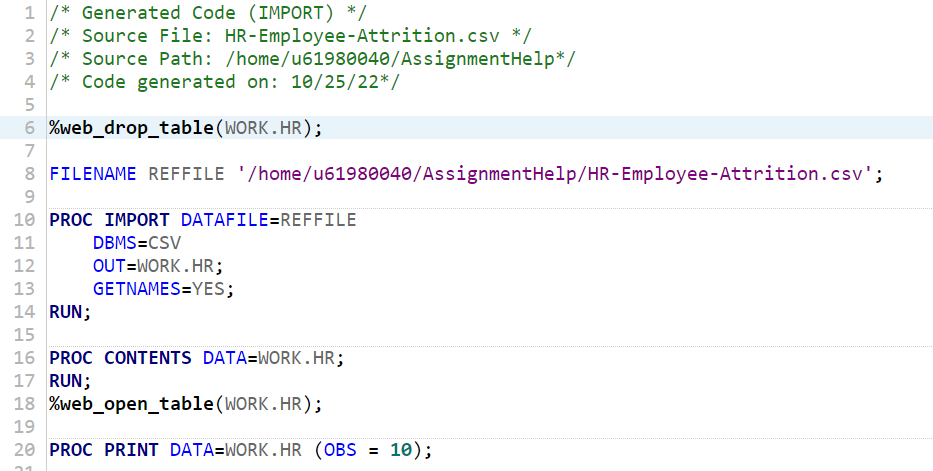
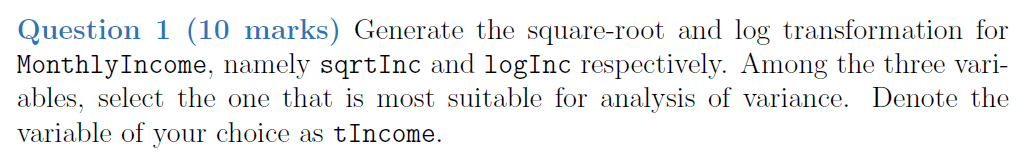
**Statistics for Data Sciences**

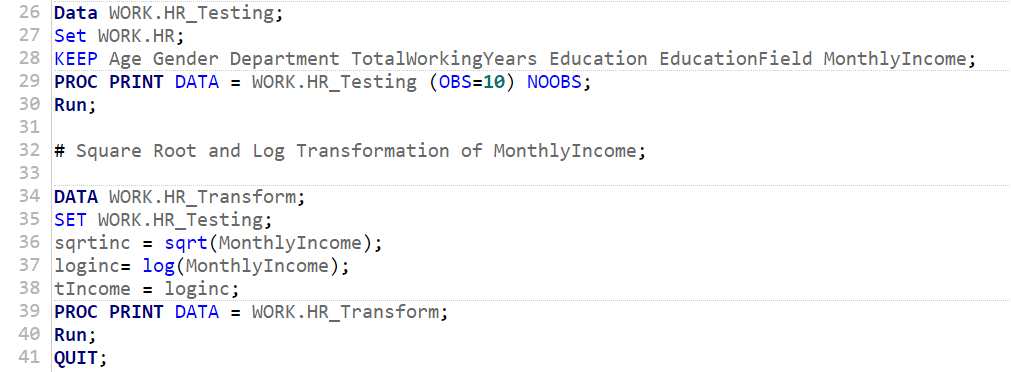
Import Data:

****

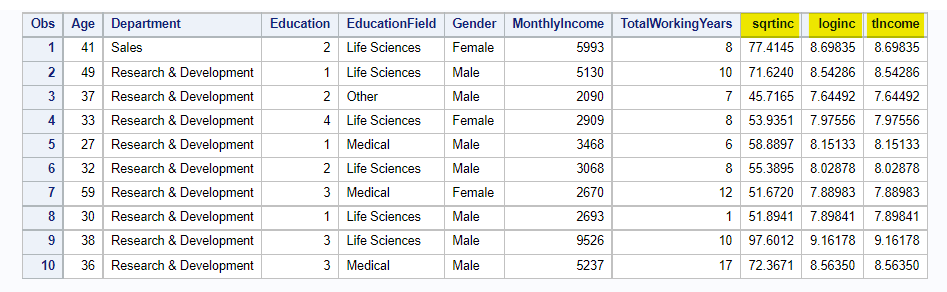
Answer 1:



Required Code for Answer 1:

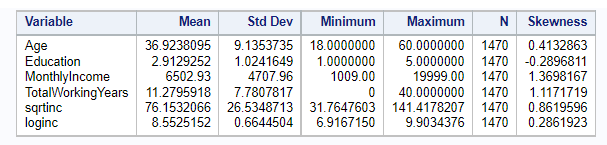


Output:

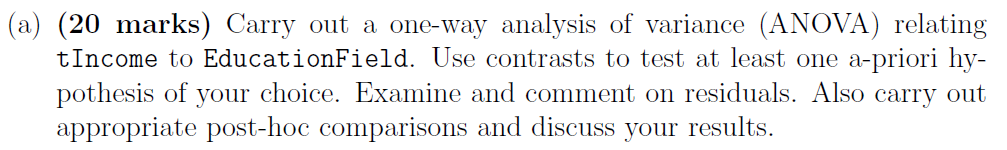


In this part, I extract the variables of interest from the original dataset for our case study. After that, the transformation was done for the variable that is generating the square-root and log transformation for MonthlyIncome, namely sqrtInc and logInc respectively (see above sas code line 36 and 37). Among the three variables, I have selected the one that is loginc for the reason that it is most suitable for analysis of variance. The skewness statistic (see below) tells that it is fairly symmetric among the three.

Statistical Summary of the variables:



Answer 2 (a):

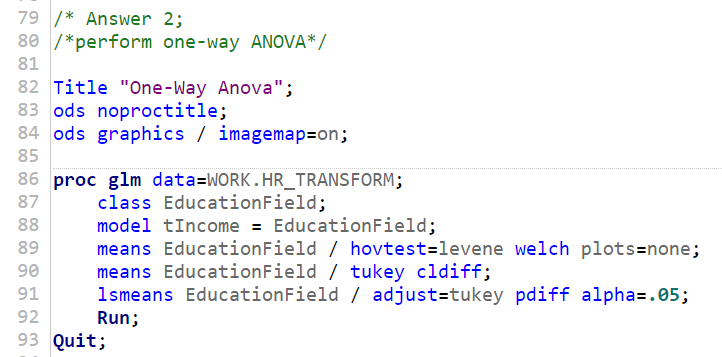


Hypothesis:

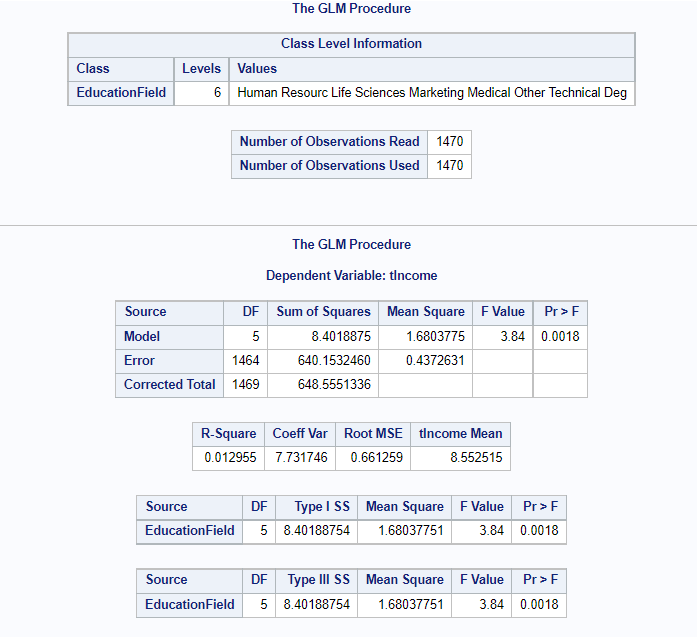
Null-Hypothesis(H0): Both variables are not significant i.e., tIncome & EducationField.

Alternative-Hypothesis(H1): Both variables have significant difference between tIncome & EducationField.

Code required for One-Way Analysis of Variance (ANOVA):



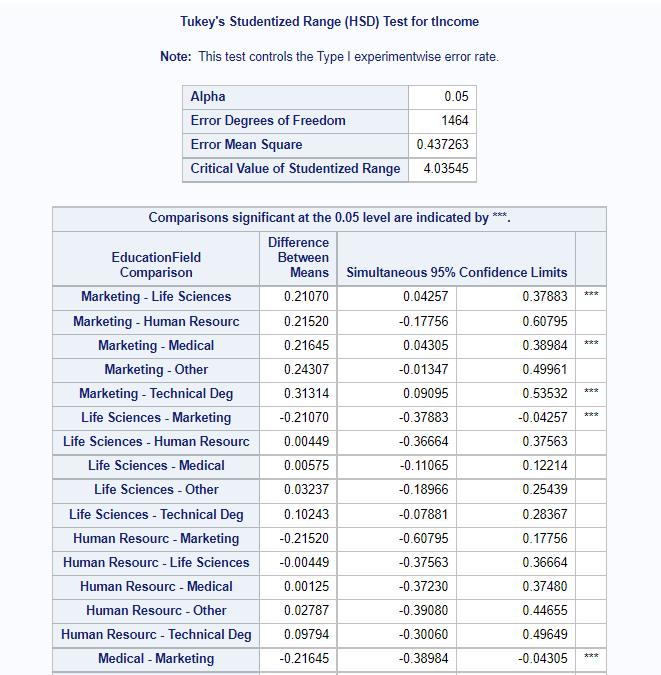
Output of One-way ANOVA:



The p-value is close to zero and significance level or alpha value is 0.05. It implies that the p-value is less than the level of significance. Thus, the null hypothesis can be rejected and in support of an alternative hypothesis. Therefore, there is a significant difference between tIncome and EducationField.

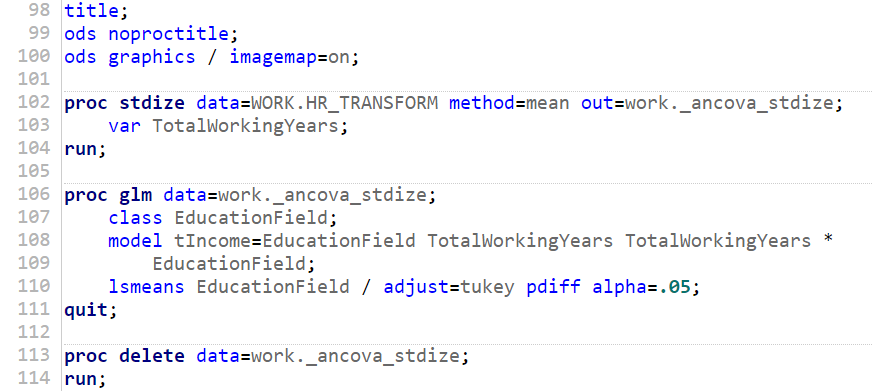
(b) Post Hoc Test:



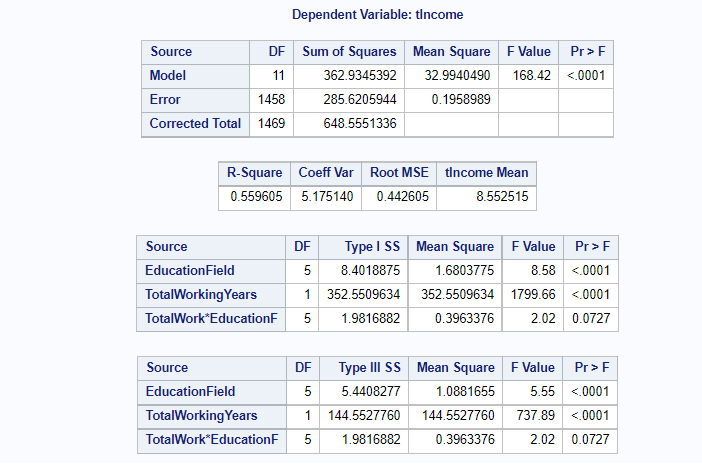


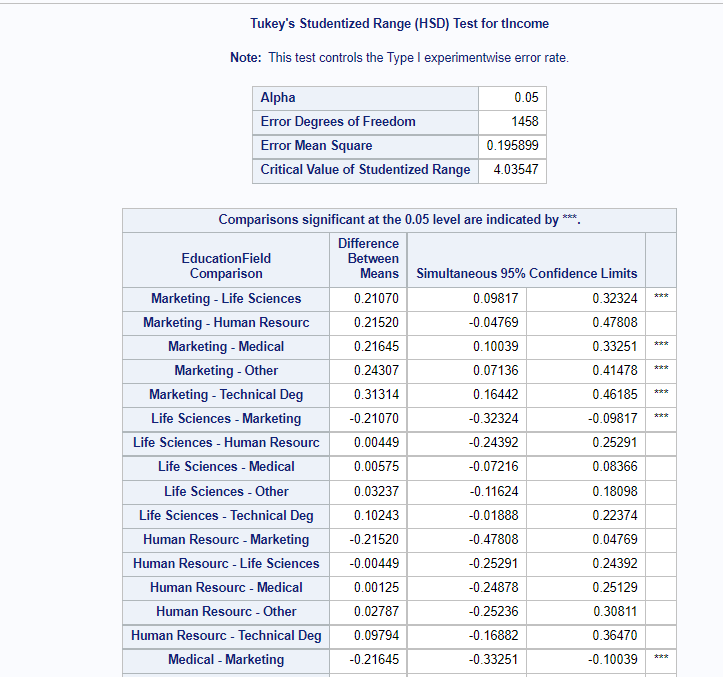
The portion (a) post hoc test states that the test is legitimate because the p-value is less than the level of significance, which denotes that there is a statistically significant difference between the groups and the mean. In simple words, this tells us that the tIncome is not equal between the level of education filed.

(c) Required SAS code for ANCOVA Analysis:

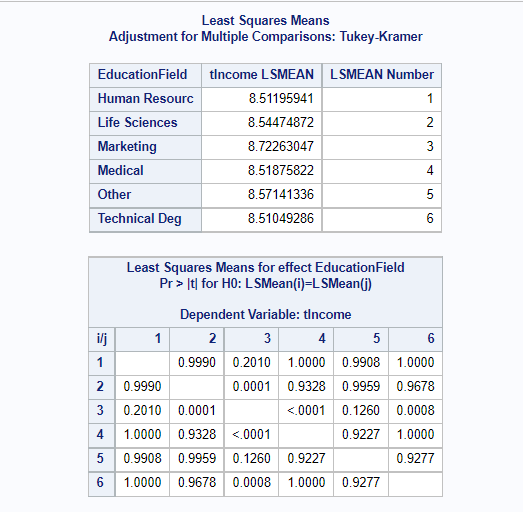


Output:

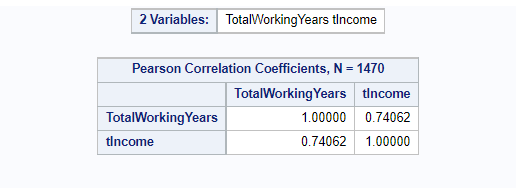








Correlation Analysis:



**Explanation**: The null hypothesis can be rejected since the p-values for the TotalWorkingYears and EducationField variables are less than 0.00 and the level of significance is 0.05, respectively. As a result, the null hypothesis cannot be accepted. However, because the interaction term's p-value is greater than 0.05, it is not statistically significant that the two variables interact.

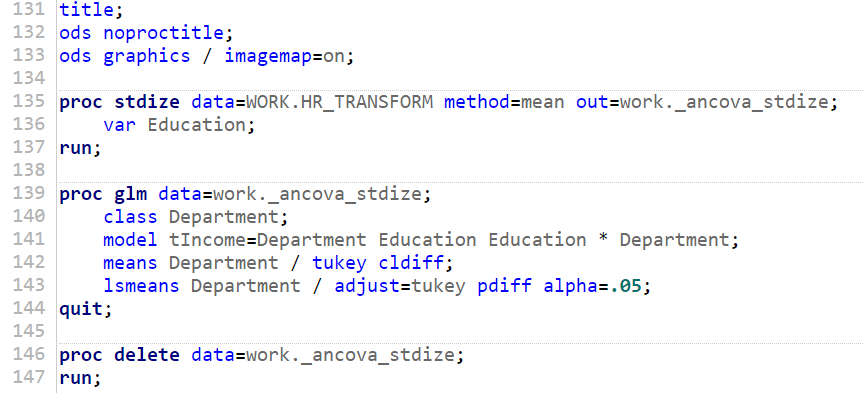
On the other hand, the post hoc test is valid for the reason that the p-value of the test is less than the level of significance which implies that there is a significant difference between the groups and the mean are statistically significant. In simple words, this tells us that the tIncome is not equal between the level of education filed. Additionally, the groups of education field i.e., marketing with life sciences, medical, technical deg, others are different.

It can be seen from the above result that there is a correlation of 0.74 between the TotalWorkingYears and the tIncome, which suggests that the association is both favorably strong and that the slope of the regression will be positive.

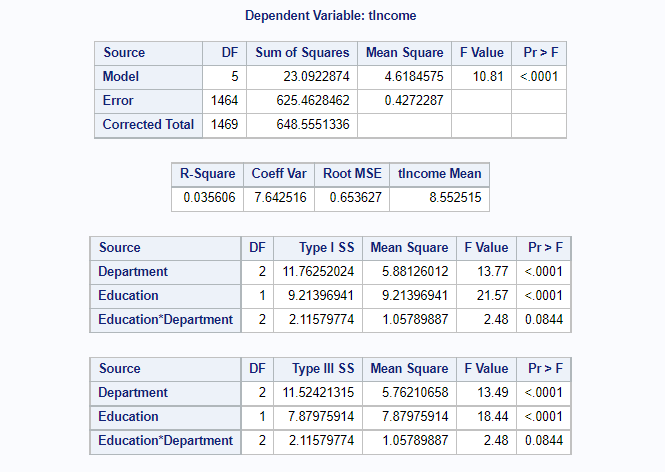
On the other hand, the R-square of the simple regression model is 0.56 (or 56%), it indicates that the 56% of the variation in the tIncome (dependent variable) can be explained in the TotalWorkingYears and EducationField (independent variables).

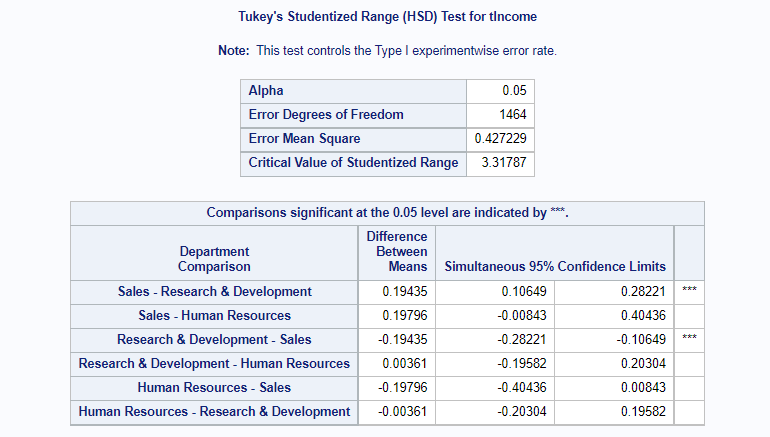
Answer 3:

Required for ANOVA model:



Output of the ANOVA model:





**Explanation**: Because the Department and Education variables' p-values are less than 0.00 and the level of significance is 0.05, the null hypothesis can be rejected because neither variable's p-value exceeds the alpha level. However, because the interaction term's p-value is greater than 0.05, it is not statistically significant that the two variables interact.

The post hoc test, however, is legitimate because the p-value of the test is less than the level of significance, which denotes that there is a statistically significant difference between the groups and the mean. Simply put, this indicates that there is a disparity in income between educational levels. The groups of departments, such as sales and research and development, are also distinct.

On the other hand, the R-square of the simple regression model is 0.04 (or 4%), it indicates that the 4% of the variation in the tIncome (dependent variable) can be explained in the Department and Education (independent variables). Thus, the R-square of the model does not appropriately explain the variation.

Moreover, the overall model is statistically significant because the p-value of the model is less than the level of significance i.e., 0.00 < 0.05.

Answer 4: In answer 1, transform a data set such as generate the square-root and log transformation for MonthlyIncome, namely sqrtInc and logInc respectively.

In the next answer, the p-value of the test is close to zero, so in that case null hypothesis is rejected. Thus, there is a significant difference between tIncome and EudcationField. In the next part, if the assumptions for ANOVA are satisfied, do not use a non-parametric method.

In part (c), the p-value of the TotalWorkingYears and EducationField variables is less than the 0.00, so the null hypothesis can be rejected. However, the interaction term of both the variables is not statistically significant. Additionally, there is a strongly positive relationship and the regression slope will be positive between the tIncome and TotalWorkingYears. Moreover, the r-square of the model is good i.e., 56%.

The null hypothesis can be rejected in response 3 since the p-values for the Department and Education variables are less than 0.00. However, there is no statistically significant interaction between the two variables. The r-square value, which is 4%, is therefore low or inadequate for the simple regression model.