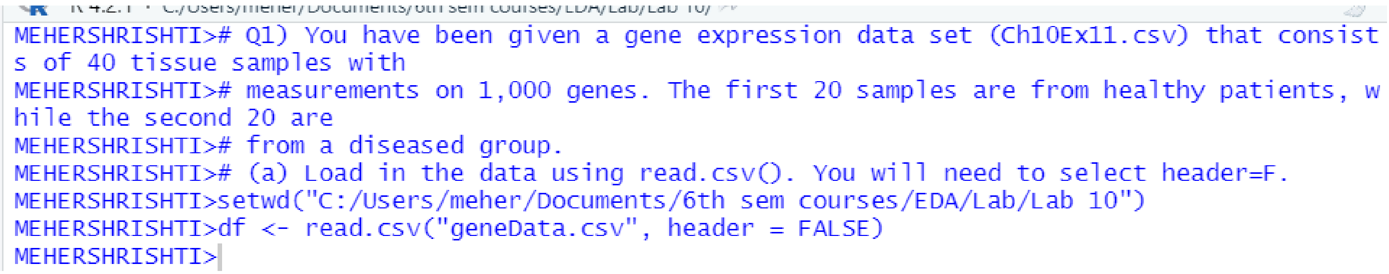
Meher Shrishti Nigam
20BRS1193

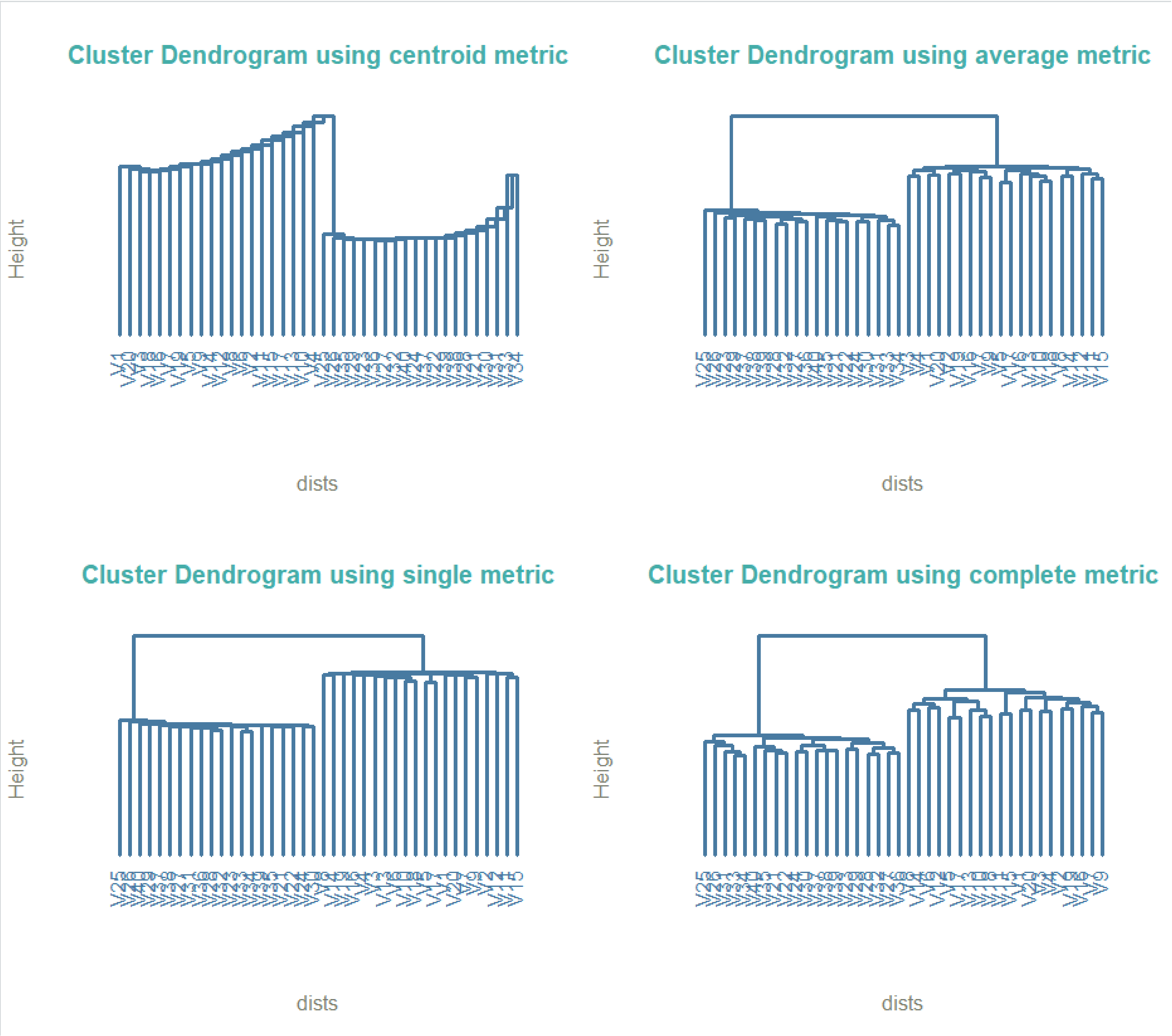
EDA LAB – 10
24 / 3 / 23

```
# Meher Shrishti Nigam

# 20BRS1193

# EDA Lab 10

options(prompt="MEHERSHRISHTI>", continue =" ")

# options(prompt=">", continue =" ")

# EDA-LAB-EXPERIMENT-10 (Date-25/3/2023)

library(dplyr)

library(ggplot2)


# Q1) You have been given a gene expression data set (Ch10Ex11.csv) that consists of 40 tissue samples with

# measurements on 1,000 genes. The first 20 samples are from healthy patients, while the second 20 are

# from a diseased group.

# (a) Load in the data using read.csv(). You will need to select header=F.

setwd("C:/Users/meher/Documents/6th sem courses/EDA/Lab/Lab 10")

df <- read.csv("geneData.csv", header = FALSE)
```

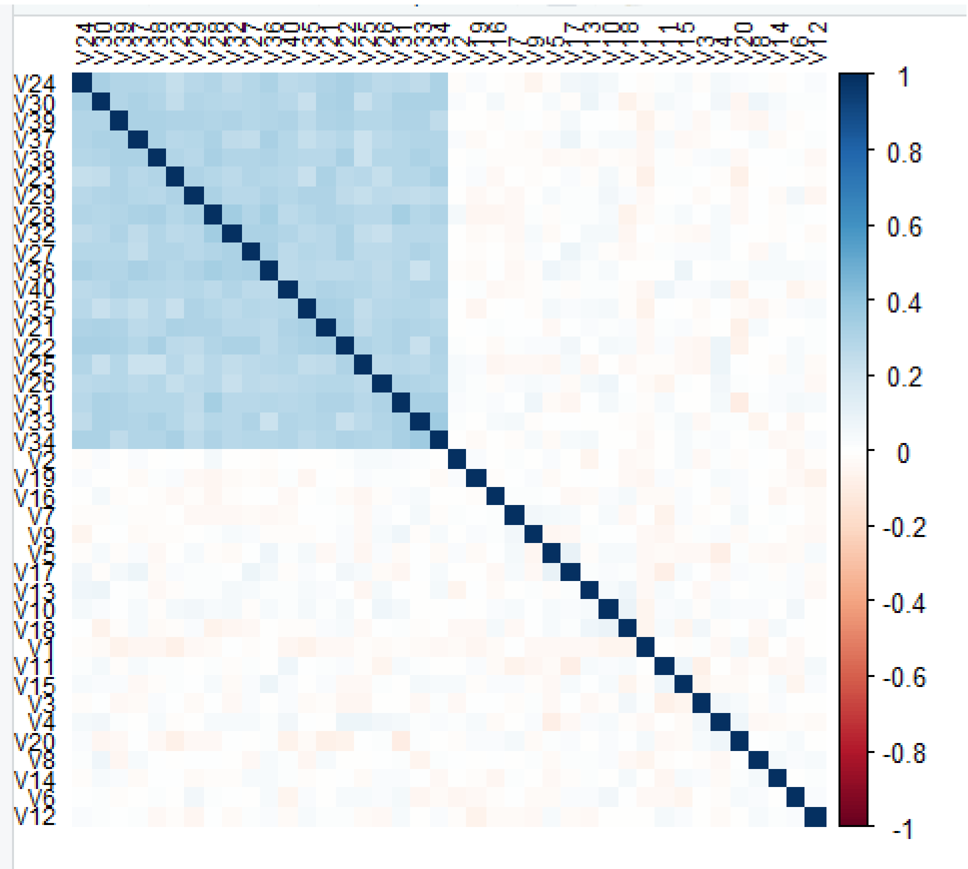

R 4.2.1 · C:/Users/meher/Documents/6th sem courses/EDA/Lab/Lab 10/

```
MEHERSHRISHTI># Q1) You have been given a gene expression data set (Ch10Ex11.csv) that consist
s of 40 tissue samples with
MEHERSHRISHTI># measurements on 1,000 genes. The first 20 samples are from healthy patients, w
hile the second 20 are
MEHERSHRISHTI># from a diseased group.
MEHERSHRISHTI># (a) Load in the data using read.csv(). You will need to select header=F.
MEHERSHRISHTI>setwd("C:/Users/meher/Documents/6th sem courses/EDA/Lab/Lab 10")
MEHERSHRISHTI>df <- read.csv("geneData.csv", header = FALSE)
MEHERSHRISHTI>
```

```
# (b) Apply hierarchical clustering to the samples using correlationbased distance, and plot the dendrogram.
# Do the genes separate the samples into the two groups? Do your results depend on the type of linkage
# used?
dists <- dist(cor(df))

methods <- c('centroid', 'average', 'single', 'complete')

par(mfrow = c(2,2))
for (method in methods) {
  clusts <- hclust(dists, method = method)

  plot(clusts,
       col = "#487AA1", col.main = "#45ADA8",
       col.lab = "#7C8071", col.axis = "#F38630",
       lwd = 3, lty = 1,
       sub = "", hang = -1,
       axes = FALSE,
       main = paste0('Cluster Dendrogram using ', method, ' metric'))
}
```

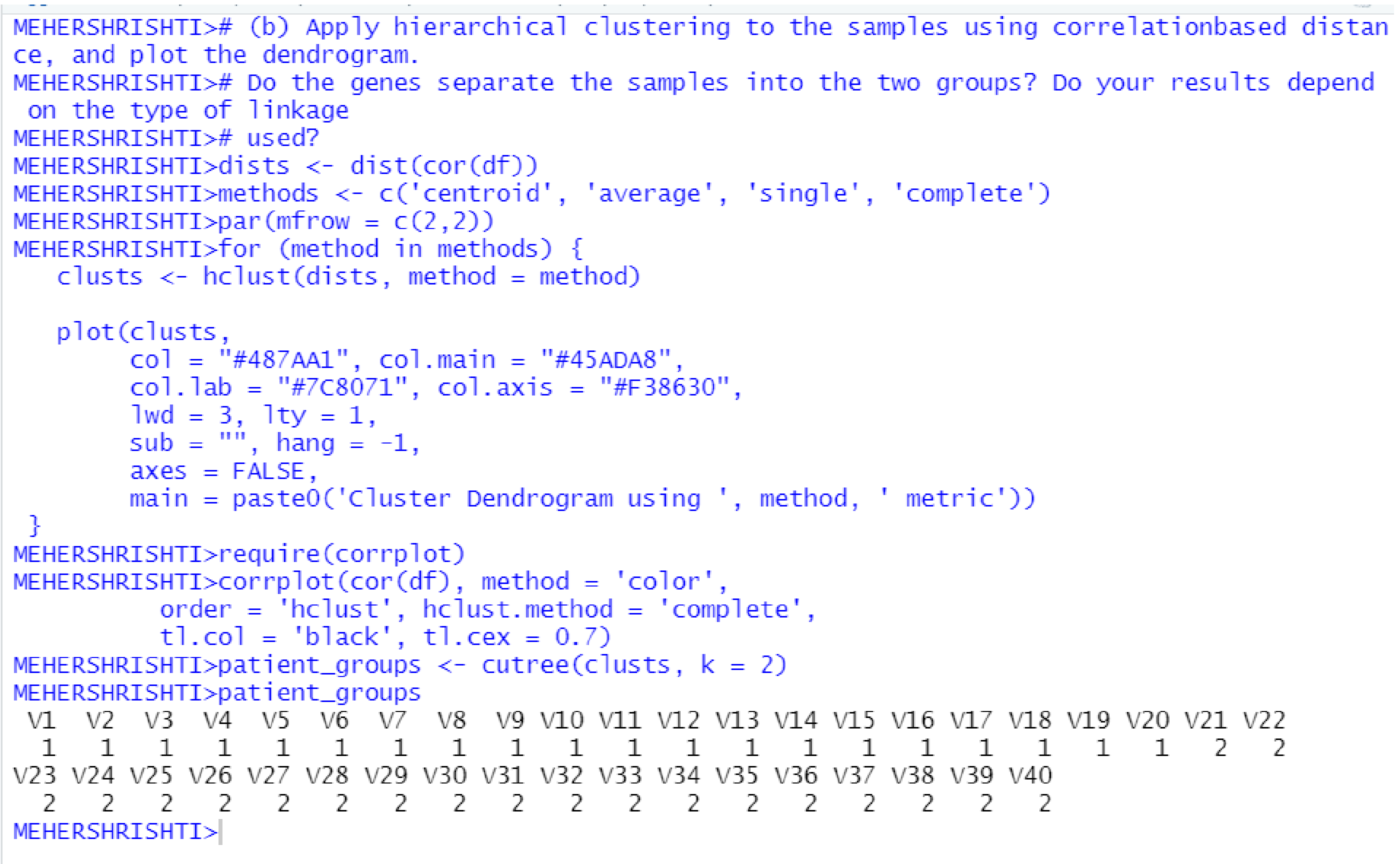


```
require(corrplot)
corrplot(cor(df), method = 'color',
        order = 'hclust', hclust.method = 'complete',
        tl.col = 'black', tl.cex = 0.7)
```



patient_groups <- cutree(clusts, k = 2)

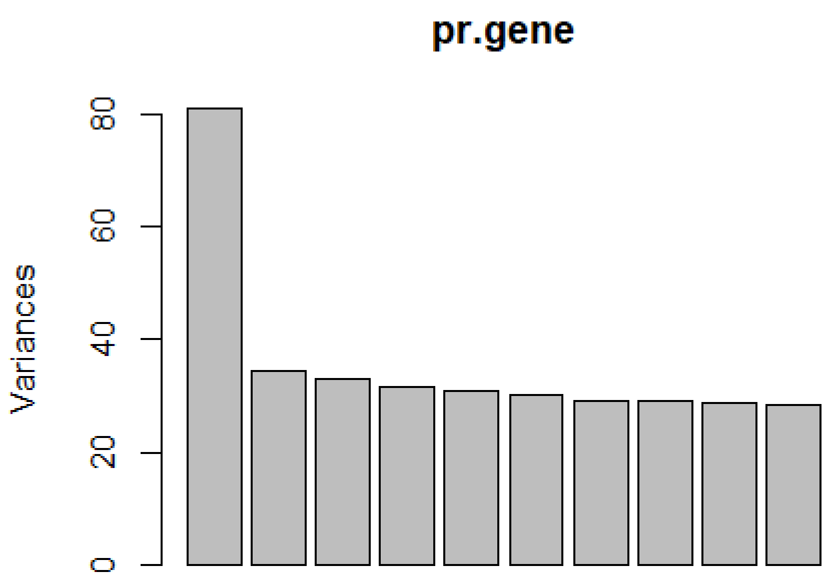
patient_groups

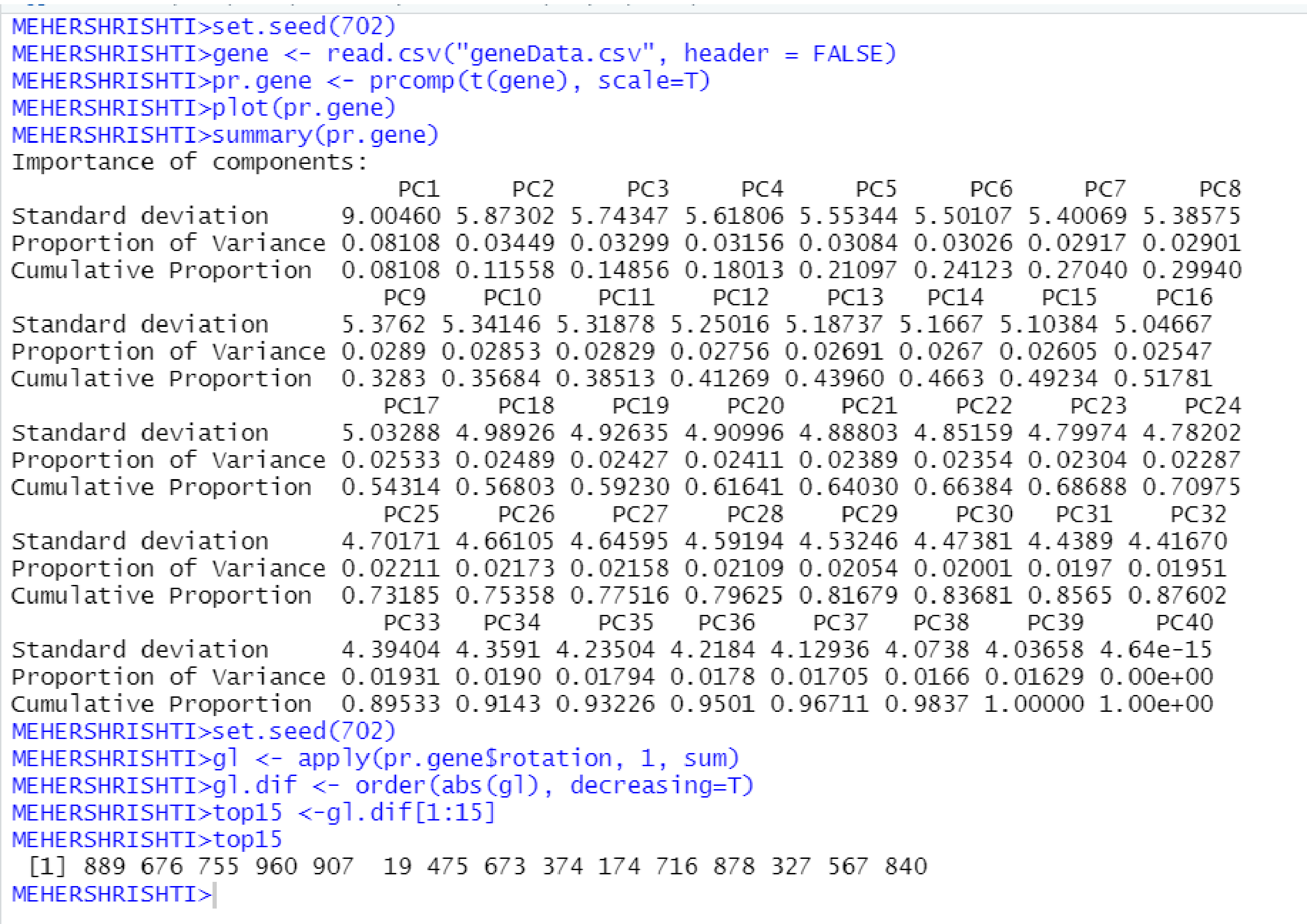


```
MEHERSHRISHTI># (b) Apply hierarchical clustering to the samples using correlationbased distan
ce, and plot the dendrogram.
MEHERSHRISHTI># Do the genes separate the samples into the two groups? Do your results depend
 on the type of linkage
MEHERSHRISHTI># used?
MEHERSHRISHTI>dists <- dist(cor(df))
MEHERSHRISHTI>methods <- c('centroid', 'average', 'single', 'complete')
MEHERSHRISHTI>par(mfrow = c(2,2))
MEHERSHRISHTI>for (method in methods) {
   clusts <- hclust(dists, method = method)

   plot(clusts,
        col = "#487AA1", col.main = "#45ADA8",
        col.lab = "#7C8071", col.axis = "#F38630",
        lwd = 3, lty = 1,
        sub = "", hang = -1,
        axes = FALSE,
        main = paste0('Cluster Dendrogram using ', method, ' metric'))
 }
MEHERSHRISHTI>require(corrplot)
MEHERSHRISHTI>corrplot(cor(df), method = 'color',
          order = 'hclust', hclust.method = 'complete',
          tl.col = 'black', tl.cex = 0.7)
MEHERSHRISHTI>patient_groups <- cutree(clusts, k = 2)
MEHERSHRISHTI>patient_groups
 V1  V2  V3  V4  V5  V6  V7  V8  V9 V10 V11 V12 V13 V14 V15 V16 V17 V18 V19 V20 V21 V22 
  1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   2   2 
V23 V24 V25 V26 V27 V28 V29 V30 V31 V32 V33 V34 V35 V36 V37 V38 V39 V40 
  2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2 
MEHERSHRISHTI>
```

The basic dendrogram hints at the variables being separated into two groups. There seems so be strong evidence for the variables belonging to two distinct groups.

```
# (c) Your collaborator wants to know which genes differ the most across the two groups. Suggest a way to
# answer this question,and apply it here.


set.seed(702)
gene <- read.csv("geneData.csv", header = FALSE)
pr.gene <- prcomp(t(gene), scale=T)
plot(pr.gene)

summary(pr.gene)

set.seed(702)
gl <- apply(pr.gene$rotation, 1, sum)
gl.dif <- order(abs(gl), decreasing=T)
top15 <-gl.dif[1:15]
top15
```



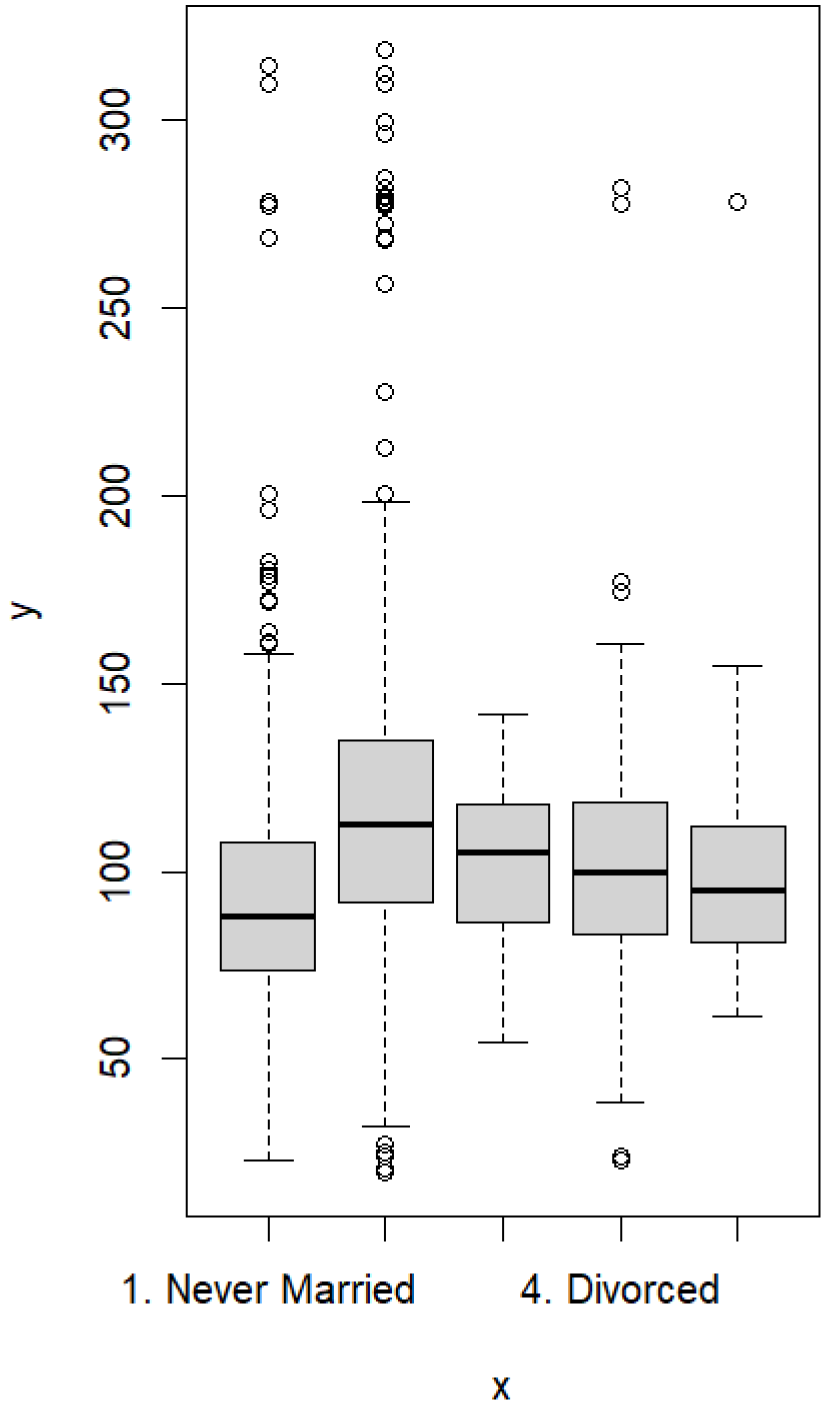


```
MEHERSHRISHTI>set.seed(702)
MEHERSHRISHTI>gene <- read.csv("geneData.csv", header = FALSE)
MEHERSHRISHTI>pr.gene <- prcomp(t(gene), scale=T)
MEHERSHRISHTI>plot(pr.gene)
MEHERSHRISHTI>summary(pr.gene)
Importance of components:
                           PC1     PC2     PC3     PC4     PC5     PC6     PC7     PC8
Standard deviation     9.00460 5.87302 5.74347 5.61806 5.55344 5.50107 5.40069 5.38575
Proportion of Variance 0.08108 0.03449 0.03299 0.03156 0.03084 0.03026 0.02917 0.02901
Cumulative Proportion  0.08108 0.11558 0.14856 0.18013 0.21097 0.24123 0.27040 0.29940
                          PC9    PC10    PC11    PC12    PC13   PC14    PC15    PC16
Standard deviation     5.3762 5.34146 5.31878 5.25016 5.18737 5.1667 5.10384 5.04667
Proportion of Variance 0.0289 0.02853 0.02829 0.02756 0.02691 0.0267 0.02605 0.02547
Cumulative Proportion  0.3283 0.35684 0.38513 0.41269 0.43960 0.4663 0.49234 0.51781
                          PC17    PC18    PC19    PC20    PC21    PC22    PC23    PC24
Standard deviation     5.03288 4.98926 4.92635 4.90996 4.88803 4.85159 4.79974 4.78202
Proportion of Variance 0.02533 0.02489 0.02427 0.02411 0.02389 0.02354 0.02304 0.02287
Cumulative Proportion  0.54314 0.56803 0.59230 0.61641 0.64030 0.66384 0.68688 0.70975
                          PC25    PC26    PC27    PC28    PC29    PC30   PC31    PC32
Standard deviation     4.70171 4.66105 4.64595 4.59194 4.53246 4.47381 4.4389 4.41670
Proportion of Variance 0.02211 0.02173 0.02158 0.02109 0.02054 0.02001 0.0197 0.01951
Cumulative Proportion  0.73185 0.75358 0.77516 0.79625 0.81679 0.83681 0.8565 0.87602
                          PC33   PC34    PC35   PC36    PC37   PC38    PC39     PC40
Standard deviation     4.39404 4.3591 4.23504 4.2184 4.12936 4.0738 4.03658 4.64e-15
Proportion of Variance 0.01931 0.0190 0.01794 0.0178 0.01705 0.0166 0.01629 0.00e+00
Cumulative Proportion  0.89533 0.9143 0.93226 0.9501 0.96711 0.9837 1.00000 1.00e+00
MEHERSHRISHTI>set.seed(702)
MEHERSHRISHTI>gl <- apply(pr.gene$rotation, 1, sum)
MEHERSHRISHTI>gl.dif <- order(abs(gl), decreasing=T)
MEHERSHRISHTI>top15 <-gl.dif[1:15]
MEHERSHRISHTI>top15
 [1] 889 676 755 960 907  19 475 673 374 174 716 878 327 567 840
MEHERSHRISHTI>
```

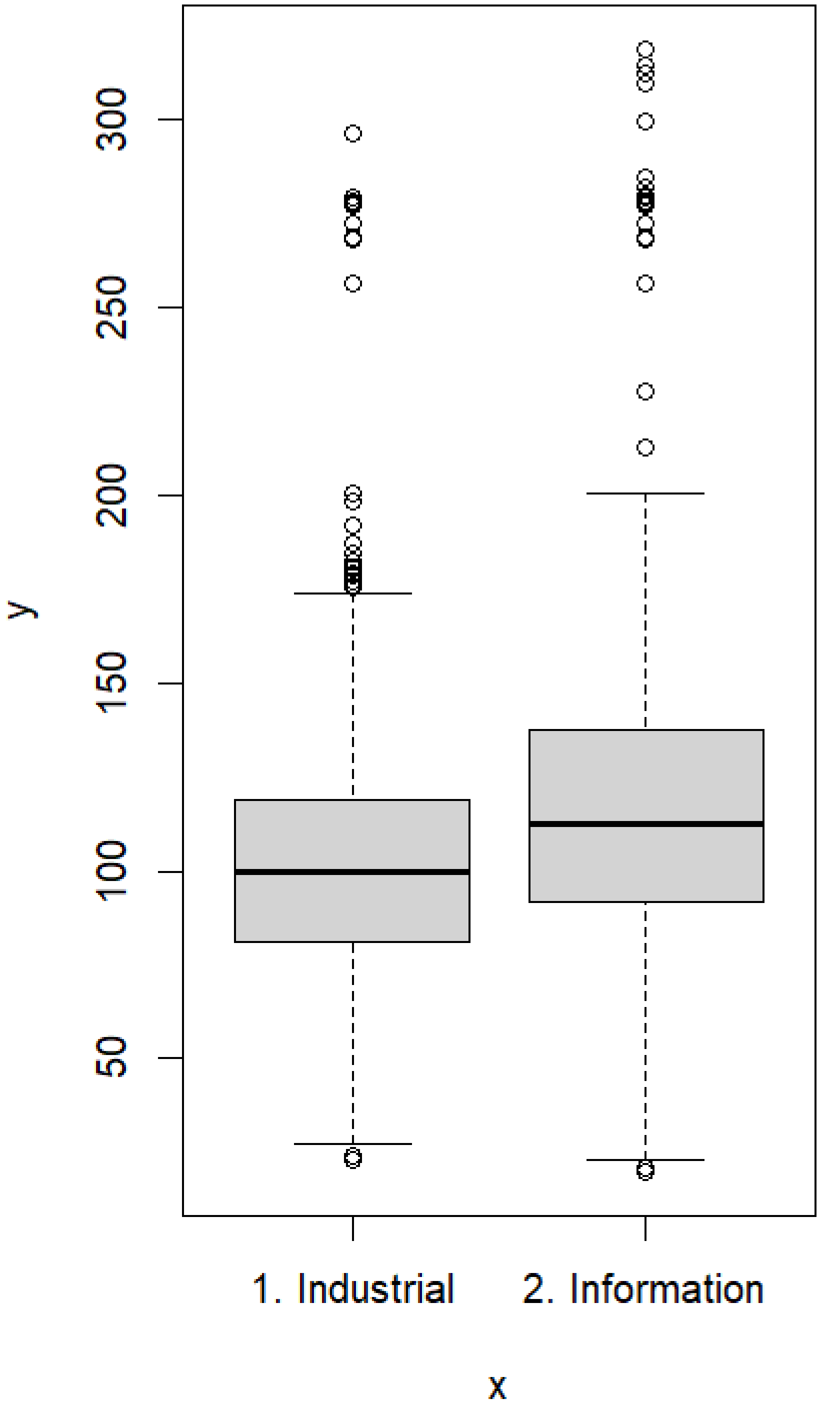
The first factor provides good separation between the two patient groups so variables that correlate higly with that factor are likely explaining the difference between diseased and healthy patients.

# Q2. The Wage data set contains a number of other features, such as marital status (maritl), job class

```
#(jobclass),and others. Explore the relationships between some of these other predictors and wage, and

# use non-linear fitting techniques in order to fit flexible models to the data. Create plots of the results

# obtained, and write a summary of your findings.


library(ISLR)

library(boot)

set.seed(1)

summary(Wage$maritl)


# table(Wage$maritl) the same with `summary`

summary(Wage$jobclass)


par(mfrow = c(1, 2))

plot(Wage$maritl, Wage$wage)
```

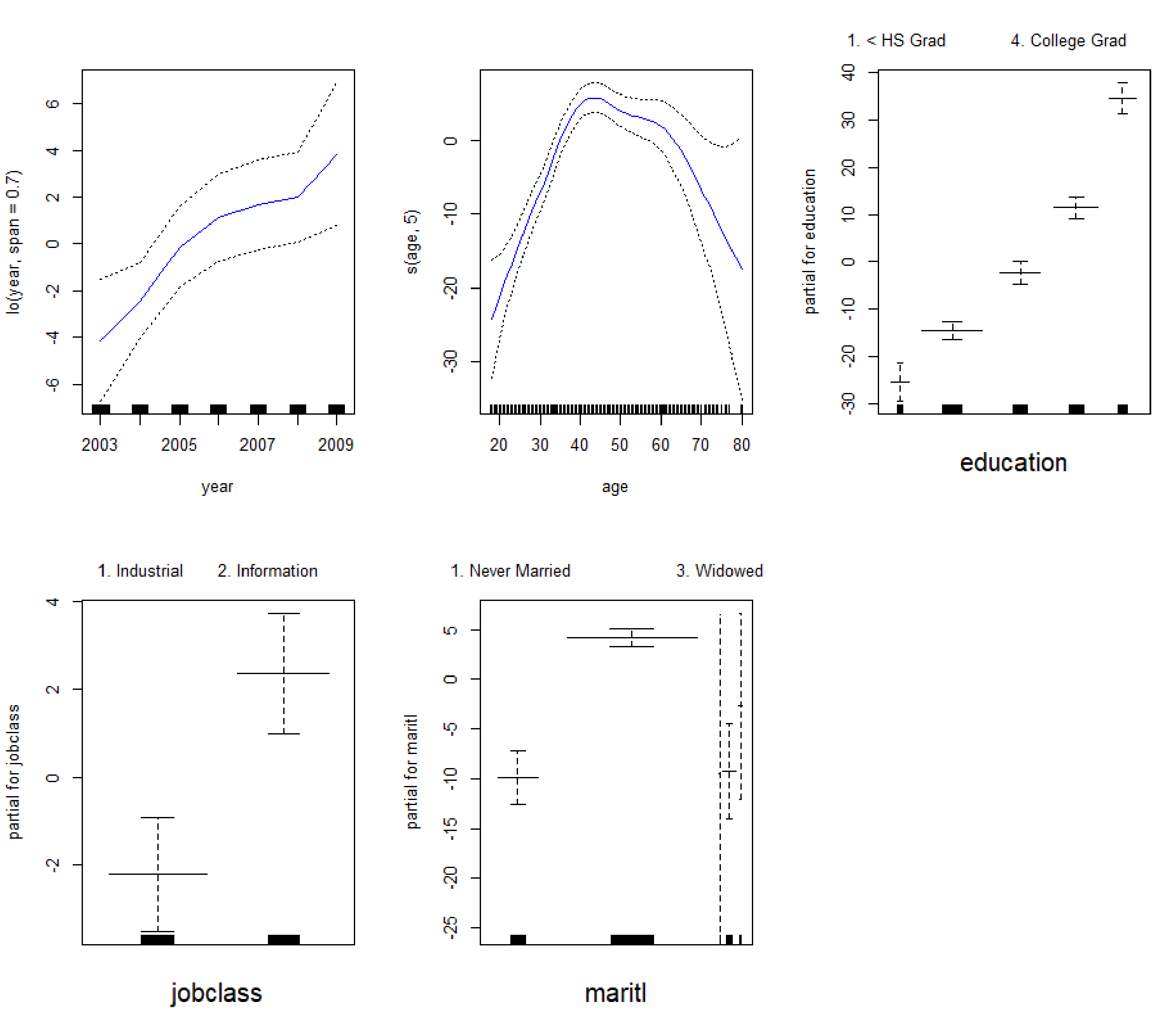


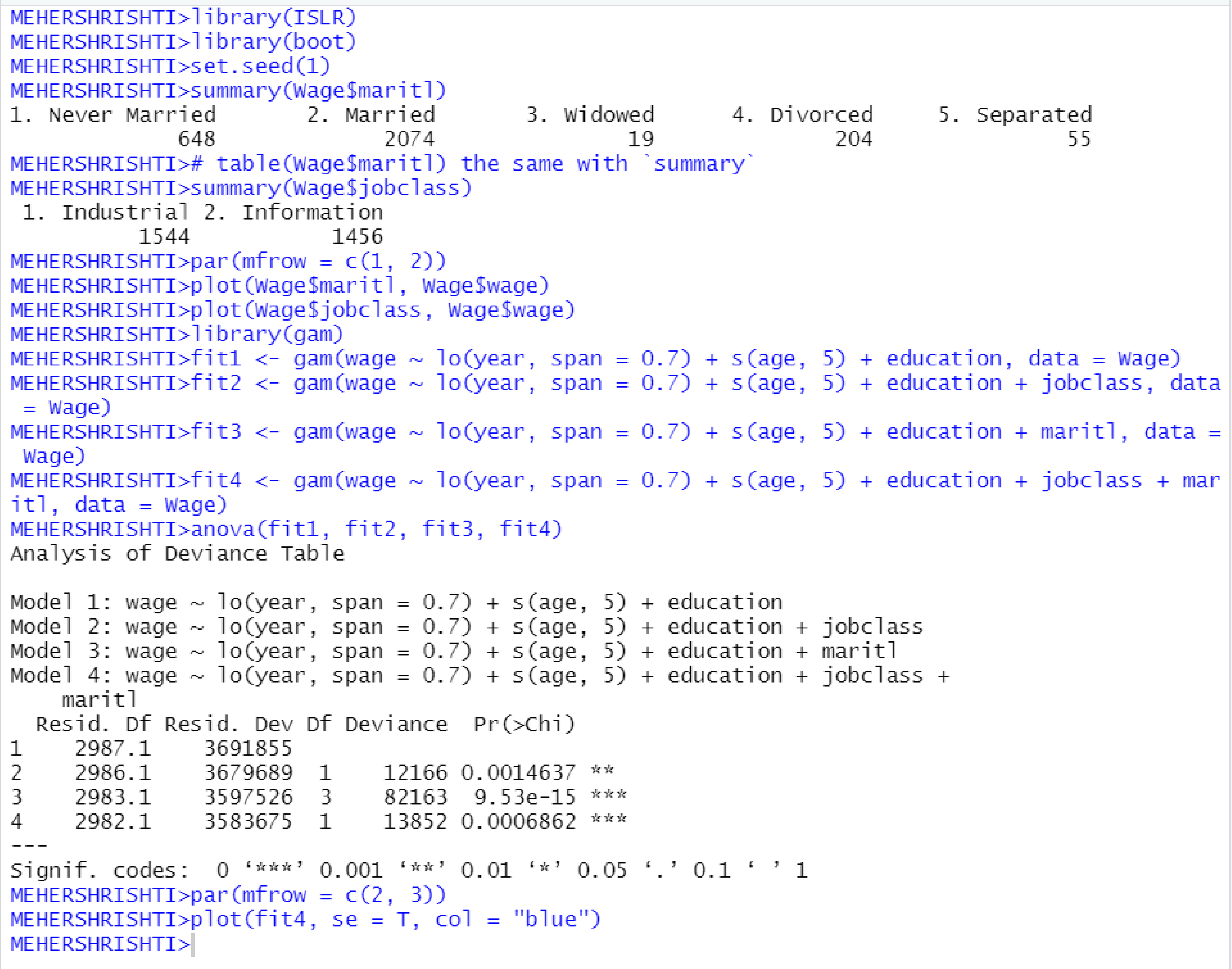
plot(Wage$jobclass, Wage$wage)



```
install.packages("gam")

library(gam)

fit1 <- gam(wage ~ lo(year, span = 0.7) + s(age, 5) + education, data = Wage)

fit2 <- gam(wage ~ lo(year, span = 0.7) + s(age, 5) + education + jobclass, data =
Wage)

fit3 <- gam(wage ~ lo(year, span = 0.7) + s(age, 5) + education + maritl, data =
Wage)
```

```
fit4 <- gam(wage ~ lo(year, span = 0.7) + s(age, 5) + education + jobclass +
maritl, data = Wage)

anova(fit1, fit2, fit3, fit4)


par(mfrow = c(2, 3))

plot(fit4, se = T, col = "blue")
```





```
MEHERSHRISHTI>library(ISLR)
MEHERSHRISHTI>library(boot)
MEHERSHRISHTI>set.seed(1)
MEHERSHRISHTI>summary(Wage$maritl)
1. Never Married       2. Married       3. Widowed      4. Divorced     5. Separated
             648             2074               19              204               55
MEHERSHRISHTI># table(Wage$maritl) the same with `summary`
MEHERSHRISHTI>summary(Wage$jobclass)
 1. Industrial 2. Information
          1544           1456
MEHERSHRISHTI>par(mfrow = c(1, 2))
MEHERSHRISHTI>plot(Wage$maritl, Wage$wage)
MEHERSHRISHTI>plot(Wage$jobclass, Wage$wage)
MEHERSHRISHTI>library(gam)
MEHERSHRISHTI>fit1 <- gam(wage ~ lo(year, span = 0.7) + s(age, 5) + education, data = Wage)
MEHERSHRISHTI>fit2 <- gam(wage ~ lo(year, span = 0.7) + s(age, 5) + education + jobclass, data
 = Wage)
MEHERSHRISHTI>fit3 <- gam(wage ~ lo(year, span = 0.7) + s(age, 5) + education + maritl, data =
 Wage)
MEHERSHRISHTI>fit4 <- gam(wage ~ lo(year, span = 0.7) + s(age, 5) + education + jobclass + mar
itl, data = Wage)
MEHERSHRISHTI>anova(fit1, fit2, fit3, fit4)
Analysis of Deviance Table

Model 1: wage ~ lo(year, span = 0.7) + s(age, 5) + education
Model 2: wage ~ lo(year, span = 0.7) + s(age, 5) + education + jobclass
Model 3: wage ~ lo(year, span = 0.7) + s(age, 5) + education + maritl
Model 4: wage ~ lo(year, span = 0.7) + s(age, 5) + education + jobclass +
    maritl
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1    2987.1    3691855
2    2986.1    3679689  1    12166 0.0014637 **
3    2983.1    3597526  3    82163  9.53e-15 ***
4    2982.1    3583675  1    13852 0.0006862 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
MEHERSHRISHTI>par(mfrow = c(2, 3))
MEHERSHRISHTI>plot(fit4, se = T, col = "blue")
MEHERSHRISHTI>
```

# Q3. Fit some of the non-linear models to the Auto data set. Is there evidence for non-linear relationships in

# this data set? Create some informative plots to justify your answer

set.seed(1)

pairs(Auto)

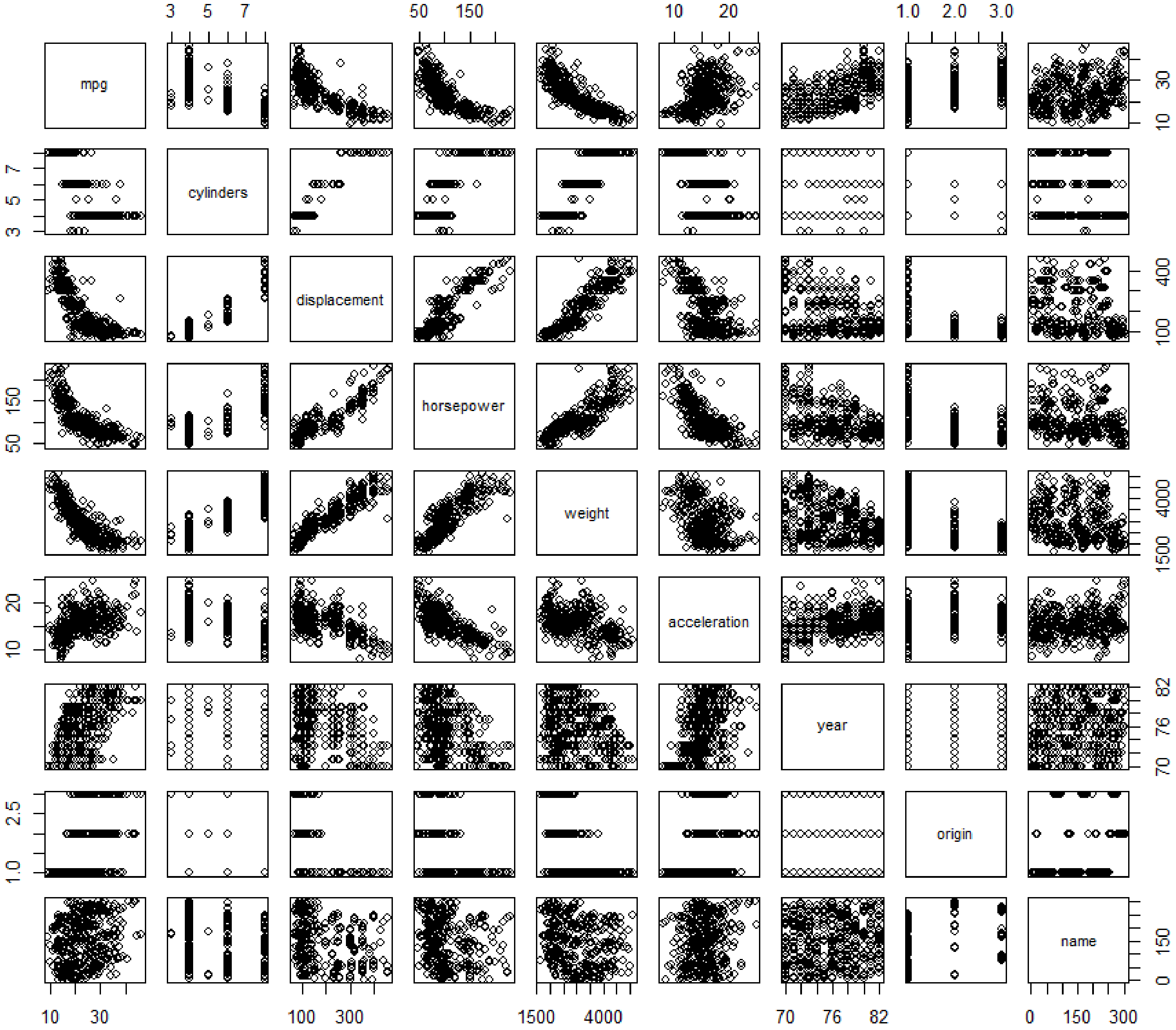
fit <- lm(mpg ~ poly(cylinders, 2) + poly(displacement, 5) + poly(horsepower, 5) + poly(weight, 5), data = Auto)

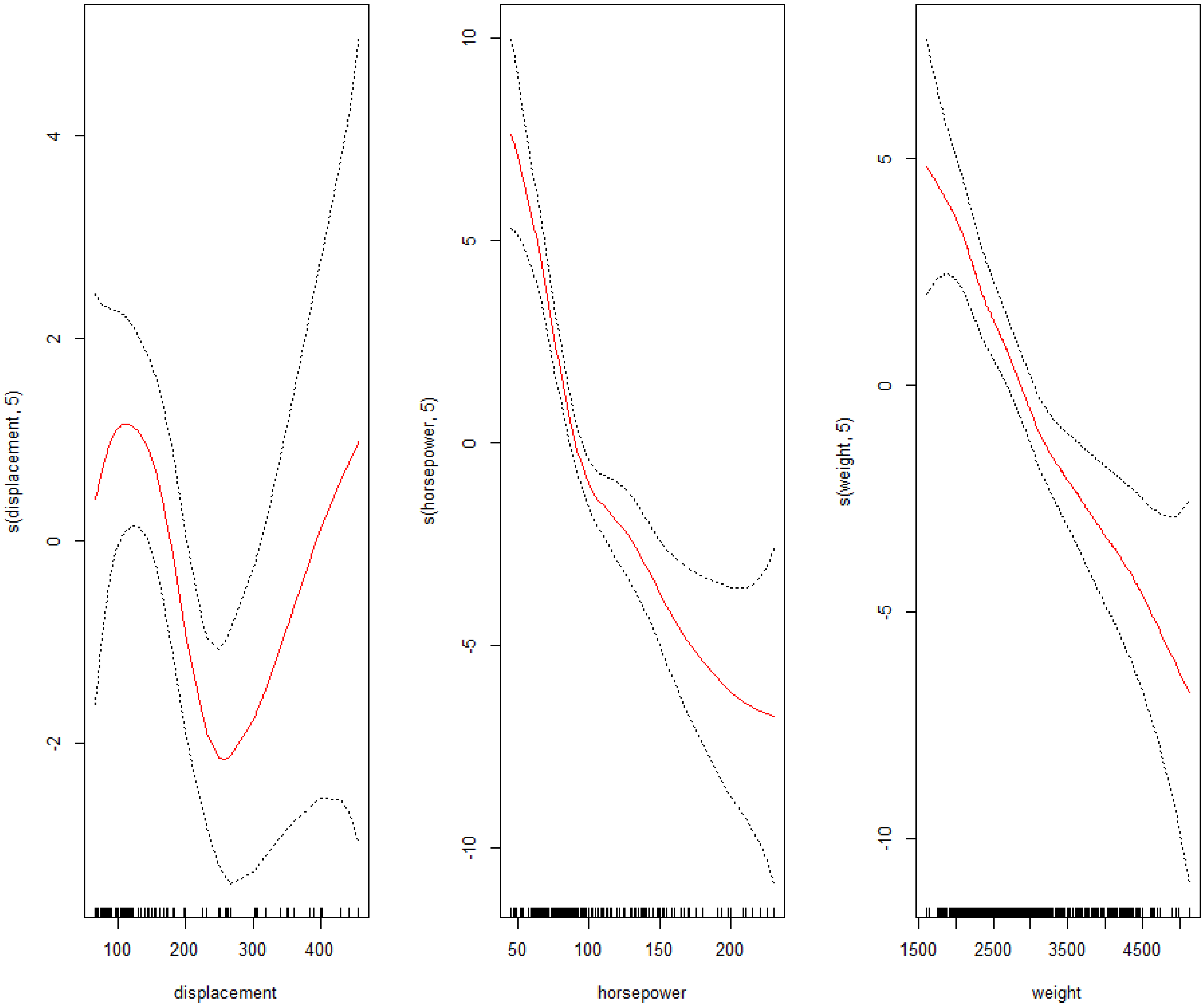
summary(fit)

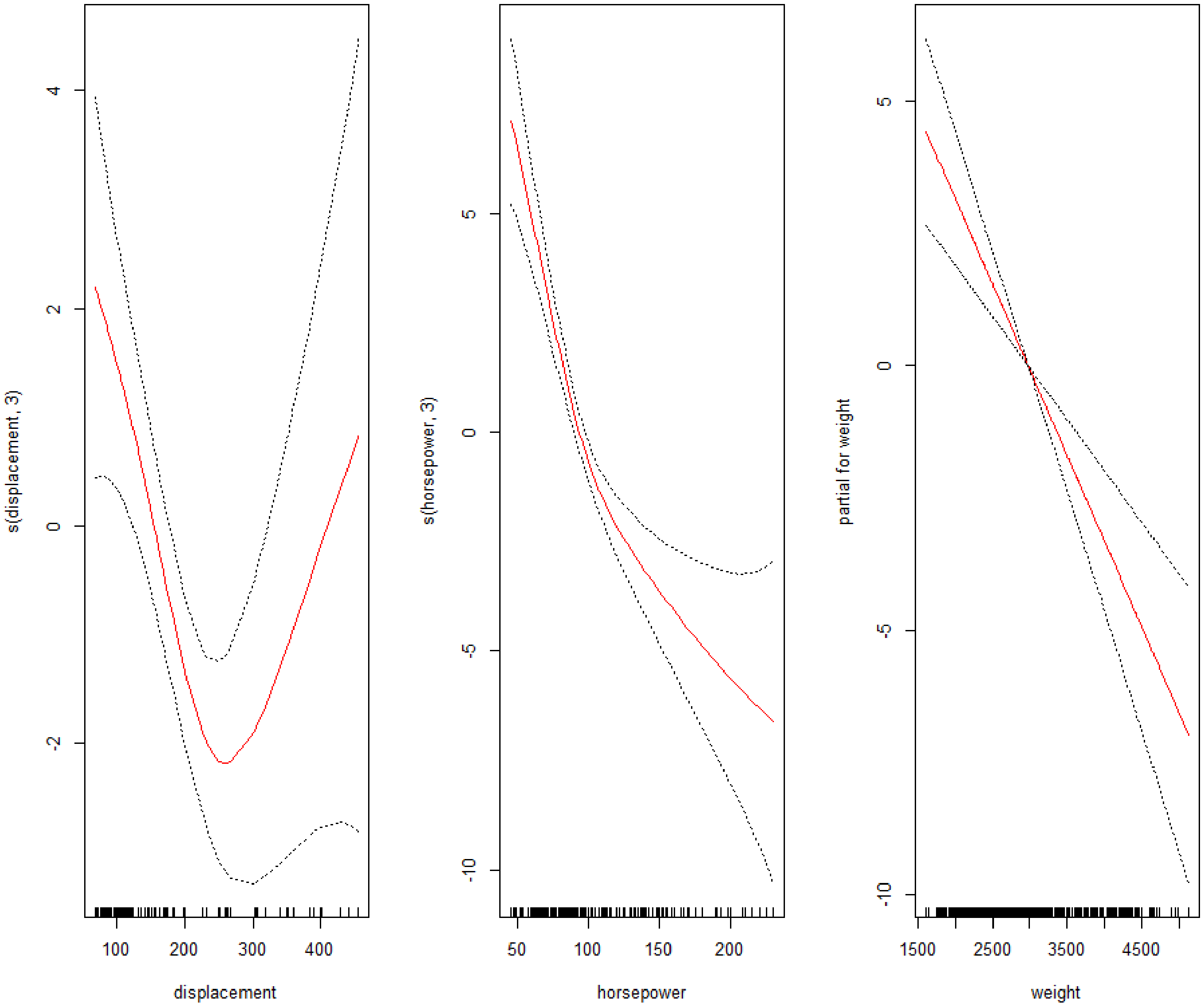
anv1 <- gam(mpg ~ displacement + horsepower + weight, data = Auto)

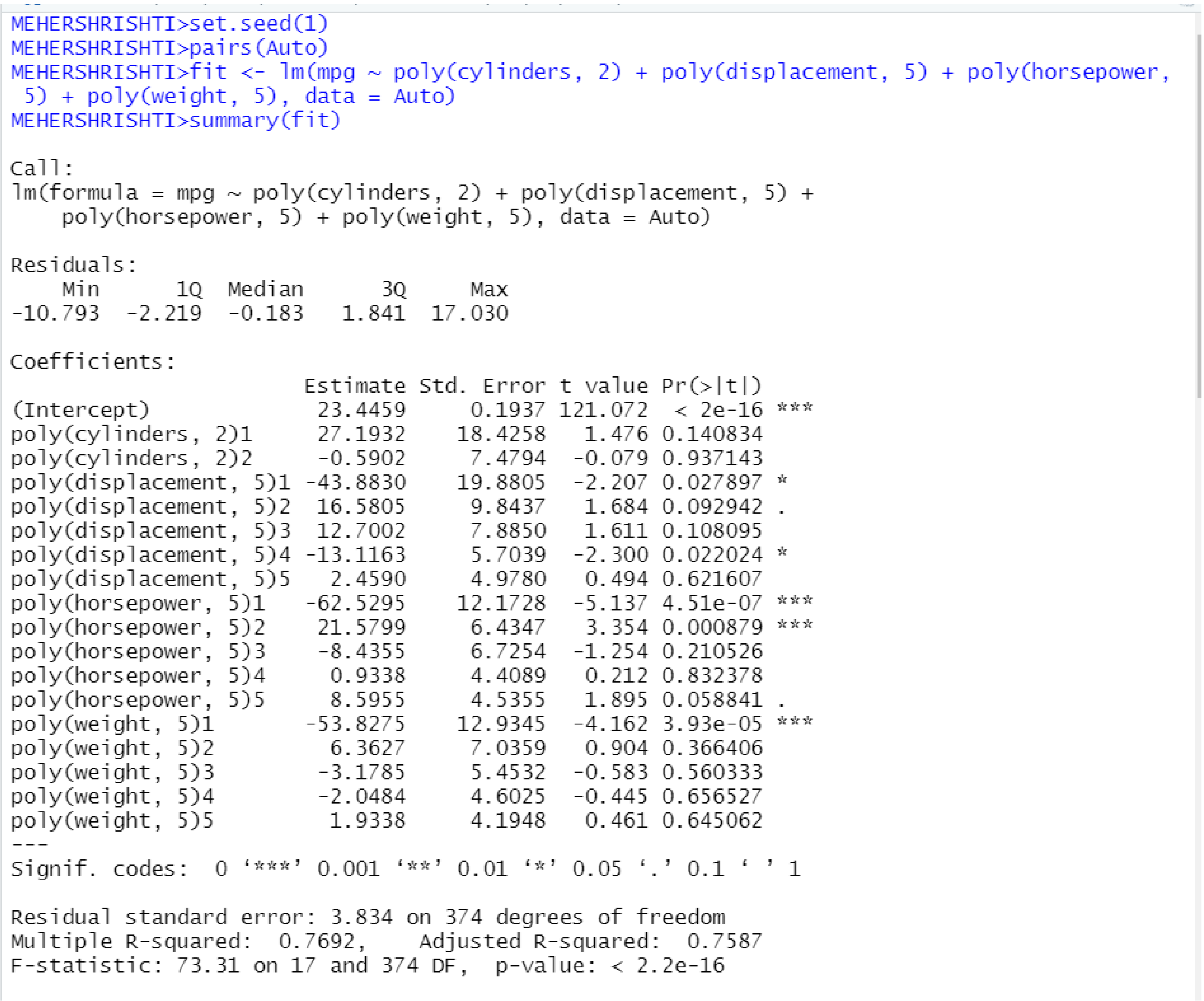
anv2 <- gam(mpg ~ displacement + s(horsepower, 2) + weight, data = Auto)

```
anv3 <- gam(mpg ~ s(displacement, 5) + s(horsepower, 5) + s(weight, 5), data = Auto)
anova(anv1, anv2, anv3, test = 'F')

summary(anv3)

par(mfrow=c(1,3))
plot.Gam(anv3, se=TRUE, col="red")

anv4 <- gam(mpg ~ s(displacement, 3) + s(horsepower, 3) + weight, data = Auto)
anova(anv4, anv3, test = 'F')

par(mfrow=c(1,3))
plot(anv4, se=TRUE, col="red")

lm1 <- glm(mpg ~ displacement + horsepower + weight, data = Auto)
lm2 <- glm(mpg ~ poly(displacement, 3) + poly(horsepower, 3) + weight, data = Auto)
lm3 <- glm(mpg ~ poly(displacement, 5) + poly(horsepower, 5) + poly(weight, 5), data = Auto)
cv.glm(Auto, lm1, K = 10)$delta[1]

cv.glm(Auto, lm2, K = 10)$delta[1]

cv.glm(Auto, lm3, K = 10)$delta[1]
```

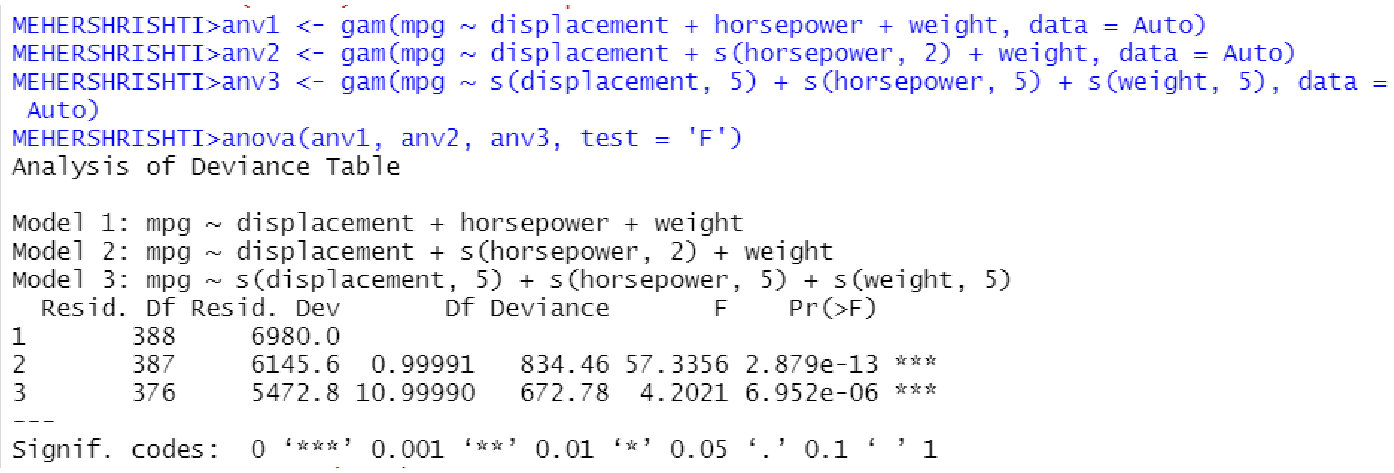

mpg
cylinders
displacement
horsepower
weight
acceleration
year
origin
name
3
5
7
50
150
10
20
1.0
2.0
3.0
30
10
7
5
3
400
100
150
50
4000
1500
20
10
82
76
70
2.5
1.0
150
0
10
30
100
300
1500
4000
70
76
82
0
150
300


s(displacement, 5)
4
2
0
-2
100
200
300
400
displacement
s(horsepower, 5)
10
5
0
-5
-10
50
100
150
200
horsepower
s(weight, 5)
5
0
-5
-10
1500
2500
3500
4500
weight

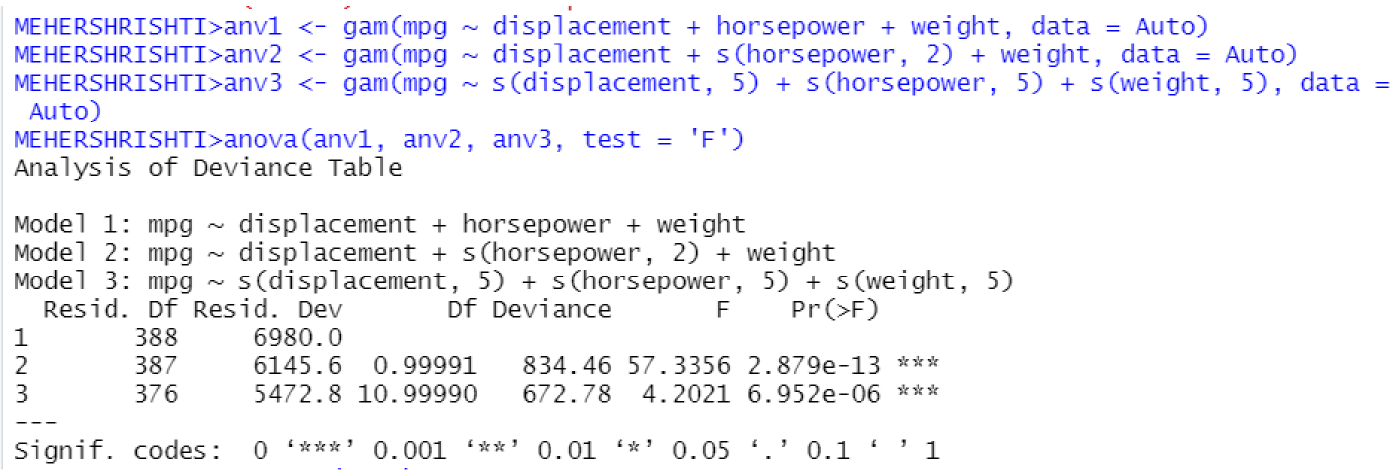

s(displacement, 3)
4
2
0
-2
100
200
300
400
displacement
s(horsepower, 3)
5
0
-5
-10
50
100
150
200
horsepower
partial for weight
5
0
-5
-10
1500
2500
3500
4500
weight



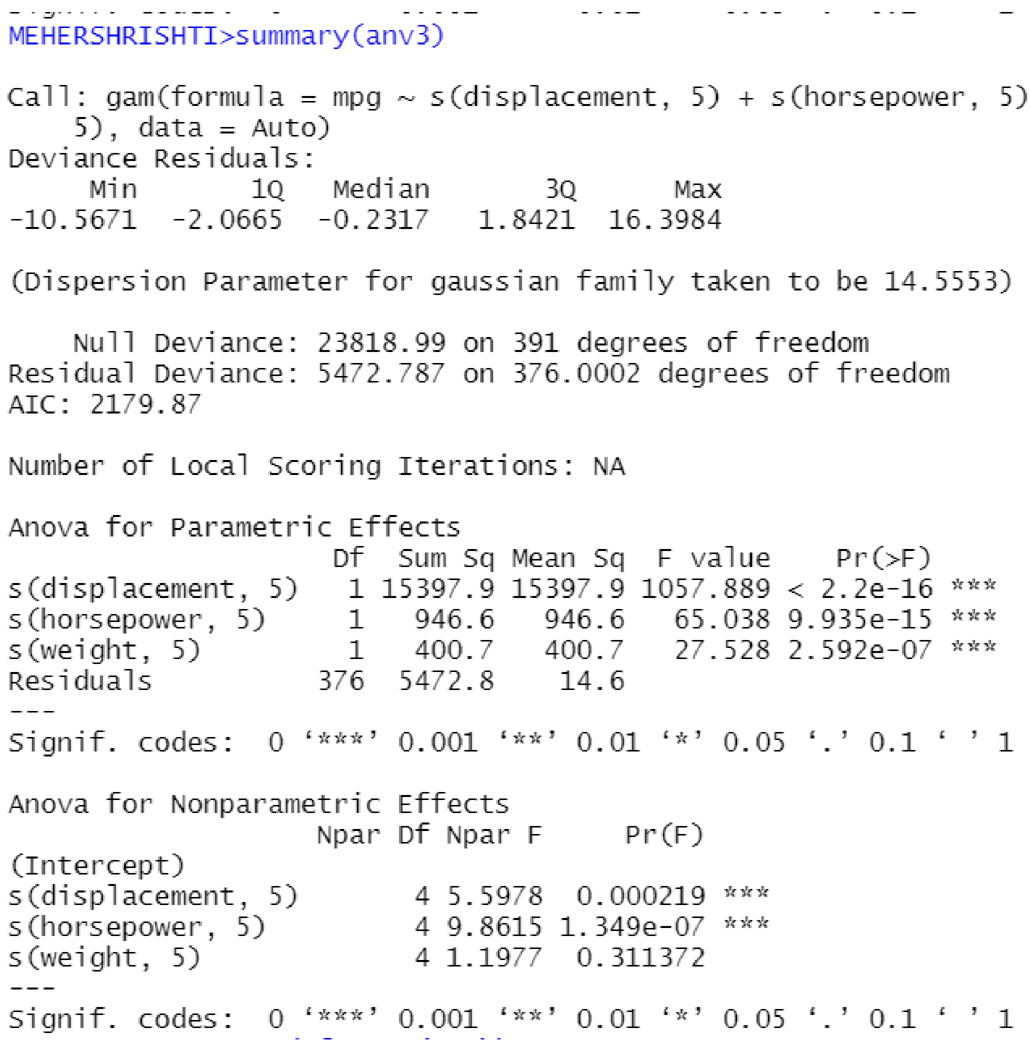
```
MEHERSHRISHTI>set.seed(1)
MEHERSHRISHTI>pairs(Auto)
MEHERSHRISHTI>fit <- lm(mpg ~ poly(cylinders, 2) + poly(displacement, 5) + poly(horsepower,
 5) + poly(weight, 5), data = Auto)
MEHERSHRISHTI>summary(fit)

Call:
lm(formula = mpg ~ poly(cylinders, 2) + poly(displacement, 5) + 
    poly(horsepower, 5) + poly(weight, 5), data = Auto)

Residuals:
    Min      1Q  Median      3Q     Max 
-10.793  -2.219  -0.183   1.841  17.030 

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)    
(Intercept)             23.4459     0.1937 121.072  < 2e-16 ***
poly(cylinders, 2)1     27.1932    18.4258   1.476 0.140834    
poly(cylinders, 2)2     -0.5902     7.4794  -0.079 0.937143    
poly(displacement, 5)1 -43.8830    19.8805  -2.207 0.027897 *  
poly(displacement, 5)2  16.5805     9.8437   1.684 0.092942 .  
poly(displacement, 5)3  12.7002     7.8850   1.611 0.108095    
poly(displacement, 5)4 -13.1163     5.7039  -2.300 0.022024 *  
poly(displacement, 5)5   2.4590     4.9780   0.494 0.621607    
poly(horsepower, 5)1   -62.5295    12.1728  -5.137 4.51e-07 ***
poly(horsepower, 5)2    21.5799     6.4347   3.354 0.000879 ***
poly(horsepower, 5)3    -8.4355     6.7254  -1.254 0.210526    
poly(horsepower, 5)4     0.9338     4.4089   0.212 0.832378    
poly(horsepower, 5)5     8.5955     4.5355   1.895 0.058841 .  
poly(weight, 5)1       -53.8275    12.9345  -4.162 3.93e-05 ***
poly(weight, 5)2         6.3627     7.0359   0.904 0.366406    
poly(weight, 5)3        -3.1785     5.4532  -0.583 0.560333    
poly(weight, 5)4        -2.0484     4.6025  -0.445 0.656527    
poly(weight, 5)5         1.9338     4.1948   0.461 0.645062    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.834 on 374 degrees of freedom
Multiple R-squared:  0.7692,	Adjusted R-squared:  0.7587 
F-statistic: 73.31 on 17 and 374 DF,  p-value: < 2.2e-16
```



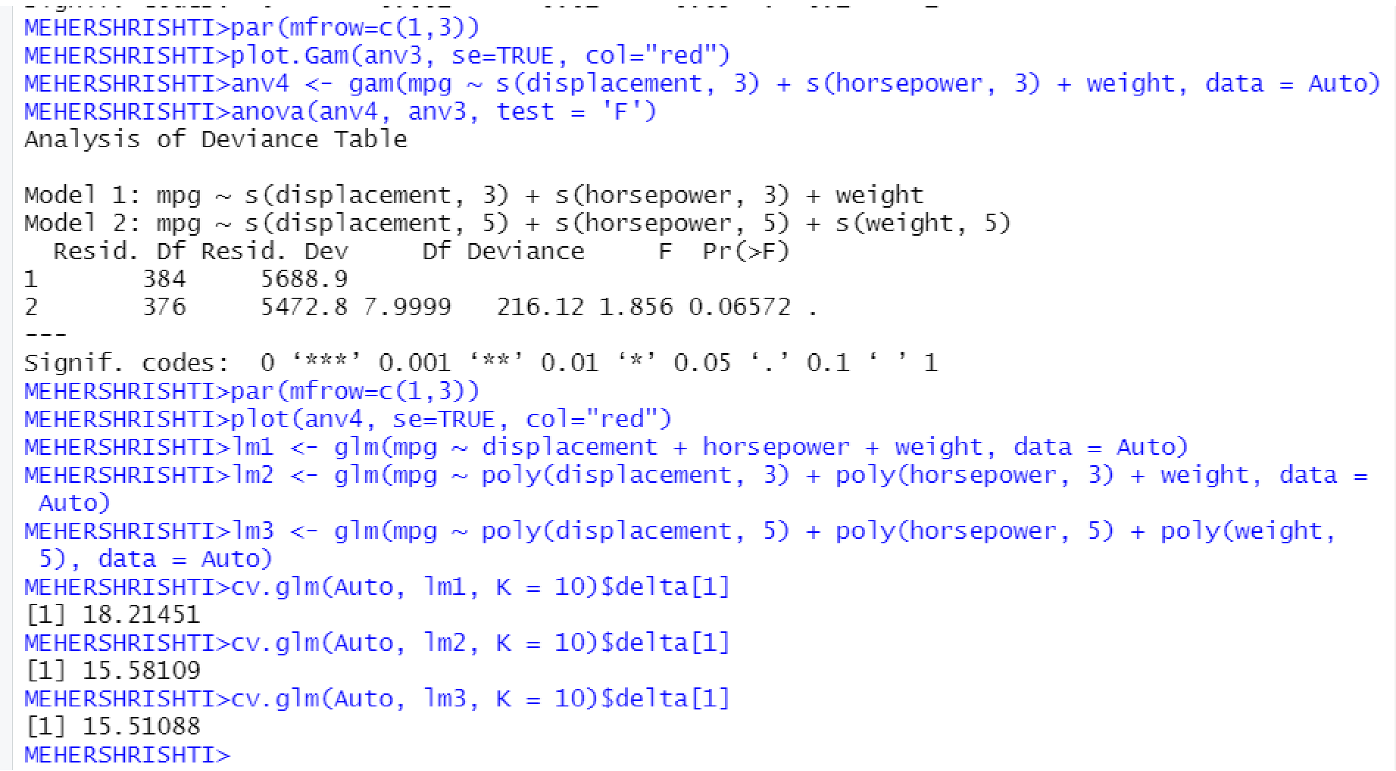
```
MEHERSHRISHTI>anv1 <- gam(mpg ~ displacement + horsepower + weight, data = Auto)
MEHERSHRISHTI>anv2 <- gam(mpg ~ displacement + s(horsepower, 2) + weight, data = Auto)
MEHERSHRISHTI>anv3 <- gam(mpg ~ s(displacement, 5) + s(horsepower, 5) + s(weight, 5), data =
 Auto)
MEHERSHRISHTI>anova(anv1, anv2, anv3, test = 'F')
Analysis of Deviance Table

Model 1: mpg ~ displacement + horsepower + weight
Model 2: mpg ~ displacement + s(horsepower, 2) + weight
Model 3: mpg ~ s(displacement, 5) + s(horsepower, 5) + s(weight, 5)
  Resid. Df Resid. Dev       Df Deviance       F    Pr(>F)    
1       388     6980.0                                        
2       387     6145.6  0.99991   834.46 57.3356 2.879e-13 ***
3       376     5472.8 10.99990   672.78  4.2021 6.952e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```



```
MEHERSHRISHTI>anv1 <- gam(mpg ~ displacement + horsepower + weight, data = Auto)
MEHERSHRISHTI>anv2 <- gam(mpg ~ displacement + s(horsepower, 2) + weight, data = Auto)
MEHERSHRISHTI>anv3 <- gam(mpg ~ s(displacement, 5) + s(horsepower, 5) + s(weight, 5), data =
 Auto)
MEHERSHRISHTI>anova(anv1, anv2, anv3, test = 'F')
Analysis of Deviance Table

Model 1: mpg ~ displacement + horsepower + weight
Model 2: mpg ~ displacement + s(horsepower, 2) + weight
Model 3: mpg ~ s(displacement, 5) + s(horsepower, 5) + s(weight, 5)
  Resid. Df Resid. Dev       Df Deviance       F    Pr(>F)
1       388     6980.0
2       387     6145.6  0.99991   834.46 57.3356 2.879e-13 ***
3       376     5472.8 10.99990   672.78  4.2021 6.952e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```



```
MEHERSHRISHTI>summary(anv3)

Call: gam(formula = mpg ~ s(displacement, 5) + s(horsepower, 5) + s(weight,
    5), data = Auto)
Deviance Residuals:
     Min       1Q   Median       3Q      Max
-10.5671  -2.0665  -0.2317   1.8421  16.3984

(Dispersion Parameter for gaussian family taken to be 14.5553)

    Null Deviance: 23818.99 on 391 degrees of freedom
Residual Deviance: 5472.787 on 376.0002 degrees of freedom
AIC: 2179.87

Number of Local Scoring Iterations: NA

Anova for Parametric Effects
                    Df  Sum Sq Mean Sq  F value    Pr(>F)
s(displacement, 5)   1 15397.9 15397.9 1057.889 < 2.2e-16 ***
s(horsepower, 5)     1   946.6   946.6   65.038 9.935e-15 ***
s(weight, 5)         1   400.7   400.7   27.528 2.592e-07 ***
Residuals          376  5472.8    14.6
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Anova for Nonparametric Effects
                   Npar Df Npar F     Pr(F)
(Intercept)
s(displacement, 5)       4 5.5978  0.000219 ***
s(horsepower, 5)         4 9.8615 1.349e-07 ***
s(weight, 5)             4 1.1977  0.311372
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```



```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
MEHERSHRISHTI>par(mfrow=c(1,3))
MEHERSHRISHTI>plot.Gam(anv3, se=TRUE, col="red")
MEHERSHRISHTI>anv4 <- gam(mpg ~ s(displacement, 3) + s(horsepower, 3) + weight, data = Auto)
MEHERSHRISHTI>anova(anv4, anv3, test = 'F')
Analysis of Deviance Table

Model 1: mpg ~ s(displacement, 3) + s(horsepower, 3) + weight
Model 2: mpg ~ s(displacement, 5) + s(horsepower, 5) + s(weight, 5)
  Resid. Df Resid. Dev     Df Deviance     F  Pr(>F)
1       384     5688.9
2       376     5472.8 7.9999   216.12 1.856 0.06572 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
MEHERSHRISHTI>par(mfrow=c(1,3))
MEHERSHRISHTI>plot(anv4, se=TRUE, col="red")
MEHERSHRISHTI>lm1 <- glm(mpg ~ displacement + horsepower + weight, data = Auto)
MEHERSHRISHTI>lm2 <- glm(mpg ~ poly(displacement, 3) + poly(horsepower, 3) + weight, data =
 Auto)
MEHERSHRISHTI>lm3 <- glm(mpg ~ poly(displacement, 5) + poly(horsepower, 5) + poly(weight,
 5), data = Auto)
MEHERSHRISHTI>cv.glm(Auto, lm1, K = 10)$delta[1]
[1] 18.21451
MEHERSHRISHTI>cv.glm(Auto, lm2, K = 10)$delta[1]
[1] 15.58109
MEHERSHRISHTI>cv.glm(Auto, lm3, K = 10)$delta[1]
[1] 15.51088
MEHERSHRISHTI>
```

The results also suggest model lm2 (same with anv4) is good enough. So the conclusion of relationships with mpg: mpg ~ displacement: cubic; mpg ~ horsepower: cubic; mpg ~ weight: linear.