**Data wrangling for Capstone Project**

The initial data set contained the following data, the 171 Post-Secondary Institutions in the State of Georgia and the 8 variables listed in the table below.

Coding for the following actions was performed using R.

**I. Renaming Columns:** There was a need to change the names of the columns because the column names were too long for ease of use. The column names for the 8 variables were shortened for coding efficiency and accuracy in describing the data.

|  |  |
| --- | --- |
| Original Column Name | Short Column Name |
| Total price for in-state students living on campus 2017-18 (DRVIC2017) | Price\_IO |
| Total price for out-of-state students living on campus 2017-18 (DRVIC2017) | Price\_OO |
| Total price for out-of-state students living off campus (not with family) 2017-18 (DRVIC2017) | Price\_OF |
| Total price for in-state students living off campus (not with family) 2017-18 (DRVIC2017) | Price\_IF |
| Full-time retention rate 2017 (EF2017D) | Reten |
| Student-to-faculty ratio (EF2017D) | SF\_Ratio |
| Total men (EF2017 All students Undergraduate total) | Men |
| Total women (EF2017 All students Undergraduate total) | Wom |

**II. Removing Missing Data:** There was 2 levels of data cleaning as it relates to missing data, based on the quantity of missing Data. The rationale for these decisions is listed below.

1. Some Institutions in the state of Georgia do not have dormitories or on campus housing for students and do not have price data for those variables. Two columns (variables) were removed based on the huge amount of missing data. The two columns removed were Total price for in state students living on campus (Price\_IO), and Total price for out of state students living on campus (Price\_OO). Sixty-five percent, i.e. 112/171, of the Institutions had missing data in these columns. The institutions either do not have on campus housing or did not submit data for variables as it relates to Total price for on campus housing in the year 2017.
2. The variable we wish to predict is Retention Rate. There were 25 Institutions which listed no data regarding Retention Rate. The value of 25 was obtained in the following manner; there was 16 Institutions which provided no data for any of the 8 variables we are considering, and there were 9 Institutions which provided no data values for Retention in the year 2017. Those Institutions were removed from selected Statistical Analyses.