Springboard Capstone Project for Introduction to Data Science: An Analysis of Online Enrollments at Small Colleges and Universities in the United States

Author: Daniel Teodorescu

Mentored by: Branko Kovac

1. The Problem

Online program offerings can bring in new students to higher education institutions that are trying to expand their enrollments. The addition of online programs benefits not only higher education institutions, but also nontraditional students since the flexibility of these programs allows students who have full-time jobs and/or family responsibilities to further their education. Anecdotal evidence shows that for small colleges and universities in the United States, adding online programs can boost enrollments and add new sources of revenues. Yet, this relationship has not been tested using large national datasets. Therefore, one goal of this project is to establish statistically whether the addition of online classes is associated with an increase in total enrollments at small colleges and universities, after controlling for relevant institutional characteristics.

Secondly, if the multiple linear regression analysis shows that there is a significant relationship between total enrollment and the offering of online classes, it is hypothesized that the addition of new online programs or courses can also improve the enrollment at small Historically Black Colleges and Universities (HBCUs), making these institutions more attractive to non-traditional students.

Therefore, a secondary goal of this project is to assess whether HBCUs are less likely than the rest of the higher education institutions in the U.S. to offer online classes or programs, after controlling for key institutional characteristics such as institution's size and location. The percentage of students enrolled in fully online programs will be compared between HBCUs and non HBCUs. Similar comparisons will be made for students enrolled in some online classes. Later in the analysis, a logistic regression model will test whether the HBCUs are less likely to offer online classes, after controlling for relevant institutional characteristics.

2. The Datasets

Two datasets were used for the purpose of this project:

1. Fall 2017 Distance Enrollment Data contains 22,590 observations representing enrollment in distance education courses at colleges and universities in the U.S. There are multiple observations per institution, representing three levels: all students, undergraduate students only, and graduate students only. The data was downloaded the IPEDS Fall Enrollment Survey collected by the National Center for Education Statistics (https://nces.ed.gov/ipeds). The following are the key variables that will be used from this data set:
    - Total number of students enrolled
    - Number of students enrolled exclusively in online courses
    - Number of students enrolled in some online classes
    - Number of students from the same state
    - Number of students from out of state
    - Number of international students
2. 2017 Institutional Characteristics Data contains variables that describe each higher education. There are 7,153 observations in the dataset, representing all higher education institutions in the U.S. and its territories. The following are key variables that will be used in the analysis:

Variable name	Variable Description	Variable Type
perc100online	% students who are enrolled only in courses that are considered online education courses.	Continuous
perc_some_online	% students who are enrolled in at least one course that is considered an online education course, but are not enrolled exclusively in online education courses.	Continuous
perc_out	% of total enrollment who are online and out of state	Continuous

perc_int	% of total enrollments who are online and international	Continuous
EFDETOT	Total students enrolled for credit during the fall	Continuous
HBCU	Indicates whether the institution is one of the Historically Black College or University (HBCU) institutions.	Categorical 1=HBCU 2=non-HBCU
HOSPITAL	Indicates whether the institution has hospital.	Categorical 1=has hospital 2= does not have hospital
MEDICAL	Indicates whether the institution offers a medical degree.	Categorical 1=offers medical degree 2=no medical degree
LOCALE	Locale identifies the geographic status of a school on an urban continuum ranging from "large city" to "rural."	Categorical 1=city 2= suburb 3=town 4=rural
GROFFER	Indicates whether the institution offers graduate programs	Categorical 1=offers graduate programs 2=no graduate programs
UGOFFER	Indicates whether the institution offers undergraduate programs	Categorical 1=offers undergraduate programs 2=no undergraduate programs
CONTROL	Indicates whether an institution is operated by publicly elected or appointed officials or by privately elected or appointed officials and derives its major source of funds from private sources.	Categorical 1=Public 2=Private, not-for-profit
HLOFFER	Highest level of offering	Categorical 1 =Postsecondary award, certificate or diploma of less than one academic year 2 = Postsecondary award, certificate or diploma of at least one but less than two academic years 3 = Associate's degree 4 = Postsecondary award, certificate or diploma of at

		least two but less than four academic years 5 = Bachelor's degree 6 =Postbaccalaureate certificate 7 =Master's degree 8 =Post-master's certificate 9 =Doctor's degree

3. Data Wrangling and Cleaning

Since the datasets were created by the National Center for Education Statistics, they were relatively comprehensive and verified. Therefore, they did not require any major transformations.

However, there were a number of variables that will not be required for the analysis, along with variables that contained NA values. Formatting was required for the variable HBCU to transform it from numeric to factor. Data wrangling also involved the merging of two datasets by UNITID. The main libraries required for the data wrangling and cleaning were acquired when the relevant data set was imported.

```
library(tidyr)
library(dplyr)

library(readxl)
hd2017 <- read_excel("CapstoneProject/hd2017.xlsx")
head(hd2017)
ef2017a_dist <- read_excel("CapstoneProject/ef2017a_dist.xlsx")

merged2017 <- merge(ef2017a_dist, hd2017, by="UNITID")
```

The following shows the steps undertaken to transform the data into a desirable format suitable for the analysis.

3.1. Subseting the Merged Dataset

The Fall 2017 merged data set contains enrollment information for all types of institutions, including public, private for profit and private not for profit. Since many of the for-profit private institutions are focused primarily on online education, they are excluded from the analysis. In addition, enrollment data are presented for all students as well as separately for undergraduates and graduate students. We are interested in the examining enrollments only for all students in aggregate.

```
#Select all students (undergraduate and graduate) and exclude for profit colleges and
universities (CONTROL=3)

merged_all_levels <- subset(merged2017, EFDELEV == "1" & CONTROL < 3 )
```

3.2. Calculating New Variables

None of the dependent variables of interest for this analysis were included in the Fall 2017 Distance Enrollment data set. Therefore, the next task was to calculate a series of variables expressed as percentages. These new variables were appended to the merged data set. In the Exploratory Data Analysis phase, these variables will be compared between HBCU and non-HBCUs and will also be dependent variables included in the multiple regression models.

```
#Calculate: a) Percent of students who are enrolled exclusively in online programs
and b) percent of students taking some online classes;

perc100online <- EFDEEXC/EFDETOT
perc_some_online <- EFDESOM/EFDETOT

#Calculate: c) Percent of online students from out of state, d) percent of online
students from outside US
perc_out <- (EFDEEX2+ EFDEEX3)/EFDETOT
perc_int<- EFDEEX4/EFDETOT

data_new <- cbind(merged_all_levels,perc100online, perc_some_online, perc_out,
perc_int)
```

3.3. Deleting Unnecessary Variables

The resulting merged data set contains 3893 observations (one observation per institution) and 91 variables. The variables that were not needed for the analysis were deleted from the dataset, yielding a revised data set with 3893 observations and 17 variables.

```
data_new <- select(data_new, INSTNM, UNITID, STABBR, CONTROL, LANDGRNT, EFDETOT,
HBCU, LOCALE, HOSPITAL, MEDICAL, GROFFER, UGOFFER, HLOFFER, perc100online,
perc_some_online, perc_out, perc_int)

str(data_new)
```

3.4. Handling Missing Values

There were no missing values for the following variables: INSTNUM, LOCALE, EFDETOT and HBCU. However, there were missing values in the two outcome variables: perc100online (989 NAs) and perc_some_online (1014 NAs). It was assumed that institutions with NAs for these variables had no students in these categories. Therefore, all NAs for these variables were recoded as 0.

In addition, some of the independent variables had codes of -1 or -2 in the original datset to deisgnte missing values. To avoid their inclusion in the analysis, there were recoded as NA.

```
data_new[is.na(data_new)] <- 0


#recode missing cases
data_new2$HOSPITAL[data_new2$HOSPITAL < 0] <- NA
data_new2$LOCALE[data_new2$LOCALE < 0] <- NA
data_new2$UGOFFER[data_new2$UGOFFER < 0] <- NA
data_new2$GROFFER[data_new2$GROFFER < 0] <- NA

data_new2$MEDICAL[data_new2$MEDICAL < 0] <- NA
data_new2$ICLEVEL[data_new2$ICLEVEL < 0] <- NA
data_new2$CONTROL[data_new2$CONTROL < 0] <- NA
data_new2$LANDGRNT[data_new2$LANDGRNT < 0] <- NA
data_new2$EFDETOT[data_new2$EFDETOT < 0] <- NA
```

3.5. Renaming Variables

Although most variables in the merged dataset have meaningful names, there was only one variables that needed to be renamed: INSTNUM (renamed to Institution_Name) The dplyr package was employed to rename the column names.

```
data_new %>% rename( Institution_Name =INSTNM)
str(data_new)
```

3.6. Formatting

The only formatting carried out on the data set involved transforming all categorical variables from a numeric to a factor format. This step was necessary in order to be able to include these variables in the regression models and conduct a bivariate analysis in

the exploratory data analysis. The code bellow shows an example of reformatting the CONTROL variable.

```
data_new2$CONTROL[data_new2$CONTROL ==1] <- "public"
data_new2$CONTROL[data_new2$CONTROL ==2 ] <- "private"
# Convert the column to a factor
data_new2$CONTROL <- factor(data_new2$CONTROL)
```

3.7. Handling Outliers

An analysis of the outliers for perc100online revealed that of 223 of the 3,983 observations were outliers; 32 of the outliers had 100% percent of students enrolled online (perc100online=1.00). The following assumption was made to further eliminate potential outliers – if an institution had 80% or more of its students enrolled exclusively in online programs, then it was considered to be primarily an online college or university. These institutions were excluded from the analysis.

In addition, since the focus of the analysis was on small colleges and universities, only institutions with enrollments between 500 and 5000 students were included in the analysis.

```
OutVals1 = boxplot(data_new$perc100online, plot=FALSE)$out
View(OutVals1)

#eliminate schools that are predominantly online – 80% or more students in online classes
data_new2 = subset(data_new2, perc100online < .80  & perc_some_online < .80)
str(data_new2)

#select small schools only
data_new2 = subset(data_new2, EFDETOT > 500  & EFDETOT  < 5000)
```

The final data set (data_new2) contained 1557 observations and 17 variables.

4. Exploratory Data Analysis

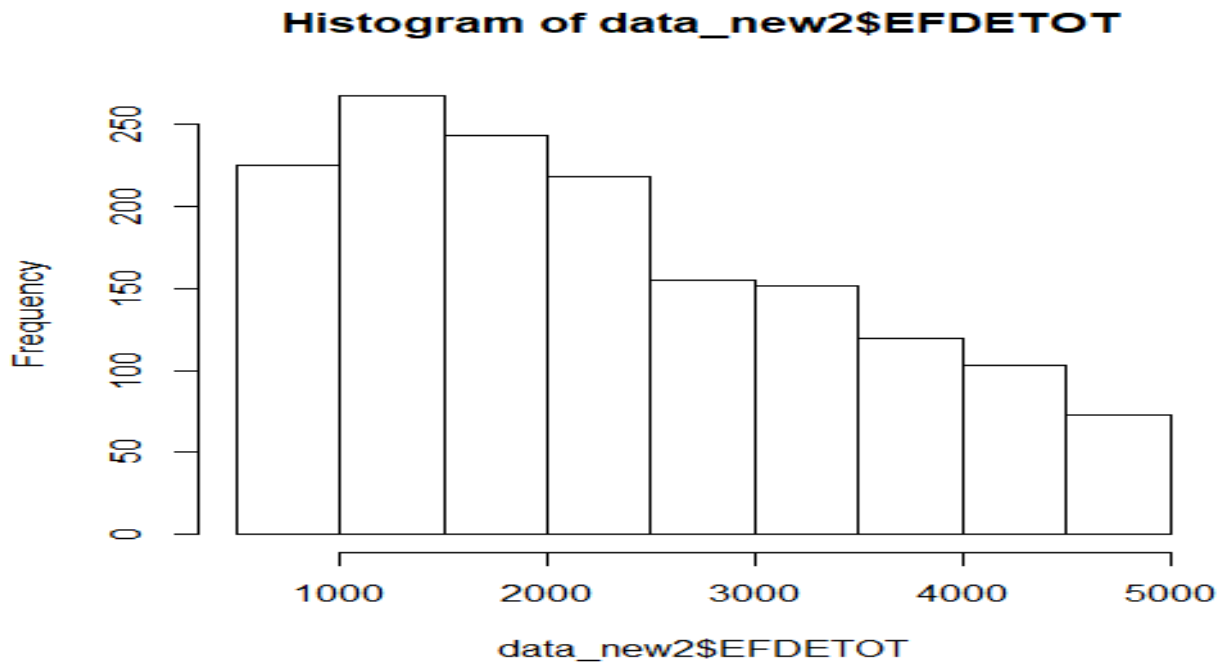
The purpose of the exploratory data analysis was to assess the relationships between percentage of students enrolled exclusively in online programs (perc100onlline) and percent of students taking some their classes online (perc_some_online) and the remaining variables in the data set to determine if there are any factors that have more of an influence on the dependent variables than others. The goal of this analysis was also to visualize the distributions of the continuous dependent variables – EFDETOT, perc100online, perc_some_online to assesses whether they are normally distributed which is a requirement for linear regression.

The following variables were investigated in the bivariate analysis: perc100online, perc_some_online, perc_out, perc_int, HBCU, EFDETOT, UGOFFER, GROFFER, CONTROL, MEDICAL, HOSPITAL, LANDGRNT, and HLOFFER. The bivariate analysis was done using bar charts to explore the relationships between a continuous variable and a factor and scatterplots for the relationships between two continuous variables.

4.1. Total Enrollments

As noted in the data wrangling section, a decision was made to focus this research on small institutions – between 500 and 5000 students. Additionally, the for profit institutions were excluded from the analysis. The median institutional size in the final dataset was 2,291 students. The histogram shows a slightly skewed distribution.

Variable	Min	Q1	Median	Mean	Q3	Max
EFDETOT	501	1290	2094	2291	3167	4999



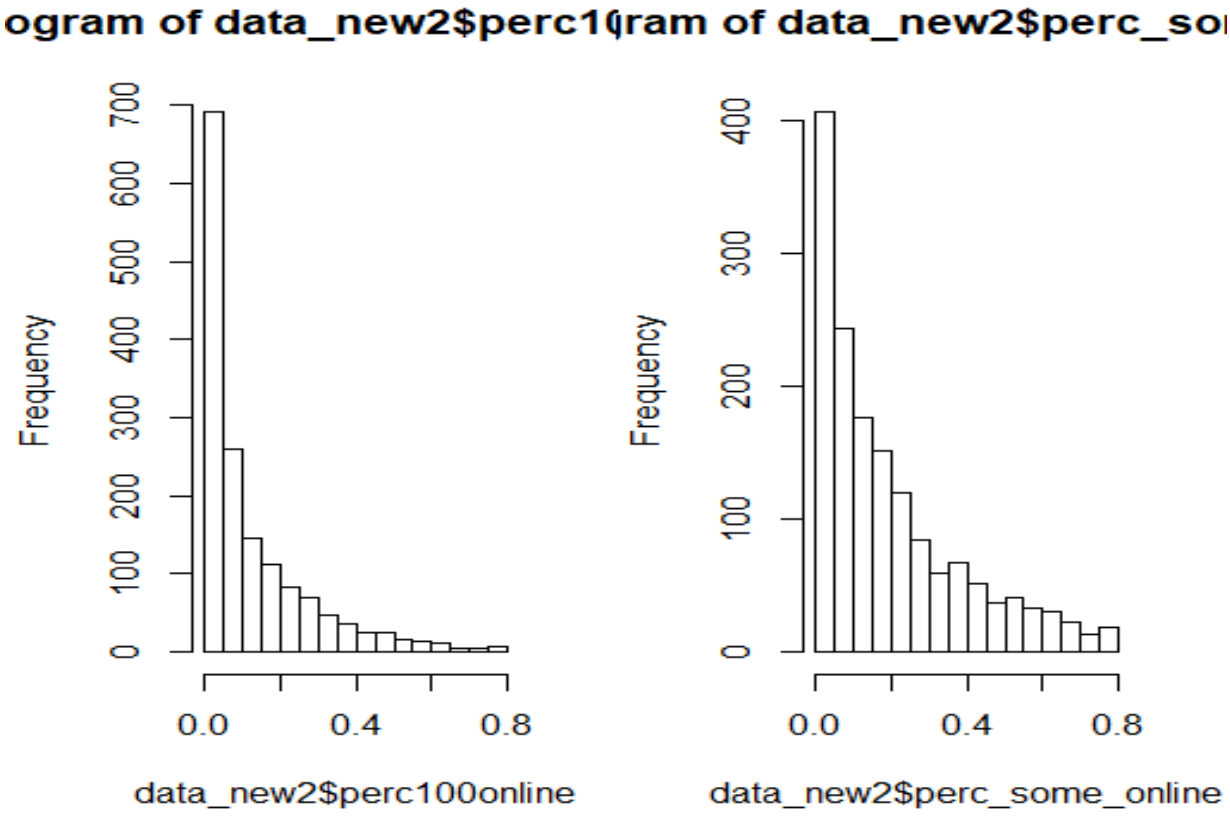
4.2. Percent Students in Online Programs or Taking an Online Course

Descriptive statistics reveal that, overall, the median percentage of students enrolled exclusively in online programs across small institutions is 7%. Additionally, the median percentage of students taking a class online is 13%.

Variable	Min	Q1	Median	Mean	Q3	Max
perc100online (percent students enrolled exclusively in online programs)	0%	1%	7%	12%	18%	79%
perc_some_online (percent students taking some of their classes online)	0%	5%	13%	20%	29%	79%

An examination of the histograms further reveals that both variables are heavily skewed, with most institutions having little or no enrollment in online courses or online

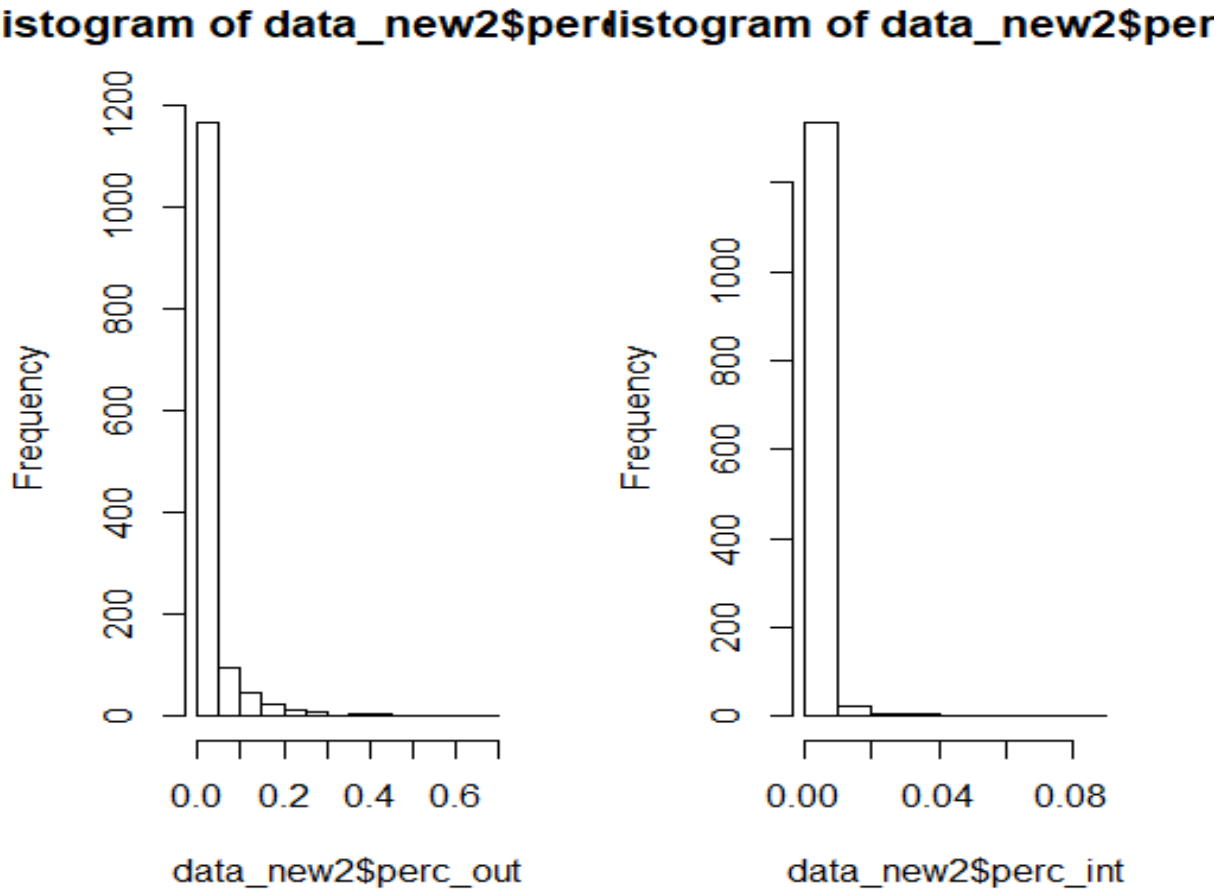
programs. Therefore, the lack of normality for these variables poses a limitation in the use of linear regression.



4.3. Out of State (perc_out) and international (perc_int) online enrollments

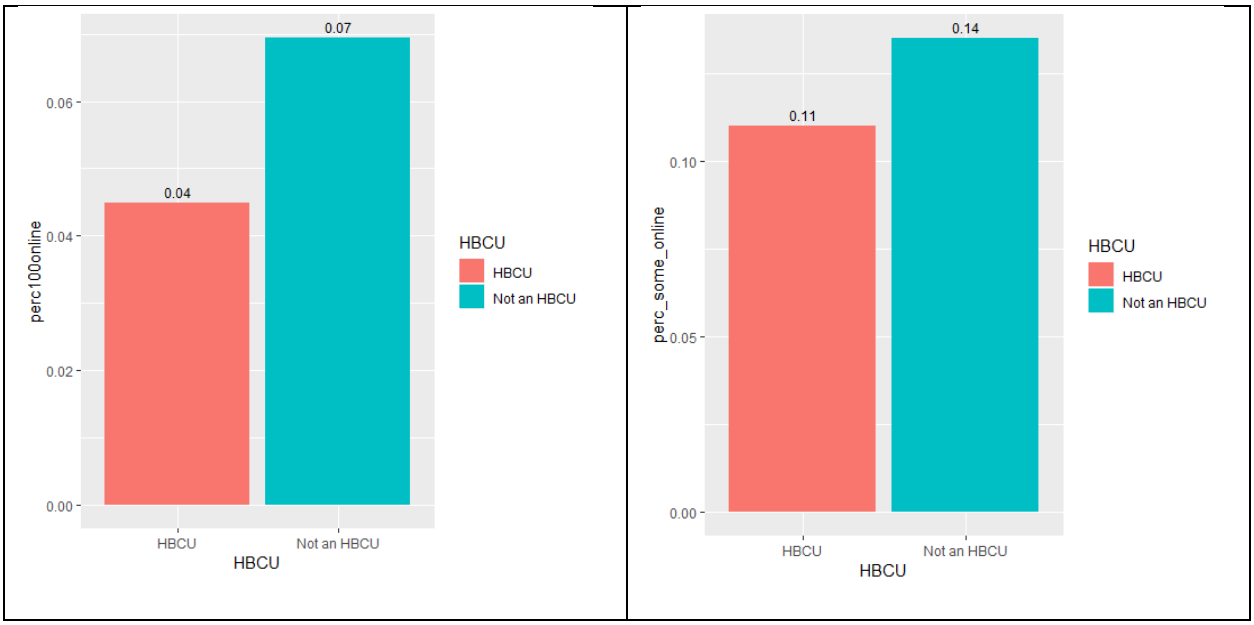
Across the small institutions examined, less than 1% of the total students are online out of state or online international students. The histograms further show that both variables are skewed, with most institutions having few out of state or international students in online courses.

Variable	Min	Q1	Median	Mean	Q3	Max
perc_out	0.0%	0.1%	0.6%	2.8%	2.4%	6.6%
perc_int	0.0%	0.0%	0.0%	0.1%	0.1%	0.9%



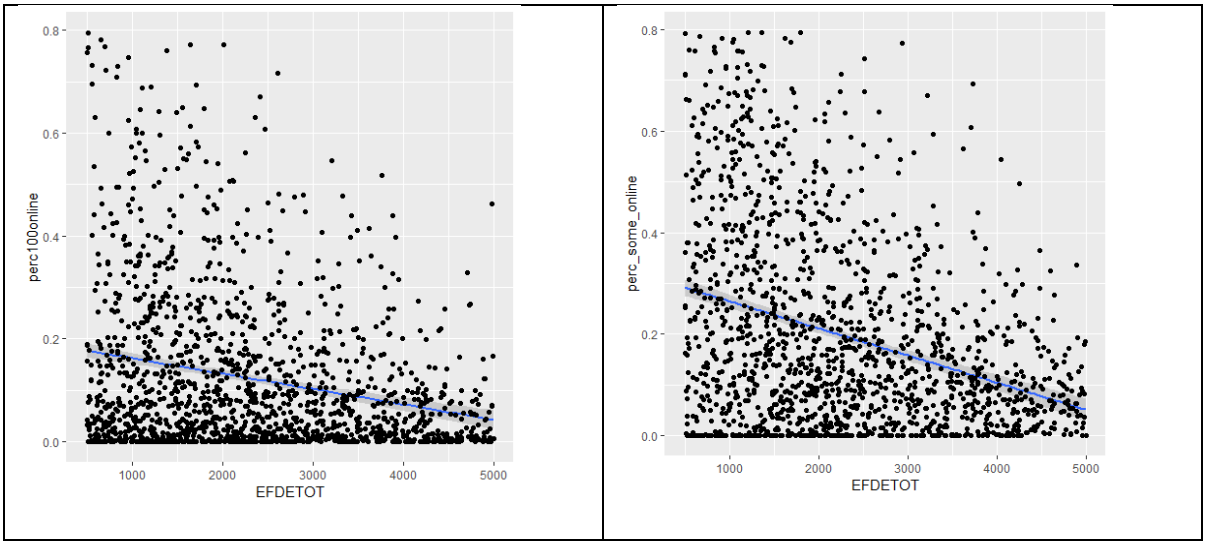
4.4. HBCU

The final data set contains 62 HBCUs and 1495 non HBCUs. Since the distributions for perc100online and perc_some_online are heavily skewed, it is indicated that medians rather than means are used in comparisons across groups of institutions. The bar charts below show the median percentage of students exclusively enrolled in online programs. As hypothesized, there is a difference between HBCUs and non HBCUs for the perc100online, suggesting that HBCUs, on average, enroll a lower percentage of students in online programs (4% vs. 7%). At the same time, the chart for perc_some_online indicates that the percentage of students taking some online classes is also lower at HBCUs than at non-HBCUs (11% vs. 14%).



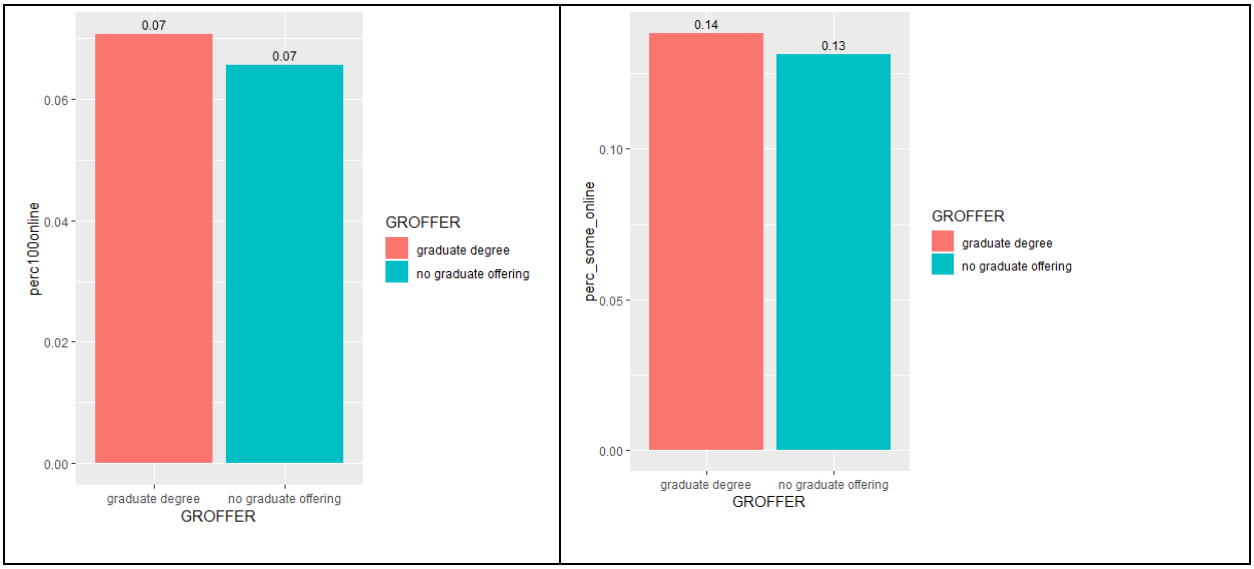
4.5. Total Enrollment (EFDETOT) and Online Enrollments

The scatterplots reveal that there is a weak linear relationship between total enrollment and percentage of students in online programs or percentage of students enrolled in some online classes. The charts also reveal that the percentage of students enrolled in online programs or in some online classes decreases with institutional size.



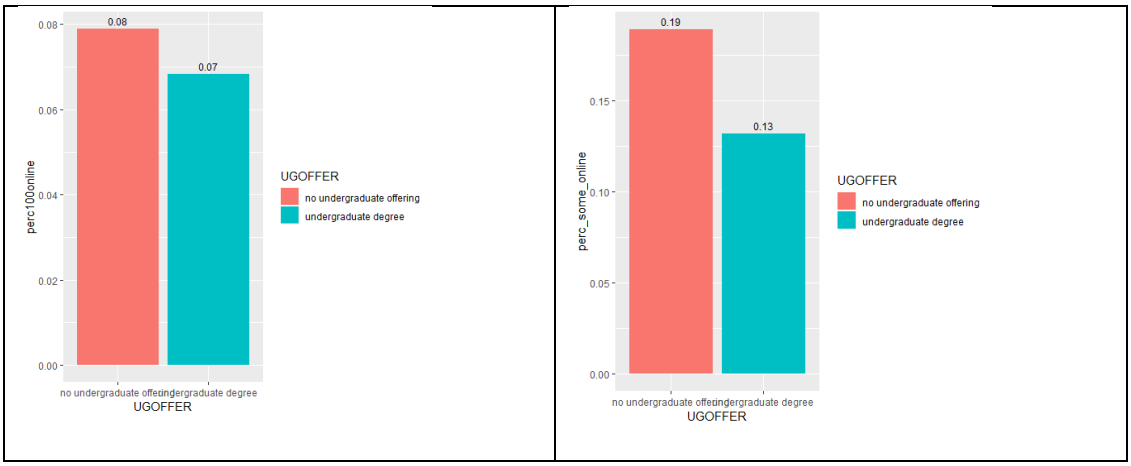
4.6. Offering Graduate Programs (GROFFER) and Online Enrollments

Institutions that offer graduate programs have about the same percentage of students enrolled exclusively in online programs as institutions that have no graduate programs (7%). This is true for the percentage of students taking some classes online (14% vs. 13%).



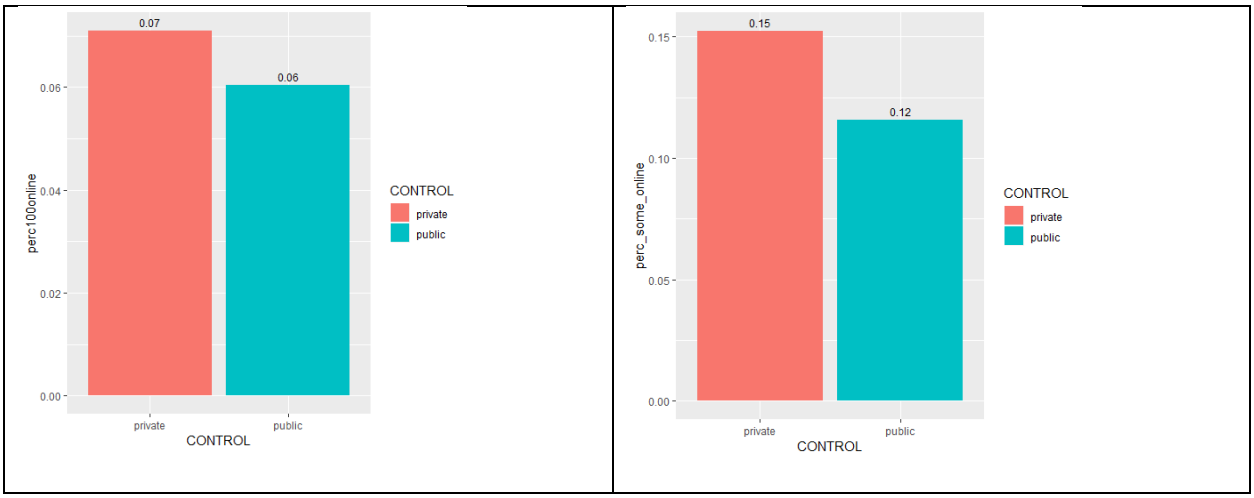
4.7. Offering Undergraduate Programs (UGOFFER) and Online Enrollments

Institutions that do not offer undergraduate programs (functioning as graduate or professional schools only) have a slightly higher percentage of students enrolled in online programs than the rest of the institutions (8% vs. 7%). They also have a higher percentage of students taking some of their classes online (19% vs. 13%).



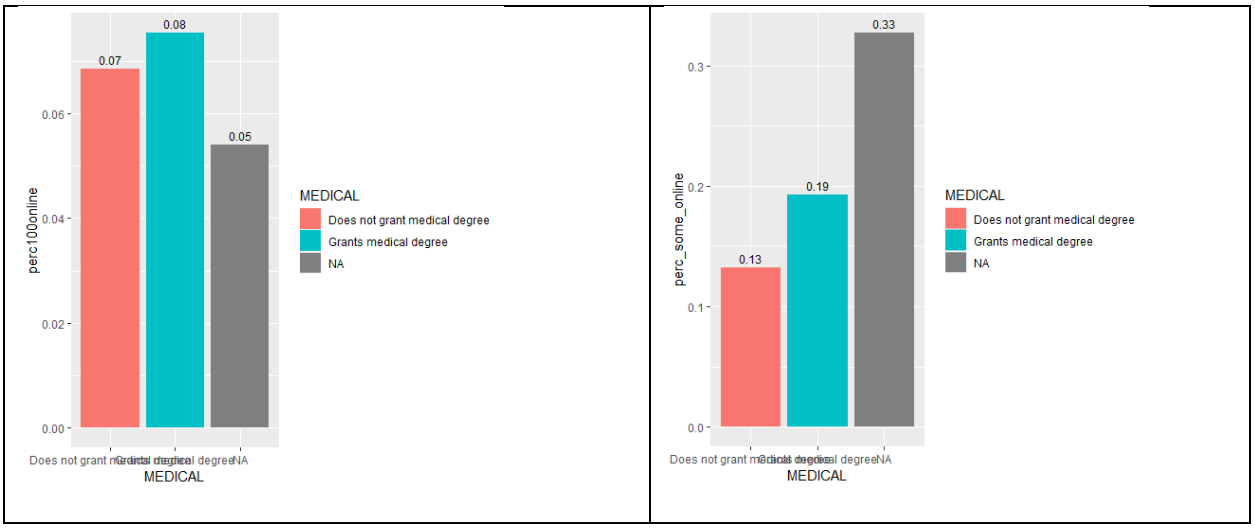
4.8. Institution control (CONTROL) and Online Enrollments

Private institutions have a slightly higher median % of students enrolled in fully online programs than public institutions (7% vs. 6%). Similarly, private institutions have a higher percentage of students taking at least one online course than public institutions (15% vs. 12%).



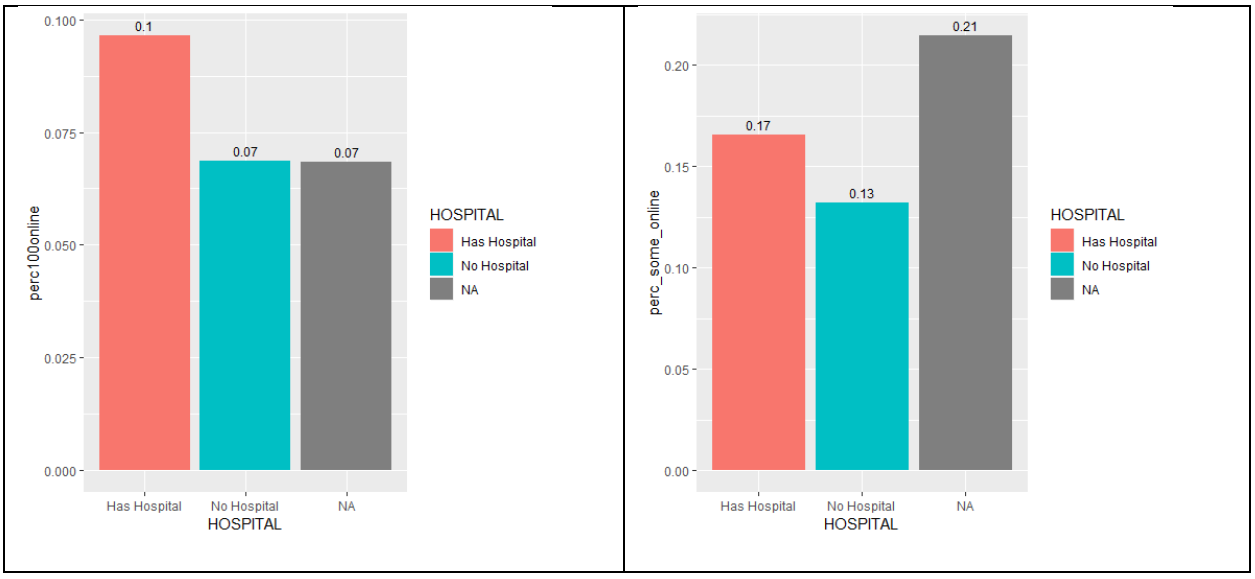
4.9. Offering medical degrees (MEDICAL) and online enrollment

Higher education institutions that offer a medical degree have a slightly higher median percentage of students enrolled in online degree programs than institutions that do not offer a medical degree (8% vs. 7%). In addition, they have a higher percentage of students taking at least a class online (19% vs 13%).



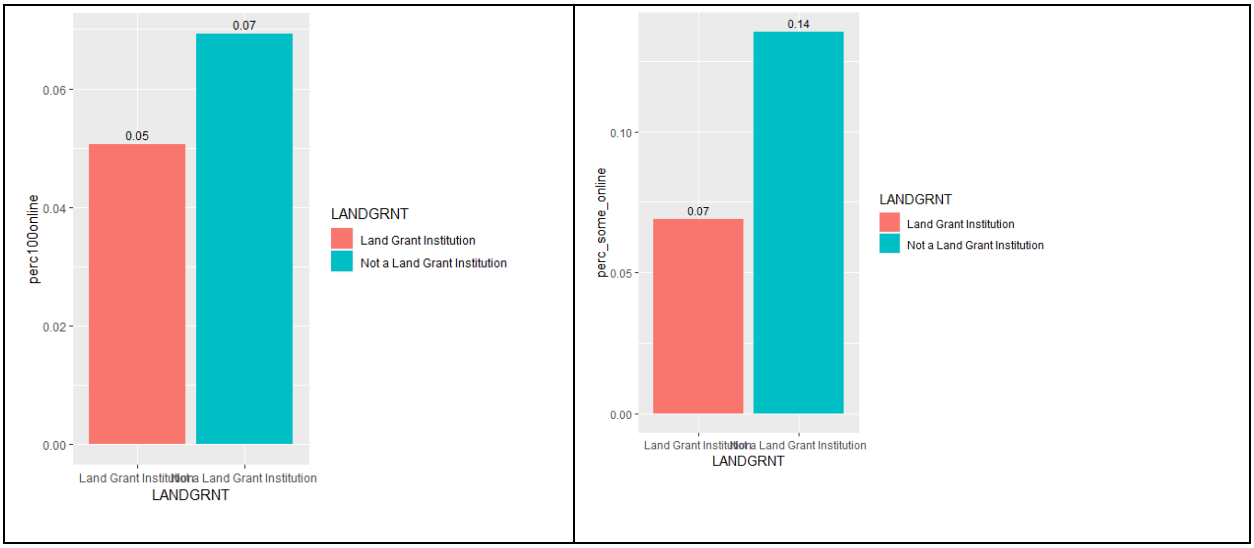
4.10. Having a hospital on campus (HOSPITAL) and online enrollment

Higher education institutions that have a hospital on their campus have a higher percentage of students enrolled in online degree programs than the rest of the institutions (10% vs. 7%). They also have a higher percentage of students taking at least a class online (17% vs. 13%).



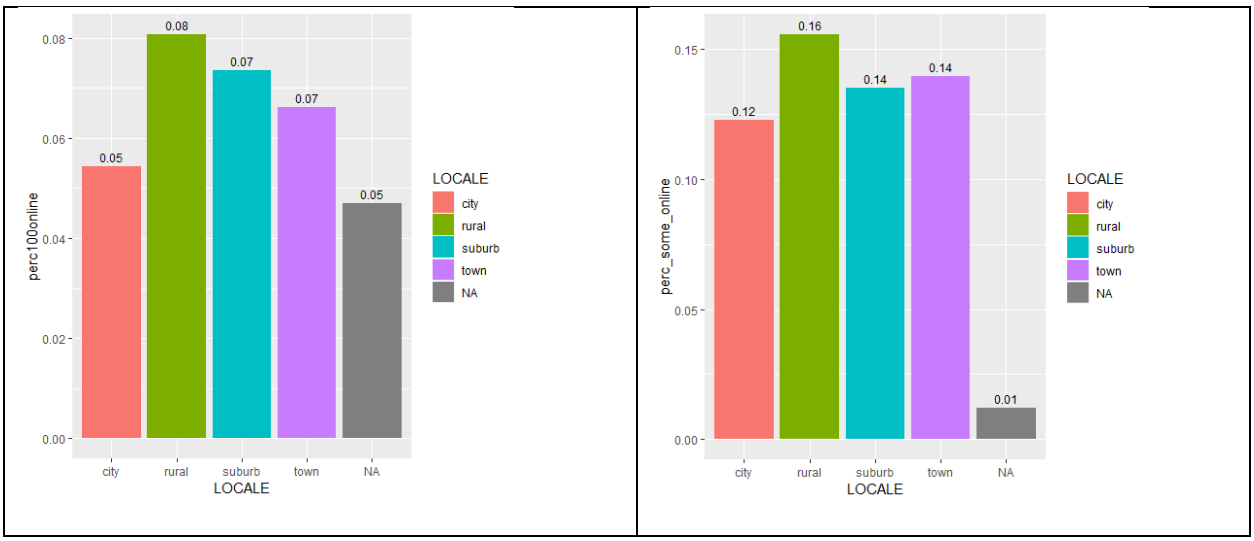
4.11. Land Grant Status (LANDGRNT) and online enrollment

Land grant institutions have a lower percentage of students enrolled in online degree programs than the rest of the institutions (5% vs. 7%). They also have a lower percentage of students taking at least a class online (7% vs. 14%).



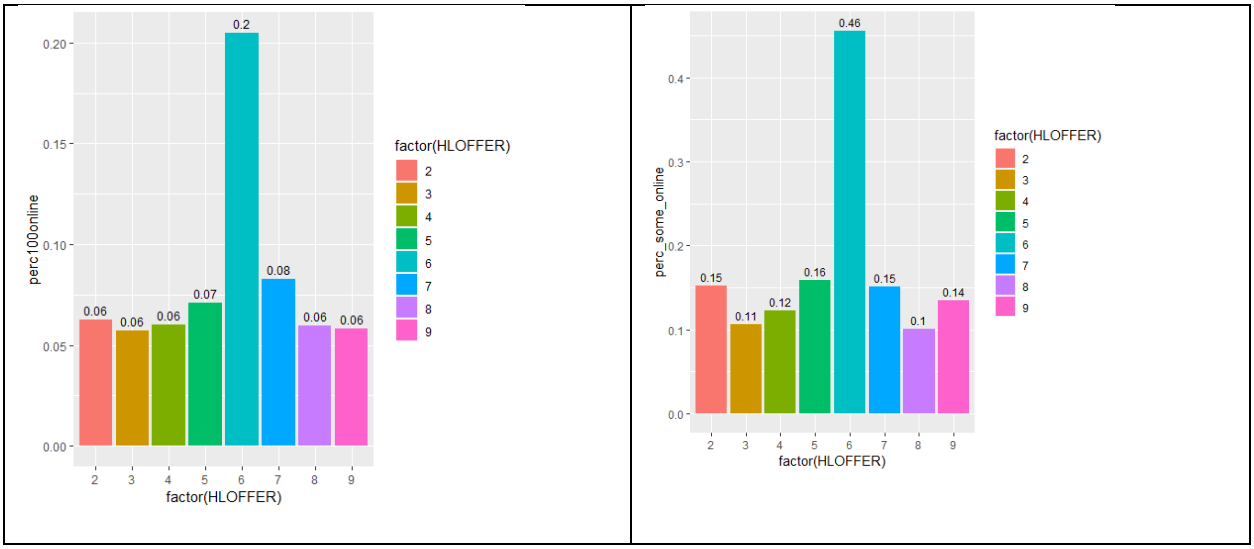
4.12. Degree of urbanization (LOCALE) and online enrollment

Higher education institutions that are located in rural areas tend to have higher percentage of students enrolled in online degree programs than urban institutions (8% vs. 5%). They also have a higher percentage of students taking at least a class online (16% vs. 12%).



4.13. Highest Degree Offered (HLOFFER) and online enrollment

Higher education institutions where the highest degree level offered is a Postbaccalaureate Certificate have a higher percentage of students enrolled in online degree programs than the rest of the institutions (20% vs. 6-8%). They also have a higher percentage of students taking at least a class online (46% vs 10-16%).



Codes for HLOFFER:
2 - Postsecondary award, certificate or diploma of at least one but less than two academic years
3 - Associate's degree
4 - Postsecondary award, certificate or diploma of at least two but less than four academic years
5 - Bachelor's degree
6 - Postbaccalaureate certificate
7 - Master's degree
8 - Post-master's certificate
9 - Doctor's degree

5. Statistical Analysis

The initial data exploration has been useful to understand trends in the data and suggested what will be worth exploring in more detail in the regression models. The initial goal of the project as discussed in the Capstone Proposal was to create a linear regression model which will determine if the percentage of online students is significantly lower at HBCUs than at other institutions, after controlling for relevant

factors. An additional goal was to explore whether, after controlling for relevant institutional characteristics, the addition of online courses leads to an increase in enrollments.

However, since the distribution of perc100online and perc_some_online are highly skewed, with most institutions having no students enrolled in fully online programs or taking online classes, linear regression models cannot be used. One of the assumptions of linear regression is that the dependent variable is normally distributed.

Therefore, the initial goal the analysis was modified. A logistic regression model was used instead to identify the most important institutional characteristics that predict whether an institution offers online courses. Additionally, the dependent variables for the logistic model was perc_some_online instead of perc100online. This is because the percentage of students in fully online programs is relatively low (7%) in the sample. This compares to 13% for the students who take some of their classes online. The variable perc_some_online was recoded as a categorical variable with two levels: 1 – offers some online classes; 2 - no online classes. The logistic regression model answers the following research question:

RQ1. After controlling for size and other relevant institutional characteristics, are HBCUs significantly less likely than other institutions to offer online classes?

In addition to the logistic model with perc_some_online as the dependent variable, a linear regression model was built to estimate whether there is a significant gain in total enrollments from offering online classes, after controlling for relevant institutional characteristics. This model answered the following research question:

RQ2. After controlling for size and other relevant institutional characteristics, does offering online classes yield a significant increase in enrollments?

For both models, the data set was split randomly into two data sets: a training data set consisting of 85% of the observations and a test data set with 15% of the observations. In terms machine learning methods, both models are considered to be supervised regression.

5.1. Logistic Regression Model

The results of the logistic model are presented below:

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.6789021	1.8839132	0.360	0.71857
CONTROLpublic	-0.2127811	0.3041801	-0.700	0.48422
MEDICALGrants medical degree	0.4737612	0.7893316	0.600	0.54837
HOSPITALNo Hospital	0.3798091	0.9030661	0.421	0.67406
LANDGRNTNot a Land Grant Institution	1.7945039	0.6476859	2.771	0.00559 **
UGOFFERundergraduate degree	0.2907398	0.5293928	0.549	0.58287
GROFFERno graduate offering	0.1877876	0.4777077	0.393	0.69424
HBCUNot an HBCU	-1.5881744	0.8048412	-1.973	0.04846 *
EFDETOT	0.0003030	0.0001008	3.008	0.00263 **
LOCALErural	0.1754421	0.3414386	0.514	0.60737
LOCALEsuburb	-0.1669362	0.2757076	-0.605	0.54486
LOCALEtown	-0.0986749	0.2680459	-0.368	0.71278
HLOFFER	0.0146464	0.1195157	0.123	0.90247

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 781.11 on 1247 degrees of freedom

Residual deviance: 757.66 on 1235 degrees of freedom

(75 observations deleted due to missingness)

AIC: 783.66

Number of Fisher Scoring iterations: 6

The estimates from the logistic regression characterize the relationship between the predictor and response variable on a log-odds scale. In order to facilitate the interpretation of the results, the coefficients were transformed using the exponential function. The new coefficients are presented in the following table:

```
##                                      Estimate   Std. Error    z value
## (Intercept)                         1.9717118 1.8839132294  0.3603680
## CONTROLpublic                       0.8083331 0.3041800934 -0.6995234
## MEDICALGrants medical degree        1.6060234 0.7893316284  0.6002055
## HOSPITALNo Hospital                 1.4620054 0.9030661387  0.4205772
## LANDGRNTNot a Land Grant Institution 6.0164894 0.6476859065  2.7706392
## UGOFFERundergraduate degree         1.3374165 0.5293927777  0.5491948
```

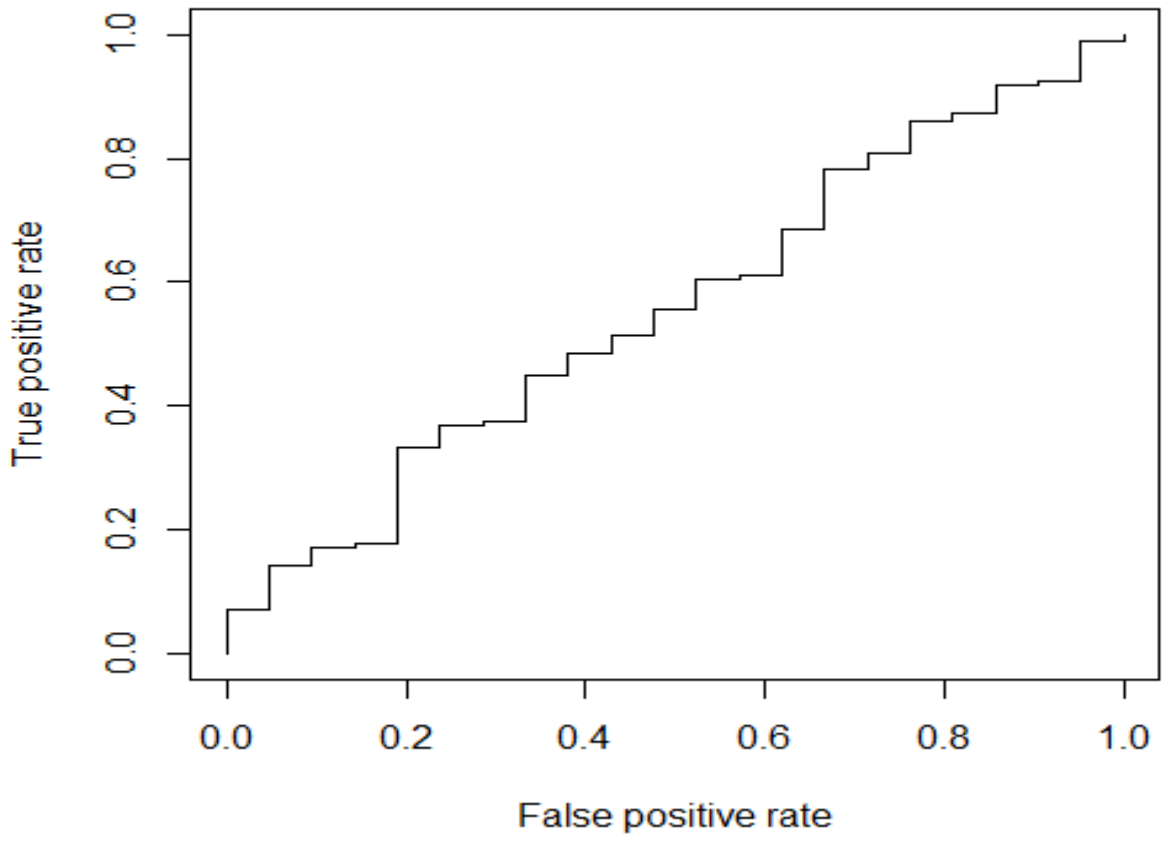
```
## GROFFERno graduate offering          1.2065772 0.4777076688  0.3931014
## HBCUNot an HBCU                      0.2042982 0.8048411665 -1.9732768
## EFDETOT                              1.0003031 0.0001007504  3.0078460
## LOCALErural                          1.1917730 0.3414385641  0.5138321
## LOCALEsuburb                         0.8462536 0.2757076417 -0.6054828
## LOCALEtown                           0.9060372 0.2680458881 -0.3681270
## HLOFFER                              1.0147542 0.1195156734  0.1225479
```

The results indicate that, after controlling for key institutional characteristics such as size, location, and highest degree offered, non HBCUs are 20% more likely than HBCUs to have students enrolled in online classes. In addition, non land grant institutions are 6 times more likely than land grant institutions to have students enrolled in online classes.

To assess the accuracy of the predictions by the model, the ROC Curve was examined.

The receiving operating characteristic (ROC) is a measure of classifier performance. Using the proportion of positive data points that are correctly considered as positive and the proportion of negative data points that are mistakenly considered as positive, we generate a graphic that shows the tradeoff between the rate at which you can correctly predict something with the rate of incorrectly predicting something. Ultimately, we're concerned about the area under the ROC curve, or AUROC. That metric ranges from 0.50 to 1.00, and values above 0.80 indicate that the model does a good job in discriminating between the two categories which comprise our target variable.

The results show an AUROC value of .56, which means that the model does a relatively poor job at discriminating between the two categories (online vs. no online students).



5.2. Linear Regression Model

A multiple linear regression model was constructed to examine the impact of offering online classes on total enrollment, after controlling for relevant institutional characteristics.

```
Call:
lm(formula = EFDETOT ~ CONTROL + MEDICAL + HOSPITAL + LANDGRNT +
    UGOFFER + GROFFER + HBCU + LOCALE + HLOFFER + some_online,
    data = train)
```

Residuals:
   Min   1Q Median   3Q   Max
-2733.2 -762.6 -153.1   657.4 3401.3

```
## Coefficients:
##                                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)                         -1732.507    585.115  -2.961  0.00313**
## CONTROLpublic                        1328.387     84.967  15.634  < 2e-16***
## MEDICALGrants medical degree            5.438    229.513   0.024  0.98110
## HOSPITALNo Hospital                   250.980    281.540   0.891  0.37286
## LANDGRNTNot a Land Grant Institution  390.944    259.801   1.505  0.13263
## UGOFFERundergraduate degree          1184.962    182.583   6.490 1.24e-10***
## GROFFERno graduate offering           222.670    147.730   1.507  0.13199
## HBCUNot an HBCU                       260.899    157.283   1.659  0.09741
## LOCALErural                          -591.555     98.955  -5.978 2.95e-09***
## LOCALEsuburb                            6.131     84.885   0.072  0.94244
## LOCALEtown                           -407.384     80.917  -5.035 5.50e-07***
## HLOFFER                               208.008     35.823   5.807 8.10e-09***
## some_onlinesome online enrollment     321.967    102.695   3.135  0.00176**
```

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1055 on 1235 degrees of freedom
  (75 observations deleted due to missingness)
Multiple R-squared: 0.2068,   Adjusted R-squared: 0.1991
F-statistic: 26.83 on 12 and 1235 DF, p-value: < 2.2e-16

The model explains 20% of the variance in total enrollment. The results suggest that, on average, after controlling for relevant institutional characteristics, offering online classes results in 322 additional students. Offering undergraduate programs adds on average 1185 students. Public institutions have on average 1328 more students than private institutions. In addition, compared to institutions located in large urban areas, institutions located in towns enroll on average 407 fewer students. Institutions located in rural areas have 592 fewer students on average than urban institutions. Lastly, the higher the degree offered by the institution, the higher the total enrollment.

The min_max accuracy rate was used to calculate the accuracy rate of linear regression model lm1. This rate was found to be 64%, which means that the model predicts with accuracy 64% of the observations in the test data set.

Conclusions

Results from the capstone project show that:

1) The logistic regression model revealed that, after controlling for size, location and other institutional characteristics, HBCUs are 20% less likely than other institutions to have students enrolled in online classes. A gap in the percentage of students taking online classes between the two groups of institutions was also found during the exploratory data analysis.

2) The multiple linear regression model showed that by offering online classes, higher education institutions can increase their enrollments by 322 students. This net gain is obtained after controlling for key factors such as institutional size, public vs. private, and location.

Recommendations

Given the positive impact of offering online classes on total enrollment, small private institutions, especially those located in rural areas and small towns should plan to expand online classes. They should also plan to offer post-bachelor's degrees since these program attract additional students.

Given that the logistic regression results and the exploratory data analysis revealed a gap between HBCUs and non HBCUs in terms of online student enrollment, HBCUs should invest more in training faculty members on creating online courses and providing the learning management systems infrastructure and technical support.

Further Work

While the variable LOCALE captures the degree of urbanization for each higher education institution, the analysis does not present possible patterns in the data based on geography. It would be interesting to explore for instance whether students at colleges and universities on the West coast are more likely to take online classes than their peers in the rest of the country. The gmap function could be used to construct a map showing levels of perc_some_online by area.

Another improvement to the analysis would be the addition of new variables to the dataset to increase the predictive accuracy of the linear and logistic regression models. For instance, the magnitude of enrollment in online classes could be associated with

financial factors such as total educational expenditures or total educational expenditures as a percentage of the total budgeted. Colleges with large instructional budgets are expected to invest more in online teaching technologies. It could also be associated with the student faculty ratio, as colleges with a high student-faculty ratio are expected to offer more online classes than colleges with a small student-faculty ratio. These variables could be extracted from the Fall Enrollment, Staff, and Finance datasets available on the National Center for Education Statistics website.