Exploring Indicative Features To Predict Stop-out in MOOCs

Han Dinh Antoine Pigeau

Ecole Polytechnique de l'Universite de Nantes

May 29, 2017

Contents

Abstract

With high popularity in massive open online courses (MOOC), there is an interest in exploring the reason why users frequently drop out of such courses, i.e., "stop-out". In this report, I will explore the indicative features that will most efficiently predict stopouts. Being able to detect a stopout before it occurs can expand our knowledge on MOOCs and how students learn online. In this report, I will build upon previous stop-out experiments by (1) find user features that have been previously tested on and strongly indicate stop-out; and (2) prove which features show discrimination between successful users and failures using our MOOC's, OpenClassrooms, massive database containing over 3 million users. This research only explores user features and their predictively. It does not perform experiments on detecting stop-out users.

Chapter 1

Introduction

Massive Open Online Courses (MOOC) are for anyone who wants to learn about a subject, whether it aids in their school/career, or if it's for their own personal knowledge. They're becoming increasingly popular to people around the world because it's conveniently accessed on the web, has little to no charges, the variety of topics are endless, and some offer licenses and certificates for users who succeed in the course requirements. There are numerous courses ranging from psychology subjects to learning computer languages. In result, each course has different teaching methods and success/fail requirements.

Although MOOCs are the ideal learning environment, there's an exceptionally high rate of users that begin the course, but do not stay active until the end of the course. In fact, a handful of students register for the course and quit soon after the first section. This rapidly growing community has inspired researchers to examine the cause of student ″stopout″. MOOC developers are concerned students stopout due to disliking something about the course, and how they can fix this. However, researchers question if students stopout because of external factors not pertaining to the MOOC, for example, heavy workloads in school. If we can automatically predict stopouts, we can possibly prevent them from occuring and have a higher percentage of students successfully completing online courses.

This introductory chapter will explain the relevancy of researching MOOCs and the details of our data source (The OpenClassrooms). Next, it'll discuss the intent of this project and, lastly, how the large amount of data is stored.

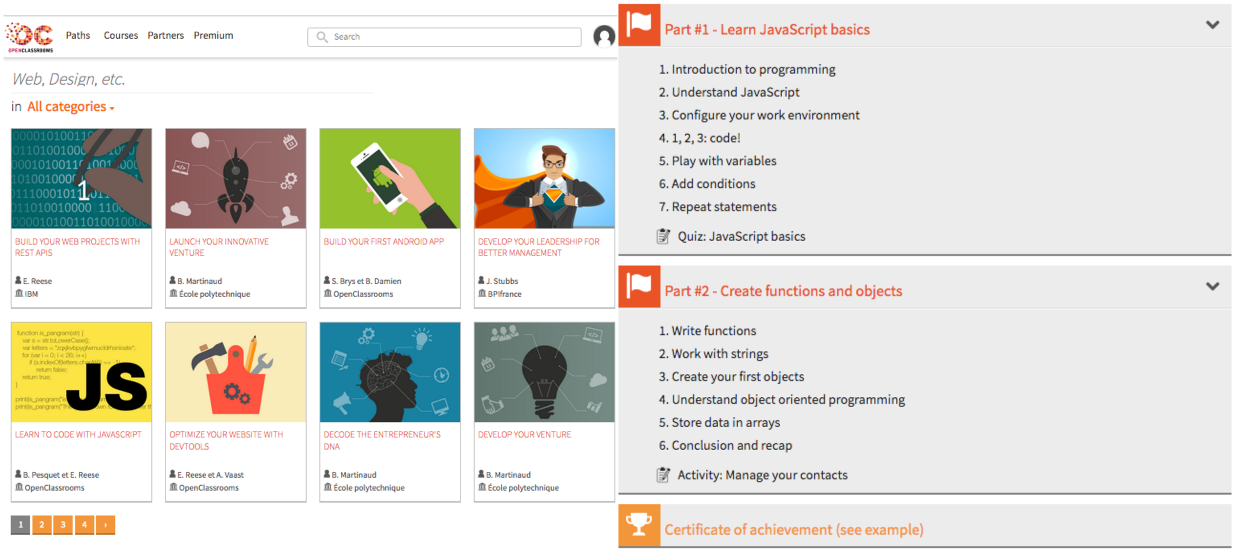
1.1 Overview: The OpenClassrooms

For my work, I will use the student data acquired from The OpenClassrooms. They are the leading digital learning platform in Europe, and over three million users per month take their technology and business related online courses. Their goal is to reach out to anyone with internet access and teach them in an intriguing manner to help them excel in their professional and personal lives. There's over 1,000 available courses including web/mobile development, designing, and entrepreneurship. In addition, since 2012, students have the option to obtain an official certificate in technology after passing a few years of courses.

Figure 1.1: The OpenClassrooms logo

1.1.1 Course Structure

The courses are created by teachers and field experts to ensure quality material. They are available online at anyday and any hour, and some are available for download as a PDF or ePub to view on tablets and e-readers.



BUILD YOUR WEB PROJECTS WITH REST APIS
E. Reese
IBM

LAUNCH YOUR INNOVATIVE VENTURE
B. Martinaud
École polytechnique

BUILD YOUR FIRST ANDROID APP
S. Brys et B. Damien
OpenClassrooms

DEVELOP YOUR LEADERSHIP FOR BETTER MANAGEMENT
J. Stubbs
BPIfrance

LEARN TO CODE WITH JAVASCRIPT
B. Pesquet et E. Reese
OpenClassrooms

OPTIMIZE YOUR WEBSITE WITH DEVTOOLS
E. Reese et A. Vaast
OpenClassrooms

DECODE THE ENTREPRENEUR'S DNA
B. Martinaud
École polytechnique

DEVELOP YOUR VENTURE
B. Martinaud
École polytechnique

(a)

Part #1 - Learn JavaScript basics

1. Introduction to programming
2. Understand JavaScript
3. Configure your work environment
4. 1, 2, 3: code!
5. Play with variables
6. Add conditions
7. Repeat statements

Quiz: JavaScript basics

Part #2 - Create functions and objects

1. Write functions
2. Work with strings
3. Create your first objects
4. Understand object oriented programming
5. Store data in arrays
6. Conclusion and recap

Activity: Manage your contacts

Certificate of achievement (see example)

(b)

Figure 1.2: In subfigure 1.2a is one view of the website's layout on a list of courses. Subfigure 1.2b shows a screen of the "Learn To Code with JavaScript" course's list of part titles.

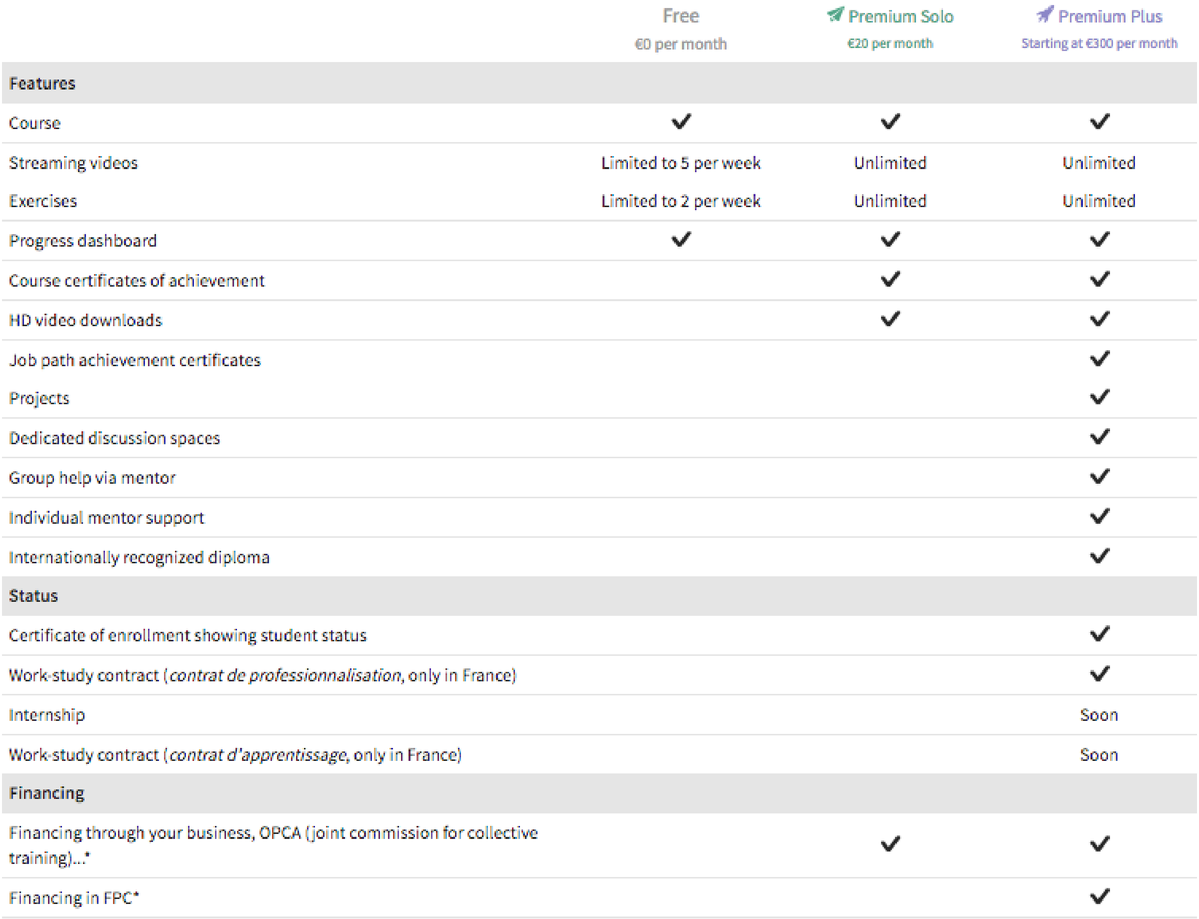
Each course contains one or multiple sections which contain text, visuals, and 3-4 minute (on average) videos. At the end of these sections, the user takes an automatically correct quiz. However, at the end of each part of certification courses, the user must pass the automatically corrected quizzes and the peer-evaluated assignments (See Figure 1.2a). There are two types of exercises: quizzes and peer assessments. A final average mark of the course quizzes determines whether the user passes or fails. Peer assessments are double-blindedly graded using a rubric made by the instructor, and then average amongst multiple students' evaluations.

For extra assistance with the material, OpenClassrooms has discussion spaces already containing millions of past users' FAQ. For Premium Plus members (defined later), private teachers are available for users that need regular mentoring; usually one session is 45 minutes to an hour per week.

Users that complete all course exercises with a final score above or equal to 70% will receive a certificate. Certificates are well-respected by companies in the field which will increase students' employment opportunities.

1.1.2 Definition of Premium User

The OpenClassrooms offers two premium membership options which differs from the free users for several reasons. Basically, premium members pay a monthly fee to have access to the certificate-earning courses and several more learning-aids, e.g., unlimited access to streaming videos and individual mentor support (see Figure 1.3 for more details).



	Free €0 per month	Premium Solo €20 per month	Premium Plus Starting at €300 per month
Features			
Course	✓	✓	✓
Streaming videos	Limited to 5 per week	Unlimited	Unlimited
Exercises	Limited to 2 per week	Unlimited	Unlimited
Progress dashboard	✓	✓	✓
Course certificates of achievement		✓	✓
HD video downloads		✓	✓
Job path achievement certificates			✓
Projects			✓
Dedicated discussion spaces			✓
Group help via mentor			✓
Individual mentor support			✓
Internationally recognized diploma			✓
Status			
Certificate of enrollment showing student status			✓
Work-study contract (*contrat de professionnalisation*, only in France)			✓
Internship			Soon
Work-study contract (*contrat d'apprentissage*, only in France)			Soon
Financing			
Financing through your business, OPCA (joint commission for collective training)...*		✓	✓
Financing in FPC*			✓

Figure 1.3: A comparison of the three different subscriptions to OpenClassrooms: free membership, Premium Solo, and Premium Plus.

All premium users can receive a Certificate of Achievement which is claimed to be recognizable across many nations and approved by OpenClassrooms' many prestigious partners, such as, Google and Microsoft. The material taught in these particular paths can be closely compared to the teachings in universities. This certificate is a respected asset to anyone's CV.

1.1.3 Definition of Session

The OpenClassrooms records the time and date each time the student is actively working on their course. A list of temporally continuous user clickstream is stored in the database called a **session**. Its a timespan where the user works continuously on the website. There will be more details on sessions in the Data Input Section 1.3.

1.2 Project Objective

The purpose of this study is to find features that can be used for future experiments to accurately predict student stopout within the OpenClassrooms website. I will approach this goal by exploring features that show discrimination between successful users and the failures. Once I discover features that set the users apart, it'll be simple to detect that a user with ”failing” features will eventually stop-out. If it is possible to predict stop-outs, then it potentially can reveal many useful things about how students learn online, the structure of various online courses, and, most importantly, how to prevent students from quitting.

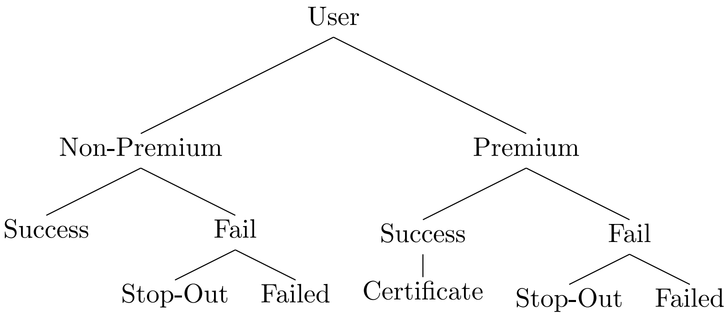
1.2.1 Goals

The goal is to find indicative features that will determine whether a student will succeed or fail a course. I will extract necessary data collected by OpenClassrooms, then I'll generate graphs to observe behavioral patterns in students that pass and fail. Next, I will describe the process of this research. It will specifically define the types of users on our MOOC, explain why I'll research features, then how to interpret the findings.

User Groups

Our definition of a **failed user** can be distinguished in two ways: one who registered for a course and finished the course with a final grade lower than 70%, and the other is one who registered for a course, showed at least one event, but stopped making action from a point during the course until the end, i.e., "stopped out". In contrast, a **successful user** is one who registers for the course, successfully completes all course requirements, and receives a final grade of 70% or above at most 1 month and a half after the end of the premium registration.

All the users are divided into 4 main *groups*: successful premium, failed premium, successful non-premium, failed non-premium.



Groups of Users To Be Tested

Figure 1.4: This tree shows the specific division of users that I will use to experiment. The groups are associated with the user's subscription type (Premium/Non-Premium) and whether they succeeded (Success) or failed (Fail). Each Fail node has two children because there are different ways to "fail" a course. The "Stop-Out" users are those who do not prevail until the end of the class, and the "Failed" users are those who completed the course, but did not receive above 70% as their final mark.

Explore features

I will need to do extensive research of scholarly articles, reports, and passed experiments related to MOOC stopouts. Although this is a fairly new research topic, there are numerous resources on the web, so my research on features will be ongoing throughout this project. I'll record each source's top indicators that have already been tested and compare them by prediction levels (high, average, low). Each MOOC that I've researched vary greatly in their teaching methods and requirements (quizzes, homework, videos, forum threads, wiki answers, and/or certificates), so not all features will not be applicable to all MOOCs. I will provide a large table of 33 features along with their descriptions, to give other researchers an opportunity to explore features that fit the MOOC in their research. Furthermore, I will determine which features can be used given only in OpenClassrooms' database for testing purposes.

Feature Interpretation

To expand on my related-works research, I'll compile graphical figures from only the user features that can be applied to OpenClassrooms' database. The graphs will be variations of histograms and box-plots displayed side-by-side for simple comparisons between successful and failed users. This way, it is easy to visually interpret the differences. Next, I'll provide a deeper analysis about the features and give reasons to why they show high discrimination between the successful users and the failures.

Lastly, a selection table will be displayed to finalize my research findings of which features are most discriminative of success and failed users in OpenClassrooms online courses. This information then can be used by future researchers to conduct experiments on predicting stop-out in OpenClassrooms, or can be adjusted to fit any other MOOC.

1.3 Data Input

The OpenClassrooms' large database offers extensive information in the form of tables about user's final grades and session durations, for example. It's important to discuss in detail what values the database contains to ensure that its possible to test the selected features.

Listed below is a summary of the data tables continuously being updated by OpenClassrooms. There is plenty of useful data that I'll be able to apply when testing various user features. Session and session exercise are two of the most important ones to take note of. Not all the values listed will be used in my experiments, but my project requires some values which will be used to apply the features to our users on OpenClassrooms.

1. **User**: holds descriptions of the user

   user ID, gender, birth date, city, country, region, zip code

2. **Exercise**: contains details of each quiz and exercise in the courses

   course ID, part ID, exercise ID, type, active

3. **Session exercise**: shows the details of each user's quiz taken and session exercise completed

   session ID, exercise ID, user ID, create date, complete date, score

4. **Premium**: stores when a user begins/ends a premium subscription

   user ID, start date, end date

5. **Session**: contains all the details of a session a user takes throughout their course(s)

   course ID, part ID, session ID, user ID, duration, start date, end date

6. **Visualization**: contains all the sessions of a user in a course

   course ID, session ID, user ID, date

7. **Follow course**: contains the date and time a user begins following a course

   course ID, user ID, date

8. **Unfollow course**: holds the date and time a user stop following a course

   course ID, user ID, date

9. **Result**: stores the user's final grade and whether they passed or failed the course

   course ID, user ID, score, date, pass/fail

For a deerper understanding of the database scalability and usability, see Table 1.1. It presents the number of users per group for each course. The table is divided in 3 parts: all users, premium users and non premium users. Each part contains the number of success, failing, dropout and failing with mark for each course. Note that a succeed premium user is a user that passes the final exam at most 1 month and a half after the end of the premium registration (that is why the number of succeed user plus the number of succeed non premium is not equal to the number of succeed user - some user succeed the final exam after our one month and a half boundary).

All	#	# Succeed	# Failing	# Dropout	# Failing with mark
Bootstrap	13045	796	12249	12237	12
Ionic	2020	46	1974	1973	1
Gestion Projet	3156	671	2485	2471	14
Rubys	306	5	301	299	2
Web	11793	3777	8016	8013	3
Twitter	1559	370	1189	1181	8
Arduino	4864	115	4749	4746	3
API REST	0	0	0	0	0
JavaScript	12829	1851	10978	10961	17
HTML5 / CSS3	0	0	0	0	0
Premium	#	# Succeed	# Failing	# Dropout	# Failing with mark
Bootstrap	2441	626	1815	1773	42
Ionic	397	25	372	363	9
Gestion Projet	1581	555	1026	980	46
Rubys	68	4	64	61	3
Web	4512	3114	1398	1296	102
Twitter	667	305	362	341	21
Arduino	533	55	478	471	7
API REST	0	0	0	0	0
JavaScript	4068	1413	2655	2532	123
HTML5 / CSS3	0	0	0	0	0
Non Premium	#	# Succeed	# Failing	# Dropout	# Failing with mark
Bootstrap	10604	138	10466	10464	2
Ionic	1623	13	1610	1610	0
Gestion Projet	1575	82	1493	1491	2
Rubys	238	0	238	238	0
Web	7281	562	6719	6717	2
Twitter	892	51	841	840	1
Arduino	4331	54	4277	4275	2
API REST	0	0	0	0	0
JavaScript	8761	328	8433	8429	4
HTML5 / CSS3	0	0	0	0	0

Table 1.1: Number of users per group for each course. From top to bottom, each sub-table represents respectively all the users, the premium users and the non premium users.

Conclusion

The OpenClassrooms is Europe's leading online learning platform. I want to find specific user features that can predict stopout in MOOCs. I will utilize previous works on online learning behaviors to discover old features that were indicative in the past and possibly propose new features that are worth testing. Then, I'll examine OpenClassrooms' database to determine which features are possible to experiment.

The next chapter will discuss past research and experiments on MOOC stopouts and online learning behaviors. I will explain why testing various features are essential to predicting stopout. It'll describe in detail several articles'/reports' indicators and the results of testing them. My overall findings will be summed up in a table to view easily. These related works will cover the feature exploration stage that will lead to a solid proposal.

Chapter 2

Bibliography

Researchers from MIT advise future experimenters to concentrate on exploring new features because it is crucial to prediction [9]. Specifically, they found after their MOOC experiments that using numerous modeling techniques, or creating behavior co variations contribute less to stopout prediction than exploring features.

This chapter is intended to provide an exploration of generalize features over all types of MOOCs. I will research numerous features for experiments in my context, and also for possible future experiments with MOOCs of different passing/failing criteria.

In this chapter, I will discuss my findings of past research on MOOCs, but focusing on the user features used to predict stop-outs. Find a summary of the features in Section 2.1.7. Lastly, I will filter the features down to ones that are applicable to OpenClassrooms' database.

2.1 Exploring Features

This section will cover my research findings of student behaviors on different MOOCs. I will explore related studies, but I'll focus on the indicators that may help us predict stopout. Some features include homework grades, clickstream data, user demographics, etc. Then, I'll summarize and gather my findings into this section. Since there are some similar features in multiple reports, I split them into related subsections: video related, user collaboration, grades & assessment, interactivity, and demographics. Please note that these features are not generalized. They're created based on the experiment's specific domain.

2.1.1 Video Related

It's a common interest to examine users' behavior in watching course videos. In article [2], they found that users who watch 25%-50% of video lectures are extremely unlikely to dropout.

In contrast, a different report [4] explored video features, and interestingly found that students who often skipped through video lectures tended to complete the course with exceptional grades. This may be explained by students who have previously learned about the subject and are searching for advanced information to further their understanding.

Another report [6] found that the video-skip feature was a strong predictor only when the particular course contained a few, shorter videos per week. Otherwise, students tend to skip through or entirely skip videos when there is more than two hours to view that week.

There are several features regarding videos in article [8]. These experiments included the total number of sessions containing videos, average number of pauses during a video, total time a user spends watching videos, percentage of sessions including video activity, and percentage of a video the user doesn't skip and actually watches. They saw that users' final scores on homework were directly affected by video watching. In fact, the percentage of videos watched feature ranked as the second-most indicative. Overall, their results revealed that all the video-related features have a large impact on overall performance.

2.1.2 User Collaboration

Researchers predicting stopout in [9] divided students into four cohorts and explored several features. Each registered user was assigned to a cohort based on their participation levels: passive collaborator, wiki contributor, forum contributor, and fully collaborative. Interestingly, they found that certain models on some cohorts worked better than others. However, the models aren't as important as selecting indicative features. Experimenters discovered that the most predictive features pertain to problem submission, e.g., submissions per correct problem, average time submission before deadline, average number of submissions. Also, the average length of user's forum posts is a helpful indicator, while the wiki edits feature is not.

As well as the video feature, experiments in [2] explore the number of forum posts submitted and number of threads viewed on the forum. Both features weren't as indicative as the video-skip feature, but the number of threads viewed may be strongly correlated to stopouts.

2.1.3 Grades & Assessments

As mentioned before, article [6] explores video-skip features, but also more: assignment performance, assignment-skip, and lag. Firstly, users' assignment performances, i.e., quiz grades, successfully predicted 95% of stopouts. However, quiz scores averaged 70% - 80%, so this may be the result of the assignments' levels of ease or that students choose to enroll in courses they have prior knowledge about. Secondly, the assignment-skip feature was predictive on courses with low work loads. Nonetheless, it was not as effective as the video-skip feature because researchers assume there is a large cohort of students interested in only viewing content rather than participating in assessments. Lastly, the lag feature was included to observe if students stopout when they fall behind on material. It tested with high recalls, especially in courses containing strongly linked sections.

Users' normalized grades were tested in report [10]. Experimenters wanted to see at which point in the course the users' grade could be determine stopout. They defined a threshold between certificate-earners and failed users. Tests revealed that the normalized grade is the second most indicative feature.

In article [8], studies included quiz and homework assessments. Similar to [6], this article tested average attempts at a quiz and the number of quizzes a student completes prior to a homework. Both of these factors showed an impact on predicting stopout, but the number of quizzes a user takes prior to a homework was the top feature throughout the entire study. Researchers suggest that it's beneficial to observe how much a user studies to prepare for the homework assessment.

Researchers were quite interested in exploring homework features, so they tested others, such as, the time between homework submission and most recent quiz attempt and the timespan from a homework submission to the most recent video activity. These features revealed to be some-what important to predicting stopout.

Other interesting features in [8] involved total number of homework-related activity and the homework save button which allows users to save a homework problem and return to it, perhaps after reviewing the material or for a more convenient time. They averaged the number of times the user used this button on each homework assignment. After experiments, researchers concluded that these two features do not have importance in stopout indication.

2.1.4 Interactivity

Three of five top features from report [10] are related to user interactivity. These researchers tested elapsed time since the last recorded action, total number of all events, and the absolute time persisted in the course. This shows features dealing with frequent activity are indicative of student stopout. Interestingly, the results showed that the top predictive feature was the time since last recorded event.

Another strong feature in [2] is how often the student checked the course progress page. The experiments showed that it is directed correlated to stopout and should be used in prediction. A user who never checks their course progress is most likely to stopout from falling behind on material, has lost interest, or is busy with external responsibilities.

Article [8] studied user activity in-between two consecutive homework assessments. These features may reveal certain patterns in behavior. With the user being aware of their previous homework grade, it may affect their study habits and/or their next homework grade. These activities included number of videos watched, average number of daily sessions, number of quizzes, average number of attempts on quizzes, and percentage of login days. Their results suggests that observing these features are not important in predicting stopout.

These researchers [8] also tested a few features concerning students' "study sessions", or a period of time a user is logged in and shows continuous activity. The time spent studying may have an impact on a student's performance. They experimented with the average number of sessions per day prior to a homework attempt, average time of each session, and the average number of logins. These three features showed no effect in their tests. The results suggest that the amount of time physically spent on the course website is not indicative of predicting stopout.

2.1.5 Demographics

Bayeck in [3] studies the roles that gender plays on MOOCs. They mention that in previous studies, online courses were male-dominant, but after their study, they found that 60% of users were female. This surprising female-dominance could have resulted for many reasons, such as, it was promoted as a group assignment and females prefer collaboration even within online settings.

A different report [5] also studied the correlation between user demographics and online learning behaviors. However, these researchers found that out of their four offered courses, mostly men participated in them. Furthermore, the highest ratio was apparent in the computer science (artificial intelligence) course with 86% of male students. Their tests also proved that age is positively correlated with breath of coverage, i.e., the older the user, the more course material they use.

Demographics may be useful for predicting stopout, but OpenClassrooms does not provide sufficient demographic information. Therefore, I will not incorporate these features in this experiment.

2.1.6 Other

In [1], researchers looked at MOOC's different individuals' behavior patterns (high- versus low-achieving students). They found that users' course registration times can sometimes predict how active and involved the student is throughout the course. Surprisingly, students who show no activity usually register up to six months early, or after the course. In other words, students who tend to register during or after the course start date mostly watch lectures and turn in a few to no assignments. The mixed results show that course registration time is not an indicative feature to stopout.

2.1.7 Table of Features

After research of several scholarly reports and articles, I've created a table that helps summarize my findings:

	Feature	Description	Level	Source
1	Time since last event	Time between a particular time t & the last recorded action	High	[10]
2	Normalized grade	Grade in the course relative to the passing threshold	High	[10]
3	Gabor filters	Measure of persistence in user event logs	High	[10]
4	Time into course	Absolute time the user survived through the course	High	[10]
5	Total # events	The total # of recorded actions	High	[10], [7]
6	Pre-deadline submission time	Duration between when student begins problems & its due date	High	[9]
7	Visits to progress page	User's frequency in visiting the course progress page	High	[2],
8	Avg lab grades	Avg of all interactive lab assignment grades relative to other users	High	[9]
9	# of forum thread views	Frequency the user views forum threads per week	High	[2]
10	Avg # problem/quiz attempts	Avg # user attempted problems/quiz	High	[9]
11	% of lecture videos watched	Amount of time user spends watching video lectures per week	High	[2],[8]
12	Avg length forum post	Avg # of lines of forums posted	High	[9]
13	# of forum posts made	Amount of posts made to the form weekly	High	[2]
14	Avg homework grades	Avg of all assignment grades relative to other users	Average	[9]
15	% Video skip	Amount the user skips through the videos (not skipping entirely)	High	[4],[6],[8]
16	Registration time	When (before/during/after) user registers for course	Average	[1]
17	NumSession	Average # of study sessions per day before a homework attempt	Low	[8], [7]
18	AvgNumLogin	# of ”work day” over the (# of ”work day” + # of ”rest day)”	Low	[8]
19	NumQuiz	# of quizzes taken before a homework attempt	High	[8]
20	VideoNumPause	Average # of pauses per video	Low	[8]
21	HWProblemSave	Average times ”save answer” button actions per assessment	Low	[8]
22	TimeHwQuiz	Time from a homework submission to most recent quiz attempt	Average	[8]
23	TimeHwVideo	Time from a homework submission to most recent video viewed	Average	[8]
24	TimePlayVideo	# of sessions containing video activity relative to total # sessions	High	[8], [7]
25	HwSessions	Total # sessions including homework-related actions	Low	[8], [7]
26	IntervalNumQuiz	Amount of quizzes taken between two homework assessments	Low	[8]
27	IntervalQuizAttempt	Average # of quizzes taken between two homework assessments	Low	[8]
28	IntervalVideo	# of videos watched between two homework assessments	Low	[8]
29	IntervalDailySession	Average # of daily sessions between two homework assessments	Low	[8]
30	IntervalLogin	% of login days between two homework assessments	Low	[8]
31	Avg # requests per session	Total # requests over total session	Unknown	[7]
32	Avg Timespan	Average time between two consecutive clicks	Unknown	[7]
33	Peer evaluation	Total # requests involving Peer Evaluation	Unknown	[7]

Table 2.1: From my research, these features are indicators used in the past to study MOOC stopouts. The features are listed by name, description, level of indication (high, average, low), and citation. Further detail and descriptions are explained in prior subsections.

2.2 Feature Testability

The OpenClassrooms' database has abundant information on students and courses, however, some features require different data that isn't recorded or given in our specific database. This section will discuss each of the features and the reasons to include, or to not include it in my research.

For each feature from the previous section, I will first give a detailed explanation of each feature and its testability within OpenClassrooms' database. Then, a simple table will summarize and simplify my explanation. I'll describe in detail the features from Table 2.2. I chose features that are appropriate for our data from OpenClassrooms' and also that could be interesting to explore.

1. **Time since last event**: I transformed this feature because OpenClassrooms' database doesn't record clickstream, but does contain sessions that the user is active during the course. *Time since last session* is the time elapsed between a user's 1-week time mark and the session immediately preceding. The 1-week mark is defined by 168 hours after the user's first session in the course.

2. **Normalized Grade**: this feature will be transformed into *final mark* because in OpenClassrooms, this mark is derived from all a user's course assessments including quizzes.

3. **Gabor filters**: this feature may be applicable to our experiments, but will not be used due to the complexity of calculating Gabor Filters.

4. **Time into course**: since OpenClassroom course do not have set course start and end dates, this feature will be tweaked to *total course duration.* It will be shown by the user's accumulated chapter and exercise session times spent on the course.

5. **Total # of events**: OpenClassrooms refers to a user's activeness as "sessions", so this feature has transformed to *total # of session.* Sessions may include any time spent on the course online platform, such as, studying the material or completing an exercise.

6. **Pre-deadline submission time**: since OpenClassrooms' does not have a specific submission deadline, this feature cannot be applied to our experiments.

7. **Visits to progress page**: this feature cannot be tested because there is no recorded data in my database from OpenClassrooms' progress pages.

8. **Average lab grades**: is a non-applicable feature because users do not complete labs or similar activities in OpenClassroom courses.

9. **# of forum thread views**: this feature will not be used because our database doesn't contain any forum thread information.

10. **Average # problem/quiz attempts**: will be applied to the OpenClassrooms' quizzes taken after each chapter. Since there is no limit to taking quizzes, this feature will be calculated by the *average # of quiz attempts.* It may be interesting to view how many times successful and failed users try a quiz.

11. **Percent of lecture videos watched**: this is a non-applicable feature because OpenClassrooms doesn't give us any data on video watching.

12. **Average length forum post**: this feature will not be used before our database doesn't contain any forum post information.

13. **# of forum posts made**: this feature will not be used because our database doesn't contain any forum post information.

14. **Average homework grades**: I will merge this feature with feature #2 because a user's final mark in the average of course exercise results.

15. **Percent video skip**: I do not have data to calculate this video features, so it cannot be experimented with.

16. **Registration time**: is a non-applicable feature because user can begin a course in OpenClassrooms' at anytime, i.e., there is no course begin date.

17. **NumSession**: this feature will be analyzed for discrimination, but to fit our context, I will test the *average number of sessions before a quiz attempt.* I want to analyze how much a user prepares or studies for a quiz prior to attempting it.

18. **AvgNumLogin**: is a non-applicable feature because the given database doesn't store the number of user logins.

19. **NumQuiz**: is a non-applicable feature because OpenClassrooms' doesn't have homework attempts.

20. **VideoNumPause**: this feature will not be used because our database doesn't contain information about video pause.

21. **HWProblemSave**: is a non-applicable feature because OpenClassrooms doesn't give data about a save button.

22. **TimeHwQuiz**: OpenClassrooms offers only a few quizzes in certain courses, so this would not be a strong indicator. Therefore, this feature will not be used in experiments.

23. **TimeHwVideo**: this is not a testable feature because OpenClassrooms doesn't give us any data on video watching.

24. **TimePlayVideo**: I do not have data to calculate this video features, so it cannot be experimented with.

25. **HwSessions**: is a non-applicable feature because our MOOC only contain quiz assessments, not homework.

26. **IntervalNumQuiz**: is a non-applicable feature because our MOOC only contain quiz assessments, not homework.

27. **IntervalQuizAttempt**: is a non-applicable feature in this context because our MOOC only contain quiz assessments, not homework.

28. **IntervalVideo**: this is not a testable feature because OpenClassrooms doesn't give us any data on video watching.

29. **IntervalDailySession**: we can use a users' sessions on OpenClassrooms and calculate the *average number of these sessions per day.*

30. **IntervalLogin**: is a non-applicable feature because OpenClassrooms' database doesn't track user login or logout.

31. **Avg # requests per session**: this feature will be analyzed using OpenClassrooms' users' *average # of accesses per session.* An access is when a user shows activity on a specific part of the course, such as, studying the material or completing an exercise.

32. **Avg timespan**: is a testable feature and will be represented as the *total inter-session time.* In other words, this feature is the total time elapsed between the end of a user session and the beginning of the proceeding session.

33. **Peer evaluation**: is a feature intended for MOOCs with several discussions and peer evaluations. OpenClassrooms does not utilize these teaching methods, so this feature will not be used to experiment.

2.2.1 Table of Testable Features

Table 2.2: This table summarizes the researched features in coordination to the testability within OpenClassrooms' database. It contains the same feature name column as Table 2.1 and the third column specifies the application to our specific database. Also, the last column lists the necessary data values to test. An "N/A" value represents that the dataset doesn't allow me to experiment with that feature. Please note that the middle column values may be slightly transform to fit OpenClassrooms' dataset and make it testable. Each feature is numbered to easily reference the detailed explanation given above.

	Feature	**Testability within OpenClassrooms**
1	Time since last event	Time since last session after 1-week
2	Normalized grade	Final mark
3	Gabor filters	N/A
4	Time into course	Total course duration
5	Total # events	Total # of sessions
6	Pre-deadline submission time	N/A
7	Visits to progress page	N/A
8	Avg lab grades	N/A
9	# of forum thread views	N/A
10	Avg # problem/quiz attempts	Average # quiz attempts
11	% of lecture videos watched	N/A
12	Avg length forum post	N/A
13	# of forum posts made	N/A
14	Avg homework grades	*Merge with #4*
15	% Video skip	N/A
16	Registration time	N/A
17	NumSession	Average # of sessions before a quiz attempt
18	AvgNumLogin	N/A
19	NumQuiz	N/A
20	VideoNumPause	N/A
21	HWProblemSave	N/A
22	TimeHwQuiz	N/A
23	TimeHwVideo	N/A
24	TimePlayVideo	N/A
25	HwSessions	N/A
26	IntervalNumQuiz	N/A
27	IntervalQuizAttempt	N/A
28	IntervalVideo	N/A
29	IntervalDailySession	Average # of sessions per day
30	IntervalLogin	N/A
31	Avg # requests per session	Avg # of accesses per session
32	Avg Timespan	Total inter-session time
33	Peer evaluation	N/A

Conclusion

After the researchers in [10] received a high amount of feedback from a dynamic survey sent to stopout users, they concluded that students drop out due to exogenous factors, e.g., too busy with work or traditional classes. Plenty of research on MOOC user behavior has been done in the past. However, none have found a single feature or a combination of features that directly predict stopout accuately.

From my research, I've compiled numerous possible features to predict stopout, and listed them in a table for easy access. It seems that course grades and video-related features are highly effective while session-related features showed no impact on overall student performance. It's important to realize that all these indicators may not be generalized to all MOOCs because each course has different teaching styles and interactive activities. For example, not all online classes will have discussion boards, therefore, it will have a completely different database than an online class that does offer them.

In the next chapter, I'll propose which user features are the strongest indicators using the related works discussed in this chapter. I'll also determine which features can be appropriately applied in my experiments with OpenClassrooms' user database.

Chapter 3

Proposal

In this chapter, we'll use the prior research of previous work in the Bibliography Chapter to look at the features in more depth. The goal of this chapter is to visualize the discriminate features that clearly separate the 4 user groups: successful premium, failed premium, successful non-premium, and failed non-premium students.

First, I'll display several histograms and boxplots for each feature which is testable within OpenClassrooms (see Table 2.2). Since there are numerous graphs, I'll then analyze each feature and mention the graphical points worth noticing.

3.1 Feature Statistics

This section displays various graphs to visually prove the following features are discriminants. The data is supplied by OpenClassrooms and contains information from over 3 million users on 8 online courses: Arduino, Bootstrap, Gestion Project, Ionic, JavaScript, Rubys, Twitter, and Web. See Table 3.1 for figure references.

Note: For each set of graphs, the data in Rubys courses for the non-premium successful users. Also, pay close attention to the captions, x-, and y-axis values to avoid graph misinterpretation.

Feature	**Figures**
Time Since Last Session (include deviation)	3.1, 3.2, 3.3, 3.4, 3.5, 3.6
Time Since Last Session (exclude deviation)	3.7, 3.8, 3.9, 3.10, 3.11, 3.12
Final Mark	3.13
Total Course Duration	3.14, 3.15, 3.16, 3.17, 3.18, 3.19
Total # of Sessions	3.20
Average # of Quiz Attempts	3.21, 3.22, 3.23, 3.24, 3.25, 3.26
Average # of Sessions Before a Quiz Attempt	3.27
Average # of Daily Sessions	3.28
Average # of Accesses Per Session	3.29
Intertime Session	3.30

Table 3.1: This table acts as a guide to help navigate through the following histograms and box-plots. Note that the feature *time since last event* is displayed in two ways: the first includes all users and data, and the other way displays the graph excluding the deviation mark. The deviation contains the users that do not return to the course one week after their first session. Without deviation, the histogram shows more detail and is worth displaying for close analysis.

3.1.1 Time Since Last Session (including deviation)

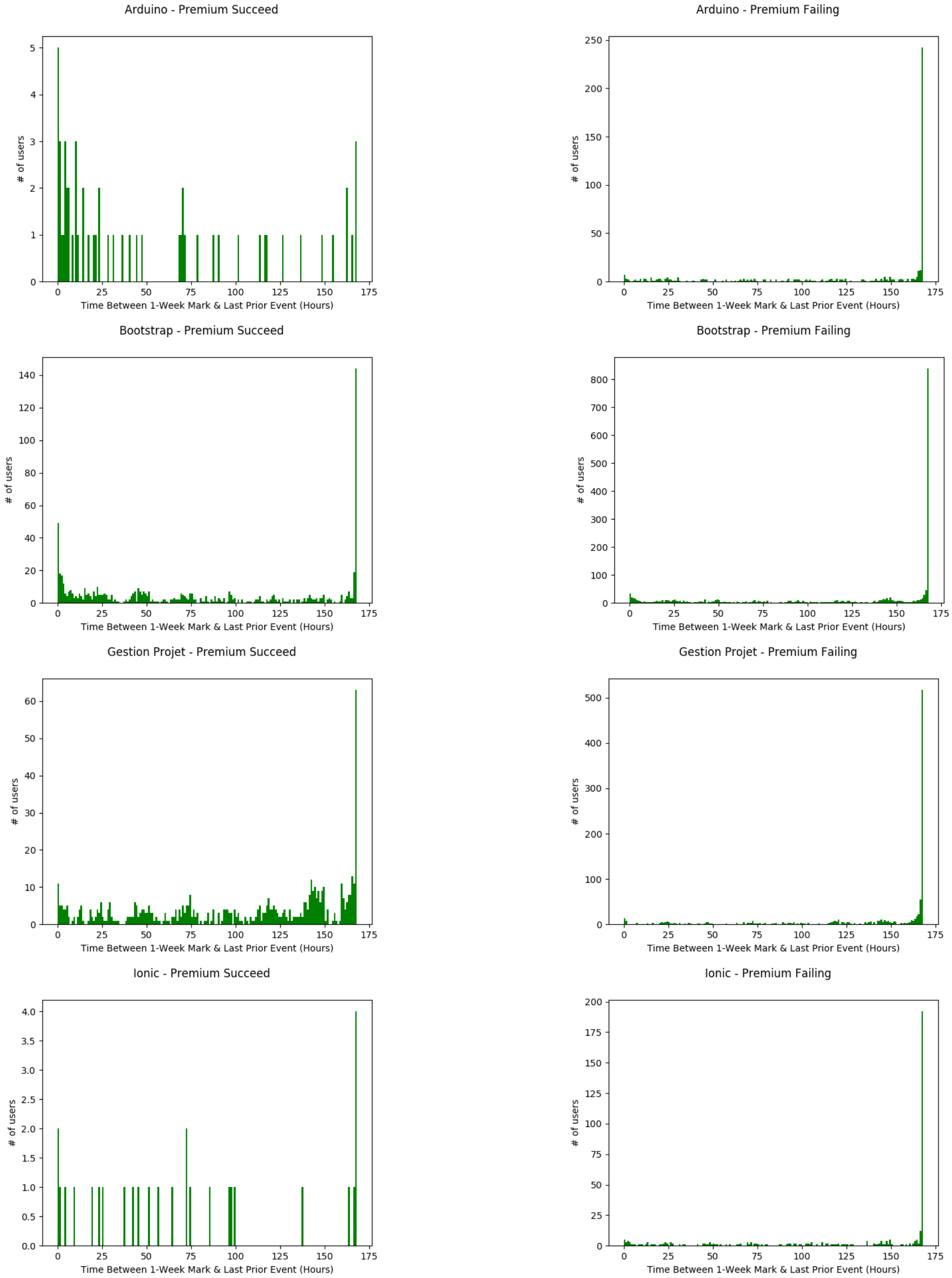


Figure 3.1: Time since last session (including deviation) for premium users: the time elapsed (hours) since the 1-week mark (1 week after the user's first session) and the most recent session proceeding the 1-week mark. The x-axis shows the time elapsed in hours, and the y-axis represents the number of students that fall under the x-axis circumstance. The graph is plotted for the courses: Arduino, Bootstrap, Gestion Project, and Ionic. The left column shows successful users and the right column displays failed users.

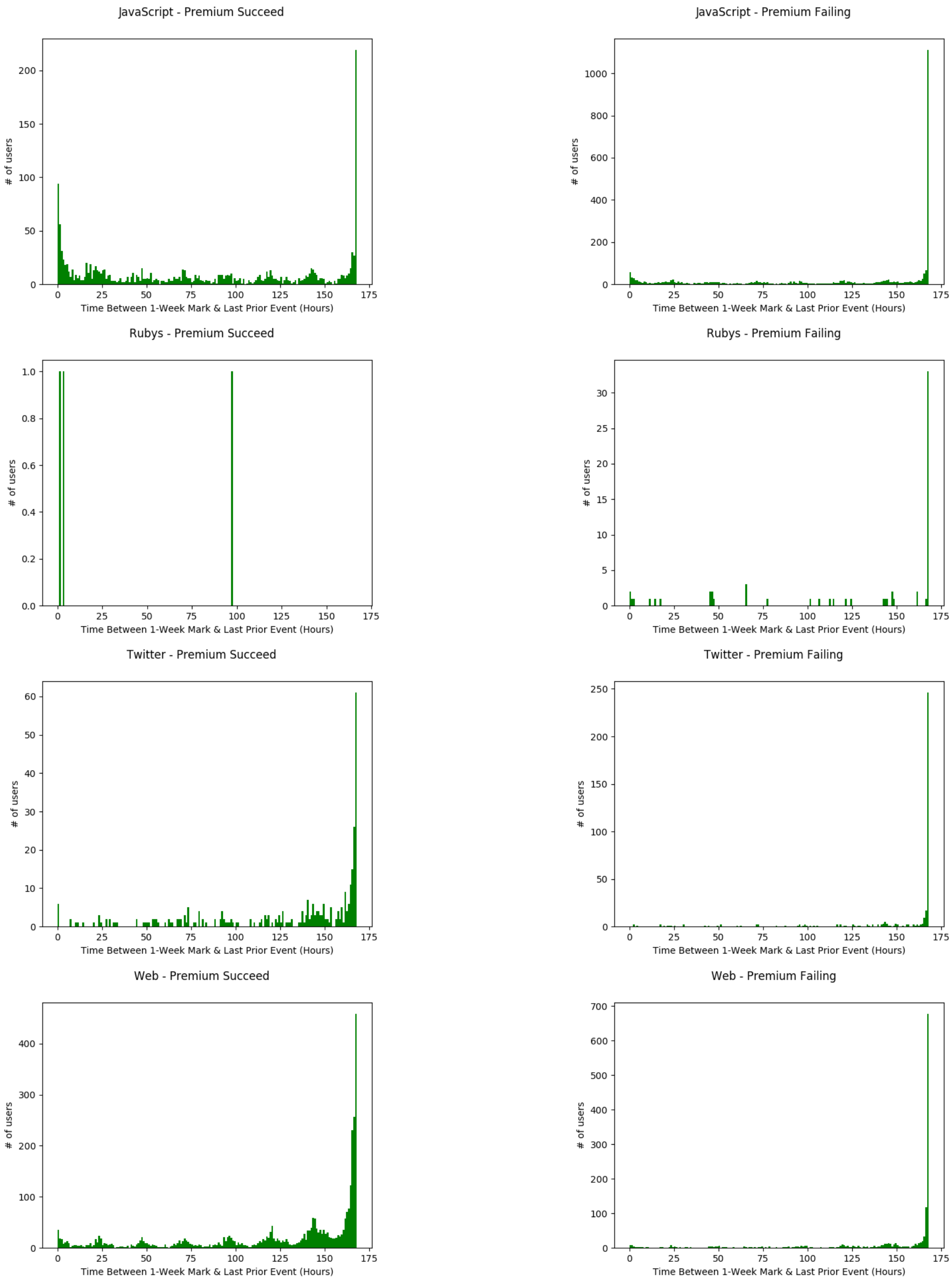


Figure 3.2: Time since last session (including deviation) for premium users: the time elapsed (hours) since the 1-week mark (1 weeks after the user's first session) and the most recent session proceeding the 1-week mark. The x-axis shows the time elapsed in hours, and the y-axis represents the number of students that fall under the x-axis circumstance. The graph is plotted for the courses: Javascript, Rubys, Twitter, and Web. The left column shows successful users and the right column displays failed users.

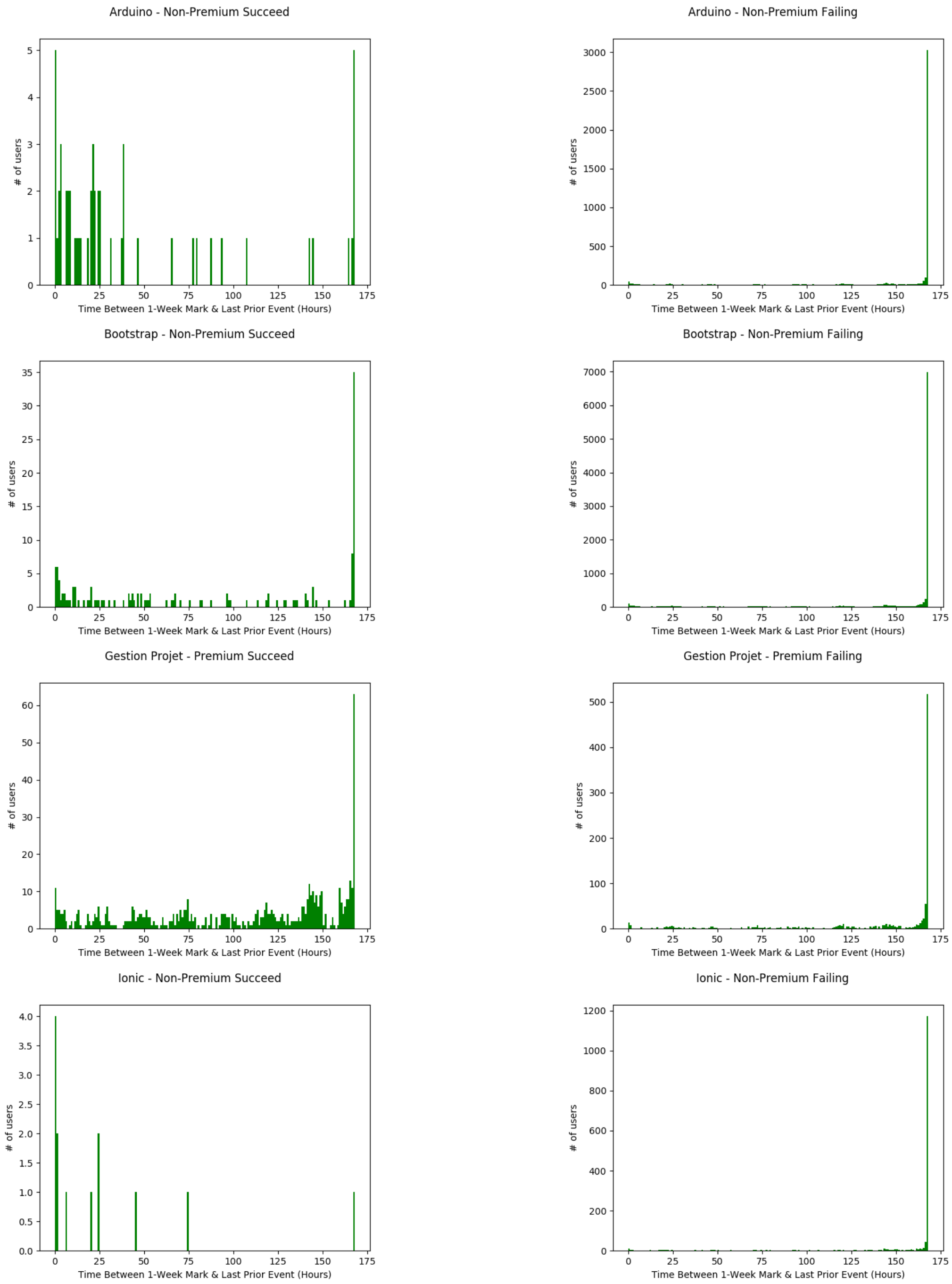


Figure 3.3: Time since last session (including deviation) for non-premium users: the time elapsed (hours) since the 1-week mark (7 days after the user's first session) and the most recent session proceeding the 1-week mark. The x-axis shows the time elapsed in hours, and the y-axis represents the number of students that fall under the x-axis circumstance. The graph is plotted for the courses: Arduino, Bootstrap, Gestion Project, and Ionic. The left column shows successful users and the right column displays failed users.

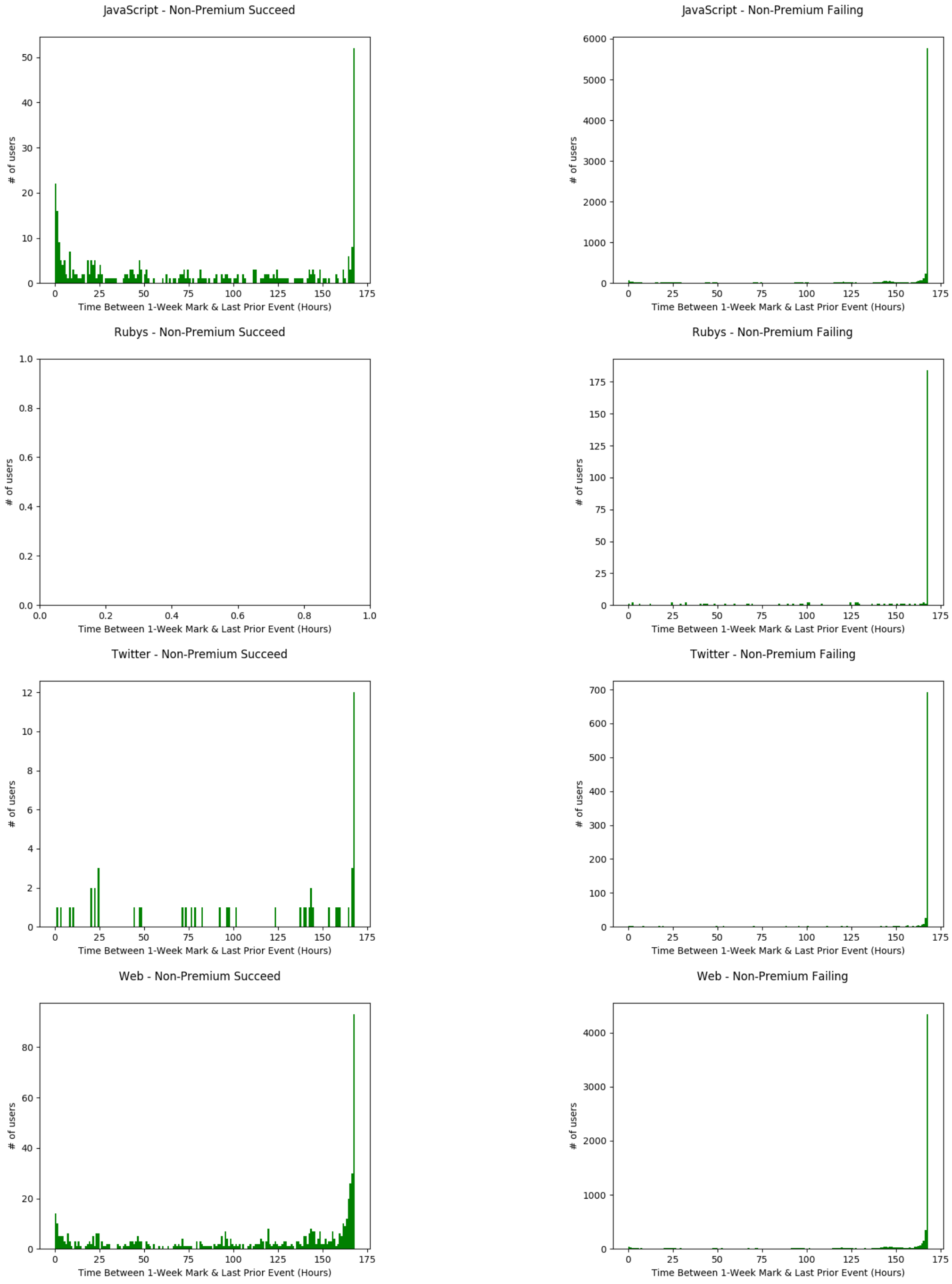


Figure 3.4: Time since last session (including deviation) for non-premium users: the time elapsed (hours) since the 1-week mark (7-days after the user's first session) and the most recent session proceeding the mark. The x-axis shows the time elapsed in hours, and the y-axis represents the number of students that fall under the x-axis circumstance. The graph is plotted for the courses: JavaScript, Rubys, Twitter, and Web. The left column shows successful users and the right column displays failed users.

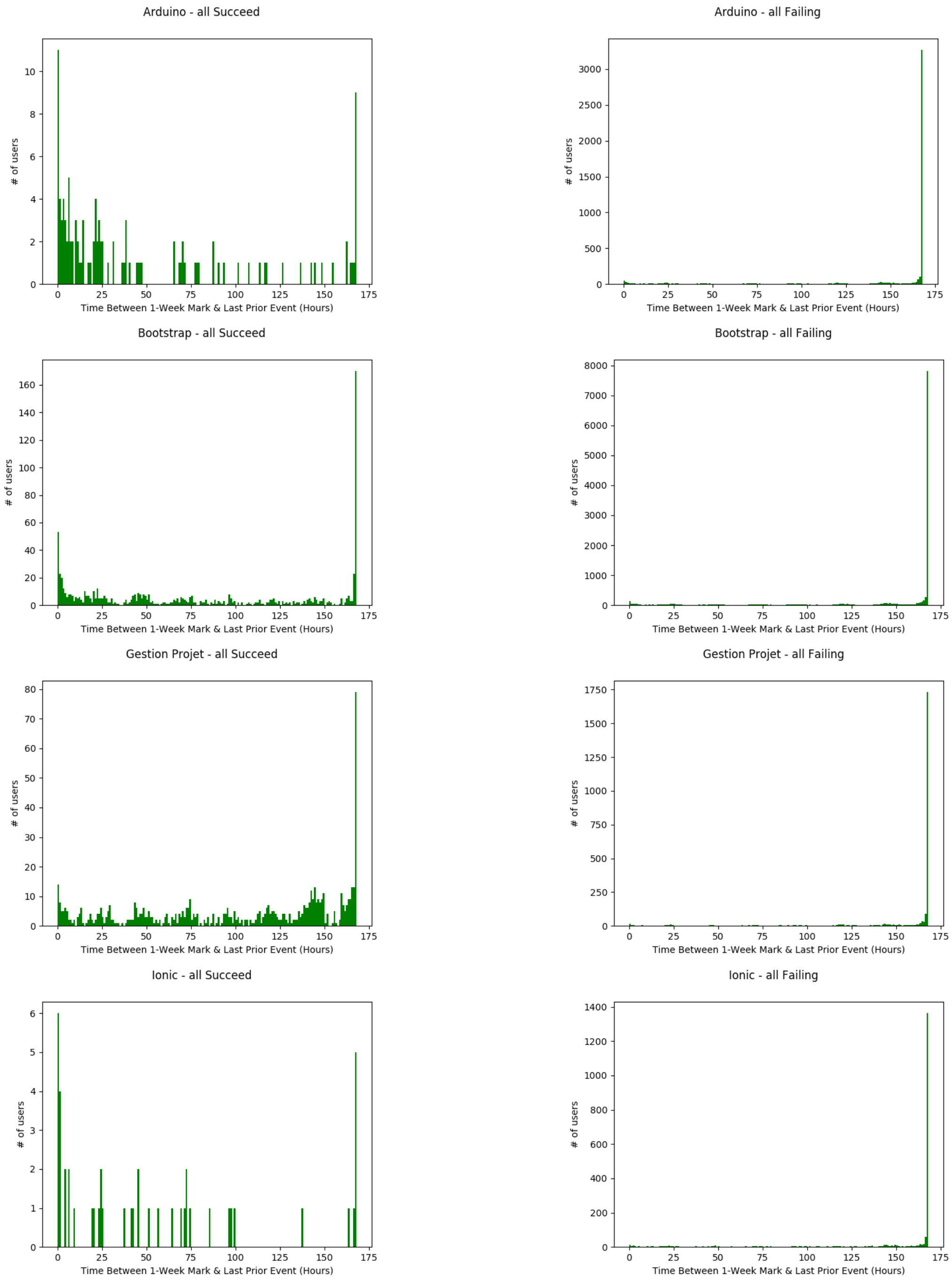


Figure 3.5: Time since last session (including deviation) for all (premium and non-premium) users: the time elapsed (hours) since the 1-week mark (1 week after the user's first session) and the most recent session proceeding the 1-week mark. The x-axis shows the time elapsed in hours, and the y-axis represents the number of students that fall under the x-axis circumstance. The graph is plotted for the courses: Arduino, Bootstrap, Gestion Project, and Ionic. The left column shows successful users and the right column displays failed users.

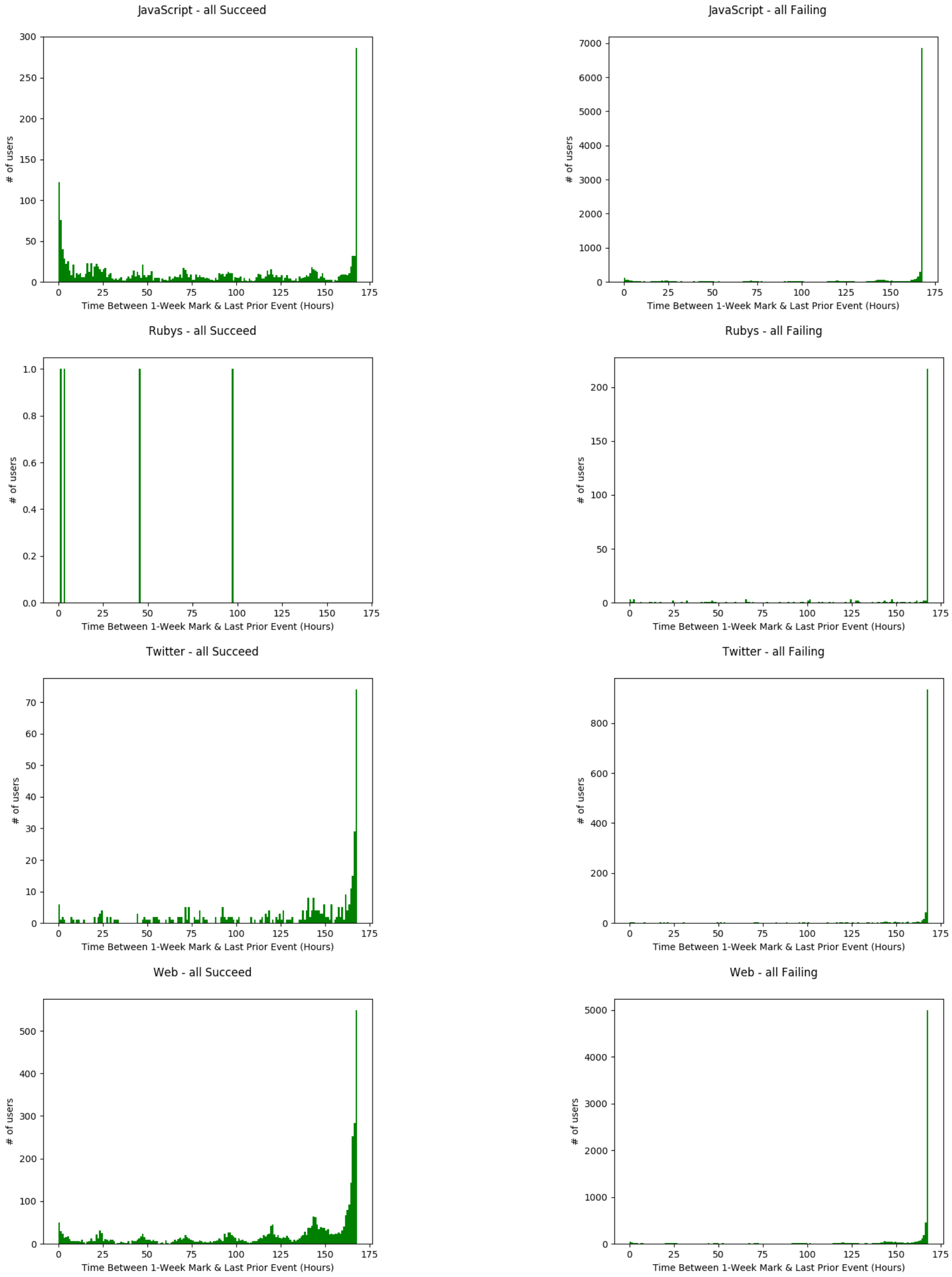


Figure 3.6: Time since last session (including deviation) for all (premium and non-premium) users: the time elapsed (hours) since the 1-week mark (1 weeks after the user's first session) and the most recent session proceeding the 1-week mark. The x-axis shows the time elapsed in hours, and the y-axis represents the number of students that fall under the x-axis circumstance. The graph is plotted for the courses: JavaScript, Ruby, Twitter, and Web. The left column shows users and the right column displays failed users.

3.1.2 Time Since Last Session (excluding deviation)

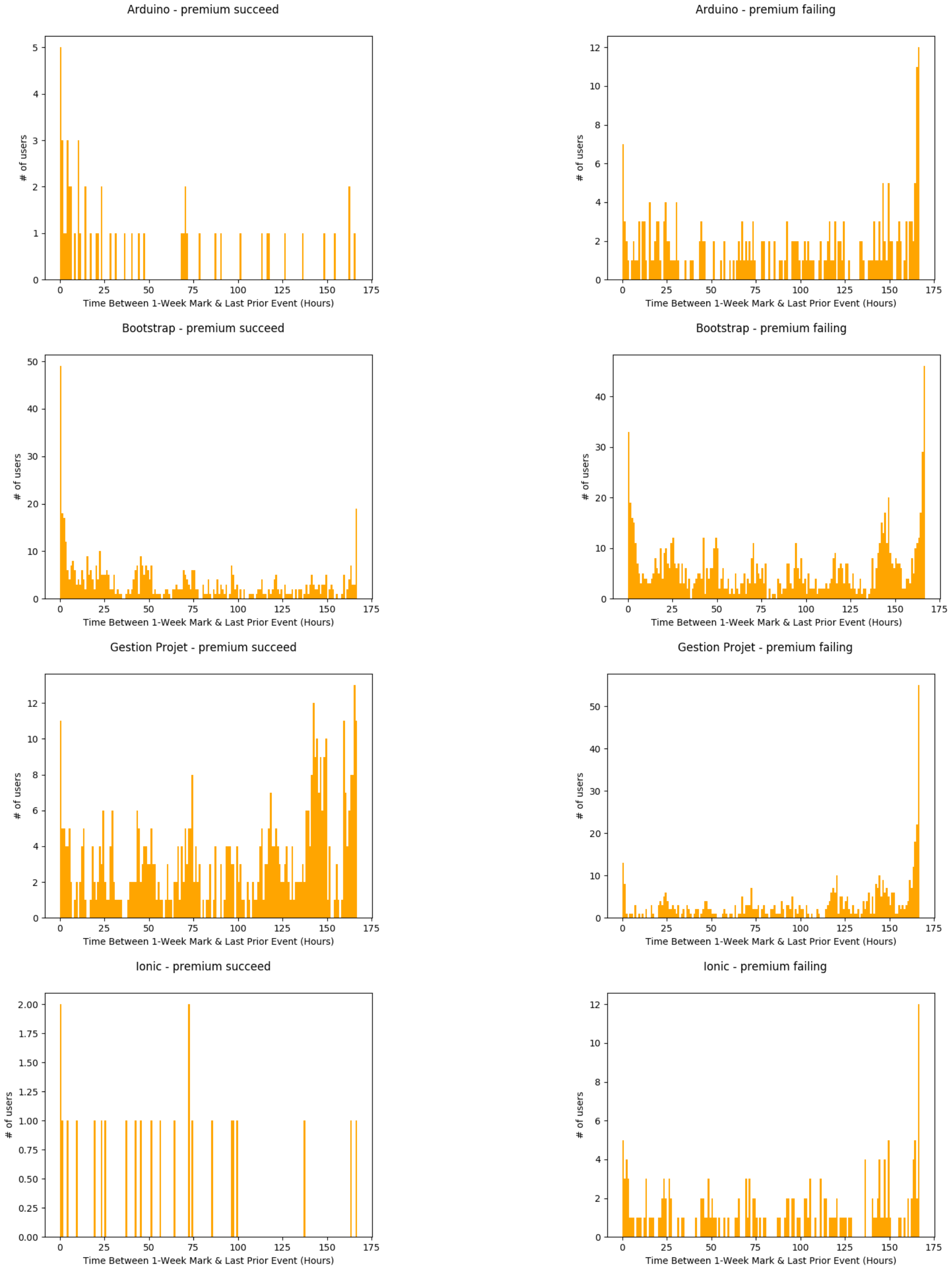


Figure 3.7: Time since last session (excluding deviation) for premium users: the time elapsed (hours) since the 1-week mark (7 days after the user's first session) and the most recent session proceeding the 1-week mark **excluding the deviation of users with 168 hours (7 days), or in other words, students that do not return after 7 days**. The graph is plotted for the courses: Arduino, Bootstrap, Gestion Project, and Ionic. The left column shows successful users and the right column displays failed users.

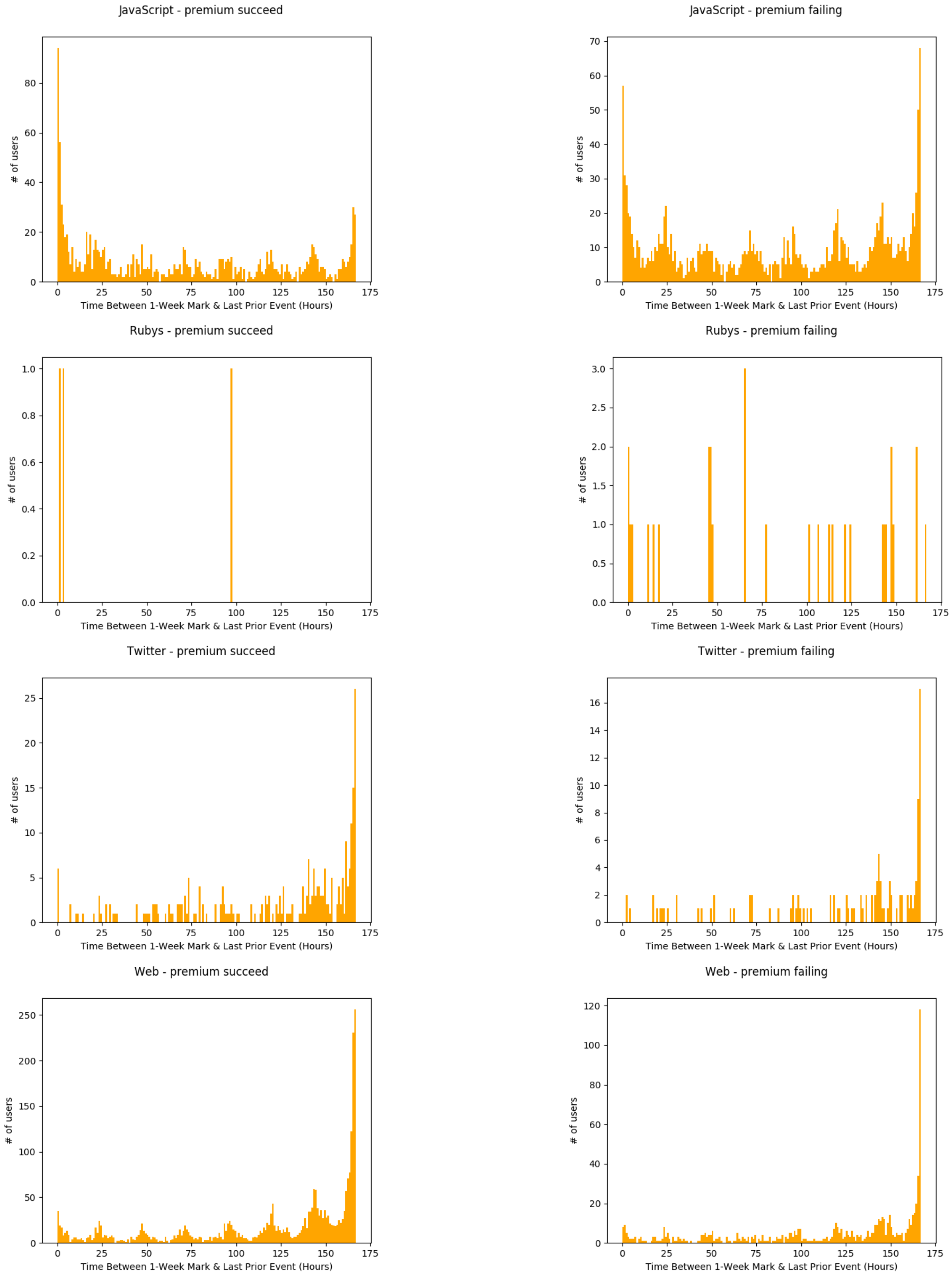


Figure 3.8: Time since last session (excluding deviation) for premium users: the time elapsed (hours) since the 1-week mark (7 days after the user's first session) and the most recent session proceeding the 1-week mark **excluding the deviation of users with 168 hours (7 days), or in other words, students that do not return after 7 days**. The graph is plotted for the courses: JavaScript, Rubys, Twitter, and Web. The left column shows successful users and the right column displays premium users.

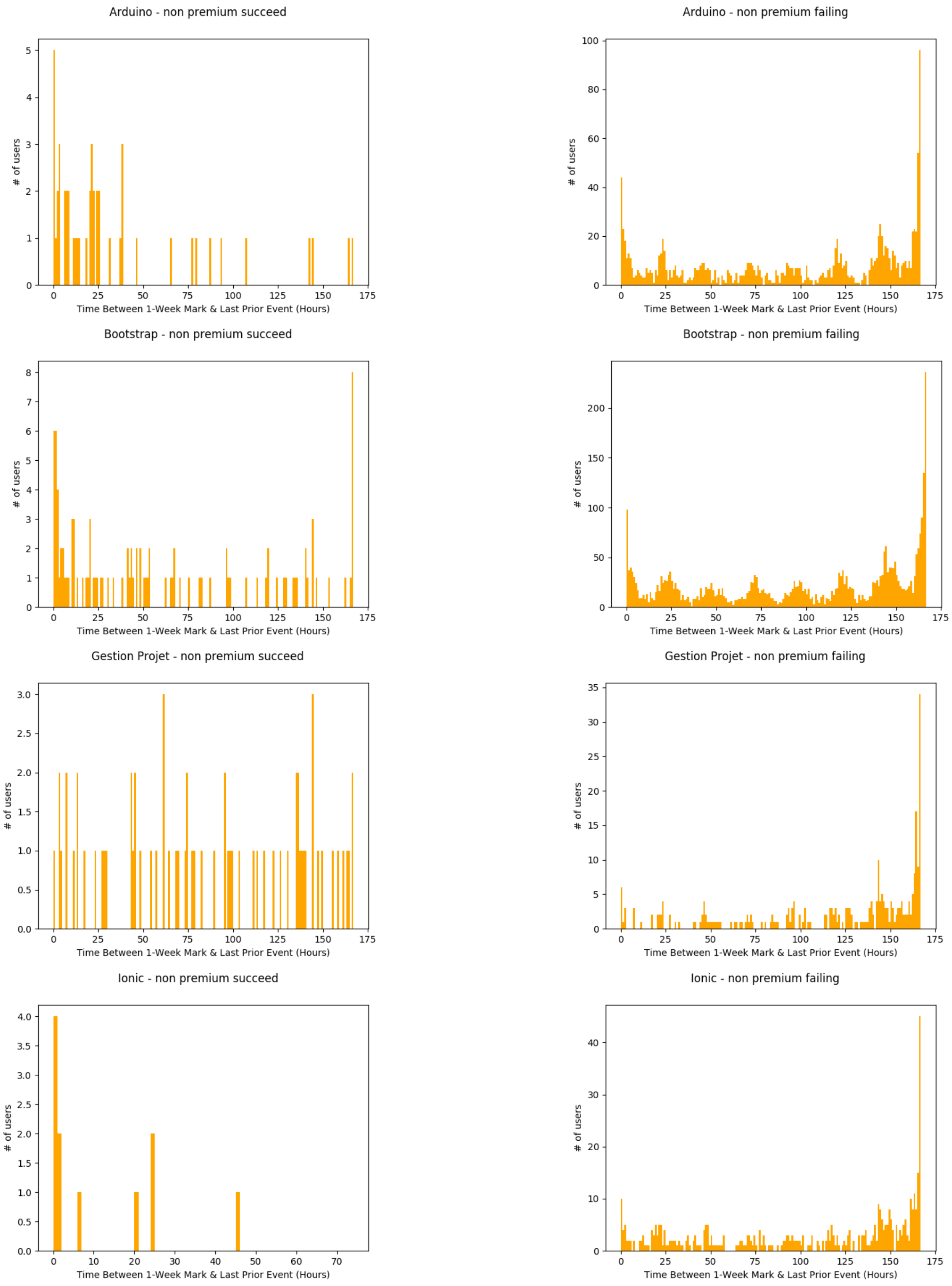


Figure 3.9: Time since last session (excluding deviation) for non-premium users: the time elapsed (hours) since the 1-week mark (7 days after the user's first session) and the most recent session proceeding the 1-week mark **excluding the deviation of users with 168 hours (7 days), or in other words, students that do not return after 7 days**. The graph is plotted for the courses: Arduino, Bootstrap, Gestion Project, and Ionic. The left column shows successful users and the right column displays failed users.

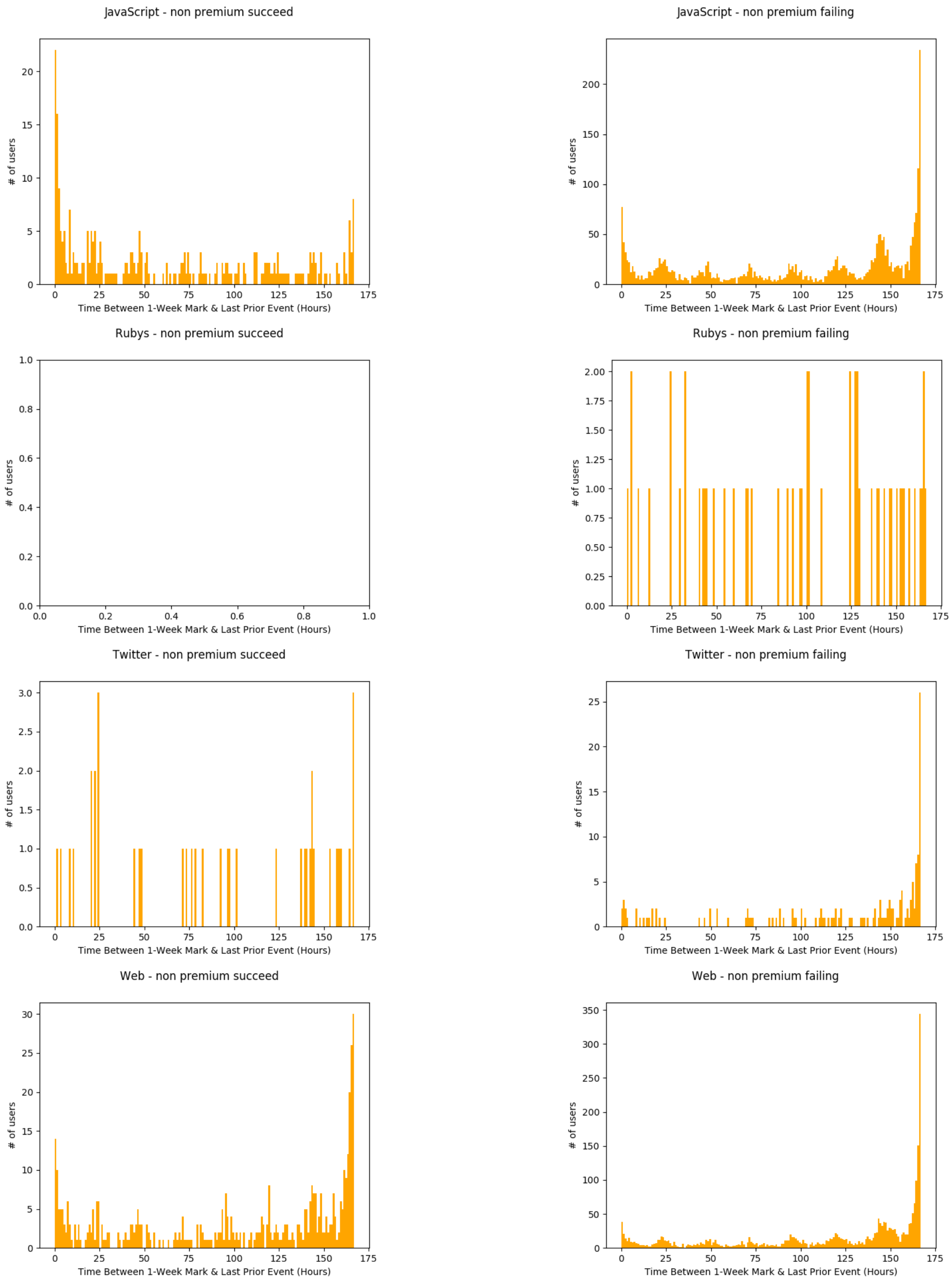


Figure 3.10: Time since last session (excluding deviation) for non-premium users: the time elapsed (hours) since the 1-week mark (7 days after the user's first session) and the most recent session proceeding the 1-week mark **excluding the deviation of users with 168 hours (7 days), or in other words, students that do not return after 7 days**. The graph is plotted for the courses: JavaScript, Rubys, Twitter, and Web. The left column shows successful users and the right column displays failed users.

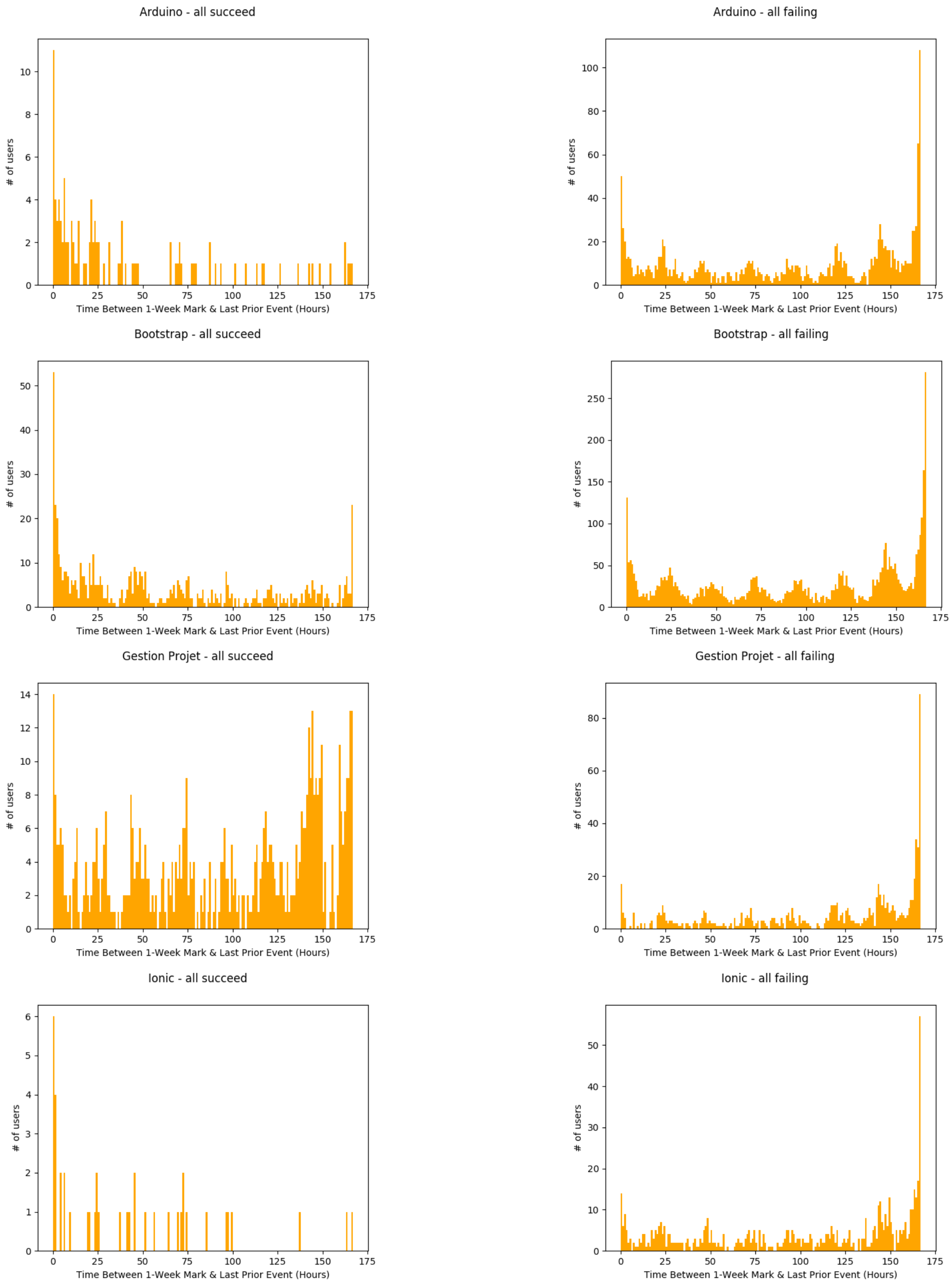


Figure 3.11: Time since last session (excluding deviation) for all (premium and non-premium) users: the time elapsed (hours) since the 1-week mark (7 days after the user's first session) and the most recent session proceeding the 1-week mark **excluding the deviation of users with 168 hours (7 days), or in other words, students that do not return after 7 days**. The graph is plotted for the courses: Arduino, Bootstrap, Gestion Project, and Ionic. The left column shows successful users and the right column displays failed users.

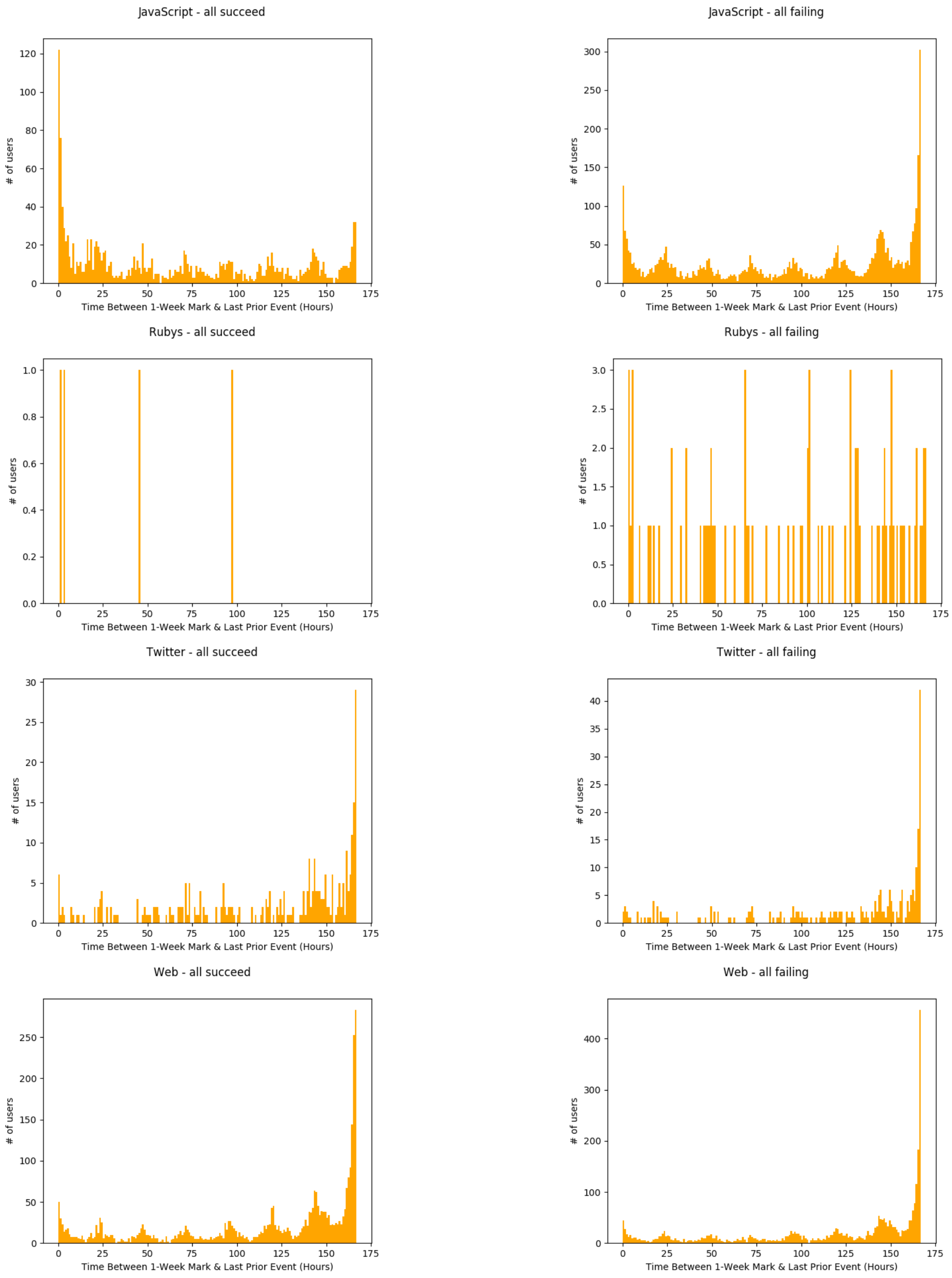


Figure 3.12: Time since last session (excluding deviation) for all (premium and non-premium) users: the time elapsed (hours) since the 1-week mark (7 days after the user's first session) and the most recent session proceeding the 1-week mark **excluding the deviation of users with 168 hours (7 days), or in other words, students that do not return after 7 days**. The graph is plotted for the courses: JavaScript, Rubys, Twitter, and Web. The left column shows successful users and the right column displays failed users.

3.1.3 Final Mark

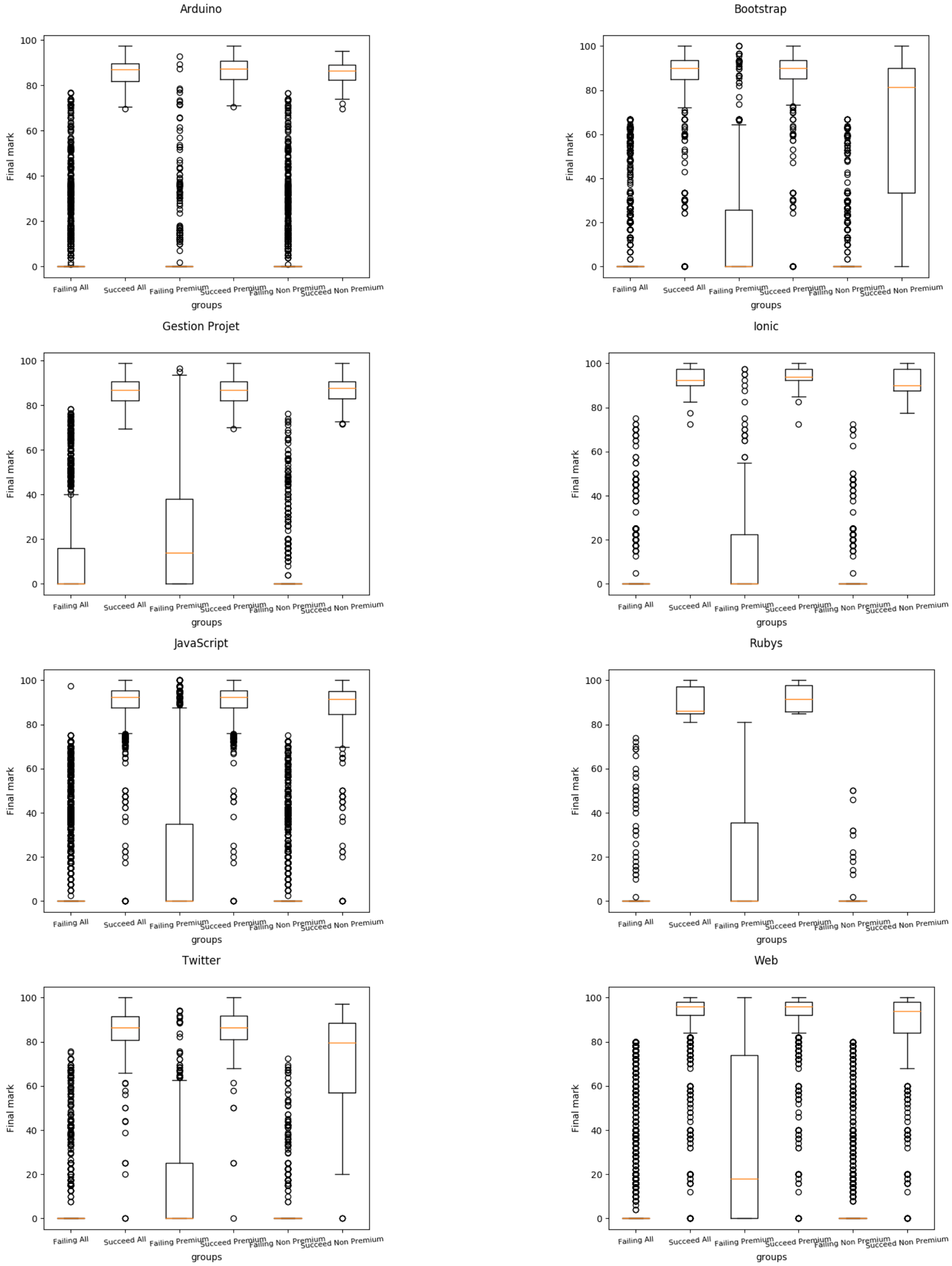


Figure 3.13: Final Mark: users' final grades for all 8 courses. For each graph, the y-axis is the final grade percentage (between 0-100), and the x-axis are the six different user groups.

3.1.4 Total Course Duration

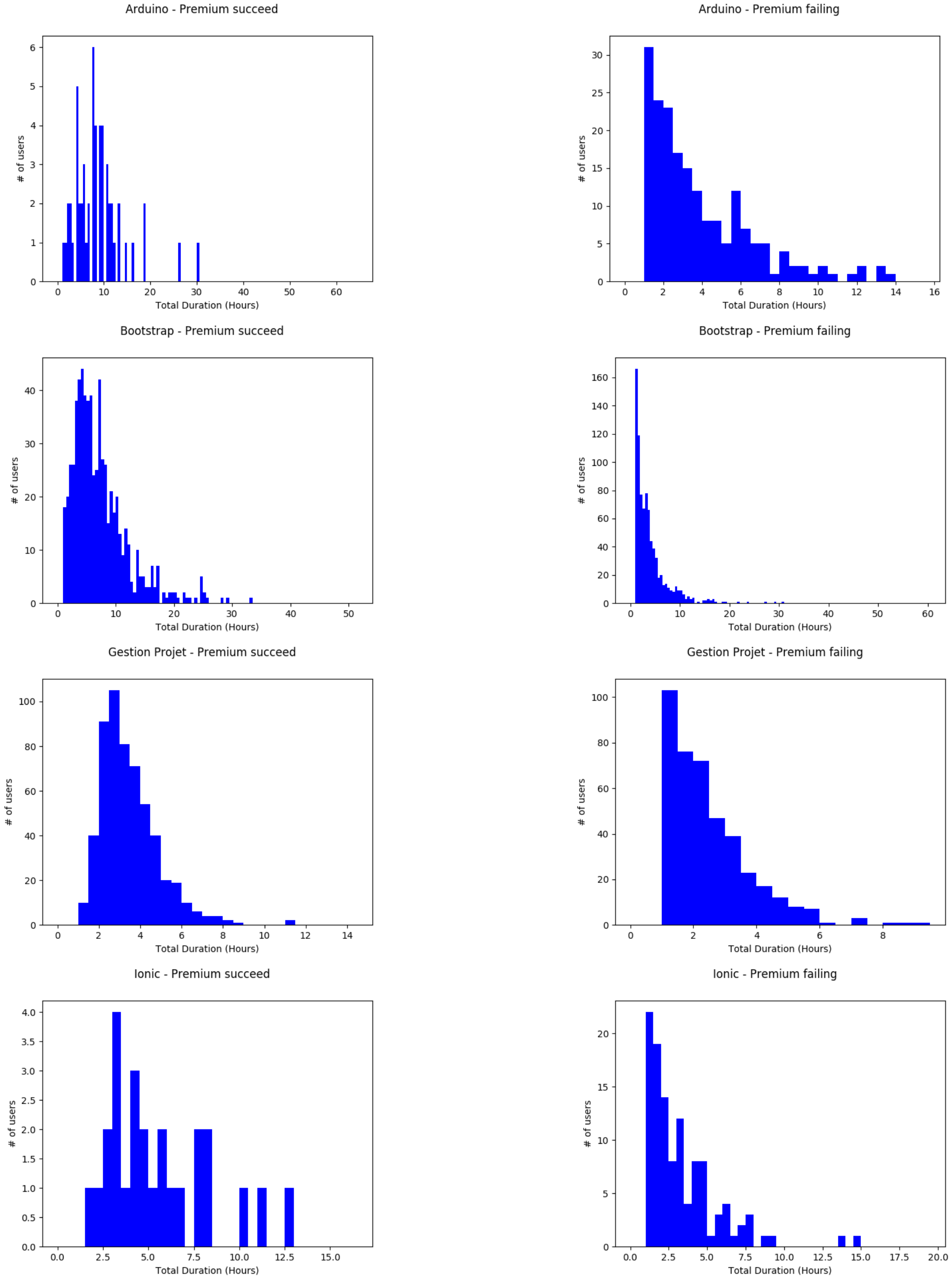
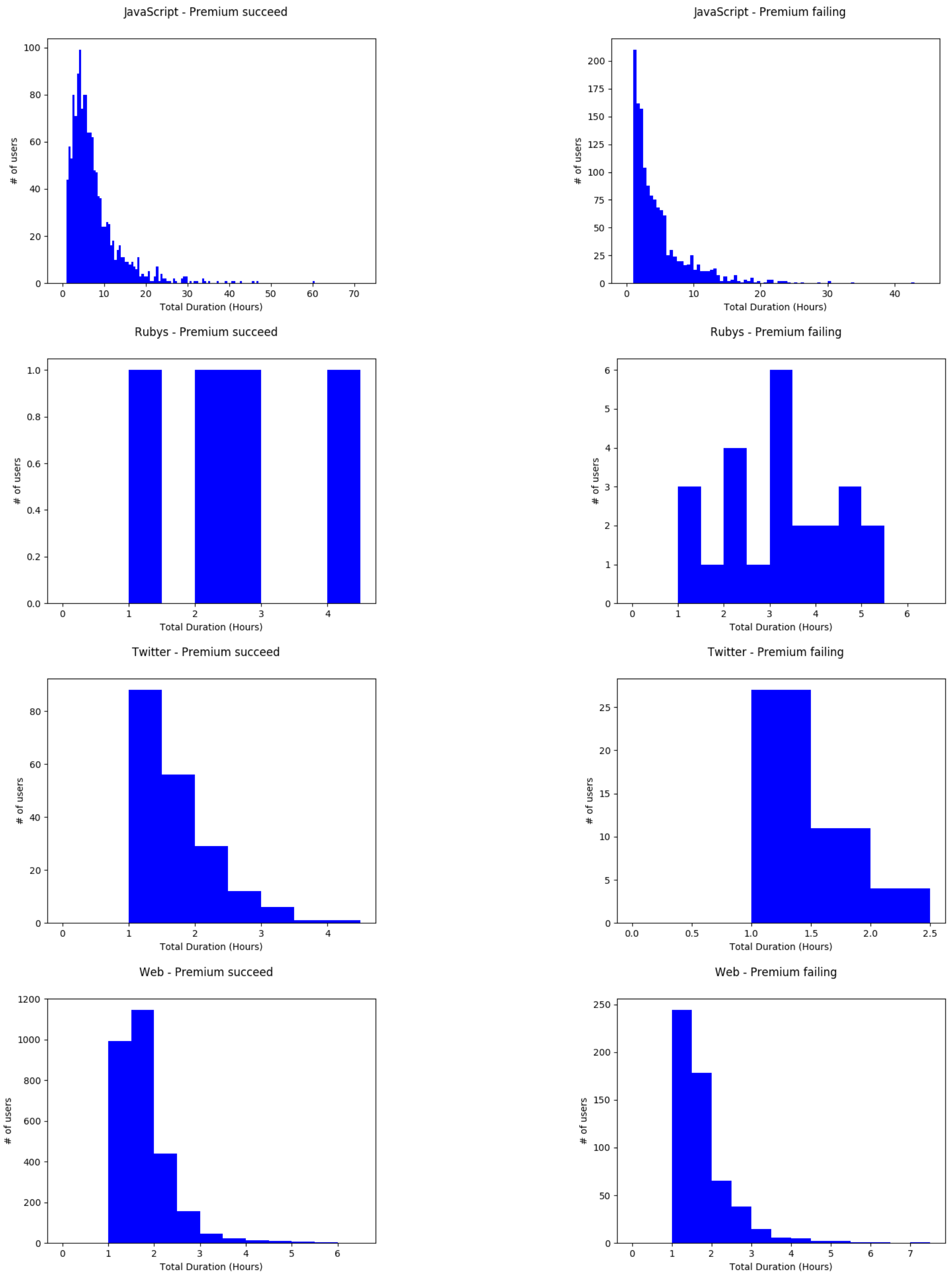


Figure 3.14: Total duration for premium users: the total amount of time users spend on the courses Arduino, Bootstrap, Gestion Project, and Ionic. The x-axis shows the duration in hours, and the y-axis represents the number of students that spent x amount of hours on the course. The left column displays the successful users and right column shows the failed users. The x-axis minimum begins at 1 to eliminate outliers less than an hour.


JavaScript - Premium succeed

JavaScript - Premium failing

Rubys - Premium succeed

Rubys - Premium failing

Twitter - Premium succeed

Twitter - Premium failing

Web - Premium succeed

Web - Premium failing

Figure 3.15: Total duration for premium users: the total amount of time users spend on the courses JavaScript, Rubys, Twitter, and Web. The x-axis shows the duration in hours, and the y-axis represents the number of students that spent x amount of hours on the course. The left column displays the successful users and right column shows the failed users.The x-axis minimum begins at 1 to eliminate outliers less than an hour.

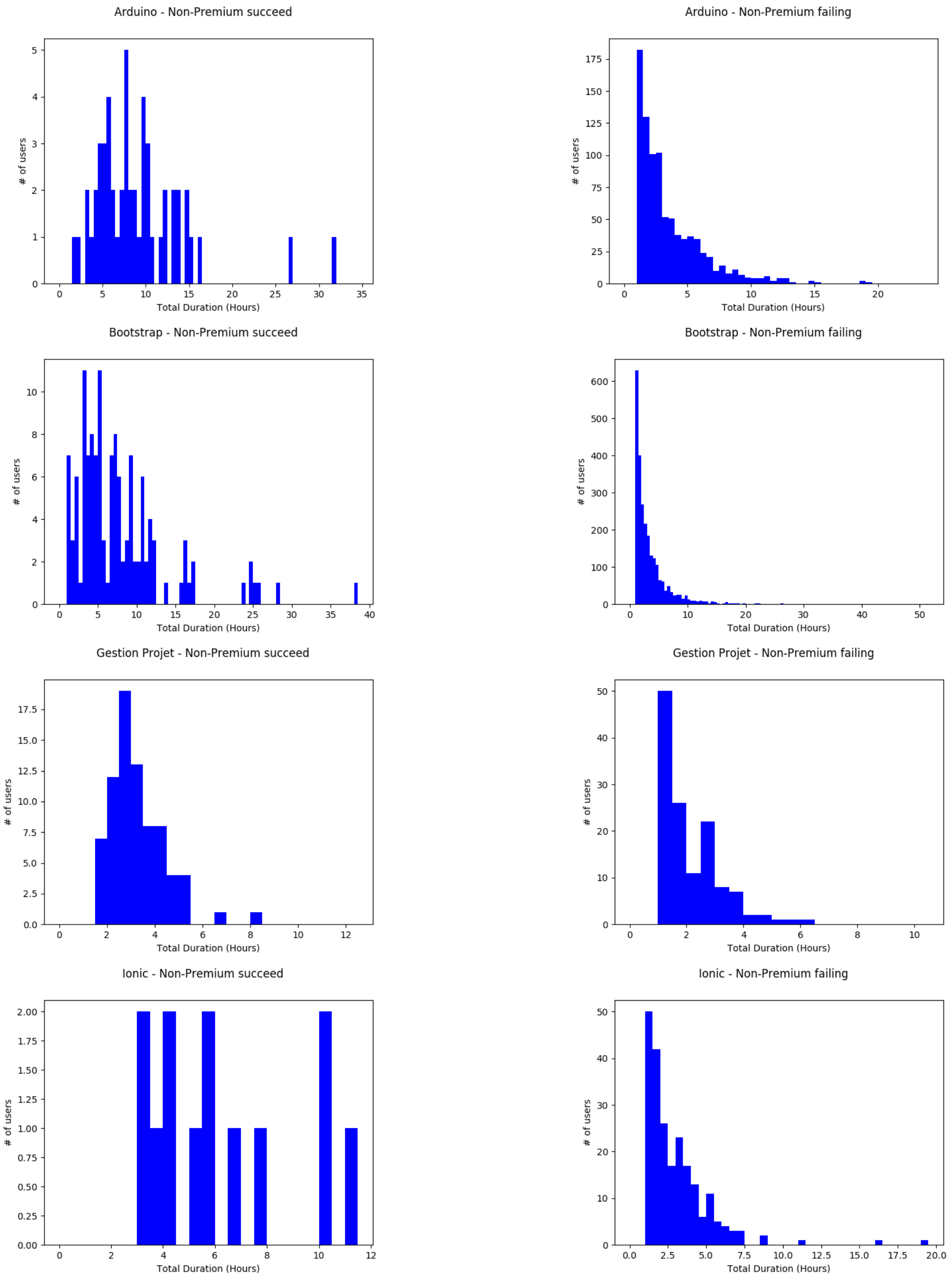


Figure 3.16: Total duration for non-premium users: the total amount of time users spend on the courses Arduino, Bootstrap, Gestion Project, and Ionic. The x-axis shows the duration in hours, and the y-axis represents the number of students that spent x amount of hours on the course. The left column displays the successful users and right column shows the failed users. The x-axis minimum begins at 1 to eliminate outliers less than an hour.

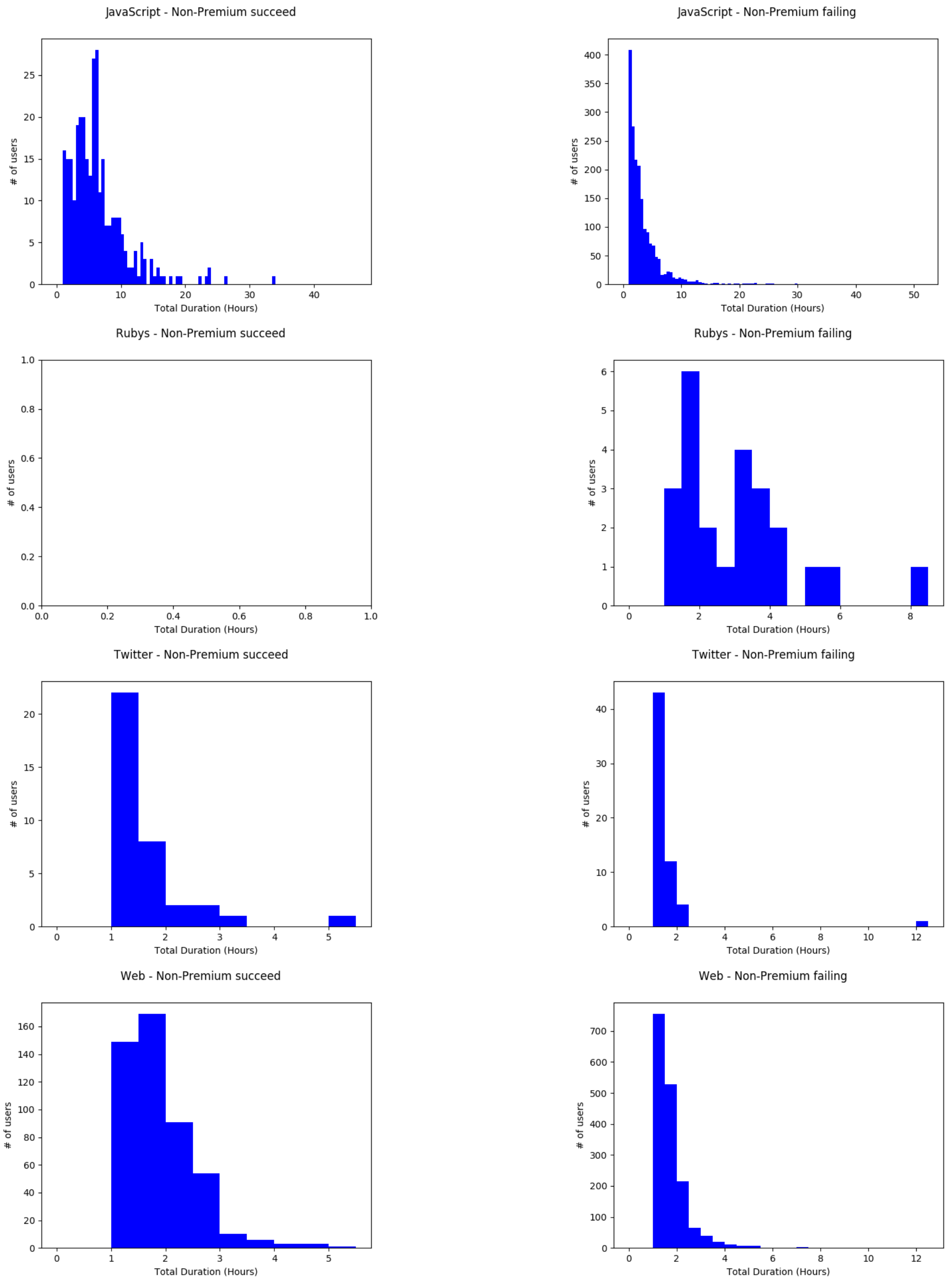


Figure 3.17: Total duration for non-premium users: the total amount of time users spend on the courses JavaScript, Rubys, Twitter, and Web. The x-axis shows the duration in hours, and the y-axis represents the number of students that spent x amount of hours on the course. The left column displays the successful users and right column shows the failed users. The x-axis minimum begins at 1 to eliminate outliers less than an hour.

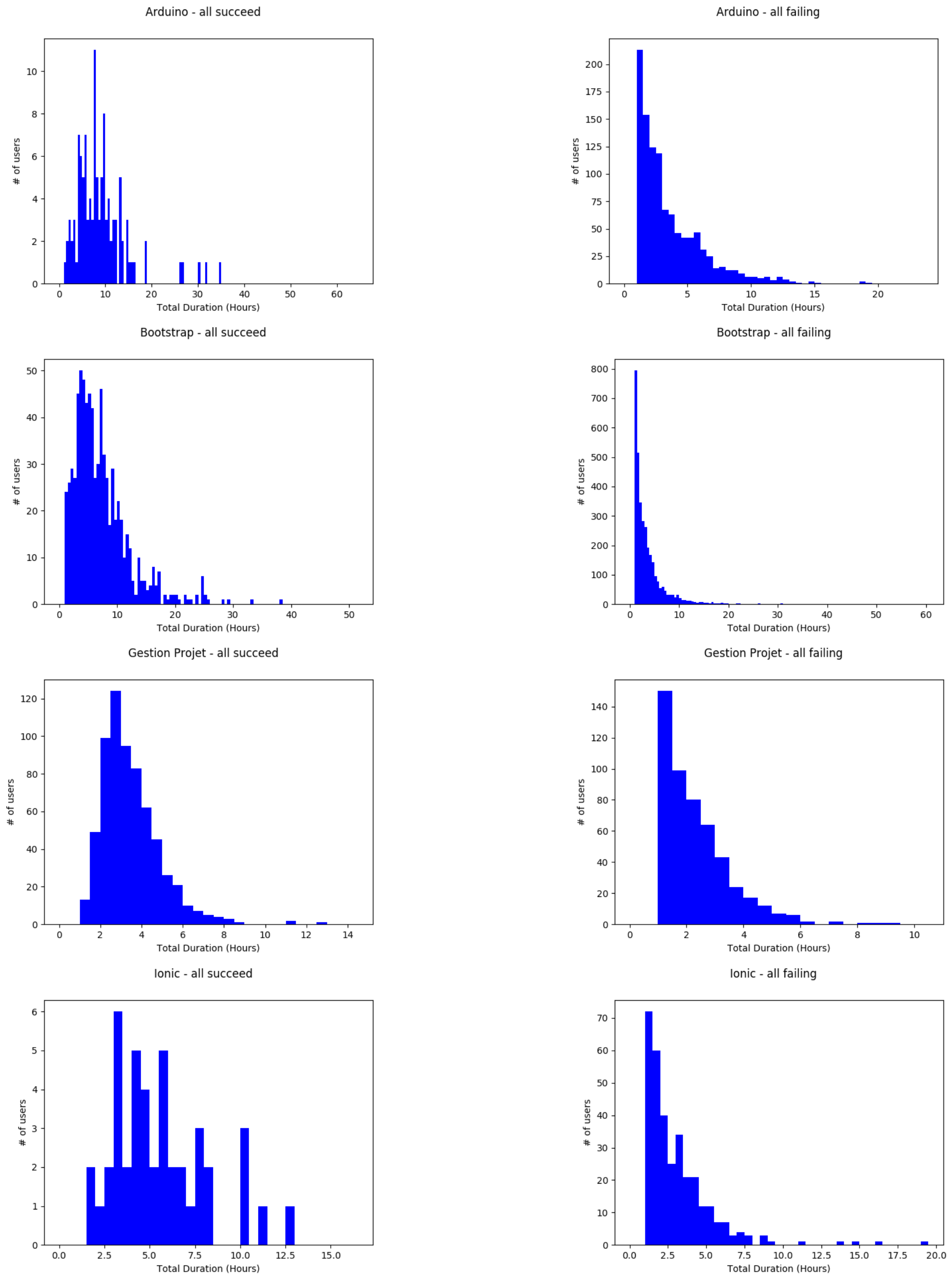


Figure 3.18: Total duration for all (premium and non-premium) users: the total amount of time users spend on the courses Arduino, Bootstrap, Gestion Project, and Ionic. The x-axis shows the duration in hours, and the y-axis represents the number of students that spent x amount of hours on the course. The left column displays successful users and right column shows the failed. The x-axis minimum begins at 1 to eliminate outliers less than an hour.

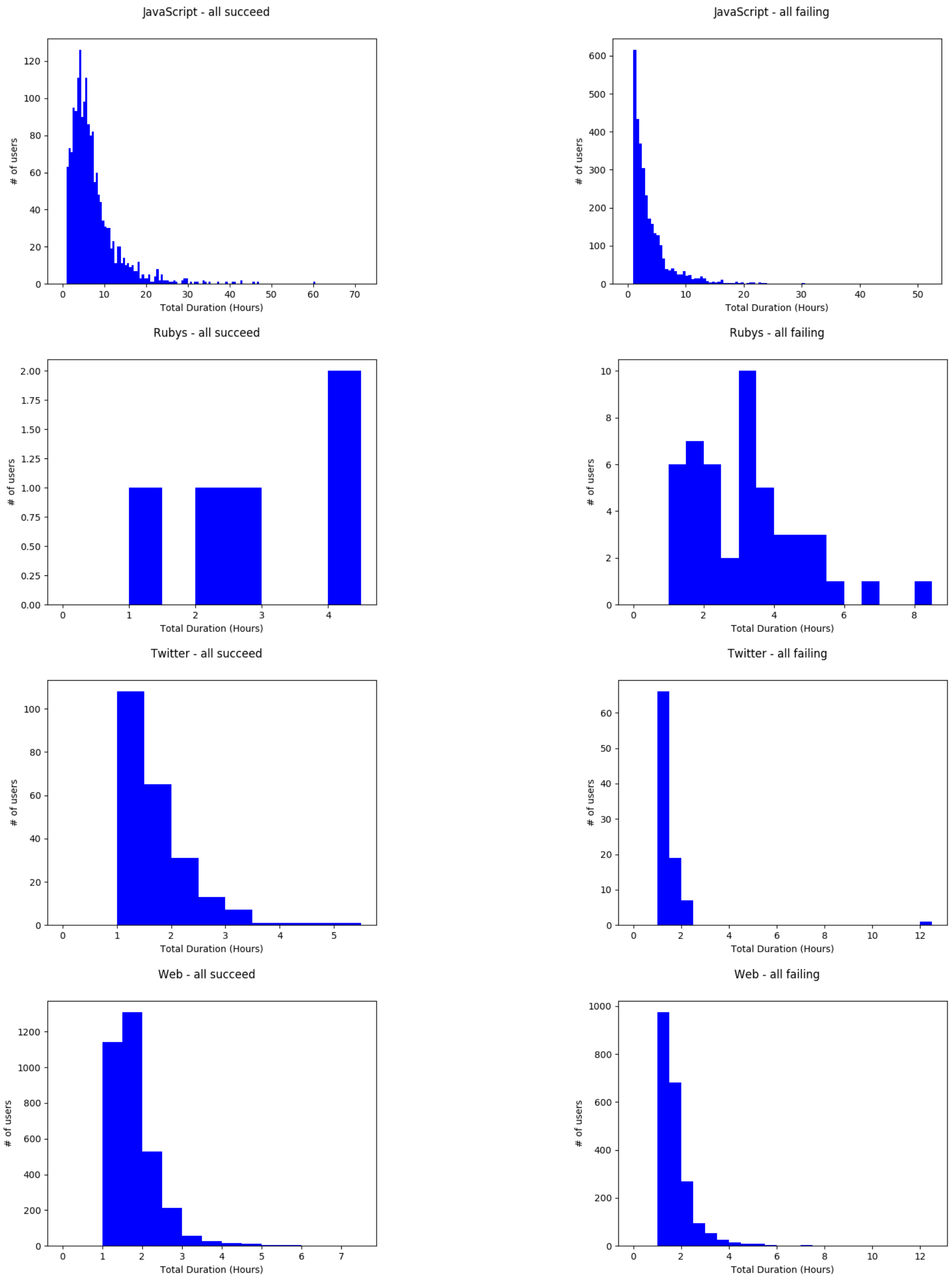


Figure 3.19: Total duration all (premium and non-premium) users: the total amount of time users spend on the courses JavaScript, Rubys, Twitter, and Web. The x-axis shows the duration in hours, and the y-axis represents the number of students that spent x amount of hours on the course. The left column displays successful users and right column shows the failed. The x-axis minimum begins at 1 to eliminate outliers less than an hour.

3.1.5 Total Number of Sessions

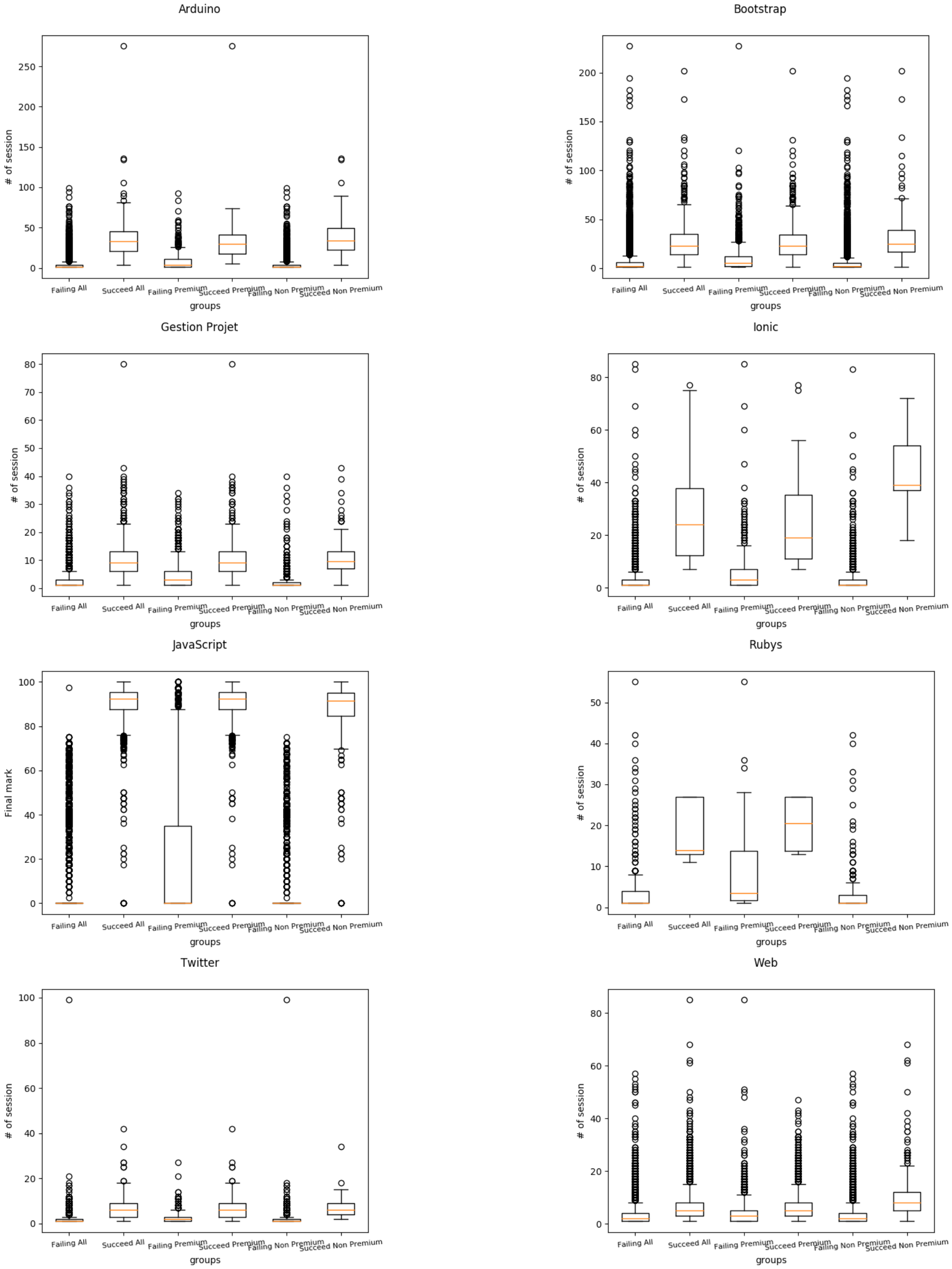


Figure 3.20: Total # of Sessions: shows the total of number of sessions in 8 courses: Arduino, Bootstrap, Gestion Project, Ionic, JavaScript, Rubys, Twitter, and Web. There is a boxplot for each group within each course (labeled on x-axis).

3.1.6 Number of Quiz Attempts

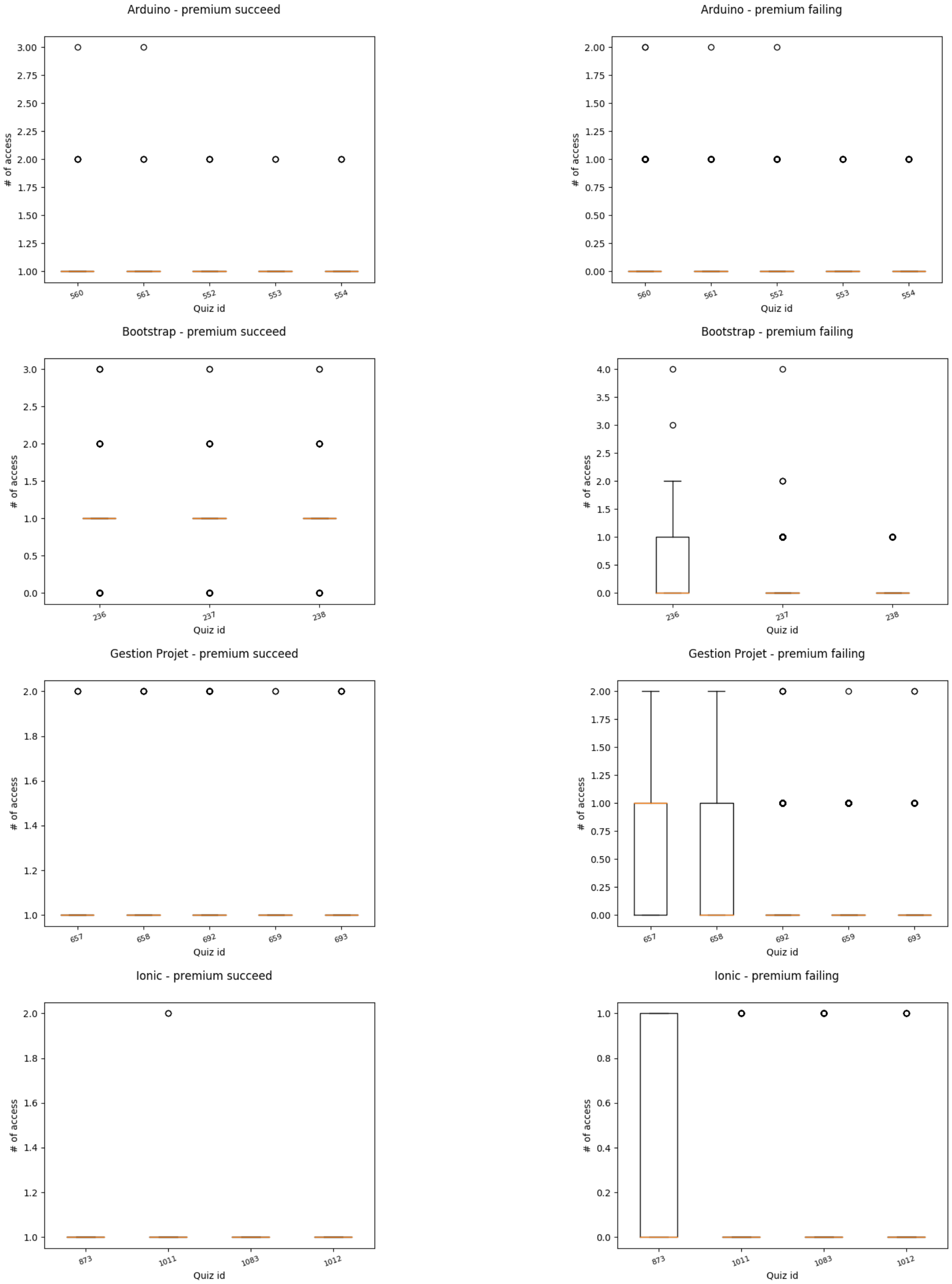
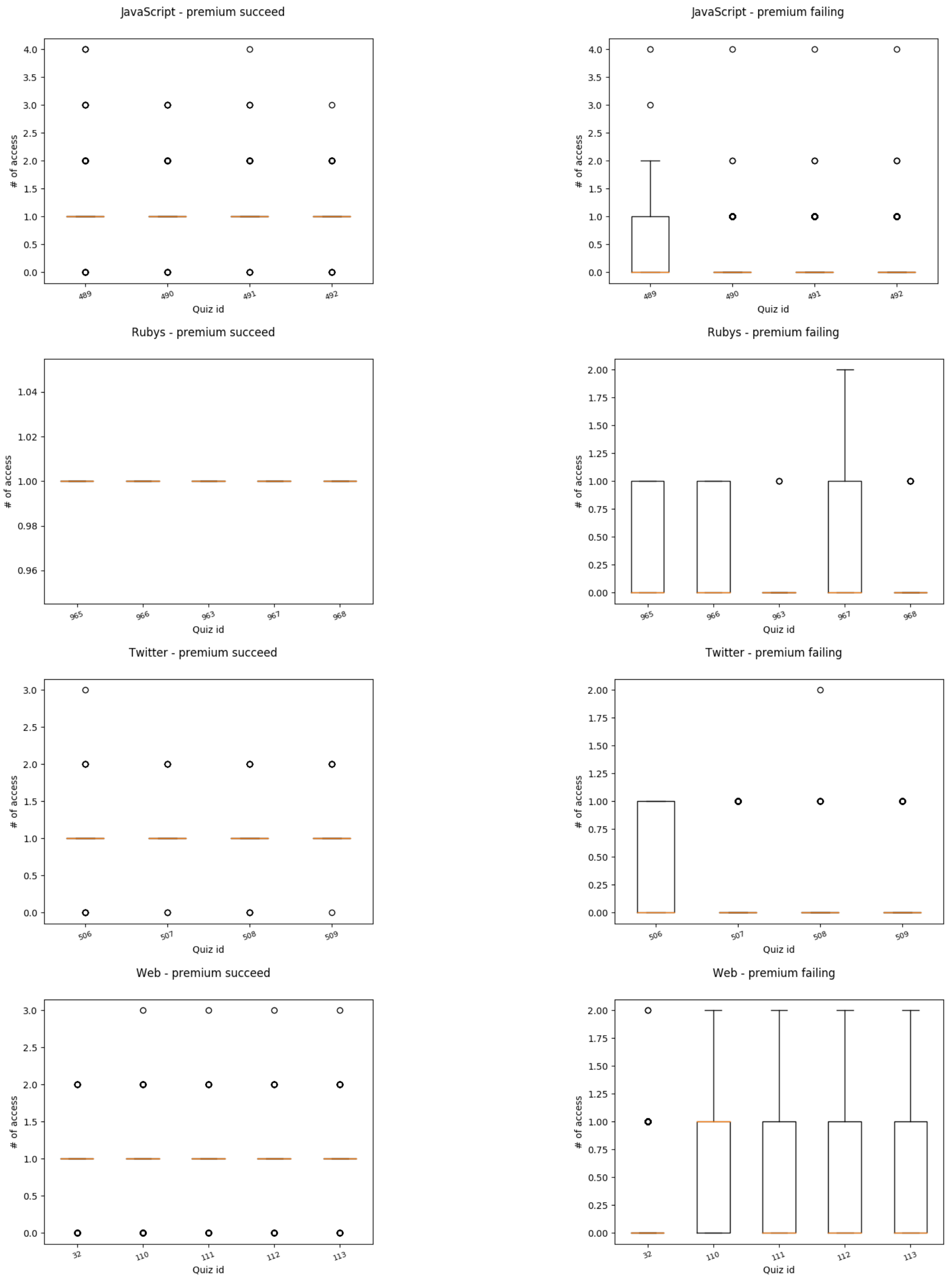


Figure 3.21: Average # of quiz attempts for premium users: graphs a box-plot for the average number (y-axis) of times the premium users attempted each quiz (x-axis) in the courses Arduino, Bootstrap, Gestion Project, and Ionic. An orange bar represents the median of users and points are outliers. The left column displays the successful users and the right shows the failed.


JavaScript - premium succeed

JavaScript - premium failing

Rubys - premium succeed

Rubys - premium failing

Twitter - premium succeed

Twitter - premium failing

Web - premium succeed

Web - premium failing

Figure 3.22: Average # of quiz attempts for premium users: graphs a box-plot for the average number (y-axis) of times the premium users attempted each quiz (x-axis) in the courses JavaScript, Rubys, Twitter, and Web. An orange bar represents the median of users and points are outliers. The left column displays the successful users and the right shows the failed.

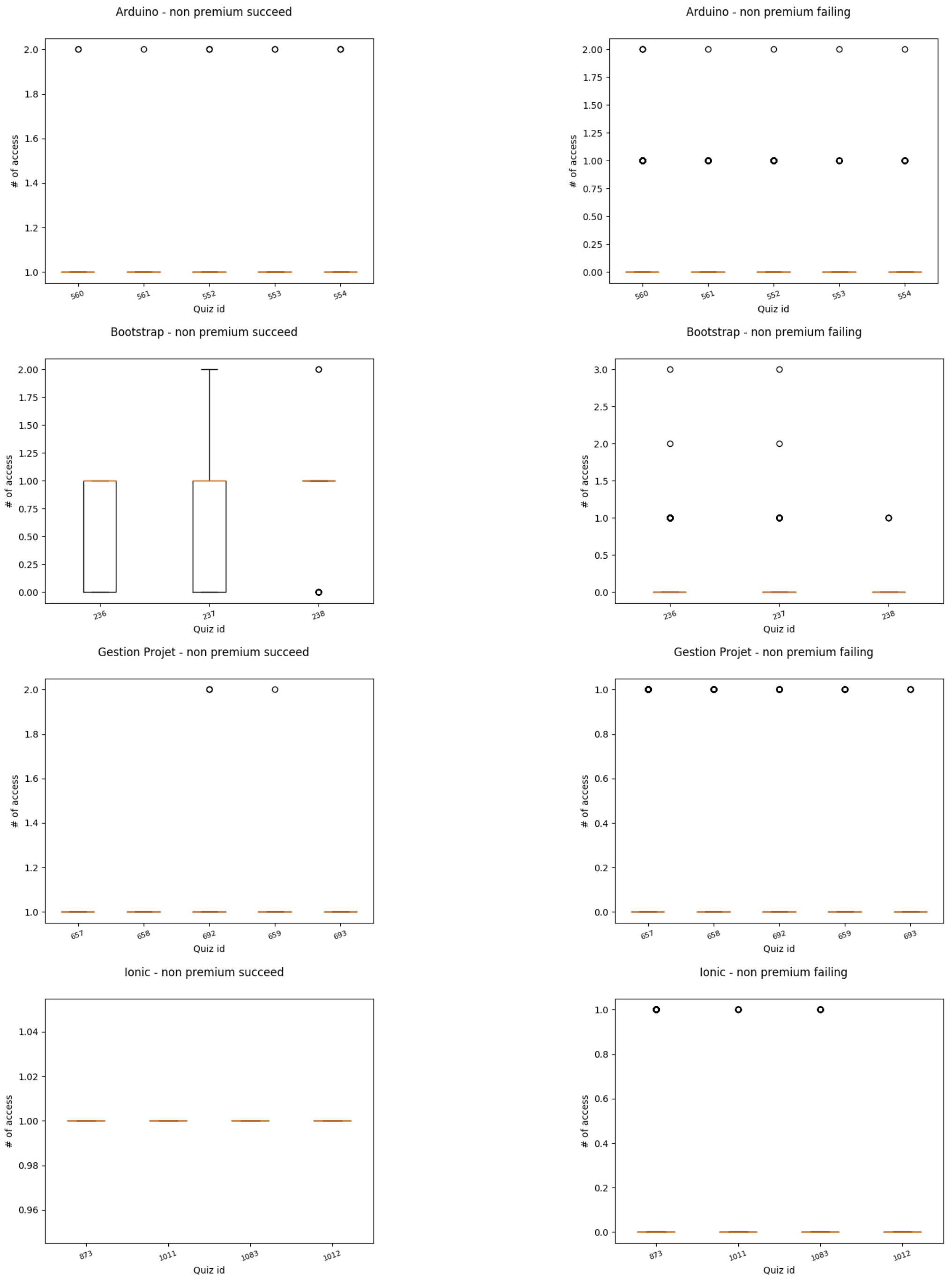


Figure 3.23: Average # of quiz attempts for non-premium users: graphs a box-plot for the average number (y-axis) of times the premium users attempted each quiz (x-axis) in the courses Arduino, Bootstrap, Gestion Project, and Ionic. An orange bar represents the median of users and points are outliers. The left column displays the successful users and the right column shows the failed.

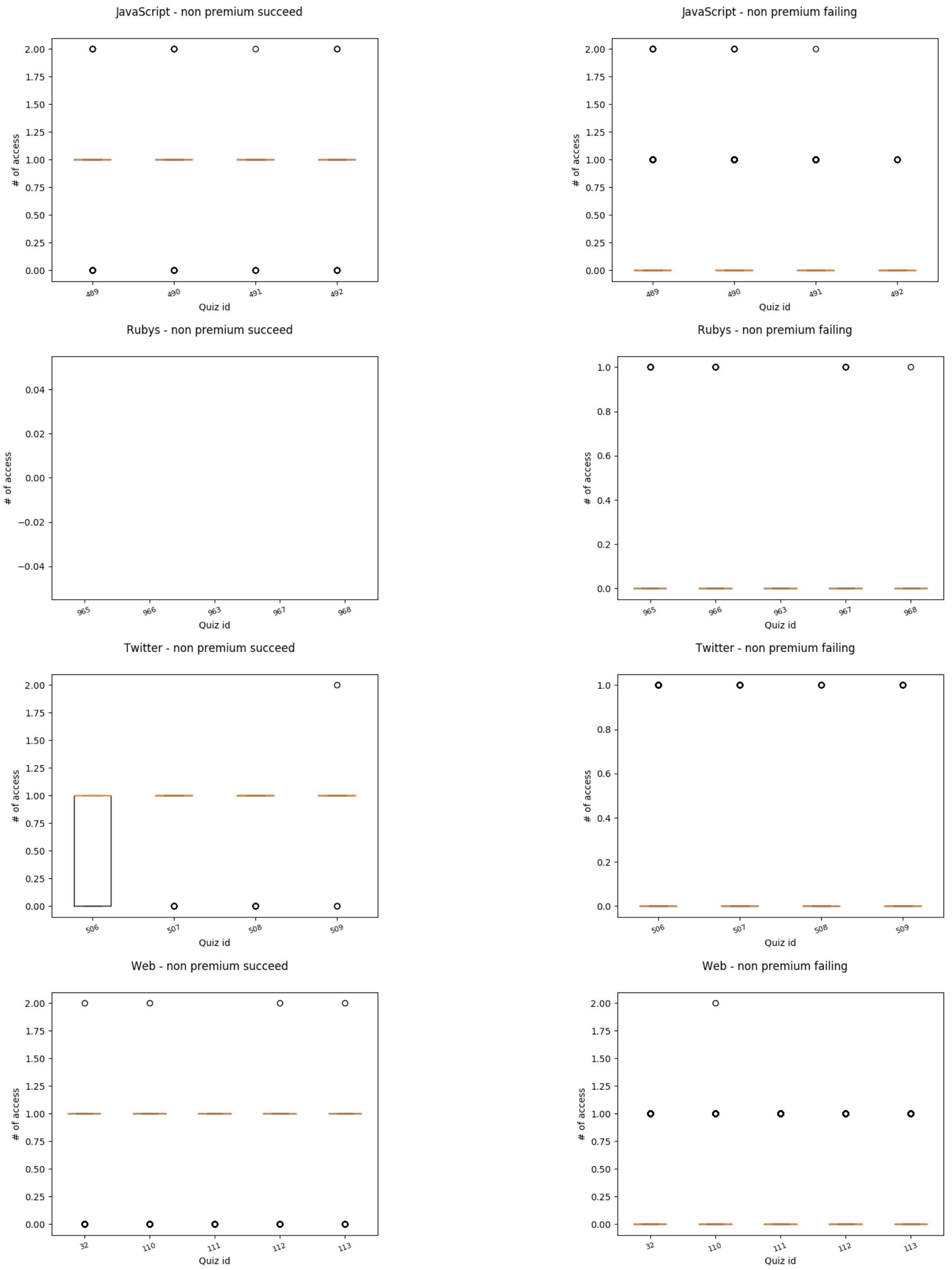


Figure 3.24: Average # of quiz attempts for non-premium users: graphs a box-plot for the average number (y-axis) of times the premium users attempted each quiz (x-axis) in the courses JavaScript, Rubys, Twitter, and Web. An orange bar represents the median of users and points are outliers. The left column displays the successful users and the right column shows the failed.

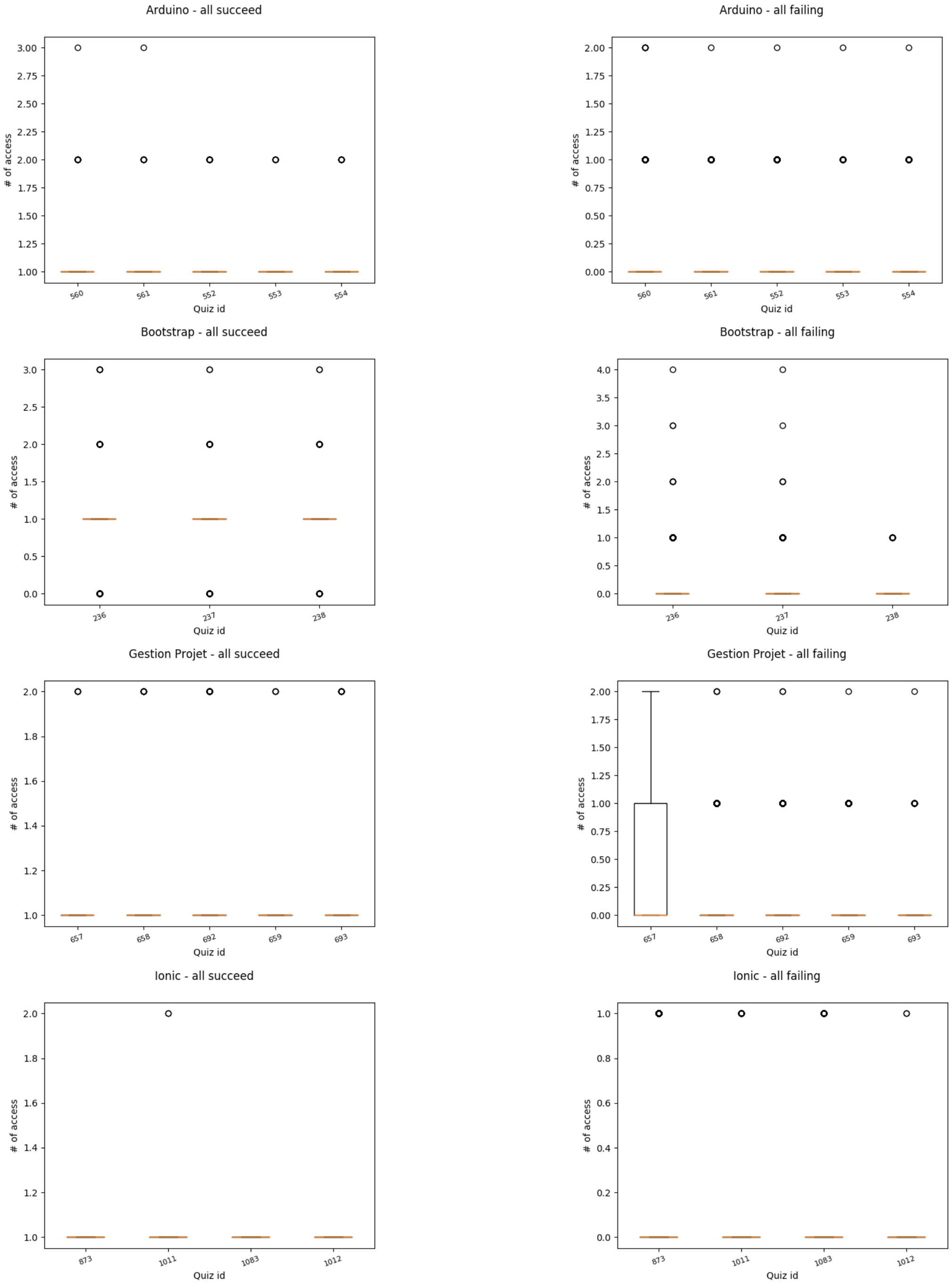


Figure 3.25: Average # of quiz attempts for all(premium and non-premium) users: graphs a box-plot for the average number (y-axis) of times the premium users attempted each quiz (x-axis) in the courses Arduino, Bootstrap, Gestion Project, and Ionic. An orange bar represents the median of users and points are outliers. The left column displays the successful users and the right column shows the failed.

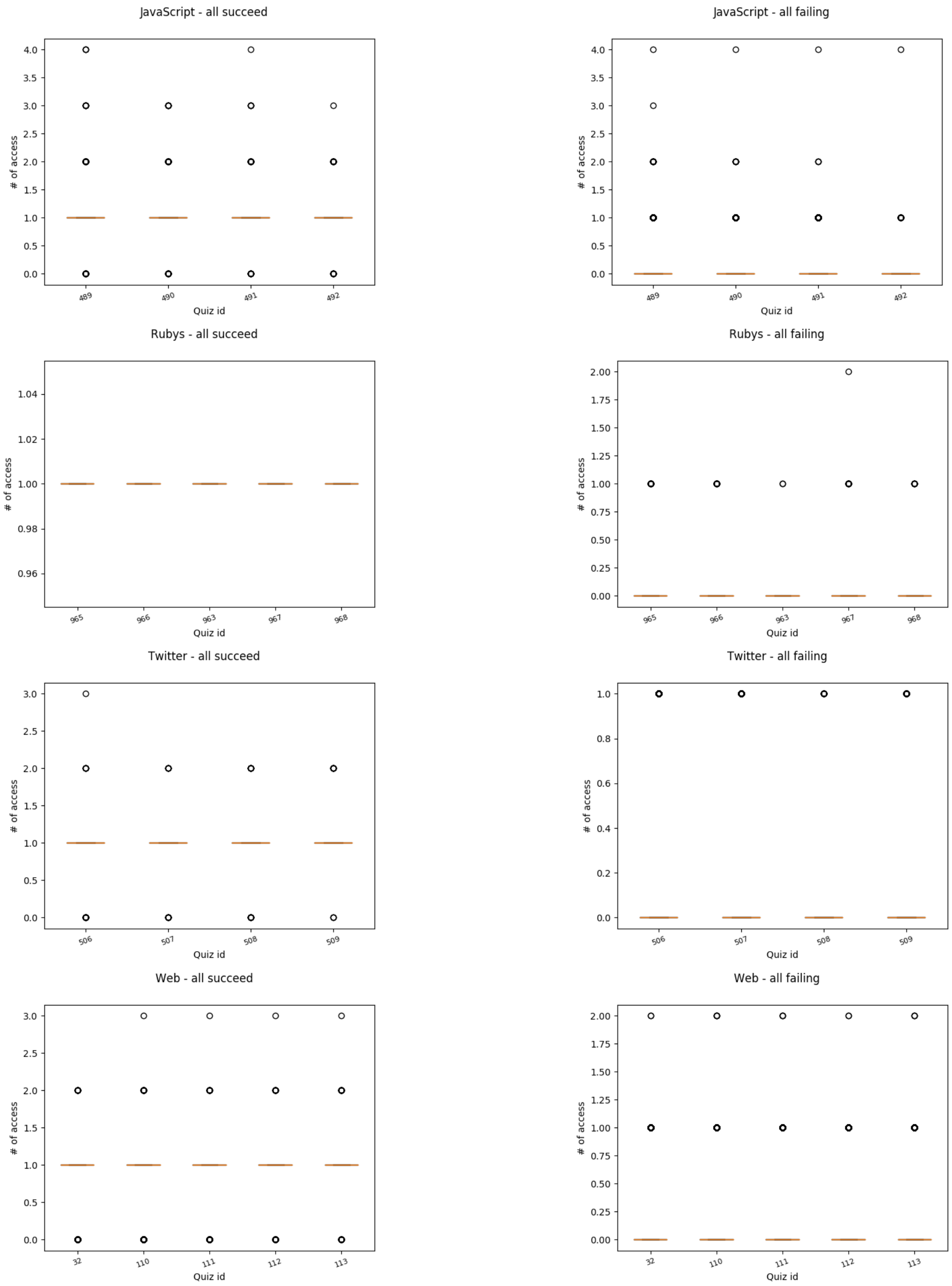


Figure 3.26: Average # of quiz attempts for all (premium and non-premium) users: graphs a box-plot for the average number (y-axis) of times the premium users attempted each quiz (x-axis) in the courses JavaScript, Rubys, Twitter, and Web. An orange bar represents the median of users and points are outliers. The left column displays the successful users and the right column shows the failed.

3.1.7 Average Number of Sessions Before a Quiz Attempt

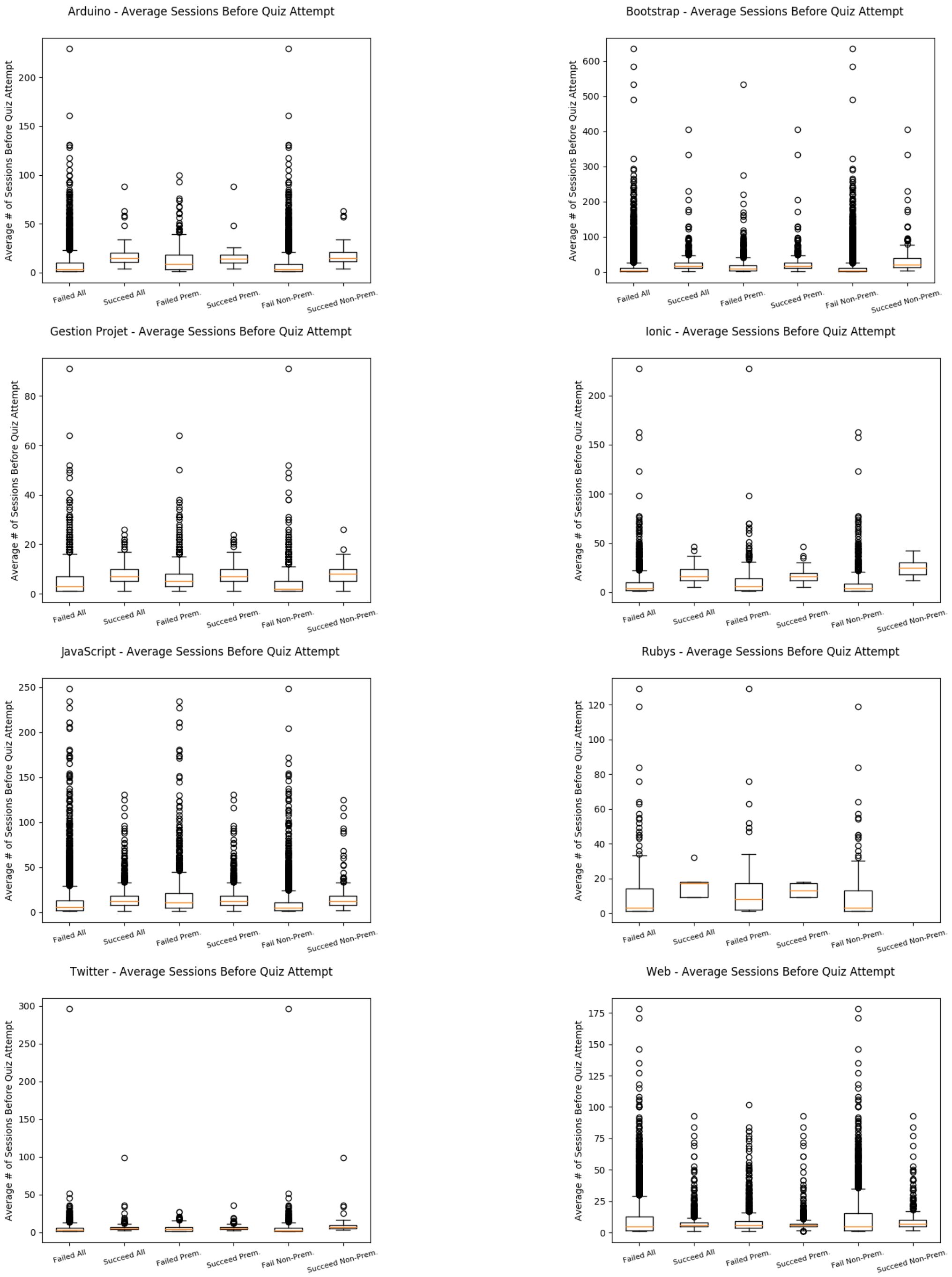


Figure 3.27: Average # of sessions before a quiz attempt: displays a box-plot for 8 courses (Arduino, Bootstrap, Gestion Project, Ionic, JavaScript, Rubys, Twitter, and Web) and for each user group, labeled by the x-axis.

3.1.8 Average Number of Daily Sessions

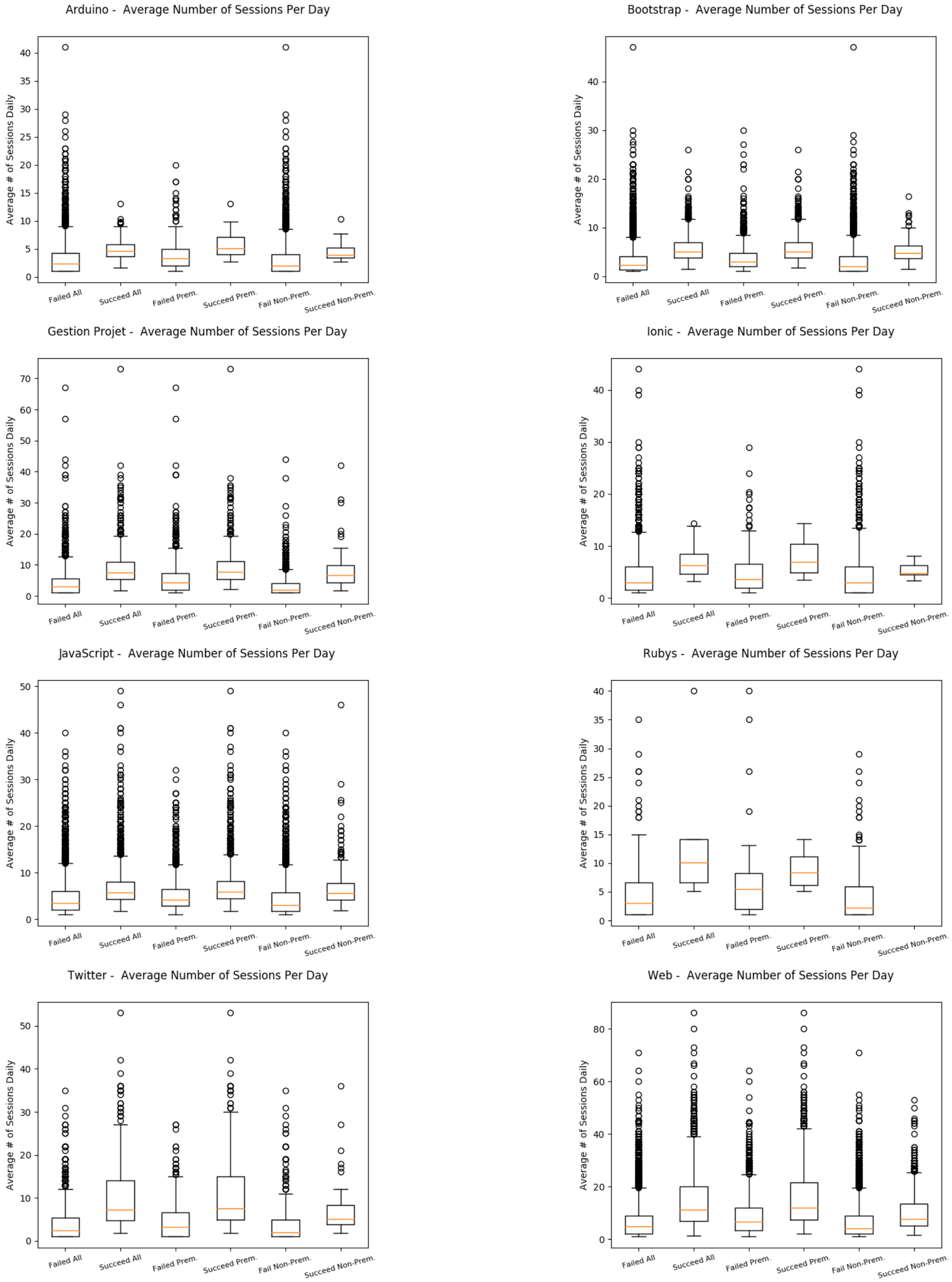


Figure 3.28: Average number of accesses per session: displays a box-plot for 8 courses (Arduino, Bootstrap, Gestion Project, Ionic, JavaScript, Rubys, Twitter, and Web) and for each user group, labeled by the x-axis.

3.1.9 Average Number of Accesses Per Session

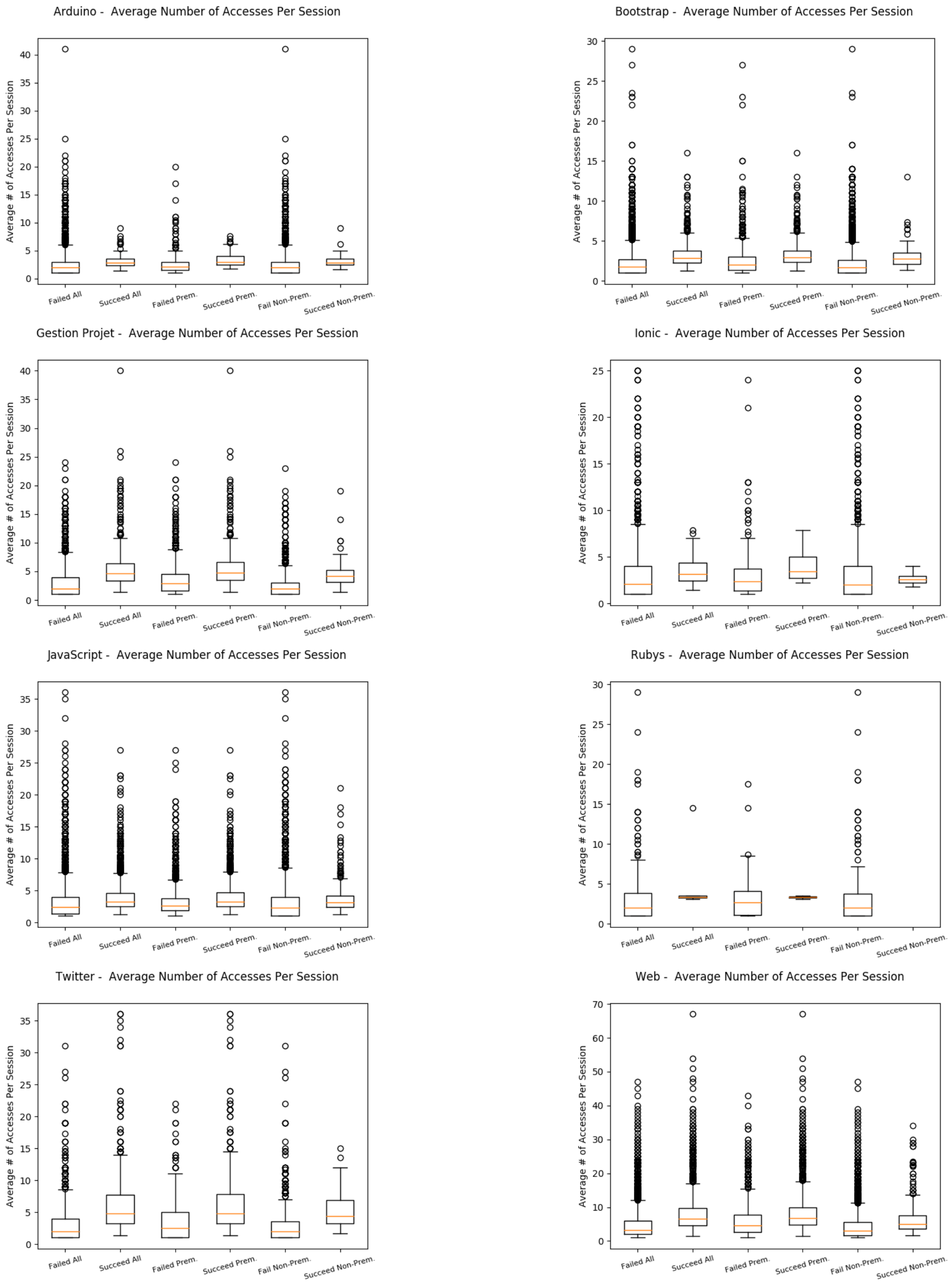


Figure 3.29: Average # of accesses per session: displays a box-plot for 8 courses (Arduino, Bootstrap, Gestion Project, Ionic, JavaScript, Rubys, Twitter, and Web) and for each user group, labeled by the x-axis.

3.1.10 Inter-Session Time

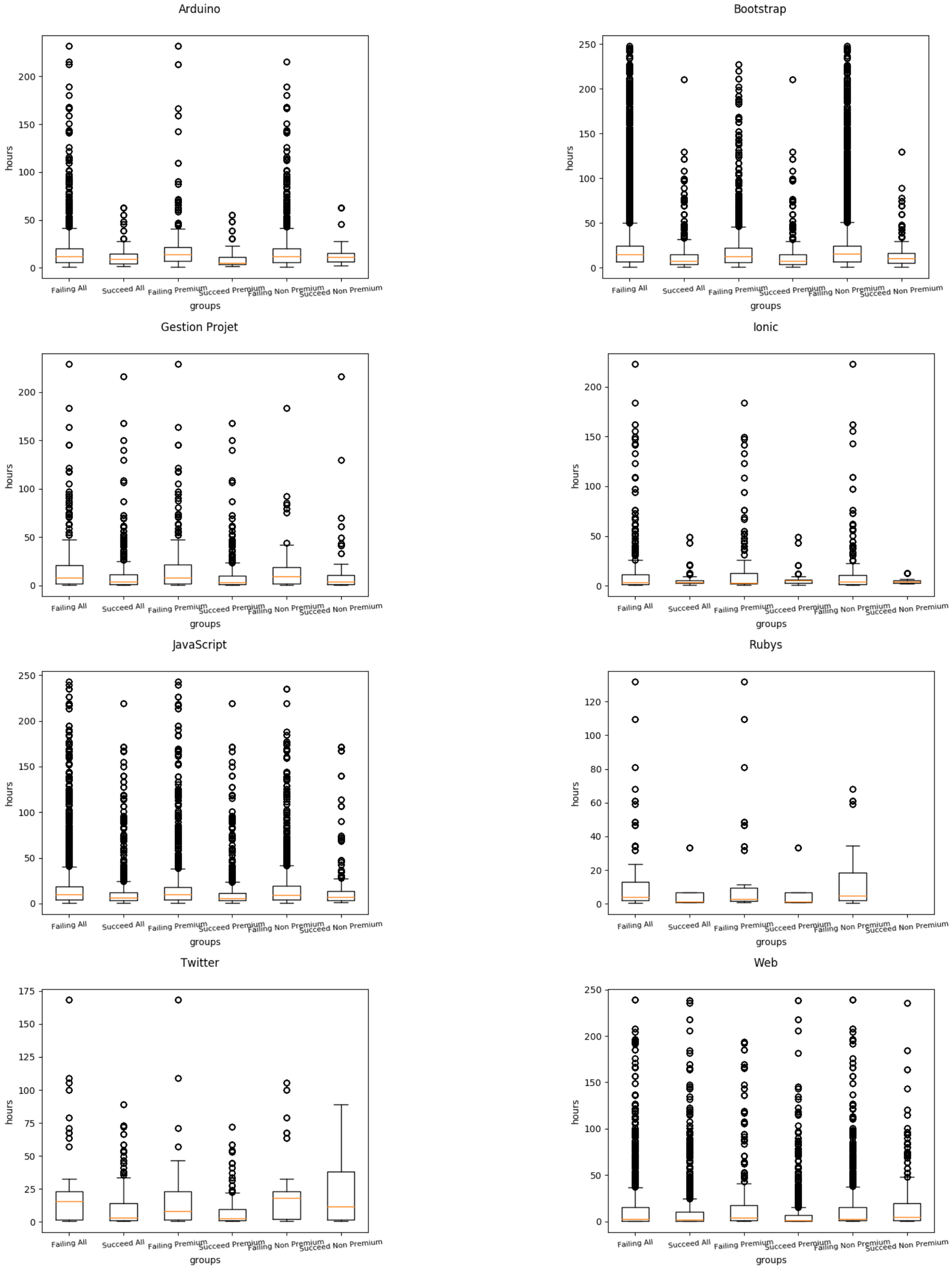


Figure 3.30: Total inter-session time: displays the time elapsed between two consecutive sessions. There's a box-plot for 8 courses (Arduino, Bootstrap, Gestion Project, Ionic, JavaScript, Rubys, Twitter, and Web) and for each user group, labeled by the x-axis. Each point represents a user's median inter-session time.

3.2 Analysis

I will continue to work on defining our discriminant features by analyzing key points in the various histograms and box-plots generated in the previous section. In this section, I will focus on pointing out the trends in each set of figures for the following features: time since last event, final mark, course duration, # of sessions, # of quiz attempts, # of sessions before a quiz attempt, # of sessions per day, # of accesses per session, and total inter-session time.

To make logical and safe observations, I typically found that Openclassrooms' courses with an abundant amount of content and subscribers, such as, Arduino, Bootstrap, JavaScript, and Web, are reliable graphs. Courses such as, Rubys, is helpful for some, but not worth mentioning for most features because the data isn't highly populous. Also, in Rubys, there is no user data in the succeed non-premium user group, so the all succeed group is inaccurate.

3.2.1 Time Since Last Session

This feature proves that successful users tend to have faster and more frequent return rates than the failures. Looking at Figures 3.5,3.6, it's apparent that the majority of failing users have only one session and do not return after one week (168 hours). Similarly, many successful users also do not revisit after a week. Since, there's a large number of users that fall under the "168 hours since last session" category, we can examine the same histograms, but without that large deviation. Looking at Figures 3.11 and3.12, Bootstrap and JavaScript show that the majority of successful users fall close to zero hours, while the failures plot close to 168 hours.

It's interesting to see in these figures that both the successful and failing groups have peaks close to zero and 168 hours. This can be explained in a few ways. The spike at zero can be caused by users frequently spending time on the course platform whether or not they pass/fail the exam, or they may drop-out after a couples hours. To explain the peak close to 168 hours in successful users, some users may begin the course then become busy with external factors that cause them to not return for a number of weeks. Others may quickly skim and complete the course within a couple hours, so they have no purpose to return to the course after. In the case of failed users, these users may be interested in just viewing the course content then immediately dropping-out after less than an hour.

In observation of premium subscriptions, the premium users tended to revisit the course more often than the non-premium. It's safe to assume that premium users are more motivated to view the material because of their subscription.

Refer to Figures 3.1, 3.2, 3.3, 3.4, 3.5, 3.6, 3.7, 3.8, 3.9, 3.10, 3.11, 3.12 to see the graphs mentioned.

3.2.2 Final Mark

The box-plots of the final mark show obvious variation between the successful and failed users. The median of successful users for all courses is 80 or higher, while the median of failed users is zero. These observations could have been made without generating visuals because the final mark is the definition of whether a student is *successful* or *failing* (see Figure 3.13).

Interestingly, the variation between successful premium and non-premium students are consistent, however, the premium failed users show noticeably higher quartile range marks than the non-premium failed users. We can assume that premium students survive throughout the course, attempt the quizzes, but receive an average below the passing threshold, while majority of the non-premium students drop out during the course and do not attempt any quizzes.

3.2.3 Total Course Duration

It's obvious that the successful users spend more time on the course than the failing users. The failing graphs are heavily skewed to the left, where the majority of students spend no more than one hour on the entire course. On the other hand, the histograms of successful users are slightly skewed to the left, where most spend well over an hour on the course. This trend can be explained by users that simply skim the course content for less than an hour. These users may already have a background on the course's topic, and are looking for new content they want to learn about. In this case, it's likely that they only care about the course content, rather than their quiz scores.

Overall, we can assume that if a user's total course duration is an hour or less, he/she will fail or drop-out of the course. *To see the histograms, refer to Figures 3.14, 3.15, 3.16, 3.17, 3.18, 3.19.*

3.2.4 Total # of Sessions

The total number of sessions a user spends on a course is an important feature. The box-plots display noticeable differences between the succeed and failed users. Successful users have overall higher medians and inter-quartile ranges than the failed (see Figure 3.20).

Interestingly, for the failing users, the non-premium subscribers have inter-quartile ranges close to zero, while the premium users appear to have have significantly higher inter-quartile ranges. This means that premium subscribers are more likely to persist, attempt the quizzes, but earn a failing score. Meanwhile, non-premium subscribers probably stop-out after a few or less sessions in the course.

3.2.5 Average # of Quiz Attempts

Looking at the amount of times a user attempts a quiz, there is an apparent contrast between the successful and failed students. The majority of successful users takes each quiz at least once, while the failed students recorded zero attempts.

Similar to the final mark feature, it is an obvious one because a user must pass the quizzes to be a successful user. However, it was interesting to find that the non-premium failed users typically have much lower numbers of attempts than the premium subscribers. For example, in the non-premium users graphs of JavaScript (See Figure 3.24), the highest outlier is 2 attempts and the 3rd-quartile is at zero. Meanwhile, the premium subscribers of JavaScript (Figure 3.22) have their highest outlier at 4 attempts and an upper quartile at 2. From this, it's safe to again assume that non-premium subscribers either drop-out, or they may simply want to scroll through the material, but do not care to take any quizzes. *To see all box-plots, refer to Figures 3.21, 3.22, 3.23, 3.24, 3.25, 3.26.*

3.2.6 Average # of Sessions Before A Quiz Attempt

The box-plots for this feature prove that the # of sessions before a quiz attempt can depict a successful user from a failed one. In this case, Arduino is a solid example because it contains 5 course quizzes. The median for all failed users is close to zero sessions while the all succeed users show a median of roughly 10 sessions before a quiz attempt. This trend continues over all 8 OpenClassrooms courses.

It is again noticeable in the Arduino course that the failed, premium subscribers have a higher number of sessions than the failed, non-premium users who mainly have zero sessions before a quiz (see Figure 3.27).

All in all, users that fail the course tend to have few sessions and jump straight to the quizzes without prior studying.

3.2.7 Average # of Daily Sessions

The number of daily sessions box-plots shows a prominent separation between the successful users from the failures. Each graph shows that the successful students have higher medians than the failed. For instance, in Web, the median of all successful users lies at around 10 daily sessions, and the all failed have a median of about 3 daily sessions. For the failures, the premium subscribers have higher medians than the non-premium subscribers in all 8 courses (see Figure 3.28). In conclusion, a user's persistence and consistency is key for passing the course.

3.2.8 Average # of Accesses Per Session

An apparent trend is shown when displaying box-plots of users' average number of accesses per session. Across all 8 OpenClassrooms courses, the successful users have higher medians than the failures (see Figure 3.29). All in all, the more times a user reads course material or completes an exercise, the more likely they are to pass.

3.2.9 Average Inter-Session Time

Based on the displays, the average amount of time elapsed between two consecutive sessions greatly varies from successful users to failed. For example, Bootstrap shows that the maximum inter-session time for all failed users is 50 hours, while the maximum time for all successful users is almost half the amount of hours of the failed. This pattern continues through all the courses, so we can conclude that users either fail or drop-out take long periods between making a session and returning to the course (refer to Figure 3.30).

Conclusion

In this chapter, I applied the research from the previous chapter to observe user features on OpenClassrooms database. I discovered interesting user behaviors with the help of graphs and descriptive analysis.

First, I compiled numerous histograms and box-plots based on specific user features and the massive data of users on OpenClassrooms' online platform. The graphs visually displayed differences between successful, failed, premium, and non-premium subscriber groups on 8 OpenClassrooms courses.

After, an analysis of the numerous graphs summarized the important points of each feature's discrimination. I found that most successful users revisit the online platform more frequently, recieve higher marks, spend more overall time on the course, have more total number of sessions, attempt quizzes more often, and spend less time between sessions compared to the failed users. It interesting to see that the premium subscriptions seemed to affect the failing users, but not on the successful users. The premium, failing users seemed to work more consistently than the non-premium, failures, but still did not care to pass the quizzes or complete the course.

This chapter will help the final selection of discriminant features in the next chapter. It will conclude this research with a final selection of predictive features, a summary of my findings with OpenClassrooms, and a description of how this information will be beneficial for future researchers and existing MOOCs that want to decrease their dropout rates. Although all MOOCs are designed differently, the selected features will be interesting for future research associated with MOOC drop out.

Chapter 4

Conclusion

I've compiled several user features that can be used by OpenClassroom researchers to help predict user stop-out. I began with a list of 33 user features that can be used more generally for other MOOCs with different success/fail criteria. I pruned the list down to 9 features applicable to OpenClassrooms and created numerous graphical figures to prove the discrimination between successful and failed students. In this chapter, I will finalize my research findings with a selection of the most-indicative user features in OpenClassroms. Then, I'll discuss how my results can be applied in future work.

4.1 Research Findings

I created a table to gather the factors in OpenClassrooms' students that significantly set apart a successful user from a failed user (see Table 4.1). Then, I will go on to describe the table and summarize my findings throughout my overall research.

4.1.1 Feature Selection

The last phase involves taking the information from analysis and choosing which features should be used in predicting student drop-out in MOOCs. It's crucial to select features that are applicable to our MOOC, show distinct discrimination between success and failed users, and help predict a user drop-out before they actually drop-out.

Feature	**Select**	**Level**
Time Since Last Session After 1-Week	Yes	Average
Final Mark	No	High
Total Course Duration	Yes	High
Total # of Sessions	Yes	High
Average # of Quiz Attempts	Yes	High
Average # of Sessions Before a Quiz Attempt	Yes	High
Average # of Daily Sessions	Yes	High
Average # of Accesses Per Session	Yes	High
Total Inter-Session Time	Yes	High

Table 4.1: This table lists the previously examined user features and whether or not it is has sufficient predictability to detect user drop-out in OpenClassrooms courses. The level column indicates the levels of discrimination.

Now, I can finally define which features help detect a failed (stop-out) user. With a time-mark from a user's first session to 1-week from that, I found that the failed users typically begin a session, then do not revisit the course within 7 days. Another trend I found was failed users normally do not spend more than an hour on the entire course. I explored two quiz-related features, which both proved to be discriminants between successful and failed users. Overall, showing zero attempts on quizzes and a low # of sessions before a quiz are noticeable factors of a user that will fail the course. The session-related features proved that not being consistently active (low averages of daily sessions, low # of accesses per session, and high inter-session times) in the online course is a large factor of failed users.

As shown on Table 4.1, the final mark showed to be an obvious discriminant feature between success and failure, however, it cannot be selected for OpenClassrooms' prediction because it is calculated at the end of the course. Ultimately, the goal is to detect drop-out users before the end, so there is no sense in using final mark as a feature for experiments. However, this feature can be useful for MOOCs that can calculate a user's mark before the end of the course. Final mark is an indicative feature of stop-out users, but not in OpenClassrooms' context.

4.2 Future Work

In Table 4.1, eight features have shown discrimination between successful and failed users. Those selected features may be used for future work that want to detect user drop-out in MOOCs. The information also can be helpful for different experiments, such as, to find reasons why students drop-out, to observe student behaviors on online courses, or to reduce stop-out in existing online courses. It's encouraged for any future MOOC-related experiments to use Table **??** and the grapical figures provided in Figure 3.1 through Figure 3.30 as aid to their own research.

Bibliography

[1] Anderson, A., Huttenlocher, D., Kleinberg, J., and Leskovec, J. (2014). Engaging with massive online courses.

[2] Balakrishnan, G. (2013). Predicting student retention in massive open online courses using hidden markov models.

[3] Bayeck, R. Y. (2016). Exploratory study of mooc learners demographics and motivation: The case of students involved in groups.

[4] Chen, Y., Chen, Q., Zhao, M., Boyer, S., Veeramachaneni, K., and Qu, H. (2016). Dropoutseer: Visualizing learning patterns in massive open online courses for dropout reasoning and prediction.

[5] Guo, P. J. and Reinecke, K. (2014). Demographic differences in how students navigate through moocs.

[6] Halawa, S., Greene, D., and Mitchell, J. (2013). Dropout prediction in moocs using learner activity features.

[7] luis (2016). Attrition and retention aspects in mooc environments.

[8] Ren, Z., Rangwala, H., and Johri, A. (2010). Predicting performance on mooc assessments using multi-regression models.

[9] Taylor, C., Veeramachaneni, K., and O'Reilly, U.-M. (2014). Likely to stop? predicting stopout in massive open online courses.

[10] Whitehill, J., Williams, J., Lopez, G., Coleman, C., and Reich, J. (2015). Beyond prediction: First steps toward automatic intervention in mooc student stopout.