Advantages and disadvantages to use factor scores as input for cluster analysis (k-means), this is also called factor-cluster segmentation

Can be done if:

* High explained variance by factors (if “only” 60% of the variance if explained by the factors, still 40 % of the variance gets lost for cluster analysis). Wir haben 50%
* Factors explain the latent dimensions on which individuals differ i.e. they characterize the structure of the data and factors account for the between cluster differences. This means that factors need to have robust loadings (i.e. factors are not highly correlated)
* Using factor scores rather than a bigger number of variables can enhance stability of the cluster solution if the ratio of sample size to active variables is small (bei uns 5/1000)
* Distance can be calculated because factor scores are numeric

Problem:

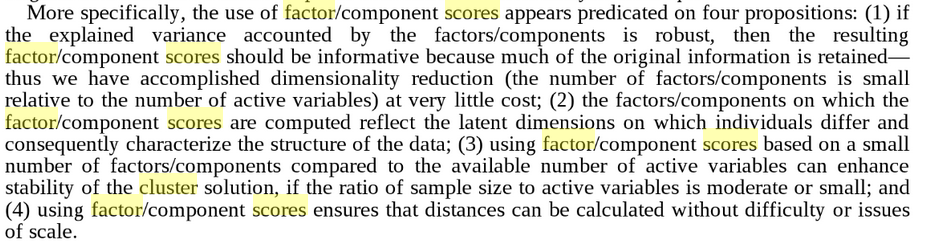
* Factor solutions are not unique. Factor analysis uses rotation and each method of rotation leads to different factor loadings which in turn lead to different clusters

References

https://stats.stackexchange.com/questions/91401/can-component-scores-be-used-for-further-analyses-e-g-cluster-analysis

https://www.researchgate.net/post/Should\_a\_factor\_analysis\_always\_precede\_a\_consensus\_cluster\_analysis\_technique

The Handbook of Marketing Research: Uses, Misuses, and Future Advances

Factor cluster technique

Example

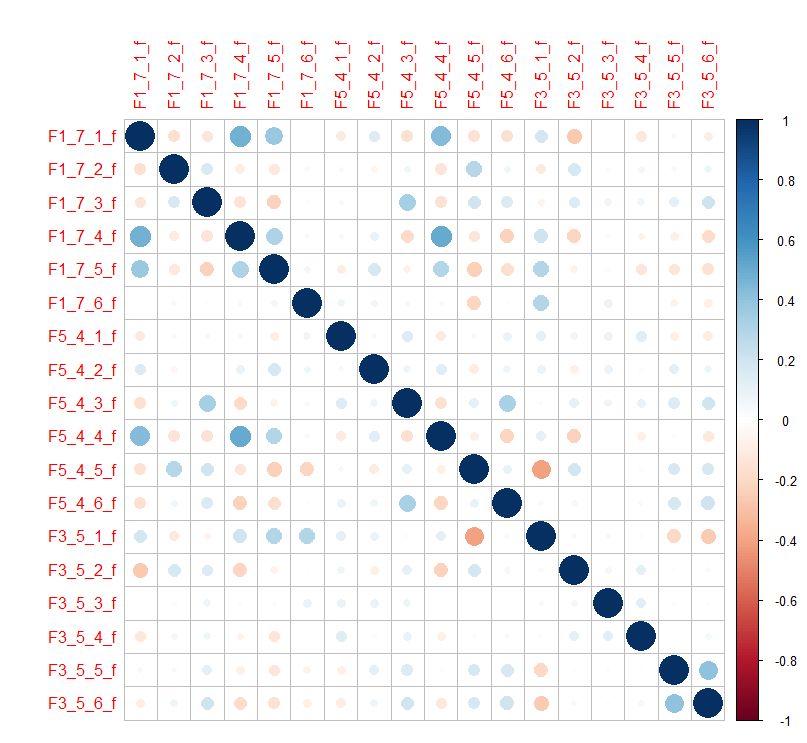
https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&ved=2ahUKEwjAlsjer4jqAhWPlhQKHZ6rBNIQFjAGegQIBhAB&url=https%3A%2F%2Fhrcak.srce.hr%2Ffile%2F17906&usg=AOvVaw3mcsEA7iy608PJOgjGvDjm

https://www.datanovia.com/en/lessons/assessing-clustering-tendency/

**does it make sense to use clustering**

https://towardsdatascience.com/when-clustering-doesnt-make-sense-c6ed9a89e9e6

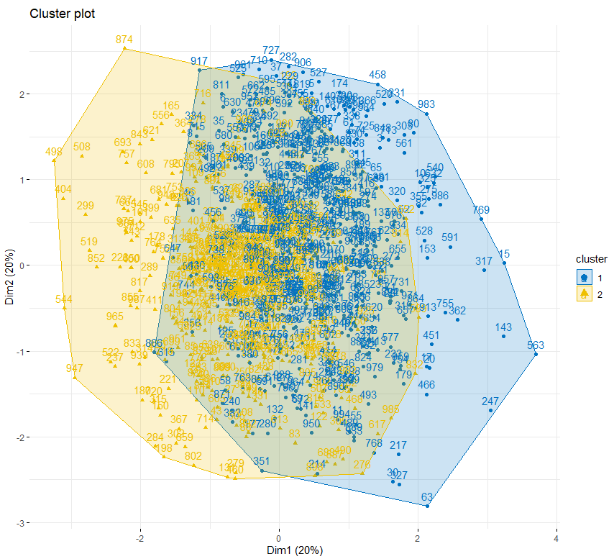
https://www.datanovia.com/en/lessons/assessing-clustering-tendency/

Summary bisherige Ergebnisse

**Faktoranalyse mit allen 16 Variablen**

* In R gleiches Ergebnis wie in SPSS (Achtung wir sollten auch stark negative Ladungen zur Beschreibung der Faktoren berücksichtigen: Bei FAC3 läd beispielsweise F5\_4\_5 stark negativ, d.h. diese Gruppe hat starke Abneigung bei dieser Frage
* Achtung: Es werden auch Variablen verwendet, die mit keiner anderen der übrigen 17 Variablen stark korreliert (< 0.3)

**Scores der Faktorenanalyse als input für Clusteranalyse mittels kmeans und Eucledian distance metric :**

* Problem: Faktoren erklären „nur“ 50% der Variation in den 18 Variablen. Das ist nicht besonders hoch um Faktor scores for Clusteranalyse zu verwenden. Das liegt daran, dass in Faktoranalyse auch Variablen verwendet wurden, die Korrelation von < 0.3 haben.
* Habe eine Funktion verwendet, die die optimale Anzahl an Cluster berechnet. Die zur Auswahl stehenden Methoden sind "silhouette" (for average silhouette width), "wss" (for total within sum of square) and "gap\_stat" (for gap statistics). Ergebnis bei gap\_stat ist 1 und bei silhouette 7 (noch genauer ansehen)
* Wählt man 2 Cluster sieht das Ergebnis folgendermaßen aus (PCA wurde für Dimensionsreduktion der Abbildung verwendet):
* Man kann daraus schon vermuten, dass die Qualität der Cluster Analyse nicht gut ist. 3 ausgewählte Validation measures ergeben:

Score Method Clusters

Connectivity 282.9861 kmeans 2 (connectivity **∈** [0, infinity)) 🡪 minimize

Dunn 0.0540 kmeans 2 (Dunn **∈** [0,infinity]) 🡪 maximize

Silhouette 0.1433 kmeans 2 (Silhouette **∈** [-1,1]) 🡪 maximize

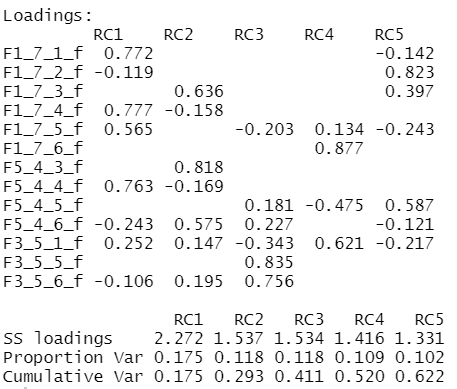
Connectivity nicht so gut

Dunn sehr schlecht

Silhouette mittel

**Faktoranalyse mit Variablen, die in Korrelationsmatrix >=0.29 haben**

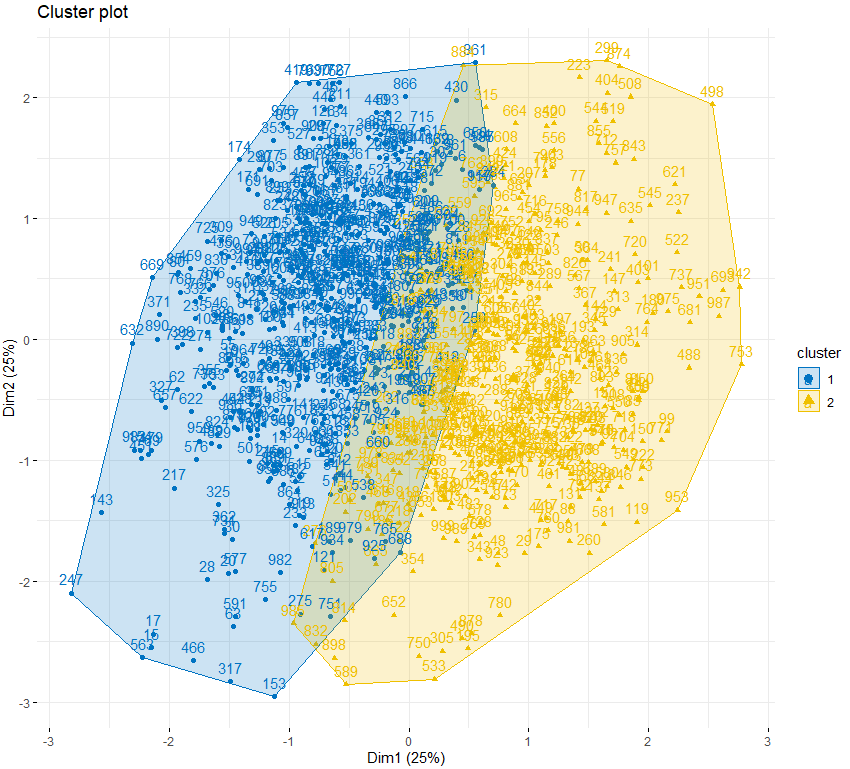
* Damit fallen weg: F5\_4\_1\_f, F5\_4\_2\_f, F3\_5\_2\_f, F3\_5\_3\_f, F3\_5\_4\_f
* Durch die Faktoren kann nun 62.2% der Variation erklärt werden



* Achtung: nur 1-4 haben eigenvalue >1
* Nun stellen sich auch die Faktoren etwas anders zusammen:
  + FAC1 (Traditionelle): Keine externe Betreuung, Betreuung in Familie, Mutter sollte nicht erwerbstätig sein, Vater soll sich um Geld kümmern
  + FAC2 (Unabhängige): Frau sollte eigenes Geld verdienen, Mutter sollte ebenfalls finanziell Verantwortung übernehmen, Mütter, die sich nur um Kind kümmern werden unglücklich
  + FAC3 (Fokus Paar): Eltern müssen sich auch Zeit für ihre Beziehung nehmen, für Kind sind auch Erfahrungen mit anderen Bezugspersonen wichtig
  + FAC4 (Übermutter): Mutter traut Vater Betreuung oft nicht zu, Vater ist nicht kompetent, Vater kann sich nicht genauso gut kümmern
  + FAC5 (partnerschaftlich): Für Entwicklung Kind egal welcher Elternteil sich kümmert, Vater kann sich genauso gut kümmern, Mutter sollte finanzielle ebenfalls Verantwortung übernehmen

**Clusteranalyse für FAC1-4**

* Wählt man 2 Cluster sieht das Ergebnis folgendermaßen aus (PCA wurde für Dimensionsreduktion der Abbildung verwendet)



* 3 ausgewählte Validation measures ergeben:

Score Method Clusters

Connectivity 232.2155 kmeans 2

Dunn 0.0276 kmeans 2

Silhouette 0.1626 kmeans 2

**Faktoranalyse mit Variablen, die in Korrelationsmatrix >=0.3 haben**

**Ändert sich nicht ausschlaggebend, wenn nur die verwendet werden (außer nur noch 3 Faktoren mit eigenvalue >1)**

* Damit fallen weg: F5\_4\_1\_f, F5\_4\_2\_f, F3\_5\_2\_f, F3\_5\_3\_f, F3\_5\_4\_f
* Und zusätzlich: F1\_7\_6\_f, F1\_7\_2\_f,F5\_4\_5\_f, F\_3\_5\_5\_f
* Durch die Faktoren kann nun 62.2% der Variation erklärt werden (wenn man 5 nehmen würde)

**Wähle je Faktor jene Variablen für die Clusteranalyse, die am höchsten laden**

* Ausgehend von Faktoranalyse mit Variablen cor > 0.29 und eigenvalue >1. Ergibt 4 Faktoren mit einer erklärten Variation von 0.55
* Variablen mit höchsten Ladungen sind:

# RC 1: F1\_7\_4\_f , F5\_4\_4\_f

# RC 2: F3\_5\_5\_f, F3\_5\_6\_f

# RC 3: F3\_5\_1\_f, F1\_7\_6\_f

# RC 4: F1\_7\_2\_f, F1\_7\_3\_f

# Loadings:

# RC1 RC2 RC3 RC4

# F1\_7\_1\_f 0.770 -0.112

# F1\_7\_2\_f -0.109 -0.190 -0.220 0.683

# F1\_7\_3\_f 0.280 0.675

# F1\_7\_4\_f 0.784 -0.153

# F1\_7\_5\_f 0.551 -0.119 0.276 -0.149

# F1\_7\_6\_f 0.623

# F5\_4\_3\_f -0.154 0.484 0.297 0.459

# F5\_4\_4\_f 0.776

# F5\_4\_5\_f -0.606 0.487

# F5\_4\_6\_f -0.280 0.510 0.208 0.190

# F3\_5\_1\_f 0.223 -0.182 0.741

# F3\_5\_5\_f 0.728 -0.253

# F3\_5\_6\_f 0.719 -0.235

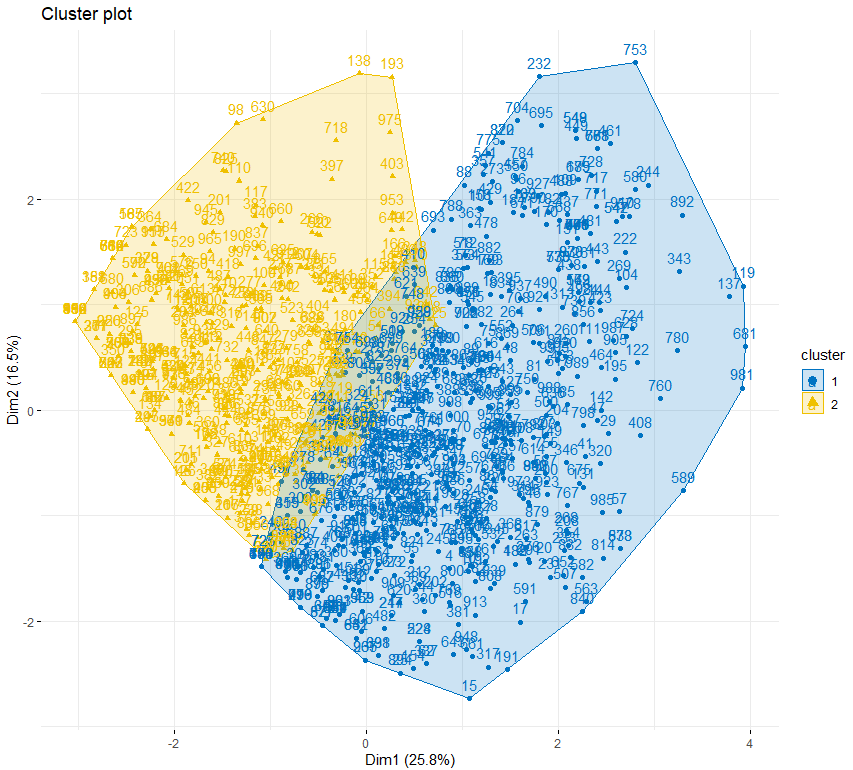
#

# RC1 RC2 RC3 RC4

# SS loadings 2.303 1.735 1.692 1.451

# Proportion Var 0.177 0.133 0.130 0.112

# Cumulative Var 0.177 0.311 0.441 0.552



F1\_7\_4\_f F5\_4\_4\_f F3\_5\_5\_f F3\_5\_6\_f F3\_5\_1\_f F1\_7\_6\_f F1\_7\_2\_f F1\_7\_3\_f

3.018519 3.530303 3.505051 3.361953 2.127946 2.392256 2.328283 2.850168

1.421182 2.231527 3.798030 3.866995 1.371921 2.172414 2.783251 3.317734

F1\_7\_4\_f F5\_4\_4\_f F3\_5\_5\_f F3\_5\_6\_f F3\_5\_1\_f F1\_7\_6\_f F1\_7\_2\_f F1\_7\_3\_f

**High high high high mid mid mid high**

**Low mid high high low mid high high**

* Ausgehend von Faktoranalyse mit allen 18 Einstellungsvariablen
* Variablen mit höchsten Ladungen sind:

# RC 1: F1\_7\_4\_f , F5\_4\_4\_f

# RC 2: F3\_5\_6\_f, F3\_5\_5\_f

# RC 3: F3\_5\_1\_f, F5\_4\_5\_f

# RC 4: F1\_7\_2\_f, F3\_5\_2\_f

# RC 5: F3\_5\_4\_f, F3\_5\_3\_f

**Clustermittelwerte für Cluster 1 und zwei**

F1\_7\_4\_f F5\_4\_4\_f F3\_5\_6\_f F3\_5\_5\_f F3\_5\_1\_f F5\_4\_5\_f F1\_7\_2\_f F3\_5\_2\_f

1.557778 2.413333 3.828889 3.802222 1.333333 3.457778 2.853333 2.582222

3.034545 3.485455 3.352727 3.478182 2.220000 2.778182 2.234545 1.814545

F3\_5\_4\_f F3\_5\_3\_f

2.702222 2.822222

2.456364 2.727273

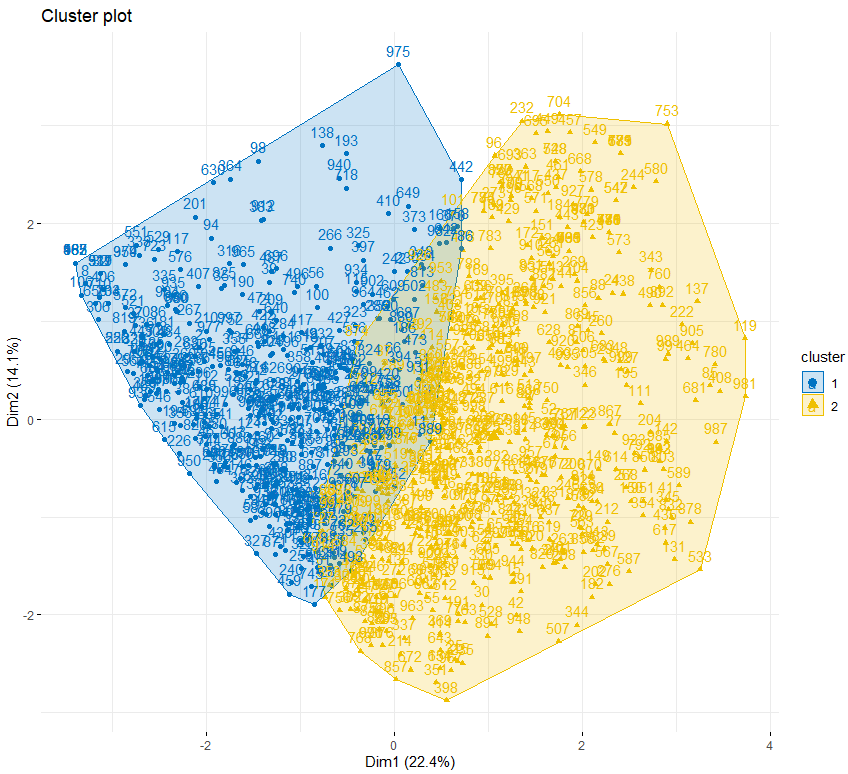
Validity measures

Score Method Clusters

Connectivity 56.7516 kmeans 2

Dunn 0.0119 kmeans 2

Silhouette 0.3150 kmeans 2



**Kreuztabelle zwischen indicator und Clusterzugehörigkeit**

1 2 3 4 5

1 22 136 40 126 126

2 179 69 189 35 78

**Clusteranalyse für weitere Variablen**

* Konto vs. Einkommensabhängig
* Einstellung zu Familie
* Partnerschaftsbonus

**Hierachische Clusteranalyse mit Grundmodell factor scores als Ausgangsmodell**

Gleiches Ergebnis wie in SPSS:

Aufteilung der Beobachtungen, wenn man 4 Cluster wählt:

R

cluster n

1 1 215

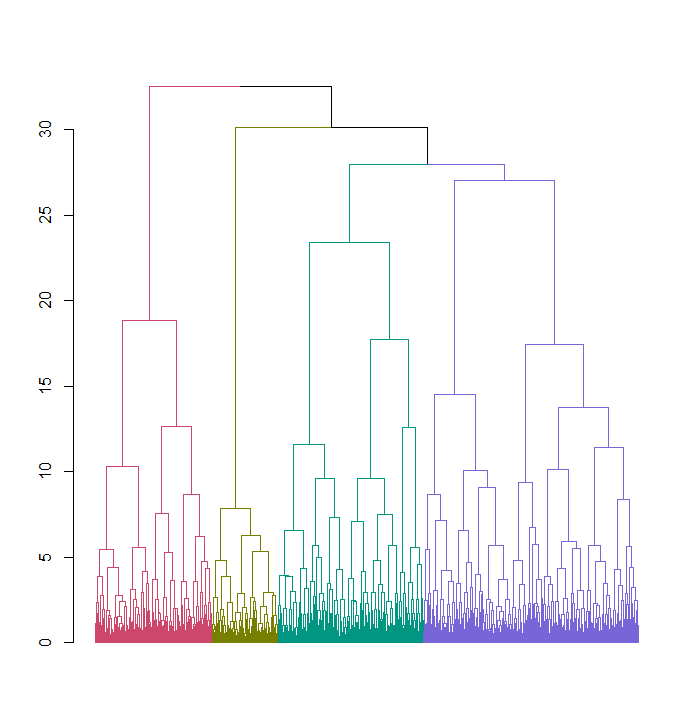
2 2 396

3 3 268

4 4 121

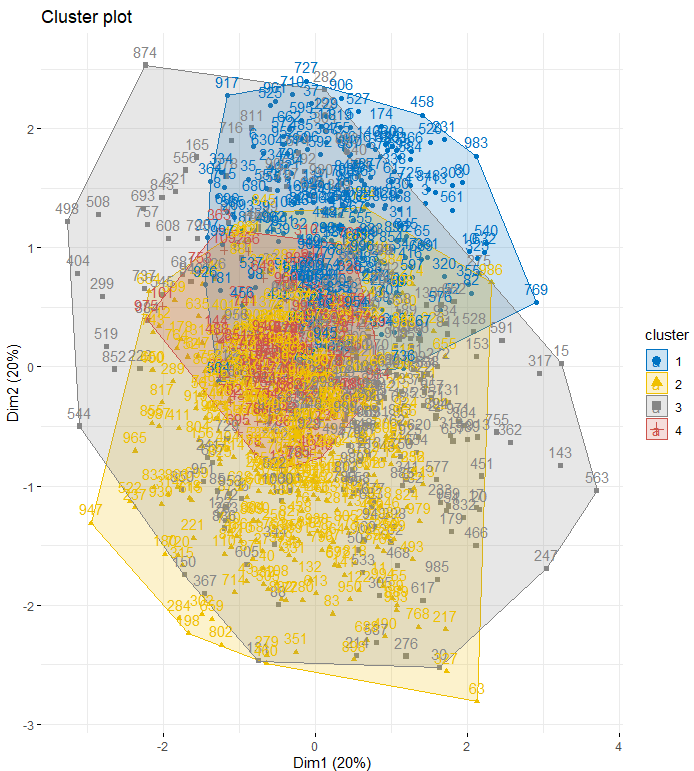
SPSS

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Ward Method** | | | | | |
|  | | Häufigkeit | Prozent | Gültige Prozente | Kumulierte Prozente |
| Gültig | 1 | 215 | 21,5 | 21,5 | 21,5 |
| 2 | 396 | 39,6 | 39,6 | 61,1 |
| 3 | 268 | 26,8 | 26,8 | 87,9 |
| 4 | 121 | 12,1 | 12,1 | 100,0 |
| Gesamt | 1000 | 100,0 | 100,0 |  |

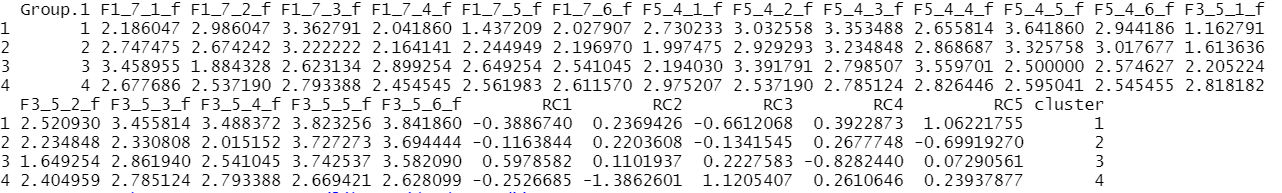


Erklärung zur Darstellung: This provides a nice illustration of the clusters. If there are more than two dimensions (variables) fviz\_cluster will perform principal component analysis (PCA) and plot the data points according to the first two principal components that explain the majority of the variance.

https://stats.stackexchange.com/questions/141280/understanding-cluster-plot-and-component-variability



Group means:



Hierachische Clusteranalyse ohne Faktoranalyse und Verwendung der Variablen:

# RC 1: F1\_7\_4\_f , F5\_4\_4\_f

# RC 2: F3\_5\_5\_f, F3\_5\_6\_f

# RC 3: F3\_5\_1\_f, F1\_7\_6\_f

# RC 4: F1\_7\_2\_f, F1\_7\_3\_f

