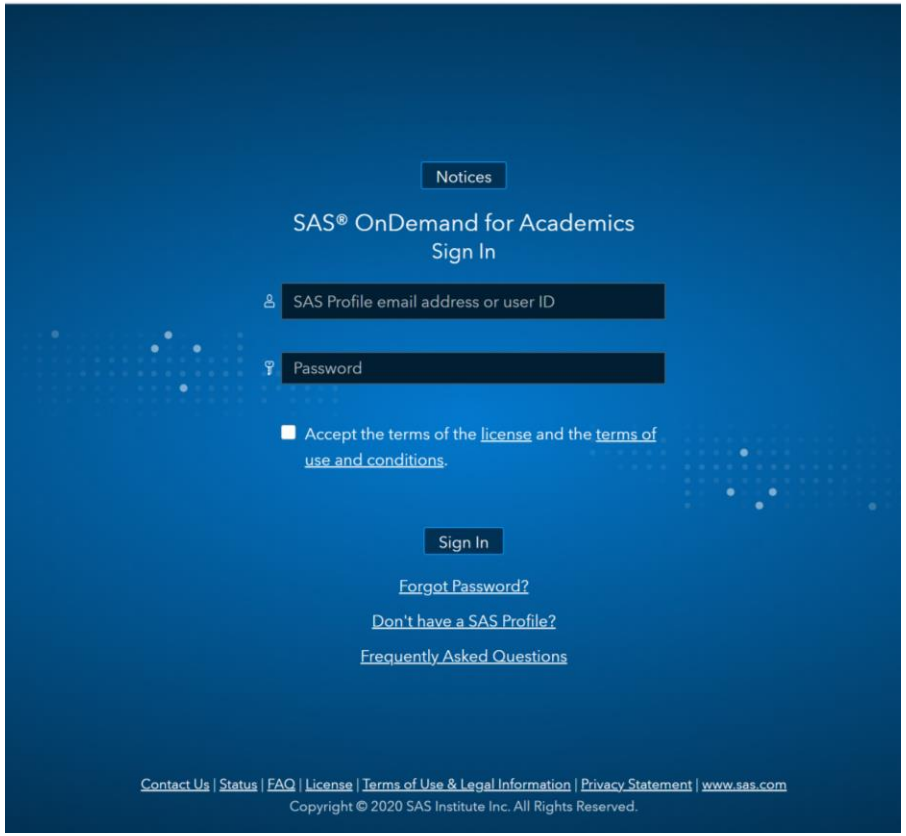
# ANLT5050

Unit 8 Assignment 2 Tutorial

# Access the SAS OnDemand for Academics Control Center

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



SAS® OnDemand for Academics
Sign In

SAS Profile email address or user ID

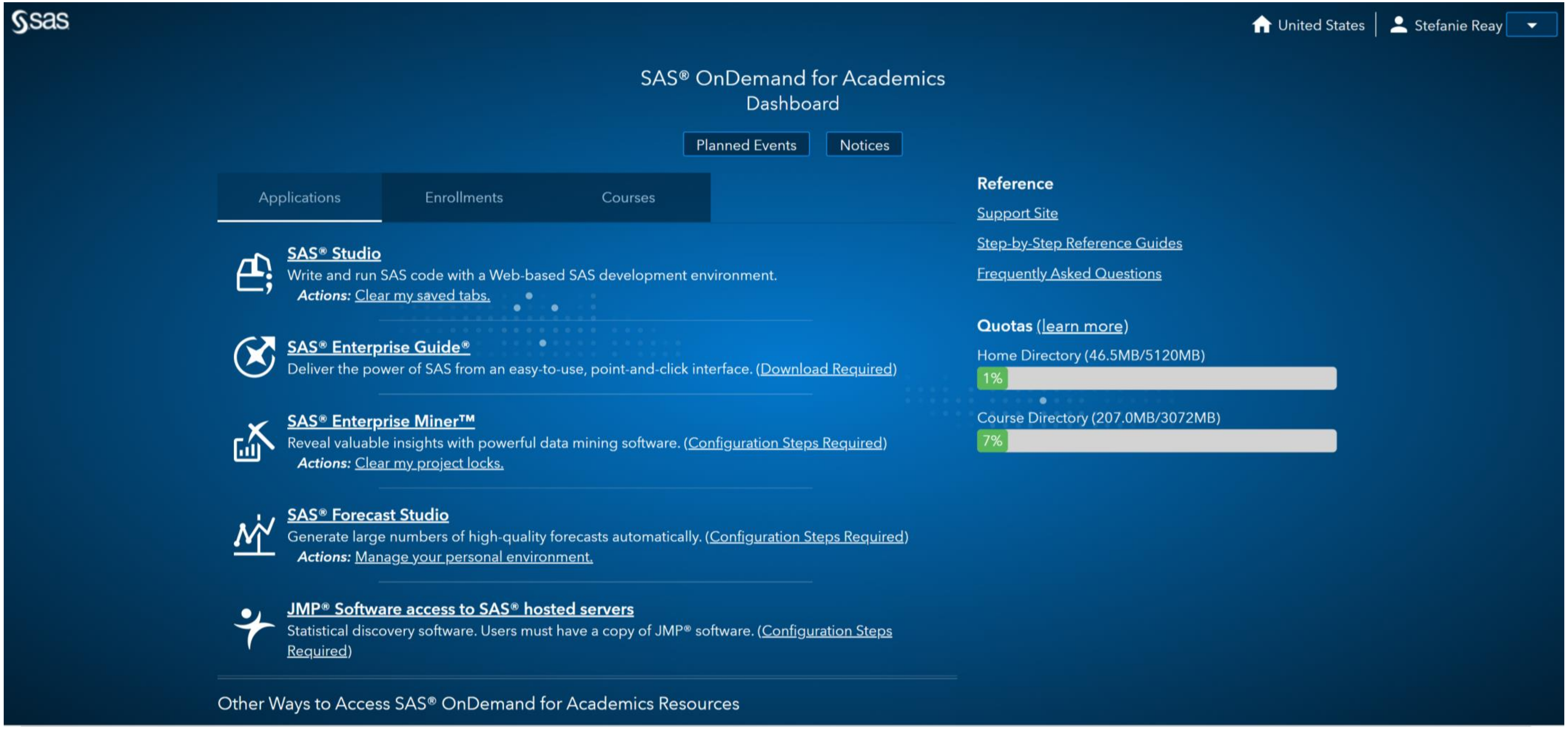
Password

□ Accept the terms of the license and the terms of use and conditions.

Notices
Sign In
Forgot Password?
Don't have a SAS Profile?
Frequently Asked Questions
Contact Us | Status | FAQ | License | Terms of Use & Legal Information | Privacy Statement | www.sas.com

Copyright © 2020 SAS Institute Inc. All Rights Reserved.

# SAS OnDemand for Academics (SODA) Control Center



SAS® OnDemand for Academics
Dashboard

**SAS® Studio**
Write and run SAS code with a Web-based SAS development environment.
*Actions:* Clear my saved tabs.

**SAS® Enterprise Guide®**
Deliver the power of SAS from an easy-to-use, point-and-click interface. (Download Required)

**SAS® Enterprise Miner™**
Reveal valuable insights with powerful data mining software. (Configuration Steps Required)
*Actions:* Clear my project locks.

**SAS® Forecast Studio**
Generate large numbers of high-quality forecasts automatically. (Configuration Steps Required)
*Actions:* Manage your personal environment.

**JMP® Software access to SAS® hosted servers**
Statistical discovery software. Users must have a copy of JMP® software. (Configuration Steps Required)

Other Ways to Access SAS® OnDemand for Academics Resources

**Reference**
Support Site
Step-by-Step Reference Guides
Frequently Asked Questions

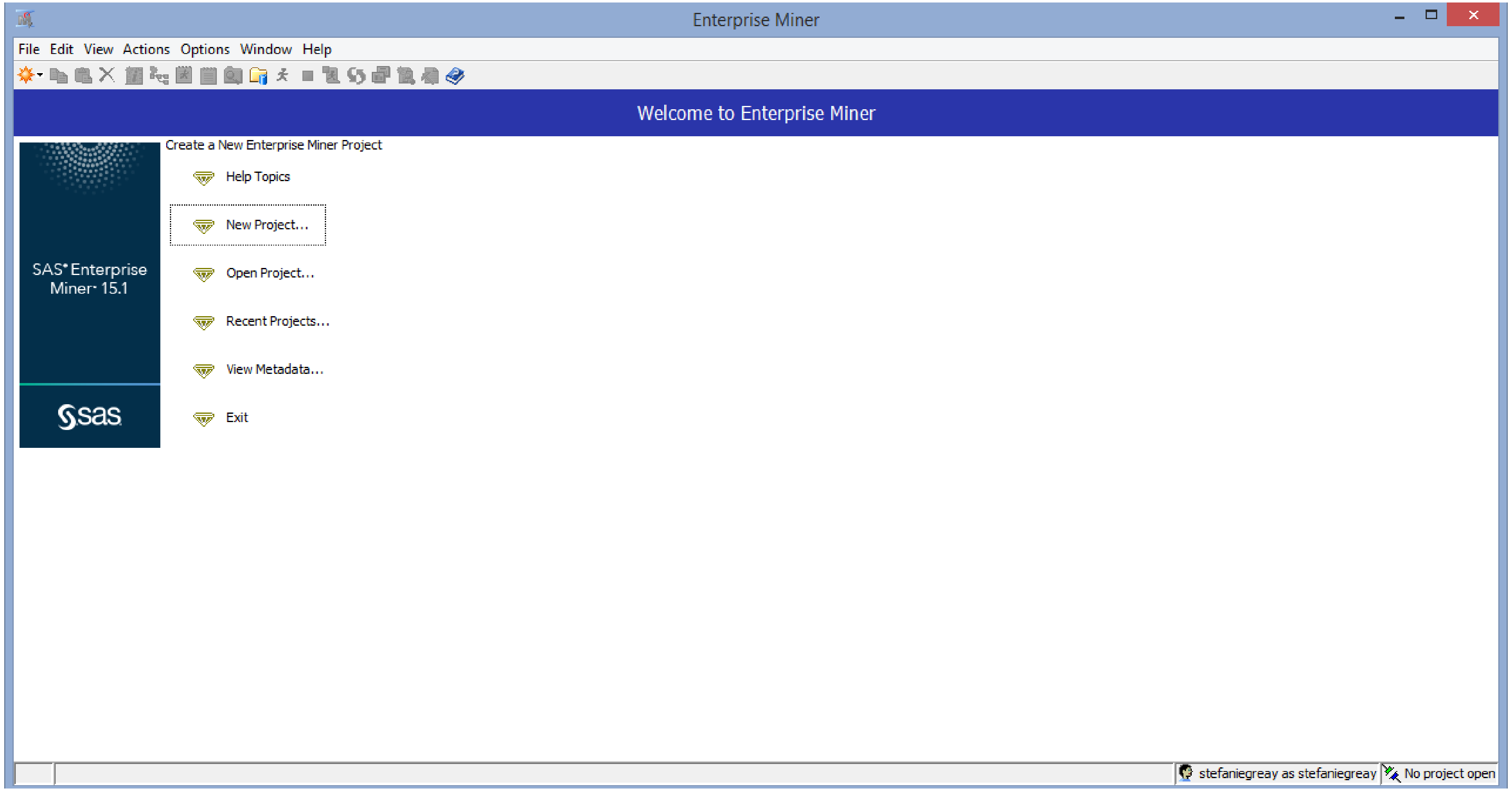
**Quotas** (learn more)
Home Directory (46.5MB/5120MB)
1%
Course Directory (207.0MB/3072MB)
7%

# 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."



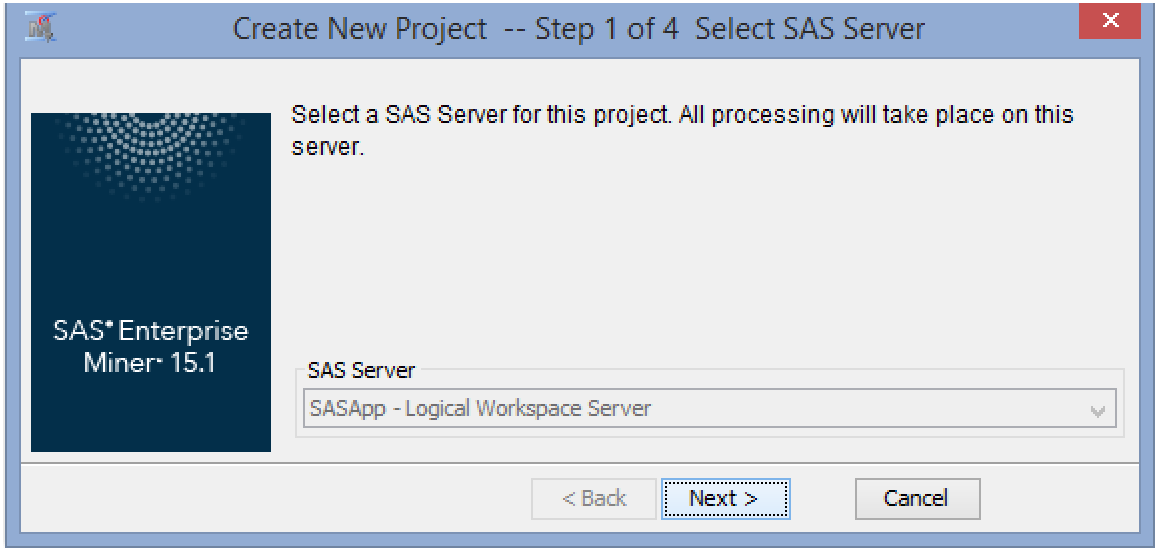
Enterprise Miner
File Edit View Actions Options Window Help
Welcome to Enterprise Miner
SAS® Enterprise Miner™ 15.1
sas

Create a New Enterprise Miner Project

- Help Topics
- New Project...
- Open Project...
- Recent Projects...
- View Metadata...
- Exit

stefaniegreay as stefaniegreay | No project open

# Click “Next>” to use the default SAS Server

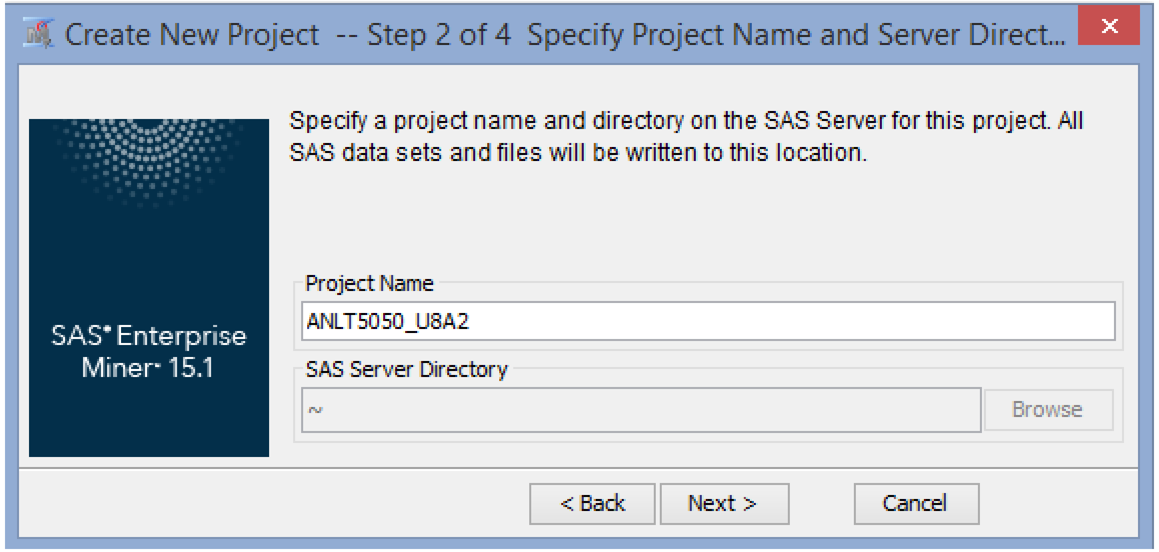


Create New Project -- Step 1 of 4 Select SAS Server

Select a SAS Server for this project. All processing will take place on this server.

SAS Server: SASApp - Logical Workspace Server

# Enter a project name and click “Next>”



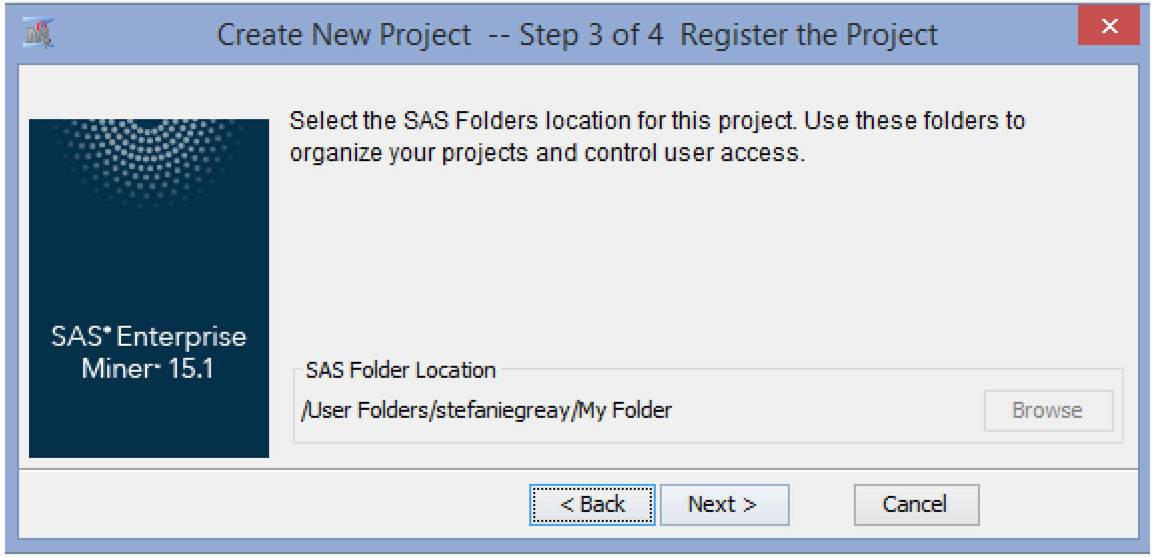
Create New Project -- Step 2 of 4 Specify Project Name and Server Direct...

Specify a project name and directory on the SAS Server for this project. All SAS data sets and files will be written to this location.

Project Name: ANLT5050_U8A2

SAS Server Directory: ~

# Click “Next>”



Create New Project -- Step 3 of 4 Register the Project

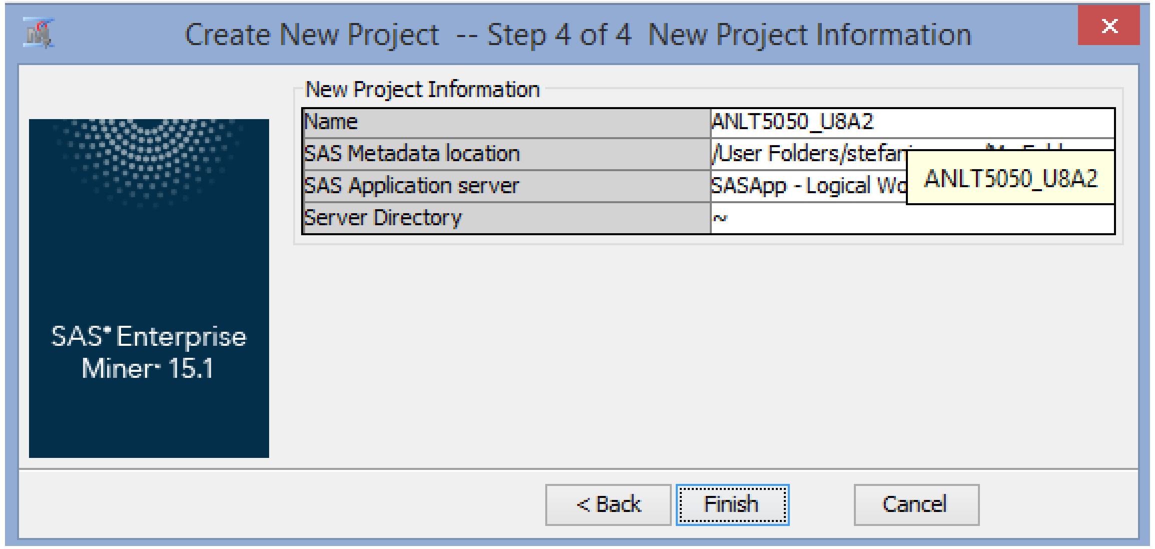
Select the SAS Folders location for this project. Use these folders to organize your projects and control user access.

SAS Folder Location

/User Folders/stefaniegreay/My Folder

© Stefanie G. Reay, MS, PhD, Capella University, 2021

# Verify your entries and click “Finish”

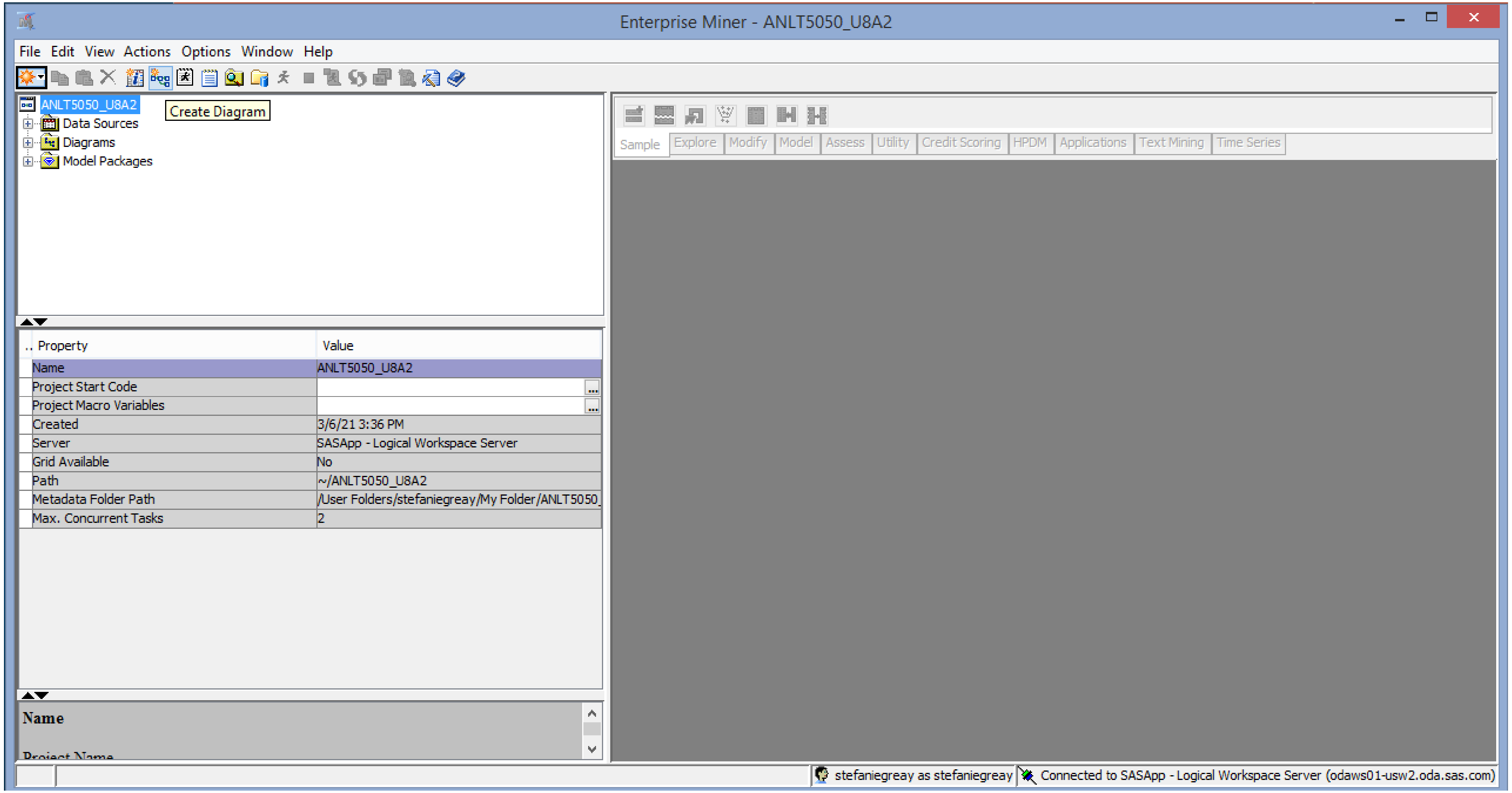

Create New Project -- Step 4 of 4 New Project Information
SAS® Enterprise Miner™ 15.1

New Project Information

| | |
|---|---|
| Name | ANLT5050_U8A2 |
| SAS Metadata location | /User Folders/stefan |
| SAS Application server | SASApp - Logical Wo |
| Server Directory | ~ |

ANLT5050_U8A2
< Back
Finish
Cancel

# Click on the “Create Diagram” icon.



| Property | Value |
|---|---|
| Name | ANLT5050_U8A2 |
| Project Start Code | |
| Project Macro Variables | |
| Created | 3/6/21 3:36 PM |
| Server | SASApp - Logical Workspace Server |
| Grid Available | No |
| Path | ~/ANLT5050_U8A2 |
| Metadata Folder Path | /User Folders/stefaniegreay/My Folder/ANLT5050_ |
| Max. Concurrent Tasks | 2 |

**Name**

Project Name

stefaniegreay as stefaniegreay

Connected to SASApp - Logical Workspace Server (odaws01-usw2.oda.sas.com)

# Name your diagram and click “OK.”

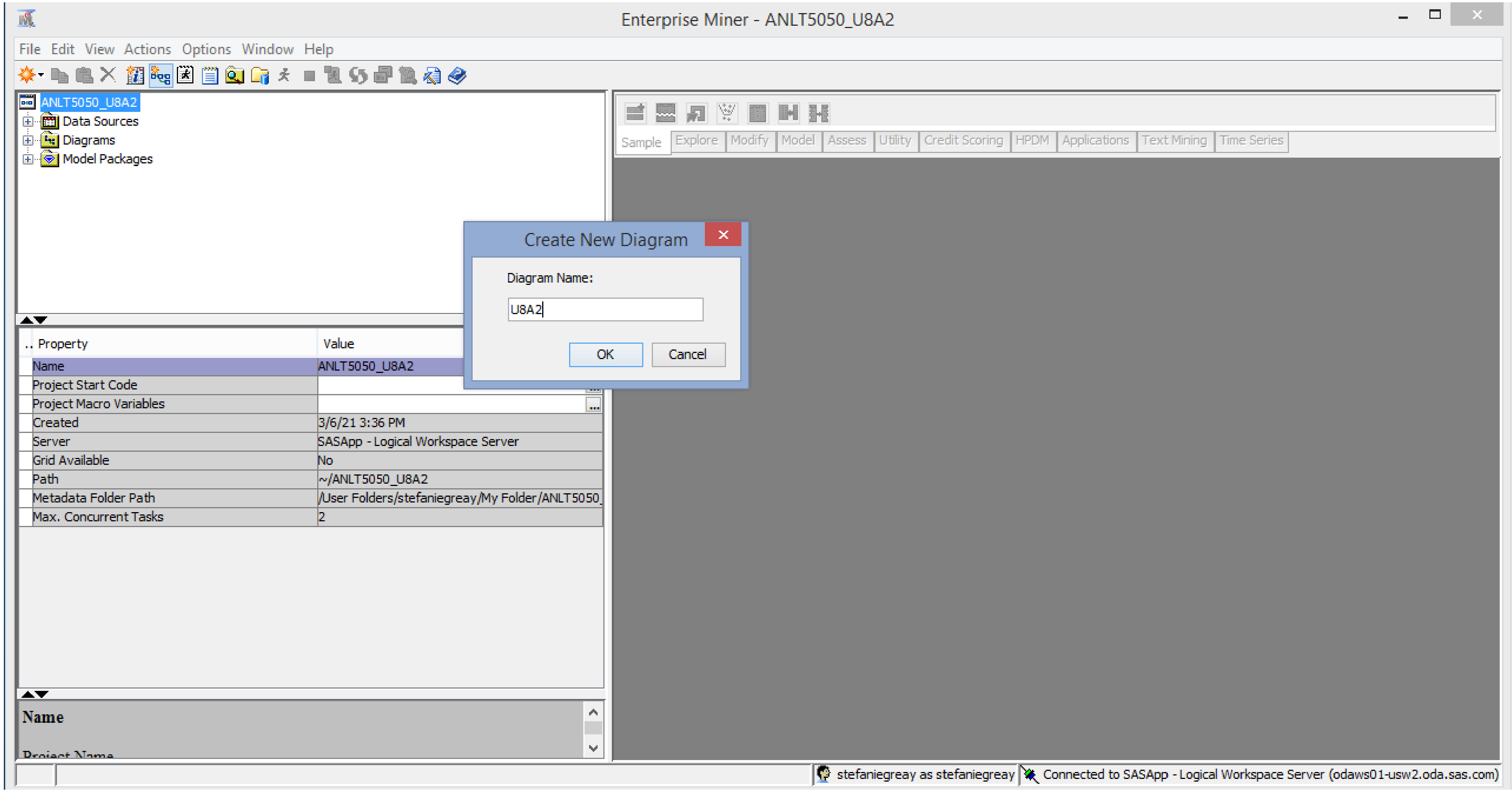
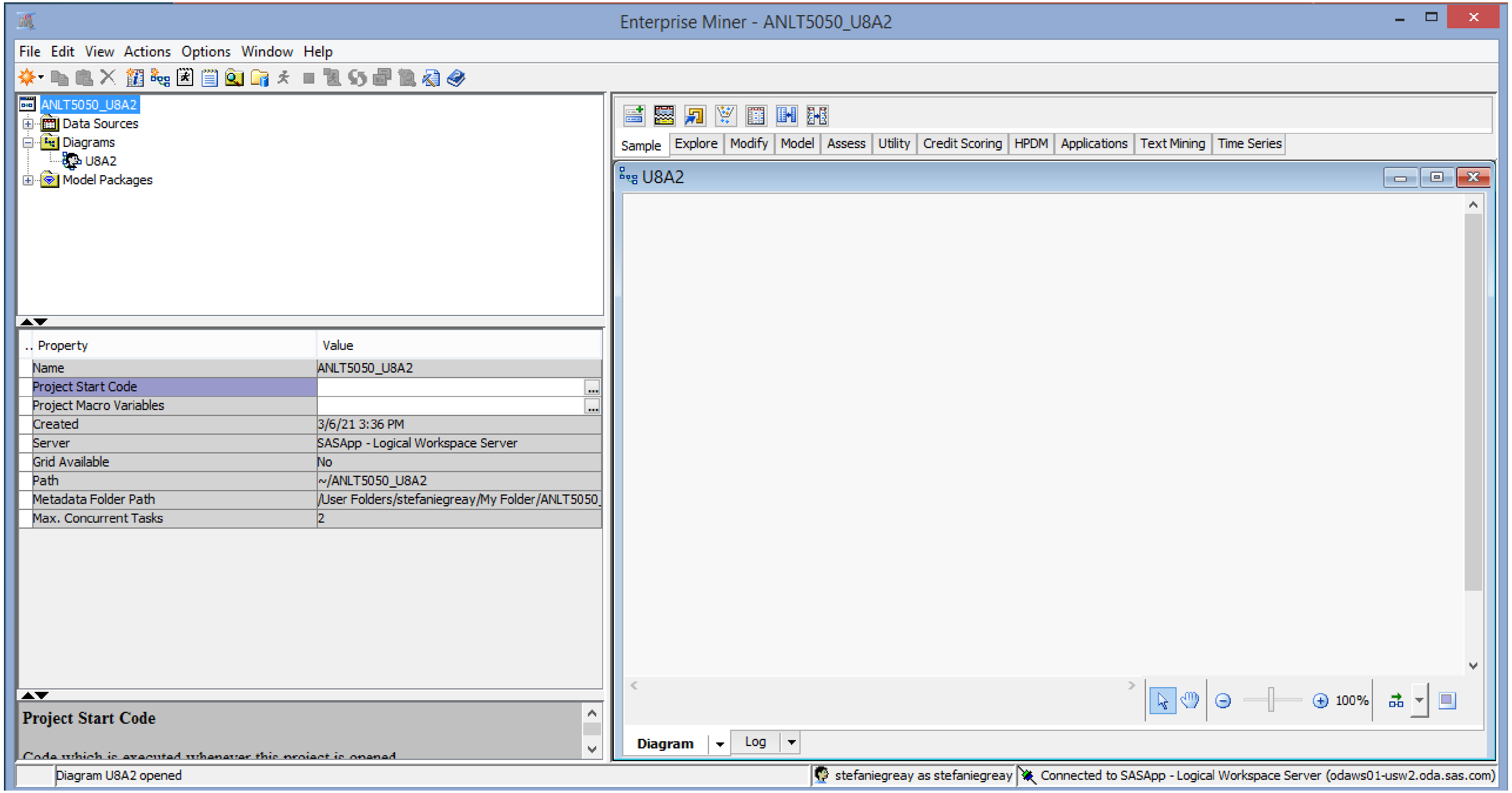


Diagram Name: U8A2

| Property | Value |
|---|---|
| Name | ANLT5050_U8A2 |
| Project Start Code | |
| Project Macro Variables | |
| Created | 3/6/21 3:36 PM |
| Server | SASApp - Logical Workspace Server |
| Grid Available | No |
| Path | ~/ANLT5050_U8A2 |
| Metadata Folder Path | /User Folders/stefaniegreay/My Folder/ANLT5050_ |
| Max. Concurrent Tasks | 2 |

# Click on the project, then click on the ellipses next to "Project Start Code."



Enterprise Miner - ANLT5050_U8A2
File Edit View Actions Options Window Help
ANLT5050_U8A2
Data Sources
Diagrams
U8A2
Model Packages
Sample Explore Modify Model Assess Utility Credit Scoring HPDM Applications Text Mining Time Series
U8A2
100%
Diagram Log

| Property | Value |
|---|---|
| Name | ANLT5050_U8A2 |
| Project Start Code | ... |
| Project Macro Variables | ... |
| Created | 3/6/21 3:36 PM |
| Server | SASApp - Logical Workspace Server |
| Grid Available | No |
| Path | ~/ANLT5050_U8A2 |
| Metadata Folder Path | /User Folders/stefaniegreay/My Folder/ANLT5050_ |
| Max. Concurrent Tasks | 2 |

**Project Start Code**

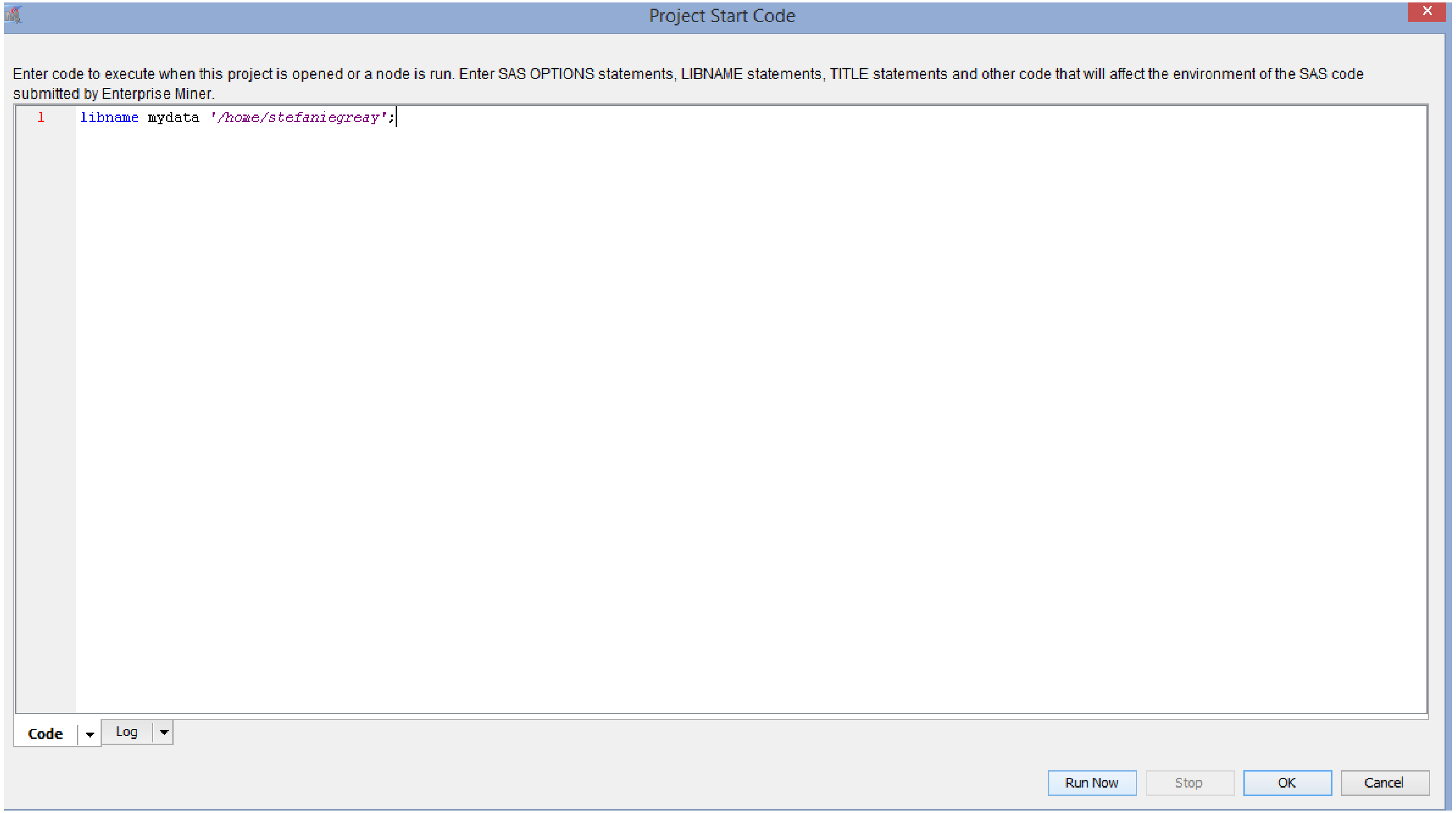
Code which is executed whenever this project is opened

Diagram U8A2 opened

stefaniegreay as stefaniegreay

Connected to SASApp - Logical Workspace Server (odaws01-usw2.oda.sas.com)

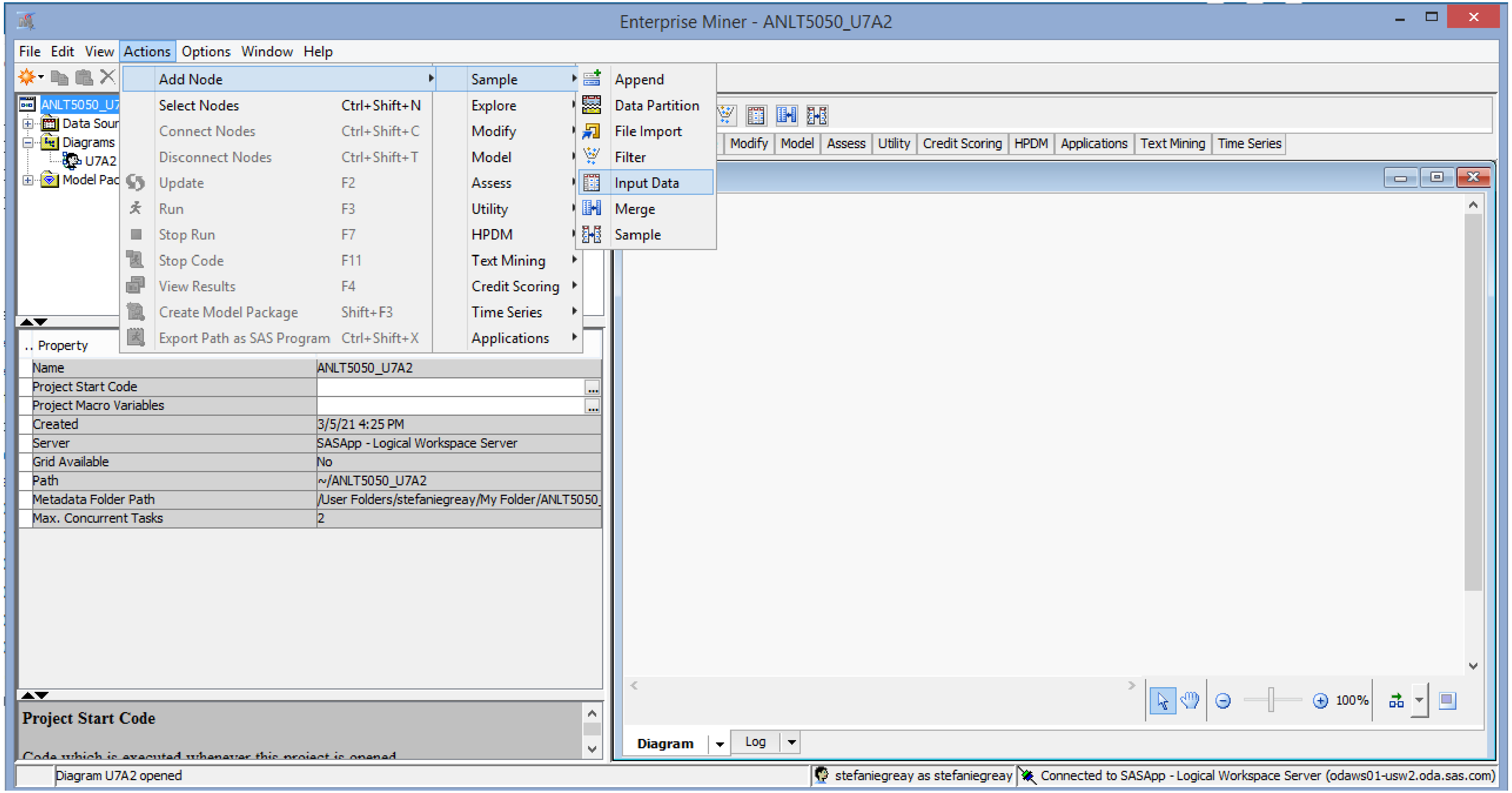
# Add the library reference for where you uploaded the dataset in SAS studio, and click "Run Now." Once it completes, click "OK."



Enter code to execute when this project is opened or a node is run. Enter SAS OPTIONS statements, LIBNAME statements, TITLE statements and other code that will affect the environment of the SAS code submitted by Enterprise Miner.

```
libname mydata '/home/stefaniegreay';
```

# Click on Actions>Add Node>Sample>Input Data



| Property | |
|---|---|
| Name | ANLT5050_U7A2 |
| Project Start Code | |
| Project Macro Variables | |
| Created | 3/5/21 4:25 PM |
| Server | SASApp - Logical Workspace Server |
| Grid Available | No |
| Path | ~/ANLT5050_U7A2 |
| Metadata Folder Path | /User Folders/stefaniegreay/My Folder/ANLT5050_ |
| Max. Concurrent Tasks | 2 |

**Project Start Code**

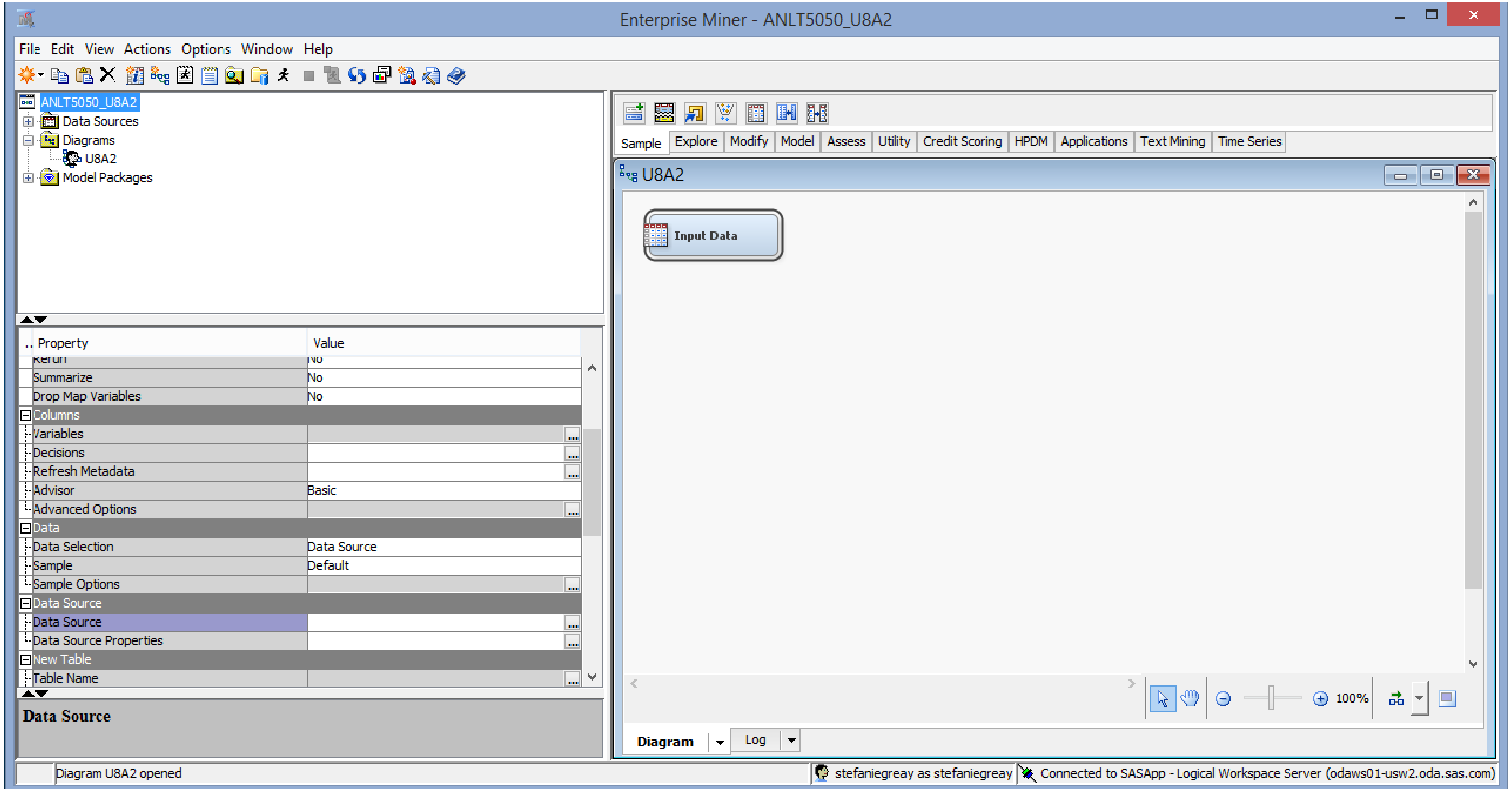
Code which is executed whenever this project is opened.

Diagram U7A2 opened

stefaniegreay as stefaniegreay

Connected to SASApp - Logical Workspace Server (odaws01-usw2.oda.sas.com)

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



| Property | Value |
|---|---|
| Rerun | No |
| Summarize | No |
| Drop Map Variables | No |
| Columns | |
| Variables | ... |
| Decisions | ... |
| Refresh Metadata | ... |
| Advisor | Basic |
| Advanced Options | ... |
| Data | |
| Data Selection | Data Source |
| Sample | Default |
| Sample Options | ... |
| Data Source | |
| Data Source | ... |
| Data Source Properties | ... |
| New Table | |
| Table Name | ... |

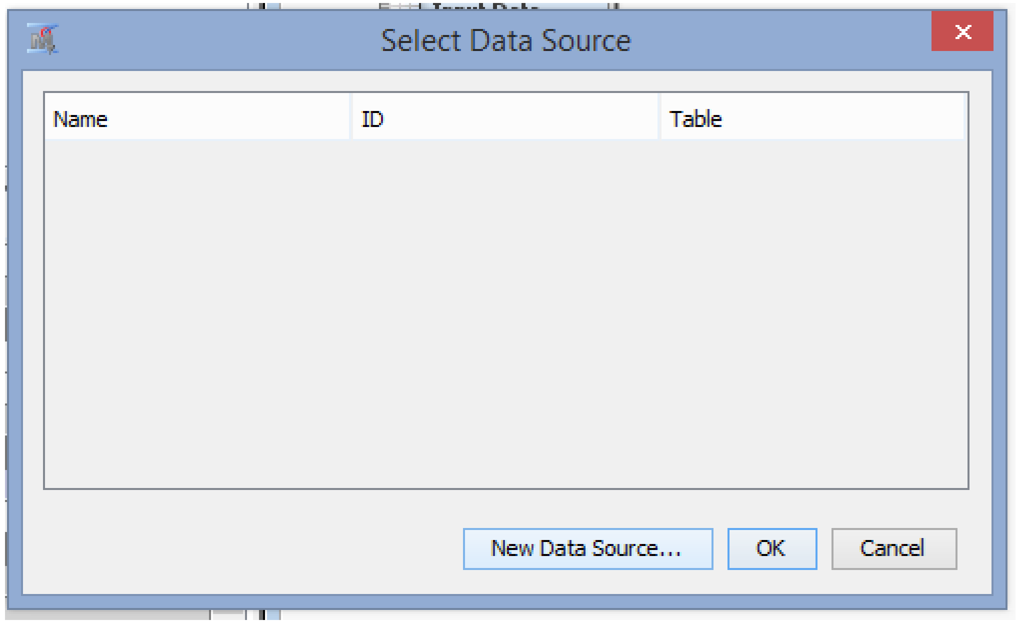
**Data Source**

Diagram U8A2 opened

stefaniegreay as stefaniegreay

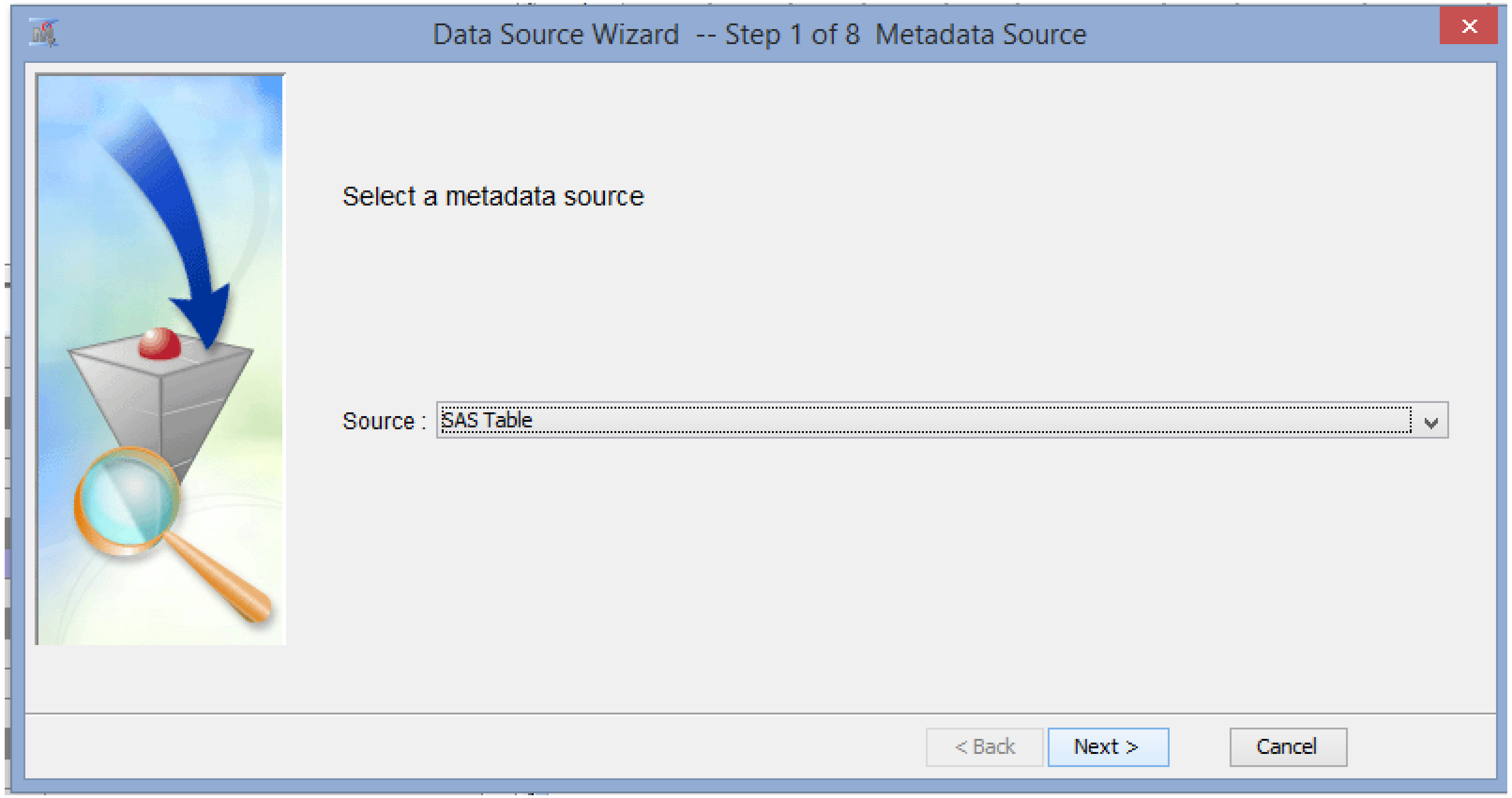
Connected to SASApp - Logical Workspace Server (odaws01-usw2.oda.sas.com)

# Click on “New Data Source”


Select Data Source

| Name | ID | Table |
|---|---|---|

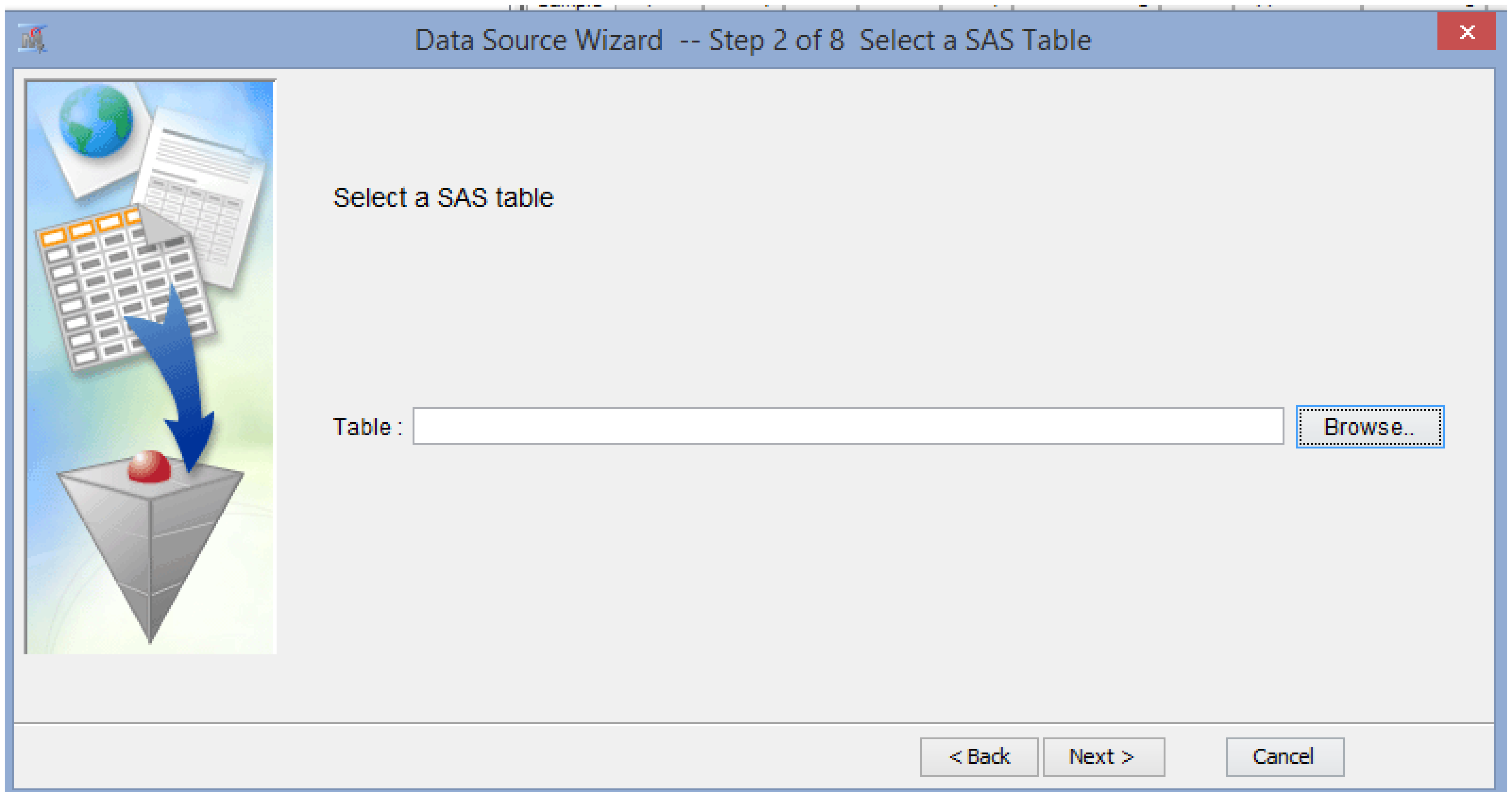
# Leave it as “SAS Table” and click “Next >”



Select a metadata source

Source : SAS Table

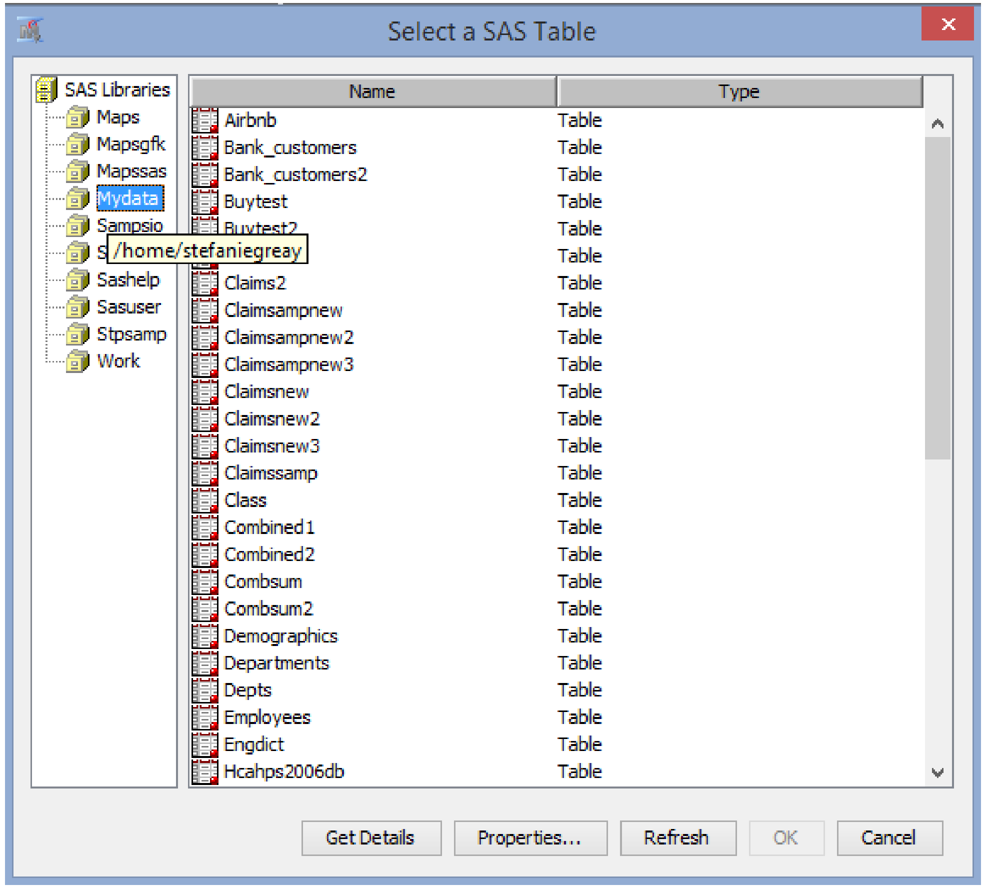
# Click on “Browse”



Select a SAS table

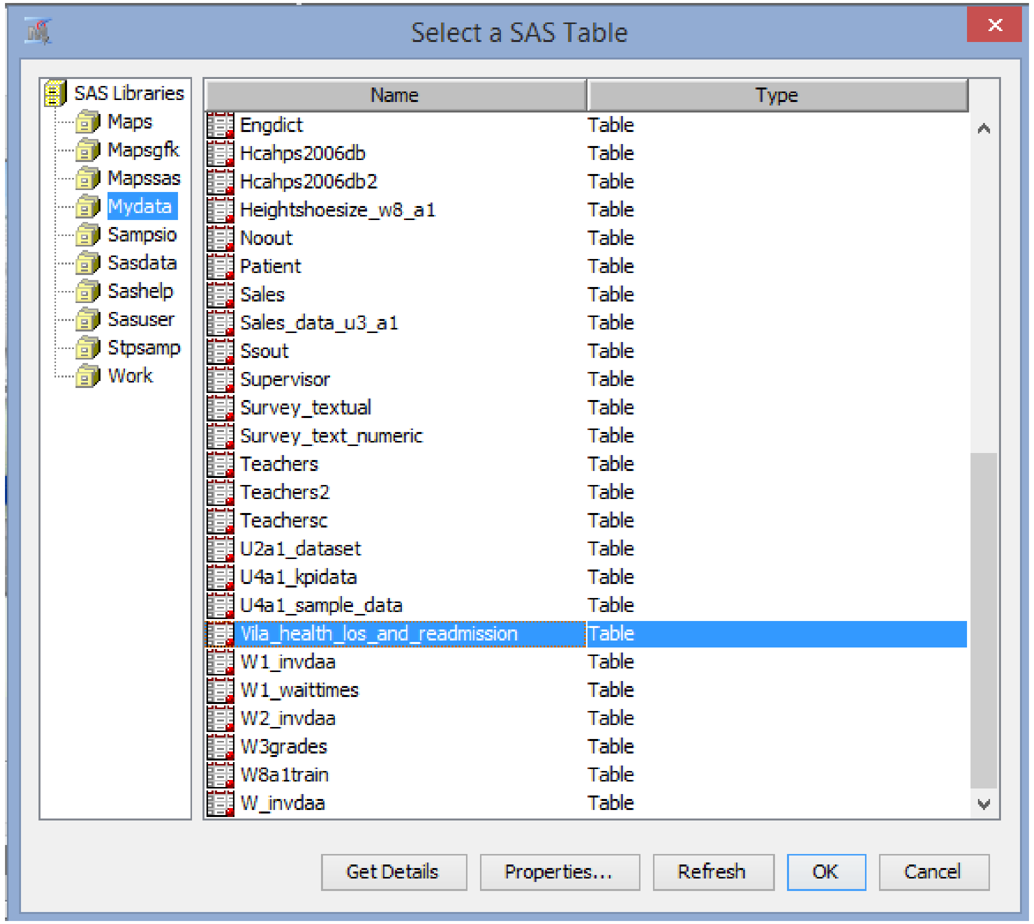
Table :

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



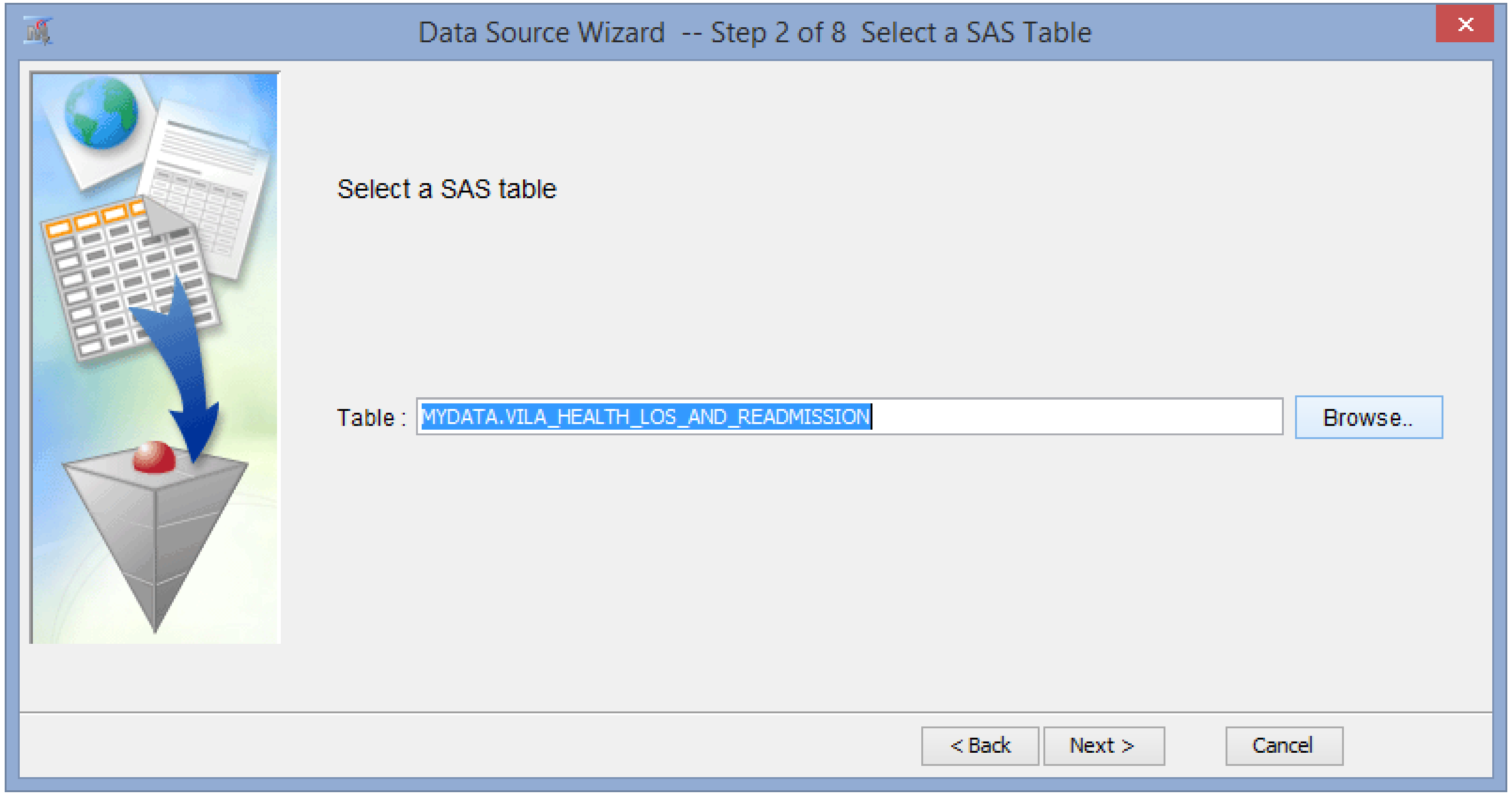
| Name | Type |
|---|---|
| Airbnb | Table |
| Bank_customers | Table |
| Bank_customers2 | Table |
| Buytest | Table |
| Buytest2 | Table |
| [illegible] | Table |
| Claims2 | Table |
| Claimsampnew | Table |
| Claimsampnew2 | Table |
| Claimsampnew3 | Table |
| Claimsnew | Table |
| Claimsnew2 | Table |
| Claimsnew3 | Table |
| Claimssamp | Table |
| Class | Table |
| Combined1 | Table |
| Combined2 | Table |
| Combsum | Table |
| Combsum2 | Table |
| Demographics | Table |
| Departments | Table |
| Depts | Table |
| Employees | Table |
| Engdict | Table |
| Hcahps2006db | Table |

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



| Name | Type |
|---|---|
| Engdict | Table |
| Hcahps2006db | Table |
| Hcahps2006db2 | Table |
| Heightshoesize_w8_a1 | Table |
| Noout | Table |
| Patient | Table |
| Sales | Table |
| Sales_data_u3_a1 | Table |
| Ssout | Table |
| Supervisor | Table |
| Survey_textual | Table |
| Survey_text_numeric | Table |
| Teachers | Table |
| Teachers2 | Table |
| Teachersc | Table |
| U2a1_dataset | Table |
| U4a1_kpidata | Table |
| U4a1_sample_data | Table |
| Vila_health_los_and_readmission | Table |
| W1_invdaa | Table |
| W1_waittimes | Table |
| W2_invdaa | Table |
| W3grades | Table |
| W8a1train | Table |
| W_invdaa | Table |

# Click “Next>”



Data Source Wizard -- Step 2 of 8 Select a SAS Table

Select a SAS table

Table : MYDATA.VILA_HEALTH_LOS_AND_READMISSION

# Verify the options and click “Next>”

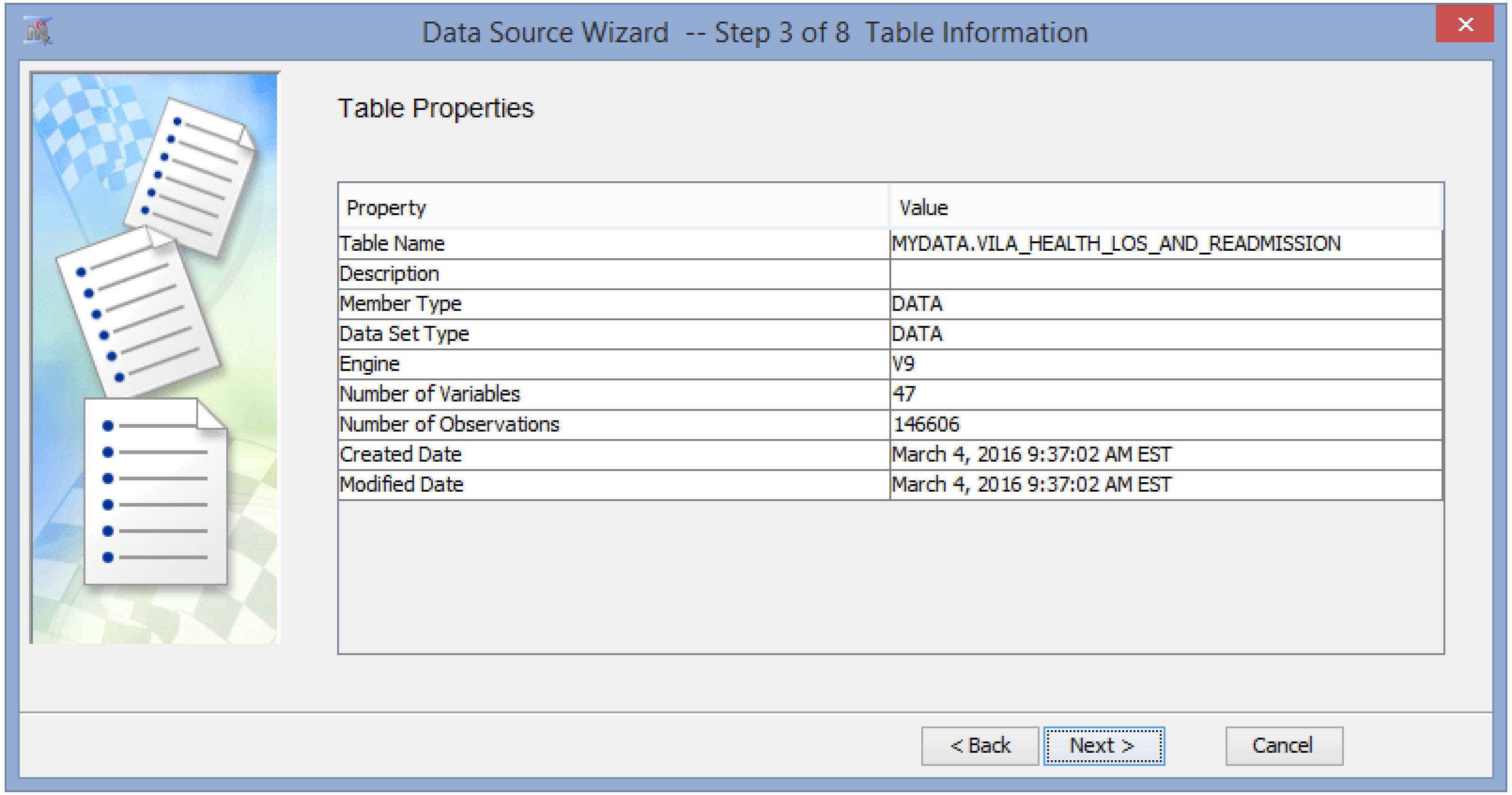
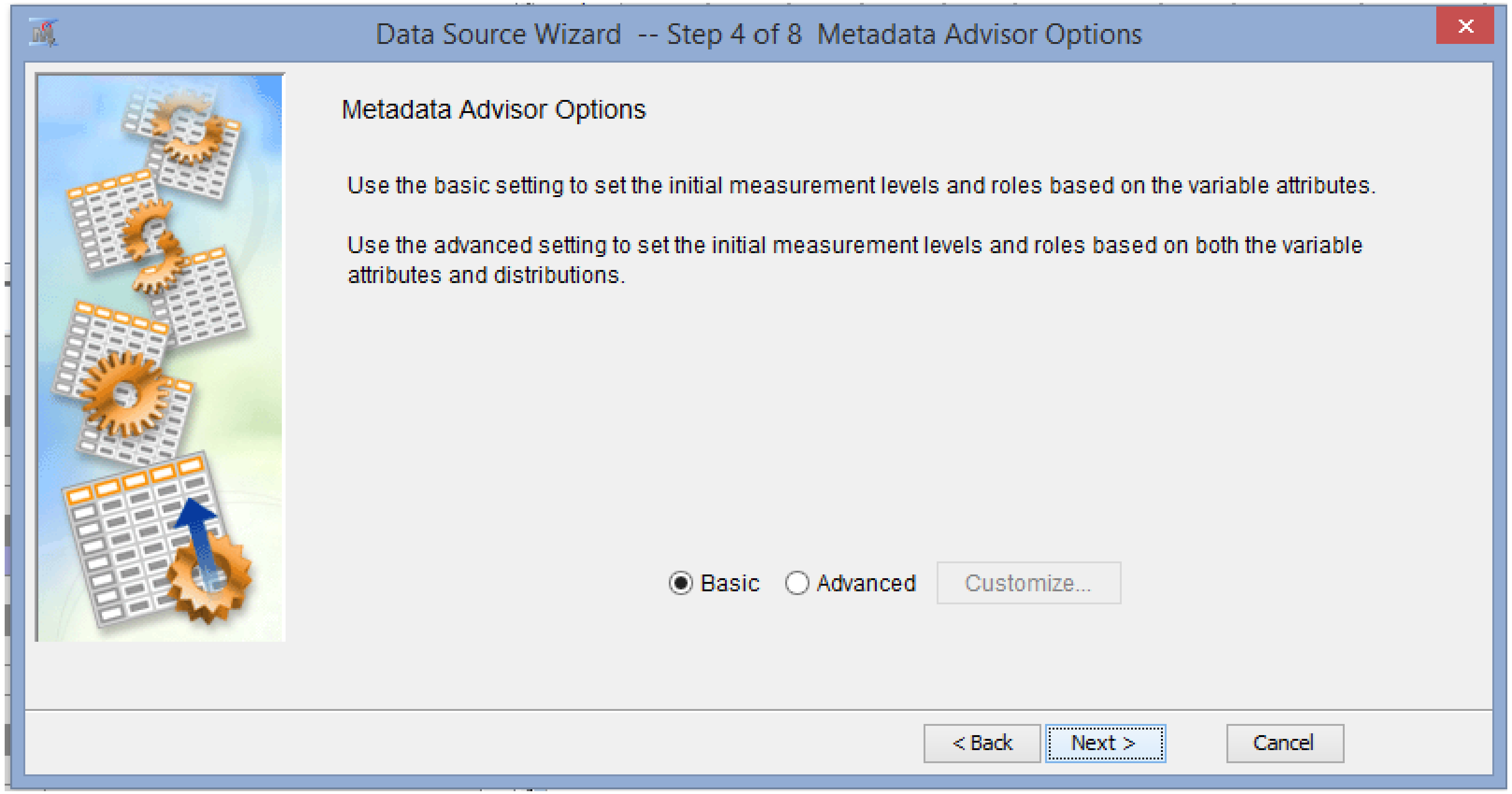


Table Properties

| Property | Value |
|---|---|
| Table Name | MYDATA.VILA_HEALTH_LOS_AND_READMISSION |
| Description | |
| Member Type | DATA |
| Data Set Type | DATA |
| Engine | V9 |
| Number of Variables | 47 |
| Number of Observations | 146606 |
| Created Date | March 4, 2016 9:37:02 AM EST |
| Modified Date | March 4, 2016 9:37:02 AM EST |

# Click “Next>”



Data Source Wizard -- Step 4 of 8 Metadata Advisor Options

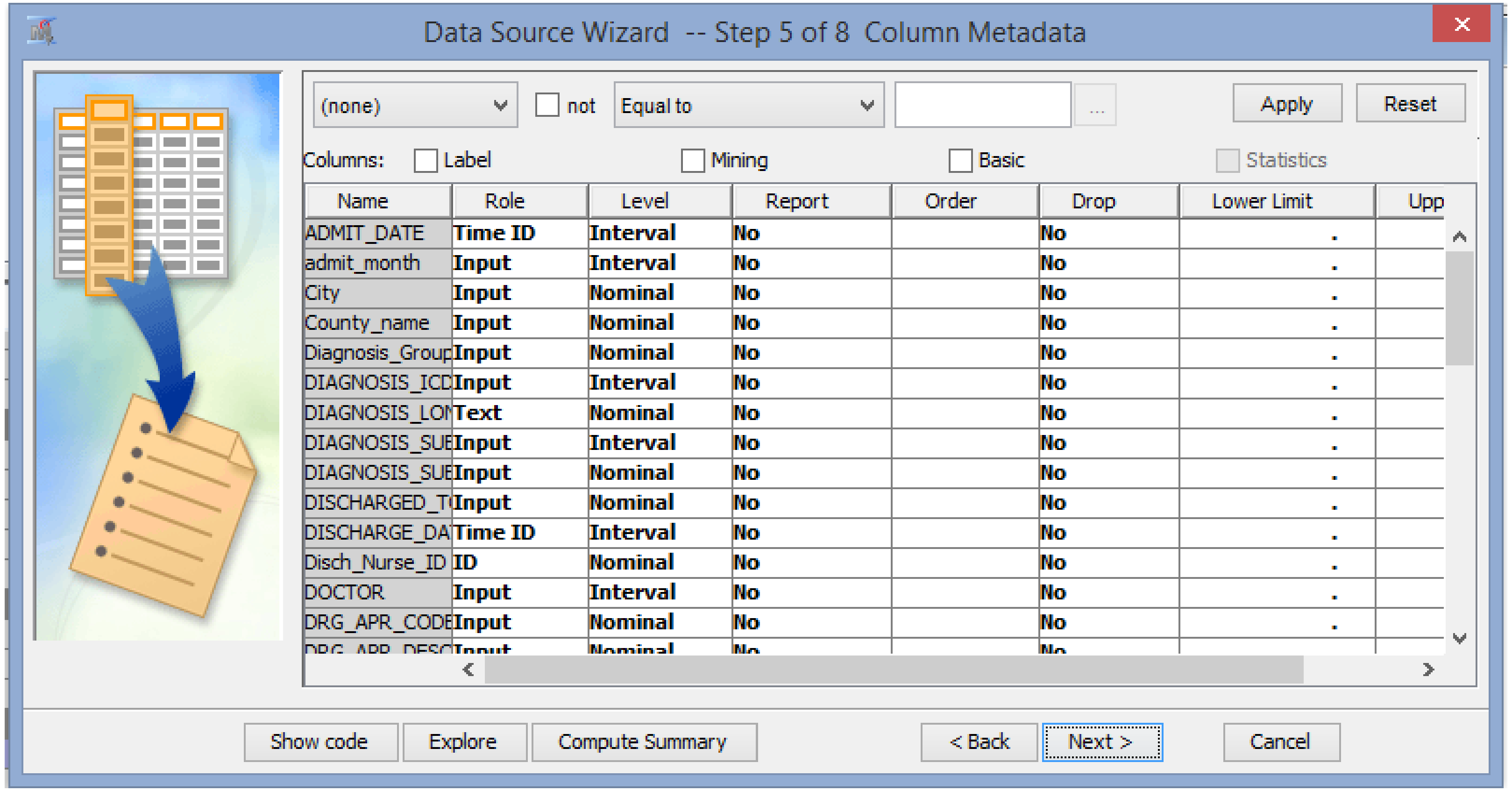
Metadata Advisor Options

Use the basic setting to set the initial measurement levels and roles based on the variable attributes.

Use the advanced setting to set the initial measurement levels and roles based on both the variable attributes and distributions.

● Basic ○ Advanced

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



Data Source Wizard -- Step 5 of 8 Column Metadata

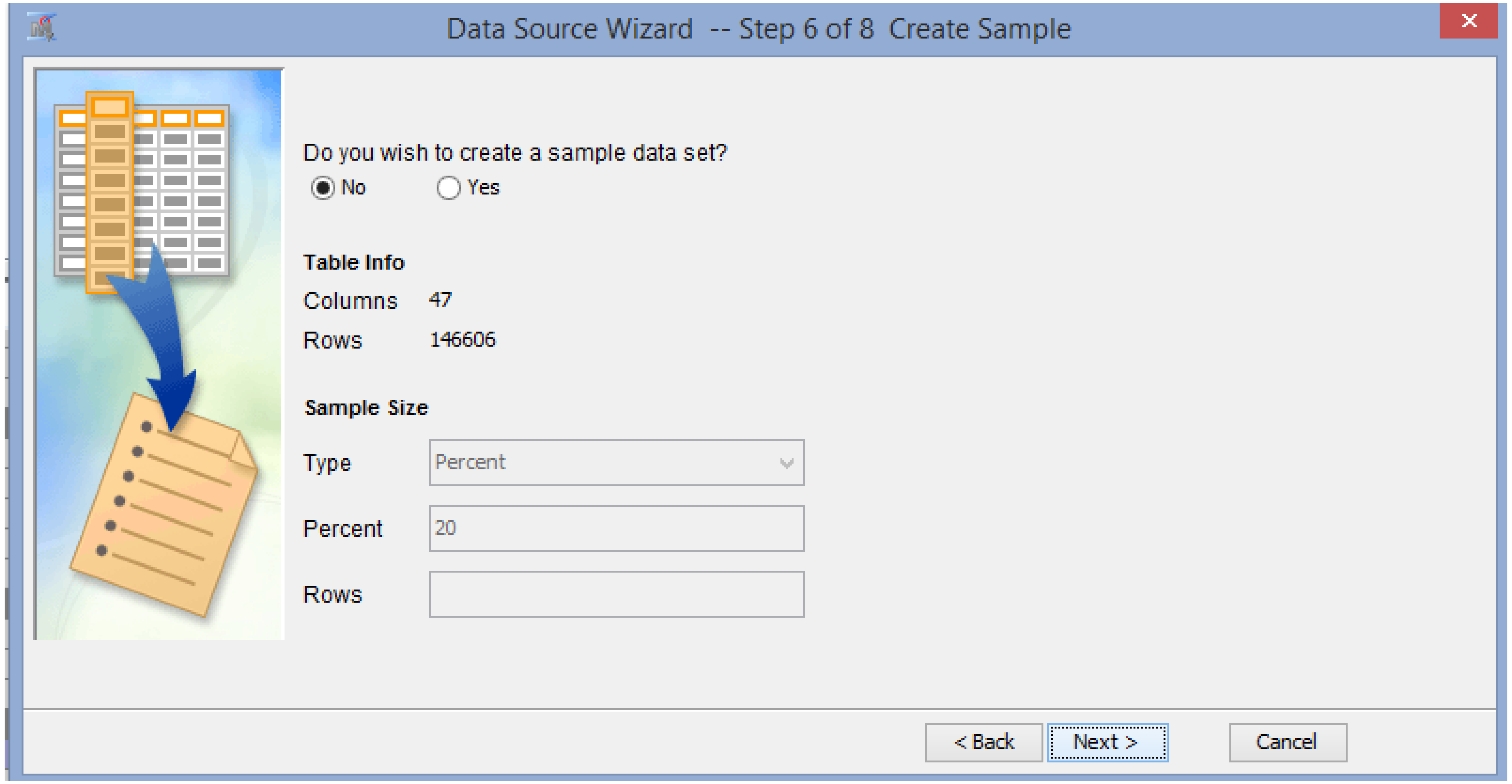
(none) ☐ not Equal to

Columns: ☐ Label ☐ Mining ☐ Basic ☐ Statistics

| Name | Role | Level | Report | Order | Drop | Lower Limit | Upp |
|---|---|---|---|---|---|---|---|
| ADMIT_DATE | Time ID | Interval | No |  | No | . |  |
| admit_month | Input | Interval | No |  | No | . |  |
| City | Input | Nominal | No |  | No | . |  |
| County_name | Input | Nominal | No |  | No | . |  |
| Diagnosis_Group | Input | Nominal | No |  | No | . |  |
| DIAGNOSIS_ICD | Input | Interval | No |  | No | . |  |
| DIAGNOSIS_LON | Text | Nominal | No |  | No | . |  |
| DIAGNOSIS_SUE | Input | Interval | No |  | No | . |  |
| DIAGNOSIS_SUE | Input | Nominal | No |  | No | . |  |
| DISCHARGED_T | Input | Nominal | No |  | No | . |  |
| DISCHARGE_DA | Time ID | Interval | No |  | No | . |  |
| Disch_Nurse_ID | ID | Nominal | No |  | No | . |  |
| DOCTOR | Input | Interval | No |  | No | . |  |
| DRG_APR_CODE | Input | Nominal | No |  | No | . |  |
| DRG_APR_DESC | Input | Nominal | No |  | No |  |  |

© Stefanie G. Reay, MS, PhD, Capella University, 2021

You may choose to sample the dataset here, or just keep the full dataset, then click “Next>.”  If you want to split into train, test, and validate, you could do this here.



Data Source Wizard -- Step 6 of 8 Create Sample

Do you wish to create a sample data set?

● No ○ Yes

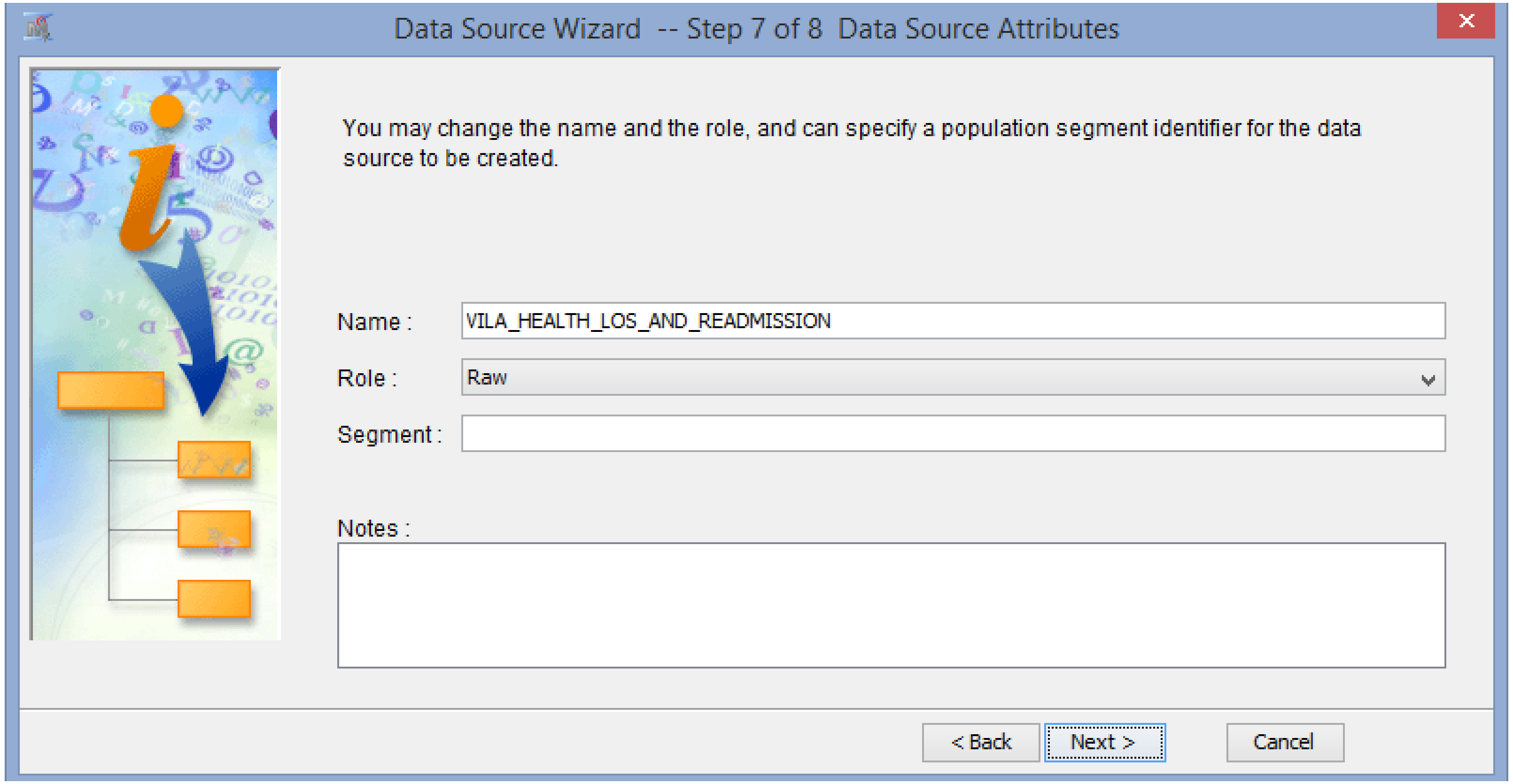
**Table Info**

| | |
|---|---|
| Columns | 47 |
| Rows | 146606 |

**Sample Size**

| | |
|---|---|
| Type | Percent |
| Percent | 20 |
| Rows | |

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



You may change the name and the role, and can specify a population segment identifier for the data source to be created.

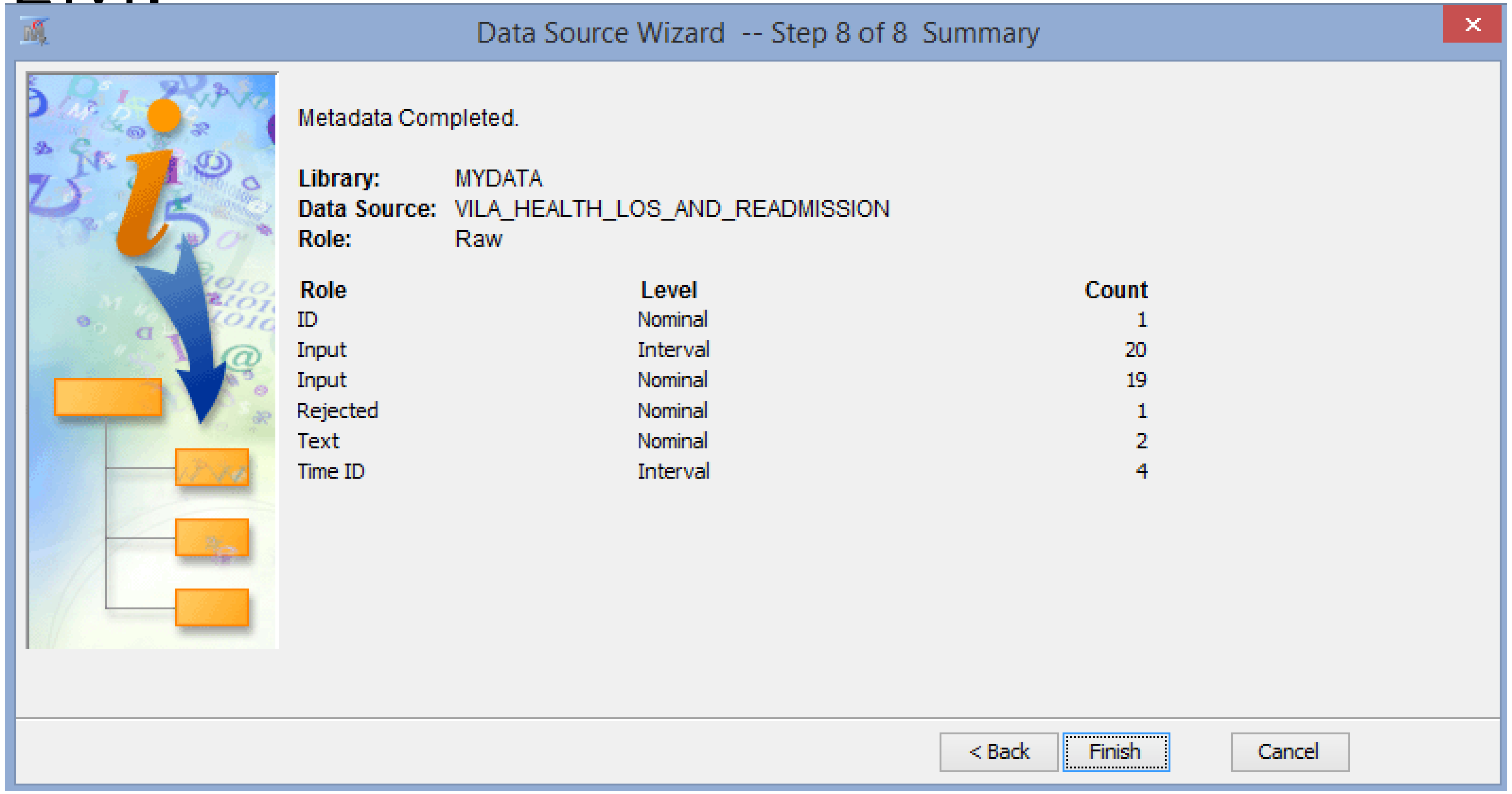
Name : VILA_HEALTH_LOS_AND_READMISSION

Role : Raw

Segment :

Notes :

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

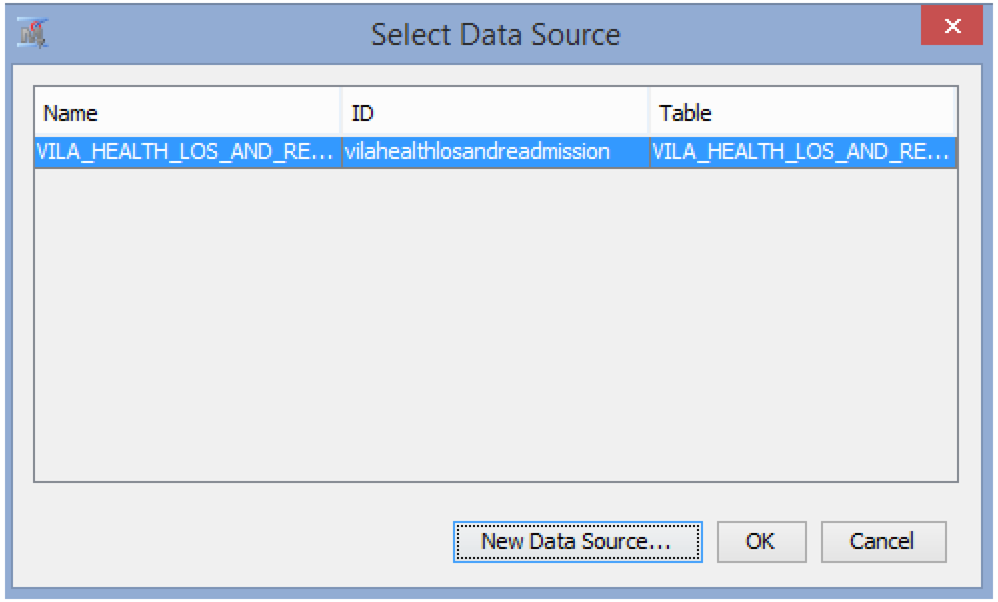


Metadata Completed.

**Library:** MYDATA
**Data Source:** VILA_HEALTH_LOS_AND_READMISSION
**Role:** Raw

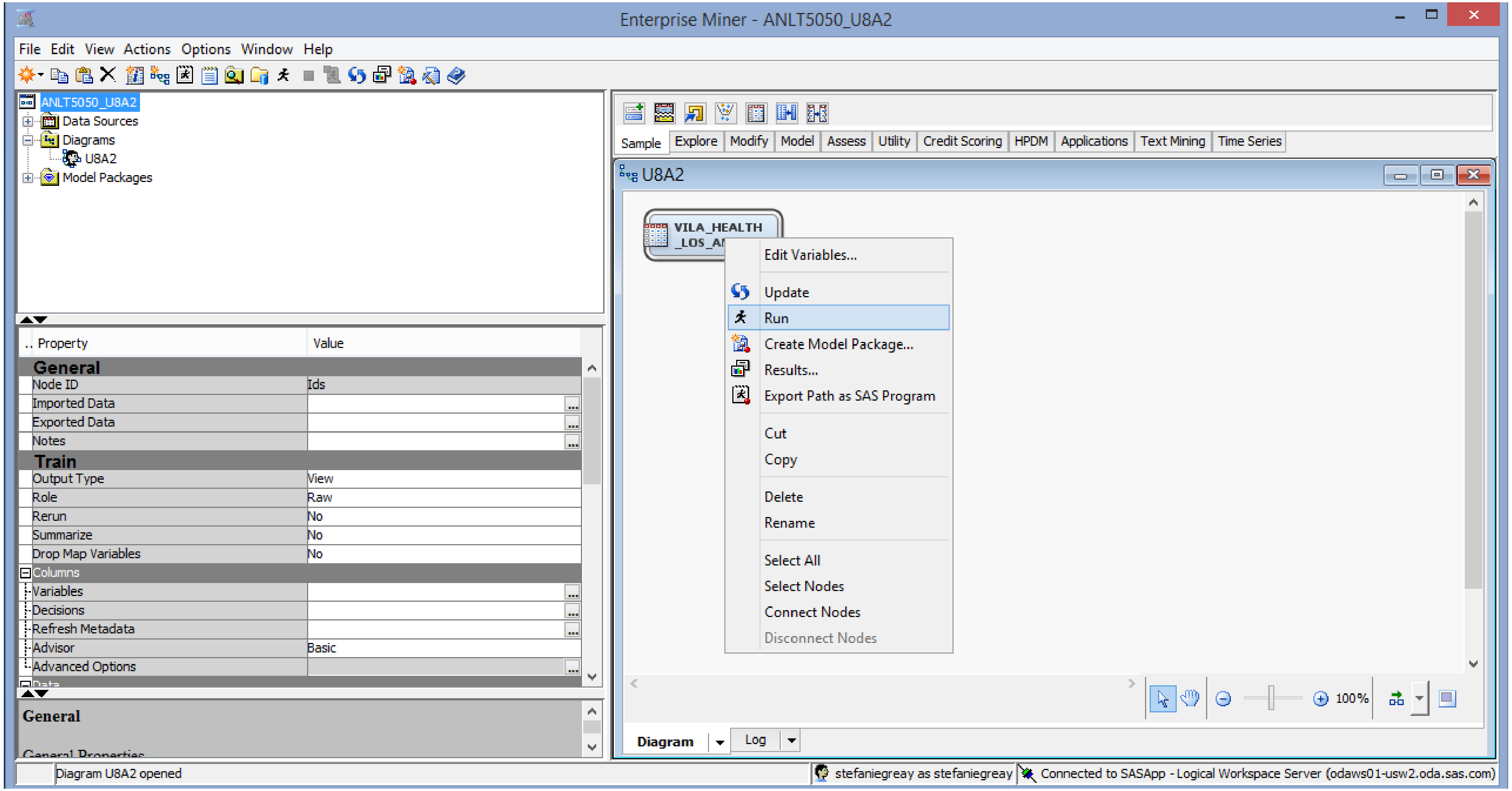
| Role | Level | Count |
|---|---|---|
| ID | Nominal | 1 |
| Input | Interval | 20 |
| Input | Nominal | 19 |
| Rejected | Nominal | 1 |
| Text | Nominal | 2 |
| Time ID | Interval | 4 |

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



| Name | ID | Table |
|---|---|---|
| VILA_HEALTH_LOS_AND_RE... | vilahealthlosandreadmission | VILA_HEALTH_LOS_AND_RE... |

# Right click on the dataset node and click "Run."



| Property | Value |
|---|---|
| **General** | |
| Node ID | Ids |
| Imported Data | ... |
| Exported Data | ... |
| Notes | ... |
| **Train** | |
| Output Type | View |
| Role | Raw |
| Rerun | No |
| Summarize | No |
| Drop Map Variables | No |
| Columns | |
| Variables | ... |
| Decisions | ... |
| Refresh Metadata | ... |
| Advisor | Basic |
| Advanced Options | ... |

General

General Properties

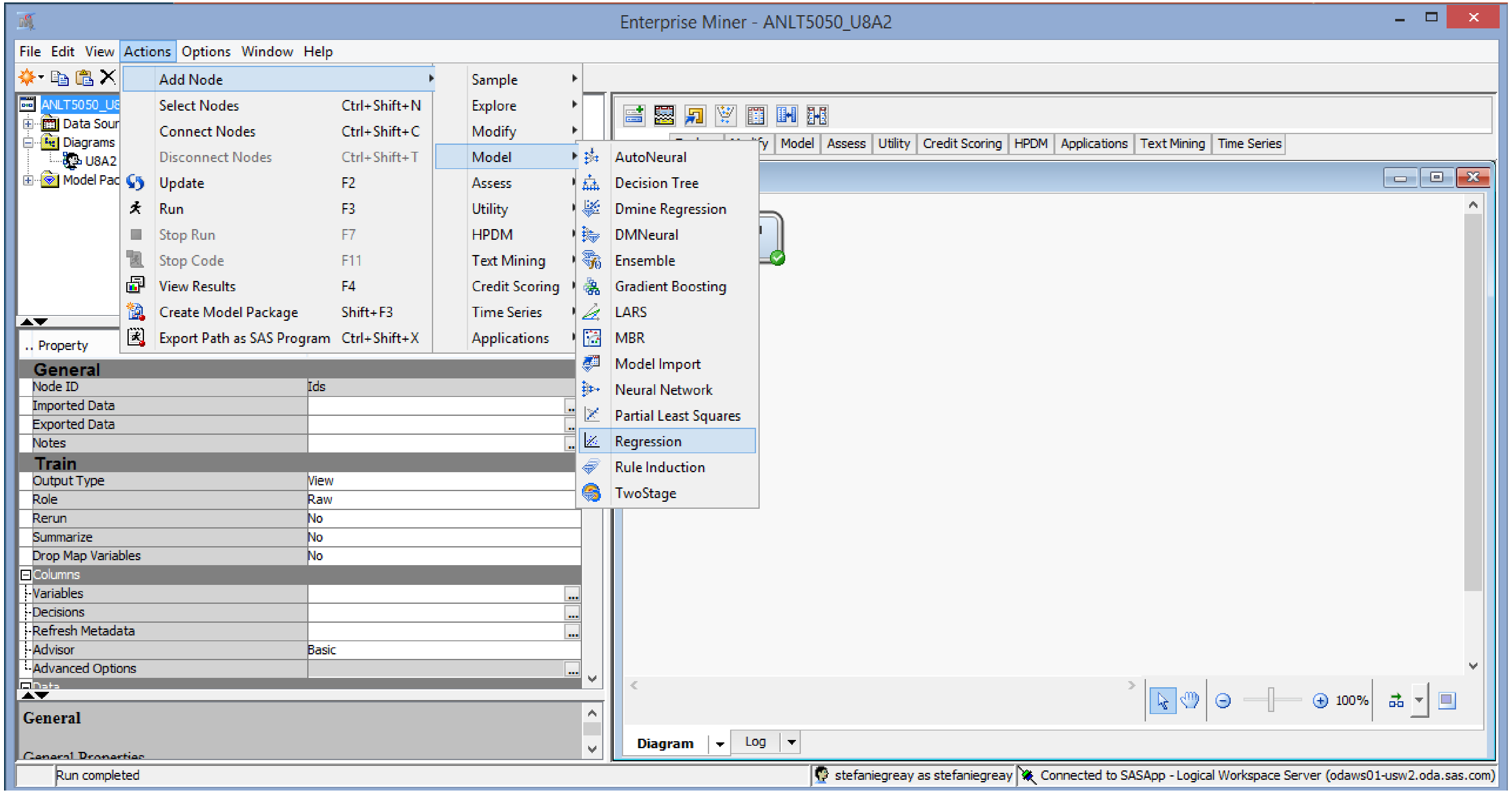
Diagram U8A2 opened

stefaniegreay as stefaniegreay

Connected to SASApp - Logical Workspace Server (odaws01-usw2.oda.sas.com)

© Stefanie G. Reay, MS, PhD, Capella University, 2021

# Click on “Actions” > “Add Node” > “Model” > “Regression”



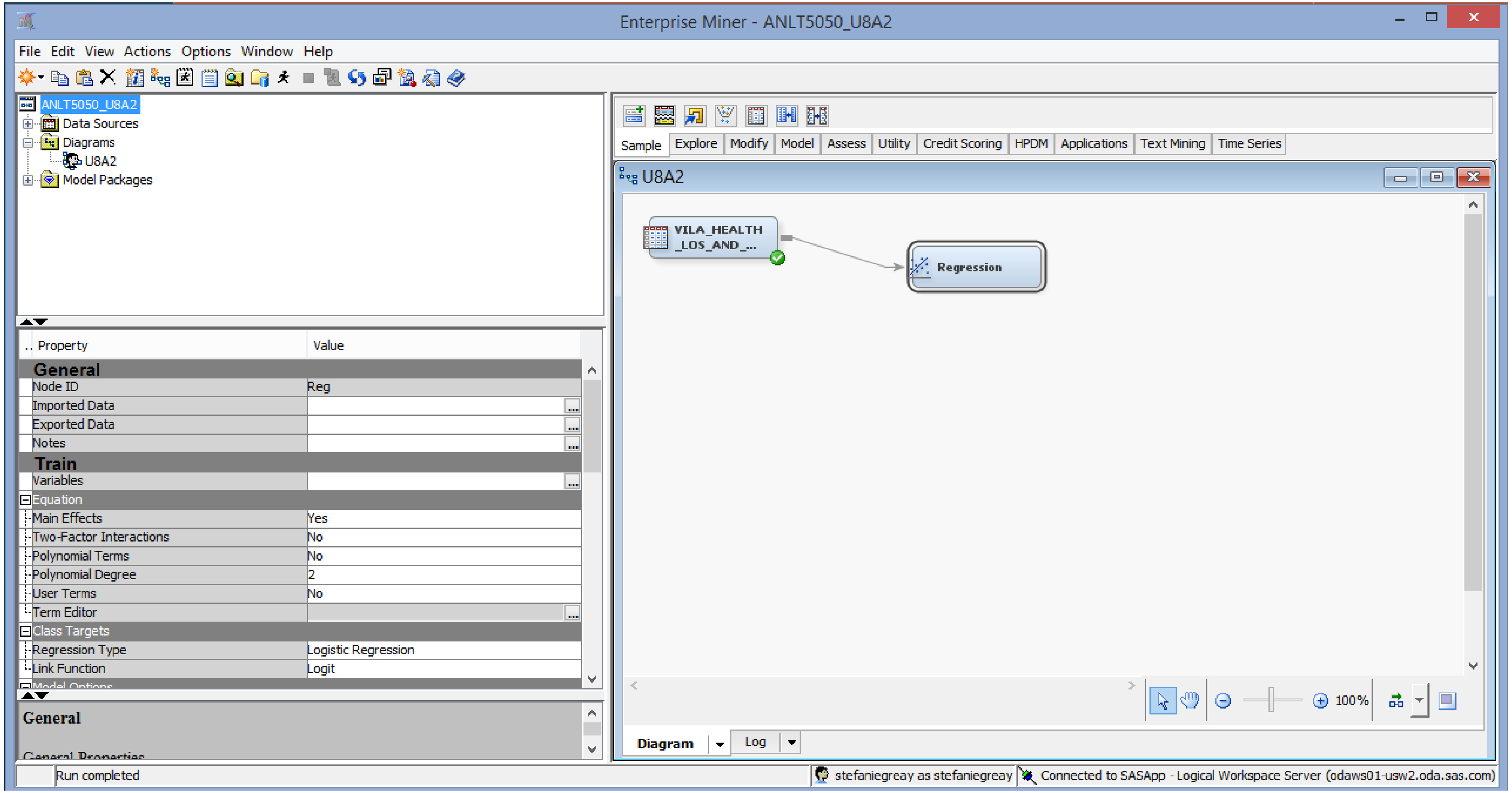
| Property | Value |
|---|---|
| **General** | |
| Node ID | Ids |
| Imported Data | |
| Exported Data | |
| Notes | |
| **Train** | |
| Output Type | View |
| Role | Raw |
| Rerun | No |
| Summarize | No |
| Drop Map Variables | No |
| Columns | |
| Variables | |
| Decisions | |
| Refresh Metadata | |
| Advisor | Basic |
| Advanced Options | |

Run completed

stefaniegreay as stefaniegreay

Connected to SASApp - Logical Workspace Server (odaws01-usw2.oda.sas.com)

# Connect the nodes



| Property | Value |
|---|---|
| **General** | |
| Node ID | Reg |
| Imported Data | ... |
| Exported Data | ... |
| Notes | ... |
| **Train** | |
| Variables | ... |
| Equation | |
| Main Effects | Yes |
| Two-Factor Interactions | No |
| Polynomial Terms | No |
| Polynomial Degree | 2 |
| User Terms | No |
| Term Editor | ... |
| Class Targets | |
| Regression Type | Logistic Regression |
| Link Function | Logit |
| Model Options | |

**General**

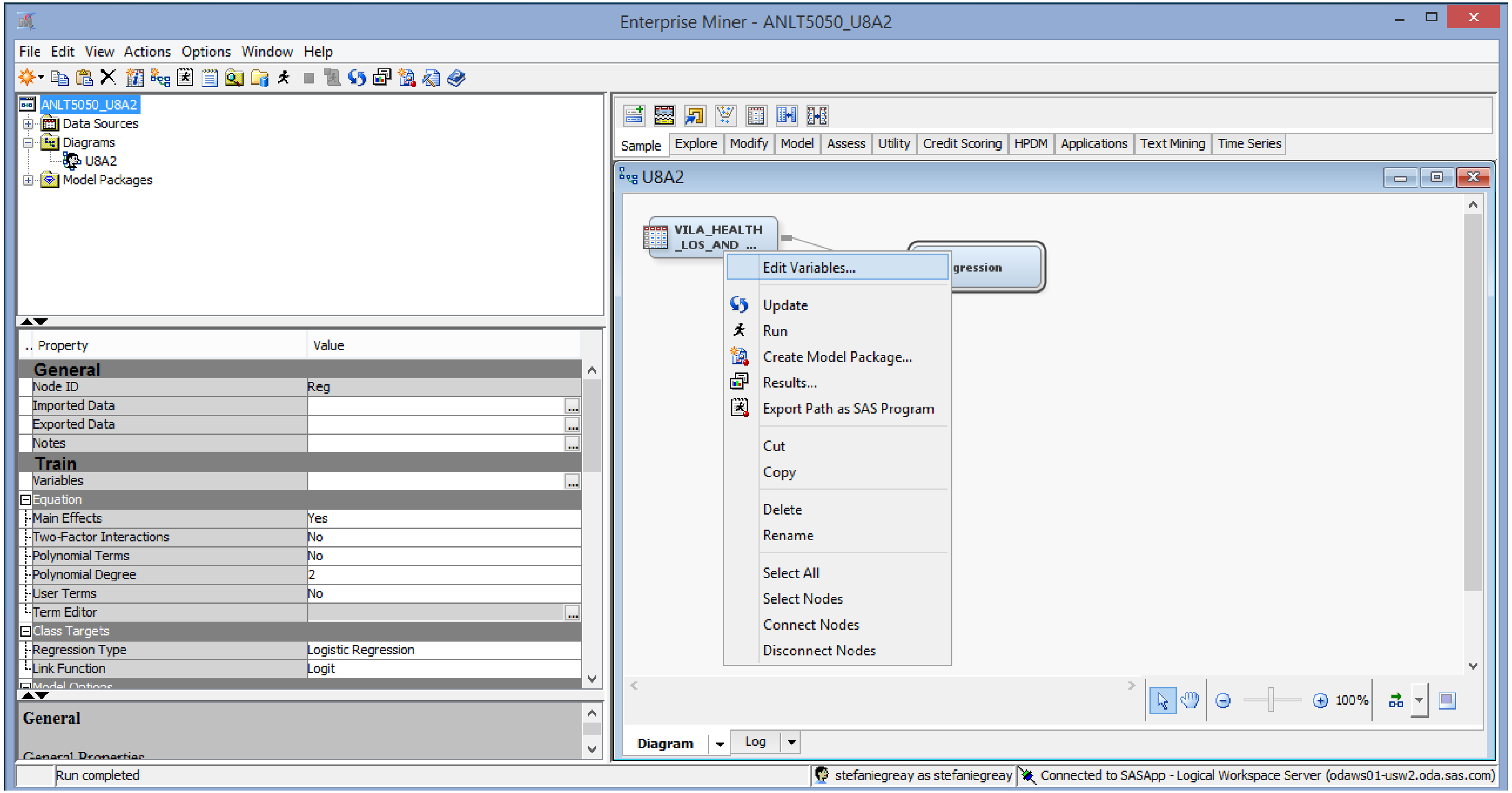
General Properties

Run completed

stefaniegreay as stefaniegreay

Connected to SASApp - Logical Workspace Server (odaws01-usw2.oda.sas.com)

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



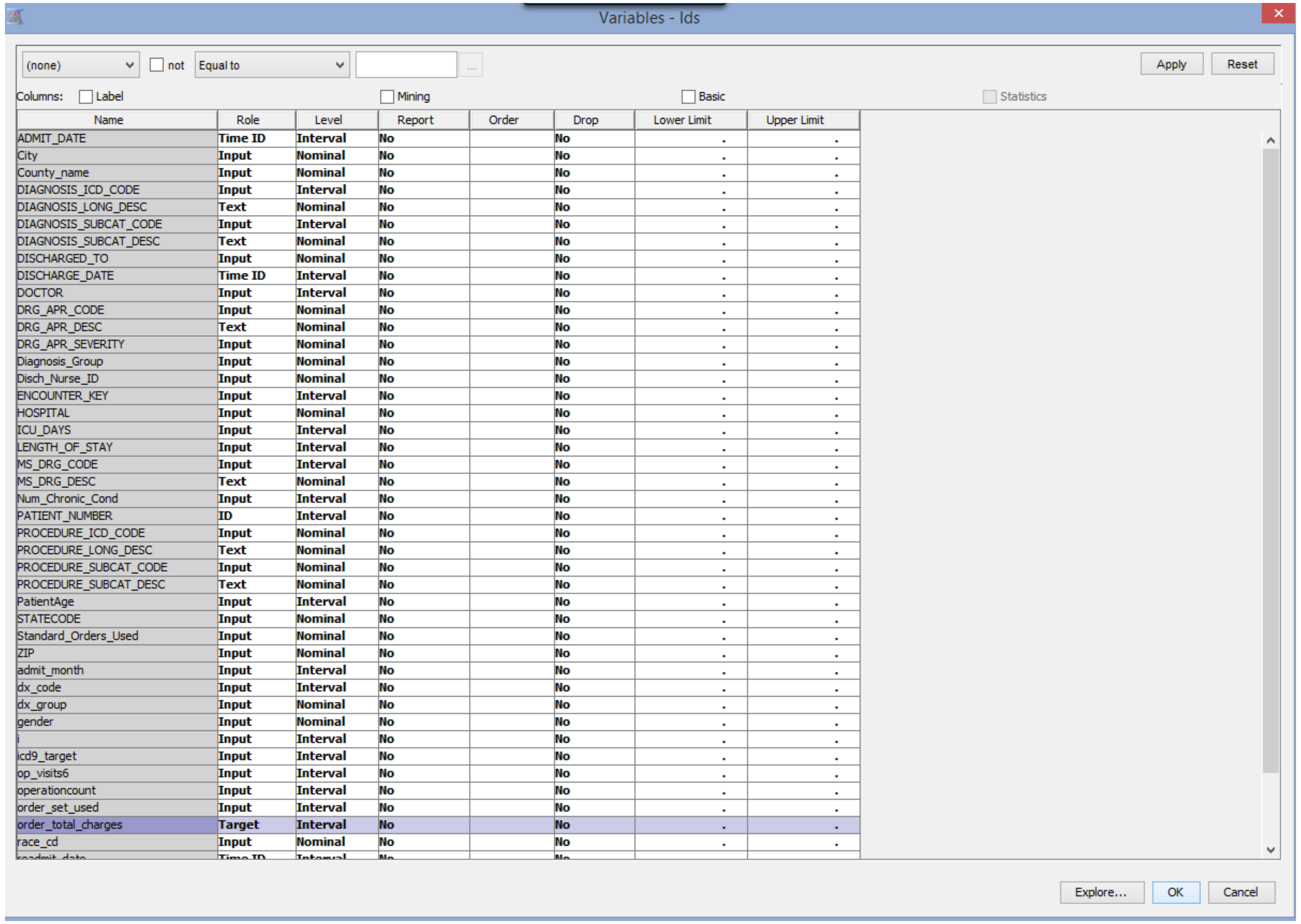
| Property | Value |
|---|---|
| **General** | |
| Node ID | Reg |
| Imported Data | ... |
| Exported Data | ... |
| Notes | ... |
| **Train** | |
| Variables | ... |
| Equation | |
| Main Effects | Yes |
| Two-Factor Interactions | No |
| Polynomial Terms | No |
| Polynomial Degree | 2 |
| User Terms | No |
| Term Editor | ... |
| Class Targets | |
| Regression Type | Logistic Regression |
| Link Function | Logit |
| Model Options | |

Run completed

stefaniegreay as stefaniegreay

Connected to SASApp - Logical Workspace Server (odaws01-usw2.oda.sas.com)

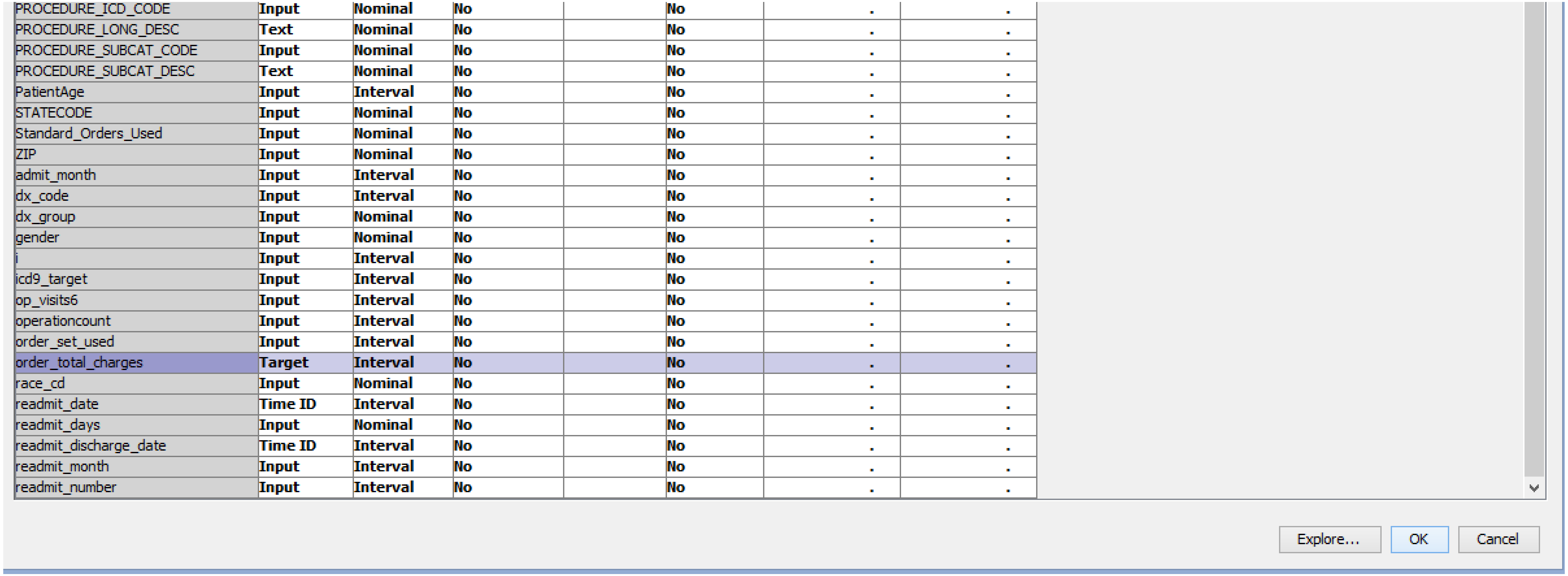
# Change the description variables to “text,” the Patient_ID variable to “ID,” the dates as “Time_ID” variables, the order_total_charges variable to “Target,” and all others to “Input”



| Name | Role | Level | Report | Order | Drop | Lower Limit | Upper Limit |
|---|---|---|---|---|---|---|---|
| ADMIT_DATE | Time ID | 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 | Time ID | 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 | Target | Interval | No | | No | . | . |
| race_cd | Input | Nominal | No | | No | . | . |
| readmit_date | Time ID | Interval | No | | No | | |

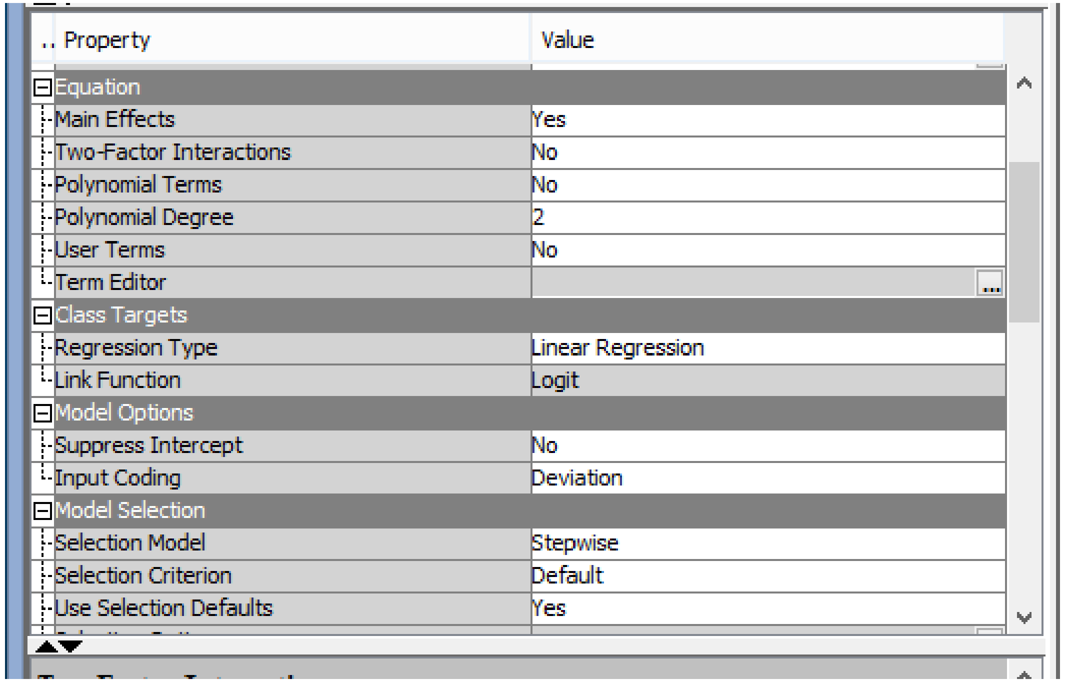
© Stefanie G. Reay, MS, PhD, Capella University, 2021

# 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 | Target | 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 | . | . |

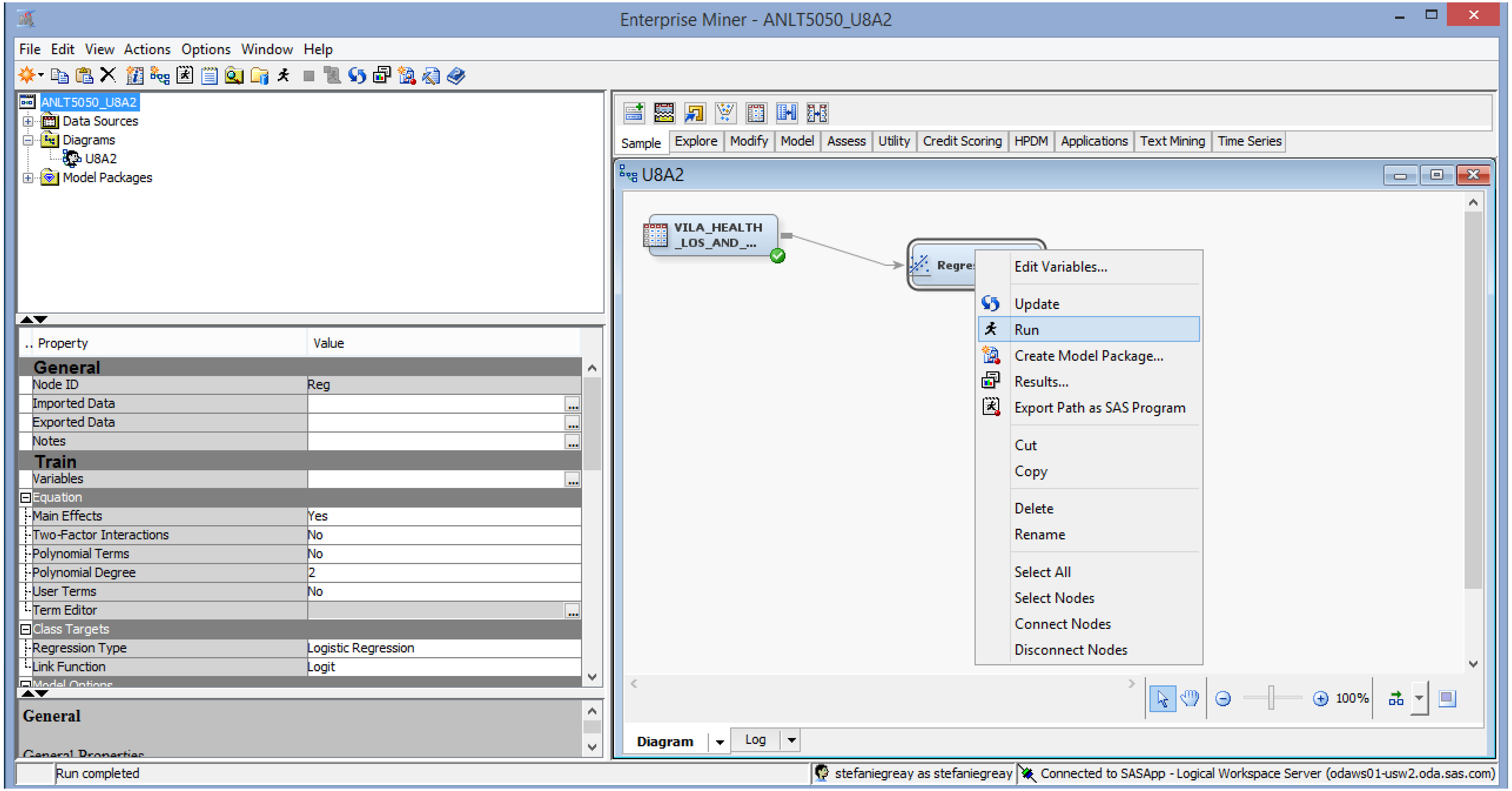
Change the regression type by right clicking on the Regression node, and scrolling down in the properties, and selecting “linear regression” as the regression type. If you wish to include interaction effects (pairwise only), you can set the Two-Factor Interactions to “Yes”. You may also opt for a variable selection method here, like Stepwise, Forward, or Backward, instead of only creating a model that contains all variables specified. Note that if you choose Two-Factor Interactions, with this many variables, you will need to edit the terms using the Term Editor to reduce them before running the node.



| Property | Value |
|---|---|
| ⊟ Equation | |
| Main Effects | Yes |
| Two-Factor Interactions | No |
| Polynomial Terms | No |
| Polynomial Degree | 2 |
| User Terms | No |
| Term Editor | ... |
| ⊟ Class Targets | |
| Regression Type | Linear Regression |
| Link Function | Logit |
| ⊟ Model Options | |
| Suppress Intercept | No |
| Input Coding | Deviation |
| ⊟ Model Selection | |
| Selection Model | Stepwise |
| Selection Criterion | Default |
| Use Selection Defaults | Yes |

© Stefanie G. Reay, MS, PhD, Capella University, 2021

# Right click on “Regression” node and click “Run”



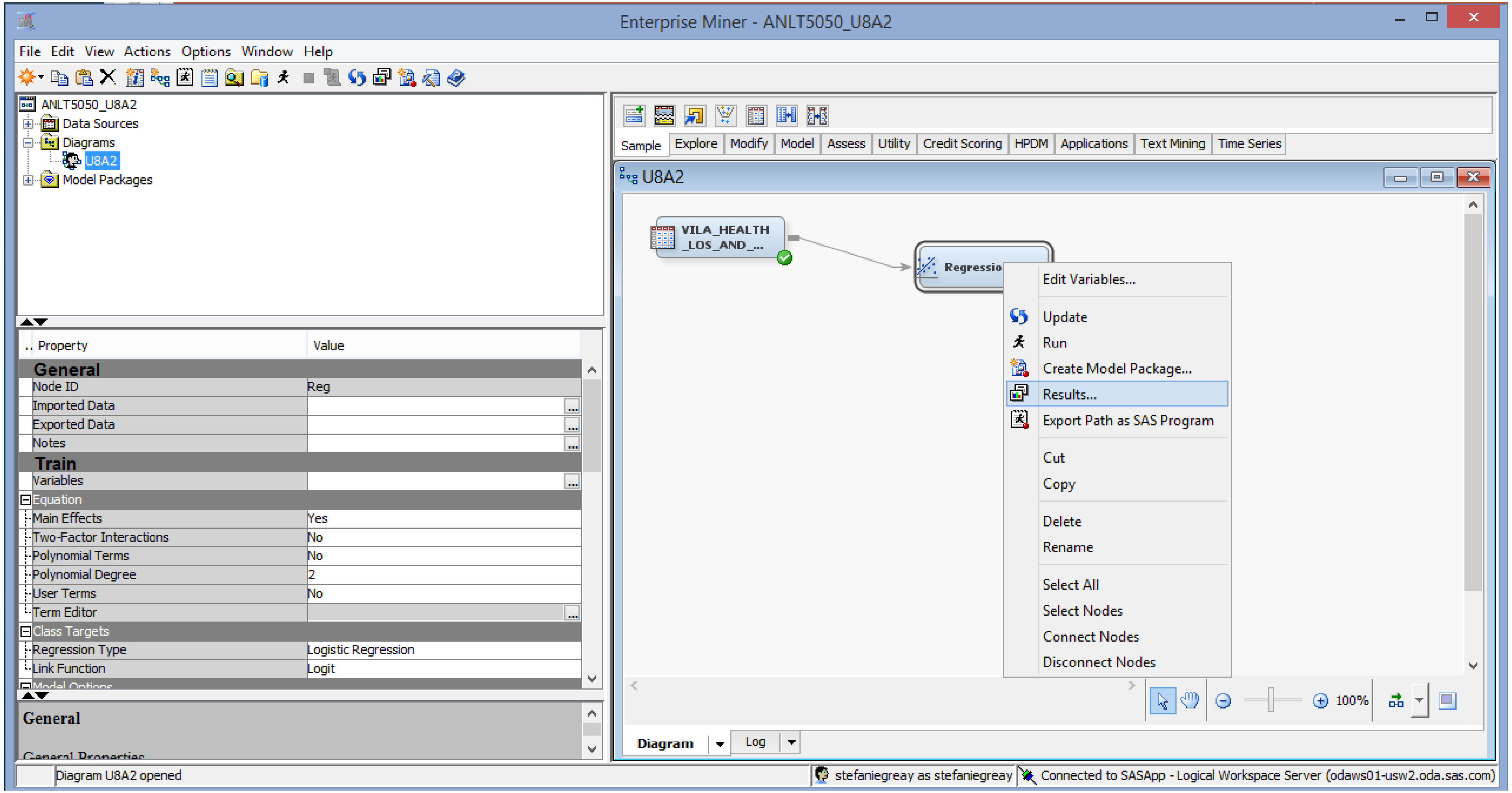
| Property | Value |
|---|---|
| **General** | |
| Node ID | Reg |
| Imported Data | ... |
| Exported Data | ... |
| Notes | ... |
| **Train** | |
| Variables | ... |
| Equation | |
| Main Effects | Yes |
| Two-Factor Interactions | No |
| Polynomial Terms | No |
| Polynomial Degree | 2 |
| User Terms | No |
| Term Editor | ... |
| Class Targets | |
| Regression Type | Logistic Regression |
| Link Function | Logit |
| Model Options | |

Run completed

stefaniegreay as stefaniegreay

Connected to SASApp - Logical Workspace Server (odaws01-usw2.oda.sas.com)

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



| Property | Value |
|---|---|
| **General** | |
| Node ID | Reg |
| Imported Data | ... |
| Exported Data | ... |
| Notes | ... |
| **Train** | |
| Variables | ... |
| ⊟Equation | |
| Main Effects | Yes |
| Two-Factor Interactions | No |
| Polynomial Terms | No |
| Polynomial Degree | 2 |
| User Terms | No |
| Term Editor | ... |
| ⊟Class Targets | |
| Regression Type | Logistic Regression |
| Link Function | Logit |
| ⊟Model Options | |

General

General Properties

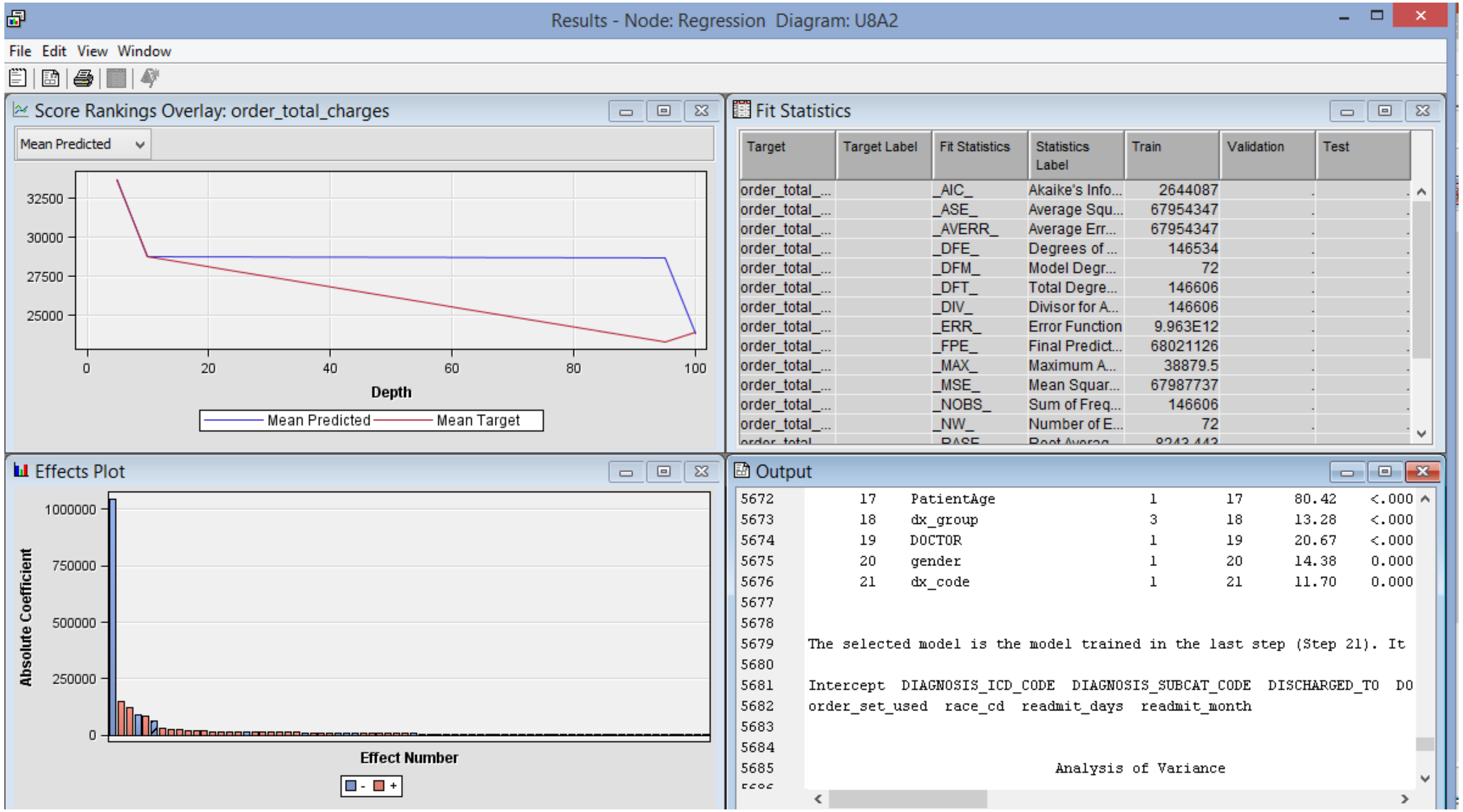
Diagram U8A2 opened | stefaniegreay as stefaniegreay | Connected to SASApp - Logical Workspace Server (odaws01-usw2.oda.sas.com)

© Stefanie G. Reay, MS, PhD, Capella University, 2021

# Considerations for fitting regression models

- The options for variable selection include: forward, backward, and stepwise.
- The process of fitting a regression model is iterative, and should not be a “set it and forget it”-type of analysis.
- SAS Enterprise Miner’s regression node is a good start for analyzing several variables, but additional analysis and evaluation of assumptions should be completed outside of the regression node.

# After you re-run the regression node (after making any adjustments to the settings), right click on the node and select "Results."



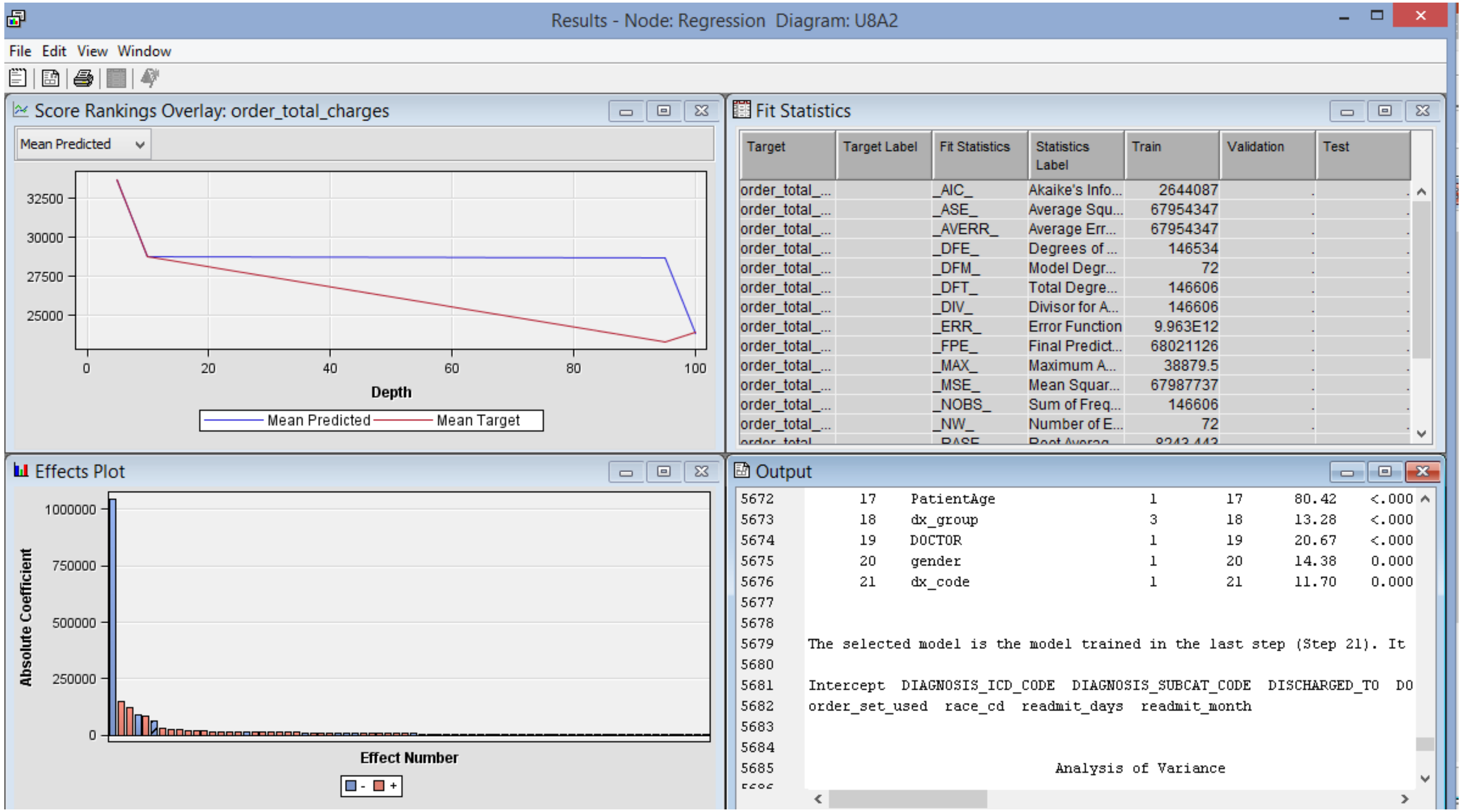
Fit Statistics

| Target | Target Label | Fit Statistics | Statistics Label | Train | Validation | Test |
|---|---|---|---|---|---|---|
| order_total_... | | _AIC_ | Akaike's Info... | 2644087 | . | . |
| order_total_... | | _ASE_ | Average Squ... | 67954347 | . | . |
| order_total_... | | _AVERR_ | Average Err... | 67954347 | . | . |
| order_total_... | | _DFE_ | Degrees of ... | 146534 | . | . |
| order_total_... | | _DFM_ | Model Degr... | 72 | . | . |
| order_total_... | | _DFT_ | Total Degre... | 146606 | . | . |
| order_total_... | | _DIV_ | Divisor for A... | 146606 | . | . |
| order_total_... | | _ERR_ | Error Function | 9.963E12 | . | . |
| order_total_... | | _FPE_ | Final Predict... | 68021126 | . | . |
| order_total_... | | _MAX_ | Maximum A... | 38879.5 | . | . |
| order_total_... | | _MSE_ | Mean Squar... | 67987737 | . | . |
| order_total_... | | _NOBS_ | Sum of Freq... | 146606 | . | . |
| order_total_... | | _NW_ | Number of E... | 72 | . | . |
| order_total_... | | _RASE_ | Root Averag... | 8243.443 | | |

Output

```
         17     PatientAge                      1          17         80.42     <.000
         18     dx_group                        3          18         13.28     <.000
         19     DOCTOR                          1          19         20.67     <.000
         20     gender                          1          20         14.38     0.000
         21     dx_code                         1          21         11.70     0.000


The selected model is the model trained in the last step (Step 21). It

Intercept  DIAGNOSIS_ICD_CODE  DIAGNOSIS_SUBCAT_CODE  DISCHARGED_TO  DO
order_set_used  race_cd  readmit_days  readmit_month


                                   Analysis of Variance
```

The Fit Statistics window shows the fit statistics for the last iteration.  If you scroll down to the bottom of the Output pane, the last iteration is shown, including the variables and the tests for contributions, etc.

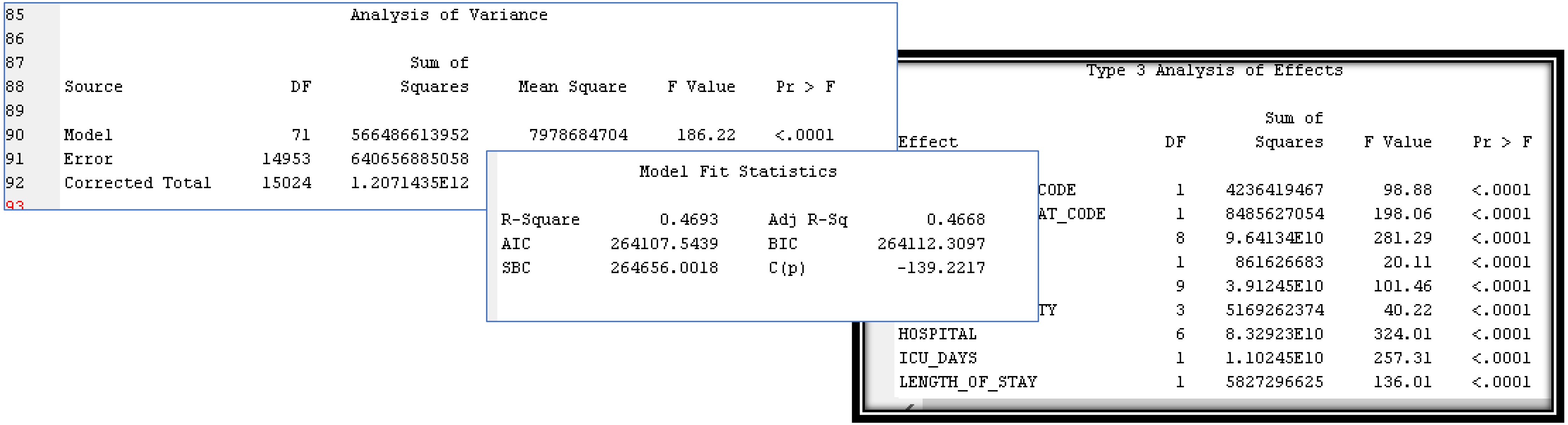


| Target | Target Label | Fit Statistics | Statistics Label | Train | Validation | Test |
|---|---|---|---|---|---|---|
| order_total_... | | _AIC_ | Akaike's Info... | 2644087 | . | . |
| order_total_... | | _ASE_ | Average Squ... | 67954347 | . | . |
| order_total_... | | _AVERR_ | Average Err... | 67954347 | . | . |
| order_total_... | | _DFE_ | Degrees of ... | 146534 | . | . |
| order_total_... | | _DFM_ | Model Degr... | 72 | . | . |
| order_total_... | | _DFT_ | Total Degre... | 146606 | . | . |
| order_total_... | | _DIV_ | Divisor for A... | 146606 | . | . |
| order_total_... | | _ERR_ | Error Function | 9.963E12 | . | . |
| order_total_... | | _FPE_ | Final Predict... | 68021126 | . | . |
| order_total_... | | _MAX_ | Maximum A... | 38879.5 | . | . |
| order_total_... | | _MSE_ | Mean Squar... | 67987737 | . | . |
| order_total_... | | _NOBS_ | Sum of Freq... | 146606 | . | . |
| order_total_... | | _NW_ | Number of E... | 72 | . | . |
| order_total_... | | _RASE_ | Root Averag... | 8243.443 | | |

```
     17    PatientAge                1        17       80.42     <.000
     18    dx_group                  3        18       13.28     <.000
     19    DOCTOR                    1        19       20.67     <.000
     20    gender                    1        20       14.38     0.000
     21    dx_code                   1        21       11.70     0.000


The selected model is the model trained in the last step (Step 21). It

Intercept  DIAGNOSIS_ICD_CODE  DIAGNOSIS_SUBCAT_CODE  DISCHARGED_TO  DO
order_set_used  race_cd  readmit_days  readmit_month


                                Analysis of Variance
```

© Stefanie G. Reay, MS, PhD, Capella University, 2021

The Analysis of Variance table provides the F-test statistic and p-value to test for significance in the overall model's ability to predict the response variable.  The Model Fit Statistics table provides a variety of model fit statistics to evaluate the fit of the model specified.  The Type 3 Analysis of Effects provides F-test statistics and p-values for testing the contribution of each individual variable.

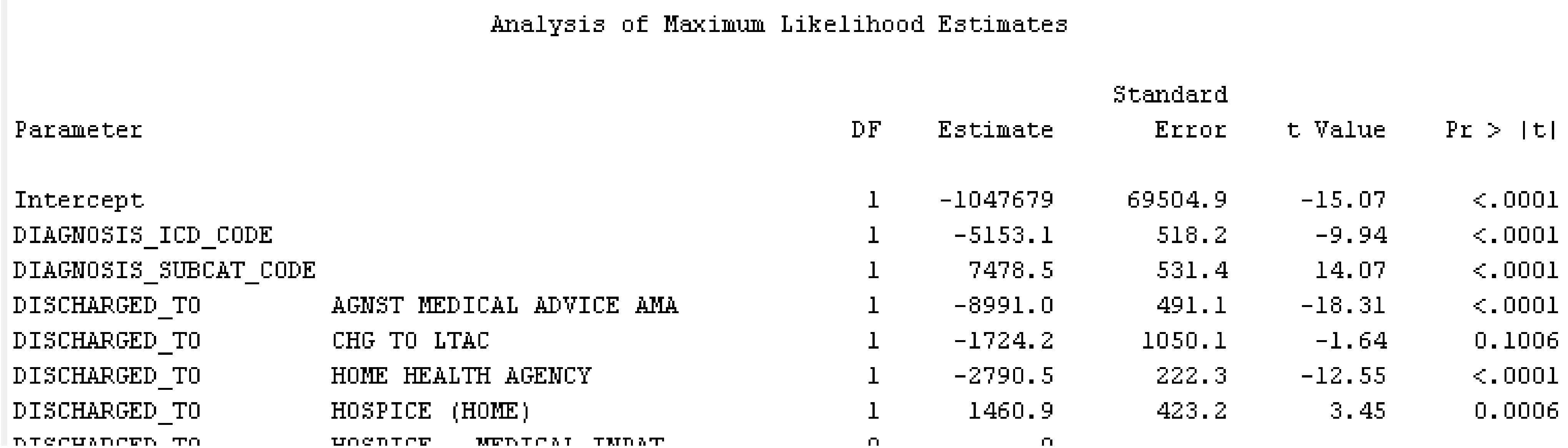


```
                         Analysis of Variance

                                   Sum of
Source                DF          Squares     Mean Square    F Value    Pr > F

Model                 71     566486613952      7978684704     186.22    <.0001
Error              14953     640656885058
Corrected Total    15024     1.2071435E12
```

```
                      Model Fit Statistics

R-Square            0.4693     Adj R-Sq          0.4668
AIC            264107.5439     BIC          264112.3097
SBC            264656.0018     C(p)           -139.2217
```

```
                     Type 3 Analysis of Effects

                                       Sum of
Effect                       DF       Squares    F Value    Pr > F

CODE                          1    4236419467      98.88    <.0001
AT_CODE                       1    8485627054     198.06    <.0001
                              8    9.64134E10     281.29    <.0001
                              1     861626683      20.11    <.0001
                              9    3.91245E10     101.46    <.0001
TY                            3    5169262374      40.22    <.0001
HOSPITAL                      6    8.32923E10     324.01    <.0001
ICU_DAYS                      1    1.10245E10     257.31    <.0001
LENGTH_OF_STAY                1    5827296625     136.01    <.0001
```

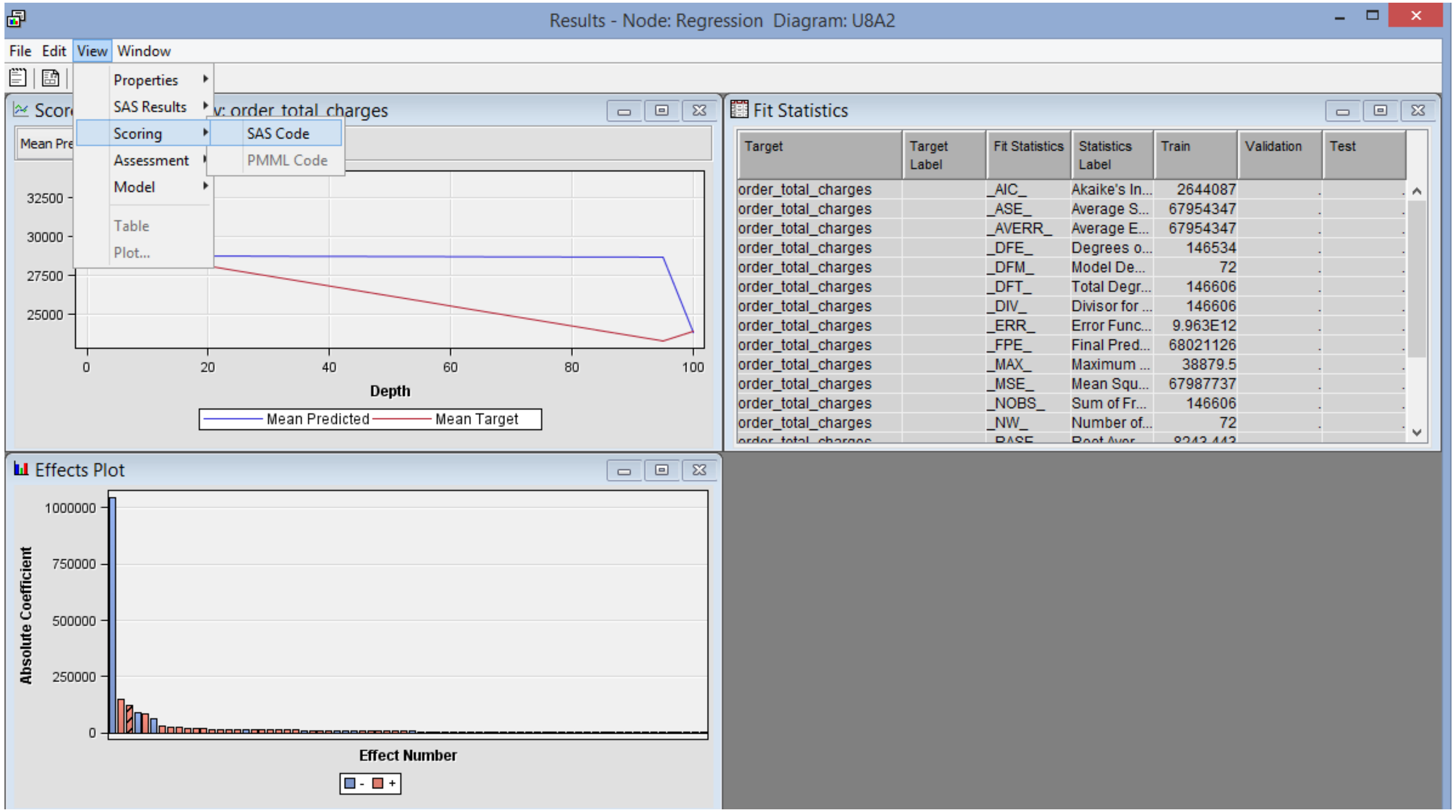
The Analysis of Maximum Likelihood Estimates provides the estimate for the coefficient for a particular variable (and the intercept), as well as the standard error, t-test statistic and p-value for the t-test for significant contribution to the model. Notice that categorical variables display as individual indicators in this table (as they are treated as separate indicator/dummy variables within the model itself…this conversion is automatic within the regression node). ALL variables should be analyzed and reported on, not just the top few.



```
                                   Analysis of Maximum Likelihood Estimates


                                                                                     Standard
Parameter                                                        DF     Estimate        Error    t Value    Pr > |t|

Intercept                                                         1     -1047679      69504.9     -15.07      <.0001
DIAGNOSIS_ICD_CODE                                                1      -5153.1        518.2      -9.94      <.0001
DIAGNOSIS_SUBCAT_CODE                                             1       7478.5        531.4      14.07      <.0001
DISCHARGED_TO             AGNST MEDICAL ADVICE AMA                1      -8991.0        491.1     -18.31      <.0001
DISCHARGED_TO             CHG TO LTAC                             1      -1724.2       1050.1      -1.64      0.1006
DISCHARGED_TO             HOME HEALTH AGENCY                      1      -2790.5        222.3     -12.55      <.0001
DISCHARGED_TO             HOSPICE (HOME)                          1       1460.9        423.2       3.45      0.0006
DISCHARGED_TO             HOSPICE - MEDICAL INPAT                 0            0
```

© Stefanie G. Reay, MS, PhD, Capella University, 2021

To get the SAS code for scoring, you can click on “View” > “Scoring”>”SAS Code.” This can be run against a test and validation dataset to test and then validate the model. (For example, if you split the dataset into 3 originally, for training, testing and validation, or on new data that comes in.)



| Target | Target Label | Fit Statistics | Statistics Label | Train | Validation | Test |
|---|---|---|---|---|---|---|
| order_total_charges | | _AIC_ | Akaike's In... | 2644087 | . | . |
| order_total_charges | | _ASE_ | Average S... | 67954347 | . | . |
| order_total_charges | | _AVERR_ | Average E... | 67954347 | . | . |
| order_total_charges | | _DFE_ | Degrees o... | 146534 | . | . |
| order_total_charges | | _DFM_ | Model De... | 72 | . | . |
| order_total_charges | | _DFT_ | Total Degr... | 146606 | . | . |
| order_total_charges | | _DIV_ | Divisor for ... | 146606 | . | . |
| order_total_charges | | _ERR_ | Error Func... | 9.963E12 | . | . |
| order_total_charges | | _FPE_ | Final Pred... | 68021126 | . | . |
| order_total_charges | | _MAX_ | Maximum ... | 38879.5 | . | . |
| order_total_charges | | _MSE_ | Mean Squ... | 67987737 | . | . |
| order_total_charges | | _NOBS_ | Sum of Fr... | 146606 | . | . |
| order_total_charges | | _NW_ | Number of... | 72 | . | . |
| order_total_charges | | _RASE_ | Root Aver... | 8243.443 | | |

# During and after the iterations, the following assumptions of regression must be validated

- There is a linear relationship between the dependent/response and independent/explanatory variables
- Multivariate normality
- No multicollinearity between the independent/explanatory variables
- No auto-correlation
- Homoscedasticity (homogeneity of variance of the residuals)

# Assumption: Linear relationship

- Linear relationships between the response/dependent and explanatory/independent variables can be checked using:
  1. A linear correlation matrix
  2. A scatterplot (or scatterplot matrix)
  3. And should also be checked for outliers using a histogram and box plot of each individual variable

# Assumption: Multivariate Normal

- Multivariate normality assumes that the residuals are normally distributed. This can be checked using all of the following:
  1. A normal probability (or normal Q-Q) plot of the residuals
  2. A histogram of the residuals
  3. A box plot of the residuals
  4. A Komogorv-Smirnov test of the residuals

# Assumption: No multicollinearity between the independent/explanatory variables

- No multicollinearity between the independent/explanatory variables can be checked using all of the following:
  1. A linear correlation coefficient matrix (including correlations between explanatory/independent variables
  2. A scatterplot matrix showing relationships between the explanatory/independent variables
  3. Tolerance
  4. Variance Inflation Factor (VIF)
  5. Condition Index

# Assumption: No autocorrelation

- No autocorrelation can be checked using all of the following:
  1. Durbin-Watson’s d test

# Assumption: Homoscedasticity (homogeneity of variance of the residuals)

- Homoscedasticity (homogeneity of variance of the residuals) can be checked using all of the following:
  1. Scatterplot of residuals vs predicted values
  2. Scatterplot of residuals vs response/dependent variable(s)
  3. Scatterplots of residuals vs each of the explanatory/independent variable(s)

  You will want to check for patterns. Patterns of any kind suggest a lack of random spread, or, in essence, a lack of homogeneity of variance within the residuals.

# SAS Documentation Reference

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

https://documentation.sas.com/?docsetId=emref&docsetTarget=n1jqzz8cssr9m2n1ktx2iyv87q56.htm&docsetVersion=14.3&locale=en