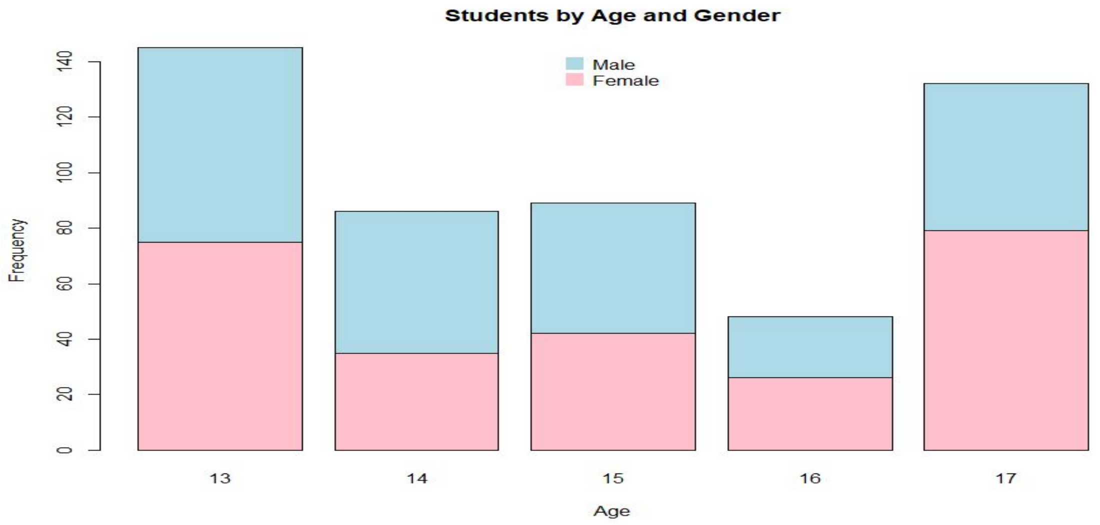
Project Goal

This executive summary outlines the current problematic trends with the *'Data Science for High School'* course and attempts to propose a solution that will remediate the issues and improve the quality of education delivered to the students. In this context, it aims to further the fourth goal laid out by the United Nations for sustainable development: "*Ensure inclusive and equitable education and promote lifelong learning opportunities for all.*"

Description of the sample data set

The given data set pertains to the historical data of the student performance in the Data Science course. The sample size is 500 students of which the ratio of male to female gender is 243:247 across the age groups of 13, 14, 15, 16 and 17 years.

The visualization of the data has been done using inherent R capabilities, imported libraries, and techniques such as sentiment analysis (with Twitter and AFINN Lists). The data set has been provided by Virtual High School, the administrator of the course.



Graph 1: Bar plot indicating ratio of registered students by age and gender

Observation: Ratio of male to female students registered is proportionate across all age groups

Age group (yrs)	Male Students	Female Students
13	70	75
14	51	35
15	47	42
16	22	26
17	53	79

Table 1: Number of male and female students registered for the course across five age groups

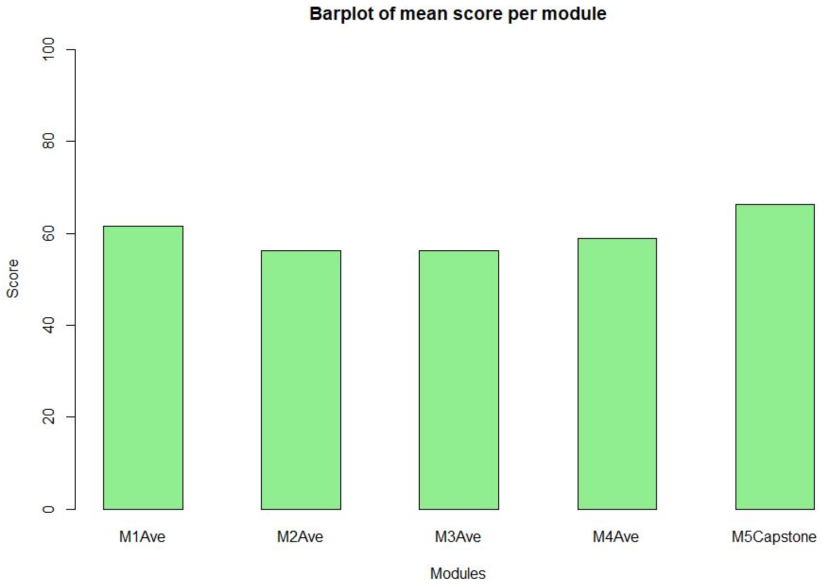
The given data set consists of the following details pertaining to each student:

- The score obtained for each assignment
- The average score of assignments in each module
- The final score that the student received in the course
- The overall rating of the course provided by the student, categorized as 'Bad', 'Ok','Good','Great', and 'Excellent'.

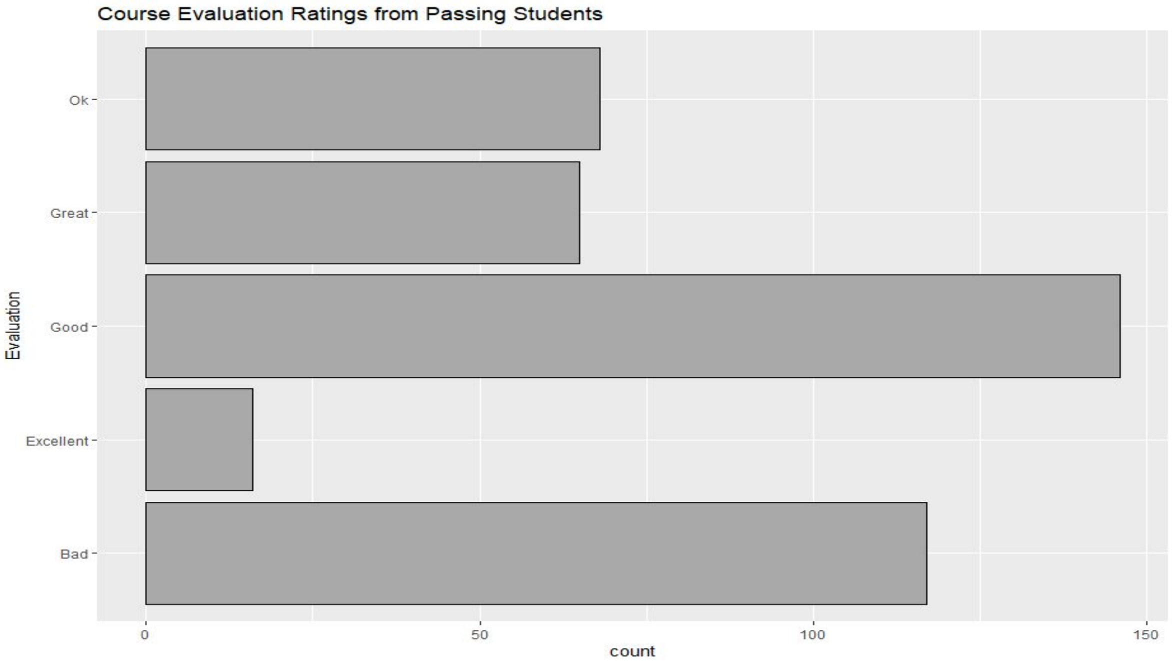
The following data was extracted to further analyze the factors contributing to the performance of the students across the different stages of the course progress.

- The mean score of all students per module
- The mean score per assignment for students that did not pass the course

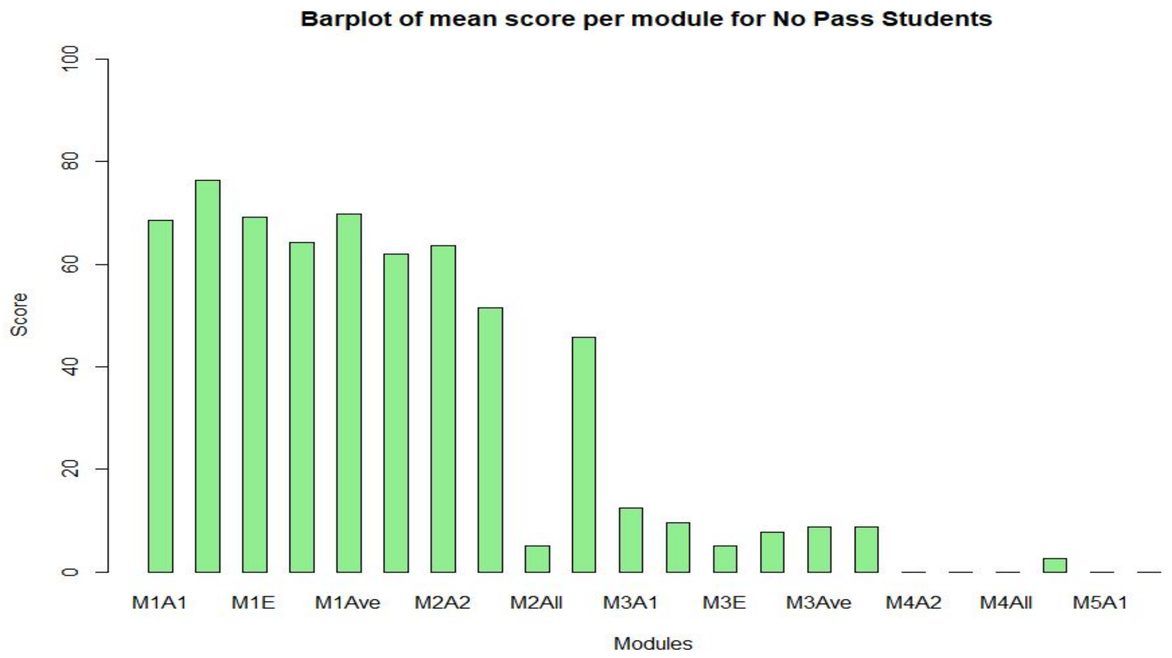
Graphs that are indicative of the above data have been provided for visualization purposes.



Graph 2: Bar plot of the average score of assignments that students received per module



Graph 3: Bar plot of student course evaluations for passing students.

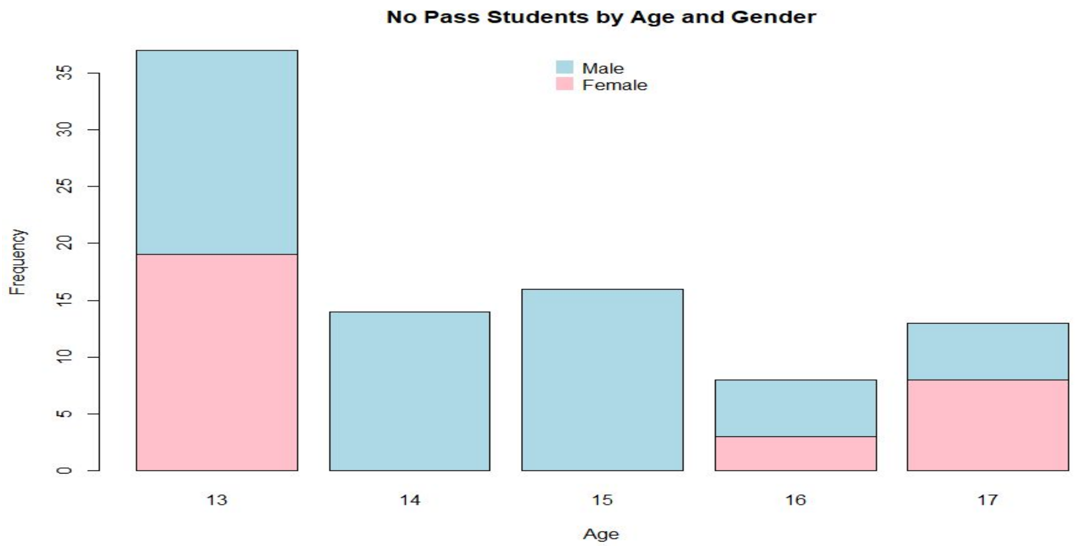


Graph 4: Bar plot of the average score per assignment that failing students received

The initial analysis of the course data has revealed the following issues.

- The qualitative and quantitative data provided indicates that the students are neither doing academically well nor feeling positive about the course. The ratings provided by the students for the course experience reflects their general displeasement with how the course went for them.
- Sentiment analysis performed using data from Twitter revealed that the general sentiment on the topics covered in the course and the field of Data Science is positive. The source of discontent for the students is not coming from the content of the course itself, but rather how the content is structured and delivered, making it seem very difficult to learn. Students are either rating the course overwhelmingly negative or are indifferent about the evaluation they provide.
- Modules (especially module 2) seem to be far too difficult as shown by the barely passing averages (median score of 57%). Only 6% of students that enroll are able to pass the class with high distinction (above 85%).

- Students' failure to complete the assignments is a bigger contributing factor to their failure in the course rather than completing the assignments and getting bad scores (below 50%). This could be driven by lack of interest.

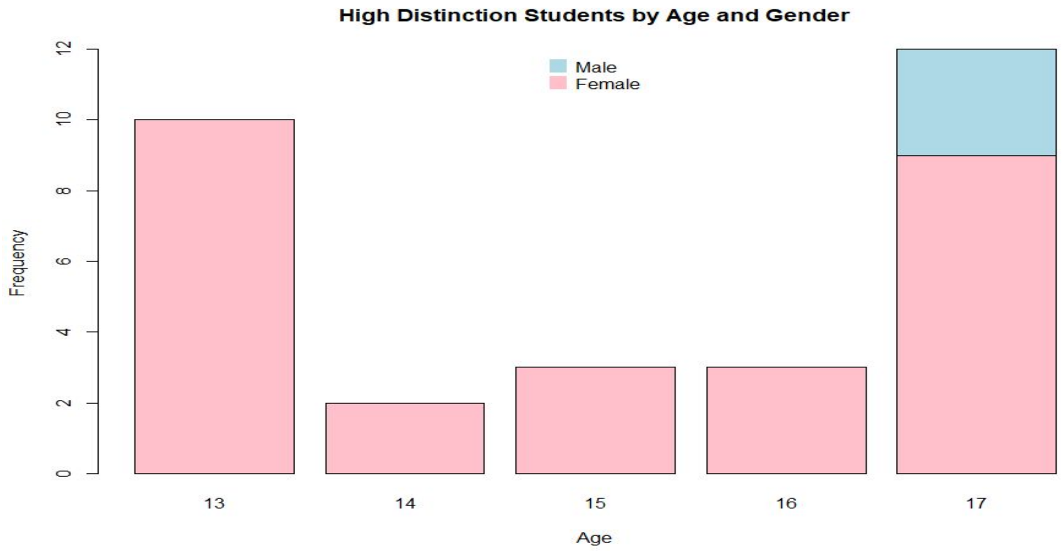


Graph 5: Barplot of students that fail the course by age and gender

Observation: Fail/Drop rate is higher among younger students, specifically male students in age groups 14,15 & 16 years

- The course drop rate among 13 year olds is significantly large. Although the percentage dropped across all age groups, except 17 year olds, is comparatively similar, this similarity could be coincidental because the population size of 14, 15 and 16 year olds is not very large. This is further compounded by the fact that the only other age group with a similar size is 17 year olds, and this group has a much lower drop rate. All of this suggests that there is the possibility of a lurking variable that specifically affects the retention rate among 13 year olds.
- As indicated in Graph 6 below, female students in all age groups outperform their male counterparts.

- The sample data set is not indicative of any data points to support the causation for the lower number of males receiving high distinction as well as a higher fail/drop rate of males in comparison to their female counterparts registering for the course.



Graph 6: Barplot of students receiving High Distinction by age and gender

Observation: Female Students outperform male students

Problem Root Cause Analysis

In order to perform a root cause analysis, the contents of the course have been evaluated from a student's perspective and the following are deemed as possible reasons for the disparity in the performance of the students.

- The course is mostly self-taught and skims through surface level topics. The quality of a student's learning seems to be dependent on how much one can utilize outside resources rather than being able to rely on the course itself.
- There are comments that explain the expected end result of certain blocks of code, but there is no clear focus on understanding the syntax or what each line of code is achieving. When it comes to writing their own code, students are required to consult outside resources to learn how to write code on their own.

- Additionally, the instructions for the unit assignments are vague at times and require a lot of assumptions, and some hyperlinks in modules are dead links. In some cases, the example code is outdated and cannot be executed successfully. For international students working across timezones, having to obtain clarifications from the instructor adds to the delay and sometimes disturbs the flow of study.
- Average grades across the board are pretty low and a strong indication that the course is too difficult. With that being said, the students that drop have relatively high scores in comparison to the rest of the students prior to dropping (shown by relatively high scores in module 1). This suggests that on top of the course being difficult, interest for the course seems to be low.
- Some of the assignments do not seem to be directly related to the flow of the course and it is hard to connect their relevance to the concepts that are covered in the course. This disconnect disturbs the flow of the course and could be a reason hindering the students' progress.
- There are 2 probable causes to explain the high drop rates with 13 year olds.
    - They may not be interested in the subject as a whole and are just exploring new interests. *Addressing this issue is beyond the scope of this project.*
    - They could be interested but they might be finding the course too difficult to follow without a firm understanding of the required foundational concepts.
- It is important to note that the given data is not indicative of any possible reasons leading to the disproportion in the number of males obtaining lower scores and having high drop rates as compared to the female students. It may just be coincidental that those who happen to do bad and/or drop are male.
- Even assuming that there is a causation between the drop rate of students and their gender, it would be difficult to identify a fix for this issue without knowing anything about the background of these students and their interests, since what works in engaging and retaining some male students may push others even further away. While there could be environmental, cultural or even biological factors that are potentially causing this disproportionate rate of drop out, it would be extremely inaccurate to extrapolate data

and determine a solution for future students with such a small sample size, limited knowledge over other potential factors, and no control over the factors considered to define the population for the sample data set.

Proposed Solutions

The following are the recommendations to remediate the above stated issues that are impacting the rates of student engagement with the course and their retention until they successfully complete it.

- Introducing more interactive content such as video lectures and/or notes that clearly demonstrate the purpose of each line of code and each function/procedure.
- Students should be given coding exercises to build code from scratch, so that they can solidify their coding skills.
- Either the hyperlinks provided have to be periodically validated or alternatively, the information that is required to be accessed via such hyperlinks can be hosted on Virtual High School servers to remove the dependency on outside resources.
- Clear explanation should be provided to connect the assignments in certain units (specifically the final assignment in unit 3) to the bigger picture of the objective that the course intends to achieve.
- Optional sections in each unit that give more in-depth explanation of the concepts introduced could be added (for ex: statistical concepts). This would provide extra support to students who need it (specifically unit 2 so that those who drop early are potentially retained).

Conclusion

In summary, the following are the key highlights of the problem analysis and the proposed solution.

- The course structure in its current state is hindering student progress leading to increased drop rate and lack of student engagement. The average student is barely able to pass the course. The students are not getting a positive experience. Male students and 13 year olds perform worse.

- Students vary in their approach of learning and to some extent adopting a student-specific personalization of teaching is required. Some students can manage to succeed with a hands-off approach to teaching, while others require a more interactive approach to get an in-depth understanding of the material to achieve equal levels of success as some of their peer counterparts.
- In order to improve the student engagement and retention rate, the 'Data Science For High School' course needs interactive content and activities that strengthen coding skills to be added as part of the course offering.
- Optional sections that go into further detail on the concepts learned can be introduced to provide extra support for students who need it.
- The dead hyperlinks have to be cleaned up and the content should be hosted on Virtual High School servers. The purpose of unit assignments should be clearly explained in context of the big picture of the course objective.

The solutions listed above further the United Nations's fourth goal to promote equality of education among students of all ages by providing quality education and improving their educational experience. When students are able to learn the course material, reproduce it on their own, and achieve better proficiency, they will be more inclined to engage in the learning process. While these improvements won't make students job-ready, students will acquire a better grasp of foundational concepts in Data Science and report a more positive learning experience. This in turn would increase their likelihood to pursue higher education in the field of Data Science and make a career for themselves in the field.