Clustering Analysis

Abby Smith

11/10/2020

```
knitr::opts_chunk$set(echo=F, cache=T, results='hide', message = F, warning = F)
```

Questions of Interest

Edovo is interested in a clustering of the course content. They have over 15,000 courses. Each course has tags attached to it, representing different topics that are covered in the course (such as "CHRISTIANITY"), but also features of the course itself (such as "INTERMEDIATE").

We have the tags for each course, and the navigation map of courses.

We are interested in:

- What courses have similar tags?
- What courses have similar features?
- What are the most important features in determining similar tags?
- How can we incorporate the navigation map in our clustering?
- What do the users look like that are in a given tag cluster? In terms of time as active user, facility, etc.

(**UPDATE 11/16**): Not only do we want a clustering analysis of users + content; we want to determine an *engagement score* for each user on the platform.

- What clusters of content are "quality"? Can we provide a quantitative metric for that?
- How can we collapse this into a score?

The Data

We get the data into a **wide** format- each row represents a course, its title, lesson count, total page count,number of items across the pages and then columns representing each tag and a corresponding column with that tag's ID. Because a course has up to 32 different tags, there are $32 \times 2 = 64$ different columns.

Question of Representation: Long vs. Wide?

Key Decisions in Clustering

- Do we one-hot encode categorical variables?
- How many clusters do we use?
- Should we use a joint dimension reduction + clustering approach?
- How to deal with many, many categorical features (tags)?

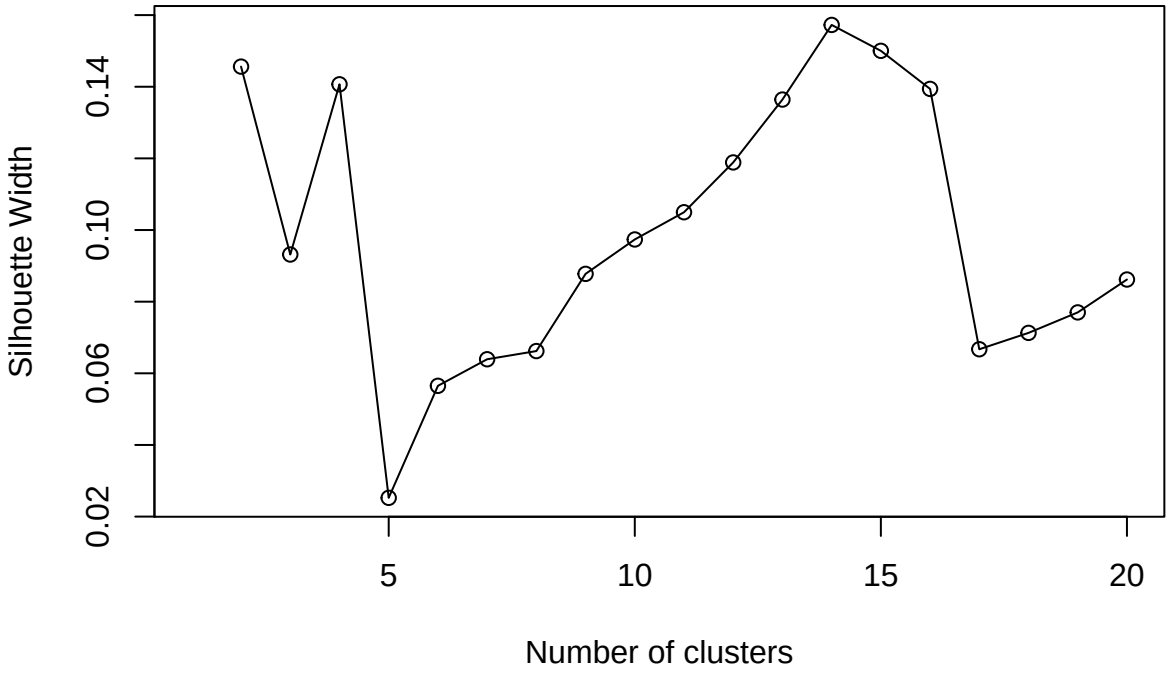
Clustering Approaches

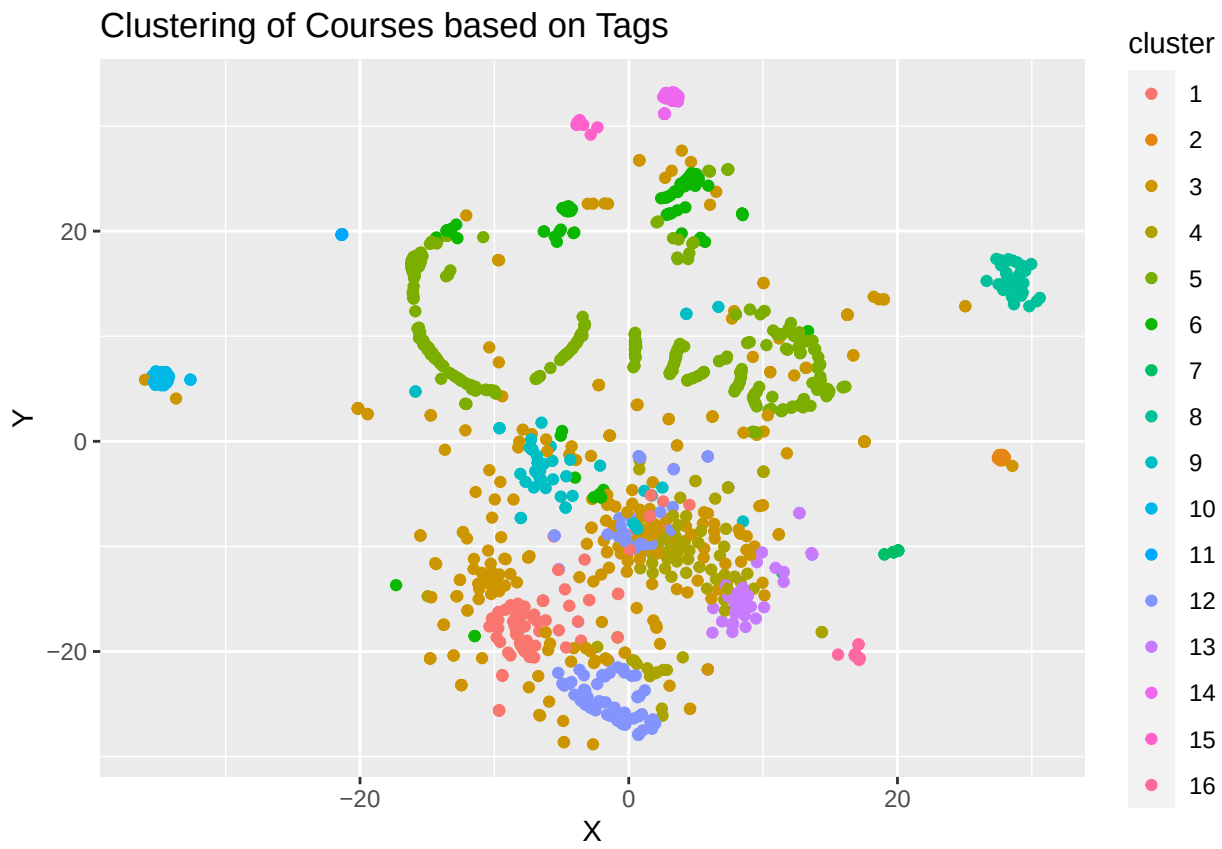
In our clustering we want to prioritize **numerical** features over **categorical**. More specifically– we want to account for hierchachy of the course in the JSON navigation map. If two users both watched just 2 pages of a very specifically tagged activity– we want to weight that higher than tags that are further up in the hiercharchy.

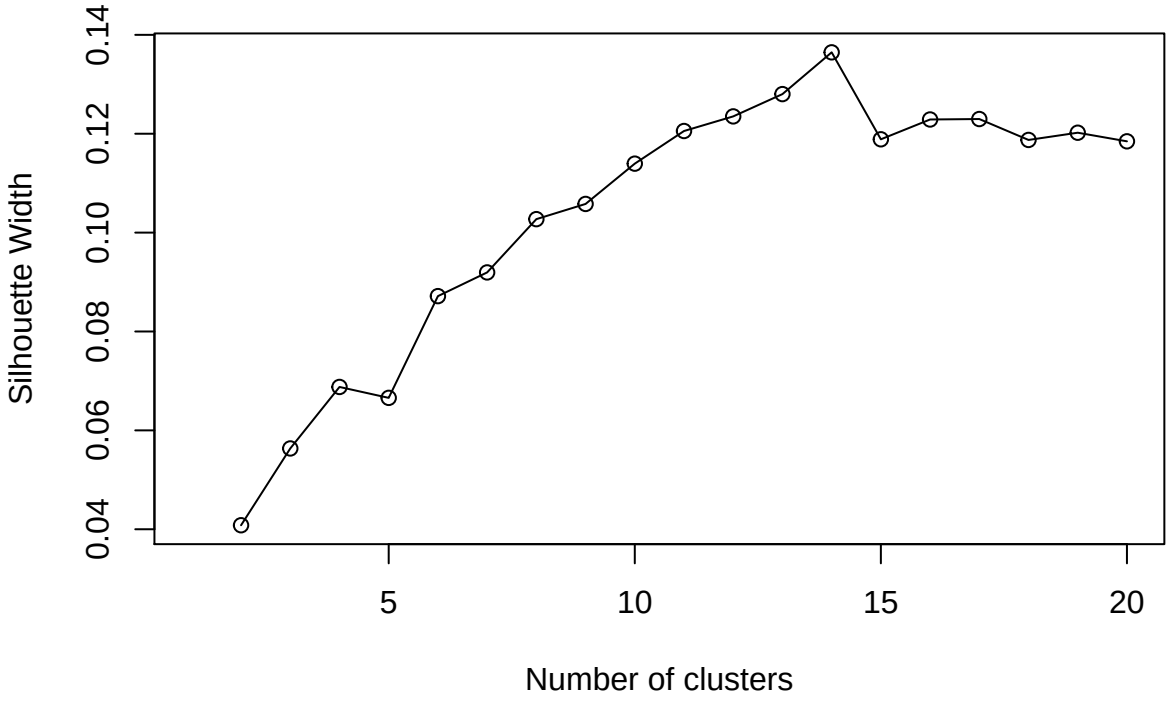
K-Means with Gower Distance

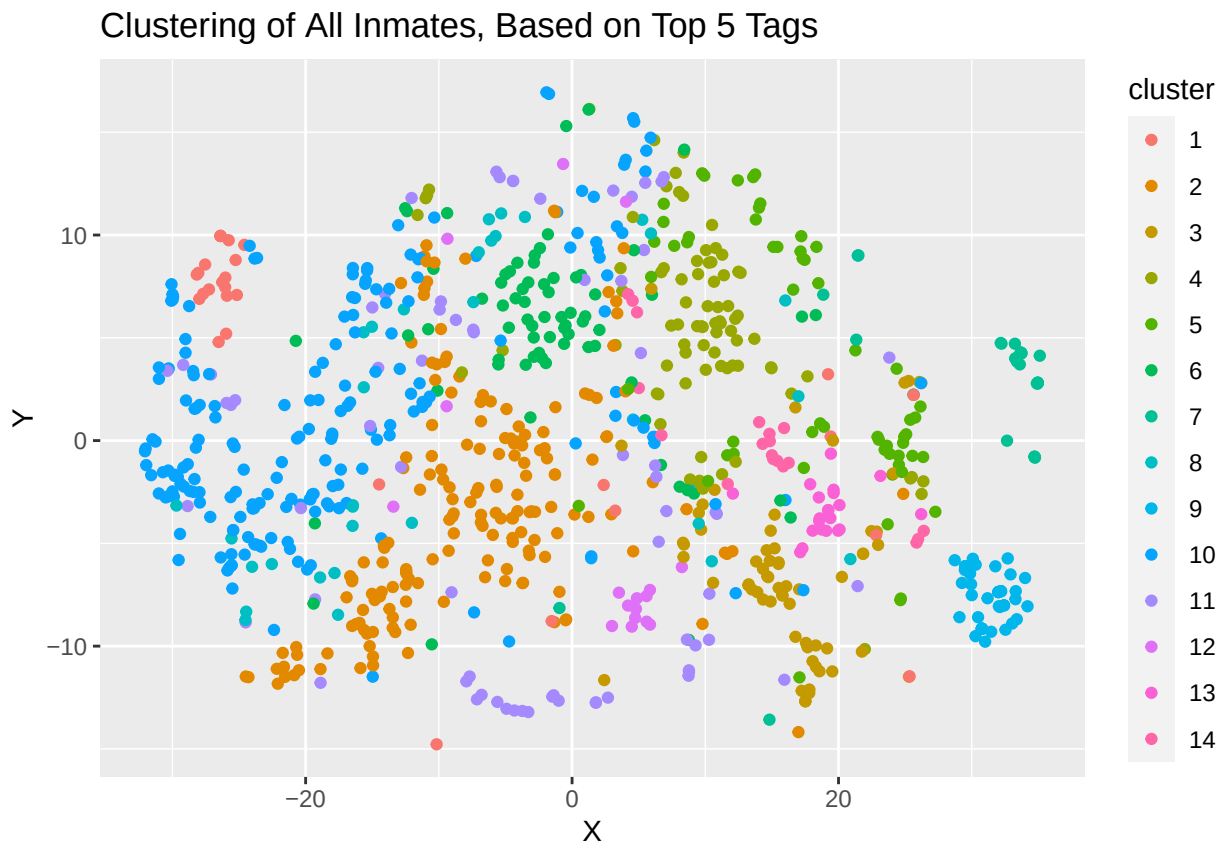
In order for a yet-to-be-chosen algorithm to group observations together, we first need to define some notion of (dis)similarity between observations. A popular choice for clustering is Euclidean distance. However, Euclidean distance is only valid for continuous variables, and thus is not applicable here. In order for a clustering algorithm to yield sensible results, we have to use a distance metric that can handle mixed data types. In this case, we will use something called Gower distance.

We will look at a few different approaches from this clustering. We will plot a low dimensional representation for each of these clusterings



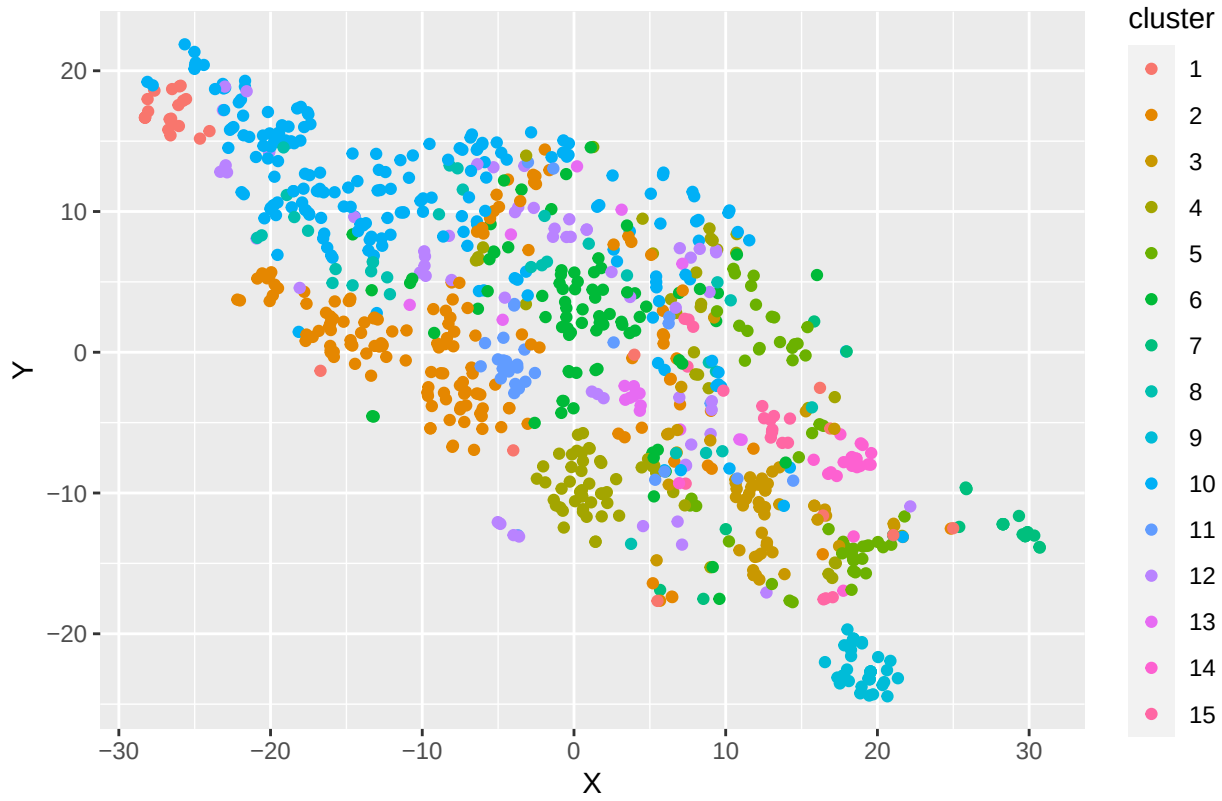

Clustering of Courses based on Tags
cluster
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
20
0
−20
Y
−20
0
20
X


Silhouette Width
0.04
0.06
0.08
0.10
0.12
0.14
5
10
15
20
Number of clusters

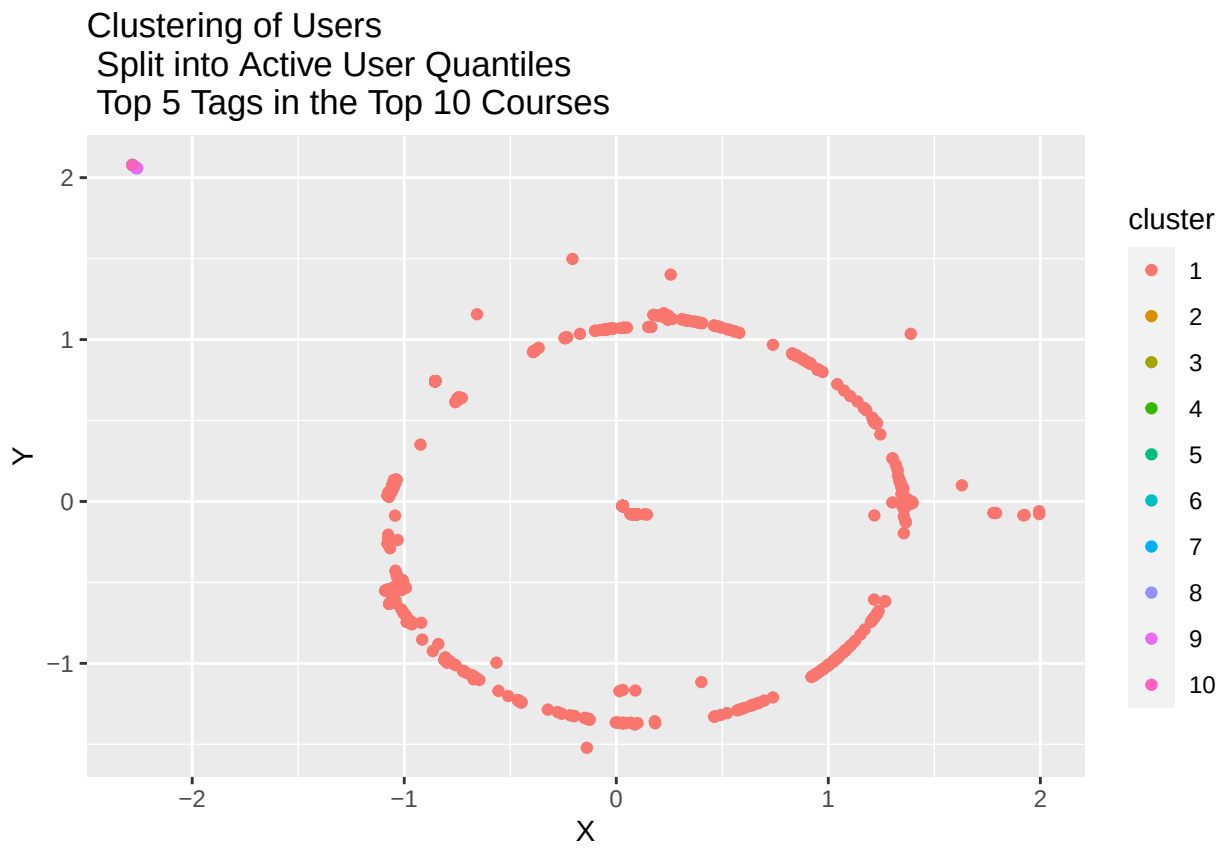


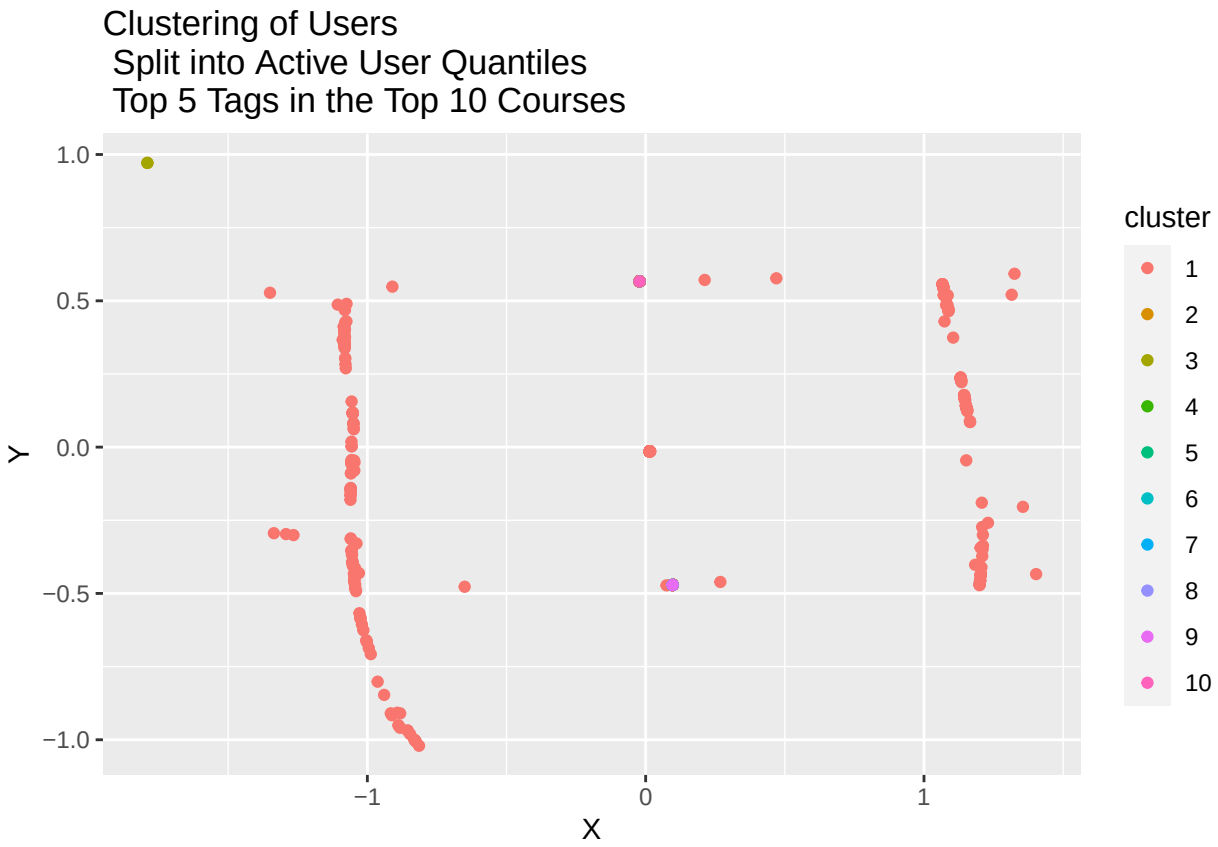
We're also curious, what do inmates in the same cluster have in common?

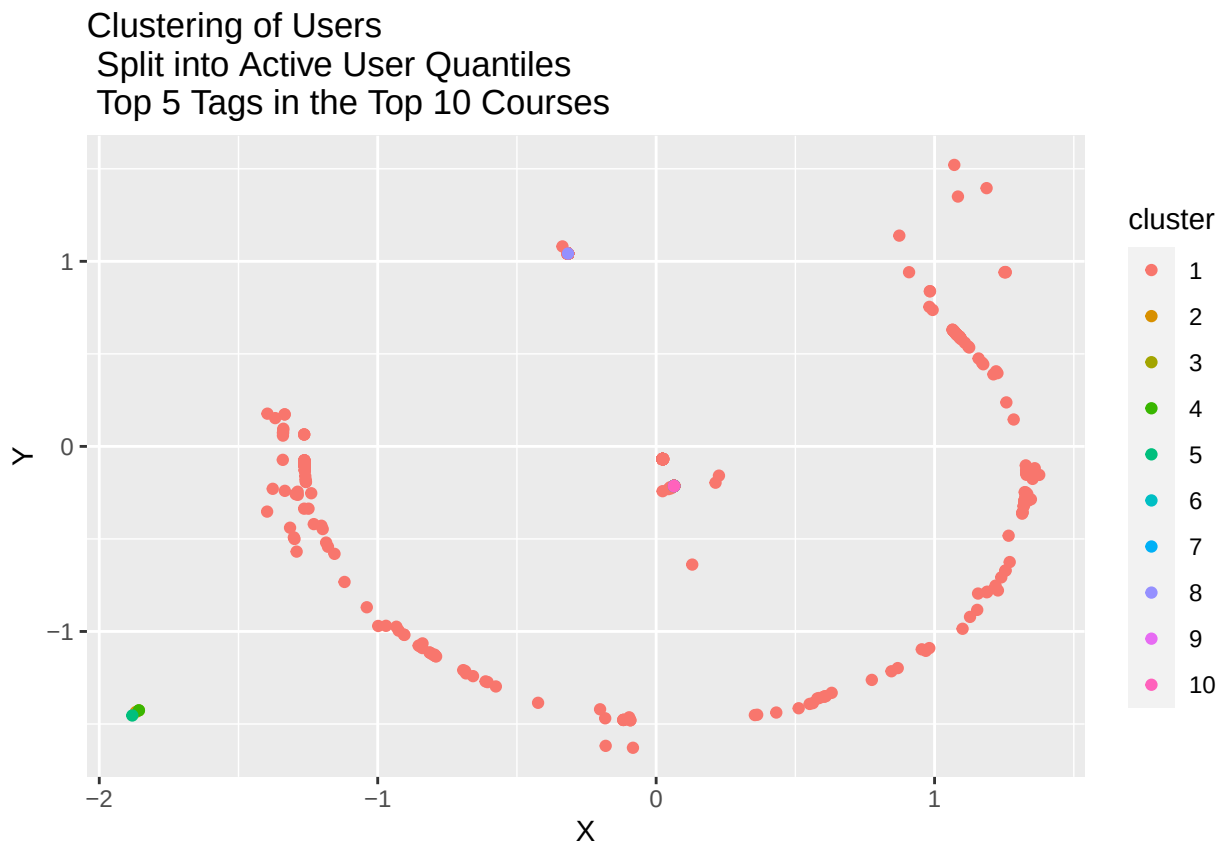
Because we're interested in the engagement patterns of users, it's probably a good idea for us to account for time as an active user in our clustering.

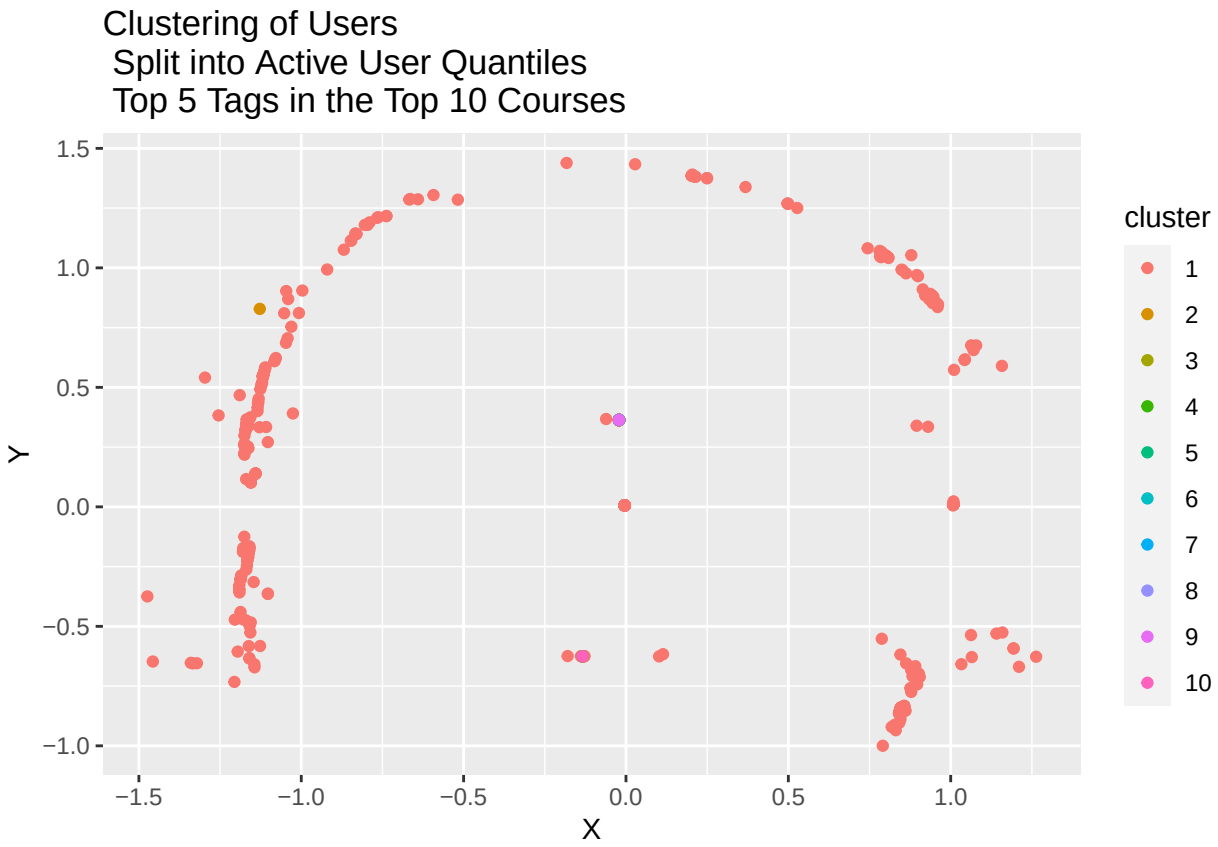


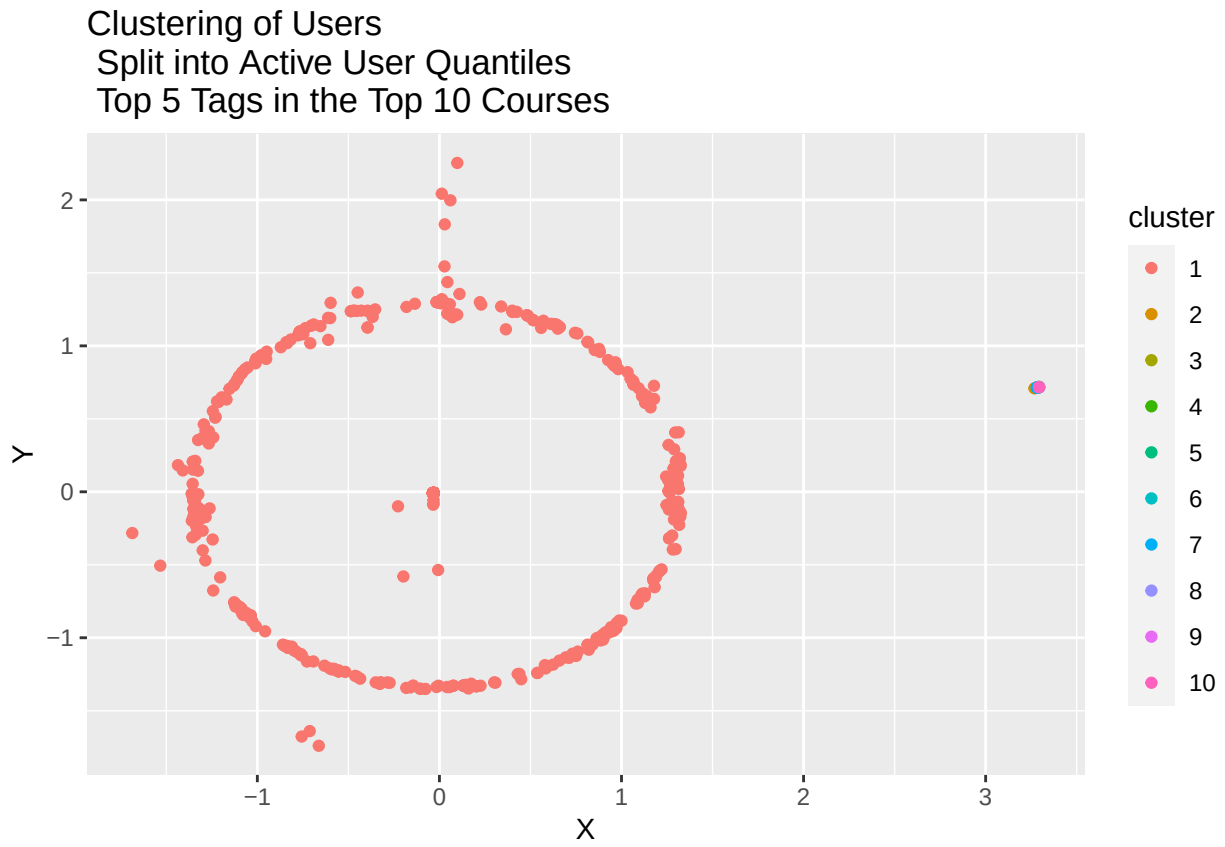
Next, we split the time as an active user up into quantiles and perform the clustering.


Clustering of Users
Split into Active User Quantiles
Top 5 Tags in the Top 10 Courses
Y
X
2
1
0
−1
−2
−1
0
1
2
cluster
1
2
3
4
5
6
7
8
9
10

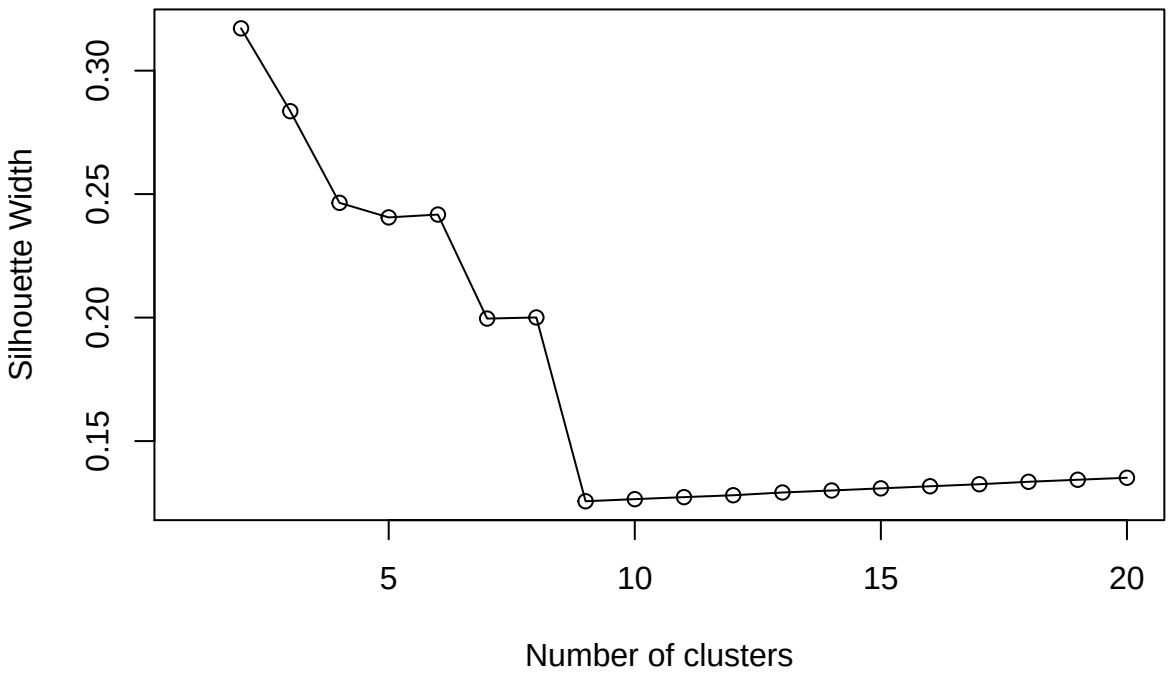

Clustering of Users
Split into Active User Quantiles
Top 5 Tags in the Top 10 Courses
1.0
0.5
0.0
−0.5
−1.0
Y
−1
0
1
X
cluster
1
2
3
4
5
6
7
8
9
10

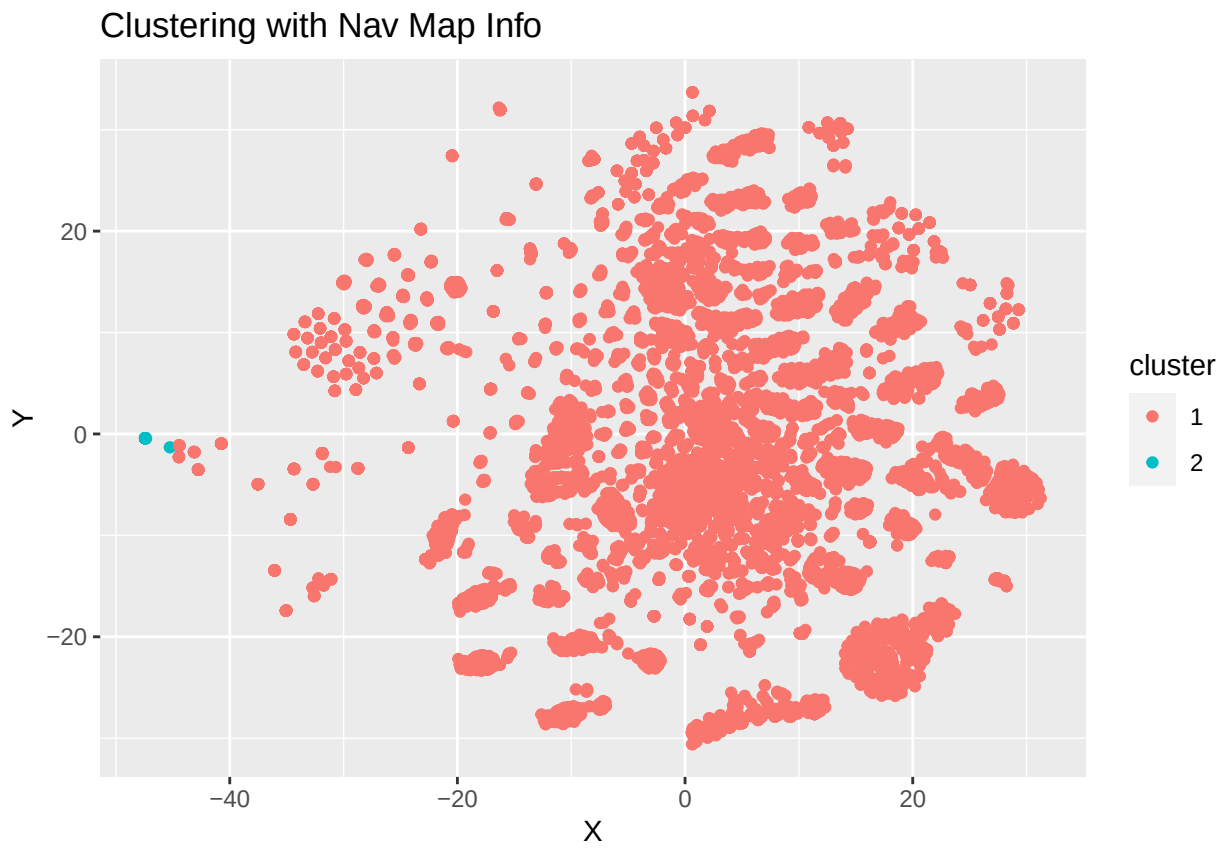

Clustering of Users
Split into Active User Quantiles
Top 5 Tags in the Top 10 Courses
Y
X
1
0
−1
−2
−1
0
1
cluster
1
2
3
4
5
6
7
8
9
10


Clustering of Users
Split into Active User Quantiles
Top 5 Tags in the Top 10 Courses
1.5
1.0
0.5
0.0
−0.5
−1.0
Y
−1.5
−1.0
−0.5
0.0
0.5
1.0
X
cluster
1
2
3
4
5
6
7
8
9
10

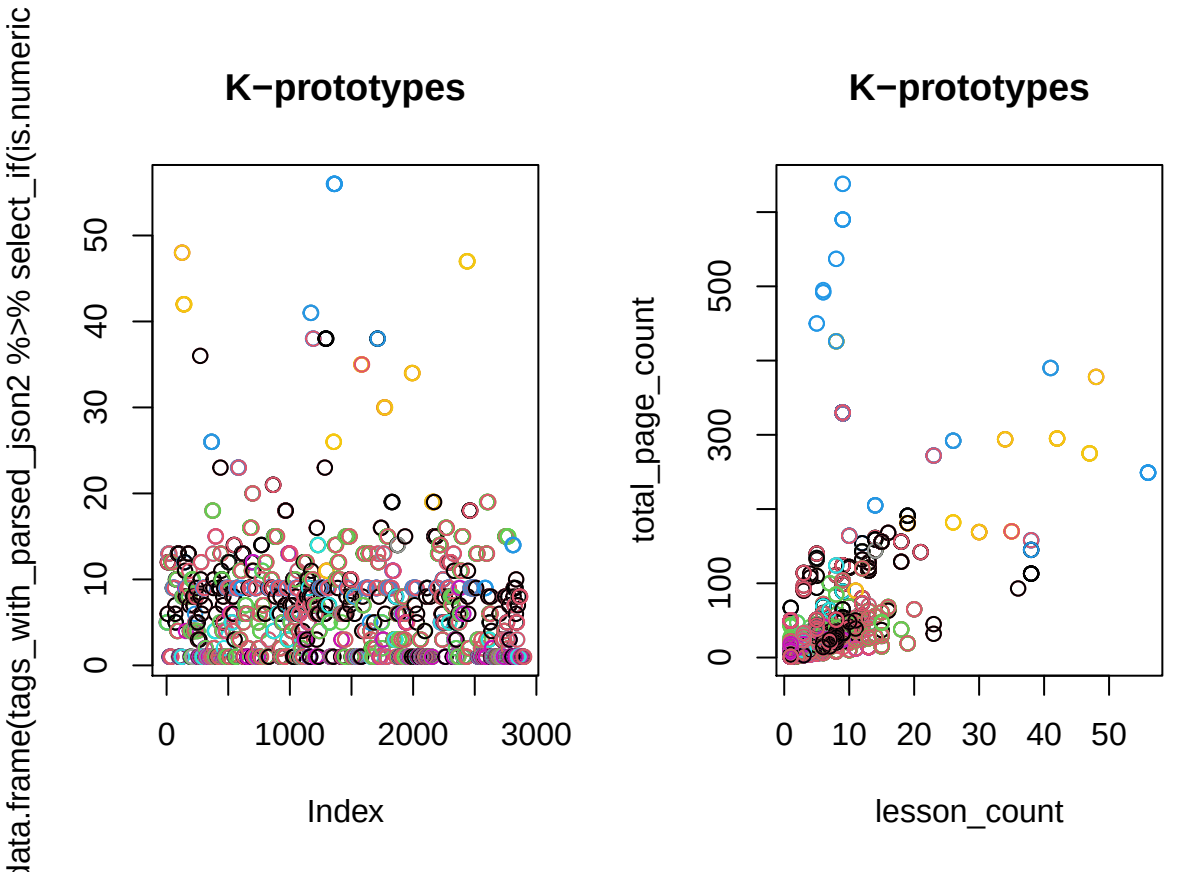


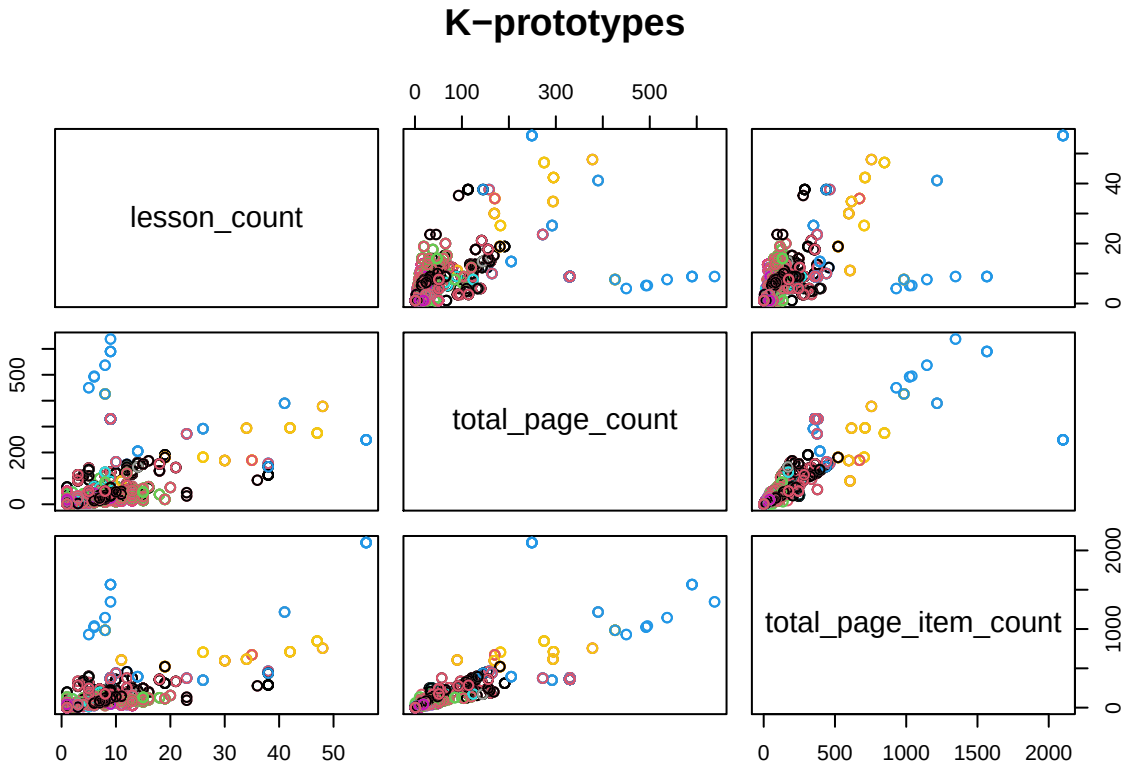
THIS IS A BIG TO-DO: Incorporating nav-map information. Here we have features like `total_pages` and `total_page_items` but we need to play around with weighting!

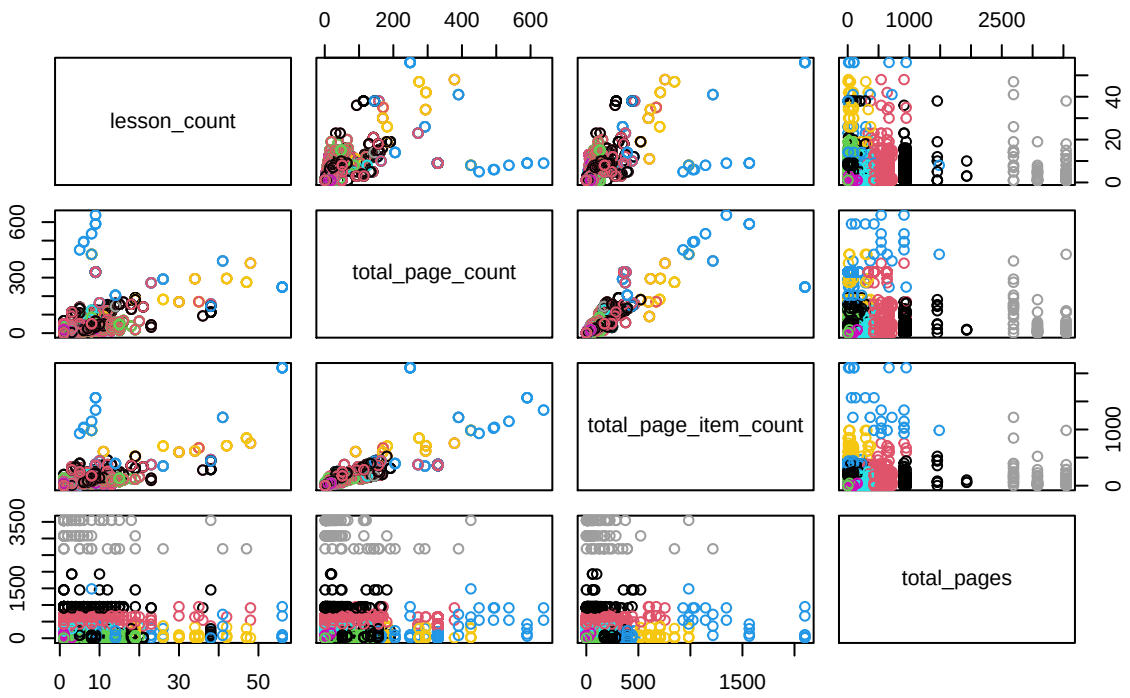

Silhouette Width
0.15
0.20
0.25
0.30
5
10
15
20
Number of clusters

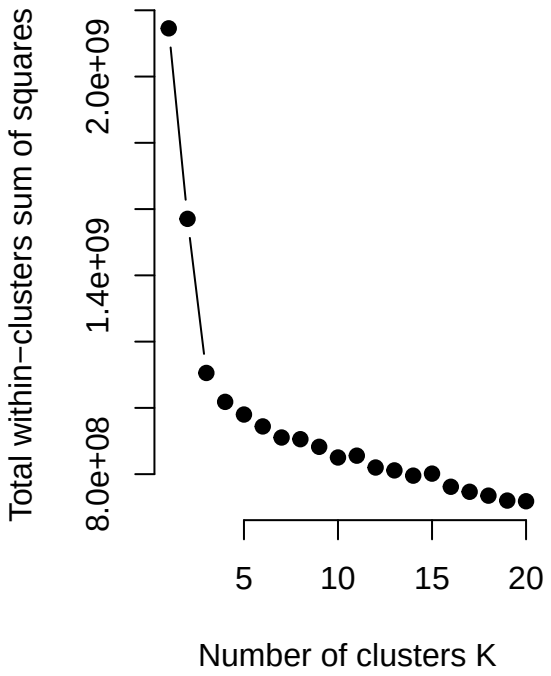

Clustering with Nav Map Info
20
0
−20
Y
−40
−20
0
20
X
cluster
1
2

K Prototypes

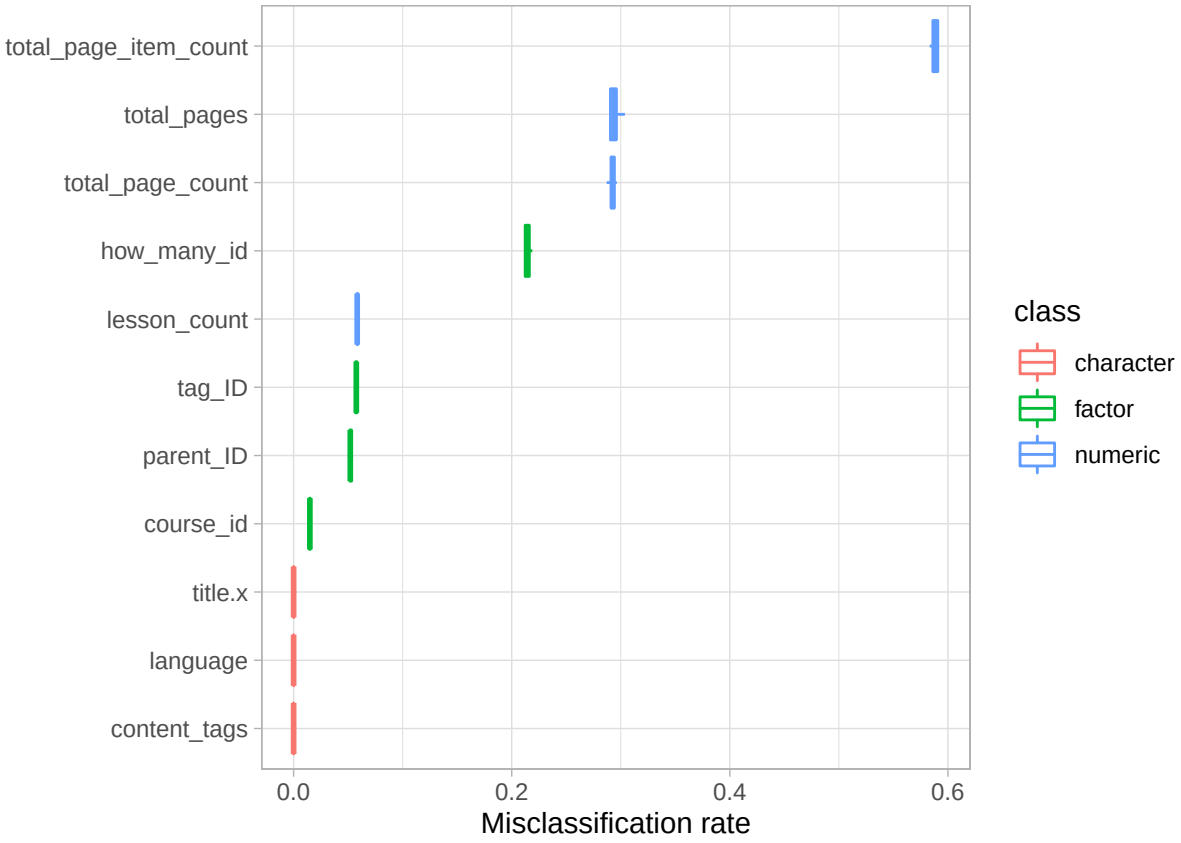





K−prototypes
lesson_count
total_page_count
total_page_item_count
total_pages
0
200
400
600
0
1000
2500
0
20
40
0
300
600
0
1000
0
1500
3500
0
10
30
50
0
500
1500


Total within−clusters sum of squares
8.0e+08
1.4e+09
2.0e+09
5
10
15
20
Number of clusters K

What features are most important?

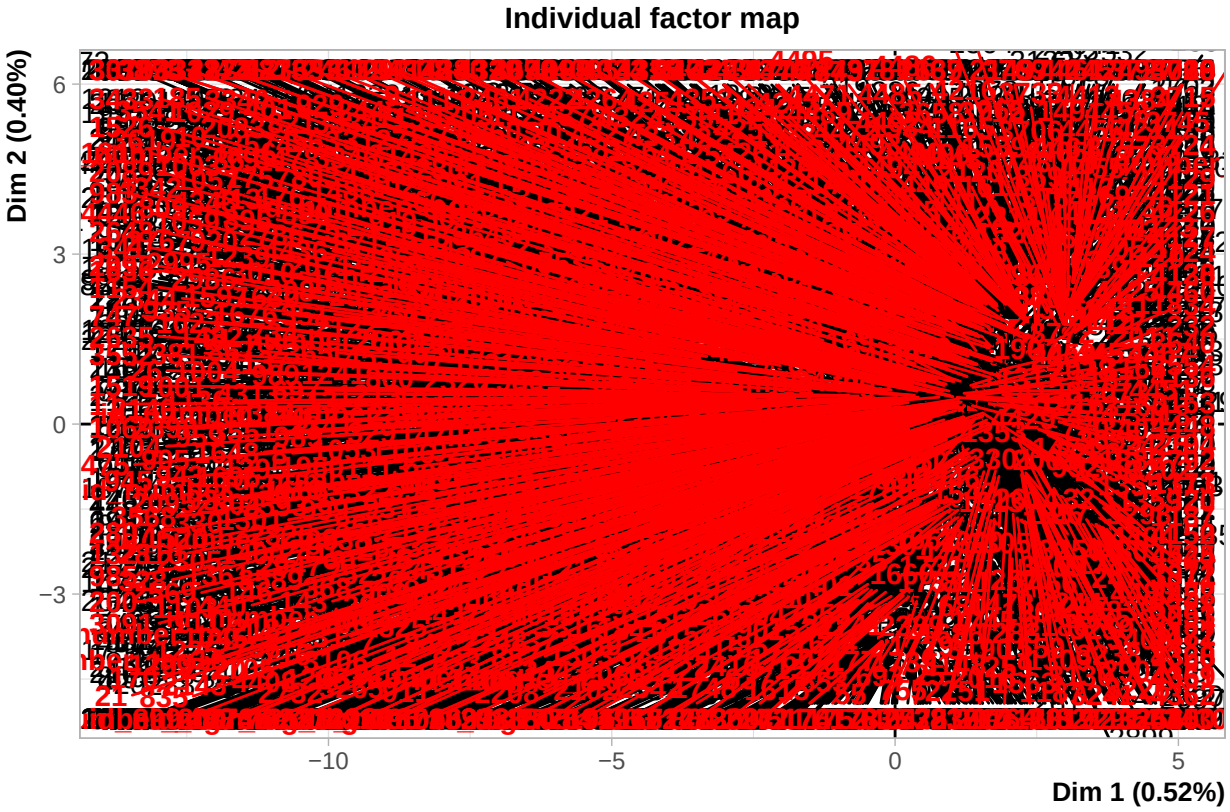


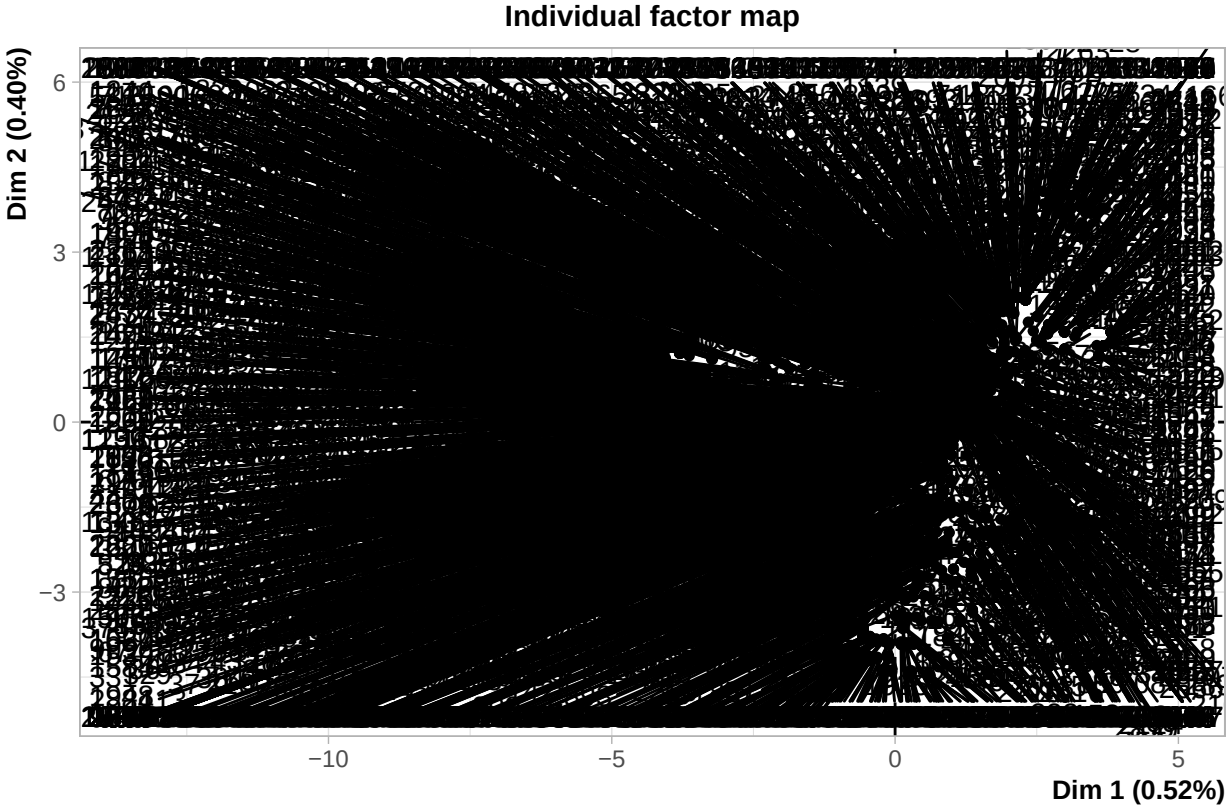
Dimension Reduction? Or at least PCA/ FAMD

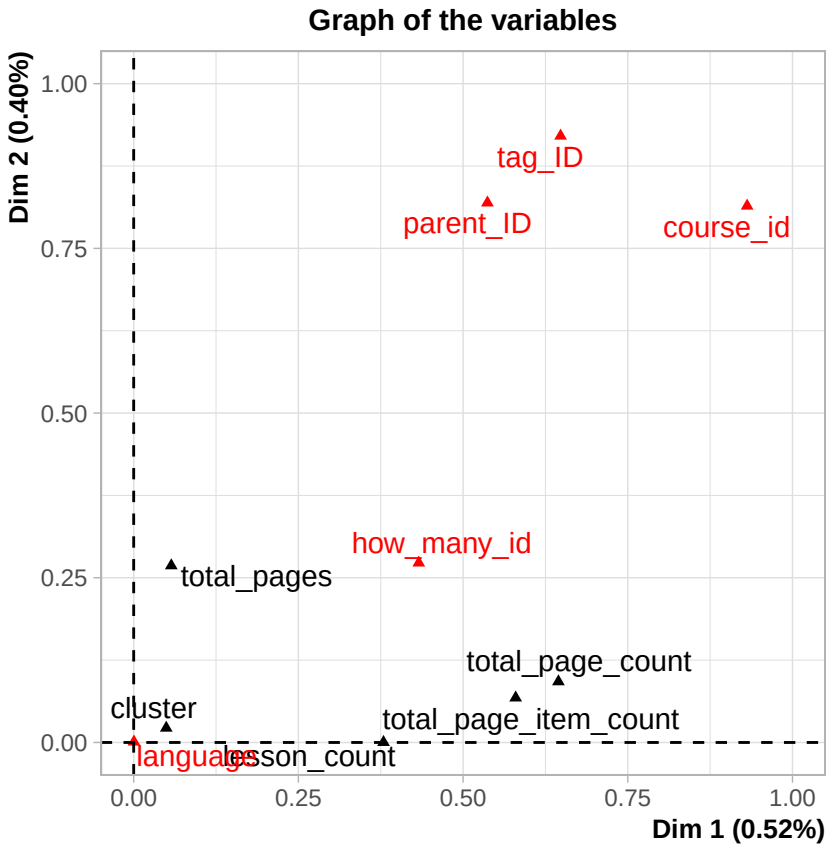
Assuming we represent the data in **wide** format– we may want to look into joint dimension reduction + clustering algorithms. UPDATE: this ay not be a good idea unless we have exclusively numerical features.

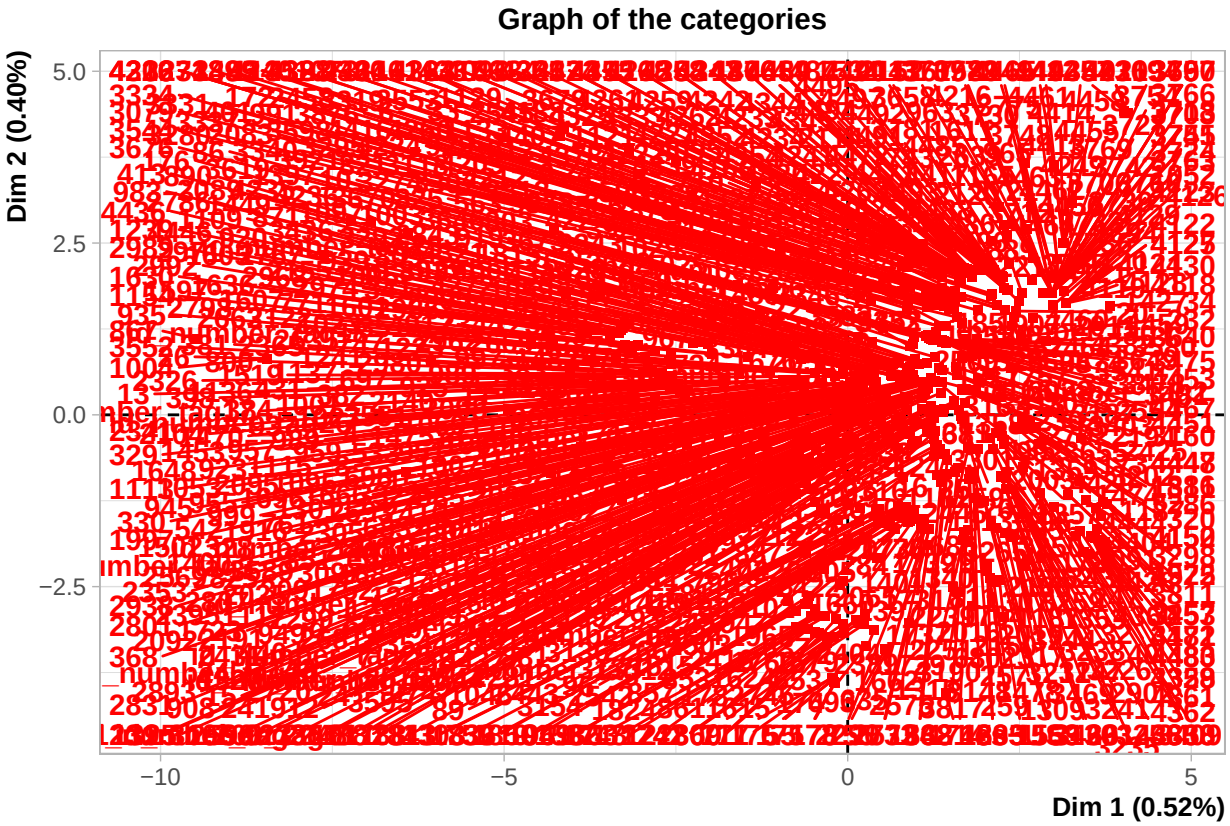
We note that the first two components when running these things capture **very little variance**. So how useful is this for anything? :'(

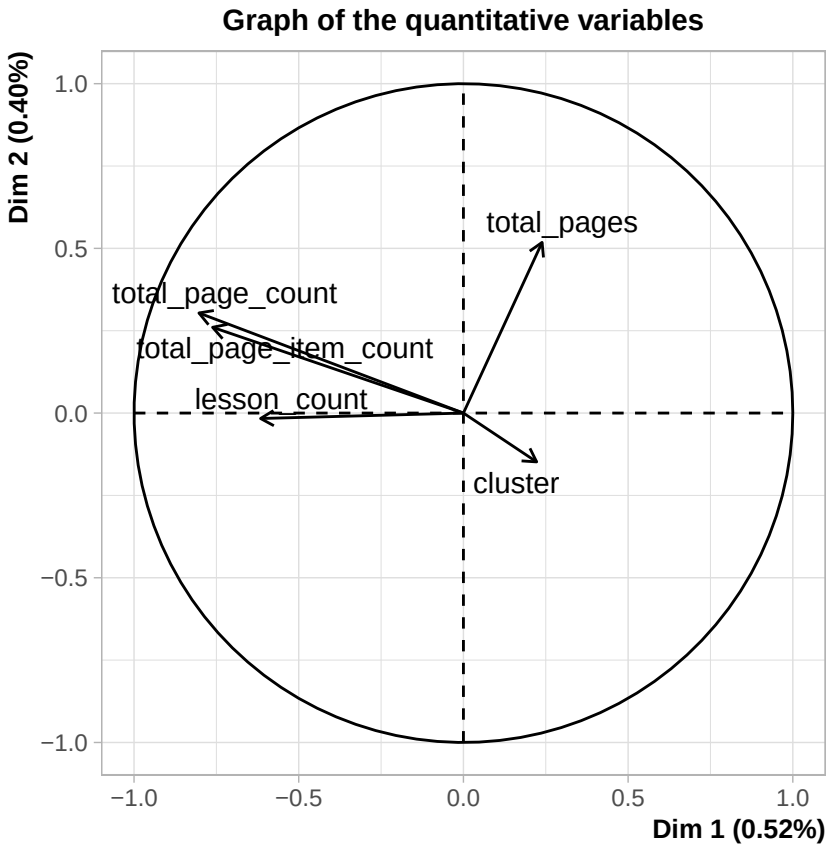
I think this is necessary if we want to make the clustering unit **INMATES**–the data frame would have many columns if we had each tag ID for each course associated with the inmate, so I think this definetely has merits.

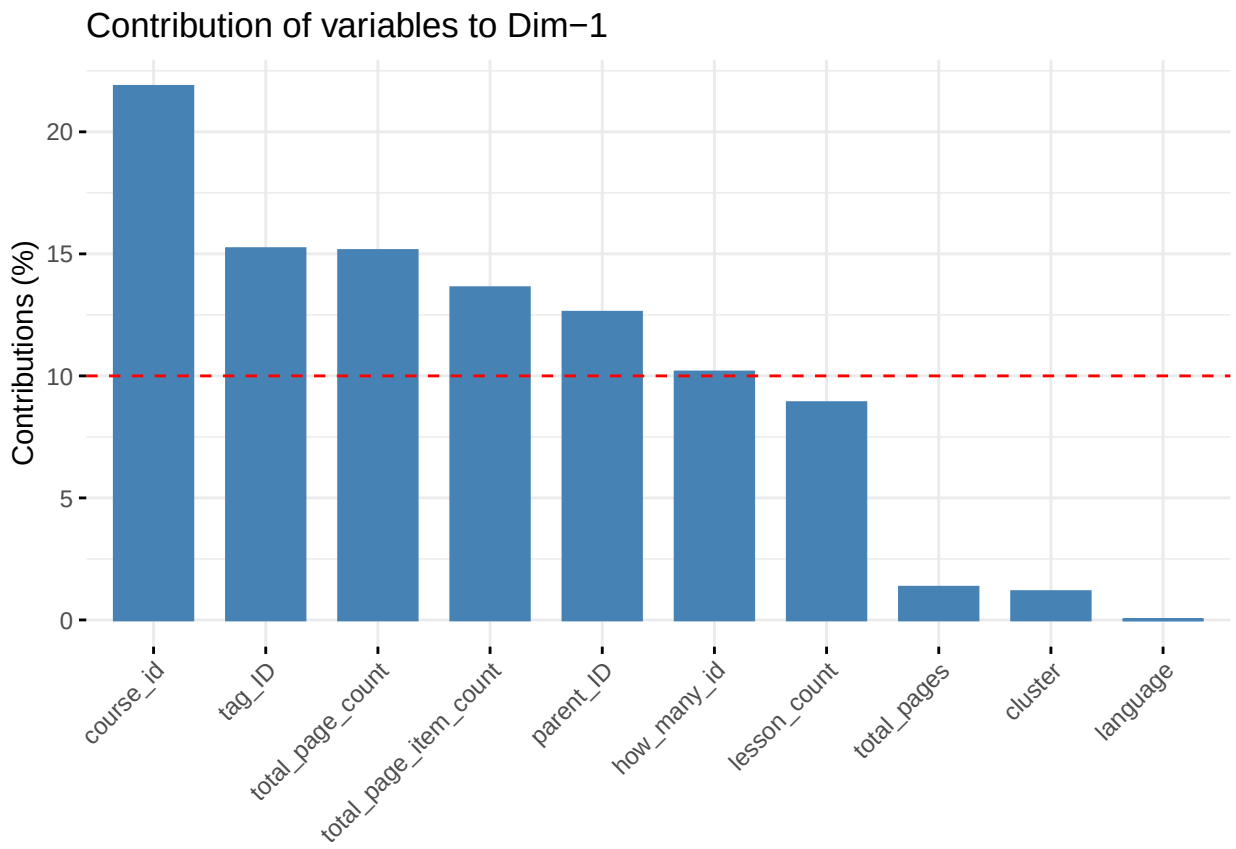

Individual factor map
Dim 2 (0.40%)
6
3
0
−3
−10
−5
0
5
Dim 1 (0.52%)

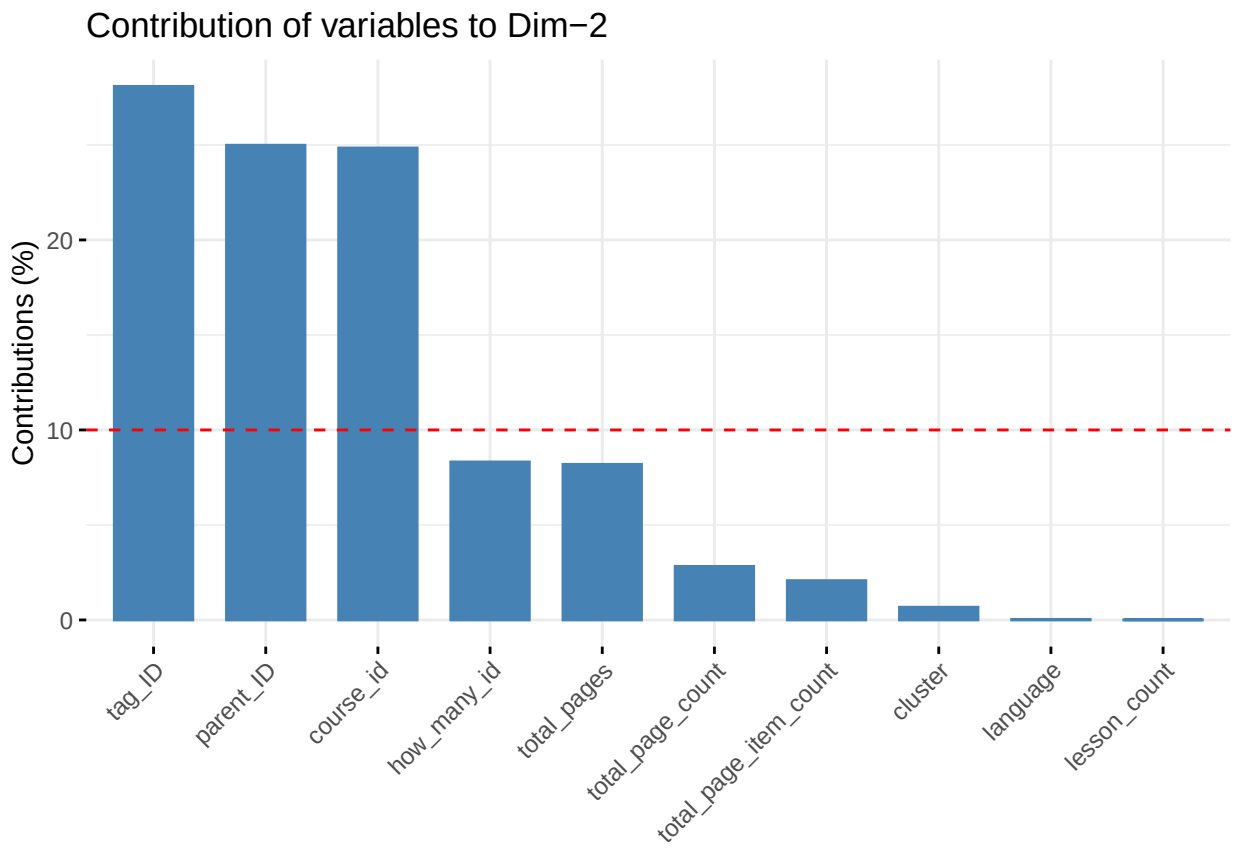

Individual factor map
Dim 2 (0.40%)
6
3
0
−3
−10
−5
0
5
Dim 1 (0.52%)


Graph of the variables
Dim 2 (0.40%)
1.00
0.75
0.50
0.25
0.00
tag_ID
parent_ID
course_id
how_many_id
total_pages
total_page_count
cluster
total_page_item_count
languages
lesson_count
0.00
0.25
0.50
0.75
1.00
Dim 1 (0.52%)


Graph of the categories
Dim 2 (0.40%)
5.0
2.5
0.0
−2.5
−10
−5
0
5
Dim 1 (0.52%)


Graph of the quantitative variables
Dim 2 (0.40%)
1.0
0.5
0.0
−0.5
−1.0
total_pages
total_page_count
total_page_item_count
lesson_count
cluster
−1.0
−0.5
0.0
0.5
1.0
Dim 1 (0.52%)


Contribution of variables to Dim−1
Contributions (%)
20
15
10
5
0
course_id
tag_ID
total_page_count
total_page_item_count
parent_ID
how_many_id
lesson_count
total_pages
cluster
language


Contribution of variables to Dim−2
Contributions (%)
20
10
0
tag_ID
parent_ID
course_id
how_many_id
total_pages
total_page_count
total_page_item_count
cluster
language
lesson_count

)