^2)   0.866793   0.118533   7.313 1.51e-12
I(mydata$mpg^3)  -0.008506   0.001474  -5.770 1.62e-08

(Intercept)     ***
mydata$mpg      ***
I(mydata$mpg^2) ***
I(mydata$mpg^3) ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 19.69 on 388 degrees of freedom
Multiple R-squared:  0.7403,	Adjusted R-squared:  0.7382
F-statistic: 368.6 on 3 and 388 DF,  p-value: < 2.2e-16
```

par(mfrow=c(2,2))
plot(fit2)



d. Perform a linear regression analysis of gas millage in terms of all other numeric variables provided.

```
> fit3 = glm(mydata$mpg ~ mydata$cylinders + mydata$displacement +
mydata$horsepower + mydata$weight +mydata$acceleration + mydata$year
,data = mydata)
> summary(fit3)

Call:
glm(formula = mydata$mpg ~ mydata$cylinders + mydata$displacement +
    mydata$horsepower + mydata$weight + mydata$acceleration +
    mydata$year, data = mydata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-8.6927  -2.3864  -0.0801   2.0291  14.3607

Coefficients:
                      Estimate Std. Error t value
(Intercept)          -1.454e+01  4.764e+00  -3.051
mydata$cylinders     -3.299e-01  3.321e-01  -0.993
mydata$displacement   7.678e-03  7.358e-03   1.044
mydata$horsepower    -3.914e-04  1.384e-02  -0.028
mydata$weight        -6.795e-03  6.700e-04 -10.141
mydata$acceleration   8.527e-02  1.020e-01   0.836
```

```
mydata$year          7.534e-01  5.262e-02  14.318
                    Pr(>|t|)
(Intercept)          0.00244 **
mydata$cylinders     0.32122
mydata$displacement  0.29733
mydata$horsepower    0.97745
mydata$weight        < 2e-16 ***
mydata$acceleration  0.40383
mydata$year          < 2e-16 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 11.8009)

    Null deviance: 23819.0  on 391  degrees of freedom
Residual deviance:  4543.3  on 385  degrees of freedom
AIC: 2088.9

Number of Fisher Scoring iterations: 2
```

e. Discuss the significance of slopes and select the best linear regression model to describe the gas millage.

From fit3 in part d the variables with the most significance are weight and year with 3 stars (***) indicating low p-value of 2e-16 and intercept with 2 stars(**) with pvalue 0.00244

```
> fit4 = lm(mydata$mpg ~  mydata$weight + mydata$year  ,data = mydata)
> summary(fit4)

Call:
lm(formula = mydata$mpg ~ mydata$weight + mydata$year, data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max
-8.8505 -2.3014 -0.1167  2.0367 14.3555

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   -1.435e+01  4.007e+00  -3.581 0.000386
mydata$weight -6.632e-03  2.146e-04 -30.911  < 2e-16
mydata$year    7.573e-01  4.947e-02  15.308  < 2e-16

(Intercept)   ***
mydata$weight ***
mydata$year   ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.427 on 389 degrees of freedom
Multiple R-squared:  0.8082,   Adjusted R-squared:  0.8072
F-statistic: 819.5 on 2 and 389 DF,  p-value: < 2.2e-16
```

f. Extend the best linear regression model selected in part e) to test whether the interaction of horsepower and number of cylinders is significant.

```
fit6 = lm(mydata$mpg ~ mydata$cylinders + mydata$displacement +
mydata$weight + mydata$year  ,data = mydata)
> summary(fit6)

Call:
lm(formula = mydata$mpg ~ mydata$cylinders + mydata$displacement +
    mydata$weight + mydata$year, data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max
-9.0169 -2.2958 -0.0967  2.0400 14.4239

Coefficients:
                      Estimate Std. Error t value
(Intercept)         -1.369e+01  4.079e+00  -3.357
mydata$cylinders    -3.217e-01  3.299e-01  -0.975
mydata$displacement  4.888e-03  6.695e-03   0.730
mydata$weight       -6.612e-03  5.735e-04 -11.531
mydata$year          7.586e-01  5.101e-02  14.872
                    Pr(>|t|)
(Intercept)         0.000868 ***
mydata$cylinders    0.330182
mydata$displacement 0.465727
mydata$weight        < 2e-16 ***
mydata$year          < 2e-16 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.432 on 387 degrees of freedom
Multiple R-squared:  0.8087,   Adjusted R-squared:  0.8067
F-statistic: 408.9 on 4 and 387 DF,  p-value: < 2.2e-16
```

g. From the models in part b), c) and f) above, report the R-squared, as a percentage and comment.

Question part		R^2
B	–	0.6059
C	-	0.8082
F	-	0.8087

h. Using the results from parts a) to g) discover the most suitable model to describe gas millage and justify.

From the three models created the model with the highest R^2value is that in part F – fit6 R^2 value was 0.8087 meaning that 80.87% of the data is covered by the model.

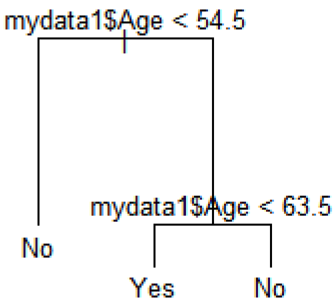
Question 2 (2+2+2+2+ 2 = 10)
This Question uses the data set Q2.csv. The data relates to a person having heart disease or not having heart disease (AHD) and many other related variables.
a. Perform a simple logistic linear regression model using the glm function in R to model the variable AHD, presence and absence of heart disease in terms of age. Is the model significant? Justify.

```
fit10 = glm(mydata1$AHD~mydata1$Age,family = "binomial",data = mydata)
> summary(fit10)

Call:
glm(formula = mydata1$AHD ~ mydata1$Age, family = "binomial",
    data = mydata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5432  -1.0745  -0.8323   1.1785   1.6997

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.88579    0.77544  -3.721 0.000198 ***
mydata1$Age  0.04887    0.01395   3.502 0.000462 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 397.31  on 288  degrees of freedom
Residual deviance: 384.29  on 287  degrees of freedom
AIC: 388.29

Number of Fisher Scoring iterations: 4
```

The model is significant and this is demonstrated through the low p-values for intercept and age both with 3 stars indicating high significance.

b. Assuming the model obtained in part a) estimate the respective probabilities of having heart disease for someone with age 60 and age 30. Compare the results and comment.

c. Construct and plot a Decision Tree to classify Heart Disease (AHD = 1 Yes, AHD = 0, No) in terms of other associated variables given in the data set.



d. Give two classification rules from the tree.

Age< 54.5
Age< 63.5
e. Construct the misclassification table and give the misclassification rate and comment.

Question 3 (2+2+2+2+2 = 10)

a. Describe briefly and compare Clustering and Principal Component Analysis.

If we have a high dimensional data set X, and a distance defined between observations; eg. Euclidean distance. The idea of agglomerative hierarchical clustering, is to gradually merge clusters together to get a hierarchy of cluster solutions.

For hierarchial clustering the method can be “single”, “average”, and “complete” and the distance between two clusters A and B is:
single — The minimum of distances between points in A and points in B
average — The average of distances between points in A and points in B
complete — The maximum of distances between points in A and points in B

PCA

Pricipal Component Analysis is a method of dimension reduction with the goal to seek a low dimensional representation of the data, that matches the complete dataset.

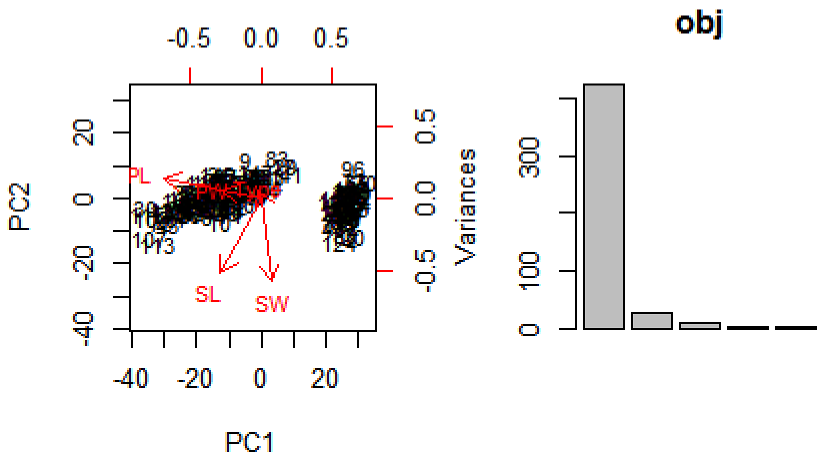
b. Use K means clustering method and identify clusters in Q3 data set starting with K=3.

```
> table(Outcome=mydata2$Type, cluster=fitted(km,"classes"))
       cluster
Outcome  1  2  3
      0  0 50  0
      1 36  0 14
      2  5  0 45
```

C. Plot using principal component command and describe the results.

mydata2 <- read.csv("Q3.csv")

```
attach(mydata2)
> summary(mydata2)
      Type           PW               PL       
 Min.   :0   Min.   : 1.00   Min.   :10.00  
 1st Qu.:0   1st Qu.: 3.00   1st Qu.:16.00  
 Median :1   Median :13.00   Median :44.00  
 Mean   :1   Mean   :11.93   Mean   :37.79  
 3rd Qu.:2   3rd Qu.:18.00   3rd Qu.:51.00  
 Max.   :2   Max.   :25.00   Max.   :69.00  
       SW              SL       
 Min.   :20.00   Min.   :43.00  
 1st Qu.:28.00   1st Qu.:51.00  
 Median :30.00   Median :58.00  
 Mean   :30.55   Mean   :58.45  
 3rd Qu.:33.00   3rd Qu.:64.00  
 Max.   :44.00   Max.   :79.00  
> obj = prcomp(mydata2[,1:5])
> biplot(obj,scale = 0)
> screeplot(obj)
```



We can see from the Screeplot above that we should use the first two components (PC1 & PC2) for the biplot. Through dimension reduction using PCA we are able to find the dominant dimensions of the dataset, reducing to only 2 dimensions by ignoring all eigenvectors with insignificant eigenvalues.

d. Use hierarchical clustering method and repeat the same in part b) to identify 3 clusters.

```
> table(Outcome=mydata2$Type, cluster=fitted(km,"classes"))
       cluster
Outcome  1  2  3
      0  0 50  0
      1 36  0 14
      2  5  0 45
```

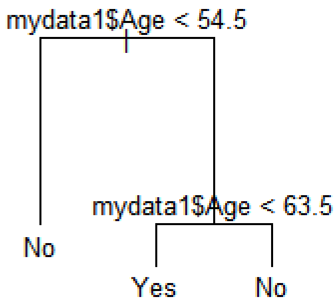
table(Outcome=mydata2$Type, cluster=fitted(km,"classes"))

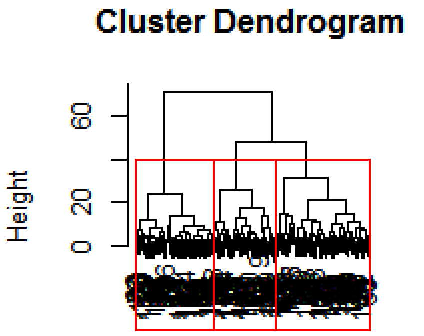
#solution for all cluster memebers
hh = hclust(dist(X),method = "complete")

cutree(hh, k=3)

plot(hh, xlab = "", sub = "Complete link cluster analysis")
rect.hclust(hh,k=3)

e. Plot using principal component command and describe the results.





Complete link cluster analysis