BUSINESS ANALYTICS
Master of Science

ASSIGNMENT I 15/11/2023

ELENI RALLI

A.M.: f2822312

```
library(foreign)
library(psych)       # for the describe function
library(DescTools)   # to calculate the mode
library(corrplot)    #for the corrplot
library(nortest)     # for lillie.test
library(lawstat)     # symmetry.test
library(moments)     #agostino.test , test for skewness (there is no normtest anymore)
library(Hmisc)      # cut function
library(car)        # for levene test
library(gplots)     # for error bars
```

#1. Read the dataset "salary.sav" and use the function str() to understand its structure.

```
> salary <- read.spss("C:\\Users\\eleni\\Downloads\\salary.sav", to.data.frame = T)
re-encoding from CP1253

> str(salary)      #it gives me the basic structure of the data
'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 O
F EMPLOYEE" "JOB SENIORITY" ...
  ..- attr(*, "names")= chr [1:11] "id" "salbeg" "sex" "time" ...
 - attr(*, "codepage")= int 1253


> head(salary)      #it gives me the top rows of the data
  id salbeg   sex time   age salnow edlevel  work          jobcat minority
1  1   8400 MALES   81 28.50  16080      16  0.25 COLLEGE TRAINEE    WHITE
2  2  24000 MALES   73 40.33  41400      16 12.50 EXEMPT EMPLOYEE    WHITE
3  3  10200 MALES   83 31.08  21960      15  4.08 EXEMPT EMPLOYEE    WHITE
4  4   8700 MALES   93 31.17  19200      16  1.83 COLLEGE TRAINEE    WHITE
5  5  17400 MALES   83 41.92  28350      19 13.00 EXEMPT EMPLOYEE    WHITE
6  6  12996 MALES   80 29.50  27250      18  2.42 COLLEGE TRAINEE    WHITE
      sexrace
1 WHITE MALES
2 WHITE MALES
3 WHITE MALES
4 WHITE MALES
5 WHITE MALES
6 WHITE MALES

> summary(salary)
```

```
#it gives me a summary or statistical description of the mean,median generally the quartiles the shape of a distribution for a given set of data , but if the varable is factor it gives the freq. of each level

       id              salbeg            sex              time             age       
 Min.   :  1.0    Min.   : 3600   MALES  :258   Min.   :63.00   Min.   :23.00  
 1st Qu.:119.2    1st Qu.: 4995   FEMALES:216   1st Qu.:72.00   1st Qu.:28.50  
 Median :237.5    Median : 6000                 Median :81.00   Median :32.00  
 Mean   :237.5    Mean   : 6806                 Mean   :81.11   Mean   :37.19  
 3rd Qu.:355.8    3rd Qu.: 6996                 3rd Qu.:90.00   3rd Qu.:45.98  
 Max.   :474.0    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  
```

#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?

```
#the essential statistics for a numerical variable are:
Mean,Median,Mode,Percentile,Quartiles (five-number summary = (minimum value, lower quartile (Q1), median value (Q2), upper quartile (Q3), maximum value)),
Standard Deviation,Variance,Range,Proportion ,Correlation(συσχέτιση)


> #keep the numeric-variables
> index <- sapply(salary, class) == "numeric"
> sal_num <- salary[index]

> head(sal_num)
  id salbeg time   age salnow edlevel  work
1  1   8400   81 28.50  16080      16  0.25
2  2  24000   73 40.33  41400      16 12.50
3  3  10200   83 31.08  21960      15  4.08
4  4   8700   93 31.17  19200      16  1.83
5  5  17400   83 41.92  28350      19 13.00
6  6  12996   80 29.50  27250      18  2.42


> sal_num <- sal_num[,-1] #without the column id

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


> describe(sal_num[,])
```

```
sal_num[, ]

 6  Variables      474  Observations
--------------------------------------------------------------------------------
salbeg
       n  missing distinct     Info     Mean      Gmd      .05      .10
     474        0       90    0.997     6806     2846     4080     4380
     .25      .50      .75      .90      .95
    4995     6000     6996    11000    13200

lowest :  3600  3900  4020  4080  4200, highest: 18000 18996 21000 24000 31992
--------------------------------------------------------------------------------
time
       n  missing distinct     Info     Mean      Gmd      .05      .10
     474        0       36    0.999    81.11    11.61       65       67
     .25      .50      .75      .90      .95
      72       81       90       94       97

lowest : 63 64 65 66 67, highest: 94 95 96 97 98
--------------------------------------------------------------------------------
age
       n  missing distinct     Info     Mean      Gmd      .05      .10
     474        0      259        1    37.19    12.83    24.42    25.19
     .25      .50      .75      .90      .95
   28.50    32.00    45.98    56.84    60.67

lowest : 23    23.25 23.33 23.42 23.58, highest: 63.75 63.83 63.92 64.25 64.5
--------------------------------------------------------------------------------
salnow
       n  missing distinct     Info     Mean      Gmd      .05      .10
     474        0      221        1    13768     6534     7797     8418
     .25      .50      .75      .90      .95
    9600    11550    14775    23757    28000

lowest :  6300  6360  6480  6540  6600, highest: 40000 41400 41500 44250 54000
--------------------------------------------------------------------------------
edlevel
       n  missing distinct     Info     Mean      Gmd      .05      .10
     474        0       10    0.917    13.49    3.128        8        8
     .25      .50      .75      .90      .95
      12       12       15       17       19

Value          8    12    14    15    16    17    18    19    20    21
Frequency     53   190     6   116    59    11     9    27     2     1
Proportion 0.112 0.401 0.013 0.245 0.124 0.023 0.019 0.057 0.004 0.002

For the frequency table, variable is rounded to the nearest 0
--------------------------------------------------------------------------------
work
       n  missing distinct     Info     Mean      Gmd      .05      .10
     474        0      208        1    7.989    8.889   0.1105   0.4200
     .25      .50      .75      .90      .95
  1.6025   4.5800  11.5600  21.6750  26.7855

lowest : 0     0.17  0.25  0.33  0.42 , highest: 36.5  37    37.58 38.33 39.67
--------------------------------------------------------------------------------
#trimmed(mean) cuts 5% from each tail--> more robust to extreme values
#the standard error (se) essentially tells us the variability of the sample, how much
#close are the values to the mean value (at the center of the distribution)

> #mode (η επικρατουσα τιμή)
> for (i in 1:6) {
+   print(Mode(sal_num[, i]))
+ }
[1] 6000
attr(,"freq")
[1] 52
[1] 81 93
attr(,"freq")
[1] 23
[1] 29.50 30.33
```

```
attr(,"freq")
[1] 6
[1] 12300
attr(,"freq")
[1] 13
[1] 12
attr(,"freq")
[1] 190
[1] 0
attr(,"freq")
[1] 24

> #Range
> for (i in 1:6) {
+   print(range(sal_num[,i ]))
+ }
[1]  3600 31992
[1] 63 98
[1] 23.0 64.5
[1]  6300 54000
[1]  8 21
[1]  0.00 39.67


> #i will create this function to check quickly all these
> mean_median_skew_kurt <- function(data_vector) {
+   if (!is.numeric(data_vector)) {
+     stop("Data vector is not numeric.")
+   }
+
+   results <- list(
+     Mean = mean(data_vector, na.rm = TRUE),
+     Median = median(data_vector, na.rm = TRUE),
+     Skewness = skewness(data_vector, na.rm = TRUE),
+     Kurtosis = kurtosis(data_vector, na.rm = TRUE)
+   )
+
+   print(results)
+
+ }

> mean_median_skew_kurt(sal_num$salbeg)
$Mean
[1] 6806.435

$Median
[1] 6000

$Skewness
[1] 2.84382

$Kurtosis
[1] 15.24727

> mean_median_skew_kurt(sal_num$time)
$Mean
[1] 81.1097

$Median
[1] 81

$Skewness
[1] -0.05240323

$Kurtosis
[1] 1.846897

> mean_median_skew_kurt(sal_num$age)
$Mean
[1] 37.18614
```

```
$Median
[1] 32

$Skewness
[1] 0.8617367

$Kurtosis
[1] 2.43167

> mean_median_skew_kurt(sal_num$salnow)
$Mean
[1] 13767.83

$Median
[1] 11550

$Skewness
[1] 2.117877

$Kurtosis
[1] 8.30863

> mean_median_skew_kurt(sal_num$edlevel)
$Mean
[1] 13.49156

$Median
[1] 12

$Skewness
[1] -0.1137455

$Kurtosis
[1] 2.725155

> mean_median_skew_kurt(sal_num$work)
$Mean
[1] 7.988608

$Median
[1] 4.58

$Skewness
[1] 1.505254

$Kurtosis
[1] 4.665567
```

- Skewness should be about 0. This measures symmetry, so a skewness near 0 indicates a symmetric distribution.
  If the distribution can be folded along a central value and both sides match, it is symmetrical and has zero skewness. If they don't match, the distribution is skewed
  A distribution is positively skewed (or right-skewed)--> the tail on the right side of the distribution is longer
  -->the mean and median will be greater than the mode.
- Kurtosis should be about 3. This measures the "tailedness" of the distribution.

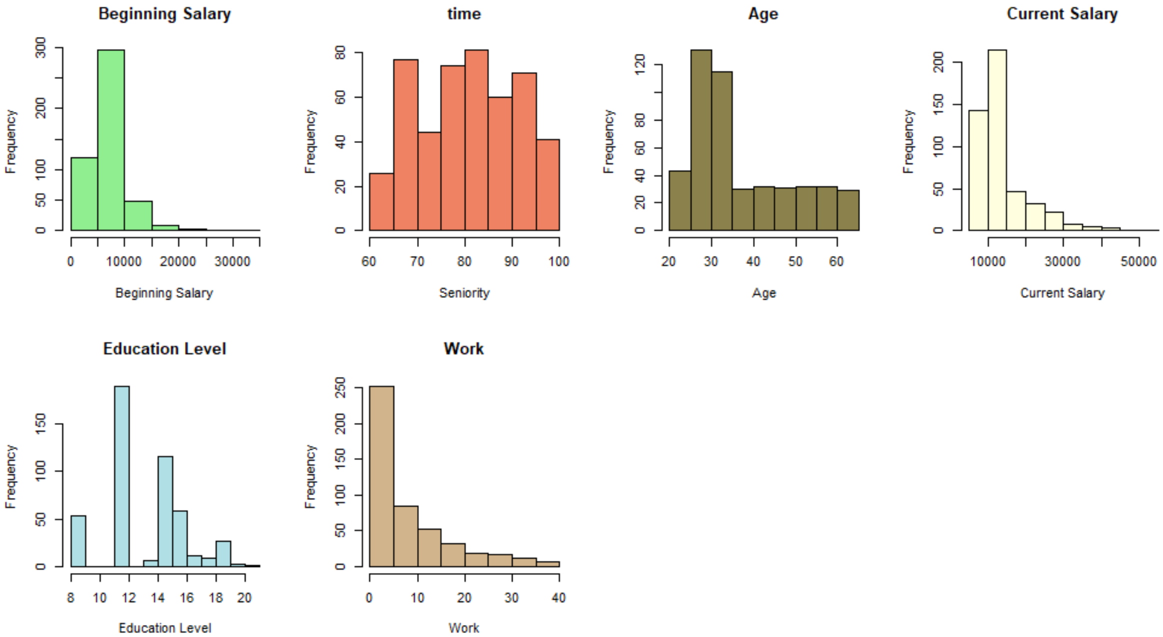
heavy tails:
in a distribution if the probability of finding observations far away from the mean is higher than it would be in a normal distribution
This characteristic can lead to a higher likelihood of outliers, which can significantly affect the mean and variance of the data.

salbeg

Skewness: 2.84382 (Highly skewed) Kurtosis: 15.24727 (High kurtosis, indicating heavy tails) Not normal
Time Skewness: -0.05240323 (Very close to 0, suggesting symmetry) Kurtosis: 1.846897 (Lower than 3, indicating lighter tails) Possibly close to normal but with lighter tails
age Skewness: 0.8617367 (Moderately skewed) Kurtosis: 2.43167 (Slightly less than 3, lighter tails) Not perfectly normal, but close
age Skewness: 0.8617367 (Moderately skewed) Kurtosis: 2.43167 (Slightly less than 3, lighter tails) Not perfectly normal, but close
salnow Skewness: 2.117877 (Highly skewed) Kurtosis: 8.30863 (High, indicating heavy tails) Not normal
edlevel Skewness: -0.1137455 (Close to 0, but slightly negatively skewed) Kurtosis: 2.725155 (Close to 3, slightly lighter tails) Possibly close to normal, but slightly negatively skewed
Work Skewness: 1.505254 (Moderately skewed) Kurtosis: 4.665567 (Slightly higher than 3, indicating slightly heavier tails) Not perfectly normal, but could be close

So, none of these variables perfectly fit a normal distribution, although some (like time, age, and edlevel) are closer than others.

```
#Diagrams that i can do for numeric variables ( i will not do all of them ) :
#Histogram,qqplot,Scatter Plot,Box Plot,Line Chart,Bar Chart,Heatmap,Density Plot,Violin Plot,Area Chart,Bubble Chart


Histogram (show the distribution of a single numeric variable )
It groups data into bins and shows the frequency of observations in each bin.

> par(mfrow = c(2, 4))
> hist(sal_num$salbeg,main = "Beginning Salary", col = "lightgreen", xlab = "Beginning Sa
lary")
> hist(salary$time, main = "time", col = "salmon2", xlab = "Seniority")
> hist(salary$age, main = "Age", col ="lightgoldenrod4"   , xlab = "Age")
> hist(salary$salnow, main = "Current Salary", col = "lightyellow", xlab = "Current Salar
y")
> hist(salary$edlevel, main = "Education Level", col = "powderblue", xlab = "Education Le
vel")
> hist(salary$work, main = "Work", col = "tan", xlab = "Work")
```

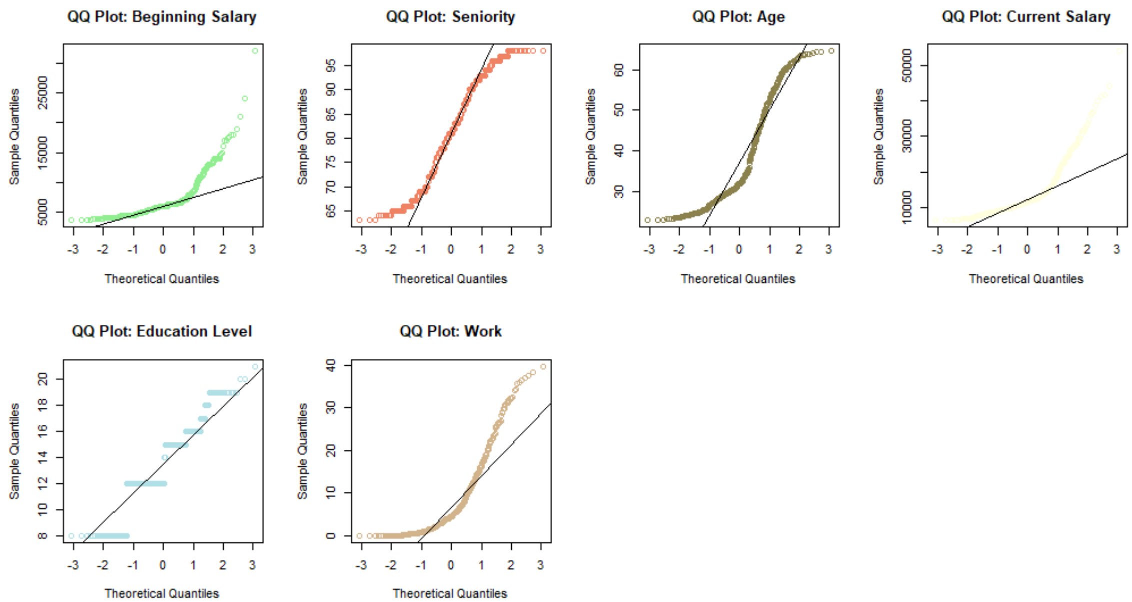


So, none of these variables seems to follow normal distribution

qqplots

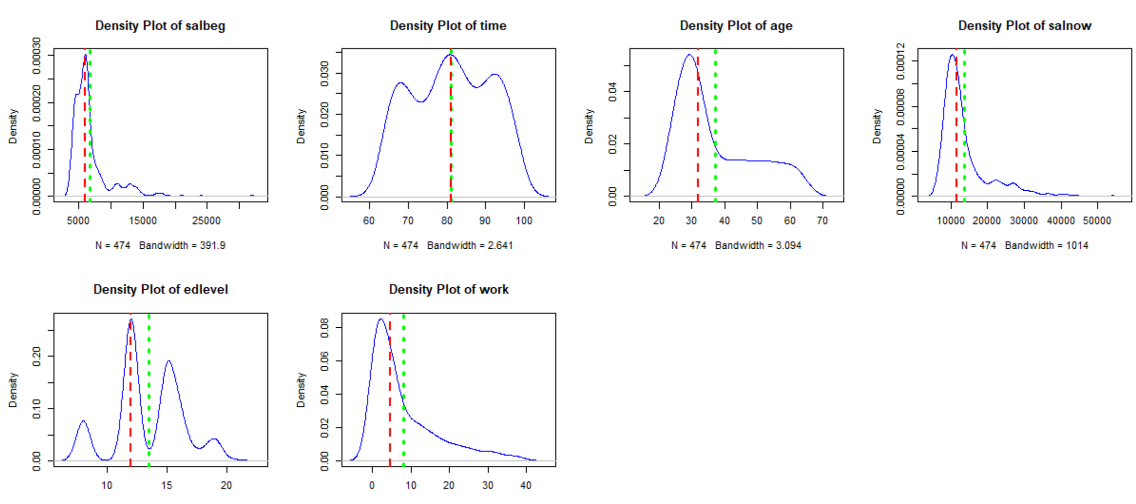
QQ plot is used to compare the distributions of two datasets or to compare the distribution of a dataset to a theoretical distribution, such as the normal distribution.

```
> par(mfrow = c(2, 4))
> qqnorm(sal_num$salbeg, main = "QQ Plot: Beginning Salary", col = "lightgreen")
> qqline(sal_num$salbeg)
nonlinear pattern --> heavy tails-->This indicates that the distribution of beginning sal
aries is not normal, with a potential right skew and outliers on the higher end.

> qqnorm(salary$time, main = "QQ Plot: Seniority", col = "salmon2")
> qqline(salary$time)
> #heavier tails than a normal distribution--> doesn't seem normally distributed
>
> qqnorm(salary$age, main = "QQ Plot: Age", col ="lightgoldenrod4")
> qqline(salary$age)
> #age is approximately normal but with some skewness or outliers.

> qqnorm(salary$salnow, main = "QQ Plot: Current Salary", col = "lightyellow")
> qqline(salary$salnow)
> # right skewness and potential outliers, not seem  normal.
>
> qqnorm(salary$edlevel, main = "QQ Plot: Education Level", col = "powderblue")
> qqline(salary$edlevel)

#a step-like pattern(discreteness)-->characteristic of discrete distributions or distribu
tions with a limited range of values--> not normally distributed.

> qqnorm(salary$work, main = "QQ Plot: Work", col = "tan")
> qqline(salary$work)
> #right tail -->the distribution of this variable is not normal
```



Density Plots:
Density plots show the distribution of a variable and can be used to identify the shape of the distribution.
The value on the y-axis for any given point on the x-axis represents the probability density of the variable at that point.
It is not a probability itself, but the height proportional to the likelihood of observing a value at or near that point.

```
> par(mfrow = c(2, 4))  # Set up a 2x2 grid for density plots

> for (col in colnames(sal_num)) {
+   mean_salbeg <- mean(sal_num[,col])
+   median_salbeg <- median(sal_num[,col])
+   plot(density(sal_num[,col ]), main = paste("Density Plot of", col), col = "blue")
+   abline(v=mean_salbeg, col="green", lwd=3, lty=3)
+   abline(v=median_salbeg, col="red", lwd=2, lty=2)
+ }

> # anyone seems NOT normally distribute
```



#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.

We want to check
Ho:μ=1000   The mean beginning salary of employees is equal to 1000 dollars.
H1:μ=!1000 The mean beginning salary of employees is not equal to 1000 dollars

salbeg is numeric variable --> I will do a hypothesis test for a sample (1 quantitative variable)

#i) Is our variable normal? - normality check (SW αν n<=50 / KS + SW αν n>50)

Shapiro-Wilk (SW- specifically designed for testing normality-check whether a dataset follows a normal distribution.)
Ho:The sample salbeg is from a normally distributed population.
H1:The sample salbeg is not from a normally distributed population.

Kolmogorov-Smirnov (KS- ks.test but we use lillie.test because It corrects for the bias that occurs in the Kolmogorov-Smirnov test due to parameter estimation)
check whether a dataset follows a specified theoretical distribution or compares two datasets to see if they come from the same distribution.
Ho:The sample data follows the specified theoretical distribution (or the two datasets are from the same distribution).
H1: not
The Lilliefors test is particularly useful for smaller sample sizes, as it is designed to be more sensitive to deviations from normality in such cases. It is often used when testing for normality.

generally Shapiro-Wilk stricter than Lilliefors (rejects H0 more easily)

```
> length(sal_num$salbeg) # n = 474
[1] 474
> # οποτε κάνω KS + SW αν n>50
```

```
> lillie.test(sal_num$salbeg)

        Lilliefors (Kolmogorov-Smirnov) normality test

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

p-value < 2.2e-16< a --> Ho rejected --> rejected the hypothesis of The sample salbeg is from a normally distributed population.

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

        Shapiro-Wilk normality test

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

p-value < 2.2e-16< a --> Ho rejected --> rejected the hypothesis of The sample salbeg is from a normally distributed population.

Our variable salbeg is NOT normal.
but n = 474 >50 , we say that our sample is large
is the mean an appropriate measure of centrality? --> This is subjective based on the verbal problem
So should I use mean or median? Accordingly, I will see if the mean and the median have a big difference (I can see it a bit from the diagrams)
if not I can t-test test for the mean value otherwise I will do Wilcoxon test test for the median

```
> mean_median_skew_kurt(sal_num$salbeg)
$Mean
[1] 6806.435

$Median
[1] 6000

$Skewness
[1] 2.84382

$Kurtosis
[1] 15.24727
```

from the Density plot before, the mean and median do not seem to be far apart , but here I see big difference
#let's also do a symmetry.test (HO of symmetry of the distribution, if rejected I go with a non-parametric test

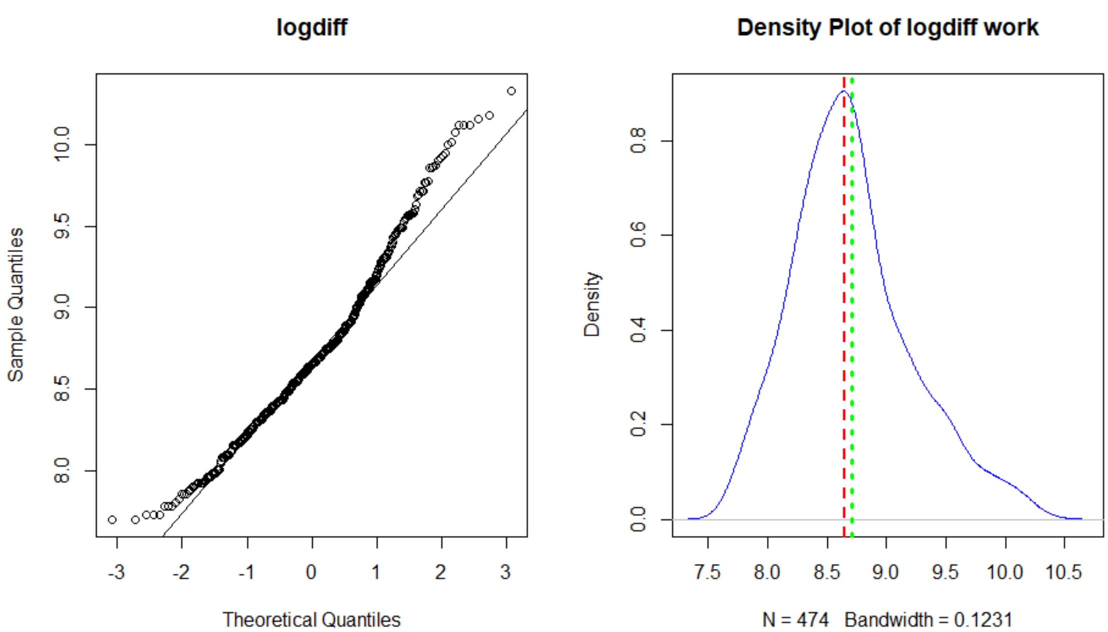
```
> symmetry.test(sal_num$salbeg) # p-value < 2.2e-16< a --> Ho rejected --> θα προχωρήσω μ
ε wilcox.test (βέβαια είναι αρκετά αυστηρο το symmetry.test)

        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
                 57

> agostino.test(sal_num$salbeg) # p-value < 2.2e-16< a --> Ho(the skewness of the dataset
is zero or close enough to zero) rejected

        D'Agostino skewness test

data:  sal_num$salbeg
skew = 2.8438, z = 14.6208, p-value < 2.2e-16
alternative hypothesis: data have a skewness
```

I go with a non-parametric test

```
H0 :median(salbeg)=1000
H1:median(salbeg)!=1000

> wilcox.test(sal_num$salbeg, mu=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 < 2.2e-16< a --> Ho rejected-->REJECT THAT The MEDIAN beginning salary of
employees is equal to 1000 dollars
```

#4. Consider the natural logarithm of 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 log(salbeg – salnow) and test if, on average, it is equal to one.). Make sure that the choice of the test is well justified.

```
> par(mfrow = c(1, 2))

Considering the log difference between current and beginning salary

> logdiff <- log(sal_num$salnow-sal_num$salbeg)

We  check :

Ho:μ(logdiff)=1   The mean of the log-transformed differences between the beginning salar
y and current salary of employees is equal to 1 dollars.

H1:μ(logdiff)!=1

> qqnorm(logdiff,main="logdiff" ); qqline(logdiff)
> plot(density(logdiff), main = paste("Density Plot of logdiff", col), col = "blue")
> abline(v=mean(logdiff), col="green", lwd=3, lty=3)
> abline(v=median(logdiff), col="red", lwd=2, lty=2)

> #βλέπω να υπάρχει κάποια απόκλιση απο την κανονικότητα ,heavy tails , no symmetry
```




```
> length(logdiff) #n=474>50 so. I do KS + SW αν n>50
[1] 474


> lillie.test(logdiff) # p-value < 2.2e-16< a --> Ho rejected --> rejected the hypothesis
of The  logdiff is from a normally distributed population.

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  logdiff
D = 0.08221, p-value = 3.986e-08
```

```
> shapiro.test(logdiff) #Ho rejected--> rejected the hypothesis of The  logdiff is from a normally distributed population.

        Shapiro-Wilk normality test

data:  logdiff
W = 0.9721, p-value = 7.441e-08
```

is the mean an appropriate measure of centrality? --> This is subjective based on the verbal problem

So should I use mean or median? Accordingly, I will see if the mean and the median have a big difference (I can see it a bit from the diagrams)

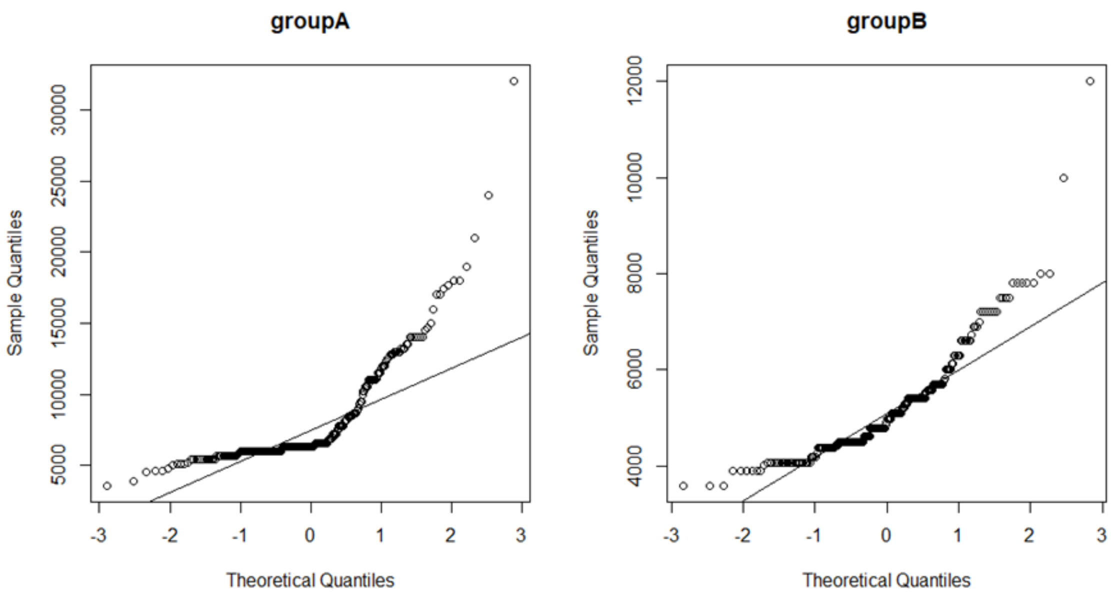
```
> mean(logdiff);median(logdiff) #Reasonably close agreement between mean and variance , o
πως φαινεται και απο το διαγραμμα
[1] 8.706788
[1] 8.648221
```

```
κάνω   παραμετρικό t.test

> t.test(logdiff, mu=1)

        One Sample t-test

data:  logdiff
t = 331.33, df = 473, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 1
95 percent confidence interval:
 8.661082 8.752494
sample estimates:
mean of x
 8.706788

# p-value < 2.2e-16< a --> Ho rejected --> rejected the hypothesis of The mean of the log-transformed differences between the beginning salary and current salary of employees is equal to 1 dollars.
```

#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.

```
we have to do with 2 independents samples ( 1 ποσοτική , 1 δίτιμη )

> groupA <- salary$salbeg[salary$sex == 'MALES']
> groupB <- salary$salbeg[salary$sex == 'FEMALES']
> n1<-length(groupA);n1 #n1=258 >50
[1] 258
> n2<-length(groupB);n2 #n2=216 >50
[1] 216

αφου n1,n2>50 θα κάνω  KS + SW

> dataset1 <- data.frame( salbeg=c(groupA, groupB),  method=factor( rep(1:2, c(n1,n2)), labels=c('MALES','FEMALES') ) )

ψαχνω αρχικά να δω αν η ποσοτική μεταβλητή είναι κανονική σε κάθε ομάδα (Tests for normality for each group) αφου n1,n2>50 θα κάνω  KS + SW

> #ας το δω ομως και γραφικά #qqplots
> par(mfrow=c(1,2))
> qqnorm(groupA , main = "groupA")
> qqline(groupA)
> qqnorm(groupB, main = "groupB")
```

```
> qqline(groupB)
```



Not seems normal,και outliers και heavytails , το groupB κανει και κάτι σαν step-like pattern(discreteness)

```
> by(dataset1$salbeg, dataset1$method, lillie.test)
dataset1$method: MALES

        Lilliefors (Kolmogorov-Smirnov) normality test

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

------------------------------------------------------------------------------------------
dataset1$method: FEMALES

        Lilliefors (Kolmogorov-Smirnov) normality test

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

# p-value < 2.2e-16< a --> Ho rejected --> rejected the hypothesis of normality in each group

> by(dataset1$salbeg, dataset1$method, shapiro.test)
dataset1$method: MALES

        Shapiro-Wilk normality test

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

------------------------------------------------------------------------------------------
dataset1$method: FEMALES

        Shapiro-Wilk normality test

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

# p-value < 2.2e-16< a --> Ho rejected --> rejected the hypothesis of normality in each group
```

```
but the samples are large n1,n2>50
so now we will check if the average is a suitable measure to describe the central positio
n for both teams:

> mean_median_skew_kurt(groupA)
$Mean
[1] 8120.558

$Median
[1] 6300

$Skewness
[1] 2.375938

$Kurtosis
[1] 11.3015

> mean_median_skew_kurt(groupB)
$Mean
[1] 5236.787

$Median
[1] 4950

$Skewness
[1] 1.754602

$Kurtosis
[1] 8.200857



> symmetry.test(groupA)# Ho rejected

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

data:  groupA
Test statistic = 13.829, p-value < 2.2e-16
alternative hypothesis: the distribution is asymmetric.
sample estimates:
bootstrap optimal m
                 28

> agostino.test(groupA)# Ho rejected

        D'Agostino skewness test

data:  groupA
skew = 2.3759, z = 10.0389, p-value < 2.2e-16
alternative hypothesis: data have a skewness

> symmetry.test(groupB)#Ho rejected

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

data:  groupB
Test statistic = 5.2527, p-value < 2.2e-16
alternative hypothesis: the distribution is asymmetric.
sample estimates:
bootstrap optimal m
                107

> agostino.test(groupB)# Ho rejected

        D'Agostino skewness test

data:  groupB
skew = 1.7546, z = 7.7789, p-value = 7.318e-15
alternative hypothesis: data have a skewness
```

So we go nonparametric wilcoxon.test tests for zero difference of medians
Ho:M1-M2=0
H1:M1-M2!=0

```
> wilcox.test(dataset1$salbeg ~ dataset1$method, mu=0)# p-value < 2.2e-16< a --> Ho rejec
ted

        Wilcoxon rank sum test with continuity correction

data:  dataset1$salbeg by dataset1$method
W = 47874, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
```

REJECT THAT The MEDIAN difference on the beginning salary (salbeg) between the two gender
s is 0. There is a significant difference

and now we will boxplot by group to show it

```
> par(mfrow=c(1,1))


> boxplot(groupA,groupB,
+         main="relation between beginning salary and sex ",las=1,
+         ylab="begging salary",
+         xlab="MALES         FEMALES",col=c("lightblue","lightpink"))
```

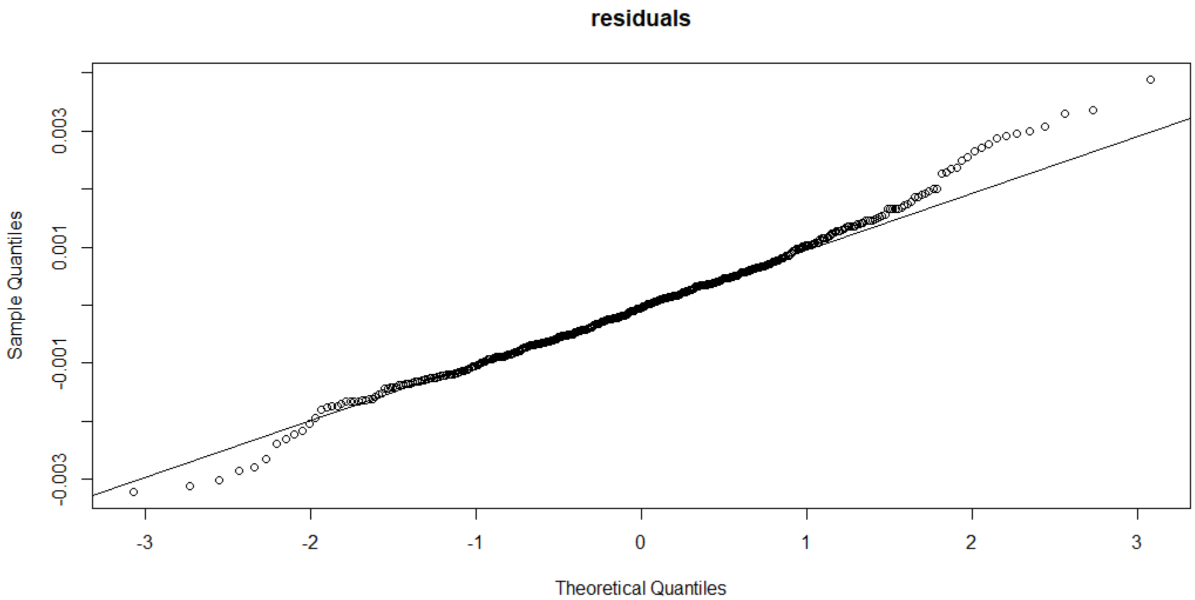
there are common points, so we cannot draw any conclusions from the diagram=
the differences in the medians are not that large, (there is considerable overlap in the gender distributions)
When the boxplots overlap significantly (although here they do not overlap significantly but at least have common points), it is more difficult to reach clear conclusions about the existence or not of significant differences between the groups based on the boxplot alone
.However, men seem to have higher begging salaries, I see that the medians of the two boxplots have a significant difference.

#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. Create the following variable with the name relSal:relsal <- ((salary$salnow-salary$salbeg)/salary$salnow)*(1/salary$time)

Investigate if, on average, the relative salary rise (relSal) 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.

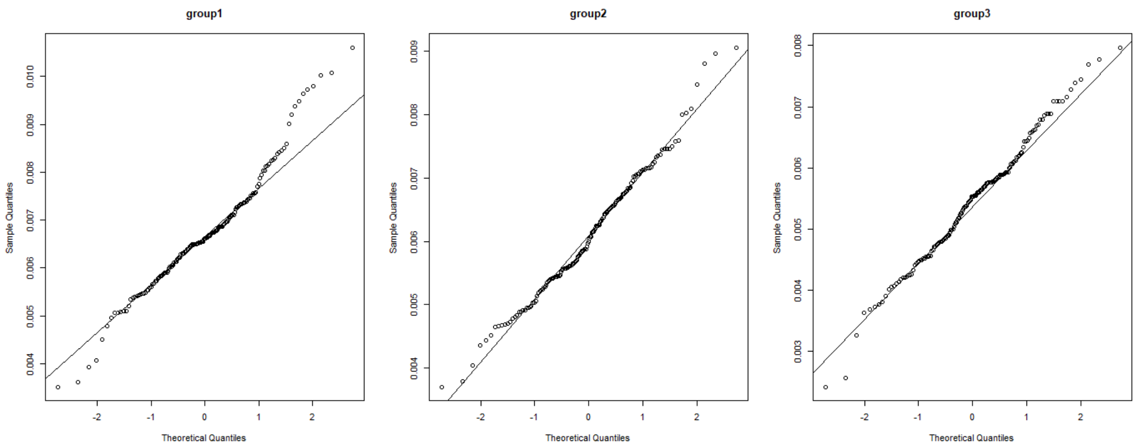
```
> salary$age_cut <- cut2(salary$age, g=3) # Cut the 'age' variable into three categories
> head(salary)
  id salbeg   sex time   age salnow edlevel  work          jobcat minority     sexrace
age_cut      relSal
1  1   8400 MALES   81 28.50  16080      16  0.25 COLLEGE TRAINEE    WHITE WHITE MALES [2
3.0,29.7) 0.005896444
2  2  24000 MALES   73 40.33  41400      16 12.50 EXEMPT EMPLOYEE    WHITE WHITE MALES [3
9.8,64.5] 0.005757395
3  3  10200 MALES   83 31.08  21960      15  4.08 EXEMPT EMPLOYEE    WHITE WHITE MALES [2
9.7,39.8) 0.006452038
4  4   8700 MALES   93 31.17  19200      16  1.83 COLLEGE TRAINEE    WHITE WHITE MALES [2
9.7,39.8) 0.005880376
5  5  17400 MALES   83 41.92  28350      19 13.00 EXEMPT EMPLOYEE    WHITE WHITE MALES [3
9.8,64.5] 0.004653535
6  6  12996 MALES   80 29.50  27250      18  2.42 COLLEGE TRAINEE    WHITE WHITE MALES [2
3.0,29.7) 0.006538532

> relSal <- ((salary$salnow-salary$salbeg)/salary$salnow)*(1/salary$time)

what we have here is 1 quantitative and 1 categorical with many levels we want to contro:
H0:μ1=μ2=μ3
H1:μi!=μj  ( για κάποιο i!=j =1,2,...,k)
let's see first if we can do this parametric control or if we should go non-parametric

to see this we first check if our residuals are normal:
here we first check if the residuals are normal

αρχικα ας δούμε αν n1,n2,n3>50
> table(salary$age_cut)

[23.0,29.7) [29.7,39.8) [39.8,64.5]
        160         156         158

> #οποτε αφου n1,n2,n3>50 θα κάνουμε KS + SW(πιο αυστηρό) για έλεγχο κανονικότητας καταλο
ίπων
> anova1 <- aov(relSal~age_cut  , data=salary )
> shapiro.test(anova1$res) #p-value = 0.0006393<a --> H0 κανονικότητας καταλοίπων rejecte
d

        Shapiro-Wilk normality test

data:  anova1$res
W = 0.98805, p-value = 0.0006393

> lillie.test(anova1$res) #p-value = 0.07976>a --> H0 dont rejected

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  anova1$res
D = 0.039118, p-value = 0.07976

> par(mfrow=c(1,1))
> qqnorm(anova1$residuals);qqline(anova1$residuals) #heavy tails
```



now that one test rejects normality and the other does not I will continue with the indication that normality will not apply to the residuals

οποτε μετά εξετάζω αν τα δείγματα >50 , ειπαμε οτι είναι >50 (για να ισχυει το ΚΟΘ)
Αρα τώρα πρέπει να δω αν είναι ο μέσος κατάλληλο μετρο περιγραφης της κεντρικής θέσης και για τις κ ομάδες

```
> #ας το δω ομως και γραφικά #qqplots

> salary$relSal <- ((salary$salnow - salary$salbeg) / salary$salnow) * (1 / salary$time)
> head(salary)
  id salbeg   sex time   age salnow edlevel  work          jobcat minority     sexrace
age_cut      relSal
1  1   8400 MALES   81 28.50  16080      16  0.25 COLLEGE TRAINEE    WHITE WHITE MALES [2
3.0,29.7) 0.005896444
2  2  24000 MALES   73 40.33  41400      16 12.50 EXEMPT EMPLOYEE    WHITE WHITE MALES [3
9.8,64.5] 0.005757395
3  3  10200 MALES   83 31.08  21960      15  4.08 EXEMPT EMPLOYEE    WHITE WHITE MALES [2
9.7,39.8) 0.006452038
4  4   8700 MALES   93 31.17  19200      16  1.83 COLLEGE TRAINEE    WHITE WHITE MALES [2
9.7,39.8) 0.005880376
5  5  17400 MALES   83 41.92  28350      19 13.00 EXEMPT EMPLOYEE    WHITE WHITE MALES [3
9.8,64.5] 0.004653535
6  6  12996 MALES   80 29.50  27250      18  2.42 COLLEGE TRAINEE    WHITE WHITE MALES [2
3.0,29.7) 0.006538532

> group1<-salary$relSal[salary$age_cut=="[23.0,29.7)"]
> group2<-salary$relSal[salary$age_cut=="[29.7,39.8)"]
> group3<-salary$relSal[salary$age_cut=="[39.8,64.5]"]

> par(mfrow=c(1,3))
> qqnorm(group1,main = "group1")
> qqline(group1)
> qqnorm(group2,main = "group2")
> qqline(group2)
> qqnorm(group3,main = "group3")
> qqline(group3)
```




```
>  mean_median_skew_kurt(group1)
$Mean
[1] 0.006725987

$Median
[1] 0.006620492

$Skewness
[1] 0.5067726

$Kurtosis
[1] 3.996509

>  mean_median_skew_kurt(group2)
$Mean
[1] 0.00609415

$Median
[1] 0.006025212

$Skewness
[1] 0.3404747

$Kurtosis
[1] 3.091786

>  mean_median_skew_kurt(group3)
$Mean
[1] 0.005429869

$Median
[1] 0.005532414

$Skewness
[1] -0.04072111

$Kurtosis
[1] 3.103651



> symmetry.test(group1)# Ho DOESNT  rejected ,suggesting that the distribution is symmetr
ic.

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

data:  group1
Test statistic = 1.5762, p-value = 0.112
alternative hypothesis: the distribution is asymmetric.
sample estimates:
bootstrap optimal m
```

```
                 55

> agostino.test(group1)# Ho rejected( reject the distribution is symmetric)

        D'Agostino skewness test

data:  group1
skew = 0.50677, z = 2.59412, p-value = 0.009483
alternative hypothesis: data have a skewness

> symmetry.test(group2)#Ho DOESNT  rejected

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

data:  group2
Test statistic = 1.1113, p-value = 0.396
alternative hypothesis: the distribution is asymmetric.
sample estimates:
bootstrap optimal m
                 17

> agostino.test(group2)# Ho DOESNT  rejected

        D'Agostino skewness test

data:  group2
skew = 0.34047, z = 1.76847, p-value = 0.07698
alternative hypothesis: data have a skewness

> symmetry.test(group3)#Ho DOESNT rejected

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

data:  group3
Test statistic = -1.7159, p-value = 0.14
alternative hypothesis: the distribution is asymmetric.
sample estimates:
bootstrap optimal m
                 99

> agostino.test(group3)# Ho DOESNT rejected

        D'Agostino skewness test

data:  group3
skew = -0.040721, z = -0.217841, p-value = 0.8276
alternative hypothesis: data have a skewness

for the group1, the 2 test doesn't agree :
The key difference is that the first test specifically checks for symmetry, while the sec
ond test checks for skewness.
A distribution can be symmetric but still not perfectly symmetric, which might explain wh
y the first test suggests
that the distribution is not perfectly symmetric but does not find strong evidence of ske
wness (p-value = 0.098).
On the other hand, the D'Agostino skewness test is more sensitive to skewness and may det
ect even mild departures from perfect symmetry (p-value = 0.009483).
For this reason, I will consider that the average is a suitable measure of describing the
neutral position and I will check for equal variances:

> #H0: The variances of the groups are equal (homogeneity of variances).
> leveneTest(relSal~age_cut, data=salary,center=mean)
Levene's Test for Homogeneity of Variance (center = mean)
       Df F value Pr(>F)
group   2  1.0088 0.3654
      471

#pvalue=0.3654>a => the H0 for Homogeneity of Variance doesnt regected

> bartlett.test(relSal~age_cut, data=salary)

        Bartlett test of homogeneity of variances
```

```
data:  relSal by age_cut
Bartlett's K-squared = 8.6022, df = 2, p-value = 0.01355
```

#p-value = 0.01355 regected if a=0.05 but for a=0.01 doesnt rejected

```
> fligner.test(relSal~age_cut, data=salary)#p-value = 0.6132 >a doesnt regected

        Fligner-Killeen test of homogeneity of variances

data:  relSal by age_cut
Fligner-Killeen:med chi-squared = 0.97816, df = 2, p-value = 0.6132
```

so, now i assume homogeneity of variances and i can do :

```
> summary(anova1) #pvalue<2e-16<a--> H0 regected
             Df    Sum Sq   Mean Sq F value Pr(>F)
age_cut       2 0.0001336 6.678e-05   56.81 <2e-16 ***
Residuals   471 0.0005537 1.180e-06
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

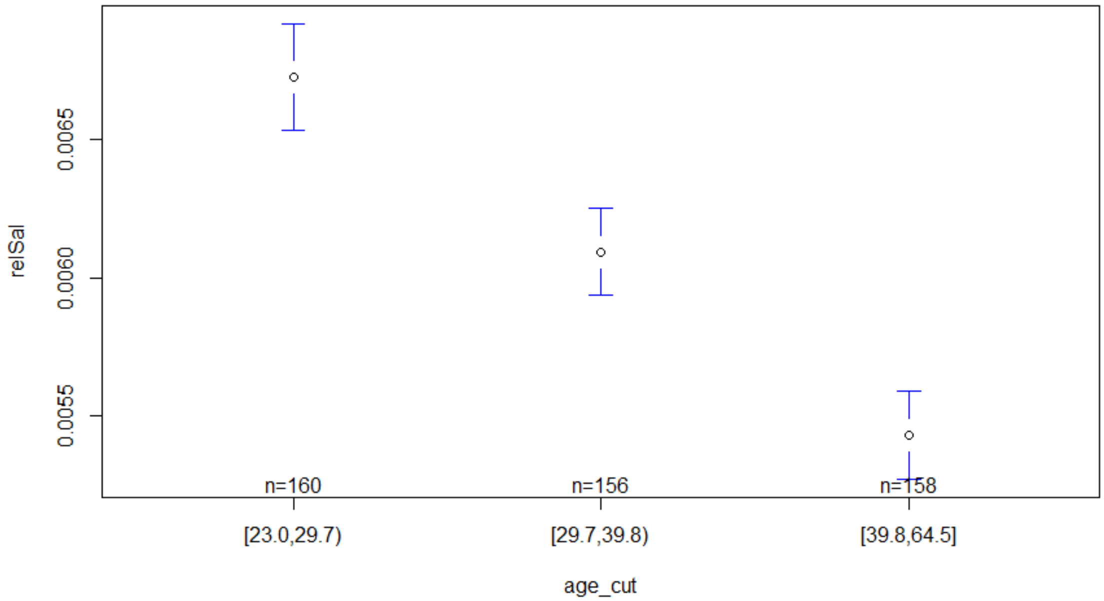
error bar ανα ομάδα για να δω ποιες είναι οι συγκεκριμένες μέσες τιμές ανα 2ομάδες που διαφέρουν
και posthoc statistic λαμβάνουμε υπόψην την πολλαπλότητα και δεν είμαστε τόσο αυστηροι

```
> pairwise.t.test(salary$relSal,salary$age_cut,p.adjust.method=p.adjust.methods[1]) #μπρούσα και benforini

        Pairwise comparisons using t tests with pooled SD

data:  salary$relSal and salary$age_cut

            [23.0,29.7) [29.7,39.8)
[29.7,39.8) 3.3e-07     -
[39.8,64.5] < 2e-16     1.8e-07

P value adjustment method: holm
```

between the group [23.0,29.7) and [29.7,39.8) pvalue<a --> ho rejected (reject that that the means of these two groups are equal),
between the group [29.7,39.8) and [39.8,64.5] pvalue=1.8e-07--> ho rejected
between the group [23.0,29.7) and [39.8,64.5] pvalue<2e-16 --> ho rejected

The conclusion from this analysis is that we reject that
there are not significant differences in the salary$relSal variable among the age groups defined by salary$age_cut.
=>strong evidence that significant differences in the salary$relSal variable among the age groups defined by salary$age_cut.

```
> par(mfrow=c(1,1))
> plotmeans(relSal~age_cut,data=salary,connect=F)
```



#error bars are short, it indicates a lower degree of variability or uncertainty in the mean.
#If two bars have non-overlapping error bars, it suggests that the means of those groups are significantly different from each other
#there is any overlap here

#7. Investigate if, on average, the relative salary rise (relSal) is the same for all job categories. 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.

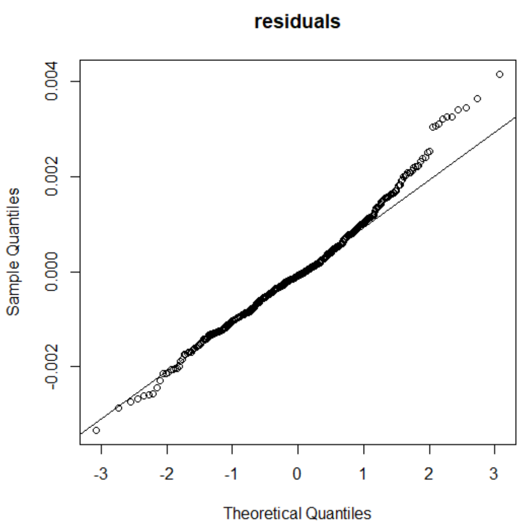
```
> levels(salary$jobcat) #7 levels
[1] "CLERICAL"         "OFFICE TRAINEE"   "SECURITY OFFICER" "COLLEGE TRAINEE"  "EXEMPT E
MPLOYEE"  "MBA TRAINEE"      "TECHNICAL"
```

what we have here is 1 quantitative and 1 categorical with many levels
we want to check:
H0:μ1=μ2=...=μ7
H1:μi!=μj (for some i!=j =1,2,...,κ=7)
let's see first if we can do this parametric check or if we should go non-parametric
 to see this we first check if our residuals are normal:

here we first check if the residuals are normal
first let's see if the sample size of each group >50

```
> table(salary$jobcat)
```

```
        CLERICAL    OFFICE TRAINEE SECURITY OFFICER  COLLEGE TRAINEE   EXEMPT EMPLOYEE 
MBA TRAINEE         TECHNICAL 
             227               136                27                41                32 
5              6 
> #οποτε αφου κάποια n< 50 θα κάνουμε  SW για έλεγχο κανονικότητας καταλοίπων

> anova2 <- aov(relSal~ jobcat , data=salary )
> shapiro.test(anova2$res)

	Shapiro-Wilk normality test

data:  anova2$res
W = 0.98351, p-value = 3.251e-05

#p-value = 3.251e-05<a --> H0 κανονικότητας καταλοίπων rejected


>par(mfrow=c(1,1))
> qqnorm(anova2$residuals , main ="residuals");qqline(anova2$residuals) #heavy tails
```



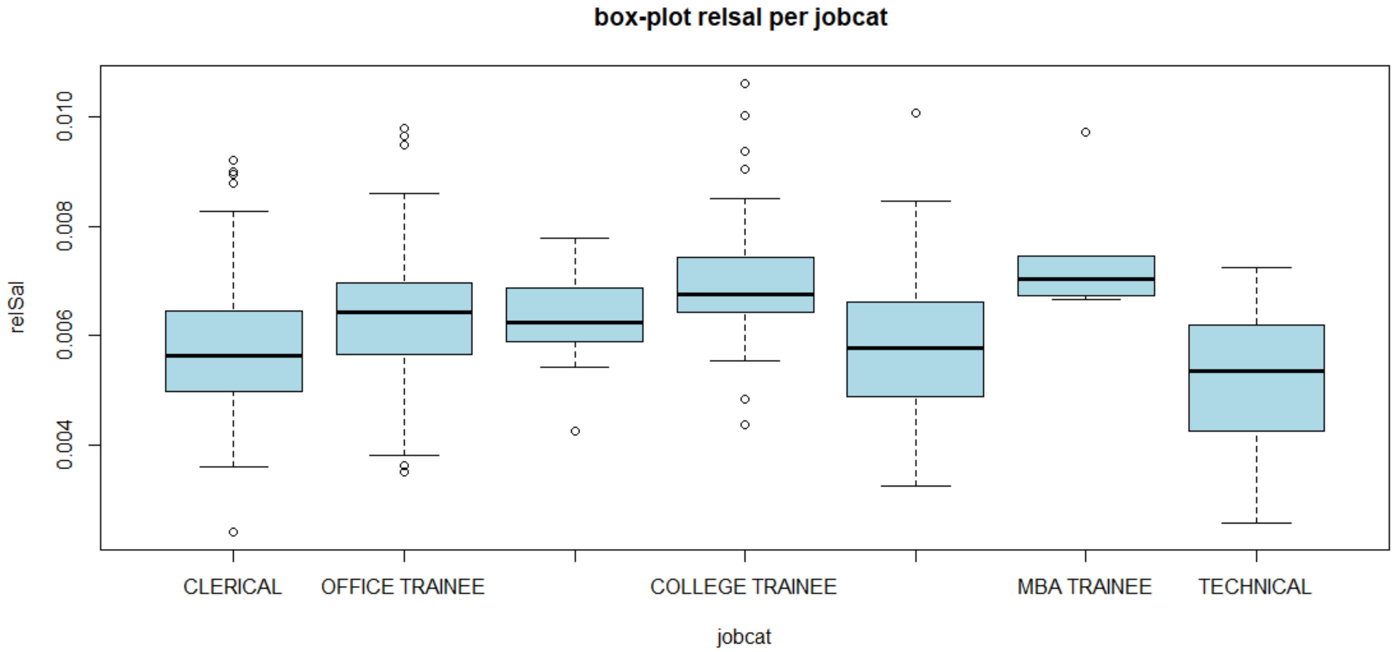
now that the test rejects normality I will continue with the indication that normality will not hold in the residuals

so then I check if the samples are >50, we said they are <50

CONSEQUENTLY WE WILL GO NON-PARAMETRICALLY, WITH A TEST OF EQUALITY OF MEDIANS (KRUSKAL WALLIS TEST)

```
H0:M1=M2=...=M7

H1:Mi!=Mj  ( για κάποιο i!=j =1,2,...,κ=7)


> kruskal.test(relSal~ jobcat , data=salary ) #p-value =1.968e-11<a --> H0 rejected

	Kruskal-Wallis rank sum test

data:  relSal by jobcat
Kruskal-Wallis chi-squared = 61.767, df = 6, p-value = 1.968e-11

This means that there are statistically significant differences in the relSal variable am
ong the different job categories (groups).
In other words, I have evidence to suggest that at least one job category has a different
median relSal value compared to the others.

and now I have to box-plot per group and pairwise.wilcox.test per 2 groups to see which a
re the specific medians that differ
BUT
```

```
> boxplot(relSal ~ jobcat, data = salary, col = "lightblue", main ="box-plot relsal per j
obcat ") # απο τα boxplot παιρνω μια εικόνα αλλα επικαλύπτονται αρκετα
>
```




```
> pairwise.wilcox.test(salary$relSal,salary$jobcat)

        Pairwise comparisons using Wilcoxon rank sum test with continuity correction

data:  salary$relSal and salary$jobcat

                 CLERICAL OFFICE TRAINEE SECURITY OFFICER COLLEGE TRAINEE
OFFICE TRAINEE   1.5e-06  -              -                -
SECURITY OFFICER 0.022    1.000          -                -
COLLEGE TRAINEE  4.0e-07  0.110          0.298            -
EXEMPT EMPLOYEE  1.000    0.276          0.322            0.014
MBA TRAINEE      0.054    0.322          0.322            1.000
TECHNICAL        1.000    0.470          1.000            0.243
                 EXEMPT EMPLOYEE MBA TRAINEE
OFFICE TRAINEE   -               -
SECURITY OFFICER -               -
COLLEGE TRAINEE  -               -
EXEMPT EMPLOYEE  -               -
MBA TRAINEE      0.260           -
TECHNICAL        1.000           0.322

P value adjustment method: holm
Warning messages:
1: In wilcox.test.default(xi, xj, paired = paired, ...) :
  cannot compute exact p-value with ties
2: In wilcox.test.default(xi, xj, paired = paired, ...) :
  cannot compute exact p-value with ties
3: In wilcox.test.default(xi, xj, paired = paired, ...) :
  cannot compute exact p-value with ties
4: In wilcox.test.default(xi, xj, paired = paired, ...) :
  cannot compute exact p-value with ties
```

Συμφωνα με την βιβλιογραφία ισχύει :
"cannot compute exact p-value with ties": This warning is common when using the Wilcoxon Rank Sum Test (or Mann-Whitney U Test) in data that contain ties. Ties occur when two or more values in the data are identical. The Wilcoxon test, being a non-parametric test, ranks the data, and ties can complicate the ranking process. When there are many ties, the

usual method of calculating exact p-values is not appropriate, and an approximation is used instead. This approximation is generally reliable, but it's important to be aware that it's being used.

"P value adjustment method: holm": This part of the output indicates that a Holm adjustment method was applied to the p-values. When conducting multiple comparisons, as in a pairwise test, there's an increased risk of committing a Type I error (false positive). The Holm method is one way to adjust for this by controlling the family-wise error rate, making the test more stringent.

These warnings don't necessarily mean that your results are invalid, but they do suggest that the p-values are approximations due to the presence of ties in your data. This is a common issue in non-parametric tests like the Wilcoxon test, especially with large datasets or datasets with many identical values.

If the presence of ties is a significant concern, or if you're working with a very large dataset, you might consider other statistical approaches or tests that are less sensitive to ties, depending on your specific research questions and data characteristics. However, in many practical scenarios, the approximated p-values provided by the Wilcoxon test with the adjustments for ties are still useful and informative.

Basically, it says that because of the ties, an approximate method has been done and for this reason there may be an error, for this reason it would be good to consult the tykey test

Based on the p-values, the conclusion is that there are statistically significant differences in relSal between the following pairs of job categories:
"OFFICE TRAINEE" and "CLERICAL" ,"OFFICE TRAINEE" and "SECURITY OFFICER" , "COLLEGE TRAINEE" and "CLERICAL"

Βέβαια είναι πιο καλό να πάρω το Turkey test καθως οταν το μέγεθος αν δειγμάτων δεν ίσο ανα τα γρκουπ λειτουργεί καλυτερα

```
> TukeyHSD(anova2)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = relSal ~ jobcat, data = salary)

$jobcat
                                          diff           lwr           upr     p adj
OFFICE TRAINEE-CLERICAL           6.264441e-04  2.621657e-04  0.0009907226 0.0000106
SECURITY OFFICER-CLERICAL         5.551545e-04 -1.287329e-04  0.0012390419 0.1991986
COLLEGE TRAINEE-CLERICAL          1.214102e-03  6.440359e-04  0.0017841678 0.0000000
EXEMPT EMPLOYEE-CLERICAL          1.796686e-04 -4.546749e-04  0.0008140122 0.9807508
MBA TRAINEE-CLERICAL              1.779230e-03  2.604018e-04  0.0032980580 0.0101862
TECHNICAL-CLERICAL               -5.819833e-04 -1.971462e-03  0.0008074957 0.8780142
SECURITY OFFICER-OFFICE TRAINEE  -7.128959e-05 -7.790806e-04  0.0006365014 0.9999426
COLLEGE TRAINEE-OFFICE TRAINEE    5.876577e-04 -1.087486e-05  0.0011861903 0.0581359
EXEMPT EMPLOYEE-OFFICE TRAINEE   -4.467755e-04 -1.106819e-03  0.0002132683 0.4131867
MBA TRAINEE-OFFICE TRAINEE        1.152786e-03 -3.769543e-04  0.0026825259 0.2806179
TECHNICAL-OFFICE TRAINEE         -1.208427e-03 -2.609826e-03  0.0001929711 0.1430631
COLLEGE TRAINEE-SECURITY OFFICER  6.589473e-04 -1.736657e-04  0.0014915603 0.2256209
EXEMPT EMPLOYEE-SECURITY OFFICER -3.754859e-04 -1.253359e-03  0.0005023871 0.8668497
MBA TRAINEE-SECURITY OFFICER      1.224075e-03 -4.115004e-04  0.0028596511 0.2887180
TECHNICAL-SECURITY OFFICER       -1.137138e-03 -2.653357e-03  0.0003790815 0.2862376
EXEMPT EMPLOYEE-COLLEGE TRAINEE  -1.034433e-03 -1.826856e-03 -0.0002420102 0.0024108
MBA TRAINEE-COLLEGE TRAINEE       5.651281e-04 -1.026218e-03  0.0021564738 0.9414864
TECHNICAL-COLLEGE TRAINEE        -1.796085e-03 -3.264484e-03 -0.0003276866 0.0059561
MBA TRAINEE-EXEMPT EMPLOYEE       1.599561e-03 -1.592555e-05  0.0032150482 0.0542858
TECHNICAL-EXEMPT EMPLOYEE        -7.616519e-04 -2.256179e-03  0.0007328751 0.7393329
TECHNICAL-MBA TRAINEE            -2.361213e-03 -4.395435e-03 -0.0003269915 0.0113468
```

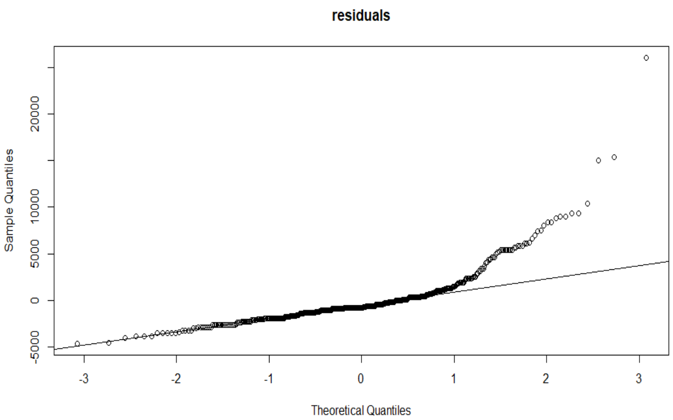
there are statistically significant differences in salary (relSal) between certain pairs of job categories,
such as "OFFICE TRAINEE"(0.0000106) vs. "CLERICAL" and "COLLEGE TRAINEE" vs. "EXEMPT EMPLOYEE."

```
However, some pairs, like "SECURITY OFFICER" vs. "CLERICAL" and "EXEMPT EMPLOYEE" vs. "OF
FICE TRAINEE," do not show statistically significant differences in salary.
```

#8. Cut the AGE variable into four categories according to quantiles. Assign the cut version of AGE into a new variable called age_cut2(Hint: it is a factor). Investigate if, on average, the begging 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.

```
> salary$age_cut2 <- cut(salary$age, quantile(salary$age, probs = c(0, 0.25, 0.5, 0.75, 1
)))
> levels(salary$age_cut2)
[1] "(23,28.5]" "(28.5,32]" "(32,46]"   "(46,64.5]"
> salary$age_cut2 <- cut(salary$age, quantile(salary$age, probs = c(0, 0.25, 0.5, 0.75, 1
)),labels = c("Q1", "Q2", "Q3", "Q4"))
> head(salary)
  id salbeg   sex time   age salnow edlevel  work          jobcat minority     sexrace
age_cut       relSal age_cut2
1  1   8400 MALES   81 28.50  16080      16  0.25 COLLEGE TRAINEE    WHITE WHITE MALES [2
3.0,29.7) 0.005896444       Q1
2  2  24000 MALES   73 40.33  41400      16 12.50 EXEMPT EMPLOYEE    WHITE WHITE MALES [3
9.8,64.5] 0.005757395       Q3
3  3  10200 MALES   83 31.08  21960      15  4.08 EXEMPT EMPLOYEE    WHITE WHITE MALES [2
9.7,39.8) 0.006452038       Q2
4  4   8700 MALES   93 31.17  19200      16  1.83 COLLEGE TRAINEE    WHITE WHITE MALES [2
9.7,39.8) 0.005880376       Q2
5  5  17400 MALES   83 41.92  28350      19 13.00 EXEMPT EMPLOYEE    WHITE WHITE MALES [3
9.8,64.5] 0.004653535       Q3
6  6  12996 MALES   80 29.50  27250      18  2.42 COLLEGE TRAINEE    WHITE WHITE MALES [2
3.0,29.7) 0.006538532       Q2

so I have a quantitative (salbeg) and a categorical salary$age_cut2 with many levels, for
precision 4 levels
we want to check

H0:μ1=μ2=...=μ4
H1:μi!=μj (for some i!=j =1,2,...,κ=4)
let's see first if we can do this parametric control or if we should go non-parametric
to see this we first check if our residuals are normal:
here we first check if the residuals are normal

first let's see if the sample size of each group >50

> table(salary$age_cut2)

 Q1  Q2  Q3  Q4 
120 117 117 119 

n1,n2,n3,n4>50 θα κάνουμε KS + SW για έλεγχο κανονικότητας καταλοίπων

> anova3 <- aov(salbeg~age_cut2  , data=salary )
> shapiro.test(anova3$res) #p-value 2.2e-16 <a --> H0 κανονικότητας καταλοίπων rejected

        Shapiro-Wilk normality test

data:  anova3$res
W = 0.74976, p-value < 2.2e-16

> lillie.test(anova3$res) #p-value 2.2e-16--> H0  rejected

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  anova3$res
```

```
D = 0.19777, p-value < 2.2e-16

> par(mfrow=c(1,1))
> qqnorm(anova3$residuals, main = "residuals");qqline(anova3$residuals) #heavy tails at r
ight
```

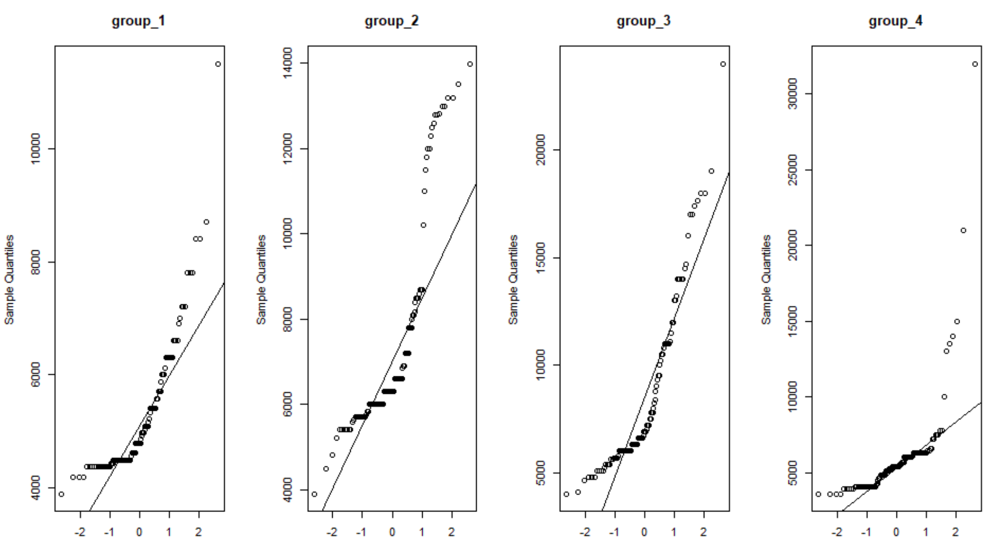


now that the tests have rejected normality i check if the samples are >50, we said they are all >50

Well, now I have to see if the average is a suitable measure of describing the central position for the k groups as well

but let's see it graphically qqplots

```
> group_1<-salary$salbeg[salary$age_cut2=="Q1"]
> group_2<-salary$salbeg[salary$age_cut2=="Q2"]
> group_3<-salary$salbeg[salary$age_cut2=="Q3"]
> group_4<-salary$salbeg[salary$age_cut2=="Q4"]
> par(mfrow=c(1,4))
> qqnorm(group_1,main = "group_1")
> qqline(group_1)
> qqnorm(group_2,main = "group_2")
> qqline(group_2)
> qqnorm(group_3,main = "group_3")
> qqline(group_3)
> qqnorm(group_4,main = "group_4")
> qqline(group_4)

> #απο τα γραφήματα βλεπω οτι τα δεδομενα δεν πλησιαζουν την κανονικότητα σε κανένα group
```




```
> mean_median_skew_kurt(group_1)
```

```
$Mean
[1] 5292.017

$Median
[1] 4800

$Skewness
[1] 2.084333

$Kurtosis
[1] 9.065441

> mean_median_skew_kurt(group_2)
$Mean
[1] 7372.376

$Median
[1] 6300

$Skewness
[1] 1.483292

$Kurtosis
[1] 3.998329

> mean_median_skew_kurt(group_3)
$Mean
[1] 8631.316

$Median
[1] 6900

$Skewness
[1] 1.420774

$Kurtosis
[1] 4.665544

> mean_median_skew_kurt(group_4)
$Mean
[1] 6002.319

$Median
[1] 5400

$Skewness
[1] 5.056125

$Kurtosis
[1] 34.72136


> symmetry.test(group_1)# Ho   rejected ,suggesting that the distribution is NOT symmetri
c.

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

data:  group_1
Test statistic = 7.1046, p-value < 2.2e-16
alternative hypothesis: the distribution is asymmetric.
sample estimates:
bootstrap optimal m
                 18

> agostino.test(group_1)# Ho rejected( reject the distribution is symmetric)

        D'Agostino skewness test

data:  group_1
skew = 2.0843, z = 6.6685, p-value = 2.584e-11
alternative hypothesis: data have a skewness
```

```
> symmetry.test(group_2)#Ho   rejected

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

data:  group_2
Test statistic = 7.9808, p-value < 2.2e-16
alternative hypothesis: the distribution is asymmetric.
sample estimates:
bootstrap optimal m
                 20

> agostino.test(group_2)# Ho   rejected

        D'Agostino skewness test

data:  group_2
skew = 1.4833, z = 5.3360, p-value = 9.503e-08
alternative hypothesis: data have a skewness

> symmetry.test(group_3)#Ho  rejected

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

data:  group_3
Test statistic = 7.1206, p-value < 2.2e-16
alternative hypothesis: the distribution is asymmetric.
sample estimates:
bootstrap optimal m
                 14

> agostino.test(group_3)# Ho rejected

        D'Agostino skewness test

data:  group_3
skew = 1.4208, z = 5.1831, p-value = 2.182e-07
alternative hypothesis: data have a skewness

> symmetry.test(group_4)#Ho rejected

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

data:  group_4
Test statistic = 4.6548, p-value < 2.2e-16
alternative hypothesis: the distribution is asymmetric.
sample estimates:
bootstrap optimal m
                 26

> agostino.test(group_4)# Ho  rejected

        D'Agostino skewness test

data:  group_4
skew = 5.0561, z = 10.2141, p-value < 2.2e-16
alternative hypothesis: data have a skewness
```

I see that in all groups the mean and median are far apart
so the mean is not a suitable measure to describe the central position, so I will go with
a non-parametric Kruskal Wallis test
WITH CHECK OF EQUALITY OF MEDIANS (KRUSKAL WALLIS TEST)

H0:M1=M2=M3=M4
H1:Mi!=Mj (για κάποιο i!=j =1,2,3,4)

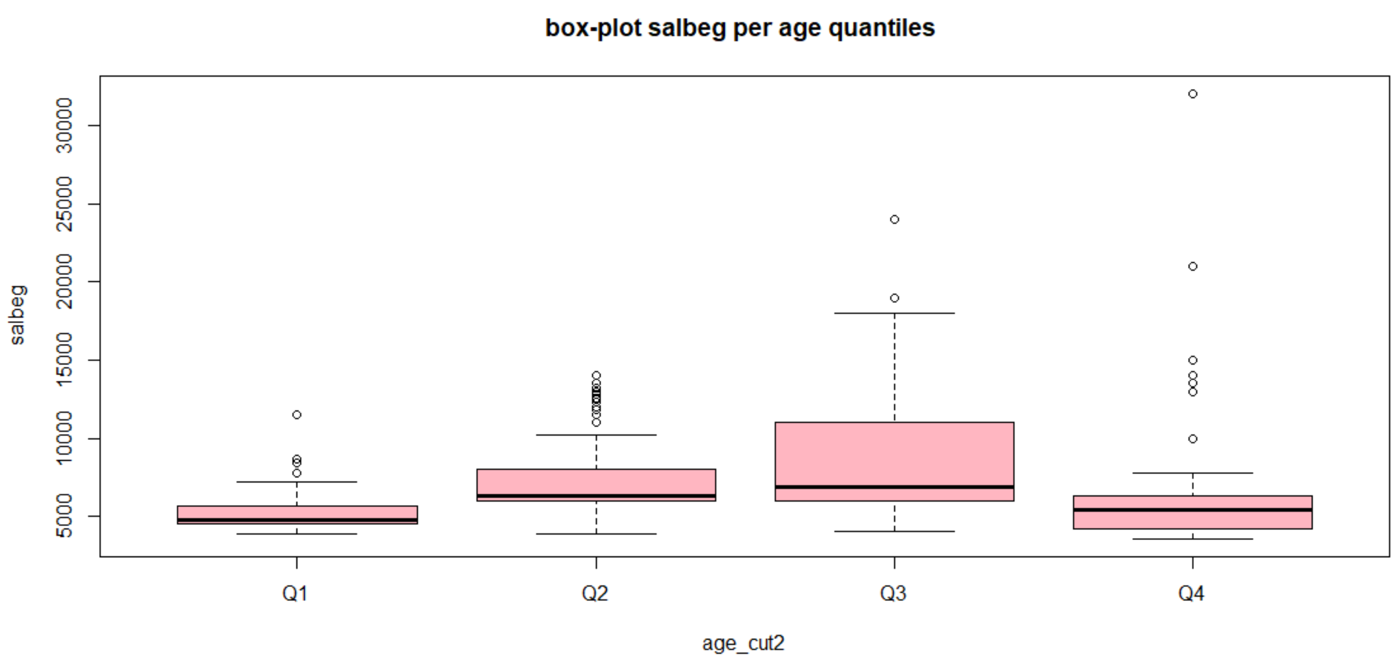
```
> kruskal.test(salbeg~age_cut2 , data=salary ) #p-value <2.2e-16<a --> H0 rejected

        Kruskal-Wallis rank sum test

data:  salbeg by age_cut2
Kruskal-Wallis chi-squared = 143.78, df = 3, p-value < 2.2e-16
```

This means that there are statistically significant differences in the salbeg variable among the different age_cut2 (groups). Levels: (23,28.5] (28.5,32] (32,46] (46,64.5]
In other words, I have evidence to suggest that at least one age_cut2 group has a different median salbeg value compared to the others.

and now I have to box-plot per group and pairwise.wilcox.test per 2 groups to see which are the specific medians that differ

```
> pairwise.wilcox.test(salary$salbeg,salary$age_cut2)

        Pairwise comparisons using Wilcoxon rank sum test with continuity correction

data:  salary$salbeg and salary$age_cut2

   Q1      Q2      Q3
Q2 < 2e-16 -       -
Q3 < 2e-16 0.11    -
Q4 0.15    4.3e-12 1.2e-13

P value adjustment method: holm

> par(mfrow=c(1,1))
> boxplot(salbeg~age_cut2, data = salary, col = "lightpink", main ="box-plot salbeg per age quantiles ")
```

απο τα boxplot παιρνω μια εικόνα αλλα επικαλύπτονται σε κάποια σημεία (έχουν κοινά σημεία)




```
> #Based on the p-values, the conclusion is that
```

there are significant differences in salary between several age_cut2 groups
(Q2 vs Q1 ,Q4 vs Q2,Q4 vs Q3),
however not all pairwise comparisons are significant

- Q3 vs. Q2: The p-value is 0.11, which is not significant . there is no significant difference in begnning salary between the "Q3"(32,46] age group and the "Q2"(28.5,32] age group after adjusting for multiple comparisons.

- Q4 vs. Q1: The p-value is 0.15, which is not significant . there is no significant difference in begging salary between the "Q4"(46,64.5] age group and the "Q1"(23,28.5] age group after adjusting for multiple comparisons.

θα κάνω και Turkey test καθως οταν το μέγεθος αν δειγμάτων δεν ίσο ανα τα γρκουπ λειτουργεί καλυτερα (αν και εδω ειναι σχεδον ίσο αλλα ας το κάνω για σιγουρια)

```
> TukeyHSD(anova3)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = salbeg ~ age_cut2, data = salary)

$age_cut2
             diff        lwr        upr     p adj
Q2-Q1  2080.3594  1113.9479  3046.7709 0.0000003
Q3-Q1  3339.2996  2372.8881  4305.7111 0.0000000
Q4-Q1   710.3027  -251.9881  1672.5934 0.2280913
Q3-Q2  1258.9402   286.4314  2231.4490 0.0050354
Q4-Q2 -1370.0567 -2338.4707  -401.6427 0.0016707
Q4-Q3 -2628.9969 -3597.4109 -1660.5829 0.0000000
```

Q2 (second age group) has a significantly higher average salary compared to Q1 (first age group) with a p-value of 0.0000003.

Q3 (third age group) also has a significantly higher average salary compared to Q1 with a p-value of 0.0000000.

There is no significant difference in average salary between Q4 (fourth age group) and Q1 (p-value = 0.2280913).

Q3 has a significantly higher average salary compared to Q2 with a p-value of 0.0050354.

Q4 has a significantly lower average salary compared to Q2 with a p-value of 0.0016707.

Q4 has a significantly lower average salary compared to Q3 with a p-value of 0.0000000

Generally , The tests agree

#9. 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.

```
I have 2 categorical variables and we want to test for equality of proportions (in indepe
ndent samples)

H0: The proportion of white male employees is equal to the proportion of white female emp
loyees. (independence of sex and color)
H1: The proportion of white male employees is not equal to the proportion of white female
employees. (sex and color dependence)


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

#prop.table(tab1) #total table proportions , i will do it with Cross table to see all the
proportions together


> library(gmodels)
> CrossTable(tab1)


   Cell Contents
|-------------------------|
|                       N |
| Chi-square contribution |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|


Total Observations in Table:  474


             |
             |     WHITE  |  NONWHITE  | Row Total |
-------------|------------|------------|-----------|
       MALES |       194  |        64  |       258 |
             |     0.271  |     0.965  |           |
             |     0.752  |     0.248  |     0.544 |
             |     0.524  |     0.615  |           |
             |     0.409  |     0.135  |           |
-------------|------------|------------|-----------|
     FEMALES |       176  |        40  |       216 |
             |     0.324  |     1.153  |           |
             |     0.815  |     0.185  |     0.456 |
             |     0.476  |     0.385  |           |
             |     0.371  |     0.084  |           |
-------------|------------|------------|-----------|
Column Total |       370  |       104  |       474 |
             |     0.781  |     0.219  |           |
-------------|------------|------------|-----------|


> # έλεγχος expected values >5

> chisq.test(tab1)$expected  # all ok

             WHITE  NONWHITE
  MALES   201.3924 56.60759
  FEMALES 168.6076 47.39241
```

```
> chisq.test(tab1,correct = F) # χ^2 τεστ p-value = 0.09948 DOESNT REJECT H0

	Pearson's Chi-squared test

data:  tab1
X-squared = 2.7139, df = 1, p-value = 0.09948


> #το chiq.test είναι προσσέγγιση του fisher.tesτ
```

```
there is no significant difference in the proportions of white male employees and white f
emale employees.
In other words, there is no significant association between gender (males or females) and
minority status (white or nonwhite).
and in other words ,there is no significant difference in the odds of being a white male
compared to being a white female.
```

#10. By making use of the factor variable minority, investigate if there are differences in the proportions among the job categories.

```
I have 2 categorical variables and we want to check for equality of proportions
H0: The proportions of minority and non-minority employees among different job categories
are the same, (there is no association between job categories and minority status).
H1: The proportions of minority and non-minority employees among different job categories
are different, (there is an association between job categories and minority status).
```

```
> tab2 <- table(salary$jobcat, salary$minority)
>
#prop.table(tab2) #total table proportions

> library(gmodels)
> CrossTable(tab2)


   Cell Contents
|-------------------------|
|                       N |
| Chi-square contribution |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|


Total Observations in Table:  474


                 |
                 |     WHITE  |  NONWHITE  | Row Total |
-----------------|------------|------------|-----------|
        CLERICAL |       160  |        67  |       227 |
                 |     1.668  |     5.936  |           |
                 |     0.705  |     0.295  |     0.479 |
                 |     0.432  |     0.644  |           |
                 |     0.338  |     0.141  |           |
-----------------|------------|------------|-----------|
  OFFICE TRAINEE |       116  |        20  |       136 |
                 |     0.912  |     3.245  |           |
                 |     0.853  |     0.147  |     0.287 |
                 |     0.314  |     0.192  |           |
                 |     0.245  |     0.042  |           |
-----------------|------------|------------|-----------|
SECURITY OFFICER |        14  |        13  |        27 |
```

```
                 |      2.376 |      8.452 |            |
                 |      0.519 |      0.481 |      0.057 |
                 |      0.038 |      0.125 |            |
                 |      0.030 |      0.027 |            |
-----------------|------------|------------|------------|
 COLLEGE TRAINEE |         40 |          1 |         41 |
                 |      1.998 |      7.107 |            |
                 |      0.976 |      0.024 |      0.086 |
                 |      0.108 |      0.010 |            |
                 |      0.084 |      0.002 |            |
-----------------|------------|------------|------------|
 EXEMPT EMPLOYEE |         30 |          2 |         32 |
                 |      1.009 |      3.591 |            |
                 |      0.938 |      0.062 |      0.068 |
                 |      0.081 |      0.019 |            |
                 |      0.063 |      0.004 |            |
-----------------|------------|------------|------------|
     MBA TRAINEE |          4 |          1 |          5 |
                 |      0.002 |      0.009 |            |
                 |      0.800 |      0.200 |      0.011 |
                 |      0.011 |      0.010 |            |
                 |      0.008 |      0.002 |            |
-----------------|------------|------------|------------|
       TECHNICAL |          6 |          0 |          6 |
                 |      0.370 |      1.316 |            |
                 |      1.000 |      0.000 |      0.013 |
                 |      0.016 |      0.000 |            |
                 |      0.013 |      0.000 |            |
-----------------|------------|------------|------------|
    Column Total |        370 |        104 |        474 |
                 |      0.781 |      0.219 |            |
-----------------|------------|------------|------------|
```

```
> # έλεγχος expected values >5

> chisq.test(tab2)$expected  # not ok

                      WHITE  NONWHITE
  CLERICAL         177.194093 49.805907
  OFFICE TRAINEE   106.160338 29.839662
  SECURITY OFFICER  21.075949  5.924051
  COLLEGE TRAINEE   32.004219  8.995781
  EXEMPT EMPLOYEE   24.978903  7.021097
  MBA TRAINEE        3.902954  1.097046
  TECHNICAL          4.683544  1.316456
Warning message:
In chisq.test(tab2) : Chi-squared approximation may be incorrect
```

```
prop.test implements the Pearson's chi-square statistics for independence p-value however
here i take
Warning message:
In prop.test(tab2) : Chi-squared approximation may be incorrect and this is why the expec
ted values of each cell are probably not greater than 5
so I can do either chisq.test with carlo simulation mode, or Fisher (surely here it will
give me an error because it takes all the possible tables
and it is not done due to size)
> fisher.test(tab2)
Error in fisher.test(tab2) :
  FEXACT error 7(location). LDSTP=18570 is too small for this problem,
  (pastp=59.7129, ipn_0:=ipoin[itp=509]=18372, stp[ipn_0]=55.1032).
Increase workspace or consider using 'simulate.p.value=TRUE'
```

carlo simulation mode : This is a more accurate method when the assumptions for the Chi-squared test ar e not met because it does not rely on the theoretical distribution of the test statistic under the null hypothesis. I nstead, it generates a distribution by simulating many datasets under the null hypothesis and calculating the te st statistic for each one. This simulated distribution is then used to estimate the p-value.

```
>
> chisq.test(tab2,simulate.p.value = T) #  p-value = 0.0004998 REJECT H0
```

```
        Pearson's Chi-squared test with simulated p-value (based on 2000 replicates)

data:  tab2
X-squared = 37.991, df = NA, p-value = 0.0004998


> fisher.test(tab2,simulate.p.value = T)

        Fisher's Exact Test for Count Data with simulated p-value (based on 2000 replicate
s)

data:  tab2
p-value = 0.0004998
alternative hypothesis: two.sided
```

So , there are statistically significant differences in the proportions of
minority and non-minority employees among the various job categories in the dataset.