# Mixed Type Data Clustering in R

A Case Study with Customer Data from an Online Fashion Retailer

Ladislaus von Bortkiewicz Chair of Statistics
Humboldt–Universität zu Berlin
SPL - Statistical Programming Languages
http://lvb.wiwi.hu-berlin.de

# Outline

- Introduction
- Conducting Cluster Analysis
- Data (and Application in R)
- Evidence
- Discussion
- Conclusion

# Introduction

- Cluster analysis
  - exploration of similarity in data
  - usually of data on numerical scale
- Application in marketing
  - e.g. Detecting customer segments
- Data
- mixed type data from a fashion retailer
- Results
- Intepretation and evaluation
  - data visualualization

# Conducting Cluster Analysis

- I. Selecting Distance/Similarity Matrix
  - (non-)euclidean, manhattan, ...
  - composite
- II. Choice of Clustering Techniques
  - Partitioning (K-Means, PAM)
  - Hierarchical (agglomerative/divisive)
  - Density
- III. Determining the Number of Clusters
- IV. Cluster Interpretation
- V. Cluster Visualization
  - scatterplots
  - dendrograms
  - heatmaps

# Raw data description

Orders from an online fashion retailer

```
'data.frame':   100000 obs. of  14 variables:
 $ order_item_id: int  1 2 3 4 5 6 7 8 9 10 ...
 $ order_date   : Factor w/ 365 levels "2012-04-01","2012-04-02",..: 157 217 304 130 169 348 278 90 162 17 ...
 $ delivery_date: Factor w/ 320 levels "?","1990-12-31",..: 107 153 211 89 134 248 193 64 1 1 ...
 $ item_id      : int  1507 1745 2588 164 1640 2378 1506 224 1970 485 ...
 $ item_size    : Factor w/ 114 levels "1","10","10+",..: 106 2 112 60 100 52 106 60 107 50 ...
 $ item_color   : Factor w/ 85 levels "?","almond","amethyst",..: 49 21 77 19 5 50 49 22 21 19 ...
 $ brand_id     : int  102 64 42 47 97 72 102 58 66 70 ...
 $ item_price   : num  24.9 75 79.9 79.9 69.9 ...
 $ user_id      : int  46943 60979 72232 41242 8810 15761 64795 23489 47837 6380 ...
 $ user_title   : Factor w/ 5 levels "Company","Family",..: 4 4 4 4 4 4 4 4 3 4 ...
 $ user_dob     : Factor w/ 12122 levels "?","1900-11-19",..: 5964 9039 1133 4571 8304 6277 1 5981 7312 3596 ...
 $ user_state   : Factor w/ 16 levels "Baden-Wuerttemberg",..: 11 4 12 16 1 10 13 10 10 10 ...
 $ user_reg_date: Factor w/ 775 levels "2011-02-16","2011-02-17",..: 1 95 713 540 336 1 663 1 572 361 ...
 $ return       : int  1 0 1 1 1 1 0 1 0 0 ...
```

Goal: Identify customer segments by age, gender, state and loyalty

# Data after manipulation

```
> str(data)
'data.frame':   28178 obs. of  6 variables:
 $ ID                 : int  6 9 13 15 23 26 27 28 30 31 ...
 $ gender             : Factor w/ 2 levels "Mr","Mrs": 2 2 2 2 2 2 2 2 2 2 ...
 $ state              : Factor w/ 16 levels "Baden-Wuerttemberg",..: 8 11 10 11 1 7
.
 $ age                : num  43 40 50 43 36 58 55 50 63 44 ...
 $ frequency.per.month: num  0.5 0.1667 0.1667 0.0833 0.0833 ...
 $ rfm                : num  7 4.67 6.67 4.67 3 ...
```

- Change identification from order_item_id to user_id
- Model based imputation of missings and anormalous values
- Drop irrelevant variables
- Create variable of interest: rfm (monthly recency, frequency and monetary value)
- Observation: also nominal scaled variables included

# Approach for PAM

- standard techniques based on euclidean metrics fail
- standard visualization fails

- Different procedure, e.g. for partitioning
  - use similarity matrix of composite metrics: **gower matrix**
  - apply partitioning around medoids algorithm **(PAM)**
  - determine number of clusters by **silhouette width**

# Results - PAM

- Draw subsample to reduce duration time of algorithm
- Calculate Gower distance matrix and summary statistics

```
> gower_dist <- daisy(sample[, -1],
+                     metric = "gower")
> summary(gower_dist) # summary statistics
499500 dissimilarities, summarized :
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  0.2790  0.3348  0.3386  0.3995  0.8411
Metric :  mixed ;  Types = N, N, I, I, I
Number of objects : 1000
```

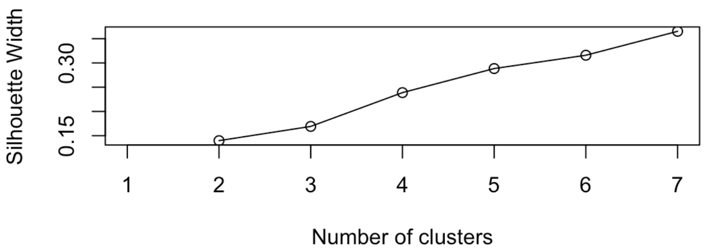
# Results - PAM

- Examples of most (dis-)similar pairs

```
> # Most similar pair of data
> sample[
+   which(gower_mat == min(gower_mat[gower_mat != min(gower_mat)]),
+         arr.ind = TRUE)[1, ], -1]
       state gender age loyalty.in.months      rfm
3316 Bavaria    Mrs  39                12 4.333333
7482 Bavaria    Mrs  40                12 4.333333
> # Most dissimilar pair of data
> sample[
+   which(gower_mat == max(gower_mat[gower_mat != max(gower_mat)]),
+         arr.ind = TRUE)[1, ], -1]
                           state gender age loyalty.in.months rfm
12296        North Rhine-Westphalia    Mrs  59                25   8
21176 Mecklenburg-Western Pomerania     Mr  24                 8   2
```

# Results - PAM

- Use silhuette width to determine number of clusters




- The higher the value the more appropriate the respective number of clusters

# Cluster Interpretation

- Check summary statistics of respective clusters

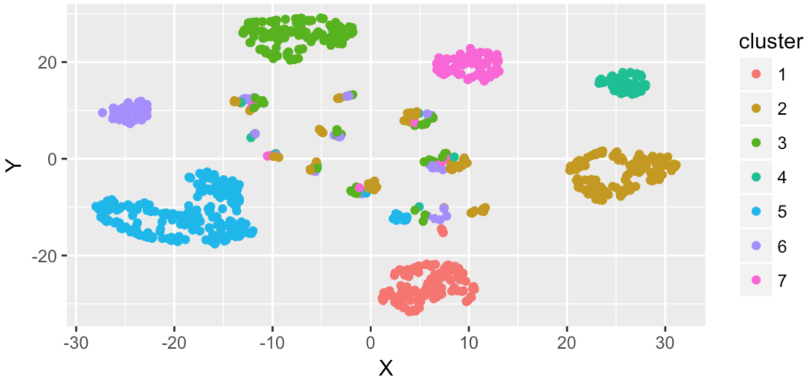
```
[[1]]
                             state       gender          age          loyalty.in.months
 Bavaria                        :133   Mr :  9    Min.   :13.00   Min.   : 6.00  
 Schleswig-Holstein             :  7   Mrs:146    1st Qu.:43.00   1st Qu.: 7.00  
 Berlin                         :  2              Median :49.00   Median :11.00  
 Brandenburg                    :  2              Mean   :47.96   Mean   :11.22  
 Hamburg                        :  2              3rd Qu.:53.50   3rd Qu.:13.00  
 Mecklenburg-Western Pomerania:  2              Max.   :76.00   Max.   :26.00  
 (Other)                        :  7                                              
      rfm           cluster 
 Min.   :1.333   Min.   :1  
 1st Qu.:3.333   1st Qu.:1  
 Median :4.333   Median :1  
 Mean   :4.594   Mean   :1  
 3rd Qu.:5.667   3rd Qu.:1  
 Max.   :8.667   Max.   :1  
```

# Cluster Interpretation

- Check medoids, i.e. representative objects of each cluster

```
> sample[pam_fit$medoids, ]
         ID                 state gender age loyalty.in.months      rfm
14992 30497               Bavaria    Mrs  49                11 4.333333
3669   7102          Lower Saxony    Mrs  46                11 6.000000
12236 24643     Baden-Wuerttemberg   Mrs  45                11 3.333333
9925  19992   Rhineland-Palatinate   Mrs  52                11 4.666667
14695 29889 North Rhine-Westphalia   Mrs  49                11 4.666667
12944 26132     Schleswig-Holstein   Mrs  46                 8 4.000000
5974  11869                 Hesse    Mrs  49                12 4.333333
> 
```

# t-SNE Visualization




- easy to see separation of clusters

# Whats else?

- presentation of hierarchical and density cluster analysis for composite data
- Comparison of the different approaches