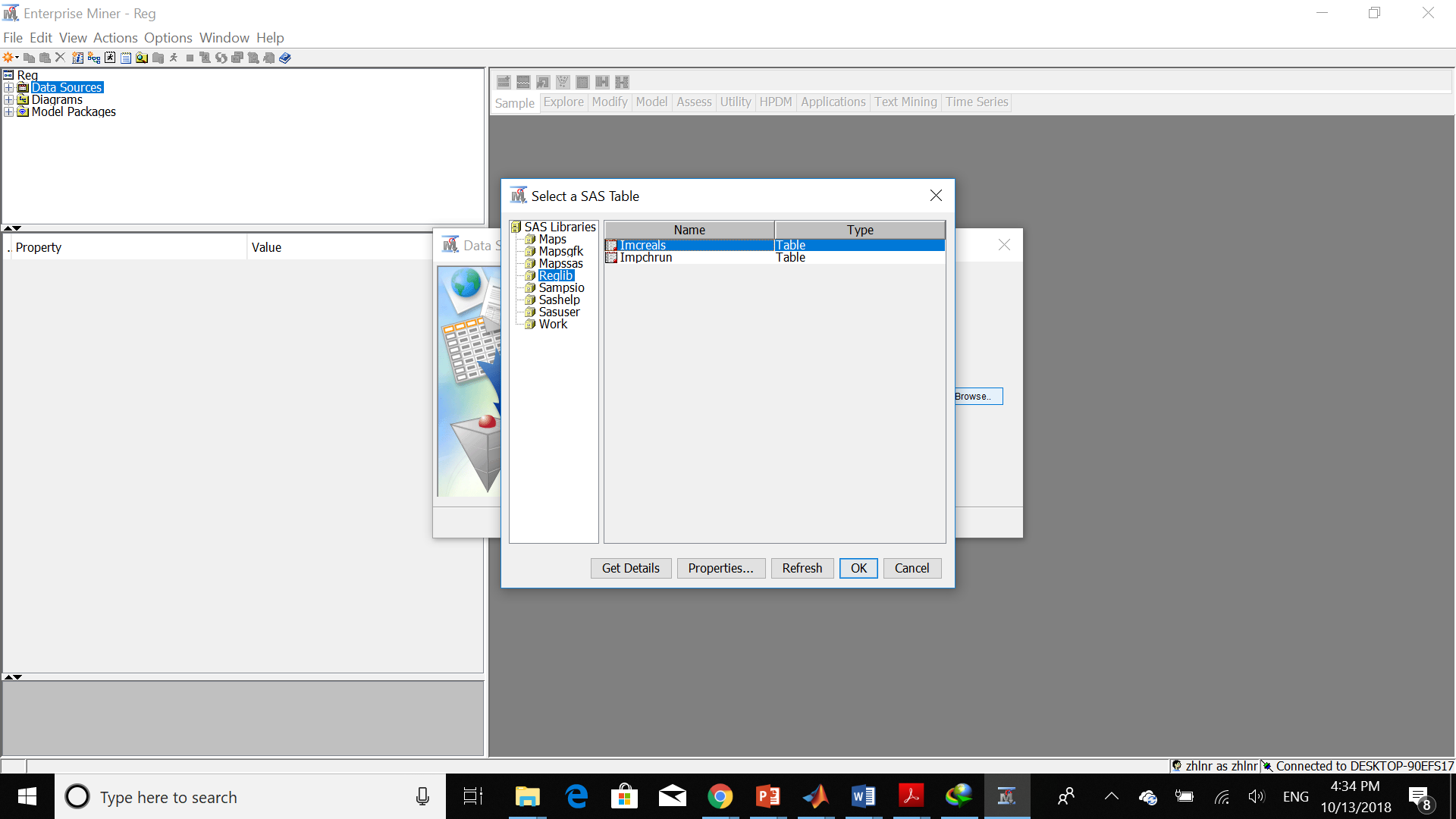
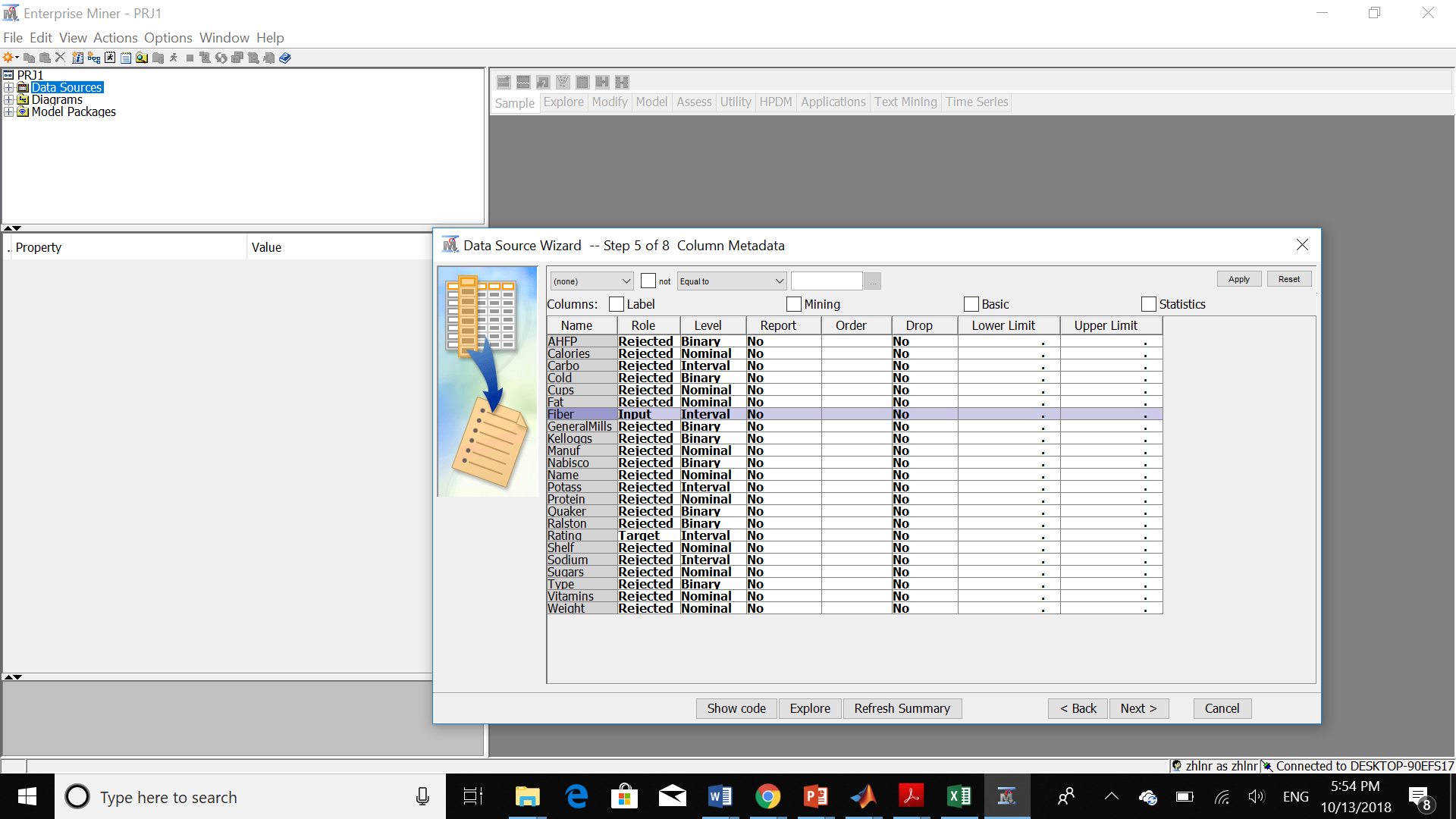
In order to answer to the two first questions of this homework, we should follow these steps:

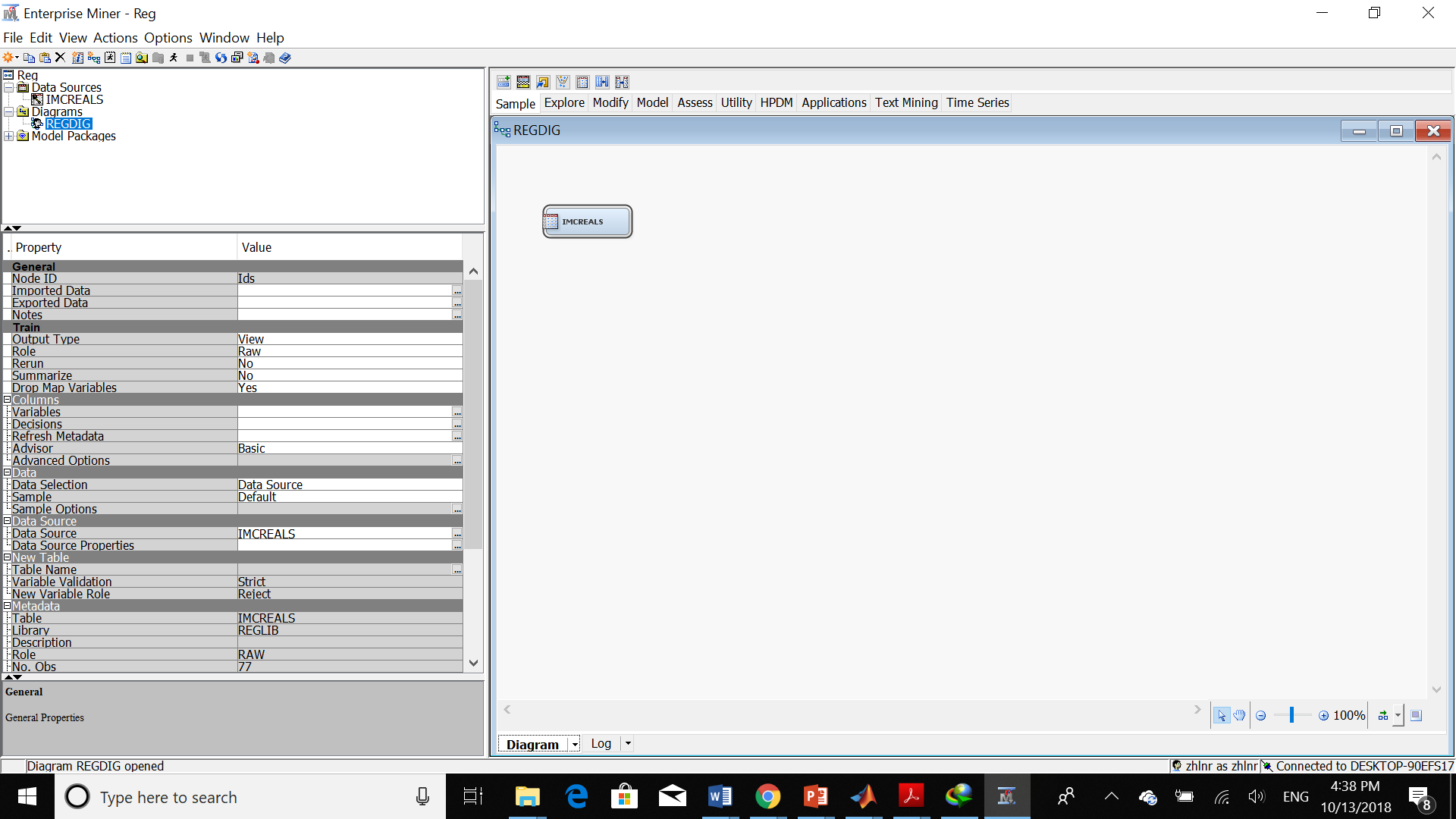
Because, it’s a new dataset we should create new data source as shown below:



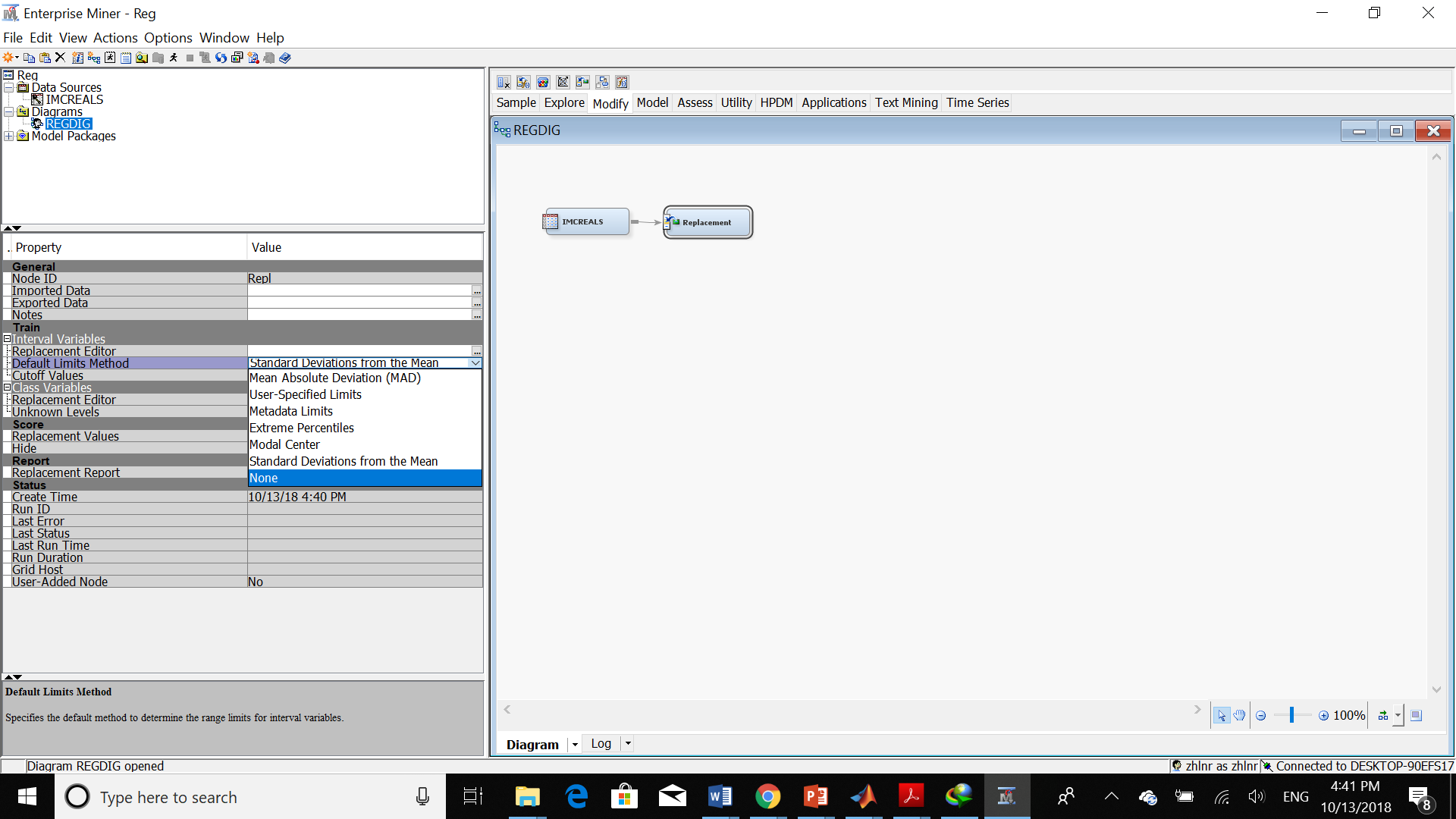
The first two questions ask us to estimate Rating based on Fiber. In order to do so, we should mark all of the other variables as rejected. Also, we should change the type of data from Nominal to Interval. We can see this step as below:

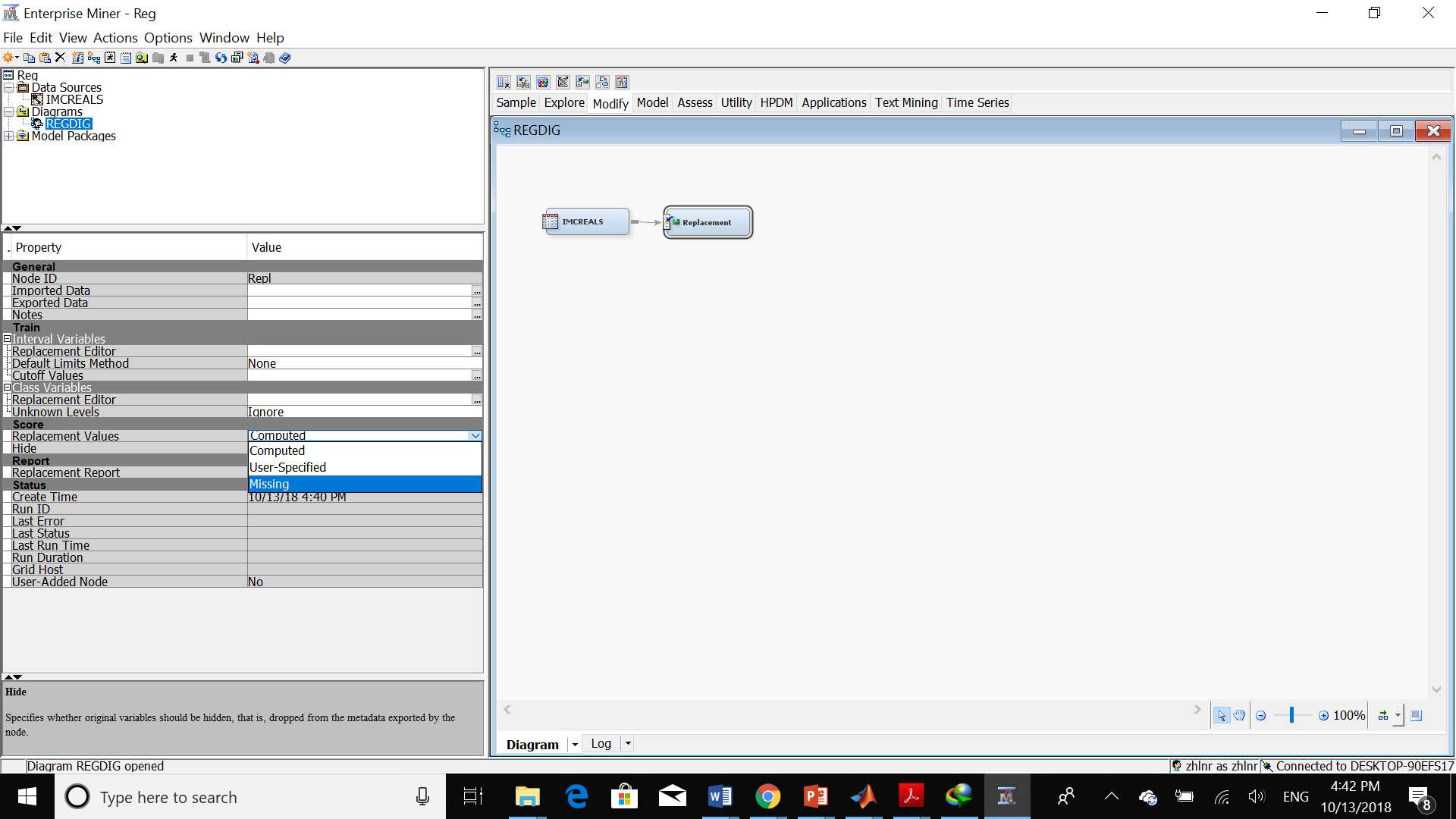


Now, we have the data set:

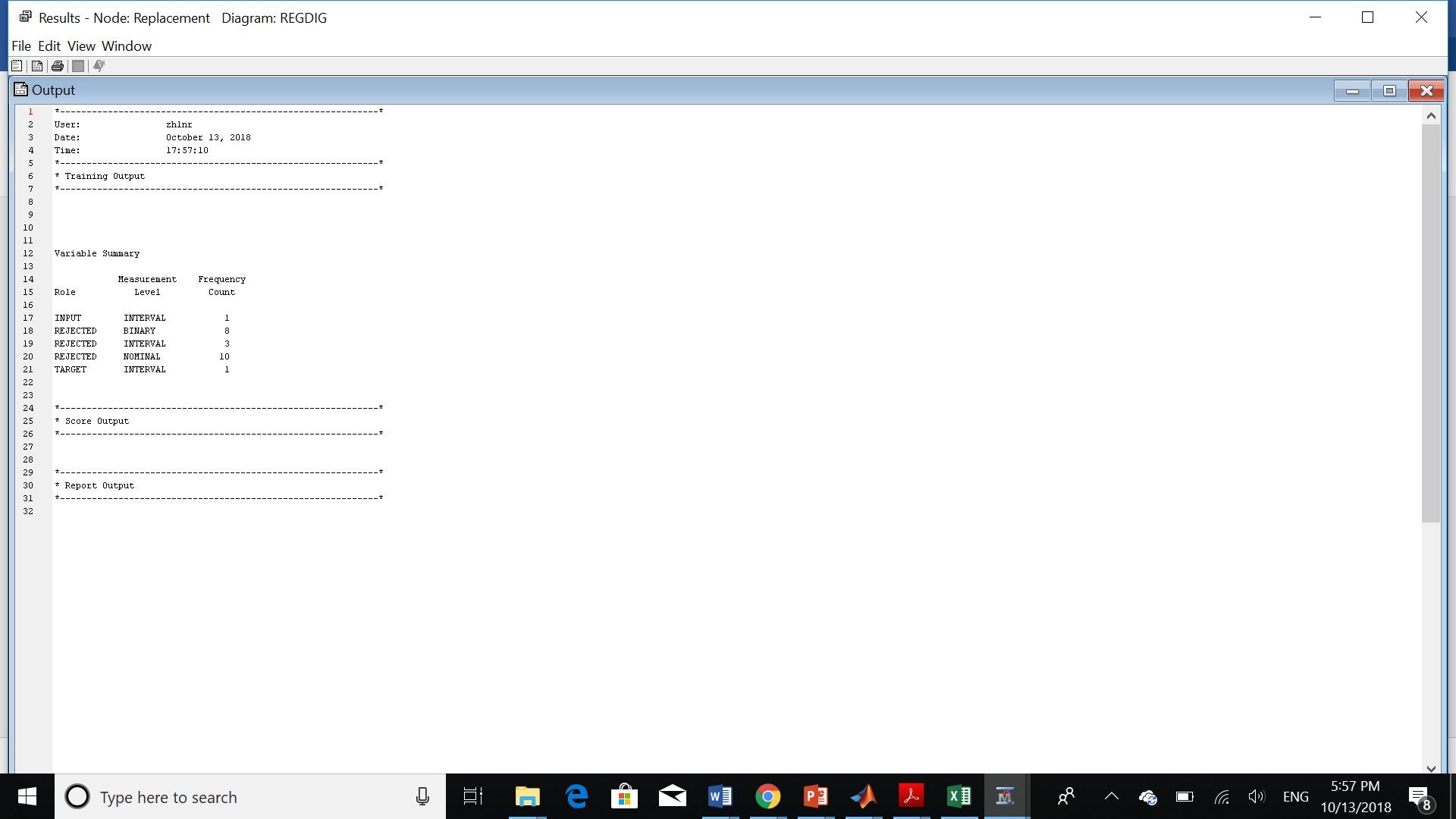


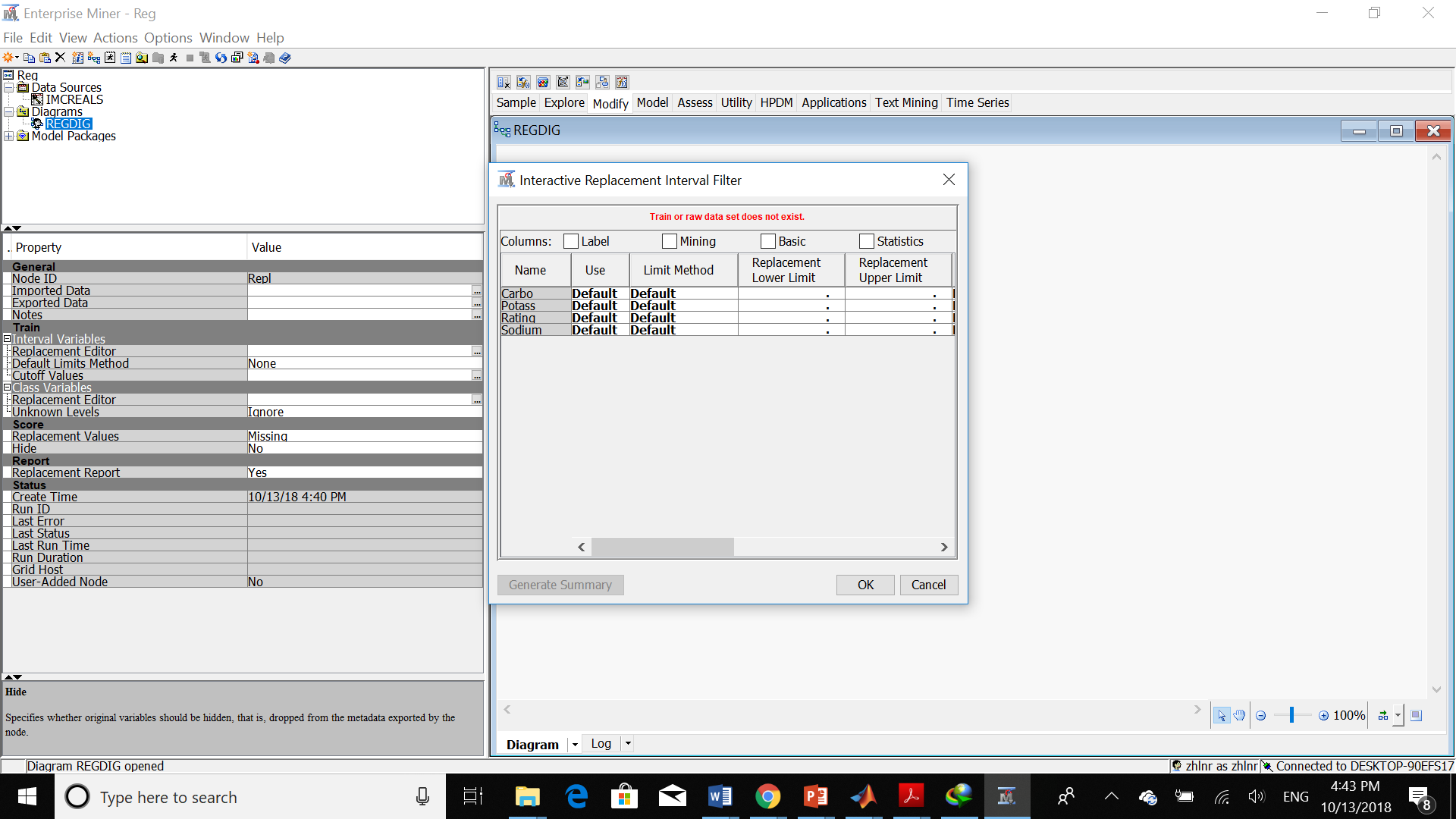
Then I created the replacement node.



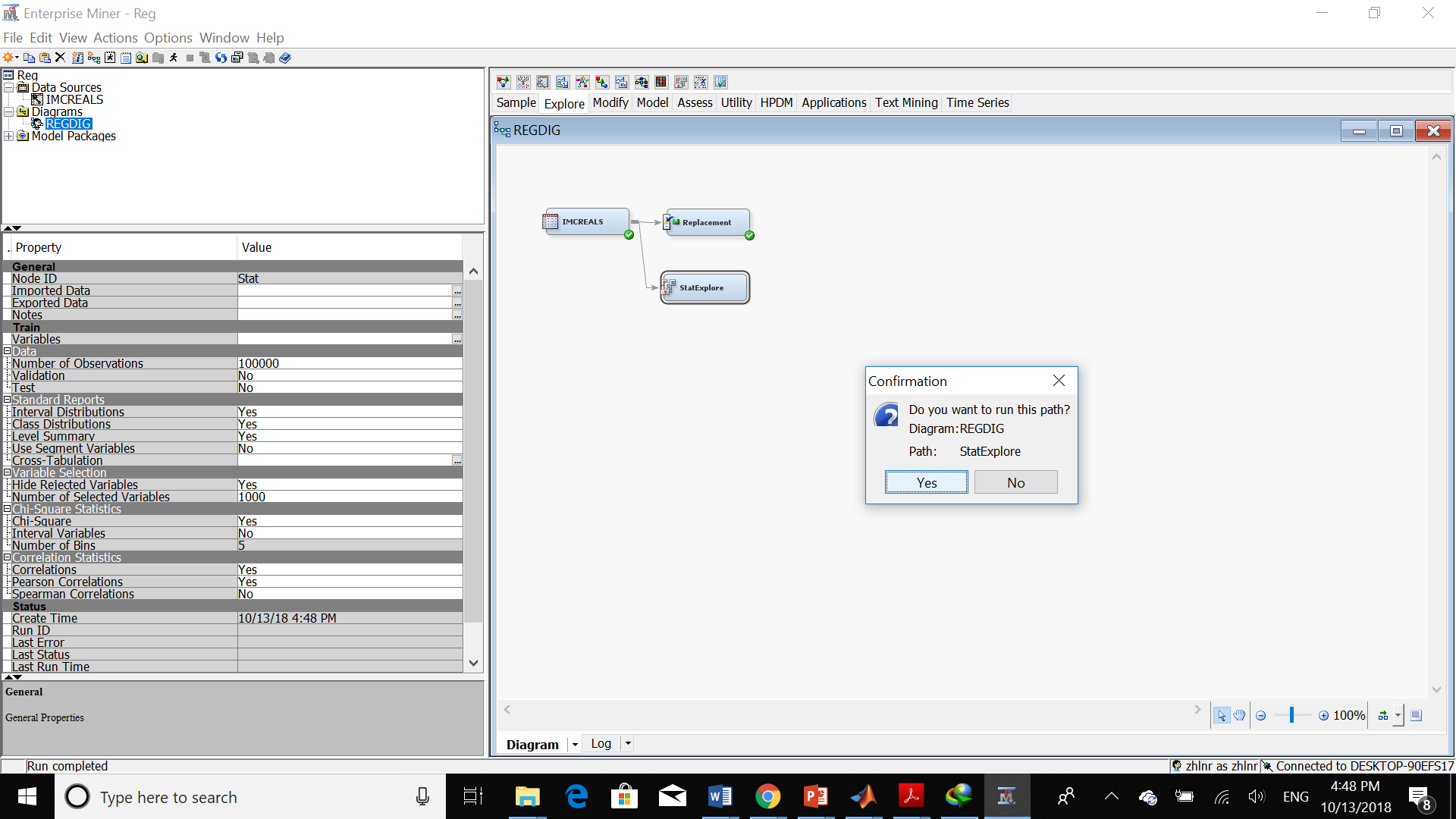


Here is the result of running replacement:

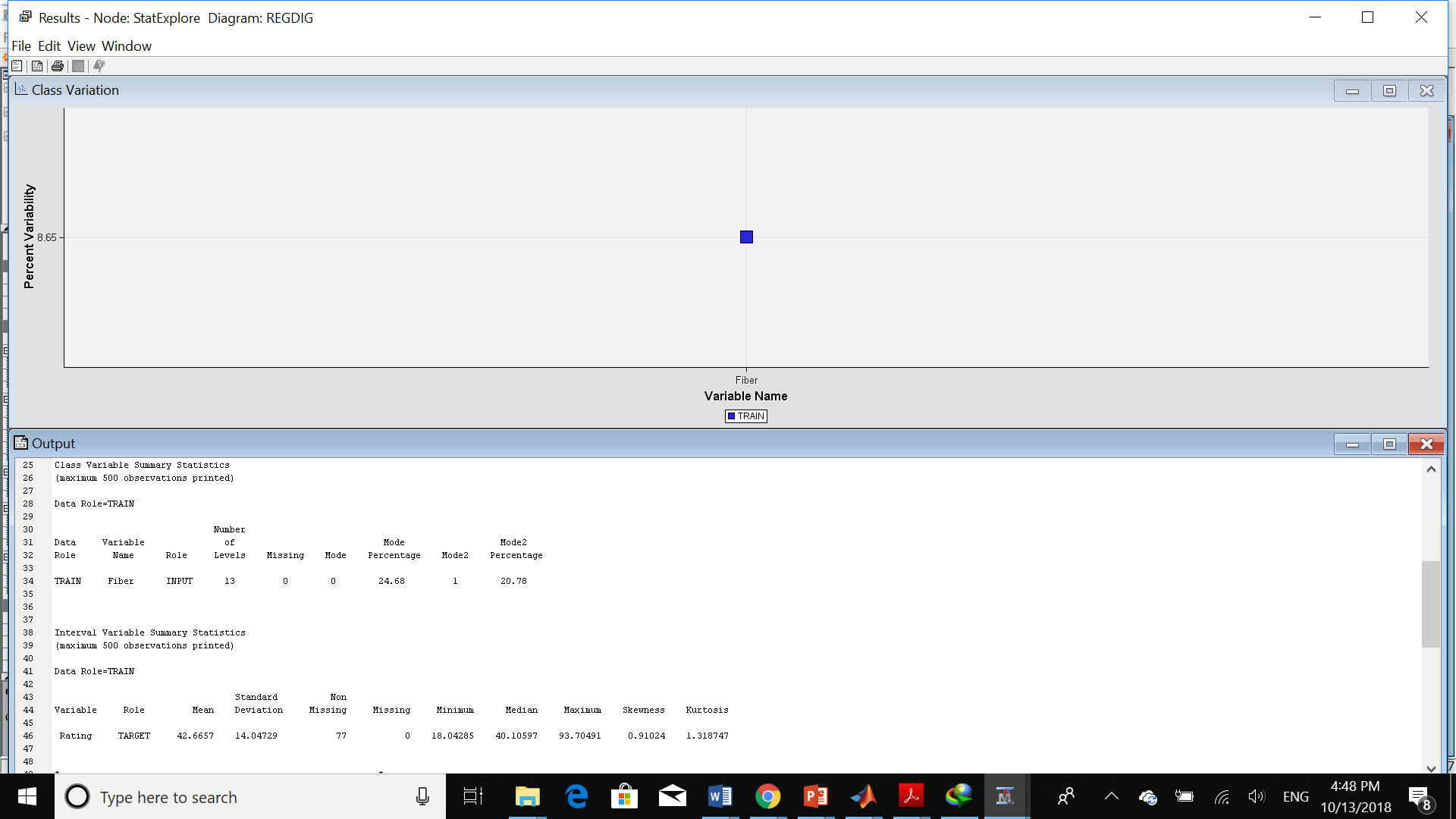




Then I create another node for StatExplore in order to see if we have any missing values.

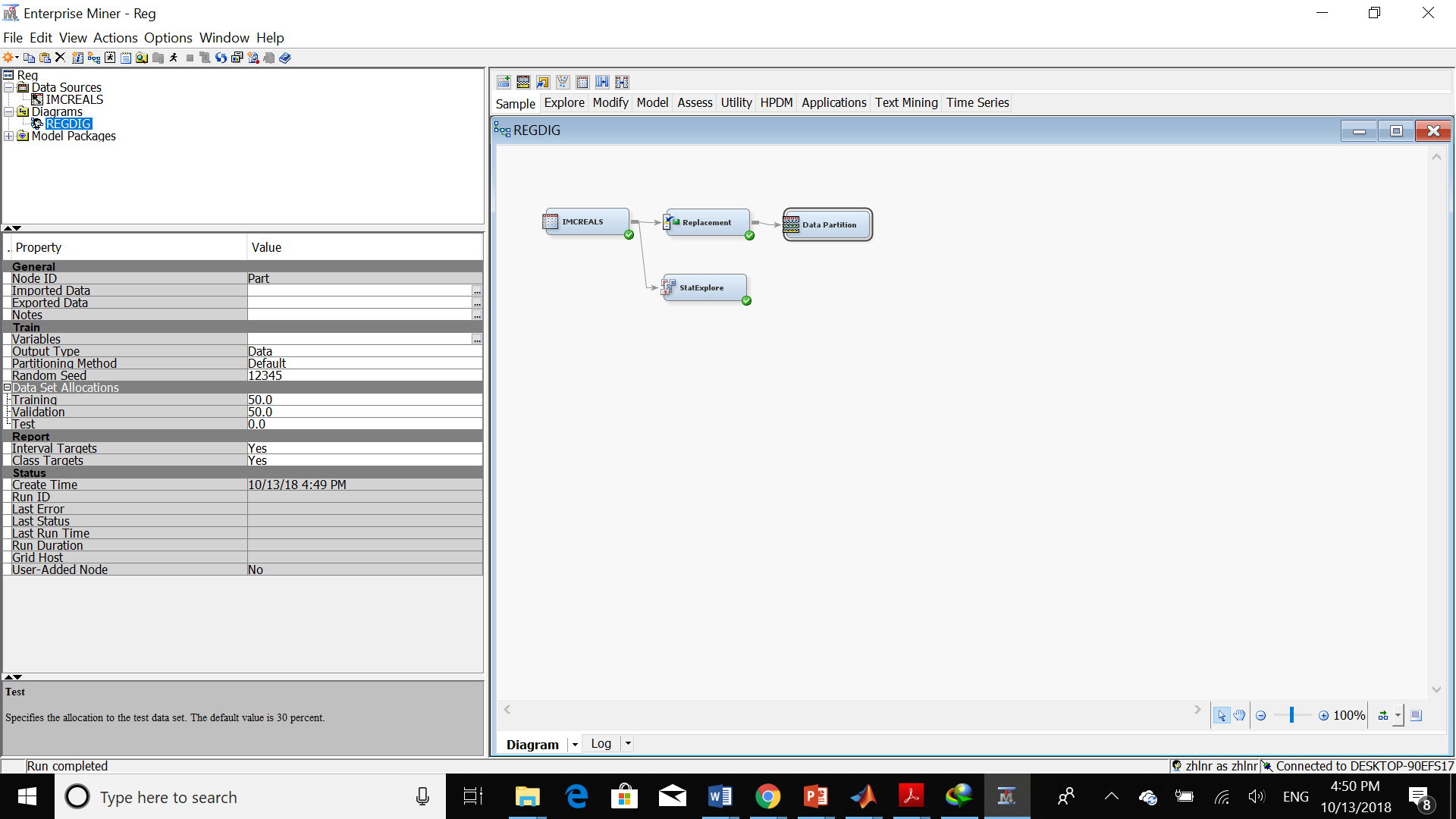


Now we can see the result:



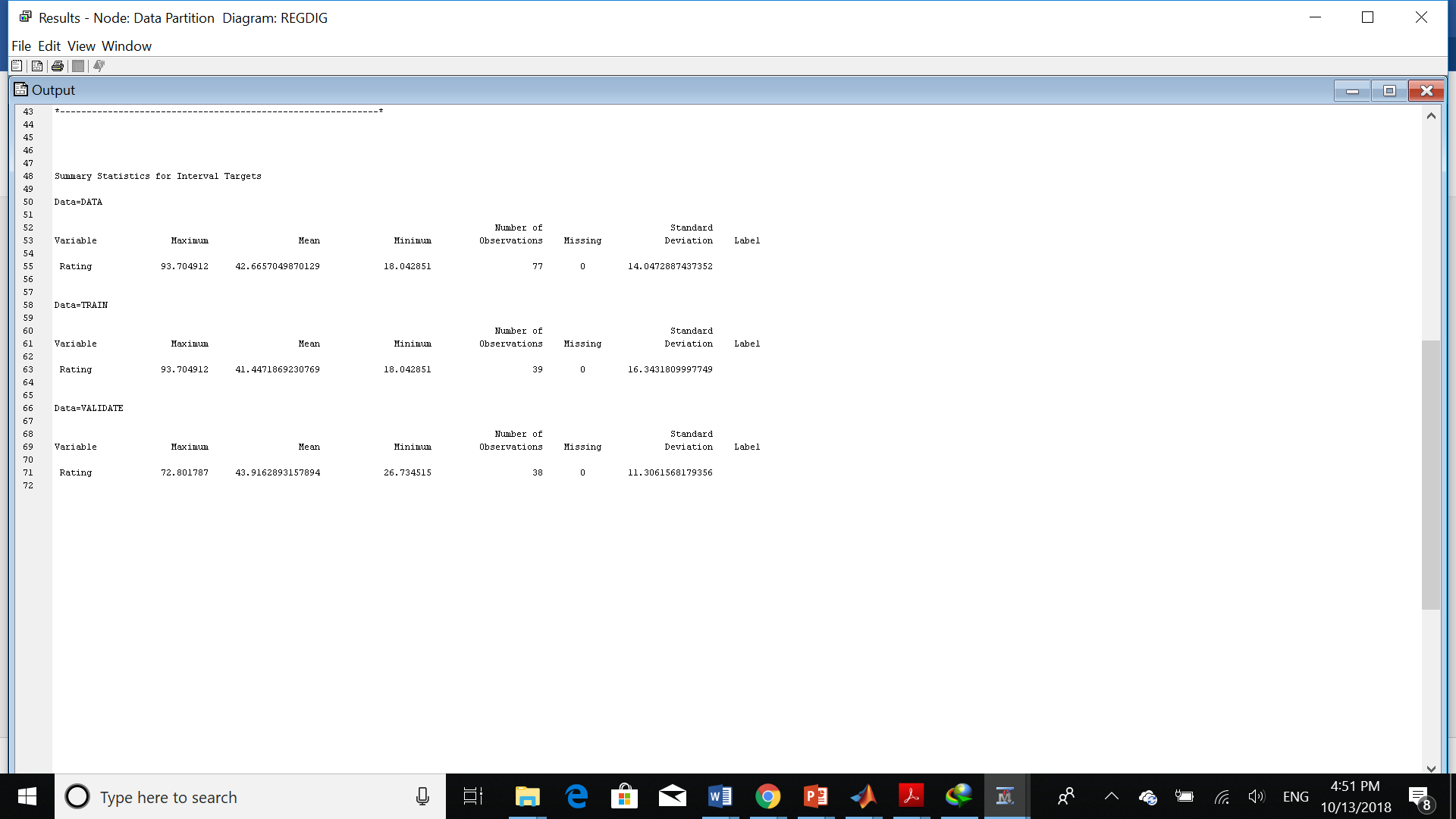
We can see that we don’t have any missing values.

Now, we should create node data partition:

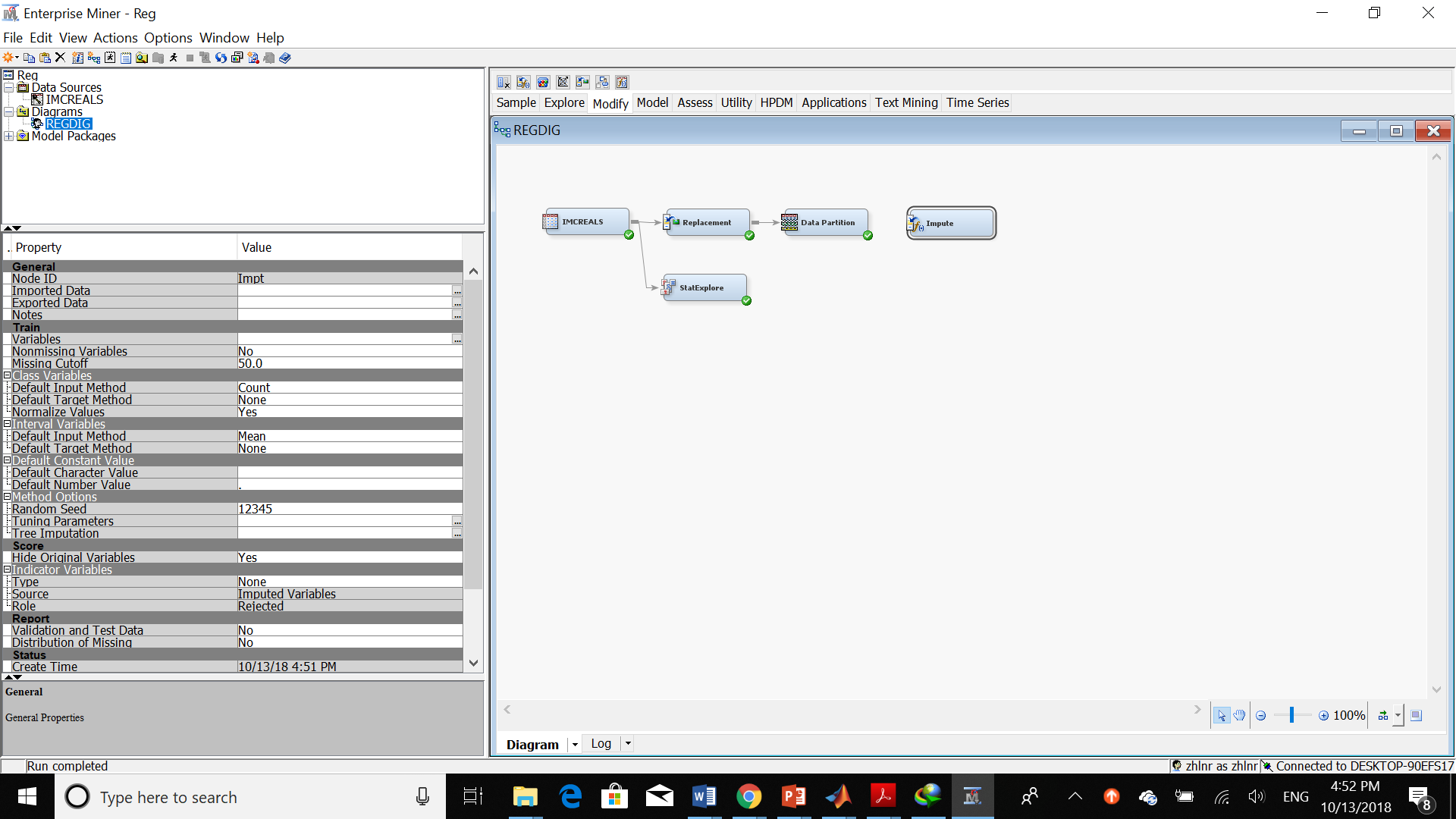


We should change the Training and Validation to 50%. Here, is the result:

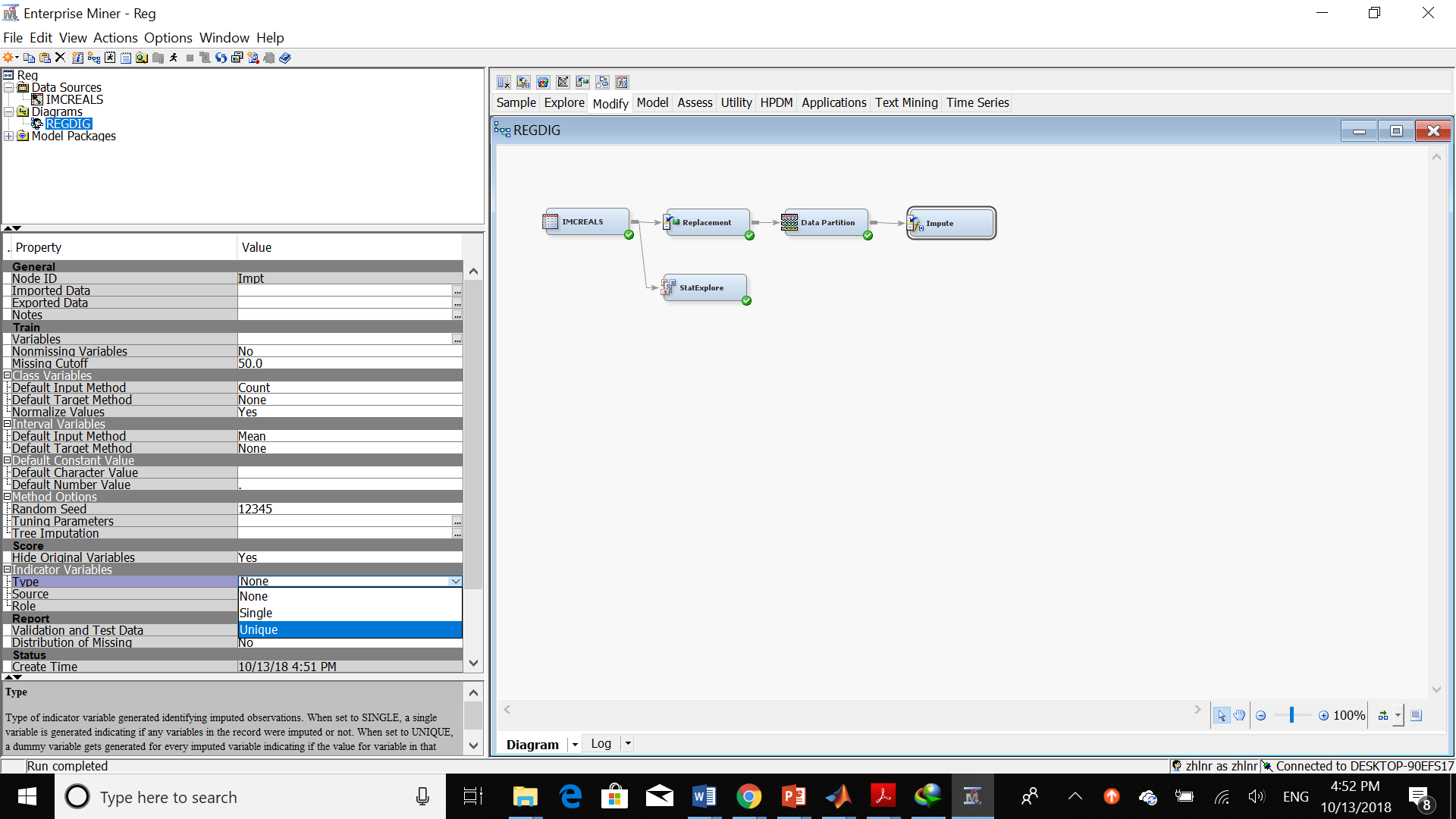




Next step is adding impute node. Although we don’t have any missing values for fiber and rating, I add this node according to the tutorial:

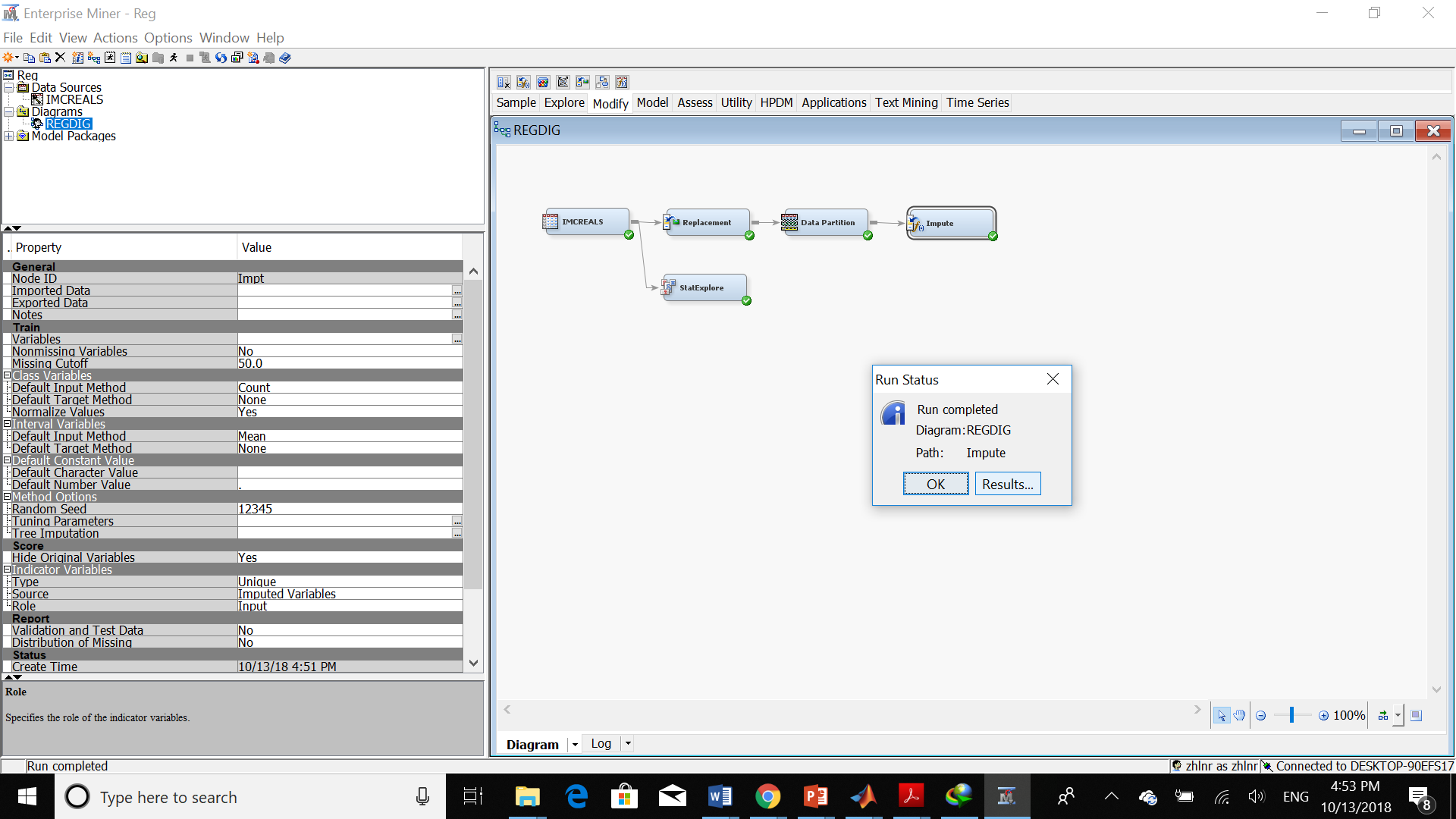


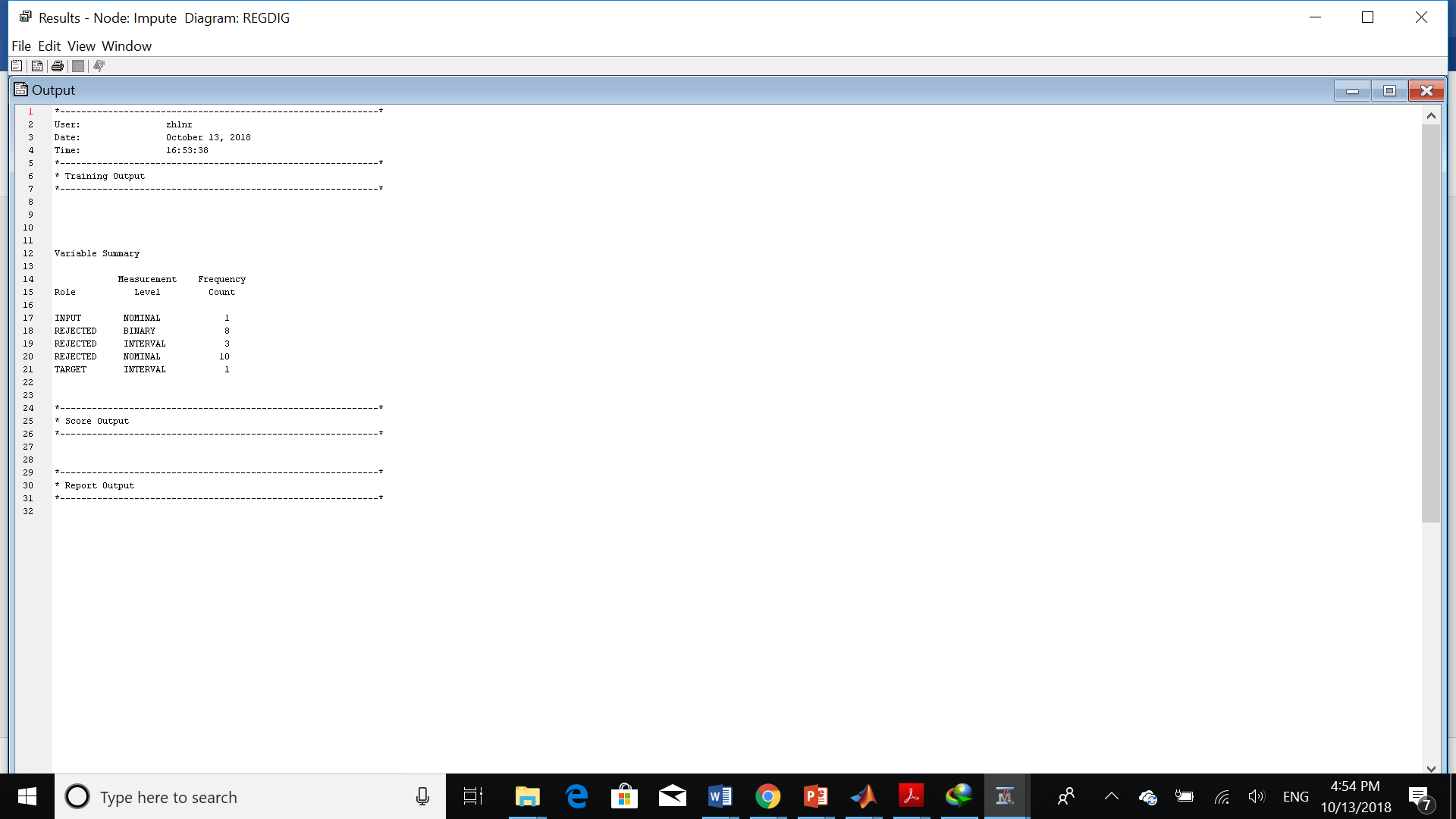
Then we should change the following things:



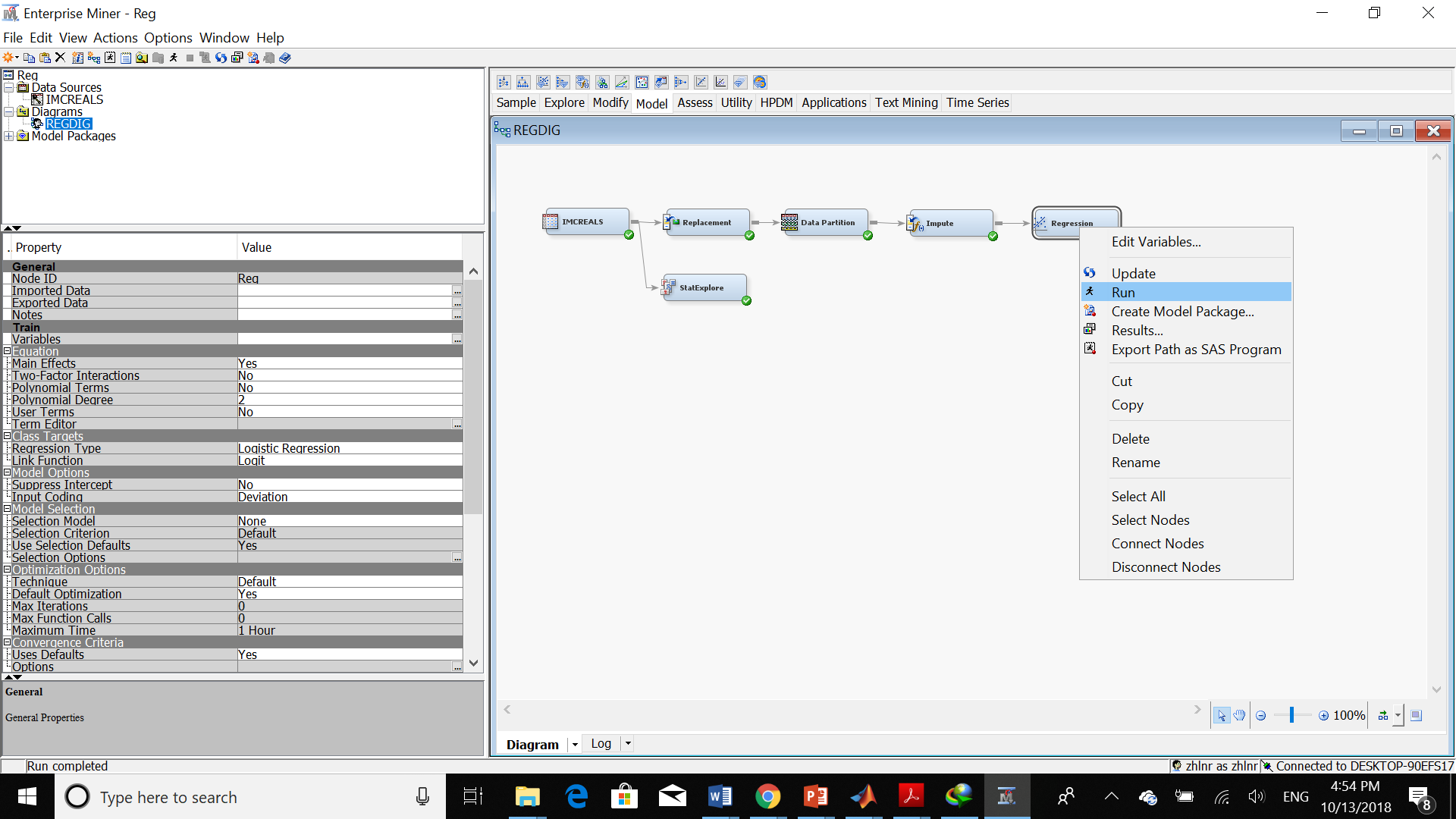


Then we should run the impute node:

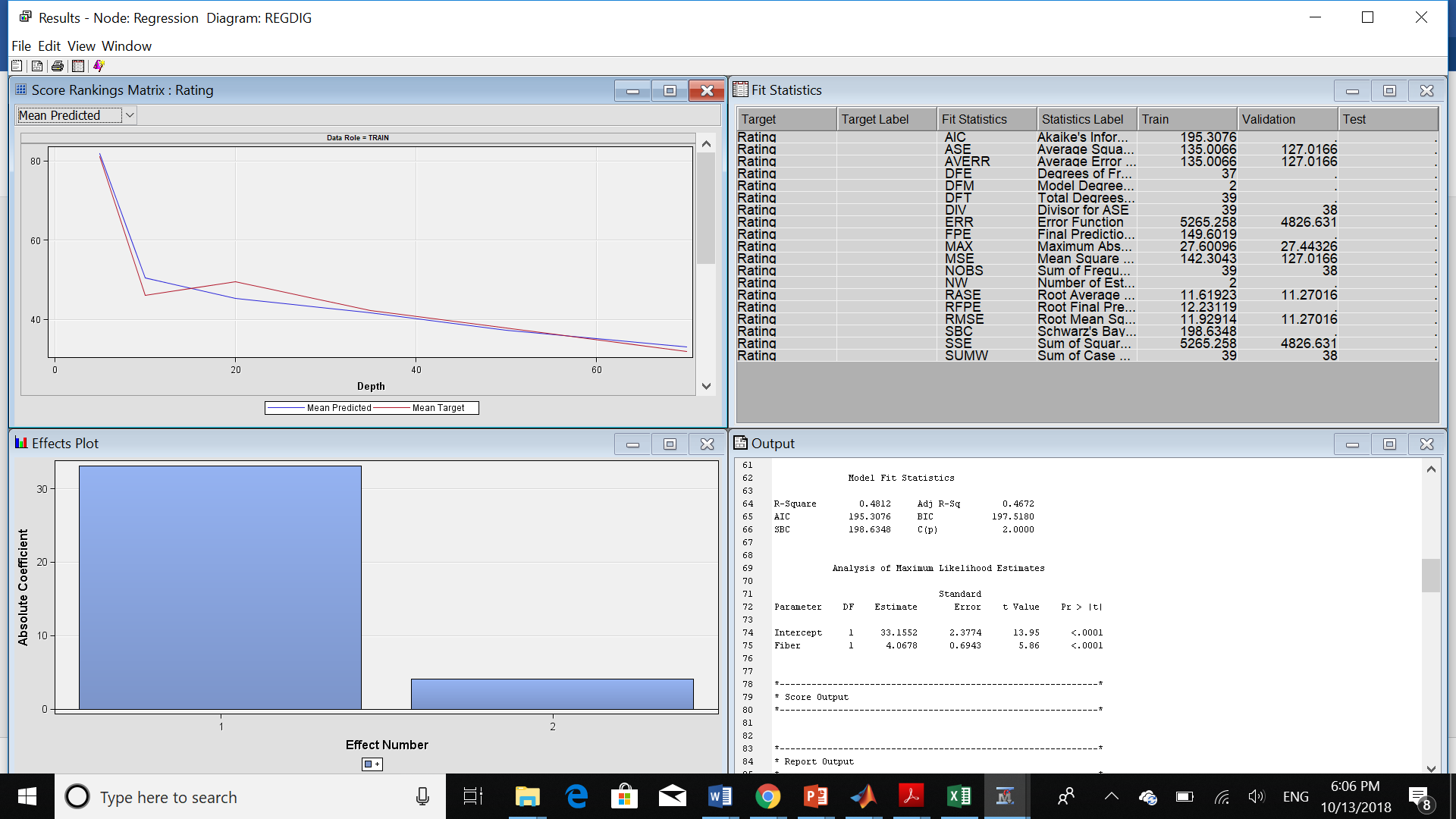




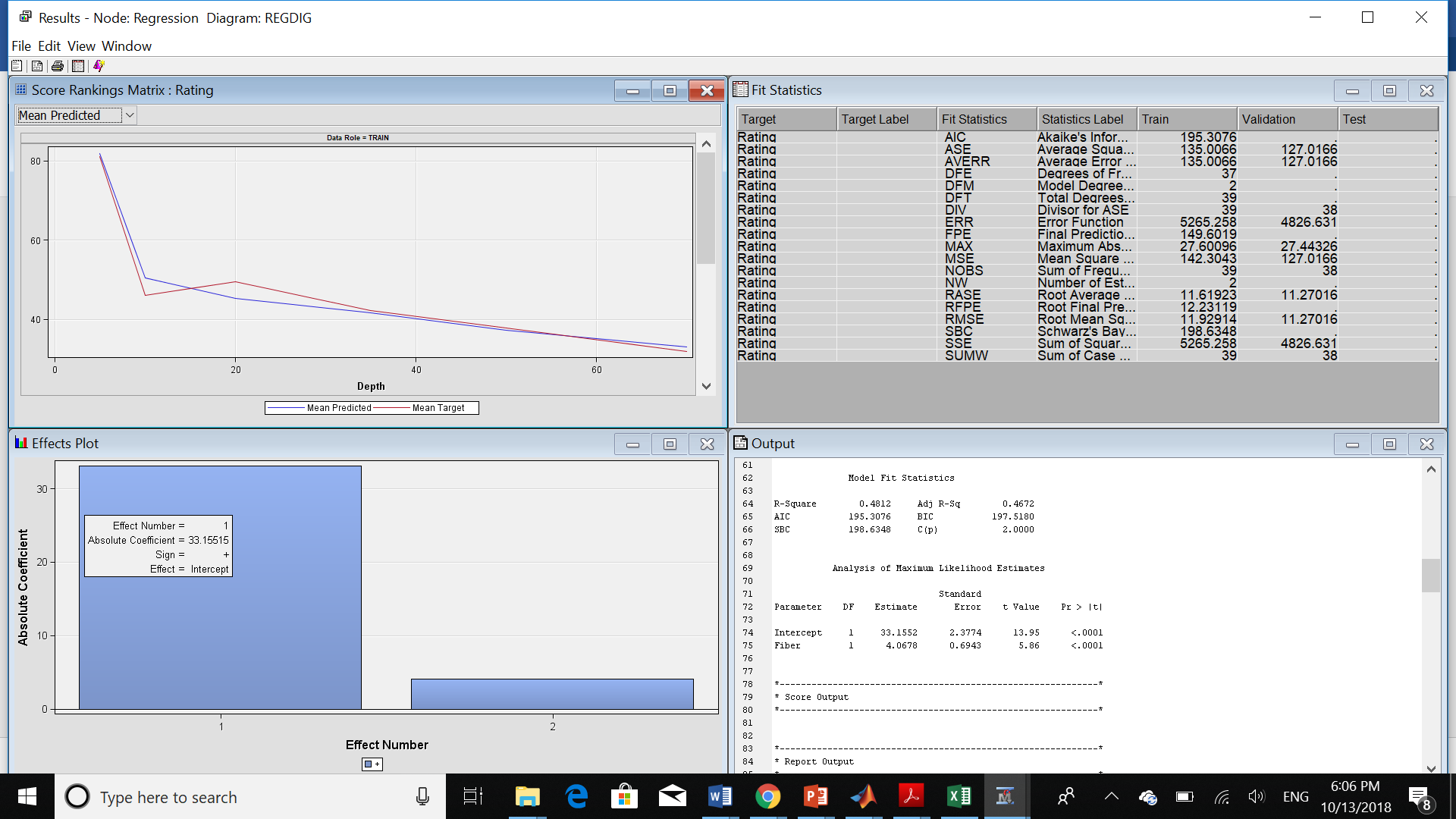
The next step is adding regression node:

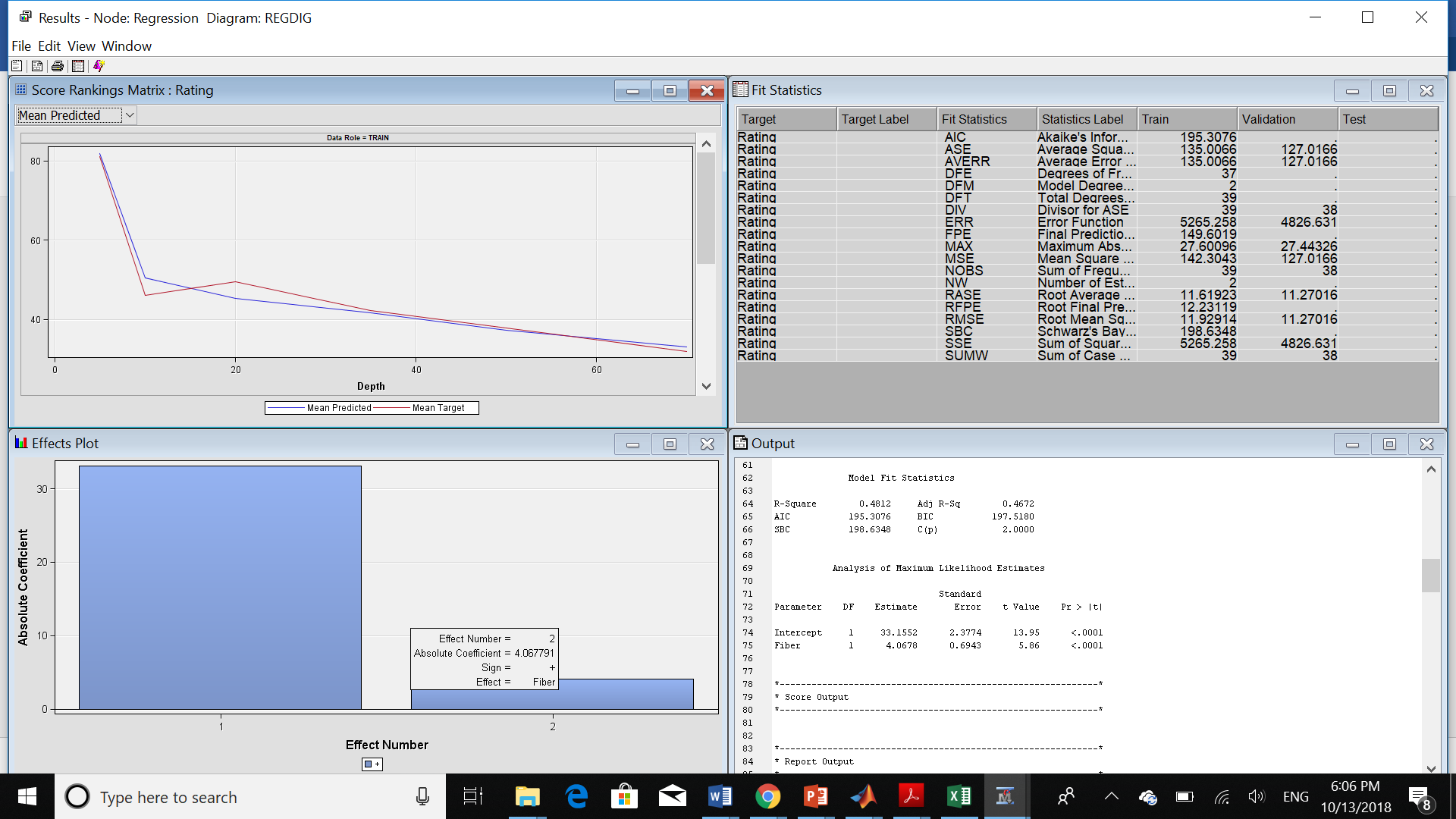


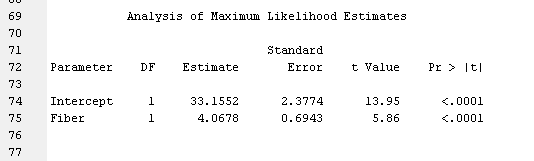
Here is the result:



We can see the coefficient as shown below:







1. **What is the estimated regression equation?**

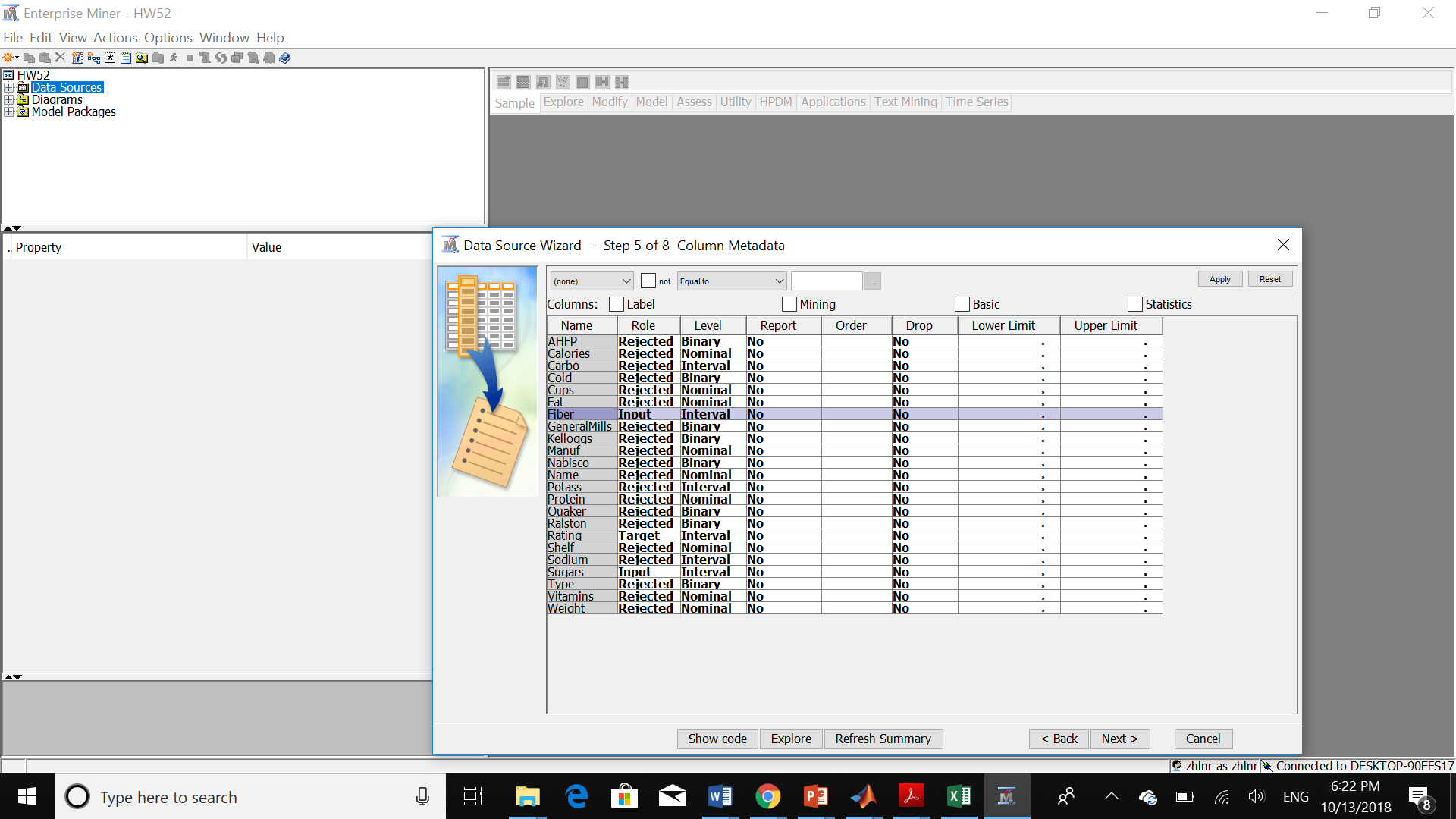
Rating = 33.1552+4.0678\*fiber

1. **Explain clearly the value of slope coefficient obtained in the regression.**

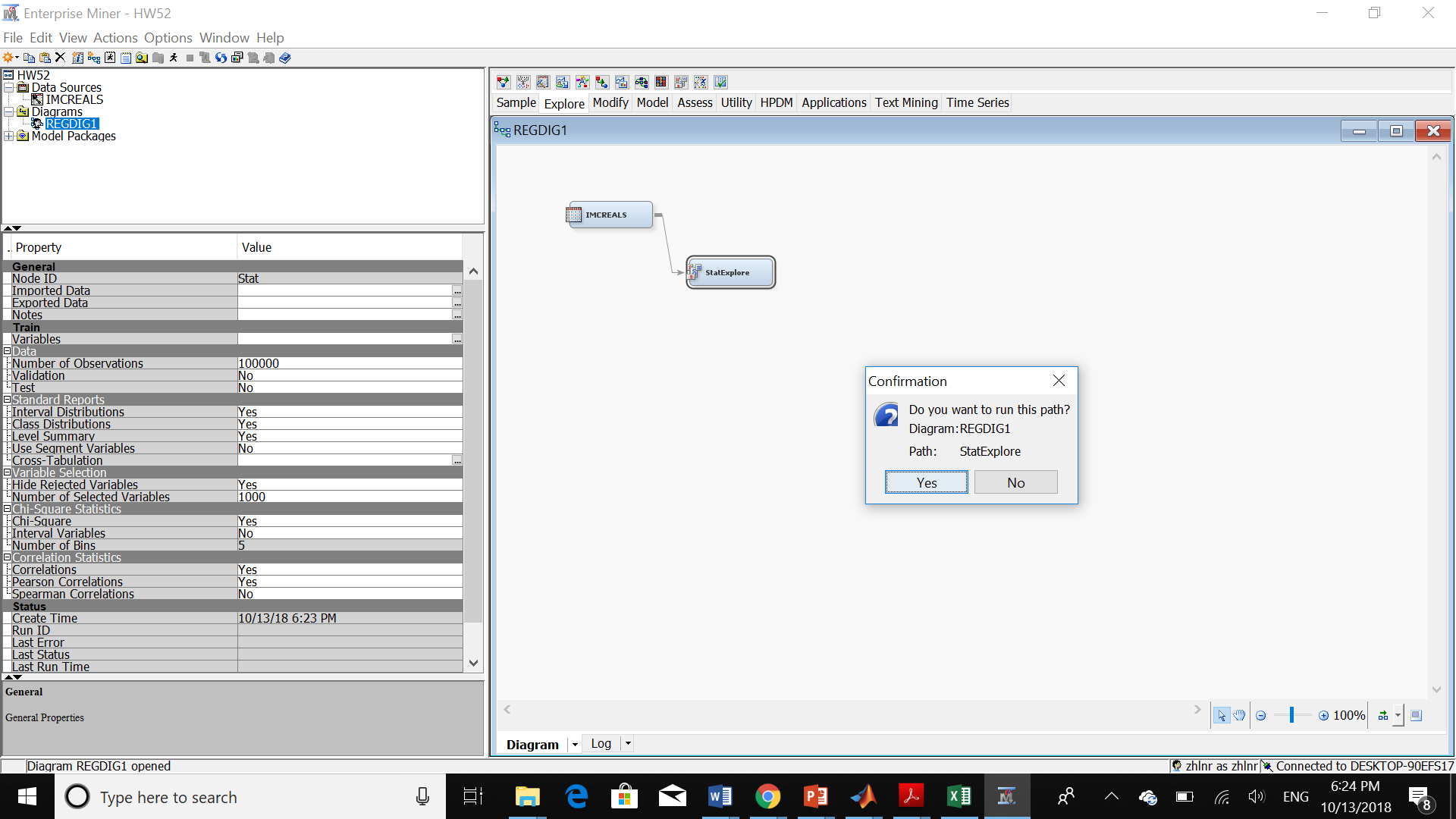
In this analysis we just calculated the rating value base on fiber value. So, we have two values for our linear regression formula. The Intercept and the coefficient for fiber which is the slope of the line. So, we can easily figure out that the slope coefficient here is 4.0678.

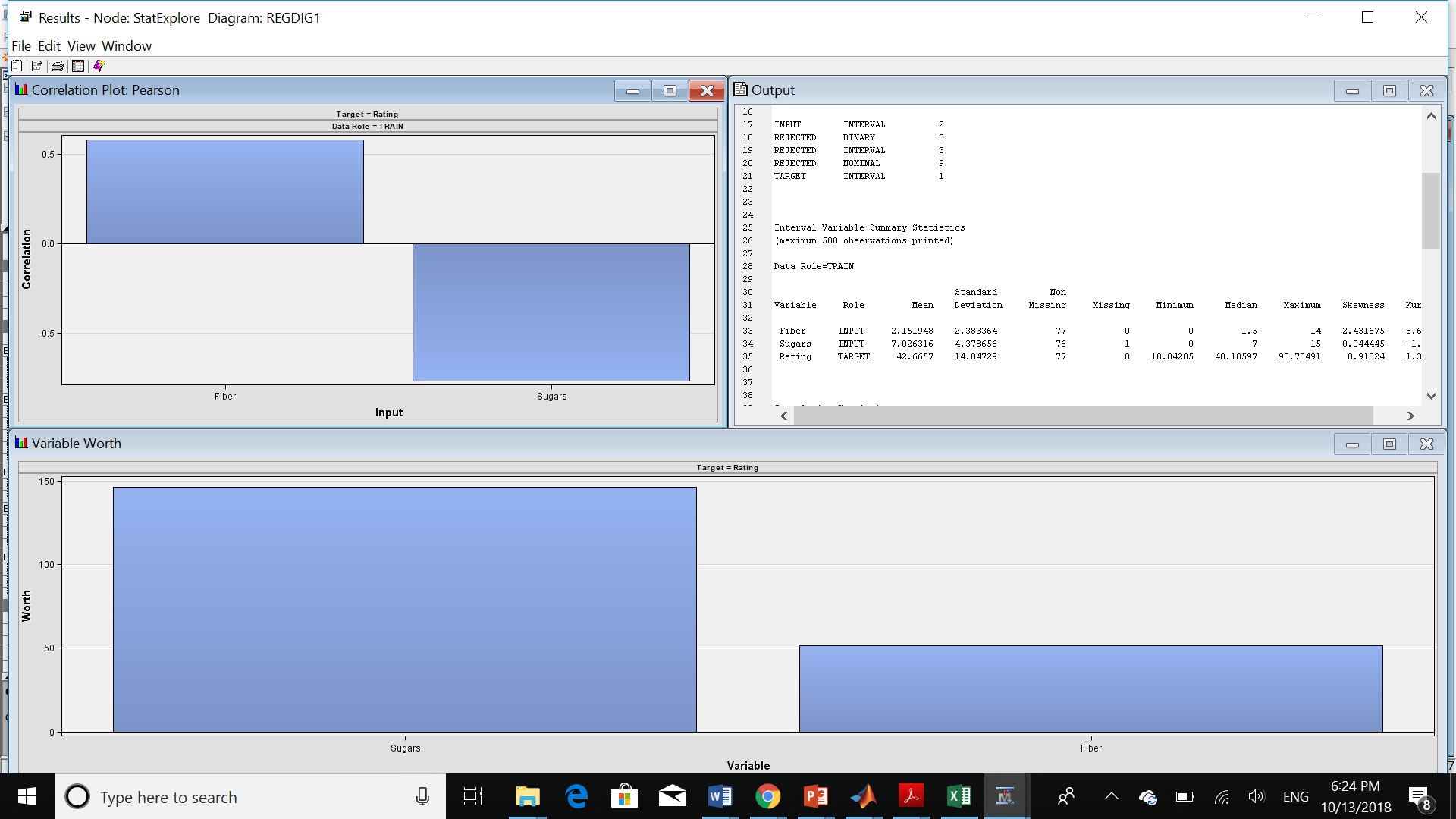
In order to answer to the next 3 questions, we should follow these steps:

Because we want to calculate the rating value based on sugar and fiber, we should change the role of the others as rejected and change the level of these variables as Interval instead of nominal.

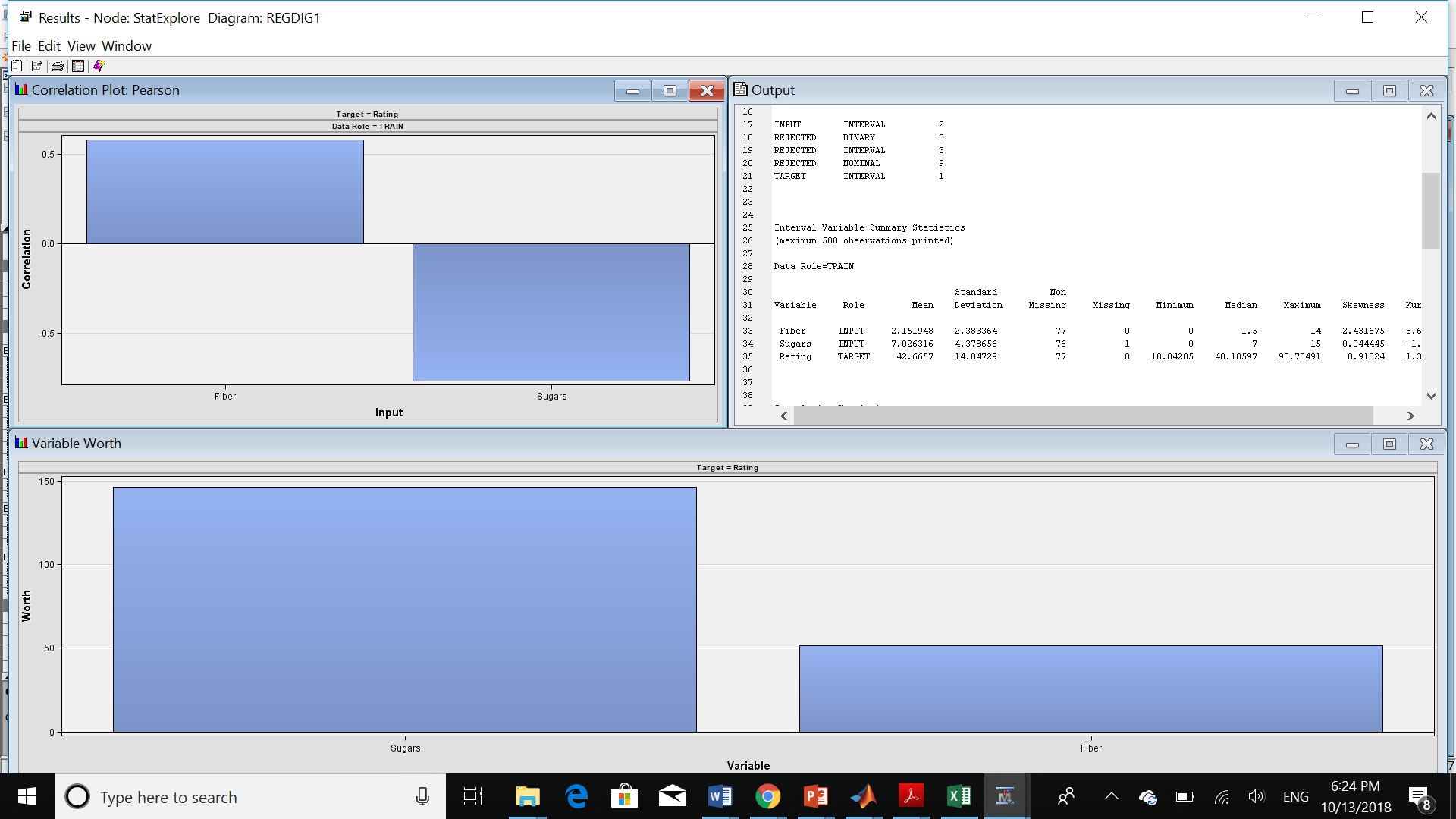


After creating the data source, I created the statexplore in order to find out if we have any missing values.

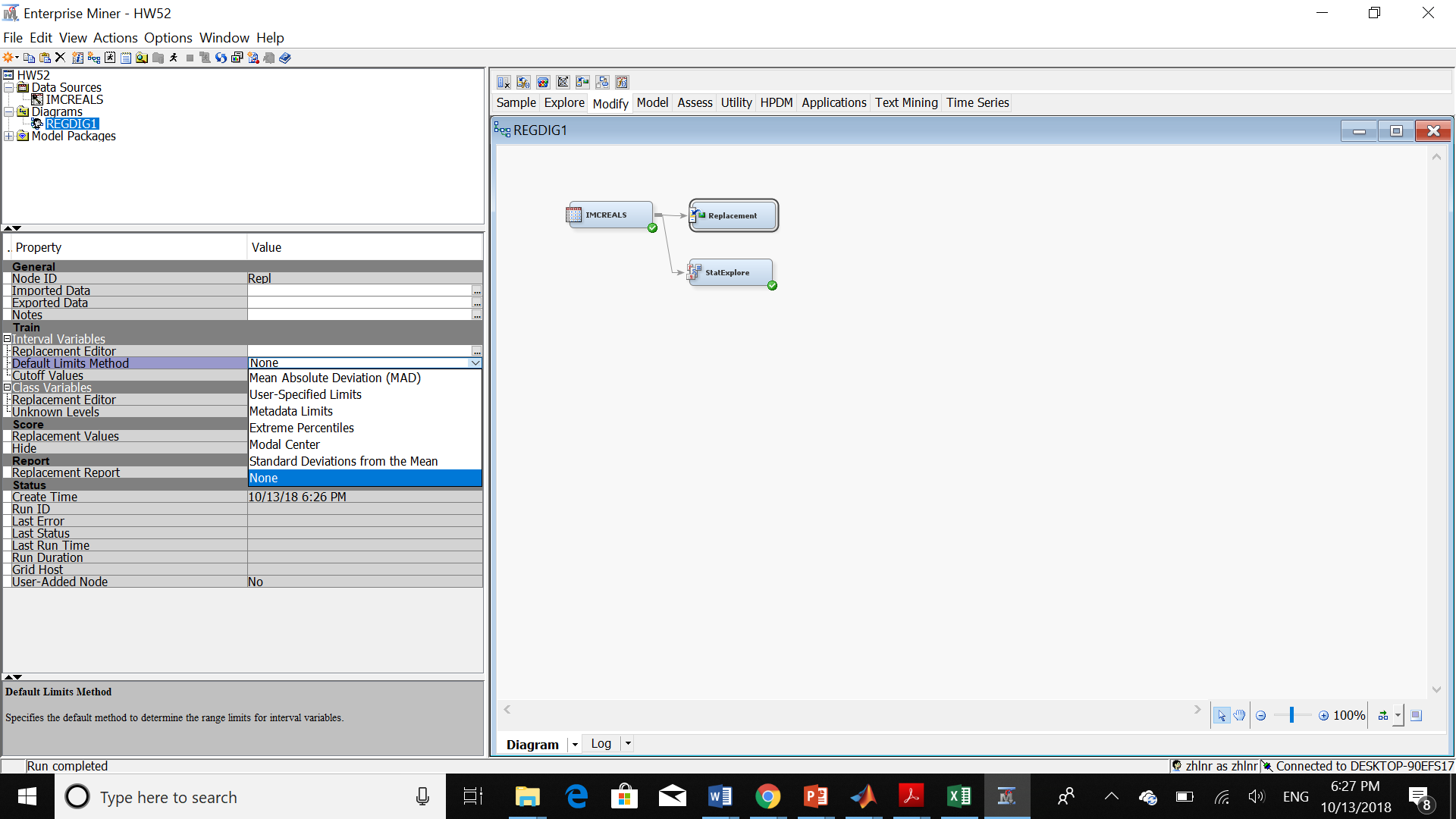


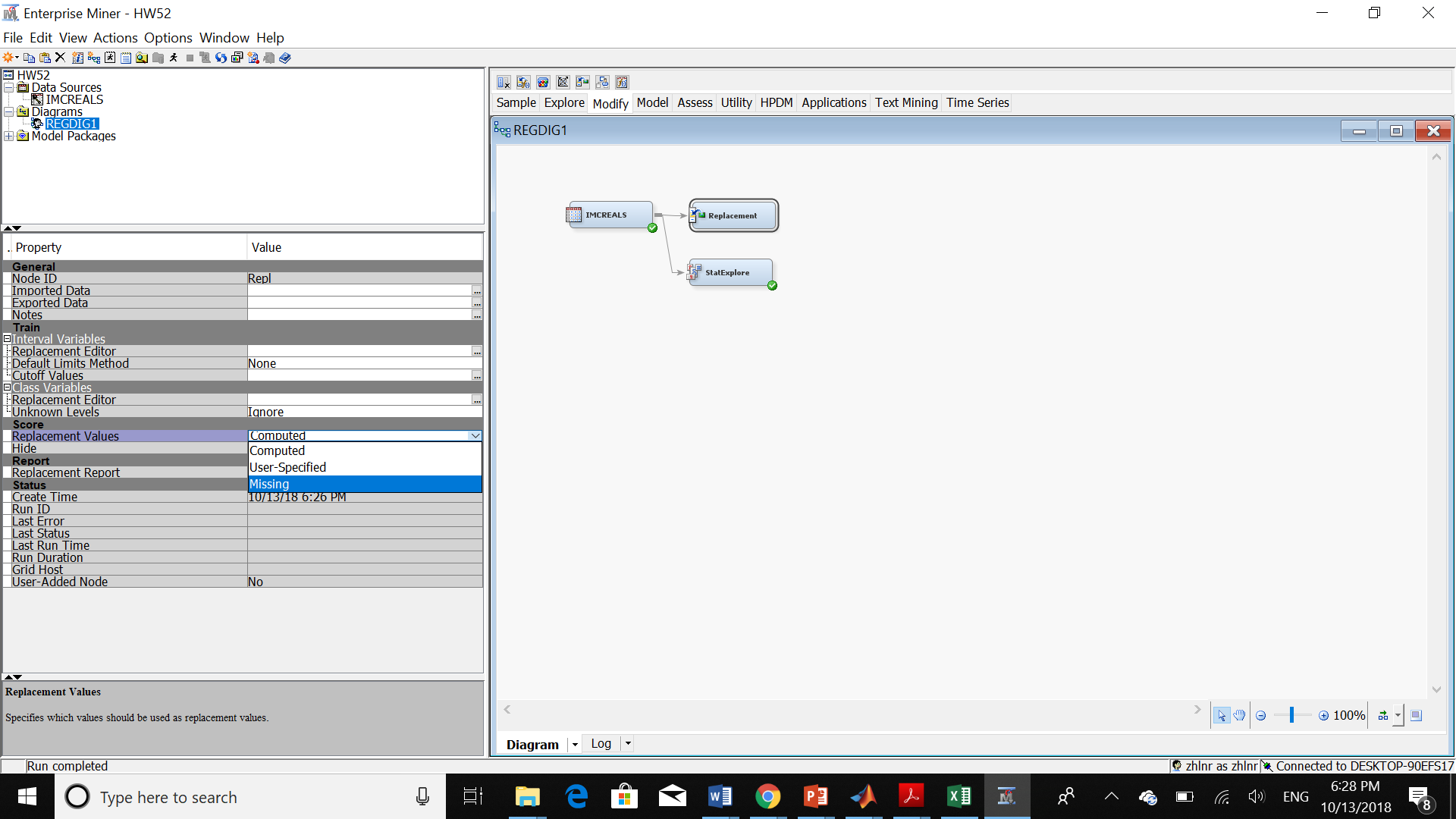


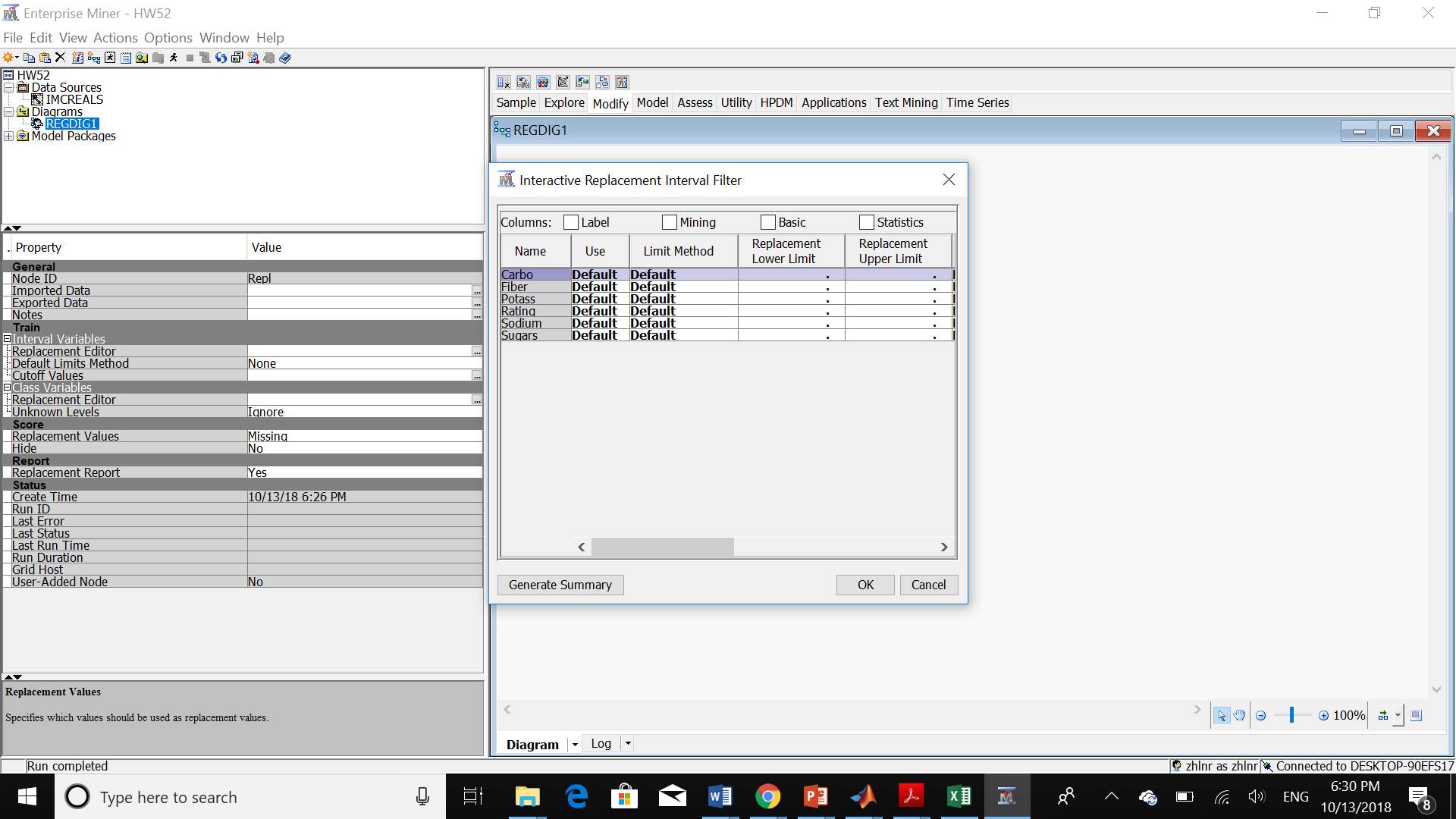
So, as we can see here we have 1 missing value for sugars.

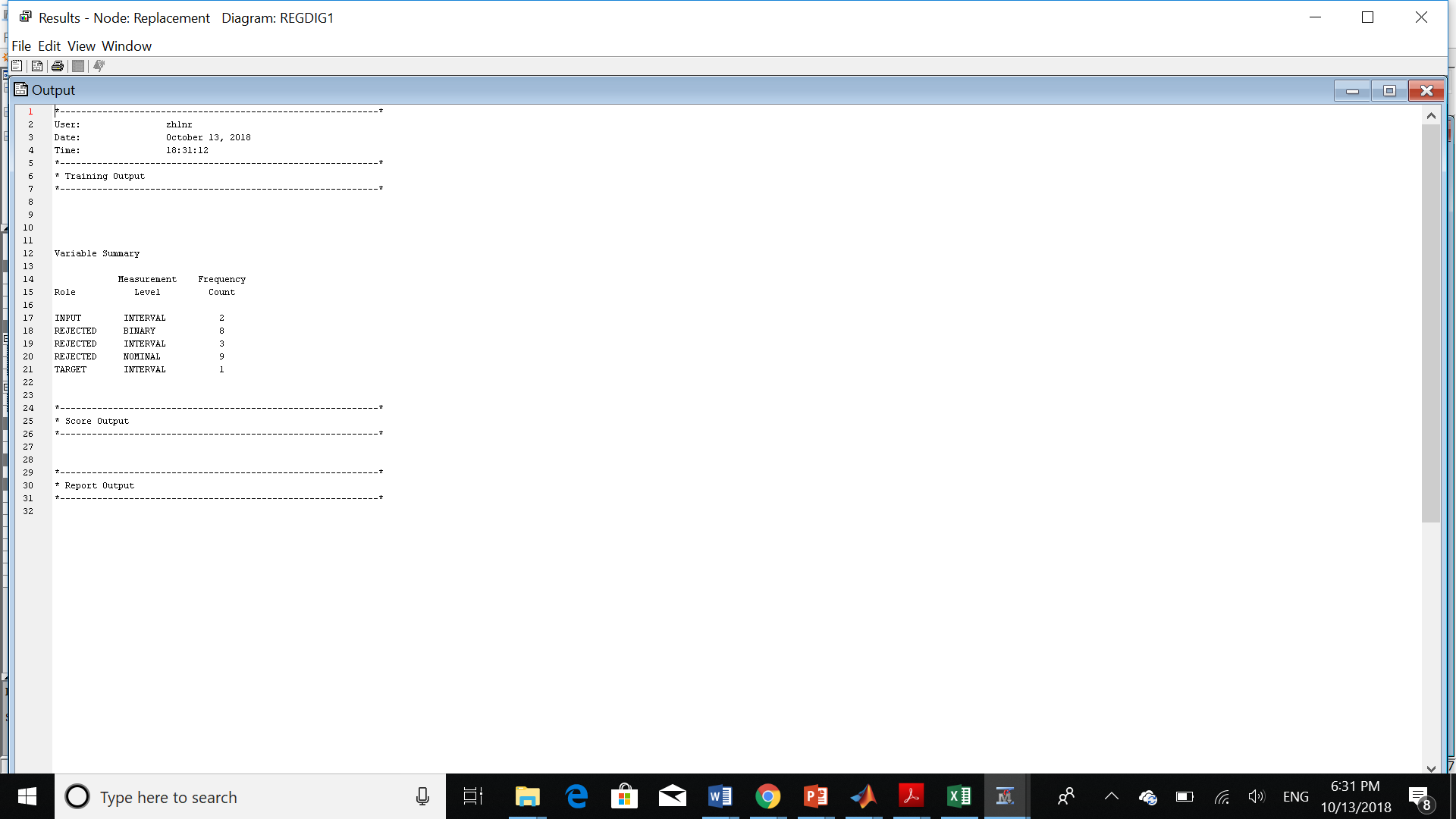


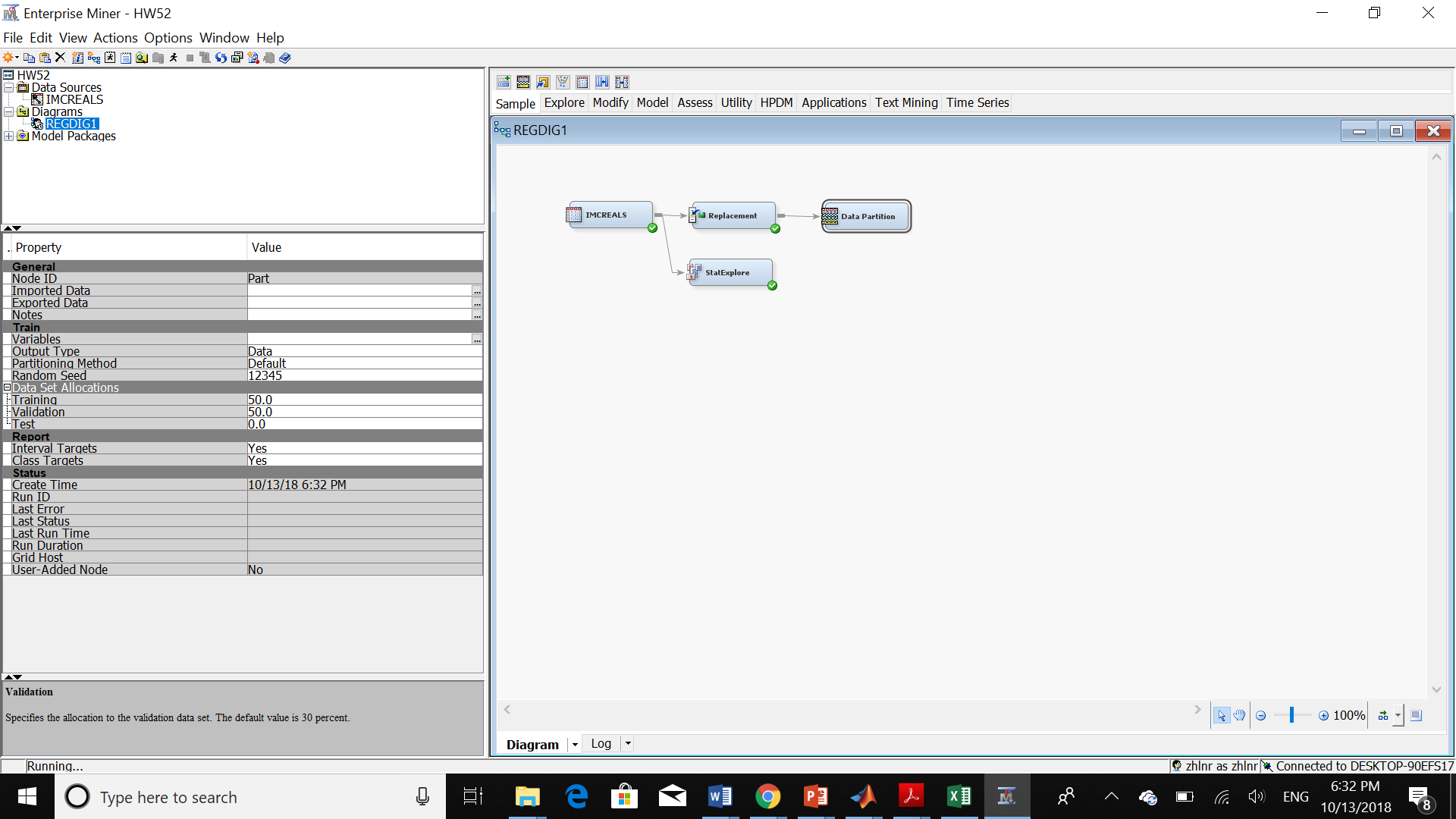
Then, we should follow the others steps as illustrated before:

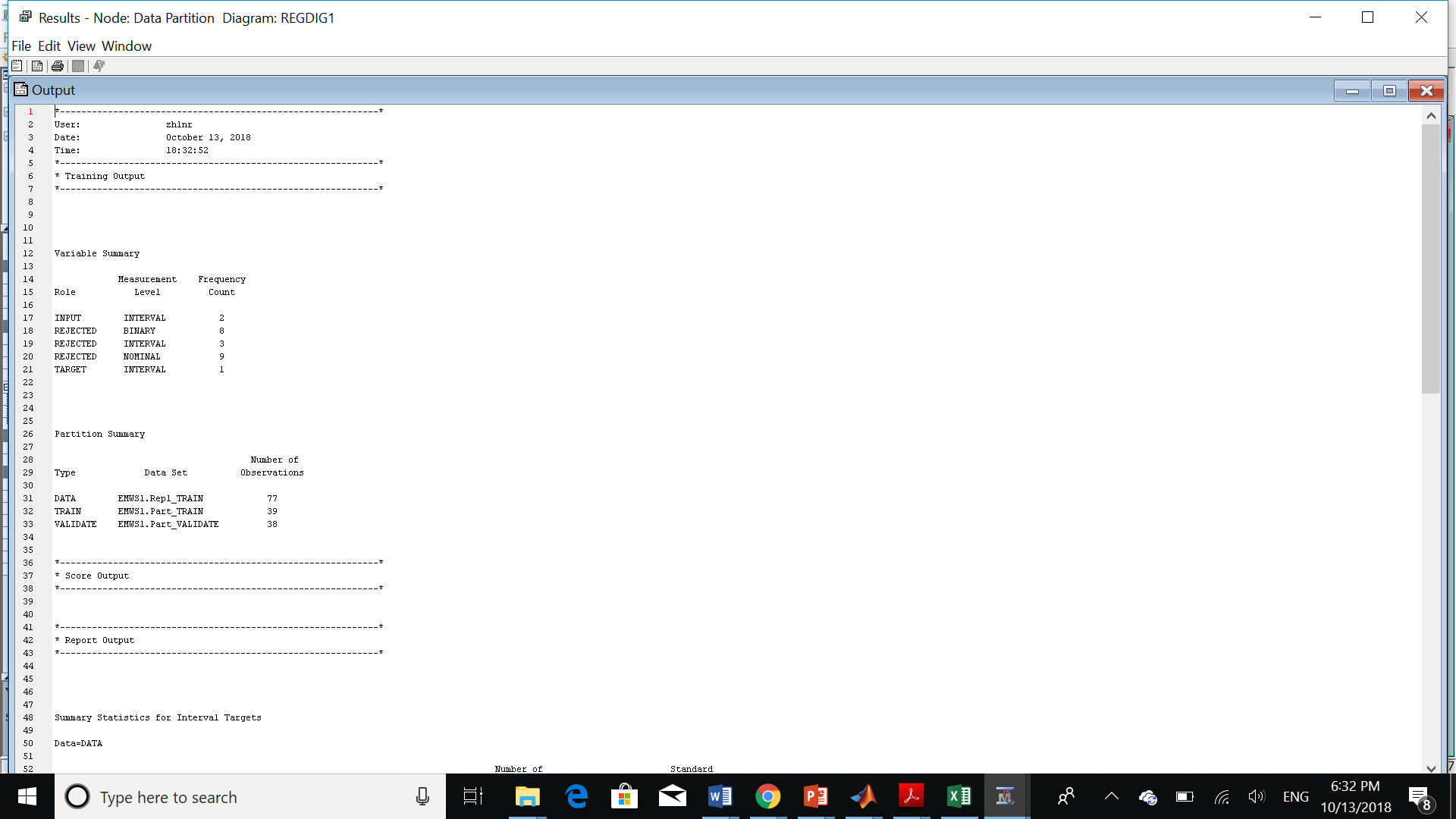


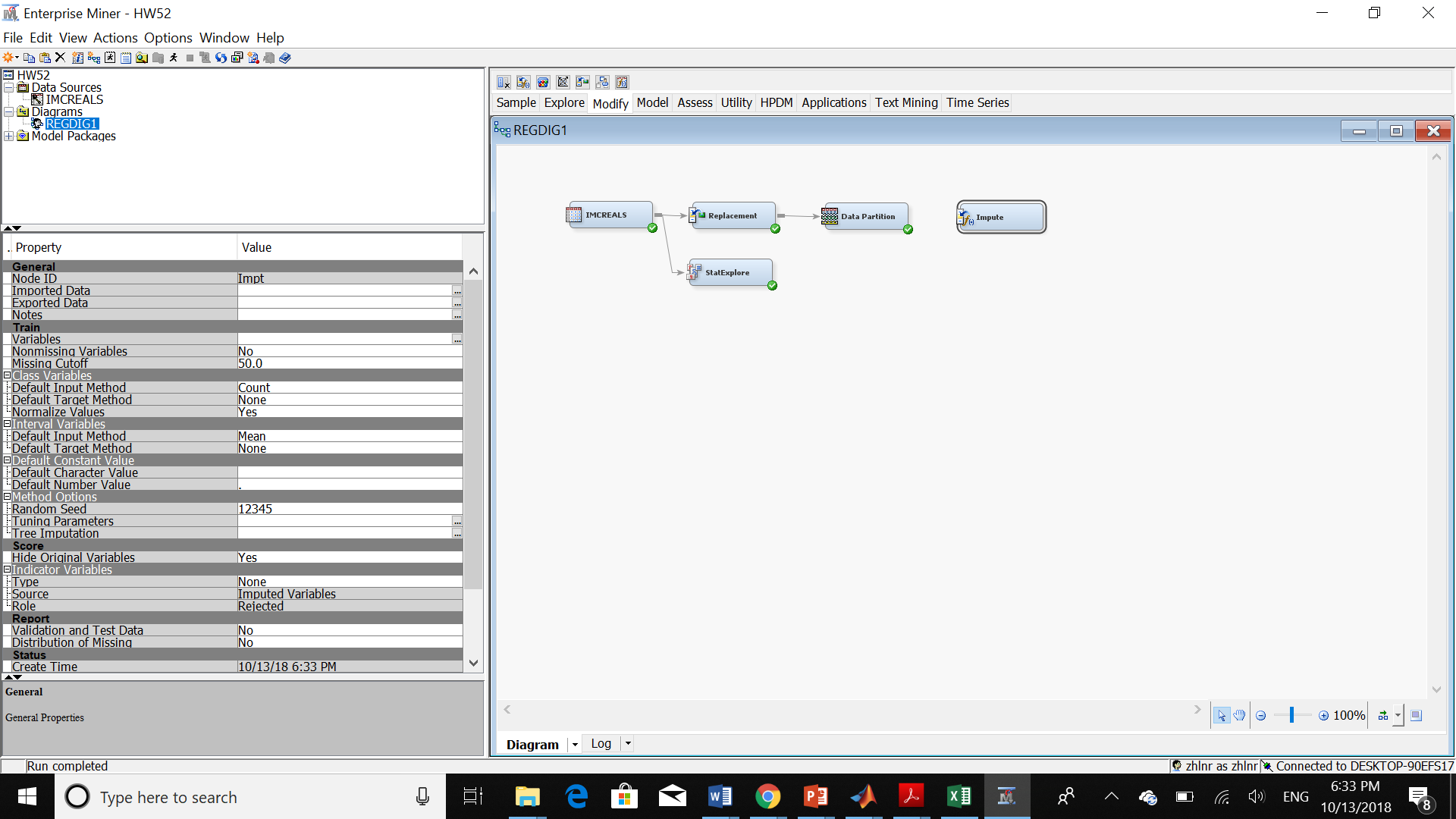


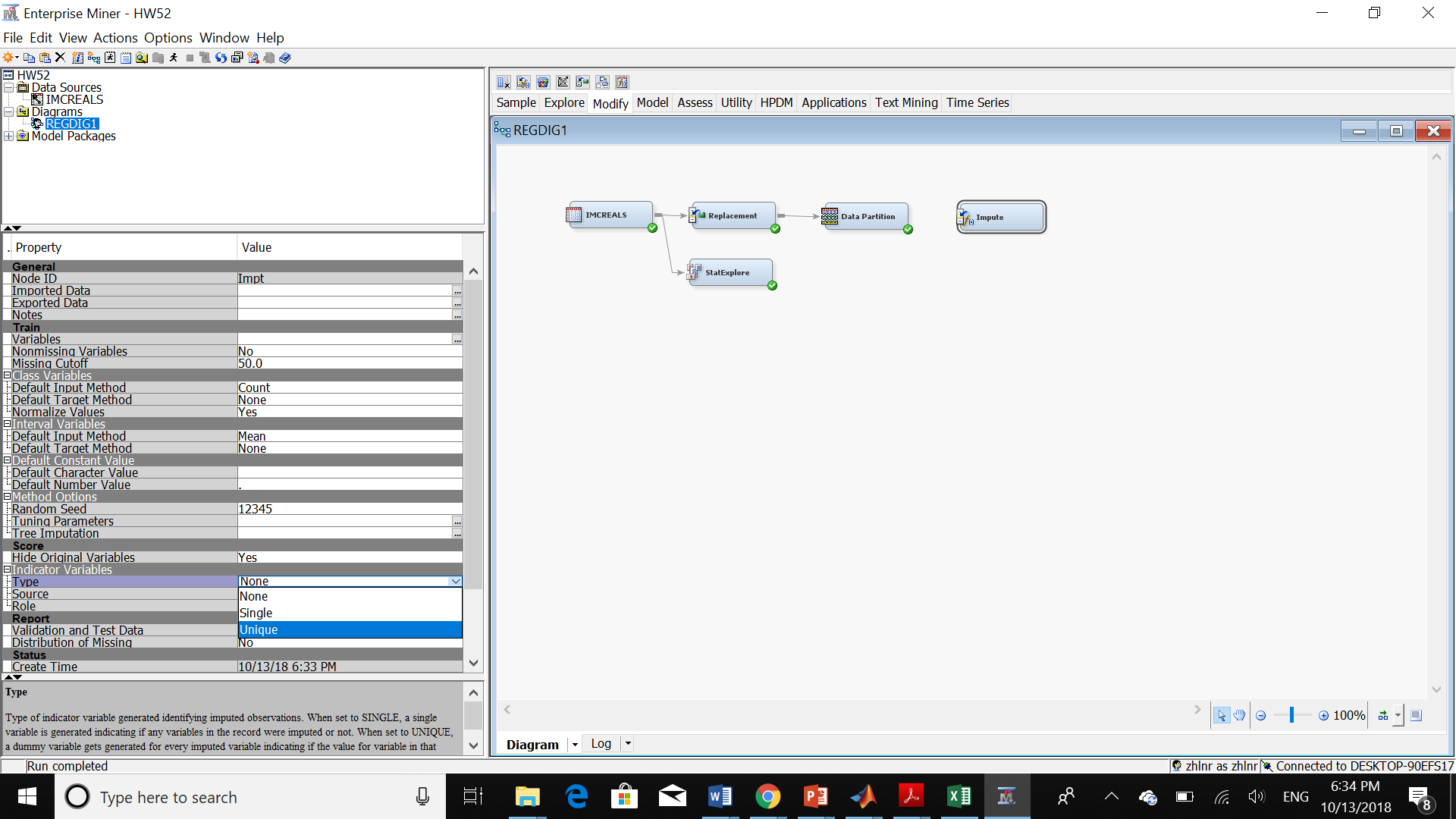


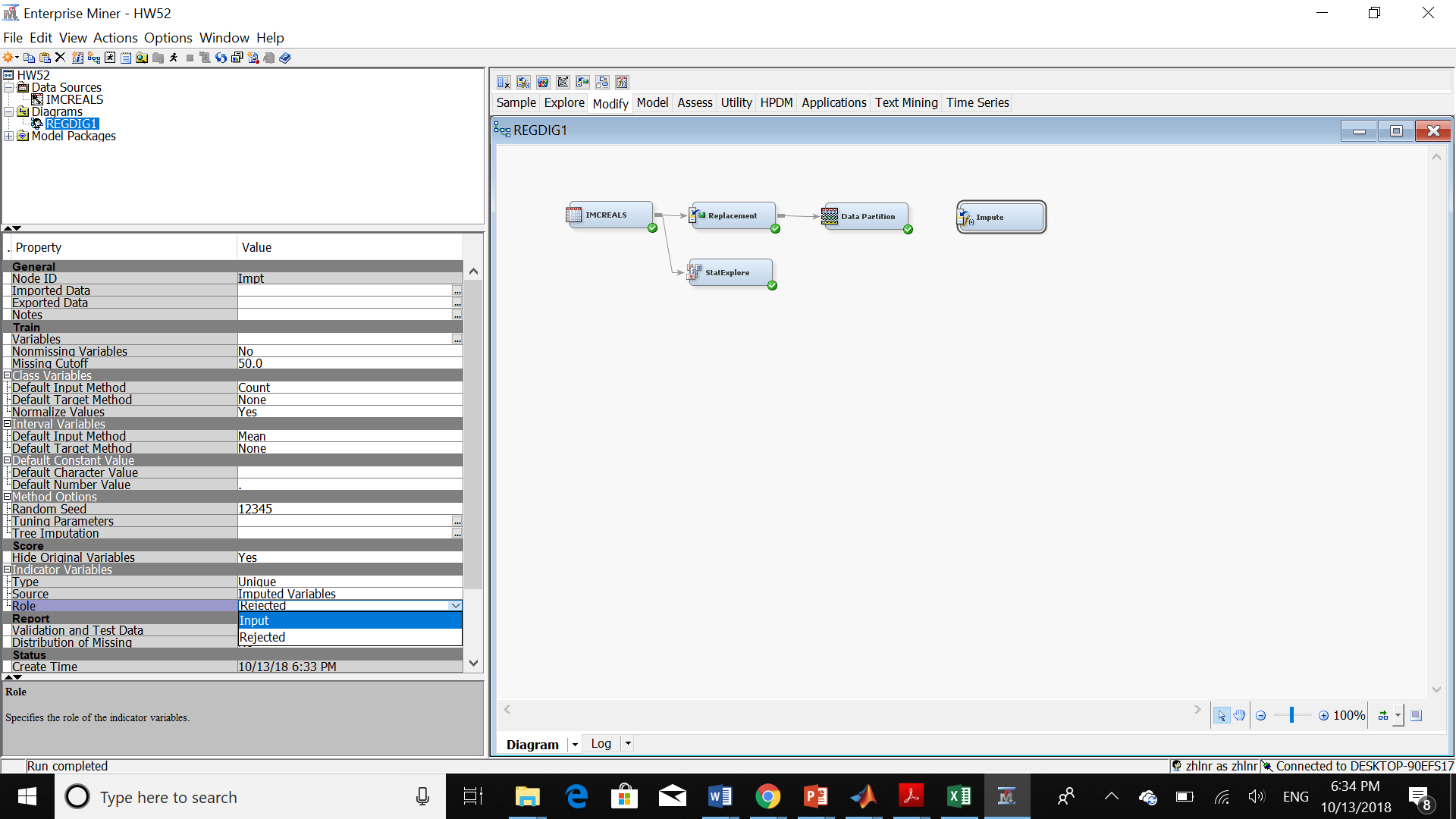


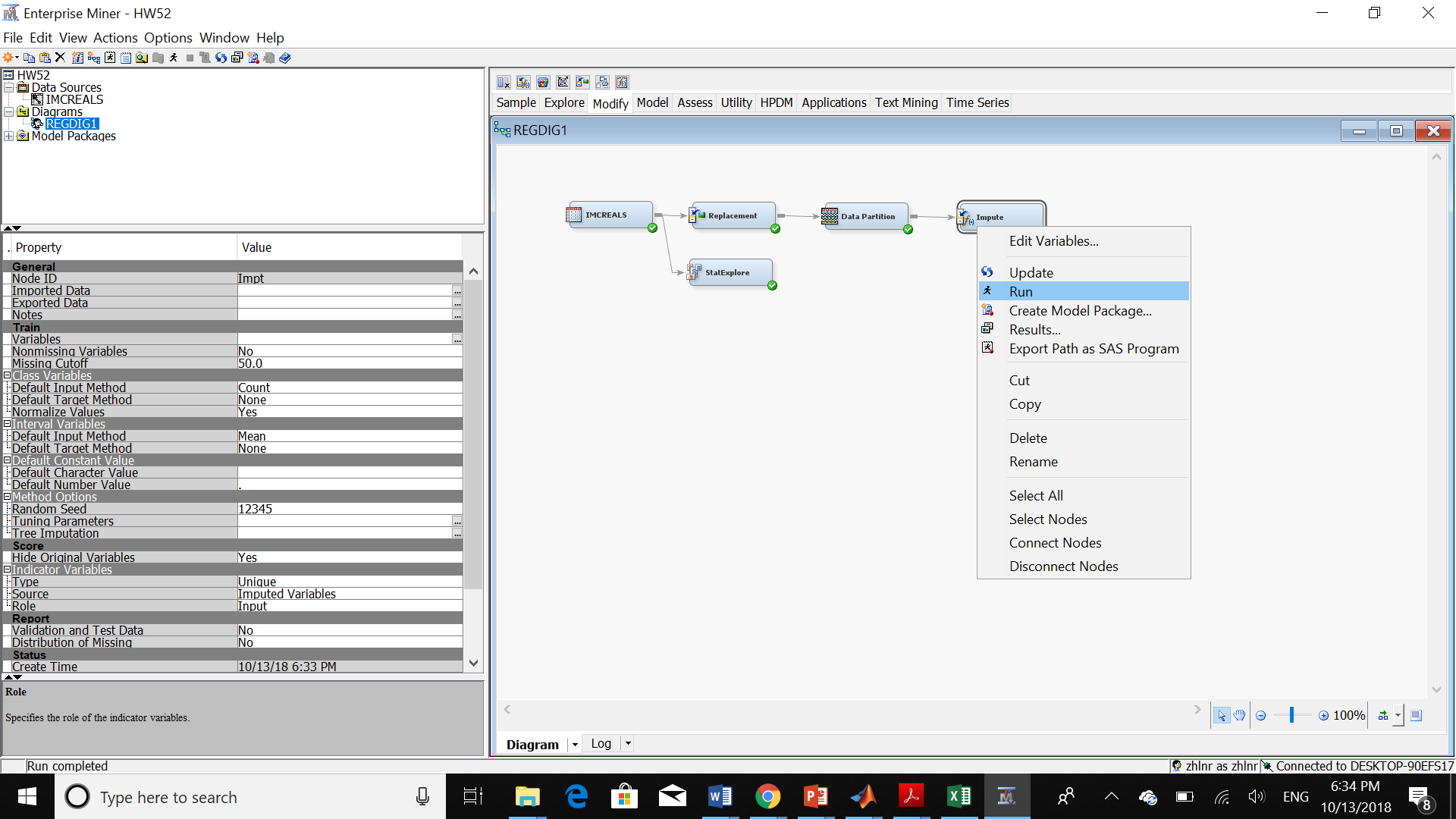


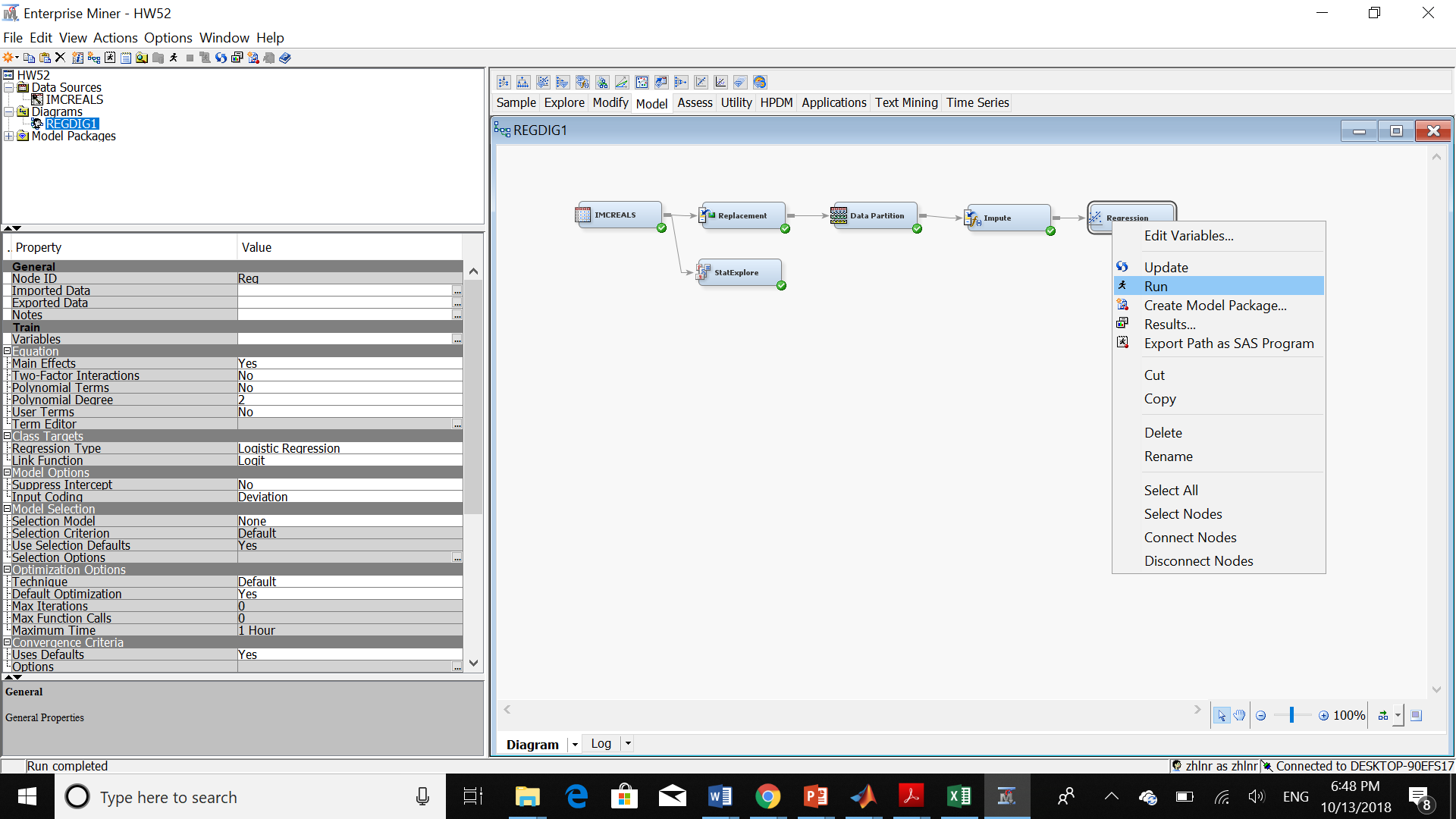




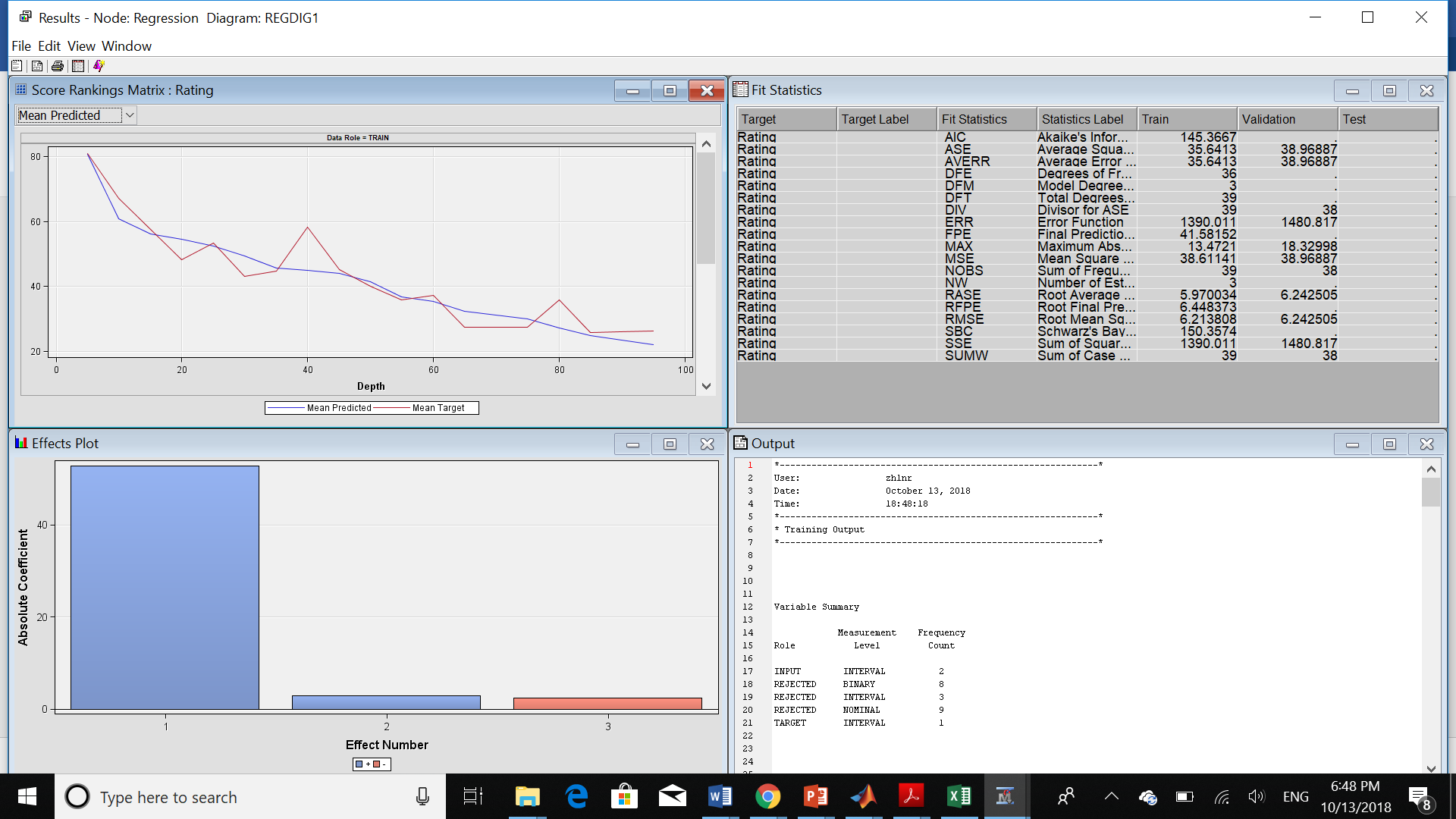








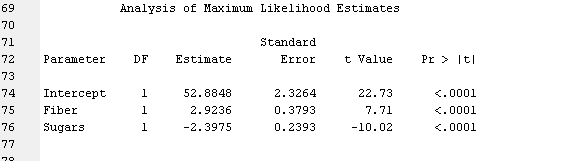
Now, we have the result of our regression:



1. **What is the estimated regression equation?**

As we can see in below image, the Intercept value is 52.8848. The coefficient for Fiber and Sugar are 2.9236 and -2.3975 respectively. As a result, the regression equation is:

Rating = 52.8848+2.9236\*fiber-2.3975\*sugars

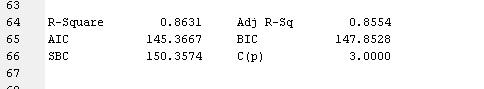


1. **Explain clearly and completely the value of the coefficient for fiber you obtained in the regression.**

As we showed in previous question: the coefficient for fiber is: 2.9236

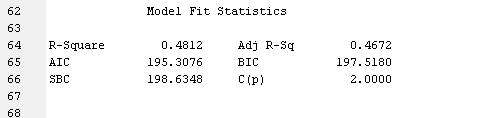
1. **Compare the r2 values from the multiple regression and the regression done earlier in the exercises. What is going on? Will this always happen?**

Here we have the r-square value for multiple regression:



So, the r2 value for multiple regression is 0.8631.

Now, we can see the r-square value for simple regression done previously:



The r-squared value is 0.4812

R-squared is a statistical measure of how close the data are to the fitted regression line. It is also known as the coefficient of determination, or the coefficient of multiple determination for multiple regression.

R-squared is always between 0 and 100%:

* 0% indicates that the model explains none of the variability of the response data around its mean.
* 100% indicates that the model explains all the variability of the response data around its mean.

In general, the higher the R-squared, the better the model fits your data.

We can see that by adding another input for predicting the training value, we have a higher r2 rate which is around double the previous method which is the simple regression.

The thing that we can conclude here is: linear regression may not estimate the distribution of many data sets accurately. In other words, if the distribution of our data set is not around a line, we cannot get an accurate result from linear regression for this type of data set. So, we need more coefficient for predicting the behavior of this data set. That’s why we can see the improvement in r2 by adding one extra input and one coefficient in the regression equation respectively.

To see if this phenomenon always happens, we should pay attention to the distribution of our data set. For example, if the distribution of data set is around a line or we can estimate a line that can fit our data, we can use a simple regression model. As a result, we can expect an acceptable r2 value. But in most data sets, the value of the target variable depends on many inputs. So, the distribution of these data sets is hard for predicting just by one input. For example, here, we can’t say the rating just depends on the amount of sugar. It depends on many other factors as well. So, adding another input variable can increase the value of r2. At the same time, we should pay attention to the fact that by adding too many variables we may experience overfitting in our regression model. Lastly, we should make a tradeoff between the amount of our r-square value and the amount of overfitting of the regression model to find out the best result.