MODUL 12
SAS® VDMML Exploratory Data Analysis (EDA)

THEME DESCRIPTION

Students well understand the benefits and functionality of SAS Visual Data Mining and Machine Learning in the SAS Visual Analytics environment, and able to describe and explore data analysis.

WEEKLY LEARNING OUTCOMES (SUB-LESSONS)

CLO-2-Sub-CLO-12:
Students better understand SAS Cloud Analytic Services (CAS) and can implement exploratory data analysis (EDA) using SAS Visual Data Mining and Machine Learning. -C2.

Through the following learning steps**:**

1. Data exploration
2. Impute Missing Observations
3. Lab Exercises

PRACTICUM SUPPORT

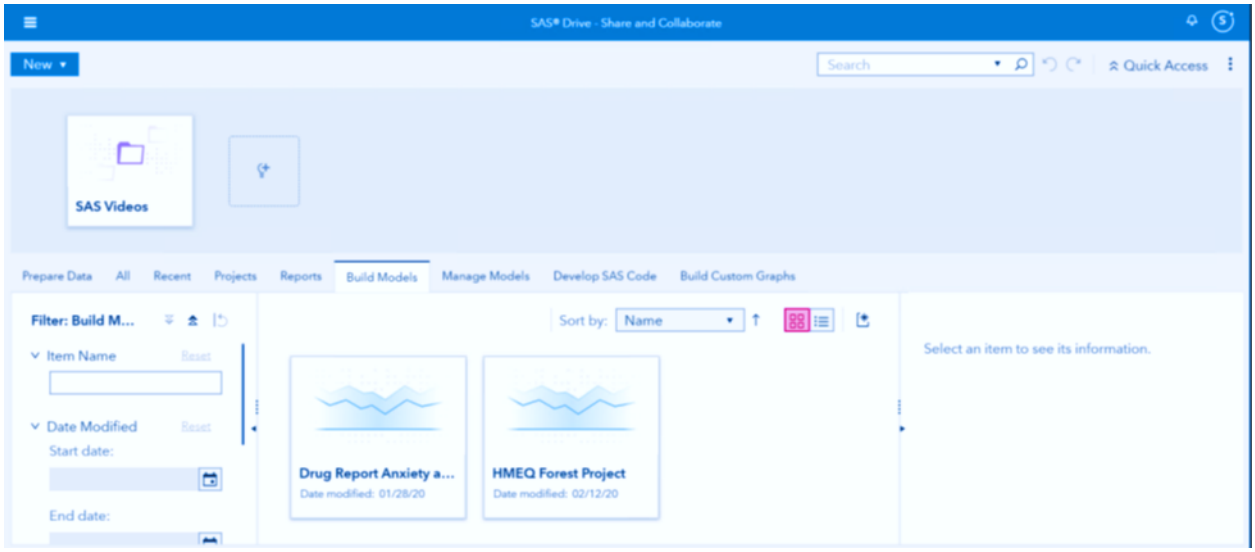
a. Windows Operating System
b. (any) Browser Application

PRACTICUM STEPS

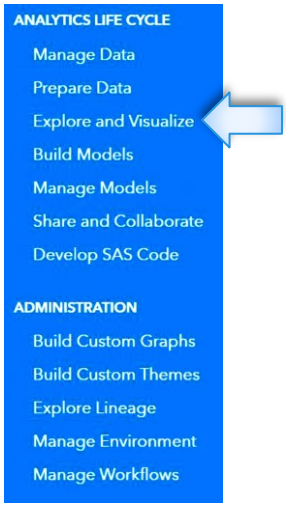
1. Data exploration

- In this demonstration, you access the SAS Visual Data Mining and Machine Learning interface and load the data.

a. Start, https://welcome.oda.sas.com
b. Open SAS®Studio, then open new tab for SAS® Drive to go to SAS Drive page.



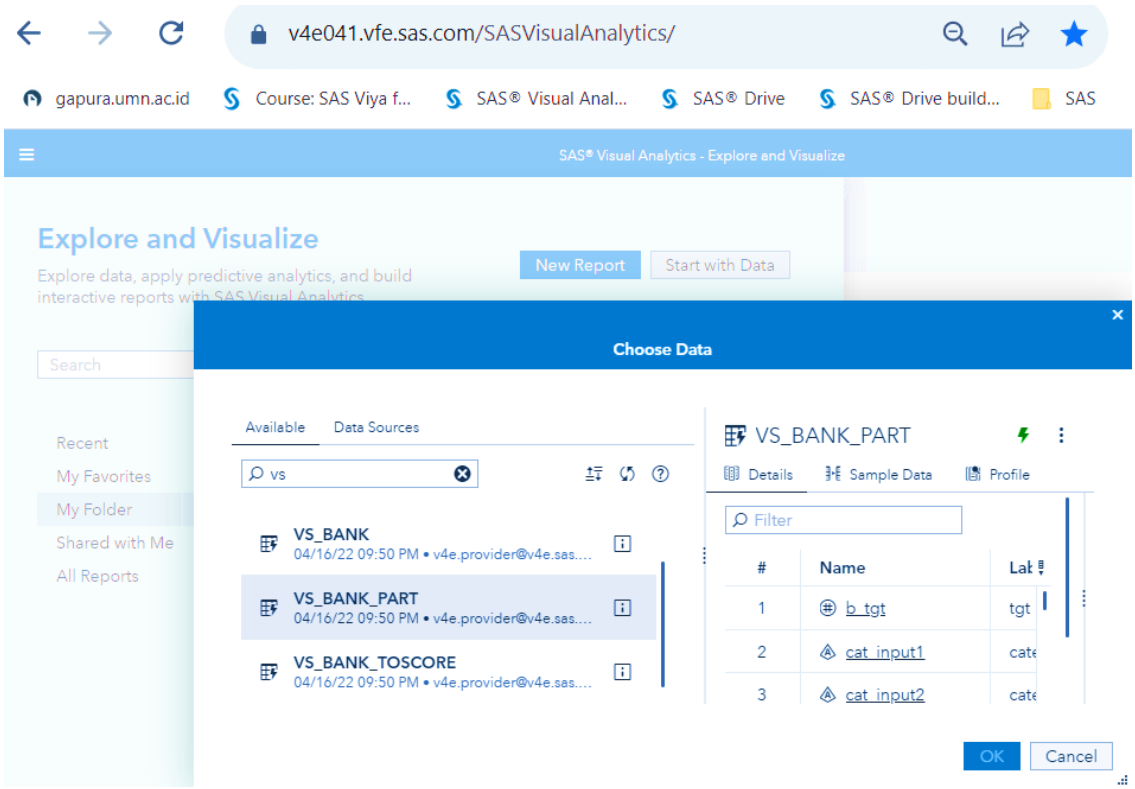
Click (Show list of applications) the upper left corner of the SAS Drive page. **Select Explore and Visualize**.



Import SAS data sets and load them to memory on a CAS server.

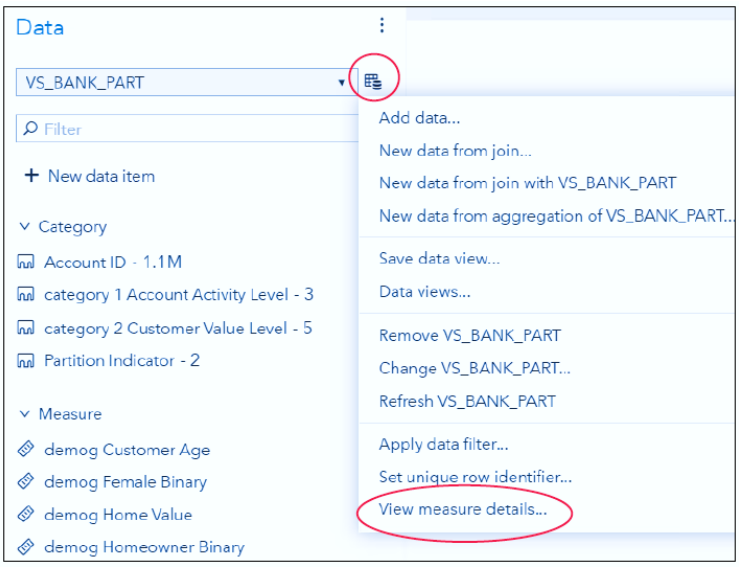
c. Click Start with Data.

d. In the Choose Data window, choose **vs_bank_part** table.

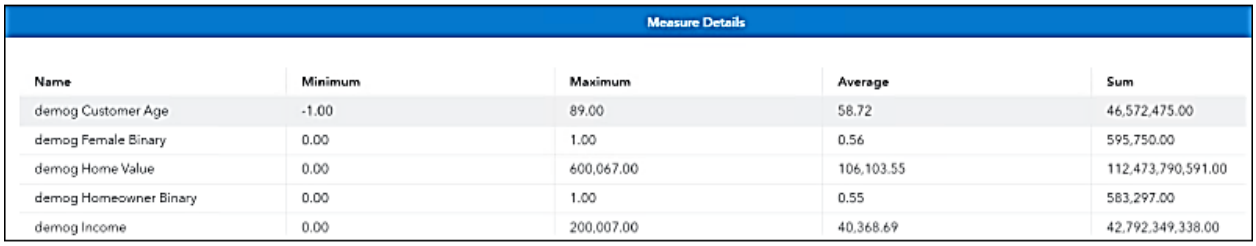


- The Data pane enables you to work with data sources, create new data items (hierarchy, calculated item, custom category, interaction effect, spline effect, partition), add a data source filter, and view and modify properties for data items.
- In this demonstration, you perform initial data exploration using Visual Analytics capabilities.
- Begin the data exploration by examining the descriptive statistics of the numeric data items.

e. In the Data pane, click (Actions) and select View measure details.



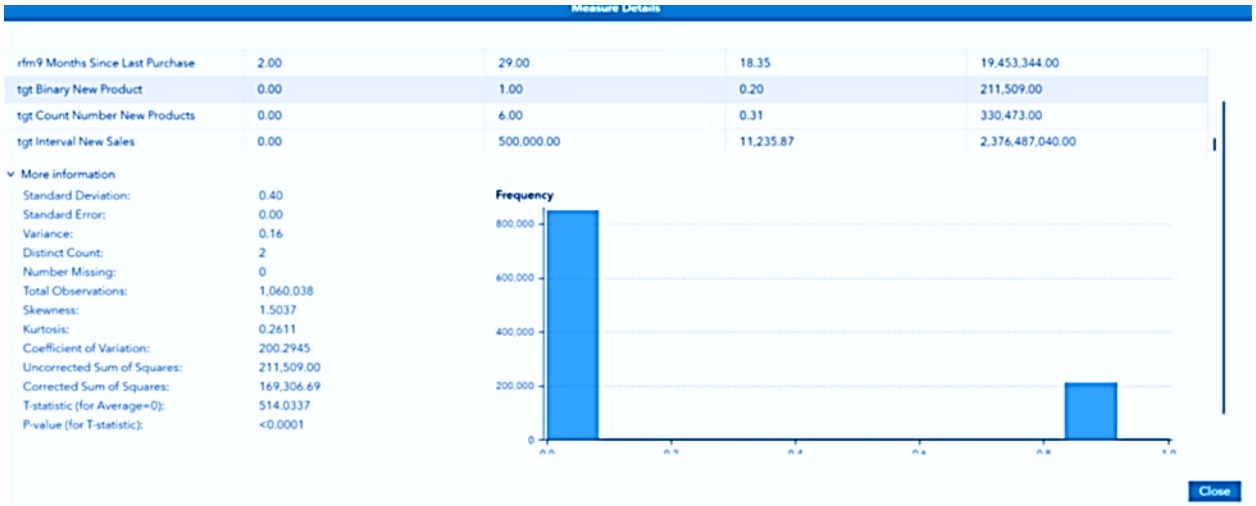
The resulting Measure Details window displays basic statistics (minimum, maximum, average, sum) for all the measures.



Measure Details

Name	Minimum	Maximum	Average	Sum
demog Customer Age	-1.00	89.00	58.72	46,572,475.00
demog Female Binary	0.00	1.00	0.56	595,750.00
demog Home Value	0.00	600,067.00	106,103.55	112,473,790,591.00
demog Homeowner Binary	0.00	1.00	0.55	583,297.00
demog Income	0.00	200,007.00	40,368.69	42,792,349,338.00

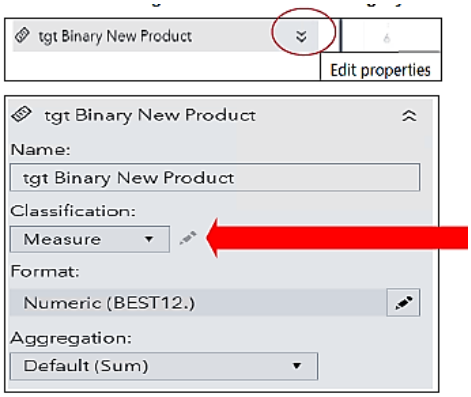
f. Select a measure in the above table to get additional statistics for that selected measure along with the plot of its distribution.



The variable tgt Binary New Product (b_tgt) is the primary dependent variable for categorical response modeling in this course. It is a binary flag that codes responders with 1 and non-responders with 0. Because it is numeric, it is treated as interval valued by default. A closer examination of the Measure Details table indicates that several observations have missing entries for some variables (demog Customer Age, RFM3). We impute the mean for some of these variables in the next few steps. Also, we can observe that both RFM1 (Average Sales Past Three Years) and RFM4 (Last Product Purchase Amount) have negative values. Customers cannot have negative amounts of past sales and product purchase. We are going to transform these values as well.

g. Select Close to exit the Measure Details window.

h. The target variable of interest (tgt Binary New Product) needs to be changed to categorical variable for subsequent modeling demonstrations. **Select Edit properties for tgt Binary New Product and change its classification to Category**.



i. Drag the tgt Binary New Product variable to the canvas to create the auto chart. A bar chart is created by default as shown on Output figure A.

j. **Screenshotted the bar chart as "IS-429 Lab Week#12A Data Explore NIMNAME"**

k. **Named your workspace tab as "Data Exploration"**

l. **Save your SAS VA (Visual Analytics) as "IS-429 Lab Week#12A EDA NIM-yourName".**

- The bar chart shows that approximately 200,000 customers have tgt Binary New Product=1.
- That is, of the total of slightly more than 1 million customers, approximately 20% made a purchase.
- Now we deal with variables having missing observations. While examining the Measure Details table, you noticed that several variables had missing observations.

2. Impute missing observations in demog Customer Age

- Let us focus on the demog Customer Age variable, which has around 266861 missing observations. The issue of missing values in data is (nearly) always present and always a concern.
- The typical approach to solving the missing data issue is imputation. Imputation is the act of plugging in a "reasonable" value for a missing obersvations.

a. On the Data pane, select New data item --> Calculated item.

b. In the Name field, enter Imp_demog_Customer_Age.

c. For the Result Type field, verify that Automatic (Numeric) is selected

d. For the Format field, click (Edit).

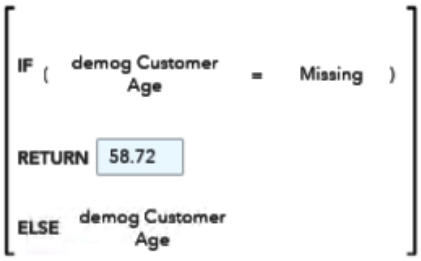
1) In the Format window, select Numeric.

2) For the Width field, verify that 12 is specified.

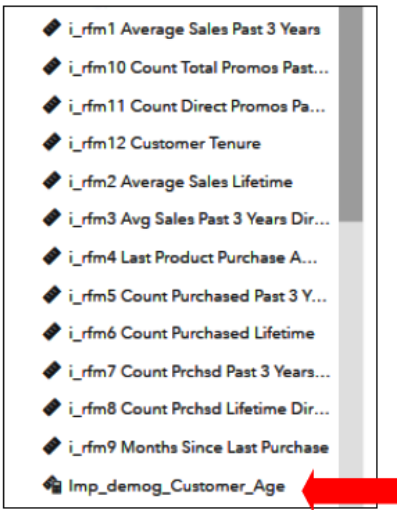
3) Click OK.

e. On the left side of the window, click Operators

f. On the left side of the window, click Operators.
g. Expand the Boolean group.
h. Double-click the IF...ELSE operator to add it to the expression.
i. Expand the Comparison group.
j. Drag x=y to the condition field in the expression. (Alternatively, you can right-click the condition box and select Use Inside --> x=y.)
k. On the left side of the window, click Data Items.
l. Expand the Numeric group.
m. Drag Demog Customer Age to the number field on the left of the equal sign.
n. Enter. (missing) in the number field on the right of the equal sign.
o. Enter 58.72 on the number field for the RETURN operator.
Note: The average for demog Customer Age is 58.72 as listed in the Measure Details table above.
p. Drag demog Customer Age to the number field for the ELSE operator.
q. The expression should resemble the following:



r. In the lower right corner of the window, click OK. The new data item is added to the Data pane.



- Similarly, other variables with missing observations can be imputed with any reasonable value supplied by analyst. For brevity, missing value imputation of other variables is not discussed.
- Recall that RFM1 (Average Sales Past Three Years) and RFM4 (Last Product Purchase Amount) have negative values. It is unlikely that the past sales and product purchase values would be negative. One of the approaches for addressing such variables is to first recode negative values to missing and then replace these missing values with the mean or any other average. The steps that follow show how this can be achieved using Visual Analytics functionality for the RFM1 variable.

Recode negative values to missing

s. Again, on the Data pane, select New data item Calculated item.
t. In the Name field, enter Recoded_rfm1.
u. For the Result Type field, verify that Automatic (Numeric) is selected.

v. For the Format field, click (Edit).
   1) In the Format window, select Numeric.
   2) For the Width field, verify that 12 is specified.
   3) Click OK.

w. On the left side of the window, click Operators.
x. Expand the Boolean group.
y. Double-click the IF...ELSE operator to add it to the expression.
z. Expand the Comparison group
aa. Drag x < y to the condition field in the expression. (Alternatively, you can right-click the condition box and select Use Inside --> x < y.)
bb. On the left side of the window, click Data Items.
cc. Expand the Numeric group.
dd. Drag rfm1 Average Sales Past 3 years to the number field on the left of the inequality sign.
ee. Enter 0 in the number field on the right of the inequality sign.
ff. Enter. (missing) in the number field for the RETURN operator.
gg. Drag rfm1 Average Sales Past 3 years to the number field for the ELSE operator.
The expression should resemble the following:

hh. In the lower right corner of the window, click OK.

- The new data item Recoded_rfm1 is added to the Data pane.
- Now the missing values created in the Recoded_rfm1 variable will be imputed using the mean of this variable. (You can again use the measure details property to see the descriptive statistics of all the variables, including this newly created data item.) The mean of Recoded_rfm1 variable is found to be 16.09, and the same value will be used for imputation.

Impute missing observations for the Recoded_rfm1 variable

ii. Repeat **steps 2.a through 2.q** to calculate a new data item, Imputed_rfm1, by replacing missing observations of the Recoded_rfm1 variable with its mean, 16.09. The final expression window should resemble the following:

In the lower right corner of the window, click OK.
The new data item, Imputed_rfm1, is added to the Data pane.

jj. In the upper left corner of the report, click (New page) next to Page 2 and **rename it into "Impute Missing Obs".**

kk. In the left pane, click the Data icon. Drag Imputed_rfm1 to the canvas.
The results are skewed to the right. For some models, it is useful to use a regularizing transformation for input variables. The log transformation regularizes this distribution.

Create a transformed variable.

ll. On the Data pane, select New data item Calculated item.

mm. In the Name f ield, enter Transformed_rfm1.

nn. For the Result Type f ield, verify that Automatic (Numeric) is selected.

oo. For the Format field, click (Edit).
1) In the Format window, select Numeric.
2) For the Width field, verify that 12 is specified.
3) Click OK.

pp. On the left side of the window, click Operators.

qq. Expand the Numeric (advanced) group.

rr. Double-click the Log operator to add it to the expression.

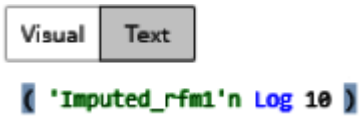
ss. On the left side of the window, click Data Items

tt. Expand the Numeric group.

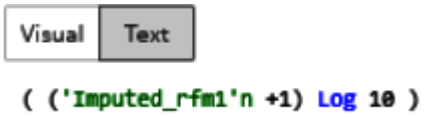
uu. Drag Imputed_rfm1 to the number field to the left of Log.

vv. Enter 10 in the second number box.
To avoid calculating the log of any zero values in the Imputed_rfm1 column, add 1 in the formula.

ww. Click the Text tab. The expression should resemble the following:



xx. Enter +1 after 'Imputed_rfm1'n and insert parentheses such that the expression resembles the following:



Alternatively, on the Visual tab, you can drag the x+y operator to the variable name to add 1 to it.

yy. Click Preview in the lower right corner of the window to preview the result to determine whether the formula has taken effect. Close the preview result window.

zz. In the lower right corner of the window, click OK.

aaa. The new data item is added to the Data pane.

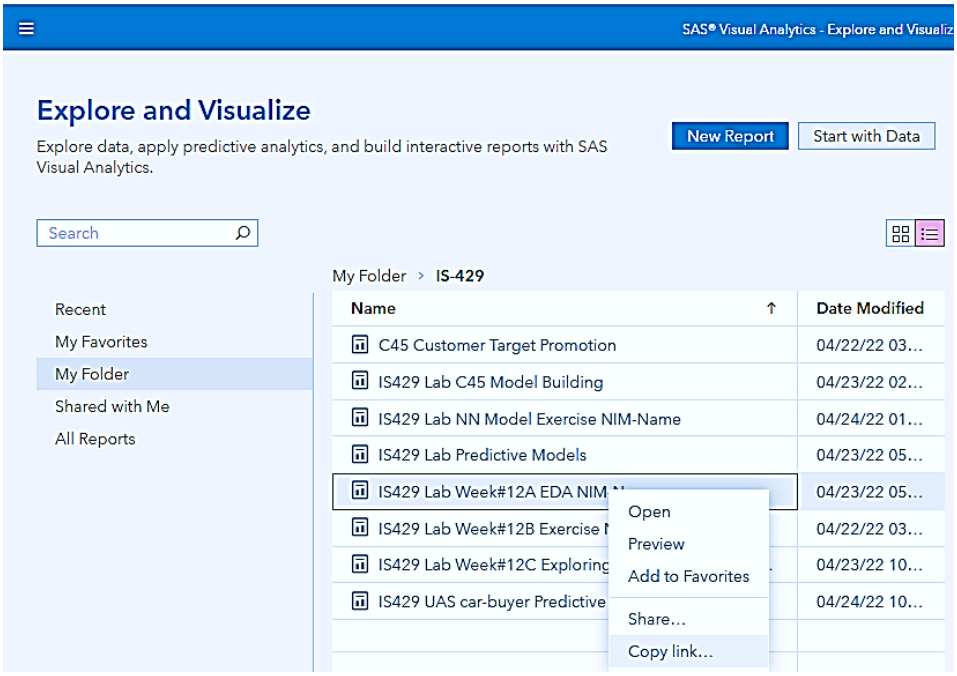
bbb. Click Page **"Impute Missing Obs"** and drag transformed_rfm1 to the right of the existing histogram on the canvas to create a new auto chart. Compare the histograms as shown on Output figure B.

ccc. **Screenshotted the histogram comparison as "IS-429 Lab Week#12B Impute NIM-yourName.jpg"**.

ddd. **Collect all the screenshots you produce into your word format file "IS-429 Lab Week#12A EDA NIM-yourName.doc/docx**

eee. **Save your VA report. Click (Menu) and select Save**.

fff. Share your VA report by "copy link" :



ggg. **Type it on your copy link into first row of your word file "IS-429 Lab Week#12A EDA NIM-yourName.doc/docx"**.

3. Lab Exercises

- You work with the PVA_PARTITION data set for practices. It contains data that represent charitable donations made to a veterans' organization.
- The data represent the results of a mail campaign to solicit donations. Solicitations involve sending a small gift to an individual and include a request for a donation.
- The data set contains the following information:
  - a flag to indicate respondents to the appeal (Target Gift Flag) and the dollar amount of their donations (Target Gift Amount)
  - respondents' PVA promotion and giving history
  - demographic data of the respondents
- In the first practice, you use SAS Visual Analytics to familiarize yourself with the data.

a. Sign in to SAS Visual Analytics. Enter your User ID field and Metadatao in the Password field.

b. Select Explore and Visualize to begin accessing and exploring the data.

c. Select the PVA_PARTITION data source.

d. Select the Data pane on the left of the canvas (if it is not open) and answer below questions:

1) Which level of the Status Category 96NK variable has the highest count? _____

2) Does the variable Age contain any missing values? If so how many? _____

- Giving name your report tab as "Data Source".
- **Screenshotted your data source figure as Output figure C into "IS-429 Lab Week#12C Exercise-C NIM-yourName.jpg.**

e. Change Target Gift Flag from a measure to a category. It is a binary indicator that represents a response to a mailing, where 1 indicates that customers did respond.

f. Answer these below questions:

1) How are responders and non-responders distributed in the data?_____

2) How many females responded to the campaign?_____

g. Save the report. Click (Menu) and select Save As. Save the report in My Folder --> My Tasks with the name "**IS-429 Lab Week#12D Exercises NIM-yourname"**. Click Save.

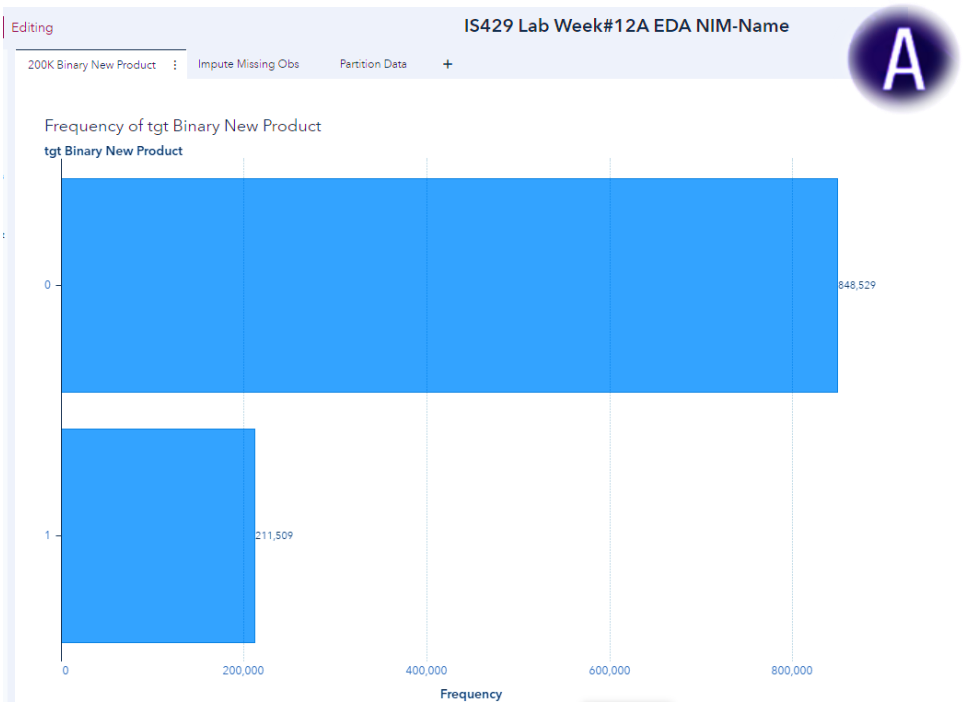
h. Collect all the screenshots you produce into your word format file "IS-429 Lab Week#12B Exercises NIM-yourName.doc/docx

i. **Share your VA report by "copy link" and type it on your copy link into firsat row of your word file "IS-429 Lab Week#12B Exercises NIM-yourName.doc/docx**".

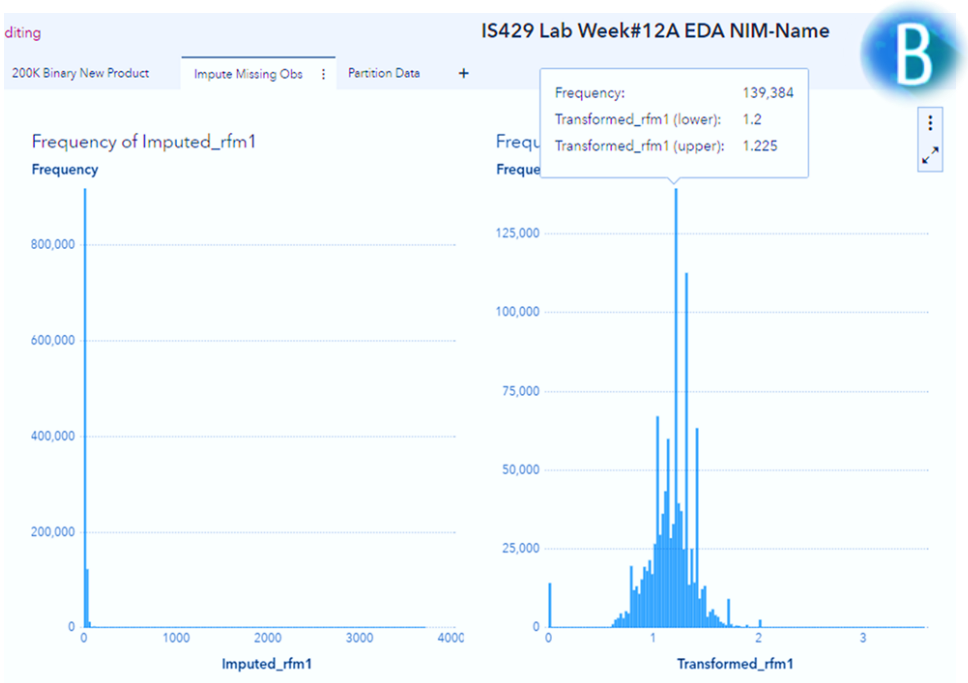
➲ Finally, today's practicum is over, **collect all your word files into format file "IS-429 Lab Week#12 EDA NIM.zip"** and **submit immediately today to e-Learning IS-429 Practicum Week#12.**

RESULTS/ OUTPUT

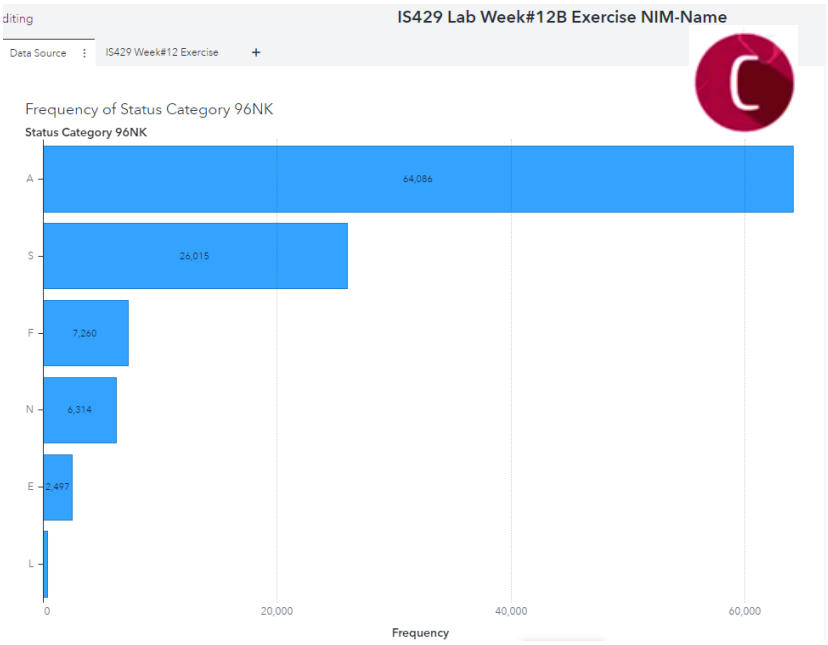
1. Data exploration

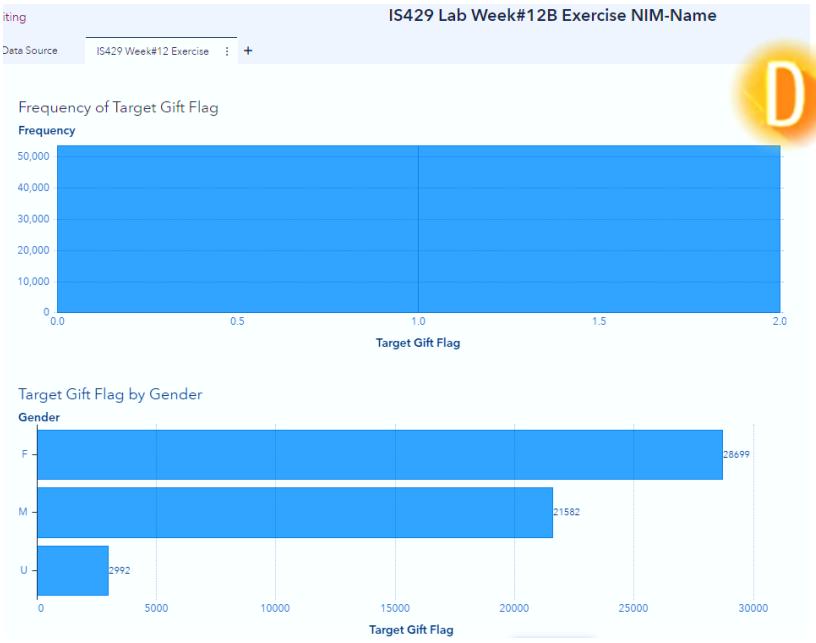


2. Impute Missing Observations



3. Lab Exercises





REFERENCE

1. SAS Institute Inc. 2020. SAS® Visual Data Mining and Machine Learning in SAS® Viya®: Interactive Machine Learning. SAS Institute Inc. Cary, NC, USA.
2. SAS Institute Inc. 2020. SAS® Viya® Programming: Getting Started. SAS Institute Inc. Cary, NC, USA.
3. SAS® Support | Documentation
4. Other additional references are excerpts from various Online Learning/websites.