Data Preparation with Claim Fraud Dataset/ SAS Enterprise Miner

Didem B. Aykurt

Colorado State University Global

MIS530; Predictive Analytics

Dr.Jennifer Catalano

January 29, 2023

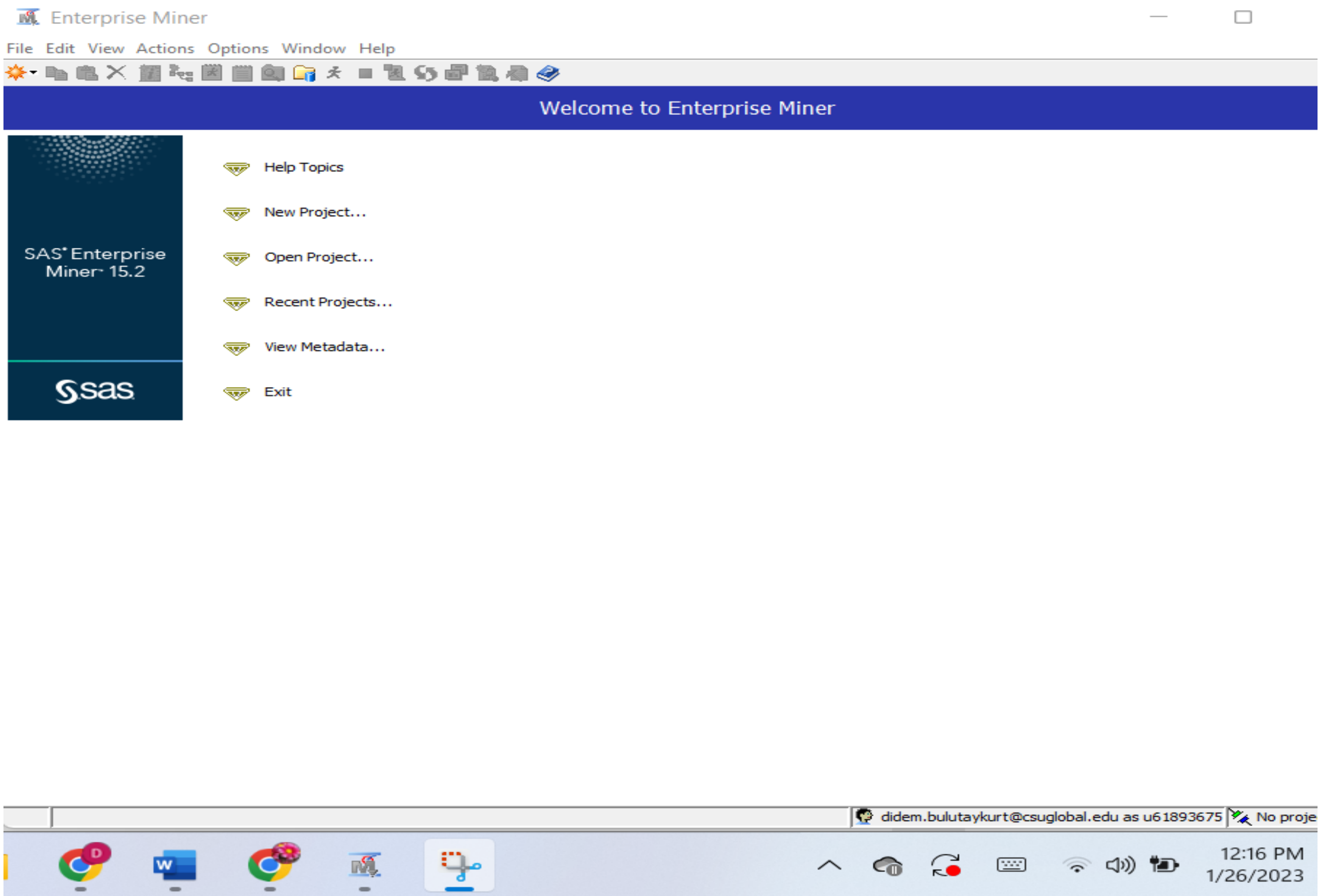
Data Preparation Using the SAS Enterprise Miner

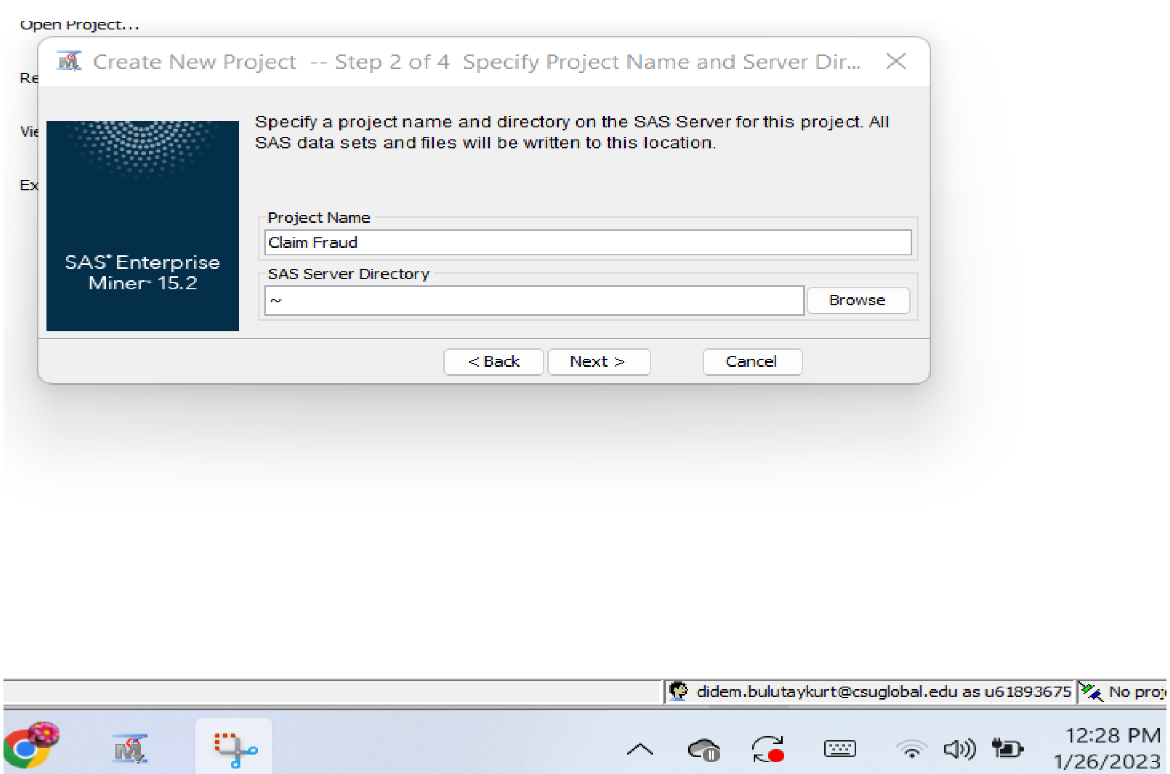
When I start to use a new programming language or software, I am always searching for why I need it or the pros and cons of this program. The first time I faced it, it was not fast and easy to use. They should update the program. Also, most learning tools were updated a few years ago; video or other devices do not support the new version. It is so complicated. The program has limited learning sources online, which tells me a lower percentage of people use this software program.

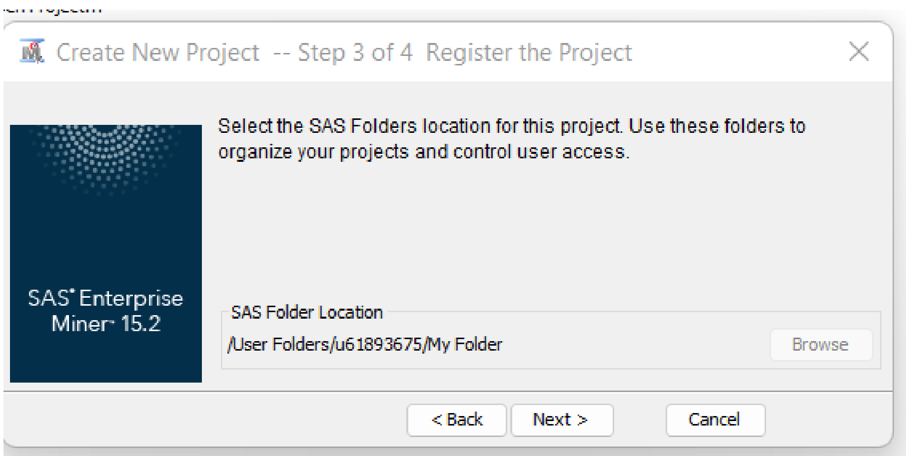
Searching for the SAS Enterprise Miner is a solution to creating accurate predictive and descriptive models on large volumes of data across different sources in the organization. (Pat Research,2021) Some business applications are for detecting fraud, minimizing risk and resource demands, reducing asset downtime and campaigns, and reducing customer attrition—the most popular category used for predictive analytics. Let’s dive deep into SAS Enterprise Miner data processing step by step. There are four components for the predictive creation model in SAS Enterprise Miner:

Creating the Project File

First launched, the Welcome screen appears. Create a new project, select New Project, and open new windows. I entered the project name “Claim Fraud,” and the SAS Server Directory showed where the file was saved. I left the automatic field location and finished.

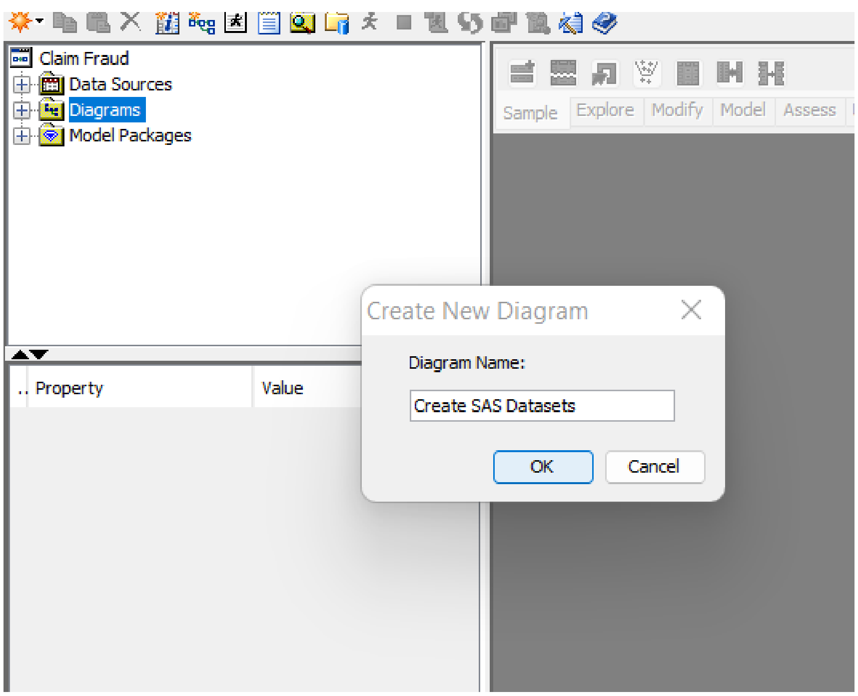

Enterprise Miner
File Edit View Actions Options Window Help
Welcome to Enterprise Miner
SAS® Enterprise Miner™ 15.2
sas
Help Topics
New Project...
Open Project...
Recent Projects...
View Metadata...
Exit
didem.bulutaykurt@csuglobal.edu as u61893675
No proje
12:16 PM
1/26/2023

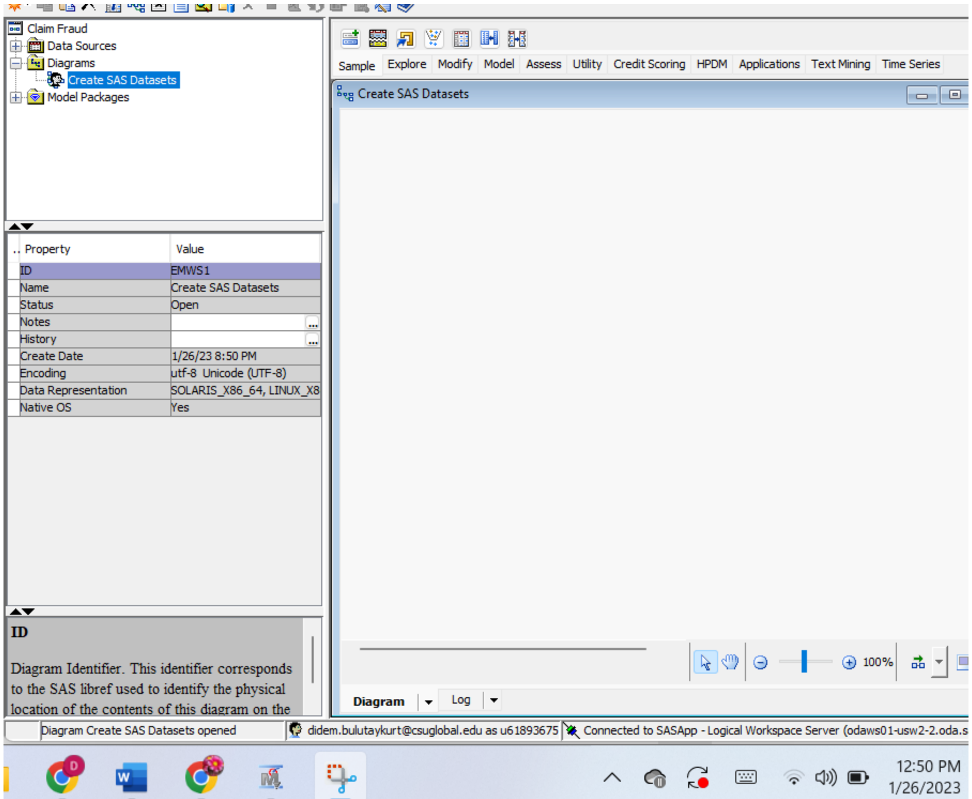

Open Project...
Create New Project -- Step 2 of 4 Specify Project Name and Server Dir...
SAS® Enterprise Miner™ 15.2
Specify a project name and directory on the SAS Server for this project. All SAS data sets and files will be written to this location.
Project Name
Claim Fraud
SAS Server Directory
~
Browse
< Back
Next >
Cancel
didem.bulutaykurt@csuglobal.edu as u61893675
No proj
12:28 PM
1/26/2023



Create the Diagram

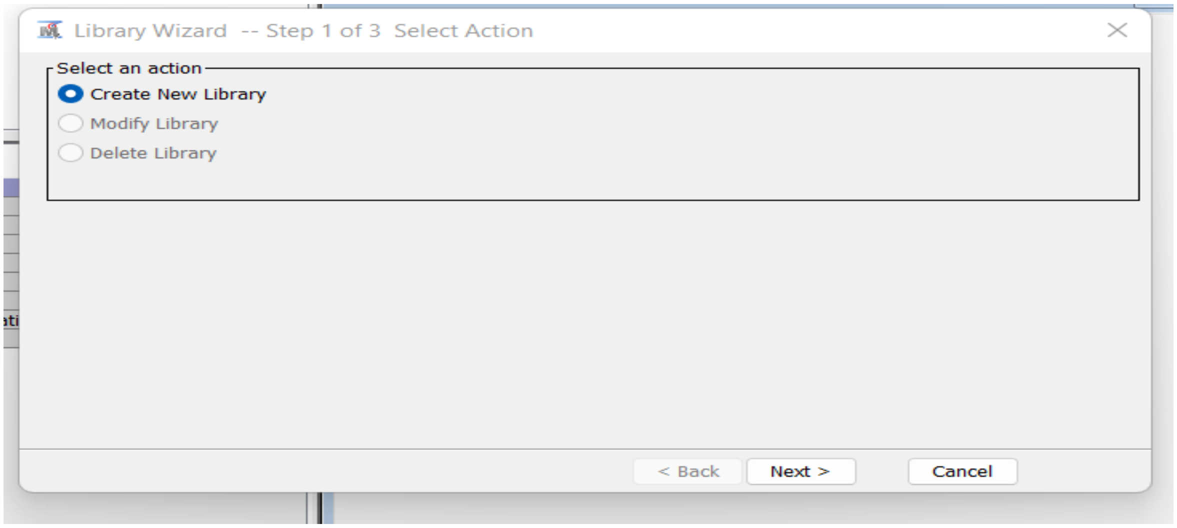
When the project window opens, the left side has a project panel list; right-click Diagrams and select Create Diagram, then pop up the new window named the diagram "Create SAS Datasets."

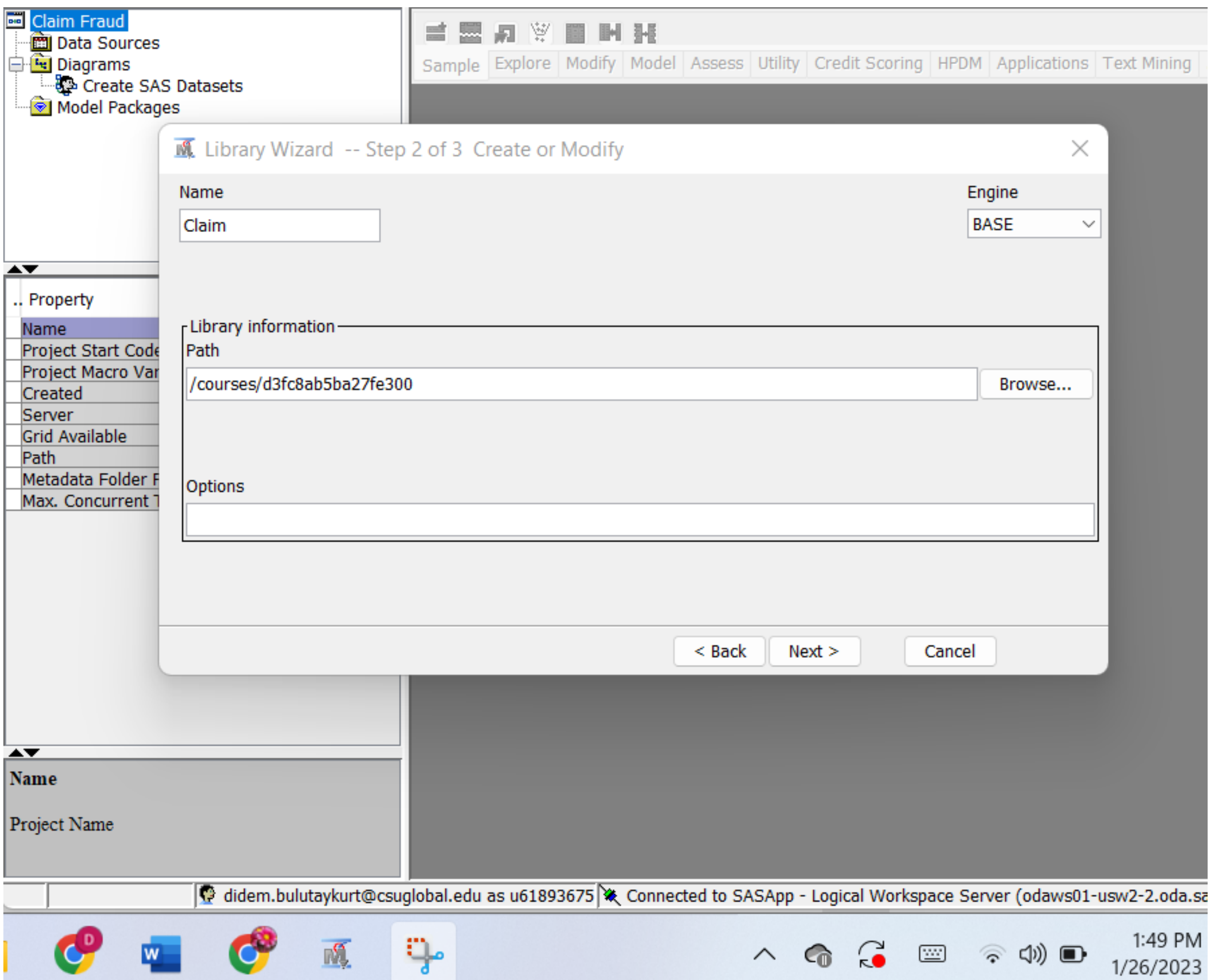




Create the Library

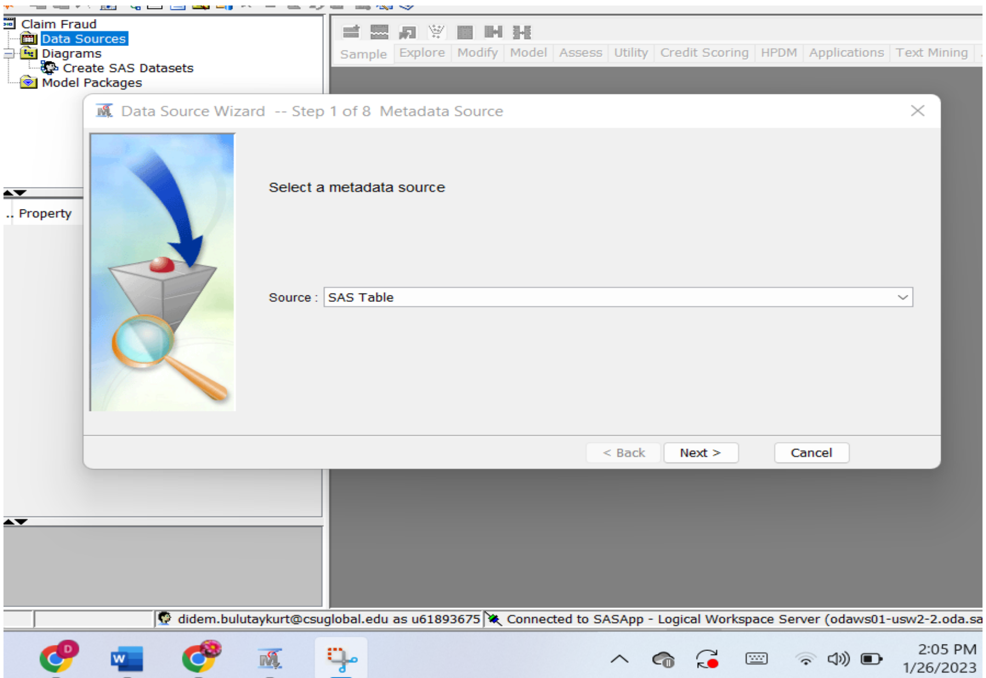
The SAS library creates the directory location of the stored SAS data files. To build a SAS library, click on the File Menu-New-Library, then the library wizard opens, and I named “Claim" the used library path as /courses/d3fc8ab5ba27fe300 next and finished. This library path already loaded the claim dataset.

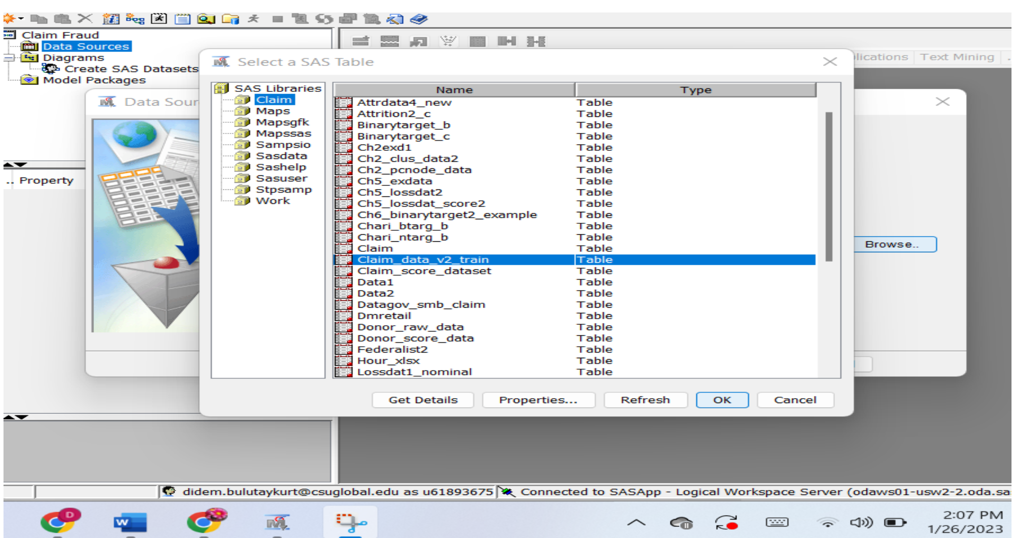

Library Wizard -- Step 1 of 3 Select Action
Select an action
Create New Library
Modify Library
Delete Library
< Back
Next >
Cancel

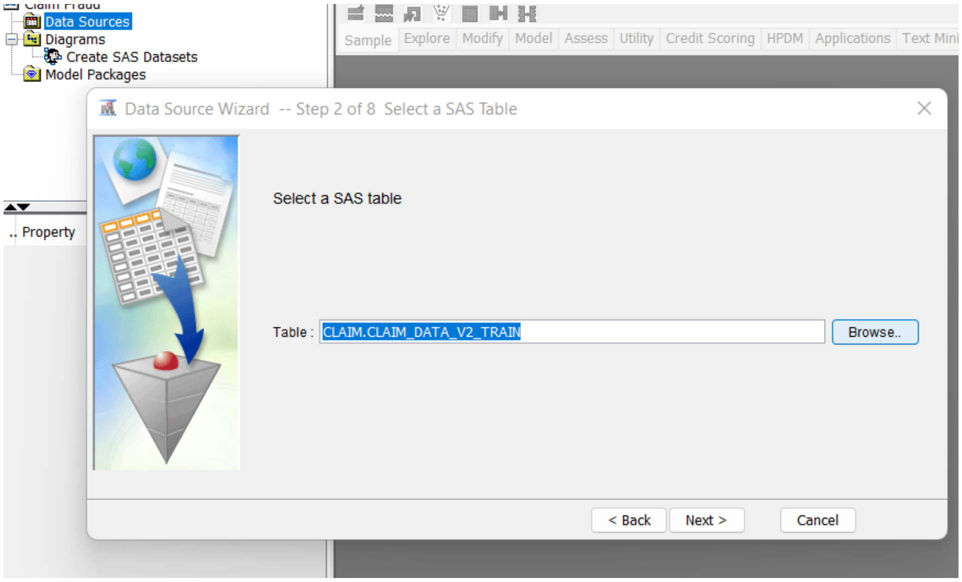

Claim Fraud
Data Sources
Diagrams
Create SAS Datasets
Model Packages
Sample
Explore
Modify
Model
Assess
Utility
Credit Scoring
HPDM
Applications
Text Mining
Library Wizard -- Step 2 of 3 Create or Modify
Name
Claim
Engine
BASE
Library information
Path
/courses/d3fc8ab5ba27fe300
Browse...
Options
< Back
Next >
Cancel
Property
Name
Project Start Code
Project Macro Var
Created
Server
Grid Available
Path
Metadata Folder F
Max. Concurrent T
Name
Project Name
didem.bulutaykurt@csuglobal.edu as u61893675
Connected to SASApp - Logical Workspace Server (odaws01-usw2-2.oda.sa
1:49 PM
1/26/2023

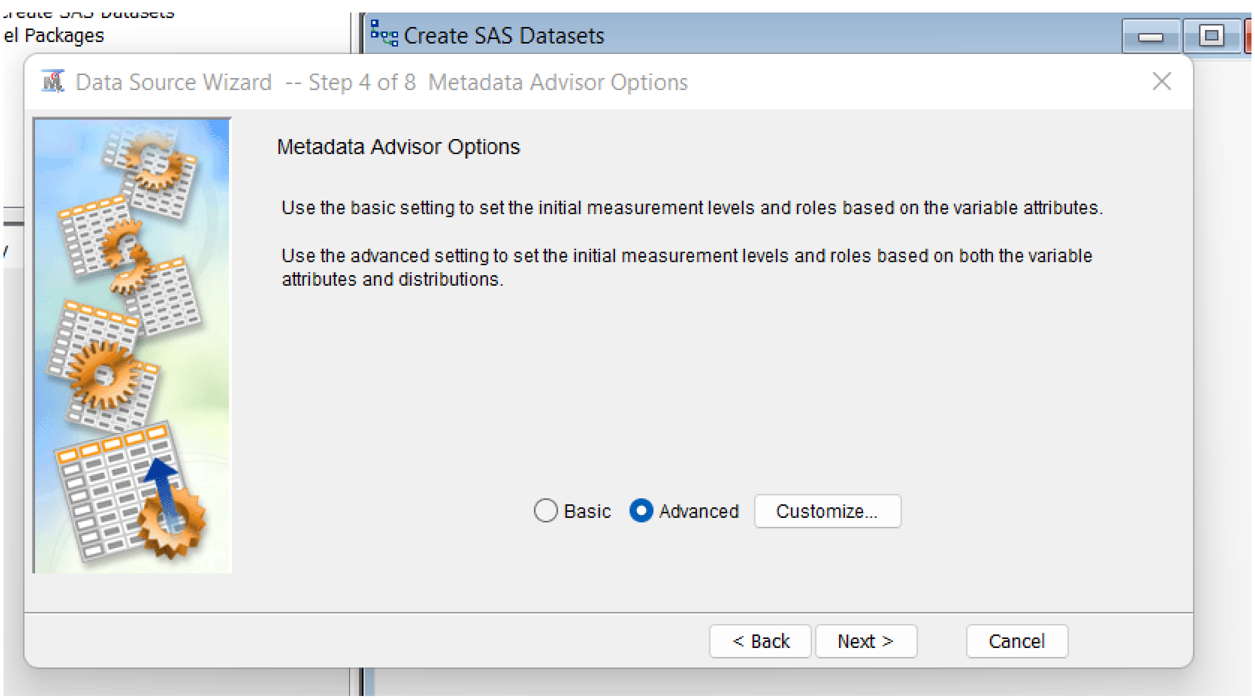
Create Data Source

Now, I have four components to start creating the data sources. I used the auto insurance claim data set that had already loaded. To make the data sources, click the File menu-New-Data Source, or the left side Project panel has Data Sources; Right click, then pop up Data Sources Wizard and select Claim.

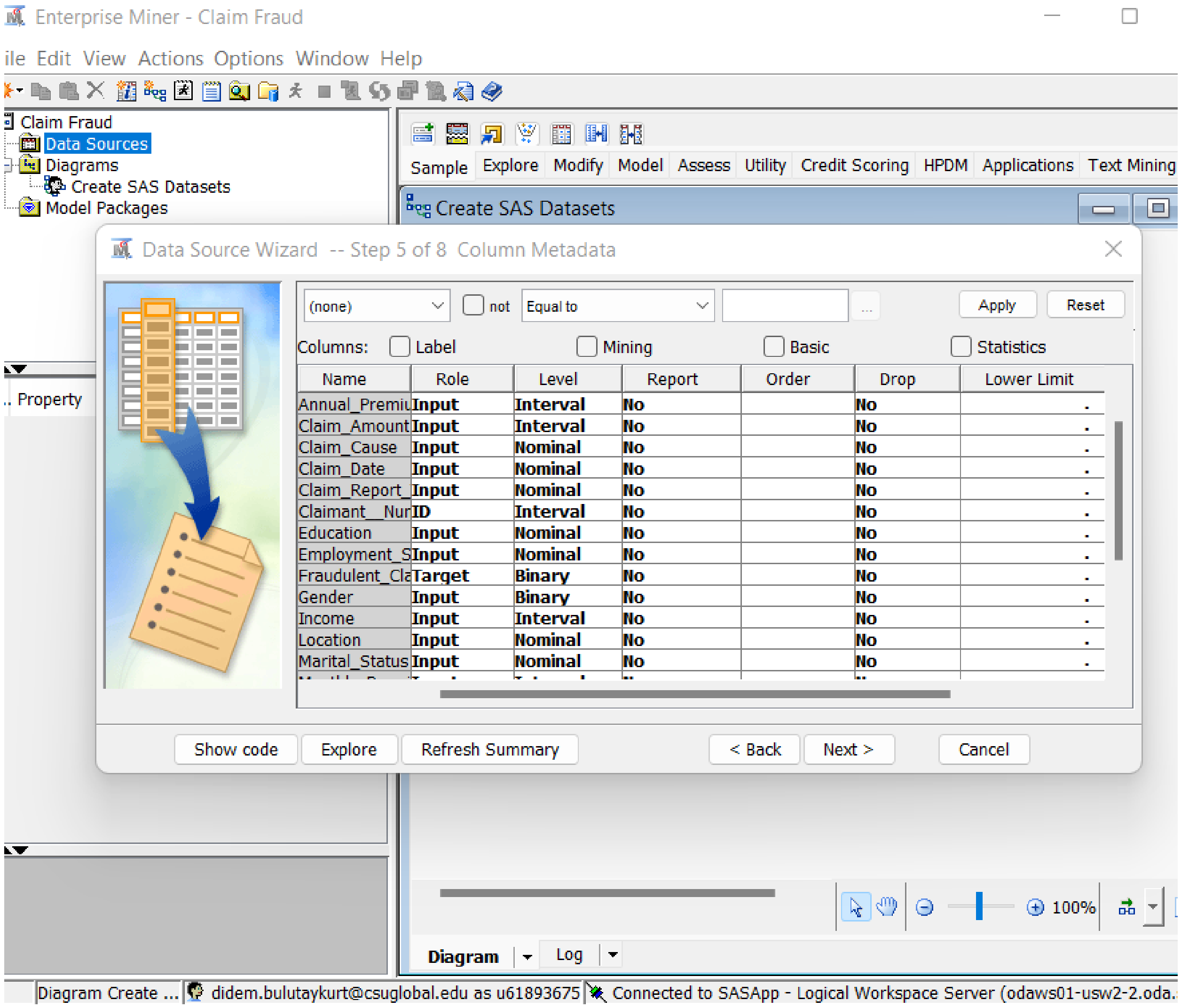


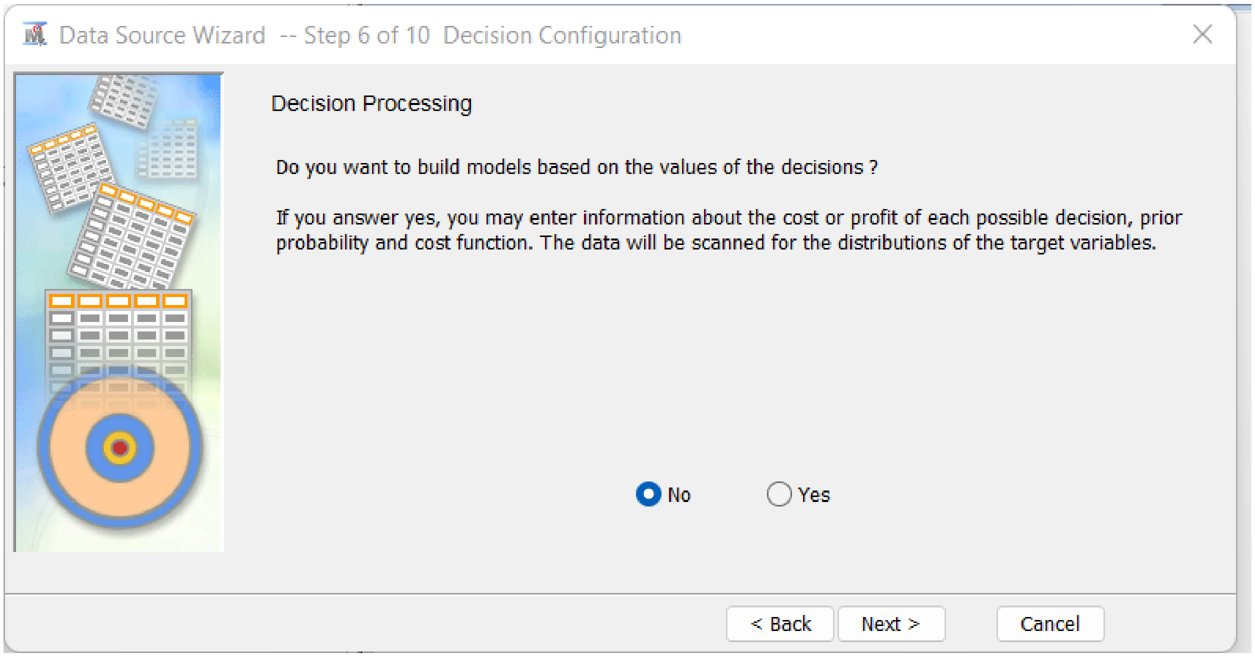


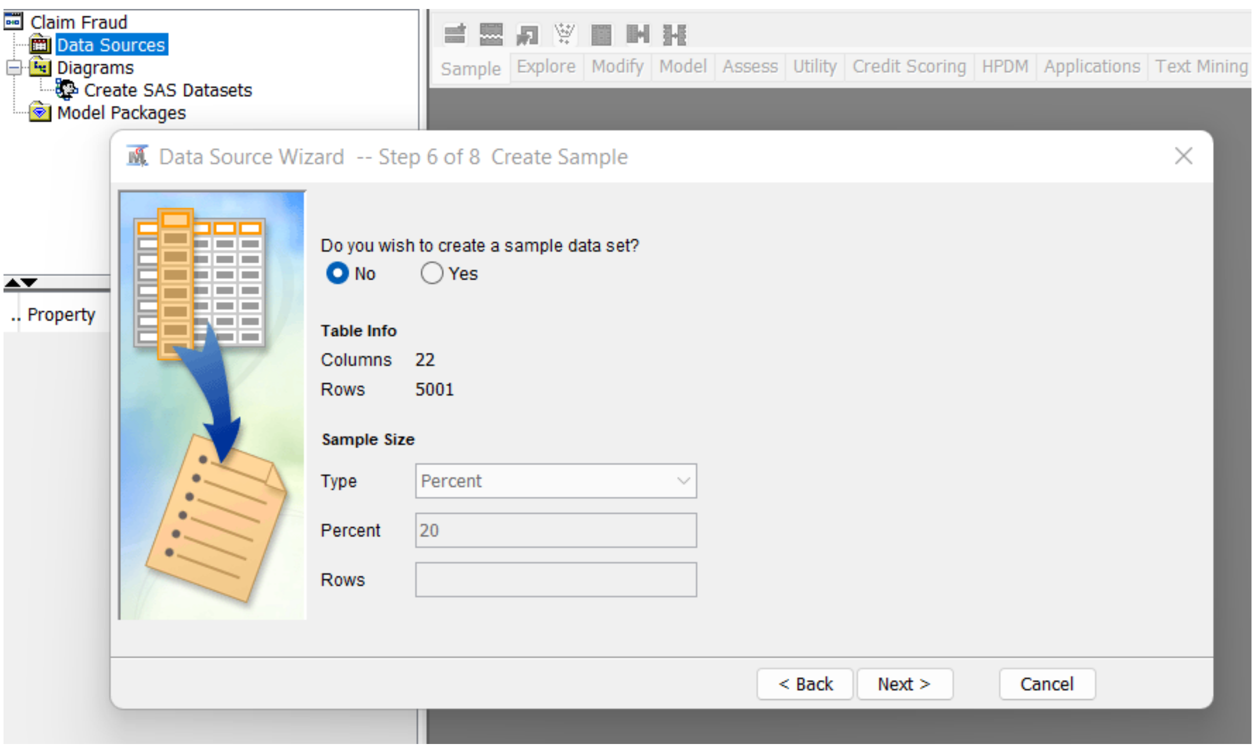


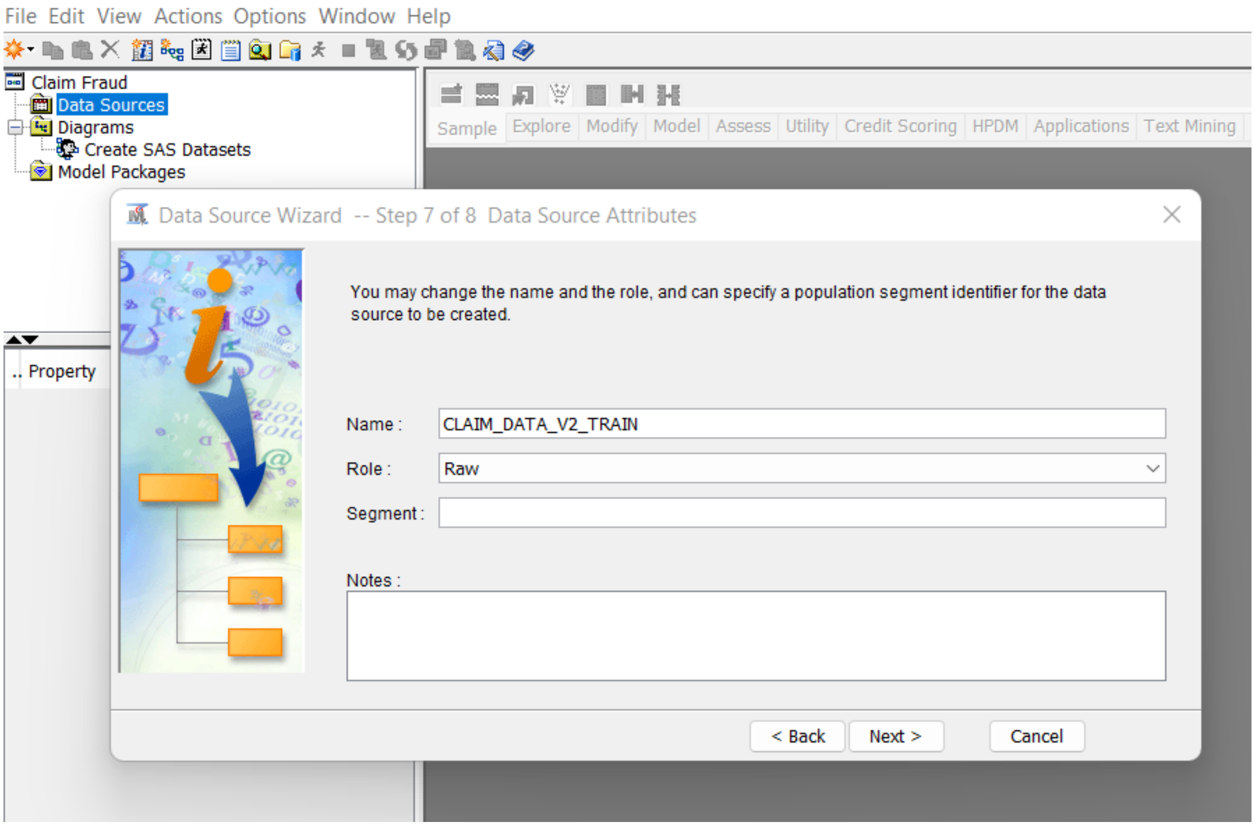


To change roles and levels, the next Metadata Advisor Option. The Fradulent_Claim variable's column Role change Input to Target as a dependent variable; it is binary (yes/no). The Claimant_Number's Role column changed Input to ID variable. The state variable will not be used in the model as changed Role Input to Reject.

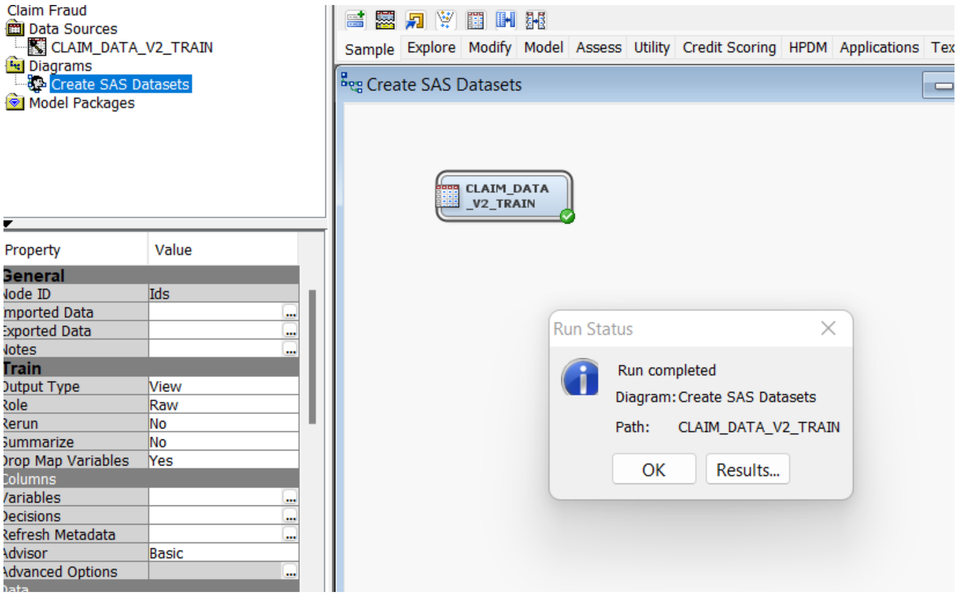



Data Source Wizard -- Step 6 of 10 Decision Configuration
Decision Processing
Do you want to build models based on the values of the decisions ?
If you answer yes, you may enter information about the cost or profit of each possible decision, prior probability and cost function. The data will be scanned for the distributions of the target variables.
No
Yes
< Back
Next >
Cancel


Claim Fraud
Data Sources
Diagrams
Create SAS Datasets
Model Packages
Sample
Explore
Modify
Model
Assess
Utility
Credit Scoring
HPDM
Applications
Text Mining
Property
Data Source Wizard -- Step 6 of 8 Create Sample
Do you wish to create a sample data set?
No
Yes
Table Info
Columns 22
Rows 5001
Sample Size
Type Percent
Percent 20
Rows
< Back
Next >
Cancel



The next step is I drag the CLAIM_DATA_V2_TRAIN from the Project Panel to the Diagram Create SAS Datasets in the workspace. Then right-click on the node and RUN. Two ways to look at results. One is that the end of the run window has a result option or right-click on the dataset and Result.



Output

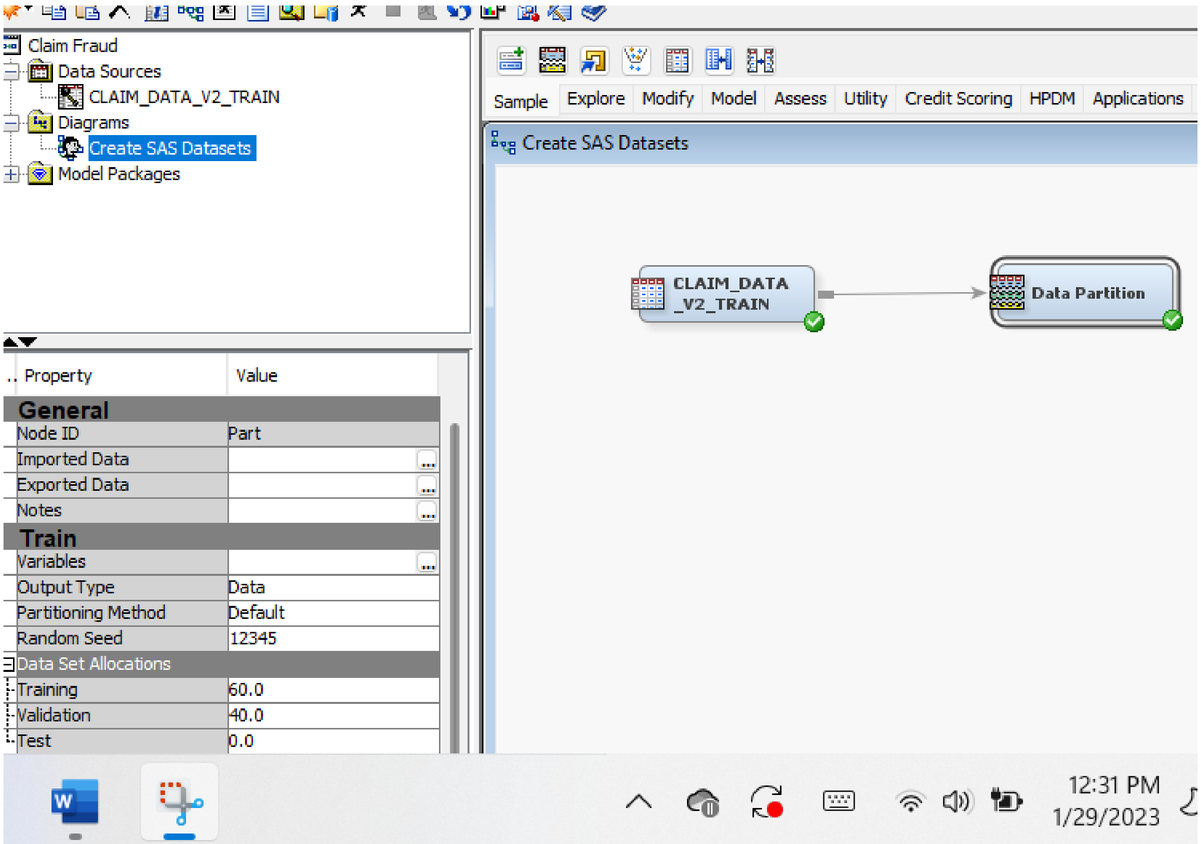
```
*------------------------------------------------------------*
User:                    u61893675
Date:                    26 January 2023
Time:                    23:39:07
*------------------------------------------------------------*
* Training Output
*------------------------------------------------------------*




Variable Summary

               Measurement    Frequency
Role              Level         Count

ID              INTERVAL          1
INPUT           BINARY            1
INPUT           INTERVAL          7
```

Variables

Variable Name	Role	Measurement Level	Order
Annual Premium	Input	Interval	
Claim Amount	Input	Interval	
Claim Cause	Input	Nominal	
Claim Date	Input	Nominal	
Claim Report Type	Input	Nominal	
Claimant Number	ID	Interval	
Education	Input	Nominal	
Employment Status	Input	Nominal	
Fraudulent Claim	Target	Binary	
Gender	Input	Binary	
Income	Input	Interval	
Location	Input	Nominal	
Marital Status	Input	Nominal	
Monthly Premium	Input	Interval	
Months Since Last Claim	Input	Interval	
Months Since Policy Inception	Input	Interval	
Outstanding Balance	Input	Interval	
State	Rejected	Nominal	

Data Partition

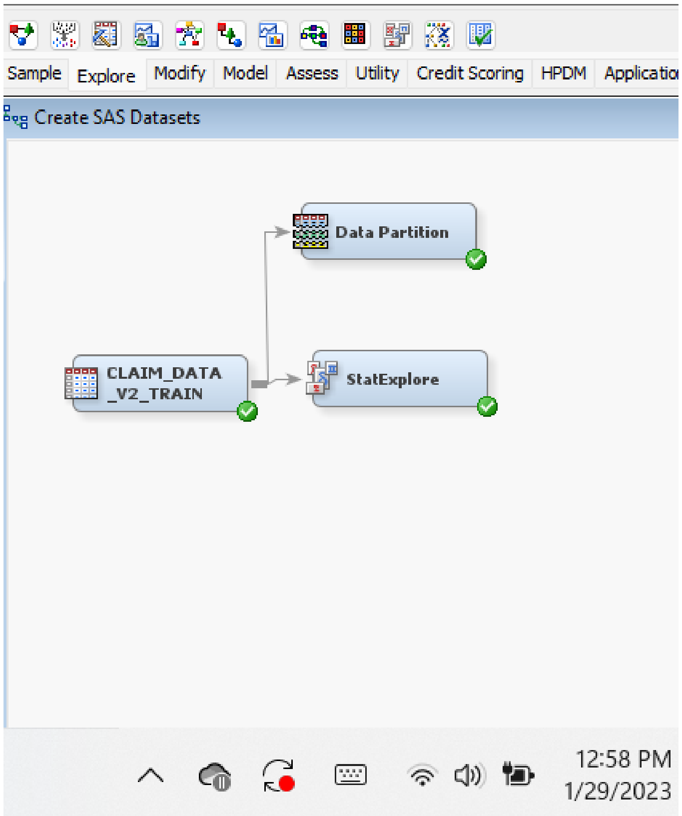
The data set divides the dataset into two or three parts: training is used to build the model, validation checks the model's accuracy, and testing partition tests the model. To create the section, click the Sample tab, then drag the Data Partition node onto the process flow. Finally, connect the Data Partition node to the CLAIM_DATA_TRAIN node. I divided the training partition into 60% and the validation partition into 40%.

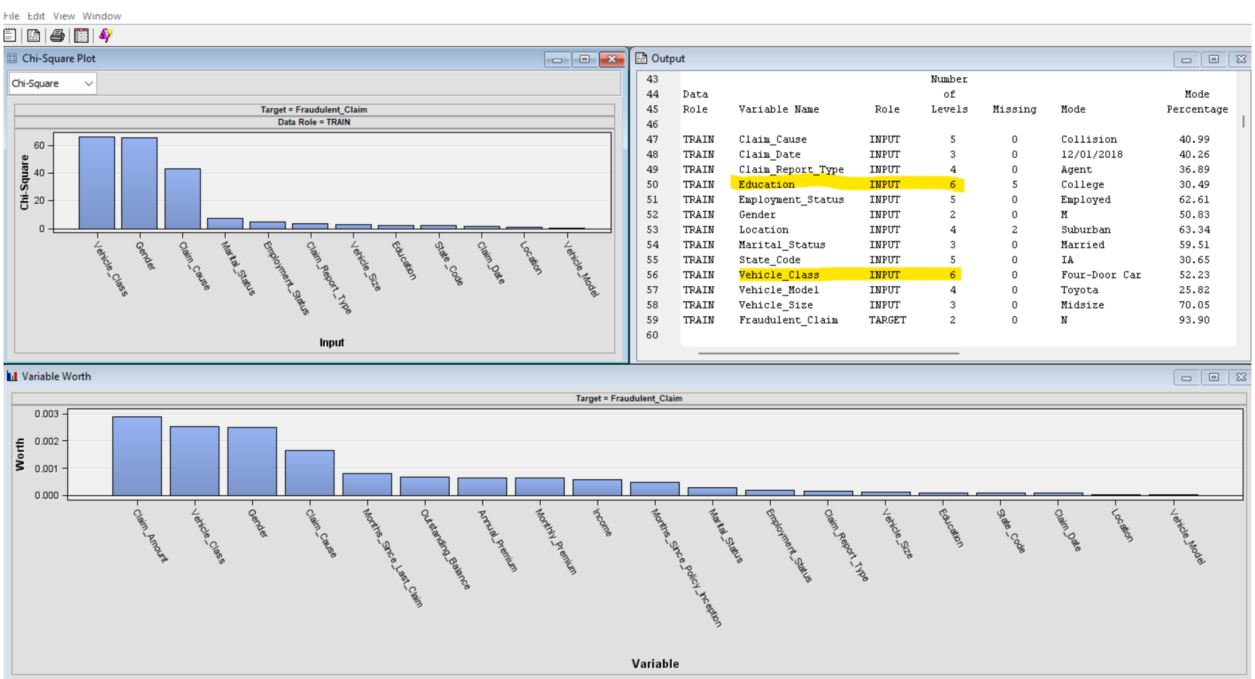


Data Exploration

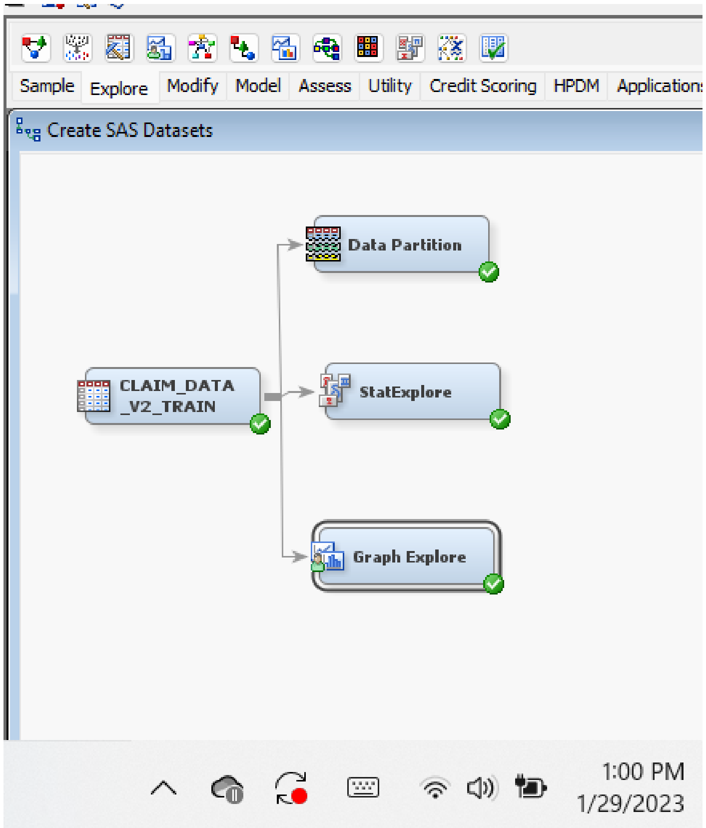
The StatExplore node shows the data summary and identifies the missing values. The Graph Explore node helps to see data point behavior by creating a histogram, stem-and-leaf plots, and box plot; both nodes are significant for descriptive statistics.

Click the Explore tab, then drag the StatExplore node onto the process flow. Finally, connect the StatExplore node to the CLAIM_DATA_TRAIN node and right-click on the node-run.

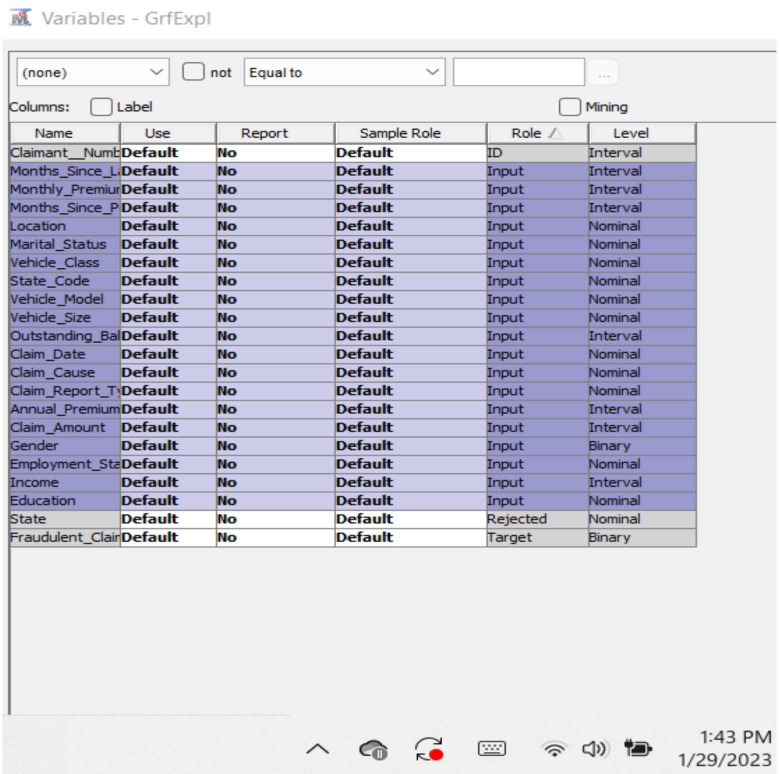




The Graph Explore node; Click the Explore tab, then drag the Graph Explore node onto the process flow. Connect the Graph Explore node to the CLAIM_DATA_TRAIN node and right-click on the node-run.



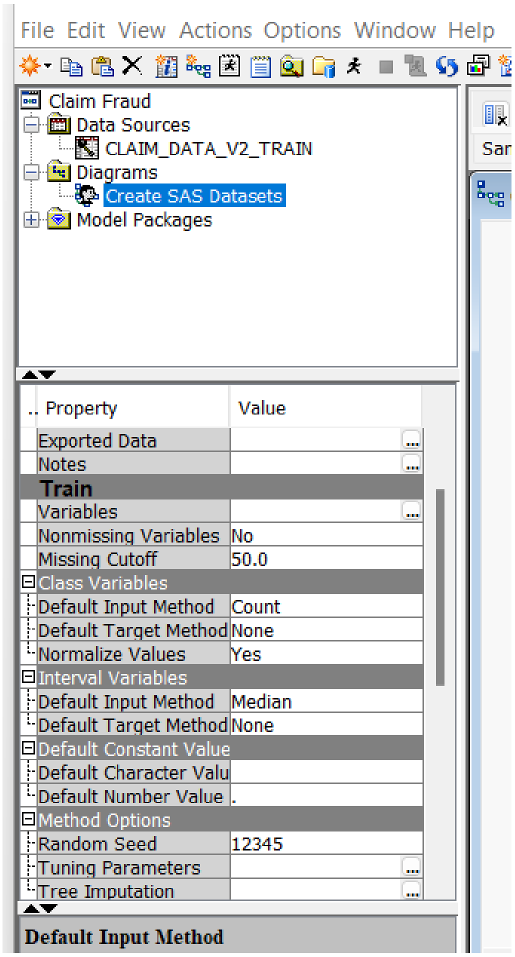
After running the Graph Explore node, right-click and select Edit Variable. For sorting by column, click on the Role column title to sort by Role. For example, to determine all the input variables you want, click the Marital_Status variable name, hold the shift down, and click on Education.

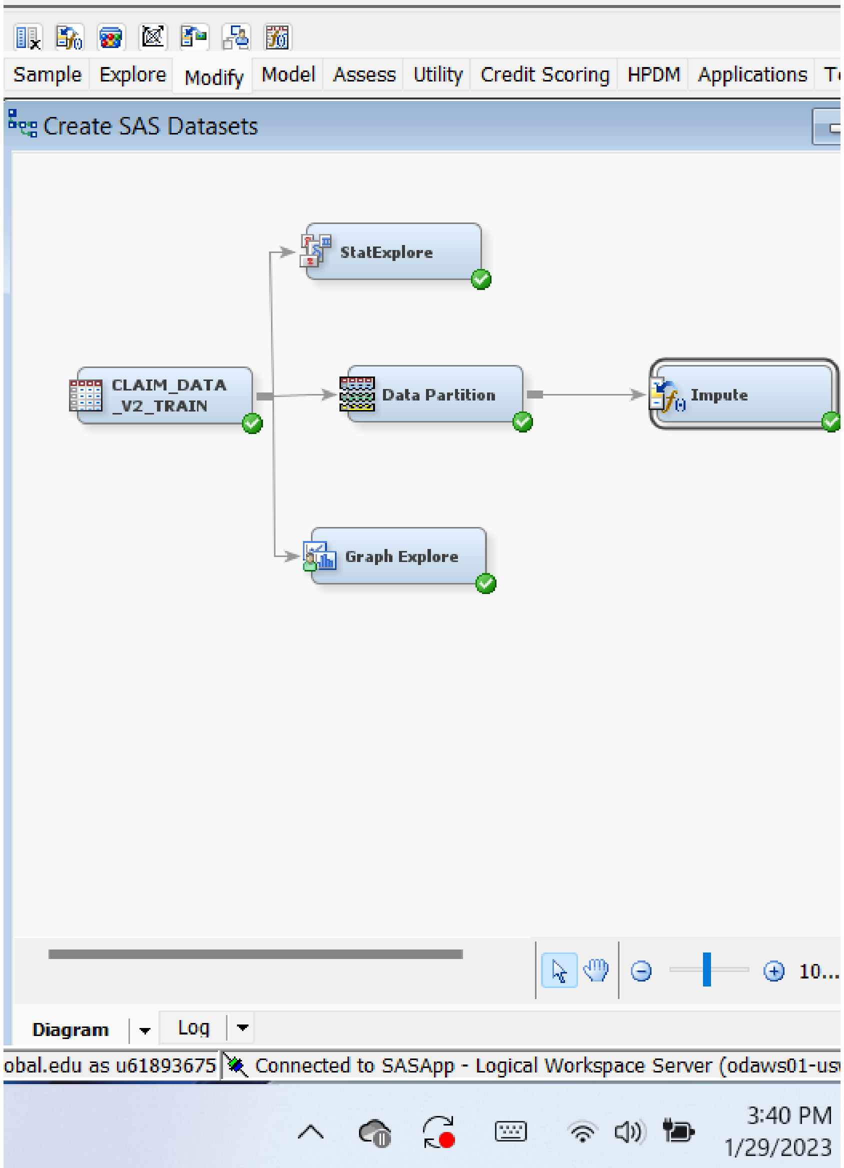


Click the Explore button, and a new window displays a histogram of all the input variables. And the Income variable has an outlier.

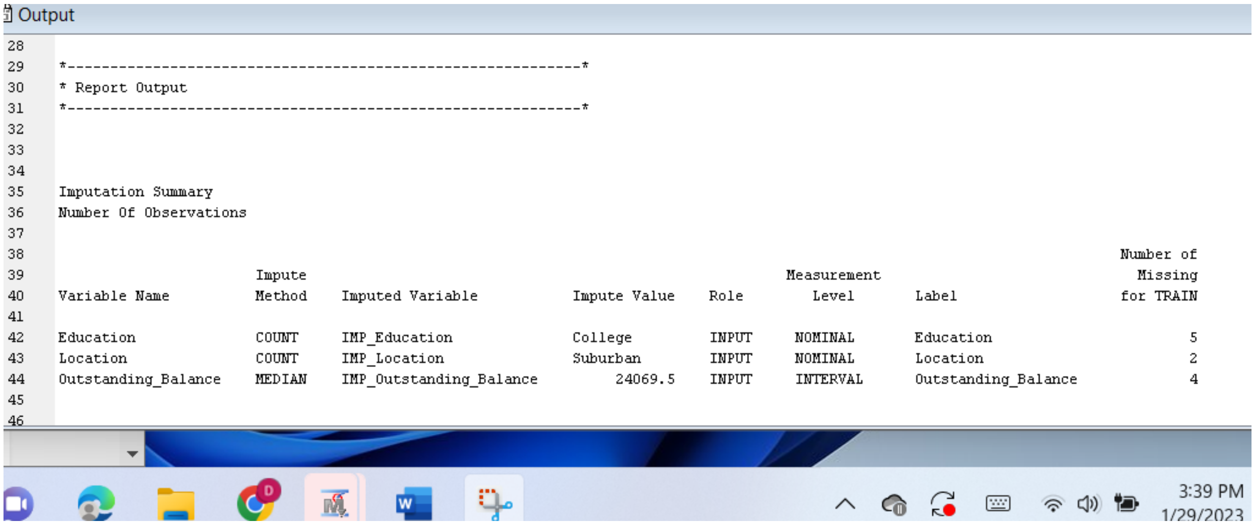
Missing Data

Missing values can cause the model result, which is why to improve data. Two methods to handle missing values; are listwise deletion and imputation, replacing the missing values with substitute values. The Impute node is used to deal with missing values. Click the Modify tab, then drag the Impute node to the process flow diagram. The Default Input Method for the Interval Variables changed from Mean to Median.





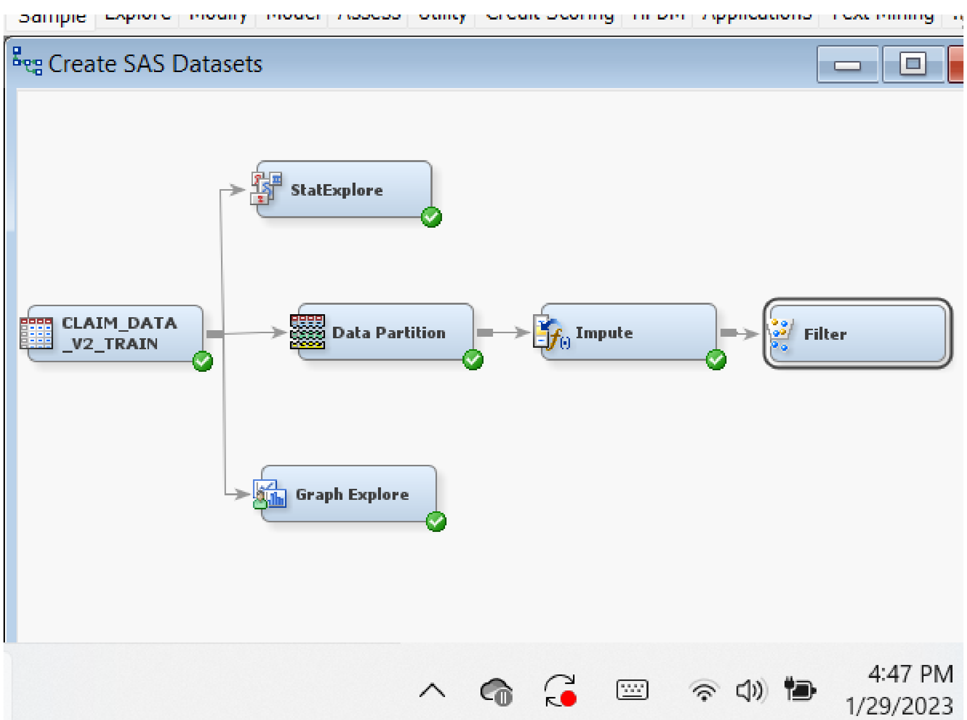
The Impute node creates a new table with a new variable replacement value for the missing data. The Impute node creates a default table as the original data set in the original dataset are not overwritten. New variables containing the impute values can identify with the prefix IMP_.



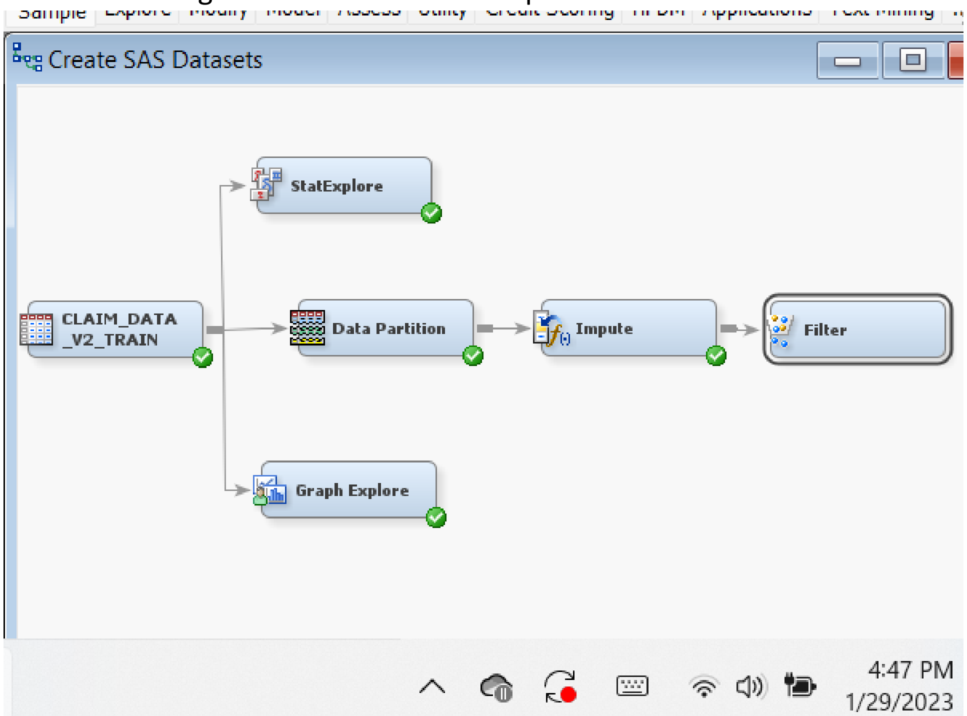
```
Output

*------------------------------------------------------------*
* Report Output
*------------------------------------------------------------*



Imputation Summary
Number Of Observations

                                                                                                              Number of
                        Impute                                                      Measurement                 Missing
Variable Name           Method    Imputed Variable            Impute Value   Role       Level      Label                   for TRAIN

Education               COUNT     IMP_Education               College        INPUT    NOMINAL     Education                      5
Location                COUNT     IMP_Location                Suburban       INPUT    NOMINAL     Location                       2
Outstanding_Balance     MEDIAN    IMP_Outstanding_Balance          24069.5   INPUT    INTERVAL    Outstanding_Balance            4
```

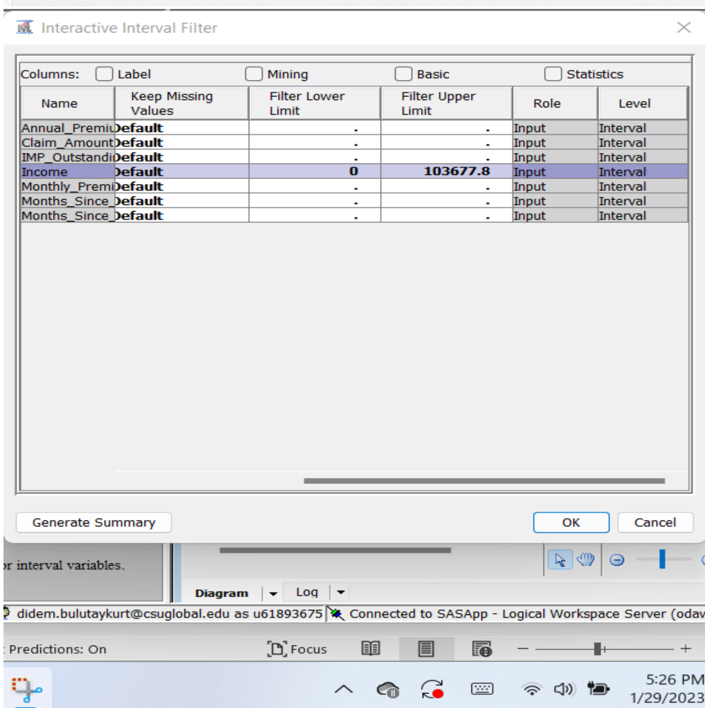

Handling Outliers

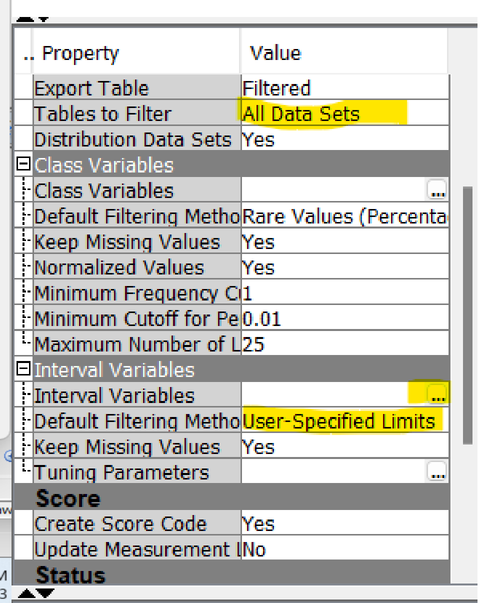
The Filter node helps to identify and eliminate outliers and filter the dataset. Filtering uses the data from the training dataset for the better result of models. Additionally, the Filter node ignores target and rejected variables. Use the Sample Tab, drag a Filter node to the process flow diagram, and connect the Impute to the Filter node.



Let's look at the Claim fraud dataset, as only Income has outliers on the Graph Explore node showed. Just the Income variable needs to be filtered; the filter setting needs to change under the Train group, Table to Filter set All Data Sets, and under the Interval Variables group, click the Interval Variables' ellipsis and set the minimum income zero and the maximum income to 103,677.78 that value probably 99.7% of the values are within three standard deviations. The Default Filtering Method is set to User-Specified Limits.







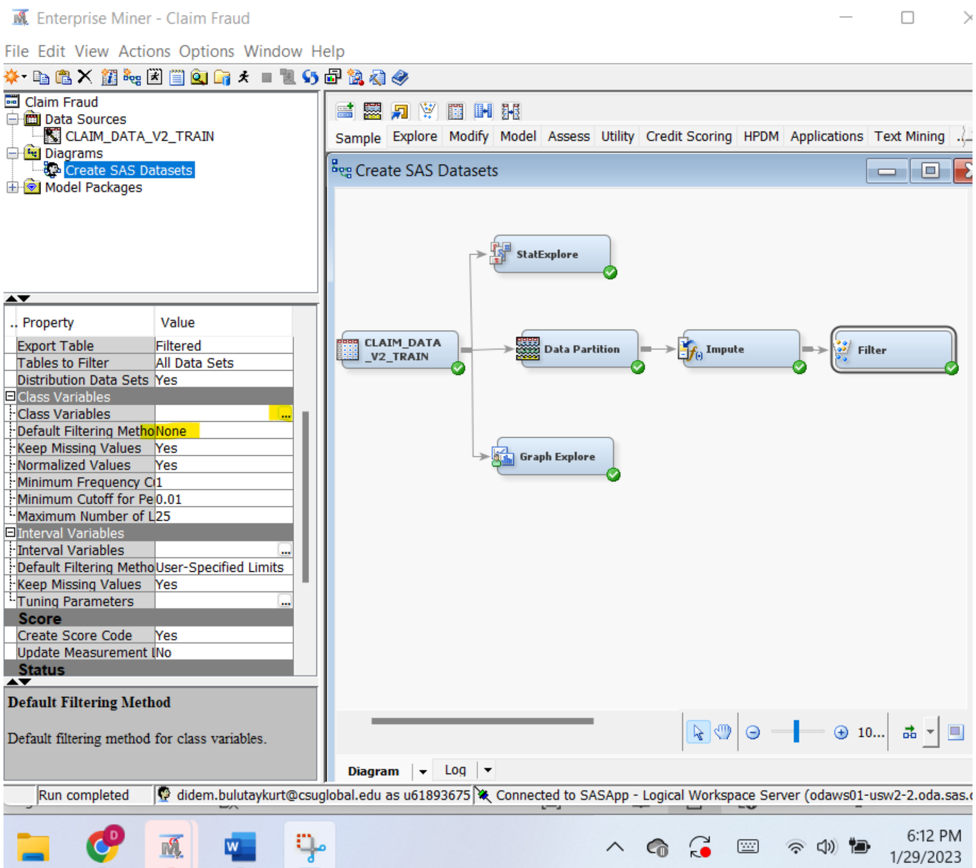
```
Output

                                              Train      Train                      Filter
 Variable       Role      Level               Count     Percent        Label        Method

Claim_Cause    INPUT    FIRE                     1    0.033356    Claim_Cause    MINPCT




Number Of Observations

Data
Role        Filtered    Excluded    DATA

TRAIN         2996          2       2998
VALIDATE      2001          2       2003



Statistics for Original and FILTERED Data
(maximum 500 observations printed)

Data Role=TRAIN Variable=Income

Statistics              Original     Filtered

Non Missing              2998.00      2996.00
Missing                     0.00         0.00
Minimum                     0.00         0.00
Maximum                933288.00     99981.00
Mean                    38326.89     38040.96
Standard Deviation      34559.26     30448.93
Skewness                    5.97         0.28
Kurtosis                  148.50        -1.11


Data Role=VALIDATE Variable=Income

Statistics                Original     Filtered

Non Missing                2003.00      2001.00
Missing                       0.00         0.00
Minimum                       0.00         0.00
Maximum                15967801.00     99960.00
Mean                      45776.10     37717.00
Standard Deviation       357290.69     30708.64
Skewness                     44.25         0.29
Kurtosis                   1972.80        -1.09


*-------------------------------------------------------------*
```

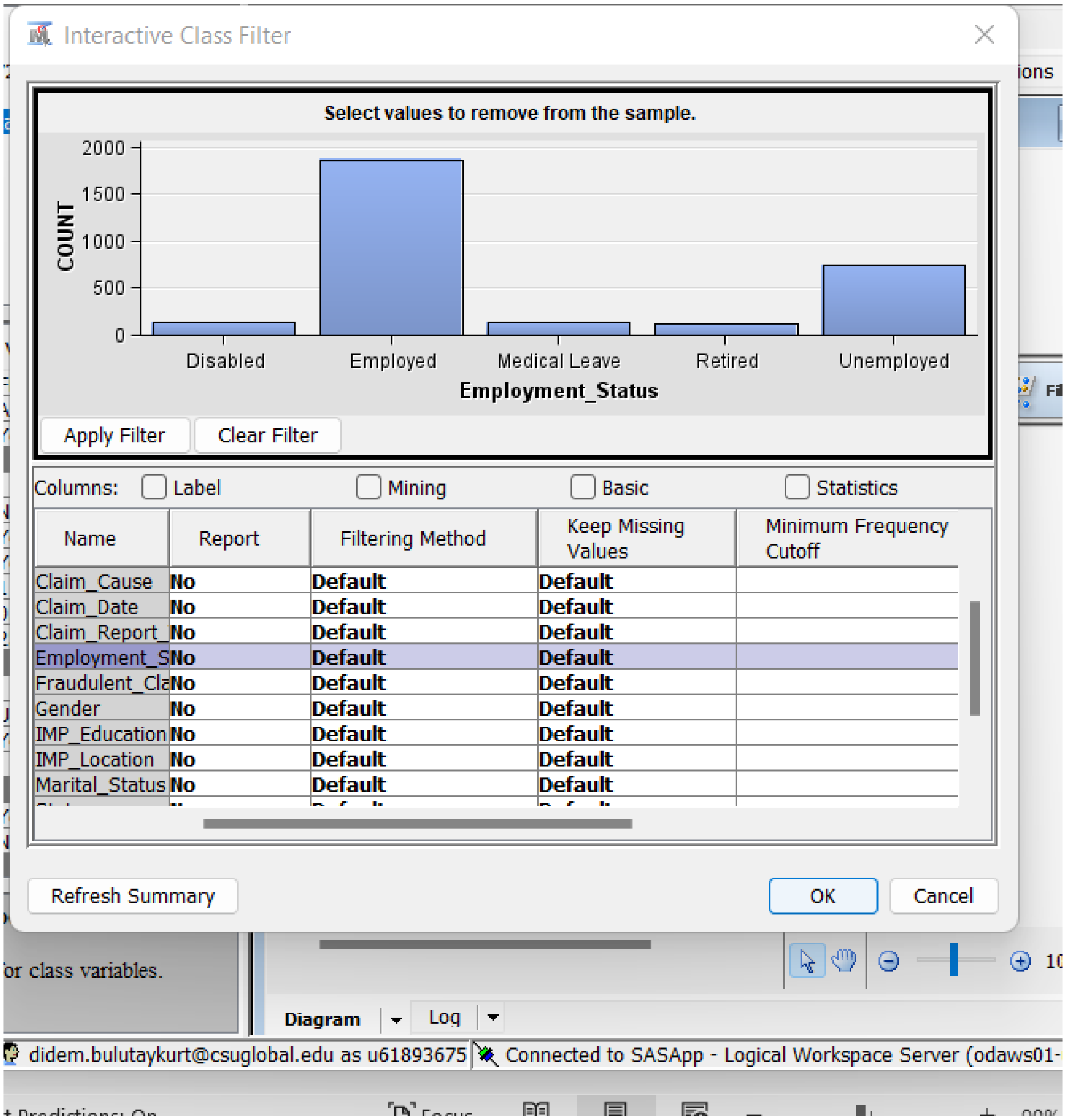
As a result, training and validation have two observations that were filtered. And the train data display that the maximum income value in the partition is $99,981, but the original maximum value is $933,288.

Categorical Variables with Too Many Levels

One of the practical uses of many categories or class variables, like zip code variables, is that they can be combined at the city or state levels and can also combine their frequency. The Replacement node connects groups and establishes different levels for the group. The Filter node can be used to set minimum frequency and several levels.

Use the same direction for the filter node, set None for the Default Filtering Method under the Class Variables group, and click on the ellipsis by Class Variables. And click on each variable to see the histogram top on the window and can update the Minimum and Maximum Frequency Cutoff.




Interactive Class Filter
Select values to remove from the sample.
COUNT
2000
1500
1000
500
0
Disabled
Employed
Medical Leave
Retired
Unemployed
Employment_Status
Apply Filter
Clear Filter
Columns:
Label
Mining
Basic
Statistics
Name
Report
Filtering Method
Keep Missing Values
Minimum Frequency Cutoff
Claim_Cause No Default Default
Claim_Date No Default Default
Claim_Report_ No Default Default
Employment_S No Default Default
Fraudulent_Cla No Default Default
Gender No Default Default
IMP_Education No Default Default
IMP_Location No Default Default
Marital_Status No Default Default
Refresh Summary
OK
Cancel
for class variables.
Diagram
Log
didem.bulutaykurt@csuglobal.edu as u61893675
Connected to SASApp - Logical Workspace Server (odaws01-

References

SAS Enterprise Miner by PAT RESEARCH, 2021. https://www.predictiveanalyticstoday.com/sas-enterprise-miner/

Richard V. McCarthy; Mary M. McCarthy; Wendy Ceccucci, 2022. *Applying Predictive Analytics Finding Value in Data*. Second edition.