LÖSUNG 11

a.

- Nach dem Öffnen der Datei MARGARINE.SAV öffnet man mit "Analysieren", "Korrelation", "Distanzen…" die Dialogbox "Distanzen". Die 10 Eigenschaftsvariablen (Streichfähigkeit, Haltbarkeit etc.) der Margarinesorten werden in das Eingabefeld "Variablen:" übertragen. Möchte man die Marken in den Ergebnisausgaben mit den Variablennamen beschriften, sollte man die Variable MARKE in das Eingabefeld "Fallbeschriftung" übertragen. Hier wird darauf verzichtet, um die Ausgabetabellen und die Grafiken weniger umfangreich zu erhalten. Die Marken werden in diesem Falle mit ihren Fallnummern angezeigt. Die Auswahl "Zwischen den Fällen" ist im Feld "Distanzen berechnen" voreingestellt und wird so belassen. Der Maßtyp "Unähnlichkeiten" ist ebenfalls im Feld „Maß“ voreingestellt und wird so belassen, da die quadrierten Euklidischen Distanzen zu diesem Typ gehören. Klicken auf die Schaltfläche "Maße…" öffnet die Unterdialogbox "Distanzen: Unähnlichkeitsmaße". Hier kann das zu berechnende Unähnlichkeitsmaß ausgewählt werden. Im Vorliegenden Fall ist der Optionsschalter „Intervall" zutreffend. In den dazugehörigen Drop-Down-Menüs mit Unähnlichkeitsmaßen wird "Quadrierte euklidische Distanz" gewählt. Ausführen mit „Weiter“ und „OK“.

- Als Ausgabe erscheint eine Distanz-Matrix („Näherungsmatrix“ genannt, wohl ein Übersetzungsfehler) mit den berechneten Distanzmaßen zwischen den 11 Margarinemarken. Jede Distanz zwischen zwei Margarinepaaren ist doppelt aufgeführt: einmal oberhalb und einmal unterhalb der Diagonalen (zur Berechnung der Distanzen s. Kapitel 17.3).

- Die kleinste Distanz (d. h. die größte Ähnlichkeit hinsichtlich der betrachteten Merkmale) besteht mit 1,099 zwischen Marke 3 (die Fallnummer 3 entspricht der Marke SB) und Marke 9 (Botteram). Die größte Distanz (d. h. die größte Unähnlichkeit) besteht mit 38,621 zwischen Marke 6 (Weihnachtsbutter) und Marke 2 (Homa).

Näherungsmatrix

	Quadrierte Euklidische Distanz										
	1	2	3	4	5	6	7	8	9	10	11
1	,000	3,792	3,794	15,198	21,442	25,484	4,882	6,025	2,268	2,909	2,113
2	3,792	,000	6,322	23,871	30,458	38,621	10,881	8,063	5,325	6,194	3,396
3	3,794	6,322	,000	14,151	24,971	28,933	3,998	3,471	1,099	2,361	1,725
4	15,198	23,871	14,151	,000	6,496	11,882	11,692	18,362	15,929	16,520	17,030
5	21,442	30,458	24,971	6,496	,000	3,606	16,410	26,957	25,334	25,906	26,768
6	25,484	38,621	28,933	11,882	3,606	,000	15,887	32,336	29,999	28,195	32,272
7	4,882	10,881	3,998	11,692	16,410	15,887	,000	6,422	5,156	3,825	6,932
8	6,025	8,063	3,471	18,362	26,957	32,336	6,422	,000	3,395	6,376	6,022
9	2,268	5,325	1,099	15,929	25,334	29,999	5,156	3,395	,000	1,564	1,118
10	2,909	6,194	2,361	16,520	25,906	28,195	3,825	6,376	1,564	,000	2,152
11	2,113	3,396	1,725	17,030	26,768	32,272	6,932	6,022	1,118	2,152	,000

Dies ist eine Unähnlichkeitsmatrix

b.

- "Analysieren", "Klassifizieren", "Hierarchische Cluster…" öffnet die Dialogbox "Hierarchische Clusteranalyse". Die 10 Variablen (Merkmale der Margarinemarken) werden in das Eingabefeld "Variable(n)" übertragen. Im Feld "Cluster" ist "Fälle" und im Feld "Anzeigen" ist "Statistik" und "Diagramme" voreingestellt. Diese Einstellungen werden belassen. Klicken der Schaltfläche "Diagramme…" öffnet die Unterdialogbox "Hierarchische Clusteranalyse: Diagramme". Durch Markieren des entsprechenden Kontrollkästchens wird ein "Dendrogramm" angefordert.

Klicken der Schaltfläche "Methode" öffnet die Dialogbox "Hierarchische Clusteranalyse: Methode…". Klicken auf den senkrechten Pfeil neben dem Eingabefeld "Cluster-Methode" öffnet eine Liste mit den verfügbaren Clusterverfahren. Gewählt wird "Nächstgelegener Nachbar" (es entspricht dem Single-Linkage-Verfahren). In Feld „Maß“ wird der Optionsschalter "Intervall" markiert und in der zugehörigen Drop-Down-Liste "Quadrierte euklidischer Distanz" gewählt. Ausführen mit „Weiter“ und „OK“.

- Die Ausgabetabelle "Zuordnungsübersicht" zeigt die einzelnen Schritte des schrittweisen Clusterbildungsprozesses. Zusammen mit der bereits oben angeführten Distanzmatrix, kann die Folge der Clusterbildung nachvollzogen werden. Im ersten Schritt werden die Marken 3 (SB) und 9 (Botteram) zu einem Cluster zusammengeführt, da zwischen diesen Marken die kleinste Distanz in Höhe von 1,099 (d.h. die größte Ähnlichkeit) besteht. Dieses Cluster erhält den Namen 3. Diesem neuen Cluster 3 wird Marke 11 (Rama) zugeordnet, weil gemäß des Verfahrens Nächstgelegener Nachbar die Distanz (= 1,118) zwischen der in Cluster 3 enthaltenen Marke 9 (Botteram) und Marke 11 (Rama) am kleinsten ist. Dieses neue Cluster behält den Namen 3. Diesem neuen Cluster wird nun die Marke 10 (Flora) zugeordnet, weil zwischen der in Cluster 3 enthaltenen Marke 9 und Marke 10 die kleinste Distanz (= 1,564) besteht. Als nächstes wird Cluster 3 die Marke 1 (Sanella) zugeordnet: Zwischen der in Cluster 3 enthaltenen Marke 11 und Marke 1 besteht die kleinste Distanz (= 2,2112). Das neue Cluster mit den nun enthaltenen Marken 3, 9, 10, 11 und 1 erhält den Namen 1. Diesem Cluster wird die Marke 8 (Becel) zugeordnet (die kleinste Distanz besteht mit 3,395 zwischen Marke 9 und Marke 8). Anschließend wird diesem Cluster 1 die Marke 2 (Homa) zugeordnet (die kleinste Distanz besteht zwischen Marke 2 und 11 mit 3,396). Da nun die kleinste Distanz (= 3,606) zwischen Marke 5 (Holländische Markenbutter) und 6 (Weihnachtsbutter) besteht, werden diese beiden Marken zu einem neuen Cluster mit dem Namen 5 zusammengeführt. Als nächstes wird dem Cluster 1 die Marke 7 (Du Darfst) zugeführt (die kleinste Distanz besteht mit 3,825 zwischen Marke 7 und Marke 10). Als nächstes wird die Marke 4 (Delicado) dem Cluster 5 (bestehend aus Marke 5 und Marke 6) zugeordnet (Distanz = 6,496). Dieses Cluster erhält den neuen Namen 4. Im letzten Schritt werden schließlich die beiden Cluster 1 (enthält die Marken 1, 2, 3, 7, 8, 9, 10, 11) und 4 (enthält die Marken 4, 5 und 6) zusammengeführt, so dass alle 11 Marken in einem Cluster vereint sind.

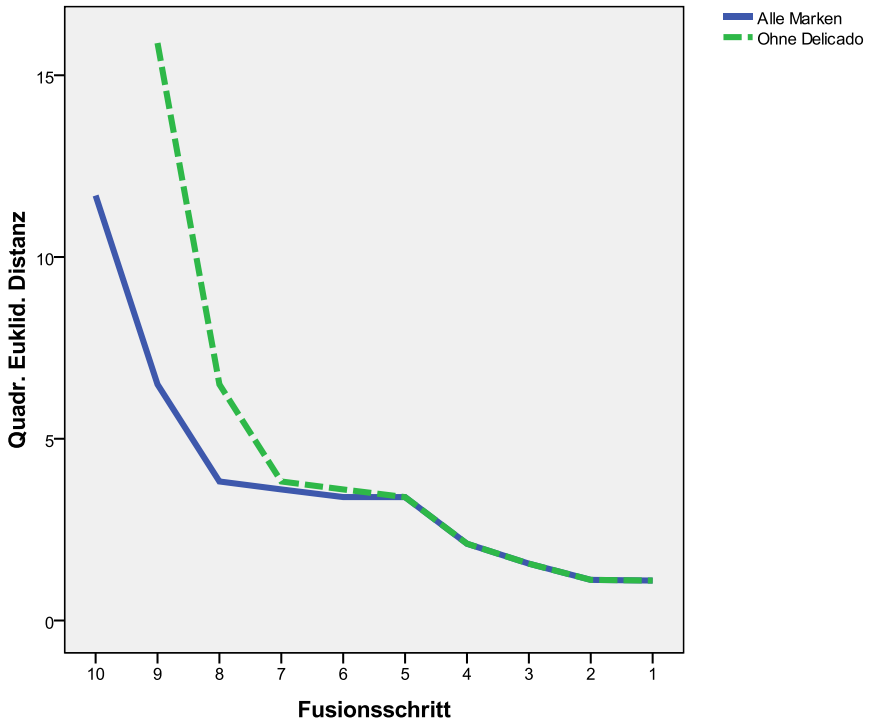
Zuordnungsübersicht

	Zusammengeführte Cluster			Erstes Vorkommen des Clusters		
Schritt	Cluster 1	Cluster 2	Koeffizienten	Cluster 1	Cluster 2	Nächster Schritt
1	3	9	1,099	0	0	2
2	3	11	1,118	1	0	3
3	3	10	1,564	2	0	4
4	1	3	2,113	0	3	5
5	1	8	3,395	4	0	6
6	1	2	3,396	5	0	8
7	5	6	3,606	0	0	9
8	1	7	3,825	6	0	10
9	4	5	6,496	0	7	10
10	1	4	11,692	8	9	0

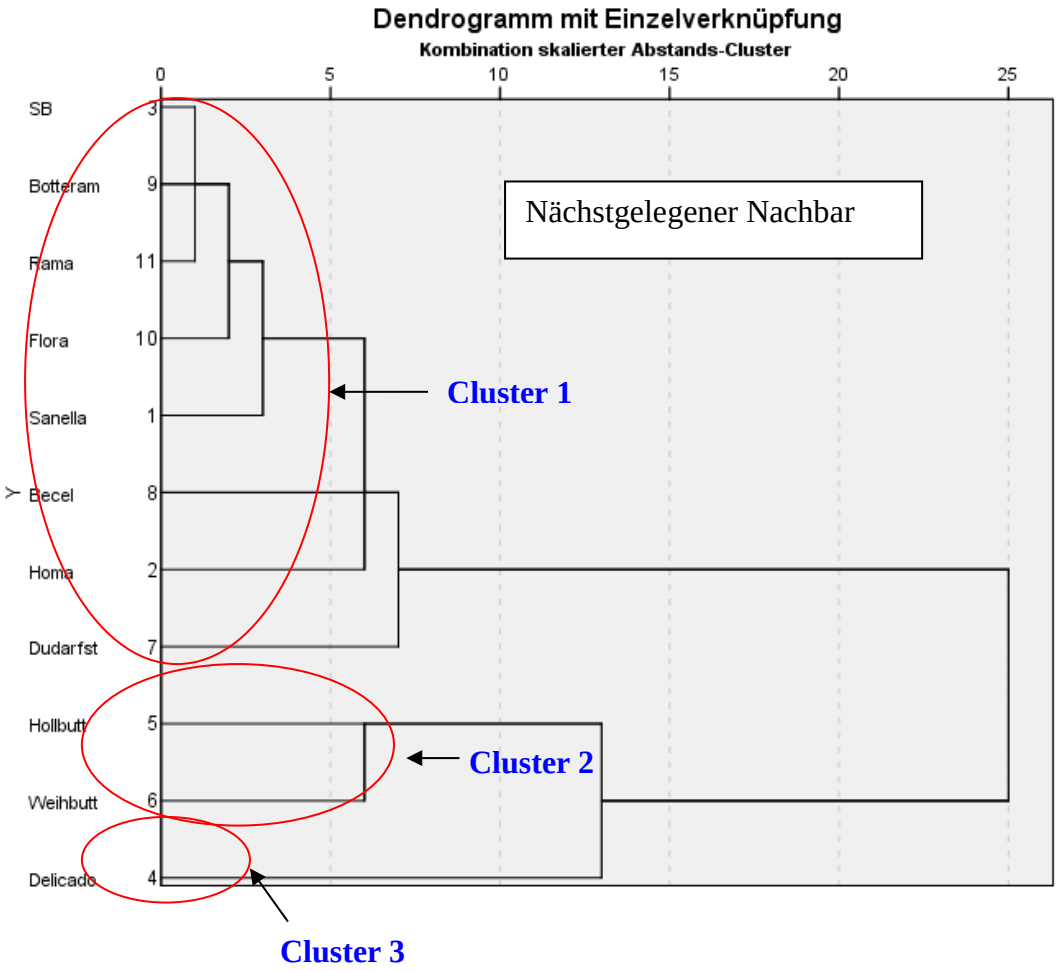
- Zu entscheiden ist, wie viele Cluster als endgültige Lösung für die Clusterung der Butter- und Margarinemarken gewählt werden sollen. Zur Entscheidungshilfe wird der Verlauf des Distanzmaßes ("Koeffizienten" genannt) im Clusterungsprozess herangezogen. Vom 8. auf den 9. Schritt (von 3 auf 2 Cluster) steigt die Distanz sehr kräftig an (von 3,825 auf 6,496).

- Dieses sieht man sehr anschaulich auch in der folgenden Grafik, in der die Distanz der einzelnen Clusterstufen auf der senkrechten Achse und die Fusionsschritte auf der waagerechten Achse abgetragen sind. Vom Übergang von 3 auf 2 Cluster besteht ein Knick im Distanzverlauf (sieht wie ein Ellbogen aus). Nach dem Ellbogen-Kriterium sind 3 Cluster die Lösung. Demnach bilden die Marken 1, 2, 3, 7, 8, 9, 10, die Marken 5 und 6 sowie die Marke 4 ein Cluster. Da ein Objekt kaum als ein Cluster betrachtet werden sollte, plädieren die Autoren Backhaus, Erichson, Plinke und Weiber dafür, die Marke 4 (Delicado) bei der Clusterung der Marken nicht zu berücksichtigen. In dieser Grafik ist der Verlauf der quadrierten Euklidischen Distanz im Fusionsprozess der Clusterbildung zu sehen. Man sieht den Ellenbogen beim 8. Fusionsschritt mit einer Clusterlösung von 3 Cluster.

  Zum Vergleich ist in der Grafik auch der Verlauf der quadrierten Distanz im Fall des unten besprochenen Ausschlusses der Marke Delicado bei der Clusterberechnung zu sehen. Der Ellbogen zeigt sich nun beim 7. Fusionsschritt mit einer Lösung von 2 Cluster.



- Die hierarchischen Clusterbildungsstufen sind auch im Dendrogramm veranschaulicht. Man sieht auch hier, dass die Marke 4 (Delicado) dem aus den Marken 5 (Holländische Markenbutter) und 6 (Weihnachtsbutter) gebildetem Cluster zugeordnet wird. Aus dem Dendrogramm kann man entnehmen, dass die Distanz (Unähnlichkeit) zwischen Marke 4 und dem Cluster, dem sie schließlich zugeordnet wird, relativ groß ist. Dieses stützt die obige Analyse, dass 3 Cluster bestehen und die Marke 4 als Ausreißer zu behandeln ist und sie deshalb bei der Clusterung der Marken nicht einbezogen werden sollte. Die quadrierte Euklidische Distanz der Marke 4 zur Marke 5 beträgt 6,496 und zur Marke 6 11,882. Vergleicht man diese Abstände mit den Distanzen der Marken innerhalb des anderen Clusters (mit den Marken 1, 2, 3, 7, 8, 9, 10 und 11), so sind diese relativ groß, aber anderseits gibt es auch im anderen Cluster zwischen einigen Marken große Distanzen. Die Distanz zwischen Marke 2 und 7 beträgt 10,881, zwischen 2 und 8 8,063, zwischen 7 und 11 6,932.



c.

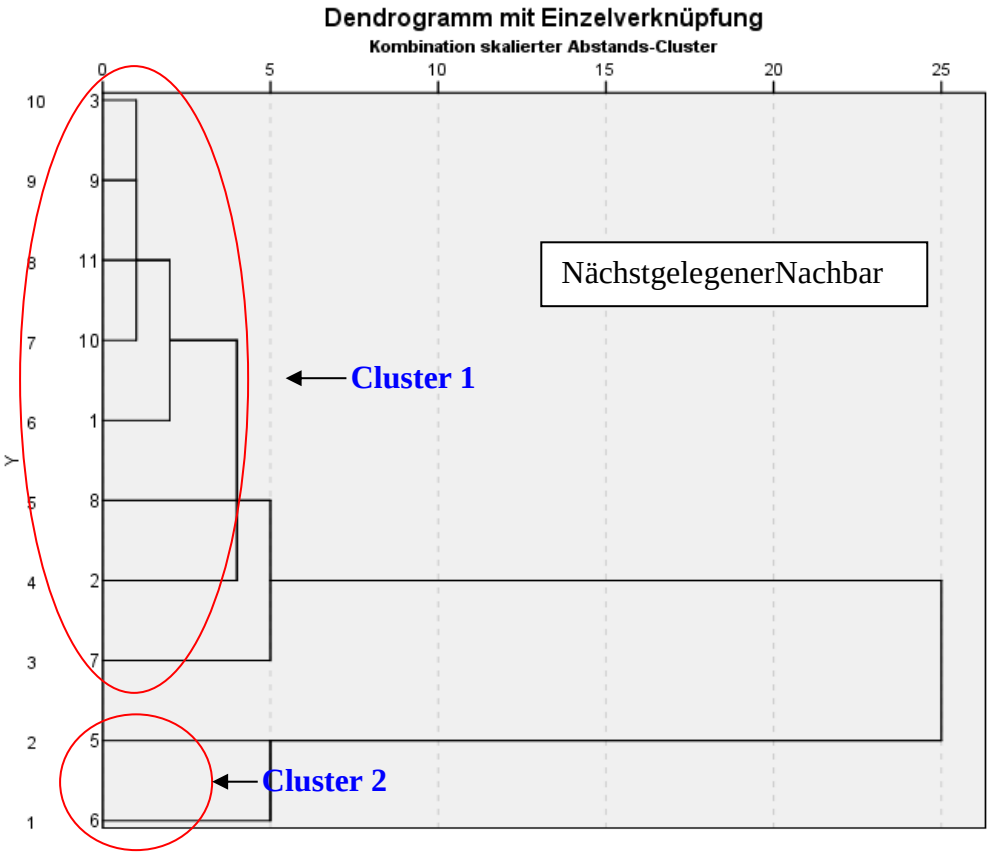
- Die Variable MARKE ist eine Stringvariable. Dies muss bei der Fallauswahl im Menü "Daten", "Fälle auswählen" dadurch berücksichtigt werden, dass der Name in Anführungszeichen gesetzt wird. Um die Marke Delicado aus der Clusterung auszuschließen: „Daten“, „Fälle auswählen…“, „Falls Bedingung zutrifft“, „Falls…“ mit der Bedingung marke ~= "Delicado" (~ ist das Symbol für das logische „nicht“).

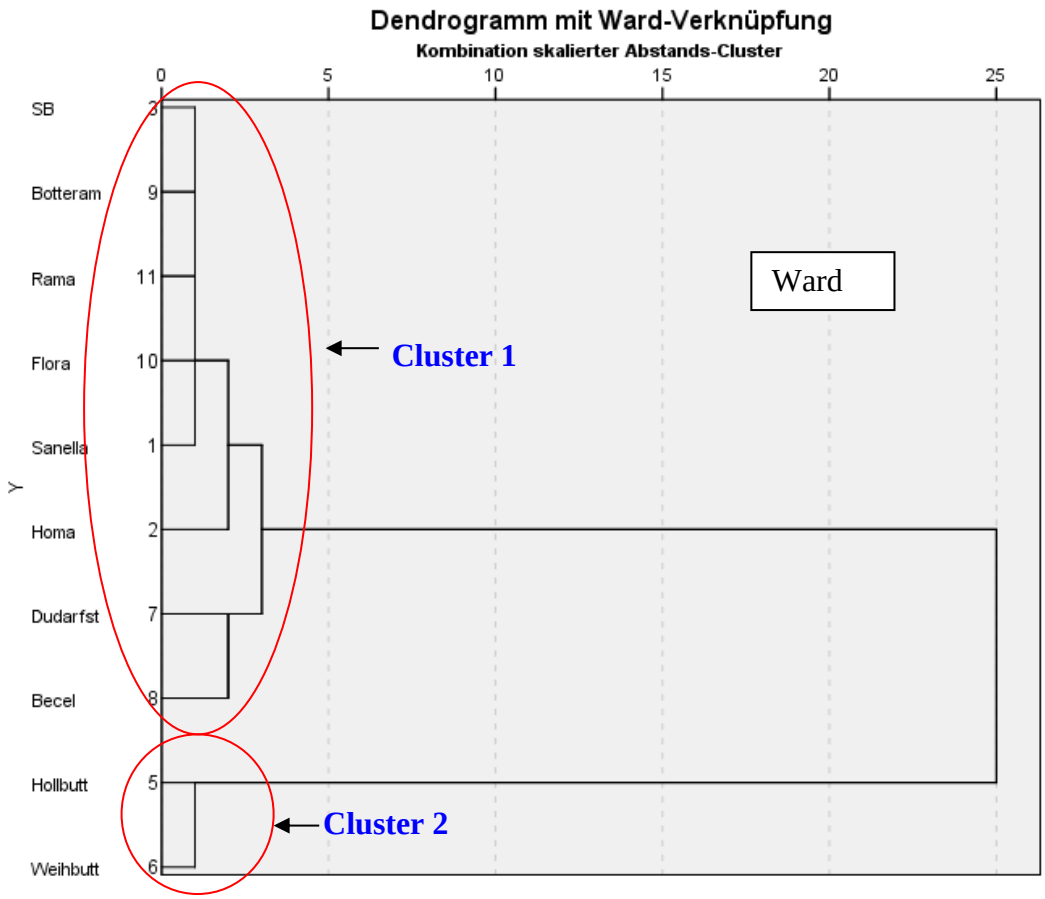
  (Alternativ kann man mit Hilfe der internen Hilfsvariablen $CASENUM zunächst eine Fallnummer erzeugen: „Transformieren“, „Variable berechnen…“, Funktionsgruppe „Alle“ wählen, „$casnum“ in das Feld „Numerischer Ausdruck“ übertragen (mit Doppelklick auf $casnum oder mit dem Pfeil übertragen), als „Zielvariable“ z.B. den Namen NR eintragen. Die Variable NR mit den Fallnummern 1 bis 11 wird den schon vorhandenen Variablen hinzugefügt. Zur Fallauswahl öffnet man mit "Daten", "Fälle auswählen" die Dialogbox "Fälle auswählen". Im Feld "Auswählen" wird "Falls Bedingung zutrifft" gewählt und mit Klicken auf die Schaltfläche "Falls" die Unterdialogbox "Fälle auswählen: Falls" geöffnet. Die Marke Delicado hat die Fallnummer 4. Zum Ausschluss der Fallnummer 4 wird die Bedingung nr ~ = 4 (~ ist das Symbol für das logische "nicht") eingetragen.)

- "Analysieren", "Klassifizieren", "Hierarchische Cluster…" öffnet die Dialogbox "Hierarchische Clusteranalyse". Die 10 Eigenschaftsvariablen (Merkmale der Margarinemarken) werden in das Eingabefeld "Variable(n)" übertragen. Im Feld "Cluster" ist der Optionsschalter "Fälle" und im Feld "Anzeigen" sind die Optionsschalter "Statistik" und "Diagramme" voreingestellt. Diese Einstellungen werden belassen. Klicken der Schaltfläche "Diagramm…" öffnet die Unterdialogbox "Hierarchische Clusteranalyse: Diagramme". Es wird durch Markieren des entsprechenden Kontrollkästchens "Dendrogramm" angefordert.

Klicken der Schaltfläche "Methode…" öffnet die Dialogbox "Hierarchische Clusteranalyse: Methode". Klicken auf den senkrechten Pfeil neben dem Eingabefeld "Clustermethode" öffnet eine Drop-Down-Liste mit den verfügbaren Clusterverfahren. Gewählt wird "Ward-Methode". Im Feld "Maß“ wird der Optionsschalter "Intervall" und in der dazugehörigen Drop-Down-Liste "Quadrierte euklidische Distanz" gewählt. Ausführen mit „Weiter“ und „OK“.

- Im Dendrogramm sind die Stufen der Clusterbildung veranschaulicht. Es werden nach dem Ellbogen-Kriterium wieder zwei Cluster gebildet. Sie enthalten die gleichen Marken wie bei Anwendung der Clustermethode Single-Linkage.
-
-

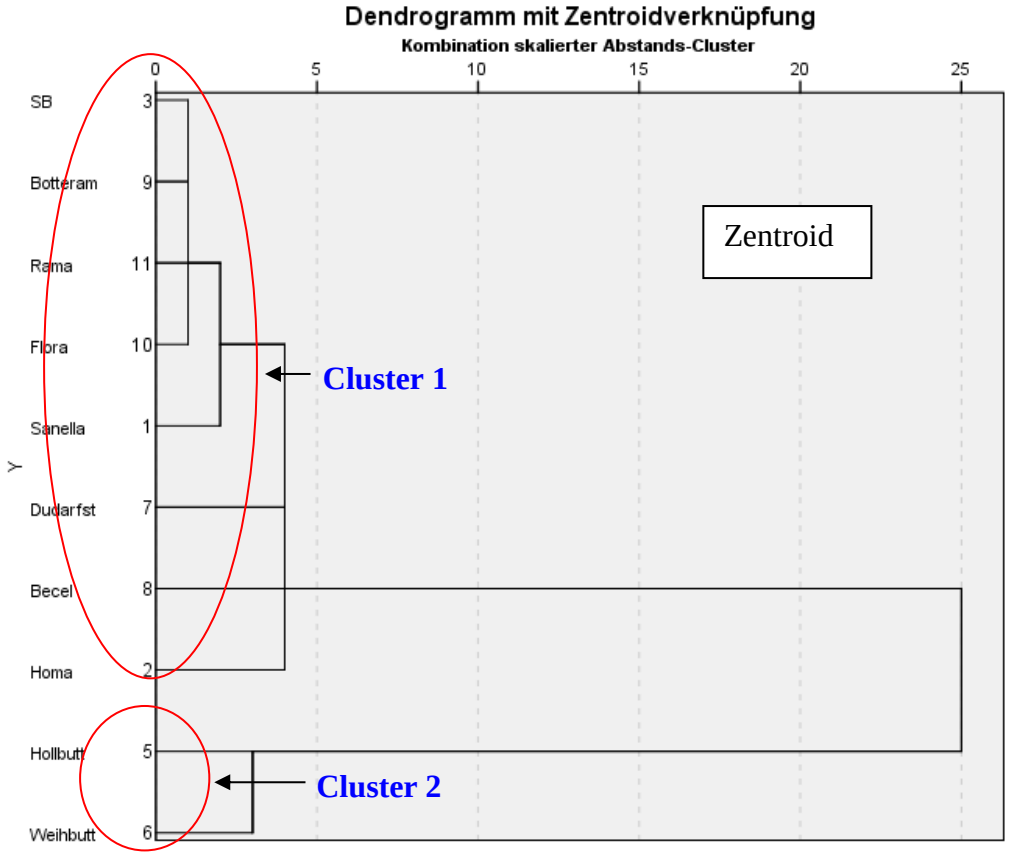




- Zur Anwendung des Clusterverfahrens Zentroid wird wie oben vorgegangen, mit dem Unterschied, dass als Methode "Zentroid-Clustering" gewählt wird. Im Dendrogramm sind die Stufen der Clusterbildung veranschaulicht. Es werden wieder zwei Cluster gebildet. Sie enthalten die gleichen Marken wie bei der Clustermethode „Durchschnittliche Verlinkung zwischen den Gruppen“.

  Im Dendrogramm sind die Stufen der Clusterbildung zu sehen. Auch bei diesem Verfahren werden nach dem Ellbogen-Kriterium zwei Cluster gebildet.

  Es zeigt sich, dass alle drei Clusterverfahren zur gleichen Clusterlösung führen.



d.

- Zunächst wird die Marke Delicado aus der Analyse ausgeschlossen (s. Lösung 11 c).

- "Analysieren", "Klassifizieren", "K-Means-Cluster…" öffnet die Dialogbox "K-Means-Clusteranalyse". Die 10 Eigenschaftsvariablen (Merkmale der Margarinemarken) werden in das Eingabefeld "Variable(n)" übertragen. Die Variable MARKE übertragen wir in „Fallbeschriftung“. Die "Anzahl der Cluster" ist auf 2 voreingestellt und wird so belassen. Im Feld "Methode" ist "Iterieren und klassifizieren" voreingestellt und wird so belassen. Klicken auf die Schaltfläche "Optionen…" öffnet die Dialogbox "K-Means-Clusteranalyse: Optionen". Neben der Voreinstellung "Anfängliche Clusterzentren" fordern wir durch Markieren des entsprechenden Kontrollkästchens "Clusterinformationen über jeden Fall" an.

- Die Ausgabetabelle "Clusterzugehörigkeit" gibt an, welche Marken jeweils in einem der beiden Cluster zusammengefasst werden. Dabei zeigt sich, dass die mit den hierarchischen Clusterverfahren gewonnene Clusterlösung bestätigt wird.

  Die Marken Sanella, Homa, SB, Du darfst, Becel, Botteram, Flora und Rama sind in einem Cluster und die Marken Hollbutt und Weihbutt in dem anderen Cluster enthalten.

  Die Distanzwerte in der Tabelle geben den Abstand zwischen einer Marke und dem Clusterzentrum des Clusters an. Kleine Distanzwerte signalisieren, dass die Marke typisch (repräsentativ) für das Cluster und große Werte, dass die Marke weniger repräsentativ für das Cluster ist.

Cluster-Zugehörigkeit

Fallnummer	marke	Cluster	Distanz
1	Sanella	2	1,150
2	Homa	2	1,897
3	SB	2	,973
5	Hollbutt	1	,949
6	Weihbutt	1	,949
7	Dudarfst	2	1,834
8	Becel	2	1,753
9	Botteram	2	,769
10	Flora	2	1,128
11	Rama	2	1,016

- Die Ausgabetabelle "Clusterzentren der endgültigen Lösung" gibt getrennt für jedes Cluster die Mittelwerte der Merkmalswerte der Marken des jeweiligen Clusters an. Cluster 1 mit den Marken Hollbutt und Weihbutt (die beiden Buttermarken) zeichnet sich z. B. durch besseren Geschmack (5,320 > 4,212), höheren Kaloriengehalt (5,278 > 3,825), größere Natürlichkeit (5,326 > 3,766) und höherem Anteil tierischer Fette aus (5,808 > 1,971).

Clusterzentren der endgültigen Lösung

	Cluster	
	1	2
streichf Streichfähigkeit	3,472	5,086
preis Preis	4,278	3,950
haltbark Haltbarkeit	3,827	4,408
ungesaet Ungesättigte Fettsäuren	3,883	3,864
backeign Back- u. Brateignung	4,284	3,731
geschmac Geschmack	5,320	4,212
kalorien Kaloriengehalt	5,278	3,825
tierfett Anteil tierischer Fette	5,808	1,971
vitaming Vitamingehalt	4,486	4,017
natuerli Natürlichkeit	5,326	3,766

- Die Distanz zwischen den Zentren der beiden Cluster wird in der nachfolgenden Tabelle angeführt.

Distanz zwischen Clusterzentren der endgültigen Lösung

Cluster	1	2
1		4,906
2	4,906	