# Where do I go to College?

December 2017

Updated on June 2018

**Neel Kumtakar**

## Introduction:

When high school students are about to graduate from their respective institutions, most of them have no idea on what college they should go to.  In order words, they are unsure of what college/university is right for them. Thus, a system of college scorecards is needed to remedy this problem. A college scorecard is just a card that has vital information about a college, such as ethnicity of students, the percentage of students that return, the rankings, etc. The database, which came from Data.Gov, contains all the data of these scorecards of most if not all the universities of the US.

The goal of the analysis is not to find out what factors will make a student choose one college over the another. But rather find a correlation or relationship between certain variables, that would in my opinion be most vital to a prospective high school graduate.

Before we develop a relationship, finding the variable that has the most weight in a student's decision is recommended. That way, we can focus on finding certain variables that might affect that variable.

In my opinion, I believe that the graduation rate of a college is enough to be a deciding factor for a student's choice in college.  While factors like prestige are enough to sway a student when choosing a college, the chance that a student graduates with his/her desired field will in my opinion, outweigh the other factors.

What factors will influence the graduation rate you might ask? We will try to investigate.

# Hypothesis-

The aims of the data set are to determine a relationship between the graduation rate and each of the following variables:

**1) Projected graduate salary**

**2) The percentage of male students at a college**

**3)  The Percentage of Female Students at a college**

**4) Student Retention rate**

**I hypothesize:**

**a) If the salary after attending is higher, the graduation rate will be higher.**
*Students will be more motivated to graduate if they are promised a high raise in salary after graduation.*

**b) If the percentage of students that return after 1 year is higher, the graduation rate will be high**

*If the students return after 1 year, they are more likely to graduate.*

**c) If the percentage of women at a college is higher, the graduation rate should be high**

*This hypothesis may or may not be true, but I believe that if you have more females at a college, the graduation rate, the chance that at least one graduates will be high.*

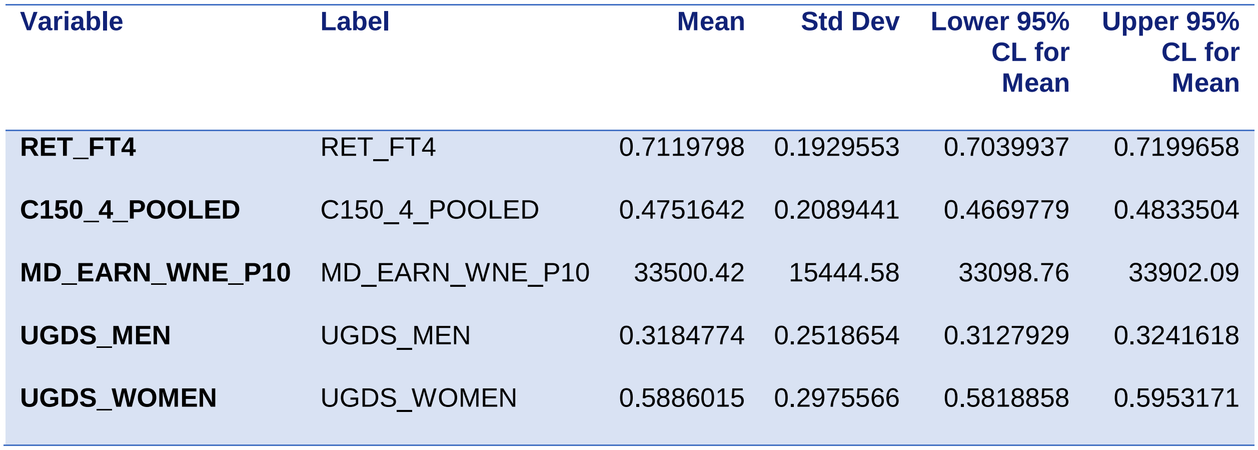
**d) If the percentage of men at a college is high, the graduation should be high.**
The reasoning for c) should apply in this case.

# Data Description:

The Data is an extremely large dataset 1777 variables. For now, we are choosing 5 to investigate out of this dataset.

| Variable Name | Description | Type |
|---|---|---|
| **C150_4_POOLED** | Student Graduation Rate | Dependent Variable |
| **UGDS_MEN** | % of male Students out of the student population | Independent Variable |
| **UGDS_WOMEN** | % of Female Students out of the student Population | Independent Variable |
| **MD_EARN_WNE_P10** | Median Earnings of a student in their 5th-6th year of postgrad | Independent Variable |
| **RET_4** | Student Retention rate- % of students that return to the college after a year | Independent Variable |

Summary data of the variables:



| Variable | Label | Mean | Std Dev | Lower 95% CL for Mean | Upper 95% CL for Mean |
|---|---|---|---|---|---|
| **RET_FT4** | RET_FT4 | 0.7119798 | 0.1929553 | 0.7039937 | 0.7199658 |
| **C150_4_POOLED** | C150_4_POOLED | 0.4751642 | 0.2089441 | 0.4669779 | 0.4833504 |
| **MD_EARN_WNE_P10** | MD_EARN_WNE_P10 | 33500.42 | 15444.58 | 33098.76 | 33902.09 |
| **UGDS_MEN** | UGDS_MEN | 0.3184774 | 0.2518654 | 0.3127929 | 0.3241618 |
| **UGDS_WOMEN** | UGDS_WOMEN | 0.5886015 | 0.2975566 | 0.5818858 | 0.5953171 |

The table above shows the standard deviation, the mean and the 95% CI for the variables.

### Retention:

Observation: RET_4, represents the college retention rate, and according to this graph, the mean is 0.712, the Standard Dev is 0.193, and the 95 CI is (0.704, 0.720).

Interpretation: The average college has a 71% retention rate, and since the deviation is about 0.2, the retention rates of the college do not differ by that much.

### Percentage of female students-

Observation: UGDS_WOMEN, represents the % of females at a college. For the Female UGDS variable, the mean is 0.59, the SD is 0.25, the CI is (0.582, 0.595).

Interpretation: In context, an average college in the US will have a student population approximately 59% female.

### Percentage of Male Students-

Observation: The UGDS Male variable, has a mean of 0.32, a SD of 0.25 and a 95% Confidence Interval of (0.313, 0.324).

Interpretation: An average college will have a 32% male student population. While these % of males and females do not add to 1, keep in mind the table does not account for the people that do not identify with a specific gender, i.e. Transgender students.

## Graduate Earnings-

**The variable MD_EARN_WNE_P10 represents the median earnings of the graduate student in his/her 5th/6th year of postgrad since graduating the college. This variable has a mean of 33500.42, a SD of 1545, and a CI of (33098.76, 33902.09). A student who has graduated college is projected to make about $33.5 K in his/her 5th/6th year of postgrad.**

**Lastly, CI50_4_PO OLED, is the graduation rate of the students graduating from a college. This variable has a mean of 0.475, a SD of 0.208 and a CI of (0.4669, 0.4833). The graduation rate for a college in the US is about 48% percent. This is not a good statistic, as this means a college student has a 48% chance of getting a bachelor's degree!**

Below is the 5-number summary: the min, first quartile, median, third quartile and max of the following variables.

| 5 number summaries of variables | | | | | |
|---|---|---|---|---|---|
| Variable name | Retention rate (RET_4,) | Graduate student Earnings (5th/ 6th year Postgrad) (MD_EARN_WNE_P10) | Graduation rate<br>(CI50_4_POOLED) | Percentage of Males<br>(UGDS_MEN) | Percent of Females<br>(UDGS_<br>WOMEN) |
| Max | 1.0000 | 250000 | 1.0000 | 1 | 1 |
| 3rd Quartile | 0.8343 | 39500 | 0.6239 | 0.46095 | 0.84125 |
| Median | 0.7442 | 31100 | 0.4610 | 0.33605 | 0.60570 |
| 1st Quartile | 0.6268 | 24100 | 0.3261 | 0.07825 | 0.47435 |
| Min | 0.0000 | 9100 | 0.0000 | 0.00000 | 0 |

## Methods:

**The statistical methods applied up to this point are the 5-number summary, the histograms, the standard deviation, the 95% CI, and later the multi regression.**

**Formula for finding the quartiles: Q= p(n+1), where p is the percentage, n is the number of elements in the set. Note that the set must be ordered lowest to highest.**

**Formula for finding mean:** $\overline{X} = (\ 1/n) \sum_{i=0}^{n} i$

**Formula for finding standard Deviation:** $\sigma = \sqrt{\frac{\sum_{i=0}^{n}(x_i - \bar{x})^2}{n}}$

## Seven Steps to building a multi regression model:

1. **Collect Data.**

2. **Hypothesize form of expectation of y, or the mean.** $E(y) = f(\bar{x})$.

3. **Estimate the Betas/ Parameters**.

4. **Estimate Errors (specific the assumption distribution).**

5**. Evaluate Model Utility.**

**6. Check if error model is acceptable.**

**7. Make inferences**. **That is, determine where the model is valid for use.**

## Results:

### According to the 7 step for building a multi-regression model:

### Step 1:

**The dataset was collected on September 28, 2015, and updated on September 29, 2017. The office of Planning, Evaluation and Policy Development of the US Department of Education collected the data, with Brian Fu, a Program/Management Analyst being the main source of contact for this dataset. The dataset is an excel spreadsheet, with a large amount of variables. For this project, the dataset was condensed and sent to SAS for analysis.**

### Step 2:

**The form of expectation for the multi regression line is**

$$\boldsymbol{y = B_0 + B_1 X_1 + B_2 X_2 + B_3 X_3 + B_4 X_4}$$

### Step 3:

**Using SAS we estimate the Coefficients (betas) of the model**

$$\boldsymbol{B_0 = -0.139,}$$

$$\boldsymbol{B_1 = 0.581,}$$

$$\boldsymbol{B_2 = -0.029,}$$

$$\boldsymbol{B_3 = -0.00692,}$$

$$\boldsymbol{B_4 = 0.00000579.}$$

## Interpretation of Betas:

- **As the retention rate increases by 1% the graduation rate increases by 58.13% while keeping other variables constant.**
- **As the percentage of male student's increases by 1%, the graduation rate decreases by 2.9% while keeping other variables constant.**
- **As the percentage rate of female students increases by 1%, the graduation rate decreases by 0.692% while keeping other variables constant.**
- **Finally, when the median earnings increases by $1, the graduation rate increases by 0.0006%, while keeping the other variables constant.**

**Parameter Estimates**

| Variable | Label | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| |
|---|---|---|---|---|---|---|
| **Intercept** | Intercept | 1 | -0.13896 | 0.02145 | -6.48 | <.0001 |
| **RET_FT4** | RET_FT4 | 1 | 0.58132 | 0.01907 | 30.48 | <.0001 |
| **UGDS_MEN** | UGDS_MEN | 1 | -0.02918 | 0.01848 | -1.58 | 0.1145 |
| **UGDS_WOMEN** | UGDS_WOMEN | 1 | -0.00692 | 0.01701 | -0.41 | 0.6842 |
| **MD_EARN_WNE_P10** | MD_EARN_WNE_P10 | 1 | 0.00000579 | 2.97251E-7 | 19.47 | <.0001 |

## Step 4:

**The error is assumed to have a normal distribution such that $N\sim(0,\sigma^2)$, with σ being the root mean square error/ Standard Deviation of the errors. Furthermore, the sum of the residuals should be 0.**

**The root mean square error of this model is 0.14047, with the Mean Square error being 0.01973. Therefore, the distribution of the error for this model is $N\sim(0, 0.01973)$.**

## Step 5:

**The equation of the distribution of the error is**

$$E(x)=\frac{1}{\sqrt{2\pi\,\sigma^2}}\;e^{-\frac{(x^2)}{2\sigma^2}}=\frac{1}{\sqrt{2\pi\,(0.14047)^2}}\;e^{-\frac{(x^2)}{2(0.14047)^2}}$$

**F-Test: Performing an F test on the model, the model is statistically significant, having a P value of less than 0.0001. Our threshold for a sufficient model is 0.05.**

**T test: T tests on the each of the betas shows that all but two of the betas (UGDS_ MEN, UGDS_ WOMEN) are statistically significant on the dependent variable (graduation rate) as p values of the betas are less than 0.05. In fact, the p values are less than 0.001.**

**R- Squared/ Adj R^2: The R^2 of the model is 0.5168 and the adjusted R^2 is 0.5159. Ideally, we would like to have these values to be 0.6 and above, but later through stepwise regression and other statistically techniques, we should expect the R^2 to increase.**

**Coefficient of Variation: Lastly, the C.V is about 28, which higher than 10, our threshold set for good model.**

## Step 6:

**Looking at the diagnostic plots, there's a couple things to glean from. Firstly, the residuals seem to be populated near the zero residual line, and mostly have a residual value between (-0.25, 0.25). The r-student plot shows that the model data is mostly within two SD's of the mean, with a couple of outliers. The residual histogram has a normal distribution.**

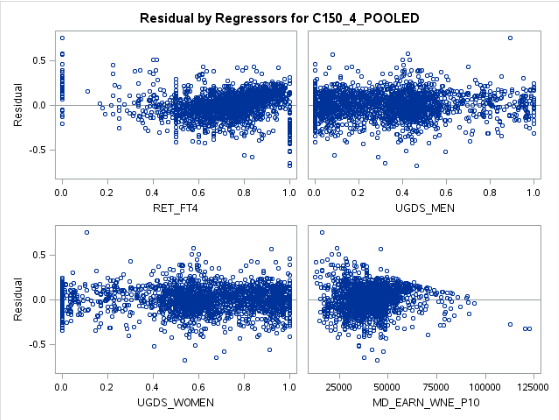
**7. None of the betas are equal to 0. But although this model has things to improve like insignificant betas, a high C.V, etc., such things can be correct through regression techniques like forward and backward selection. Conclusively The model is valid for future use.**

**B1= Retention Rate, B2= % of female students, B3=% of male students, B4= Graduate Earnings**

**The multi regression equation of the data is:**

$$\boldsymbol{Y = -0.139 + 0.581\,X_1 - 0.03\,X_2 + -0.007\,X_3 + 0.000005\,X_4}$$

**Out of all the independent variables, RET_FT4, which represents the retention rate has the greatest beta value. Additionally, it also has the largest t value. This evidence suggests that the College Retention rate is might be more significant on the dependent variable, the College graduation rate than the other independent variables.**



The residuals of the three variables are randomly scattered and show no trend. Mostly they're populated near the zero residual line which is good, because we want the difference of the predicted and the actual value to be as close as possible. However, the Retention variable, the

# Before Logistic Regression:

Before we perform the logistic regression, let’s see if there is any multicollinearity between the independent variable terms.

**Pearson Correlation Coefficients**
**Prob > |r| under H0: Rho=0**
**Number of Observations**

| | MD_EARN_WNE_P10 | RET_FT4 | C150_4_POOLED | UGDS_MEN | UGDS_WOMEN |
|---|---|---|---|---|---|
| MD_EARN_WNE_P10<br>MD_EARN_WNE_P10 | 1.00000<br>5682 | 0.38286<br><.0001<br>2031 | 0.50665<br><.0001<br>2288 | -0.07535<br><.0001<br>5682 | 0.09119<br><.0001<br>5682 |
| RET_FT4<br>RET_FT4 | 0.38286<br><.0001<br>2031 | 1.00000<br>2245 | 0.59841<br><.0001<br>2162 | -0.08068<br>0.0001<br>2245 | 0.07634<br>0.0003<br>2245 |
| C150_4_POOLED<br>C150_4_POOLED | 0.50665<br><.0001<br>2288 | 0.59841<br><.0001<br>2162 | 1.00000<br>2505 | -0.08071<br><.0001<br>2505 | 0.08113<br><.0001<br>2505 |
| UGDS_MEN<br>UGDS_MEN | -0.07535<br><.0001<br>5682 | -0.08068<br>0.0001<br>2245 | -0.08071<br><.0001<br>2505 | 1.00000<br>7544 | -0.45152<br><.0001<br>7544 |
| UGDS_WOMEN<br>UGDS_WOMEN | 0.09119<br><.0001<br>5682 | 0.07634<br>0.0003<br>2245 | 0.08113<br><.0001<br>2505 | -0.45152<br><.0001<br>7544 | 1.00000<br>7544 |

Note that all the p values from all the known interactions between the variables, are below 0.05. These p values are statistically significant. If a P value is greater than 0.05, there might be some multicollinearity happening between the two variables. But overall, all the independent variable seems to show no correlation or relationship between each other, which is a good thing. We will now use the VIF Ridge regression method to verify our findings.

**Parameter Estimates**

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| | Variance Inflation |
|---|---|---|---|---|---|---|
| Intercept | 1 | -0.13896 | 0.02145 | -6.48 | <.0001 | 0 |
| RET_FT4 | 1 | 0.58132 | 0.01907 | 30.48 | <.0001 | 1.18215 |
| UGDS_MEN | 1 | -0.02918 | 0.01848 | -1.58 | 0.1145 | 2.08614 |
| UGDS_WOMEN | 1 | -0.00692 | 0.01701 | -0.41 | 0.6842 | 2.08337 |
| MD_EARN_WNE_P10 | 1 | 0.00000579 | 2.97251E-7 | 19.47 | <.0001 | 1.18260 |

The VIF of each variable doesn't seem to be irregular and below 10. Multicollinearity doesn't seem to occur within the independent variables.

## Performing the logistic regression:

The dependent variable is the graduation rate, known as C150_4_POOLED. But this binary variable is actually going to be coded. The coded variable which will take the place of C150_4_POOLED, and will be known as Gcode. Gcode has two outcomes, 1 and 0. The condition of Gcode is as follows. We set our cutoff at 0.5 for graduation rate.

Gcode $= \begin{cases} 0, & Graduation\ rate < 0.5 \\ 1, & Graduation\ rate > 0.5 \end{cases}$

**Definitions In Context:**

True Positive (TP)- A college that has a graduation rate of above 0.5 is declared to be a good college- No Error

False Positive (FP)- A college that has a graduation rate of below 0.5 is declared a good college- Type 1 Error

True Negative (TN)- A college that has a graduation rate if below 0.5 is declared a poor college- No Error

False Negative (FN)- A college that has a graduation rate of above 0.5 is declared a poor college. Type 2 Error

**Recall that $\boldsymbol{Sensitivity = \frac{TP}{TP+FN}}$ and $\boldsymbol{Specificity = \frac{TN}{TN+FP}}$ .**

**Sensitivity is the amount of correct number of positive tests, the scenarios where a college has a graduation rate of above 0.5 is considered good, over the total number of tests. Specificity is the amount of correct negative tests, the scenarios where a college whose graduation rate is below 0.5 is declared poor, over the total amount of tests.**

**Logistic Regression General model:**

$$\mathbf{P(y = 1) = \pi = \frac{exp(B_0 + B_1\boldsymbol{X_1} + B_2\boldsymbol{X_2} + B_3\boldsymbol{X_3} + B_4\boldsymbol{X_4})}{1 + exp(B_0 + B_1\boldsymbol{X_1} + B_2\boldsymbol{X_2} + B_3\boldsymbol{X_3} + B_4\boldsymbol{X_4})}}$$

## Stepwise Regression-

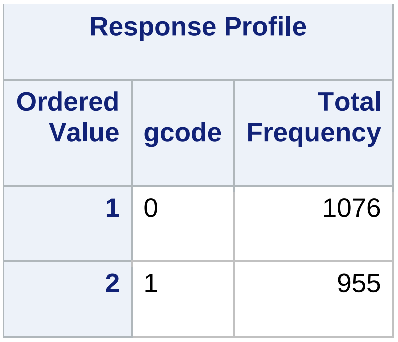
**After performing the logistic regression, we will need to perform stepwise techniques such as forward selection, backward selection and stepwise selection. Forward selection has the model initially start with 0 variables and variables are later added to the model, when SAS thinks it's significant. Backward selection is where the model has all the variables, and SAS removes the variables that are insignificant to the model. After performing forward, backward and stepwise selection with the logistic model, we will use the R^2 value, AIC, SIC to evaluate the best model.**

## Results:

### Logistic Regression-

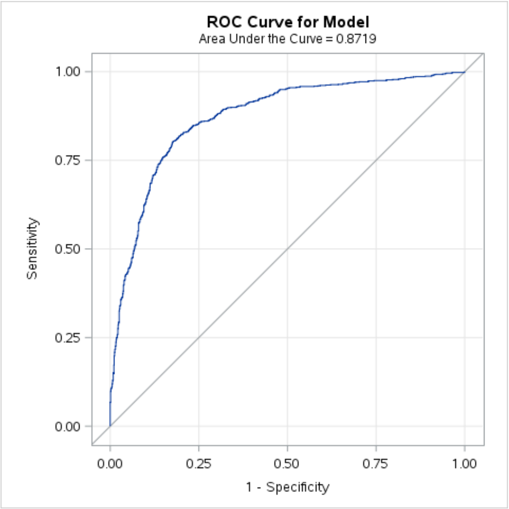
The total number of tests was 2031. The AUC or the area under the ROC Curve is about 0.87. This AUC value is a good AUC value as it is close to 1. But more importantly, the probability of choosing a true positive test against choosing a negative test is 0.8719.

$$\overline{\boldsymbol{Y}} = \frac{\boldsymbol{exp}(\,\mathbf{-9.8315+9.2064}\ \ \boldsymbol{X_1}\mathbf{-0.1784}\ \boldsymbol{X_2}\mathbf{+0.1328}\ \boldsymbol{X_3}\mathbf{+0.000074}\ \boldsymbol{X_4})}{\mathbf{1+exp(-9.8315+9.2064}\ \ \boldsymbol{X_1}\mathbf{-0.1784}\ \boldsymbol{X_2}\mathbf{+0.1328}\ \boldsymbol{X_3}\mathbf{+0.000074}\ \boldsymbol{X_4})}$$



Response Profile

| Ordered Value | gcode | Total Frequency |
|---|---|---|
| 1 | 0 | 1076 |
| 2 | 1 | 955 |



| Analysis of Maximum Likelihood Estimates | | | | |
|---|---|---|---|---|
| Parameter | Estimate | Standard Error | Wald Chi-Square | Pr > ChiSq |
| Intercept | -9.8315 | 0.5459 | 324.3194 | <.0001 |
| RET_FT4 | 9.2064 | 0.5433 | 287.1592 | <.0001 |
| UGDS_MEN | -0.1784 | 0.3303 | 0.2918 | 0.5891 |
| UGDS_WOMEN | 0.1328 | 0.3036 | 0.1912 | 0.6619 |
| MD_EARN_WNE_P10 | 0.000074 | 7.048E-6 | 109.1609 | <.000 |

## Stepwise methods- Logistic Regression: Choosing the best model

| Type of Model | Variables | AIC | SC | R Squared |
|---|---|---|---|---|
| Logistic | All Variables | 1901.434 | 1929.516 | 0.3633 |
| Logistic Forward | Retention, Med Earnings | 1899.120 | 1915.968 | 0.3628 |
| Logistic Backward | Retention, Med Earnings | 1899.120 | 1915.968 | 0.3628 |
| Logistic Stepwise | Retention, Med Earnings | 1899.120 | 1915.968 | 0.3212 |

## Additional Discussion of Results:

Interpretation of Betas in Logistic Model

**1) RET_4:**

$\mathbf{B_1 = 9.2064.} \qquad \boldsymbol{e^{B_1} - 1 = 9960.}$

**For every 1% increase in Retention, it's estimated that the odds of a good college are to increase by 996000% while keeping UGDS_WOMEN, ) UGDS_MEN, MD_EARN_WNE_P10 fixed. (This seems very unrealistic, but it does show that the retention variable has a lot of influence on Gcode).**

**2) UGDS_MEN:**

$\mathbf{B_2 = -0.1784.} \qquad \boldsymbol{e^{B_2} - 1 = -0.1633.}$

**For every 1% increase in the percentage of male students, it is estimated that the odds of a good college are to decrease by 16.33%, while keeping UGDS_WOMEN, RET_4, MD_EARN_WNE_P10 fixed.**

**3) UGDS_WOMEN**

$\boldsymbol{B_3 = 0.1328} \qquad \boldsymbol{e^{B_3} - 1 = 0.142.}$

**For every 1% increase in the percentage of female students, it is estimated that the odds of a good college are to increase by 14.20% while keeping UGDS_MEN, RET_4, MD_EARN_WNE_P10 fixed** .

**4) MD_EARN_WNE_P10:**

$\boldsymbol{B_4 = 0.00074} \quad \boldsymbol{e^{B_4} - 1 = 7.4\, x\, 10^{-5}.}$

**For every $1 increase in median graduate earnings, it is estimated that the odds of a good college are to increase by 0.0074%, while keeping UGDS_MEN, RET_4, UGDS_WOMEN fixed**

## Stepwise Regression

**Evaluating AIC, SC, and $R^2$, it would seem that either the Logistic Backward or the logistic forward selection model is the best. Why? Both have the lowest AIC and SC values. Although Logistic Stepwise model has the same AIC/ SC values as the forward and the backward method models, it has a lower $R^2$, compared to the two. Also, although one can see that the logistic model has the highest $R^2$, it is only superior in only one criteria. We are evaluating three.**

**Interaction term: Interestingly enough, retention and the median earnings variable were the only two variables in the forward/backward logistic model. it would be nice to experiment with an interaction term containing the two. The logistic model with only the retention variable, med earnings variable and interaction term (med earnings x retention) interestingly has an $R^2$ of 0.3950, AIC of 1795.61, an SC of 1818.08. Perhaps this is the best model, due to having the lowest AIC, SC value and the highest $R^2$, value.**

## Best model and the complete model:

**Best** model: (forward logistic model with retention x median earnings interaction term):

$B_0$ = intercept, $B_1$ = retention , $B_2$ = Median Graduate earnings, $B_3$ = interaction term

$$\boldsymbol{\pi}^* = \mathbf{B_0} + \mathbf{B_1}\boldsymbol{X_1} + \mathbf{B_2}\boldsymbol{X_2} + \mathbf{B_3}\boldsymbol{X_3}$$

$$\boldsymbol{\pi}^* = \mathbf{5.9612} + \mathbf{11.4474}\,\boldsymbol{X_1} + \mathbf{0.000379}\,\boldsymbol{X_2} + \mathbf{-0.00059}\,\boldsymbol{X_3}$$

**Complete** model: $B_0 =$ intercept, $B_1$ = retention, $B_2$ = Median Graduate earnings,

$B_3$= % of men students, $B_4$ = % of women students

$$\boldsymbol{\pi}^* = \mathbf{B_0} + \mathbf{B_1}\boldsymbol{X_1} + \mathbf{B_2}\boldsymbol{X_2} + \mathbf{B_3}\boldsymbol{X_3} + \mathbf{B_4}\boldsymbol{X_4}$$

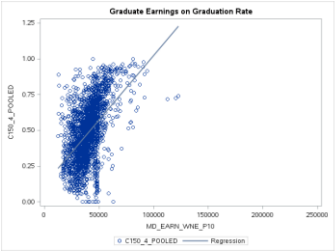
$$\boldsymbol{\pi}^* = \mathbf{9.8315} + \mathbf{9.2064}\ \boldsymbol{X_1} - \mathbf{0.1784}\,\boldsymbol{X_2} + \mathbf{0.1328}\ \boldsymbol{X_3} + \mathbf{0.000074}\,\boldsymbol{X_4}$$

## Conclusion:

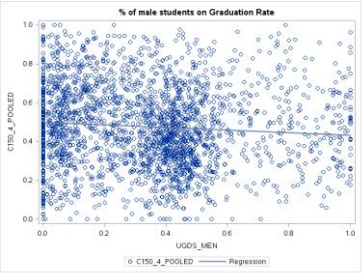
Referring back to beginning, based on the results, we can answer each hypothesis. **Firstly, if the salary after attending is high, does that mean that the graduation rate is also high**? Yes. A single regression between salary and graduation rate shows a positive regression line, and between $75-$100K in terms of median graduate earnings, the graduation rate is 0.8 or higher. **If the retention of students increases, does that mean the graduation rate is bound to increase**? Yes. Performing a singular regression shows a positive regression line and when the retention is between 0.8-1.0 the graduation rate is 0.6-1.

**Lastly, the two other hypothesis (refer to beginning) are actually incorrect according to the data.** The single regression line on the percentage of men vs the graduation rate reveals a negative regression line, and when the % of men students is between 0.8 and 1 (all men's college), the graduation rate is approx. 0.4 according to the regression line. On the other hand, when doing a single regression with % of women vs graduation rate, the regression line is a positive regression line, but when the percentage of women is between 0.8-1, the graduation rate is 0.45. While these questions are answered, the bigger question of this data analysis is: *what variable(s) influenced the data set the most?*

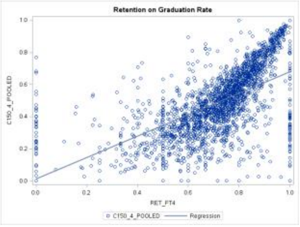
Selection methods reveal that the **retention and the median earnings** were the variables that weren't cut off and remained in the last model of each selection (forward, backward, stepwise). Plus, in the multi regression, they had the highest beta values, t values and were statistically significant on the graduation rate variable. Thus, these variables mostly likely had a higher weight on the multi regression and logistic stepwise models as a whole. But could these variables be the top factors that might sway a HS student from one college to the other? Possibly.



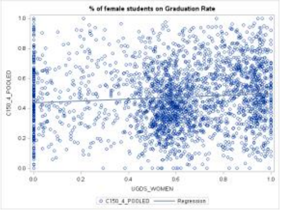
Median Earnings vs Graduation Rate



% of male student's vs Graduation Rate



Retention rate vs Graduation Rate



% of female students vs Graduation Rate

# References:

Fu, Brian. "Article Title." *College Scorecard.* Office of Planning, Evaluation, and Policy Development, September 28, 2015, Web. December 8, 2017
https://catalog.data.gov/dataset/college-scorecard

# Appendix:

```
proc import datafile='/folders/myfolders/sasuser.v94/Project1/ProjectDataZ.xlsx'
dbms= xlsx

out=ProN;

/*proc contents data=ProN; */

proc means data=ProN mean std clm;

var RET_FT4;

var C150_4_POOLED;

var MD_EARN_WNE_P10;

var UGDS_MEN;

var UGDS_WOMEN;

run;

/*Finding the univariate data */

proc Univariate data=ProN;

var RET_FT4;

var C150_4_POOLED;

var MD_EARN_WNE_P10;

var UGDS_MEN;

var UGDS_WOMEN;

run;
```

```
/* generating the histograms for all variables*/
proc Univariate data=ProN;
histogram RET_FT4;
histogram C150_4_POOLED;
histogram UGDS_MEN;
histogram UGDS_WOMEN;
histogram MD_EARN_WNE_P10;
run;
/* When you perform the multi regression*/
proc reg data=ProN;
model C150_4_POOLED= RET_FT4 UGDS_MEN UGDS_WOMEN
MD_EARN_WNE_P10;
ods graphics on;
run;
/* Finding the scatter and best fit line of the single regression models */
proc sgplot data=ProN;
title 'Retention on Graduation Rate ';
scatter X= RET_FT4  Y=C150_4_POOLED ;
reg X = RET_FT4  Y=C150_4_POOLED;
run;
proc sgplot data=ProN;
title 'Graduate Earnings on Graduation Rate ';
scatter X= MD_EARN_WNE_P10 Y=C150_4_POOLED ;
reg X = MD_EARN_WNE_P10 Y=C150_4_POOLED;
```

```
run;
proc sgplot data=ProN;
title '% of male students on Graduation Rate ';
scatter X= UGDS_MEN Y=C150_4_POOLED ;
reg X = UGDS_MEN Y=C150_4_POOLED;
run;
proc sgplot data=ProN;
title '% of female students on Graduation Rate ';
scatter X= UGDS_WOMEN Y=C150_4_POOLED ;
reg X = UGDS_WOMEN Y=C150_4_POOLED;
run;

/*Logistic Regression*/
proc corr data=ProN;
/*analysing Multicollinerity thru Proc Corr*/
proc reg data= ProN outvif outest=ProNest;
/*analysing MultiCollinearity thru VIF*/
model C150_4_POOLED= RET_FT4 UGDS_MEN UGDS_WOMEN
MD_EARN_WNE_P10 /vif ridge=(0.01 0.02 0.03 0.04 0.05 0.06 0.07 0.08 0.09 0.1
1.1 1.2 1.3 1.4 1.5);;
run;
/*coding the dependent variable*/
data ProN2;
set ProN;
if C150_4_POOLED > 0.50 then gcode=1;
```

```
else gcode=0;
run;
PROC LOGISTIC data= ProN2 alpha=.05
plots(only)= (effect oddsratio roc);
model gcode (event= '1') = RET_FT4 UGDS_MEN UGDS_WOMEN
MD_EARN_WNE_P10
/ clodds=wald clparm=wald Rsquare;
run;
/*forward, backward, stepwise selection on logistic model*/
proc LOGISTIC data=ProN2;
model gcode = RET_FT4 UGDS_MEN UGDS_WOMEN MD_EARN_WNE_P10
/selection=forward Rsquare;
run;
proc LOGISTIC data=ProN2;
model gcode = RET_FT4 UGDS_MEN UGDS_WOMEN MD_EARN_WNE_P10
/selection=backward Rsquare;
run;
proc LOGISTIC data=ProN2;
model gcode = RET_FT4 UGDS_MEN UGDS_WOMEN MD_EARN_WNE_P10
/selection=stepwise Rsquare;
run;
/*experiment with interaction term*/
data ProN3;
set ProN2;
Interaction= RET_FT4*MD_EARN_WNE_P10;
```

```
proc logistic data=ProN3;
model gcode= RET_FT4 MD_EARN_WNE_P10 Interaction / Rsquare;
run;
```