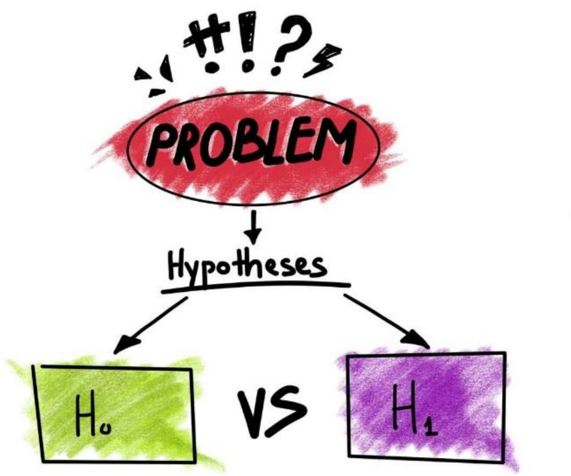


Assignment 1

FOR LAB ON HYPOTHESIS TESTING

Panagiotis Vaidomarkakis | Statistics for Business Analytics I (Part Time) | 22.11.2022

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

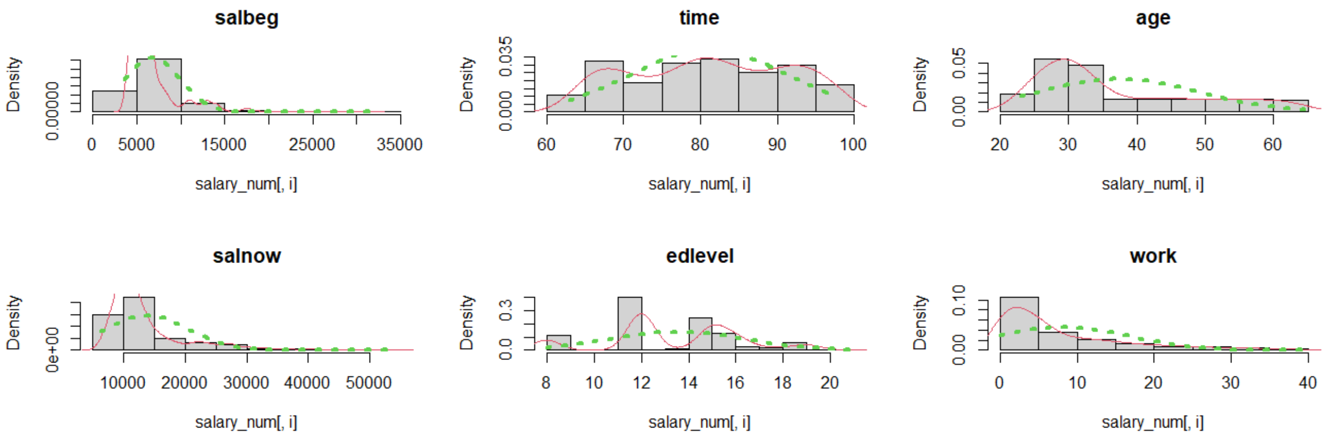
```
> str(salary)
'data.frame':	474 obs. of  11 variables:
 $ id      : num  1 2 3 4 5 6 7 8 9 10 ...
 $ salbeg  : num  8400 24000 10200 8700 17400 ...
 $ sex     : Factor w/ 2 levels "MALES","FEMALES": 1 1 1 1 1 1 1 1 1 1 ...
 $ time    : num  81 73 83 93 83 80 79 67 96 77 ...
 $ age     : num  28.5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

After reading the dataset “salary.sav” as a data frame, using the str() function we can see that it has 11 variables, numerical variables and 4 factor variables and all the observations are 474. Moreover, we have variable labels which help us understand the data. Furthermore, we have ids of the employees, the beginning salary, the gender, the job seniority, the age of the employees, the current salary, educational level of the employees, their working experience, their working position, minority classification and gender and race classification. Only gender and the last three variables are categorical variables which are factors in our data frame. Age and working experience are decimal variables and ids, beginning salary, seniority, current salary and education level are integer variables.

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?

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

After creating a new data frame in order to have only numeric values, we get a summary statistics and we’ve made histograms in order to see normally distributed attributes as you can see below:



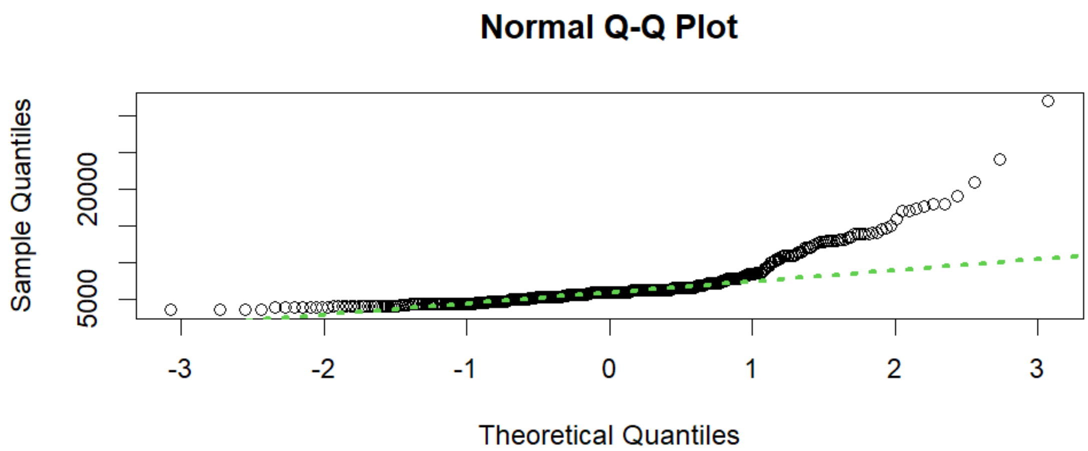
As we can see from the chart, none of the above is normally distributed variable.

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.

In order to assume normality at beginning salary, we need to test both Kolmogorov-Smirnov test and Shapiro-Wilk test.

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

        Shapiro-Wilk normality test

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

> lillie.test(salary_num$salbeg)

        Lilliefors (Kolmogorov-Smirnov) normality test

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

> ks.test(salary_num$salbeg, 'pnorm')

        Asymptotic one-sample Kolmogorov-Smirnov test

data:  salary_num$salbeg
D = 1, p-value < 2.2e-16
alternative hypothesis: two-sided
```

In all 3 tests, p-value is much less than 0.05 so we reject the H_o hypothesis that the beginning salary of all employees are normally distributed.



Q-Q Plot doesn't help us with the normality either. Our sample is big enough (n>50) so we need to see at least if the mean is sufficient descriptive measure for central location.

```
> mean(salary_num$salbeg)
[1] 6806.435
> median(salary_num$salbeg)
[1] 6000
> symmetry.test(salary_num$salbeg)

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

data:  salary_num$salbeg
Test statistic = 10.18, p-value < 2.2e-16
alternative hypothesis: the distribution is asymmetric.
sample estimates:
bootstrap optimal m
                 72
```

As we can see, both the comparison of mean and median as far as the symmetry test didn't help us to assume that mean is sufficient descriptive measure(if the symmetry test failed then the mean and the median are far away from each other). So, we will use Wilcoxon test for one sample in order to test the medians.

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

        Wilcoxon signed rank test with continuity correction

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

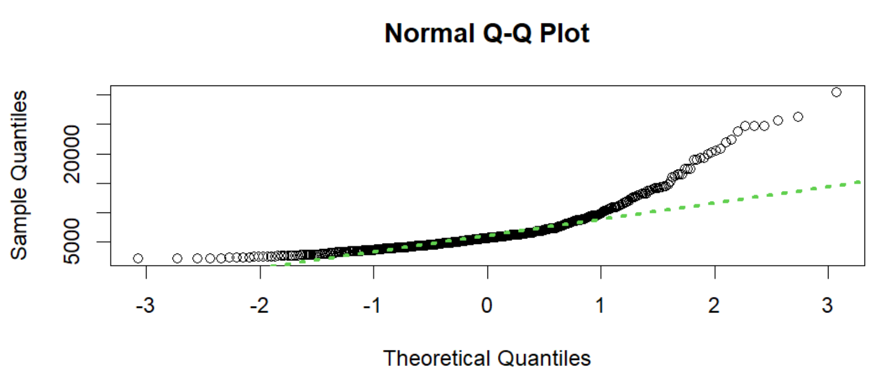
So, we reject the H_o hypothesis that the beginning salary for a typical emploee is equal to 1000$.

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.

In order to assume normality at the difference between the current salary and the beginning salary, we need to test both Kolmogorov-Smirnov test and Shapiro-Wilk test.

```
> shapiro.test(salary_num$saldiff)

        Shapiro-Wilk normality test

data:  salary_num$saldiff
W = 0.78168, p-value < 2.2e-16

> lillie.test(salary_num$saldiff)

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  salary_num$saldiff
D = 0.186, p-value < 2.2e-16

> ks.test(salary_num$saldiff,'pnorm')

        Asymptotic one-sample Kolmogorov-Smirnov test

data:  salary_num$saldiff
D = 1, p-value < 2.2e-16
alternative hypothesis: two-sided
```

In all 3 tests, p-value is much less than 0.05 so we reject the H_0 hypothesis that the salary difference at all employees is normally distributed.



Q-Q Plot doesn't help us with the normality either. Our sample is big enough (n>50) so we need to see at least if the mean is sufficient descriptive measure for central location.

```
> mean(salary_num$saldiff)
[1] 6961.392
> median(salary_num$saldiff)
[1] 5700
> symmetry.test(salary_num$saldiff)

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

data:  salary_num$saldiff
Test statistic = 10.536, p-value < 2.2e-16
alternative hypothesis: the distribution is asymmetric.
sample estimates:
bootstrap optimal m
                263
```

As we can see, both the comparison of mean and median as far as the symmetry test didn't help us to assume that mean is sufficient descriptive measure. So, we will use Wilcoxon test for one sample in order to test the medians.

```
> wilcox.test(salary_num$saldiff, mu=0)

        Wilcoxon signed rank test with continuity correction

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

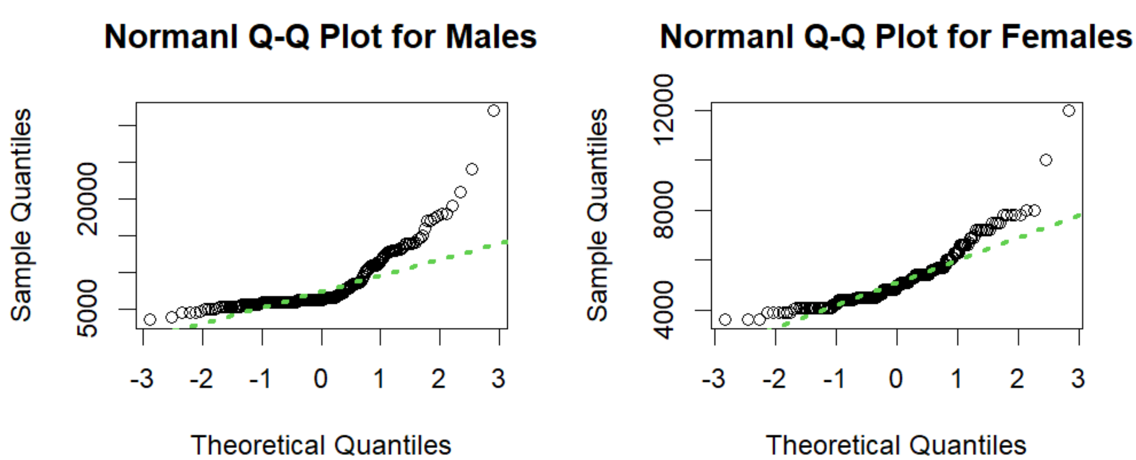
So, we reject the H_0 hypothesis and we say that there is a difference between the beginning salary and the current salary in every emploee.

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.

After creating a new data frame in order to keep only gender and beginning salary, we need to test both Kolmogorov-Smirnov test and Shapiro-Wilk test for every gendet(MALES and FEMALES).

```
> by( salary_gen$salbeg, salary_gen$sex, lillie.test)
salary_gen$sex: MALES

        Lilliefors (Kolmogorov-Smirnov) normality test

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

--------------------------------------------------
salary_gen$sex: FEMALES

        Lilliefors (Kolmogorov-Smirnov) normality test

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

> by( salary_gen$salbeg, salary_gen$sex, shapiro.test)
salary_gen$sex: MALES

        Shapiro-Wilk normality test

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

--------------------------------------------------
salary_gen$sex: FEMALES

        Shapiro-Wilk normality test

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

In 2 tests, p-value is much less than 0.05 so we reject the H_0 hypothesis that the beginning salary of males and the beginning salary of females at all employees is normally distributed.



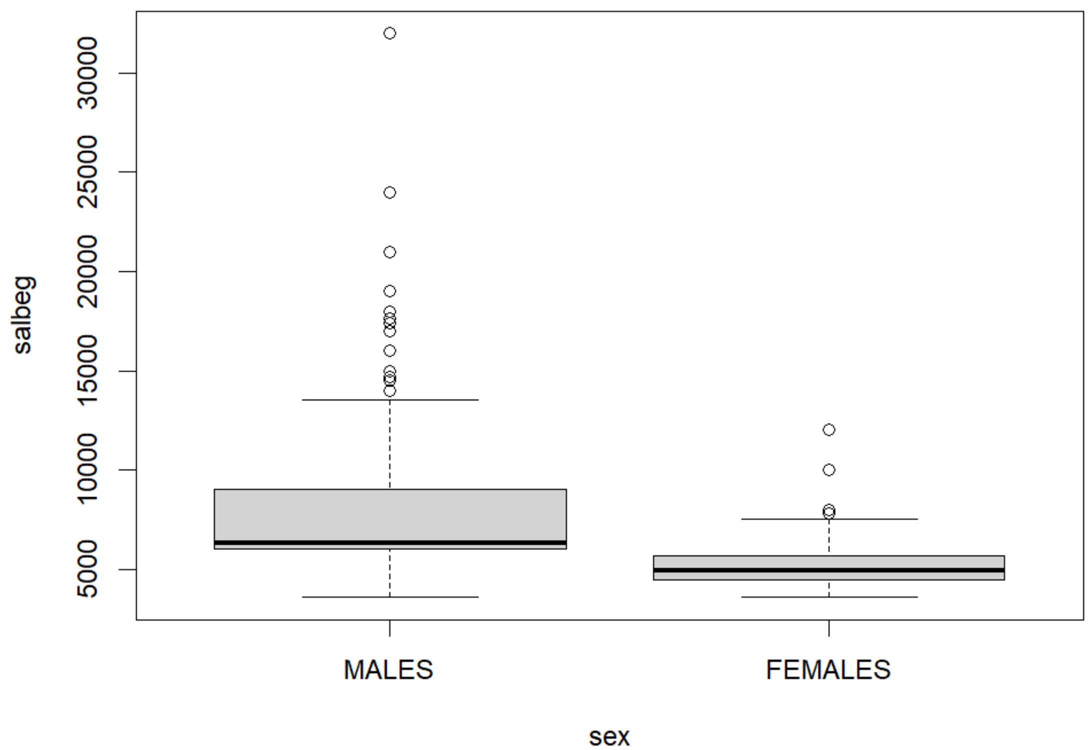
Q-Q Plot doesn't help us with the normality of the males and the females either. Our sample is big enough (n>50) so we need to see at least if the mean is sufficient descriptive measure for central location.

```
> mean(salary_gen$salbeg[which(salary_gen$sex=='MALES')])
[1] 8120.558
> median(salary_gen$salbeg[which(salary_gen$sex=='MALES')])
[1] 6300
> symmetry.test(salary_gen$salbeg[which(salary_gen$sex=='MALES')])

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

data:  salary_gen$salbeg[which(salary_gen$sex == "MALES")]
Test statistic = 13.829, p-value < 2.2e-16
alternative hypothesis: the distribution is asymmetric.
sample estimates:
bootstrap optimal m
                 35

> mean(salary_gen$salbeg[which(salary_gen$sex=='FEMALES')])
[1] 5236.787
> median(salary_gen$salbeg[which(salary_gen$sex=='FEMALES')])
[1] 4950
> symmetry.test(salary_gen$salbeg[which(salary_gen$sex=='FEMALES')])

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

data:  salary_gen$salbeg[which(salary_gen$sex == "FEMALES")]
Test statistic = 5.2527, p-value < 2.2e-16
alternative hypothesis: the distribution is asymmetric.
sample estimates:
bootstrap optimal m
                 33
```

As we can see, both the comparison of mean and median in every gender as far as the symmetry test didn't help us to assume that mean is sufficient descriptive measure. So, we will use Wilcoxon test for one sample in order to test the medians.

```
> wilcox.test(salary_gen$salbeg[which(salary_gen$sex=='MALE
S')],salary_gen$salbeg[which(salary_gen$sex=='FEMALES')])

        Wilcoxon rank sum test with continuity correction

data:  salary_gen$salbeg[which(salary_gen$sex == "MALES")] and
 salary_gen$salbeg[which(salary_gen$sex == "FEMALES")]
W = 47874, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
```

So, we reject the H_0 hypothesis and we say that there is a difference between the beginning salary of males and females.



In this boxplot, we can see that in general, the lowest value in the boxplots is the same so in the low paid employees, there isn't any difference at any gender. The majority of male employees are starting at a higher salary that the majority of the female employees. Moreover, as we can assume from the boxplot, there are more outliers at males than females which means that more males tend to start at higher positions which have higher beginning salary, but that's only an assumption.

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.

After "cutting" our sample in three groups based on age, we do an ANOVA (ANalysis Of VAriance) testing in order to see if there is a difference in the beginning salary among the three groups based on age.

```
> age_cut <- cut2(salary_num$age,g=3)
> head(age_cut)
[1] [23.0,29.7) [39.8,64.5] [29.7,39.8) [29.7,39.8)
[5] [39.8,64.5] [23.0,29.7)
Levels: [23.0,29.7) [29.7,39.8) [39.8,64.5]
> salary_num <- data.frame(salary_num,age_cut)
>
> anova1 <- aov(salbeg~age_cut, data = salary_num)
> anova1
Call:
   aov(formula = salbeg ~ age_cut, data = salary_num)

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

Residual standard error: 3018.581
Estimated effects may be unbalanced
> summary(anova1)
             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
```

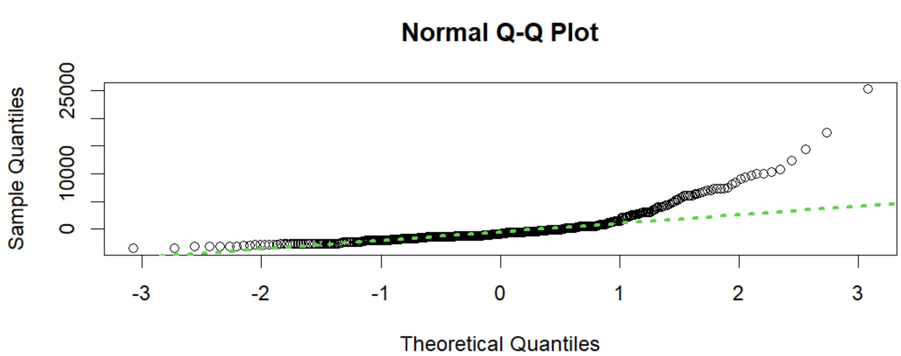
After that, we need to test both Kolmogorov-Smirnov test and Shapiro-Wilk test for the residuals to see if they are normally distributed.

```
> 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-Wilk normality test

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

In 2 tests above, p-value is much less than 0.05 so we reject the H_0 hypothesis that the beginning salary in the residuals of the anova is normally distributed.

```
> bartlett.test(salbeg~age_cut, data = salary_num)

        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 = salary_num)

        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, data = salary_num)
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
```

In the 3 tests above, we can see that the Homoscedasticity (or the homogeneity of variances) is rejected either.



Q-Q Plot doesn't help us with the normality of the residuals in the ANOVA either. Each group is big enough (n>50) so we need to see at least if the mean is sufficient descriptive measure of central location for all groups.

```
> mean(salary_num$salbeg[which(salary_num$age_cut=='[23.0,29.7)')])
[1] 5767.788
> median(salary_num$salbeg[which(salary_num$age_cut=='[23.0,29.7)')])
[1] 5370
> symmetry.test(salary_num$salbeg[which(salary_num$age_cut=='[23.0,29.7)')])

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

data:  salary_num$salbeg[which(salary_num$age_cut == "[23.0,29.7)")]
Test statistic = 4.5813, p-value < 2.2e-16
alternative hypothesis: the distribution is asymmetric.
sample estimates:
bootstrap optimal m
                 70
> mean(salary_num$salbeg[which(salary_num$age_cut=='[29.7,39.8)')])
[1] 7997.795
> median(salary_num$salbeg[which(salary_num$age_cut=='[29.7,39.8)')])
[1] 6600
> symmetry.test(salary_num$salbeg[which(salary_num$age_cut=='[29.7,39.8)')])

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

data:  salary_num$salbeg[which(salary_num$age_cut == "[29.7,39.8)")]
Test statistic = 9.0684, p-value < 2.2e-16
alternative hypothesis: the distribution is asymmetric.
sample estimates:
bootstrap optimal m
                 24

>
> mean(salary_num$salbeg[which(salary_num$age_cut=='[39.8,64.5]')])
[1] 6681.949
> median(salary_num$salbeg[which(salary_num$age_cut=='[39.8,64.5]')])
[1] 5700
> symmetry.test(salary_num$salbeg[which(salary_num$age_cut=='[39.8,64.5]')])

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

data:  salary_num$salbeg[which(salary_num$age_cut == "[39.8,64.5]")]
Test statistic = 6.556, p-value < 2.2e-16
alternative hypothesis: the distribution is asymmetric.
sample estimates:
bootstrap optimal m
                 49
```

As we can see, both the comparison of mean and median in every group as far as the symmetry test didn't help us to assume that mean is sufficient descriptive measure. We will try to see the same tests on the residuals of the ANOVA.

```
> mean(anova1$residuals)
[1] 8.428944e-13
> median(anova1$residuals)
[1] -877.7875
> symmetry.test(anova1$residuals)

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

data:  anova1$residuals
Test statistic = 11.477, p-value < 2.2e-16
alternative hypothesis: the distribution is asymmetric.
sample estimates:
bootstrap optimal m
                 91
```

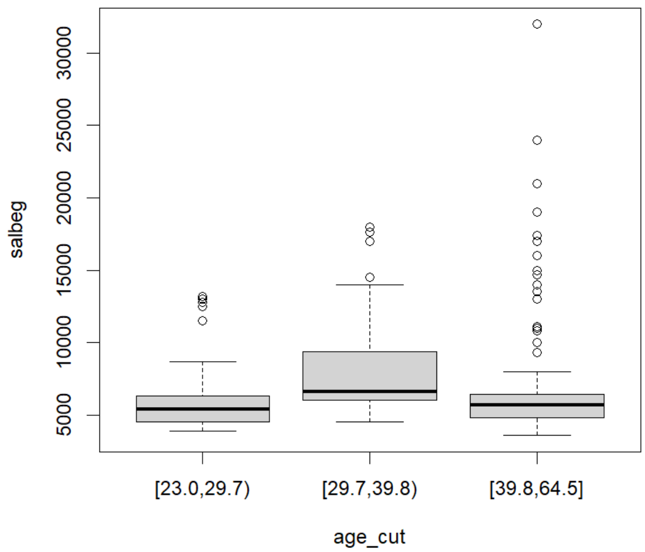
As we can see, both the comparison of mean and median in the residuals as far as the symmetry test didn't help us to assume that mean is sufficient descriptive measure. So, we will use Kruskal-Wallis test for one sample in order to test the equality of medians.

```
> kruskal.test(salbeg~age_cut, data = salary_num)

        Kruskal-Wallis rank sum test

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

So, we reject the H_o hypothesis and we say that there is a difference between the beginning salary in some of the groups. We will try to find out in what group we have differences or if we have differencies between all of them using pairwise Wilcox rank in order to do multiple comparisons between all groups.

```
> pairwise.wilcox.test(salary_num$salbeg,salary_num$age_cut)

        Pairwise comparisons using Wilcoxon rank sum test with continuity correction

data:  salary_num$salbeg and salary_num$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
```

We observe that in the first and the third group don't have any significant difference because p-value is more than 0.05(0.089 to be precise) but there might be a difference in the 2nd group and we will use boxplot in order to see if there is actual difference.



As we can see, the 2nd group is higher than the rest of them (the median also is higher) which can answer our question. On average, there are differences between the beginning salary in the groups. Specifically, in the age between 29,7 and 39,8, you will probably take a better beginning salary in this company.

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.

In order to answer this question, we need to make a table with the gender of the employees and the minority of them as well.

```
> CrossTable(salary$sex, salary$minority, digits = 1, format = "SPSS", prop.r = T,
+            prop.c = F, prop.t = F, prop.chisq = F, chisq = T, fisher = T, mcnemar = F)

   Cell Contents
|-------------------------|
|                   Count |
|             Row Percent |
|-------------------------|

Total Observations in Table:  474

             | salary$minority
  salary$sex |     WHITE  |  NONWHITE  | Row Total |
-------------|-----------|-----------|-----------|
       MALES |       194  |        64  |       258 |
             |     75.2% |     24.8% |     54.4% |
-------------|-----------|-----------|-----------|
     FEMALES |       176  |        40  |       216 |
             |     81.5% |     18.5% |     45.6% |
-------------|-----------|-----------|-----------|
Column Total |       370  |       104  |       474 |
-------------|-----------|-----------|-----------|
```

Then we need to make a Pearson Chi-squared test and a Fisher test in order to come with a conclusion.

```
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
```

As we can see from the above, it is the first time that we don't reject the H_0 hypothesis. Therefore, the proportion of white male employees might be equal to the proportion of white female employees and in conclusion, these are independent groups.

Code:

```
library(foreign)

library(psych)

library(Hmisc)

# Question 1

salary <- read.spss("salary.sav",to.data.frame=TRUE)

str(salary)

# Question 2

numeric.only <- sapply(salary, class) == "numeric"

numeric.only

salary_num <- salary[numeric.only]

salary_num <- salary_num[,-1]
```

```
head(salary_num)

summary(salary_num)

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=100)

  ynorm <- dnorm( index, mean=mean(salary_num[,i]),

                 sd(salary_num[,i]) )

  lines( index, ynorm, col=3, lty=3, lwd=3 )

}

dev.off()

# Question 3

library('nortest')

shapiro.test(salary_num$salbeg)

lillie.test(salary_num$salbeg)

ks.test(salary_num$salbeg, 'pnorm')

qqnorm(salary_num$salbeg)

qqline(salary_num$salbeg,col=3, lty=3, lwd=3 )

library(lawstat)

mean(salary_num$salbeg)

median(salary_num$salbeg)

symmetry.test(salary_num$salbeg)

wilcox.test(salary_num$salbeg, mu=1000)

# so we reject the Ho that the avg of beginning salary of a typical

# employee is equal to 1000$

# Question 4

salary_num$saldiff <- salary_num$salnow - salary_num$salbeg

shapiro.test(salary_num$saldiff)

lillie.test(salary_num$saldiff)

ks.test(salary_num$saldiff,'pnorm')
```

```
qqnorm(salary_num$saldiff)

qqline(salary_num$saldiff,col=3, lty=3, lwd=3 )

mean(salary_num$saldiff)

median(salary_num$saldiff)

symmetry.test(salary_num$saldiff)

wilcox.test(salary_num$saldiff, mu=0)

# so we reject the Ho that there isn't any significant difference

# between beggining salary and current salary

# Question 5

salary_gen <- salary[,c("salbeg","sex")]

by( salary_gen$salbeg, salary_gen$sex, lillie.test)

by( salary_gen$salbeg, salary_gen$sex, shapiro.test)

par(mfrow=c(1,2))

qqnorm(salary_gen$salbeg[which(salary_gen$sex=='MALES')], main = 'Normanl Q-Q Plot for Males')

qqline(salary_gen$salbeg[which(salary_gen$sex=='MALES')],col=3, lty=3, lwd=3 )

qqnorm(salary_gen$salbeg[which(salary_gen$sex=='FEMALES')], main = 'Normanl Q-Q Plot for Females')

qqline(salary_gen$salbeg[which(salary_gen$sex=='FEMALES')],col=3, lty=3, lwd=3 )

dev.off()

mean(salary_gen$salbeg[which(salary_gen$sex=='MALES')])

median(salary_gen$salbeg[which(salary_gen$sex=='MALES')])

symmetry.test(salary_gen$salbeg[which(salary_gen$sex=='MALES')])

mean(salary_gen$salbeg[which(salary_gen$sex=='FEMALES')])

median(salary_gen$salbeg[which(salary_gen$sex=='FEMALES')])

symmetry.test(salary_gen$salbeg[which(salary_gen$sex=='FEMALES')])

wilcox.test(salary_gen$salbeg[which(salary_gen$sex=='MALES')],salary_gen$salbeg[which(salary_gen$sex=='FEMALES')])

# so we reject the Ho that the median in beginning salary of males

# is equal to the median in the beginning salary of females

boxplot(salbeg~sex, data = salary_gen)

boxres<-boxplot(salbeg~sex, data = salary_gen, plot=F)

out.index <- which(salary_gen$salbeg %in% boxres$out)

# Question 6

age_cut <- cut2(salary_num$age,g=3)

head(age_cut)
```

```
salary_num <- data.frame(salary_num,age_cut)

anova1 <- aov(salbeg~age_cut, data = salary_num)

anova1

summary(anova1)

lillie.test(anova1$residuals)

shapiro.test(anova1$residuals)

bartlett.test(salbeg~age_cut, data = salary_num)

fligner.test(salbeg~age_cut, data = salary_num)

library(car)

leveneTest(salbeg~age_cut, data = salary_num)

qqnorm(anova1$residuals)

qqline(anova1$residuals,col=3, lty=3, lwd=3 )

mean(anova1$residuals)

median(anova1$residuals)

symmetry.test(anova1$residuals)

mean(salary_num$salbeg[which(salary_num$age_cut=='[23.0,29.7)')])

median(salary_num$salbeg[which(salary_num$age_cut=='[23.0,29.7)')])

symmetry.test(salary_num$salbeg[which(salary_num$age_cut=='[23.0,29.7)')])

mean(salary_num$salbeg[which(salary_num$age_cut=='[29.7,39.8)')])

median(salary_num$salbeg[which(salary_num$age_cut=='[29.7,39.8)')])

symmetry.test(salary_num$salbeg[which(salary_num$age_cut=='[29.7,39.8)')])

mean(salary_num$salbeg[which(salary_num$age_cut=='[39.8,64.5]')])

median(salary_num$salbeg[which(salary_num$age_cut=='[39.8,64.5]')])

symmetry.test(salary_num$salbeg[which(salary_num$age_cut=='[39.8,64.5]')])

median(anova1$residuals)

median(anova1$residuals)

symmetry.test(anova1$residuals)

kruskal.test(salbeg~age_cut, data = salary_num)

pairwise.wilcox.test(salary_num$salbeg,salary_num$age_cut)

# so we reject the Ho that the median in beginning salary

# is equal to all three groups

boxplot(salbeg~age_cut, data = salary_num)

boxres<-boxplot(salbeg~age_cut, data = salary_num, plot=F)
```

```
out.index <- which(salary_num$salbeg %in% boxres$out)

# Question 7

tab1 <- table(salary$sex, salary$minority)

tab1

prop.table(tab1)

prop.table(tab1, 1)

prop.table(tab1, 2)

prop.test(tab1)

chisq.test(tab1)

# We don't reject the Ho so these are independent groups

#Second way of showing results

require(gmodels)

CrossTable(salary$sex, salary$minority)

CrossTable(salary$sex, salary$minority, digits = 1, format = "SPSS")

CrossTable(salary$sex, salary$minority, digits = 1, format = "SPSS", prop.r = T,

      prop.c = F, prop.t = F, prop.chisq = F, chisq = T, fisher = T, mcnemar = F)

#Third way of showing results

require(sjPlot)

require(sjstats)

sjt.xtab(salary$sex, salary$minority, show.cell.prc = F, show.col.prc = F, show.exp = F)

sjp.xtab(salary$sex, salary$minority)
```