BUSINESS ANALYTICS PROJECT

Author:

- Magno Alessandro: 4478234

Data source:

- Aulaweb Database course log 2019
- Aulaweb Database course log 2020
- Aulaweb Database course log 2021

Aim of the project

The project focuses on a Process and Data Mining activity on educational data, starting with Aulaweb logs from previous years' database course. The goal is to interpret student learning behavior through raw data.

Preprocessing Log Data

Using Tableau Prep Builder, I cleaned the three log data and kept the following variables:

- Time: The date they accessed it
- User: The id of the students
- Action: The action that the student has done
- Information: Some additional information of the action

I decided to focus the analyses on those actions performed by students that can affect their performance (i.e. final grade).

Actions considered relevant to the students’ performance (log 2019):

- Assignment view
- Course view
- Create subscription
- Download archive
- Forum view discussion
- Quiz attempt
- Quiz close attempt
- Quiz review
- Quiz view
- Quiz view summary

Actions considered relevant to the students' performance (log 2020):

- Comment view
- Course view
- Create subscription
- Forum add discussion
- Forum add post
- Forum view discussion
- Quiz attempt
- Quiz close attempt
- Quiz review
- Quiz view
- Quiz view summary
- Upload file

Actions considered relevant to the students' performance (log 2021):

- Course view
- Create subscription
- Forum add discussion
- Forum add post
- Forum view discussion
- Quiz attempt
- Quiz close attempt
- Quiz review
- Quiz view
- Quiz view summary
- Upload file

I had to further clean the 2019 data logs with a script in python, to remove students who only took the databases 2 course.

Time Analysis

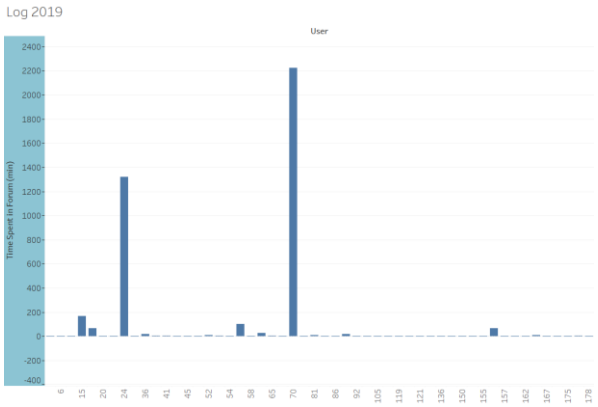
I decided to formulate queries to obtain aggregated results. I calculated the following variables related to the time spent working:

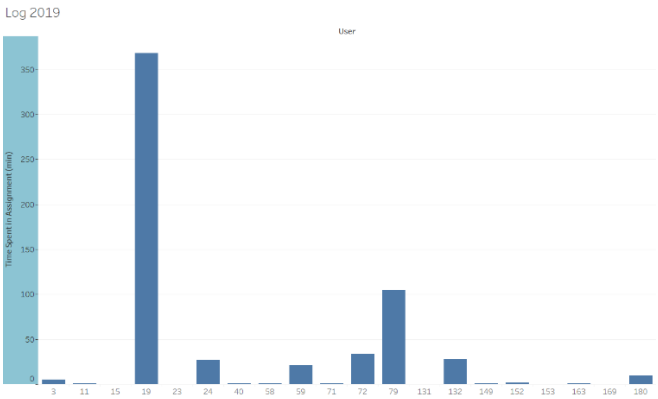
- **Time theory**: total time spent on theoretical components of the content, calculated as the sum of periods of any action related to resource
- **Time quiz**: total time spent in instructional tasks, calculated as the sum of periods of any action related to quiz
- **Time forum**: total time spent on forums, calculated as the sum of periods of any action related to forum

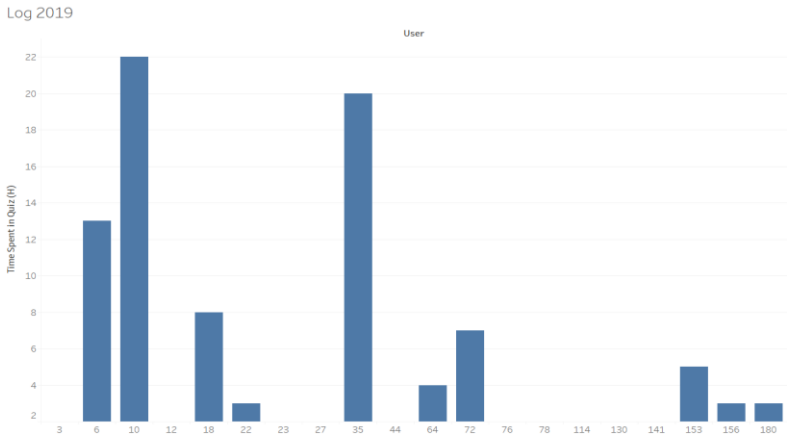
- **Time assignment**: total time spent on “hands in” activity, calculated as the sum of periods of any action related to assignment

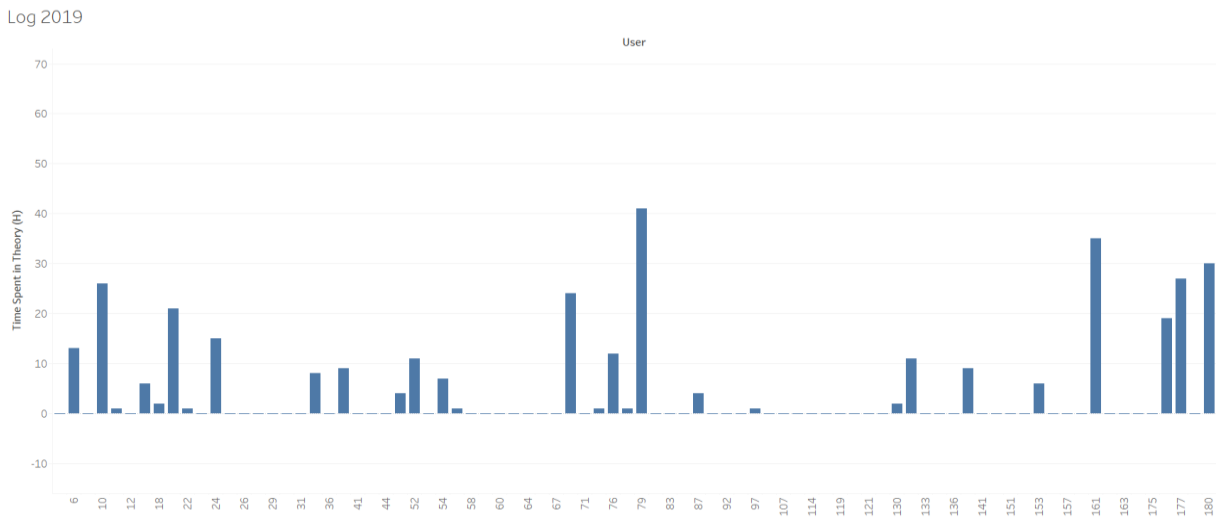
All these working times are totally theoretical, because these took place outside of teaching hours; thus, while using Aulaweb, the students could simultaneously be working or surfing the Internet, so must be considered as approximations.

Distribution of the previous variables

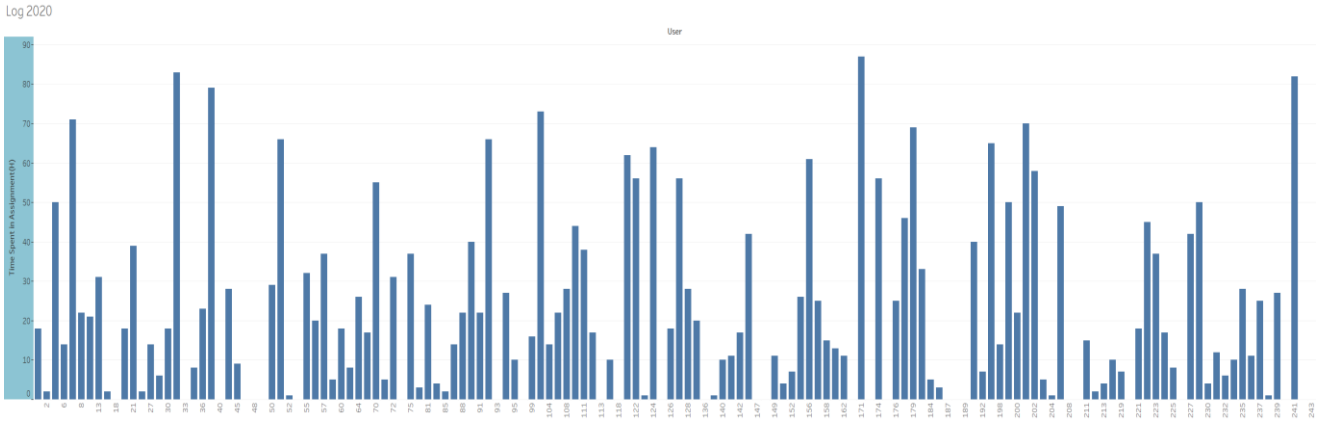


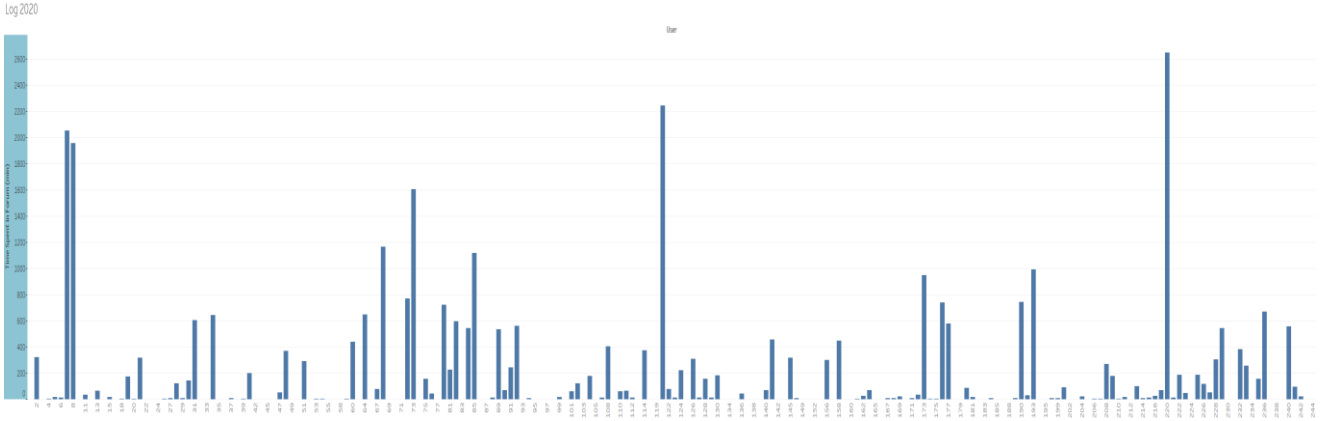


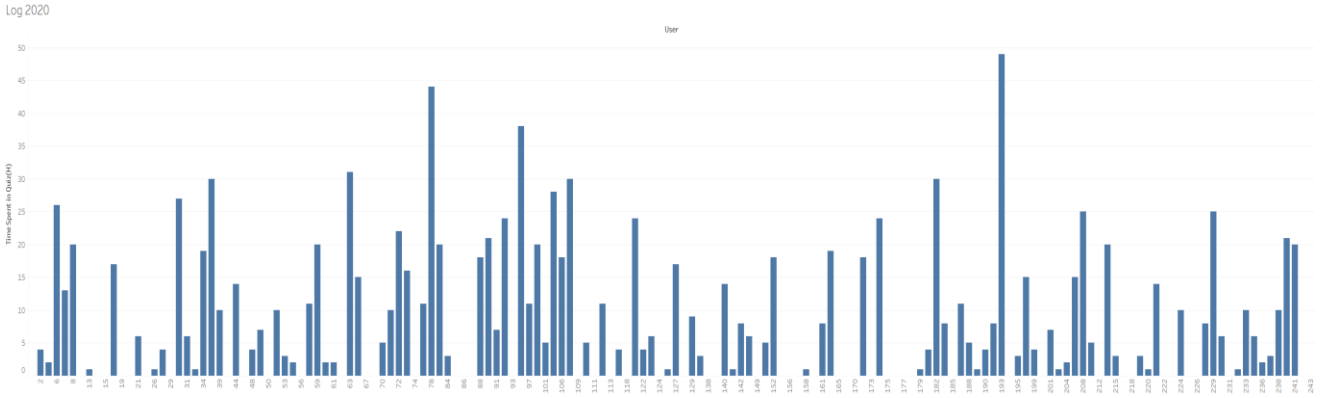


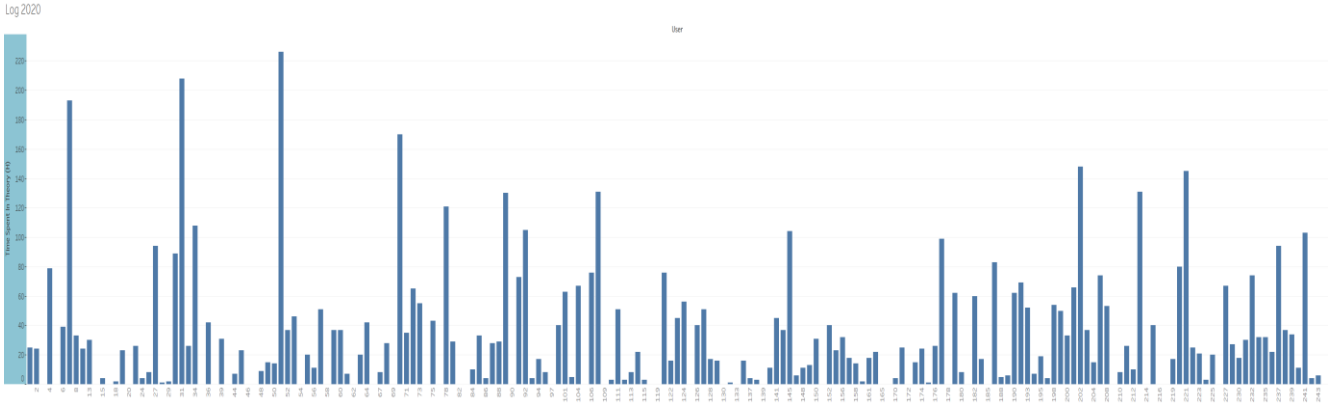


I calculated the time spent on Quiz and the time spent on Theory in hours, and the Forum and Assignment times in minutes to avoid having too many null values.

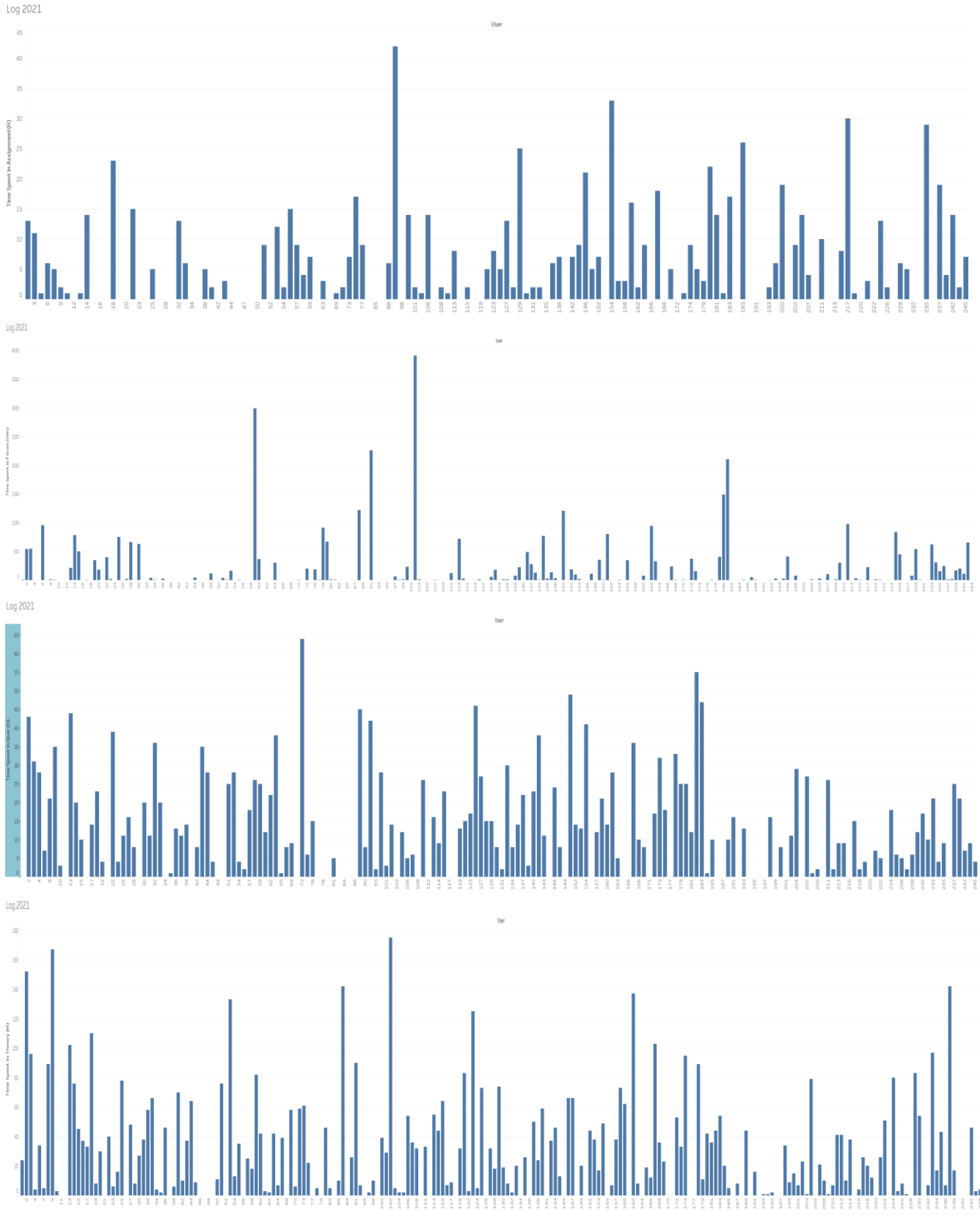








For the 2020 and 2021 logs, I calculated the time spent in Theory, Quiz, Assignment in hours, while Forum in minutes.



As you can see from the distributions, the 2020 and 2021 log files contain more data than 2019, this is due in my opinion to 2 factors:

- the extracted data period of 2019 does not cover the main course period (February-May) where most students attend
- the 2020 and 2021 logs, instructional activity was conducted online due to the pandemic

Clustering Approach for Grouping Log Data

I have chosen to apply two different approaches. A first approach uses all event log data to disclose a process model of a student's behavior. A second approach applies clustering first in order to group students with similar marks or characteristics, then, it implements process mining to discover more specific models of the student's behavior. Since I did not have the students' final grades available, I grouped them using a clustering algorithm based on their interactions with the Aulaweb course. To do this I used Weka, which is a data mining tool applied for the study's clustering. Weka is a collection of machine learning algorithms for data mining tasks. The Weka system has several clustering algorithms, I used the expectation-maximization (EM) clustering algorithm. This algorithm is used in statistics to find maximum likelihood estimators of parameters in probabilistic models that rely on unobservable variables. I have selected this specific algorithm because it is well-know clustering algorithm that does not require the user to specify the number of clusters. My objective is to group together students who have similar characteristics when using Aulaweb.

Values (mean ± std.dev.) of the centroids of each cluster

Log Data 2019

Attribute	Cluster 0	Cluster 1
Time Theory	11.79±12.27	0.83±2.40
Time Quiz	3.66±5.94	0.008±0.092
Time Forum	155.85±489.35	1.85±4.13
Time Assignment	23.42±72.90	0.14±0.39

I obtained two clusters with the following distribution of students:

- Cluster 0: 26 students
- Cluster 1: 56 students

Clustering algorithms provide a highly interpretable result model by means of the values of each cluster centroid. The centroid represents the most typical student in a cluster and it does not necessarily describe any give case in that cluster.

Log Data 2020

Attribute	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
Time Theory	84.38± 37.06	164.9± 51.85	26.96± 21.25	44.87± 23.62	19.55± 10.97	1.63± 2.71	51.55± 20.33
Time Quiz	19.68± 12.66	10.85± 7.73	8.44± 9.44	7.74± 6.76	1.36± 1.54	0±9.41	9.97± 8.60
Time Forum	425.89± 272.53	852.22± 909.87	4.11± 5.89	90.22± 94.16	358.66± 312.09	21.53± 46.27	1584.81± 800.53
Time Assignment	28.93± 20.26	66.56± 9.59	11.60± 10.41	47.77± 21.16	10.51± 10.25	0±0	7.99±9.57

I obtained six clusters with the following distribution of students:

- Cluster 0: 6 students
- Cluster 1: 53 students
- Cluster 2: 28 students
- Cluster 3: 22 students
- Cluster 4: 67 students
- Cluster 5: 3 students
- Cluster 6: 18 students

Log Data 2021

Attribute	Cluster 0	Cluster 1	Cluster 2
Time Theory	34.45±22.96	4.00±5.88	84.03±41.45
Time Quiz	13.79±9.48	0.58±1.70	29.20±14.28
Time Forum	125.98±227.86	43.24±98.37	687.21±902.43
Time Assignment	4.52±5.71	0±0.0001	11.81±10.35

I obtained three clusters with the following distribution of students:

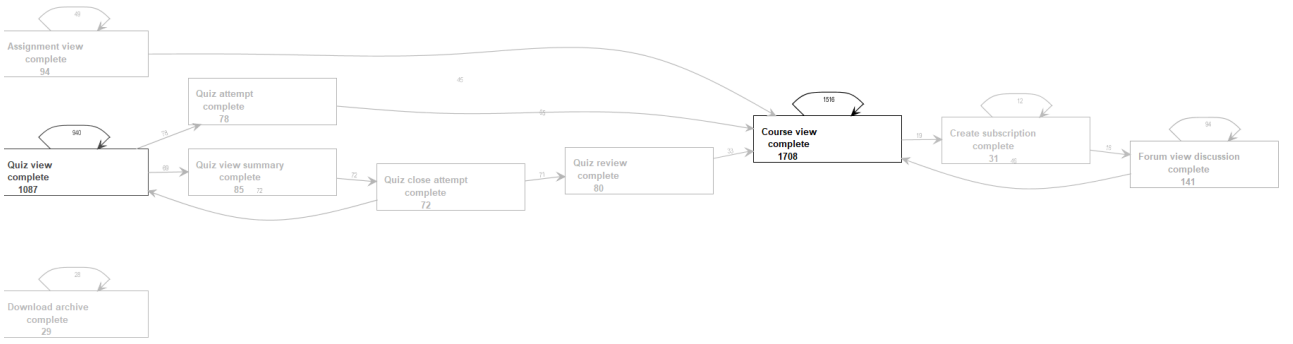
- Cluster 0: 86 students
- Cluster 1: 99 students
- Cluster 2: 37 students

Discovered models using ProM

Among the wide variety of algorithms, I applied the robust algorithm Heuristic Miner to investigate the processes in the users' behavior. In this context, Heuristic Miner can be used to express the main behavior registered in an event log. It focuses on the control-flow perspective and generates a process model in the form of a Heuristics Net for the given event log. Therefore, the Heuristic Miner algorithm was designed to make use of a frequency-based metric that is less sensitive to noise and the incompleteness of the logs. As quality measures, I used fitness and the default threshold parameters of the Heuristic Miner algorithm. I applied Heuristic Miner using the ProM tool over the previously obtained log data sets in order to discover students' process models and workflows. I applied the algorithm to each of these logs:

- Log data 2019
- Cluster 0-1 2019
- Log data 2020
- Cluster 1-2-3-4 2020
- Log data 2021
- Cluster 0-1-2 2021

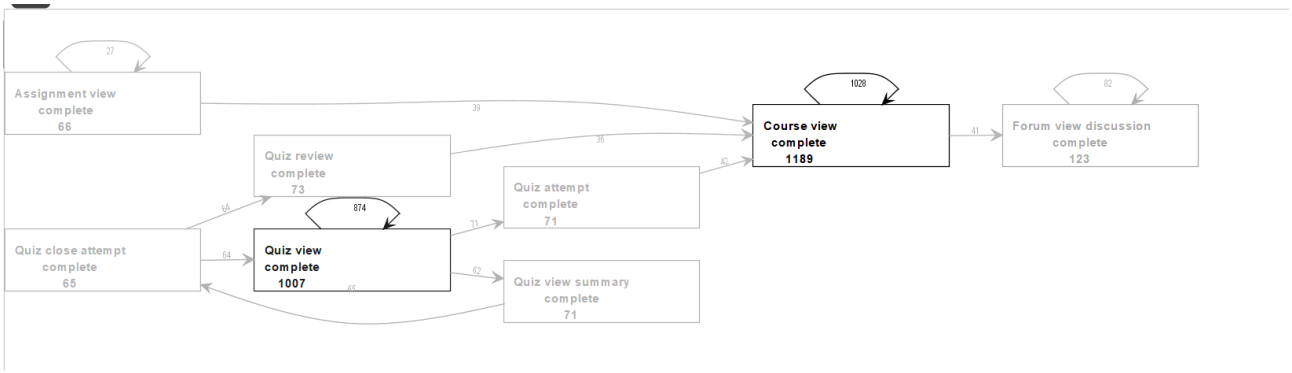
Log data 2019



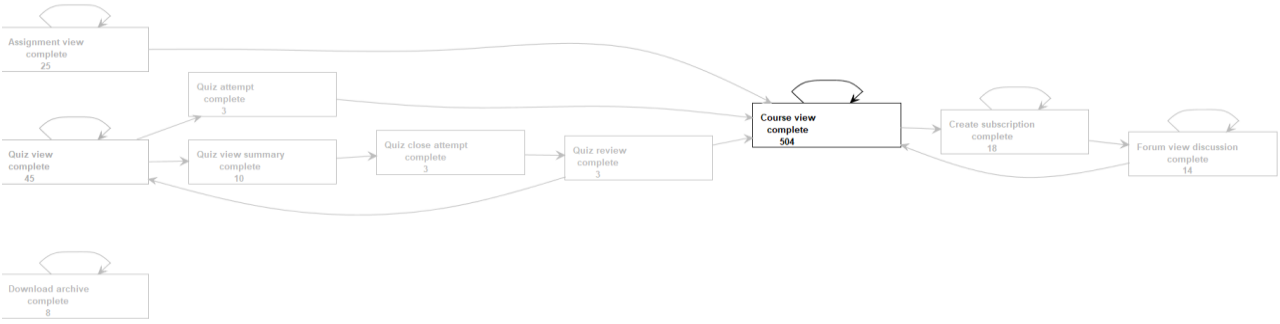
The model discovered by the Heuristics Miner algorithm is a heuristic network that is a cyclic, directed graph representing the most common behaviors of students browsing the course. In this graph, the square boxes represent the actions of the students when interacting with Aulaweb's interface, and the arcs/links represent dependences/relations between actions.

The previous figure shows the heuristic network obtained when using the log file with all students. We can see that there are two subnets that most of the students follow in the course. The first one concerns actions regarding quiz, course and forum. The separate one at the bottom concerns the download of the archive containing the slides. It is important to note that these networks show the general behavior of all the students (fail and pass students mixed), soI cannot draw conclusions with the information obtained

Cluster 0 data 2019



Cluster 1 data 2019

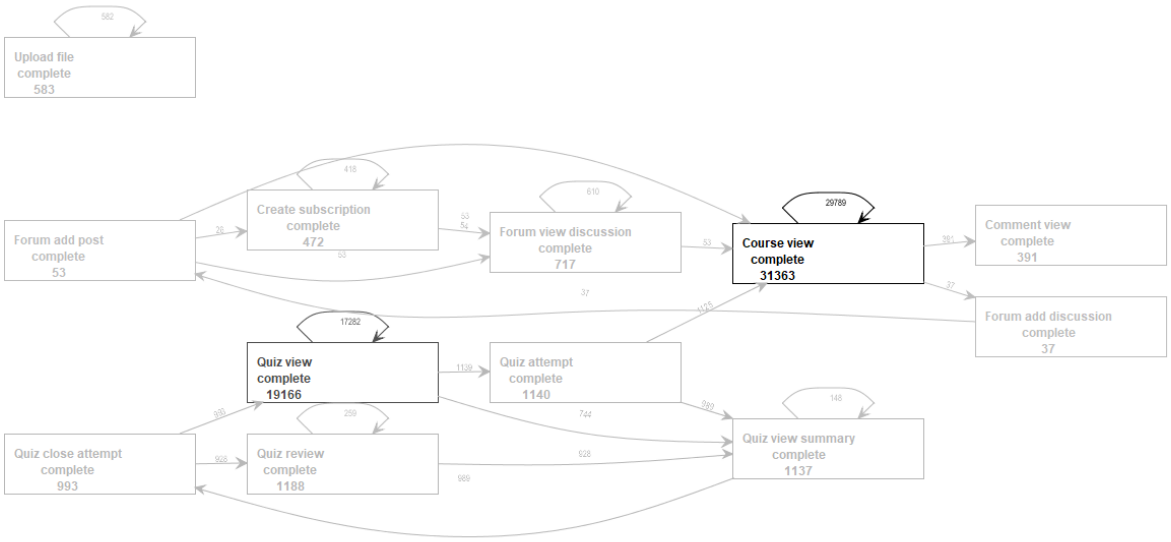


With the separation of students into the two clusters, seeing the models, I can speculate that it is quite likely that most of the students belonging to cluster1 passed the exam by seeing a good activity in quiz and forum actions.

While the students who belong to cluster2, most likely did not pass the exam having as major activity course view actions.

It is also interesting to see how the heuristic net of students who probably failed is much smaller than those of the heuristic net for all students and for probably passing students.

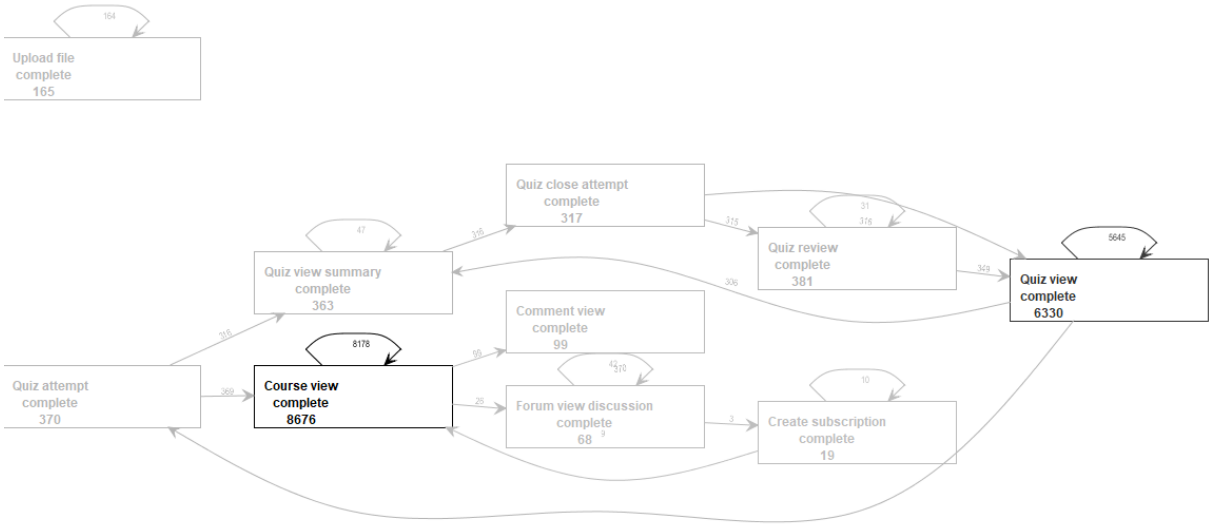
Log data 2020



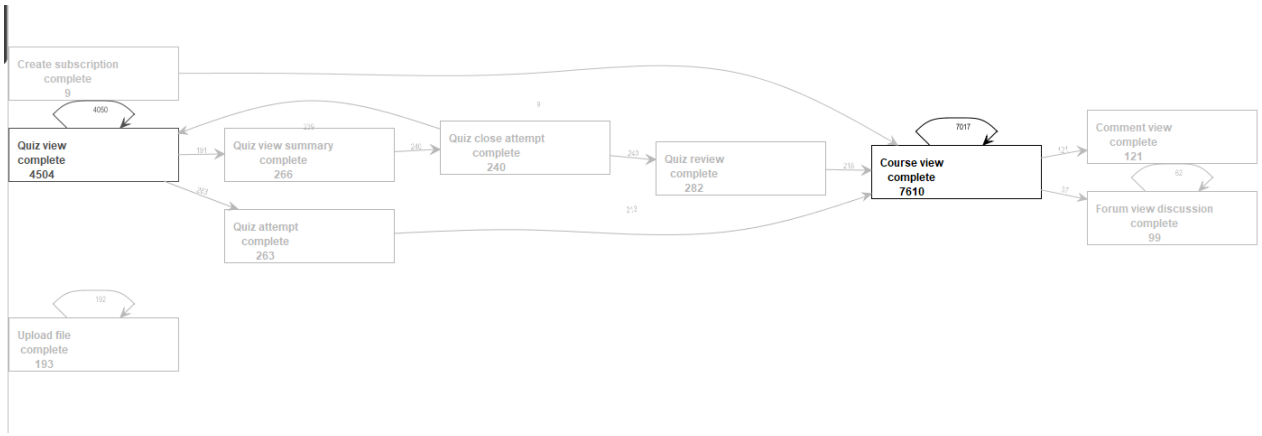
The heuristic net of all students of year 2020 is higher than 2019. This is because the 2020 log contains more data than the 2019 log.

Using Weka's EM algorithm I obtained 6 clusters (not specifying the number of clusters), since some have a very low distribution of students I decided to show only the four clusters with more students.

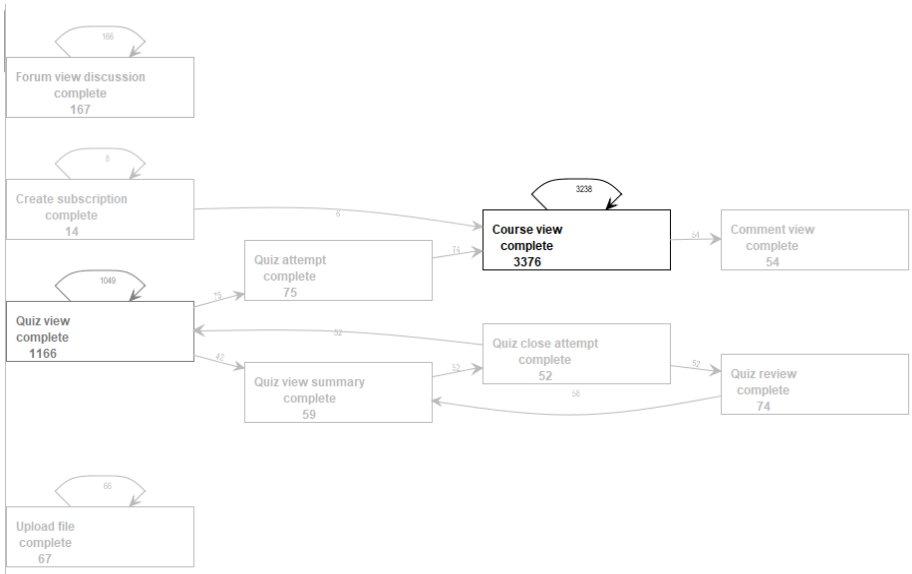
Cluster 1 data 2020



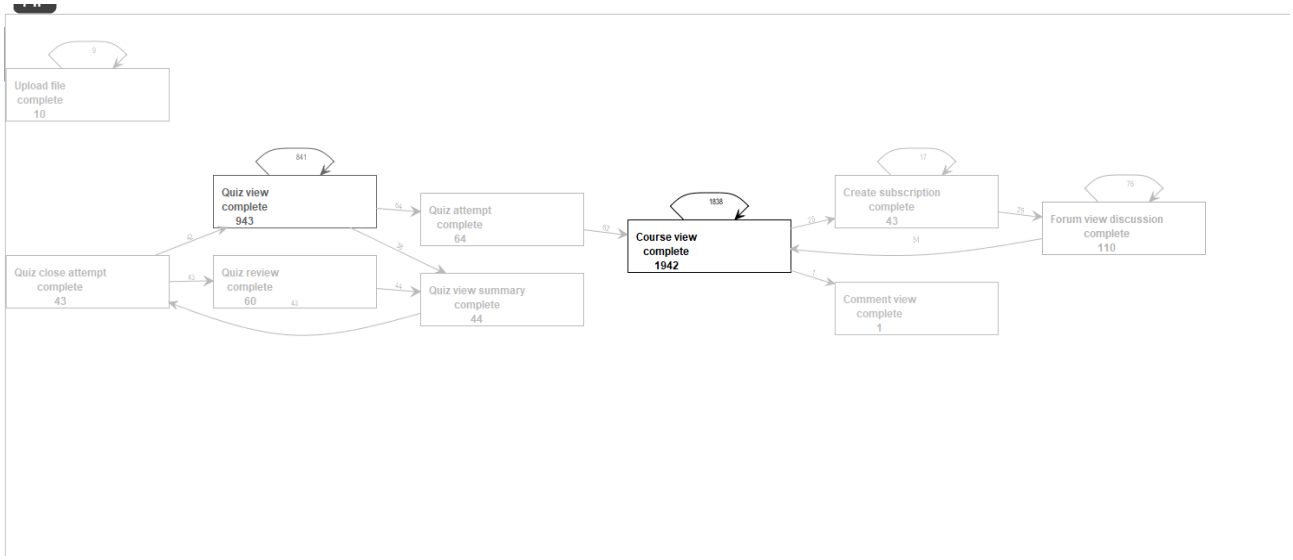
Cluster 2 data 2020



Cluster 3 data 2020

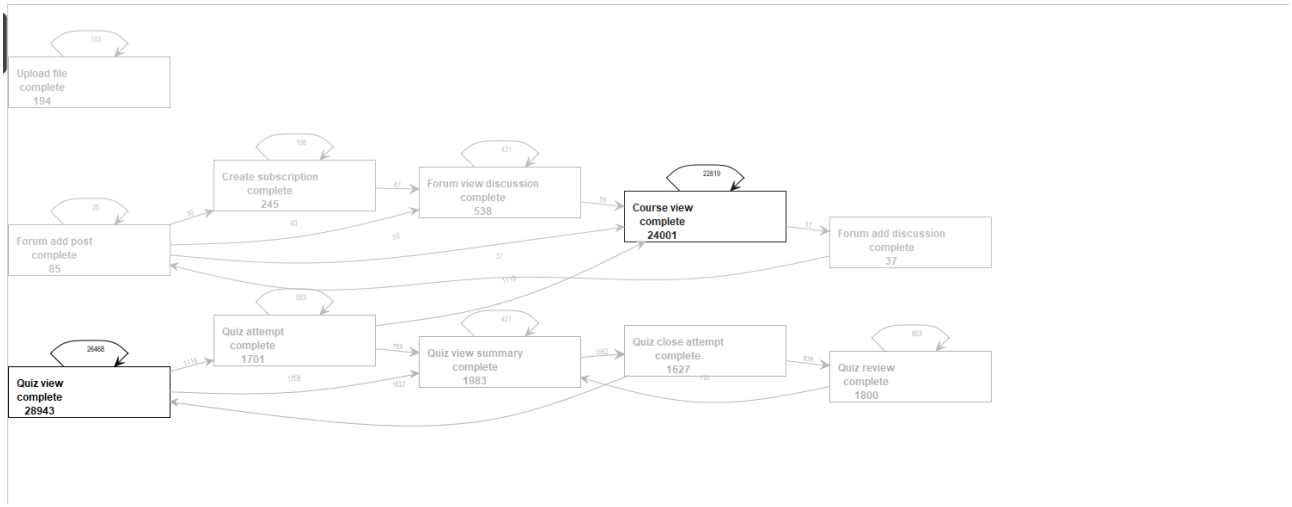


Cluster 4 data 2020

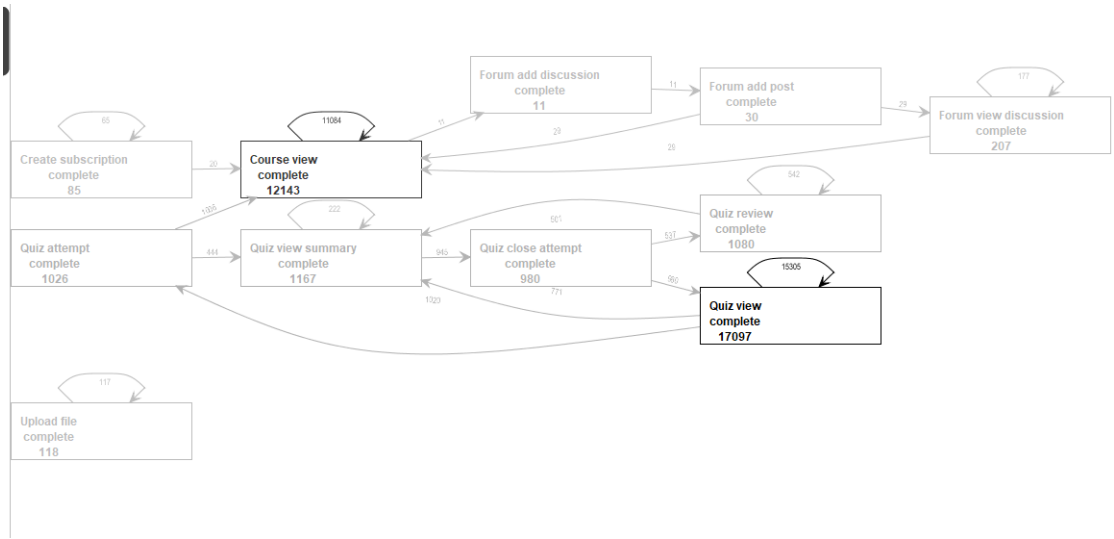


Seeing the great activity in the quizzes, in the assignments (upload files), in the forum of the students belonging to cluster 1 and cluster 2, I can assume that they have passed the course even with high grades, compared to the students belonging to cluster 3 and 4 who perhaps passed the course but with lower grades.

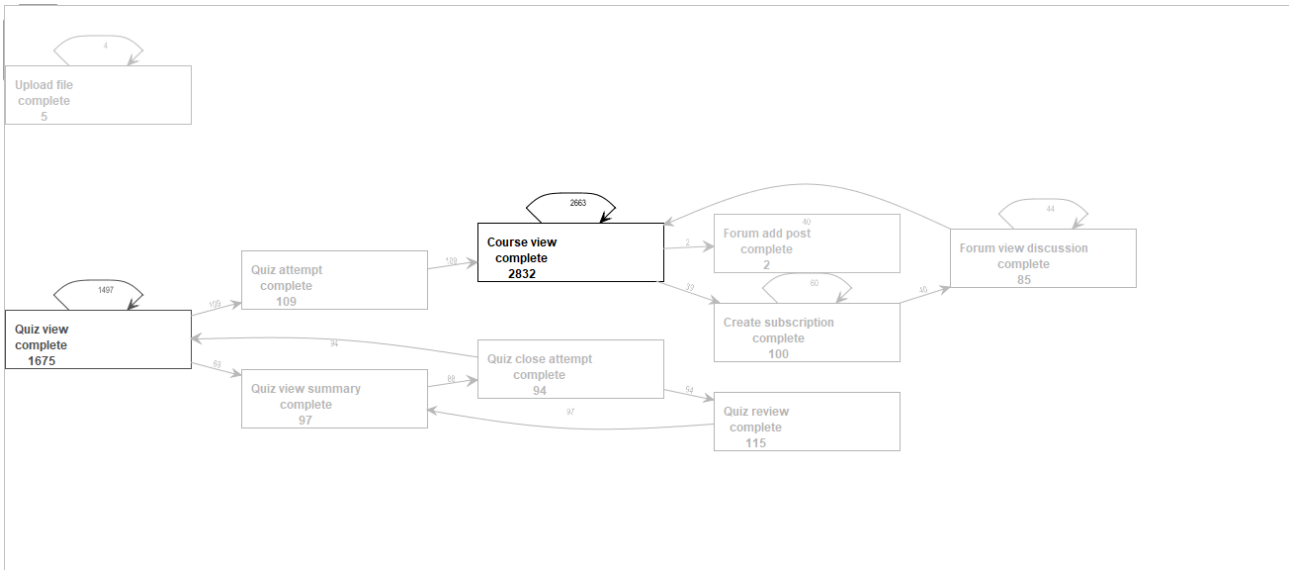
Log data 2021



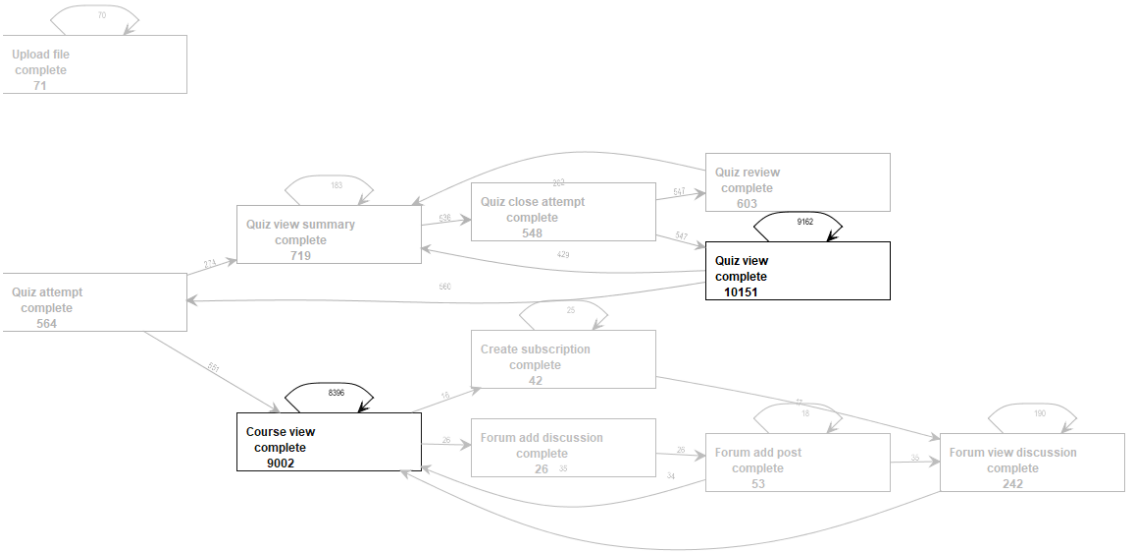
Cluster 0 data 2021



Cluster 1 data 2021



Cluster 2 data 2021



Regarding these last logs of 2021, I find it difficult to hazard hypothesis because there are a lot of relations/dependences between the actions that make the models harder to interpret. Seeing cluster 1 smaller than the others I could say that perhaps most of the students who did not pass the course belong to it.