بسم الله الرحمن الرحيم

Data Analysis Report

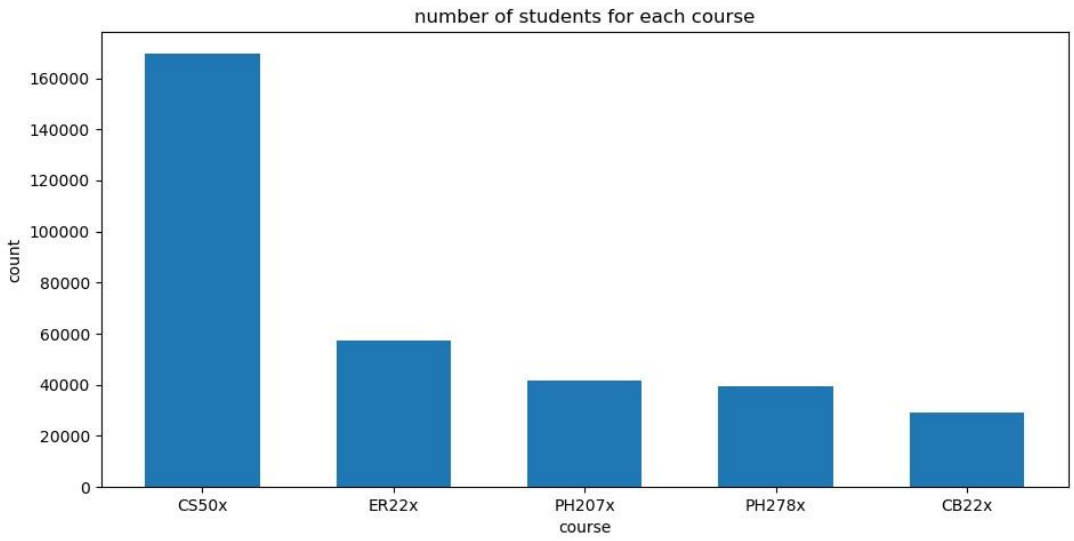
The data provided by Harvard indicates that the courses and students who registered can be analyzed as coming in the following charts and conclusions.

Before the analyses, we need to demonstrate that the courses in the data sets have ids different than their names, so here is the mapping:

Course Id	Course Full Title	Category
HarvardX/PH207x/2012_Fall	Health in Numbers: Quantitative Methods in Clinical & Public Health Research	Public Health
HarvardX/PH278x/2013_Spring	Human Health and Global Environmental Change	Public Health
HarvardX/CS50x/2012	Introduction to Computer Science I	Computer Science
HarvardX/CB22x/2013_Spring	The Ancient Greek Hero	Classical Studies
HarvardX/ER22x/2013_Spring	Justice	Ethics and Moral Reasoning

Now, let us see the data patterns and what they are telling us:

1. **Course Popularity**



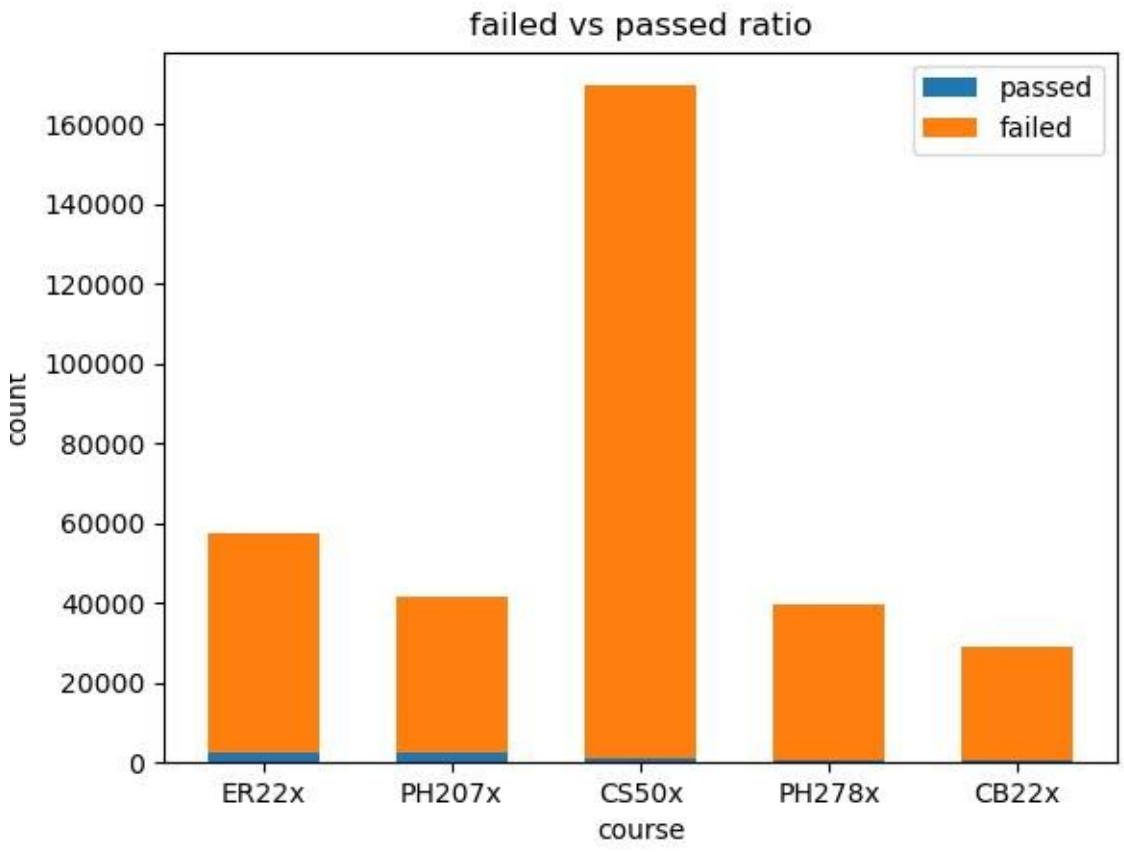
This chart demonstrates the order of courses from the most registered in, to the least. The graph is skewed and indicates that the **most popular** course is the CS50x with more than 160,000 students.

In addition, the **second most popular** course is PH207x, with almost 60,000 students.

It is worth mentioning that the courses PH278x and PH278x, which belong to the **same category**, have almost the **same popularity** in terms of registered students.

Finally, it is notable that students seem to be not less interested in classical studies, since the registered students in the CB22x course are around 30,000, showing the least number of registrations.

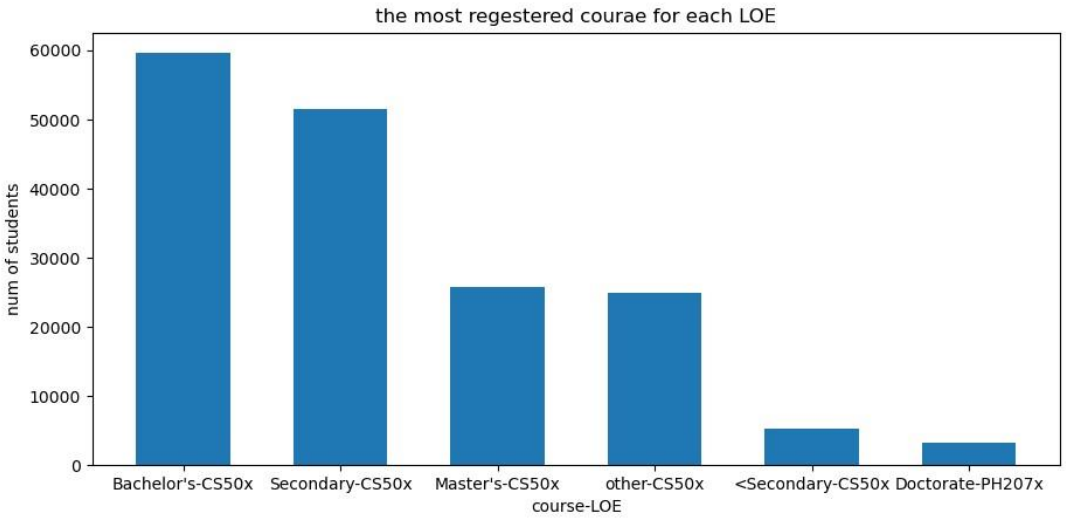
2. **Passed and Failed Ratios:**



The graph shows clearly that the students who **passed** form a **very small ratio**, which approaches zero in some courses. While the students who **failed** form the **vast majority** of the students. This could be an indicator that the courses are either very hard or are specified for professionals.

Also, this can be a sign that these courses are very effective for students who passed.

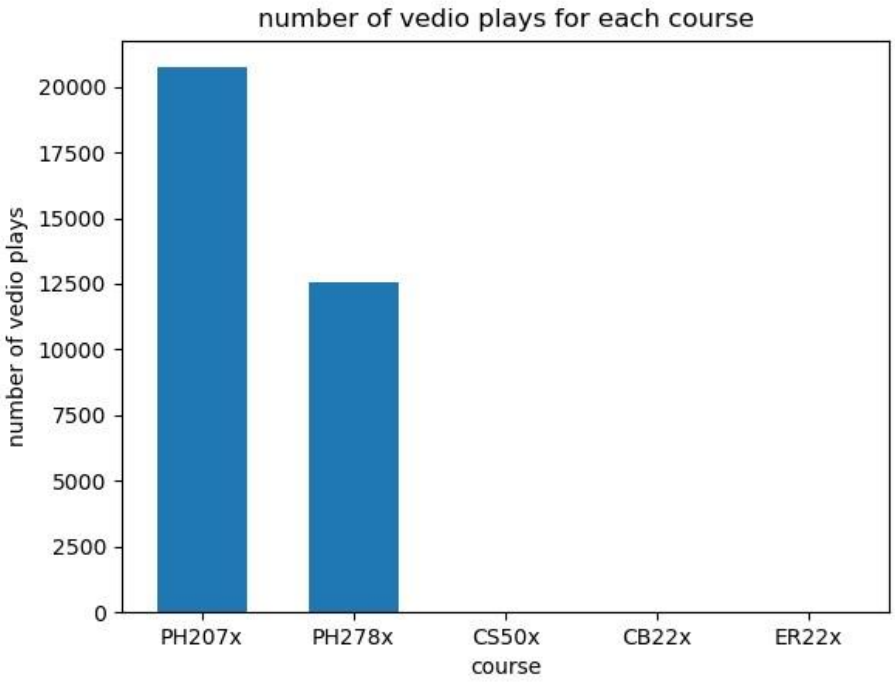
3. **Most Popular Course Amongst Each Level of Education (LOE):**



This chart has clear information which is that the **most popular course among all levels** of education - **except the Doctorate** level- is the **CS50x** course. As for **Doctorate**, the **most popular is PH207x**.

This can be an indicator that **most students** of different levels of education **have an interest** in the **Computer Science track**. This information can be a sign that this course has **great weight** in the **recommendation** formula.

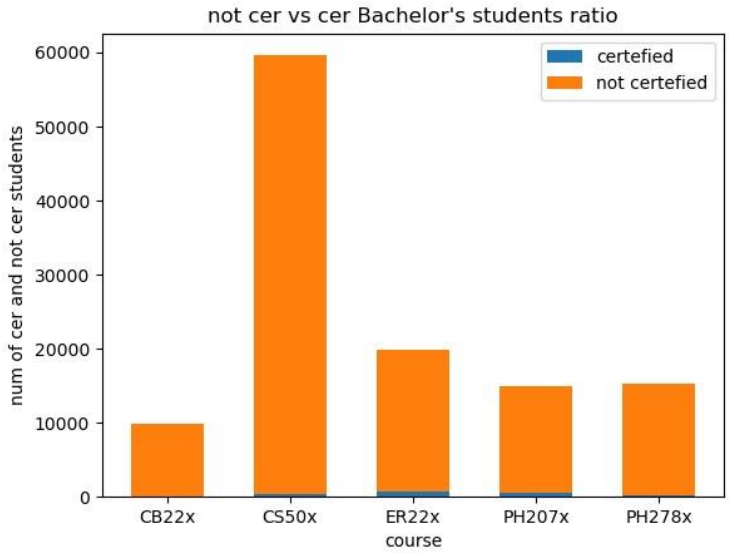
4. **Video Plays Count Per Course:**



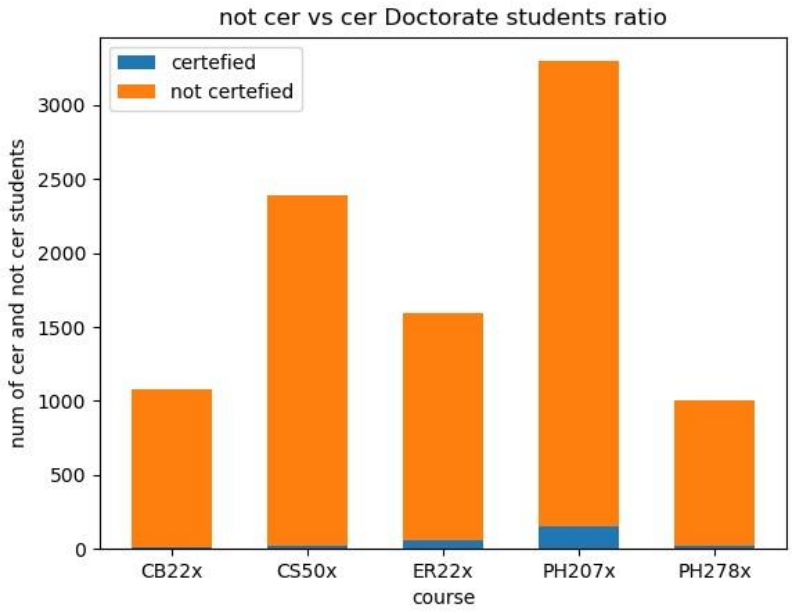
The previous chart shows that the summation of video plays among all students for the Public Health track are clearly higher than the other fields.

It is also worth noting that the other courses literally have ZERO video plays. This can be an indicator that these courses are preferred to be studied theoretically over through videos.

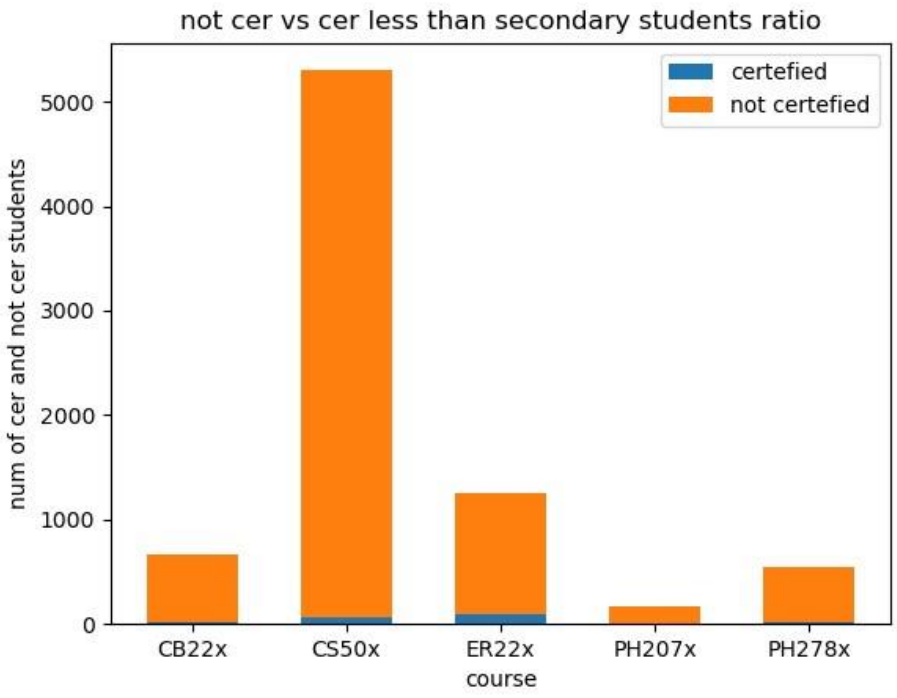
5. **Certified VS Not Certified Bachelor's Students among all courses:**



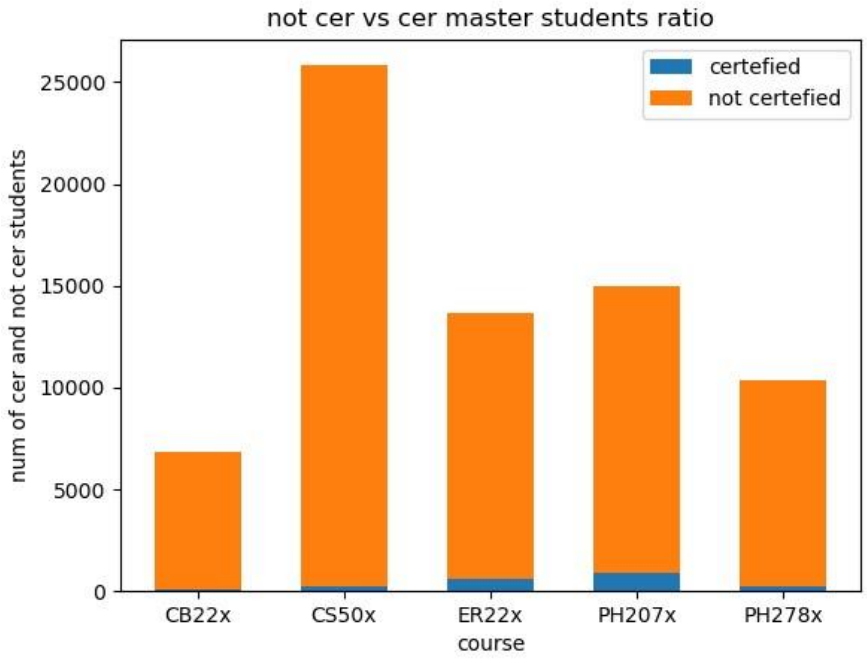
6. **Certified VS Not Certified Doctorate Students among all courses:**



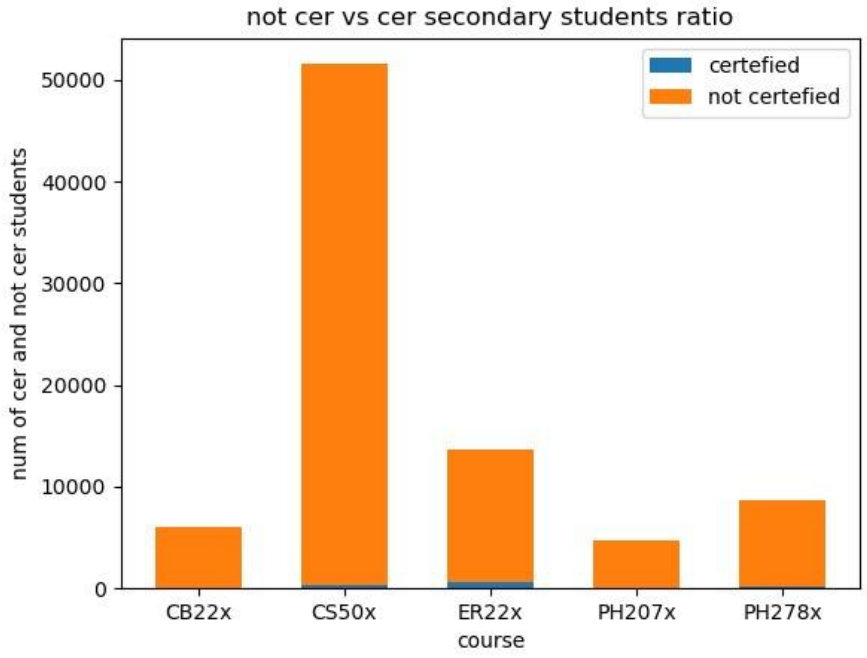
7. Certified VS Not Certified Less Than Secondary Students among all courses:



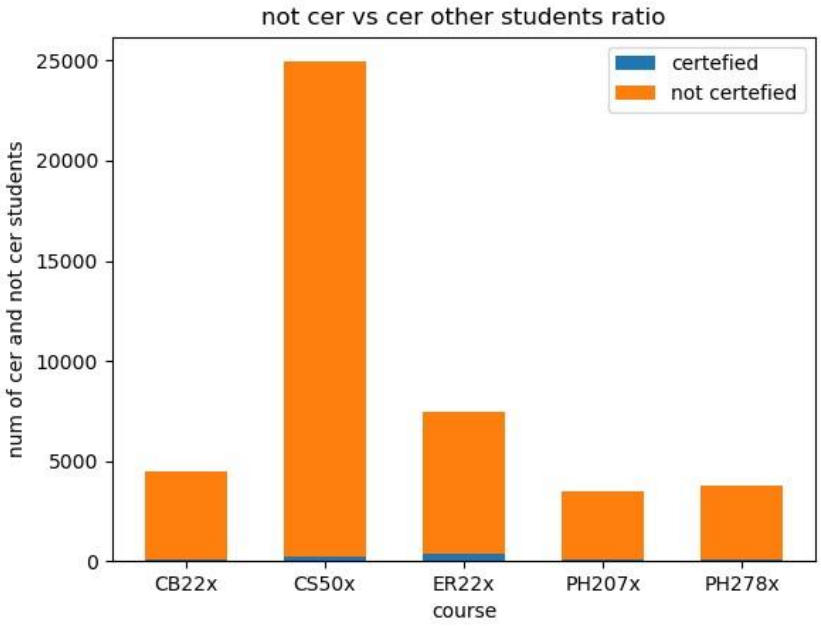
8. Certified VS Not Certified Master Students among all courses:



9. Certified VS Not Certified Secondary Students among all courses:



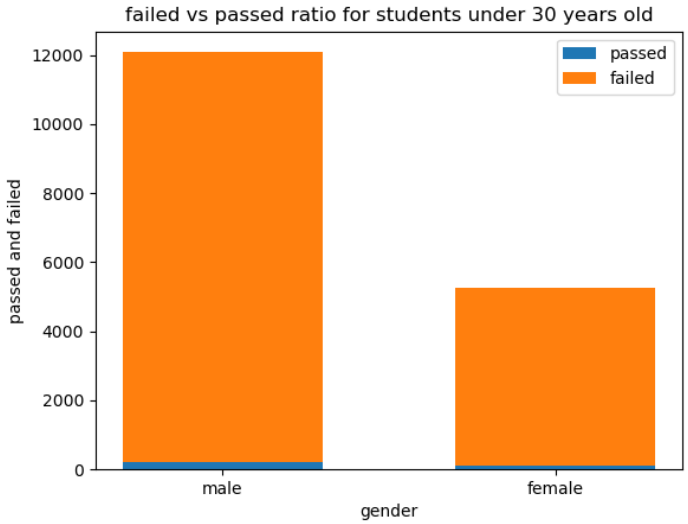
10. Certified VS Not Certified Students From Other Levels among all courses:

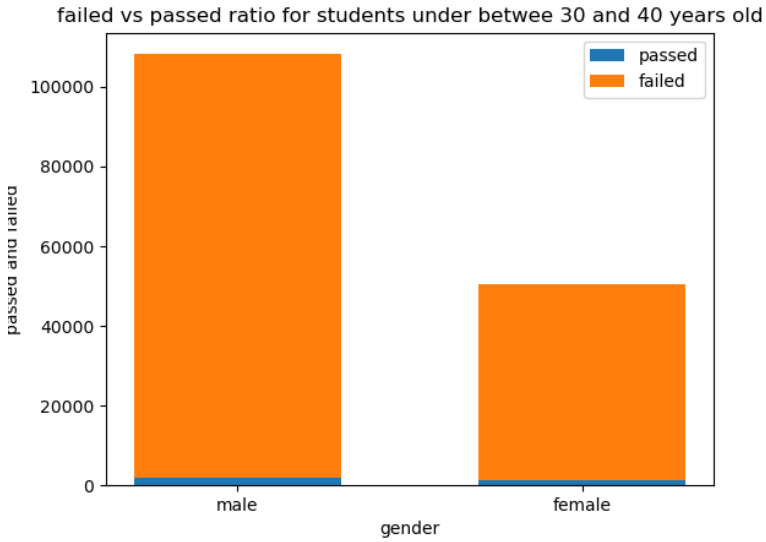


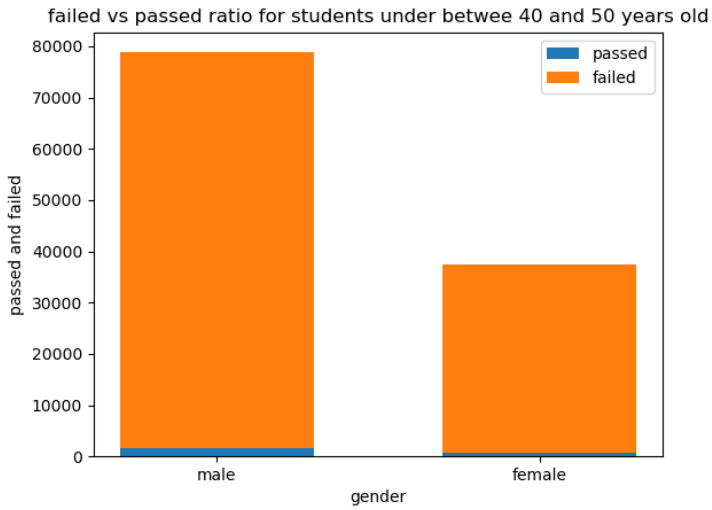
Charts from 5-10 show that the **certified** students from all levels form **a very small ratio** compared to those who are not certified. This can be notable for all courses.

This can be a strong sign that students who tend to register in courses must have a strong foundation in the associated topic. In addition, they must study very well to be able to achieve the maximum usefulness of the course, and to increase their chance to pass and get certified.

11. Certified vs Not certified males and females student from every age stage:

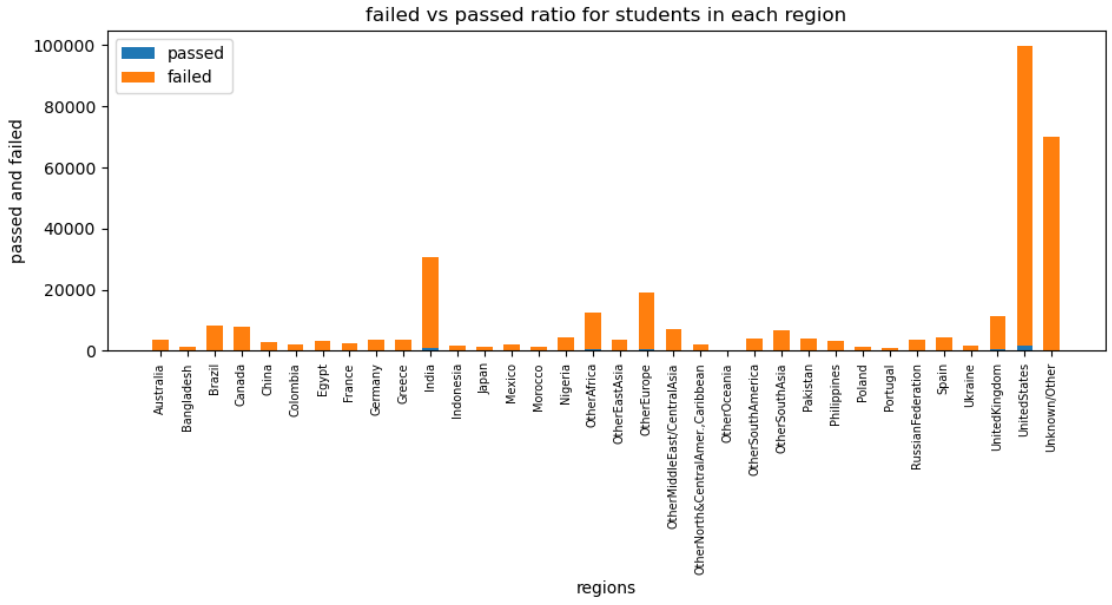

failed vs passed ratio for students under 30 years old
passed
failed
12000
10000
8000
6000
4000
2000
0
passed and failed
male
female
gender


failed vs passed ratio for students under betwee 30 and 40 years old
passed
failed
100000
80000
60000
40000
20000
0
male
female
gender



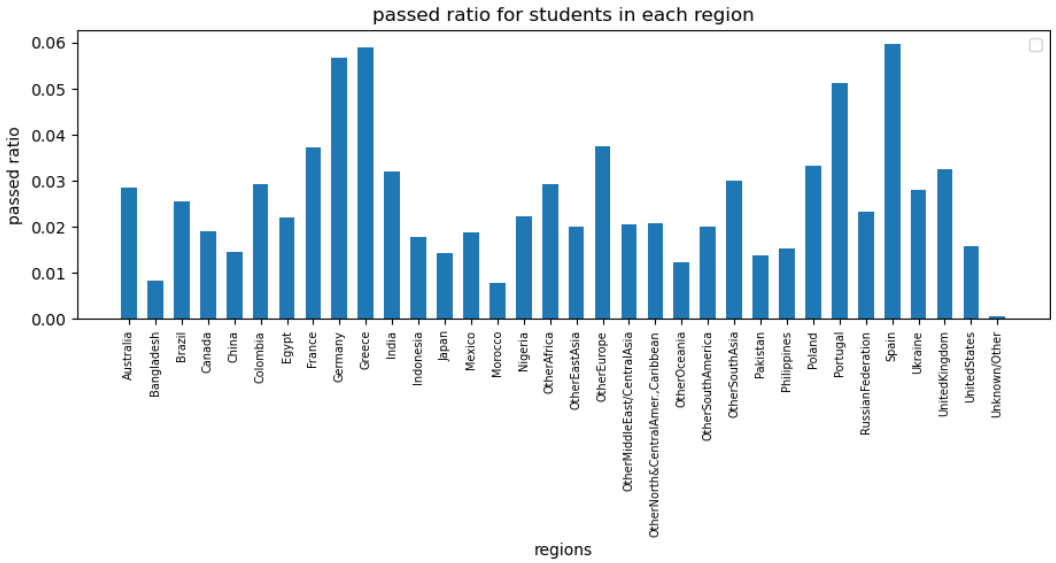
The charts above show that the number of male students is larger than the female students but the ratio of passing is almost equal which leads us to expect that the gender will not effect on the recommendation system.

12. the number of passed vs failed students from each region:



We can see that the widest audience from the US, followed by countries not on the list. This maybe means that people don't prefer to share personal information such as this about themselves.

13. The passing ratio in each country:



The regions with the highest passing ratios are Germany, Greece, Portugal, Spain, the countries with the lowest passing ratios are Bangladesh, Morocco, Pakistan. This leads us to conclude that the weakness or strength of the Internet infrastructure affects the process of learning.