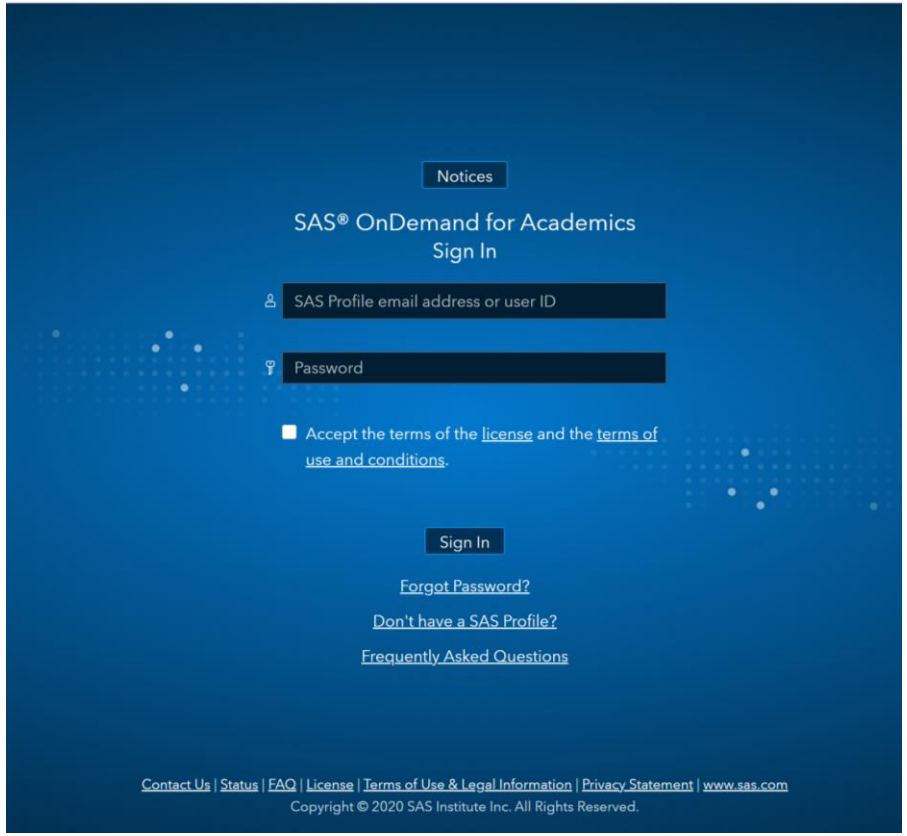
ANLT5050

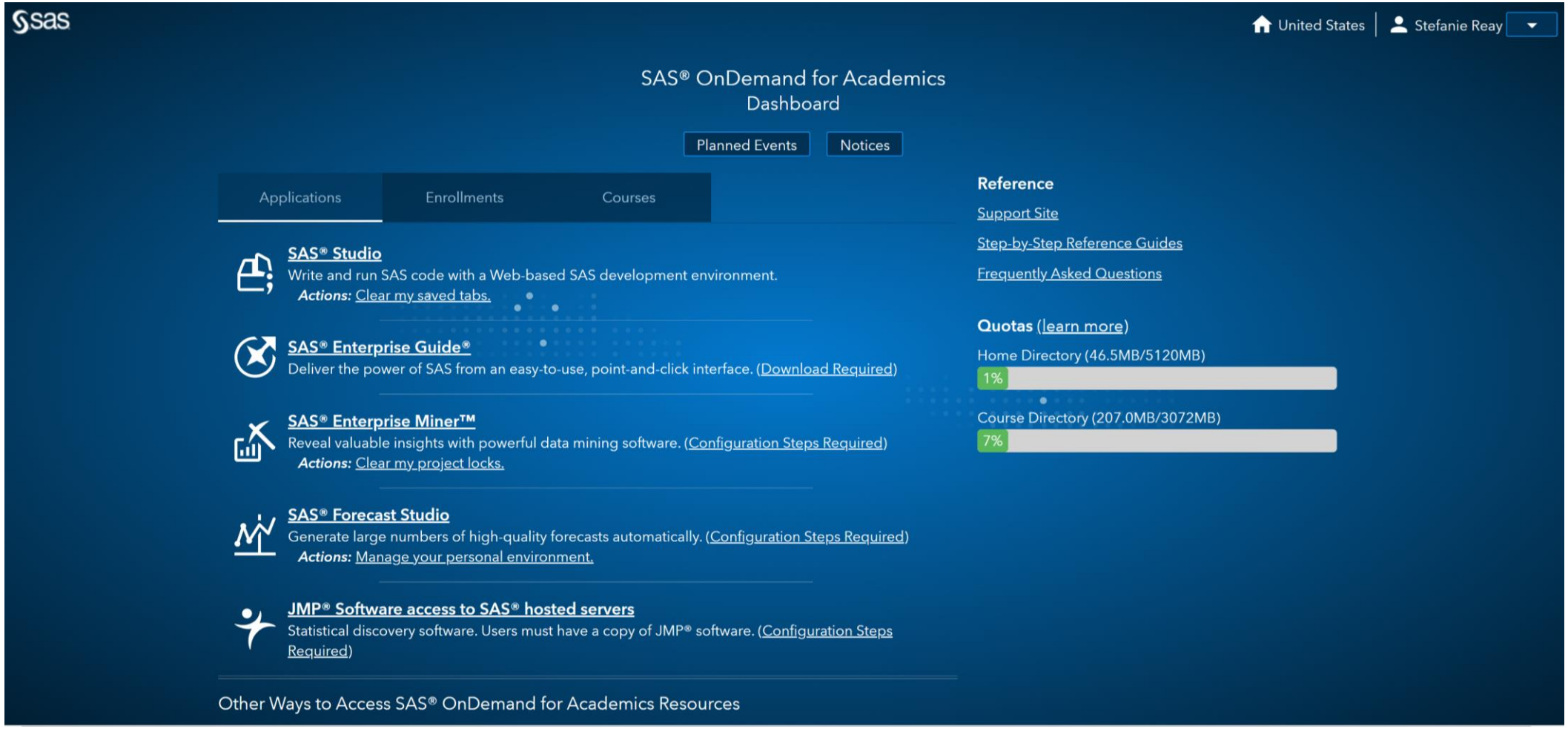
Unit 6 Assignment 1 Tutorial

Access the SAS OnDemand for Academics Control Center

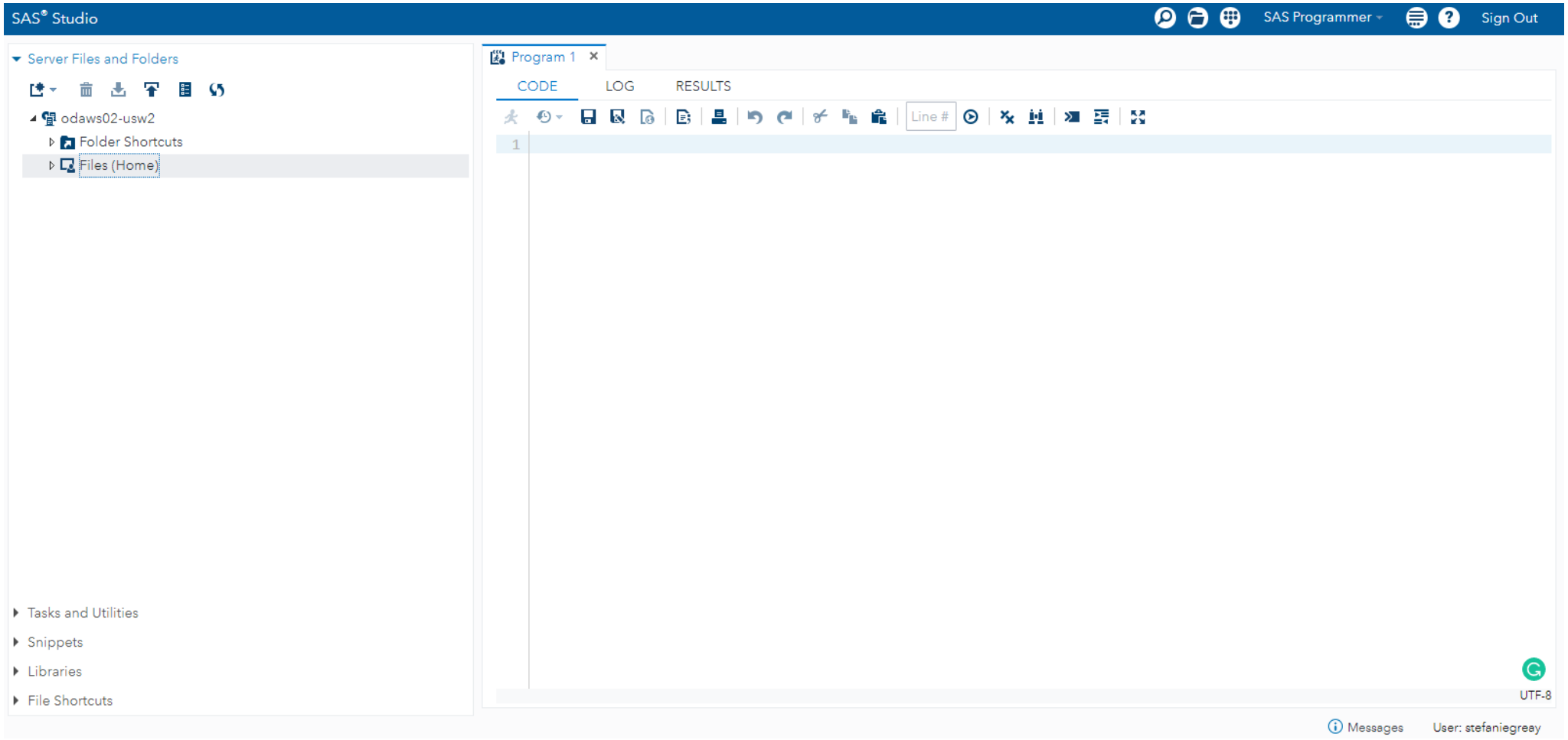
https://odamid.oda.sas.com/SASODAControlCenter



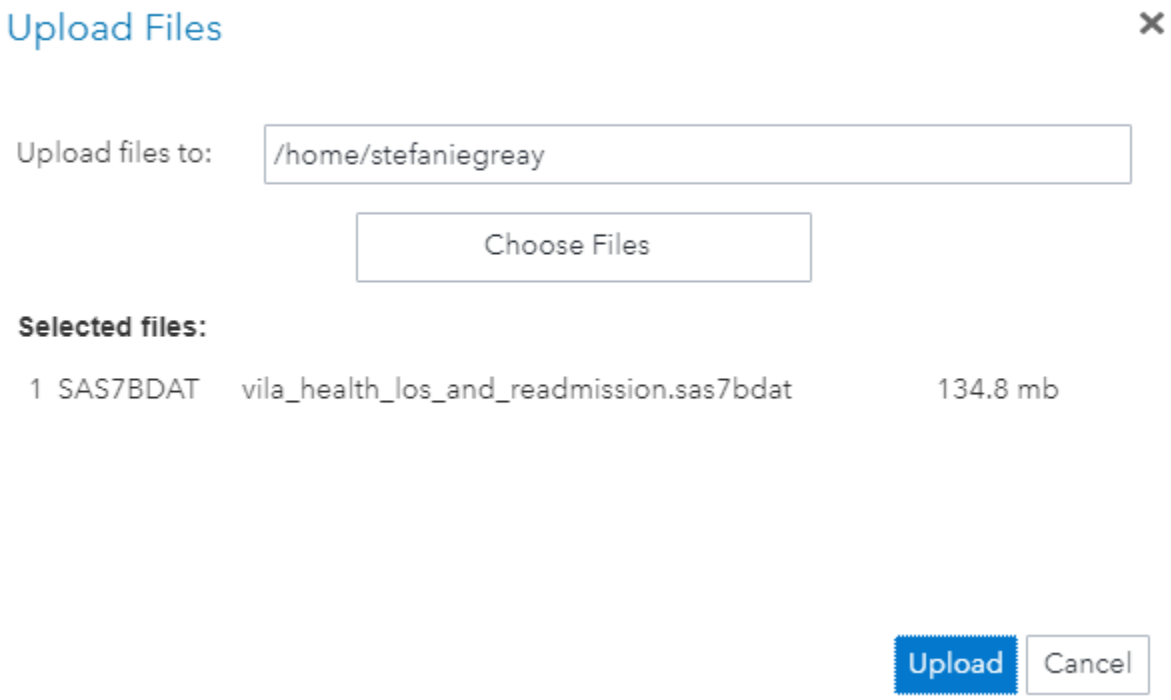
SAS OnDemand for Academics (SODA) Control Center



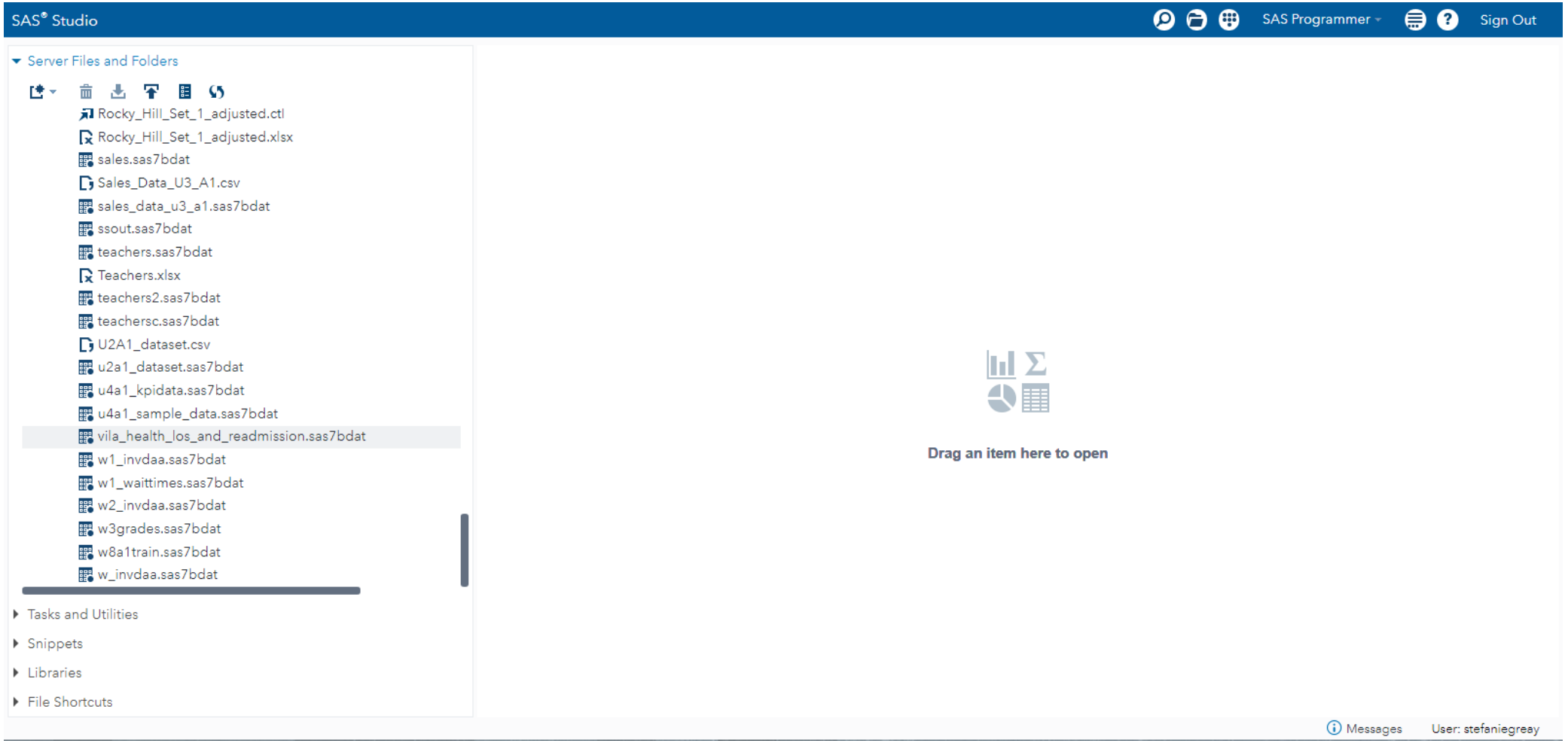
To upload the dataset to the SAS server, open SAS Studio, then click on “Files (Home)” and click the upload button.



Click on “Choose Files” to browse to the file you want to upload, then click “Upload.”



Verify that the upload was successful by scrolling down in your Files(Home) area.

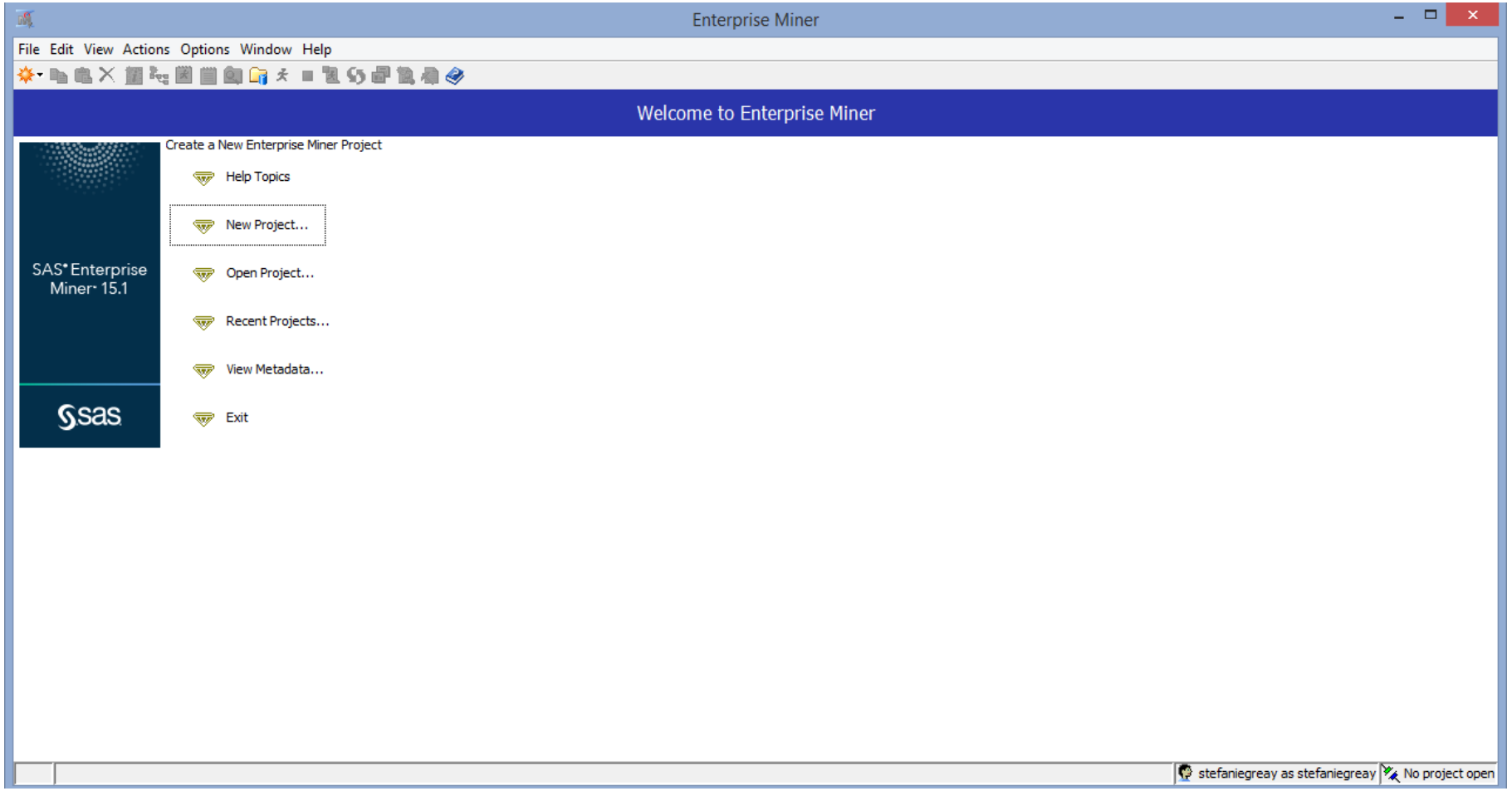


SAS Enterprise Miner Instructions

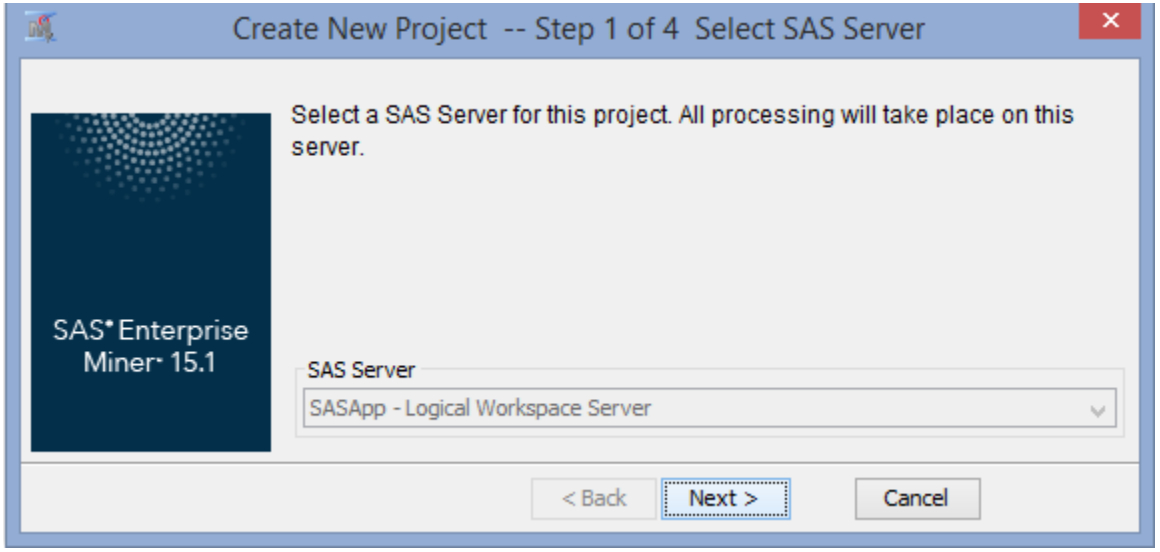
The following slides provide instructions on how to complete this task in SAS Enterprise Miner.

Once you have uploaded the dataset for this unit onto the SAS servers using SAS Studio, you may proceed from here using SAS Enterprise Miner.

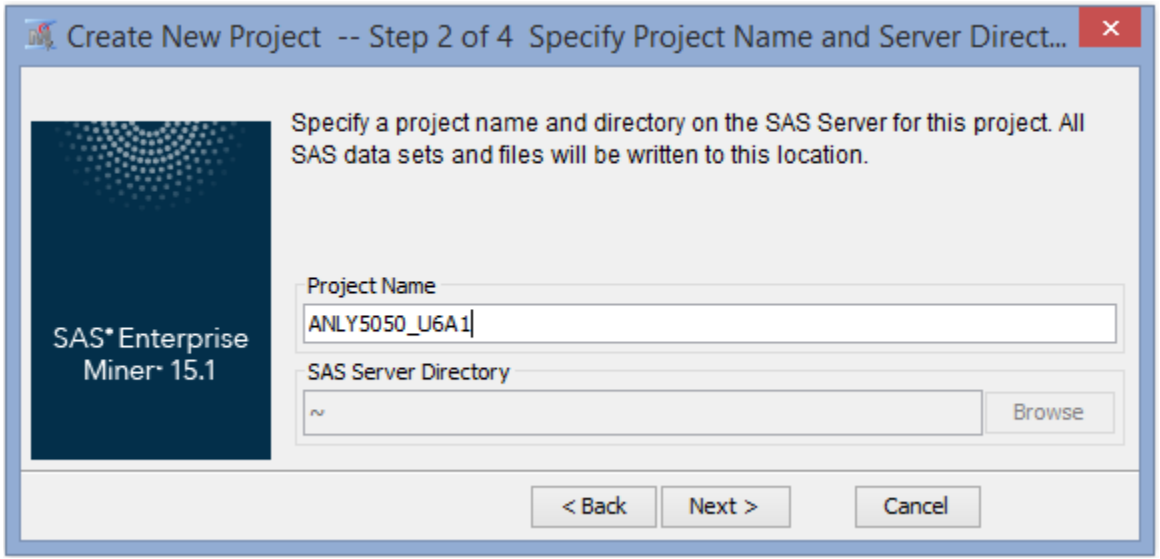
Once you download and start SAS Enterprise Miner, open a new project by clicking on “New Project.”



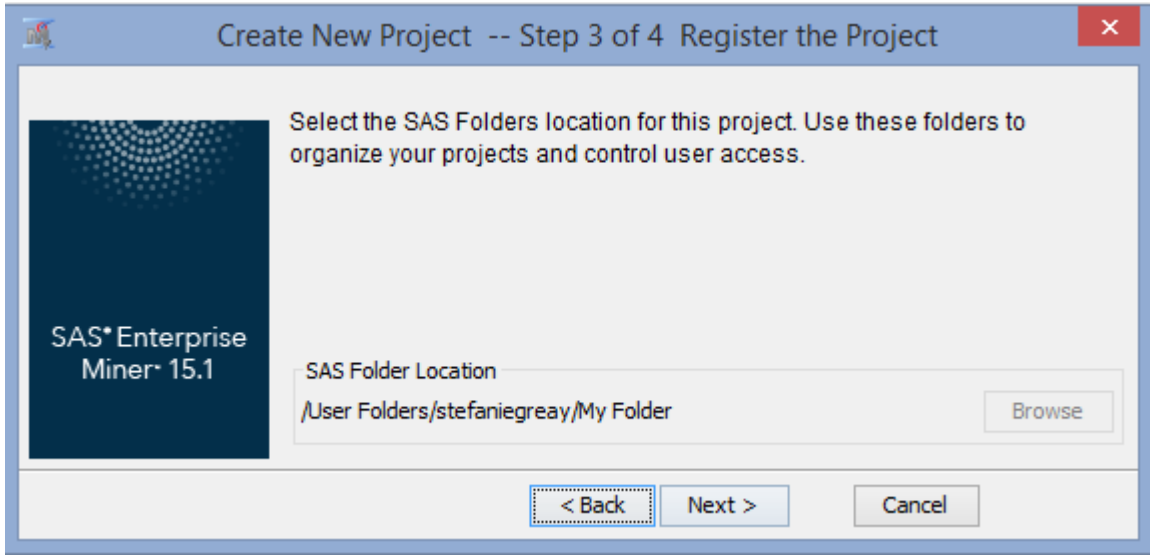
Click “Next>” to use the default SAS Server



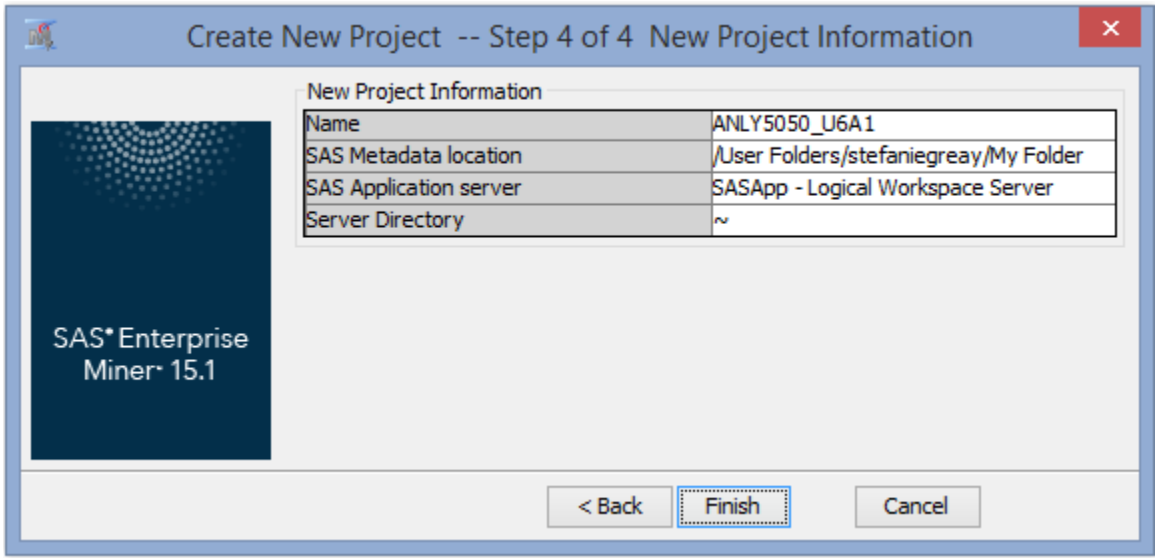
Enter a project name and click “Next>”



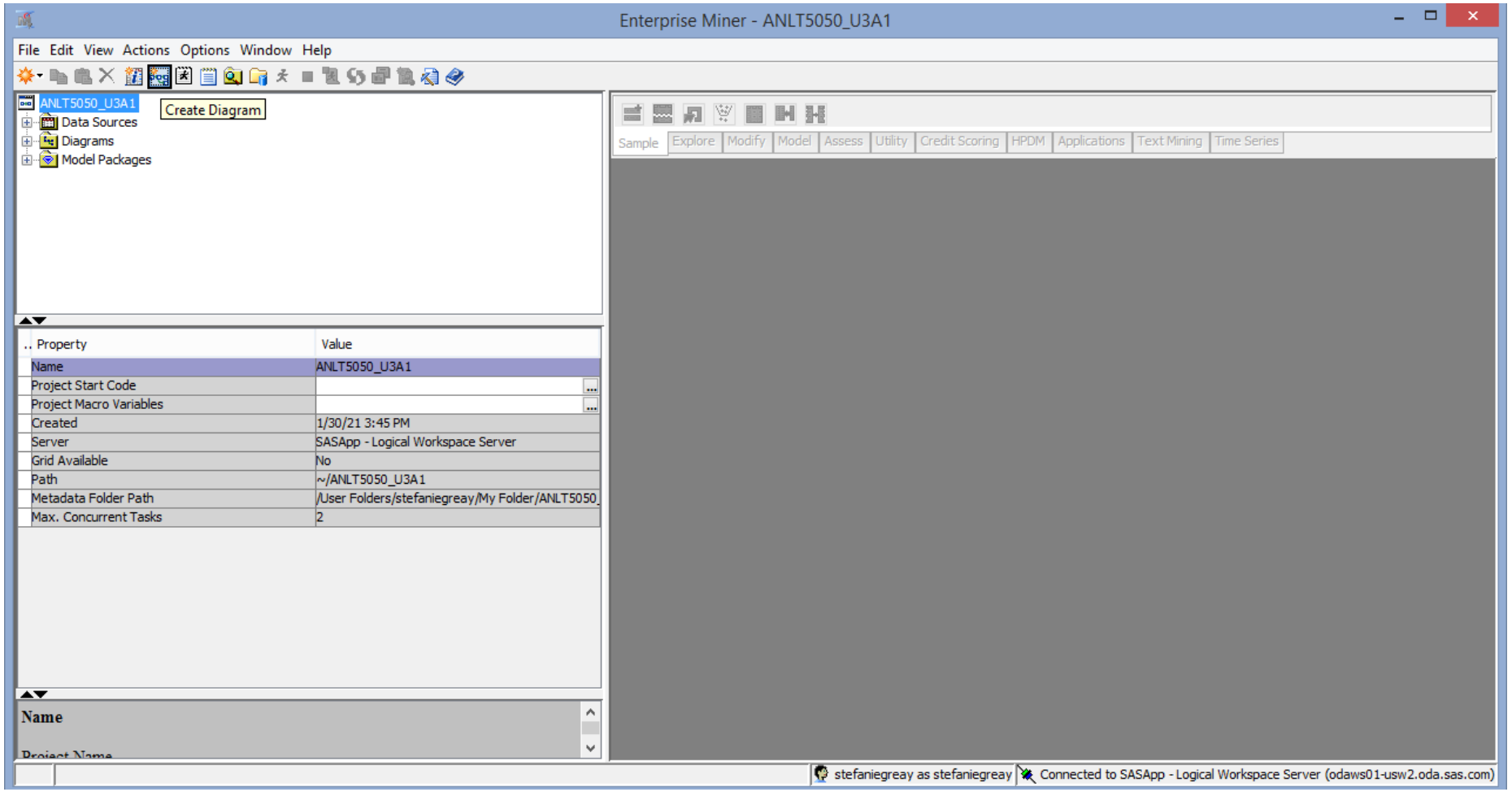
Click “Next>”



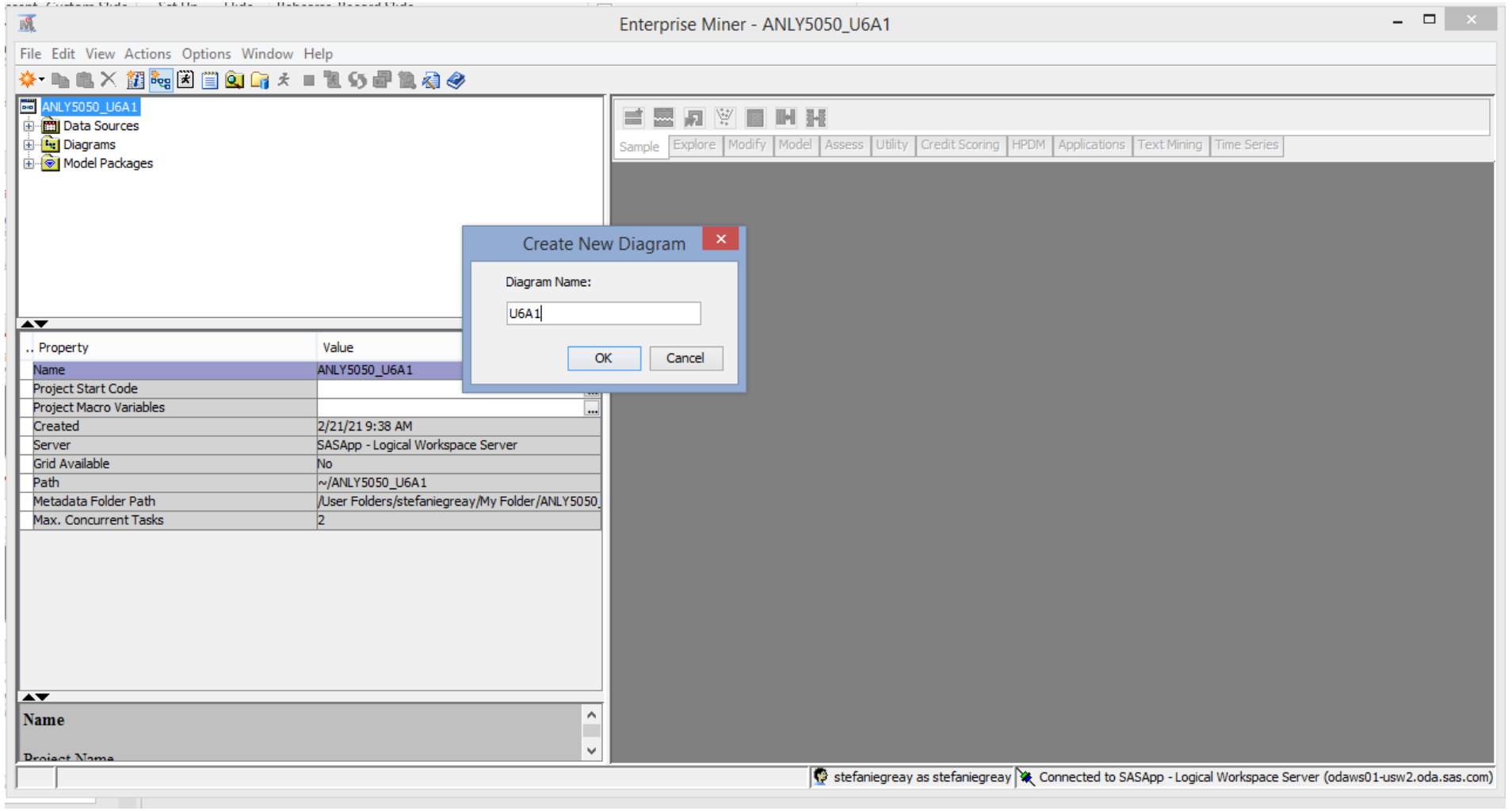
Verify your entries and click “Finish”



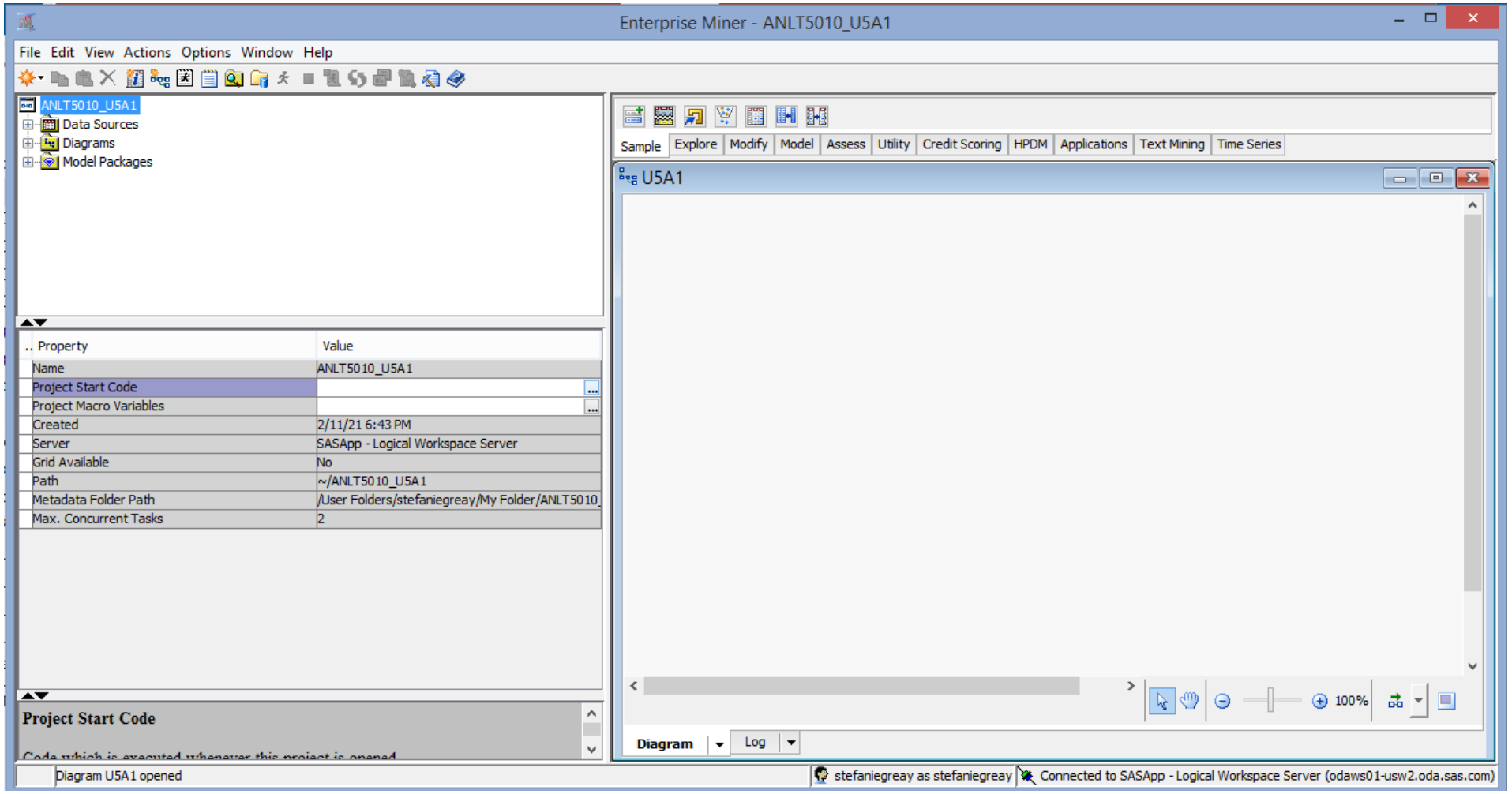
Click on the “Create Diagram” icon.



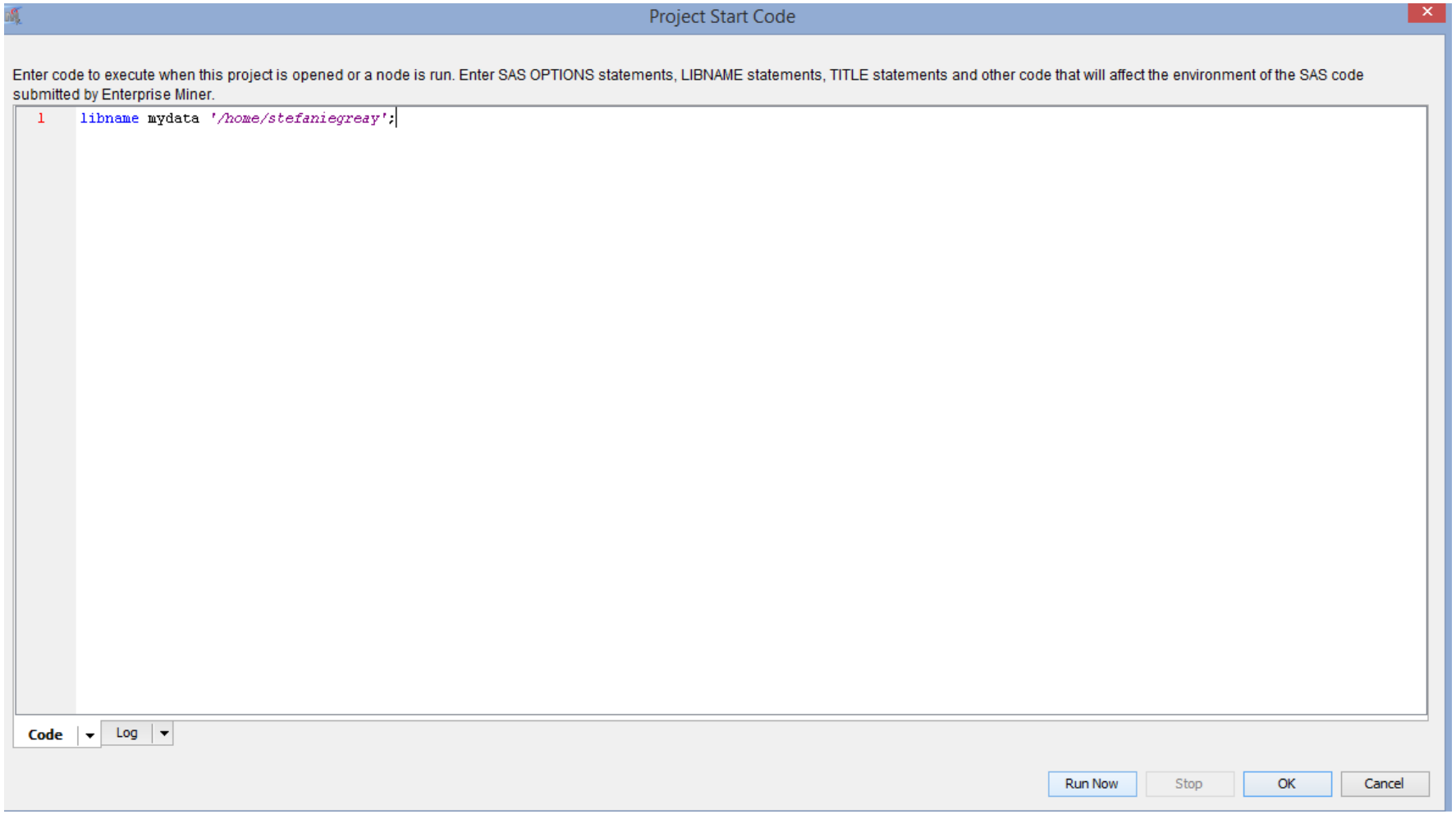
Name your diagram and click “OK.”



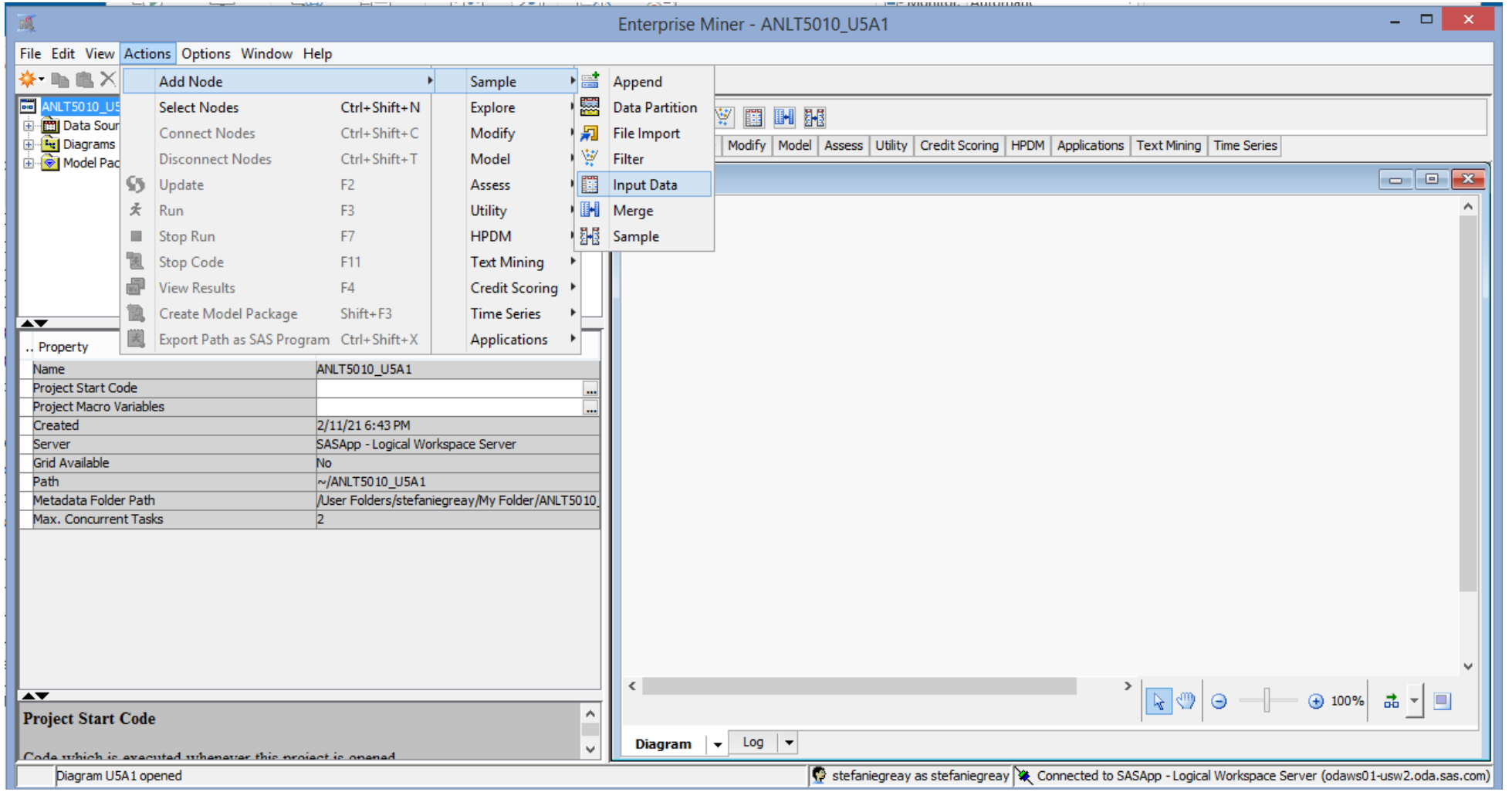
Click on the project, then click on the ellipses next to “Project Start Code.”



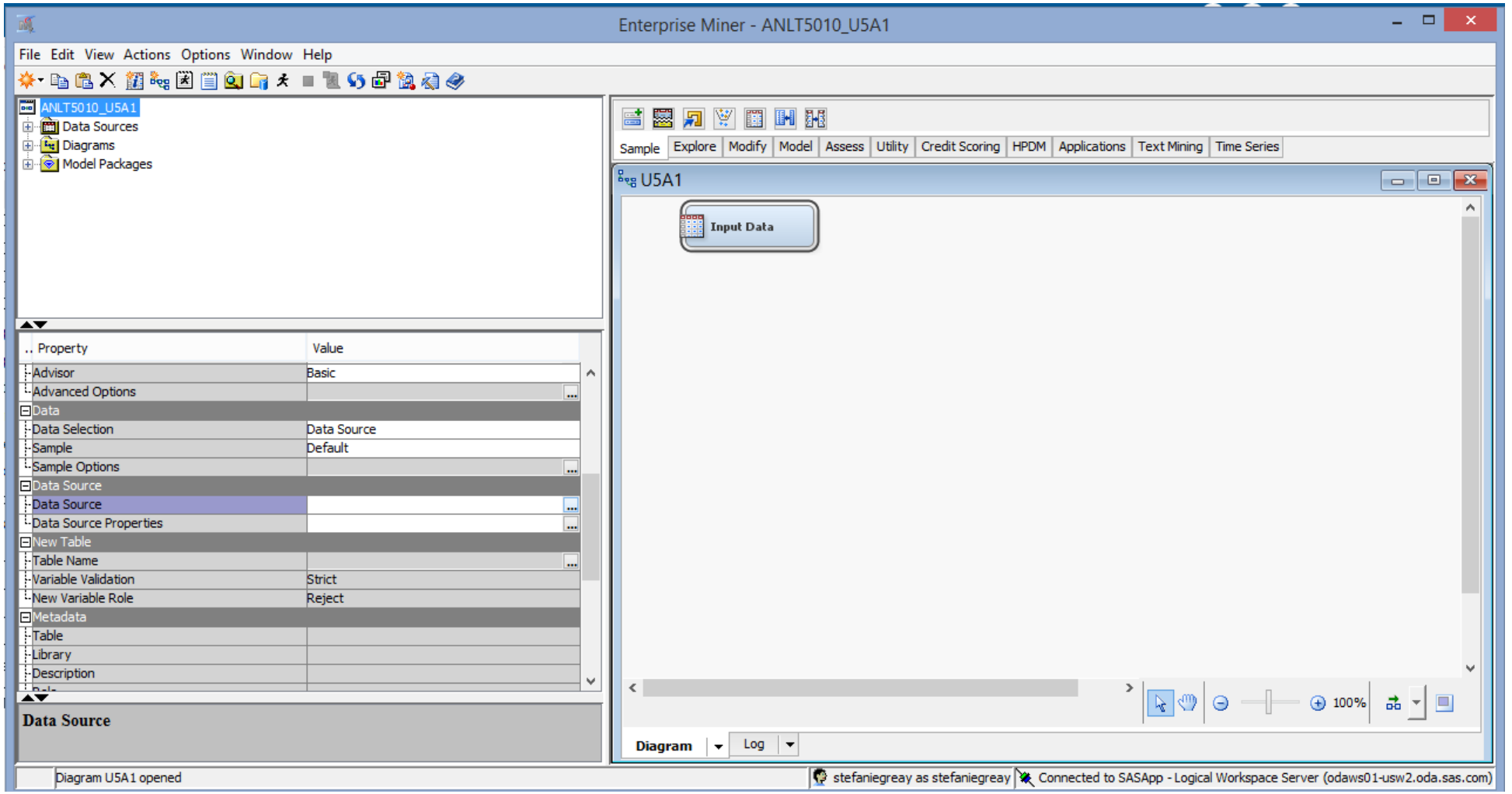
Add the library reference for where you uploaded the dataset in SAS studio, and click “Run Now.” Once it completes, click “OK.”



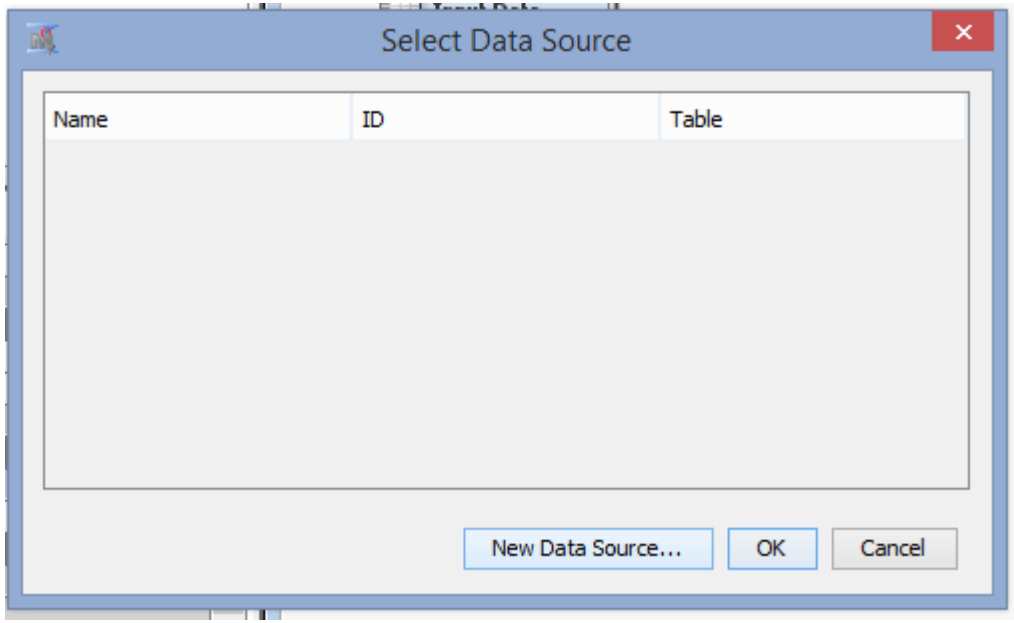
Click on Actions>Add Node>Sample>Input Data



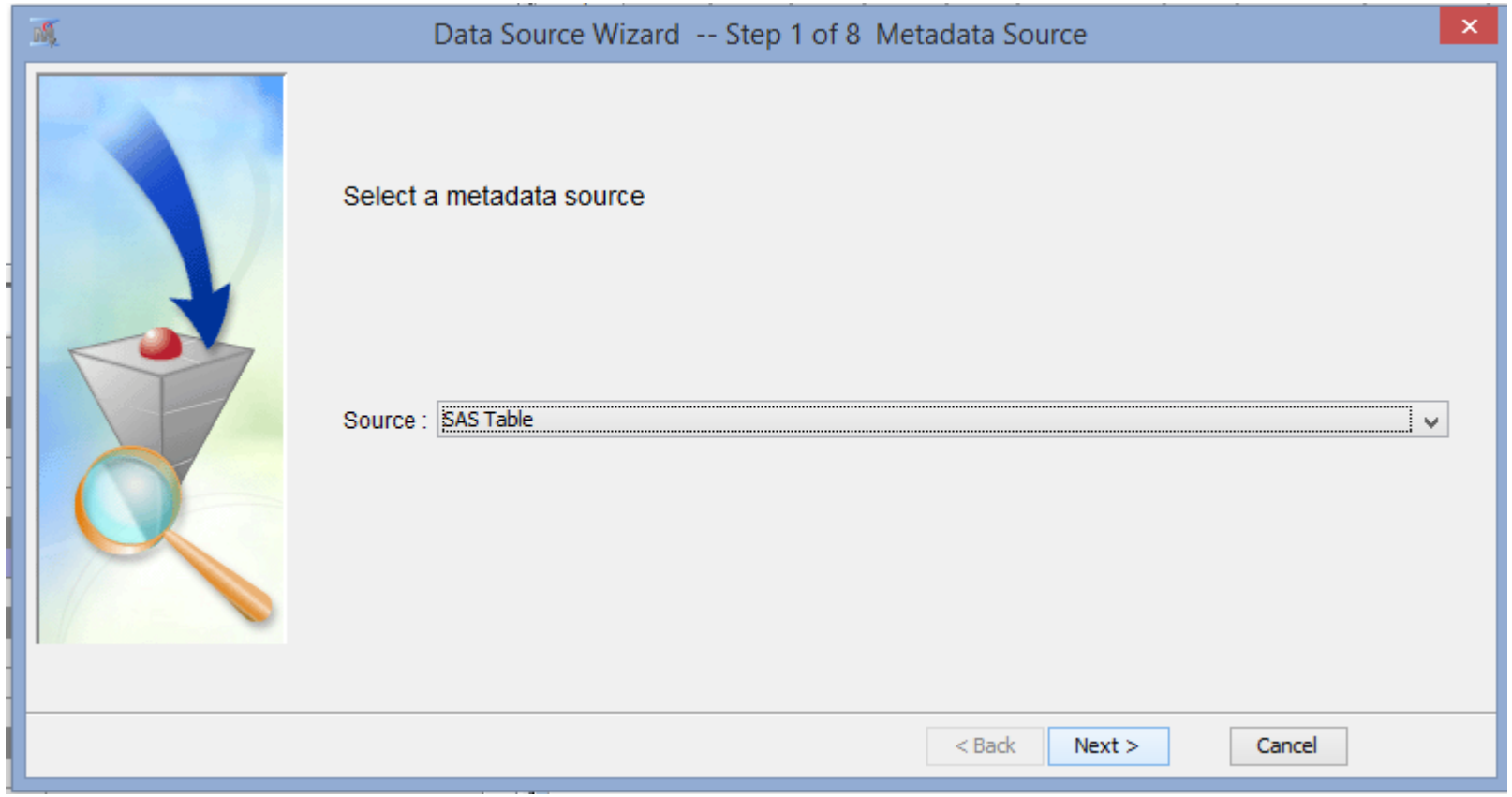
Click the ellipses (3 dots) next to “Data Source.”



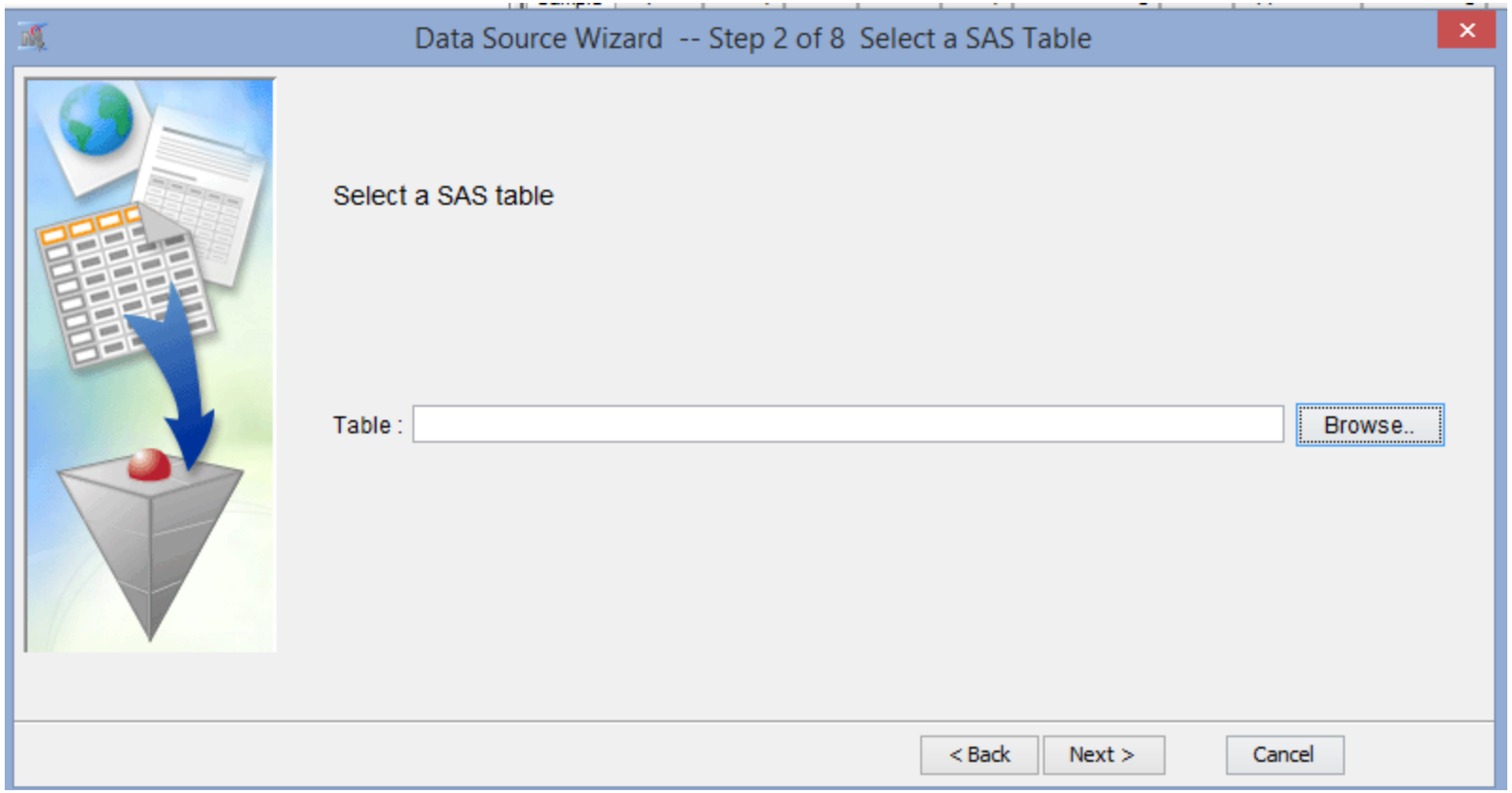
Click on “New Data Source”



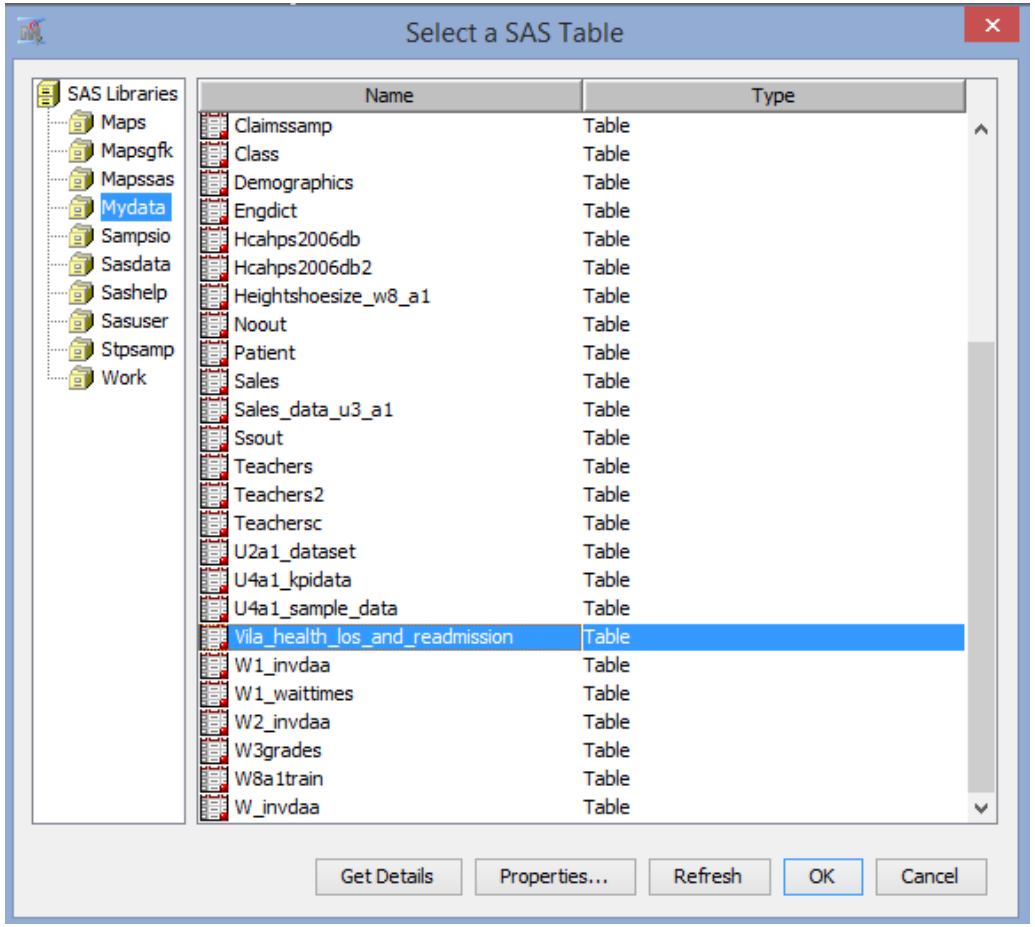
Leave it as “SAS Table” and click “Next >”



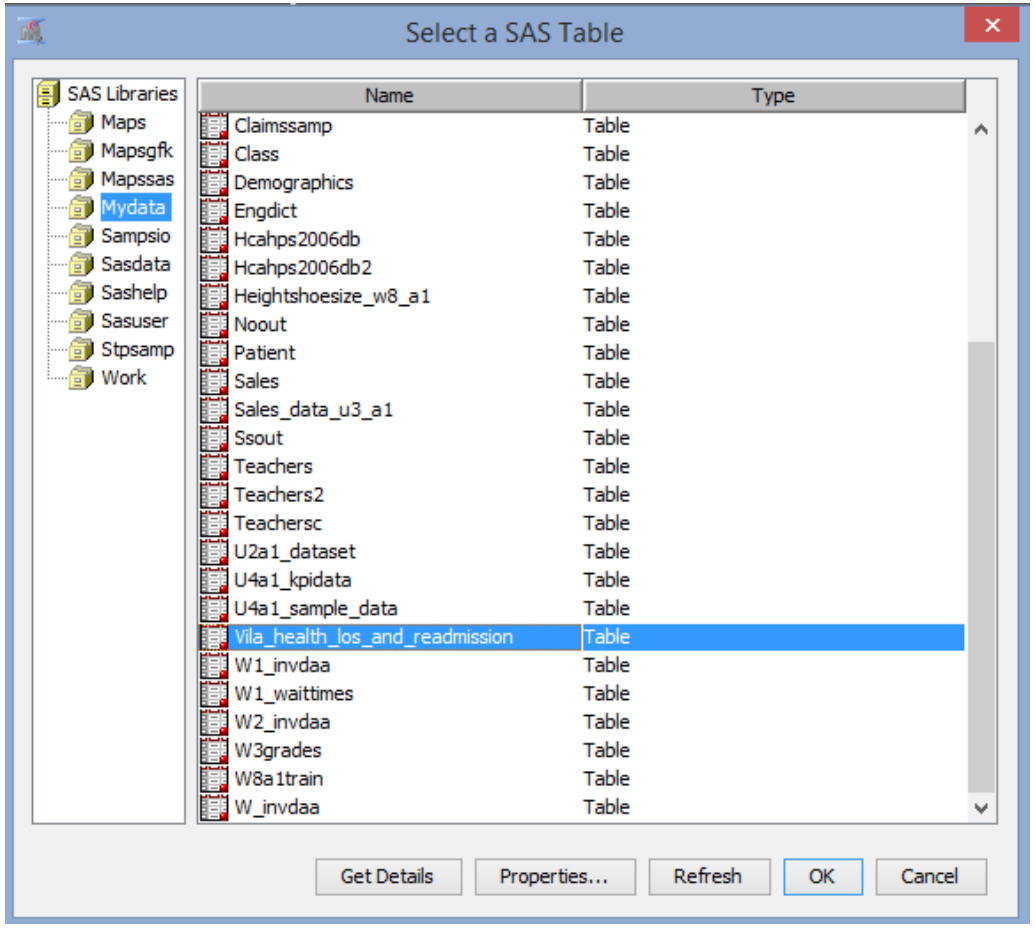
Click on “Browse”



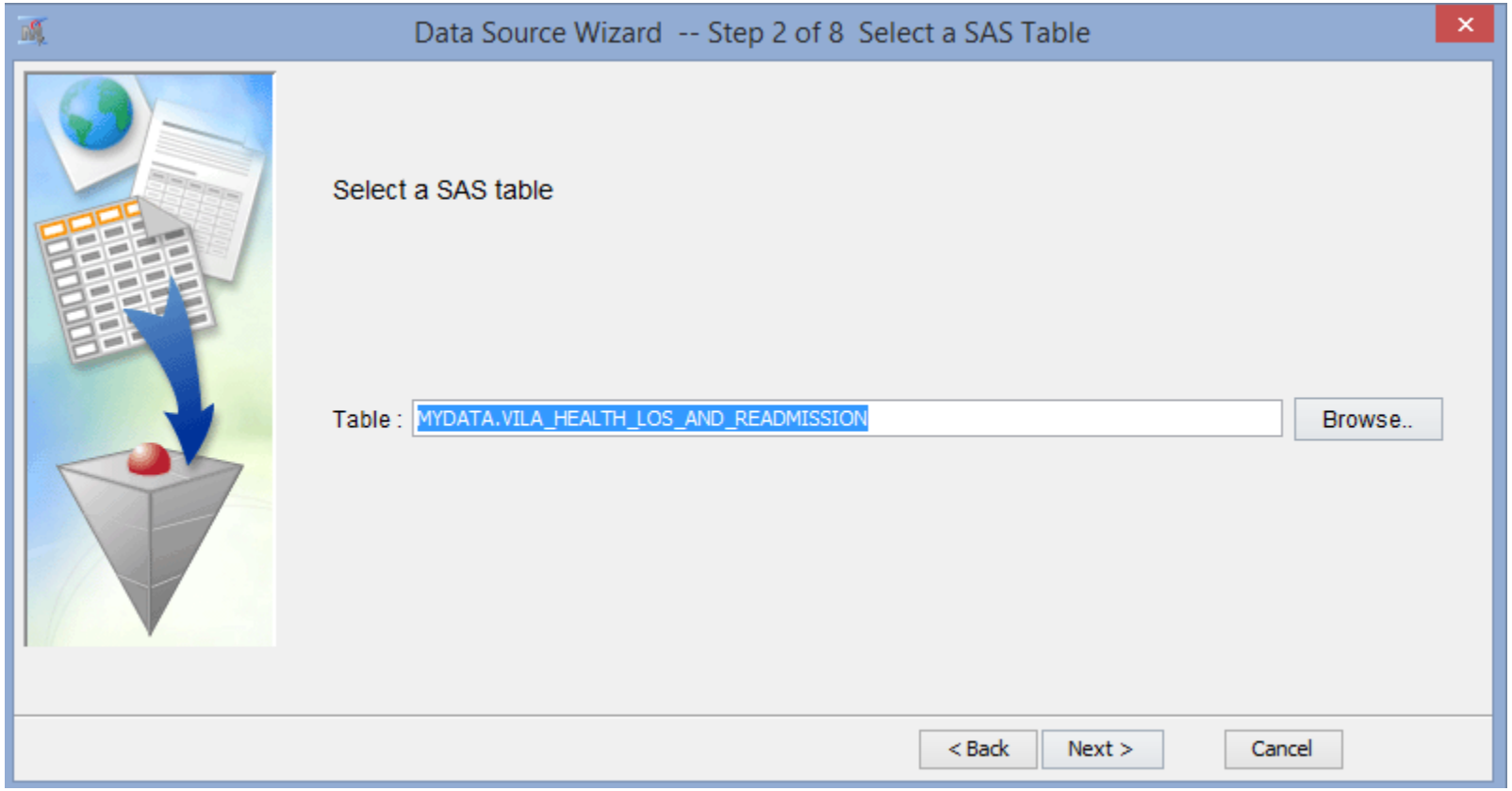
Double click on the libname you just set up in the project startup code.



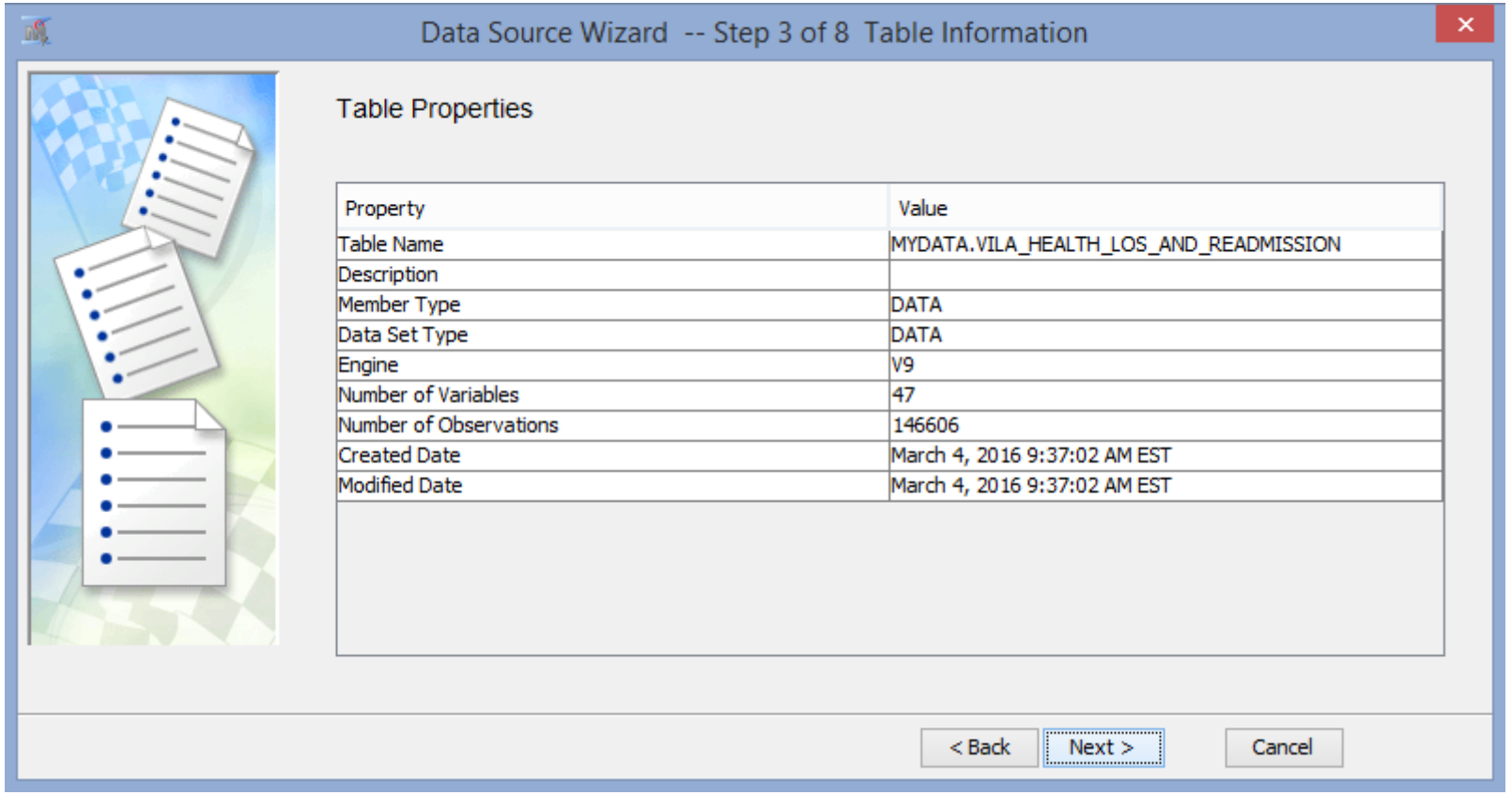
Double click to select the dataset for this unit, and click “OK”



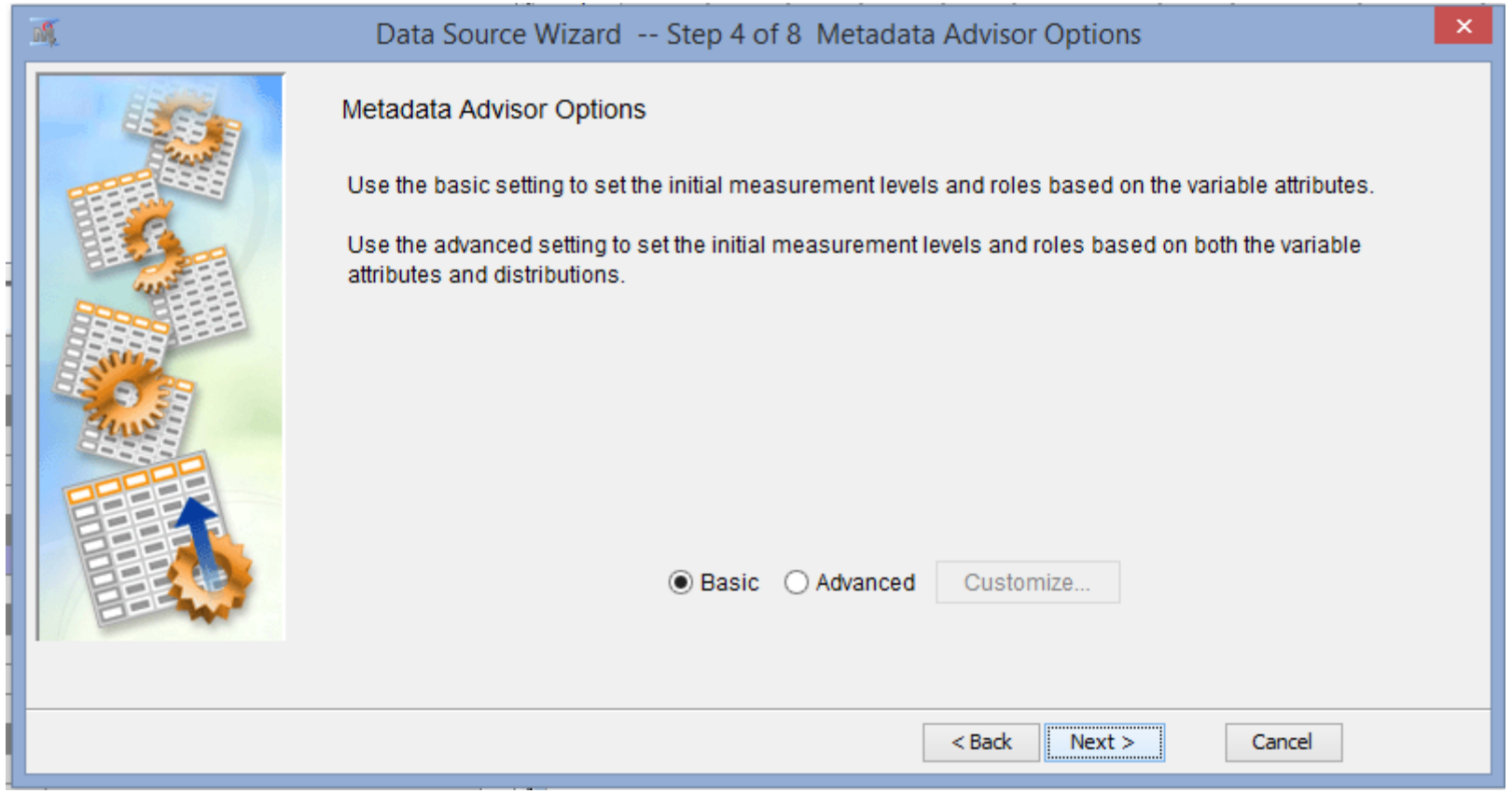
Click “Next>”



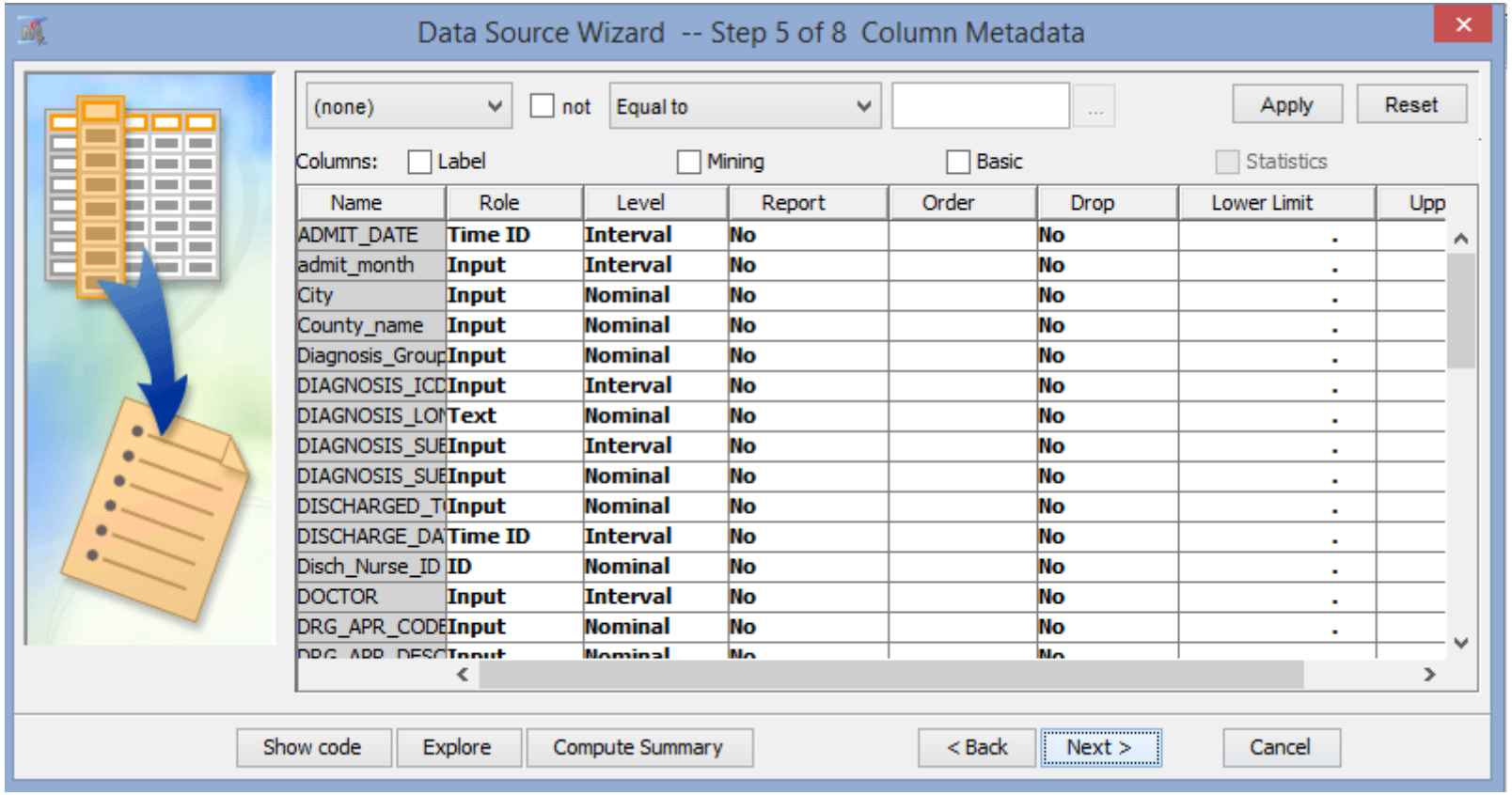
Verify the options and click “Next>”



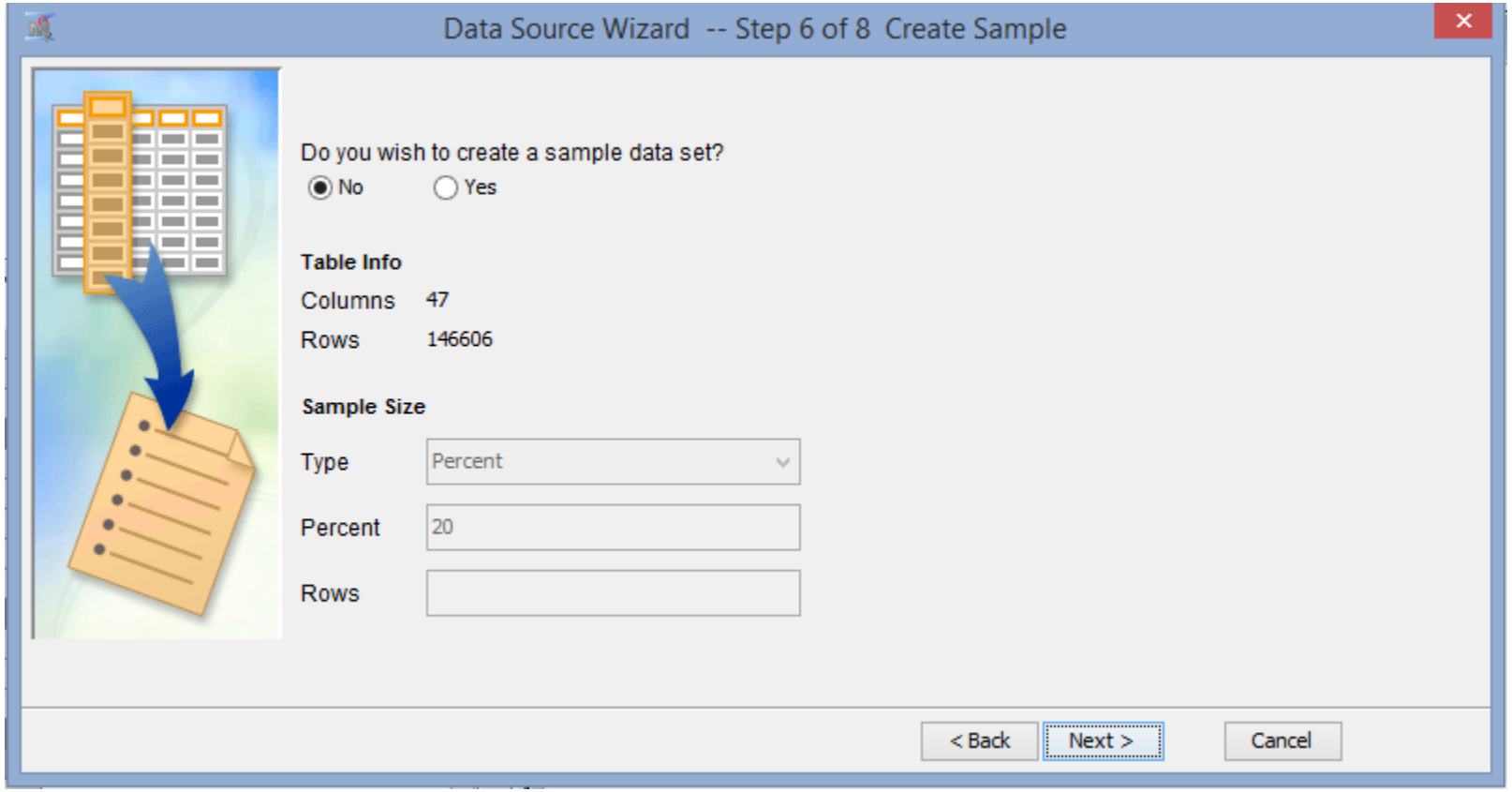
Click “Next>”



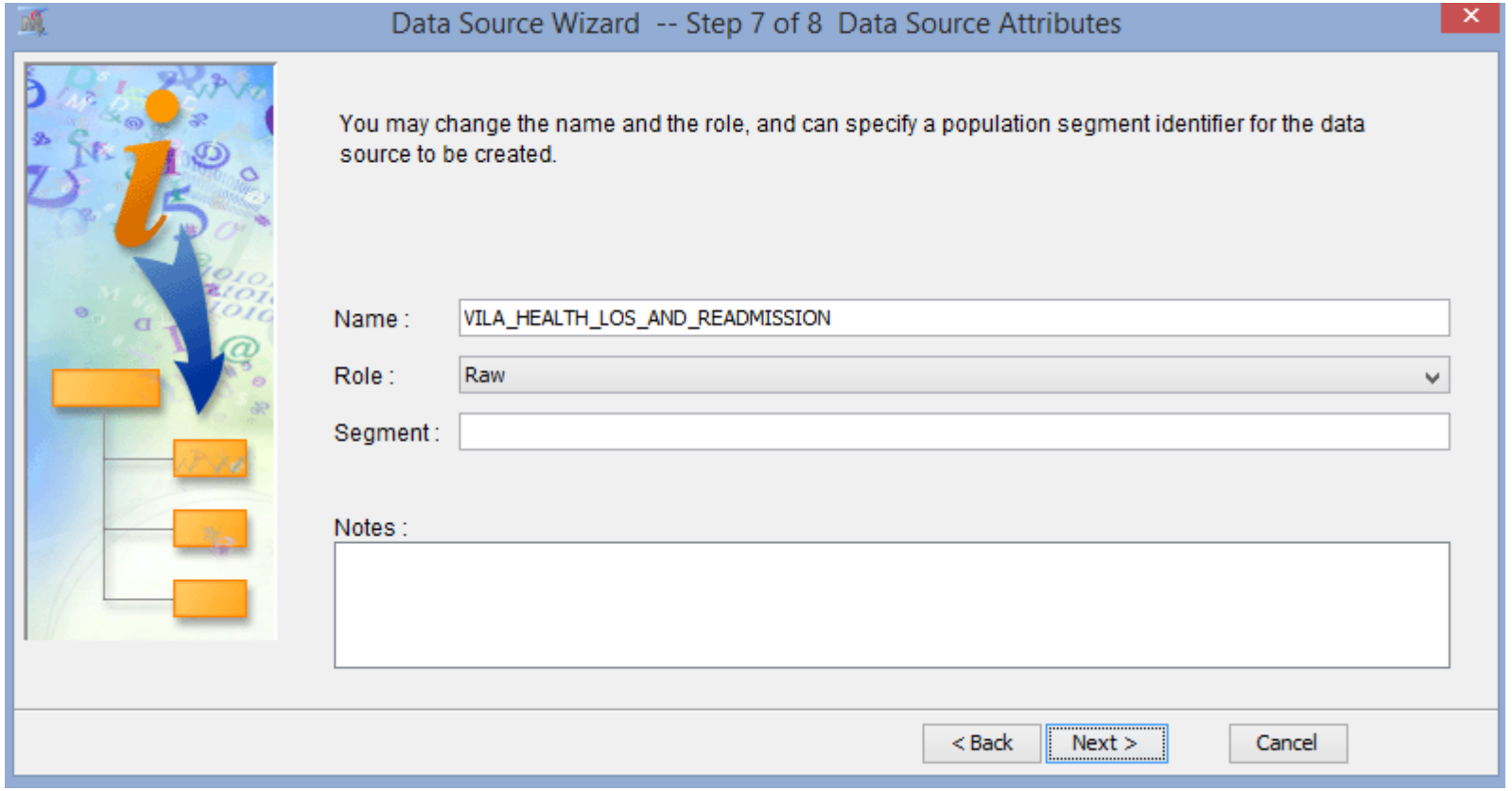
Verify the variables and settings, adjust if necessary, and then click “Next>”



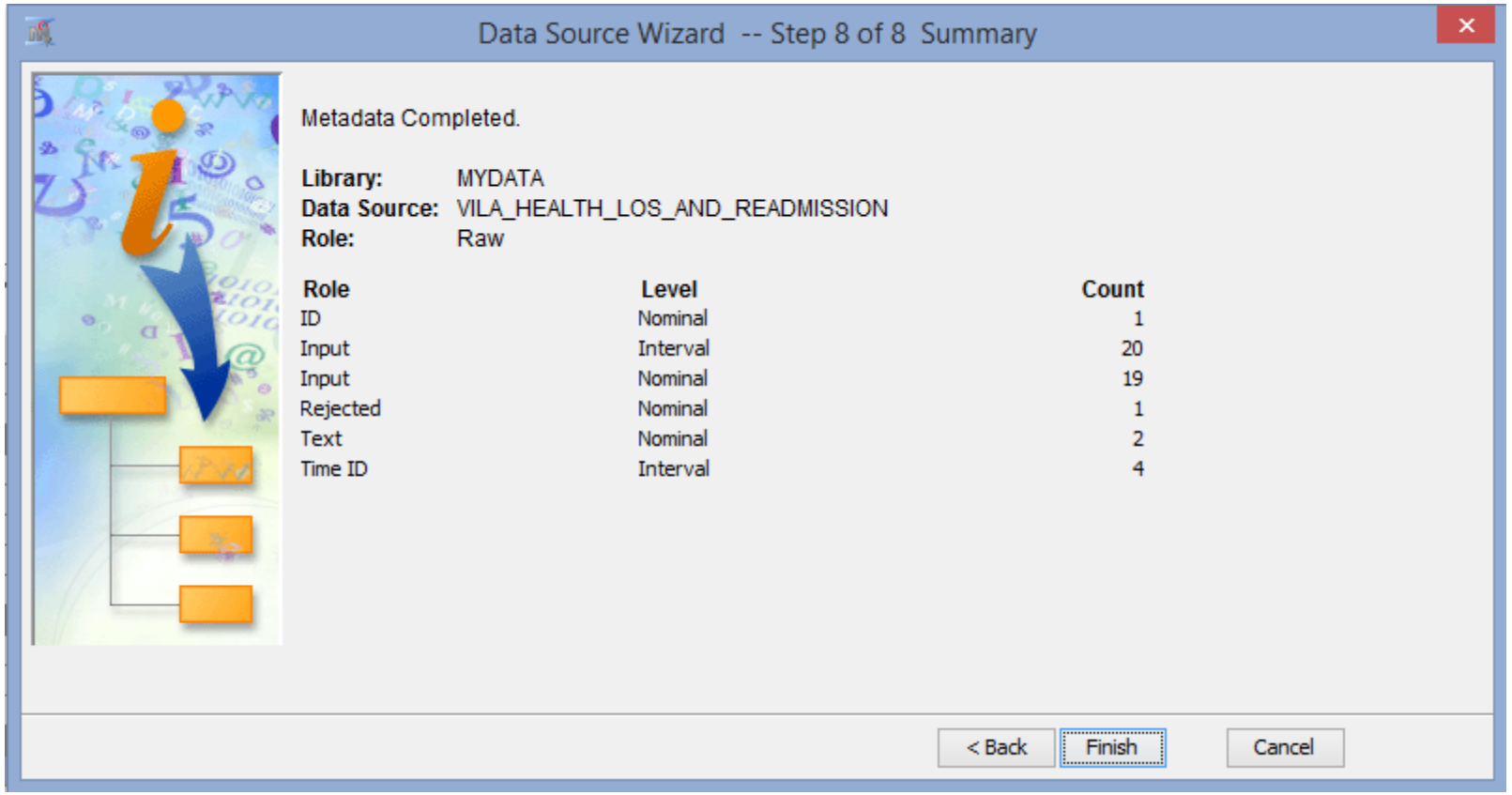
You may choose to sample the dataset here, or just keep the full dataset, then click “Next>”



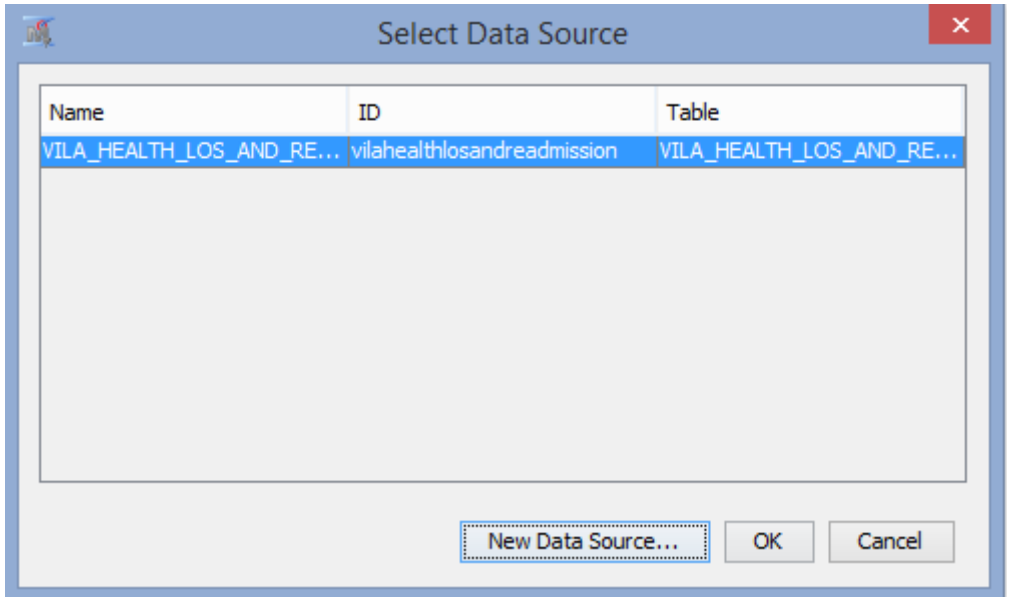
You may choose to adjust the role of the dataset, or leave it as the default, then click “Next>”



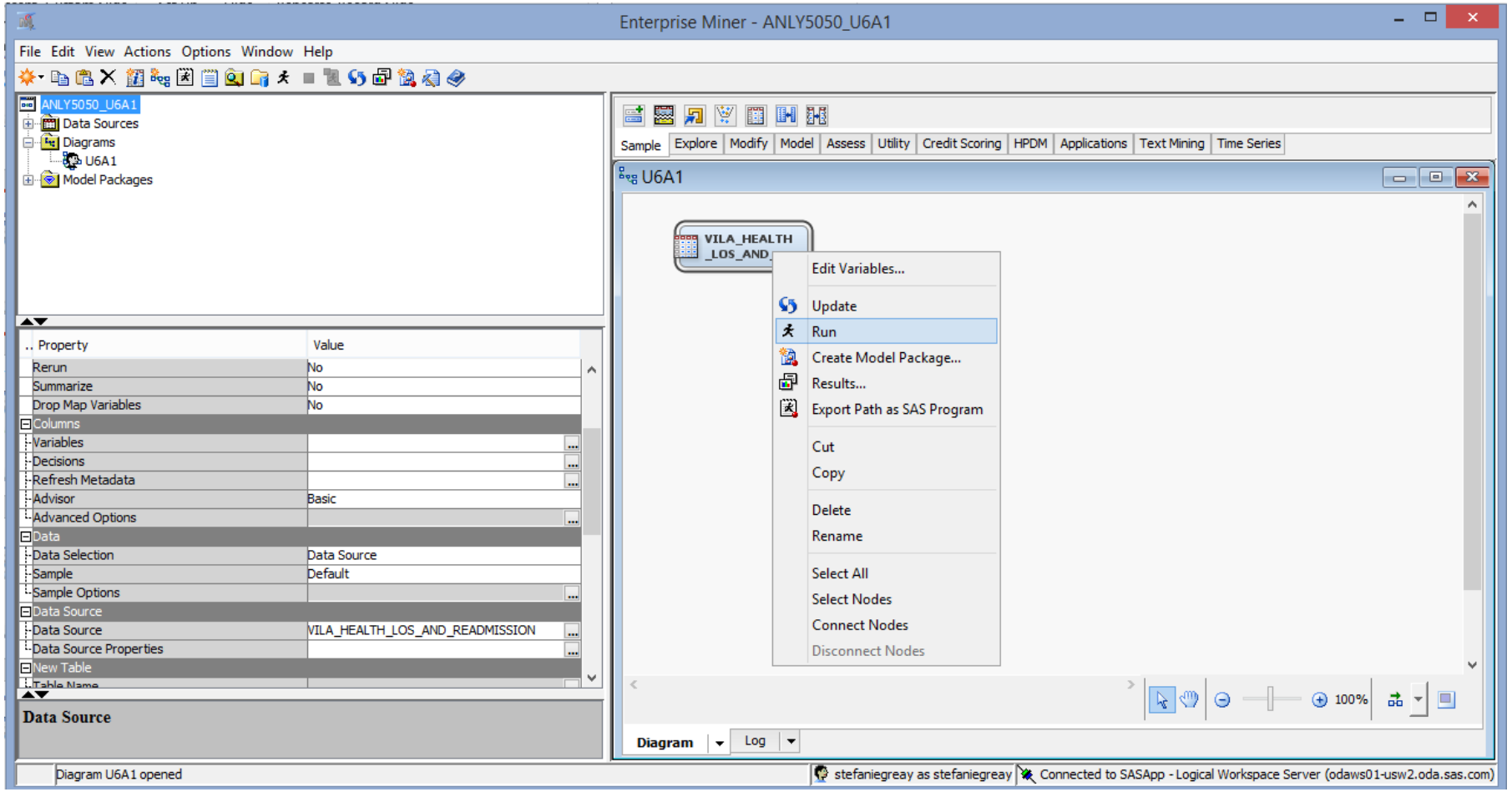
Click “Finish” to finish the data source registration within EM.



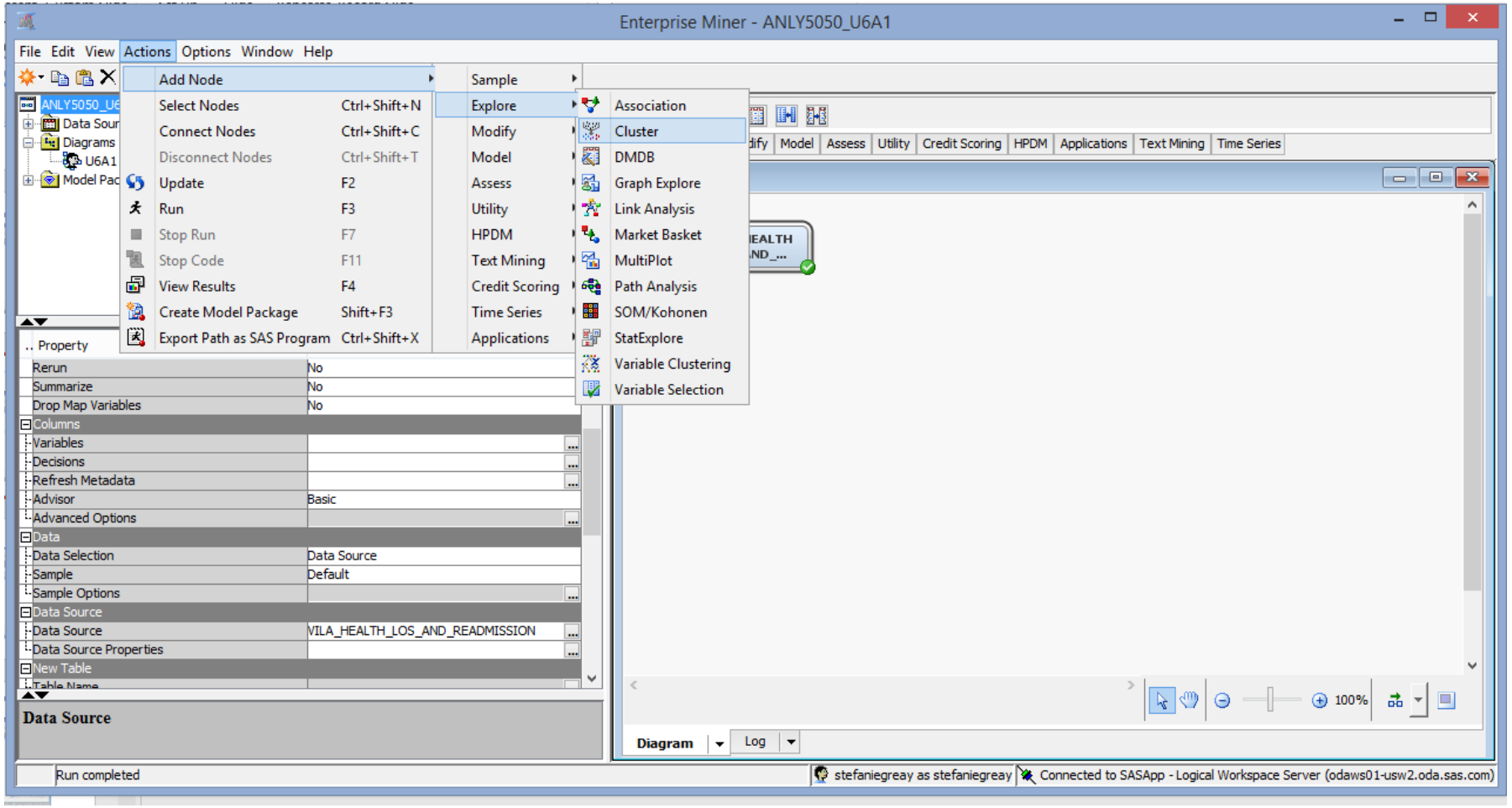
Click “OK” to complete the process. The name of the node should then change to the name of the dataset.



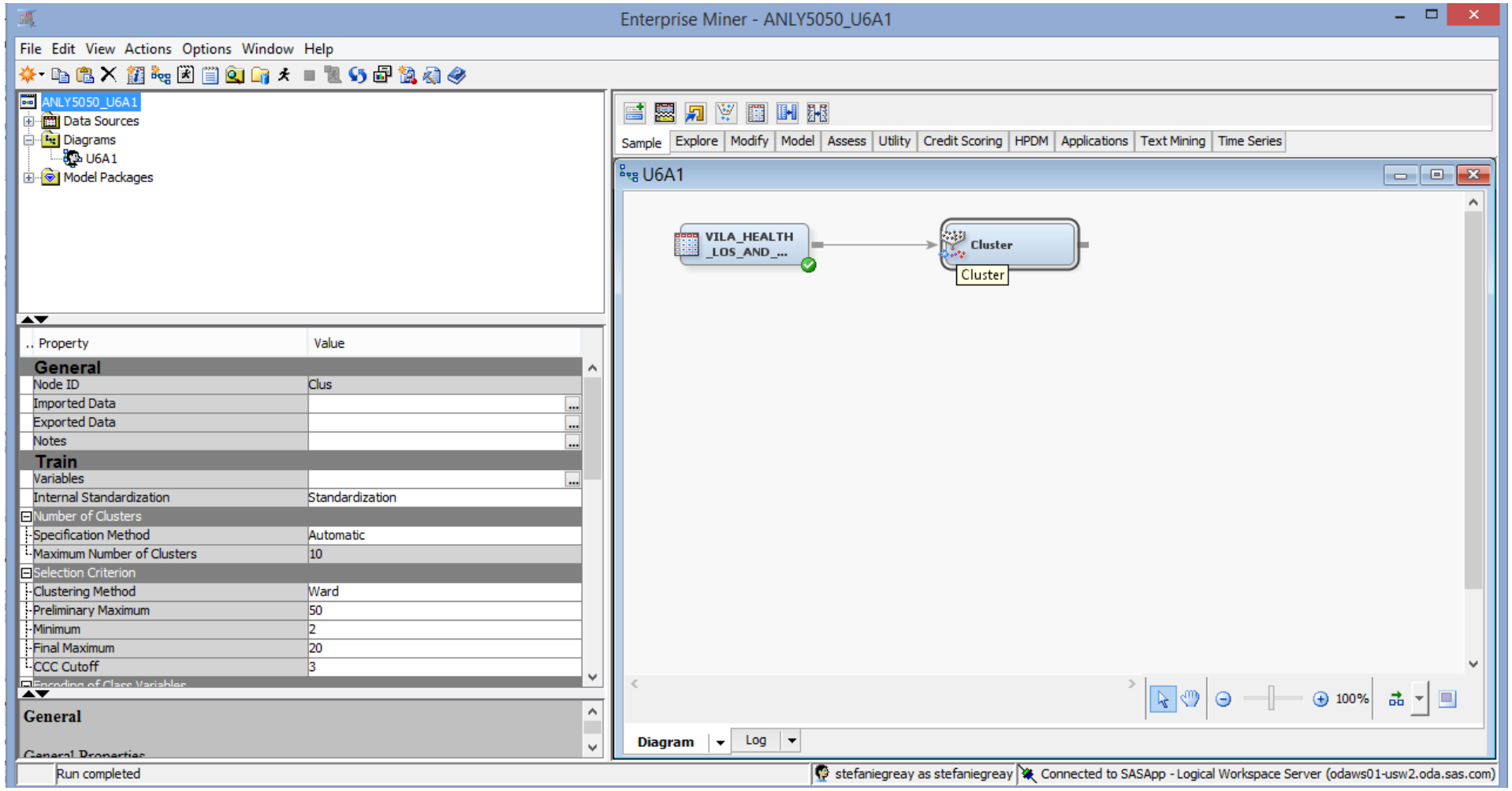
Right click on the dataset node and click “Run.”



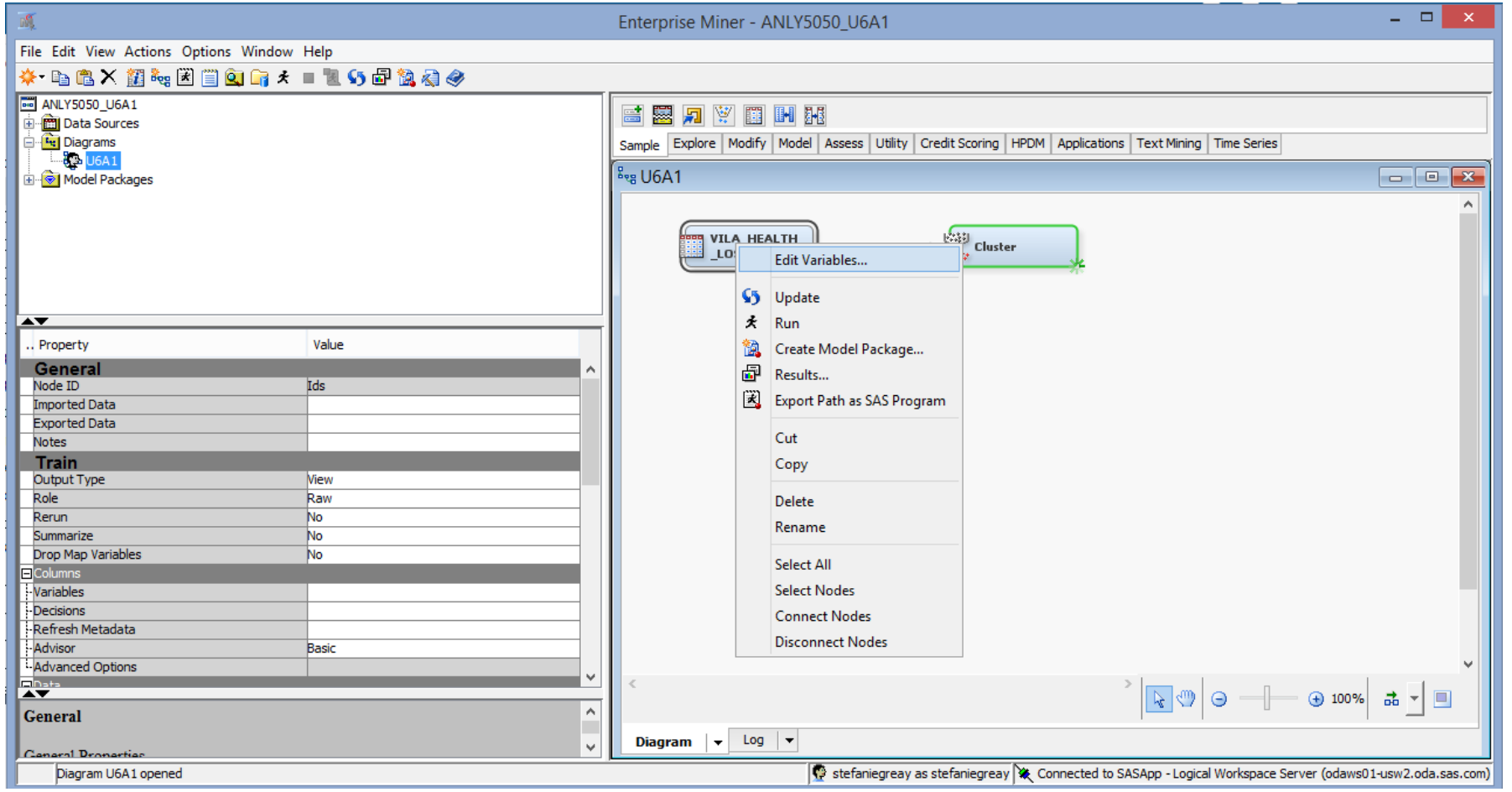
Click on “Actions” > “Add Node” > “Explore” > “Cluster”



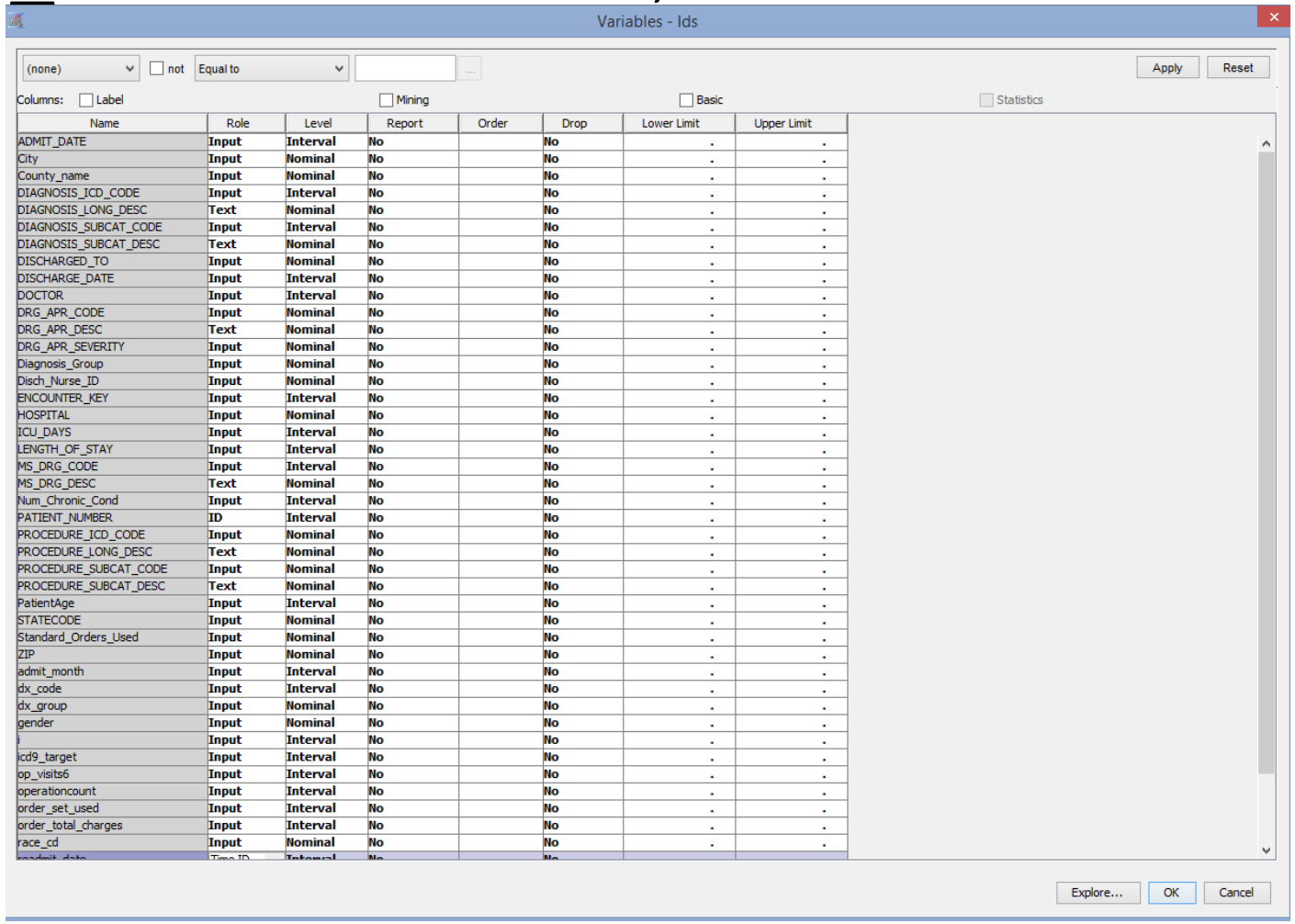
Connect the nodes



Right click on the dataset node and choose “edit variables.”

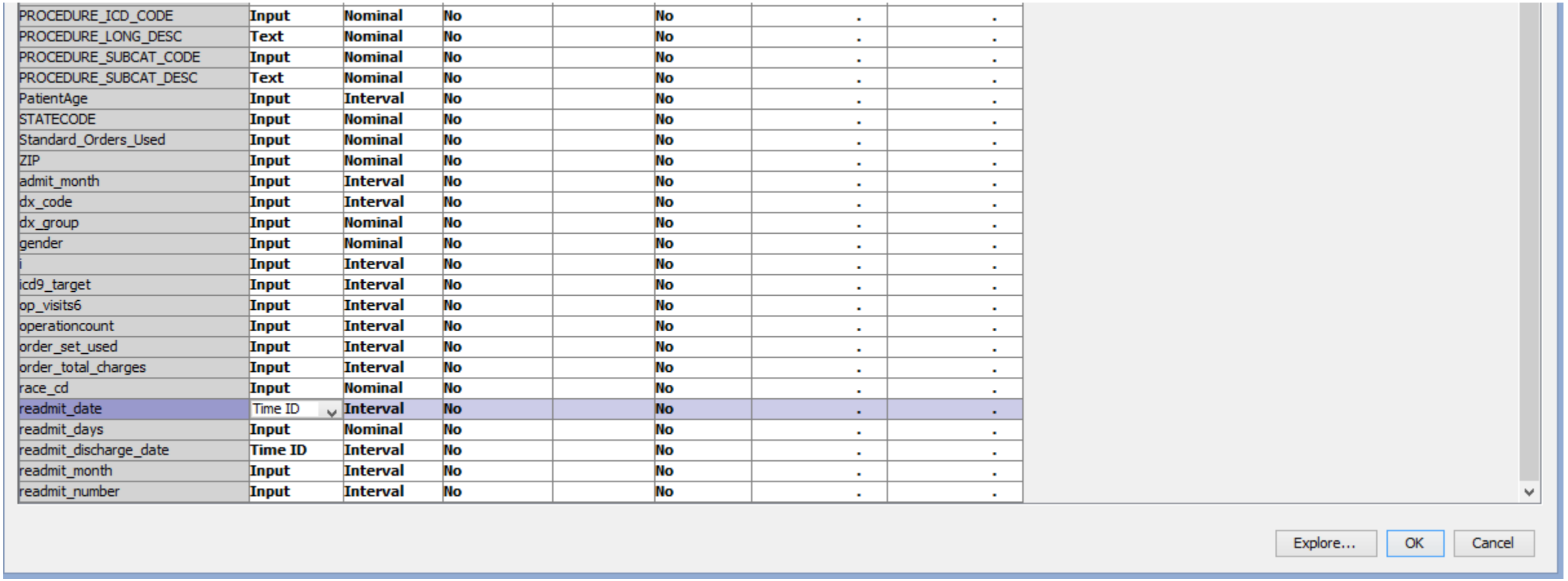


Change the description variables to “text,” the Patient_ID variable to “ID,” and all others to “Input”



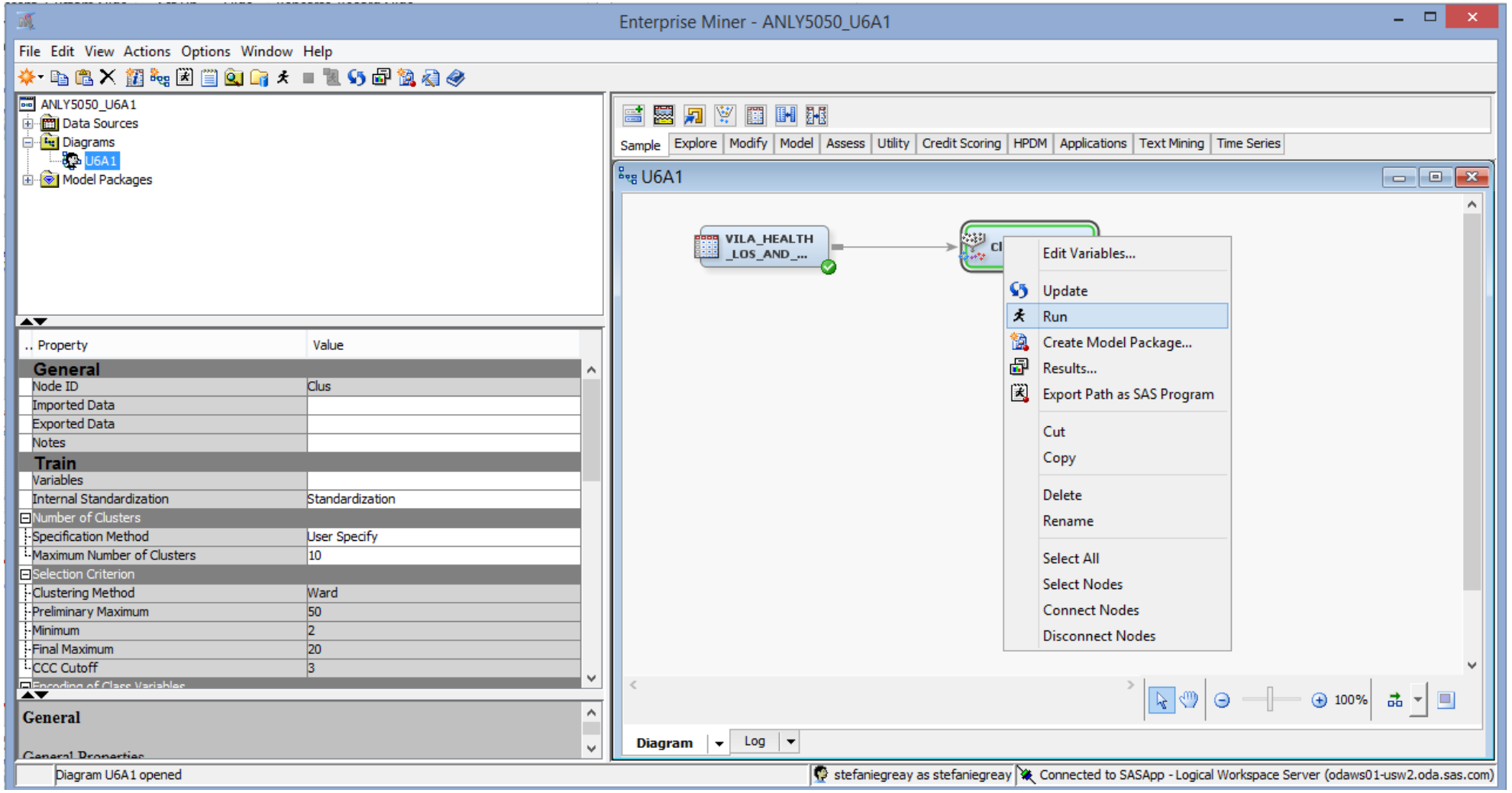
Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
ADMIT_DATE	Input	Interval	No		No	.	.
City	Input	Nominal	No		No	.	.
County_name	Input	Nominal	No		No	.	.
DIAGNOSIS_ICD_CODE	Input	Interval	No		No	.	.
DIAGNOSIS_LONG_DESC	Text	Nominal	No		No	.	.
DIAGNOSIS_SUBCAT_CODE	Input	Interval	No		No	.	.
DIAGNOSIS_SUBCAT_DESC	Text	Nominal	No		No	.	.
DISCHARGED_TO	Input	Nominal	No		No	.	.
DISCHARGE_DATE	Input	Interval	No		No	.	.
DOCTOR	Input	Interval	No		No	.	.
DRG_APR_CODE	Input	Nominal	No		No	.	.
DRG_APR_DESC	Text	Nominal	No		No	.	.
DRG_APR_SEVERITY	Input	Nominal	No		No	.	.
Diagnosis_Group	Input	Nominal	No		No	.	.
Disch_Nurse_ID	Input	Nominal	No		No	.	.
ENCOUNTER_KEY	Input	Interval	No		No	.	.
HOSPITAL	Input	Nominal	No		No	.	.
ICU_DAYS	Input	Interval	No		No	.	.
LENGTH_OF_STAY	Input	Interval	No		No	.	.
MS_DRG_CODE	Input	Interval	No		No	.	.
MS_DRG_DESC	Text	Nominal	No		No	.	.
Num_Chronic_Cond	Input	Interval	No		No	.	.
PATIENT_NUMBER	ID	Interval	No		No	.	.
PROCEDURE_ICD_CODE	Input	Nominal	No		No	.	.
PROCEDURE_LONG_DESC	Text	Nominal	No		No	.	.
PROCEDURE_SUBCAT_CODE	Input	Nominal	No		No	.	.
PROCEDURE_SUBCAT_DESC	Text	Nominal	No		No	.	.
PatientAge	Input	Interval	No		No	.	.
STATECODE	Input	Nominal	No		No	.	.
Standard_Orders_Used	Input	Nominal	No		No	.	.
ZIP	Input	Nominal	No		No	.	.
admit_month	Input	Interval	No		No	.	.
dx_code	Input	Interval	No		No	.	.
dx_group	Input	Nominal	No		No	.	.
gender	Input	Nominal	No		No	.	.
i	Input	Interval	No		No	.	.
icd9_target	Input	Interval	No		No	.	.
op_visits6	Input	Interval	No		No	.	.
operationcount	Input	Interval	No		No	.	.
order_set_used	Input	Interval	No		No	.	.
order_total_charges	Input	Interval	No		No	.	.
race_cd	Input	Nominal	No		No	.	.

Second part of list (for reference). Leave the readmit_date and readmit_discharge_date as Time ID variables. Click "OK" once you finish editing.

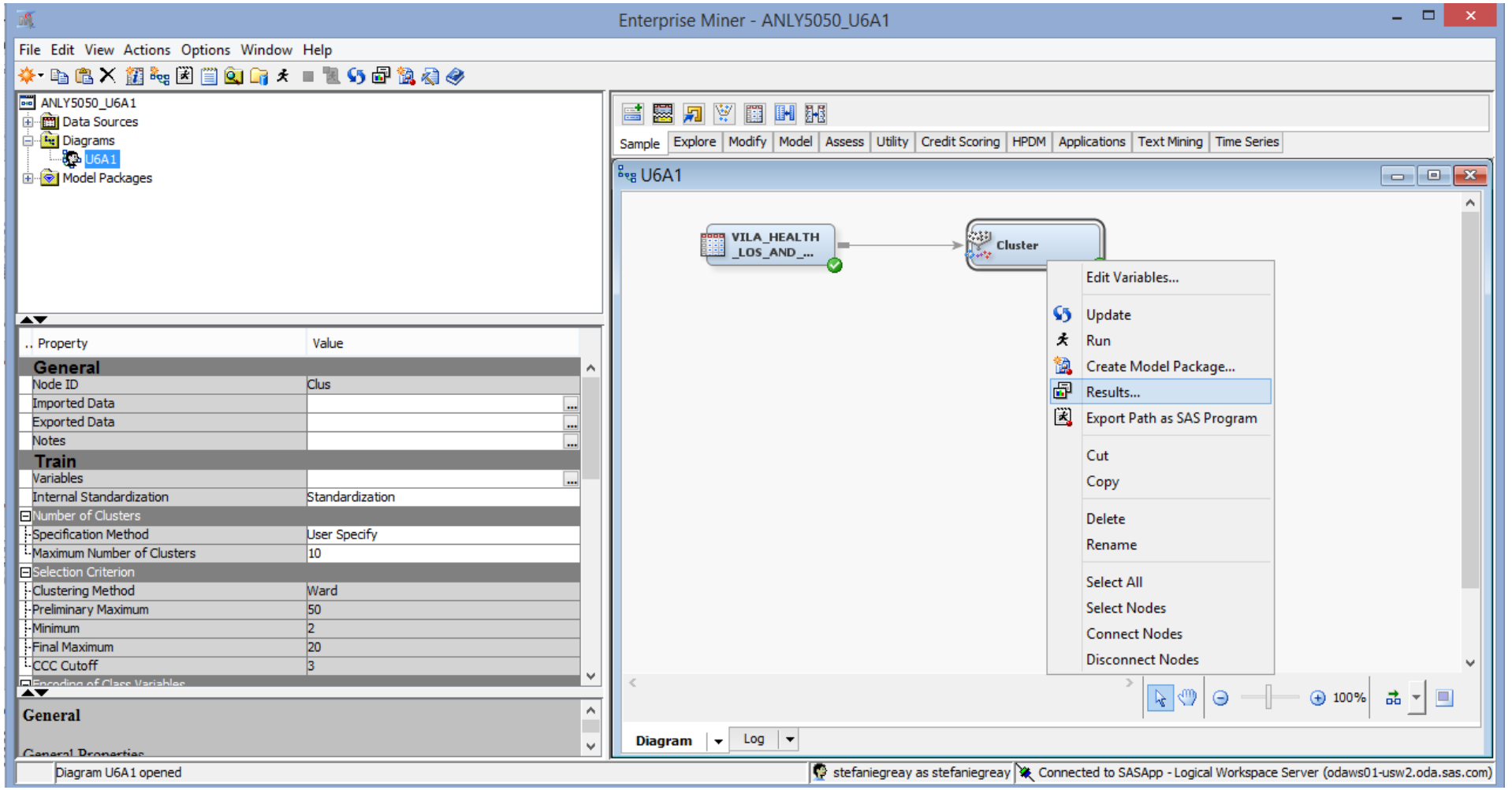


PROCEDURE_ICD_CODE	Input	Nominal	No		No	.	.
PROCEDURE_LONG_DESC	Text	Nominal	No		No	.	.
PROCEDURE_SUBCAT_CODE	Input	Nominal	No		No	.	.
PROCEDURE_SUBCAT_DESC	Text	Nominal	No		No	.	.
PatientAge	Input	Interval	No		No	.	.
STATECODE	Input	Nominal	No		No	.	.
Standard_Orders_Used	Input	Nominal	No		No	.	.
ZIP	Input	Nominal	No		No	.	.
admit_month	Input	Interval	No		No	.	.
dx_code	Input	Interval	No		No	.	.
dx_group	Input	Nominal	No		No	.	.
gender	Input	Nominal	No		No	.	.
i	Input	Interval	No		No	.	.
icd9_target	Input	Interval	No		No	.	.
op_visits6	Input	Interval	No		No	.	.
operationcount	Input	Interval	No		No	.	.
order_set_used	Input	Interval	No		No	.	.
order_total_charges	Input	Interval	No		No	.	.
race_cd	Input	Nominal	No		No	.	.
readmit_date	Time ID	Interval	No		No	.	.
readmit_days	Input	Nominal	No		No	.	.
readmit_discharge_date	Time ID	Interval	No		No	.	.
readmit_month	Input	Interval	No		No	.	.
readmit_number	Input	Interval	No		No	.	.

Right click on “Cluster” node and click “Run”



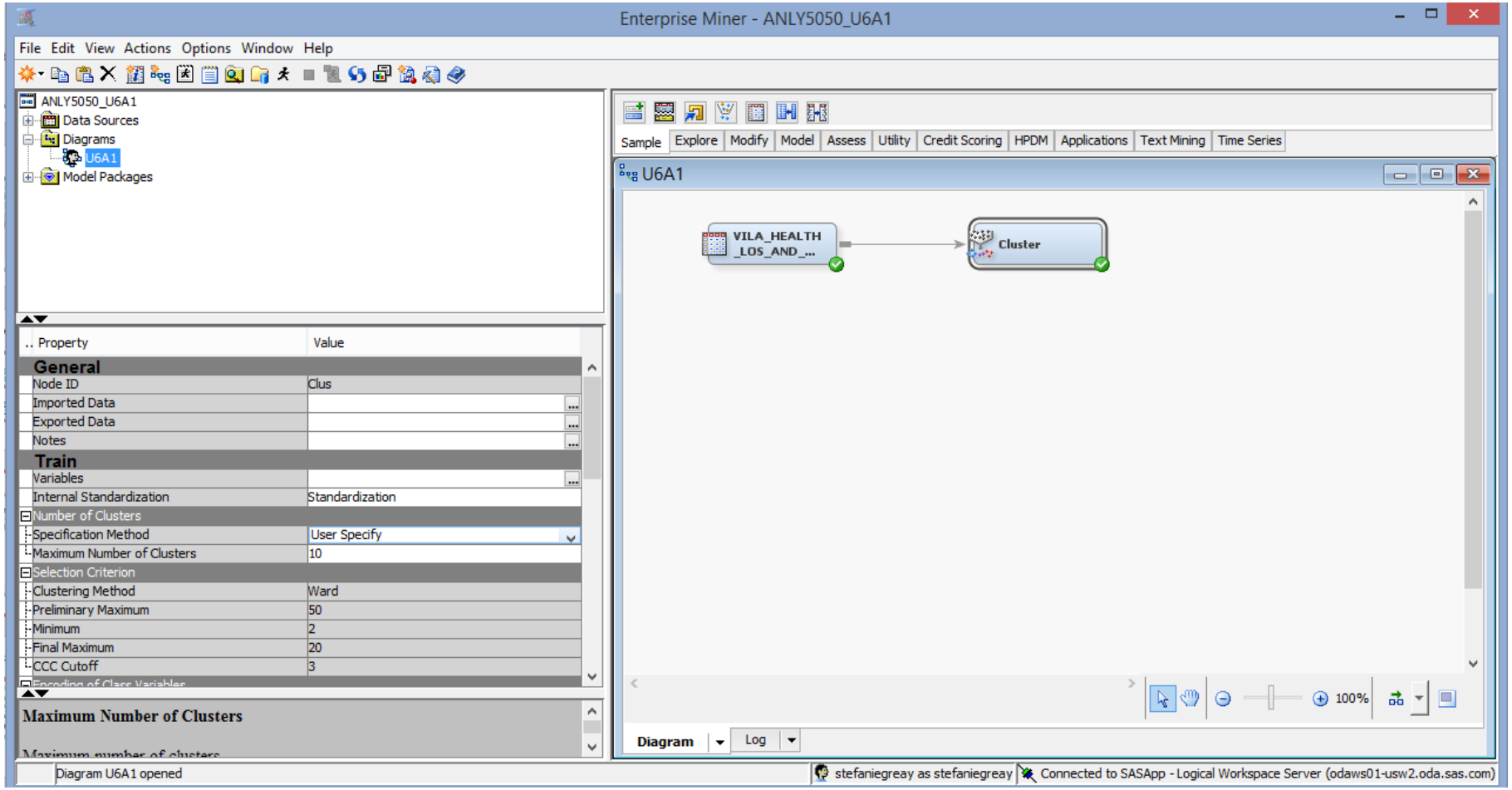
Right click on the “Cluster” node and click “Results” to view the results.



Selecting the “Optimal” Number of Clusters

- There are many options to select the optimal number of clusters, and it is an art, not a science.
- My preferred method is using the CCC in SAS (the Cubic Cluster Criterion), and plot the change to identify a “good” stopping point for the number of clusters. More details about this criterion are available here: https://documentation.sas.com/?docsetId=emref&docsetTarget=n1dm4owbc3ka5jn11yjkod7ov1va.htm&docsetVersion=14.3&locale=en
- SAS Enterprise Miner has many options for allowing it to select the optimal number of clusters, as well.

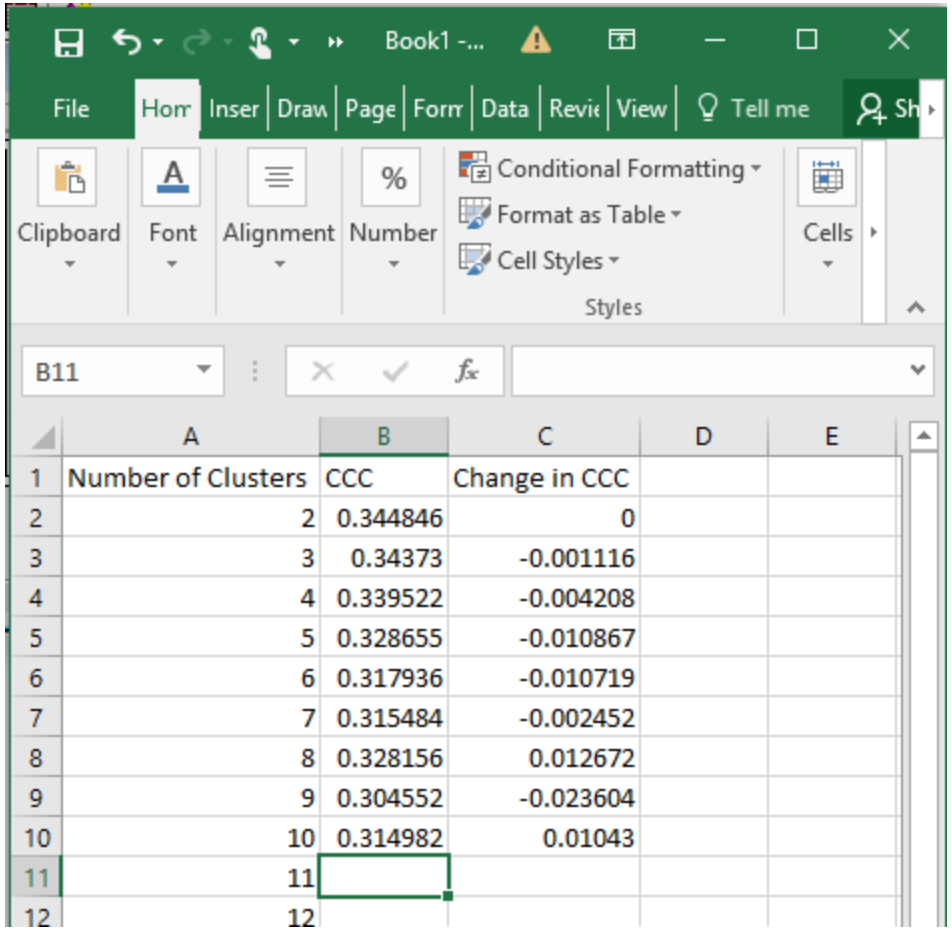
To manually specify the number of clusters to create, choose “User Specify” under the Specification Method and enter the desired number of clusters directly below.



Using the CCC change to select the “Optimal” Number of Clusters

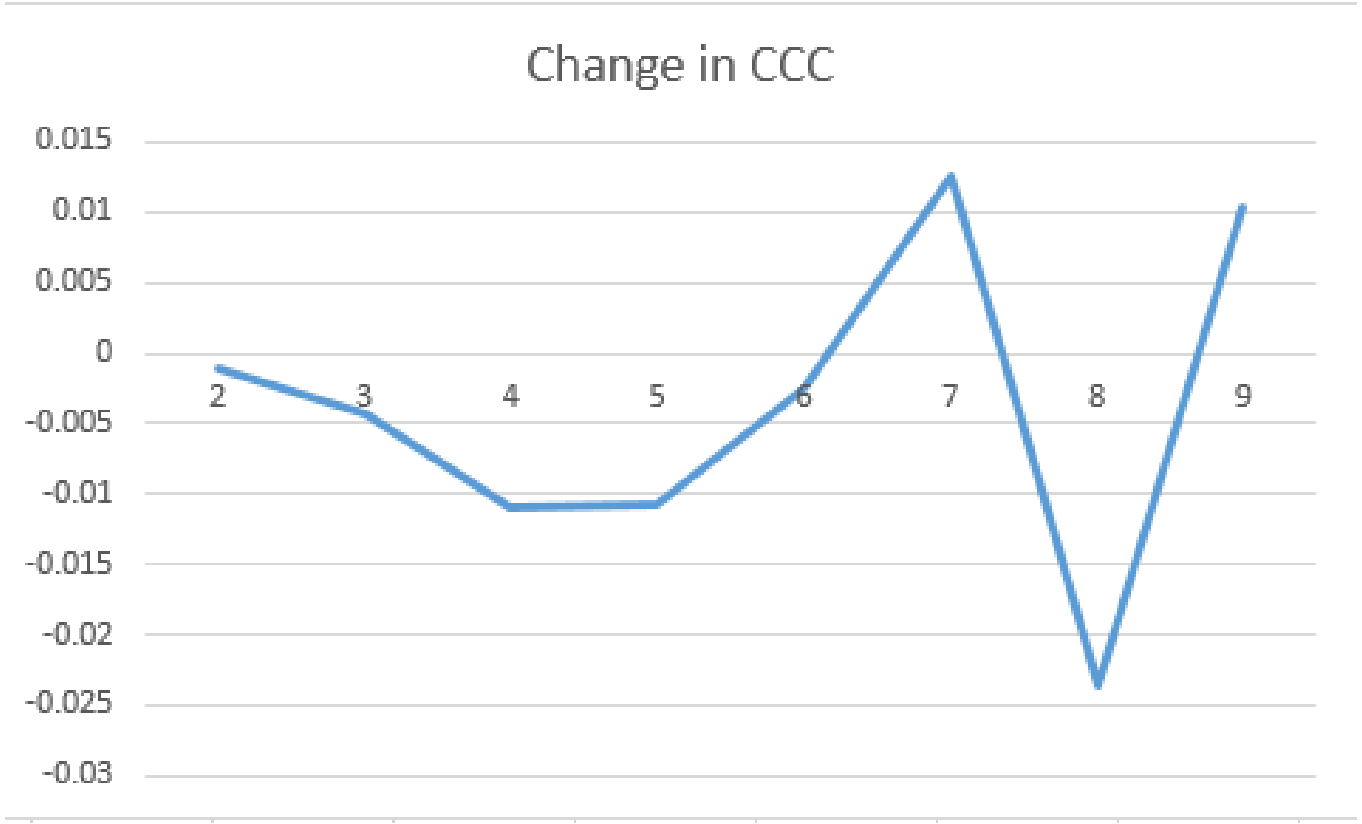
- To use the change in CCC to select the “optimal” number of clusters manually, set the specification method to “User Specify” and start at number of clusters equal to 2.
- Run the cluster node and go to the results.
- In the results, note the CCC in the top right pane.
- Create a spreadsheet with “number of clusters,” “CCC,” and “Change in CCC” as columns.
- Re-run this process, entering the number of clusters and associated CCC for each run, starting at 2 and increasing by 1 each run.
- Calculate the change in CCC value for each run by taking the CCC for each run and subtracting the CCC for the previous run.

The sheet should look something like this.



	A	B	C	D	E
1	Number of Clusters	CCC	Change in CCC		
2	2	0.344846	0		
3	3	0.34373	-0.001116		
4	4	0.339522	-0.004208		
5	5	0.328655	-0.010867		
6	6	0.317936	-0.010719		
7	7	0.315484	-0.002452		
8	8	0.328156	0.012672		
9	9	0.304552	-0.023604		
10	10	0.314982	0.01043		
11	11				
12	12				

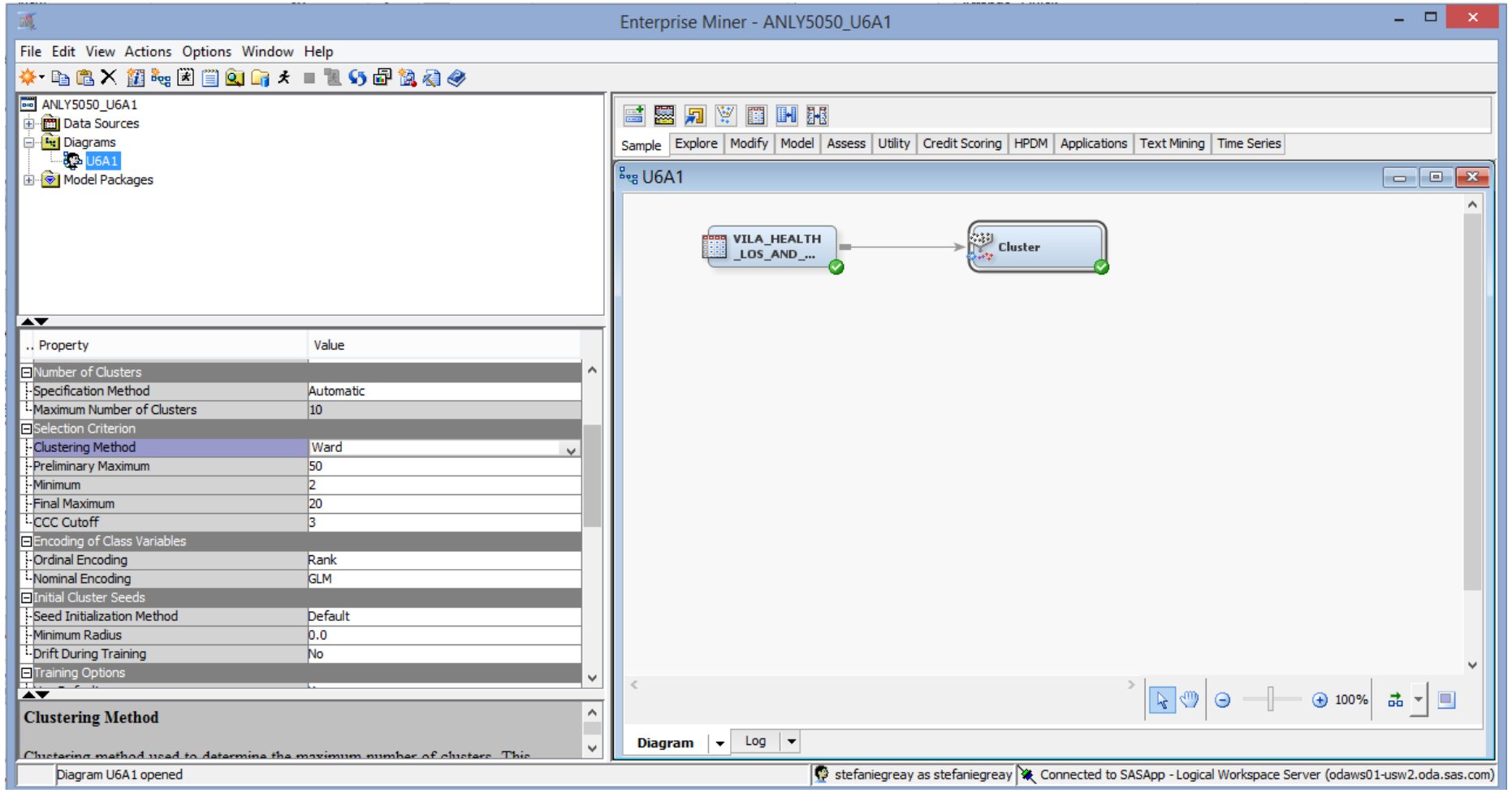
You can then graph (using a line graph) the change in CCC column, which should look something like this. One of the “elbows” or points of great change would be a good option for selecting a number of clusters.



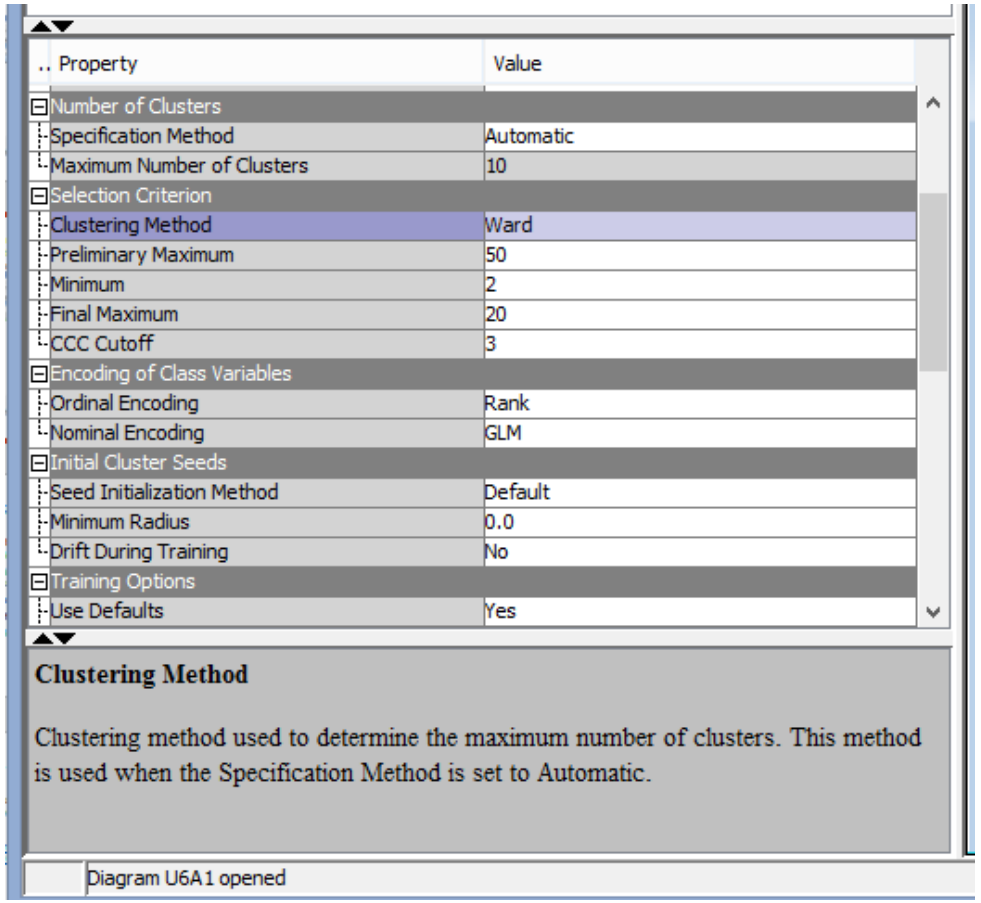
After number of cluster identification/selection

- After you select or identify a number of clusters, re-run the cluster node with that specific number of clusters
- Then view/review the results to analyze and interpret the clusters themselves
- Two areas you will definitely want to review are:
  - The variables that contribute to the creation of each of the clusters
  - The aspects of those variables for the observations in each cluster (i.e. frequencies for categorical variables and summaries for numeric variables)

To allow SAS Enterprise Miner to determine the number of clusters to create, choose “Automatic” under the Specification Method and enter the selection options in the Selection Criteria section below that.



If you click on the grey area of the specific option, a summary of that property is shown in the pane below the property pane we are editing. SAS's product help documentation for the Cluster Node goes into more detail for each of these options.



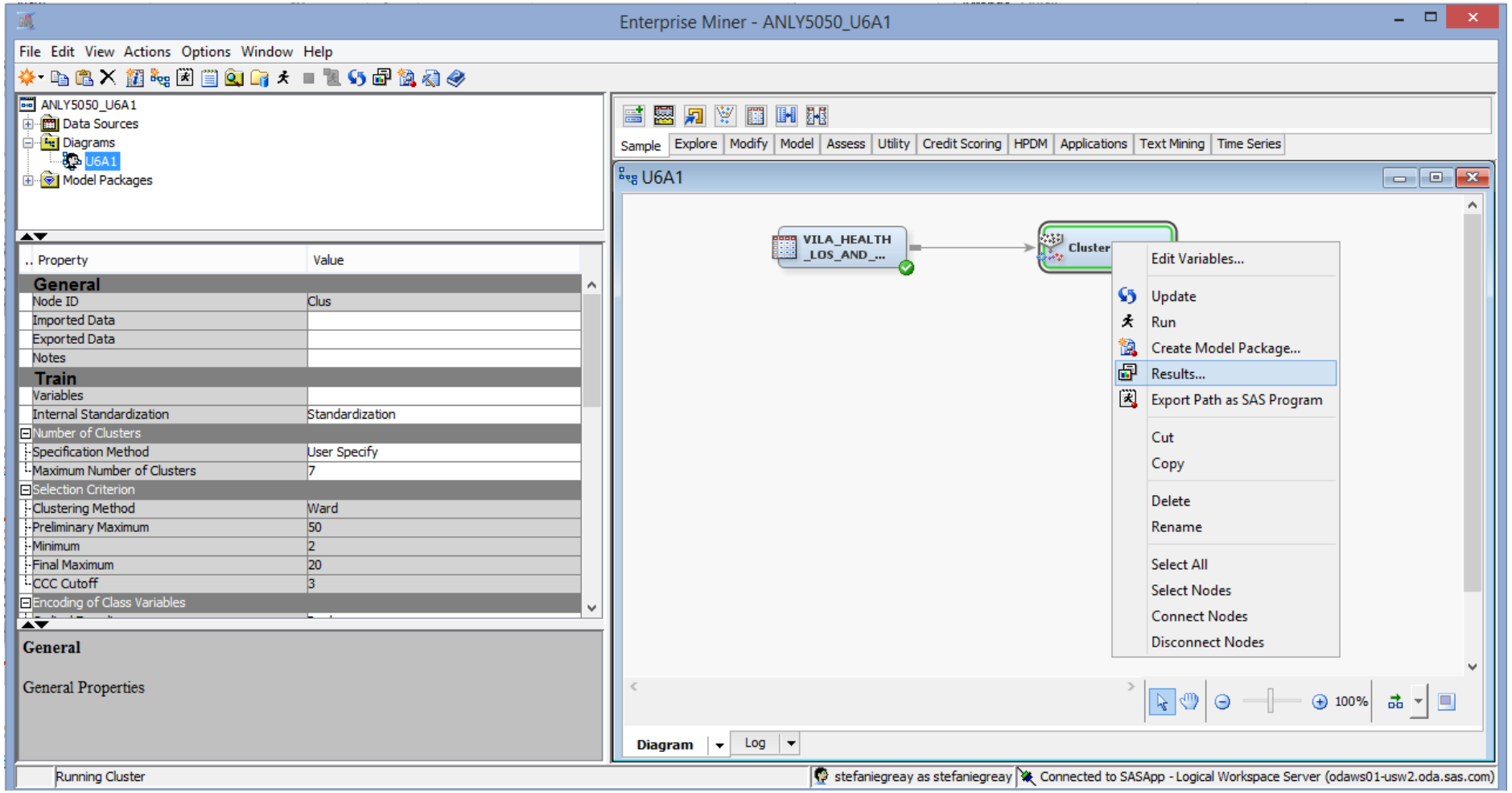
Considerations and Standards

1) Sample size and whether or not to subset into train/test/validate
2) Supervised vs Unsupervised
3) Variables to include (and whether to reduce the variables first, by using PCA, for example)
4) Number of clusters
5) Interpreting the clusters

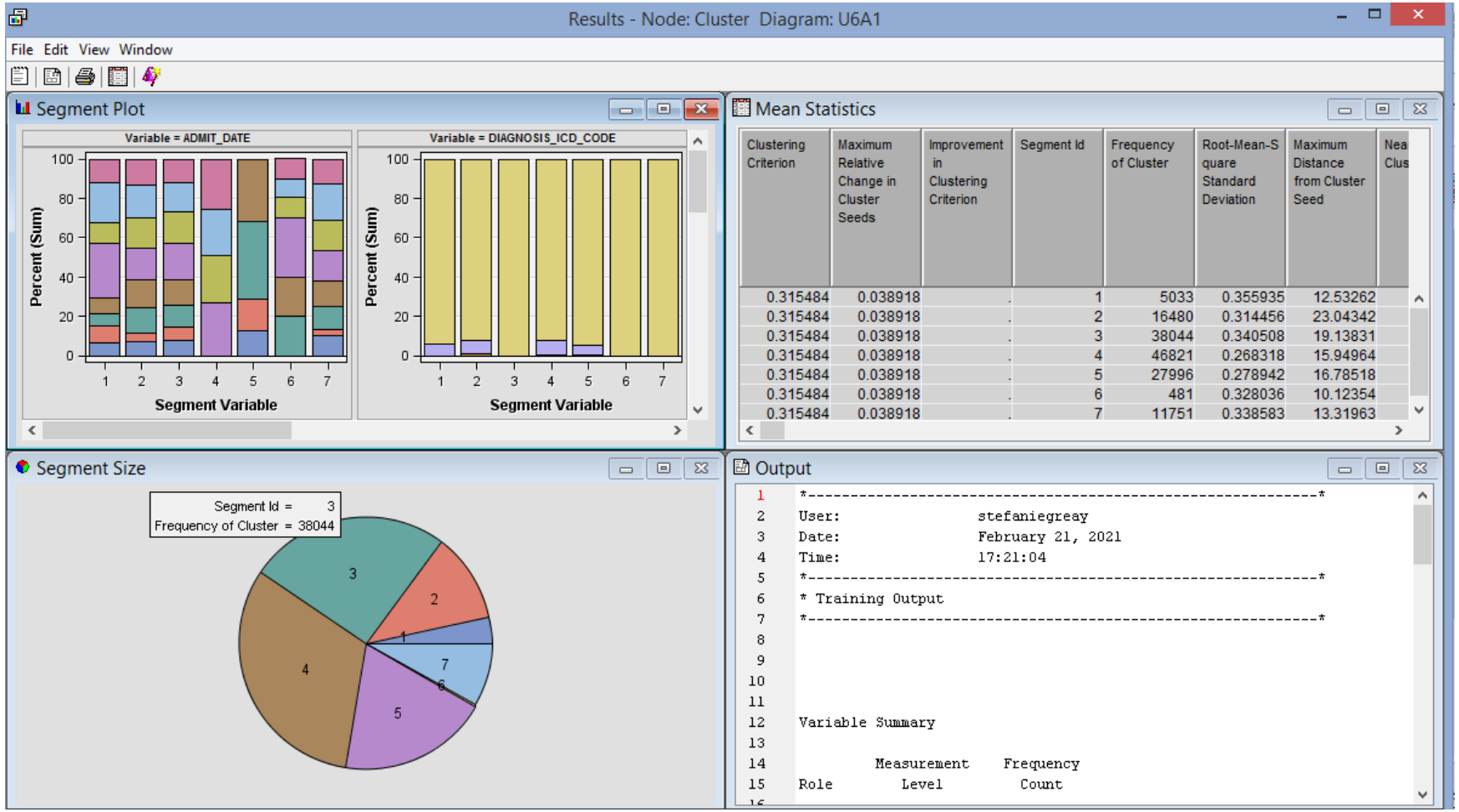
It's a bit older but here is a good review of some of the standards used in marketing (and whether they are supported in academic research):

Dolnicar, S. (2002). A review of unquestioned standards in using cluster analysis for data-driven market segmentation. Faculty of Commerce-Papers, 273. http://ro.uow.edu.au/cgi/viewcontent.cgi?article=1286&context=commpapers

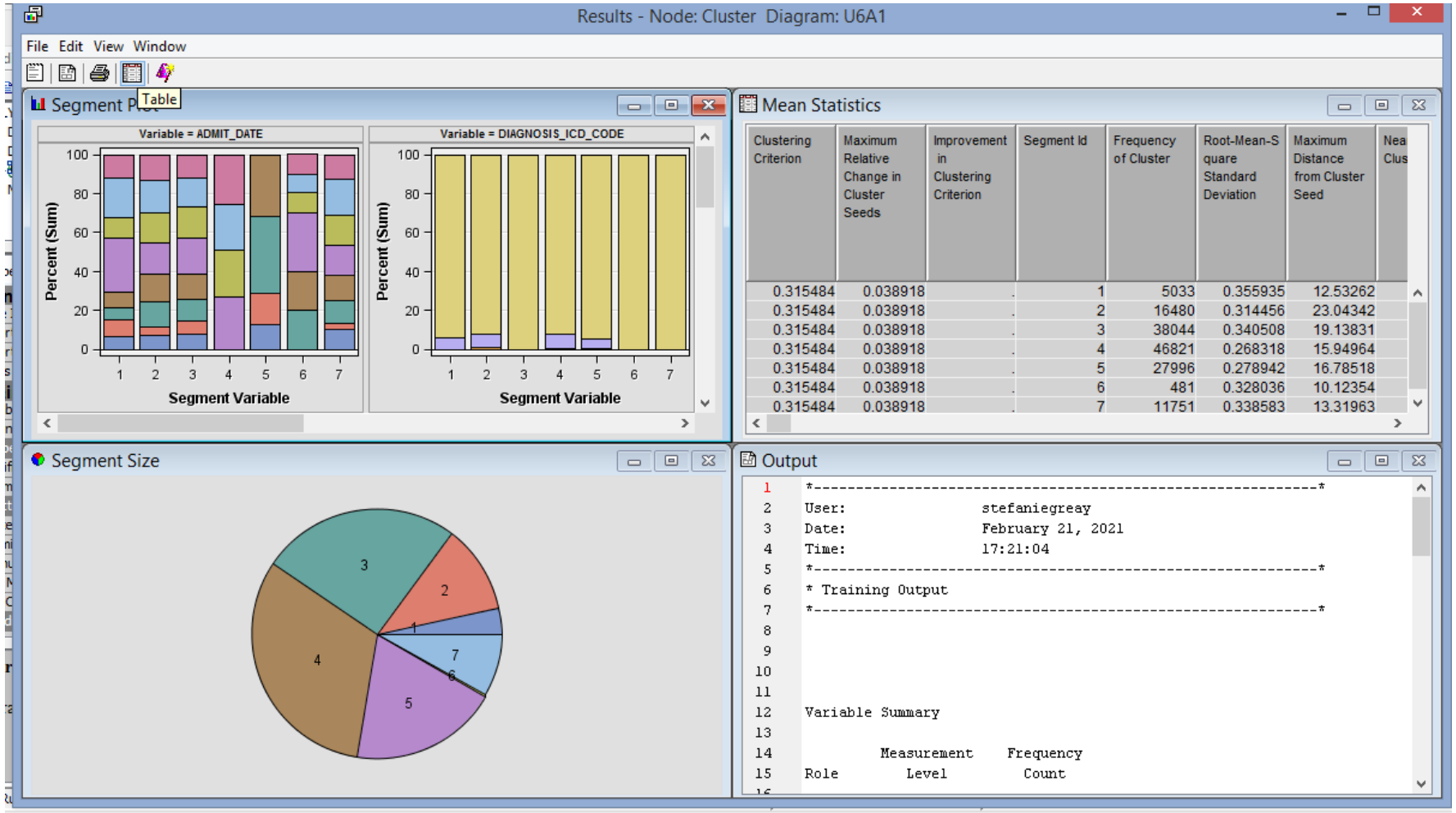
To review the results of the cluster analysis, right click on the “Cluster” node and select “Results.”



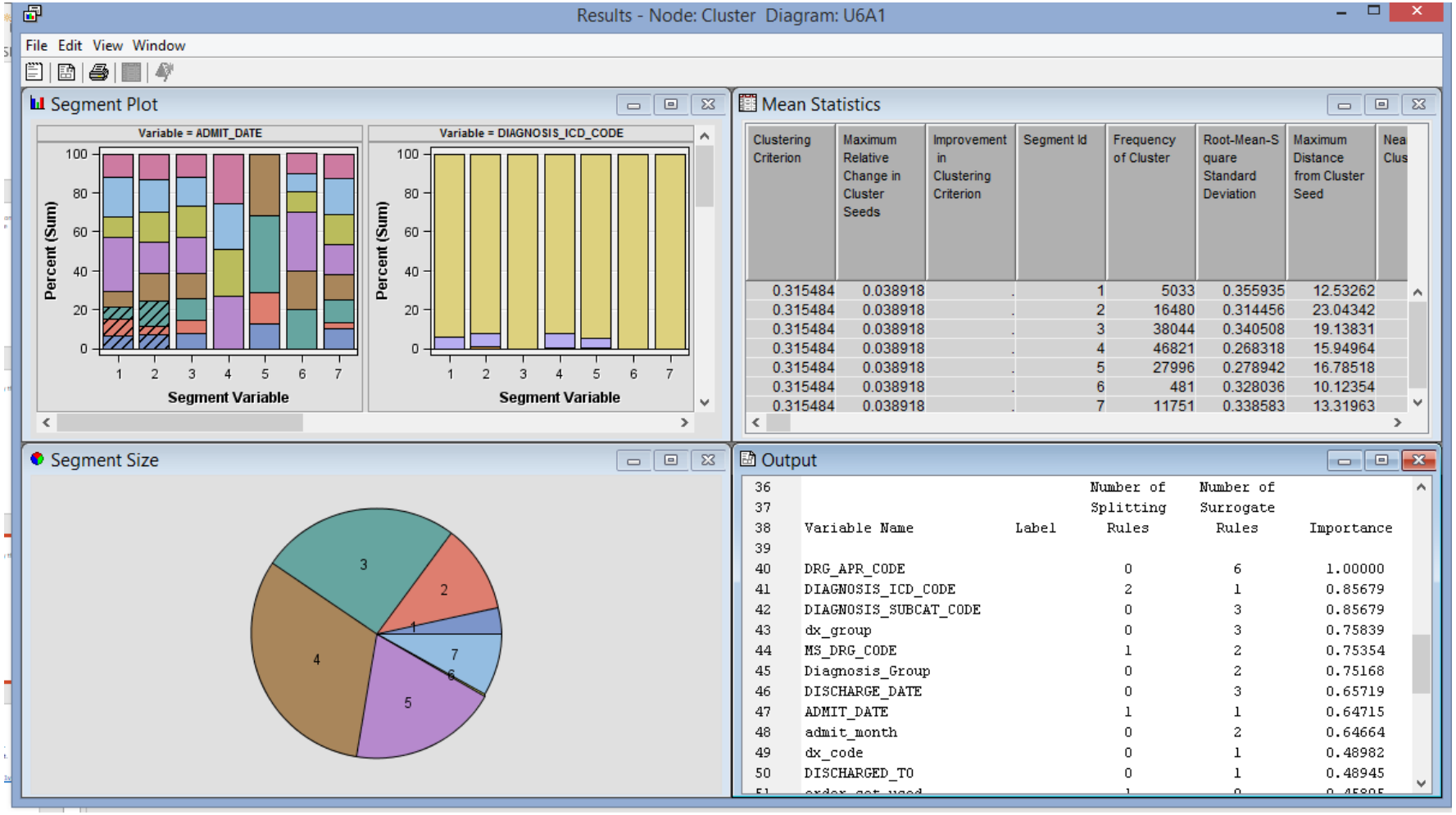
The segment chart in the bottom left will show how many observations fall into each segment/cluster if you hover over each slice, so will the top right table. Use this to summarize the size of the clusters.



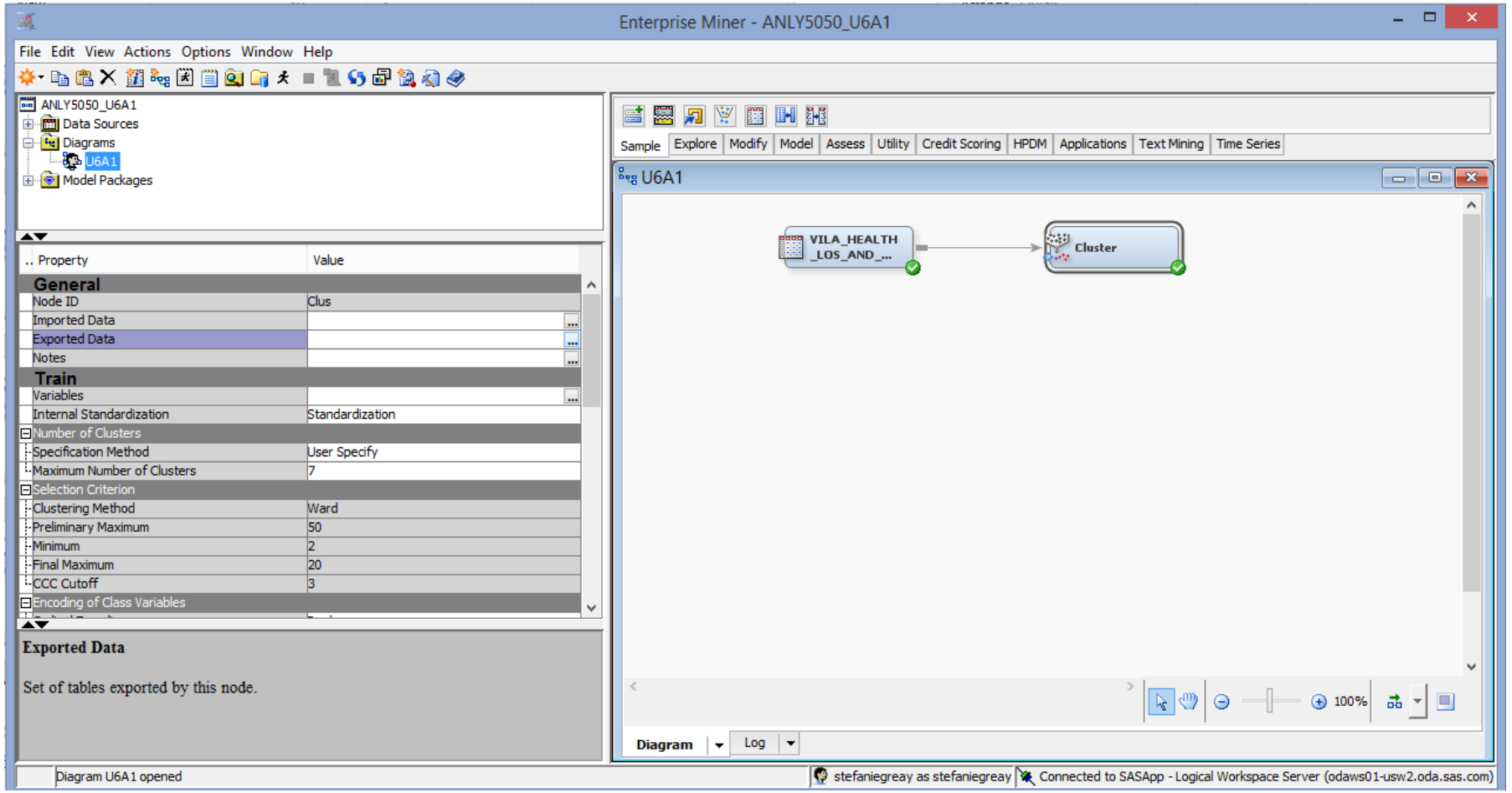
Clicking on “Table” (the picture of a table) on the top left will display the table of the segment plot.



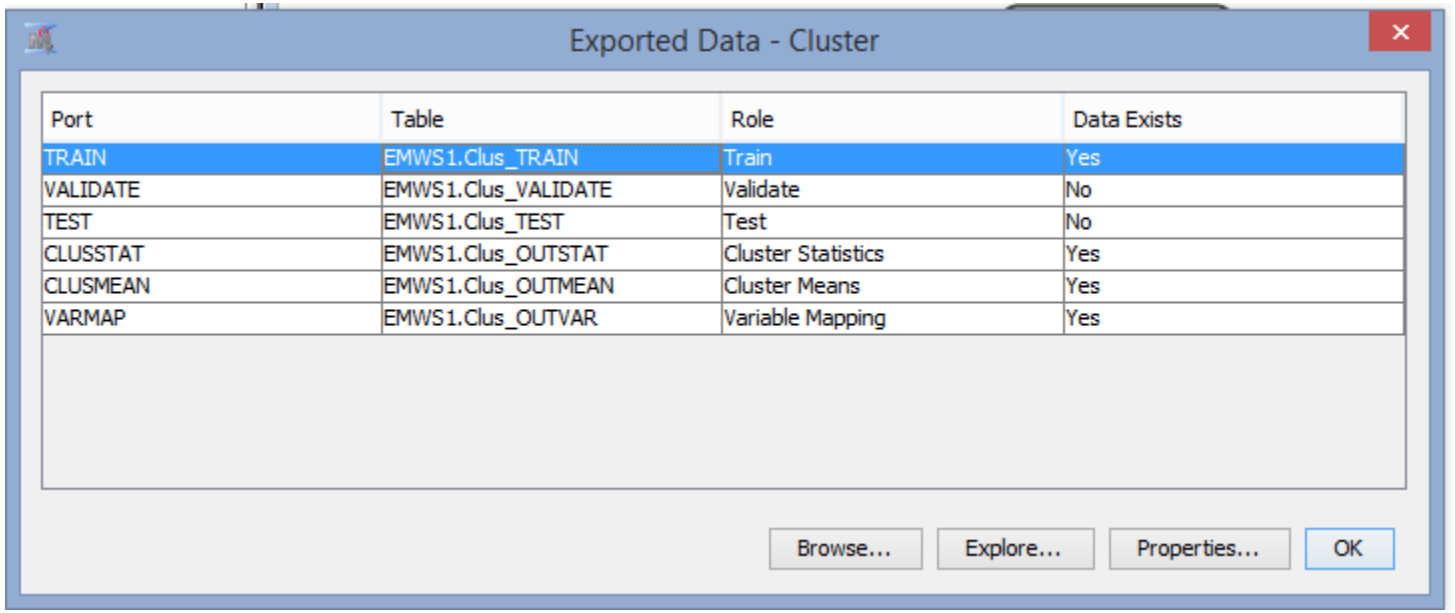
Scrolling down in the Output pane, you can see the importance of the variables in creating the clusters.



To view the cluster numbers merged back to the original dataset, so that you can identify and analyze the observations within the clusters, right click on the Cluster node, then click the elipses next to output data.



The EMWS1.Clus_TRAIN dataset has the original data with the cluster number and description included in it. You can explore that by clicking on “Explore,” review the properties of it by clicking on “Properties,” or may access this data in a code node.



Analyzing the clusters

1) You may use a data step to create a separate dataset with the observations from each cluster, and analyze it that way, or
2) You may use proc freq and proc univariate with the appropriate types of variables with by statements for the cluster variable, which has a label of “Segment ID” but you can check the specific variable name in the properties of that dataset.
3) You can technically do this for every variable used in the creation of the clusters, but it’s often only useful to do so for the top x number of contributing variables, as this usually is most helpful for explaining/describing each cluster and it’s contents.

SAS Documentation Reference

The link below brings you to the SAS Documentation on the Cluster Node, which has an example, including interpretation of the output.

https://documentation.sas.com/?docsetId=emref&docsetTarget=n1vjatb74dundbn12d2ecb09juak.htm&docsetVersion=14.3&locale=en