Technology Arts Sciences Cologne
Faculty of Economics, Business and Law
Prof. Dr. Arrenberg
Room 221, Tel. 39 14
jutta.arrenberg@th-koeln.de

Exercises Quantitative Methods

Worksheet: Cluster Analysis

Exercise 12.1 (Martens, Seite 229, Statistische Datenanalyse mit SPSS für Windows)
We want to cluster 40 vehicles due to the values of the variables speed (Geschwindigkeit km/h), cubic capacity (Hubraum in ccm), luggage compartment (Kofferraum in Liter), output (Leistung in kW) and payload (Zuladung in kg).

The greatest jump in the dendrogram provides five cluster. The cluster centers are:

Final Cluster Center

	Cluster				
	1	2	3	4	5
Hubraum	2 723	3 350	1 557	2 385	4 803
Geschwindigkeit	189	246	177	175	204
Kofferraum	3 620	368	638	1 850	1 753
Leistung	113	201	69	94	196
Zuladung	708	425	432	633	579

The memberships are:

Cluster 1 Renault Espace V6; Chrysler Voyager

Cluster 2 Alfa 156 2,5 V6; Alfa Spider 3,0 V6; Audi S8 Allrad; Bentley Arnage; BMW Z3 Coupé; Chrysler 300 M; Ferrari F355; Jaguar XJ 8 3,2; Maserati Quattroporte; Porsche 911 Carrera; Mercedes S 300

Cluster 3 Audi A4 1,8; Citrœn Saxo 1,1i X; Fiat Punto 60 SX; Fiat Coupé 1,8 16V; Ford Ka 1,3; Honda Civic 1,4i; Mazda 323 F 1,5; Mazda MX5; Smart; Opel Corsa 1.0; Opel Vectra CDX; Peugeot 206; Peugeot 406 SV 2,0; Renault Twingo; Volvo C 70 T5; VW Polo SDL; VW Beetle

Cluster 4 Citrœn Xantia 1,9 T; Ford Galaxy 2,3 GLX; Honda Shuttle 2,3i; Mercedes C 220 T; Mitsubishi Pajero; Nissan Patrol 2,8

Cluster 5 BMW 750i; Jeep Grand Cherokee; Puch G 500; Toyota Landcruiser

Exercise 12.2
Please open the file Pisa_00_03_06_09_12.sav

We want to cluster the participating countries of the PISA survey 2012 due to a cluster analysis:

1. Hierarchical cluster analysis (no fixed number of clusters)
2. Hierarchical cluster analysis with fixed number of clusters
   The greatest jump in the dendrogram provides the number of clusters.
3. K-means Cluster analysis

Exercise 12.3

a) What are the levels of a variable in a hierarchical cluster analysis?

b) What are the levels of a variable in a k-means cluster analysis?

c) A telephone provider wants to cluster the clients into groups due to the values of the following variables of the last month:

   - Calls in the provider network
   - National calls not in the provider network
   - International calls
   - Short messages services in the provider network
   - National short messages services not in the provider network
   - International short messages services

   The values of the variables are listed in the file *Telefon.sav*.

   1. Please run a hierarchical cluster analysis for all six variables. How do you get the number of clusters?
   2. Please run a k-means cluster analysis for all six variables. The number of clusters is the finding of c.1). What are the final cluster centers? Please comment the clusters.

Exercise 12.4

The values of the variables tax per year, mileage in liter pro 100 km, cubic capacity in ccm, horsepower, exhaust fumes in g/km and price in Euro are listed in the file *auto.sav*. The observed vehicles are:

Vehicle	Cluster membership
Audi S4 Avant A8	
Opel Astra 1.5 Turbo	
VW Golf GT 1.4 TSI	
Ford Mondeo 2.0 TDCi	
Mazda 6 2.0 CD	
Opel Vectra 1.9 CDTI	
Peugeot 407 HDI	
Toyota Avensis 2.0 D-4D	
VW Passat 1.9 TDI	
BMW 530i	
Mercedes E350	
BMW X5 4.4i	
Porsche Cayenne S	
Audi Q7 4.2 FSI	
Daewoo Kalos 1.4 SE	
Hyundai Getz 1.3 GLS	
Ford Fiesta 1.4 i	
Daihatsu cuore	
Fiat Panda	
Daihatsu Sirion	
Nissan Micra	
Suzuki Swift	
Ford Mondeo ST220	
Opel vectra GTS	

We want to cluster the vehicles:

a) Please delete the vehicle Porsche 911 Carrera 4S and run hierarchical cluster analysis.

   1. How to determine the number of clusters?
   2. How many clusters?

b) Please delete the vehicle Porsche 911 Carrera 4S and run a k-means cluster analysis. The number of clusters is the result of a).

   1. Please plot the cluster membership of every vehicle in the table above.
   2. Please comment the p-values in the ANOVA-table.

Exercise 12.5

Please open the file *Kriminalität.sav*. Delete the case number 9 "'District of Columbia"'.

a) Run a hierarchical cluster analysis with all variables. What is the number of clusters?

b) Run a k-means cluster analysis with the number of clusters under a). Commet the clusters.

c) Plot a scatterplot of all cases with the first principal component as the x-axis and the second principal component as the y-axis. Mark the clusters.

Technology Arts Sciences Cologne
Faculty of Economics, Business and Law
Prof. Dr. Arrenberg
Room 221, Tel. 39 14
jutta.arrenberg@th-koeln.de

Exercises Quantitative Methods

Mock Exam

Exercise

Please open the file *1991 U.S. General Society Survey.sav*. Run a K-means cluster analysis with the variables

- Letztes abgeschlossenes Schuljahr
- Letztes abgeschlossenes Schuljahr des Vaters
- Letztes abgeschlossenes Schuljahr der Mutter
- Letztes abgeschlossenes Schuljahr des Partners

where three clusters are generated.

a) Please show the average value of the four considered variables among the three clusters:

Average Value

	Cluster		
	1	2	3
Letztes abgeschlossenes Schuljahr			
Letztes abgeschlossenes Schuljahr des Vaters			
Letztes abgeschlossenes Schuljahr der Mutter			
Letztes abgeschlossenes Schuljahr des Partners			

b) Please comment the three clusters.

c) Please understand the variable "'Allgemeine Zufriedenheit"' as a scale leveled variable. Are there significant differences of the degree of satisfaction across the three clusters? Please run a test to show this.

Technology Arts Sciences Cologne
Faculty of Economics, Business and Law
Prof. Dr. Arrenberg
Room 221, Tel. 39 14
jutta.arrenberg@th-koeln.de

Exercises Quantitative Methods

Problem:
How to recode a scale leveled variable into an ordinal leveled variable with categories of about the same sizes?

a) Manual Recoding

1. Please open the file *German_credit.sav*
   We want to group the variable "Duration in months" into four classes of about the same number of cases in each group/class.
2. Analyze → Descriptive Statistics → Frequencies ...
3. Variable(s)= "Duration in months"
   Click "ok"
   In the output table of SPSS please look in the column "Cumulative Percent" for the values 0.25 and 0.50 and 0.75. The associated values are 11 months (with cumulative percent 18.0), 18 months (with cumulative percent 54.6) and 24 months (with cumulative percent 77.0)
4. Now recode the variable "Duration in months" with
   Transform → Recode into different variables ...
   into an ordinal leveled variable with the four groups:
   1. class: up to 11 months
   2. class: over 11 up to 18 months
   3. class: over 18 up to 24 months
   4. class: more than 24 months

   The number of the cases are:
   1. class: 180 cases
   2. class: 366 cases
   3. class: 224 cases
   4. class: 230 cases.

b) Automatic Recoding

1. Please open the file *German_credit.sav*
   We want to group the variable "Duration in months" into four classes of about the same number of cases in each group/class.
2. Transform → Rank Cases ...
3. Variable(s)= "Duration in months"
4. Click "Rank Types ..."

5. Select "Ntiles = 4"
   Click "Continue"
6. Click "ok"
   The associated numbers of the class of each case are listed in the input table in the column "Nduration (label: Percentile Group of Duration)".
   The number of the cases are:
   1. class: 180 cases
   2. class: 366 cases
   3. class: 224 cases
   4. class: 230 cases.

Technology Arts Sciences Cologne
Faculty of Economics, Business and Law
Prof. Dr. Arrenberg
Room 221, Tel. 39 14
jutta.arrenberg@th-koeln.de

Exercises Quantitative Methods

Problem:
How to select cases

a) of a numeric variable?

1. Please open the file *1991 US.sav*
2. Data → Select Cases
3. Select "If condition is satisfied"
   Click "If ..."
   Gender = 1
   Continue
4. ok

b) of a string variable?
With quotation marks

1. Please open the file *Dauer_Museum.sav*
2. Data → Select Cases
3. Select "If condition is satisfied"
   Click "If ..."
   Museum = "ML"
   Continue
4. ok

Technology Arts Sciences Cologne
Faculty of Economics, Business Administration and Law
Prof. Dr. Arrenberg
Room 221, Tel. 39 14
jutta.arrenberg@th-koeln.de

Exercises Quantitative Methods

Worksheet: Recall

Example (c.f. Anderson et al. page 742)
Three college admission test preparation programmes are being evaluated. The scores obtained by a sample of 20 people who used the test preparation programmes provided the following data:

Programme		
A	B	C
540	450	600
400	540	630
490	400	580
530	410	490
490	480	590
610	370	620
	550	570

Use a test to determine whether there is a significant difference among the three preparation programmes.

Example (c.f. Anderson et al. page 742)
Condé Nast Traveler Magazine conducts an annual survey of its readers in order to rate the top 80 cruise ships in the world. With 100 the highest possible rating, the overall ratings for a sample of ships from Holland America, Princess and Royal Caribbean cruise lines are shown here:

Holland America		Princess		Royal Caribbean	
Ship	Rating	Ship	Rating	Ship	Rating
Amsterdam	84.5	Coral	85.1	Adventure	84.8
Maasdam	81.4	Dawn	79.0	Jewel	81.8
Ooterdam	84.0	Island	83.9	Mariner	84.0
Volendam	78.5	Princess	81.1	Navigator	85.9
Westerdam	80.9	Star	83.7	Serenade	87.4

Use a test to determine whether the overall ratings among the three cruise lines differ significantly.

Exercise 12.1

Final Cluster Center

	Cluster				
	1	2	3	4	5
Cubic Capacity	2 723	3 350	1 557	2 385	4 803
Top Speed	189	246	177	175	204
Luggage Compartment	3 620	368	638	1 850	1 753
Horsepower	113	201	69	94	196
Payload	708	425	432	633	579

Cluster 1: Giants with four wheels

Cluster 2: Runabouts

Cluster 3: Inefficient vehicles

Cluster 4: Worm gear

Cluster 5: Gasoline guzzler