Impact of Covid-19 Pandemic on Student Participation in an Intro CS MOOC

by

Christopher Glendon Matthew Mauck

B.S. Computer Science and Engineering
Massachusetts Institute of Technology 2020

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Engineering in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2022

Author .....
Department of Electrical Engineering and Computer Science
January 21, 2022

Certified by .....
Ana Bell, PhD
Lecturer, edX Instructor
Thesis Supervisor

Accepted by .....
Katrina LaCurts
Chair, Master of Engineering Thesis Committee

Impact of Covid-19 Pandemic on Student Participation in an Intro CS MOOC

by

Christopher Glendon Matthew Mauck

Submitted to the Department of Electrical Engineering and Computer Science
on January 21, 2022, in partial fulfillment of the
requirements for the degree of
Master of Engineering in Electrical Engineering and Computer Science

Abstract

The impact of the COVID pandemic spreads far and wide, encompassing nearly all aspects of society. One important community that has been forced to enter uncharted territory is academia. Although many students and instructors were subjected to new tools such as virtual lectures, one platform that remained unchanged throughout is the MOOC (Massive Open Online Course) platform edX: a platform that enables students around the world to engage in academia through an online, virtual environment. In efforts to analyze the impacts of the pandemic, this thesis will provide a data-driven survey of the landscape of the introductory computer science course, titled 6.00.1x Introduction to Computer Science and Programming, offered on the edX platform. With enrollment ranging from thirty thousand to one hundred thousand students per run, this edX class provides many individuals with their first taste of computer programming. This large enrollment count provides ample amounts of granular data in efforts to survey pre-covid, beginning of covid, and steady-state covid class runs in the 2019, 2020, and 2021 years respectively. We first aim to take a high level overview of the differences and similarities created by the onset of the pandemic. Then, using various tools and techniques, take a deeper dive into specific aspects of student involvement and interaction to gain useful insights. Finally, we will use these findings to promote and support future iterations of the edX class.

Thesis Supervisor: Ana Bell, PhD
Title: Lecturer, edX Instructor

Acknowledgments

First and foremost I would like to thank God for providing me with countless blessings, immeasurable support, and placing incredible people into my life. I want to thank my parents Dr. Michael and Sherri, my brother Matthew, my incredible girlfriend Christina, my friends Keilon, Eric, Jack and Nick, and my extended family for their unwavering love and support throughout my childhood and into my academic career. In loving memory of my Grandmas Mimi and Nana. These people have molded me into the gentleman, athlete, and student that I am today and I owe them more than I can put in words.

Next, I would like to thank Professor Ana Bell for this opportunity she so graciously provided me with. I appreciate the willingness to be flexible and easy going, especially during this time of the pandemic with virtual meetings and unforeseen and unplanned circumstances. This research aligned well with my interest in data science and modeling, and I am grateful to have helped in producing something of value during a tough time for us all. I would also like to thank all of the teachers, coaches, and supporters I had growing up in West Palm Beach, Florida as well as the professors I have met at MIT for all of their priceless instruction, advice, and wisdom.

Lastly, I would like to thank all of my new friends and teammates that I have had the honor to meet here at MIT who have all made this journey incredibly enjoyable. My journey would not have been anything like it was without the MIT football team and coaches who have allowed me to challenge myself on the field and leave MIT with two NEWMAC Championships. The impressive people at MIT are what drew me to this second-to-none community, and the life-long friendships and connections I have made is what I will cherish the most as I transition into my career in Dallas, Texas.

Contents

1 Introduction **9**

1.1 Motivation 9

1.2 Class Environment 9

1.3 Pandemic Impact 10

1.4 Related Work 11

1.5 Intentions 12

1.6 Data 13

1.7 Limitations 14

1.8 Methods 14

1.8.1 Conventions 14

1.8.2 Data Processing 15

2 Results **17**

2.1 Demographics 17

2.1.1 Age 17

2.1.2 Gender 19

2.1.3 Education 20

2.2 Enrollment 22

2.2.1 By Semester 22

2.2.2 By Country 24

2.3 Engagement 27

2.3.1 Activity 27

2.3.2 Chapters 30

2.3.3 Video Interaction 31

2.3.4 Practice Problems 35

2.4 Performance 36

2.4.1 Pass/Fail 36

2.4.2 Correlations 38
2.5 Machine Learning 42
2.5.1 Feature Visualization 42
2.6 Modeling 50
2.6.1 Model Selection and Preprocessing 50
2.6.2 Model Results 51

3 Conclusion **53**

4 Resources **55**

List of Figures

2-1 Age of Enrolled Students 17
2-2 Percentage of College Aged Students 18
2-3 Enrollment by Gender 19
2-4 Enrollment by Gender by Pandemic Timing 20
2-5 Enrollment by Education Level 20
2-6 Enrollment by Education Level by Pandemic Timing 21
2-7 Enrollment per Semester 22
2-8 Verified Enrollment per Semester 24
2-9 Top Countries by Enrollment 25
2-10 Top 3 Countries by Enrollment by Pandemic Timing 26
2-11 Verified Days Active Boxplot by Pandemic Timing 28
2-12 Verified Average Total Time Spent 29
2-13 Verified Average Total Time Spent per Semester 30
2-14 Verified Average Chapters Visited Boxplot per Semester 31
2-15 Verified Average Videos Viewed Boxplot per Semester 32
2-16 Verified Average Videos Viewed by Pandemic Timing 33
2-17 Verified Average Videos Events per Semester 33
2-18 Verified Percentage of Videos Watched per Semester 34
2-19 Verified Average Problem Checks per Semester 35
2-20 Verified Average Problem Checks by Pandemic Timing 36
2-21 Percentage of Passing Verified Students per Semester 37
2-22 Feature Correlation Matrix 39
2-23 Feature Correlation to Passing Grade by Pandemic Timing 41
2-24 Correlation Variance by Pandemic Timing 41
2-25 Top Features Scatter Matrices 43
2-26 Top 4 Features with Pass/Fail 45
2-27 Top 4 Features Pass/Fail Separate 46

2-28 Top 4 Features with Pass/Fail by Pandemic Timing 47
2-29 Video and Problem Features with Pass/Fail by Pandemic Timing . . 48
2-30 Weak Feature Clustering 49
2-31 GBC Model Confusion Matrices per Pandemic Timing Subset 51

List of Tables

1.1 Semester to Pandemic Timing 13

2.1 Top 5 Enrollment Increases by Counry 26
2.2 Top 5 Enrollment Decreases by Counry 26
2.3 Correlation Strength Value Description 38
2.4 Gradient Boosting Classifier Parameters 51
2.5 GBC Model Performance over Pandemic Timing Subsets 51

Chapter 1

Introduction

1.1 Motivation

As an individual impacted by disruption to the educational environment myself, there is an element of personal motivation that I exhibit to use my knowledge to produce something meaningful from this unfortunate event. Over the course of the pandemic, virtual classes became the new normal which means I also have hands-on experience with virtual learning. Coupling both of these aspects with my coursework and industry experience with computer and data science, I feel that I have the opportunity to provide a meaningful analysis that could potentially drive improvements to future online classes that are exposed to adverse conditions.

At a high level, the tenure, consistency, and enrollment count of this intro CS MOOC allows for a considerable amount of user collected data for analysis. Because of its longevity, we have an established baseline of pre-covid class runs that can be used in direct comparison to beginning-of Covid and steady-state Covid class runs. On top of this, its enrollment count, running from 30k to 100k per semester, will provide a data stream that should be free from too much noise that would prevent an accurate analysis.

1.2 Class Environment

6.0001 and 6.0002 are the foundation classes for aspiring computer scientists, taught by Dr.Ana Bell, PhD here at MIT. The combination of these two classes, named 6.00, are designed to provide an introductory experience to the world of programming in python. In efforts to provide this same content to those not enrolled

here, MIT provides an online variation on the edX platform, in the form of a MOOC, or Massive Open Online Course. This provides anybody with an internet connection the ability to learn python at an introductory level, taught by some of the best professors in the world.

This online course started running in 2014 and is currently still offered as of Fall '21, having 22 runs over this span between the fall, spring, and summer semesters. Enrollment varies from thirty thousand to one hundred thousand per run, with eighty thousand plus already enrolled in Fall '22, and an average of approximately sixty thousand.

Students have the opportunity to pay an enrollment fee in order to access exam assessments, but payment is not required to view lecture videos, do exercises, and problem sets. For the purpose of this thesis, we will be primarily looking at those students who paid for those additional features. This is because we believe that payment entices students to gain the most from the course, whereas many that sign up for the free version do not have that extra incentive to complete the course in its entirety.

This is backed by the statistic that shows that average completion rate for all enrollments from start to finish varies from just 3% to 5%, whereas the average completion rate for those who pay for a verified certificate is 52%. By focusing on this demographic specifically, we feel that we will get the best look into the impacts of the pandemic.

1.3 Pandemic Impact

One focus of this thesis will be to quantify the impact of the pandemic on virtual learning, specifically in this edX class. Some hypotheses that may be affected by the disruption the pandemic caused are listed below:

- Did the pandemic create more time to work on this MOOC?
- Does more free time imply more engagement?

- Are students more or less inclined to work during the pandemic?
- Did the pandemic increase college aged enrollment?
- How does enrollment size change each semester?
- Does the passing rate increase at any point in the pandemic?

1.4 Related Work

Since the inception of this intro CS MOOC, there have been various thesis projects completed that look at the course through different lenses. The majority of these, after discussion with the course instructor, seemed to focus on student performance, class structure, and academic material. On the contrary, I plan to briefly address these topics, but instead look through the lens of the pandemic and provide a survey of the class' current landscape.

Vonder Haar looked at understanding learner engagement and the effect of the course structure [7]. The body of this work analyzed the changes of the class that took place between the class runs and did an excellent job of breaking down how they affected performance throughout. One interesting analysis this study showed was the impact of some key course-specific parameters such as instructor led vs. self-paced as well as introductory vs advanced.

Bajwa takes a different approach, and looks down to the keyboard stroke level to model student trajectories throughout individual problems [5]. The body of this work continues previous work on the "doer effect" and aims to track student engagement and performance at extremely granular levels, as opposed to the "macro level" that others already have [5].

In the most recent work, Blackwell and Wiltrout discuss the impact of COVID on the engagement and attainment in a similar introductory, online biology MOOC offered at MIT [6]. In this piece, they look at certain metrics such as video clickthrough, problem attempts, and forum engagement in efforts to determine how learning habits changed due to the pandemic. They discovered that the courses enrollment spiked

once the pandemic hit, yet the increase was not sustained. Due to the pandemic itself, they identified that "learner fatigue, changing lifestyle, [and] loss of motivation ... proved detrimental to performance" [6]. This paper does an excellent job addressing engagement and attainment in particular, and the themes found throughout offered great inspiration to this thesis directly.

1.5 Intentions

Due to the pandemic and its impact detailed above, this thesis intends to serve as a survey of the current landscape of this intro CS MOOC. Because we have data from three points in time throughout the pandemic, before, beginning, and in the thick of it, we believe that considerable takeaways can be discovered in efforts to drive improvements and protocols for online learning moving forward.

We first aim to address high level impacts of the pandemic on the course. Looking at data aggregated at the course level, analysis will be done to identify similarities, differences, correlations, and other relationships between the three points in time. This high level analysis will be the starting point and foundation for further, more in depth discovery.

Once the foundation is set, we then will incorporate more granular data in order to determine deeper levels of understanding on how students interacted with the course before and during the pandemic course runs. This will begin to set the foundation for classification models.

When both the high level and deeper dive data discoveries are complete, we then turn to develop a classification model and other ML approaches in efforts to identify student and course patterns. This will be an iterative process, and will adapt to and incorporate discoveries along the way.

1.6 Data

The data being used for this thesis is sourced from the MIT Institutional Research database, retrieved from Google's BigQuery. This department collects user data for all of the edX classes and stores them in a digestible format. We will be using the fall, spring, and summer course runs in 2019, 2020, and 2021, respectively. The data is well maintained, consistent, and maintains the same data columns for all of the class runs we are using. Some discrepancies do exist, which will be addressed in further sections.

We define Pre-Pandemic or Pre-Covid to be before the CDC declared COVID to be a pandemic. All of the semesters listed in this category started and finished before anybody really knew what COVID was. The Start of the Pandemic contains two semesters, where during the Spring '20 the CDC declared COVID a pandemic. We consider the Spring '20 to Summer '20 as the proposed inflection point. Because the pandemic was declared half way through Spring '20, we know that the following semester is the first semester to start and end during the pandemic. Finally, the Steady State Pandemic are those semesters starting with Fall '20 and ending with Summer '21. At this point in time, the pandemic had been active for quite some time. The mapping from class run to point in the pandemic and data descriptions are below.

Pre-Pandemic	**Start of Pandemic**	**Steady-State Pandemic**
Spring '19	Spring '20	Fall '20
Summer '19	Summer '20	Spring '21
Fall '19		Summer '21

Table 1.1: Map from semester label to pandemic timing. The pandemic began half way through the Spring '20 semester.

person_course: this dataset consists of data aggregated at the course level. It specifies statistics such as engagement duration, frequency, and diversity. This will be helpful to analyze how the pandemic affected users throughout the duration of the course.

person_course_day: this dataset consists of data aggregated at the individual day level. Similar statistics as dataset above, except details engagement per day. This will be helpful for classifying students with an ML classifier as this provides us a very granular stream to produce solid features.

course_problem: this dataset consists of problem specific data aggregated at the course level. It identifies each problem throughout the course and provides certain stats like average score, number of times attempted, and whether it was A/B tested. This will be helpful to identify impacts of the pandemic on certain problem types.

1.7 Limitations

The scope of this thesis covers only the semesters surrounding the pandemic and will not be addressing semesters prior to Spring '19. This is mainly because using multiple years and dozens of semesters would take away from the focus of this work, which is to address the impact of the pandemic itself, not to analyze the class itself.

Due to only limiting the scope of this thesis to a discrete sample space of class runs throughout the course of the pandemic, validating causations to discrepancies in the data itself won't be entirely possible. This thesis is meant to be a discovery of the landscape developed by the pandemic, with commentary addressing these discrepancies, as opposed to a data driven analysis of the class data.

1.8 Methods

1.8.1 Conventions

Each semester will be labeled as {Summer, Fall, Spring } + {'19, '20, '21}. As for timing, the CDC declared the COVID pandemic on March 11, 2020 which falls in the middle of the Spring '20 semester [3]. We expected going into this discovery that the Spring to Summer of 2020 would be the inflection point for many of the metrics we are analyzing.

The course data is partitioned into three types of students:

1. Audit: students who just made an account and registered for the course
2. Verified: students who paid a fee to have access to exams and a certificate if they pass the course
3. Credit: students using the class to fulfill a requirement at their university

For the beginning of this discovery, we will include all student types mentioned above in order to look at the class from a high level, top down frame of reference. This will provide us with a good foundation to identify demographic and enrollment trends shaped by the pandemic.

Then, to address engagement and performance trends, we will focus on the verified subset of students as they provide the least noisy and most dense data. The monetary fee provides some incentive for completion, and the associated data streams are large and dense enough to look for differences. We will also exclude instructors and community TAs (teaching assistants) at all steps in this process.

The associated graphs will also be labeled with which subset they are representing, if applicable, in order to distinguish which group of students we are looking at. This is important to note because many of the audit students' interactions and engagement are not consistent within the course, and many do not complete the course. The associated data is very sparse and is difficult to find any meaningful patterns. On the other hand, the credit students are a very small sample space in which patterns and finding meaningful trends, patterns, and discrepancies would be difficult. All of the calculations were computed using Pandas, directly on the datasets listed above.

1.8.2 Data Processing

For the most part, the data sets were consistent throughout. For some of the columns, small variations existed that required some minor cleaning before analysis. One simple example of this was the "grade" column. For 2019 and part of 2020, this column consisted of blank entries and grades strictly greater than 0.55, which was

the passing threshold. For the later part of 2020 and 2021, this column had zeroes instead of blanks, and grades in the range [0,1]. This made direct comparison of grades from semester to semester difficult, and ultimately disallowed the grade field to be considered as a continuous value. Instead, we treated the grade value as a binary pass or fail in order to keep consistency from semester to semester.

This theme was consistent throughout all of the data, where we identified discrepancies in how the values were recorded, then established a way to ensure and maintain consistency in effort to not misrepresent anything.

Once all of the semester-to-semester discrepancies were accounted for, we aggregated the data sets into a few different groups: by semester type (spring, summer, fall), by time frame relative to pandemic (before, during, steady-state), and finally into one large collective. By doing this, we were able to isolate each group independently and look at them in more detail. Grouping by semester type gave us a good look into how students of the same semester were affected through the pandemic. Grouping by time frame relative to pandemic allowed us to see changes that COVID imparted on students as they were adjusting to the new normal, and grouping all class runs together gave us an excellent high level overview of the class as a whole.

Chapter 2

Results

2.1 Demographics

To establish a foundation, we first looked at the demographics of the class aggregated over all of the semester runs. The following analysis on all demographic data will be of the same aggregate. The demographic data includes age, gender, location, and education metrics. These data points were taken from an aggregate of all the course runs, combining all of them together for analysis. We filtered these to include all three of the student types listed above students, excluding staff and assistants.

2.1.1 Age

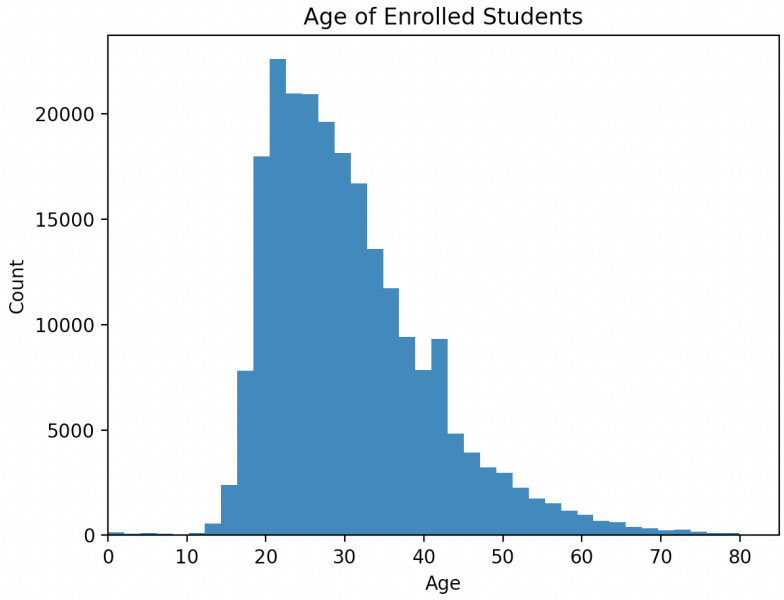

Figure 2-1: Age distribution of all students and all semesters, n=60.

As expected, the age distribution peaks around the college aged group, with a

mean of 30.6, a st. dev. of 10.6, and a mode of 21. This makes perfect sense, as college aged students are most likely to engage with a college level class. This does show, however, that this class has applications to those of all ages, spanning from high school students, all the way to elderly adults.

When looked through a tighter lens at each individual course run, the general shape of the distribution remained the same, and each class run exhibited similar age distributions irrespective of the pandemic timing.

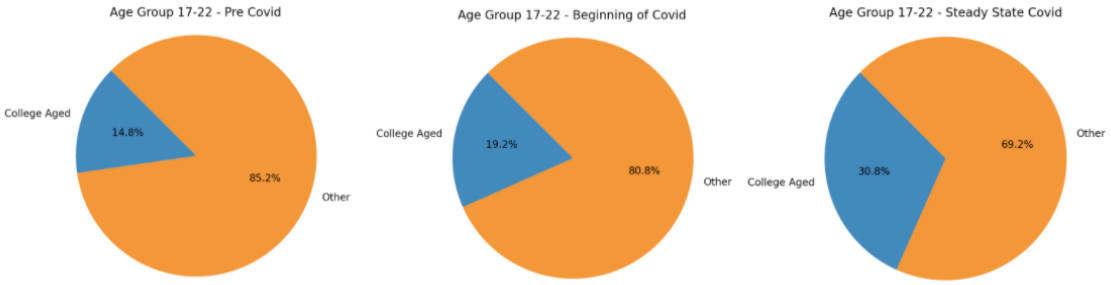

Figure 2-2: Percentage of students age 17-22, by pandemic timing. The blue sections represent the portion of students between the ages of 17 and 22, while the orange sections represent the remainder.

When aggregated by the pandemic timing, we found that at the beginning of and during the steady-state pandemic, we saw a 5% and 10% increase, respectively, in college aged students (defined as ages 17 to 22), enrolled in the course. This was expected, as more college students became exposed to online classes, while the older and younger age groups had to focus on prioritizing other tasks more important than online education. During these times, college students around the world were displaced from the physical classroom and placed into the virtual classroom, giving them more time and accessibility to classes such as this one. Going from 15% to 30%, a 100% increase, over the course of the pandemic illustrates that this course became much more popular for the college aged demographic.

Overall, this age data gives us a good idea of which age groups are interacting with the course, at what times throughout the pandemic. This may have an effect on how the course is perceived at the onset, as well as how the users engage with and are retained throughout. Knowing the composition of the class may give instructors

ideas on how to structure the course.

2.1.2 Gender

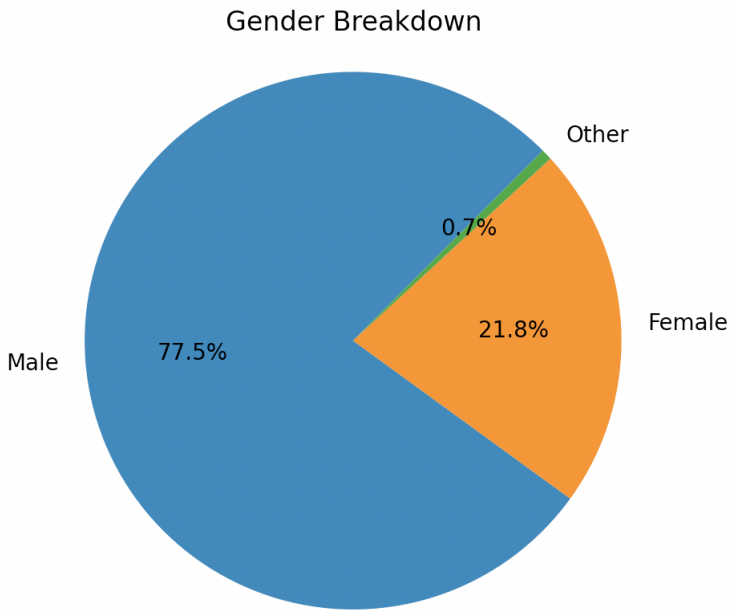

Figure 2-3: Percentage of male, female, and other students, all semesters. Blue represents male, orange represents females, and green represents other.

Looking at the aggregate of all semesters, including pandemic semesters, we see males account for 77.5% of enrollees, females for 21.8%, and others for less than a percentage point. This is somewhat promising, as this class has a female enrollment rate of 3% higher than the United States female CS degree earners, where females earn 18% of computer science degrees [4]. As more initiatives and programs are developed to increase female involvement in STEM, we hope to see these percentages grow closer to one another.

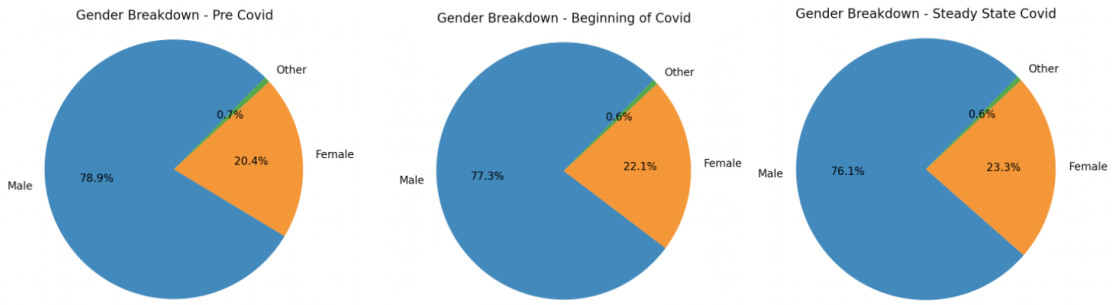

Figure 2-4: Percentage of male, female, and other students by pandemic timing groups. Blue represents male, orange represents females, and green represents other.

Aggregating by pandemic timing, we see slight variations in the gender breakdown. From pre-covid to steady-state, female involvement steadily increased by three percentage points, while male involvement decreased by the same margin.

2.1.3 Education

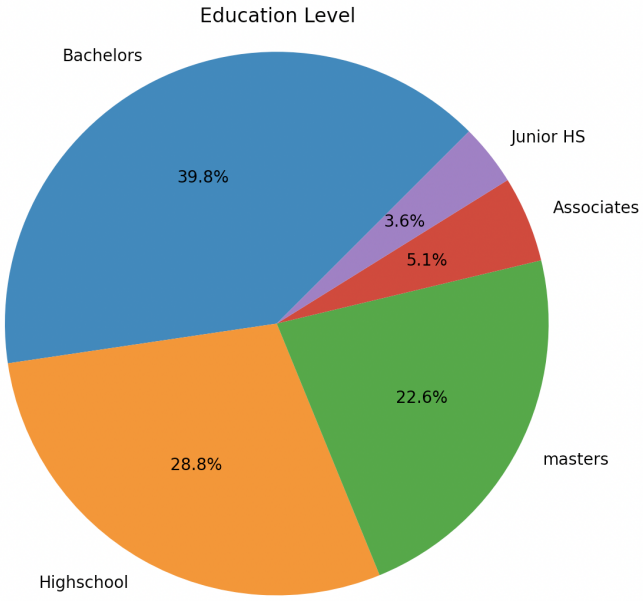

Figure 2-5: Percentage of each education level enrolled, all semesters. Blue represents bachelors, orange represents highschool, green represents masters, red represents associates, and purple represents junior highschool.

We see that when aggregated for all semesters, including pandemic semesters, the

Bachelors education group comes in at nearly 40%, taking the majority, followed by Highschool at 28%, and Masters in 3rd at 22%. This was expected as this course is aimed for college students, pursuing at least a bachelor's degree. This shows in the data, as Bachelor and Master degree level education account for almost two-thirds of the enrollment base. It was interesting to find almost a third of the users were of the high school education level. This is promising because this means that this course is reaching students at a time when they are considering their career path. With this course, they are exposed to the field of Computer Science which could persuade them to pursue further education in CS. We also see that a small but non-negligible amount of Junior HS and community college students are enrolling in the course which can also be used as evidence for the importance of this intro CS MOOC.

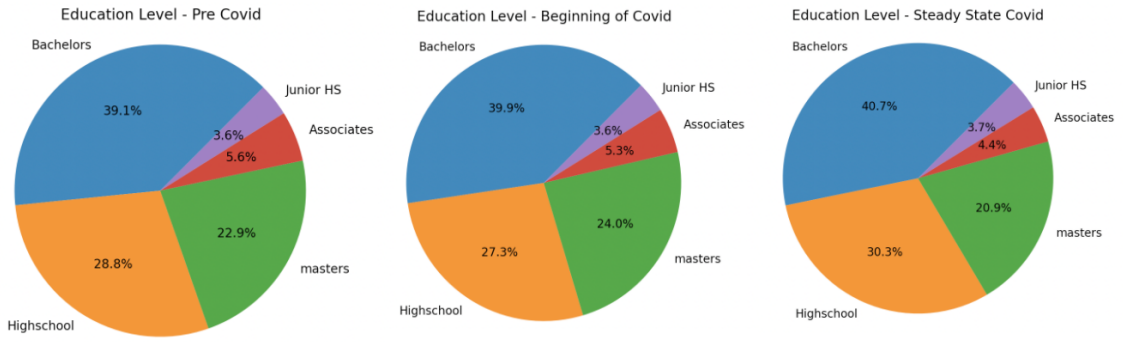

Figure 2-6: Percentage of each education level enrolled, by pandemic timing. Blue represents bachelors, orange represents highschool, green represents masters, red represents associates, and purple represents junior highschool.

When aggregating by pandemic timing, we see small changes in the education levels, none of which are significant. We do see a small 2% overall increase in Highschool students, paired with a similarly sized decrease in Masters students. Bachelor and Associate students vary by less than two percentage points, while Junior HS students are nearly equivalent throughout the pandemic.

This correlates nicely with the age data because we see that the Highschool and Bachelor percentages increase, while Masters, Junior HS and Associates decrease. More students of the 17 to 22 age group enrolled over the course of the pandemic, taking away from other groups. Again, this increase is expected, as this demographic

became incentivized and in some cases forced to take online classes as a result of the pandemic.

2.2 Enrollment

Now that we have analyzed the overall demographic data, let's take a look at enrollment data in efforts to try and determine how the pandemic impacted student registration. For this section, enrollment data will be filtered to all students, including all three types mentioned above. This is because we want a high level, top down look at the class and including all student types is necessary for this. These metrics are crucial to understanding the impact the pandemic had on the enrollment numbers for this MOOC. By analyzing such metrics, we hope to illustrate how the pandemic affected this MOOC through the lens of how many students were signed up to take it.

2.2.1 By Semester

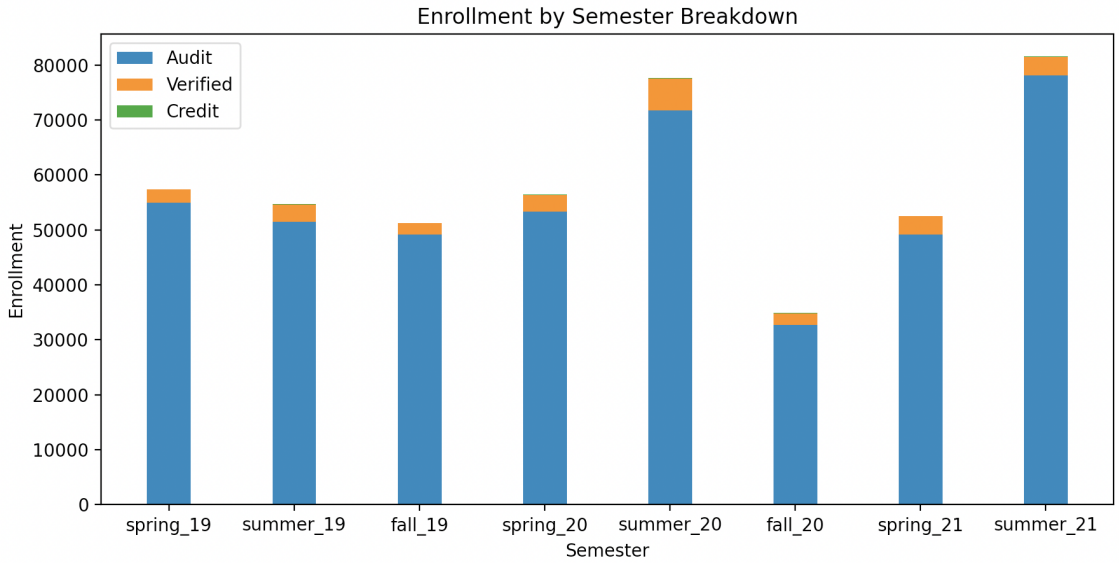

Figure 2-7: Total enrollment count per semester. Blue is audit students, orange is verified, and green is credit.

Taking into account that the pandemic shut down schools during the Spring '20

semester across the world, including MIT, we see an immediate 35% increase in enrollment in the following Summer ‘20 semester. This is probably a direct result of COVID shutdowns, as students were driven to online instruction the following summer semester. Summer ‘20 was the first semester to start after the pandemic was declared, and many schools did not offer in person instruction, driving many students to virtual classes such as this one.

It’s also important to note that the Summer ‘20 semester began prior to 50% of elementary and highschools, and thousands of college universities declaring the following Fall ‘20 semester to be taught entirely virtual [9]. Following the initial upwards spike reaction to the declaration of the pandemic in the Summer ‘20 semester, we then have a sharp 55% decrease the next Fall ‘20 semester. This was most likely due to this semester being the first school-year semester after the declaration, which led many higher education students to take gap semesters, light loads, and sought out alternate forms of education. Many students were unsure if they wanted to continue online instruction without having other options. At the same time, those students who normally would have taken on the challenge of taking an online class, might have refrained from doing so due to the uncertainty of the pandemic.

In the following two semesters, as students began to adjust to the “new normal”, we see the enrollment increase by 35% and 60% the following two semesters, ultimately returning to similar levels as the Summer ‘20 semester.

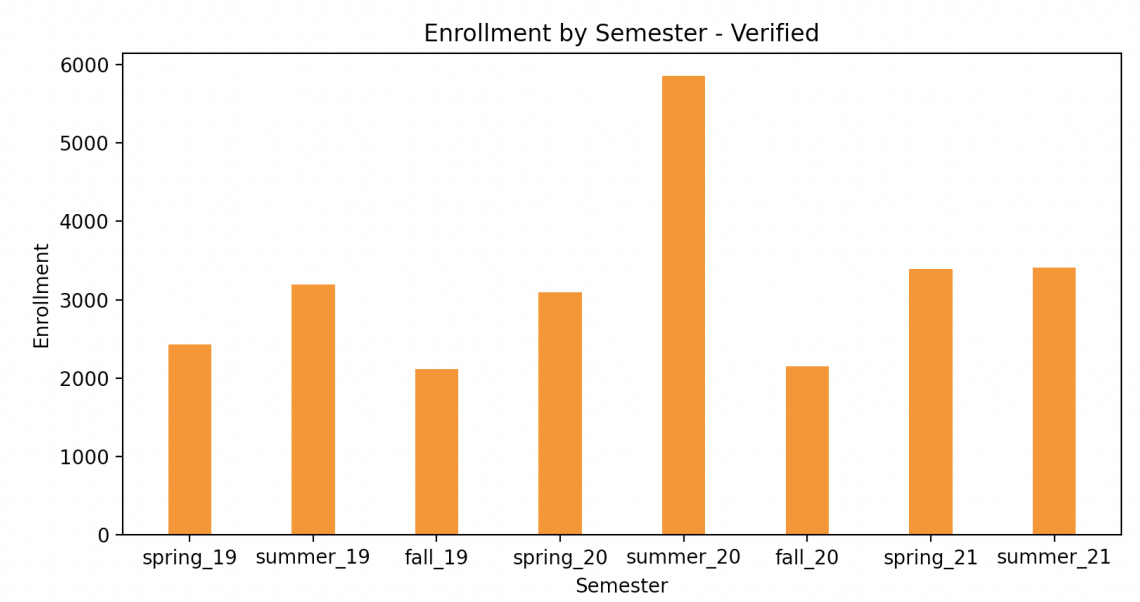

Figure 2-8: Total enrollment count per semester, verified students only.

If we zoom into the orange section from the chart above, we see the verified student enrollment follows the similar pattern as the audit group except from Spring '21 to Summer '21. Between these two semesters, the audit enrollment spikes upwards while the verified enrollment remains at the same level as the prior semester. This could be caused by an overall increase in curiosity in computer science, as curious but not dedicated students would choose the audit track as opposed to the verified track.

2.2.2 By Country

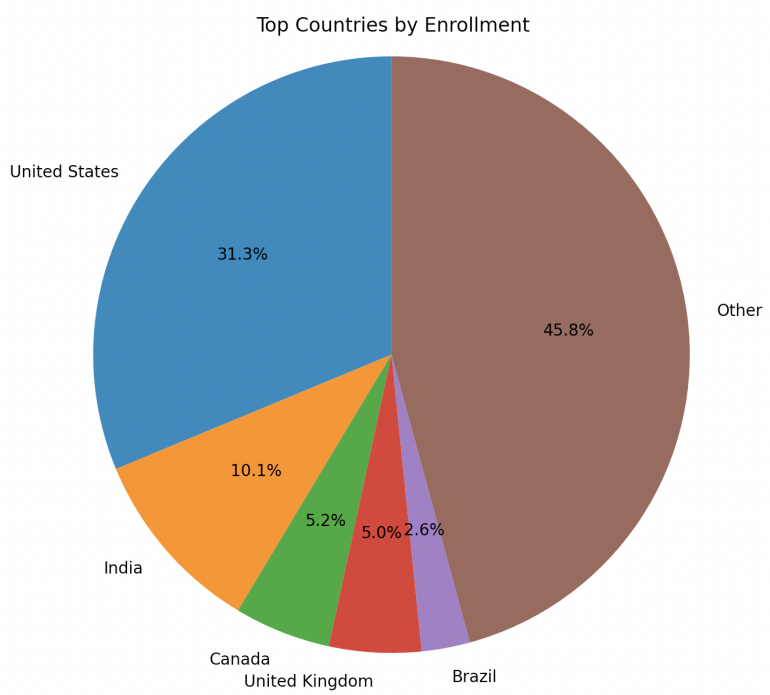

Figure 2-9: Percentage of enrollment by country, all semesters. Blue is United States, orange is India, green is Canada, red is United Kingdom, purple is Brazil, and brown is other.

As expected, the US contributes to nearly a third of all students, with India and Canada coming in at 10.1% and 5.2%, respectively. Outside of the top five countries, we see that 45% of all enrollment is accounted for by countries with increasingly small numbers, with some countries like Turks and Caicos and Eritrea having only one student. This is a great metric that illustrates how much reach online classes such as this one can have, with thousands of students enrolling from countries with low human development indexes [2].

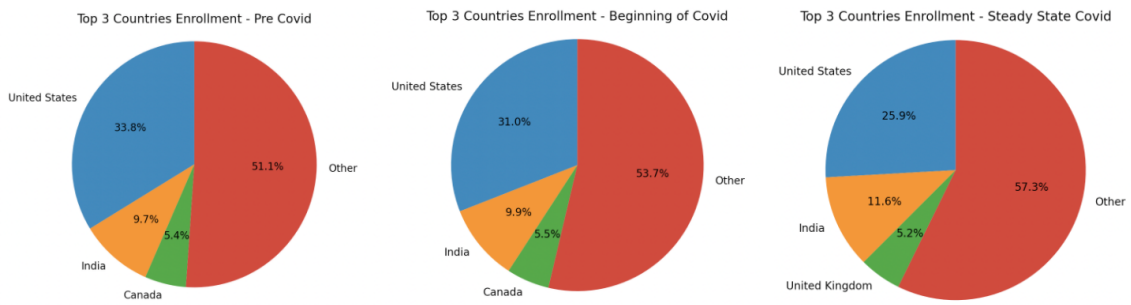

Figure 2-10: Percentage of top 3 countries by enrollment, by pandemic timing. Blue is United States, orange is India, green is Canada, and red is other.

After aggregating by pandemic timing, we notice an interesting trend. We see that the United States decreased its enrollment dominance by 8%, while the other group increased by 6%. This shows that the pandemic increased virtual schooling in smaller countries and provided an opportunity for increased learning. We see from the charts below that the United States' enrollment decreased by over 60% through the pandemic

Rank	**Country**	**Pre-Covid**	**Steady-State**	**Delta**	**% Change**
1	Vietnam	5	488	483	9660%
2	Russia	498	963	465	93.3%
3	Taiwan	100	325	225	225%
4	Korea	326	475	149	45.7%
5	Bangladesh	393	476	83	21.1%

Table 2.1: Top 5 countries with largest enrollment increase from pre-covid semesters to steady state semesters.

Rank	**Country**	**Pre-Covid**	**Steady-State**	**Delta**	**% Change**
1	United States	49,887	16,312	-33,575	-67.3%
2	India	14,345	7,321	-7,024	-48.9%
3	Canada	8,047	2,722	-5,325	-66.1%
4	UK	7,301	3,285	-4,016	-55%
5	Indonesia	3,170	654	-2,516	-79.36%

Table 2.2: Top 5 countries with largest enrollment decrease from pre-covid semesters to steady state semesters.

The charts above provide more specific data to accompany the enrollment pie

charts. We see that Vietnam, Russia, Taiwan, and Korea increased their enrollments by hundreds of students from beginning to steady-state Covid. This corroborates with another analysis done on an astronomy MOOC during the pandemic, which showed similar enrollment trends, where Bangladesh increased their enrollment due to the pandemic while the United States, Canada, and the UK decreased their enrollment [8]. These countries contributed to the 'other' section increasing over the course of the pandemic. On the other hand, the data shows the United States, India, Canada, the UK, Indonesia, and multiple other countries' enrollments declining by the thousands.

2.3 Engagement

Now that we have looked at the class demographics as a whole as well as how they have shifted throughout the pandemic, we now will turn our focus to how students engaged with the class. We aim to determine the impacts of the pandemic and its new normal on how students went about completing, or not completing, this course. Certain data points such as video use, problem interaction, and activity durations will give us a great opportunity to identify patterns caused by the COVID disruption. For this section, the majority of analysis will be on the verified student population, for reasons listed in earlier sections.

2.3.1 Activity

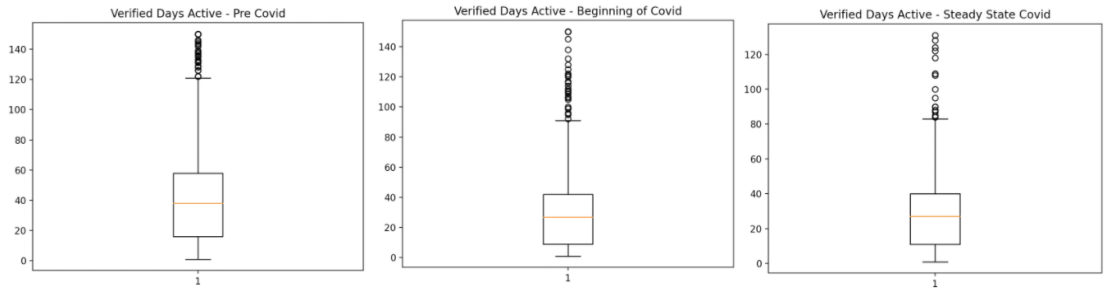

Figure 2-11: Boxplot representing the number of days active of verified students, by pandemic timing. Orange line denotes median, box lines and dashes denote upper and lower quantiles, circles denote outliers.

We define a day to be active if the student logged in to the course on that day. This doesn't necessarily mean they interacted with the course, just that they took the effort to log into their course platform. This is a good starting point, as it gives us an idea how often the students are thinking about and taking the time to log into the online platform. We see from the charts that the average student spent 36 days active in the pre-covid runs, 25 days active at the beginning of Covid, and 26 days active during the steady state. This obviously denotes a decrease in activity within the course, but we need to take a deeper look. We also notice that the plots contain many outliers, with some days entering into the hundreds. These outliers are probably due to the fact that verified students continue to have access to the course material after the course is finished. These outliers show that at the time of this data being recorded, those respective students continue to interact with the course after the course ended, therefore adding active days.

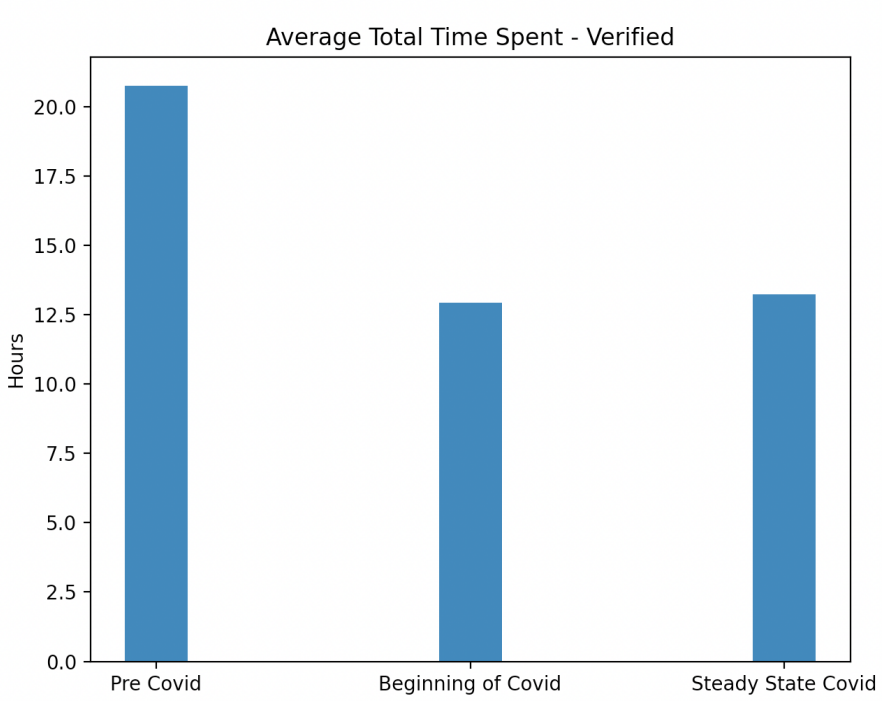

Figure 2-12: Average total time spent, in hours, per verified student. Derived from sum of consecutive events, with a 5 min max cutoff.

We then analyzed the total elapsed time students spent on the course. This metric is calculated by summing the time difference of consecutive interaction events, with a 5 minute max cutoff. Simply put, it's the time students spent using the course platform over the course of the semester. We see that prior to COVID, students were averaging over twenty hours of interaction time with the course. This decreased to thirteen hours of interaction at the beginning of and into steady state pandemic. This denotes the interaction with the course decreased, but does not indicate if the pandemic was the only cause, or a cause at all.

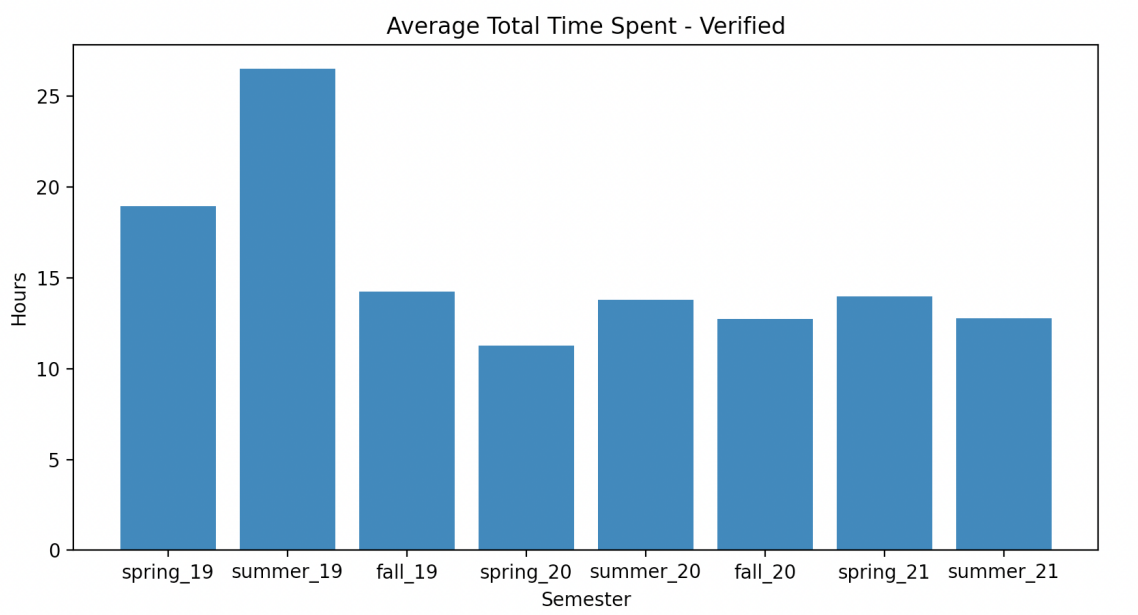

Figure 2-13: Average total time spent, in hours, per verified student, per semester. Derived from sum of consecutive events, with a 5 min max cutoff.

Interestingly, if we look at the same data by semester, we see that the decline in interaction actually began prior to the pandemic, between the Summer '19 to the Fall '19 semesters. This could have been due to changes not identifiable by our data.

Although overall it looks like COVID didn't cause a decrease, we can still look at the semesters going into the pandemic. Starting from Summer '19, we see a downward trend of total time spent, which bottoms during the Spring '20 semester when COVID struck, at eleven hours. Immediately after in the following Summer '20 semester, time spent increased by 27% to fourteen hours, and remained above the Spring '20 low. This could be a result of the pandemic, as those who chose to spend the time and money to take the course, actually had more time to focus on course material. It's possible that without the pandemic, time spent would have continued to decrease or remain at the low levels, continuing the trend of the previous semesters. This is one example of the limitation of sample size and scope of this work.

2.3.2 Chapters

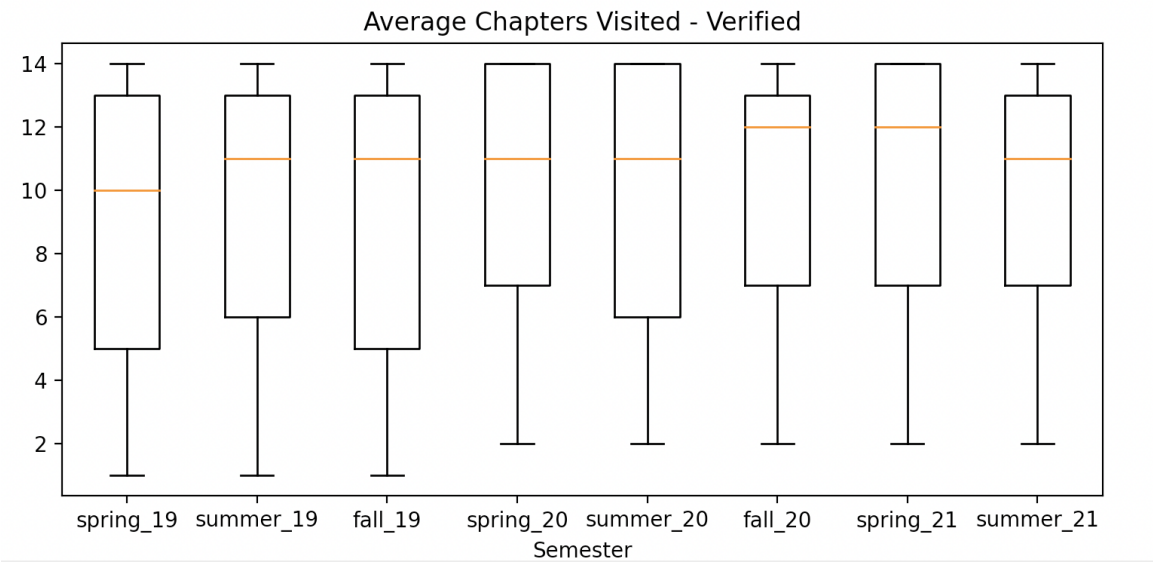

Figure 2-14: Boxplot representing the number of chapters visited by verified students, by semester. Orange line denotes median, box lines and dashes denote upper and lower quantiles, circles denote outliers.

Having a total chapter count of fourteen, we don't see any glaringly obvious discrepancies or patterns about the average number of chapters completed over each semester. It makes intuitive sense that visiting more chapters in the course generally means more effort and therefore a higher chance of passing the course. It can be seen slightly that semesters within the pandemic, namely Fall '20 and onwards, do have slightly higher medians, even if they are only one or two chapters worth. It is also worth noting that the lower fourth quartiles also are higher during the pandemic versus prior to.

Both of these could indicate that the pandemic and its associated effects created more time to explore more chapters. Seeing a general increase in chapter viewing, albeit small, still is a positive overall as students are engaging with the course in a meaningful and more frequent manner. We will see later that this metric of chapter visiting is very highly correlated with passing the class.

2.3.3 Video Interaction

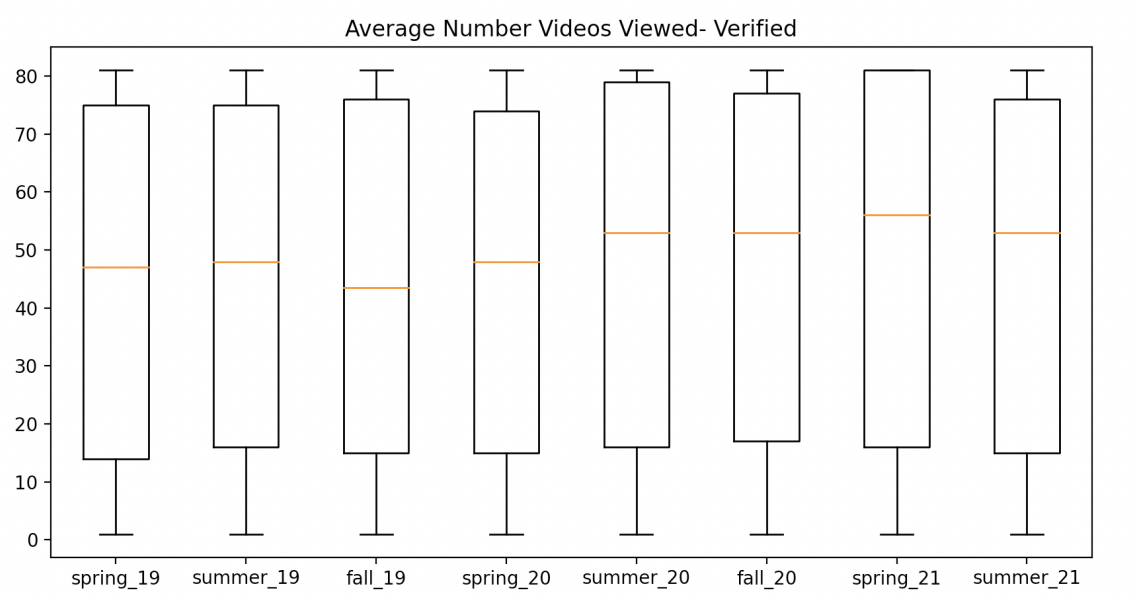

Figure 2-15: Boxplot representing the number of videos viewed by verified students, by semester. Orange line denotes median, box lines and dashes denote upper and lower quantiles, circles denote outliers.

One element of the course that provides a solid insight into user engagement is student video statistics and interactions. Video watching usually requires attention and significant time dedication in order to get the most out of watching them. Similar to the data already presented, we see a similar trend in the average number of unique videos watched per verified student. From the Spring '20 to Summer '20 semester we see a 13% increase in this metric, which is similar to the time data we looked at previously.

The data also shows that the average and upper quartile values are higher throughout the pandemic than they were prior to. If students are watching more videos during the pandemic than prior to, this also means they are engaging with and interacting with the course more frequently, for higher durations of time. These findings also support the claim that the pandemic increased student interaction, through allowing for more time to partake in the course.

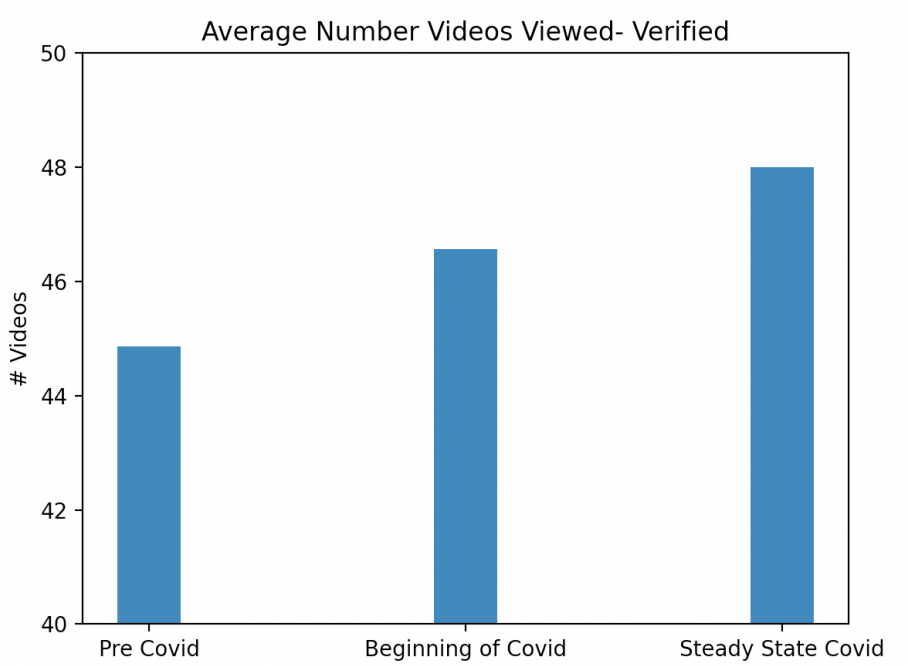

Figure 2-16: Average number of unique videos viewed, per verified student.

If we aggregate by pandemic timing, it's easy to see that holistically, students watched more videos heading into and during the pandemic. Talking numbers, students watched 44.8, 46.5, and 48 videos in the pre, beginning of, and steady state COVID stages, respectively.This represents a 7% increase across the range.

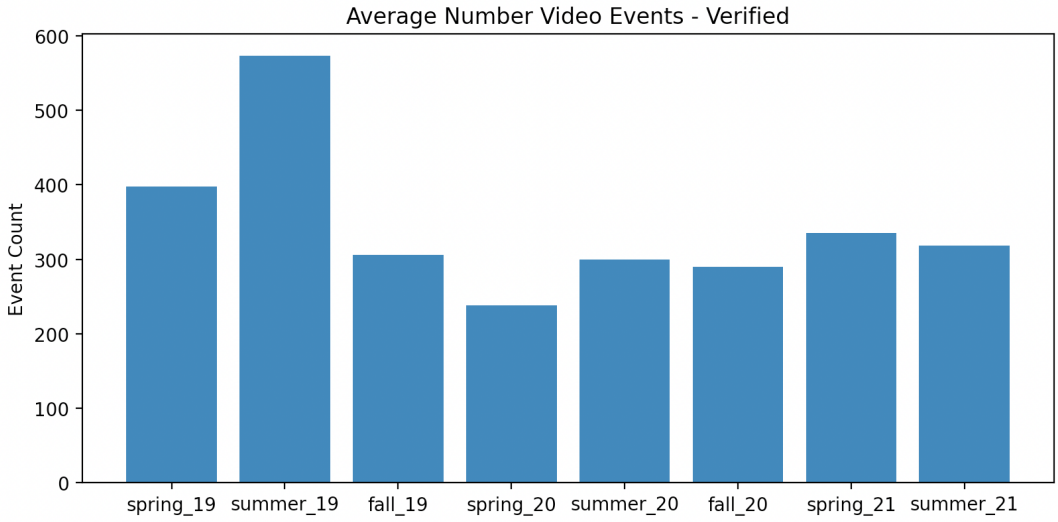

Figure 2-17: Average number of video events (play, pause, scrub, etc), per verified student.

If we analyze the total number of video events, defined as play, pause, seek, etc, we see a direct correlation to the average total time spent graph. Interestingly, we find that the number of video events follows the pattern of the number of videos watched only during the pandemic semesters, and deviates significantly prior to. Between the Spring '19 and Fall '19 semesters, the number of videos watched stays relatively the same when compared to the number of video events. We see the same sharp increase followed by a sharp decrease as was found in the total time spent metric. This still makes sense, as more interactions with the videos also means more time spent on the course.

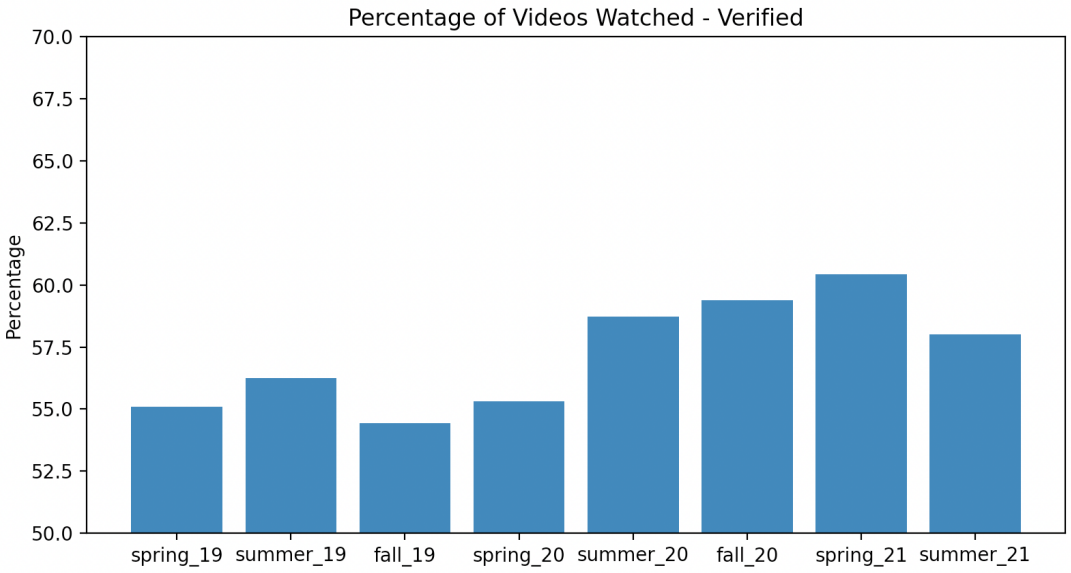

Figure 2-18: Percentage of total videos watched, per verified student, by semester.

We also looked at the total percentage of videos watched over each of the semesters, and found trends that continue to match the data already explored. It is clearly seen that students watched a higher percentage of videos during the pandemic as opposed to prior, as well as a large change between the Spring '20 and Summer '20 semesters. For the first, we see an increase from 54% to 61% of videos watched between the Fall '19 and Spring '21 semesters. For the second, we see a 6.3% increase at the pandemic inflection point, from Spring '20 to Summer '20. These patterns continue to illustrate that the pandemic may have caused an increase in engagement with the course.

2.3.4 Practice Problems

Another indication of student engagement is their interaction with the practice problems provided for their learning. These are intended to be done at the students' pace and can be completed at any time throughout the course. In the following sections we will take a similar approach as we did with the videos, identifying patterns throughout the different course runs.

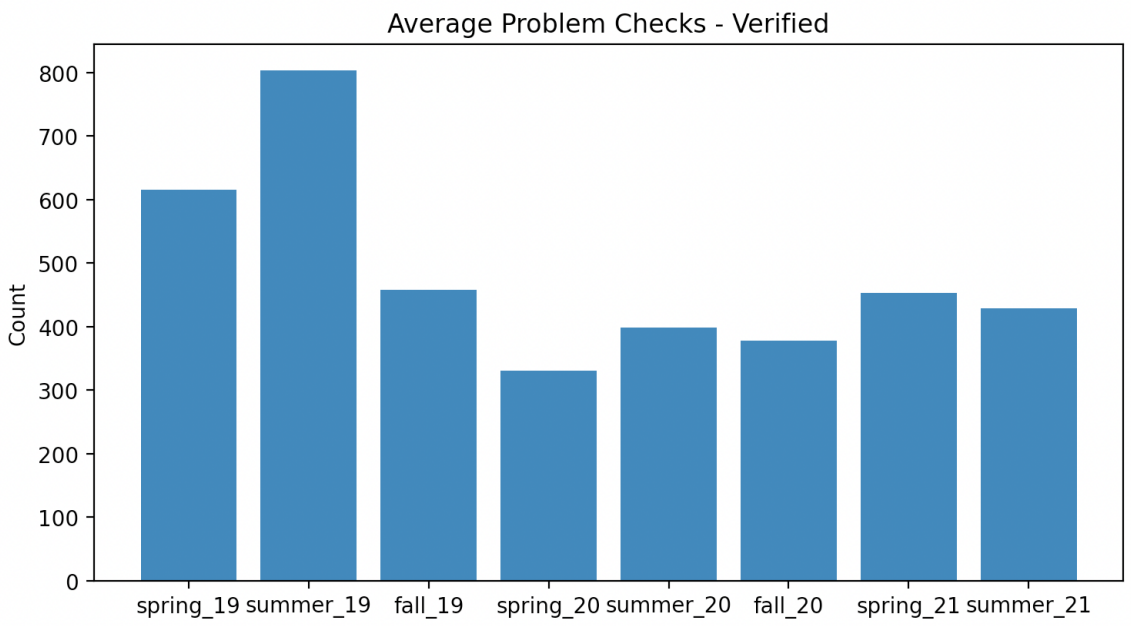

Figure 2-19: Average number of problem checks, per verified student, per semester.

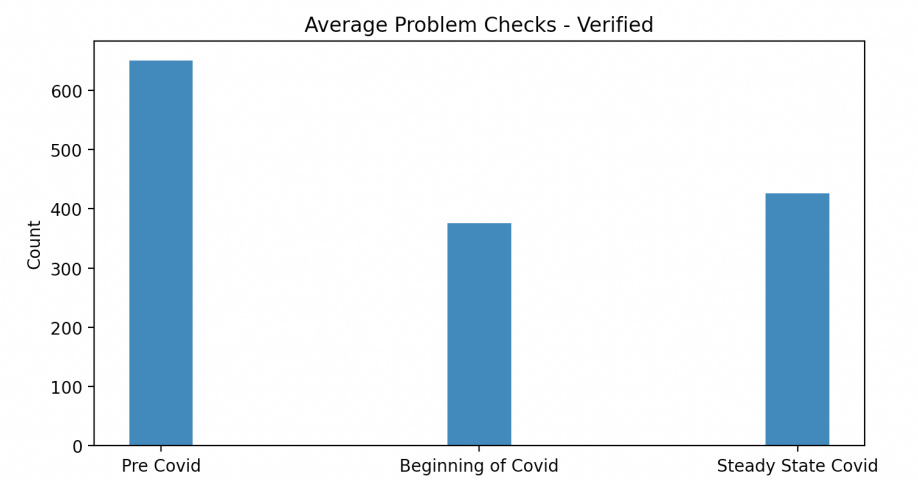

Figure 2-20: Average number of problem checks per verified student, by pandemic timing.

Remarkably, we see the average number of problem checks follow the same pattern as the number of all video events and total time spent. We see the same downward trend, bottoming out in Spring '20, then recovering 23% in the following semester. We think the same predictions we have already mentioned apply to this pattern as well.

2.4 Performance

Now that we have analyzed video and problem statistics, we now move on in efforts to determine how the pandemic affected the performance of the students. In this course, verified students are evaluated on a pass/fail basis, needing a 55% in the course to pass and receive the certificate. As mentioned in a section earlier, some of the semesters' grades are recorded as continuous in the range [0,1], and some as one of 0, 0.55. For this reason, we will transform the data into a binary pass or fail for our analysis.

2.4.1 Pass/Fail

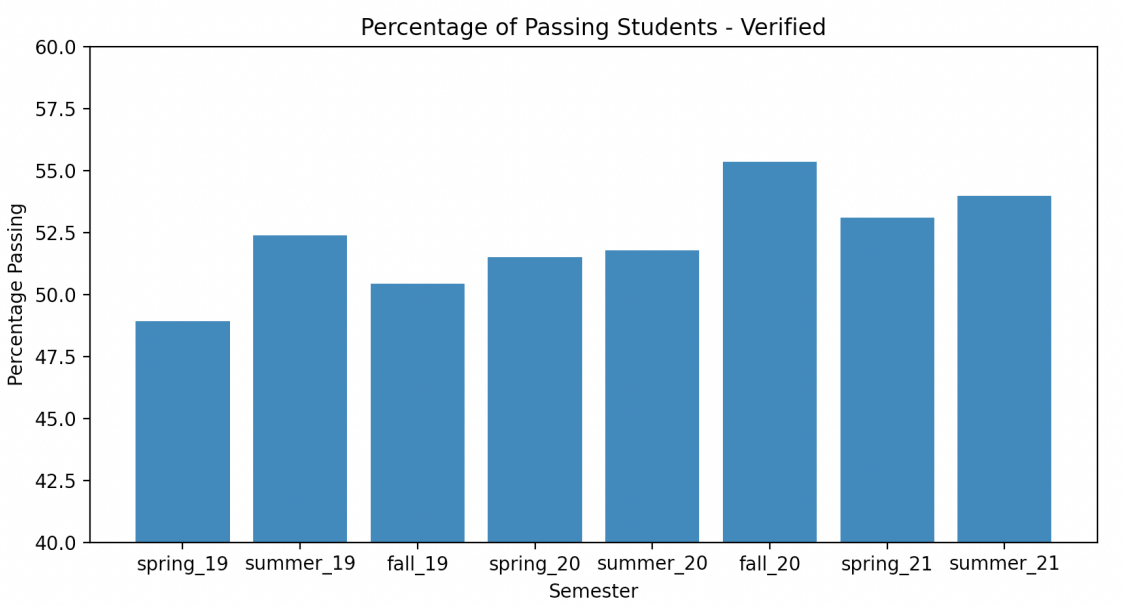

Figure 2-21: Percentage of passing verified students, per semester.

Charting the percentage of passing students each semester, we see that all of the pandemic semesters are higher than the non pandemic semesters. At the inflection point, we see a trivial increase of passing students, however, from the Summer '20 to Fall '20 semester, we see an 8% increase in the number of passing students. The following semesters decrease from this peak, yet stay on the general upward trend of the data.

This increase in the passing rate could be due to the pandemic creating the upward trends in the other associated data, such as video and problem engagement. This makes sense: if students are dedicating more time spent in the course, watching, interacting with, and completing videos and practice problems, it's expected that more of them would also pass the course. On the other hand, factors mentioned above such as material changes and the lessening of grading standards due to COVID could also explain the increase in the passing rate. In the following section, we aim to correlate these metrics with the performance of the students.

2.4.2 Correlations

In order to determine which of the metrics above were meaningful in the increase of the passing rate, we computed correlation matrices. Within the matrix, the intersecting value between the two metrics provides a numerical way to measure this correlation. The higher the value, the higher those two metrics are correlated, and the same exists for lower values.

As well as the above, the correlation matrices will determine which of these parameters will be good features for our machine learning models. The subset of parameters with high correlation values should be a strong feature set to train on, while those with weak values will be excluded from the model. We utilize the Pearson method, the default within pandas, for computing pairwise correlation, using the following table to discern the values [1].

Coefficient Value	**Strength of Association**
$0.1 < \|r\| < 0.3$	small correlation
$0.3 < \|r\| < 0.5$	moderate correlation
$0.5 < \|r\|$	strong correlation

Table 2.3: Quantitative correlation value to qualitative strength association.

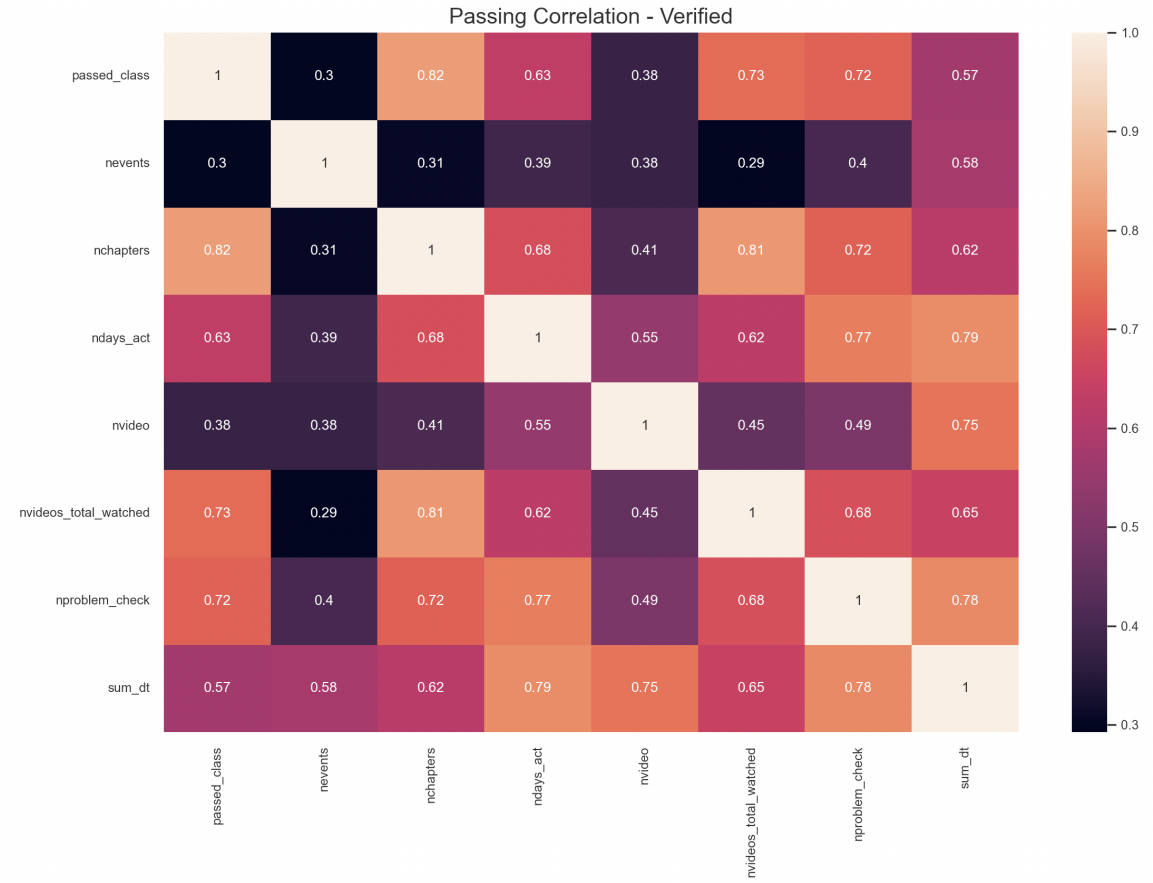

Figure 2-22: Correlation matrix between features: passing the class (*passed_class*), number of total interaction events (*nevents*), number of chapters visited (*nchapters*), number of days active (*ndays_act*), percentage of videos watched (*nvideo*), number of videos watched (*nvideos_total_watched*), number of problem checks (*nproblem_check*), and total time spent in course (*sum_dt*). Higher values represent higher correlations between the two respective features. The lighter the shaded color, the higher the correlation. The darker the shade, the lower the correlation.

We first look at the correlation of parameters taken of the aggregate of the class as a whole to give us a basic understanding of how the intro MOOC operates.

- *passed_class*: can take on two values, one if the student passed the class with a 55% or higher, or zero if the student did not pass the class
- *nevents*: total number of tracking log events for a given student
- *nchapters*: total number of chapters visited by the student
- *ndays_act*: total number of days a student is active in the course

- *nvideo*: total number of video events (play, pause, click, scrub, etc) by the student
- *nvideo_total_watched*: total percentage of all videos watched by the student
- *nproblem_check*: total number of problem checks by the student
- *sum_dt*: total elapsed time (in seconds) spent by the student on this course, based on time difference of consecutive events, with five min max cutoff

Looking at the top row, we can observe the correlation values that affect if the students passed the class or not. The highest value we see, 0.82, comes from the *nchapters* parameter which we determine to be a very strong correlation. This tells us that students that had a higher number of chapters completed were more likely to also pass the class. The lowest value we see is 0.3, which comes from the *nevents* parameter, which tells us that there does not exist a strong correlation between raw activity in the course and passing the class.

Other parameters that exhibit very strong correlations are the total percentage of videos watched (*nvideo_total_watched*), number of problem checks (*nproblem_check*), number of days the student is active (*ndays_act*), and the total time spent on the course (*sum_dt*), with correlation values of 0.73, 0.72, 0.63, and 0.57 respectively. All of these high valued parameters will be used as features in our machine learning model as they exhibit excellent predictive abilities with regards to passing the class. Only one other parameter exhibited a small to medium correlation at 0.38, which was the total number of video interactions (*nvideo*).

We notice from these values that the two parameters that measure total interaction, namely interaction with the videos and interaction with the course, are meaningless in regards to whether or not a student passed the class. On the other hand, parameters that measure more in depth involvement with the course, like chapters and videos completed, are crucial in passing the course at the end. This can be directly applied to future course runs, as we see that students with high chapter completion and higher percentage of videos watched are more likely to pass.

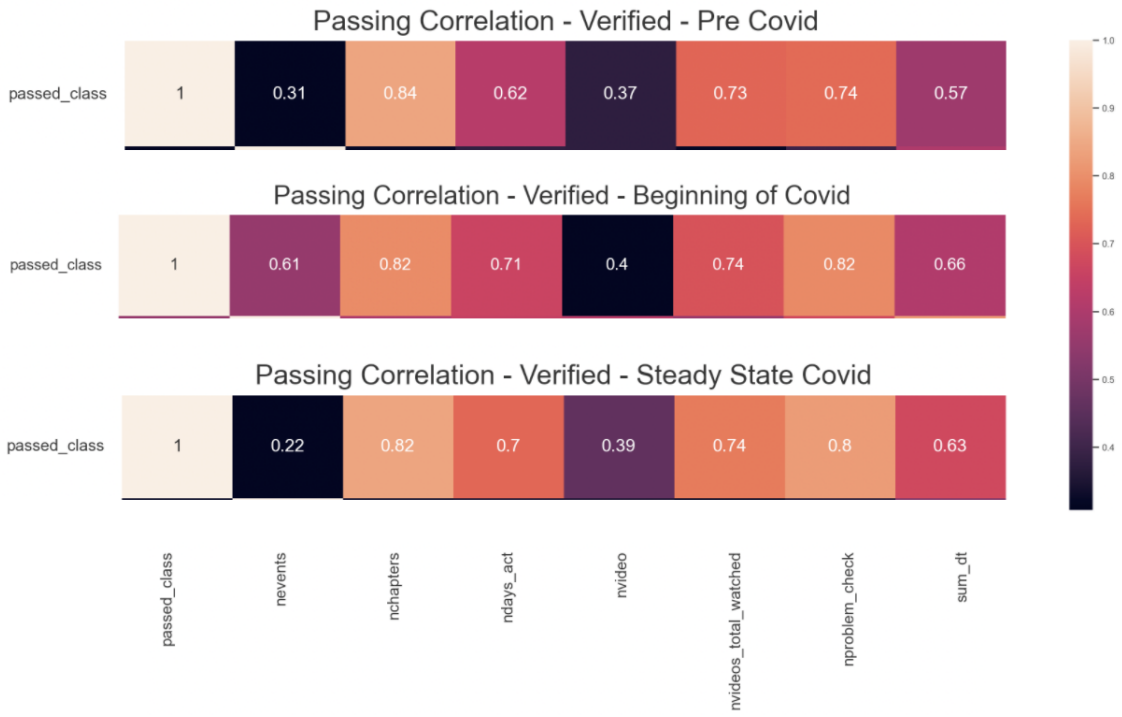

Figure 2-23: Correlation matrix between the bottom row of features and leftmost *passed_class* feature, displaying the change in these values by each pandemic timing group. Same correlation and coloring rules apply as above.

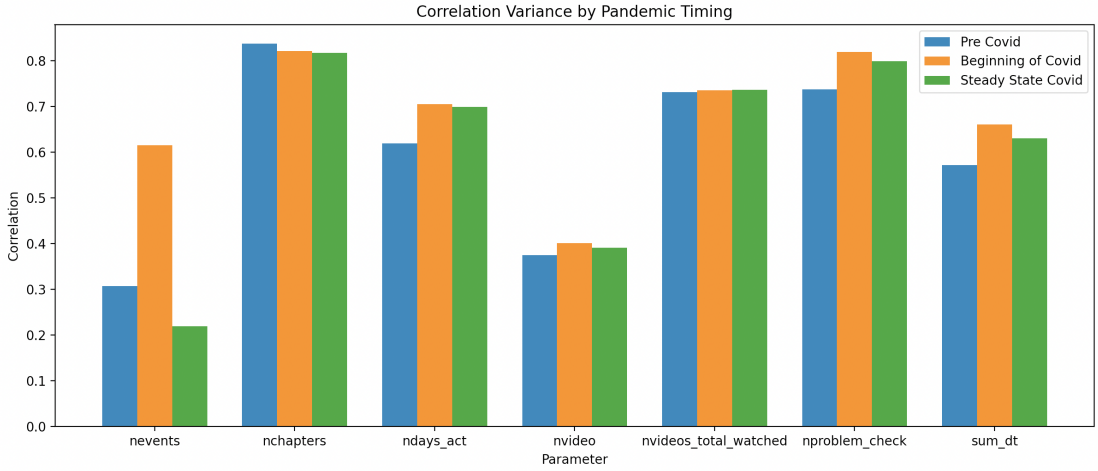

Figure 2-24: Change in correlation values with respect to *passed_class* feature, by semester, verified students. Blue represents pre-covid, orange beginning of covid, and green steady state.

Now that we have a baseline established, let's take a look at how these correlations shift over the course of the pandemic. For the most part, the correlations remained

unchanged, with a few exceptions. The most noticeable change was the *nevents* parameter which went from having a small to medium correlation of 0.31 during the pre-covid period, to a strong correlation of .61 (100% increase) in the beginning of covid period, then back down to a small correlation of 0.22 in the steady state covid (177% decrease).

The parameters *ndays_ active*, *nvideo*, *nproblem_ check*, and *sum_ dt* all follow a similar pattern as *nevents*, just with smaller magnitudes. All of them increase from before to beginning, then decrease to steady state. This pattern could be attributed to the pandemic itself, as more free time for learning at the start of the pandemic naturally produced a higher interaction rate with the class, and consequently a higher passing rate as well. This also follows the general trend of the student passing rates as the pandemic progressed.

2.5 Machine Learning

Now that we have developed a high level understanding of the class, we use more complex methods, namely machine learning, to continue our search for the impacts of the pandemic. For this approach, using the features highly correlated with passing the class,outlined above, we train an ML model on the pre-covid data set, using the pass/fail metric to determine accuracy. Then, using this model, we predict the following semesters' pass/fail rates to determine if students impacted by the pandemic differ from those prior in any meaningful ways.

2.5.1 Feature Visualization

We now have established *nevents*, *ndays_ active*, *nvideo*, *nproblem_ check*, and *sum_ dt* as a strong correlative feature set with passing the class. Given these features, we utilized various techniques to visualize them in efforts to illustrate patterns.

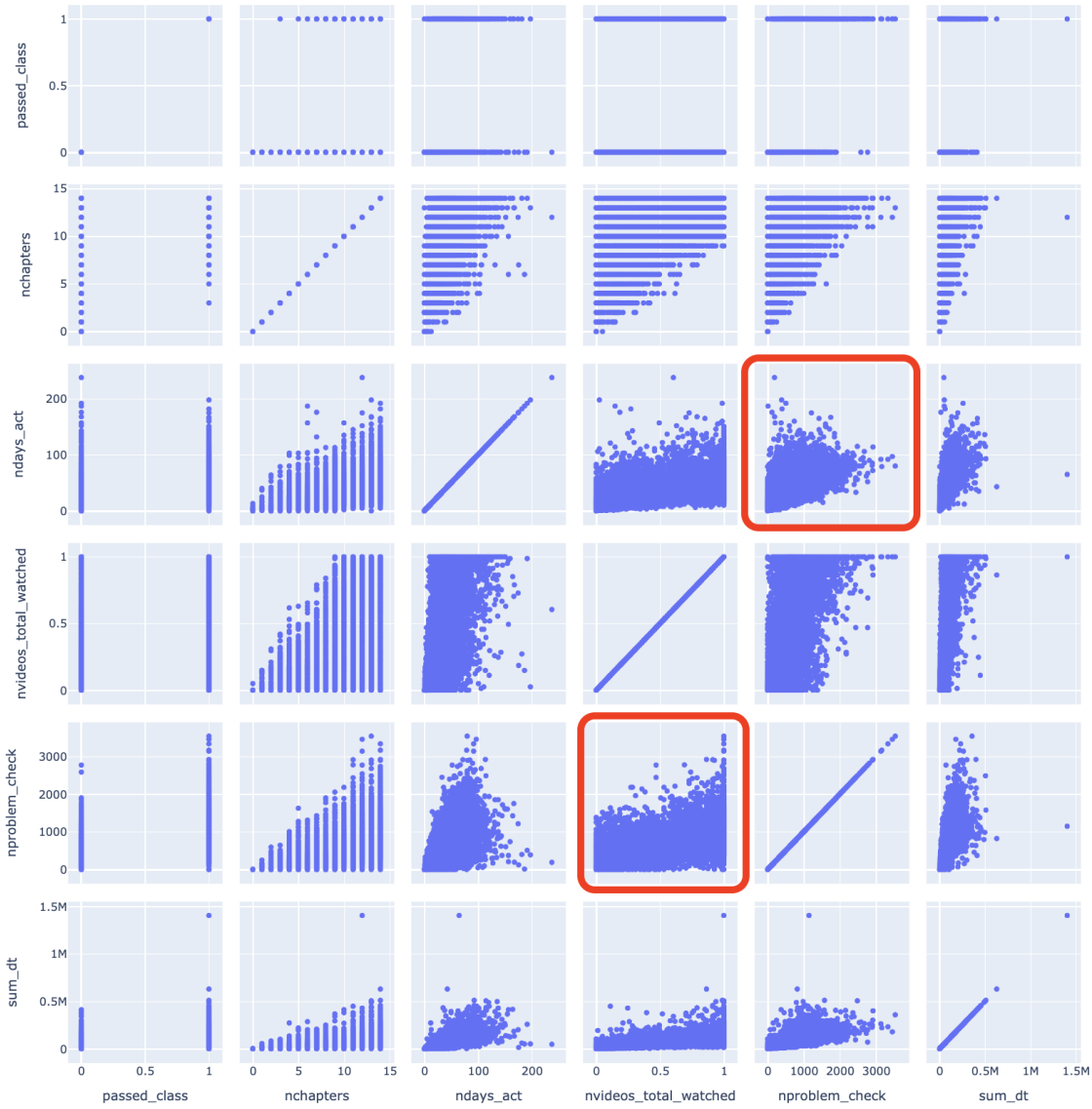

Figure 2-25: Scatter matrices of all pairs of selected features. X axis is denoted on the lower row and Y axis denoted by vertical row on the left. Intersection of features denotes data points of each plotted against one another. Red selections denotes the feature pairs we will use for visualization.

The above chart provides us with pairwise scatter matrices of all the selected features. This gives us an illustration of how each pair of features affects one another, and helps us to select a few to take a deeper look at. For our purposes, we want a few features that we can select to display the interaction of features accordingly. We need to first determine which features to display on the axes, then secondly in what relation to graph them in order to produce a clusterable sample space.

The *nchapters* feature is discrete on the range [0, 14] which produces charts with linear lines of data, which is not good for what we need. Removing that as an option, we see the selected plots in red have just what we're looking for. All of the axes *nproblem_check*, *ndays_act*, and *nvideos_total_watched* are continuous, and graphing them as they are within the scatter plot provides graphs that look appealing to cluster.

Figure 2-26: Top four features selected from figure 22, from all semesters. Number of problems checked on the x axis, with the number of days active on the y axis. The size of the data point represents the number of chapters visited. Red denotes students that passed and blue denotes those that failed. Red = 13,380. Blue = 12,250.

We see in the above chart that there exists a clear trend and grouping within the student data. As the number of problem checks and number of days active increase, the number of passing students does as well. The size of the data point also represents the number of chapters completed. It's a little tough to see because of how dense the plot is, but as the points progress in both positive directions, the size of them

also increases, indicating that the number of chapters completed follows the similar pattern as mentioned above. The number of passing students (red) is 13,380 while the number of failing students (blue) is 12,250.

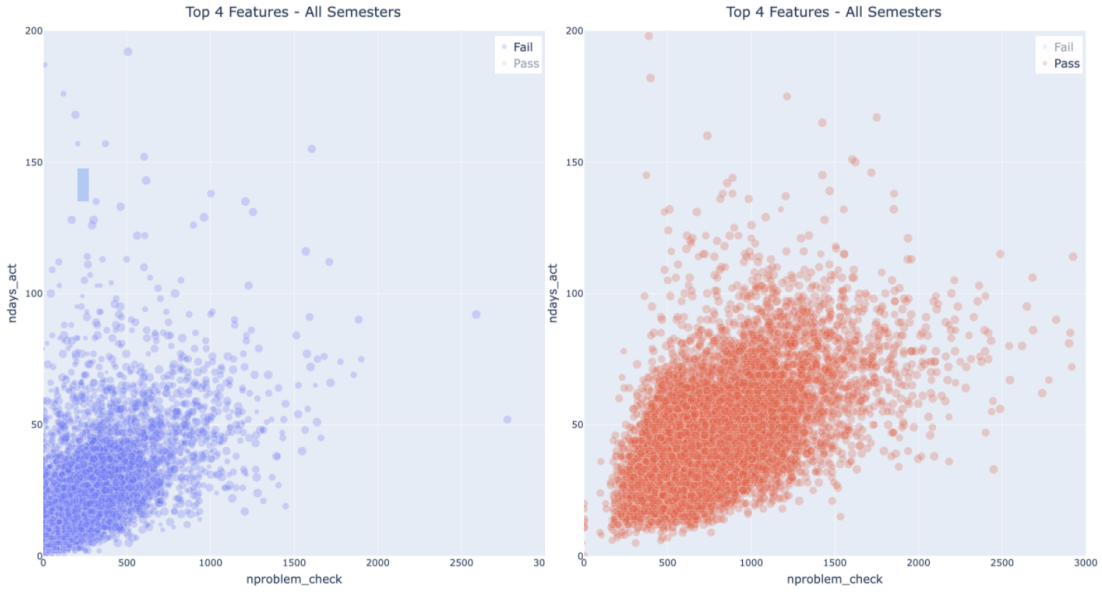

Figure 2-27: Same data as above figure, with passing and failing students separated. Red = 13,380. Blue = 12,250.

Here we have the same data visualized, split to show just the fail data points on the left and just the pass points on the right. We can see that a solid overlap occurs within the data, roughly between 250 and 750 on the x axis and 15 and 35 on the y axis. Students in this range could be considered within the average, as this is where approximately half of the overall set pass and half of them fail. This view also gives us an idea of where the centroids of both datasets are located. It's clearly identifiable that the fail group has a centroid with lower values for both of the axes, approximately at (250, 35). On the other hand, the pass group has a centroid with relatively larger values, at approximately (1000, 50).

As with most large datasets, we do see some interesting outliers on both sides of the pass, fail divide. On the fail side, we see a number of data points that are larger in (x, y) value than the centroid of the pass data. These students had problem-check and days-active values that were relatively high, yet still did not pass the class.

Qualitatively, these are students who were engaged with and interacted with the class frequently, watching videos and completing practice problems, yet did not complete enough assessments to earn a passing grade. We also see students who passed the class with smaller (x, y) values than the centroid of the fail subset. These students had (x, y) values that were relatively low, yet still completed enough to pass. These are probably students that had some CS knowledge coming into the course and were able to do the graded portion without learning much of the material again.

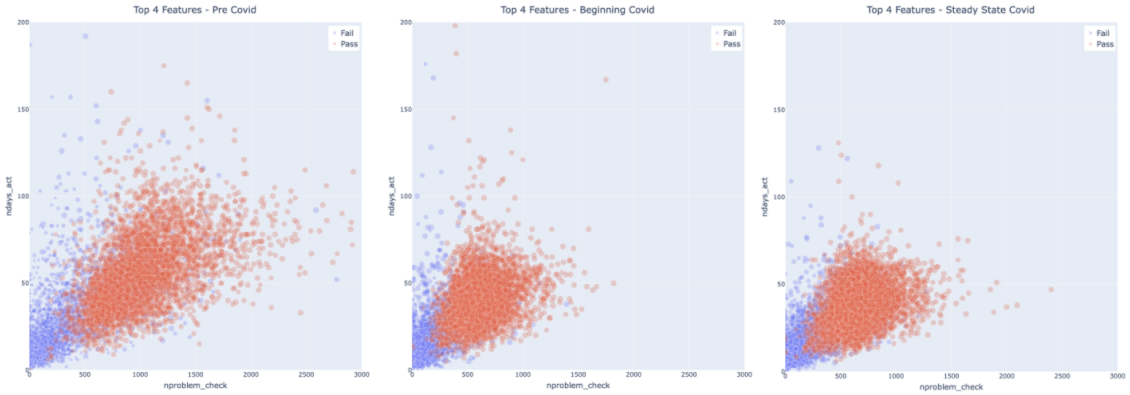

Figure 2-28: Top four features selected from figure 2-25, by pandemic timing. Number of problems checked on the x axis, with the number of days active on the y axis. The size of the data point represents the number of chapters visited. Red denotes students that passed and blue denotes those that failed. From left to right: Blue 3,811 Red 3,929; Blue 4,321 Red 4,623; Blue 3,118 Red 4,828.

Here we aggregate by pandemic timing to see how these groups shifted over the course of the pandemic. We can immediately see a leftward and downward shift as well as a densification of the data within the beginning and steady state groups when compared to the pre covid group. This indicates an overall decrease in the problem-check and days-active metrics; students were completing less problems and interacting with the course less. It's also noticeable that the data becomes more dense and the number of outliers decrease from pre covid to beginning.

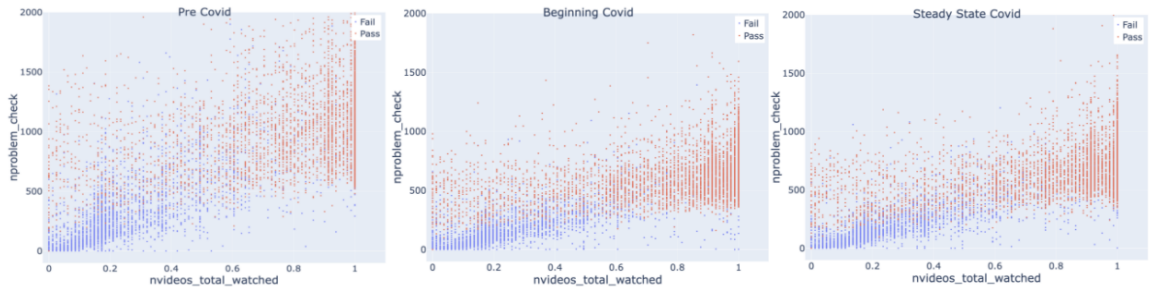

Figure 2-29: Video and Problem features selected from figure 2-25, by pandemic timing. Number of problems checked on the y axis, with the percentage of videos watched on the x axis. Red denotes students that passed and blue denotes those that failed. From left to right: Blue 3,811 Red 3,929; Blue 4,321 Red 4,623; Blue 3,118 Red 4,828.

Here we see the trend between number of total videos watched and number of problems checked, aggregated by pandemic timing. We see the same pattern as in figure 2-28, with the densification of the data between the pre-covid and during covid semesters. Its also pretty clear that as you progress in each positive direction, the color shifts from blue to red which indicates that more students are passing. In this set, we also observe that the estimated centroids of the pass/fail subsets are further spread apart than in figure 2-28.

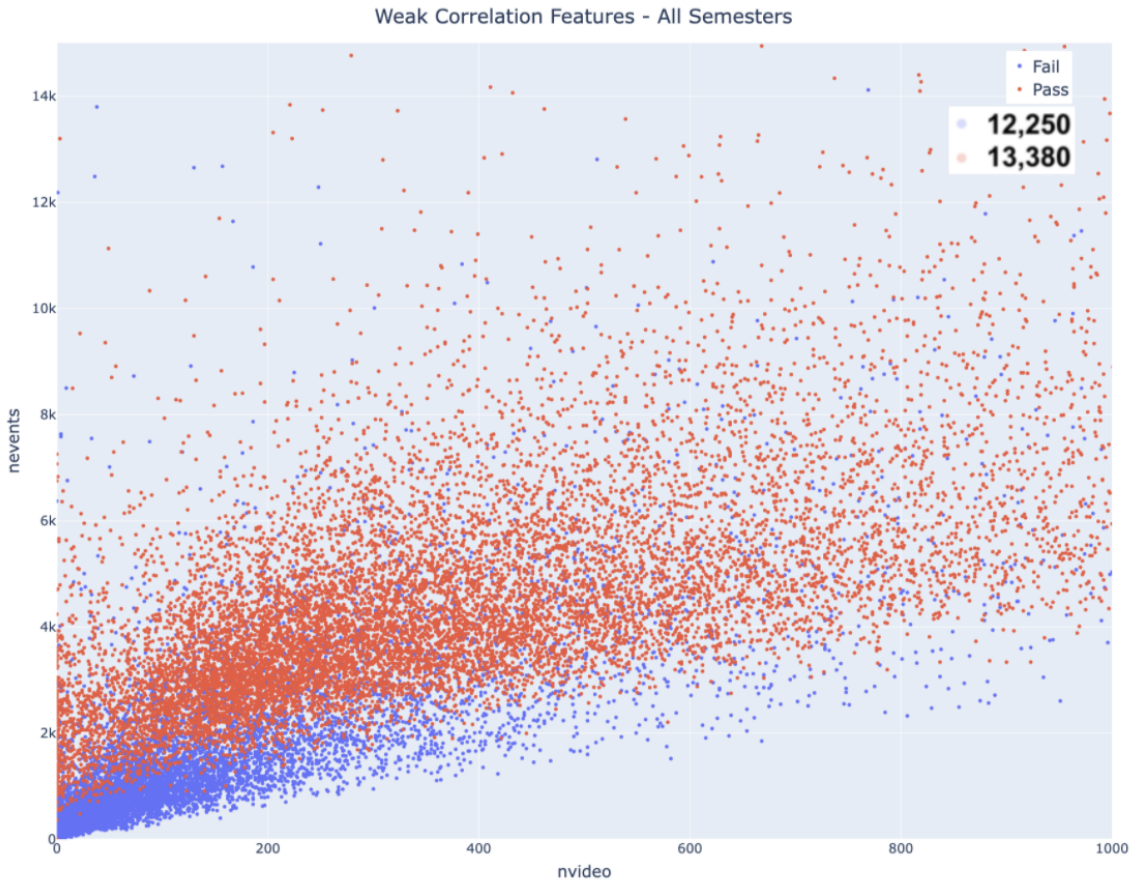

Figure 2-30: Weak feature clustering from all semesters. Number of video events on the x axis and number of total log events on the y axis. Notice there is no noticeable trends. Red denotes students that passed and blue denotes those that failed. Red = 13,380. Blue = 12,250.

Just for fun, the above graph shows two features, number of video events and number of total events that had relatively low correlation values from figure 2-22. Visually its easy to see the data is much more dissipated. The red and blue data points spread across the entire range of x and y values, indicating a low correlation to passing or failing. This is in stark contrast to figure 2-28, which has distinctive clusters and associated centroids.

Effectively, these clustering plots show that students lessened their involvement and interaction with the course after the pandemic began, yet maintained the ability to pass the course as we know from previous data that shows that the passing rate marginally increased. This can be attributed to students becoming more efficient in

engaging with the course. Due to the pandemic and its effects, students had more time to spend on the course, with less outside-course interruptions like commuting and social activities. As a result, students spent effectively less time than they would have otherwise to pass the course.

2.6 Modeling

Now that we have visualized some of our features and how they interact with others, we aim to use machine learning to try and understand the impact of the pandemic on student performance. At a high level, we train our selected model on the pre-covid subset of data, using the five strongest metrics as features, and the pass/fail metric as labels. This should cause the model to learn the patterns of the first group. Then, using this pretrained model, we predict if a given student will pass or fail in the beginning of covid and steady state covid subsets. Using the associated accuracy's and other performance metrics, we will draw conclusions on how the pandemic did or did not affect the model, therefore also affecting the students and metrics.

2.6.1 Model Selection and Preprocessing

We first preprocessed all of the data, using sklearn's MinMaxScaler which scales all of the data between zero and one. This was done in order to normalize the data that directly affects model performance. We then split the pre-covid data into training and testing subsets so that our model does not over or underfit. Using the training and testing data, we tried five different classification models: K Neighbors, Decision Trees, Random Forest, Ada Boost, and Gradient Boost. The highest performing model after tuning each models' parameters was the Gradient Boosting Classifier, which performed at 94% accuracy. The parameters are listed below.

Model	Gradient Boosting Classifier
n_ estimators	76
learning_ rate	0.55
max_ depth	76
accuracy	94.1%

Table 2.4: Parameter assignment of selected gradient boosting classifier.

2.6.2 Model Results

Parameter	Pre-Covid	Beginning	Steady State
Accuracy	94.1%	94.0%	93.4%
Precision	92.2%	91.9%	91.0%
Recall	96.6%	96.9%	97.4%
F1 Score	94.4%	94.4%	94.1%

Table 2.5: Gradient boosting classification model performance metrics for each pandemic timing group. Model was trained on pre-covid group only.

We see that over all three semesters, the model did an excellent job predicting which students pass and which fail. All of the performance metrics remained basically the same across all three of the groupings. We do not see any discrepancies that we were trying to find as a result of the pandemic.

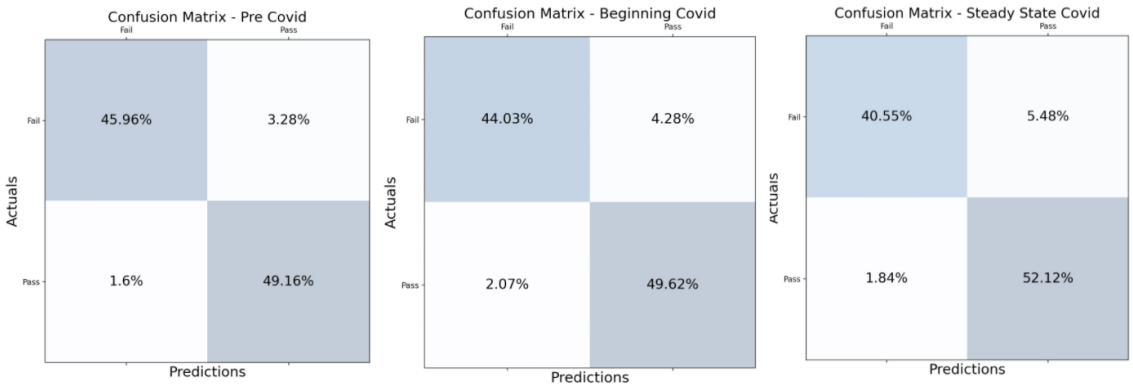

Figure 2-31: Confusion matrix created by gradient boosting classifier, by each pandemic timing group. Top left number in each is true positive, bottom right is true negative, top right is false positive, and bottom left is false negative.

Looking at the confusion matrices gives us another illustration of how the model

performed over the three subsets. We see similar metrics for each of the four performance values over each of the semesters.

One possible explanation is that any discrepancies caused by the pandemic were not of enough magnitude to disrupt the gradient boosting model. If the changes in the data were too low, the model would not be affected and it would continue predicting outputs at a high rate. It's also possible that even if considerable changes in the data were present, the model was able to learn well enough, and have enough classifiers to overcome the changes.

Chapter 3

Conclusion

All in all, this thesis provides us with a thorough data discovery and landscape analysis of this introductory computer science MOOC over the course of the pandemic. Using a plentiful amount of graphs and illustrations along the way, we looked at metrics as simple as age and as complex as feature correlations to address patterns caused by COVID over the previous eight semesters. In doing so, we found that COVID did seem to have an influence in some areas of the course, while others maybe not so much.

We first looked at demographic data like age, education levels, and gender to try and understand the composition of the students as a whole. We saw that the majority of students taking this class were of the ages 17 to 22, male, and from the United States. As the pandemic progressed, we saw a 100% increase in the number of college and late-highschool aged and educated students enrolled in the course, indicating an increased curiosity in the class and computer science field as a whole, caused by COVID.

Then we analyzed the enrollment data and found an immediate spike in enrollment in the first semester that began after the pandemic was declared, solidifying our conjecture that the pandemic increased student enrollment in this course. We also discovered that enrollment outside of the United States increased during the pandemic, showing promising signs that this class is expanding its reach across the world.

Diving into the class itself, we found mixed results when it came to student interaction and engagement in the course. Some metrics like number of videos viewed and days active in the course tended to increase as a result of the pandemic, while others like total time spent on the course and problems checked were lower than pre-

pandemic levels. However, when it came to performance, the percentage of students that passed the class did steadily increase throughout the pandemic, which is also promising.

Focusing on student performance, we determined a high correlation between five or so factors and passing, which could help future runs of the course increase student involvement and performance. Due to the pandemic, we also noticed that videos watched, total time spent, and days active in the course increased their correlation to passing.

We concluded our analysis by using machine learning in efforts to find further impacts and disruptions caused by the pandemic. Training on the pre-covid data set only, our model was able to accurately predict which students passed with the same accuracy for all three data subgroups: pre-pandemic, beginning of pandemic, and steady state pandemic. This probably means that the pandemic did not disrupt the overall data enough for the model to have trouble with classification.

As a whole, this thesis provides evidence that the pandemic did have a measurable impact on this introductory computer science MOOC. Fortunately, the data shows that the course benefited overall from the pandemic, pulling students from across the world who spent more meaningful time on the course which resulted in higher passing rates as the pandemic progressed.

Chapter 4

Resources

The majority of the work was done on a Macbook Pro, using Google Sheets and Overleaf for composition, Microsoft Excel for data viewing, and Terminal and Sublime for code editing and execution. All of the data was transferred and stored on an MIT dropbox instance and downloaded locally for easy and quick access.

The following python packages were used: pandas, matplotlib, datetime, seaborn, numpy, and sklearn. Within sklearn, the DecisionTreeClassifier, RandomForestClassifier, AdaBoostClassifier, GradientBoostingClassifier, and KNeighborsClassifier were explored.

Bibliography

[1] Pearson's correlation using stata. *Laers Statistics*, 2018. https://statistics.laerd.com/stata-tutorials/pearsons-correlation-using-stata.php.

[2] Third world countries. *World Population Review*, 2021. https://worldpopulationreview.com/country-rankings/third-world-countries.

[3] Cdc museum covid-19 timeline. *Centers for Disease Control and Prevention*, Jan 2022. https://www.cdc.gov/museum/timeline/covid19.html.

[4] Statistics. *National Girls Collaborative*, 2022. https://ngcproject.org/statistics.

[5] Ayesha R. Bajwa. Analyzing student learning trajectories in an introductory programming mooc. Master's thesis, Massachusetts Institute of Technology, 2019.

[6] Virginia Katherine Blackwell and Mary Ellen Wiltrout. Learning during covid-19. *EMOOCs 2021*, 2021:219 – 236, 2021.

[7] Vonder Haar and Christine M. Understanding learner engagement and the effect of course structure in massive open online courses. Master's thesis, Massachusetts Institute of Technology, 2020.

[8] Chris Impey and Martin Formanek. Moocs and 100 days of covid: Enrollment surges in massive open online astronomy classes during the coronavirus pandemic. *Social Sciences Humanities Open*, 4(1):100177, 2021.

[9] Steve Liesman. Half of u.s. elementary and high school students will study virtually only this fall, study shows. *CNBC*, Aug 2020. https://www.cnbc.com/2020/08/11/half-of-us-elementary-and-high-school-students-will-study-virtually-only-this-fall-study-shows.html.