Athens University of Economics and Business

School of Business

Department of Management Science & Technology

Master of Science in Business Analytics

Program:	Full-time
Quarter:	1st (Fall Quarter)
Course:	Statistics for Business Analytics I
Assignment №:	1
Students (Registration №):	Souflas Eleftherios-Efthymios (f2822217)

Statistics for Business Analytics I
Lab Assignment #1

Student: Souflas Eleftherios-Efthymios

Question 1

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

Output

```
file <- "path\\to\\salary.sav"
# install.packages("foreign")
library(foreign)
salary <- read.spss(file, to.data.frame=TRUE)
str(salary)
> str(salary)
'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 [1:11] "EMPLOYEE CODE" "BEGINNING SALARY" "SEX
OF EMPLOYEE" "JOB SENIORITY" ...
  ..- attr(*, "names")= chr [1:11] "id" "salbeg" "sex" "time" ...
 - attr(*, "codepage")= int 1253
```

Comment

Salary is a data frame containing 474 observations of the following 11 variables:

- id, a numeric vector, labeled as employee code
- salbeg, a numeric vector, labeled as beginning salary
- sex, a factor, labeled as sex of employee, with two levels: 0 as males and 1 as females
- time, a numeric vector, labeled as job seniority
- age, a numeric vector, labeled as age of employee
- salnow, a numeric vector, labeled as current salary
- edlevel, a numeric vector, labeled as educational level (not clear though what value depicts which educational level)
- work, a numeric vector, labeled as work experience
- jobcat, a factor, labeled as employment category, with the following seven levels:

VALUE	LEVEL
1	CLERICAL
2	OFFICE TRAINEE
3	SECURITY OFFICER
4	COLLEGE TRAINEE
5	EXEMPT EMPLOYEE
6	MBA TRAINEE
7	TECHNICAL

- minority, a factor, labeled as minority classification, with two levels: 0 as white and 1 as nonwhite

- sexrace, a factor, labeled as sex & race classification, with the following four levels:

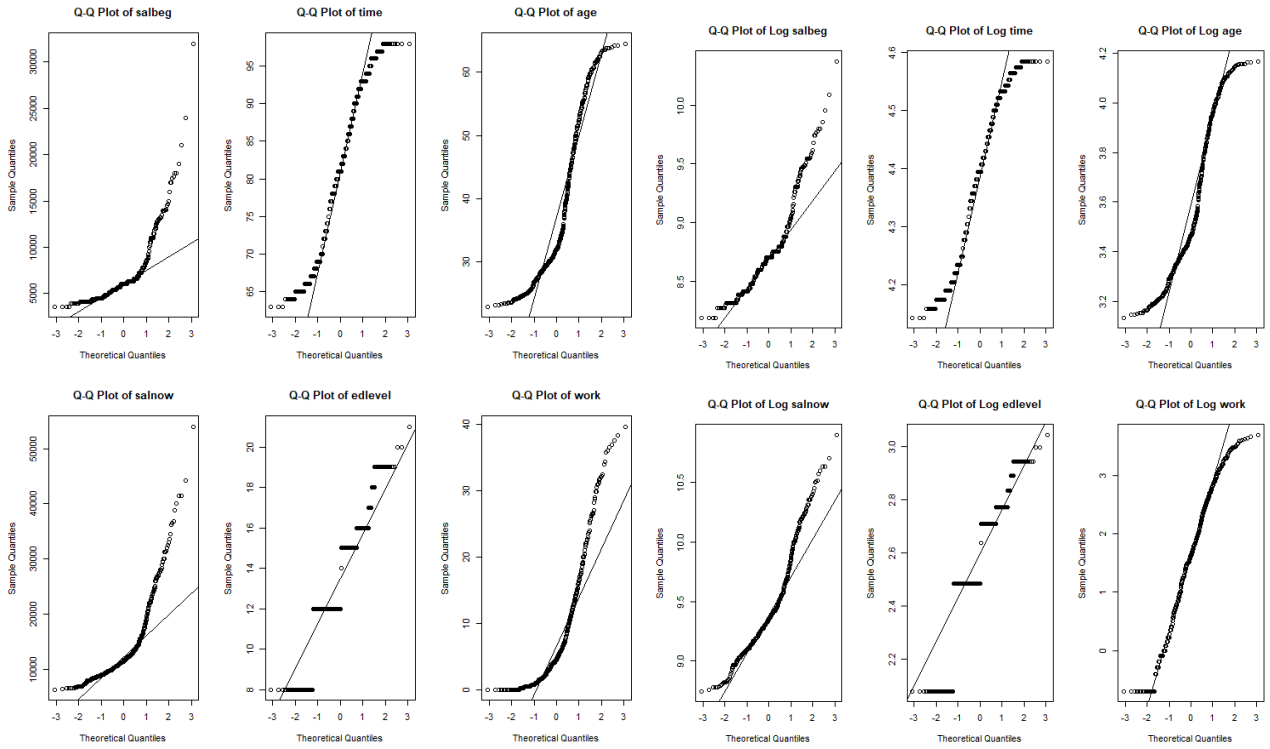
VALUE	LEVEL
1	WHITE MALES
2	MINORITY MALES
3	WHITE FEMALES
4	MINORITY FEMALES

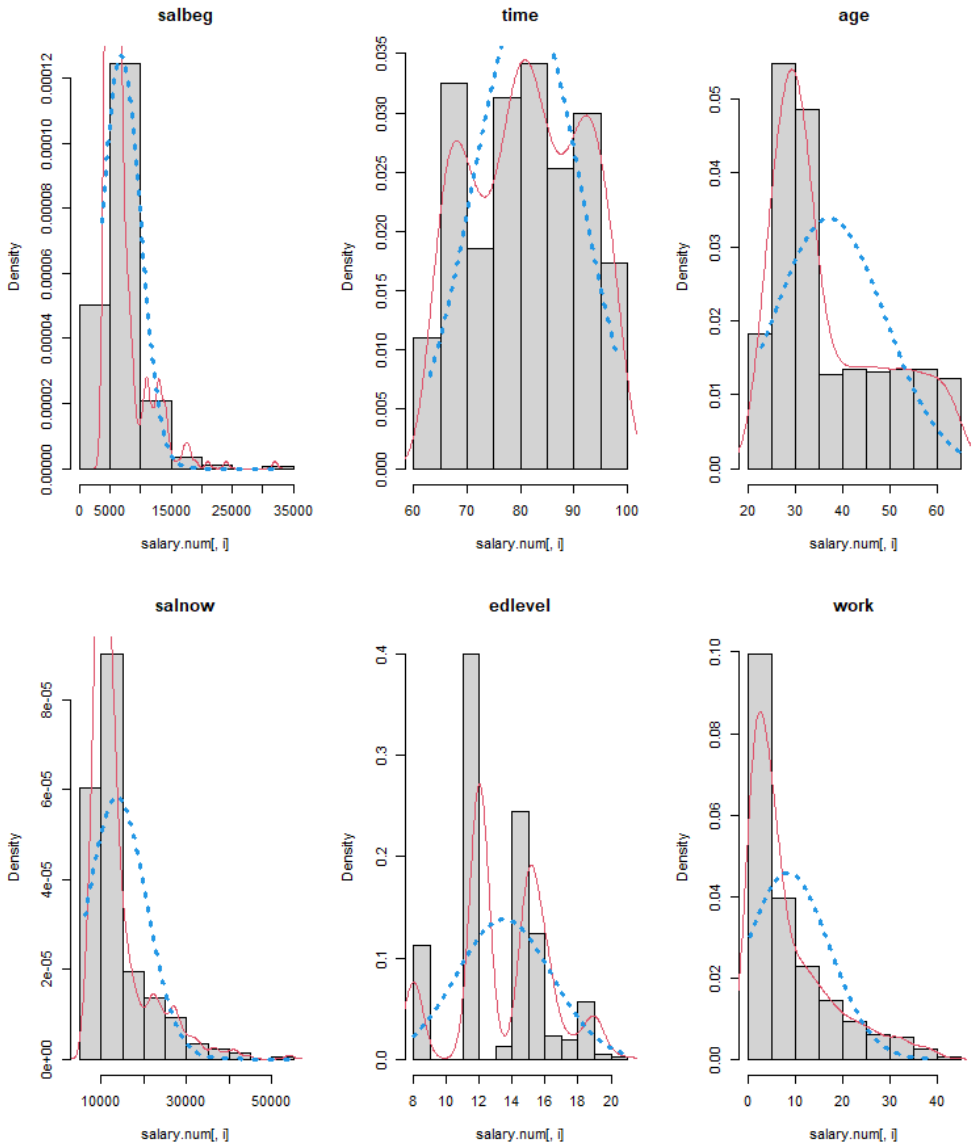
Question 2

Get that summary statistics of the numerical variables in the dataset and visualize their distribution (e.g. use histograms etc). Which variables appear to be normally distributed? Why?

Output

```
> summary
     salbeg          time            age           salnow         edlevel
 Min.   : 3600   Min.   :63.00   Min.   :23.00   Min.   : 6300   Min.   : 8.00
 1st Qu.: 4995   1st Qu.:72.00   1st Qu.:28.50   1st Qu.: 9600   1st Qu.:12.00
 Median : 6000   Median :81.00   Median :32.00   Median :11550   Median :12.00
 Mean   : 6806   Mean   :81.11   Mean   :37.19   Mean   :13768   Mean   :13.49
 3rd Qu.: 6996   3rd Qu.:90.00   3rd Qu.:45.98   3rd Qu.:14775   3rd Qu.:15.00
 Max.   :31992   Max.   :98.00   Max.   :64.50   Max.   :54000   Max.   :21.00
      work
 Min.   : 0.000
 1st Qu.: 1.603
 Median : 4.580
 Mean   : 7.989
 3rd Qu.:11.560
 Max.   :39.670
```





Comment

Id, although numeric, is not statistically significant to be visualized as every employee has a unique id in ascending order and thus it is excluded.

Normal (or Gaussian) distribution is a type of continuous probability distribution for a real-valued random variable. From examining the Q-Q Plot and the Log Q-Q Plot, we can securely mention that the edlevel (education level) is not a continuous, but a discrete variable. It takes values from 8 to 21 to depict the educational level of the employee. If we said that, value 8 is equal to “primary school education” and 21 to “postdoctoral education” and the middle values to the intermediate stages of education, then the variable would definitely be a factor, just like sex or minority.

1. A variable that is normally distributed has a histogram and density plot that are bell-shaped, with only one peak, and is symmetric around the mean.
2. Q-Q plot, shows the distribution of the data against the expected normal distribution. For normally distributed data, observations should lie approximately on a straight line. If all the points fall approximately along this reference line, we can assume normality.

From summary statistics, we get that only the time variable’s mean is almost equal to the median. But that alone is not enough to assume normality.

Q-Q Plots

Observations should lie approximately on the straight Q-Q line. As all the points do not fall approximately along this reference line, we cannot assume normality, by only viewing Q-Q Plots for any variable.

Log Q-Q Plots

As natural logarithm of 0 is undefined and the value of 0 is present in the observations of variable work, we add to the variable work a small positive value (0.5). As before, even after log transformation, a large number of points do not fall on the Q-Q line, even for variables time and work.

Histograms and Density Plots (red) against Normal Distribution Density Plot (blue)

As before, the two density plots (Observed Values vs Expected if Normality was assumed) are not close. Even variable time that seems symmetric, is not bell-shaped and has many peaks.

To conclude, we cannot assume normality for any variable.

Question 3

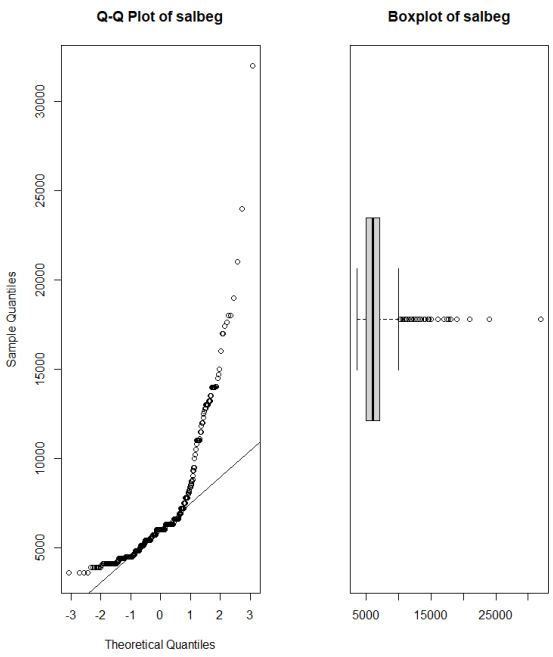
Use the appropriate test to examine whether the beginning salary of a typical employee can be considered to be equal to 1000 dollars. How do you interpret the results? What is the justification for using this particular test instead of some other? Explain.

Output

```
# install.packages("nortest")
# install.packages("lawstat")
library(nortest)
library(lawstat)
single_cont <- function(data, column, condition, sl = 0.05){
  var1 <- data[,column]
  par(mfrow=c(1,2))
  # Can we assume normality?
  test1 <- 1
  if (length(var1) > 50){
    test1 <- lillie.test(var1)$p.value
  }
  test2 <- shapiro.test(var1)$p.value
  qqnorm(var1, main = paste("Q-Q Plot of" , names(data)[column]))
  qqline(var1)
  if ((test1 < sl) | (test2 < sl)){
    # Is the sample large?
    if (length(var1) > 50){
      # Is the mean a sufficient descriptive measure for central location?
      test3 <- symmetry.test(var1)$p.value
      boxplot(var1, main = paste("Boxplot of" , names(data)[column]), horizontal=TRUE)
    }
    if ((length(var1) <= 50) || (test3 < sl)){
      final <- wilcox.test(var1, mu = condition)$p.value
    }
  }
  if (((test1 >= sl) && (test2 >= sl)) || (test3 >= sl)){
    final <- t.test(var1, mu = condition)$p.value
  }
  if (final < sl){
    message <- paste("We reject the null hypothesis. P-value =", round(final,2))
  } else {
    message <- paste("We cannot reject the null hypothesis. P-value =", round(final,2))
  }
  return(message)
  par(mfrow=c(1,1))
}
single_cont(salary, 2, 1000)
```

```
> single_cont(salary, 2, 1000)

[1] "We reject the null hypothesis. P-value = 0"
```



Comment

Hypothesis test for a single continuous variable.

Variable: salbeg

Null Hypothesis: $H_0: \mu = 1000$

Alternative: $H_1: \mu \neq 1000$

I believe constructing a function to handle each case explains all choices of tests.

After we call the function, with blue color is highlighted the path that it will follow for this specific case. Firstly, we check if the variable is normally distributed. But, as we have seen from the previous question, it is not. After the tests produce a p-value equal to below 0.05 and the Q-Q Plot does not provide us with enough evidence to assume normality, we reject the assumption. Then, because the sample's population is above 50, we must check if the distribution is symmetric. In order to prove that, we take advantage of boxplot and symmetry test (R library: lawstat). Examining the boxplot and the p-value of symmetry test (<0.05), we are indicated that the mean is not a sufficient descriptive measure for central location. So, we make use of the Wilcoxon test for one sample, in order to check the assumption for the median value. The message, that the test produces, is that the p-value is equal to a very small value below 0.05.

Thus, we reject the null hypothesis and we state that the beginning salary of a typical employee cannot be considered to be equal to 1000 dollars. From examining the plots above, we can securely mention that the beginning salary can be considered above 1000 dollars (in fact close to 5000-6000 dollars).

Question 4

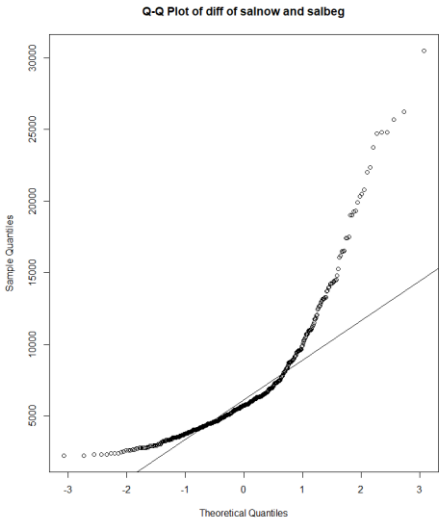
Consider the difference between the beginning salary (salbeg) and the current salary (salnow). Test if the there is any significant difference between the beginning salary and current salary. (Hint: Construct a new variable for the difference (salnow – salbeg) and test if, on average, it is equal to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

Output

```
x <- salary$salnow - salary$salbeg
length(x) # 474
lillie.test(x)
> lillie.test(x)
        Lilliefors (Kolmogorov-Smirnov) normality test
data:  x
D = 0.186, p-value < 2.2e-16
```

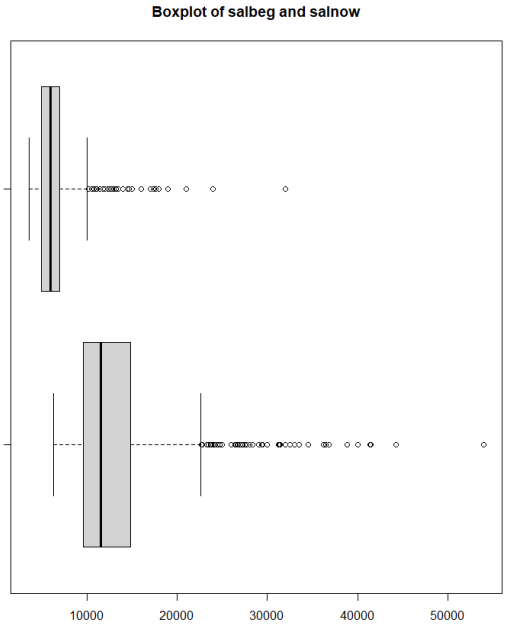
```
shapiro.test(x)
> shapiro.test(x)
        Shapiro-Wilk normality test
data:  x
W = 0.78168, p-value < 2.2e-16
```

```
qqnorm(x, main = paste("Q-Q Plot of diff of" , names(salary)[6], "and",
names(salary)[2]))
qqline(x)
```

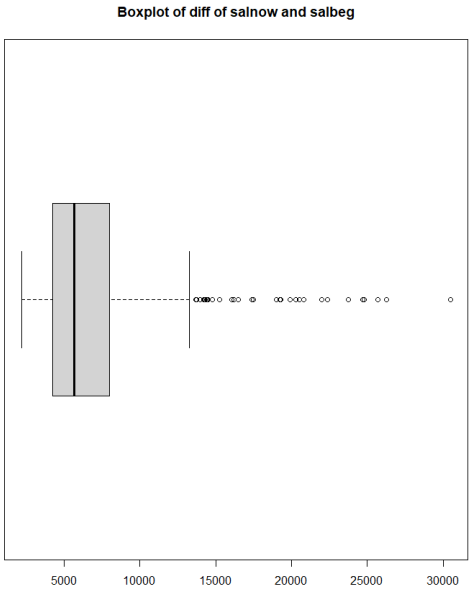



```
symmetry.test(x)
> symmetry.test(x)
        m-out-of-n bootstrap symmetry test by Miao, Gel, and Gastwirth (2006)
data:  x
Test statistic = 10.536, p-value < 2.2e-16
alternative hypothesis: the distribution is asymmetric.
sample estimates:
bootstrap optimal m
                 72
```

```
boxplot(salary$salnow, salary$salbeg, main = paste("Boxplot of" , names(salary)[2],
"and", names(salary)[6]), horizontal=TRUE)
```



```
wilcox.test(x, mu = 0)
> wilcox.test(x, mu = 0)
        Wilcoxon signed rank test with continuity correction
data:  x
V = 112575, p-value < 2.2e-16
alternative hypothesis: true location is not equal to 0
mean(salary$salnow)/mean(salary$salbeg)
median(salary$salnow)/median(salary$salbeg)
> mean(salary$salnow)/mean(salary$salbeg)
[1] 2.022766
> median(salary$salnow)/median(salary$salbeg)
[1] 1.925
boxplot(x, main = paste("Boxplot of diff of" ,  names(salary)[6], "and",
names(salary)[2]), horizontal=TRUE)
```



Comment

Hypothesis test for two dependent samples (difference between the two dependent values-measurements). We eliminate correlation by using the difference of each pair:

$$\Delta_i = x_{1i} - x_{2i}$$

We test if the mean of the difference is zero or not

Variable: diff of salnow and salbeg

Null Hypothesis: $H_0: \mu = 0$

Alternative: $H_1: \mu \neq 0$

The sample is large (above 50), so we implement Kolmogorov-Smirnov and Shapiro-Wilk Tests. Both tests give us a p-value a lot less than 0.05 and together with examining the Q-Q plot, we cannot assume normality. So, because the sample is large, we must test if the mean is a sufficient descriptive measure of central location for the difference. We cannot assume symmetry, as again, together with examining the boxplots of the original variables (many outliers on the upper limit – positive skewed), p-value of symmetry test is a lot less than 0.05. Thus, we test for zero median difference with the use of Wilcoxon Test for dependent samples. The resulting p-value is a lot less than 0.05.

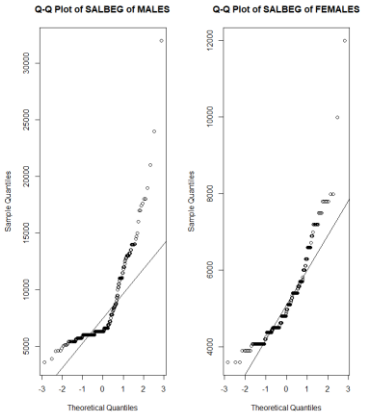
So, we reject the null hypothesis. There is significant difference between the beginning and the current salary and in fact, if we examine the boxplot of the difference, the current salary is greater than the beginning salary by more than 5000 dollars.

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.

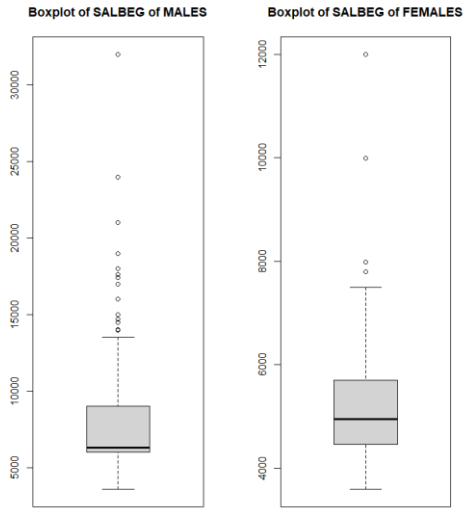
Output

```
dataset1 <- salary[,2:3]
n1 <- nrow(subset(dataset1, dataset1$sex=="MALES"))
n2 <- nrow(subset(dataset1, dataset1$sex=="FEMALES"))
n1 > 50
n2 > 50
by(dataset1$salbeg, dataset1$sex, lillie.test)
> by(dataset1$salbeg, dataset1$sex, lillie.test)
dataset1$sex: MALES
       Lilliefors (Kolmogorov-Smirnov) normality test
data:  dd[x, ]
D = 0.25863, p-value < 2.2e-16
------------------------------------------------------------------------------------------
dataset1$sex: FEMALES
       Lilliefors (Kolmogorov-Smirnov) normality test
data:  dd[x, ]
D = 0.14843, p-value = 1.526e-12
by(dataset1$salbeg, dataset1$sex, shapiro.test)
> by(dataset1$salbeg, dataset1$sex, shapiro.test)
dataset1$sex: MALES
       Shapiro-Wilk normality test
data:  dd[x, ]
W = 0.73058, p-value < 2.2e-16
------------------------------------------------------------------------------------------
dataset1$sex: FEMALES
       Shapiro-Wilk normality test
data:  dd[x, ]
W = 0.85837, p-value = 2.98e-13
males <- subset(dataset1, dataset1$sex=="MALES")
females <- subset(dataset1, dataset1$sex=="FEMALES")
par(mfrow=c(1,2))
qqnorm(males$salbeg, main = paste("Q-Q Plot of" , toupper(names(dataset1)[1]), "of",
levels(dataset1$sex)[1]))
qqline(males$salbeg)
qqnorm(females$salbeg, main = paste("Q-Q Plot of" , toupper(names(dataset1)[1]), "of",
levels(dataset1$sex)[2]))
qqline(females$salbeg)
par(mfrow=c(1,1))
```

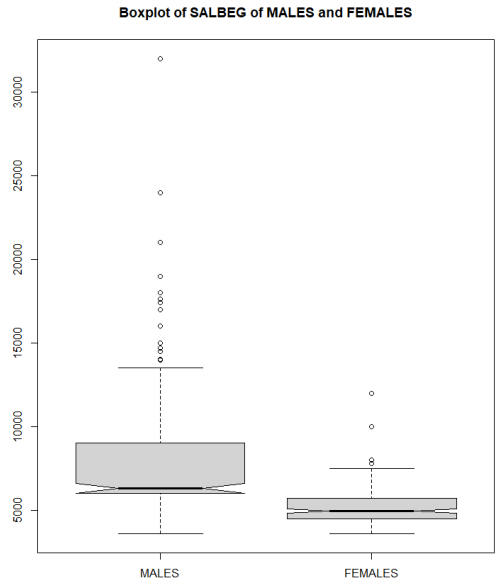



```
by(dataset1$salbeg, dataset1$sex, symmetry.test)
> by(dataset1$salbeg, dataset1$sex, symmetry.test)
       dataset1$sex: MALES
              m-out-of-n bootstrap symmetry test by Miao, Gel, and Gastwirth (2006)
data:  dd[x, ]
Test statistic = 13.829, p-value < 2.2e-16
alternative hypothesis: the distribution is asymmetric.
sample estimates:
bootstrap optimal m
                 44
------------------------------------------------------------------------------------------
```

```
dataset1$sex: FEMALES
       m-out-of-n bootstrap symmetry test by Miao, Gel, and Gastwirth (2006)
data:  dd[x, ]
Test statistic = 5.2527, p-value < 2.2e-16
alternative hypothesis: the distribution is asymmetric.
sample estimates:
bootstrap optimal m
                 75
par(mfrow=c(1,2))
boxplot(males$salbeg, main = paste("Boxplot of" , toupper(names(dataset1)[1]), "of",
levels(dataset1$sex)[1]))
boxplot(females$salbeg, main = paste("Boxplot of" , toupper(names(dataset1)[1]), "of",
levels(dataset1$sex)[2]))
par(mfrow=c(1,1))
```




```
# Wilcoxon rank-sum test (or Mann-Whitney)
wilcox.test(males$salbeg, females$salbeg)
> wilcox.test(males$salbeg, females$salbeg)
       Wilcoxon rank sum test with continuity correction
data:  males$salbeg and females$salbeg
W = 47874, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
boxplot(males$salbeg, females$salbeg, names=levels(dataset1$sex)[1:2],
        main = paste("Boxplot of" , toupper(names(dataset1)[1]), "of",
levels(dataset1$sex)[1], "and", levels(dataset1$sex)[2]),
        notch = TRUE)
```



Comment

Hypothesis test for two samples - Testing for the association between a continuous and a categorical variable of two levels.

Variables: continuous: salbeg,
categorical: sex

Measurements of the variable salbeg in two sex groups (male/women) of different research units.

Null Hypothesis: $H_0: \mu_1 = \mu_2$

Alternative: $H_1: \mu_1 \neq \mu_2$

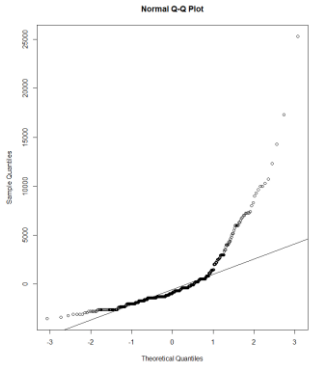
As each sample's population is greater than 50 (it is also sufficient to only one be above 50), we test normality with Q-Q Plot, Kolmogorov-Smirnov and Shapiro-Wilk Tests in each sample. As all p-values are less than 0.05, together with examining the Q-Q plot, we cannot assume normality in either of the samples and because we have large overall sample population, we test if the mean is a sufficient descriptive measure of central location for either of the two samples. We cannot assume symmetry, as together with examining the boxplots for each level (males/females), where we observe that many outliers lie on the upper limit – positive skewed, p-value of symmetry test is a lot less than 0.05 for each level. Thus, we test for zero difference between the medians with the use of Wilcoxon Rank-Sum Test for independent samples. The resulting p-value is a lot less than 0.05.
So, we reject the null hypothesis. There is significant difference between the beginning salary of men and women and in fact, by examining the boxplot, the beginning salary of men is greater than the respective of women.

Question 6

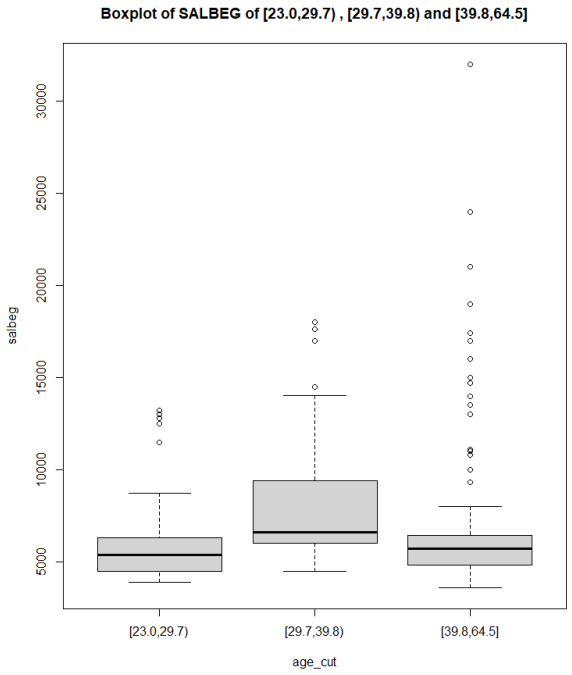
Cut the AGE variable into three categories so that the observations are evenly distributed across categories (Hint: you may find the cut2 function in Hmisc package to be very useful). Assign the cut version of AGE into a new variable called age_cut. Investigate if, on average, the beginning salary (salbeg) 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.

Output

```
# install.packages("Hmisc")
library(Hmisc)
salary$age_cut <- cut2(salary$age, g = 3)
dataset2 <- salary[, c(2,12)]
nrow(dataset2) > 50
anova1 <- aov(salbeg~age_cut, data = dataset2)
lillie.test(anova1$residuals)
> lillie.test(anova1$residuals)
       Lilliefors (Kolmogorov-Smirnov) normality test
data:  anova1$residuals
D = 0.21891, p-value < 2.2e-16
shapiro.test(anova1$residuals)
> shapiro.test(anova1$residuals)
       Shapiro-Wilk normality test
data:  anova1$residuals
W = 0.71244, p-value < 2.2e-16
qqnorm(anova1$residuals)
qqline(anova1$residuals)
```



```
bartlett.test(salbeg~age_cut, data = dataset2)
> bartlett.test(salbeg~age_cut, data = dataset2)
       Bartlett test of homogeneity of variances
data:  salbeg by age_cut
Bartlett's K-squared = 83.024, df = 2, p-value < 2.2e-16
fligner.test(salbeg~age_cut, data = dataset2)
> fligner.test(salbeg~age_cut, data = dataset2)
       Fligner-Killeen test of homogeneity of variances
data:  salbeg by age_cut
Fligner-Killeen:med chi-squared = 6.777, df = 2, p-value = 0.03376
library(car)
leveneTest(salbeg~age_cut, data = dataset2)
> leveneTest(salbeg~age_cut, data = dataset2)
Levene's Test for Homogeneity of Variance (center = median)
       Df F value   Pr(>F)
group   2  5.5026 0.004342 **
      471
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
boxplot(salbeg~age_cut, data = dataset2, main =
          paste("Boxplot of" , toupper(names(dataset2)[1]), "of",
levels(dataset2$age_cut)[1], ",",
                levels(dataset2$age_cut)[2], "and",levels(dataset2$age_cut)[3]))
```




```
# Kruskal-Wallis Test (Equality of medians)
kruskal.test(salbeg~age_cut, data = dataset2)
> kruskal.test(salbeg~age_cut, data = dataset2)
       Kruskal-Wallis rank sum test
data:  salbeg by age_cut
Kruskal-Wallis chi-squared = 92.742, df = 2, p-value < 2.2e-16
# We reject the null hypothesis
pairwise.wilcox.test(dataset2$salbeg, dataset2$age_cut)
> pairwise.wilcox.test(dataset2$salbeg, dataset2$age_cut)
       Pairwise comparisons using Wilcoxon rank sum test with continuity correction
data:  dataset2$salbeg and dataset2$age_cut
            [23.0,29.7) [29.7,39.8)
[29.7,39.8) < 2e-16     -
[39.8,64.5] 0.089       8.9e-12
P value adjustment method: holm
boxplot(salbeg~age_cut, data = dataset2, main =
          paste("Boxplot of" , toupper(names(dataset2)[1]), "of",
levels(dataset2$age_cut)[1], ",",
                levels(dataset2$age_cut)[2], "and",levels(dataset2$age_cut)[3]))
```

Comment

Hypothesis test for three samples - Testing for the association between a continuous and a categorical variable of three levels.

Variables: continuous: salbeg,
categorical: age_cut

Measurement of the variable salbeg in 3 age categories of different employees.

Null Hypothesis: $H_0: \mu_1 = \mu_2 = \mu_3$

Alternative: $H_1: \mu_k \neq \mu_j, for\ some\ k \neq j\ \in \{1,2,3\}$

ANOVA: Analysis of Variance

Assumptions:

- Residuals' normality or the sample size to be large (n>50)
- Equal variances

As total samples size (n1 + n2 + n3) is above 50, the first assumption is satisfied, but if we test for residuals' normality with Q-Q Plot and Kolmogorov-Smirnov and Shapiro-Wilk Tests we can observe that the residuals' normality assumption is rejected (p-value < 0.05). The same happens with the homoscedasticity assumption after the implementation of Bartlett's, Fligner-Killeen and Levene's tests (p-value < 0.05). So, we cannot implement the ANOVA F-Test and the total samples size is large. Therefore, we must check if the mean is a sufficient descriptive measure of central location for all groups. The boxplot leads us to reject symmetry assumption, as there are a lot of outliers on the upper bound – positive skewed. Consequently, we must check for the equality of medians with the Kruskal-Wallis's test and after we implemented the test, the p-value was below 0.05.

Thus, we reject the null hypothesis. There is significant difference in the beginning salary of at least one age group and in fact, after the implementation of the Pairwise Wilcox Test and the examination of boxplot for each level, the beginning salary of the 2nd age group ([29.7, 39.8)) is greater than the respective of the other two. With yellow color, in the result of the implementation of the Pairwise Wilcox Test, is highlighted the pairwise comparison of the first and the third age group, from the result of which we cannot reject the hypothesis of equality of beginning salary of these groups. With light blue color, is highlighted the pairwise comparison of the second age group with the other two groups, from the result of whom we reject the hypothesis of equality of beginning salary of these groups (the second age group's beginning salary differs from the other two).

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.

Output

```
tab1 <- table(salary$sex, salary$minority)
> tab1
         WHITE NONWHITE
  MALES     194       64
  FEMALES   176       40
prop.table(tab1, 2)
> prop.table(tab1, 2)
             WHITE  NONWHITE
  MALES   0.5243243 0.6153846
  FEMALES 0.4756757 0.3846154
prop.test(tab1)
> prop.test(tab1)
          2-sample test for equality of proportions with continuity correction
data:  tab1
X-squared = 2.3592, df = 1, p-value = 0.1245
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.14102693  0.01527327
sample estimates:
```

```
   prop 1    prop 2
0.7519380 0.8148148
chisq.test(tab1)
> chisq.test(tab1)
        Pearson's Chi-squared test with Yates' continuity correction
data:  tab1
X-squared = 2.3592, df = 1, p-value = 0.1245
fisher.test(tab1)
> fisher.test(tab1)
        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
```

Comment

Hypothesis test for two categorical variables - Testing for the association between two categorical variables (independent samples) - Testing for the equality of proportions/probabilities in independent groups/samples.

Variables: categorical: minority, sex

Null Hypothesis: $H_0: \pi_{white\ males} = \pi_{white\ females}$

Alternative: $H_1: \pi_{white\ males} \neq \pi_{white\ females}$

We use table function to build a contingency table of the counts at each combination of factor levels. We then create the total table proportions. Then, we implement the Pearson's Chi-squared test (with the use of prop.test and chisq.test) and the Fisher's Exact test statistics to check for independence. All tests produce p-values above 0.05 and so we cannot reject the null hypothesis. Thus, we cannot reject the hypothesis that the proportion of white male employees is equal to the proportion of white female employees, or otherwise we state that there is independence between gender and minority of employees.

Code

```
#########################################################################################################################################
#########################################################################################################################################
##                                                                                                                                     ##
##                                                    ANSWER TO QUESTION 1                                                             ##
##                                                                                                                                     ##
#########################################################################################################################################
#########################################################################################################################################

file <- "path\\to\\salary.sav"
# install.packages("foreign")
library(foreign)
salary <- read.spss(file, to.data.frame=TRUE)
str(salary)
# salary is a data frame containing 474 observations of the following 11 variables:
# id, a numeric vector, labeled as employee code
# salbeg, a numeric vector, labeled as beginning salary
# sex, a factor, labeled as sex of employee, with two levels: 0 as males and 1 as females
# time, a numeric vector, labeled as job seniority
# age, a numeric vector, labeled as age of employee
# salnow, a numeric vector, labeled as current salary
# edlevel, a numeric vector, labeled as educational level (not clear though what value depicts which educational level)
# work, a numeric vector, labeled as work experience
# jobcat, a factor, labeled as employment category, with the following seven levels:
#         value            level
#             1         CLERICAL
#             2   OFFICE TRAINEE
#             3 SECURITY OFFICER
#             4  COLLEGE TRAINEE
#             5  EXEMPT EMPLOYEE
#             6      MBA TRAINEE
#             7        TECHNICAL
# minority, a factor, labeled as minority classification, with two levels: 0 as white and 1 as nonwhite
# sexrace, a factor, labeled as sex & race classification, with the following four levels:
#         value            level
#             1      WHITE MALES
#             2   MINORITY MALES
#             3    WHITE FEMALES
#             4 MINORITY FEMALES


#########################################################################################################################################
#########################################################################################################################################
##                                                                                                                                     ##
##                                                    ANSWER TO QUESTION 2                                                             ##
##                                                                                                                                     ##
#########################################################################################################################################
#########################################################################################################################################

# Id, although numeric, is not statistically significant to be visualized as every employee has a unique id in ascending order.
nums <- sapply(salary,class)=='numeric'
nums[1] <- FALSE
salary.num <- salary[,nums]
summary <- summary(salary.num)
# install.packages("psych")
library(psych)
describe <- round(t(describe(salary.num)),2)
# Histogram and Q-Q plot of a normal distribution (rnorm)
# Normal (or Gaussian) distribution is a type of continuous probability distribution for a real-valued random variable
# 1. A variable that is normally distributed has a histogram that is bell-shaped, with only one peak, and is symmetric
# around the mean.
# 2. (Q-Q) plot, shows the distribution of the data against the expected normal distribution. For normally distributed data,
# observations should lie approximately on a straight line. As all the points fall approximately along this reference line,
# we can assume normality.
set.seed(1)
norm <- rnorm(1000)
round(t(describe(norm)),2)
hist(norm, freq = FALSE)
lines(x = density(norm), col = "red")
qqnorm(norm)
qqline(norm)

# All plots in a function, press enter for “changing pages”
eda.plots <- function(data, ask=F){
  graphics.off()
  numeric.only <- sapply(data,class)=='numeric'
  y <- data[,numeric.only]
  n<-ncol(y)
  for (i in 1:n){
    if (!ask) win.graph()
    par(mfrow=c(2,2), ask=ask)
    y1 <- y[,i]
    vioplot(y1)
    hist(y1, probability=TRUE, main=names(y)[i])
    lines(density(y1), col=2)
    qqnorm(y1, main=names(y)[i])
    qqline(y1)
    boxplot(y1, main=names(y)[i], horizontal=TRUE)
  }
}
# install.packages("vioplot")
library(vioplot)
eda.plots(salary.num, ask=T)
graphics.off()
```

```
# Q-Q Plots
# Observations should lie approximately on the straight Q-Q line. As all the points do not fall approximately along this
# reference line, we can not assume normality, by only viewing Q-Q Plots.
par(mfrow=c(2,3))
for (i in 1:length(salary.num)){
  qqnorm(salary.num[,i], main = paste("Q-Q Plot of" , names(salary.num)[i]))
  qqline(salary.num[,i])
}

# Log Q-Q Plots
# As natural logarithm of 0 is undefined we add to the variable work a small positive value.
# As before, even after log transformation, a large amount of points do not fall on the Q-Q line, even for variables time, work
sapply(salary.num,min)
salary.num[,6] <- salary.num[,6]+0.5
salary.num.log<-log(salary.num)
p<-ncol(salary.num.log)
par(mfrow=c(2,3))
for (i in 1:p){
  qqnorm(salary.num.log[,i], main = paste("Q-Q Plot of Log" , names(salary.num)[i]))
  qqline(salary.num.log[,i])
}

# Histograms and Density Plots (red) against Normal Distribution Density Plot (blue)
# As before, the two density plots are not even close. Even variable time that seems symmetric, is not bell-shaped and
# has many peaks.
p<-ncol(salary.num)
par(mfrow=c(2,3))
for (i in 1:p){
  hist(salary.num[,i], main=names(salary.num)[i], probability=TRUE)
  lines(density(salary.num[,i]), col=2)
  index <- seq(min(salary.num[,i]), max(salary.num[,i]), length.out=500)
  ynorm <- dnorm(index, mean=mean(salary.num[,i]), sd(salary.num[,i]) )
  lines(index, ynorm, col=4, lty=3, lwd=3 )
}
par(mfrow=c(1,1))


##########################################################################################################################################
##########################################################################################################################################
##                                                                                                                                      ##
##                                                    ANSWER TO QUESTION 3                                                              ##
##                                                                                                                                      ##
##########################################################################################################################################
##########################################################################################################################################

# Hypothesis test for a single continuous variable
# Variable: salbeg
# Null Hypothesis:  h0: μ = 1000
# Alternative:      h1: μ != 1000

# install.packages("nortest")
# install.packages("lawstat")
# I believe constructing a function to handle the answer explains all cases and choices of tests
library(nortest)
library(lawstat)
single_cont <- function(data, column, condition, sl = 0.05){
  var1 <- data[,column]
  par(mfrow=c(1,2))
  # Can we assume normality?
  test1 <- 1
  if (length(var1) > 50){
    test1 <- lillie.test(var1)$p.value
  }
  test2 <- shapiro.test(var1)$p.value
  qqnorm(var1, main = paste("Q-Q Plot of" , names(data)[column]))
  qqline(var1)
  if ((test1 < sl) | (test2 < sl)){
    # Is the sample large?
    if (length(var1) > 50){
      # Is the mean a sufficient descriptive measure for central location?
      test3 <- symmetry.test(var1)$p.value
      boxplot(var1, main = paste("Boxplot of" , names(data)[column]), horizontal=TRUE)
    }
    if ((length(var1) <= 50) || (test3 < sl)){
      final <- wilcox.test(var1, mu = condition)$p.value
    }
  }
  if (((test1 >= sl) && (test2 >= sl)) || (test3 >= sl)){
    final <- t.test(var1, mu = condition)$p.value
  }
  if (final < sl){
    message <- paste("We reject the null hypothesis. P-value =", round(final,2))
  } else {
    message <- paste("We cannot reject the null hypothesis. P-value =", round(final,2))
  }
  return(message)
  par(mfrow=c(1,1))
}
single_cont(salary, 2, 1000)


##########################################################################################################################################
##########################################################################################################################################
##                                                                                                                                      ##
##                                                    ANSWER TO QUESTION 4                                                              ##
##                                                                                                                                      ##
##########################################################################################################################################
```

```
#########################################################################################################################################

# Hypothesis test for difference between the two dependent values/measurements
# Variable: diff of salnow and salbeg
# Null Hypothesis:  h0: μ = 0
# Alternative:      h1: μ != 0

x <- salary$salnow - salary$salbeg
length(x)
# Above 50, so KS + SW
lillie.test(x)
shapiro.test(x)
qqnorm(x, main = paste("Q-Q Plot of diff of" , names(salary)[6], "and", names(salary)[2]))
qqline(x)
# So we cannot assume normality and the sample is large
symmetry.test(x)
boxplot(salary$salnow,salary$salbeg, main = paste("Boxplot of" , names(salary)[2], "and", names(salary)[6]), horizontal=TRUE)
# So the mean is not a sufficient descriptive measure of central location for the difference as it is not symmetric
wilcox.test(x, mu = 0)
mean(salary$salnow)/mean(salary$salbeg)
median(salary$salnow)/median(salary$salbeg)
boxplot(x, main = paste("Boxplot of diff of" ,  names(salary)[6], "and", names(salary)[2]), horizontal=TRUE)
# So we reject the null hypothesis. There is significant difference between the beginning and current salary and in fact the
# current salary is about the double of the beginning


#########################################################################################################################################
#########################################################################################################################################
##                                                                                                                                     ##
##                                                    ANSWER TO QUESTION 5                                                             ##
##                                                                                                                                     ##
#########################################################################################################################################
#########################################################################################################################################

# Hypothesis test for two samples - Testing for the association between a continuous and a categorical variable of two levels
# Variables: continuous: salbeg, categorical: sex
# Null Hypothesis:  h0: μ1 = μ2
# Alternative:      h1: μ1 != μ2

dataset1 <- salary[,2:3]
n1 <- nrow(subset(dataset1, dataset1$sex=="MALES"))
n2 <- nrow(subset(dataset1, dataset1$sex=="FEMALES"))
n1 > 50
n2 > 50
# As at least 1 of the 2 samples amount is more than 50 (the other one is also above 50), we test normality with KS, SW, Q-Q Plot
by(dataset1$salbeg, dataset1$sex, lillie.test)
by(dataset1$salbeg, dataset1$sex, shapiro.test)
males <- subset(dataset1, dataset1$sex=="MALES")
females <- subset(dataset1, dataset1$sex=="FEMALES")
par(mfrow=c(1,2))
qqnorm(males$salbeg, main = paste("Q-Q Plot of" , toupper(names(dataset1)[1]), "of", levels(dataset1$sex)[1]))
qqline(males$salbeg)
qqnorm(females$salbeg, main = paste("Q-Q Plot of" , toupper(names(dataset1)[1]), "of", levels(dataset1$sex)[2]))
qqline(females$salbeg)
par(mfrow=c(1,1))
# So we cannot assume normality for either of the two samples and both samples are large
by(dataset1$salbeg, dataset1$sex, symmetry.test)
par(mfrow=c(1,2))
boxplot(males$salbeg, main = paste("Boxplot of" , toupper(names(dataset1)[1]), "of", levels(dataset1$sex)[1]))
boxplot(females$salbeg, main = paste("Boxplot of" , toupper(names(dataset1)[1]), "of", levels(dataset1$sex)[2]))
par(mfrow=c(1,1))
# So the mean is not a sufficient descriptive measure of central location for either of the two samples
# Wilcoxon rank-sum test (or Mann-Whitney)
wilcox.test(males$salbeg, females$salbeg)
mean(males$salbeg)/mean(females$salbeg)
median(males$salbeg)/median(females$salbeg)
boxplot(males$salbeg, females$salbeg, names=levels(dataset1$sex)[1:2],
        main = paste("Boxplot of" , toupper(names(dataset1)[1]), "of", levels(dataset1$sex)[1], "and", levels(dataset1$sex)[2]),
        notch = TRUE)
# So we reject the null hypothesis. There is significant difference between the beginning salary of men and women and in fact the
# beginning salary of men is greater than the respective of women


#########################################################################################################################################
#########################################################################################################################################
##                                                                                                                                     ##
##                                                    ANSWER TO QUESTION 6                                                             ##
##                                                                                                                                     ##
#########################################################################################################################################
#########################################################################################################################################

# install.packages("Hmisc")
library(Hmisc)
salary$age_cut <- cut2(salary$age, g = 3)
# Hypothesis test for three samples - Testing for the association between a continuous and a categorical variable of three levels
# Variables: continuous: salbeg, categorical: age_cut
# Null Hypothesis:  h0: μ1 = μ2 = μ3
# Alternative:      h1: μk != μj, for some k != j ε {1,2,3}
# ANOVA: Analysis of Variance

dataset2 <- salary[, c(2,12)]
nrow(dataset2) > 50

# ASSUMPTIONS:
# Residuals' normality or the sample size to be large (n>50)
# Equal variances
# As total number of samples (n1 + n2 + n3) above 50, we test normality with KS, SW, Q-Q Plot
```

```
anova1 <- aov(salbeg~age_cut, data = dataset2)
lillie.test(anova1$residuals)
shapiro.test(anova1$residuals)
qqnorm(anova1$residuals)
qqline(anova1$residuals)
bartlett.test(salbeg~age_cut, data = dataset2)
fligner.test(salbeg~age_cut, data = dataset2)
library(car)
leveneTest(salbeg~age_cut, data = dataset2)
# So we cannot assume neither normality nor homoscedasticity for either of the residuals and the total number of samples is large
boxplot(salbeg~age_cut, data = dataset2, main =
          paste("Boxplot of" , toupper(names(dataset2)[1]), "of", levels(dataset2$age_cut)[1], ",",
                levels(dataset2$age_cut)[2], "and",levels(dataset2$age_cut)[3]))
# So the mean is not a sufficient descriptive measure of central location for either of the three samples
# Kruskal-Wallis Test (Equality of medians)
kruskal.test(salbeg~age_cut, data = dataset2)
# We reject the null hypothesis
pairwise.wilcox.test(dataset2$salbeg, dataset2$age_cut)
boxplot(salbeg~age_cut, data = dataset2, main =
          paste("Boxplot of" , toupper(names(dataset2)[1]), "of", levels(dataset2$age_cut)[1], ",",
                levels(dataset2$age_cut)[2], "and",levels(dataset2$age_cut)[3]))
# So we reject the null hypothesis. There is significant difference in the beginning salary of at least one age group
# and in fact the beginning salary of the 2nd age group ([29.7, 39.8)) is greater than the respective of the other two


##############################################################################################################################
##############################################################################################################################
##                                                                                                                          ##
##                                                  ANSWER TO QUESTION 7                                                    ##
##                                                                                                                          ##
##############################################################################################################################
##############################################################################################################################

# Hypothesis test for two categorical variables
# Testing for the association between two categorical variables (independent samples)
# Variables: categorical: minority, categorical: sex
# Null Hypothesis:  h0: nmales = nfemales
# Alternative:      h1:  nmales != nfemales
tab1 <- table(salary$sex, salary$minority)
prop.table(tab1, 2)
prop.test(tab1)
chisq.test(tab1)
fisher.test(tab1)
# We cannot reject the null hypothesis that the proportion of white male employees is equal to the proportion of white female
# employees, or otherwise we state that there is independence between gender and minority of employees.
```