

**Irene Hyunyou Choi**

***Data Analytics Assignment for Adobe Systems Incorporated Customer and Product Analytics Team***

**Basic Analysis using Shiny Tool,**

**Decision Tree Model, and**

**Logistic Regression Model**

**Documentation**

*Open University Learning Analytics Dataset Analysis*

***Version: DRAFT 1***

***1/4/2018***

***Assignment Started Date:*** *December 30, 2017*

***Assignment Completion Date:*** *January 4, 2018*

***Contents***

[1 Executive Summary 4](#_Toc502881700)

[1.1 Background 4](#_Toc502881701)

[1.2 Summary Model Development Process and Drivers 4](#_Toc502881702)

[2 Model Theory and Approach 5](#_Toc502881703)

[2.1 Model Purpose 5](#_Toc502881704)

[2.2 Candidate Modeling Approaches 5](#_Toc502881705)

[2.3 Selected Modeling Approach 5](#_Toc502881706)

[2.3.1 Decision Tree 5](#_Toc502881707)

[2.3.2 Logistic Regression 5](#_Toc502881708)

[3 Model Development Data 7](#_Toc502881709)

[3.1 Dataset Overview 7](#_Toc502881710)

[3.1.1 Courses.csv 7](#_Toc502881711)

[3.1.2 Assessment.csv 7](#_Toc502881712)

[3.1.3 Vle.csv 7](#_Toc502881713)

[3.1.4 StudentInfo.csv 7](#_Toc502881714)

[3.1.5 StudentRegistration.csv 7](#_Toc502881715)

[3.1.6 StudentAssessment.csv 8](#_Toc502881716)

[3.1.7 StudentVle.csv 8](#_Toc502881717)

[3.2 Dataset size 8](#_Toc502881718)

[3.3 Dataset Manipulation (Dataset Merge, Missing Value Fill, Variable Creation) 8](#_Toc502881719)

[3.3.1 df\_assessments\_courses.csv 8](#_Toc502881720)

[3.3.2 df\_assessments.csv 8](#_Toc502881721)

[3.3.2.1 Submission variable 8](#_Toc502881722)

[3.3.3 df\_studentVle\_v1.csv 8](#_Toc502881723)

[3.3.3.1 Frequency variable 9](#_Toc502881724)

[3.3.4 df\_studentReg\_v1.csv 9](#_Toc502881725)

[3.3.5 df\_studentInfo.csv 9](#_Toc502881726)

[3.3.6 df\_coursesInfo.csv 9](#_Toc502881727)

[3.3.7 shiny\_studentinfo.csv 9](#_Toc502881728)

[3.4 Development Data Structure 10](#_Toc502881729)

[4 Basic Analysis 11](#_Toc502881730)

[4.1 Final result interpretation for each presentation 11](#_Toc502881731)

[4.2 Correlation between courses and results 12](#_Toc502881732)

[4.3 Analysis between student demography and final result 14](#_Toc502881733)

[4.3.1 Plot 1 – student result per module 14](#_Toc502881734)

[4.3.2 Plot 2 – Final Result for each demographic group within a Module ID 15](#_Toc502881735)

[4.3.3 Plot 3 – Assessment Type and Period Information for each presentation 16](#_Toc502881736)

[5 Model Estimation/Development 17](#_Toc502881737)

[5.1 Decision Tree 17](#_Toc502881738)

[5.1.1 Model fit 17](#_Toc502881739)

[5.1.2 Feature Selection Model fit 17](#_Toc502881740)

[5.1.3 Feature Selection improvement 17](#_Toc502881741)

[5.1.4 Conclusion 19](#_Toc502881742)

[5.2 Logistic Regression 19](#_Toc502881743)

[5.2.1 Dataset Preprocessing 19](#_Toc502881744)

[5.2.2 Outline of Variable Selection Procedure 19](#_Toc502881745)

[5.2.3 Model fit 20](#_Toc502881746)

[5.2.4 Model prediction 20](#_Toc502881747)

[5.2.5 Manual Bootstrapping 20](#_Toc502881748)

[6 Data Limitation and Potential Suggestions 22](#_Toc502881749)

[6.1 Data limitation 22](#_Toc502881750)

[6.2 Potential Suggestion 23](#_Toc502881751)

# Executive Summary

## Background

Irene Hyunyou Choi (“I”, or “Model Developer”) analyzed Open University Learning Analytics Dataset (“OULAD”) to provide the result of the analysis and to develop a recommender system for technical and non-technical individuals at Adobe Digital Incorporated (“Relevant Individuals”). The results of these analysis are intended to identify highest correlating factors related to the student results.

Specifically, the objective of this document is to:

1. provide basic analysis of the dataset using correlation and Shiny tool;
2. describe the analysis process for logistic regression;
3. describe the feature selection process using decision tree; and
4. discuss the findings attained by basic analysis, logistic regression and decision tree.

## Summary Model Development Process and Drivers

In order to build a comprehensive and accurate model, the Model Developer considered the most efficient way to combine the dataset. Dataset with two different base was built, student information base and course information base, respectively.

Below table shows the selected model drivers for the Decision Tree Model and Logistic Regression Model. Model drivers of these two different model includes broad indicators of student learning activity and student demography information.

Table 1: Macroeconomic drivers for selected models

|  |  |
| --- | --- |
| **Model** | **Model Drivers** |
| Decision Tree | * Studied\_credits * Count\_Date * Sum\_click\_sum * Avg\_date\_submission |
| Logistic Regression | * Gender * Studied\_credits * Highest\_education * Imd\_band * Num\_of\_prev\_attempts * Module\_presentation\_length * Frequency * Disability * Age\_band * Date\_registration |

# Model Theory and Approach

## Model Purpose

The models discussed in this document are part of a data analyzing framework. The purpose of the Models is to find significant variables for prediction of the student’s final result in the module.

## Candidate Modeling Approaches

A range of candidate approaches are used to find prediction models across the multiple dataset. The table below provides an overview of the most common methods applied.

Table 2: Candidate Approaches Advantages & Disadvantages

|  |  |  |
| --- | --- | --- |
| **c** | **Advantages** | **Disadvantages** |
| **Linear Regression:** A simple regression determining relationship between an independent variable and a dependent variable. | * Simple, transparent, straightforward. * Low model risk due to low complexity. | * Low accuracy in a dataset with large number of variables |
| **Multivariate Regression:** A complex version of the Linear Regression that determines relationship between multiple independent variable and a dependent variable. | * Relatively simple, transparent, straightforward * Able to find relative impact of multiple variables in a dataset | * Require all the assumption test * Low accuracy |
| **Decision Tree:** A decision support tool that uses model of decisions and the possible consequences. | * Simple to use * Easy to understand if there are not too many branches | * Inadequate in applying regression and predicting continuous values * Large decision tree with multiple branches are hard to comprehend |
| **Logistic Regression:** A regression that measures a correlation for a dichotomous dependent variable for multiple independent variables. | * Possible to interpret the relationships of dependent and independent variables | * Overfitting * Multicollinearity |

## Selected Modeling Approach

### Decision Tree

Decision Tree is tree figure graph for decision model. Algorithms of decision tree commonly start from target variable, and choose the most related variables first and subsetted to next related variables. These for the most part measure the homogeneity of the objective variable in the subsets. There are many different algorithms in decision tree model. I chose the entrophy. Entrophy is a measure of how frequently an arbitrarily picked elements. User could set up the max tree depth, and minimum number of cases for parent node and child node. I set up the max tree depth at five. The pruning process might be required. Pruning decision tree means reducing the size of tree by removing node of the tree that provides neglectable relationship. Pruning tree helped to avoid the high risk of overfitting, so it possibly decreased error rate and captured only significant results.

### Logistic Regression

Logistic regression is a statistical method for analyzing a dataset in which there are one or more independent variables that determine a binary outcome. The target variable in dataset is a binary variable categorized into either 0 or 1, and there are many independent variables. Logistic regression computes each dependent variable's coefficients, which means it could find out how the variables associated with target variable either positively or negatively affect the target variable. There are p-value for each variables, which could be used to find out how the variables are associated target variable. This model could compute the confidence interval for our estimate values using standard error result table. I used both forward and both stepwise selection. The process of the forward selection is that started with the entire least squares model containing all p predictors, and then iteratively removes the least useful predictor, one at a time.

# Model Development Data

## Dataset Overview[[1]](#footnote-2)

The OULAD contains data about courses, students and the interactions with Virtual Learning Environment (VLE). The dataset is divided into three schema and further divided into seven csv files. Please see below for the dataset divisions

1. Module Presentation
2. Courses;
3. Assessment;
4. Vle;
5. Student-Demographics
6. StudentInfo;
7. Student-Activities
8. StudentRegistration;
9. StudentAssessment; and
10. StudentVle.

Below subsections will describe each csv files.

### Courses.csv

The courses.csv contains information about the modules, and their presentation and length of each module-presentation.

### Assessment.csv

The assessment.csv contains information about the modules, their presentation, assessment in each module-presentation, including the identification number for each assessments, type of assessment, date of the final submission the modules, and weight of the assessment in proportion to the final exam weight.

### Vle.csv

The vle.csv contains information about the module, their presentation, identification number for each material, role associated with the material, and week from and until which the material is planned to be used.

### StudentInfo.csv

The studentinfo.csv contains information about module, their presentation, identification number of the student, gender, student’s living location, education level on entry, Index of Multiple Depravation (“IMD”) of the location, band of student age, number of attempts on the module for each student, total number of credit that the student is currently taking, materials in the vle, including the module, their presentation, identification number for each material, role associated with the material, and week from and until which the material is planned to be used.

### StudentRegistration.csv

The studentRegistration.csv contains information about module, their presentation, identification number of the student, date of registration and date of unregistration for students who have dropped the module.

### StudentAssessment.csv

The studentAssessment.csv contains information about identification number of the student, identification number of the assessment, indication of assessment result transferal, and score of the assessment.

### StudentVle.csv

The studentVle.csv contains information about module, their presentation, identification number of the student, identification number for the VLE material, date of interaction with the VLE material, and number of times a student interacted with the VLE materials.

## Dataset size

Table 3: Dataset Overview

|  |  |  |  |
| --- | --- | --- | --- |
| **Name** | **Observations** | **Variables** | **Missing variable(count)** |
| Assessments.csv | 206 | 6 | date(11) |
| Courses.csv | 22 | 3 |  |
| studentAssessment.csv | 173,912 | 5 | score(173) |
| studentInfo.csv | 32,593 | 12 |  |
| studentRegistration.csv | 32,593 | 5 | date\_registration(45)  date\_unregistration(22521) |
| studentVle.csv | 106,552,806 | 6 |  |
| Vle.csv | 6,364 | 6 | week\_from(5243)  week\_to(5243) |

## Dataset Manipulation (Dataset Merge, Missing Value Fill, Variable Creation)

Below are description of the newly formed dataset files as results of merges and data cleaning. These merges allowed for the Model Developer to fill out missing values for some variables and allowed the Model Developer to create new variables that might act as a predictor of the student results.

### df\_assessments\_courses.csv

df\_assessments\_courses.csv is a new csv file which combines information of assessments.csv and couses.csv. It contains information about the assessment and length of each module-presentation. By combining the two files, length variable in courses.csv was able to be used for filling the missing value in the date variable in assessments dataset.

### df\_assessments.csv

df\_assessments.csv is a new csv file which combines courses.csv, assessments.csv, and studentAssessmet.csv. With the new csv file, the submission variable was created.

#### Submission variable

Submission variable contains information about the variance in submission date in relation to assessment due date. This variable is used as one of the variables in student activity.

### df\_studentVle\_v1.csv

df\_studentVle\_v1.csv summarizes the studentVle.csv dataset. It contains sum click and total visit date of each Vle for each student. Also, it computes the frequency of the Vle visits.

#### Frequency variable

Frequency variable used for one of the factors of student activity.

### df\_studentReg\_v1.csv

df\_studentReg\_v1.csv combined with studentRegistration.csv and student final result variable in studentInfo.csv. It contained the missing value in date\_registration and date\_unregistration variable, so the date\_unregistration row dropped. This dataset is used to measure the correlation between registration date variable and the student final result variable.

### df\_studentInfo.csv

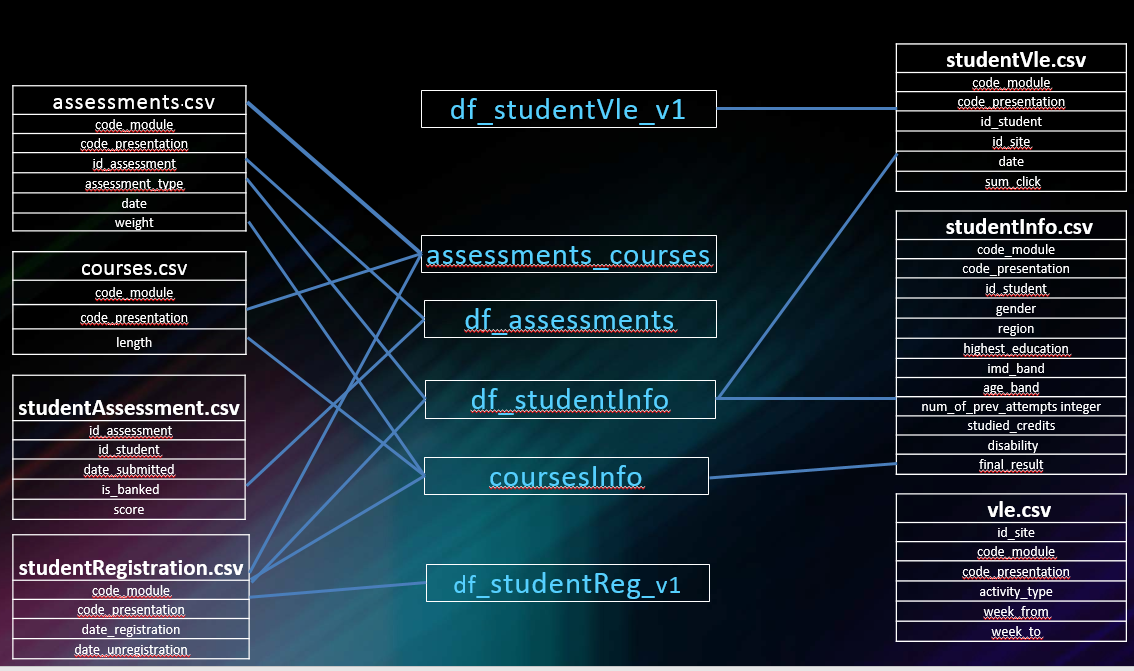
df\_studentInfo.csv combines studentInfo.csv, studentRegistration.csv, courses.csv, and df\_studentVle\_v1.csv. In this dataset, the categorical variables in studentInfo.csv is converted to numerical variables. Unnecessary variables such as X, date\_unregistration are dropped in this dataset. This dataset serves as the main dataset for student base analysis in statistical analysis and model development.

### df\_coursesInfo.csv

df\_courseInfo.csv is created to find the relationship between module-presentation components such as length, number of assessments, and weight of assessments and final results. This dataset is the main dataset used for course base statistical analysis.

### shiny\_studentinfo.csv

shiny\_studentInfo.csv contains same variables as df\_studentInfo.csv but has an additional id column for shiny performance. To run a shiny, I named shiny dataset and defined the module ID.

Figure 1: Data Merge Process

## Development Data Structure

The development data for decision tree and logistic regression models are structured as shown below:

Table 4: Data Structure Example

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| … | highest\_education | imd\_band | age\_band | num\_of\_prev\_attempts | … | final\_result.x | date\_registration | module\_presentation\_length | count\_date | … |
| … | 3 | 3 | 1 | 0 | … | 163 | 0 | 0.003 | 4 | … |
| … | 2 | 9 | 2 | 0 | … | 0 | 0 | 0 | 4 | … |
| … | 3 | 4 | 1 | 0 | … | 0 | 0 | 0 | 4.6 | … |
| … | 4 | 7 | 1 | 0 | … | 0 | 0 | 0 | 4.6 | … |
| … | 2 | 4 | 2 | 0 | … | 317 | 0 | 0.021 | 2.6 | … |
| … | 3 | 3 | 1 | 0 | … | 0 | 0 | 0 | 2.6 | … |
| … | … | … | … | … | … | … | … | … | … | … |

# Basic Analysis

## Final result interpretation for each presentation

Figure 2 and Figure 3 show the total number of final result for each presentation and proportion of the four different type of final results. Figure 2 and Figure 3 contains visualized information about 22 different presentations.

Module-presentation AAA 2014B and AAA2014J are the smallest modules with student count less than 400. CCC 2014B, FFF 2014B, BBB 2014B, FFF 2014J and BBB 2014J are module-presentations with over 2000 students. The average student for each presentation is about 1400.

The final results is divided into four types, including, distinction (“D”), fail (“F”), pass (“P”) and withdraw (“W”). Final result D is always accounts for the smallest proportion of students.

Figure 2: Total Number of Each Presentation and Final Results

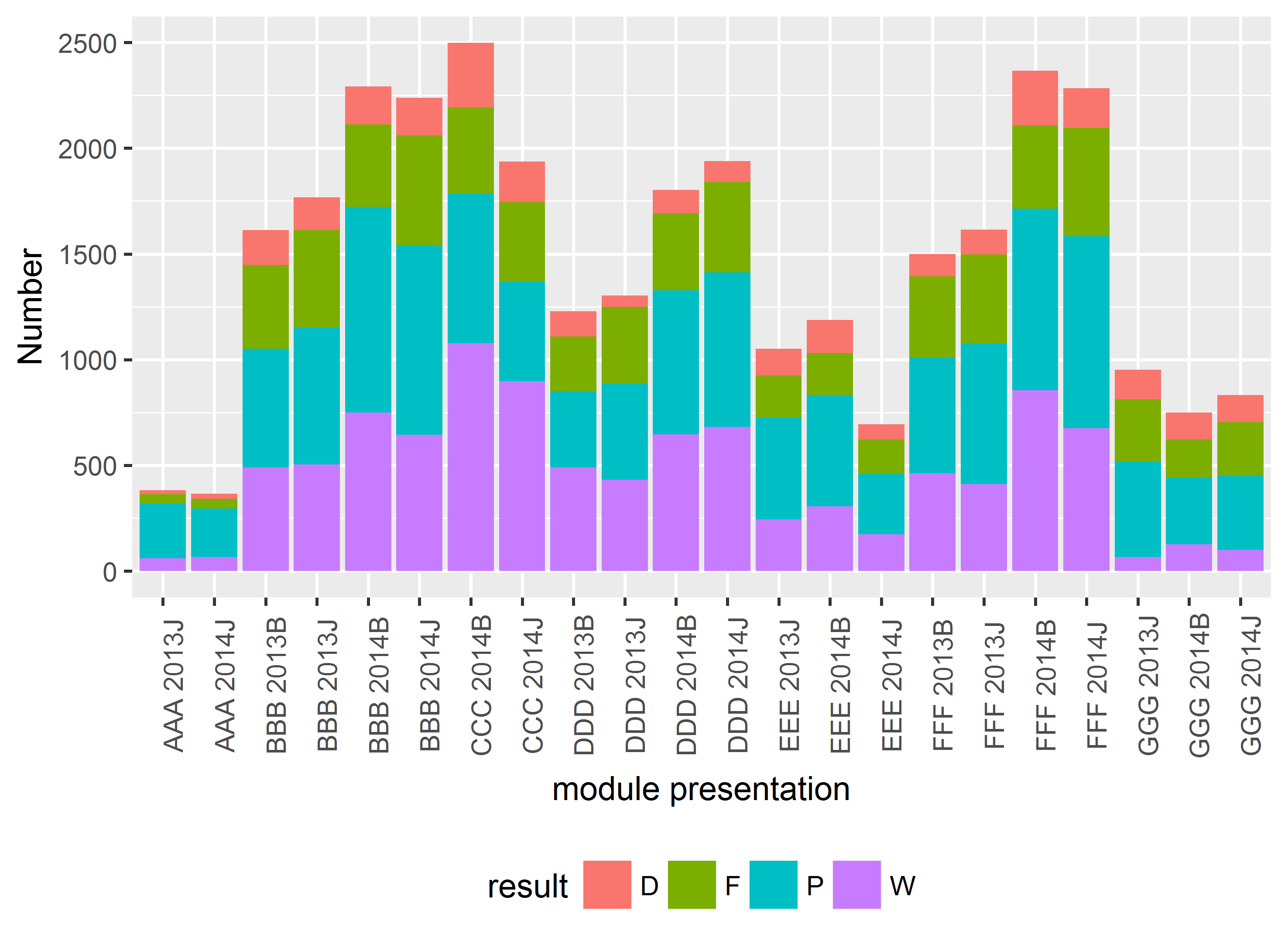
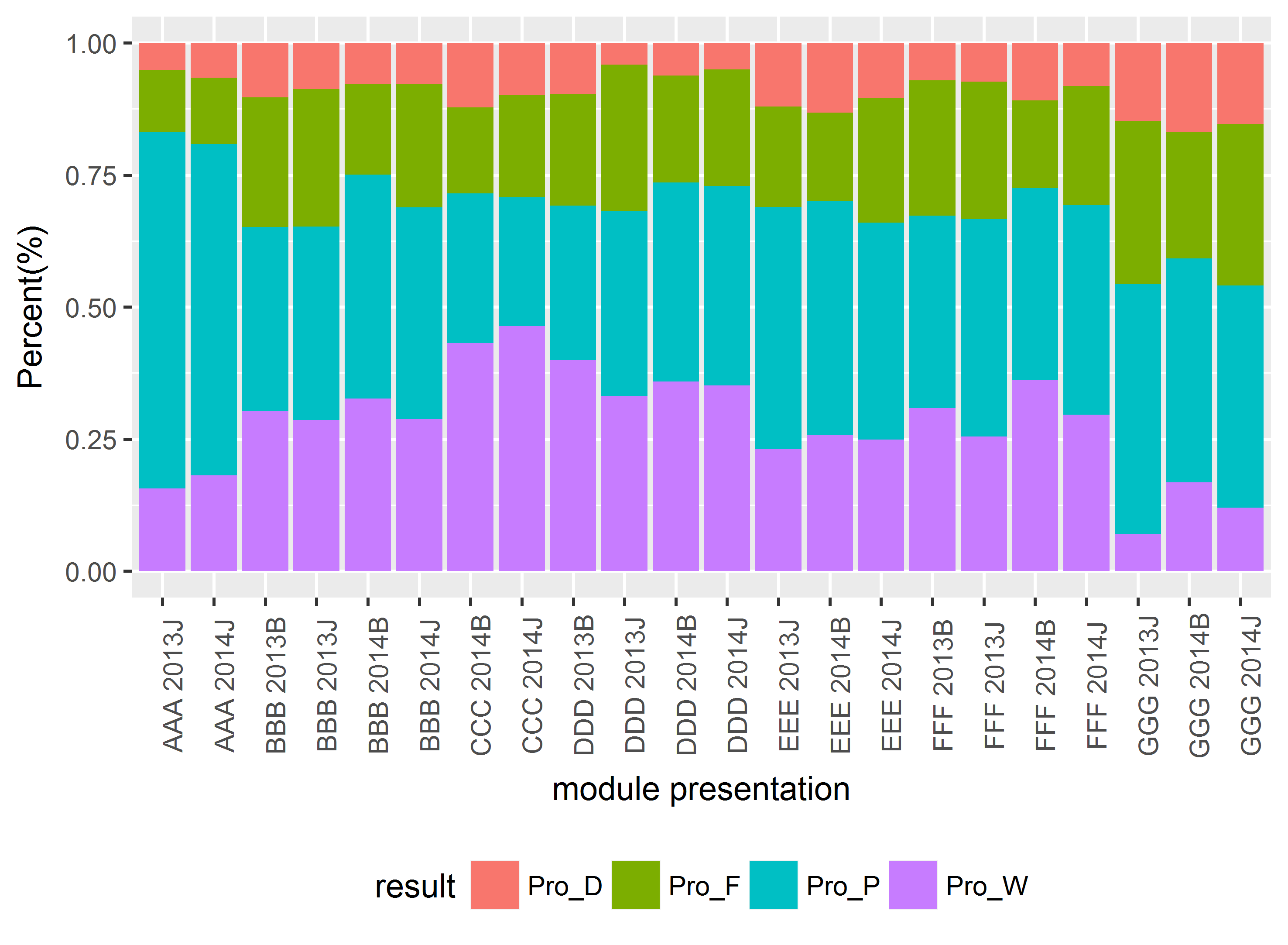


Figure 3: Proportion of Final Results per Presentation Module



Using this data, I was able to pick out presentations with highest proportion for each final results. To find the cause of different results per presentation, further analysis was done using the top 5 module-presentations with highest proportion for each final results.

Table 5: Top 5 presentation for each group

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Final Result | First | Second | Third | Fourth | Fifth |
| Distinction | GGG 2014B | GGG 2014J | GGG 2013J | EEE 2014B | CCC 2014B |
| Fail | GGG 2013J | GGG 2014J | DDD 2013J | FFF 2013 J | BBB 2013J |
| Pass | AAA 2013J | AAA 2014J | GGG 2013J | EEE 2013J | EEE 2014B |
| Withdrawal | CCC 2014J | CCC 2014B | DDD 2013B | FFF 2013 B | DDD 2014B |

## Correlation between courses and results

Please see below for variable explanations:

* Number of Assessment per type:
  + Tutor Marked Assessment (“TMA”);
  + Computer Marked Assessment (“CMA”); and
  + Exam.
* Total weight for each assessment type in each module-presentation
  + weighted\_TMA;
  + weighted\_CMA; and
  + weighted\_Exam.
* Proportion of each assessment type per module-presentation
  + Pro\_P;
  + Pro\_W;
  + Pro\_D; and
  + Pro\_F.

As shown in Figure 4 and 5, Exam, weighted\_CMA and weighted\_Exam are highly correlated with Pro\_W. This means that the number of exam in each module-presentation and weight of exam and weight of CMA all have an effect on proportion of withdrawal. As a result, we can conclude that the assessment type and number of exam will affect to the decision of withdrawal for students. Number of assessment for TMA has a negative correlation with the distinction student results. This means that students will have harder time getting distinction in a class with more TMA assessment type.

Figure 5 also shows some obvious characteristics such as the positive correlation between TMA and weighted\_TMA and Exam and weighted\_Exam and negative correlation between Pro\_W and Pro\_P. Although these obvious characteristics are not a ground breaking findings, they help show legitimacy of the dataset and analysis.

Figure 4: correlation plot

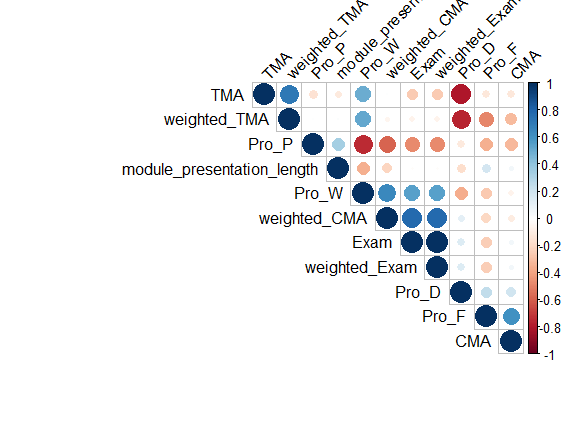
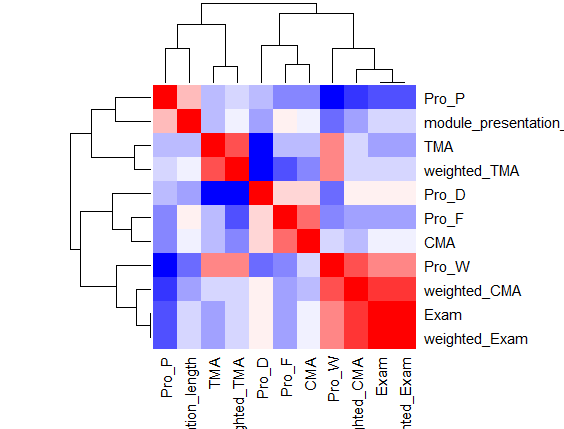


Figure 5: Heatmap plot



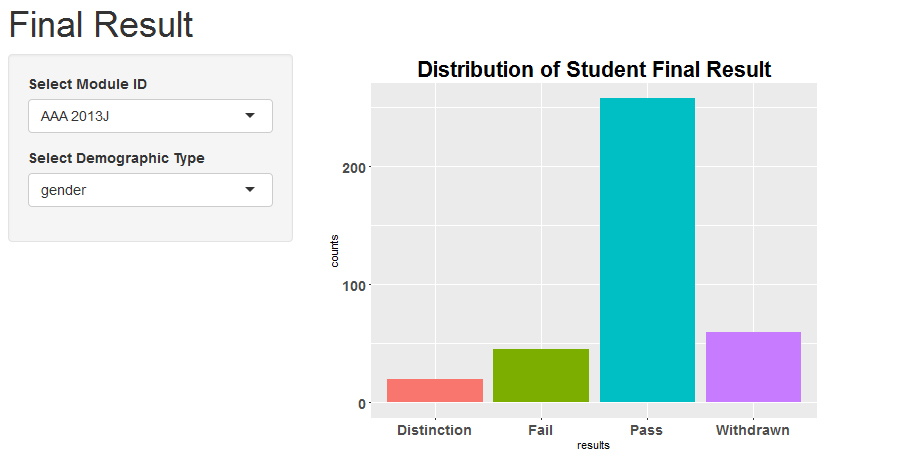
## Analysis between student demography and final result

In this session, Shiny tool was used to visualize the final result for each demography information in each presentation. The final product of the shiny tool with the link provided below, allows for users to select the Module ID[[2]](#footnote-3) and demographic type. The user is then able to see the final result of each module segmented out by the demographic type selected by the user.

### Plot 1 – student result per module

Plot 1 shows the total number of each final result for selected Module ID. User is to select the module ID from the drop down menu to automatically show the first bar graph in Shiny

Figure 6: Example of Student Result per Module



### Plot 2 – Final Result for each demographic group within a Module ID

Plot 2 shows the students’ final results for each Module ID segmented by demographic information selected by the user. On top of selecting the Module ID as discussed in Plot 1, the user is to select a demographic type such as gender, region, highest\_education, imd\_band, age\_band, and disability. This will generate second group of bar graphs that shows final results for each Module ID per demographic segmentation. The below graphs shows example of demographic segmentations for a Module ID.

Figure 7: Final Result for Highest Education Demographic Group within a Module ID

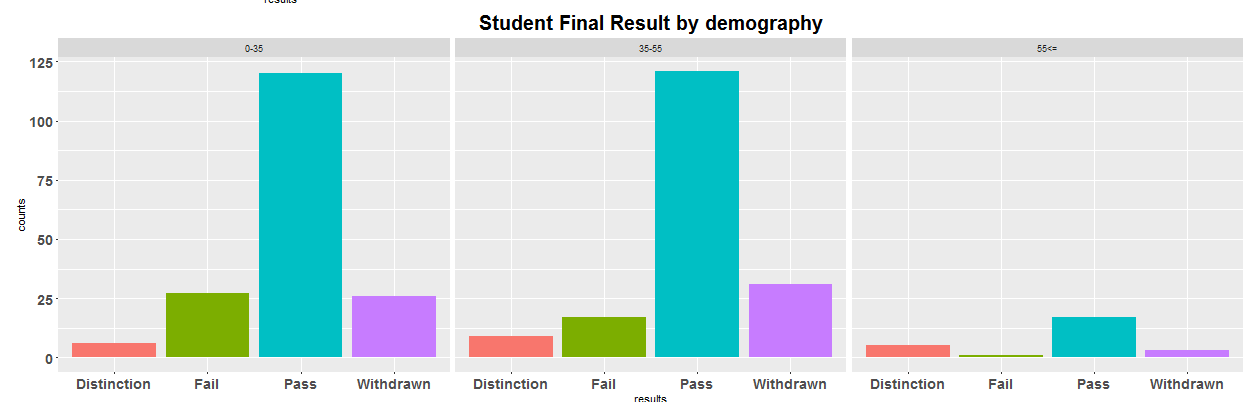


Figure 8: Final Result for Highest Education Demographic Group within a Module ID

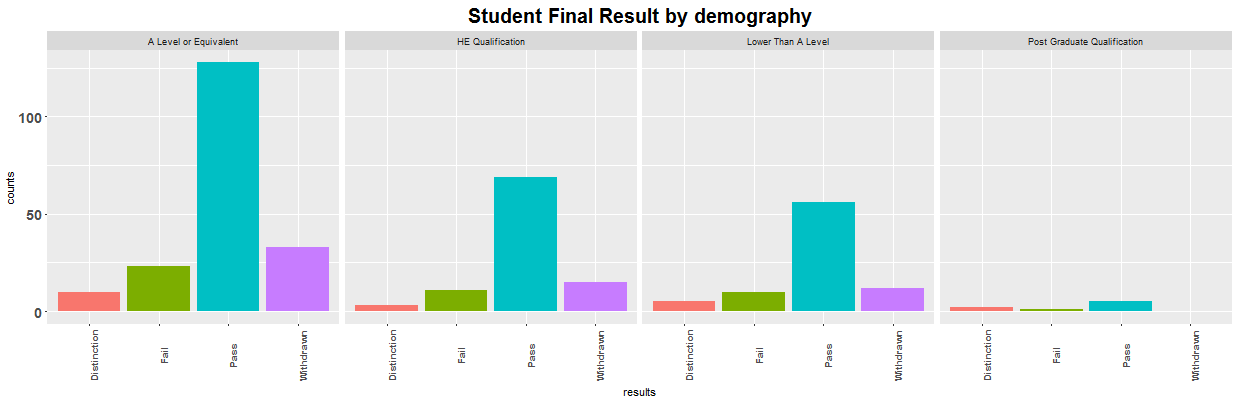
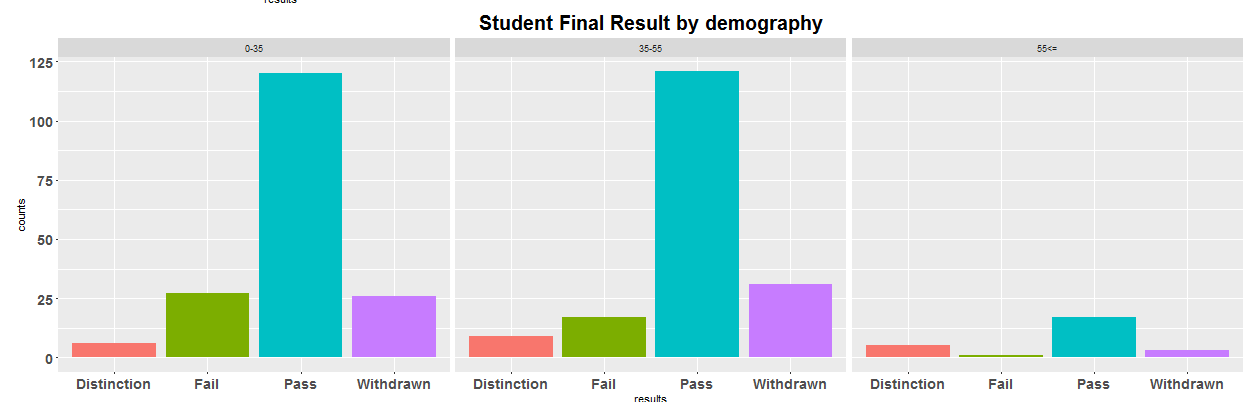


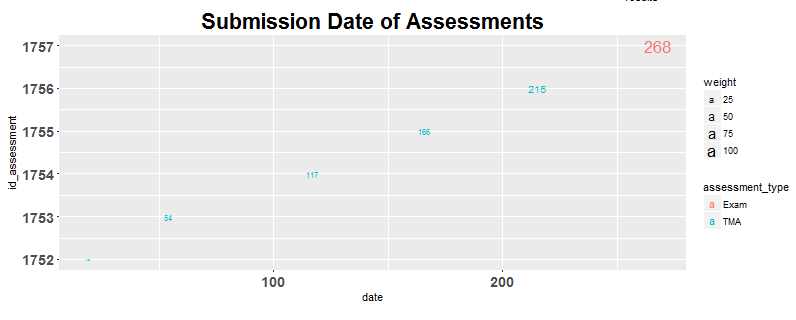
Figure 9: Final Result for Age Demographic Group within a Module ID



### Plot 3 – Assessment Type and Period Information for each presentation

Plot 3 provides information about the assessment within each module ID. This graph shows the user the unique ID number of the assessment, assessment type and assessment date. The size of dot also provides information about the weight of each assessments. The different color represents the different assessment type such as TMA, CMA, and Exam.

Figure 10: Assessment Type, Period, and Weight information for Assessments within a Module ID



With this information, the user is able to visually analyze the effect of demographics in each module ID and simultaneously look at the assessment characteristics in each module ID.

# Model Estimation/Development

The purpose of the model development is to find the significant predictors.

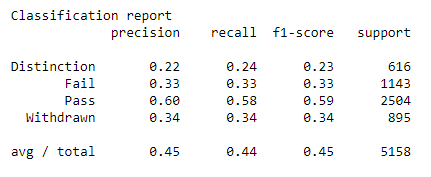
## Decision Tree

The dataset was split between the training and testing dataset; training accounted for 80% and testing accounted for 20%. Then I trained the decision tree using the training data with the entropy criterion.

### Model fit

I fitted the test dataset into the trained decision tree model, and attained the below result with 45% accuracy.

Figure 11: Initial Model Fit



### Feature Selection Model fit

To improve the model performance, I selected the top 30% of the most important feature using a chi2 test. The same training and test dataset was used. The top 30% translates to top four in the current dataset; these top four features were “studied\_credits”, “count\_date”, “sum\_click\_sum”, and “avg\_date\_submission”. The dataset consists of two different segment of variables: student demographic information and student activities. The result of the feature selection showed that the student’s activities during the presentation period had more effect on the final result. Please see below for the explanation of the top four features:

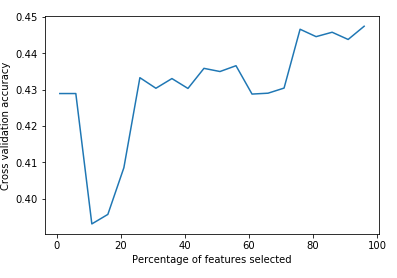
* studied\_credits: number of credits that a student is currently enrolled in
* count\_date: number of total days visits to VLE materials
* sum\_click-sum: number of total clicks to VLE materials
* avg\_date\_submission: average deviation date from the assessment due date.

This shows that the student activities have more significant effect on the final results. I fitted the decision tree classifier with new features, however the result of accuracy was at 0.46, which had no significant difference.

### Feature Selection improvement

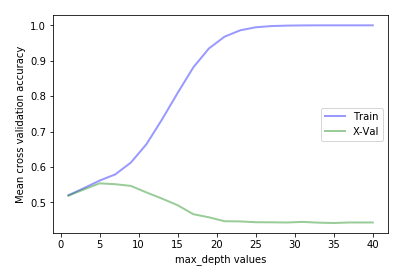
In order to perform the feature selection more systematically, I needed to find the best percentile using cross-validation. Below plot show the relationship between plot percentile of the features and cross-validation scores. The least crow validation accuracy is met when around 10% of features are selected. After 20% the accuracy stay in the stable stage.

Figure 12: Feature Selection Plot



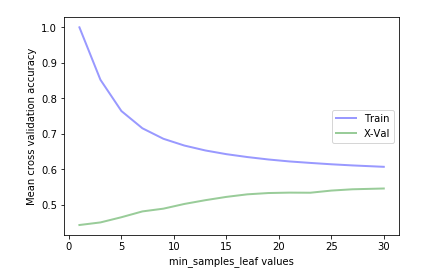
In order to increase the accuracy of the model an evenly spaced range of numbers in a specified interval was created. To create an evenly spaced range of numbers in a specified interval, I created the Kfold methodology and found the max\_depth value. I needed a more systematic way to explore the space of values for each parameter. The following is a general function that performs cross-validation using a range of values for a specified parameter of a model. The max\_depth of 5 was determined to create the best outcome. The larger values seem to lead to over-fitting. Using this finding, I fitted the model again and attained 56% accuracy.

Figure 13: Mean Cross Validation Accuracy vs. Max Depth Values



Another parameter of decision tree that's important is the min number of samples allowed at a leaf node. The below graph showed that the Train and X Variable Value starts to merge at min\_samples\_leaf of 11. Thus min\_samples\_leaf of 11 was selected. The accuracy was improved to 53%.

Figure 14: Mean Cross Validation Accuracy for each Samples Leaf Value



### Conclusion

The most important finding from the decision tree model was that the student activities were selected as important predictors for the final results. This shows that action of each student is more important than students’ characteristics. The parameter optimizations allowed for the accuracy to improve from 45% to 53%. However, the limitation of the decision tree is that I am not able to determine whether the variable is positively or negatively correlated with the student results. Thus, my next model is logistic regression which can fill the missing information from the decision tree.

## Logistic Regression

### Dataset Preprocessing

First step of the dataset preprocessing was to convert the region variable type from categorical to numerical. After this, the target variable of model was set as the students’ final result. I then converted the target variable type from categorical to binary; Pass/Distinction to “1” and Fail and Withdraw to “0”.

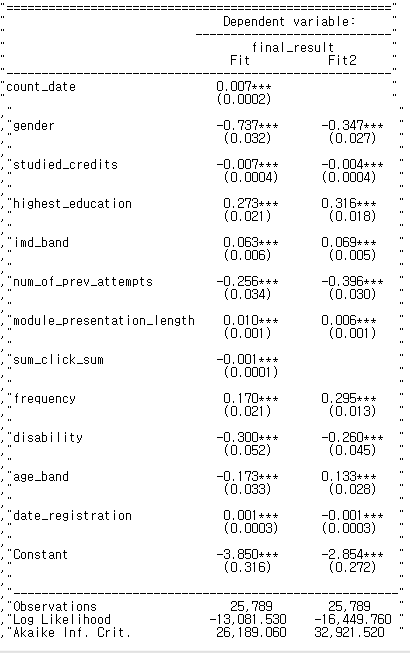
### Outline of Variable Selection Procedure

After the dataset preprocessing, I used the forward and stepwise variable selection process. Followings variables were selected as the predictors:

* count\_date;
* gender;
* studied\_credits;
* highest\_education;
* imd\_band;
* num\_of\_prev\_attempts;
* module\_presentation\_length;
* sum\_click\_sum;
* frequency;
* disability;
* age\_band; and
* date\_registration.

### Model fit

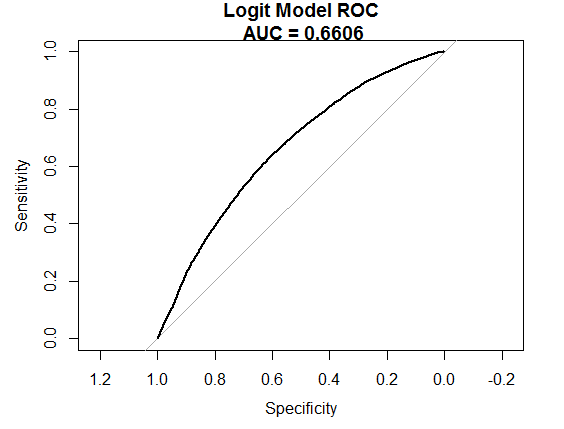
I fitted the logistic regression model with the selected variables. In logistic regression, avoiding excessive collinearity between the predictors is desired in order to prevent instability of the estimated coefficients. To test for the presence of severe multicollinearity in the final models, Variance Inflation Factor (VIF) statistics were assessed. The VIF measures how much the variance of the estimated regression coefficients are inflated in comparison to when the predictors are not linearly related. I used a criteria of VIF = 5. As the results of the VIF analysis, I excluded the two variable with highest value of multicollinearity: sum\_click\_sum (VIF = 11.41) and count\_date (VIF = 8.34).



### Model prediction

The test set was predicted with the fitted model. As a result, 66% accuracy was attained. This is a higher accuracy than the Decision Tree model.

Figure 15: Logistic Regression Model ROC



### Model outcome analysis

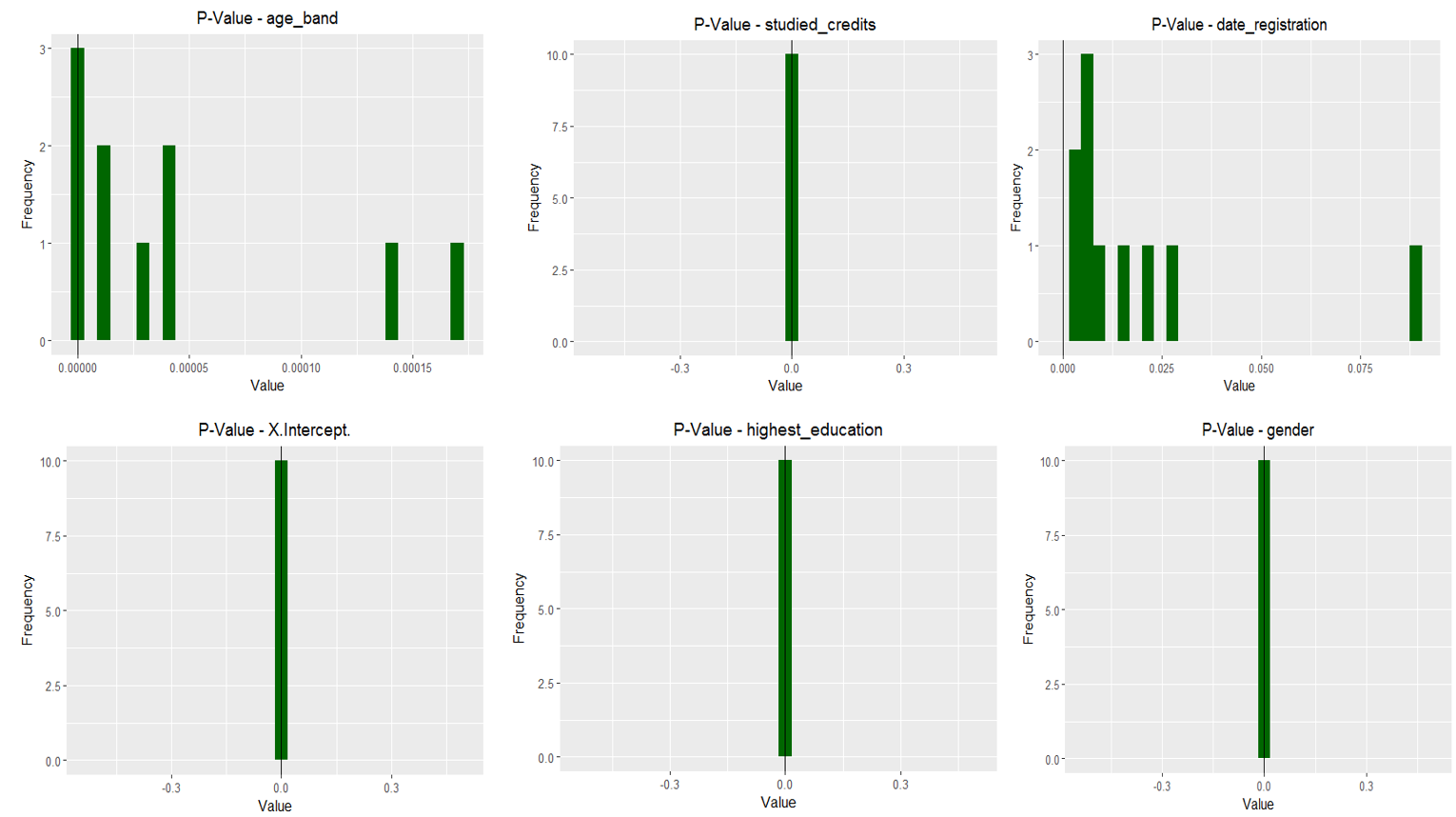
Highest education and frequency of access to the VLE have a positive impact on the final result.

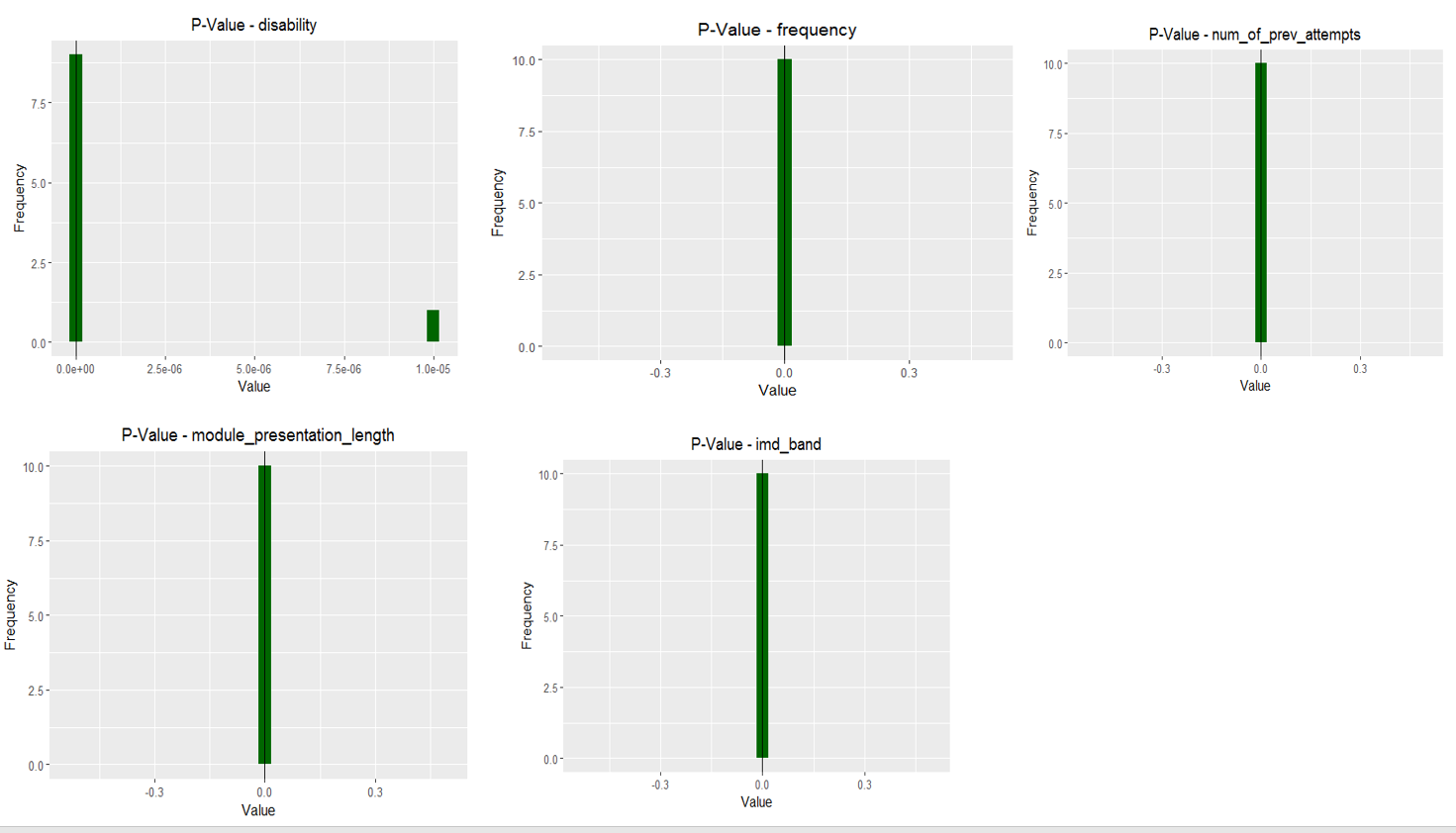
A student who is male or has a disability will get low final result than others. The number of previous attempts has negative impact on the final result.

### Manual Bootstrapping

Manual Bootstrapping was performed to check the coefficients stability and p values. Age band and date registration variables does not stable as other variables. It also supported by the model fit comparison table. Those variables’ signs are different than fit and fit2.

Figure 16: P-Value





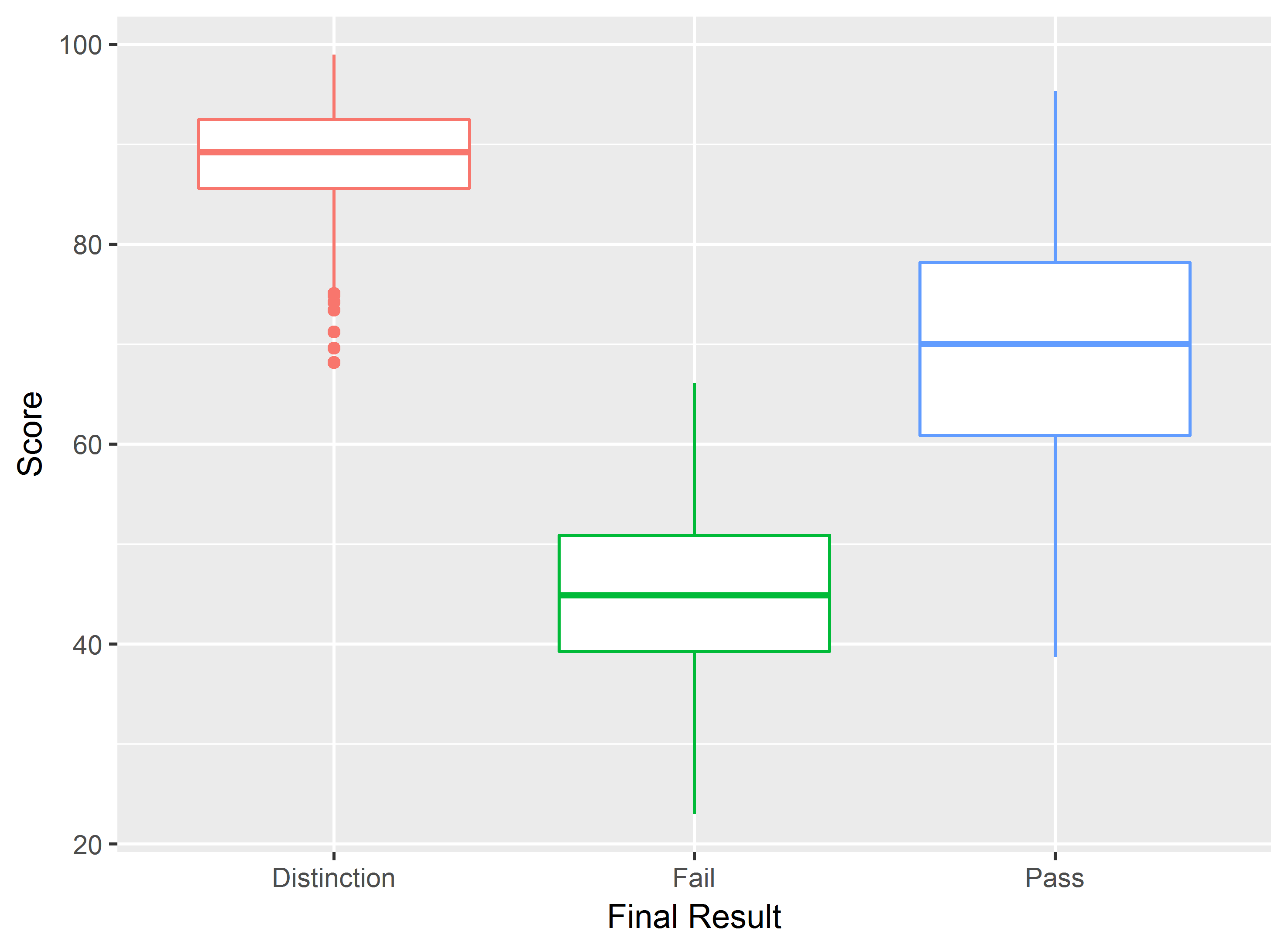
# Data Limitation and Potential Suggestions

## Data limitation

In the initial attempt to analyze the dataset, the assessment.csv, studentAssessment.csv and final result of each student taken from studentInfo.csv was used to find out the correlation between evaluation method and the final result of the students. In the process, I discovered that the most students did not have an exam score. In order to find out the proportion of the missing score, I calculated the total score base for each module-presentation. When I filtered out students who had less than total score base for the class, there were significant number of students who did not have a full total score even though they got pass or distinction as a final result. In fact, the exam scores for all students in 18 presentations out of 22 presentations were missing. In a class where the exam weight is between 33.3% and 50%, high missing value for exam score is a crucial data limitation. The total number of students with a full assessment score was only 19,033, which is about 10% of the total dataset. Due to the lack of data, I was not able to determine how the method of assessment influenced the final result.

In addition to this, the score ranges overlap for different score groups. This analysis can be better understood by looking at the boxplot below. The box plot represented that score range of each groups (distinction, pass, and fail). The below box plot was created using student scores manually calculated by multiplying the weight of each assessment to student’s score on each assessment for all students who were not missing any assessment scores. It makes sense that the median value of distinction is higher than others. However, the score range of failed group extends up to over 60 while the pass group extends to below 40. There is no clear metric differentiator between different score groups. To better understand the assessment structure and to create an accurate analysis, data description regarding the scoring must be explained in more detail and be able to explain the overlap of scores between different result groups. .

Figure 17: Box plot of Score Range for Final Result Type



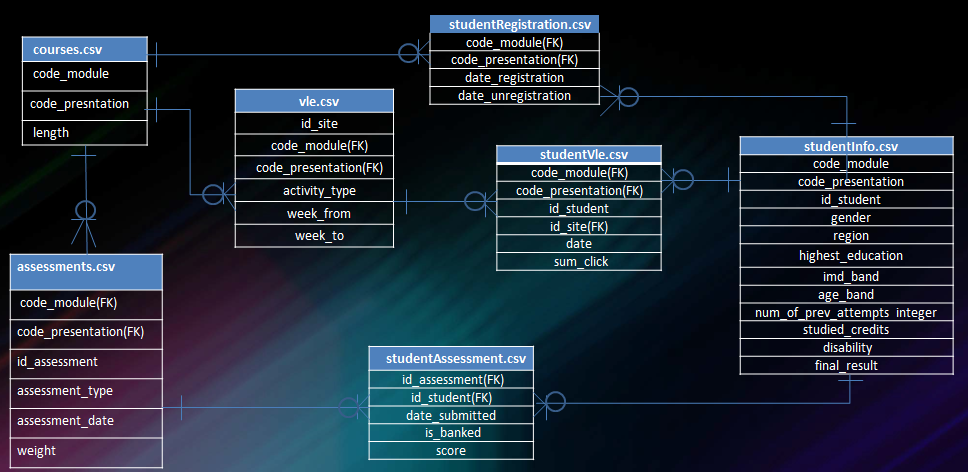
## Potential Suggestion

To overcome the data limitation, the data collection process should be updated as the below table. First, in ERD, master data is mainly code-like data, the courses.csv should consider as the top-level Master Data of ERD. However, it does not make sense that the relationship established between the courses table and the studentInfo table because they have exactly same key. This is because the code\_module and code\_presentation data values stored in the (studentRegistration table) can be joined by making it clear whether they are from the courses table or studentInfo. For example, in a given ERD, the studentInfo table and the studentRegistration table keys have the same configuration. It could be considered as a single table.

Secondly, studentAssessment table has an issue. Looking at the ERD posted on the website, the studentAssessment table is supposed to be related to the assessments table. Thus, the code\_module and code\_presentation columns that exist as keys in the assessments table must exist as foreign keys in the studentAssessment table, but this is missing. If these two columns, code\_module and code\_presentation, which exist as keys in the assessments table do not exist in the studentAssessment table, it could not be applied to join the function.

Finally, it is not a good idea to use “date” as the data column name because “date” is reserved word. Therefore, it is preferable to use assessment\_date for the date column existing in the assessments table, and convert the date present in the studentVle table to studentVle\_date.

Figure 18: Suggest ERD



1. Kuzilek, J. et al. Open University Learning Analytics dataset Sci. Data 4:170171 doi: 10.1038/sdata.2017.171 (2017). [↑](#footnote-ref-2)
2. Module ID is a combination of code module and code presentation [↑](#footnote-ref-3)