Project Goal

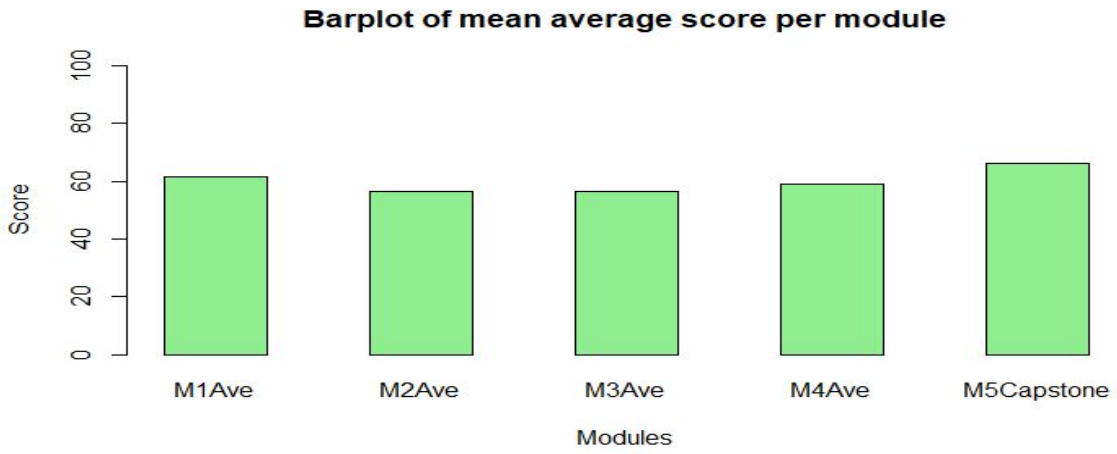
This executive summary outlines the current problematic trends with the *'Data Science for High School'* course and attempts to propose a solution that will remediate the issues and improve the quality of education delivered to the students. In this context, it aims to further the fourth goal laid out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

The given sample set pertaining to the historical data of the student performance in the Data Science course has been analyzed, and 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.

Analysis of the given sample data set

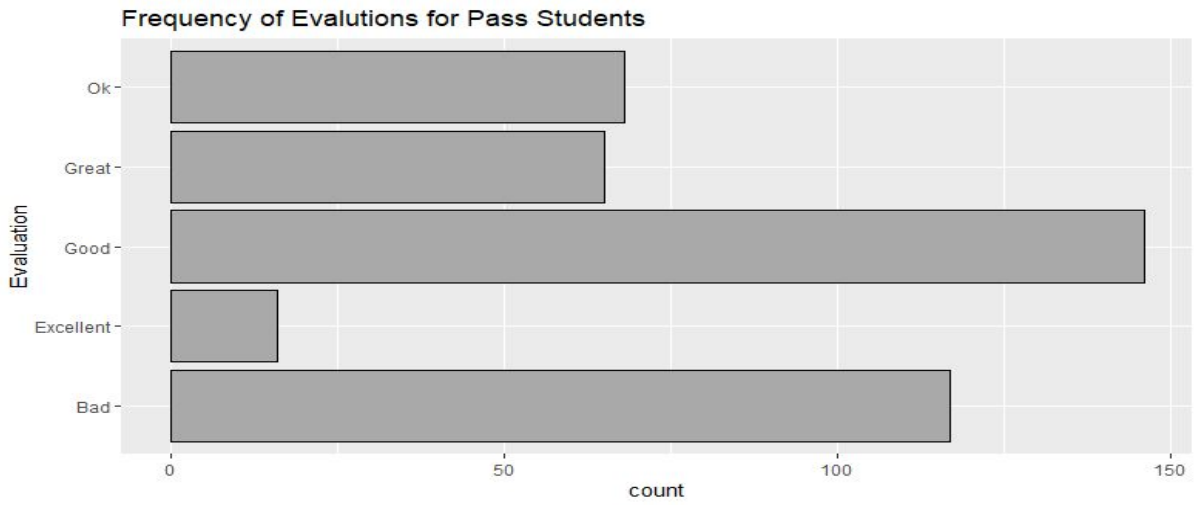
The initial analysis of the course data revealed the following issues.

- The qualitative and quantitative data provided is indicative that the students are neither doing academically well nor feeling positive about the course.
- 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%).
- While quantitatively the data indicated that too many students are unsuccessful, and the ratings provided by the students for the course experience reflects their general displeasement with how the course went for them.



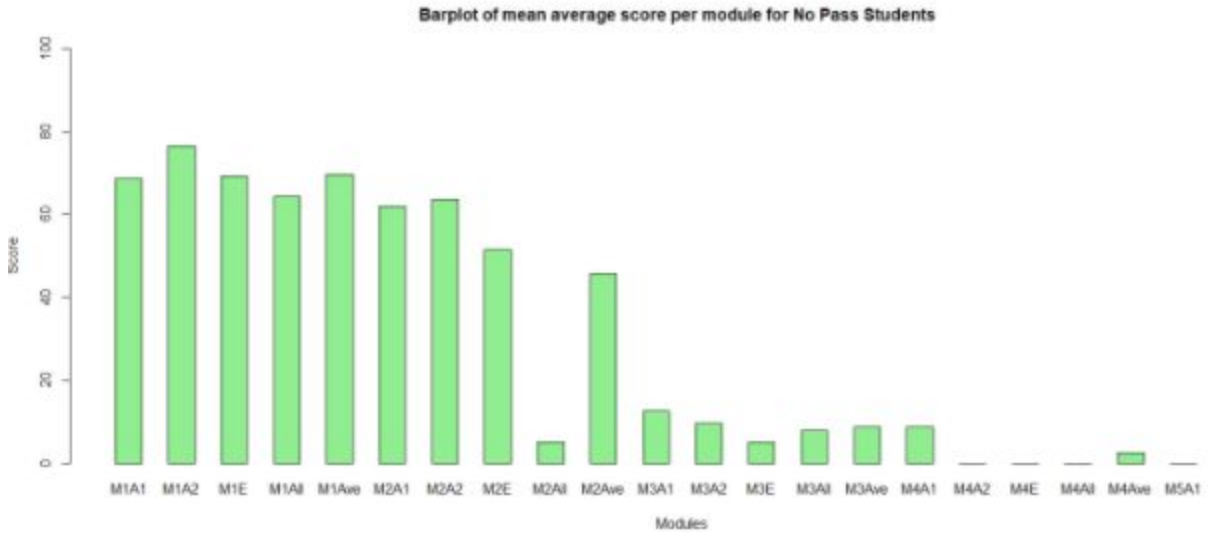
Graph 1: Bar plot indicating the average score of assignments students received per module

- General sentiment on the topics covered in the course and the field of Data Science is positive, so the problem isn't due to the topics. To reiterate, the source of discontent for the students isn't 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 rate the course overwhelmingly negative and indifferent on their evaluations.

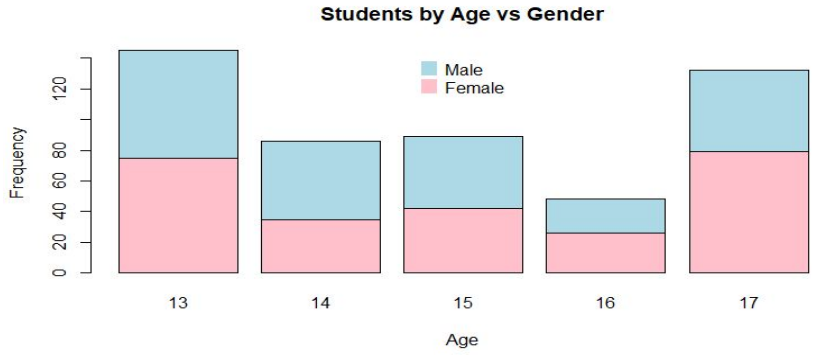


Graph 2: Bar plot of student course evaluations for passing students.

- Students' failure to complete the assignments is the 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 and/or engagement.

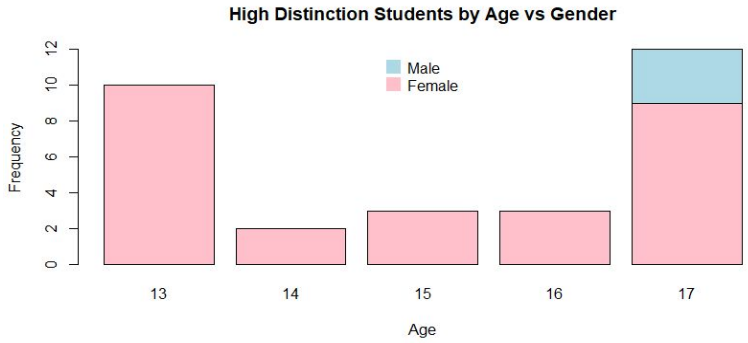


Graph 3: Bar plot of the average score received per assignment by students that did not pass the course



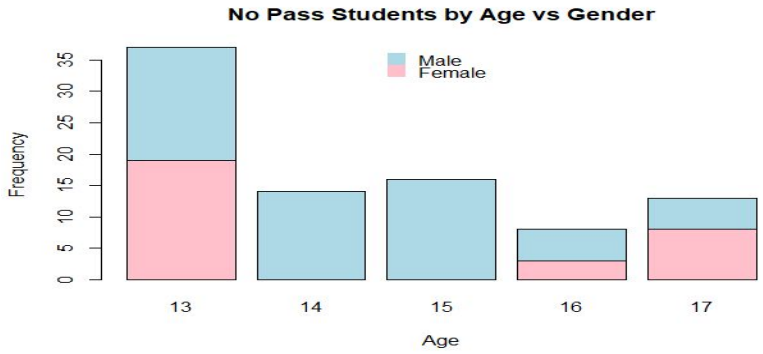
Graph 4: Bar plot indicating ratio of registered students by age and gender

Observation: Ratio of male to female students registered is proportionate



Graph 5: Barplot of students receiving High Distinction by age and gender

Observation: Female Students outperform male students



Graph 6: Barplot of students that fail the course by age and gender

Observation: Fail/Drop rate is higher among male students in age groups 13,14,15 and 16 years

- A significant number of 13 year olds dropped the course. While the percentage dropped for all age groups except 17 year olds is comparatively similar as per the given data, it is important to note that the population size of 14, 15 and 16 year olds is not large and hence the similarity in percentage dropped could be coincidental. This is further compounded by the fact that the only other age with a similar initial registration number is 17 year olds, and 17 year olds have a much lower drop rate. All of this suggests that there is a possibility of underlying causes that specifically affect 13 year old retention rate.
- 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.

Problem Root Cause Analysis

In order to perform a root cause analysis, the contents of the course have been evaluated from the students' 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 they 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, out of the students that drop, their scores are relatively high in comparison to the rest of the students prior to dropping (shown by relatively high scores in module 1). This suggests that 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 for them to understand without a firm understanding of the required foundational concepts. *A possible means to address this issue is by conducting a course entrance test for the students to evaluate their understanding of the prerequisite concepts, thereby determining their eligibility to register for the course.*

- 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 males may push other males even further away. While there may 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 the sample.

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 supposed to be accessed via the hyperlinks can be hosted on VHS 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. Males and 13 year olds perform worse.
- Students vary in their approach to learning and to some extent adopting a student-specific personalized approach to the teaching approach 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.
- The dead hyperlinks have to be cleaned up and the content should be hosted on VHS servers. The purpose of unit assignment should be clearly explained in context of the big picture of the course objectives.
- Optional sections that go into further detail on concepts learned can be introduced to provide extra support for students who need it

The solutions listed above further the UN's fourth goal by improving the educational experience of the students through improved quality of the education delivered to them and promoting equality of education amongst all ages. When students are able to learn the material easier, reproduce it on their own, and achieve better proficiency in the course material, they will be more inclined to be engaged 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 which 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.