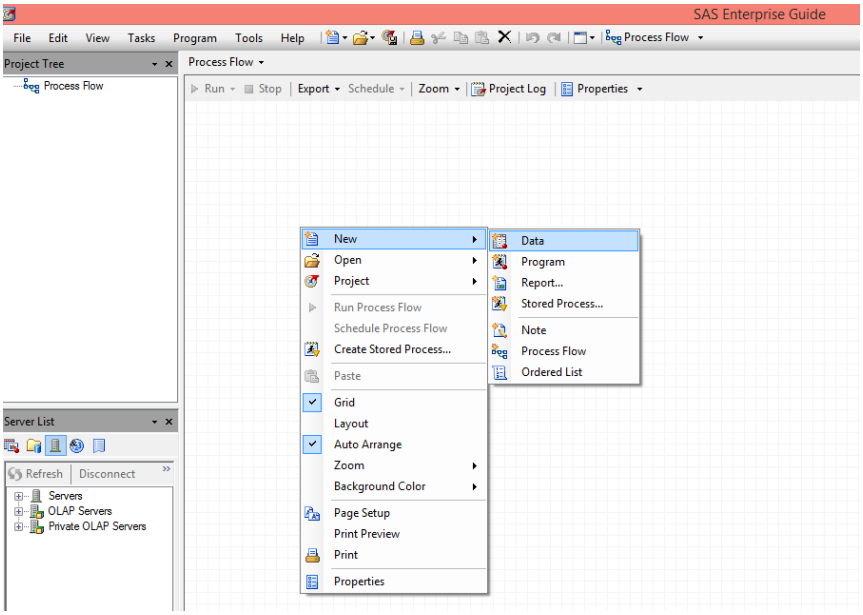
Harsh Agarwal
Roll No: B004
B.Tech CS-B

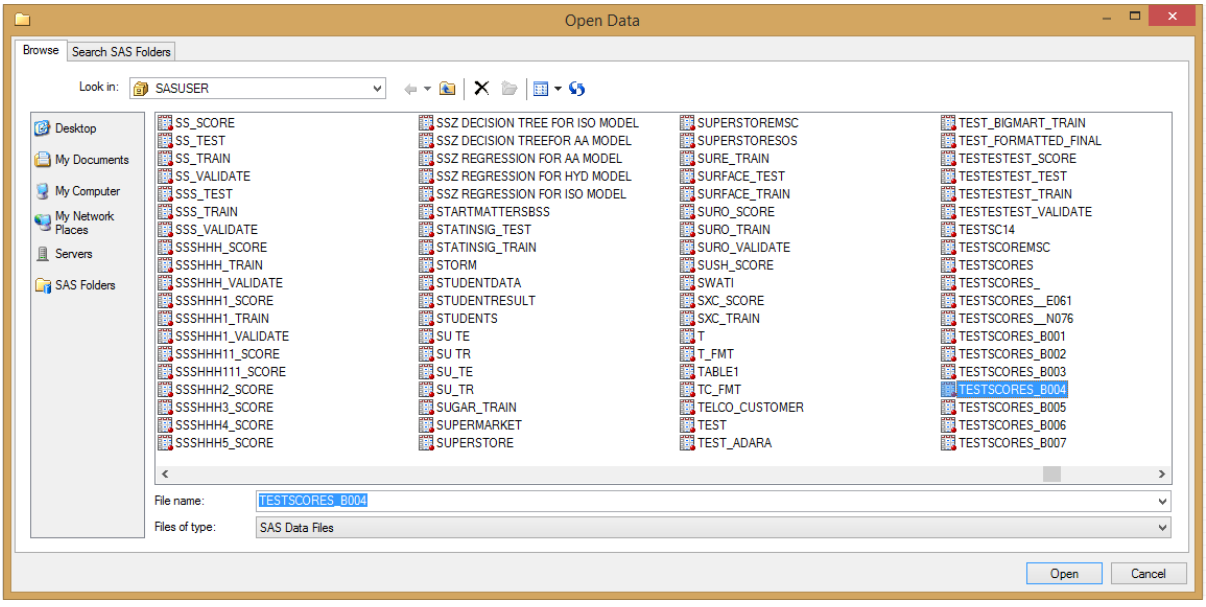
Assignment 1

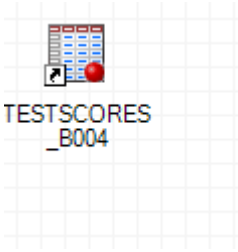
Aim: - In this assignment, we have implemented the following:

1) Filter and Sort
2) Statistical Analysis
3) Distribution Analysis
4) Confidence Intervals
5) t Test – One sample and Two sample

Dataset Used: SAT Test Score







TESTSCORES_B004

Filter and Sort | Query Builder | Data | Describe | Graph | Analyze | Export | Send To

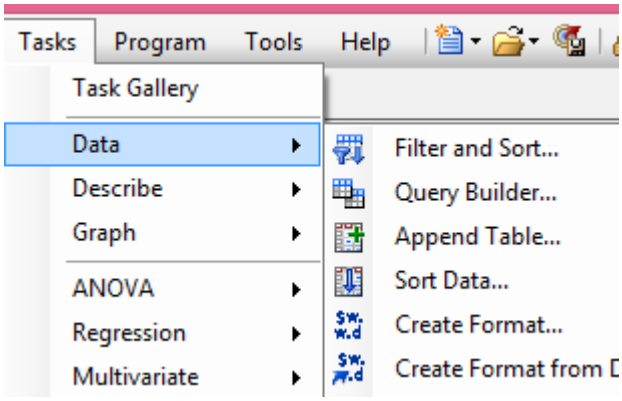
	Gender	SATScore	IDNumber
1	Male	1170	61469897
2	Female	1090	33081197
3	Male	1240	68137597
4	Female	1000	37070397
5	Male	1210	64608797
6	Female	970	60714297
7	Male	1020	16907997
8	Female	1490	9589297
9	Male	1200	93891897
10	Female	1260	85859397
11	Male	1150	38152597
12	Female	1390	99108497
13	Male	1240	59666697
14	Female	1370	70847197
15	Male	1140	47613397
16	Female	1160	53750297
17	Male	1050	95948597
18	Female	1110	3873197
19	Male	1100	25756097
20	Female	1080	43493297
21	Male	1120	27543197
22	Female	1080	26212897
23	Male	1050	8945097
24	Female	1200	51799397
25	Male	1600	39196697
26	Female	1100	48154497
27	Male	1050	55189597
28	Female	1060	46028397
29	Male	1140	75332897
30	Female	1100	29520797
31	Male	1340	55983497
32	Female	1240	93236497
33	Male	1090	6975697
34	Female	1180	29686297
35	Male	1170	76815697
36	Female	1130	64045497
37	Male	1290	9880297
38	Female	1380	23048597
39	Male	1010	76058697

Filter and Sort

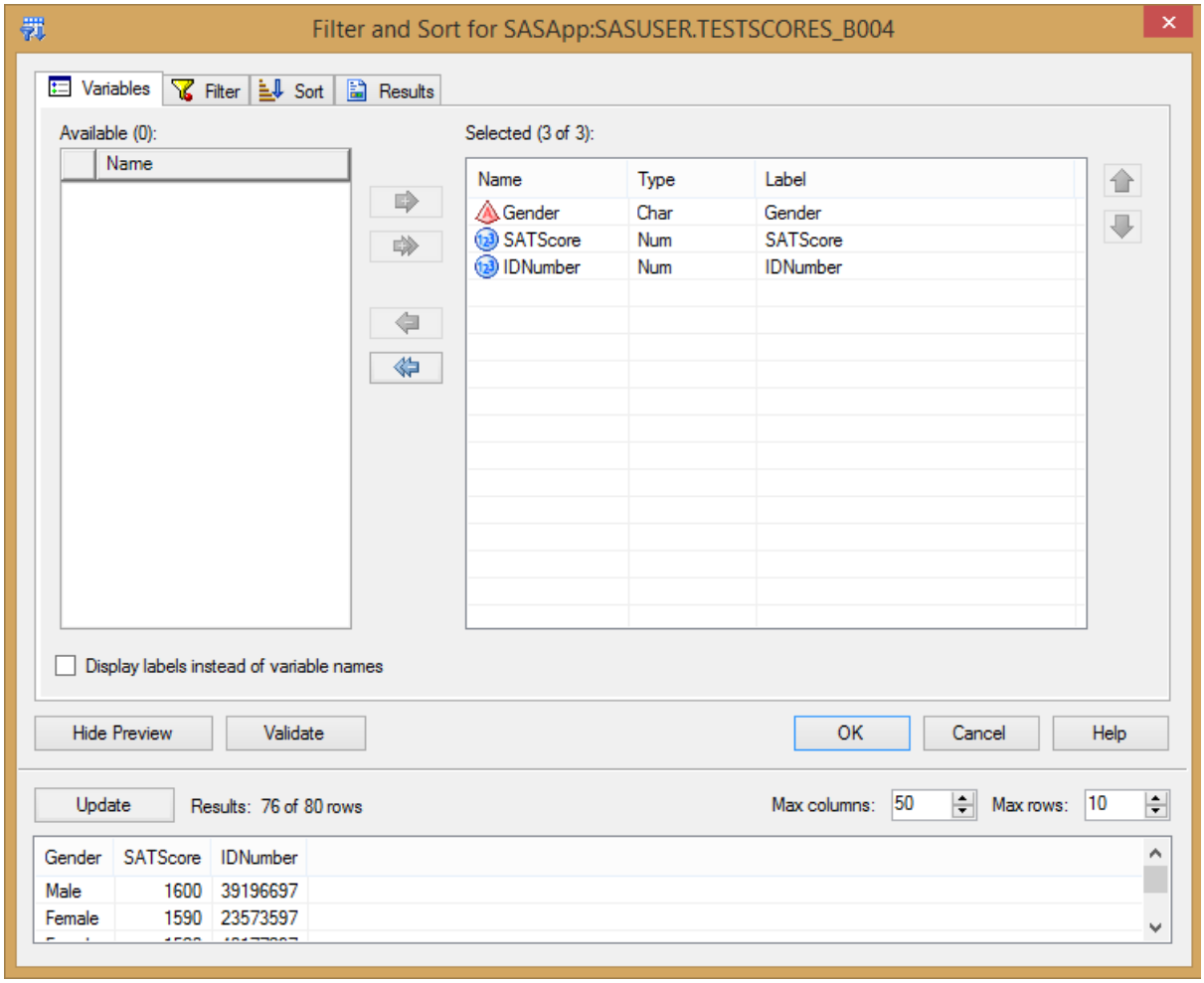
- **Tasks > Data > Filter and Sort**
- **Select all variables > Add filters wrt variables selected**
- **Sort by SAT Score > Ascending**

Steps:

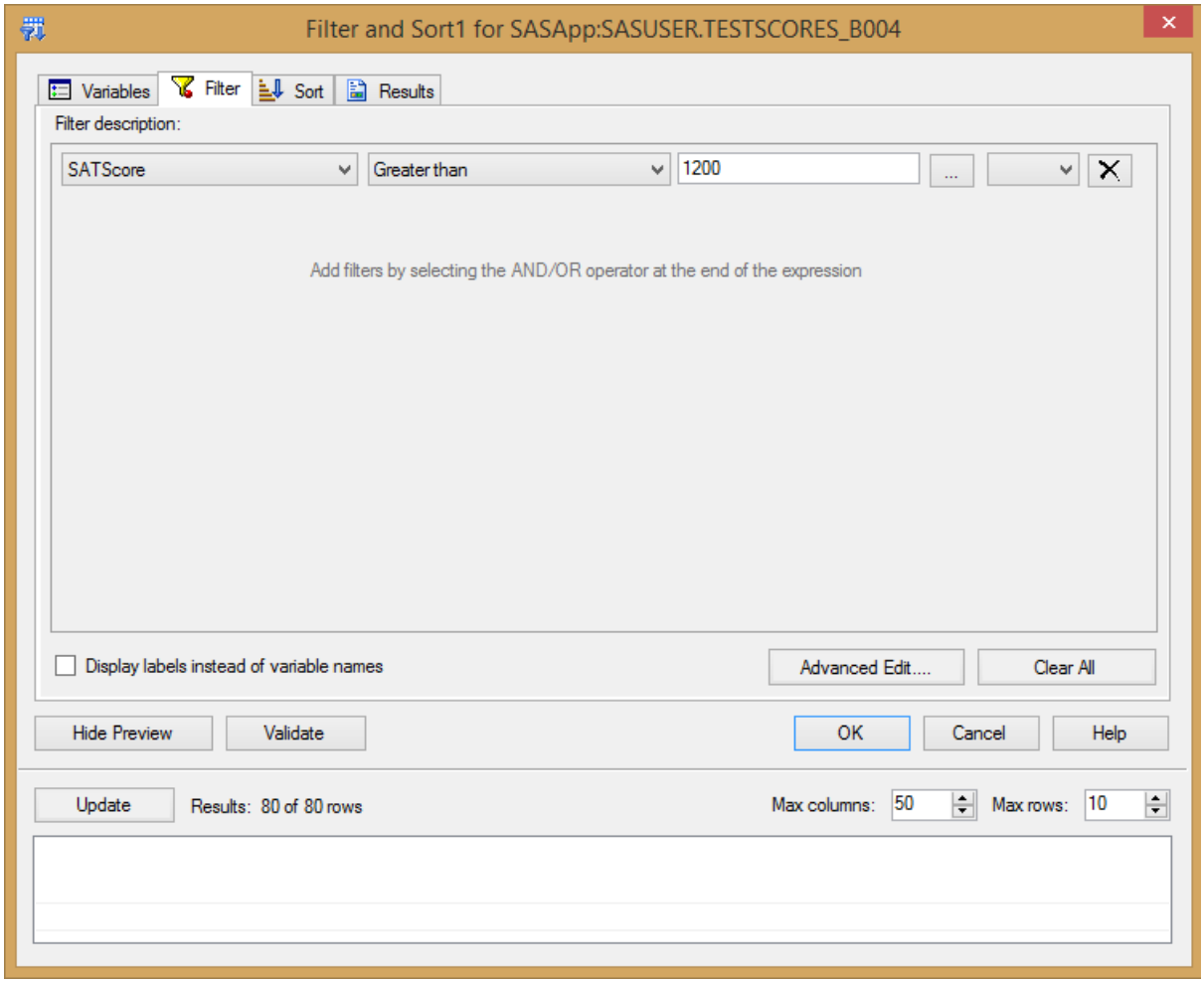
1)



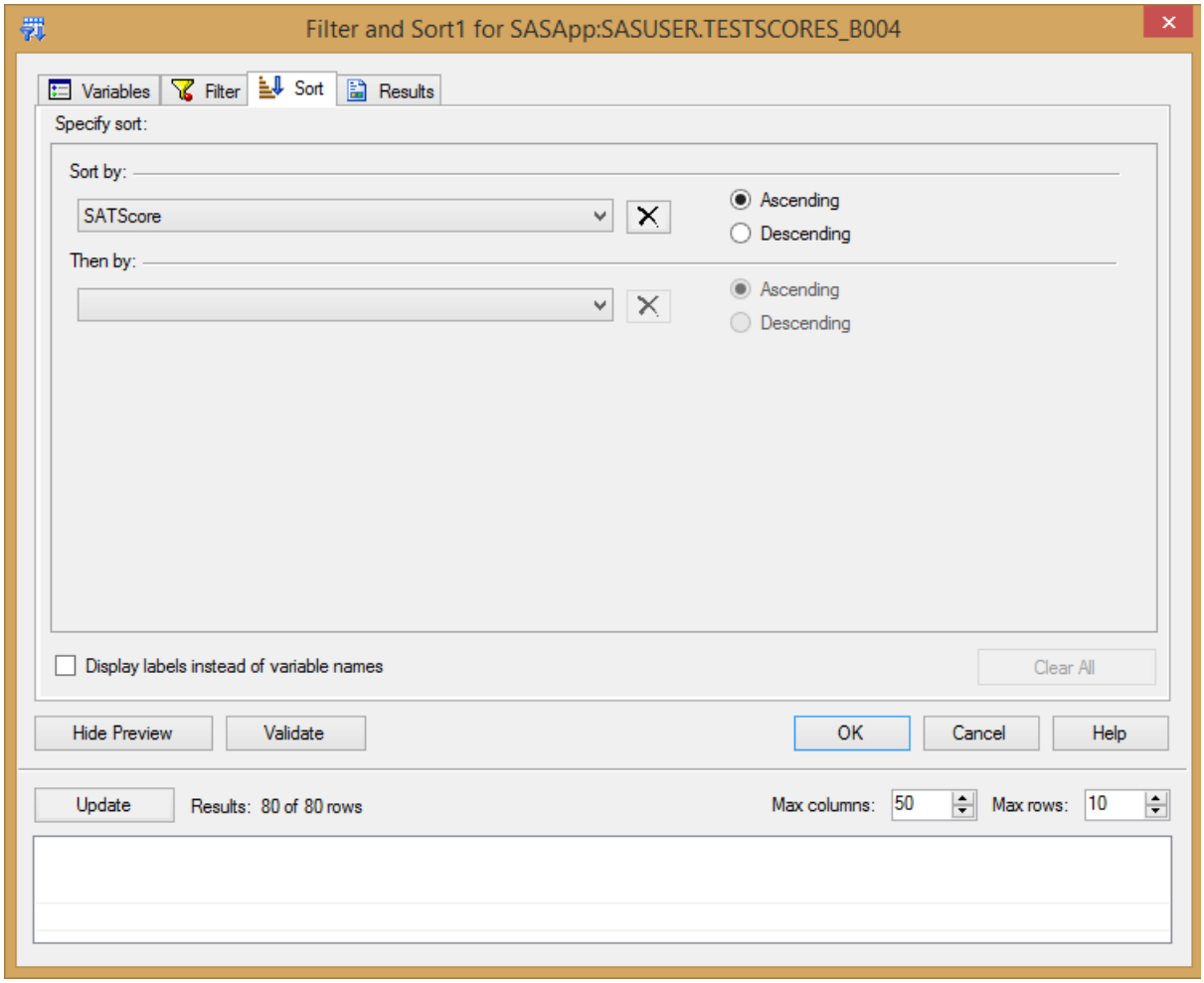
2)


Filter and Sort for SASApp:SASUSER.TESTSCORES_B004
Variables
Filter
Sort
Results
Available (0):
Name
Selected (3 of 3):
Name
Type
Label
Gender
Char
Gender
SATScore
Num
SATScore
IDNumber
Num
IDNumber
Display labels instead of variable names
Hide Preview
Validate
OK
Cancel
Help
Update
Results: 76 of 80 rows
Max columns:
50
Max rows:
10
Gender
SATScore
IDNumber
Male
1600
39196697
Female
1590
23573597

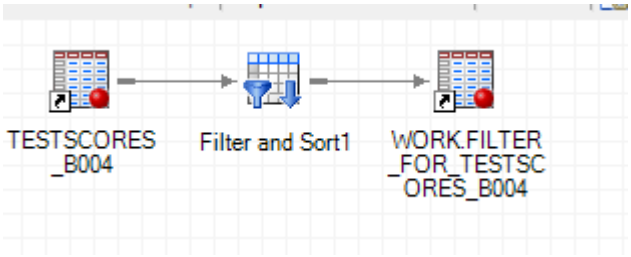
3)


Filter and Sort1 for SASApp:SASUSER.TESTSCORES_B004
Variables
Filter
Sort
Results
Filter description:
SATScore
Greater than
1200
...
Add filters by selecting the AND/OR operator at the end of the expression
Display labels instead of variable names
Advanced Edit....
Clear All
Hide Preview
Validate
OK
Cancel
Help
Update
Results: 80 of 80 rows
Max columns:
50
Max rows:
10

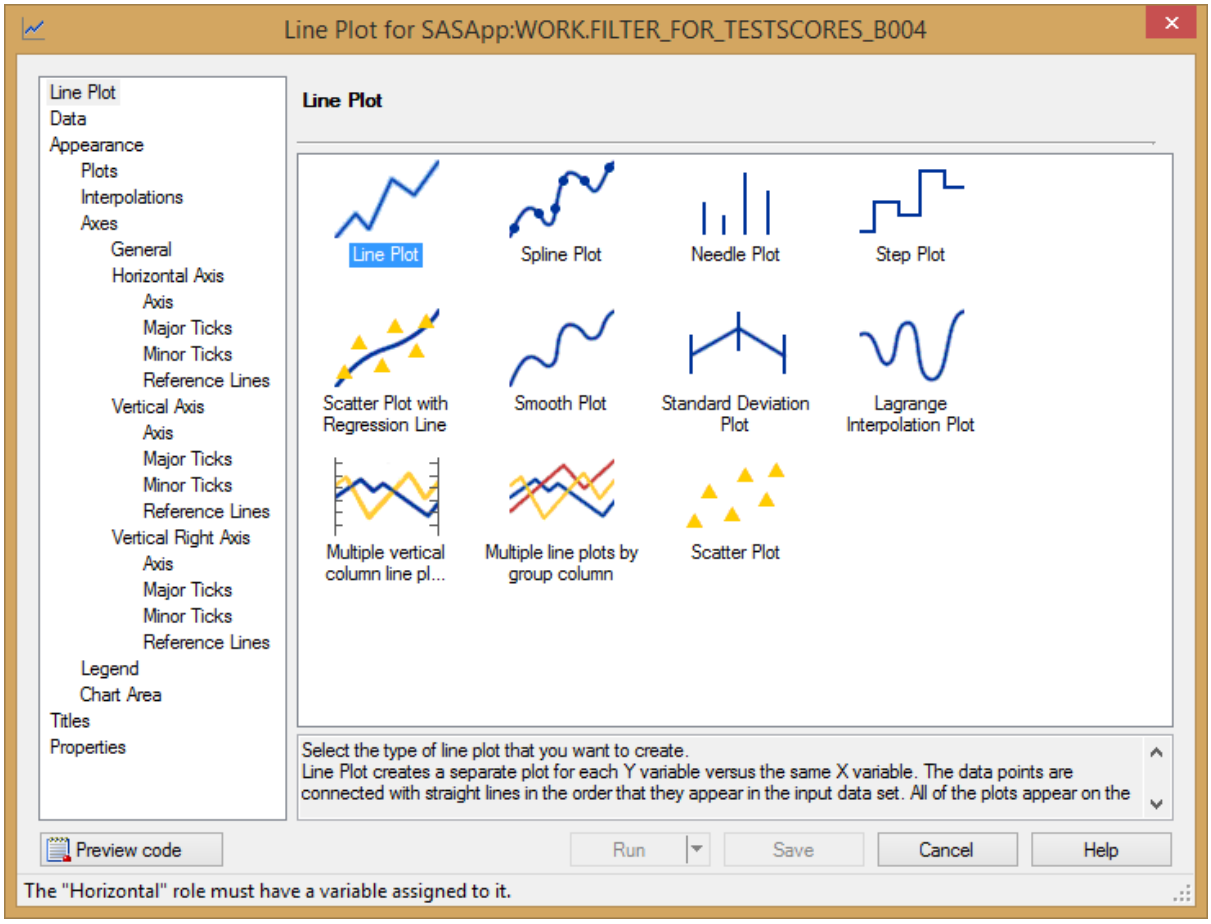
4)


Filter and Sort1 for SASApp:SASUSER.TESTSCORES_B004
Variables
Filter
Sort
Results
Specify sort:
Sort by:
SATScore
Ascending
Descending
Then by:
Ascending
Descending
Display labels instead of variable names
Clear All
Hide Preview
Validate
OK
Cancel
Help
Update
Results: 80 of 80 rows
Max columns:
50
Max rows:
10

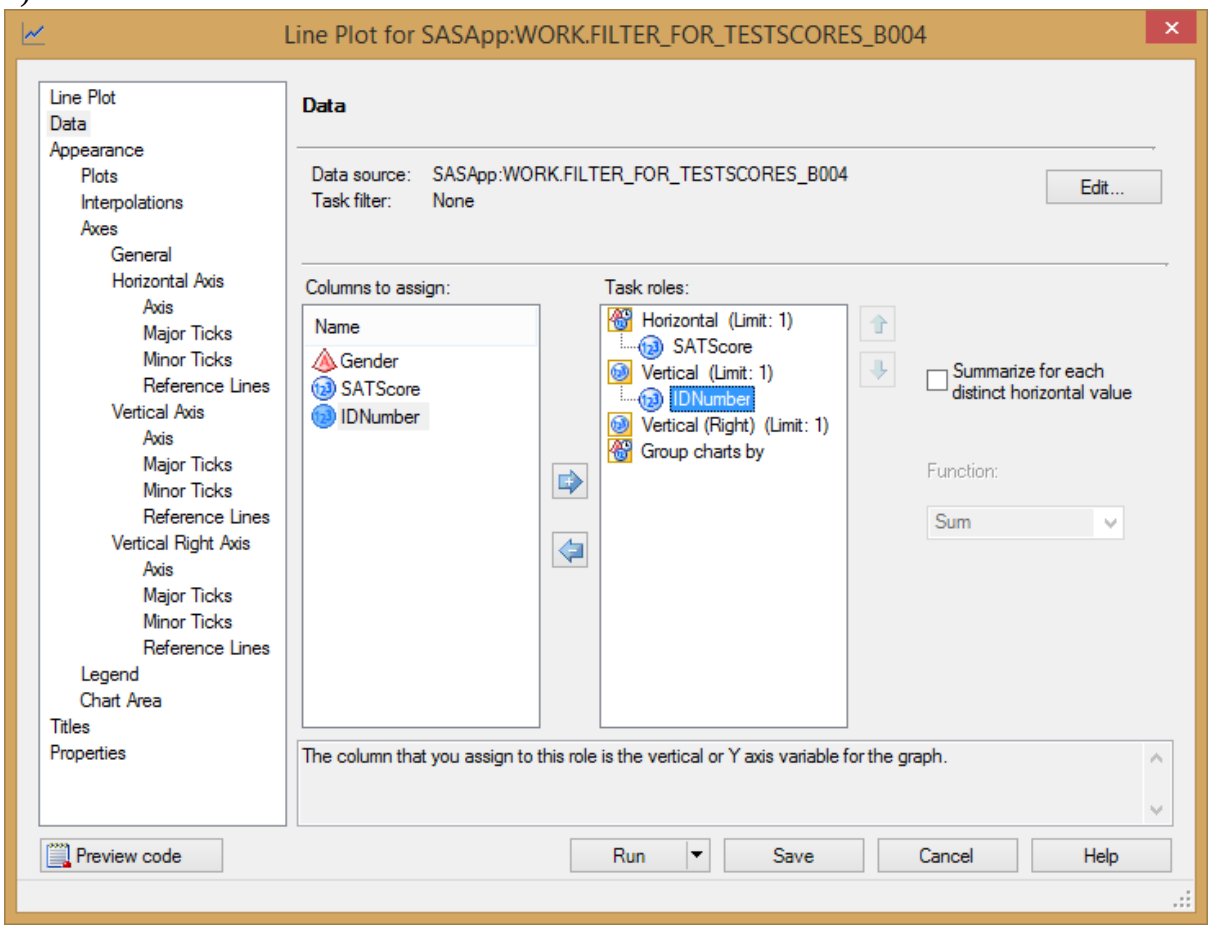
5)


TESTSCORES
_B004
Filter and Sort1
WORK.FILTER
_FOR_TESTSC
ORES_B004

6)


Line Plot for SASApp:WORK.FILTER_FOR_TESTSCORES_B004
Line Plot
Data
Appearance
Plots
Interpolations
Axes
General
Horizontal Axis
Axis
Major Ticks
Minor Ticks
Reference Lines
Vertical Axis
Axis
Major Ticks
Minor Ticks
Reference Lines
Vertical Right Axis
Axis
Major Ticks
Minor Ticks
Reference Lines
Legend
Chart Area
Titles
Properties
Line Plot
Line Plot
Spline Plot
Needle Plot
Step Plot
Scatter Plot with Regression Line
Smooth Plot
Standard Deviation Plot
Lagrange Interpolation Plot
Multiple vertical column line pl...
Multiple line plots by group column
Scatter Plot
Select the type of line plot that you want to create.
Line Plot creates a separate plot for each Y variable versus the same X variable. The data points are connected with straight lines in the order that they appear in the input data set. All of the plots appear on the
Preview code
Run
Save
Cancel
Help
The "Horizontal" role must have a variable assigned to it.

7)

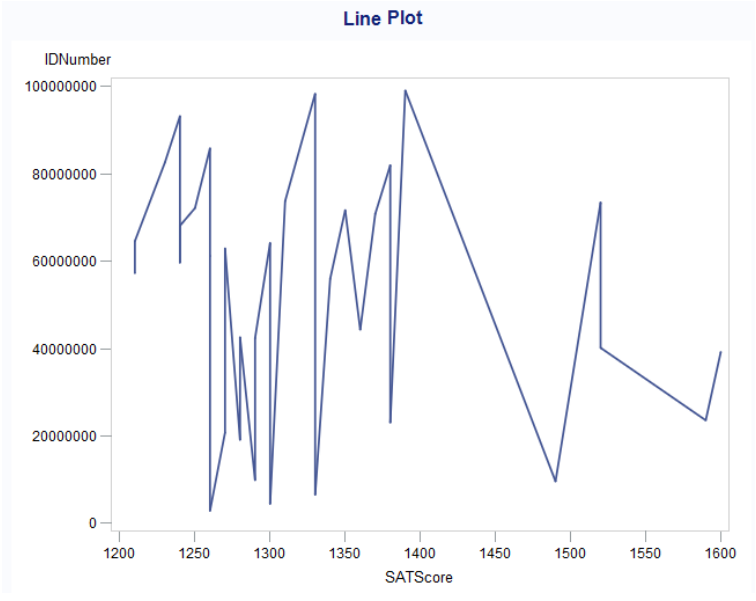

Line Plot for SASApp:WORK.FILTER_FOR_TESTSCORES_B004
Line Plot
Data
Appearance
Plots
Interpolations
Axes
General
Horizontal Axis
Axis
Major Ticks
Minor Ticks
Reference Lines
Vertical Axis
Axis
Major Ticks
Minor Ticks
Reference Lines
Vertical Right Axis
Axis
Major Ticks
Minor Ticks
Reference Lines
Legend
Chart Area
Titles
Properties
Data
Data source: SASApp:WORK.FILTER_FOR_TESTSCORES_B004
Task filter: None
Edit...
Columns to assign:
Name
Gender
SATScore
IDNumber
Task roles:
Horizontal (Limit: 1)
SATScore
Vertical (Limit: 1)
IDNumber
Vertical (Right) (Limit: 1)
Group charts by
Summarize for each distinct horizontal value
Function:
Sum
The column that you assign to this role is the vertical or Y axis variable for the graph.
Preview code
Run
Save
Cancel
Help

Results:

Input Data | Code | Log | Output D

Modify Task | Filter and Sort | Que

	Gender	SATScore
1	Male	1210
2	Male	1210
3	Female	1230
4	Female	1240
5	Male	1240
6	Male	1240
7	Female	1250
8	Female	1260
9	Male	1260
10	Female	1260
11	Female	1270
12	Female	1270
13	Male	1280
14	Female	1280
15	Male	1290
16	Male	1290
17	Female	1300
18	Female	1300
19	Male	1310
20	Male	1330
21	Female	1330
22	Male	1340
23	Female	1350
24	Male	1360
25	Female	1370
26	Female	1380
27	Female	1380
28	Female	1390
29	Female	1490
30	Female	1520
31	Female	1520
32	Female	1590
33	Male	1600

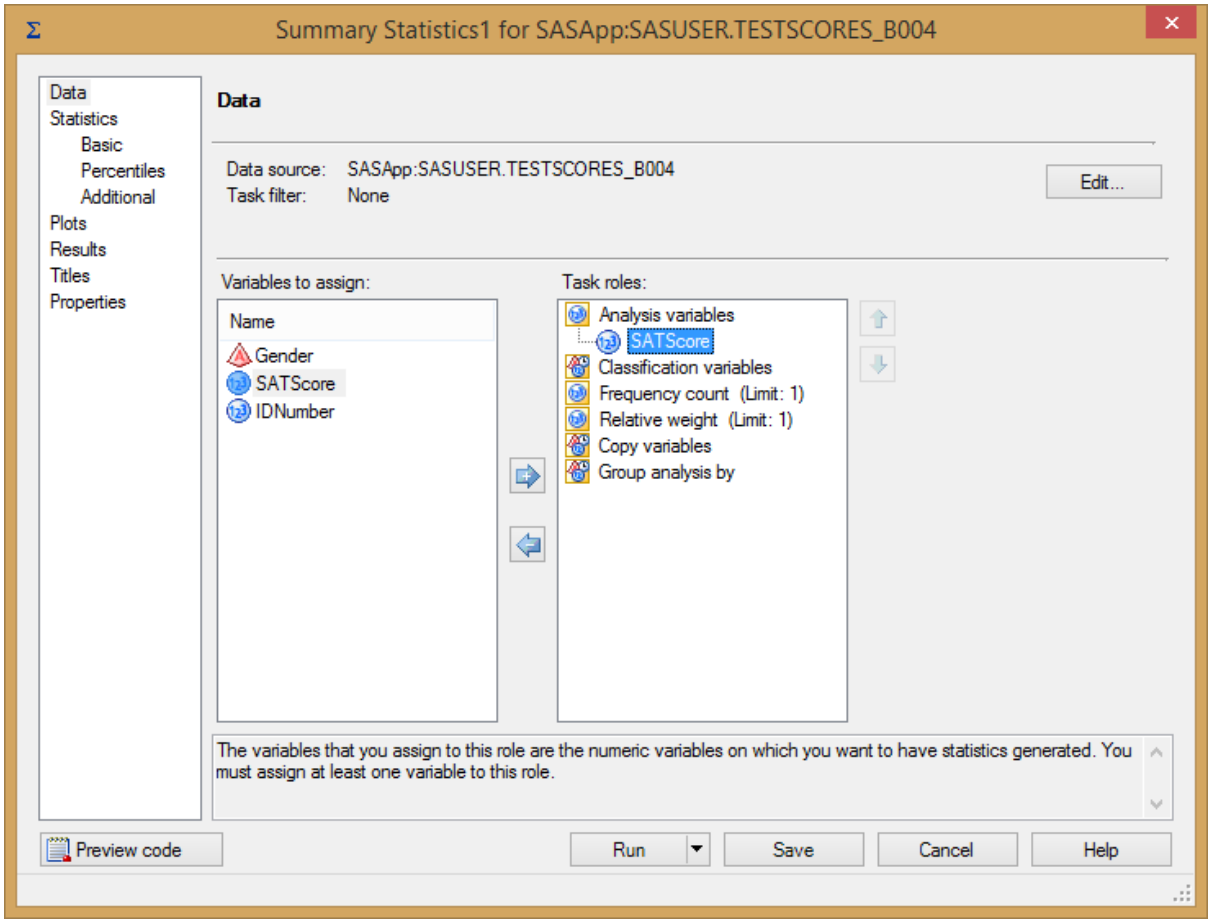


Summary Statistics

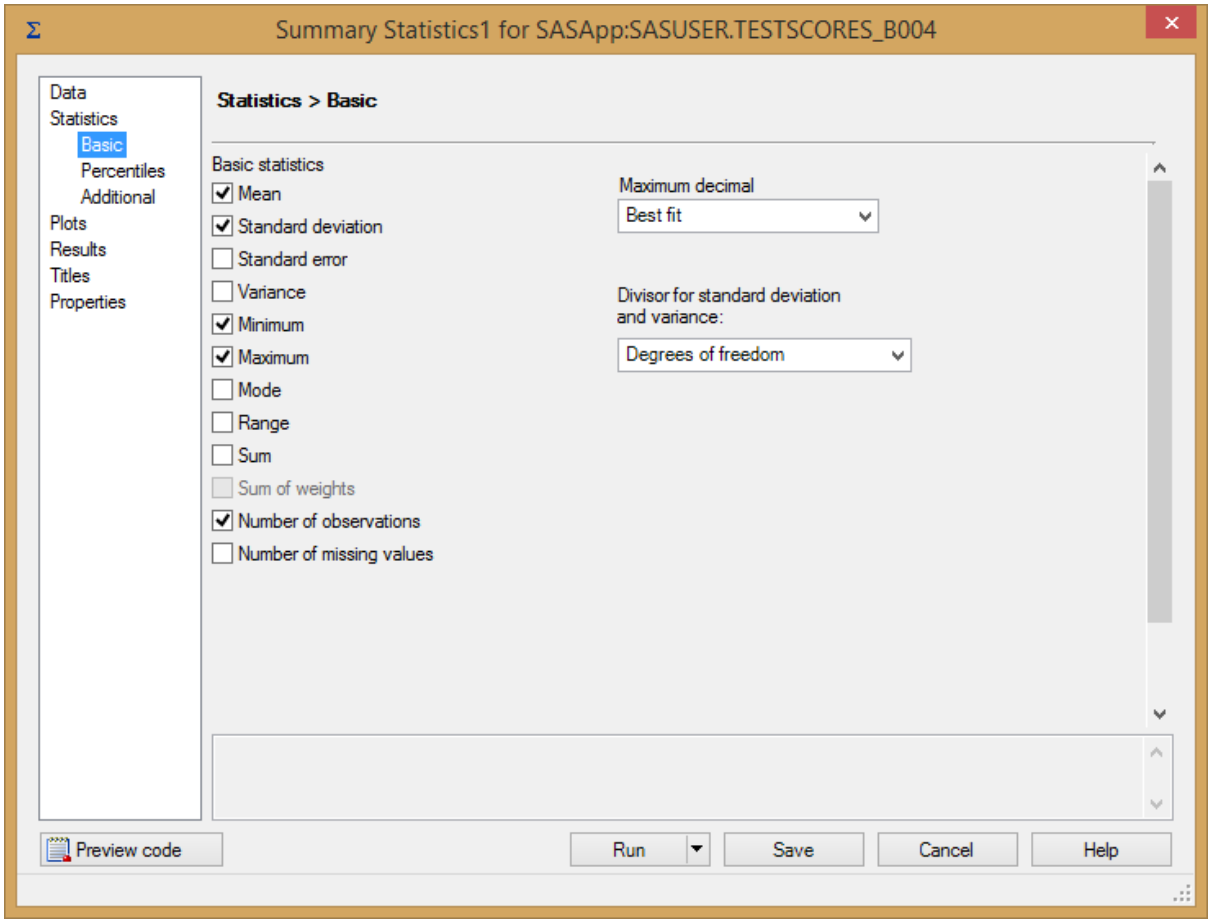
- **Tasks > Describe > Summary Statistics**
- **Data > Analysis Variable > SAT Score**
- **Statistics > Percentile > Lower Quartile, Upper Quartile**
- **Plots > Histogram, Box, Whisker**

Steps :

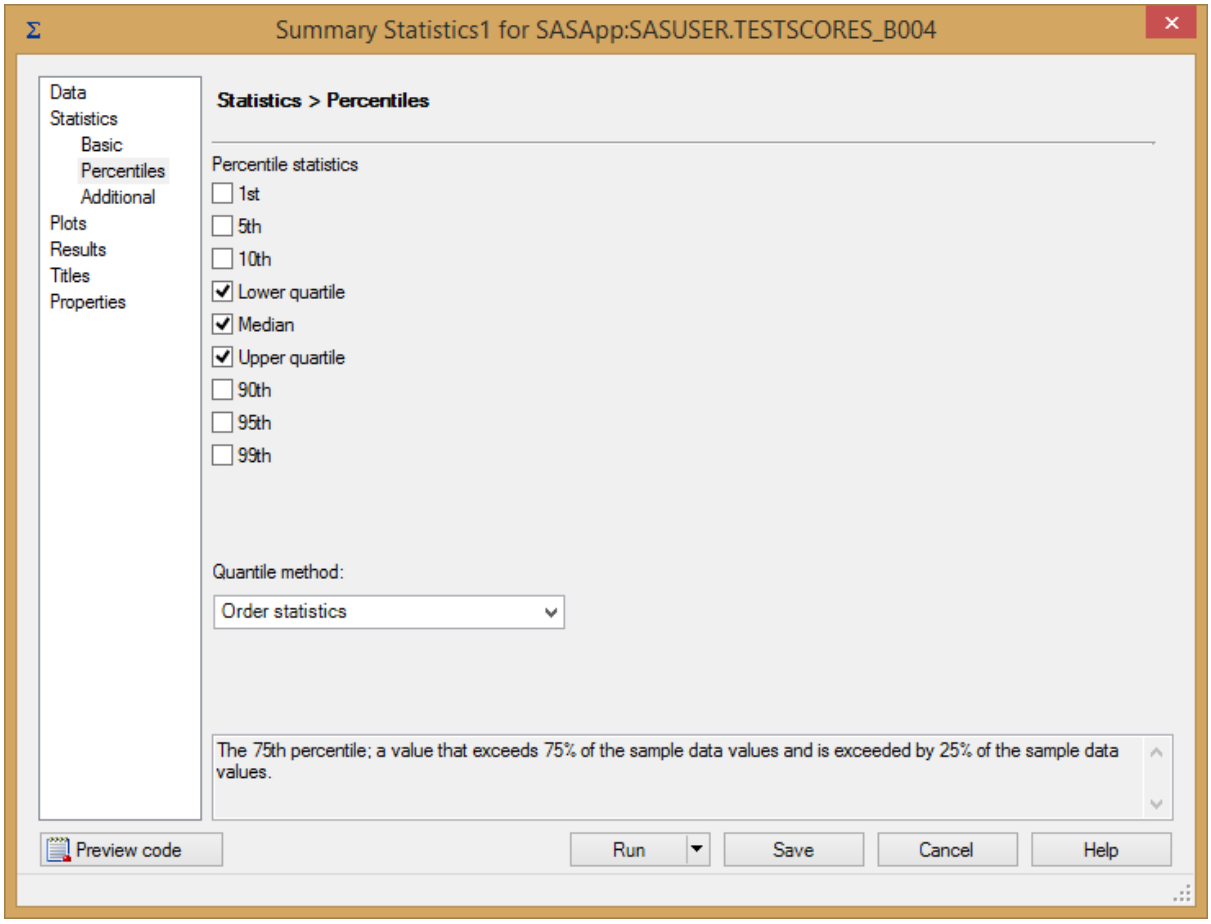
1)



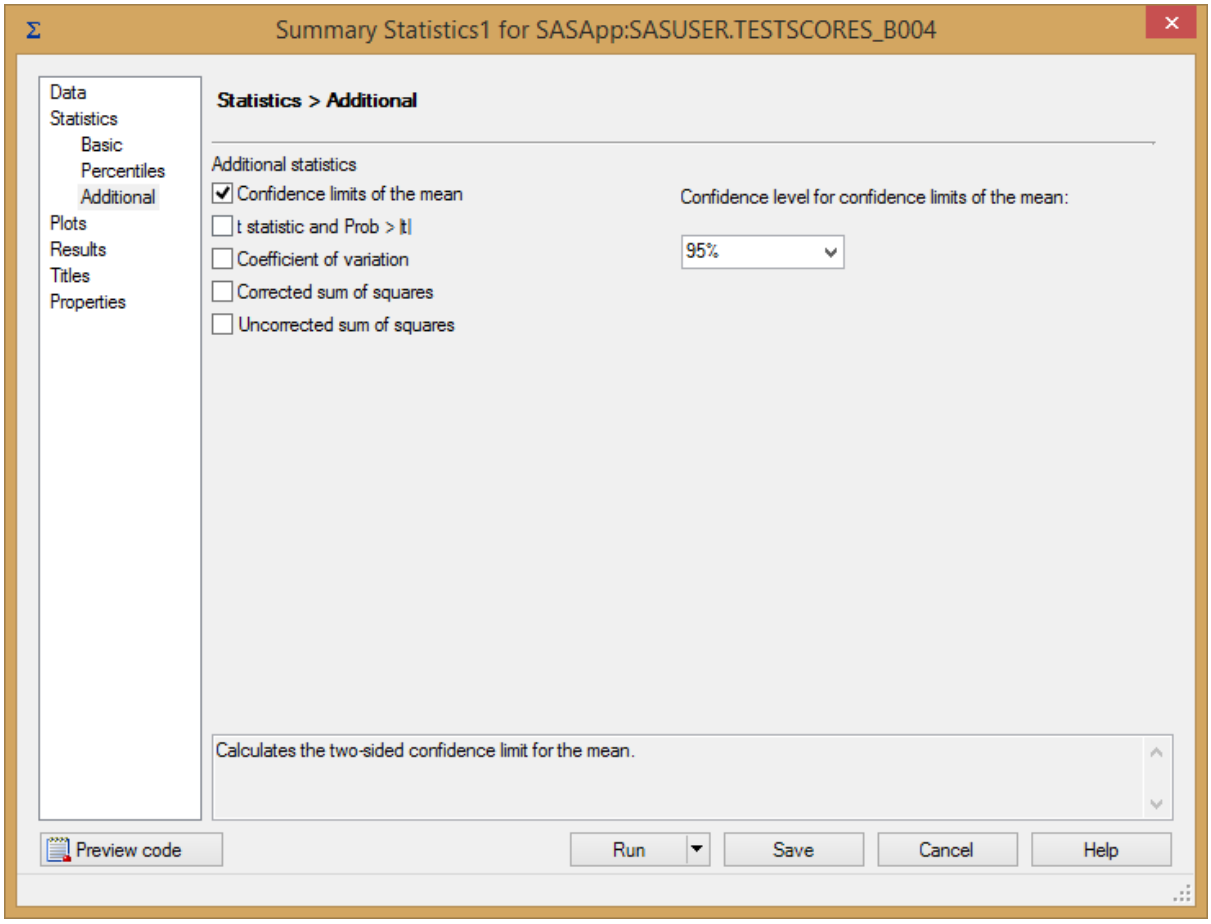
2)


Summary Statistics1 for SASApp:SASUSER.TESTSCORES_B004
Data
Statistics
Basic
Percentiles
Additional
Plots
Results
Titles
Properties
Statistics > Basic
Basic statistics
Mean
Standard deviation
Standard error
Variance
Minimum
Maximum
Mode
Range
Sum
Sum of weights
Number of observations
Number of missing values
Maximum decimal
Best fit
Divisor for standard deviation and variance:
Degrees of freedom
Preview code
Run
Save
Cancel
Help

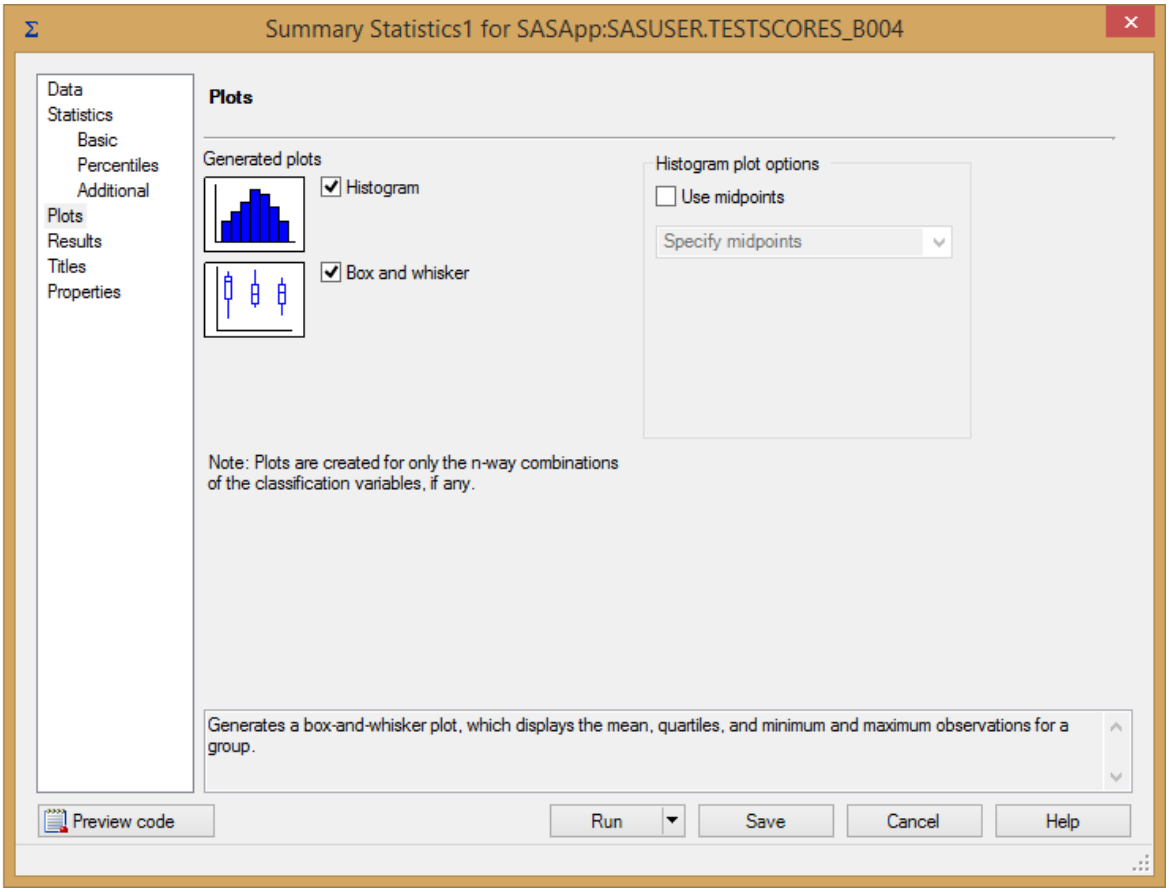
3)


Summary Statistics1 for SASApp:SASUSER.TESTSCORES_B004
Data
Statistics
Basic
Percentiles
Additional
Plots
Results
Titles
Properties
Statistics > Percentiles
Percentile statistics
1st
5th
10th
Lower quartile
Median
Upper quartile
90th
95th
99th
Quantile method:
Order statistics
The 75th percentile; a value that exceeds 75% of the sample data values and is exceeded by 25% of the sample data values.
Preview code
Run
Save
Cancel
Help

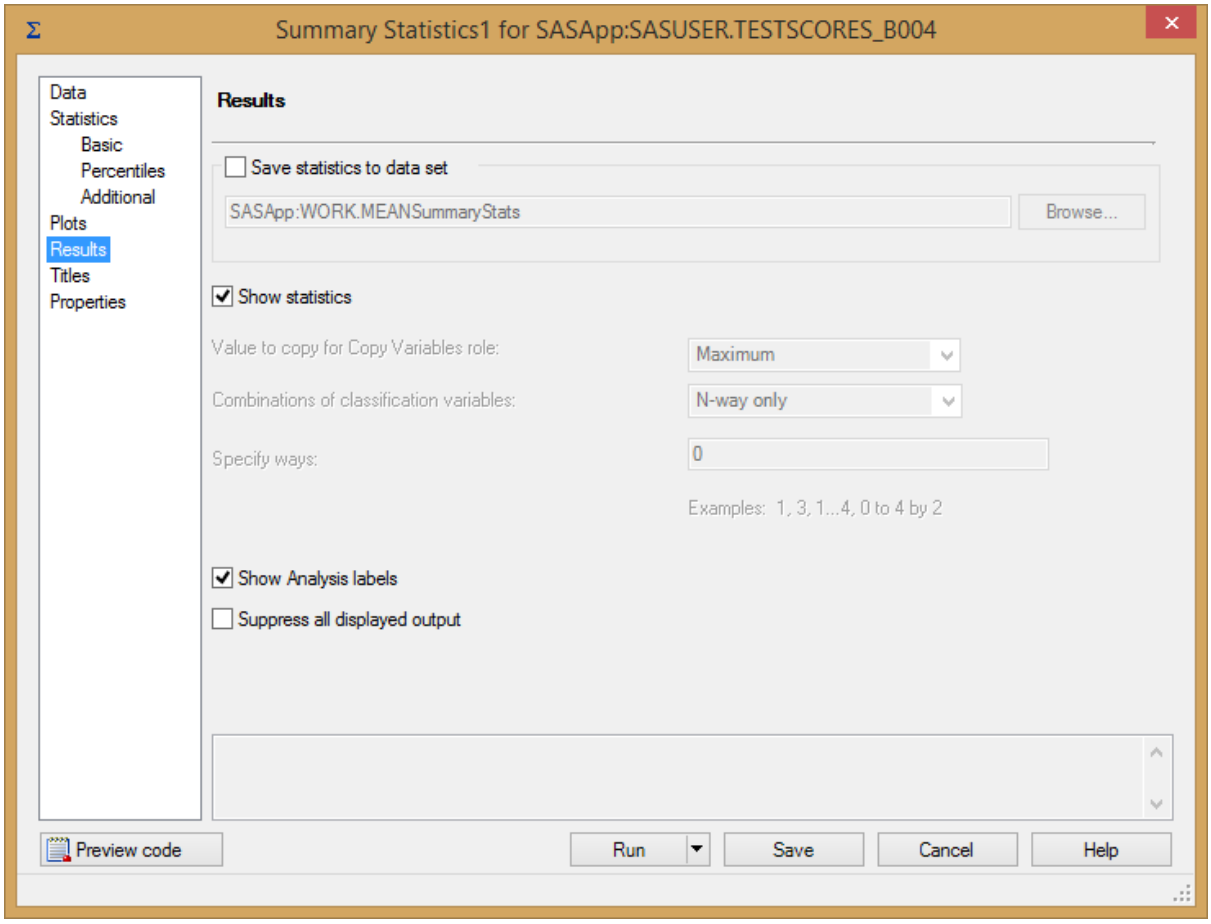
4)


Summary Statistics1 for SASApp:SASUSER.TESTSCORES_B004
Data
Statistics
Basic
Percentiles
Additional
Plots
Results
Titles
Properties
Statistics > Additional
Additional statistics
Confidence limits of the mean
t statistic and Prob > |t|
Coefficient of variation
Corrected sum of squares
Uncorrected sum of squares
Confidence level for confidence limits of the mean:
95%
Calculates the two-sided confidence limit for the mean.
Preview code
Run
Save
Cancel
Help

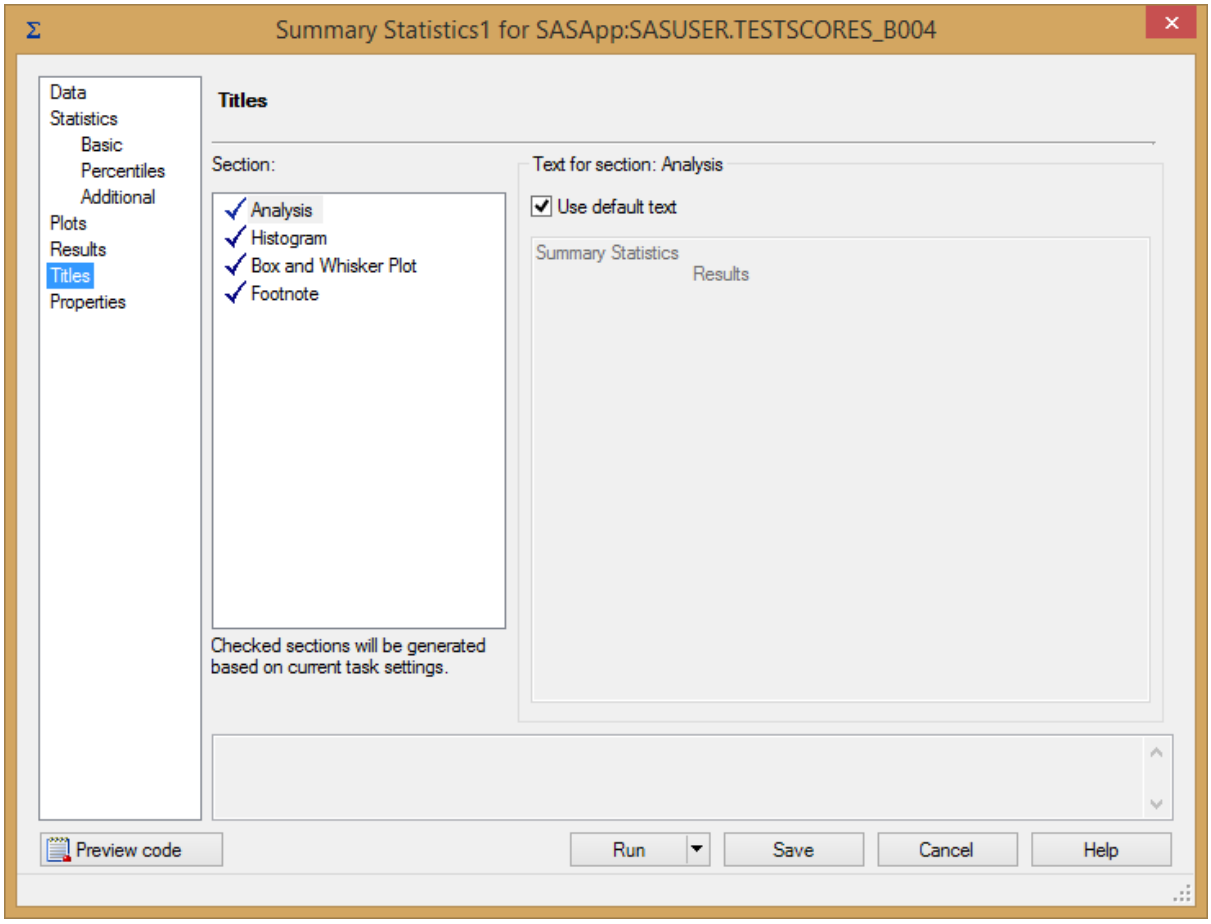
5)


Summary Statistics1 for SASApp:SASUSER.TESTSCORES_B004
Data
Statistics
Basic
Percentiles
Additional
Plots
Results
Titles
Properties
Plots
Generated plots
Histogram
Box and whisker
Histogram plot options
Use midpoints
Specify midpoints
Note: Plots are created for only the n-way combinations of the classification variables, if any.
Generates a box-and-whisker plot, which displays the mean, quartiles, and minimum and maximum observations for a group.
Preview code
Run
Save
Cancel
Help

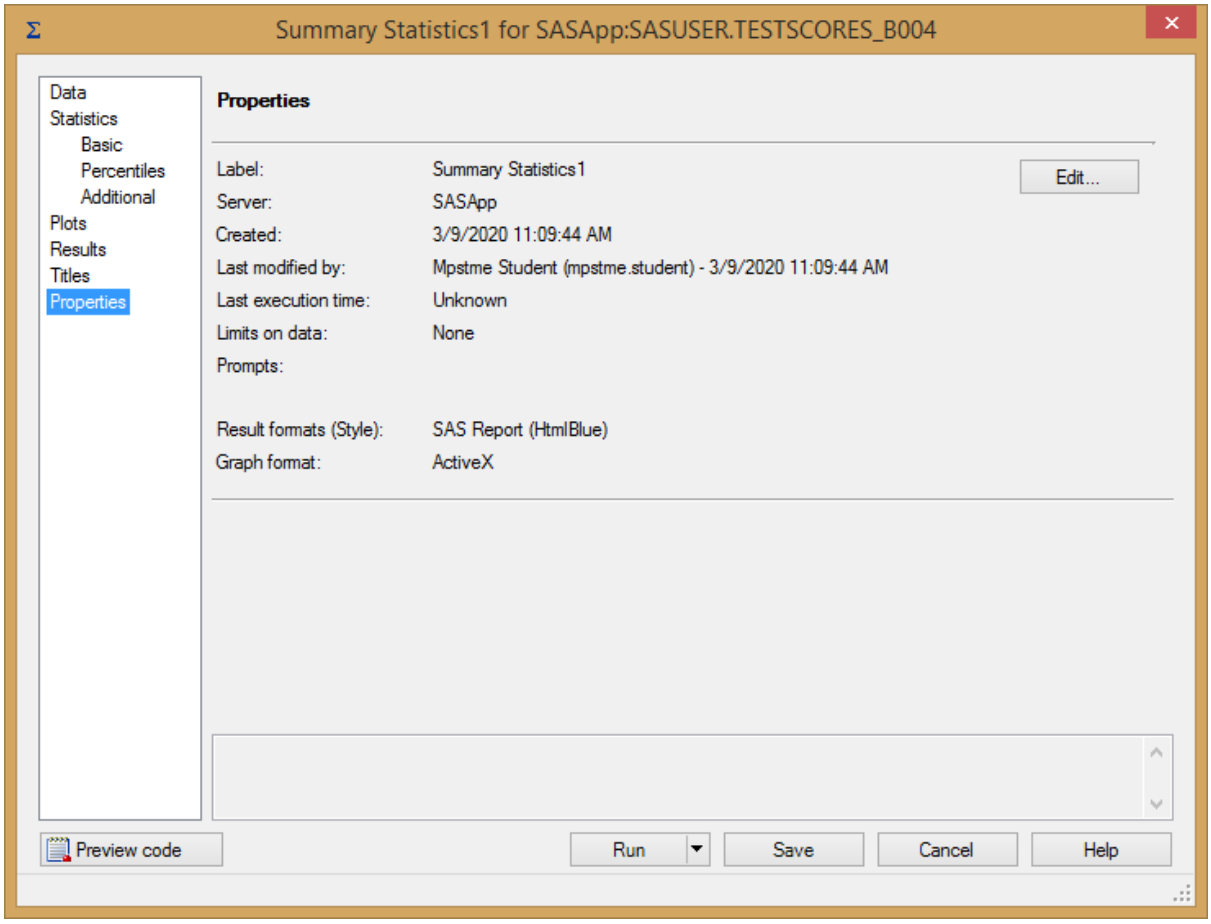
6)


Summary Statistics1 for SASApp:SASUSER.TESTSCORES_B004
Data
Statistics
Basic
Percentiles
Additional
Plots
Results
Titles
Properties
Results
Save statistics to data set
SASApp:WORK.MEANSummaryStats
Browse...
Show statistics
Value to copy for Copy Variables role:
Maximum
Combinations of classification variables:
N-way only
Specify ways:
0
Examples: 1, 3, 1...4, 0 to 4 by 2
Show Analysis labels
Suppress all displayed output
Preview code
Run
Save
Cancel
Help

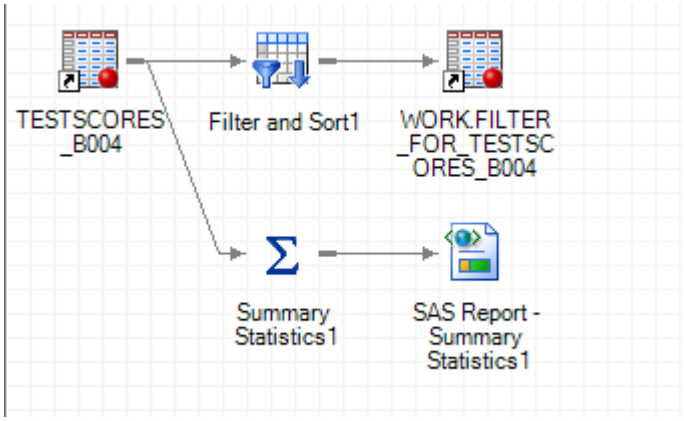
7)


Summary Statistics1 for SASApp:SASUSER.TESTSCORES_B004
Data
Statistics
Basic
Percentiles
Additional
Plots
Results
Titles
Properties
Titles
Section:
Analysis
Histogram
Box and Whisker Plot
Footnote
Checked sections will be generated based on current task settings.
Text for section: Analysis
Use default text
Summary Statistics
Results
Preview code
Run
Save
Cancel
Help

8)



Results and Conclusion:



Summary Statistics

Results

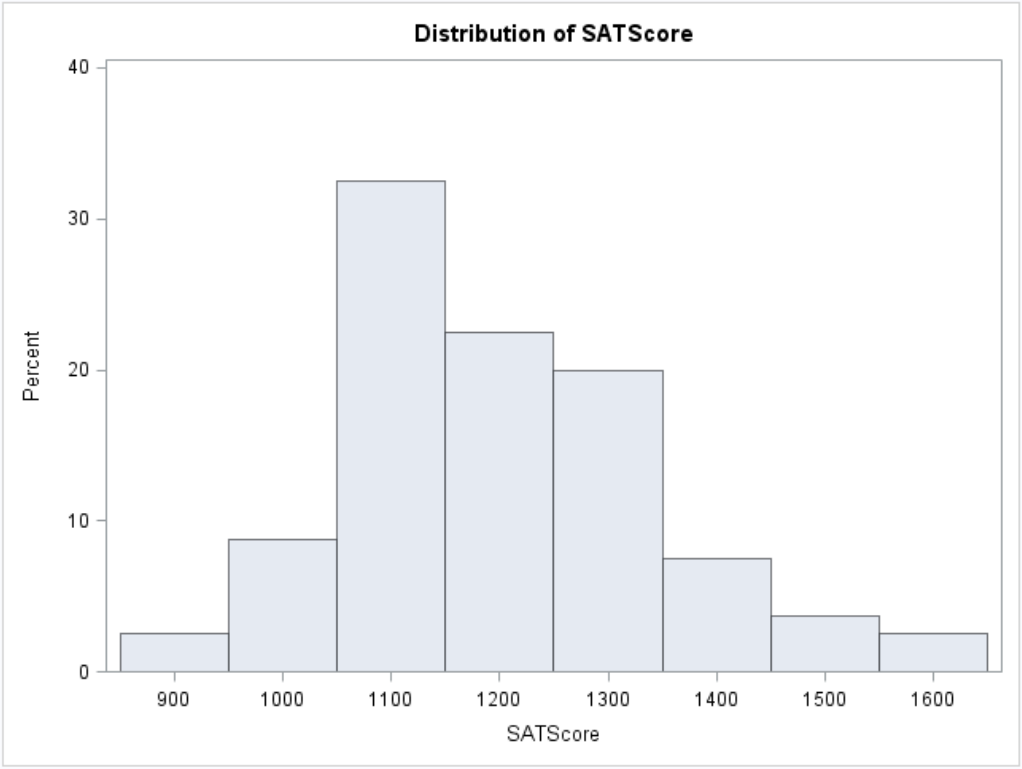
The MEANS Procedure

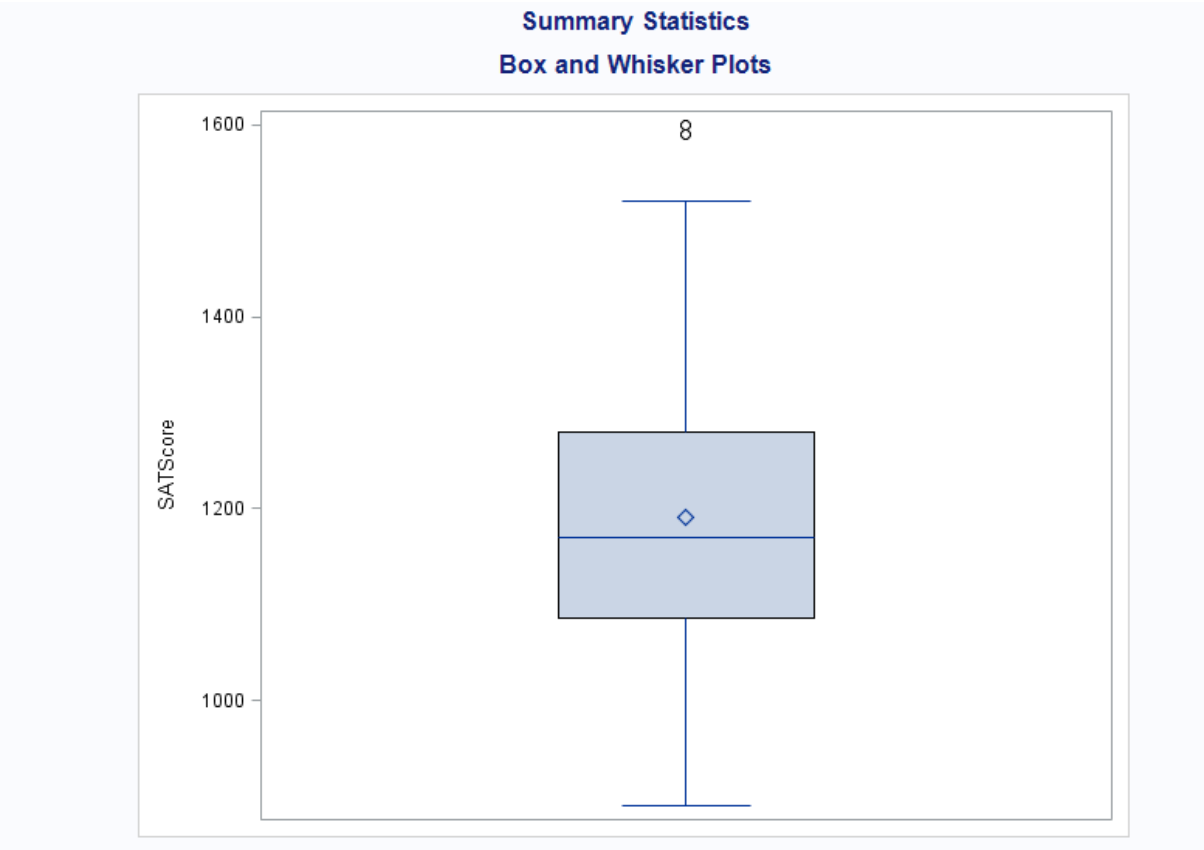
Analysis Variable : SATScore											
Mean	Std Dev	Minimum	Maximum	Range	N	N Miss	Lower Quartile	Median	Upper Quartile	Lower 95% CL for Mean	Upper 95% CL for Mean
1190.63	147.0584466	890.0000000	1600.00	710.0000000	80	0	1085.00	1170.00	1280.00	1157.90	1223.35

Summary Statistics

Histograms

The UNIVARIATE Procedure





Results and Conclusion:

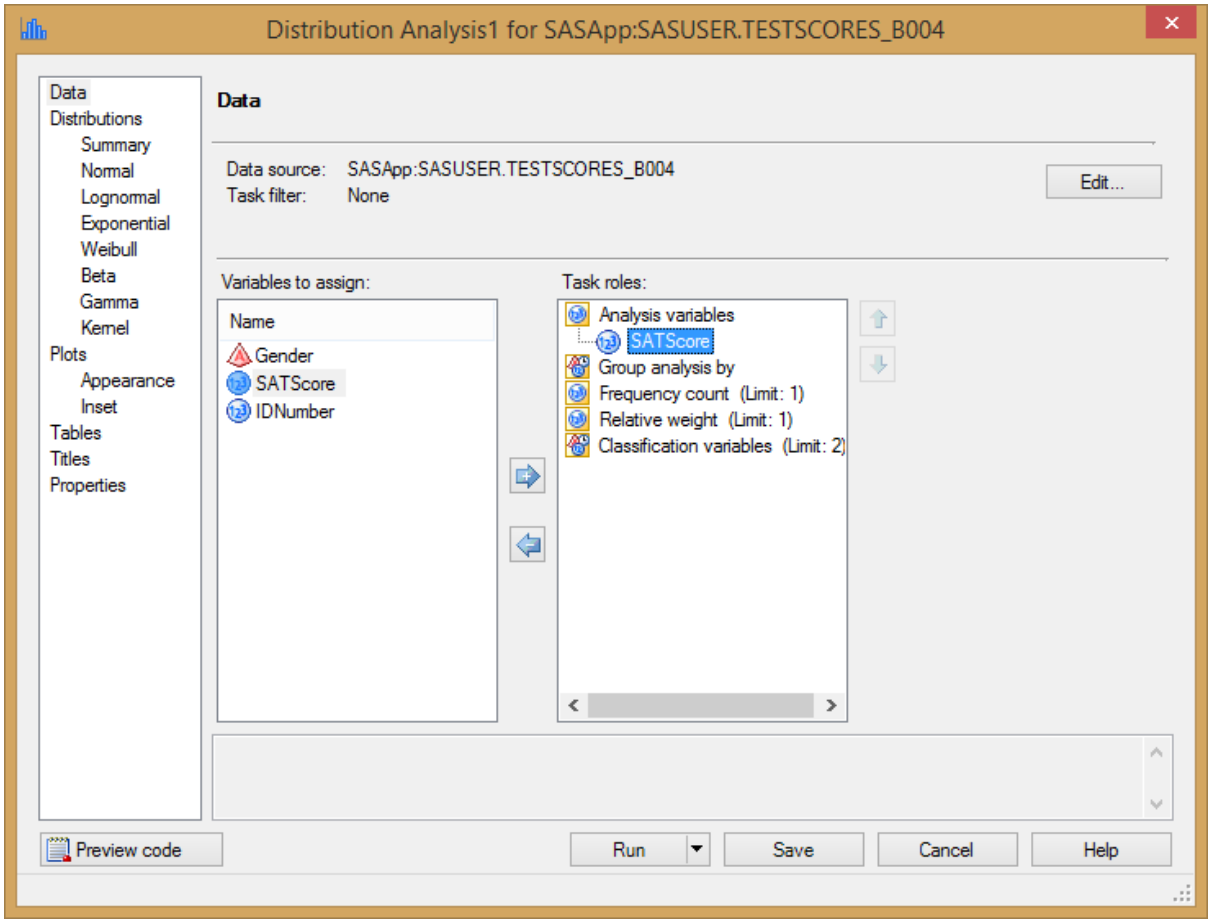
- The data ranges from a minimum of 890 to a maximum of 1600. It has 80 datapoints
- The mean and median are similar which indicates that the data points are symmetric

Distribution Analysis

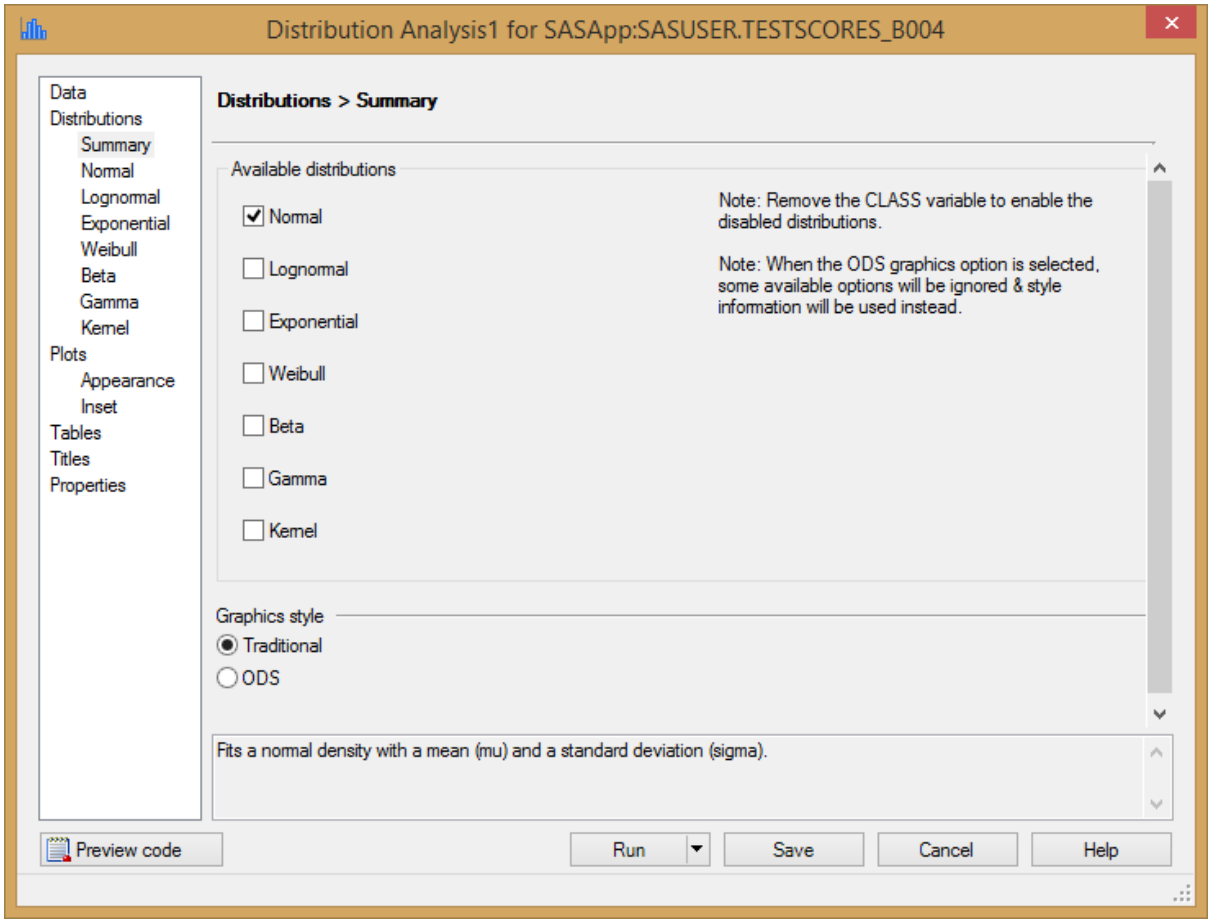
- **Tasks > Describe > Distribution Analysis**
- **Data > Analysis Variable > SAT Score**
- **Summary > Normal > Apply Distribution to all variables**
- **Appearance > Histogram, Probability, Box**
- **Tables > Basic Confidence Intervals, Basic Measures, Extreme Values, Quartiles**

Steps :

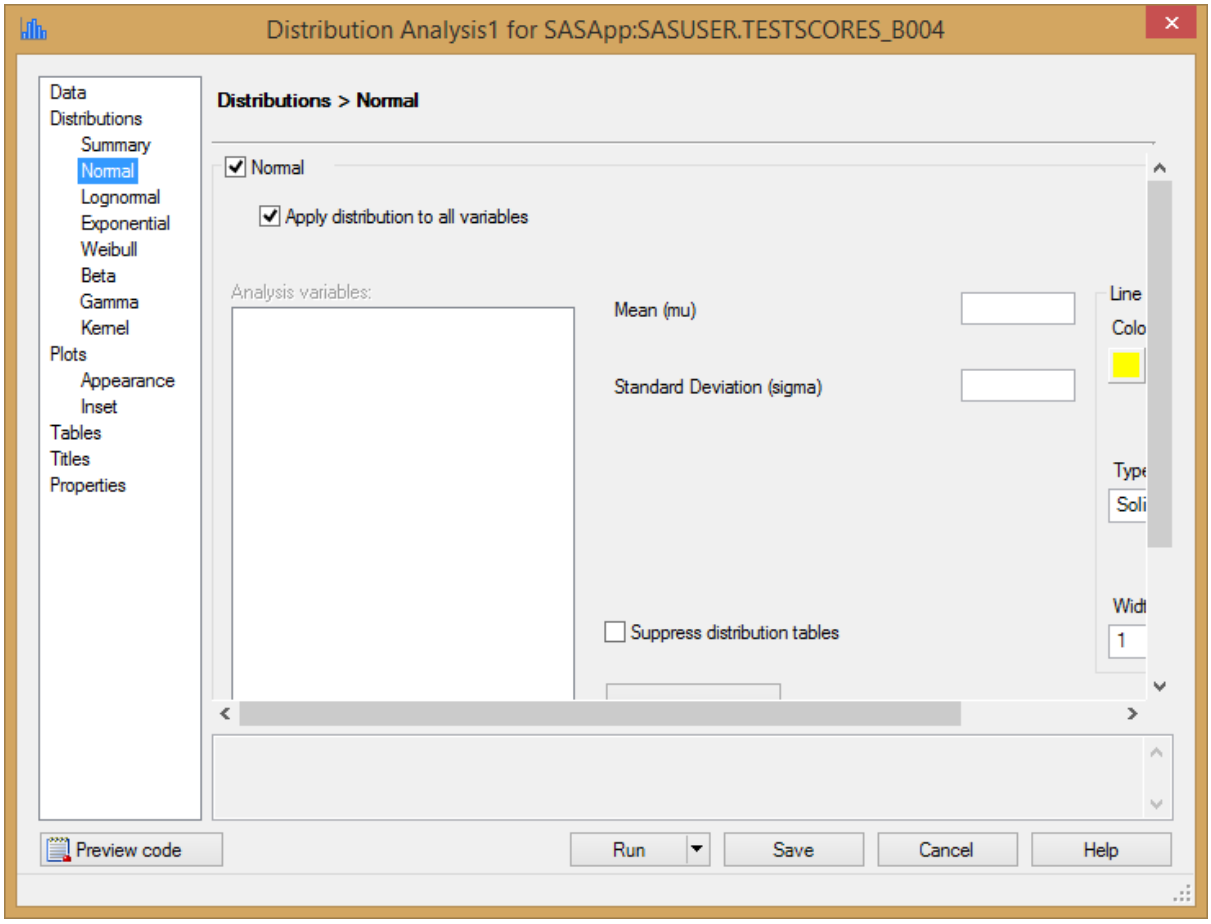
1)


Distribution Analysis1 for SASApp:SASUSER.TESTSCORES_B004
Data
Distributions
Summary
Normal
Lognormal
Exponential
Weibull
Beta
Gamma
Kernel
Plots
Appearance
Inset
Tables
Titles
Properties
Data
Data source: SASApp:SASUSER.TESTSCORES_B004
Task filter: None
Edit...
Variables to assign:
Name
Gender
SATScore
IDNumber
Task roles:
Analysis variables
SATScore
Group analysis by
Frequency count (Limit: 1)
Relative weight (Limit: 1)
Classification variables (Limit: 2)
Preview code
Run
Save
Cancel
Help

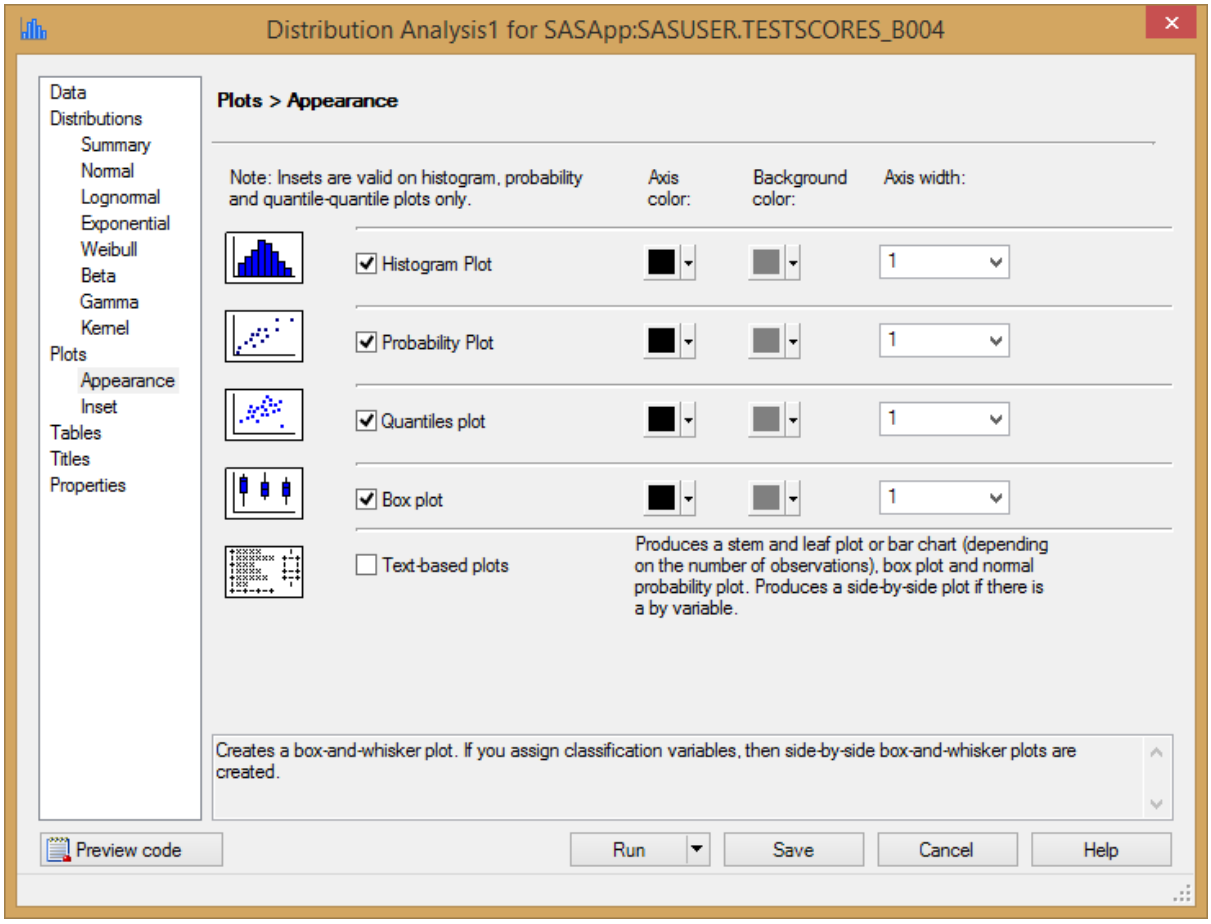
2)


Distribution Analysis1 for SASApp:SASUSER.TESTSCORES_B004
Data
Distributions
Summary
Normal
Lognormal
Exponential
Weibull
Beta
Gamma
Kernel
Plots
Appearance
Inset
Tables
Titles
Properties
Distributions > Summary
Available distributions
Normal
Lognormal
Exponential
Weibull
Beta
Gamma
Kernel
Note: Remove the CLASS variable to enable the disabled distributions.
Note: When the ODS graphics option is selected, some available options will be ignored & style information will be used instead.
Graphics style
Traditional
ODS
Fits a normal density with a mean (mu) and a standard deviation (sigma).
Preview code
Run
Save
Cancel
Help

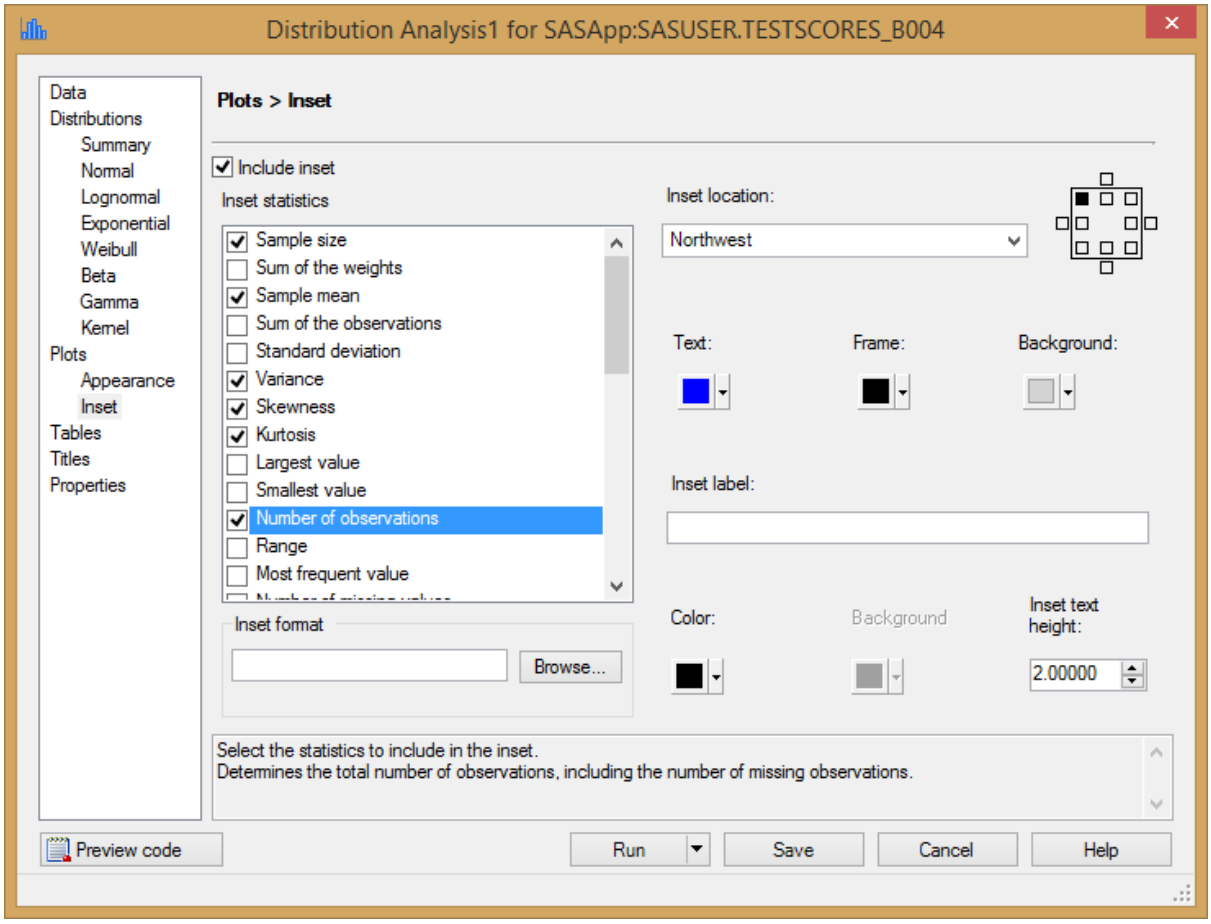
3)


Distribution Analysis1 for SASApp:SASUSER.TESTSCORES_B004
Data
Distributions
Summary
Normal
Lognormal
Exponential
Weibull
Beta
Gamma
Kernel
Plots
Appearance
Inset
Tables
Titles
Properties
Distributions > Normal
Normal
Apply distribution to all variables
Analysis variables:
Mean (mu)
Standard Deviation (sigma)
Suppress distribution tables
Line
Colo
Type
Soli
Widt
1
Preview code
Run
Save
Cancel
Help

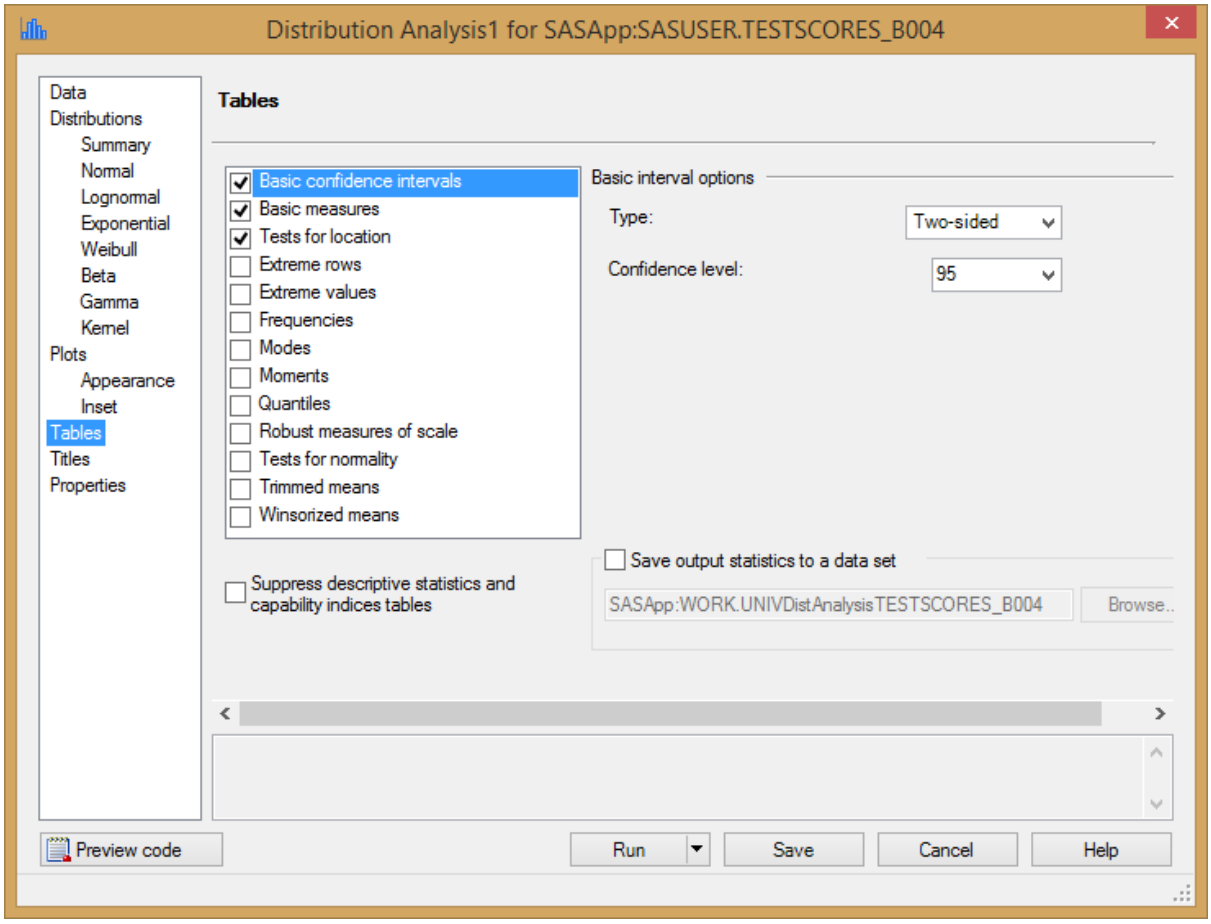
4)


Distribution Analysis1 for SASApp:SASUSER.TESTSCORES_B004
Data
Distributions
Summary
Normal
Lognormal
Exponential
Weibull
Beta
Gamma
Kernel
Plots
Appearance
Inset
Tables
Titles
Properties
Plots > Appearance
Note: Insets are valid on histogram, probability and quantile-quantile plots only.
Axis color:
Background color:
Axis width:
Histogram Plot
Probability Plot
Quantiles plot
Box plot
Text-based plots
Produces a stem and leaf plot or bar chart (depending on the number of observations), box plot and normal probability plot. Produces a side-by-side plot if there is a by variable.
Creates a box-and-whisker plot. If you assign classification variables, then side-by-side box-and-whisker plots are created.
Preview code
Run
Save
Cancel
Help

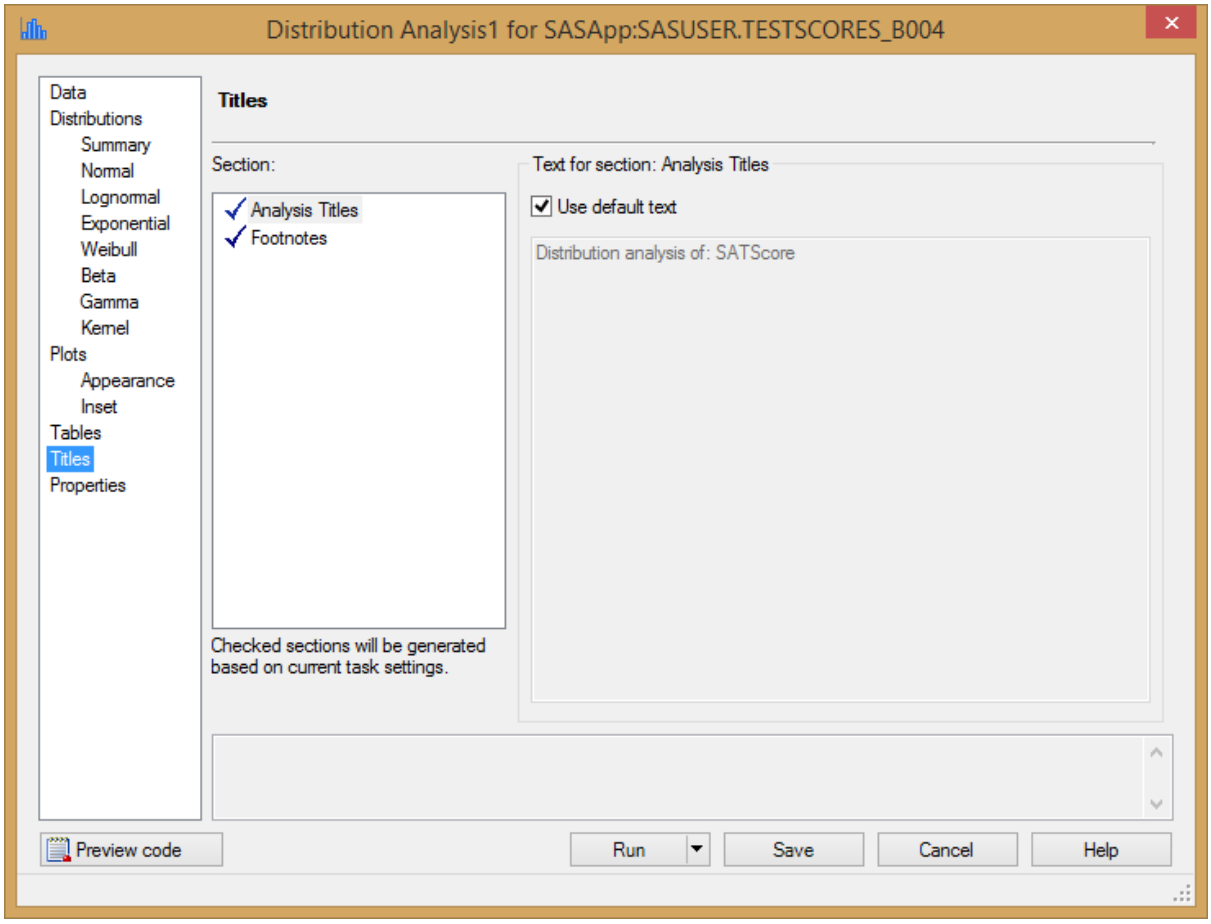
5)

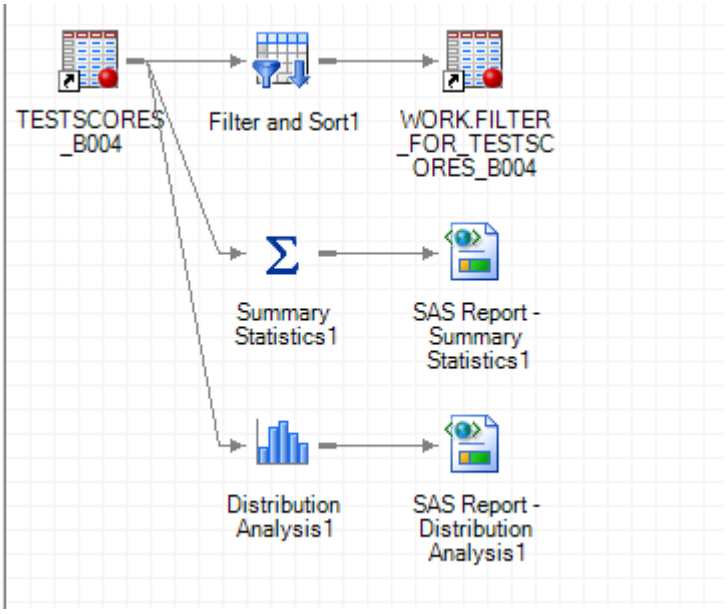

Distribution Analysis1 for SASApp:SASUSER.TESTSCORES_B004
Data
Distributions
Summary
Normal
Lognormal
Exponential
Weibull
Beta
Gamma
Kernel
Plots
Appearance
Inset
Tables
Titles
Properties
Plots > Inset
Include inset
Inset statistics
Sample size
Sum of the weights
Sample mean
Sum of the observations
Standard deviation
Variance
Skewness
Kurtosis
Largest value
Smallest value
Number of observations
Range
Most frequent value
Inset format
Browse...
Inset location:
Northwest
Text:
Frame:
Background:
Inset label:
Color:
Background
Inset text height:
2.00000
Select the statistics to include in the inset.
Determines the total number of observations, including the number of missing observations.
Preview code
Run
Save
Cancel
Help

6)


Distribution Analysis1 for SASApp:SASUSER.TESTSCORES_B004
Data
Distributions
Summary
Normal
Lognormal
Exponential
Weibull
Beta
Gamma
Kernel
Plots
Appearance
Inset
Tables
Titles
Properties
Tables
Basic confidence intervals
Basic measures
Tests for location
Extreme rows
Extreme values
Frequencies
Modes
Moments
Quantiles
Robust measures of scale
Tests for normality
Trimmed means
Winsorized means
Basic interval options
Type:
Two-sided
Confidence level:
95
Save output statistics to a data set
SASApp:WORK.UNIVDistAnalysisTESTSCORES_B004
Browse...
Suppress descriptive statistics and capability indices tables
Preview code
Run
Save
Cancel
Help

7)


Distribution Analysis1 for SASApp:SASUSER.TESTSCORES_B004
Data
Distributions
Summary
Normal
Lognormal
Exponential
Weibull
Beta
Gamma
Kernel
Plots
Appearance
Inset
Tables
Titles
Properties
Titles
Section:
Analysis Titles
Footnotes
Checked sections will be generated based on current task settings.
Text for section: Analysis Titles
Use default text
Distribution analysis of: SATScore
Preview code
Run
Save
Cancel
Help



Results and Conclusion:

Distribution analysis of: SATScore

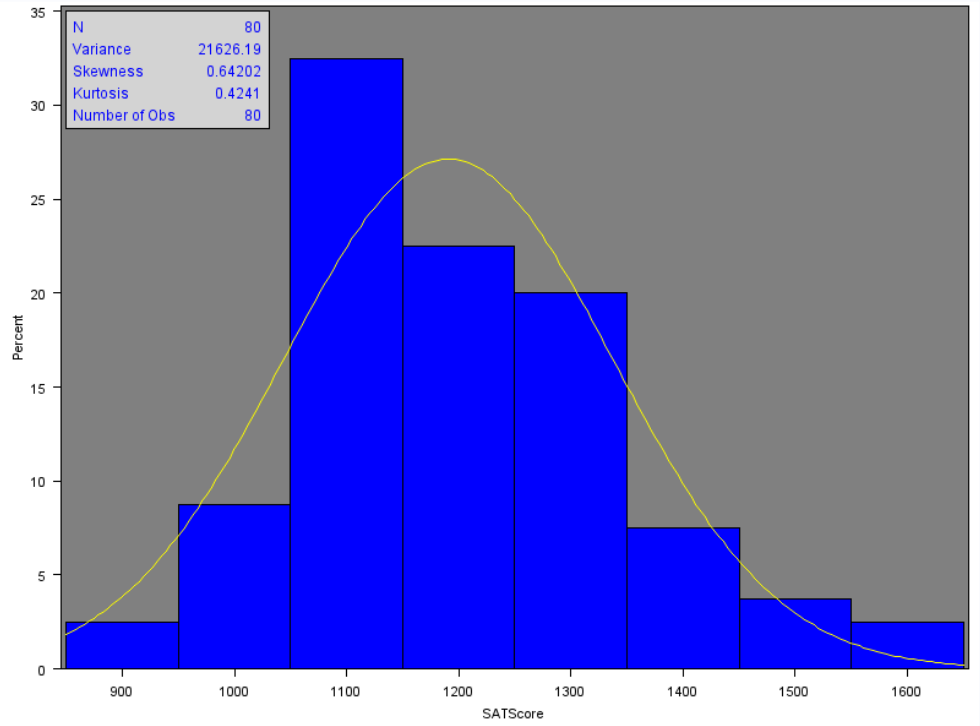
The UNIVARIATE Procedure
Variable: SATScore

Basic Statistical Measures			
Location		Variability	
Mean	1190.625	Std Deviation	147.05845
Median	1170.000	Variance	21626
Mode	1050.000	Range	710.00000
		Interquartile Range	195.00000

Basic Confidence Limits Assuming Normality			
Parameter	Estimate	95% Confidence Limits	
Mean	1191	1158	1223
Std Deviation	147.05845	127.27215	174.18670
Variance	21626	16198	30341

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	72.41525	Pr > |t|	<.0001
Sign	M	40	Pr >= |M|	<.0001
Signed Rank	S	1620	Pr >= |S|	<.0001

The UNIVARIATE Procedure

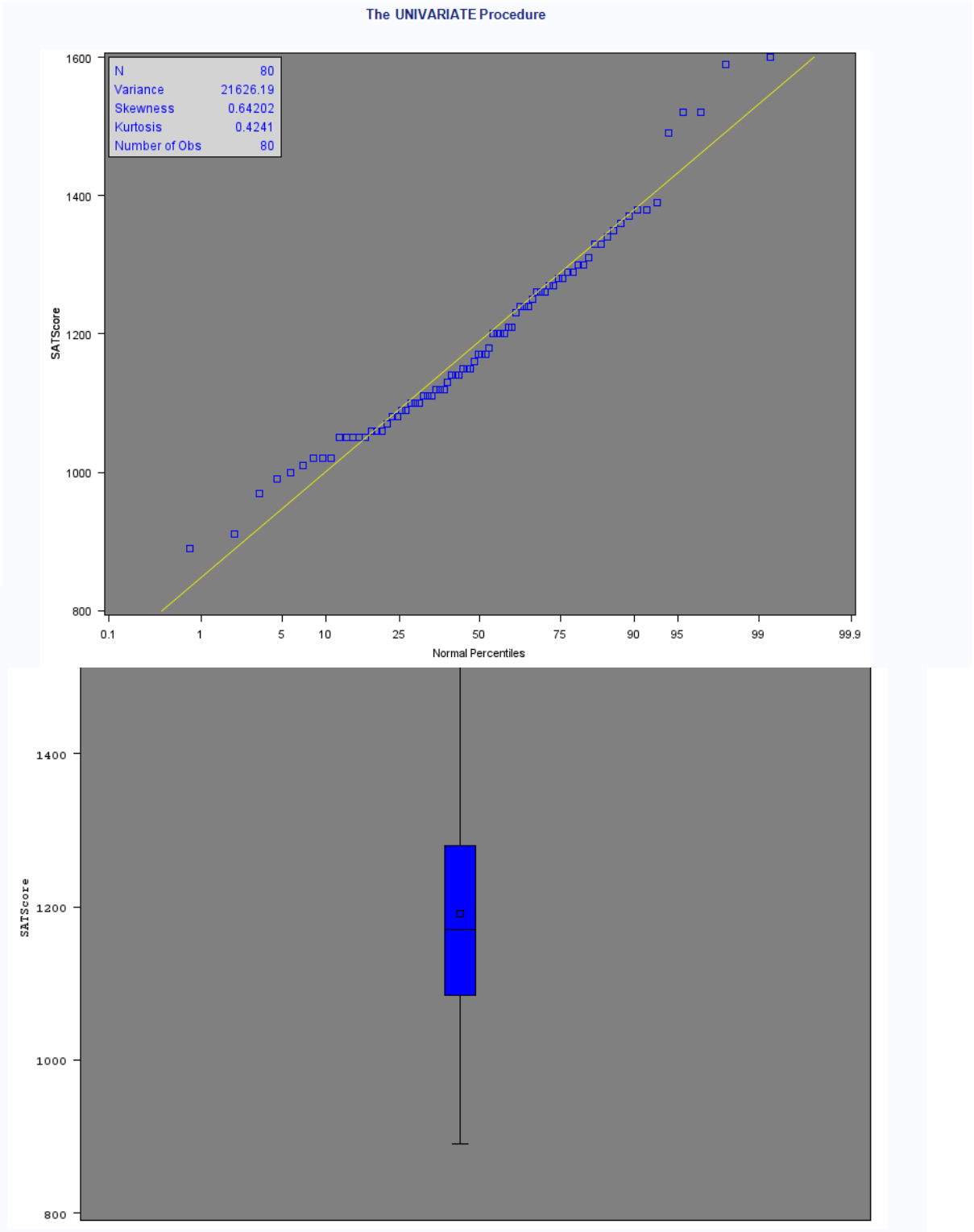


The UNIVARIATE Procedure
Fitted Normal Distribution for SATScore

Parameters for Normal Distribution		
Parameter	Symbol	Estimate
Mean	Mu	1190.625
Std Dev	Sigma	147.0584

Goodness-of-Fit Tests for Normal Distribution				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.08382224	Pr > D	>0.150
Cramer-von Mises	W-Sq	0.09964577	Pr > W-Sq	0.114
Anderson-Darling	A-Sq	0.70124822	Pr > A-Sq	0.068

Quantiles for Normal Distribution		
	Quantile	
Percent	Observed	Estimated
1.0	890.000	848.516
5.0	995.000	948.735
10.0	1020.000	1002.162
25.0	1085.000	1091.436
50.0	1170.000	1190.625
75.0	1280.000	1289.814
90.0	1375.000	1379.088
95.0	1505.000	1432.515
99.0	1600.000	1532.734



Results and Conclusion:

- Mean of the data is 1190.625 is almost equal to the median (1170), which shows that the distribution is fairly symmetric.

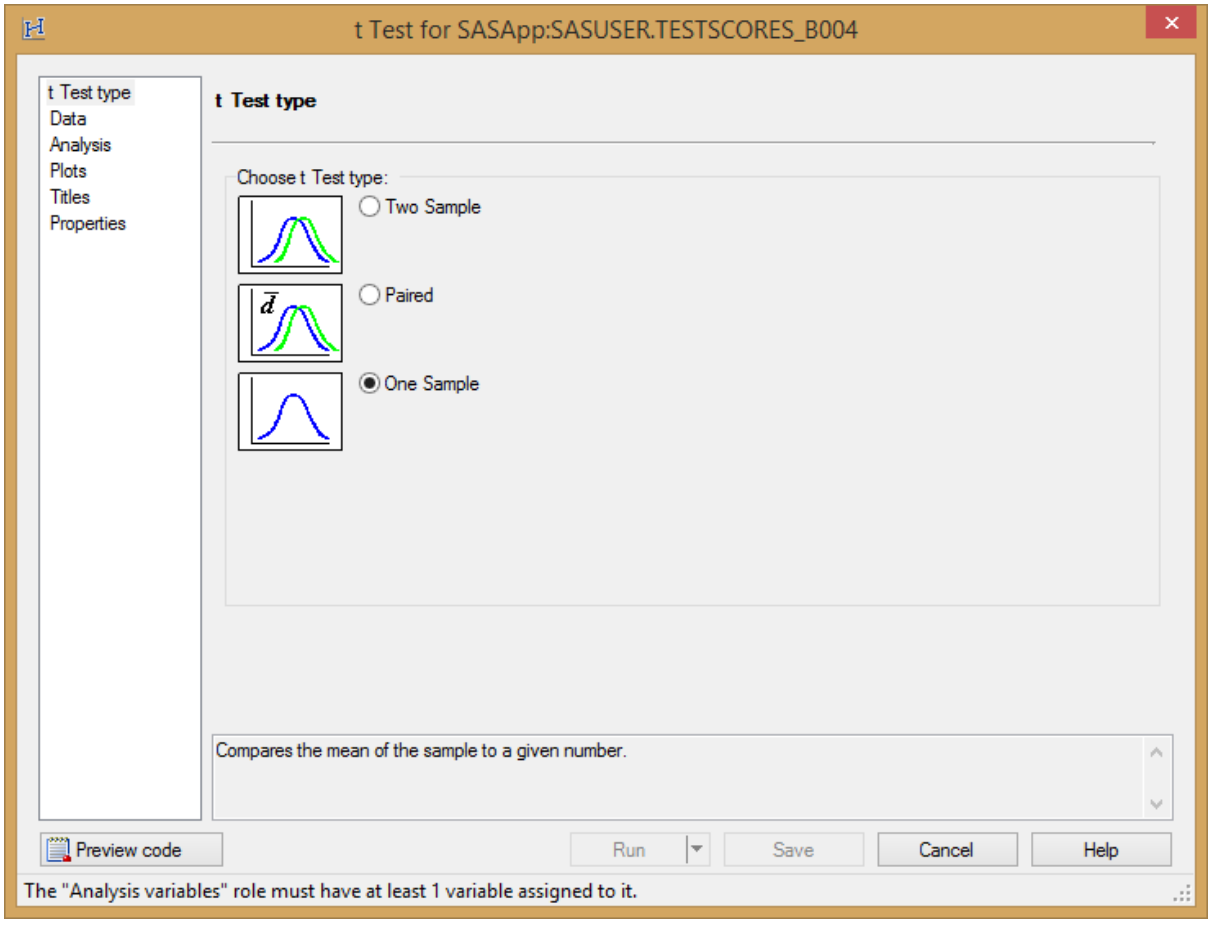
- Standard deviation of the data is oberseved to be 147.05, which means that the average variability around the mean is approximately 147.
- Distribution is slightly skewed to the right since skewness is positive .
- Distribution has slightly heavier tails than the normal distribution since kurtosis is positive and equal to +0.42.
- Range of the student score observed is from 890 to 1600

t-Test One Sample

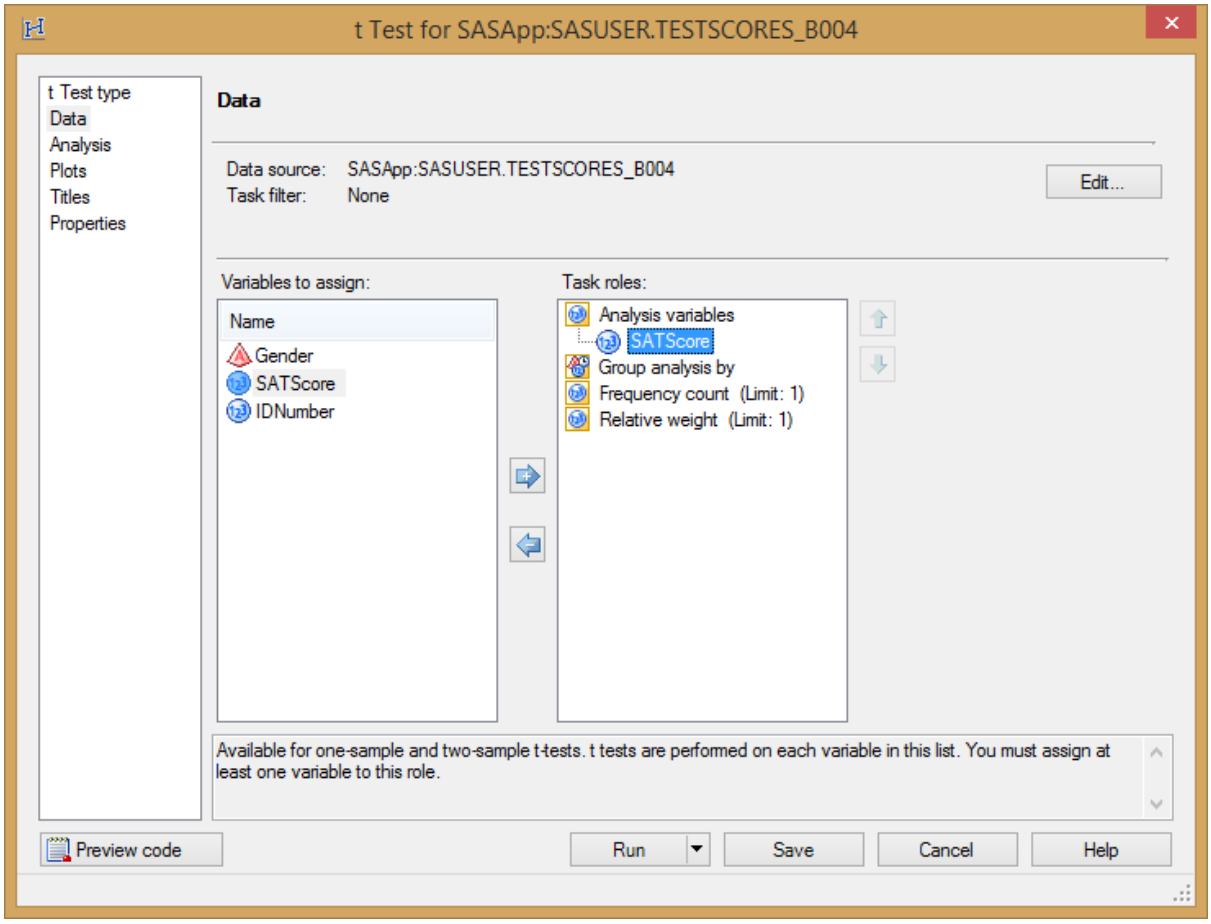
- **Tasks > ANOVA > tTest**
- **Type > One Sample**
- **Data > Analysis variable > SAT Score**
- **Analysis (h0) = 1200**
- **Confidence = 95**
- **Plot > Summary**

Steps:

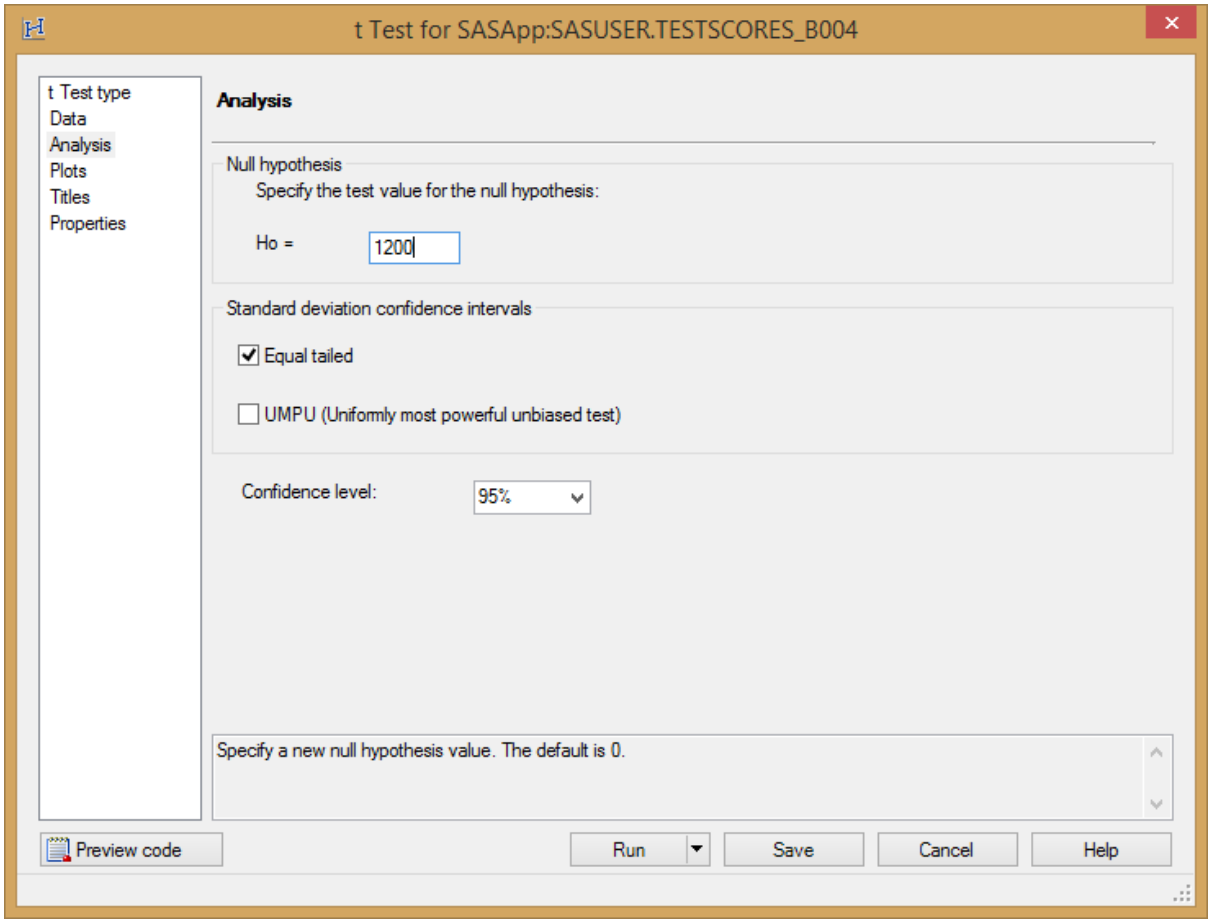
1)



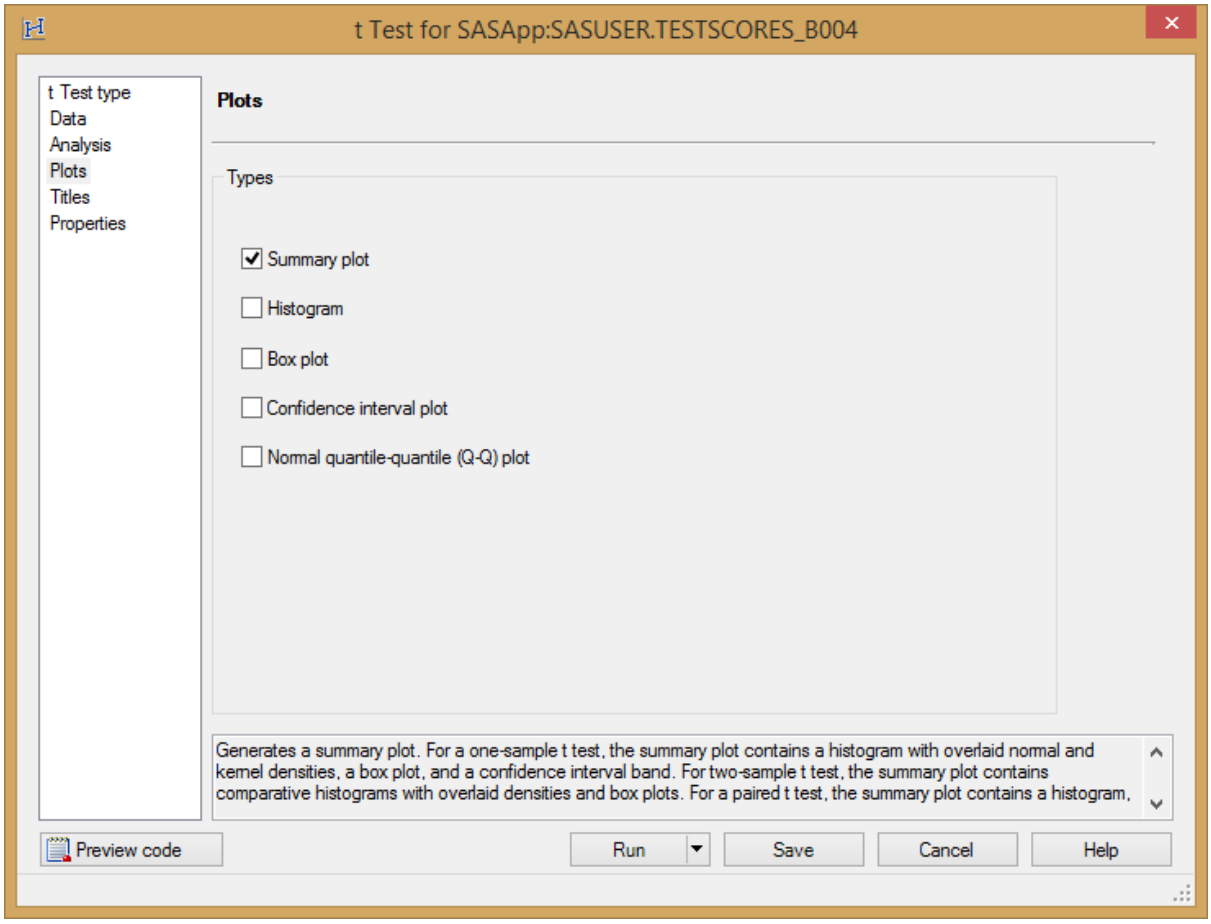
2)


t Test for SASApp:SASUSER.TESTSCORES_B004
t Test type
Data
Analysis
Plots
Titles
Properties
Data
Data source: SASApp:SASUSER.TESTSCORES_B004
Task filter: None
Edit...
Variables to assign:
Name
Gender
SATScore
IDNumber
Task roles:
Analysis variables
SATScore
Group analysis by
Frequency count (Limit: 1)
Relative weight (Limit: 1)
Available for one-sample and two-sample t-tests. t tests are performed on each variable in this list. You must assign at least one variable to this role.
Preview code
Run
Save
Cancel
Help

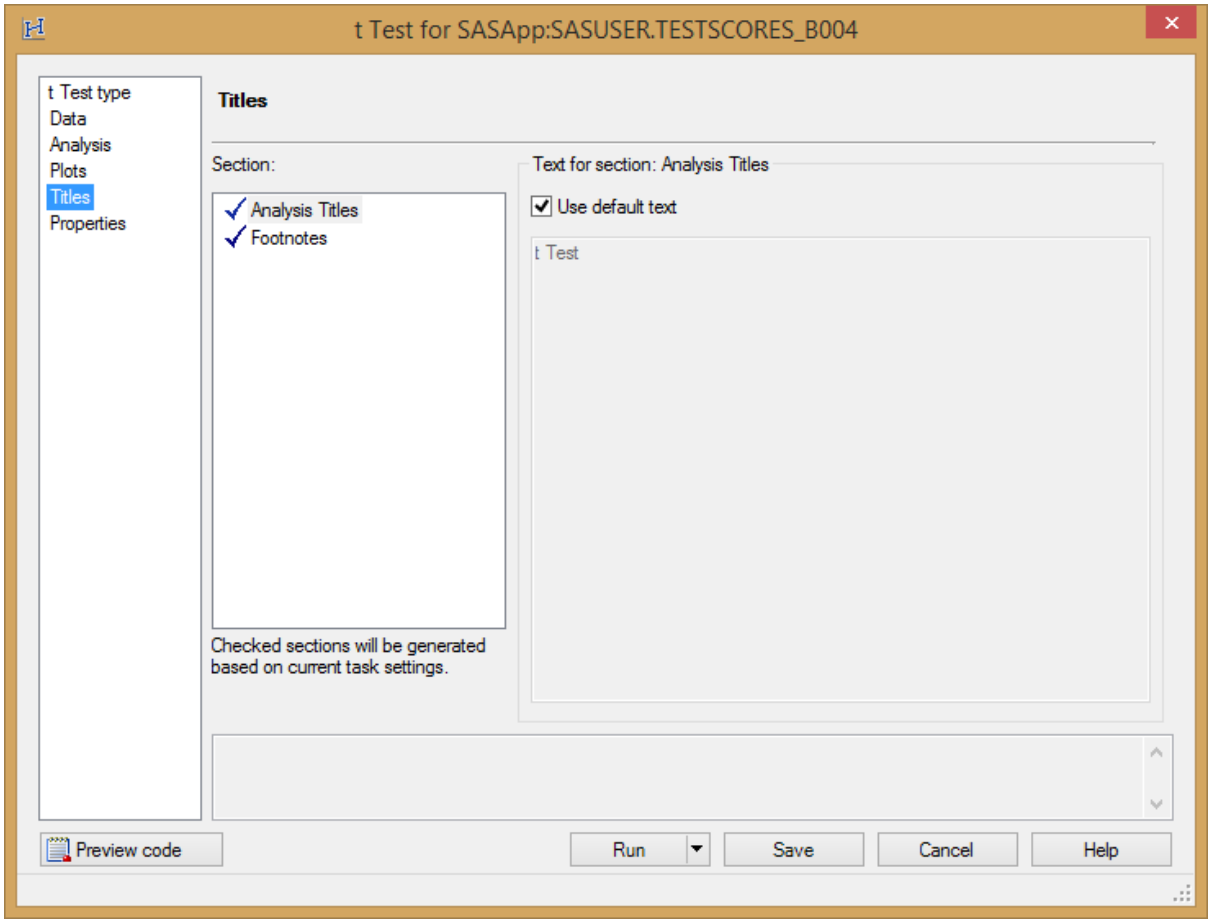
3)


t Test for SASApp:SASUSER.TESTSCORES_B004
t Test type
Data
Analysis
Plots
Titles
Properties
Analysis
Null hypothesis
Specify the test value for the null hypothesis:
Ho =
1200
Standard deviation confidence intervals
Equal tailed
UMPU (Uniformly most powerful unbiased test)
Confidence level:
95%
Specify a new null hypothesis value. The default is 0.
Preview code
Run
Save
Cancel
Help

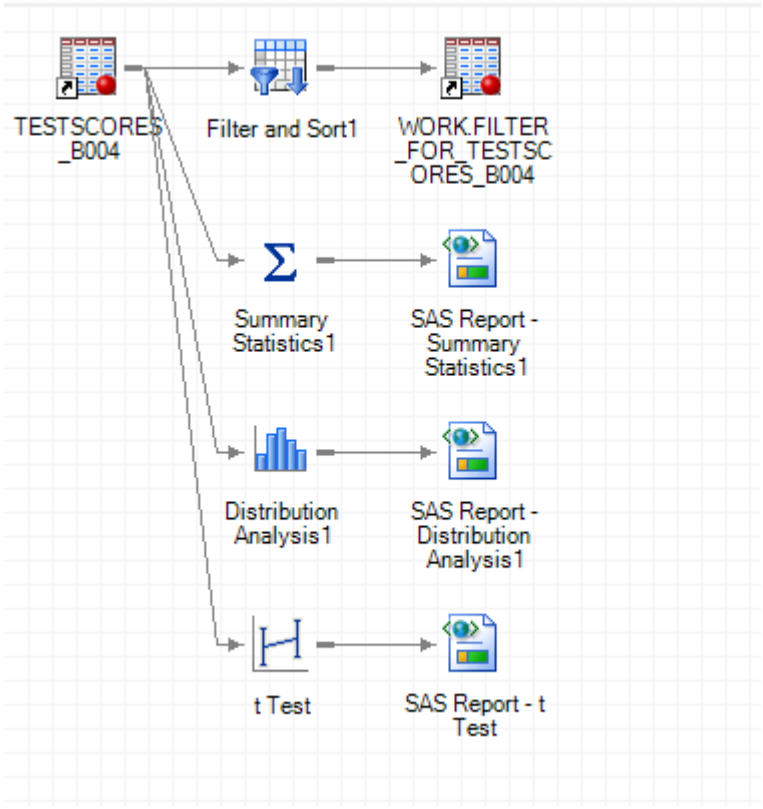
4)


t Test for SASApp:SASUSER.TESTSCORES_B004
t Test type
Data
Analysis
Plots
Titles
Properties
Plots
Types
Summary plot
Histogram
Box plot
Confidence interval plot
Normal quantile-quantile (Q-Q) plot
Generates a summary plot. For a one-sample t test, the summary plot contains a histogram with overlaid normal and kernel densities, a box plot, and a confidence interval band. For two-sample t test, the summary plot contains comparative histograms with overlaid densities and box plots. For a paired t test, the summary plot contains a histogram,
Preview code
Run
Save
Cancel
Help

5)



Results and Conclusion:



t Test

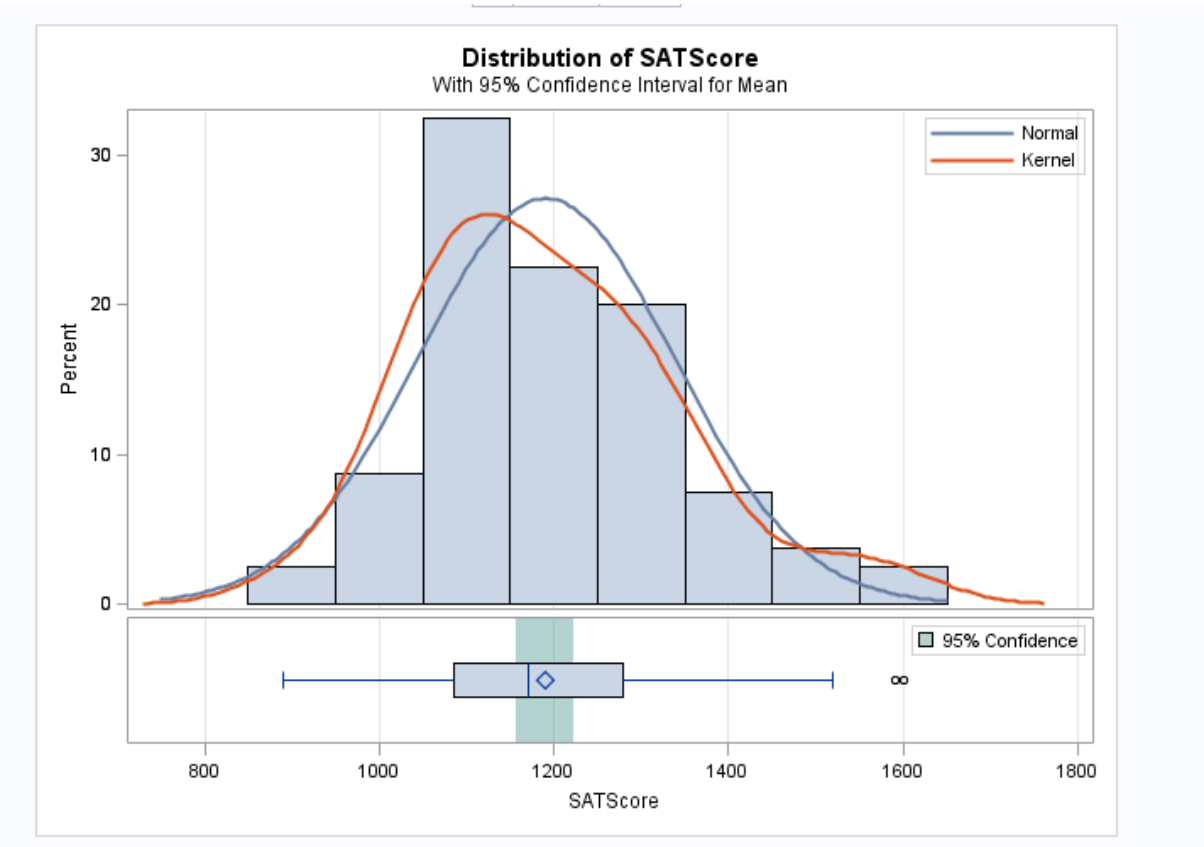
The TTEST Procedure

Variable: SATScore

N	Mean	Std Dev	Std Err	Minimum	Maximum
80	1190.6	147.1	16.4416	890.0	1600.0

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
1190.6	1157.9	1223.4	147.1	127.3	174.2

DF	t Value	Pr > |t|
79	-0.57	0.5702



Results and Conclusion:

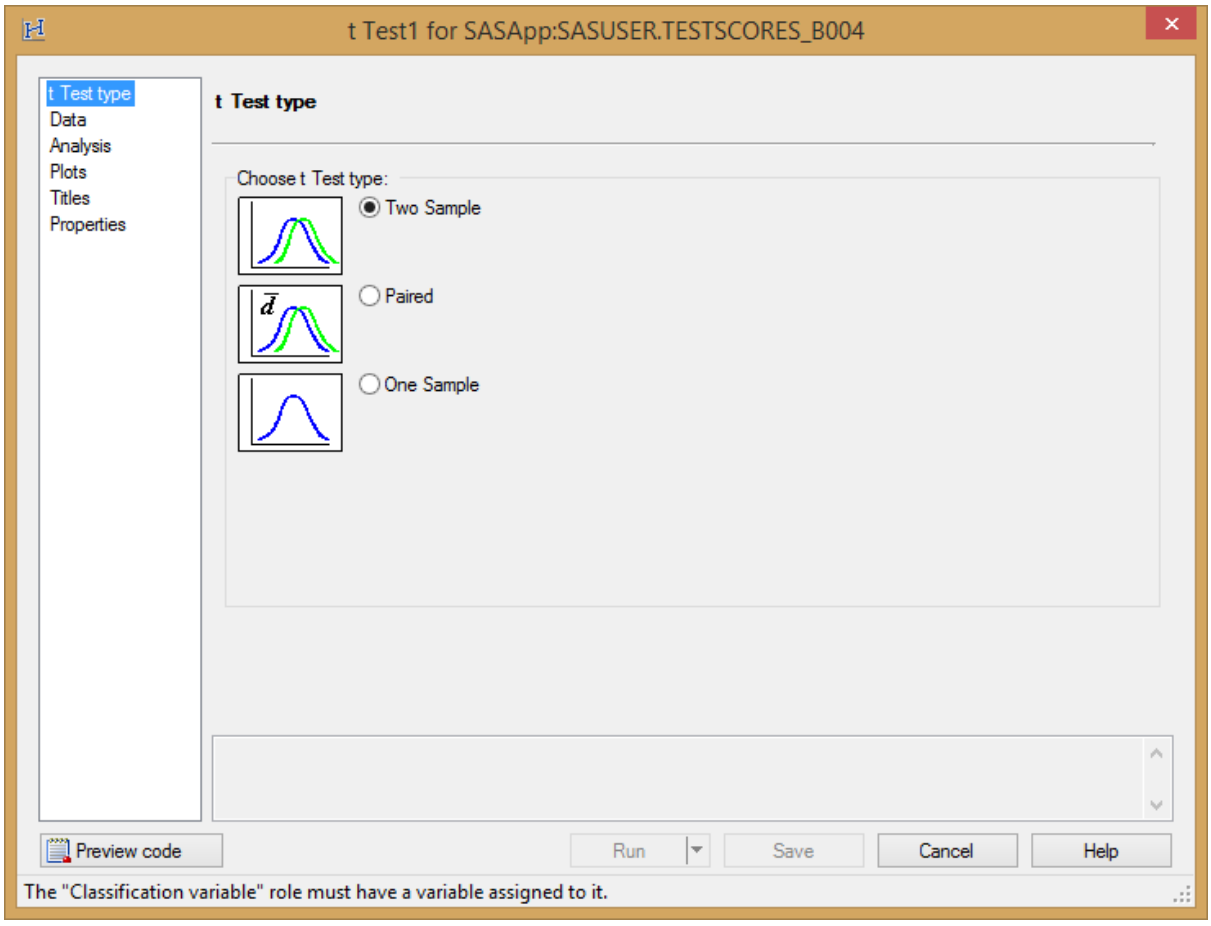
- p value is 0.5702. Since it is more than the significance level of alpha=0.5, this value is not statistically significant.
- Histogram along with the normal and kernel density curves are produced on one plot, along with a horizontal box plot with a shaded confidence interval for the mean included.
- Normal curve line indicated by blue is slightly shifted to right as compared to the kernel red line. The mean value is observed to be closer to 1200.

t-Test Two Sample

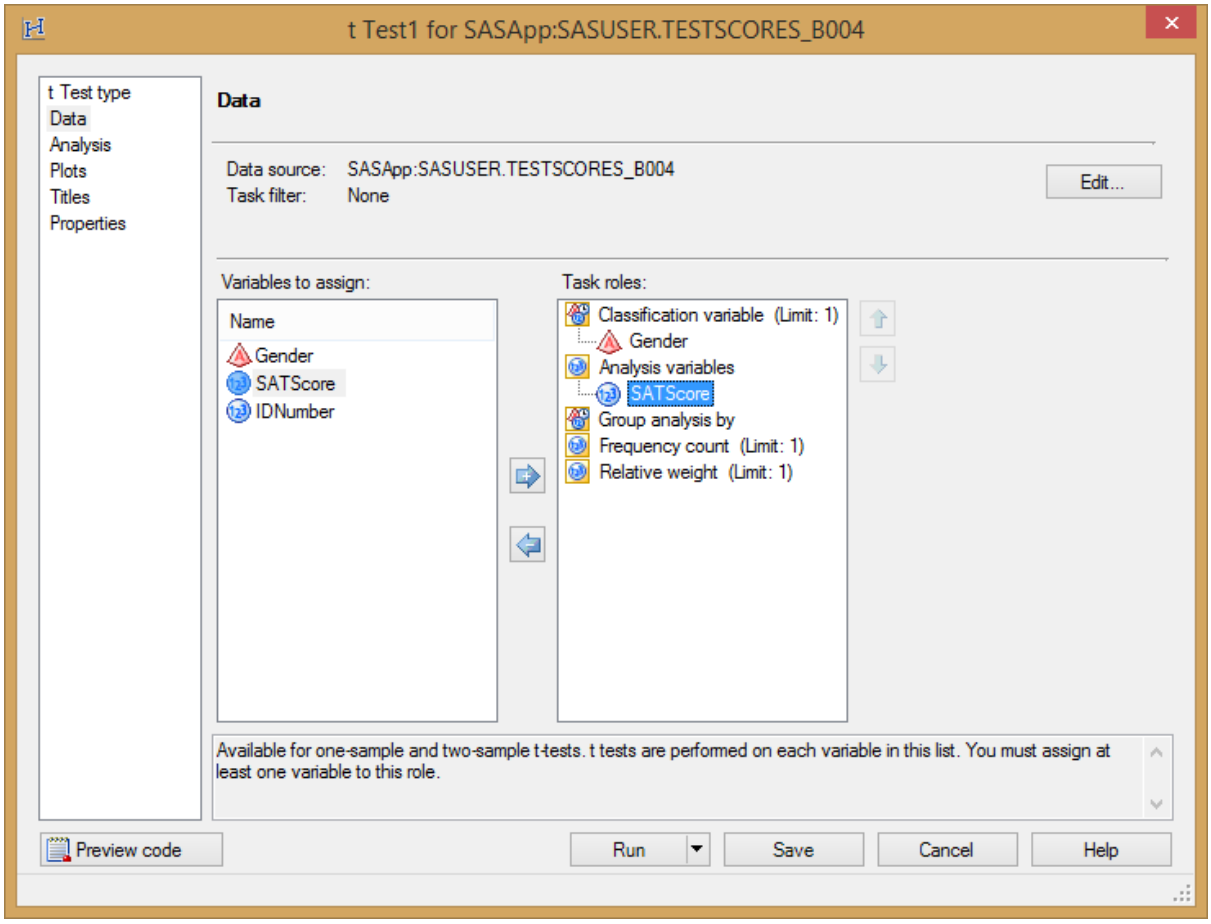
- **Tasks > ANOVA > tTest**
- **Type > Two Sample**
- **Data > Analysis variable > SAT Score**
- **Data > Classification Variable > Gender**
- **Analysis (h0) = 0**
- **Confidence = 95**
- **Plot > Summary**

Steps:

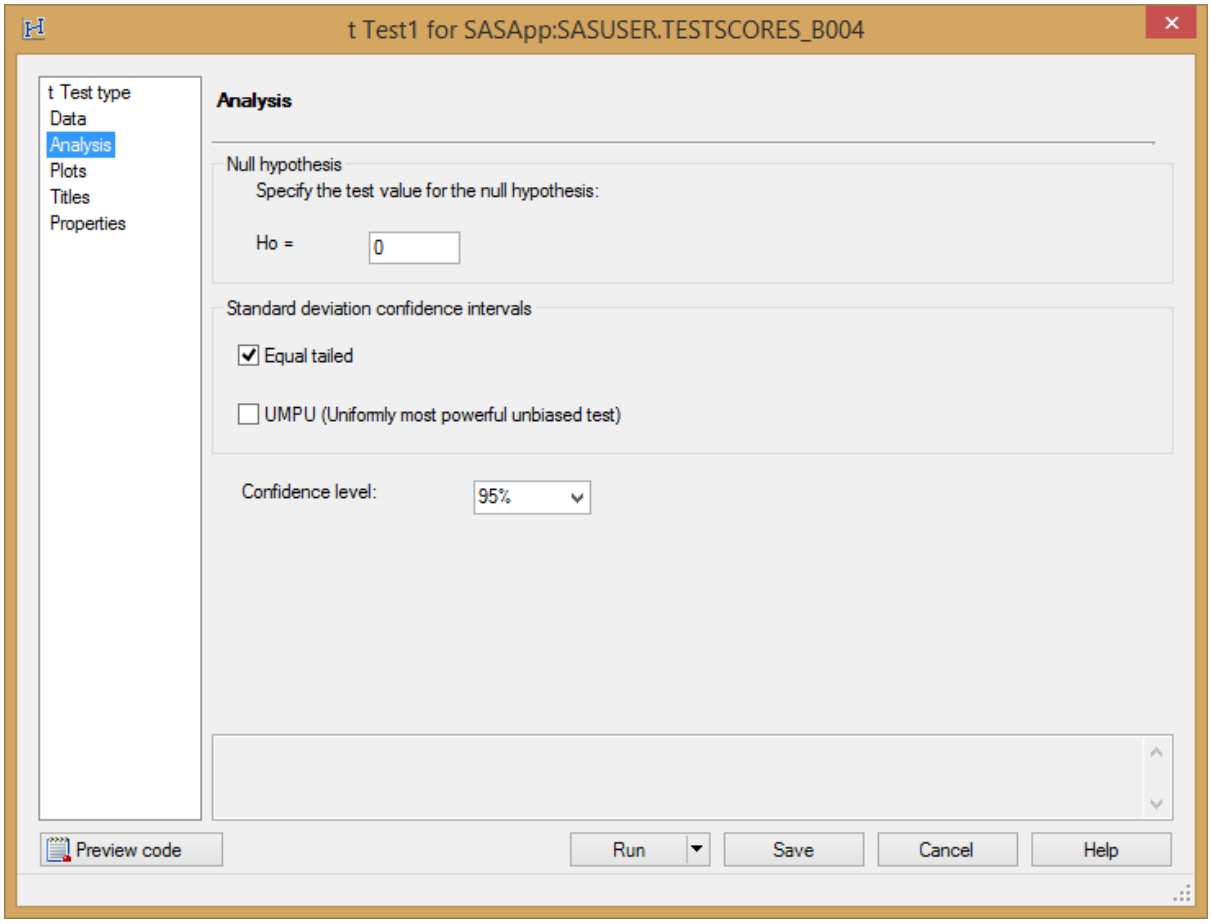
1)



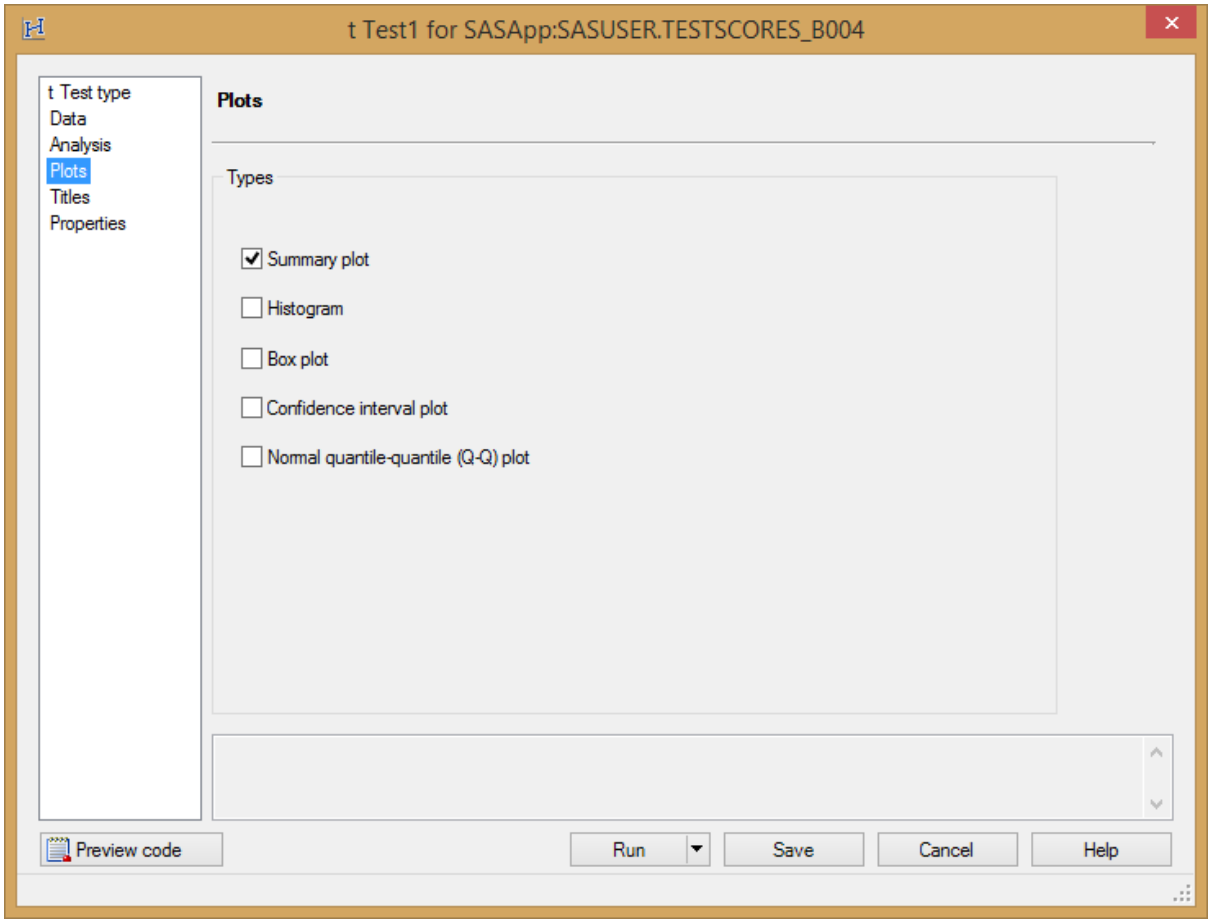
2)


t Test1 for SASApp:SASUSER.TESTSCORES_B004
t Test type
Data
Analysis
Plots
Titles
Properties
Data
Data source: SASApp:SASUSER.TESTSCORES_B004
Task filter: None
Edit...
Variables to assign:
Name
Gender
SATScore
IDNumber
Task roles:
Classification variable (Limit: 1)
Gender
Analysis variables
SATScore
Group analysis by
Frequency count (Limit: 1)
Relative weight (Limit: 1)
Available for one-sample and two-sample t-tests. t tests are performed on each variable in this list. You must assign at least one variable to this role.
Preview code
Run
Save
Cancel
Help

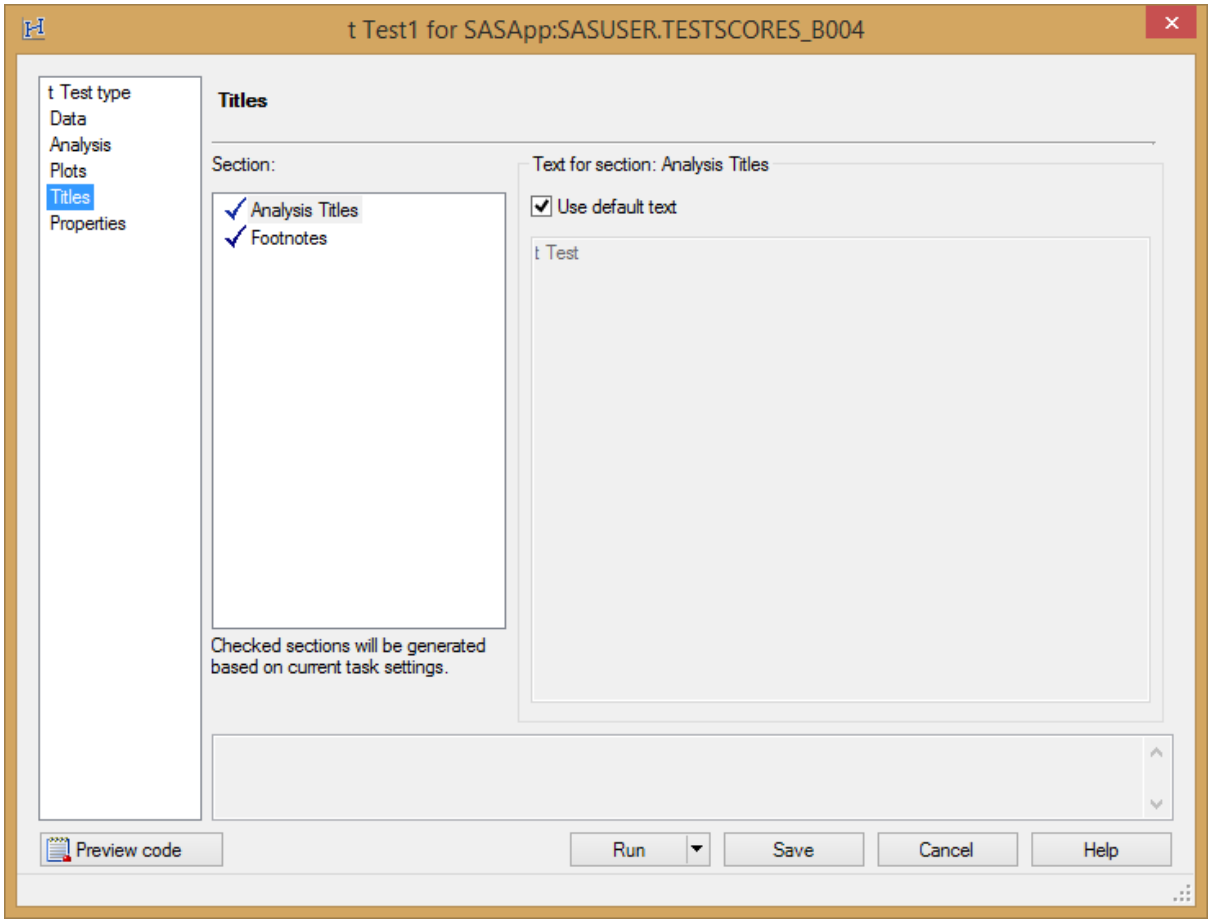
3)


t Test1 for SASApp:SASUSER.TESTSCORES_B004
t Test type
Data
Analysis
Plots
Titles
Properties
Analysis
Null hypothesis
Specify the test value for the null hypothesis:
Ho =
0
Standard deviation confidence intervals
Equal tailed
UMPU (Uniformly most powerful unbiased test)
Confidence level:
95%
Preview code
Run
Save
Cancel
Help

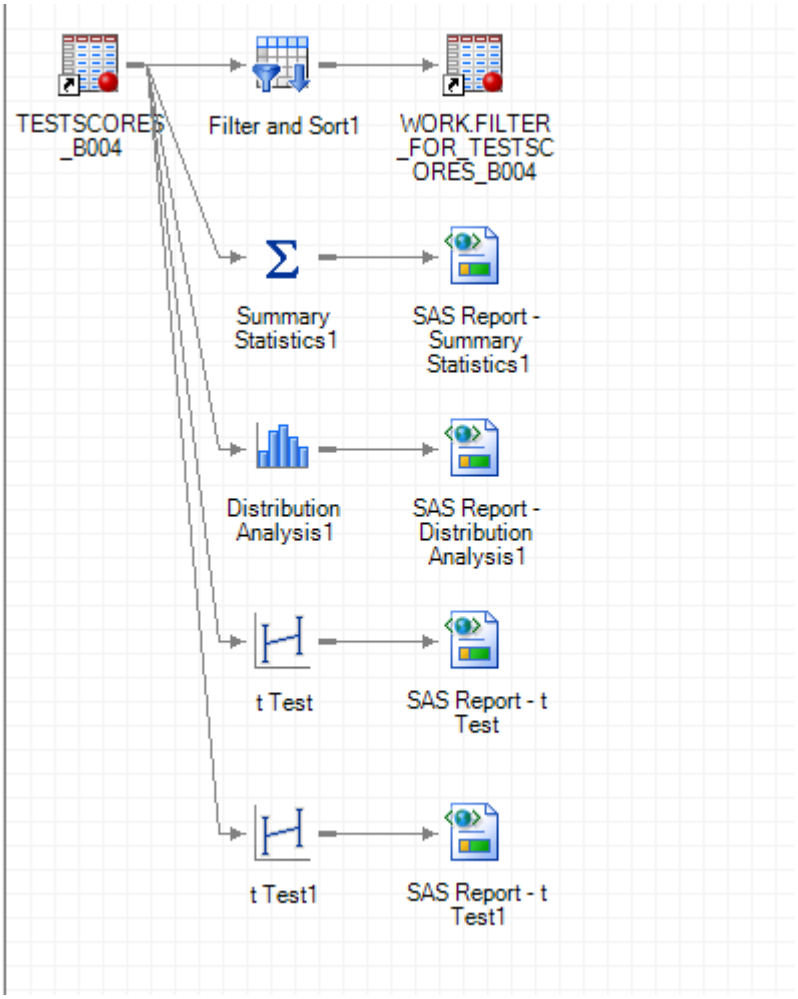
4)


t Test1 for SASApp:SASUSER.TESTSCORES_B004
t Test type
Data
Analysis
Plots
Titles
Properties
Plots
Types
Summary plot
Histogram
Box plot
Confidence interval plot
Normal quantile-quantile (Q-Q) plot
Preview code
Run
Save
Cancel
Help

5)



Results and Conclusion:



t Test

The TTEST Procedure

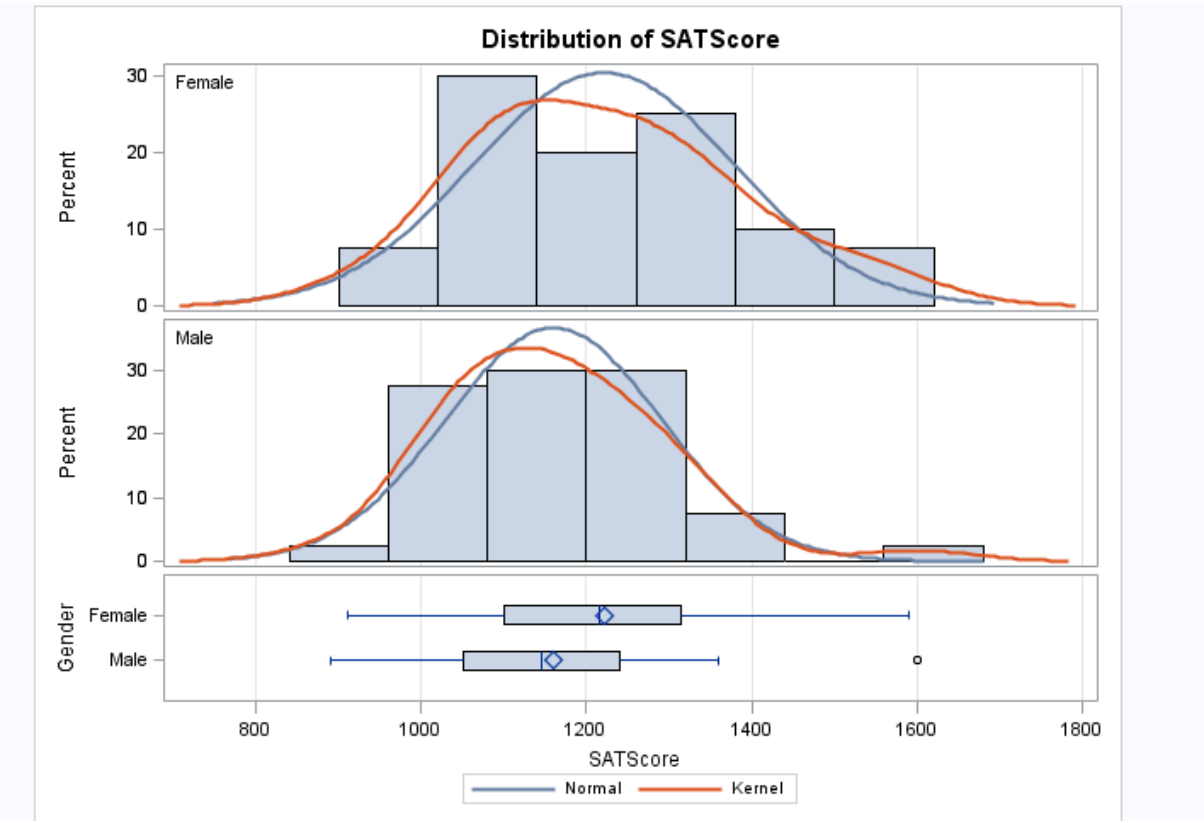
Variable: SATScore

Gender	N	Mean	Std Dev	Std Err	Minimum	Maximum
Female	40	1221.0	157.4	24.8864	910.0	1590.0
Male	40	1160.3	130.9	20.7008	890.0	1600.0
Diff (1-2)		60.7500	144.8	32.3706		

Gender	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
Female		1221.0	1170.7	1271.3	157.4	128.9	202.1
Male		1160.3	1118.4	1202.1	130.9	107.2	168.1
Diff (1-2)	Pooled	60.7500	-3.6950	125.2	144.8	125.2	171.7
Diff (1-2)	Satterthwaite	60.7500	-3.7286	125.2			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	78	1.88	0.0643
Satterthwaite	Unequal	75.497	1.88	0.0644

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	39	39	1.45	0.2545



Results and Conclusion:

- Here two sample t test is conducted between Gender and SAT Score
- The classification variable is taken as Gender and the Analysis Variable is taken as SAT Score to understand the correlation between the scores obtained and the test takers gender
- The t-test value for both Pooled and Satterthwaite method gave us 1.88 respectively and the corresponding p value was 0.064 for both methods.
- Because the p-value is more than the confidence interval of 0.05 we can accept the null hypothesis and the distribution plot can be used to understand the distribution of the SAT Score by both male and female test takers.