2022

How COVID-19 Impacted Tertiary Education: Predicting the Educational Infrastructure of the Future

Alexander Muehleisen

Claremont McKenna College

How COVID-19 Impacted Tertiary Education: Predicting the Educational Infrastructure of the Future

submitted to
Professor Mark Huber

by
Alexander Muehleisen

for
Senior Thesis in Data Science
April 25, 2022

Abstract

In the following paper, data pertaining to educational shifts during the COVID-19 pandemic is gathered and analyzed. The pandemic shook the foundations of working life for many and the subsequent shift towards hybrid and work-from-home environments set the scene for perennial changes to the work-life balance. The impact of the COVID-19 pandemic on students is measured in this paper through literature review, data analysis, and regression analysis on student assessment scores. The pandemic shed light on the fact that education systems in the United States and around the world have not changed in hundreds of years. The rise of Zoom and other online-based educational resources has given way to new education models that offer hope for an improved educational infrastructure of the future. One of these models, mastery learning, is explored throughout this paper. The paper first identifies some of the existing problems with the implementation of mastery learning, before attempting to dispel these issues by analyzing current education models and the impact of COVID-19 on learning. Eventually, a educational framework for tertiary institutions of the future is suggested, and details of the framework are supported by findings from analysis conducted in this paper. These findings include incorporating the importance of confidence and maintaining a balanced work-life schedule into future developments into the education industry.

Acknowledgements

This thesis would not have been possible without the continued support of my thesis reader, Professor Mark Huber, who guided and mentored me throughout this process. Professor Huber, thank you so much for all of your help, patience, and feedback.

Contents

1 Introduction

1.1 Problem Description

The COVID-19 pandemic shook the foundations of working life for many and the subsequent shift towards hybrid and work-from-home environments set the foundation for perennial changes to the work-life balance. Students, in particular, were acutely affected. As schools shut, students were forced to quickly adjust to an online learning environment taught by teachers and professors who were just as new to online-learning technology as the students. Beyond the number of issues percolating around this change, the COVID-19 pandemic shed light on the fact that education systems in the United States and around the world have not changed in hundreds of years. Once initial problems were resolved, the simplicity and convenience of software like Zoom, Microsoft Teams, and other online-based educational resources began to offer hope for improvements to be made to existing models of education.

1.1.1 Contemporary Educational Frameworks: The Prussian Model

Educational frameworks at tertiary institutions have followed the same architecture for hundreds of years (Khan, 2012). The current model, eponymously named the Prussian Model, finds its origins in 18th century Prussia and is now ubiquitous across schooling systems in the United States and the rest of the world. The Prussians introduced many of the norms and orthodoxies that set the foundation of the education system today, including 'primary school' and 'secondary school', a system of 'grades', the sorting of disciplines into 'subjects', the division of the day into 'periods', and the termination of a student's academic career after they complete a certain number of 'grades' (Khan, 2012). Moreover, while most countries' citizens are proponents of a tax-supported and compulsory education system, many of them may not have guessed that these innovations originate from Prussia (Khan, 2012). Of course, this is by no means the only influence on contemporary school systems, but the legacy of the Prussian Model is important here due to its historical contextualization. Industrialization and the Napoleonic Wars coincided with, or perhaps led to, Prussian desire for an education system wherein "loyal and tractable" citizens could be churned out to play a role in war or industrial progression (Khan 2012, p.47). Although there were other considerations, the cornerstone of the system was manufacturing loyal citizens who would not question the chain of command, current leadership, or authority.

1.1.2 Moving Past the Prussian Model

Irrespective of ill-intention or ulterior motive behind the Prussian Model, its development has unequivocally worked quite successfully for the past two centuries. However, two centuries is a long span of time and myriad changes have taken place in the world since the Model was established. Education has failed to keep up with this ever-changing industry climate, which calls for more independent thinking and less rote memorization than ever before. Furthermore, the industry has failed to keep up with the cultural shifts that pervade today's society. Perhaps most important, burgeoning technological advancements have rendered educational systems' lack of transformation and innovation painfully obvious. New generations of students need to be versed in things like digital literacy, research methods, and basic computer science operations in addition to or instead of the traditional school curriculum. As technology continues to grow, it makes sense to teach its applications and build it into the curriculum as much as possible. COVID-19 became the impetus to make this happen. Online-learning companies like Zoom, Coursera, and Khan Academy have become integrated into all facets of school, work, and everyday life. As of now, the future of working life seems to be some combination of in-person and online and it is time for the education systems of the world to follow suit and evolve accordingly.

1.1.3 Personal Connection and Khan Academy

The motivation this paper began with a personal interest in Khan Academy – a non-profit educational organization founded by Sal Khan in 2008. The idea for the non-profit grew organically for Khan, who started making YouTube videos to tutor his cousin in math, before quitting his job as a financial analyst and founding Khan Academy. Taking on the ambitious mission of providing a 'free, world-class education to

anyone, anywhere,' (Khan, 2012) Khan Academy now provides access to 10,000 videos, 3000 articles, and 50,000 exercises – all unique and readily accessible from anywhere in the world. The philosophy behind the education model is Sal Khan's belief in 'teaching for mastery'; meaning to teach until a student masters a concept, instead of teaching within a set term and ignoring student mistakes along the way. The accumulation of these mistakes can lead to a drop of confidence for some, and, worse, a complete disinterest in learning for others. Mastery learning offers a solution to this that can now be achieved through the growing branches of technology.

1.1.4 Mastery Learning

As a concept, mastery learning has been around for decades (Teacher Experiment Academy, 2019). However, attempts at its implementation have been a different story (Roseman University of Health Sciences, 2022). Technology changes this problem. With the advent and success of Khan Academy and other Massive Open Online Courses (MOOCs) like edX and Coursera, mastery learning now seems more realistic and achievable than ever before. Moreover, the shift towards hybrid work-and-learning environments created by the pandemic presents another opportunity to push towards mastery learning. Although technology has become more and more of a part of everyday schooling and working lives, the pandemic massively accelerated this shift (Ercolano, 2020).

The key concept behind mastery learning is giving the student as much time as needed to completely 'master' a concept, and, eventually, an entire course. 'Mastering,' in this sense, can be treated as an antonym to achieving anything lower than 100 percent on a concept test. Within a traditional style of teaching, a student might score 70 out of 100 on a given concept test. Although the student has technically passed, a score of 70 essentially represents some sort of misunderstanding of 30 percent of the concepts tested in the test. Due to the time constraints of semesters and terms, the class will be forced to move on despite this student not understanding close to one third of the tested material. Moreover, there are likely several students who scored similar or even lower scores, meaning that a good proportion of the class does not have a solid grasp of course material. Most classes build up content linearly, especially subjects like mathematics, so these students will be at a huge disadvantage if the next focus of the class is on a concept that progresses directly from the last of the last one. In the end, they might feel discouraged enough to study or try on the next test. Even if students remain spirited, they simply may not have the aptitude to perform well on subsequent tests after missing so much fundamental content early on.

On the other hand, students within a framework of mastery learning would not move on to new content until they have mastered the previous content. Time is less important and students are encouraged to traverse the subject material at their own pace. This method removes the 'holes' that are created when students have to learn course material that builds off a previous concept.

1.1.5 The University of the Future

Ultimately, this paper hopes to present some prediction of what a tertiary institution of the future might look like. Current education models will be analyzed to reveal which parts of contemporary education work well and which parts do not work well. Further, the impact of COVID-19 on education will be analyzed to discover what can be learned from all the technological changes involved in shifting towards an online learning environment. Finally, the performance of online learning platforms and MOOCs will be analyzed to suggest how technology can be appropriately integrated into the educational infrastructure of the future.

1.2 Data Gathering Process

The datasets used in analyses were taken from various sources. The search for the data predominantly consisted of parsing through government-affiliated websites, as well as larger organizations and nonprofits that track data within industries like education. UNESCO and UNICEF are good examples of this. To aid the search, tools, and references such as Google Scholar, Google Dataset Search, and government indexes were used. Additionally, websites dedicated to statistics and the collection of data, like Kaggle, were used and referenced. Similar studies into educational frameworks and the education industry as a whole were also looked at and taken into consideration. Only .csv files were used as data in this paper, so in most cases, the

datasets were simply imported into RStudio using the *readr* package's `read_csv` function. Otherwise, the data was loaded into Python using `read_csv` from Pandas. A complete and descriptive list of the tables used can be found below:

1. High School Longitudinal Study (HSLS) [2009]
    - 23,503 observations, 9,614 columns
    - Data collected by the National Center for Education Statistics (NCES)
    - A nationally representative longitudinal study of over 23,000 9th graders from 944 schools in 2009.
    - Data collected through surveys of students, parents, math and science teachers, school administrators, and school counselors
2. UNESCO Global Dataset on the Duration of School Closures (2020-2021)
    - 210 observations, 11 columns
    - Data collected by UNESCO Institute for Statistics (UIS)
    - Full and Partial School Closures at a nationwide level due to COVID-19
    - Data collected through surveys on National Education Responses to COVID-19
3. California Assessment of Student Performance and Progress (CAASPP) [2016-2019, 2021]
    - 4,294,551 observations, 39 columns
    - Data collected by the English Language Proficiency Assessments for California (ELPAC)
    - A standardized assessment for students in California in grades 3, 4, 5, 6, 7, 8, and 11. Consists of English, Mathematics, Science, and other assessment tests.
4. Distance Learning Data from the Digest of Education Statistics (2012-2020)
    - 18 observations, 3 columns
    - Data collected by the National Center for Education Statistics (NCES)
    - The number and percentage of students enrolled in degree-granting postsecondary institutions, by distance education participation
5. Open University Learning Analytics Dataset (OULAD) [2017]
    - 207,242 observations, 9 columns
    - Data collected by Open University Analyze
    - Data about courses, students, and their interactions with a Virtual Learning Environment (VLE) for seven selected courses (called modules).
    - The dataset consists of 7 separate tables connected using unique identifiers.

The High School Longitudinal Study is a nationally representative, longitudinal study of over 23,000 9th graders from 944 schools in 2009. The first follow-up took place in 2012, and the second took place in 2016. The study includes assessments in algebraic skills, reasoning, and problem-solving, as well as surveys of students, their parents, their teachers, and school administrators. It also includes background, household, and demographic data. It is the only ongoing study of school-based longitudinal studies. For the purpose of this analysis, only base-year data from 2009 will be used.

Established in 2014, the California Assessment of Student Performance and Progress (CAASPP) replaced the previous assessment system and implemented a series of programs to promote high-quality teaching in the state. The CAASPP system operates through three assessment branches: Science, English Language Arts/Literacy (ELA) and Mathematics, and Spanish Reading/Language Arts. Each branch encompasses various assessments and student participation requirements, including the California Science Test, Smarter Balanced Assessments, and the California Spanish test, respectively. California Alternative Assessments, offered to students with significant cognitive disabilities, are also incorporated into the first two branches. The system aims to assist teachers, administrators, students, and parents through these assessment tests. Additionally, through different assessment approaches and the collection of data from these assessments, the CAASPP hopes to promote high-quality teaching and learning.

The UNESCO Global Dataset on the Duration of School Closures was collected as a part of the UNESCO Institute for Statistics' response to the COVID-19 pandemic. As a part of this response, UNESCO, the World Bank Group (WBG), and UNICEF have launched the Framework for Reopening Schools, a flexible tool for policymakers and planners that highlights which factors will ensure a successful transition back to in-person

learning. These partners have also created various resources, surveys, and resources available to the general public. The hope of the project is to create a space for discussion and the exchange of personal experiences. Since the outbreak of the pandemic, UNESCO has monitored the daily statuses of school closures and the online/hybrid alternatives explored by schooling systems around the world. The data reflects the status of school closures across the globe by region and country.

The Digest of Education Statistics covers the field of American education from pre-kindergarten to graduate school and provides a comprehensive collection of statistical information on the field. A set of tables contains this broad set of data, which was collected from numerous sources and includes results from government and private surveys, as well as activities carried out by the National Center for Education Statistics. To qualify for inclusion in the Digest, material must be broad in scope and nationwide in coverage. The data contains information on a variety of subjects within the field of education, including the number of schools and colleges, teachers, enrollments, and graduates. Additionally, the study looks at educational attainment, finances, federal funds for education, libraries, and international education. The dataset on Distance Learning was extracted from the Digest of Education Statistics, and tracks the rising popularity of distance learning over the last decade.

The information gathered in the Open University Learning Analytics Dataset (OULAD) examines student interactions with a Virtual Learning Environment (VLE). Data within the Analytics project covers student, course, and interaction information collected on seven modules (courses). Specifically, variables in the dataset include student scores on assessment exams, weightings of assessment exams, module lengths, module presentation styles, and the number of interactions a student had with a module. Observations were taken of residents in the United Kingdom. The dataset is anonymized in order to adhere to privacy concerns.

2 Analysis

2.1 Structure of Analysis

Analysis will adhere to the following structure:

1. Multiple Linear Regression Analysis of Confidence on Student Performance
2. Multiple Linear Regression Analysis of External Factors on Student Performance
3. Predicting Student Performance through Classification
4. Trends in the impact of COVID-19 on Education
5. Random Forest Model on the Importance of factors in Online-Learning

These specific interest clusters and analytic techniques were chosen based on inspection of the data and research into good predictors of, and detriments to, student performance. This is certainly not the only lens to analyze student performance through. All of the datasets contain various other observations of assessments in interest clusters that were not analyzed in this paper. For future research into similar variables or other variables from the study, please reference the websites and homepages of the NCES, the UIS, the ELPAC, and Open University. In this paper, the data will be analyzed in RStudio.

2.2 R Libraries

The following R packages were used in analysis:

```
library(tidyverse)
library(tidyr)
library(ggplot2)
library(dplyr)
library(knitr)
library(readr)
```

```
library(reticulate)
library(magrittr)
library(randomForest)
library(datasets)
library(caret)
```

The *reticulate* package serves as a Python interpreter in R. However, Python must already be installed on the respective OS to use the package. *Mini-conda*, a small, bootstrap version of Python can alternatively be installed through *reticulate*.

2.3 Python Packages

Some analysis will be performed in Python, which is possible in R through the *reticulate* package. Therefore, the following libraries and modules will also need to be installed to run the appropriate regressions and models. They are installed using the `py_install` function from the *reticulate* package. As they are already installed on this OS, they are commented out, but still included in this paper for reference.

```
#py_install("pandas")
#py_install("scikit-learn")
#py_install("spicy")
#py_install("seaborn")
#py_install("matplotlib")
```

The libraries mentioned above can be imported into R, as follows:

```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import warnings
```

```
from sklearn.model_selection import train_test_split
from sklearn.model_selection import GridSearchCV
from sklearn.linear_model import LinearRegression
from sklearn.linear_model import LogisticRegression
from sklearn.linear_model import Ridge
from sklearn.preprocessing import StandardScaler
from sklearn.preprocessing import PolynomialFeatures
from sklearn.preprocessing import KBinsDiscretizer
from sklearn.pipeline import make_pipeline
from sklearn.svm import SVC
from sklearn.metrics import confusion_matrix
```

2.4 Contemporary Educational Infrastructures

The rudiments of future development in the education industry will be fabricated from existing models of education that exist in the world today. Inspecting and analyzing these models and infrastructures is crucial. The first analysis will cover the High School educational framework in the United States, specifically looking at data collected from surveys and assessments of 9th graders around the country. This data was collected in the High School Longitudinal Study in 2009, which was carried out by the NCES. Regression analyses will be conducted to estimate the effect of one or more explanatory/feature variables on a chosen dependent/target variable. The analysis of current models of education should reveal the kind of actions, reflections, and mindsets that lead toward higher levels of learning and better assessment results.

2.5 The Role of Confidence in Student Performance

2.5.1 Read-in HSLS Data

The data can be imported into R using the *reticulate* package and Pandas' *pandas.read_csv* function in Python.

```
hsls_data = pd.read_csv('hsls_17_student_pets_sr_v1_0.csv')
hsls_data
```

```
##         STU_ID  SCH_ID  ...  W5W1W2W3W4PSRECORDS199  W5W1W2W3W4PSRECORDS200
## 0        10001      -5  ...                     0.0             2053.407870
## 1        10002      -5  ...                     0.0                0.000000
## 2        10003      -5  ...                     0.0                0.000000
## 3        10004      -5  ...                     0.0                0.000000
## 4        10005      -5  ...                     0.0                0.000000
## ...        ...     ...  ...                     ...                     ...
## 23498    35202      -5  ...                     0.0                0.000000
## 23499    35203      -5  ...                     0.0                0.000000
## 23500    35204      -5  ...                     0.0                0.000000
## 23501    35205      -5  ...                     0.0              862.728189
## 23502    35206      -5  ...                     0.0                0.000000
##
## [23503 rows x 9614 columns]
```

2.5.2 Descriptive Statistics

To measure what aspects of the current education model work well, an indicator of strong educational success in our dataset is the first thing that needs to be found. This will represent the target column. Following, a feature matrix needs to be established, consisting of either one or more predictors from the dataset. After analyzing the data, there are a few assessment measurements that might be helpful to consider as potential target variables. The Mathematics Theta Score (**X1TXMTH**) is particularly promising as an indicator of student performance. Referring to the HSLS Online Codebook (HSLS, 2009), the score provides an estimate of mathematical achievement in the population of 9th graders. Moreover, mathematics scores usually serve as good indicators of student performance due to their universality and therefore potential for cross-analysis.

For this regression, the Mathematics Theta Score (**X1TXMTH**) will be used as the target column. To properly predict the target column, the following columns have been identified as potential predictors: Scale of student's mathematics identity (**X1MTHID**); Scale of student's mathematics self-efficacy (**X1MTHEFF**); Scale of student's interest in Fall 2009 math course (**X1MTHINT**). The columns are relatively similar and have been intentionally chosen as such to directly measure the effect of confidence in student studies. All three columns are continuous, and descriptive statistics are explored below. Descriptive statistics for the target variable are also included.

A summary of the data can be drawn with the `describe` function from Pandas. The function lists the number of observations, the mean, and the standard deviation for each column, which provides a better understanding of the distribution of the data. Minimum and maximum values are also noted, as well as interquartile ranges for each variable.

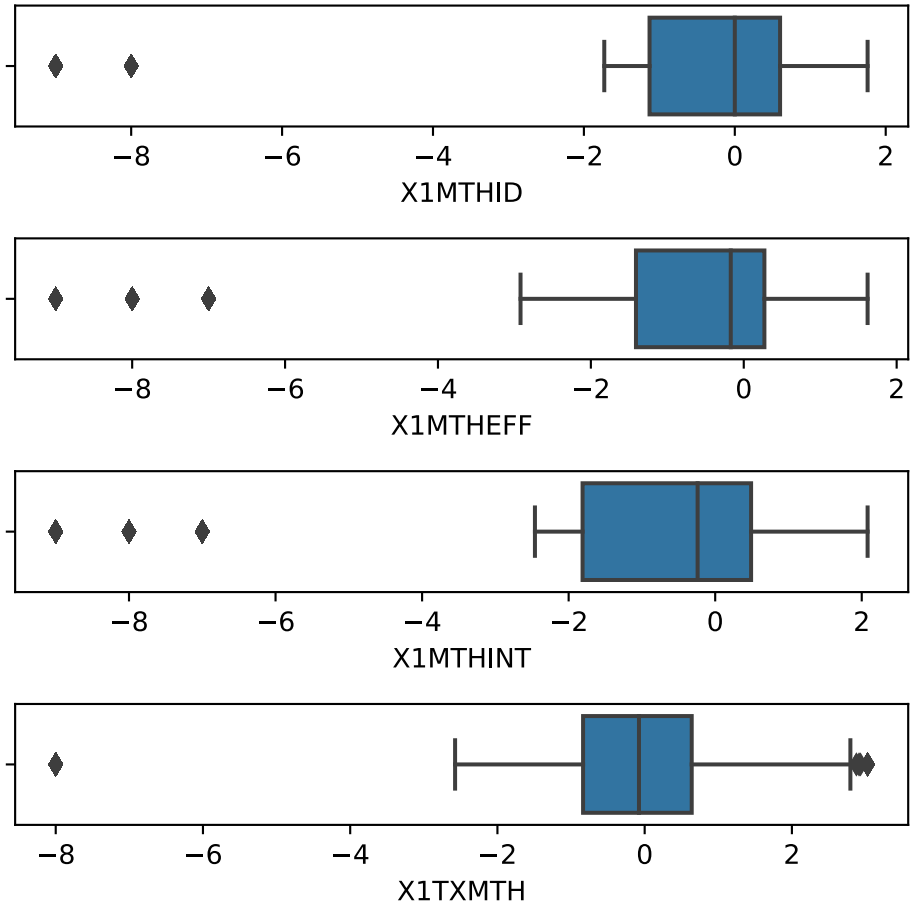
```
summary1 = hsls_data[['X1MTHID', 'X1MTHEFF', 'X1MTHINT', 'X1TXMTH']].describe()
print(summary1)
```

```
##            X1MTHID      X1MTHEFF      X1MTHINT       X1TXMTH
## count  23503.000000  23503.000000  23503.000000  23503.000000
## mean      -0.775739     -1.515598     -1.660666     -0.669273
## std        2.625987      3.237070      3.354718      2.453563
## min       -9.000000     -9.000000     -9.000000     -8.000000
## 25%       -1.130000     -1.410000     -1.810000     -0.836500
```

```
## 50%         0.000000     -0.170000     -0.240000     -0.077400
## 75%         0.600000      0.270000      0.490000      0.639000
## max         1.760000      1.620000      2.080000      3.028300
```

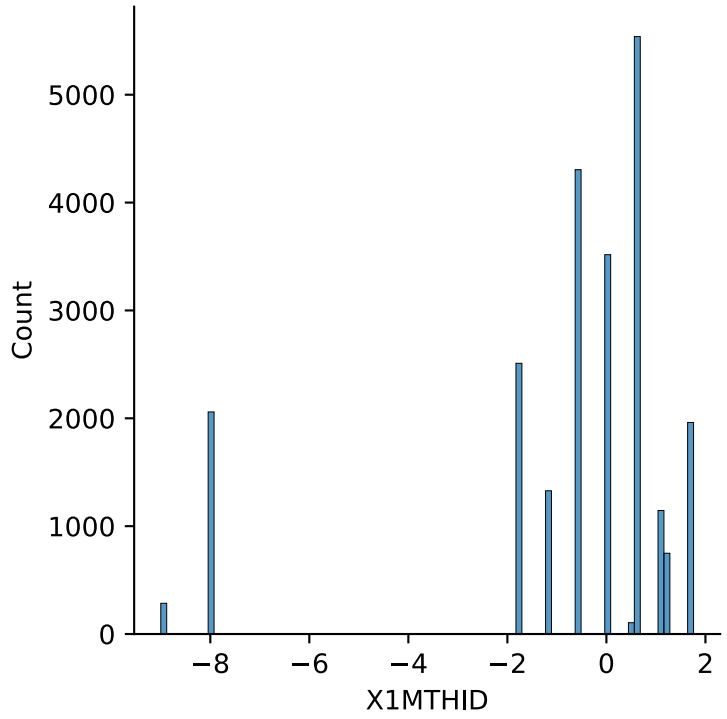
Next, the data and the feature matrix need to be checked for outliers. This can be achieved using a `subplot` from *matplotlib* and the `boxplot` function from the *seaborn* package.

```
warnings.filterwarnings('ignore')

fig, axs = plt.subplots(4, figsize = (5,5))
plt1 = sns.boxplot(hsls_data['X1MTHID'], ax = axs[0])
plt2 = sns.boxplot(hsls_data['X1MTHEFF'], ax = axs[1])
plt3 = sns.boxplot(hsls_data['X1MTHINT'], ax = axs[2])
plt4 = sns.boxplot(hsls_data['X1TXMTH'], ax = axs[3])

plt.tight_layout()
plt.show()
```

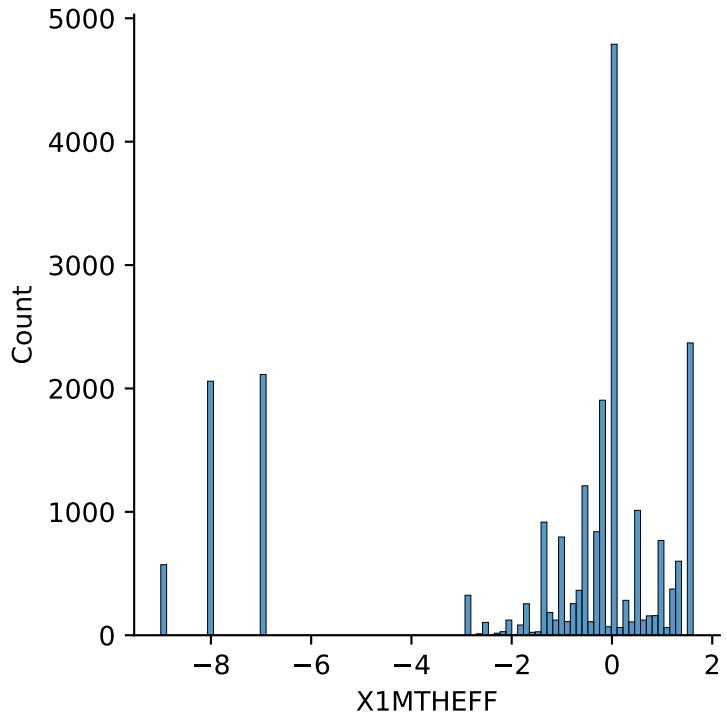


The boxplots illustrate the number of extreme values present in the data, so the next step should be to analyze each variable to find out what is going wrong. Plotting the distributions with `displot` from *seaborn* will be helpful here.

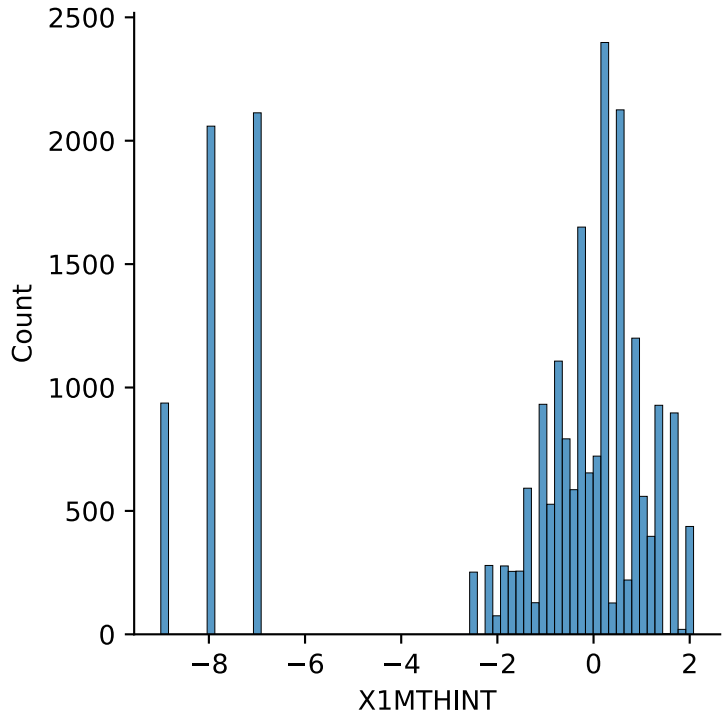
```
sns.displot(hsls_data['X1MTHID'], height=4)
```



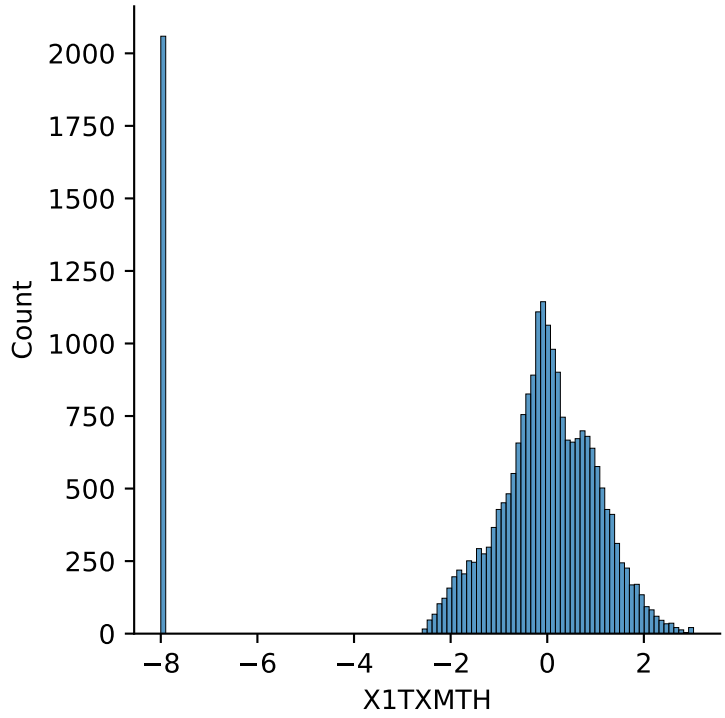

```
sns.displot(hsls_data['X1MTHEFF'], height=4)
```




```
sns.displot(hsls_data['X1MTHINT'], height=4)
```




```
sns.displot(hsls_data['X1TXMTH'], height=4)
#plt.show()
```



The plots above highlight the data clusters around values −7 and −8. To note the significance of these values, the High School Longitudinal Study Codebook may be helpful. The Codebook reveals that any non-response for a student in a given column is assigned a value of −8 for that observation. Further, for

columns **X1MTHEFF** and **X1MTHINT**, any student not taking a Math course in the Fall of 2009 is assigned a value of -7 for that respective column.

2.5.3 Data Cleaning

Using the information found above, the data cleaning process for these variables should be relatively simple. Removing any observation below -7 from these columns is the only step that is necessary, and it can be achieved with the following code chunk:

```
hsls_rna = hsls_data[(hsls_data['X1MTHID']>-7) &
(hsls_data['X1MTHEFF']>-7) &
(hsls_data['X1MTHINT']>-7) &
(hsls_data['X1TXMTH']>-7)]
```

```
print(hsls_data.shape)
```

```
## (23503, 9614)
```

```
print(hsls_rna.shape)
```

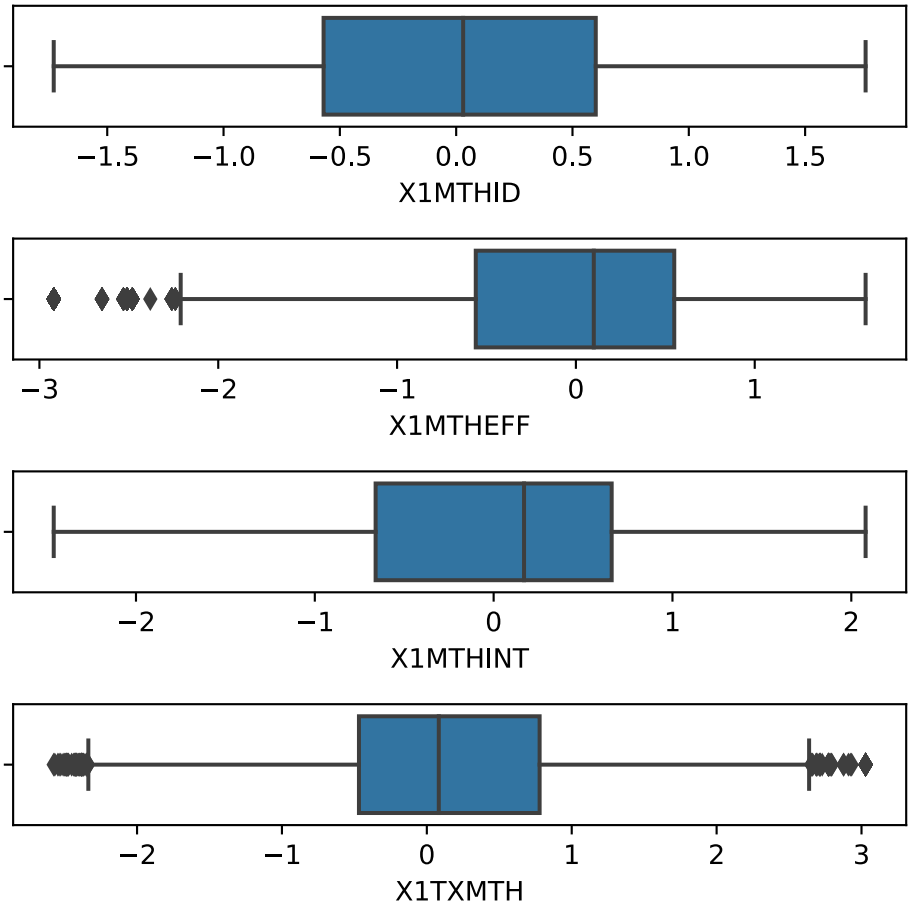
```
## (17960, 9614)
```

Now that the missing data has been removed, the relationship between the target column and the feature matrix should be plotted to try to get a better picture of how the variables interact with each other. The data exploration steps from above are repeated to do this.

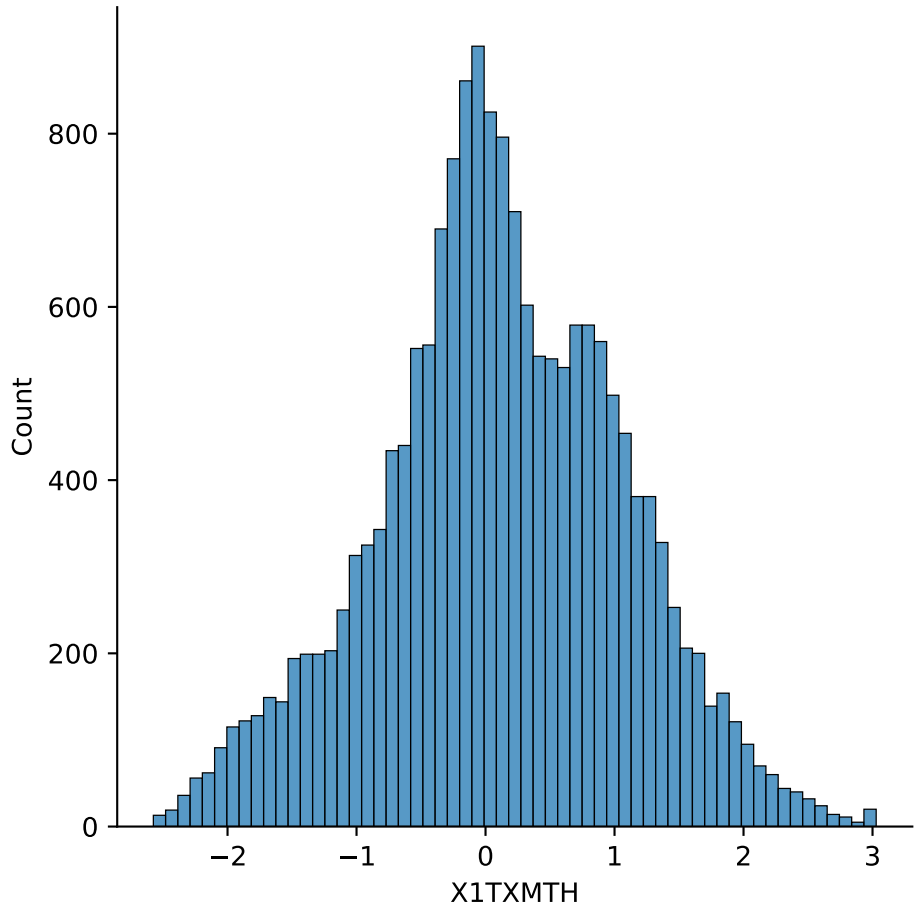
```
summary2 = hsls_rna[['X1MTHID', 'X1MTHEFF', 'X1MTHINT', 'X1TXMTH']].describe()
print(summary2)
```

```
##             X1MTHID      X1MTHEFF      X1MTHINT       X1TXMTH
## count  17960.000000  17960.000000  17960.000000  17960.000000
## mean       0.058670      0.044429      0.037540      0.114428
## std        1.002214      0.992231      0.992286      0.948320
## min       -1.730000     -2.920000     -2.460000     -2.575100
## 25%       -0.570000     -0.560000     -0.660000     -0.468450
## 50%        0.030000      0.100000      0.170000      0.082550
## 75%        0.600000      0.550000      0.660000      0.778400
## max        1.760000      1.620000      2.080000      3.028300
```

```
warnings.filterwarnings('ignore')

fig, axs = plt.subplots(4, figsize = (5,5))
plt10 = sns.boxplot(hsls_rna['X1MTHID'], ax = axs[0])
plt20 = sns.boxplot(hsls_rna['X1MTHEFF'], ax = axs[1])
plt30 = sns.boxplot(hsls_rna['X1MTHINT'], ax = axs[2])
plt40 = sns.boxplot(hsls_rna['X1TXMTH'], ax = axs[3])

plt.tight_layout()
plt.show()
```




```
sns.displot(hsls_rna['X1TXMTH'])
#plt.show()
```

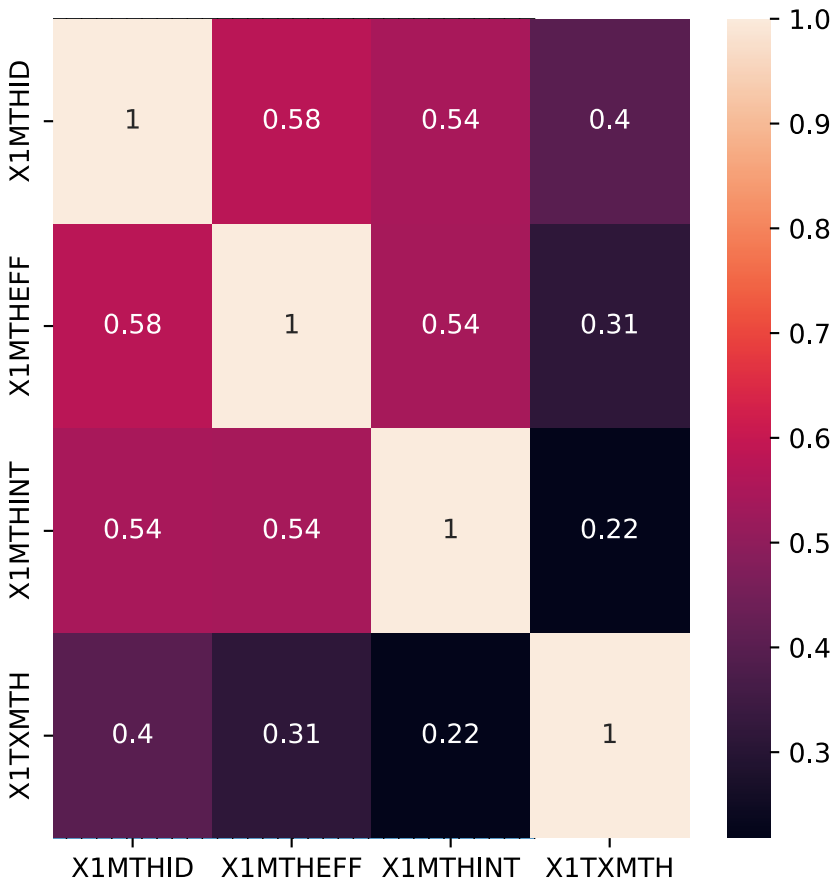


The normality of the data now that the missing values have been removed is also important to note here. This should be expected, given the nature of the **X1TXMTH** variable. Referring to the HSLS Codebook, this information is 'norm-referenced,' meaning it gives an estimate of achievement relative to the population. Given this information, a normally-distributed plot for this variable is expected, and can be seen a lot more clearly when the extreme values are removed from the data.

2.5.4 Heatmap

A correlogram can reveal the correlation between each of our identified variables by displaying a 2D correlation matrix between selected variables. The plot provides a clear visualization of the strength of relationships between variables. A score of '1' inside of the correlation matrix means that there is a perfect, positive relationship between two variables on its axes, whereas a score of '0' would indicate that there is no relationship between the variables.

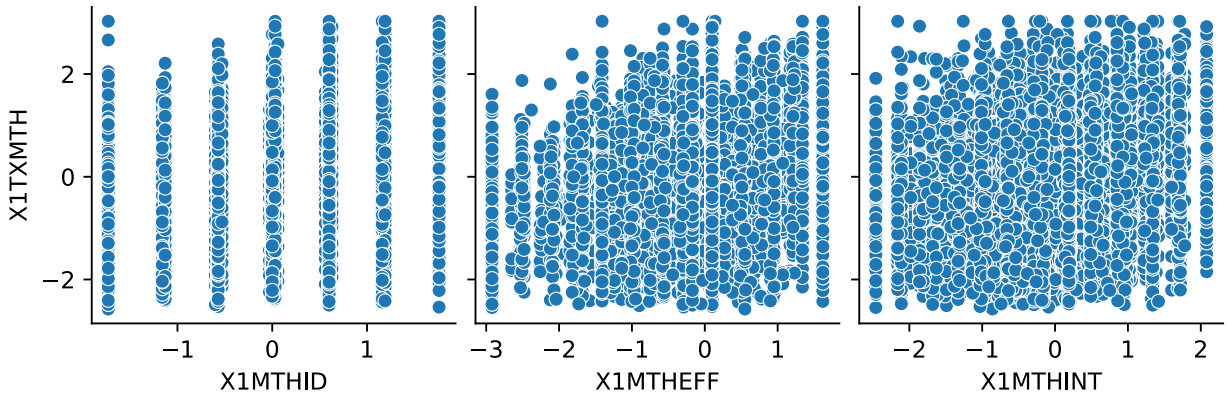
```
warnings.filterwarnings('ignore')

sns.heatmap(hsls_rna[['X1MTHID', 'X1MTHEFF', 'X1MTHINT', 'X1TXMTH']].corr(), annot = True)
plt.rcParams["figure.figsize"] = (7,10)
plt.show()
```



The correlogram reveals that there is a relatively strong, positive relationship between **X1MTHID** and **X1MTHEFF**. Most interesting to this analysis, however, is the relationship between the target variable, **X1TXMTH**, and the feature variables. The heatmap revealed a positive correlation between all of these variables. The strongest relationship was found between **X1TXMTH** and **X1MTHID**, with a correlation coefficient of 0.4. The weakest relationship was found to be between **X1TXMTH** and **X1MTHINT**, with a low correlation coefficient of 0.22.

Overall, an important feature of the heatmap to note is that all of the measured correlations are positive and below a score of 0.5. The reason for this may become clearer after plotting the relationships between the feature variables and the target variables. The `pairplot` function from the *seaborn* package can be used for this.

```
sns.pairplot(hsls_rna, x_vars=['X1MTHID', 'X1MTHEFF', 'X1MTHINT'],
y_vars='X1TXMTH', kind='scatter')
#plt.show()
```



The `pairplot` function plots pairwise relationship between variables in the data. If there were any sort of correlation between the feature variables and the target variable, it would become clear through the shape of the scatterplots above. Instead, no clear pattern is discernible between any of the variables. Nevertheless, this confirms the relatively low correlation coefficients between the variables, especially in the case of **X1MTHINT**, which seems to have no correlation to **X1TXMTH** when observed visually. On the other hand, some slope can be observed on the first two pairwise plots, which is the reason for the higher correlation coefficients for **X1MTHID** and **X1MTHEFF**. However, again, neither of these plots display a particularly promising relationship between any of the variables, which is the reason for all the coefficients being below 0.5. The reason behind the positive correlation coefficients can also be seen, as when one of the feature variables tends to increase, so does the target variable.

2.5.5 Multiple Linear Regression of Confidence on Student Performance

Linear regression is a useful tool for analyzing and predicting a numeric response. It is also flexible and perfect for this situation with multiple feature columns. A multiple linear regression will therefore be run to model the linear relationship between a single dependent variable (**X1TXMTH**) and three independent variables (**X1MTHID**, **X1MTHEFF**, **X1MTHINT**). The analysis should reveal whether the feature matrix does a good job of predicting the target variable and which variables within the feature matrix are significant predictors of the dependent variable.

2.5.5.1 Initial Linear Model First, the identified predictor variables are extracted from the table and assigned to variable `X`. Similarly, the target variable is extracted from the table and assigned to variable `y`. The `StandardScalar` function from the *scikit-learn* library is used to transform the data to follow a Standard Normal Distribution, making the mean equal to 0 and scaling the data to unit variance. Next, the `train_test_split` function is used to split the data into a training and a testing set, using a test size of 20 percent of the data. The model will be trained using the training set and tested against the testing set to measure its predictive accuracy. The `LinearRegression` function from *scikit-learn* will be used to create the linear model and fit it to the training data.

```
X = hsls_rna[['X1MTHID', 'X1MTHEFF', 'X1MTHINT']]
y = hsls_rna[['X1TXMTH']]
X_scaled = StandardScaler().fit_transform(X)
Xtrain,Xtest,ytrain,ytest = train_test_split(X_scaled, y, test_size=0.2, random_state=10)
lin_mod = LinearRegression()
lin_mod.fit(Xtrain, ytrain)
```

```
## LinearRegression()
```

```
lin_mod_score = lin_mod.score(Xtest, ytest)
lin_mod_score
```

```
## 0.17070201307778643
```

```
coefficients = lin_mod.coef_
coefficients
```

```
## array([[ 0.32421736,  0.13233443, -0.04138233]])
```

Using `lin_mod.score`, the R^2 score for our model is measured. In this case, the score was only `0.1707`, meaning that under a linear model, approximately 17 percent of the variance for the dependent variable can be explained by the independent variables. This is not very promising. The coefficients of our model reveal the significance of each column used in the feature matrix. Since **X1MTHID** had the highest score (`0.3242`), it is worth noting that this column had the most influence in predicting the target variable out of the three predictors.

2.5.5.2 Adding Polynomial Features

The created linear model is not very accurate at all right now, but it could potentially be improved by adding polynomial features. This is done by raising existing features to an exponent to allow the model to identify nonlinear patterns in the data. Since `lin_mod_score` is not particularly strong, adding such features could help improve the model significantly. In this case, a pipeline is created using *scikit-learn's* `make_pipeline` function. The pipe adds cubic features (polynomial features of degree 3) and then fits a Linear Regression Model. Its accuracy is measured by the `polyscore` variable.

```
polypipe = make_pipeline(PolynomialFeatures(4),LinearRegression())
polypipe.fit(Xtrain, ytrain)
```

```
## Pipeline(steps=[('polynomialfeatures', PolynomialFeatures(degree=4)),
##                 ('linearregression', LinearRegression())])
```

```
polyscore = polypipe.score(Xtest, ytest)
polyscore
```

```
## 0.16848539412056862
```

The scoring variable reveals that adding polynomial features to the model does not improve it. In fact, it actually makes our model worse, with our scoring variable for the model deceasing from `0.1707` to `0.1684`. There are likely several issues with the created model, and other methods of improving the model should be explored.

2.5.5.3 Adding Regularization

To further improve the model, regularization can be added. Regularization is a method to avoid overfitting the data. By adding regularization, the coefficient estimates are constrained as to be shrunk toward zero. It reduces the complexity of the regression, thus avoiding overfitting.

A new Polynomial Pipeline should be created, but this time with an unspecified degree. The pipe should also include *skikit-learn's* `Ridge` function, which is a Linear Regressor with L2 Regularization. Additionally, a Grid Search can be set up to test various values to find the best value for the model's degree and alpha components.

```
polypipe2 = make_pipeline(PolynomialFeatures(), Ridge())
grid = {'polynomialfeatures__degree':[1, 2, 3, 4, 5],'ridge__alpha':[.001,.01,.1, 1, 10, 100]}
search = GridSearchCV(polypipe2, grid)
search.fit(Xtrain, ytrain)
```

```
## GridSearchCV(estimator=Pipeline(steps=[('polynomialfeatures',
##                                         PolynomialFeatures()),
##                                        ('ridge', Ridge())]),
##              param_grid={'polynomialfeatures__degree': [1, 2, 3, 4, 5],
##                          'ridge__alpha': [0.001, 0.01, 0.1, 1, 10, 100]})
```

```
best_score = search.best_estimator_.score(Xtest, ytest)
best_score
```

```
## 0.16989367524503984
```

```
best_degree = search.best_params_
best_degree
```

```
## {'polynomialfeatures__degree': 3, 'ridge__alpha': 10}
```

The variable `best_score` uses the `best_estimator_.score` to calculate the best R^2 obtained by any of the searches carried out by the Grid Search. The `best_params` function calculates the parameters that were used to achieve the best score, based on the parameters entered into the Grid Search. In this case, that was the degree of the polynomial and the alpha of the Ridge. Together, they reveal that adding regularization and the Grid Search yielded a best R^2 score of 0.1698 for this model, which was obtained with a polynomial degree of 3 and an alpha of 10 for the Ridge. Unfortunately, this is hardly an improvement on the previous `polyscore`, meaning that overfitting was probably not the issue here. Ultimately, only about 17 percent of the variance of Mathematics Scores can be explained by the independent variables that measured confidence.

2.6 The Influence of External Factors on Student Performance

Since the linear regression of confidence on student performance did not yield favorable results, the next step in the analysis is to run a Multiple Linear Regression on three new variables: **X1FAMINCOME**, **X1DUALLANG**, and **X1TMEFF**, which measure the Total Family Income from all Sources in 2008, whether a student's first language was only-English or English and an additional language, and the scale of the math teacher's self-efficacy, respectively.

2.6.1 Descriptive Statistics

For the second regression, the Mathematics Theta Score (**X1TXMTH**) will again be used as the target column. The variables for this regression have been chosen to play counterpart to some of the observations made with the selected variables in the first regression. Since the first regression focused on factors the student can adjust themselves, the second regression will investigate the relationship between student performance and external factors that may affect that performance. Therefore, the feature matrix will consist of variables: **X1FAMINCOME**, **X1DUALLANG**, and **X1TMEFF**.

These variables were chosen to represent the influence of external factors on a student's education and testing scores. The first predictor, **X1FAMINCOME**, is a measure of a student's family's socioeconomic status. The second feature variable determines also interprets the background of a student. It assesses the impact of learning when English is not your first language. The final predictor, **X1TMEFF**, represents the abilities of Mathematics teachers, as well as their confidence in themselves. All three variables work together to analyze the impact of external factors on student performance in Mathematics.

Again, the `describe` function from Pandas is used to compute descriptive statistics for the feature matrix and the target column.

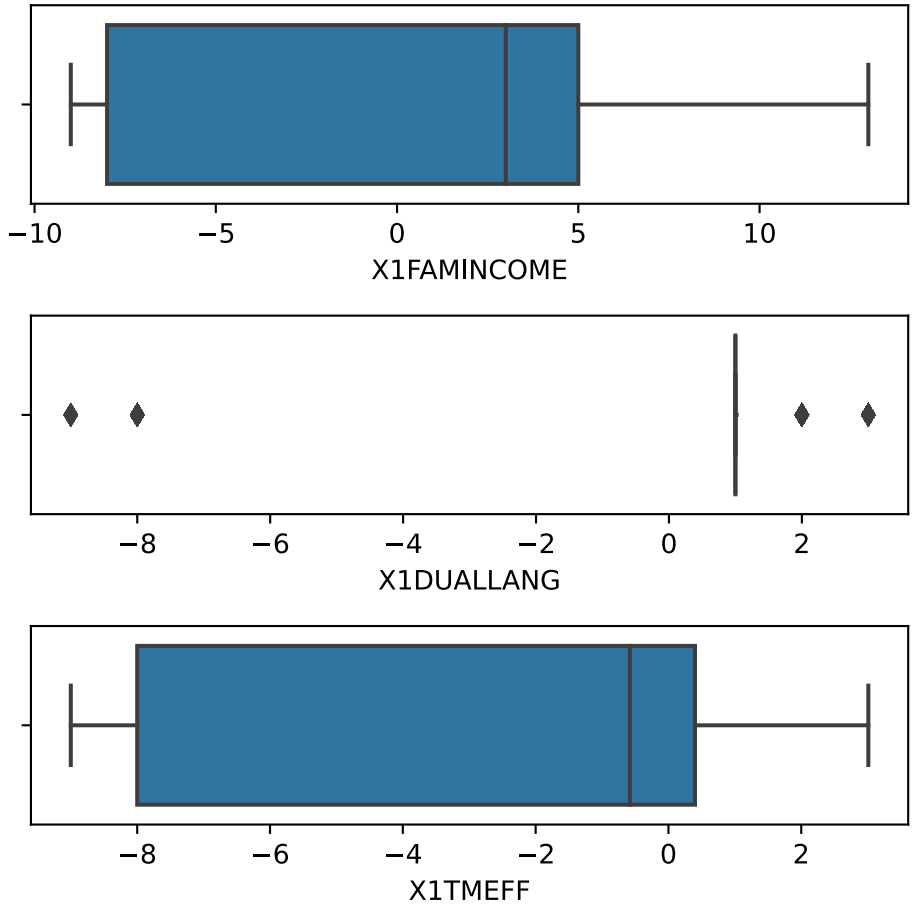
```
summary3 = hsls_data[['X1FAMINCOME', 'X1DUALLANG', 'X1TMEFF', 'X1TXMTH']].describe()
print(summary3)
```

```
##         X1FAMINCOME    X1DUALLANG       X1TMEFF       X1TXMTH
## count  23503.000000  23503.000000  23503.000000  23503.000000
## mean       1.009531      0.409863     -2.897429     -0.669273
## std        6.268507      2.679317      4.039161      2.453563
## min       -9.000000     -9.000000     -9.000000     -8.000000
## 25%       -8.000000      1.000000     -8.000000     -0.836500
## 50%        3.000000      1.000000     -0.580000     -0.077400
## 75%        5.000000      1.000000      0.400000      0.639000
## max       13.000000      3.000000      3.010000      3.028300
```

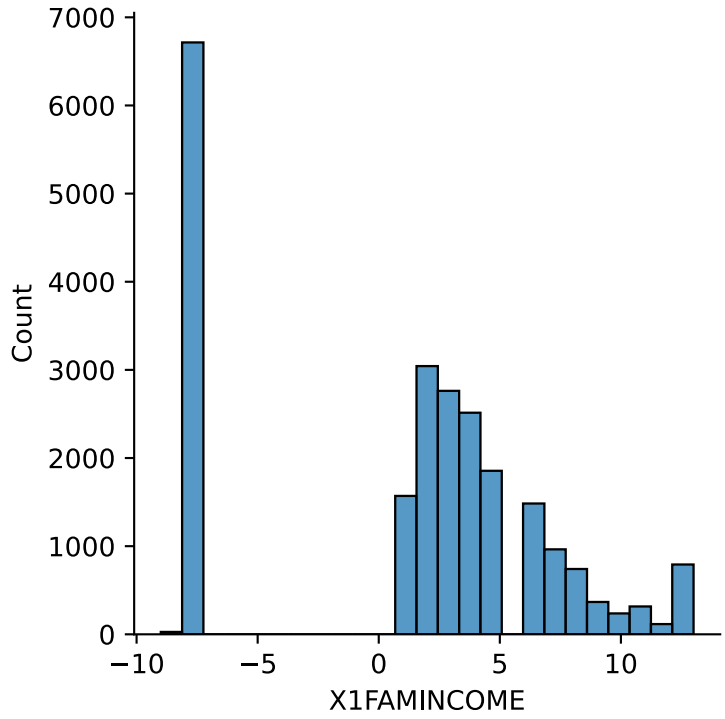
The data needs to be checked for outliers again. The same process of plotting the distribution of the variables will be undertaken using the same functions from *matplotlib* and *seaborn*. Since the Math Theta Score column

(**X1TXMTH**) has already been cleaned, it will be left out of this process.

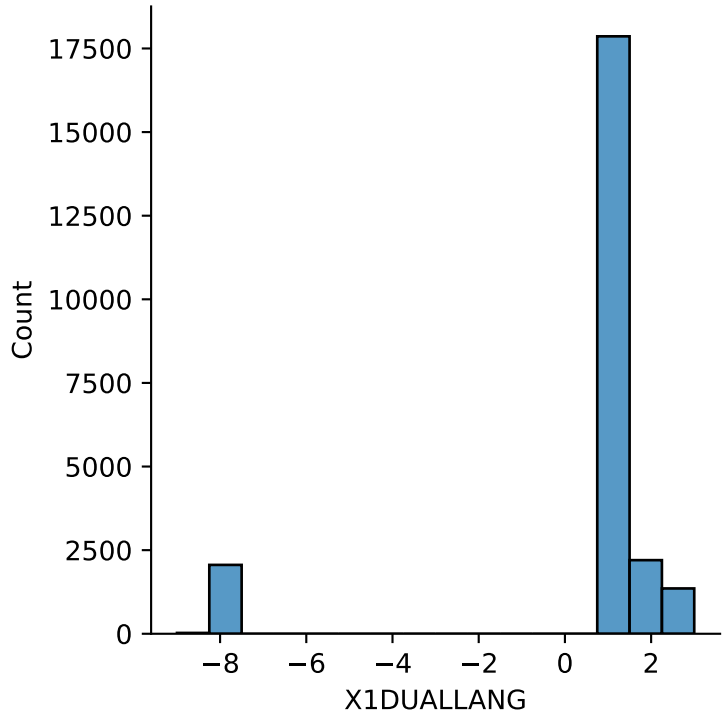
```
warnings.filterwarnings('ignore')

fig, axs = plt.subplots(3, figsize = (5,5))
plt5 = sns.boxplot(hsls_data['X1FAMINCOME'], ax = axs[0])
plt6 = sns.boxplot(hsls_data['X1DUALLANG'], ax = axs[1])
plt7 = sns.boxplot(hsls_data['X1TMEFF'], ax = axs[2])

plt.tight_layout()
plt.show()
```



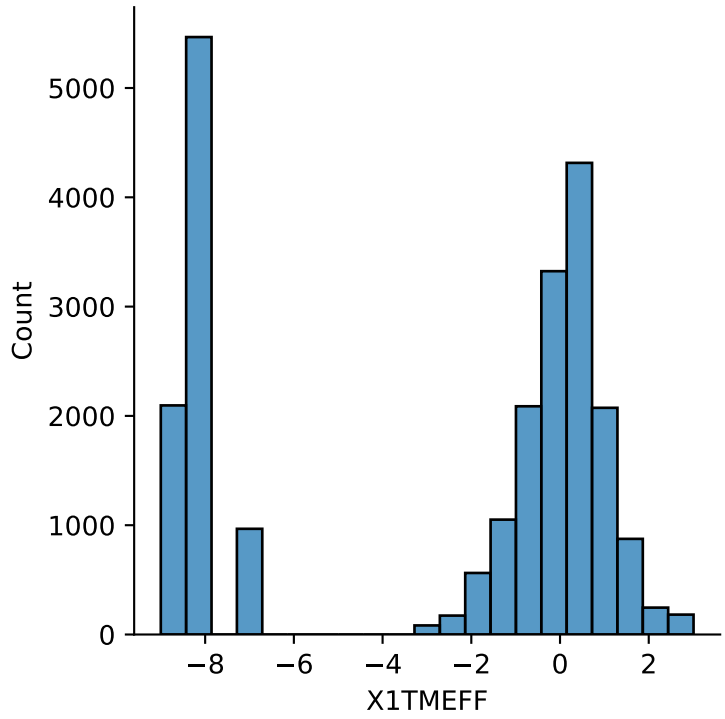

```
sns.displot(hsls_data['X1FAMINCOME'], height=4)
```




```
sns.displot(hsls_data['X1DUALLANG'], height=4)
```




```
sns.displot(hsls_data['X1TMEFF'], height=4)
#plt.show()
```



The boxplots and histograms illustrate the number of extreme values present in the data, which are centered around -8 and -9, as they were in the last regression.

2.6.2 Data Cleaning

The plots above provide a solid base to work from to start the cleaning process. They reveal that like our previous variable set, all of these column place missing and non-response observations at a value below -7. We can confirm this by checking the HSLS Online Codebook again. After doing so, the data cleaning process for this feature matrix will end up looking quite similar to the last cleaning process.

```
hsls_rna2 = hsls_data[(hsls_data['X1FAMINCOME']>-7) &
(hsls_data['X1DUALLANG']>-7) &
(hsls_data['X1TMEFF']>-7) &
(hsls_data['X1TXMTH']>-7)]
```

```
print(hsls_data.shape)
```

```
## (23503, 9614)
```

```
print(hsls_rna.shape)
```

```
## (17960, 9614)
```

```
print(hsls_rna2.shape)
```

```
## (10813, 9614)
```

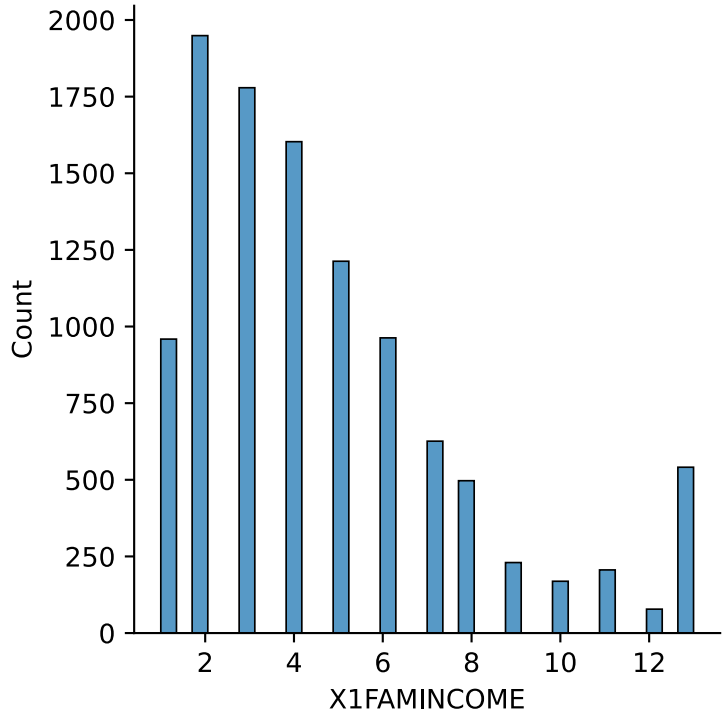
Now that the missing data has been removed, a better picture of the relationship between the target column and the feature matrix can be mapped out.

```
summary4 = hsls_rna2[['X1FAMINCOME', 'X1DUALLANG', 'X1TMEFF', 'X1TXMTH']].describe()
print(summary4)
```

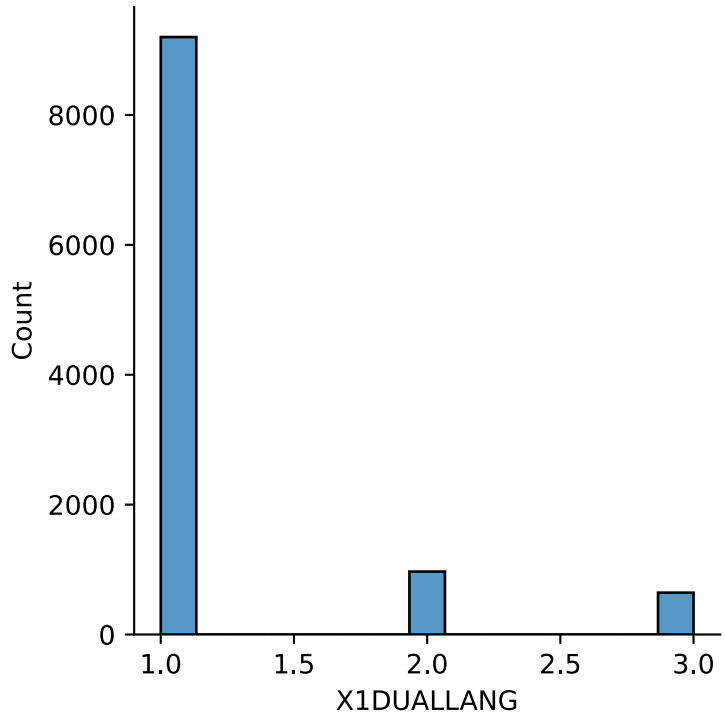
```
##         X1FAMINCOME    X1DUALLANG       X1TMEFF       X1TXMTH
```

```
## count  10813.000000  10813.000000  10813.000000  10813.000000
## mean       4.698234      1.208730      0.100157      0.146253
## std        3.076385      0.533202      0.947971      0.958043
## min        1.000000      1.000000     -3.260000     -2.567200
## 25%        2.000000      1.000000     -0.450000     -0.436900
## 50%        4.000000      1.000000      0.200000      0.114200
## 75%        6.000000      1.000000      0.660000      0.812100
## max       13.000000      3.000000      3.010000      3.028300
```

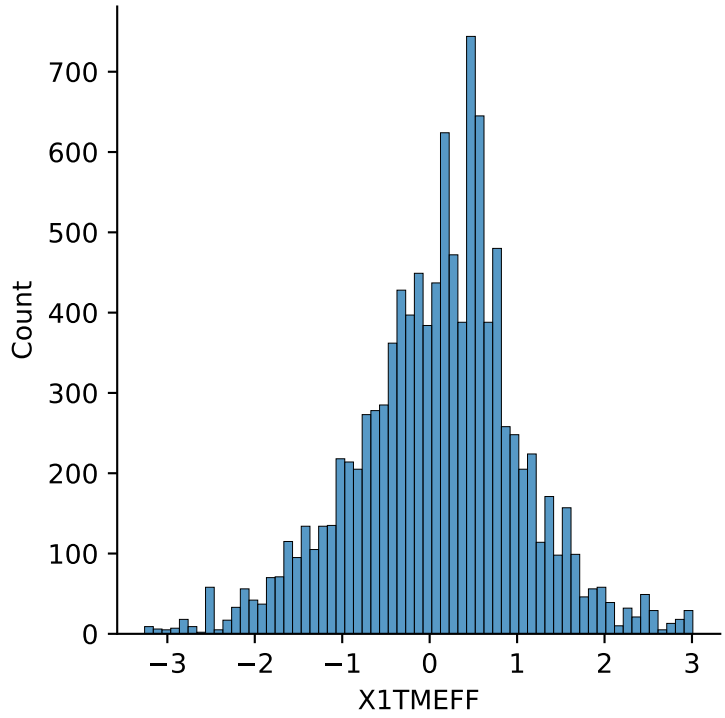
```
sns.displot(hsls_rna2['X1FAMINCOME'], height=4)
```




```
sns.displot(hsls_rna2['X1DUALLANG'], height=4)
```



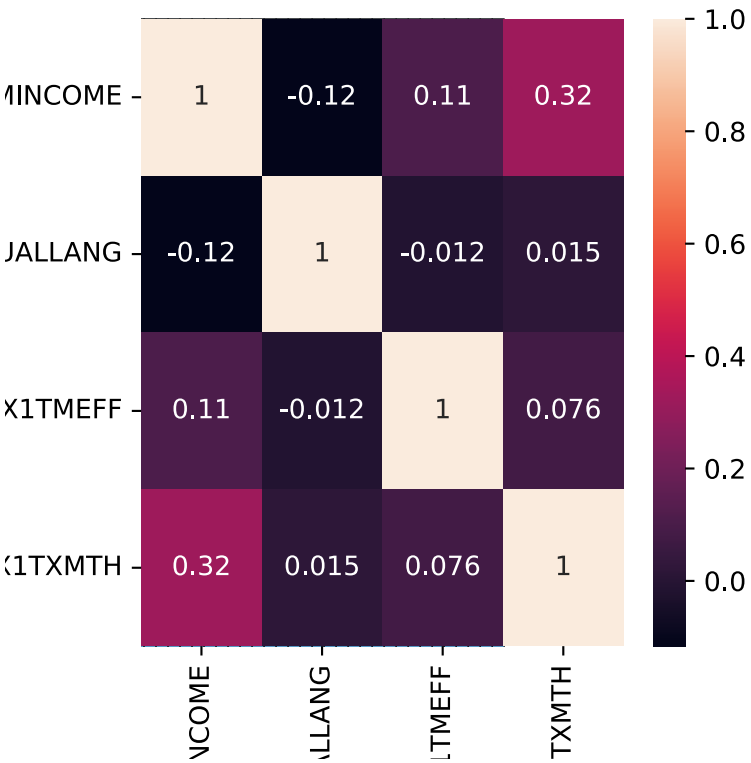

```
sns.displot(hsls_rna2['X1TMEFF'], height=4)
#plt.show()
```



2.6.3 Heatmap

Reviewing the heatmap again for the new feature matrix is a great place to start before diving into analysis.

```
warnings.filterwarnings('ignore')

sns.heatmap(hsls_rna2[['X1FAMINCOME', 'X1DUALLANG', 'X1TMEFF', 'X1TXMTH']].corr(), annot = True)
plt.rcParams["figure.figsize"] = (7,10)
plt.show()
```



The relationship between the feature variables and the target column seems to generally be weaker than when the heatmap was created for the first regression on confidence. However, there is a moderate, positive relationship between **X1FAMINCOME** and the target variable, **X1TXMTH** (R score of 0.32). The other feature variables, **X1DUALLANG**, and **X1TMEFF**, hardly had any correlation with the target variable. The R scores for these relationships was less than 0.1, which is very low. The reasons for these low R scores will have to be explored as the regression progresses.

2.6.4 Multiple Linear Regression of External Factors on Student Performance

The linear regression process will now be run again, this time estimating the effect of external factors, such as financial situation, background, and teacher expectations, on student performance.

2.6.4.1 Initial Linear Model First, the identified predictor variables are extracted from the table and assigned to variable `Xb`. Similarly, the target variable is extracted from the table and assigned to variable `yb`. The `StandardScalar` function and the `train_test_split` function from the *scikit-learn* library are both again used to normalize the data and then split it into training and testing sets. The model will be trained using the training set and tested against the testing set to measure its predictive accuracy. The `LinearRegression` function from *scikit-learn* will again be used to create the linear model and fit it to the training data.

```
Xb = hsls_rna2[['X1FAMINCOME', 'X1DUALLANG', 'X1TMEFF']]
yb = hsls_rna2[['X1TXMTH']]
Xb_scaled = StandardScaler().fit_transform(Xb)
Xbtrain,Xbtest,ybtrain,ybtest = train_test_split(Xb_scaled, yb, test_size=0.2, random_state=10)
```

```
lin_mod_b = LinearRegression()
lin_mod_b.fit(Xbtrain, ybtrain)
```

```
## LinearRegression()
```

```
lin_mod_b_score = lin_mod_b.score(Xbtest, ybtest)
lin_mod_b_score
```

```
## 0.107484573036288
```

```
coefficients_b = lin_mod_b.coef_
coefficients_b
```

```
## array([[0.31089137, 0.05655984, 0.03946021]])
```

Using `lin_mod_b.score`, the R^2 score for our model is measured. In this case, the score was `0.1074`, meaning that about 10 pecent of the variance for the Mathematics Theta Score (**X1TXMTH**) can be explained by the independent variables under a linear model. Furthermore, the coefficients of our model reveal the significance of each column used in the feature matrix. These are found using `lin_mod_b.coef_`. The **X1FAMINCOME** had the highest value at `0.3108`. This coefficient is significantly larger than the coefficients for either of our other predictors (`0.05655`, `0.03946`), so this is worth noting and coming back to later.

2.6.4.2 Adding Polynomial Features

The model is not very accurate, but it looks like it could be relying heavily on the **X1FAMINCOME** column. It might be able to be improved by adding polynomial features. *Scikit-learn's* `make_pipeline` function can be used again here to add polynomial features (again of degree 3) and then fit a Linear Regression Model. Its accuracy is measured by the `polyscore_b` variable.

```
polypipe_b = make_pipeline(PolynomialFeatures(4),LinearRegression())
polypipe_b.fit(Xbtrain, ybtrain)
```

```
## Pipeline(steps=[('polynomialfeatures', PolynomialFeatures(degree=4)),
##                 ('linearregression', LinearRegression())])
```

```
polyscore_b = polypipe_b.score(Xbtest, ybtest)
polyscore_b
```

```
## 0.11843104300450269
```

The new polyscore variable reveals that adding polynomial features has improved our model, from an R^2 score of `0.8183` to `0.8618`.

2.6.4.3 Adding Regularization

To further improve the model, regularization can be added. Again, a new Polynomial Pipeline will be created with unspecified degree and with the L2 Regularization of *skikit-learn's* `Ridge` function. The Grid Search will also be set up again to test various values for the model's best degree, as well as the Ridge's optimal alpha.

```
polypipe2_b = make_pipeline(PolynomialFeatures(), Ridge())
grid_b = {'polynomialfeatures__degree':[1, 2, 3, 4, 5],'ridge__alpha':[.001,.01,.1, 1, 10, 100, 1000]}
search_b = GridSearchCV(polypipe2_b, grid_b)
search_b.fit(Xbtrain, ybtrain)
```

```
## GridSearchCV(estimator=Pipeline(steps=[('polynomialfeatures',
##                                         PolynomialFeatures()),
##                                        ('ridge', Ridge())]),
##              param_grid={'polynomialfeatures__degree': [1, 2, 3, 4, 5],
##                          'ridge__alpha': [0.001, 0.01, 0.1, 1, 10, 100, 1000]})
```

```
best_score_b = search_b.best_estimator_.score(Xbtest, ybtest)
best_score_b
```

```
## 0.11724163378113184
```

```
best_degree_b = search_b.best_params_
best_degree_b
```

```
## {'polynomialfeatures__degree': 3, 'ridge__alpha': 100}
```

`best_score_b` uses the `best_estimator_.score` function on the `search_b` variable, which holds the Grid Search fitted on the training sets, to calculate the best R^2 obtained by any search. The `best_params` function calculates the parameters that were used to achieve the best score, based on the parameters entered into the Grid Search. In this case, that was the degree of the polynomial and the alpha of the Ridge. Together, they reveal that adding regularization and the Grid Search yielded a best R^2 score of 0.8602 for this model, which was obtained with a polynomial degree of 3 and an alpha of 100.

2.7 Predicting Mathematics Scores through Classification

Both sets of feature variables returned good results, so the next step in the analysis involves focusing in on one interest cluster and trying to improve those results further. Accordingly, since the regression on confidence yielded slightly higher values for the best R^2 score ($0.8807 > 0.8602$), the variables from the first regression will be used as a part of further analysis.

The regressions just run can be turned into a classification problem using `KBinsDiscretizer` from *skikit-learn*. This function creates a new variable, `y_bins`, which contains a value (2, 1, or 0) reflecting whether the Math Assessment Score for a particular student falls into the top, middle, or bottom third of all test score observations in the data.

```
kb = KBinsDiscretizer(n_bins=3,encode='ordinal')
ybins = kb.fit_transform(hsls_rna[['X1TXMTH']])[:,0]
```

2.7.1 Adding a Support Vector Machine

A Support Vector Machine (SVM) is a flexible supervised machine learning method used for both regression and classification. In this analysis, it will be used for the latter. A Linear Support Vector Classifier (SVC) [`LinearSVC`] will be used to fit the data to return a best-fit 'hyperplane,' which essentially categorizes the data.

Splitting the new data with `y_bins` into training and testing sets, the next step is to set up a new model to analyze the data for classification analysis. This can be achieved with a Support Vector Machine from *skikit-learn*, using a polynomial kernel. Initial analysis will be trialed with an SVC with a degree of 2 and a level of regularization (C) of 10.

```
X2train,X2test,y2train,y2test = train_test_split(X_scaled, ybins, test_size=0.2, random_state=10)

svc = SVC(kernel='poly', degree=3, C=10)
svc.fit(X2train, y2train)
```

```
## SVC(C=10, kernel='poly')
```

```
SVCscore = svc.score(X2test, y2test)
SVCscore
```

```
## 0.44097995545657015
```

The scoring variable, `SVCscore`, is `0.4409` which is not great, but quite a big improvement over what the scoring variable was before. Another alternative is Logistic Regression, which might help improve the model even further.

2.7.2 Logistic Regression

Logistic regression is another good option for classification problems, and *skikit-learn's* `LogisticRegression` function can be used to achieve this. With the help of this function, a predictive analysis will be run to explain the relationship between the three independent variables and the target variable. Ultimately, the probabilities behind these relationships will be estimated using a logistic regression equation.

```
logr = LogisticRegression()
logr.fit(X2train, y2train)
```

```
## LogisticRegression()
```

```
logrscore = logr.score(X2test, y2test)
logrscore
```

```
## 0.45712694877505566
```

The scoring variable, `logrscore`, is again higher than we had last, now at `0.4571`. This is a decent model and explains that under a logistic model, approximately 45 percent of the variance for the dependent variable can be explained by the independent variables. This is a significant improvement from our initial linear model on this feature matrix, which achievced an R^2 score of only `0.1707`.

2.7.3 Confusion Matrix

Constructing a confusion matrix for classification problems can be extremely beneficial to look at how the trained model predicts the testing data. It can be constructed using the `confusion_matrix` function from *skikit-learn*.

```
confusion_matrix(y2test, logr.predict(X2test))
```

```
## array([[693, 191, 310],
##        [515, 182, 517],
##        [232, 185, 767]], dtype=int64)
```

The diagonals of the confusion matrix display true positives (testing values that the model correctly predicted), while the values in the corners represent false positives (cases in which the model incorrectly predicted the test data). It seems like the model did fairly well overall, but did a lot better at predicting assessment scores in the top and bottom thirds, as opposed to the middle third.

2.7.4 Takeaways

The models created pointed towards an emphasis on the intangible. Often, it is easy to only focus on what is visible or on what is quantifiable, like hours spent in the library, or hours spent with a teacher one-on-one. However, this analysis demonstrates the importance of often-overlooked qualities like confidence.

Although correlation does not prove causation, the created models demonstrate that the explanatory variables predicted the dependent variable quite well, indicating that things like identifying with a subject, being confident about that subject, and displaying interest in that subject can go a long way towards improving learning assessment results. Of course, the obvious confounding variable here is the opposite - students with good test scores have always scored well on assessments, and therefore identify more with a certain subject or have more confidence in a certain subject.

2.8 The Impact of COVID-19 on Education

2.8.1 School Closures

After the initial outbreak of COVID-19, most countries around the world decided to temporarily close schools and suspend in-person instruction, in favor of a new remote-learning model. By the end of April 2020,

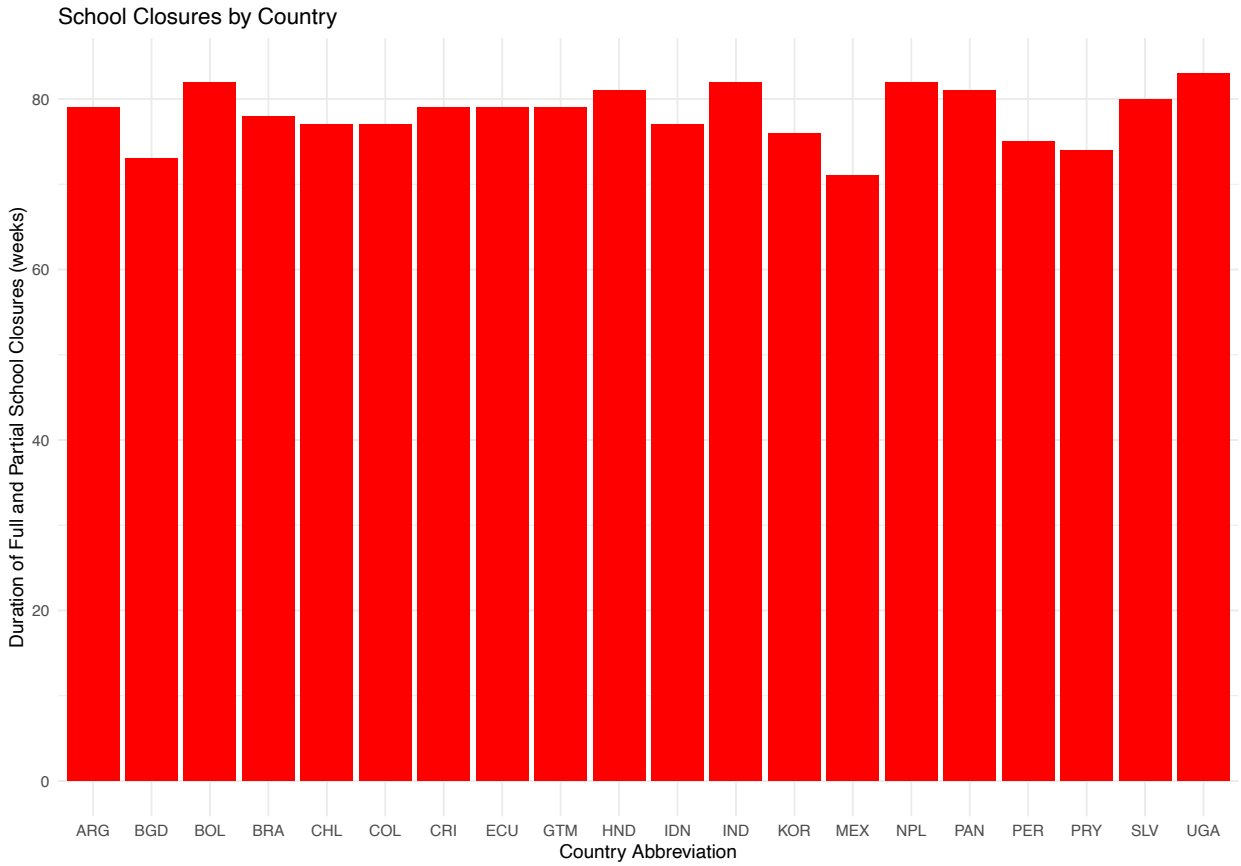
UNESCO estimated that educational facilities had been shut down in 186 countries, affecting close to 74 percent of enrolled students worldwide (UNESCO, 2020).

2.8.1.1 Less Time Spent in Learning Children and young adults spend almost all of their formal education in schools. On a very basic level, a school closure means no place for students to learn. Of course, there are many other ways for students to learn. Online teaching is the immediate and direct substitute for in-person teaching, however, questions about the effectiveness of online learning have remained for years. Not to mention the effectiveness of forced online learning, in which students and teachers only have limited time to adjust to a new, completely virtual learning environment. Students can also learn independently, but this comes down to the student and it is conceivable that the majority of students would opt to do something other than study with their free time. This brings us back to the notion that less time spent in learning results in some sort of interrupted learning or learning-loss. This is supported by a School Barometer survey conducted on Austrian, German, and Swiss students aged between 10 and 19, which found that weekly learning time during the COVID-19 lockdown was reduced by 4 to 8 hours (Huber et al., 2020).

2.8.2 Trends in the Data

```
school_closure_data <- read.csv("duration_school_closures.csv")
```

```
school_closure_data %>%
  arrange(desc(Duration.of.FULL.and.PARTIAL.school.closures..in.weeks.)) %>%
  head(20) %>%
  ggplot(aes(ISO, Duration.of.FULL.and.PARTIAL.school.closures..in.weeks.)) +
  geom_col(show.legend = TRUE,  fill='red') +
  theme_minimal() +
  labs(x = "Country Abbreviation",
       y = "Duration of Full and Partial School Closures (weeks)",
       title = "School Closures by Country")
```



The UNESCO Global Dataset on the Duration of School Closures (2021), supports these trends. In the plot above, the most affected countries, in terms of school closures, are plotted. All of these countries had over 65 weeks of school closures across the years of the pandemic, representing a huge amount of time away from in-person instruction.

2.8.2.1 Inequality in Learning Although COVID-19 school closures brought on a sudden and sharp break from regular school cycles, they did not represent the first extended break from school that data has been collected on. Almost always unrecognized as a gap in education, Summer Break can also lead to a learning-loss or decline in learning, as it exacerbates inequality among students from different socioeconomic backgrounds (Downey et al., 2018). Usually, schools provide equal access to educational resources for all students, regardless of background. This changes when summer break comes around. Students from higher socioeconomic statuses retain access to top-quality materials and opportunities like summer camps. On the other hand, students from lower socioeconomic statuses do not have access to the same resources or opportunities and suffer as a consequence of this (Meyer et al., 2017). This learning-loss can be extended to a similar situation, which is the schooling interruption brought on by the COVID-19 pandemic. In fact, given the number of weeks schools were fully or partially closed during quarantine, the learning deficit caused by COVID-19 might be even more disastrous. As shown in the figure above, countries like Brazil, India, and South Korea experienced well over 70 weeks without school due to the pandemic.

2.8.2.2 Amount of time at Home with Children Another factor to consider when dealing with the socioeconomic status of a student's family is the amount of time parents can spend at home with their children. Considering parents are often a child's first resource for learning, time spent at home with parents is critical for a student's successful education. Moreover, students that do not possess independent learning skills are at a clear disadvantage if parents are not around during summer break or a COVID-induced break from school. Parents of lower socioeconomic status are often unable to spend as much time at home as a parent in a more advantageous situation. Both parents might have to work, or either parent may have to

work long or irregular hours. This can be extremely detrimental to a student's learning and might have other serious ramifications as well.

2.8.2.3 Digital Resources at Home A final aspect of school closures to consider is the disparity between digital resources at school and digital resources at home. Students from higher socioeconomic statuses are much more likely to have access to computers and the internet, while students from lower socioeconomic backgrounds may not be so lucky. Teacher Tapp, an app that asks daily questions to thousands of teachers in the United Kingdom, found that about 10% of students did not have access to a device or the internet after the first week UK school closures (Wilson, 2020). The distance between low and high socioeconomic families is considerably large and growing all over the world, especially in countries like the United Kingdom and the United States. Without access to proper digital resources at home, students from low socioeconomic backgrounds are at risk of falling behind their high socioeconomic status counterparts.

2.8.3 COVID-19 Impact on Student Assessment Exams

To qualitatively measure the impact of COVID-19 on the United States' education system, student assessment scores taken from the California Assessment of Student Performance and Progress (CAASPP) between 2016 and 2021 are reviewed in the following analysis.

2.8.3.1 Read-in CAASPP Data First, CAASPP data is imported into r using `read_csv` from the *readr* package in R. There are five separate .csv files for the CAASPP assessment scores, each containing data on Number and Literacy scores from a specific year. The datasets from 2016 to 2019 and the dataset for 2021 are available for use. The 2020 dataset is not available due to complications in collecting the data caused by COVID-19. This is important to note as the analysis continues to progress.

```
numlit21_data <- read.csv("NumLit_Scores_2021.csv")
numlit19_data <- read.csv("NumLit_Scores_2019.csv")
numlit18_data <- read.csv("NumLit_Scores_2018.csv")
numlit17_data <- read.csv("NumLit_Scores_2017.csv")
numlit16_data <- read.csv("NumLit_Scores_2016.csv")
```

The datasets are joined together using a full_join, and the joined dataset is named joined_data. The column that is most interesting from this dataset is the **Mean Scale Score** variable, which contains aggregate values of student scores in Numeracy and Literacy measured tests. We want to do some data transformations with this column, so it needs to be converted to a double. After doing so, all the missing values in the data are changed to NA, which will help manipulate and filter the data. This concludes the data cleaning process for this dataset.

```
joined_data <- full_join(numlit21_data, numlit19_data)
joined_data <- full_join(joined_data, numlit18_data)
joined_data <- full_join(joined_data, numlit17_data)
#joined_data <- full_join(joined_data, numlit16_data)
```

2.8.3.2 Cleaning, Transforming, and Joining the Data If any of the joins raise an error, as `numlit16data` did here, simply comment out the line and check the data for inconsistencies. In this case, some of the column data types in `numlit16data` did not match the corresponding column data types in `joined_data`. This issue can be resolved by transforming the data so that each column data type is the same in each dataset. The type of the data can be found with the `sapply` function in R.

```
numlit16_data$Students.Tested <- as.factor(numlit16_data$Students.Tested)
numlit16_data$Students.with.Scores <- as.factor(numlit16_data$Students.with.Scores)
numlit16_data$Total.Tested.At.Entity.Level <- as.factor(numlit16_data$Total.Tested.At.Entity.Level)
numlit16_data$Total.Tested.with.Scores <- as.factor(numlit16_data$Total.Tested.with.Scores)
numlit16_data$CAASPP.Reported.Enrollment <- as.factor(numlit16_data$CAASPP.Reported.Enrollment)
```

After making these changes, the `numlit16data` can be successfully joined with the other datasets as follows:

```
joined_data <- full_join(joined_data, numlit16_data)
```

The last thing that needs to be done to prepare the data for analysis is simply making sure the data type of the target variable can be used with aggregate functions in R. Additionally, changing the data type of the column to *double* forces R to convert any missing values to NA. This is advantageous in terms of the filtering and sorting that will be done later.

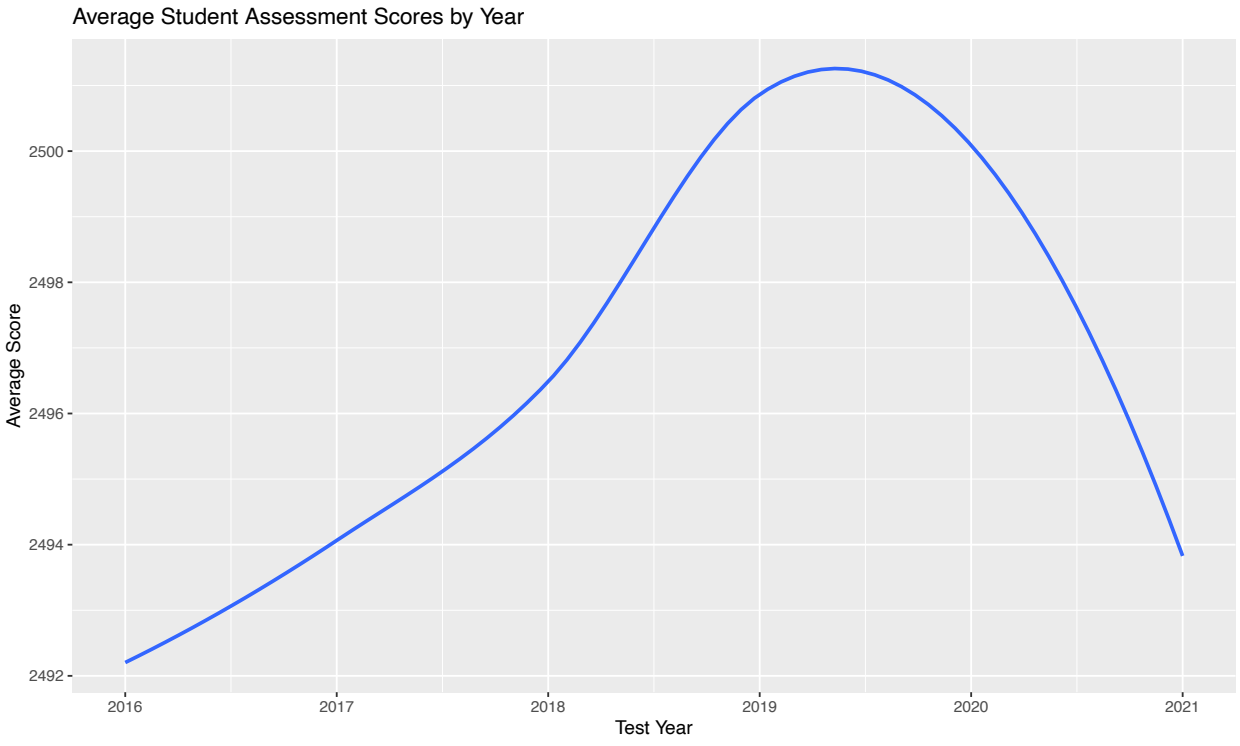
```
joined_data$Mean.Scale.Score <- as.double(joined_data$Mean.Scale.Score)
```

```
## Warning: NAs introduced by coercion
```

The `NAs introduced by coercion` warning is displayed after running this command, which confirms that R has converted missing values in the data to NAs.

2.8.3.3 Plotting Student Performance To look at how COVID-19 affected student performance, `NA` values are first filtered out of the target variable, **Mean Scale Score**, before plotting that Score versus Testing Year. This is achieved using the `geom_smooth` function from the *ggplot2* library.

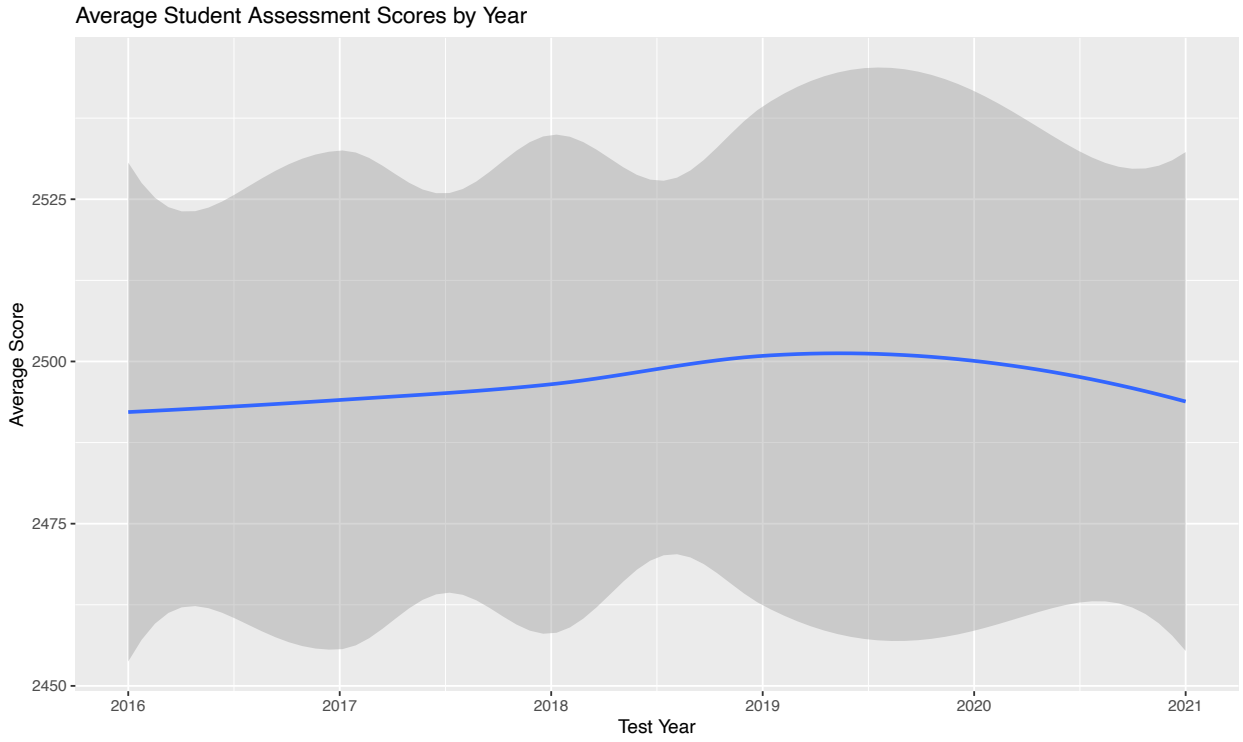
```
joined_data %>%
  filter(!is.na(Mean.Scale.Score)) %>%
  group_by(Test.Year, Grade) %>%
  summarise(Average.Score = mean(Mean.Scale.Score)) %>%
  ggplot(aes(x = Test.Year, y = Average.Score)) +
  geom_smooth(se = FALSE) +
  labs(x = "Test Year",
       y = "Average Score",
       title = "Average Student Assessment Scores by Year")
```



The data reveals that there was a slight dip in performance in 2021. Student performance is increasing until 2019, after which the scores from 2021 drop off significantly. Of course, this is only one data point, so it is

not particularly convincing. Our results can be confirmed by plotting the same graph, this time including a confidence interval around the `geom_smooth` layer.

```
### with Confidence Interval around Smooth
joined_data %>%
  filter(!is.na(Mean.Scale.Score)) %>%
  group_by(Test.Year, Grade) %>%
  summarise(Average.Score = mean(Mean.Scale.Score)) %>%
  ggplot(aes(x = Test.Year, y = Average.Score)) +
  geom_smooth() +
  labs(x = "Test Year",
       y = "Average Score",
       title = "Average Student Assessment Scores by Year")
```



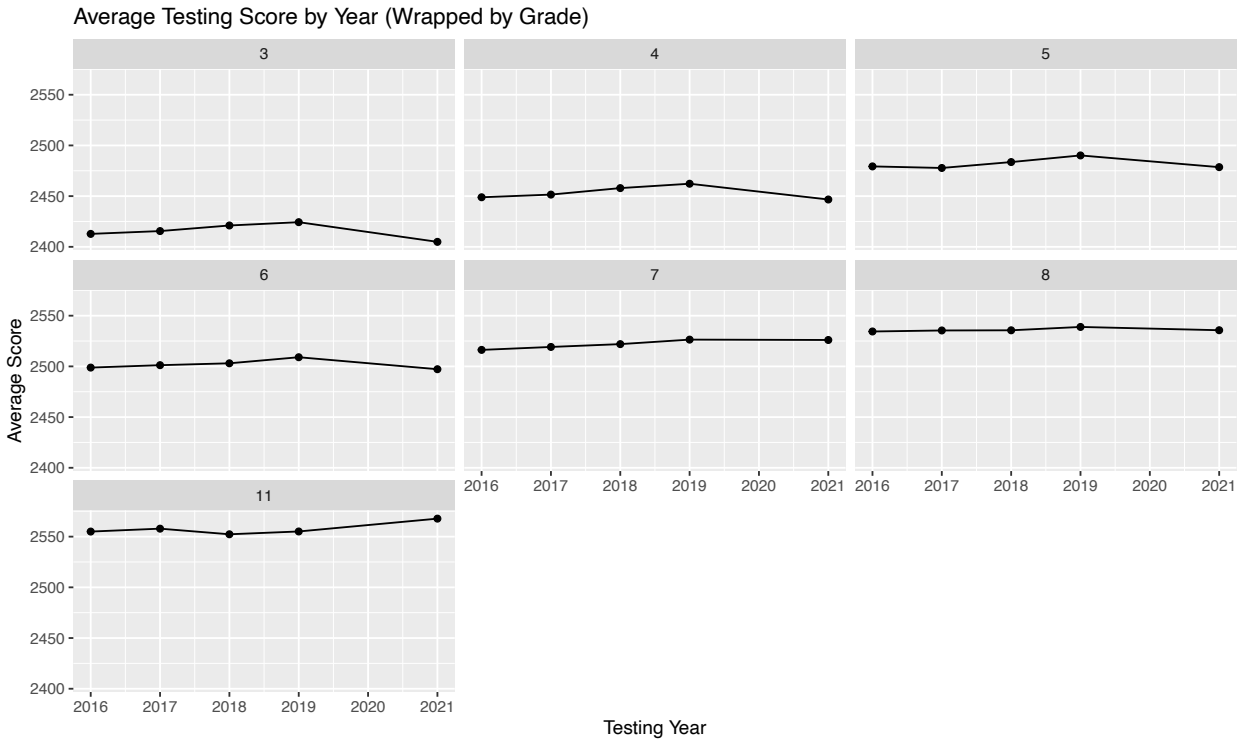
This confirms the drop-off in results from 2019 to 2021, but the confidence interval on this figure is very broad, so findings should be taken with a grain of salt. Nevertheless, the reason for the drop-off can be inferred to be due to the effect of school closures and the potential build-up of learning-loss over this schooling interruption. To dive further into this, a good place to start might be to split up students by grade. This will determine whether the drop in performance is common across all schooling grades or not.

```
plt_data1 <- joined_data %>%
  filter(!is.na(Mean.Scale.Score)) %>%
  group_by(Test.Year, Grade) %>%
  summarise(Average.Score = mean(Mean.Scale.Score))
head(plt_data1, 10)
```

Test.Year	Grade	Average.Score
2016	3	2412.762
2016	4	2448.868
2016	5	2479.319

Test.Year	Grade	Average.Score
2016	6	2498.812
2016	7	2516.256
2016	8	2534.412
2016	11	2554.996
2017	3	2415.537
2017	4	2451.510
2017	5	2477.770

```
plt_data1 %>%
  ggplot(aes(x = Test.Year, y = Average.Score)) +
  facet_wrap(vars(Grade), ncol = 3) +
  geom_point() +
  geom_line() +
  labs(x = "Testing Year",
       y = "Average Score",
       title = "Average Testing Score by Year (Wrapped by Grade)")
```



It looks like 3rd graders had the largest drop-off in performance, followed by 4th, 5th, and 6th graders. Inspecting the plots more closely, this COVID-related drop-off in performance from 2019 to 2021 seems to decrease as students get older. To look into this further, the difference between scores in 2021 and 2019 can be measured and the difference between them can be plotted.

```
head(plt_data1, 10)
```

Test.Year	Grade	Average.Score
2016	3	2412.762
2016	4	2448.868

Test.Year	Grade	Average.Score
2016	5	2479.319
2016	6	2498.812
2016	7	2516.256
2016	8	2534.412
2016	11	2554.996
2017	3	2415.537
2017	4	2451.510
2017	5	2477.770

```
plt_data2 <- plt_data1 %>%
  filter(Test.Year == 2019) %>%
  mutate(`2019 Test Scores` = Average.Score)
plt_data3 <- plt_data1 %>%
  filter(Test.Year == 2021) %>%
  mutate(`2021 Test Scores` = Average.Score)
```

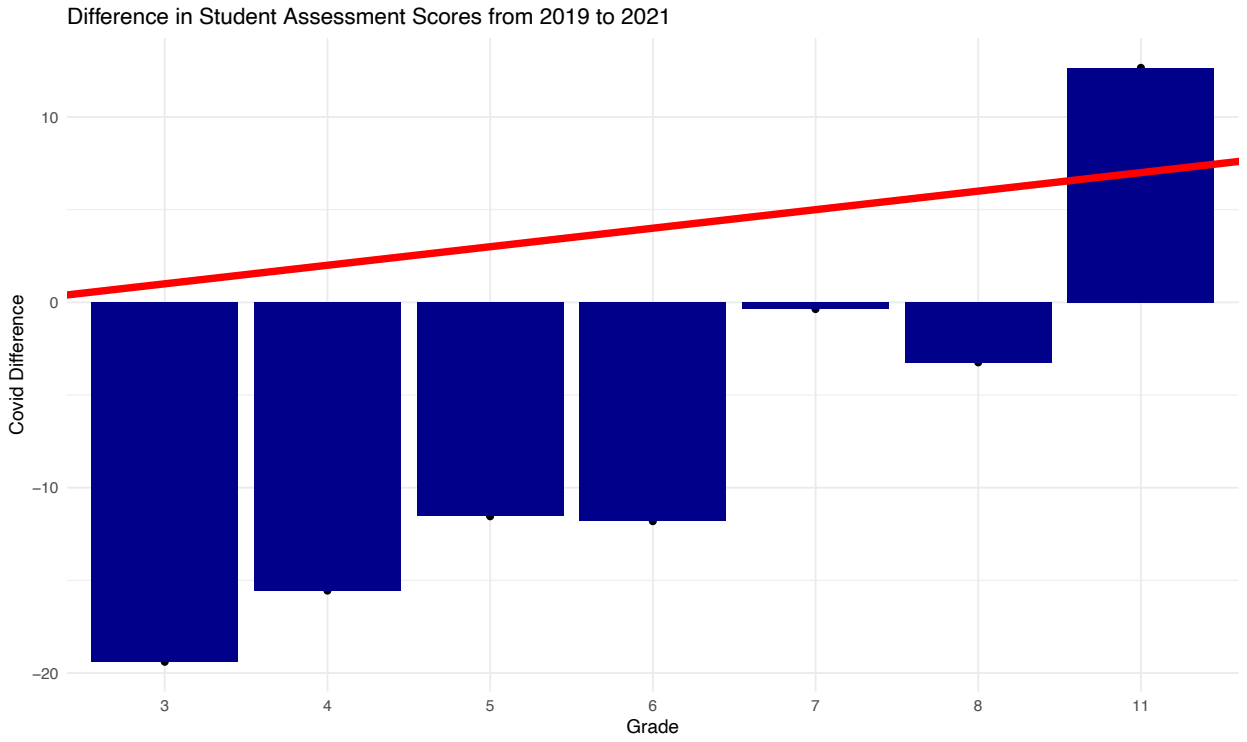
```
var1 <- 1:7
plt_data4 <- tibble(
  `2019 Test Scores` = var1,
  `2021 Test Scores` = var1
)
```

```
plt_data4$`2019 Test Scores` <- paste(plt_data2$`2019 Test Scores`)
plt_data4$`2021 Test Scores` <- paste(plt_data3$`2021 Test Scores`)
```

```
plt_data4$`Covid Difference` <- as.numeric(plt_data4$`2021 Test Scores`) - as.numeric(plt_data4$`2019 T
plt_data4$`Grade` <- c(3, 4, 5, 6, 7, 8, 11)
plt_data4$Grade <- as.factor(plt_data4$Grade)
plt_data4 <- select(plt_data4, Grade, `2019 Test Scores`, `2021 Test Scores`, `Covid Difference`)
plt_data4
```

Grade	2019 Test Scores	2021 Test Scores	Covid Difference
3	2424.30556790733	2404.91226681741	-19.3933011
4	2462.22358338054	2446.67024376576	-15.5533396
5	2490.12463034231	2478.58055400112	-11.5440763
6	2509.01751918159	2497.2096549436	-11.8078642
7	2526.39075912769	2526.0265128593	-0.3642463
8	2538.89379582756	2535.65321482602	-3.2405810
11	2555.09558582294	2567.74859840661	12.6530126

```
plt_data4 %>%
  ggplot(aes(Grade, `Covid Difference`)) +
  geom_point() +
  geom_col(fill = 'darkblue') +
  geom_abline(color = 'red', size = 2) +
  theme_minimal() +
  labs(title = "Difference in Student Assessment Scores from 2019 to 2021")
```



After some more data manipulation, the difference in scores from 2019 to 2021 can be plotted against students' grade levels. As speculated, the older a student is, the less affected they seem to be from the quarantine schooling interruption. This may be because older students are more independent and are therefore also more likely to be independent learners. Older students are also more mature and more likely to have to study or work towards future professional goals. In their time away from school, students in grade 11 would most likely have studied a lot more than 3rd graders, so their accumulated learning-loss might be a lot less than that of younger students.

2.9 Online Courses and Distance Learning

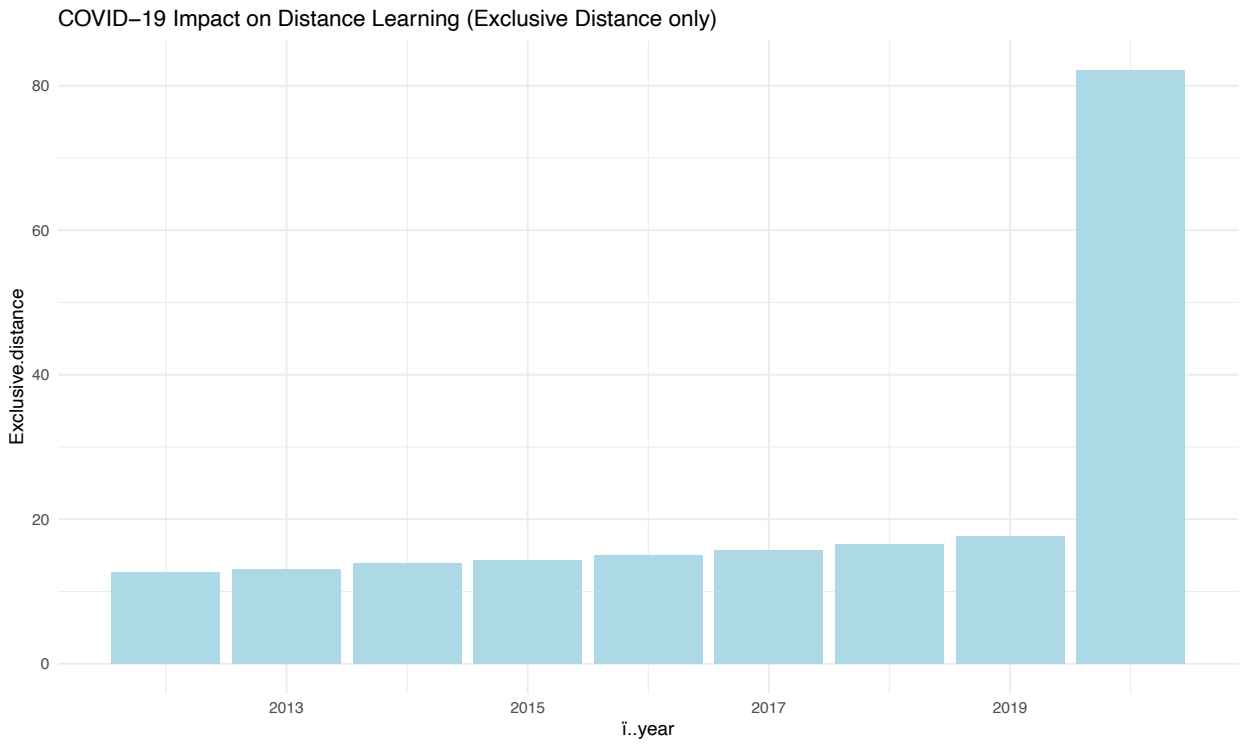
The popularity of MOOCs and online learning platforms have been steadily increasing since the late 2000s. Distance Learning Data taken from the Digest of Education Statistics can help analyze the popularity of such programs in recent years.

2.9.1 Read-in Education Digest Data

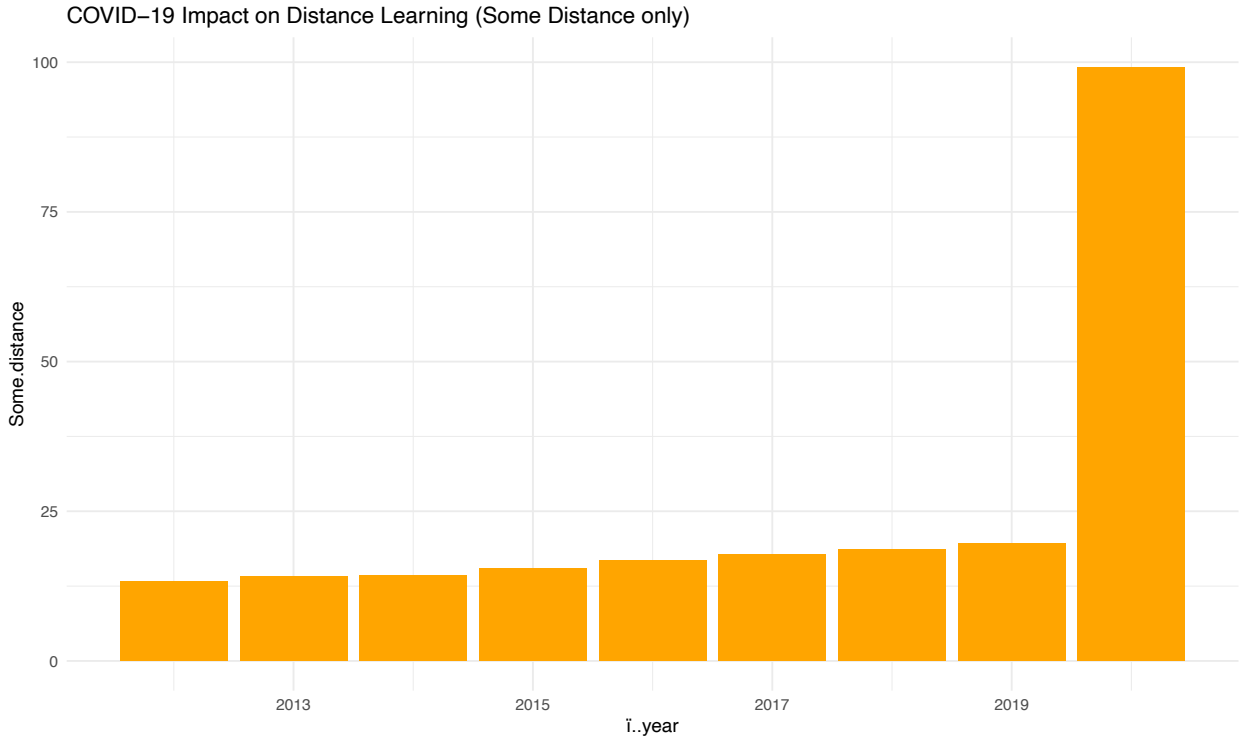
Again, the data is read into R with `read_csv`.

```
distance_learning_numbers <- read.csv("distance_learning_numbers.csv")
```

```
distance_learning_numbers %>%
  ggplot(aes(x=`ï..year`, y=Exclusive.distance)) +
  geom_bar(position="stack", stat="identity", fill='lightblue') +
  theme_minimal() +
  labs(title = "COVID-19 Impact on Distance Learning (Exclusive Distance only)")
```




```
distance_learning_numbers %>%
  ggplot(aes(x=`ï..year`, y=Some.distance)) +
  geom_bar(position="stack", stat="identity", fill = 'orange') +
  theme_minimal() +
  labs(title = "COVID-19 Impact on Distance Learning (Some Distance only)")
```

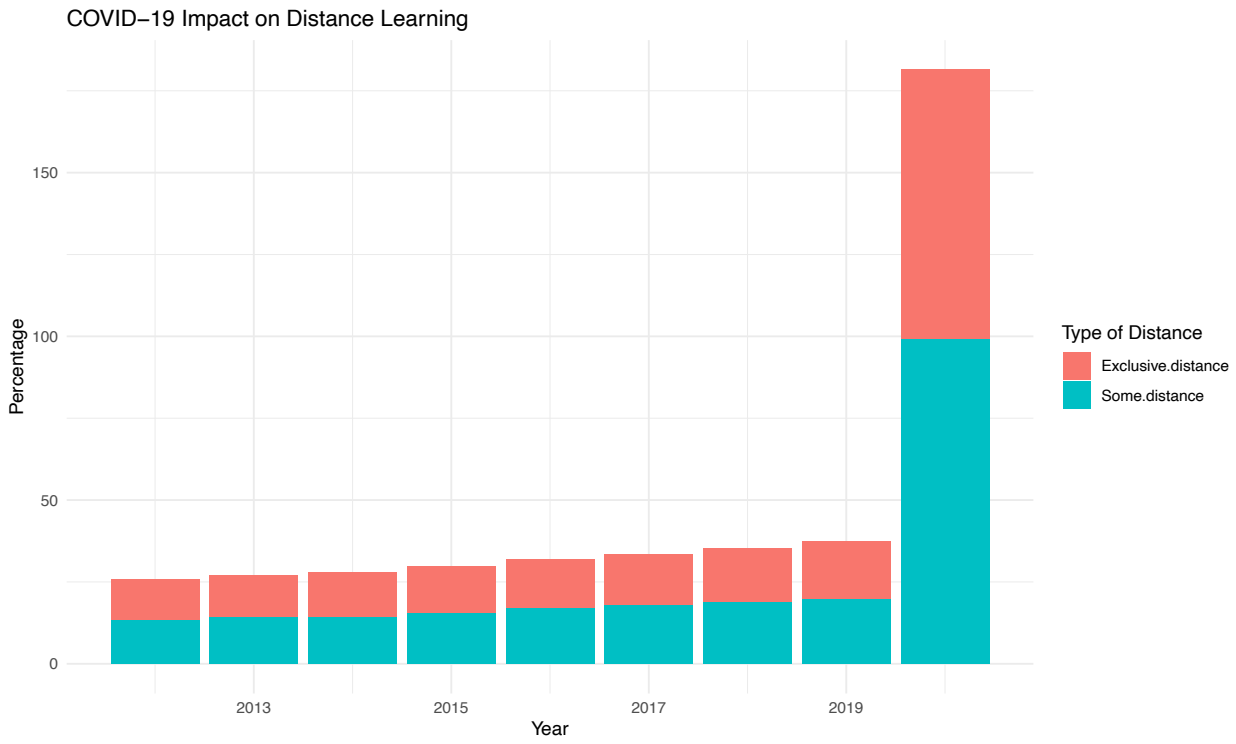


Placing the plots together in a stacked bar chart might be visually helpful. However, some data transformation needs to be performed to make this possible. Looking at the data, to distinguish between the type of distance (some distance vs. exclusive distance), the table needs to be made longer to include this information as a new column. This can be done using the `pivot_longer` function from the *tidyr* library.

```
distance_learning_data <- distance_learning_numbers %>%
  pivot_longer(!`ï..year`, names_to = "Type of Distance", values_to = "Percentage")

names(distance_learning_data)[names(distance_learning_data)=="ï..year"] <- "Year"
distance_learning_data
```

Year	Type of Distance	Percentage
2012	Exclusive.distance	12.6
2012	Some.distance	13.3
2013	Exclusive.distance	13.1
2013	Some.distance	14.1
2014	Exclusive.distance	13.9
2014	Some.distance	14.2
2015	Exclusive.distance	14.3
2015	Some.distance	15.4
2016	Exclusive.distance	15.0
2016	Some.distance	16.8
2017	Exclusive.distance	15.7
2017	Some.distance	17.8
2018	Exclusive.distance	16.6
2018	Some.distance	18.7
2019	Exclusive.distance	17.6
2019	Some.distance	19.7
2020	Exclusive.distance	82.2
2020	Some.distance	99.2

```
distance_learning_data %>%
  ggplot(aes(x=Year, y=Percentage, fill = `Type of Distance`)) +
  geom_bar(stat="identity") +
  theme_minimal() +
  labs(title = "COVID-19 Impact on Distance Learning")
```



This data reveals the prevalence of distance learning in the United States, as measured by the NCES. The sharp trend upward in popularity coincides with the rise of the COVID-19 pandemic, which saw numbers rise up to 99 percent for some type of distance learning, and 82 percent for schools adopting exclusive distance learning structures. While the data is not available for 2021 and 2022, the popularity of these distance learning courses has been trending upwards since 2012 and is likely that this trend would have continued even without COVID-19. However,the influence of COVID-19 on these numbers is unquestionable, and the United States can expect to see these numbers, especially for partial distance learning models, remain high for the next few years.

2.9.2 Distance Learning Performance Numbers

2.9.3 Read-in Open University Data

Using `read_csv`:

```
assessments <- read_csv("assessments.csv")
courses <- read_csv("courses.csv")
studentAssessment <- read_csv("studentAssessment.csv")
studentInfo <- read_csv("studentInfo.csv")
studentRegistration <- read_csv("studentRegistration.csv")
studentVle <- read_csv("studentVle.csv")
vle <- read_csv("vle.csv")
```

2.9.4 Data Cleaning

Looking at the tables, *studentAssessment* seems the most promising, as it includes the `id_assessment` and `id_student` keys, as well as a `score` column. The keys will be especially helpful for joins, while the `score` column may prove to be an excellent target variable to evaluate student performance with.

```
head(studentAssessment, 10)
```

id_assessment	id_student	date_submitted	is_banked	score
1752	11391	18	0	78
1752	28400	22	0	70
1752	31604	17	0	72
1752	32885	26	0	69
1752	38053	19	0	79
1752	45462	20	0	70
1752	45642	18	0	72
1752	52130	19	0	72
1752	53025	9	0	71
1752	57506	18	0	68

Having data for the type of assessment might be helpful, so joining the *studentAssessment* table with the *assessments* table is a good idea. This can be accomplished with an `inner_join` from the *dplyr* package, which will join the tables only on matching values in both tables on the `id_assessment` key.

```
oulad_data <- studentAssessment %>% inner_join(assessments, by="id_assessment")
```

This process can be repeated with the *studentInfo* table, this time attempting to incorporate information on the students into the joined table.

```
oulad_data <- oulad_data %>% inner_join(studentInfo, by="id_student")
```

The *studentVLE* table contains information on the number of interactions each student made with an online module, but only contains an observation for the number of clicks in a given day. These values can be summed by grouping the data on `id_student` and then adding together the total clicks on a given day for each student. This is accomplished with the `sum` function from R.

```
studentVle2 <- studentVle %>% group_by(id_student) %>% summarize(sum_click = sum(sum_click))
```

The *studentVLE* table can now be joined with the rest of our data.

```
oulad_data <- oulad_data %>% inner_join(studentVle2, by="id_student")
```

The joined table, `oulad_data`, now contains all the information necessary to run an analysis. However, it also might contain a few insignificant columns that were added as a part of the joining process. These columns can be removed with a simple `select` statement.

```
oulad_data <- oulad_data %>% select(-date_submitted, -is_banked,
                                    -code_module.x, -code_presentation.x,
                                    -code_module.y, -code_presentation.y,
                                    -date, -gender, -region, -age_band,
                                    -disability, -num_of_prev_attempts,
                                    -imd_band)
```

2.9.5 Random Forest Model

Random Forest is an example of a supervised machine learning algorithm, which means that it is trained using labeled data and a simple method for computational complexity. It is quite a common algorithm because of its high accuracy and pertinence to both regression and classification problems. Since the dataset being used here provides information on the final result of each module for each student, placed into three categories (pass, fail, distinction), a Random Forest Model of classification seems like a good option here. This model combines the output of multiple decision trees into a single result, meaning several models are created and tested, before a single, final numerical is given.

To prepare to build the model, the first thing to do is make sure that our target column is a factor, so that it can be classified correctly.

```
oulad_data$final_result <- as.factor(oulad_data$final_result)
```

Next, the data should be split into training and testing sets, which can be done with the `sample` function. The testing data will represent 30% of the overall data in this case.

```
set.seed(223)
ind <- sample(2, nrow(oulad_data), replace = TRUE, prob = c(0.7, 0.3))
oulad_train <- oulad_data[ind==1,]
oulad_test <- oulad_data[ind==2,]
```

Checking the data for missing values reveals that quite a few values are missing.

```
sum(!complete.cases(oulad_train))
```

```
## [1] 155
```

```
sum(!complete.cases(oulad_train$score))
```

```
## [1] 155
```

The Random Forest model can be built using the `randomForest` function from the *randomForest* package in R. The number of trees is set to 500 and the 'importance' parameter is set to TRUE. As noted above, there a quite a few missing values in the dataset. The last parameter, 'na.action' is set to 'omit', in order to remove these straggling NAs from the data.

```
rf_model <- randomForest(final_result ~ weight+studied_credits+sum_click,
                    data=oulad_train,
                    importance=TRUE,
                    ntree = 500,
                    na.action = na.omit)
```

Next, the model is used to predict training values. A confusion matrix can be made to observe the accuracy of the model.

```
p2_train <- predict(rf_model, oulad_train)
confusionMatrix(p2_train, oulad_train$final_result)
```

```
## Confusion Matrix and Statistics
##
##              Reference
## Prediction    Distinction  Fail  Pass Withdrawn
##   Distinction         593    18    60        17
##   Fail                 86  3643  1016      1550
##   Pass              20490 18529 81177     14041
##   Withdrawn            63   944   488      2470
##
## Overall Statistics
##
##                Accuracy : 0.6053
##                  95% CI : (0.6028, 0.6078)
##     No Information Rate : 0.5699
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.1458
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
## Statistics by Class:
```

```
##
##                      Class: Distinction Class: Fail Class: Pass
## Sensitivity                     0.027930     0.15747      0.9811
## Specificity                     0.999234     0.97827      0.1503
## Pos Pred Value                  0.861919     0.57871      0.6047
## Neg Pred Value                  0.857167     0.85967      0.8571
## Prevalence                      0.146241     0.15934      0.5699
## Detection Rate                  0.004084     0.02509      0.5591
## Detection Prevalence            0.004739     0.04336      0.9246
## Balanced Accuracy               0.513582     0.56787      0.5657
##                      Class: Withdrawn
## Sensitivity                   0.13663
## Specificity                   0.98824
## Pos Pred Value                0.62295
## Neg Pred Value                0.88948
## Prevalence                    0.12452
## Detection Rate                0.01701
## Detection Prevalence          0.02731
## Balanced Accuracy             0.56243
```

Finally, the model is used to predict testing values. A confusion matrix can again be made to observe the accuracy of the model.

```
p2_test <- predict(rf_model, oulad_test)
confusionMatrix(p2_test, oulad_test$final_result)
```

```
## Confusion Matrix and Statistics
##
##              Reference
## Prediction    Distinction  Fail  Pass Withdrawn
##   Distinction         171     3    42         9
##   Fail                 46  1348   471       721
##   Pass               8861  7948 34503      6225
##   Withdrawn            34   499   258       918
##
## Overall Statistics
##
##                Accuracy : 0.5953
##                  95% CI : (0.5914, 0.5991)
##     No Information Rate : 0.5684
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.1247
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
## Statistics by Class:
##
##                      Class: Distinction Class: Fail Class: Pass
## Sensitivity                     0.018766     0.13758      0.9781
## Specificity                     0.998980     0.97631      0.1400
## Pos Pred Value                  0.760000     0.52127      0.5997
## Neg Pred Value                  0.855398     0.85791      0.8294
## Prevalence                      0.146833     0.15789      0.5684
## Detection Rate                  0.002756     0.02172      0.5560
```

```
## Detection Prevalence           0.003626     0.04167      0.9272
## Balanced Accuracy              0.508873     0.55694      0.5591
##                      Class: Withdrawn
## Sensitivity                   0.11660
## Specificity                   0.98540
## Pos Pred Value                0.53716
## Neg Pred Value                0.88475
## Prevalence                    0.12687
## Detection Rate                0.01479
## Detection Prevalence          0.02754
## Balanced Accuracy             0.55100
```

The Confusion Matrix reveals that the overall accuracy of the model was `0.5953`, which is fairly high and a good fit to the data. It also reveals that the model was best at predicting final results that claimed 'distinction' status, followed by those who simply passed. the other important numbers to note here are the 'importance' numbers, which will be explored in the next section.

2.9.6 Variable Importance

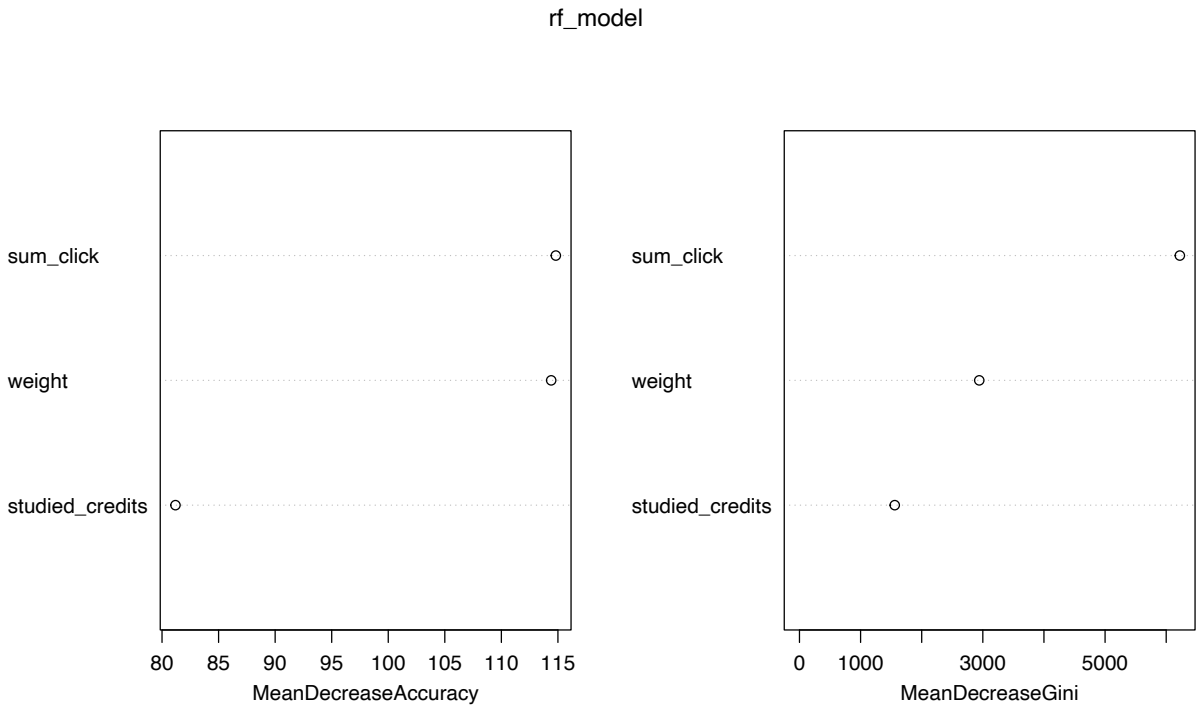
The importance of feature variables used in the model can be calculated using the `importance` function from the *randomForest* library.

```
importance(rf_model)
```

```
##                 Distinction     Fail      Pass Withdrawn MeanDecreaseAccuracy
## weight             33.28651  5.63510  97.89542  68.02863            114.40428
## studied_credits    50.96570 24.57228  65.46359  76.26718             81.18703
## sum_click          60.12304 87.24680 102.64519  92.22498            114.81500
##                 MeanDecreaseGini
## weight                  2941.490
## studied_credits         1558.818
## sum_click               6223.309
```

These values can be visualized by plotting them with the `varImpPlot` command.

```
varImpPlot(rf_model)
```



The plot reveals that the number of clicks a student makes on a module has the highest level of importance in regard to creating the model on the data. This indicates that the column is the most valuable predictor of the dependent variable.

3 Designing the Tertiary Education Institution of the Future

3.1 Core Infrastructure

The university of the future should incorporate the findings of this paper and be premised on the theory of mastery learning. At its heart, the university's infrastructure will take on a similar shape to Khan Academy's, focusing on individualized attention and individualized student learning paths. Moreover, Khan Academy's 'Knowledge Map' should be borrowed and internalized. The 'Knowledge Map' features a sort of dependency tree with lessons as nodes of the tree and connections between lessons demonstrated with attached lines. This sort of visualization is advantageous to the traditional, linear way of looking at lessons, in the sense that the 'Knowledge Map' gives a much more comprehensive and holistic overview of what exactly was learned, and how it connects with related and upcoming material. Nevertheless, linearity within specific subjects should certainly be kept, as having a clear idea of where a given lesson falls in the expanded vacuum of a given subject.

3.2 Class Structure

The class structure of the institution will predominantly include the watching of online videos. There are several reasons for this. First and foremost, online videos allow students to review course material as often as desired. Students learn at different rates, as observed in this paper, and to cater to those differences, schools need individualized tracks of learning for each student. This is only feasible in the real world with the use of technology. The other advantage of the online videos is that it offers an opportunity to standardize the level of teaching across an entire institution. Professors can work together to establish superlative pedagogical

theories and practices, and students can benefit from access that combined, greater knowledge of the school's faculty.

Of course, the learning environments created by in-person campuses are irreplaceable, so the videos will have to be supplemented in some way. The first of such ways would be through one-on-one sessions with faculty or teaching assistants; the second might be group discussions about specific lessons. These sessions would be entirely optional, so students who like to learn independently will not have their learning styles impacted. Moreover, this framework allows those students to move ahead at their own pace, not having to be held back by the speed of the class if they choose to move faster. Finally, this grants students an opportunity to teach other students themselves and gain an even greater understanding of the course material. In fact, teaching is one of the best ways for students to learn. Younger students need a lot more guidance than older students, as seen in this paper's analyses, and once students mature, they will be able to operate as independently as they like.

3.3 Projects and Internships

Alongside online videos, 'projects' will consume most of a student's time at the university. These will be the most analogous to the traditional 'class' component of education. Instead of collecting course credits that count towards graduations, students at this institution will take on various 'projects' of their choosing, in order to collect 'project credits'. 'Projects' do not finish after a certain time-period like classes do. Instead, their lifespan is determined by students and students can choose how much time they spend on a specific 'project' before moving on. These could also take on the form of internships, in an effort to encourage collaboration between students and researchers, or real-world laborers in any field. The point of projects is to teach the fundamental skills that will serve students well throughout the entire course of their life. Instead of traditional university classes, which start and end within 6 months, students taking on 'projects' will learn the process of starting a project/task, planning out all the details for it, and maintaining the dedication and persistence required to see the project through. Along the way, they will find new ways of researching things, gathering information, and things that can only be taught by experience. They should have the chance to bring in professionals and people of importance that relate to their project, and in this way they can already begin to create that professional network that will become invaluable in their later life and early professional career. Moreover, since students have as much time as they want to work on 'projects' of their choosing, they can feel confident in the choices they are making and take all the time they need to learn and master what is in front of them. Students should work with each other and students should be encouraged to seek out help from other students with their 'projects'. For example, if a student needs to build a website for a 'project', but is not necessarily interested in building that website themselves, the institution should create a network for them to work with each other and even get paid for their services. In many ways, these projects are similar to consultant cases and engagements.

3.4 On-Campus Recruiting

Instead of on-campus recruiting, the university of the future could have a student showcase of projects that they have been working on. Employers would walk around and ask students questions about their projects, instead of the other way around. This is not only a great opportunity for students to show off their hard work and gain experience in a formal setting, but also a great opportunity for employers to get a better sense of students they might want to hire.

3.5 Funding and Donations

Instead of asking parents for donations in a very generic fashion, the university of the future could support a 'Sponsored Projects' page, wherein students can post project ideas with detailed descriptions. Parents would be able to scroll through a comprehensive list of these projects, pick out specific ones that pique their interest, and have the option of donating or supporting that project. This platform should be open to parents and alumni, who are great resources that most universities do not take enough advantage of. They can support a project, follow its progress, and even contribute and talk to students involved if they want to.

3.6 A True, Full-time Education

No extended Summer Break. Online instruction allows students to take breaks whenever they need to. This eliminates the need for a long summer break and the extended interruption of schooling. Further, students with fewer digital resources at home have the opportunity to continue to learn with the help of the institution's resources year-round, if they choose to.

3.7 Use of Technology

Technology should be used in as many ways as possible. The most basic way it will be used is to provide each student with a personalized learning path that they can follow and map out on software. This would be somewhat similar to an account built with Khan Academy or a different MOOC. Additionally, this should be accompanied by an iOS/Android application where students can conduct similar learning activities, as well as view course information on an Instagram-style platform.

3.8 Enrollment

University is expensive. Even exorbitantly so. Tuition should be free, and in the institution of the future, it would be. Students would have to pay room and board themselves, if they choose to live on the campus provided by the institution. Students who choose not to live on campus should be able to attend the institution completely virtually, and their entire college experience would be free if they did this.

3.9 Accreditation

The key here is that this institution of the future should represent the first step towards a new style of education that champions mastery learning instead of traditional learning. The current accreditation process in most countries is centered around traditional learning, so accreditation will most likely be impossible early on. However, the necessity of the college degree has been decreasing in recent years, as companies expand their search to non-college-graduates as well. Certification programs that grant certificates to individuals who can prove their prowess in certain subjects are expanding and are now being offered by firms like Google and Microsoft. Other companies like Tesla have publicly stated their propensity to hire anyone with ability, regardless of college degree or not. These advancements exemplify the shift towards a hiring-environment in which university degrees have less power, and ability is supreme.

3.10 Flexibility

The COVID-19 pandemic emphasized the necessity of maintaining flexibility within educational models and infrastructures. Given the removal of time-constraints and the malleable nature of the institution's framework, students are able to choose the type of education they want, and tailor the institution's resources to meet their needs. Usually, if a student becomes ill or falls victim to some sort of emergency, they would need to skip class and thus miss out on invaluable information contained in the lecture they missed. They have to juggle the ramifications of their own personal situation, in addition to the stress caused by having to catch up on work missed for each class. In the tertiary education institution of the future, this would not be an issue, as students learn at their own pace. They choose their own hours of work, meaning higher efficiency and less stress for all. Recorded lectures are always available and they can review them at any time.

3.11 Outreach

The applications of this model of learning are endless, especially for developing countries and countries with struggling education systems around the world. Again, one of the key advantages of a system like this is that anyone enrolled in the institution has access to lecture recordings from the best professors and minds available. While students in such countries may not have had access to this quality of teaching in the past, it is possible with a hybrid tertiary institution like the one suggested.

4 Conclusion

4.1 Key Findings and Results

The analyses performed in this paper revealed several interesting factors in the creation of an educational infrastructure. One of these factors was confidence. Confidence in oneself and one's abilities is incredibly important, especially in the pursuit of academic excellence. Students who identified with or who demonstrated strong levels of self-efficacy and interest in a given subject seemed to score higher scores in mathematics assessments than students who did not. This was confirmed by the accuracy of the first multiple linear regression model on confidence. Further, students who came from favorable backgrounds, school, and living situations, were found to generally score higher on assessments than those with less favorable situations.

The analysis of how COVID-19 impacted schooling demonstrated that students of different ages learn very differently from one another, and interruptions to their learning may have varied effects, depending on the age and grade level of the student. Plotting trends in student assessment data and grouping by grade level illustrated that older students seemed to deal with the COVID learning interruption a lot better than younger students.

Finally, looking into data from the online schooling realm revealed that interaction with the course material is critically important for success in an online environment. Interestingly, it was not the number of credits they earned through study that was most important for performance, it was the number of times they interacted and 'clicked' on material on the webpage.

4.2 Future Improvements

4.2.1 Cohesion

Datasets containing both assessment scores *and* some sort of background or socioeconomic information were generally quite hard to find. For cohesiveness throughout the paper, it might have been more advantageous to continue to work with a single dataset, like the High School Longitudinal Study, and implement various methods of analysis to look at the data.

4.2.2 Redundencies

Some features were added to models in an attempt to improve them but did not yield favorable results. Although turning the initial regression on high school data into a classification problem was an interesting idea, the SVC did not improve the model, and the Logistic Regression failed to improve on results already obtained as well. This beckons the question of whether these features should have been included in this paper at all. These features were also added as an example of what else can be done with the data, so they have been left in place in the paper as a placeholder for future analysis.

4.2.3 Application of Models

The analysis of the High School Longitudinal Study (HSLS) was quite interesting and could have been augmented in several ways. Follow-up data could have been used in analysis and other facets of the data could have been explored in more detail. In the end, quite a few models were created to predict student assessment testing scores but none were applied beyond examining and verifying relationships between variables. The application of these models and their predictions could be an exciting addition to an updated version of this paper.

References

California Department of Education. 2019. "The California Assessment of Student Performance and Progress." https://www.cde.ca.gov/ta/tg/ai/cefcaaspp.asp.

Cellini, Stephanie. 2021. "How Does Virtual Learning Impact Students in Higher Education?" https://www.brookings.edu/blog/brown-center-chalkboard/2021/08/13/how-does-virtual-learning-impact-students-in-higher-education/.

Chaturvedi, Kunal. 2021. "COVID and Its Impact on Education, Social Life and Mental Health of Students - A Survey, Children, and Youth Services Review" 121 (105886). https://doi.org/10.1016/j.childyouth.2020.105866.

Di Pietro, G. 2020. "The Likely Impact of COVID on Education - Reflections Based on the Existing Literature and Recent International Datasets," 1–20. https://doi.org/10.2760/126686.

Downey, Douglas. 2018. "Inequality in Reading and Math Skills Forms Mainly Before Kindergarten - A Replication, and Partial Correction, of "Are Schools the Great Equalizer?"" 91 (323-357). https://doi.org/10.1177/0038040718801760.

Ercolano, Patrick. 2020. "COVID-19 Is Transforming How Companies Use Digital Technology." Johns Hopkins University. https://hub.jhu.edu/2020/07/27/digital-technology-in-business-joel-le-bon/#:~:text=COVID%2D19%20made%20organizations%20who,go%2Dto%2Dcustomers%20strategy.

Huber, Stephen Gerhard. 2020. "COVID-19 and Schooling - Evaluation, Assessment and Accountability in Times of Crises—Reacting Quickly to Explore Key Issues for Policy, Practice and Research with the School Barometer." 32 (237-270). https://doi.org/10.1007/s11092-020-09322-y.

Khan, Salman. 2012. *The One World Schoolhouse - Education Reimagined*. Grand Central Publishing.

Kuzilek, J. 2017. "Open University Learning Analytics Dataset" 4. https://analyse.kmi.open.ac.uk/open_dataset.

Meyer, F. 2017. "Subjectivity of Teacher Judgments - Exploring Student Characteristics That Influence Teacher Judgments of Student Ability." 65. https://doi.org/10.1016/j.tate.2017.02.021.

National Center for Education Statistics. 2009. "High School Longitudinal Study of 2009." https://nces.ed.gov/surveys/hsls09/.

Orlov, George. 2021. "Learning During the Pandemic - It Is Not Who You Teach, but How You Teach" 202 (109812). https://doi.org/10.1016/j.econlet.2021.109812.

Roseman University of Health Sciences. 2022. "Roseman University Six-Point Mastery Learning Model." https://www.roseman.edu/about-roseman-university/six-point-mastery-learning-model/.

Teacher Experiment Academy. 2019. "Mastery Learning - Individual Learning Approach for Faculty and Students." https://tea.dtei.uci.edu/resources/mastery-learning/.

UNESCO. 2022. "Adverse Consequences of School Closures." https://en.unesco.org/covid19/educationresponse/consequences.

UNICEF. 2020. "Education and COVID-19." https://data.unicef.org/topic/education/covid-19/.

Wells, Ester. 2021. "Education in Isolation." https://news.medill.northwestern.edu/chicago/education-in-isolation-create-community-over-curriculum-educators-say/.

Wilson, Mariella. 2020. "The Coronavirus Will Widen the Education Gap in the UK." https://www.aljazeera.com/opinions/2020/4/12/the-coronavirus-will-widen-the-education-gap-in-the-uk/.