Descriptive Analytics with Claim Fraud Dataset/ SAS Enterprise Miner

Didem B. Aykurt

Colorado State University Global

MIS530; Predictive Analytics

Dr.Jennifer Catalano

February 5, 2023

Descriptive Analytics

Descriptive analytics is the essential step of data analysis. This stage has a few critical topics; the summary measures of central tendency include the mean, median, and mode and historical data to understand better standard deviation, variance, range, and the kurtosis and skewness and prepare the data for predictive analytics. For example, if a variable is highly skewed, the variable may need to be normalized to produce a more accurate model.

Additionally, this project will touch on statistical correlation methods to develop and prepare the dataset for predictive models. Discuss the main topics with the Claim Fraud Dataset's summary statistics in the StatExplore node results Output Window.

Interval Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN

Variable	Role	Mean	Standard Deviation	Non Missing	Missing	Minimum	Median	Maximum	Skewness	Kurtosis
Annual_Premium	INPUT	1129.889	312.4143	2998	0	600	1128	1680	0.04341	-1.1642
Claim_Amount	INPUT	797.4217	688.7673	2998	0	189.8684	578.1016	7422.852	3.05889	13.69692
Income	INPUT	38326.89	34559.26	2998	0	0	34510	933288	5.967957	148.4955
Monthly_Premium	INPUT	94.15744	26.03452	2998	0	50	94	140	0.04341	-1.1642
Months_Since_Last_Claim	INPUT	15.06571	11.1	2998	0	0	14	60	0.685541	0.173394
Months_Since_Policy_Inception	INPUT	48.27085	28.1729	2998	0	0	48	99	0.046468	-1.14612
Outstanding_Balance	INPUT	23799.85	13826.52	2994	4	27	24052	47996	0.003073	-1.18935

2:24 PM
2/2/2023

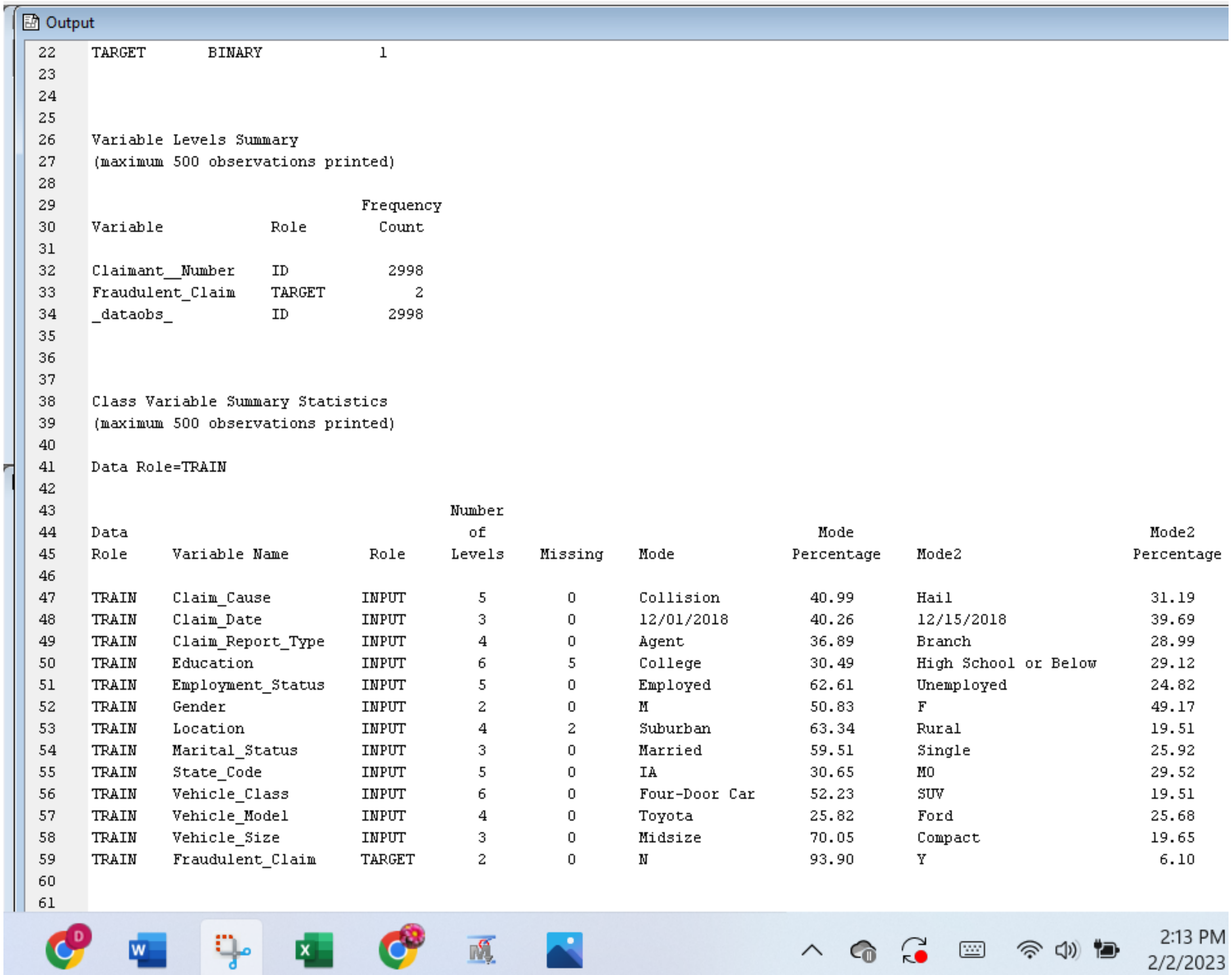
Figure 1: StatExplore Output Window- class variable summary statistics

Mean

The first thing to check the mean value is to see an average data point for each variable. Outliers and high values affect the mean values. For example, in the claim fraud dataset, 'Income' has the highest mean of $30326.89, which shows the average income is a little high, and 'Annual_Prenium' has a higher value of 1129.889.

Median

The median is preferably measured at a distance from the mean if there are outliers or high variability in the dataset. The data is usually distributed if the mean and median length are small. Outliers and high values do not affect the median.



```
Output

TARGET       BINARY             1



Variable Levels Summary
(maximum 500 observations printed)

                              Frequency
Variable            Role        Count

Claimant__Number    ID           2998
Fraudulent_Claim    TARGET          2
_dataobs_           ID           2998



Class Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN

                                         Number
Data                                       of                                 Mode                                Mode2
Role     Variable Name         Role      Levels    Missing    Mode            Percentage    Mode2                 Percentage

TRAIN    Claim_Cause           INPUT        5         0       Collision          40.99      Hail                     31.19
TRAIN    Claim_Date            INPUT        3         0       12/01/2018         40.26      12/15/2018               39.69
TRAIN    Claim_Report_Type     INPUT        4         0       Agent              36.89      Branch                   28.99
TRAIN    Education             INPUT        6         5       College            30.49      High School or Below     29.12
TRAIN    Employment_Status     INPUT        5         0       Employed           62.61      Unemployed               24.82
TRAIN    Gender                INPUT        2         0       M                  50.83      F                        49.17
TRAIN    Location              INPUT        4         2       Suburban           63.34      Rural                    19.51
TRAIN    Marital_Status        INPUT        3         0       Married            59.51      Single                   25.92
TRAIN    State_Code            INPUT        5         0       IA                 30.65      MO                       29.52
TRAIN    Vehicle_Class         INPUT        6         0       Four-Door Car      52.23      SUV                      19.51
TRAIN    Vehicle_Model         INPUT        4         0       Toyota             25.82      Ford                     25.68
TRAIN    Vehicle_Size          INPUT        3         0       Midsize            70.05      Compact                  19.65
TRAIN    Fraudulent_Claim      TARGET       2         0       N                  93.90      Y                         6.10

2:13 PM
2/2/2023
```

Figure 2: StatExplore Output Window- Class variable summary statistics

Mode

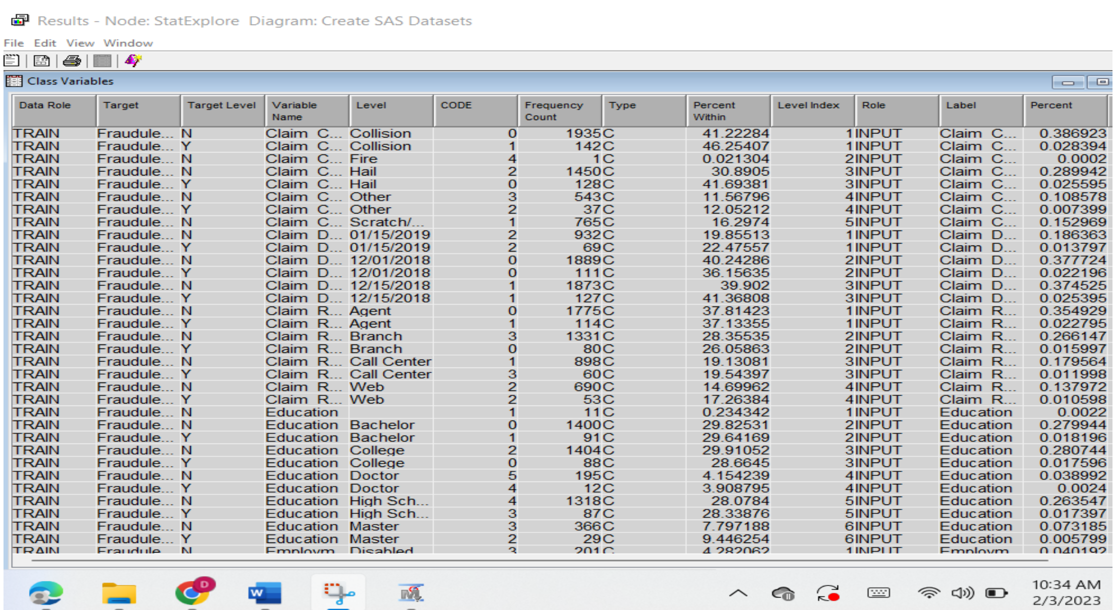
The mode is the most frequent data point that is rarely used, primarily as if categorical or class variables have missing values and can be replaced with mode values. The highest percentage of mode values for the 'Employement_Status' category 'Employed' at 62.61 and the 'Vehicle_Size' category 'midsize' at 70.05.

Variance and Distribution

The result of variance and distribution shows how data points spread that measure range differ between the minimum and maximum values. Outliers and high values affect the range. The most helpful measure is the sample variance, the average squared deviations of each observation from the mean. Another standard deviation measure is the square root of the variance and is in the same units of measurement as the original data.

Figure 1 shows the summary statistics of the interval variables. The 'Claim_Amount' mean was $797.4217, with a standard deviation of $688.7673 and a median of $578.1016.

We can see all the class variables with target variables by Train dataset; click View tab- Summary Statistics- Class Variables from the StatExplore Output Window.



Data Role	Target	Target Level	Variable Name	Level	CODE	Frequency Count	Type	Percent Within	Level Index	Role	Label	Percent
TRAIN	Fraudule...	N	Claim C...	Collision	0	1935	C	41.22284	1	INPUT	Claim C...	0.386923
TRAIN	Fraudule...	Y	Claim C...	Collision	1	142	C	46.25407	1	INPUT	Claim C...	0.028394
TRAIN	Fraudule...	N	Claim C...	Fire	4	1	C	0.021304	2	INPUT	Claim C...	0.0002
TRAIN	Fraudule...	N	Claim C...	Hail	2	1450	C	30.8905	3	INPUT	Claim C...	0.289942
TRAIN	Fraudule...	Y	Claim C...	Hail	0	128	C	41.69381	3	INPUT	Claim C...	0.025595
TRAIN	Fraudule...	N	Claim C...	Other	3	543	C	11.56796	4	INPUT	Claim C...	0.108578
TRAIN	Fraudule...	Y	Claim C...	Other	2	37	C	12.05212	4	INPUT	Claim C...	0.007399
TRAIN	Fraudule...	N	Claim C...	Scratch/...	1	765	C	16.2974	5	INPUT	Claim C...	0.152969
TRAIN	Fraudule...	N	Claim D...	01/15/2019	2	932	C	19.85513	1	INPUT	Claim D...	0.186363
TRAIN	Fraudule...	Y	Claim D...	01/15/2019	2	69	C	22.47557	1	INPUT	Claim D...	0.013797
TRAIN	Fraudule...	N	Claim D...	12/01/2018	0	1889	C	40.24286	2	INPUT	Claim D...	0.377724
TRAIN	Fraudule...	Y	Claim D...	12/01/2018	0	111	C	36.15635	2	INPUT	Claim D...	0.022196
TRAIN	Fraudule...	N	Claim D...	12/15/2018	1	1873	C	39.902	3	INPUT	Claim D...	0.374525
TRAIN	Fraudule...	Y	Claim D...	12/15/2018	1	127	C	41.36808	3	INPUT	Claim D...	0.025395
TRAIN	Fraudule...	N	Claim R...	Agent	0	1775	C	37.81423	1	INPUT	Claim R...	0.354929
TRAIN	Fraudule...	Y	Claim R...	Agent	1	114	C	37.13355	1	INPUT	Claim R...	0.022795
TRAIN	Fraudule...	N	Claim R...	Branch	3	1331	C	28.35535	2	INPUT	Claim R...	0.266147
TRAIN	Fraudule...	Y	Claim R...	Branch	0	80	C	26.05863	2	INPUT	Claim R...	0.015997
TRAIN	Fraudule...	N	Claim R...	Call Center	1	898	C	19.13081	3	INPUT	Claim R...	0.179564
TRAIN	Fraudule...	Y	Claim R...	Call Center	3	60	C	19.54397	3	INPUT	Claim R...	0.011998
TRAIN	Fraudule...	N	Claim R...	Web	2	690	C	14.69962	4	INPUT	Claim R...	0.137972
TRAIN	Fraudule...	Y	Claim R...	Web	2	53	C	17.26384	4	INPUT	Claim R...	0.010598
TRAIN	Fraudule...	N	Education		1	11	C	0.234342	1	INPUT	Education	0.0022
TRAIN	Fraudule...	N	Education	Bachelor	0	1400	C	29.82531	2	INPUT	Education	0.279944
TRAIN	Fraudule...	Y	Education	Bachelor	1	91	C	29.64169	2	INPUT	Education	0.018196
TRAIN	Fraudule...	N	Education	College	2	1404	C	29.91052	3	INPUT	Education	0.280744
TRAIN	Fraudule...	Y	Education	College	0	88	C	28.6645	3	INPUT	Education	0.017596
TRAIN	Fraudule...	N	Education	Doctor	5	195	C	4.154239	4	INPUT	Education	0.038992
TRAIN	Fraudule...	Y	Education	Doctor	4	12	C	3.908795	4	INPUT	Education	0.0024
TRAIN	Fraudule...	N	Education	High Sch...	4	1318	C	28.0784	5	INPUT	Education	0.263547
TRAIN	Fraudule...	Y	Education	High Sch...	3	87	C	28.33876	5	INPUT	Education	0.017397
TRAIN	Fraudule...	N	Education	Master	3	366	C	7.797188	6	INPUT	Education	0.073185
TRAIN	Fraudule...	Y	Education	Master	2	29	C	9.446254	6	INPUT	Education	0.005799
TRAIN	Fraudule...	N	Employm...	Disabled	3	201	C	4.282062	1	INPUT	Employm...	0.040192

Figure 3: StatExplore node-class variable results

The class variable table shows a count of each class variable by target variable of Y or N as 2952 observations had an Employed for the 'Employemet_Status' variable and with no (N) 'Fraudulent_Claim' and 174 observations had an Employed level for the 'Employemet_Status' variable with yes (Y) 'Fraudulent_Claim.'

Skewness

Skewness tells that the dataset has an asymmetric distribution or is not symmetrical. There are three different distributions. One zero skew means the distribution is balanced (mean=median). Another negative skew is when the skewness value is negative (mean<median), and a positive skew is when the skewness value is positive(mean>median). For example, the skewness of temp shows a negative number of -0.33, which tells the left skew.

Figure 1 shows the average 'Annual_Prenium, 'Monthly_Prenium,' 'Months_Science_Last_Claim' and 'Months_Science_Policy_Inception" have close values for mean, and the median would have a normal distribution. 'Outstanding_Balance' has a left-skewed distribution. And 'Claim_Amount' and 'Income' have a right-skewed distribution.

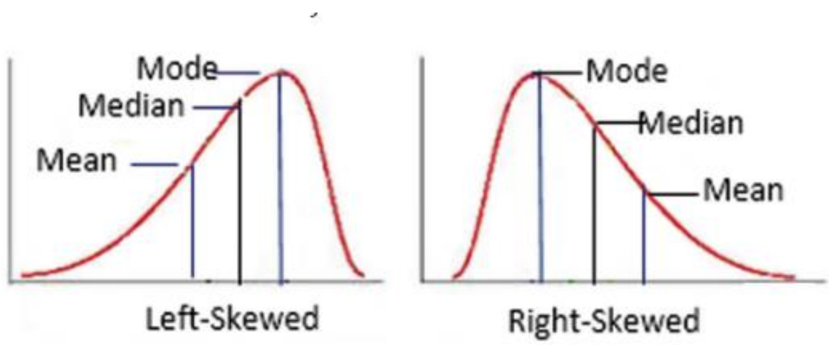


Fig. 3.10 Skewness

$$\frac{n}{(n-1)(n-2)}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s}\right)$$

Where n is the sample size,
x_i is the i^{th} value of the variable,
$\bar{x}$ is the sample average, and
s is the sample standard deviation

Fig. 3.11 Skewness Formula

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

Kurtosis

Kurtosis shows the variable's probability or frequency, which also helps to compare which variable has a heavy distribution tail with three kurtosis types. I found so many different ranges people use. I want to use zero for the normal kurtosis distribution because I check outliers, and the best explanation for the case outliers is the zero number for kurtosis. Medium tails are **mesokurtic(kurtosis=0)**, low kurtosis is **platykurtic(kurtosis<0)**, and high kurtosis is **leptokurtic(kurtosis>0).**

Figure 1 shows the 'Moths_Scince_Last_Claim' kurtosis value was 0.05, close to the zero would be mesokurtic, and the 'Income' kurtosis of 4790, which is pretty high, is leptokurtic.

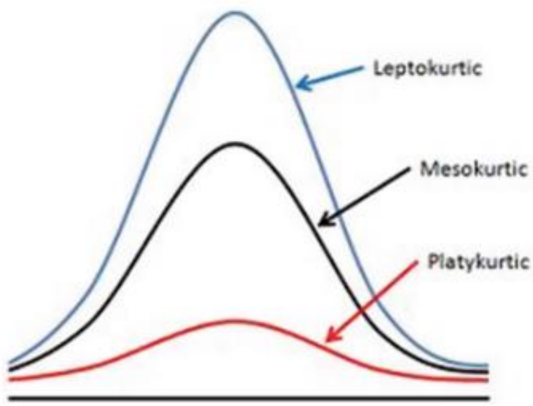


Fig. 3.12 Kurtosis taken from https://www.bogleheads.org/wiki/Excess_kurtosis

$$\frac{n(n+1)}{(n-1)(n-2)(n-3)}\sum_{i=1}^{n}\left(\frac{(x_i-\bar{X}}{S}\right)^4-\frac{3(n-1)^2}{(n-2)(n-3)}$$

Where n is the sample size,
x_i is the i^{th} value of the variable,
$\bar{x}$ is the sample average, and
s is the sample standard deviation

Fig. 3.13 Kurtosis formula

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

The best way to solve skewness and kurtosis problems with transformations is when there is a relatively wide range of values instead of a relatively small range. The log transformation transforms skewed data to follow an approximately normal distribution. "For the variable claim_amount, the skewness value was 2.922 and kurtosis 12.62. Notice that the skewness and kurtosis values for income were $68.48667 and $4790.542, respectively. Income is highly right-skewed with a leptokurtic shape. Income should be transformed to provide a more accurate model. The Transform node can be used to modify the income variable. "(McCarthy,2022)

Click the Modify tab, drag and drop Transform Variables on the process diagram, and connect the Filter node.

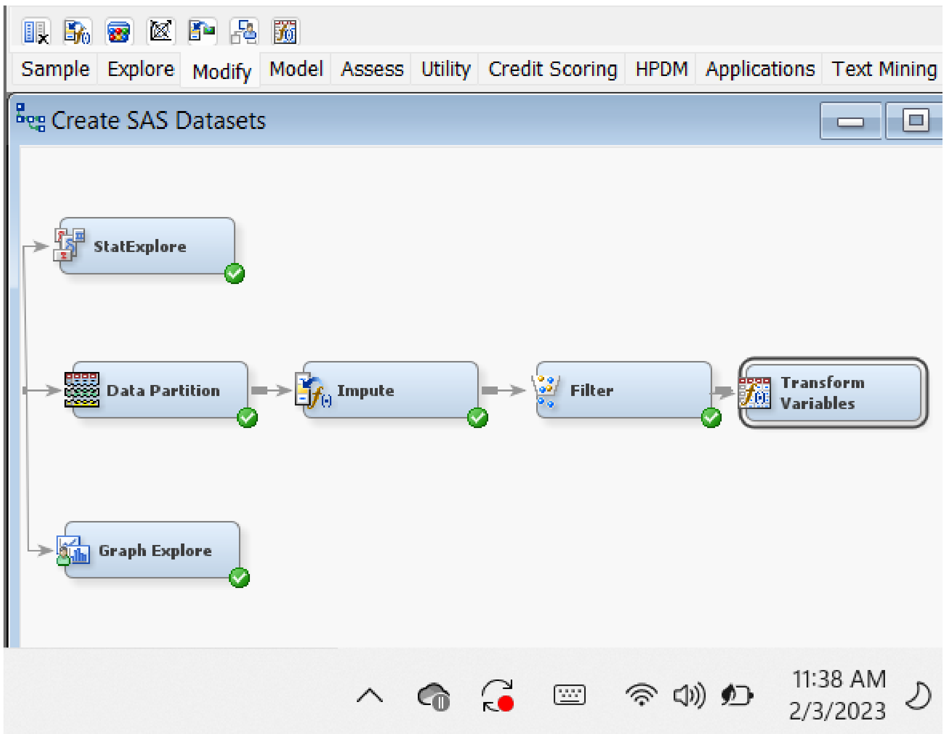


Figure 4: Transform Variables node

The left side has a Transform Properties window; click to ellipse Formulas on the Train group.

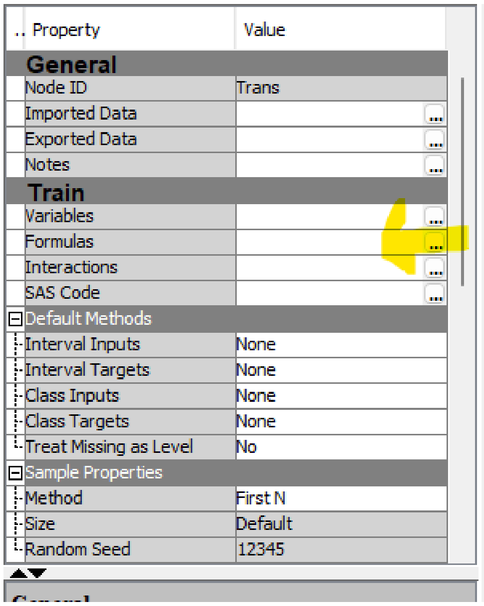


Figure 5: Transform Variables properties

Then, a new window will pop up, and you can choose any variable to see the histogram; the income variable has a right-skewed.

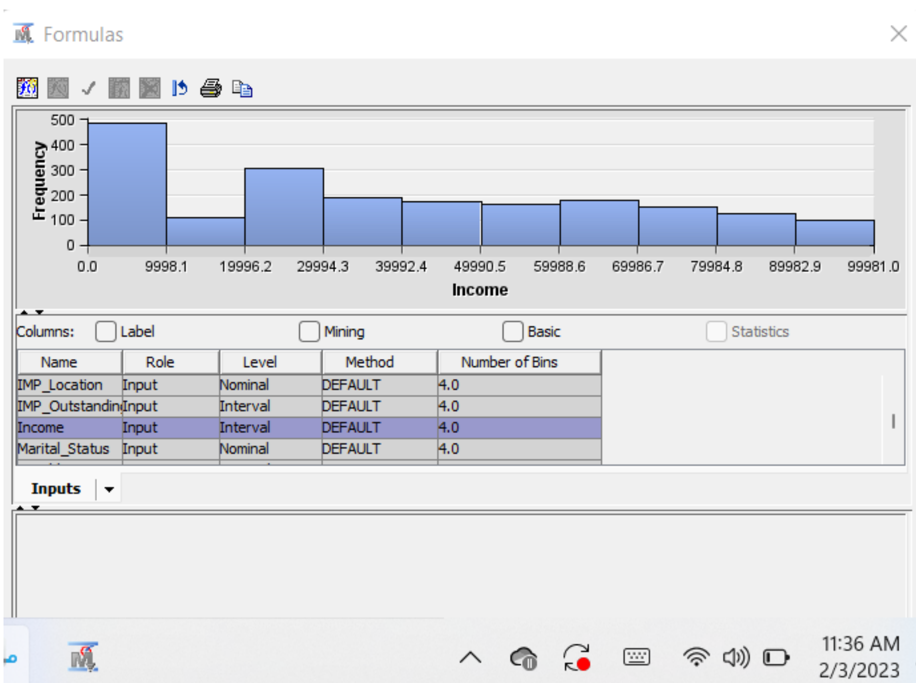


Figure 6: Formulas window- Income distribution

Click the ellipse button on the Train group's Variable; a new Variables-Trans window will open and change the method from Default to Log for the Income variable. Next, we should update the Transform Properties window, set the Default Methods for Interval Inputs to None, and Run the Transform Variables node.

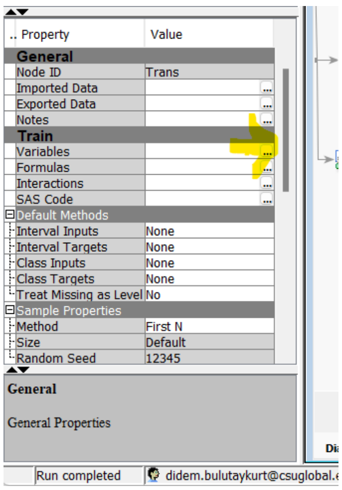


Figure 7: Transform Properties Window

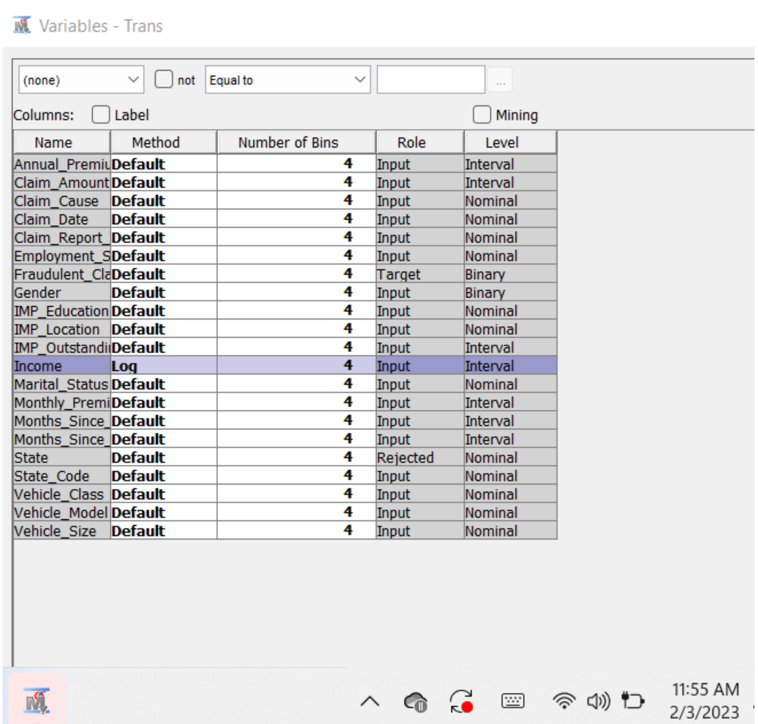
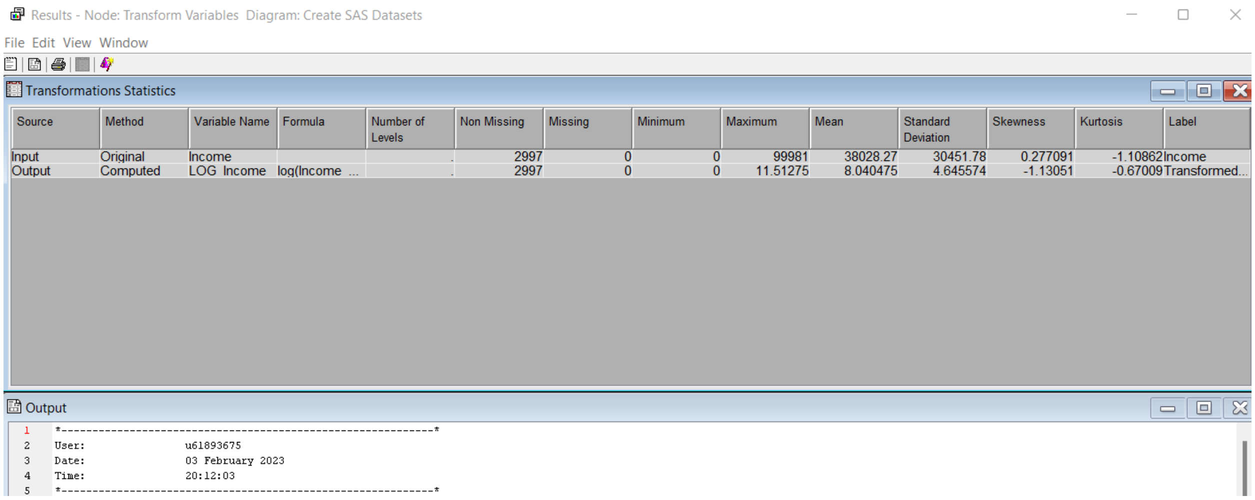


Figure 8: Transform Variables window.

The result of the Transform node is that the Income skew ness value is close to zero now that

Income is a normal distribution variable.


Results - Node: Transform Variables Diagram: Create SAS Datasets

File Edit View Window

Transformations Statistics

Source	Method	Variable Name	Formula	Number of Levels	Non Missing	Missing	Minimum	Maximum	Mean	Standard Deviation	Skewness	Kurtosis	Label
Input	Original	Income		.	2997	0	0	99981	38028.27	30451.78	0.277091	-1.10862	Income
Output	Computed	LOG_Income	log(Income ...	.	2997	0	0	11.51275	8.040475	4.645574	-1.13051	-0.67009	Transformed...

Output

```
1  *------------------------------------------------------------*
2  User:                u61893675
3  Date:                03 February 2023
4  Time:                20:12:03
5  *------------------------------------------------------------*
```

Figure 9: Results of Transform node

Covariance and Correlation

The covariance measures two variables, x (input, independent variable) and y (target, dependent variable), so the correlation (r) ranges from -1 to +1. If the covariance is more significant than zero, a positive relation means two variables move in the same direction; with less than zero negative relation, two variables move in the opposite direction with equal zero X and Y nonrelation independent. " The covariance value is the product of the two variables and is not a standardized unit of measurement. So, measuring the degree to which the variables move together is impossible." (McCarthy,2022)

The square correlation (r^2) or the coefficient of determination measures the percent of the variation in the target variable by the input variable. The R-square range is 0% to 100%. Mostly, use the scatter plot chart to see the plot of the two variables.

If two variables have a strong correlation, there should be some multicollinearity that negatively affects the predictive model. We can solve this problem with SAS Enterprise Miner; click the Explore tab, drag and drop the Variable Cluster node on the diagram, and Run. See the Result table; click View tab- Model- Variable Correlation, then pop up the Variable Correlation window. There is a tab icon showing the list. Again, there is no collinearity to warrant concern.

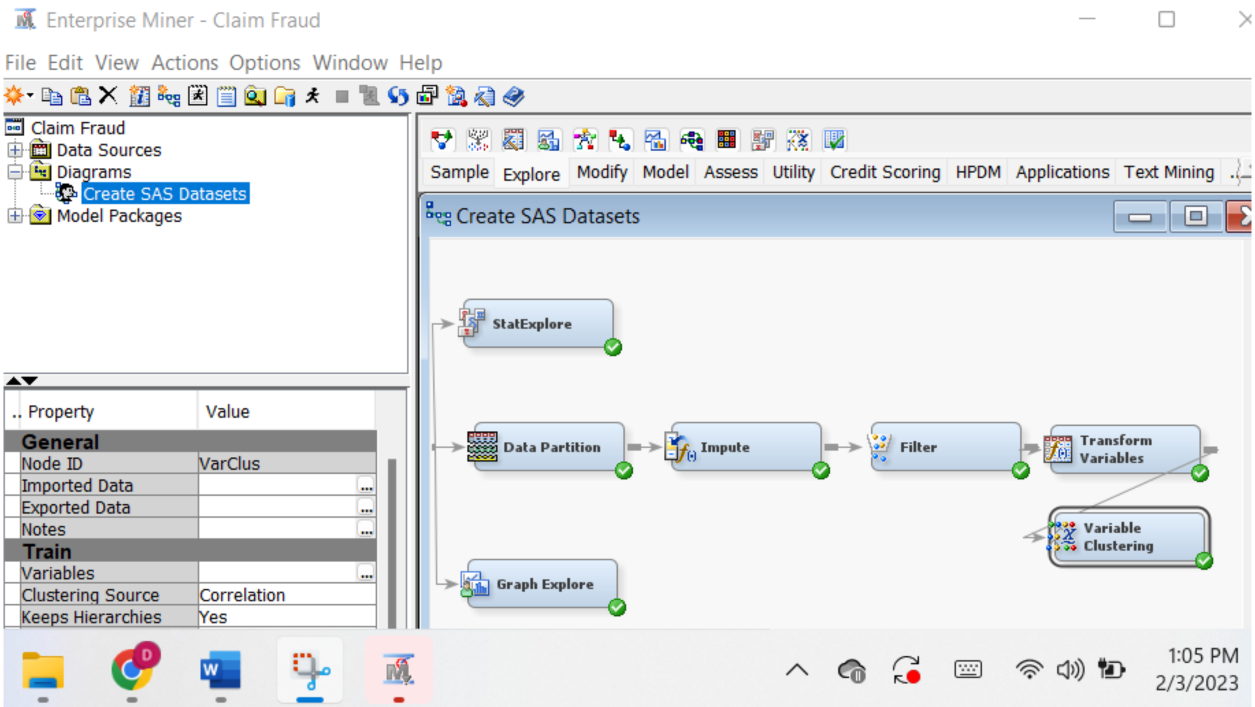


Figure 10: Variable Cluster node

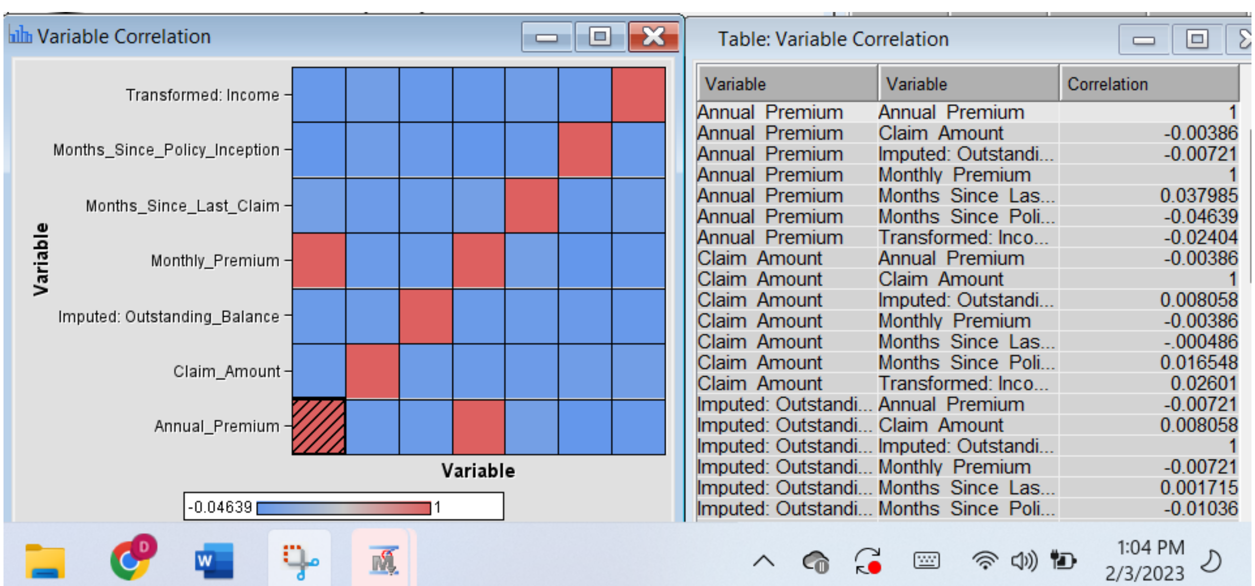


Figure 10: Variable correlation matrix and table result- Variable Cluster

Variable Reduction

This is the excellent part I read because when completed claim fraud data hadn't multicollinearity, I asked if had how to solve it. That part answered my question. If the dataset has multicollinearity or more variables, reducing the number of variables can decrease multicollinearity, redundancy, and irrelevancy and improve the model result. Variable Clustering and Principal Component analysis solve the multicollinearity problem.

Variable Clustering

Variable clustering measures the correlations and covariances between the input variables and creates close data point groups or similar variables. The main aim is to reduce the correlation within the groups.

Points (4,8) and (5,6) are the closest, so they are combined to form cluster 2. The centroid for cluster 2 is (4.5, 7) (Fig. 3.33).

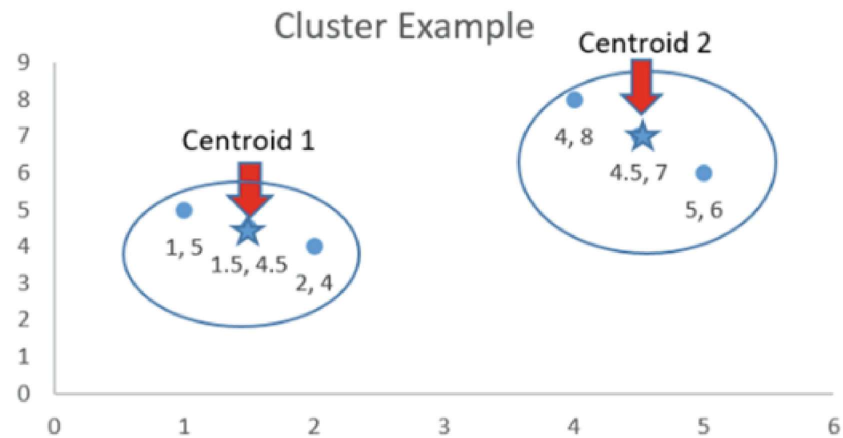


Fig. 3.33 Scatter plot with two clusters

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

Figure 3.33 shows two clusters as a question of when to stop the combining clusters. This might be a reason to pre-define the number of clusters or set the max distance between the group and the points.

The Variable Cluster node in SAS Enterprise Miner can create different clusters and select the representative variables from the cluster. If the dataset has more than 30 variables, the Train group should be set to Yes on the Keep Hierarchies, and Two Stage Clustering should be set to Yes. This method for identifying the variables passed to the subsequent node will filter the best variables in each cluster with the min r-square ratio value. Two-stage variable clustering should be used if the dataset has more than 100 variables and more than 100,00 observations.

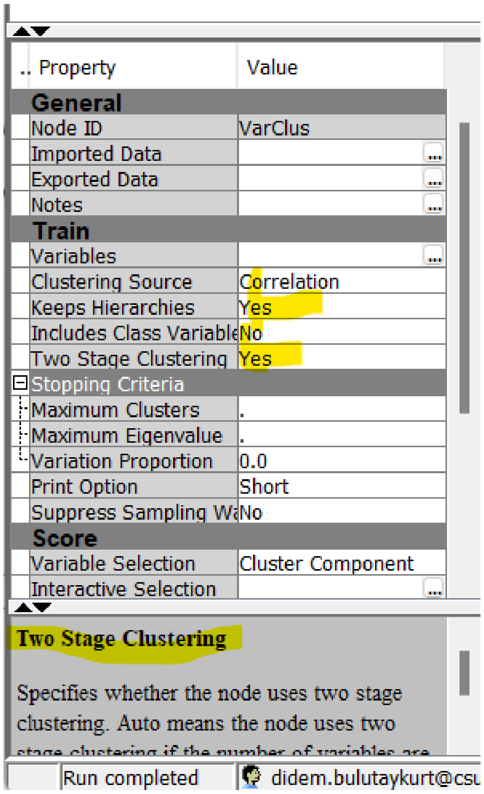


Figure 11: Variable Cluster node properties for more than 30 variables.

The Cluster Plot results from the Variable Cluster node show the objects' hierarchical relationship. The diagram read left to right; a long line means a more significant difference. The relationship between the LOG_INCOME and CLAIM_AMOUNT, MONTHLY_PREMIUM, and MONTHLY_SINCE_POLICY_INCEPT, and ANNUAL_PREMIUM are most similar and are first joined together. IMP_OUTSTANDING_BALANCE and MOUNTHLY_SINCE_LAST_CLAIM are connected to the cluster, meaning the variables more like each other than any variable or cluster joins at a significant level.

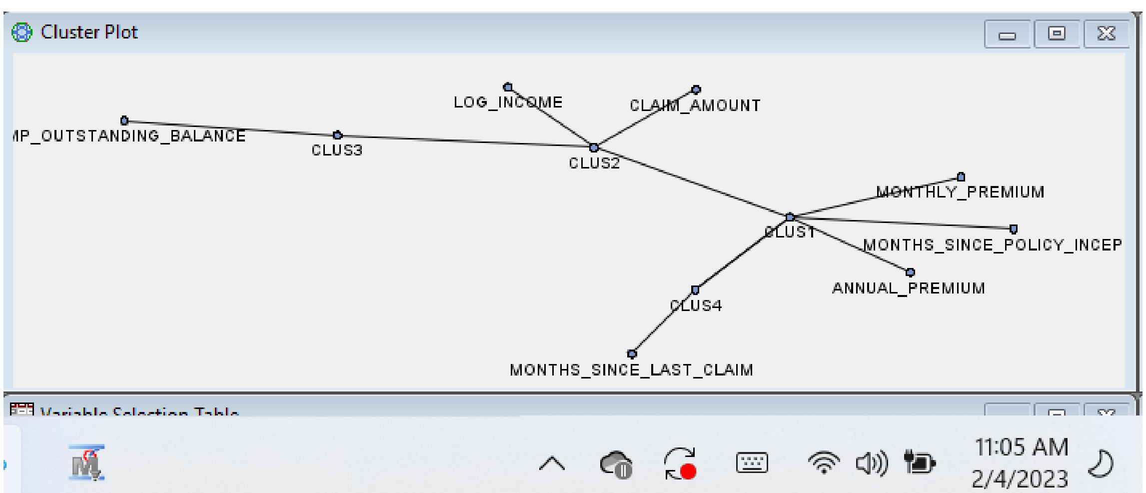


Figure 12: Claim Fraud cluster plot

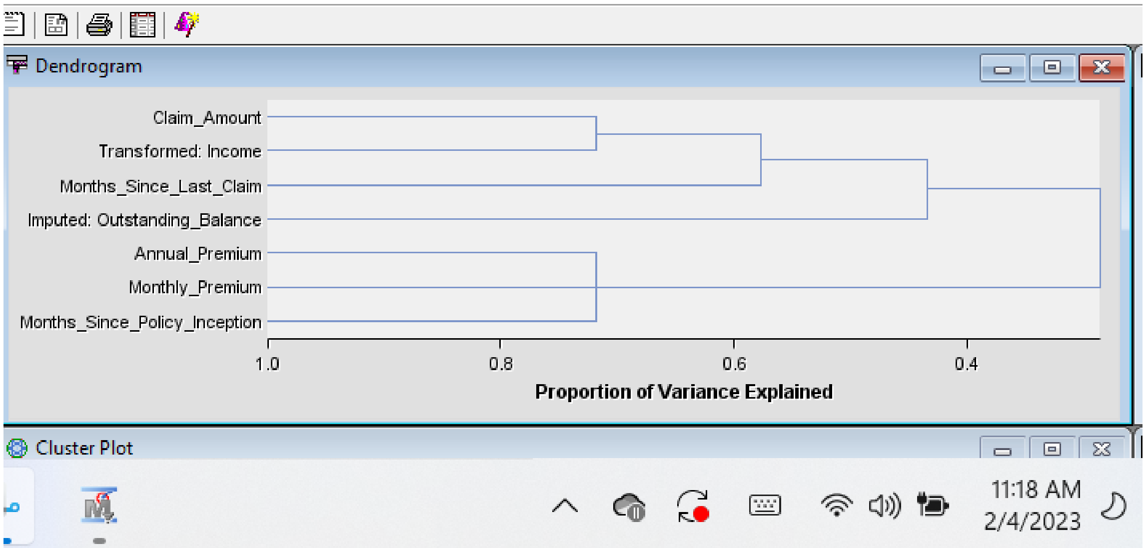


Figure 13: Claim Fraud dataset dendrogram

Principal Component Analysis

This is another variable reduction strategy. It is used when several redundant variables or variables correlate with one another and may measure the same construct. Principal component analysis mathematically manipulates the input variables and develops fewer artificial variables.

Drag and drop The Principal Components node to the diagram work area on the Modify tab. The node result shows a scatter plot and Eigenvalue to see the relation. "The plot shows the Eigenvalues on the y-axis and the number of principal components on the x-axis. It will always be a downward curve; the point where the slope of the curve flattens indicates the number of principal components needed. "(McCarthy, 2022)

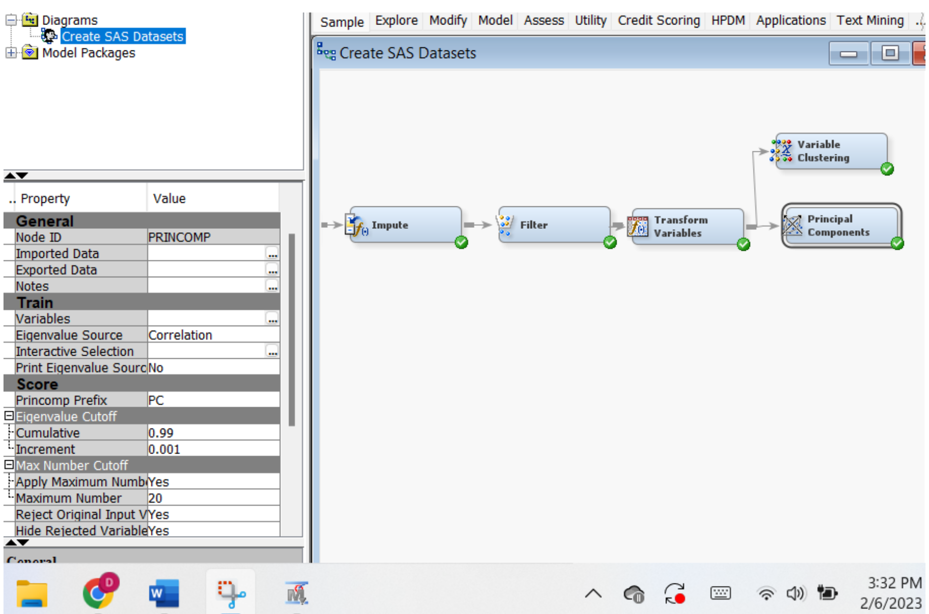


Figure 14: Principal Component node properties

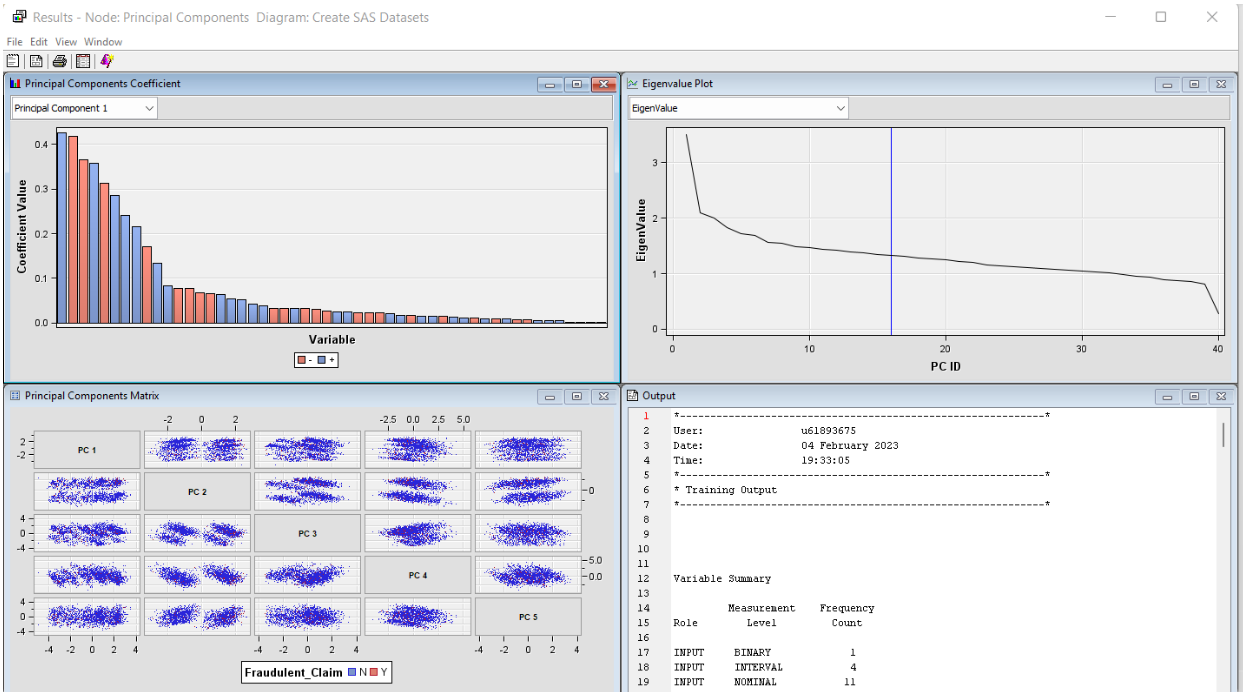


Figure 15: Principal Component node result.

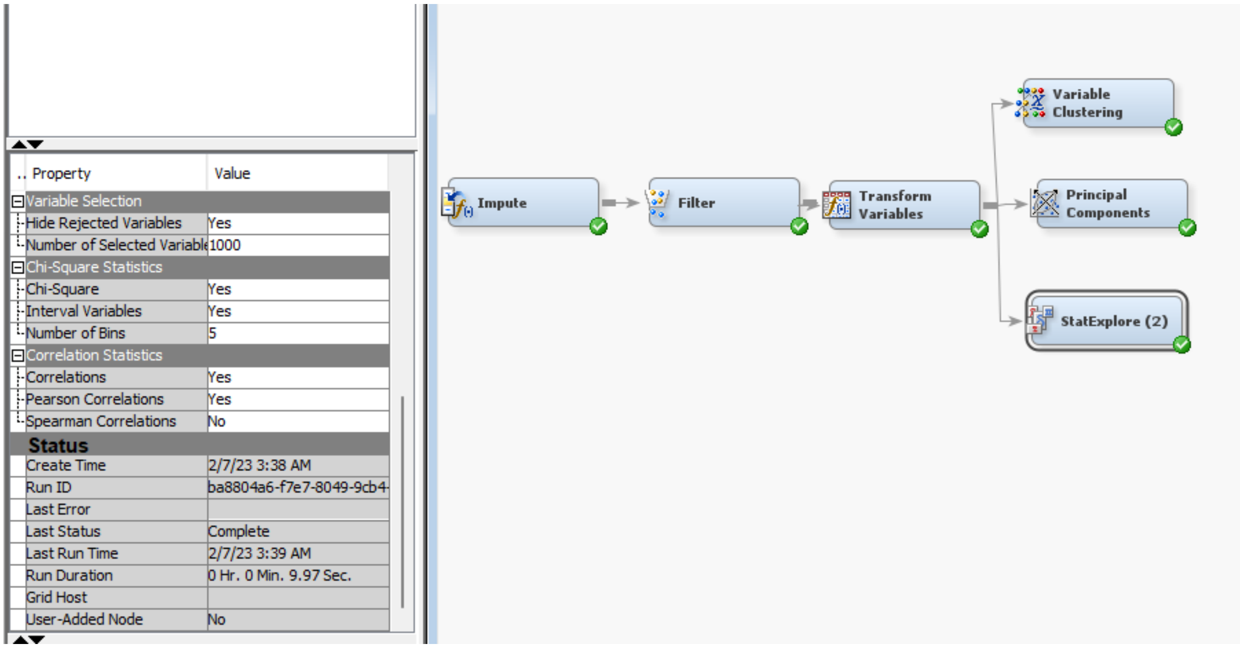
Hypothesis Testing and Chi-Square

The hypothesis testing helps to create business questions with null hypothesis Ho should be tested, and alternative hypothesis AH_A is opposite of the null hypothesis. There are two types of errors; a type I error refers to alpha as the significant level. If the p-value is smaller or equal to alpha 0.05, reject the null hypothesis and accept the alternative hypothesis. Do not reject the null hypothesis if the p-value is higher or equal to a significant level.

Ho: Gender and claim fraud are independent.

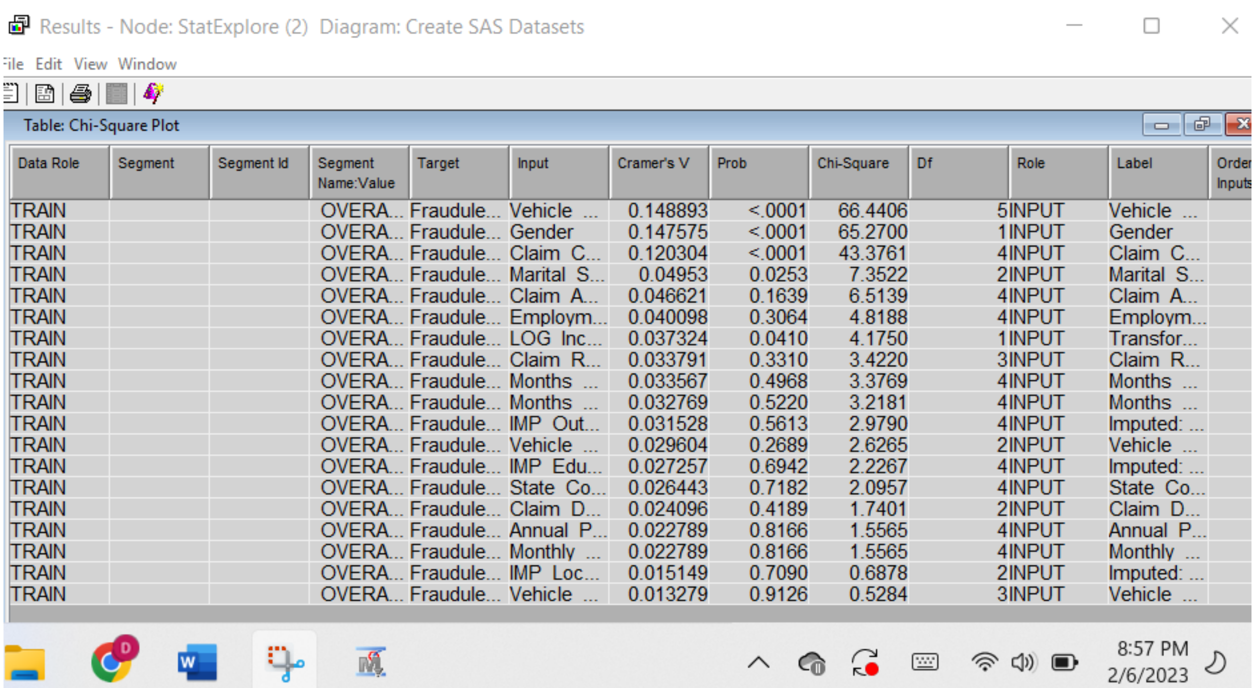
H_A: Gender and claim fraud are not independent.

The Chi-Square test determines if there is a significant relation between two categorical variables. In the SAS Enterprise Miner, drag and drop StatExplorer in the Model tab on diagrams workplace and set the Interval Variables in the Chi-Square Statistic group Yes. Figure 17 shows the output of the gender p-value at .000, lower than 0.05, which means rejecting the null hypothesis.



Property	Value
Variable Selection	
Hide Rejected Variables	Yes
Number of Selected Variabl	1000
Chi-Square Statistics	
Chi-Square	Yes
Interval Variables	Yes
Number of Bins	5
Correlation Statistics	
Correlations	Yes
Pearson Correlations	Yes
Spearman Correlations	No
Status	
Create Time	2/7/23 3:38 AM
Run ID	ba8804a6-f7e7-8049-9cb4-
Last Error	
Last Status	Complete
Last Run Time	2/7/23 3:39 AM
Run Duration	0 Hr. 0 Min. 9.97 Sec.
Grid Host	
User-Added Node	No

Figure 16: StatExplore node for Chi-Square test in Explore tab.



Data Role	Segment	Segment Id	Segment Name:Value	Target	Input	Cramer's V	Prob	Chi-Square	Df	Role	Label	Order Inputs
TRAIN			OVERA...	Fraudule...	Vehicle ...	0.148893	<.0001	66.4406	5	INPUT	Vehicle ...	
TRAIN			OVERA...	Fraudule...	Gender	0.147575	<.0001	65.2700	1	INPUT	Gender	
TRAIN			OVERA...	Fraudule...	Claim C...	0.120304	<.0001	43.3761	4	INPUT	Claim C...	
TRAIN			OVERA...	Fraudule...	Marital S...	0.04953	0.0253	7.3522	2	INPUT	Marital S...	
TRAIN			OVERA...	Fraudule...	Claim A...	0.046621	0.1639	6.5139	4	INPUT	Claim A...	
TRAIN			OVERA...	Fraudule...	Employm...	0.040098	0.3064	4.8188	4	INPUT	Employm...	
TRAIN			OVERA...	Fraudule...	LOG Inc...	0.037324	0.0410	4.1750	1	INPUT	Transfor...	
TRAIN			OVERA...	Fraudule...	Claim R...	0.033791	0.3310	3.4220	3	INPUT	Claim R...	
TRAIN			OVERA...	Fraudule...	Months ...	0.033567	0.4968	3.3769	4	INPUT	Months ...	
TRAIN			OVERA...	Fraudule...	Months ...	0.032769	0.5220	3.2181	4	INPUT	Months ...	
TRAIN			OVERA...	Fraudule...	IMP Out...	0.031528	0.5613	2.9790	4	INPUT	Imputed: ...	
TRAIN			OVERA...	Fraudule...	Vehicle ...	0.029604	0.2689	2.6265	2	INPUT	Vehicle ...	
TRAIN			OVERA...	Fraudule...	IMP Edu...	0.027257	0.6942	2.2267	4	INPUT	Imputed: ...	
TRAIN			OVERA...	Fraudule...	State Co...	0.026443	0.7182	2.0957	4	INPUT	State Co...	
TRAIN			OVERA...	Fraudule...	Claim D...	0.024096	0.4189	1.7401	2	INPUT	Claim D...	
TRAIN			OVERA...	Fraudule...	Annual P...	0.022789	0.8166	1.5565	4	INPUT	Annual P...	
TRAIN			OVERA...	Fraudule...	Monthly ...	0.022789	0.8166	1.5565	4	INPUT	Monthly ...	
TRAIN			OVERA...	Fraudule...	IMP Loc...	0.015149	0.7090	0.6878	2	INPUT	Imputed: ...	
TRAIN			OVERA...	Fraudule...	Vehicle ...	0.013279	0.9126	0.5284	3	INPUT	Vehicle ...	

Figure 17: Chi-Square Plot window output.

Reference

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

Farhad Malik, 2019. Is there A Statistical Method To Test A Claim? https://medium.com/fintechexplained/is-there-a-statistical-method-to-test-a-claim-8d847adabd81