https://datascience-project.eu/

Training Fiche Vorlage

Titel	Cluster-Analyse
Schlüsselwörter (Meta-Tags)	Statistische Einheiten, Cluster, Intra-Cluster, Inter-Cluster, Dissimilaritätsindex, Merge-Distanz, Dendogramm.
Sprache	Deutsch
Zielsetzungen / Lernziele / Lernergebnisse	**Ziel dieses Moduls ist es, die Technik der Clusteranalyse einzuführen und zu erklären.** **Am Ende dieses Moduls werden Sie in der Lage sein:** - **Die Logik der Clusteranalyse zu verstehen** - **Die Anforderungen zu kennen** - **Eine Clusteranalyse durchzuführen**

Lehrgang:	
Datenwissenschaftliche Kompetenz	
Modul Datenvisualisierung und visuelle Analyse	X
Einführung in die Datenwissenschaft für Human- und Sozialwissenschaften	
Datenwissenschaft für den guten Zweck	
Datenjournalismus und Geschichtenerzählen	

Beschreibung	In diesem Modul lernen wir die mehrdimensionale Analysetechnik der Clusteranalyse kennen, die auch automatische Gruppenanalyse genannt wird. Clusteranalysen dienen dazu, statistische Einheiten, die gemeinsame Merkmale aufweisen, zu gruppieren und sie nicht a priori definierten Kategorien zuzuordnen. Die gebildeten Gruppen müssen innerhalb (Intra-Cluster) möglichst homogen und außerhalb (Inter-Cluster) heterogen sein. Die Anwendungsmöglichkeiten dieser Art von Analyse sind vielfältig: Clusteranalysen werden z.B. in der Informatik, Medizin, Biologie, oder im Marketing genutzt.

Co-funded by the Erasmus+ Programme of the European Union

Mit Unterstützung des Programms Erasmus+ der Europäischen Union. Dieses Dokument und sein Inhalt spiegeln die
Die Verantwortung für den Inhalt dieser Veröffentlichung trägt allein der Verfasser; die Kommission haftet nicht für die weitere Verwendung der darin enthaltenen Angaben.

	Im letzten Teil des Moduls lernen wir, wie wir die Software R für Clusteranalysen einsetzen können.
Inhalt in 3 Ebenen gegliedert	**1. EINLEITUNG** Clusteranalysen dienen dazu, statistische Einheiten, die gemeinsame Merkmale aufweisen, zu gruppieren und sie nicht a priori definierten Kategorien zuzuordnen. Die gebildeten Gruppen müssen innerhalb (Intra-Cluster) möglichst homogen und außerhalb (Inter-Cluster) heterogen sein. Clusteranalysen sind Verfahren, die im Wesentlichen aus vier Phasen bestehen: - Auswahl der Variablen - Datenerhebung - Datenverarbeitung - Überprüfung und Verwendung der Ergebnisse **2. ANFORDERUNGEN DER CLUSTERANALYSE** In der Clusteranalyse können verschiedene Arten von Variablen verwendet werden: - Deskriptive Variablen (Beispiel: demografische, sozioökonomische, geografische) - Variablen, welche Aufschluss über das Verhalten geben (d. h. diejenigen Variablen, die die Fragen beantworten: was, wann, wo, wie und warum) Wir sprechen in diesem Zusammenhang also über qualitative und quantitative Variablen. Die für die Clusteranalyse zur Verfügung stehende Stichprobe muss hinreichend zahlreich, identifizierbar, und stabil sein. Sie sollte außerdem leicht zugänglich und ausreichend ergiebig sein. **3. Wie man eine Clusteranalyse durchführt** **3.1 Dissimilaritätsmatrix (oder Distanzmatrix), D** Wir gehen von unserer **X-Datenmatrix** mit nxp Dimensionen aus und wandeln sie in eine **D-Distanzmatrix** mit nxn Dimensionen um. Die letztere ist nützlich um zu wissen, wie viele statistische Einheiten sich voneinander unterscheiden und daher nützlich für die Auswahl welcher Variablen in der Analyse berücksichtigt werden sollen.

Co-funded by the Erasmus+ Programme of the European Union

Mit Unterstützung des Programms Erasmus+ der Europäischen Union. Dieses Dokument und sein Inhalt spiegeln die
Die Verantwortung für den Inhalt dieser Veröffentlichung trägt allein der Verfasser; die Kommission haftet nicht für die weitere Verwendung der darin enthaltenen Angaben.

$$\mathbf{X} = \begin{pmatrix} x_{1,1} & & x_{1,p} \\ & x_{i,k} & \\ x_{n,1} & & x_{n,p} \end{pmatrix} \Longrightarrow \mathbf{D} = \begin{pmatrix} d_{1,1} & & d_{1,n} \\ & d_{i,j} & \\ d_{n,1} & & d_{n,n} \end{pmatrix}$$

Wie wir sehen, ist die Matrix **D** eine symmetrische Matrix, die entlang der Hauptdiagonalen den Wert 0 hat, da der Abstand eines Punktes zu sich selbst gleich Null ist.

Um die Abstände zwischen den Punkten zu berechnen, wird der Index $d_{i,j}$verwendet, welcher den Grad der Ähnlichkeit zwischen i und j misst.

Es gibt verschiedene Indizes, die wir zur Berechnung dieser Abstände verwenden können. Welcher Index für unsere Zwecke am besten geeignet, hängt von der Art der Variablen ab, die wir verwenden.

3.2 Entfernungen

- Bei der Verwendung **quantitativer Variablen** beziehen wir uns auf den **Grad der Unähnlichkeit** . Es gibt mehrere Möglichkeiten, ihn zu berechnen:
  Euklidischer Abstand:
  Diese bezieht sich auf den Satz von Pythagoras und ist empfindlich gegenüber Ausreißern. Er wird wie folgt berechnet:

$$d_{i,j} = \left[\sum_k (x_{i,k} - x_{j,k})^2\right]^{\frac{1}{2}}$$

  Manhattan-Distanz:
  Diese Methode wird auch City Block genannt und erweist sich als robuster als der euklidische Abstand, weshalb diese Metrik, wenn möglich, bevorzugt verwendet wird. Diese Methode wird wie folgt berechnet:

$$d_{i,j} = \sum_k |x_{i,k} - x_{j,k}|$$

Co-funded by the Erasmus+ Programme of the European Union
Mit Unterstützung des Programms Erasmus+ der Europäischen Union. Dieses Dokument und sein Inhalt spiegeln die
Die Verantwortung für den Inhalt dieser Veröffentlichung trägt allein der Verfasser; die Kommission haftet nicht für die weitere Verwendung der darin enthaltenen Angaben.

Bei der Berechnung von Entfernungen werden immer die Maßeinheiten der Variablen berücksichtigt. Der Effekt der Messung kann durch die Normierung der **X-Matrix** in der **Z-Matrix** eliminiert werden:

$$Z_k = \frac{(X_k - M_k)}{S_k}$$

Sobald die Matrix standardisiert ist, werden wir sie natürlich verwenden, um den Unähnlichkeitsindex zu berechnen. Die Manhattan-Distanz ergibt sich aus:

$$d_{i,j} = \sum_k \frac{1}{S_k} |z_{i,k} - z_{j,k})|$$

Dabei ist $\frac{1}{S_k}$ die Gewichtung.

Eine Standardisierung wird durchgeführt, wenn wir allen Variablen das gleiche Gewicht geben wollen. Wird hingegen eine Variable stärker gewichtet als die anderen, so wird keine Standardisierung vorgenommen.

- Bei der Verwendung von **Binärvariablen**, d. h. **von** Variablen, die nur zwei Modi haben (wenn wir von Modi sprechen, bedeutet dies, dass die uns zur Verfügung stehenden Variablen **qualitative Variablen** sind). Die Modi der Binärvariablen haben den Status 0 und 1. Mit dieser Art von Variablen berechnen wir **den Grad der Ähnlichkeit**, d. h. die Ähnlichkeit zwischen i und j.
  Binäre Variablen werden unterteilt in:
  Symmetrische Variablen BS, **BS:** die beiden Zustände (0 und 1) haben die gleiche Bedeutung.
  Asymmetrische Binärvariablen, BA: dem Zustand 1 wird mehr Bedeutung beigemessen als dem Zustand 0.

  M-Koeffizient (Simple Matching):
  Er wird für **symmetrische binäre Variablen** verwendet und durch Addition der Konkordanzhäufigkeiten und der Diskordanzhäufigkeiten berechnet und dann durch die Gesamtzahl geteilt.

  $$s = \frac{(a+d)}{p}$$

Co-funded by the Erasmus+ Programme of the European Union

Mit Unterstützung des Programms Erasmus+ der Europäischen Union. Dieses Dokument und sein Inhalt spiegeln die
Die Verantwortung für den Inhalt dieser Veröffentlichung trägt allein der Verfasser; die Kommission haftet nicht für die weitere Verwendung der darin enthaltenen Angaben.

Jaccard-Index:
Dieser Index wird für **asymmetrische binäre Variablen** verwendet und wird berechnet, indem die Konkordanzhäufigkeit durch die Differenz zwischen der Gesamt- und der Diskordanzhäufigkeit geteilt wird.

$$s = \frac{a}{(p-d)}$$

3.3 Arten von Clustern

Es gibt verschiedene Arten von Clustern, je nachdem, welchen Ansatz wir bei der Erstellung von Gruppen verfolgen wollen.

Hierarchische Algorithmen führen sukzessive Zusammenführungen oder Aufteilungen von Daten durch. Nachdem ein Objekt einem Cluster beigetreten ist, ist seine Zuordnung unwiderruflich. Wir unterscheiden zwischen folgenden hierarchischen Clustern:

- **Agglomerative Cluster (bottom-up**):
  Ziel ist es, die vielen Cluster zu gruppieren und einen einzigen Cluster zu erhalten, der alle von Anfang an vorhandenen Cluster enthält.

- **Divisive Cluster (top-down):**
  In diesem Fall gehen wir von einem einzigen Cluster aus, und das Ziel ist es, diesen in viele Cluster aufzuteilen.

3.4) Arten von Verbindungen zwischen statistischen Einheiten

Cluster können durch verschiedene Arten von Verbindungen gebildet werden:

- **Single-Linkage**
- **Complete-Linkage**
- **Average-Linkage**

Single-Linkage verwendet die Technik des "nächsten Nachbarn". Der Grad der Nähe zwischen zwei Gruppen wird unter Berücksichtigung des Mindestabstands zwischen den Punkten ermittelt. Mit anderen Worten, es werden die Einheiten berücksichtigt, die einander am nächsten sind. Diese Verknüpfung ist zwar

Co-funded by the Erasmus+ Programme of the European Union
Mit Unterstützung des Programms Erasmus+ der Europäischen Union. Dieses Dokument und sein Inhalt spiegeln die
Die Verantwortung für den Inhalt dieser Veröffentlichung trägt allein der Verfasser; die Kommission haftet nicht für die weitere Verwendung der darin enthaltenen Angaben.

rechnerisch am schnellsten zu erreichen, schafft jedoch Gruppen, die untereinander zu homogen sind.

Complete-Linkage verwendet stattdessen die Technik des "weitesten Nachbarn". Sie berücksichtigt die Ähnlichkeiten/Entfernungen zwischen den am weitesten entfernten Gruppen (also denjenigen, die einander weniger ähnlich sind). In der Praxis bedeutet dies, dass der kleinste maximale Abstand zwischen den Punkten berücksichtigt wird. Diese Verbindung ist zwar rechnerisch am langsamsten, führt aber zu sehr heterogenen Gruppen auf der Außenseite und homogenen Gruppen auf der Innenseite.

Average-Linkage bei der Bildung von Clustern verwendet den minimalen durchschnittlichen Abstand. In der Praxis wird zunächst der durchschnittliche Abstand zwischen allen Beobachtungen berechnet und dann der kleinste Abstand berücksichtigt. Auch diese Verknüpfung ist rechnerisch langsam, aber sie ist robust und weniger anfällig für Ausreißer.

Die **Ward-Methode** kann bei quantitativen Daten verwendet werden. Diese Technik minimiert die Varianz innerhalb von Gruppen, indem sie diese homogenisiert. In der Praxis maximiert diese Methode die interne Homogenität (bzw. minimiert die interne Heterogenität) und maximiert die externe Heterogenität.

3.5 Dendogramm und Fusionsabstand

Sobald die Verbindung, die die Daten in unserem Besitz am besten repräsentiert, ausgewählt wurde, wird das **Dendogramm** berechnet. Anhand dieses **Baumdiagramms** können wir die Verteilung der statistischen Einheiten veranschaulichen. Bei jedem Schritt vergrößert sich der Abstand zwischen den Clustern, so dass es notwendig ist, eine **Stoppregel** zu wählen**.** Diese Regel erlaubt es uns, die Anzahl der Gruppen zu wählen, die wir erhalten möchten. Die Baumschnitttechnik lässt sich anhand des Graphen der **Fusionssabstände** (oder -höhen) anwenden, der anzeigt, wo Cluster gebildet werden. Grafisch wird der Punkt betrachtet, an dem ein größerer Anstieg zu verzeichnen ist. Dieser Teil wird dann in dem der R-Software gewidmeten Teil des Moduls aufgegriffen.

4. Beispiel mit R-Software

Die Clusteranalyse zielt darauf ab, die bestmögliche Verteilung einer Gruppe von Elementen in Bezug auf Anzahl und Zusammensetzung zu ermitteln, so dass diese innerhalb der Gruppe so homogen wie möglich und so verschieden wie möglich voneinander sind. Diese Konstruktionen können sowohl in Bezug auf die Wahl der

Co-funded by the Erasmus+ Programme of the European Union

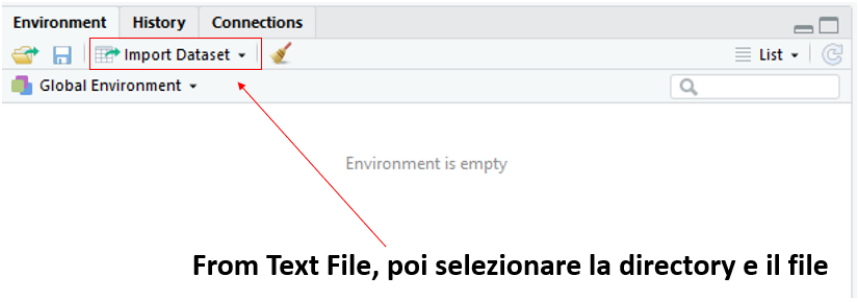
Mit Unterstützung des Programms Erasmus+ der Europäischen Union. Dieses Dokument und sein Inhalt spiegeln die
Die Verantwortung für den Inhalt dieser Veröffentlichung trägt allein der Verfasser; die Kommission haftet nicht für die weitere Verwendung der darin enthaltenen Angaben.

Gruppierungsstrategien als auch in Bezug auf das für die Messung der Ähnlichkeit/Unähnlichkeit gewählte Kriterium durchgeführt werden.

Wir verwenden folgenden Datensatz:

Nazioni	Cereali	Riso	Patate	Zucchero	Verdure	Vino	Carne	Latte	Burro	Uova
Belgio	72,2	4,2	98,8	40,4	103,2	20,9	102	80	7,7	14,2
Danimarca	70,5	2,2	57	39,5	50	22	105,8	145,2	4,1	14,3
Germania	71,3	2,3	74,1	37,1	83,1	22,8	97,2	90,7	6,9	14,8
Grecia	109,8	5,4	90	30	229,5	25,3	77,1	63,1	0,9	11,3
Spagna	71,4	5,8	107,8	26,8	191,7	43	102,1	98,4	0,6	15,3
Francia	73	4,3	78,2	34,1	95	64,5	110,5	98,9	8,9	15
Irlanda	93,4	3,2	151,5	34,8	55	3,9	105	185,9	3,4	11,4
Italia	110,2	4,8	38,6	27,9	181,9	61,6	88	65	2,4	11,1
Olanda	54,6	5	86,7	39,7	99	14	89,4	136,2	5,4	10,7
Portogallo	86	5,7	106,6	29,4	100	57	75,5	96	1,5	7,7
RegnoUnito	74,3	4,5	94,1	39,8	60	10,4	74,4	129,3	3,2	10,8
Austria	68,7	4,2	62,6	37,1	81,9	34,3	93,4	121,3	4,3	13,4
Finlandia	70,1	5,4	61,6	35,7	52,6	10,2	65	208,4	5,8	10,9
Islanda	79,7	1,9	50,2	54,9	50	6,2	71,7	205,6	4,6	11,3
Norvegia	76,9	3,5	73,2	37,3	48,3	6,6	54,9	176,5	2,1	11,3
Svezia	69,3	4,3	70	37,5	48,5	12,3	60,5	154,1	5,7	12,9

Wir importieren den Datensatz:



Bei den Optionen für den ***Zeilennamen*** wählen wir die Formulierung: "***use first column***", um die Bezeichnungen der Spalten und damit auch der Variablen in den Diagrammen anzuzeigen.

Bei den Optionen für ***decimal*** wählen wir: "***comma***".

Wir weisen **X** als Objekt mit folgendem Befehl den in der Analyse verwendeten Datensatz zu:

X<-as.matrix(nome_del_dataset)

Dann standardisieren wir die X-Matrix durch:

Co-funded by the Erasmus+ Programme of the European Union

Mit Unterstützung des Programms Erasmus+ der Europäischen Union. Dieses Dokument und sein Inhalt spiegeln die
Die Verantwortung für den Inhalt dieser Veröffentlichung trägt allein der Verfasser; die Kommission haftet nicht für die weitere Verwendung der darin enthaltenen Angaben.

Z<Skala(X)

Als Nächstes berechnen wir den Abstand zwischen den Elementen. Wir können entweder den euklidischen Abstand oder die Manhattan-Distanz verwenden.

Die entsprechenden Befehle sind:

d<-dist(Z)

D<-round(D,2)

d_m<-dist(Z, method="manhattan")

d_m<-round(d_m, 2)

NB: Mit dem Befehl Runden können wir auf die gewünschte signifikante Zahl aufrunden, in diesem Fall auf die Sekunde.

Dann geht es um die Wahl der Verbindung (Linkage) zwischen den Elementen.

Beginnen wir mit **Single-Linkage**:

hc_s<-hclust(d,method="single")

Mit dem Befehl können wir eine **Zusammenfassung der Ergebnisse** der Einfachbindung anzeigen:

summary(hc_s)

Dann können wir das **Dendogramm** mit der Plot-Funktion visualisieren:

plot(hc_s)

Um zu entscheiden, wo das Baumdiagramm geschnitten werden soll, verwenden wir den Befehl **cutree.** Durch die Darstellung des Fusionspunktes im **Scree-Plot** erfahren wir, wie viele Gruppen wir durch die Fusionsabstandsverknüpfung am besten erhalten sollen. Die Befehle dazu lauten:

n<-nrow(X)

n_clus<-seq(n-1,1)

hc_s$merge

hc_s$height

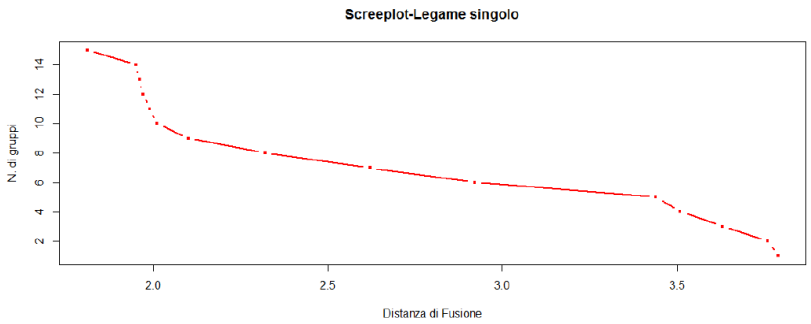
Co-funded by the Erasmus+ Programme of the European Union

Mit Unterstützung des Programms Erasmus+ der Europäischen Union. Dieses Dokument und sein Inhalt spiegeln die
Die Verantwortung für den Inhalt dieser Veröffentlichung trägt allein der Verfasser; die Kommission haftet nicht für die weitere Verwendung der darin enthaltenen Angaben.

d_fus_s<-hc_s$height

plot(d_fus_s,n_clus, "b", main="Screeplot Single-Linkage", xlab="Fusionsabstand", ylab="Anzahl der Gruppen",cex=0.6, col="red",lwd=2.5)

Grafisch:



Wenn wir nun die Fusionspunkte (hc_s$merge) und die Höhen (hc_s$height) sehen wollen, verwenden wir den Befehl **cbind**, um sie zusammen zu visualisieren. Der Befehl $merge zeigt für jeden Schritt des Gruppierungsalgorithmus das Paar von Elementen an, die entsprechend der gewählten Verknüpfung zusammengeführt wurden. Werte, denen ein "-" vorangestellt ist, stehen für ein einzelnes Element, während positive Werte jene Cluster darstellen, welche in den vorherigen Schritten gebildet wurden.

So wird im ersten Schritt der erste Cluster aus dem Paar (13, 16) gebildet, das den Modellen Finnland und Schweden entspricht, während der dritte Cluster (Schritt 10) aus den Elementen des Clusters 2 (Griechenland, Italien) plus dem Element 1 (Frankreich) gebildet wird. Das Feld $height gibt den Abstand an, der für die Fusion zwischen den Elementen/Gruppen berücksichtigt wird.

cbind(hc_s$merge,hc_s$height)

Co-funded by the Erasmus+ Programme of the European Union

Mit Unterstützung des Programms Erasmus+ der Europäischen Union. Dieses Dokument und sein Inhalt spiegeln die
Die Verantwortung für den Inhalt dieser Veröffentlichung trägt allein der Verfasser; die Kommission haftet nicht für die weitere Verwendung der darin enthaltenen Angaben.

```
> cbind(hc_s$merge,hc_s$height)
      [,1] [,2] [,3]
 [1,]  -13  -16 1.81
 [2,]   -2   -3 1.95
 [3,]   -1    2 1.96
 [4,]  -15    1 1.97
 [5,]  -11    4 1.99
 [6,]   -9    5 2.01
 [7,]  -12    3 2.10
 [8,]    6    7 2.32
 [9,]   -6    8 2.62
[10,]   -4   -8 2.92
[11,]  -14    9 3.44
[12,]   -7   11 3.51
[13,]  -10   12 3.63
[14,]   10   13 3.76
[15,]   -5   14 3.79
```

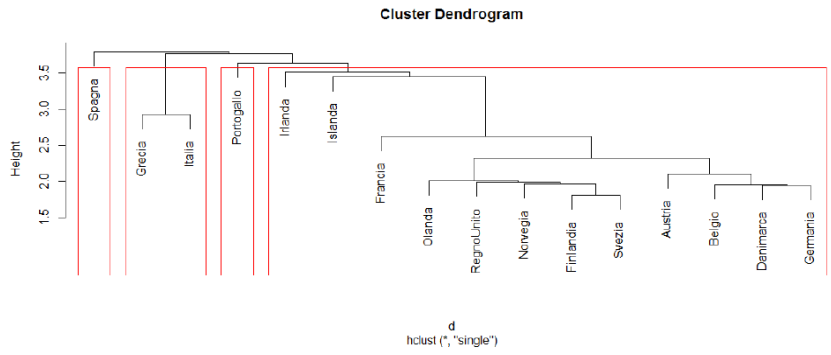
Zum Schneiden des Baums verwenden wir den Befehl **cutree**, unter **k** setzen wir den Punkt, an dem der Fusionsabstand einen horizontalen Trend annimmt:

groups <- cutree(hc_s, k=4)

plot(hc_s)

rect.hclust(hc_s, k=4, border="red")

Das Dendogramm wird uns folgendes Ergebnis anzeigen:



Man kann sagen, dass diese Art der Fusion nicht gut ist, weil es Cluster gibt, die einzelne Elemente enthalten, und einen Cluster, der innerhalb des Clusters zu homogen ist.

Mit den anderen Links verfahren wir auf die gleiche Weise.

Complete-Linkage:

Co-funded by the Erasmus+ Programme of the European Union

Mit Unterstützung des Programms Erasmus+ der Europäischen Union. Dieses Dokument und sein Inhalt spiegeln die
Die Verantwortung für den Inhalt dieser Veröffentlichung trägt allein der Verfasser; die Kommission haftet nicht für die weitere Verwendung der darin enthaltenen Angaben.

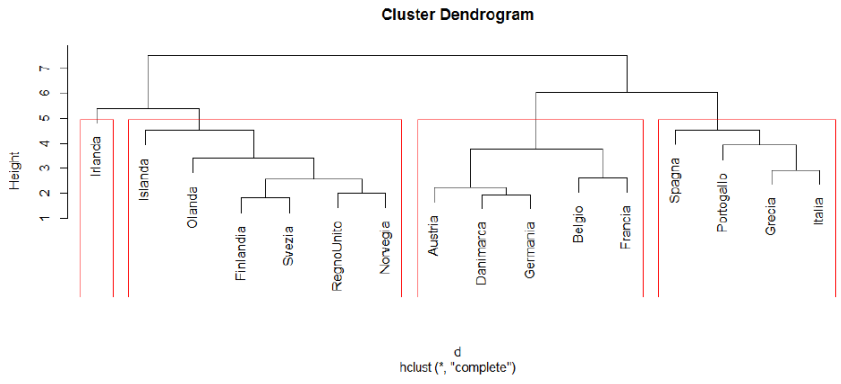
```
hc_c<-hclust(d,method="compl")

summary(hc_c)

plot(hc_c)

n<-nrow(X)

n_clus<-seq(n-1,1)

hc_c$merge

hc_c$height

d_fus_c<-hc_c$height
```

Mit folgenden Befehlen erhalten wir den Scree-Plot der Fusionsabstände für Complete-Linkage:

```
plot(d_fus_c,n_clus, "b", main="Screeplot Complete-Linkage",
xlab="Fusionsabstand", ylab="Anzahl der Gruppen",cex=0.6, col="red",lwd=2.5)

cbind(hc_c$merge,hc_c$height)
```

Beim Schneiden des Baumdiagramms für Complete-Linkage auf k, werden wir die Figur entsprechend dem Scree-Plot der Fusionsabstände zuordnen:

```
groups <- cutree(hc_c, k=4)

plot(hc_c)

rect.hclust(hc_c, k=4, border="red")
```



Co-funded by the Erasmus+ Programme of the European Union

Mit Unterstützung des Programms Erasmus+ der Europäischen Union. Dieses Dokument und sein Inhalt spiegeln die
Die Verantwortung für den Inhalt dieser Veröffentlichung trägt allein der Verfasser; die Kommission haftet nicht für die weitere Verwendung der darin enthaltenen Angaben.

Average-Linkage:

hc_a<-hclust(d,method="average")

summary(hc_a)

plot(hc_a)

n<-nrow(X)

n_clus<-seq(n-1,1)

hc_a$merge

hc_a$height

d_fus_a<-hc_a$height

Scree-Plot der Fusionsabstände für Average-Linkage:

plot(d_fus_a,n_clus, "b", main="Screeplot Average-Linkage", xlab="Fusionsabstand", ylab="Anzahl der Gruppen",cex=0.6, col="red",lwd=2.5)

cbind(hc_a$merge,hc_a$height)

Wenn wir den Baum bei Average-Linkage auf k schneiden, werden wir die Zahl entsprechend dem scree-Plot der Fusionsabstände zuweisen:

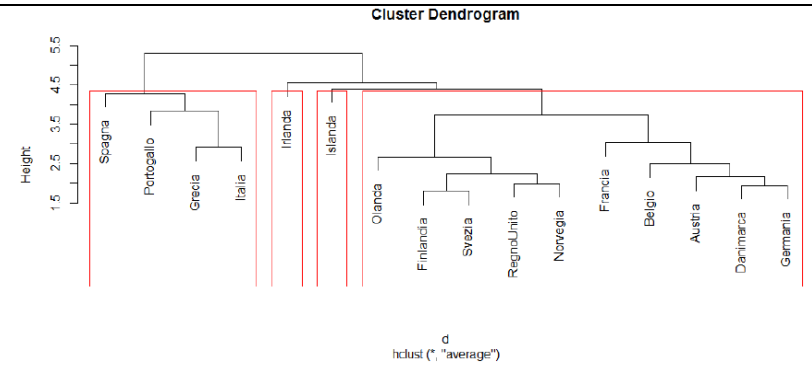
groups <- cutree(hc_a, k=4)

plot(hc_a)

rect.hclust(hc_a, k=4, border="red")

Co-funded by the Erasmus+ Programme of the European Union

Mit Unterstützung des Programms Erasmus+ der Europäischen Union. Dieses Dokument und sein Inhalt spiegeln die
Die Verantwortung für den Inhalt dieser Veröffentlichung trägt allein der Verfasser; die Kommission haftet nicht für die weitere Verwendung der darin enthaltenen Angaben.

	
Selbstbeurteilung (Multiple-Choice-Fragen und Antworten)	1. Die Distanzmatrix: A) hat auf der großen Diagonale alle 0 B) hat auf der größten Diagonale alle 1 C) hat auf der größten Diagonale die Abstände zwischen i und j 2. Welche Arten von Variablen können in der Clusteranalyse verwendet werden? A) nur qualitative Variablen B) nur quantitative Variablen C) sowohl qualitative als auch quantitative Variablen 3. Was ist das Ziel der Clusteranalyse? A) Zusammenfassung von statistischen Einheiten nach gemeinsamen Merkmalen B) Erstellen einer Linearkombinationen von Ausgangsvariablen C) Die Anzahl der Variablen zur Erklärung eines Phänomens reduzieren
Ressourcen (Videos, Verweislinks)	
Verwandtes Material	
Verwandte PPT	

Co-funded by the Erasmus+ Programme of the European Union

Mit Unterstützung des Programms Erasmus+ der Europäischen Union. Dieses Dokument und sein Inhalt spiegeln die
Die Verantwortung für den Inhalt dieser Veröffentlichung trägt allein der Verfasser; die Kommission haftet nicht für die weitere Verwendung der darin enthaltenen Angaben.

Literaturverzeichnis	Johnson, S. C. (1967). Hierarchical clustering schemes, Psychometrika, 32, 241-254. Pollice, A. (2013). Statistica multivariata, http://www.uniba.it/ricerca/dipartimenti/dse/dipartimento/personal/personale-docente/umfrage/stat_mult/disp10.pdf Ward, J. H., Jr. (1963). Hierarchical Grouping to Optimize an Objective Function, Journal of American Statistical Association, 58, 236-244.
Zur Verfügung gestellt von	[Unisalento]

Co-funded by the Erasmus+ Programme of the European Union

Mit Unterstützung des Programms Erasmus+ der Europäischen Union. Dieses Dokument und sein Inhalt spiegeln die
Die Verantwortung für den Inhalt dieser Veröffentlichung trägt allein der Verfasser; die Kommission haftet nicht für die weitere Verwendung der darin enthaltenen Angaben.