Background

In this project, our client, an Online Education Service Provider which offers certificate programs, was looking for a software application which could track their student's success post-graduation offering their graduate students surveys to measure their success to use within their Marketing Campaigns.

The institution offered training to excel in various examinations (i.e., licensing, standardized tests, etc.) To know whether or not students would perform better at a standardized test after completing one of their preparatory courses, a group of 65 students were given a test before and then again after having gone through their training. I was provided with their results and tested their claim at 5% level of significance (please reference the excel spreadsheet "Training Prep Course").

Null Hypothesis H0: $\text{mu\_d} \leq 0$
Alternate Hypothesis Ha: $\text{mu\_d} > 0$

The next component to the analysis was to assess the distribution on stipends at the institution. Our client was concerned that there might be a stipend offering gap based on gender. They cited that there were many studies that showed women earned less than men. This, however, remained a very hot topic across industries with many people having criticized the studies they were siting, ignoring factors that could be the reason behind gender pay gap (i.e., women working part time, women being less represented in high-paying jobs in the tech industry, etc.).

Our client wanted to make sure that for their Marketing Campaign their graduate students who had "assistant" positions at their institution were treated fairly without a gender bias. The stipend (known as "assistantship") the teaching and research assistants received consisted of a tuition waiver plus some cash amount, available to the graduate student for up to six years.

Our client conducted a study and provided data on 1,000 graduate student assistants that they had compiled. For each student, the following information was recorded:

- Assistantship (Stipend): Annual stipend in dollars;
- Gender: F, female; or M, male;
- Field of Study: S, science, including the natural sciences and engineering; A, arts, including both liberal and fine arts, and humanities;
- Year of Study: year of the student's graduate studies at the beginning of their contract period;
- Marital Status: M, married; or S, single

At first, I completed an analysis to see if the there was a gender gap in pay. The following table was the result of this analysis:

t-Test: Two-Sample Assuming Unequal Variances		
	Male Assistantship (Stipend)	*Female Assistantship (Stipend)*
Mean	25532.27985	24421.446
Variance	1343904.439	1109965.485
Observations	482	518
Hypothesized Mean Difference	0	
df	971	
t Stat	15.81913563	
P(T<=t) one-tail	1.26433E-50	
t Critical one-tail	1.646424413	
P(T<=t) two-tail	2.52866E-50	
t Critical two-tail	1.962410103	

However, there was expressed concern that my first analysis was too simplistic since a pay gap could be due to the field of study. So, I conducted a second analysis on the field of study which I had the following results:

t-Test: Two-Sample Assuming Unequal Variances		
	Assistantship (Stipend) For ART	*Assistantship (Stipend) for Science*
Mean	24041.62203	25952.36501
Variance	621483.0801	615193.5853
Observations	521	479
Hypothesized Mean Difference	0	
df	992	
t Stat	-38.39036076	
P(T<=t) one-tail	1.1775E-198	
t Critical one-tail	1.646391129	
P(T<=t) two-tail	2.3551E-198	
t Critical two-tail	1.962358258	

There was still much uncertainty about any conclusion that could be made about the issue of gender gap in student compensation based on the results of studies that only focused on one of the characteristics of the student body at a time and did not consider the impact of other variables.

I was asked to help in bringing more clarity to this question. In my Initial, first step I ran a multiple linear regression analysis and determined that the four combined predictor variables: **Years of Study**, **Gender**, **Field of Study**, and **Marital Status** explained the variations in the assistantship (stipends) through its Adjusted R^2 of 82.98%, making this a very good predictive model. I then looked at the p-values of each predictor and discovered that, with the exception of **Marital Status** (which happened to have a p-value of 0.7626), all were considerably lower than either 0.01 or 0.05 levels of significance, demonstrating that **Years of Study**, **Gender**, and **Field of Study** had a highly significant relationship with assistantship (stipends).

Next, I decided to omit the predictor: **Marital Status** (due to its high p-value and low relationship to assistantship/stipends) and re-ran my multiple linear regression analysis with only the three predictor

variables: **Years of Study**, **Gender**, and **Field of Study**. My Adjusted R^2 of 82.99%, increased ever so slightly. And all of the p-values for each of these independent variables remained below the alpha level of significance.

In the end, my analysis contradicts the original belief that women earn less than men when it pertains to gender roles and pay. Based on my analysis for our client, graduate students who had "assistant" positions saw female assistants receiving approximately $770 more than their male colleagues in assistantship (stipends). Moreover, students, whether male or female, enrolled in a Science Field of Study earned approximately $2,535 more in stipends than their Art counterparts.

		1-YEAR OF STUDY			
		FEMALE SCIENCE	**MALE SCIENCE**	**FEMALE ART**	**MALE ART**
	Coefficients				
Intercept	24669.32789	1	1	1	1
Year of Study	406.3940826	1	1	1	1
Gender	769.6133725	1	0	1	0
Field Of Study	-2534.642018	0	0	1	1
Assistantship (Stipend)		**$ 25,845.34**	**$ 25,075.72**	**$ 23,310.69**	**$ 22,541.08**

In conclusion, I learned that using linear regression analysis, as compared to the two-sample t-tests that I used on my original study, was a better predictive model because it allows for multiple predictor variables to be analyzed, simultaneously, in relation to a response variable, in order to determine their level of influence (in my case the amount of stipends to be received). Overall, our client should have had some concern and awareness regarding gender gap when it came to compensation of its graduate assistants. More importantly, they realized that there existed a great disparity in awarding stipend amounts between students in a science field versus ones studying the arts. An avenue I recommended our client should explore, to further study and/or address the gap attributed due to gender and/or field of study when awarding stipends, was to perhaps include a predictor variable which accounted for the number of applicants based on gender for science versus art. It could've been that fewer applicants were applying to one program over another. A second factor to consider, was the acceptance rate into either program/field of study based on gender. It could've simply been that it was more difficult to get admitted into one program over the other which might've explain the disproportion of assistantships (stipends) awarded for either field of study.

INITIAL SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.911320442
R Square	0.830504948
Adjusted R Square	0.829823561
Standard Error	510.2459532
Observations	1000

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	4	1269309173	317327293.3	1218.844465	0
Residual	995	259049178.1	260350.9328		
Total	999	1528358351			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	24671.79799	44.5034718	554.3791751	0	24584.46656	24759.12942
Year of Study	404.6772711	13.8519017	29.21456416	5.5668E-136	377.4949775	431.8595646
Gender	770.1602668	51.69665314	14.89768138	1.84465E-45	668.7132863	871.6072474
Field Of Study	-2534.94402	51.50886975	-49.21373799	8.0647E-269	-2636.022504	-2433.865536
Martial Status	13.88502043	45.95585875	0.3021382	0.762609855	-76.29650633	104.0665472

REFINED SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.91131191
R Square	0.830489398
Adjusted R Square	**0.829978823**
Standard Error	510.0131356
Observations	1000

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	1269285407	423095135.5	1626.579553	0
Residual	996	259072944.9	260113.3985		
Total	999	1528358351			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	24669.32789	43.72614973	564.1779129	0	24583.52194	24755.13384
Year of Study	406.3940826	12.62703894	32.18443252	**2.2543E-156**	381.61543	431.1727352
Gender	769.6133725	51.64138087	14.90303628	**1.71118E-45**	668.2749795	870.9517656
Field Of Study	-2534.642018	51.4756722	-49.23960989	**4.2787E-269**	-2635.655232	-2433.628803