/\* SAS example code for cluster analysis \*/

/\* PROC CLUSTER performs many hierarchical methods \*/

DATA FOODDATA;

INPUT ObsNum Food $ Energy Protein Fat Calcium Iron;

cards;

1 BB 340 20 28 9 2.6

2 HR 245 21 17 9 2.7

3 BR 420 15 39 7 2.0

4 BS 375 19 32 9 2.5

5 BC 180 22 10 17 3.7

6 CB 115 20 3 8 1.4

7 CC 170 25 7 12 1.5

8 BH 160 26 5 14 5.9

9 LL 265 20 20 9 2.6

10 LS 300 18 25 9 2.3

11 HS 340 20 28 9 2.5

12 PR 340 19 29 9 2.5

13 PS 355 19 30 9 2.4

14 BT 205 18 14 7 2.5

15 VC 185 23 9 9 2.7

16 FB 135 22 4 25 0.6

17 AR 70 11 1 82 6.0

18 AC 45 7 1 74 5.4

19 TC 90 14 2 38 0.8

20 HF 135 16 5 15 0.5

21 MB 200 19 13 5 1.0

22 MC 155 16 9 157 1.8

23 PF 195 16 11 14 1.3

24 SC 120 17 5 159 0.7

25 DC 180 22 9 367 2.5

26 UC 170 25 7 7 1.2

27 RC 110 23 1 98 2.6

;

run;

/\* Calculating std. deviations of each variable \*/

PROC MEANS DATA= FOODDATA STDDEV;

OUTPUT OUT = mysds STDDEV(Energy Protein Fat Calcium Iron)

= sde sdp sdf sdc sdi;

run;

/\* Dividing each variable by its sd \*/

DATA FOODSUM;

IF \_N\_ =1 THEN SET MYSDS;

SET FOODDATA;

run;

DATA FOODSTD;

SET FOODSUM;

Energystd=Energy/sde; Proteinstd=Protein/sdp; Fatstd=Fat/sdf;

Calciumstd=Calcium/sdi; Ironstd = Iron/sdi;

KEEP Energystd Proteinstd Fatstd Calciumstd Ironstd Food;

run;

proc print data = foodstd;

run;

PROC CLUSTER DATA = FOODSTD METHOD = COMPLETE OUTTREE=TREE;

/\* Other options (plus MANY others): \*/

\*PROC CLUSTER DATA = FOODSTD METHOD = AVERAGE;

\*PROC CLUSTER DATA = FOODSTD METHOD = SINGLE;

\*PROC CLUSTER DATA = FOODSTD METHOD = WARD;

ID Food;

run;

PROC TREE horizontal; /\* Omit the word "horizontal" if you want a vertical dendrogram \*/

ID Food;

run;

/\* Printing out the 5-cluster solution \*/

PROC TREE noprint out=part nclusters=5;

ID Food;

run;

PROC SORT;

BY cluster;

run;

proc print;

ID Food;

BY cluster;

run;

/\* Standardization by using z-scores can be done automatically, for example: \*/

PROC CLUSTER DATA = FOODDATA METHOD = COMPLETE STANDARD;

run;

/\*\*\*\*\*\*\*\*\*\* Cars example \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

DATA carsdata;

INPUT Carname $19. mpg disp hp drat wt qsec;

cards;

Mazda RX4 21.0 160.0 110 3.90 2.620 16.46

Mazda RX4 Wag 21.0 160.0 110 3.90 2.875 17.02

Datsun 710 22.8 108.0 93 3.85 2.320 18.61

Hornet 4 Drive 21.4 258.0 110 3.08 3.215 19.44

Hornet Sportabout 18.7 360.0 175 3.15 3.440 17.02

Valiant 18.1 225.0 105 2.76 3.460 20.22

Duster 360 14.3 360.0 245 3.21 3.570 15.84

Merc 240D 24.4 146.7 62 3.69 3.190 20.00

Merc 230 22.8 140.8 95 3.92 3.150 22.90

Merc 280 19.2 167.6 123 3.92 3.440 18.30

Merc 280C 17.8 167.6 123 3.92 3.440 18.90

Merc 450SE 16.4 275.8 180 3.07 4.070 17.40

Merc 450SL 17.3 275.8 180 3.07 3.730 17.60

Merc 450SLC 15.2 275.8 180 3.07 3.780 18.00

Cadillac Fleetwood 10.4 472.0 205 2.93 5.250 17.98

Lincoln Continental 10.4 460.0 215 3.00 5.424 17.82

Chrysler Imperial 14.7 440.0 230 3.23 5.345 17.42

Fiat 128 32.4 78.7 66 4.08 2.200 19.47

Honda Civic 30.4 75.7 52 4.93 1.615 18.52

Toyota Corolla 33.9 71.1 65 4.22 1.835 19.90

Toyota Corona 21.5 120.1 97 3.70 2.465 20.01

Dodge Challenger 15.5 318.0 150 2.76 3.520 16.87

AMC Javelin 15.2 304.0 150 3.15 3.435 17.30

Camaro Z28 13.3 350.0 245 3.73 3.840 15.41

Pontiac Firebird 19.2 400.0 175 3.08 3.845 17.05

Fiat X1-9 27.3 79.0 66 4.08 1.935 18.90

Porsche 914-2 26.0 120.3 91 4.43 2.140 16.70

Lotus Europa 30.4 95.1 113 3.77 1.513 16.90

Ford Pantera L 15.8 351.0 264 4.22 3.170 14.50

Ferrari Dino 19.7 145.0 175 3.62 2.770 15.50

Maserati Bora 15.0 301.0 335 3.54 3.570 14.60

Volvo 142E 21.4 121.0 109 4.11 2.780 18.60

;

run;

/\* Calculating ranges of each variable \*/

PROC MEANS DATA= CARSDATA RANGE;

OUTPUT OUT = myranges RANGE(mpg disp hp drat wt qsec)

= rmpg rdi rhp rdr rwt rqs;

run;

/\* Dividing each variable by its range \*/

DATA carssum;

IF \_N\_ =1 THEN SET MYRANGES;

SET carsdata;

run;

DATA carsstd;

SET carssum;

mpgstd=mpg/rmpg; dispstd=disp/rdi; hpstd=hp/rhp;

dratstd=drat/rdr; wtstd=wt/rwt; qsecstd=qsec/rqs;

KEEP mpgstd dispstd hpstd dratstd wtstd qsecstd Carname;

run;

proc print data = carsstd;

run;

/\* Single Linkage \*/

PROC CLUSTER DATA = CARSSTD METHOD = SINGLE OUTTREE=TREE;

ID Carname;

run;

PROC TREE horizontal; /\* Omit the word "horizontal" if you want a vertical dendrogram \*/

ID Carname;

run;

/\* Average Linkage \*/

PROC CLUSTER DATA = CARSSTD METHOD = AVERAGE OUTTREE=TREE;

ID Carname;

run;

PROC TREE horizontal; /\* Omit the word "horizontal" if you want a vertical dendrogram \*/

ID Carname;

run;

/\* Printing out the 2-cluster solution for average linkage\*/

PROC TREE noprint out=part nclusters=2;

ID Carname;

run;

PROC SORT;

BY cluster;

run;

proc print;

ID Carname;

BY cluster;

run;