# Assignment_1

*Burcin_Sarac_FT18*

This is an R Markdown document. I have to begin with setting my working directory to read data. And I also add the necessary packages at this stage.

```
setwd("E:/dersler/Statistics 1/assignment 3")
library(tidyverse)
library(foreign)
library(nortest)
library(Hmisc)
library(car)
library(gmodels)
```

# Q1

```
salary <- read.spss("salary.sav", to.data.frame = T)
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  "EMPLOYEE CODE" "BEGINNING SALARY" "SEX OF EMPLOYEE" "JOB SENIORITY" ...
##   ..- attr(*, "names")= chr  "id" "salbeg" "sex" "time" ...
##  - attr(*, "codepage")= int 1253
```

There are 474 observations and 11 variables listed in the dataset. $sex, $jobcat, $minority, $sexrace are categorical variables and their datatype are “Factor”. The remaining variables are numerical variables and their type are “num”.

# Q2

I dropped id column because it is duplicated with row names and also check if there is any NA variables in dataset, before getting summary statistics.

```
salary <- salary[,-1]
sum(which(is.na(salary)))
```

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

It seems there is not any missing data, so I continue with summary statistics;

```
summary(salary)
```

```
##      salbeg           sex           time            age       
##  Min.   : 3600   MALES  :258   Min.   :63.00   Min.   :23.00  
##  1st Qu.: 4995   FEMALES:216   1st Qu.:72.00   1st Qu.:28.50  
##  Median : 6000                 Median :81.00   Median :32.00  
##  Mean   : 6806                 Mean   :81.11   Mean   :37.19  
##  3rd Qu.: 6996                 3rd Qu.:90.00   3rd Qu.:45.98  
##  Max.   :31992                 Max.   :98.00   Max.   :64.50  
##                                                               
##      salnow         edlevel           work                     jobcat   
##  Min.   : 6300   Min.   : 8.00   Min.   : 0.000   CLERICAL        :227  
##  1st Qu.: 9600   1st Qu.:12.00   1st Qu.: 1.603   OFFICE TRAINEE  :136  
##  Median :11550   Median :12.00   Median : 4.580   SECURITY OFFICER: 27  
##  Mean   :13768   Mean   :13.49   Mean   : 7.989   COLLEGE TRAINEE : 41  
##  3rd Qu.:14775   3rd Qu.:15.00   3rd Qu.:11.560   EXEMPT EMPLOYEE : 32  
##  Max.   :54000   Max.   :21.00   Max.   :39.670   MBA TRAINEE     :  5  
##                                                   TECHNICAL       :  6  
##      minority               sexrace   
##  WHITE   :370   WHITE MALES     :194  
##  NONWHITE:104   MINORITY MALES  : 64  
##                 WHITE FEMALES   :176  
##                 MINORITY FEMALES: 40  
##                                       
##                                       
## 
```

From data summary, both max data values in begining salary and current salary might be an outlier. For checking this I try to find if it is same row or not.

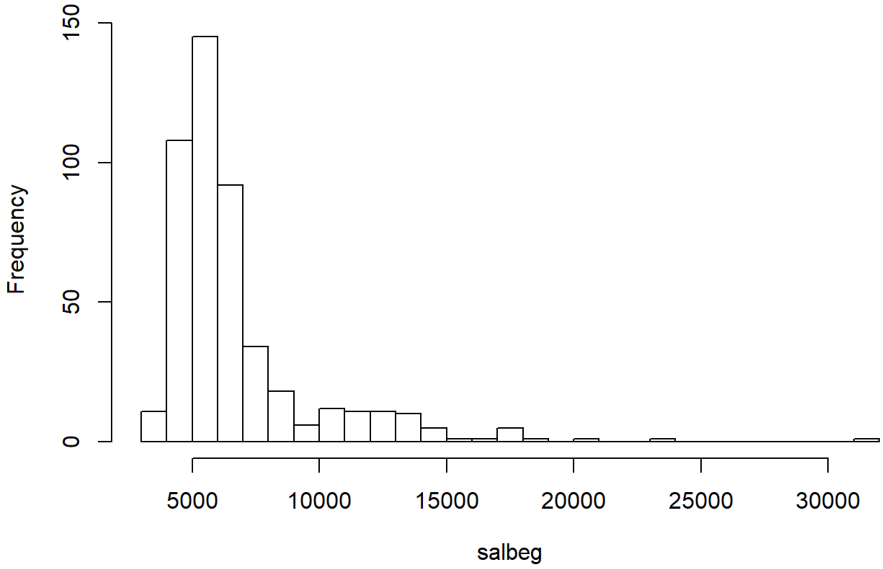
```
salary[which(salary$salbeg==max(salary$salbeg)),]
```

```
##    salbeg   sex time   age salnow edlevel  work    jobcat minority
## 56  31992 MALES   96 49.58  54000      19 16.58 TECHNICAL    WHITE
##        sexrace
## 56 WHITE MALES
```

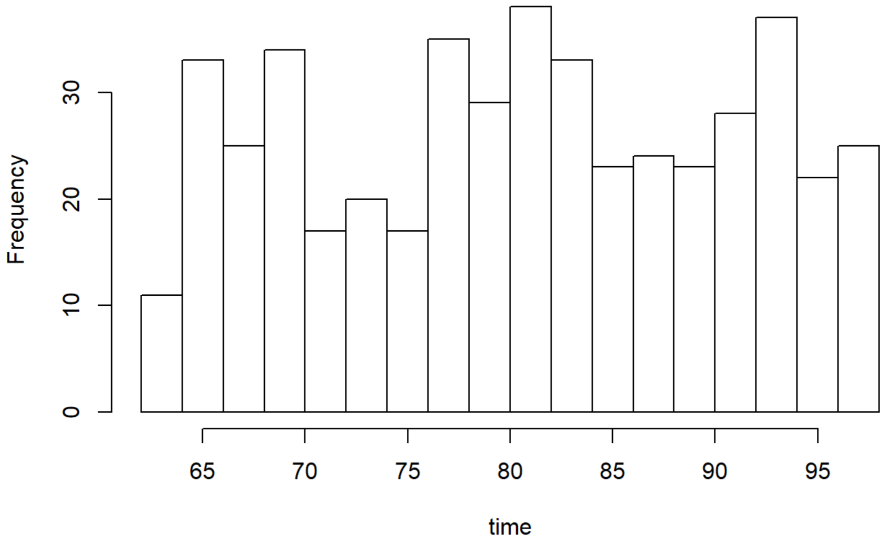
It seems correct, because this row both includes max beginning and current salary values.

After that, histograms drawn for all numerical variables below;

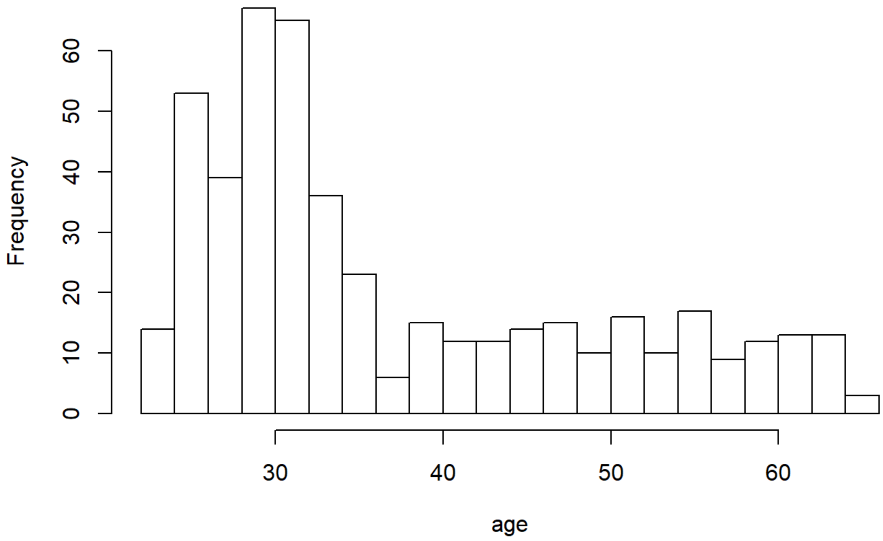
```
for(i in 1:ncol(salary)){
    if (class(salary[,i]) == "numeric"){
    hist(salary[i])
}}
```



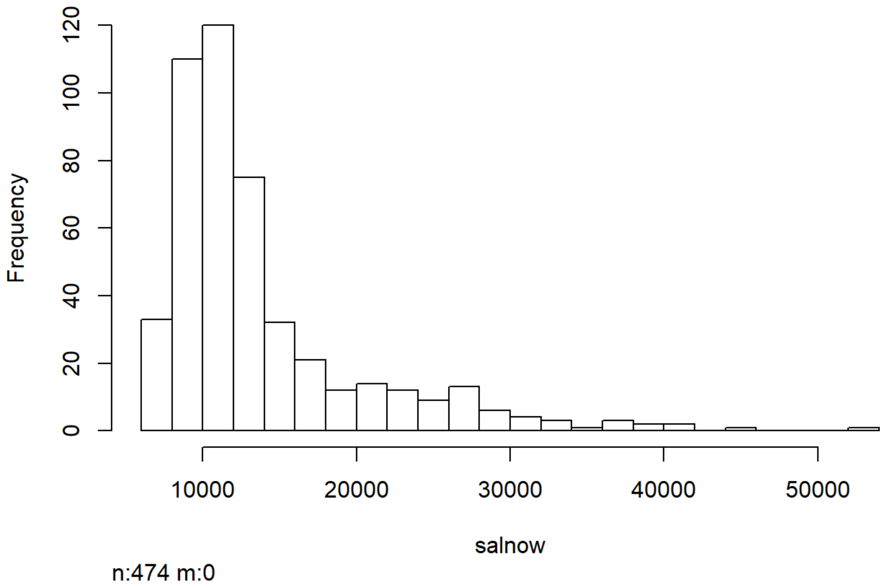
n:474 m:0

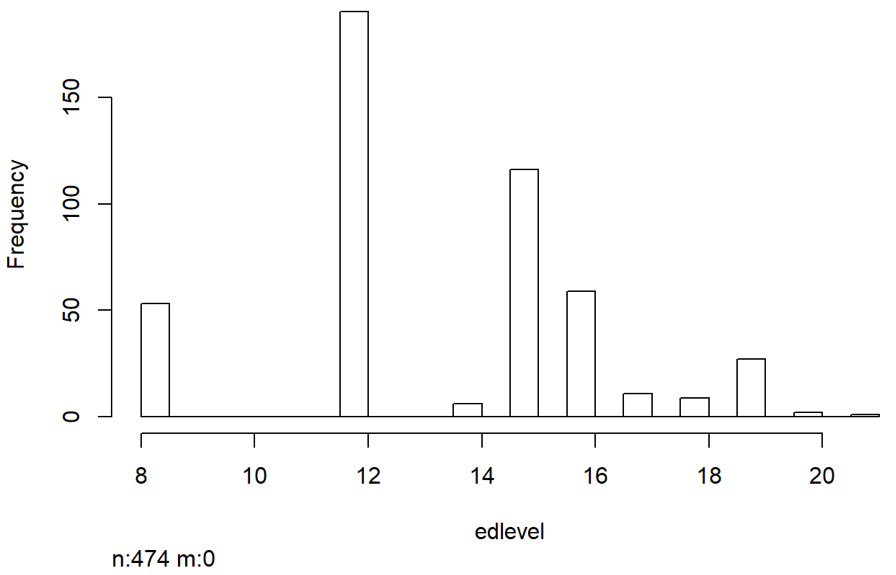


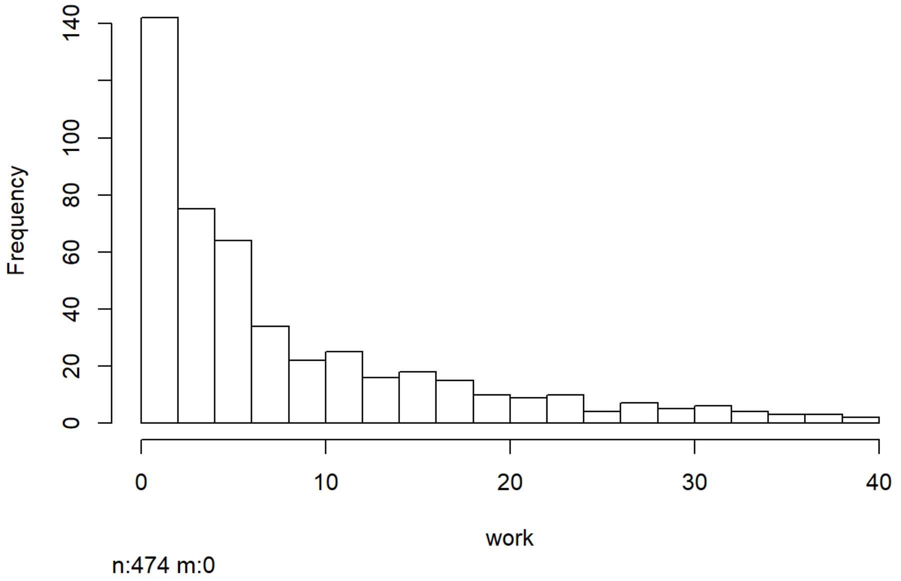
n:474 m:0



n:474 m:0


Frequency
0
20
40
60
80
100
120
10000
20000
30000
40000
50000
salnow
n:474 m:0


Frequency
0
50
100
150
8
10
12
14
16
18
20
edlevel
n:474 m:0


Frequency
0
20
40
60
80
100
140
0
10
20
30
40
work
n:474 m:0

It seen from summary of data and histograms, I can roughly say that, none of numerical variables distibuted normally.

# Q3

Firstly, for analyzing begining salary, I create a new data called “beginning” includes only beginning salary data, and test for normality with Liliefors(Kolmogorov-Smirnov) and Shapiro tests.

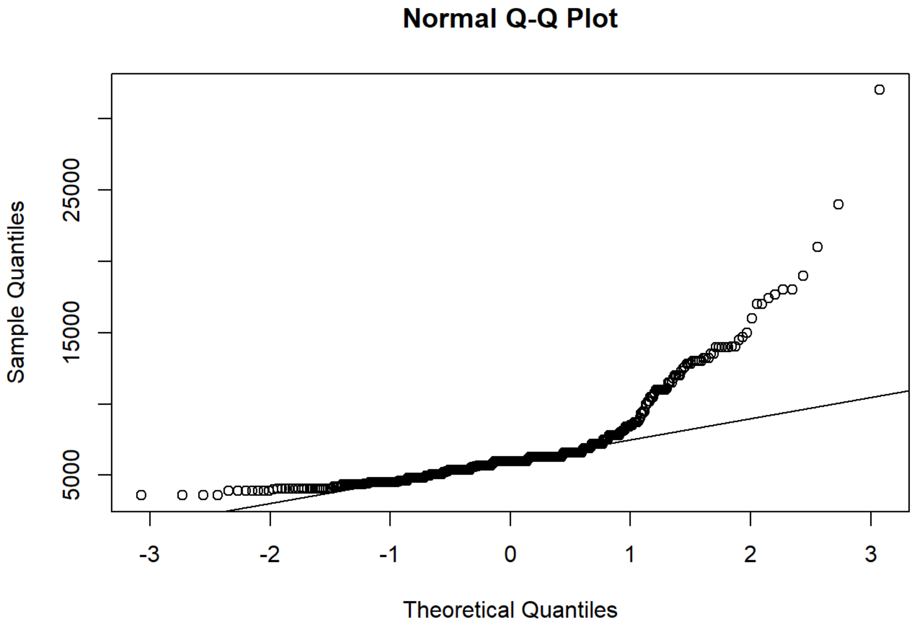
```
beginning <- salary$salbeg
lillie.test(beginning)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  beginning
## D = 0.25188, p-value < 2.2e-16
```

```
shapiro.test(beginning)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  beginning
## W = 0.71535, p-value < 2.2e-16
```

```
qqnorm(beginning)
qqline(beginning)
```



Both tests and QQ Plot shows that,

P-value is very small and it means I reject the null Hypothesis, which assumes data is normally distributed. So because of the beginning salary does not normally distributed,I need to check sample size and location of mean and median if they are close or not.

```
multifunctions <- function(x){
  c(length=length(x),mean=mean(x),median=median(x))
}
multifunctions(beginning)
```

```
##   length     mean   median
##  474.000 6806.435 6000.000
```

The sample size is larger than 50 and as long as this is a subjective decision to make, I assume the values of mean and median are closely located. So I will use one sample t-test for my continous variable although it does not normally distributed. This time my null hypothesis is mu=1000, and I will do two tailed test, which means my alternative hypothesis is mean!=1000

```
t.test(beginning, mu=1000)
```

```
## 
##  One Sample t-test
## 
## data:  beginning
## t = 40.154, df = 473, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 1000
## 95 percent confidence interval:
##  6522.289 7090.581
## sample estimates:
## mean of x 
##  6806.435
```

According to t-test results, P-Value is remarkably small, which means I reject my null hypothesis. With this result, I can say that “At 95% siginificance level beginning salary of a typical employee does not equal to 1000 dollars.”

# Q4

This time for testing difference between beginning salary and the current salary, I need to create a dataset only includes difference between them.

```
diff <- salary$salnow - salary$salbeg
```

And after this, according to set up a hypothesis for two dependent samples, I check normality with Liliefors(Kolmogorov-Smirnov) and Shapiro tests as before and also draw a QQ Plot for checking.

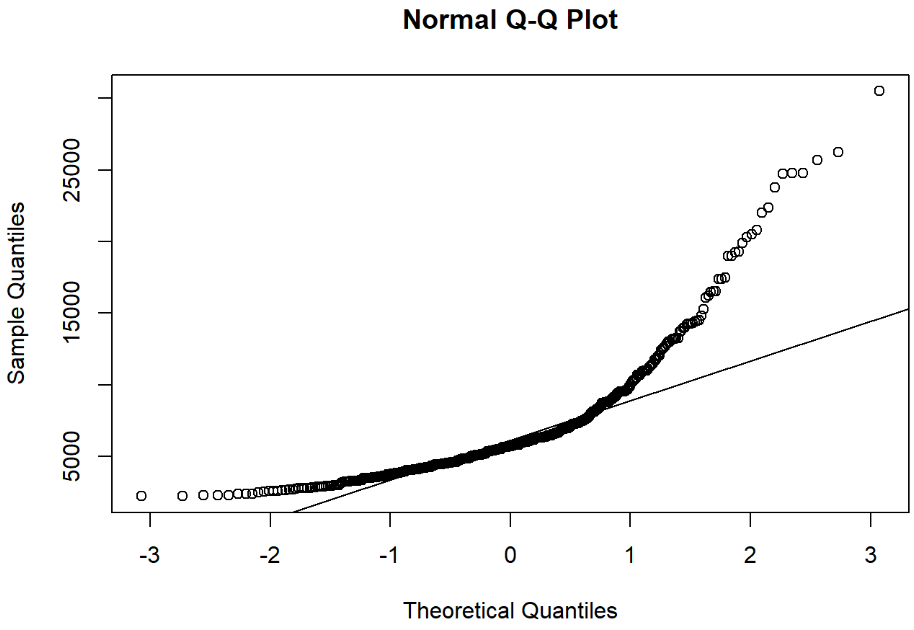
```
lillie.test(diff)
```

```
## 
##  Lilliefors (Kolmogorov-Smirnov) normality test
## 
## data:  diff
## D = 0.186, p-value < 2.2e-16
```

```
shapiro.test(diff)
```

```
## 
##  Shapiro-Wilk normality test
## 
## data:  diff
## W = 0.78168, p-value < 2.2e-16
```

```
qqnorm(diff)
qqline(diff)
```



It seem from normality tests that difference data is not normally distibuted. So I again check the sample size and location of mean and median;

```
multifunctions(diff)
```

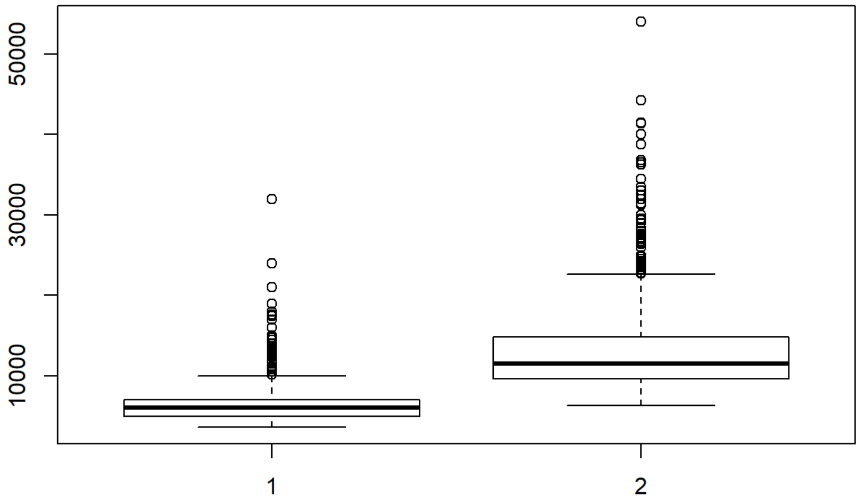
```
##   length     mean   median 
##  474.000 6961.392 5700.000
```

Different from before, their locations are not closed. So I will use Wilcoxon Test for my hypothesis. In this case due to checking any significant difference between beginning and current salaries, my null hypothesis is mu=0, means that there is not any significant difference between beginning and current salaries and alternative hypothesis is mu!=0.

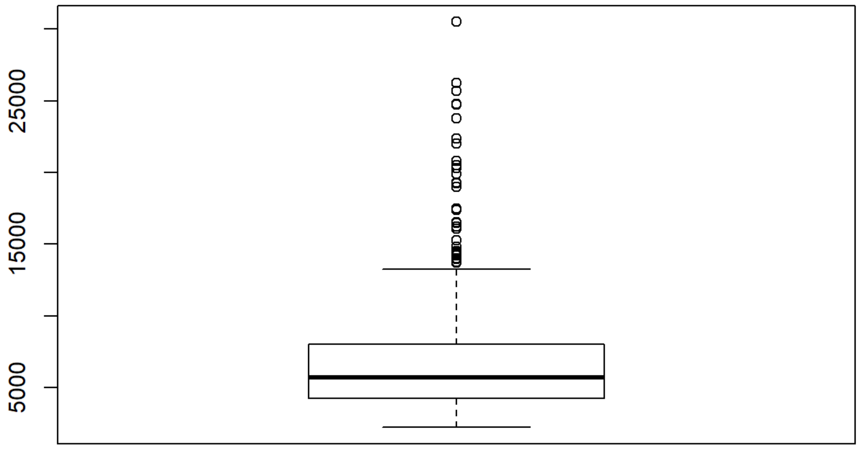
```
wilcox.test(diff, mu=0)
```

```
## 
##  Wilcoxon signed rank test with continuity correction
## 
## data:  diff
## V = 112580, p-value < 2.2e-16
## alternative hypothesis: true location is not equal to 0
```

```
boxplot(salary$salbeg, salary$salnow)
```



```
boxplot(diff)
```



According to Wilcoxon-test results, P-Value is remarkably small, which means I reject my null hypothesis.It can also be seen from boxplots that the beginning salary plot is positioned in a lower level than current salary plot and for the second boxplot, there is a significant difference can observed as well. With this result, I can say that “At 95% siginificance level there is a significant difference between the beginning salary and current salary.”

# Q5

In this question, it is needed to comparene categorical and one continous variable. Before testing our null hypothesis, which assumes there is not any differences between genders in terms of the beginning salary, I need to test normality in each sample. Additionally, I prefer to separate beginning salaries between males and females for easy interpretion.

```
males <- salary$salbeg[salary$sex=="MALES"]
females <- salary$salbeg[salary$sex=="FEMALES"]
by(salary$salbeg, salary$sex, lillie.test)
```

```
## salary$sex: MALES
## 
##  Lilliefors (Kolmogorov-Smirnov) normality test
## 
## data:  dd[x, ]
## D = 0.25863, p-value < 2.2e-16
## 
## -------------------------------------------------------- 
## salary$sex: FEMALES
## 
##  Lilliefors (Kolmogorov-Smirnov) normality test
## 
## data:  dd[x, ]
## D = 0.14843, p-value = 1.526e-12
```

```
by(salary$salbeg, salary$sex, shapiro.test)
```

```
## salary$sex: MALES
## 
##  Shapiro-Wilk normality test
## 
## data:  dd[x, ]
## W = 0.73058, p-value < 2.2e-16
## 
## -------------------------------------------------------- 
## salary$sex: FEMALES
## 
##  Shapiro-Wilk normality test
## 
## data:  dd[x, ]
## W = 0.85837, p-value = 2.98e-13
```

According to the normality tests, due to low p-value, I reject null hypothesis, which assumes both samples are normally distributed. After determine that the samples are not normally distibuted, I check sample size and mean if it is a sufficient measure for central location on both samples.

```
multifunctions(males)
```

```
##   length     mean   median 
##  258.000 8120.558 6300.000
```

```
multifunctions(females)
```

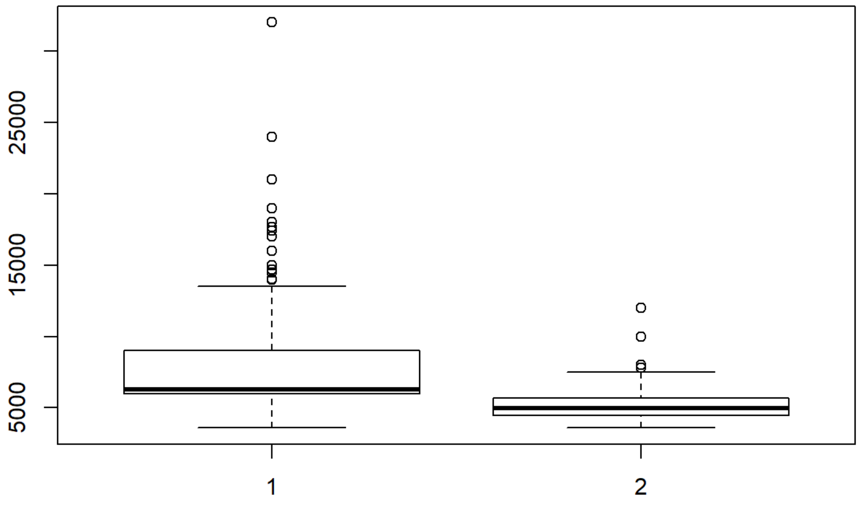
```
##   length     mean   median 
##  216.000 5236.787 4950.000
```

Samples are large enough, however difference between mean and median is large, so I will test for zero difference between medians through Wilcoxon Rank-sum test.

```
by(salary$salbeg, salary$sex, wilcox.test)
```

```
## salary$sex: MALES
## 
##  Wilcoxon signed rank test with continuity correction
## 
## data:  dd[x, ]
## V = 33411, p-value < 2.2e-16
## alternative hypothesis: true location is not equal to 0
## 
## -------------------------------------------------------- 
## salary$sex: FEMALES
## 
##  Wilcoxon signed rank test with continuity correction
## 
## data:  dd[x, ]
## V = 23436, p-value < 2.2e-16
## alternative hypothesis: true location is not equal to 0
```

```
boxplot(males, females)
```



Due to Wilcoxon test, p-value is far lower than significance level. So I reject null hypothesis, which assumes medians of beginning salary of males and beginning salary of females equal to zero. With this I can say that, “At 95% siginificance level there is a significant difference between the beginning salary between two genders.”

# Q6

For cutting age variable into 3 ranges and doing it equally, I use cut2 function and create a age_cut column to keep it. And this time I will compare beginning salary as one continous variable in terms of age_cut levels as 3 categorical variables, because of this I will use anova for testing normality.

```
salary$age_cut <- cut2(salary$age, g=3)
anova <- aov(salbeg~age_cut, salary)
summary(anova)
```

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

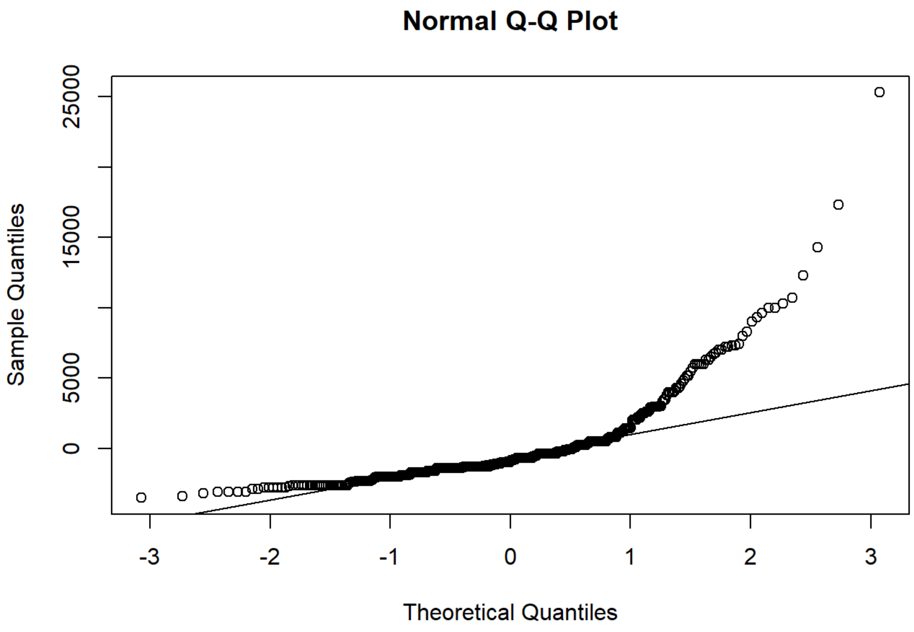
```
lillie.test(anova$residuals)
```

```
## 
##  Lilliefors (Kolmogorov-Smirnov) normality test
## 
## data:  anova$residuals
## D = 0.21891, p-value < 2.2e-16
```

```
shapiro.test(anova$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  anova$residuals
## W = 0.71244, p-value < 2.2e-16
```

```
qqnorm(anova$residuals)
qqline(anova$residuals)
```



According to p-values in normality tests, I reject null hypothesis, in other words the samples are not normally distributed. So I check size of samples and as I know they are big enough to continue testing, I will also check the locations of mean and median in all ranges.

```
young <- salary$age[salary$age_cut=="[23.0,29.7)"]
middle <- salary$age[salary$age_cut=="[29.7,39.8)"]
old <- salary$age[salary$age_cut=="[39.8,64.5]"]

multifunctions(young)
```

```
##    length      mean    median
## 160.00000  26.71631  27.04000
```

```
multifunctions(middle)
```

```
##    length      mean    median
## 156.00000  32.79603  32.04000
```

```
multifunctions(old)
```

```
##    length      mean    median
## 158.00000  52.12304  51.96000
```

It seems that mean and median values are close in both samples, so I will test if variances are equal or not with Bartlett test, Fligner-Killeen test and Levene’s test.

```
bartlett.test(salbeg~age_cut, salary)
```

```
##
##  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, salary)
```

```
##
##  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
```

```
leveneTest(salbeg~age_cut, salary)
```

```
## 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
```

Due to lower p-values than significance level, I again reject my null hypothesis, which assumes equal variances. So after determine that variances are not equal, I will continue with one-way test and Kruskal Wallis rank sum test.

```
oneway.test(salbeg~age_cut, salary, var.equal=FALSE)
```

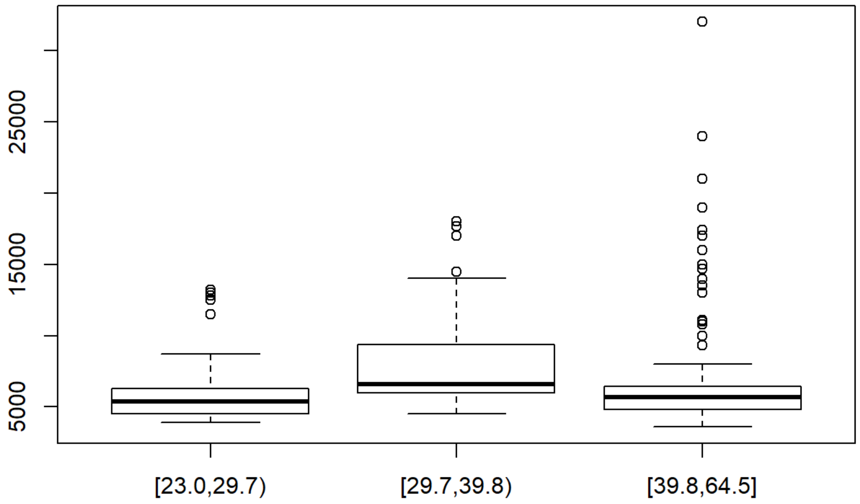
```
##
##  One-way analysis of means (not assuming equal variances)
##
## data:  salbeg and age_cut
## F = 32.752, num df = 2.00, denom df = 284.42, p-value = 1.582e-13
```

```
kruskal.test(salbeg~age_cut, salary)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  salbeg by age_cut
## Kruskal-Wallis chi-squared = 92.742, df = 2, p-value < 2.2e-16
```

The p-value again lower than significance level, which leads me to reject null hypothesis assumed equal means with unequal variances. So I continue with Pairwise comparison test.

```
boxplot(salary$salbeg~salary$age_cut)
```



```
pairwise.t.test(salary$salbeg, salary$age_cut, pool.sd = F)
```

```
##
##  Pairwise comparisons using t tests with non-pooled SD
##
## data:  salary$salbeg and salary$age_cut
##
##             [23.0,29.7) [29.7,39.8)
## [29.7,39.8) 8.8e-14     -
## [39.8,64.5] 0.0085      0.0017
##
## P value adjustment method: holm
```

Again Pairwise test shows that p-value is lower. So I reject null hypothesis, which assumes age level has not significant effect in beginning salary. So I can say that “At 95% siginificance level there is a significant difference between the beginning salary between three age groups, which age ranges are [23.0,29.7),[29.7,39.8) and [39.8,64.5].” Moreover due to pairwise comparison table, I can say that there is a significant difference between [29.7,39.8) and [23.0,29.7) age ranges with a very low p-value, but there is less significant difference between ages [23.0,29.7) and [39.8,64.5].

# Q7

In this question I will test if proportions are equal with “white” and “males” data to “white” and “females” data in sex and minority columns accordingly. For testing this two categorical variables, I will first check the number of samples via contingency and probability tables.

```
(tab <- table(salary$sex, salary$minority))
```

```
##
##           WHITE NONWHITE
##   MALES     194       64
##   FEMALES   176       40
```

```
prop.table(tab)
```

```
##
##                WHITE   NONWHITE
##   MALES   0.40928270 0.13502110
##   FEMALES 0.37130802 0.08438819
```

```
prop.table(tab,1)
```

```
##
##               WHITE  NONWHITE
##   MALES   0.7519380 0.2480620
##   FEMALES 0.8148148 0.1851852
```

```
prop.table(tab,2)
```

```
##
##               WHITE  NONWHITE
##   MALES   0.5243243 0.6153846
##   FEMALES 0.4756757 0.3846154
```

After checking data via tables, it seems that the probabilities are close and for the first thought assumption in the null hypothesis might be true, but I will decide it at later stages of analysis. Now I will use prop.test to implement the Pearson’s chi-square statistics for independence and also check the Pearson’s Chi-squared test with Yates continuity correction via chisq.test().

```
prop.test(tab)
```

```
##
##  2-sample test for equality of proportions with continuity
##  correction
##
## data:  tab
## 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(tab)
```

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

It seems from both tests that p-value is larger than significance level(0.05), so I cannot reject null hypothesis. However, although R did not give any warning and all expected values seems larger than 5 in Chi-squared test, because of Chi-squared test is an aproximation test, I would like to do Fisher test as well and I will also prepare a crosstable of tests.

```
fisher.test(tab)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  tab
## 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
```

```
CrossTable(tab, digits = 1, format = "SPSS", prop.r = T,
           chisq = T, fisher = T)
```

```
## 
##    Cell Contents
## |-------------------------|
## |                   Count |
## | Chi-square contribution |
## |             Row Percent |
## |          Column Percent |
## |           Total Percent |
## |-------------------------|
## 
## Total Observations in Table:  474 
## 
##              | 
##              |    WHITE  | NONWHITE  | Row Total | 
## -------------|-----------|-----------|-----------|
##        MALES |      194  |       64  |      258  | 
##              |      0.3  |      1.0  |           | 
##              |     75.2% |     24.8% |     54.4% | 
##              |     52.4% |     61.5% |           | 
##              |     40.9% |     13.5% |           | 
## -------------|-----------|-----------|-----------|
##      FEMALES |      176  |       40  |      216  | 
##              |      0.3  |      1.2  |           | 
##              |     81.5% |     18.5% |     45.6% | 
##              |     47.6% |     38.5% |           | 
##              |     37.1% |      8.4% |           | 
## -------------|-----------|-----------|-----------|
## Column Total |      370  |      104  |      474  | 
##              |     78.1% |     21.9% |           | 
## -------------|-----------|-----------|-----------|
## 
##  
## Statistics for All Table Factors
## 
## 
## Pearson's Chi-squared test 
## ------------------------------------------------------------
## Chi^2 =  2.713926     d.f. =  1     p =  0.0994759 
## 
## Pearson's Chi-squared test with Yates' continuity correction 
## ------------------------------------------------------------
## Chi^2 =  2.359218     d.f. =  1     p =  0.1245446 
## 
##  
## Fisher's Exact Test for Count Data
## ------------------------------------------------------------
## Sample estimate odds ratio:  0.6894628 
## 
## Alternative hypothesis: true odds ratio is not equal to 1
## p =  0.1186169 
## 95% confidence interval:  0.429148 1.098149 
## 
## Alternative hypothesis: true odds ratio is less than 1
## p =  0.06183135 
## 95% confidence interval:  0 1.023731 
## 
## Alternative hypothesis: true odds ratio is greater than 1
## p =  0.9611881 
## 95% confidence interval:  0.4617367 Inf 
## 
## 
##  
##        Minimum expected frequency: 47.39241 
```

All tests supports the decision via p-values which is about, not to reject null hypothesis.In other words, in 95% significance level I assume that the gender and minority are independent variables, which means gender does not effect minority of employee and the proportions of “White males” and “white females” data are equal.