Song Recommender

MathStat Application

MathStat

Professor Tillmann Schwörer

Summer Semester 2020

Alexander P. Kleine::	MatNr. 934944
Robert Weber::	MatNr. 936729

Overview

- The Task
- Obtaining Data from a Music Test Survey
- Project Vision
- Wrangling… Wrangling… Wrangling
- EDA – Explorative Data Analysis (k-Means Clustering)
- Recommenderlab Package (Recommender System)
- RStudio – Guideline Markdown

The Chosen Task

Setup:

Your team collects a selection of jokes (~ 25), ideally jokes that appeal to different humour preferences. Each course participant must rate a mandatory subset of jokes (~15). Additionally each participant rates a subset of the non-mandatory jokes.

Possible analytic questions:

- **Distances**: Measure distances of jokes and/or persons in terms of joke ratings, using alternative distance measures
- **Recommendations**: Make an individualised joke recommendation.
  Recommend one of the bonus jokes that the closest neighbour liked
- **Clustering**: Use k-means or hierarchical clustering to identify homogenous clusters of jokes or persons **Principle component analysis/Singular value decomposition**: Find a lower dimensional repre- sentation (latent factors) of persons and/or jokes.

Songs, movies, or books analysis The same setup and the analytic questions as above can be addressed using an alternative domain, e.g.: Songs, Movies, Books

Note: It won’t be feasible to get a full set of ratings for 15 movies or books from each participant. Hence, you will have a *sparse* rating matrix. This makes some of the analytic questions more difficult or unfeasable (notably clustering and PCA/SVD). However, it is possible to measure distances for each pair of movie or book raters, based on the subset of movies/books that both have rated. And thus, to make personalized recommendations.

Obtaining Data from a CATI/CAWI Music Test

Choosing the Domain: **Songs**

Using **current Music Test Data** from a music research unit

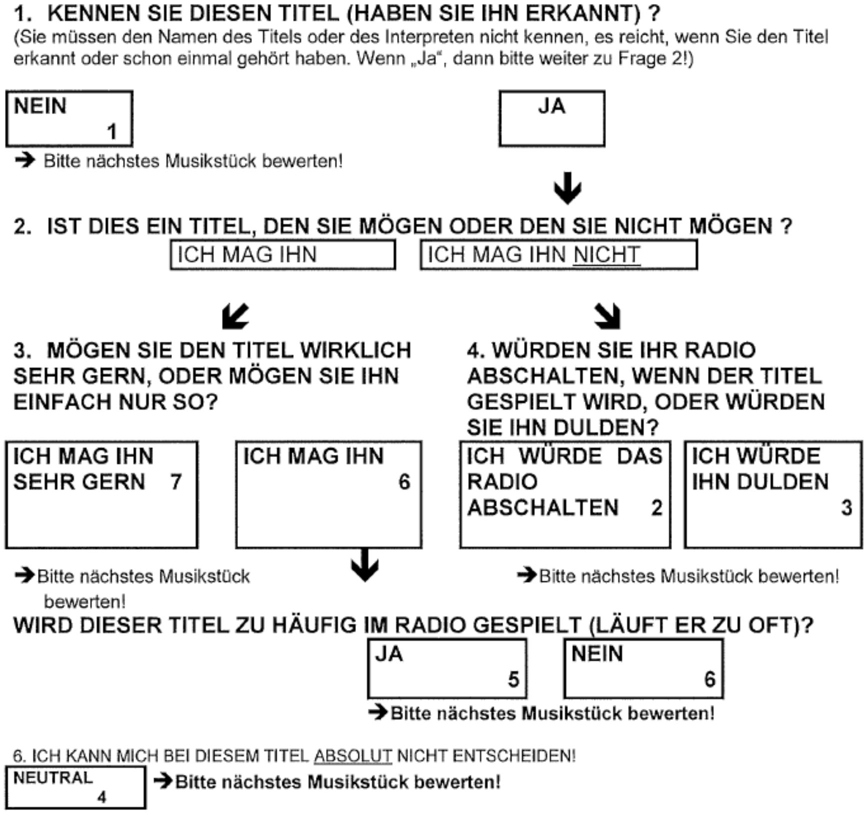
As this data is used to make playlist recommendations or rather **to curate playlists** that are supposed to generate a mixture of maximum number of positive responses while trying to keep negs to a minimum…

We wanted to use this dataset to create a recommender system

DATENWERFT
Markt- und
Meinungsforschung
Rostock

Obtaining Data from a CATI/CAWI Music Test

- Data generated from CATI/CAWI Music Tests
- 120 test persons / 50 song ratings each (resulting in 6,000 entries)
- We do have further data on the tested person (ZAG, listening type, main station)
- The response: „I do not know this song“ is present 495 times (8.25 %)



1. KENNEN SIE DIESEN TITEL (HABEN SIE IHN ERKANNT) ?
(Sie müssen den Namen des Titels oder des Interpreten nicht kennen, es reicht, wenn Sie den Titel erkannt oder schon einmal gehört haben. Wenn „Ja“, dann bitte weiter zu Frage 2!)

NEIN 1 | JA

➔ Bitte nächstes Musikstück bewerten!

2. IST DIES EIN TITEL, DEN SIE MÖGEN ODER DEN SIE NICHT MÖGEN ?
ICH MAG IHN | ICH MAG IHN NICHT

3. MÖGEN SIE DEN TITEL WIRKLICH SEHR GERN, ODER MÖGEN SIE IHN EINFACH NUR SO?

ICH MAG IHN SEHR GERN 7 | ICH MAG IHN 6

➔Bitte nächstes Musikstück bewerten!

4. WÜRDEN SIE IHR RADIO ABSCHALTEN, WENN DER TITEL GESPIELT WIRD, ODER WÜRDEN SIE IHN DULDEN?

ICH WÜRDE DAS RADIO ABSCHALTEN 2 | ICH WÜRDE IHN DULDEN 3

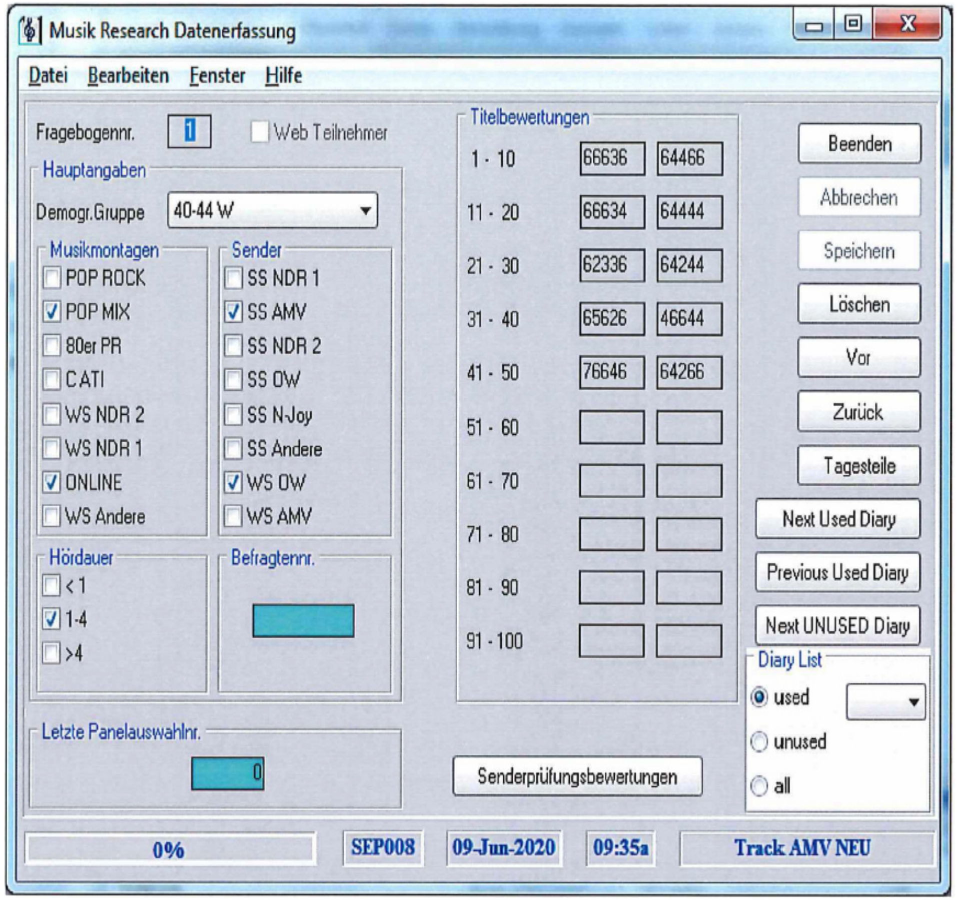
➔Bitte nächstes Musikstück bewerten!

WIRD DIESER TITEL ZU HÄUFIG IM RADIO GESPIELT (LÄUFT ER ZU OFT)?
JA 5 | NEIN 6

➔Bitte nächstes Musikstück bewerten!

6. ICH KANN MICH BEI DIESEM TITEL ABSOLUT NICHT ENTSCHEIDEN!
NEUTRAL 4 ➔Bitte nächstes Musikstück bewerten!

Obtaining Data from a CATI/CAWI Music Test



Fragebogennr. 1 □ Web Teilnehmer

Hauptangaben

Demogr.Gruppe 40-44 W

Musikmontagen
□ POP ROCK
☑ POP MIX
□ 80er PR
□ CATI
□ WS NDR 2
□ WS NDR 1
☑ ONLINE
□ WS Andere

Sender
□ SS NDR 1
☑ SS AMV
□ SS NDR 2
□ SS OW
□ SS N-Joy
□ SS Andere
☑ WS OW
□ WS AMV

Hördauer
□ <1
☑ 1-4
□ >4

Befragtennr.

Letzte Panelauswahlnr. 0

Titelbewertungen

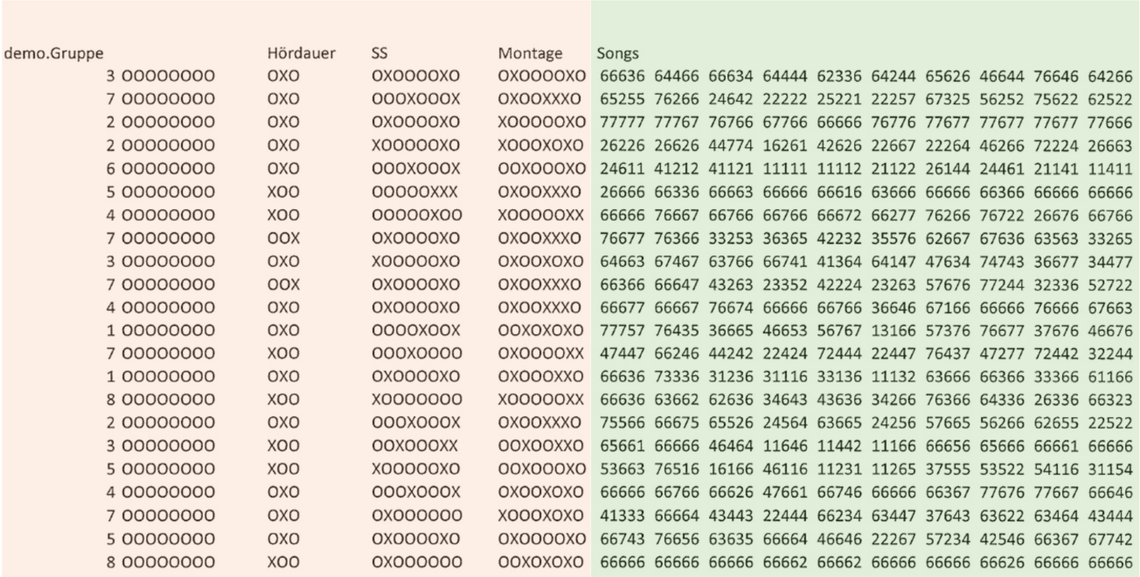
1 - 10	66636	64466
11 - 20	66634	64444
21 - 30	62336	64244
31 - 40	65626	46644
41 - 50	76646	64266
51 - 60		
61 - 70		
71 - 80		
81 - 90		
91 - 100		

Diary List
● used
○ unused
○ all

0% | SEP008 | 09-Jun-2020 | 09:35a | Track AMV NEU

Data about the tested person

Data about the tested songs



demo.Gruppe		Hördauer	SS	Montage	Songs									
3	OOOOOOOO	OXO	OXOOOOXO	OXOOOOXO	66636	64466	66634	64444	62336	64244	65626	46644	76646	64266
7	OOOOOOOO	OXO	OOOXOOOX	OXOOXXXO	65255	76266	24642	22222	25221	22257	67325	56252	75622	62522
2	OOOOOOOO	OXO	OXOOOOXO	XOOOOOXO	77777	77767	76766	67766	66666	76776	77677	77677	77677	77666
2	OOOOOOOO	OXO	XOOOOOXO	XOOOXOXO	26226	26626	44774	16261	42626	22667	22264	46266	72224	26663
6	OOOOOOOO	OXO	OOOXOOOX	OOXOOOXO	24611	41212	41121	11111	11112	21122	26144	24461	21141	11411
5	OOOOOOOO	XOO	OOOOOXXX	OXOOXXXO	26666	66336	66663	66666	66616	63666	66666	66366	66666	66666
4	OOOOOOOO	XOO	OOOOOXOO	XOOOOOXX	66666	76667	66766	66766	66672	66277	76266	76722	26676	66766
7	OOOOOOOO	OOX	OXOOOOXO	OXOOXXXO	76677	76366	33253	36365	42232	35576	62667	67636	63563	33265
3	OOOOOOOO	OXO	XOOOOOXO	OXOOXOXO	64663	67467	63766	66741	41364	64147	47634	74743	36677	34477
7	OOOOOOOO	OOX	OXOOOOXO	OXOOXXXO	66366	66647	43263	23352	42224	23263	57676	77244	32336	52722
4	OOOOOOOO	OXO	OXOOOOXO	OXOOXXXO	66677	66667	76674	66666	66766	36646	67166	66666	76666	67663
1	OOOOOOOO	OXO	OOOOXOOX	OOXOXOXO	77757	76435	36665	46653	56767	13166	57376	76677	37676	46676
7	OOOOOOOO	XOO	OOOXOOOO	OXOOOOXX	47447	66246	44242	22424	72444	22447	76437	47277	72442	32244
1	OOOOOOOO	OXO	OXOOOOXO	OXOOOXXO	66636	73336	31236	31116	33136	11132	63666	66366	33366	61166
8	OOOOOOOO	XOO	XOOOOOOO	XOOOOOXX	66636	63662	62636	34643	43636	34266	76366	64336	26336	66323
2	OOOOOOOO	OXO	OOOXOOOX	OXOOXXXO	75566	66675	65526	24564	63665	24256	57665	56266	62655	22522
3	OOOOOOOO	XOO	OOXOOOXX	OOXOOXXO	65661	66666	46464	11646	11442	11166	66656	65666	66661	66666
5	OOOOOOOO	XOO	XOOOOOXO	OOXOOOXO	53663	76516	16166	46116	11231	11265	37555	53522	54116	31154
4	OOOOOOOO	OXO	OOOXOOOX	OXOOXOXO	66666	66766	66626	47661	66746	66666	66367	77676	77667	66646
7	OOOOOOOO	OXO	OXOOOOOO	XOOOXOXO	41333	66664	43443	22444	66234	63447	37643	63622	63464	43444
5	OOOOOOOO	OXO	OXOOOOXO	OXOOOOXO	66743	76656	63635	66664	46646	22267	57234	42546	66367	67742
8	OOOOOOOO	XOO	OXOOOOOO	OOXOXOXO	66666	66666	66666	66662	66662	66666	66666	66626	66666	66666

A subset of raw data before wrangling

Decision concerning the Scope of the Project

Most Common Types of Recommender Systems:

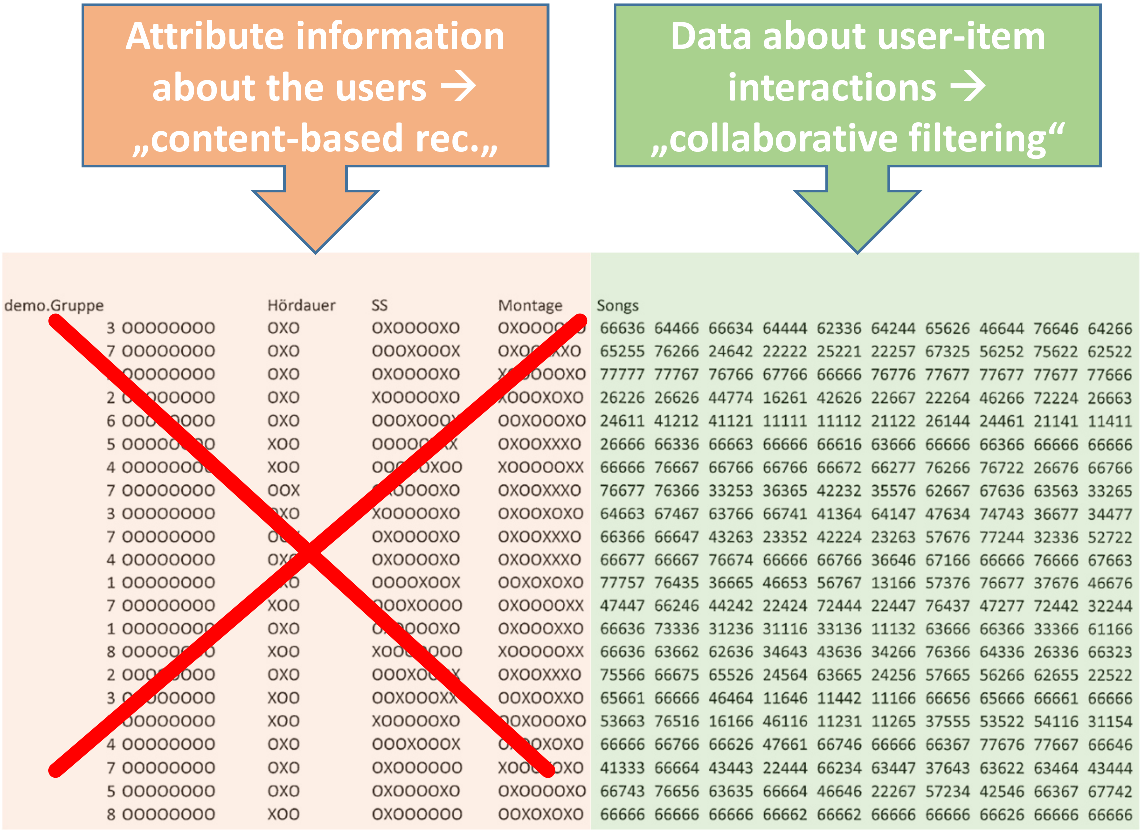
- **Collaborative Filtering Methods**: as in neighbourhood models → a ratings matrix includes dependencies between individual items
- **Content-Based Methods:** User‘s interest can be modeled on the basis of properties (or attributes) of the items they have rated in the past.
- **Knowledge-Based Methods**: users interactively specify their interests after which this is combined with domain knowledge to provide recommendations

While our data is used in a hybridized manner in practise,

we are going to focus on a „Collaborative Filtering Method“.

Decision concerning the Scope of the Project

- Even though we do have (social) data about the tested person, we are going to assume, that we don‘t.
- It means that we intentionally take out information and context, which will in turn leave us with a more straightforward approach.
- This takes scope out of this project, while also having practical relevance, since we cannot neccessarily assume to have this information about the audience, we are going to make recommendations to.
- **Collaborative Filtering** refers to the use of ratings from multiple users **to predict missing ratings**.



Attribute information about the users → „content-based rec.„

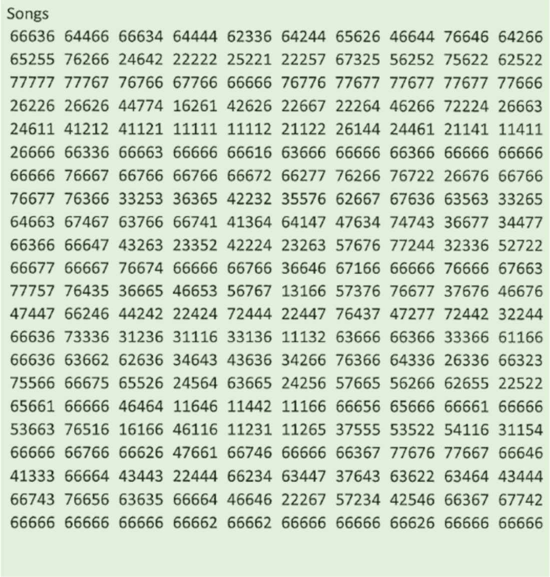
Data about user-item interactions → „collaborative filtering“

demo.Gruppe		Hördauer	SS	Montage	Songs									
3	00000000	OXO	OXOOOOXO	[illegible]	66636	64466	66634	64444	62336	64244	65626	46644	76646	64266
7	00000000	OXO	OOOXOOOX	[illegible]	65255	76266	24642	22222	25221	22257	67325	56252	75622	62522
[illegible]	[illegible]	OXO	OXOOOOXO	[illegible]	77777	77767	76766	67766	66666	76776	77677	77677	77677	77666
2	[illegible]	OXO	XOOOOOXO	[illegible]	26226	26626	44774	16261	42626	22667	22264	46266	72224	26663
6	[illegible]	OXO	[illegible]	OOXOOOXO	24611	41212	41121	11111	11112	21122	26144	24461	21141	11411
5	[illegible]	XOO	[illegible]	OXOOXXXO	26666	66336	66663	66666	66616	63666	66666	66366	66666	66666
4	[illegible]	XOO	[illegible]	XOOOOOXX	66666	76667	66766	66766	66672	66277	76266	76722	26676	66766
7	00000000	OOX	[illegible]	OXOOXXXO	76677	76366	33253	36365	42232	35576	62667	67636	63563	33265
3	00000000	[illegible]	XOOOOOXO	OXOOXOXO	64663	67467	63766	66741	41364	64147	47634	74743	36677	34477
7	00000000	[illegible]	OXOOOOXO	OXOOXXXO	66366	66647	43263	23352	42224	23263	57676	77244	32336	52722
4	00000000	[illegible]	OXOOOOXO	OXOOXXXO	66677	66667	76674	66666	66766	36646	67166	66666	76666	67663
1	00000000	[illegible]	OOOOXOOX	OOXOXOXO	77757	76435	36665	46653	56767	13166	57376	76677	37676	46676
7	00000000	XOO	[illegible]	OXOOOOXX	47447	66246	44242	22424	72444	22447	76437	47277	72442	32244
1	[illegible]	OXO	[illegible]	OXOOOXXO	66636	73336	31236	31116	33136	11132	63666	66366	33366	61166
8	[illegible]	XOO	[illegible]	XOOOOOXX	66636	63662	62636	34643	43636	34266	76366	64336	26336	66323
2	[illegible]	OXO	[illegible]	OXOOXXXO	75566	66675	65526	24564	63665	24256	57665	56266	62655	22522
3	[illegible]	XOO	[illegible]	OOXOOXXO	65661	66666	46464	11646	11442	11166	66656	65666	66661	66666
[illegible]	[illegible]	XOO	XOOOOOXO	[illegible]	53663	76516	16166	46116	11231	11265	37555	53522	54116	31154
4	00000000	OXO	OOOXOOOX	[illegible]	66666	66766	66626	47661	66746	66666	66367	77676	77667	66646
7	00000000	OXO	OXOOOOOO	[illegible]	41333	66664	43443	22444	66234	63447	37643	63622	63464	43444
5	00000000	OXO	OXOOOOXO	OXOOOOXO	66743	76656	63635	66664	46646	22267	57234	42546	66367	67742
8	00000000	XOO	OXOOOOOO	OOXOXOXO	66666	66666	66666	66662	66662	66666	66666	66626	66666	66666

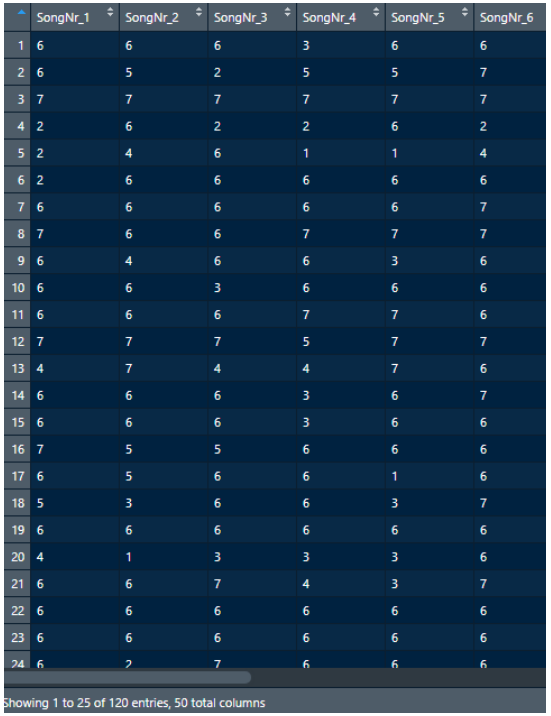
A subset of raw data before wrangling

Wrangling… wrangling… wrangling

- After wrangling, we are left with 50 variables (Songs) and 120 observations (Scores) of these variables.
- The two possible Scores that we have to give special consideration are: „1“ and „4“
- 1: is translated as „I do not know this song“ (missing value?); there are 494 of these in the data
- 4: is translated as „I can‘t say.“ (neutral), 292 times present
- There seems to be a group of people that either are ignorant of the tested music, don‘t like it or can‘t make up their mind.



```
Songs
 66636  64466  66634  64444  62336  64244  65626  46644  76646  64266
 65255  76266  24642  22222  25221  22257  67325  56252  75622  62522
 77777  77767  76766  67766  66666  76776  77677  77677  77677  77666
 26226  26626  44774  16261  42626  22667  22264  46266  72224  26663
 24611  41212  41121  11111  11112  21122  26144  24461  21141  11411
 26666  66336  66663  66666  66616  63666  66666  66366  66666  66666
 66666  76667  66766  66766  66672  66277  76266  76722  26676  66766
 76677  76366  33253  36365  42232  35576  62667  67636  63563  33265
 64663  67467  63766  66741  41364  64147  47634  74743  36677  34477
 66366  66647  43263  23352  42224  23263  57676  77244  32336  52722
 66677  66667  76674  66666  66766  36646  67166  66666  76666  67663
 77757  76435  36665  46653  56767  13166  57376  76677  37676  46676
 47447  66246  44242  22424  72444  22447  76437  47277  72442  32244
 66636  73336  31236  31116  33136  11132  63666  66366  33366  61166
 66636  63662  62636  34643  43636  34266  76366  64336  26336  66323
 75566  66675  65526  24564  63665  24256  57665  56266  62655  22522
 65661  66666  46464  11646  11442  11166  66656  65666  66661  66666
 53663  76516  16166  46116  11231  11265  37555  53522  54116  31154
 66666  66766  66626  47661  66746  66666  66367  77676  77667  66646
 41333  66664  43443  22444  66234  63447  37643  63622  63464  43444
 66743  76656  63635  66664  46646  22267  57234  42546  66367  67742
 66666  66666  66666  66662  66662  66666  66666  66626  66666  66666
```

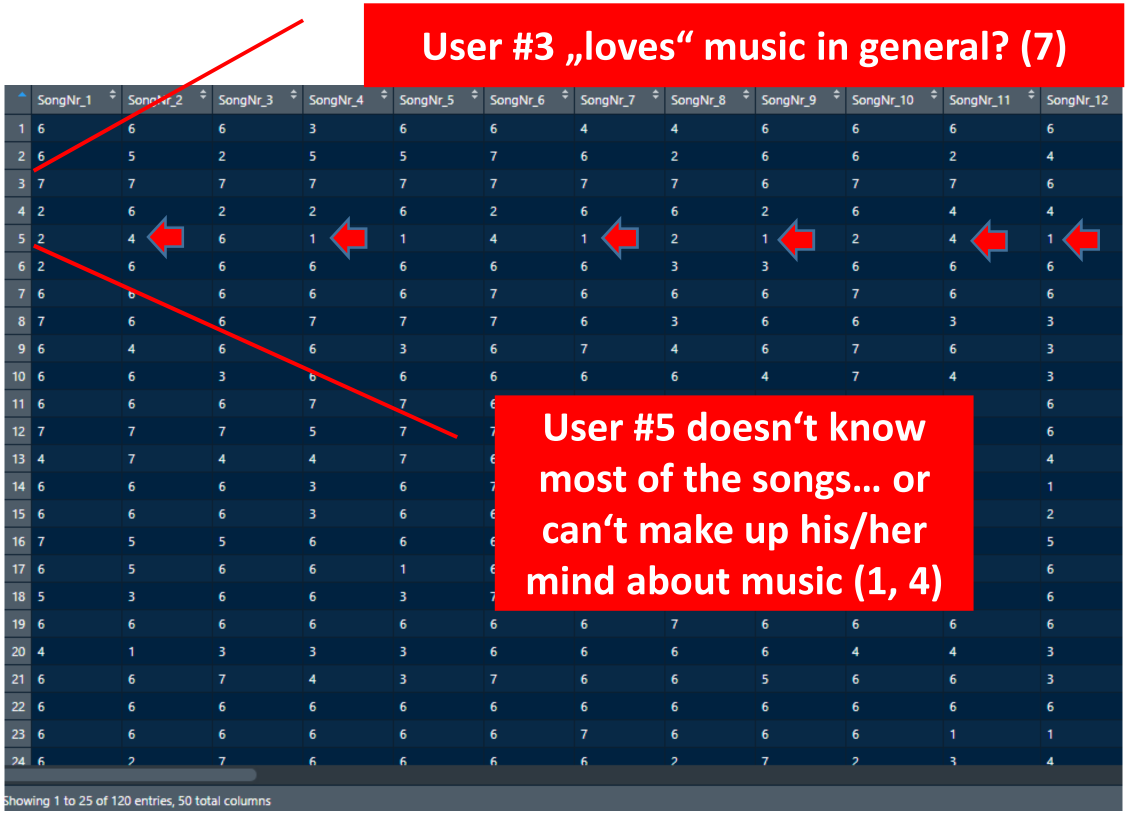


	SongNr_1	SongNr_2	SongNr_3	SongNr_4	SongNr_5	SongNr_6
1	6	6	6	3	6	6
2	6	5	2	5	5	7
3	7	7	7	7	7	7
4	2	6	2	2	6	2
5	2	4	6	1	1	4
6	2	6	6	6	6	6
7	6	6	6	6	6	7
8	7	6	6	7	7	7
9	6	4	6	6	3	6
10	6	6	3	6	6	6
11	6	6	6	7	7	6
12	7	7	7	5	7	7
13	4	7	4	4	7	6
14	6	6	6	3	6	7
15	6	6	6	3	6	6
16	7	5	5	6	6	6
17	6	5	6	6	1	6
18	5	3	6	6	3	7
19	6	6	6	6	6	6
20	4	1	3	3	3	6
21	6	6	7	4	3	7
22	6	6	6	6	6	6
23	6	6	6	6	6	6
24	6	2	7	6	6	6

Showing 1 to 25 of 120 entries, 50 total columns

Wrangling… wrangling… wrangling

- After wrangling, we are left with 50 variables (Songs) and 120 observations (Scores of tested people) of these variables.
- For the next steps it is assumed that the market for attention to popular music is efficient, i.e. that the score „1" can be interpreted as a maximum disinterest in that particular song.
- Scores of 4 then represent true indifference and we have a true scale of 1:7.
- First glances at the test scores let us to postulate at least three patterns: Very positive, very negative and mixed scores.
- The question, whether the first two are essentially similar and just shifted patterns will have to be adressed later.
- **Let's have a look at a particular song's test pattern (the one with the highest mean score) on the next slide before going on to distances**



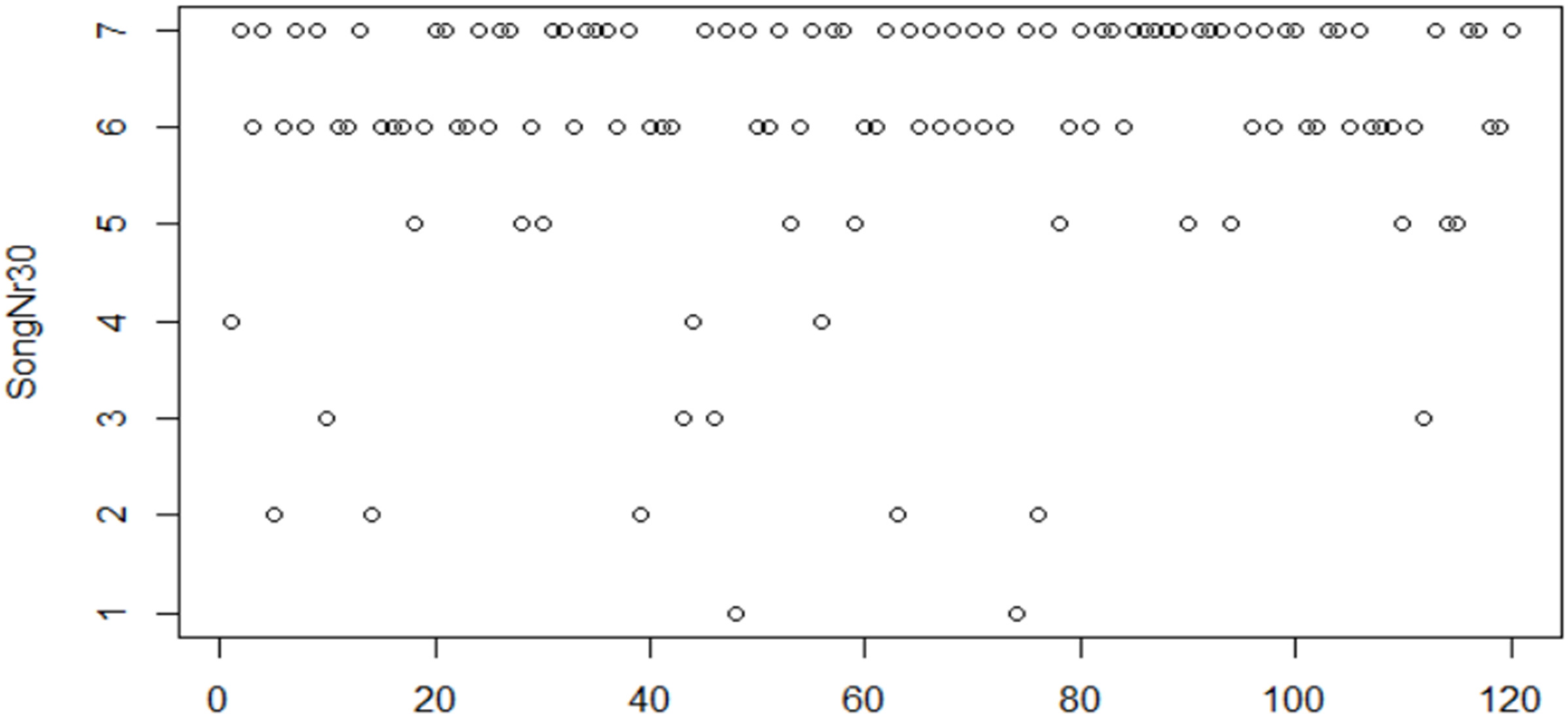
User #3 „loves" music in general? (7)

	SongNr_1	SongNr_2	SongNr_3	SongNr_4	SongNr_5	SongNr_6	SongNr_7	SongNr_8	SongNr_9	SongNr_10	SongNr_11	SongNr_12
1	6	6	6	3	6	6	4	4	6	6	6	6
2	6	5	2	5	5	7	6	2	6	6	2	4
3	7	7	7	7	7	7	7	7	6	7	7	6
4	2	6	2	2	6	2	6	6	2	6	4	4
5	2	4	6	1	1	4	1	2	1	2	4	1
6	2	6	6	6	6	6	6	3	3	6	6	6
7	6	6	6	6	6	7	6	6	6	7	6	6
8	7	6	6	7	7	7	6	3	6	6	3	3
9	6	4	6	6	3	6	7	4	6	7	6	3
10	6	6	3	6	6	6	6	6	4	7	4	3
11	6	6	6	7	7	6	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	6
12	7	7	7	5	7	7	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	6
13	4	7	4	4	7	6	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	4
14	6	6	6	3	6	7	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	1
15	6	6	6	3	6	6	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	2
16	7	5	5	6	6	6	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	5
17	6	5	6	6	1	6	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	6
18	5	3	6	6	3	7	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	6
19	6	6	6	6	6	6	6	7	6	6	6	6
20	4	1	3	3	3	6	6	6	6	4	4	3
21	6	6	7	4	3	7	6	6	5	6	6	3
22	6	6	6	6	6	6	6	6	6	6	6	6
23	6	6	6	6	6	6	7	6	6	6	1	1
24	6	2	7	6	6	6	6	2	7	2	3	4

Showing 1 to 25 of 120 entries, 50 total columns

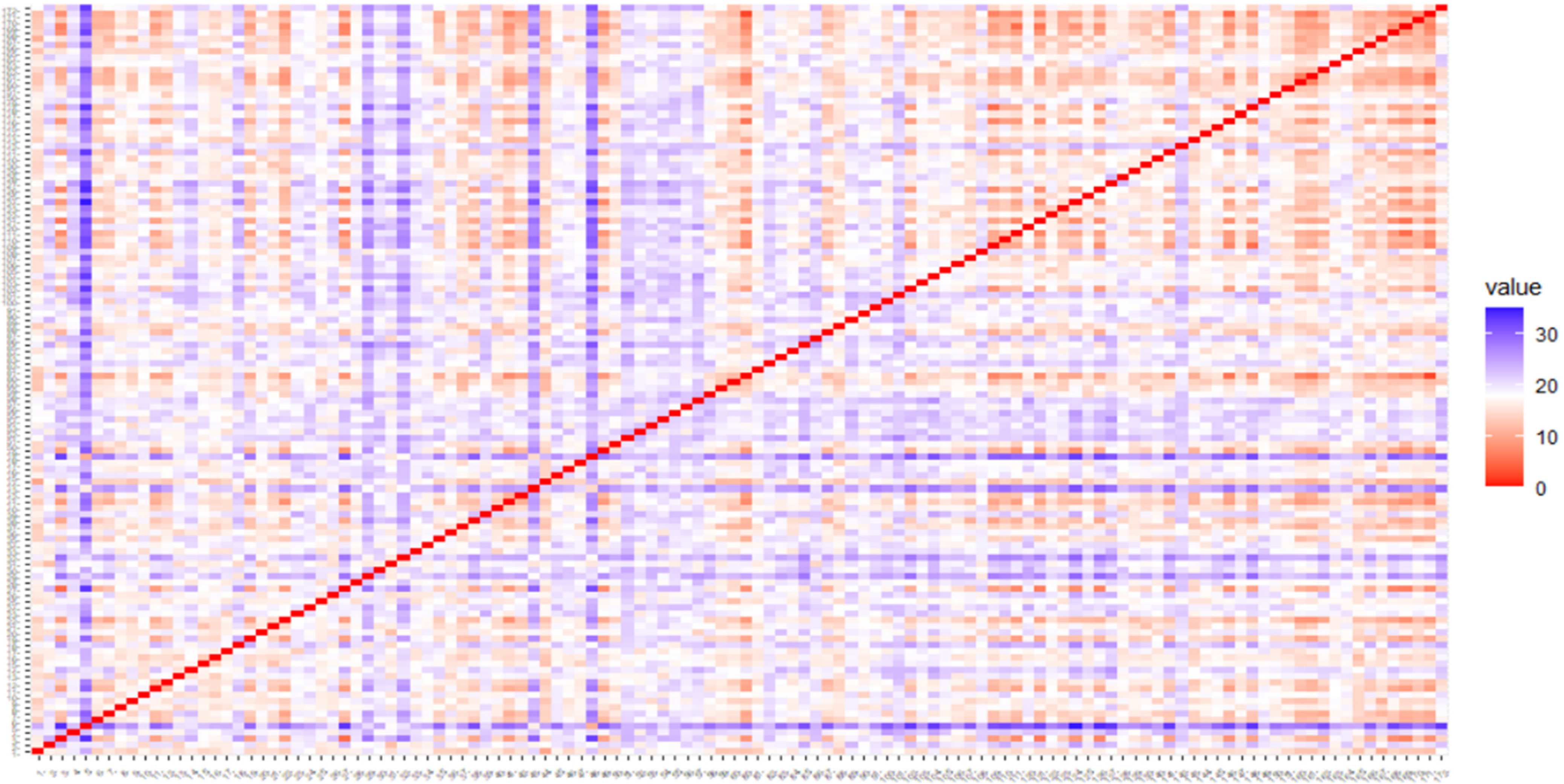
User #5 doesn't know most of the songs… or can't make up his/her mind about music (1, 4)

Visual: (120) Scores of SongNr_30

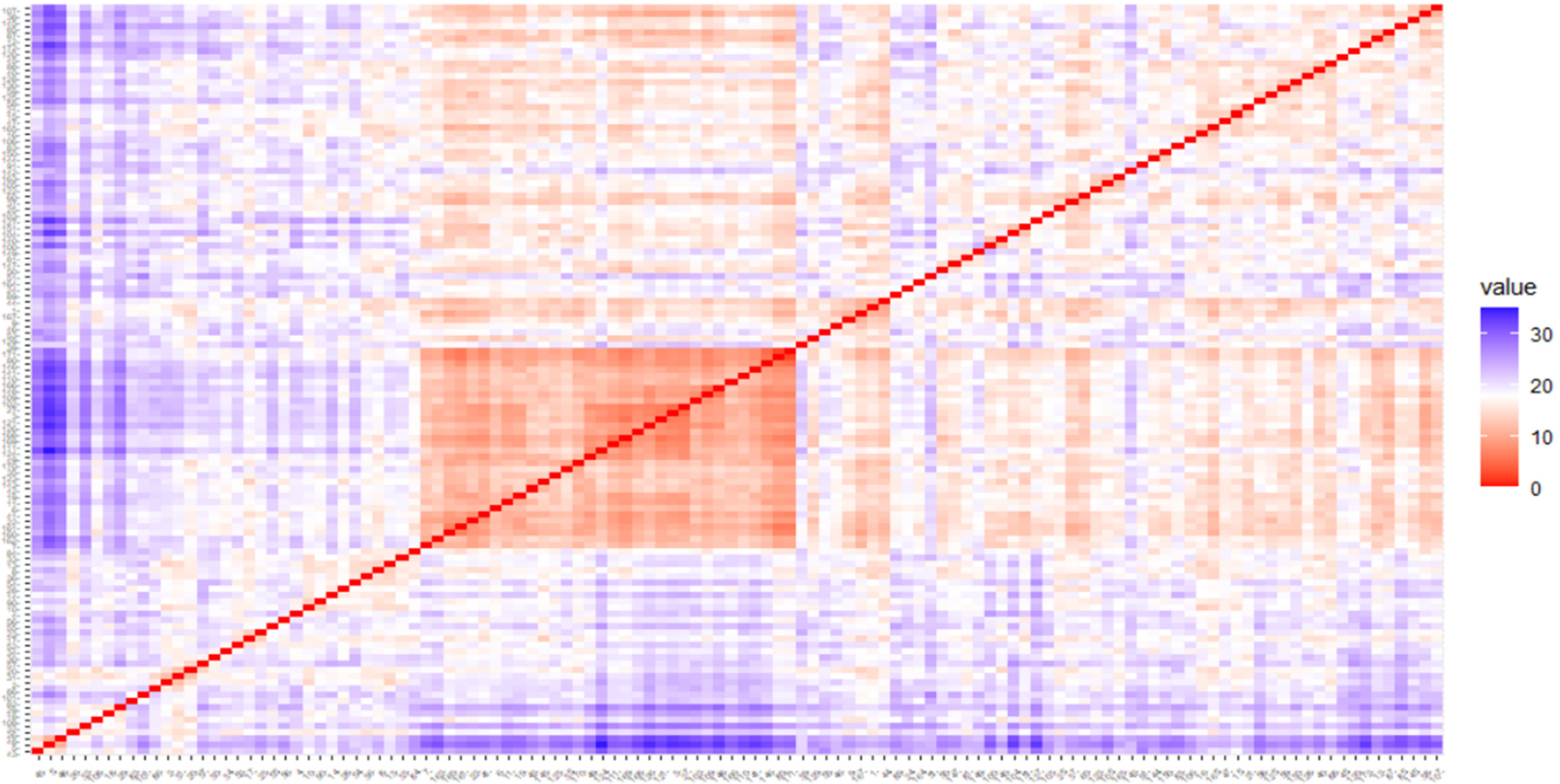


(Dynoro, In My Mind: https://open.spotify.com/track/sDtfiOiYSSu01r78p6bZLw**)**

Distance Visualisation (fviz_dist, „Euclidean“, unordered) (invariant to scaling)

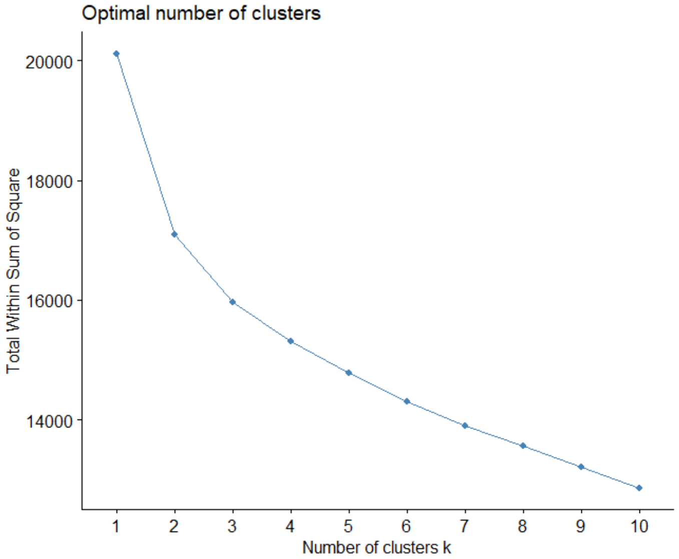


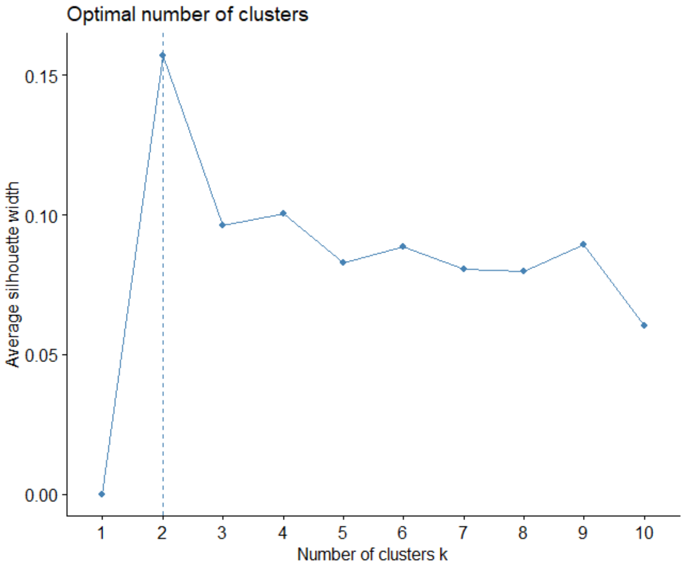
Ordered Dissimilarity Image (ODI) (not completely invariant to scaling)



Clustering Approaches

- We do not have to standardize as all data is on the same scale
- k-Means Clustering uses Euclidean distances, which should give a first idea about clustering possibilities
- Using the Elbow Approach (via fviz_nbclust, method = „wss“), the optimal number of clusters could be 3. This would split the data into three clusters of sizes 25, 42, 53.
- Using the „Silhouette Method“ (againg via fviz_nbclust), the optimal number of clusters should be 2. This splits the data into two groups of almost equal size (58, 62).





Overview

- The Task
- Obtaining Data from a Music Test Survey
- Project Vision
- Wrangling… Wrangling… Wrangling
- EDA – Explorative Data Analysis (k-Means Clustering)
- **Recommenderlab Package (Recommender System)**
- **RStudio – Guideline Markdown**