The CONTENTS Procedure

Data Set Name	WORK.HEART	**Observations**	303
Member Type	DATA	**Variables**	14
Engine	V9	**Indexes**	0
Created	10/17/2023 15:39:50	**Observation Length**	112
Last Modified	10/17/2023 15:39:50	**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	/saswork/SAS_workB0BD000177E5_odaws01-usw2.oda.sas.com/SAS_workD8C5000177E5_odaws01-usw2.oda.sas.com/heart.sas7bdat
Release Created	9.0401M7
Host Created	Linux
Inode Number	537031022
Access Permission	rw-r--r--
Owner Name	u62339736
File Size	256KB
File Size (bytes)	262144

Alphabetic List of Variables and Attributes

#	Variable	Type	Len	Format	Informat
1	age	Num	8	BEST12.	BEST32.
12	ca	Num	8	BEST12.	BEST32.
5	chol	Num	8	BEST12.	BEST32.
3	cp	Num	8	BEST12.	BEST32.
9	exang	Num	8	BEST12.	BEST32.

Alphabetic List of Variables and Attributes					
#	**Variable**	**Type**	**Len**	**Format**	**Informat**
6	fbs	Num	8	BEST12.	BEST32.
10	oldpeak	Num	8	BEST12.	BEST32.
7	restecg	Num	8	BEST12.	BEST32.
2	sex	Num	8	BEST12.	BEST32.
11	slope	Num	8	BEST12.	BEST32.
14	target	Num	8	BEST12.	BEST32.
13	thal	Num	8	BEST12.	BEST32.
8	thalach	Num	8	BEST12.	BEST32.
4	trestbps	Num	8	BEST12.	BEST32.

1. Read the file in SAS and display the contents using the PROC IMPORT and PROC PRINT procedures, print only the first 10 observations. (3 points)

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
6	57	1	0	140	192	0	1	148	0	0.4	1	0	1	1
7	56	0	1	140	294	0	0	153	0	1.3	1	0	2	1
8	44	1	1	120	263	0	1	173	0	0	2	0	3	1
9	52	1	2	172	199	1	1	162	0	0.5	2	0	3	1
10	57	1	2	150	168	0	1	174	0	1.6	2	0	2	1

2.Perform basic Data analysis using PROC Means (2 points).

The MEANS Procedure

Variable	N	Mean	Std Dev	Minimum	Maximum
age	303	54.3663366	9.0821010	29.0000000	77.0000000
sex	303	0.6831683	0.4660108	0	1.0000000
cp	303	0.9669967	1.0320525	0	3.0000000
trestbps	303	131.6237624	17.5381428	94.0000000	200.0000000
chol	303	246.2640264	51.8307510	126.0000000	564.0000000
fbs	303	0.1485149	0.3561979	0	1.0000000
restecg	303	0.5280528	0.5258596	0	2.0000000
thalach	303	149.6468647	22.9051611	71.0000000	202.0000000
exang	303	0.3267327	0.4697945	0	1.0000000
oldpeak	303	1.0396040	1.1610750	0	6.2000000
slope	303	1.3993399	0.6162261	0	2.0000000
ca	303	0.7293729	1.0226064	0	4.0000000
thal	303	2.3135314	0.6122765	0	3.0000000
target	303	0.5445545	0.4988348	0	1.0000000

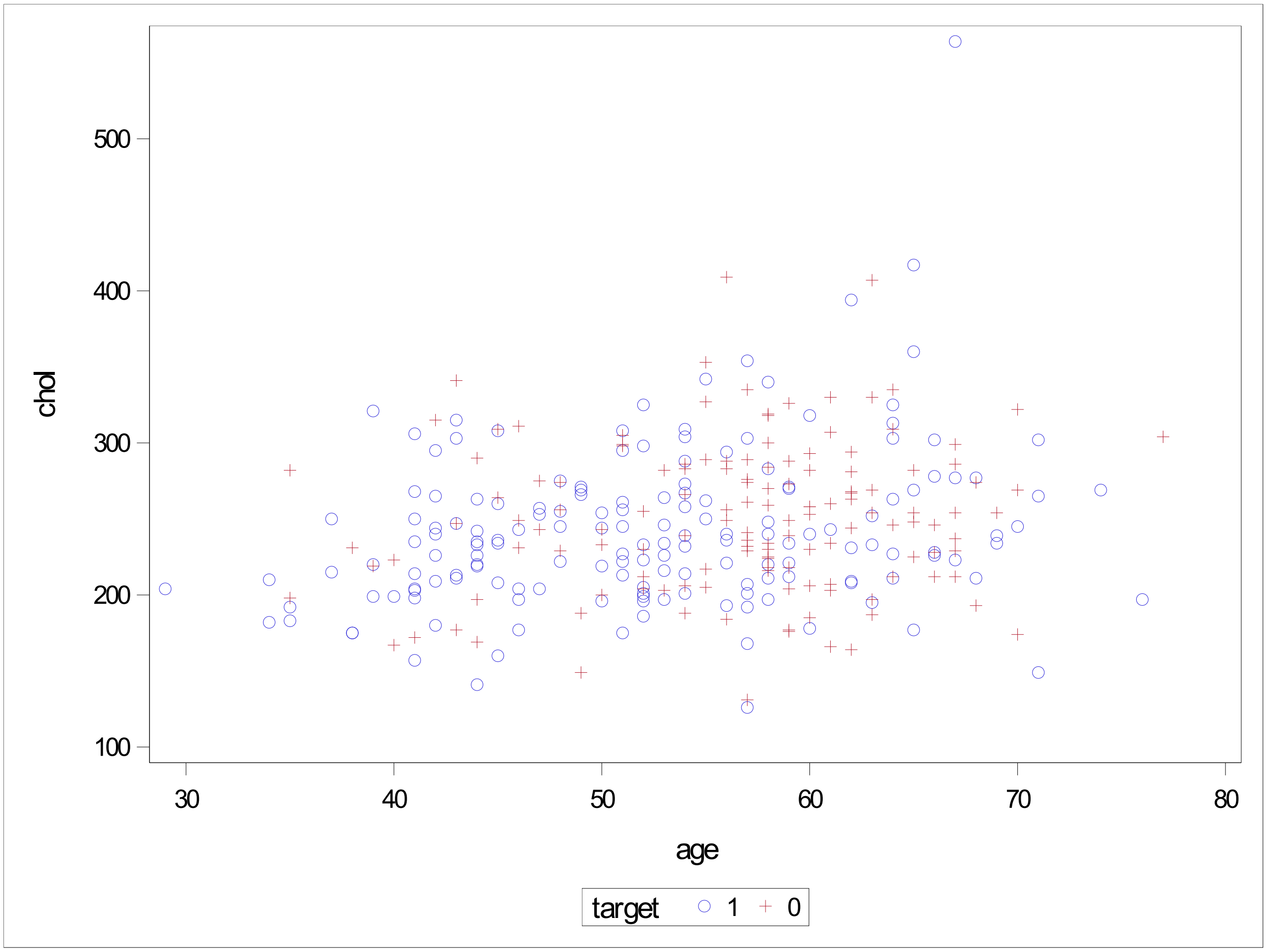
3.Apply standardization to your dataset (to all the attributes) using stdize procedure and print the data (obs=10) (2 points).

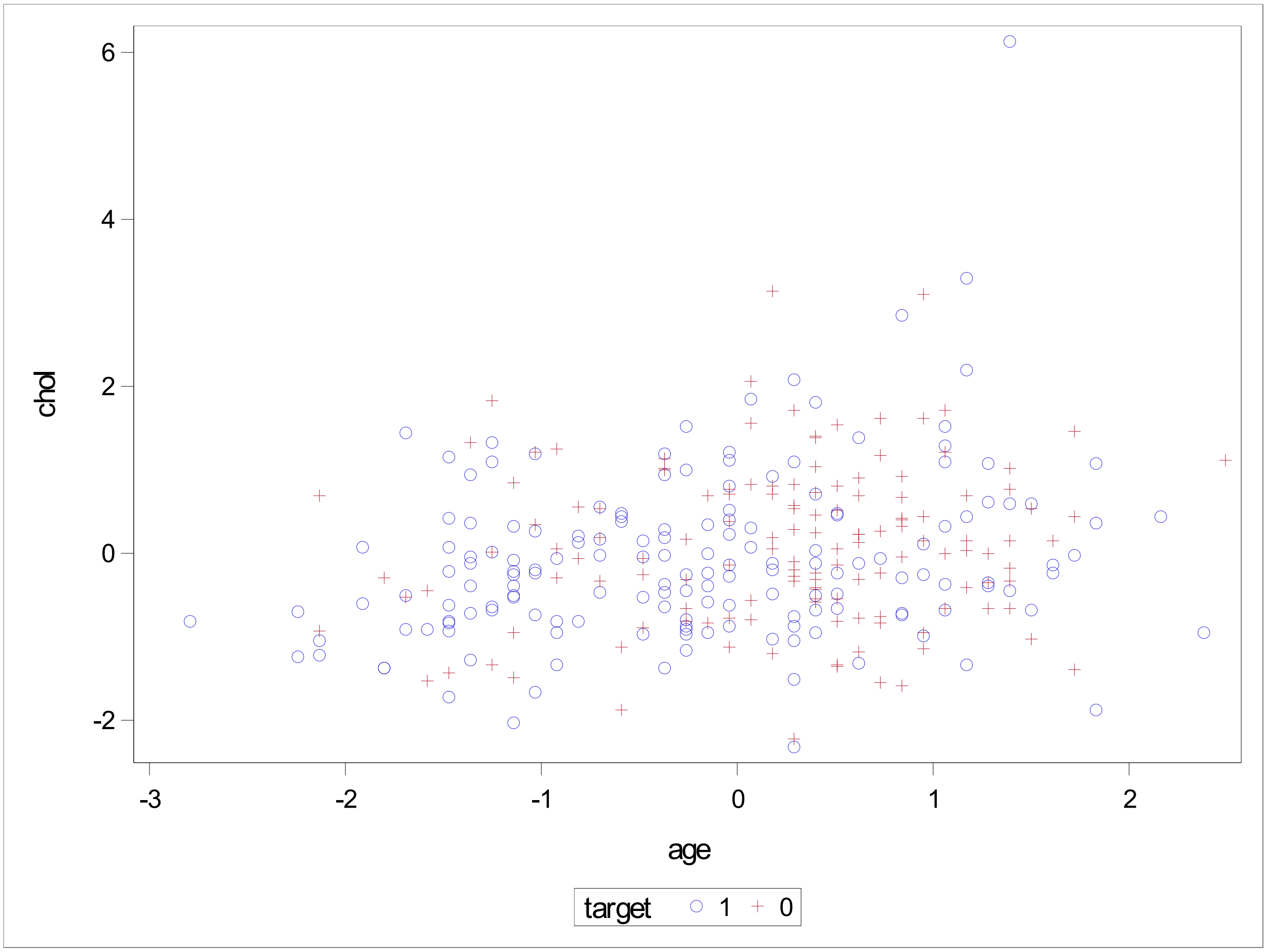
Obs	age	sex	cp	trestbps	chol	fbs	restecg
1	0.9506240215	0.6798805249	1.9698642473	0.7626940758	-0.255910365	2.3904835162	-1.004170712
2	-1.912149695	0.6798805249	1.0009212815	-0.092584625	0.0720802521	-0.416944799	0.8974775738
3	-1.471722969	-1.465992382	0.0319783157	-0.092584625	-0.815423771	-0.416944799	-1.004170712
4	0.1798772518	0.6798805249	0.0319783157	-0.662770426	-0.198029668	-0.416944799	0.8974775738
5	0.2899839332	-1.465992382	-0.93696465	-0.662770426	2.078611086	-0.416944799	0.8974775738
6	0.2899839332	0.6798805249	-0.93696465	0.4776011755	-1.046946559	-0.416944799	0.8974775738
7	0.1798772518	-1.465992382	0.0319783157	0.4776011755	0.9209971433	-0.416944799	-1.004170712
8	-1.141402925	0.6798805249	0.0319783157	-0.662770426	0.3228966063	-0.416944799	0.8974775738
9	-0.260549474	0.6798805249	1.0009212815	2.3021957372	-0.911891599	2.3904835162	0.8974775738
10	0.2899839332	0.6798805249	1.0009212815	1.047786976	-1.509992136	-0.416944799	0.8974775738

Obs	thalach	exang	oldpeak	slope	ca	thal	target
1	0.0154172814	-0.695480041	1.0855422911	-2.270822075	-0.713248971	-2.145323783	1
2	1.6307737425	-0.695480041	2.1190672376	-2.270822075	-0.713248971	-0.512074772	1
3	0.9758995015	-0.695480041	0.3103985813	0.9747396642	-0.713248971	-0.512074772	1
4	1.2378491979	-0.695480041	-0.206363892	0.9747396642	-0.713248971	-0.512074772	1
5	0.5829749569	1.4331103867	-0.37861805	0.9747396642	-0.713248971	-0.512074772	1
6	-0.071899284	-0.695480041	-0.550872207	-0.648041205	-0.713248971	-2.145323783	1
7	0.1463921296	-0.695480041	0.2242715024	-0.648041205	-0.713248971	-0.512074772	1
8	1.0195577842	-0.695480041	-0.895380523	0.9747396642	-0.713248971	1.1211742386	1
9	0.5393166742	-0.695480041	-0.464745129	0.9747396642	-0.713248971	1.1211742386	1
10	1.063216067	-0.695480041	0.482652739	0.9747396642	-0.713248971	-0.512074772	1

4.Apply k-means clustering using fastclus procedure of SAS. Scatter plot your cluster labels (use y=chol and x=age) to visualize and compare with the original data labels. Assuming that you do not know the exact number of clusters in the dataset, try k=2, 3, 4, 5 and evaluate the solutions. Choose the best K value based on an appropriate evaluation metric (e.g. the total within-cluster sum of squares). (8 points)

Answer:
According RMS Std Deviation that measures the degree of homogeneity between the clusters. The RMS values need to be similar for a good clustering solution I notice that RMS values for K=2 are mostly similar compare to other k values, in K=3 also similarity is high and mostly their similarity are the same(k=2 & k=3) because K=2 has less cluster
I choose K=2.


chol
500
400
300
200
100
30
40
50
60
70
80
age
target ○ 1 + 0



The FASTCLUS Procedure
Replace=FULL Radius=0 Maxclusters=2 Maxiter=100 Converge=0.02

Convergence criterion is satisfied.

Criterion Based on Final Seeds =	0.9192

Cluster Summary						
Cluster	Frequency	RMS Std Deviation	Maximum Distance from Seed to Observation	Radius Exceeded	Nearest Cluster	Distance Between Cluster Centroids
1	128	0.9964	5.6412		2	2.8453
2	175	0.8640	7.0174		1	2.8453

Pseudo F Statistic =	54.14

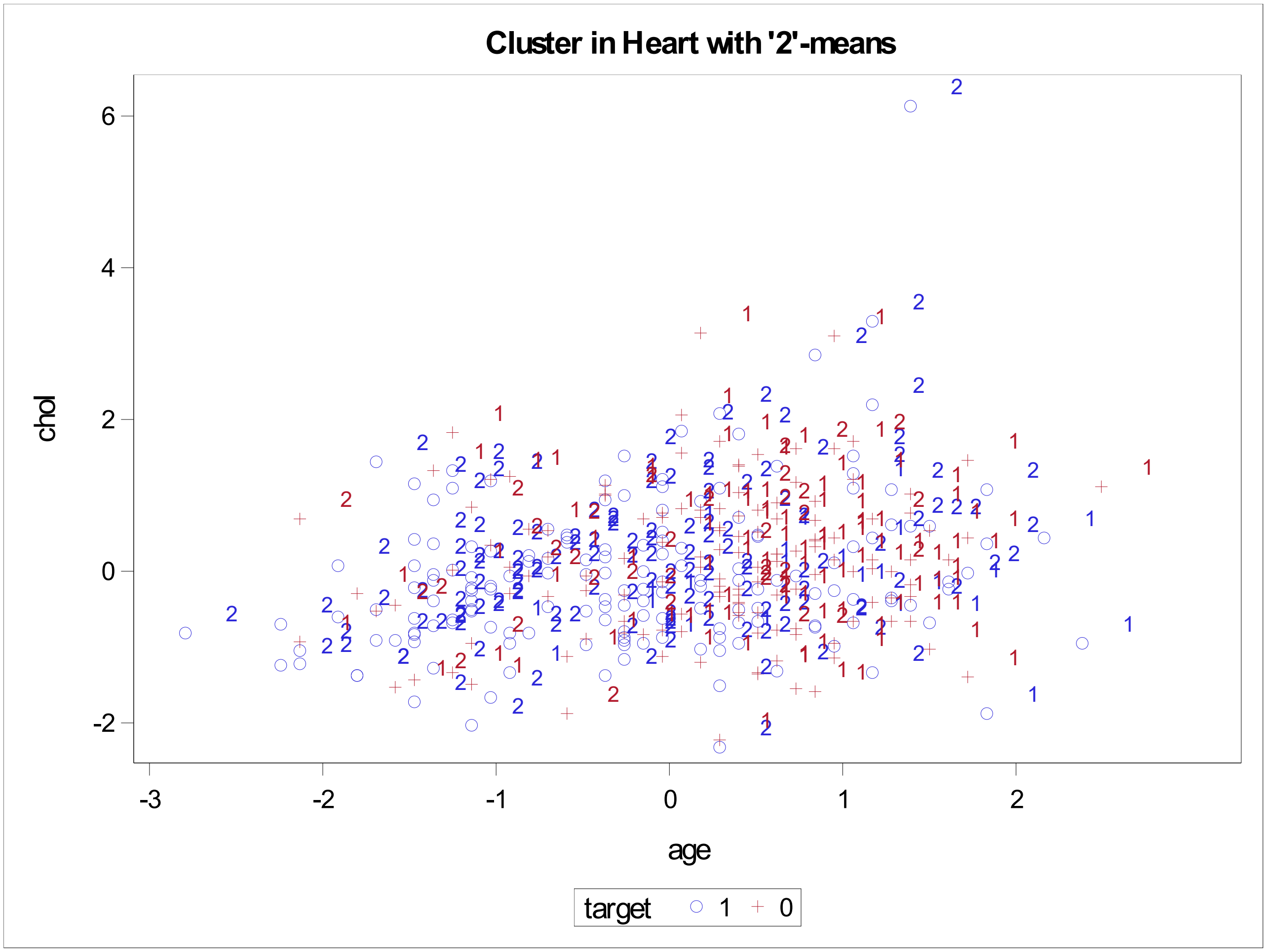
Cluster in Heart with '2'-means

Observed Over-All R-Squared =	0.15244

Approximate Expected Over-All R-Squared =	0.06399

Cubic Clustering Criterion =	32.474

WARNING: The two values above are invalid for correlated variables.



The FASTCLUS Procedure

Replace=FULL Radius=0 Maxclusters=3 Maxiter=100 Converge=0.02

Convergence criterion is satisfied.

Criterion Based on Final Seeds =	0.8851

Cluster in Heart with '3'-means

Cluster Summary

Cluster	Frequency	RMS Std Deviation	Maximum Distance from Seed to Observation	Radius Exceeded	Nearest Cluster	Distance Between Cluster Centroids
1	132	0.8296	5.1995		3	2.2894
2	98	0.9592	5.5813		3	3.1067
3	73	0.8956	6.3447		1	2.2894

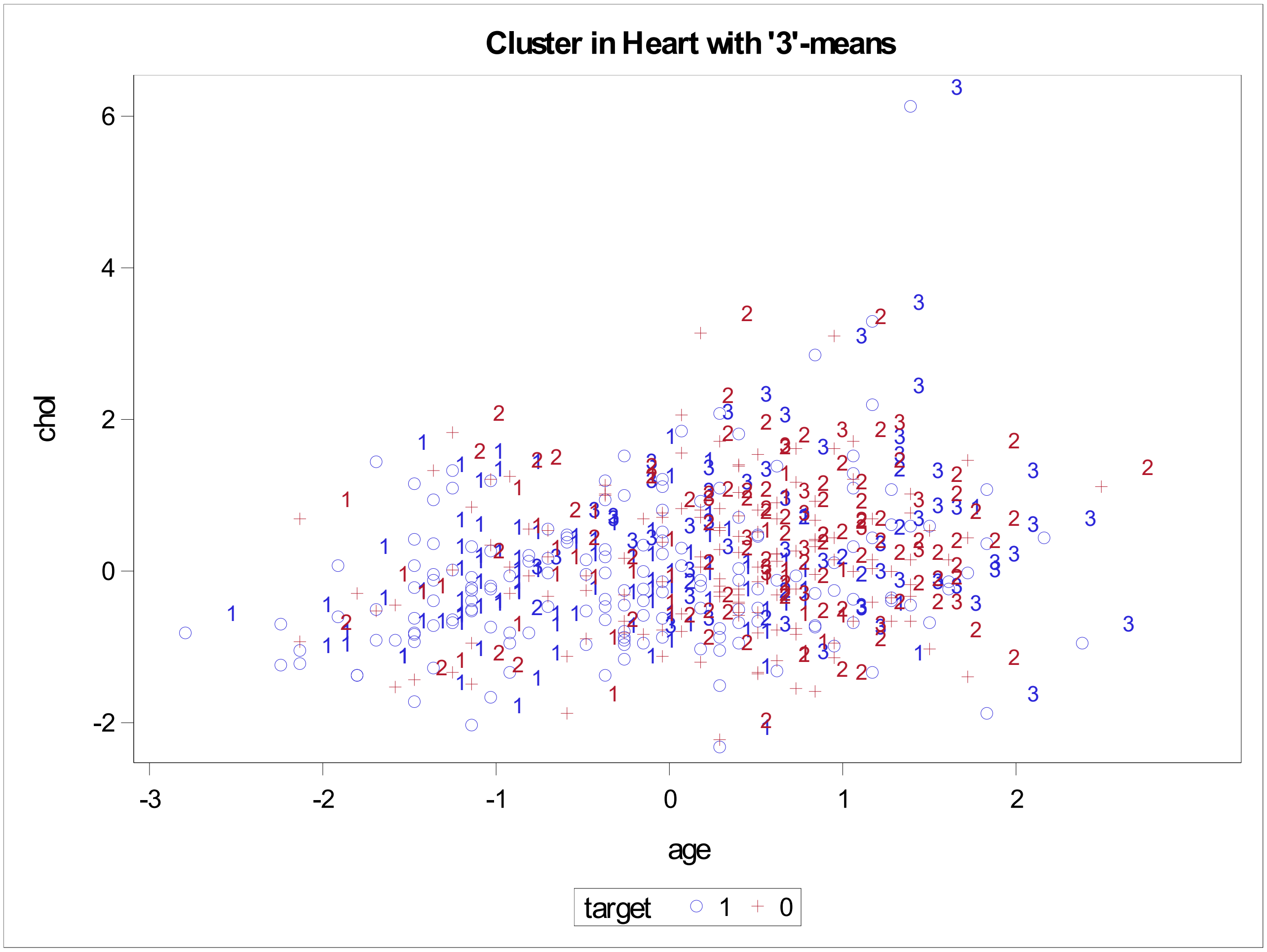
Pseudo F Statistic = 41.01

Observed Over-All R-Squared = 0.21469

Approximate Expected Over-All R-Squared = 0.11370

Cubic Clustering Criterion = 28.311

WARNING: The two values above are invalid for correlated variables.



The FASTCLUS Procedure

Replace=FULL Radius=0 Maxclusters=4 Maxiter=100 Converge=0.02

Convergence criterion is satisfied.

Criterion Based on Final Seeds =	0.8650

Cluster Summary						
Cluster	Frequency	RMS Std Deviation	Maximum Distance from Seed to Observation	Radius Exceeded	Nearest Cluster	Distance Between Cluster Centroids
1	66	0.9477	5.5343		2	2.3308
2	67	0.9436	4.6948		1	2.3308
3	116	0.8030	4.3887		4	2.4577
4	54	0.8138	6.1487		3	2.4577

Cluster in Heart with '4'-means

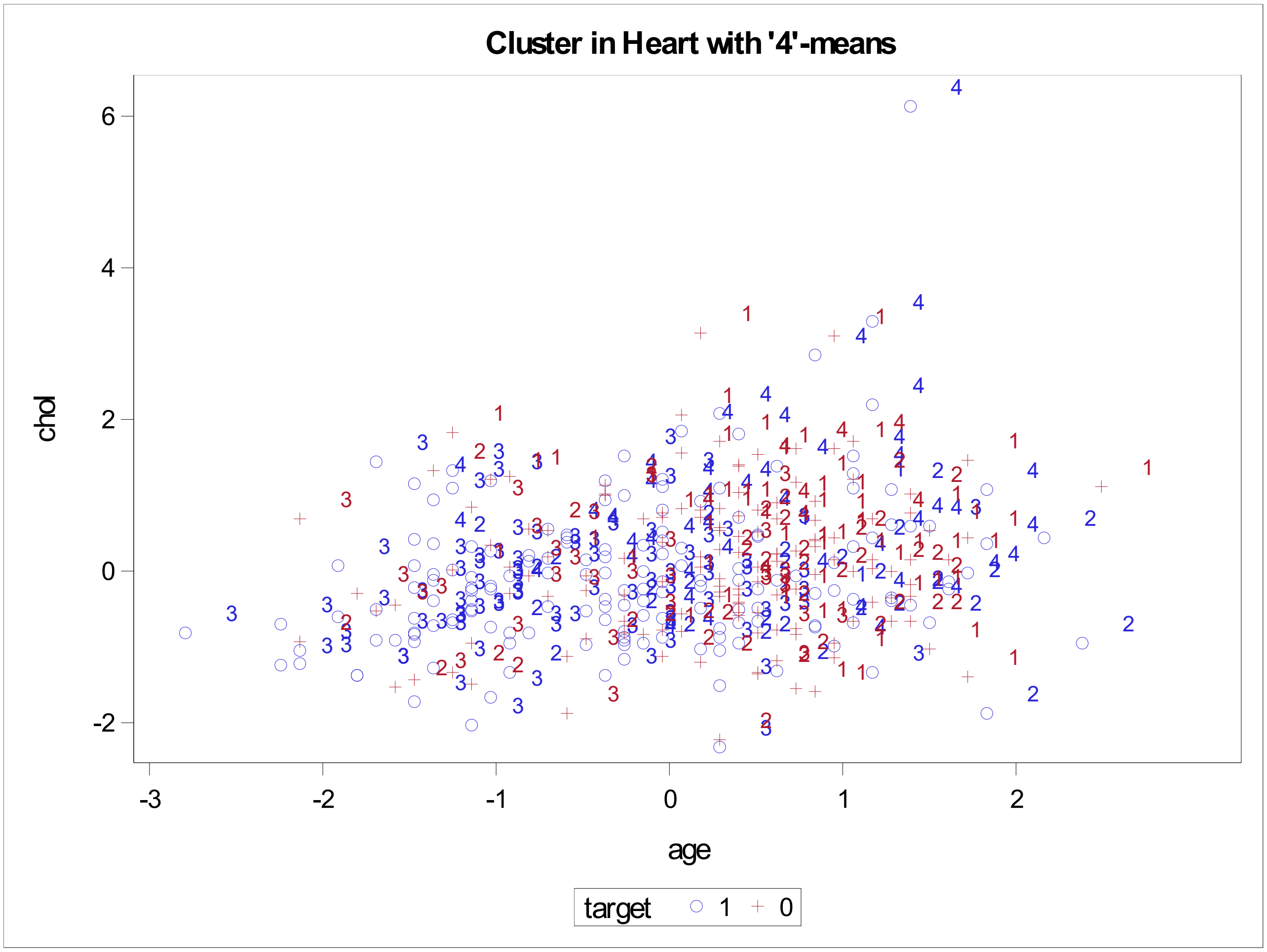
Pseudo F Statistic =	33.28

Observed Over-All R-Squared =	0.25035

Approximate Expected Over-All R-Squared =	0.15513

Cubic Clustering Criterion =	23.673

WARNING: The two values above are invalid for correlated variables.



The FASTCLUS Procedure
Replace=FULL Radius=0 Maxclusters=5 Maxiter=100 Converge=0.02

Convergence criterion is satisfied.

Cluster in Heart with '5'-means

Criterion Based on Final Seeds =	0.8391

Cluster Summary						
Cluster	**Frequency**	**RMS Std Deviation**	**Maximum Distance from Seed to Observation**	**Radius Exceeded**	**Nearest Cluster**	**Distance Between Cluster Centroids**
1	48	0.8996	4.5125		4	2.6280
2	66	0.8294	4.2676		3	2.3901
3	42	1.0578	5.3096		2	2.3901
4	88	0.7307	4.1303		5	2.5136
5	59	0.8107	6.2073		4	2.5136

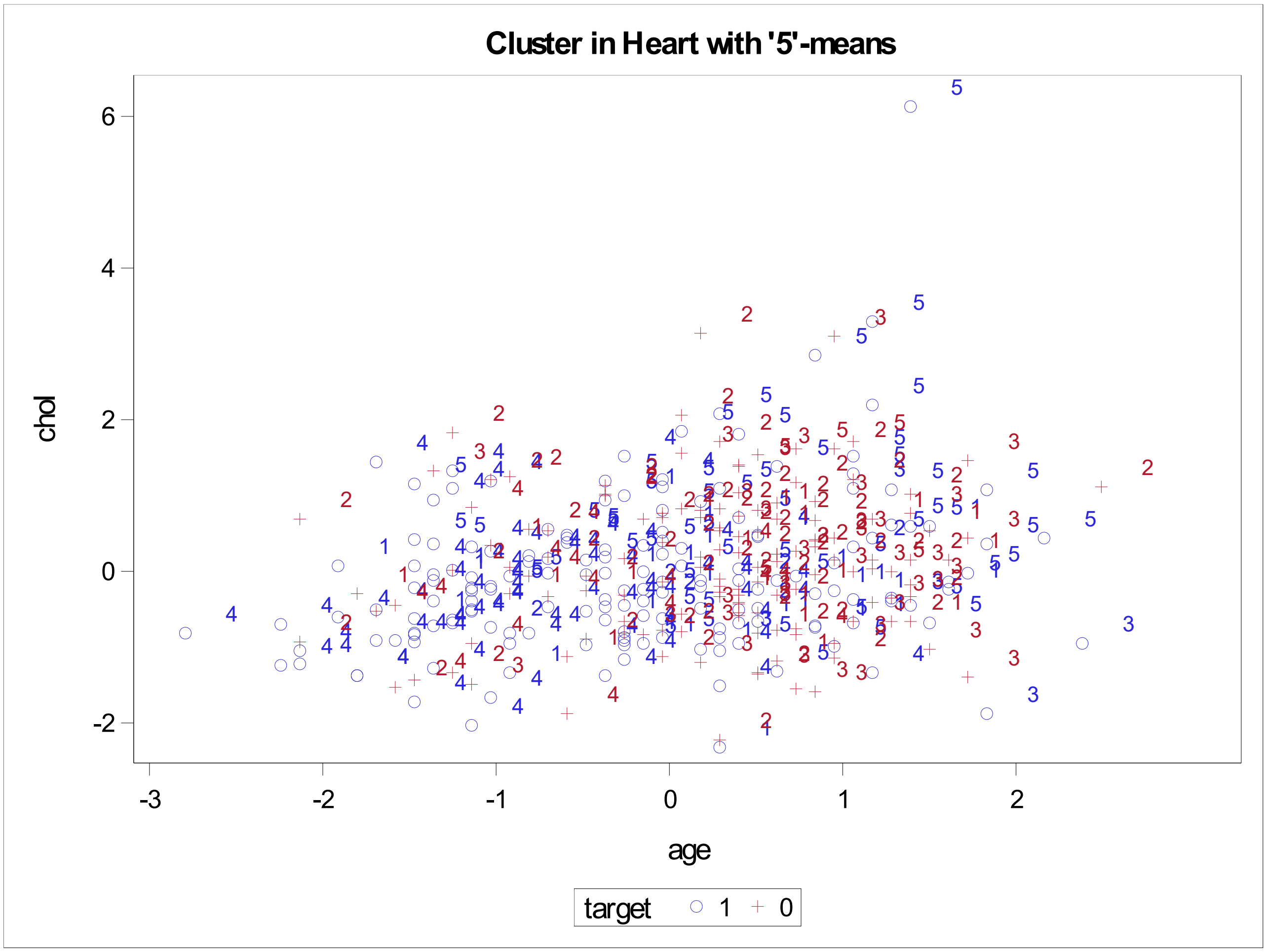
Pseudo F Statistic =	31.03

Observed Over-All R-Squared =	0.29404

Approximate Expected Over-All R-Squared =	0.19065

Cubic Clustering Criterion =	24.430

WARNING: The two values above are invalid for correlated variables.


Cluster in Heart with '5'-means
chol
6
4
2
0
-2
age
-3
-2
-1
0
1
2
target ○ 1 + 0

```
options validvarname=V7;
proc import Datafile="/home/u62339736/Big_data_lab/assignment/heart.csv" out=heart
dbms=csv ;
run;
ods Rtf  file="/home/u62339736/Big_data_lab/assignment/ heart_assignment.rtf"
startpage=no ;
proc contents data=heart;
run;
proc print data=heart(obs=10);
run;
proc means data=heart;
run;
proc stdize data=heart out=heart_std method=std;
  var age sex cp trestbps chol fbs restecg thalach exang oldpeak slope ca thal;
run;
proc print data=heart_std(obs=10);
run;
proc sgplot data=heart;
  scatter x= age   y=chol  /group= target;
run;
proc sgplot data=heart_std;
  scatter x= age   y=chol/group= target ;
run;
%macro doFASTCLUS;
       %do k=2 %to 5;
              title "Cluster in Heart with '&k'-means";

              proc fastclus data=heart_std out=cluster_assignment&k maxiter=100
                            maxclusters=&k
                 summary;
       var age sex cp trestbps chol fbs restecg thalach exang oldpeak slope ca thal;
                 run;

              proc sgplot;
                     scatter x=age y=chol/datalabel=cluster group= target;
              run;

       %end;
%mend;
%doFASTCLUS;
ods Rtf  close;
```