**Springboard-Capstone-Project**

Springboard Capstone Project for Foundations of Data Science Course

**The Problem**

For many small colleges and universities, adding online programs can boost enrollments and add new sources of revenues. Online program offerings can bring in additional students from out of state as well as international students. The addition of online programs can benefit not only higher education institutions, but also nontraditional students. One of the main attractions of these programs is that it allows students who have full-time jobs and/or family responsibilities to further their education.

The addition of the new online programs will improve the enrollment at HBCUs, making these institutions more attractive to non-traditional students.

The primary goal of this project is to assess the extent to which HBCUs have expanded their online enrollments. Specifically, the percentage of students enrolled in fully online programs is compared between HBCUs and non HBCUs. Similar comparisons are made for students enrolled in some online classes between HBCUs and non-HBCUs. Comparisons will also be made for percentage of students from out of state and percent of international students.

**The Datasets**

Two datasets will be used for the purpose of this project:

1. Fall 2017 Distance Enrollment Data contains 22,590 observations representing enrollment in distance education programs and courses at colleges and universities in the U.S. There are multiple observations for each institution, representing enrollment data for the following categories: All students total, Undergraduate total, Degree/certificate-seeking total, Non-degree/certificate-seeking, and Graduate students.

The data was downloaded the IPEDS Fall Enrollment Survey collected by the National Center for Education Statistics. The following are the main 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 out of state
  + Number of international students

1. 2017 Institutional Characteristics Data contains variables describing each higher education institution. The following are key variables that will be used in the analysis:
   * Control (public, private for profit, or private non-profit)
   * Locale
   * Size
   * HBCU

**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.

**Subseting the 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 for all students in aggregate. During Exploratory Data Analysis, the resulting data set could be further subset first to select only HBCUs and later to select non-HBCUs public or private not for profit.

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

merged\_all\_levels <- subset(merged2017, EFDELEV == "1" & CONTROL < 3 )

**Calculating New Variables**

None of the variables of interest for this analysis were included in the 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)

**Deleting Unnecessary Variables**

The resulting merged data set contained 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 10 variables.

data\_new <- select(data\_new, INSTNM, UNITID, STABBR,EFDETOT, HBCU, LOCALE, perc100online, perc\_some\_online, perc\_out, perc\_int)

str(data\_new)

**Handling Missing Values**

There were no missing values for the following variables: INSTNUM, LOCALE, EFDETOT and HBCU. However, there were missing values in the outcome variables: perc100online (989 NAs), perc\_some\_online (1014 NAs), perc\_out (1304 NAs), and perc\_int (1304 NAs). It was assumed that institutions with NAs for these variables had no students in these four categories. Therefore, all NAs for these variables were recoded as 0.

summary(data\_new)

data\_new[is.na(data\_new)] <- 0

summary(data\_new)

**Renaming Variables**

Although most variables in the merged dataset have meaningful names, there were two variables that needed to be renamed: INSTNUM (renamed to Institution\_Name) and EFDETOT (renamed to Total\_Enrollment. The dplyr package was employed to rename the column names.

data\_new %>% rename( Institution\_Name =INSTNM, Total\_Enrollment= EFDETOT )

str(data\_new)

**Formatting**

The only formatting carried out on the data set involved transforming the HBCU variable from a numeric to a factor format. This step was necessary in order to be able to conduct exploratory data analysis for the two groups of institutions.

factor\_hbcu <- factor(data\_new$HBCU)

**Handling Outliers**

An analysis of the outliers for perc100online revealed that of 223 of the 3,983 were outliers; 32 of the outliers had 100% percent of students enrolled online (perc100online=1.00). Since the analysis should exclude exclusively online institutions, all outlier values of 1.00 for perc100online were changed to NA.

OutVals = boxplot(data\_new$perc100online, plot=FALSE)$out

View(OutVals)

perc100online[perc100online =1] <- NA