Name: Amit Birajdar

Class: BTech CS-B

Roll No.: B014

PREDICTIVE MODELLING - ASSIGNMENT 1

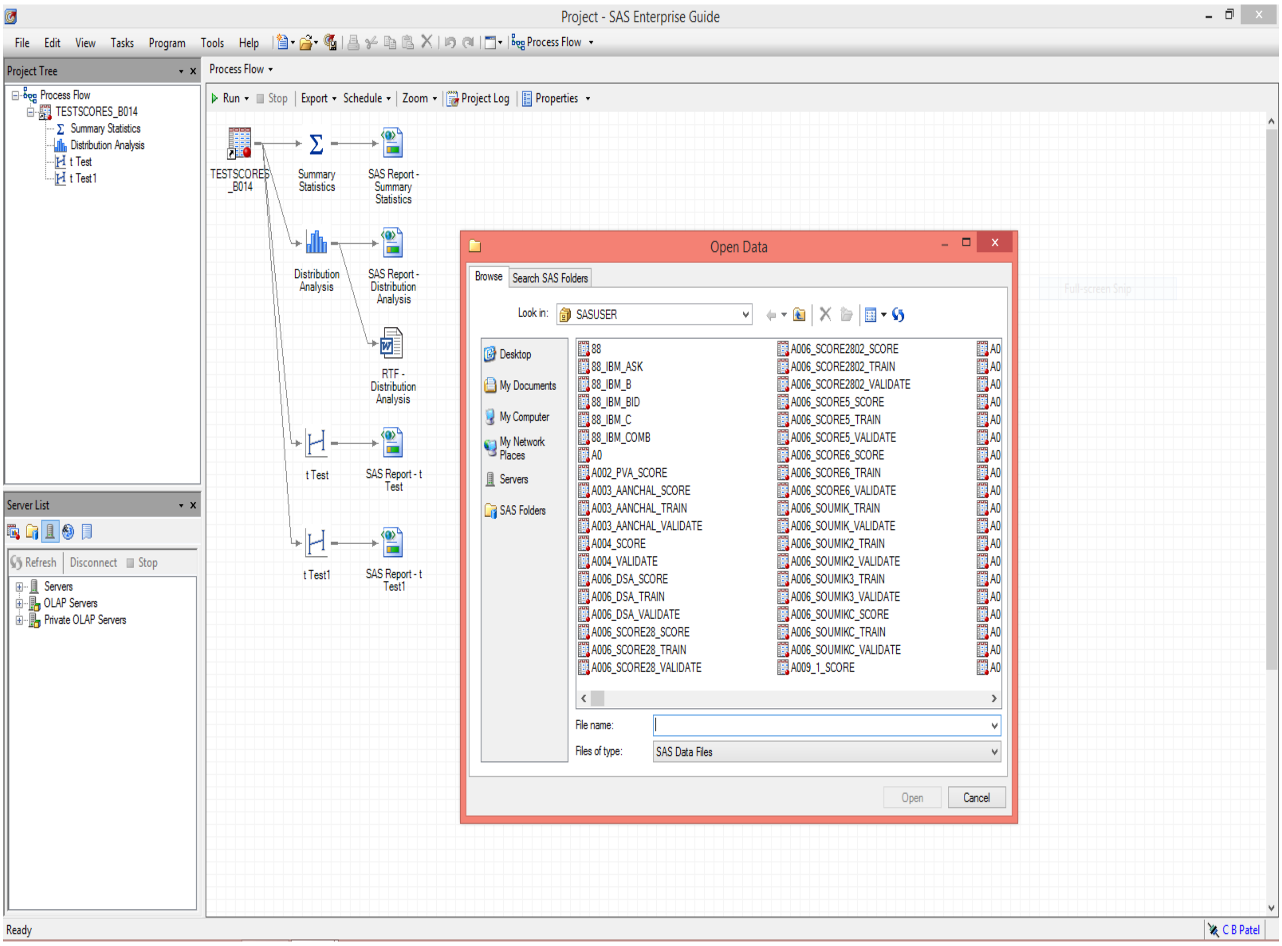
Aim: To implement:

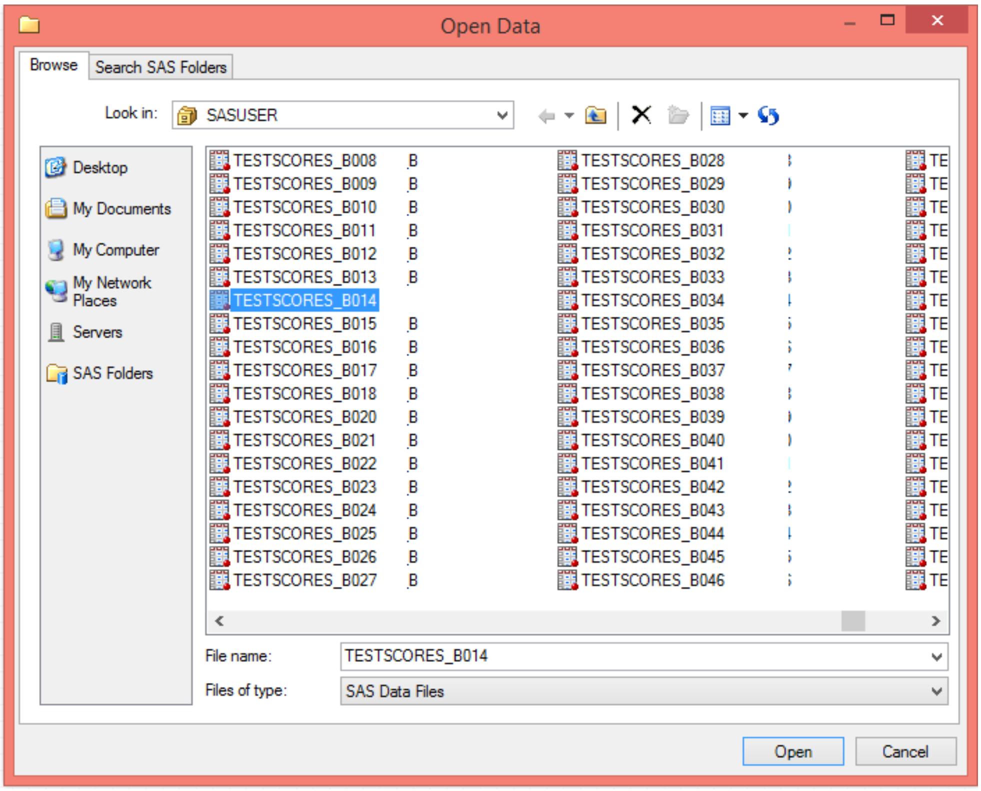
1. Filter and Sort
2. Statistical Analysis
3. Distribution Analysis
4. Confidence Intervals
5. t-test – One sample and Two sample

Data-set used: SAT Test Score

Importing a data source:

File -> Open -> Data -> Servers -> Local -> SASUSER library -> TESTSCORES_B014





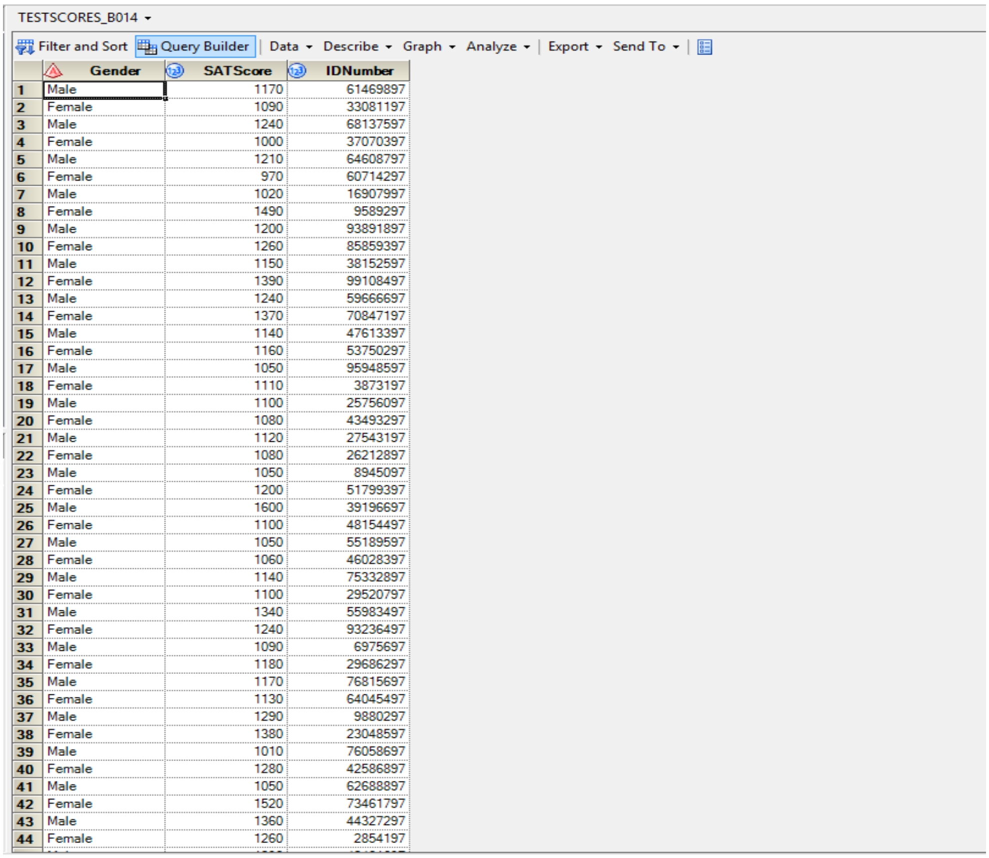
Look in: SASUSER

TESTSCORES_B008
TESTSCORES_B009
TESTSCORES_B010
TESTSCORES_B011
TESTSCORES_B012
TESTSCORES_B013
TESTSCORES_B014
TESTSCORES_B015
TESTSCORES_B016
TESTSCORES_B017
TESTSCORES_B018
TESTSCORES_B020
TESTSCORES_B021
TESTSCORES_B022
TESTSCORES_B023
TESTSCORES_B024
TESTSCORES_B025
TESTSCORES_B026
TESTSCORES_B027
TESTSCORES_B028
TESTSCORES_B029
TESTSCORES_B030
TESTSCORES_B031
TESTSCORES_B032
TESTSCORES_B033
TESTSCORES_B034
TESTSCORES_B035
TESTSCORES_B036
TESTSCORES_B037
TESTSCORES_B038
TESTSCORES_B039
TESTSCORES_B040
TESTSCORES_B041
TESTSCORES_B042
TESTSCORES_B043
TESTSCORES_B044
TESTSCORES_B045
TESTSCORES_B046

File name: TESTSCORES_B014

Files of type: SAS Data Files

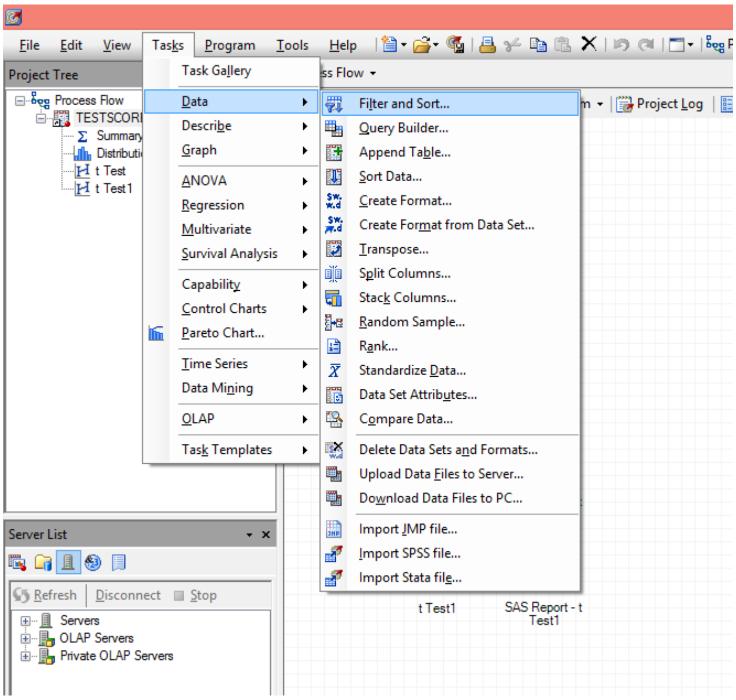
Double click on the data set in the process flow window to view its contents.



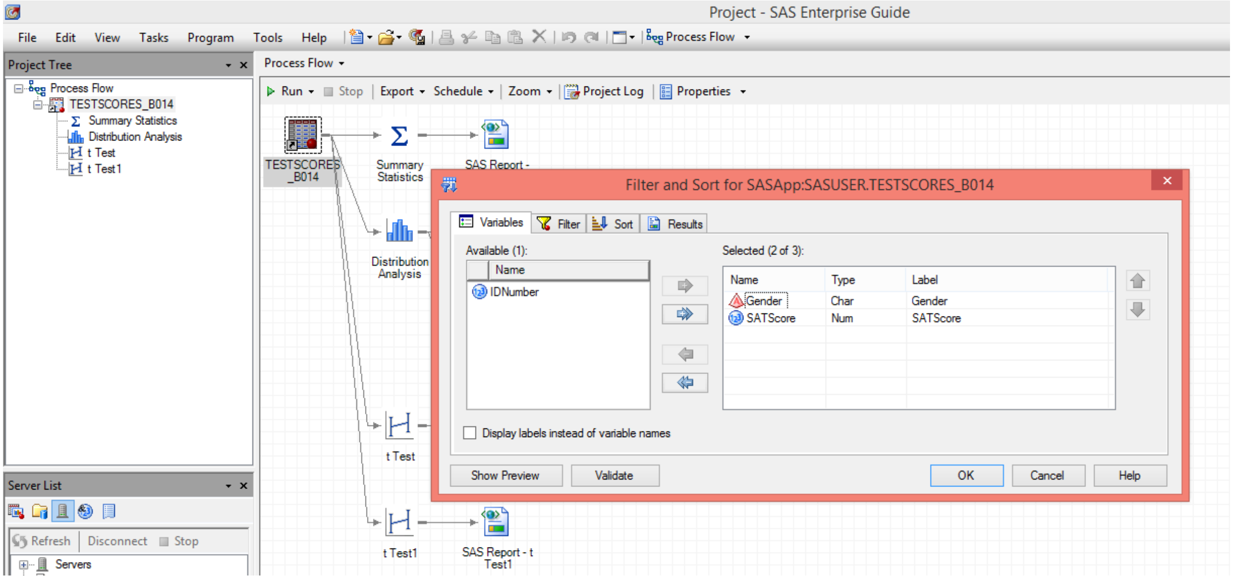
	Gender	SATScore	IDNumber
1	Male	1170	61469897
2	Female	1090	33081197
3	Male	1240	68137597
4	Female	1000	37070397
5	Male	1210	64608797
6	Female	970	60714297
7	Male	1020	16907997
8	Female	1490	9589297
9	Male	1200	93891897
10	Female	1260	85859397
11	Male	1150	38152597
12	Female	1390	99108497
13	Male	1240	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
40	Female	1280	42586897
41	Male	1050	62688897
42	Female	1520	73461797
43	Male	1360	44327297
44	Female	1260	2854197

Filter and Sort:

1. Select data-set on the process window
2. Click on ‘Tasks’ in the menu bar
    a. Select Data -> Filter and Sort (this opens a filter and sort wizard)

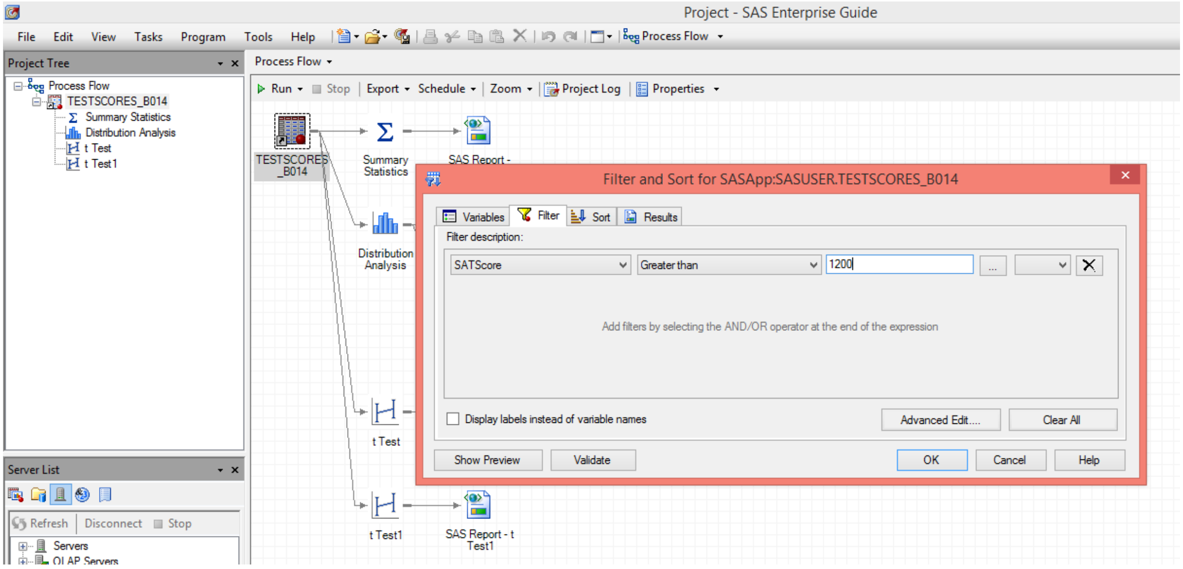


Select variables on the basis of which you want to filter the data-set



Name	Type	Label
Gender	Char	Gender
SATScore	Num	SATScore

Set filter conditions by selecting variable, filtering condition and value.



Filter and Sort for SASApp:SASUSER.TESTSCORES_B014

Variables | Filter | Sort | Results

Filter description:

SATScore | Greater than | 1200

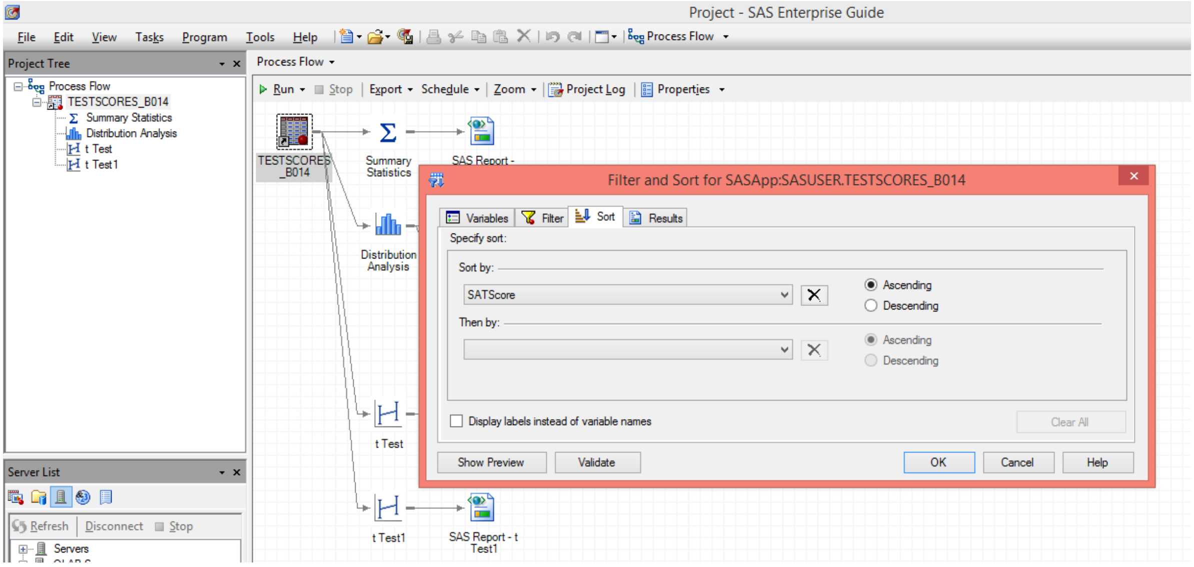
Add filters by selecting the AND/OR operator at the end of the expression

☐ Display labels instead of variable names

Advanced Edit.... | Clear All

Show Preview | Validate | OK | Cancel | Help

Select sort and apply sorting conditions



Filter and Sort for SASApp:SASUSER.TESTSCORES_B014

Variables | Filter | Sort | Results

Specify sort:

Sort by:

SATScore ● Ascending ○ Descending

Then by:

● Ascending ○ Descending

☐ Display labels instead of variable names

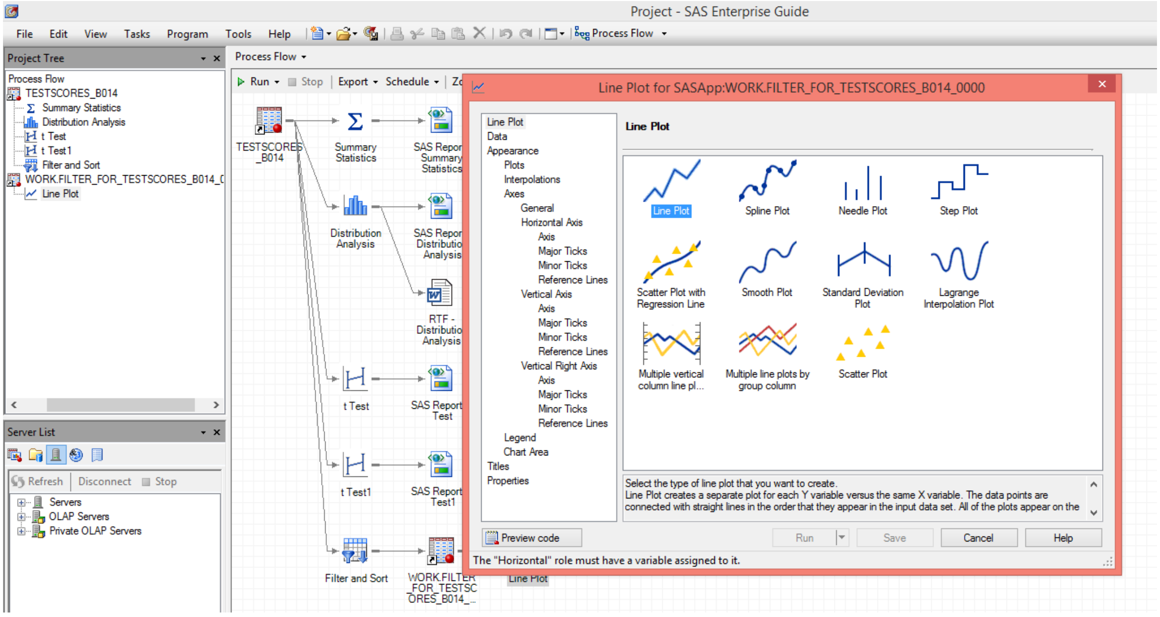
Clear All

Show Preview | Validate | OK | Cancel | Help

When you are done, click ‘OK’ to apply the filters and sorting conditions to the data-set.

You will observe a new data-set being created in the process window that is derived from the main data-set after applying filter and sort function. Double click on this new data-set to view its contents.

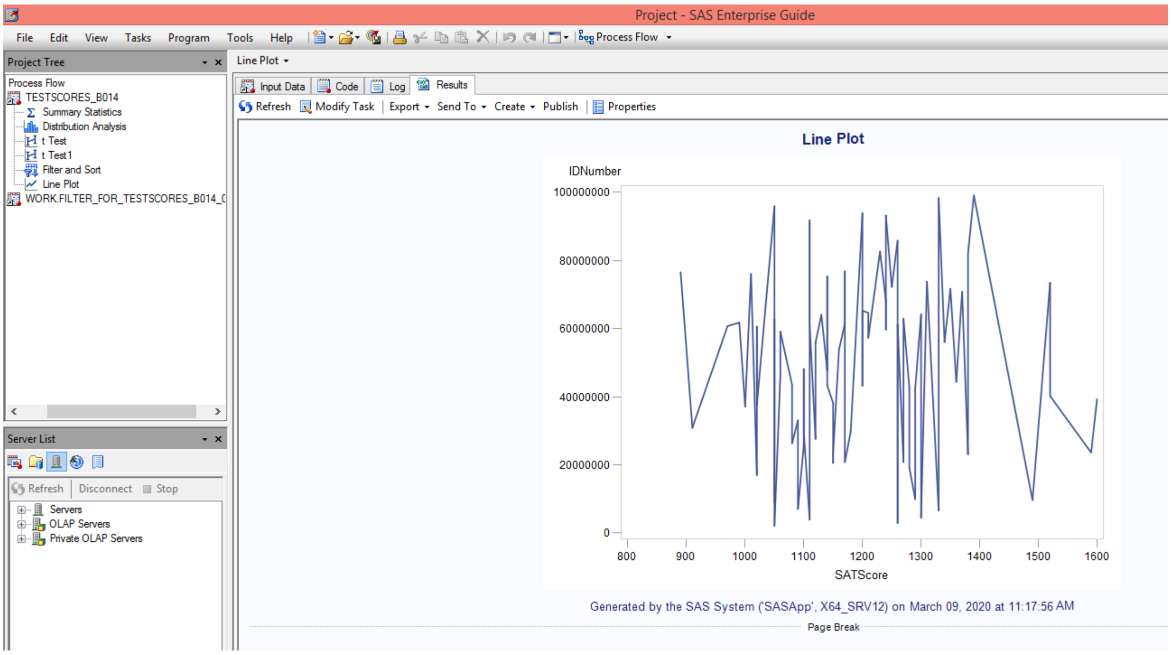
Line Plot:



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

The "Horizontal" role must have a variable assigned to it.

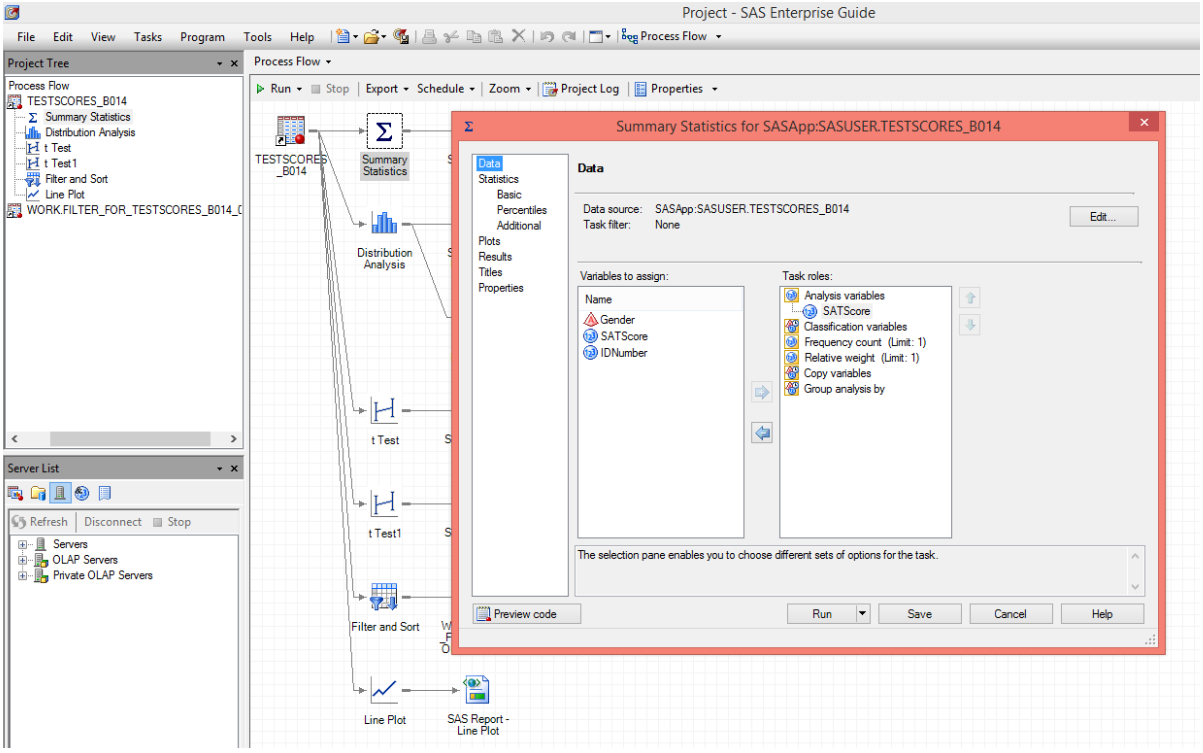
View line plot by double clicking the node on the process window.



Generated by the SAS System ('SASApp', X64_SRV12) on March 09, 2020 at 11:17:56 AM

Summary Statistics:

Select data-set -> Tasks –> Describe -> Summary statistics
This will open a summary statistics wizard.
Select the variables for analysis



Summary Statistics for SASApp:SASUSER.TESTSCORES_B014

Data

Data source: SASApp:SASUSER.TESTSCORES_B014
Task filter: None

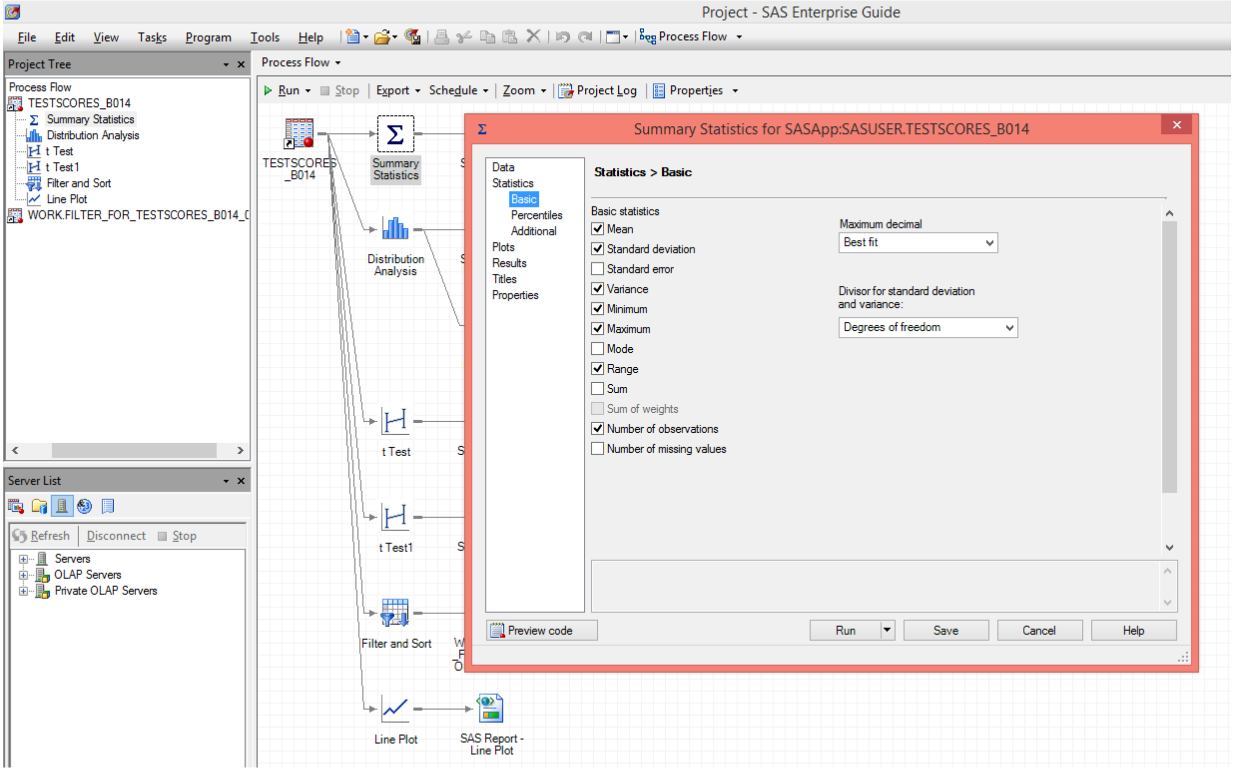
Variables to assign:

- Name
- Gender
- SATScore
- IDNumber

Task roles:

- Analysis variables
  - SATScore
- Classification variables
- Frequency count (Limit: 1)
- Relative weight (Limit: 1)
- Copy variables
- Group analysis by

The selection pane enables you to choose different sets of options for the task.



Summary Statistics for SASApp:SASUSER.TESTSCORES_B014

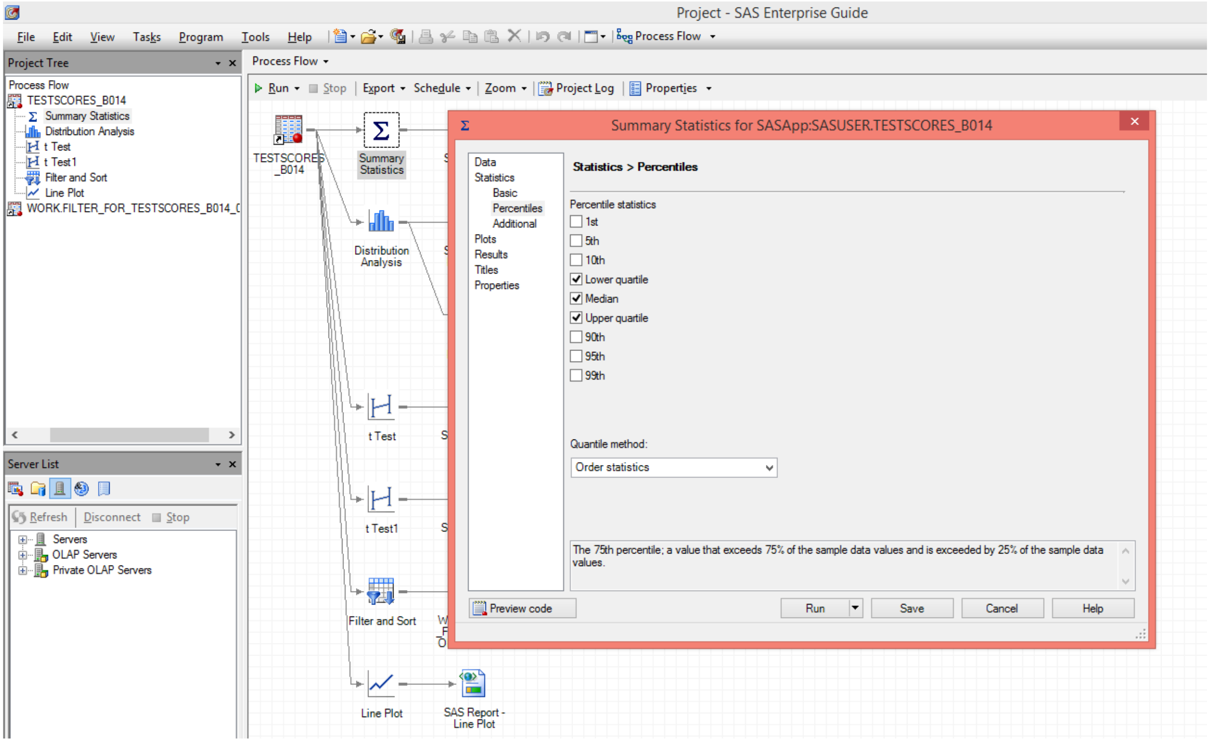
Statistics > Basic

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



Summary Statistics for SASApp:SASUSER.TESTSCORES_B014

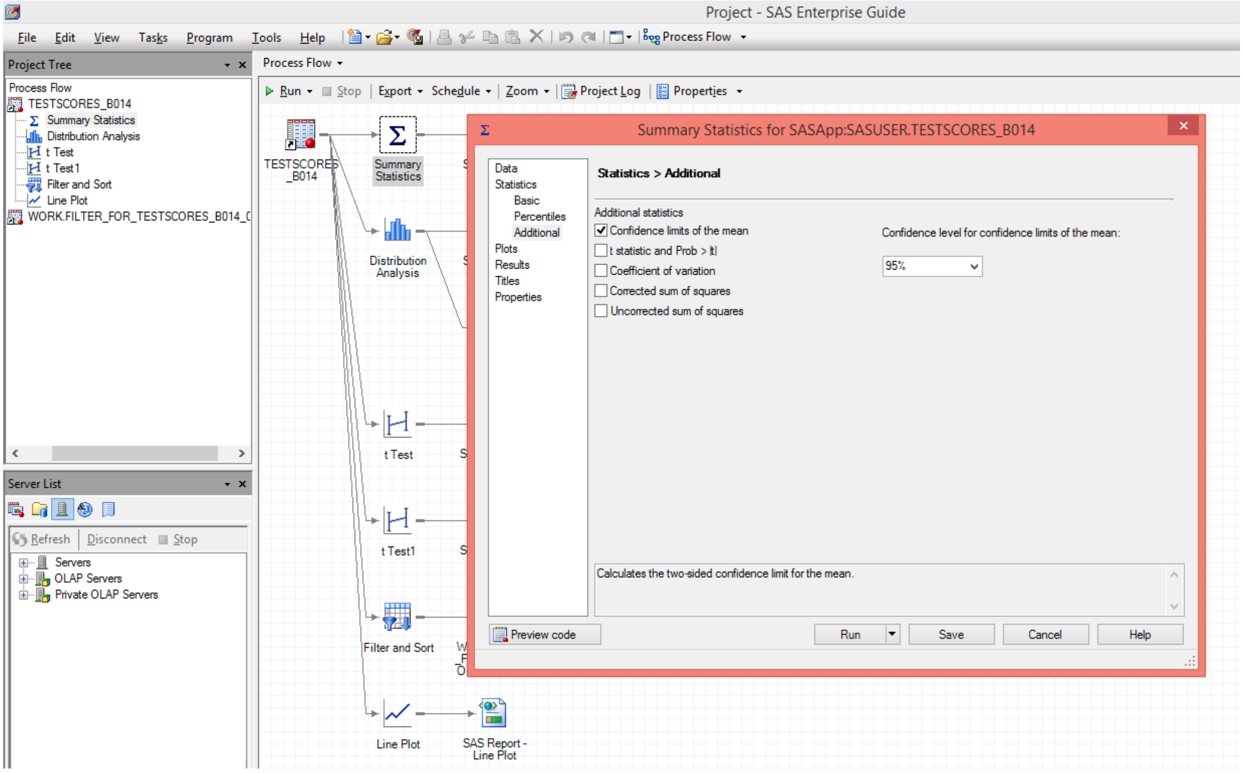
Statistics > Percentiles

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.



Summary Statistics for SASApp:SASUSER.TESTSCORES_B014

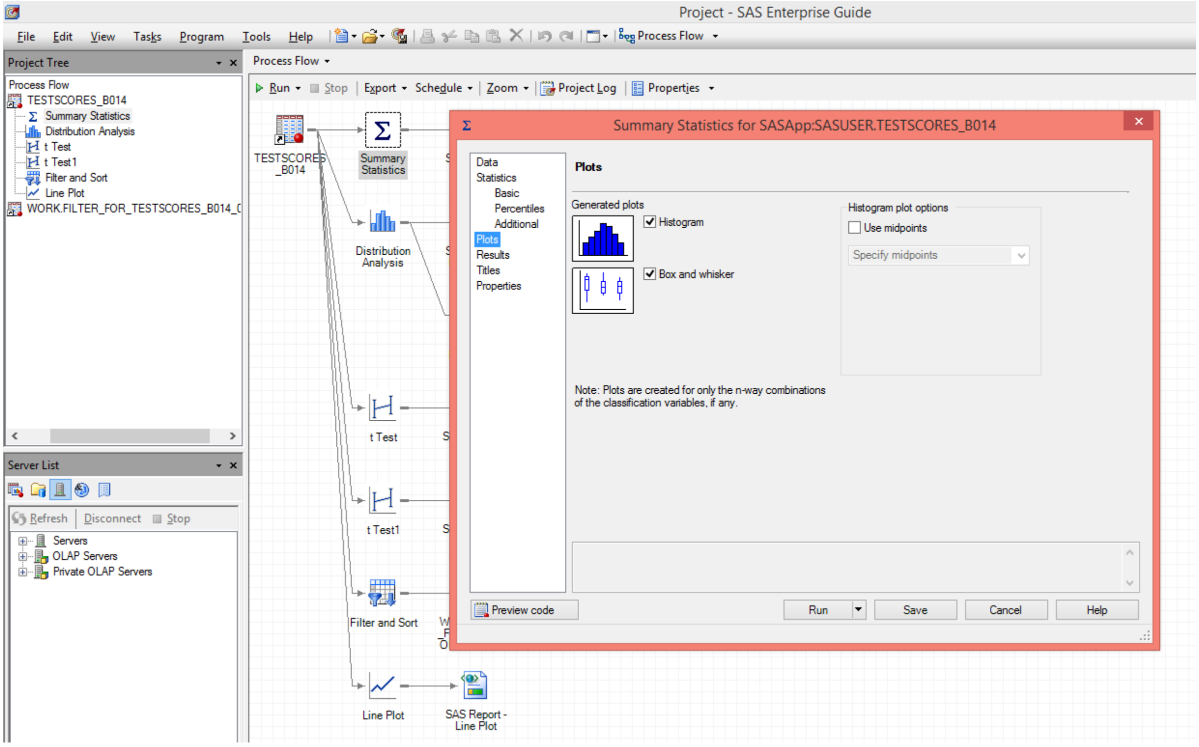
Statistics > Additional

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.



Summary Statistics for SASApp:SASUSER.TESTSCORES_B014

Plots

Generated plots

☑ Histogram

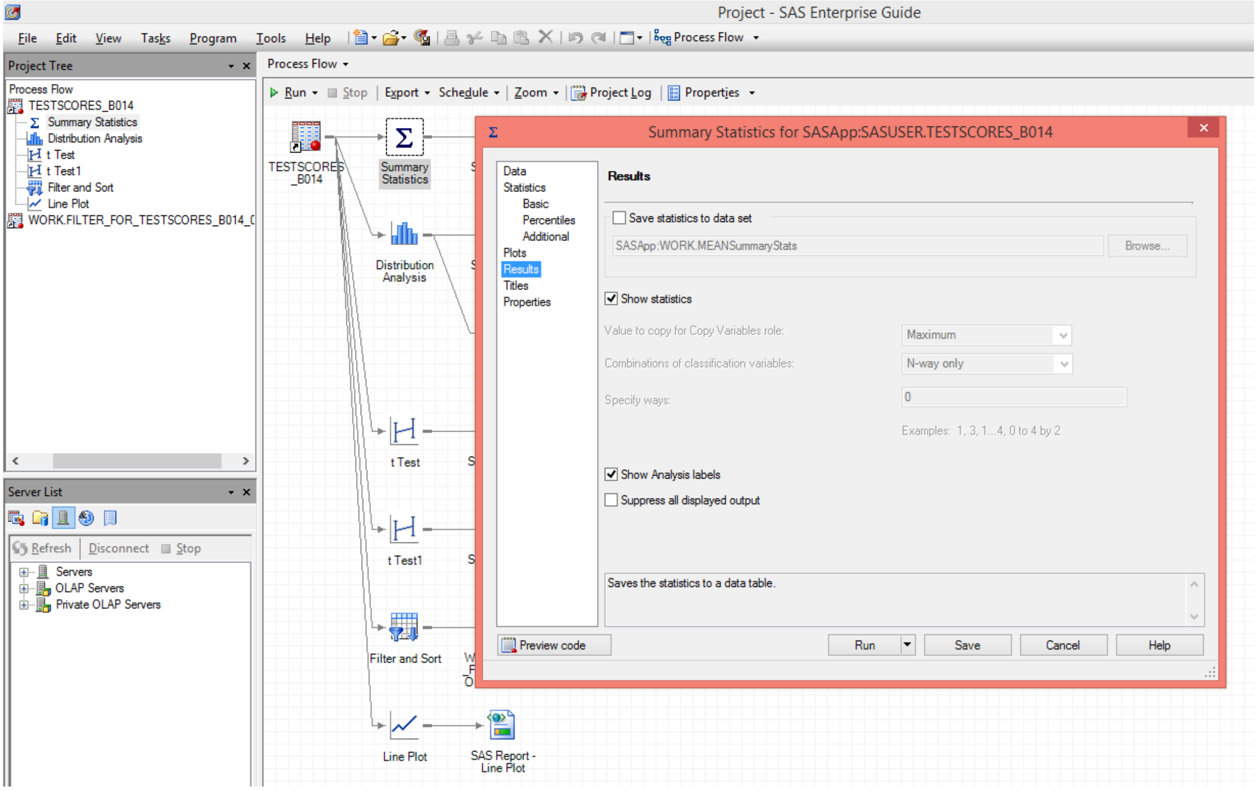
☑ Box and whisker

Histogram plot options

□ Use midpoints

Specify midpoints

Note: Plots are created for only the n-way combinations of the classification variables, if any.



Summary Statistics for SASApp:SASUSER.TESTSCORES_B014

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

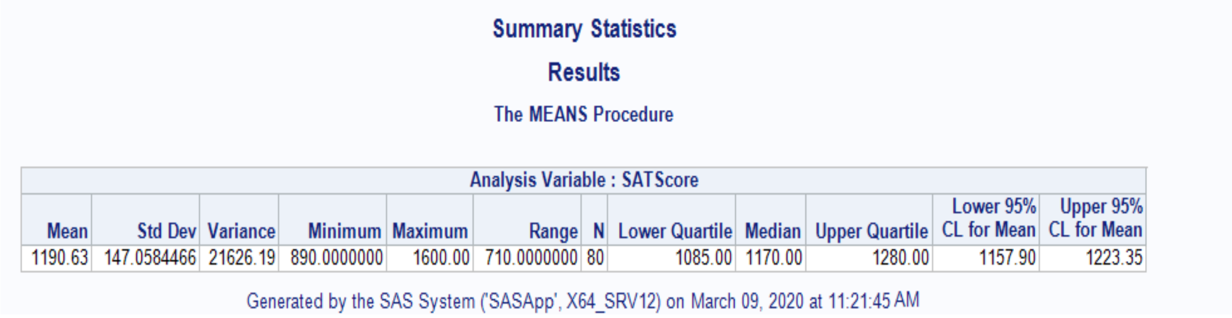
Examples: 1, 3, 1...4, 0 to 4 by 2

☑ Show Analysis labels

□ Suppress all displayed output

Saves the statistics to a data table.

Summary statistics result shows the selected values, ie mean, standard deviation, variance, min,mac, range, quartiles and confidence levels for the selected data-set.



Summary Statistics

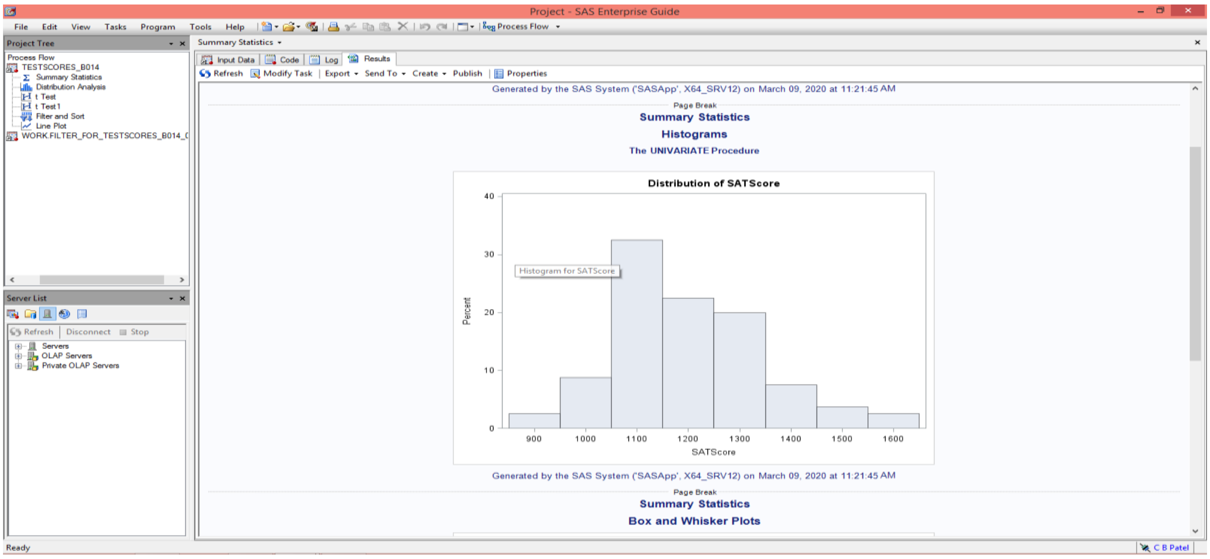
Results

The MEANS Procedure

Analysis Variable : SATScore											
Mean	Std Dev	Variance	Minimum	Maximum	Range	N	Lower Quartile	Median	Upper Quartile	Lower 95% CL for Mean	Upper 95% CL for Mean
1190.63	147.0584466	21626.19	890.0000000	1600.00	710.0000000	80	1085.00	1170.00	1280.00	1157.90	1223.35

Generated by the SAS System ('SASApp', X64_SRV12) on March 09, 2020 at 11:21:45 AM

The histogram plot shows the percentage distribution of SAT scores.



Generated by the SAS System ('SASApp', X64_SRV12) on March 09, 2020 at 11:21:45 AM

Summary Statistics

Histograms

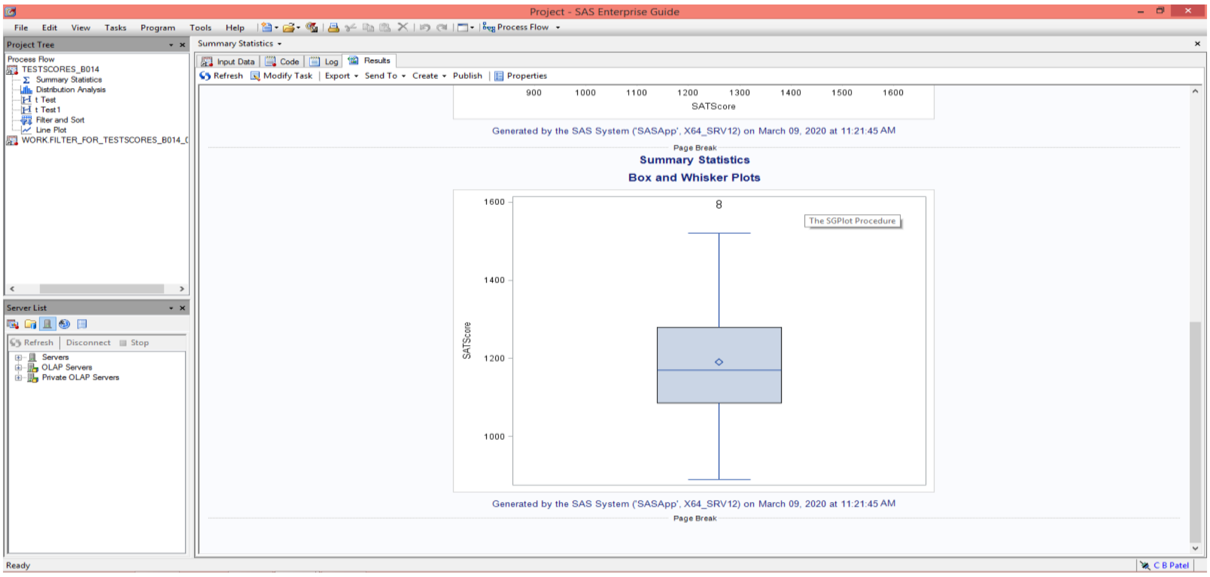
The UNIVARIATE Procedure

Generated by the SAS System ('SASApp', X64_SRV12) on March 09, 2020 at 11:21:45 AM

Summary Statistics

Box and Whisker Plots

Box plot indicates the minimum, 1st quartile (25th percentile), median, 3rd quartile (75th percentile) and maximum values for the data-set.



Generated by the SAS System ('SASApp', X64_SRV12) on March 09, 2020 at 11:21:45 AM

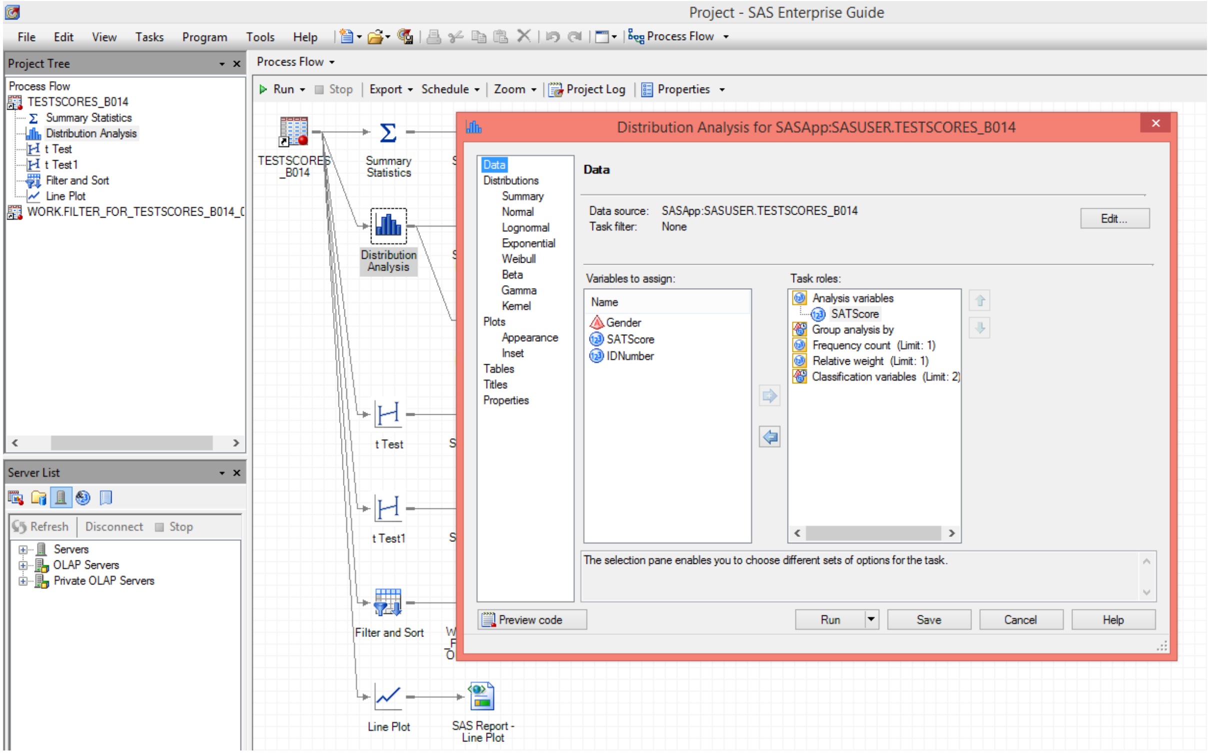
Summary Statistics

Box and Whisker Plots

Generated by the SAS System ('SASApp', X64_SRV12) on March 09, 2020 at 11:21:45 AM

Distribution Analysis:

Select data-set -> Tasks –> Describe -> Distribution Analysis

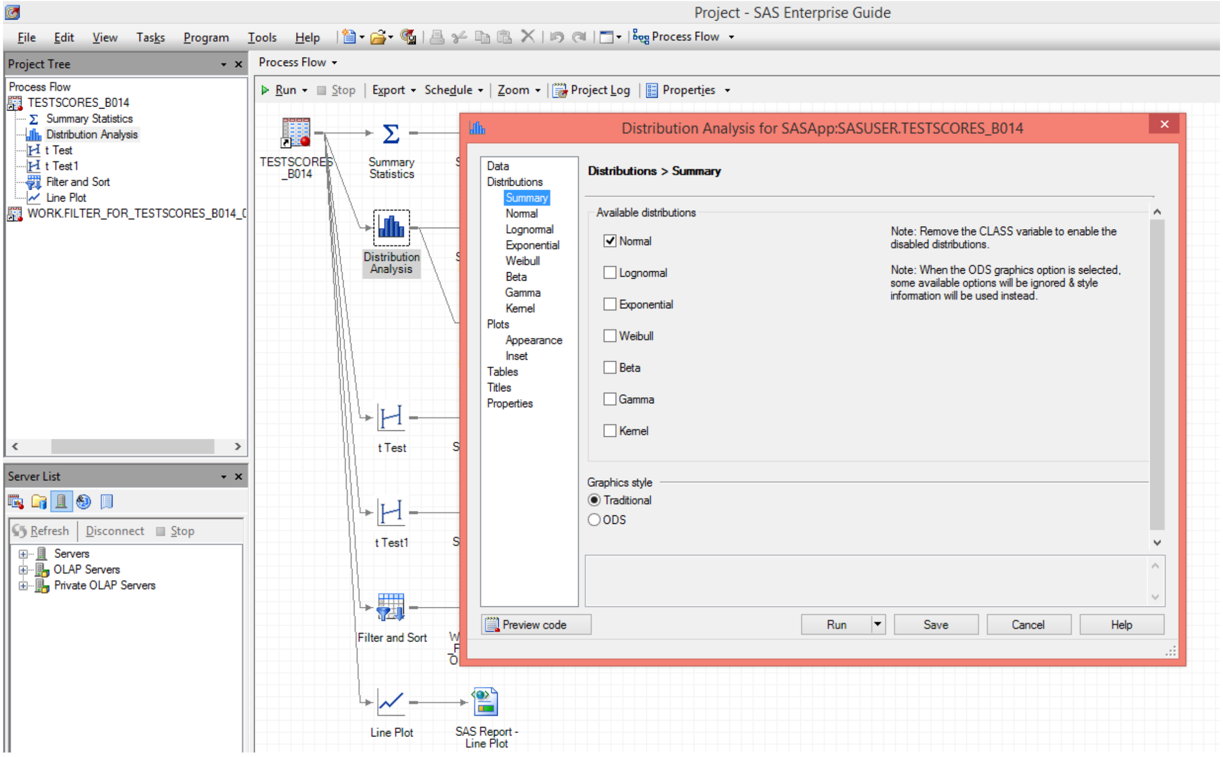


Data

Data source: SASApp:SASUSER.TESTSCORES_B014
Task filter: None

Task roles:
- Analysis variables
  - SATScore
- Group analysis by
- Frequency count (Limit: 1)
- Relative weight (Limit: 1)
- Classification variables (Limit: 2)

The selection pane enables you to choose different sets of options for the task.



Distributions > Summary

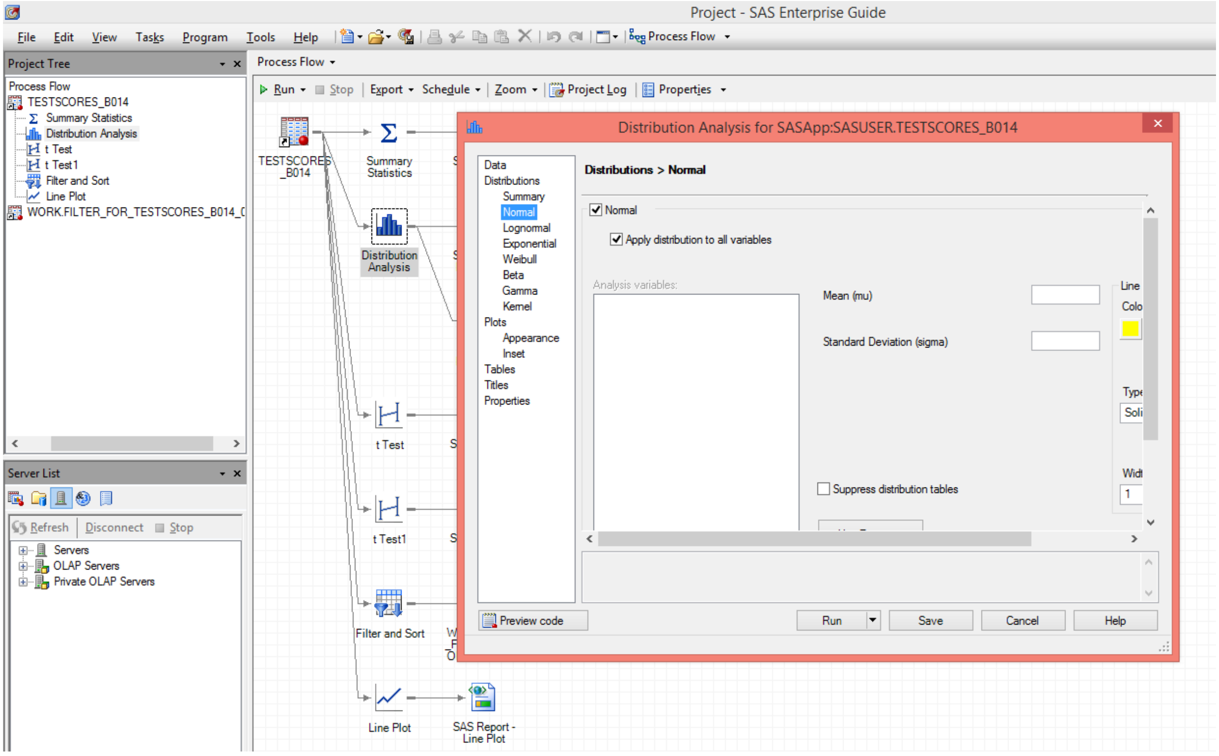
Available distributions

☑ 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



Distributions > Normal

☑ Normal

☑ Apply distribution to all variables

Analysis variables:

Mean (mu)

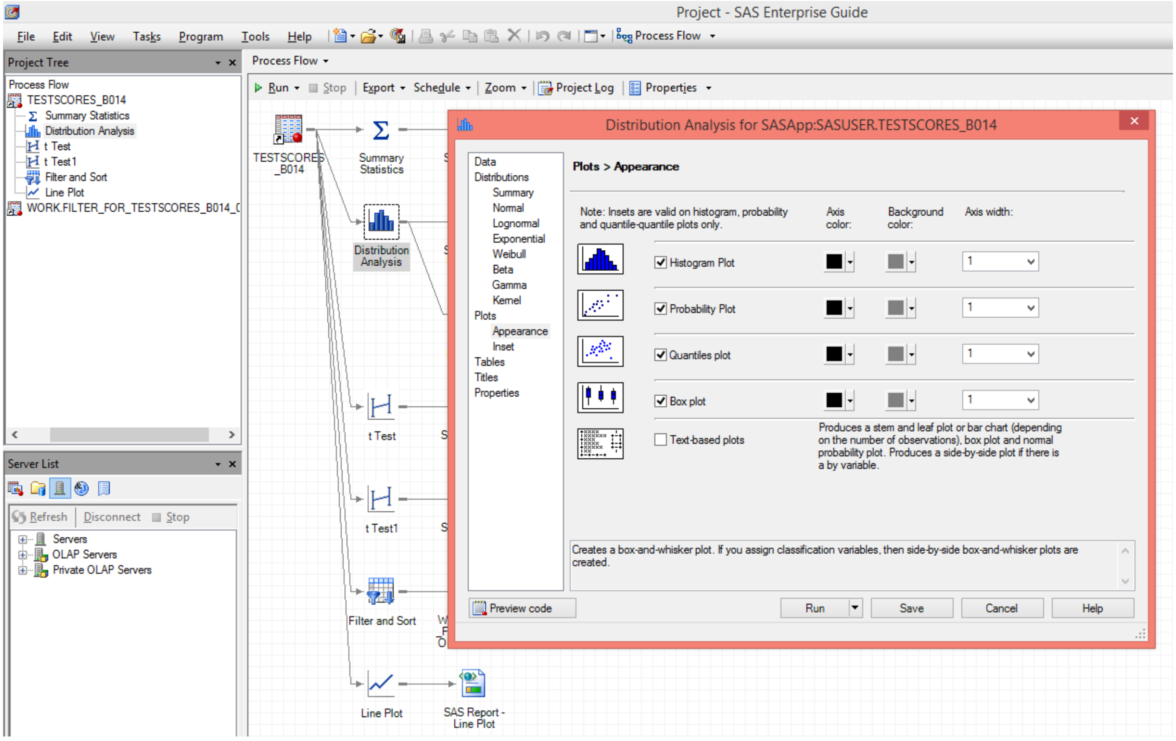
Standard Deviation (sigma)

Line Colo

Type Soli

Widt 1

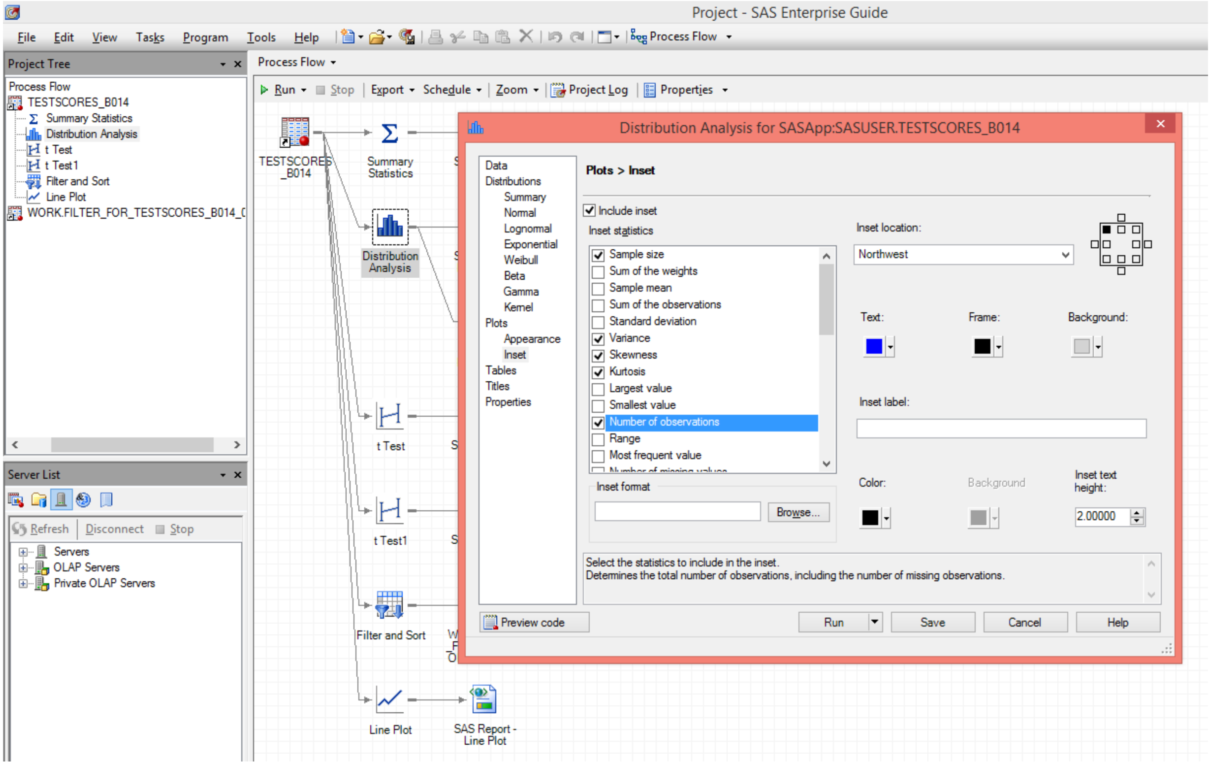
□ Suppress distribution tables



Plots > Appearance

Note: Insets are valid on histogram, probability and quantile-quantile plots only.	Axis color:	Background color:	Axis width:
☑ Histogram Plot	■	■	1
☑ Probability Plot	■	■	1
☑ Quantiles plot	■	■	1
☑ Box plot	■	■	1
□ 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.



Distribution Analysis for SASApp:SASUSER.TESTSCORES_B014

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 location: Northwest

Text: Frame: Background:

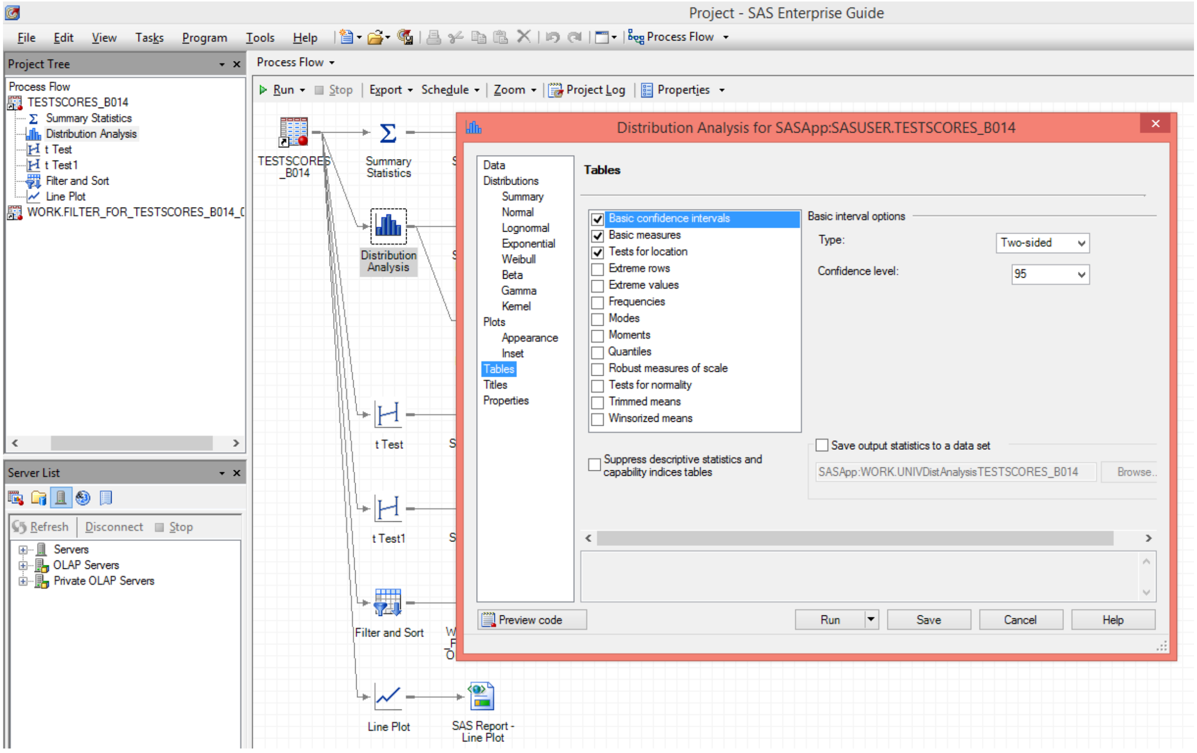
Inset label:

Inset format

Browse...

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.



Distribution Analysis for SASApp:SASUSER.TESTSCORES_B014

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

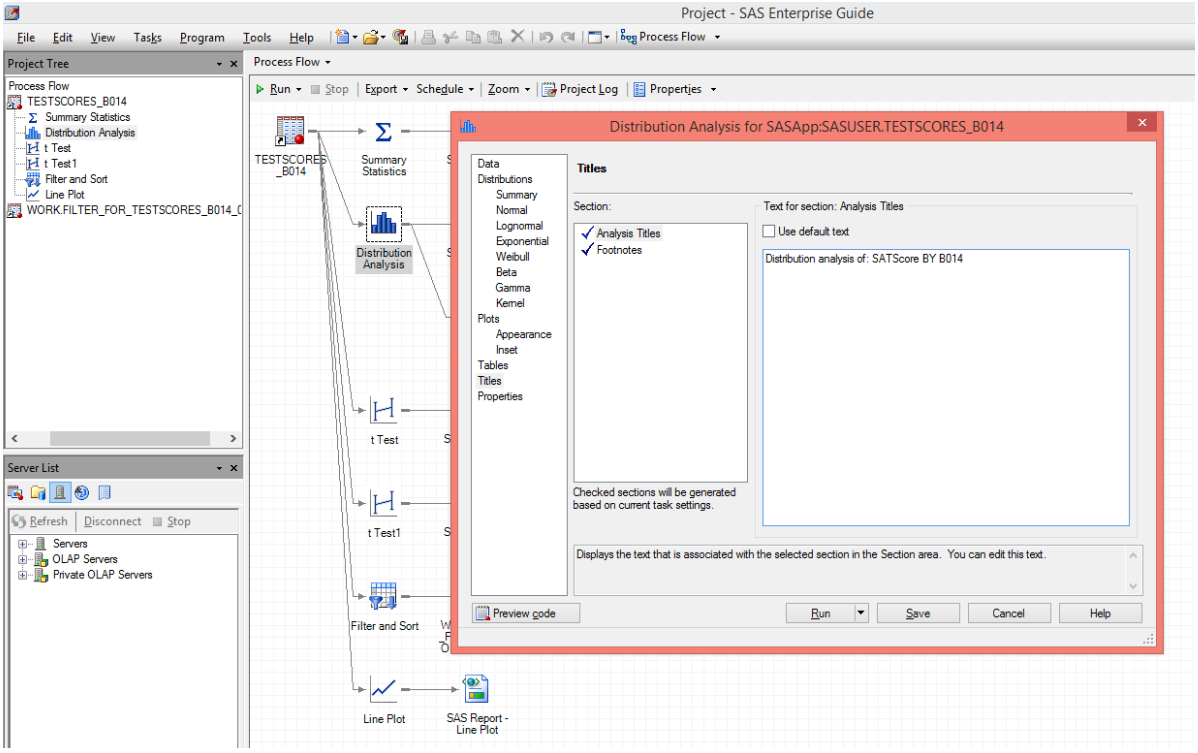
Type: Two-sided

Confidence level: 95

☐ Save output statistics to a data set

SASApp:WORK.UNIVDistAnalysisTESTSCORES_B014 Browse...

☐ Suppress descriptive statistics and capability indices tables



Distribution Analysis for SASApp:SASUSER.TESTSCORES_B014

Titles

Section:

✓ Analysis Titles

✓ Footnotes

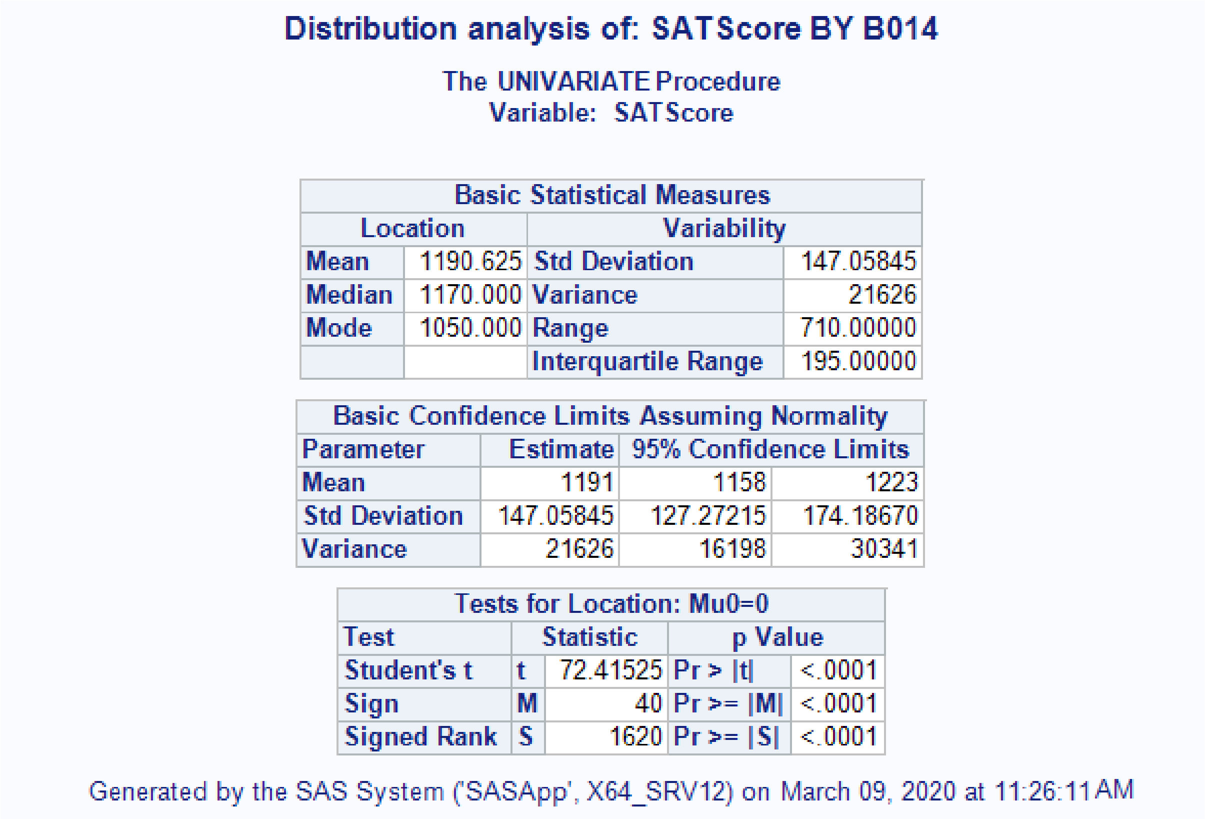
Checked sections will be generated based on current task settings.

Text for section: Analysis Titles

□ Use default text

Distribution analysis of: SATScore BY B014

Displays the text that is associated with the selected section in the Section area. You can edit this text.



Distribution analysis of: SATScore BY B014

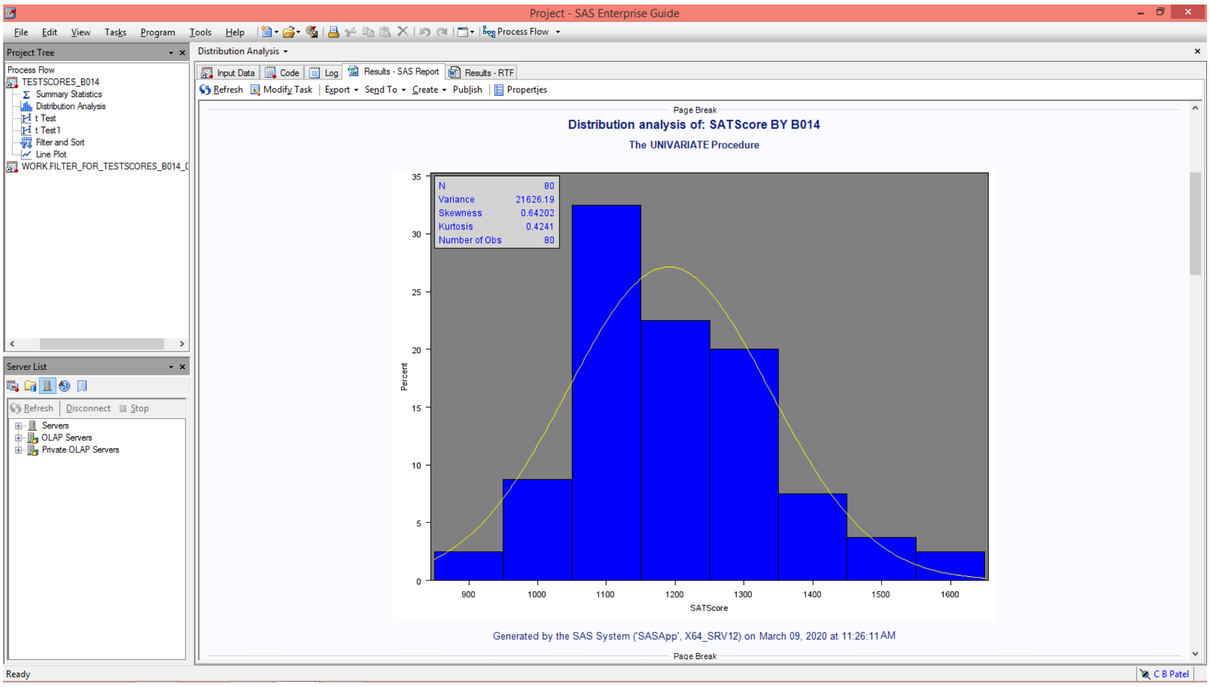
The UNIVARIATE Procedure
Variable: SATScore

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

Generated by the SAS System ('SASApp', X64_SRV12) on March 09, 2020 at 11:26:11 AM

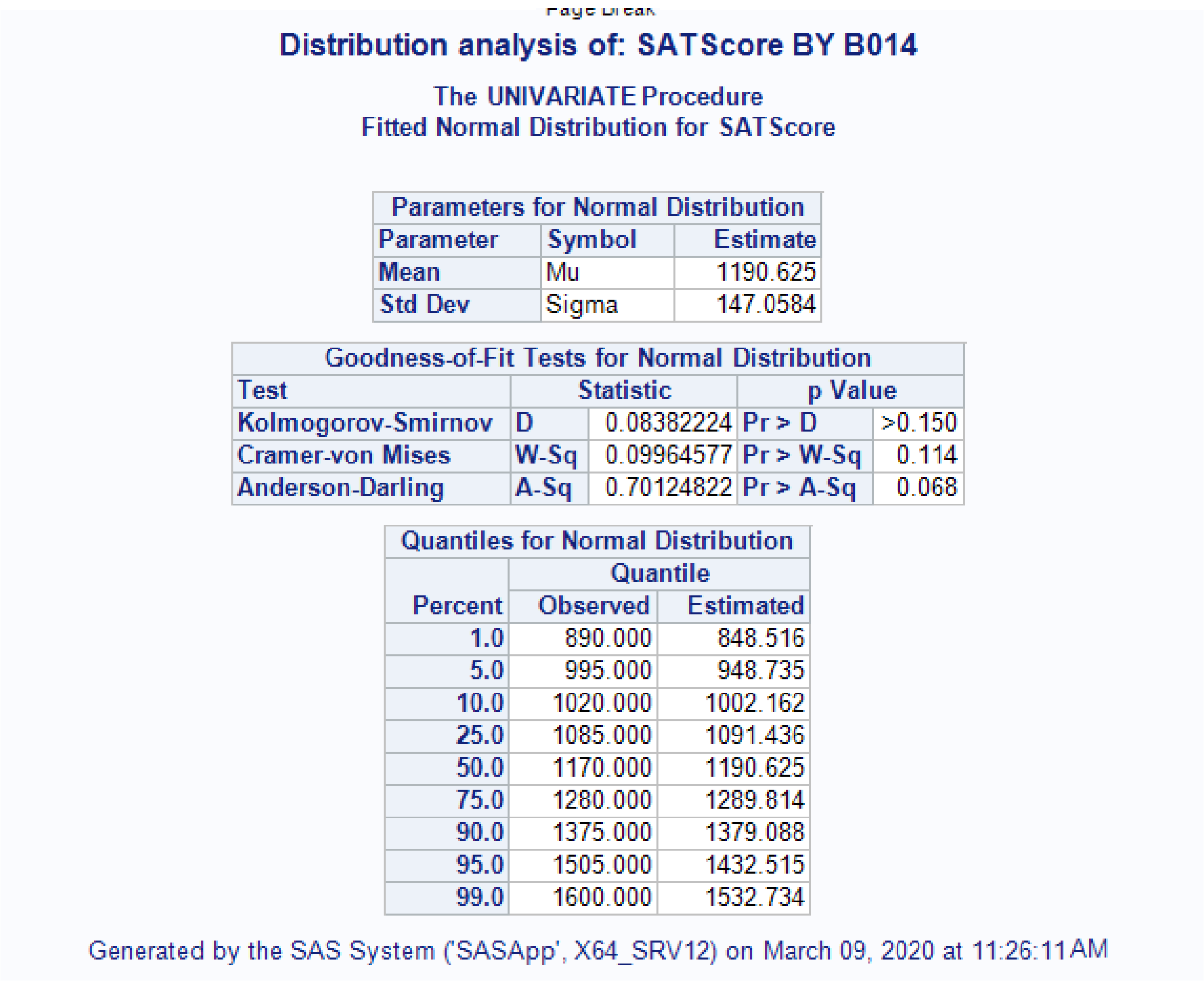


Distribution analysis of: SATScore BY B014

The UNIVARIATE Procedure

N	80
Variance	21626.19
Skewness	0.64202
Kurtosis	0.4241
Number of Obs	80

Generated by the SAS System ('SASApp', X64_SRV12) on March 09, 2020 at 11:26:11 AM



Page Break

Distribution analysis of: SATScore BY B014

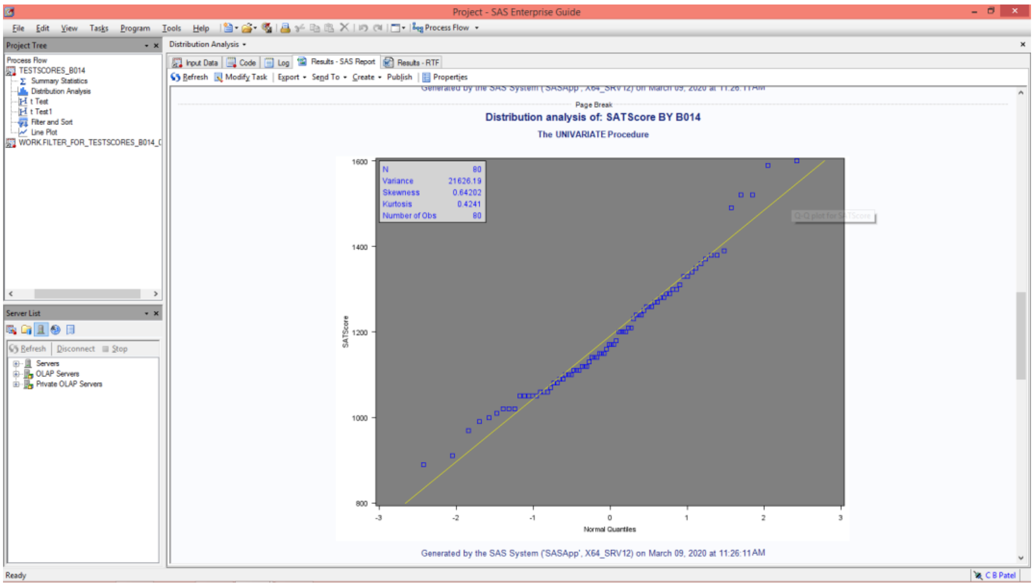
The UNIVARIATE Procedure
Fitted Normal Distribution for SATScore

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

Generated by the SAS System ('SASApp', X64_SRV12) on March 09, 2020 at 11:26:11 AM



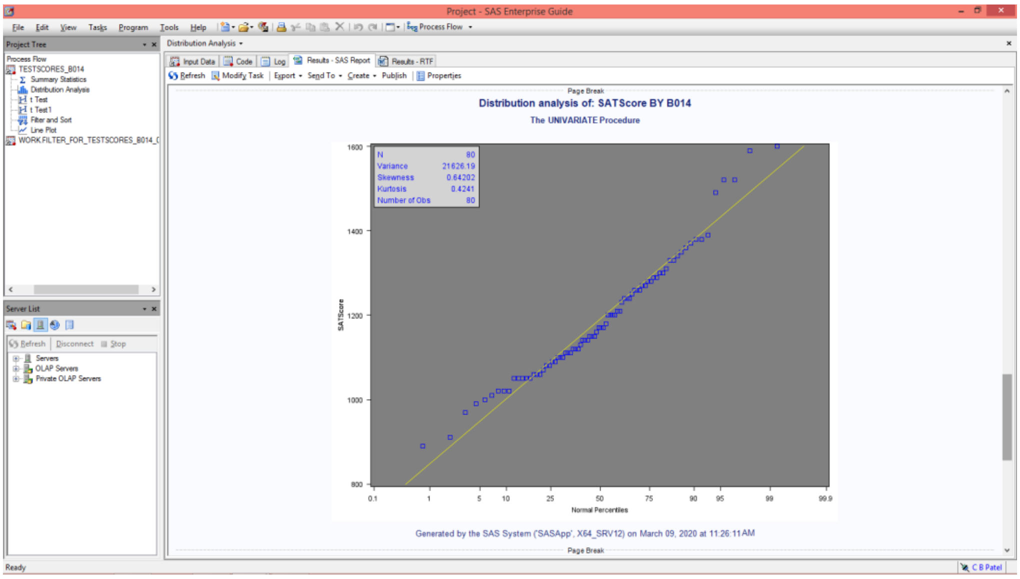
Generated by the SAS System ('SASApp', X64_SRV12) on March 09, 2020 at 11:26:11 AM

Distribution analysis of: SATScore BY B014

The UNIVARIATE Procedure

N	80
Variance	21626.19
Skewness	0.64202
Kurtosis	0.4241
Number of Obs	80

Generated by the SAS System ('SASApp', X64_SRV12) on March 09, 2020 at 11:26:11 AM

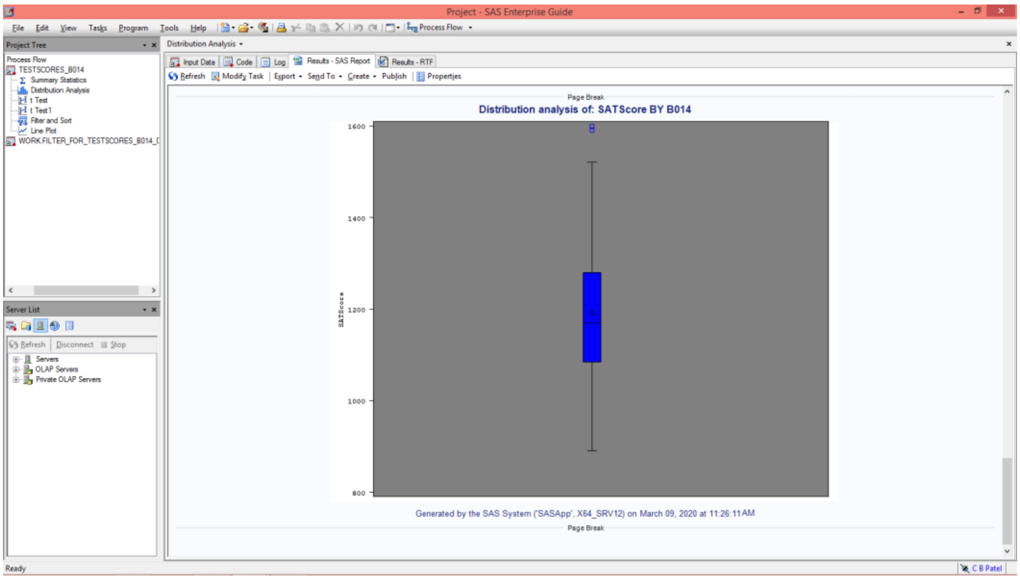


Distribution analysis of: SATScore BY B014

The UNIVARIATE Procedure

N	80
Variance	21626.19
Skewness	0.64202
Kurtosis	0.4241
Number of Obs	80

Generated by the SAS System ('SASApp', X64_SRV12) on March 09, 2020 at 11:26:11 AM

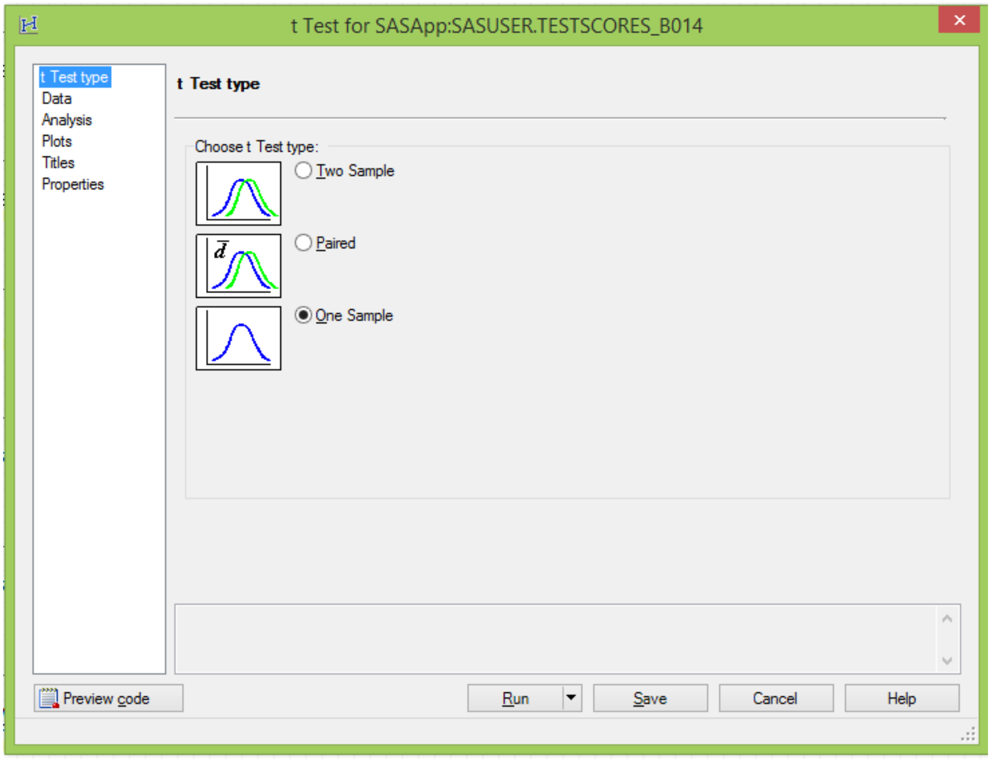


Distribution analysis of: SATScore BY B014

Generated by the SAS System ('SASApp', X64_SRV12) on March 09, 2020 at 11:26:11 AM

T-Test:

One Sample



t Test for SASApp:SASUSER.TESTSCORES_B014

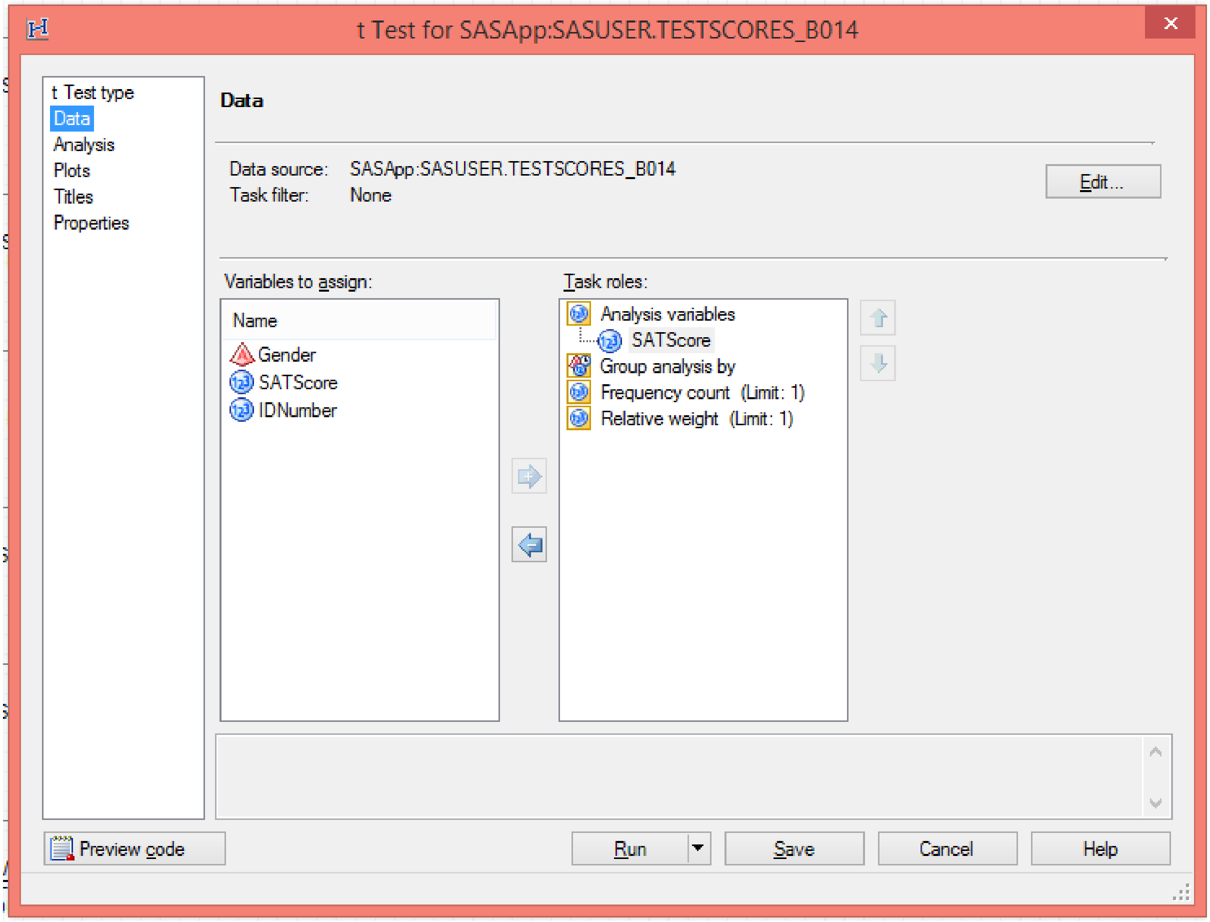
t Test type

Choose t Test type:

○ Two Sample

○ Paired

● One Sample



t Test for SASApp:SASUSER.TESTSCORES_B014

Data

Data source: SASApp:SASUSER.TESTSCORES_B014

Task filter: None

Variables to assign:

Name

Gender

SATScore

IDNumber

Task roles:

Analysis variables

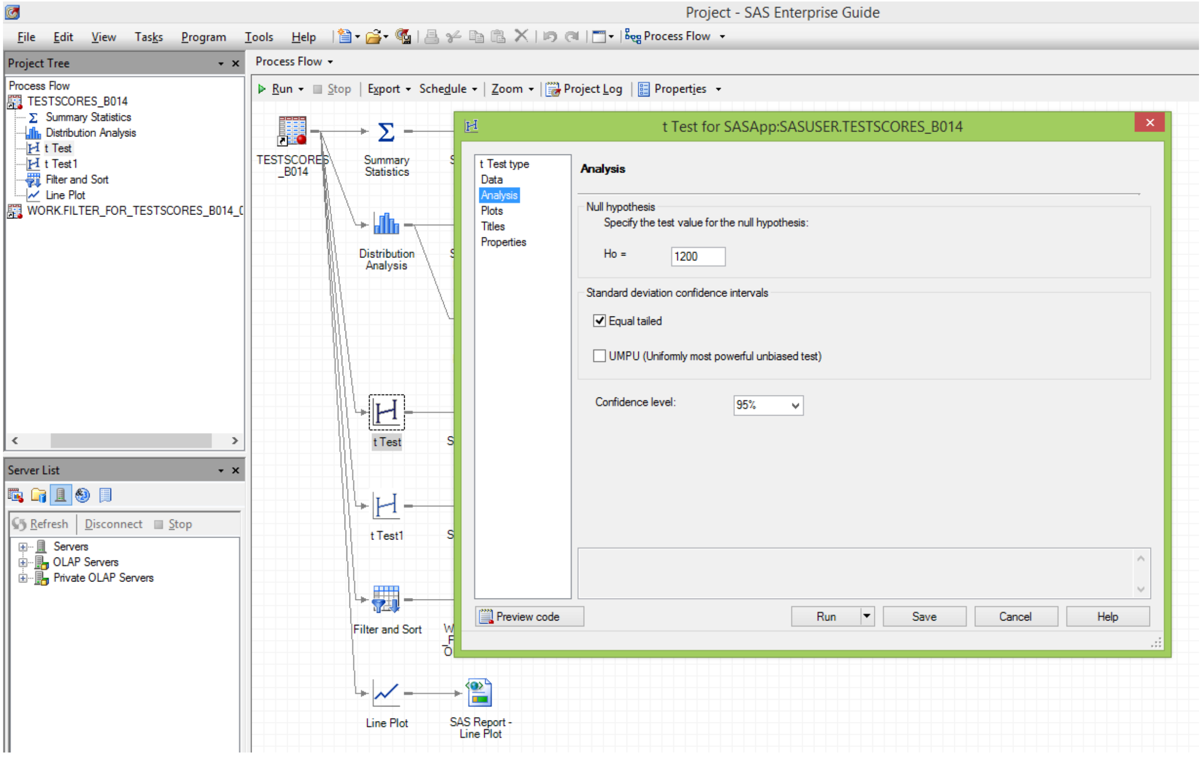
SATScore

Group analysis by

Frequency count (Limit: 1)

Relative weight (Limit: 1)

Enter the null hypothesis condition (value of Ho) and specify confidence level.



Analysis

Null hypothesis

Specify the test value for the null hypothesis:

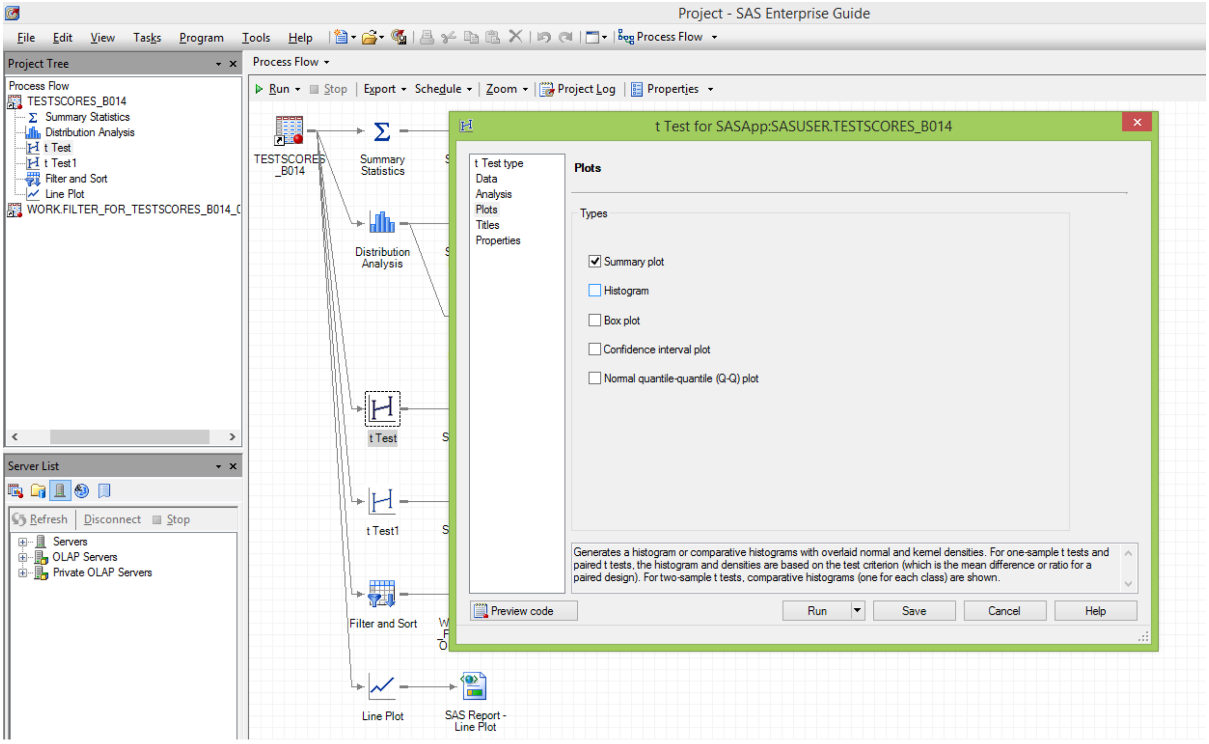
Ho = 1200

Standard deviation confidence intervals

☑ Equal tailed

☐ UMPU (Uniformly most powerful unbiased test)

Confidence level: 95%



Plots

Types

☑ Summary plot

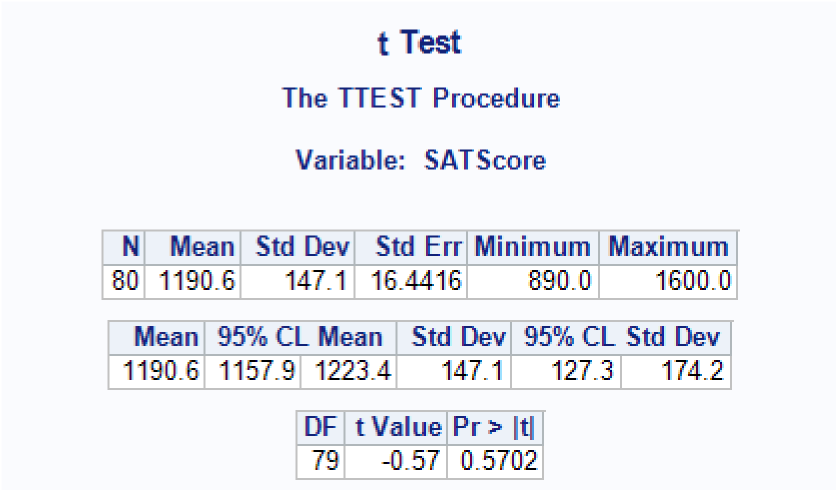
☐ Histogram

☐ Box plot

☐ Confidence interval plot

☐ Normal quantile-quantile (Q-Q) plot

Generates a histogram or comparative histograms with overlaid normal and kernel densities. For one-sample t tests and paired t tests, the histogram and densities are based on the test criterion (which is the mean difference or ratio for a paired design). For two-sample t tests, comparative histograms (one for each class) are shown.



t Test

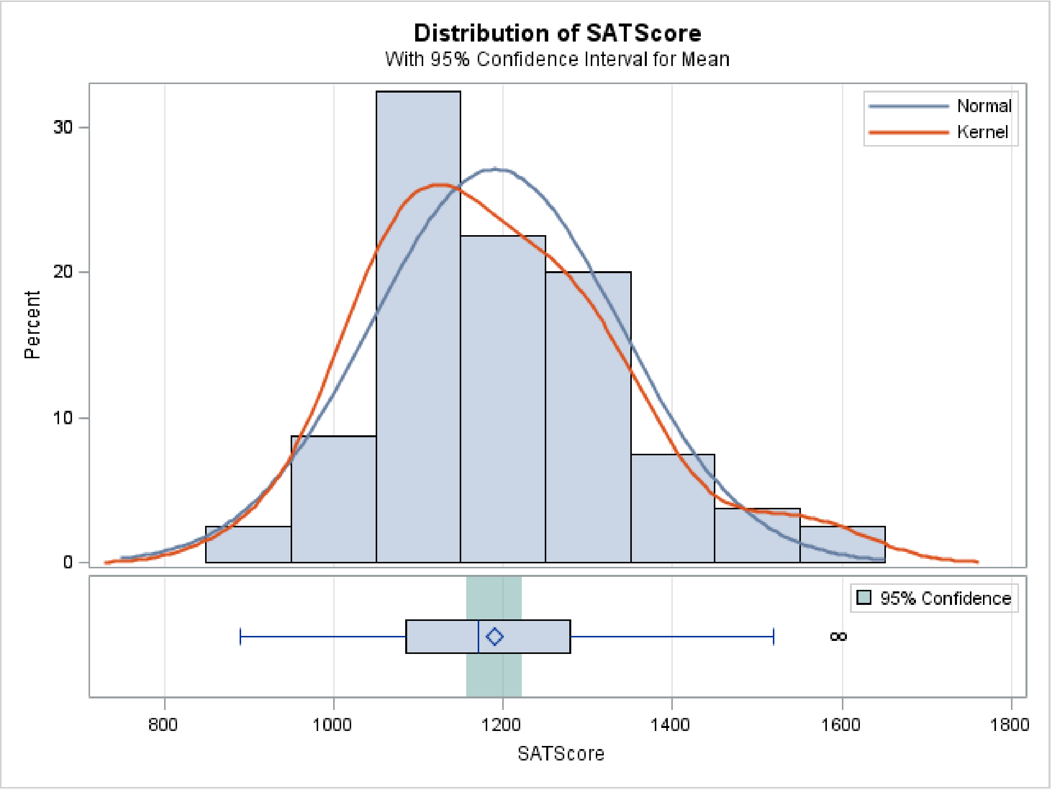
The TTEST Procedure

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

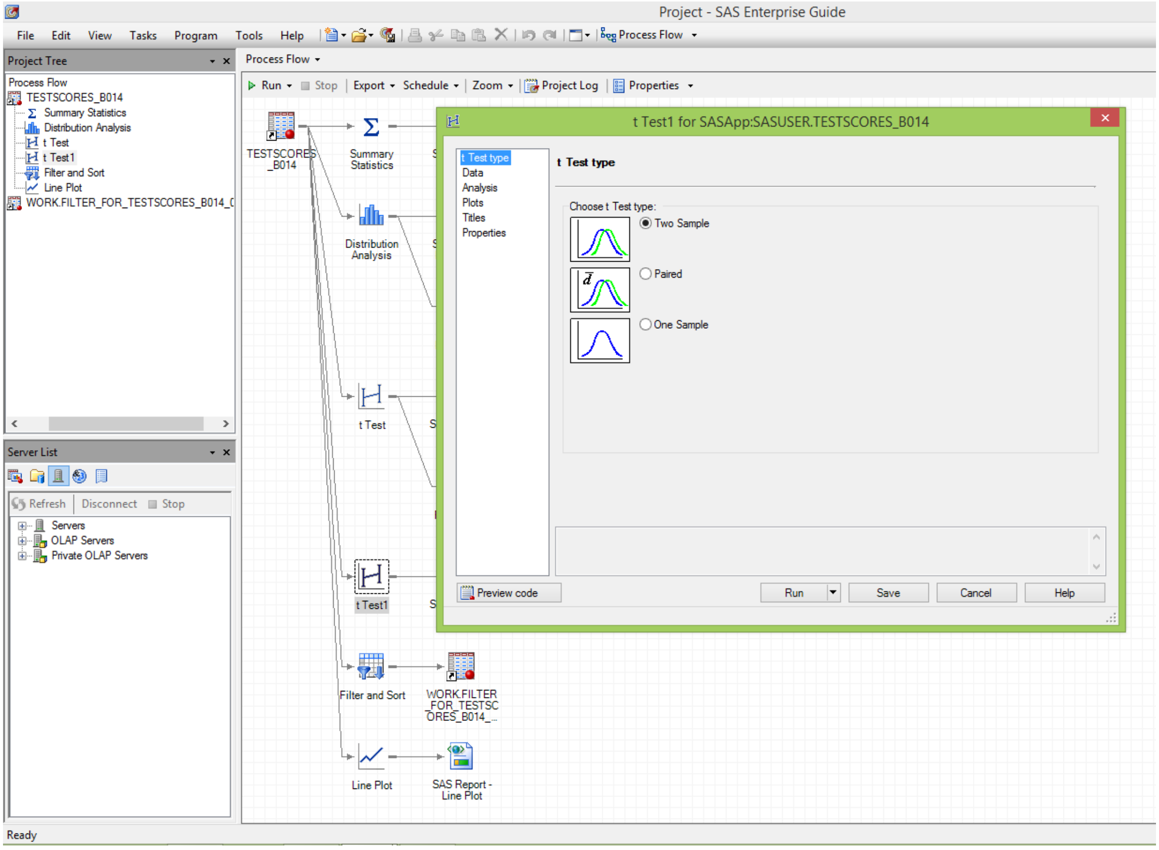


Generated by the SAS System ('SASApp', X64_SRV12) on March 09, 2020 at 11:31:25 AM

Page Break

T-test will either accept or fail to accept a claim (null hypothesis) based on the comparison of P value and alpha value (confidence interval)

Two sample:



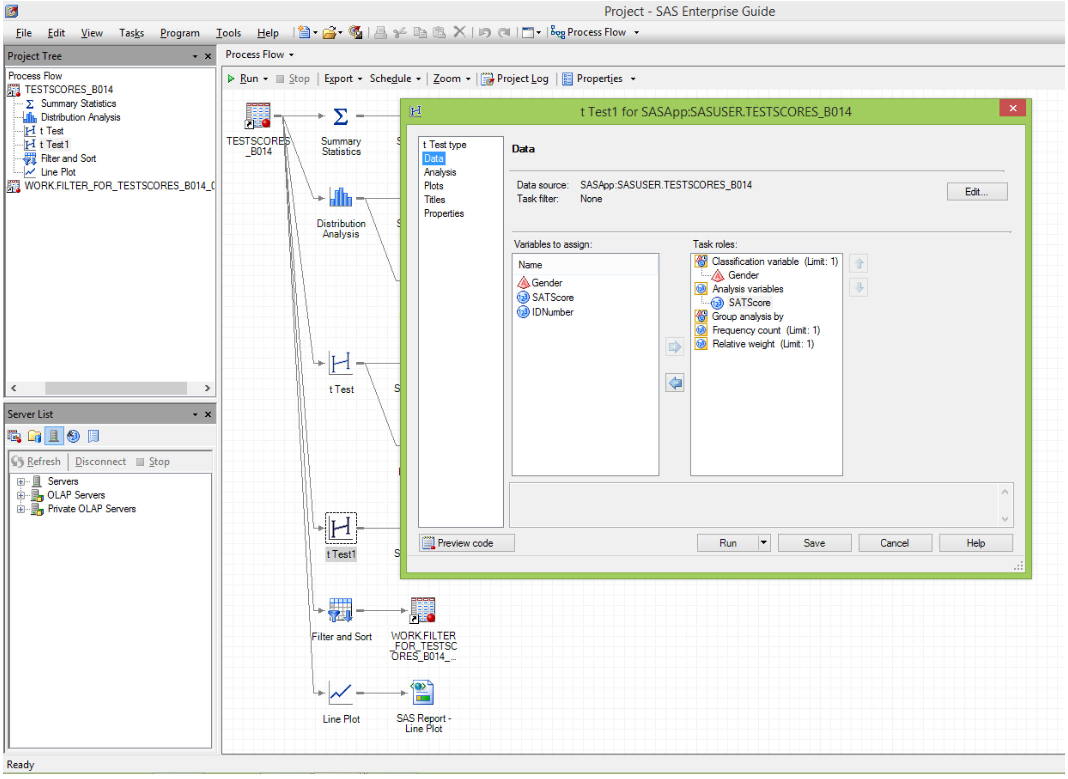
t Test type

Choose t Test type:

● Two Sample

○ Paired

○ One Sample



Data

Data source: SASApp:SASUSER.TESTSCORES_B014

Task filter: None

Variables to assign:

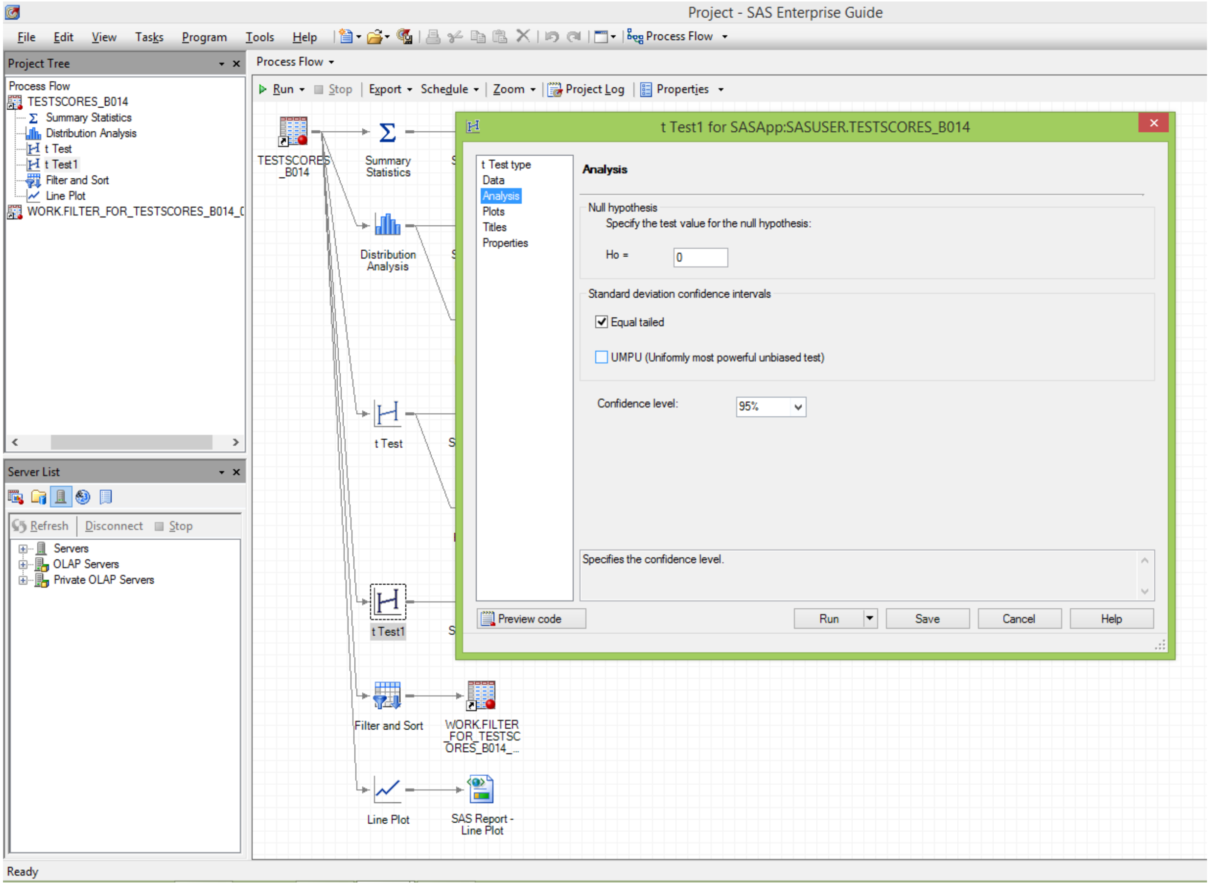
Name

- Gender
- SATScore
- IDNumber

Task roles:

- Classification variable (Limit: 1)
  - Gender
- Analysis variables
  - SATScore
- Group analysis by
- Frequency count (Limit: 1)
- Relative weight (Limit: 1)

Ho is 0 to indicate that the null hypothesis is: values for the two variables are equal.



t Test1 for SASApp:SASUSER.TESTSCORES_B014

Analysis

Null hypothesis

Specify the test value for the null hypothesis:

Ho = 0

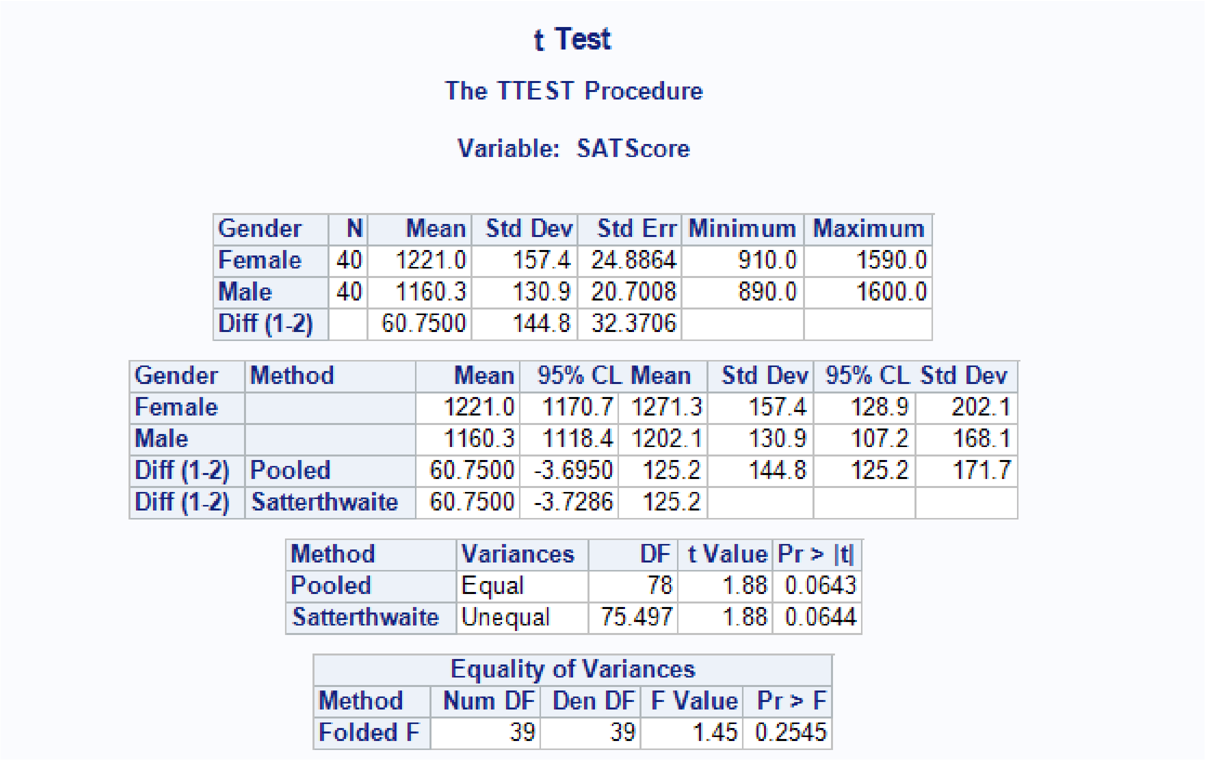
Standard deviation confidence intervals

☑ Equal tailed

□ UMPU (Uniformly most powerful unbiased test)

Confidence level: 95%

Specifies the confidence level.



t Test

The TTEST Procedure

Variable: SATScore

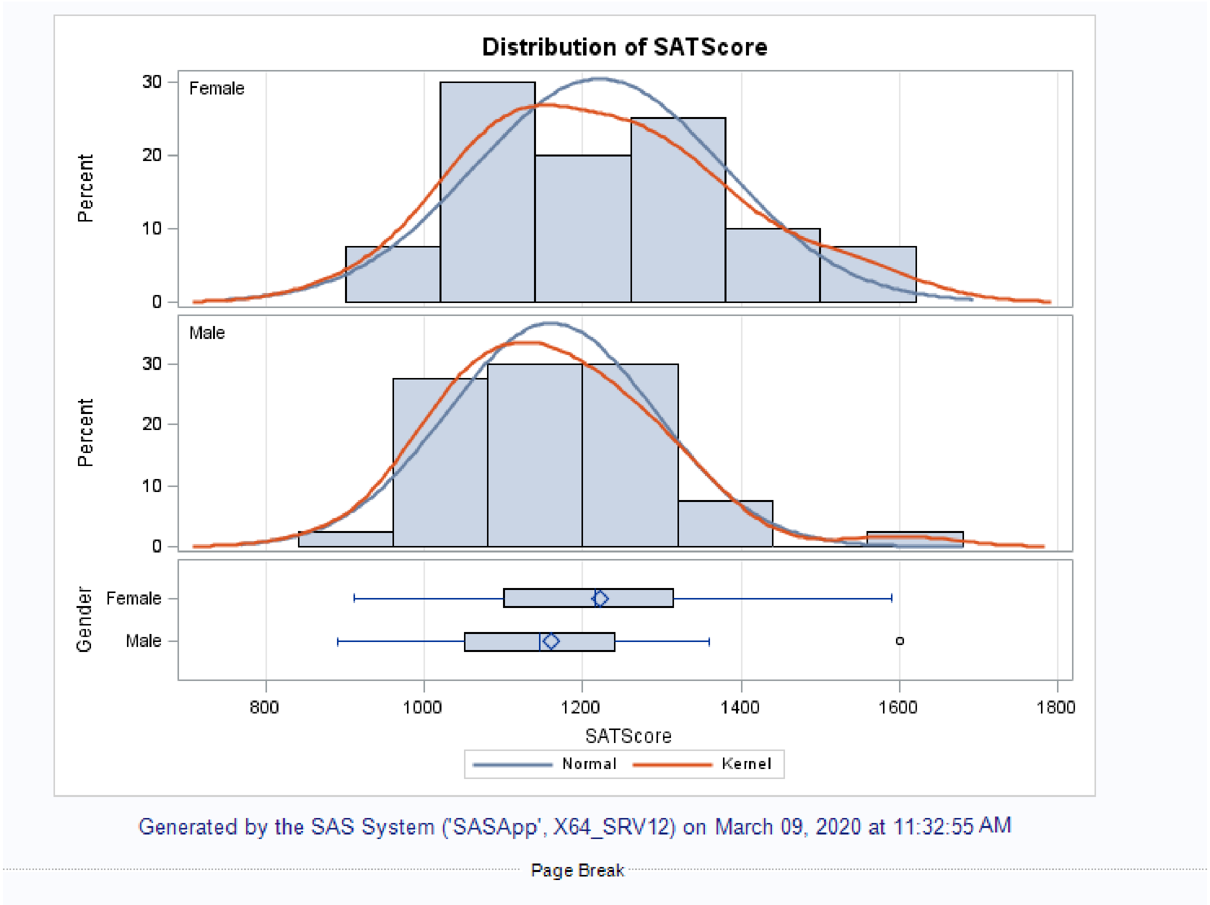
Gender	N	Mean	Std Dev	Std Err	Minimum	Maximum
Female	40	1221.0	157.4	24.8864	910.0	1590.0
Male	40	1160.3	130.9	20.7008	890.0	1600.0
Diff (1-2)		60.7500	144.8	32.3706		

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

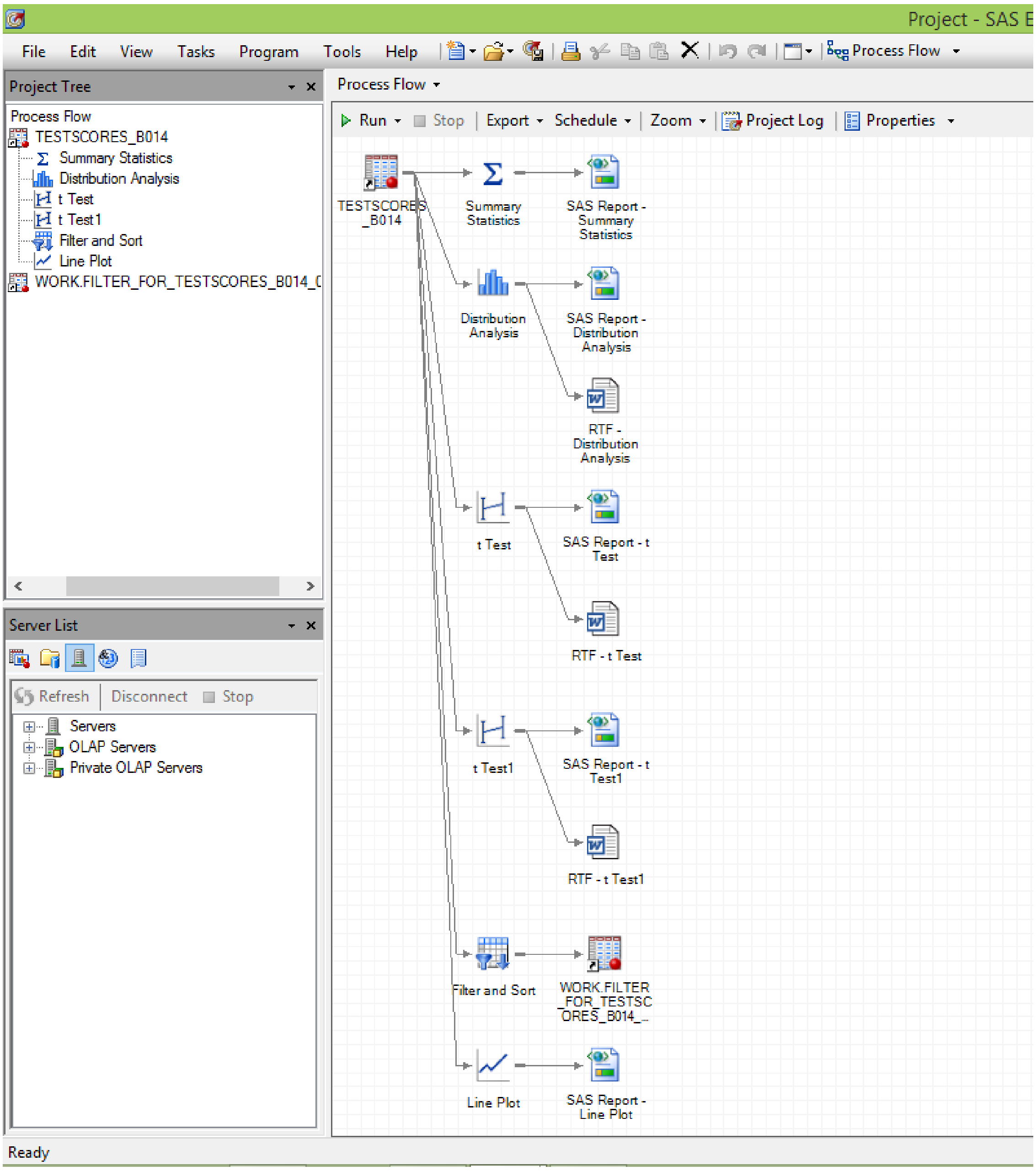


Generated by the SAS System ('SASApp', X64_SRV12) on March 09, 2020 at 11:32:55 AM

Page Break

You either accept or fail to accept the null hypothesis based on comparison of the probability value obtained and the alpha value (confidence value).

Final process flow:


Project - SAS E
File
Edit
View
Tasks
Program
Tools
Help
Process Flow
Project Tree
Process Flow
TESTSCORES_B014
Summary Statistics
Distribution Analysis
t Test
t Test1
Filter and Sort
Line Plot
WORK.FILTER_FOR_TESTSCORES_B014_(
Server List
Refresh
Disconnect
Stop
Servers
OLAP Servers
Private OLAP Servers
Process Flow
Run
Stop
Export
Schedule
Zoom
Project Log
Properties
TESTSCORES_B014
Summary Statistics
SAS Report - Summary Statistics
Distribution Analysis
SAS Report - Distribution Analysis
RTF - Distribution Analysis
t Test
SAS Report - t Test
RTF - t Test
t Test1
SAS Report - t Test1
RTF - t Test1
Filter and Sort
WORK.FILTER_FOR_TESTSCORES_B014_...
Line Plot
SAS Report - Line Plot
Ready