Hypothesis Testing Assignment In R

1. QUESTIONS

Question 1: Read the dataset "salary.sav" as a data frame and use the function str() to understand its structure.

The structure of the imported SPSS file is as shown below:

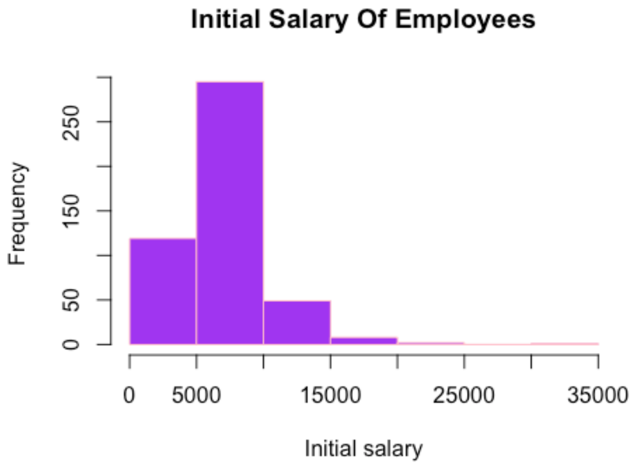
```
'data.frame': 474 obs. of  11 variables:
$ id      : num  1 2 3 4 5 6 7 8 9 10 ...
$ salbeg  : num  8400 24000 10200 8700 17400 ...
$ sex     : Factor w/ 2 levels "MALES","FEMALES": 1 1 1 1 1 1 1 1 1 1 ...
$ time    : num  81 73 83 93 83 80 79 67 96 77 ...
$ age     : num  28.5 40.3 31.1 31.2 41.9 ...
$ salnow  : num  16080 41400 21960 19200 28350 ...
$ edlevel : num  16 16 15 16 19 18 15 15 15 12 ...
$ work    : num  0.25 12.5 4.08 1.83 13 ...
$ jobcat  : Factor w/ 7 levels "CLERICAL","OFFICE TRAINEE",..: 4 5 5 4 5 4 1 1 1 3
...
$ minority: Factor w/ 2 levels "WHITE","NONWHITE": 1 1 1 1 1 1 1 1 1 1 ...
$ sexrace : Factor w/ 4 levels "WHITE MALES",..: 1 1 1 1 1 1 1 1 1 1 ...
- attr(*, "variable.labels")= Named chr  "EMPLOYEE CODE" "BEGINNING SALARY" "SEX OF
EMPLOYEE" "JOB SENIORITY" .....
- attr(*, "names")= chr  "id" "salbeg" "sex" "time" ...
- attr(*, "codepage")= int 1253
```

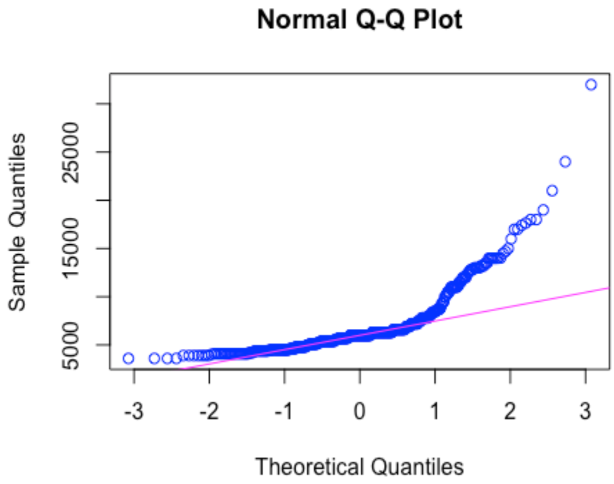
Question 2: Get that summary statistics of the numerical variables in the dataset and visualize their distribution (e.g. use histograms etc). Which variables are normally distributed? Why?

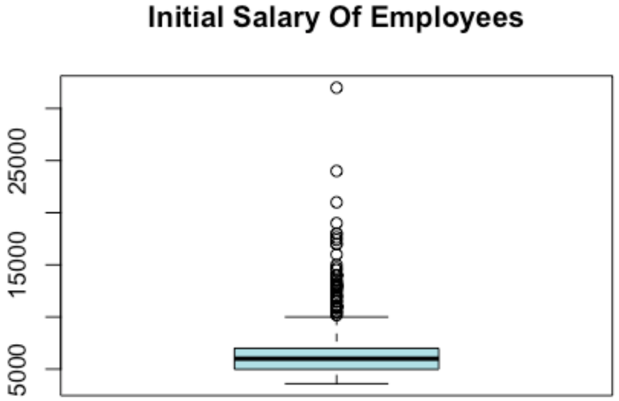
The summary statistics per numerical variable are shown below:

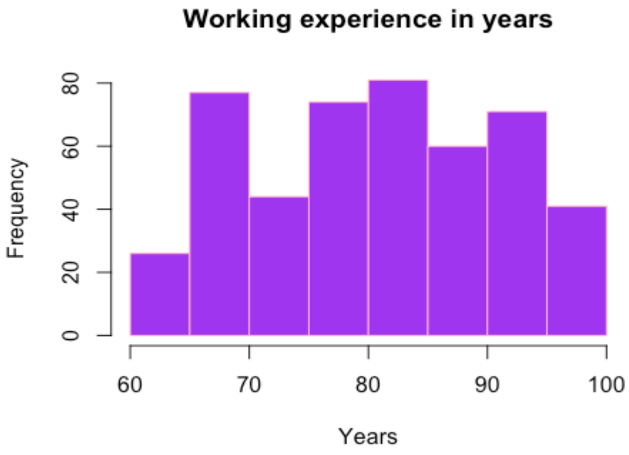
```
 • Initial Salary (mySPSSData$salbeg)
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
3600    4995    6000    6806    6996   31992
 • Time (mySPSSData$time)
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
63.00   72.00   81.00   81.11   90.00   98.00
 • Age (mySPSSData$age)
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
23.00   28.50   32.00   37.19   45.98   64.50
 • Current Salary (mySPSSData$salnow)
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
6300    9600   11550   13768   14775   54000
 • Education Level (mySPSSData$edlevel)
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
8.00   12.00   12.00   13.49   15.00   21.00
 • Work (mySPSSData$work)
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.000   1.603   4.580   7.989  11.560  39.670
```

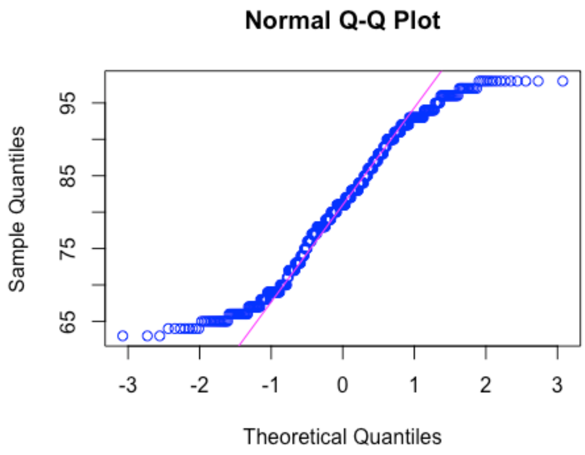
The distribution for each of these variables is visualized as follows:

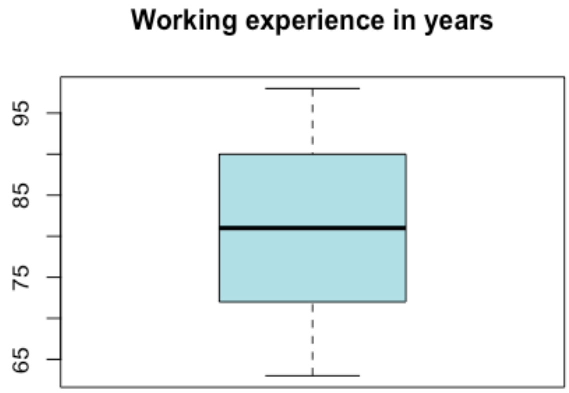

Initial Salary Of Employees
Frequency
0
50
150
250
0
5000
15000
25000
35000
Initial salary

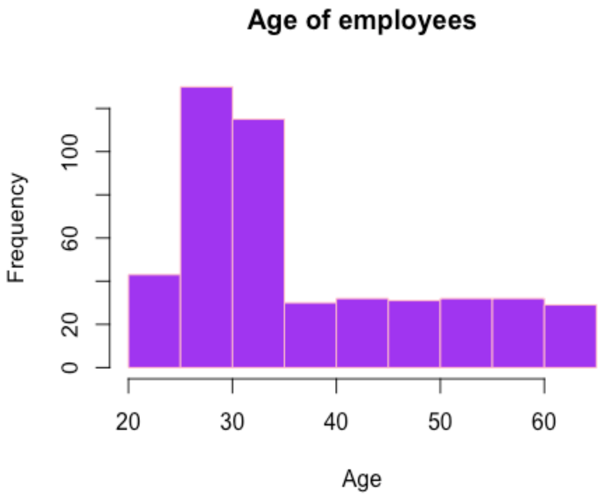

Normal Q-Q Plot
Sample Quantiles
5000
15000
25000
-3
-2
-1
0
1
2
3
Theoretical Quantiles

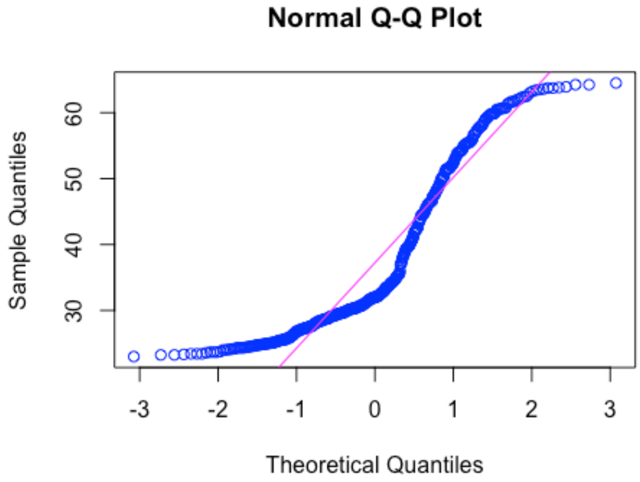

Initial Salary Of Employees
5000
15000
25000

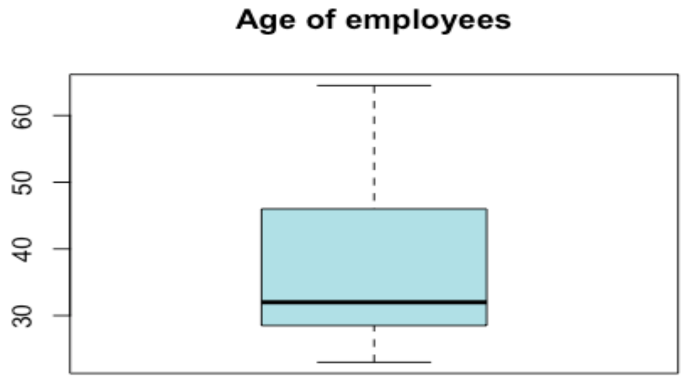

Working experience in years
Frequency
0
20
40
60
80
60
70
80
90
100
Years

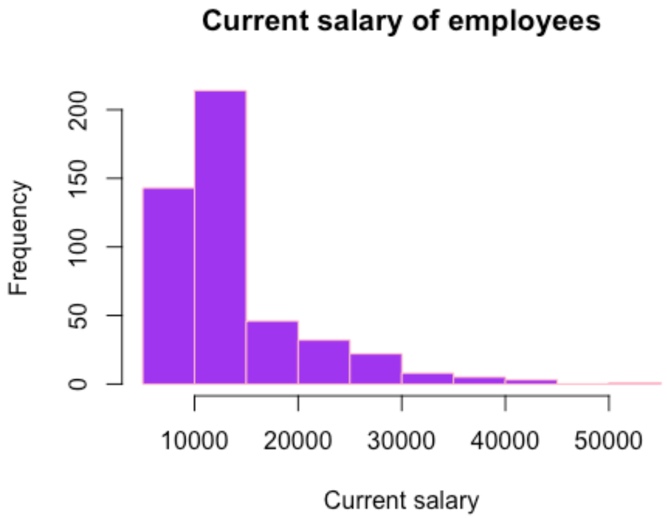

Normal Q-Q Plot
Sample Quantiles
65
75
85
95
-3
-2
-1
0
1
2
3
Theoretical Quantiles

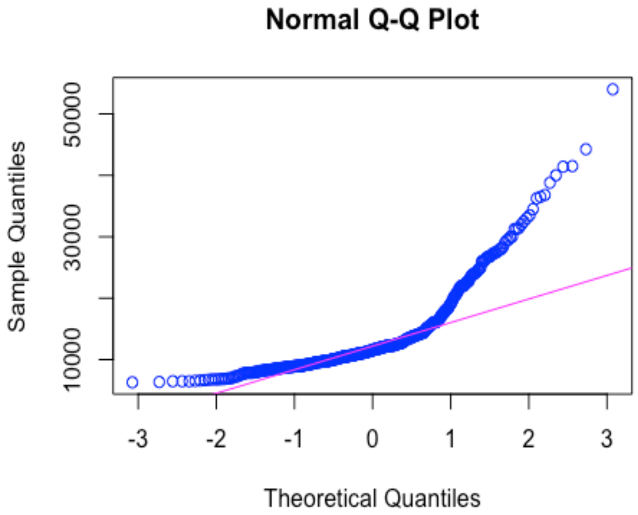

Working experience in years
95
85
75
65

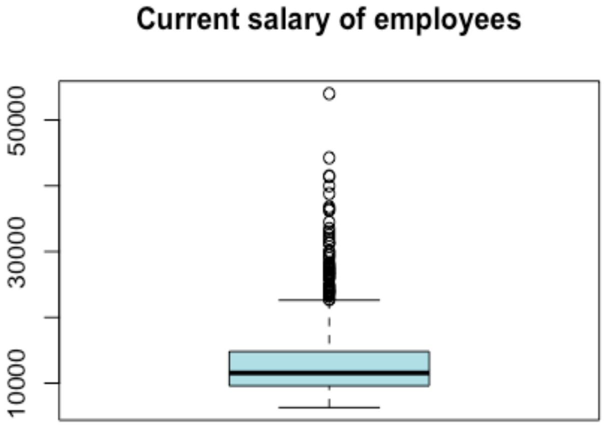

Age of employees
Frequency
0
20
60
100
20
30
40
50
60
Age

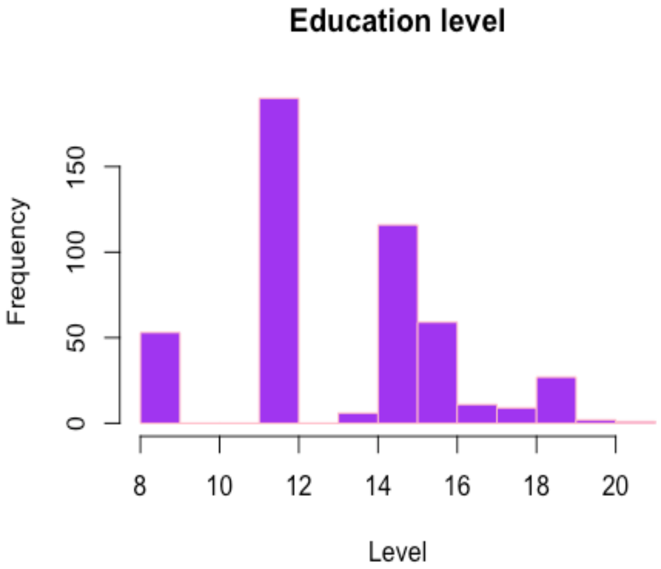

Normal Q-Q Plot
Sample Quantiles
30
40
50
60
-3
-2
-1
0
1
2
3
Theoretical Quantiles

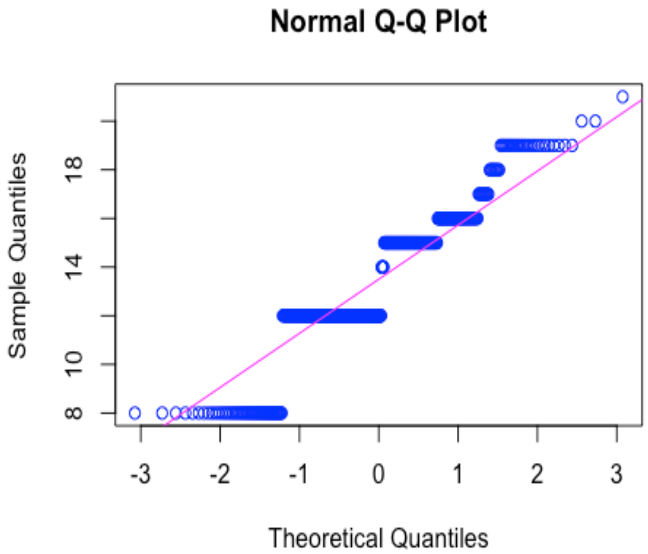

Age of employees
60
50
40
30

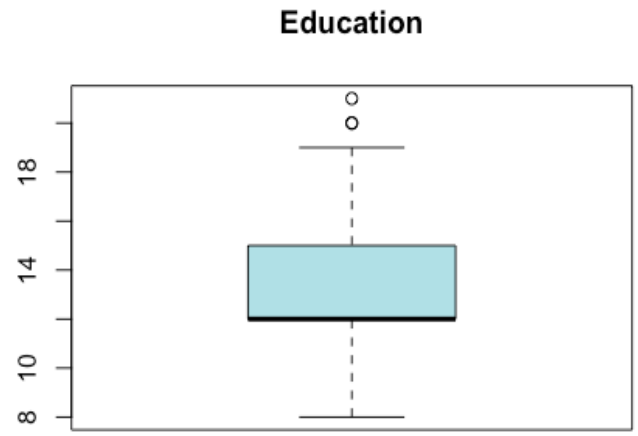

Current salary of employees
Frequency
0
50
100
150
200
10000
20000
30000
40000
50000
Current salary

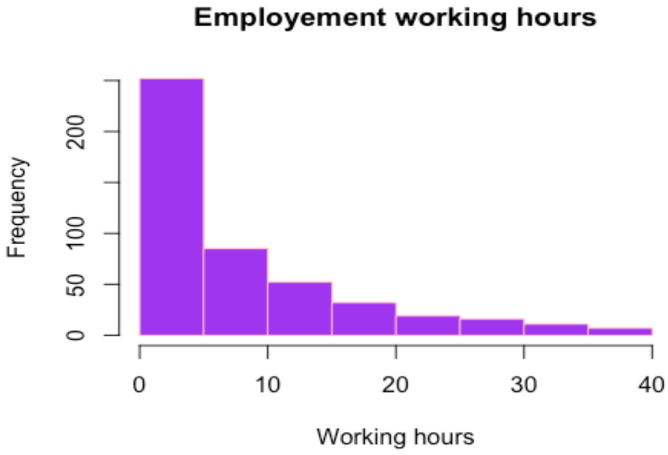

Normal Q-Q Plot
Sample Quantiles
10000
30000
50000
-3
-2
-1
0
1
2
3
Theoretical Quantiles

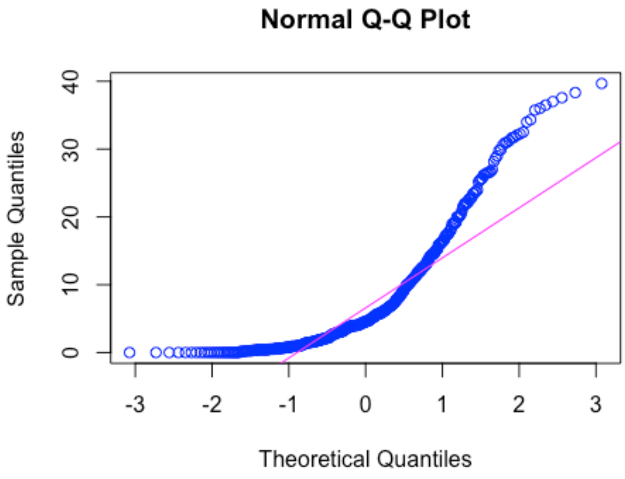

Current salary of employees
10000
30000
50000

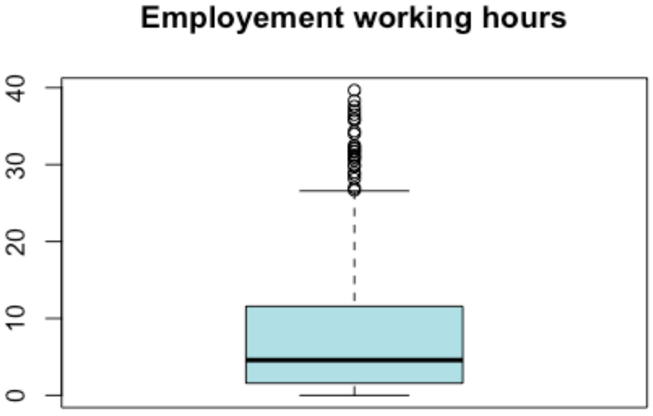

Education level
Frequency
0
50
100
150
8
10
12
14
16
18
20
Level


Normal Q-Q Plot
Sample Quantiles
8
10
14
18
-3
-2
-1
0
1
2
3
Theoretical Quantiles


Education


Employement working hours


Normal Q-Q Plot


Employement working hours

Based on the Q-Q plots presented above it seems that only the variable ***time*** follows normal distribution since the majority of the data points for this variable are on the line of the diagram which simulates the normal distribution. The other Q-Q plots that are related with the rest numeric variables are slightly or 'heavily' skewed which leads to the conclusion that they are not most probably normally distributed. For the variables of the initial salary, the current salary and the work a significant number of outliers exist in the relevant box plots which indicates the presence of many observations that are numerically distant from the majority of the rest data. Also note that no summary statistics and diagrams were calculated and created respectively for the numerical variable ***id*** since this variable is not of statistical importance.

Question 3: Use the appropriate test to examine whether the beginning salary of a typical employee can be considered to be equal to 1000 dollars. How do you interpret the results? What is the justification for using this particular test instead of some other? Explain.

As already mentioned in question 2, the variable ***salbeg*** does not seem to be normally distributed based on the generated Q-Q plot diagram. However, we will check the validity of this statement by performing the Shapiro-Wilk and the Lilliefors (Kolmogorov-Smirnov) normality test since the given sample includes 474 observations (i.e. greater than 50).

```
          Shapiro-Wilk normality test

             data:  initialSalary
        W = 0.71535, p-value < 2.2e-16

  Lilliefors (Kolmogorov-Smirnov) normality test

             data:  initialSalary
        D = 0.25188, p-value < 2.2e-16
```

Based on the retrieved results since p-value < 2.2e-16 < 0.05, we verify that the variable of the initial salary is not normally distributed. After this we check whether the mean of the salbeg variable can be considered a sufficient descriptive measure for the central location (since n > 50).

```
m-out-of-n bootstrap symmetry test by Miao, Gel, and Gastwirth (2006)

                  data:  initialSalary
        Test statistic = 10.18, p-value < 2.2e-16
   alternative hypothesis: the distribution is asymmetric.
                   sample estimates:
                bootstrap optimal m 64
```

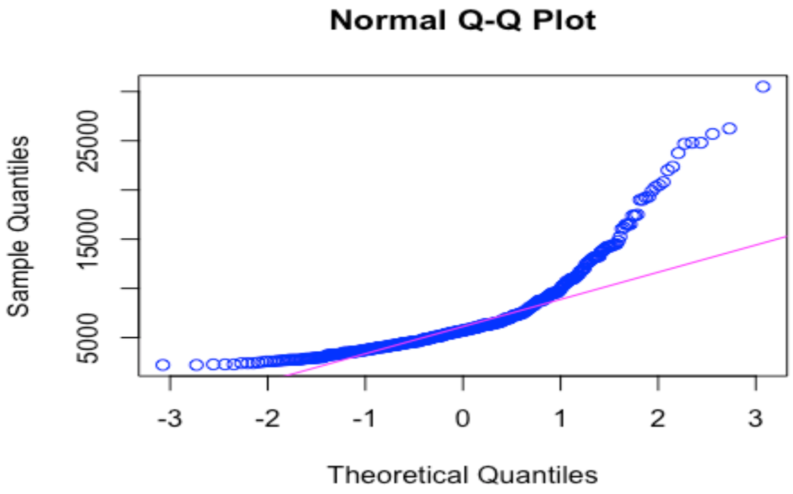
As can be seen in the results of the symmetry test we conclude that the distribution of the ***salbeg*** variable is asymmetric, and this is why we will proceed to a non-parametric hypothesis testing (Wilcoxon sign rank test) for the median of this variable with the null hypothesis to be **H_o: μ = 1.000** against the alternative hypothesis H_1: **μ ≠ 1.000**.

```
                Wilcoxon signed rank test with continuity correction

                              data:  initialSalary
                           V = 112580, p-value < 2.2e-16
              alternative hypothesis: true location is not equal to 1000
```

Since p-value < 2.2e-16 < 0.05, the null hypothesis is rejected in favor of the alternative hypothesis. This is why the beginning salary of the employees cannot be equal to 1.000 dollars.

Question 4: Consider the difference between the beginning salary (salbeg) and the current salary (salnow). Test if the there is any significant difference between the beginning salary and current salary (Hint: Construct a new variable for the difference (salnow – salbeg) and test if, on average, it is equal to zero.). Make sure that the choice of the test is well justified.

As already mentioned in question 2, the variables ***salbeg*** and ***salnow*** do not seem to be normally distributed and so does the difference ***salnow*** – ***salbeg*** (based on the generated Q-Q plot diagram). However, we will check the validity of this statement by performing the Shapiro-Wilk and the Lilliefors (Kolmogorov-Smirnov) normality test on this difference since the given sample includes 474 observations (i.e. greater than 50).



```
Shapiro-Wilk normality test

data:  diff
W = 0.78168, p-value < 2.2e-16

Lilliefors (Kolmogorov-Smirnov) normality test

data:  diff
D = 0.186, p-value < 2.2e-16
```

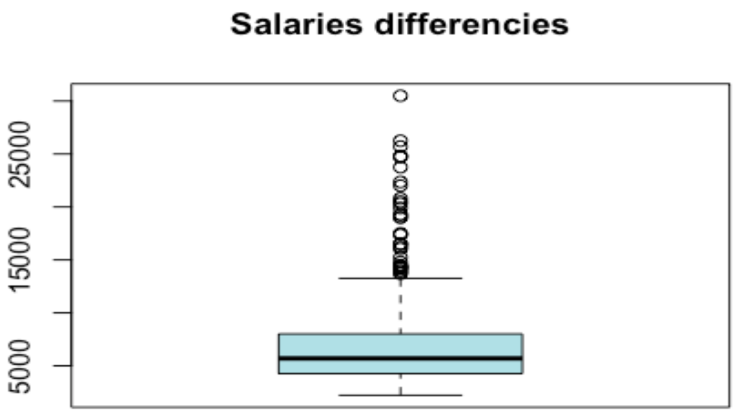
Based on the retrieved results since p-value < 2.2e-16 < 0.05, we verify that the new variable of the salaries difference (salnow – salbeg) is not normally distributed. Now we will check whether the mean of the difference can be considered a sufficient descriptive measure for the central location (since n > 50).

```
m-out-of-n bootstrap symmetry test by Miao, Gel, and Gastwirth (2006)

data:  diff
Test statistic = 10.536, p-value < 2.2e-16
alternative hypothesis: the distribution is asymmetric.
sample estimates:
bootstrap optimal m 115
```

As can be seen in the results of the symmetry test we conclude that the distribution of the salaries difference variable is asymmetric, and this is why we will proceed to a non-parametric hypothesis testing (Wilcoxon sign rank test) for the median of this variable with the null hypothesis to be **H_o: $\mu = 0$** against the alternative hypothesis **H_1: $\mu \neq 0$**.

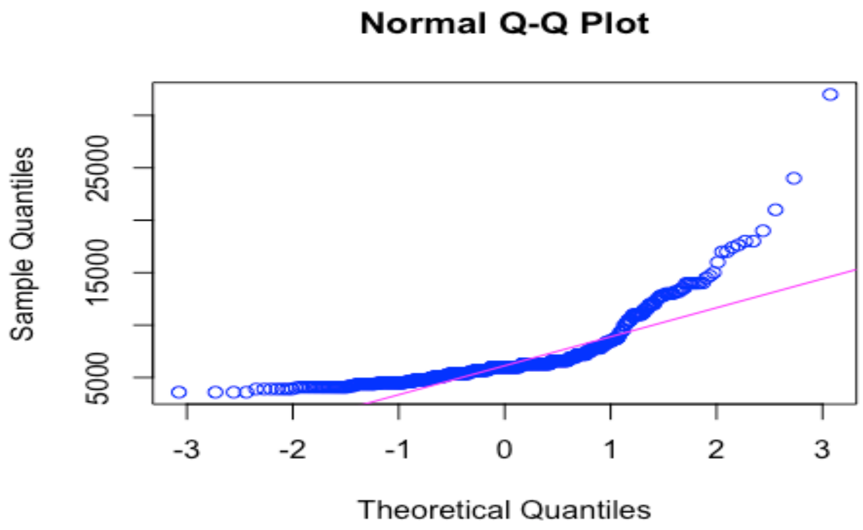
```
Wilcoxon signed rank test with continuity correction

data:  diff
V = 112580, p-value < 2.2e-16
alternative hypothesis: true location is not equal to 0
```

Since p-value < 2.2e-16 < 0.05, the null hypothesis is rejected in favor of the alternative hypothesis. This is why we can consider that the beginning salary cannot be equal to the current salary of the employees. This significant error between the beginning and current salary can be depicted in the below boxplot.



Question 5: Is there any difference on the beginning salary (salbeg) between the two genders? Give a brief justification of the test used to assess this hypothesis and interpret the results.

In this question we will examine the hypothesis testing between the continuous variable ***salbeg*** and the categorical variable ***sex*** for which two groups have been specified, male and female employees. For this reason, a data frame has been created including the beginning salary of the male and female employees which the salary not to be normally distributed, as mentioned in question 2 based on the generated Q-Q plot diagram. However, we will check the validity of this statement by performing the Shapiro-Wilk and the Lilliefors (Kolmogorov-Smirnov) normality test since the given sample includes 474 observations (i.e. greater than 50).



```
allSalaries$sex: Males

Lilliefors (Kolmogorov-Smirnov) normality test

data:  dd[x, ]
D = 0.25863, p-value < 2.2e-16

--------------------------------------------------------

allSalaries$sex: Females

Lilliefors (Kolmogorov-Smirnov) normality test

data:  dd[x, ]
D = 0.14843, p-value = 1.526e-12

allSalaries$sex: Males

Shapiro-Wilk normality test

data:  dd[x, ]
W = 0.73058, p-value < 2.2e-16

--------------------------------------------------------

allSalaries$sex: Females

Shapiro-Wilk normality test

data:  dd[x, ]
W = 0.85837, p-value = 2.98e-13
```

Based on the retrieved results since p-value < 0.05, we verify that normal distribution does not exist. Now we will check whether the mean of the beginning salary can be considered a sufficient descriptive measure for the central location of both groups (since n_M, $n_F > 50$ where n_M is the sample of the male employees and n_F is the sample of the female employees).

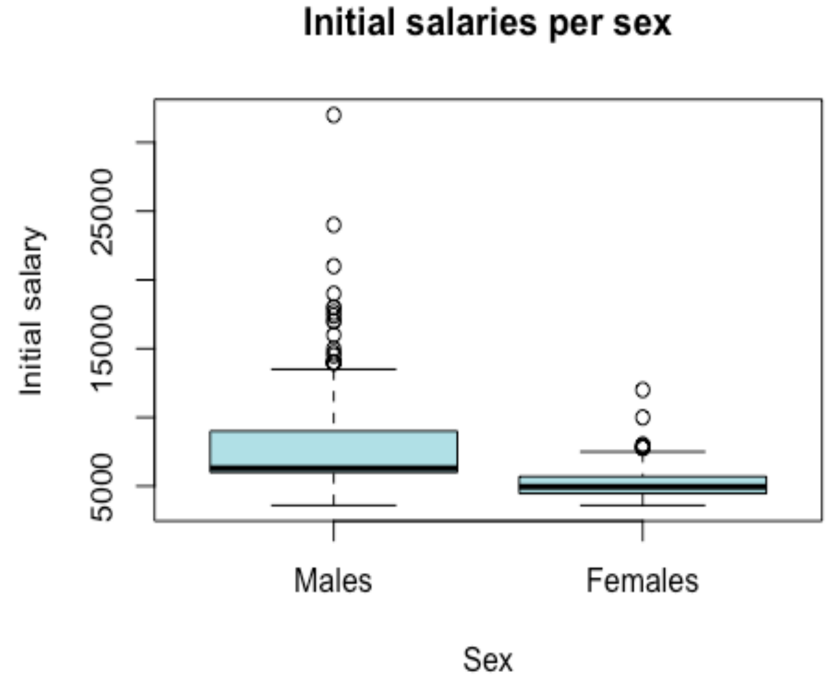
```
m-out-of-n bootstrap symmetry test by Miao, Gel, and Gastwirth (2006)

data:  allSalaries$salaries
Test statistic = 10.18, p-value < 2.2e-16
alternative hypothesis: the distribution is asymmetric.
sample estimates:
bootstrap optimal m 130
```

As can be seen in the results of the symmetry test we conclude that the distribution of the initial salary variable is asymmetric, and this is why we will proceed to a non-parametric hypothesis testing

(Wilcoxon sign rank test) for the median of this variable with the null hypothesis to be **H_0: μ1 = μ2 and** against the alternative hypothesis **H_1: μ1 ≠ μ2**.

```
                 Wilcoxon signed rank test with continuity correction

                             data:  allSalaries$salaries
                            V = 112580, p-value < 2.2e-16
               alternative hypothesis: true location is not equal to 0
```

Since p-value < 2.2e-16 < 0.05, the null hypothesis is rejected in favor of the alternative hypothesis. This means that there is a difference on the beginning salary between the male and female employees which is also depicted in the boxplot below.



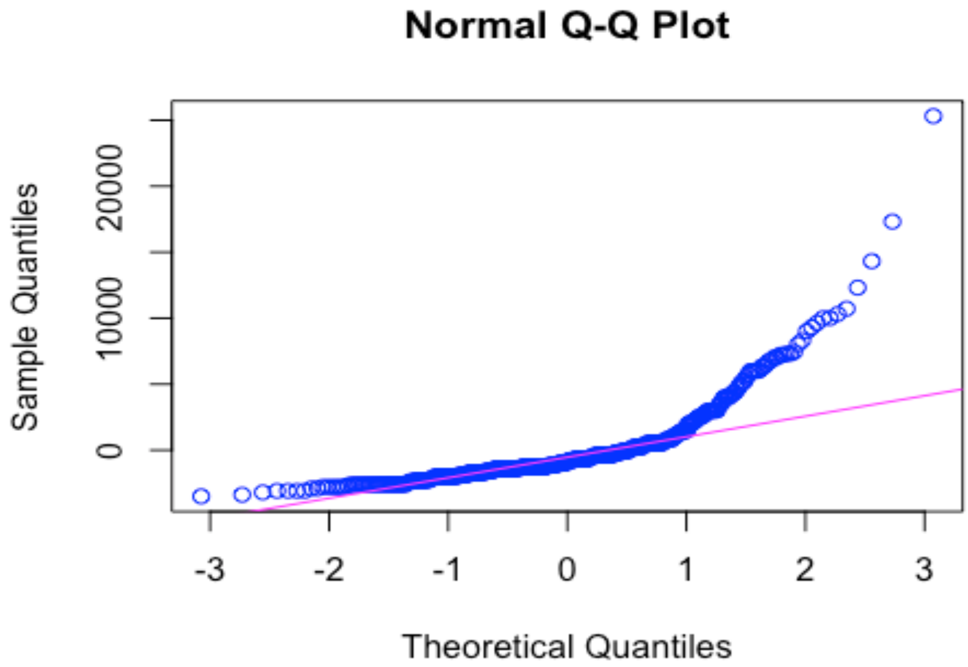
Question 6: Cut the AGE variable into three categories so that the observations are evenly distributed across categories (Hint: you may find the cut2 function in Hmisc package to be very useful). Assign the cut version of AGE into a new variable called age_cut. Investigate if, on average, the beginning salary (salbeg) is the same for all age groups. If there are significant differences, identify the groups that differ by making pairwise comparisons. Interpret your findings and justify the choice of the test that you used by paying particular attention on the assumptions.

In this question we will examine the hypothesis testing ANOVA between the quantitative variable ***salbeg*** and the 3 groups of the categorical variable ***age*** which include the same number of observations (474 / 3 = 158). The results of the ANOVA analysis are presented below:

```
             Df    Sum Sq   Mean Sq F value   Pr(>F)    
age_cut       2 3.965e+08 198235718   21.76 9.18e-10 ***
Residuals   471 4.292e+09   9111833                     
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We will check if the assumption of the ANOVA testing regarding the normality of the residuals is valid by:

A) Depicting the relevant Q-Q plot diagram which implies that the ANOVA residuals are not normally distributed.



B) Performing the Shapiro-Wilk and the Lilliefors (Kolmogorov-Smirnov) normality tests for the residuals since the given samples of the 3 groups include 158 observations each (i.e. greater than 50) which verifies the above statement about the non-normal distribution of the residuals since p-value < 2.2e-16 < 0.05.

```
                    Lilliefors (Kolmogorov-Smirnov) normality test

                          data:  anova1$residuals
                        D = 0.21891, p-value < 2.2e-16

                          Shapiro-Wilk normality test

                          data:  anova1$residuals
                        W = 0.71244, p-value < 2.2e-16
```

Since the total sample n = 474 > 50 and the sample of each age group n1, n2, n3 = 158 > 50 we will proceed to the examination of the mean as a sufficient descriptive measure for the central location of all the age groups.

```
m-out-of-n bootstrap symmetry test by Miao, Gel, and Gastwirth (2006)

data:  anova1$residuals
Test statistic = 11.477, p-value < 2.2e-16
alternative hypothesis: the distribution is asymmetric.
sample estimates:
bootstrap optimal m 51
```

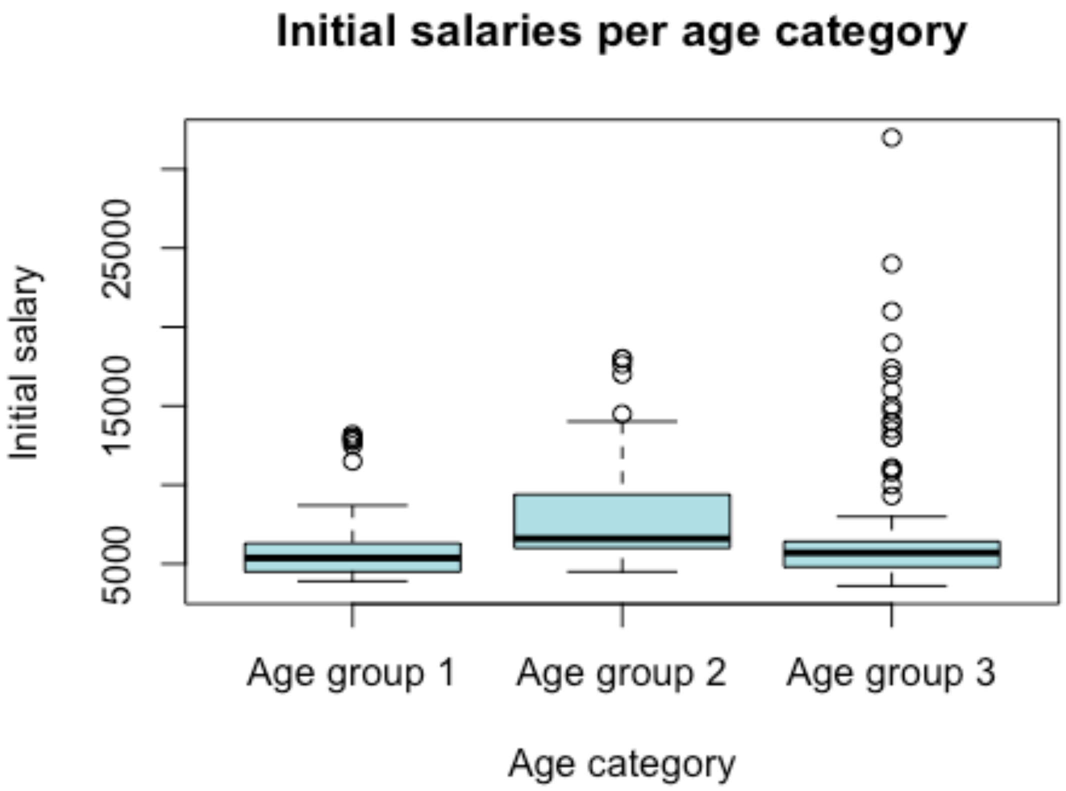
As can be seen in the results of the symmetry test we conclude that the distribution of the residuals is asymmetric, and this is why we will proceed to a non-parametric hypothesis testing (Kruskal-Wallis rank test) for the equality of the medians with the null hypothesis to be **H_o: $\mu_1 = \mu_2 = \mu_3$** against the alternative hypothesis **H_1: $\mu_1 \neq \mu_2 \neq \mu_3$.**

```
                           Kruskal-Wallis rank sum test

                         data:  intialSalary by age_cut
              Kruskal-Wallis chi-squared = 92.742, df = 2, p-value < 2.2e-16
```

Since p-value < 2.2e-16 < 0.05, the null hypothesis is rejected in favor of the alternative hypothesis. This means that some medians between the age groups differ. In order to identify which age groups present these significant differences as far as the initial salary is concerned, we will proceed to the multiple comparisons test.

```
Pairwise comparisons using Wilcoxon rank sum test

data:  allData$intialSalary and allData$age_cut

            Age group 1 Age group 2
Age group 2 < 2e-16     -
Age group 3 0.089       8.9e-12

P value adjustment method: holm
```

Based on the above results we conclude that ***age group 1 - age group 2*** and ***age group 2 – age group 3*** have differences in the initial salary while the group pair 1 – 3 do not have any significant differences. The below boxplot depicts graphically the initial salary per age category as well.



Question 7: By making use of the factor variable minority, investigate if the proportion of white male employees is equal to the proportion of white female employees.

As requested, we will test the equality of the proportion of the independent groups that include white males and white females which means that the null hypothesis is $\mathbf{H_o}$: $\boldsymbol{\pi_{whitemales}= \pi_{whitefemales}}$

against the alternative hypothesis **H_1: $\pi_{whitemales} \neq \pi_{whitefemales}$.** We will firstly create the following tables:

```
The total table proportion:
            MALES     FEMALES
WHITE    0.40928270 0.37130802
NONWHITE 0.13502110 0.08438819

The row proportions table:
            MALES   FEMALES
WHITE    0.5243243 0.4756757
NONWHITE 0.6153846 0.3846154

The column proportions table:
            MALES   FEMALES
WHITE    0.7519380 0.8148148
NONWHITE 0.2480620 0.1851852
```

```
Pearson's Chi-squared test with simulated p-value (based on 2000 replicates)

data:  tab1
X-squared = 2.7139, df = NA, p-value = 0.1119

Fisher's Exact Test for Count Data

data:  tab1
p-value = 0.1186
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
0.429148 1.098149
sample estimates:
odds ratio
0.6894628
```

Since the p-values retrieved from the chisq.test and the Fisher's exact tests performed above are greater than 0.05, we fail to reject the null hypothesis. Thus, we can consider that the proportion of white male employees is equal to the proportion of white female employees.

2. R CODE REFERENCE

# Question 1

```
# Activate the 'foreign' library
install.packages("foreign")
library(foreign)

# Read the SPSS data
mySPSSData <- read.spss("salary.sav", to.data.frame = T)
str(mySPSSData)
```

# Question 2

```
summary(mySPSSData$salbeg)
summary(mySPSSData$time)
summary(mySPSSData$age)
summary(mySPSSData$salnow)
summary(mySPSSData$edlevel)
summary(mySPSSData$work)



# salbeg (initial salary)
  hist(mySPSSData$salbeg,
  main='Initial Salary Of Employees',
  xlab='Initial salary',
  ylab='Frequency',
  border='pink',
  col='purple')
  qqnorm(mySPSSData$salbeg, col=4); qqline(mySPSSData$salbeg, col=6);
  boxplot(mySPSSData$salbeg, horizontal=FALSE, main = 'Initial Salary Employees', col=c('powderblue'))

# time (working experience)
```

```
   hist(mySPSSData$time,
    main='Working experience in years',
    xlab='Years',
    ylab='Frequency',
    border='pink',
    col='purple')

qqnorm(mySPSSData$time, col=4); qqline(mySPSSData$time, col=6);
boxplot(mySPSSData$time, horizontal=FALSE, main = 'Working experience in years',
col=c('powderblue'))

# age
   hist(mySPSSData$age,
    main='Age of employees',
    xlab='Age',
    ylab='Frequency',
    border='pink',
    col='purple')

qqnorm(mySPSSData$age, col=4); qqline(mySPSSData$age, col=6);
boxplot(mySPSSData$age, horizontal=FALSE, main = 'Age of employees',
col=c('powderblue'))



# salnow (current salary)
    hist(mySPSSData$salnow,
    main='Current salary of employees',
    xlab='Current salary',
    ylab='Frequency',
    border='pink',
    col='purple')

qqnorm(mySPSSData$salnow, col=4); qqline(mySPSSData$salnow, col=6);
boxplot(mySPSSData$salnow, horizontal=FALSE, main = 'Current salary of employees',
col=c('powderblue'))
```

```
# education level
   hist(mySPSSData$edlevel,
    main='Education level',
    xlab='Level',
    ylab='Frequency',
    border='pink',
    col='purple')

qqnorm(mySPSSData$edlevel, col=4); qqline(mySPSSData$edlevel, col=6);
boxplot(mySPSSData$edlevel, horizontal=FALSE, main = 'Education', col=c('powderblue'))



# work (employment working hours)
    hist(mySPSSData$work,
    main='Employement working hours',
    xlab='Working hours',
    ylab='Frequency',
    border='pink',
    col='purple')

qqnorm(mySPSSData$work, col=4); qqline(mySPSSData$work, col=6);



boxplot(mySPSSData$work, horizontal=FALSE, main = 'Employement working hours',
col=c('powderblue'))
```

# Question 3

```
install.packages("nortest")
library(nortest)
initialSalary <- mySPSSData$salbeg
nrow(filter(mySPSSData, is.na(mySPSSData$salbeg)))

# Testing for normality of one quantitative variable (n>50)
qqnorm(mySPSSData$salbeg, col=4); qqline(mySPSSData$salbeg, col=6);
```

```
shapiro.test(initialSalary) # Normality rejected
lillie.test(initialSalary) # Normality rejected

# Testing the sufficiency of the mean for central location
library(lawstat)
symmetry.test(initialSalary, option = c("MGG", "CM", "M"), side = c("both", "left", "right"),
 boot = TRUE, B = 1000, q = 8/9) # Symmetry rejected

# Non-parametric hypothesis testing
wilcox.test(initialSalary, mu=1000) # Ho hypothesis rejected
```

# Question 4

```
library(nortest)
diff <- mySPSSData$salnow - mySPSSData$salbeg
nrow(filter(mySPSSData, is.na(mySPSSData$salbeg) & is.na(mySPSSData$salnow)))

# Testing for normality of two dependent variables (n>50)
qqnorm(diff, col=4); qqline(diff, col=6);
mean(diff); median(diff)
shapiro.test(diff)
lillie.test(diff)



# Testing the sufficiency of the mean for central location for the difference
symmetry.test(diff, option = c("MGG", "CM", "M"), side = c("both", "left", "right"),
boot = TRUE, B = 1000, q = 8/9) # Symmetry rejected

# Non-parametric hypothesis testing
wilcox.test(diff) # Ho hypothesis rejected
boxplot(diff, horizontal=FALSE, main = 'Salaries differencies', col=c('powderblue'))
```

# Question 5

```
   library(nortest)
   nrow(filter(mySPSSData, is.null(mySPSSData$salbeg)))

   # Testing for normality of two independent samples
   salMales <- subset(mySPSSData$salbeg, mySPSSData$sex=='MALES')
   salFemales <- subset(mySPSSData$salbeg, mySPSSData$sex=='FEMALES')
   nM<-length(salMales)
   nF<-length(salFemales)

   allSalaries <- data.frame(salaries=c(salMales, salFemales), sex=factor(rep(1:2, c(nM,nF)),
    labels=c('Males','Females')))

# Testing for normality of each group
   qqnorm(allSalaries$salaries, col=4); qqline(diff, col=6);
   by(allSalaries$salaries, allSalaries$sex, lillie.test) # Normality rejected
   by(allSalaries$salaries, allSalaries$sex, shapiro.test) # Normality rejected

# Testing the sufficiency of the mean for central location of both groups (nM, nF > 50)
   install.packages("lawstat")
   library(lawstat)
   symmetry.test(allSalaries$salaries, option = c("MGG", "CM", "M"), side = c("both", "left",
    "right"), boot = TRUE, B = 1000, q = 8/9) # Sufficiency rejected

 # Non-parametric hypothesis testing

   wilcox.test(allSalaries$salaries) # Ho hypothesis rejected
   boxplot(data=allSalaries, salaries~sex ,horizontal=FALSE, xlab='Sex',
   ylab='Initial salary', main = 'Initial salaries per sex', col=c('powderblue'))
```

# Question 6

```
# Hypothesis testing for multiple samples
install.packages("Hmisc")
library(Hmisc)
age_cut <- cut2(mySPSSData$age, m=158, g=3, levels.mean = FALSE, oneval = TRUE,
 onlycuts = FALSE)
intialSalary <- mySPSSData$salbeg
age_cut <- factor(age_cut, labels = paste('Age group '))
allData <- data.frame(intialSalary=intialSalary, age_cut=age_cut)

# Conducting ANOVA
anova1 <- aov(intialSalary~age_cut, data=allData)
anova2 <- oneway.test(intialSalary~age_cut, data=allData)
summary(anova1)

# Normality of the residuals
library(nortest)
lillie.test(anova1$residuals) # Normality rejected
shapiro.test(anova1$residuals) # Normality rejected
qqnorm(anova1$residuals, col=4); qqline(anova1$residuals, col=6)
# Testing the sufficiency of the mean for central location of all age groups (n> 50, n of each
 group = 158)
library(lawstat)
symmetry.test(anova1$residuals, option = c("MGG", "CM", "M"), side = c("both", "left",
 "right"), boot = TRUE, B = 1000, q = 8/9) # Sufficiency rejected

# Non-parametric hypothesis testing
kruskal.test(intialSalary~age_cut, data=allData) # Ho hypothesis rejected



# Multiple comparisons testing using unequal variances
pairwise.wilcox.test(allData$intialSalary, allData$age_cut)
```

```
boxplot(data=allData, intialSalary~age_cut ,horizontal=FALSE, xlab='Age category',
ylab='Initial salary', main = 'Initial salaries per age category', col=c('powderblue'))
```

# Question 7

```
# Testing for the association between two categorical variables
tab1 <- table(mySPSSData$minority, mySPSSData$sex)
prop.table(tab1) # Total Table Proportions
prop.table(tab1, 1)  # Row Proportions
prop.table(tab1, 2) # Column Proportions

# Implementing Chi squared test and Fisher exact test
chisq.test(tab1, correct=FALSE, simulate.p.value = TRUE) # Ho hypothesis approved
fisher.test(tab1) # Ho hypothesis approved
```