Student Retention at Open University: Business Insights and Potential Solutions

Shiwen Xu

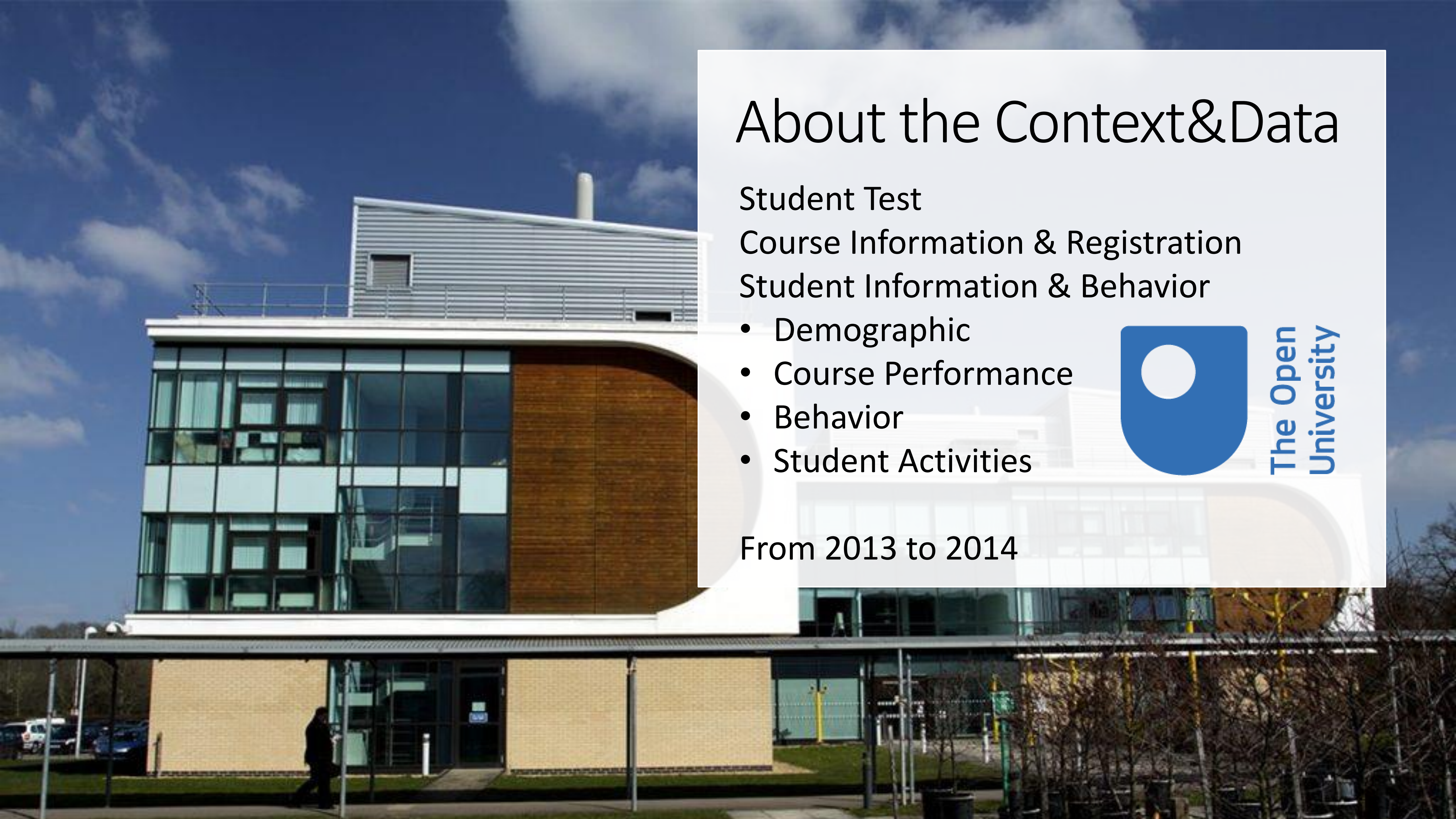
Presentation to Adobe Customer and Product Analytics Team

Agenda

- Outline of Steps in Formulating the Problem and Exploring the Data
- Motivation and Key Question: How to Improve Student Retention (Reduce Withdrawal)?
- Explore 3 Underlying Drivers of Course Withdrawal
- Prediction Model
- Regression Analysis
- Summary, Recommendations, and Next Steps
- Appendix: additional analysis

Steps in Analyzing Data and Identifying Problems

1. Understand the business context and Identify the problem
2. Prepare Data Infrastructure
    1. Upload each CSV files to Google Cloud Storage
    2. Build a Database in Google BigQuery using Relational Data Model
3. Explore Data and Identify the problem with the largest impact
    1. SQL in Google BigQuery
    2. R for further anaysis
    3. Identify student retention as a major problem
4. Explore 3 Underlying Drivers of Course Withdrawal
5. Prediction Model
6. Regression Analysis
7. Recommendation on Potential Solution



About the Context&Data

Student Test

Course Information & Registration

Student Information & Behavior

- Demographic
- Course Performance
- Behavior
- Student Activities

From 2013 to 2014

The Open University

Motivation: why focusing on student retention

Goal of Open University:

- Scale by enrolling more students (Acquisition)
- Engage each student (Retention)

Student Acquisition and Retention at University and Course Level

	Acquisition	Retention
University-level	Sign-up Online	Actively Taking Courses or Get Degree
Course-level	Enroll into a Course	Complete a Course

Motivation: why focusing on student retention

Why:

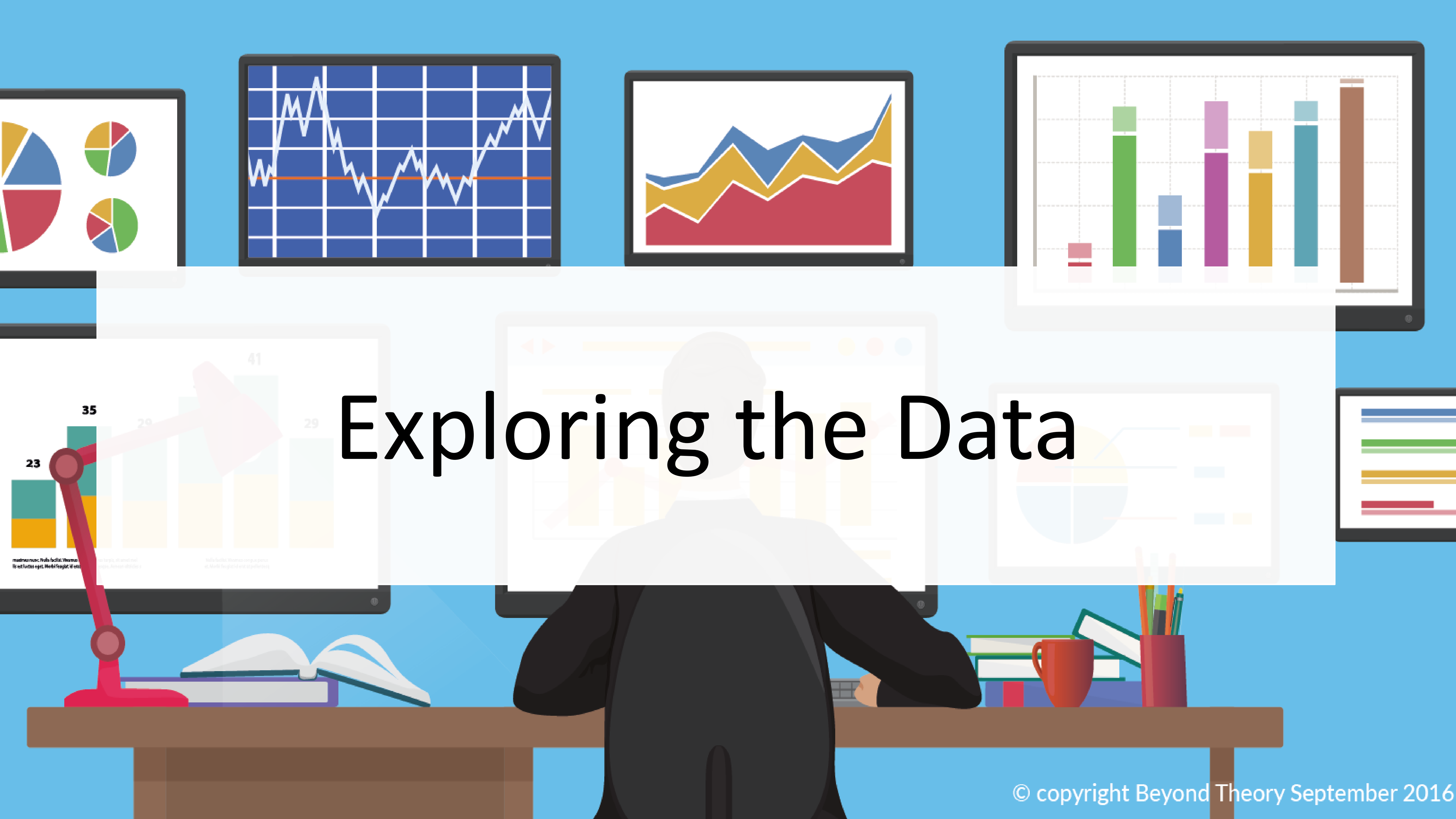
- Business angle:
  - In distance learning, customer acquisition is relatively simple (no geographic constraint)
  - But student retention may face a fundamental challenge due to the lack of physical interaction
- Data angle:
  - We do not have data on customer acquisition (e.g. advertising channel & spending)
  - Thus can make little data-driven recommendations along this line.

	Acquisition	Retention
University-level	Sign-up Online	Actively Taking Courses or Get Certificate
Course-level	Enroll into a Course	Complete a Course

Problems & Insights

Student Retention

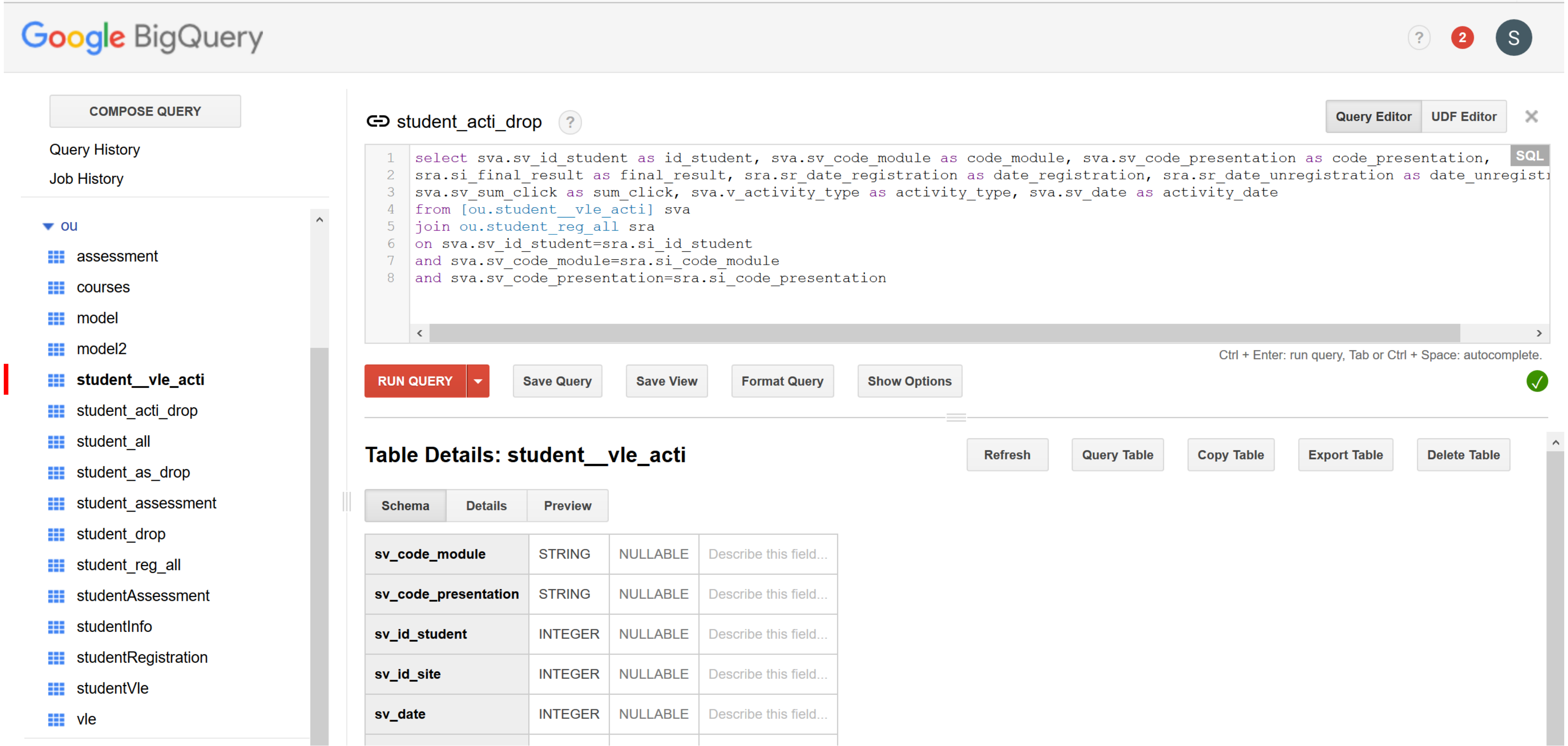
- Long-term: not finish the degree / complete enough courses
- Short-term: not complete the course (Withdrawal)
  - Course Design
  - Student Education Level
  - Student Activities



Exploring the Data

Prepare Data Infrastructure (Google BigQuery)

SQL Queries are included in the Notepad.



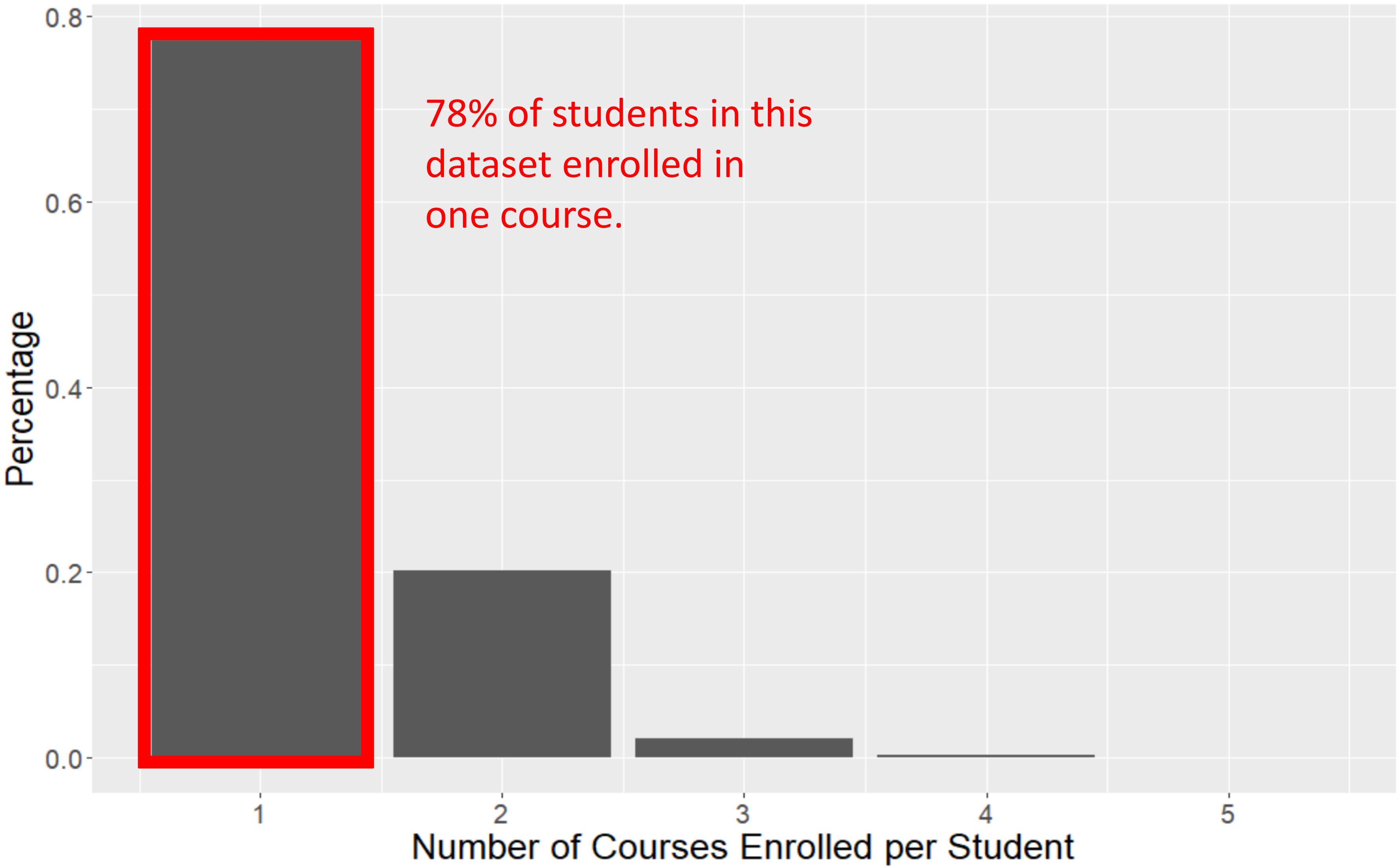
```sql
select sva.sv_id_student as id_student, sva.sv_code_module as code_module, sva.sv_code_presentation as code_presentation,
sra.si_final_result as final_result, sra.sr_date_registration as date_registration, sra.sr_date_unregistration as date_unregist
sva.sv_sum_click as sum_click, sva.v_activity_type as activity_type, sva.sv_date as activity_date
from [ou.student__vle_acti] sva
join ou.student_reg_all sra
on sva.sv_id_student=sra.si_id_student
and sva.sv_code_module=sra.si_code_module
and sva.sv_code_presentation=sra.si_code_presentation
```

Ctrl + Enter: run query, Tab or Ctrl + Space: autocomplete.

Table Details: student__vle_acti

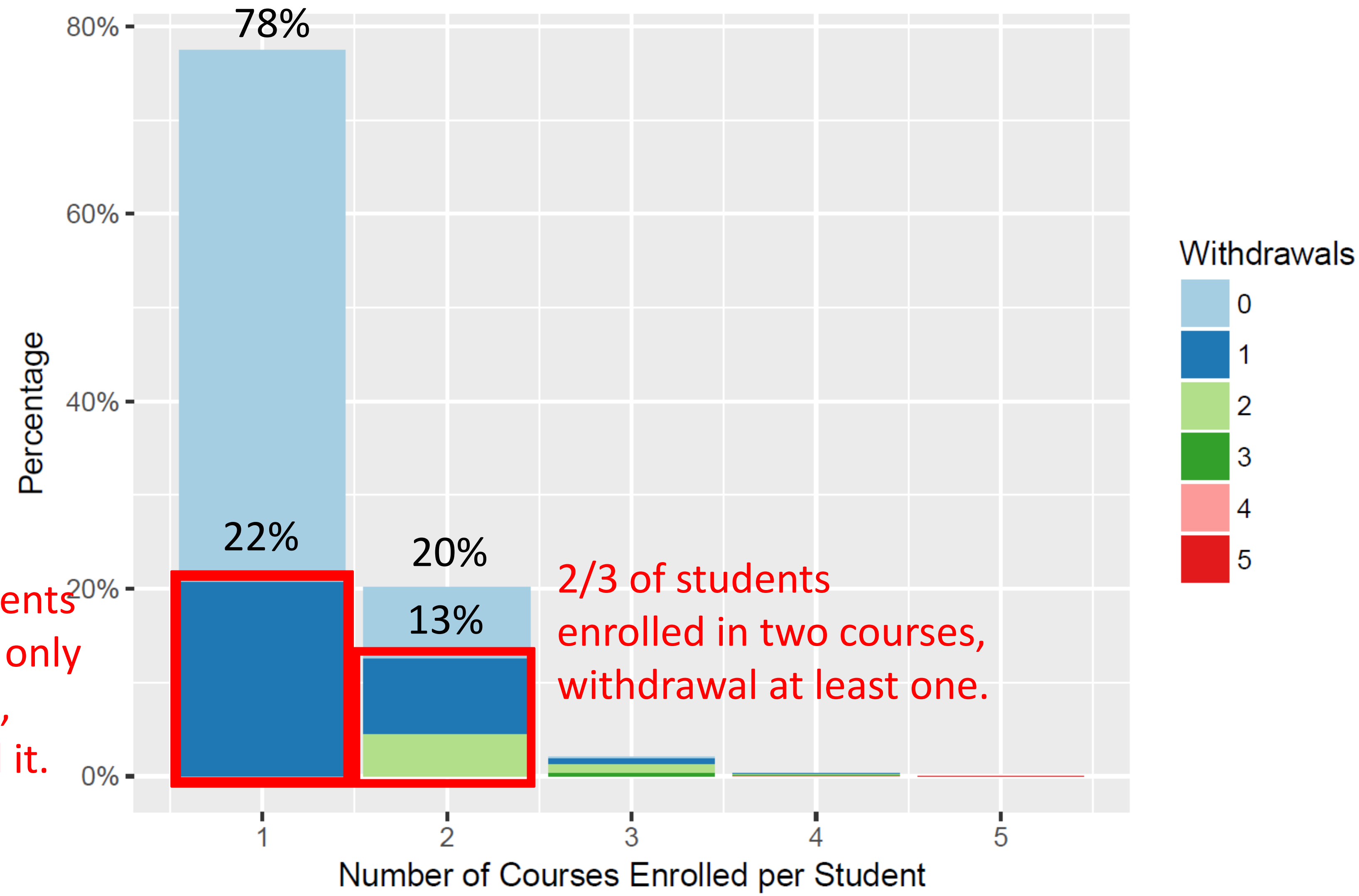
sv_code_module	STRING	NULLABLE	Describe this field...
sv_code_presentation	STRING	NULLABLE	Describe this field...
sv_id_student	INTEGER	NULLABLE	Describe this field...
sv_id_site	INTEGER	NULLABLE	Describe this field...
sv_date	INTEGER	NULLABLE	Describe this field...

The Distribution of Number of Courses Enrolled Per Students during the Data Period



78% of students in this dataset enrolled in one course.

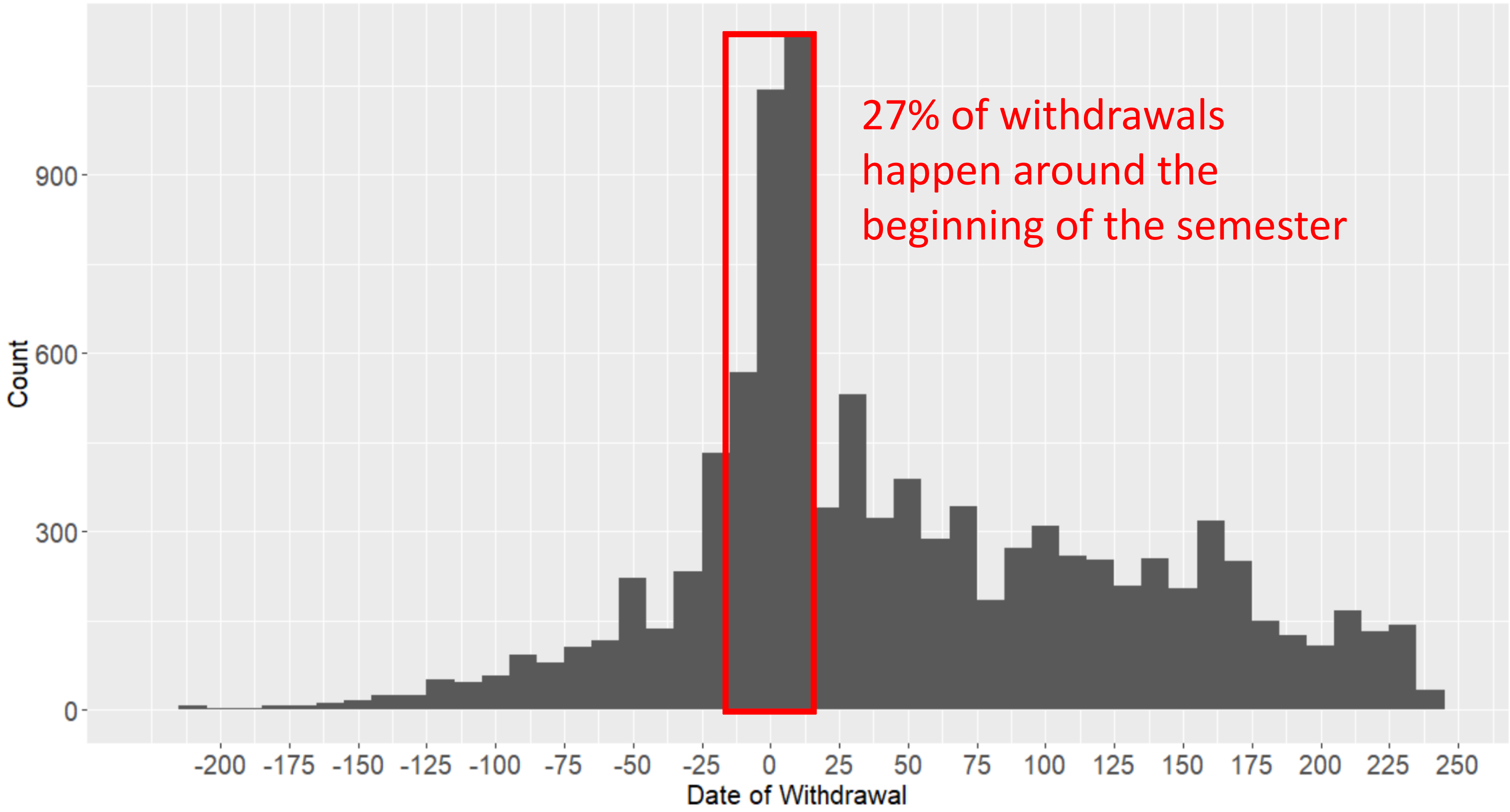
The Distribution of Number of Courses Enrolled Per Students Decomposed by Course Withdrawal or not



1/3 of students enrolled in only one course, withdrawal it.

2/3 of students enrolled in two courses, withdrawal at least one.

The Distribution of Withdrawal



27% of withdrawals happen around the beginning of the semester

Retention Problem

Two Aspects of the Retention Problem (Conclude from Slides Above)

- Students at Open University only register for a small number of courses (mostly one)
- Even worse, a big portion of students drop the course

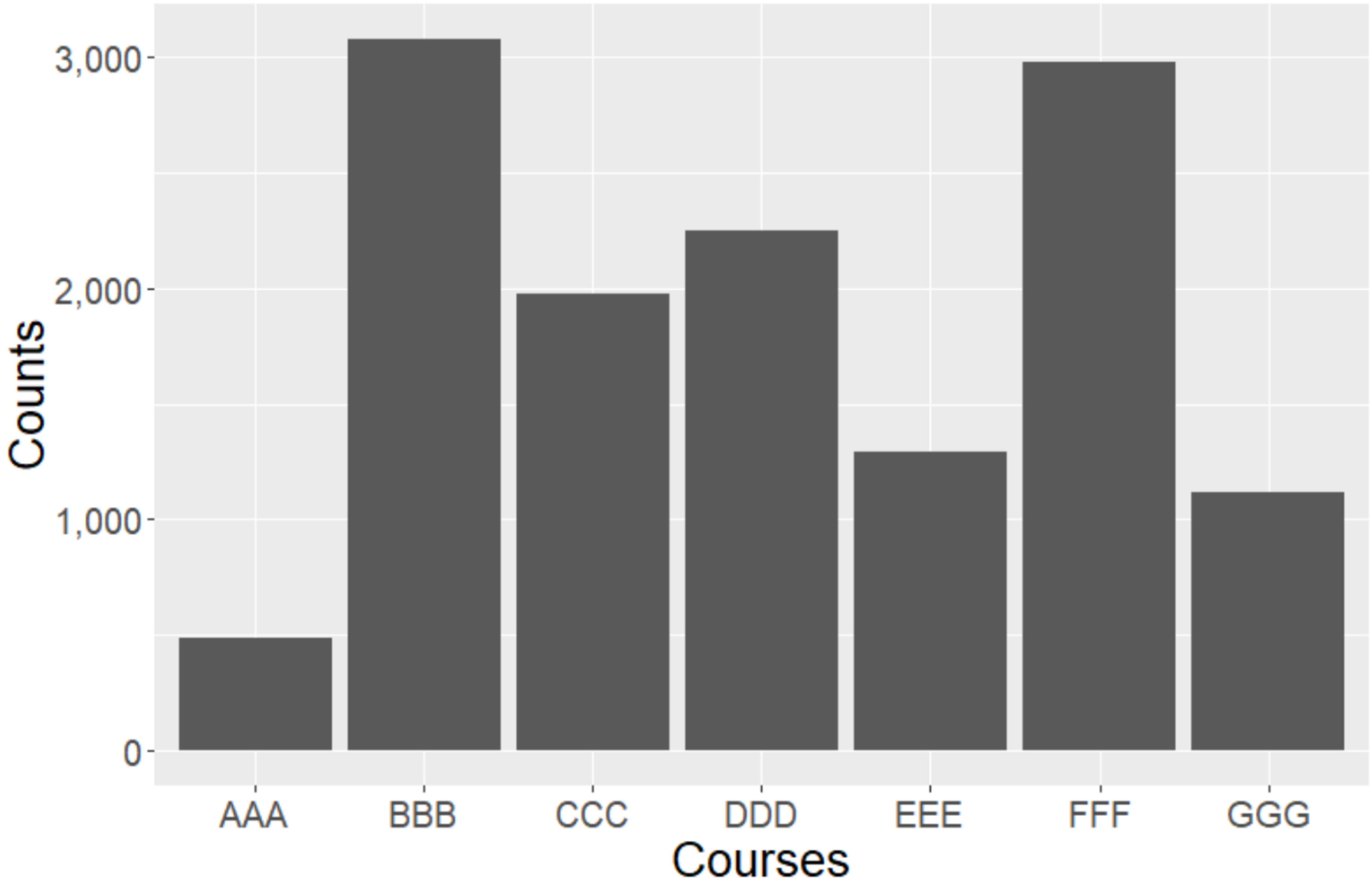
Why?

I found three aspects related with this problem:

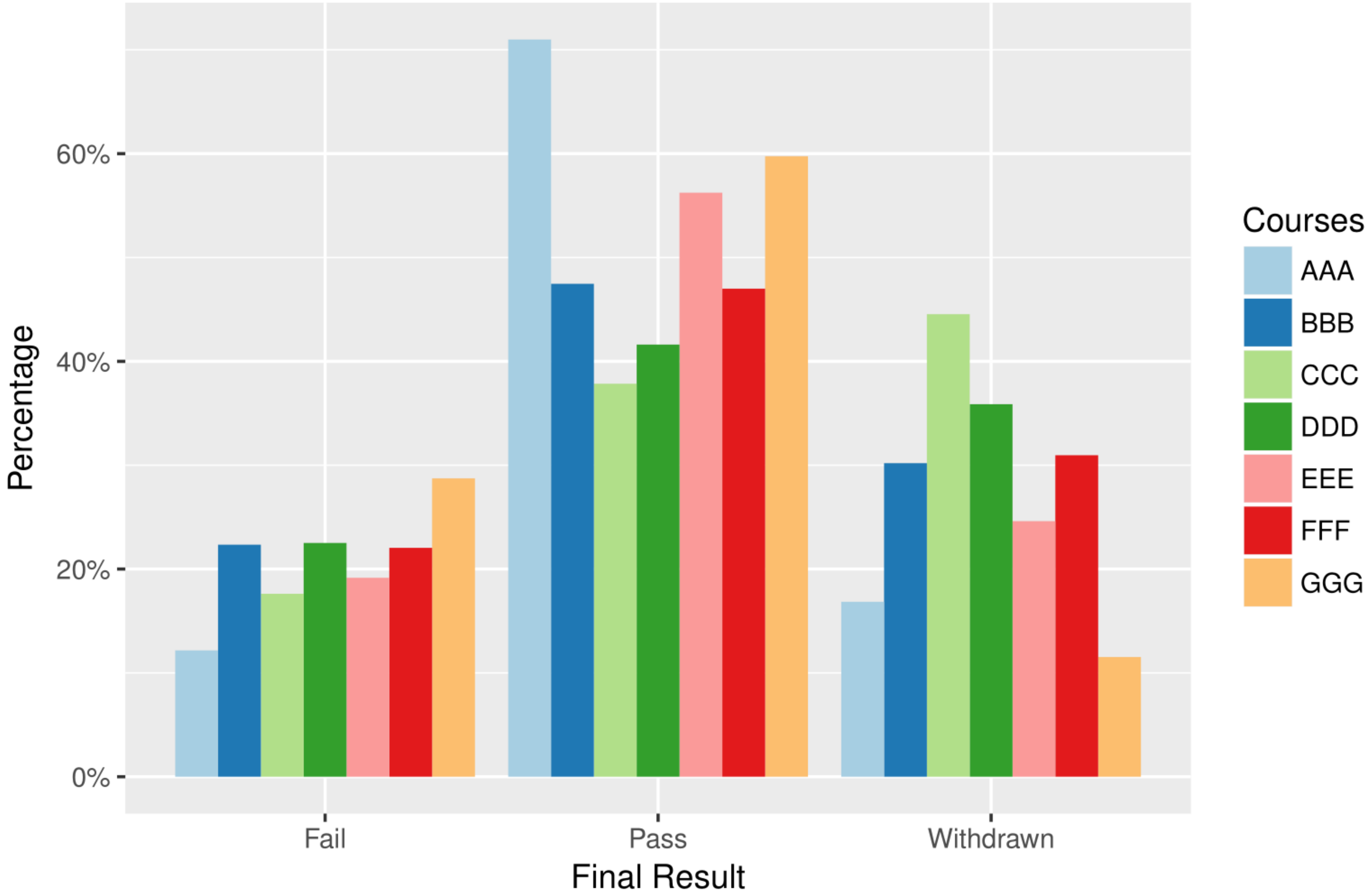
1. Course design
2. Education Background
3. Student Activity

Whether the course withdrawal caused by the course design?

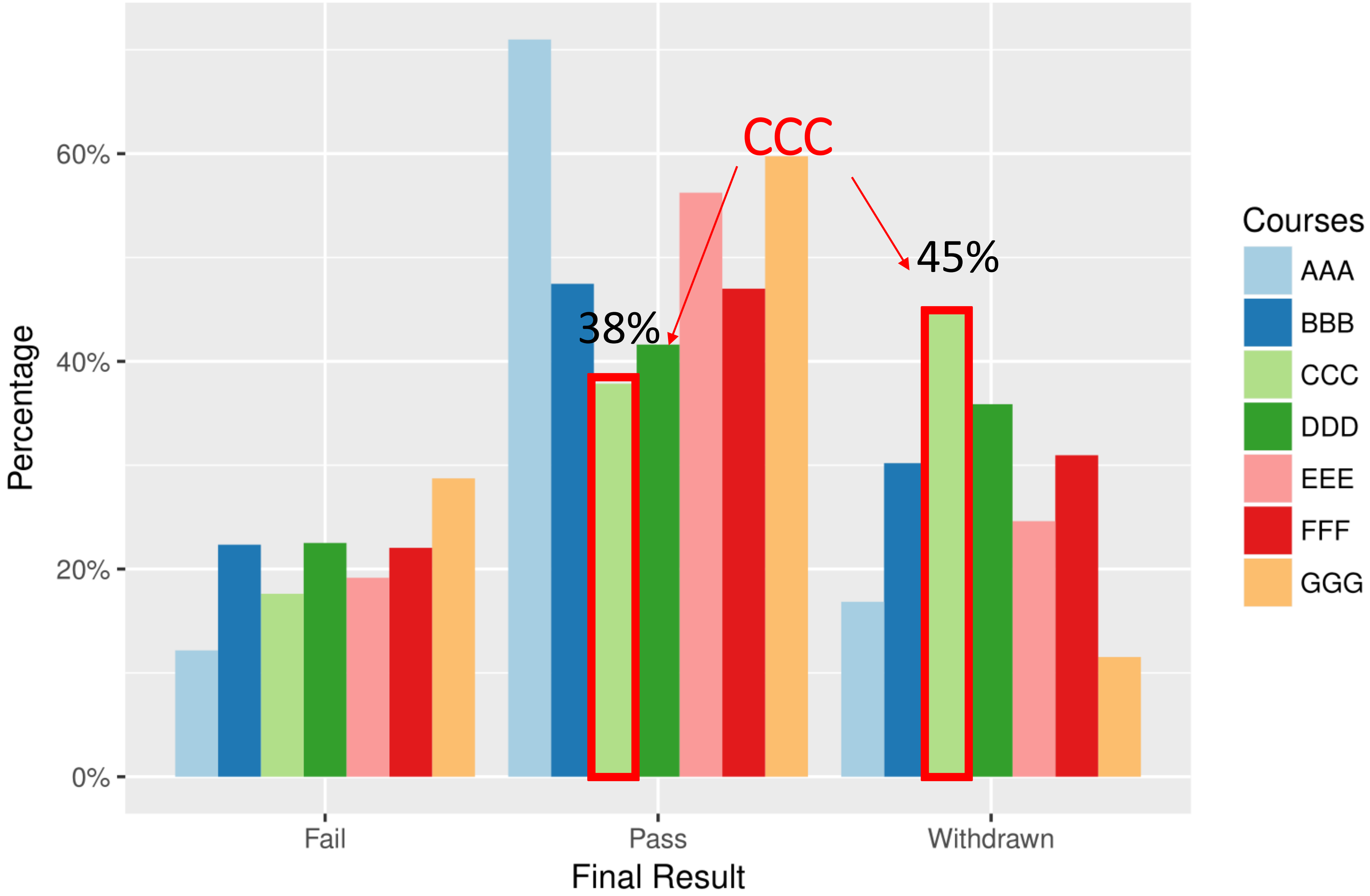
The Total Enrollment for Courses



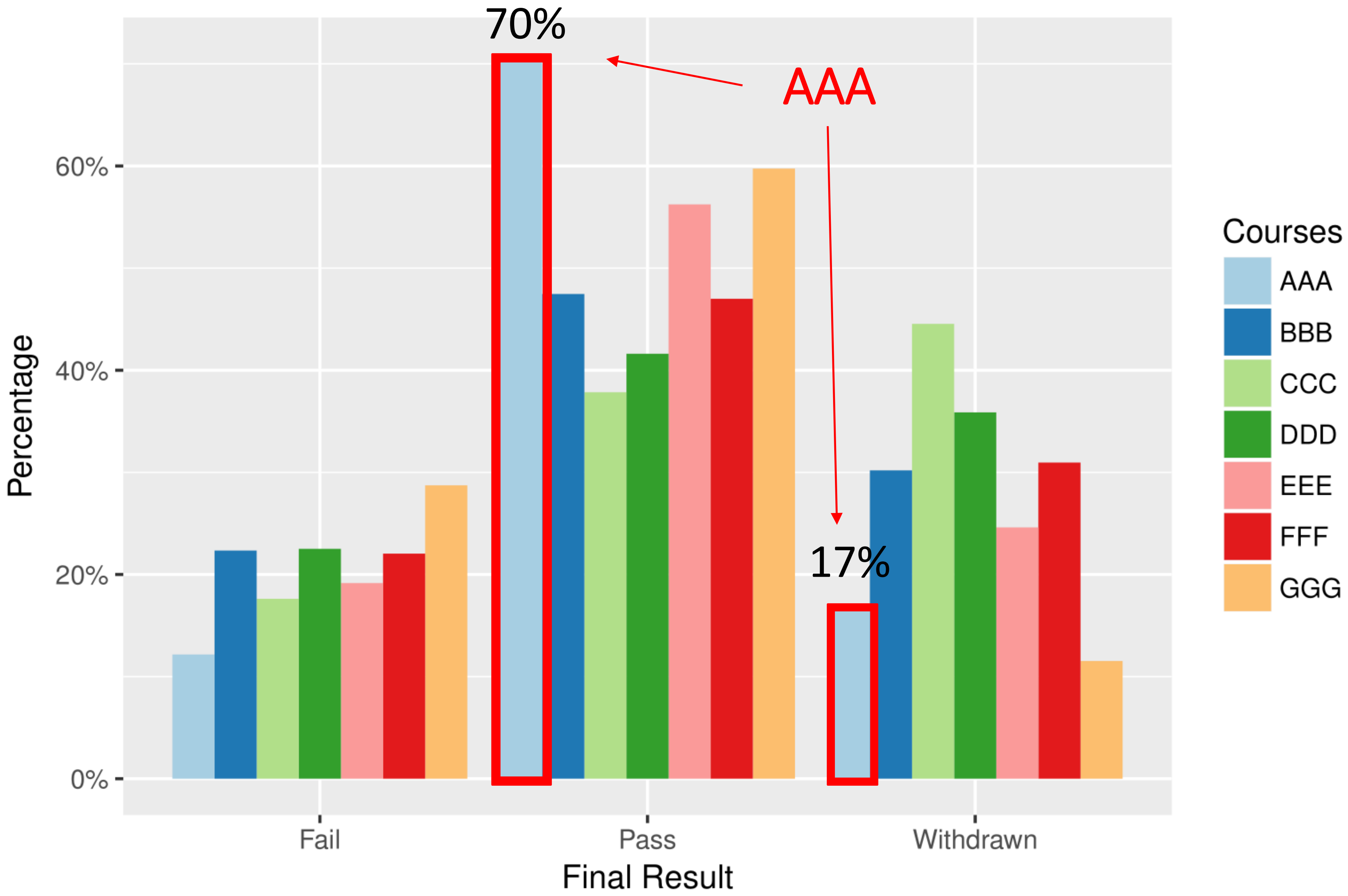
Percentage of Student Performance for Each Course



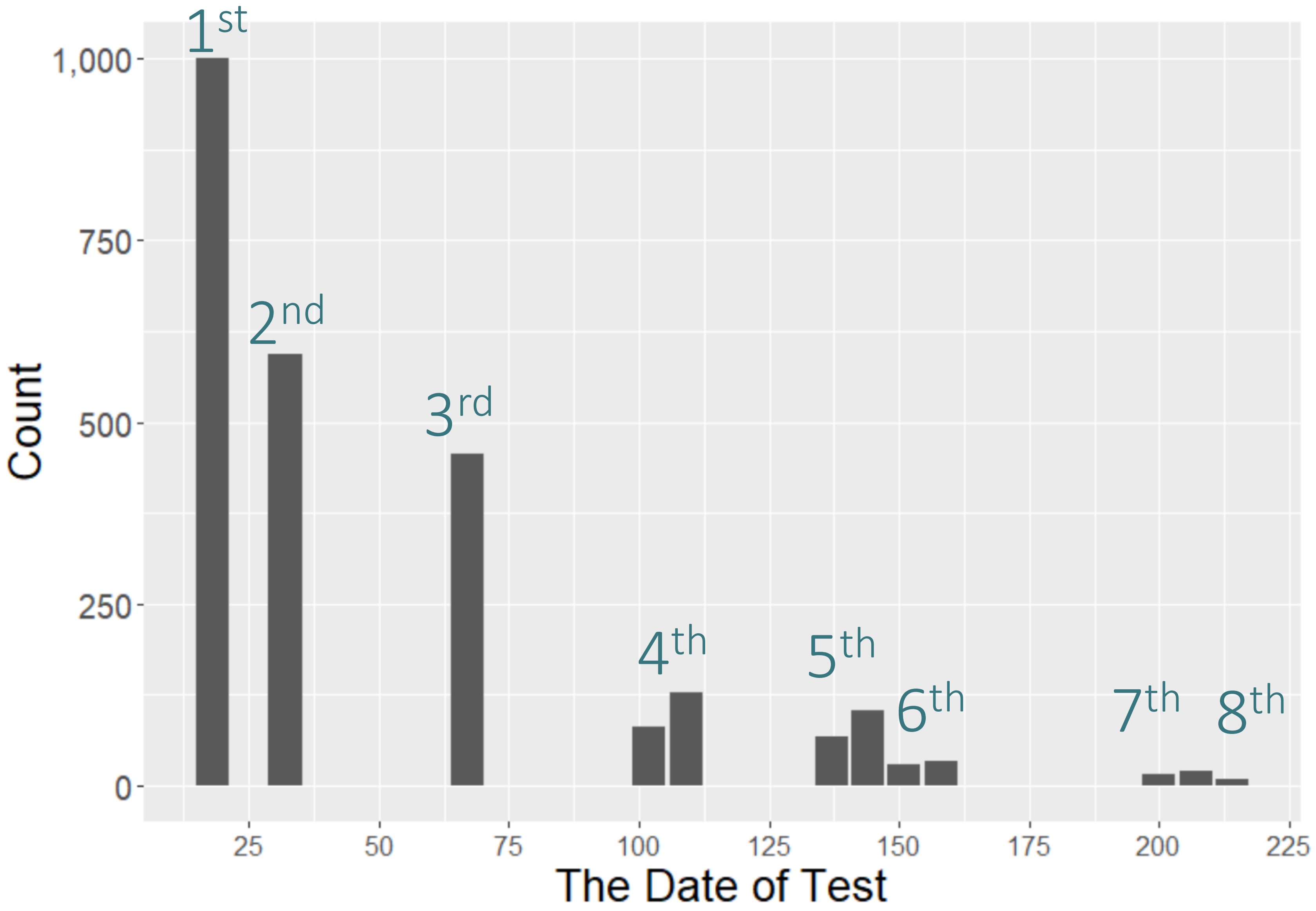
Percentage of Student Performance for Each Course



Percentage of Student Performance for Each Course



The Number of Students Taking Tests for Course CCC



First Test: CMA, 2%

Second Test: TMA, 9%

Third Test: CMA, 7%

Fourth Test: CMA, 8%

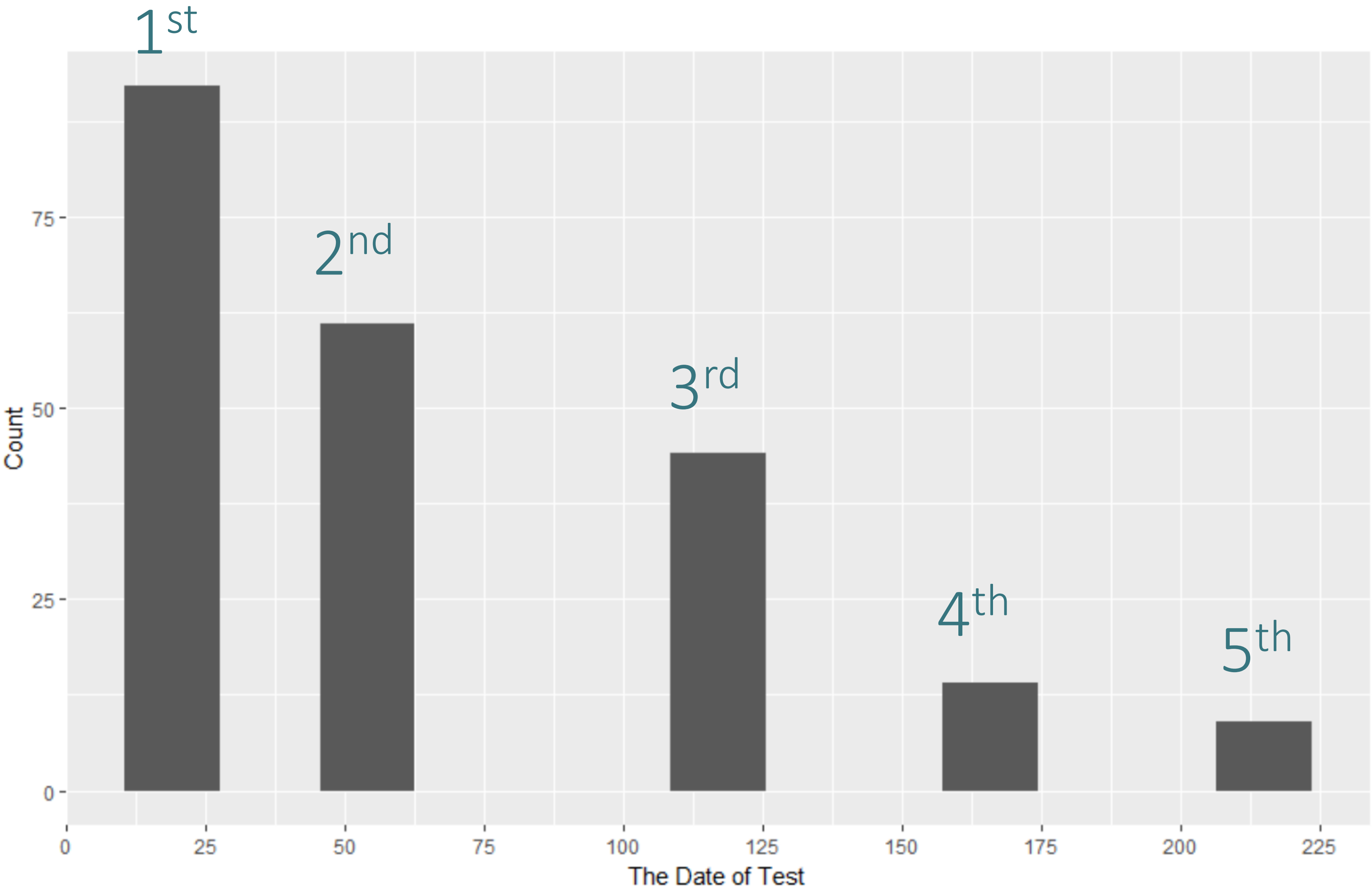
Fifth Test: TMA, 22%

Sixth Test: TMA 22%

Seventh Test: TMA 22%

Eighth Test: CMA 8%

The Number of Students Taking Tests for Course AAA



First Test: TMA, 10%

Second Test: TMA, 20%

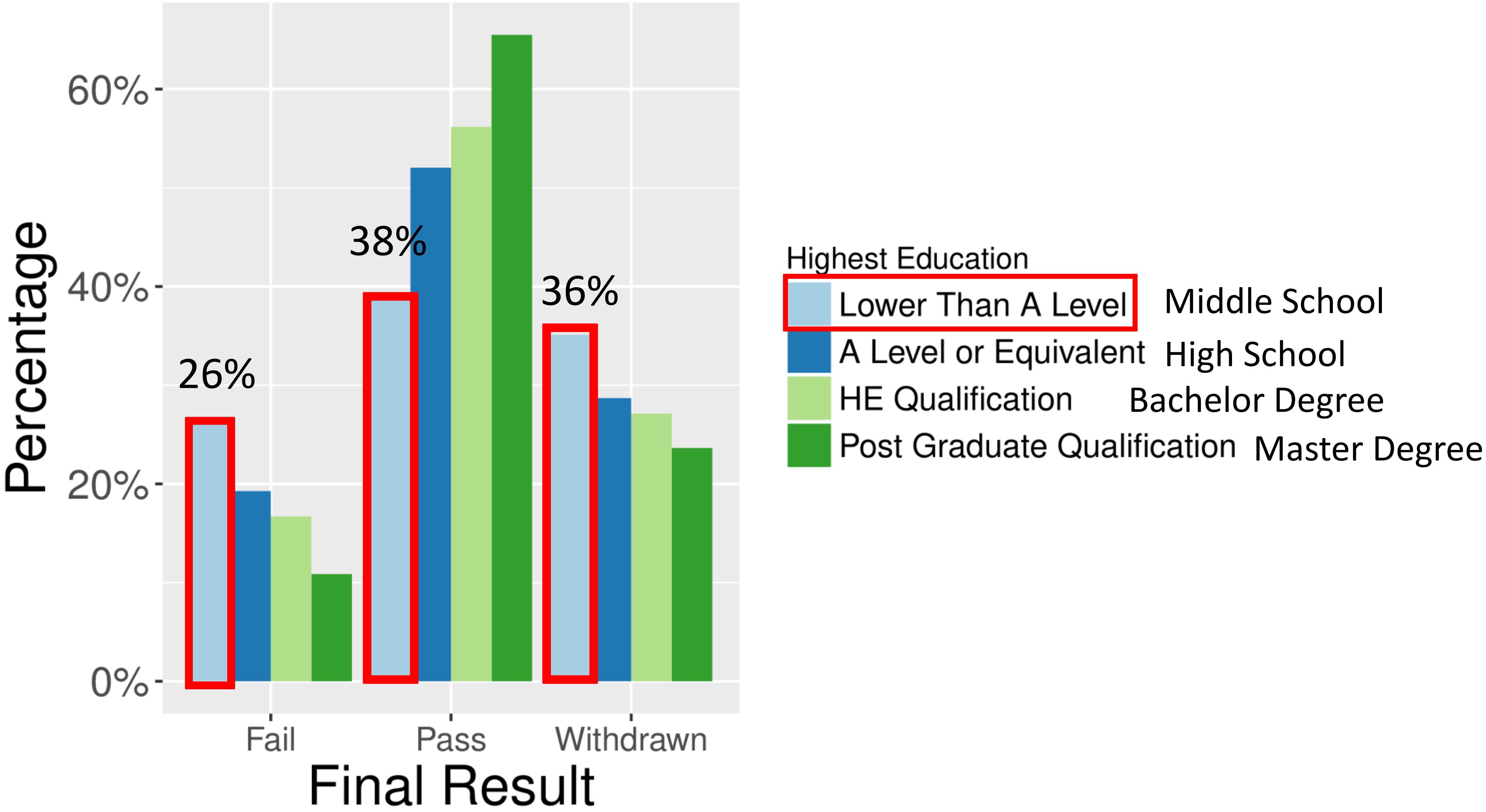
Third test: TMA, 20%

Fourth TMA, 20%

Fifth TMA, 30%

Whether the course withdrawal caused by the diverse education background?

Student Education Background by Student Performance



26%

38%

36%

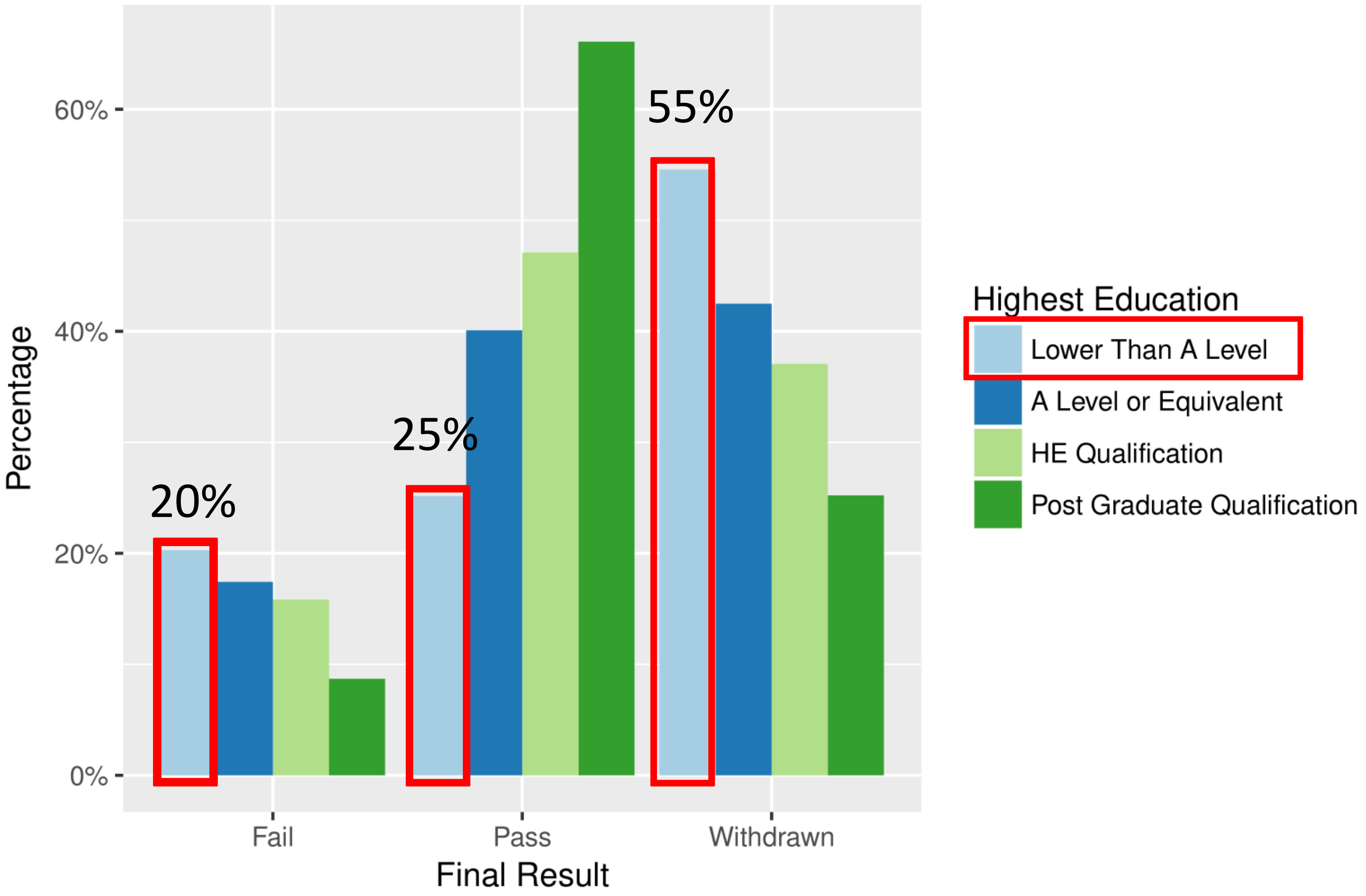
Lower Than A Level: Middle School

A Level or Equivalent: High School

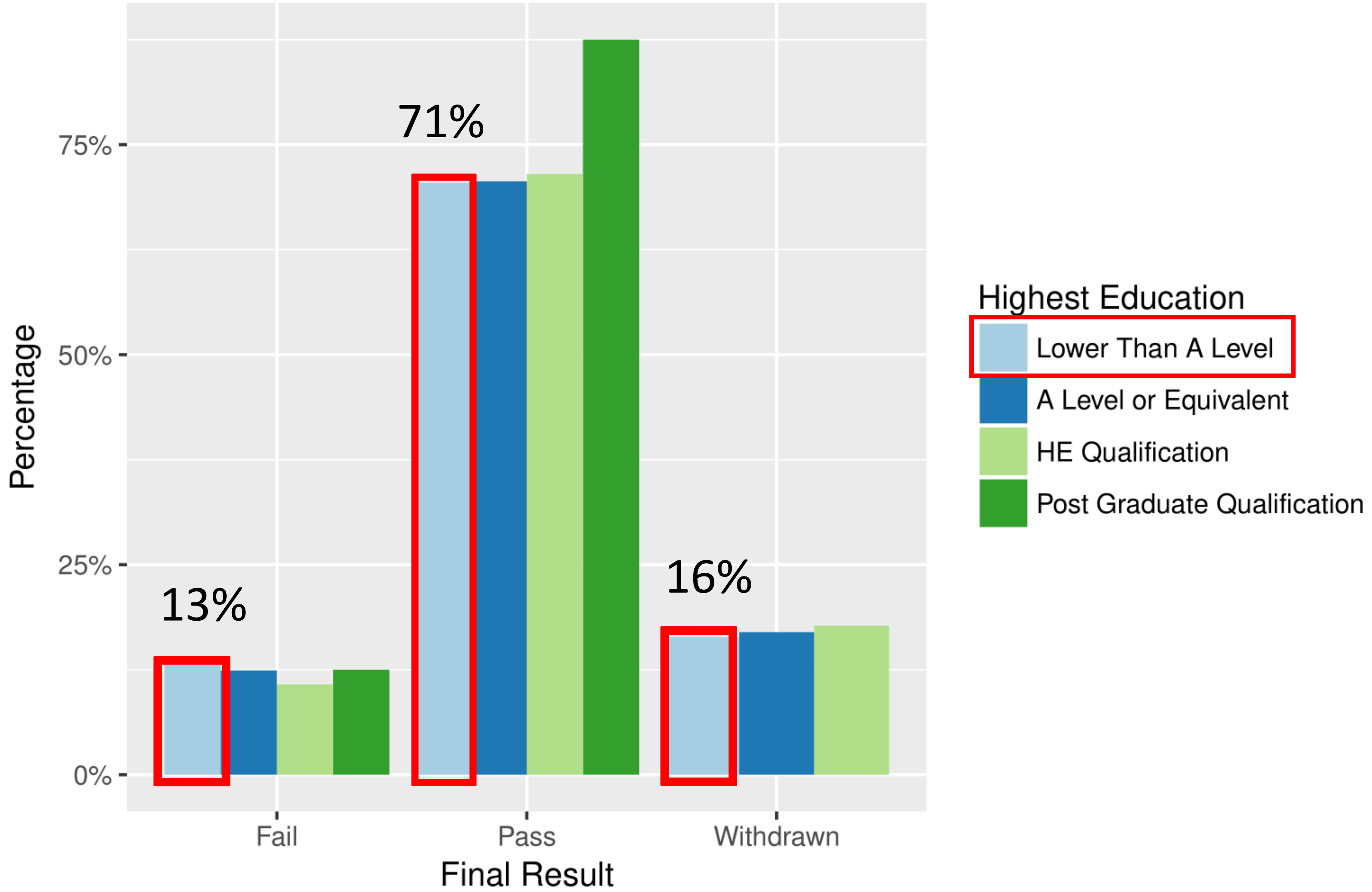
HE Qualification: Bachelor Degree

Post Graduate Qualification: Master Degree

Student Performance of Each Education Level for Course CCC

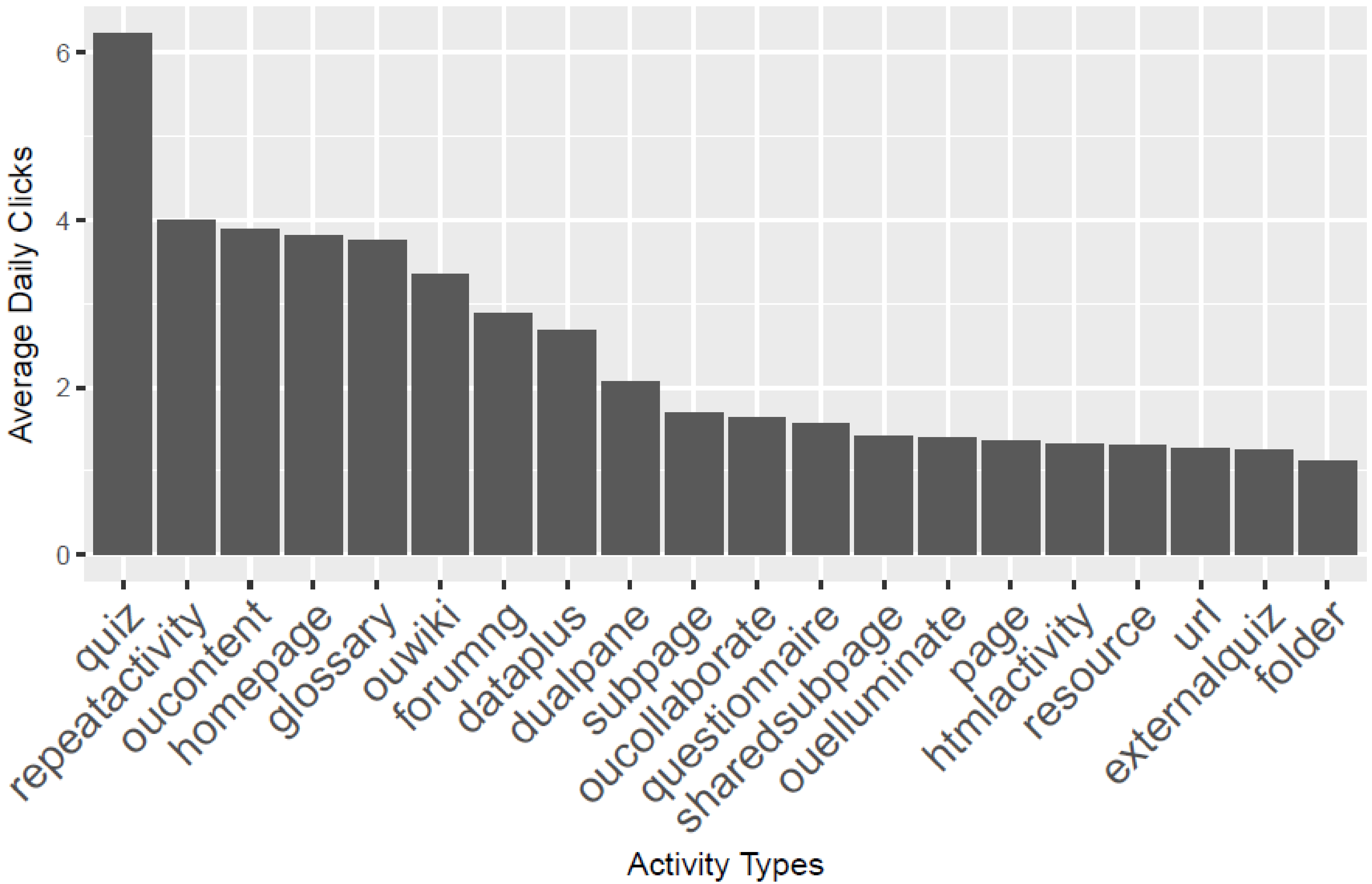


Student Performance of Each Education Level for Course AAA

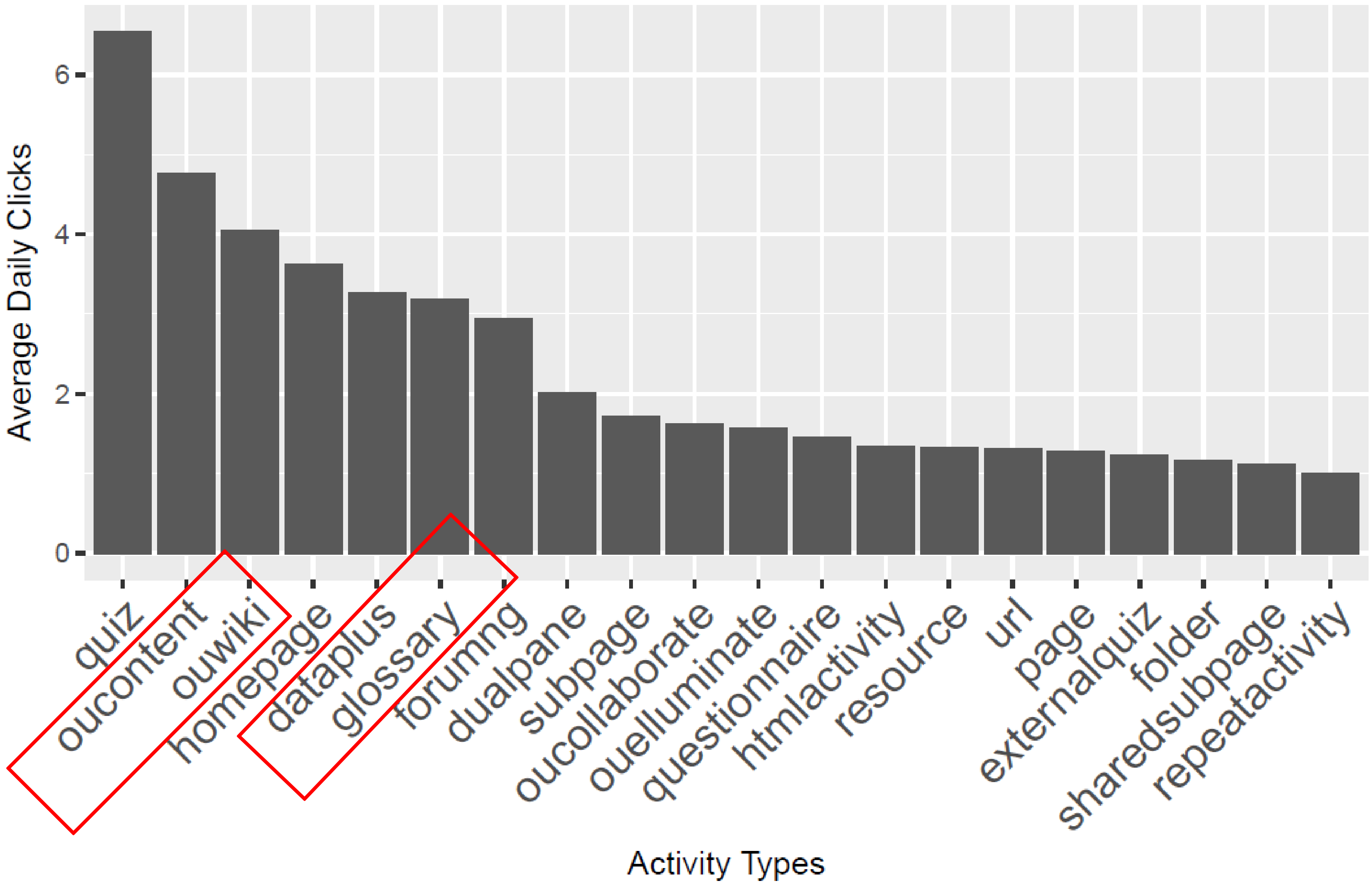


Whether course withdrawal caused by the student activity?

Average Daily Activity for Withdrawal by Activity Types



Average Daily Activity for Complete by Activity Types



Decision Tree Model:
Predicting Students at Risk of Course Withdrawal

Baseline Accuracy: when predicting all student WILL NOT drop the class (drop=0)

75.48%

Decision Tree Model:
Predicting Students at Risk of Course Withdrawal

Outcome:

Drop=0/1 (after a student register a specific course in a specific semester)

Predictors (Ranked by Importance):

Total click (up to registration),
Current credits,
Date of registration,
Highest education,
Number of previous attempts

Training dataset:

75% of the total observations, 21,921 rows

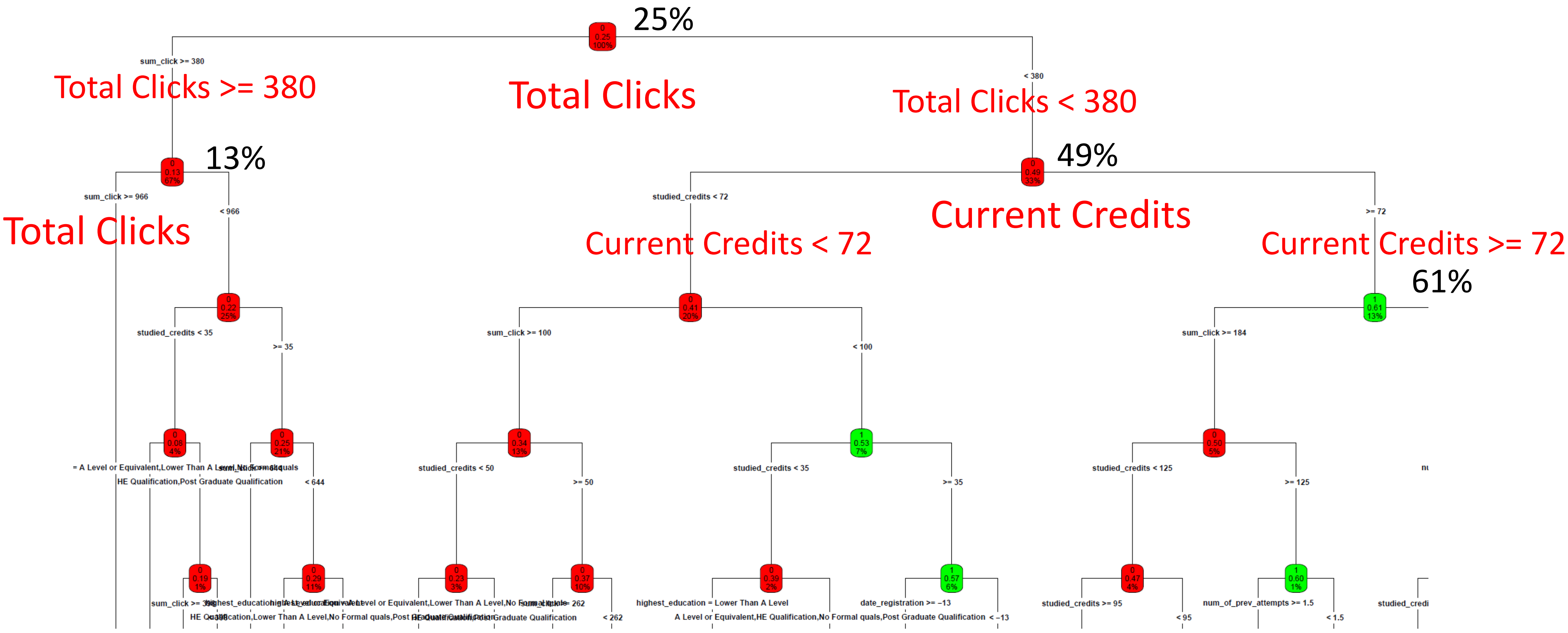
Testing:

25% of the total observations, 7307 rows

	Confusion Matrix - Testing			
	Prediction			
Actual		0	1	
	0	5118	1035	6153
	1	414	740	1154
		5532	1775	7307

Accuracy: 80.17%

Decision Tree Model: Predicting Students at Risk of Course Withdrawal

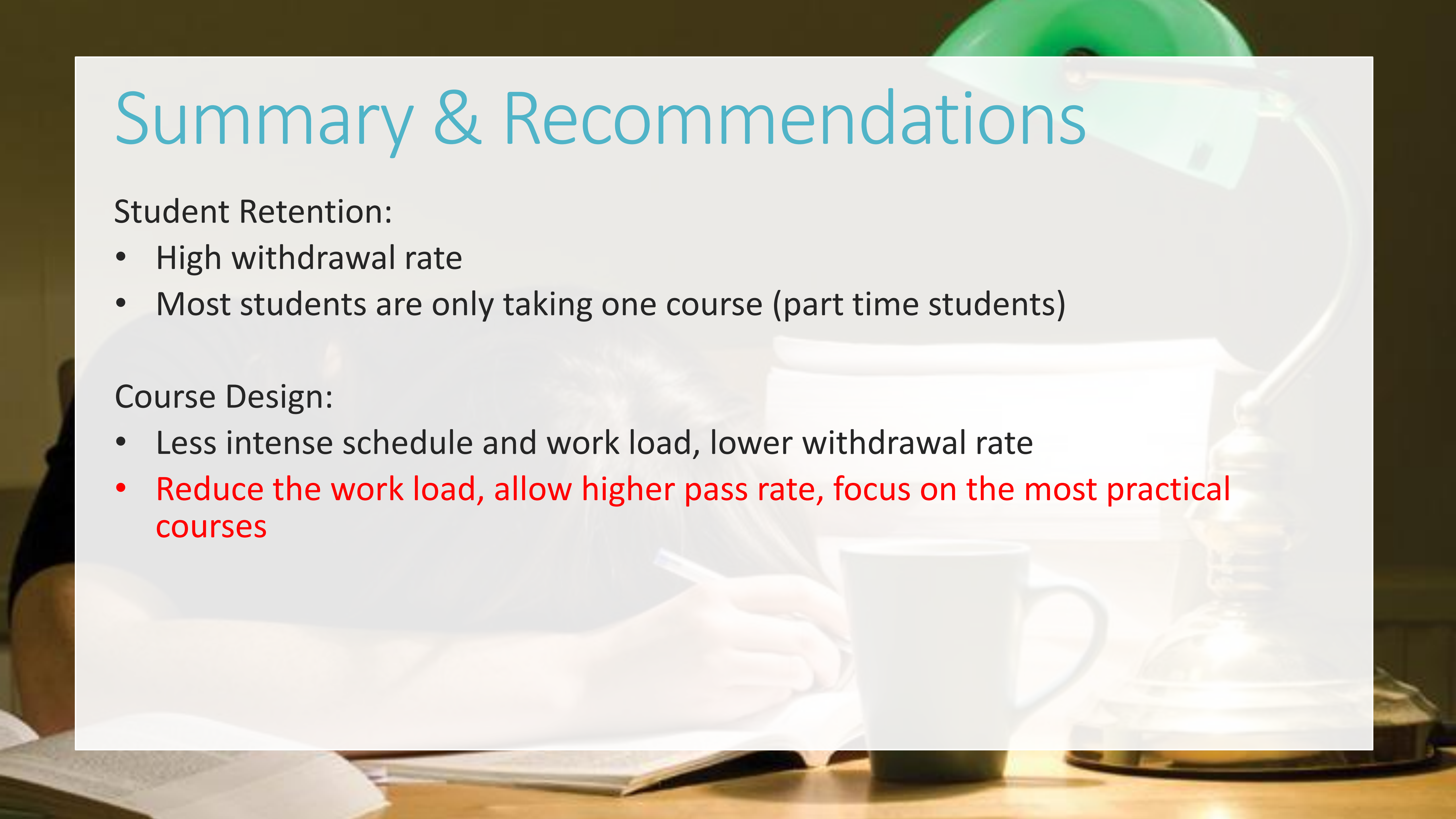


....

It is not a full image of the decision tree model.

Regression Analysis

Withdraw	Coefficients	Standard Error	t	p-value (P>t)	Confidence	Interval
Past Behavior						
Total Clicks	-0.000074	0.000001	-54.219	0.0000	-0.000077	-0.000072
Current Credits	0.0017	0.000061	27.522	0.0000	0.0016	0.0018
Number of Previous Attempts	-0.0154	0.005113	-3.012	0.0026	-0.0254	-0.0054



Summary & Recommendations

Student Retention:

- High withdrawal rate
- Most students are only taking one course (part time students)

Course Design:

- Less intense schedule and work load, lower withdrawal rate
- Reduce the work load, allow higher pass rate, focus on the most practical courses

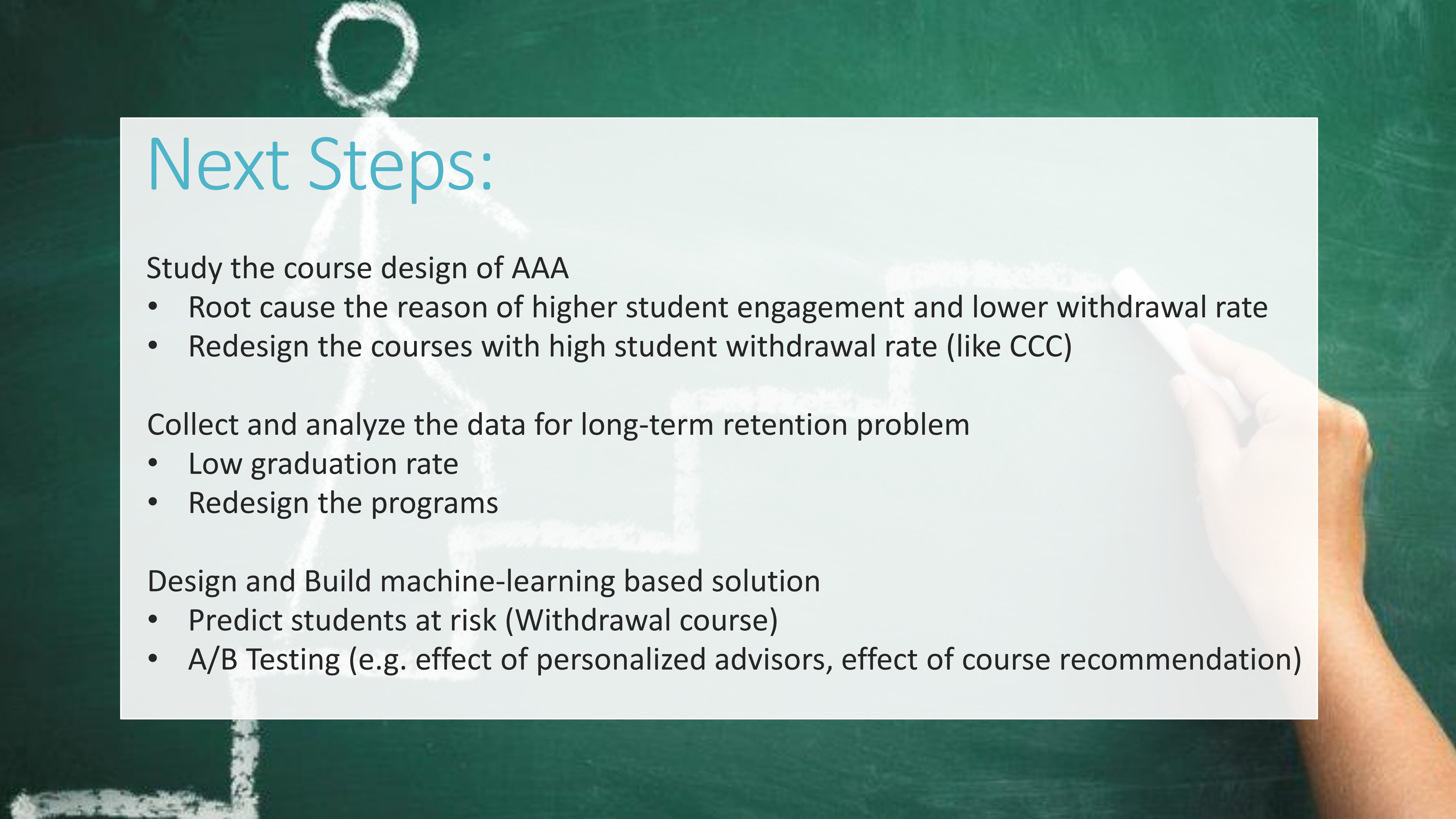
Summary & Recommendations Cont.

Diverse Education Level:

- Student from different education background perform differently
- Students from lower education background perform worse in demanding courses
- Personalized advisors
- Provide instruction of registering the most suitable courses
- Provide additional guidance during the semester (before and after exams)

Student Activities:

- Students who are more active, taking fewer credits, having previous attempts are less likely to drop
- Machine learning model to predict students at risk



Next Steps:

Study the course design of AAA

- Root cause the reason of higher student engagement and lower withdrawal rate
- Redesign the courses with high student withdrawal rate (like CCC)

Collect and analyze the data for long-term retention problem

- Low graduation rate
- Redesign the programs

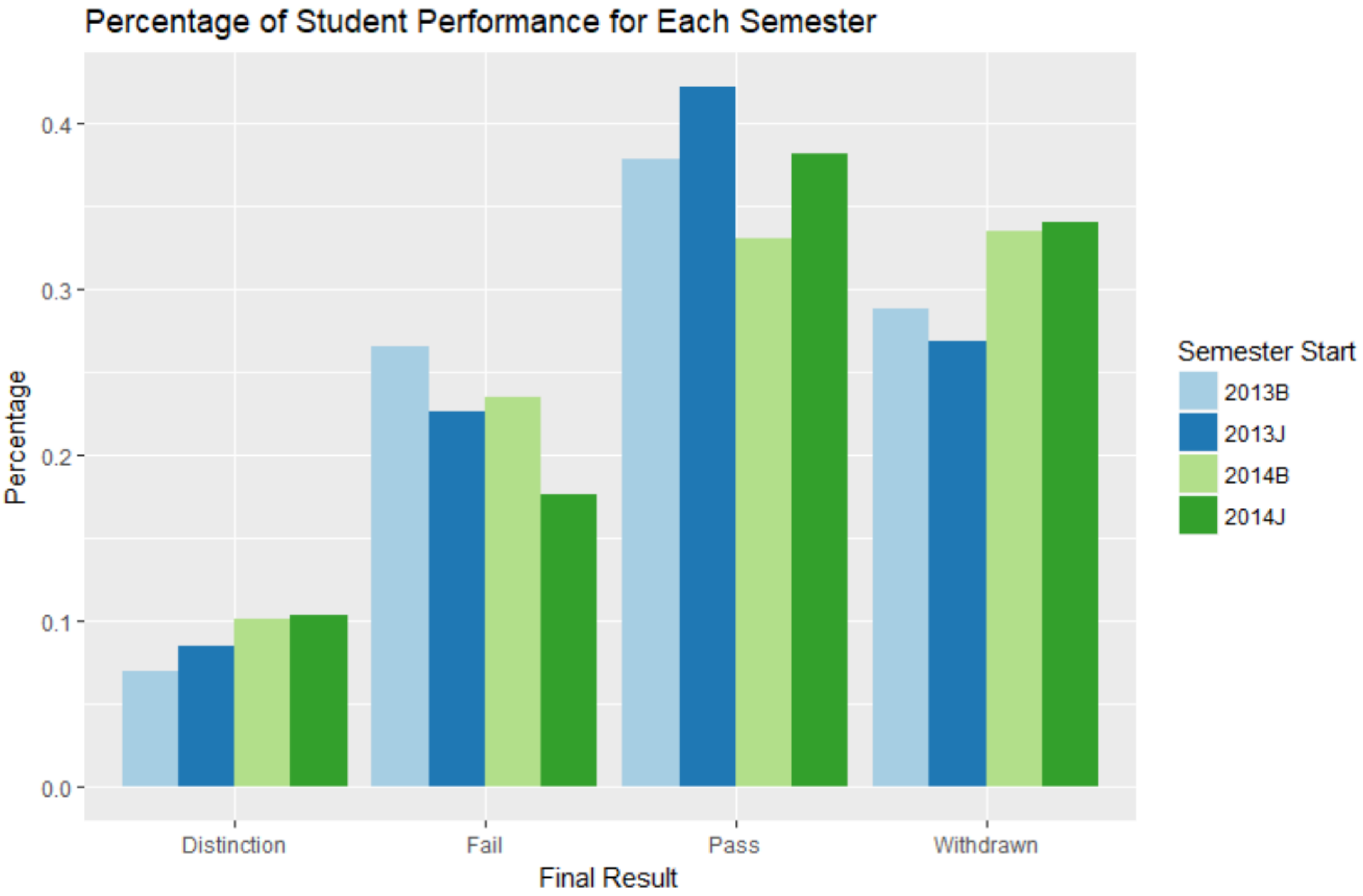
Design and Build machine-learning based solution

- Predict students at risk (Withdrawal course)
- A/B Testing (e.g. effect of personalized advisors, effect of course recommendation)

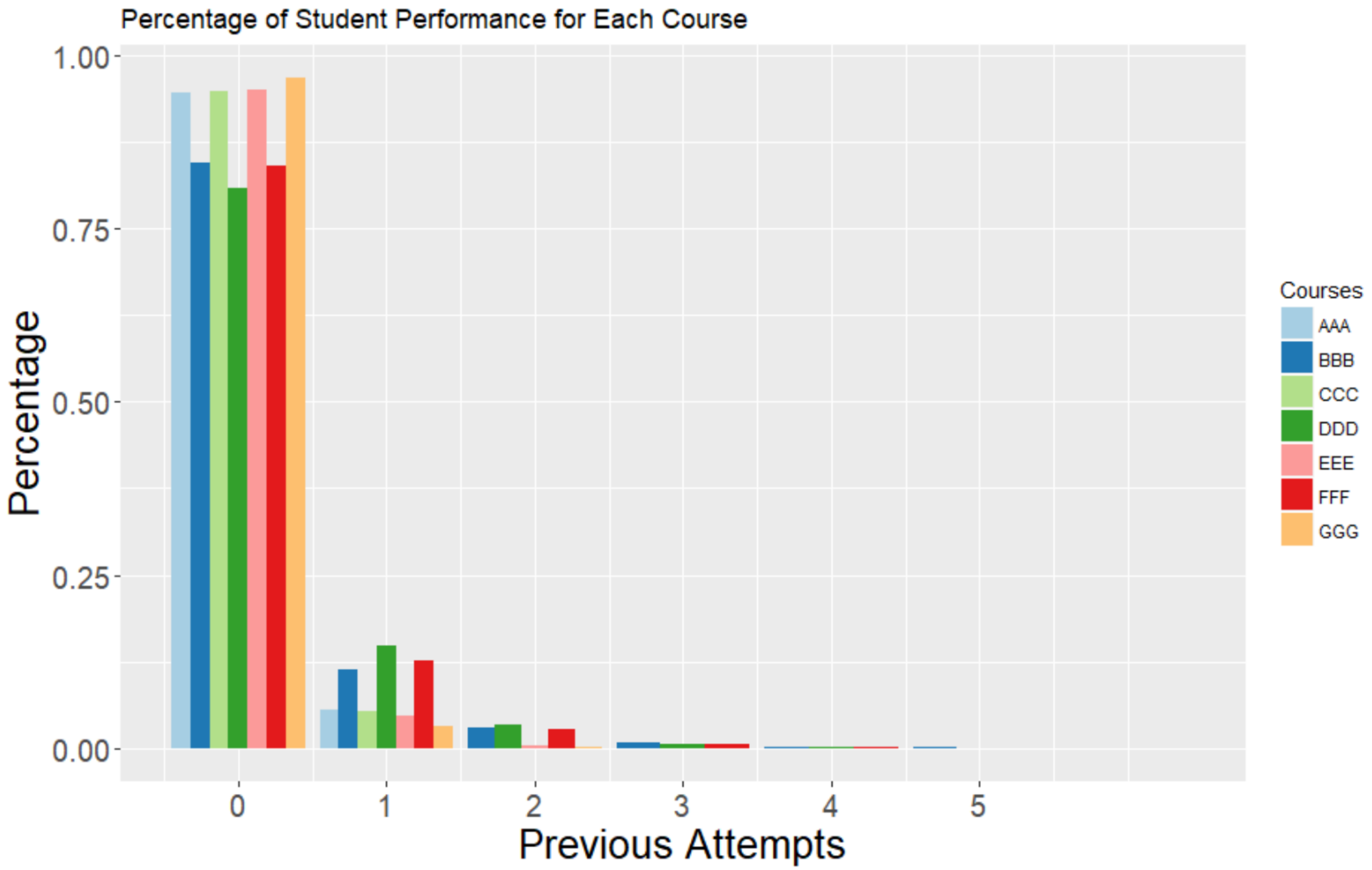
Thank you!

Feedback/Questions?

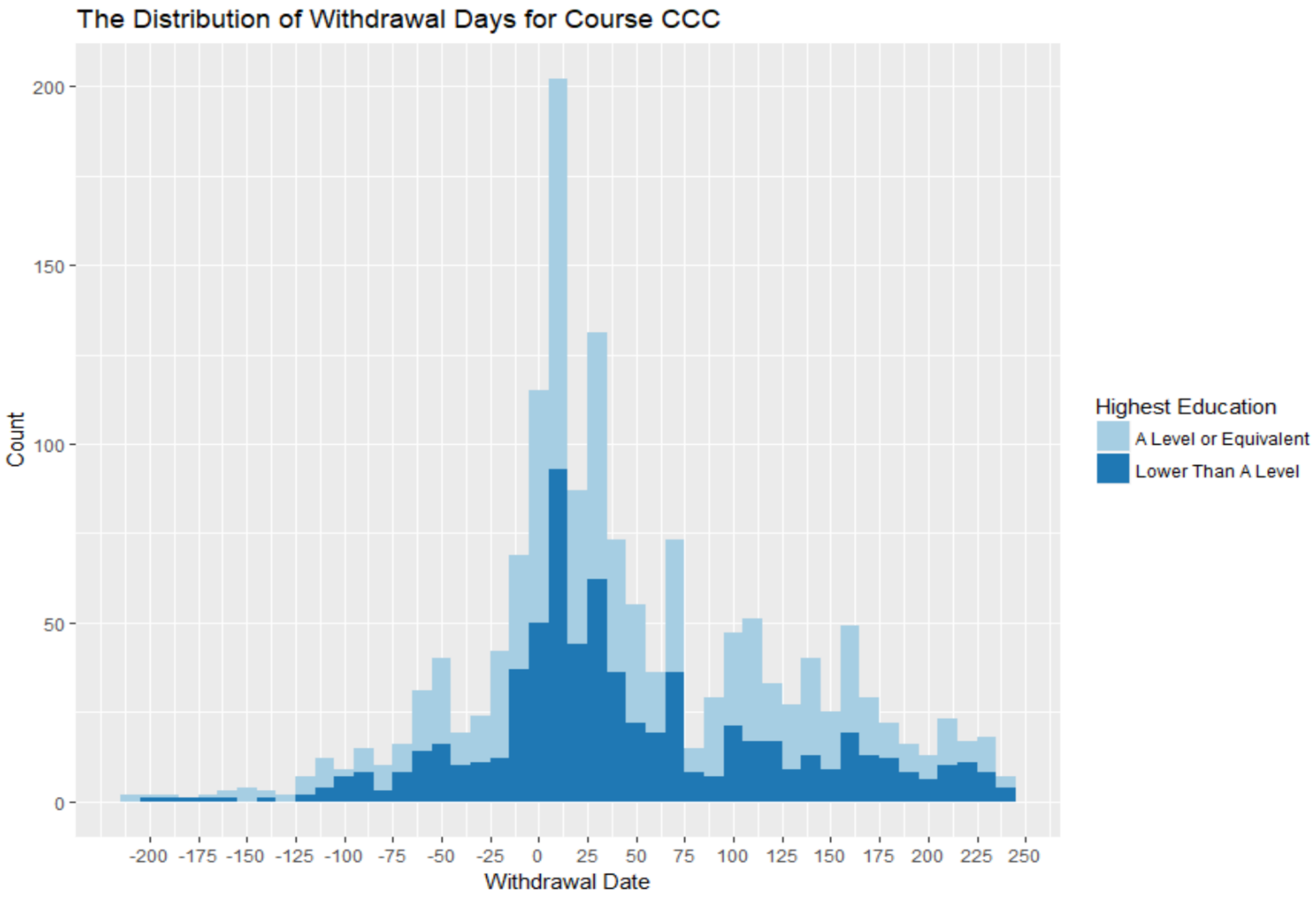
Appendix A


Percentage of Student Performance for Each Semester
Percentage
0.4
0.3
0.2
0.1
0.0
Distinction
Fail
Pass
Withdrawn
Final Result
Semester Start
2013B
2013J
2014B
2014J

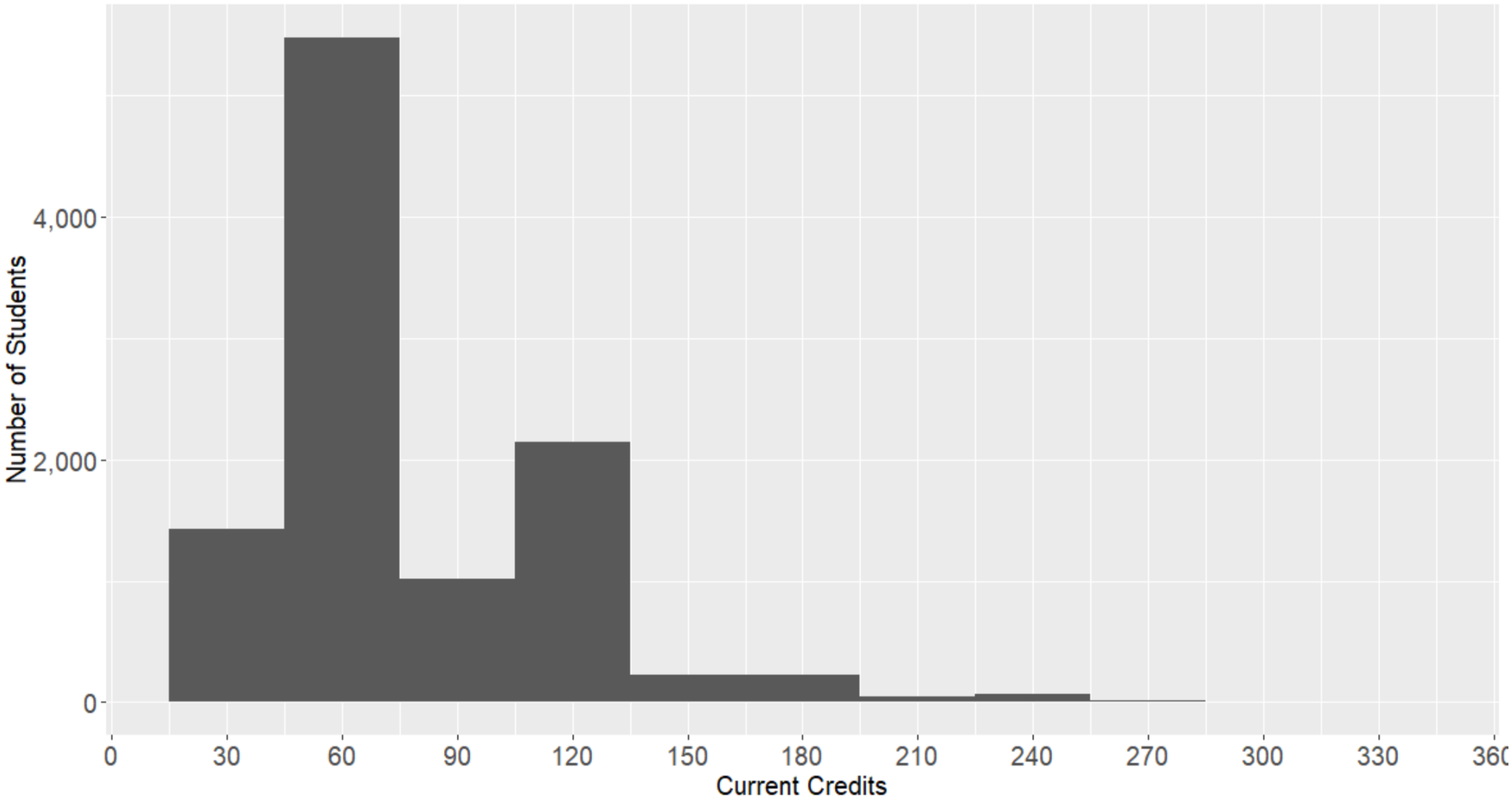
Appendix B


Percentage of Student Performance for Each Course
Percentage
1.00
0.75
0.50
0.25
0.00
0
1
2
3
4
5
Previous Attempts
Courses
AAA
BBB
CCC
DDD
EEE
FFF
GGG

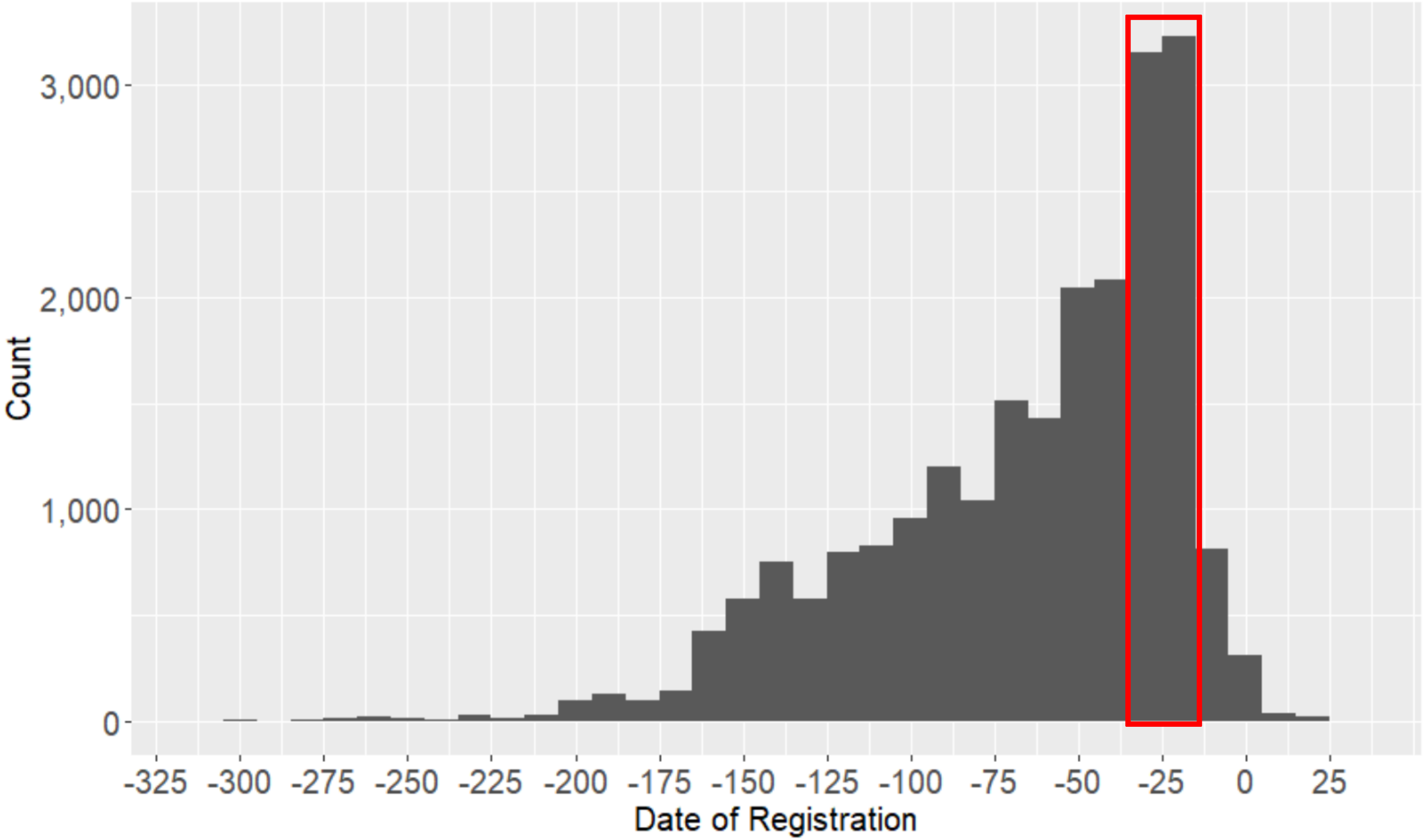
Appendix c


The Distribution of Withdrawal Days for Course CCC
Count
200
150
100
50
0
-200
-175
-150
-125
-100
-75
-50
-25
0
25
50
75
100
125
150
175
200
225
250
Withdrawal Date
Highest Education
A Level or Equivalent
Lower Than A Level

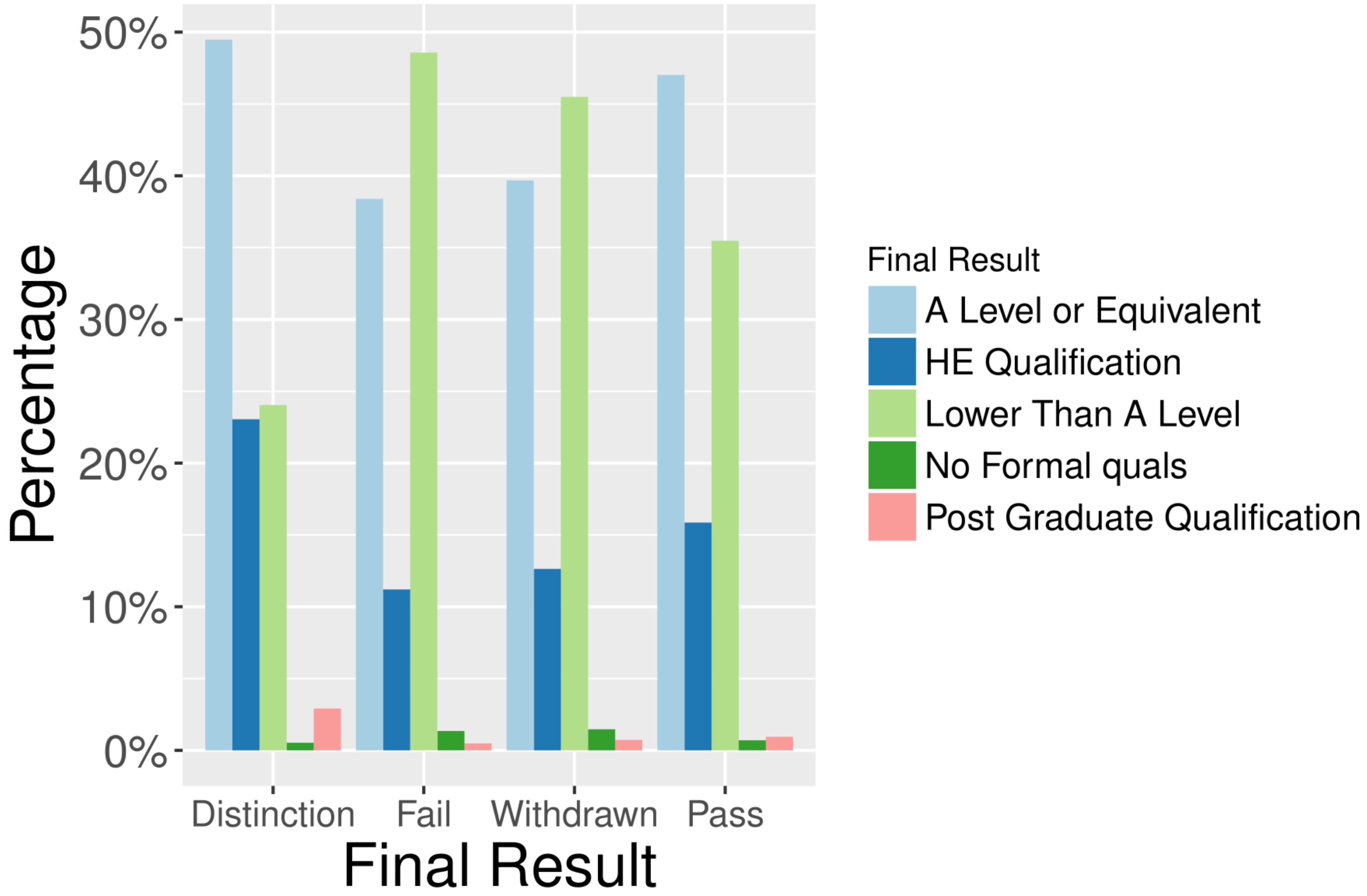
Appendix D: The Distribution of Credit Taking per Student for Semester October 2014



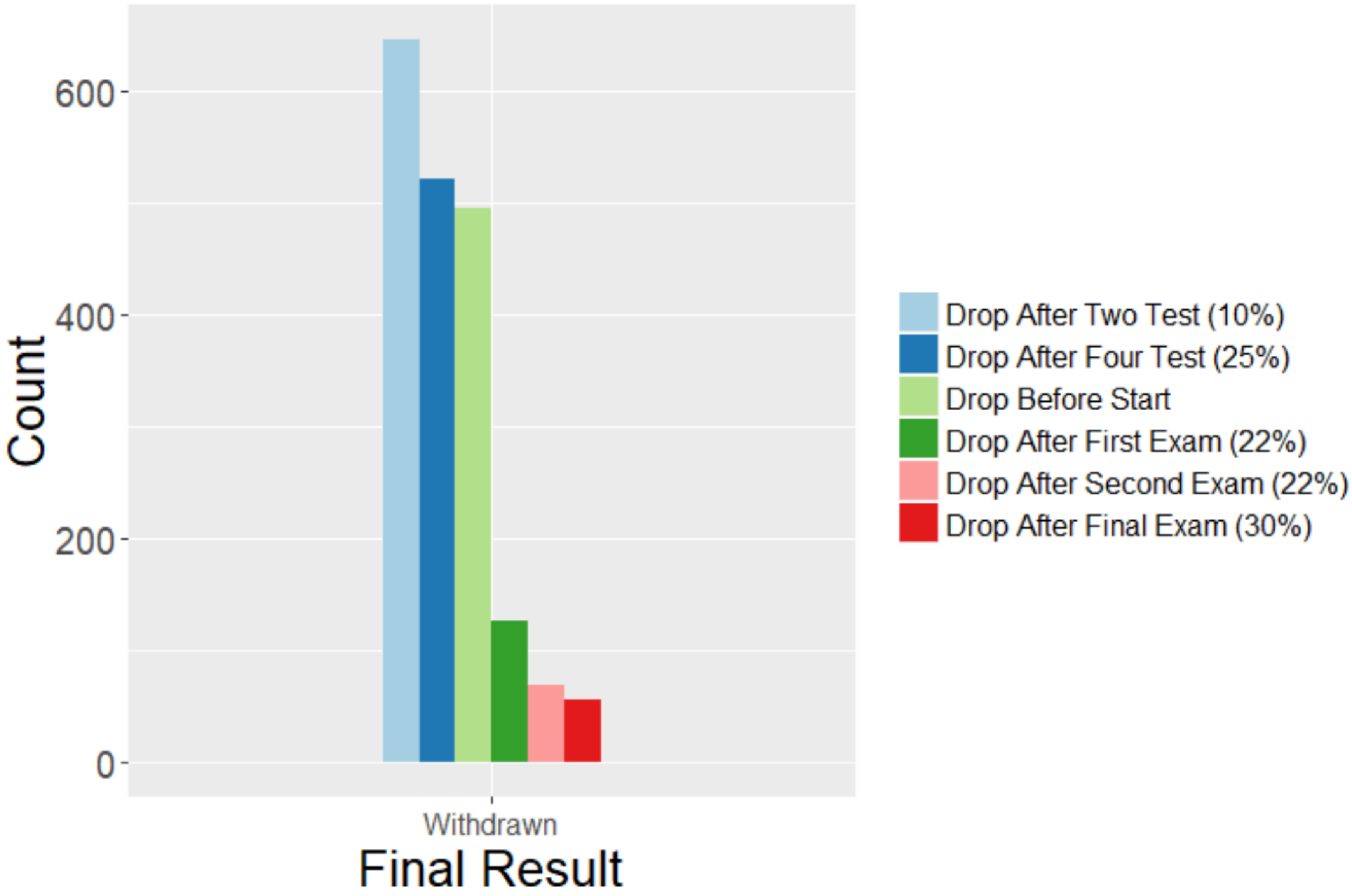
The Distribution of When Student Register the Courses



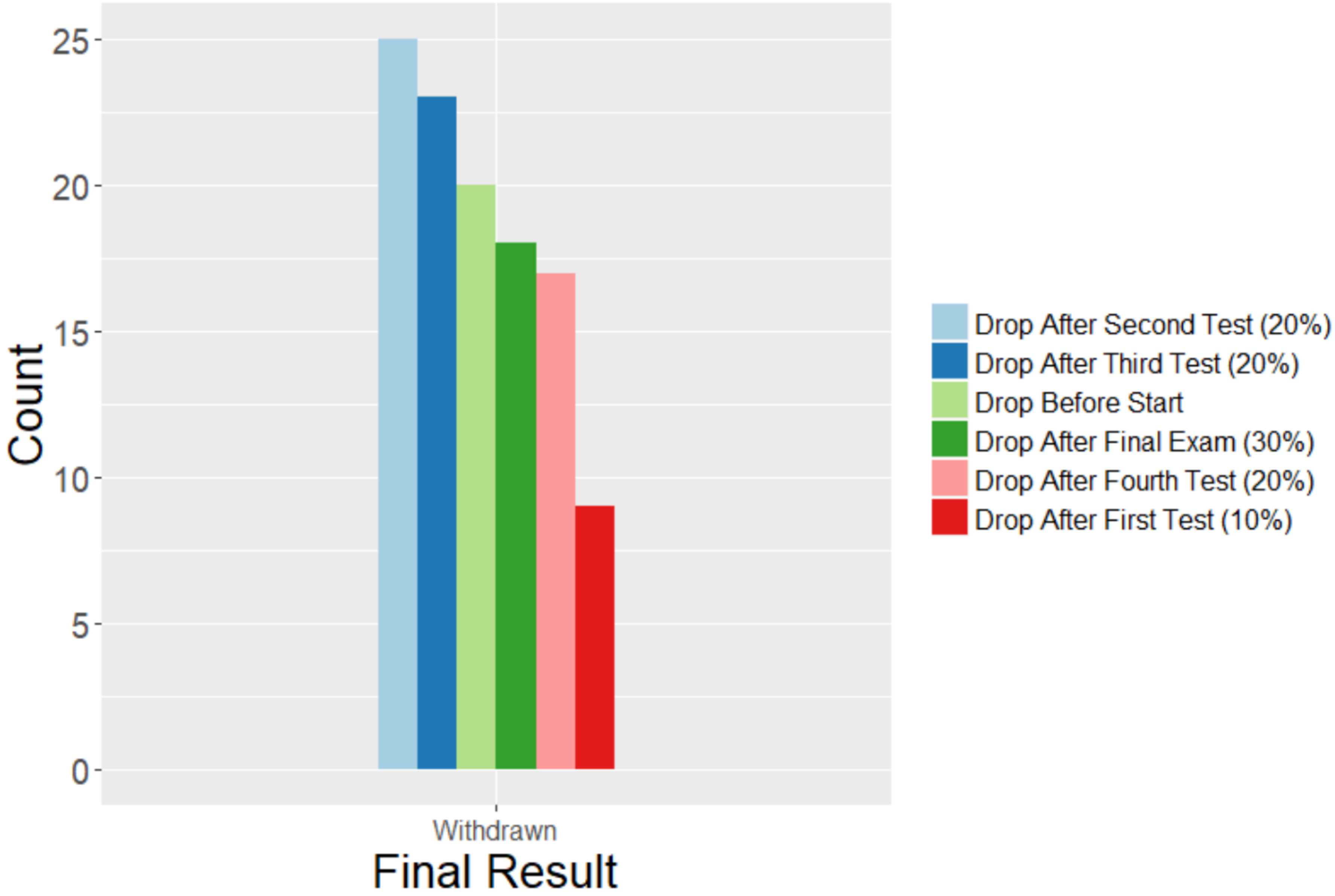
Student Education Background by Student Performance



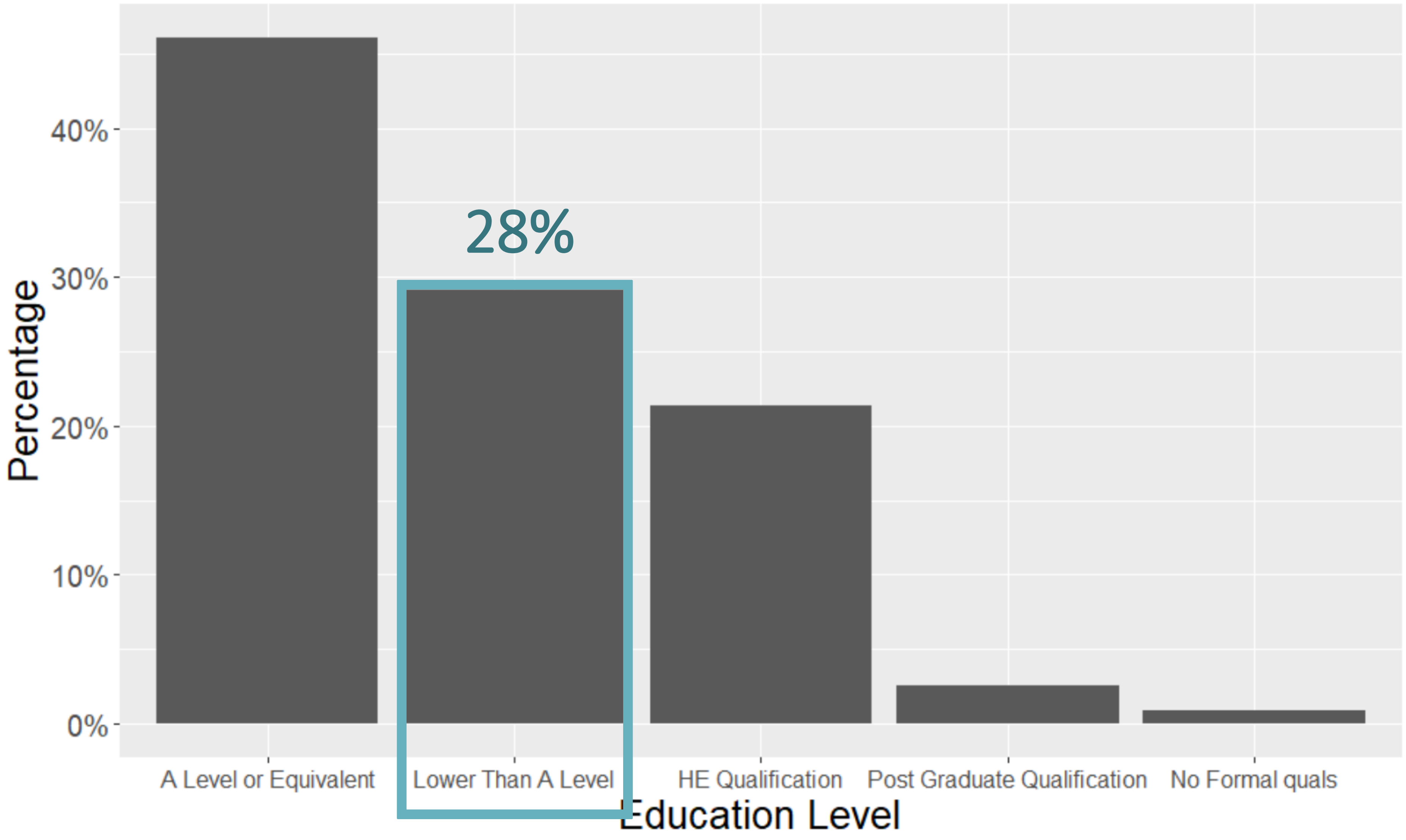
The Number of Student Drop After Test for Course CCC



The Number of Student Drop After Test for Course AAA



Student Education background in % for Course CCC



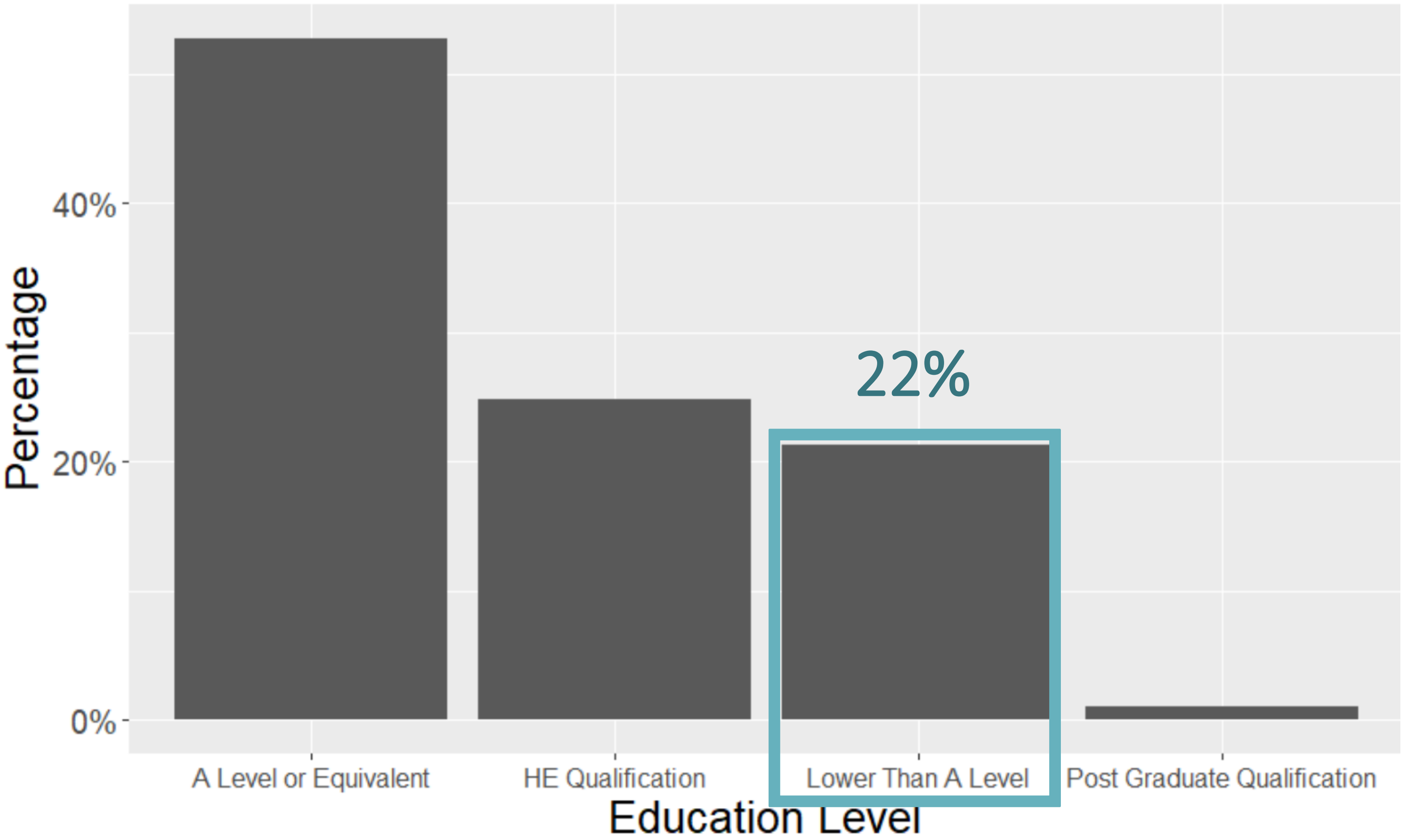
A Level: High School Graduate

Lower Than A Level: Middle School Graduate

HE Qualification: Bachelor’s Degree

Post Graduate: Master’s Degree or higher

Student Education background in % for Course AAA



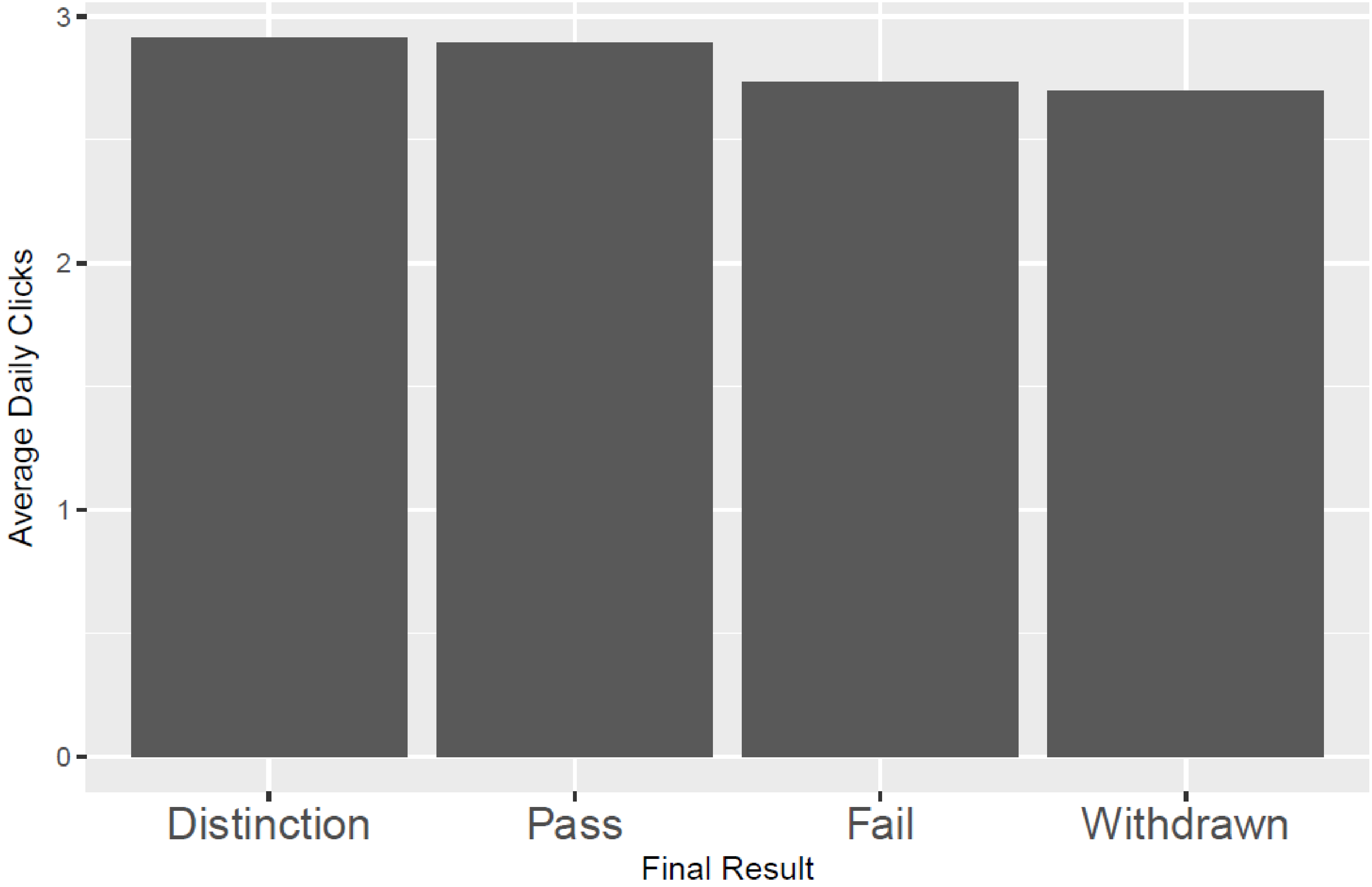
A Level: High School Graduate

Lower Than A Level: Middle School Graduate

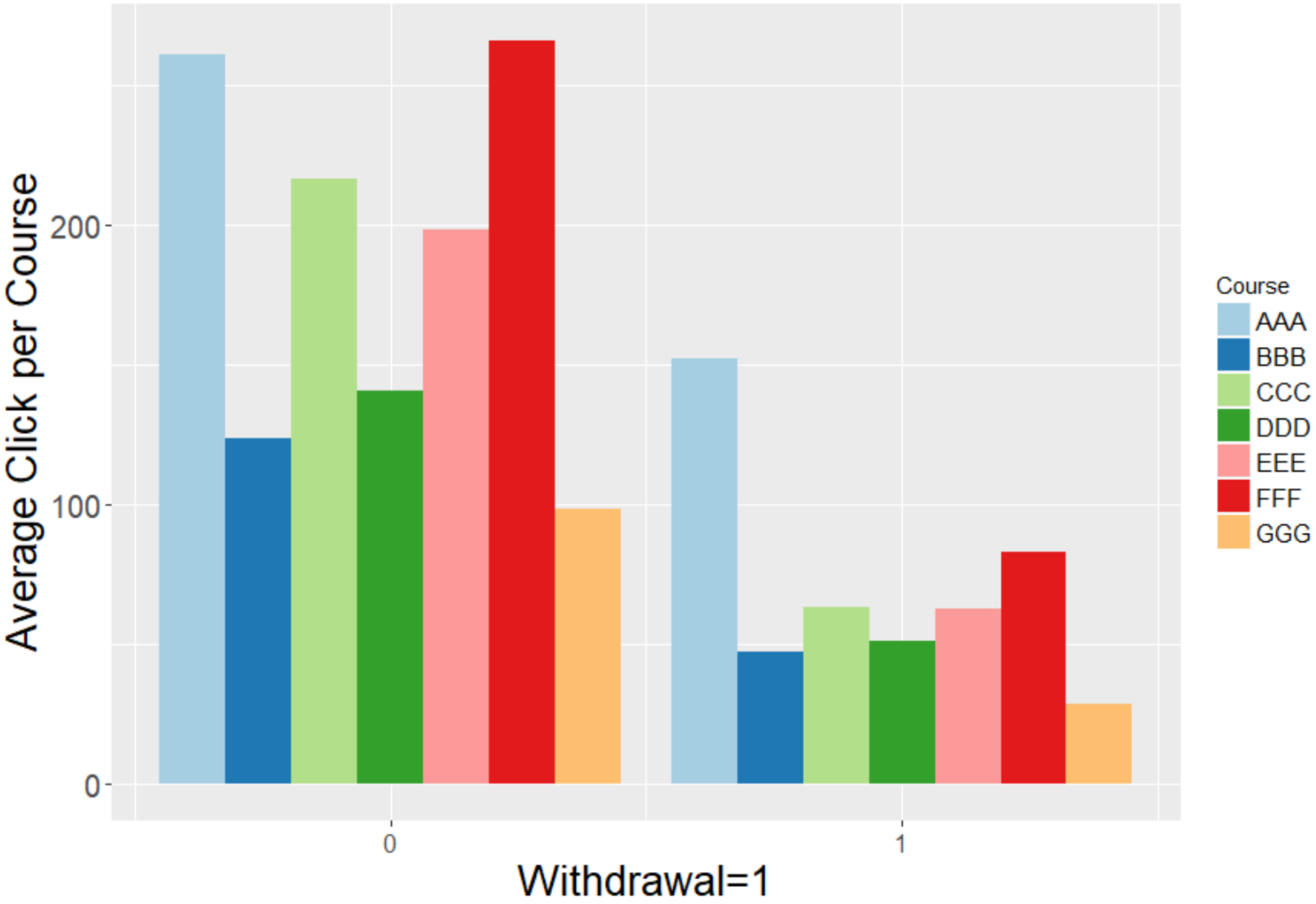
HE Qualification: Bachelor’s Degree

Post Graduate: Master’s Degree or higher

Average Daily Clicks between Different Final Results



The Average Clicks per Course Between Withdrawal or Not



Regression Analysis

Withdraw	Coefficients	Standard Error	t	p-value (P>t)	Confidence	Interval
Student Education (A level omitted)						
HE Qualification	-0.0001911	0.0069394	-0.03	0.978	-0.0137926	0.0134105
Lower Than A Level	0.0687588	0.0050924	13.5	0	0.0587774	0.0787401
No Formal quals	0.1287799	0.0231421	5.56	0	0.0834204	0.1741395
Graduate Qualification	-0.0194275	0.0229277	-0.85	0.397	-0.0643668	0.0255118
Course (AAA omitted)						
BBB	-0.0573664	0.0155173	-3.7	0	-0.0877809	-0.0269518
CCC	0.1871593	0.0161013	11.62	0	0.1556	0.2187185
DDD	0.0810767	0.015573	5.21	0	0.0505529	0.1116004
EEE	0.0296011	0.0164495	1.8	0.072	-0.0026406	0.0618428
FFF	0.1456321	0.0154433	9.43	0	0.1153626	0.1759016
GGG	-0.1369594	0.0171699	-7.98	0	-0.1706132	-0.1033056
Semester (2013B omitted)						
2013J	-0.0198419	0.0076611	-2.59	0.01	-0.0348581	-0.0048258
2014B	-0.0113292	0.0080602	-1.41	0.16	-0.0271275	0.0044691
2014J	0.0155069	0.0075429	2.06	0.04	0.0007224	0.0302914
Past Behavior						
Total Number of Clicks in VLE	-0.0000885	1.46E-06	-60.63	0	-0.0000913	-0.0000856
studied_credits	0.0011571	0.0000637	18.16	0	0.0010322	0.001282
num_of_prev_attempts	-0.0167268	0.0050063	-3.34	0.001	-0.0265394	-0.0069143