WQD7005 DATA MINING

2022/2023 SEMESTER 2

CREDIT SCORE CLASSIFICATION

GROUP ASSIGNMENT 1

DM G1(RL) – Group 1		
No.	**MATRIC ID**	**NAME**
1	22056764	Devayani A/P Balkrishnan
2	S2195613	Li Xin Qi
3	S2155659	Lim Yu Xuan
4	S2192763	Navaneeta A/P P Shanmugam

Table of Contents

1 Introduction

In this modern age, having a loan provides individual the access to cash for purchasing and gain ownership on assets instantly. In exchange, the individual will bear the financial commitment with a repayment of principal and interest to the lender over the years. However, giving out loan is a risky business to the lender as it imposes credit risk when borrowers fail to repay their loans, resulting in loan default. Thus, credit scoring plays a vital role in allowing lenders to assess the risk prior to approving any loan application, especially in recent years with the dramatic growth in consumer credit.

Credit scoring is a score that represents the likelihood of an individual repaying loans. It is a formal statistical method used for differentiating credit applicants into 'good' and 'bad' risk classes (Li et al., 2004).

From a borrower's standpoint, building a good credit scoring will not only allow their loan application to get approved easier as compared to a riskier borrower, but it also qualifies borrowers for lower interest rates. Having a good credit score reflects the individual as a trustworthy borrower and thus would enjoy more offers and rewards. Without a good credit score, it makes individuals vulnerable in getting funds for unexpected financial emergencies that would likely experience greater financial distress.

A few decades ago, financial institutions granted credit to individuals based on human judgement on loan default land experience in the industry (Hand & Henley, 1997). However, with the emergence of underwriting technologies and the economic pressure from the surging credit demand, financial institutions started to develop and leverage on credit scoring as a statistical model for making credit granting decisions. This shows a transition of assessment and pricing credit risk that are more data driven. With credit scoring classification, it optimizes financial institutions the ability to screen high-risk borrowers and target more generous facilities to lower-risk borrowers. These in turn enable financial institutions in minimizing default cost while increasing sales probability and overall expected profits (Einav et al., 2013).

In this project, we aim to build a credit scoring classification model for a finance company by leveraging on their customer's credit-related information. The expected outcome of the project is to segregate these customers into different credit score brackets that will enable the company to assess their customer's credit risk level. The analysis will be done and visualized based on data mining steps developed by SAS institute named SEMMA, which follows with the process sequence of Sample, Explore, Modify, Model and Assess.

2 Analysis Goal

The analysis goal of this study is to evaluate the borrower's creditworthiness and classify them via credit scoring based on the credit-related information. This study is examined by building a credit scoring classification model with different analytics models that best classify borrowers' credit scoring.

3 Accessing and Assaying Prepared Data

3.1 Data Acquisition

The datasets were acquired from https://www.kaggle.com/datasets/parisrohan/credit-score-classification?resource=download. It was uploaded by Rohan Paris in July 2022 under the title 'Credit score classification'. The purpose of the datasets was to create a machine learning model that can categorize the credit score when given information about a person's credit history. Only the training dataset (train.csv) was chosen to be used in the model creation as it has the column denoting the credit score. The testing dataset (test.csv) will be used during the testing phase of our models.

3.2 Analysis of Data (Number of attributes/rows, anomalies, target features)

There are 28 attributes in the training dataset and 27 in the test dataset, with the credit score column being the difference. The training dataset has 100,000 rows while the testing dataset has 50,000 rows. The training dataset spans from January to August for each customer while the testing dataset covers the rest of the year, from September to December. Both datasets have a total of 12,500 unique customers and their data. The target feature of the data is the credit score which has three classifications, namely Poor, Standard, or Good.

3.3 Column Metadata

The table below describes the columns in the datasets.

Table 3.3.1 Column Metadata

Column	Description
ID	Represents a unique identification of an entry
Customer_ID	Represents a unique identification of a person
Month	Represents the month of the year
Name	Represents the name of a person
Age	Represents the age of the person
SSN	Represents the social security number of a person
Occupation	Represents the occupation of the person
Annual_Income	Represents the annual income of the person
Monthly_Inhand_Salary	Represents the monthly base salary of a person
Num_Bank_Accounts	Represents the number of bank accounts a person holds
Num_Credit_Card	Represents the number of other credit cards held by a person
Interest_Rate	Represents the interest rate on credit card
Num_of_Loan	Represents the number of loans taken from the bank
Type_of_Loan	Represents the types of loan taken by a person
Delay_from_due_date	Represents the average number of days delayed from the payment date
Num_of_Delayed_Payment	Represents the average number of payments delayed by a person
Changed_Credit_Limit	Represents the percentage change in credit card limit
Num_Credit_Inquiries	Represents the number of credit card inquiries

Credit_Mix	Represents the classification of the mix of credits
Outstanding_Debt	Represents the remaining debt to be paid (in USD)
Credit_Utilization_Ratio	Represents the utilization ratio of credit card
Credit_History_Age	Represents the age of credit history of the person
Payment_of_Min_Amount	Represents whether only the minimum amount was paid by the person
Total_EMI_per_month	Represents the monthly EMI payments (in USD)
Amount_invested_monthly	Represents the monthly amount invested by the customer (in USD)
Payment_Behaviour	Represents the payment behavior of the customer (in USD)
Monthly_Balance	Represents the monthly balance amount of the customer (in USD)
Credit_Score	Represents the bracket of credit score (Poor, Standard, Good)

4 Methodology: SEMMA

SAS Institute demonstrates data mining using the SEMMA methodology to uncover hidden information from massive amounts of data and improve business decision-making. The SEMMA process is well-known in handling business problems across various sectors as it aids businesses to gain a competitive edge, enhance performance, and provide customers with more beneficial services. The diagram below illustrates each phase in SEMMA in detail.

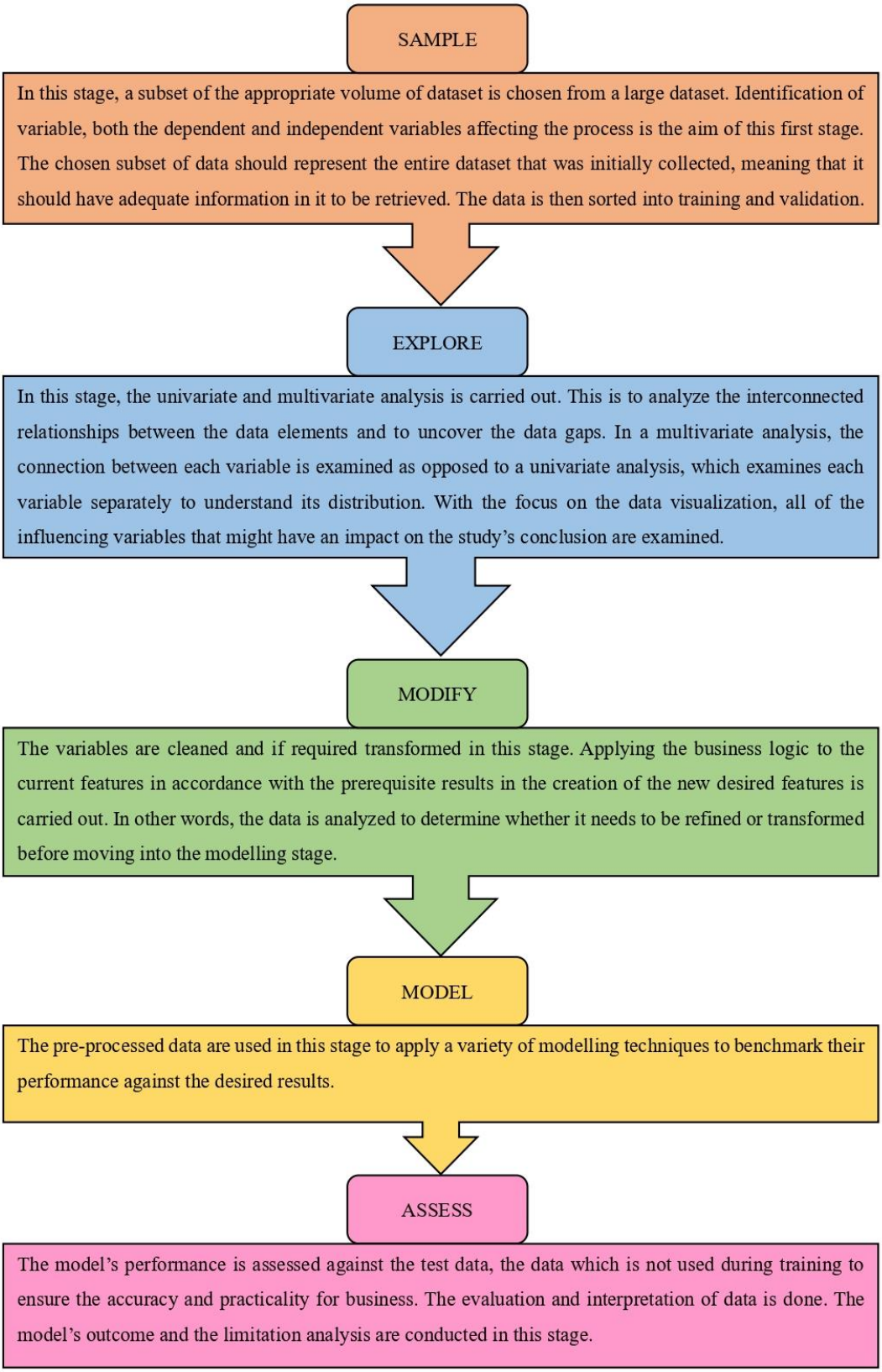


Figure 4.1.1 SEMMA

4.1 Project Setup

Setting up a new SAS Enterprise Miner project is required prior to applying the SEMMA process on the dataset. The step-by-step details include creating project, diagram, and library, and importing dataset are shown and described in the diagrams below.

1. A new SAS Enterprise Miner project is created by setting the project name "Credit Score Classification" and location of the project (Figures 7.1.1 - 7.1.4).
2. A SAS Enterprise Miner diagram workspace named "SE" is created to contain and display the steps involved in this project (Figures 7.1.5 - 7.1.6).
3. The training dataset, "train.csv" is uploaded to SAS server via SAS Studio located in SAS On Demand for Academics by selecting the "Upload" button to browse from local files. The uploaded dataset is ready for access in SAS Enterprise Miner to perform sampling and exploring techniques (Figures 7.1.7 - 7.1.11).
4. The File Import node located on the Sample tab of the toolbar is dragged to the SE diagram workspace to import the dataset from SAS server to SAS Enterprise Miner (Figures 7.1.12 - 7.1.18).
5. After importing the dataset, the Save Data node in the Utility tab is dragged to the SE diagram, then connected to the File Import node to save the dataset as a SAS table (Figures 7.1.19 - 7.1.23).
6. A library named "Dataset" is created with its path set as the location of the saved dataset to store the dataset (Figures 7.1.24 - 7.1.27).

4.2 SAMPLE

Once the project setup is completed, the dataset is ready for sampling process. The following describes the steps involved in the Sample stage, starting from creating and loading data sources to data sampling.

1. The data source is created via the Data Source Wizard by first setting the source as SAS Table (Figures 7.2.1 - 7.2.2).
2. The SAS table named "Em_save_train" located in the "Dataset" library is selected as our data source (Figures 7.2.3 - 7.2.6).
3. The information of the chosen data source like table name, number of variable and observations are displayed as shown in (Figure 7.2.7).
4. The advanced setting is used to modify the initial measurement levels and roles of the variables in the data source (Figure 7.2.8).
5. The figures below show the default setting (Figure 7.2.9) and modified setting (Figure 7.2.10) of the measurement levels and roles, respectively.

   Figure 4.2.1 below demonstrates and highlights the modifications applied to the column metadata, especially on the reclassification of variables role and level. The Customer ID and SSN variables are dropped as they are not required for exploring and modelling stage later.

Variables	Role	Level
Age	Input	Nominal
Amount_invested_monthly	Input	Nominal
Annual_Income	Input	Nominal
Changed_Credit_Limit	Input	Nominal
Credit_History_Age	Input	Nominal
Credit_Mix	Input	Nominal
Credit_Score	Input	Nominal
Credit_Utilization_Ratio	Input	Interval
Customer_ID	ID	Nominal
Delay_from_due_date	Input	Interval
ID	ID	Nominal
Interest_Rate	Input	Interval
Month	Input	Nominal
Monthly_Balance	Input	Interval
Monthly_Inhand_Salary	Input	Interval
Name	Rejected	Nominal
Num_Bank_Accounts	Input	Interval
Num_Credit_Card	Input	Interval
Num_Credit_Inquiries	Input	Interval
Num_of_Delayed_Payment	Rejected	Nominal
Num_of_Loan	Rejected	Nominal
Occupation	Input	Nominal
Outstanding_Debt	Rejected	Nominal
Payment_Behaviour	Input	Nominal
Payment_of_Min_Amount	Input	Nominal
SSN	Rejected	Nominal
Total_EMI_per_month	Input	Interval
Type_of_Loan	Text	Nominal

Variables	Role	Level
Age	Input	Nominal
Amount_invested_monthly	Input	Nominal
Annual_Income	Input	Ordinal
Changed_Credit_Limit	Input	Nominal
Credit_History_Age	Input	Nominal
Credit_Mix	Input	Ordinal
Credit_Score	Target	Ordinal
Credit_Utilization_Ratio	Input	Interval
Customer_ID	Input	Nominal
Delay_from_due_date	Input	Interval
ID	ID	Nominal
Interest_Rate	Input	Interval
Month	Input	Nominal
Monthly_Balance	Input	Interval
Monthly_Inhand_Salary	Input	Interval
Name	Rejected	Nominal
Num_Bank_Accounts	Input	Interval
Num_Credit_Card	Input	Interval
Num_Credit_Inquiries	Input	Interval
Num_of_Delayed_Payment	Input	Nominal
Num_of_Loan	Input	Nominal
Occupation	Input	Nominal
Outstanding_Debt	Input	Nominal
Payment_Behaviour	Input	Nominal
Payment_of_Min_Amount	Input	Nominal
SSN	Rejected	Nominal
Total_EMI_per_month	Input	Interval
Type_of_Loan	Input	Nominal

Figure 4.2.1 Reclassification of Variables Roles and Level

However, the level of the variables after reclassification still does not reflect the actual situation due to dirty and noisy data in the dataset. The variables level can only be classified accurately after the data cleaning process which will be applied at Modify stage in the next project phase. Figure 4.2.2 below demonstrates the correct level of variables of dataset.

Variables	Level
Age	Nominal
Amount_invested_monthly	Interval
Annual_Income	Interval
Changed_Credit_Limit	Interval
Credit_History_Age	Ordinal
Credit_Mix	Ordinal
Credit_Score	Ordinal
Credit_Utilization_Ratio	Interval
Customer_ID	Nominal
Delay_from_due_date	Interval
ID	Nominal
Interest_Rate	Interval
Month	Nominal
Monthly_Balance	Interval
Monthly_Inhand_Salary	Interval
Name	Nominal
Num_Bank_Accounts	Interval
Num_Credit_Card	Interval
Num_Credit_Inquiries	Interval
Num_of_Delayed_Payment	Interval
Num_of_Loan	Interval
Occupation	Nominal
Outstanding_Debt	Interval
Payment_Behaviour	Nominal
Payment_of_Min_Amount	Nominal
SSN	Nominal
Total_EMI_per_month	Interval
Type_of_Loan	Nominal

Figure 4.2.2 Correct Variables Level

6. The decision processing is not specified as illustrated in (Figure 7.2.11).
7. The sample dataset is not created via the Data Source Wizard (Figure 7.2.12), but instead the data sampling process is applied after defining the data source (where detailed discussion is provided at point 12).
8. No changes have been made to the name and role of the data source (Figure 7.2.13).
9. The summary of the data source is shown in (Figure 7.2.14). As observed, there are 16 nominal variables (1 ID, 13 Input, 2 Rejected), 9 interval variables with Input role and 3 ordinal variables (2 Input, 1 Target) in the dataset.
10. The data source named “EM_SAVE_TRAIN” is created successfully and dragged to the SE diagram (Figures 7.2.15 - 7.2.17).
11. The node is then renamed to “Training Data” and ready for data sampling (Figures 7.2.18 - 7.2.22).
12. Data sampling is performed by dragging the Sample node in the Sample tab (Figures 7.2.23 - 7.2.24) to the SE diagram and connecting it to the “Training Data” node. The stratified equal-sized sampling method is applied with a random seed 12345, and the size of the training dataset is set to 30% (Figures 7.2.25 - 7.2.29).

4.3 EXPLORE

In the Explore phase, data exploration is performed to identify patterns, trends, and relationship between variables, as well as discover anomalies and outliers in the dataset that require further investigation in later phase. Common Exploratory Data Analysis (EDA) techniques include summary statistics, univariate, bivariate and multivariate analysis are applied using the nodes under the Explore tab in SAS Enterprise Miner. The findings of the analyses are discussed in detail later in this section, alongside some screenshots of SAS Enterprise Miner.

4.3.1 Statistical Summary

To generate the summary statistics of the dataset, the StatExplore node is used by dragging the node to the SE diagram and connecting to the Sample node (Figures 7.3.1 - 7.3.4). Figure 4.3.1.1 below shows the summary statistics of the class variables while Figure 4.3.1.2 below shows the summary statistics of the interval variables.

Class Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN

Data Role	Variable Name	Role	Number of Levels	Missing	Mode	Mode Percentage	Mode2	Mode2 Percentage
TRAIN	Age	INPUT	513	0	25	2.95	28	2.94
TRAIN	Amount_invested_monthly	INPUT	513	19	__10000__	4.50		3.42
TRAIN	Annual_Income	INPUT	513	0	12613.56	0.61	75881.16	0.61
TRAIN	Changed_Credit_Limit	INPUT	513	0	_	3.11	5.71	0.86
TRAIN	Credit_History_Age	INPUT	404	0	NA	9.10	15 Years and 11 Months	0.53
TRAIN	Credit_Mix	INPUT	4	0	Good	32.16	Standard	29.42
TRAIN	Month	INPUT	8	0	June	12.72	July	12.65
TRAIN	Num_of_Delayed_Payment	INPUT	259	2103		7.01	10	5.18
TRAIN	Num_of_Loan	INPUT	159	0	3	14.66	2	14.36
TRAIN	Occupation	INPUT	16	0	_______	6.98	Scientist	6.53
TRAIN	Payment_Behaviour	INPUT	7	0	Low_spent_Small_value_payments	24.41	High_spent_Medium_value_payments	17.85
TRAIN	Payment_of_Min_Amount	INPUT	3	0	Yes	46.16	No	41.79
TRAIN	Type_of_Loan	INPUT	513	287		14.65	Mortgage Loan	2.40
TRAIN	Credit_Score	TARGET	3	0	Good	33.33	Poor	33.33

Figure 4.3.1.1 Class Variables Summary Statistics

Based on Figure 4.3.1.2, there are 14 class variables and out of these variables, 11 of them do not contain missing values, whereas the Amount invested monthly, Num of Delayed Payment, and Type of Loan variables have missing values. It is noticeable that the Age, Amount invested monthly, Annual Income, Changed Credit Limit, Credit History Age, Num of Delayed Payment, Num of Loan, and Type of Loan variables have extremely high number of levels due to dirty and noisy data, and large numbers of distinct values. The Age, Amount invested monthly, Annual Income, Changed Credit Limit, Credit History Age, Num of Delayed Payment, and Num of Loan variables are classified as class variables, which are supposed to be classified as interval variables. The Type of Loan variable possess high cardinality, and high dimensional data can result in a range of issues such as overfitting, low model performance and interpretability. These issues will be properly addressed and discussed further in the Modify phase.

Interval Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN

Variable	Role	Mean	Standard Deviation	Non Missing	Missing	Minimum	Median	Maximum	Skewness	Kurtosis
Credit_Utilization_Ratio	INPUT	32.28744	5.135271	30000	0	20.25707	32.3135	50	0.041792	-0.93116
Delay_from_due_date	INPUT	20.0766	15.03858	30000	0	-5	16	67	1.024543	0.430127
Interest_Rate	INPUT	77.20173	494.2043	30000	0	1	12	5775	8.630536	77.49108
Monthly_Balance	INPUT	409.4905	221.5765	29624	376	0.503582	341.2423	1566.613	1.585072	2.823866
Monthly_Inhand_Salary	INPUT	4328.088	3296.71	25470	4530	303.6454	3173.971	15204.63	1.114643	0.521351
Num_Bank_Accounts	INPUT	16.77553	117.1833	30000	0	-1	5	1798	11.2261	133.0324
Num_Credit_Card	INPUT	22.56953	131.3644	30000	0	0	5	1499	8.394718	72.82476
Num_Credit_Inquiries	INPUT	28.20697	197.0245	29439	561	0	5	2594	9.652007	97.51068
Total_EMI_per_month	INPUT	1432.826	8401.09	30000	0	0	70.58768	82236	7.04676	51.34881

Figure 4.3.1.2 Interval Variable Summary Statistics

Based on Figure 4.3.1.2, the Monthly Balance, Monthly Inhand Salary, and Num Credit Inquiries variables have missing values, while the remaining 6 variables have no missing values. It is not reasonable that the Delay from due date and the Num Bank Accounts variables contain negative values (with minimum values -5 and –1 respectively), and the Num Bank Accounts, Num Credit Card and Interest rate variables have extremely high values (with maximum values 1798, 1499 and 5775 respectively).

Ob...	ID	Cu...	M...	Na...	Age	SSN	O...	Ann...	Mont...	Nu...	Num_...	Intere...	Nu...	Type...	Dela...	Nu...	Ch...	Nu...	Cred...	Outst...	Credi...	Cre...	Pay...	Total...	Amo...	Paym...	Monthl...	Credit_S...
1	0x1602	CU...	Jan...	Aaro...	23	821-0...	Scie...	19114...	1824...	3	4	34	Auto L...	37	11.27	4_	809.98	26.822...	22 Ye...	No	49.57...	80.415...	High_sp...	312.494...	Good
2	0x1603	CU...	Feb...	Aaro...	23	821-0...	Scie...	19114...	.	3	4	34	Auto L...	-1	11.27	4	Good	809.98	31.944...	NA	...	No	49.57...	118.28...	Low_sp...	284.629...	Good
3	0x1604	CU...	Mar...	Aaro...	-500	821-0...	Scie...	19114...	.	3	4	34	Auto L...	37	_	4	Good	809.98	28.609...	22 Ye...	No	49.57...	81.699...	Low_sp...	331.209...	Good
4	0x1605	CU...	Apri...	Aaro...	23	821-0...	Scie...	19114...	.	3	4	34	Auto L...	54	6.27	4	Good	809.98	31.377...	22 Ye...	No	49.57...	199.45...	Low_sp...	223.451...	Good
5	0x1606	CU...	May...	Aaro...	23	821-0...	Scie...	19114...	1824...	3	4	34	Auto L...	6	11.27	4	Good	809.98	24.797...	22 Ye...	No	49.57...	41.420...	High_sp...	341.48...	Good
6	0x1607	CU...	Jun...	Aaro...	23	821-0...	Scie...	19114...	.	3	4	34	Auto L...	84	9.27	4	Good	809.98	27.262...	22 Ye...	No	49.57...	62.430...	!@9#%8...	340.479...	Good
7	0x1608	CU...	July...	Aaro...	23	821-0...	Scie...	19114...	1824...	3	4	34	Auto L...	38_	11.27	4	Good	809.98	22.537...	22 Ye...	No	49.57...	178.34...	Low_sp...	244.565...	Good
8	0x1609	CU...	Aug...	...	23	#F%$...	Scie...	19114...	1824...	3	4	34	Auto L...	36	11.27	4	Good	809.98	23.93...	NA	...	No	49.57...	24.785...	High_sp...	358.124...	Standard
9	0x160e	CU...	Jan...	Rick ...	28_	004-0...	___...	34847...	3037...	2	4	61	Credit-...	34	5.42	2	Good	605.03	24.464...	26 Ye...	No	18.81...	104.29...	Low_sp...	470.690...	Standard
10	0x160f	CU...	Feb...	Rick ...	28	004-0...	Tea...	34847...	3037...	2	4	61	Credit-...	71	7.42	2	Good	605.03	38.550...	26 Ye...	No	18.81...	40.391...	High_sp...	484.591...	Good
11	0x1610	CU...	Mar...	Rick ...	28	004-0...	Tea...	34847...	3037...	2	1385	61	Credit-...	3-1	5.42	2_	605.03	33.224...	26 Ye...	No	18.81...	58.515...	High_sp...	466.466...	Standard
12	0x1611	CU...	Apri...	Rick ...	28	004-0...	Tea...	34847...	.	2	4	61	Credit-...	33_	5.42	2	Good	605.03	39.182...	26 Ye...	No	18.81...	99.306...	Low_sp...	465.676...	Good
13	0x1612	CU...	May...	Rick ...	28	004-0...	Tea...	34847...	3037...	2	4	61	Credit-...	31	6.42	2	Good	605.03	34.977...	26 Ye...	No	18.81...	130.11...	Low_sp...	444.867...	Good
14	0x1613	CU...	Jun...	Rick ...	28	004-0...	Tea...	34847...	3037...	2	4	61	Credit-...	30	5.42	2	Good	605.03	33.38...	27 Ye...	No	18.81...	43.477...	High_sp...	481.505...	Good
15	0x1614	CU...	July...	Rick ...	28	004-0...	Tea...	34847...	.	2	4	61	Credit-...	34	5.42	2	Good	605.03	31.131...	27 Ye...	NM	18.81...	70.101...	High_sp...	464.880...	Good
16	0x1615	CU...	Aug...	Rick ...	28	004-0...	Tea...	34847...	3037...	2	4	61	Credit-...	34	5.42	2	Good	605.03	32.933...	27 Ye...	No	18.81...	218.90...	Low_sp...	356.078...	Good
17	0x161a	CU...	Jan...	Lang...	34	486-8...	___...	14316...	121...	1	5	83	Auto L...	58	7.1	3	Good	1303.01	28.616...	17 Ye...	No	246.9...	168.41...	!@9#%8...	1043.31...	Good
18	0x161b	CU...	Feb...	...	34	486-8...	Eng...	14316...	121...	1	5	83	Auto L...	136	7.1	3	Good	1303.01	41.702...	17 Ye...	No	246.9...	232.86...	High_sp...	998.869...	Good
19	0x161c	CU...	Mar...	Lang...	34	486-8...	___...	14316...	.	1	5	83	Auto L...	87	11.1	_	Good	1303.01	26.519...	17 Ye...	No	246.9...	__100...	High_sp...	715.741...	Good
20	0x161d	CU...	Apri...	Lang...	34	486-8...	Eng...	14316...	121...	1	5	83	Auto L...	85	9.1	3_	1303.01	39.501...	NA	...	No	246.9...	825.21...	Low_sp...	426.513...	Good
21	0x161e	CU...	May...	Lang...	34	486-8...	___...	14316...	121...	1	5	83	Auto L...	105	7.1	3	Good	1303.01	31.376...	18 Ye...	No	246.9...	430.94...	Low_sp...	810.782...	Good
22	0x161f	CU...	Jun...	Lang...	34	486-8...	Eng...	14316...	121...	1	5	8967	Auto L...	86	7.1	3	Good	1303.01	39.783...	18 Ye...	No	246.9...	257.80...	High_sp...	963.921...	Good
23	0x1620	CU...	July...	...	34	486-8...	Eng...	14316...	121...	1	5	83	Auto L...	86	7.1	3	Good	1303.01	38.068...	18 Ye...	No	246.9...	263.17...	High_sp...	968.555...	Standard
24	0x1621	CU...	Aug...	Lang...	34	486-8...	Eng...	14316...	121...	1	5	83	Auto L...	86	7.1	3	Good	1303.01	38.374...	18 Ye...	No	246.9...	__100...	High_sp...	895.494...	Standard
25	0x1626	CU...	Jan...	Jaso...	54	072-3...	Entr...	30689...	2612...	2	5	41	Not Sp...	06	1.99	4	Good	632.46	26.544...	17 Ye...	No	16.41...	81.228...	Low_sp...	433.60...	Standard
26	0x1627	CU...	Feb...	Jaso...	54	072-3...	Entr...	30689...	2612...	2	5	41	Not Sp...	53	1.99	4	Good	632.46	35.279...	17 Ye...	No	16.41...	124.88...	Low_sp...	409.951...	Standard
27	0x1628	CU...	Mar...	Jaso...	55	072-3...	Entr...	30689...	2612...	2	5	41	Not Sp...	39	1.99	4	Good	632.46	32.301...	17 Ye...	NM	16.41...	83.406...	High_sp...	411.427...	Standard
28	0x1629	CU...	Apri...	Jaso...	55	072-3...	Entr...	30689...	2612...	2	5	41	Not Sp...	76	-2.01	4	Good	632.46	38.132...	17 Ye...	No	16.41...	272.33...	Low_sp...	262.499...	Standard
29	0x162a	CU...	May...	Jaso...	55	072-3...	Entr...	30689...	2612...	2	5	41	Not Sp...	56	-1.01	4	Good	632.46	41.154...	17 Ye...	No	16.41...	__100...	Low_sp...	359.374...	Standard
30	0x162b	CU...	Jun...	Jaso...	55	#F%$...	___...	30689...	2612...	2	5	41	Not Sp...	56	-3.01	4_	632.46	27.445...	17 Ye...	No	16.41...	84.952...	High_sp...	419.880...	Standard
31	0x162c	CU...	July...	Jaso...	55	072-3...	Entr...	30689...	2612...	2	5	41	Not Sp...	5	1.99	4	Good	632.46	26.056...	17 Ye...	No	16.41...	71.283...	Low_sp...	443.549...	Standard
32	0x162d	CU...	Aug...	Jaso...	55	072-3...	Entr...	30689...	2612...	2	5	4-100	Not Sp...	49	1.99	4	Good	632.46	27.332...	17 Ye...	No	16.41...	125.61...	High_sp...	379.216...	Standard
33	0x1632	CU...	Jan...	Dee...	21	615-0...	Dev...	35547...	2853...	7	5	50	...	5	2.58	4	Stand...	943.86	39.797...	30 Ye...	Yes	0276.72...	!@9#%8...	288.605...	Standard
34	0x1633	CU...	Feb...	Dee...	21	615-0...	Dev...	35547...	.	7	5	50	...	9	2.58	4	Stand...	943.86	27.020...	30 Ye...	NM	074.443...	High_sp...	460.887...	Standard
35	0x1634	CU...	Mar...	Dee...	21	615-0...	Dev...	35547...	2853...	7	5	5-100	...	512	2.58	4	Stand...	943.86	23.462...	30 Ye...	Yes	0173.13...	Low_sp...	392.192...	Standard
36	0x1635	CU...	Apri...	Dee...	21	615-0...	Dev...	35547...	2853...	7	5	50	...	115	2.58	4_	943.86	28.924...	30 Ye...	Yes	096.785...	High_sp...	438.545...	Standard
37	0x1636	CU...	May...	Dee...	21	615-0...	Dev...	35547...	2853...	7	5	50	...	917	2.58	4_	943.86	41.776...	31 Ye...	Yes	062.723...	High_sp...	482.607...	Standard
38	0x1637	CU...	Jun...	Dee...	21	615-0...	Dev...	35547...	.	7	5	50_	...	515	2.58	4	Stand...	943.86	29.217...	31 Ye...	Yes	037.643...	High_sp...	497.687...	Standard

Figure 4.3.1.3 Overview of Data

Based on Figure 4.3.1.3, some of the numerical variables such as the Age variable consists of '_' character, which makes it unable to classify as interval variable, but instead character variable. In fact, the Credit History Age variable consists of missing values, but the missing values are labelled as 'NA' that is not recognized as missing values in SAS Enterprise Miner. Furthermore, the Payment of Min Amount column has a new label besides the 'Yes' and 'No' label, and the Credit Mix column also contains '_' character in the data. All missing values, noisy and inconsistent data will be cleaned and transformed in the next phase.

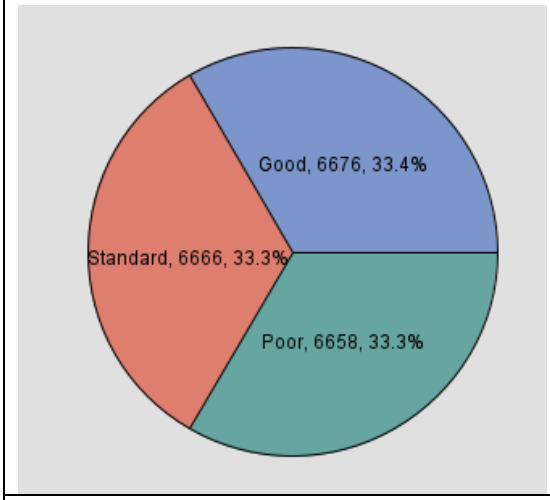
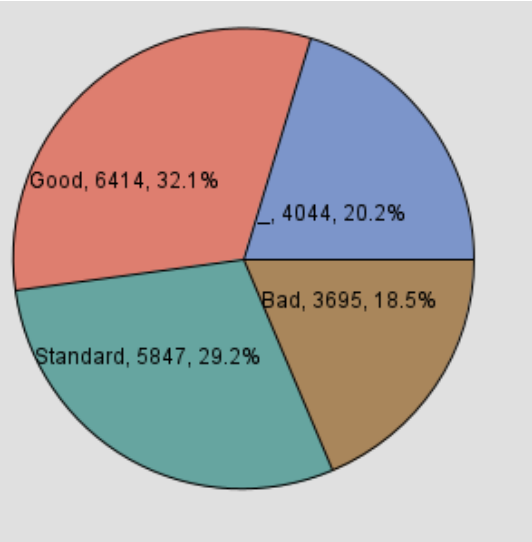
4.3.2 Graph Visualization

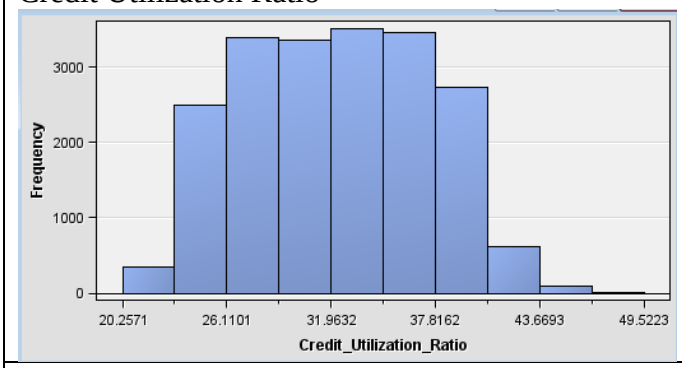
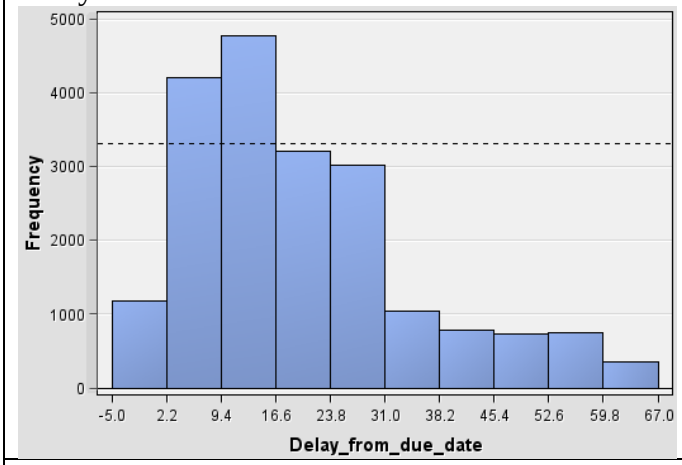
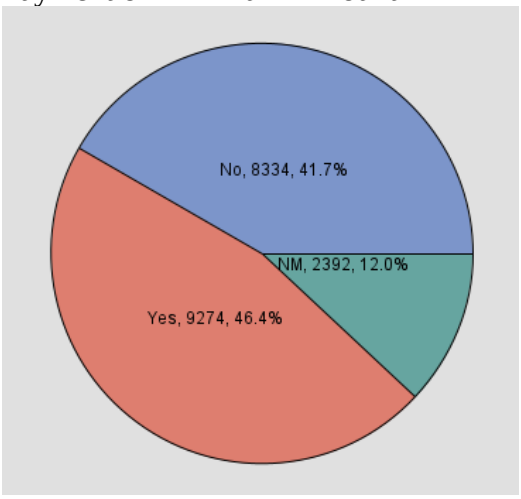
In the graph visualization, univariate, bivariate and multivariate analysis are performed. The data exploration will not reflect with the original dataset as the proportion is adjusted based on sampling set. The GraphExplore, StatExplore, Multiplot, and Variable Clustering nodes are used by dragging the nodes to the SE diagram and connecting to the Sample node (shown in Appendix).

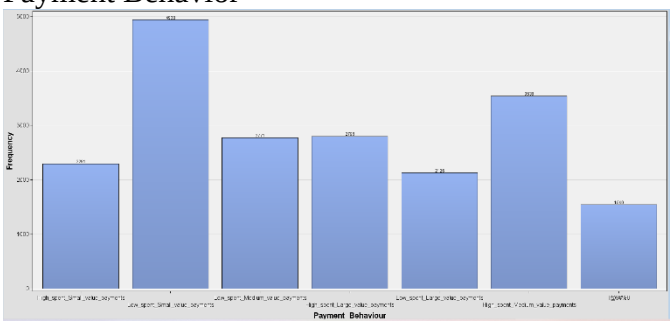
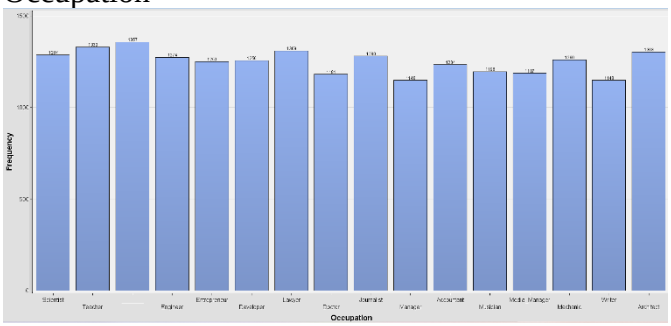
4.3.2.1 Univariate Analysis

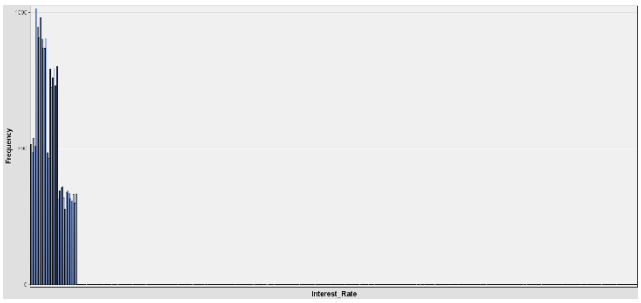
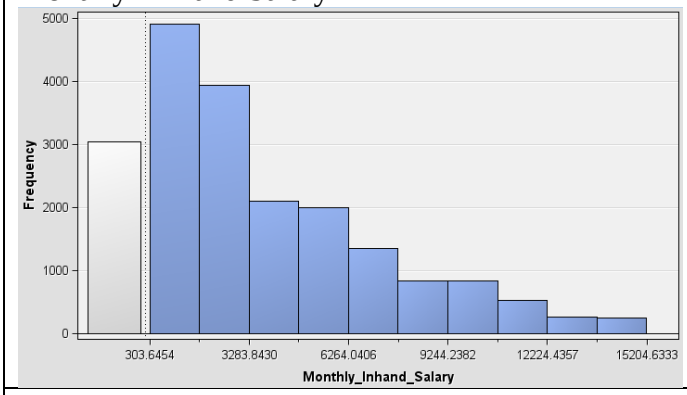
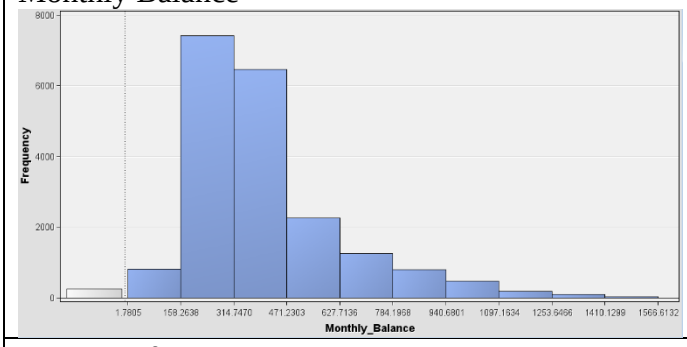
Univariate analysis is conducted by analyzing a single variable in the data using common techniques such as histogram, box plot etc. This analysis aims to describe the data, compare the spread of the variable as well as to find patterns and anomalies in the data.

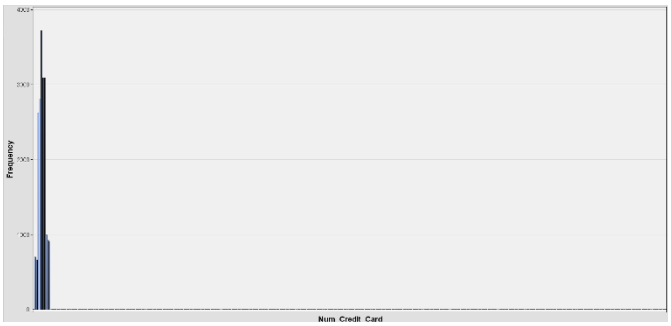
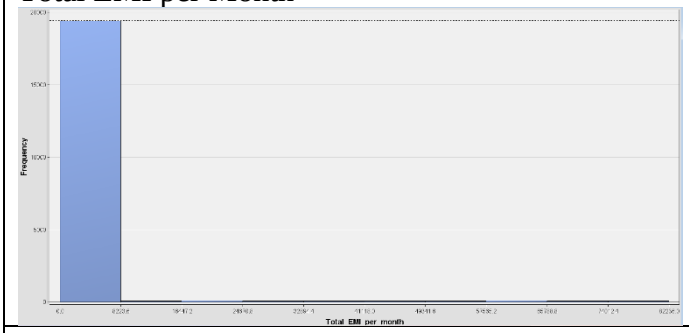
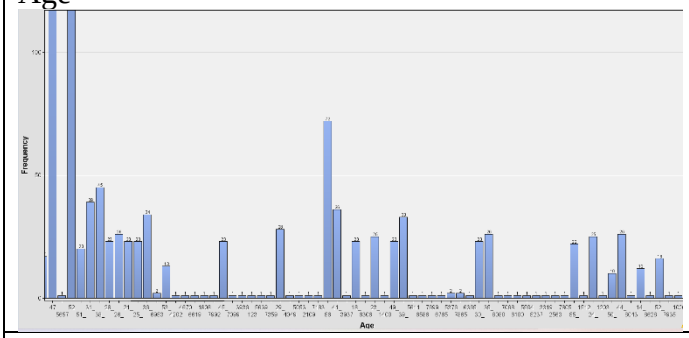
In Figure 4.3.2.1 below, pie charts, bar graphs and histograms are presented to describe the patterns of each variable of the dataset.

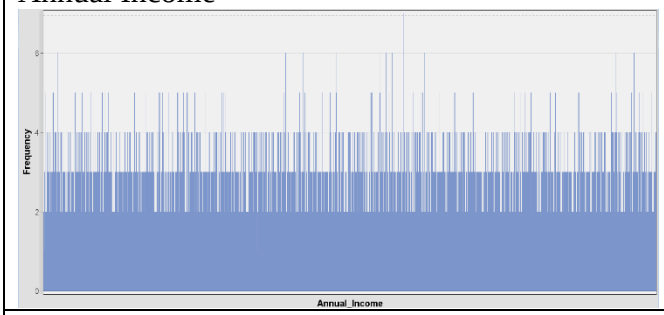
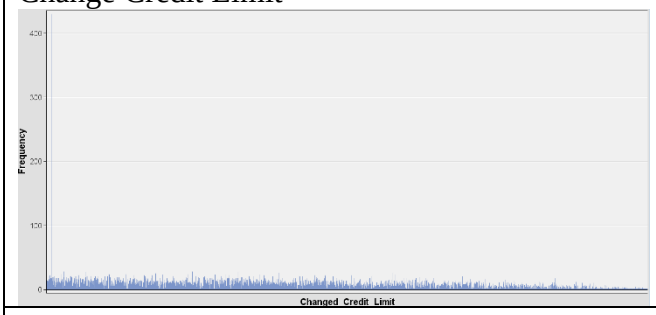
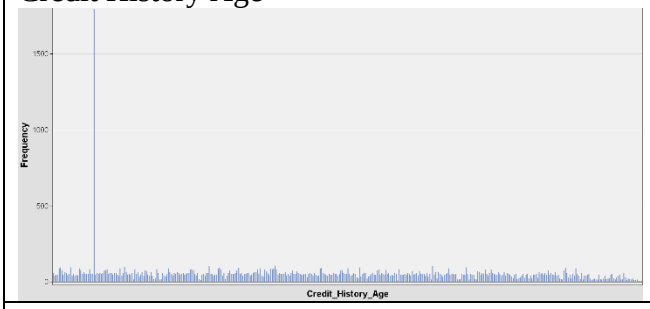
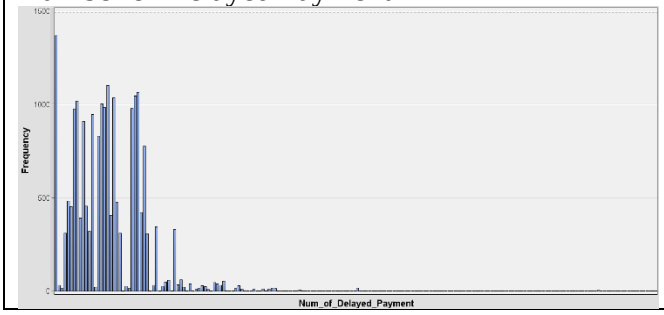
Variables	**Findings**
Credit Score 	Class target variable which uses stratified sampling and break down into 3 categories with different proportions and percentage: • Good – 6,676 (33%) • Standard – 6,666 (34%) • Poor – 6,658 (33%)
Credit Mix 	In credit mix variable, there are 4 categories with different proportions and percentage as below: • Standard – 5,847 (29%) • Good – 6,414 (32%) • Missing value – 4,044 (20%) • Bad – 3,695 (19%)

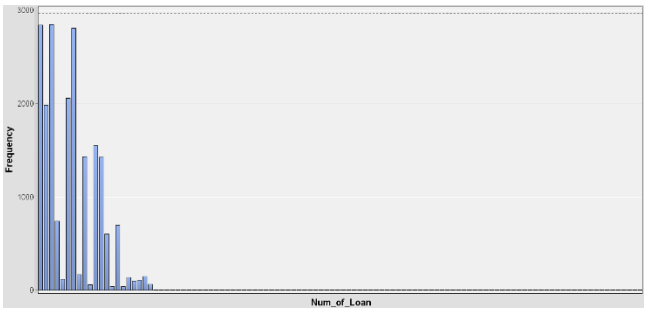
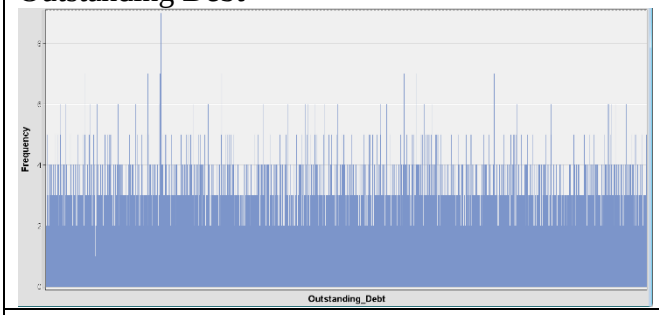
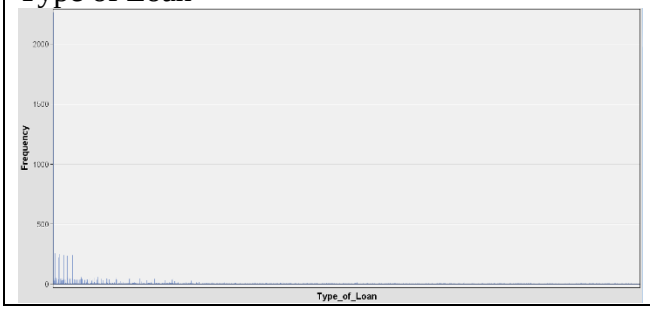
Credit Utilization Ratio 	Based on credit utilization ratio variable, most borrowers' credit utilization ratio falls between 26.11 - 37.82, followed by the range between 23.18 - 26.11, and range between 37.82 - 40.74, with 20.23 as the lowest and 49.5 as the highest ratio.
Delay from Due Date 	Most borrowers delay their monthly instalment for 9 - 17 days (about 2 and a half weeks), followed by 2- 9 days and 17 - 24 days (about 3 and a half weeks) from their due date. The maximum day from delaying their payment is 67 days (about 2 months), which is more than 2 months. The minimum days is -5 days which is unusual, as this may indicate that borrowers make advanced payment.
Payment of Minimum Amount 	Categorized into 3 categories: • Made minimum amount payment 9,274 (46%) • Did not make minimum amount payment 8,334 (42%) • Not mentioned 2,392 (12%)

Payment Behavior 	Payment behavior is a breakdown into spending power and value of payments categories. The categories are ranked from highest to lowest frequency as below: • Low spending with small value payment 4,938 (25%) • High spending with medium value payment 3,538 (18%) • High spending with large value payment 2,796 (14%) There are 1,540 invalid values found in this variable.
Occupation 	The occupation across the borrowers is fairly distributed, with the occupation types ranked from highest to lowest frequency: • Teacher 1,330 (7%) • Lawyer 1,309 (7%) • Architect 1,303 (7%) There are unknown values amounting to 1,357 (7%) of the dataset.
Number of Bank Accounts 	The variable shows that mostly the number of bank accounts a borrower has lies between 1-10 accounts. Unusual pattern detected in the variable. Reason being are due to: • Dirty data is found, which the data type is detected as character, rather than numeric. • Noisy data is detected with 835 invalid values showing 0 number of accounts, which is unusual as each borrower must have at least 1 bank account to obtain a loan.

Interest Rate 	The variable shows that the interest rate mostly falls between 0-34%, with 8% as the most common interest rate for 1,081 borrowers. Unusual pattern detected in the variable. Reason being are due to: • Noisy data detected with maximum interest rate of 5,775% shown which is abnormal as the maximum value for interest rate should be 100%.
Monthly In Hand Salary 	• Right-skewed distribution with 4,898 borrowers (24%) having a monthly in hand salary between $303 and $1,794, and 3,936 (20%) having salary in between $1,793 and $3,284. • 3,042 missing values detected in this variable.
Monthly Balance 	• 7,420 borrowers (37%) fall in the range of $158 and $315 monthly balance, followed by 6,466 borrowers (32%) with monthly balance between $315 and $471. • 255 missing values are found in the variable.
Number of Credit Inquiries 	• Highly skewed distribution. • 2,378 values fall under 4 credit enquiries. • There are 1,536 and 379 with no enquiries and missing values. • Outliers detected as with maximum value of 2,592 credit inquiries, which is unusual.

Number of Credit Cards 	• Highly skewed distribution. • 3,726 borrowers (19%) own 5 credit cards. • Most borrowers own between 0 and 10 credit cards. • Outliers detected with 1,498 as the maximum number of credit cards owned by a single borrower, which is very unusual.
Total EMI per Month 	• Highly skewed distribution. • 19,397 of the Estimated monthly instalment (EMI) records (97%) lie between $0 and $8224. • Outliers detected with a maximum value of $82,236 EMI.
Age 	Unusual pattern in the data due to dirty data as the data type is detected as character, rather than numeric. Data cleaning is required to reflect the accuracy of the age distribution.
Amount Invested Monthly 	Unusual pattern in the data due to dirty data as the data type is detected as character, rather than numeric. Data cleaning is required to reflect the accuracy of the amount invested monthly distribution.

Annual Income 	Unusual pattern in the data due to dirty data as the data type is detected as character, rather than numeric. Data cleaning is required to reflect the accuracy of the annual income distribution.
Change Credit Limit 	Unusual pattern in the data due to dirty data as the data type is detected as character, rather than numeric. Data cleaning is required to reflect the accuracy of the change credit limit distribution.
Credit History Age 	Unusual pattern in the data due to dirty data as the data type is detected as character, rather than numeric. Data cleaning is required to reflect the accuracy of the credit history age distribution.
Number of Delayed Payment 	Unusual pattern in the data due to dirty data as the data type is detected as character, rather than numeric. 1,416 records shown as missing values. Data cleaning is required to reflect the accuracy of the number of delayed payment distribution.

Number of Loan 	Unusual pattern in the data due to dirty data as the data type is detected as character, rather than numeric. 978 is shown as the maximum value of loan, which is unusual and inaccurate as the number of loans is not arranged according to numerical order. Data cleaning is required to reflect the accuracy of the number of loan distribution.
Outstanding Debt 	Unusual pattern in the data due to dirty data as the data type is detected as character, rather than numeric. Data cleaning is required to reflect the accuracy of the outstanding debt distribution.
Type of Loan 	Unusual pattern in the data due to dirty data as the data type is detected as character, rather than numeric. Data cleaning is required to reflect the accuracy of the loan type distribution.

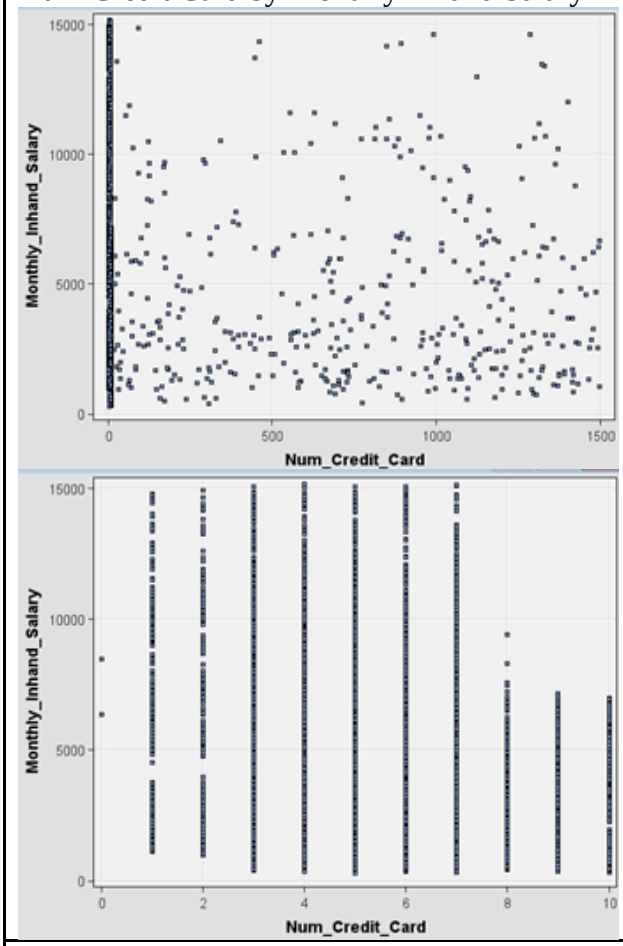
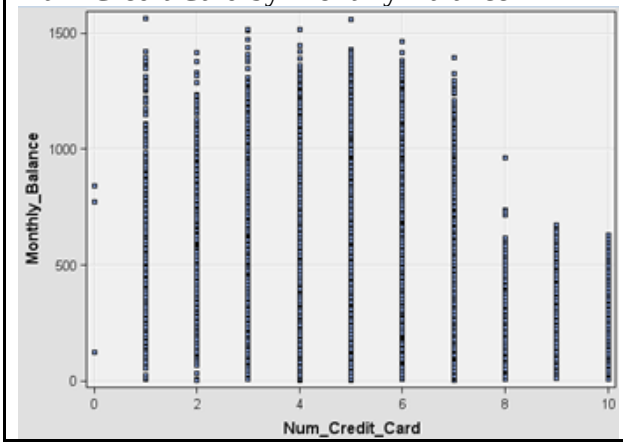
4.3.2.2 Bivariate Analysis

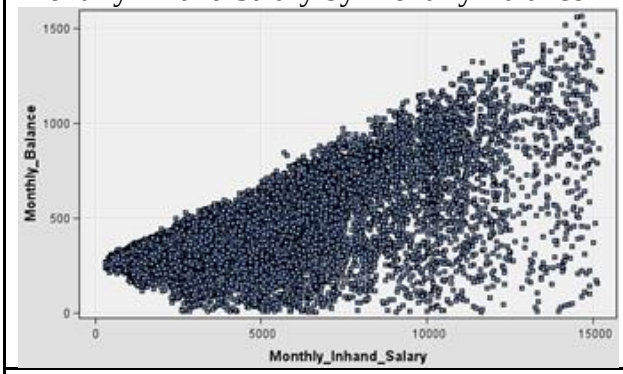
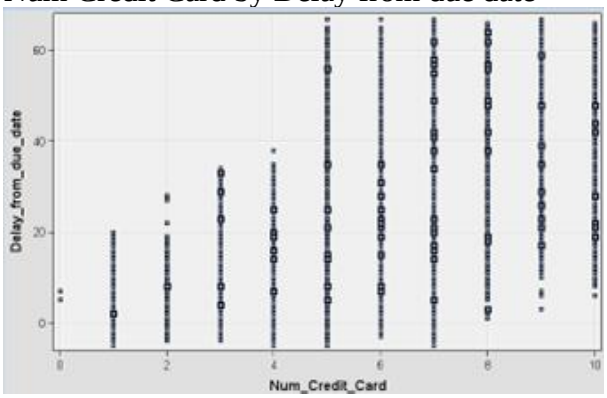
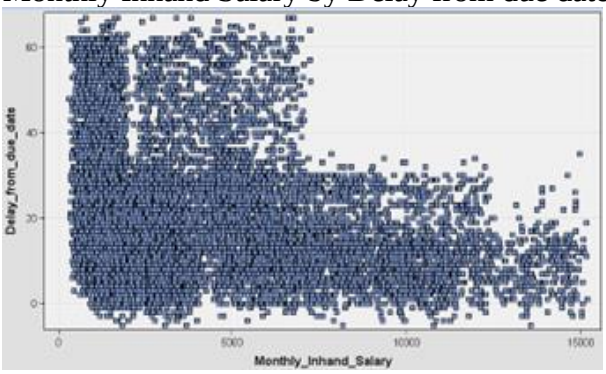
Unlike univariate analysis, bivariate analysis is a statistical method used to examine the relationship between two variables by analyzing the degree of association or correlation between them and identifying patterns or trends in the data. In this project, we focus on exploring the relationship between two input variables and examining how each continuous input variable is connected to the target variable with scatter plots and bar plots to illustrate these relationships. By conducting this bivariate analysis, valuable insights on how the variables are related can be obtained, thus informed decisions can be made based on the findings.

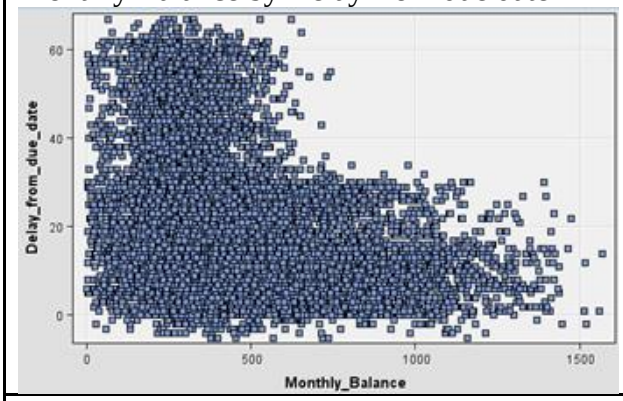
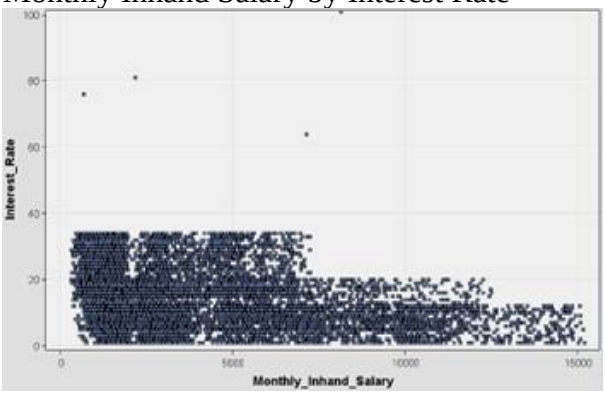
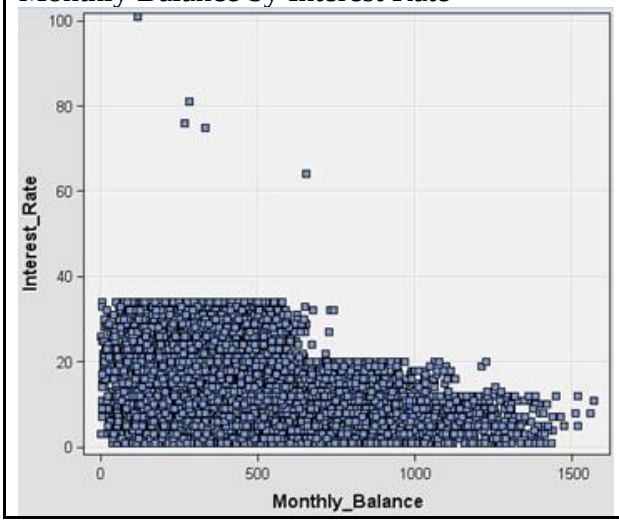
4.3.2.2.1 Bivariate Analysis on Input Variables

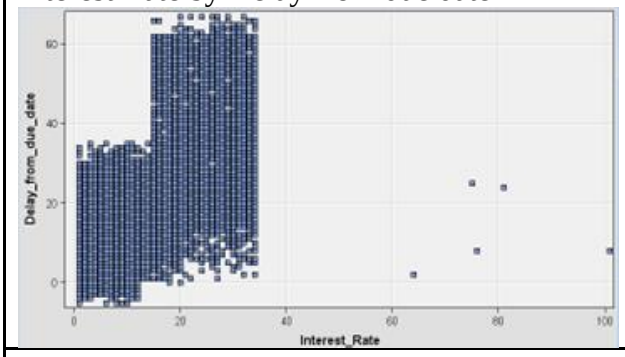
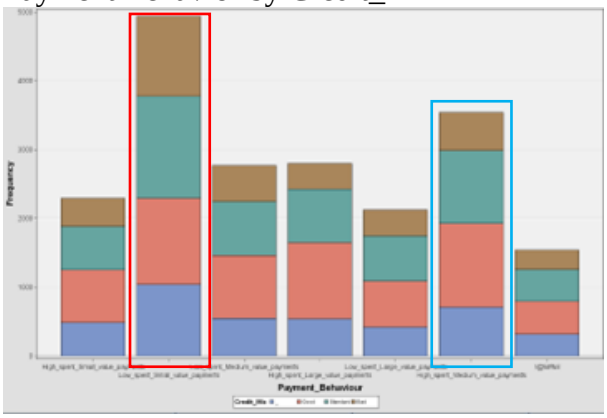
Identifying relationships between features is crucial in selecting pertinent input variables for a machine learning model, guiding feature engineering efforts, improving model performance, as well as interpreting the analysis results for better decision making.

Figure 4.3.2.2.1 below displays and describes the connection between two input variables in the dataset.

Bivariate Analysis Graphs	**Findings**
Num Credit Card by Monthly Inhand Salary 	Based on the first plot, it is observed that the Num Credit Card variable contains a lot of outliers, and most customers fall in the range of 0 to 10 credit cards. Focusing on customers with credit cards less than or equal to 10, there is a huge drop in the monthly in-hand salary for customers owned more than 7 credit cards.
Num Credit Card by Monthly Balance 	It is clearly seen that the monthly balance experienced a huge fall for customers owned more than 7 credit cards, showing a similar trend described in Num Credit Card and Monthly Inhand Salary. This indicates there may be a strong connection between monthly in-hand salary and monthly balance, but more study is needed to assure their relationship.

Monthly Inhand Salary by Monthly Balance 	It is proven that the higher the monthly in-hand salary, the higher the monthly balance of the customer.
Num Credit Card by Delay from due date 	Customers tend to pay their credit card debt later if they have more credit cards. The number of days delayed from the payment due date increases sharply if the customer holds more than 4 credit cards. This is reasonable as taking multiple credit cards can make debt repayments unsustainable. Since there is some connection between the number of credit cards and delay from due date, and number of credit card and monthly in-hand salary, then it is worth to investigate the relationship between monthly in-hand salary and delay from due date.
Monthly Inhand Salary by Delay from due date 	It is observed that the number of customers with monthly in-hand salary more than 2000 and pay loan later than 40 days from due date is reduced, and all customers with monthly in-hand salary greater than 7000 settled their debt in less than 40 days from due date. Customers with higher monthly in-hand salary pay their debt earlier, as compared to those with lower monthly in-hand salary.

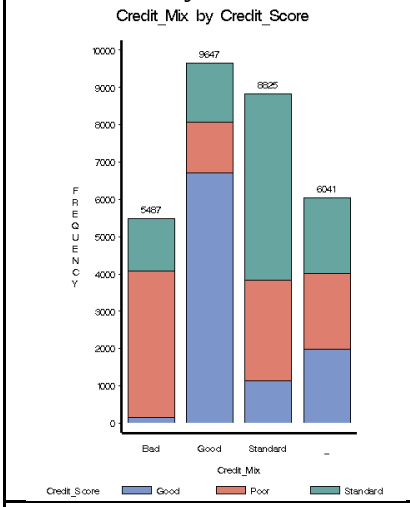
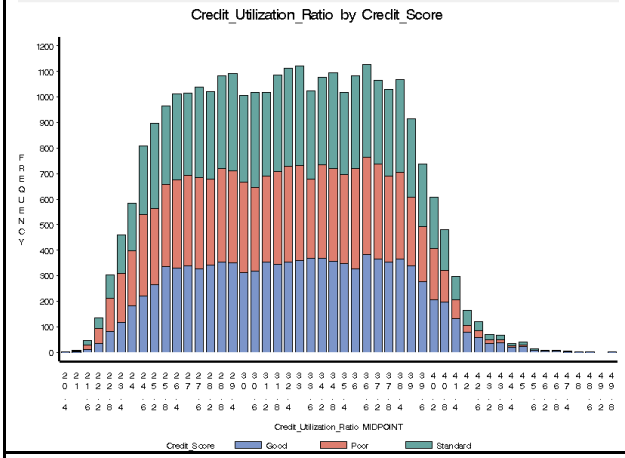
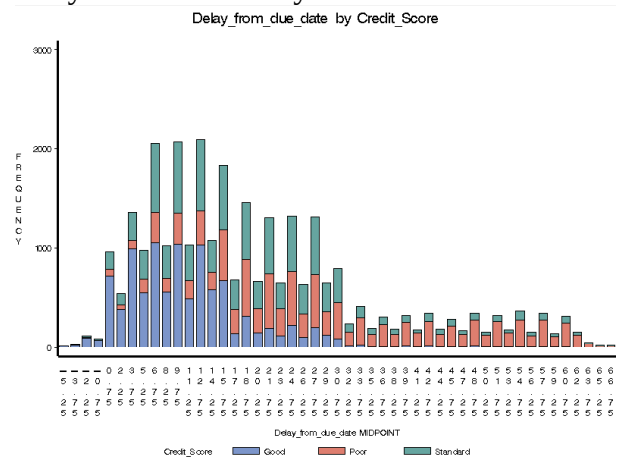
Monthly Balance by Delay from due date 	Similar pattern is observed in this plot as compared to the monthly in-hand salary and delay from due date scatter plot. There is also a huge drop in the number of customers who make payment 40 days after the due date if their monthly balance is more than 500.
Monthly Inhand Salary by Interest Rate  Monthly Balance by Interest Rate 	Similar patterns are observed in these plots as compared to the monthly in-hand salary and delay from due date scatter plot. The interest rate falls by half for customers with monthly in-hand salary more than 7000 and monthly balance more than 500.

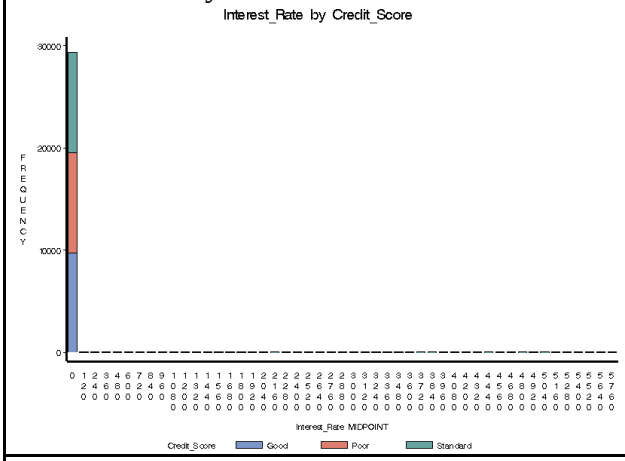
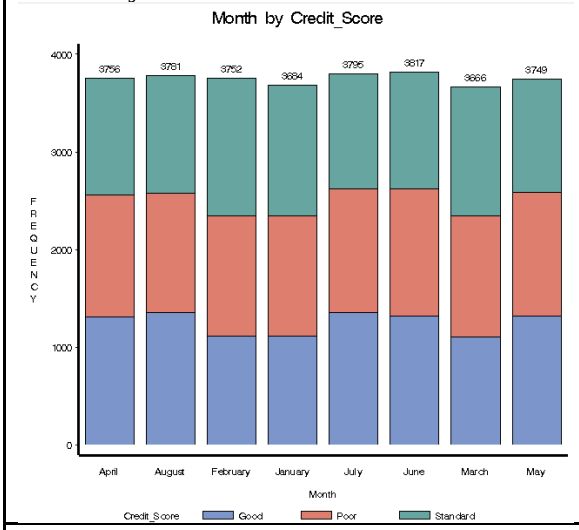
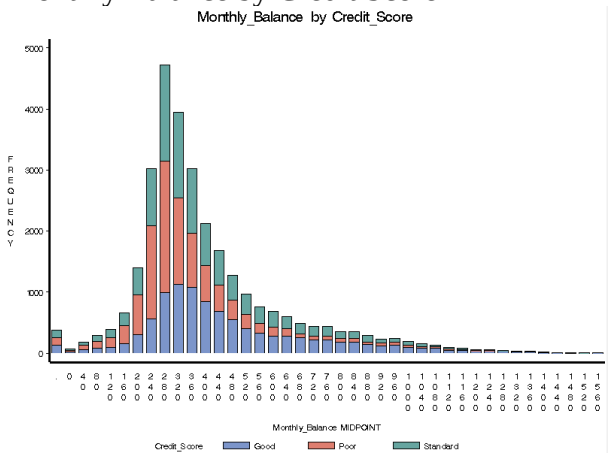
Interest Rate by Delay from due date 	Customers with higher interest rates on their credit cards are suffering from paying their credit card debt as those with credit card interest higher than 20% delayed their payment twice times more than those with credit card interest less than 20%.
Payment of Min Amount by Credit Mix 	Most customers who paid only minimum monthly payments on their credit card have standard to low credit mix score while customers who made payment more than required minimum amount or full amount have standard to good score.
Payment Behavior by Credit_Mix 	The proportion of customers with poor credit mix score is significantly higher if their payment behavior is low spent and small value payment (red box), compared to customers with low spent and large value payment (blue box). Therefore, it is said that credit card payment is one of the essential elements affecting credit mix score.

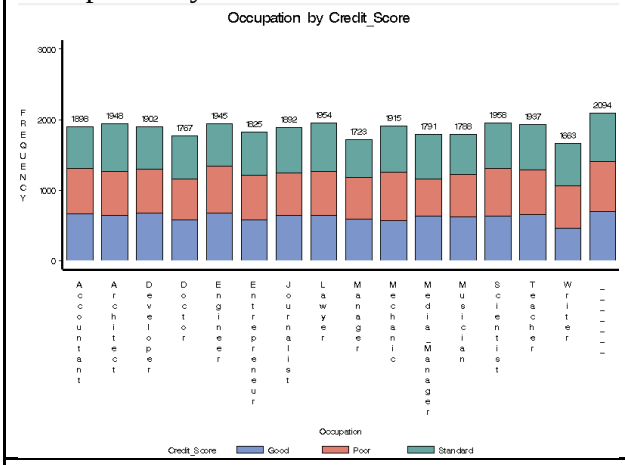
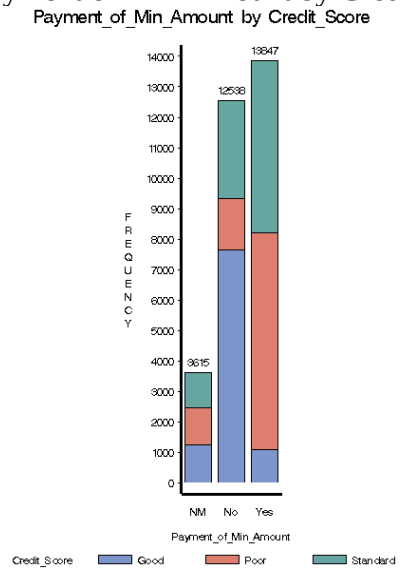
4.3.2.2.2 Bivariate Analysis on Input and Target Variables

Bivariate analysis serves as one of the tools for determining whether there is a correlation between the input features and the target variable. This is vital in a predictive modelling context as the strength of the relationship will significantly impact the accuracy and interpretation of the model. A feature with strong positive or negative correlation to the target variable is treated as a dependable predictor, and thus it is worth to incorporate into the model training stage to boost the accuracy. Conversely, a feature with weak to no correlation is not useful to the predictive model as it will not enhance the model accuracy.

Figure 4.3.2.2.2 below illustrates and explains the relationship between input variables and target variable.

Bivariate Analysis Graphs	Findings
Credit Mix by Credit Score 	The credit mix is related to credit score as shown in the graph where most customers with good credit mix also obtained good credit score.
Credit Utilization Ratio by Credit Score 	The credit utilization ratio shows a normal distribution with almost equal portion of good, standard, and poor credit score at most of the bins.
Delay from due date by Credit Score 	The proportion of customers with standard and good credit score reduced, and the number of customers with poor credit score rose when the number of days delayed from due date increased. This indicates that customers with good credit scores pay their credit card debt much earlier than those with poor credit score.

Interest Rate by Credit Score 	The bar plot did not show any meaningful results due to outliers in interest rate.
Month by Credit Score 	There is no significant difference in the amount of customers in various credit scores across months.
Monthly Balance by Credit Score 	The plot illustrates a right-skewed normal distribution with a small spike at monthly balance less than 0. Customers with poor credit scores are relatively high when the monthly balance is closer to mean, but the number of customers drops drastically when the monthly balance gets higher. This shows the connection between monthly balance and credit score although it is not that obvious as compared to other plots.

Occupation by Credit Score 	There is no obvious difference in the distribution of credit score in each segment of the customers' occupation.
Payment of Min Amount by Credit Score 	The number of customers with standard to poor credit scores and paying only minimum payment is significantly higher. Hence, payment of minimum amount is also directly affecting the credit score.

4.3.2.3 Multivariate Analysis

Multivariate analysis is a statistical method that studies simultaneous effect of more than two independent variables and the relationships among the datasets.

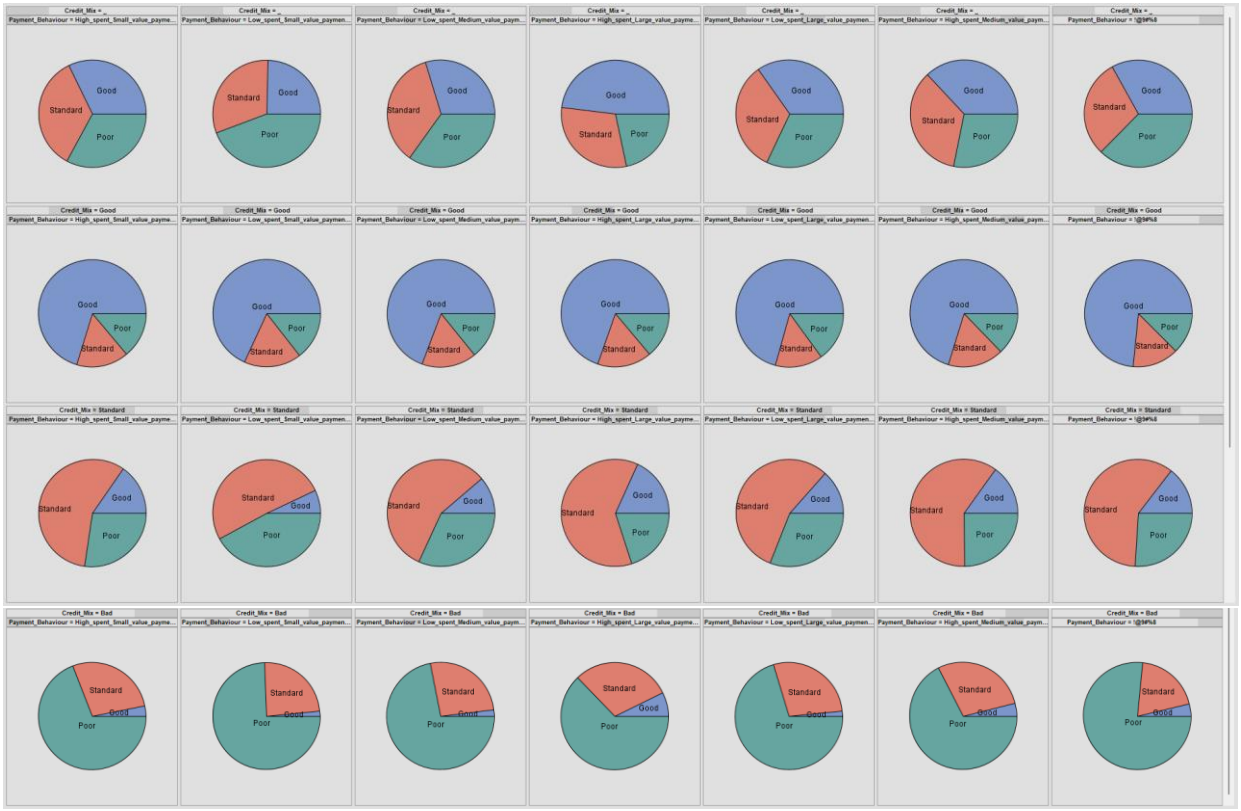


Figure 4.3.2.3.1 Lattice Pie Charts

The Lattice Bar Charts show how the categories Payment_Behaviour, and Credit_Mix, correlate with Credit_Score. Generally, having a good Credit_Mix gives a good Credit_Score, standard Credit_Mix gives a standard Credit_Score, and poor Credit_Mix gives a poor Credit_Score regardless of the Payment_Behaviour. When the Credit_Mix is undefined, High_spent Payment_Behaviour does not really affect the Credit_Score unless Large_value_payments are made, which increases the likelihood of a good Credit_Score. Low_spent Payment Behaviour also does not affect Credit_Score unless Small_value_payments are made, which results in poor Credit_Score.

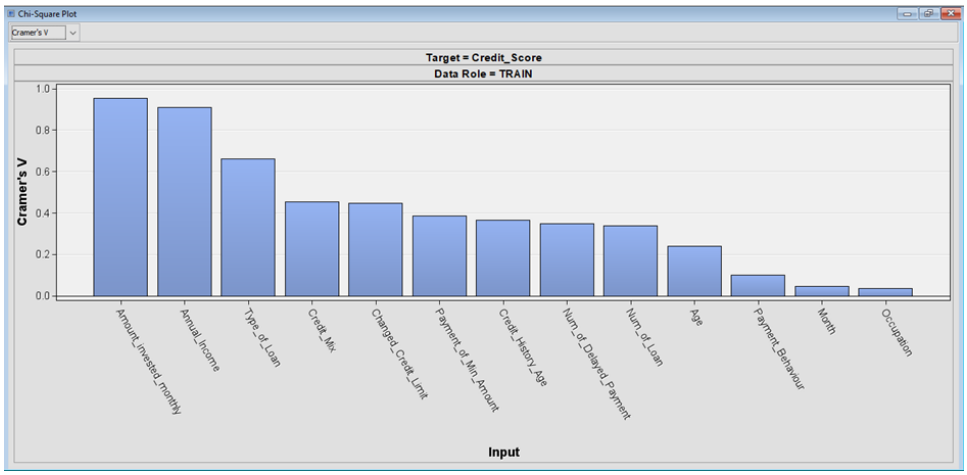


Figure 4.3.2.3.2 Cramer's V Statistics with respect to the Target Variable

Figure 4.3.2.3.2 shows Cramer's V measure in which the 0 value indicates no association whereas the 1 indicates perfect association between the two variables. The threshold value is set to 0.1. Thus, Amount invested monthly, Annual Income, Type of Loan, Credit Mix, Changed Credit Limit, Payment of Min Amount, Credit History Age, Num of Delayed Payment, Num of Loan, Age, and Payment Behaviour show association with the target variable which is the Credit Score. Meanwhile, Month and Occupation show nearly no association with Credit Score.

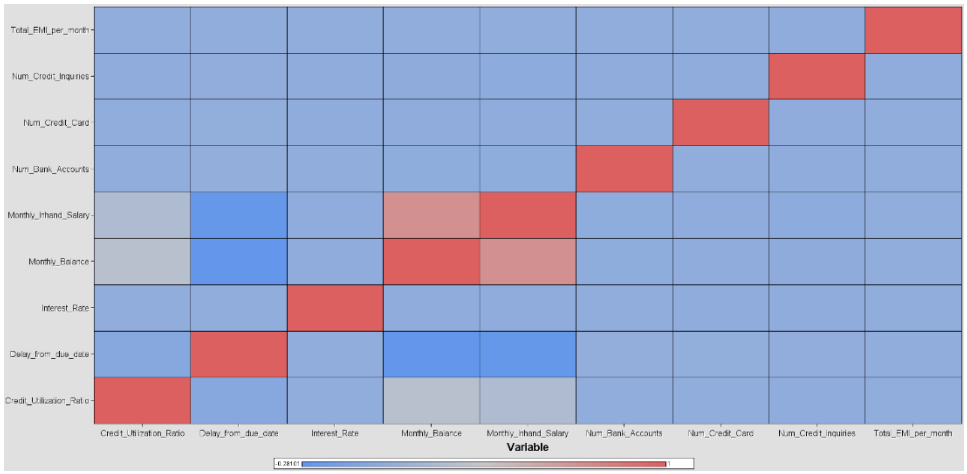


Figure 4.3.2.3.3 Correlation Matrix

Table 4.3.2.3 Correlation Table

Variable 1	Variable 2	Correlation
Credit_Utilization_Ratio	Credit_Utilization_Ratio	1.000
Credit_Utilization_Ratio	Delay_from_due_date	-0.055
Credit_Utilization_Ratio	Interest_Rate	0.001
Credit_Utilization_Ratio	Monthly_Balance	0.258
Credit_Utilization_Ratio	Monthly_Inhand_Salary	0.196

Credit_Utilization_Ratio	Num_Bank_Accounts	0.001
Credit_Utilization_Ratio	Num_Credit_Card	0.002
Credit_Utilization_Ratio	Num_Credit_Inquiries	-0.002
Credit_Utilization_Ratio	Total_EMI_per_month	0.004
Delay_from_due_date	Credit_Utilization_Ratio	-0.055
Delay_from_due_date	Delay_from_due_date	1.000
Delay_from_due_date	Interest_Rate	0.009
Delay_from_due_date	Monthly_Balance	-0.281
Delay_from_due_date	Monthly_Inhand_Salary	-0.265
Delay_from_due_date	Num_Bank_Accounts	0.023
Delay_from_due_date	Num_Credit_Card	0.016
Delay_from_due_date	Num_Credit_Inquiries	0.011
Delay_from_due_date	Total_EMI_per_month	0.002
Interest_Rate	Credit_Utilization_Ratio	0.001
Interest_Rate	Delay_from_due_date	0.009
Interest_Rate	Interest_Rate	1.000
Interest_Rate	Monthly_Balance	-0.002
Interest_Rate	Monthly_Inhand_Salary	-0.005
Interest_Rate	Num_Bank_Accounts	0.002
Interest_Rate	Num_Credit_Card	-0.007
Interest_Rate	Num_Credit_Inquiries	-0.002
Interest_Rate	Total_EMI_per_month	0.001
Monthly_Balance	Credit_Utilization_Ratio	0.258
Monthly_Balance	Delay_from_due_date	-0.281
Monthly_Balance	Interest_Rate	-0.002
Monthly_Balance	Monthly_Balance	1.000
Monthly_Balance	Monthly_Inhand_Salary	0.705
Monthly_Balance	Num_Bank_Accounts	-0.012
Monthly_Balance	Num_Credit_Card	-0.005
Monthly_Balance	Num_Credit_Inquiries	-0.015
Monthly_Balance	Total_EMI_per_month	0.007
Monthly_Inhand_Salary	Credit_Utilization_Ratio	0.196
Monthly_Inhand_Salary	Delay_from_due_date	-0.265
Monthly_Inhand_Salary	Interest_Rate	-0.005
Monthly_Inhand_Salary	Monthly_Balance	0.705
Monthly_Inhand_Salary	Monthly_Inhand_Salary	1.000
Monthly_Inhand_Salary	Num_Bank_Accounts	-0.011
Monthly_Inhand_Salary	Num_Credit_Card	-0.002
Monthly_Inhand_Salary	Num_Credit_Inquiries	-0.015
Monthly_Inhand_Salary	Total_EMI_per_month	0.008
Num_Bank_Accounts	Credit_Utilization_Ratio	0.001

Num_Bank_Accounts	Delay_from_due_date	0.023
Num_Bank_Accounts	Interest_Rate	0.002
Num_Bank_Accounts	Monthly_Balance	-0.012
Num_Bank_Accounts	Monthly_Inhand_Salary	-0.011
Num_Bank_Accounts	Num_Bank_Accounts	1.000
Num_Bank_Accounts	Num_Credit_Card	0.003
Num_Bank_Accounts	Num_Credit_Inquiries	-0.006
Num_Bank_Accounts	Total_EMI_per_month	0.003
Num_Credit_Card	Credit_Utilization_Ratio	0.002
Num_Credit_Card	Delay_from_due_date	0.016
Num_Credit_Card	Interest_Rate	-0.007
Num_Credit_Card	Monthly_Balance	-0.005
Num_Credit_Card	Monthly_Inhand_Salary	-0.002
Num_Credit_Card	Num_Bank_Accounts	0.003
Num_Credit_Card	Num_Credit_Card	1.000
Num_Credit_Card	Num_Credit_Inquiries	-0.005
Num_Credit_Card	Total_EMI_per_month	0.000
Num_Credit_Inquiries	Credit_Utilization_Ratio	-0.002
Num_Credit_Inquiries	Delay_from_due_date	0.011
Num_Credit_Inquiries	Interest_Rate	-0.002
Num_Credit_Inquiries	Monthly_Balance	-0.015
Num_Credit_Inquiries	Monthly_Inhand_Salary	-0.015
Num_Credit_Inquiries	Num_Bank_Accounts	-0.006
Num_Credit_Inquiries	Num_Credit_Card	-0.005
Num_Credit_Inquiries	Num_Credit_Inquiries	1.000
Num_Credit_Inquiries	Total_EMI_per_month	-0.005
Total_EMI_per_month	Credit_Utilization_Ratio	0.004
Total_EMI_per_month	Delay_from_due_date	0.002
Total_EMI_per_month	Interest_Rate	0.001
Total_EMI_per_month	Monthly_Balance	0.007
Total_EMI_per_month	Monthly_Inhand_Salary	0.008
Total_EMI_per_month	Num_Bank_Accounts	0.003
Total_EMI_per_month	Num_Credit_Card	0.000
Total_EMI_per_month	Num_Credit_Inquiries	-0.005
Total_EMI_per_month	Total_EMI_per_month	1.000

The correlation matrix and table can be used to determine the variables to be dropped based on the correlation coefficients. The correlation matrix takes the outliers into consideration when calculating the correlation. The closer the number is to 1 or -1, the stronger the correlation. The closest positive correlation value obtained was 0.7 from Monthly In-hand Salary and Monthly Balance. However, since none of the values were in the range of |0.9| to |1|, none of the variables were dropped.

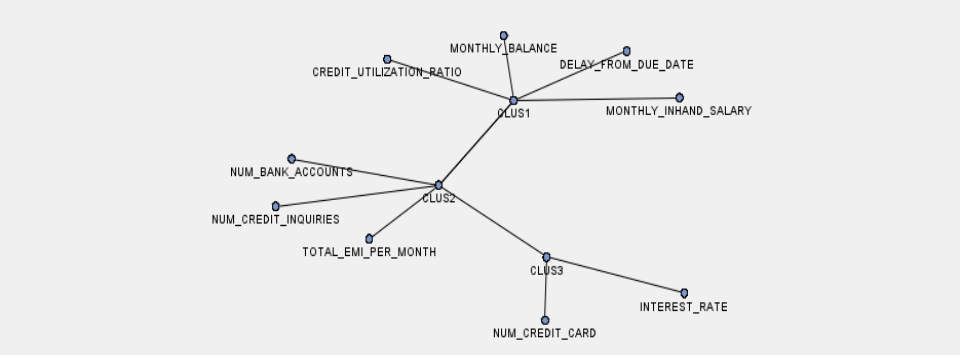


Figure 4.3.2.3.4 Cluster Plot

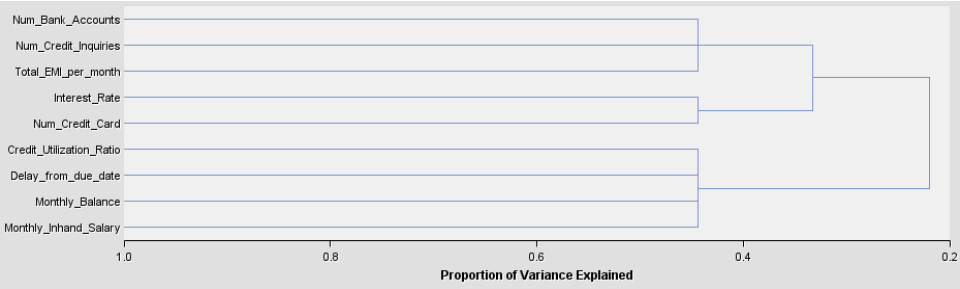


Figure 4.3.2.3.5 Dendrogram of Variance Proportion

Variable Selection Table

Cluster	Variable	R-Square With Own Cluster Component	Next Closest Cluster	R-Square with Next Cluster Component	Type	Label	1-R2 Ratio	Variable Selected
CLUS1	MONTHLY BALANCE	0.781366	CLUS2	3.433E-5	Variable		0.218641	YES
CLUS2	NUM CREDIT INQUIRIES	0.416614	CLUS1	.0002716	Variable		0.583544	YES
CLUS3	INTEREST RATE	0.503625	CLUS1	2.494E-5	Variable		0.496388	YES
CLUS1	CLUS1	1	CLUS2	2.474E-5	ClusterComp	Cluster 1	0	NO
CLUS1	MONTHLY INHAND SALA..	0.744971	CLUS2	5.799E-5	Variable		0.255044	NO
CLUS1	DELAY FROM DUE DATE	0.262602	CLUS2	5.051E-5	Variable		0.737435	NO
CLUS1	CREDIT UTILIZATION RA..	0.189093	CLUS2	1.582E-5	Variable		0.81092	NO
CLUS2	CLUS2	1	CLUS1	2.474E-5	ClusterComp	Cluster 2	0	NO
CLUS2	NUM BANK ACCOUNTS	0.338754	CLUS1	.0002425	Variable		0.661406	NO
CLUS2	TOTAL EMI PER MONTH	0.254022	CLUS1	4.678E-5	Variable		0.746013	NO
CLUS3	CLUS3	1	CLUS2	3.477E-6	ClusterComp	Cluster 3	0	NO
CLUS3	NUM CREDIT CARD	0.503625	CLUS1	4.742E-5	Variable		0.496399	NO

Figure 4.3.2.3.6 Variable Selection Table with Best Variable

Figure 4.3.2.3.4 and Figure 4.3.2.3.5 depict the variable clustering of the dataset prior to any data cleaning and transformation. There were three clusters made due to each cluster satisfies the stopping criteria specified in the Variation Proportion property. The first cluster is made of Credit_Utilization_Ratio, Monthly_Balance, Delay_from_due_date, Monthly_Inhand_Salary. The second cluster consists of Num_Bank_Accounts, Num_Credit_Inquiries, and Total_EMI_per_month. The last cluster has Num_Credit_Card and Interest_Rate. Monthly_Balance, Num_Credit_Inquiries, and Interest_Rate were the best variable in each cluster with the lowest 1–R^2 Ratio.

4.3.2.3.1 Interesting Visualization

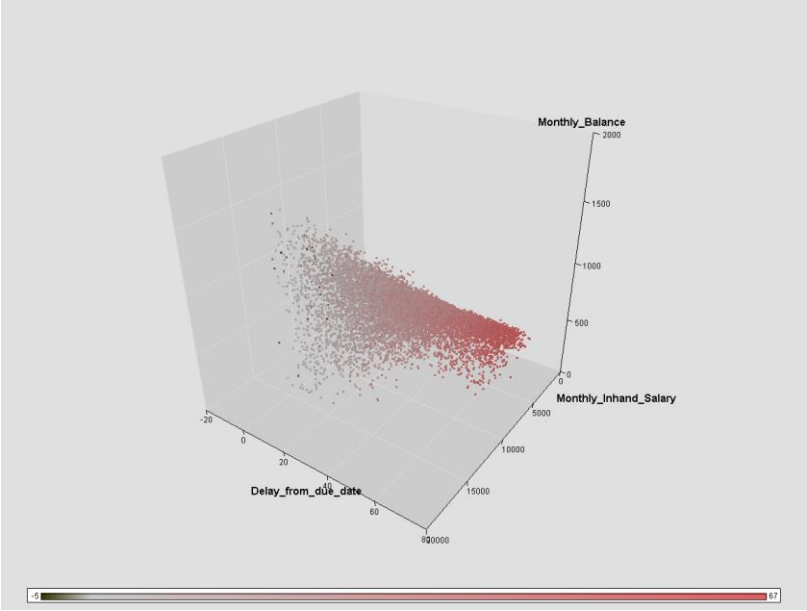


Figure 4.3.2.3.1.1 3-dimensional Scatter Plot

The Scatter plot depicts the relationship between Monthly_Inhand_Salary (x-axis), Delay_from_due_date (y-axis, colour), and Monthly_Balance (z-axis). From the plot it can be seen that Monthly_Inhand_Salary and Monthly_Balance increase proportionally. As for Delay_from_due_date, it decreases as the Monthly_Inhand_Salary and Monthly_Balance increase.

5 Conclusion

In the sample stage, the roles, and variable types of the dataset before and after making manual adjustment are shown in the table below as a comparison.

Table 5.1 Revised Metadata Level Setting

Role	Variable Type	Count (Before)	Count (After)
Input	Ordinal	-	2
Input	Interval	9	9
Input	Nominal	19	16
Target	Ordinal	-	1

In the exploration stage, several data error types are detected and listed in the table below.

Table 5.2 Column and its Data Type Error

Column	Data Type Error
Name	Incomplete
Age	Noisy
Occupation	Noisy, Incomplete
Annual_Income	Noisy
Monthly_Inhand_Salary	Incomplete
Num_Bank_Accounts	Noisy
Num_Credit_Card	Noisy
Num_of_Loan	Noisy, Inconsistent
Type_of_Loan	Incomplete
Num_of_Delayed_Payment	Incomplete, Noisy
Changed_Credit_Limit	Incomplete, Noisy
Num_Credit_Inquiries	Incomplete
Credit_Mix	Incomplete
Outstanding_Debt	Noisy
Credit_History_Age	Incomplete, Noisy
Payment_of_Min_Amount	Incomplete
Amount_invested_monthly	Incomplete, Noisy
Payment_Behaviour	Noisy
Monthly_Balance	Incomplete
Interest_Rate	Noisy
Total_EMI_per_Month	Noisy

The common data type errors found in this dataset are incomplete due to missing values, inconsistent due to negative values that are supposed to be positive value, and lastly noisy data due to outliers. These invalid data types will be corrected in the Modify stage by applying data transformation techniques like data cleaning, binning etc., which will be described in detail in the next project phase.

The primary objective of this project is to build a credit scoring classification model based on borrower's credit related information. To accomplish this goal, the dataset was explored through various statistical techniques, including univariate, bivariate and multivariate analysis for deeper understanding of the dataset. The key findings of the analyses allow us to identify strongly correlated features and can be used as guidance to develop the credit score classification model. The insights are as follows:

1. This project is conducted with 100,000 rows that consist of 12,500 unique customers. Most customers with different types of occupation have roughly a credit card utilization of 32% from their credit limit, and mostly delay payment from due date by 2 weeks. Most customers behave

in a way, i.e., to make low spending with small value payment, followed by high spending with medium payment and high spending with large value payment.

2. The monthly in-hand salary variable shows significant positive correlation with monthly balance variable, but both variables show negative correlation with the number of credit cards variable. This indicates customers with lower monthly in-hand salary and monthly balances hold more credit cards.
3. Customers who owned more than 5 credit cards are more likely to delay their payment more, which is probably due to higher outstanding balance to pay off. The more the days delayed from payment due date, the higher the interest rates charged for credit card outstanding, which will negatively impact customers' credit score.
4. Borrowers with more money in hand have less delay in payment due to their strong financial affordability. When borrowers have more money in hand, their loan interest rate is normally lesser. This is observed that the interest rate dropped by half when monthly in hand salary is greater than $7,000.
5. Customers who tend to make only minimum payment instead of full payment and make low credit card payment usually have a lower credit mix score, as compared to those customers who paid full amount.
6. Based on Lattice pie charts, it can be observed that larger payment value affects credit score, as compared to spending frequency. From Cramer's V statistic, individual's amount invested monthly indicates strongest correlation with credit score, followed by annual income and type of loan, respectively. From the correlation matrix table, the monthly in-hand salary and monthly balance shows a positive correlation to credit score with a correlation value of more than 0.7.
7. In overall, credit mix, monthly balance, delay from payment due date (in days) and payment of only minimum amount show correlation with the credit score. There may be more variables showing correlation with the credit score after applying data transformation in the Modify phase.

6 References

Einav, L., Jenkins, M. A., & Levin, J. (2013). The impact of credit scoring on consumer lending. *The RAND Journal of Economics*, *44*(2), 249–274. https://doi.org/10.1111/1756-2171.12019

Hand, D. J., & Henley, W. (1997). Statistical Classification Methods in Consumer Credit Scoring: a Review. *Journal of the Royal Statistical Society*, *160*(3), 523–541. https://doi.org/10.1111/j.1467-985x.1997.00077.x

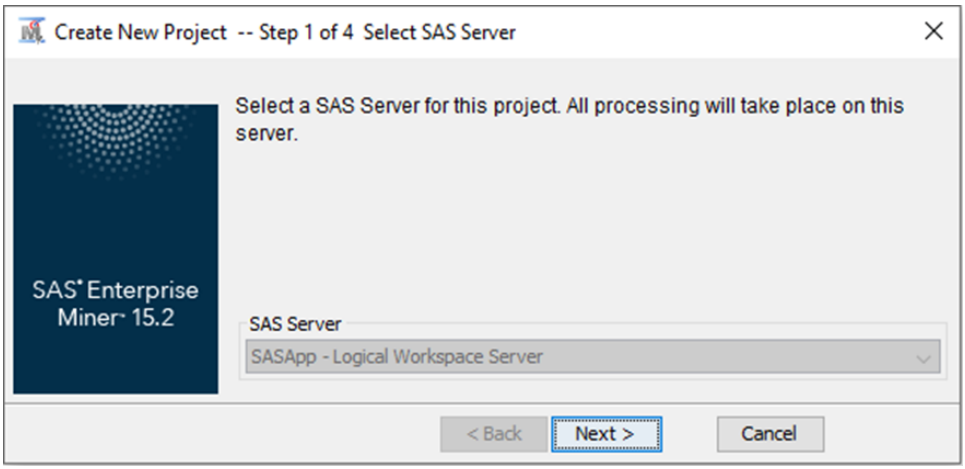
Jha, V. (2020, September 5). *Semma model*. GeeksforGeeks. https://www.geeksforgeeks.org/semma-model/

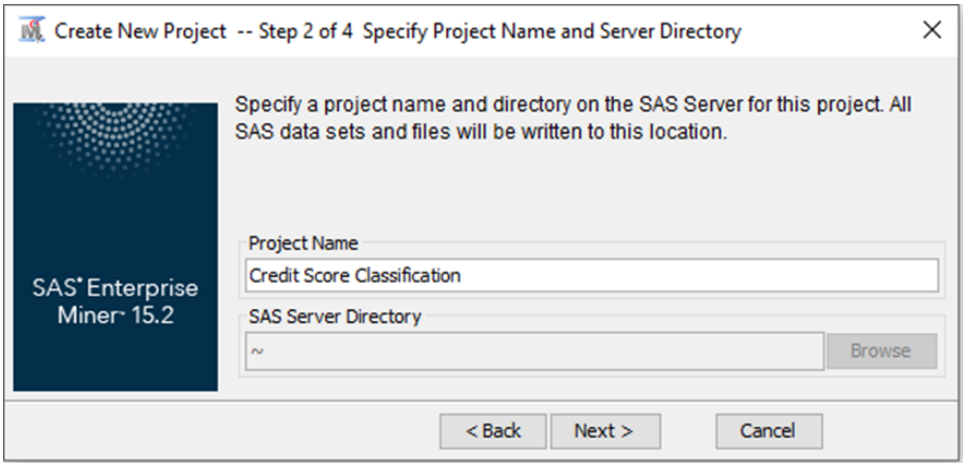
Li, X., Ying, W., Tuo, J., Li, B., & Liu, W. (2004). Applications of classification trees to consumer credit scoring methods in commercial banks. *Systems, Man and Cybernetics*. https://doi.org/10.1109/icsmc.2004.1401175

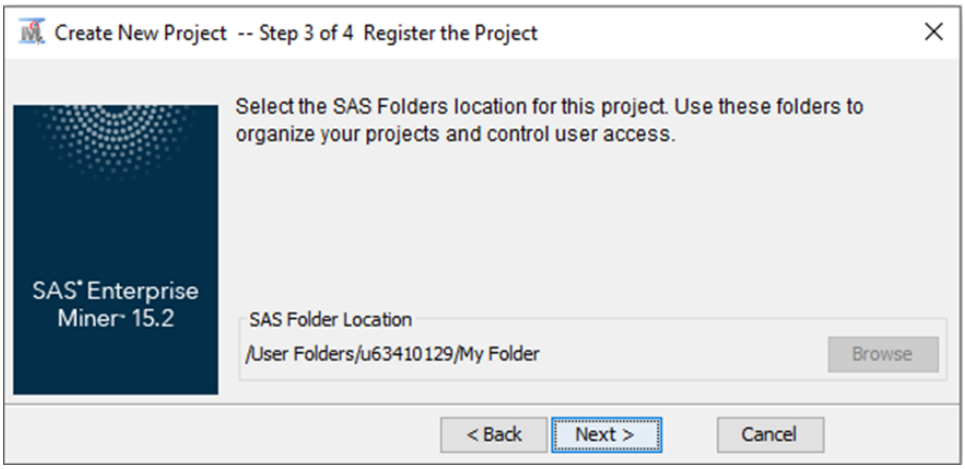
7 Appendix (SAS Enterprise Miner Screenshots)

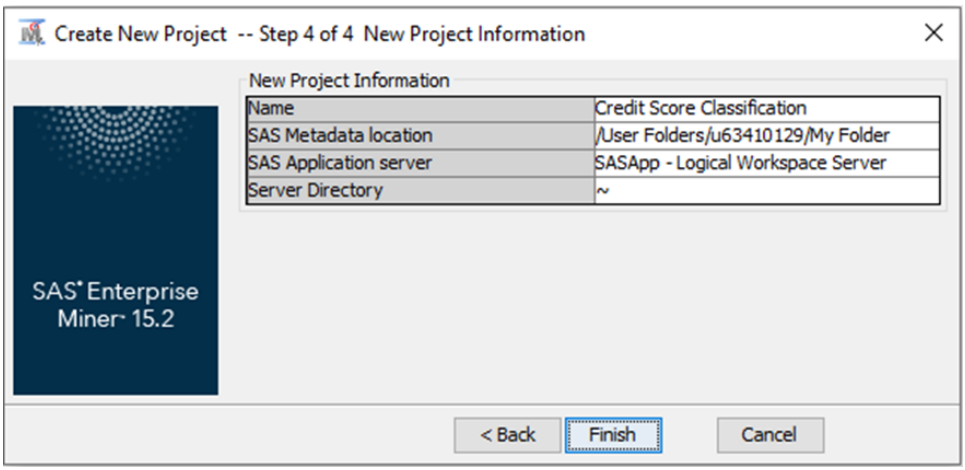
7.1 Project Setup

Figures 7.1.1 - 7.1.4 Create Project

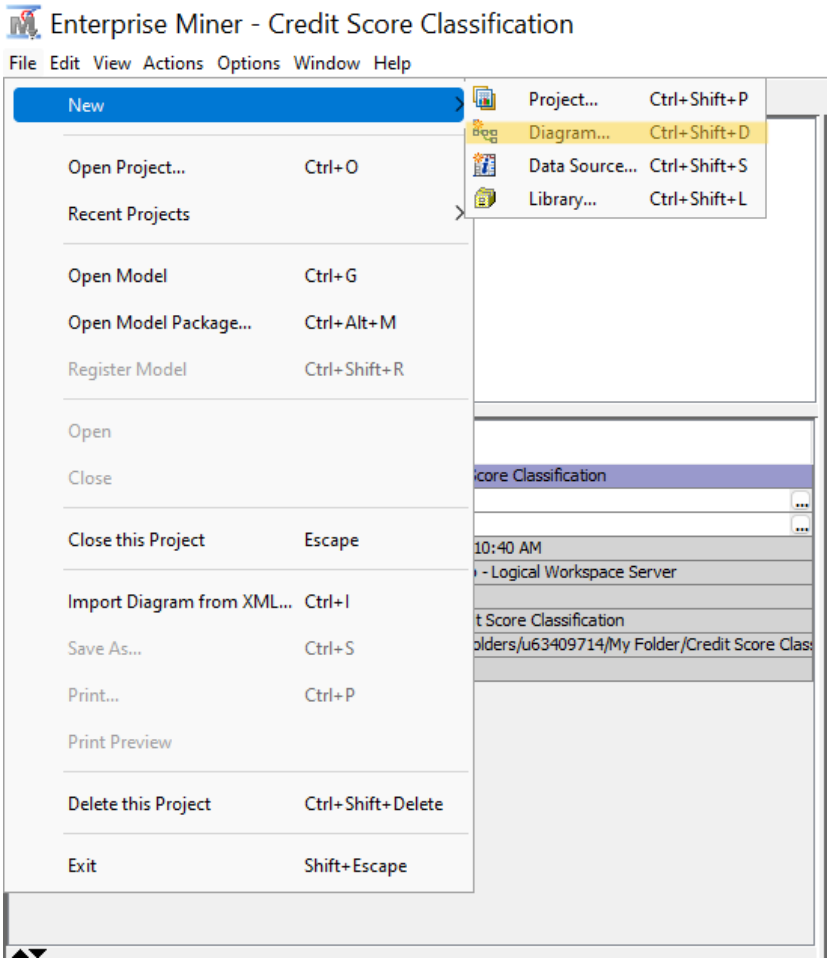


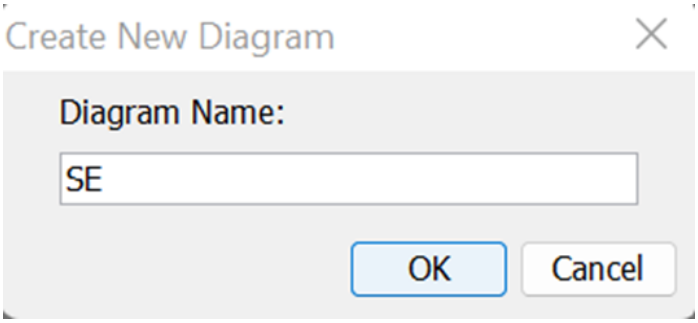




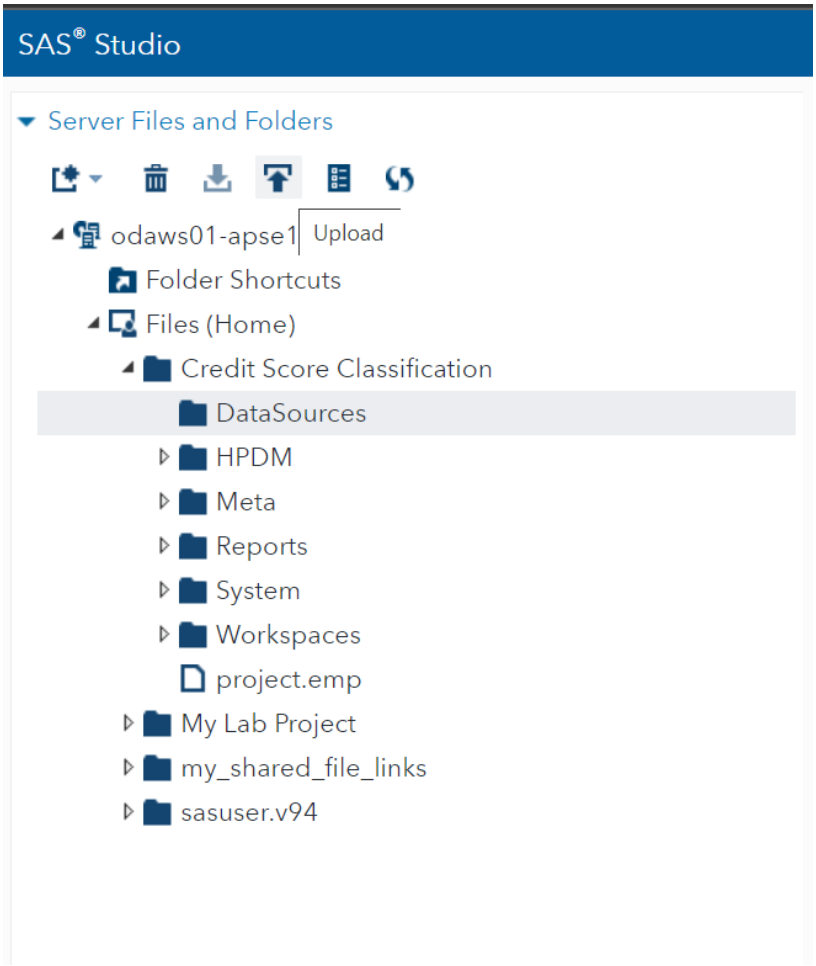


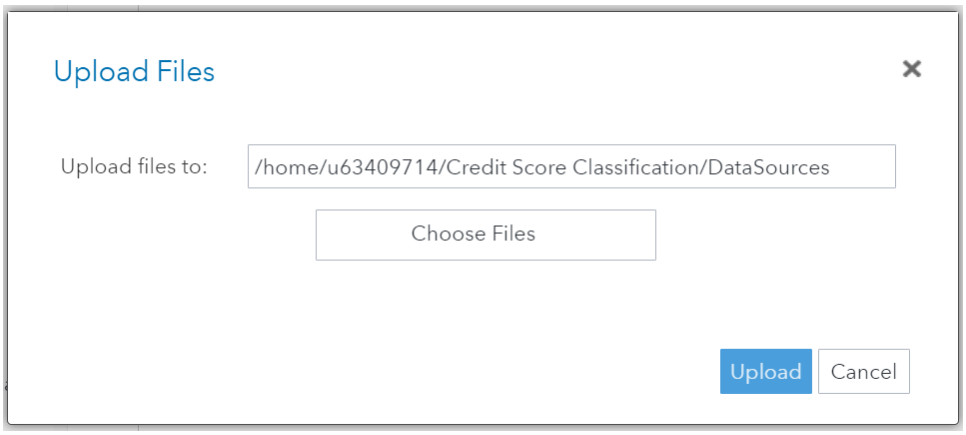
Figures 7.1.5 & 7.1.6 Create Diagram

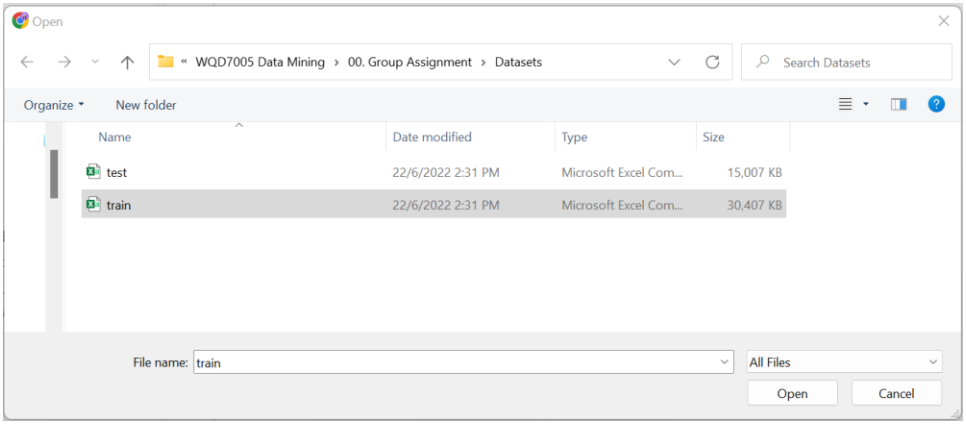


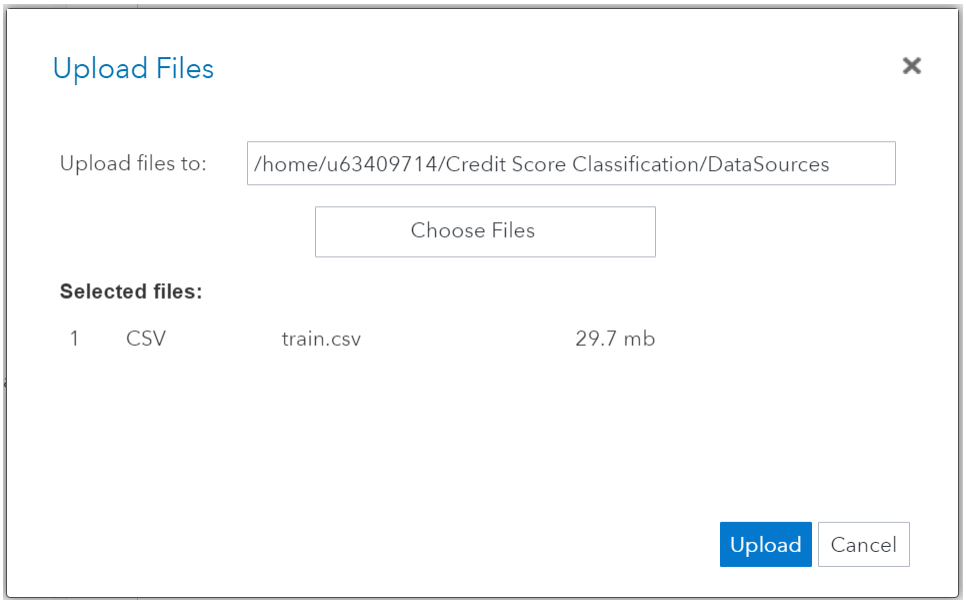


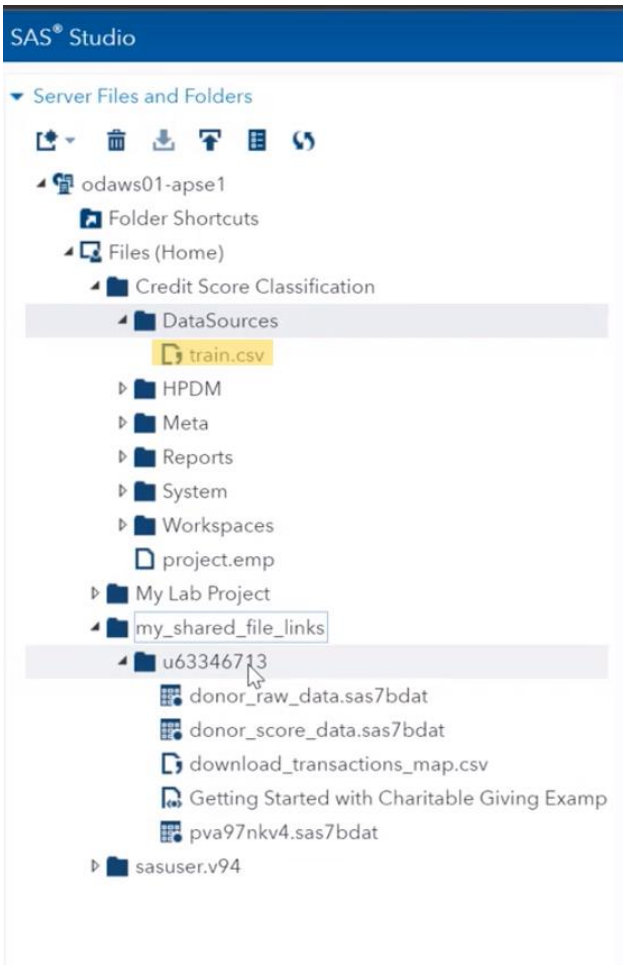
Figures 7.1.7 - 7.1.11 Upload Data to SAS Studio



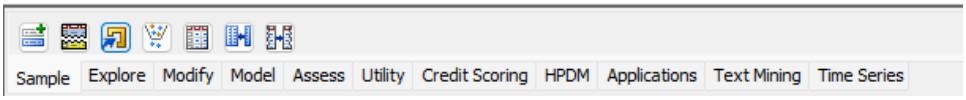

Upload Files
Upload files to:
/home/u63409714/Credit Score Classification/DataSources
Choose Files
Upload
Cancel


Open
WQD7005 Data Mining › 00. Group Assignment › Datasets
Search Datasets
Organize
New folder
Name
Date modified
Type
Size
test
22/6/2022 2:31 PM
Microsoft Excel Com...
15,007 KB
train
22/6/2022 2:31 PM
Microsoft Excel Com...
30,407 KB
File name: train
All Files
Open
Cancel

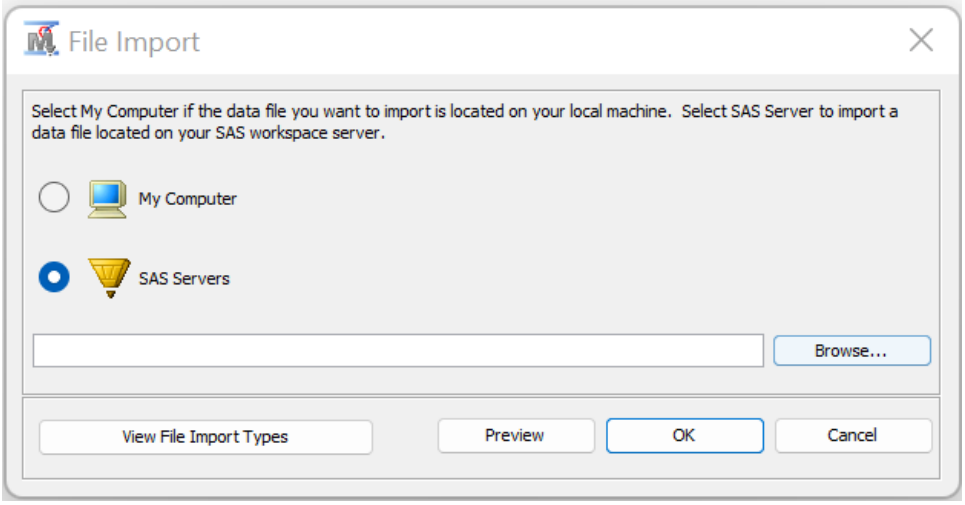

Upload Files
Upload files to:
/home/u63409714/Credit Score Classification/DataSources
Choose Files
Selected files:
1 CSV train.csv 29.7 mb
Upload
Cancel

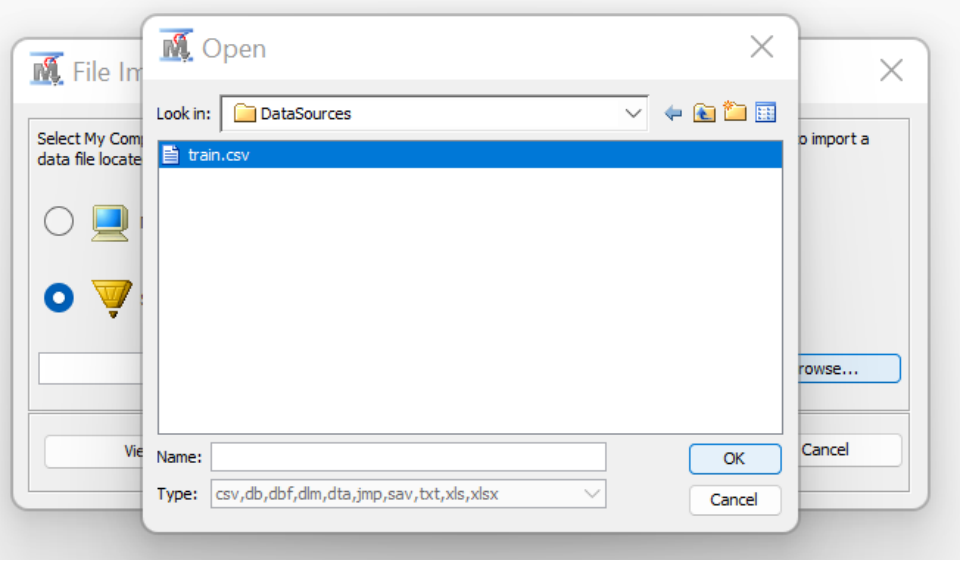


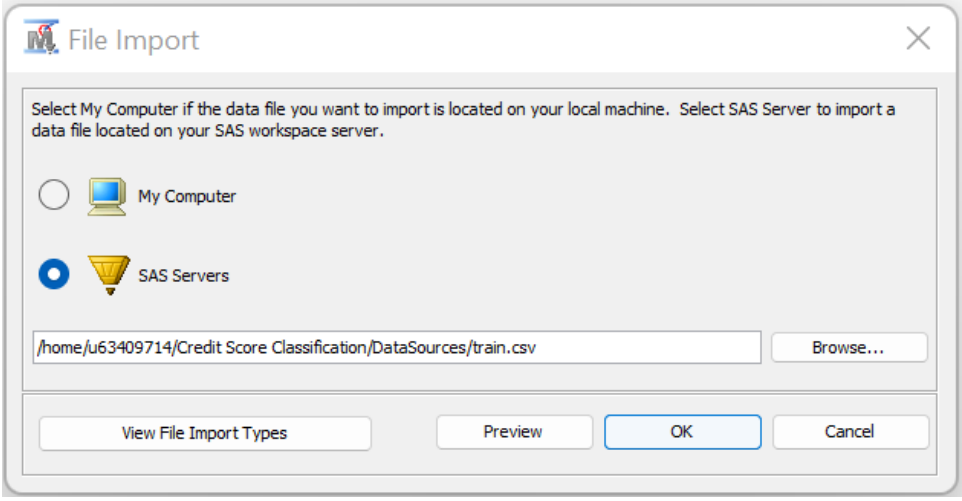
Figures 7.1.12 - 7.1.18 Import File to Enterprise Miner



Property	Value
General	
Node ID	FIMPORT
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Import File	
Maximum Rows to Import	1000000
Maximum Columns to Import	10000
Delimiter	,
Name Row	Yes
Number of Rows to Skip	0
Guessing Rows	500
File Location	Local
File Type	XLS
Advanced Advisor	No
Rerun	No
Score	
Role	Train

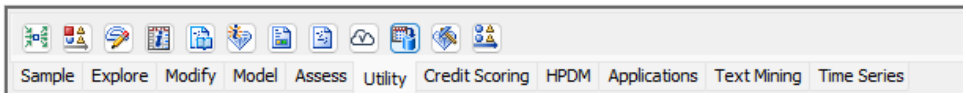


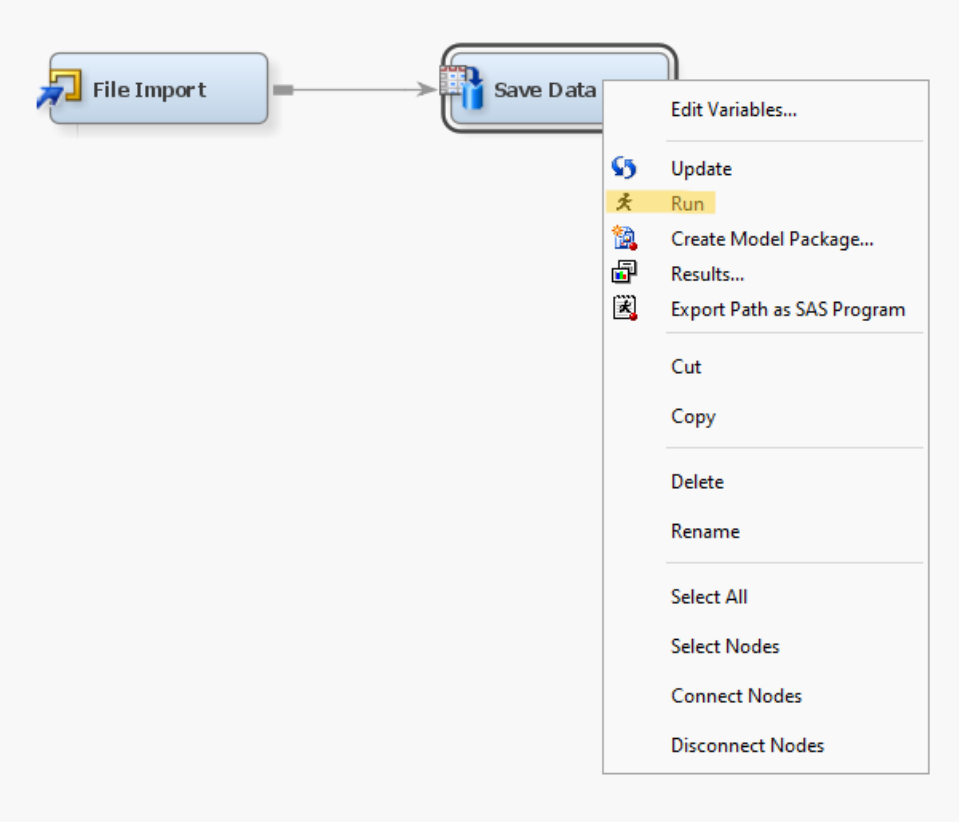

File Im
Open
Look in:
DataSources
train.csv
Select My Com
data file locate
o import a
rowse...
Vie
Cancel
Name:
Type:
csv,db,dbf,dlm,dta,jmp,sav,txt,xls,xlsx
OK
Cancel

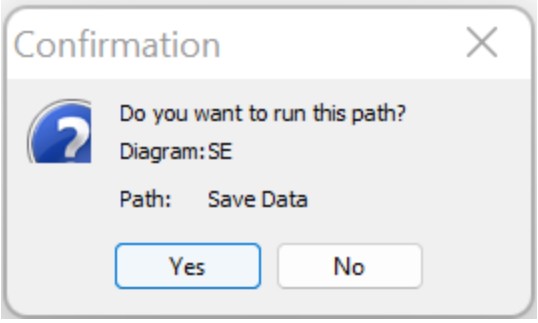

File Import
Select My Computer if the data file you want to import is located on your local machine. Select SAS Server to import a data file located on your SAS workspace server.
My Computer
SAS Servers
/home/u63409714/Credit Score Classification/DataSources/train.csv
Browse...
View File Import Types
Preview
OK
Cancel

Property	Value
General	
Node ID	FIMPORT
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Import File	/home/u63409714/Credit Score Classification/Dat
Maximum Rows to Import	1000000
Maximum Columns to Import	10000
Delimiter	,
Name Row	Yes
Number of Rows to Skip	0
Guessing Rows	500
File Location	Remote
File Type	csv
Advanced Advisor	No
Rerun	No
Score	
Role	Train

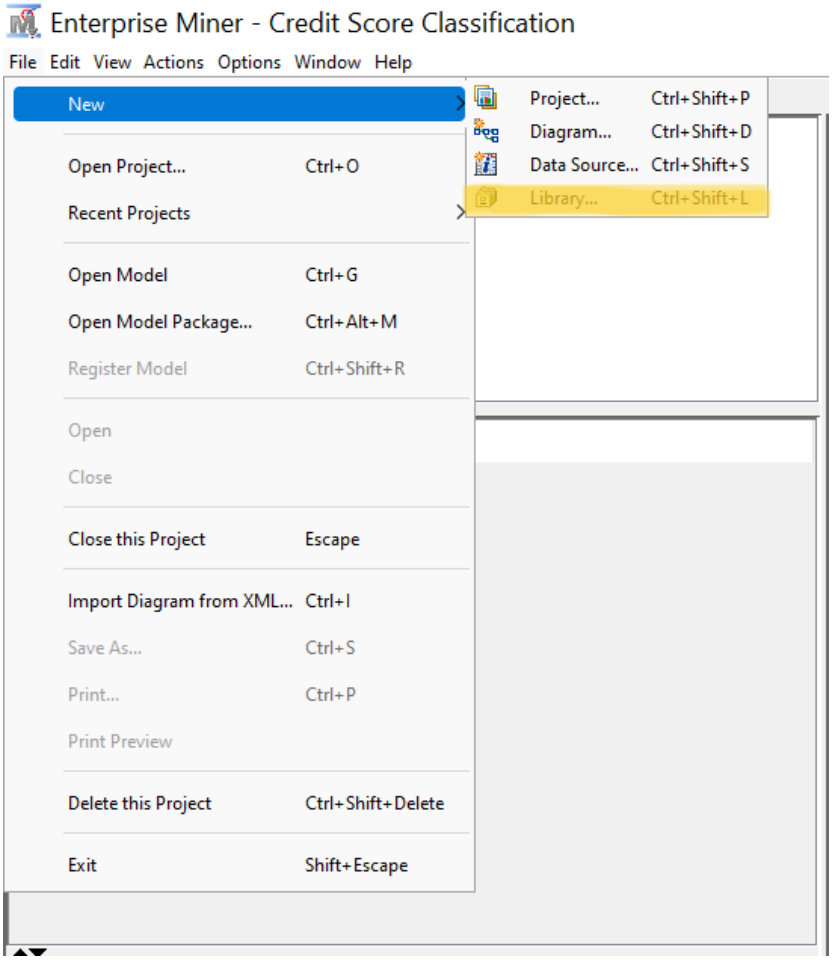
Figures 7.1.19 - 7.1.23 Save as SAS Table

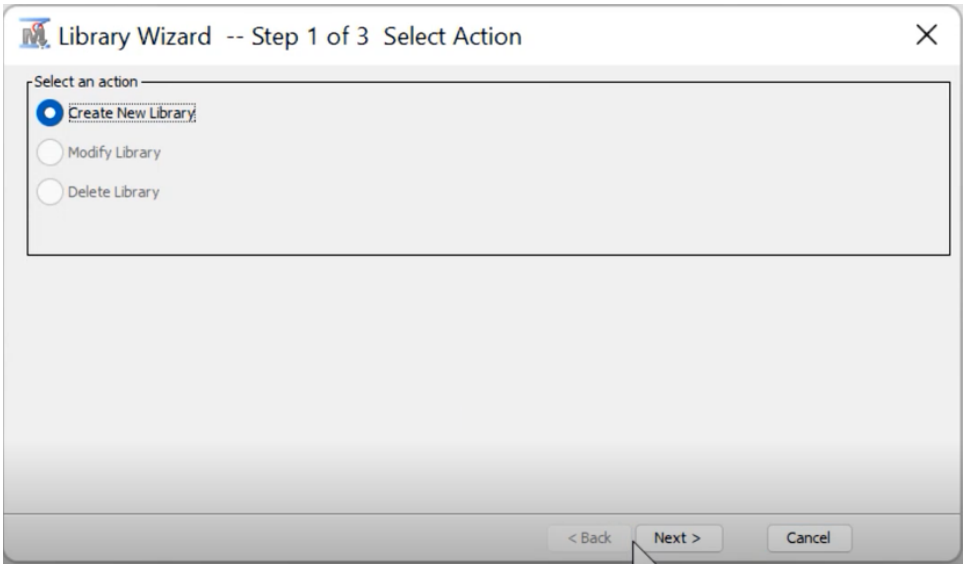


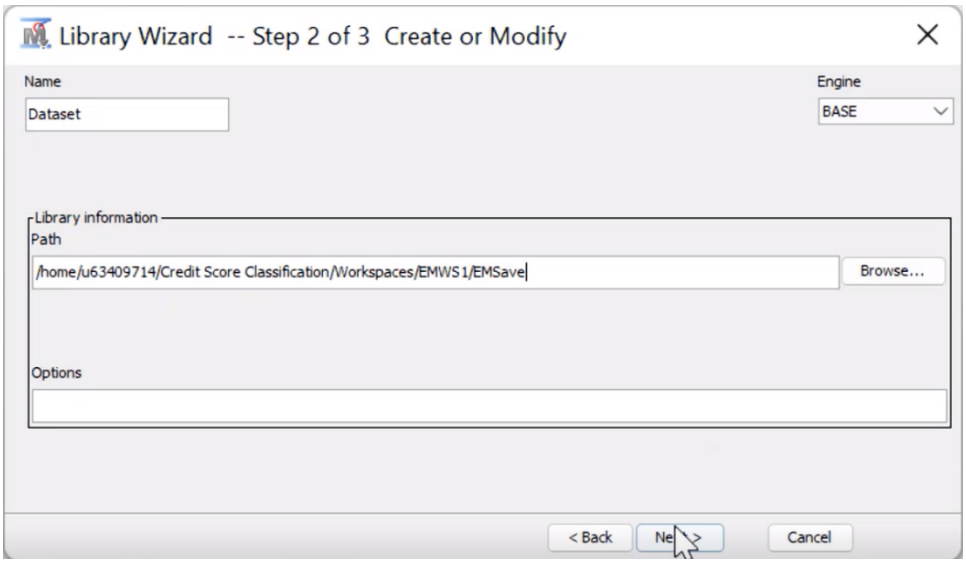


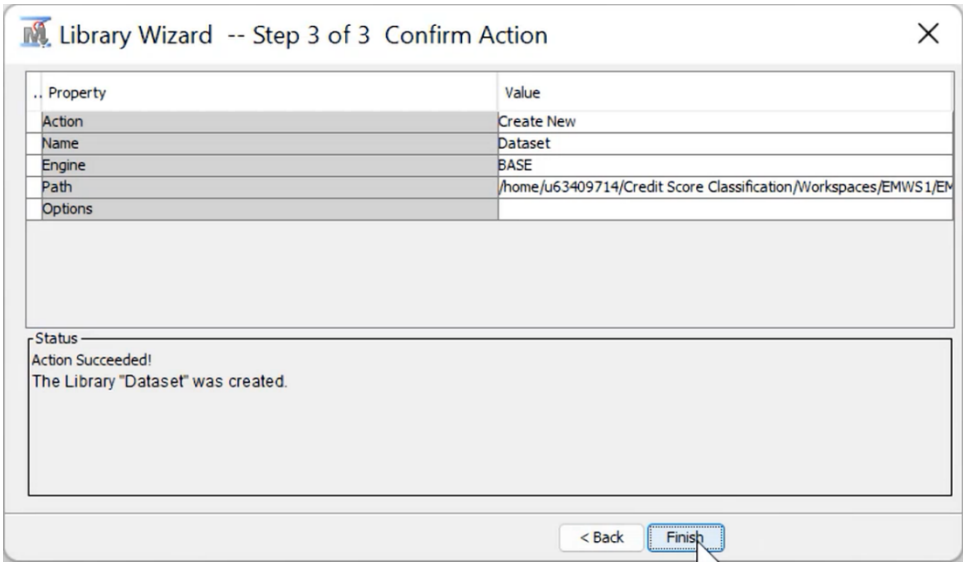


Figures 7.1.24 - 7.1.27 Create Library


Enterprise Miner - Credit Score Classification
File Edit View Actions Options Window Help
New
Open Project... Ctrl+O
Recent Projects
Open Model Ctrl+G
Open Model Package... Ctrl+Alt+M
Register Model Ctrl+Shift+R
Open
Close
Close this Project Escape
Import Diagram from XML... Ctrl+I
Save As... Ctrl+S
Print... Ctrl+P
Print Preview
Delete this Project Ctrl+Shift+Delete
Exit Shift+Escape
Project... Ctrl+Shift+P
Diagram... Ctrl+Shift+D
Data Source... Ctrl+Shift+S
Library... Ctrl+Shift+L

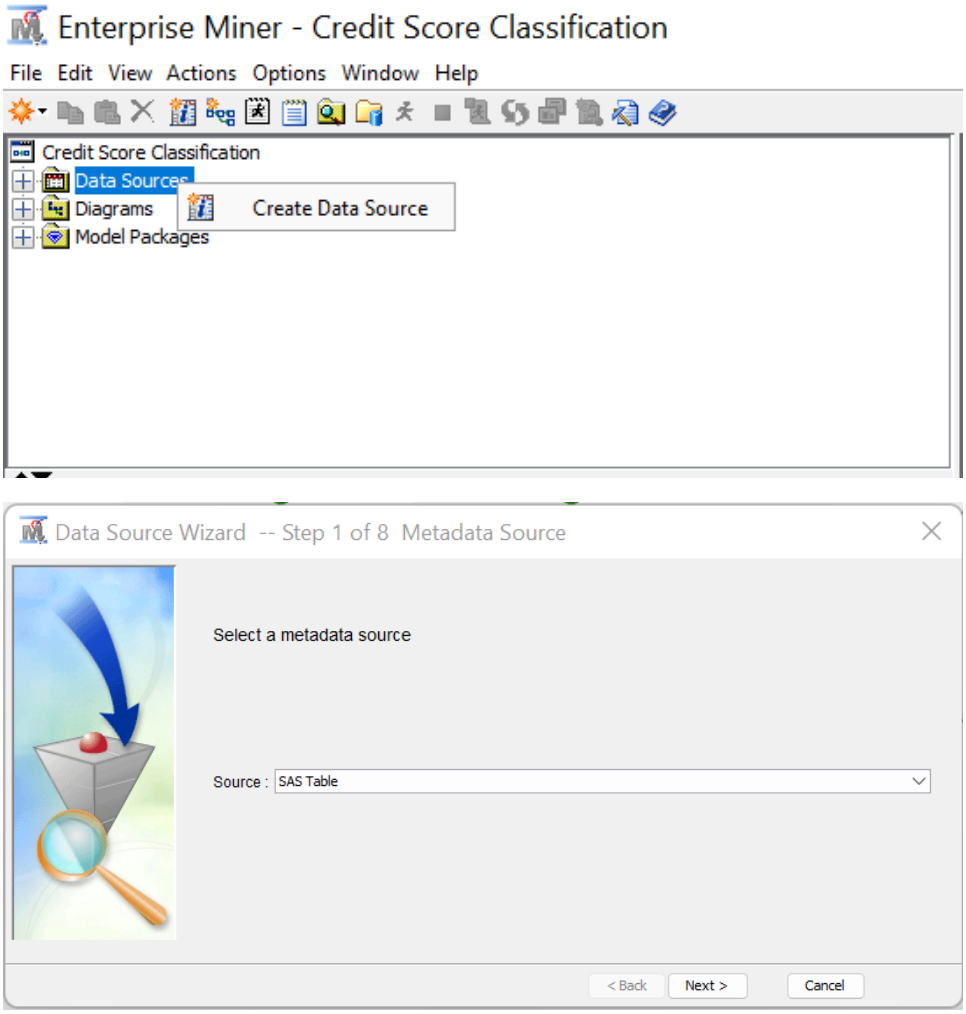

Library Wizard -- Step 1 of 3 Select Action
Select an action
Create New Library
Modify Library
Delete Library
< Back
Next >
Cancel



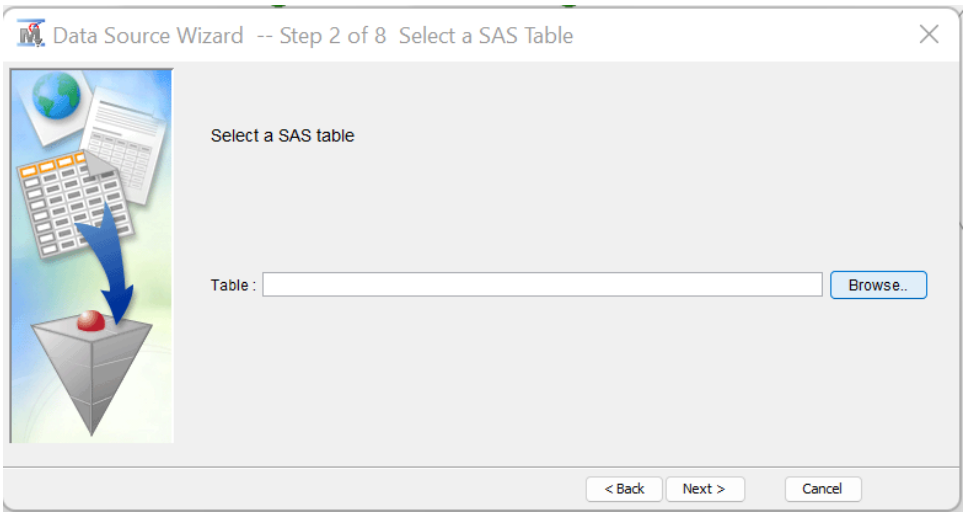


7.2 Sample

Figures 7.2.1 - 7.2.2 Create Data Source



Figures 7.2.3 - 7.2.6 Select SAS Table



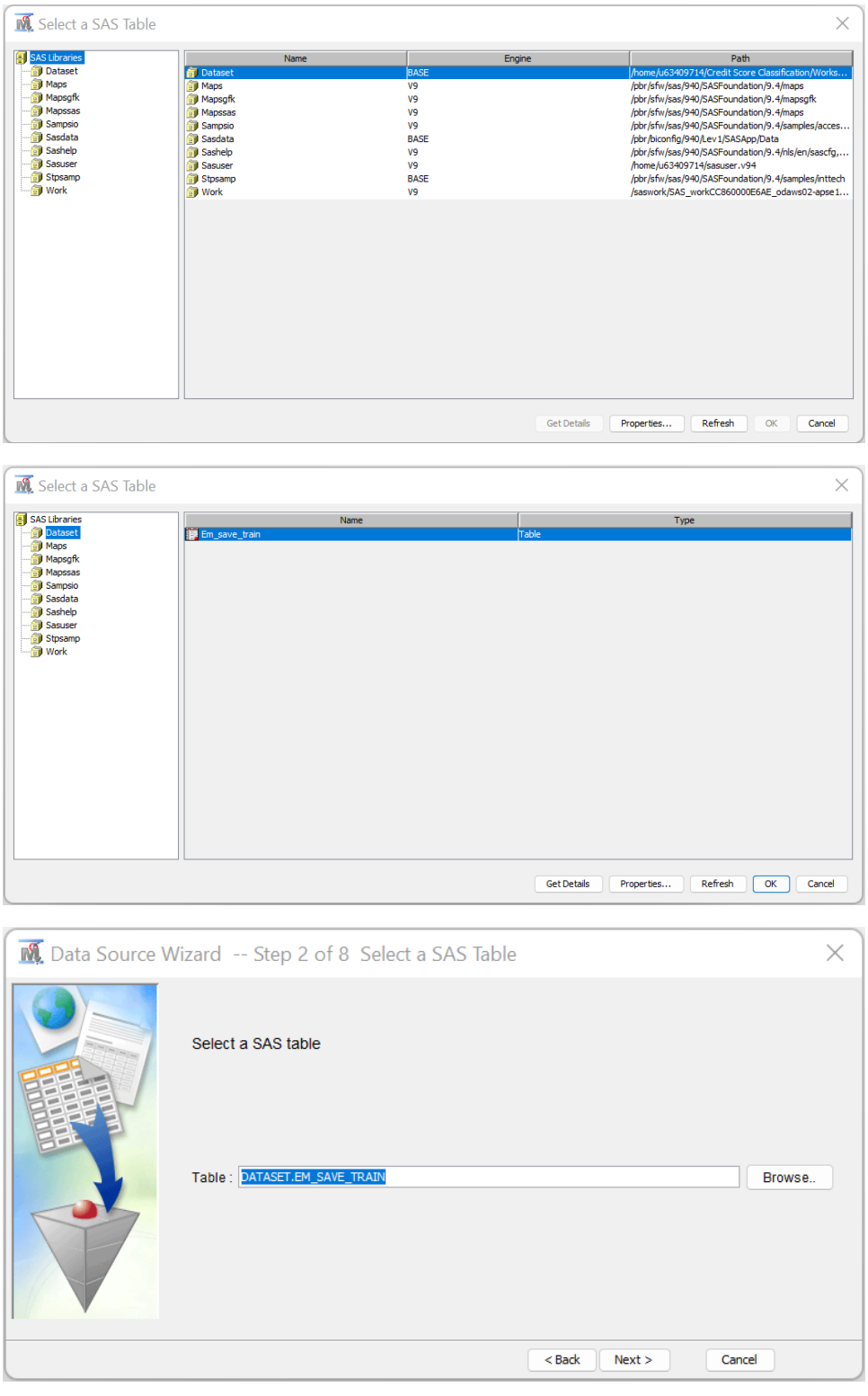


Figure 7.2.7 Information of Selected SAS Table

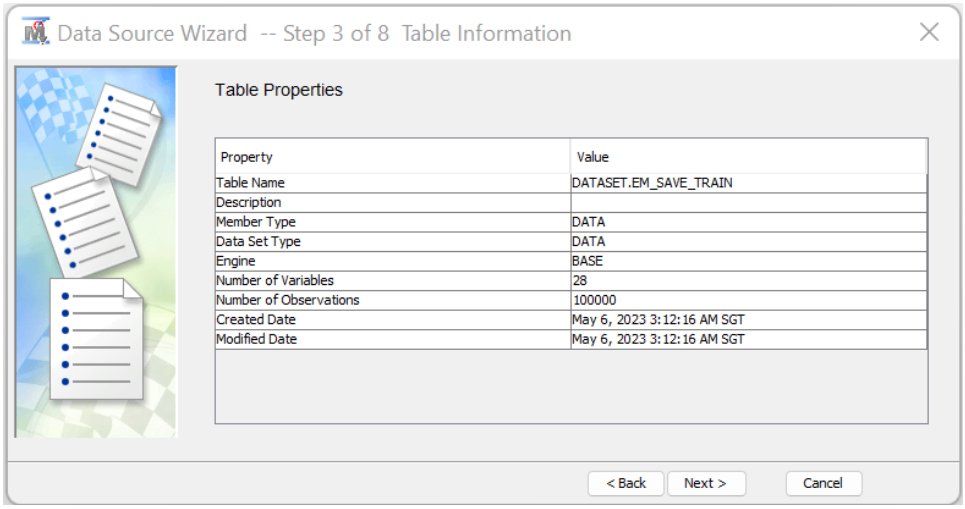


Figure 7.2.8 Metadata Advanced Setting

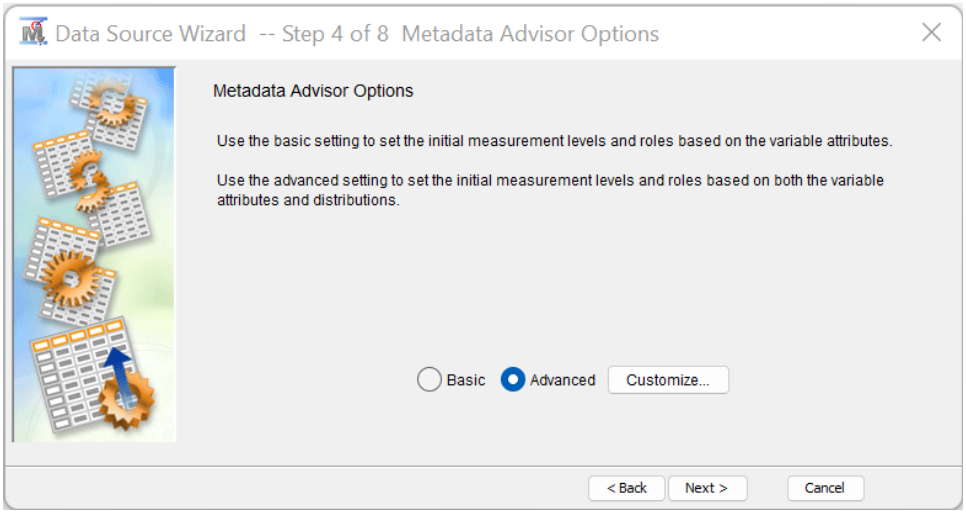


Figure 7.2.9 Default Metadata Setting

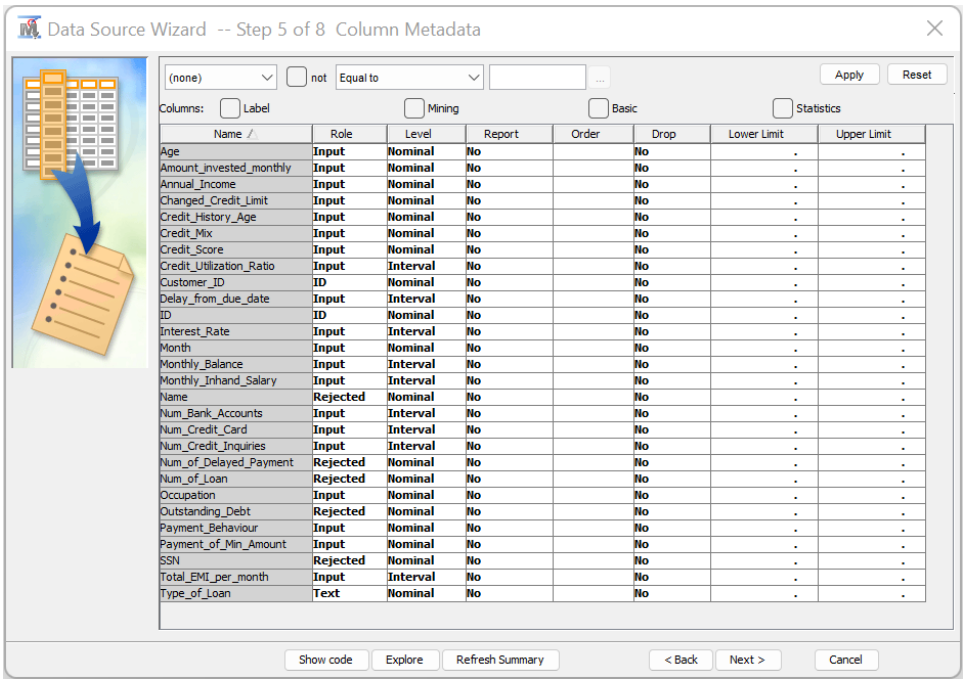


Figure 7.2.10 Modified Metadata Setting

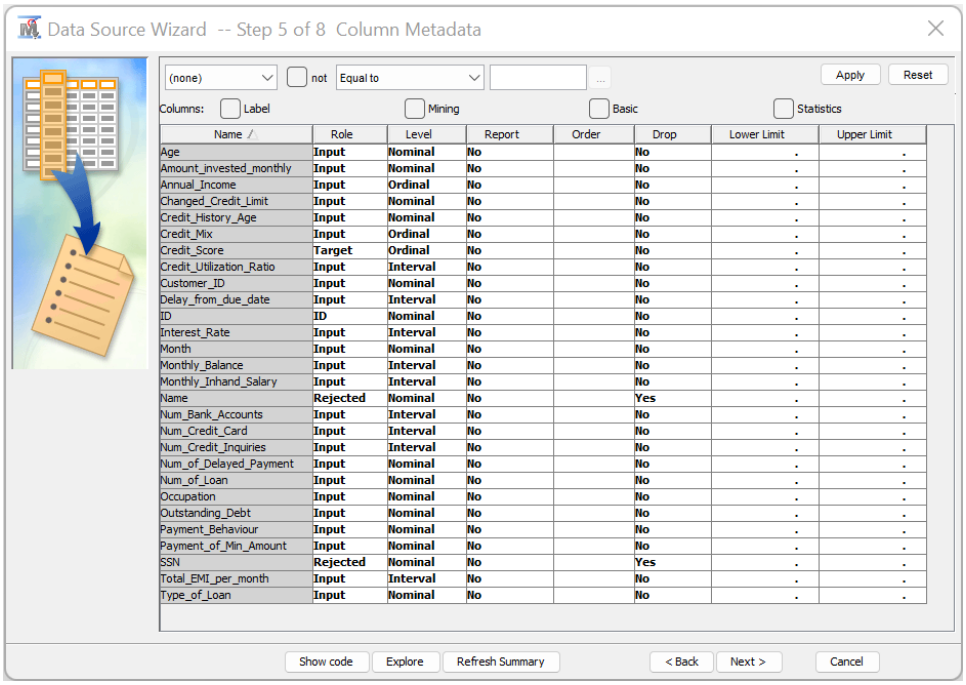


Figure 7.2.11 Unspecified Decision Processing

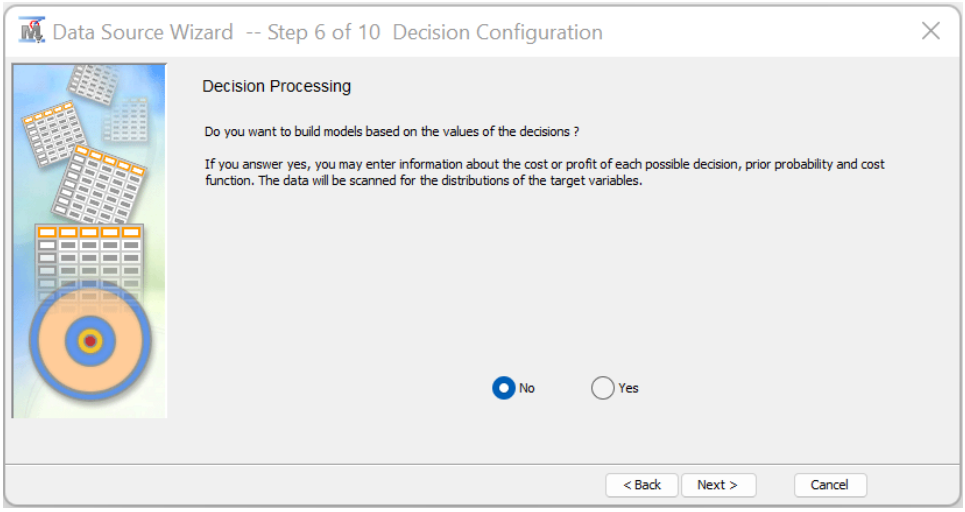


Figure 7.2.12 Did Not Create Sample Dataset

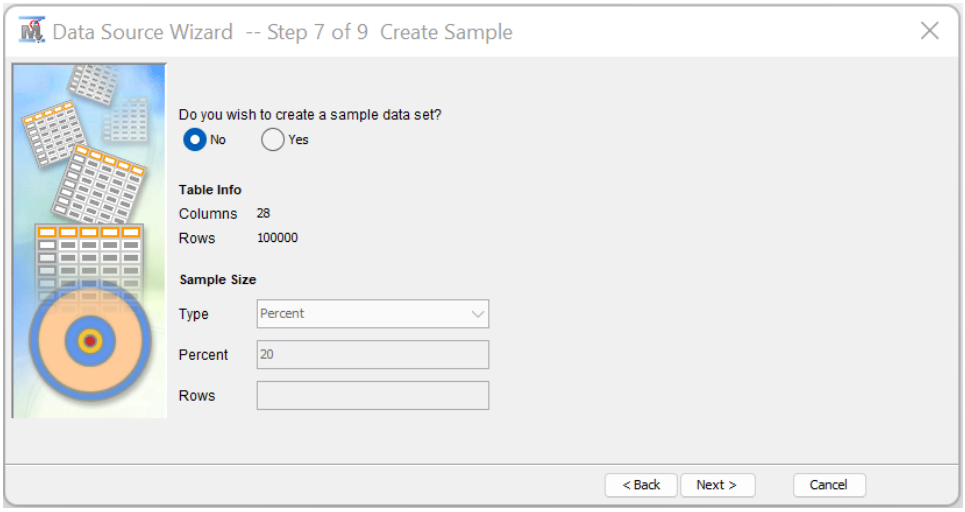


Figure 7.2.13 Name and Role of Data Source

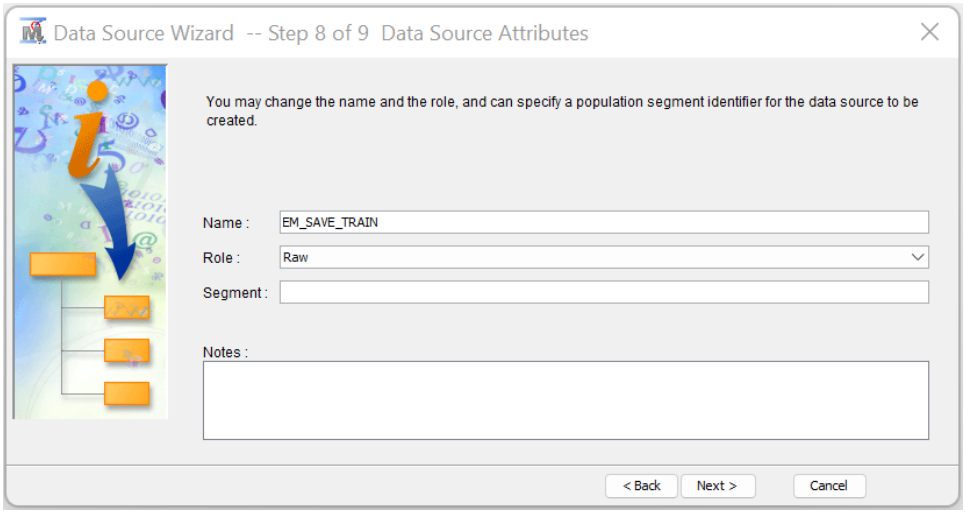
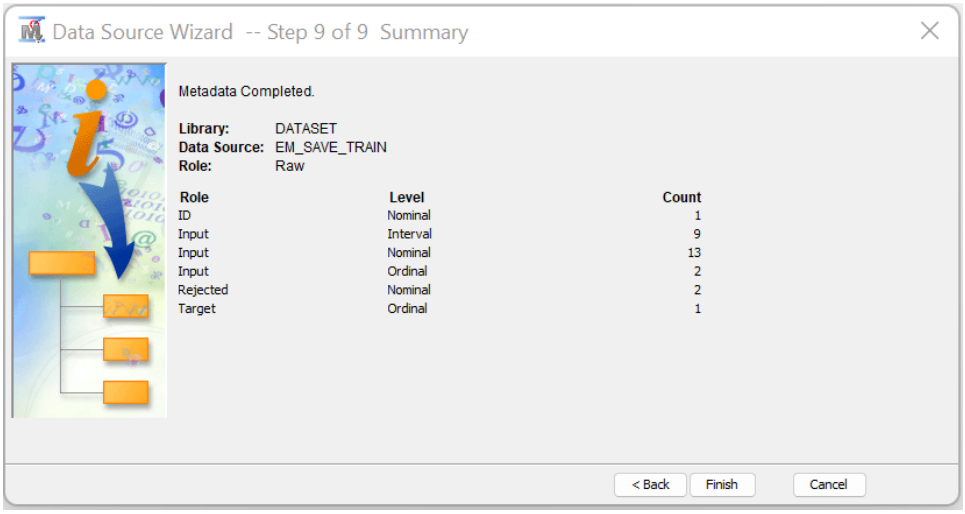
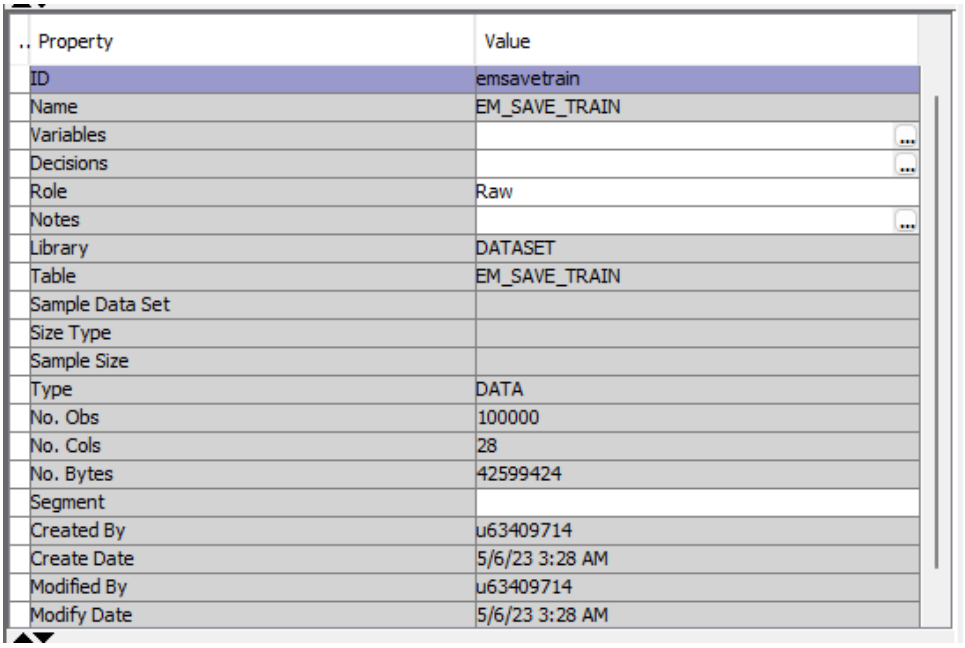


Figure 7.2.14 Data Source Summary

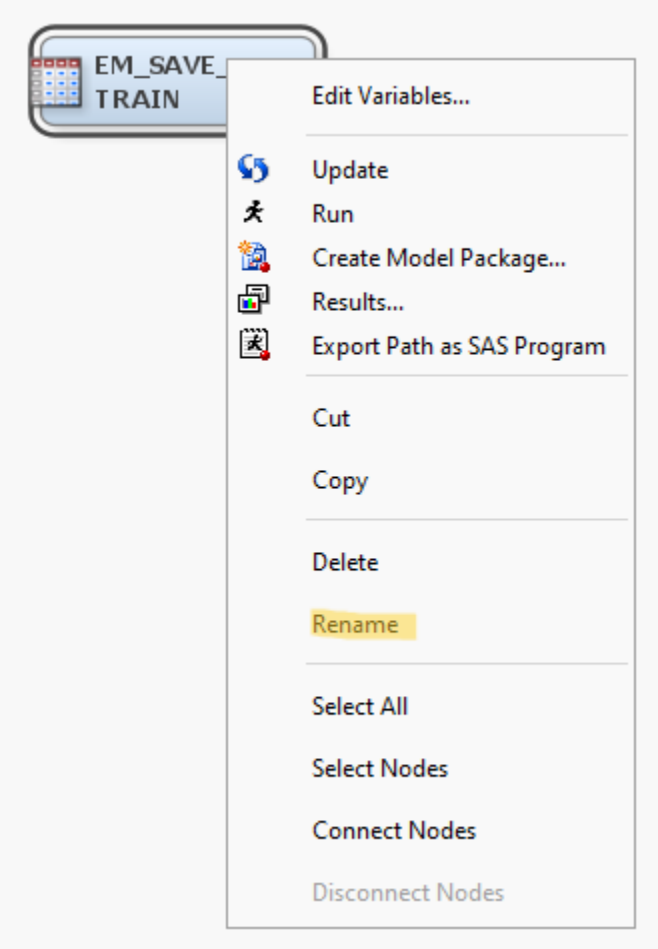


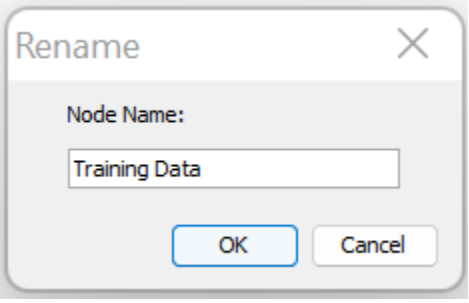
Figures 7.2.15 - 7.2.17 Use Created Data Source

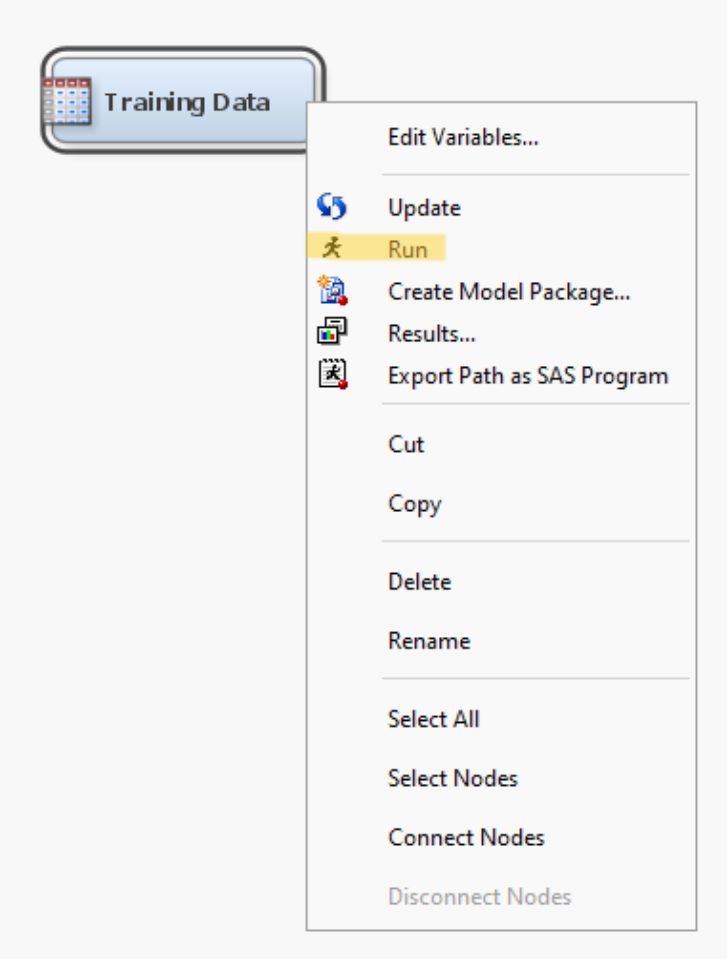


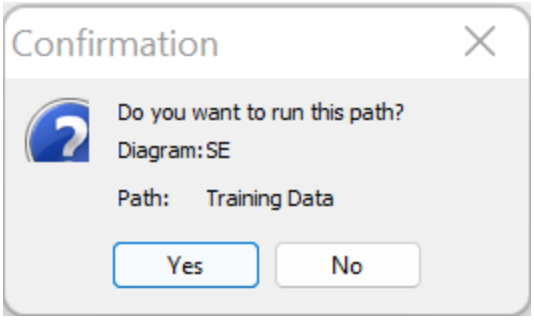
Property	Value
ID	emsavetrain
Name	EM_SAVE_TRAIN
Variables	
Decisions	
Role	Raw
Notes	
Library	DATASET
Table	EM_SAVE_TRAIN
Sample Data Set	
Size Type	
Sample Size	
Type	DATA
No. Obs	100000
No. Cols	28
No. Bytes	42599424
Segment	
Created By	u63409714
Create Date	5/6/23 3:28 AM
Modified By	u63409714
Modify Date	5/6/23 3:28 AM

Figures 7.2.18 - 7.2.22 Rename Data Source

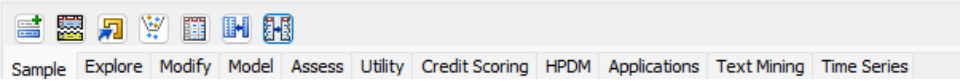








Figures 7.2.23 - 7.2.24 Sample Node in Sample Tab



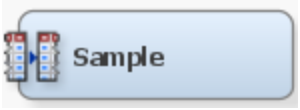
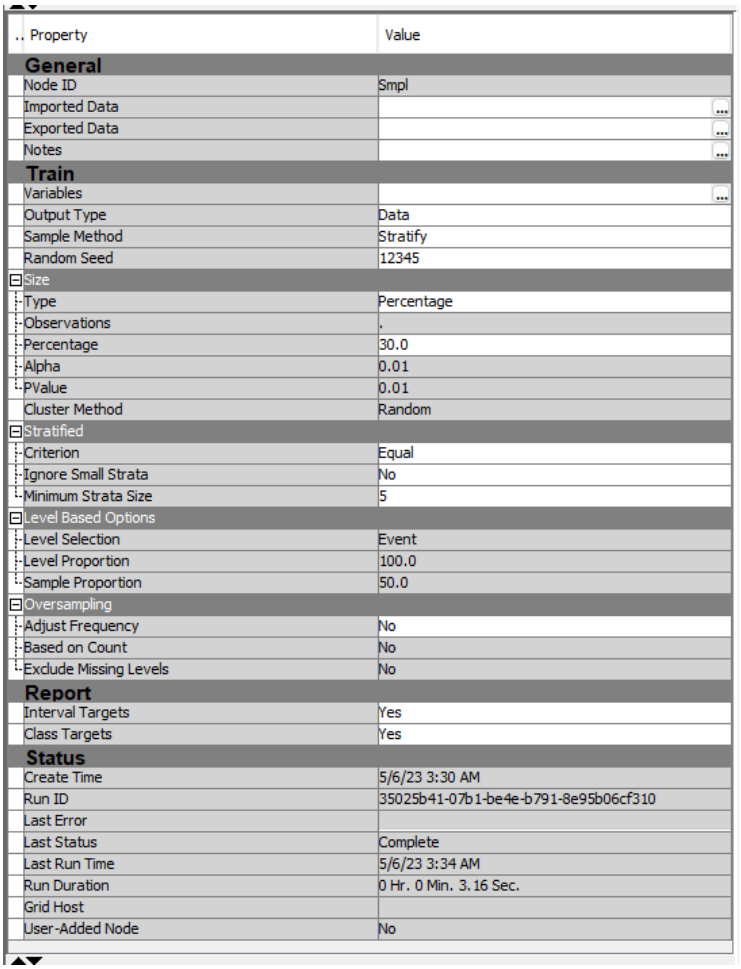


Figure 7.2.25 Default Setting of Sample Node

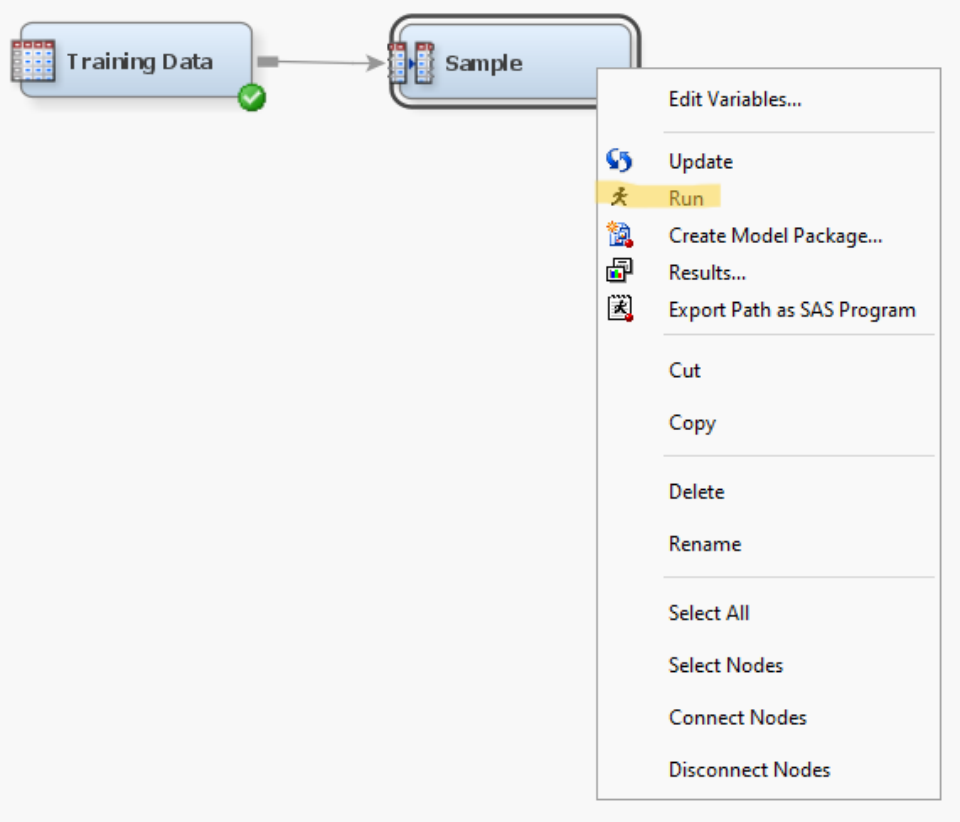
Property	Value
General	
Node ID	Smpl
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Output Type	Data
Sample Method	Default
Random Seed	12345
Size	
Type	Percentage
Observations	.
Percentage	10.0
Alpha	0.01
PValue	0.01
Cluster Method	Random
Stratified	
Criterion	Proportional
Ignore Small Strata	No
Minimum Strata Size	5
Level Based Options	
Level Selection	Event
Level Proportion	100.0
Sample Proportion	50.0
Oversampling	
Adjust Frequency	No
Based on Count	No
Exclude Missing Levels	No
Report	
Interval Targets	Yes
Class Targets	Yes
Status	
Create Time	5/8/23 10:09 AM
Run ID	
Last Error	
Last Status	
Last Run Time	
Run Duration	
Grid Host	
User-Added Node	No

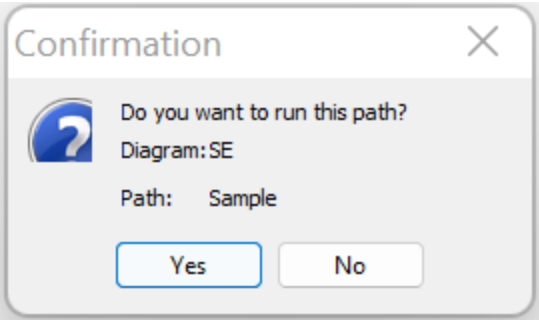
Figure 7.2.26 Modified Setting of Sample Node



Property	Value
General	
Node ID	Smpl
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Output Type	Data
Sample Method	Stratify
Random Seed	12345
Size	
Type	Percentage
Observations	.
Percentage	30.0
Alpha	0.01
PValue	0.01
Cluster Method	Random
Stratified	
Criterion	Equal
Ignore Small Strata	No
Minimum Strata Size	5
Level Based Options	
Level Selection	Event
Level Proportion	100.0
Sample Proportion	50.0
Oversampling	
Adjust Frequency	No
Based on Count	No
Exclude Missing Levels	No
Report	
Interval Targets	Yes
Class Targets	Yes
Status	
Create Time	5/6/23 3:30 AM
Run ID	35025b41-07b1-be4e-b791-8e95b06cf310
Last Error	
Last Status	Complete
Last Run Time	5/6/23 3:34 AM
Run Duration	0 Hr. 0 Min. 3.16 Sec.
Grid Host	
User-Added Node	No

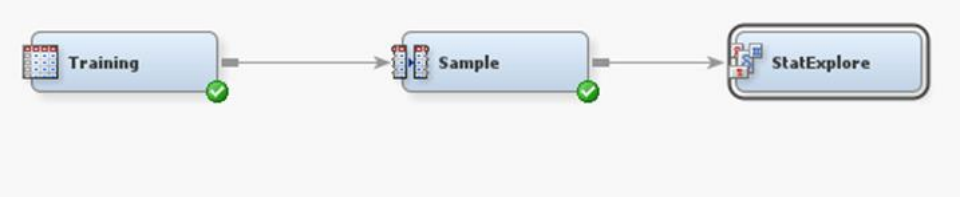
Figures 7.2.27 - 7.2.29 Perform Data Sampling

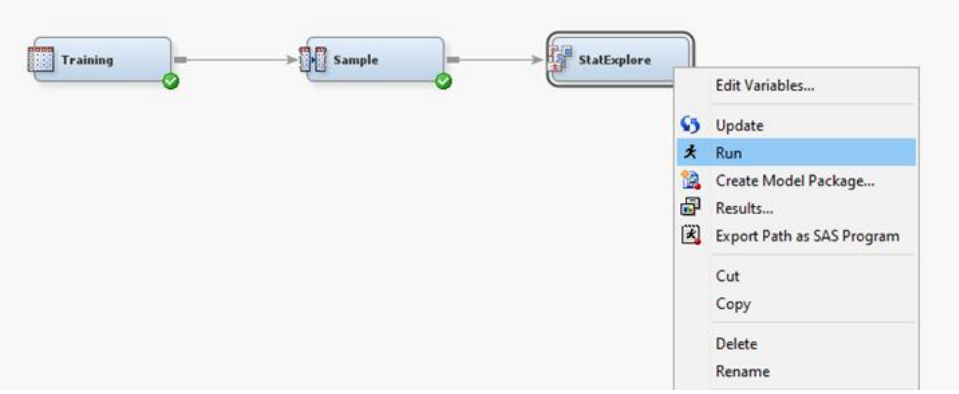


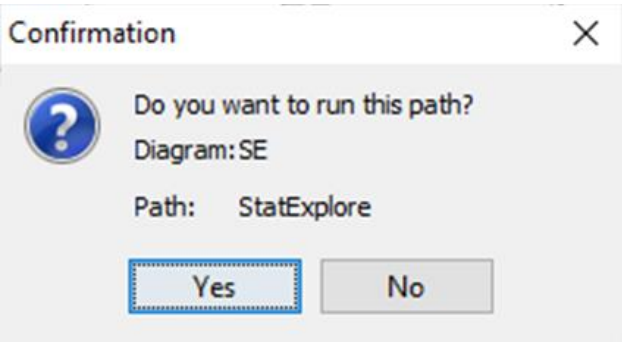


7.3 Explore

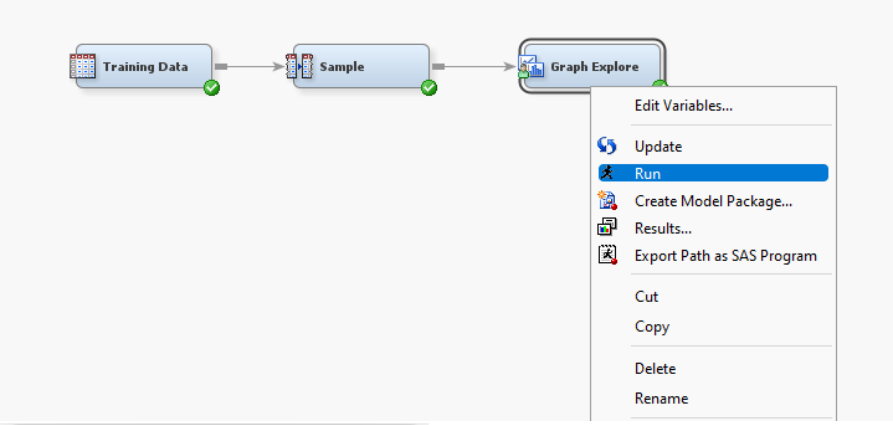
Figures 7.3.1 - 7.3.4 Generate summary statistics

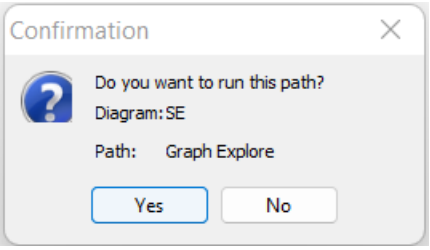


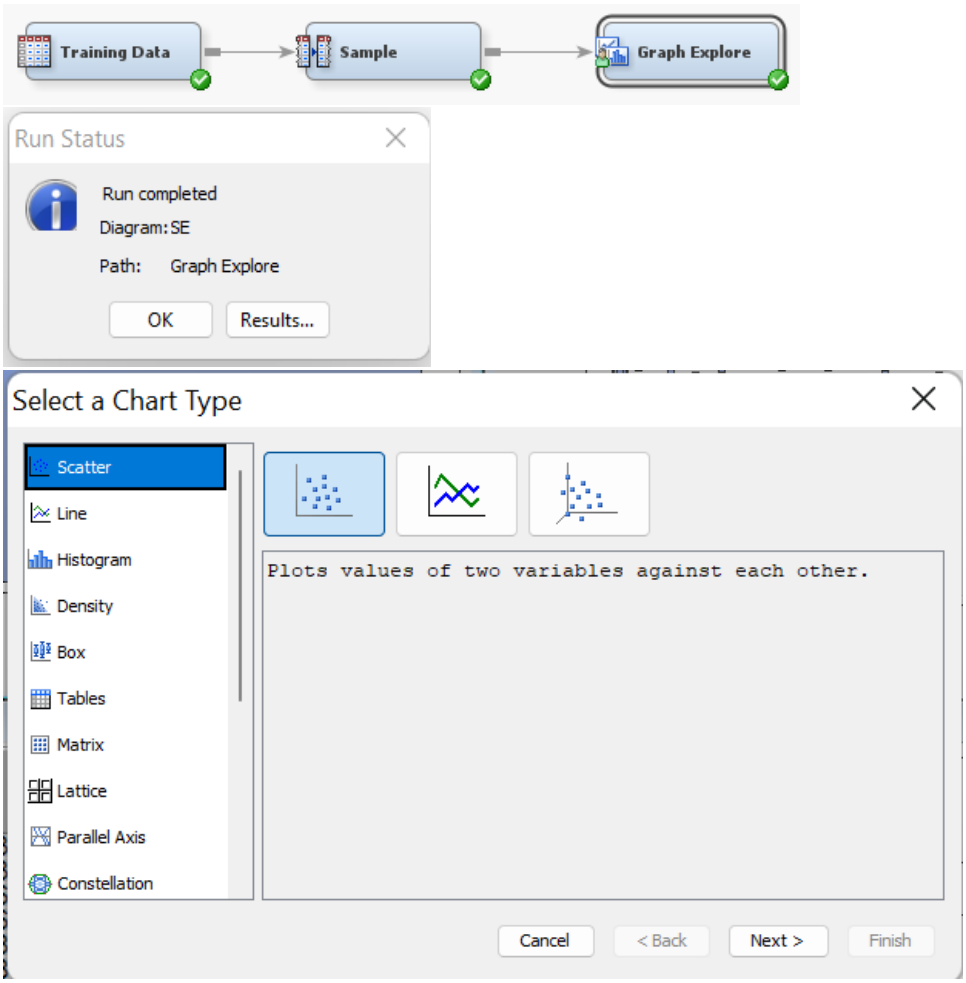




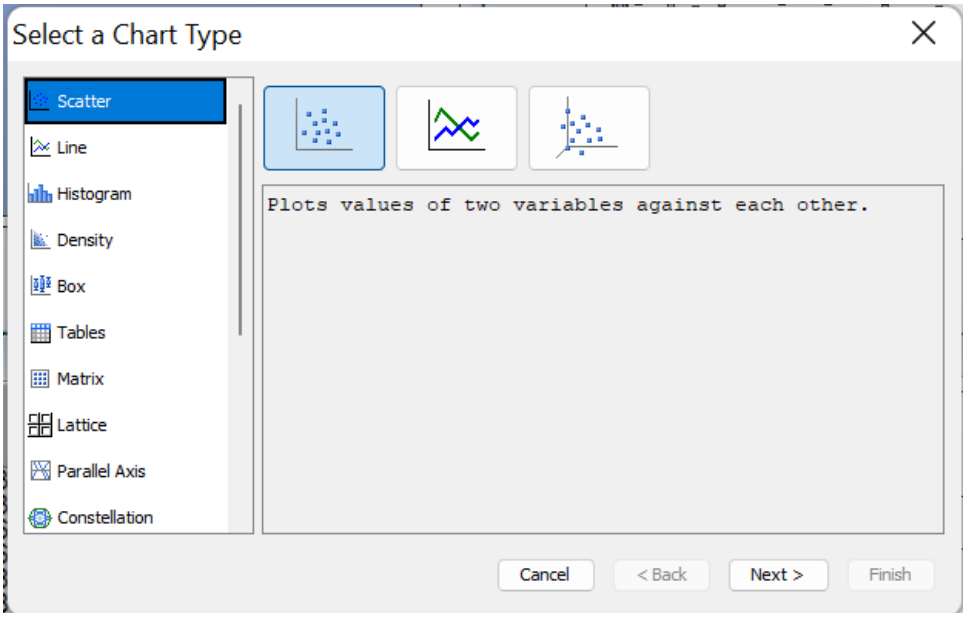
Figures 7.3.5 - 7.3.9: Univariate: Generate graphs (from GraphExplore node results page)

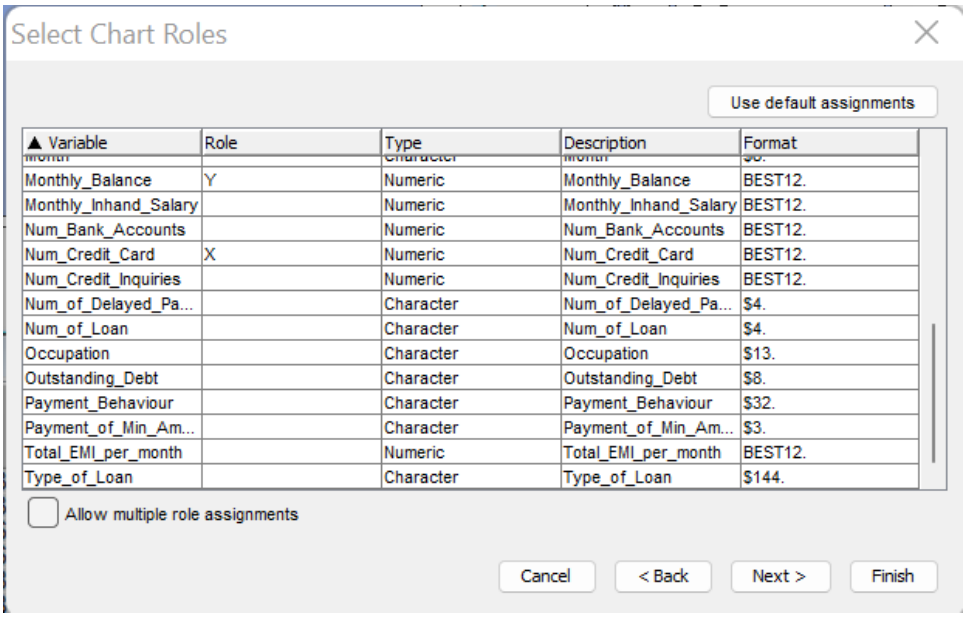




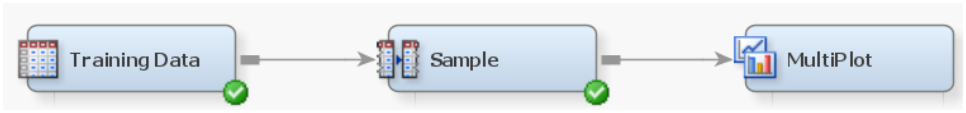


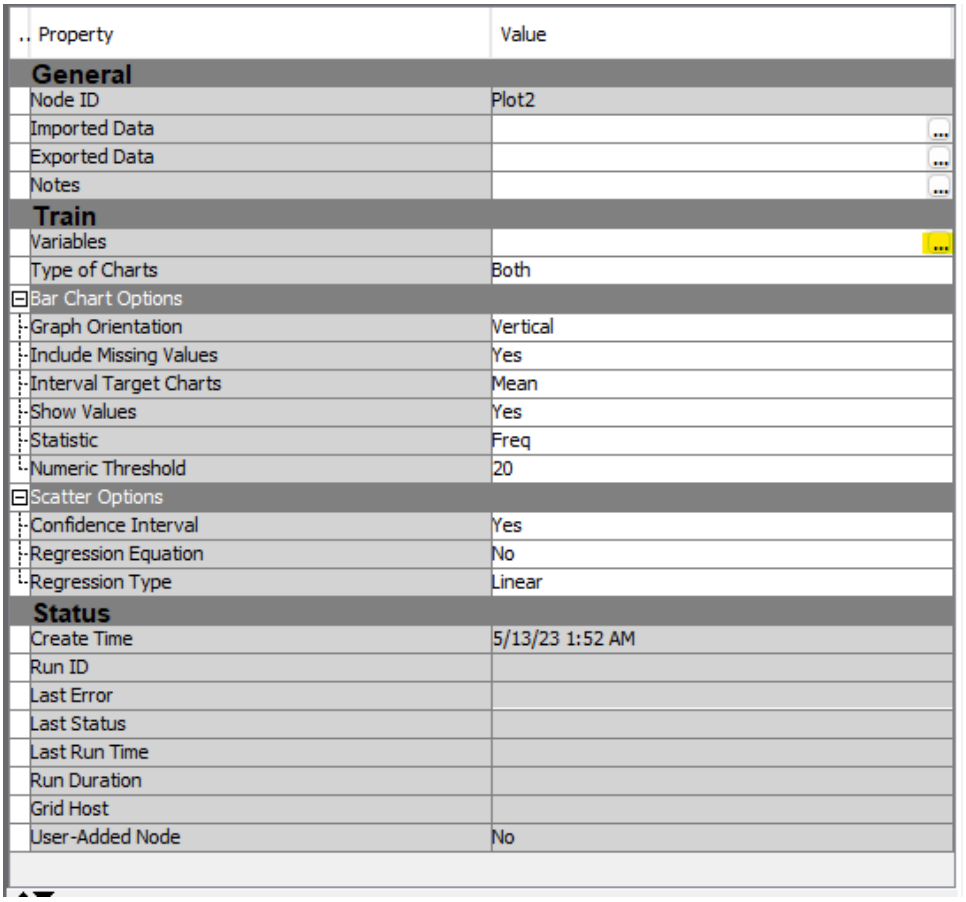
Figures 7.3.10 - 7.3.11: Bivariate (Input Variables): Generate Plot (from GraphExplore node results page)



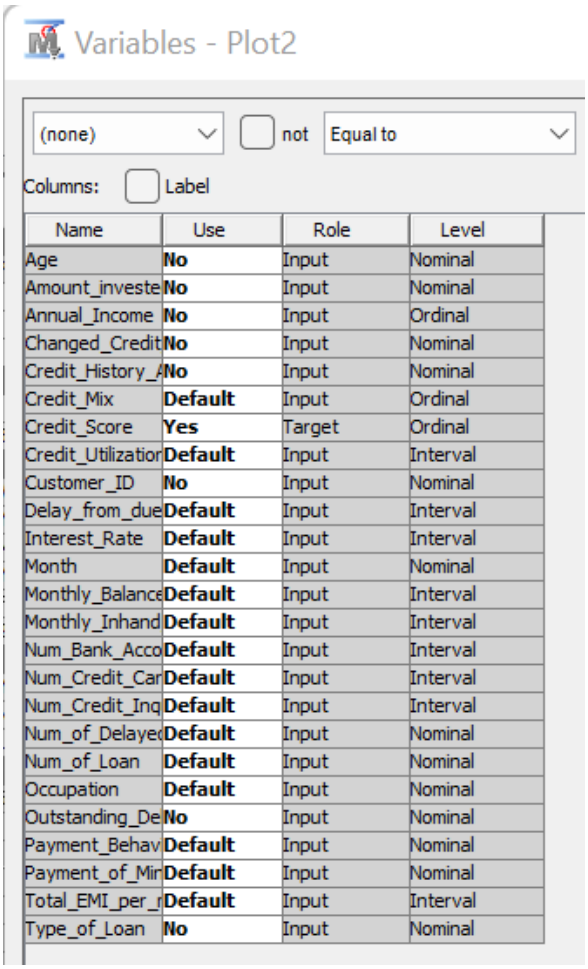


Figures 7.3.12 - 7.3.19: Bivariate (Input and Target Variables): Generate Plot

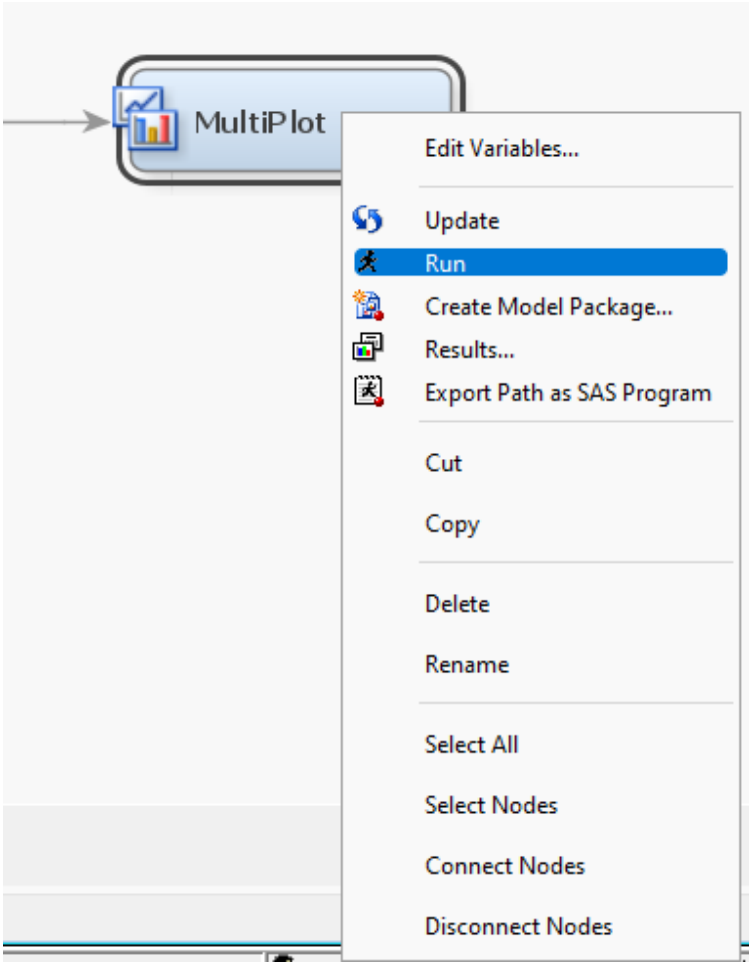


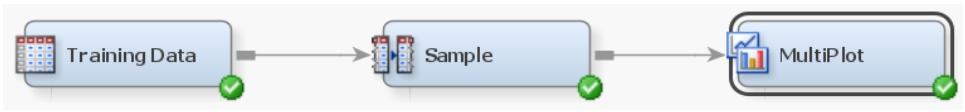


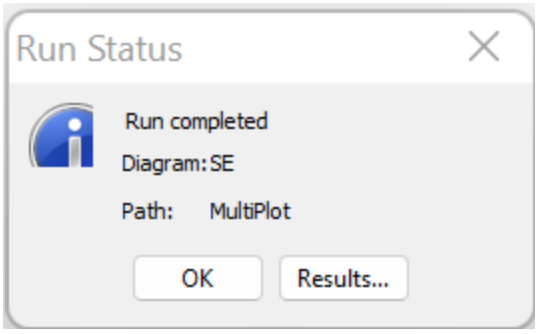
Property	Value
General	
Node ID	Plot2
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Type of Charts	Both
Bar Chart Options	
Graph Orientation	Vertical
Include Missing Values	Yes
Interval Target Charts	Mean
Show Values	Yes
Statistic	Freq
Numeric Threshold	20
Scatter Options	
Confidence Interval	Yes
Regression Equation	No
Regression Type	Linear
Status	
Create Time	5/13/23 1:52 AM
Run ID	
Last Error	
Last Status	
Last Run Time	
Run Duration	
Grid Host	
User-Added Node	No

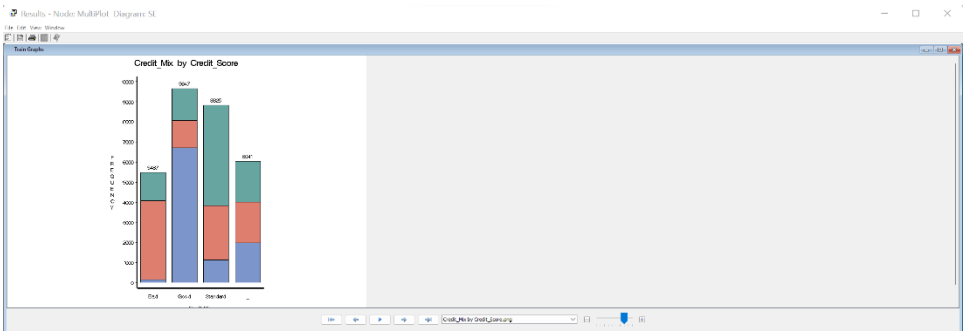


Name	Use	Role	Level
Age	**No**	Input	Nominal
Amount_investe	**No**	Input	Nominal
Annual_Income	**No**	Input	Ordinal
Changed_Credit	**No**	Input	Nominal
Credit_History_A	**No**	Input	Nominal
Credit_Mix	**Default**	Input	Ordinal
Credit_Score	**Yes**	Target	Ordinal
Credit_Utilizatior	**Default**	Input	Interval
Customer_ID	**No**	Input	Nominal
Delay_from_due	**Default**	Input	Interval
Interest_Rate	**Default**	Input	Interval
Month	**Default**	Input	Nominal
Monthly_Balance	**Default**	Input	Interval
Monthly_Inhand	**Default**	Input	Interval
Num_Bank_Acco	**Default**	Input	Interval
Num_Credit_Car	**Default**	Input	Interval
Num_Credit_Inq	**Default**	Input	Interval
Num_of_Delaye	**Default**	Input	Nominal
Num_of_Loan	**Default**	Input	Nominal
Occupation	**Default**	Input	Nominal
Outstanding_De	**No**	Input	Nominal
Payment_Behav	**Default**	Input	Nominal
Payment_of_Mir	**Default**	Input	Nominal
Total_EMI_per_r	**Default**	Input	Interval
Type_of_Loan	**No**	Input	Nominal

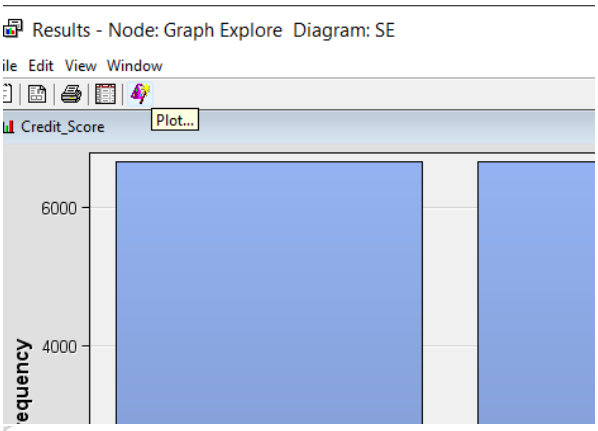

MultiPlot
Edit Variables...
Update
Run
Create Model Package...
Results...
Export Path as SAS Program
Cut
Copy
Delete
Rename
Select All
Select Nodes
Connect Nodes
Disconnect Nodes

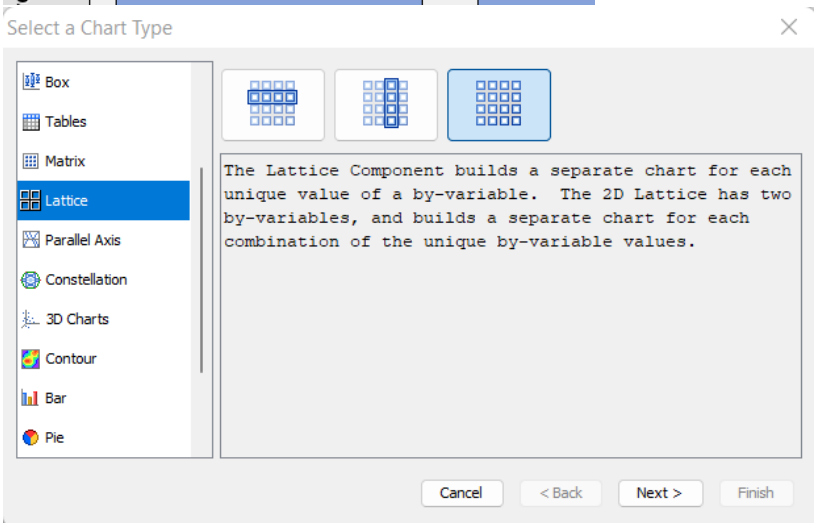

Training Data
Sample
MultiPlot

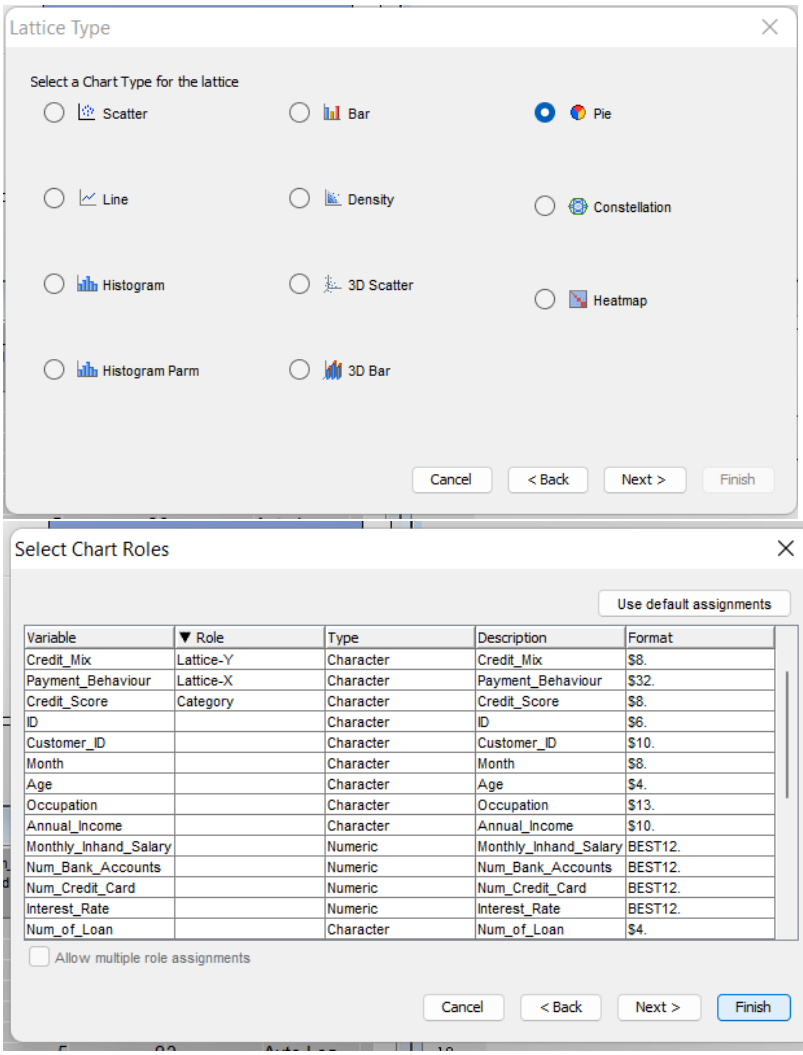

Run Status
Run completed
Diagram: SE
Path: MultiPlot
OK
Results...



Figures 7.3.20 - 7.3.23: Multivariate: Generate a Lattice Plot (from GraphExplore node results page)

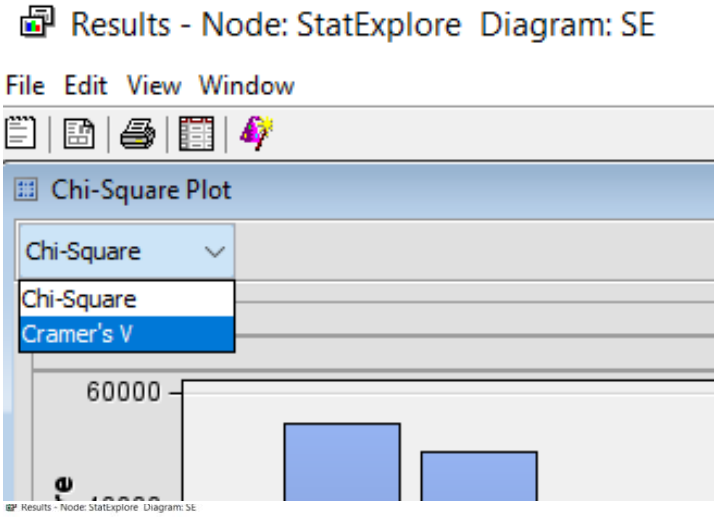


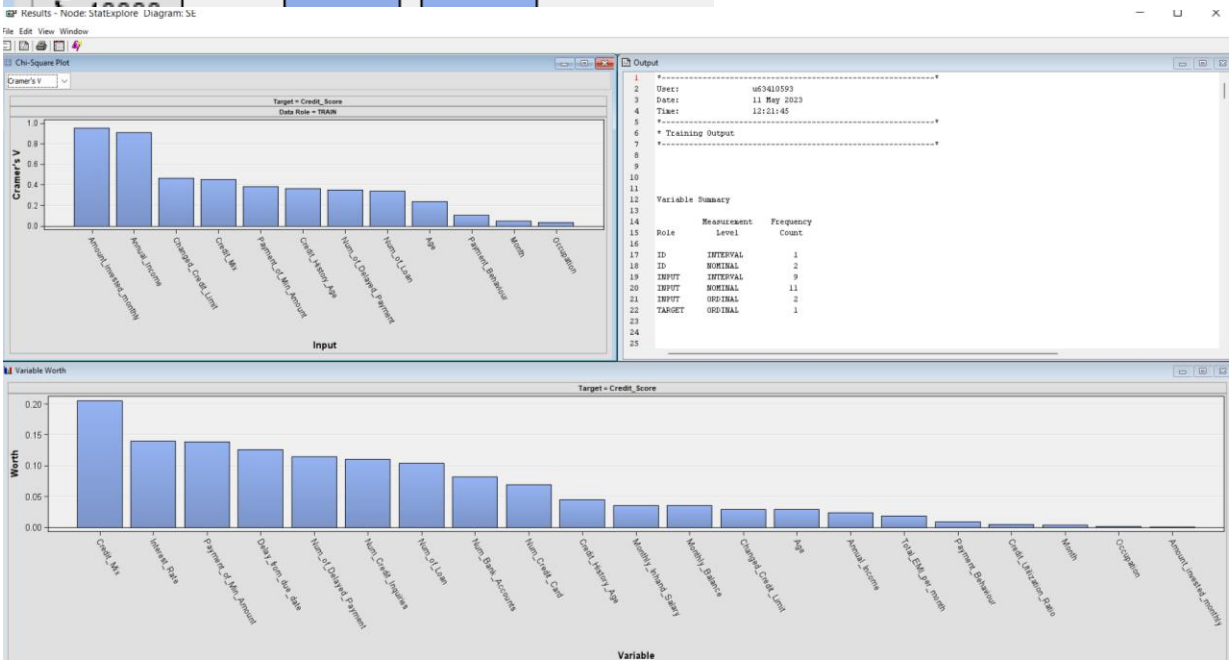



Lattice Type
Select a Chart Type for the lattice
Scatter
Bar
Pie
Line
Density
Constellation
Histogram
3D Scatter
Heatmap
Histogram Parm
3D Bar
Cancel
< Back
Next >
Finish

Select Chart Roles
Use default assignments
Variable
Role
Type
Description
Format
Credit_Mix
Lattice-Y
Character
Credit_Mix
$8.
Payment_Behaviour
Lattice-X
Character
Payment_Behaviour
$32.
Credit_Score
Category
Character
Credit_Score
$8.
ID
Character
ID
$6.
Customer_ID
Character
Customer_ID
$10.
Month
Character
Month
$8.
Age
Character
Age
$4.
Occupation
Character
Occupation
$13.
Annual_Income
Character
Annual_Income
$10.
Monthly_Inhand_Salary
Numeric
Monthly_Inhand_Salary
BEST12.
Num_Bank_Accounts
Numeric
Num_Bank_Accounts
BEST12.
Num_Credit_Card
Numeric
Num_Credit_Card
BEST12.
Interest_Rate
Numeric
Interest_Rate
BEST12.
Num_of_Loan
Character
Num_of_Loan
$4.
Allow multiple role assignments
Cancel
< Back
Next >
Finish

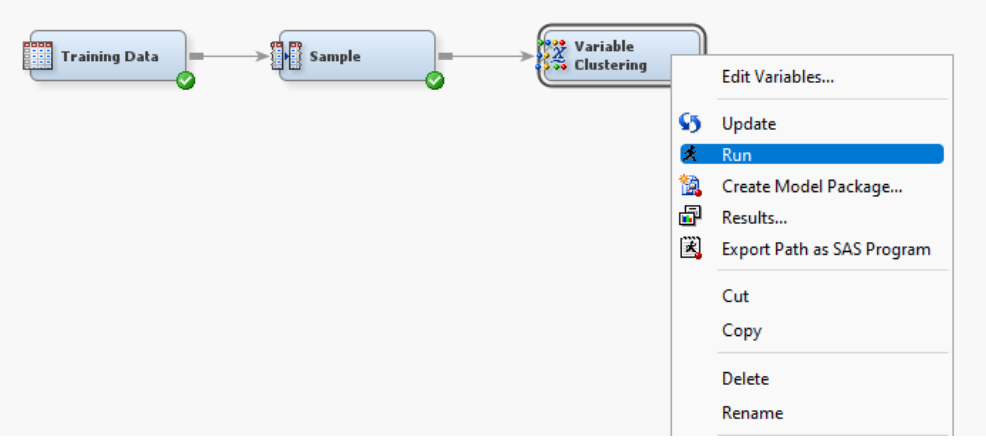
Figures 7.3.24 - 7.3.25: Multivariate: Generate Cramer's V (from StatExplore node results page)

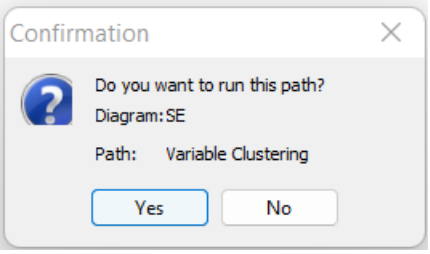


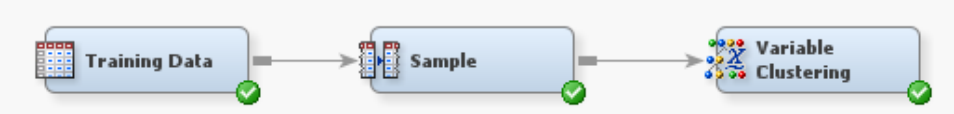


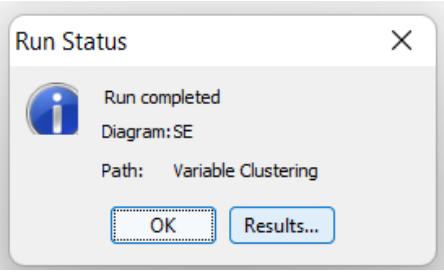
Figures 7.3.26 - 7.3.32: Multivariate: Variable Clustering

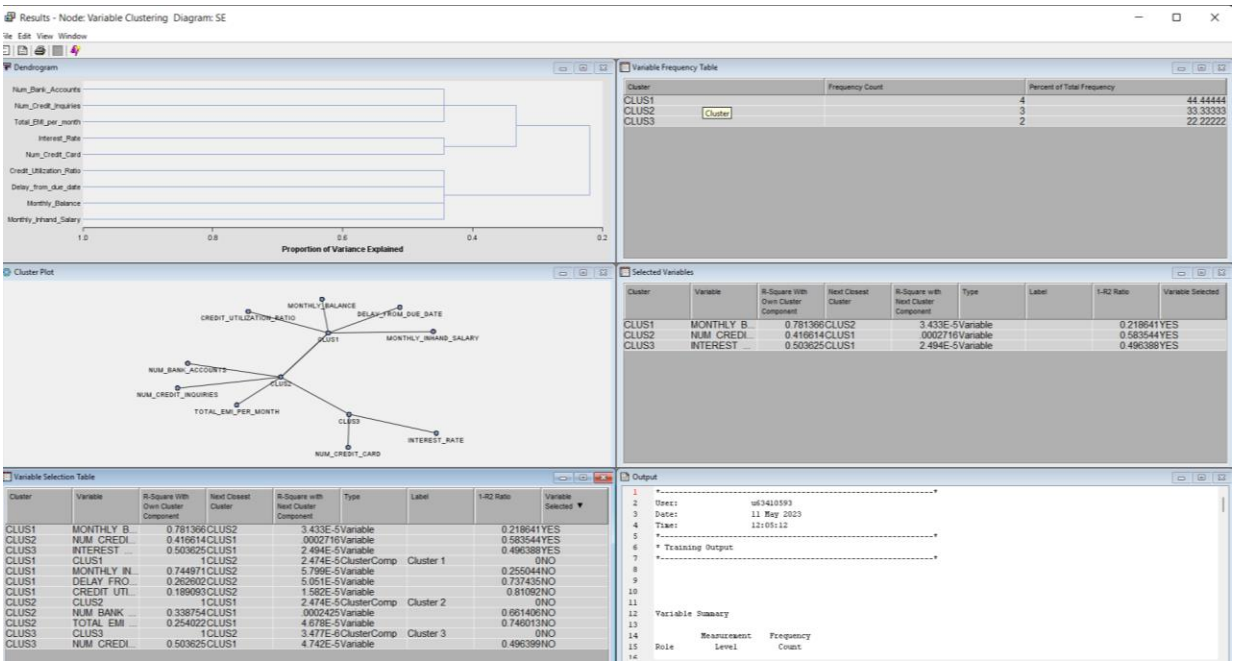
Score	
Variable Selection	Best Variables
Interactive Selection	
Hides Rejected Variables	Yes


Training Data
Sample
Variable Clustering
Edit Variables...
Update
Run
Create Model Package...
Results...
Export Path as SAS Program
Cut
Copy
Delete
Rename

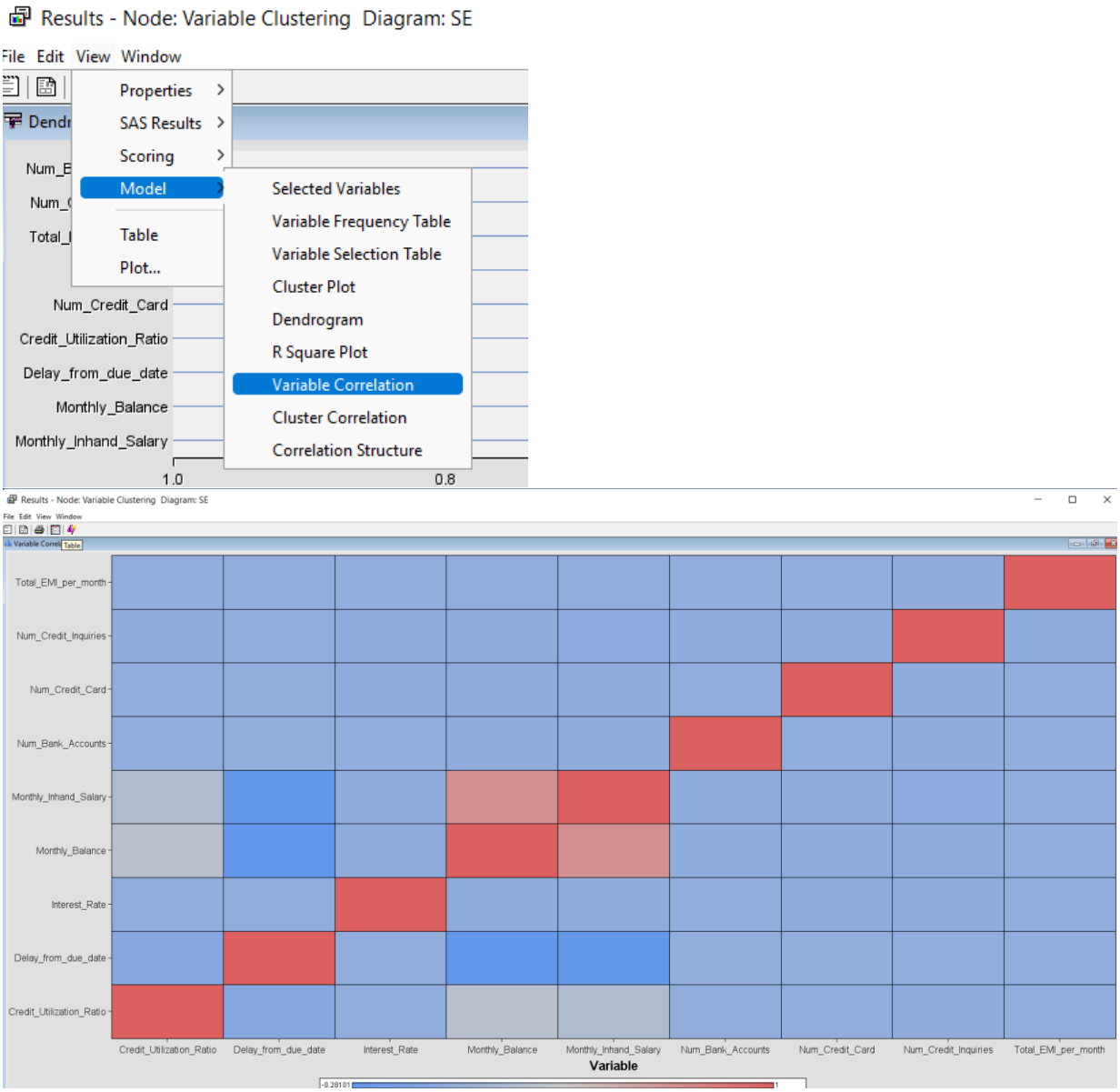

Confirmation
Do you want to run this path?
Diagram: SE
Path: Variable Clustering
Yes
No


Training Data
Sample
Variable Clustering

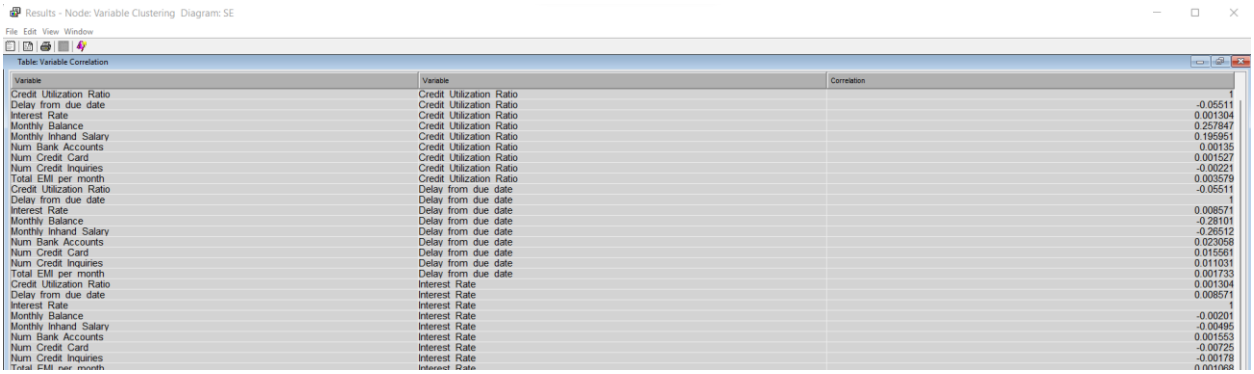

Run Status
Run completed
Diagram: SE
Path: Variable Clustering
OK
Results...



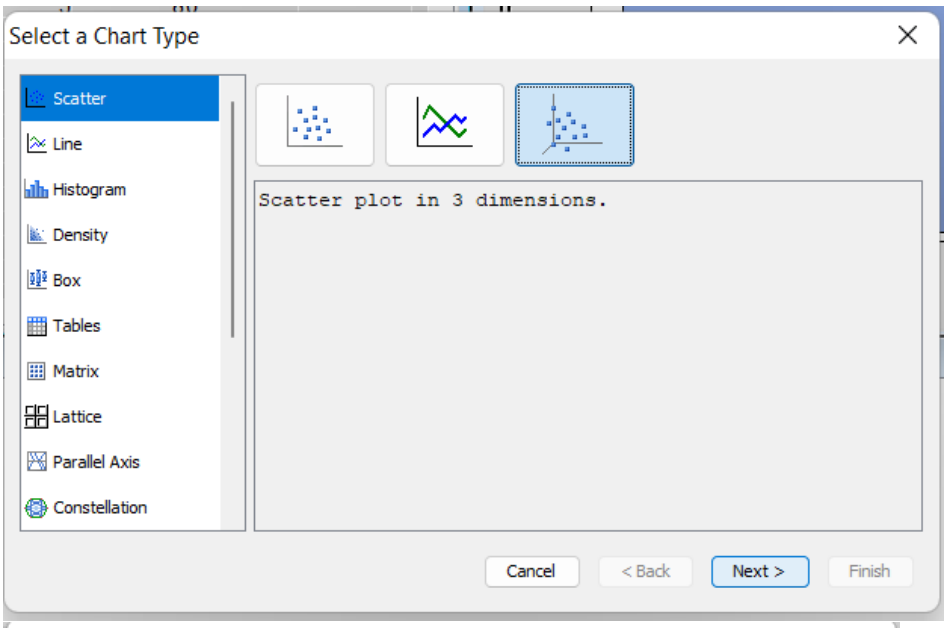
Figures 7.3.33 - 7.3.35: Multivariate: Correlation Matrix (from Variable Clustering node results page)



The correlation table is generated from the correlation matrix page.



Figures 7.3.36 - 7.3.39: Visualizing 3-D Scatter plot (from GraphExplore node plot page)



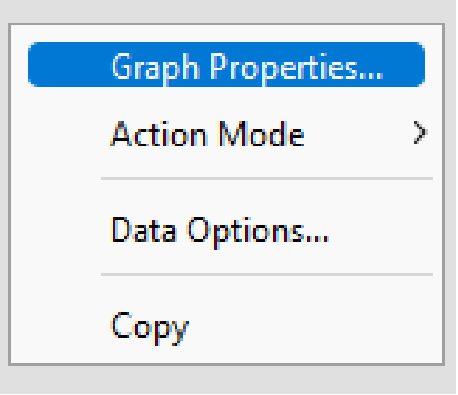
Select Chart Roles

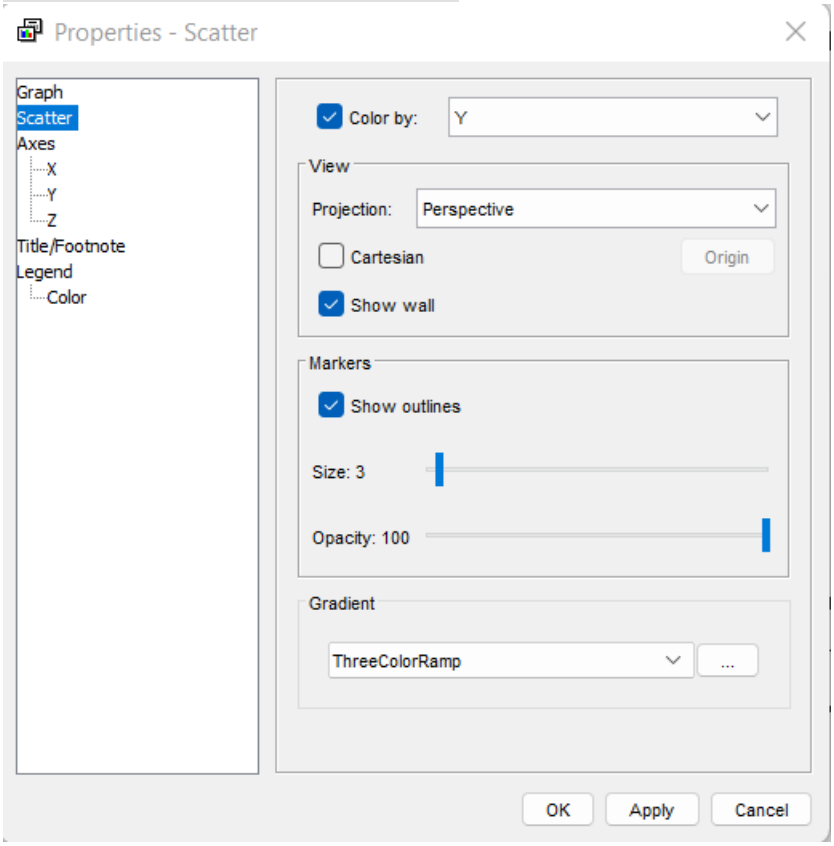
Use default assignments

Variable	▼ Role	Type	Description	Format
Monthly_Balance	Z	Numeric	Monthly_Balance	BEST12.
Delay_from_due_date	Y	Numeric	Delay_from_due_date	BEST12.
Monthly_Inhand_Salary	X	Numeric	Monthly_Inhand_Salary	BEST12.
ID		Character	ID	$6.
Customer_ID		Character	Customer_ID	$10.
Month		Character	Month	$8.
Age		Character	Age	$4.
Occupation		Character	Occupation	$13.
Annual_Income		Character	Annual_Income	$10.
Num_Bank_Accounts		Numeric	Num_Bank_Accounts	BEST12.
Num_Credit_Card		Numeric	Num_Credit_Card	BEST12.
Interest_Rate		Numeric	Interest_Rate	BEST12.
Num_of_Loan		Character	Num_of_Loan	$4.
Type_of_Loan		Character	Type_of_Loan	$144.

Allow multiple role assignments

Cancel < Back Next > Finish


Graph Properties...
Action Mode
Data Options...
Copy


Properties - Scatter
Graph
Scatter
Axes
X
Y
Z
Title/Footnote
Legend
Color
Color by:
Y
View
Projection:
Perspective
Cartesian
Origin
Show wall
Markers
Show outlines
Size: 3
Opacity: 100
Gradient
ThreeColorRamp
...
OK
Apply
Cancel