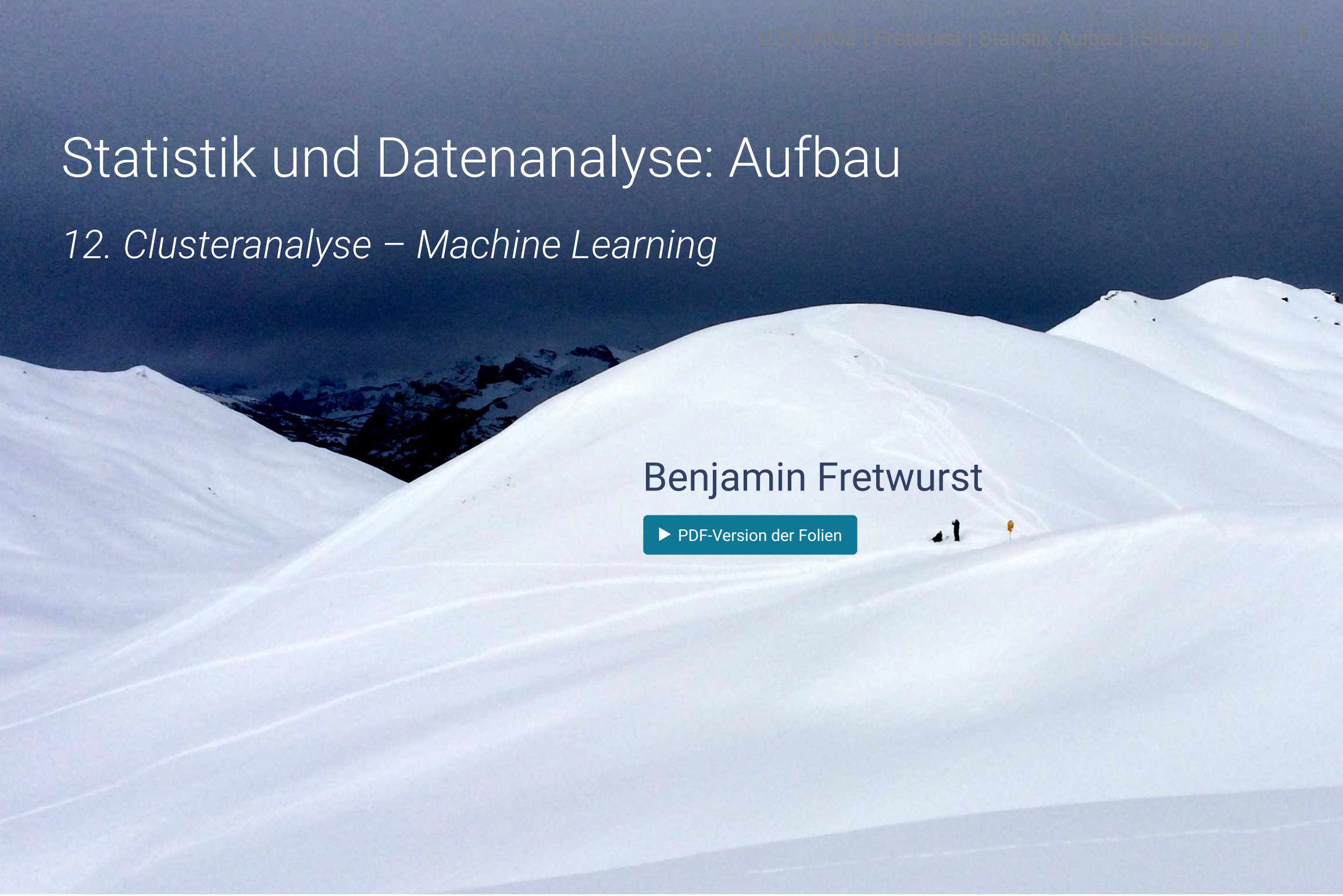
Statistik und Datenanalyse: Aufbau

12. Clusteranalyse – Machine Learning



Benjamin Fretwurst

▶ PDF-Version der Folien

Inhalt

Orga

Photo: Lisa Vogel

Was könnten wir die letzten zwei Sitzungen machen?

Version 1

- Übung zur Clusteranalyse (wie im Plan in Präsenz)
- Zusammenfassung / Wiederholung (online only)

Version 2 (beides online only)

- Fokussierte Wiederholung (Dummys & Interaktionen)
- LEFs durchgehen (online only)
- Übung zur Clusteranalyse (im Begleittext)

Lernziele

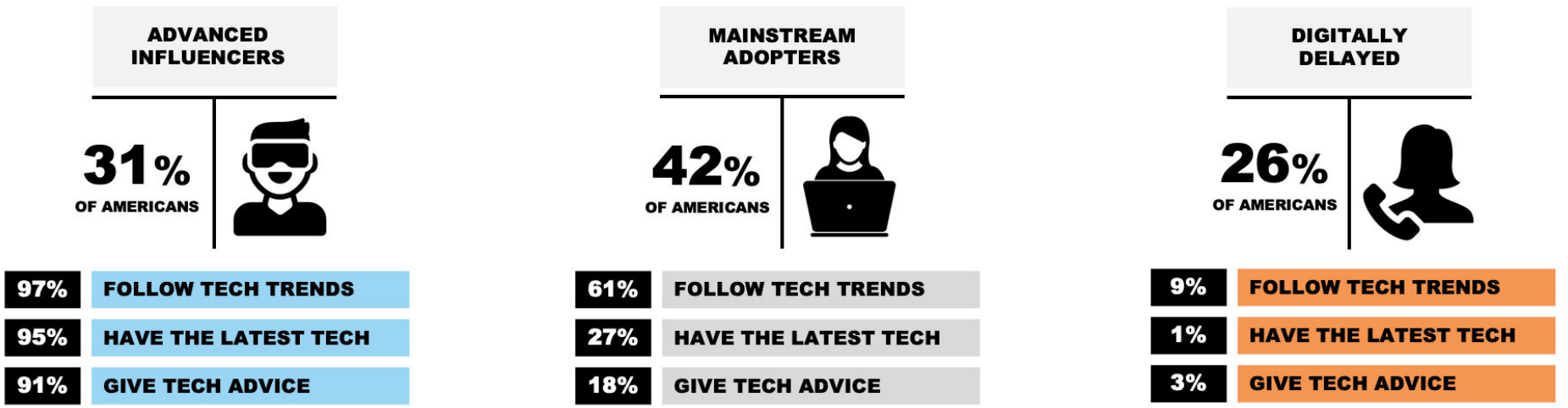
Gruppensuche – Clusteranalysen

- Problemstellung
- Anwendungsmöglichkeiten
- Vorgehensweise
- Varianten
- Umsetzung in R (im Begleittext)

1 Clusteranna
nalyse

Beispiel

Es wurde erst eine Faktorenanalyse über viele Items zur Techniknutzung gemacht und anhand der Faktoren Gruppen gebildet.



		GENDER		AGE				US REGION				HH INCOME			POLITICS		EDUCATION	
TABLE SHOWS %	TOTAL	FEMALE	MALE	16-29	30-44	45-59	60+	SOUTH	WEST	MIDWEST	NTH-EAST	<$40K	$40-$99K	$100K+	LIB	CON	UNI GRAD	OTHER
ADVANCED INFLUENCER	31.4	24.4	38.7	39.7	47.7	25.9	13.9	32.9	29.2	27.4	35.7	23.4	34.3	38.6	37.9	31.4	38.5	28.2
MAINSTREAM ADOPTER	42.3	44.3	40.2	42.0	36.9	47.0	43.3	41.3	44.1	45.4	38.5	45.8	39.3	42.1	40.9	38.0	37.2	44.5
DIGITALLY DELAYED	26.3	31.3	21.1	18.3	15.4	27.1	42.8	25.8	26.7	27.2	25.8	30.8	26.5	19.3	21.2	30.6	24.2	27.2
SAMPLE BASE (N)	2,025	1,024	992	486	501	491	547	766	479	427	353	736	797	492	491	739	615	1,410

TECHNOLOGY SEGMENTS IN AMERICA

Anwendungen

1. **Mustererkennung in grossen Datenmengen**
2. **Marketing und Kundensegmentierung**
3. **Biologie und Bioinformatik**
    - Klassifizierung von Genexpressionsdaten
4. **Medizin und Gesundheitswesen**
    - Patientenklassifizierung für personalisierte Medizin
5. **Bildverarbeitung zu Segmentierung**
6. **Finanzwesen**
    - Risikobewertung, Betrugserkennung
7. **Sozialwissenschaften: Gruppen anhand von Meinungen, Einstellungen, Handeln**
8. **Maschinelles Lernen**
    - Initialisierung von Clustern als Ausgangspunkte für Algorithmen
9. **Umweltwissenschaften**
    - Klassifizierung von Umweltdaten, Identifikation von Umweltclustern und Überwachung über die Zeit.
10. **Textanalyse und Natural Language Processing (NLP)**
    - Gruppierung von Dokumenten basierend auf dem Inhalt, Identifikation von Themenclustern und Sentimentanalyse.

1.1 Problemstellung und Vorgehen

Problemstellung

Wie können Fälle in einem Datensatz nach einer oder mehreren Variablen gruppiert werden?

Grundsätzliches Vorgehen

Wir suchen Gruppen (Cluster) von Fällen, die sich untereinander so stark wie möglich ähneln (homoge Cluster) und so stark von den anderen Gruppen unterscheiden wie möglich. Es geht also um Segmentierung anhand von Mustern in den Daten – Clusteranalyse ist Mustererkennung.

Zugehörigkeit des Verfahrens

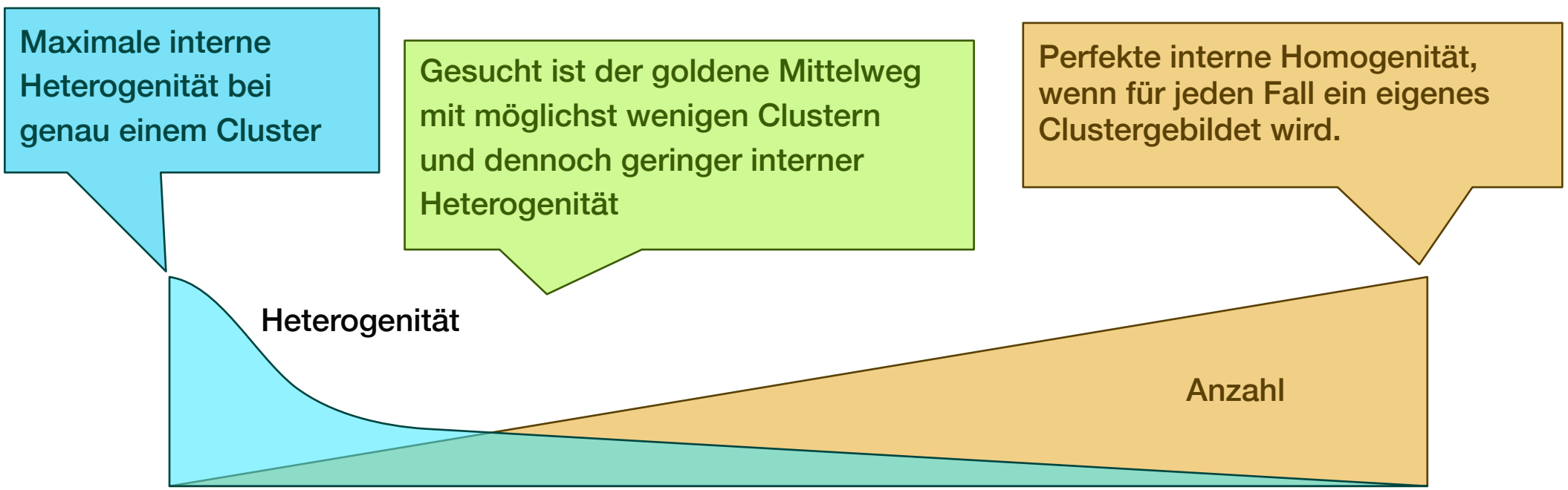
Die Clusteranalyse gehört zu den explorativen Verfahren.

Im Kontext von ML wird sie als «unsupervised learning» behandelt.

Optimierungsproblem der Clusteranalyse

Ziel

Wir wollen eine starke Vereinfachung, also wenige Cluster und gleichzeitig wenig Heterogenität in den Clustern, die aber mit der Zahl der Cluster abnimmt.



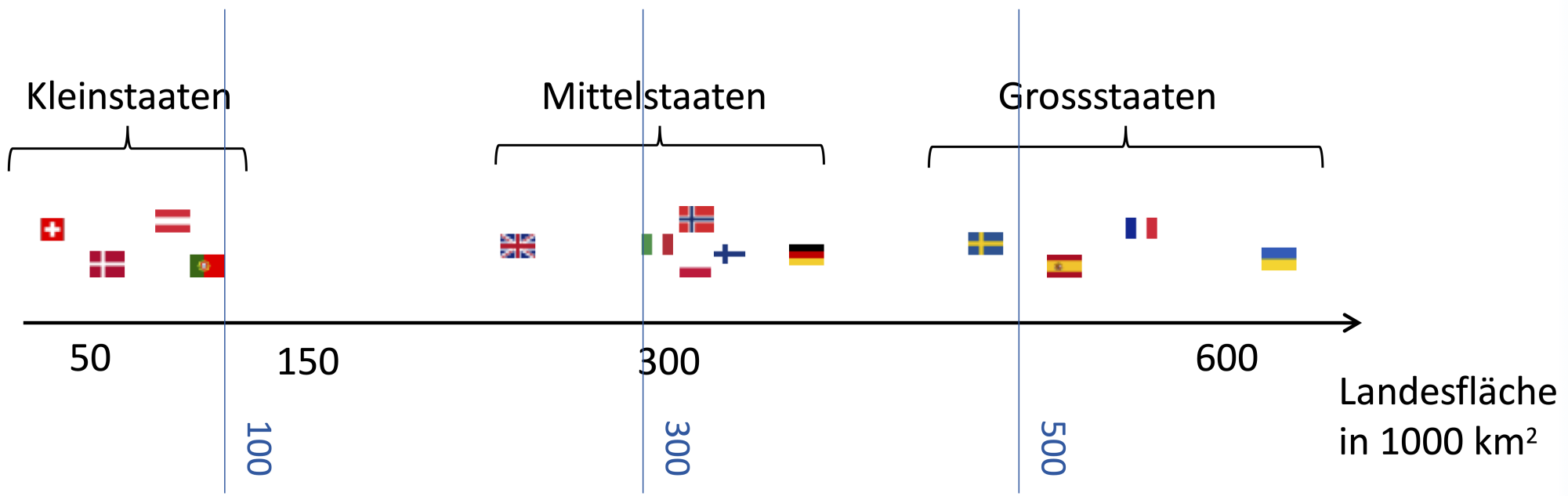
1.2 Ablauf einer Clusteranalyse

1. Auswahl der Clustervariablen (ggf. vorher Faktorenanalyse)
2. Bestimmung der Ähnlichkeiten
3. Auswahl des Fusionsalgorithmus
4. Bestimmung der Clusterzahl
5. Interpretation einer Cluster-Lösung

Monothetische CA

Vorgeehen

Alle Fälle werden bezüglich eines einzigen Merkmals geclustert. Die Grenzwerte zwischen den Clustern sind nicht zwingend gleichmässig verteilt.



Nobelpreisverdächtige Schokolade

Polythetische CA

Alle Fälle werden bezüglich ihrer Werte auf mehreren Merkmalen geclustert.

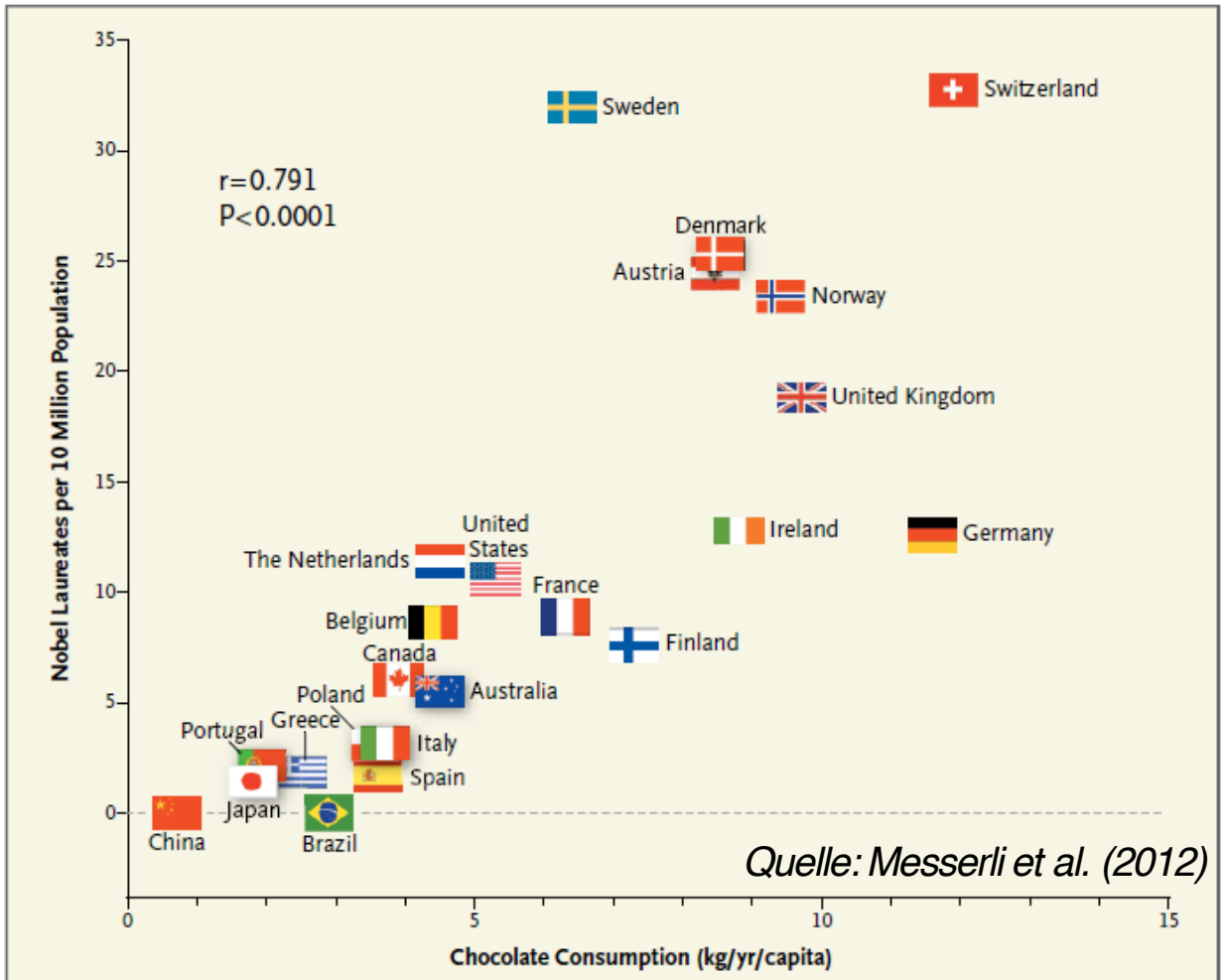


Figure 1. **Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.**

1.3 Voraussetzungen

Nicht zu viele fehlende Werte

Zu viele fehlende Werte verfälschen die Clusterbildung.

Skalenniveau

Das Skalenniveau spielt keine Rolle. Die Algorithmen können Cluster anhand von metrischen, dichotomenn oder kategorialen (mehrere Kategorien) Variablen extrahieren.

Fallzahl

Es braucht relativ grosse Stichproben. Bei kleinen Stichproben sind die Clusteranalysen recht ungenau.

Ähnliche Skalierung der Variablen

Haben die Variablen sehr unterschiedliche Skalierungen (zB Geschlecht dichotom und Alter in Jahren), dann ist eine vorherige z-Transformation der Variablen sinnvoll.

1.4 Vorgehensweise

Proximitätsmasse festlegen

Ähnlichkeits- bzw. Distanzmasse wählen

Distanzmasse

- euklidische Distanzen bei metrischen Variablen
- M-Koeffizient bei dichotomen Variablen (Übereinstimmungen)

Ähnlichkeitsmasse

- (Q-)Korrelationskoeffizient bei stetigen Variablen
- χ^2-Quadrat-Homogenität bei kategoriellen Merkmalen

Cluster-Algorithmen wählen

- Hierarchische Clusteranalyse mit Single-Linkage- oder Complete-Linkage-Verfahren
- Partionierende Clusteranalyse mit K-Means-Algorithmus

Unterschiede zur Faktorenanalyse

Zentraler Unterschied: Dimension, in welche die Daten vereinfacht werden.

		Faktor A					Faktor B						Faktor C			Faktor D				
	Fall	A1	A2	A3	A4	A5	B1	B2	B3	B4	B5	B6	C1	C2	C3	D1	D2	D3	D4	...
Cluster 1	100	1	2	3	2	1	4	1	3	3	3	2	4	2	2	2	2	4	5	
	101	3	3	2	2	3	4	1	4	3	2	2	3	2	2	3	1	3	3	
	102	3	3	5	4	2	3	3	3	5	1	5	5	2	2	2	2	3	4	
	103	2	4	4	3	5	4	3	1	1	4	3	4	3	2	4	4	3	4	
	104	4	5	3	5	3	3	2	1	2	4	4	2	2	4	1	2	4	5	
	105	3	1	2	4	4	1	5	3	4	2	2	1	4	1	2	2	2	4	
Cluster 2	106	4	2	2	3	1	2	3	5	1	4	3	1	3	1	2	4	2	2	
	107	3	3	5	5	3	2	5	2	4	4	4	2	1	2	4	2	4	4	
	108	2	2	5	4	2	5	3	3	3	3	1	4	2	2	4	2	3	3	
	109	4	1	4	5	3	2	3	1	1	4	2	4	4	4	2	2	4	5	
	110	3	4	4	3	5	4	2	2	5	4	3	2	1	2	2	4	2	1	
	111	4	1	3	2	3	5	3	2	2	4	3	4	5	5	4	4	3	3	
	...																			

Unterschiede zur Faktorenanalyse

Faktorenanalyse

- Zusammenfassen vieler Items (Variablen) zu wenigen Faktoren / Indizes
- Erkennen von Dimensionen hinter Antwort-Mustern als Korrelation der Items
- Prüfen des Zusammenhangs zwischen Gruppen von Items

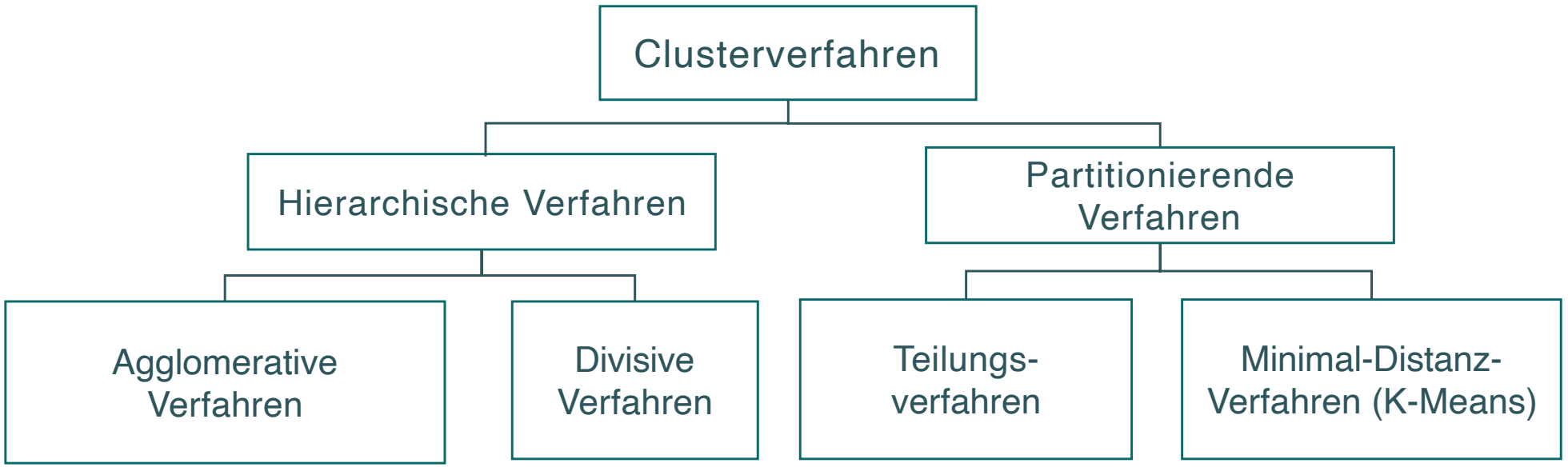
Clusteranalyse

- Zusammenfassen vieler Fälle zu wenigen Gruppen
- Erkennen von Typen mit ähnlichem Antwort-Verhalten
- Prüfen der Heterogenität der untersuchten Stichprobe oder Teilendavon

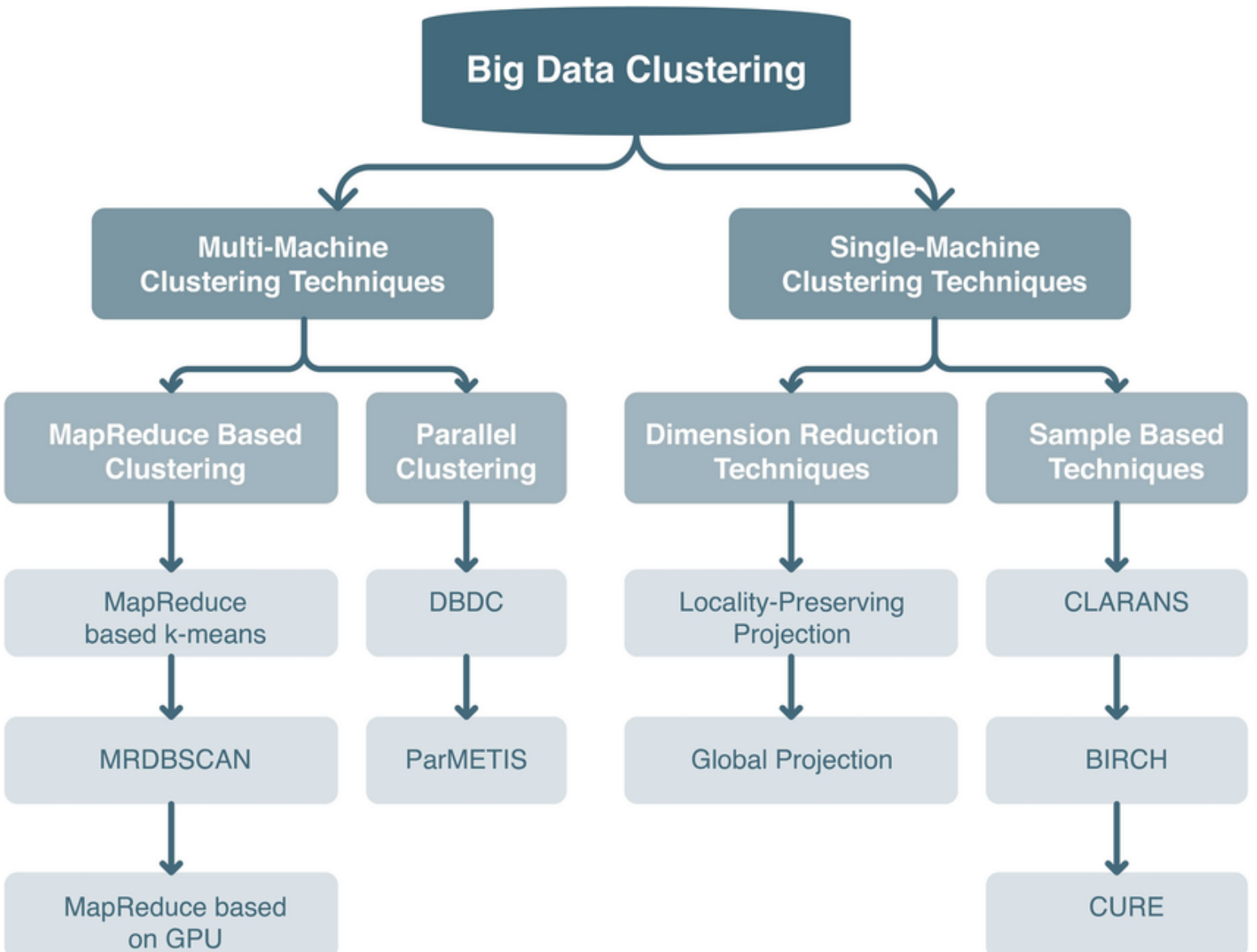
Faktorenanalyse vor Clusteranalyse

Da Clusteranalysen ein verzerrendes Gewicht auf hoch korrelierte Variablen legen, führt man häufig vor der Clusteranalyse Faktorenanalysen durch.

Systematik der Clusteranalyseverfahren



Big-Data-Clustering-Taxonomy



2 Hierarchische Clusteranalyse

Ablauf

1. Jeder Fall ist sein eigenes Cluster (2468 Befragte → 2468 Cluster)
2. Für jede Paarung werden die Ähnlichkeitsmasse berechnet
3. Die beiden Fälle geringster Distanz werden zusammengelegt
4. Es werden wieder die Abstände zwischen dem neuen Cluster und den übrigen berechnet
5. Schritt 3 und 4 so lange wiederholen bis alle Objekte in einem Cluster sind

Das sind bei 2468 Fällen 2467 Fusionierungsschritte und schon im ersten Schritt 3'044'278 Vergleiche, also um die 4.6^{12} Vergleiche.

2.1 Vorteile und Nachteile

Vorteile

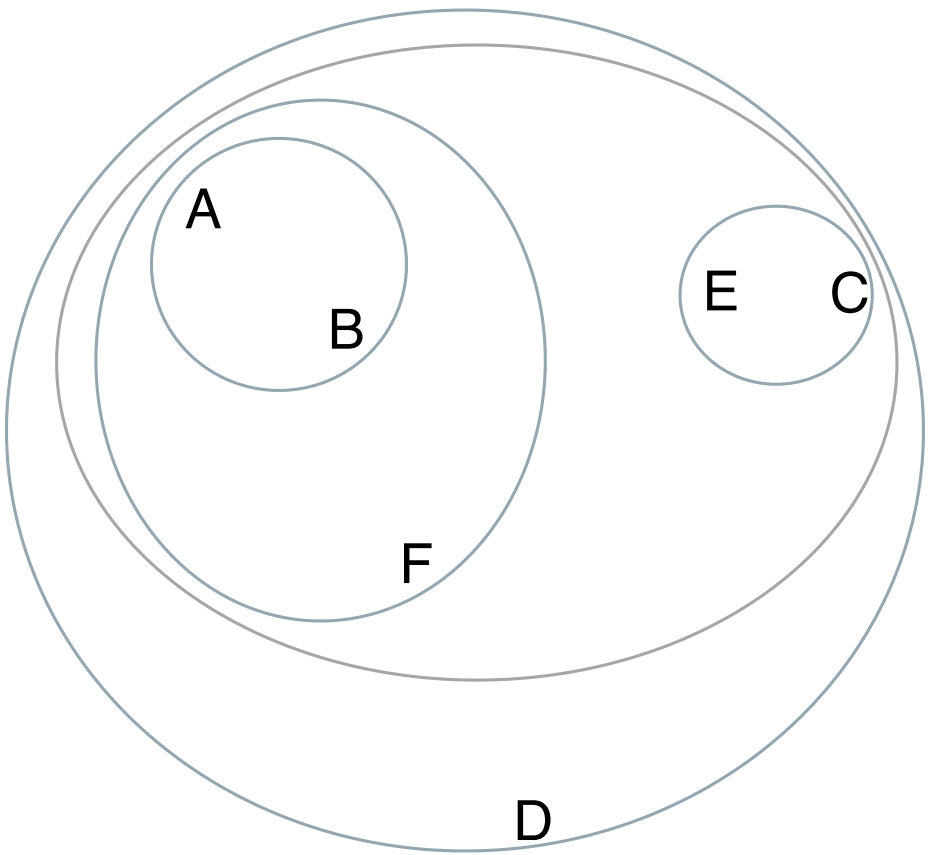
- Kann auch mit kategorialen Variablen und ordinalen gerechnet werden.
- Einfach identifizierbare Ausreisser
- Kann gut angepasst werden

Nachteile

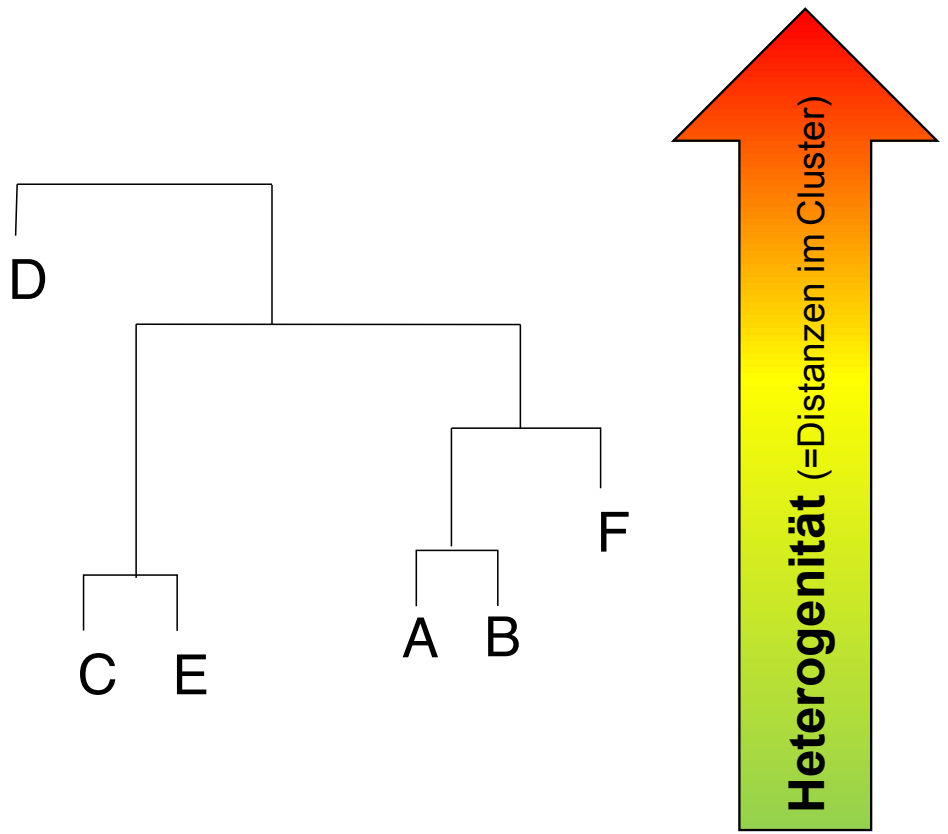
- Es müssen immer jeweils paarweise Distanzen auf jeder Stufe berechnet werden
- Sehr rechenaufwendig und damit viel langsamer als k-means-Cluster
- Viele Masse und Kennwerte

2.2 Cluster-Dendrogramm der hierachischen CA

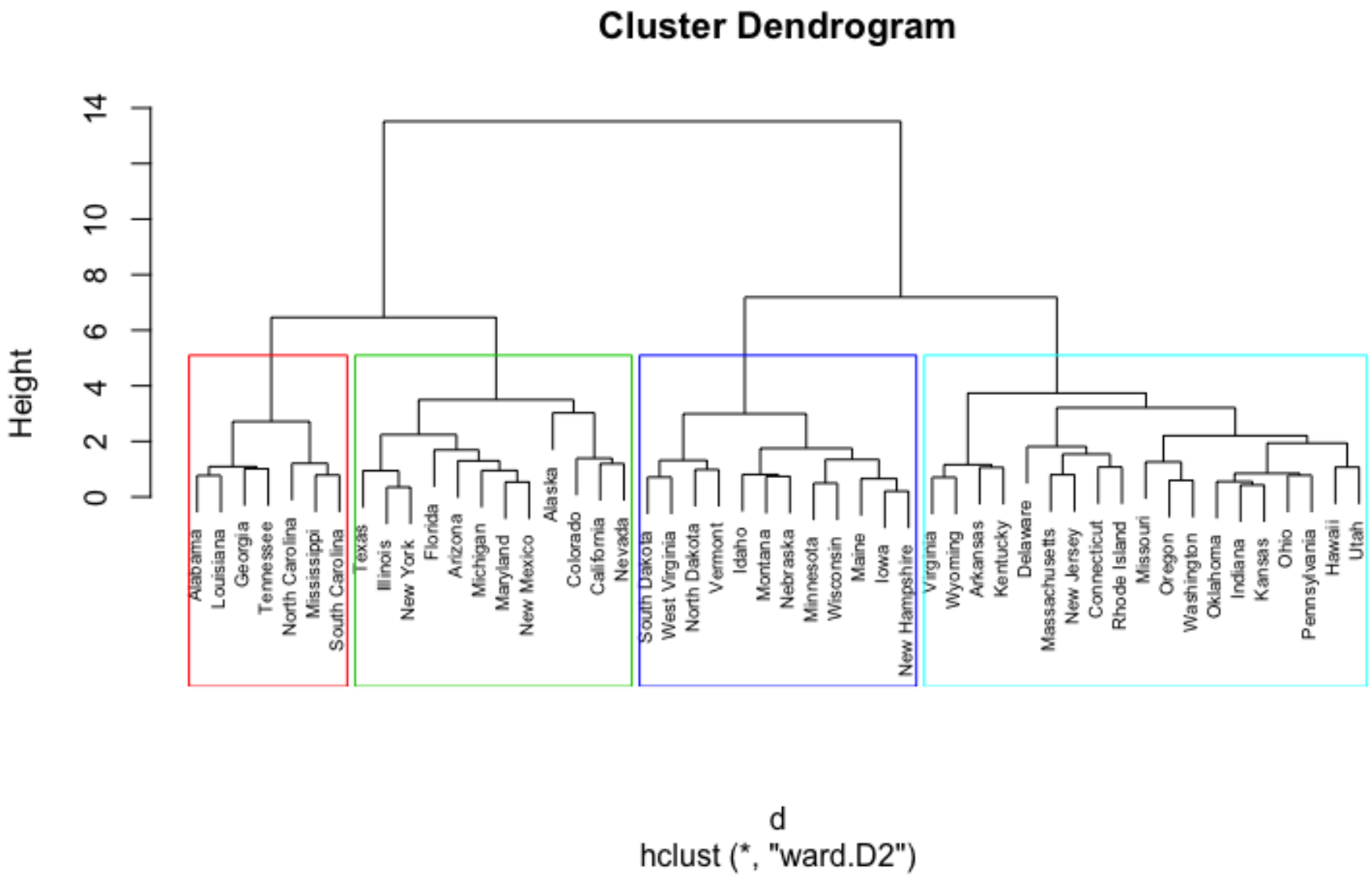
Elemente:



Dendrogramm:

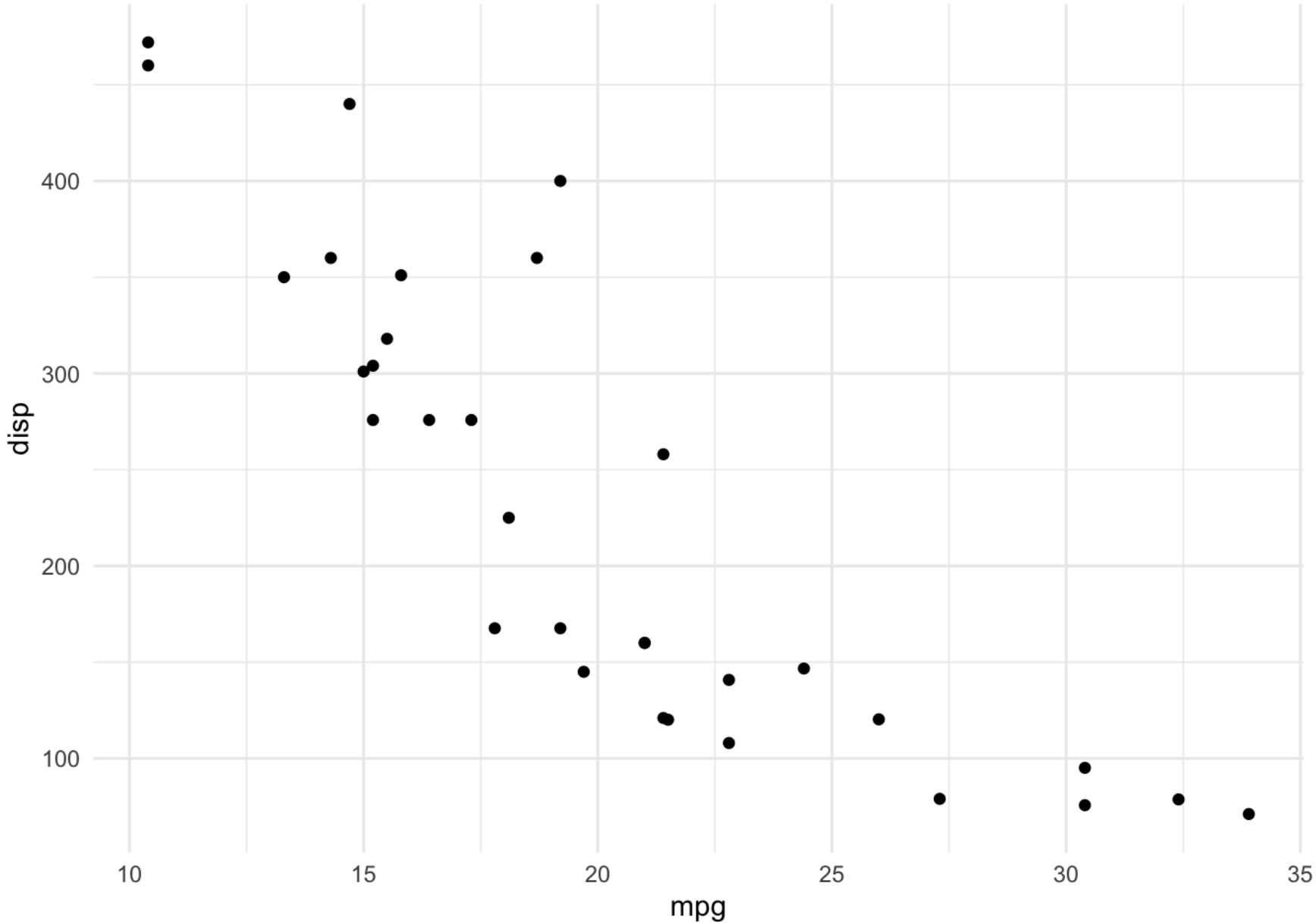


Beispiel

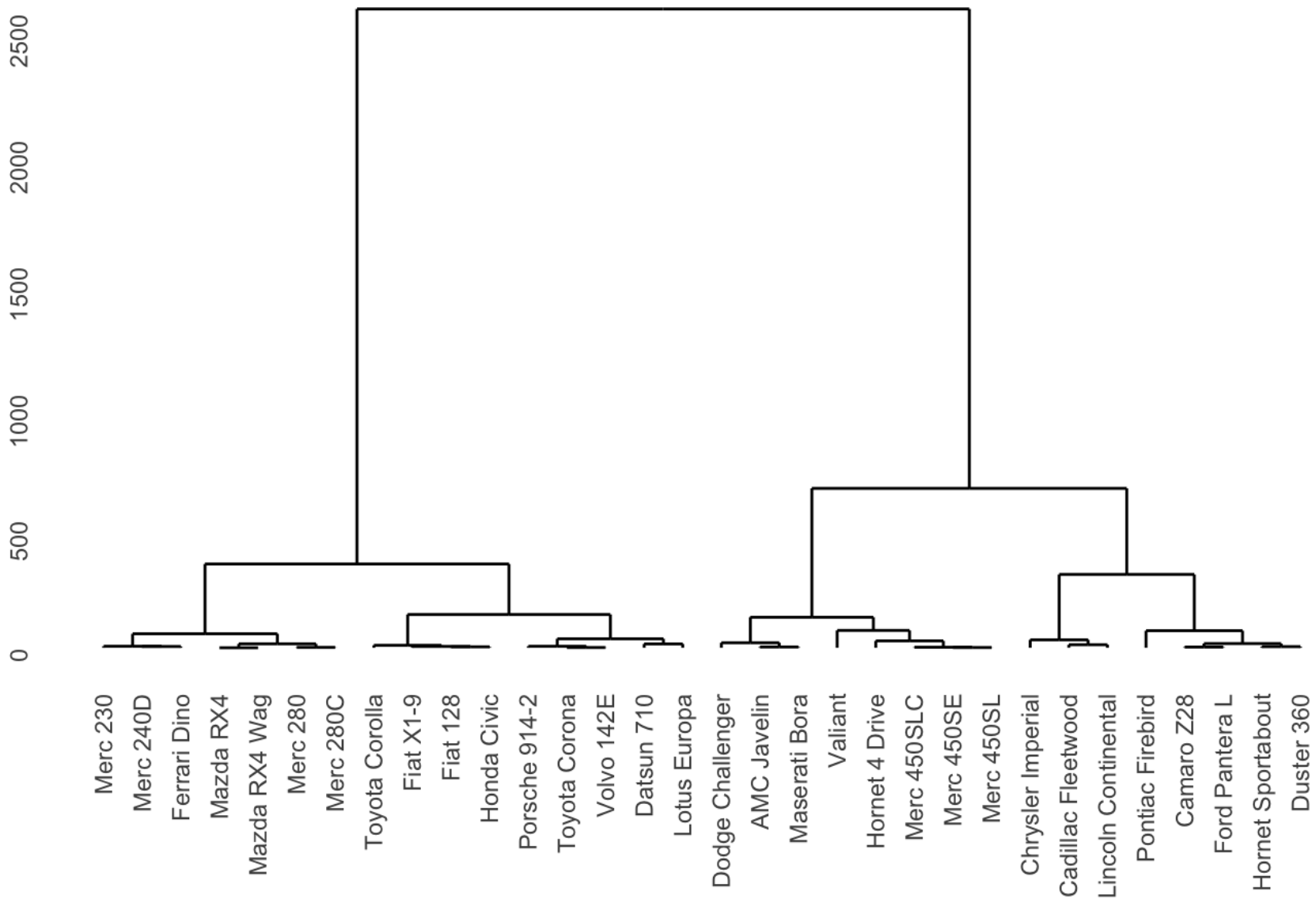

Cluster Dendrogram
Height
0
2
4
6
8
10
14
Alabama
Louisiana
Georgia
Tennessee
North Carolina
Mississippi
South Carolina
Texas
Illinois
New York
Florida
Arizona
Michigan
Maryland
New Mexico
Alaska
Colorado
California
Nevada
South Dakota
West Virginia
North Dakota
Vermont
Idaho
Montana
Nebraska
Minnesota
Wisconsin
Maine
Iowa
New Hampshire
Virginia
Wyoming
Arkansas
Kentucky
Delaware
Massachusetts
New Jersey
Connecticut
Rhode Island
Missouri
Oregon
Washington
Oklahoma
Indiana
Kansas
Ohio
Pennsylvania
Hawaii
Utah
d
hclust (*, "ward.D2")

Hierarchische Clusteranalyse in R

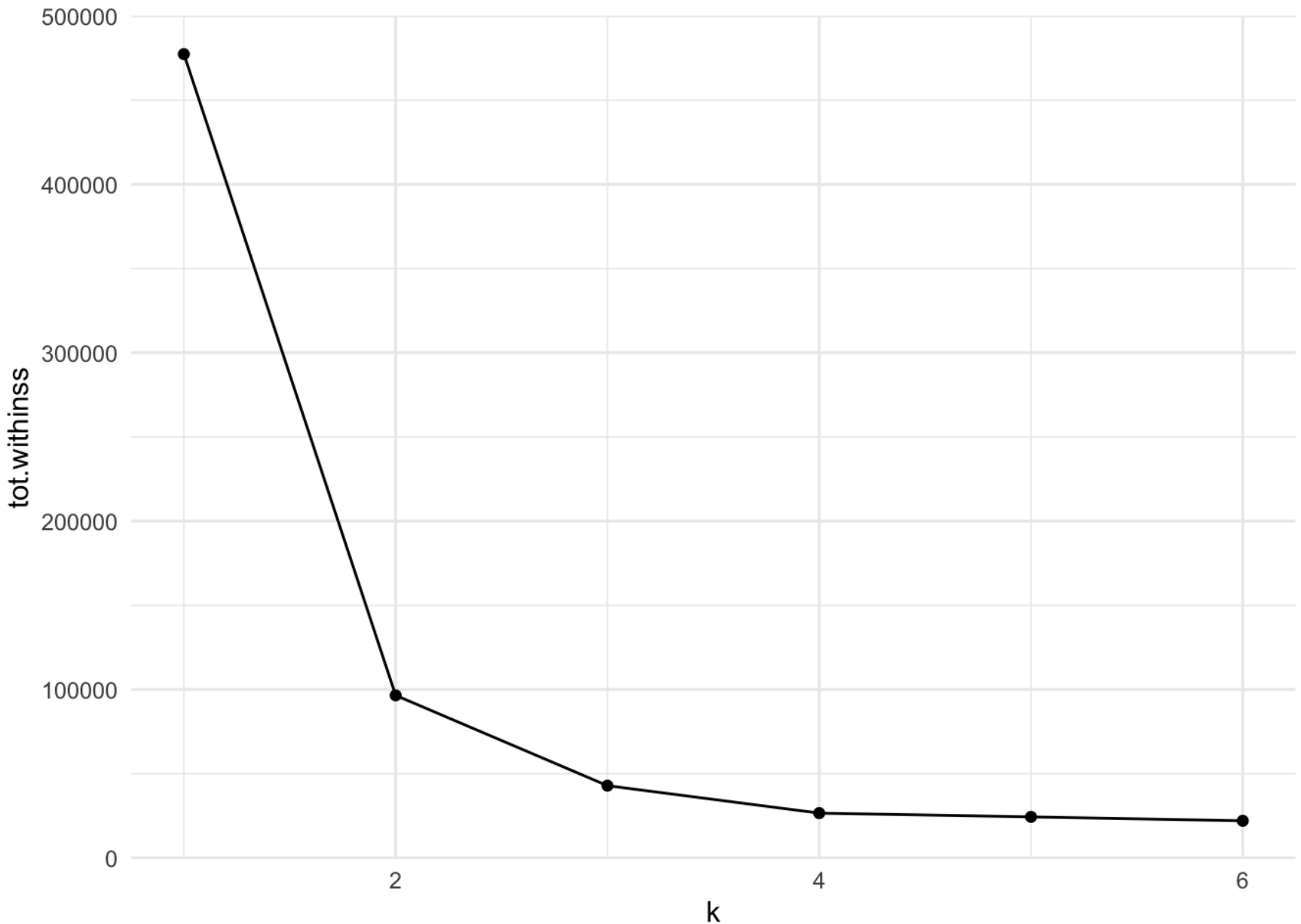
Autos nach Verbrauch (Miles/Gallon und Hubraum)



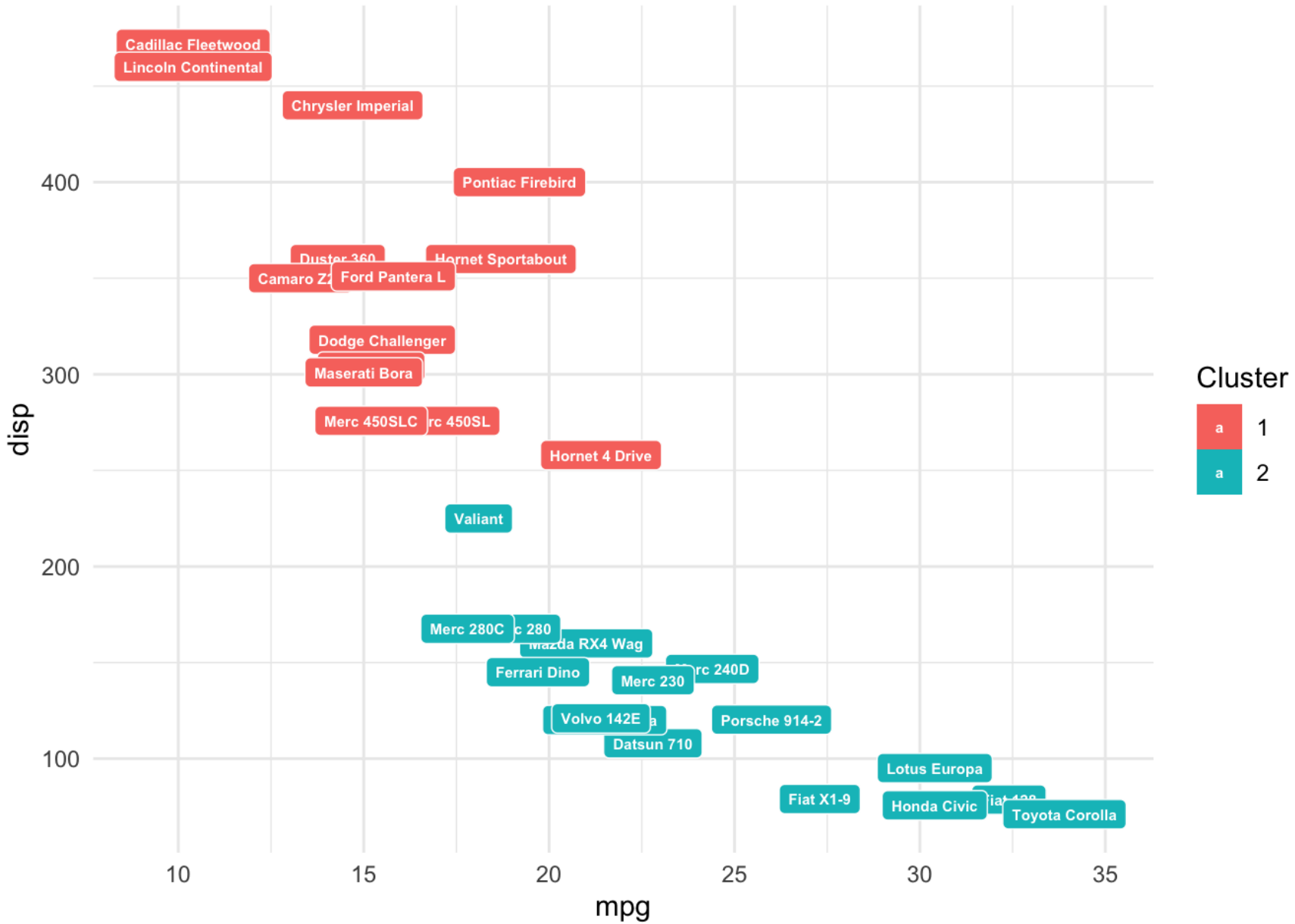
Dendrogramm



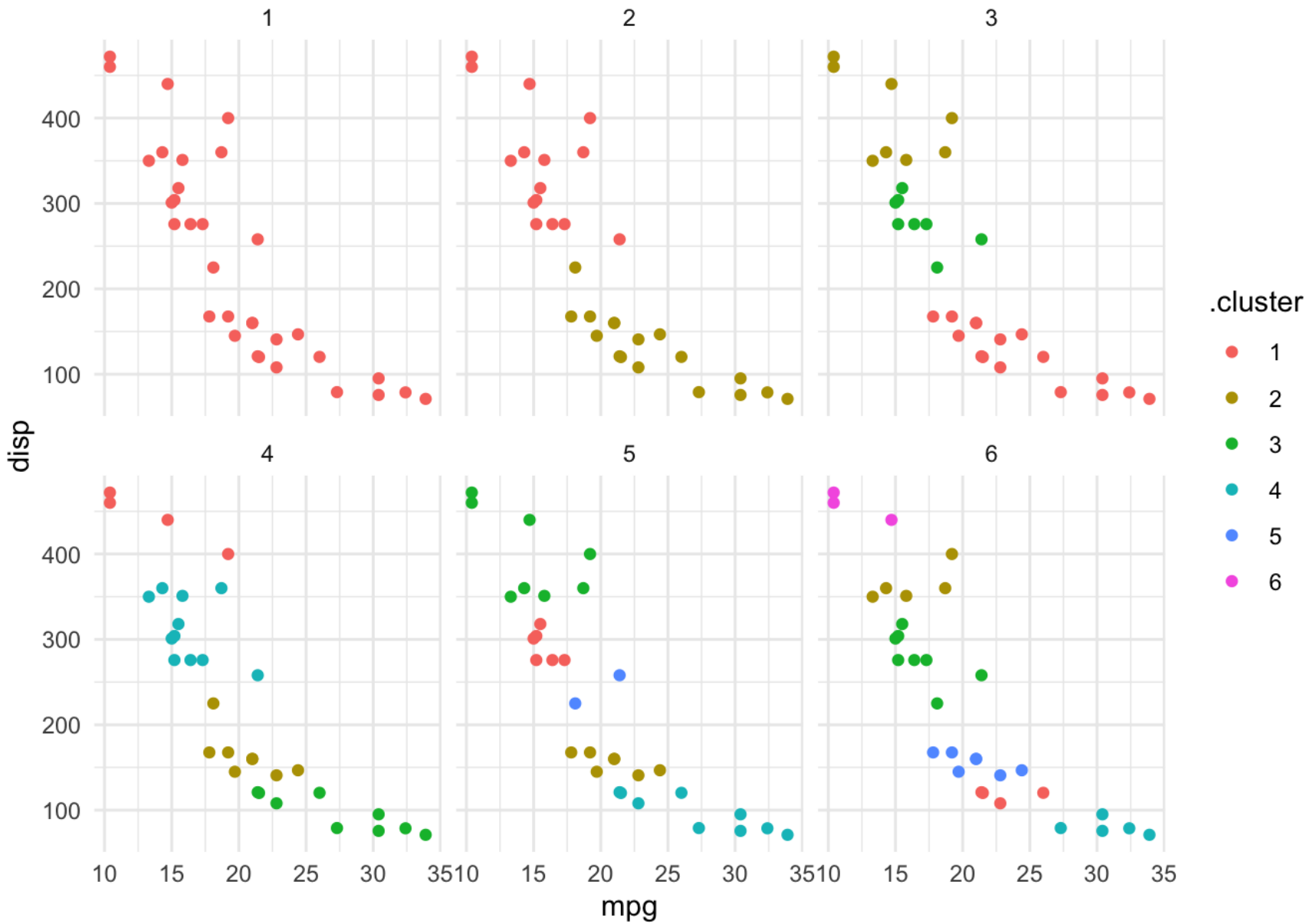
Screeplot der Quadratsumme innerhalb der Cluster



Cluster


Cadillac Fleetwood
Lincoln Continental
Chrysler Imperial
Pontiac Firebird
Hornet Sportabout
Ford Pantera L
Dodge Challenger
Maserati Bora
Merc 450SLC
Hornet 4 Drive
Valiant
Merc 280C
Ferrari Dino
Merc 230
Volvo 142E
Porsche 914-2
Datsun 710
Lotus Europa
Fiat X1-9
Honda Civic
Toyota Corolla
400
300
200
100
disp
10
15
20
25
30
35
mpg
Cluster
1
2

Vergleich nach Clustergrössen



3 K-Mean-Clustering

Ablauf

0. Clustervariablen festlegen
1. Bestimmung der Anzahl Cluster
2. zufällige Startpartition der Cluster anlegen
3. nächste Elemente der Cluster bestimmen
4. Clusterzentren neu ausrichten
5. nächste Elemente der Cluster bestimmen
6. **iterativ Clusterzentren immer wieder neu ausrichten (4.) und zugehörige Elemente bestimmen (5.)**
7. wenn sich nichts mehr tut, enden
8. Interpretation der Clusterlösung

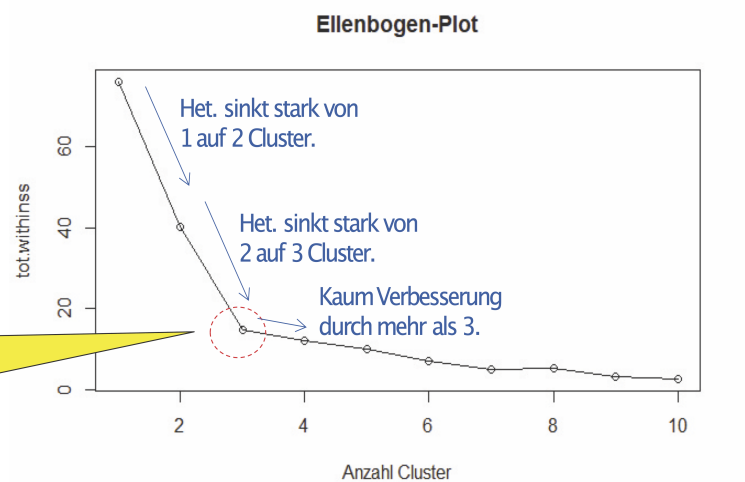
3.1 Voraussetzung

- Die Variablen für die Clusteranalyse müssen metrisch skaliert sein. Kategoriale und ordinale Variablen (mit Informationsverlust) können aber in Dummys umgewandelt werden.
- Die Variablen sollten ähnliche Standardabweichungen haben, können dafür aber z-transformiert werden oder Faktoren einer FA sein (die sind z-transformiert)
- Die Variablen sollten nicht zu stark korrelieren. Bei höheren Korrelationen bietet sich eine vorherige FA an.

3.2 Clusterzahl bestimmen

Es werden k-mean-Clusteranalysen für 1-viele durchgeführt und dann der Knick (Ellbogen) gesucht.

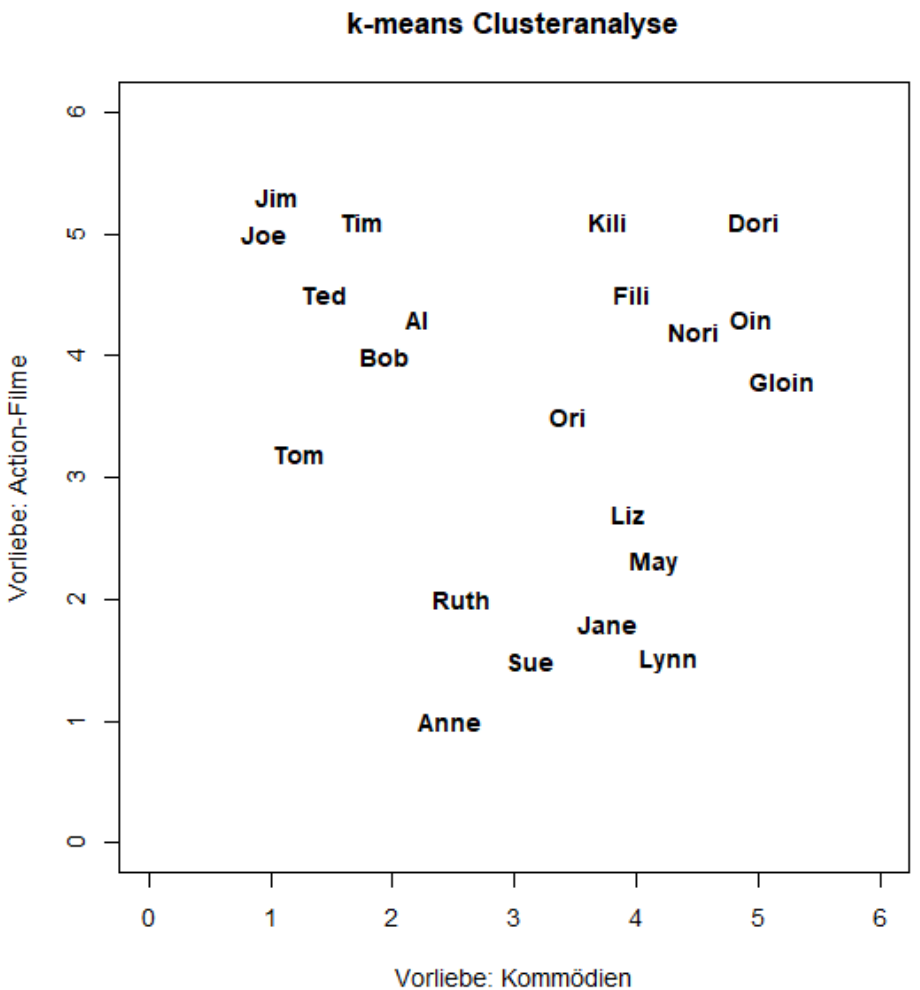
Hier ist der Ellenbogen. Also gibt es **genau drei Cluster**. Die Lösung, die oben schrittweise gefunden wurde, ist also die beste Clusterlösung.
Achtung: Die korrekte Anzahl ist genau der Knick. Nicht wie beim Scree–Kriterium!



Gütemass der Lösung als R^2

R^2 als $\frac{between_{SS}}{total_{SS}}$ wie Varianzaufklärung.

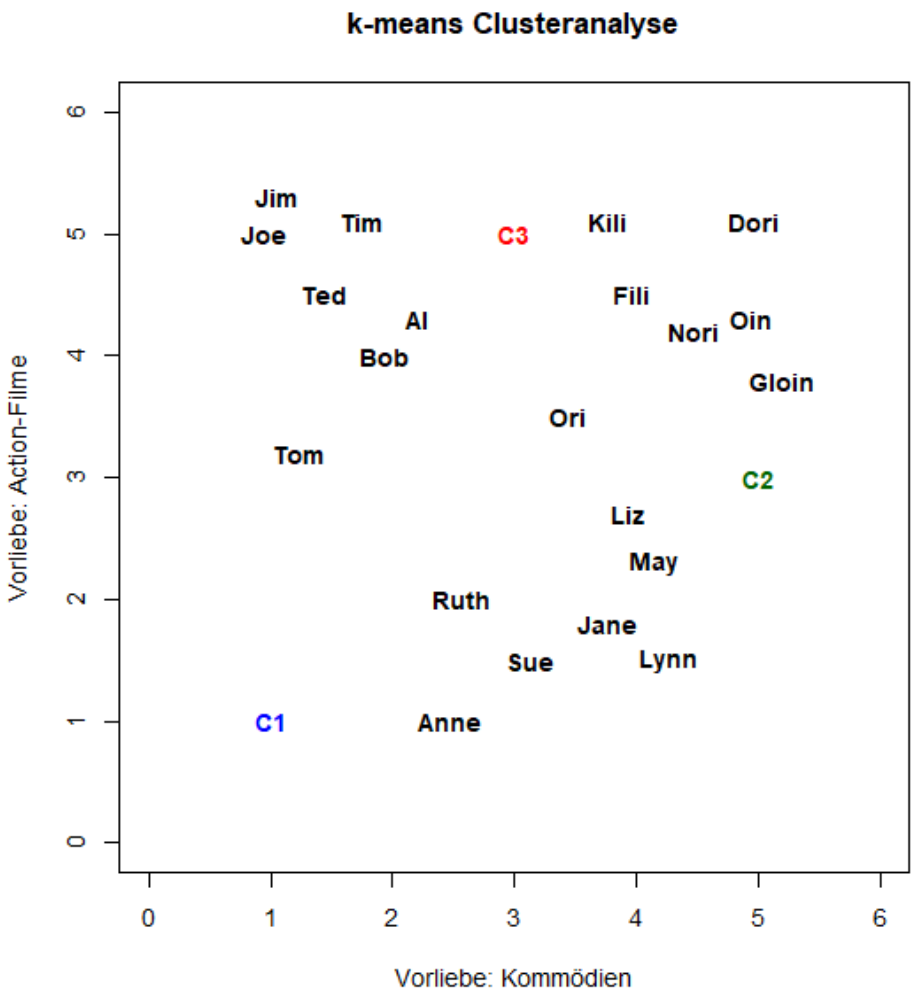
3.3 K-Means-Clusteranalyse Iterationen



Ausgangslage für die Cluster-Analyse

- 21 Elemente, die bezüglich zwei metrischen Merkmalen geclustert werden sollen.
- Für jedes Element wurde jedes Merkmal gemessen, so können die Fälle auf einem Koordinatensystem eingetragen werden.
- Als Distanzmass wird die euklidische Distanz verwendet. Das ist genau die Distanz, die man von Auge sieht.

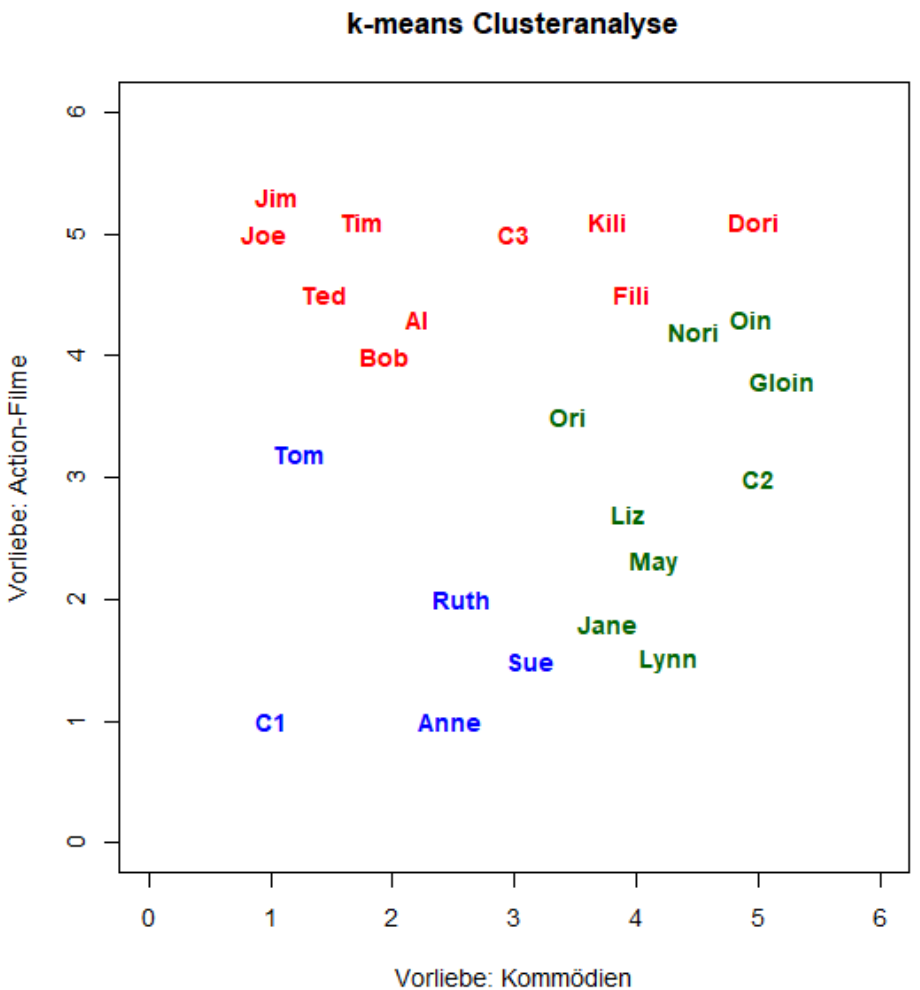
K-Means-Cluster-Algorithmus



Schritt 1: Cluster-Zentren

- Es wird eine feste Anzahl (k) von Clusterzentren definiert, die irgendwo zufällig verteilt werden
- Die Cluster-Zentren müssen nicht in der Nähe der tatsächlichen Cluster sein.
- Hier heissen die Zentren: C1, C2 und C3 und sind absichtlich ausserhalb der von Auge sichtbaren Cluster gesetzt.

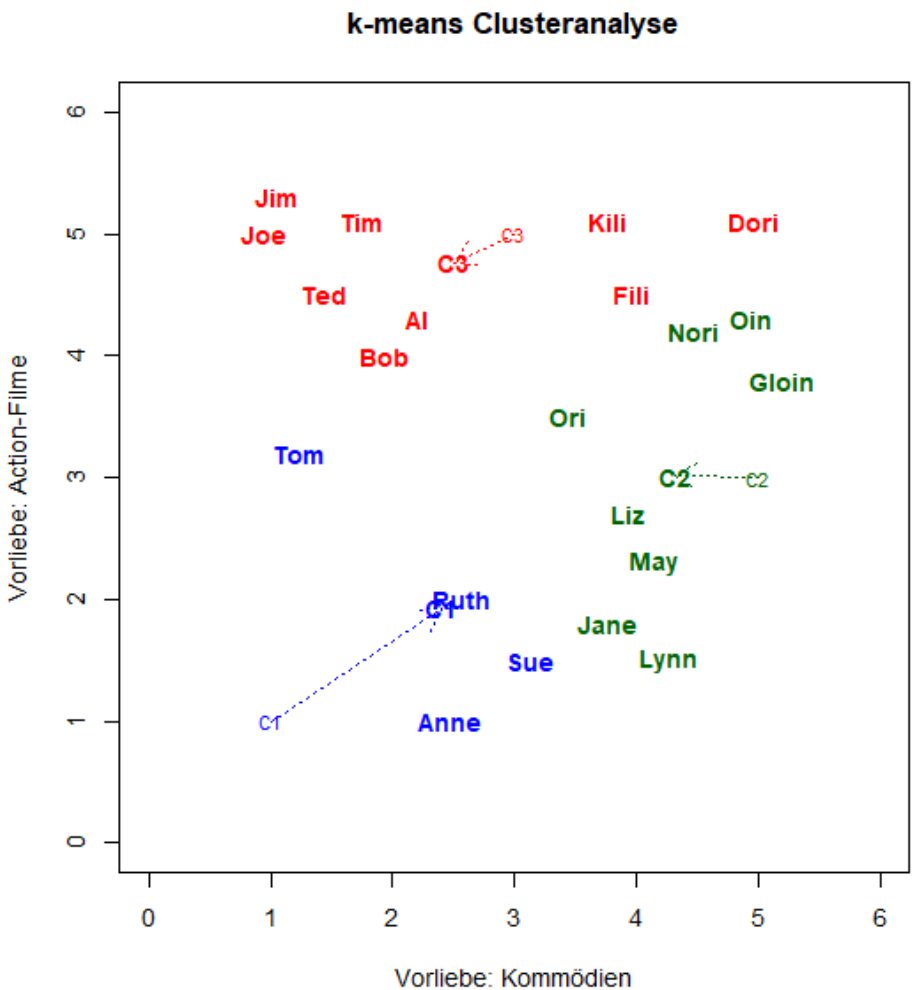
K-Means-Cluster-Algorithmus



Schritt 2: Elemente zuordnen

- Jedes Element wird dem Cluster zugeordnet, zu dessen Zentrum es den geringsten Abstand hat.
- Dies ist die erste Lösung der k-Means Analyse (schlechte Lösung)
- Die Cluster-Zentren liegen nun aber nicht im Zentrum der Cluster

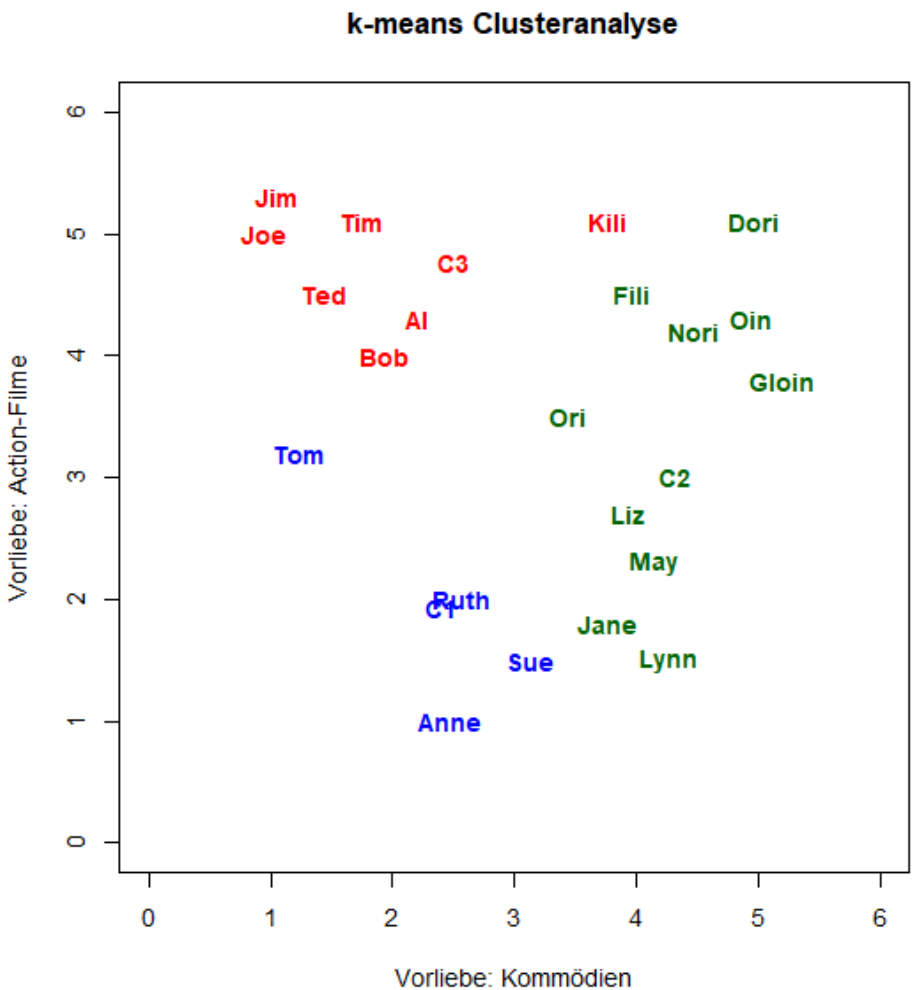
K-Means-Cluster-Algorithmus



Schritt 3: Zentren zentrieren

- Die tatsächlichen Zentren der Cluster werden berechnet, indem man den Mittelpunkt aller Elemente berechnet, die zu jedem Cluster gehören.
- Die Cluster-Zentren werden dort hin verschoben.
- Nun haben sie sich aber näher an einige Elemente bewegt und sich von anderen entfernt

K-Means-Cluster-Algorithmus

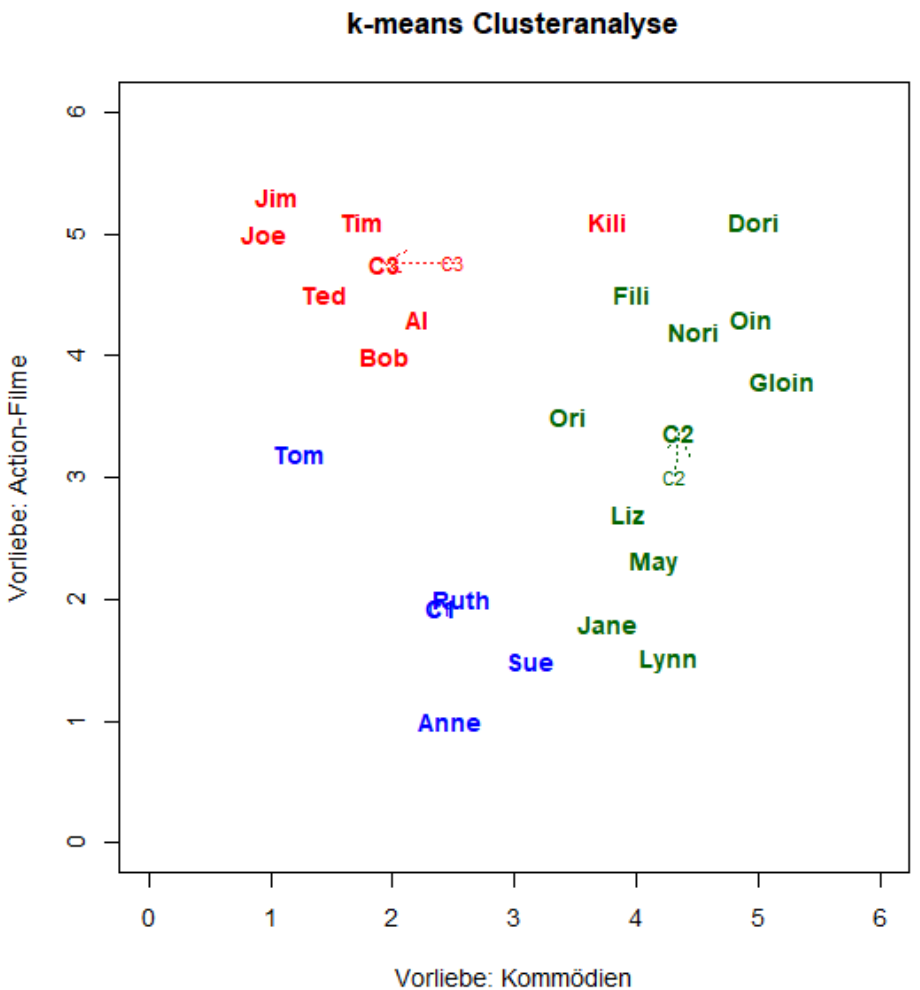


nochmals Schritt 2:

Elemente zuordnen

- Jedes Element wird nun wieder dem Zentrum zugeordnet, das ihm jetzt am nächsten liegt.
- Dori und Fili haben von C3 nach C2 gewechselt
- Nun sind die beiden Clusterzentren C2 und C3 schon wieder am falschen Ort

K-Means-Cluster-Algorithmus

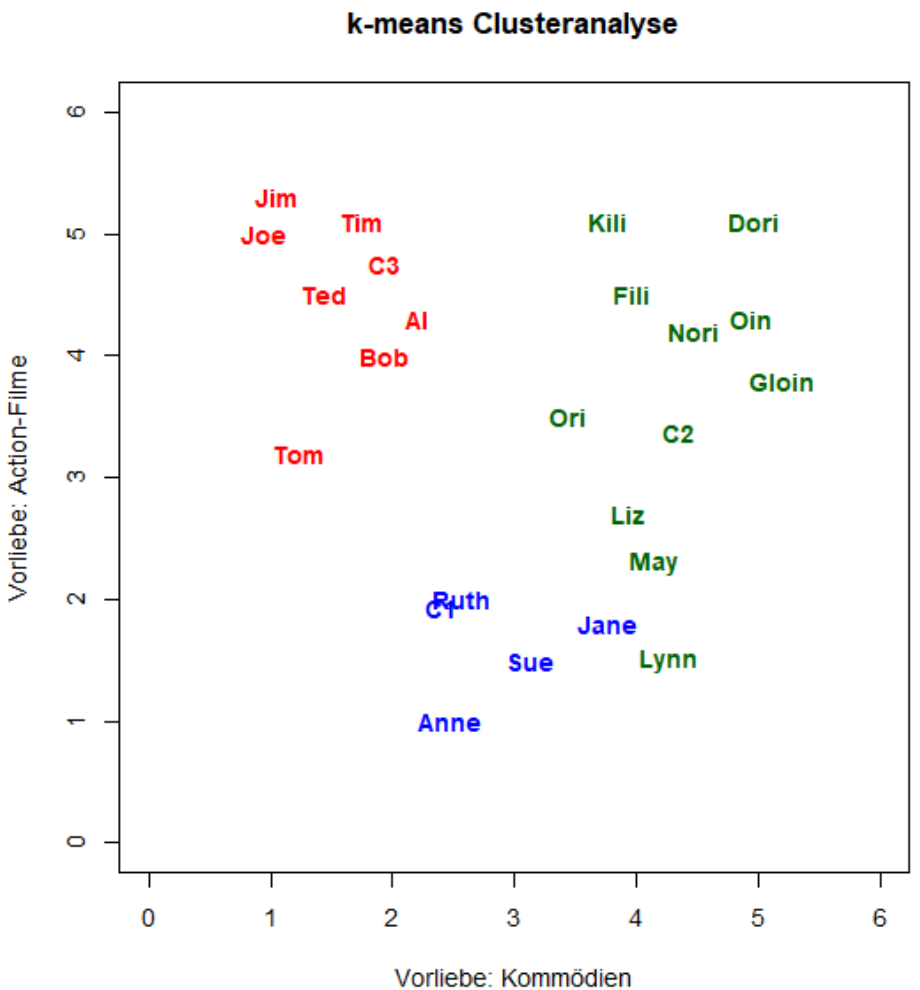


nochmals Schritt 3:

Zentren korrigieren

- Die Clusterzentren C2 und C3 verschieben sich noch einmal, damit sie wieder genau in der Mitte ihres Clusters liegen.
- Nun stimmen die Elemente wieder nicht mehr

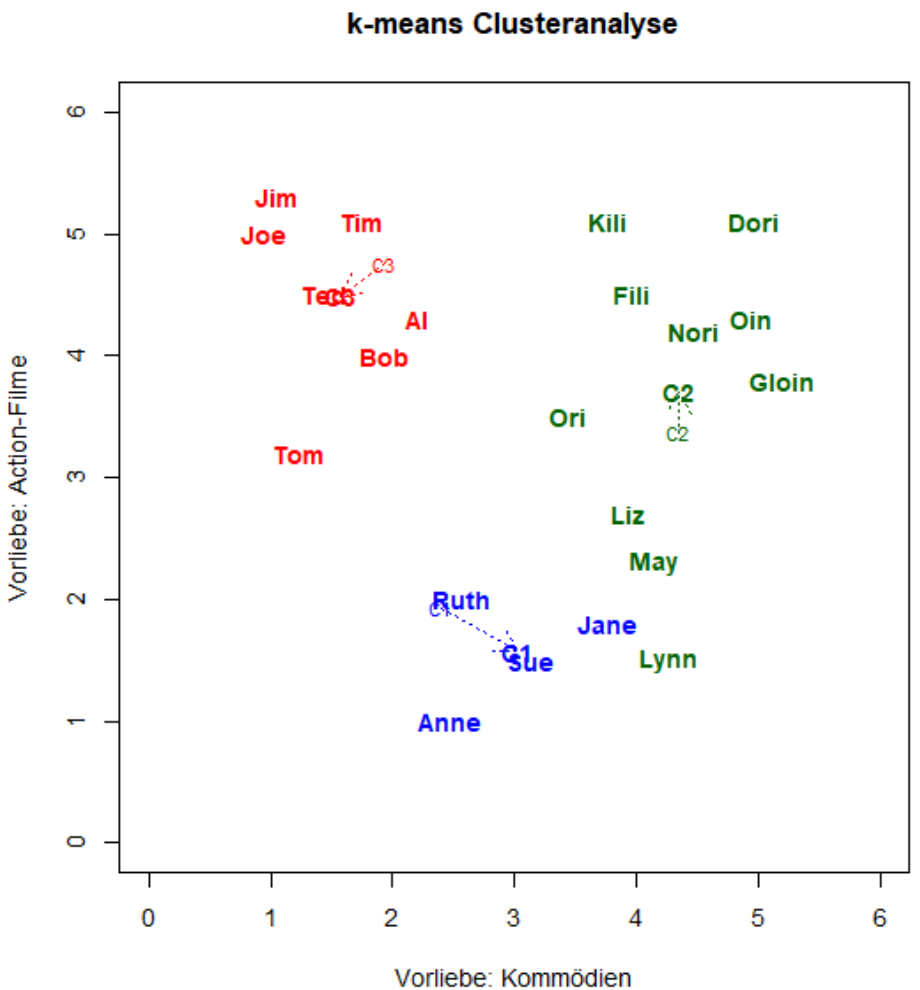
K-Means-Cluster-Algorithmus



nochmals Schritt 2: Elemente zuordnen

- Jane ist jetzt im Cluster 1, Kili im Cluster 2, Tom in Cluster 3
- Damit stimmen die Zentren wieder nicht mehr

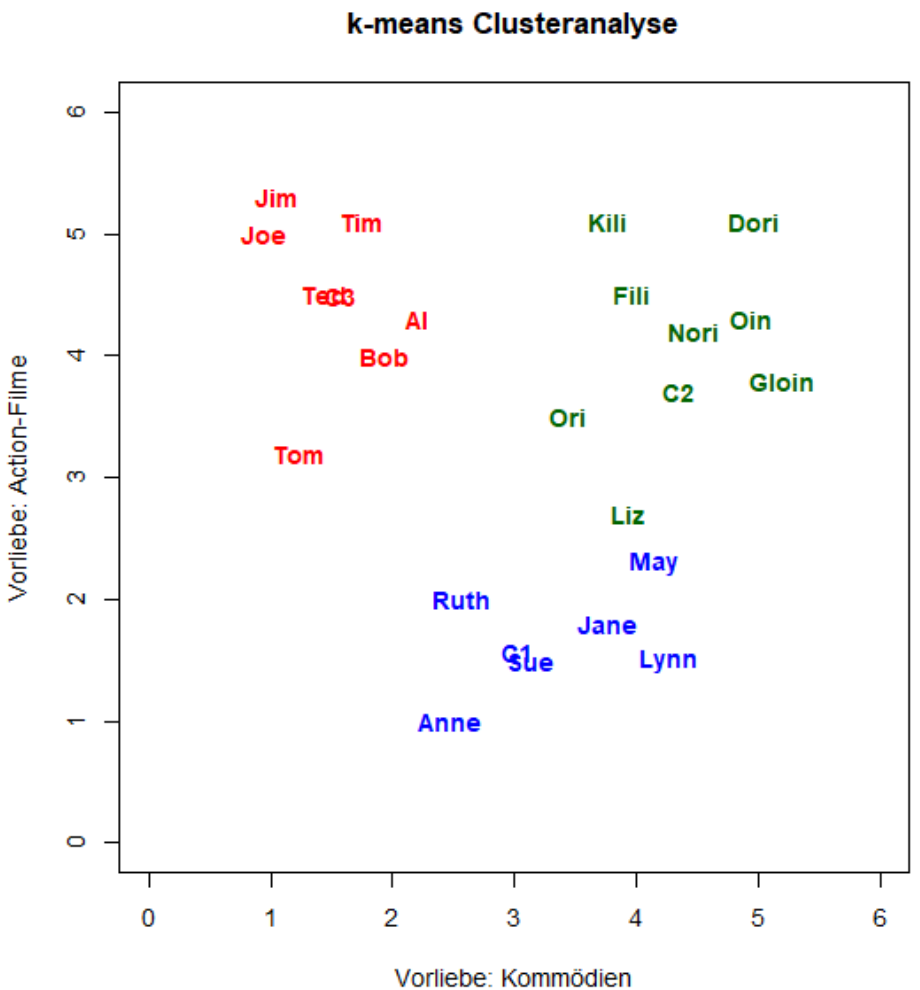
K-Means-Cluster-Algorithmus



nochmals Schritt 3: Zentren korrigieren

- Alle Cluster-Zentren verschieben sich etwas, um wieder in der Mitte der Elemente zu liegen.
- Nun stimmen die Elemente wieder nicht mehr.

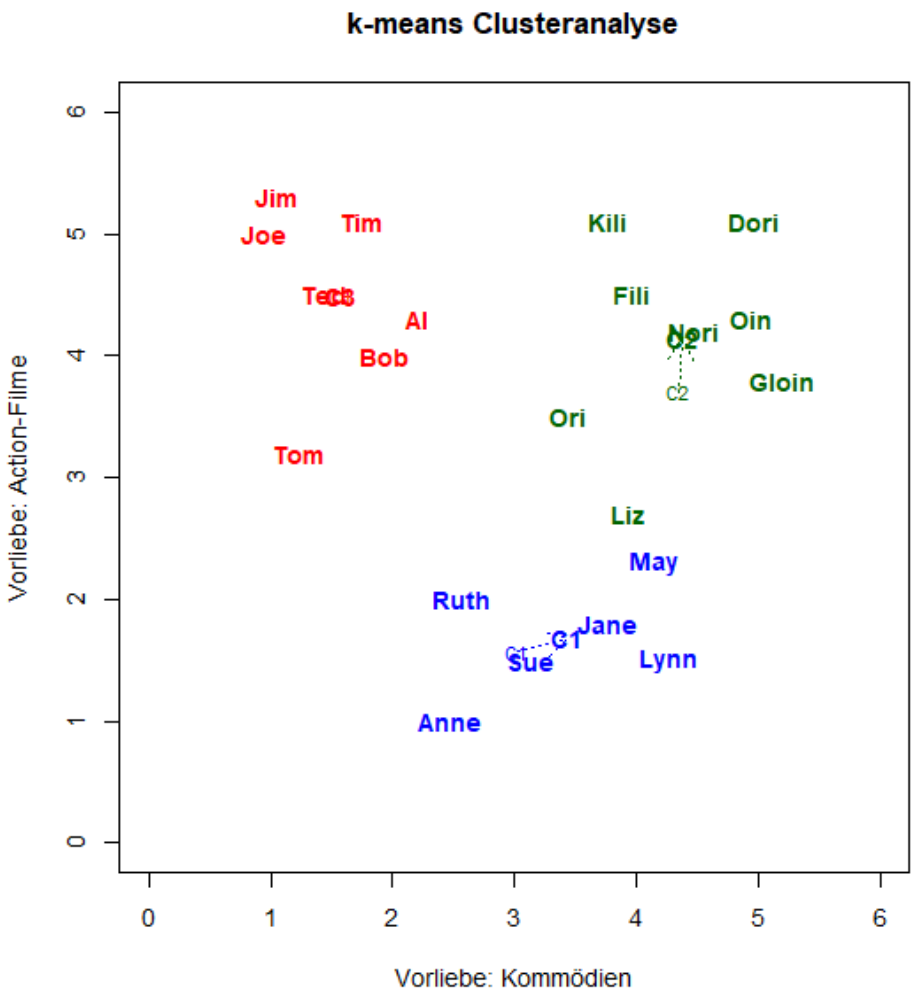
K-Means-Cluster-Algorithmus



nochmals Schritt 2: Elemente zuordnen

- May und Lynn gehören neu zu Cluster 1 und nicht mehr zu Cluster 2.
- Damit stimmen C1 und C2 wieder nicht mehr!

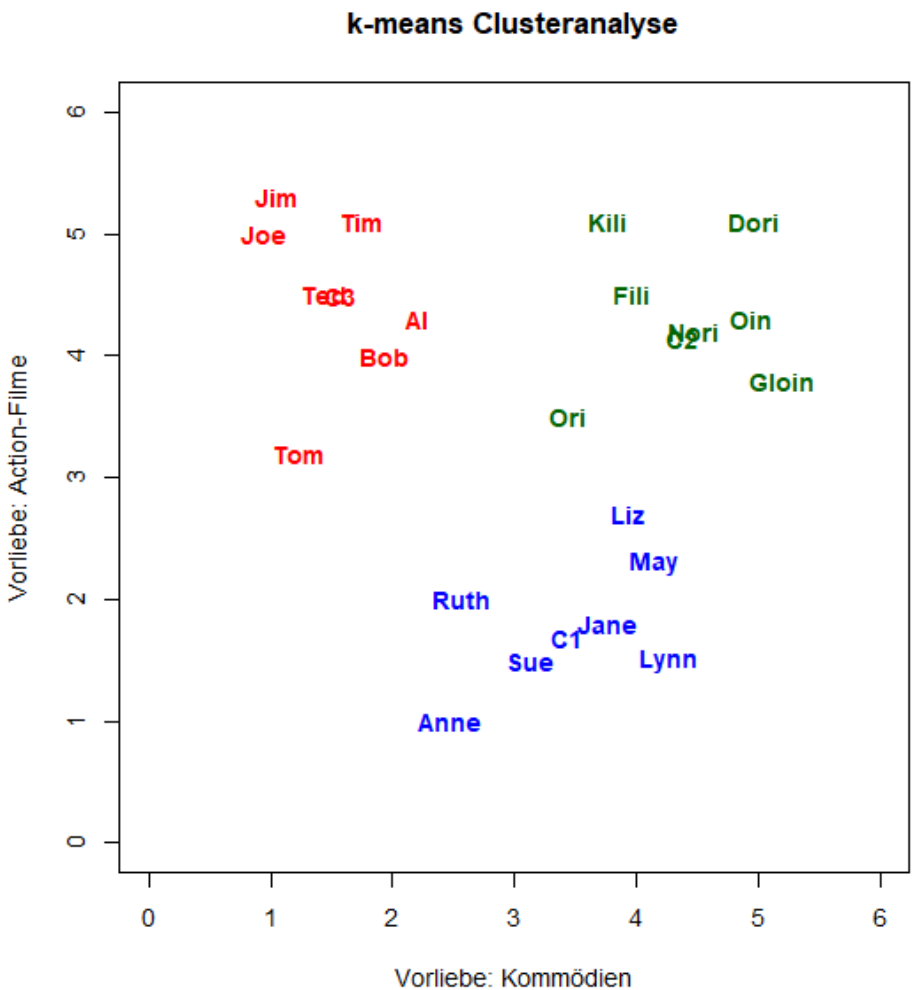
K-Means-Cluster-Algorithmus



nochmals Schritt 3: Zentren korrigieren

- C3 bleibt, wo es ist. Aber die anderen beiden müssen korrigiert werden.
- Nun stimmen deren Elemente wieder nicht mehr

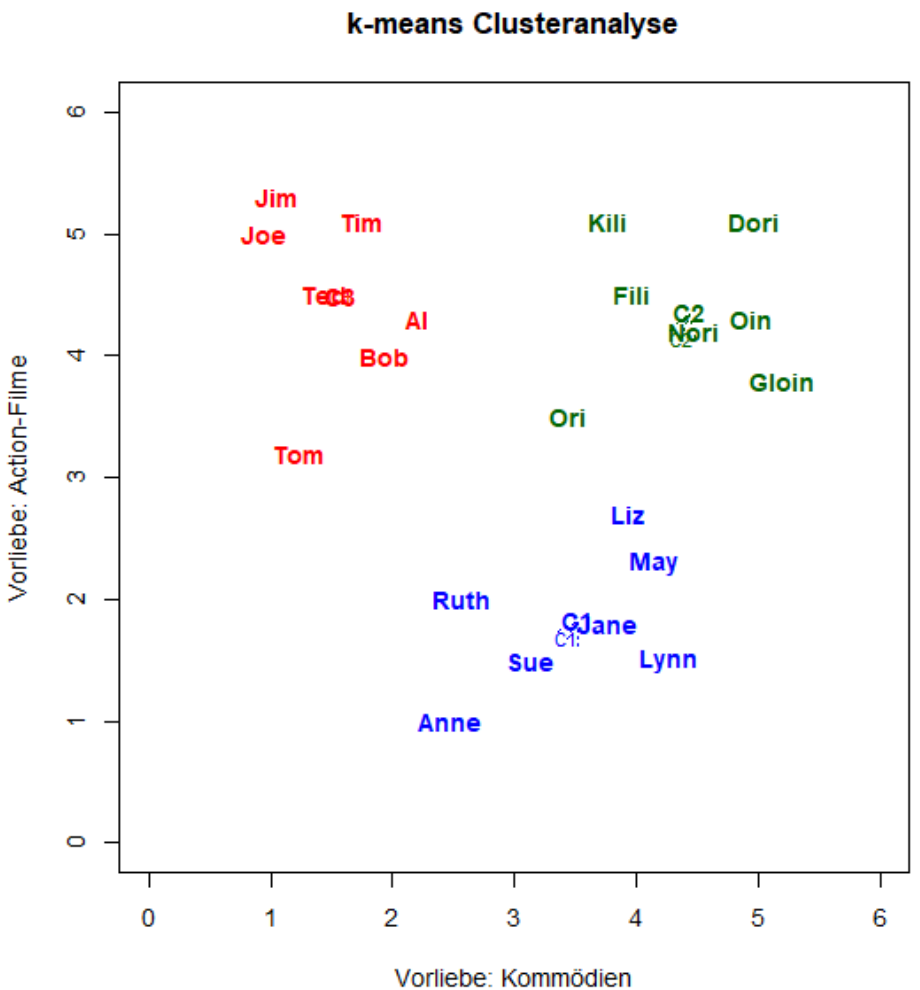
K-Means-Cluster-Algorithmus



nochmals Schritt 2: Elemente zuordnen

- Liz gehört neu zu C1, alle anderen stimmen soweit.
- Nun stimmen die Zentren nicht mehr

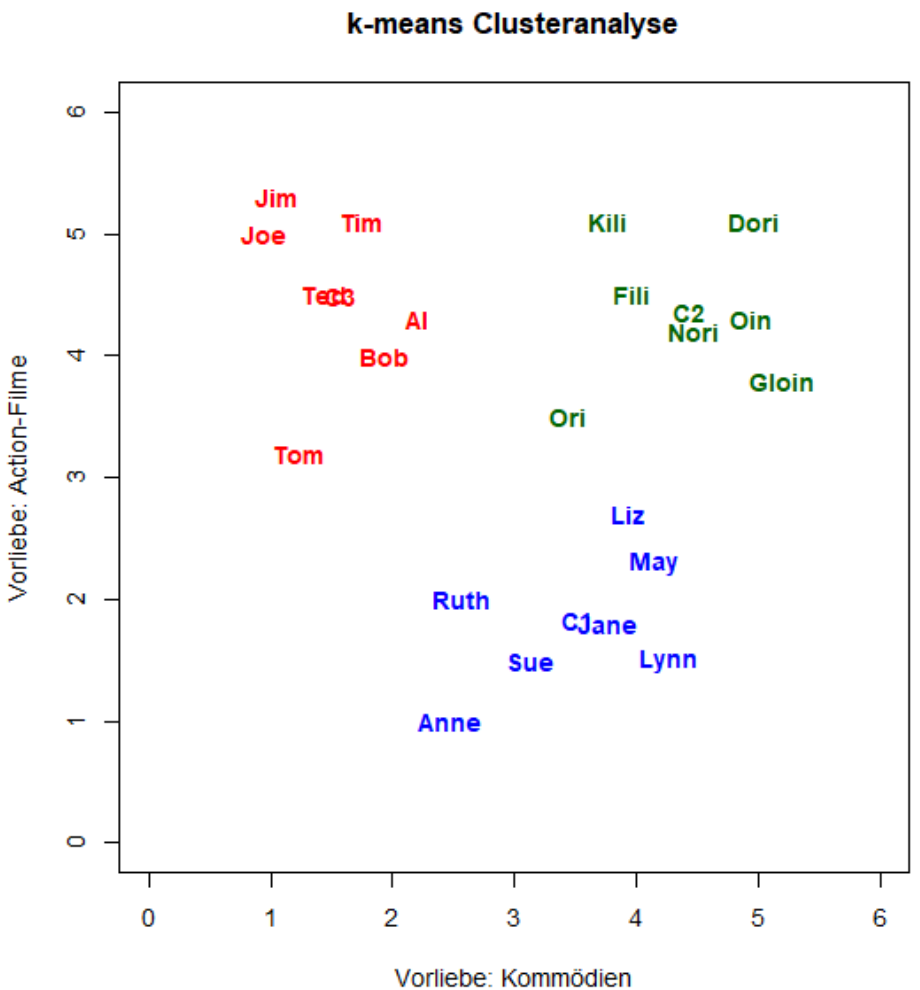
K-Means-Cluster-Algorithmus



nochmals Schritt 3: Zentren korrigieren

- C1 und C2 werden ganz wenig korrigiert.
- Stimmen die Elemente noch?

K-Means-Cluster-Algorithmus



nochmals Schritt 2: Elemente zuordnen

- Es gibt keine Änderungen mehr. Alle Elemente bleiben in den Clustern, denen sie zugeordnet waren.
- Jetzt stimmen auch die Zentren
- Endlösung ist erreicht: Die drei Zentren liegen genau im Zentrum der Elemente, die ihnen am nähesten sind.

Der k-means Algorithmus

In Worten

Am Anfang die Anzahl (k) der Cluster festlegen. Dann werden Clusterzentren zufällig in den Variablenraum gelegt. Dann wird jeder Fall seinem nächsten Clusterzentrum zugeordnet. Dann das Clusterzentrum in den Mittelpunkt seines Clusters verlegt. Wieder werden alle Fälle ihren jeweils nächsten Clustern zugeordnet. Dann wieder Verschiebund der Clusterzentren usw. bis kein Fall mehr sein Cluster wechselt und keine Verschiebung der Clusterzentren mehr stattfindet.

Gütemass ist die Summe quadrierter Clusterzentrenabweichungen.

Clusteranalyse schöner mit `factoextra`

- Code

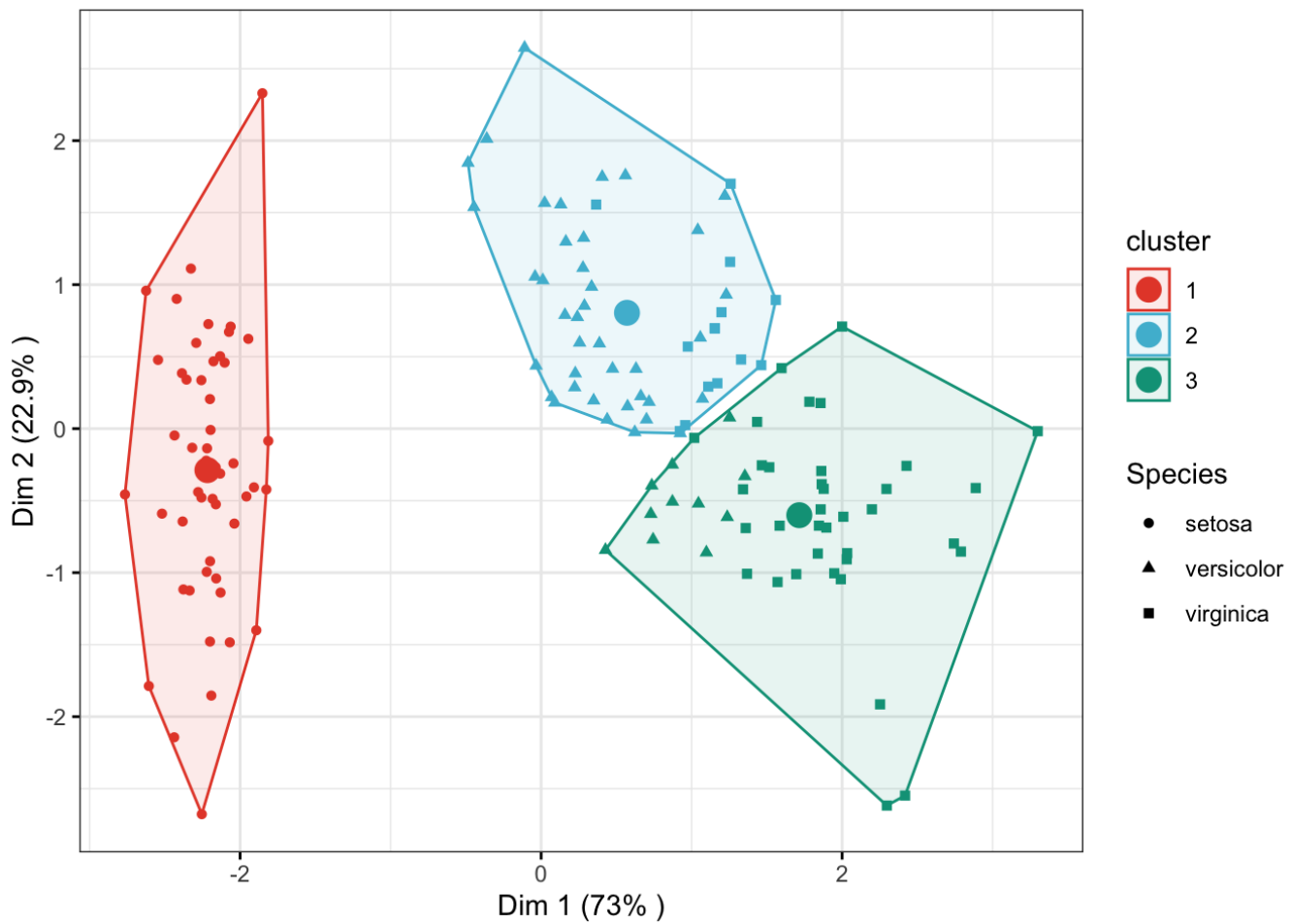


Figure 1: Pflanzencluster

Lesen Sie!

Emmer/Füting/Vowe (2006): “Wer kommuniziert wie über politische Themen?”

Wer kommuniziert wie über politische Themen?

Eine empirisch basierte Typologie individueller politischer Kommunikation

Martin Emmer/Angelika Füting/Gerhard Vowe

Der Aufsatz enthält eine empirisch basierte Typologie der individuellen politischen Kommunikation in Deutschland. Dafür wurden mittels einer Clusteranalyse Personen aufgrund bestimmter Merkmale so gruppiert, dass sich die Angehörigen einer Gruppe untereinander in ihrer politischen Kommunikation möglichst stark ähneln und von den Angehörigen anderer Gruppen möglichst stark unterscheiden. Die Basis bildet ein deutschlandweit repräsentativer Datensatz von 2003, der detailliert die politische Kommunikation der Bevölkerung beschreibt – von der Nutzung politischer Medienangebote (z. B. Schauen von TV-Nachrichten) über die interpersonale Kommunikation zu politischen Themen bis zu partizipativer politischer Kommunikation (z. B. Teilnahme an Unterschriftensammlungen). Es konnten fünf Typen ermittelt werden: der „passive Mainstreamer“ (größte Gruppe mit 43 % der Bevölkerung), der „eigennützige Interessenvertreter“, der „bequeme Moderne“, der „traditionelle Engagierte“ und der „organisierte Extrovertierte“ (kleinste Gruppe mit 9 %). Die Etiketten machen das jeweilige Kommunikationsprofil der Typen deutlich. Diese Typologie kann über die deskriptive Funktion hinaus dazu dienen, kommunikationstheoretisch und -politisch relevante Phänomene wie „Wissenskluft“, „Digital Divide“ oder „Politikverdrossenheit“ weiter aufzuklären.

Schlagwörter: Typologie, Politische Kommunikation, Gesellschaftsstruktur, Sekundäranalyse, Internet, Partizipation, Interpersonale Kommunikation, Mediennutzung, Clusteranalyse

Take Home – Ausblick – Vokabeln

Take Home

Clusteranalysen

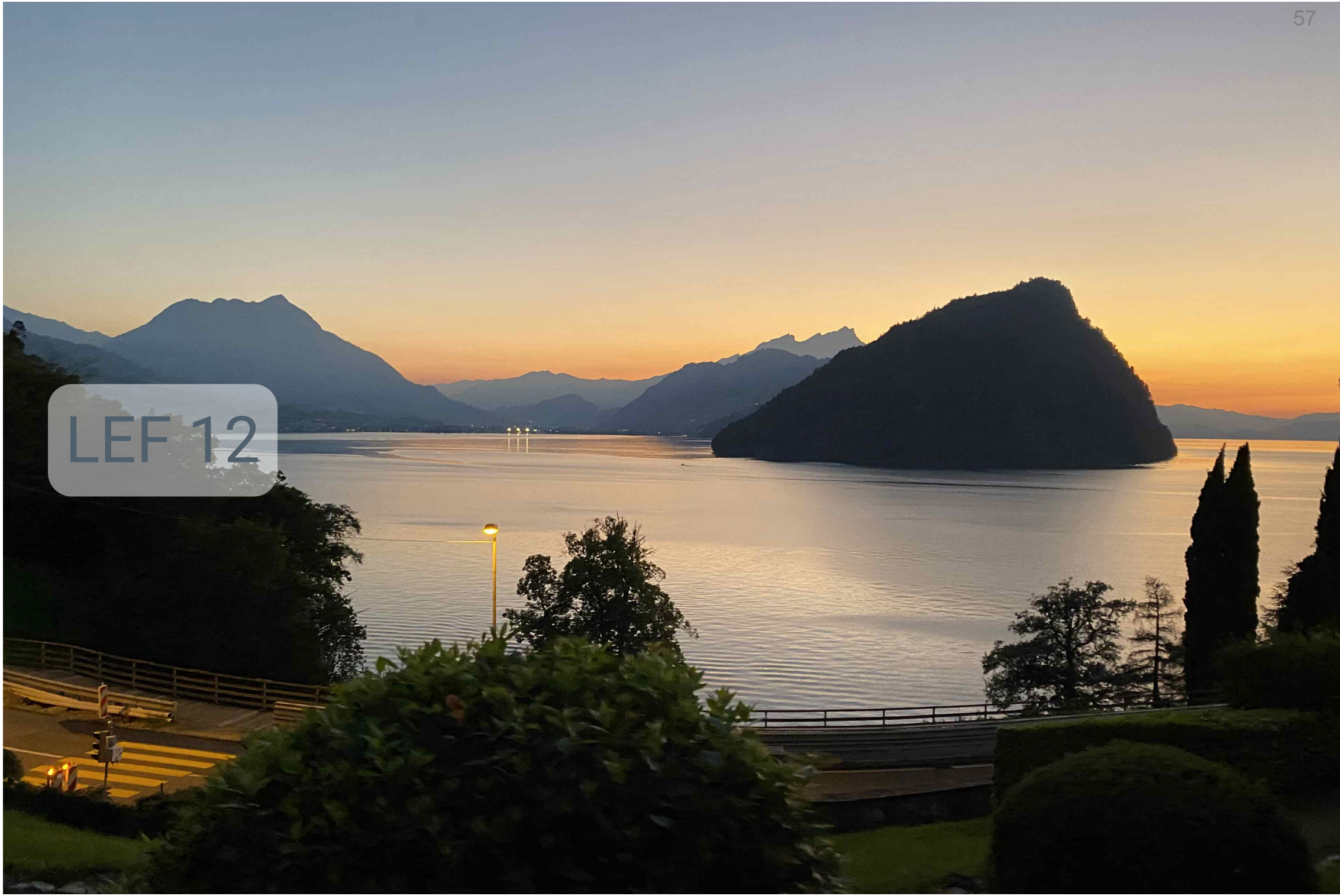
Mit Clusteranalysen versucht man, Fälle anhand von Variablen zusammenzufassen. Es sollen möglichst nicht zu viele Cluster sein und in den Clustern die Fälle ähnlich und die Cluster selbst unähnlich.

k-Mean-Cluster

- Kann man mit metrischen machen (auch Dummys).
- Der Algorithmus findet iterativ die beste (kleinste mittlere Distanzen) Zuordnung aller Fälle zu eine Anzahl vorgegebener Cluster (k).
- Mit dem Ellenbogenkriterium kann man bestimmen, welche Anzahl Cluster k optimal ist.
- Geht nur mit metrischen, nicht mit ordinalen oder kategorialen
- Findet immer eine Lösung, auch bei reiner Zufallsverteilung.

Ausblick

Wir könnten uns mit Clusteranalysen in R beschäftigen (im Sinne einer Übung) oder auf kompliziertere Sachverhalte eingehen. Etwas, was Sie gerne nochmal erklärt bekommen hätten.


LEF 12

Essayfragen 12

E12.1 a) Was ist das Ziel einer Clusteranalyse? b) Was ist der Unterschied zu einer Faktorenanalyse?

E12.2 Welches Skalenniveau wird benötigt, wenn bei einer Clusteranalyse «Clusterzentren» berechnet werdenn sollen?

E12.3 Wie ist der grobe Ablauf einer k-means-Clusteranalyse?

E12.4 Warum werden Clusteranalysen im Kontext von ML als «unsupervised learning» bezeichnet?

E12.5 Warum kann man nicht einfach sagen, dass die Clusterlösung mit der kleinsten Heterogenität (mittlere Distanz zu Clusterzentren) die beste Lösung ist?

E12.6 Warum werden oft Faktorenanalysen vor den Clusteranalysen durchgeführt?

E12.7 Was verbirgt sich hinter der Abkürzung «ML»?

E12.8 Wozu dient bei einer Clusteranalyse das «Ellbogenkriterium»?

E12.9 Warum sind hierarchische Clusteranalysen für Big Data oder zum Beispiel Bildanalysen ungeeignet?

MC-Fragen 12

MC 12.1.

MC 12.1: Sind folgende Aussagen richtig oder falsch?

richtig	falsch	Aussagen
○	○	Clusteranalysen dienen der Zusammenfassung von Fällen in Gruppen.
○	○	Mit Clusteranalysen werden Variablen zusammengefasst.
○	○	Mit der Clusteranalyse werden möglichst unähnliche Fälle in Clustern zusammengefasst.
○	○	Die Clusteranalyse ist ein exploratives Verfahren.

Punkte: 0

MC 12.2.

MC 12.2: Sind folgende Aussagen richtig oder falsch?

richtig	falsch	Aussagen
○	○	Die Clusteranalyse wird im ML-Kontext als «unsupervised learning» behandelt.
○	○	Ziel der Clusteranalyse ist eine Reduktion der Fälle auf wenige untereinander homegene Cluster bei möglichst geringer Homogenität innerhalb der Cluster.
○	○	Häufig werden Clusteranalysen vor Faktorenanalysen durchgeführt, damit letztere besser fitten.
○	○	Damit die Variablen nicht hoch korrelieren, werden vor CAs oft FAs durchgeführt.

Punkte: 0

MC 12.3.

MC 12.3: Sind folgende Aussagen richtig oder falsch?

richtig	falsch	Aussagen
○	○	Clusteranalysen sind vor allem für die Analyse fehlender Werte geeignet.
○	○	Es gibt Clusteranalyseverfahren, die mit allen möglichen Skalenniveaus gut klarkommen.
○	○	Clusteranalysen funktionieren besonders gut mit kleinen Fallzahlen.
○	○	Es kann immer höchstens so viele Cluster geben, wie es Variablen gibt.

Punkte: 0

MC 12.4.

MC 12.4: Sind folgende Aussagen richtig oder falsch?

richtig	falsch	Aussagen
○	○	Proximitätsmasse sind Ähnlichkeits- beziehungsweise Distanzmasse.
○	○	Euklidische Distanzen können nur für metrische Variablen festgestellt werden.
○	○	Für eine Clusteranalyse werden immer mindestens so viele Variablen benötigt, wie man Cluster extrahieren will.
○	○	Manche Clusteralgorithmen nehmen als Ähnlichkeitsmass auch die Korrelation.

Punkte: 0

MC 12.5.

MC 12.5: Sind folgende Aussagen richtig oder falsch?

richtig	falsch	Aussagen
○	○	k-means-Cluster kann nur auf metrische (und Dummys) angewendet werden.
○	○	Bei k-means-Cluster sind ähnliche Standardabweichungen der Variablen erwünscht.
○	○	Bei k-means-cluster müssen hohe Korrelationen zwischen den Variablen vorliegen.
○	○	Mit dem Ellbogenkriterium werden Clusterzentren einander zugeordnet.

Punkte: 0

MC 12.6.

MC 12.6: Mal was zu R?

richtig	falsch	Aussagen
○	○	Clusteranalysen haben eher deskriptiven Analysegehalt.
○	○	Clusteranalysen werden häufig eingesetzt um Typologien zu bilden.
○	○	Hierarchische Clusteranalysen sind so rechenaufwendig, dass selbst moderne Rechner eher verrottet sind als sie mit der Analyse fertig werden.
○	○	Die Clusterzentren der k-means-Cluster korrelieren hoch miteinander.

Punkte: 0

Insgesamt 0 von 12 Punkten, was 0% und etwa einer 1 entspricht.