STAT 762 Assignment 1

Francis Tang UPI: ftan638

Due: 26 March 2019

Question 1

(a) Create a data frame in R. Make sure that each column in your data frame has been specified as the appropriate class. Take an initial look at the data and comment on what you find.

Here I convert the columns 'age', 'cad.dur', and 'choleste' into numerical values. And convert binary values 'tvdlm' and 'sex' into factors.

```
# read data as a new dataframe from txt file
acath.df <- read.csv("~/Desktop/STATS 762/acath.txt", sep="")
str(acath.df)
```

```
## 'data.frame':    1490 obs. of  5 variables:
##  $ sex     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ age     : int  73 68 41 58 81 58 47 66 48 67 ...
##  $ cad.dur : int  132 85 15 7 2 79 6 8 69 48 ...
##  $ choleste: int  268 120 247 168 246 221 272 257 236 274 ...
##  $ tvdlm   : int  1 1 0 0 1 1 0 0 1 1 ...
```

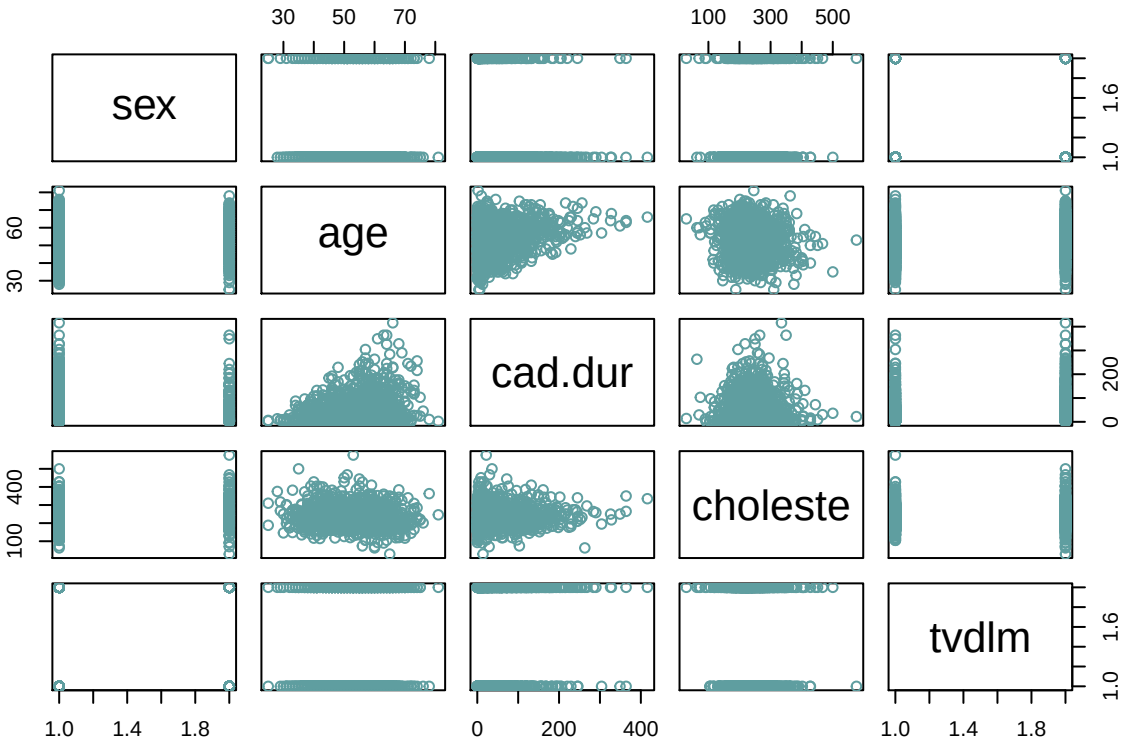
```
# convert three columns into numerical class
acath.df$cad.dur <- as.numeric(acath.df$cad.dur)
acath.df$choleste <- as.numeric(acath.df$choleste)
acath.df$age <- as.numeric(acath.df$age)
acath.df$sex <- as.factor(acath.df$sex)
acath.df$tvdlm <- as.factor(acath.df$tvdlm)
str(acath.df)
```

```
## 'data.frame':    1490 obs. of  5 variables:
##  $ sex     : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ age     : num  73 68 41 58 81 58 47 66 48 67 ...
##  $ cad.dur : num  132 85 15 7 2 79 6 8 69 48 ...
##  $ choleste: num  268 120 247 168 246 221 272 257 236 274 ...
##  $ tvdlm   : Factor w/ 2 levels "0","1": 2 2 1 1 2 2 1 1 2 2 ...
```

```
# get a summary of each column
summary(acath.df)
```

```
##  sex           age           cad.dur          choleste   tvdlm
##  0:1219   Min.   :25.00   Min.   :  0.00   Min.   : 29   0:767
##  1: 271   1st Qu.:46.00   1st Qu.:  6.00   1st Qu.:200   1:723
##           Median :53.00   Median : 23.00   Median :230
##           Mean   :52.29   Mean   : 43.96   Mean   :235
##           3rd Qu.:59.00   3rd Qu.: 61.00   3rd Qu.:265
##           Max.   :81.00   Max.   :416.00   Max.   :576
```

```
# get a paired plot
pairs(acath.df, col = "cadetblue")
```

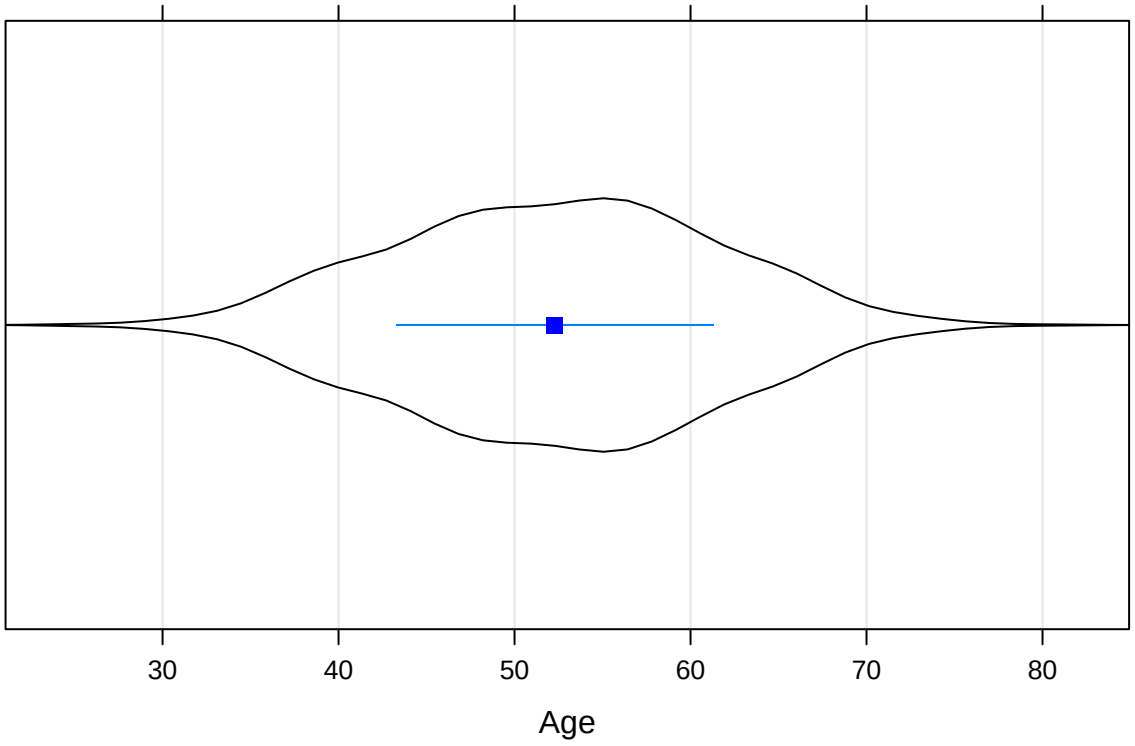


We got to know the mean, median and quatiles of the variables. And the paired plot gave us an rough overview of the relationships between these variables. There might be some correlations among age and cad.dur, choleste and cad.dur. We will analyse them deeper in the next questions.

For the numeric variables we may want to look at some standard plots such as violin plots:

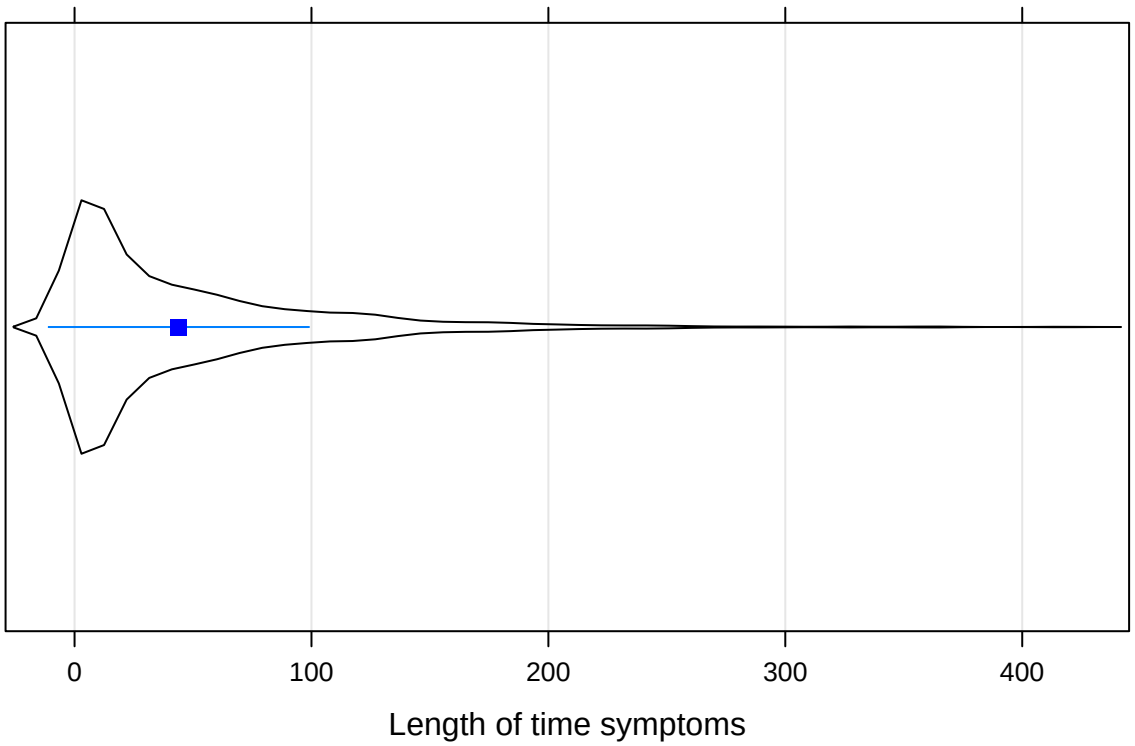
```
library(lattice)
library(violinmplot)
violinmplot(acath.df$age,
            main="Violin plot for age", xlab="Age")
```

```
## Warning in bwplot.numeric(x = x, data = data, panel = panel.violinm, ...):
## explicit 'data' specification ignored
```



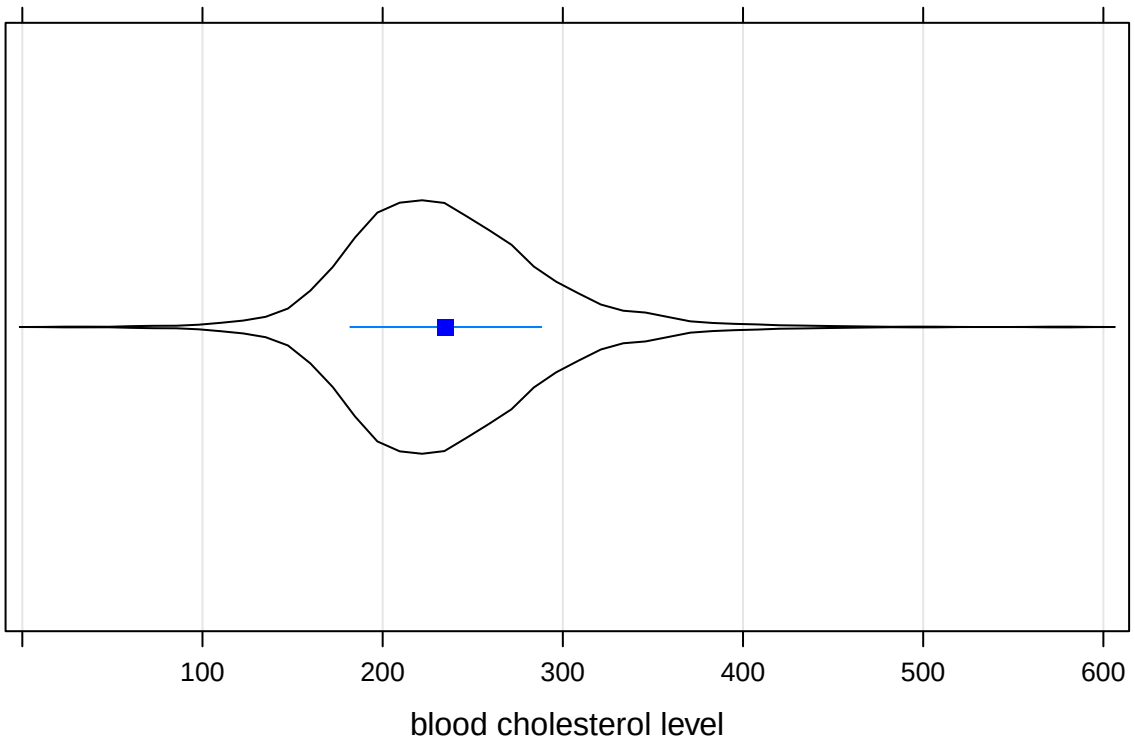

```
violinmplot(acath.df$cad.dur,
            main="Violin plot for the length of time symptoms",
            xlab="Length of time symptoms")
```

```
## Warning in bwplot.numeric(x = x, data = data, panel = panel.violinm, ...):
## explicit 'data' specification ignored
```




```
violinmplot(acath.df$choleste,
            main="Violin plot for blood cholesterol level",
            xlab="blood cholesterol level")
```

```
## Warning in bwplot.numeric(x = x, data = data, panel = panel.violinm, ...):
## explicit 'data' specification ignored
```



The violin plots gave us a primary illustration of how the variables distribute. For example, most of the age are between [40,60], the length of time symptoms distribute very unevenly by having a very dense distribution around [0,100]. Also, cholesteral level distributes mostly around [150,300] with small outliers.

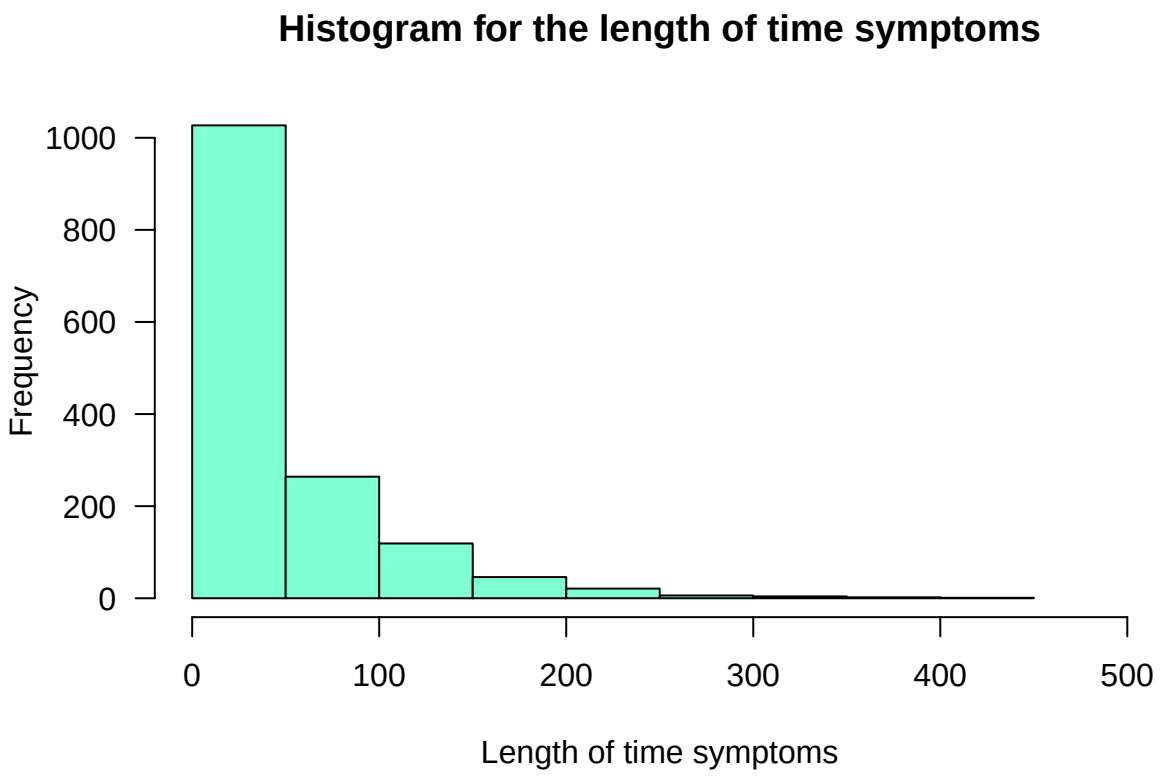
(b) The variables age, cad.dur and cholest are numerical. Suppose we wish to investigate the distribution of each of these characteristics for the given data. For each characteristic produce:

i. A plot that explores the distribution of that characteristic.

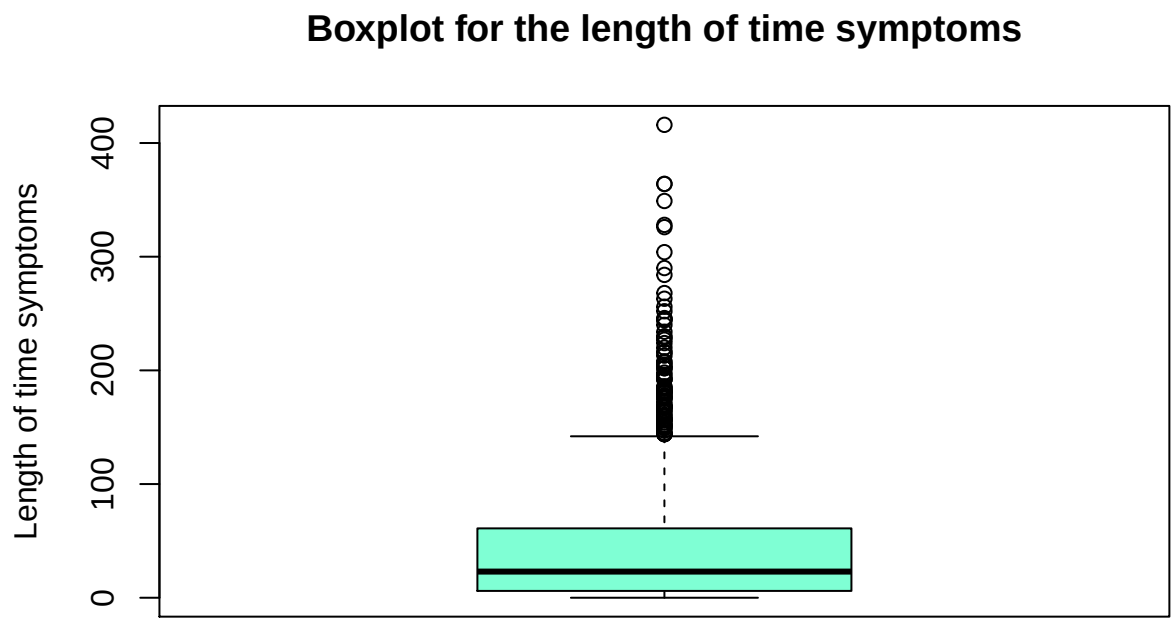
ii. A short (1–3 sentences) description of the distribution.

For the length of time symptoms, the plots below have shown that most of the data distribute densely around [0,100] with some outliers between [100,450].

```
hist(acath.df$cad.dur,
     main="Histogram for the length of time symptoms",
     xlab="Length of time symptoms",
     col="aquamarine",
     xlim=c(0,500),
     las=1,
     breaks=10)
```




```
boxplot(acath.df$cad.dur,
     main="Boxplot for the length of time symptoms",
     ylab="Length of time symptoms",
     col="aquamarine")
```



For the patients' age, the plots below have shown that the distribution of age is pretty much looking alike a normal distribution. The mean and median are around 55 with data distributes evenly between 25 and 81.

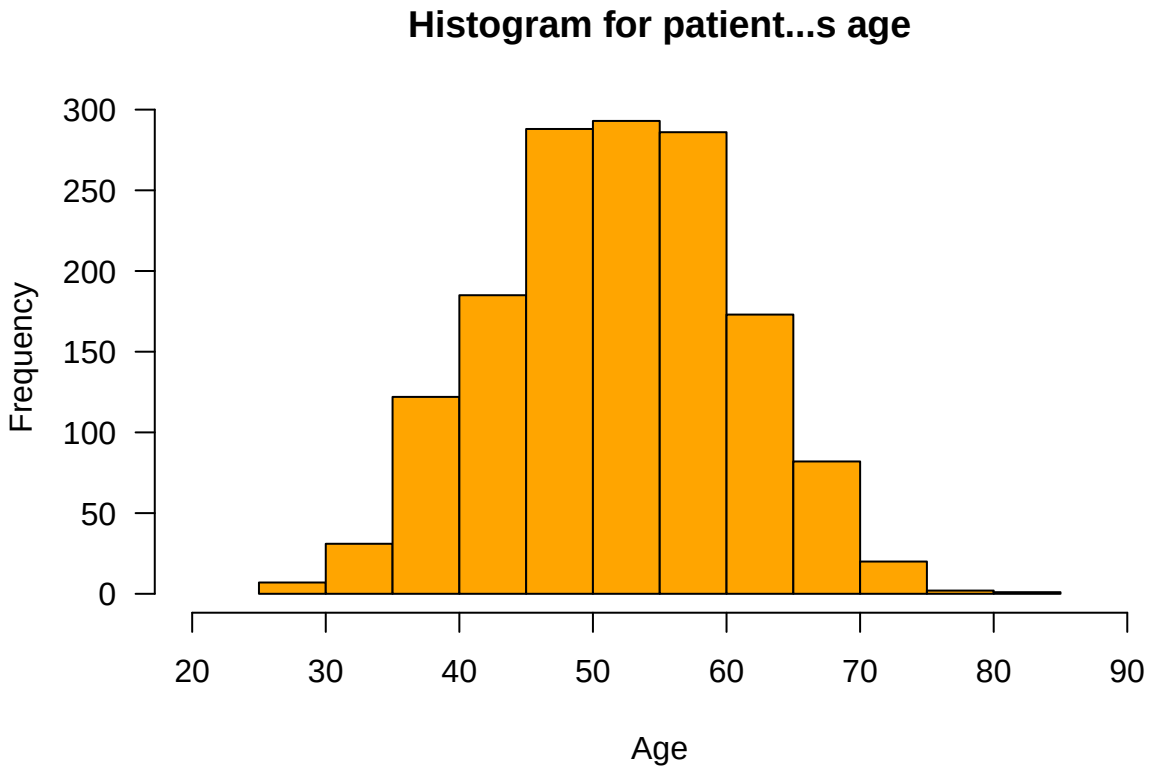
```
hist(acath.df$age,
     main="Histogram for patient's age",
```

```
     xlab="Age",
     col="orange",
     xlim=c(20,90),
     las=1,
     breaks=14)
```

```
## Warning in title(main = main, sub = sub, xlab = xlab, ylab = ylab, ...):
## conversion failure on 'Histogram for patient's age' in 'mbcsToSbcs': dot
## substituted for <e2>
```

```
## Warning in title(main = main, sub = sub, xlab = xlab, ylab = ylab, ...):
## conversion failure on 'Histogram for patient's age' in 'mbcsToSbcs': dot
## substituted for <80>
```

```
## Warning in title(main = main, sub = sub, xlab = xlab, ylab = ylab, ...):
## conversion failure on 'Histogram for patient's age' in 'mbcsToSbcs': dot
## substituted for <99>
```



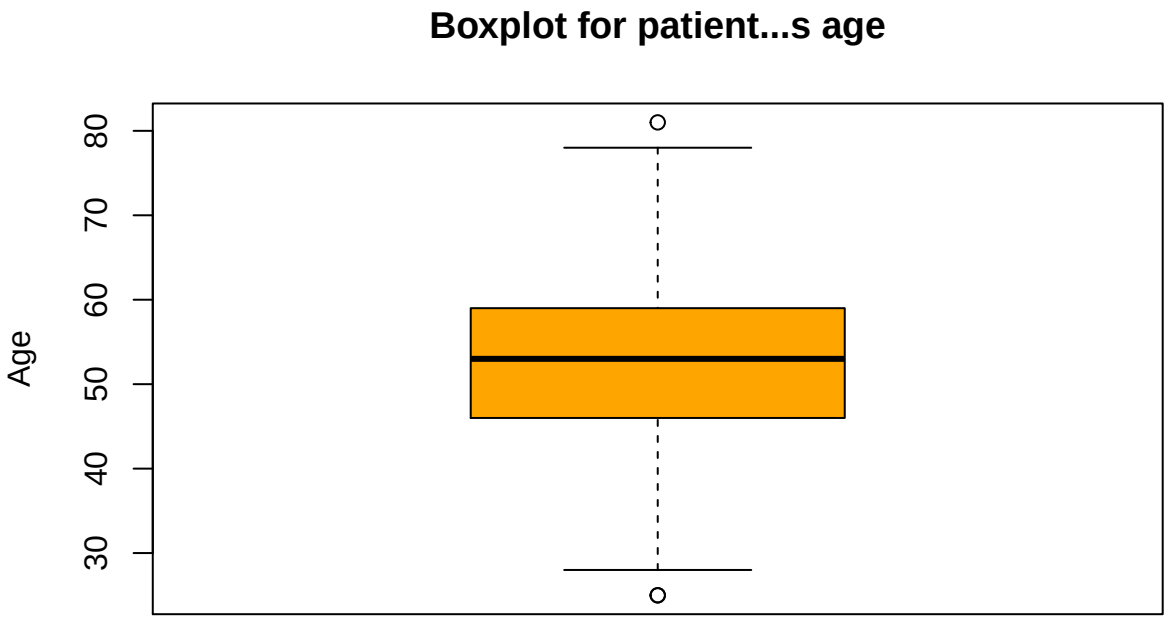

```
boxplot(acath.df$age,
     main="Boxplot for patient's age",
     ylab="Age",
     col="orange")
```

```
## Warning in title(main = "Boxplot for patient's age", ylab = "Age"):
## conversion failure on 'Boxplot for patient's age' in 'mbcsToSbcs': dot
## substituted for <e2>
```

```
## Warning in title(main = "Boxplot for patient's age", ylab = "Age"):
## conversion failure on 'Boxplot for patient's age' in 'mbcsToSbcs': dot
## substituted for <80>
```

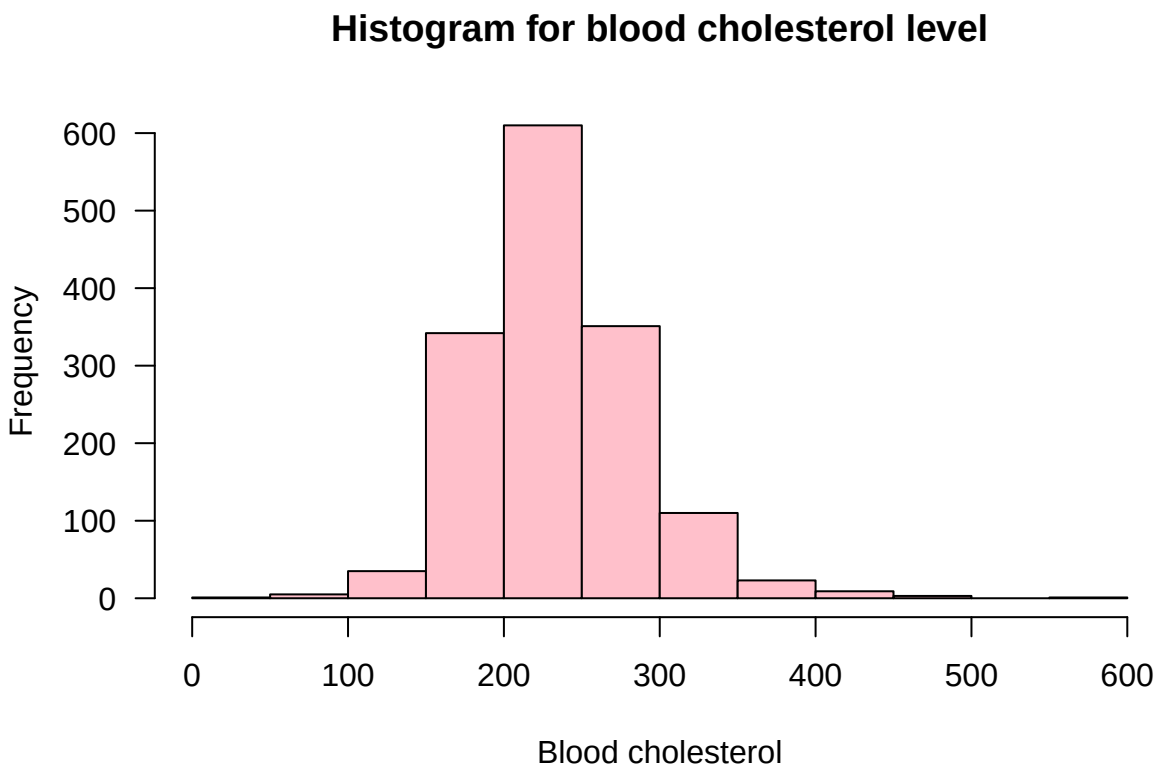
```
## Warning in title(main = "Boxplot for patient's age", ylab = "Age"):
```

```
## conversion failure on 'Boxplot for patient's age' in 'mbcsToSbcs': dot
## substituted for <99>
```

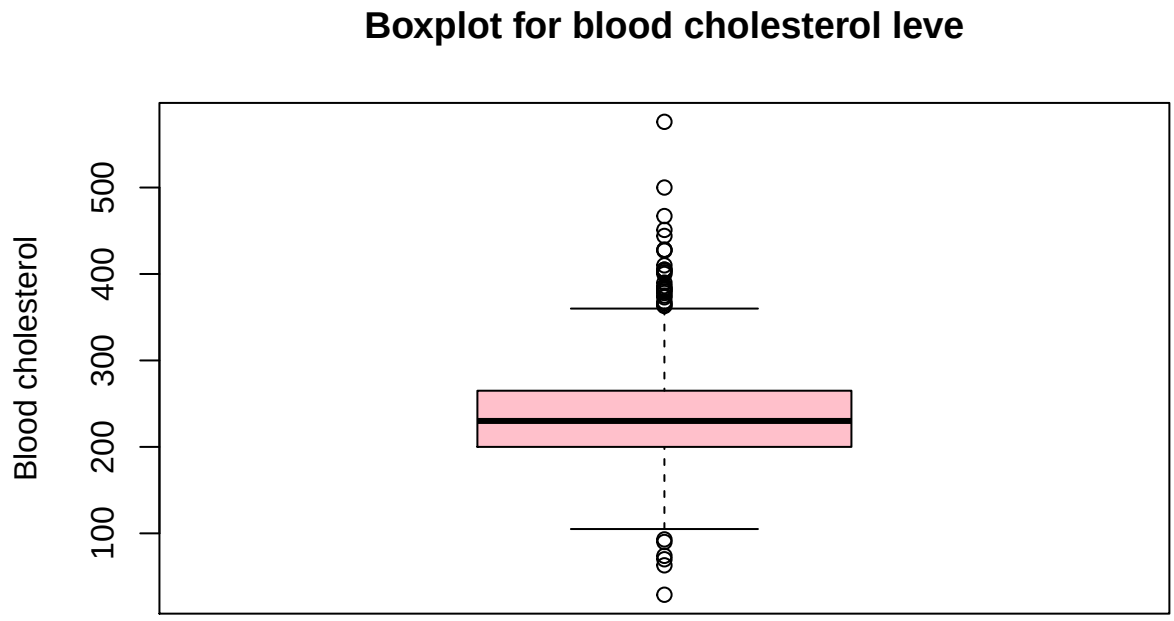


For blood cholesterol level, the plots below have shown that the blood cholesterol level mostly distributes around [150,300]. A few outliers also exist between [0,150] and [400,600].

```
hist(acath.df$choleste,
     main="Histogram for blood cholesterol level",
     xlab="Blood cholesterol",
     col="pink",
     xlim=c(0,600),
     las=1,
     breaks=12)
```




```
boxplot(acath.df$choleste,
     main="Boxplot for blood cholesterol leve",
     ylab="Blood cholesterol",
     col="pink")
```

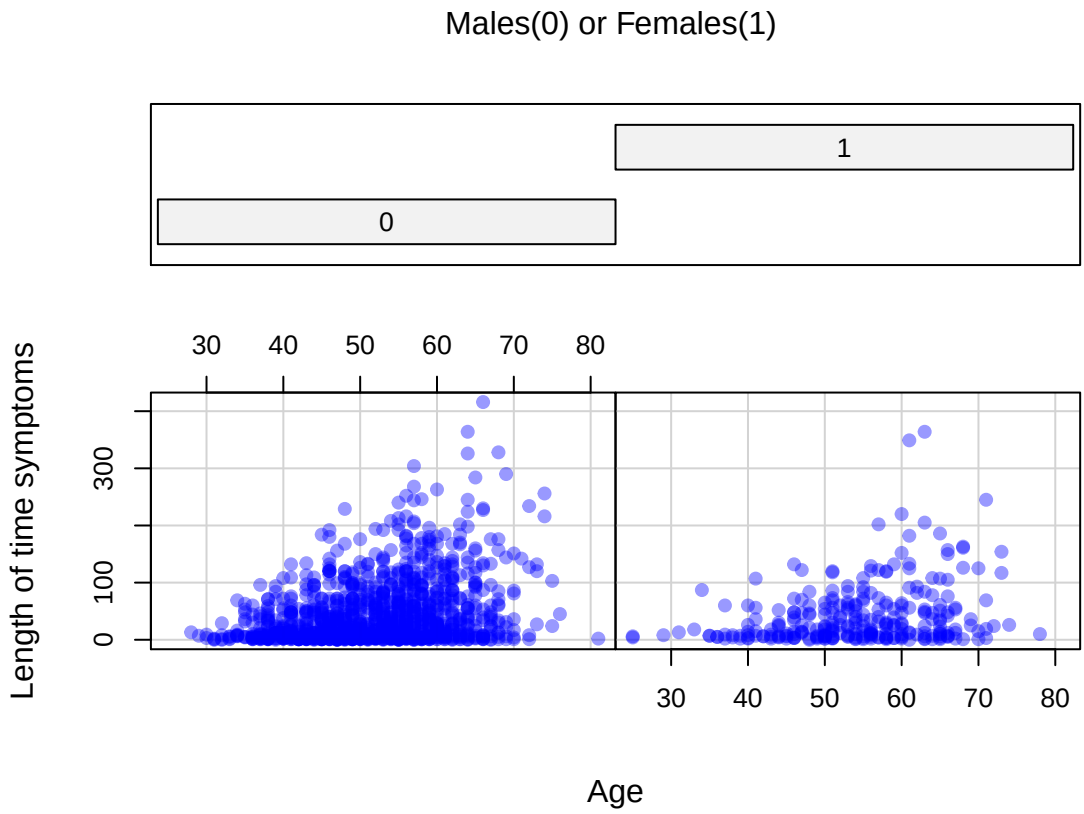


(c) Suppose we wish to investigate how cad.dur is related to age and whether this relationship is different for females and males. Create a suitable plot for this purpose. Briefly describe what you learn from your plot.

Conditional plots are used to how the relationship between two variables is affected by another variable. The conditional plot below gave us an idea that the people from [40,70] age group usually have a longer

time of symptoms. But sex may be a moderate influence, because female appears to have potential less time symptons than male. Although this might be false or biased because of the sample amount difference between male and female.

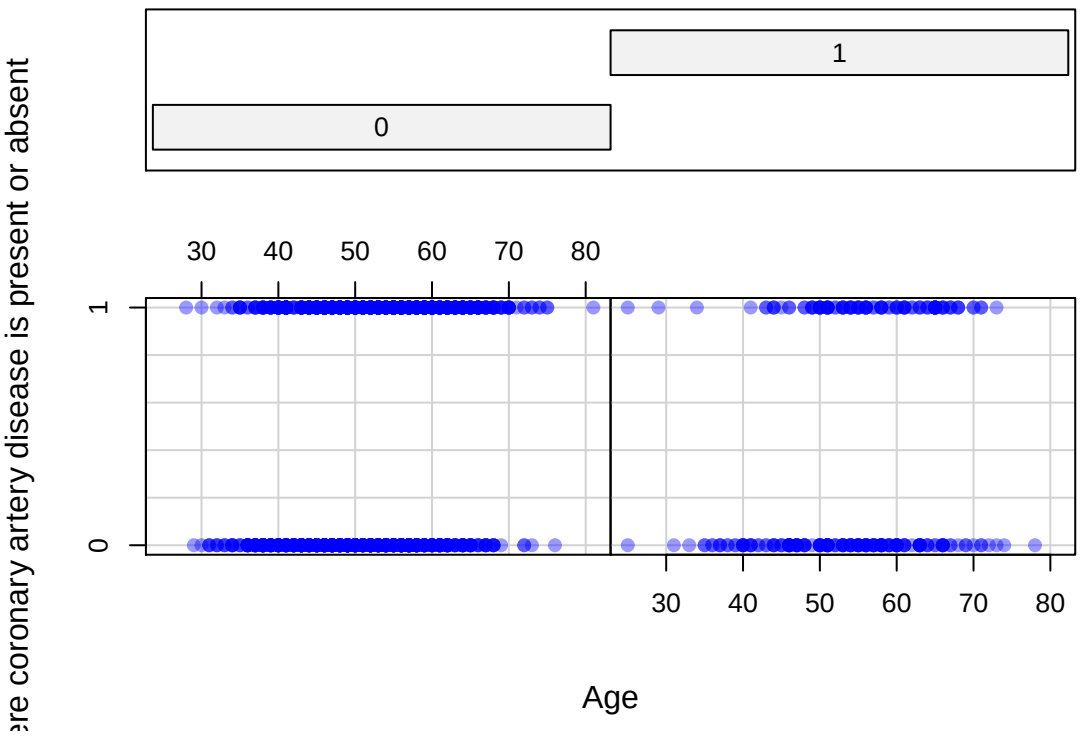
```
#acath.lm<-lm(cad.dur ~ age + sex, data=acath.df)
#summary(acath.lm)
mycol=rgb(0,0,1,alpha=.4)
with(acath.df,coplot(cad.dur~age|sex,pch=19,col=mycol,
                     xlab=c("Age","Males(0) or Females(1) "),
                     ylab="Length of time symptoms"))
```



(d) Suppose we wish to investigate how tvdlm is related to age and whether this relationship is different for females and males.

i. First try creating a plot that leaves age as a numeric variable.

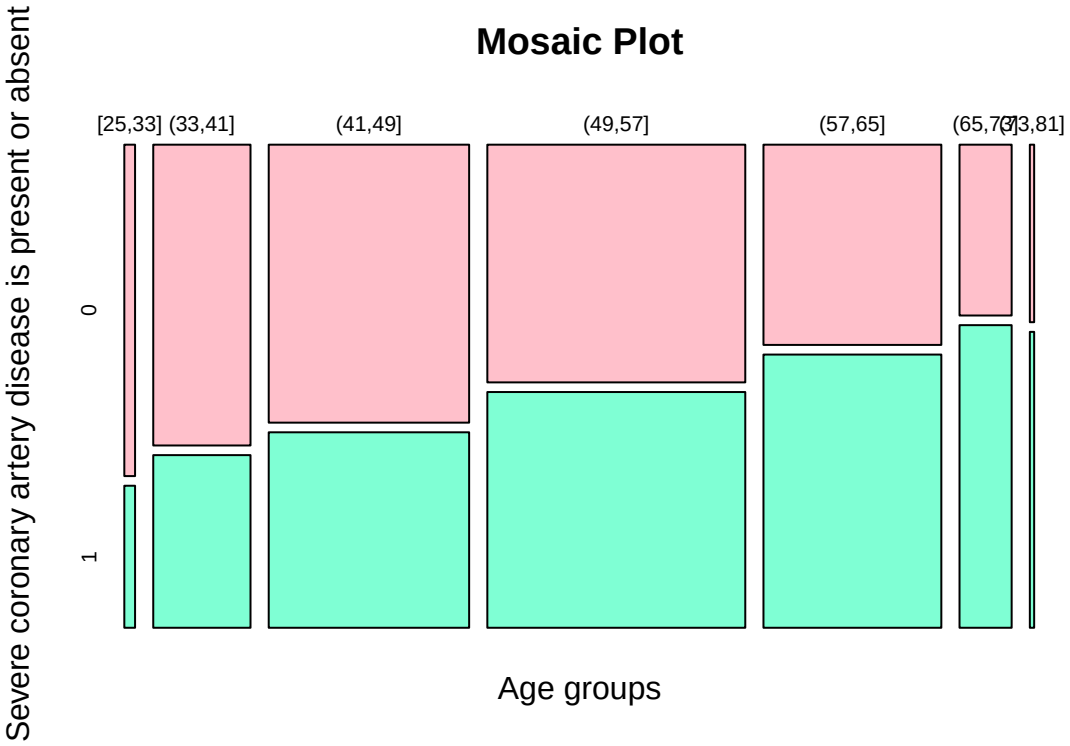
```
mycol=rgb(0,0,1,alpha=.4)
with(acath.df,coplot(tvdlm~age|sex,pch=19,col=mycol,
                     xlab=c("Age","Males(0) or Females(1) "),
                     ylab="Severe coronary artery disease is present or absent"))
```



ii. Now try converting age to a categorical variable where each level is a different age group. It is up to you decide the ranges for your groups. Note that the cut function in R will make task much easier (refer to its help page in R). Create a suitable mosaic plot.

Here I divide the age into 7 different groups with 56 age difference. Each group contains a 8-year-old age difference. It is very clear that older people usually has more present status on severe coronary artery disease than younger people.

```
acath.df$age.group <- cut(acath.df$age, seq(min(acath.df$age),
max(acath.df$age), by = 8), include.lowest=TRUE)
with(acath.df,mosaicplot(table(age.group,tvdlm),col=c("pink","aquamarine"),
                         main="Mosaic Plot",
                         ylab = "Severe coronary artery disease is present or absent",
                         xlab = "Age groups"))
```



Briefly, summarize what you learn from your plots. Which plot did you find more useful in answering this question?

Answer: Mosaic plot is very clear for identify the relationship between age and tvdlm. It points out older can somehow brings more chance of a present status of tvdlm. Conditional plot has a better feature of pointing out how gender plays a role in the relationship between age and tvdlm. And it shows us a potential relationship that females may have less chance of having present for tvdlm, especially among the very young and old people.

(e) Now try fitting some logistic regression models where tvdlm is the response of interest.

i. Fit a logistic regression model that estimates the probability that severe coronary artery disease is present using the other variables as regressors. For this model include all of the other regressors but do not include any interactions. Describe the impact that each of the explanatory variables has on the probability that severe coronary artery disease is present.

```
acath.glm = glm(tvdlm~cad.dur+sex+age+choleste, family=binomial, data=acath.df)
summary(acath.glm)
```

```
##
## Call:
## glm(formula = tvdlm ~ cad.dur + sex + age + choleste, family = binomial,
##     data = acath.df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0506  -1.0867  -0.7265   1.1472   2.0535
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.825543   0.445518  -6.342 2.27e-10 ***
## cad.dur      0.005723   0.001132   5.058 4.25e-07 ***
```

```
## sex1        -0.656261   0.146746  -4.472 7.75e-06 ***
## age          0.037397   0.006534   5.723 1.04e-08 ***
## choleste     0.002900   0.001052   2.757  0.00583 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2064.3  on 1489  degrees of freedom
## Residual deviance: 1959.5  on 1485  degrees of freedom
## AIC: 1969.5
##
## Number of Fisher Scoring iterations: 4
```

According to the output above, all four regressors are related to severe coronary artery disease is present. While age has the most significance, we can believe that when age grows, the chance of being present will grow as well. cad.dur and sex also have significant influence on the 'present' status, while female has a less chance of being 'present' - indicates a nagetive relation, cad.dur has a mild positive relation with being 'present'. At last, choleste has a less significance but still strong enough to determine its relationship with being 'present'. This relation is mildly positive.

ii. Now check for possibility that gender interacts with one or more of the other explanatory variables. If you find evidence that one or more such interactions exist, explain the impact they have on the way the factors involved affect the response.

```
# first convert boolean values back to numerical
#acath.df$sex <- as.numeric(acath.df$sex)
#acath.df$tvdlm <- as.numeric(acath.df$tvdlm)
acath.glm = glm(tvdlm~cad.dur * choleste * age * sex, family = binomial, data = acath.df)
summary(acath.glm)
```

```
##
## Call:
## glm(formula = tvdlm ~ cad.dur * choleste * age * sex, family = binomial,
##     data = acath.df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0567  -1.0690  -0.7225   1.1594   2.3134
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -4.531e+00  2.165e+00  -2.093   0.0364 *
## cad.dur                  -2.331e-02  4.760e-02  -0.490   0.6244
## choleste                  1.058e-02  9.041e-03   1.171   0.2418
## age                       7.082e-02  4.149e-02   1.707   0.0879 .
## sex1                     -4.683e+00  4.905e+00  -0.955   0.3398
## cad.dur:choleste          1.259e-04  1.978e-04   0.637   0.5244
## cad.dur:age               4.682e-04  8.384e-04   0.558   0.5765
## choleste:age             -1.589e-04  1.763e-04  -0.901   0.3675
## cad.dur:sex1              2.703e-02  1.056e-01   0.256   0.7979
## choleste:sex1             1.524e-02  1.911e-02   0.797   0.4252
## age:sex1                  8.239e-02  8.726e-02   0.944   0.3451
## cad.dur:choleste:age     -1.867e-06  3.501e-06  -0.533   0.5938
## cad.dur:choleste:sex1    -1.429e-04  3.963e-04  -0.361   0.7184
## cad.dur:age:sex1         -6.061e-04  1.753e-03  -0.346   0.7295
```

```
## choleste:age:sex1         -2.822e-04  3.435e-04  -0.821   0.4114
## cad.dur:choleste:age:sex1  2.349e-06  6.560e-06   0.358   0.7203
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2064.3  on 1489  degrees of freedom
## Residual deviance: 1939.6  on 1474  degrees of freedom
## AIC: 1971.6
##
## Number of Fisher Scoring iterations: 4
```

```
anova(acath.glm)
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: tvdlm
##
## Terms added sequentially (first to last)
##
##
##                          Df Deviance Resid. Df Resid. Dev
## NULL                                      1489     2064.3
## cad.dur                   1   55.941      1488     2008.3
## choleste                  1    1.528      1487     2006.8
## age                       1   26.699      1486     1980.1
## sex                       1   20.644      1485     1959.5
## cad.dur:choleste          1    0.686      1484     1958.8
## cad.dur:age               1    0.617      1483     1958.2
## choleste:age              1    5.462      1482     1952.7
## cad.dur:sex               1   11.381      1481     1941.3
## choleste:sex              1    0.113      1480     1941.2
## age:sex                   1    0.561      1479     1940.7
## cad.dur:choleste:age      1    0.292      1478     1940.3
## cad.dur:choleste:sex      1    0.041      1477     1940.3
## cad.dur:age:sex           1    0.002      1476     1940.3
## choleste:age:sex          1    0.579      1475     1939.7
## cad.dur:choleste:age:sex  1    0.127      1474     1939.6
```

Both the ANOVA table and the logistic regression model provide the idea that there is no interaction existing in this case according to logistic regression model, there is no significant evidence ($p < .05$) that an iteraction exists among these regressors.

Question 2

(a) Find the orthogonal projection matrix for the subspace of R5 spanned by v1 and v2 (call this subspace S12).

```
v1 = c(1, 1, 0, 1, 1)
v2 = c(3, 2, 1, 2, 3)
V = cbind(v1,v2)
# projection1 is the orthogonal projection matrix for S12
projection1 = V%*%solve(t(V)%*%V)%*%t(V)
projection1
```

```
##       [,1]   [,2]  [,3]   [,4]  [,5]
## [1,] 0.375  0.125  0.25  0.125 0.375
## [2,] 0.125  0.375 -0.25  0.375 0.125
## [3,] 0.250 -0.250  0.50 -0.250 0.250
## [4,] 0.125  0.375 -0.25  0.375 0.125
## [5,] 0.375  0.125  0.25  0.125 0.375
```

(b) Find the orthogonal projection matrix for S|_ 12 (the orthogonal compliment of S12).

```
# projection2 is orthogonal projection matrix for the orthogonal compliment of S12
projection2 = diag(rep(1, 5)) - projection1
projection2
```

```
##        [,1]   [,2]  [,3]   [,4]   [,5]
## [1,]  0.625 -0.125 -0.25 -0.125 -0.375
## [2,] -0.125  0.625  0.25 -0.375 -0.125
## [3,] -0.250  0.250  0.50  0.250 -0.250
## [4,] -0.125 -0.375  0.25  0.625 -0.125
## [5,] -0.375 -0.125 -0.25 -0.125  0.625
```

(c) Find the projection of v3 onto S12 and onto S|_12. Show that these two vectors are orthogonal to each other and that the sum of their squared lengths is equal to the squared length of v3.

```
# generate a projection of v3 onto S12: projection3
v3 = c(2, 2, 0, -3, 1)
projection3 = projection1%*%v3
projection3
```

```
##             [,1]
## [1,] 1.00000e+00
## [2,] 1.44329e-15
## [3,] 1.00000e+00
## [4,] 1.44329e-15
## [5,] 1.00000e+00
```

```
# generate a projection of v3 onto
# the orthogonal compliment of S12: projection4
projection4 = projection2%*%v3
projection4
```

```
##               [,1]
## [1,]  1.000000e+00
## [2,]  2.000000e+00
## [3,] -1.000000e+00
## [4,] -3.000000e+00
## [5,] -1.776357e-15
```

```
# vectors projection3 and projection4 are
# orthogonal only if projection3 ^ t * projection4 = 0

# round to 5 decimal places
round(t(projection3)%*%projection4, 5)
```

```
##      [,1]
## [1,]    0
```

```
round(sum(projection3^2) + sum(projection4^2), 5) == round(sum(v3^2),5)
```

```
## [1] TRUE
```

(d) Find two vectors that form an orthogonal basis for S12.

The following work was done under the reference of:

https://yutsumura.com/find-an-orthonormal-basis-of-the-given-two-dimensional-vector-space/.

First, they are not orthogonal as the dot product is 10:

```
# another vector u1 which is perpendicular to v1 needs to be found
v1%*%v2
```

```
##      [,1]
## [1,]   10
```

Let us first find an orthogonal basis for S12 by the Gram-Schmidt orthogonalization process.

Let w1 := v1. Next, let w2 := v2 + av1, where a is a scalar to be determined so that w1 * w2 = 0.

As w1 and w2 is orthogonal, we have:

w1 * w2 = 0 = v1 * v2 + av1 * v1 = 10 + 4a

It follows that a = -5/2 and:

```
w1 <- v1
w2 <- v2 - 5/2 * v1
w2
```

```
## [1]  0.5 -0.5  1.0 -0.5  0.5
```

Now, to avoid fractions in our computation, let us consider 2w2, instead of w2. Note that the scaling does not change the orthogonality.

We have:

```
2 * w2
```

```
## [1]  1 -1  2 -1  1
```

Thus the set {w1, 2w2} is an orthogonal basis for S12.

However, the length of these vectors are not 1 as we see:

```
lengthw1 = sqrt(1^2 + 1^2 + 0^2 + 1^2 + 1^2)
lengthw1
```

```
## [1] 2
```

```
length2w2 = sqrt(1^2 + (-1)^2 + 1^2 + (-1)^2 + 1^2)
length2w2
```

```
## [1] 2.236068
```

Now it suffices to normalize the vectors w1, 2w2 to obtain an orthonormal basis.

Therefore, the set below is an orthonormal basis for S12.

```
cat("The two vectors form an orthogonal basis for S12 are\n",
    lengthw1 * w1, " and ", length2w2 * w2)
```

```
## The two vectors form an orthogonal basis for S12 are
##  2 2 0 2 2  and  1.118034 -1.118034 2.236068 -1.118034 1.118034
```