**Exploratory Data Analysis**

The purpose of the exploratory data analysis was to assess the relationship between percentage of students enrolled exclusively in online programs (perc100onlline), percent of and the remaining variables in the data set to determine if there are any factors that have more of an influence on this variable than others. The goal of this analysis of was to examine the more influential variables in the data set and also to visualize the distributions of continuous dependent variables – perc100online and perc\_some\_online.

The following variables were investigated in this section: perc100online, perc\_some\_online, perc\_out, perc\_int, HBCU, EFDETOT, GROFFER, CONTROL, . This was done via bar boxlpots to explore the relationships between a continuous variable and a factor, and scatterplots for the relationships between two continuous variables.

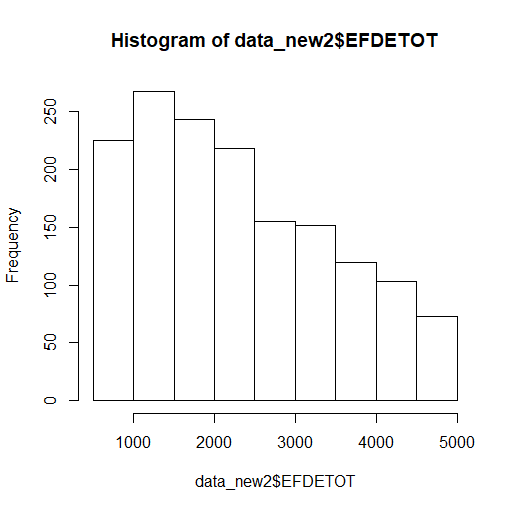
A definition for each variable is included in the following table:

|  |  |  |
| --- | --- | --- |
| 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 |

**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 pro

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

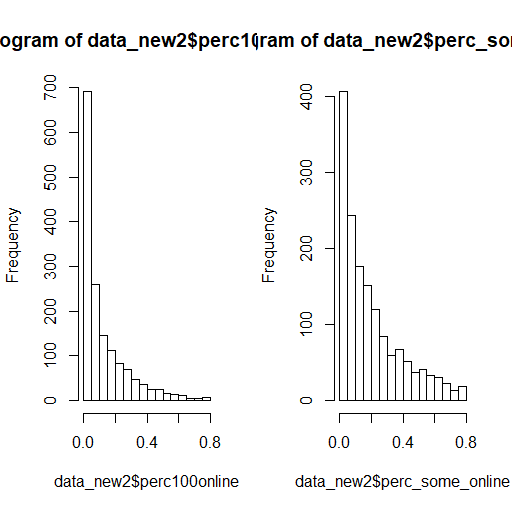


**Percent Students in Online Programs or Taking an Online Course**

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% |

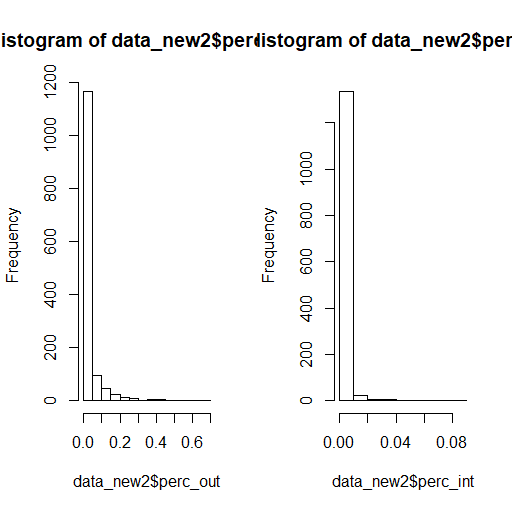
An examination of the histograms further reveals that both variables are skewed, with most institutions having little or no enrollment in online courses or online programs. The lack of normality for these variables poses a limitation in the use of linear regression.



**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 insi

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **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% |



**HBCU**

The final data set contains 62 HBCUs and 1495 non HBCUs. 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%).

|  |  |
| --- | --- |
|  |  |

**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.

|  |  |
| --- | --- |
|  |  |

**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%).

|  |  |
| --- | --- |
|  |  |

**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%).

|  |  |
| --- | --- |
|  |  |

**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%).

|  |  |
| --- | --- |
|  |  |

**Offering medical degrees (MEDICAL) and online enrollment**

Higher education institutions that offer a medical degree have a slightly higher 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%).

|  |  |
| --- | --- |
|  |  |

**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%).

|  |  |
| --- | --- |
|  |  |

**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%).

|  |  |
| --- | --- |
|  |  |

**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%).

|  |  |
| --- | --- |
|  |  |

**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

**Summary of Data Exploration and Next Steps**

The initial exploration has been useful to understand trends in the data and suggest 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. Since the distribution of perc100online is highly skewed, with most institutions having no students enrolled in fully online programs, one cannot use a linear regression model. One of the assumptions of linear regression is that the dependent variable is normally distributed. Therefore, the initial goal the analysis was modified to the following: a logistic regression model will be used instead to identify the institutional characteristics that predict whether an institution offers online courses. Additionally, the logistic model will be conducted for 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.

In addition the logistic model with perc\_some\_online as the dependent variable, a linear regression model will be built to estimate the gain in total enrollments from offering online programs, after controlling for relevant institutional characteristics.