DATA PROGRAMMING WITH SAS FINAL ASSIGNMENT

DATA ANALYSIS I - DATA ANALYSIS II - TASKS DEMONSTRATION

KANDURI SAKETH SAI NIGAM
STUDENT NO. 22201204

DATA ANALYSIS I - 1. IMPORTING HEART-DISEASE DAT AND PRINTING 1ST FIVE ROWS OF DATASET

Obs	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
1	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
2	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
3	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
4	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
5	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1

The Heart-Disease Dataset is imported and Top 5 rows of the dataset are displayed.

The CONTENTS Procedure

Data Set Name	S40840.HEARTDISEASEDATASET	**Observations**	303
Member Type	DATA	**Variables**	14
Engine	V9	**Indexes**	0
Created	07/08/2023 00:57:35	**Observation Length**	112
Last Modified	07/08/2023 00:57:35	**Deleted Observations**	0
Protection		**Compressed**	NO
Data Set Type		**Sorted**	NO
Label			
Data Representation	SOLARIS_X86_64, LINUX_X86_64, ALPHA_TRU64, LINUX_IA64		
Encoding	utf-8 Unicode (UTF-8)		

Engine/Host Dependent Information	
Data Set Page Size	131072
Number of Data Set Pages	1
First Data Page	1
Max Obs per Page	1168
Obs in First Data Page	303
Number of Data Set Repairs	0
Filename	/home/u63443094/sasuser.v94/heartdiseasedataset.sas7bdat
Release Created	9.0401M7
Host Created	Linux
Inode Number	187958645
Access Permission	rw-r--r--
Owner Name	u63443094
File Size	256KB
File Size (bytes)	262144

Variables in Creation Order					
#	**Variable**	**Type**	**Len**	**Format**	**Informat**
1	age	Num	8	BEST12.	BEST32.
2	sex	Num	8	BEST12.	BEST32.
3	cp	Num	8	BEST12.	BEST32.
4	trestbps	Num	8	BEST12.	BEST32.
5	chol	Num	8	BEST12.	BEST32.
6	fbs	Num	8	BEST12.	BEST32.
7	restecg	Num	8	BEST12.	BEST32.
8	thalach	Num	8	BEST12.	BEST32.
9	exang	Num	8	BEST12.	BEST32.
10	oldpeak	Num	8	BEST12.	BEST32.
11	slope	Num	8	BEST12.	BEST32.

The CONTENTS Procedure

Variables in Creation Order					
#	**Variable**	**Type**	**Len**	**Format**	**Informat**
12	ca	Num	8	BEST12.	BEST32.
13	thal	Num	8	BEST12.	BEST32.
14	target	Num	8	BEST12.	BEST32.

The HeartDiseaseDataset dataset has 303 observations (rows) and 14 variables (columns). The dataset contains 14 variables in the order in which they were produced.

DATA ANALYSIS I - 3. FREQUENCY TABLE FOR FEW CATEGORICAL VARIABLE

The FREQ Procedure

sex	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	96	31.68	96	31.68
1	207	68.32	303	100.00

cp	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	143	47.19	143	47.19
1	50	16.50	193	63.70
2	87	28.71	280	92.41
3	23	7.59	303	100.00

target	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	138	45.54	138	45.54
1	165	54.46	303	100.00

The output summarizes the distribution of the Sex, Chest Pain, and Target Categorical variables in the dataset, displaying the count and proportion of each category, as well as cumulative counts and proportions up to each category. This information can help you comprehend the data and how different categories are distributed within each variable. The variable sex is divided into two groups (0 and 1), with 96 occurrences accounting for 31.68% and 207 occurrences accounting for 68.32% of the total. The cumulative frequency for category 0 is 96 (31.68%), whereas it is 303 (100.00%) for category 1. The variable cp is divided into four groups (0, 1, 2, and 3), with relative frequencies of 143 (47.19%), 50 (16.50%), 87 (28.71%), and 23 (7.59%). For categories 0, 1, 2, and 3, the cumulative frequencies are 143 (47.19%), 193 (63.70%), 280 (92.41%), and 303 (100.00%), respectively. Finally, the variable target is divided into two groups (0 and 1), with frequencies of 138 (45.54%) and 165 (54.46%), respectively. For categories 0 and 1, the total frequencies are 138 (45.54%) and 303 (100.00%), respectively. This frequency analysis provides useful information about the distribution of each variable, including the number and percentage of occurrences for each category, as well as cumulative percentages, which can be useful for understanding the data patterns and characteristics related to heart disease in the dataset.

DATA ANALYSIS I - 4. SUMMARY STATISTICS FOR NUMERICAL VARIABLES(SERUM CHOLESTORAL IN MG/DL)

The MEANS Procedure

Analysis Variable : chol				
N	Mean	Std Dev	Minimum	Maximum
303	246.2640264	51.8307510	126.0000000	564.0000000

There are 303 data points in the chol variable that reflect cholesterol levels. The average cholesterol level is about 246.26, with a standard deviation of 51.83 showing some data variability. The range of cholesterol values is between 126 to 564. This data gives an overview of the distribution of cholesterol levels in the dataset and might help you understand the cholesterol profile of the participants in the research.

The UNIVARIATE Procedure
Variable: thalach

Moments			
N	303	**Sum Weights**	303
Mean	149.646865	**Sum Observations**	45343
Std Deviation	22.9051611	**Variance**	524.646406
Skewness	-0.5374097	**Kurtosis**	-0.0619699
Uncorrected SS	6943881	**Corrected SS**	158443.215
Coeff Variation	15.3061417	**Std Error Mean**	1.31586712

Basic Statistical Measures			
Location		**Variability**	
Mean	149.6469	**Std Deviation**	22.90516
Median	153.0000	**Variance**	524.64641
Mode	162.0000	**Range**	131.00000
		Interquartile Range	33.00000

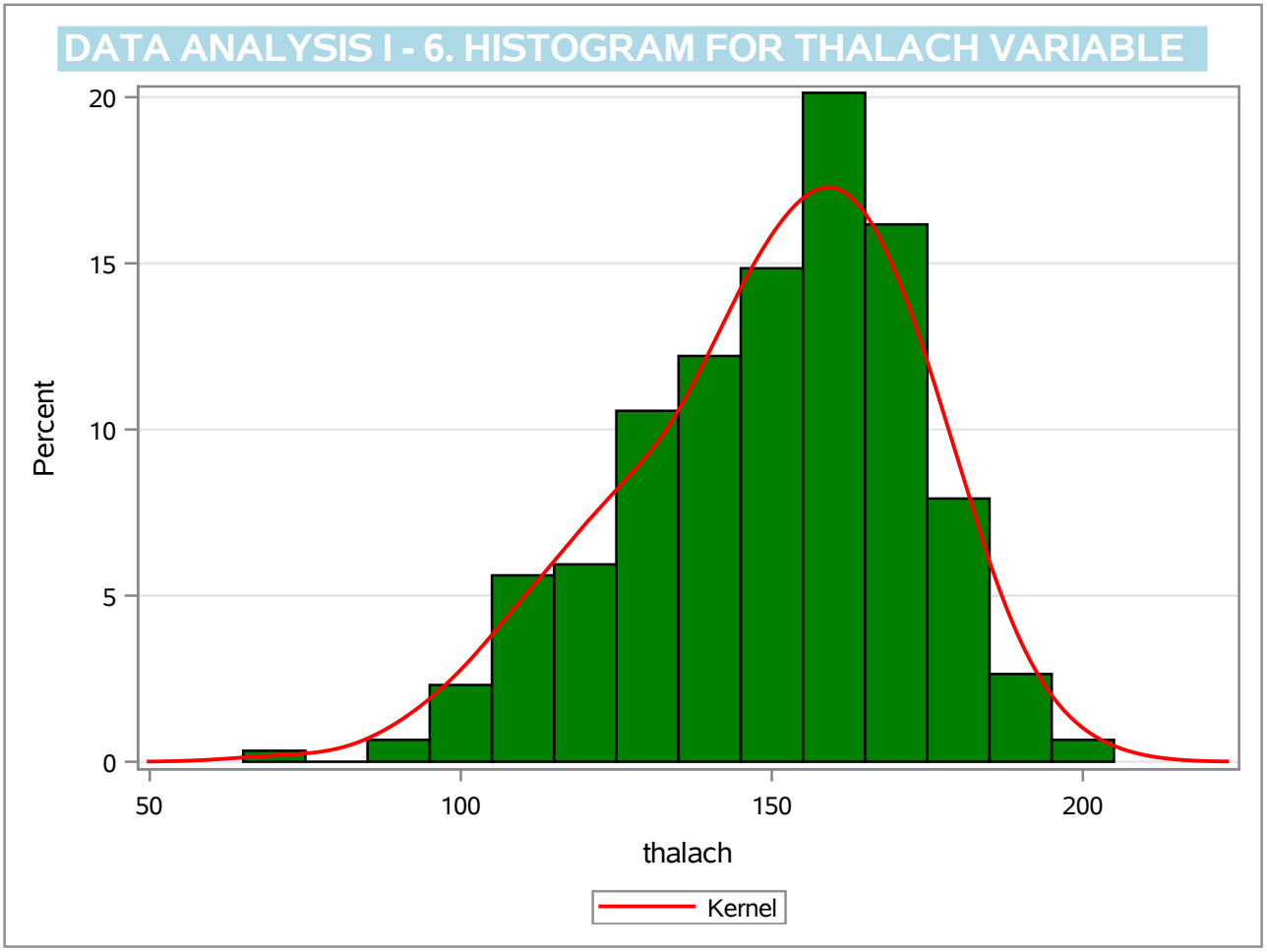
Tests for Location: Mu0=0				
Test	**Statistic**		**p Value**	
Student's t	**t**	113.7249	**Pr > |t|**	<.0001
Sign	**M**	151.5	**Pr >= |M|**	<.0001
Signed Rank	**S**	23028	**Pr >= |S|**	<.0001

Quantiles (Definition 5)	
Level	**Quantile**
100% Max	202
99%	192
95%	182
90%	177
75% Q3	166
50% Median	153
25% Q1	133
10%	116
5%	108
1%	95
0% Min	71

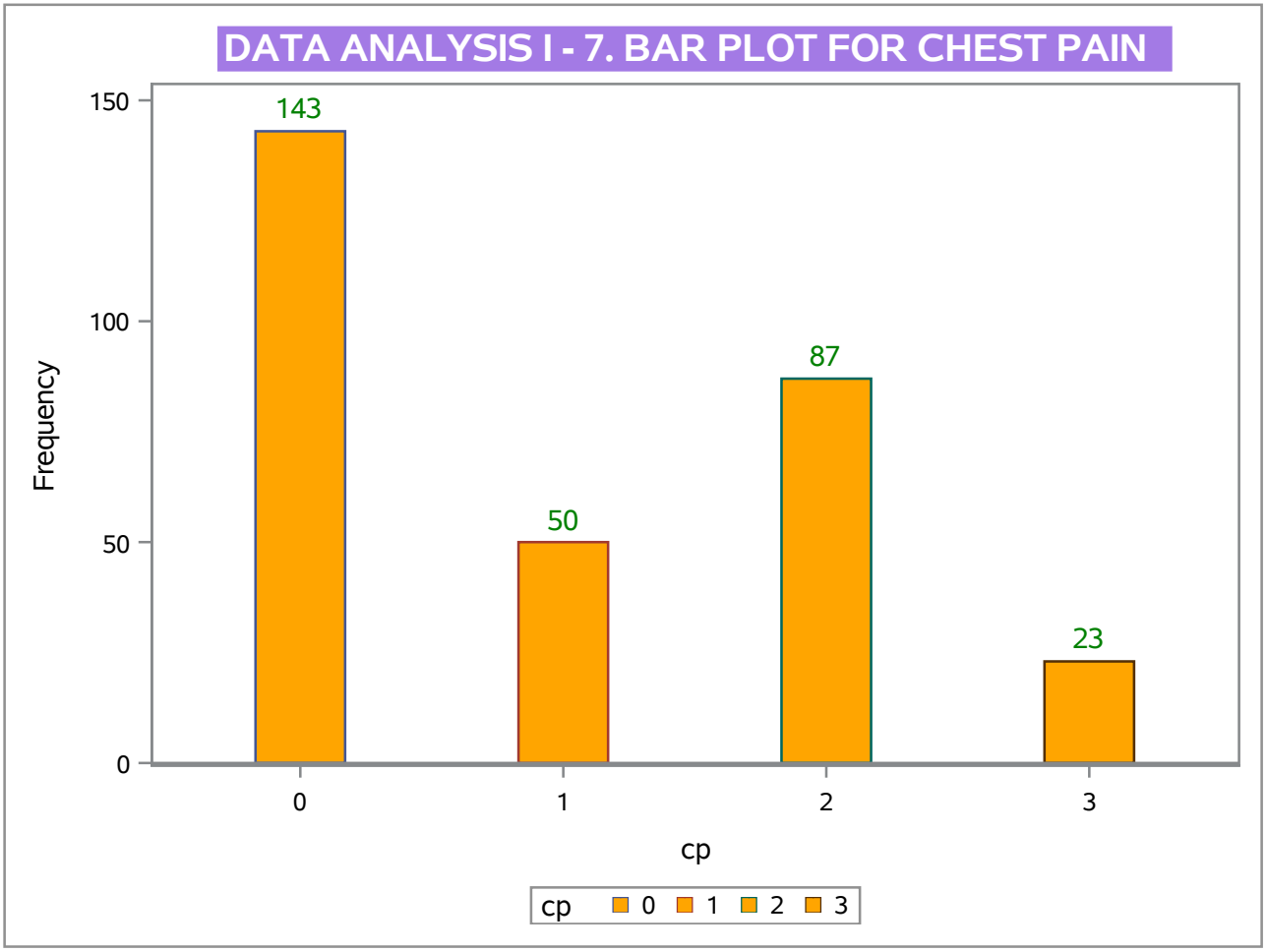
The UNIVARIATE Procedure
Variable: thalach

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
71	273	190	63
88	244	192	126
90	298	194	104
95	263	195	249
96	234	202	73

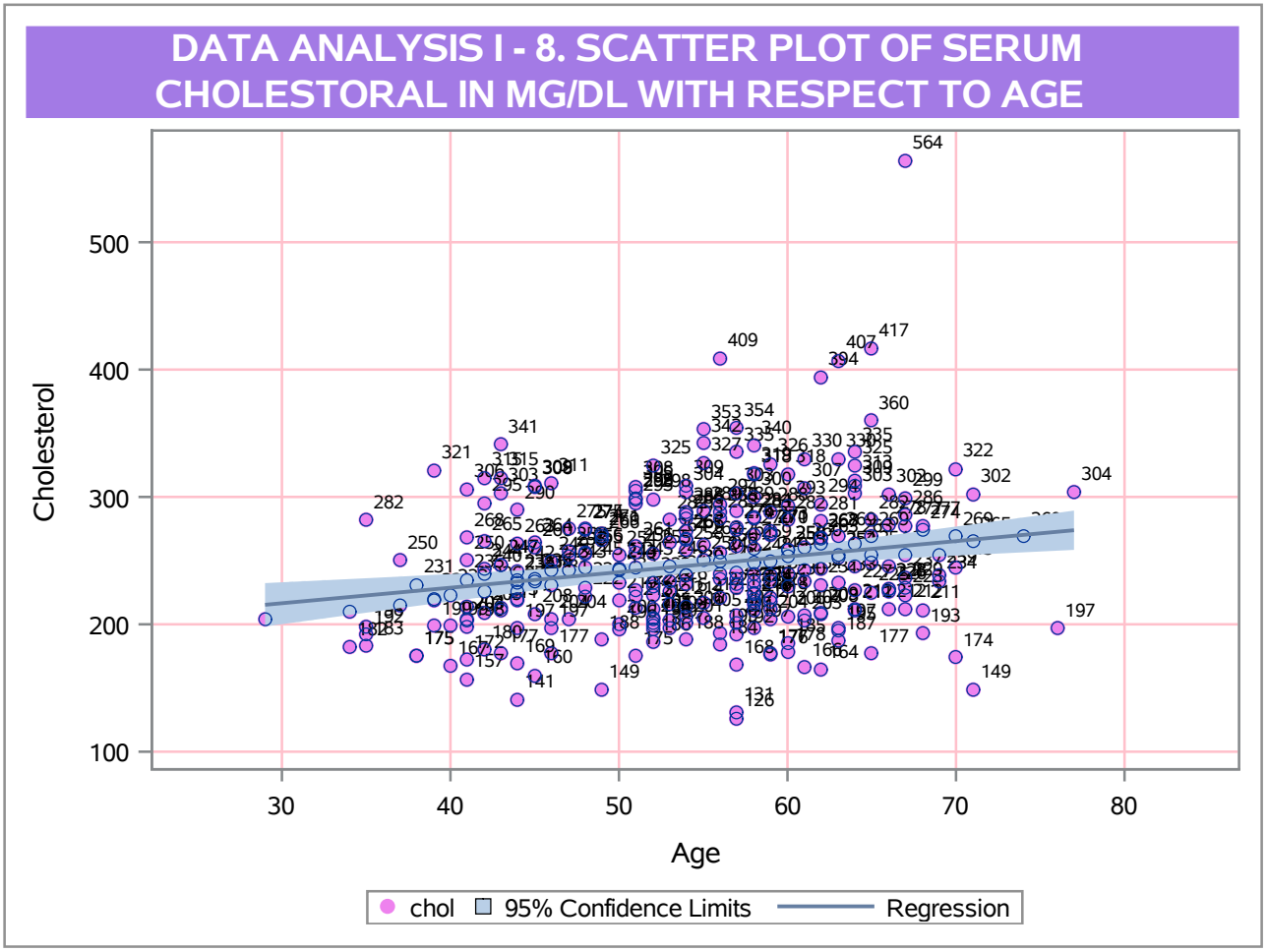
The highest heart rate obtained during exercise is represented by the thalach variable. There are 303 observations of the variable thalach in the dataset. The mean (average) thalach value is around 149.65, while the standard deviation is at 22.91, indicating a moderate dispersion of data points around the mean. The variance is determined to be 524.65, which further demonstrates the data's variability. The skewness of around -0.54 shows that the distribution has a minor leftward asymmetry, whilst the negative kurtosis of approximately -0.06 suggests that the peak is somewhat flatter than in a normal distribution. The thalach value range is 131, with a minimum of 71 and a high of 202. The median (50th percentile) value is 153, which is close to the mean, indicating a fairly symmetric distribution. The spread between the 25th and 75th percentiles is represented by the interquartile range (IQR) of 33. Several location tests, including the Student's t-test, sign test, and signed-rank test, all produce very modest p-values (0.0001), showing strong evidence to reject the null hypothesis that thalach's mean is zero. The quantiles at different levels give information on the distribution of data at various percentiles, with greater quantile values correlating to higher heart rate levels. Overall, this study gives a thorough knowledge of the distribution of maximal heart rate (thalach) reached during exercise in the sample, including its central tendency, variability, and extreme values.



The output plot for the variable thalach from the dataset HeartDiseaseDataset is a graph with a histogram and an overlay density plot. The larger tail on one side of the figure indicates that the data are skewed. Since the longer tail is on the right, the data are positively skewed, and the graphic shows that the distribution is more peaked and has heavier tails.Between 100-200, the majority of the values of the thalach exists and low between 50-100.



By displaying the frequency of each type of chest pain, the bar chart will let you compare the distribution of the CP variable. As there are four distinct categories, Cluster 0 has the largest number of patients reporting chest discomfort (143), while Cluster 3 has the lowest number (23).



The plotted points show a slender linear relationship between age and cholesterol levels. This implies that age and cholesterol levels are positively correlated. We also notice a few outliers. The scatter plot includes people of all ages, ranging from those who are younger (about 30 years old) to those who are older (up to 77 years old). Individual cholesterol levels range from 130 to 565, depending on the person.

DATA ANALYSIS II - 1. Read the dataset erasmus.csv into SAS and call the resulting table erasmus, saving it in the s40840 library. The file contains column names on the first row, with the first observation starting on the second row. You should ensure your code will overwrite any previous object of the same name. (a) Print the first 4 rows of the resulting erasmus table.

Obs	academic_year	duration	nationality	gender	age	sending_country	sending_city	receiving_country	receiving_city
1	2014-2015	1	AT	Female	13	AT	Dornbirn	AT	Dornbirn
2	2014-2015	1	AT	Female	14	AT	Dornbirn	AT	Dornbirn
3	2014-2015	1	AT	Female	15	AT	Dornbirn	AT	Dornbirn
4	2014-2015	1	AT	Male	14	AT	Dornbirn	AT	Dornbirn

The first four rows of the Erasmus dataset, which I successfully imported, will be displayed in the Output table.

DATA ANALYSIS II - 1(b) The duration variable is stored in months. Find the mean duration spent in the programme by students of Irish nationality ('IE'). How many students of Irish nationality are in the dataset?

The MEANS Procedure

Analysis Variable : duration
Mean
1.4148282

DATA ANALYSIS II - 1(b) The duration variable is stored in months. Find the mean duration spent in the programme by students of Irish nationality ('IE'). How many students of Irish nationality are in the dataset?

The FREQ Procedure

nationality	Frequency	Percent	Cumulative Frequency	Cumulative Percent
IE	2765	100.00	2765	100.00

Obs	nationality	COUNT
1	IE	2765

In general, the study indicates that there are 2765 students from Ireland ('IE') in the dataset, with an average duration of 1.4148282 months in the program.

DATA ANALYSIS II - 1(c) One student is older than all other participants. What is the age of this student? In what city did this student study? In what academic year did they start?

MaxAge
80

Obs	age
1	80

City where the Oldest Student Studied

Obs	sending_city
1	Valencia

Obs	academic_year
1	2018-2019

As a whole, this study finds the dataset's oldest student, who is 80 years old. They studied at Valencia and began their academic career in 2018-2019.

DATA ANALYSIS II - 1(d) Create a table of the nationality variable for students who are not from Ireland (that is, their nationality is not 'IE') and whose receiving city is Dublin. The table should be ordered from highest to lowest frequency. What is the most frequent nationality of non-Irish students who studied in Dublin?

Obs	academic_year	duration	nationality	gender	age	sending_country	sending_city	receiving_country	receiving_city
1	2014-2015	5	AT	Female	16	AT	Wien	IE	Dublin
2	2014-2015	5	AT	Female	17	AT	Wien	IE	Dublin
3	2014-2015	5	AT	Female	18	AT	Wien	IE	Dublin
4	2014-2015	5	AT	Male	17	AT	Wien	IE	Dublin
5	2014-2015	5	AT	Male	18	BE	Gent	IE	Dublin

Obs	nationality	COUNT	PERCENT
1	UK	37	17.2093
2	CZ	14	6.5116
3	IT	14	6.5116
4	BE	13	6.0465
5	ES	11	5.1163
6	NO	11	5.1163
7	PL	11	5.1163
8	AM	10	4.6512
9	AT	9	4.1860
10	DE	9	4.1860
11	NL	8	3.7209
12	EL	7	3.2558
13	FR	6	2.7907
14	AL	4	1.8605
15	CA	4	1.8605
16	HU	4	1.8605
17	IS	4	1.8605
18	LU	4	1.8605
19	RO	4	1.8605
20	SO	4	1.8605
21	TR	4	1.8605
22	FI	3	1.3953
23	IN	3	1.3953
24	NG	3	1.3953
25	JM	2	0.9302
26	LT	2	0.9302
27	ZM	2	0.9302
28	ZW	2	0.9302
29	CL	1	0.4651
30	GM	1	0.4651
31	KR	1	0.4651
32	RS	1	0.4651
33	SK	1	0.4651
34	US	1	0.4651

The 1st table shows the top-5 rows of the table which has data of Dublin and Non - IE.With 37 students, the United Kingdom (UK) is the most common nationality among non-Irish students who studied in Dublin.

DATA ANALYSIS II - 1(e) In a single table, print the summary statistics for the age variable, divided into groups by both gender and academic year. Which cohort had the greatest mean age?

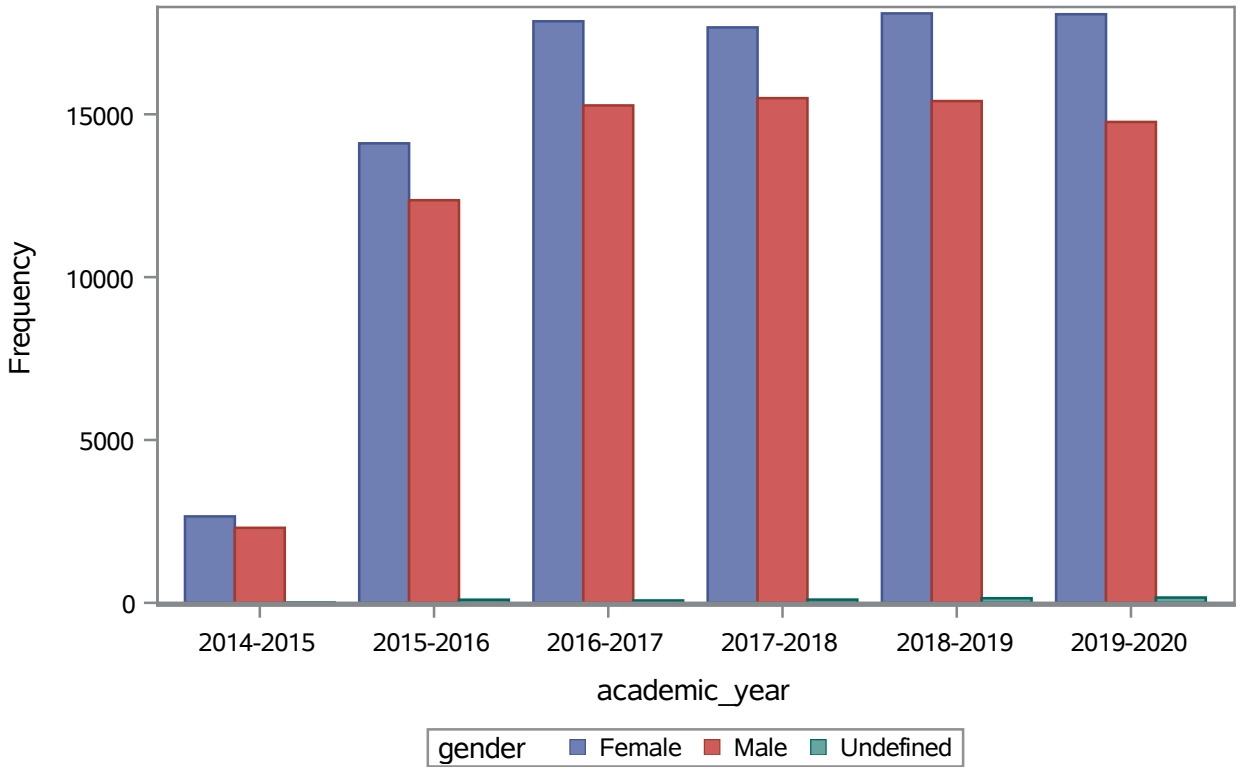
The MEANS Procedure

Analysis Variable : age							
gender	academic_year	N Obs	N	Mean	Std Dev	Minimum	Maximum
Female	2014-2015	2654	2654	23.9412208	9.7880831	10.0000000	77.0000000
	2015-2016	14108	14108	24.0651403	10.3145164	10.0000000	79.0000000
	2016-2017	17855	17855	24.7527303	10.6134282	10.0000000	79.0000000
	2017-2018	17667	17667	24.8643233	10.8716454	11.0000000	78.0000000
	2018-2019	18097	18097	25.0214953	11.2577428	10.0000000	79.0000000
	2019-2020	18071	18071	24.6354380	10.9034254	10.0000000	79.0000000
Male	2014-2015	2303	2303	24.0990013	9.9221870	11.0000000	75.0000000
	2015-2016	12361	12361	24.4508535	10.7722089	10.0000000	78.0000000
	2016-2017	15272	15272	24.8666186	10.6889371	12.0000000	79.0000000
	2017-2018	15495	15495	24.5256534	10.9204191	10.0000000	79.0000000
	2018-2019	15406	15406	24.7404907	11.2338354	10.0000000	80.0000000
	2019-2020	14765	14765	24.4332543	10.7642953	10.0000000	79.0000000
Undefined	2014-2015	9	9	18.0000000	5.4083269	12.0000000	28.0000000
	2015-2016	96	96	20.6250000	8.4868319	11.0000000	59.0000000
	2016-2017	73	73	20.7123288	4.7330970	14.0000000	38.0000000
	2017-2018	99	99	21.2424242	7.1272557	13.0000000	47.0000000
	2018-2019	142	142	24.9718310	11.3405965	13.0000000	62.0000000
	2019-2020	162	162	21.9320988	7.4751153	13.0000000	56.0000000

Obs	gender	academic_year	MeanAge
1	Female	2018-2019	25.021495275

Female cohort had the greatest mean age in the 2018-2019 academic year, with a mean age of 25.021495275.

DATA ANALYSIS II - 1(f) Produce a clustered bar chart of the 'academic year' variable, clustered by gender. Describe the resulting plot.



The bar chart clearly shows a considerable increase, more than tripling, in the number of students participating in the Erasmus program between academic years. 2014-2015 and 2015-2016. However, from 2016 through 2020, the total student population remained largely stable, with little growth. Notably, the overall number of female students attending the program outnumbers male students by a little margin.

DATA ANALYSIS II - 2. For this question, create a subset of the erasmus dataset which contains only those individuals whose receiving country is Ireland ('IE'). Call this subset erasmus2 and use this subset for all of the follwing parts:(a) Conduct a univariate analysis of the age variable for those individuals inerasmus2. Write a short description of your findings, including key statistics and discussion of any plots produced.

The UNIVARIATE Procedure
Variable: age

Moments			
N	2757	**Sum Weights**	2757
Mean	24.4084149	**Sum Observations**	67294
Std Deviation	10.2274589	**Variance**	104.600916
Skewness	1.43706983	**Kurtosis**	1.43734204
Uncorrected SS	1930820	**Corrected SS**	288280.125
Coeff Variation	41.9013646	**Std Error Mean**	0.19478224

Basic Statistical Measures			
Location		**Variability**	
Mean	24.40841	**Std Deviation**	10.22746
Median	20.00000	**Variance**	104.60092
Mode	16.00000	**Range**	56.00000
		Interquartile Range	11.00000

Tests for Location: Mu0=0				
Test	**Statistic**		**p Value**	
Student's t	**t**	125.3113	**Pr > |t|**	<.0001
Sign	**M**	1378.5	**Pr >= |M|**	<.0001
Signed Rank	**S**	1900952	**Pr >= |S|**	<.0001

Quantiles (Definition 5)	
Level	**Quantile**
100% Max	69
99%	55
95%	47
90%	41
75% Q3	28
50% Median	20
25% Q1	17
10%	16
5%	15
1%	14
0% Min	13

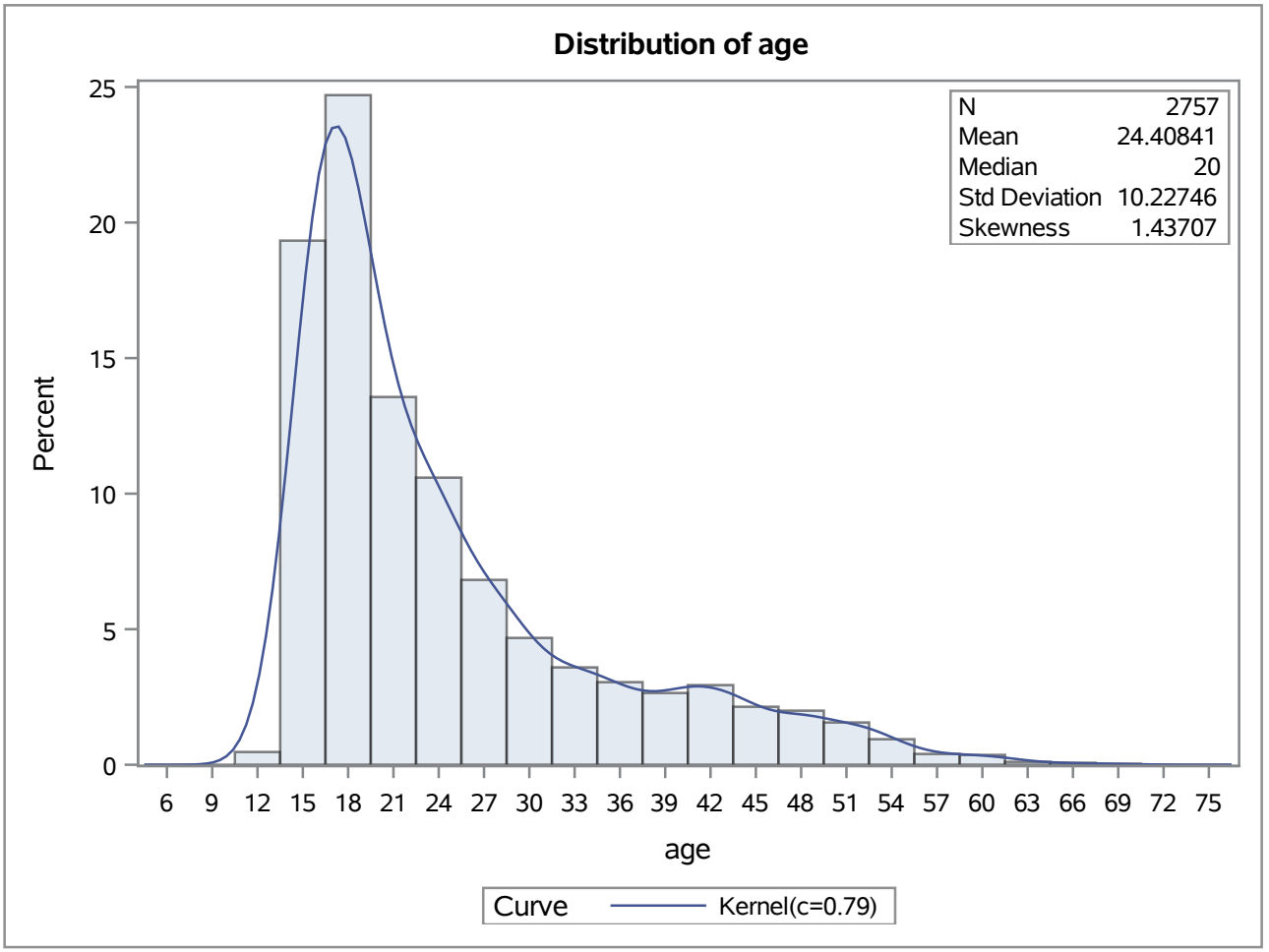
DATA ANALYSIS II - 2. For this question, create a subset of the erasmus dataset which contains only those individuals whose receiving country is Ireland ('IE'). Call this subset erasmus2 and use this subset for all of the follwing parts:(a) Conduct a univariate analysis of the age variable for those individuals inerasmus2. Write a short description of your findings, including key statistics and discussion of any plots produced.

The UNIVARIATE Procedure
Variable: age

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
13	2623	62	1696
13	2434	63	1716
13	1842	65	514
13	1544	66	1717
13	1538	69	1062

DATA ANALYSIS II - 2. For this question, create a subset of the erasmus dataset which contains only those individuals whose receiving country is Ireland ('IE'). Call this subset erasmus2 and use this subset for all of the follwing parts:(a) Conduct a univariate analysis of the age variable for those individuals inerasmus2. Write a short description of your findings, including key statistics and discussion of any plots produced.

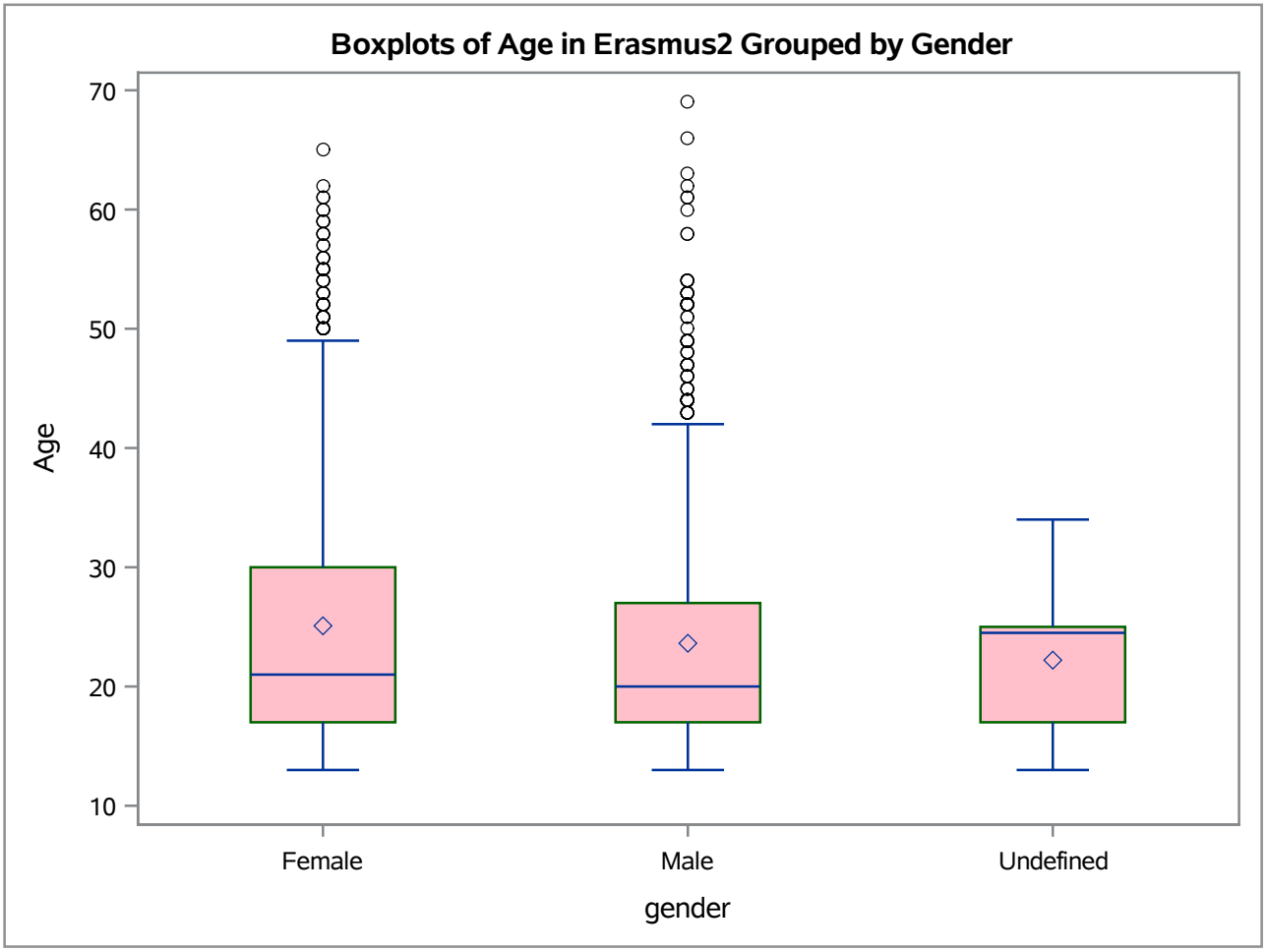
The UNIVARIATE Procedure



The resulting plot is a histogram for the Age variable with a kernel distribution line.The data set's Age variable has a mean value of 24.41 years and a median value of 20 years, showing a right-skewed distribution with several outliers that bring the mean over the median. The standard deviation of 10.23 shows a substantial spread around the mean. The age range runs from 13 years as the lowest value to 69 years as the highest value, illustrating the dataset's variety. Overall, the majority of people are younger, as demonstrated by the median and the leftward tail of the distribution, with a few elderly people contributing to the lengthier right tail.

DATA ANALYSIS II - 2(b) Create boxplots of the age variable in erasmus2, grouped by gender. Ensure the plot is neat with an appropriate title etc. Comment on the resulting plot.

Obs	age	gender
1	16	Female
2	17	Female
3	18	Female
4	18	Female
5	22	Female



According to the box plot, Irish guys participating in the Erasmus program are 18.5 years old on average. The average age of Irish women participating in the Erasmus program is 20. According to the plot, males have bigger unusual ages than women, such as those above 40; and on average, more women than men engage in the Erasmus program.The data's right-skewed distribution for both men and women suggests that a sizable number of persons are older than average.

DATA ANALYSIS II - 2(c) Conduct a hypothesis test to see if there is a statictially significant difference between the mean ages of female and male students, using as your sample data those students in erasmus2. State your hypotheses carefully, check all assump- tions necessary, run your chosen test, comment on the resulting plots and state your conclusion clearly. Use a significance level of α = 0.01.

The TTEST Procedure

Variable: age

gender	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
Female		1496	25.0909	10.6835	0.2762	13.0000	65.0000
Male		1237	23.6257	9.6632	0.2748	13.0000	69.0000
Diff (1-2)	Pooled		1.4652	10.2343	0.3933		
Diff (1-2)	Satterthwaite		1.4652		0.3896		

gender	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
Female		25.0909	24.5491	25.6327	10.6835	10.3139	11.0807
Male		23.6257	23.0867	24.1647	9.6632	9.2969	10.0599
Diff (1-2)	Pooled	1.4652	0.6940	2.2364	10.2343	9.9700	10.5132
Diff (1-2)	Satterthwaite	1.4652	0.7013	2.2291			

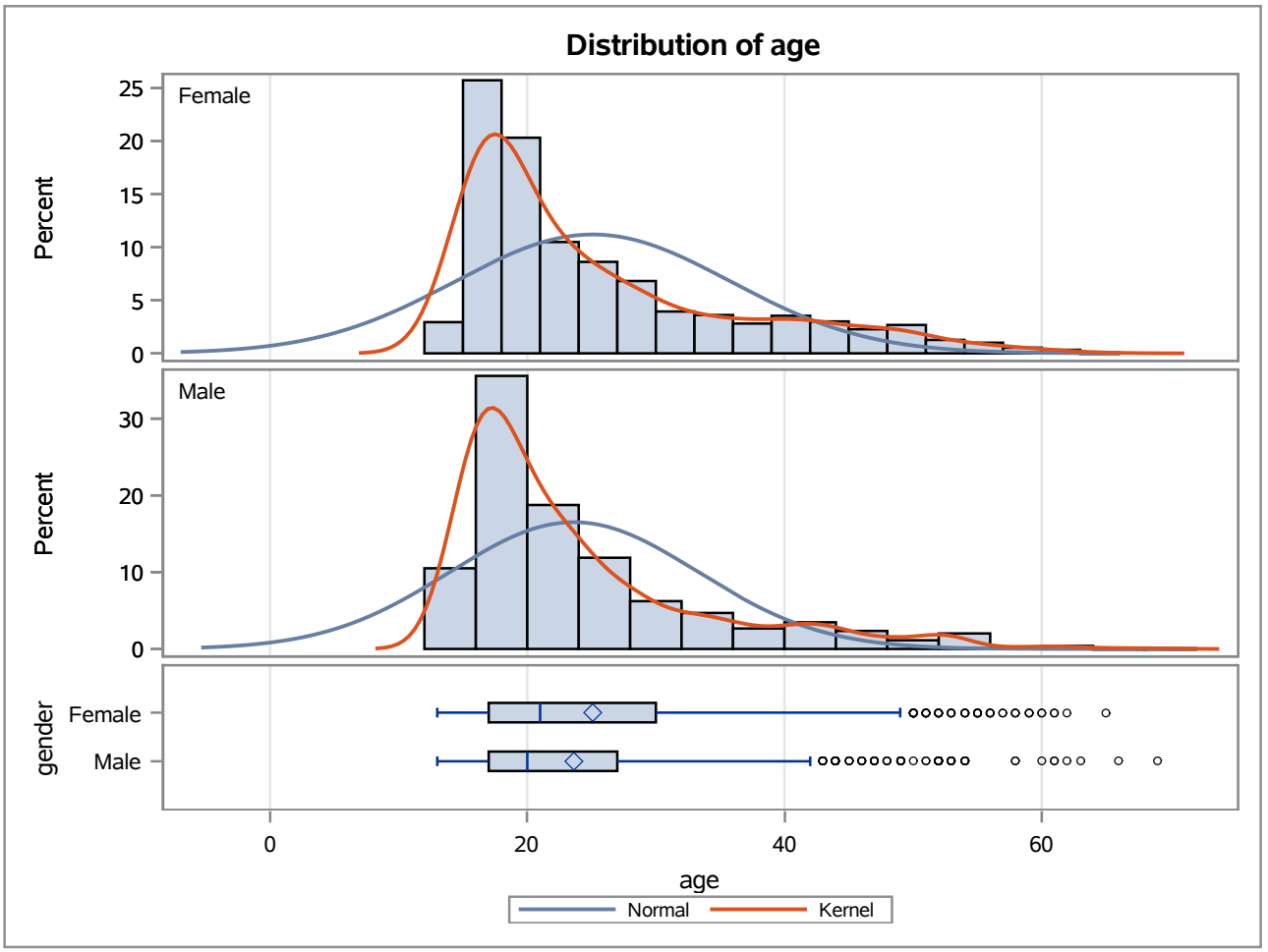
Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	2731	3.73	0.0002
Satterthwaite	Unequal	2709.1	3.76	0.0002

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	1495	1236	1.22	0.0002

DATA ANALYSIS II - 2(c) Conduct a hypothesis test to see if there is a statictially significant difference between the mean ages of female and male students, using as your sample data those students in erasmus2. State your hypotheses carefully, check all assump- tions necessary, run your chosen test, comment on the resulting plots and state your conclusion clearly. Use a significance level of α = 0.01.

The TTEST Procedure

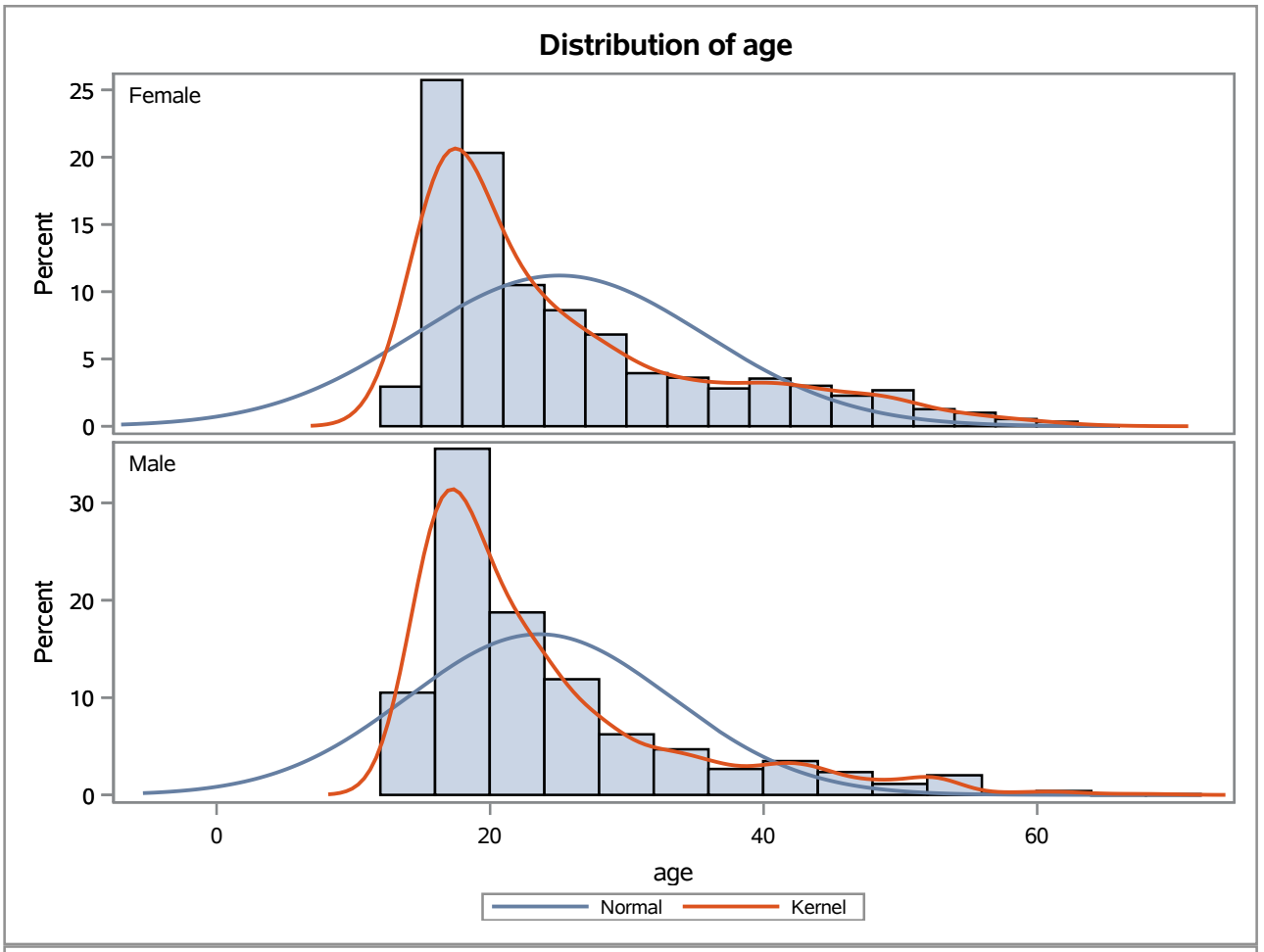
Variable: age

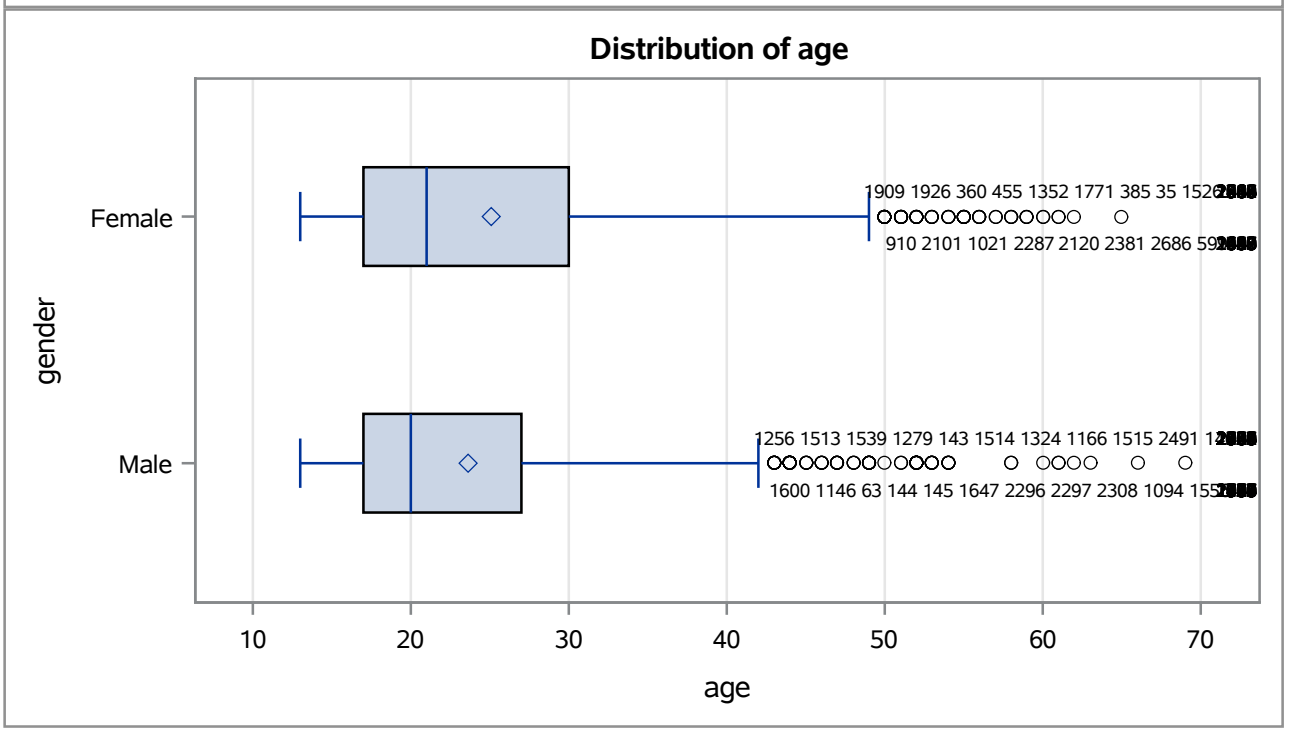


DATA ANALYSIS II - 2(c) Conduct a hypothesis test to see if there is a statictially significant difference between the mean ages of female and male students, using as your sample data those students in erasmus2. State your hypotheses carefully, check all assump- tions necessary, run your chosen test, comment on the resulting plots and state your conclusion clearly. Use a significance level of α = 0.01.

The TTEST Procedure

Variable: age

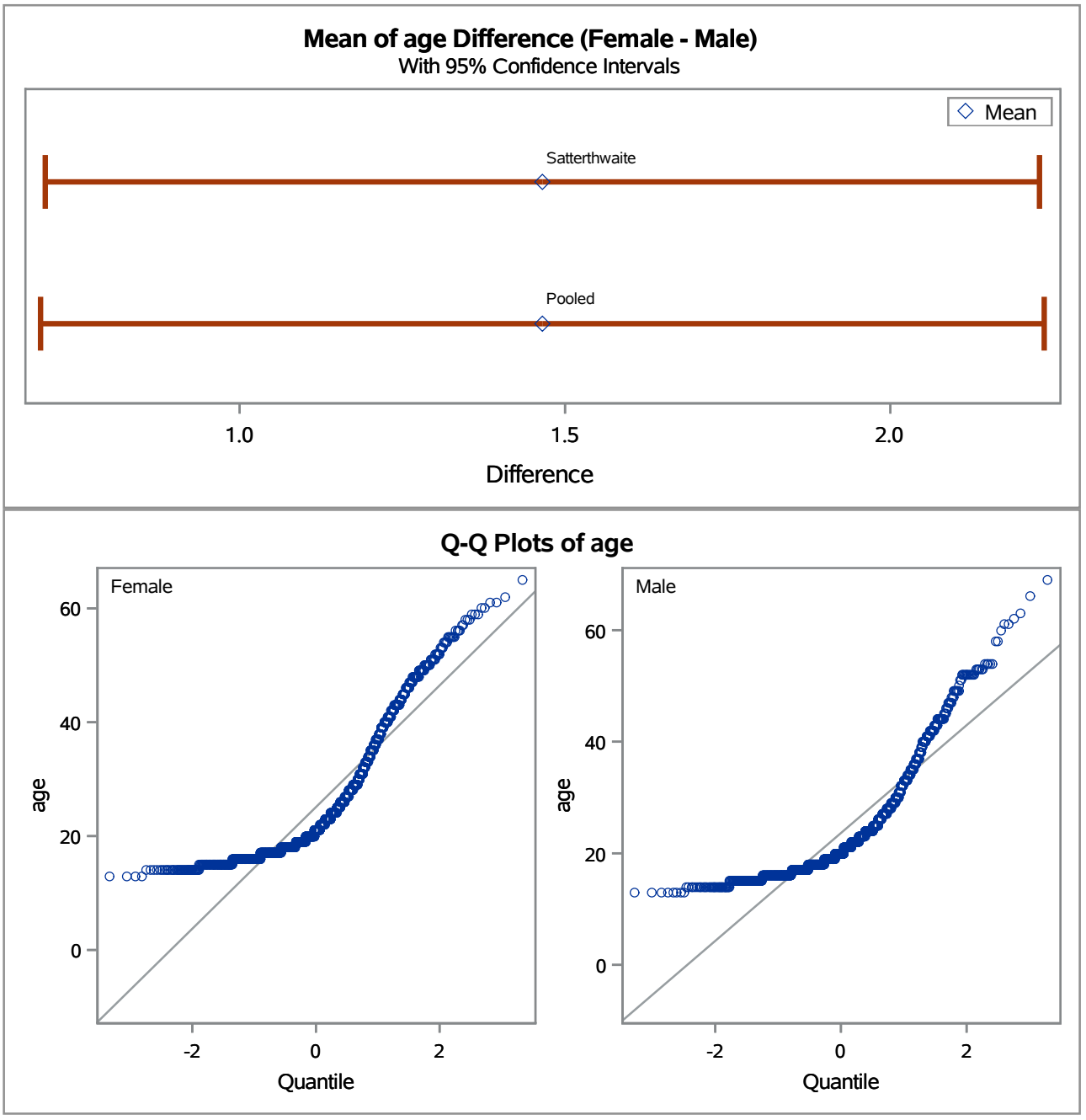




DATA ANALYSIS II - 2(c) Conduct a hypothesis test to see if there is a statictially significant difference between the mean ages of female and male students, using as your sample data those students in erasmus2. State your hypotheses carefully, check all assump- tions necessary, run your chosen test, comment on the resulting plots and state your conclusion clearly. Use a significance level of α = 0.01.

The TTEST Procedure

Variable: age



A hypothesis test is run on the Erasmus2 dataset to see whether there is a statistically significant difference in the mean ages of male and female students. The level of significance (alpha) is set at 0.01. The plots illustrate that both the male and female distributions are right-skewed, but their maxima are different. The null hypothesis (H0) asserts that the average age of male and female students in Ireland's program is the same. The alternative hypothesis (H1) is expressed as (female = male), indicating that the average age of male and female students in Ireland's program differs.Because there might be unequal variances, a two-sample t-test was used, and degrees of freedom (DF)

were calculated using two different approaches. The P-Value for both processes is close to 0.0002, and the t-test values for the pooled approach are about 3.73 and 3.76 for the Satterthwaite method. We reject the null hypothesis (H0) since the p-values for both techniques are less than the significance threshold (= 0.01). The findings show that the mean ages of male and female Erasmus2 students differed statistically considerably. According to the Q- Q plot, the distribution of male and female data is not completely normally distributed and contains some outliers or extremes.On the mean difference plot, the estimated mean difference between the two groups will be displayed as a point.The confidence interval of the mean difference plot excludes 0, showing that the difference in means is statistically significant at the 95% level of assurance. In this case, there is evidence to suggest that the means of the two groups are distinct. The average age of students is believed to be roughly 25.09 years for females and 23.63 years for males, based on the boxplot result shown above. As a result, according to the Erasmus dataset, the t-test results show that female students are older than male students among those studying in Ireland. The hypothesis test and summary data show a statistically significant age difference between male and female students.

Obs	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
1	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
2	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
3	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
4	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
5	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1

PURPOSE OF MULTIVARIATE ANALYSIS:

Using the Multivariate Analysis task, you may analyze and examine multivariate data using a wide range of tools and methods. In a multivariate analysis, many variables are analyzed concurrently to uncover links, patterns, and variances in the data. When working with datasets including several dependent and independent variables, it is very helpful. Principal Component Analysis, Factor Analysis, Canonical Correlation Analysis, Discriminant Analysis, Multidimensional Scaling (MDS), and others are a few of the task's primary functions.

DEMONSTRATION OF MULTIVARIATE ANALYSIS:

1. LOAD THE DATASET

2. PERFORM MULTIVARIATE ANALYSIS

2.1 FACTOR ANALYSIS

2.2 CANONICAL CORRELARION ANALYSIS

Now we are going to perform MULTIVARIATE ANALYSIS by applying few functionalities to the above printed HEARTDISEASEDATASET

TASKS DEMONSTRATION

Input Data Type	Raw Data
Number of Records Read	303
Number of Records Used	303
N for Significance Tests	303

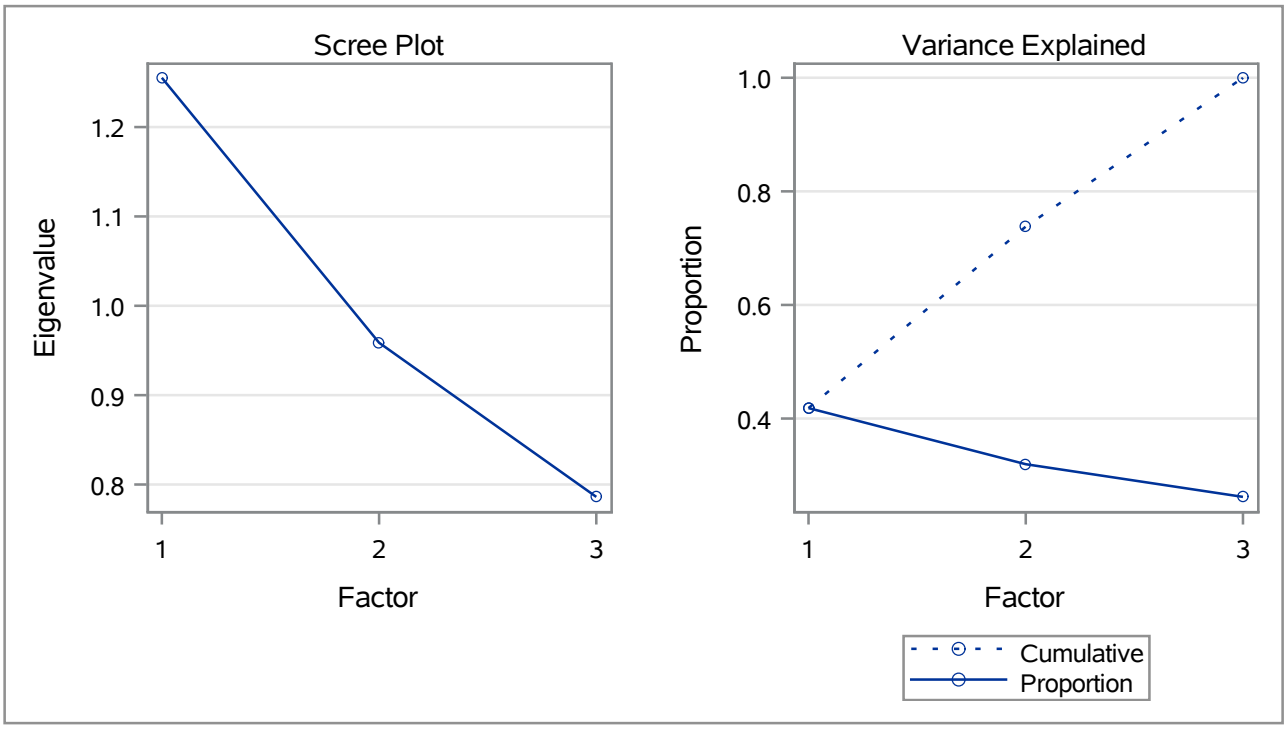
TASKS DEMONSTRATION

Initial Factor Method: Principal Components

Prior Communality Estimates: ONE

Eigenvalues of the Correlation Matrix: Total = 3 Average = 1				
	Eigenvalue	Difference	Proportion	Cumulative
1	1.25520023	0.29654246	0.4184	0.4184
2	0.95865777	0.17251577	0.3196	0.7380
3	0.78614200		0.2620	1.0000

3 factors will be retained by the NFACTOR criterion.



Factor Pattern			
	Factor1	Factor2	Factor3
age	0.73098	0.28071	-0.62198
cp	-0.41915	0.90750	0.02735
chol	0.73836	0.23726	0.63129

Variance Explained by Each Factor		
Factor1	Factor2	Factor3
1.2552002	0.9586578	0.7861420

Final Communality Estimates: Total = 3.000000		
age	cp	chol
1.0000000	1.0000000	1.0000000

PURPOSE OF FACTOR ANALYSIS:

TASKS DEMONSTRATION

In order to identify the common factors that contribute to the variability of the variables age, cp (type of chest pain), and chol (serum cholesterol), factor analysis is a statistical technique used to investigate the underlying structure or latent variables that explain the correlations among observed variables of the HEARTDISEASEDATASET dataset. We specify that we wish to extract three components by using nfactors=3.

1. Initial Factor Method:(Principal Components):

The selected technique (Principal Components) and the previous communality estimates (ONE) are likely included in this table, which also provides information on the initial setup of the factor analysis.

2. Eigenvalues of the Correlation Matrix:

The correlation matrix's eigenvalues are displayed in this table. Each eigenvalues magnitude, the difference between subsequent eigenvalues, the percentage of variance explained by each eigenvalue, and the total percentage of variance explained are all given.

3. Scree and Variance Plots:

To identify the amount of components to maintain, this part could contain visual representations like a scree plot and variance plots.

4. Factor Pattern:

The factor loadings, which represent the strength of each observed variable's relationship to each factor, are shown in this table as the factor pattern. For the variables age, cp, and chol, it displays the factor loadings for three factors (Factor1, Factor2, and Factor3) in this instance.

5. Variance Explained by Each Factor:

The variation that each component (component 1, Factor 2, and Factor 3) explains is shown in this table. It is equivalent to the eigenvalues from the table of the correlation matrix's eigenvalues.

6. Final Communality Estimates:

The final communality estimates for each observed variable (age, cp, and chol) are shown in this table following factor analysis. Communality is the percentage of each variable's variation that can be explained by the components that were kept.

TASKS DEMONSTRATION

Canonical Correlation Analysis

	Canonical Correlation	Adjusted Canonical Correlation	Approximate Standard Error	Squared Canonical Correlation	Eigenvalues of Inv(E)*H = CanRsq/(1-CanRsq)			
					Eigenvalue	Difference	Proportion	Cumulative
1	0.657509	0.645905	0.032666	0.432319	0.7616	0.6367	0.8070	0.8070
2	0.333148	0.302888	0.051157	0.110987	0.1248	0.0795	0.1323	0.9393
3	0.208305	0.182460	0.055047	0.043391	0.0454	0.0335	0.0481	0.9874
4	0.108262	.	0.056869	0.011721	0.0119	0.0118	0.0126	0.9999
5	0.007757	.	0.057540	0.000060	0.0001		0.0001	1.0000

	Test of H0: The canonical correlations in the current row and all that follow are zero				
	Likelihood Ratio	Approximate F Value	Num DF	Den DF	Pr > F
1	0.47709030	9.61	25	1089.9	<.0001
2	0.84041912	3.29	16	898.82	<.0001
3	0.94533981	1.86	9	718.1	0.0542
4	0.98821977	0.88	4	592	0.4758
5	0.99993983	0.02	1	297	0.8937

Multivariate Statistics and F Approximations					
S=5 M=-0.5 N=145.5					
Statistic	**Value**	**F Value**	**Num DF**	**Den DF**	**Pr > F**
Wilks' Lambda	0.47709030	9.61	25	1089.9	<.0001
Pillai's Trace	0.59847793	8.08	25	1485	<.0001
Hotelling-Lawley Trace	0.94367402	11.02	25	709.85	<.0001
Roy's Greatest Root	0.76155147	45.24	5	297	<.0001
NOTE: F Statistic for Roy's Greatest Root is an upper bound.					

TASKS DEMONSTRATION

Canonical Correlation Analysis

Raw Canonical Coefficients for the VAR Variables					
	V1	V2	V3	V4	V5
age	-0.003068818	0.1022872322	-0.030255278	-0.053004917	0.0346560387
sex	-0.88674657	0.4654122447	1.6793500925	-0.818496541	-0.597020081
cp	0.5551114917	0.196140006	-0.00488485	-0.024639294	-0.834442976
chol	-0.002941831	0.006809234	0.0082121866	0.0166008756	-0.003371204
thalach	0.0222357871	0.0140269653	0.0198514925	-0.010033861	0.0364867044

Raw Canonical Coefficients for the WITH Variables					
	W1	W2	W3	W4	W5
trestbps	0.0080532659	0.0458735742	-0.02789457	0.0003992744	0.0201329204
restecg	0.0672743155	-0.690092369	-0.688778724	-1.202090067	1.1516933745
exang	-0.909296199	-0.768061519	-0.878723679	1.6494087498	0.8461091694
thal	-0.072301719	0.3396351375	1.3508599496	0.1356870912	1.0396258958
target	1.4521084769	-0.303049377	0.1220629695	1.548308229	0.9762033603

TASKS DEMONSTRATION

Canonical Correlation Analysis

Standardized Canonical Coefficients for the VAR Variables					
	V1	V2	V3	V4	V5
age	-0.0279	0.9290	-0.2748	-0.4814	0.3147
sex	-0.4132	0.2169	0.7826	-0.3814	-0.2782
cp	0.5729	0.2024	-0.0050	-0.0254	-0.8612
chol	-0.1525	0.3529	0.4256	0.8604	-0.1747
thalach	0.5093	0.3213	0.4547	-0.2298	0.8357

Standardized Canonical Coefficients for the WITH Variables					
	W1	W2	W3	W4	W5
trestbps	0.1412	0.8045	-0.4892	0.0070	0.3531
restecg	0.0354	-0.3629	-0.3622	-0.6321	0.6056
exang	-0.4272	-0.3608	-0.4128	0.7749	0.3975
thal	-0.0443	0.2080	0.8271	0.0831	0.6365
target	0.7244	-0.1512	0.0609	0.7724	0.4870

TASKS DEMONSTRATION

Canonical Structure

Correlations Between the VAR Variables and Their Canonical Variables					
	V1	V2	V3	V4	V5
age	-0.2621	0.8411	-0.4417	-0.1667	0.0309
sex	-0.4310	0.0314	0.7056	-0.4930	-0.2689
cp	0.7576	0.1958	0.0769	-0.1077	-0.6085
chol	-0.1258	0.4897	0.2079	0.8373	0.0055
thalach	0.7096	-0.0021	0.5240	-0.0373	0.4696

Correlations Between the WITH Variables and Their Canonical Variables					
	W1	W2	W3	W4	W5
trestbps	0.0006	0.8564	-0.4332	0.0248	0.2799
restecg	0.1494	-0.4524	-0.2787	-0.5827	0.5964
exang	-0.7457	-0.1717	-0.2759	0.4999	0.2975
thal	-0.3734	0.2398	0.6947	-0.0144	0.5659
target	0.9106	-0.2315	-0.0222	0.3176	0.1263

Correlations Between the VAR Variables and the Canonical Variables of the WITH Variables					
	W1	W2	W3	W4	W5
age	-0.1723	0.2802	-0.0920	-0.0180	0.0002
sex	-0.2834	0.0105	0.1470	-0.0534	-0.0021
cp	0.4981	0.0652	0.0160	-0.0117	-0.0047
chol	-0.0827	0.1632	0.0433	0.0906	0.0000
thalach	0.4665	-0.0007	0.1092	-0.0040	0.0036

Correlations Between the WITH Variables and the Canonical Variables of the VAR Variables					
	V1	V2	V3	V4	V5
trestbps	0.0004	0.2853	-0.0902	0.0027	0.0022
restecg	0.0982	-0.1507	-0.0581	-0.0631	0.0046
exang	-0.4903	-0.0572	-0.0575	0.0541	0.0023
thal	-0.2455	0.0799	0.1447	-0.0016	0.0044
target	0.5987	-0.0771	-0.0046	0.0344	0.0010

PURPOSE OF CANONICAL CORRELATION ANALYSIS:

The Canonical Correlation Analysis (CCA) tasks goal is to use Canonical Correlation Analysis to examine the connections between two sets of variables in a dataset. Users can do multivariate analysis using the CCA task in SAS Studio to see whether there are any significant correlations between the two sets of variables and to select linear combinations (canonical variables) that maximize the correlation between them.

1. Canonical Correlation:

TASKS DEMONSTRATION

The canonical correlations discovered during the study are displayed in this table. Canonical correlation coefficients, adjusted canonical correlation coefficients, approximative standard errors, squared canonical correlations, and eigenvalues of the product of the inverse of the error covariance matrix (E) and the correlation matrix between the canonical variables (H) are all included in this table. It also has a H0 test in it: The current row and all those that follow have zero canonical correlations.

2. Eigenvalues of Inv(E)*H:

For each canonical correlation, this table displays the eigenvalues, differences between subsequent eigenvalues, proportions, and cumulative proportions of variance explained.

3. Multivariate Statistics and F Approximations:

The multivariate statistics and F-approximations in this table are used to evaluate the significance of the canonical correlations.

4. Raw Canonical Coefficients for the VAR Variables:

The variables in the VAR set's raw canonical coefficients are provided in this table. The linear combinations (canonical variables) obtained from the VAR set that maximize the canonical correlation with the WITH set are shown by these coefficients.

5. Raw Canonical Coefficients for the WITH Variables:

This table offers the raw canonical coefficients for the variables in the WITH set, much as Table 4.

6. Standardized Canonical Coefficients for the VAR Variables:

The variables in the VAR set's standardized canonical coefficients are shown in this table. The relative weights of the individual variables in the canonical variables may be more easily compared thanks to standardized coefficients.

7. Standardized Canonical Coefficients for the WITH Variables:

This table displays the standardized canonical coefficients for the variables in the WITH set, similar to Table 6.

8. Canonical Structure:

The correlations between the original variables and their equivalent canonical variables are displayed in this table. It sheds light on the connections between the original variables and the standard variables.

9. Correlations Between the VAR Variables and the Canonical Variables of the WITH Variables:

TASKS DEMONSTRATION

The correlations between the canonical variables obtained from the WITH set and the original variables in the VAR set are shown in this table. The correlations between the original variables in the WITH set and the canonical variables deduced from the VAR set are also shown in this table.