Online Learning Outcomes:
Impacts of Demographic, Socioeconomic, and Student Behavior

INTRODUCTION

The world is currently undergoing the largest online learning academic endeavor ever recorded. Around the globe, students and teachers from kindergarten through college, with little warning, have migrated their classrooms from physical to virtual environments. The shear timeline and volume of this shift is astonishing. However, will this new-normal compromise the quality of education? Are there student characteristics or behaviors that will lead to certain educational outcomes? Does a student's socioeconomic status impact their ability to succeed in this online environment? Does the digital divide impact student outcomes? These are some of the important questions we hope to shed light on with our research. We believe there is an opportunity to show correlations between a student's demographics, their socioeconomic status, and their educational behaviors. Through regression analysis, the aim of this report is to determine how demographic, regional socioeconomic and academic behavior factors impact a student's online performance. Specifically, we wanted to know how these factors impact two outcomes: one, a student's average academic numeric score across all classes; and second, how these factors impact the likelihood a student will withdraw from online education. As such, we built OLS Regression to inform question one (numeric score) and Logistic Regression to answer question two (likelihood of withdrawal).

LITERATURE REVIEW

Following is a brief overview of the research on online learning discovered during our research. The most comprehensive study we discovered was the Evaluation of Evidence-Based Practices in Online Learning report publish by the U.S. Department of Education in 2010[i]. This study found that students in online conditions performed modestly better, on average, than those learning the same material through traditional face-to-face instruction. Another study titled "Learning on Demand" by I.E Elaine Allen and Jeff Seaman, both professors at Babson College, in 2010 found that poor economic factors greatly impact the demand for online study more so than that for corresponding face-to-face offerings[ii]. This study helps better understand the link between economic conditions and online learning, but did little to explore how online learning impacts academic outcomes. Generally, what we learned is that there are several studies related to the subject of online learning, but most center on cost effectiveness, teacher effectiveness, comparing learning styles between online and face-to-face instruction, or measuring demand for online courses.

DATA

Source and Collection

We identified a database that contains anonymized data about courses, students and their interactions with a Virtual Learning Environment (VLE) for seven courses. The database is open to the public through the Open University located in Milton Keynes, United Kingdom. The Open University is the "leading university for flexible, innovative teaching and world-leading research" in the United Kingdom and in 157 countries worldwide. The dataset is called the Open University Learning Analytics Dataset (OULAD).

The OULAD Dataset contains information on students, courses, assessments, registration, and interactions with the online learning environment for 32,594 student records and 173,913 academic assessments across 22 courses. The data is organized in a 3rd normal form database. A detailed entity relationship diagram (ERD) can be found in ANNEX 1. We downloaded the database tables and moved them into second normal form to develop the datasets needed to conduct this analysis. A complete Data Dictionary for these tables can be seen in ANNEX 2. However, some of the more important features and their definitions are listed in the table below:

Feature	Description
id_student	A unique identification number for the student.
gender	The student's gender.
region	Identifies the geographic region, where the student lived while taking the module-presentation.
highest_education	Highest student education level on entry to the module presentation.
imd_band	Specifies the Index of Multiple Depravation Band of each location the student lived during the module-presentation (This is a composite value based on the Income, Employment, Health deprivation and Disability, Education Skills and Training, Barriers to Housing and Services, Crime, and Living Environment of each location.
age_band	Band of the student's age.
num_of_prev_attempts	The number times the student has attempted this module.
studied_credits	The total number of credits for the modules the student is currently studying.
disability	Indicates whether the student has declared a disability.
final_result	Student's final result in the module-presentation.
date_registration	The date of student's registration on the module presentation, this is the number of days measured relative to the start of the module-presentation (e.g. the negative value -30 means that the student registered to module presentation 30 days before it started).
sum_click_per_course	The number of times a student interacts with the material
avg_score_all_assessments	The student's average score across all assessment. The range is from 0 to 100. The score lower than 40 is interpreted as Fail.

One key feature that deserves special attention is the imd_band. The Index of Multiple Deprivation (imd_band) feature is a composite feature derived from a collection of socioeconomic indicators about a given region. There are seven domains that build this composite value and they include a region's income, employment, education, health crime, barriers to housing, and living environment.[iii] Each IMD value represents a UK government qualitative measure of English local councils. The IMD ranks each area in England from 1st (most deprived) to 32,844th (least deprived) and bins them in 10 percentage categories reflected in tour dataset. In the context of this analysis, the imd_band feature records where each student resided during their online study. This value provides details on understanding the relationship between a student's socioeconomics and educational outcomes.

Data Augmentation

Although the OULAD dataset provided excellent features for modeling, we added several features that we felt are important in explaining variations in online educational outcomes and allow us to better generalize our findings.

The first additional feature we added was a measure of population for each region, and with that we then calculated and added each region's population density to the dataset (ANNEX 3). The region feature included in the OULAD dataset is helpful, but we wanted to develop features that make comparing named regions in the dataset with similar regions around the world. The population and population density features are more informative than the original region feature and allows us to better understand relationship between a student's demographics (where they come from) and online educational outcomes.

Next we added three features that helped us understand the amount broadband penetration existed in each region. Online study requires a sufficient and dependable internet connection. There has been a lot of discussion on how the digital divide (those with good internet vs those without) impacts daily lives. We wanted to see how it impacts educational outcomes. As such, we included a median and average download speed for each region, as well as the average data usage per region calculated from data included in a Connected Nations Report on the state of the UK's communications infrastructure [iv].

Lastly, we added a feature that helps translate the UK academic system into a standard of measurement consistent with the US education system. Adding these features allowed us to better understand the socioeconomic environments students come from and make some of the original features more universally understandable. A table outlining the details of each feature we created is seen below:

Feature's Added	Description
population	A region's population
population density	A region's population / area
region_broadband_median_download_speed_mbitpers	Median download speed of a region as reported by Connected Nations Report of 2017 as megbits per second
region_broadband_average_download_speed_Mbitspers	Average download speed of a region as reported by Connected Nations Report of 2017 as megbits per second
region_broadband_avg_data_usage_gbs	Average data usage by region as reported by Connected Nations Report of 2017 as Bibabits per second
highest_eduation_us_equivalent	Conversion of UK school system credintials to US standards

In summary, the collection of features from the OULAD database, coupled with the augmented data we added, provides a valuable dataset with information along three key areas of: a student's Demographic information, their Socioeconomic information (to include access to broadband internet), and their academic behaviors.

METHOD

With the merging of the OULAD database and our augmentation features complete, we then organized a specific dataset to answer our research questions. The dataset used to conduct exploratory data analysis and build the models that help answer research questions one and two is titled DS1 Unique Student and is included as part of this research package.

Procedurally, the analysis team worked geographically separated and performed the analysis via collaborative sessions on Microsoft Teams. The initial download and processing of the data used MS excel but its size and complexity required migration to a more robust software system. As such, we used the R statistical software suite (version 3.6.2 (2019-12-12) -- "Dark and Stormy Night") for modeling support for research question one. A complete list of R packages used to establish the analytical environment for this research is included in ANNEX 4. For research question 2 we used Minitab, and Python to conduct our analysis.

Once the data was cleaned, organized, we split our datasets randomly using a hold-out data validation method with an 80% train and 20% test. With processing and business rules established, the analysis team worked individually on developing regression models that helped shed light on each research question. Once all models were built and evaluated independently, the analysis team shared their process and models discussing the strengths and weakness before the group decided which to include as part of this report. This method allowed each analyst to creatively explore different modeling approaches and discover and share their techniques with others. It also resulted in this report offering a "collective best" model for each research question.

ANALYSIS

Data Cleaning and Treatment for Missing Values

The data cleaning phase of this analysis revealed the dataset had two features with missing data. The first, and most substantial feature missing data was the sum_clicks_percourse feature. For our purposes this is a critically important feature in understanding the academic behavior of each student. By recording how often a student interacted with the Virtual Learning Environment (VRE – online academic portal), we are provided a broad measure of how involved they were with each course. Generally speaking, this is the equivalent of classroom participation in a physical academic setting. For reasons not included in the description of the data files, 85% of the observations were missing data for this feature. This discovery lead us to make a develop business that classify a student as "involved." To effectively analyze what factors impact online student outcomes, we decided to condition the data to include only students that were actually involved. The is little mystery in trying to determine an educational outcome of student that simply registers for a course and then never attends. We

determined “involvement” required a student to make at least 1 click in a course. As such, we removed all observations that had less than 1 click.

The next feature with missing values was date_registration. This feature does not include an actual date, but rather a day scale that indicated how soon before, or how soon after a student registers for a course in reference to the course’s start date. For example, if a student registered for a course 1 week before the start of the course, her value would be -7. Although not the case in all instances, this feature can generally be understood as a sense of a student’s enthusiasm for a particular course, or more generally the focus they bring to their academic pursuit. It is reasonable to believe that a student that registers early, compared to a student that registers closer to the start date of the course has a more disciplined approach to their education, which could lead to better academic outcomes. There were approximately 33 observations missing this data. After cleaning these the sum_clicks_percourse and date_registration features, our dataset was reduced from 24,690 observations to 3,594 observations.

Data Organization

Before moving to our exploratory data phase, we partitioned the cleaned dataset into three distinct datasets. The first was the further reduction of students we considered involved. Each student in the dataset has a final_result feature. This feature is a categorical variable that identifies whether a student has “Passed with Distinction”, “Passed”, “Failed”, or “Withdrew.” A breakdown of each category can be seen to right.

final_result Status Breakdown				
	Distinction	Fail	Pass	Withdrawn
Count	338	822	1778	656
% of Data	9%	23%	49%	18%

To answer our first research question, we determined that the student has to meet two criteria: register for and complete the course. Again, there is no mystery in determining the outcome of a student that withdraws from a class. Given this, we decided to partition the student dataset to include only students that “Passed with Distinction”, “Passed”, or “Failed” and titled this dataset “stdataInv”. This partition reduced the cleaned dataset to 2,938 observations.

The second partition of the cleaned dataset was conducted to help answer our second research question of understanding the demographic, socioeconomic, and academic behavior have on a student’s decision to remain engaged in online academics. A student’s decision to withdraw is very important in answering this question. As such, we partitioned a second set of data that included all observations from the cleaned dataset but was organized to categorize a student as “Withdrew” or “Did not Withdraw” maintaining 3,291 observations.

Exploratory Data Analysis

The exploratory data phase consisted of three distinct activities. The first was to provide summary statistics for each of the features in the dataset, the second activity was to create visualizations for both numeric and categorical features, and lastly was to examine the scatterplots and correlations between each feature. Each of these actives helped us better understand the features, their relationships, and provided us some initiation as we prepared to move into the modeling phase.

Summary Statistics:

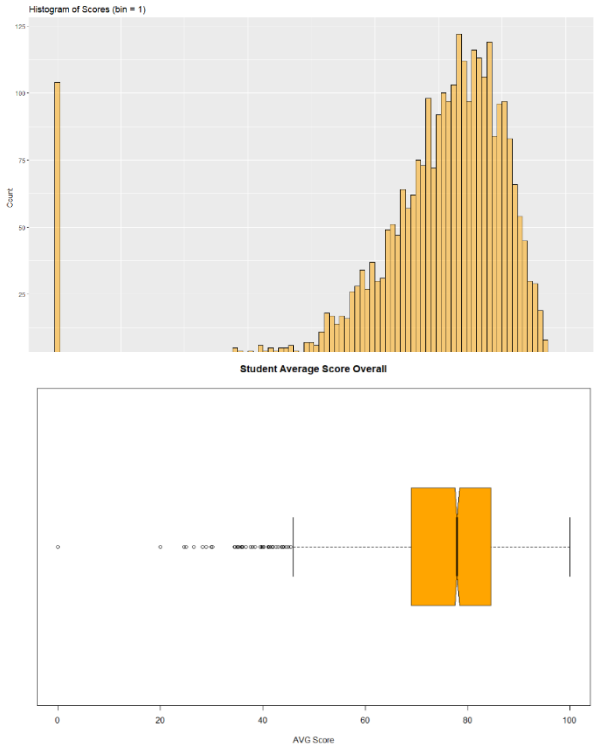
Summary Statistics for the stdataInv dataset	gender	avg_score_all_assessments	population_millions	region_density	region_broadband_median_download_speed_mbitpers
	F:2336	Min. : 0.00	Min. :2.314	Min. : 180.3	Min. :17.50
	M: 602	1st Qu.: 69.00	1st Qu.:4.800	1st Qu.: 574.9	1st Qu.:20.00
		Median : 77.95	Median :5.480	Median : 844.6	Median :23.16
		Mean : 73.76	Mean :5.603	Mean : 1875.9	Mean :26.87
		3rd Qu.: 84.55	3rd Qu.:7.000	3rd Qu.: 1169.7	3rd Qu.:29.94
		Max. :100.00	Max. :9.130	Max. :14675.5	Max. :50.00
highest_eduation_us_equivalent	**imd_band**	**age_band**	**num_of_prev_attempts**	**studied_credits**	**region_broadband_average_download_speed_Mbitspers**
High School (with Honors) :1348	0-10 :340	0-35 :1818	Min. :0.0000	Min. : 60.00	Min. :25.70
Highschool (non-honors) :1123	20-30 :363	35-55 :1085	1st Qu.:0.0000	1st Qu.: 60.00	1st Qu.:26.40
Less than Highschool Diploma: 21	30-40 :375	55 or older: 35	Median :0.0000	Median : 60.00	Median :31.62
Masters Degree or Higher : 12	40-50 :370		Mean :0.1617	Mean : 82.39	Mean :37.93
Some Undergrad College : 434	50-60 :329		3rd Qu.:0.0000	3rd Qu.:120.00	3rd Qu.:44.32
	70-80 :334		Max. :5.0000	Max. :420.00	Max. :63.80
	80-90 :276				
	90-100 :245				
disability	**final_result**	**num_courses**	**date_registration**	**sum_clicks_per course**	**region_broadband_avg_data_usage_gbs**
N:2693	Distinction: 338	Min. :1.000	Min. :-260.00	Min. : 1	Min. :148.6
Y: 245	Fail : 822	1st Qu.:1.000	1st Qu.: -88.00	1st Qu.: 221	1st Qu.:170.1
	Pass :1778	Median :1.000	Median : -52.00	Median : 498	Median :185.5
	Withdrawn : 0	Mean :1.011	Mean : -62.98	Mean : 955	Mean :190.5
		3rd Qu.:1.000	3rd Qu.: -29.00	3rd Qu.: 1067	3rd Qu.:214.1
		Max. :3.000	Max. : 69.00	Max. :15716	Max. :253.6

Data Visualizations and Transformations

The following section provides graphic depictions of the features in the dataset. Each feature has been aggregated and summarized except for two: our dependent variable ang_score_all_assessments and sum_clicks_percourse. These distributions required additional depth in analysis and is it is offered below.

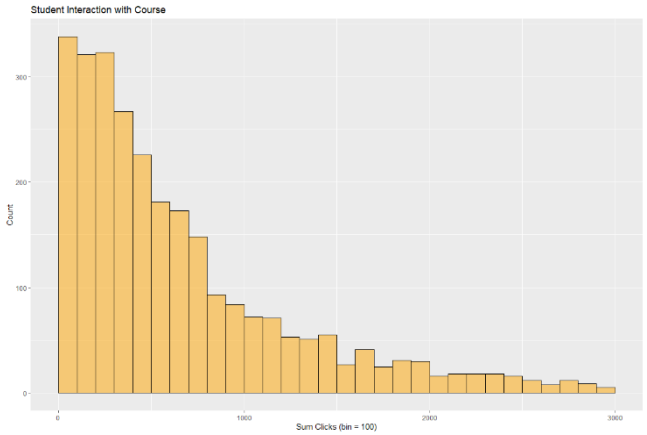
Average Score all Assessments:

This distribution looks like what we would expect, however there is a significant portion of data with zeros for their scores. To better examine this constructed a boxplot. This graphic makes clear there are several low scores below the lower fence. To Determine that cut-off we calculated the lower fence to identify outliers. Our five-number summary shows that Min = 0, Q1 = 69, Median = 77.95, Q3 = 84.55, and Max = 100. This results in an IQR of 15.55 and a Lower Fence value of 45.675. This means values less than 45.675 are technically classified as outliers. We were reluctant to remove observations from our data, however to remain consistent with our methodology to model "involved" student we remove all observations with a score of 0. In an effort to make a more precise model, we determined it unsound to model unengaged students as determined by the lack of basic effort it would take to get a 0 score as this suggest there was no effort taken on behalf of the student.

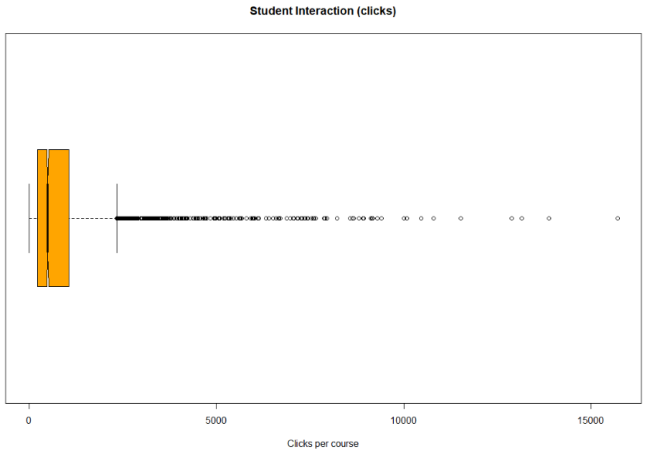


Sum Clicks:

Another feature that required additional analysis is the sum_clicks_percouse. As seen in the image to the right, this distribution is extremely right skewed with a long tail.



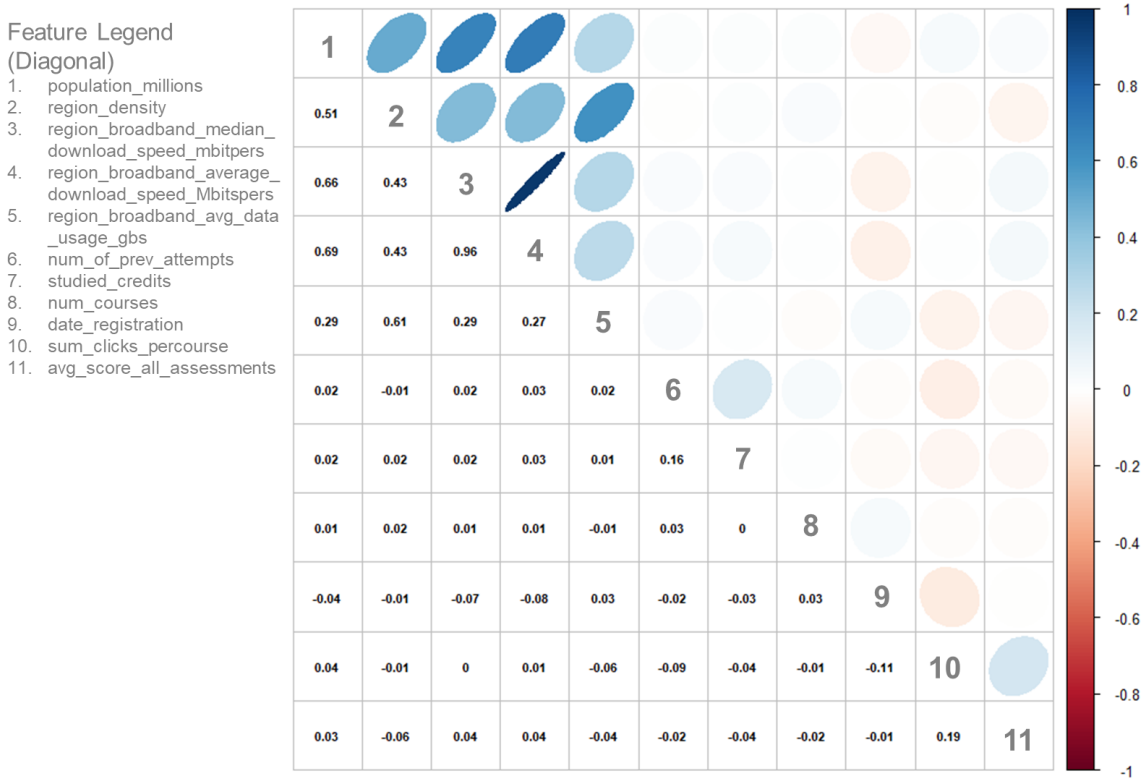
This plot shows there are several observations with excessive clicks. To investigate further we developed a boxplot. The boxplot helps confirm the identification of potential outliers. More formally, we conducted an Upper Fence calculation. Our five number summary consists of Min = 1, Q1 = 221, Median = 498, Q3 = 1067, Max = 15,716. This results in an IQR of 882.8 and an Upper Fence value of 2,431.4. These means that any value above 2,431.4 is considered an outlier. Further analysis shows that values above this value constitute nearly 9% of the data. Despite being outliers, we elected to keep these upper values in the data since the quantity of values at this number are so high. This suggest this is actual student behavior and not an error in recording.



The remainder of the features in the dataset did not require further analysis beyond basic histograms and have been excluded from the body of this report, but are included in ANNEX 5.

Correlation Analysis

The correlation plot below shows the correlation between numeric features in the dataset. The color intensity and shape of circle indicate the strength of the correlation. Also, in the lower portion of the chart, are the actual coefficient values. A legend has been placed to the left to identify the features in the diagonal of the matrix. As seen below, the correlation between region median and average broadband speeds (3 and 4 respectively) are highly correlated as we would suspect. Given median download speed has a more normal distribution, we will use this feature rather than average download speed. Additionally, we see the correlation between 3 and 1 (median download speed and population) are relatively strongly correlated suggesting that higher populated areas also have higher broadband speeds.



Additional Scatterplot Analysis can be found in ANNEX 6.

Modeling

Model 1

Model 1 was designed to help answer our first research question: *"How do demographic, regional socioeconomic and academic behavior factors impact a student's online performance?* To model this we used the stdatInvTrain data set with all features except the following:

- id_student - This variable provides no information and is simply a unique identifier for each student
- region - since Region is such a localized value, we will remove it and use pop density in its place
- final_result - Final result is nearly the same as our dependent variable but in a categorical form
- region_broadband_average_download_speed_Mbitspers - Average and median broadband download speed explain the same thing but given median download speed has a more normal distribution we will use that instead

In total we examined four versions of model 1: Null, Full, a reduce Model 1.2, and a Stepwise model. We began by estimating the Null model which is basically the average of avg_score_all_assessments. The output is below:

```
Call:
lm(formula = avg_score_all_assessments ~ 1, data = stdataInvTrain)

Residuals:
    Min      1Q  Median      3Q     Max
-56.495  -6.095   1.905   8.123  23.505

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  76.4953     0.2354   324.9   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.21 on 2266 degrees of freedom
```

Next we built our Full model. The output of this model is below:

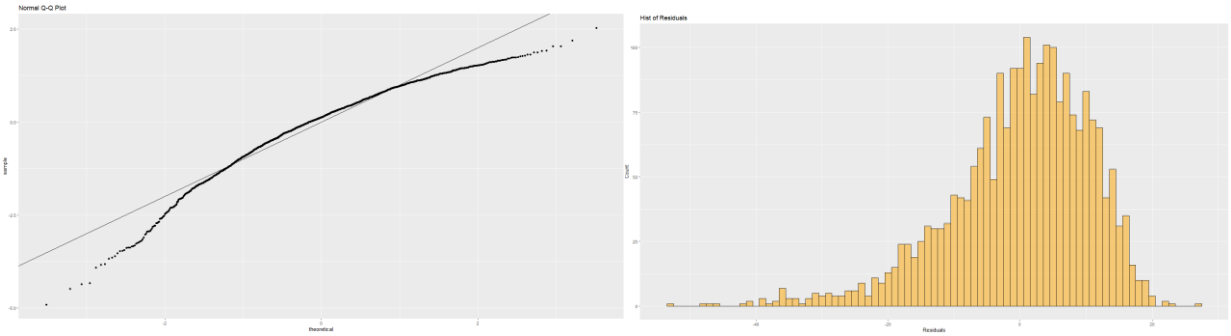
```
Call:
lm(formula = avg_score_all_assessments ~ . - id_student - region -
    final_result - region_broadband_average_download_speed_Mbitspers,
    data = stdataInvTrain)

Residuals:
    Min      1Q  Median      3Q     Max
-52.546  -5.694   1.362   7.544  26.987

Coefficients:
                                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                                       7.500e+01  3.332e+00  22.510  < 2e-16 ***
genderM                                          -3.755e+00  5.792e-01  -6.483 1.10e-10 ***
population_millions                               3.175e-01  1.580e-01   2.010 0.044583 *
region_density                                   -4.667e-04  8.495e-05  -5.494 4.38e-08 ***
region_broadband_median_download_speed_mbitpers   4.421e-02  3.136e-02   1.410 0.158736
region_broadband_avg_data_usage_gbs               1.645e-02  1.008e-02   1.632 0.102725
highest_eduation_us_equivalentHighschool (non-honors)      -2.499e+00  5.019e-01  -4.980 6.85e-07 ***
highest_eduation_us_equivalentLess than Highschool Diploma -9.758e-01  2.577e+00  -0.379 0.704986
highest_eduation_us_equivalentMasters Degree or Higher      1.914e+00  3.875e+00   0.494 0.621341
highest_eduation_us_equivalentSome Undergrad College       -7.911e-01  6.974e-01  -1.134 0.256780
imd_band20-30                                     2.714e+00  9.292e-01   2.921 0.003522 **
imd_band30-40                                     2.028e+00  9.290e-01   2.184 0.029102 *
imd_band40-50                                     3.266e+00  9.209e-01   3.547 0.000398 ***
imd_band50-60                                     4.436e+00  9.583e-01   4.629 3.88e-06 ***
imd_band60-70                                     4.354e+00  9.671e-01   4.502 7.09e-06 ***
imd_band70-80                                     3.024e+00  9.686e-01   3.122 0.001821 **
imd_band80-90                                     4.940e+00  1.020e+00   4.842 1.37e-06 ***
imd_band90-100                                    4.562e+00  1.067e+00   4.277 1.98e-05 ***
age_band35-55                                     7.208e-01  4.894e-01   1.473 0.140931
age_band55 or older                              -3.724e+00  2.249e+00  -1.656 0.097804 .
num_of_prev_attempts                              7.393e-01  4.893e-01   1.511 0.130975
studied_credits                                  -3.823e-03  6.237e-03  -0.613 0.539941
disabilityY                                      -6.769e-01  8.272e-01  -0.818 0.413252
num_courses                                      -5.516e+00  2.366e+00  -2.331 0.019831 *
date_registration                                 9.730e-03  5.289e-03   1.840 0.065972 .
sum_clicks_percourse                              1.194e-03  1.662e-04   7.186 9.07e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.75 on 2241 degrees of freedom
Multiple R-squared:  0.09056, Adjusted R-squared:  0.08041
F-statistic: 8.926 on 25 and 2241 DF,  p-value: < 2.2e-16
```

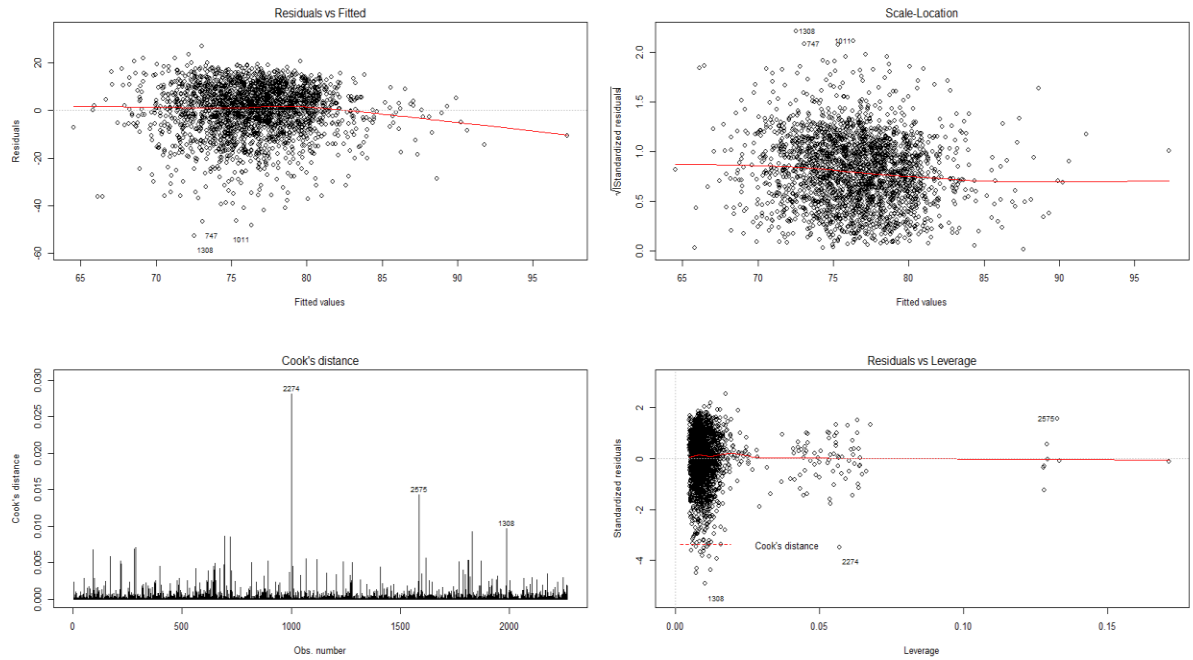
The Full model is significant and has an R^2 values of 0.0905, an Adjusted R^2 of 0.0804, and a standard error of 10.75. From the Residuals section we see they are not centered on zero, but they are not from it. Also, we see that they skew more negative than positive. A better review of the residuals can be seen below in the Normal Q-Q plot.



The Q-Q plot shows relatively normal residuals throughout most of the fits but both tails are challenge slightly in terms of normality. The histogram of the residuals below above right shows the skew of the residuals more clearly. Absent the long left tail, the distribution appears relatively normal.

The graphs show that although not perfect, the residuals are relatively normal. Next we can check the assumption of equal variance (homoscedasticity) and determine if the residuals are spread equally along the ranges of predictors. The Scale location graph in the upper right corner of the figure below looks fair in that the line is roughly horizontal, however there appears to be more residuals, and more dispersion above the line. Finally the last two graphs at the bottom help us identify influential cases. Although it looks as though there may be a few

observations that require additional investigation, it does not appear that any of them are providing excessive leverage on the model.



Finally we examined the possibility of multicollinearity by calculating the Variance Inflation Factor of the coefficients.

	GVIF	Df
gender	1.076490	1
population_millions	2.090064	1
region_density	2.051915	1
region_broadband_median_download_speed_mbitpers	1.925253	1
region_broadband_avg_data_usage_gbs	1.690691	1
highest_eduation_us_equivalent	1.170920	4
imd_band	1.192867	8
age_band	1.236880	2
num_of_prev_attempts	1.065781	1
studied_credits	1.051476	1
disability	1.030967	1
num_courses	1.008222	1
date_registration	1.035585	1
sum_clicks_percourse	1.137220	1

We can see in the readout above that most of the coefficients are fine as the VIF is 1, and none exceed 10. However, as we determined before, we can see that population_millions and region_density both score just over two. This is to be expected as population density is derived from the population. As such, we will remove population and only model population density in Model 1.2.

In model 1.2 we removed several of the features determined to not be significant in the full model, and we also interacted gender with disability to see if there is a difference in the performance of students with disabilities by gender. Model 1.2 can be seen below:

```
Call:
lm(formula = avg_score_all_assessments ~ imd_band + region_broadband_median_download_speed_mbitpers +
    region_broadband_avg_data_usage_gbs + region_density + highest_eduation_us_equivalent +
    age_band + disability + disability * gender + gender + sum_clicks_percourse +
    date_registration, data = stdataInvTrain)

Residuals:
    Min      1Q  Median      3Q     Max
-52.626  -5.736   1.481   7.507  28.896

Coefficients:
                                                        Estimate Std. Error t value Pr(>|t|)
(Intercept)                                            7.013e+01  2.158e+00  32.497  < 2e-16 ***
imd_band20-30                                          2.728e+00  9.298e-01   2.934 0.003380 **
imd_band30-40                                          2.067e+00  9.291e-01   2.224 0.026231 *
imd_band40-50                                          3.234e+00  9.212e-01   3.511 0.000455 ***
imd_band50-60                                          4.422e+00  9.576e-01   4.617 4.10e-06 ***
imd_band60-70                                          4.250e+00  9.668e-01   4.396 1.15e-05 ***
imd_band70-80                                          2.889e+00  9.669e-01   2.987 0.002844 **
imd_band80-90                                          5.019e+00  1.020e+00   4.922 9.17e-07 ***
imd_band90-100                                         4.554e+00  1.066e+00   4.272 2.01e-05 ***
region_broadband_median_download_speed_mbitpers        8.283e-02  2.538e-02   3.264 0.001116 **
region_broadband_avg_data_usage_gbs                    1.523e-02  1.005e-02   1.515 0.129875
region_density                                        -4.207e-04  8.075e-05  -5.209 2.07e-07 ***
highest_eduation_us_equivalentHighschool (non-honors) -2.455e+00  4.989e-01  -4.921 9.25e-07 ***
highest_eduation_us_equivalentLess than Highschool Diploma -7.999e-01  2.574e+00  -0.311 0.755981
highest_eduation_us_equivalentMasters Degree or Higher 2.381e+00  3.871e+00   0.615 0.538624
highest_eduation_us_equivalentSome Undergrad College  -6.907e-01  6.966e-01  -0.991 0.321583
age_band35-55                                          7.460e-01  4.886e-01   1.527 0.126908
age_band55 or older                                   -3.482e+00  2.250e+00  -1.547 0.121904
disabilityY                                           -1.570e+00  8.791e-01  -1.786 0.074302 .
genderM                                               -4.128e+00  5.947e-01  -6.942 5.05e-12 ***
sum_clicks_percourse                                   1.204e-03  1.656e-04   7.270 4.93e-13 ***
date_registration                                      9.082e-03  5.285e-03   1.719 0.085824 .
disabilityY:genderM                                    6.615e+00  2.516e+00   2.630 0.008605 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.75 on 2244 degrees of freedom
Multiple R-squared:  0.08852, Adjusted R-squared:  0.07959
F-statistic: 9.906 on 22 and 2244 DF,  p-value: < 2.2e-16
```

Model 1.2 is significant and has an R^2 values of 0.0885, and Adjusted R^2 of 0.0795, and a standard error of 10.75. From the Residuals section we see they are pretty similar to the Full model, but are slightly more distributed. Interestingly though, it turns out our interaction term of gender and disability is significant.

Finally, we also execute a Stepwise procedure that iterated on models in both directions between the Null and the Full model. Before executing, we updated the full model by removing population_millions as a result of our VIF analysis, and added the disability*gender interaction term based on what we learned in model 1.2 into. After 12 steps, the Stepwise procedure offers the following model as the best model.

```
Call:
lm(formula = avg_score_all_assessments ~ sum_clicks_percourse +
    gender + imd_band + highest_eduation_us_equivalent + region_density +
    population_millions + num_courses + age_band + region_broadband_avg_data_usage_gbs +
    date_registration + num_of_prev_attempts, data = stdataInvTrain)

Residuals:
    Min      1Q  Median      3Q     Max
-52.360  -5.663   1.346   7.585  27.092

Coefficients:
                                                    Estimate Std. Error t value Pr(>|t|)
(Intercept)                                         7.469e+01  3.275e+00  22.806  < 2e-16 ***
sum_clicks_percourse                                1.193e-03  1.661e-04   7.182 9.29e-13 ***
genderM                                            -3.747e+00  5.785e-01  -6.477 1.15e-10 ***
imd_band20-30                                       2.681e+00  9.289e-01   2.886 0.003934 **
imd_band30-40                                       2.072e+00  9.285e-01   2.232 0.025744 *
imd_band40-50                                       3.293e+00  9.207e-01   3.577 0.000355 ***
imd_band50-60                                       4.511e+00  9.570e-01   4.714 2.57e-06 ***
imd_band60-70                                       4.448e+00  9.655e-01   4.607 4.31e-06 ***
imd_band70-80                                       3.145e+00  9.658e-01   3.256 0.001146 **
imd_band80-90                                       5.052e+00  1.018e+00   4.962 7.50e-07 ***
imd_band90-100                                      4.748e+00  1.061e+00   4.476 8.00e-06 ***
highest_eduation_us_equivalentHighschool (non-honors) -2.502e+00  5.007e-01  -4.998 6.25e-07 ***
highest_eduation_us_equivalentLess than Highschool Diploma -9.271e-01  2.575e+00  -0.360 0.718906
highest_eduation_us_equivalentMasters Degree or Higher  1.706e+00  3.870e+00   0.441 0.659284
highest_eduation_us_equivalentSome Undergrad College -8.331e-01  6.952e-01  -1.198 0.230923
region_density                                     -4.589e-04  8.464e-05  -5.421 6.54e-08 ***
population_millions                                 4.489e-01  1.279e-01   3.510 0.000458 ***
num_courses                                        -5.489e+00  2.364e+00  -2.322 0.020307 *
age_band35-55                                       7.080e-01  4.877e-01   1.451 0.146787
age_band55 or older                                -3.687e+00  2.248e+00  -1.640 0.101114
region_broadband_avg_data_usage_gbs                 1.798e-02  1.003e-02   1.793 0.073145 .
date_registration                                   9.431e-03  5.279e-03   1.787 0.074116 .
num_of_prev_attempts                                6.863e-01  4.827e-01   1.422 0.155176
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.75 on 2244 degrees of freedom
Multiple R-squared:  0.08934, Adjusted R-squared:  0.08041
F-statistic: 10.01 on 22 and 2244 DF,  p-value: < 2.2e-16
```

The Step model is significant and has an R^2 values of 0.0893, and Adjusted R^2 of 0.0804, and a standard error of 10.75. From the Residuals section we can see they are similar to the other models. Based on adjusted R^2 values, the Full and Step model are the lead models. However, to arrive at our final model we cross validated using our test data and compare each model's PMSE.

Model 1 Results

As a result of our cross validation, we report our model 1.2 as our final model for research question 1. Although it has a slightly lower R^2 and Adjusted R^2, it has a the lowest PMSE and less coefficients than the full model making it slightly more parsimonious. Below is a performance table summarizing the outcomes of each model.

Model	R^2	AdJ R^2	MSE	PMSE	Ratio Rule PMSE/MSE < 2, model is fine	# Coefficients
Full	0.09056	0.08041	115.4	134.9982	1.1698	13
M1.2	0.08852	0.07959	115.6	134.5639	1.1640	11
Step	0.08934	0.08041	115.7	135.3245	1.1696	11

Confidence Intervals

The confidence intervals for the estimates in our final model are below:

	2.5 %	97.5 %
(Intercept)	65.8997494802	74.3640361344
imd_band20-30	0.9046952857	4.5512721569
imd_band30-40	0.2445586152	3.8885147109
imd_band40-50	1.4277742693	5.0406879077
imd_band50-60	2.5438655791	6.2996479532
imd_band60-70	2.3542169586	6.1458940668
imd_band70-80	0.9924609685	4.7846756100
imd_band80-90	3.0195034197	7.0186076001
imd_band90-100	2.4634736727	6.6435341143
region_broadband_median_download_speed_mbitpers	0.0330625153	0.1326051639
region_broadband_avg_data_usage_gbs	-0.0044808505	0.0349344562
region_density	-0.0005790010	-0.0002622959
highest_eduation_us_equivalentHighschool (non-honors)	-3.4334672295	-1.4766579296
highest_eduation_us_equivalentLess than Highschool Diploma	-5.8471181040	4.2472684044
highest_eduation_us_equivalentMasters Degree or Higher	-5.2103029880	9.9713888301
highest_eduation_us_equivalentSome Undergrad College	-2.0568083531	0.6754605279
age_band35-55	-0.2120568475	1.7040937201
age_band55 or older	-7.8940857450	0.9306285750
disabilityY	-3.2936525660	0.1542324323
genderM	-5.2943845724	-2.9620088123
sum_clicks_percourse	0.0008793684	0.0015289706
date_registration	-0.0012810710	0.0194455506
disabilityY:genderM	1.6821067319	11.5485829574

As seen above, there are several estimates that have very small coefficients which means that regardless of their significance, they have little influence on the independent variable. The table below offers a brief interpretation of what each one means when all other variables are held constant:

Model 1 Interpretation

Estimate	Coeff	Interpretation
70.13189	Intercept	The average score for a Female, with no disability, age band 0-35, with highschool honors as highest education level, from a region with an IMD_band of 0 to 10
2.72798	imd_band20-30	All else being equal, the estimated increase to the average score if the student's region is imd_band 20-30
2.06654	imd_band30-40	All else being equal, the estimated increase to the average score if the student's region is imd_band 30-40
3.23423	imd_band40-50	All else being equal, the estimated increase to the average score if the student's region is imd_band 40-50
4.42176	imd_band50-60	All else being equal, the estimated increase to the average score if the student's region is imd_band 50-60
4.25006	imd_band60-70	All else being equal, the estimated increase to the average score if the student's region is imd_band 60-70
2.88857	imd_band70-80	All else being equal, the estimated increase to the average score if the student's region is imd_band 70-80
5.01906	imd_band80-90	All else being equal, the estimated increase to the average score if the student's region is imd_band 80-90
4.55350	imd_band90-100	All else being equal, the estimated increase to the average score if the student's region is imd_band 90-100
0.08283	region _broadband_median_download_speed_mbitpers	All else being equal, the estimated increase to the average score for every 1 unit increase region _broadband_median_download_speed_mbitpers
0.01523	region_broadband_avg_data_usage_gbs	All else being equal, the estimated increase to the average score for every 1 unit increase in region_broadband_avg_data_usage_gbs
-0.00042	region_density	All else being equal, the estimated decrease to the average score for every 1 unit increase in region _broadband_median_download_speed_mbitpers
-2.45506	highest_eduation_us_equivalentHighschool (non-honors)	All else being equal, the estimated decrease to the average score if the student's highest_eduation_us_equivalentHighschool (non-honors)
-0.79992	highest_eduation_us_equivalentLess than Highschool Diploma	All else being equal, the estimated decrease to the average score if the student's highest_eduation_us_equivalentLess than Highschool Diploma
2.38054	highest_eduation_us_equivalentMasters Degree or Higher	All else being equal, the estimated increase to the average score highest_eduation_us_equivalentMasters Degree or Higher
-0.69067	highest_education_us_equivalentSome Undergrad College	All else being equal, the estimated decrease to the average score if the student's highest_education_us_equivalentSome Undergrad College
0.74602	age_band35-55	All else being equal, the estimated increase to the average score if the student's age_band35-55
-3.48173	age_band55 or older	All else being equal, the estimated decrease to the average score age_band55 or older
-1.56971	disabilityY	All else being equal, the estimated decrease to the average score if the student's has a disability
-4.12820	genderM	All else being equal, the estimated decrease to the average score if the student's his male
0.00120	sum_clicks_percourse	All else being equal, the estimated increase to the average score for every 1 unit increase sum_clicks_percourse
0.00908	date_registration	All else being equal, the estimated change to the average score for each additional day registered early (negative input) or late(positive input) relative to the start of the module-presentation
6.61534	disabilityY:genderM	All else being equal, the estimated amount more that disability will decrease the average score if the students is male and has a disability, in addition to the main effects of male and disability

Model 1 Key Findings

The key findings below are conditioned on the data from this sample, and each finding requires holding all other variables constant.

- Being from a wealthier region has advantages.
  - As suspected, students from wealthier regions (IMD Bands from 80 – 100) generally have better academic outcomes and can expect to earn on average nearly an additional 5 points on their average score across all assessments.

- Access to broadband makes a difference.
  - For every one unit increase in median broadband speed of a region, those student can expect to increase academic outcomes by 0.08 points. Using the min and max values in this dataset means that a student from a region with a median of 50 Mbps (max) compared to similar student from a region with a median of 17.5 Mbps (min), can expect to earn on average 2.6 additional points on their average score across all assessments.

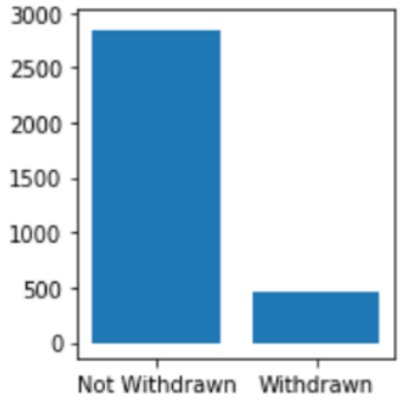
- Gender matters.
  - Also, Males have lower academic outcomes

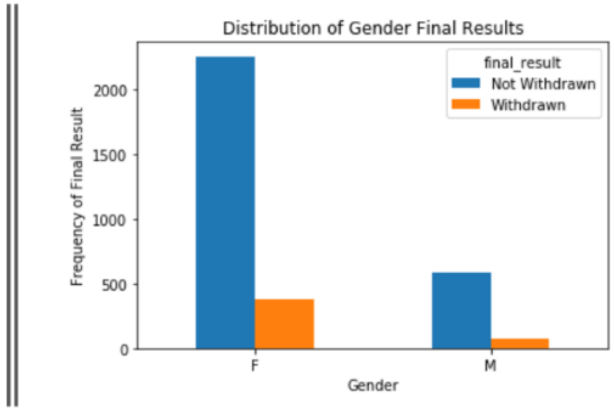
- On average, middle age students (35- 55) perform better than their younger and older counterparts, with older student (56 plus) have the worse of the three age groups.

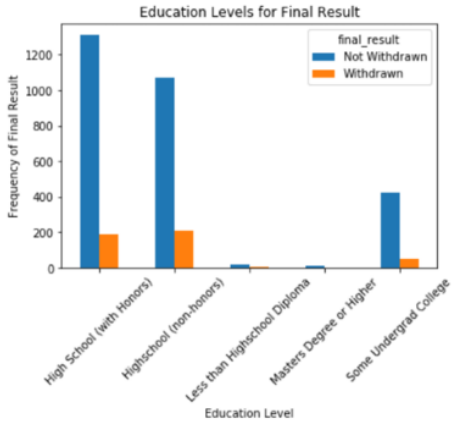
- Interestingly, students with master's degree typically achieve better academic outcomes but students with high school diplomas (with Honors) perform better than students with some undergrad college, and not surprisingly, students with less than a high school diploma.

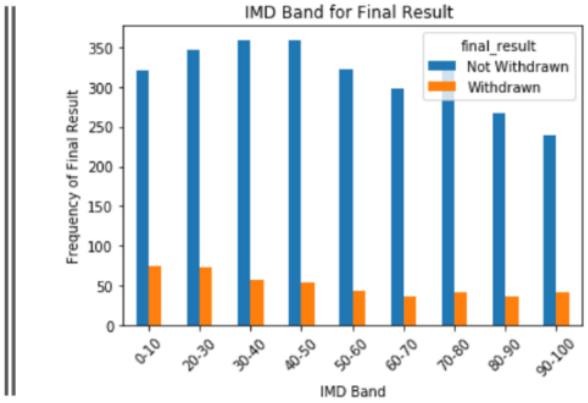
Model 2

Model 2 was designed to help answer our second research question: *"How do demographic, regional socioeconomic and academic behavior factors impact a student's likelihood of withdrawing of not withdrawing from an academic program?* This model built off the primary dataset but unlike the data used for model 1 (stdataInv), this data set did not partition filter out student's with a "withdrawn" final result. This dataset used this status to create a new column that would serve as the target feature in the logistic regression model. Each student was assigned a class: "withdrawn" if they withdrew, and "Not Withdrawn" if their final_result status was "Distinction", "Pass" or "Fail." This dataset is titled "stDataWithdrawn" which includes 3,291 observations, then split as 80% train and 20% test. A visualization of the target feature breakdown can be seen below:









We noticed our dataset was imbalanced being that there were approximately 2,800 students categorized as not withdrawn and 500 categorized as withdrawn.

We knew with our regression model, solving using imbalanced data would make our model very susceptible to overfitting for the majority class causing our test error to be high. We used a method of over-sampling the minority class called Synthetic minority oversampling technique (SMOTE). This method works by creating synthetic samples from the minor class (Not Withdrawn) instead of creating copies. Randomly choosing one of the k-nearest-neighbors and using it to create a similar, but randomly tweaked, new observations. From here we had an even proportion for our prediction class.

We went on to run a stepwise regression of this data to calculate the r-squared value, used python machine learning module called scikit-learn to find other meaningful results like accuracy, f-1 score, precision, recall, ROC curve and confusion matrix.

Initial Model Evaluation

In our initial binary logistic regression on withdrawn/not withdrawn, we implemented a stepwise model using Minitab. The features that were used in the model were region broadband median download speed, region broadband average data usage, number of previous attempts, studied credits, number of courses, sum clicks per course, gender, highest education completed, imd band, age band, and disability. Although all of these variables were included in the model, many of their p-values deemed the variables insignificant. This model achieved a low R-squared value of 41.74%.

Feature Selection and Model Refinement

After the first model achieved a low R-squared value in Minitab, we decided to switch to python for our next logistic model to try and improve upon the model's performance. The python model included region density, number of previous attempts, sum of clicks per course, gender, highest education completed, imd band, and age. We were able to reduce the quantity of predictors by removing those that Minitab labeled as insignificant. As a result, the model achieved an accuracy of 82%.

Report Final Model

Variable	Coefficient	Interpretation
Region Broadband Median Download Speed	-0.000542	For each additional MB/sec increase in median download speed there is an expected change in log odds of -0.000542
Region Broadband Average Data Usage	-0.007508	For each additional MB increase in average data usage there is an expected change in log odds of –0.0075
Number of Previous Attempts	-1.033509	For each additional attempt there is an expected change in log odds of –1.034
Number of Courses	2.844610	For each additional course there is an expected change in log odds of 2.845
Sum Clicks per Course	-0.000142	For each additional click there is an expected change in log odds of -0.000142
Gender (M=1)	-2.069016	If the student is male, there is an expected change in log odds of -2.069
Highest Education (Some Undergrad College)	-2.166684	For a student that has completed some college there will be an expected change in log odds of –2.167
Highest Education (high school non-honors)	-0.292734	For a student that has completed some college there will be an expected change in log odds of –0.293
IMD-Band (20-30)	-2.089706	Students in this imd band will have an expected change of 2.089 lower than students in the 0-10 band
IMD-Band (30-40)	-2.003591	Students in this imd band will have an expected change of 2.004 lower than students in the 0-10 band
IMD-Band(40-50)	-1.906014	Students in this imd band will have an expected change of 1.906 lower than students in the 0-10 band
IMD-Band (50-60)	-1.905764	Students in this imd band will have an expected change of 1.906 lower than students in the 0-10 band
IMD-Band (60-70)	-2.250452	Students in this imd band will have an expected change of 2.250 lower than students in the 0-10 band
IMD-Band (70-80)	-1.964996	Students in this imd band will have an expected change of 1.965 lower than students in the 0-10 band
IMD-Band (80-90)	-1.905181	Students in this imd band will have an expected change of 1.905 lower than students in the 0-10 band
IMD-Band (90-100)	-1.181011	Students in this imd band will have an expected change of 1.181 lower than students in the 0-10 band
Age_Band_35-55	-0.698614	Students in this age band will have an expected change of 0.699 lower than students in the other bands
Disability (Yes=1)	-0.957344	Students with a disability will have an expected change in log odds that is 0.957 lower than those without disabilities
Studied Credits	0.003859	For each additional credit a student took the expected change in log odds increases by 0.003
Intercept	0.16868674	All students have initial log odds of 0.169

Model 2 Results

Our initial stepwise regression yielded an r-squared value of 41.74 percent. Using more advanced techniques, we were able to evaluate our model performance. The first was using accuracy of the logistic regression classifier. Accuracy of logistic regression classifier on test set: 0.82. You need to understand that r-squared is a measure of explanatory power, not fit. You can generate lots of data with low r-squared because we don't expect models (especially in social or behavioral sciences) to include all the relevant predictors to explain an outcome variable.

The next goal was to make a confusion matrix to see how well our model predicted each output.

Confusion Matrix:

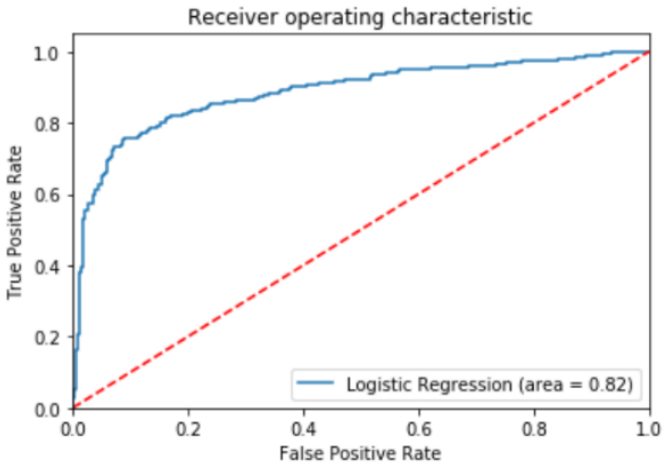
```
[[392  71]
 [ 90 349]]
```

Using scikit-learn we can also look at the classification report to further understand the results displayed by the confusion matrix.

	precision	recall	f1-score
0	0.81	0.85	0.83
1	0.83	0.79	0.81
accuracy			0.82
macro avg	0.82	0.82	0.82
weighted avg	0.82	0.82	0.82

Model 1 Interpretation

Of the entire test set, we can predict with 82% accuracy students who will withdraw. The last important evaluation criterion was using the ROC curve which measures the true positive cases over that false positive. This is similar to the classifier for logistic regression, and the interpretation is the same.



CONCLUDING REMARKS

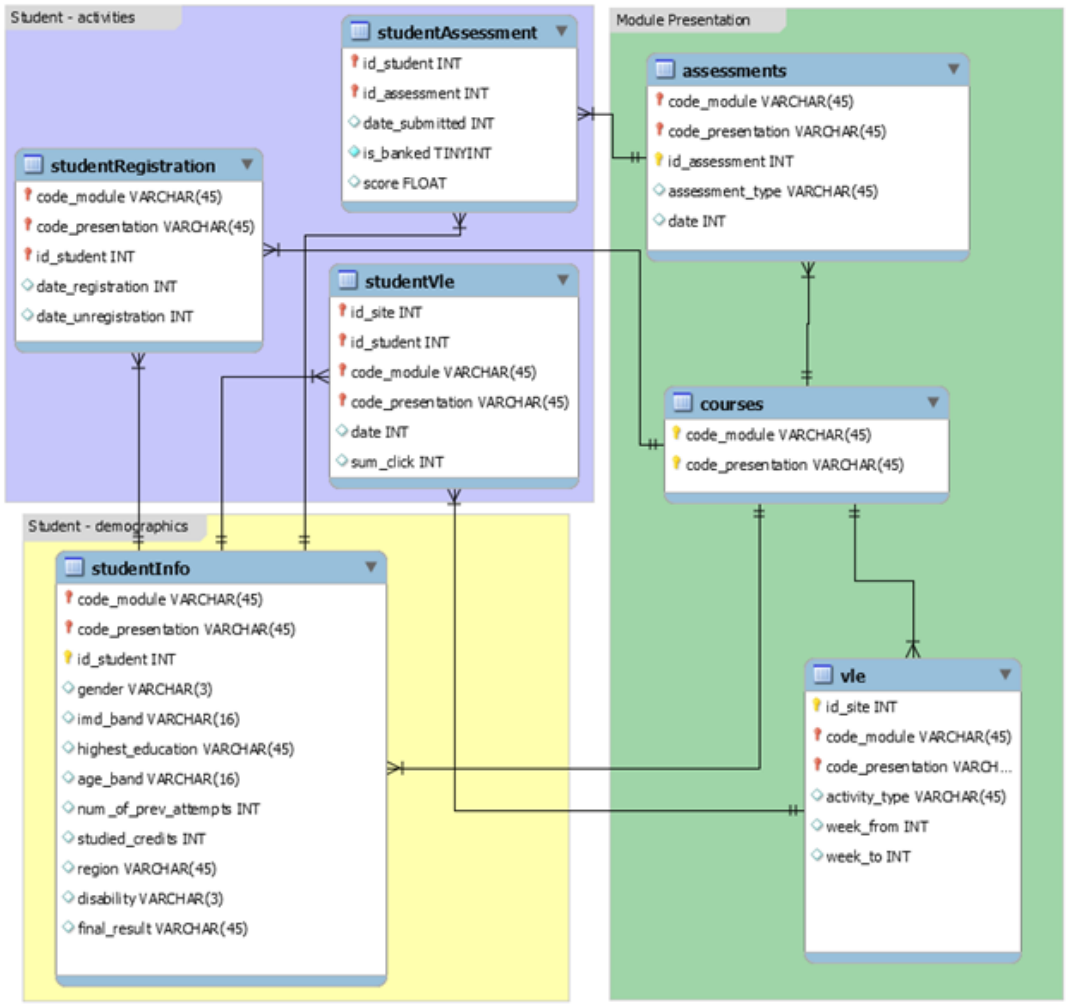
We can conclude that students who participate in online classes can be successful at fairly high rate and our analysis has shown that a student's demographics, their socioeconomic status and their academic behavior play a role in determining academic outcomes. Our findings show that being from a wealthier region has advantages, and how access to broadband matters. We've demonstrated that number of courses is a lead indicators in helping to predict if student will withdraw from a program before completing. This findings are novel, and given the current state of affairs with the worlds migration to online learning, these findings are also timely and relevant.

However our research is not without limitations. One limitation is that the sample we studied consists of students that self-selected to participate in online education. This self-selection may bias the results of our findings in that theses students are already pre-dispositioned to be more successful since this took this endeavor on willingly. This is very different than the current situation where many students many students around the world were forced into online learning. Another limitation is that this analysis did not model the effectiveness of the teachers. An effective learning environment requires both student and teachers that are both interested in positive outcomes. This analysis was a step in the right direction but much more could be done as the limitations stated above hinder the amount of inference this analysis provides.

However, with much of the world now conducting online study as a result of COVID-19, so many more datasets should become available for further research where students that did not elect to study online can be examined, or where students of younger age groups can be studied. With the future of online learning continuing to grow, researching this topic remains a worthy academic endeavor. One area of additional study would be to compare the scores/grades and the withdrawal rates prior to the global switch from face-to-face to online once the world resumes campus classrooms. This would help better assess the effectiveness of online educations.

ANNEXES

ANNEX 1: Entity Relationship Diagram



ANNEX 2: Data Dictionary Original File

Student Table

This file contains demographic information about the students together with their results. File contains the following columns:

- code_module – an identification code for a module on which the student is registered.
- code_presentation - the identification code of the presentation during which the student is registered on the module.
- id_student – a unique identification number for the student.
- gender – the student's gender.
- region – identifies the geographic region, where the student lived while taking the module-presentation.
- highest_education – highest student education level on entry to the module presentation.
- imd_band – specifies the Index of Multiple Depravation Band of each location the student lived during the module-presentation (This is a composite value based on the Income, Employment, Health deprivation and Disability, Education Skills and Training, Barriers to Housing and Services, Crime, and Living Environment of each location.
- age_band – band of the student's age.
- num_of_prev_attempts – the number times the student has attempted this module.
- studied_credits – the total number of credits for the modules the student is currently studying.
- disability – indicates whether the student has declared a disability.
- final_result – student's final result in the module-presentation.

Course Table
File contains the list of all available modules and their presentations. The columns are:

- code_module – code name of the module, which serves as the identifier.
- length - length of the module-presentation in days.

Assessment Table
This file contains information about assessments in module-presentations. Usually, every presentation has a number of assessments followed by the final exam. CSV contains columns:

- code_module – identification code of the module, to which the assessment belongs.
- code_presentation - identification code of the presentation, to which the assessment belongs.
- id_assessment – identification number of the assessment.
- assessment_type – type of assessment. Three types of assessments exist: Tutor Marked Assessment (TMA), Computer Marked Assessment (CMA) and Final Exam (Exam).
- date – information about the final submission date of the assessment calculated as the number of days since the start of the module-presentation. The starting date of the presentation has number 0 (zero).
- weight - weight of the assessment in %. Typically, Exams are treated separately and have the weight 100%; the sum of all other assessments is 100%.

Virtual Learning Environment (VLE)

This file contains information about the available materials in the VLE. Typically these are html pages, pdf files, etc. Students have access to these materials online and their interactions with the materials are recorded. The vle.csv file contains the following columns:

- id_site – an identification number of the material.
- code_module – an identification code for module.
- code_presentation - the identification code of presentation.
- activity_type – the role associated with the module material.
- week_from – the week from which the material is planned to be used.
- week_to – week until which the material is planned to be used.

Student Registration Table

This file contains information about the time when the student registered for the module presentation. For students who unregistered the date of unregistration is also recorded. File contains five columns:

- code_module – an identification code for a module.
- code_presentation - the identification code of the presentation.
- id_student – a unique identification number for the student.
- date_registration – the date of student's registration on the module presentation, this is the number of days measured relative to the start of the module-presentation (e.g. the negative value -30 means that the student registered to module presentation 30 days before it started).
- date_unregistration – date of student unregistration from the module presentation, this is the number of days measured relative to the start of the module-presentation. Students, who completed the course have this field empty. Students who unregistered have Withdrawal as the value of the final_result column in the studentInfo.csv file.

Student Assessment Table

This file contains the results of students' assessments. If the student does not submit the assessment, no result is recorded. The final exam submissions is missing, if the result of the assessments is not stored in the system. This file contains the following columns:

- id_assessment – the identification number of the assessment.
- id_student – a unique identification number for the student.
- date_submitted – the date of student submission, measured as the number of days since the start of the module presentation.
- is_banked – a status flag indicating that the assessment result has been transferred from a previous presentation.
- score – the student's score in this assessment. The range is from 0 to 100. The score lower than 40 is interpreted as Fail. The marks are in the range from 0 to 100.

Student VLE Table

This file contains information about each student's interactions with the materials in the VLE. This file contains the following columns:

- code_module – an identification code for a module.
- code_presentation - the identification code of the module presentation.
- id_student – a unique identification number for the student.
- id_site - an identification number for the VLE material.
- date – the date of student's interaction with the material measured as the number of days since the start of the module-presentation.
- sum_click – the number of times a student interacts with the material in that day.

ANNEX 3 : Regional Characteristics

Feature	Levels / Values	Consists of	Population (Millions)	Area (sq mi)	Density (Pop / Area)
Region	East Anglian	Norfolk, Suffolk, Cambridgeshire	6.235	7382	844.62
	East Midlands	East Midlands	4.811	6034	797.32
	Ireland	Ireland	4.904	27133	180.74
	London	London	8.908	607	14675.45
	North	North East, North West, Yorkshire	14.933	14414	1036.01
	North West	North West	7.052	5469	1289.45
	Scotland	Scotland	5.424	30090	180.26
	South	South West England, South East England, London, East of England	27.945	23955	1166.56
	South East	South East	8.635	7373	1171.17
	South West	South West	5.289	9200	574.89
	Wales	Wales	3.139	8023	391.25
	West Midlands	West Midlands	5.713	5000	1142.60
	Yorkshire	Yorkshire	5.300	4531	1169.72

ANNEX 4 : R packages used in the analysis

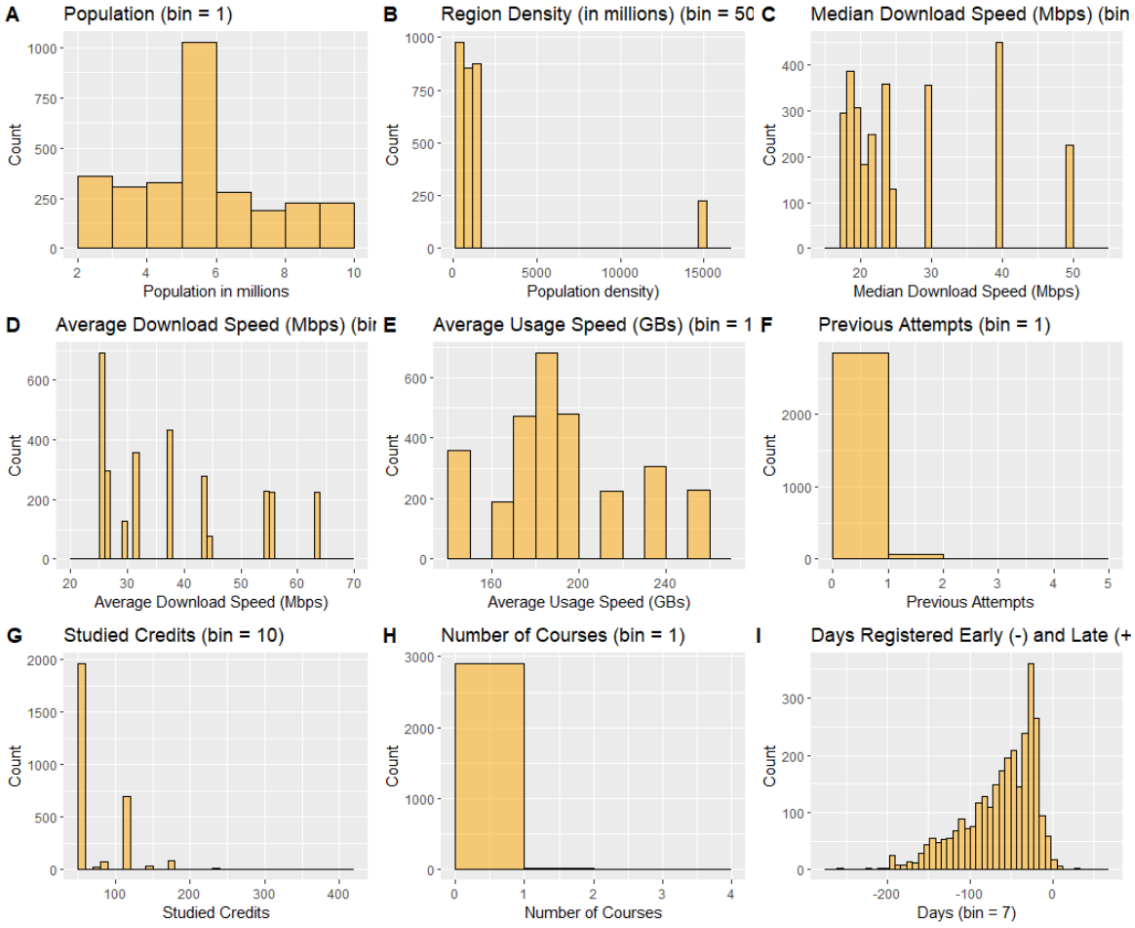
```
library(data.table)
library(tidyverse)
library(psych)
library(readxl)
library(car)
library(ggplot2)
library(mice)
library(VIM)
library(ggpubr)
library(gpairs)
library(corrplot)
library(coefplot)
```

Python Packages:

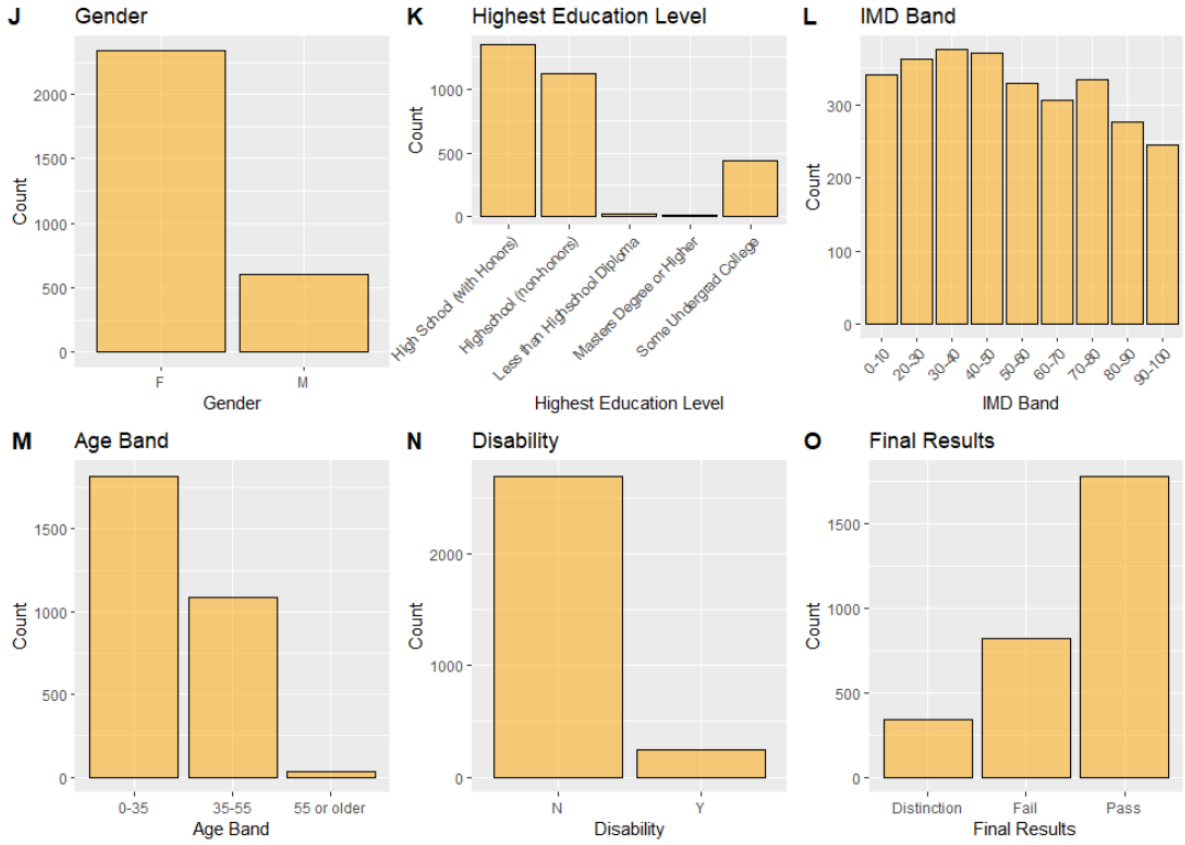
```
skLearn
matplotlib
pandas
numpy
imblearn
```

ANNEX 5: Supporting Visualizations

Numeric Features:

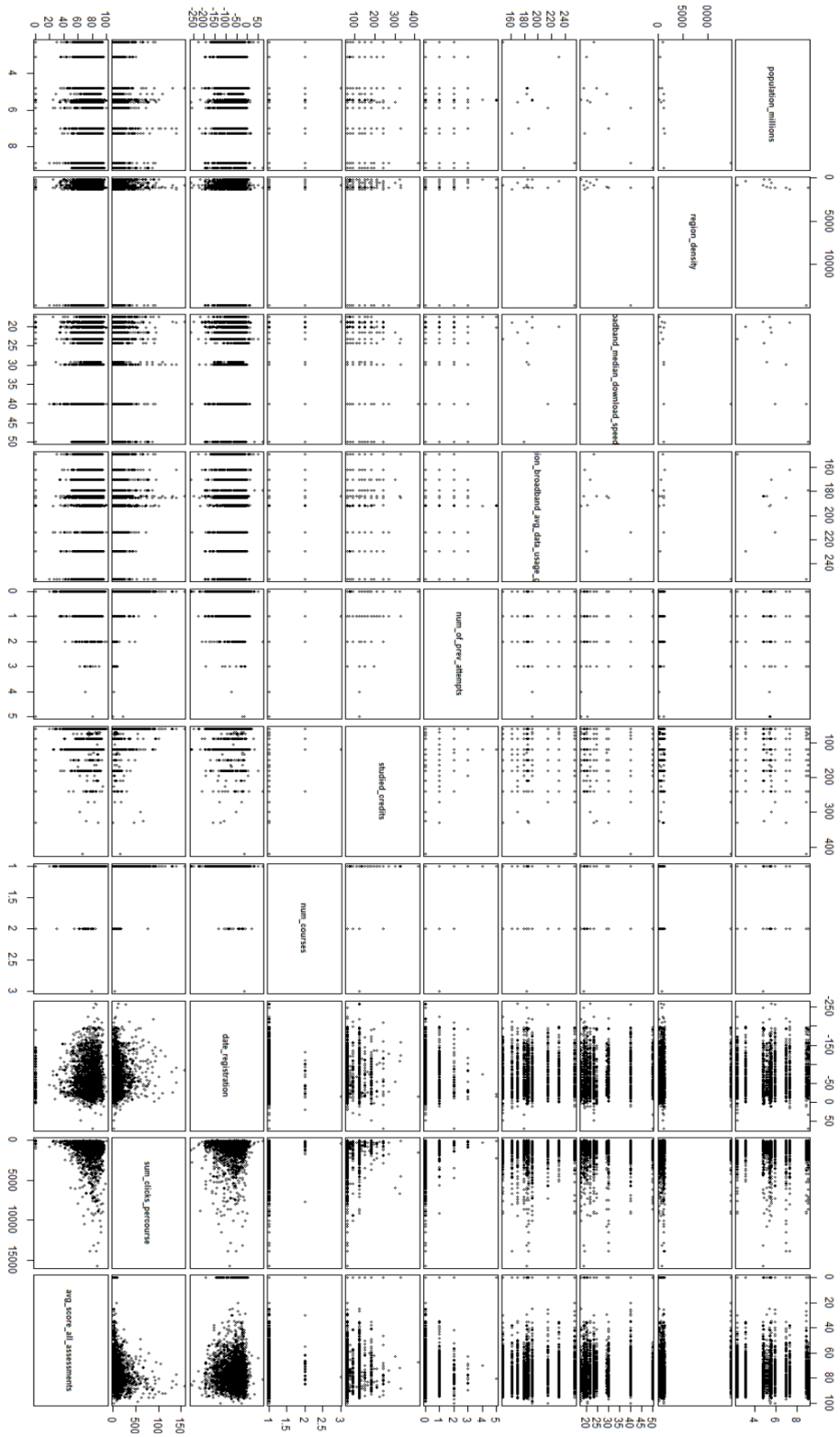


Categorical Features:



ANNEX 6: Scatterplot Analysis

Although not very informative at an aggregated level in this written output, the scatterplot matrix below shows there are no problematic relationships between the features.



Endnotes

[i] US Dept Education: Evaluation of Evidence-Based Practices in Online Learning
https://www2.ed.gov/rschstat/eval/tech/evidence-based-practices/finalreport.pdf

[ii] "Learning on Demand" by I.E Elaine Allen and Jeff Seaman
https://files.eric.ed.gov/fulltext/ED529931.pdf

[iii] Socioeconomic Reginal Bands and Metrics
https://www.gov.uk/government/statistics/english-indices-of-deprivation-2015

[iv] Connected Nations report on regional broadband statistics
https://www.ofcom.org.uk/research-and-data/multi-sector-research/infrastructure-research/connected-nations-2017/data-downloads