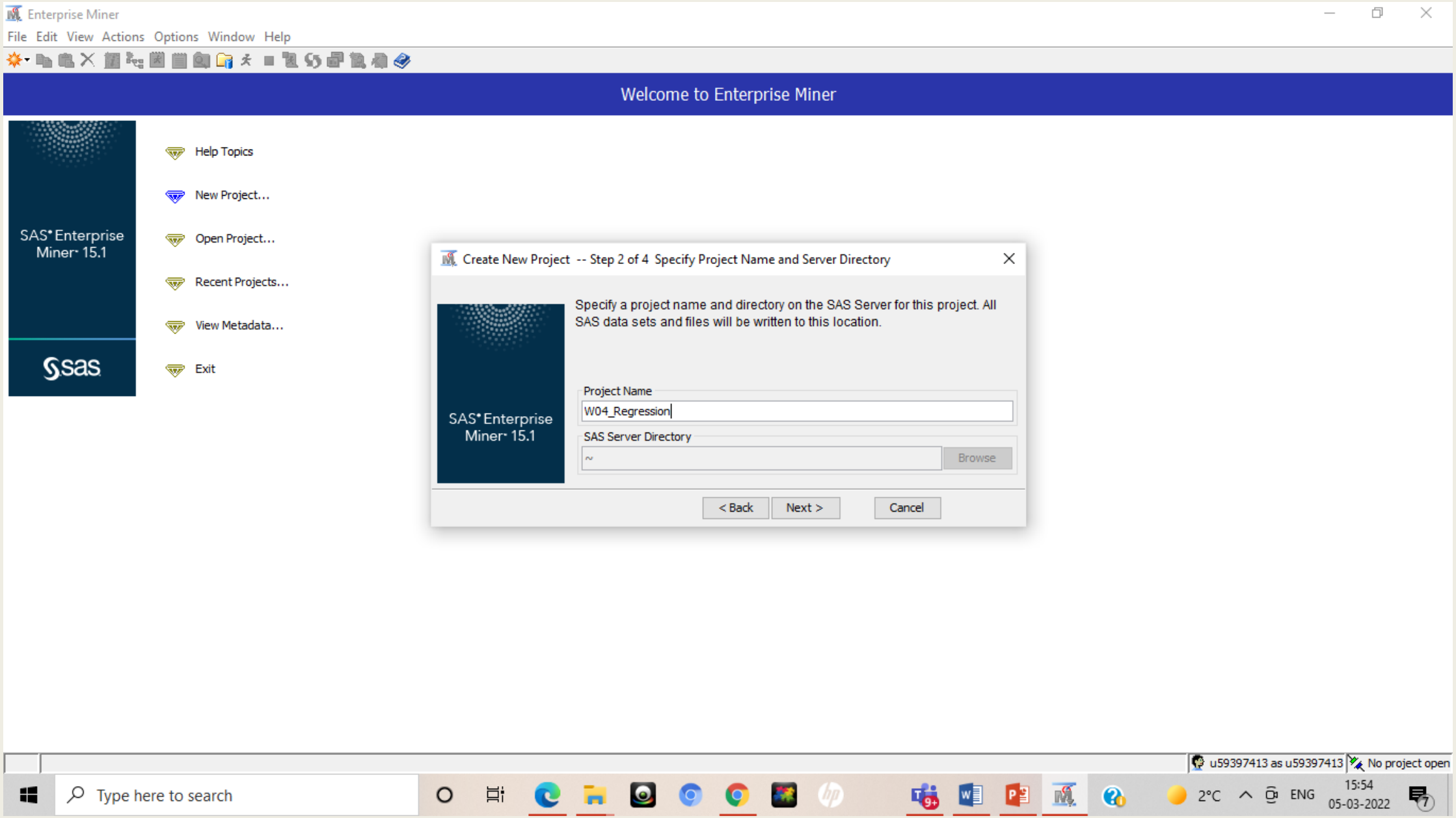
WORKSHOP 4

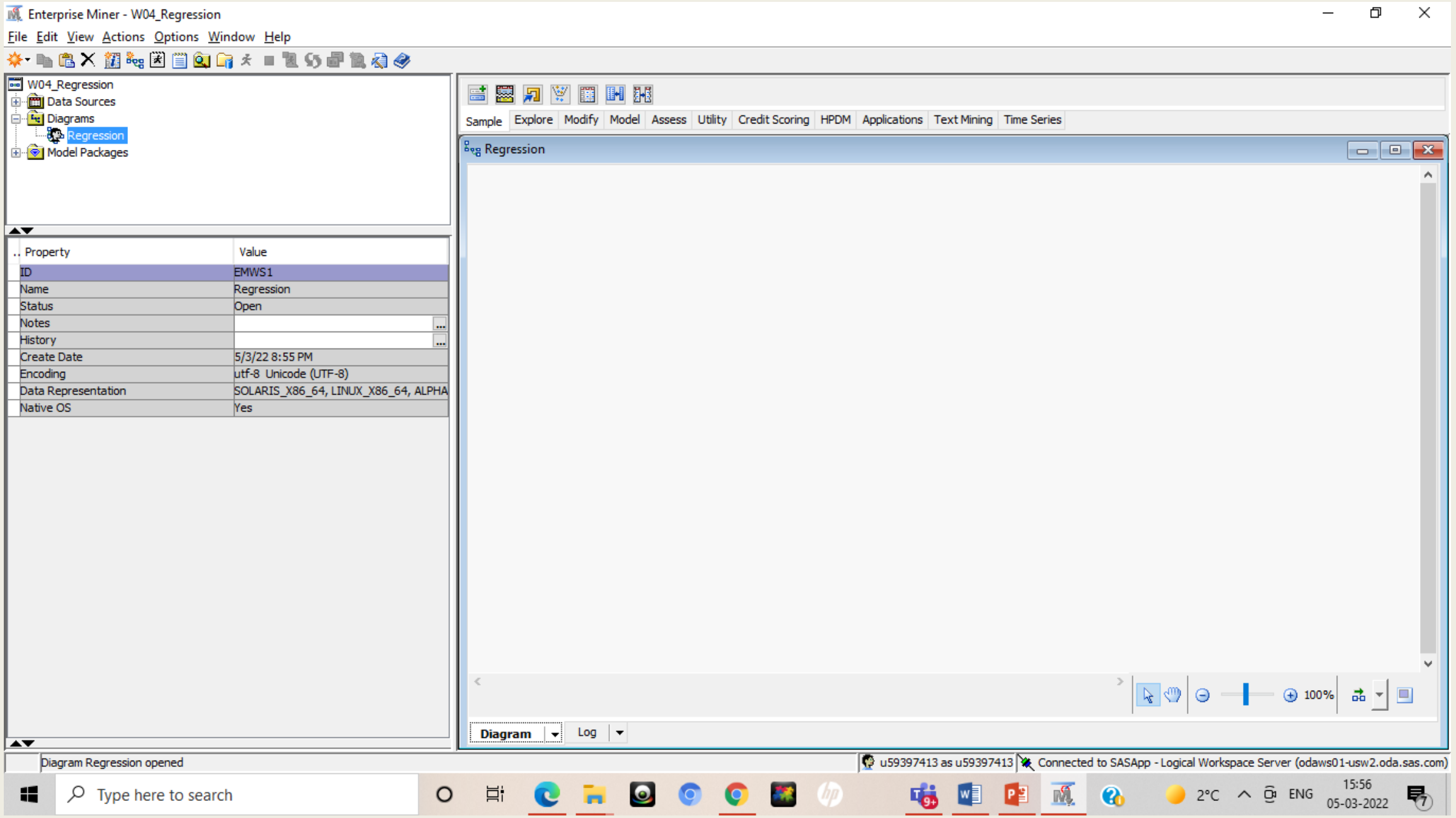
By Group 5-
Anand Mohan Thakur
Josh Shaji
Poonam Bhaliyan
Poornima Singh
Prateek Ramjanam Singh

PART I: LINEAR REGRESSION

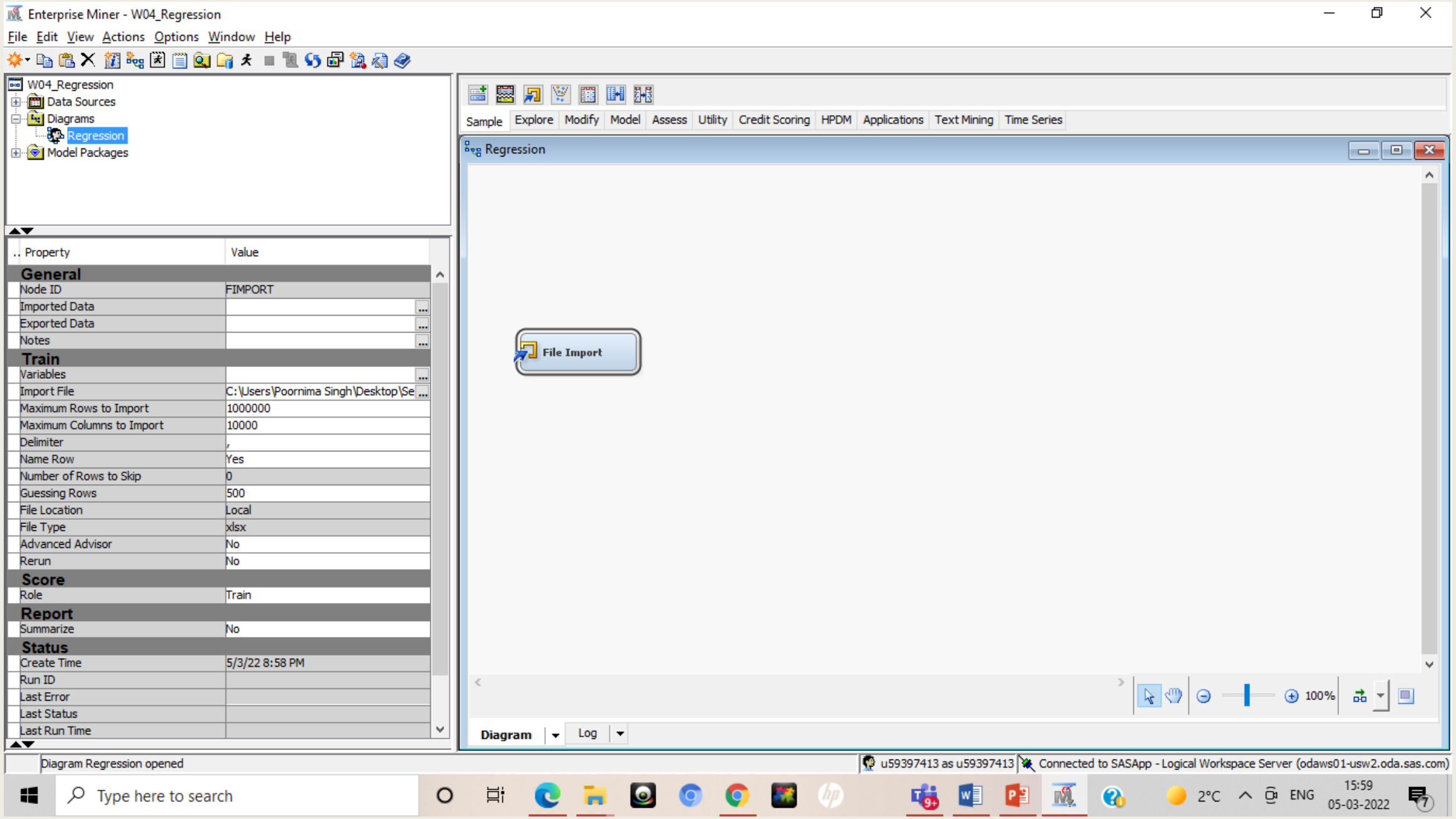
Answer 1: Start a new project named *W04_Regression*



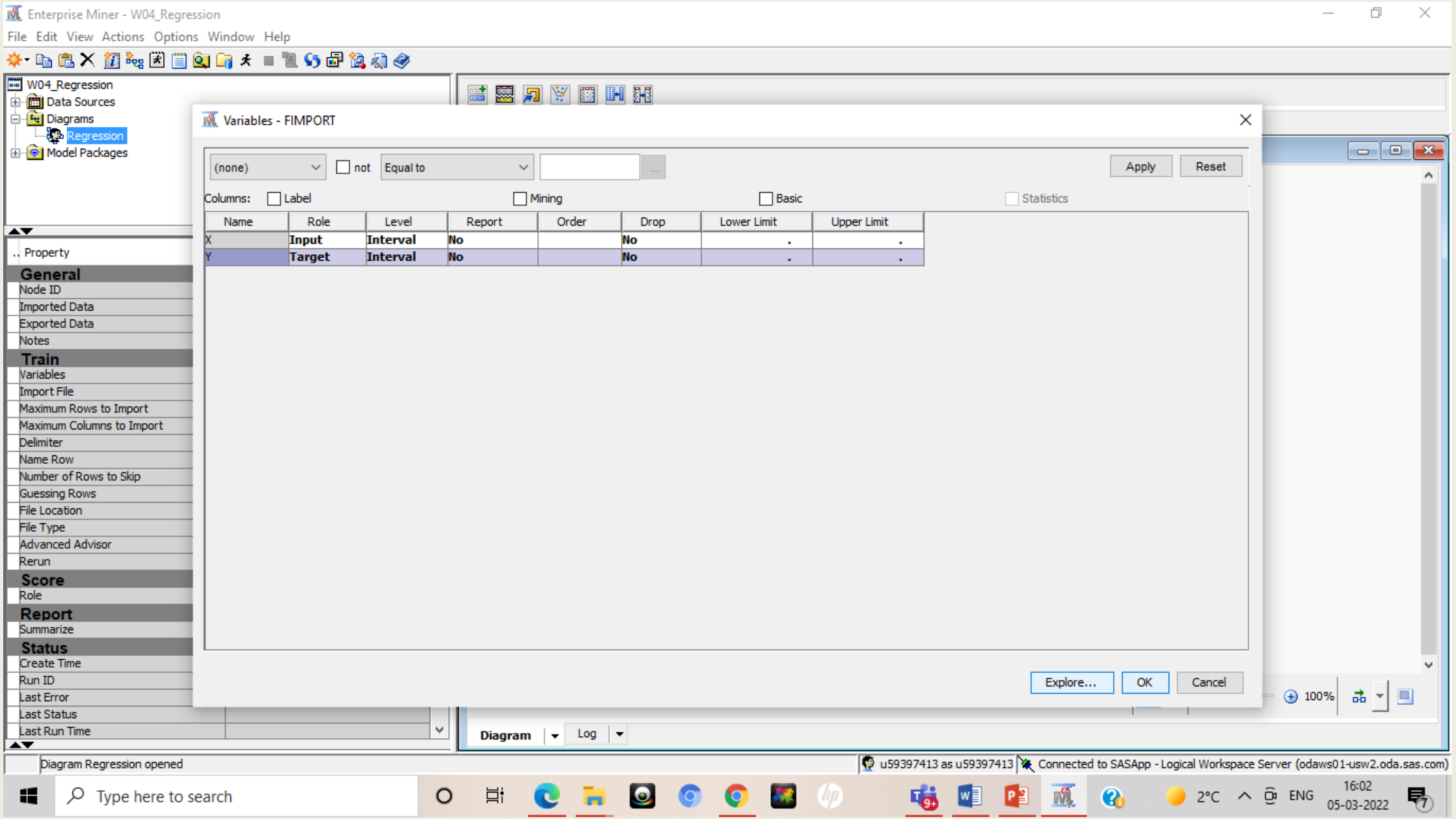
Answer 2: Create a diagram and name it as *Regression*



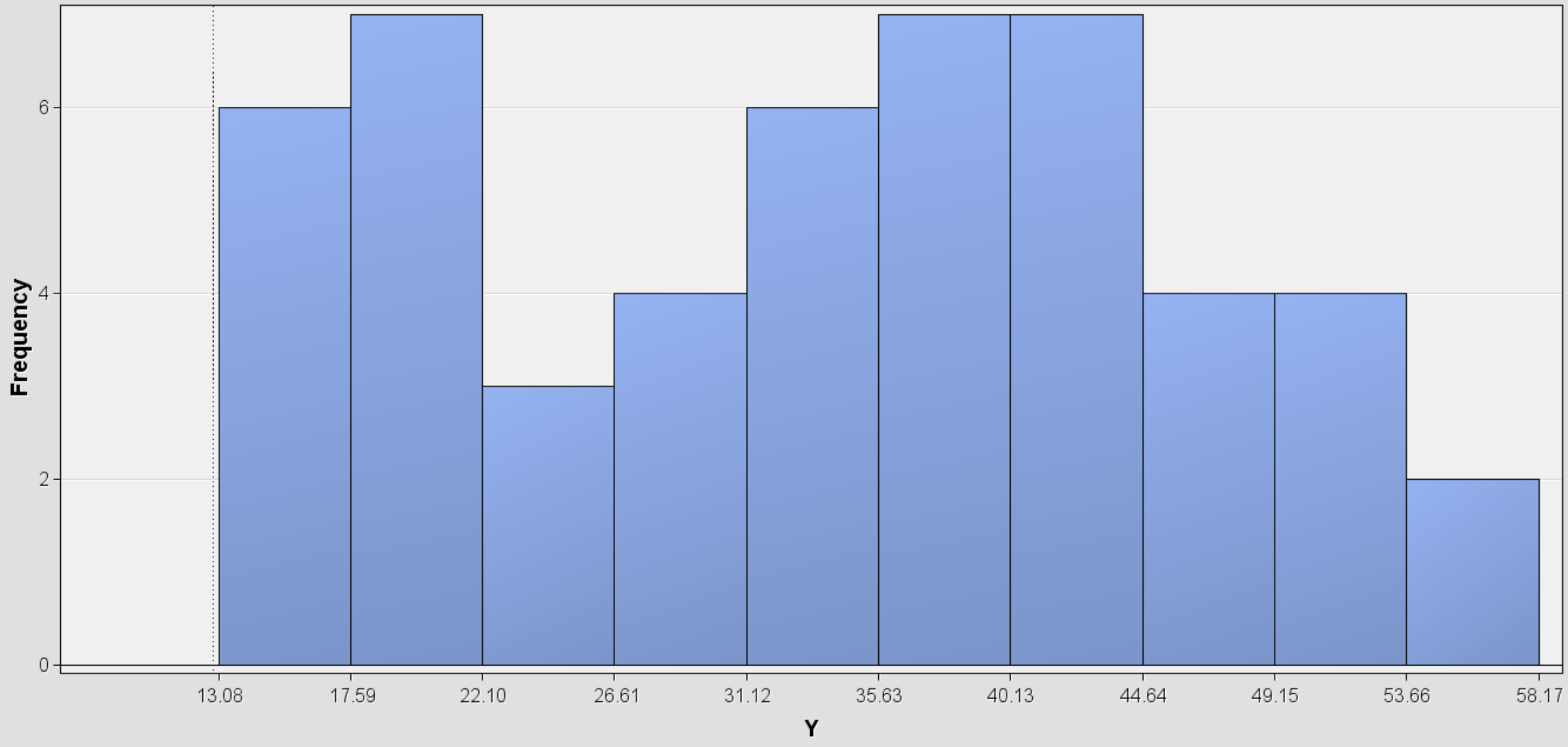
Answer 3: Importing the file W04_2D.xlsx using File Import



Answer 4: Click on Y. Choose Target as the Role for Y. Then click Explore

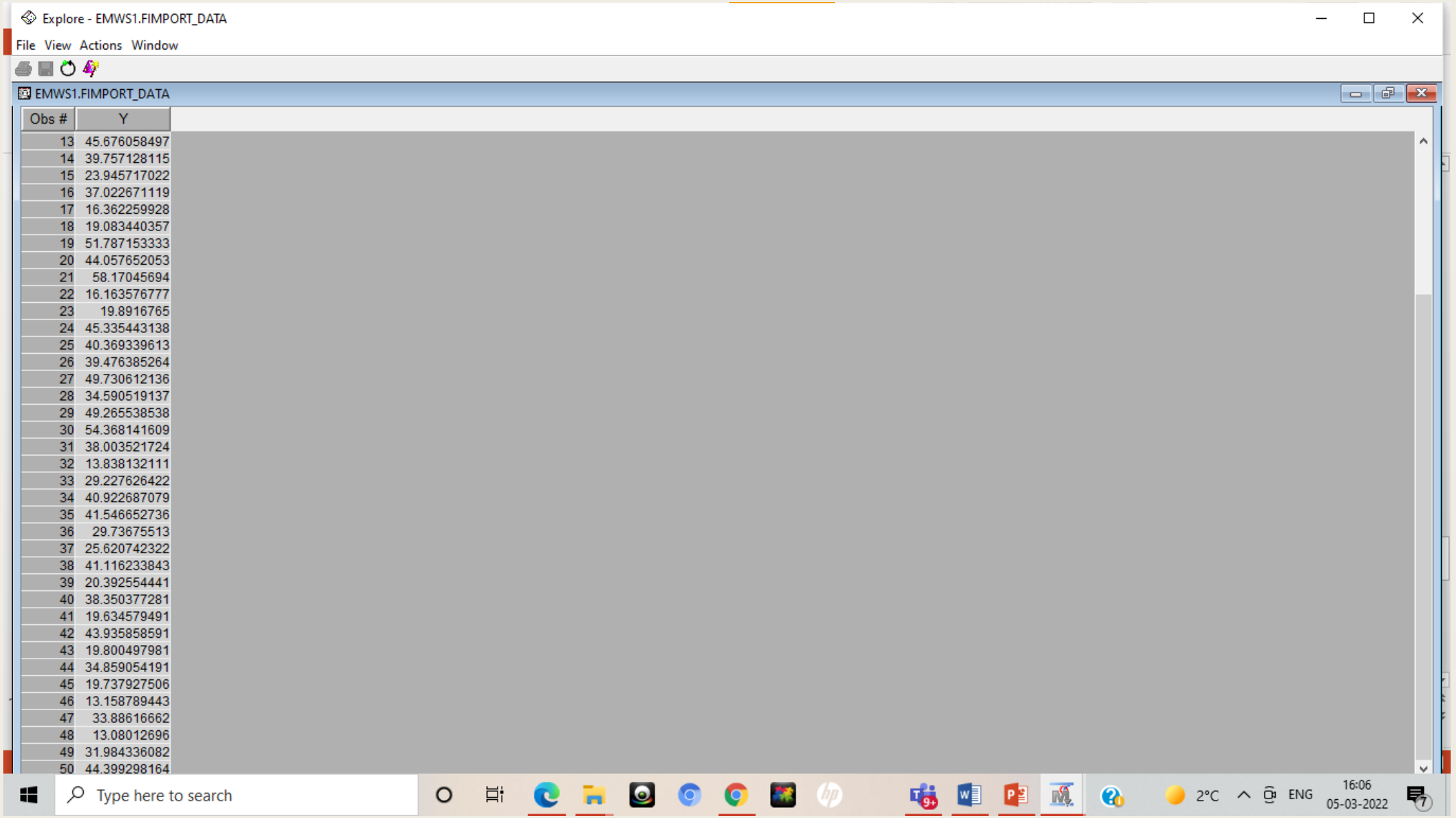


Answer 4: Paste the frequency plot of Y here

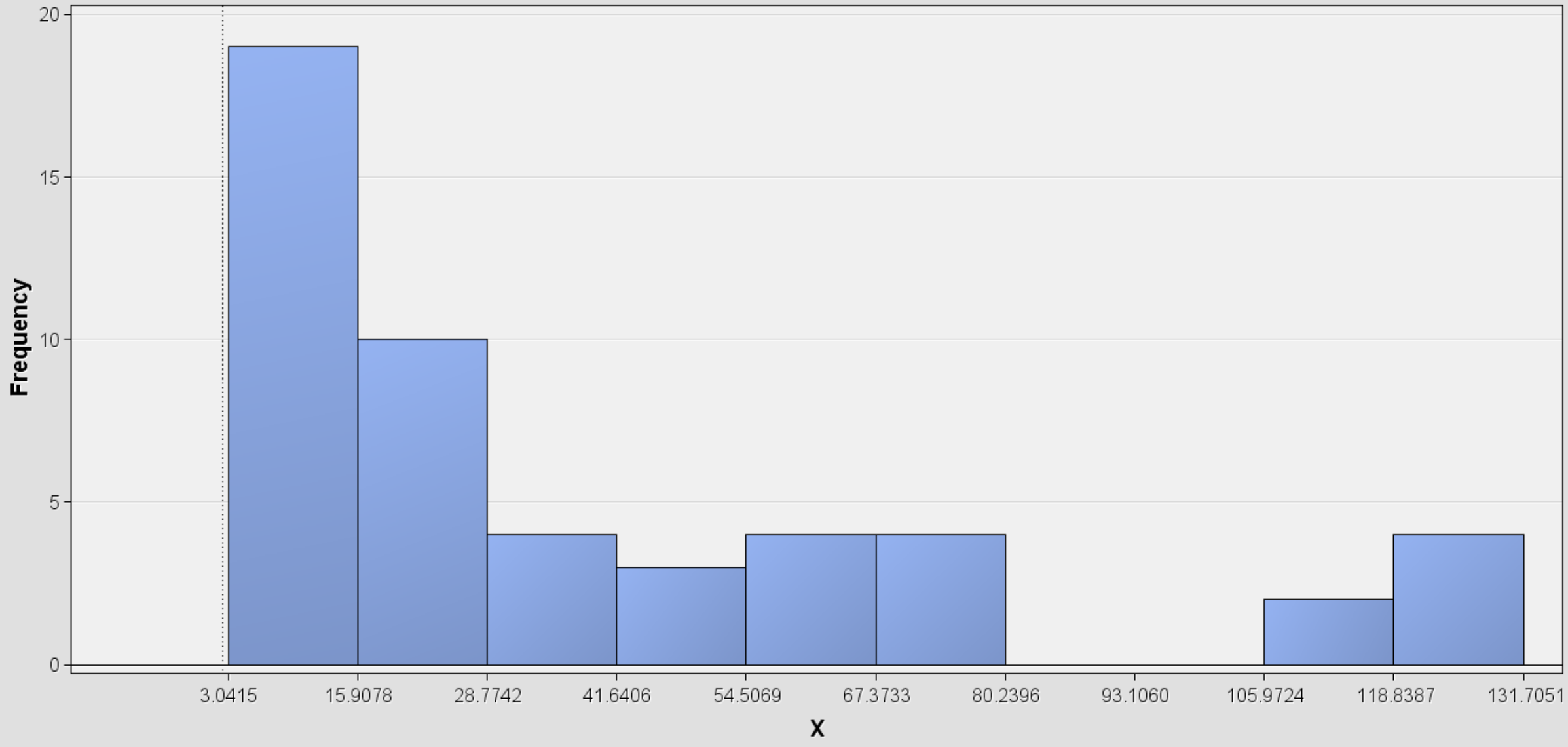


Answer 4: How many rows were imported?

50 rows were imported



Answer 4: Paste the frequency plot of X here



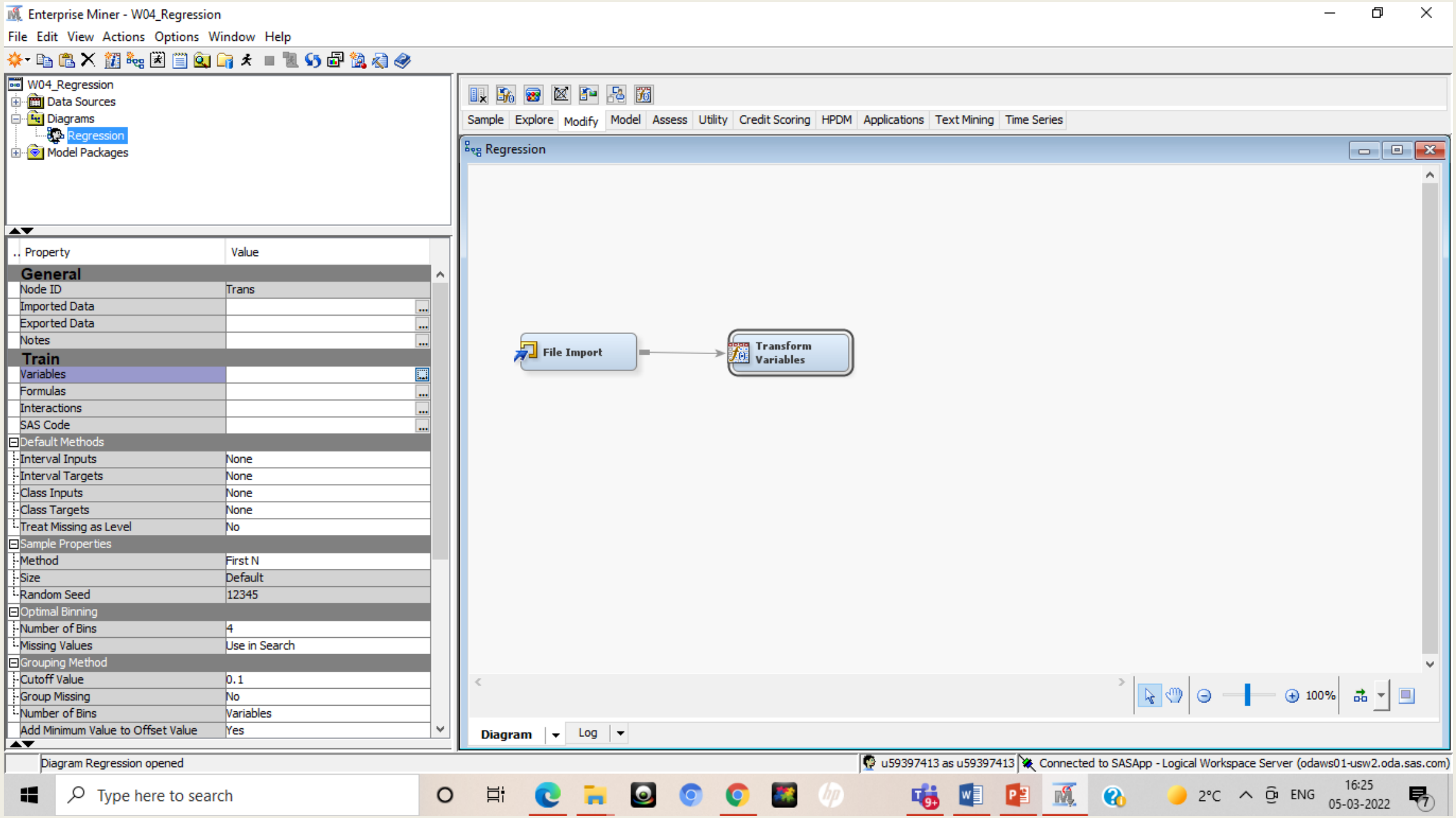
Answer 4: Is the distribution of X suitable for a linear regression model?

The frequency plot of variable X is right skewed which indicates that :

1. There can be a presence of certain amount of outliers that can affect the analysis and would further require complexity to remove the abnormality.
2. There can be an influence on the distribution of the residuals(the difference between estimated value of the regression and the observed dataset) which can lead to non normality.

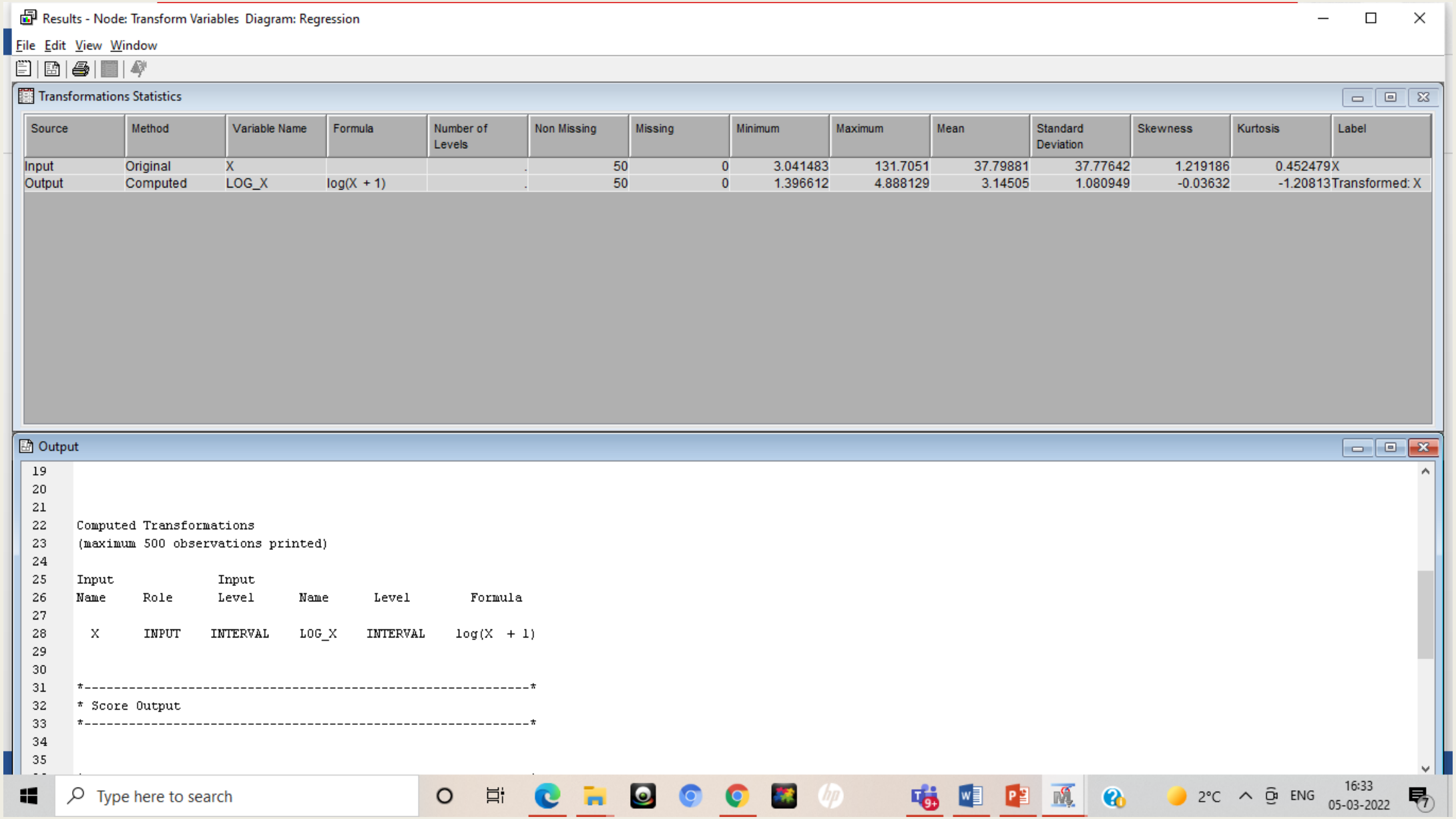
Hence, it is recommend to transform the skewed variables(left or right) before performing regression analysis in order to eliminate the harmful effects.

Answer 5: Transformation of Variable X

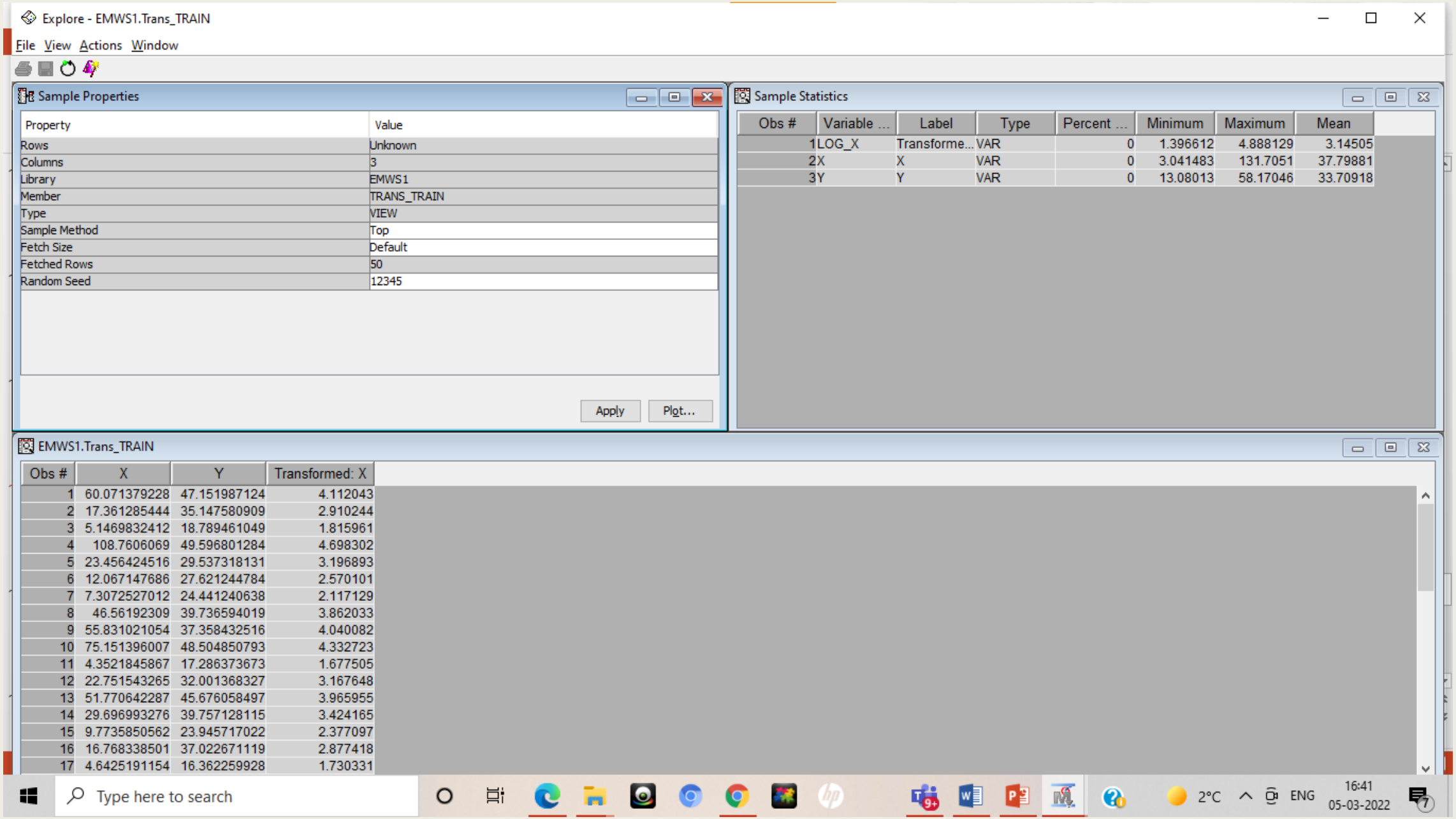


Answer 6: Transformation of Variable X and formula selected

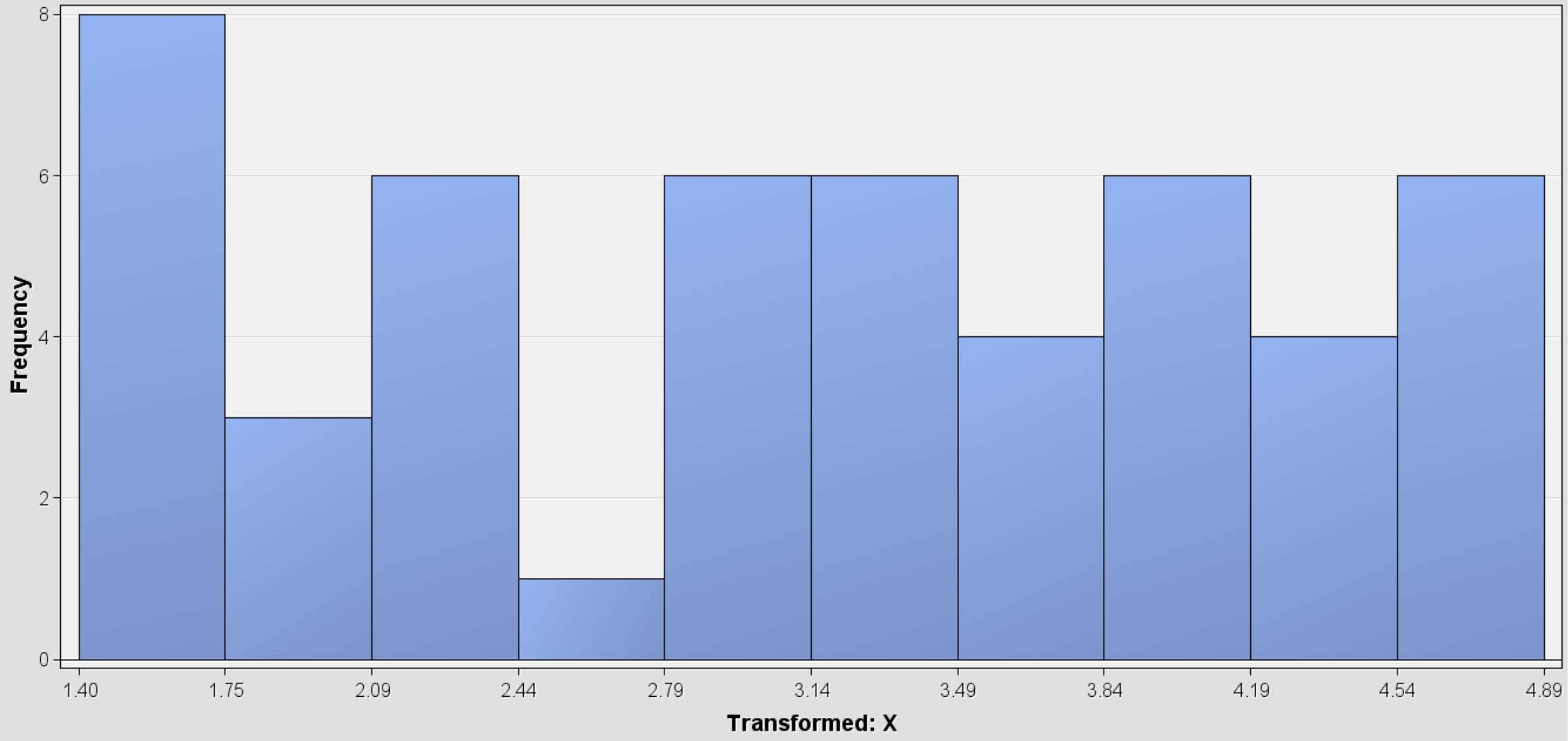
The formula used is log (X+1)



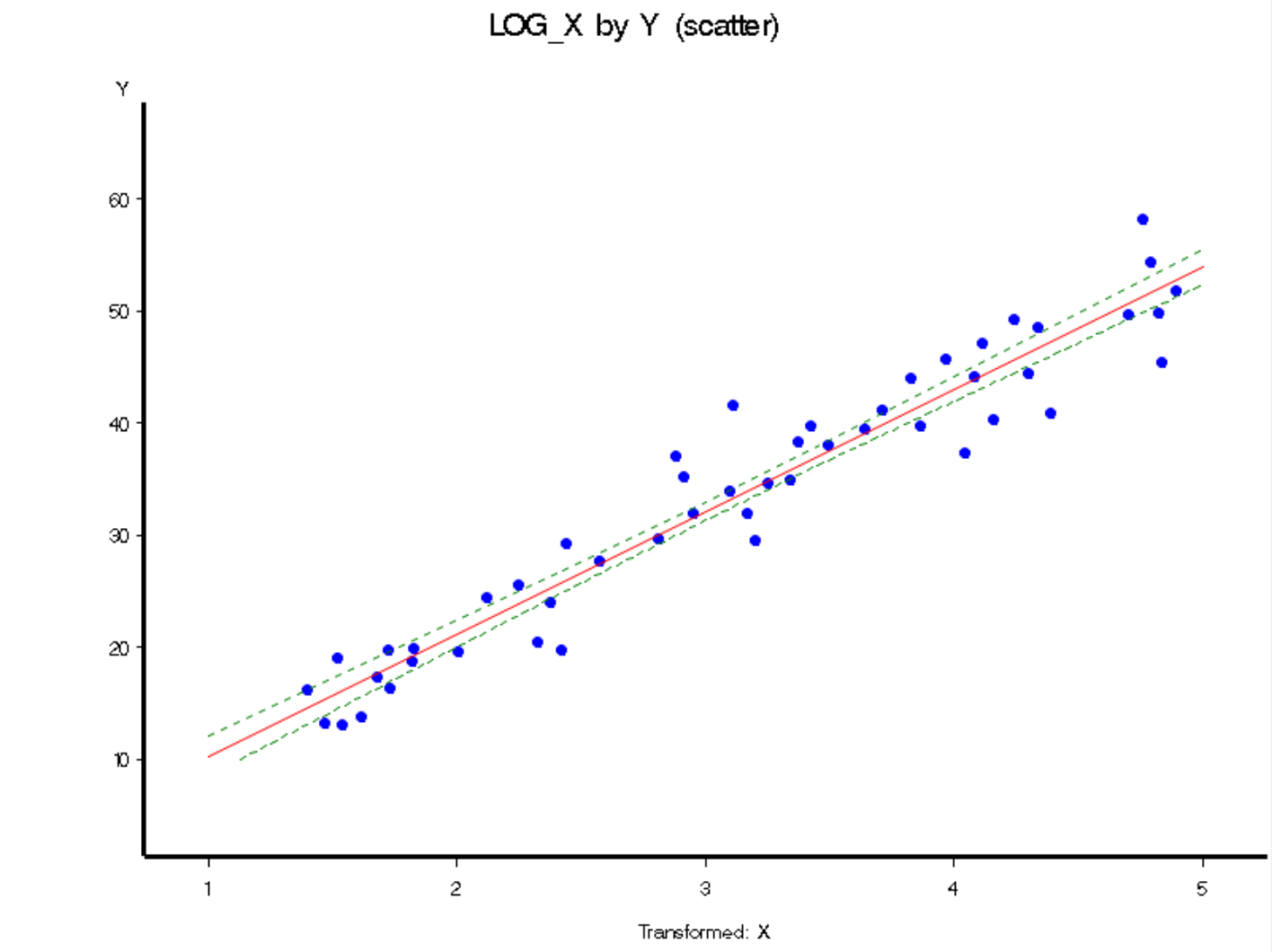
Answer 7: Explore of the Transformed of Variable X



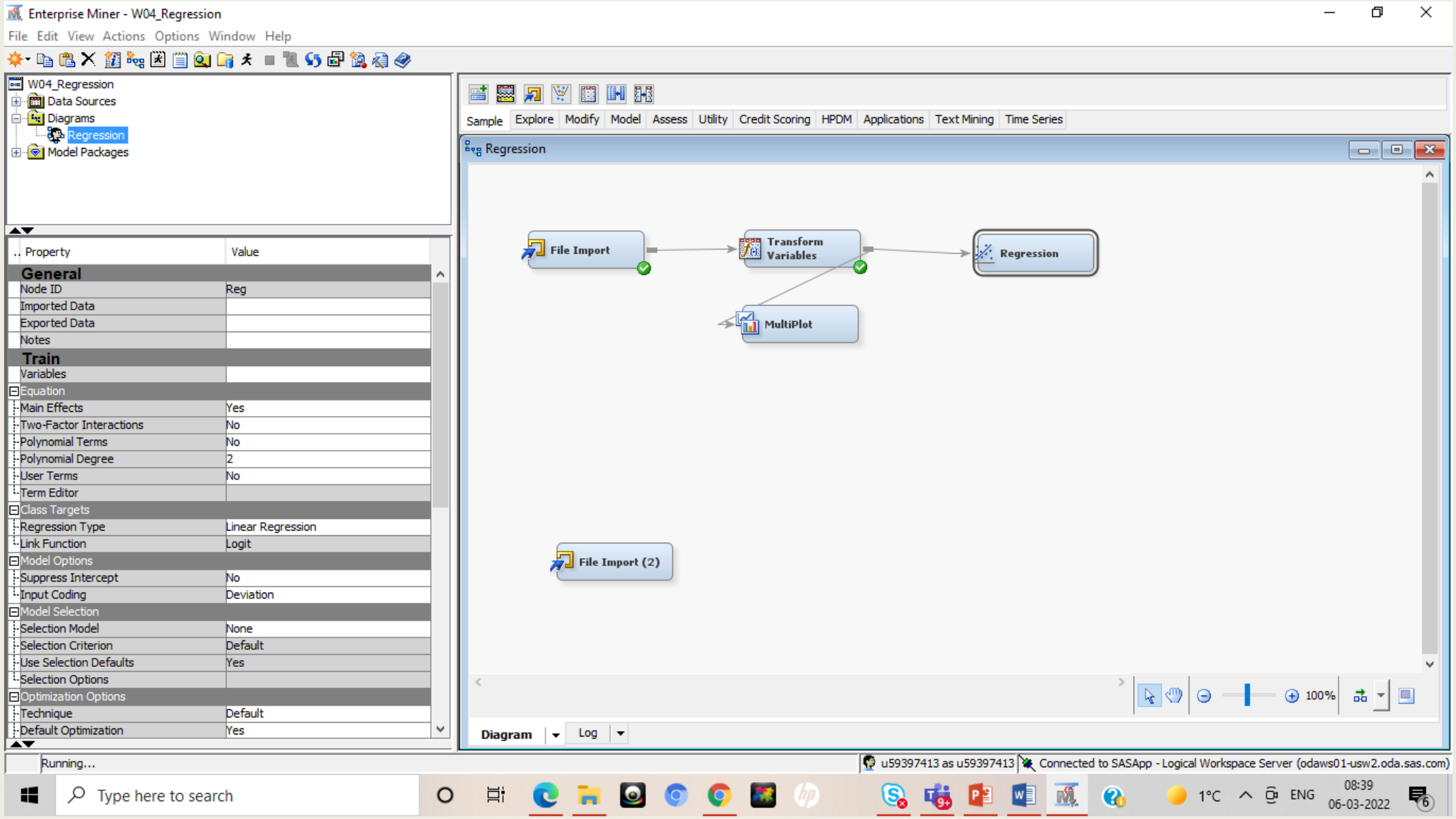
Answer 8: Histogram Plot of Transformed X



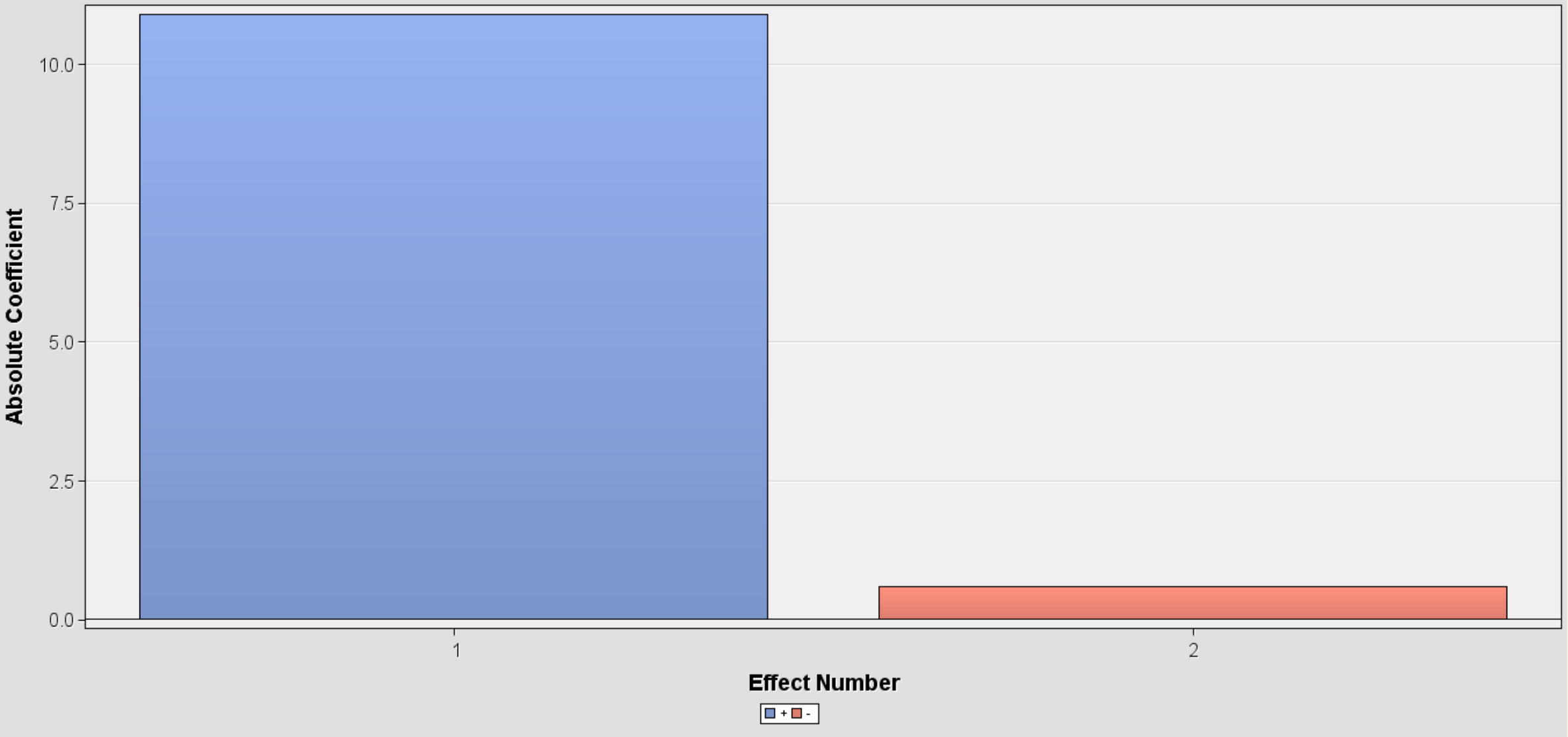
Answer 10: Scatter Plot between log X and Y



Answer 12: Drag the Regression node from Model tab and connect to the Transform node

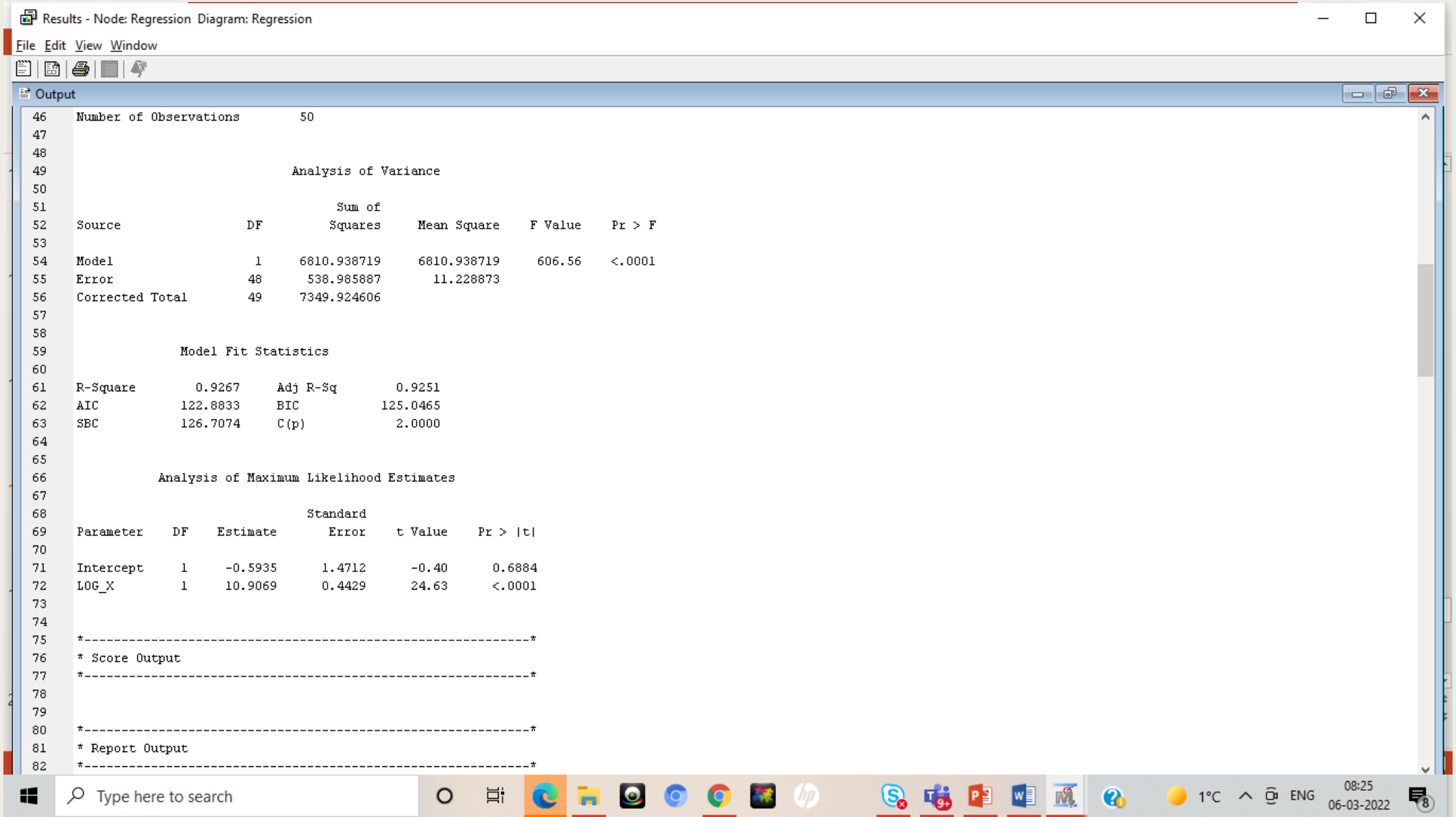


Answer 13: Linear Regression model – Effects Plot



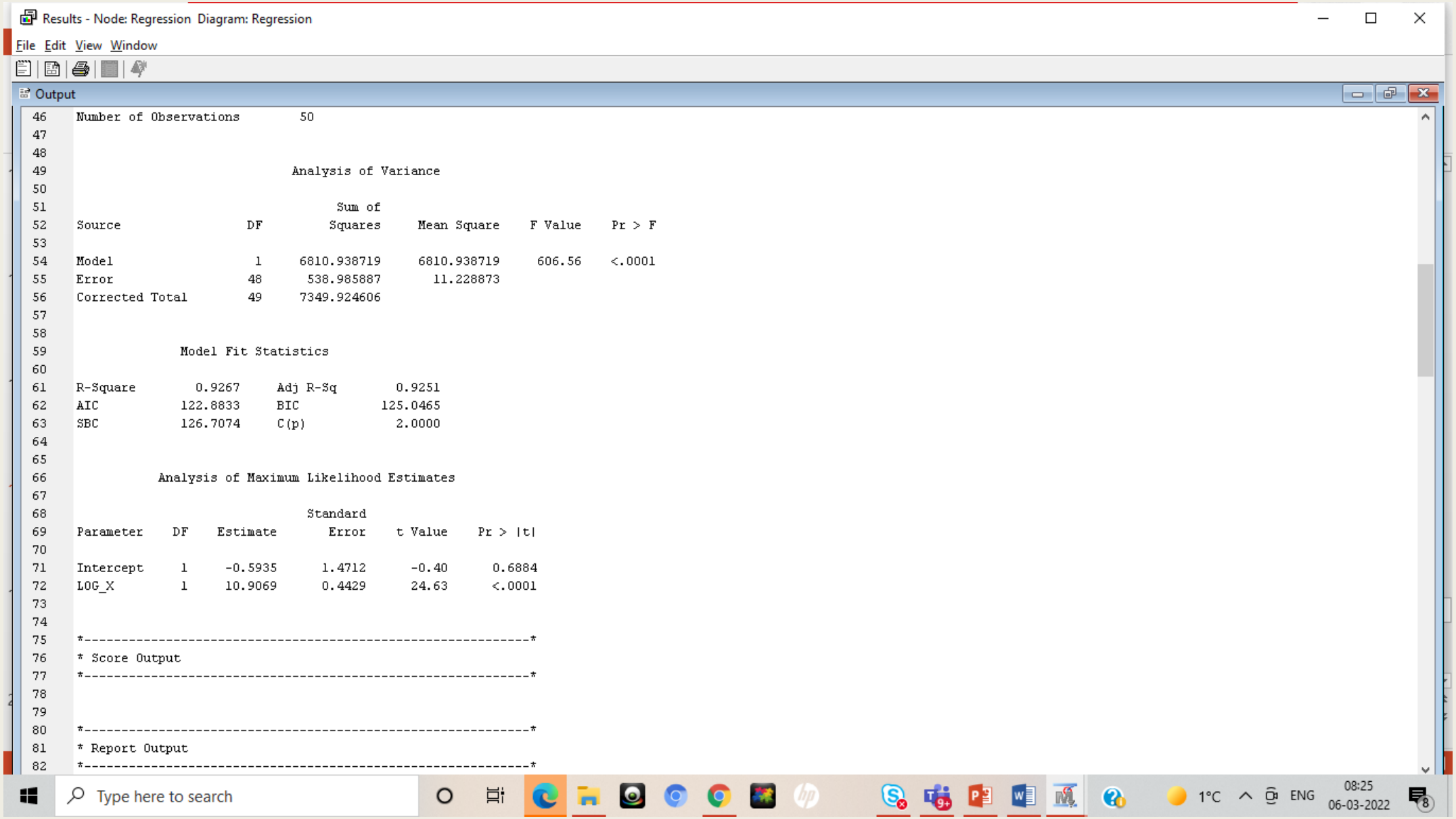
Answer 13: Linear Regression model – Coefficient for log X

The coefficient for log X is 10.9069



Answer 13: Linear Regression model – intercept

The intercept is -0.5935



Answer 13: Linear Regression model – Predicted Y for X = 10

The linear equation obtained after running the regression model is:

$Y = -0.5935 + (10.9069 * \log X)$

Now for predicting value of Y at X = 10, we have :

$Y = -0.5935 + (10.9069 * \log 10)$

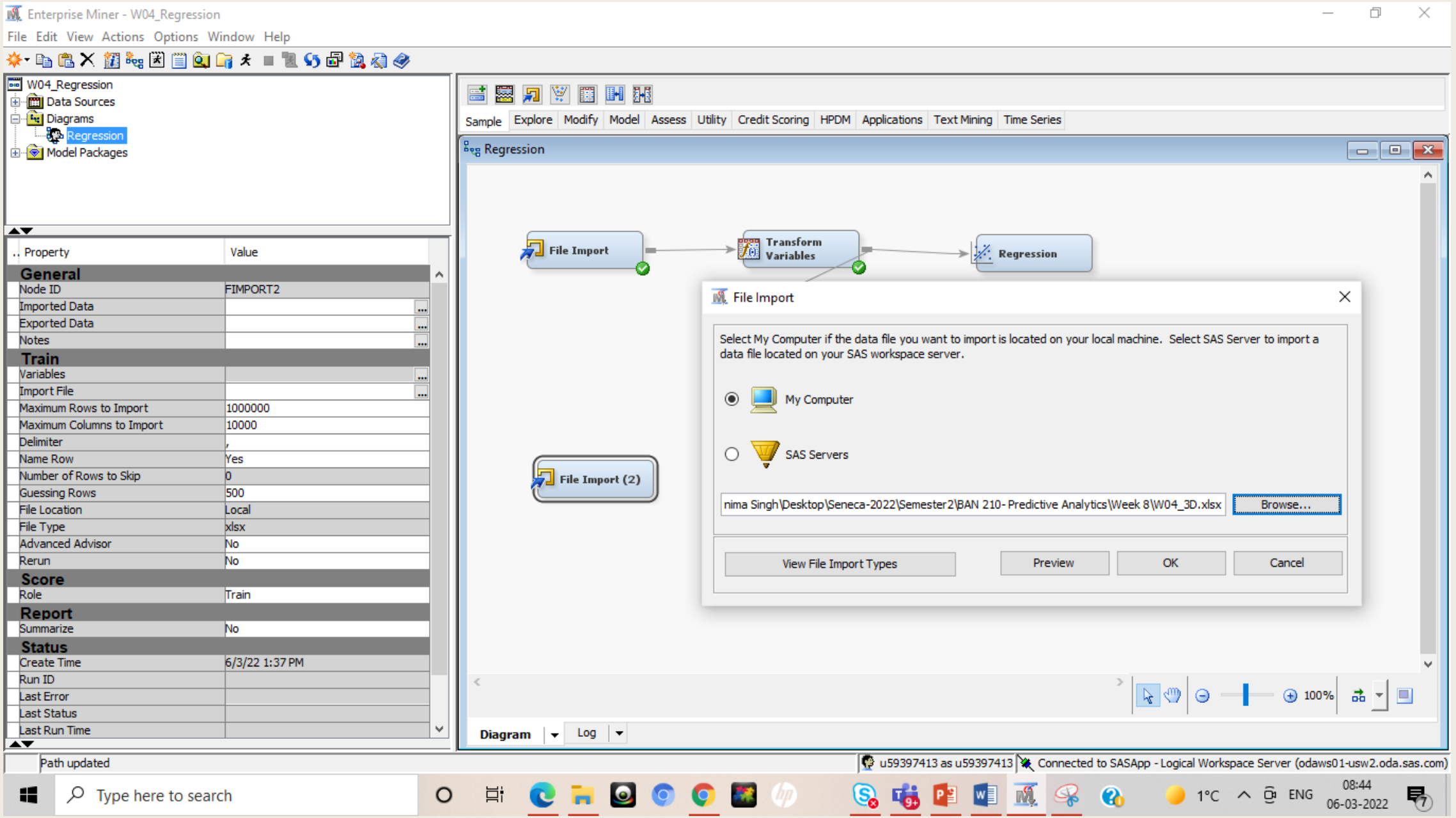
$Y = -0.5935 + (10.9069 * 2.302585)$

$Y = 24.5205$

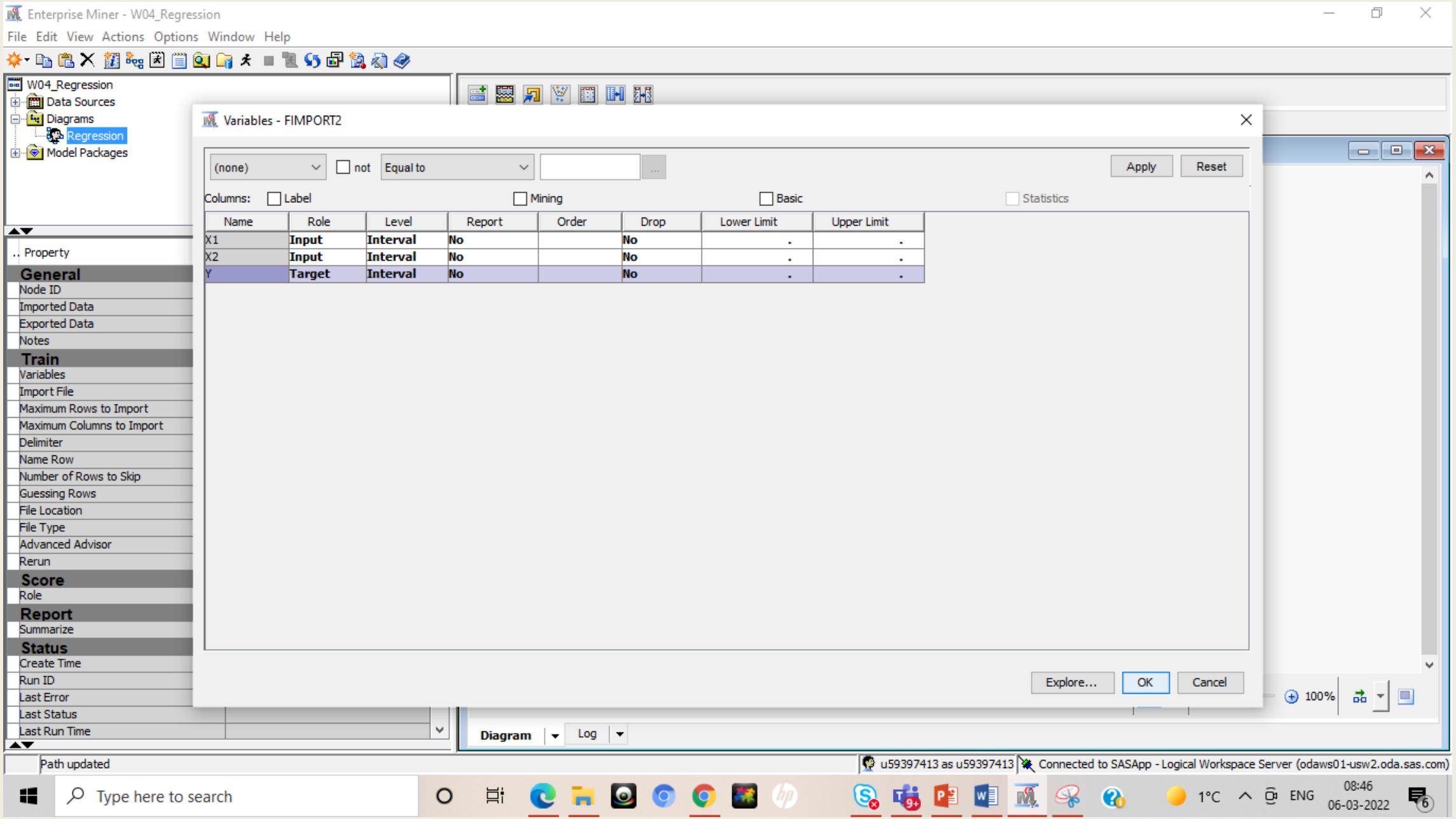
The predicted value of Y is **24.5205**

PART II: MULTIVARIATE REGRESSION

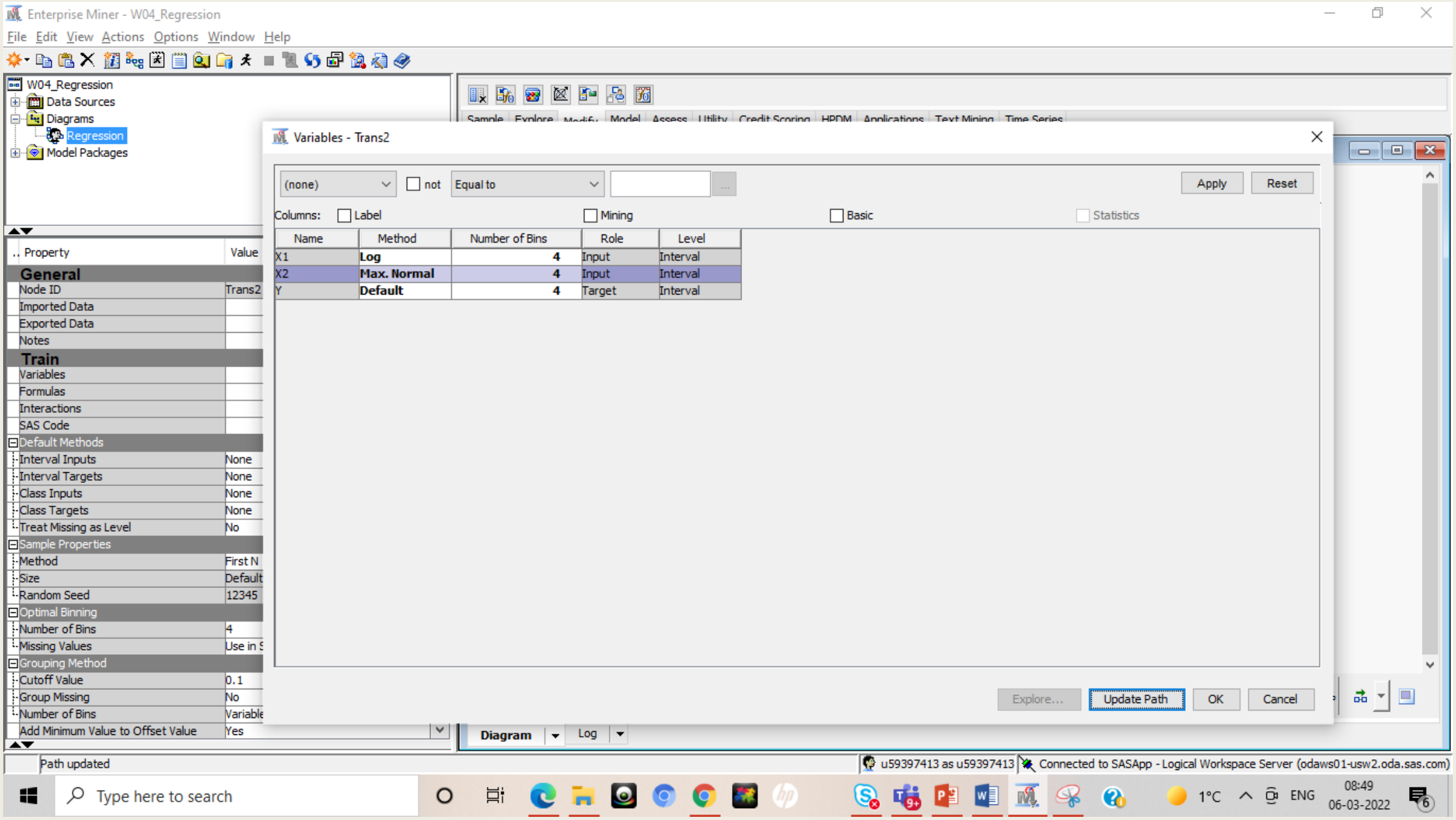
Answer 15: Importing the file W04_3D.xlsx using File Import



Answer 16: Click on Y. Choose Target as the Role for Y

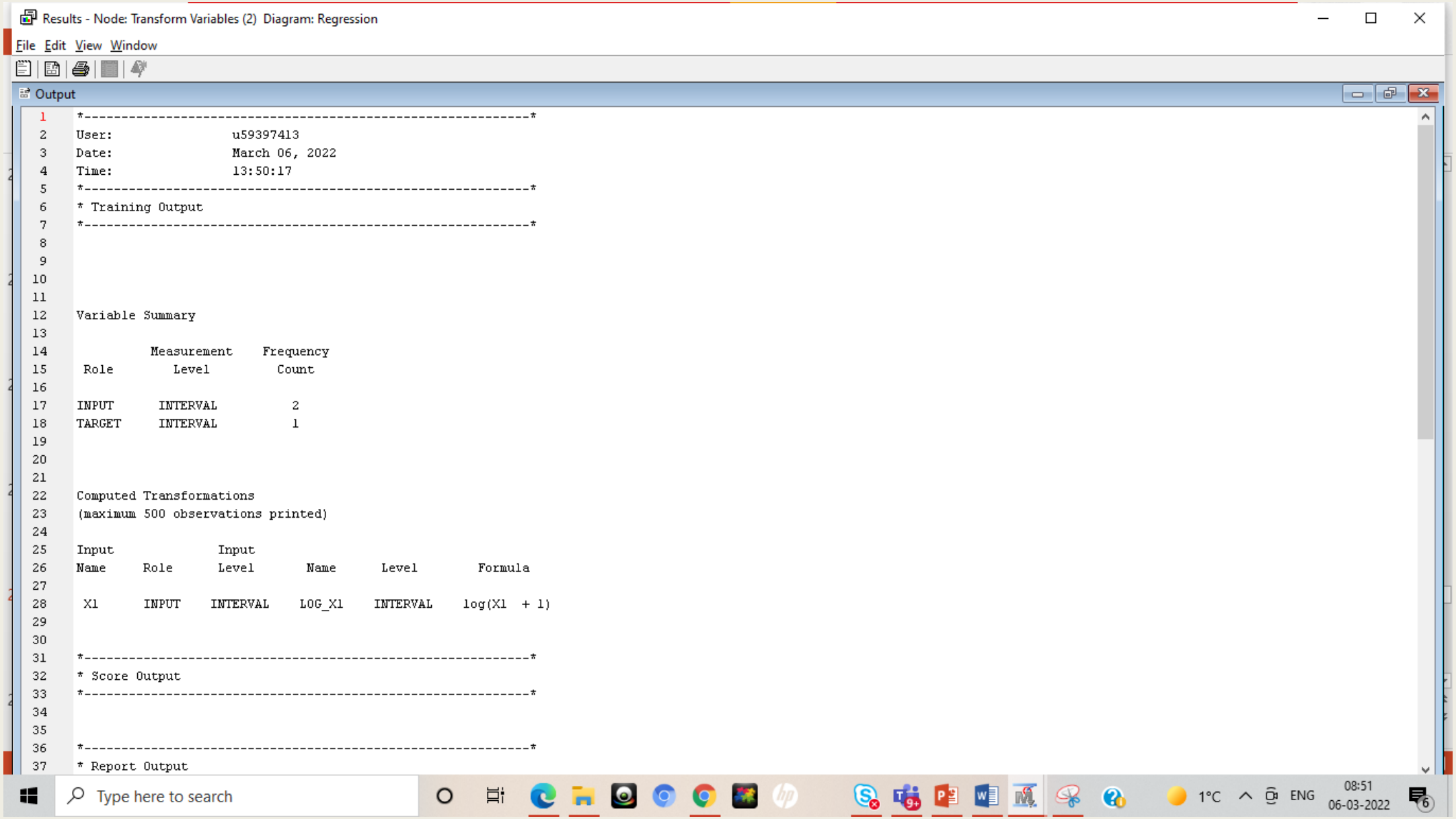


Answer 17: Transformation of Variable

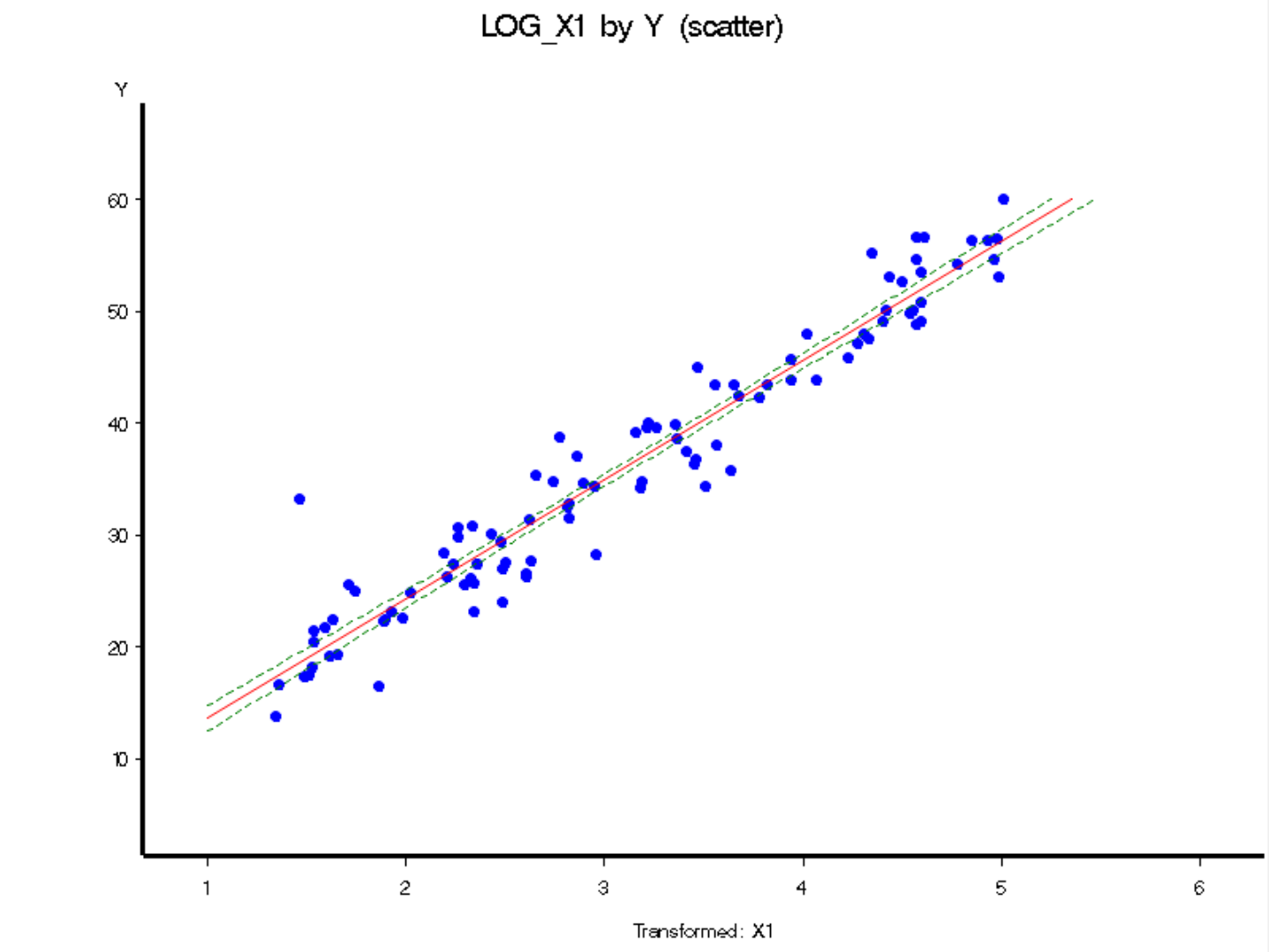


Answer 18: Transformation of Variable X1 and X2 and formula selected

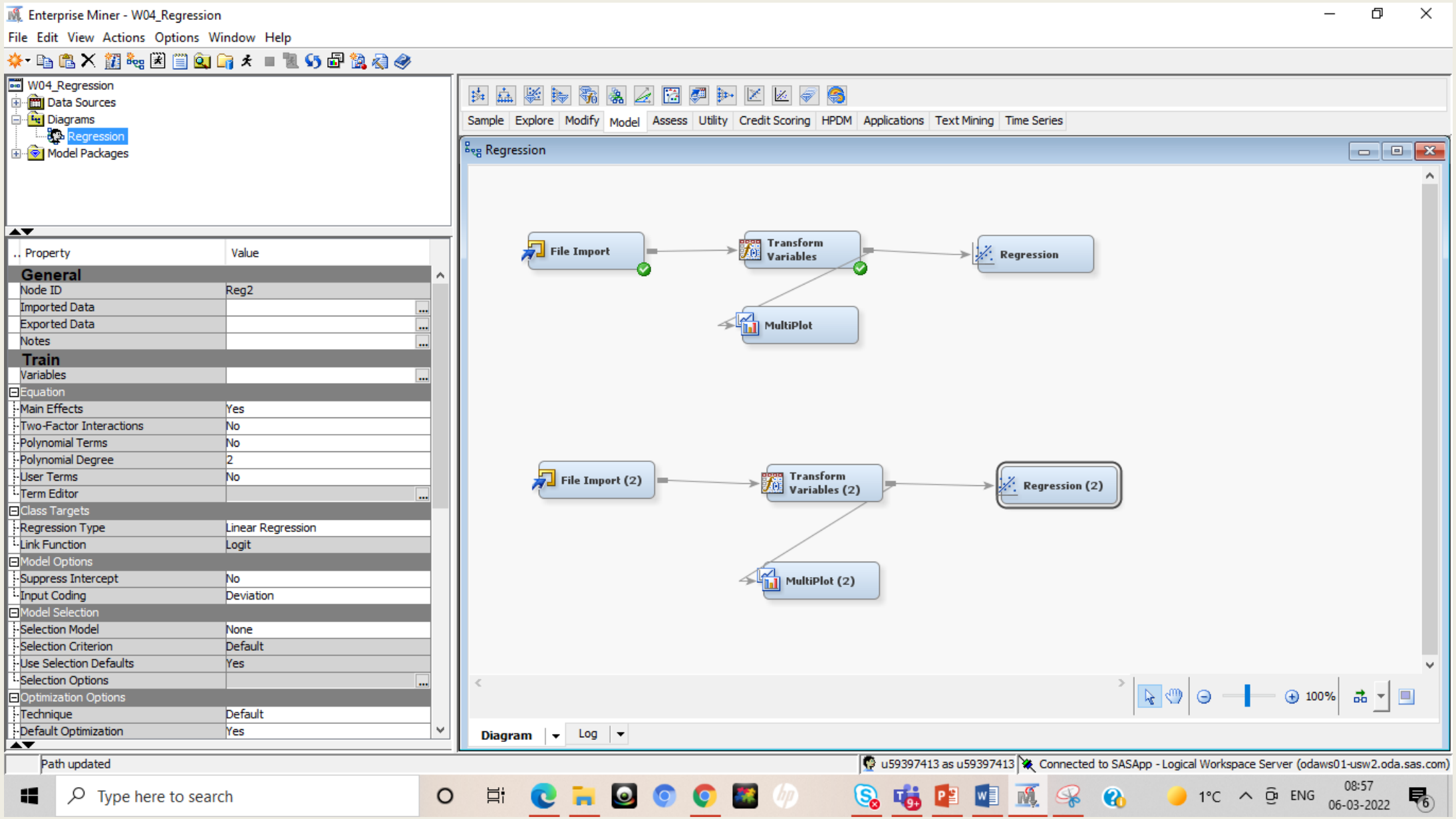
The formula selected is log (X1 +1)



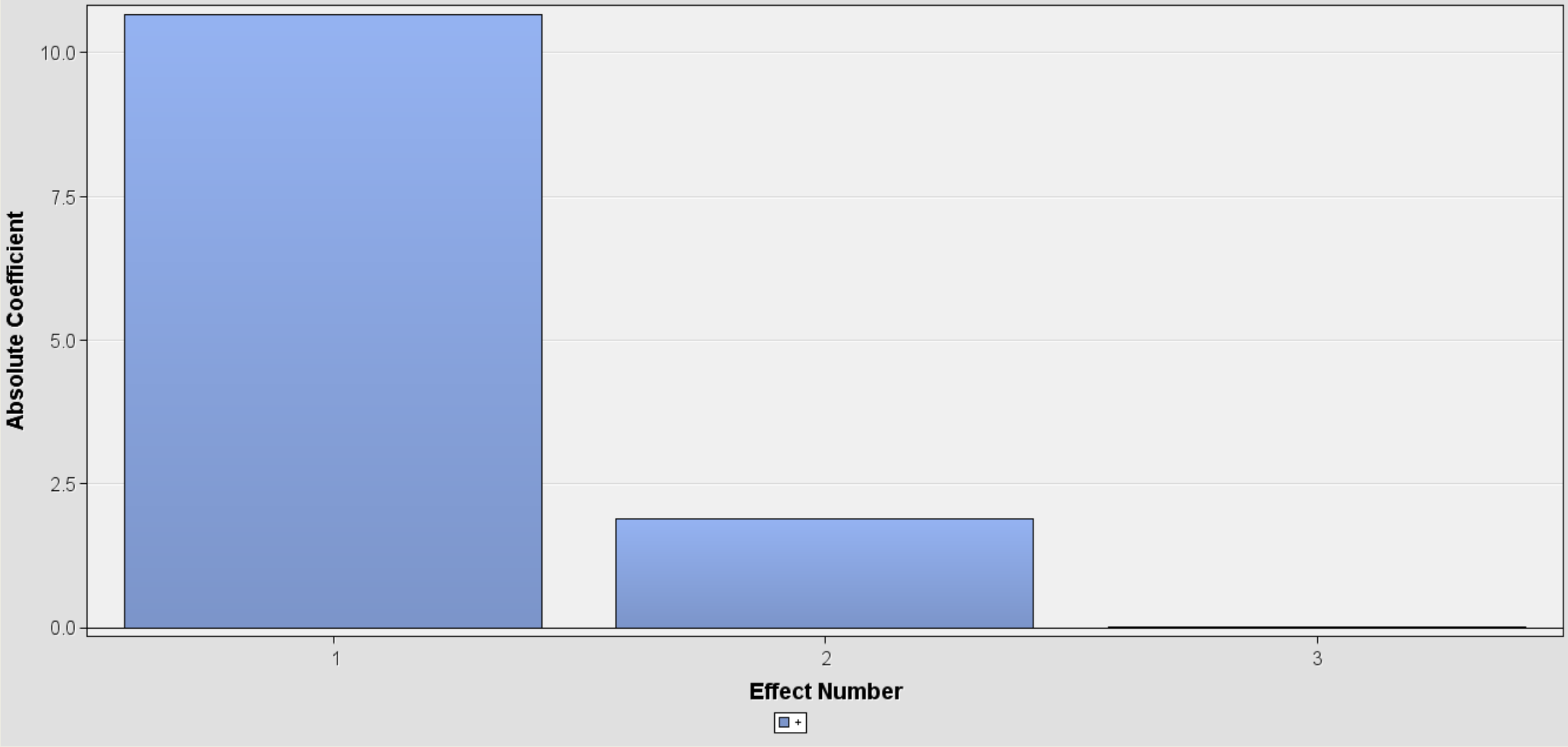
Answer 19: Scatter Plot between log X1 and X2



Answer 20: Drag the Regression node from Model tab and connect to the Transform node.

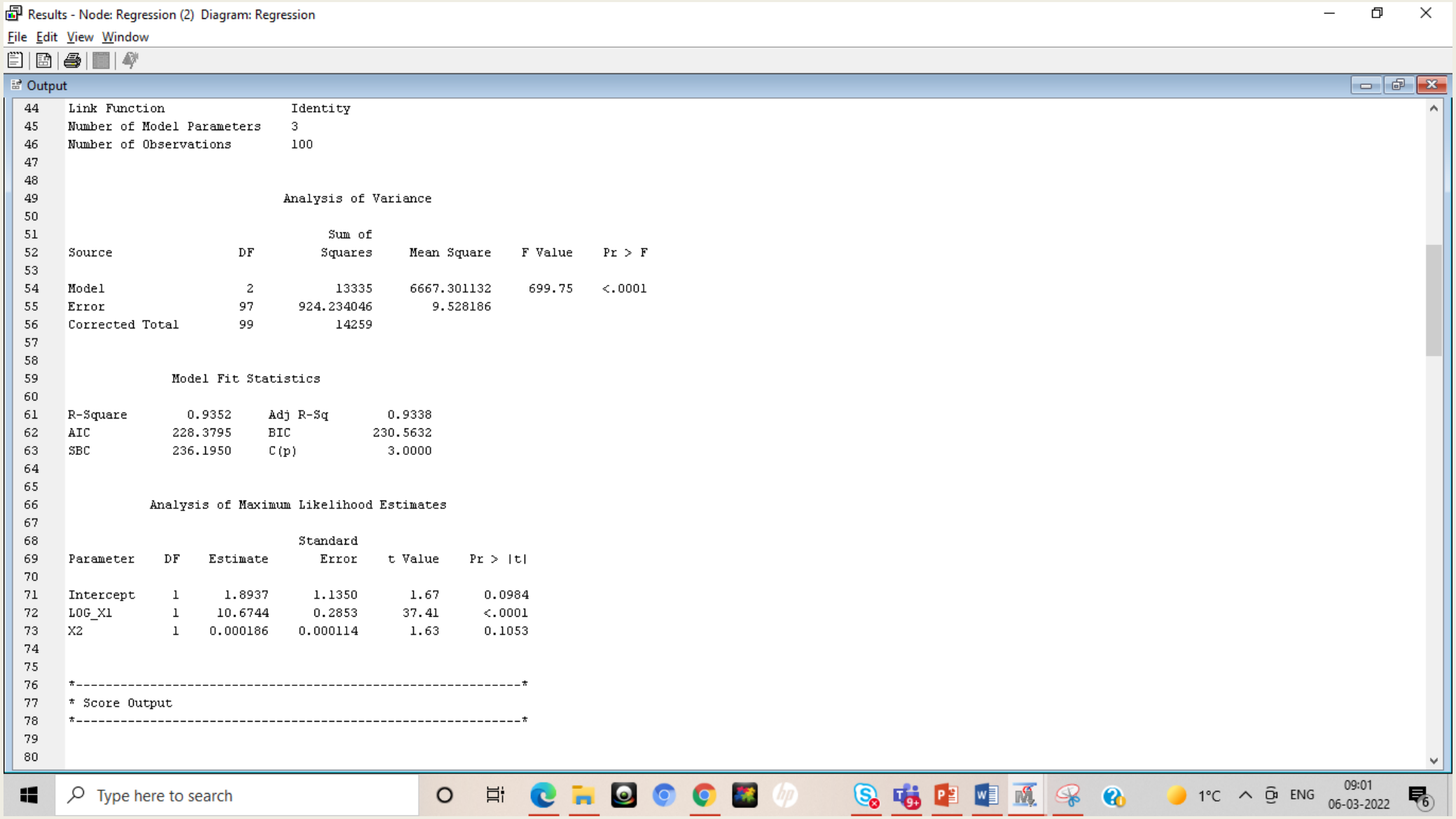


Answer 21: Linear Regression model – Effects Plot



Answer 21: Linear Regression model – the formula for the model

The formula is Y = 1.8937 + (10.6744 * log X1) + (0.000186 * X2)



Answer 21: Linear Regression model – Predicted Y for X1 = 10 and X2 =5000

Y = 1.8937 + (10.6744 * log X1) + (0.000186 * X2)

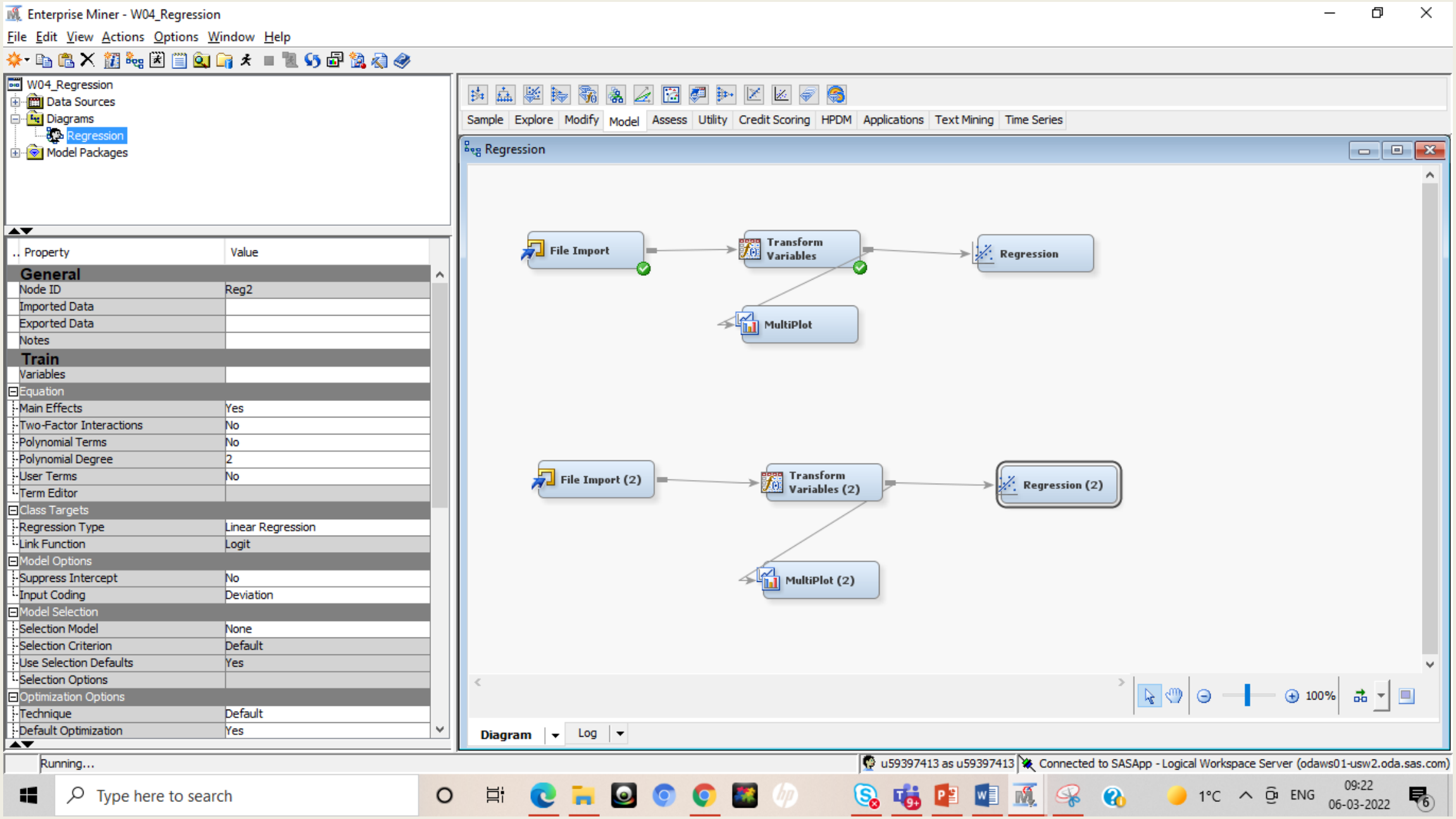
Y = 1.8937 + (10.6744 * log 10) + (0.000186 * 5000)

Y = 1.8937 + (10.6744 * 2.302585) + (0.000186 * 5000)

Y = 27.4024133

The predicted value of Y is **27.4024133**

Answer 22: Paste the diagram here



PART III: REGRESSION ASSESSMENT

Answer 23: Importing the file W04_3D.xlsx using File Import

First Regression Model

Model Fit Statistics

R-Square	0.9267	Adj R-Sq	0.9251
AIC	122.8833	BIC	125.0465
SBC	126.7074	C(p)	2.0000

Fit Statistics	Statistics Label	Train
AIC	Akaike's Information Criterion	122.883
ASE	Average Squared Error	10.780

Second Regression Model

Model Fit Statistics

R-Square	0.9352	Adj R-Sq	0.9338
AIC	228.3795	BIC	230.5632
SBC	236.1950	C(p)	3.0000

Fit Statistics	Statistics Label	Train
AIC	Akaike's Information Criterion	228.380
ASE	Average Squared Error	9.242

	R squared	Average squared error
First regression model	0.9267	10.780
Second regression model	0.9352	9.242

- R-squared value determine how close the data are to the fitted regression line and average squared error is the mean square error(of the predicted and the observed value). The more the value of the R-Square , the better the model will behave.
- It is indicated that the addition of more independent variables can lead to a better analysis and hence good R-square value. However, comparing the univariate and the multivariate regression using different dataset would not result in best result as these are tow different statistical approach.
- The better fit on training data does not normally comply that a resulted model would be performing best on test or validation dataset as the model would be performing on different dataset and hence could result in different statistics.

GROUP WORK DECLARATION

We, **Group 5 (Anand Mohan Thankur, Josh Shaji, Poonam Bhaliyan, Prateek Ramjanam Singh, and Poornima Singh)** declare that the attached assignment is our own work in accordance with the Seneca Academic Policy. We have not copied any part of this assignment, manually or electronically, from any other source including web sites, unless specified as references. We have not distributed our work to other students.

	Name	Task(s)
1	Anand Mohan Thakur (149200206)	Consolidated the Workshop together on MS Teams
2	Josh Shaji (133557215)	Consolidated the Workshop together on MS Teams
3	Poonam Bhaliyan (121114219)	Consolidated the Workshop together on MS Teams
4	Prateek Ramjanam Singh (124483215)	Consolidated the Workshop together on MS Teams
5	Poornima Singh (125638213)	Consolidated the Workshop together on MS Teams

THANK YOU