CIND 119: Introduction to Big Data Analytics Assignment 1:

Perform K-Means clustering on a dataset and analyze the results with using SAS

By Stephanie Boissonneault
Oct 4th 2023

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)

To display the contents of the data, I used PROC CONTENTS, and to display the first ten observations I used PROC PRINT.

SAS® Studio

Assignment_1.sas

CODE LOG RESULTS OUTPUT DATA

```sas
/* Generated Code (IMPORT) */
/* Source File: heart.csv */
/* Source Path: /home/u63574148/Assignement 1 */
/* Code generated on: 9/27/23, 8:02 AM */
%web_drop_table(WORK.IMPORT);

/* Questions #1 */
FILENAME REFFILE '/home/u63574148/Assignment 1/heart.csv';

PROC IMPORT DATAFILE=REFFILE
    DBMS=CSV
    OUT=heart;
    GETNAMES=YES;
RUN;

PROC CONTENTS DATA=heart;
RUN;

PROC PRINT DATA=heart(obs=10);
title "Heart Dataset";
RUN;
```

Results from printing the contents:

The CONTENTS Procedure

Data Set Name	WORK.HEART	Observations	303
Member Type	DATA	Variables	14
Engine	V9	Indexes	0
Created	10/03/2023 20:28:26	Observation Length	112
Last Modified	10/03/2023 20:28:26	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_workF8DC0001F6DD_odaws02-usw2-2.oda.sas.com/SAS_work2ABC0001F6DD_odaws02-usw2-2.oda.sas.com/heart.sas7bdat
Release Created	9.0401M7
Host Created	Linux
Inode Number	1946160047
Access Permission	rw-r--r--
Owner Name	u63574148
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.
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.

Results from printing the first ten observations:

Heart 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
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).

SAS® Studio

Assignment_1.sas

CODE LOG RESULTS OUTPUT D

```
/* Questions #2 */

PROC MEANS DATA=heart;
RUN;
```

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

Based on the means procedure, I can note that all variables have been measured on a total sample size of 303. In other words, there are 303 observations for each available. It can be concluded that the sample ranges from age 29 to 77 based on the minimum and maximum recorded values. While the age range is widespread across 9 standard deviations from the mean, the average age of the sample is 54. The average resting blood pressure on admission to the hospital (trestbps) is 131.6 and the average serum cholesterol (chol) is 246. The data for chol (SD = 51.8) is more widespread from the mean than that of trestbps (SD = 17.5). The maximum heart rate achieved (thalach) is spread across 22 standard deviations from the mean and is spread over a range of 131 (Max 202 - Min 71). Sex, cp, fbs, restecg, exang, slope, thal and target are categorical variables whose categories have been assigned a number for the purpose of analysis. If we look at sex, for example, the minimum and maximum values of 0 and 1 are present because the values 0 and 1 have been assigned to label “Female” and “Male”. The target value is assigned 0 or 1 based on whether there is the presence of heart disease.

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

SAS® Studio

*Assignment_1.sas | *Lab 4 iris | Lab_4.sas

CODE | LOG | RESULTS

```
/* Question #3 */

PROC STDIZE method=std out=stan_heart;
var age sex cp trestbps chol fbs restecg thalach exang oldpeak slope ca thal target;
RUN;

PROC PRINT DATA=stan_heart(obs=10);
title "Standardized Heart Dataset";
RUN;

PROC MEANS DATA=stan_heart;
RUN;
```

Standardized Heart Dataset

Obs	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
1	0.9506240215	0.6798805249	1.9698642473	0.7626940758	-0.255910365	2.3904835162	-1.004170712	0.0154172814	-0.695480041	1.0855422911	-2.270822075	-0.713248971	-2.145323783	0.9130188171
2	-1.912149695	0.6798805249	1.0009212815	-0.092584625	0.0720802521	-0.416944799	0.8974775738	1.6307737425	-0.695480041	2.1190672376	-2.270822075	-0.713248971	-0.512074772	0.9130188171
3	-1.471722969	-1.465992382	0.0319783157	-0.092584625	-0.815423771	-0.416944799	-1.004170712	0.9758995015	-0.695480041	0.3103985813	0.9747396642	-0.713248971	-0.512074772	0.9130188171
4	0.1798772518	0.6798805249	0.0319783157	-0.662770426	-0.198029668	-0.416944799	0.8974775738	1.2378491979	-0.695480041	-0.206363892	0.9747396642	-0.713248971	-0.512074772	0.9130188171
5	0.2899839332	-1.465992382	-0.93696465	-0.662770426	2.078611086	-0.416944799	0.8974775738	0.5829749569	1.4331103867	-0.37861805	0.9747396642	-0.713248971	-0.512074772	0.9130188171
6	0.2899839332	0.6798805249	-0.93696465	0.4776011755	-1.046946559	-0.416944799	0.8974775738	-0.071899284	-0.695480041	-0.550872207	-0.648041205	-0.713248971	-2.145323783	0.9130188171
7	0.1798772518	-1.465992382	0.0319783157	0.4776011755	0.9209971433	-0.416944799	-1.004170712	0.1463921296	-0.695480041	0.2242715024	-0.648041205	-0.713248971	-0.512074772	0.9130188171
8	-1.141402925	0.6798805249	0.0319783157	-0.662770426	0.3228966063	-0.416944799	0.8974775738	1.0195577842	-0.695480041	-0.895380523	0.9747396642	-0.713248971	1.1211742386	0.9130188171
9	-0.260549474	0.6798805249	1.0009212815	2.3021957372	-0.911891599	2.3904835162	0.8974775738	0.5393166742	-0.695480041	-0.464745129	0.9747396642	-0.713248971	1.1211742386	0.9130188171
10	0.2899839332	0.6798805249	1.0009212815	1.047786976	-1.509992136	-0.416944799	0.8974775738	1.063216067	-0.695480041	0.482652739	0.9747396642	-0.713248971	-0.512074772	0.9130188171

Standardized Heart Dataset

The MEANS Procedure

Variable	N	Mean	Std Dev	Minimum	Maximum
age	303	1.392359E-17	1.0000000	-2.7930031	2.4921176
sex	303	8.793846E-17	1.0000000	-1.4659924	0.6798805
cp	303	6.35218E-16	1.0000000	-0.9369647	1.9698642
trestbps	303	-7.73858E-16	1.0000000	-2.1452535	3.8987160
chol	303	-9.12361E-17	1.0000000	-2.3203219	6.1302599
fbs	303	7.694615E-17	1.0000000	-0.4169448	2.3904835
restecg	303	1.568236E-16	1.0000000	-1.0041707	2.7991259
thalach	303	-4.76333E-16	1.0000000	-3.4335871	2.2856480
exang	303	1.159322E-15	1.0000000	-0.6954800	1.4331104
oldpeak	303	8.207589E-17	1.0000000	-0.8953805	4.4444984
slope	303	0	1.0000000	-2.2708221	0.9747397
ca	303	5.631725E-16	1.0000000	-0.7132490	3.1983246
thal	303	-5.23967E-17	1.0000000	-3.7785728	1.1211742
target	303	3.971887E-16	1.0000000	-1.0916529	0.9130188

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 withincluster sum of squares). (8 points)

SAS® Studio

*Assignment_1.sas

CODE LOG RESULTS OUTPUT DATA

```
/* Questions #4 */

%macro doFASTCLUS;
    %do k=2 %to 5;
            title'Clusters in HEART with' &k'-Means';

            PROC FASTCLUS
                    data=stan_heart
                    out=clust_solution
                    maxiter=100
                    maxclusters=&k
                    summary;

            PROC SGPLOT data=clust_solution;
                scatter x=age y=chol
                /datalabel=cluster group=target;
            RUN;
    %end;
%MEND;
%doFASTCLUS;

PROC PRINT DATA=clust_solution;
RUN;
```

Clusters in HEART with 2-Means

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

Convergence criterion is satisfied.

Criterion Based on Final Seeds =	0.8997

Cluster Summary

Cluster	Frequency	RMS Std Deviation	Maximum Distance from Seed to Observation	Radius Exceeded	Nearest Cluster	Distance Between Cluster Centroids
1	138	0.9875	6.2728		2	3.7737
2	165	0.8252	7.9654		1	3.7737

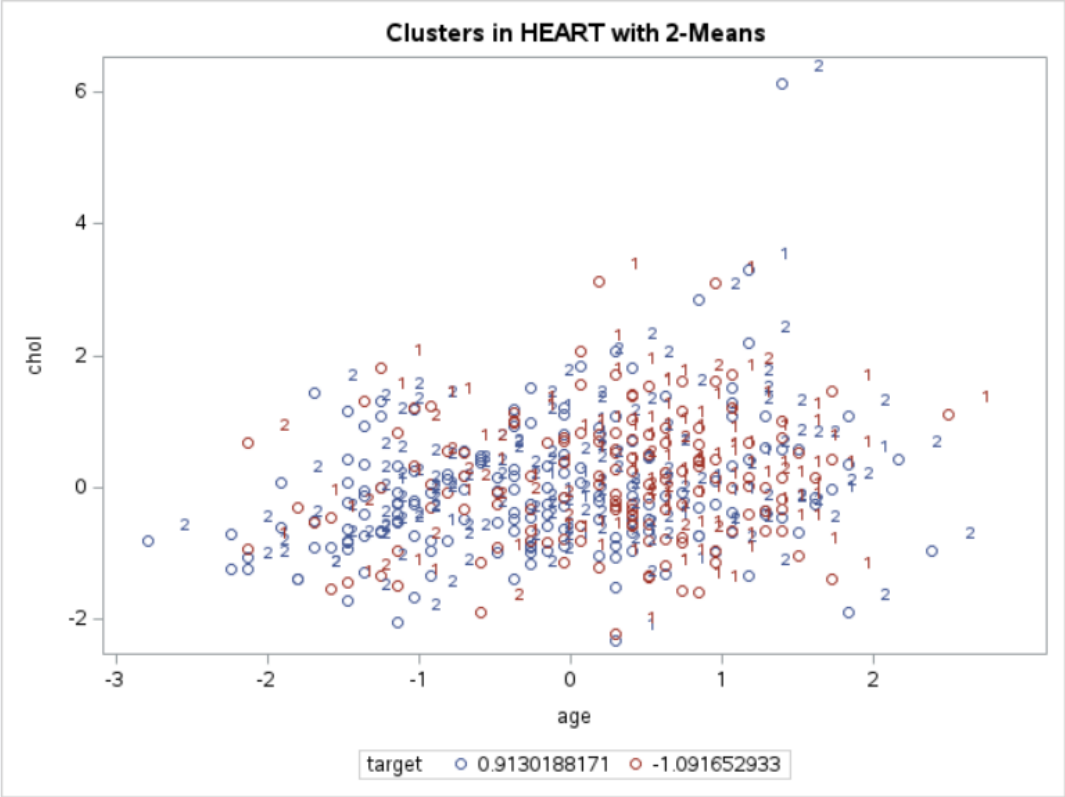
Pseudo F Statistic =	82.08

Observed Over-All R-Squared =	0.21426

Approximate Expected Over-All R-Squared =	0.09086

Cubic Clustering Criterion =	31.506

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



Clusters in HEART with 3-Means

The FASTCLUS Procedure
Replace=FULL Radius=0 Maxclusters=3 Maxiter=100 Converge=0.02

Convergence criterion is satisfied.

Criterion Based on Final Seeds =	0.8652

Cluster Summary						
Cluster	Frequency	RMS Std Deviation	Maximum Distance from Seed to Observation	Radius Exceeded	Nearest Cluster	Distance Between Cluster Centroids
1	123	0.9568	6.2385		3	3.7391
2	70	0.7992	7.3744		3	2.6246
3	110	0.8072	5.2246		2	2.6246

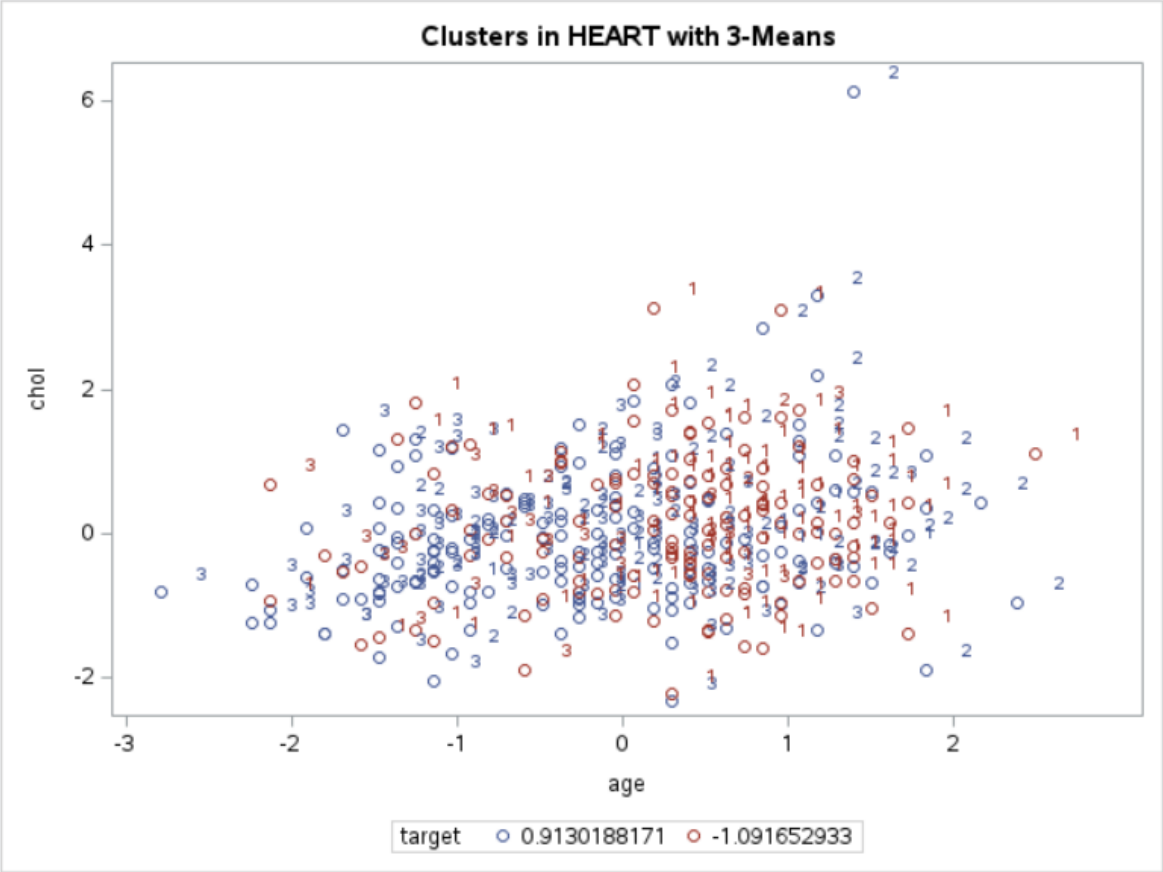
Pseudo F Statistic =	56.48

Observed Over-All R-Squared =	0.27352

Approximate Expected Over-All R-Squared =	0.13020

Cubic Clustering Criterion =	35.859

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



Clusters in HEART with 4-Means

The FASTCLUS Procedure
Replace=FULL Radius=0 Maxclusters=4 Maxiter=100 Converge=0.02

Convergence criterion is satisfied.

Criterion Based on Final Seeds =	0.8666

Cluster Summary

Cluster	Frequency	RMS Std Deviation	Maximum Distance from Seed to Observation	Radius Exceeded	Nearest Cluster	Distance Between Cluster Centroids
1	92	0.9570	6.2092		2	2.7519
2	75	0.9172	5.3965		4	2.6346
3	4	0.9261	4.1236		4	4.6874
4	132	0.7782	5.1664		2	2.6346

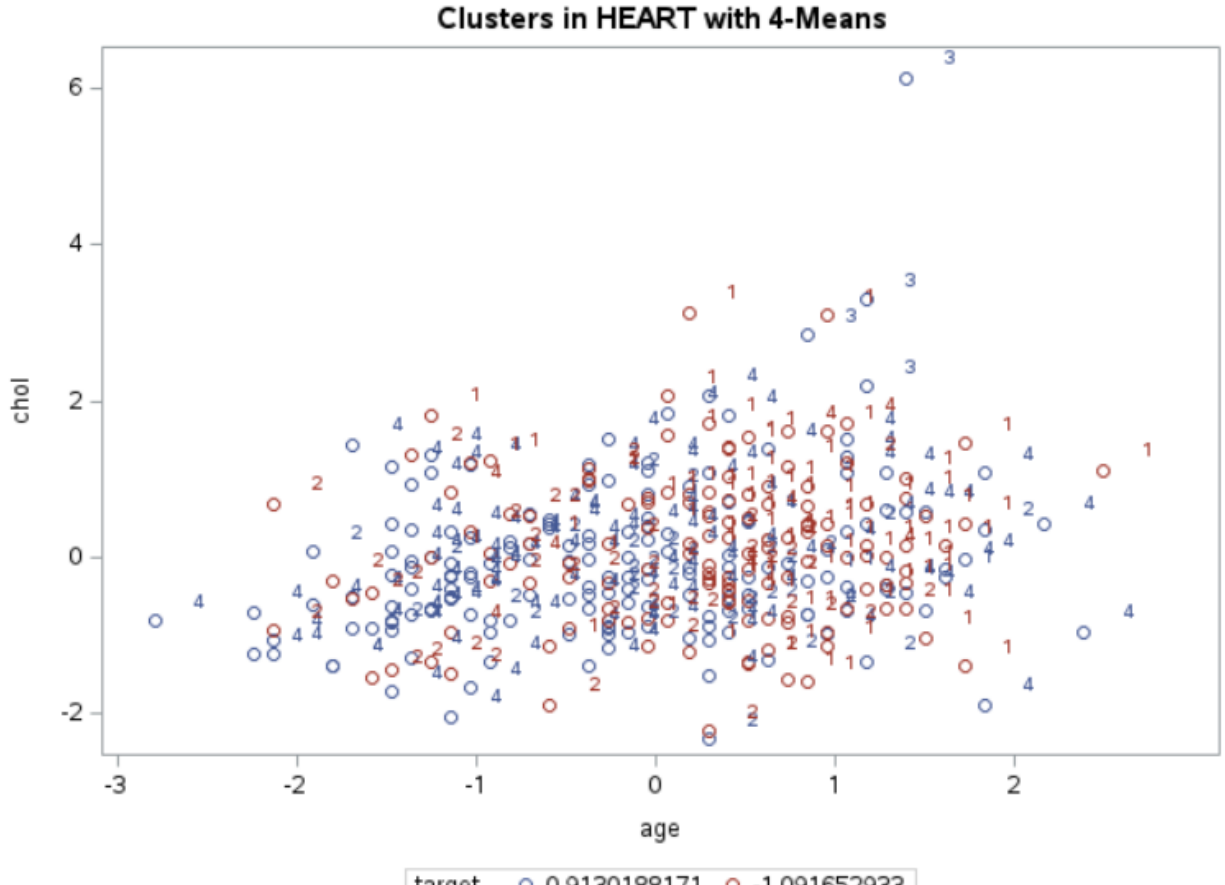
Pseudo F Statistic =	37.09

Observed Over-All R-Squared =	0.27123

Approximate Expected Over-All R-Squared =	0.16224

Cubic Clustering Criterion =	26.155

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



Clusters in HEART with 5-Means

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

Convergence criterion is satisfied.

Criterion Based on Final Seeds =	0.7980

Cluster Summary						
Cluster	Frequency	RMS Std Deviation	Maximum Distance from Seed to Observation	Radius Exceeded	Nearest Cluster	Distance Between Cluster Centroids
1	67	0.7714	7.4513		5	2.5692
2	33	0.8655	4.9115		3	3.7218
3	54	0.8931	5.5185		4	3.0373
4	56	0.8210	6.1686		3	3.0373
5	93	0.7399	4.6849		1	2.5692

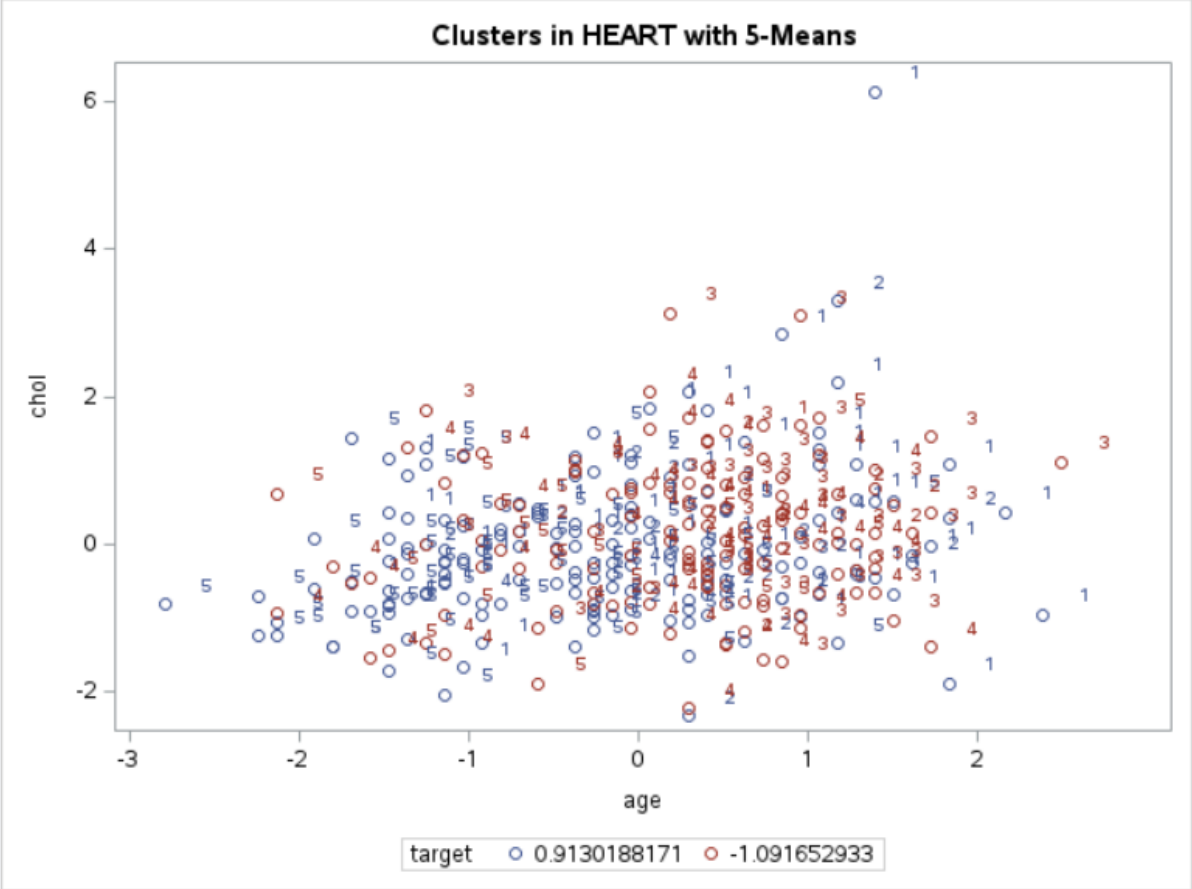
Pseudo F Statistic =	46.03

Observed Over-All R-Squared =	0.38190

Approximate Expected Over-All R-Squared =	0.18979

Cubic Clustering Criterion =	48.640

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



To choose the best k value, I evaluated and compared the average RMS Standard Deviation to select the best K value in clustering. I would select the K value of 2 because it has the highest average RMS Std Deviation, and therefore the best value for K in clustering.

Calculations for finding the average RMS:

K value	**RMS Std Deviations**	**Average RMS Std Deviation**
2	(0.9875 + 0.8252)/2	**0.9063**
3	(0.9568 + 0.7992 + 0.8072)/3	0.8544
4*	(0.9570 + 0.9172 + 0.9261 + 0.7782)/4	0.8946
5	(0.7714 + 0.8655 + 0.8931 + 0.8210 + 0.7399)	0.8182

***Also noted: I noticed that cluster #3 from the K-means 4 table had a frequency of 4. This very low number as opposed to the observations in other clusters may be due to the clustering of outliers as shown in the sgplot. The clusters for the k means 4 is also the least homogeneous

```
/* Generated Code (IMPORT) */
/* Source File: heart.csv */
/* Source Path: /home/u63574148/Assignement 1 */
/* Code generated on: 9/27/23, 8:02 AM */
%web_drop_table(WORK.IMPORT);

/* Questions #1 */
FILENAME REFFILE '/home/u63574148/Assignment 1/heart.csv';

PROC IMPORT DATAFILE=REFFILE
    DBMS=CSV
    OUT=heart;
    GETNAMES=YES;
RUN;

PROC CONTENTS DATA=heart;
RUN;

PROC PRINT DATA=heart(obs=10);
title "Heart Dataset";
RUN;

/* Questions #2 */

PROC MEANS DATA=heart;
RUN;

/* Question #3 */

PROC STDIZE method=std out=stan_heart;
var age sex cp trestbps chol fbs restecg thalach exang oldpeak slope ca thal target;
RUN;

PROC PRINT DATA=stan_heart(obs=10);
title "Standardized Heart Dataset";
RUN;

PROC MEANS DATA=stan_heart;
RUN;

/* Questions #4 */

%macro doFASTCLUS;
    %do k=2 %to 5;
            title'Clusters in HEART with' &k'-Means';

            PROC FASTCLUS
                    data=stan_heart
                    out=clust_solution
                    maxiter=100
                    maxclusters=&k
                    summary;

            PROC SGPLOT data=clust_solution;
                scatter x=age y=chol
                /datalabel=cluster group=target;
            RUN;
    %end;
%MEND;
%doFASTCLUS;

PROC PRINT DATA=clust_solution;
RUN;
```