Assignment #4

BIOINF 525: Module 2

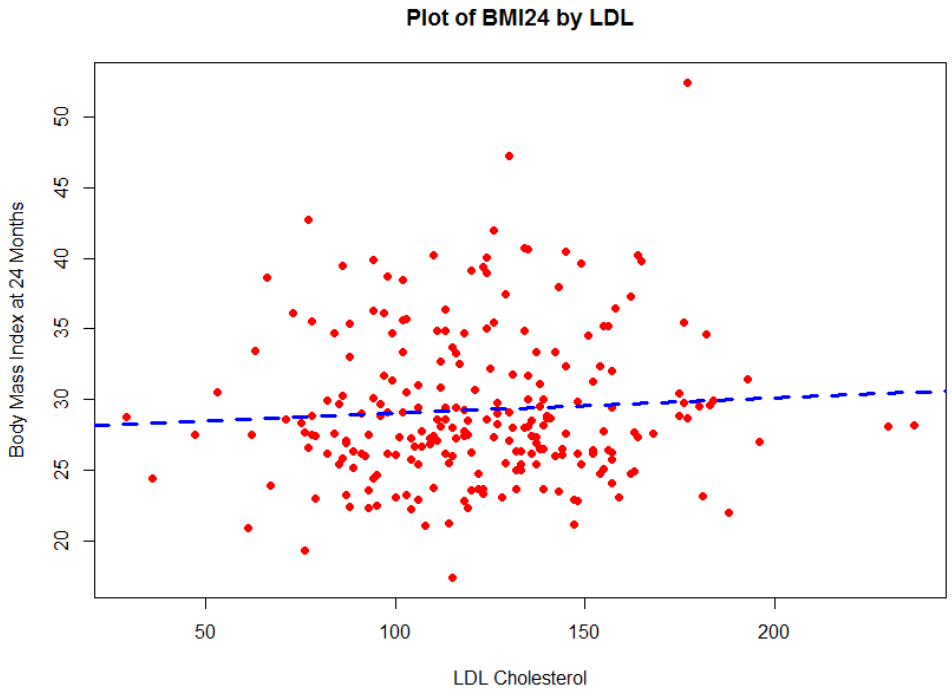
Use the data set from the TROPHY study to answer the following questions. Return the R code, the output, and a brief description of the results.

1. Test whether LDL predict BMI24: H_0: $\rho = 0$ vs. H_A: $\rho \neq 0$.

    a. Use a scatterplot to visually display the relationship between LDL and BMI24. Add a line (best linear fit) to the scatterplot to help identify any linear pattern between LDL and BMI24.

    ```
    plot(LDL, BMI24, xlab="LDL Cholesterol",  ylab="Body Mass Index at 24 Months",
    main="Plot of BMI24 by LDL", pch=16,col="red")

    abline(lm(BMI24~LDL),col="blue",lty=2,lwd=3)
    ```



b. Calculate the Pearson correlation and the corresponding 95%CI between BMI24 and LDL. Is the correlation different from zero? H_0: $\rho = 0$ vs. H_A: $\rho \neq 0$?

```
cor.test(LDL,BMI24)

        Pearson's product-moment correlation

data:  LDL and BMI24
t = 1.0095, df = 229, p-value = 0.3138
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.06306007  0.19397059
sample estimates:
      cor
0.06655937
```

cor=0.6
95%CI=(-0.06.0.19)
p=.3138 > 0.05

Do not reject the H_0. LDL is not related to BMI24.

c. Use simple linear regression to predict BMI24 based on LDL. What is the R-square for the model?

```
out=lm(BMI24~LDL); summary(out)

Multiple R-squared:  0.00443,
Adjusted R-squared:  8.268e-05
```

The multiple R-square is very small indicating that only 0.443% of the variance of BMI24 can be explained by LDL in this linear regression modal.

2. Test whether the association between LDL and BMI24 is modified by treatment (Candesartan). That is, is the relationship between LDL and BMI24 different between treatment group and placebo group H_0: $\rho_1 = \rho_2$ vs. H_A: $\rho_1 \neq \rho_2$.

    a. Use one overlapping scatterplot to visually display the relationship between LDL and BMI24 for Candesartan and Placebo groups. Add a line (best linear fit) for each group and a legend for each line.

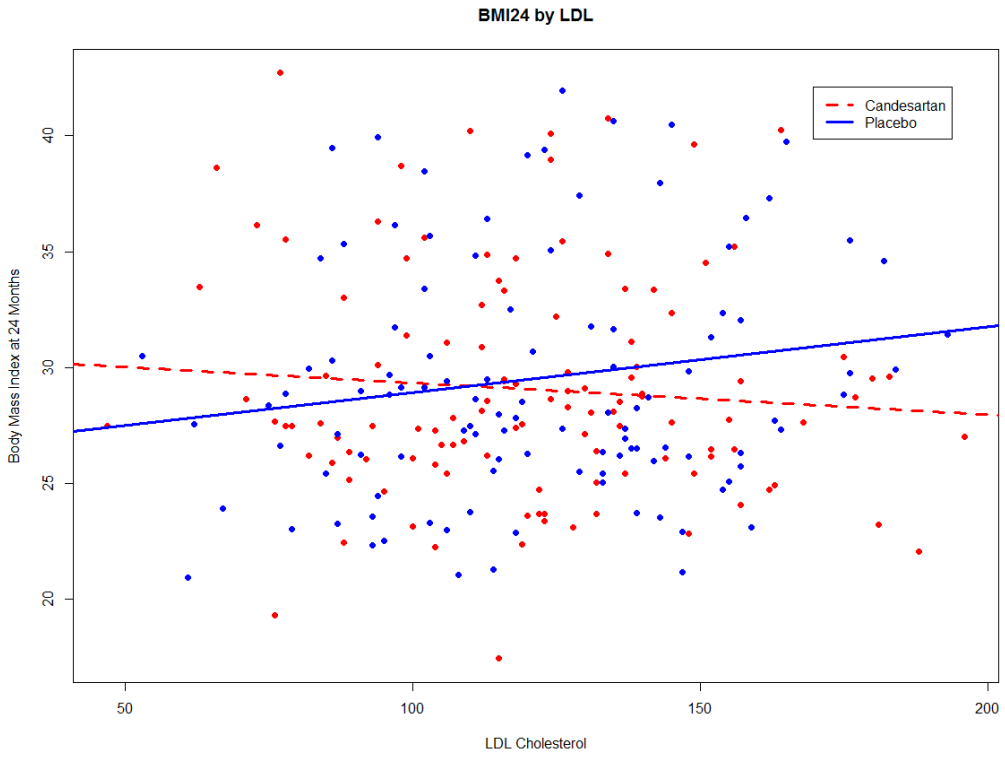
```
plot(LDL[Trt==1],BMI24[Trt==1], xlab='LDL Cholesterol',
    ylab='Body Mass Index at 24 Months', main='BMI24 by LDL', col="red",pch=16)

abline(lm(BMI24[Trt==1]~LDL[Trt==1]),col="red",lty=2,lwd=3)

points(LDL[Trt==2],BMI24[Trt==2],col="blue",pch=16)

abline(lm(BMI24[Trt==2]~LDL[Trt==2]),col="blue",lty=1,lwd=3)

legend(locator(1),c("Candesartan","Placebo"),lty=c(2,1), col=c("red","blue"), lwd=c(3,3))
```



b. Test if the relationship between LDL and BMI24 is different for Candesartan and Placebo. Use multiple regression with BMI24 as outcome, and LDL, Trt, and LDL*Trt as predictors to test this. Is the interaction term (Trt*LDL) significant (p<.05)?

```
out=lm(BMI24~LDL+Trt.c+Trt.c*LDL);summary(out)
```

```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      30.71483    2.06533  14.872   <2e-16 ***
LDL              -0.01365    0.01662  -0.821   0.4124
Trt.cPlacebo     -4.63505    2.73764  -1.693   0.0918 .
LDL:Trt.cPlacebo  0.04197    0.02181   1.924   0.0556 .
```

The interaction term is not significant since p-value=0.0556 > 0.05, though the p-value is close to being <.05.Thus, there is no significant difference between Candesartan and Placebo groups on the association (slopes) of LDL with BMI24.

c. Repeat b), but adjusted for Age and Insulin (add Age and Insulin in the model in b.). Is the interaction Trt*LDL significant, after adjusting for Age and Insulin?

```
out=lm(BMI24~LDL+Trt.c+Trt.c*LDL+Age+Insulin);summary(out)
```

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      28.456620   2.684960  10.599  < 2e-16 ***
LDL              -0.011590   0.015672  -0.740   0.4603
Trt.cPlacebo     -5.241299   2.580048  -2.031   0.0434 *
Age              -0.004158   0.038932  -0.107   0.9150
Insulin           0.224284   0.040267   5.570 7.24e-08 ***
LDL:Trt.cPlacebo  0.043245   0.020539   2.106   0.0364 *
```

After adjusting for Age and Insulin, the p-value for the interaction term (LDL by Trt.c) is p=.0364 < 0. 05. Thus, there is a significant difference between Candesartan and Placebo groups on the association (slopes) of LDL with BMI24 after adjusting for Age and Insulin.

Only one of questions 3) and 4) is required (your pick), the other one is optional.

3. Principal Component Analysis
   a. Run a principal component to reduce the following multivariate data:

Insulin
Gluc_fast
Triglyceride
HDL
LDL

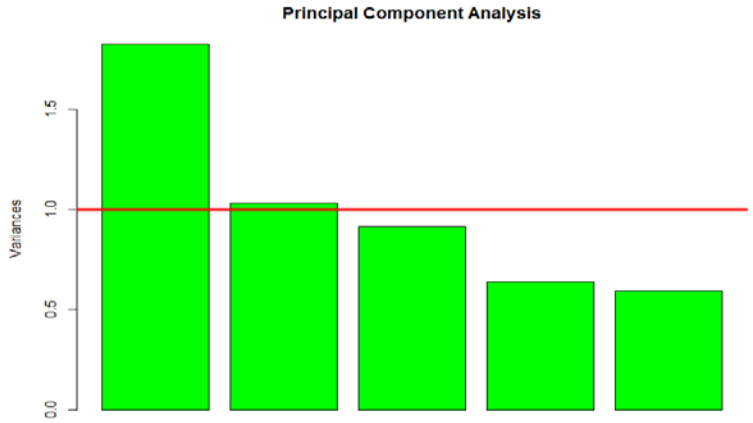
```
data=cbind(Insulin,Gluc_fast,Triglyceride,HDL,LDL)
outpc=prcomp(data,scale=T)
output
```

```
Standard deviations:
[1] 1.3504849 1.0154627 0.9564417 0.7978242 0.7705335

Rotation:
                    PC1        PC2        PC3        PC4         PC5
Insulin      -0.5249003  0.3090300 -0.2790241  0.1176622 -0.73299474
Gluc_fast    -0.4790702  0.1661122 -0.5888989 -0.1132128  0.61909565
Triglyceride -0.4954280 -0.2685043  0.4035067  0.6927943  0.19918561
HDL           0.4829422 -0.0118514 -0.5576990  0.6742998 -0.03029752
LDL          -0.1276297 -0.8970352 -0.3185871 -0.1967086 -0.19709499
```

   b. Plot the variance of principal components (PCs) generated in a). Use the variance > 1 criteria to select PCs to use in the next analysis.

```
plot(outpc, col='green', main='Principal Component Analysis')
abline(a=1,b=0,col="red",lwd=3,lty=3)
```



PC1 and PC2 with variance > 1 are selected for further analysis.

c. What % of the variance is explained (alone and combined) by PCs chosen in part b)?

summary(outpc)

```
Importance of components:
                          PC1    PC2    PC3    PC4    PC5
Standard deviation     1.3505 1.0155 0.9564 0.7978 0.7705
Proportion of Variance 0.3648 0.2062 0.1830 0.1273 0.1187
Cumulative Proportion  0.3648 0.5710 0.7540 0.8813 1.0000
```

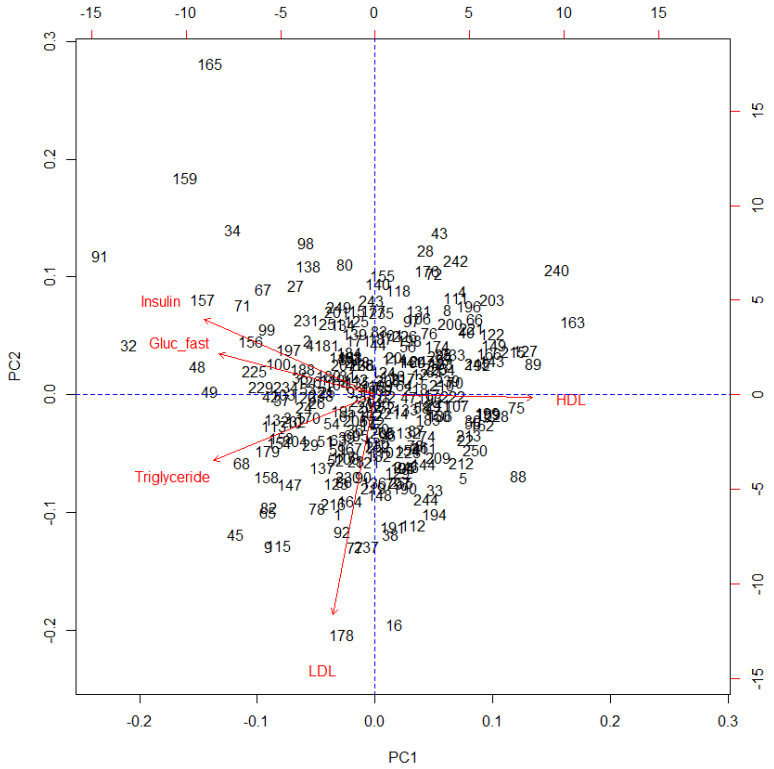
PC1 alone explains 36.5% of the variance and PC2 alone explains 20.6% of the variance. Combined PC1 and PC2 explain 57.1% of the variance

d. Submit a biplot for PC1 and PC2. Based on the biplot identify which of the variables (Insulin, Gluc_fast, Triglyceride, HDL, and LDL) contributes least in PC1, but most in PC2?

biplot(outpc,choice=c(1,2),cex=0.5)

abline(a=0,b=0,col="blue",lty=2)

abline(a=0,b=99999,col="blue",lty=2)



LDL contributes the least for PC1 but the most for PC2.

e. Run a multiple regression to test whether PC1 and PC2 predict BMI24? Does PC1 or PC2 predict BMI24?

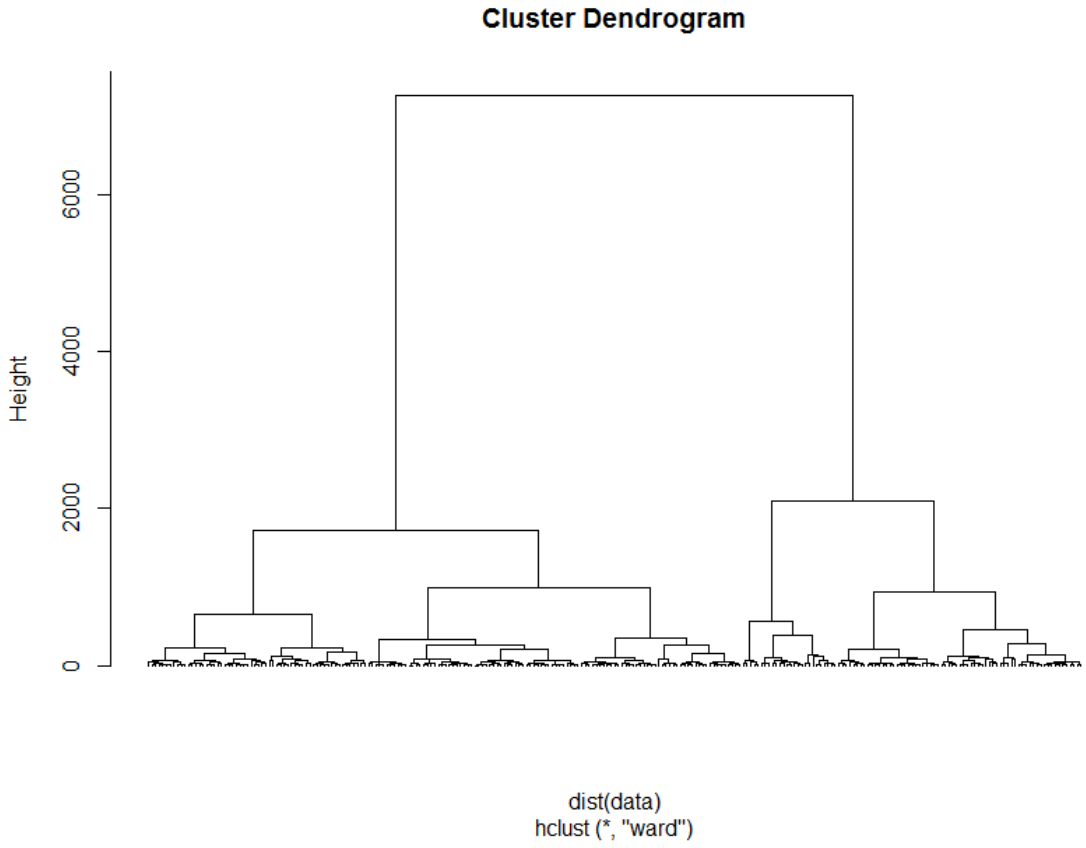
```
PC1=predict(outpc)[,1];PC2=predict(outpc)[,2]
summary(lm(BMI24~PC1+PC2))
```

```
Coefficients:
            Estimate Std. Error t value   Pr(>|t|)
(Intercept)  29.3644     0.3367  87.216   < 2e-16 ***
PC1          -1.0882     0.2495  -4.361  1.96e-05 ***
PC2           0.1037     0.3281   0.316     0.752
```

The p-value for PC1 is p= 1.96e-5 < .05, hence it predicts BMI24. Per one unit increase in PC1 the BMI24 decreases by -1.0882. The p-value for PC2 is p=.752, hence PC2 does not predict BMI24 after adjusting for PC1.

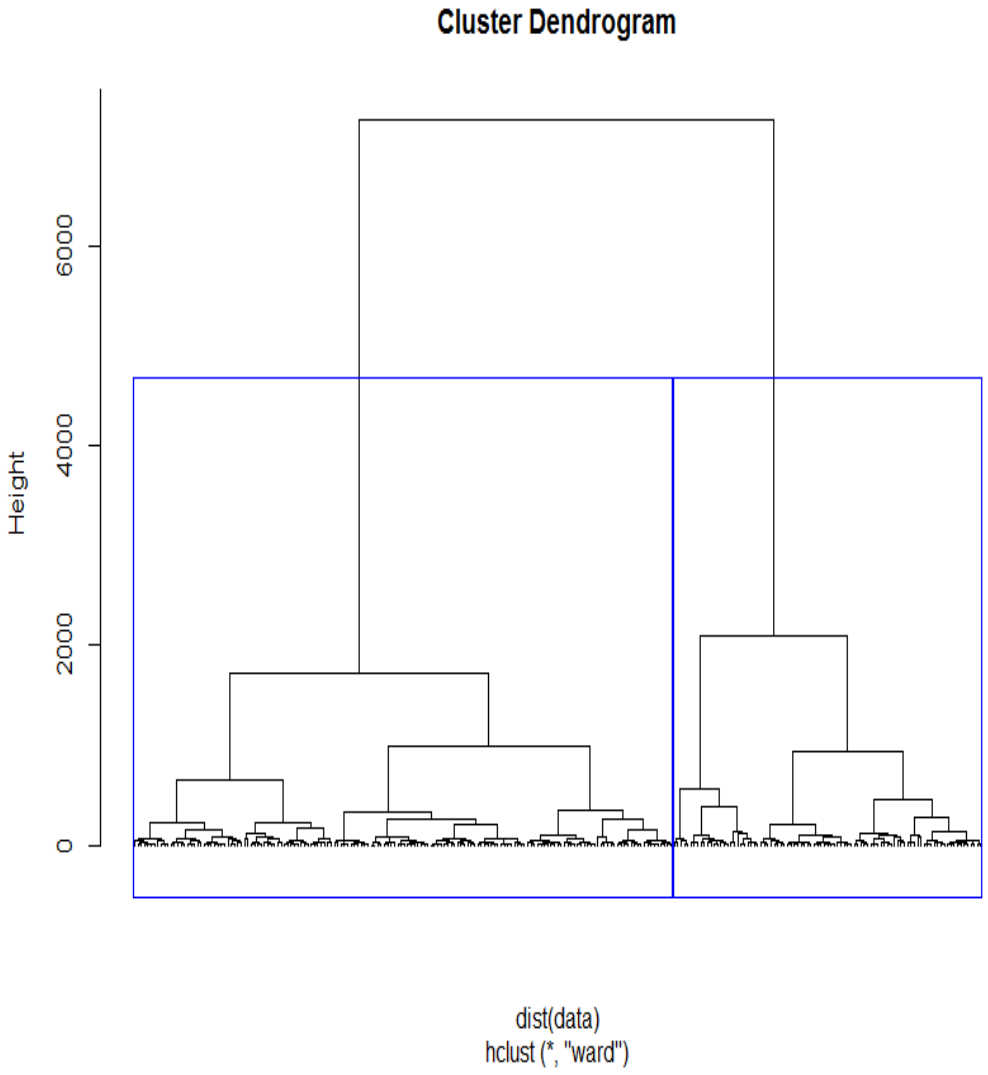
4. Cluster Analysis
   a. Run a hierarchical clustering to group subjects based on similarities on the following variables: BMI, Insulin, Gluc_fast, Triglyceride, HDL, and LDL. Create the Dendrogram

```
data <- cbind(BMI,Insulin,Gluc_fast,Triglyceride,HDL,LDL)
hpg <- hclust(dist(data),method='complete')
plot(hpg,hang=-1,labels=F)
```



b. Identify two clusters on the Dendrogram by using “cutree” function. Save a cluster variable that indicates cluster membership for each subject.

```
rec.hclust(hpg,k=2,border='blue')
cluster=cutree(hpg,k=2)
table(cluster)
```




```
cluster
  1   2
 93 162
```

c. Describe the characteristics of each cluster by populating the following table

```
colMeans(data[cluster==1,]);
colMeans(data[cluster==2,])
var=diag(var(data[cluster==1,]));sqrt(var)
var=diag(var(data[cluster==2,]));sqrt(var)
```

	Cluster 1 (n=93)	Cluster 2 (n=162)
Insulin	13.2±7.8	9.9±8.2
Gluc_fast	97.7±16.5	92.9±9.8
Triglyceride	223.0±61.0	100.8±29.3
HDL	44.9±10.5	51.3±15.2
LDL	132.1±31.1	117.9±31.6

Based on the data from the table cluster 2 is healthier than cluster 1.

d. Use the t-test to test whether BMI24 different between two identified clusters?

```
t.test(BMI24~cluster)
```

```
        Welch Two Sample t-test

data:  BMI24 by cluster
t = 2.6333, df = 138.607, p-value = 0.009416
alternative hypothesis: true difference in means is not equal
to 0
95 percent confidence interval:
 0.5012378 3.5223964
sample estimates:
mean in group 1 mean in group 2
       30.61182        28.60001
```

The p-value from the t-test is p=.009<.05. There is a difference on BMI24 by cluster, where subjects from cluster 2 have on average a lower BMI24.

5. Create a Heatmap for BMI, Insulin, Gluc_fast, Triglyceride, HDL, and LDL. Create the Dendrogram. Use several option for optimal display.

```
library(gplots)
library(RColorBrewer)
selcol2 <- colorRampPalette(brewer.pal(9,"Set1"))
clustcol = selcol2(2)
heatmap.2(data1, col=rainbow(200), scale="column", xlab='Biomarkers',
        ylab='Sample', main='Heat Map', tracecol='black',
        RowSideColors=clustcol[cluster], margins=c(10,10), labRow=F)
```