# Learning Analytics in Secondary Education: An Iterative Process to Determine Students' Ideal Subject Combination

Peh Zhan Hao, Singapore Management University
Heng Kok Chin, Singapore Management University
Tan Yong Kiong, Alson, Singapore Management University

## ABSTRACT

In today's educational world, more and more educational institutions are incorporating technology into teaching and learning to enrich students' learning experiences and improve teachers' pedagogical practices. The dilemma that our project sponsor face is how to aptly establish the right criteria to recommend the right subject combinations to students, so as to improve their learning outcomes. For instance, it is difficult for teachers to decide whether or not to recommend students to take on the subject combinations of Double Science or Triple Science. Should schools determine the capability of students based on their overall examination grades, or should they base their decisions on their individual subject grades (such as Mathematics or Science)? Often, many parents believe that their child is capable of coping with Double or Triple Science combinations, even if their Secondary 2 results show otherwise. Without proper analytical evidence, it is difficult for teachers to convince parents that the recommended subject combination is the better option for their child. We performed an in-depth analysis of students' academic performance using an extensive set of data extracted from our sponsor's database, comprising of students' historical examination results. We also developed an application which offers our sponsor the prescriptive solutions they need, to support their decision making. This paper aims to illustrate how the application of learning analytics in educational institutions can generate useful and actionable insights that can aid teachers in their decision making. We share our views on how learning analytics and visualization techniques can drive better predictions of students' success and enrich students' learning experiences.

## INTRODUCTION

Our project sponsor is a heartland neighbourhood secondary school located in the North region of Singapore. The school is committed to providing an ideal learning environment and experiences for its students. Despite having a comprehensive set of past students' data, the school lacks the expertise to analyse the data in a way that can aid in their decision making. This paper aims to illustrate how the integration of learning analytics can aid teachers in their decision making and improve students' learning experiences.

## MOTIVATION

The role of data analytics is becoming even more relevant and important, given the rise of Learning Analytics. Learning analytics seek to improve teaching and learning through the targeted analysis of students' academic performance data [1][2]. By analysing the past data of students' examination results using various data analysis and visualization techniques, it enables the school to discover useful patterns and relationships within the data. These insights equip the school with the intelligence that would enable them to better understand students' performance and make informed decisions in their curriculum to refine their pedagogical strategies and optimize student performance [3].

## SUMMARY

The focus of our paper is to discuss how, by applying learning analytics in education institutions, will enable decision makers to make more data-driven decisions, which could significantly impact student performance and overall learning experiences. We will discuss our methodology process, followed by application development, and the design considerations and best practices that we have learnt through our analysis.

## LITERATURE REVIEW

In recent years, there has been the development of several dashboard applications to support teaching and learning in the education sector. Particularly, the advent of learning dashboards help teachers improve their knowledge of students by providing tools for the review of analysis of students' history [4]. Such dashboards often provide teachers with the graphical representation of a student or a course, often in the form of bar charts and matrices. Such visualization techniques enable them to make flexible, data-driven decision making. While previous research focuses on the evaluation of course activity and teaching practices, none has been done specifically in analysing students' examination scores. As such, our study will attempt to fill in the gap and shed light on how learning dashboards can be applied to generate insights of students' academic performance, thereby improving their learning outcome.

However, the challenge lies in visualizing and generating insights from the large datasets and sources. To tackle this problem, we explored the use of two visualization methods: box-and-whisker plot and tableplot. The boxplot is a simple yet powerful tool for displaying a single group of data, allowing the user to easily study the summary of the distribution [5]. On the other hand, the tableplot can display the aggregated distribution patterns of numerous variables in one single figure [6]. The tableplot is a valuable tool for inspecting statistical data, especially when there are numerous variables and a large dataset involved. We will demonstrate how the use of these methods can aid teachers in analysing students' performance in our paper and through our application.

## METHODOLOGY

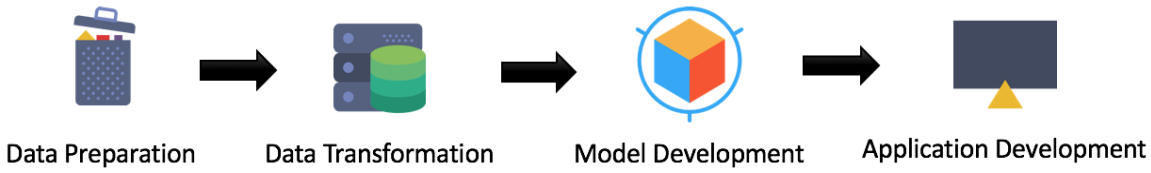



**Figure 1. Data Methodology Process**

Figure 1 shows the data methodology process that we have adopted for our study. The dataset comprises of 3 batches of historical students' data from our project sponsor, consisting of students who took their GCE 'O' Level examinations and graduated in years 2014, 2015 and 2016. The student records comprised of their respective subject results for each of the continual and semester examinations (i.e. CA1, SA1, CA2 and SA2) from Secondary 1 to 4, as well as their PSLE and GCE 'O' Level results. To protect the confidentiality of students and to ensure accuracy of the results, the names of students have been coded by our Project Sponsor.

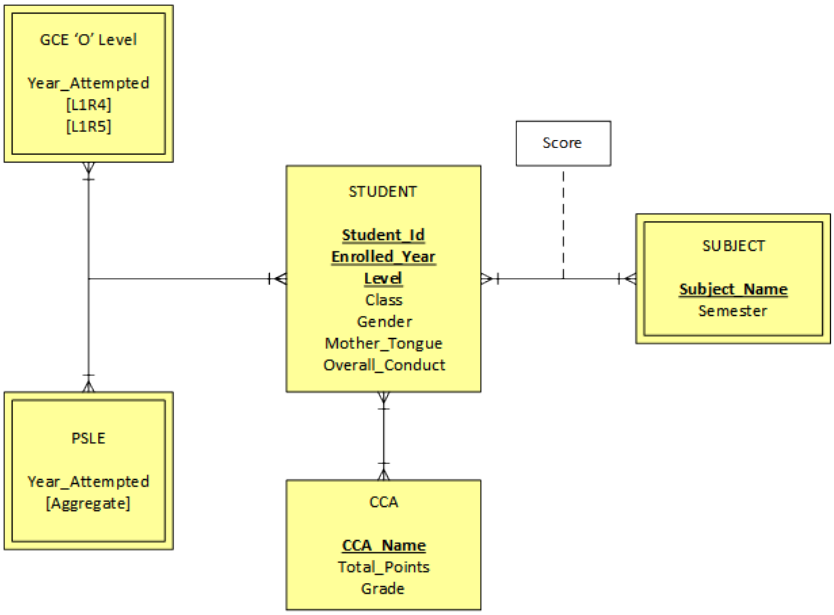
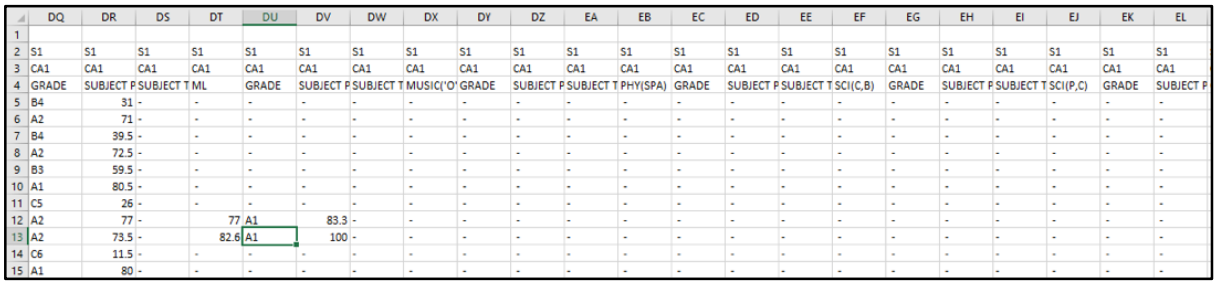



**Figure 2. Entity-Relationship Diagram of the Data**

## DATA PREPARATION

Before proceeding with the analysis, we had to perform significant data cleaning and preparation to ensure the consistency of our variables and better interpretation of the dataset, given the nature of our data. Our initial data was provided on a per batch per examination basis (i.e. CA1 Batch of 2014, SA1 Batch of 2014 etc.), which required us to combine the data together. Also, as the subject combinations of the students differed from one another, we had to categorize them into their respective subject combination to proceed with our analysis. As a result, we spent a considerable amount of time and effort in cleaning and transforming the datasets.

In addition, the data consisted of many redundant columns with missing values which we had to eliminate, as seen in Figure 3. To ensure the consistency of our analysis, we also eliminated the data of 7 students who previously retained (did not take the same GCE 'O' Level examination as their other peers from the same batch) as their records contained numerous missing data.



| | DQ | DR | DS | DT | DU | DV | DW | DX | DY | DZ | EA | EB | EC | ED | EE | EF | EG | EH | EI | EJ | EK | EL |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | | | | | | | | | | | | | | | | | | | | | | |
| 2 | S1 | S1 | S1 | S1 | S1 | S1 | S1 | S1 | S1 | S1 | S1 | S1 | S1 | S1 | S1 | S1 | S1 | S1 | S1 | S1 | S1 | S1 |
| 3 | CA1 | CA1 | CA1 | CA1 | CA1 | CA1 | CA1 | CA1 | CA1 | CA1 | CA1 | CA1 | CA1 | CA1 | CA1 | CA1 | CA1 | CA1 | CA1 | CA1 | CA1 | CA1 |
| 4 | GRADE | SUBJECT P | SUBJECT T | ML | GRADE | SUBJECT P | SUBJECT T | MUSIC('O' | GRADE | SUBJECT P | SUBJECT T | PHY(SPA) | GRADE | SUBJECT P | SUBJECT T | SCI(C,B) | GRADE | SUBJECT P | SUBJECT T | SCI(P,C) | GRADE | SUBJECT P |
| 5 | B4 | 31 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| 6 | A2 | 71 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| 7 | B4 | 39.5 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| 8 | A2 | 72.5 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| 9 | B3 | 59.5 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| 10 | A1 | 80.5 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| 11 | C5 | 26 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| 12 | A2 | 77 | - | 77 | A1 | 83.3 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| 13 | A2 | 73.5 | - | 82.6 | A1 | 100 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| 14 | C6 | 11.5 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| 15 | A1 | 80 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |

**Figure 3. Redundant Columns and Missing Data**

We used SAS JMP Pro to facilitate us in the cleaning and transformation of data, by joining the different tables together to obtain a complete dataset. Following which, JMP Pro offers a simple and efficient way of identifying and elimination of missing values and redundant columns. Lastly, we filtered the data and kept the relevant columns that were needed for our analysis which will be explained more in detail below.

## DATA TRANSFORMATION

The next step is to identify the attributes which are relevant for our model, and remove those that are either redundant or irrelevant. Table 1 shows the current criteria adopted by our sponsor in determine which subject combination to recommend and offer to their students.

| Triple Pure Sciences | Double Pure Sciences | 1 Pure + 1 Combined Sciences | Combined Science |
|---|---|---|---|
| • Top 45 in the level<br>• Achieved Mathematics and Science scores >= 70% | • Top 110 in the level<br>• Achieved Mathematics and Science scores >= 67% | • Mathematics and Science scores possibly around 60 - 65% | • Rest of the students |

**Table 1. Current Subject Combination Criteria**

Given that the sponsor base their decisions only on the students' overall Secondary 2 Mathematics and Science scores, we wanted to find out if the other subjects have any influence on students' 'O' Level performance. As such, we have expanded the criteria by selecting all their overall Secondary 2 subject scores to include in our analysis for developing our regression model.

## MODEL DEVELOPMENT

To determine which subjects are good predictors of students' 'O' Level performance, we performed a multivariate regression analysis with the students' respective subjects scores. By applying a regression analysis, we identified the relevant subjects that are significant in predicting students' 'O' Level L1R4 and L1R5 scores. Figure 4 shows the results of our analysis.

## ANALYSIS

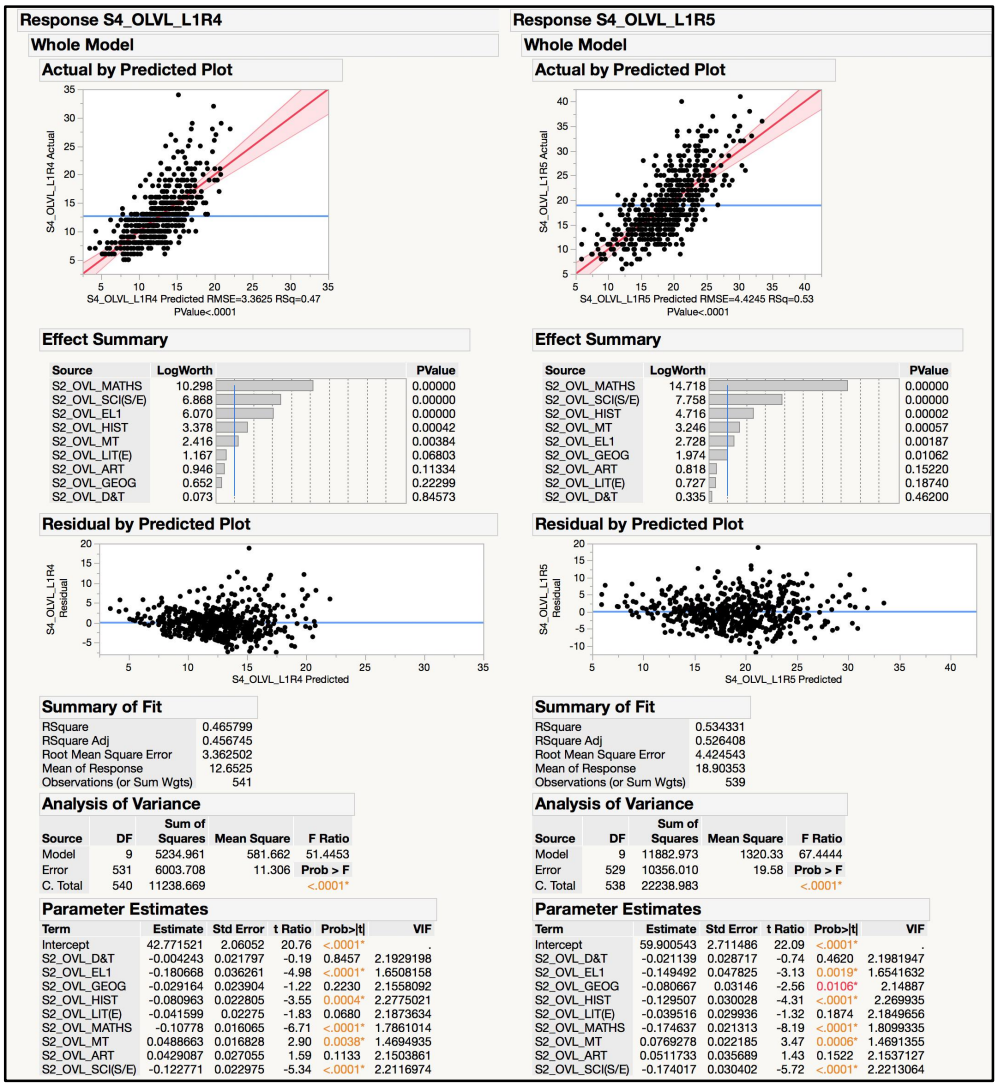



**Figure 4. Multivariate Regression Analysis Results**

The first step in analysing the results is to check for multicollinearity - a phenomenon where two or more variables in a regression model are highly correlated, which could affect the variance and stability of the regression coefficients [7]. Since the Variance Inflation Factor (VIF) of all our variables are below 8 (according to the general rule of thumb), we can conclude that there is no multicollinearity in our multivariate regression model.

The Fit Model equation for predicting **L1R4** would look like this:
42.8 – (0.00424 * D&T) – (0.181 * EL1) – (0.0292 * GEOG) – (0.0810 * HIST) – (0.0416 * LIT(E)) – (0.108 * MATHS) + (0.0489 * MT) + (0.0429 * ART) – (0.123 * SCI(S/E))

On the other hand, the Fit Model equation for predicting **L1R5** would look like this:
59.9 – (0.00211 * D&T) – (0.149 * EL1) – (0.0807 * GEOG) – (0.130 * HIST) – (0.0395 * LIT(E)) – (0.175 * MATHS) + (0.0769 * MT) + (0.0512 * ART) – (0.174 * SCI(S/E))

From the results, we also concluded that the following variables are significant (p-values < 0.05) in predicting both the GCE 'O' Level L1R4 and L1R5 scores: **Mathematics, Science, English, History and Mother Tongue**. Our two models achieved an **adjusted R-Squared value** of **45.67%** and **52.64%** respectively, which suggests that our model explains around half the variability of our data around the mean. Although our models achieved rather low R-Squared values, this is common in the fields of psychology in the prediction of human behaviour [8]. Furthermore, our predictors have very low p-values, which suggests that they are statistically significant.

## APPLICATION DEVELOPMENT

Having developed our regression model in determining students' 'O' Level performance, we proceeded with the development of a web application that would enable teachers to easily and interactively achieve the following outcomes:

1. Determine which is the ideal subject combination for a student and;
2. Analyse the current academic standing of that student compared to his or her peers.

Having identified the relevant attributes that are statistically significant from our regression model, the next step is to design a model that would enable us to determine the range of possible outcomes of students' 'O' Level performance, based on their subject combinations.

### MONTE CARLO SIMULATION

To simulate students' future performance, we have performed a Monte Carlo simulation to estimate the 'O' Level performance of students. Monte Carlo simulation is a type of simulation where repeated random sampling and statistical analysis are performed to compute the results [9]. To apply this in determining the ideal subject combination for a student, we can break it down in the following steps [10]:

1. First, using the results of a current student, we obtained a group of students in the past who had similar Secondary 2 results.

2. With this group of students, we further split them into groups according to their subject combinations and compute the mean and standard deviation for each subject combination.

3. For each subject combination, we generated S independent samples of N random data which are normally distributed using the mean and standard deviation from each subject combination.

4. A sample mean and standard deviation was calculated for each sample and at the end of generating all the samples, an overall mean and standard deviation of all the samples was computed.

5. Using a confidence level of 95%, a confidence interval was calculated for each subject combination using the following formula:

$$\overline{x} \pm z^* \frac{\sigma}{\sqrt{n}}$$

6. With a range of estimated 'O' Level performance for each subject combination, determining the ideal subject combination will then be based on which subject combination will offer the lowest L1R4 or L1R5 scores as well as the spread of the confidence interval.

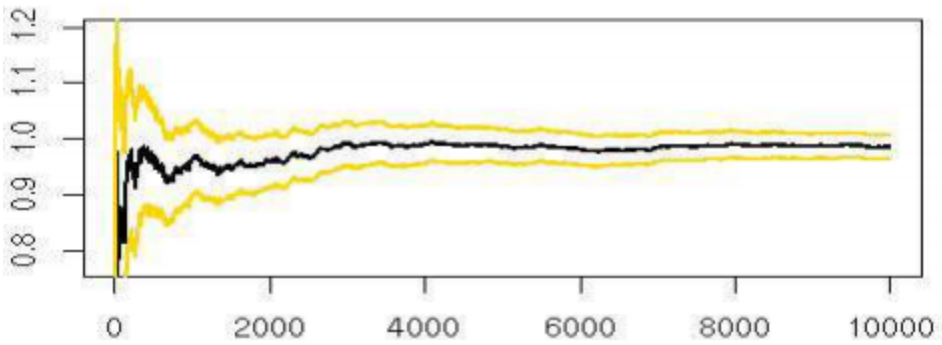



**Figure 5: An Example of how Convergence Occur as Number of Repetitions Increases**

By performing Monte Carlo simulation, we are trying to obtain a range of values that the results will eventually converge towards. This convergence is valid by the Strong Law of Large Numbers which state that as more repetitions are performed, the results will be closer to the expected value [11], making the analysis more accurate.

## DESIGN PRINCIPLES & ARCHITECTURE

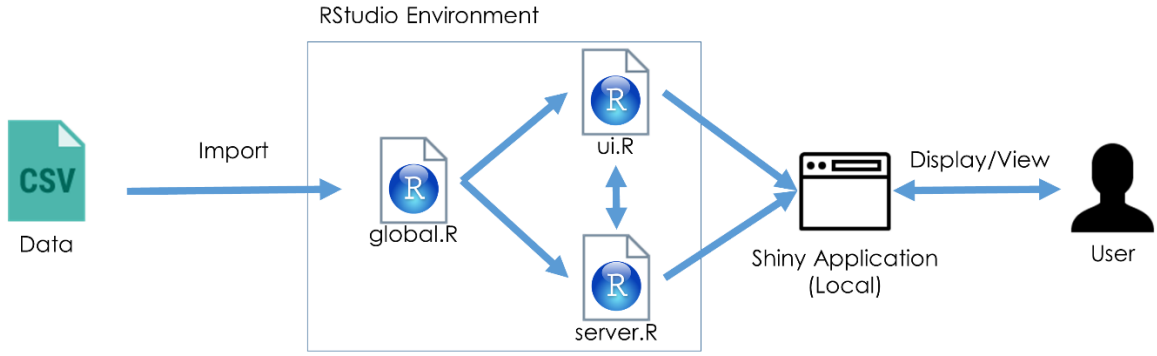



**Figure 6: Architecture Overview of Our Application**

After designing our model, we proceed to the development of our web application. We have decided to build the application using R, a language and environment for statistical computing and graphics. Specifically, we have chosen the Shiny package, an open-source R package that offers an elegant and powerful web framework which allows us to build interactive web applications.

The overall architecture of the application is quite simple and can be seen in Figure 6. Data provided by the sponsor in .csv files will be imported into RStudio. The three main files that we would be using are the global.R, ui.R and the server.R files. global.R controls variables that are used by both ui.R and server.R. ui.R focuses on the input elements and output elements that the user will see and interact with. server.R handles the processing and backend handling of the inputs entered by the user and returns an output to ui.R. With these three files, it creates the Shiny application which the user interacts with.

The general layout of the web application is to have a navigation toolbar at the top, which is what users are familiar with. There will be a similar layout within each of the tabs for each function. The layout would have two main portions. The left sidebar would be used to collect user inputs and the right main area would be used for the displaying of the visualization. The web application will aim to be simple and intuitive to use, promoting interactions between the user and the web application to better suit the users' needs and requirements.

The next step is to provide decision makers with the visualization they need to analyse the results. Figure 7 shows Version 1.0 of the R Shiny application that we have developed. The left sidebar of the web application allows the user to input a student's Secondary 2 overall scores for Mathematics, Science and English, and pre-define a value for the range (i.e. the spread of student's scores from the input score to take into consideration). With the input values, the application would search through the dataset for students whose performance meets the criteria that falls within the range, and return the average 'O' Level L1R4 and L1R5 scores of those students.

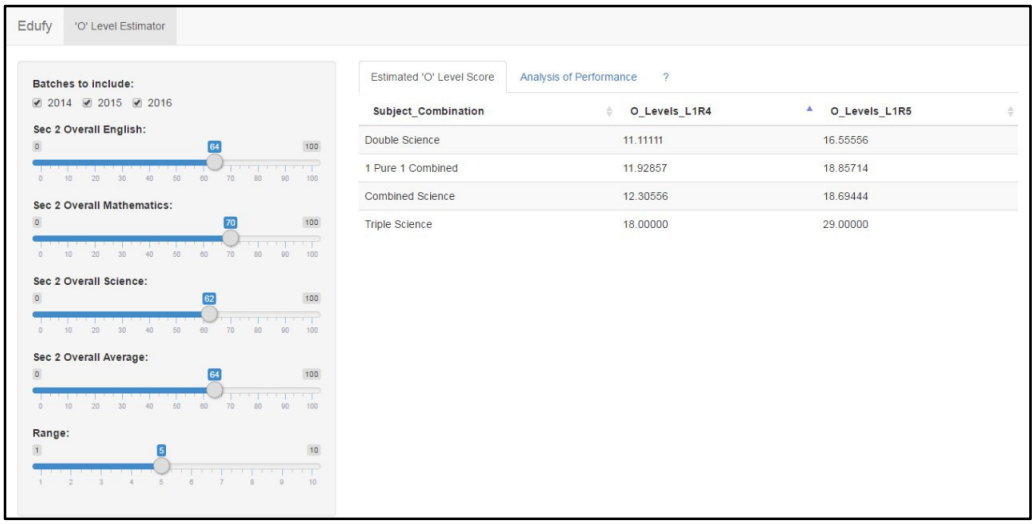



**Figure 7: R Shiny Application - 'O' Level Performance Estimator (Version 1.0)**

In addition, our application aims to allow teachers to be able to view the relative performance of a student, as compared to his or her peers of the same batch. Given the input values, the application would generate a bar chart with the input score as compared to the average subject scores for all students of the same batch as shown in Figure 8.


**Figure 8: R Shiny Application - Analysis of Student's Performance (Version 1.0)**

However, upon reviewing the application with our supervisor, we realized that there were several shortcomings in our design considerations. Firstly, the user experience (UX) of the application could be improved by allowing the user to select or key in the student's ID, instead of having to manually input their respective subject scores. This change can be done by simply pre-storing and loading the new data (i.e. student subject scores), and retrieving the values based on the student's ID. This change would make the application much more user-friendly and convenient for the user.

In addition, there were also limitations in the graphical representation of our data. By representing the student's relative performance in the form of bar chart, the user is only able to see the differences in the frequency (or raw count) on the Y-axis (vertical). In situations where the student's score is similar or equal to the average score, the bar chart would not offer much information. To provide the user with more quantitative information of the student's relative performance, it would be more appropriate to represent the data in the form of a box-and-whisker plot, where the end user would be able to know more descriptive information (i.e. the first quartile, the median, the third quartile, and the maximum value).

## MODEL ITERATION

With these design considerations in mind, we performed the next iteration of our application.

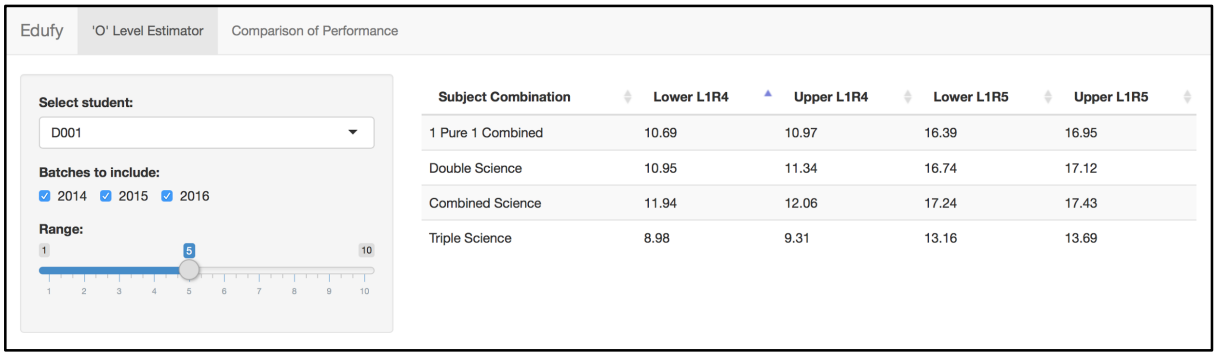



**Figure 9: Web Application - 'O' Level Estimator (Version 2.0)**

Based on Figure 9, the user has an improved user experience as he or she can easily select a student's ID from a dropdown list, and decide which batches of students' data and range to be included in the Monte Carlo simulation. The resulting data table on the right side displays the lower and upper bounds of the L1R4 and L1R5 score that the student is likely to obtain, based on the respective subject combinations.

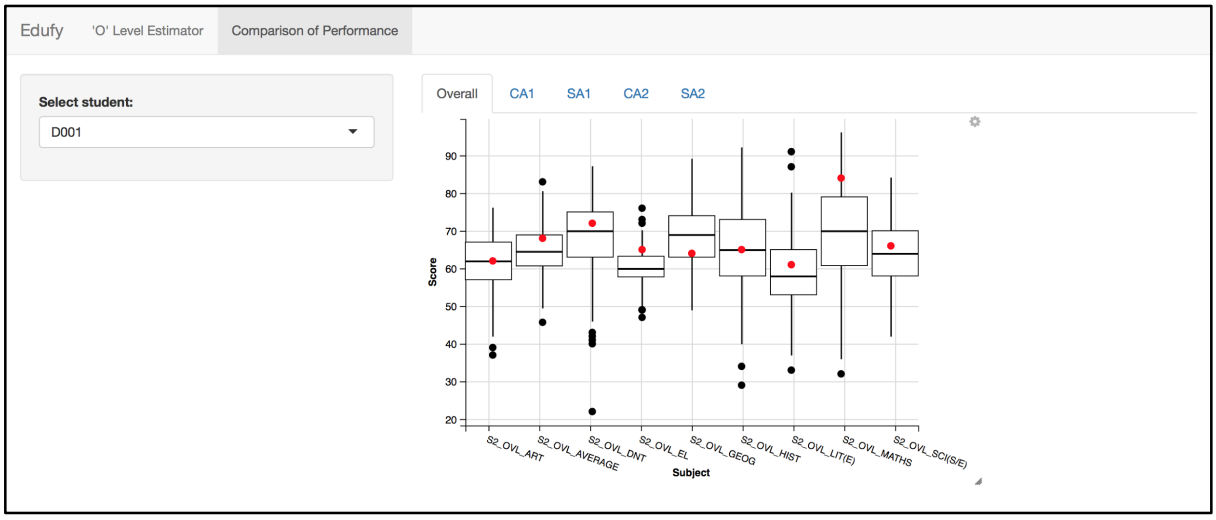



**Figure 10: Web Application - Comparison of Performance (Version 2.0)**

From Figure 10, the user is also able to view a student's relative performance for each of the student's examinations throughout his or her Secondary 2 journey in the form of a box-and-whisker plot. The red dot represents the student's performance while the middle line of the box represents the median score of students of the same batch. By representing the data in the form of a box-and-whisker plot, it allows the user to have a quick visualization of the relative performance of the student compared to his or her peers.

We continued to iterate and improved on our existing application based on feedback from our supervisor and sponsor, leading to Version 3.0 of the application.

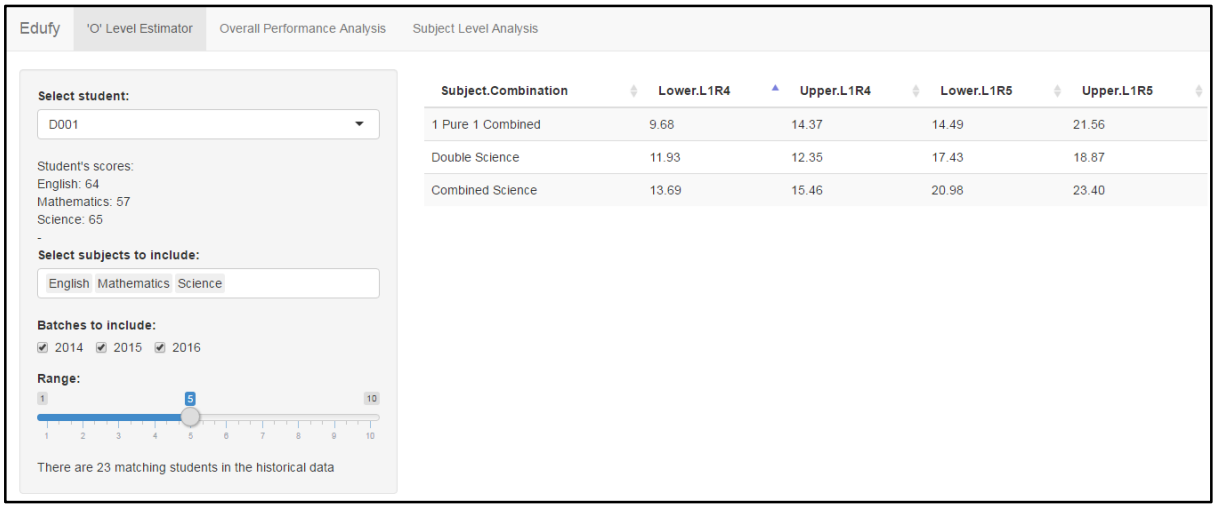



**Figure 11: Web Application - 'O' Level Estimator (Version 3.0)**

Several improvements were made to the 'O' Level Estimator (Figure 11). Firstly, we allowed the user to select which subjects to include when trying to sieve out past students with similar scores. This can be useful to the user to experiment and try out different combinations of subjects that might be significant in determining the L1R4 and L1R5 scores. Secondly, after selecting the subjects to be included, the scores of the selected student for those subjects are displayed for reference. Lastly, the number of past students who have similar scores falling within the range is also displayed for reference.

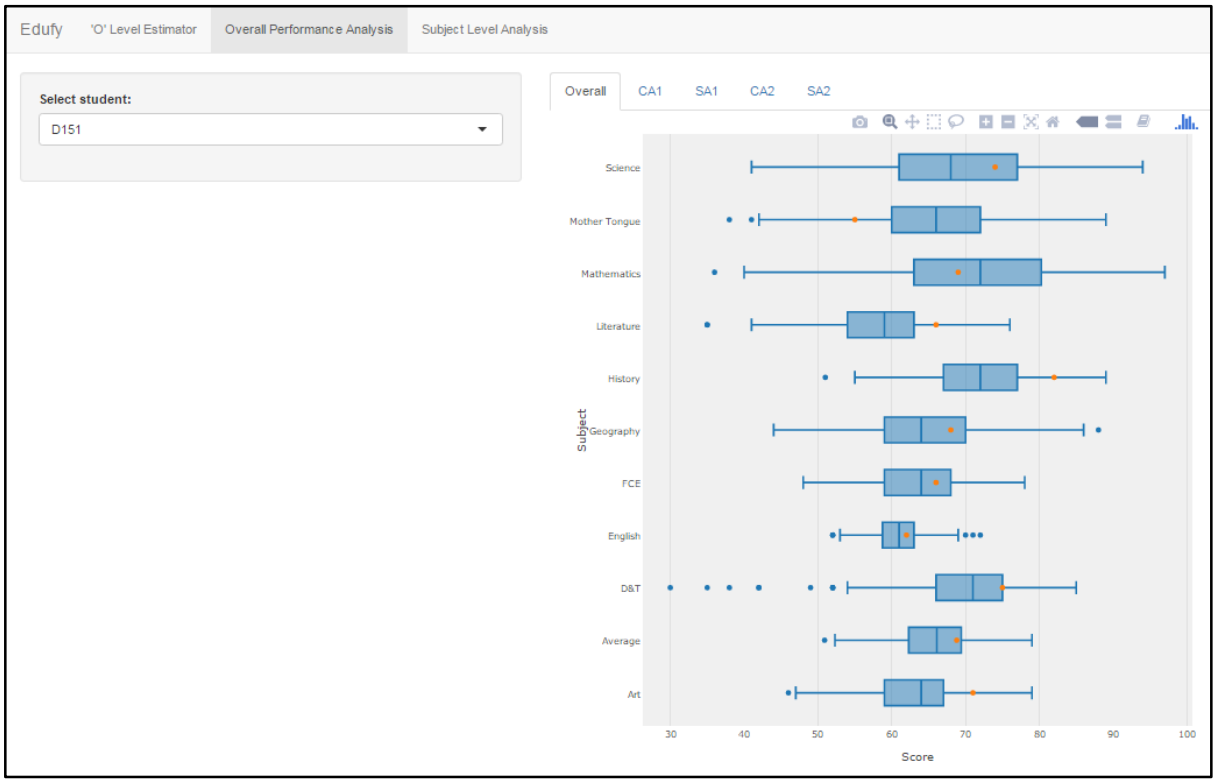



**Figure 12: Web Application - Overall Performance Analysis (Version 3.0)**

The box-and-whisker plot for the 'Overall Performance Analysis' was changed to a horizontal orientation from the original vertical orientation to signify progress (from left to right) rather than superiority (from down to up). The colour of the dots that represents the scores of the selected student was also changed from red to orange as red implies failure or a bad score. The overall look was also changed to be more visually appealing (Figure 12).

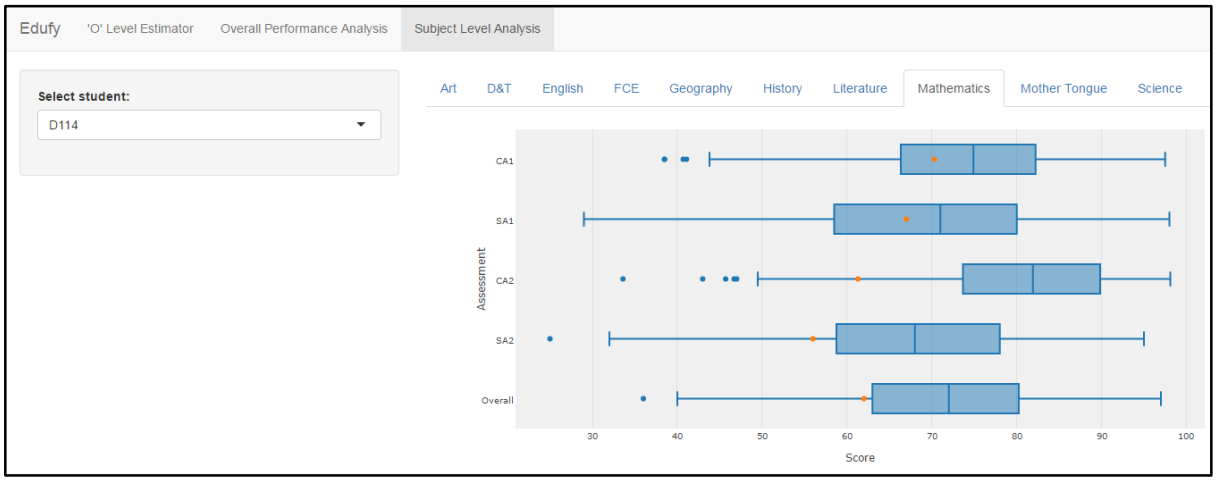



**Figure 13: Web Application - Subject Combination Analysis (Version 3.0)**

A new visualization called 'Subject Level Analysis' (Figure 13) was added for this version. This visualization allows the user to view how well a student performs for each of the subjects across the semesters in Secondary 2. This can be useful in assessing the consistency of a student's performance. Other subjects can be selected via tabs above the box-and-whisker plot.

In Version 4.0 of our web application, we added a new visualization called the tableplot (Figure 14). With this visualization, our sponsor can better adjust the criteria for the Secondary 2 subject combinations. Users can select the subjects to compare across, sort by one of the columns and set certain criteria (such as Mathematics more than 70 marks) and the visualization will update itself to show how many students fit the criteria.

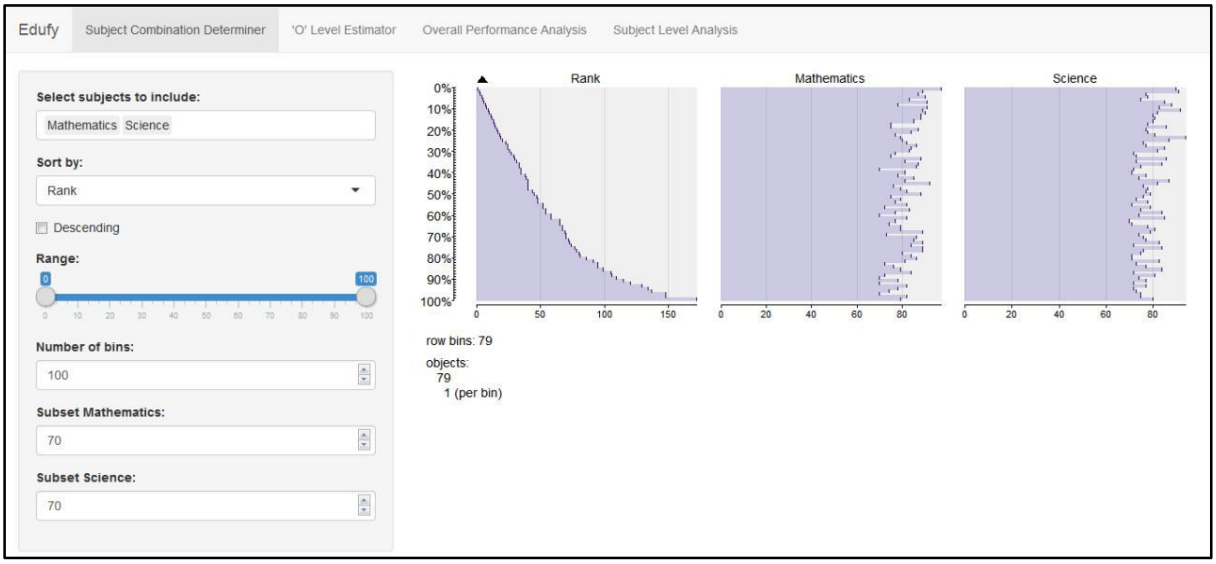



**Figure 14: Web Application - Tableplot (Version 4.0)**

We have also improved the box-and-whisker plot to make it more visually appealing. It also has a more informative tooltip as compared to the previous versions. By presenting it in a neater way, users can better understand how to utilize this box-and-whisker plot.

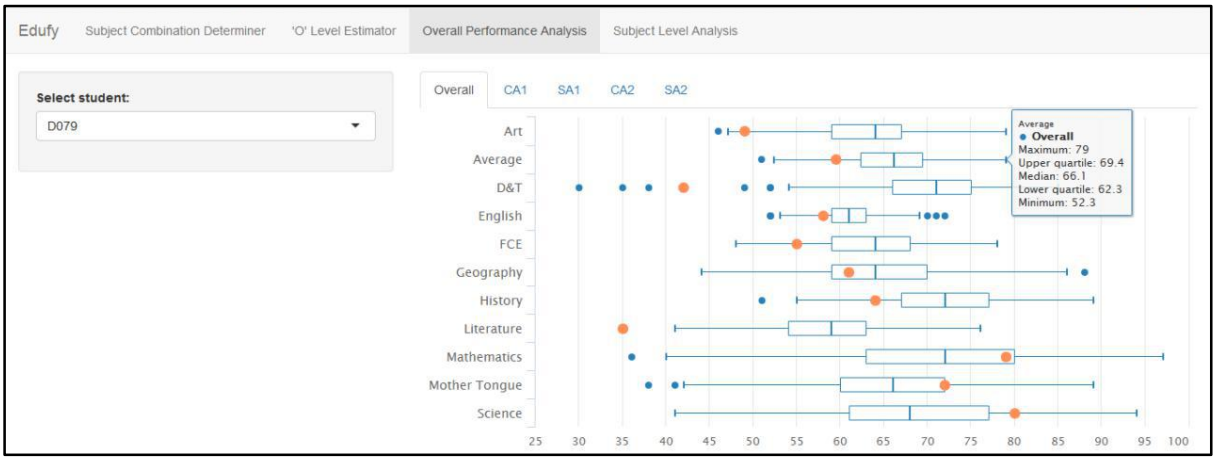



**Figure 15: Web Application – Improved Box-and-Whisker Plot (Version 4.0)**

## USE CASES OF APPLICATION

To better visualize the usefulness of this web application, we have developed use cases to illustrate how this tool can help teachers and decision makers make better, data-driven recommendations on subject combinations.

### A. SUBJECT COMBINATION DETERMINER


**Figure 16. Subject Combination Determiner – Landing Page**

If a teacher wishes to determine the number of students to be allocated into the Triple Science Combination, he or she can simply select the subjects to include, and drag the desired cut-off for each of the subjects as seen in Figure 16.


**Figure 17. Subject Combination Determiner – Selected Subject Cut-off Grades**

Next, the teacher can select the desired cut-off for each of the subjects as seen in Figure 18. The tableplots will automatically compute and display the number of students that qualify for the Triple Science combination. This can be performed for the other subject combinations, which provides the teacher with a useful tool to determine the number of students for each combination based on their resources such as manpower.

### A. 'O' LEVEL ESTIMATOR

The 'O' Level Estimator is used on Student D105 as a use case for the rest of the analysis. When this student is selected on the application, the Lower and Upper L1R4 & L1R5 are generated. Also, the student's scores for the respective subjects selected are also known.

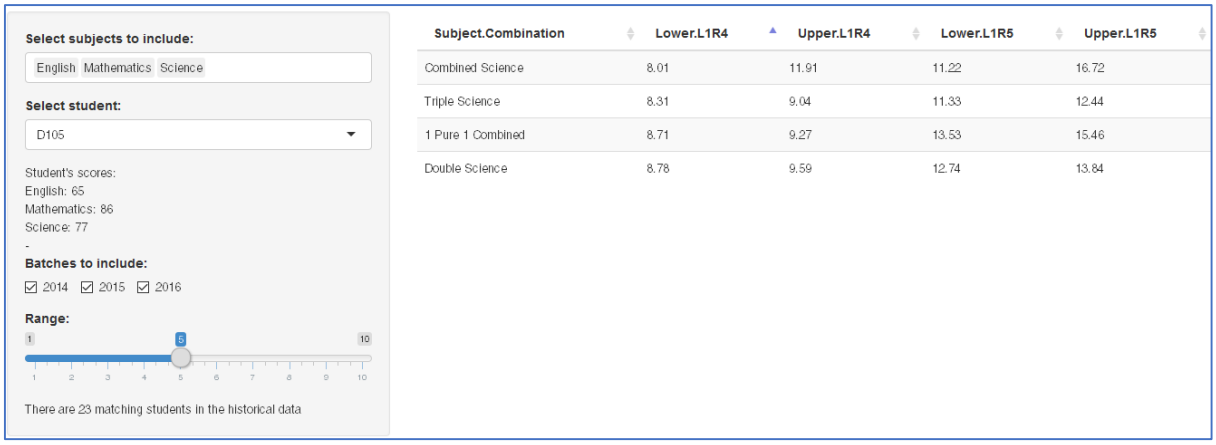



**Figure 18. 'O' Level Estimator – Student D105 selected**

As seen in Figure 18, based on the 23 matching students in the historical data which is within 5 marks range of each subject score, Combined Science provides the best in the lower L1R4 and L1R5 at 8.01 and 11.22 respectively. However, Triple Science provides the best in terms of upper L1R4 and L1R5 at 9.04 and 12.44 respectively. This analysis illustrates that although Combined Science seems to be the better choice amongst the four subject combinations in the best case scenario, Triple Science provides the better option with a small range of the eventual 'O' Level results based on historical data.

If the user wishes to extend the range to 10, meaning the students are matched in the historical data within 10 marks of each subject, there will be 112 matching students.

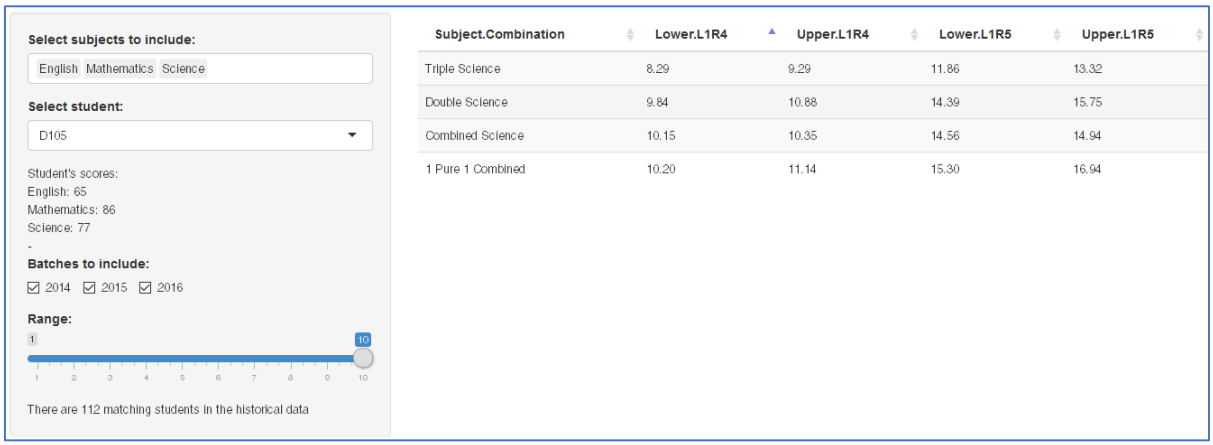



**Figure 19. 'O' Level Estimator – Student D105 selected with Range 10**

As seen in Figure 19, the Triple Science provides the best lower and upper L1R4 and L1R5 amongst the four subject combinations at 8.29 – 9.29 and 11.86 – 13.32.

### B. OVERALL PERFORMANCE ANALYSIS

Taking D105 as a use case, we will show the application usability with regard to the overall performance analysis.

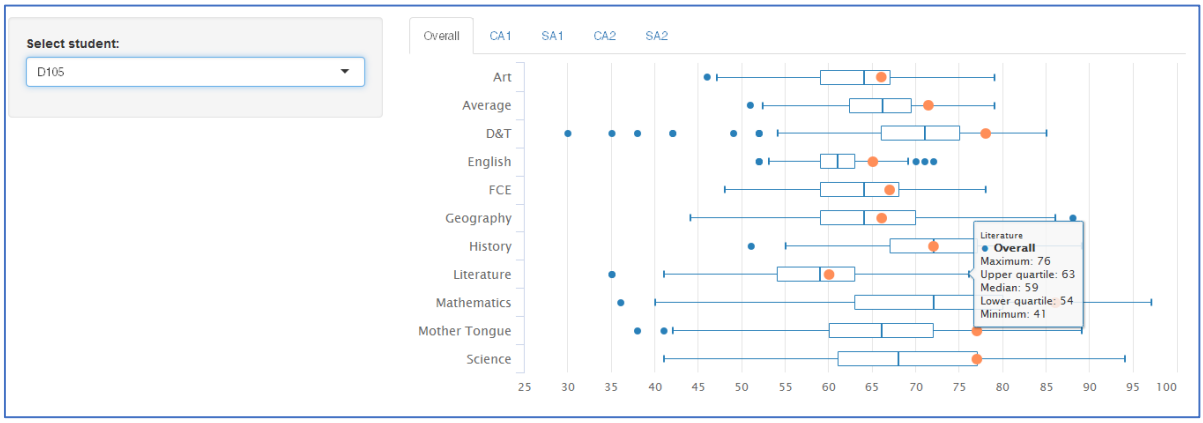



**Figure 20. Overall Performance Analysis – Student D105 selected showing Literature score**

With this application, we can observe that student D105 is an above average student, scoring in the upper 50% of the student cohort. We will observe his Literature score for a more in-depth analysis. D105 scored 60 marks for his Literature. This puts him in the upper 50% of his student cohort. We are able to find the interquartile range by subtracting the upper quartile (75% percentile) from the lower quartile (25% percentile).

Interquartile range = 63 – 54 = 11

The lower and upper inner fences are denoted on the boxplot as the two tailends excluding the outliers. The calculation are as follow:

Lower inner fence = Q1 – 1.5*IQR
= 54 – 1.5*11
= 37.5

Upper inner fence= Q3 + 1.5*IQR
= 63 + 1.5*11
= 79.5

Since 35 marks is lower than the lower inner fence, it is displayed as an outlier.

## DISCUSSION

To sum up, these are the visualizations that our web application can generate and provide:

- Tableplot - This shows the ranking of students and their performance across significant subjects, all in one glance. Users can select the subjects to consider and display, sort these subjects in ascending/descending order, zoom in to look at a range of students and to subset and filter out students that do not meet the input criteria

- Data Table - This shows the estimated ‘O’ Level scores of current students, using results of past students as a basis. Users can select a student (data is pre-loaded) and view their current scores for significant subjects, select which are the subjects to consider for the analysis, the amount of data to include and the range of scores to include

- Box-and-whisker Plot - This shows the performance of a student in relative to his/her batchmates across all subjects for an assessment period. Users can select a student to view his/her performance compared to the rest of the students

- Box-and-whisker Plot - This shows the performance of a student in relative to his/her batchmates across all assessment periods for a subject. Users can select a student to view his/her performance compared to the rest of the students

For our sponsor, each of the above-mentioned visualization can provide them different insights.

- Utilizing the tableplot, our sponsor can re-assess the school's current criteria for the subject combinations. They can experiment the subjects that they want to consider and subset the scores for each subject to determine the criteria. At the end, they can see the number of students that fits the criteria. This can help them to decide if the criteria set is too stringent (resulting in too little students fitting the criteria) or too strict (resulting in too many students fitting the criteria)

- The data table allows our sponsor to look at individual student's performance and based on historical results, estimate (based on certain significant subjects) the likely range of 'O' Level scores that the student will obtain. This is also based on a what-if analysis on each of the subject combinations. What is the likely 'O' Level L1R4 score if this student chooses the 'Double Science' subject combination as compared to the 'Combined Science' subject combination?

- The two box-and-whisker plots helps identify students who are not performing so well as compared to their peers. It can either be the case that the students are consistently not performing well or the case where the performance of a student suddenly drops, signifying that there is a potential problem with his learning experience. With early identification of these symptoms, our sponsor can act fast to rectify these symptoms and help the students perform better

In our case, our analysis is only applied to Secondary Education. Across the education sector, our study can be replicated to analyse results from Primary Education and Tertiary Education with a few tweaks and adjustments. There are also criteria in Primary Education (EM1, EM2, EM3) and Tertiary Education (Art/Science streams) that could employ the use of tableplots. Cut-off points for entry to certain educational institutions could also make use of tableplots, which excels at analysing large datasets. Our 'O' Level Estimator can also be used to predict PSLE scores and GPA scores.

Looking beyond the education sector, the concept of using tableplot in determining criteria can be applied to public policies where subsidies and benefits are given to citizens who fall under a certain tier. The number of citizens that could benefit can be determined using tableplot and setting certain requirements (such as income level, age group) which subsets the data. The concept of estimation based on historical results can also be applied to other areas such as project management. Given the data of projects (personnel, resources, time) managed in the past, companies could estimate the time necessary to complete a project given a team of people. This can help in better allocation of manpower and planning.

## CONCLUSION

Integrating technology into Secondary Education offers significant value to educational institutions. By applying learning analytics, the systematic collection and analysis of data helps drive predictions for student success that are actionable and can be refined overtime.

Interactive visual analytics such as the use of box-and-whisker plots can empower teachers to gain insights and make informed, data-driven decisions that will optimize student performance. In addition, parents can make up for the lack of direct interaction with their child's academic progress, and benefit from such analysis and tools to better monitor their child's performance.

## LIMITATIONS OF WORK

The application has provided teachers and students with a better understanding of the future 'O' Level scores of the current students using the profiles and past students from different subject combinations. The application also provided a comparison of a student with the entire cohort. However, it does not specifically answer the question of which subject combination should a student take. This is because similar scoring profile of the past students' Secondary 2 results may not be available for all four subject combinations due to insufficient data. Since the Monte Carlo simulation is based on historical Secondary 2 results, the insufficient data results in a not-so accurate analysis of the student's L1R4/L1R5 results. The available data are not extensive enough to distinguish

the pros and cons of students taking different subject combinations, which would aid in the decision making of teachers, students and parents.

## FUTURE DIRECTIONS

In addition to the analysis on solely Secondary 2 results, the application can be used for other Secondary and Primary Levels. For example, Primary 4 teachers can utilize the application to predict the students' Primary School Leaving Examination (PSLE) results by the end of Primary 6. Teachers can also use this application to review the results on an individual and cohort-wide scale for each academic level.

The Ministry of Education (MOE) can also employ this application on a nationwide scale, by incorporating and comparing the results from different secondary schools. In conjunction with the goal of MOE where "Every School is a Good School", this application can help to identify the strengths and weaknesses of each school, be it neighbourhood or the more prestigious ones. From there, educators can adjust its academic policies to improve the teaching efficiency of schools to go towards the goal.

This web application can be improved by including tutorials and simple instructions in helping users understand the purpose and functions of this web application. There could also be an import function that allows for on-the-go exploration of data by users.

## SOFTWARE USED

For our initial data cleaning, we used Microsoft Excel to transform and recode our data to prepare it for our analysis. To perform our exploratory data analysis, we have used JMP Pro 13 - a predictive analytics software created by SAS®. Besides the robustness and extensiveness of JMP Pro, it also offers a user-friendly interface, visualization techniques and memory storage functionality which simplifies the effort and reduces the time taken by the user to run the analysis, thereby allowing users to generate graphs efficiently and interactively. To develop the web application, we have chosen to use R language with the RStudio Integrated Development Environment which is an open-source software used for statistical analysis. With the Shiny package from R, we managed to develop a simple yet useful web application for our sponsor.

## ACKNOWLEDGEMENTS

Firstly, we would like to thank our project sponsor for kindly providing us with the access to their school's data and the support from the staff involved in the project. We also thank Professor Kam Tin Seong (Associate Professor of Information Systems; Senior Advisor, SIS) for his valuable advice and guidance throughout this project.

## REFERENCES

[1] Elias, T. (2011). *Learning analytics: Definitions, processes and potential*. Unpublished Internal Whitepaper of Athabasca University, Canada. Retrieved from https://landing.athabascau.ca/mod/file/download.php?file_guid=43713

[2] Fritz, J. (2010). Classroom walls that talk: Using online course activity data of successful students to raise self-awareness of underperforming peers. *The Internet and Higher Education, 49*, 89–97.

[3] Larusson, J., White, B., & SpringerLink. (2014). *Learning analytics : From research to practice , New York : Springer*.

[4] Haythornthwaite, Caroline, De Laat, Maarten, Dawson, Shane, Verbert, Katrien, Duval, Erik, Klerkx, Joris, . . . Santos, José Luis. (2013). Learning Analytics Dashboard Applications. American Behavioral Scientist, 57(10), 1500-1509.

[5] Benjamini, Y. (1988). Opening the Box of a Boxplot. The American Statistician, 42(4), 257-262.

[6] Martijn Tennekes, Edwin De Jonge, & Piet J. H. Daas. (2013). Visualizing and Inspecting Large Datasets with Tableplots. Journal of Data Science, 11(1), 43-58.

[7] O'brien, R. (2007). A Caution Regarding Rules of Thumb for Variance Inflation Factors. *Quality & Quantity, 41*(5), 673-690.

[8] Frost, J. (2013). Regression Analysis: How Do I Interpret R-squared and Assess the Goodness-of-Fit? Retrieved from http://blog.minitab.com/blog/adventures-in-statistics-2/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit

[9] Raychaudhuri, S. (2008). *Introduction to Monte Carlo simulation*. Retrieved from http://www.informs-sim.org/wsc08papers/012.pdf

[10] Zhang, J. T. (2011). Simulation studies in statistics. Retrieved from http://www.stat.nus.edu.sg/~zhangjt/teaching/ST2137/lecture/lec%2011.pdf

[11] Robert, C. P., & Casella, G. (2011). Introducing monte carlo methods with r. Retrieved from http://inst-mat.utalca.cl/jornadasbioestadistica2011/doc/CursoCasella/%20UseR-SC-10-B-Part1.pdf