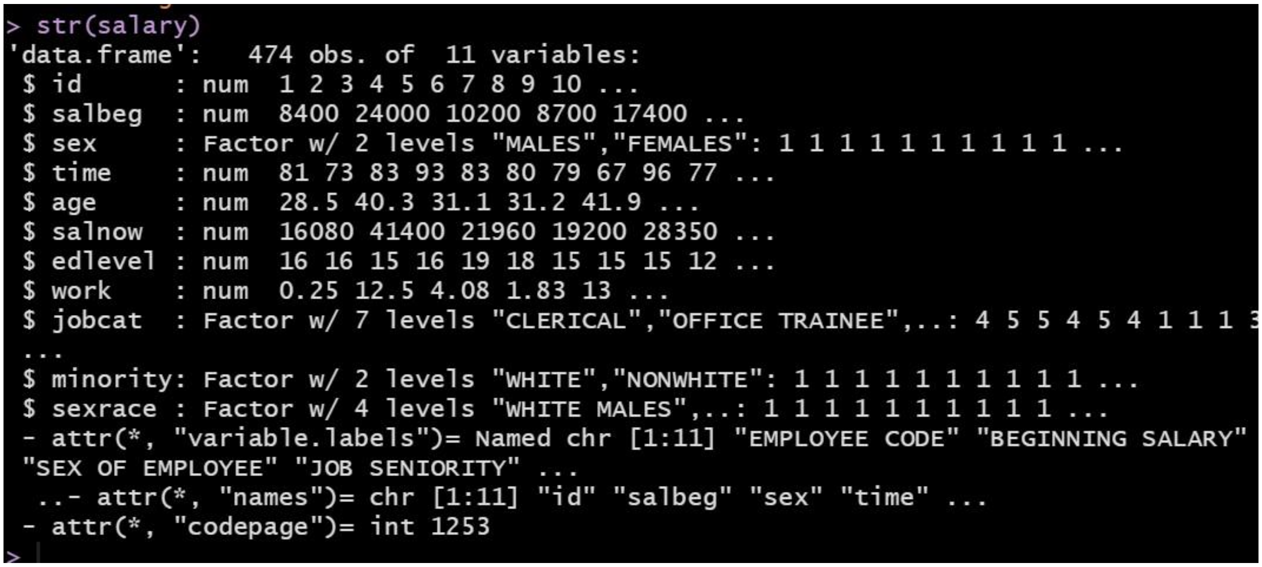
STATISTICS FOR BA I – R ASSIGNMENT 3

Question 1 - Read file 'salary' and identify its features using function str()

First, we use the function ‘read.spss’ to read the file and save it into a new variable called ‘Salary’. Using the str() function we observe that the dataset contains 474 observations (lines) and 11 variables (columns).

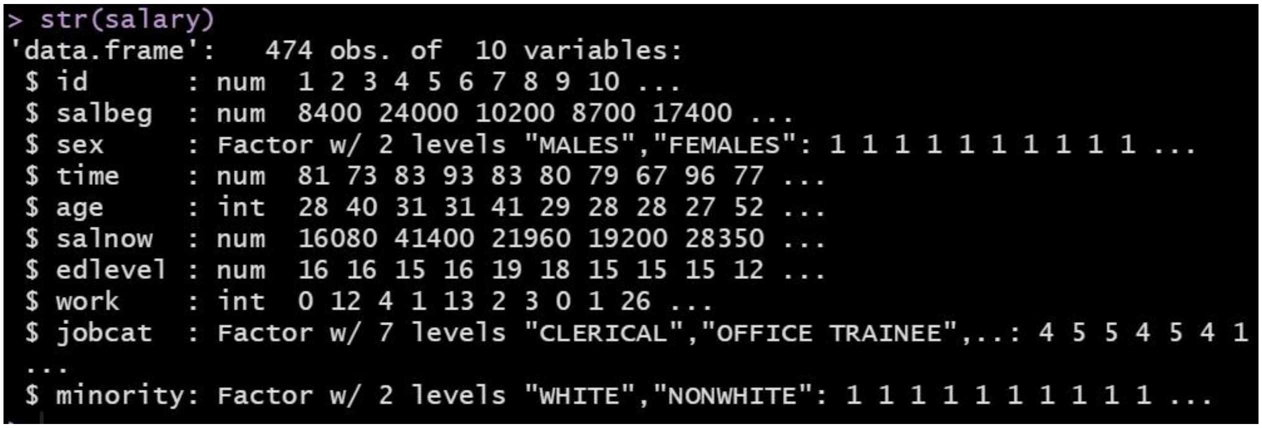


```
> str(salary)
'data.frame':   474 obs. of  11 variables:
 $ id      : num  1 2 3 4 5 6 7 8 9 10 ...
 $ salbeg  : num  8400 24000 10200 8700 17400 ...
 $ sex     : Factor w/ 2 levels "MALES","FEMALES": 1 1 1 1 1 1 1 1 1 1 ...
 $ time    : num  81 73 83 93 83 80 79 67 96 77 ...
 $ age     : num  28.5 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
>
```

The columns that are included are the following:

- ID of employee (numeric)
- Salbeg - Starting salary of employee (numeric)
- sex – Gender of employee (factor)
- time – seniority (numeric)
- age of employee (numeric)
- salnow – Current salary (numeric)
- edlevel – Education level (numeric)
- work – Years of work experience (numeric)
- jobcat – Job description (factor)
- minority – ethnicity (factor)
- sexrace – gender and ethnicity (factor)

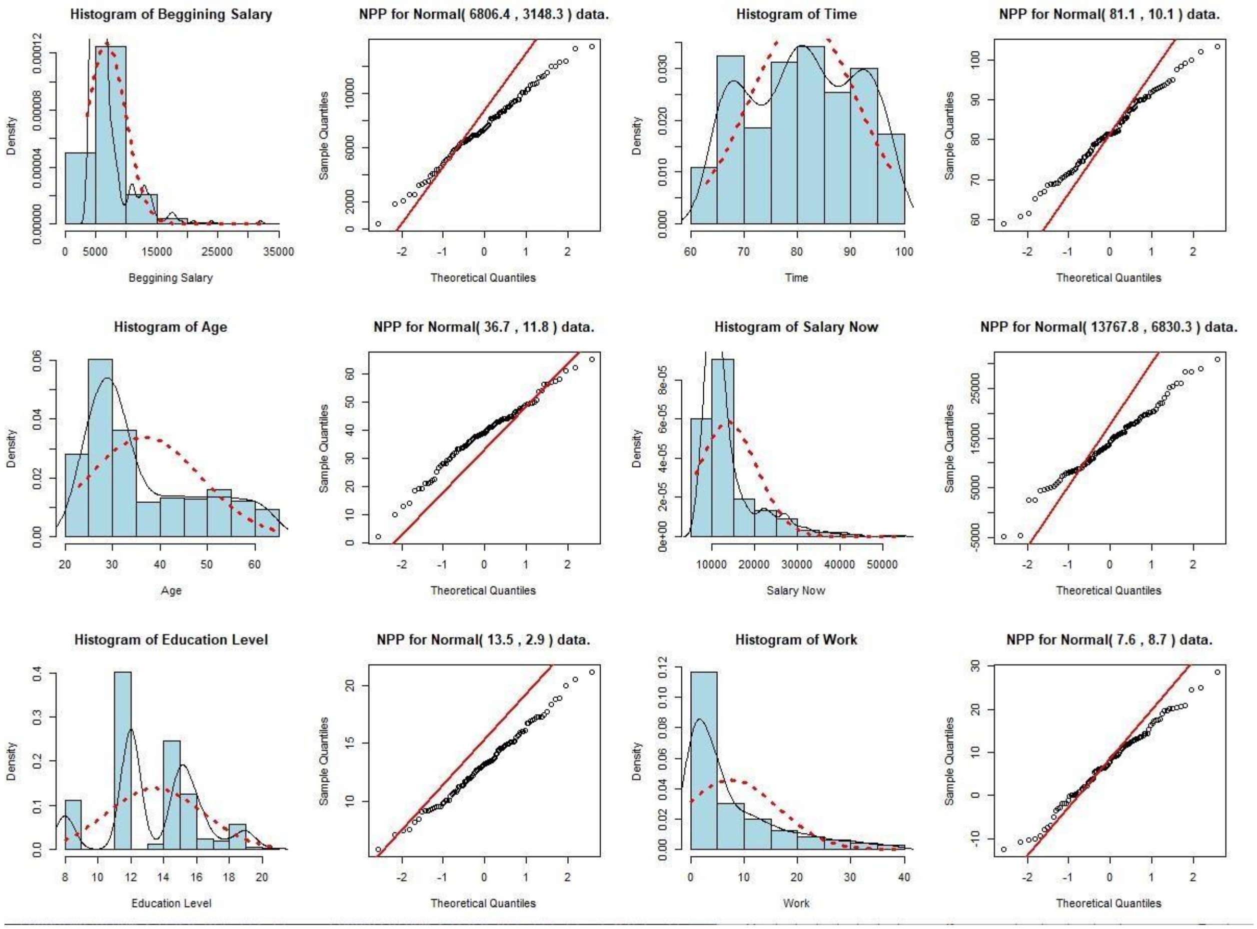
It is noticed after examination that the variable sexrace is redundant because it consists of two variables that already exist in the dataset. We can make better use of the other two variables, so the variable sexrace is dropped. Also, the variables age and work have decimal values in most observations, so we transformed them to values with integer type. The rest of the variables did not appear to need any kind of transformation.



```
> str(salary)
'data.frame':	474 obs. of  10 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     : int  28 40 31 31 41 29 28 28 27 52 ...
 $ salnow  : num  16080 41400 21960 19200 28350 ...
 $ edlevel : num  16 16 15 16 19 18 15 15 15 12 ...
 $ work    : int  0 12 4 1 13 2 3 0 1 26 ...
 $ jobcat  : Factor w/ 7 levels "CLERICAL","OFFICE TRAINEE",..: 4 5 5 4 5 4 1
 ...
 $ minority: Factor w/ 2 levels "WHITE","NONWHITE": 1 1 1 1 1 1 1 1 1 1 ...
```

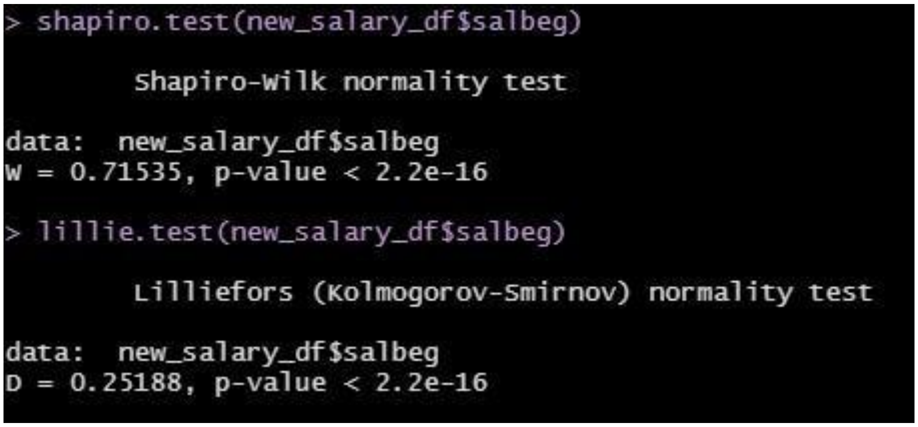
Question 2 - Get summary of numeric values in the data frame and visualize them

We create a dummy variable that consists of the data classes of the original datatype 'Salary'. We find the class of each variable using the sapply() function. Then, we create a new data frame that consists only of the variables that have a 'numeric' or 'integer' class and we drop the id column. Then, we use the function summary() to observe the summarized contents of each of the variables collected above. We observe that the variables salnow, salbeg and work seem to have outliers inside their values. The rest of the variables do not seem to have such values. Finally, we use the functions hist(), lines(), qqnorm(), qqlines() inside a for loop to produce the below outcome of graphs. None of those variables, appear to be normally distributed because they don't have the 'bell' shape of the normal distributions on the following graphs. Time variable has a shape close to normal, but its tails are large, so it is probably following a student distribution.



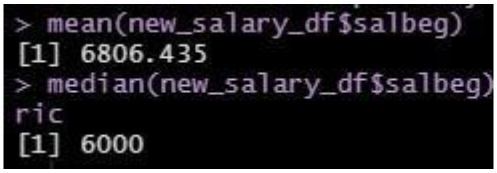
Question 3 - Examine if the starting salary can be 1000 dollars. Interpret the results.

Considering that we examine one sample, we first examine its normality using the Shapiro.test() and the Lillie.test() functions. In both cases, with a confidence level of 5% we reject the null hypothesis that the sample is normal. (p-value < 0.05)



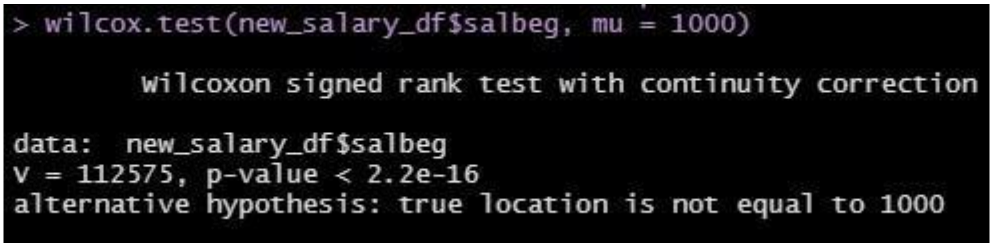
```
> shapiro.test(new_salary_df$salbeg)

        Shapiro-Wilk normality test

data:  new_salary_df$salbeg
W = 0.71535, p-value < 2.2e-16

> lillie.test(new_salary_df$salbeg)

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  new_salary_df$salbeg
D = 0.25188, p-value < 2.2e-16
```

So, in the next step we examine whether it's a large sample. Using the length() function, we observe that it consists of 474 observations, so it's considered a large sample (>50). So, we must examine its symmetry/skewness. We check if the mean is close to the median using the functions mean() and median() respectively. Their difference is more than 800 so we don’t consider the sample to be symmetric.



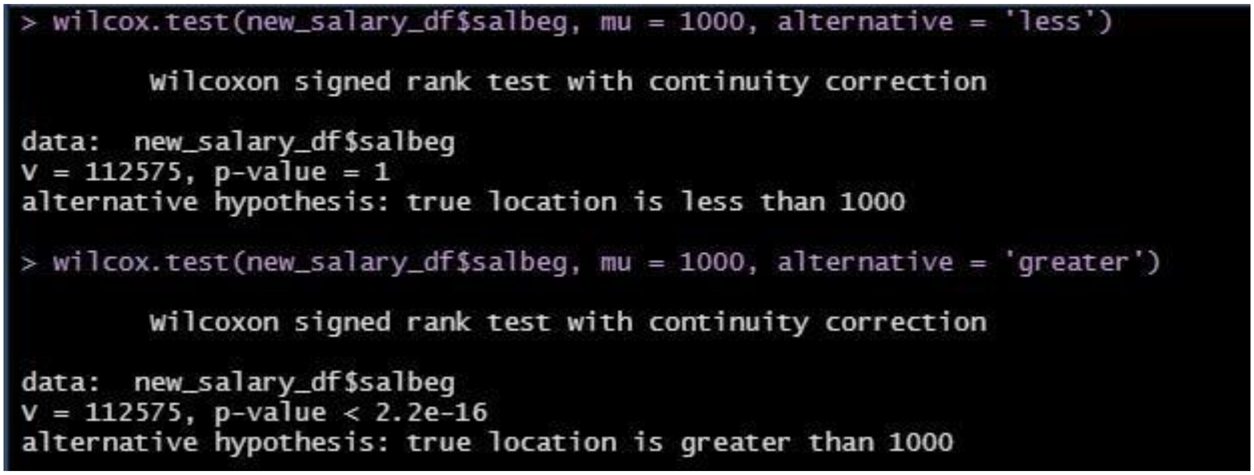
```
> mean(new_salary_df$salbeg)
[1] 6806.435
> median(new_salary_df$salbeg)
ric
[1] 6000
```

So, in the final step we use the non-parametric Wilcoxon test for one sample to examine if the median starting salary of the population can be equal to 1000.



```
> wilcox.test(new_salary_df$salbeg, mu = 1000)

        Wilcoxon signed rank test with continuity correction

data:  new_salary_df$salbeg
V = 112575, p-value < 2.2e-16
alternative hypothesis: true location is not equal to 1000
```

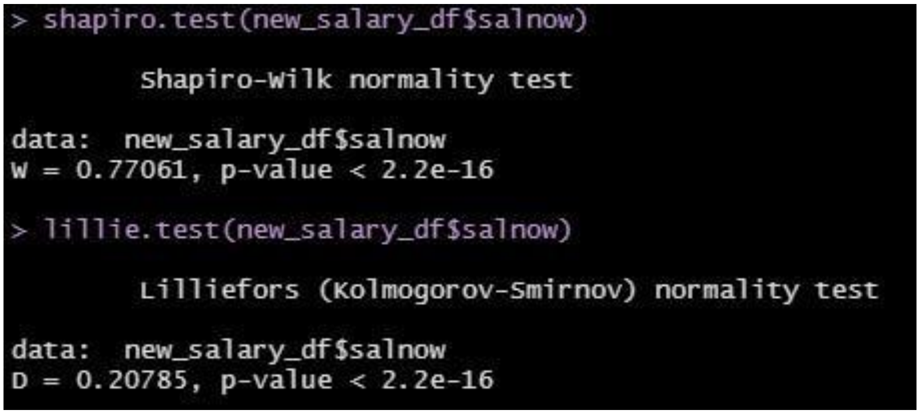
With a 5% confidence level we reject the null hypothesis that the median starting salary of the population is 1000. (p-value < 0.05) Using 2 more Wilcoxon test, with different alternatives (‘less’, ‘greater’) we realize that the median starting salary will be less than 1000. (for less – p-value =1 > 0.05)



```
> wilcox.test(new_salary_df$salbeg, mu = 1000, alternative = 'less')

        Wilcoxon signed rank test with continuity correction

data:  new_salary_df$salbeg
V = 112575, p-value = 1
alternative hypothesis: true location is less than 1000

> wilcox.test(new_salary_df$salbeg, mu = 1000, alternative = 'greater')

        Wilcoxon signed rank test with continuity correction

data:  new_salary_df$salbeg
V = 112575, p-value < 2.2e-16
alternative hypothesis: true location is greater than 1000
```

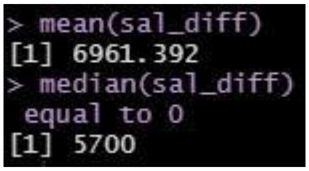
Question 4 - Consider the difference between salbeg and salnow. Test the difference's significance

First, we assign the difference of the variables salbeg and salnow for every observation into a new observation called ‘sal_diff’. We know from **‘Question 3’** that the variable salbeg does not follow a normal distribution. So, even though it is not necessary, we examine the normality of the variable salnow using the functions Lillie.test() and Shapiro.test(). In both cases, with a confidence level of 5%, we reject the null hypothesis that the variable salnow follows a normal distribution, because (p-value < 0.05).



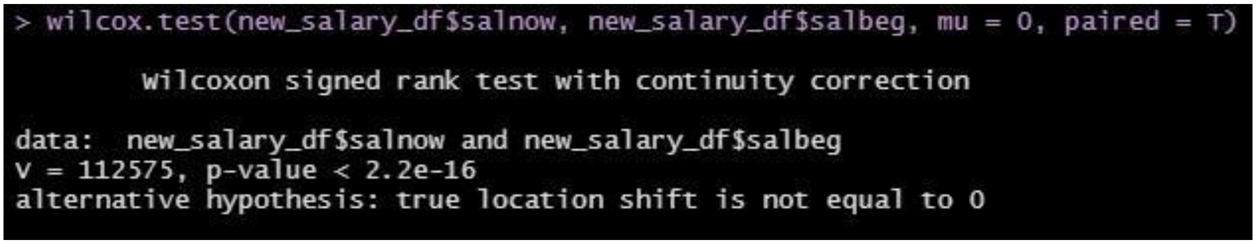
```
> shapiro.test(new_salary_df$salnow)

        Shapiro-Wilk normality test

data:  new_salary_df$salnow
W = 0.77061, p-value < 2.2e-16

> lillie.test(new_salary_df$salnow)

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  new_salary_df$salnow
D = 0.20785, p-value < 2.2e-16
```

In the next step, we examine the new variable's size using the length() function. We observe that it consists of 474 observations, so it is considered a large sample (n > 50). We examine the difference's symmetry, by observing how close are its mean and median, using the mean() and median() functions respectively.



```
> mean(sal_diff)
[1] 6961.392
> median(sal_diff)
 equal to 0
[1] 5700
```

Their difference is considered large, so in the next step we use the non-parametric Wilcoxon paired test, to examine if the median difference between the starting and the current salaries for the population can be equal to 0. We use the paired version of the Wilcoxon test because we examine the same variable for the same employees after a given time (a certain amount of years).

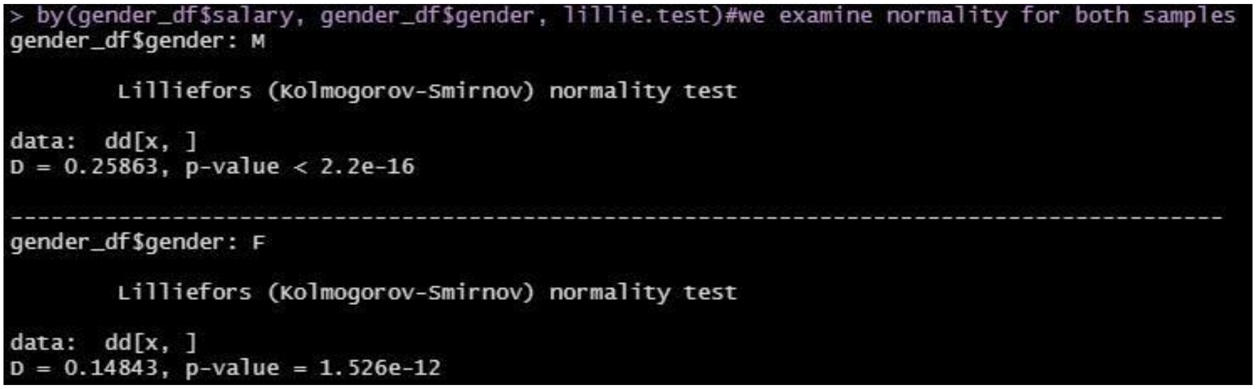


```
> wilcox.test(new_salary_df$salnow, new_salary_df$salbeg, mu = 0, paired = T)

        Wilcoxon signed rank test with continuity correction

data:  new_salary_df$salnow and new_salary_df$salbeg
V = 112575, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
```

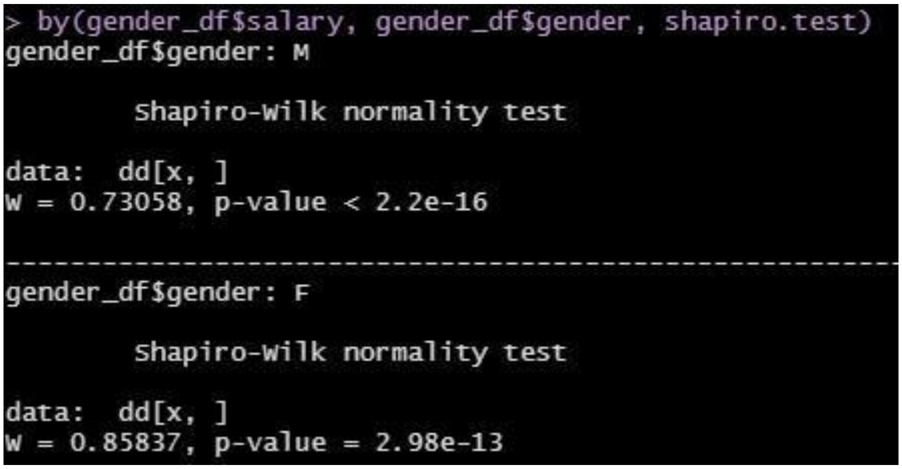
Given that the p-value is less than 0.05, with a confidence level of 5% we reject the null hypothesis that the median starting salary is equal to the median current salary for the population.

Question 5 - Test if there is any difference in the starting salary between the two genders

First, we create two new variables that each contain the starting salaries for men and women respectively. Then, using the function data.frame() we combine the two variables into a common data frame. Using the by function, we apply the Shapiro.test() and Lillie.test() function to the above data frame to examine their normality.

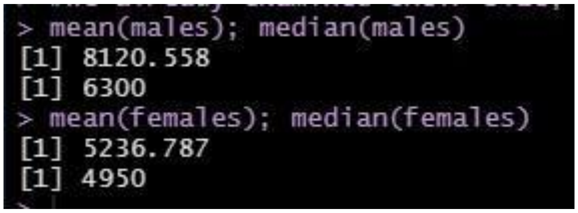


```
> by(gender_df$salary, gender_df$gender, lillie.test)#we examine normality for both samples
gender_df$gender: M

        Lilliefors (Kolmogorov-Smirnov) normality test

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

-------------------------------------------------------------------------------------
gender_df$gender: F

        Lilliefors (Kolmogorov-Smirnov) normality test

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



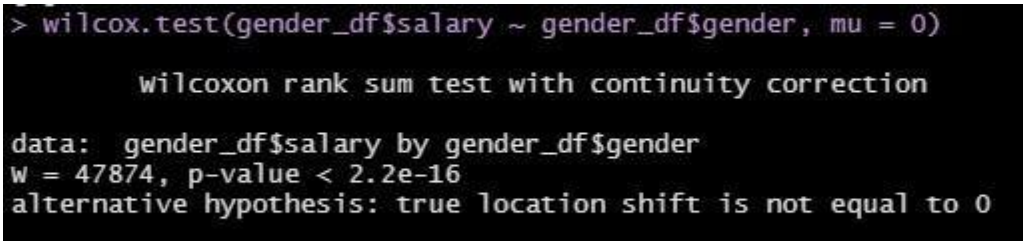
```
> by(gender_df$salary, gender_df$gender, shapiro.test)
gender_df$gender: M

        Shapiro-Wilk normality test

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

-----------------------------------------------------------
gender_df$gender: F

        Shapiro-Wilk normality test

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

In both cases, for both new variables, with a confidence level of 5%, we reject the null hypothesis that they follow a normal distribution, given that they have (p-values < 0.05) in every case. In the next step, we examine their symmetry using the mean() and median() functions for each of them separately.



```
> mean(males); median(males)
[1] 8120.558
[1] 6300
> mean(females); median(females)
[1] 5236.787
[1] 4950
```

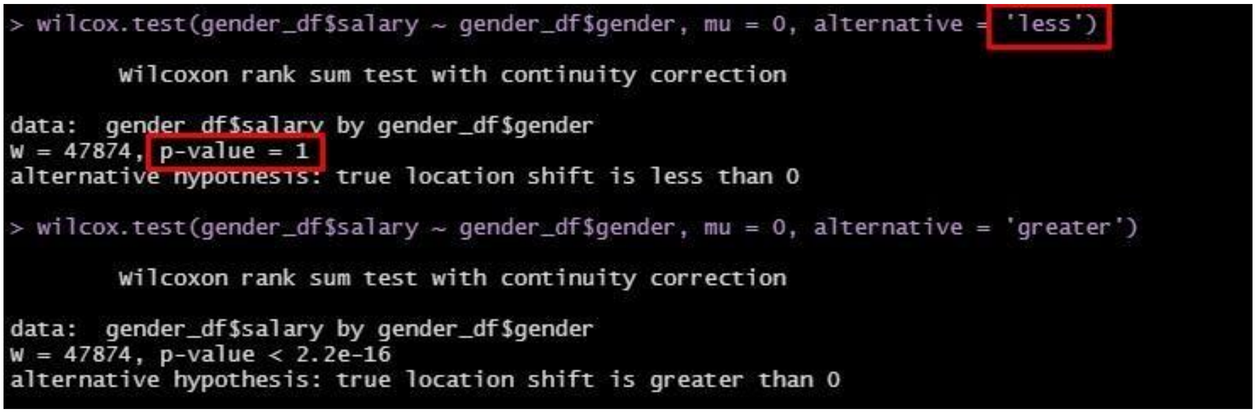
While the values for the starting salaries of the females seem to be symmetric, the values for men do not. So, we use the non-parametric Wilcoxon test to examine if the median starting salary is equal for both genders.



```
> wilcox.test(gender_df$salary ~ gender_df$gender, mu = 0)

        Wilcoxon rank sum test with continuity correction

data:  gender_df$salary by gender_df$gender
W = 47874, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
```

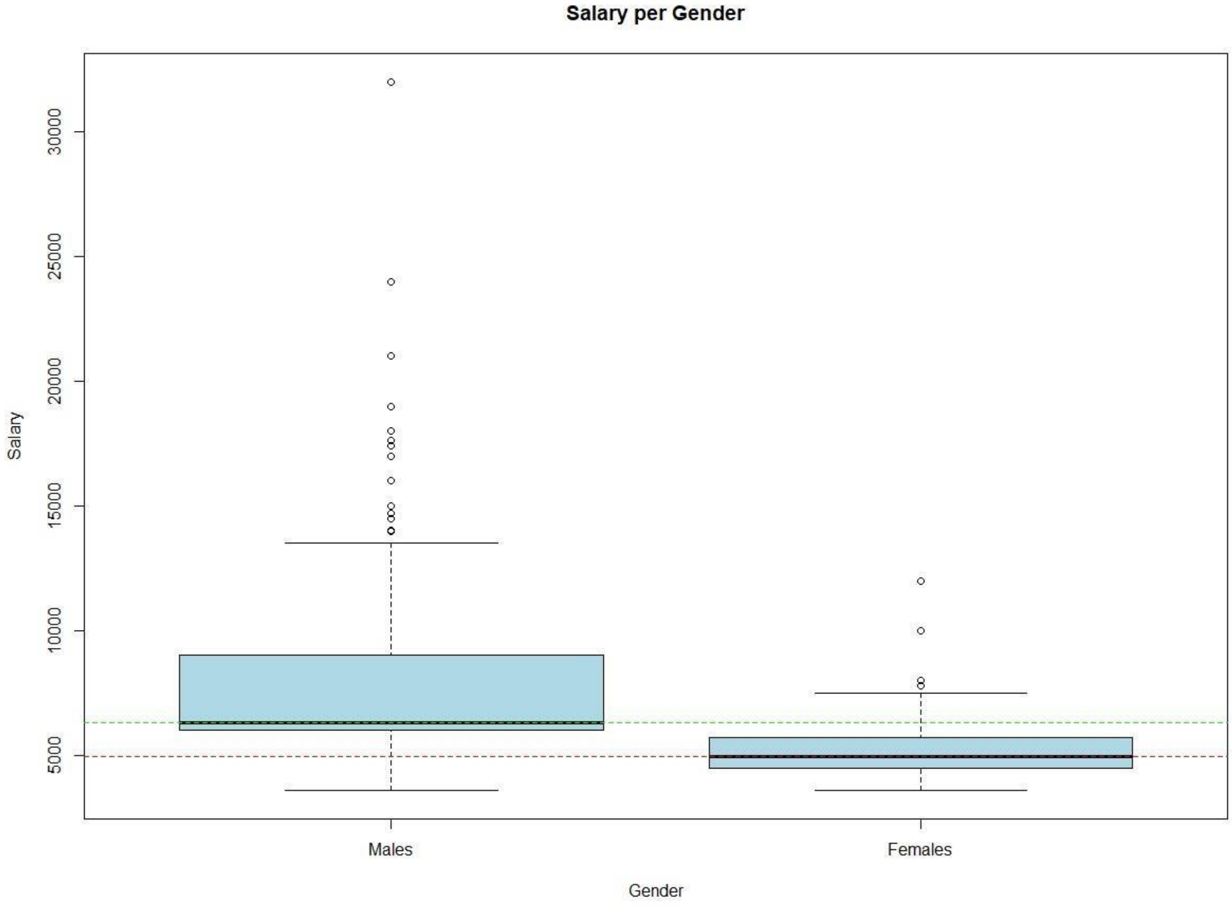
Given that the (p-value < 0.05) and with a confidence level of 5% we reject the null hypothesis that the median starting salary is equal for both genders. By experimenting (changing the alternative option in

Wilcoxon test to ‘less’, ‘greater’) we observe that the median starting salary is greater for men, in comparison to women.



```
> wilcox.test(gender_df$salary ~ gender_df$gender, mu = 0, alternative = 'less')

        Wilcoxon rank sum test with continuity correction

data:  gender_df$salary by gender_df$gender
W = 47874, p-value = 1
alternative hypothesis: true location shift is less than 0

> wilcox.test(gender_df$salary ~ gender_df$gender, mu = 0, alternative = 'greater')

        Wilcoxon rank sum test with continuity correction

data:  gender_df$salary by gender_df$gender
W = 47874, p-value < 2.2e-16
alternative hypothesis: true location shift is greater than 0
```

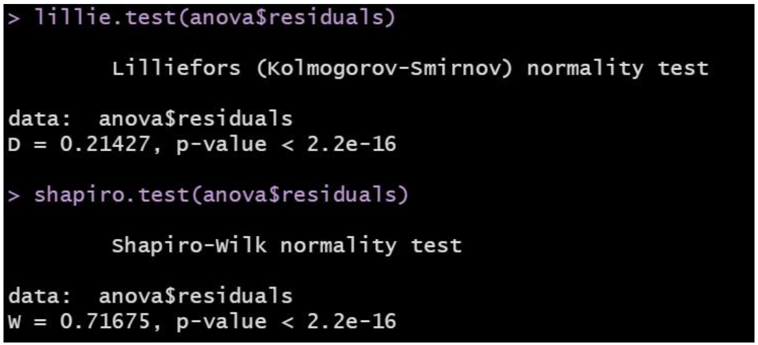
To confirm the above assumption, we plot boxplots for the starting salaries of men and women separately.



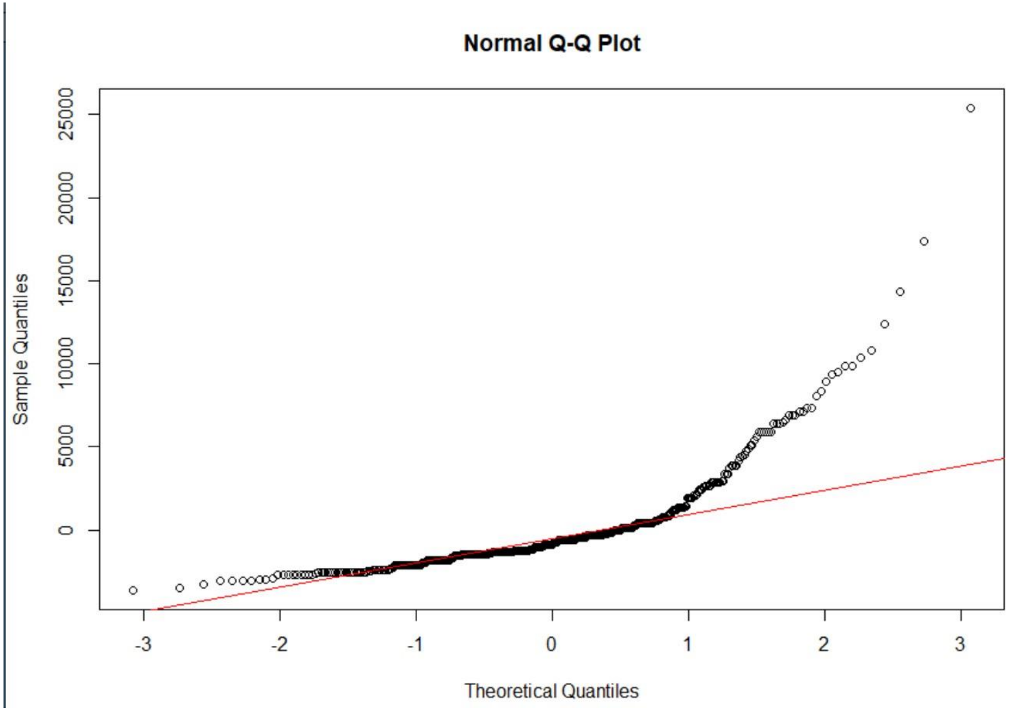
Indeed, we observe that the median of each gender is not included in the values of the other, so they can’t be equal. For more precision, we drew horizontal lines on the value of the median of each gender.

Question 6 - Cut the age variable to 3 categories. Examine if the starting salary is the same for all age groups

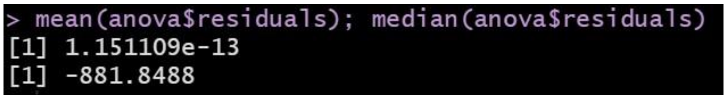
First, we use the function cut2() to cut the sample into 3 new subsamples. Then, we assign the starting salaries of each group to 3 new variables, named accordingly. We examine normality for each one of these 3 new variables.



```
> lillie.test(anova$residuals)

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  anova$residuals
D = 0.21427, p-value < 2.2e-16

> shapiro.test(anova$residuals)

        Shapiro-Wilk normality test

data:  anova$residuals
W = 0.71675, p-value < 2.2e-16
```

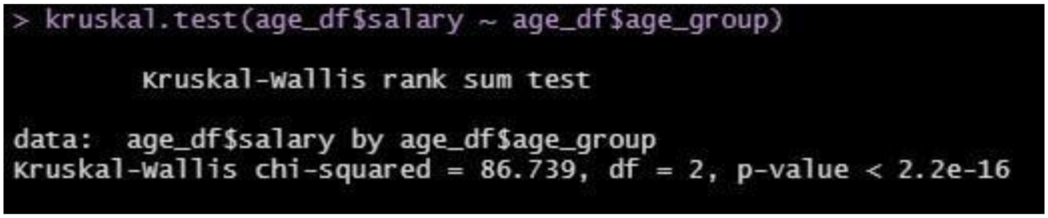


The normality tests for the residuals provide a (p-value < 0.05), so with a confidence level of 5% we reject the null hypothesis that they follow a normal distribution. So, we examine their symmetry.



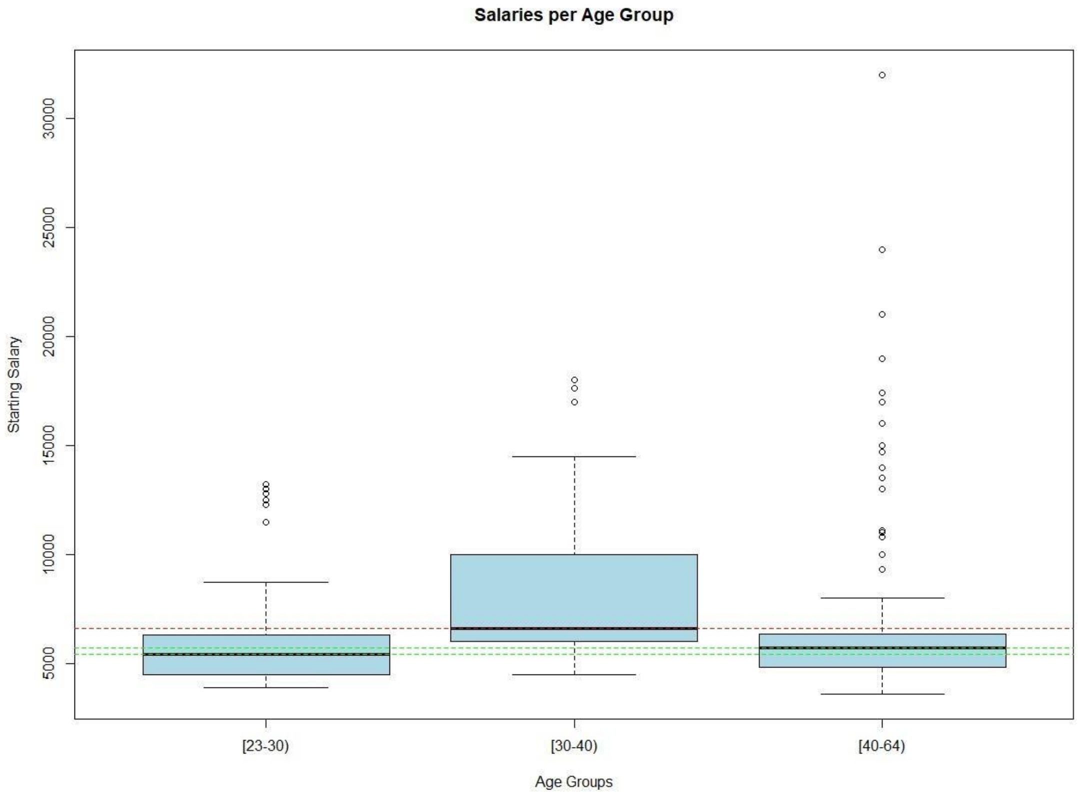
```
> mean(anova$residuals); median(anova$residuals)
[1] 1.151109e-13
[1] -881.8488
```

Given that the difference between the mean and the median is great, we reject symmetry hypothesis of the residuals. So we apply a Kruskal-wallis test on the variables.

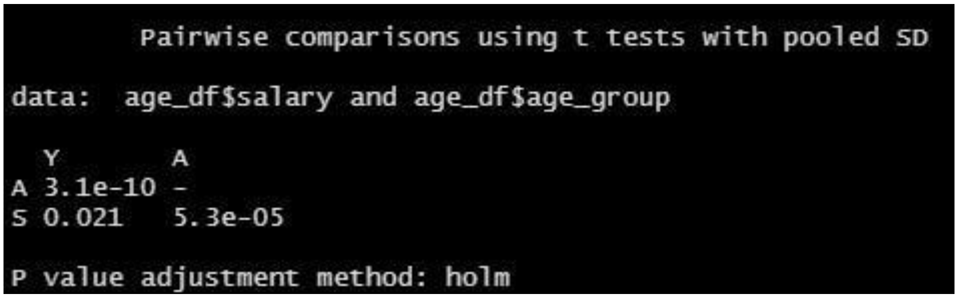


```
> kruskal.test(age_df$salary ~ age_df$age_group)

        Kruskal-Wallis rank sum test

data:  age_df$salary by age_df$age_group
Kruskal-Wallis chi-squared = 86.739, df = 2, p-value < 2.2e-16
```

We observe that (p-value < 0.05), so with a confidence level of 5% we reject the null hypothesis that the median of starting salary is equal for all age groups throughout the population. To confirm the above outcome, we plot boxplots for each age group. We also plotted horizontal lines at the values of the median of each age group, for more precision on our assumptions.



We also used the non-parametric pairwise.t.test to examine all the pairwise comparisons of the 3 variables.

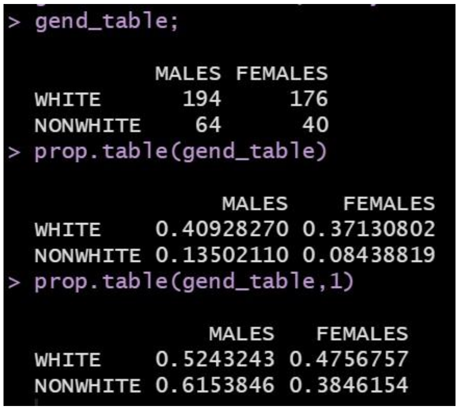


```
        Pairwise comparisons using t tests with pooled SD

data:  age_df$salary and age_df$age_group

  Y       A
A 3.1e-10 -
S 0.021   5.3e-05

P value adjustment method: holm
```

We observe that with a confidence level of 1%, we do not reject the null hypothesis that the median starting salary for young adults can be equal to the median starting salary of the senior adults of the population. The hypothesis of equality for the rest pairs of the variables are rejected with a confidence level of 5%, because they all have a p-value < 0.05.

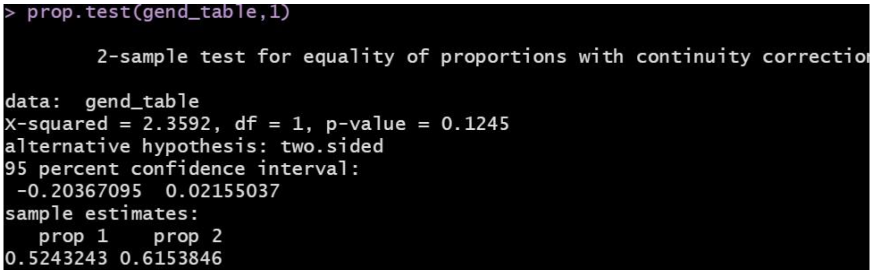
Question 7 - Examine if the proportion of the white males is equal to the proportion of white females

First, we created a table containing all frequencies of gender and minority values in the sample. Then we use the prop.table() functions to examine the percentile frequencies of each corresponding pair. At first, we observe that the number of white males in the company is greater than the number of white females.



```
> gend_table;

          MALES FEMALES
  WHITE     194     176
  NONWHITE   64      40
> prop.table(gend_table)

               MALES    FEMALES
  WHITE    0.40928270 0.37130802
  NONWHITE 0.13502110 0.08438819
> prop.table(gend_table,1)

              MALES   FEMALES
  WHITE    0.5243243 0.4756757
  NONWHITE 0.6153846 0.3846154
```

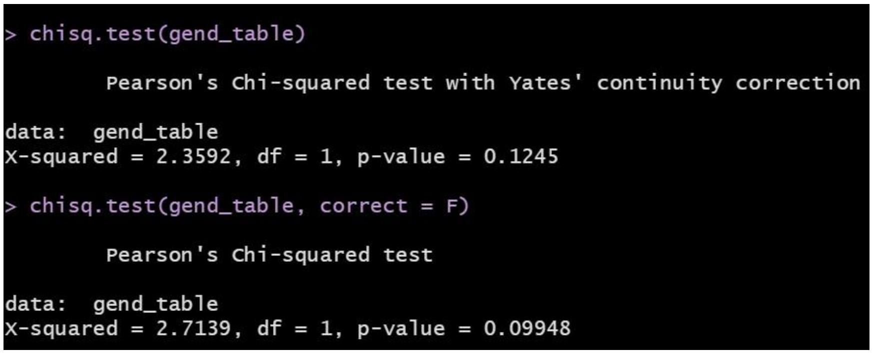
Afterwards, we use the prop.test to examine the independence between the proportion of white females with white males.



```
> prop.test(gend_table,1)

        2-sample test for equality of proportions with continuity correctio

data:  gend_table
X-squared = 2.3592, df = 1, p-value = 0.1245
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.20367095  0.02155037
sample estimates:
   prop 1    prop 2
0.5243243 0.6153846
```

We observe that the p-value for prop 1 is greater than 0.05, so with a confidence level of 5% we do not reject the null hypothesis that the two proportions are independent.

Then we use multiple chisq.tests() to examine their independence.



```
> chisq.test(gend_table)

        Pearson's Chi-squared test with Yates' continuity correction

data:  gend_table
X-squared = 2.3592, df = 1, p-value = 0.1245

> chisq.test(gend_table, correct = F)

        Pearson's Chi-squared test

data:  gend_table
X-squared = 2.7139, df = 1, p-value = 0.09948
```

In all cases, we observe that the p-value is greater than 0.05, so with a confidence level of 5% we do not reject the null hypothesis that the proportion of white men employees is independent to the proportion of the white female employees in the population.

R-Code

```
require(foreign)
library('lawstat')
library('nortest') library('Hmisc')
library('gmodels')
#Question 1 - read file 'salary' and identify its features using function str() salary
<- read.spss("C:\\Users\\user\\Downloads\\salary.sav", to.data.frame = T)
str(salary)
#salary is a dataframe that consists of 474 observations and 11 variables

#the dataset has numeric values in the 'age' & 'work' variables, while they should appear in integer
format
salary$age <- as.integer(salary$age)
salary$work <- as.integer(salary$work)

#sexrace variable is unnecessary in the dataset because it consists of 2 other variables that
#exist in the current dataset. These variables are 'sex' and 'minority'. So, we can drop the variable
'sexrace' salary <-
salary[,-11]
str(salary)


#question 2 - get summary of numeric values in the dataframe and visualize them df_class
<- sapply(salary, class) # find variables that have a numeric type

new_salary_df <- salary[, (df_class == 'integer') | (df_class == 'numeric')] #keep variables with numeric
type
new_salary_df <- new_salary_df[,-1] #remove id variable

summary(new_salary_df) #salary of numeric variables
#salbeg (beggining salary) seems to have some outliers. Max value has a great difference from the
variable's mean and median

coln <- c('Beggining Salary', 'Time', 'Age', 'Salary Now', 'Education Level', 'Work') #giving proper names to
the variables

par(mfrow = c(3,4)) #modify plot window to show all histograms at once for (i
in 1:ncol(new_salary_df)){ #an iteration that goes through all the variables
  hist(new_salary_df[,i], xlab = coln[i], probability = TRUE, col = 'lightblue', main = paste('Histogram
of',coln[i]))
  lines(density(new_salary_df[,i])) #lines that depict the actual variables   index
<- seq( min(new_salary_df[,i]), max(new_salary_df[,i]),length.out=100)    ynorm
<- dnorm(index, mean=mean(new_salary_df[,i]), sd(new_salary_df[,i]) )
#create an index that shows a normal distribution with mean and sd of each
variable   lines( index, ynorm, col='red', lty=3, lwd=3 )
  #lines that depict a normal distribution according to the above index for each variable
qqnorm(rnorm(100,mean = mean(new_salary_df[,i]), sd = sd(new_salary_df[,i]))
        , main = paste('NPP for Normal(',
round(mean(new_salary_df[,i]),1),',',round(sd(new_salary_df[,i]),1),') data.'))
  qqline(rnorm(100,mean = new_salary_df[,i], sd = sd(new_salary_df[,i])),col="red",lty=1,lwd=2)
}

#observing the graphs, we we can not assume normality for none of the variables examined
#because their shaped differ greatly from the normal distribution's bell shape
#and their points do not match with the points of the qqplot graph


#Question 3 - examine if the beggining salary can be 1000 dollars. Interpret the results. #considering
that we examine one variable(salbeg), we first must examine its normality using kolmogorov and
shapiro tests

shapiro.test(new_salary_df$salbeg) lillie.test(new_salary_df$salbeg)
#examining the above outcome, we observe that in both cases the p-value is way less than 0.05. #in
other words, there is enough evidence to reject the null hypothesis that the variable salbeg is
following a normal distribution

#so the next step is to examine whether the sample is large
length(new_salary_df$salbeg) #the sample has 474 variables, so it is considered a large sample

#we examine the sample's symmetry/skewness mean(new_salary_df$salbeg)
median(new_salary_df$salbeg) #the difference between the mean and the median is large enough to
consider the sample assymetric

#we use the wilcoxon test for one sample wilcox.test(new_salary_df$salbeg,
mu = 1000)
#p-value is less than 0.05, so there is enough evidence to reject the null hypothesis that the median of
the population's beggining salary is equal to 1000$

wilcox.test(new_salary_df$salbeg, mu = 1000, alternative = 'less') wilcox.test(new_salary_df$salbeg,
mu = 1000, alternative = 'greater')
#considering the above wilcoxon tests, with a 5% confidence level we assume that the beggining salary
of the population is less than 1000$



#Question 4 - consider the difference between salbeg and salnow. test the difference's significance
sal_diff <- new_salary_df$salnow - new_salary_df$salbeg
```

```
#we know from question 3 that the variable salbeg doesn't follow a normal distribution, so we examine
whether the sample is large shapiro.test(new_salary_df$salnow) lillie.test(new_salary_df$salnow)
#the variable salnow, also does not follow a normal distribution - for reference
#so, the difference of these two, also does not a follow a normal distribution

length(sal_diff) #474 observations, so it's a large sample, so we use examine their symmetry

mean(sal_diff)
median(sal_diff) #the difference between mean and median is large, so we use the wilcoxon test to test
if the difference is equal to 0

wilcox.test(new_salary_df$salnow, new_salary_df$salbeg, mu = 0, paired = T)
#given that the p-value is less than 0.05, we have enough evidence to reject the null hypothesis that the
difference
#of the starting and current salary is 0. So, we can assume that the current salaries are different than the
starting ones

wilcox.test(sal_diff, mu = 0, alternative = 'greater') wilcox.test(sal_diff,
mu = 0, alternative = 'less')
#given the above wilcoxon test, with a 5% confidence level we can assume that the median difference in
starting and current
#salaries of the population is more than 0 so the current salaries, in average are greater than the starting
ones
par(mfrow = c(1,1))
boxplot(new_salary_df$salbeg, new_salary_df$salnow, names = c('Starting Salary', 'Current Salary')
        ,ylab = 'Salary', col = 'lightblue')
abline(h = median(new_salary_df$salbeg), col = 'red', lty = 2)  abline(h
= median(new_salary_df$salnow), col = 'green', lty = 2)
#by observing the boxplots, we can confirm that the median starting salary of the population is not the
#same with the median current salary for the population



#Question 5 - test if there is any difference in the starting salary between the two genders
males <- salary[which(salary$sex == 'MALES'),2] females <- salary[which(salary$sex ==
'FEMALES'),2]

n1 <- length(males)  n2
<- length(females)
gender_df <- data.frame( salary=c(males, females),  gender=factor( rep(1:2, c(n1,n2)), labels=c('M','F') ) )
by(gender_df$salary, gender_df$gender, lillie.test)#we examine normality for both samples
by(gender_df$salary, gender_df$gender, shapiro.test)
#p-value is less than 0.05, so we reject the null hypothesis that the samples are normally distributed
(both of them)
```

```
#we already examined their size, and both of them are large samples (>50), so in the next step we check
their symmetry mean(males); median(males) mean(females); median(females)
#while the sample of females' salaries is symmetric, the sample of males isn't. so we use wilcoxon test
#to test whether their starting salaries are equal

wilcox.test(gender_df$salary ~ gender_df$gender, mu = 0)
#p-value < 0.05, so with a confidence level of 5% we reject the null hypothesis that the median starting
#salary of the population is the same between genders

wilcox.test(gender_df$salary ~ gender_df$gender, mu = 0, alternative = 'less')
wilcox.test(gender_df$salary ~ gender_df$gender, mu = 0, alternative = 'greater')
#given the above p-values, with a confidence level of 5% we can assume that the starting salaries for the
males
#of the population are greater than those of women par(mfrow
= c(1,1))
boxplot(males,females, names = c('Males', 'Females'),xlab = 'Gender', ylab = 'Salary', main = 'Salary per
Gender'
        ,col = 'lightblue')
abline(h = median(males), col = 'green', lty = 2) abline(h
= median(females), col = 'red', lty = 2)
#observing the boxplots we can assume that the median starting salary is not the same
#for the two genders in the population


#Question 6 - cut the age variable to 3 categories. examine if the beggining salary is the same for all age
groups
age_cut <- cut2(salary$age, g = 3)
young_ad <- salary[which(salary$age >= 23 & salary$age<30),2]; len1 <- length(young_ad) adult_pers
<- salary[which(salary$age >= 30 & salary$age<40),2]; len2 <- length(adult_pers) senior_pers <-
salary[which(salary$age >= 40),2]; len3 <- length(senior_pers)

age_df <- data.frame(salary = c(young_ad, adult_pers, senior_pers), age_group = factor(rep(1:3,
c(len1,len2,len3)), labels=levels(age_cut))) anova <- aov(salary ~ age_group,data = age_df)
lillie.test(anova$residuals) shapiro.test(anova$residuals)
qqnorm(anova$residuals) qqline(anova$residuals, col =
'red')
#we examine normality for the residuals of the salaries of the age groups
#(p-value < 0.05), so with a confidence level of 5% we reject the null hypothesis
#that the residuals follow a normal distribution

#we already confirmed that all 3 of the samples are large (>50), so in the next step we examine their
symmetry
mean(anova$residuals); median(anova$residuals)
```

```
#the difference between the mean and the median is large, so there is no symmetry in the model

#so we examine if the 3 samples have equal medians using the kruskal-wallis test
kruskal.test(age_df$salary ~ age_df$age_group)
#given that the p-value is (<0.05), with a confidence level of 5% we reject the null hypothesis that the
starting
#salary is the same for each age group.  boxplot(age_df$salary ~ age_df$age_group, names = c('[23-
30)','[30-40)','[40-64]'),ylab = 'Starting Salary'
        ,xlab = 'Age Groups', main = 'Salaries per Age Group', col = 'lightblue')
abline(h = median(young_ad), col = 'green', lty = 2)
abline(h = median(adult_pers), col = 'red', lty = 2) abline(h
= median(senior_pers), col = 'green', lty = 2)
pairwise.t.test(age_df$salary,age_df$age_group)
#observing the boxplots and the outcome of the pairwise.t.test, we can assume that with a confidence
level of 1%
#seniors and young adults have a mean equal starting salary (p-value > a = 0.01). adults don't have equal
mean salary
#with other age groups

#Question 7 - examine if the proportion of the white males is equal to the proportion of white females
gend_table <- table(salary$minority, salary$sex) gend_table; prop.table(gend_table)
prop.table(gend_table,1)
#at first glance, we observe that the proportion of white men is greater than the proportion
#of white women in the sample  prop.test(gend_table,1)
#observing the prop.test outcome, with a confidence level of 5% we do not reject
#the null hypothesis that the proportion of white males is independent to the
#proportion of white females of the population

chisq.test(gend_table)
chisq.test(gend_table, correct = F)
#through the chisq.test we observe that (p-value > 0.05) so we do not reject the null hypothesis
#that the proportion of white male employees is independent to the proportion of the white female
employees
```