Kapitel VII

Clusteranalyse

Lösungen

A. Rechenaufgaben

B. SPSS-Aufgaben

1.

1. (Standardisierung der Variablen *ccm, kW* und *Preis*):

Analysieren -> Deskriptive Statistiken -> Deskriptiv…

Übertragen Sie die Variablen *ccm, kW* und *Preis* in das Feld *„Variable(n)"*.

Markieren Sie das Feld „*Standardisierte Werte als Variable speichern*" .

OK

Fahrzeug	zccm	zkW	zPreis
Audi A4	-0.0690	0.0948	0.2663
BMW 320i	0.5145	0.9543	0.6505
Citroen C5	0.1453	0.2134	-0.1220
Ford Mondeo	-0.2769	-0.3497	-0.2565
Jaguar S-Type	1.2062	1.6062	1.4464
Mercedes E 240	1.4184	1.1024	1.5941
Opel Corsa	-1.5478	-1.1202	-1.2787
Peugeot 307	0.1453	-0.7942	-0.7275
Renault Clio	-1.6539	-1.4758	-1.2286
VW Golf	0.1178	-0.2312	-0.3439

2.

- *Analysieren -> Klassifizieren -> Clusterzentrenanalyse…*
- Übertragen Sie die die standardisierten Variablen *zccm, zkW* und *zPreis* in das Feld „*Variablen)*" und die Variable *fahrzeug* in das Feld „*Fallbeschriftung*".
- Wählen Sie als Anzahl der Cluster den Wert 3.

3.

Wahlen Sie „*Iterieren*"
(Die Anzahl der Iterationen ist bei *SPSS* laut Vorgabe 10 begrenzt, jedoch kann dieser Wert erweiter werden.). Wir bleiben bei 10.

Weiter

4.

- *Speichern*
- Markieren Sie „*Cluster-Zugehörigkeit*"
- *Weiter*

5.

- *Optionen*

- Markieren Sie „*Anfängliche Clusterzentren*“, *ANOVA-Tabelle*„“ und „*Cluster-Informationen für jeden Fall*“

- *Weiter, OK.*

Output und Interpretation der Ergebnisse:

- Die Tabelle „***Anfängliche Clusterzentren***“ zeigt diejenigen Gruppencentroide, die *SPSS* nach einer Durchsicht der Daten als sinnvolle Startwerte für eine Zuordnung der anderen Beobachtungen auswählt. Alle weiteren Beobachtungen werden aufgrund ihrer Lage zu diesen Gruppenmittelwerten einem dieser Gruppenmittelwerte zugeordnet. Auf diese Weise entstehen im weiteren Ablauf Cluster.

Anfängliche Clusterzentren

	Cluster		
	1	2	3
Z-Wert: Kilowatt	1,41837	-1,65389	,11776
Z-Wert: Hubraum	1,10244	-1,47584	-,23116
Z-Wert(Preis)	1,59411	-1,22860	-,34393

- Das Iterationsprotokoll meldet bereits in der Tabelle „***Iterationsprotokoll*** “nach der 2. Iteration eine Konvergenz des Algorithmus. Das bedeutet, dass von den anfänglich zehn vorgegebenen Iterationen nur zwei benötigt werden, um eine Lösung zu produzieren:

Iterationsprotokoll[a]

	Änderung in Clusterzentren		
Iteration	1	2	3
1	,534	,187	,151
2	,000	,000	,000

a. Konvergenz wurde aufgrund geringer oder keiner Änderungen der Clusterzentren erreicht. Die maximale Änderung der absoluten Koordinaten für jedes Zentrum ist ,000. Die aktuelle Iteration lautet 2. Der Mindestabstand zwischen den anfänglichen Zentren beträgt 2,339.

Umgruppierungen der einzelnen Objekte würde zu keiner nennenswerten Verbesserungen mehr führen.

- Die jetzt entstandenen Cluster haben einen neuen Gruppencentroid, der in der Tabelle „***Clusterzentren der endgültigen Lösung***“ für jede Cluster und jede Variable ausgegeben wird:

Clusterzentren der endgültigen Lösung

	Cluster		
	1	2	3
Z-Wert: Kilowatt	1,04637	-1,60084	,01252
Z-Wert: Hubraum	1,22098	-1,29803	-,21337
Z-Wert(Preis)	1,23032	-1,25365	-,23673

- Die Clusterung für jeden einzelnen Fall zeigt die Tabelle “***Cluster-Zugehörigkeit***“ in zusammengefasster Form:

Cluster-Zugehörigkeit

Fallnummer	Fahrzeugtyp	Cluster	Distanz
1	Audi A4	3	,596
2	BMW 320i	1	,831
3	Citroen C5	3	,461
4	Ford Mondeo	3	,321
5	Jaguar S-Type	1	,470
6	Mercedes E 240	1	,534
7	Opel Corsa	2	,187
8	Peugeot 307	3	,772
9	Renault Clio	2	,187
10	VW Golf	3	,151

- Um die Gruppierung zu überprüfen, ist in der Prozedur *Clusterzentrenanalyse* eine Varianzanalyse (*ANOVA*) integriert, die die Cluster als abhängige Variable, und die einzelnen Variablen als unabhängige (metrische) Variable betrachtet. Ziel ist zu untersuchen, ob sich die gefundenen Cluster signifikant voneinander unterscheiden (siehe Tabelle „***ANOVA***“):

ANOVA

	Cluster		Fehler			
	Mittel der Quadrate	df	Mittel der Quadrate	df	F	Sig.
Z-Wert: Kilowatt	4,205	2	,084	7	49,964	,000
Z-Wert: Hubraum	4,035	2	,133	7	30,362	,000
Z-Wert(Preis)	3,982	2	,148	7	26,922	,001

Die F-Tests sollten nur für beschreibende Zwecke verwendet werden, da die Cluster so gewählt wurden, daß die Differenzen zwischen Fällen in unterschiedlichen Clustern maximiert werden. Dabei werden die beobachteten Signifikanzniveaus nicht korrigiert und können daher nicht als Tests für die Hypothese der Gleichheit der Clustermittelwerte interpretiert werden.

Für die Interpretation dieser Tabelle gilt die gleiche Logik, wie für alle behandelten Testverfahren in *SPSS*. Hier wurde die Cluster als Gruppen betrachtet.
In der Tabelle wird die gesamte Streuung in einen erklärten Teil (zwischen den Clustern) und in einen unerklärten Teil aufgeteilt (in den Clustern). Der Quotient aus diesen beiden Größen ergibt schließlich die Prüfgröße:

$$F_{prüf} = \frac{\text{Varianz zwischen den Clustern}}{\text{Varianz in den Clustern}}.$$

Entscheidend ist hier wieder, dass der Wert in der Spalte *Sig* jeweils Null oder nur unwesentlich von Null verschieden ist. Dann wird angenommen, dass sich die ermittelten Cluster signifikant unterscheiden.

- In der Tabelle „***Anzahl der Fälle in jedem Cluster***“ erschienen die Häufigkeiten der Fälle in jedem der 3 Cluster.

Anzahl der Fälle in jedem Cluster

Cluster	1	3,000
	2	2,000
	3	5,000
Gültig		10,000
Fehlend		,000

(Letzte Aktualisierung 12.03.2012)