Open University Virtual Learning Environment Analysis

1. Introduction

The Open University (OU) has introduced a new virtual learning environment (VLE) that is designed to enhance the learning experience through online tools and resources. In this report we are going to answer questions about VLE effectiveness in improving students results and predicting students grades . This analysis involves data preprocessing, exploratory data analysis (EDA), feature engineering, model training, and evaluation.

Two primary **questions** in this report are:

(1) Is the VLE improving students' grades?
(2) Can we predict students' grades using data from the VLE?

To answer these questions, related data sets downloaded from OU website encompassing seven interconnected .CSV files (tables) and after preprocessing being analysed. These data sets include:

- courses.csv: That Lists all modules and their presentations, identified by code_module and code_presentation.
- assessments.csv: Provides details about assessments for each module-presentations, including assessment type (TMA, CMA, or Exam), submission dates, and weightings.
- vle.csv: Contains information about materials in the VLE, such as id_site, activity_type, and the weeks during which the materials are intended for use.
- studentInfo.csv: includes demographic information about students like gender, region, highest education level, and final results.
- studentAssessment.csv: Records student assessment scores and links students to specific assessments.
- studentVle.csv: Logs all student interactions with VLE materials, detailing the number of clicks per day.
- studentRegistration.csv: student registration statuses and dates.

The main indicator of interaction with VLE is the daily sum of clicks, after cleaning, wrangling and consolidating above tables first we performed a hypothesis test as follows:

Hypothesis: If there is a relationship between students grades and total sum of clicks then students with different interaction have different grades.

Then we develop a linear regression model to predict the students' grades. Because Different module presentation categories have different behaviours, in this report only the result of linear regression model related to module FFF, 2013 presentation mentioned. But with the code and guidelines provided in the notebook we can analyse and create different models associated with any module/presentation.

To merge the above-mentioned tables, we used key identifiers to connect tables to each other as follows:

Table studentInfo can be linked to studentAssessment, studentVle and studentRegistration tables using column id_student. Table courses links to the assessments, studentRegistration, vle and studentInfo using identifier columns code_module and code_presentation. Finally, assessments table links to studentAssessment using id_assessment and vle to studentVle using id_site. [1]

It is worth mentioning that an extensive data cleaning performed and related datasets all merged and most of the time in the research spend in this area. Then using EDA techniques and considering study questions, related insights about the students results, demographics and VLE interactions is investigated and visualized. Other statistical methods like pearsonR and k test also employed, to analyse VLE interaction data, module-specific patterns, and assessment types to extract meaningful conclusions. The main assumption in our analysis is that because, different OU module's structured to have assessment types including TMA, CMA and Exams, we only using TMA and CMA assessments information to calculate student grades (i.e., multiplying to its related assessment weights for each student). In modelling part, a Linear regression model developed, and it was showed that we can effectively predict students' grades.

---

2. Data Exploration

2.1. Data Cleaning and Wrangling

The dataset comprises 7 tables, including student demographics, module registrations, assessment scores, and VLE activity logs. It requires significant processing and transformation to extract features before building prediction models. Initial exploration by help of summary statistic functions revealed missing data, which were handled mainly by removal or substitution from other tables.

The bellow table shows the summary of structure of tables and their related Null values

Table	Row, Cols	Missing rows	Column names
assessments	206, 6	11	['code_module', 'code_presentation', 'id_assessment', 'assessment_type', 'date', 'weight']
courses	22, 3	0	['code_module', 'code_presentation', 'module_presentation_length']
studentAssessment	173912, 5	173	['id_assessment', 'id_student', 'date_submitted', 'is_banked', 'score']
studentInfo	32593, 12	1111(?) + 3516 (Oct._2020)	['code_module', 'code_presentation', 'id_student', 'gender', 'region', 'highest_education', 'imd_band', 'age_band', 'num_of_prev_attempts', 'studied_credits', 'disability', 'final_result']
studentRegistration	32593, 5	22521	['code_module', 'code_presentation', 'id_student', 'date_registration', 'date_unregistration']
studentVLE	10655280, 6	0	['code_module', 'code_presentation', 'id_student', 'id_site', 'date', 'sum_click']
vle	6364, 6	5243	['id_site', 'code_module', 'code_presentation', 'activity_type', 'week_from', 'week_to']

Consideration for data cleaning and wrangling are as follows:

a. student_assessment dataset:

- The score column in the studentAssessment table contains some items indicated as **'?'** , which amount to 173 rows. There is only a few of these NaNs across this feature (173 compared to 173912 rows), therefore for calculating score for score analysis I decided to remove these NaNs. And change their Dtype to float
- Removing rows with their score flag being 'is_banked,

b. Assessment dataset

- Handling NaNs related to 'date's of Exams (with "module_presentation_length"), 11 dates substituted.
- Merging assessments table and student_assessment (assessment_type = 'Exam' is problematic)
- There are some missing keys in assessments table and student_assessment (Missing keys: [1757, 1763, 14990, 15002, 15014, 15025, 40087, 40088, 30713, 30718, 30723, 34872, 34885, 34898, 34911, 37424, 37434, 37444]) and these missing keys are related to assessment_type ['Exam']
- Creating columns for different assessment types scores and their related weighted scores.
- Removing 'Exam' assessments, (only considering TMA and CMA assessments for calculating final score)

  Based on descriptions on Open university website about weight of the assessment in % in the dataset it is mentioned that "Typically, Exams are treated separately and have the weight 100%; the sum of all other assessments is 100%. so, if I multiply each score with related weight and sum them up for 'CCC', 'DDD' modules I will get sum

of scores for each students that is calculated from 200. (100 for sum (TMA and CMA) and 100 for Exam). Also, in previous sections it was shown that some of the keys in assessment table are missing in student_assessment table. Therefore, only weighted score out of TMA and CMA scores was considered in this report. [3]
Also in the EDA section it was detected that no weights (neither Exam nor TMA and CMA) for module GGG was recorded.

- Creating a new feature related to timeliness in submitting assignments (avg_timliness)
  date column in assessments.csv gives information about the final submission date of the assessment calculated as the number of days since the start of the module-presentation. The starting date of the presentation has number 0 (zero).
  date_submitted column in studentAssessment.csv is the date of student submission, measured as the number of days since the start of the module presentation.
  now we create a new column named submission_timeliness that if it is positive means student submitted assignment without delays, and if negative shows delays for submission

c. The vle dataset

because we have 'code_module','code_presentation' in student_vle table I removed these two column from vle as they are redundant
week_from and week_to are almost empty columns, 5243 out of 6364 are nulls. I assumed these two features are not important for my analysis, so I removed week_from and week_to columns
Duplicated base on id_sites checked to make sure it acts as a key perfectly, and there was no duplicates.

d. The student_vle dataset

- merging vle and student_vle
- Create new feature: avearge weekly clicks and total sum_clicks for each student/module/presentation/activity_type
- using features that are time_series not used in the analysis. So, while grouping based on id_student/code_module/code_presentation/activity_type to group based on sum or mean date column lost.
- The date column in vle_merged shows days since the start of the course. Dividing date (days from registration) by 7 (date // 7) groups the data into weeks.
  Assumption: The date // 7 shows that day 0 is the first recorded activity and is part of the first week of the course. This is a simple assumption to reach the week grouping process and doesn't necessarily align with the usual calendar week that starts from Monday to Sunday. Then sum_click by week gives the total interaction for each week for calculating the mean (sum_click by week and id_student gives per-student weekly averages)
- Creating separate column for sum of clicks for different categories of activity_type
- instead of using
  vle_merged=pd.get_dummies(vle_merged,columns=["activity_type"]) code we first creat activity_columns then concatenate them to merged table, because we need activity_columns names to iterate over and put related sum_click values in the True cells
- Merging assessment_merged and vle_merged_df tables

e. The courses datset

- Adding content of courses dataset to merged_df (our first consolidated merged table from 4 vle, student_vle, assessments, and student_assessment)

f. The student_info dataset

- Removing sensitive features from student_info table
- To make the analysis simpler and unbiased, gender, region, imd_band and disability features are NOT used in the analysis. Instead, analysis focus is on academic and interaction data.
- One-Hot coding for highest_eduacation and age_band
- Merging the student_info_c to merged_df dataset
- Removing rows with final result indicated as 'Withdrawn':
  Because final score was calculated based on assignment scores (Exam scores ignored for the reasons mentioned earlier) for DDD and CCC modules, as a result although the student final_result is withdrawn but we still have a calculated final_score for these modules. These students with withdrawn status, quitted the modules at different dates during the semester(presentation), so their final_score has no value for our modeling. Therefore, rows with final_result being showed as Withdrawn removed. size before removing Withdrawal was (25521, 39) and after removing Withdrawal we remained with 20964 rows

g. The student_registration dataset

Because the information in this table may not add value to our analysis it has been decided not to include it.

h. Final all merged dataset

Includes 20964 entries, total 39 columns, and no nulls

i. Removing repeatactivity_sum and sharedsubpage_sum features

- Only 2 records were associated with repeatactivity_clicks with std=0
- sharedsubpage_sum with mean of almost zero and std close to zero was not considered in model

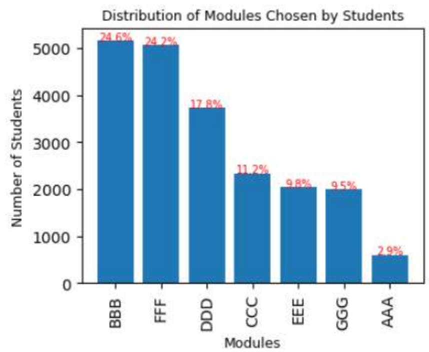
2.2. Summary of the Dataset

The final data set created by merging 6 different tables and an extensive data cleaning and wrangling includes 23415 rows, that shows number of students being nalysed, and it contains 37 columns as follows:

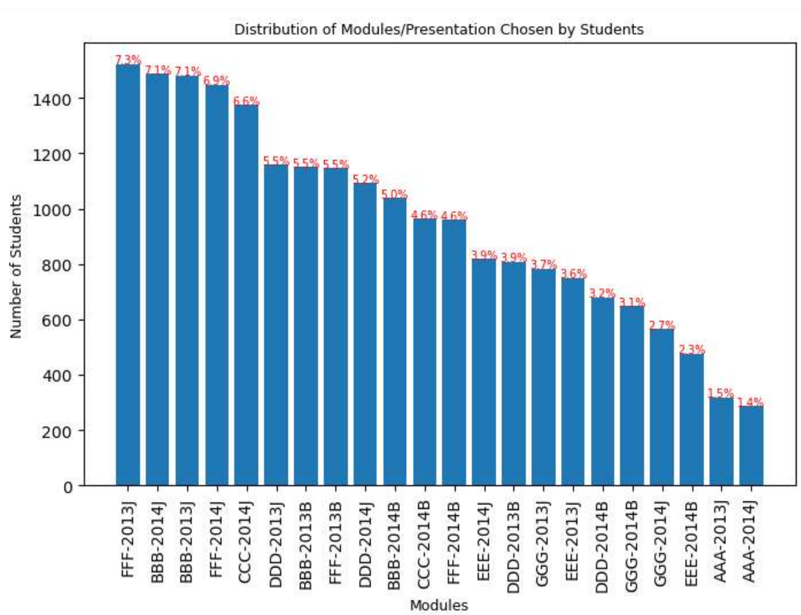
['code_module', 'code_presentation', 'id_student', 'final_score', 'avg_timeliness', 'total_clicks', 'avg_week_clicks', 'dataplus_clicks', 'dualpane_clicks', 'externalquiz_clicks', 'folder_clicks', 'forumng_clicks', 'glossary_clicks', 'homepage_clicks', 'htmlactivity_clicks', 'oucollaborate_clicks', 'oucontent_clicks', 'ouelluminate_clicks', 'ouwiki_clicks', 'page_clicks', 'questionnaire_clicks', 'quiz_clicks', 'resource_clicks', 'subpage_clicks', 'url_clicks', 'module_presentation_length', 'num_of_prev_attempts', 'studied_credits', 'final_result', 'A_Level', 'HE_Qual', 'Lower_A_level', 'No_formal', 'Post_Graduate', 'age_band_0-35', 'age_band_35-55', 'score_segment']

2.3. Important aspects of the dataset and features

There are 7 unique modules in the data set (i.e ['AAA' 'BBB' 'CCC' 'DDD' 'EEE' 'FFF' 'GGG']) and 4 unique presentations including ['2013J' '2014J' '2013B' '2014B']. However, the Number of unique module+presentations are 22.



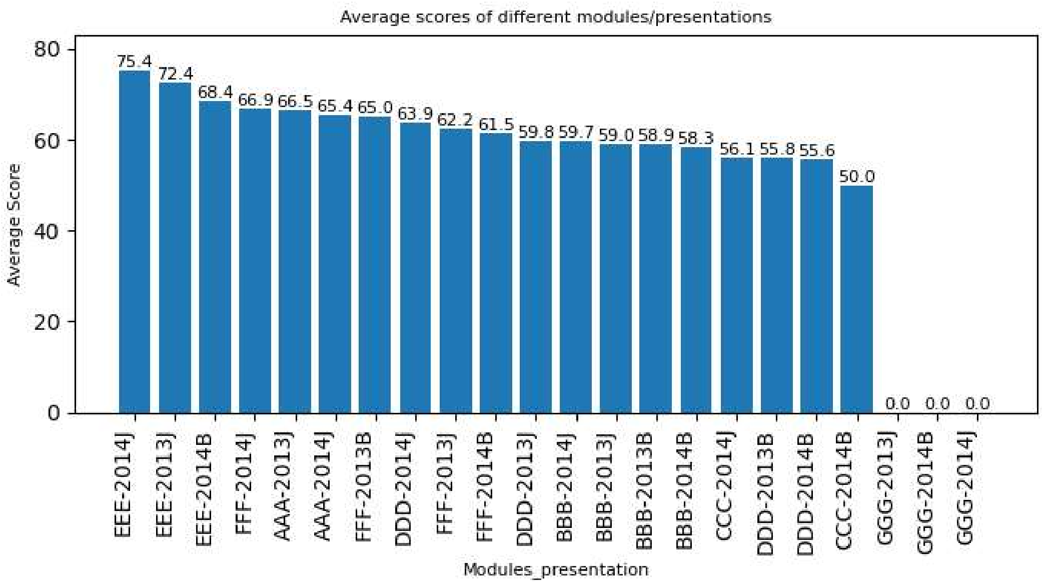
The distribution of Modules Chosen by Students Considering the time it has been offered (presented) is shown in following bar chart:



2.3.1 Top 5 Modules Chosen by Students:

	code_module_presentation	students_counts	percent
0	FFF-2013J	1522	7.260065
1	BBB-2014J	1487	7.093112
2	BBB-2013J	1481	7.064492
3	FFF-2014J	1448	6.907079
4	CCC-2014J	1375	6.558863

In the following bar chart, the Distribution of modules average score is shown:



Above bar chart indicates that the average scores for module GGG are zero. In data cleaning and wrangling section, it was described that for calculation consistency we did not consider exam type assessments for calculating final score. Also, it was detected that there are missing keys in student_ssessment dataset that were related to Exam assessments. Further investigation in the assessment table on module GGG shows that TMA and CMA weights are zero, however Exam weights are 100%. However, id_assessments keys related to these Exam weights are missing in the studenet_assessment table.

2.3.2 Top 5 Modules with the Highest and Lowest Average Scores:

Top 5 modules with the highest average score are:

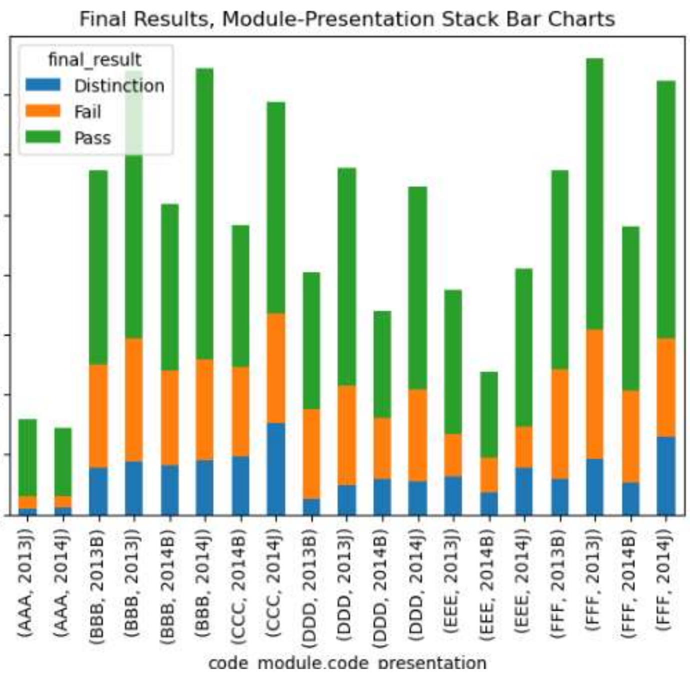
	code_module_presentation	Average Score
0	EEE-2014J	75.364683
1	EEE-2013J	72.392277
2	EEE-2014B	68.446303
3	FFF-2014J	66.861188
4	AAA-2013J	66.520376

Top 5 modules with the lowest average score are:

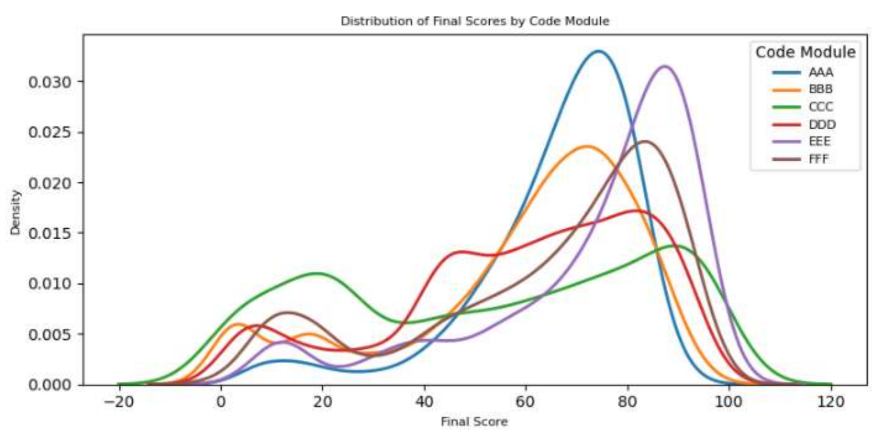
	code_module_presentation	Average Score
0	CCC-2014B	49.963575
1	DDD-2014B	55.611728
2	DDD-2013B	55.803156
3	CCC-2014J	56.131164
4	BBB-2014B	58.285472

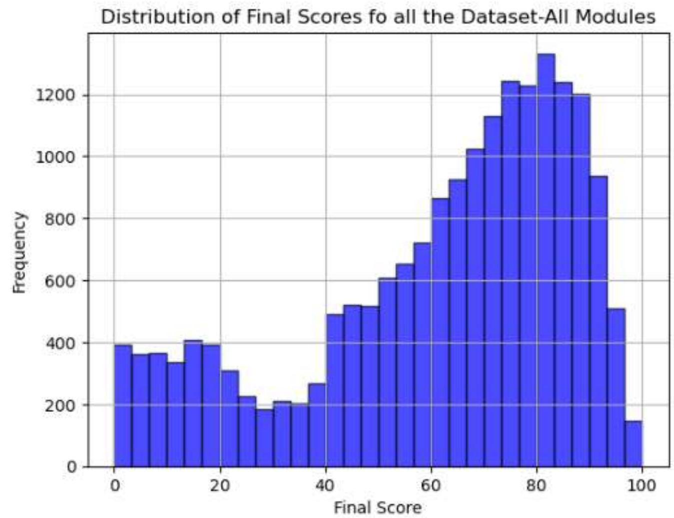
2.3.3 Distribution of final results:

Distribution of final results that are shown as Pass, Fail, or distinction is shown in the following diagram. It should be mentioned that final results that are indicated as Withdrawn are removed from the dataset because they considered irrelevant to our analysis.



And the following distribution plots shows the distribution of scores that was calculated based on assessments and related weights for each students.



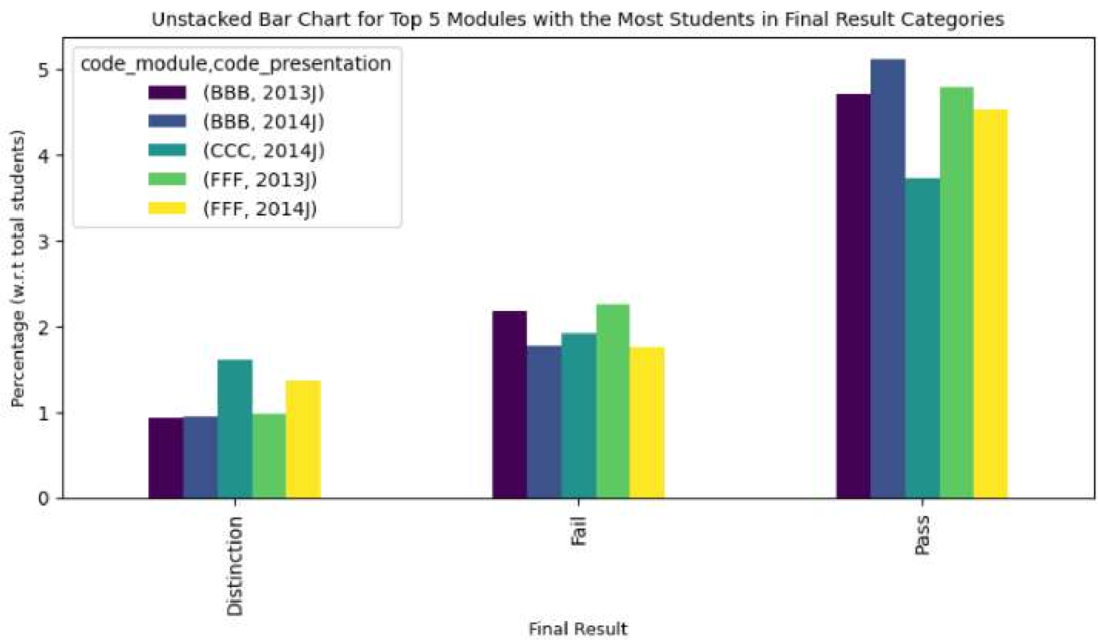


Performing Kolmogorov-Smirnov test of normality on final scores data shows a Test Statistic: 0.98 and P-value: 0.0 therefore we reject the null hypothesis: The data is not normally distributed.

2.3.4 The top 5 modules with the greatest number of fails

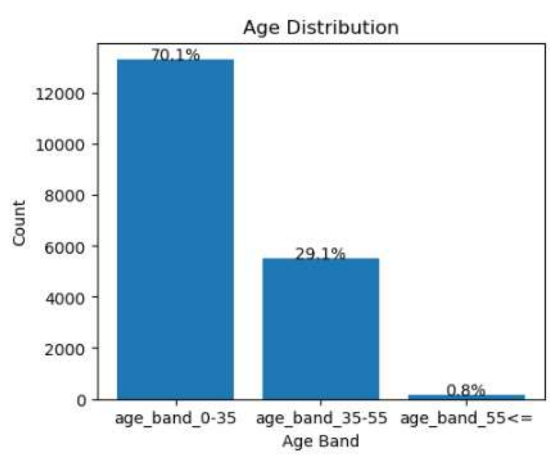
```
The top 5 modules with the most number of fails are:
```

	code_module_presentation	final_result	student_counts
0	FFF-2013J	Fail	428
1	BBB-2013J	Fail	411
2	FFF-2013B	Fail	366
3	CCC-2014J	Fail	363
4	BBB-2013B	Fail	347



2.3.5 Age Distribution of Students:

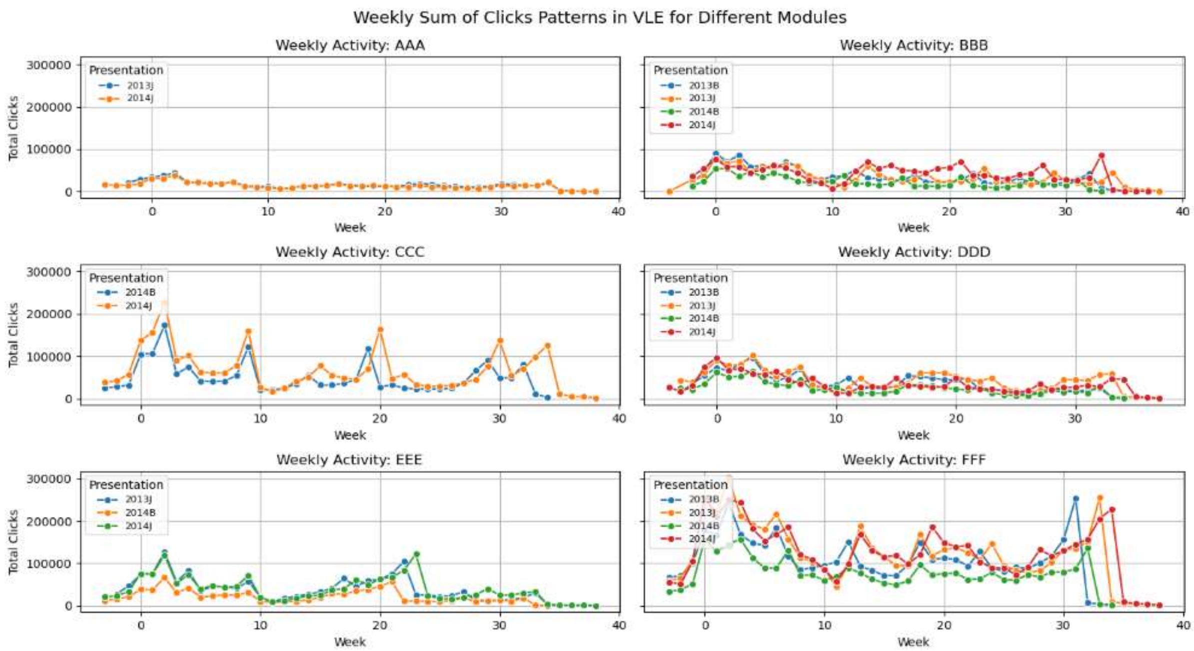
- The age distribution revealed a predominantly adult learner base, with peaks at the 0-35 age band (70%) and 29% for 35_55 band , highlighting the university's reach among working professionals and non-traditional students. There is small portion for age greater than 55 (<1%).



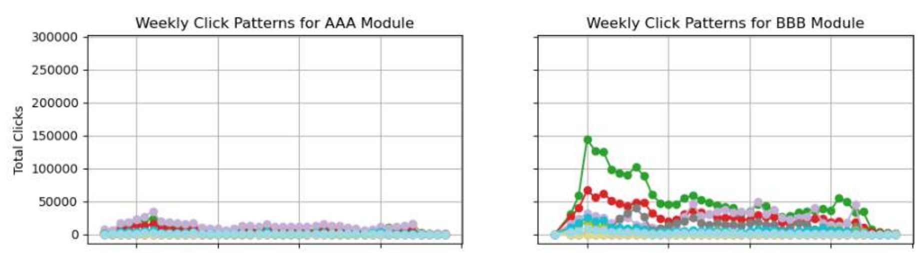
To make our modelling simpler the age band greater than 55 was removed from the dataset.

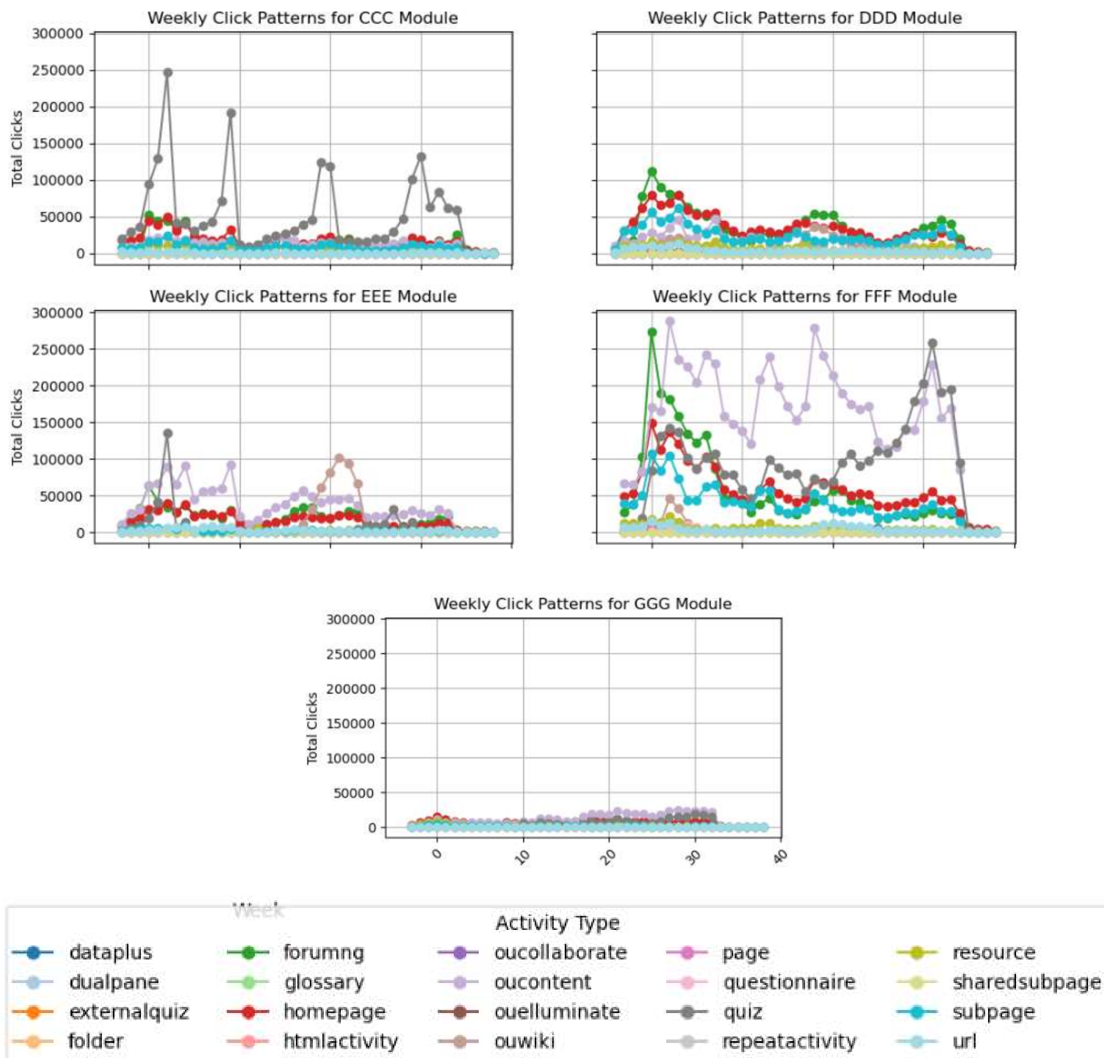
2.3.6 Weekly Activity Patterns in the VLE:

- Students displayed consistent engagement during weekdays, with peaks around assignment deadlines, emphasizing the VLE's role in structured learning.

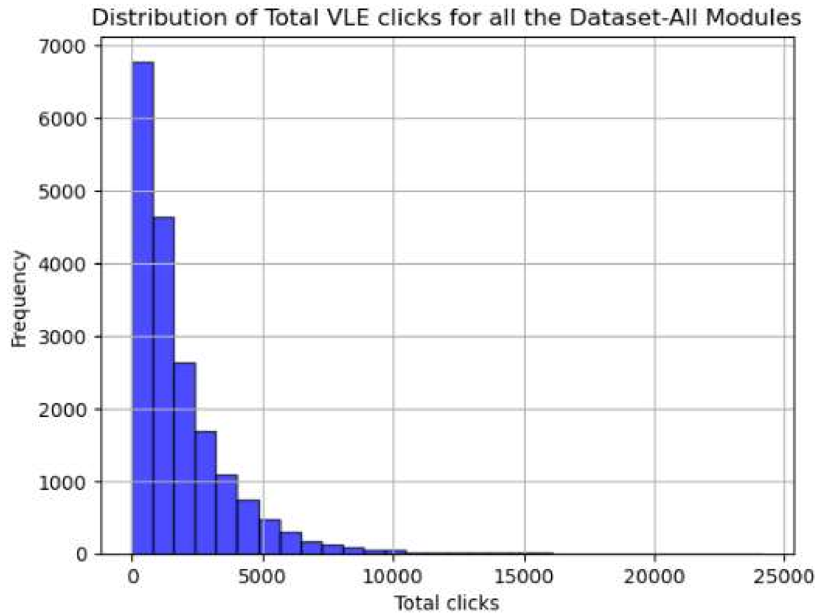


Also weekly click patterns for different modules explored and is shown in following charts:

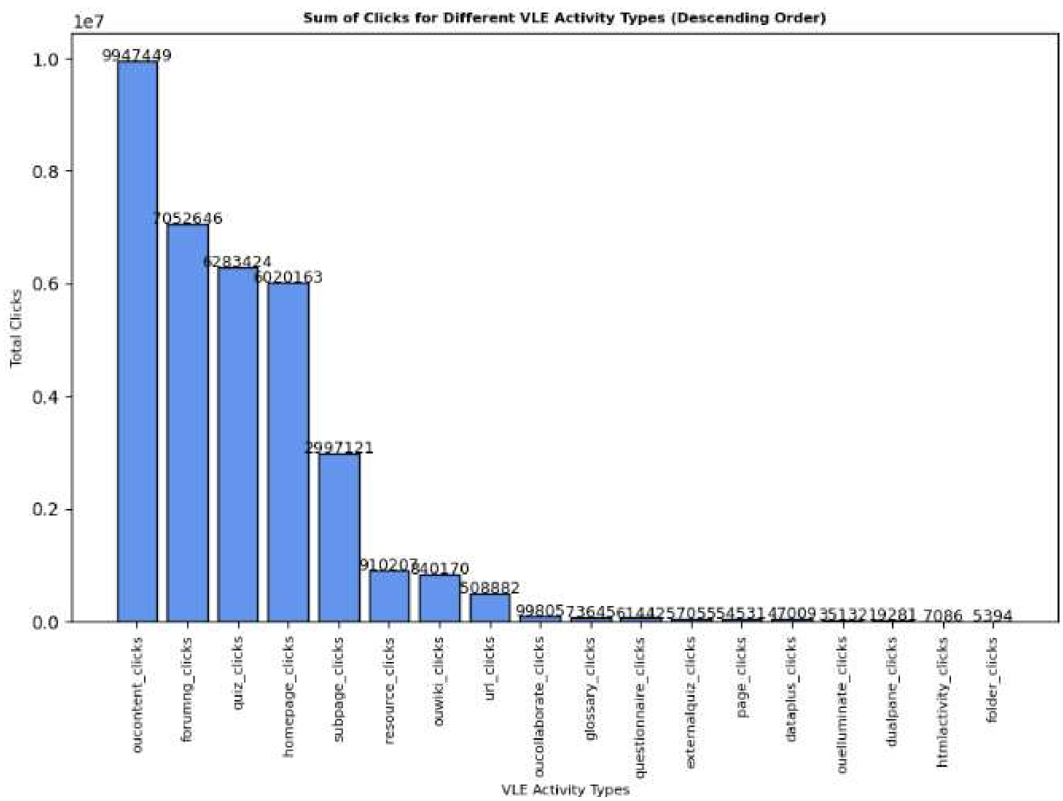



Weekly Click Patterns for CCC Module
Weekly Click Patterns for DDD Module
Weekly Click Patterns for EEE Module
Weekly Click Patterns for FFF Module
Weekly Click Patterns for GGG Module
Total Clicks
300000
250000
200000
150000
100000
50000
0
0
10
20
30
40
Week
Activity Type
dataplus
dualpane
externalquiz
folder
forumng
glossary
homepage
htmlactivity
oucollaborate
oucontent
ouelluminate
ouwiki
page
questionnaire
quiz
repeatactivity
resource
sharedsubpage
subpage
url

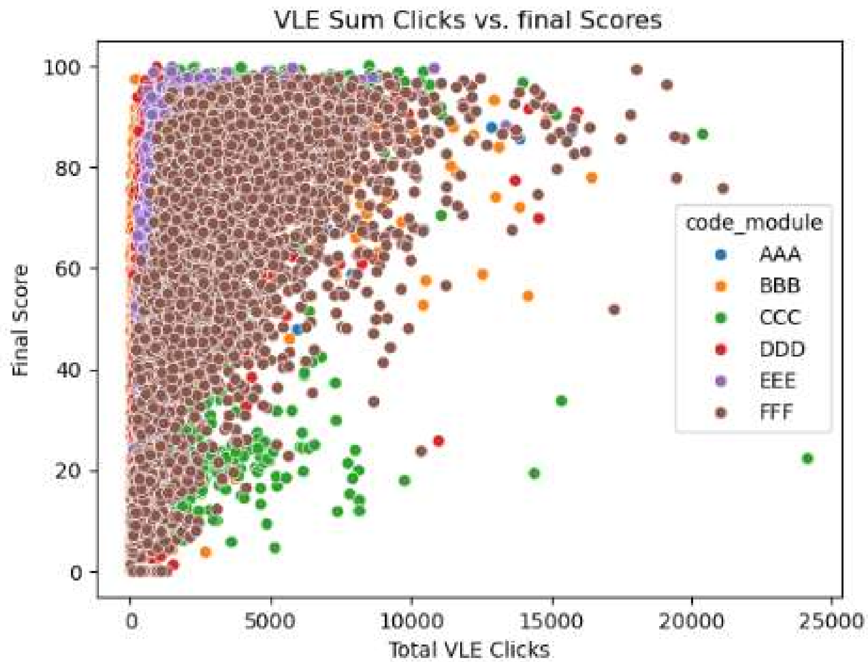
2.3.7 Distribution of total clicks



2.3.8 Distribution of sum of clicks for VLE activities

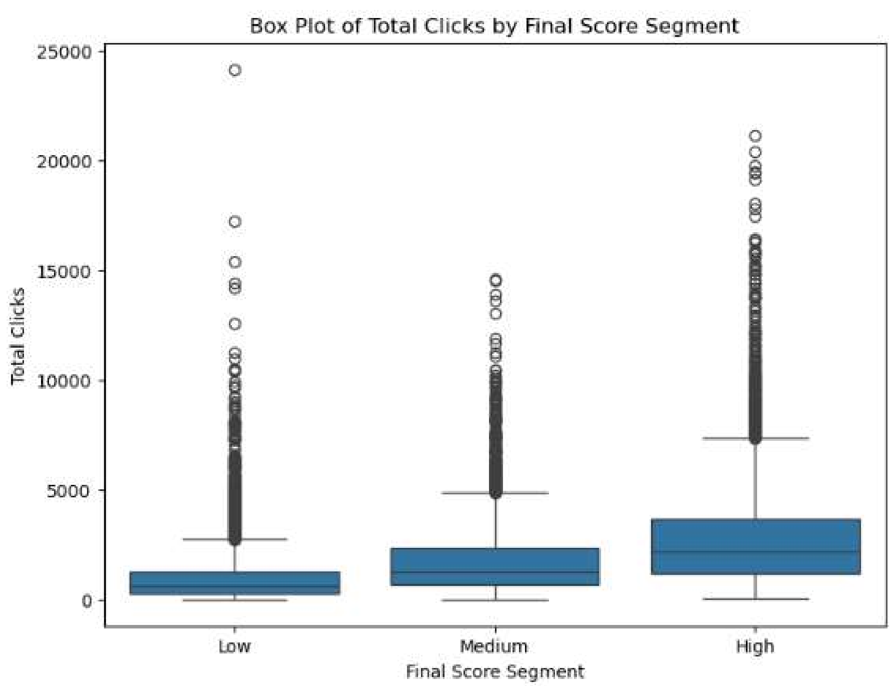


2.3.9 Scatter plot Clicks vs Scores on all the data



2.3.10 Scatter Plot for module FFF_2013J (we use this data set for modelling)

2. Box plot of different categories of scores versus their associated total clicks. Dividing calculated final scores into 3 bins if between 0-60 it labelled as low, if between 60 to 75 as medium and else high we draw this box plot that was the basis for our hypothesis testing.



3. Methods

Linear regression was used to model the relationship between students' grades and their VLE engagement metrics. Feature selection techniques, including Recursive Feature Elimination (RFE) and Sequential Feature Selection, were applied to identify the most predictive variables. Also to answer the first research question a hypothesis tests also performed.

3.1. Hypothesis Testing:

Our main hypothesis was that **If there is a relationship between students grades and total sum of clicks then students with different interaction have different grades**.

Null hypothesis (H0): There is no significant difference in total_clicks between the different final_score segments.
Alternative hypothesis (H1): There is a significant difference in total_clicks between the different final_score segments.

An ANOVA test was conducted to examine whether there is a significant difference in total_clicks across final_score segments (Low, Medium, High). The test yielded an F-statistic of **1696.94** and a p-value of **0.0**. Since the p-value is less than 0.05, we reject the null hypothesis, indicating that there is a statistically significant difference in total_clicks among the score segments. This result suggests that students in different final_score categories (Low, Medium, High) have distinct patterns of total engagement as measured by clicks.

3.2. Predictive Modelling

To be able to perform a linear regression modelling some criteria's must be met especially the normality of the features, or at least their symmetry.

3.2.1. Skewness of features

At first it was decided to use the main consolidated dataset for modelling by linear regression, but checking for the skewness of features, showed that skew values of the features range widely and some of them are highly skewed (greater than 1000) . This can affect the performance of linear regression. Linear regression assume that predictors have normal distribution or at least are symmetrically distributed. But, after checking the skewness of a subset of dataset (i.e., filtering for only module FFF, 2013), it was showed less skewness but still needed some transformation.

- **Feature Engineering:**
    - Log transformations and scaling were applied to skewed variables to enhance model accuracy.

    Transformations on Skewed Predictors was applied using Log transform and Yeo-Johnson transformation for columns with zeros or negative values (as we have zero values in the dataset). after above transformation the kurtosis also checked.

homepage_clicks, ouwiki_clicks, oucollaborate_clicks, num_of_prev_attempts, No_formal, and Post_Graduate, url_clicks, studied_credits, and quiz_clicks had Leptokurtic behaviour (High Kurtosis > 3). But A_Level, Lower_A_level, age_band_0-35, and age_band_35-55 features were Platykurtic (Negative Kurtosis < 0).

3.2.2. Investigate multicollinearity among features:

- Multicollinearity among predictors was assessed through personr with PermutationMethod at 0.05 significance level. A loop was coded to iterate over all the feature pairs and check for their correlation. The results of this method indicate that for the specified feature pairs, the null hypothesis(no correlation) could not be rejected. This means there is insufficient evidence to suggest these feature pairs are significantly correlated at 0.05 significance level. After extracting a list of pairs of highly correlated features, it was decided which feature to remove. In our subset of dataset FFF_2013J, it is decided to remove these features: glossary_clicks, No_formal and Post_Graduate.

3.2.3. Feature selection

For all possible number of features in our dataset of 32 features, the Score and features chosen using sequential Feature Selector (Forward).

3.2.4. Testing assumptions

- The residuals of the regression model were tested for normality using the Lilliefors.

4. Results

Performing SequentialFeatureSelector algorithm for different tolerance (i.e, 0.1, 0.01 and 0.001) selected the following features for our dataset of FFF,2013J.

```
Tolerance: 0.1
Features chosen: ['total_clicks']
Score: 0.3427041110405078
```

```
Tolerance: 0.01
Features chosen: ['total_clicks' 'folder_clicks' 'forumng_clicks' 'oucollaborate_clicks'
 'oucontent_clicks' 'quiz_clicks' 'resource_clicks' 'subpage_clicks'
 'Lower_A_level']
Score: 0.5474009067280481
```

```
Tolerance: 0.001
Features chosen: ['avg_timeliness' 'total_clicks' 'avg_week_clicks' 'folder_clicks'
 'forumng_clicks' 'homepage_clicks' 'oucollaborate_clicks'
 'oucontent_clicks' 'page_clicks' 'questionnaire_clicks' 'quiz_clicks'
 'resource_clicks' 'subpage_clicks' 'num_of_prev_attempts' 'Lower_A_level']
Score: 0.5779903266246215
```

Interpretation of these features shows that, from tol=0.1 to tol=0.01, the R^2 score improves significantly from 34% to 0.54% and from tol=0.01 to tol=0.001, the improvement is less than before(0.54% to 57%), and model selects nine features (compared to 5). Because we need our model for interpretation not performance, it is decided to choose Tolerance: 0.001 with nine features.

4.1.1. Feature Importance and Model Calibration

After performing Linear Regression using statsmodels library, with the 9 features [avg_timeliness + total_clicks + avg_week_clicks + folder_clicks + forumng_clicks + homepage_clicks + oucollaborate_clicks + oucontent_clicks + page_clicks+ questionnaire_clicks+ quiz_clicks+subpage_clicks+num_of_prev_attempts+Lower_A_level]

The OLS regression results were as follows:

- **R-squared: 0.549**
- **F_static : 131.0**
- **Intercept: 61.14 with confidence interval between 55.8 and 66.435**

OLS Regression Results

Dep. Variable:	final_score	**R-squared:**	0.549
Model:	OLS	**Adj. R-squared:**	0.545
Method:	Least Squares	**F-statistic:**	131.0
Date:	Mon, 13 Jan 2025	**Prob (F-statistic):**	1.84e-248
Time:	01:31:53	**Log-Likelihood:**	-6519.8
No. Observations:	1522	**AIC:**	1.307e+04
Df Residuals:	1507	**BIC:**	1.315e+04
Df Model:	14		
Covariance Type:	nonrobust		

The following features were identified as significant predictors of final grades:

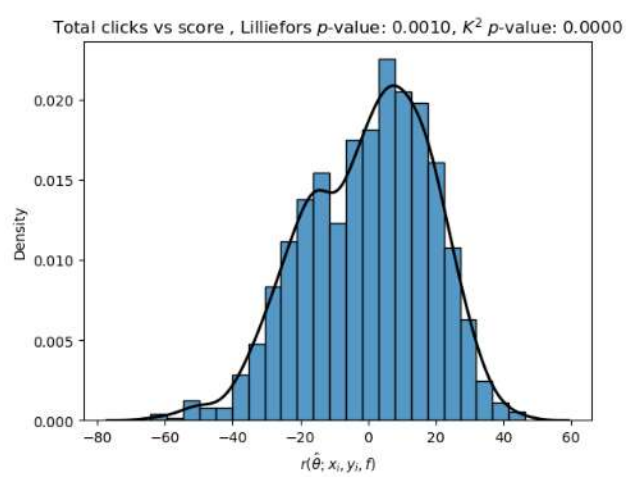
- **Total Clicks:** Higher engagement has a -0.0421 coefficient suggesting not important. But its subcategories like folder_clicks and questionnaire click are more important.
- **Forum Participation:** Active participation in forums was a key predictor of success. With forumng having the highest coefficient of 10.97
- **Demographics:** Lower previous educational qualifications having negatively impacts final grades

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	61.1435	2.698	22.665	0.000	55.852	66.435
Lower_A_level[T.True]	-6.2868	0.935	-6.720	0.000	-8.122	-4.452
avg_timeliness	-0.0421	0.016	-2.577	0.010	-0.074	-0.010
total_clicks	-0.0069	0.001	-5.506	0.000	-0.009	-0.004
avg_week_clicks	-3.4655	0.870	-3.984	0.000	-5.172	-1.759
folder_clicks	3.0698	0.283	10.845	0.000	2.515	3.625
forumng_clicks	10.9704	0.810	13.545	0.000	9.382	12.559
homepage_clicks	0.0120	0.003	3.518	0.000	0.005	0.019
oucollaborate_clicks	-0.3200	0.066	-4.825	0.000	-0.450	-0.190
oucontent_clicks	0.0151	0.001	10.235	0.000	0.012	0.018
page_clicks	-0.2911	0.086	-3.403	0.001	-0.459	-0.123
questionnaire_clicks	0.2892	0.055	5.223	0.000	0.181	0.398
quiz_clicks	0.0168	0.002	10.154	0.000	0.014	0.020
subpage_clicks	-0.0115	0.006	-1.903	0.057	-0.023	0.000
num_of_prev_attempts	-7.0948	1.134	-6.259	0.000	-9.318	-4.871

Model Performance

- The final regression model explained approximately 55% of the variance in students' grades (ΔR-squared = 0.549).
- Key features included total clicks, forum interactions, and the number of previous attempts.
- RMSE was 7.32, indicating reasonable prediction accuracy for the diverse student population.

Visualizations



5. Discussions

For this part it is explained that how we can answer the two main questions of the assignment, and we will discuss some recommendations.

A. Is the VLE improving students' grades?

As it was explained in previous sections VLE consists of various online activities (such as forum, page, and quiz clicks), which are used as predictors in the model to estimate the students' final_score. The key findings are as follows:

- Total clicks, forum clicks, homepage clicks, oucollaborate clicks, and oucontent clicks are all significant predictors of final score, with p-values close to 0 (showing statistical significance).
- The coefficients for these variables indicate their direction and strength of impact on final_score:
  - Forumng_clicks: A positive coefficient (10.97) indicates that an increase in forum clicks is associated with an increase in final score.
  - Homepage_clicks: Similarly, a positive coefficient (0.012) suggests that more homepage interactions are linked to higher final scores.
  - Oucontent_clicks: Also positively correlated, with a coefficient of 0.015, indicating a positive relationship with final scores.
  - Total_clicks and avg_week_clicks: These show negative coefficients, which could imply that students who engage too much in these activities may be detracting from their performance, or it may be a proxy for students spending time on the platform without being engaged in productive activities like forum_ng.

The VLE activities, especially forum engagement, homepage, and content interaction, do show a significant relationship with students' grades. However, the negative relationship between total_clicks and avg_week_clicks with final scores could suggest that the quality of interaction is more important than quantity. Therefore, simply using the VLE more may not guarantee better grades for students.

B. Can we predict students' grades?

The R-squared value of 0.549 indicates that approximately 54.9% of the variance in final_score can be explained by the selected predictors in the model. While this suggests a moderate fit, there is still considerable unexplained variance, meaning there are other factors that influence final scores that are not captured by this model.

- Adjusted R-squared (0.545) accounts for the number of predictors and still indicates a reasonable fit.
- F-statistic: The very small p-value for the F-statistic (1.84e-248) indicates that the overall model is statistically significant, meaning the predictors are collectively useful for predicting final scores.

So, is possible to predict students' grades with this model, but the predictions may not be highly accurate, given the relatively moderate R-squared. The model can provide useful insights into factors influencing grades, but further improvements in feature selection or adding more predictors may enhance prediction accuracy.

C. Recommendations:

- Promoting VLE: Since interactions with the VLE, particularly in forums and content areas, are positively associated with higher final scores, promoting these activities can help improve students' grades. So for the course FFF_2013J students should be encouraged to actively participate in discussions and engage with content more meaningfully, rather than simply increasing the volume of activity (clicks).
- Assessment Types:
  The negative relationship between num_of_prev_attempts and final score (-7.09) suggests that students who have made more previous attempts may be struggling, which could indicate a need for better support or alternative assessment methods for these students.
  Also, the negative correlation with total clicks and avg_week_clicks could suggest that a more structured or focused approach of using the VLE might be a better option than leaving students to freely explore the platform without a structured plane around using the contents.
- Improvement through Focused Engagement: The model's coefficients for VLE activity suggest that structured interactions on the platform, such as participating in discussions or engaging with content, are more beneficial than unstructured browsing or interacting with multiple resources.

6. Conclusion

With this analysis we demonstrated that the VLE for open university significantly enhances learning outcomes and provided predictive insights using linear regression. For OU to maximize its impact and enhancing students’ grades, it should actively promote VLE engagement and tailor support to various student needs with different background. Each module has different structure and criteria, and a tailored VLE interaction plan can help students.