STATISTICS FOR BUSINESS ANALYTICS I
LAB ASSIGNMENT – VRETTEAS STYLIANOS – p2822003

CASE: The salary data frame contains information about 474 employees hired by a Midwestern bank between 1969 and 1971. It was created for an Equal Employment Opportunity (EEO) court case involving wage discrimination. The file contains beginning salary (SALBEG), salary now (SALNOW), age of respondent (AGE), seniority (TIME), gender (SEX coded 1 = female, 0 = male)
among other variables.

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

Structure of the salary object created by SPSS file salary.sav using str()

```
'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 above output indicates that salary object is a dataframe with 474 observations of 11 variables. Below there are the names of each variable alongside its data type (numeric – factor) and the first observations. For the factor data type there is also the information about the levels.

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?

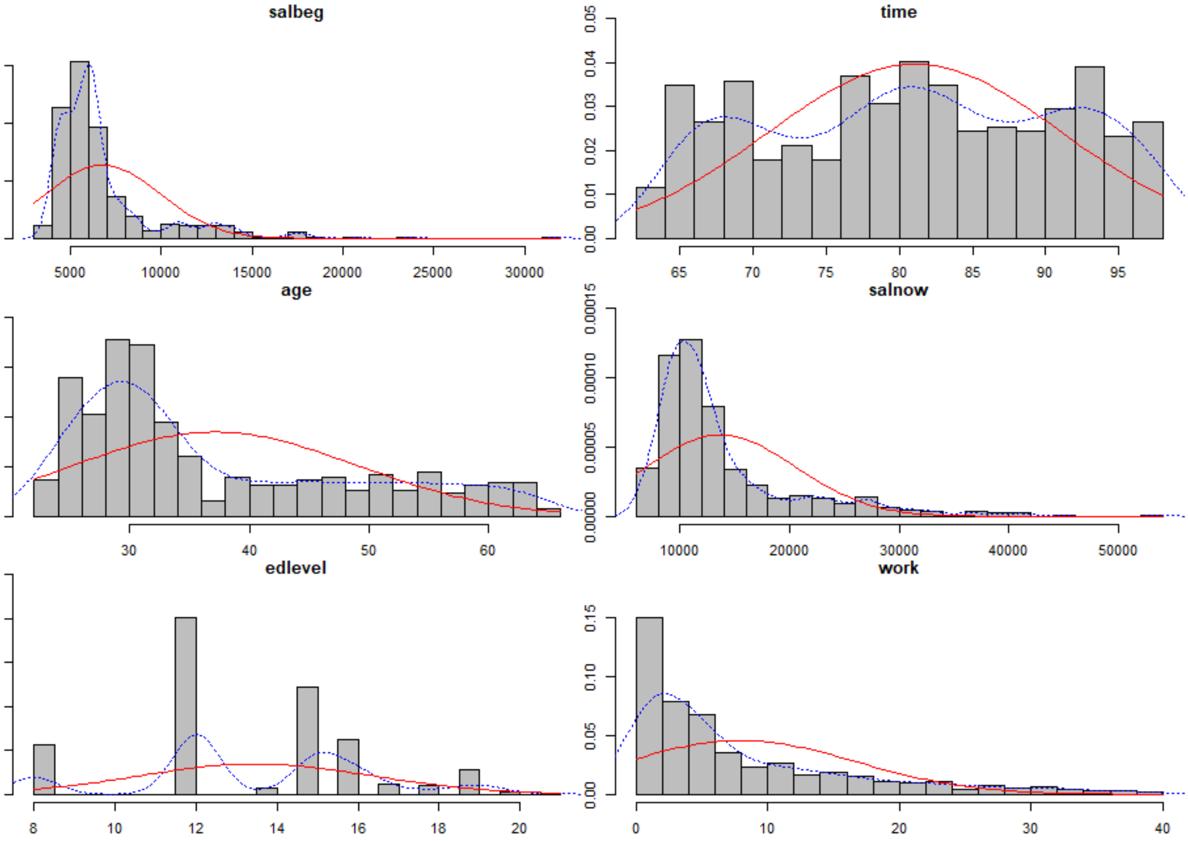
Summary statistics about the **numerical** variables.

```
$salbeg
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   3600    4995    6000    6806    6996   31992 

$time
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  63.00   72.00   81.00   81.11   90.00   98.00 

$age
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  23.00   28.50   32.00   37.19   45.98   64.50 

$salnow
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   6300    9600   11550   13768   14775   54000 

$edlevel
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   8.00   12.00   12.00   13.49   15.00   21.00 

$work
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  0.000   1.603   4.580   7.989  11.560  39.670 
```

The numeric variables are providing info about the beginning salary , time, age, current salary , education level and work status.

The summary function, provides this information about the quantiles, the min, max, median and mean about the data. From this, we can have a first taste for the distribution of each variable.

Histograms for the numerical variables in order to have a first look of their fit.

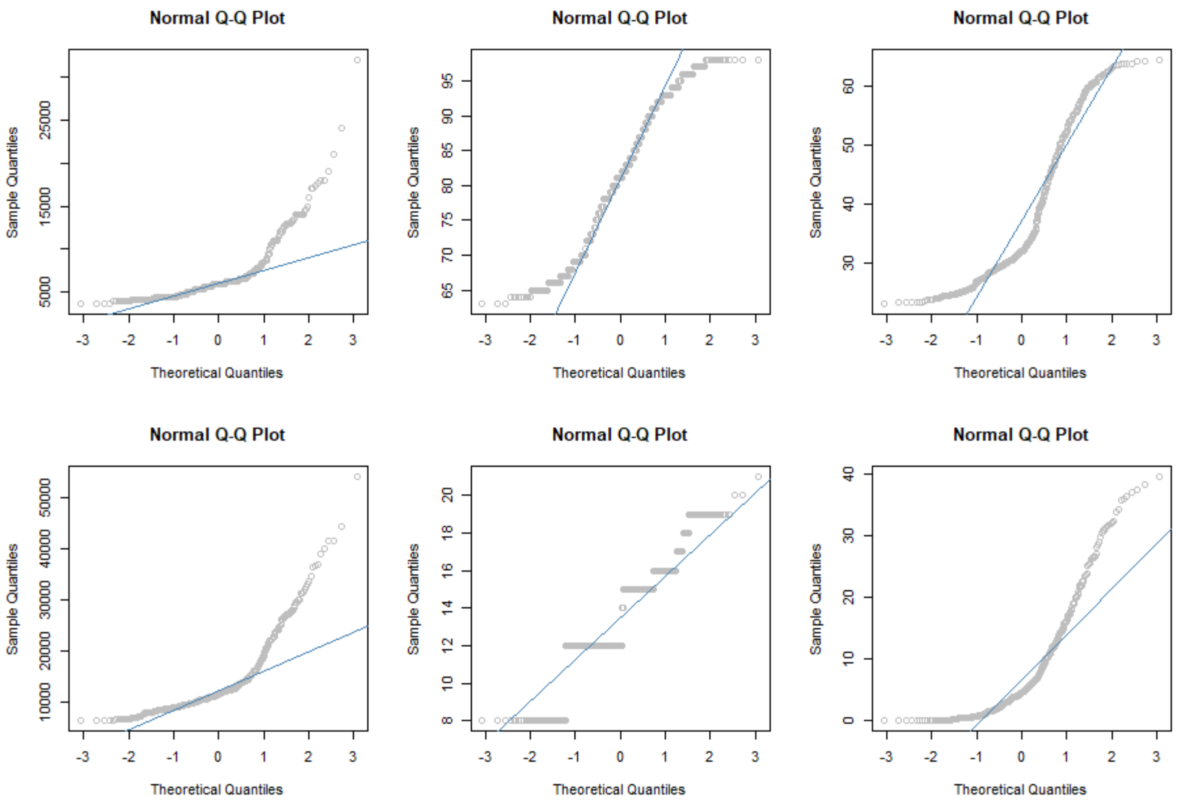


The above diagrams were made with the use of the library "psych" and the function multi.hist.

The blued dotted line shows the probability density line of the distribution. The red line indicates the normal fit.

From the first look we can assume, that no numeric variable of the data frame has a normal distribution but we will do further investigation about this.

QQ – plots for the numerical variables in order to have a visual investigation of the normality assumption.



A QQ plot is a comparison plot of the quantiles of two distributions.

The comparison is between the observed quantiles of the data (depicted as dots/circles) with the quantiles that we would except if the data were normally distributed
For the normal distribution we depict this with a line passing from the 1st and 3rd quartiles of a theoretically assumed normal distribution.
If the distribution of the observed data is “close” to the line we can assume a normal distribution but we have to investigate this further.
From the above QQ-plots we can assume again that the distribution of each numerical variable is not normal.

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.

Since the question is referring to the typical employee, we have to check if the mean of the variable salbeg is equal to 1000 dollars.

For choosing the correct test for one sample, we have to check several assumptions in the process.

First, we check normality for the salbeg variable, we will do further tests than the above visualizations with histograms and qq plots.
We use both Shapiro-Wilks test and Kolmogorov-Smirnov (because sample size is above 50).

SW – test

```
    Shapiro-Wilk normality test


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

p-value < 0.05, so we reject the null hypothesis and we can assume that the sample did not came for a normally distributed population.

Kolmogorov-Smirnov normality test

```
    Lilliefors (Kolmogorov-Smirnov) normality test


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

p-value < 0.05, so we reject the null hypothesis and we can assume that the sample did not came for a normally distributed population.

Since we assume no normality for the data and our sample size is above 50 observations, we shall check if the mean is a sufficient descriptive measure for the central location.

This can be estimated with the help of the lawstat library and the function symmetry.test

Symmetry - test

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


data:  sal_num$salbeg
Test statistic = 10.18, p-value < 2.2e-16
alternative hypothesis: the distribution is asymmetric.
sample estimates:
bootstrap optimal m
                 91
```

p-value for the symmetry test is below 0.05 so we reject the null hypothesis and we can assume the alternative hypothesis that the distribution is asymmetric.

Since we assume that the distribution is asymmetric, we choose to test for the medians with **Wilcoxon test for one sample.**

Ho: M=1000
H1: M≠1000

```
    Wilcoxon signed-rank test with continuity correction


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

p-value is less than 0.05 so we reject the Ho and assume the alternative hypothesis that the median is not equal to 1000.

Finally, we can assume that the beginning salary of a typical employee , cannot be considererd equal to 1000 dollars in a confidence level of 95% .

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.

In this occasion, we have two samples. The 1st is the observations of the beginning salary and the 2nd the observations for the current salaries of the employees. The two samples are depended because the employees are the same.

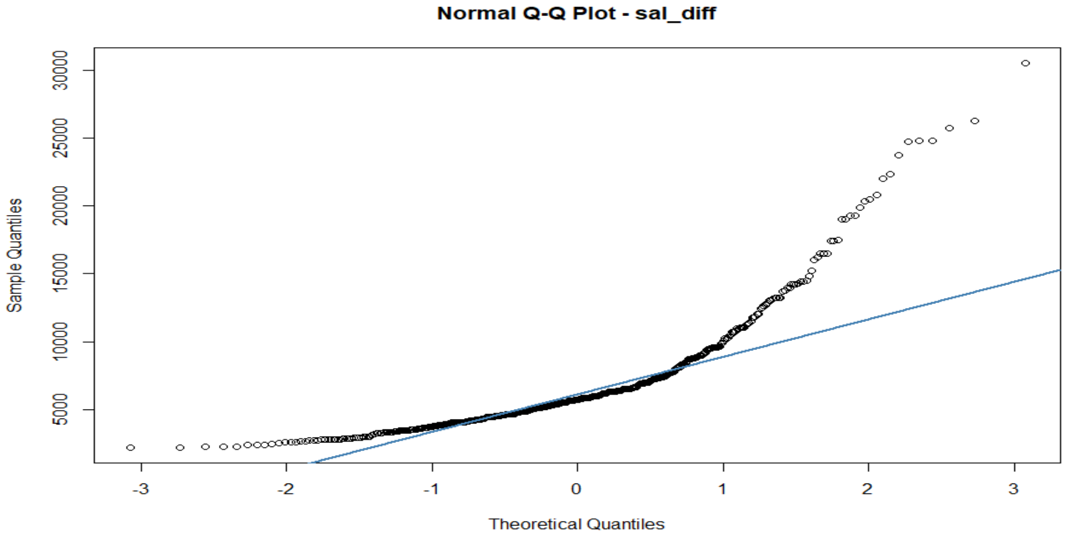
In order to check if there is a significant difference between these two, we have to construct a new variable "sal_diff "with the difference in the two samples in order to eliminate the correlation between them. We have to examine again one sample but we check now if the mean for the difference is zero or not.

Can we assume normality for "sal_diff" data?
We use both Shapiro-Wilks test and Kolmogorov-Smirnov (because sample size for sal_diff is above 50).

```
   Shapiro-Wilk normality test


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


    Lilliefors (Kolmogorov-Smirnov) normality test


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

In both tests p-value is below 0.05 so we can reject the null hypothesis that the data came from a normal distributed population.

In addition, we implement a QQ-plot in order to visualize the assumption.



The dots/circles are far from the theoretical normal distributed line.

Furthermore, since normality is rejected and the sample size is bigger than 50 observations, we have to do a symmetry test in order to check if the mean is a sufficient measure for the central location.

Symmetry test

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

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

p-value for the symmetry test is less than 0.05, so we reject the symmetry assumption.

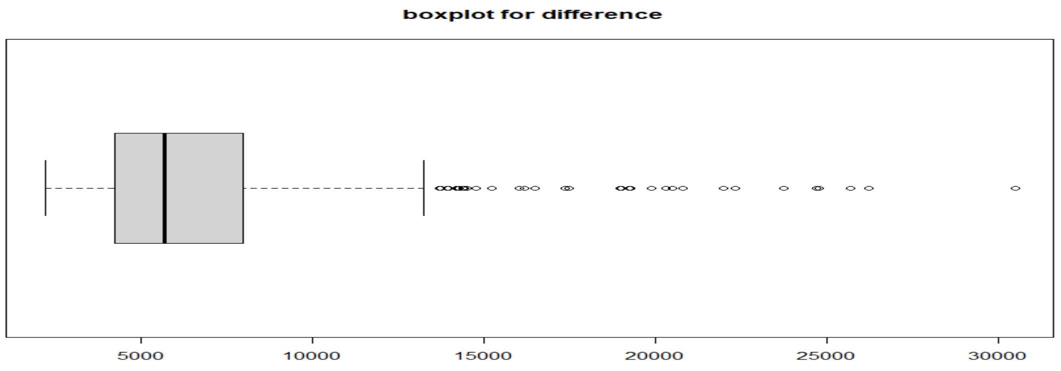
Since the symmetry assumption is rejected too, we will use the Wilcoxon test for dependent samples in order to check zero median difference.

$H0: M\Delta = 0$
$H1: M\Delta \neq 0$

```
Wilcoxon signed rank test with continuity correction


data:  sal_diff
V = 112575, p-value < 2.2e-16
alternative hypothesis: true location is not equal to 0
```

p-value is less than 0.05, so we reject the null hypothesis and we assume the alternative hypothesis that the medians are different.

Since we reject Ho, we visualize the result with a boxplot for the sal_diff variable.



Finally, we conclude from the above that there is a statistically significant difference between the beginning salary and the current salary in a confidence level of 95% because the test shows that the median of their difference is not equal to zero.

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.

For the question above we will implement a hypothesis test between the continuous variable salbeg and the categorical variable sex (specifically, sex is a binary variable since it has only two factors).

We are interested to test for differences on the values of the quantitative variable for the two groups (are the means or the medians equal?) This is the independent samples t-test. It examines the relationship between the binary and the numeric variable, since if the means-medians on average are the same then the state of the binary does influence the mean or median.

For this, we construct a new dataset in R with the salbeg variable in one column and the gender info at the 2nd .

First, we check normality of each group (with both tests SW – KS)

Shapiro – Wilk normality test

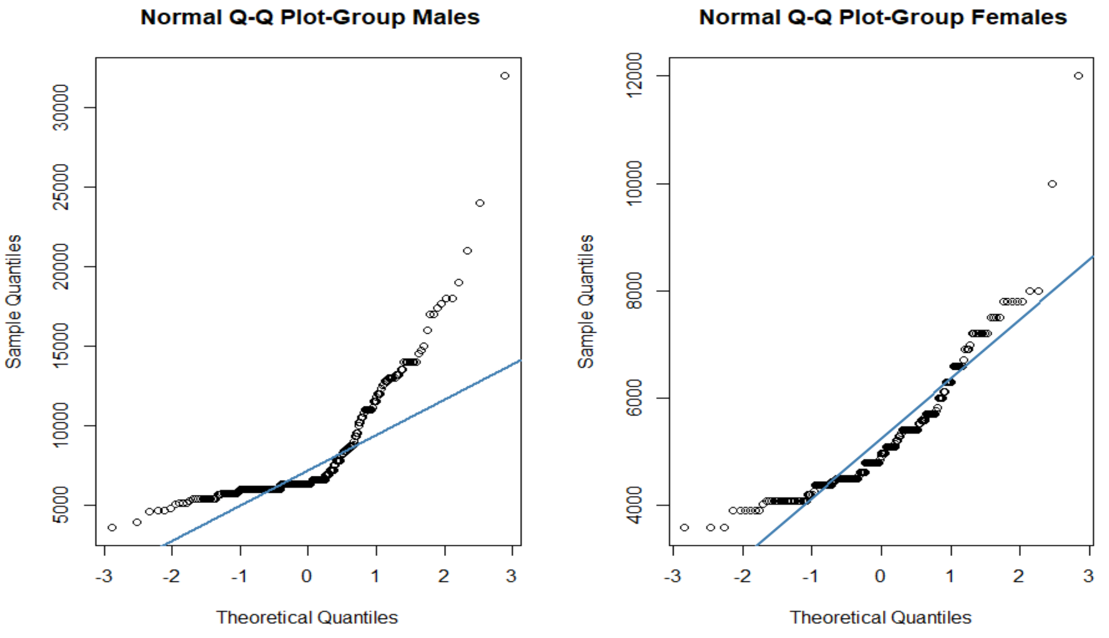
```
dataset5$gender: MALES

    Shapiro-Wilk normality test

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

--------------------------------------------------------------------------
------------
dataset5$gender: FEMALES

    Shapiro-Wilk normality test

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

Kolmogorov-Smirnov test - because (n1 & n2 >50)

```
dataset5$gender: MALES

    Lilliefors (Kolmogorov-Smirnov) normality test

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

--------------------------------------------------------------------------
------------
dataset5$gender: FEMALES

    Lilliefors (Kolmogorov-Smirnov) normality test

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

We reject the null hypothesis in both tests for each group thus we assume no normality.

QQ – plot for each group to have a visualization of normality



Then because, (n1 & n2 >50) we implement a symmetry test to check if the mean is a good measure for the central location.

Symmetry test

```
dataset5$gender: 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
                 31


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


dataset5$gender: 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
                 53
```

Both tests for symmetry fail for each group , p-value <0.05 thus we reject the null hypothesis and we assume the alternative with the asymmetric distributions.

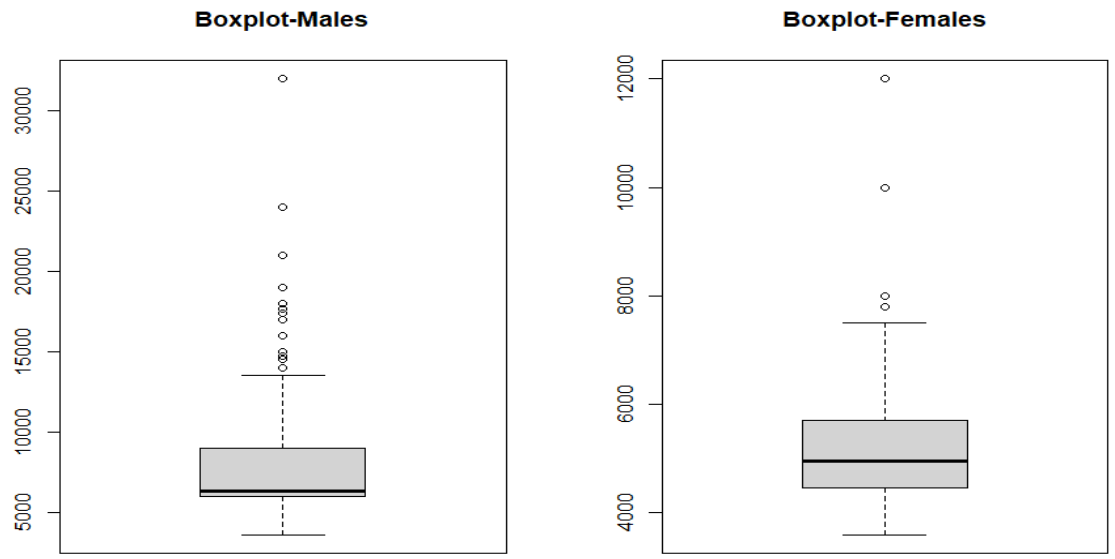
Since the means are not a good measure for the central location , we will implement Wilcoxon rank sum test with continuity correction for the medians of the two groups with the below hypothesis.

H0: M1=M2
H0: M1≠M2

```
Wilcoxon rank sum test with continuity correction


data:  group_females and group_males
W = 7854, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
```

p-value <00.05 so we reject the null hypothesis for equal medians and we assume there is a significant difference between the medians.

Since we rejected the null hypothesis of the above test, we visualize the result in order to have a better look of the means.



Finally, since the Wilcoxon rank sum test for the medians failed, we can assume that there is significance difference between the beginning salaries of males and females employees at a statistical significant level of 95%.

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 testthat you used by paying particular attention on the assumptions.

We should first, create the age_cut variable with the help of the Hmisc package. The age sample is splitted into three groups with the use of the cut2 function. The age groups are 23-30 , 30-40 and 40-64.50 respectively, thus their names “Younger” , “Middle” , “ Older” .

Dataframe “dataset6” is created with the sal_beg variable in the first column and the age-groups in the second.

We examine for a possible relationship between a quantitative variable and a categorical.

The hypothesis test is called ANOVA (analysis of variation)

We check for possible differences between the means of groups.

- We assume residuals normality, if the sample size is above 50.
- Equal variances

If the above assumptions are rejected, a non-parametric test will be used to check for differences in the medians of the groups. (Kruskal – Wallis test).

Results of ANOVA

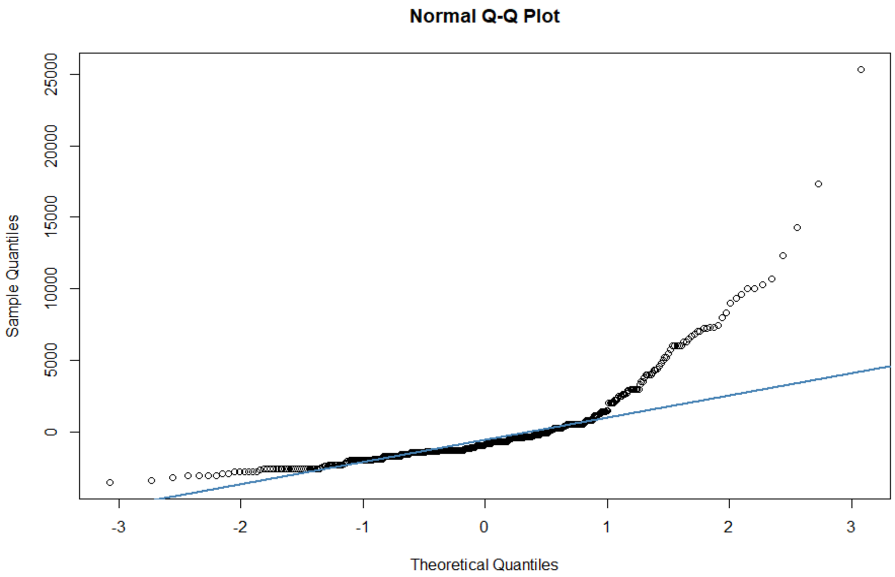
```
Call:
   aov(formula = salaries6 ~ age_cut, data = dataset6)

Terms:
                   age_cut  Residuals
Sum of Squares   396471437 4291673358
Deg. of Freedom          2        471

Residual standard error: 3018.581
Estimated effects may be unbalanced
```

Residuals normality – check

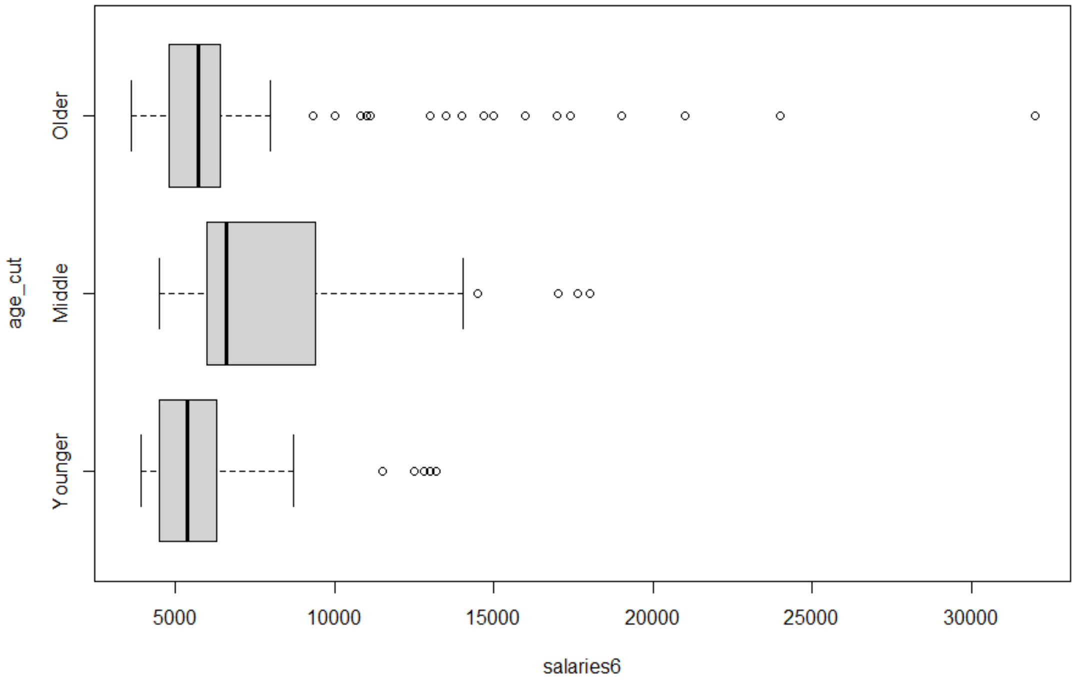
```
    Lilliefors (Kolmogorov-Smirnov) normality test

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
```

In both test we reject the null hypothesis (p-value < 0.05), so we assume no normality for the residuals.

QQ plot for the residuals



Furthermore, we check symmetry to see if the means are a good measure for central location .
From the boxplot below we can see that no age group distribution is symmetrical.



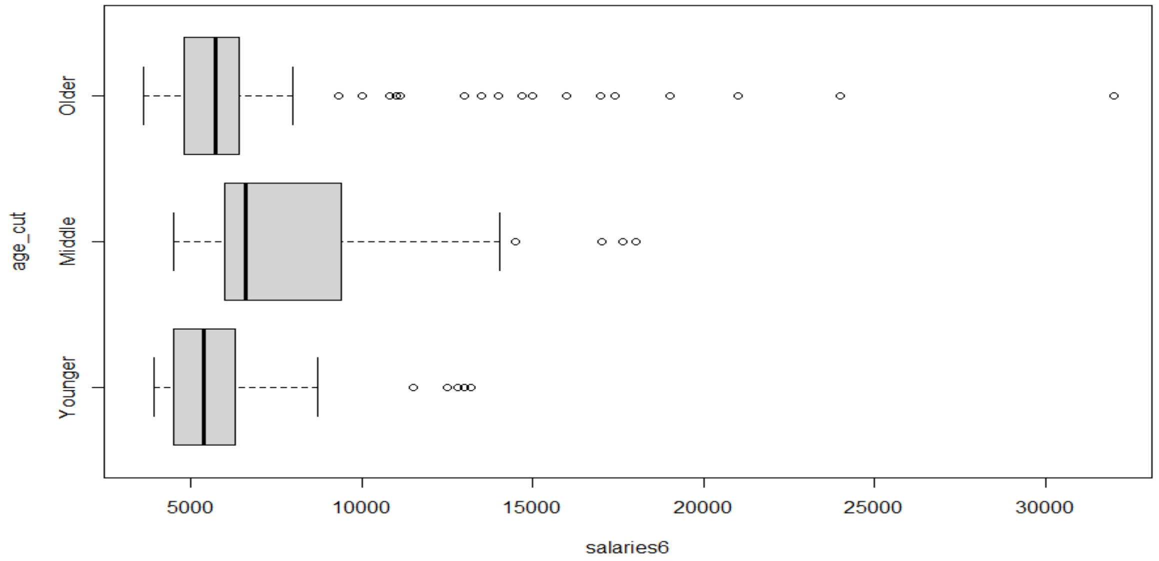
From the above, we assume that the correct test for this case is the non-parametric - Kruskal Wallis test for the medians.

H_0: $M_1=M_2=\ldots=M_K$
H_1: $M_k \neq M_j$ for some k ≠j in {1,2, … K}.

```
   Kruskal-Wallis rank sum test


data:  salaries6 by age_cut
Kruskal-Wallis chi-squared = 92.742, df = 2, p-value < 2.2e-16
```

p-value is less than 0.05 so we reject the null hypothesis of Kruskal Wallis test thus we assume that the age group medians are not significant equal but we do not know which group differs most.

Since we reject the null hypothesis, we implement the same boxplot in order to check the above result.



Finally, we can assume that the medians of the age groups are not equal so there is significant difference between them in a confidence level of 95%. From the boxplot we see that the median for the “Middle” aged group slightly differs from the medians of the other two groups.

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.

In this question we have to examine the equality of proportions in two independent groups, the white male employees and the white female employees. We first construct the proportions table (calculated by the frequencies table) with the cell – rows and columns percentages respectively.

```
        WHITE NONWHITE # cell percentages
MALES   40.93    13.50
FEMALES 37.13     8.44

        WHITE NONWHITE # row percentages
MALES   75.19    24.81
FEMALES 81.48    18.52

        WHITE NONWHITE # column percentages
MALES   52.43    61.54
FEMALES 47.57    38.46
```

Then we implement Pearson's Chi- Squared for independence.
Ho: whitemales = πfemales vs H1: πwhite-males ≠πfemales

```
   Pearson's Chi-squared test with Yates' continuity correction
data:  tab1
X-squared = 2.3592, df = 1, p-value = 0.1245

0.6894628
```

From the above test, we do not reject the null hypothesis for the equality of the proportions since p-value is greater than 0.05

We finally, assume there is no difference between the proportion of the white males group and the white females group

R – CODE

```
library(foreign)
library(psych)
library(nortest)
library(lawstat)
library(Hmisc)

### task 1
salary<- read.spss(file.choose(), to.data.frame = T)
str(salary)


#### task 2

index <- sapply(salary,class) == "numeric" # extract the numeric
sal_num <- salary[index] # get numeric data
sal_num$id <- NULL # remove the id column
lapply(sal_num, summary) # summaries


multi.hist(sal_num,dcol= c("blue","red"),dlty=c("dotted", "solid")) #
histograms with normal curves
y<-sal_num
p<-ncol(y)
par(mfrow=c(2,3)) ## qq plots
for (i in 1:p){
  qqnorm(y[,i])
  qqline(y[,i])
}


### task 3

par(mfrow=c(1,1))
qqnorm(sal_num$salbeg)
qqline(sal_num$salbeg, col="steelblue", lwd=2)
shapiro.test(sal_num$salbeg) # SW test
```

```
lillie.test(sal_num$salbeg) # KS test
symmetry.test(sal_num$salbeg) # symmetry test
wilcox.test(sal_num$salbeg, mu=1000)
boxplot(sal_num$salbeg, main = "Boxplot for sal_num$salbeg",
horizontal=TRUE)



### task 4

sal_diff <- sal_num$salnow - sal_num$salbeg #creation of difference
variable
qqnorm(sal_diff)
qqline(sal_diff, col="steelblue", lwd=2)
shapiro.test(sal_diff)
lillie.test(sal_diff)
symmetry.test(sal_diff)
wilcox.test(sal_diff, mu=0)
boxplot(sal_diff, main = " boxplot for difference", horizontal = TRUE )



### task 5

group_males <- salary$salbeg[which(salary$sex == "MALES")]
group_females <- salary$salbeg[which(salary$sex == "FEMALES")]
n1 <- length(group_males)
n2 <- length(group_females)
dataset5 <- data.frame(sal_beg = c(group_males, group_females),
                       gender = factor( rep(1:2, c(n1,n2)),
                       labels=c('MALES','FEMALES') ) )

par(mfrow=c(1,2))
qqnorm(dataset5$sal_beg[which(dataset5$gender== "MALES")], main = " Normal
Q-Q Plot-Group Males")
qqline(sal_num$salbeg[which(dataset5$gender== "MALES")], col="steelblue",
lwd=2)
qqnorm(dataset5$sal_beg[which(dataset5$gender== "FEMALES")], main = "
Normal Q-Q Plot-Group Females")
```

```
qqline(sal_num$salbeg[which(dataset5$gender== "FEMALES")],
col="steelblue", lwd=2)

by(dataset5$sal_beg, dataset5$gender, shapiro.test) #  normality test for
both groups
by(dataset5$sal_beg, dataset5$gender, lillie.test)  #  normality test for
both groups
by(dataset5$sal_beg, dataset5$gender, symmetry.test) # symmetry test for
both groups

wilcox.test(group_females,group_males)

boxplot(dataset5$sal_beg[which(dataset5$gender== "MALES")], main =
"Boxplot-Males", horizontal = FALSE)
boxplot(dataset5$sal_beg[which(dataset5$gender== "FEMALES")], main =
"Boxplot-Females", horizontal = FALSE)



### task 6

salaries6<- salary$salbeg
age_cut<- cut2(sal_num$age,g=3)
levels(age_cut)<-c("Younger","Middle","Older")
dataset6 <- data.frame(sal_beg = salaries6, Age_Group = age_cut )
#anova
anova1 <- aov(salaries6 ~ age_cut, data = dataset6)
names(anova1)
summary(anova1)
lillie.test(anova1$residuals) # KS test
shapiro.test(anova1$residuals) # SW test
par(mfrow=c(1,1))
qqnorm(anova1$residuals) ## qq plot
qqline(anova1$residuals, col="steelblue", lwd = 2)
boxplot(salaries6~age_cut, dataset6) # symmetry check
kruskal.test(salaries6~age_cut)
boxplot(salaries6~age_cut, dataset6, horizontal = TRUE)
```

```
# task 7

tab1 <- table(salary$sex,salary$minority)#frequency-table
round(100*prop.table(tab1),2)    # cell percentages
round(100*prop.table(tab1, 1),2) # row percentages
round(100*prop.table(tab1, 2),2) # column percentages
chisq.test(tab1) # Pearson - chi_squared test
```