# GOODNESS OF FIT USING SPSS

# WHAT IS GOODNESS OF FIT ?

A goodness-of-fit is a statistical technique . It is Applied to measure “how well the actual(observed) data points Fit into a Machine Learning model ” . It summarizes the divergence between actual observed data points and expected data points in context to a statistical or machine learning model.

Assessment of divergence between the observed data points and model predicted data points is critical to understand , a decision made o poorly fitting models might be badly misleading. A seasoned practitioner must examine the fitment of actual and model predicted data points.

# WHY DO WE TEST GOODNESS OF FIT ?

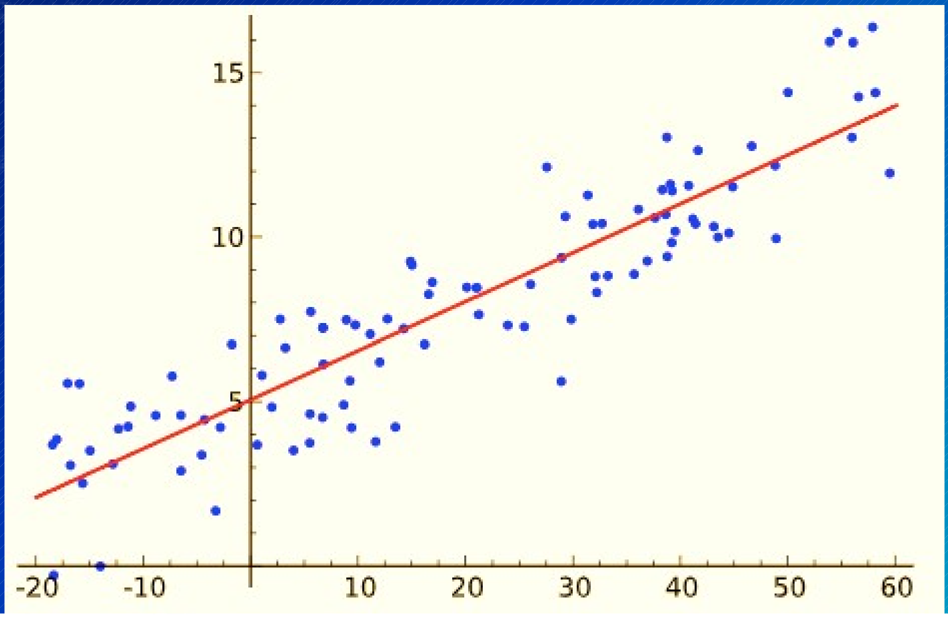
Goodness-of-fit tests are statistical tests to determine values match those predicted by the model . Goodness-of-fit tests are frequently applied in business decision making .

# WHAT ARE THE MOST COMMMON GOODNESS OF FIT TEST ?

There are multiple methods for determining goodness-of-fit. Some of the most popular methods used in statistics include the chi-square, the Kolmogorov - Smirnov test, the Anderson-Darling test and the Shipiro - Wilk test.

# FOR EXAMPLE :

The below image depict the linear regression function . The Goodness-of-fit tests here will compare the actual observed values denoted by bule dots to the predicted vaules denoted by the red regression line .



# EXPLAIN TWO TEST ONLY

“

1 . CHI – SQUARE TEST GOODNESS OF FIT USING SPSS

2 . NORMAL DISTRIBUTION FIT  USING SPSS

# 1. CHI – SQUARE TEST FOR GOODNESS OF FIT USING SPSS

Chi-Square goodness of fit test is a non-parametric test that is used to find out how the observed value of a given phenomena is significantly different from the expected value. In Chi-Square goodness of fit test, the term goodness of fit is used to compare the observed sample distribution with the expected probability distribution. Chi-Square goodness of fit test determines how well theoretical distribution (such as normal, binomial, or Poisson) fits the empirical distribution. In Chi-Square goodness of fit test, sample data is divided into intervals. Then the numbers of points that fall into the interval are compared, with the expected numbers of points in each interval.

# The chi – square test for a goodness of fit test is

$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

$x^2_c$ =chi – square goodness of fit test

$O_i$ **=an obsrved count for bin** $i$

$E_i$ **=an expected count for bin** $i$ **, asserted by the null hypothesis.**

# The Expected frequency is calculated by

$$E_i = \left(F(Y_u) - F(Y_l)\right) N$$

F =the cumulative distribution function for the probability distribution being tested

$Y_u$ =the upper limit for class I

$Y_i$ =the lower limit for class i , and

N =the sample size

# Application of Chi-square as the goodness of fit

The Chi-square is applied to establish or refute that a relationship exists between actual observed values and predicted values. The chi-squared test is a very useful tool for predictive analytics professionals. It is used very commonly in Clinical research, Social sciences, and Business research .

It is also right tail test.

# Procedure for Chi-Square Goodness of Fit Test:

## HYPOTHESIS :

**Null Hypothesis : $H_0$ There is no diference between observed and expected hypothesis**

**Alternative Hypothesis : $H_1$ There is a significant difference between observed and expected hypothesis**

# Example of chi – square goodness of fit test in SPSS

A shop owner claims that an epual numbers of customers come into his shop each weekday. To test his Hypothesis , a researcher records the number of customers that come into the shop on a given week and Find the following.

Monday-50 customers
Tuesday-60 customers
Wednesday-40 customers
Thursday-47 customers
Friday-53 customers

Use the following steps to perform a Chi-Square goodness of fit test in SPSS to Determine if the data is consistent with the shop owner's claims.

# Procedure for chi-square test

Step 1 : Open SPSS softwear and select variable view and enter the variables.

Step 2 : Select the data view and enter the data.

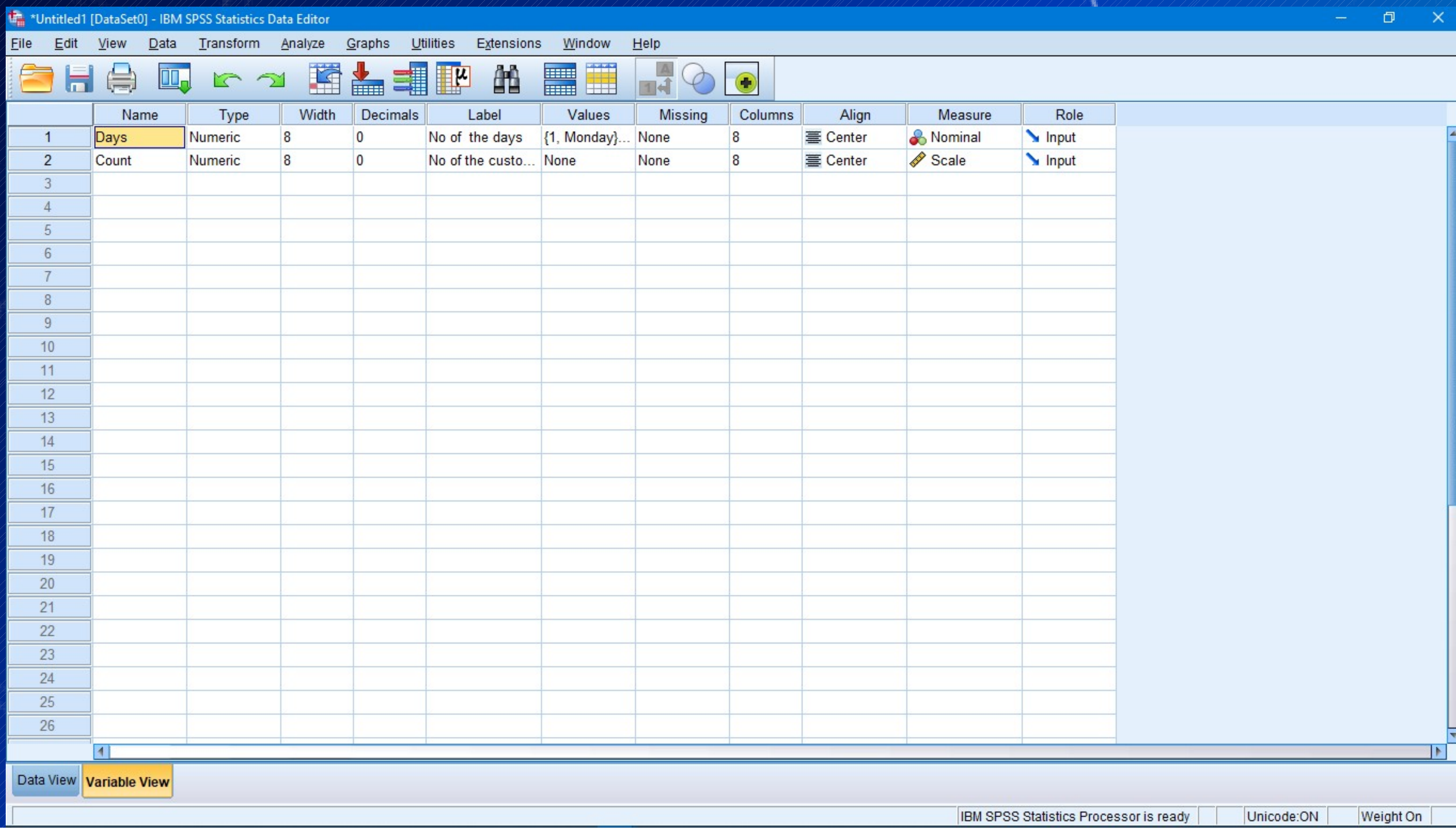
Step 3 : Select the analysis and click Non-parametric test.

Step 4 : Click legacy dialogs and select chi-square and transform the test variables list box into the no.of.customers.

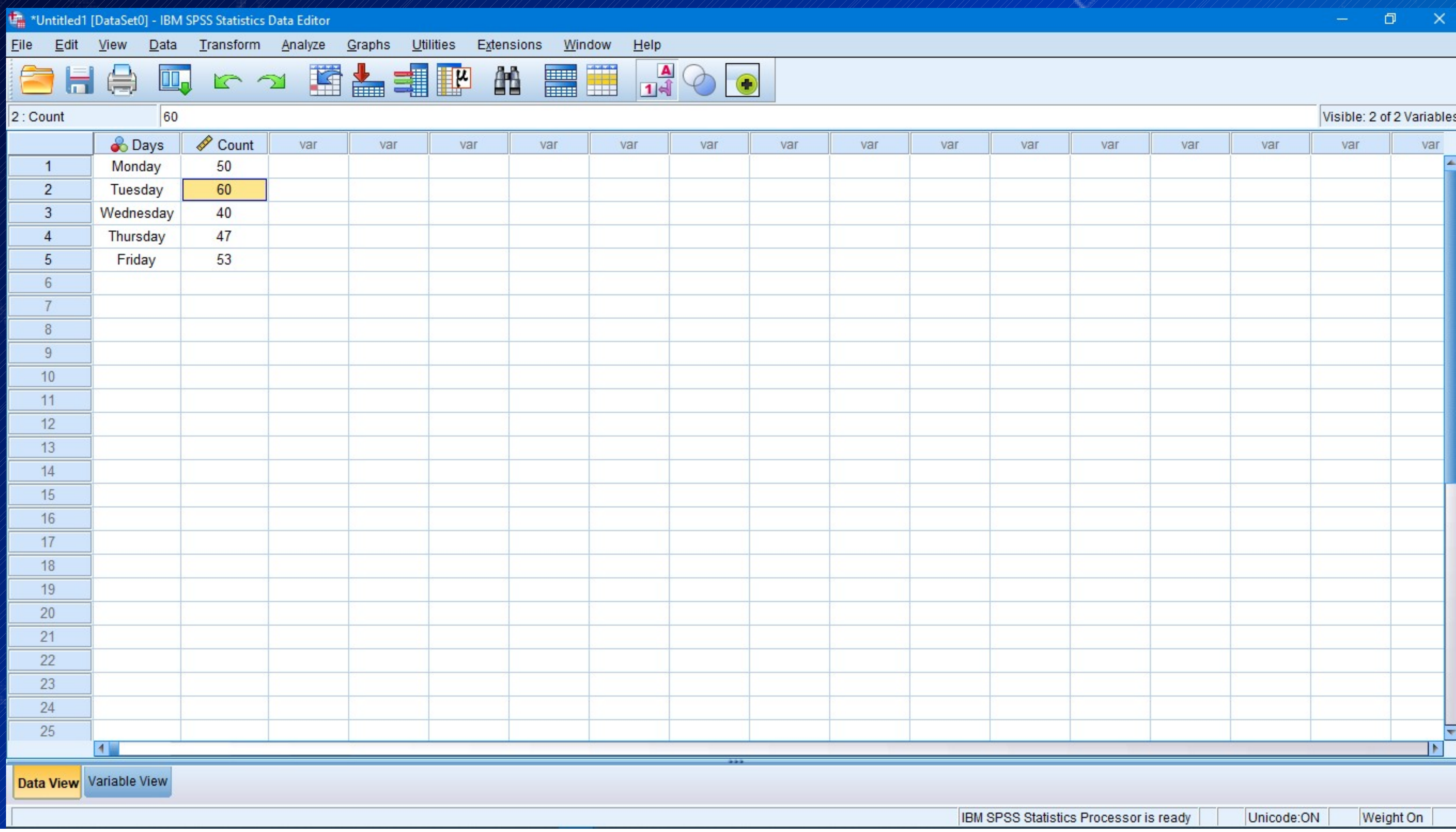
Step 5 : Select option and tick the discriptive and select continue.

Step 6 : Finally click ok to get output.

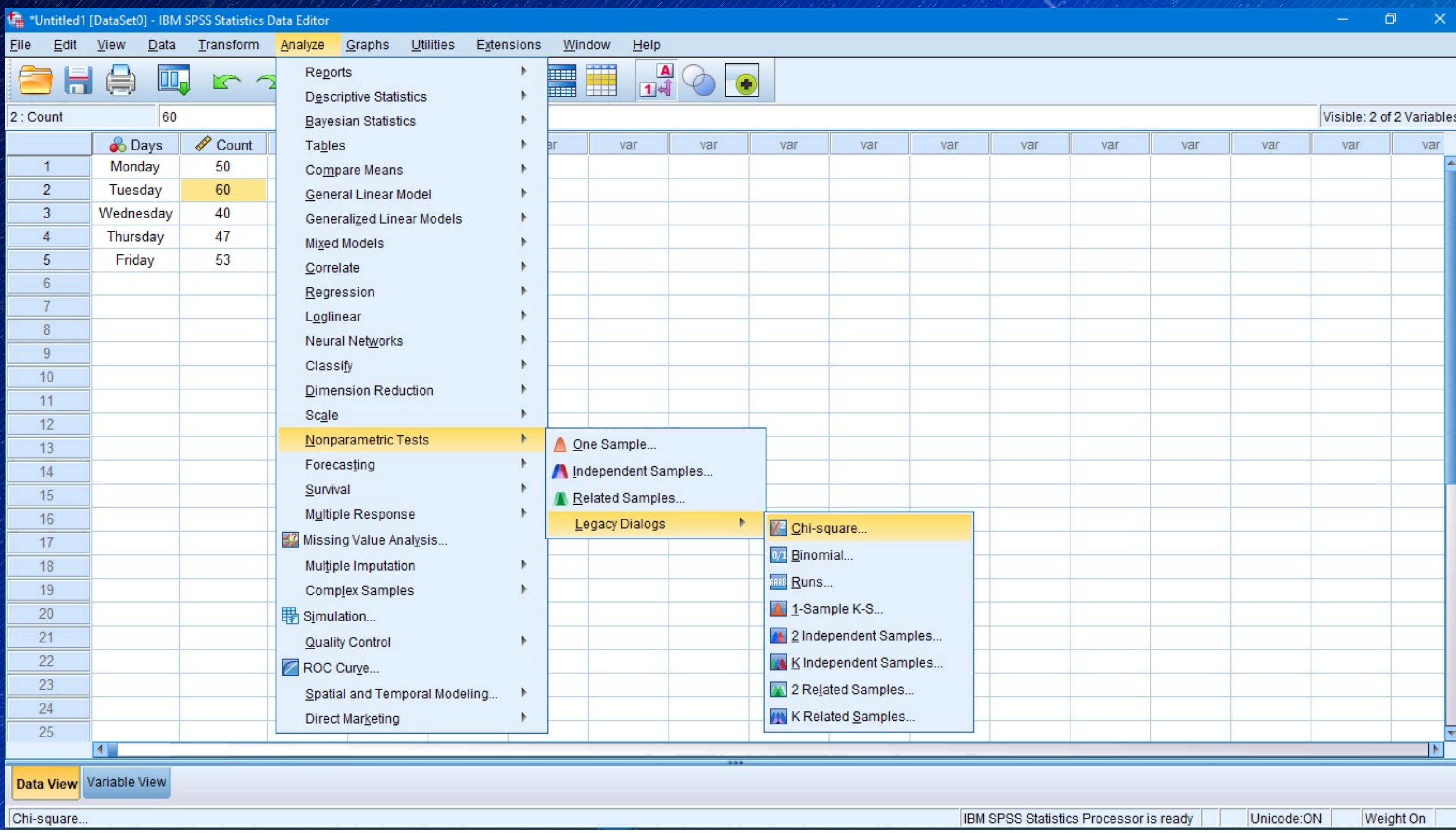
# Step 1 : Enter the variable view and enter variables in SPSS



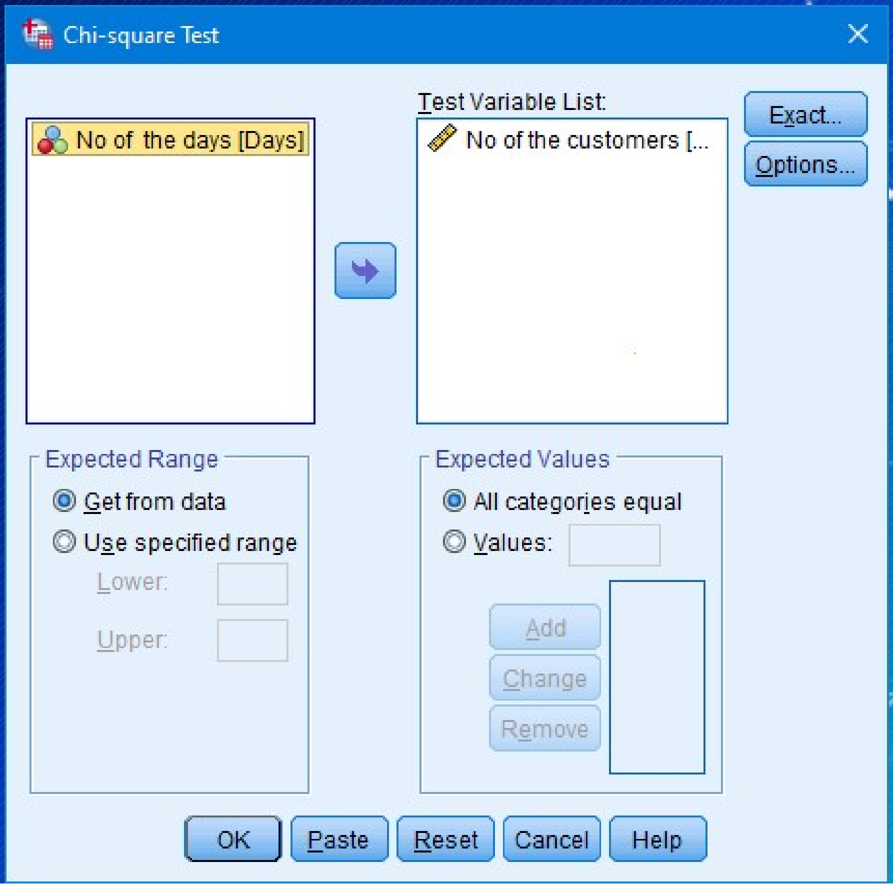
# Step 2 : Select data view and enter data in SPSS



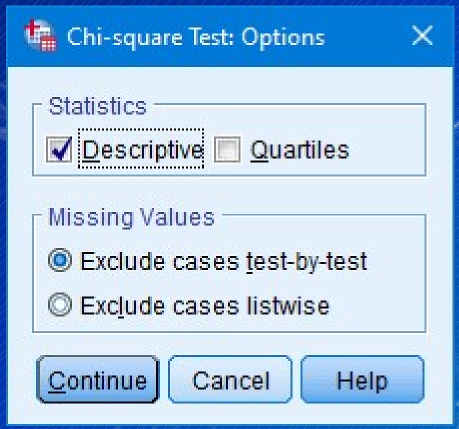
# Step 3 : Select the analysis and click Non-parametric test and Click legacy dialogs and select chi-square.



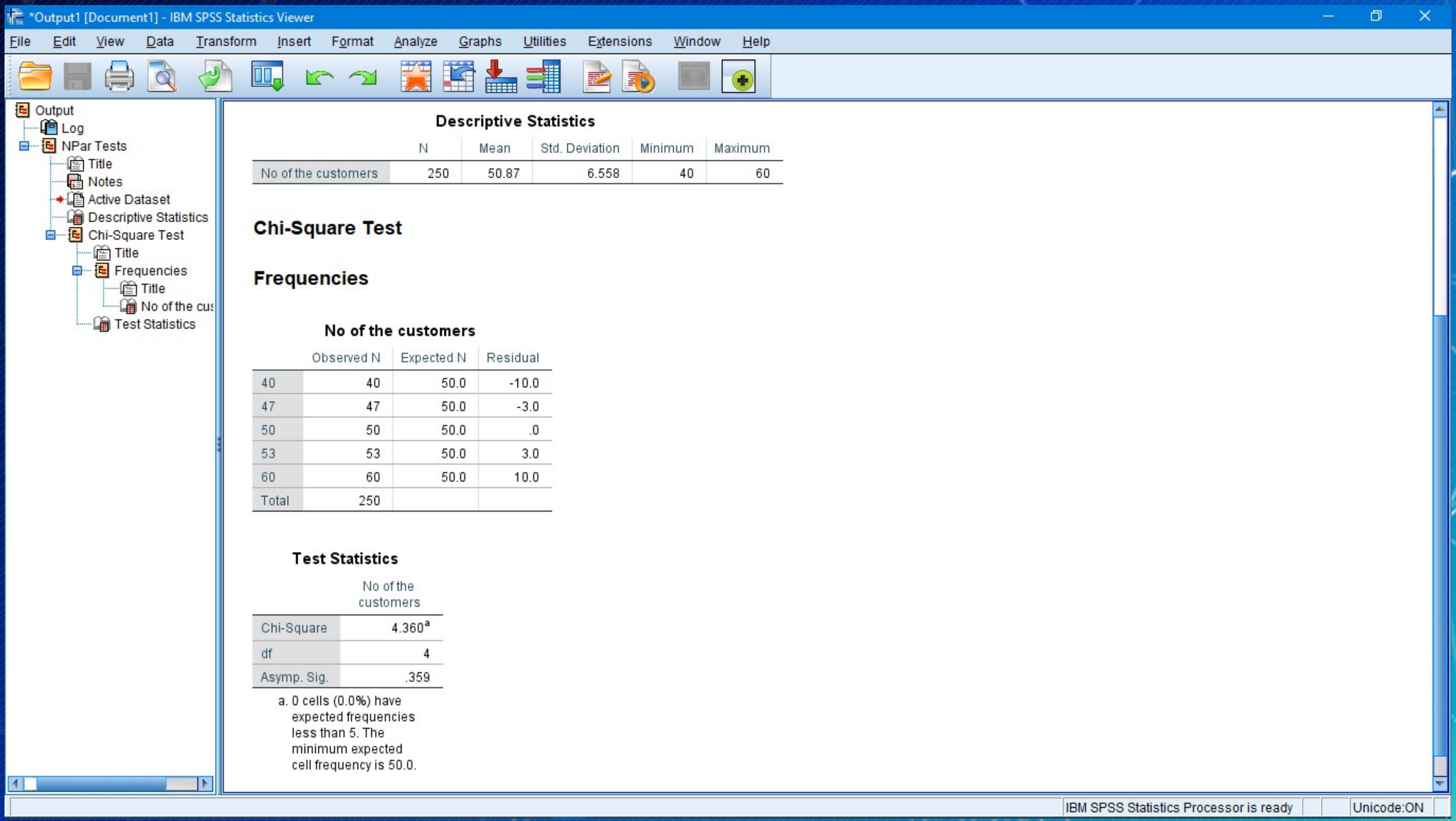
# Step 4 : transform the test variables list box into the no . Of . customers



Step 5 : Select option and tick the discriptive and select continue and click ok.



# OUTPUT :


*Output1 [Document1] - IBM SPSS Statistics Viewer
File Edit View Data Transform Insert Format Analyze Graphs Utilities Extensions Window Help
Output
Log
NPar Tests
Title
Notes
Active Dataset
Descriptive Statistics
Chi-Square Test
Title
Frequencies
Title
No of the cus
Test Statistics
Descriptive Statistics
N | Mean | Std. Deviation | Minimum | Maximum
No of the customers | 250 | 50.87 | 6.558 | 40 | 60
Chi-Square Test
Frequencies
No of the customers
Observed N | Expected N | Residual
40 | 40 | 50.0 | -10.0
47 | 47 | 50.0 | -3.0
50 | 50 | 50.0 | .0
53 | 53 | 50.0 | 3.0
60 | 60 | 50.0 | 10.0
Total | 250
Test Statistics
No of the customers
Chi-Square | 4.360a
df | 4
Asymp. Sig. | .359
a. 0 cells (0.0%) have expected frequencies less than 5. The minimum expected cell frequency is 50.0.
IBM SPSS Statistics Processor is ready
Unicode:ON

# Conclusion :

**Chi-Square:** The Chi-Square test statistic, found to be 4.36.

**df :** The degrees of freedom, calculated as #categories-1 = 5-1 = 4.

**Asymp. Sig:** The p-value that corresponds to a Chi-Square value of 4.36 with 4 degrees of freedom, found to be .359. This value can also be found by using the Chi-Square Score to P Value Calculator.

Since the p-value (.359) is not less than 0.05, we fail to reject the null hypothesis. This means we do not have sufficient evidence to say that the true distribution of customers is different from the distribution that the shop owner claimed

# 2 . NORMAL DISTRIBUTION FIT USING IN SPSS

Normality test using SPSS :

An normality test is used to determine whether sample data has been drawn from a normal distribute population.

What is Normal distribution ?

The normal distribution is always symmetrical about the mean which look like a “bell curve”.

When testing for normality:

- Probabilities > 0.05 indicate that the data are normal.
- Probabilities < 0.05 indicate that the data are NOT normal.

# Normal Distribution Formula:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

where

σ is a population standard deviation;
μ is a population mean;
x is a value or test statistic;
e is a mathematical constant of roughly 2.72;
π a mathematical constant of roughly 3.14.

**The following numerical and visual output must be investigated :**

- Skewness and kurtosis z-values

(should be somewhere in the span of -1.96 to +1.96)

- The shapiro-wilk test p-value

(should be above 0.05)

- Histograms , normal Q-Q plots and Box plots

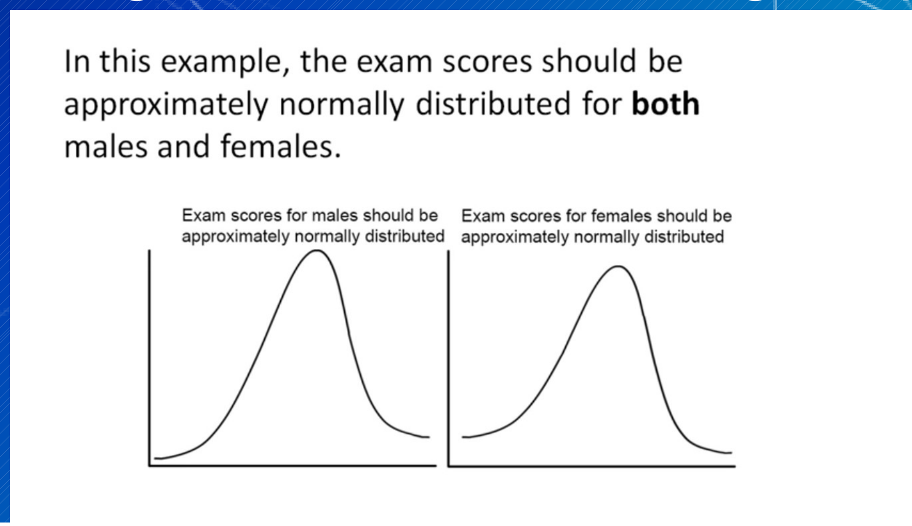
(should be indicated that our data are approximately normal Distributed).

**In a many statistical analysis,there are dependent variables and independent variable :**

Dependent variable= a that variable depend on other factors.

For example : exam scores,as a variable,may be change depending on the student gender.

Independent variable= a variable that does not depend on the other factor.For example : gender does not change depending on exam scores.



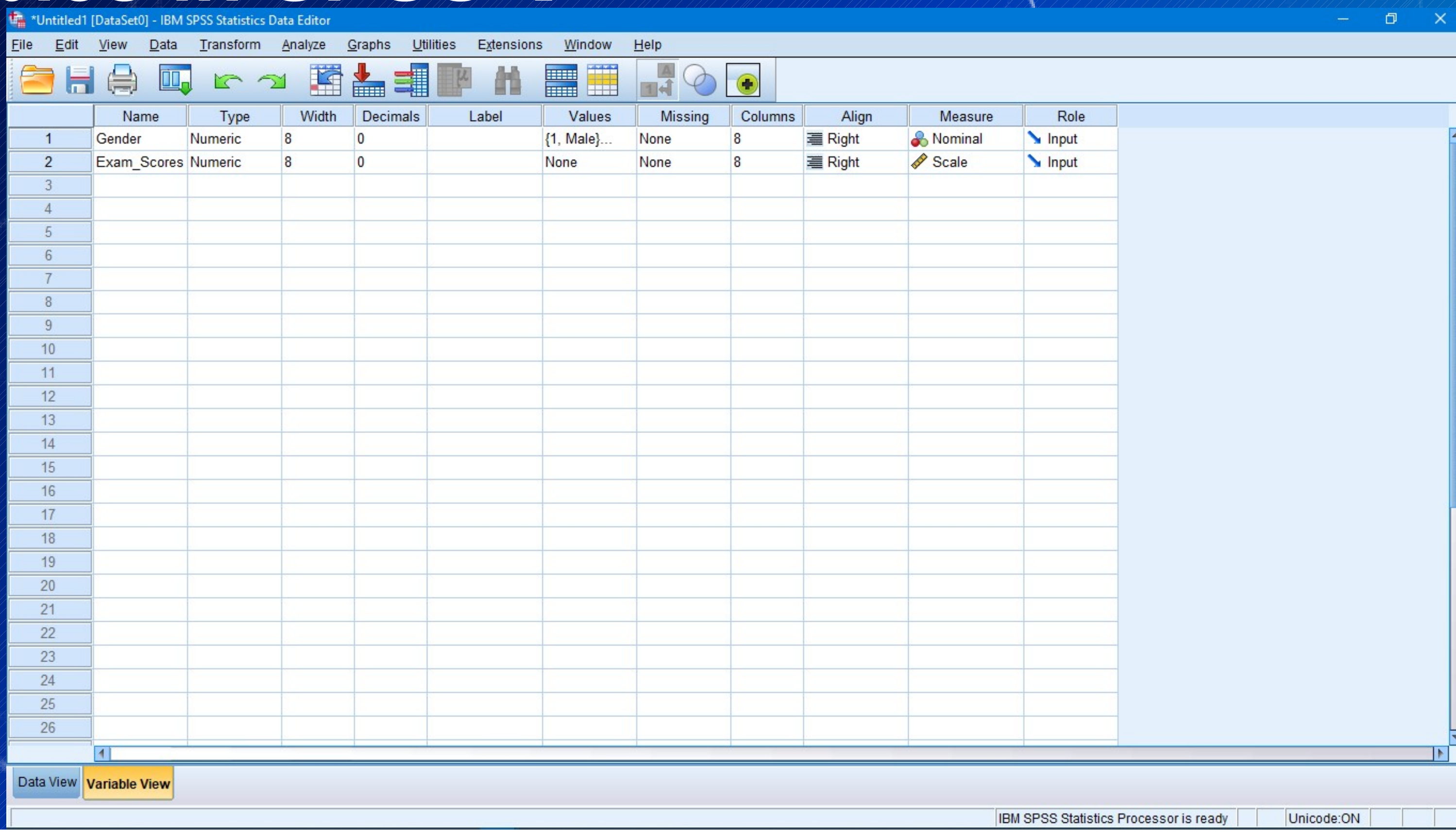
# Example:

The students

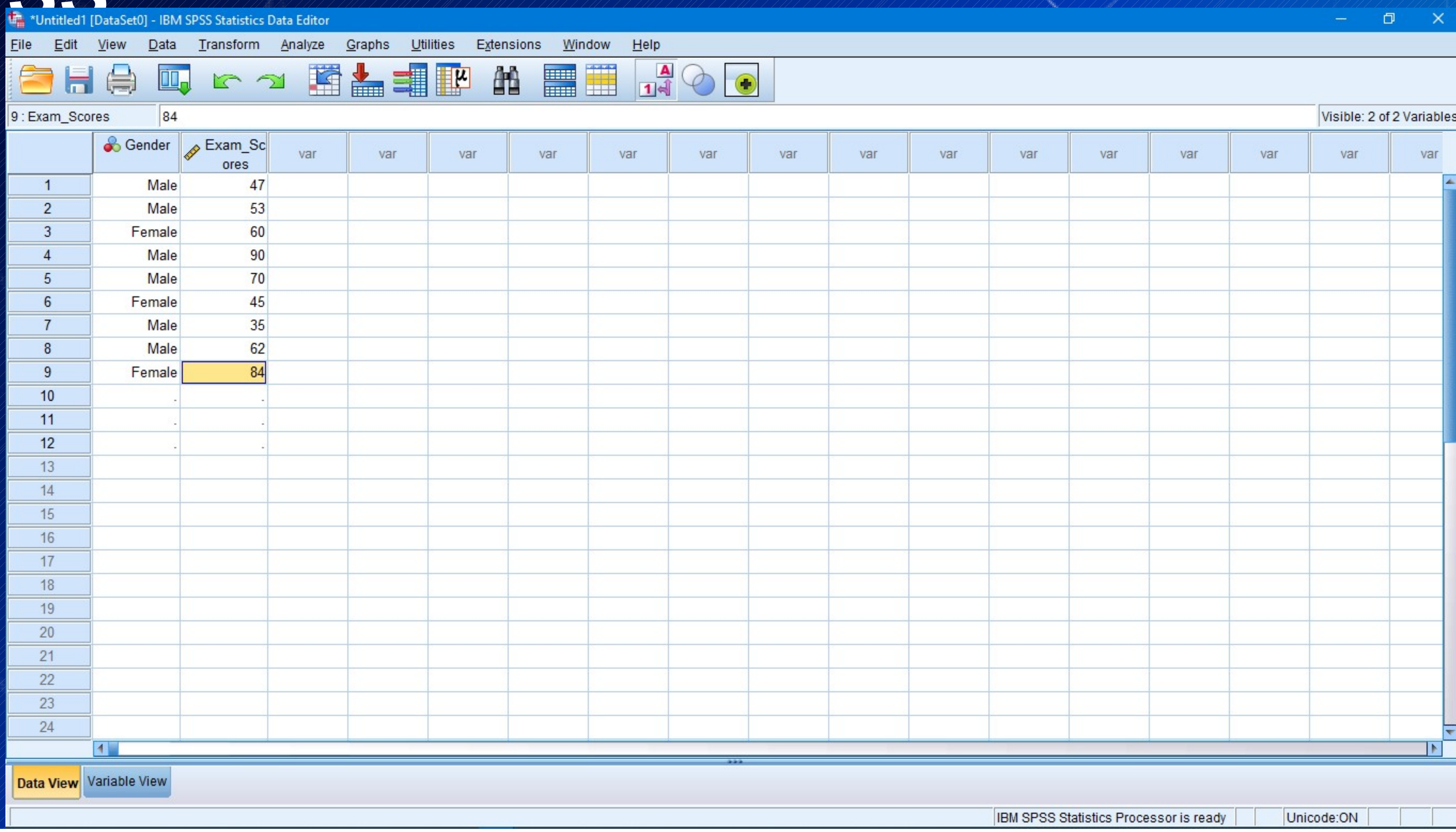
Gender = male and female

Exam scores = 47,53,60,90,70,45,35,62,84,

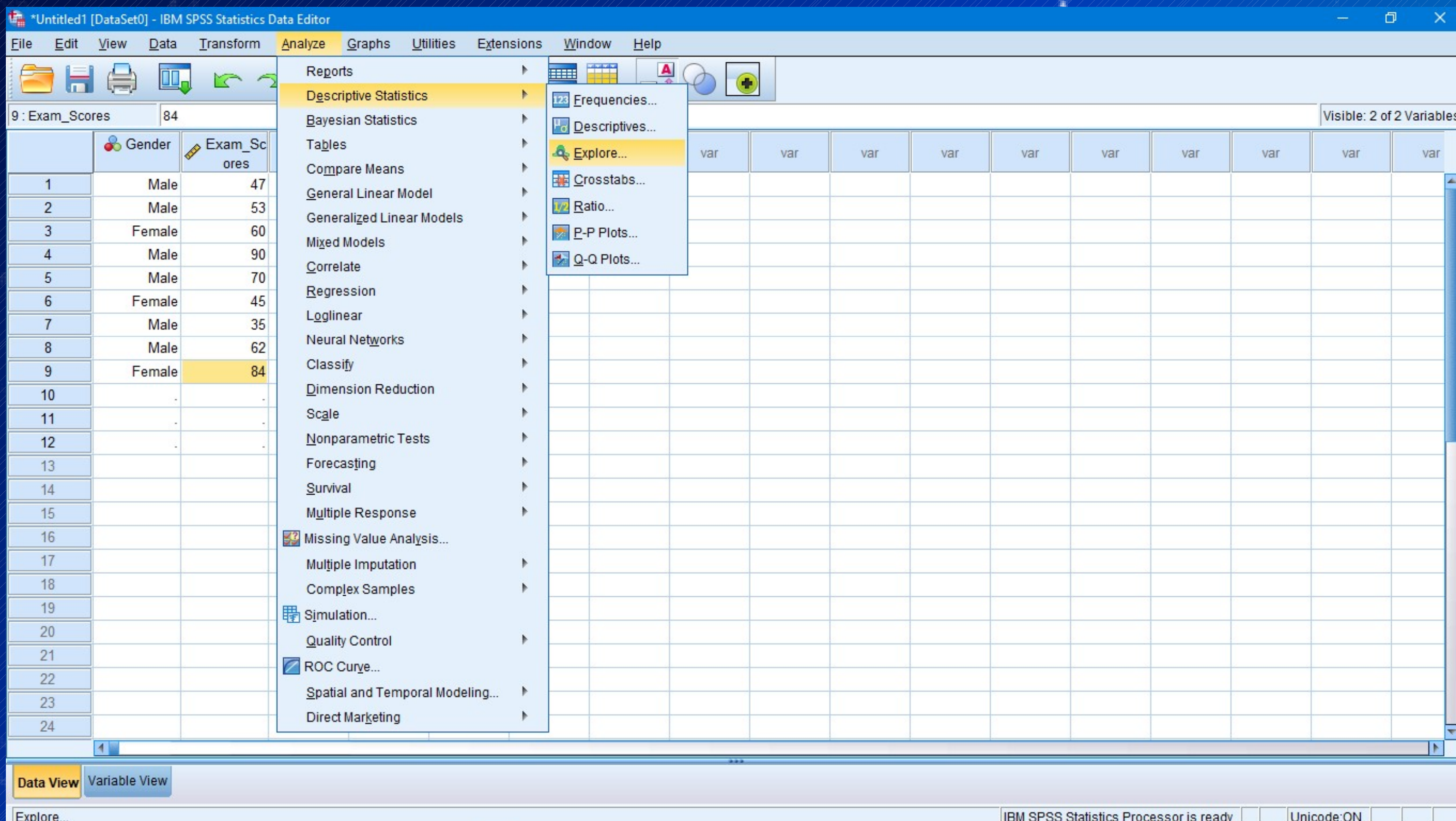
# Step 1：Enter the variable view and enter variables in SPSS：



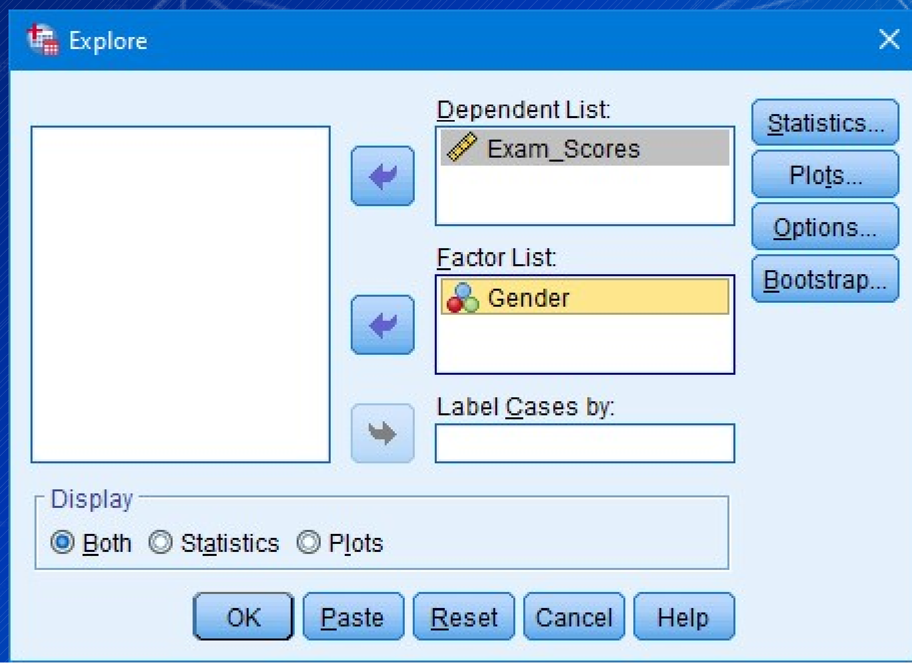
# Step 2 : Select data view and enter data in SPSS



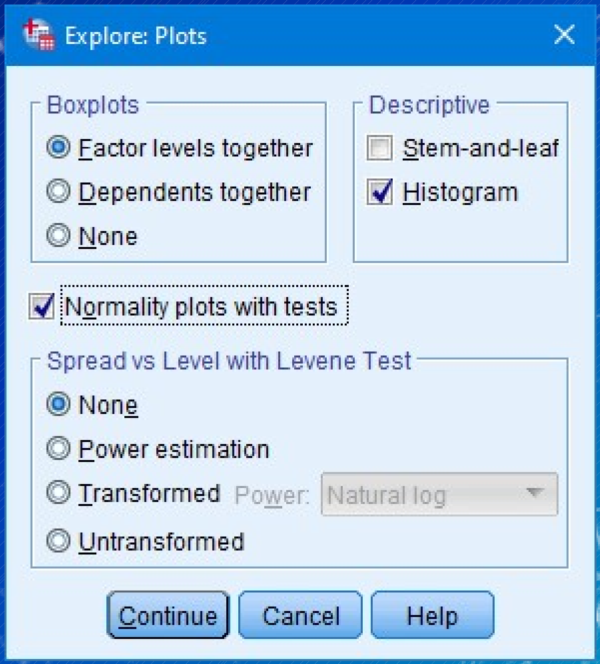
# Step 3 : Select the analysis and click discriptive statistics and select explore :



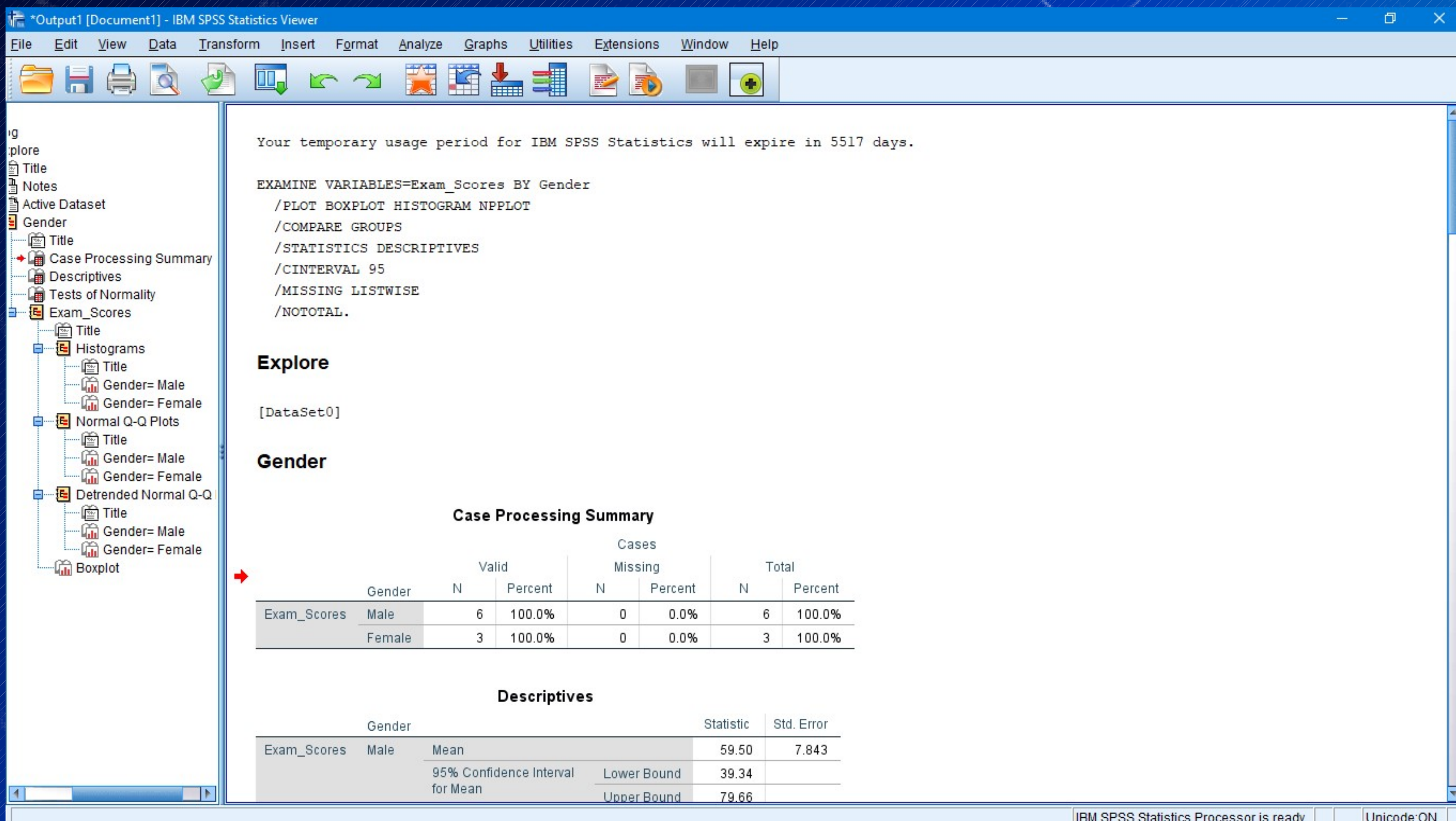
Step 4： transform the Dependent list box into the Exam_scores and transform the Factor list box into the Gender



# Step 5 : Open polts and ticks histogram and normality plots with test and continue then click ok .



# OUTPUT :


*Output1 [Document1] - IBM SPSS Statistics Viewer

File Edit View Data Transform Insert Format Analyze Graphs Utilities Extensions Window Help

Your temporary usage period for IBM SPSS Statistics will expire in 5517 days.

```
EXAMINE VARIABLES=Exam_Scores BY Gender
  /PLOT BOXPLOT HISTOGRAM NPPLOT
  /COMPARE GROUPS
  /STATISTICS DESCRIPTIVES
  /CINTERVAL 95
  /MISSING LISTWISE
  /NOTOTAL.
```

**Explore**

[DataSet0]

**Gender**

**Case Processing Summary**

| | | Cases | | | | | |
|---|---|---|---|---|---|---|---|
| | | Valid | | Missing | | Total | |
| | Gender | N | Percent | N | Percent | N | Percent |
| Exam_Scores | Male | 6 | 100.0% | 0 | 0.0% | 6 | 100.0% |
| | Female | 3 | 100.0% | 0 | 0.0% | 3 | 100.0% |

**Descriptives**

| | Gender | | | Statistic | Std. Error |
|---|---|---|---|---|---|
| Exam_Scores | Male | Mean | | 59.50 | 7.843 |
| | | 95% Confidence Interval for Mean | Lower Bound | 39.34 | |
| | | | Upper Bound | 79.66 | |

*Output1 [Document1] - IBM SPSS Statistics Viewer

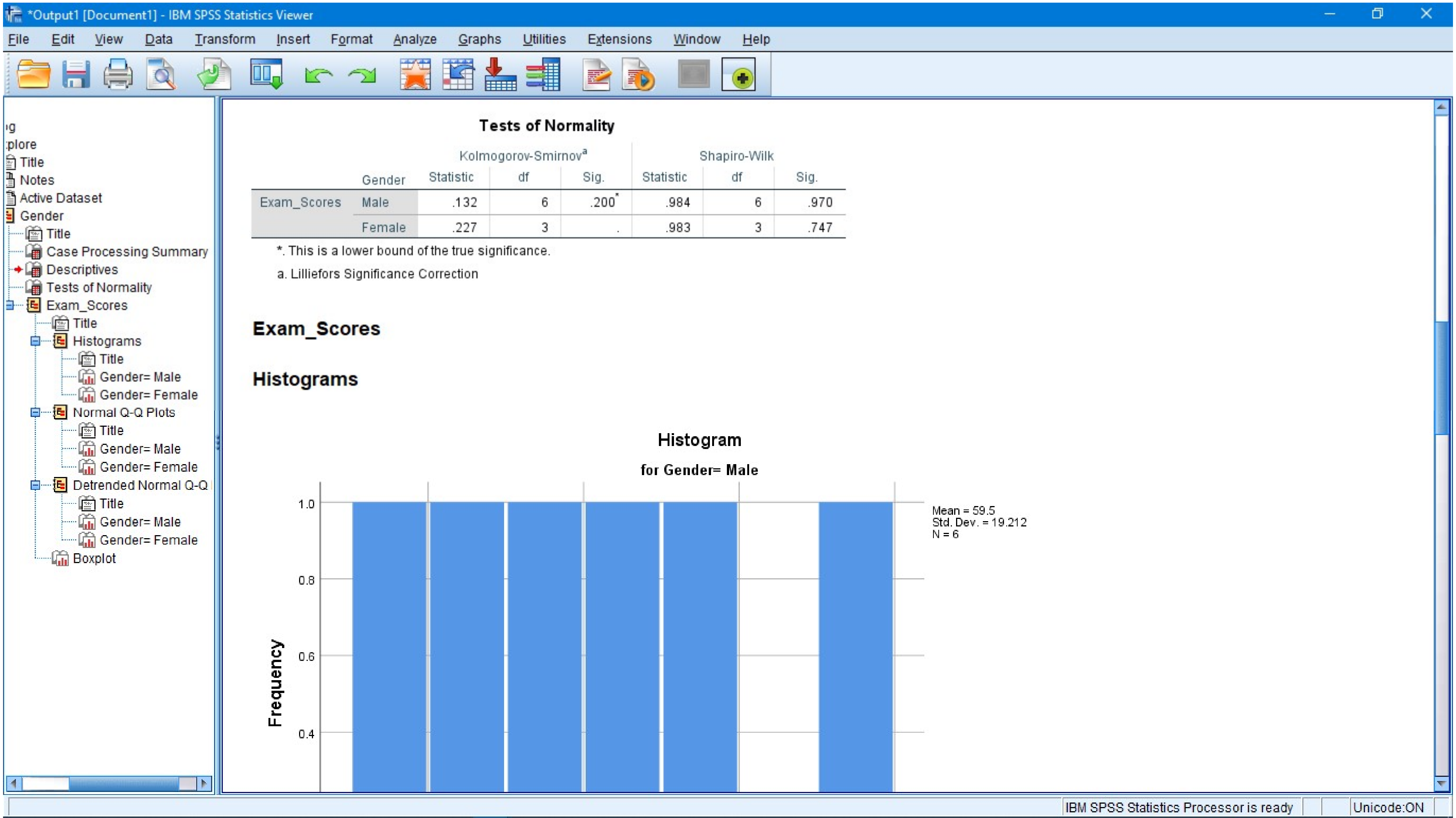
File Edit View Data Transform Insert Format Analyze Graphs Utilities Extensions Window Help

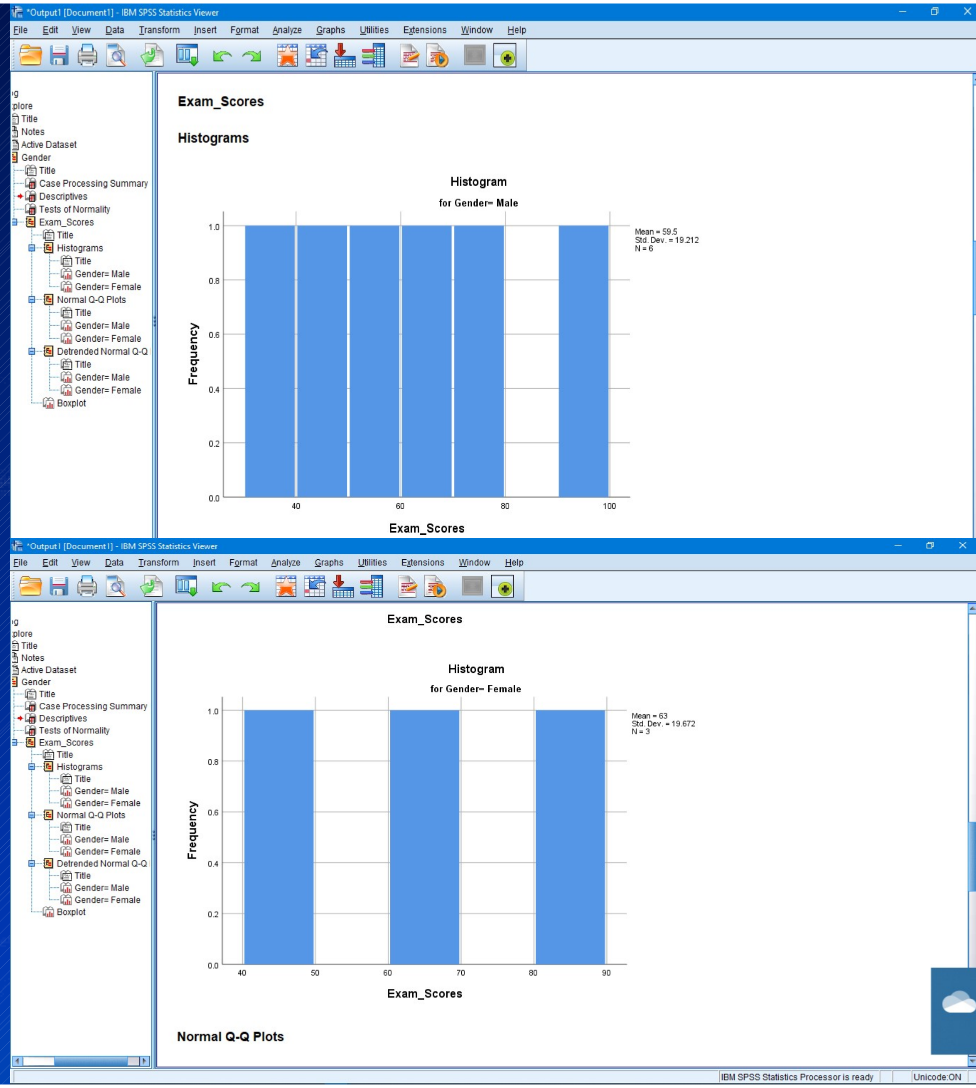
- g
- plore
  - Title
  - Notes
  - Active Dataset
  - Gender
    - Title
    - Case Processing Summary
    - Descriptives
    - Tests of Normality
    - Exam_Scores
      - Title
      - Histograms
        - Title
        - Gender= Male
        - Gender= Female
      - Normal Q-Q Plots
        - Title
        - Gender= Male
        - Gender= Female
      - Detrended Normal Q-Q
        - Title
        - Gender= Male
        - Gender= Female
      - Boxplot

**Descriptives**

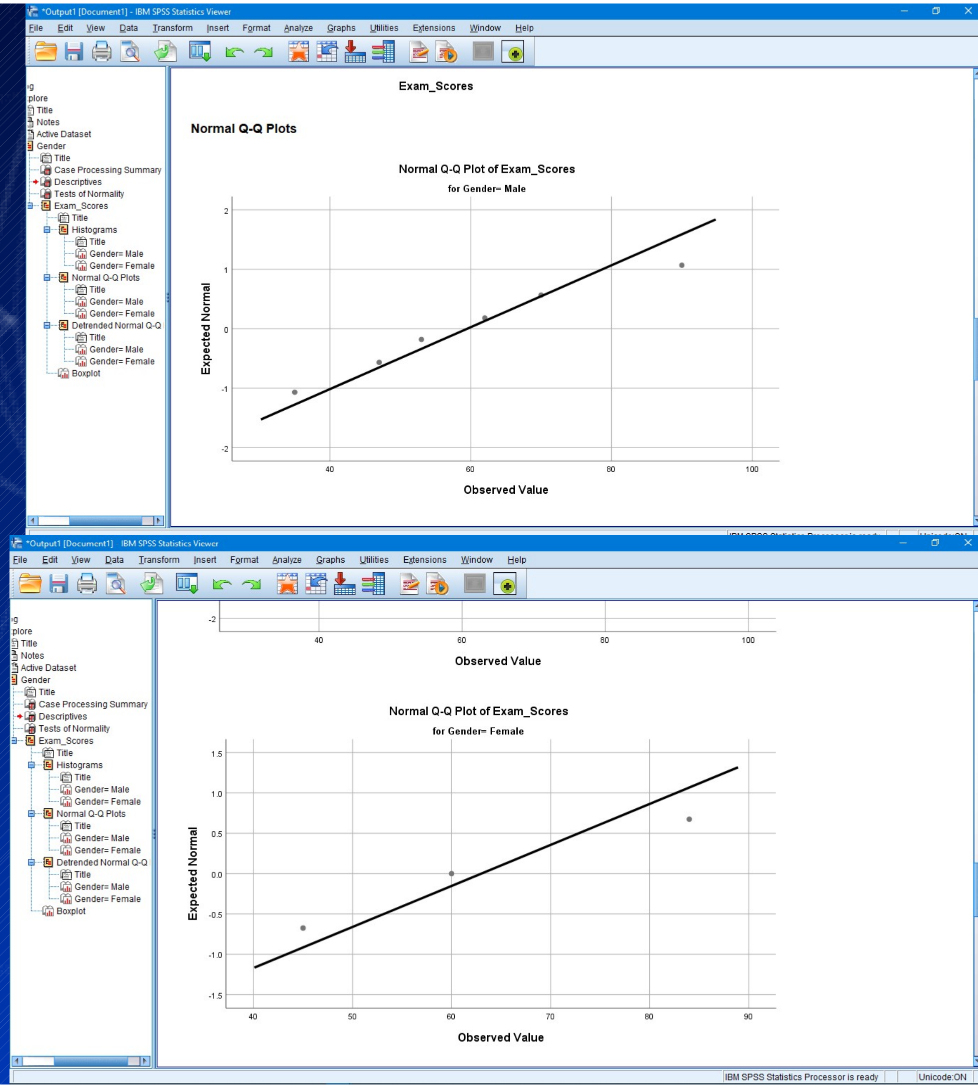
| | Gender | | | Statistic | Std. Error |
|---|---|---|---|---|---|
| Exam_Scores | Male | Mean | | 59.50 | 7.843 |
| | | 95% Confidence Interval for Mean | Lower Bound | 39.34 | |
| | | | Upper Bound | 79.66 | |
| | | 5% Trimmed Mean | | 59.17 | |
| | | Median | | 57.50 | |
| | | Variance | | 369.100 | |
| | | Std. Deviation | | 19.212 | |
| | | Minimum | | 35 | |
| | | Maximum | | 90 | |
| | | Range | | 55 | |
| | | Interquartile Range | | 31 | |
| | | Skewness | | .534 | .845 |
| | | Kurtosis | | .245 | 1.741 |
| | Female | Mean | | 63.00 | 11.358 |
| | | 95% Confidence Interval for Mean | Lower Bound | 14.13 | |
| | | | Upper Bound | 111.87 | |
| | | 5% Trimmed Mean | | . | |
| | | Median | | 60.00 | |
| | | Variance | | 387.000 | |
| | | Std. Deviation | | 19.672 | |
| | | Minimum | | 45 | |
| | | Maximum | | 84 | |
| | | Range | | 39 | |
| | | Interquartile Range | | . | |
| | | Skewness | | .670 | 1.225 |
| | | Kurtosis | | . | . |

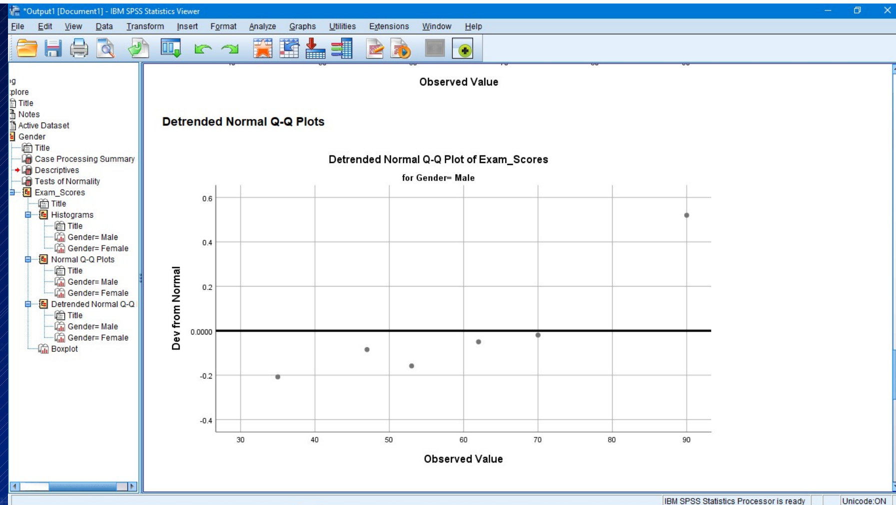
IBM SPSS Statistics Processor is ready | Unicode:ON

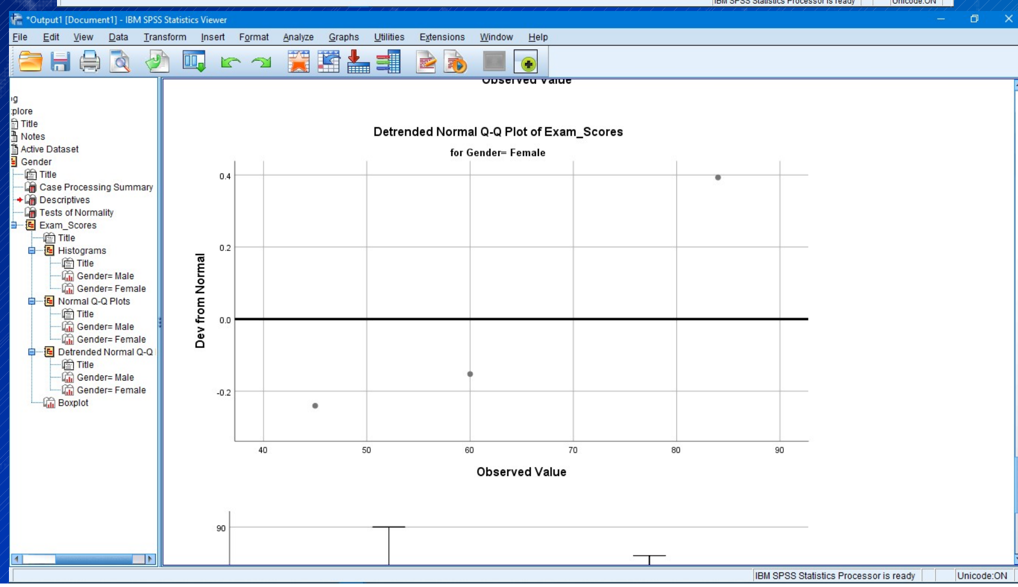

*Output1 [Document1] - IBM SPSS Statistics Viewer
File Edit View Data Transform Insert Format Analyze Graphs Utilities Extensions Window Help
Title
Notes
Active Dataset
Gender
Title
Case Processing Summary
Descriptives
Tests of Normality
Exam_Scores
Title
Histograms
Title
Gender= Male
Gender= Female
Normal Q-Q Plots
Title
Gender= Male
Gender= Female
Detrended Normal Q-Q
Title
Gender= Male
Gender= Female
Boxplot
Tests of Normality
Kolmogorov-Smirnov[a]
Shapiro-Wilk
Gender Statistic df Sig. Statistic df Sig.
Exam_Scores Male .132 6 .200* .984 6 .970
Female .227 3 . .983 3 .747
*. This is a lower bound of the true significance.
a. Lilliefors Significance Correction
Exam_Scores
Histograms
Histogram
for Gender= Male
Mean = 59.5
Std. Dev. = 19.212
N = 6
Frequency
1.0
0.8
0.6
0.4
IBM SPSS Statistics Processor is ready
Unicode:ON

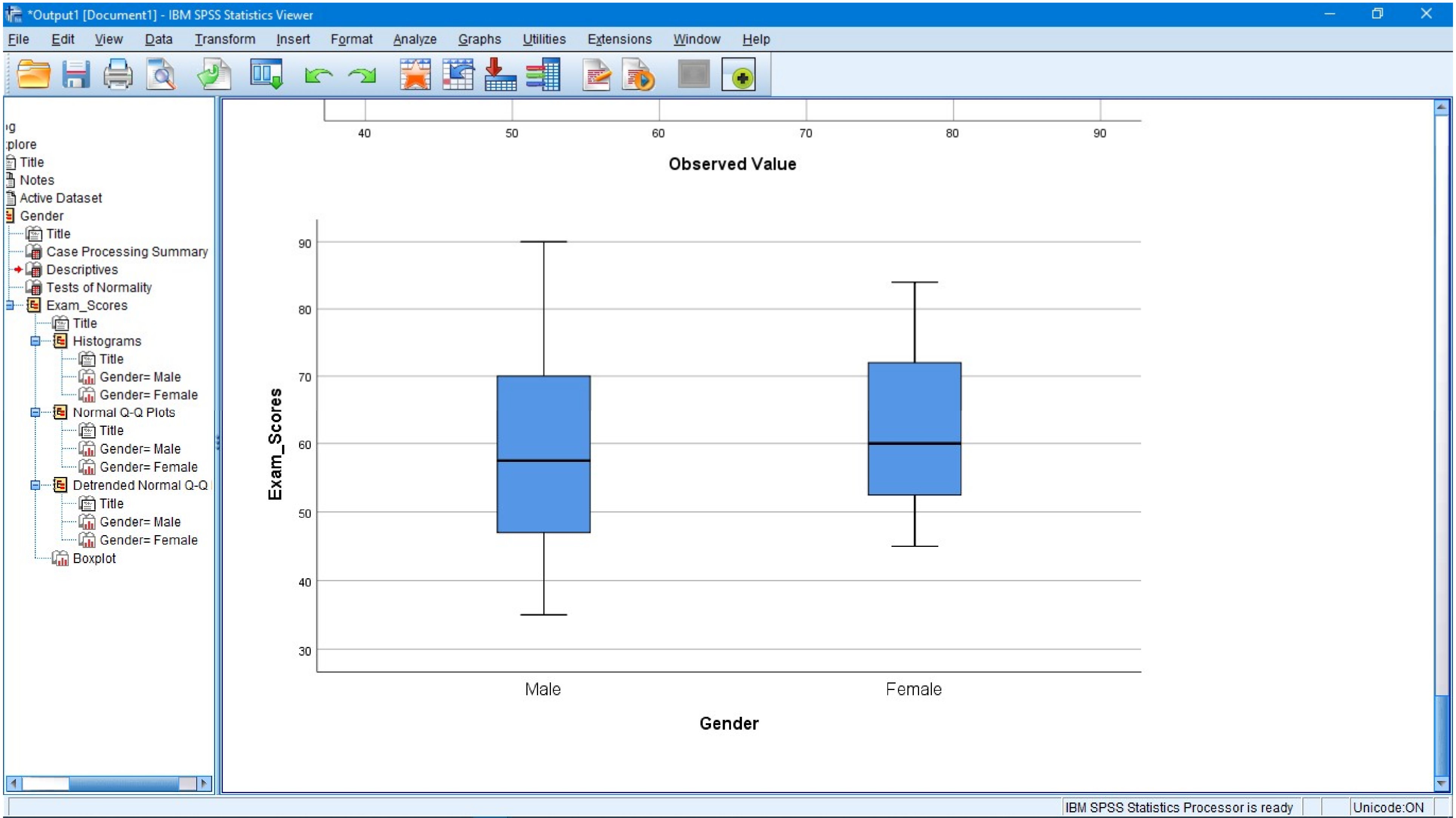


Exam_Scores
Histograms
Histogram
for Gender= Male
Mean = 59.5
Std. Dev. = 19.212
N = 6
Frequency
Exam_Scores
Exam_Scores
Histogram
for Gender= Female
Mean = 63
Std. Dev. = 19.672
N = 3
Frequency
Exam_Scores
Normal Q-Q Plots


*Output1 [Document1] - IBM SPSS Statistics Viewer
Exam_Scores
Normal Q-Q Plots
Normal Q-Q Plot of Exam_Scores
for Gender= Male
Expected Normal
Observed Value
Normal Q-Q Plot of Exam_Scores
for Gender= Female
Expected Normal
Observed Value
IBM SPSS Statistics Processor is ready
Unicode:ON


*Output1 [Document1] - IBM SPSS Statistics Viewer
File Edit View Data Transform Insert Format Analyze Graphs Utilities Extensions Window Help
Observed Value
Detrended Normal Q-Q Plots
Detrended Normal Q-Q Plot of Exam_Scores
for Gender= Male
Dev from Normal
Observed Value
IBM SPSS Statistics Processor is ready
Unicode:ON


*Output1 [Document1] - IBM SPSS Statistics Viewer
File Edit View Data Transform Insert Format Analyze Graphs Utilities Extensions Window Help
Title
Notes
Active Dataset
Gender
Case Processing Summary
Descriptives
Tests of Normality
Exam_Scores
Histograms
Gender= Male
Gender= Female
Normal Q-Q Plots
Detrended Normal Q-Q
Boxplot
Detrended Normal Q-Q Plot of Exam_Scores
for Gender= Female
Dev from Normal
Observed Value
IBM SPSS Statistics Processor is ready
Unicode:ON


*Output1 [Document1] - IBM SPSS Statistics Viewer
File Edit View Data Transform Insert Format Analyze Graphs Utilities Extensions Window Help
g
plore
Title
Notes
Active Dataset
Gender
Title
Case Processing Summary
Descriptives
Tests of Normality
Exam_Scores
Title
Histograms
Title
Gender= Male
Gender= Female
Normal Q-Q Plots
Title
Gender= Male
Gender= Female
Detrended Normal Q-Q
Title
Gender= Male
Gender= Female
Boxplot
40 50 60 70 80 90
Observed Value
90
80
70
60
50
40
30
Exam_Scores
Male
Female
Gender
IBM SPSS Statistics Processor is ready
Unicode:ON

# Conclusion :

The skewness and kurtosis measure should be as to zero as possible in spss.

As a consequence you must divide the measure by its standard error.and you need to do this hand using a calculator.

This will give the z-vaule, which should be somewhere between -1.96 to +1.96.

Male : to calculate the skewness z-vaule , divide the skewness by its standard error.

0.534/0.845 = 0.63

To value is 0.63 neither below -1.96 nor above +1.96

0.245/1.741 = 0.14

**Female:** 0.670/1.225 = 0.54

All the z-values are within +/- 1.96.

Conclusion : Regarding skewness and kurtosis .our example data are little skewed and kurtotic, for both males and females, but it does not differ significantly from normality.

The null hypothesis for the test of normality is that the data are normally distributed.

In spss p-value is labeled by “sig”.

Both p-values are above 0.05.we keep null hypothesis.