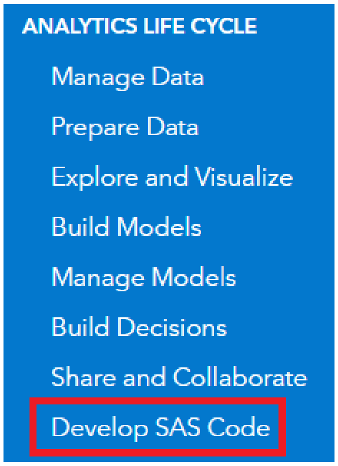
# In Praise of Data Prep: Good Data = Better Models

## Overview

- **Synopsis**
    - There is a classic saying among data scientists: garbage in = garbage out. Put simply, your models are only as good as the data that underlies them. So, from SAS Studio Tasks to Visualizations in SAS Visual Analytics to automated data wrangling in SAS Model Studio, learn how to use SAS Viya to better understand your data before you rush into the modeling process.
        - In this SAS On-the-Job, you'll assume the role of Professional Data Wrangler at iLink Telecom, Inc.  Data for this project feed into a larger effort by the company to identify which customers are most likely to leave the company for another wireless provider (i.e., churn). Better understanding the data is the first – and critical – step in that process.
        - This workshop focuses on data exploration – and preparation – tools within a broader SAS Viya for Learners tour.
    - Since we're doing a deep dive on the data, a detailed data dictionary can be found in the appendix.
- **Tour Overview**
    - Part I: SAS Studio Tasks
    - Part II: SAS Visual Analytics
    - Part III: SAS Model Studio

## Part I: SAS Studio Tasks

- **Objective**
    - Outliers + statistical checks with SAS Studio Tasks
- **SAS Viya for Learners Setup**
    - Access SAS Studio



ANALYTICS LIFE CYCLE

Manage Data

Prepare Data

Explore and Visualize

Build Models

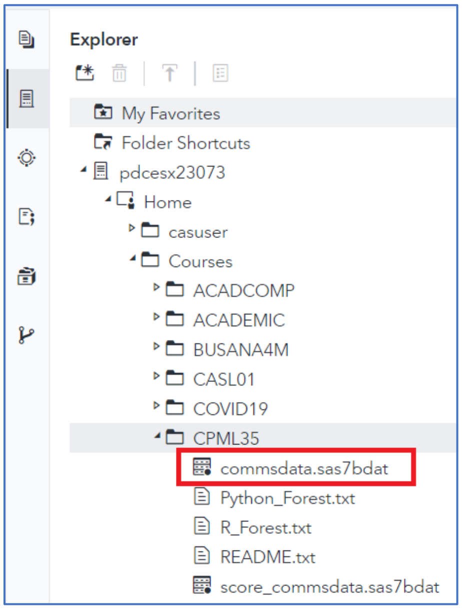
Manage Models

Build Decisions

Share and Collaborate

Develop SAS Code

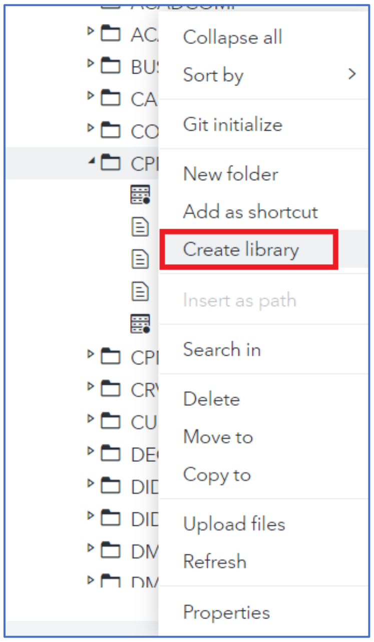
    - Find course data



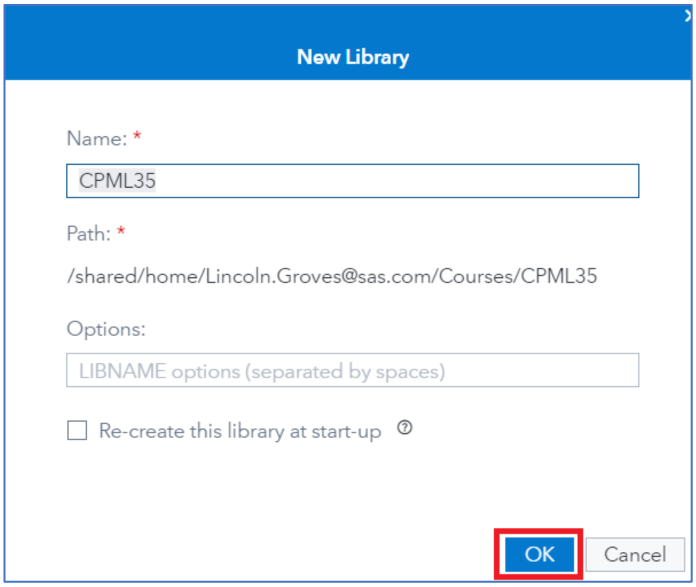
Explorer

My Favorites
Folder Shortcuts
pdcesx23073
Home
casuser
Courses
ACADCOMP
ACADEMIC
BUSANA4M
CASL01
COVID19
CPML35
commsdata.sas7bdat
Python_Forest.txt
R_Forest.txt
README.txt
score_commsdata.sas7bdat

- Create a **SAS Library**
  - **Part I**: Find **Create library** option



  - **Part II**: Accept the **New Library** defaults



New Library

Name: *

CPML35

Path: *

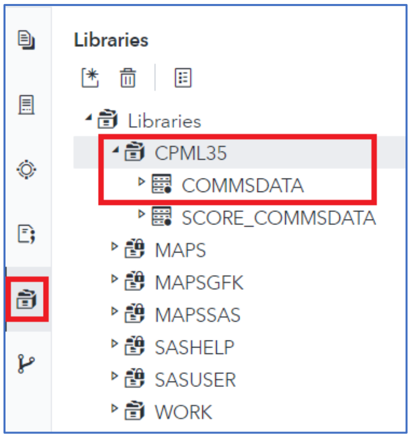
/shared/home/Lincoln.Groves@sas.com/Courses/CPML35

Options:

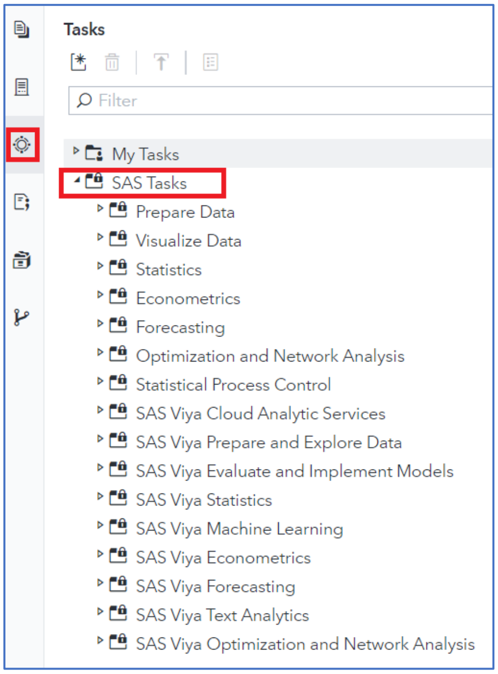
LIBNAME options (separated by spaces)

□ Re-create this library at start-up

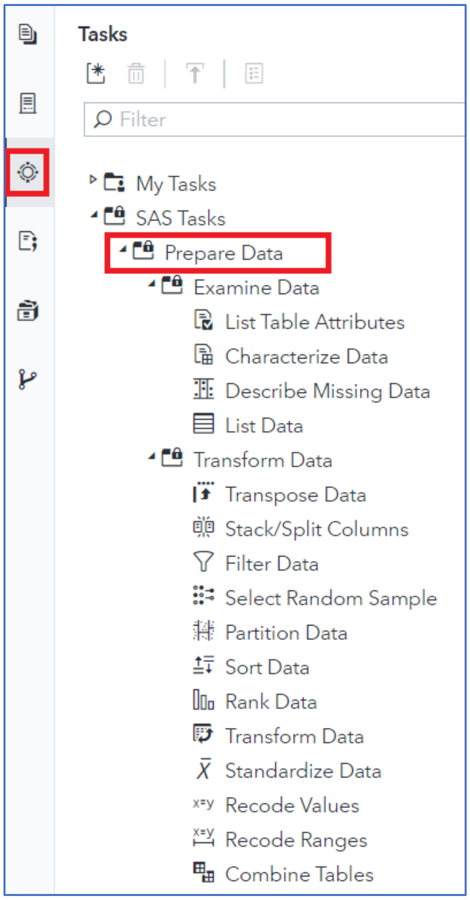
- **Part III**: Check to see if all is good:



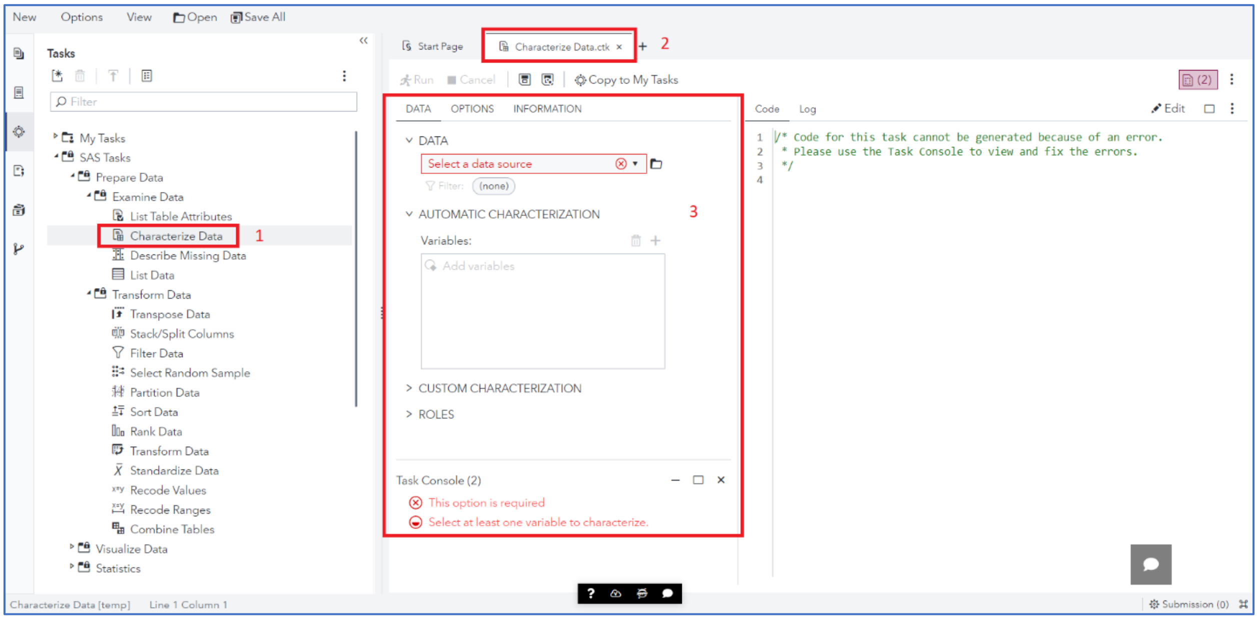
- **Welcome to SAS Studio Tasks!**
  - What is a SAS Studio Tasks? (thank you, ChatGPT)
    - *In SAS Studio, tasks refer to pre-defined, point-and-click operations or workflows that guide users through specific data analysis or data manipulation processes.*
    - *A SAS Studio task provides a visual and user-friendly way to perform various analytical tasks without the need for writing SAS code manually. Each task is designed to address a specific analytical need or process, such as data import, data exploration, statistical analysis, data transformation, and reporting.*
    - *SAS Studio tasks are particularly useful for users who are new to SAS or prefer a graphical interface over writing code. They offer a simplified approach to utilizing SAS functionality and enable users to leverage the power of SAS software without requiring extensive programming knowledge.*
  - Explore a bit:


Tasks
Filter
My Tasks
SAS Tasks
Prepare Data
Visualize Data
Statistics
Econometrics
Forecasting
Optimization and Network Analysis
Statistical Process Control
SAS Viya Cloud Analytic Services
SAS Viya Prepare and Explore Data
SAS Viya Evaluate and Implement Models
SAS Viya Statistics
SAS Viya Machine Learning
SAS Viya Econometrics
SAS Viya Forecasting
SAS Viya Text Analytics
SAS Viya Optimization and Network Analysis

- Locate the **Prepare Data** task:



- Let's get to know our data a bit better. Open the **Characterize Data** tasks:

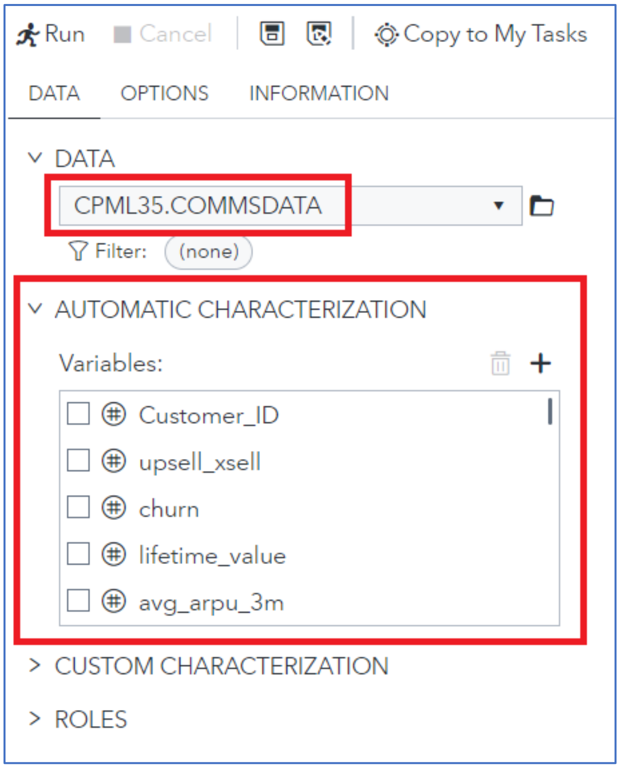


```
/* Code for this task cannot be generated because of an error.
 * Please use the Task Console to view and fix the errors.
 */
```

Task Console (2)

- This option is required
- Select at least one variable to characterize.

- Start with the following settings:



DATA

CPML35.COMMSDATA

Filter: (none)

AUTOMATIC CHARACTERIZATION

Variables:

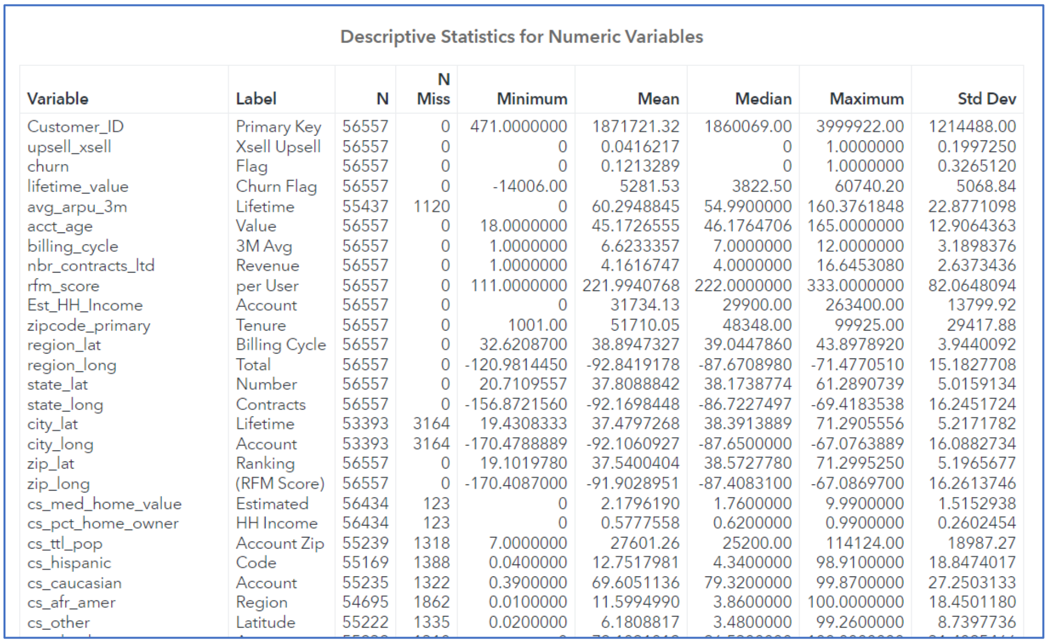
- □ Customer_ID
- □ upsell_xsell
- □ churn
- □ lifetime_value
- □ avg_arpu_3m

CUSTOM CHARACTERIZATION

ROLES

- Submit the code:

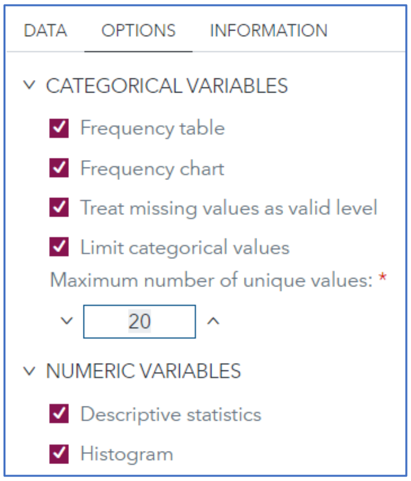
- Examine the output:



Descriptive Statistics for Numeric Variables

| Variable | Label | N | N Miss | Minimum | Mean | Median | Maximum | Std Dev |
|---|---|---|---|---|---|---|---|---|
| Customer_ID | Primary Key | 56557 | 0 | 471.0000000 | 1871721.32 | 1860069.00 | 3999922.00 | 1214488.00 |
| upsell_xsell | Xsell Upsell | 56557 | 0 | 0 | 0.0416217 | 0 | 1.0000000 | 0.1997250 |
| churn | Flag | 56557 | 0 | 0 | 0.1213289 | 0 | 1.0000000 | 0.3265120 |
| lifetime_value | Churn Flag | 56557 | 0 | -14006.00 | 5281.53 | 3822.50 | 60740.20 | 5068.84 |
| avg_arpu_3m | Lifetime | 55437 | 1120 | 0 | 60.2948845 | 54.9900000 | 160.3761848 | 22.8771098 |
| acct_age | Value | 56557 | 0 | 18.0000000 | 45.1726555 | 46.1764706 | 165.0000000 | 12.9064363 |
| billing_cycle | 3M Avg | 56557 | 0 | 1.0000000 | 6.6233357 | 7.0000000 | 12.0000000 | 3.1898376 |
| nbr_contracts_ltd | Revenue | 56557 | 0 | 1.0000000 | 4.1616747 | 4.0000000 | 16.6453080 | 2.6373436 |
| rfm_score | per User | 56557 | 0 | 111.0000000 | 221.9940768 | 222.0000000 | 333.0000000 | 82.0648094 |
| Est_HH_Income | Account | 56557 | 0 | 0 | 31734.13 | 29900.00 | 263400.00 | 13799.92 |
| zipcode_primary | Tenure | 56557 | 0 | 1001.00 | 51710.05 | 48348.00 | 99925.00 | 29417.88 |
| region_lat | Billing Cycle | 56557 | 0 | 32.6208700 | 38.8947327 | 39.0447860 | 43.8978920 | 3.9440092 |
| region_long | Total | 56557 | 0 | -120.9814450 | -92.8419178 | -87.6708980 | -71.4770510 | 15.1827708 |
| state_lat | Number | 56557 | 0 | 20.7109557 | 37.8088842 | 38.1738774 | 61.2890739 | 5.0159134 |
| state_long | Contracts | 56557 | 0 | -156.8721560 | -92.1698448 | -86.7227497 | -69.4183538 | 16.2451724 |
| city_lat | Lifetime | 53393 | 3164 | 19.4308333 | 37.4797268 | 38.3913889 | 71.2905556 | 5.2171782 |
| city_long | Account | 53393 | 3164 | -170.4788889 | -92.1060927 | -87.6500000 | -67.0763889 | 16.0882734 |
| zip_lat | Ranking | 56557 | 0 | 19.1019780 | 37.5400404 | 38.5727780 | 71.2995250 | 5.1965677 |
| zip_long | (RFM Score) | 56557 | 0 | -170.4087000 | -91.9028951 | -87.4083100 | -67.0869700 | 16.2613746 |
| cs_med_home_value | Estimated | 56434 | 123 | 0 | 2.1796190 | 1.7600000 | 9.9900000 | 1.5152938 |
| cs_pct_home_owner | HH Income | 56434 | 123 | 0 | 0.5777558 | 0.6200000 | 0.9900000 | 0.2602454 |
| cs_ttl_pop | Account Zip | 55239 | 1318 | 7.0000000 | 27601.26 | 25200.00 | 114124.00 | 18987.27 |
| cs_hispanic | Code | 55169 | 1388 | 0.0400000 | 12.7517981 | 4.3400000 | 98.9100000 | 18.8474017 |
| cs_caucasian | Account | 55235 | 1322 | 0.3900000 | 69.6051136 | 79.3200000 | 99.8700000 | 27.2503133 |
| cs_afr_amer | Region | 54695 | 1862 | 0.0100000 | 11.5994990 | 3.8600000 | 100.0000000 | 18.4501180 |
| cs_other | Latitude | 55222 | 1335 | 0.0200000 | 6.1808817 | 3.4800000 | 99.2600000 | 8.7397736 |

        - Are missing values an issue?
        - Yes!  Let’s return to that shortly.
    - That’s way too much to process!  Let’s change a few **Options**:



DATA OPTIONS INFORMATION

CATEGORICAL VARIABLES

☑ Frequency table

☑ Frequency chart

☑ Treat missing values as valid level

☑ Limit categorical values

Maximum number of unique values: *

20

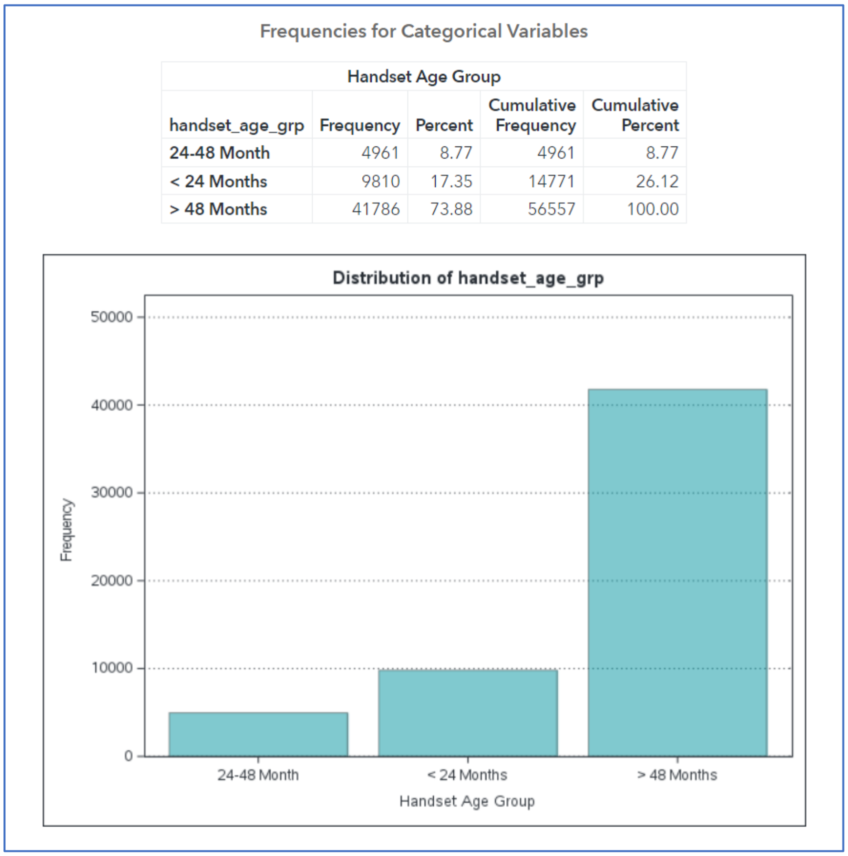
NUMERIC VARIABLES

☑ Descriptive statistics

☑ Histogram

    - Let’s reduce the number of variables:
        - Categorical
            - handset_age_grp
        - Numeric
            - churn
            - lifetime_value
            - ever_days_over_plan
            - ever_times_over_plan
            - equip_age
            - avg_days_susp
            - curr_days_susp
            - times_susp
            - MB_Data_Usg_M04
            - seconds_of_data_norm
        - *Hint: you can use the search bar to help locate the variable*

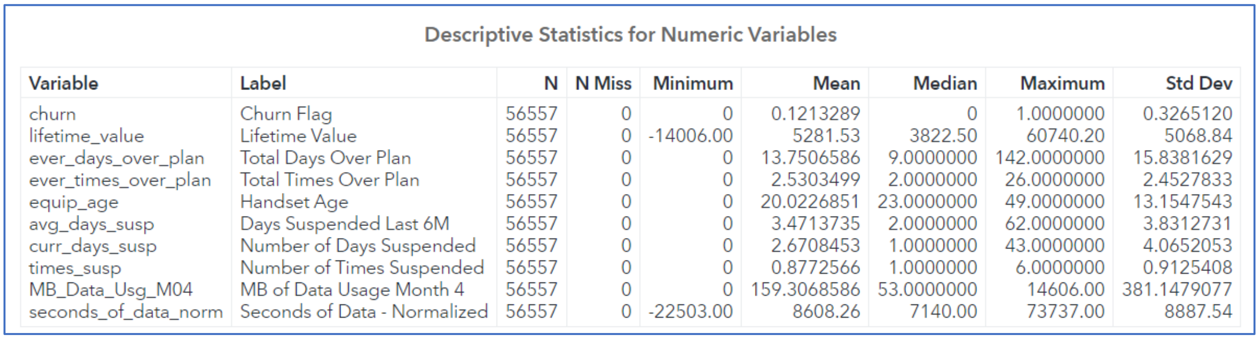
- Resubmit and examine the new output:



Frequencies for Categorical Variables

| Handset Age Group | | | | |
|---|---|---|---|---|
| handset_age_grp | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
| 24-48 Month | 4961 | 8.77 | 4961 | 8.77 |
| < 24 Months | 9810 | 17.35 | 14771 | 26.12 |
| > 48 Months | 41786 | 73.88 | 56557 | 100.00 |

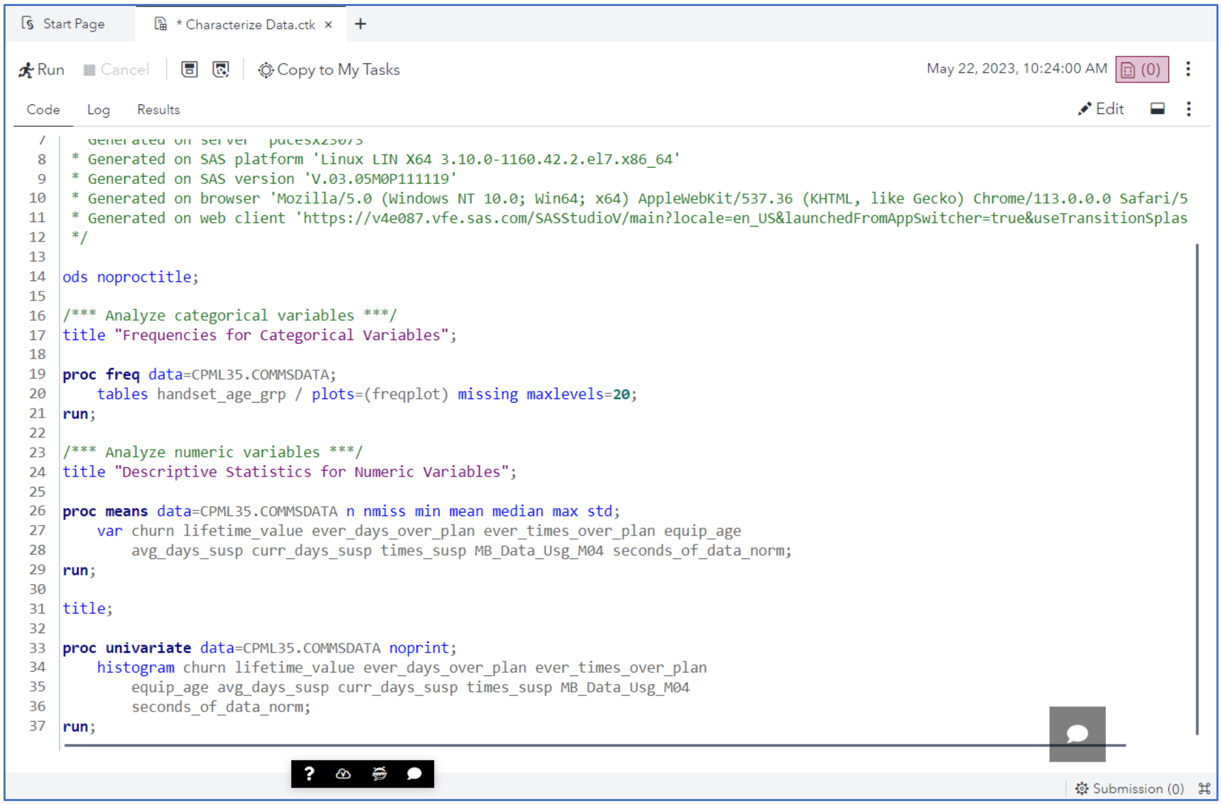
- Examine the descriptive statistics in detail. Does the underlying data make sense?



Descriptive Statistics for Numeric Variables

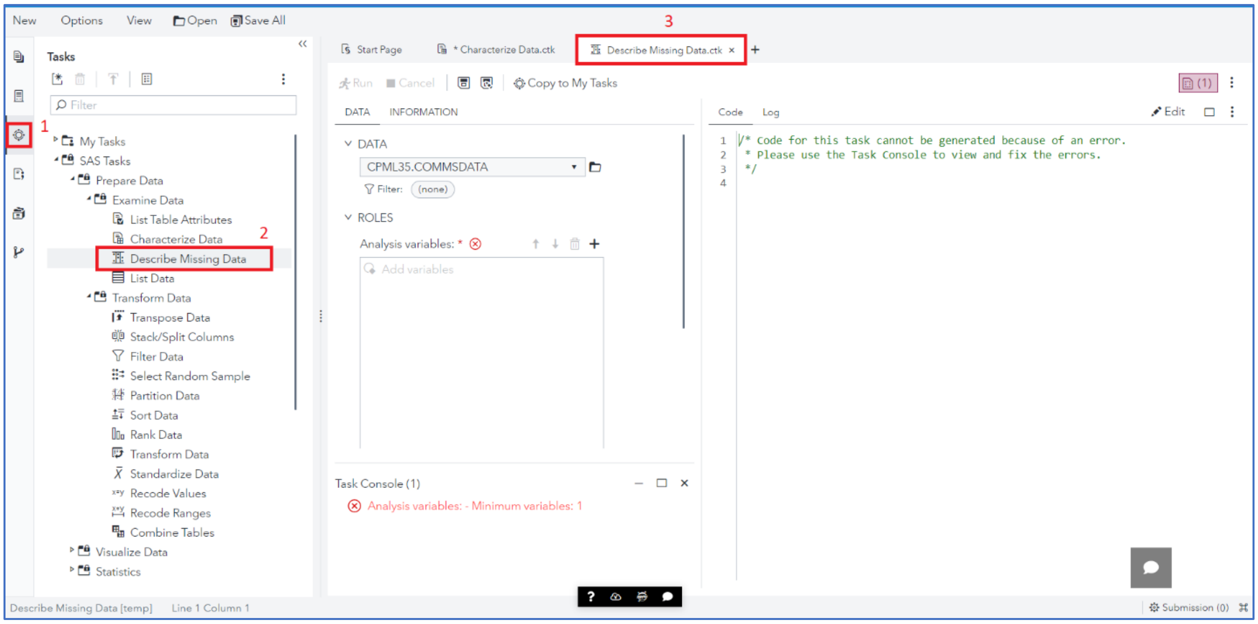
| Variable | Label | N | N Miss | Minimum | Mean | Median | Maximum | Std Dev |
|---|---|---|---|---|---|---|---|---|
| churn | Churn Flag | 56557 | 0 | 0 | 0.1213289 | 0 | 1.0000000 | 0.3265120 |
| lifetime_value | Lifetime Value | 56557 | 0 | -14006.00 | 5281.53 | 3822.50 | 60740.20 | 5068.84 |
| ever_days_over_plan | Total Days Over Plan | 56557 | 0 | 0 | 13.7506586 | 9.0000000 | 142.0000000 | 15.8381629 |
| ever_times_over_plan | Total Times Over Plan | 56557 | 0 | 0 | 2.5303499 | 2.0000000 | 26.0000000 | 2.4527833 |
| equip_age | Handset Age | 56557 | 0 | 0 | 20.0226851 | 23.0000000 | 49.0000000 | 13.1547543 |
| avg_days_susp | Days Suspended Last 6M | 56557 | 0 | 0 | 3.4713735 | 2.0000000 | 62.0000000 | 3.8312731 |
| curr_days_susp | Number of Days Suspended | 56557 | 0 | 0 | 2.6708453 | 1.0000000 | 43.0000000 | 4.0652053 |
| times_susp | Number of Times Suspended | 56557 | 0 | 0 | 0.8772566 | 1.0000000 | 6.0000000 | 0.9125408 |
| MB_Data_Usg_M04 | MB of Data Usage Month 4 | 56557 | 0 | 0 | 159.3068586 | 53.0000000 | 14606.00 | 381.1479077 |
| seconds_of_data_norm | Seconds of Data - Normalized | 56557 | 0 | -22503.00 | 8608.26 | 7140.00 | 73737.00 | 8887.54 |

  - Nope! What on earth are those negative values?
- While in the neighborhood, let’s examine the code:



```
 * Generated on server 'pdcesx23073'
 * Generated on SAS platform 'Linux LIN X64 3.10.0-1160.42.2.el7.x86_64'
 * Generated on SAS version 'V.03.05M0P111119'
 * Generated on browser 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/113.0.0.0 Safari/5
 * Generated on web client 'https://v4e087.vfe.sas.com/SASStudioV/main?locale=en_US&launchedFromAppSwitcher=true&useTransitionSplas
 */

ods noproctitle;

/*** Analyze categorical variables ***/
title "Frequencies for Categorical Variables";

proc freq data=CPML35.COMMSDATA;
    tables handset_age_grp / plots=(freqplot) missing maxlevels=20;
run;

/*** Analyze numeric variables ***/
title "Descriptive Statistics for Numeric Variables";

proc means data=CPML35.COMMSDATA n nmiss min mean median max std;
    var churn lifetime_value ever_days_over_plan ever_times_over_plan equip_age
        avg_days_susp curr_days_susp times_susp MB_Data_Usg_M04 seconds_of_data_norm;
run;

title;

proc univariate data=CPML35.COMMSDATA noprint;
    histogram churn lifetime_value ever_days_over_plan ever_times_over_plan
        equip_age avg_days_susp curr_days_susp times_susp MB_Data_Usg_M04
        seconds_of_data_norm;
run;
```

- **More SAS Studio Tasks**
  - Let's **Describe Missing Data ➔** focus on Census Data
  - Find + open our task:



DATA

CPML35.COMMSDATA

Filter: (none)

ROLES

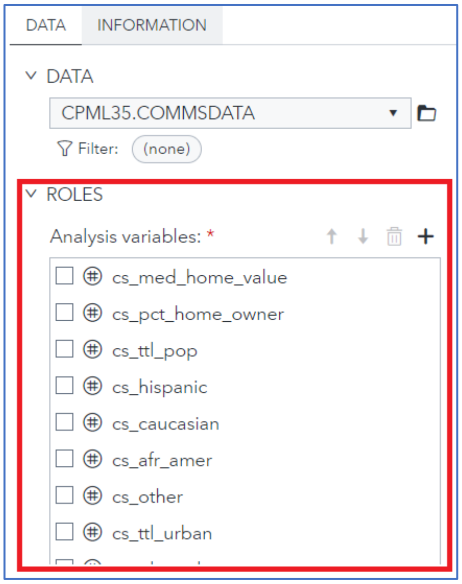
Analysis variables: *

Task Console (1)

Analysis variables: - Minimum variables: 1

```
/* Code for this task cannot be generated because of an error.
 * Please use the Task Console to view and fix the errors.
 */
```

- For the Analysis Variables, select all the variables beginning with *cs_*



DATA INFORMATION

DATA

CPML35.COMMSDATA

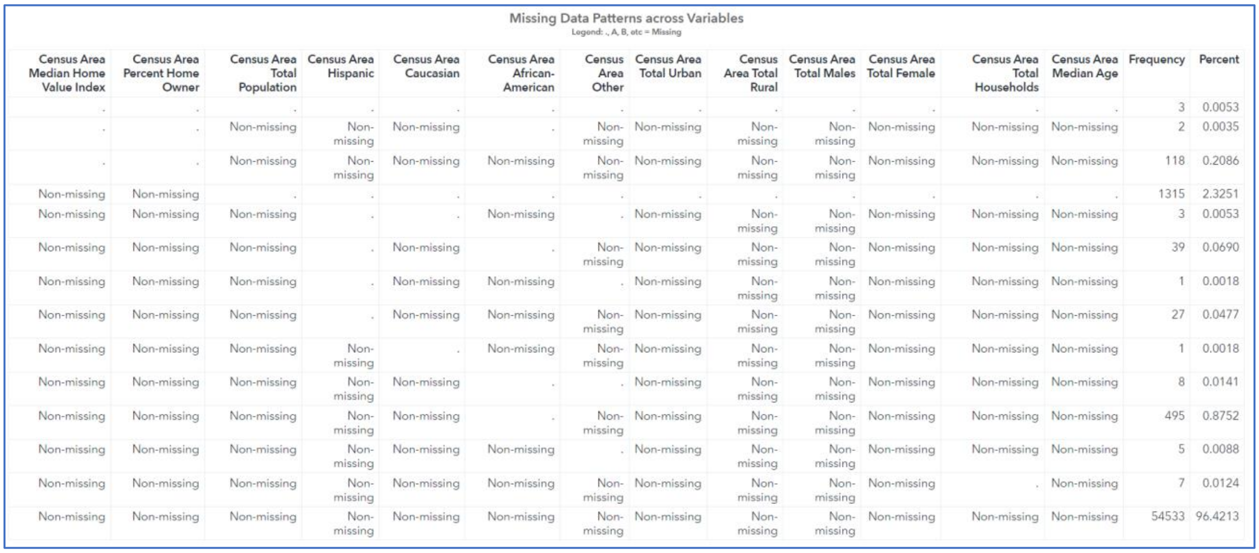
Filter: (none)

ROLES

Analysis variables: *

- □ cs_med_home_value
- □ cs_pct_home_owner
- □ cs_ttl_pop
- □ cs_hispanic
- □ cs_caucasian
- □ cs_afr_amer
- □ cs_other
- □ cs_ttl_urban

- **Run** the code!
- From our output, examine the pattern of missingness:



Missing Data Patterns across Variables

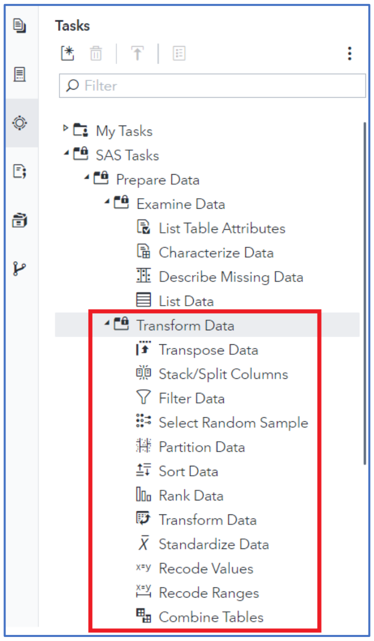
Legend: ., A, B, etc = Missing

| Census Area Median Home Value Index | Census Area Percent Home Owner | Census Area Total Population | Census Area Hispanic | Census Area Caucasian | Census Area African-American | Census Area Other | Census Area Total Urban | Census Area Total Rural | Census Area Total Males | Census Area Total Female | Census Area Total Households | Census Area Median Age | Frequency | Percent |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| . | . | . | . | . | . | . | . | . | . | . | . | . | 3 | 0.0053 |
| . | . | Non-missing | Non-missing | Non-missing | . | Non-missing | Non-missing | Non-missing | Non-missing | Non-missing | Non-missing | Non-missing | 2 | 0.0035 |
| . | . | Non-missing | Non-missing | Non-missing | Non-missing | Non-missing | Non-missing | Non-missing | Non-missing | Non-missing | Non-missing | Non-missing | 118 | 0.2086 |
| Non-missing | Non-missing | . | . | . | . | . | . | . | . | . | . | . | 1315 | 2.3251 |
| Non-missing | Non-missing | Non-missing | . | . | Non-missing | . | Non-missing | Non-missing | Non-missing | Non-missing | Non-missing | Non-missing | 3 | 0.0053 |
| Non-missing | Non-missing | Non-missing | . | Non-missing | . | Non-missing | Non-missing | Non-missing | Non-missing | Non-missing | Non-missing | Non-missing | 39 | 0.0690 |
| Non-missing | Non-missing | Non-missing | . | Non-missing | Non-missing | . | Non-missing | Non-missing | Non-missing | Non-missing | Non-missing | Non-missing | 1 | 0.0018 |
| Non-missing | Non-missing | Non-missing | . | Non-missing | Non-missing | Non-missing | Non-missing | Non-missing | Non-missing | Non-missing | Non-missing | Non-missing | 27 | 0.0477 |
| Non-missing | Non-missing | Non-missing | Non-missing | . | Non-missing | Non-missing | Non-missing | Non-missing | Non-missing | Non-missing | Non-missing | Non-missing | 1 | 0.0018 |
| Non-missing | Non-missing | Non-missing | Non-missing | Non-missing | . | . | Non-missing | Non-missing | Non-missing | Non-missing | Non-missing | Non-missing | 8 | 0.0141 |
| Non-missing | Non-missing | Non-missing | Non-missing | Non-missing | . | Non-missing | Non-missing | Non-missing | Non-missing | Non-missing | Non-missing | Non-missing | 495 | 0.8752 |
| Non-missing | Non-missing | Non-missing | Non-missing | Non-missing | Non-missing | . | Non-missing | Non-missing | Non-missing | Non-missing | Non-missing | Non-missing | 5 | 0.0088 |
| Non-missing | Non-missing | Non-missing | Non-missing | Non-missing | Non-missing | Non-missing | Non-missing | Non-missing | Non-missing | Non-missing | . | Non-missing | 7 | 0.0124 |
| Non-missing | Non-missing | Non-missing | Non-missing | Non-missing | Non-missing | Non-missing | Non-missing | Non-missing | Non-missing | Non-missing | Non-missing | Non-missing | 54533 | 96.4213 |

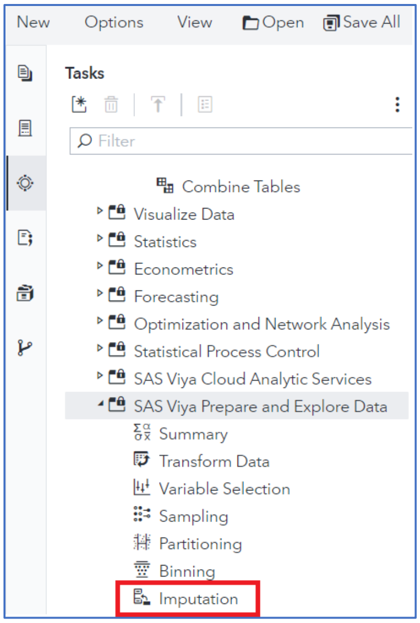
  - Notice any interesting trends?
  - When and why could data be missing?

- Great – the data need to be fixed. How do we do that in SAS Studio?
  - **Option 1**: code
    - But, you've gotta know how to write SAS code
  - **Option 2**: more SAS Studio tasks

    - Find the options under **Transform Data**:



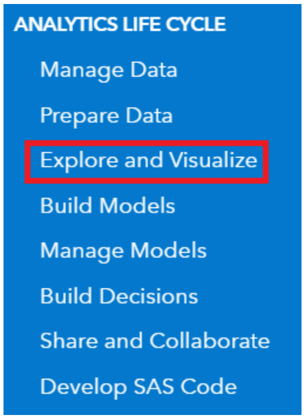
    - Some greatest hits:
        - Transform Data
        - Standardize Data
        - Recode Values
        - Combine Tables
    - *Note: where is imputation? Sorry, SAS 9… it's not a readily available task!*
- **Option 3**:
    - Upload data to CAS
    - Then use SAS Studio Tasks specific to SAS Viya, such as Imputation:



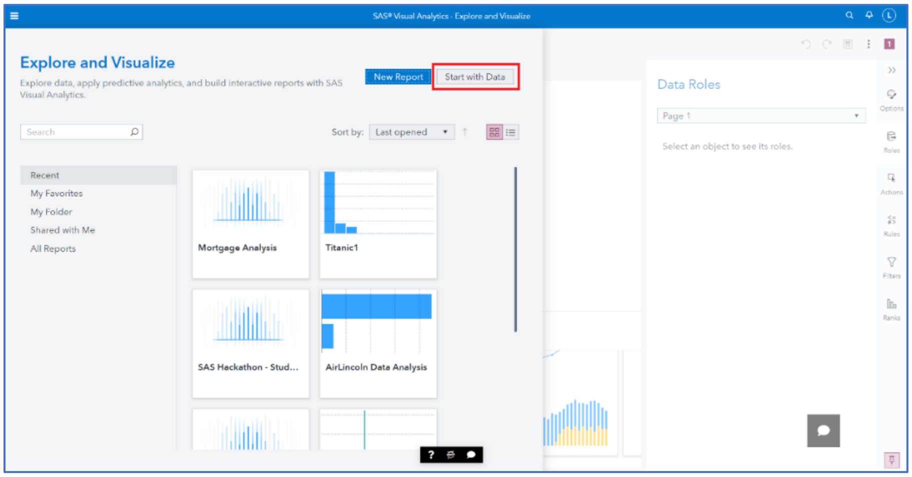
- Where can I learn more?
    - SAS Programming 1: Essentials
    - SAS Programming 2: Data Manipulation Techniques

## Part II: SAS Visual Analytics

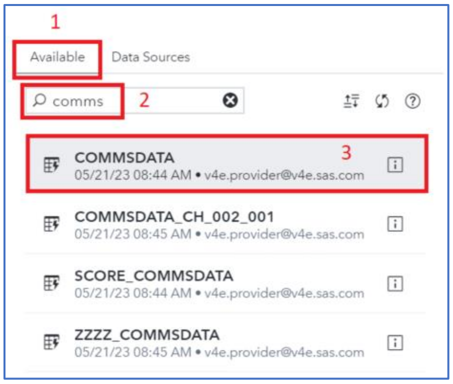
- **Objective**
  - Better understand the data investigation tools available in SAS Visual Analytics
  - Examine data visually in a dashboard
  - Use Auto chart functions – and other tools – to simplify the data preprocessing stage
- **Setup**
  - Move on over to SAS Visual Analytics



  - Create a new report utilizing **Start with Data**:



  - Load our data:



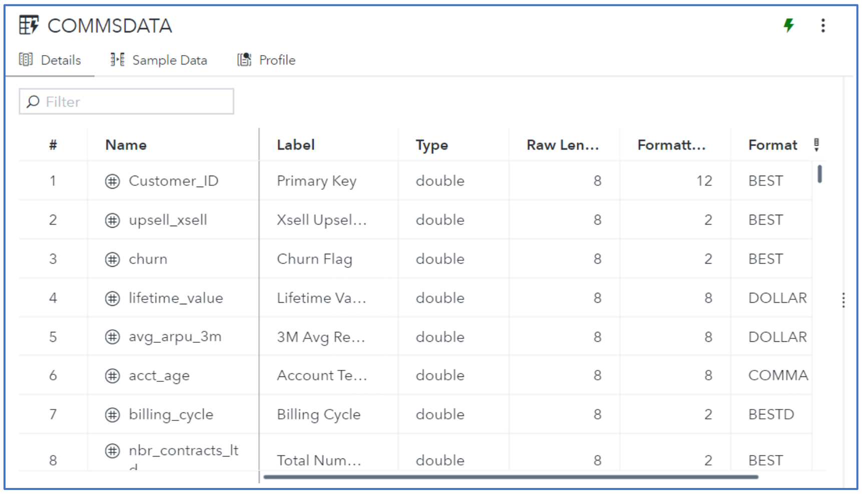
COMMSDATA
05/21/23 08:44 AM • v4e.provider@v4e.sas.com

COMMSDATA_CH_002_001
05/21/23 08:45 AM • v4e.provider@v4e.sas.com

SCORE_COMMSDATA
05/21/23 08:44 AM • v4e.provider@v4e.sas.com

ZZZZ_COMMSDATA
05/21/23 08:45 AM • v4e.provider@v4e.sas.com

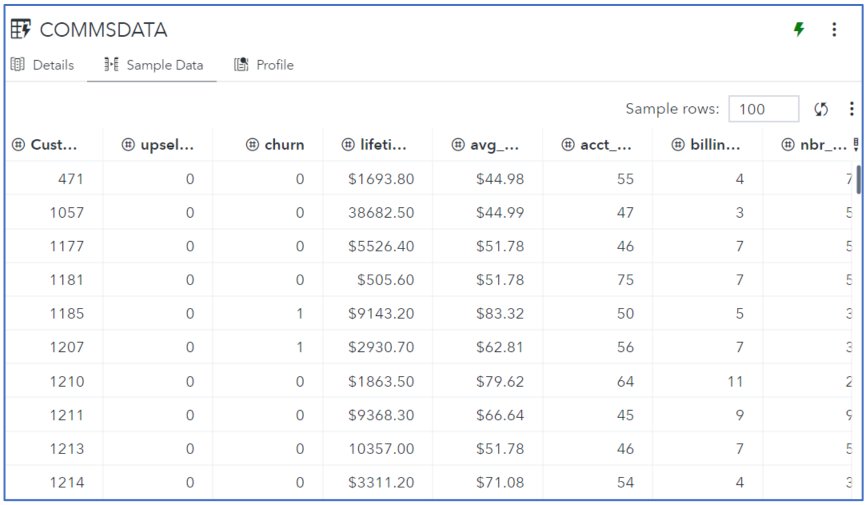
- Examine measure **Details**:



COMMSDATA

| # | Name | Label | Type | Raw Len… | Formatt… | Format |
|---|---|---|---|---|---|---|
| 1 | Customer_ID | Primary Key | double | 8 | 12 | BEST |
| 2 | upsell_xsell | Xsell Upsel… | double | 8 | 2 | BEST |
| 3 | churn | Churn Flag | double | 8 | 2 | BEST |
| 4 | lifetime_value | Lifetime Va… | double | 8 | 8 | DOLLAR |
| 5 | avg_arpu_3m | 3M Avg Re… | double | 8 | 8 | DOLLAR |
| 6 | acct_age | Account Te… | double | 8 | 8 | COMMA |
| 7 | billing_cycle | Billing Cycle | double | 8 | 2 | BESTD |
| 8 | nbr_contracts_ltd | Total Num… | double | 8 | 2 | BEST |

- Explore **Sample Data**:

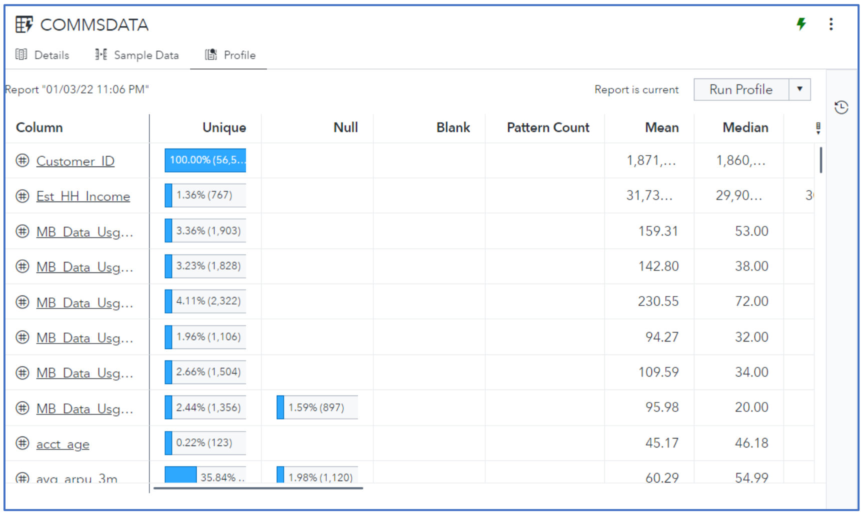


COMMSDATA

Sample rows: 100

| Cust… | upsel… | churn | lifeti… | avg_… | acct_… | billin… | nbr_… |
|---|---|---|---|---|---|---|---|
| 471 | 0 | 0 | $1693.80 | $44.98 | 55 | 4 | 7 |
| 1057 | 0 | 0 | 38682.50 | $44.99 | 47 | 3 | 5 |
| 1177 | 0 | 0 | $5526.40 | $51.78 | 46 | 7 | 5 |
| 1181 | 0 | 0 | $505.60 | $51.78 | 75 | 7 | 5 |
| 1185 | 0 | 1 | $9143.20 | $83.32 | 50 | 5 | 3 |
| 1207 | 0 | 1 | $2930.70 | $62.81 | 56 | 7 | 3 |
| 1210 | 0 | 0 | $1863.50 | $79.62 | 64 | 11 | 2 |
| 1211 | 0 | 0 | $9368.30 | $66.64 | 45 | 9 | 9 |
| 1213 | 0 | 0 | 10357.00 | $51.78 | 46 | 7 | 5 |
| 1214 | 0 | 0 | $3311.20 | $71.08 | 54 | 4 | 3 |

- Run/examine a **Profile**:



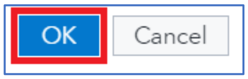
COMMSDATA

Report "01/03/22 11:06 PM"

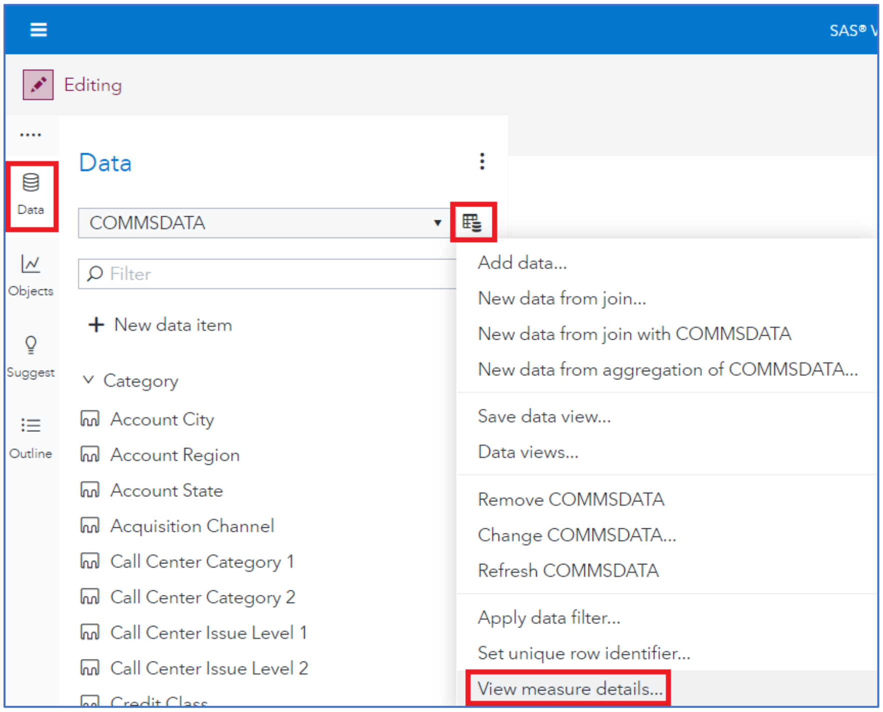
Report is current

| Column | Unique | Null | Blank | Pattern Count | Mean | Median |
|---|---|---|---|---|---|---|
| Customer_ID | 100.00% (56,5… | | | | 1,871,… | 1,860,… |
| Est_HH_Income | 1.36% (767) | | | | 31,73… | 29,90… |
| MB_Data_Usg… | 3.36% (1,903) | | | | 159.31 | 53.00 |
| MB_Data_Usg… | 3.23% (1,828) | | | | 142.80 | 38.00 |
| MB_Data_Usg… | 4.11% (2,322) | | | | 230.55 | 72.00 |
| MB_Data_Usg… | 1.96% (1,106) | | | | 94.27 | 32.00 |
| MB_Data_Usg… | 2.66% (1,504) | | | | 109.59 | 34.00 |
| MB_Data_Usg… | 2.44% (1,356) | 1.59% (897) | | | 95.98 | 20.00 |
| acct_age | 0.22% (123) | | | | 45.17 | 46.18 |
| avg_arpu_3m | 35.84% .. | 1.98% (1,120) | | | 60.29 | 54.99 |

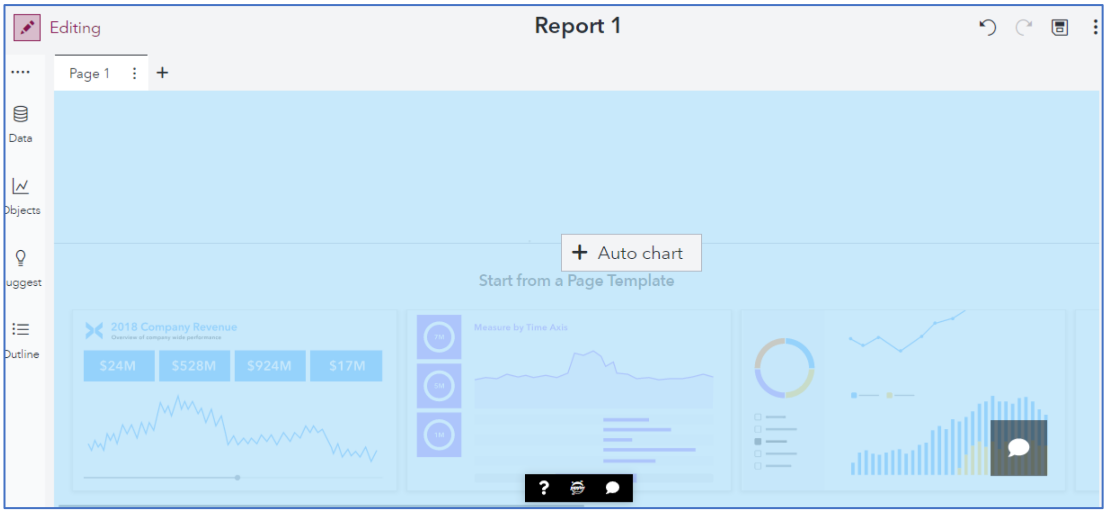
- Click Ok to (finally) load our data



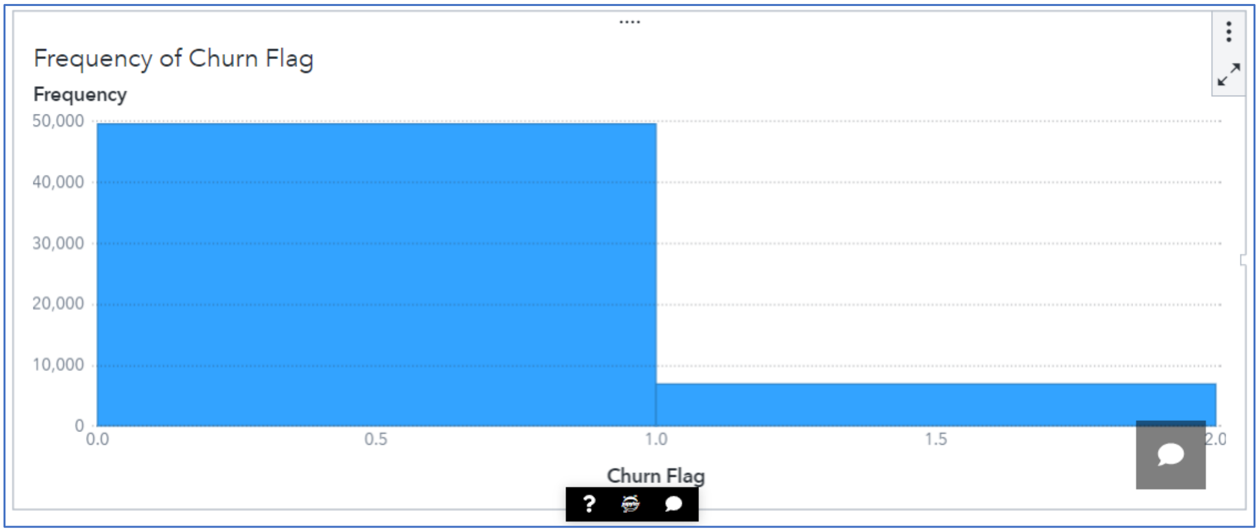
- Want even more preliminary statistics?
  - **Data ➔ Actions ➔ View measure details…**



- But, **View measure details…** aren't permanent in the dashboard. So, let's plot instead. Start with the outcome variable **Churn**. Drag-and-drop Churn on the canvas:

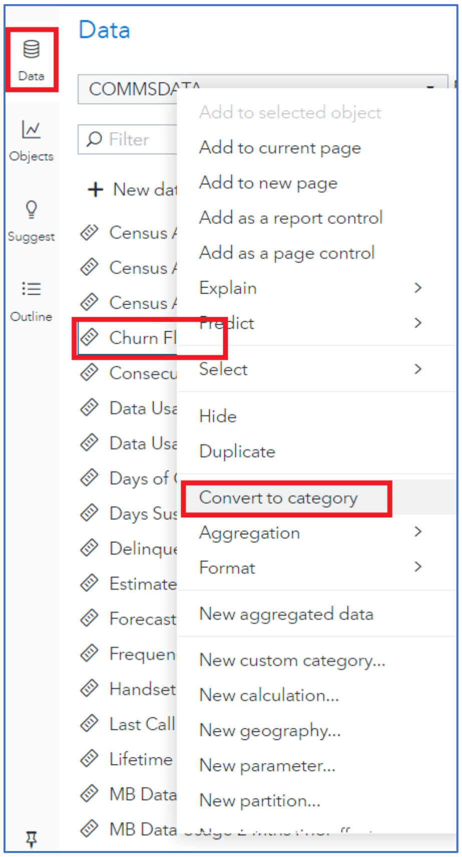


- Examine the output. Notice anything weird?

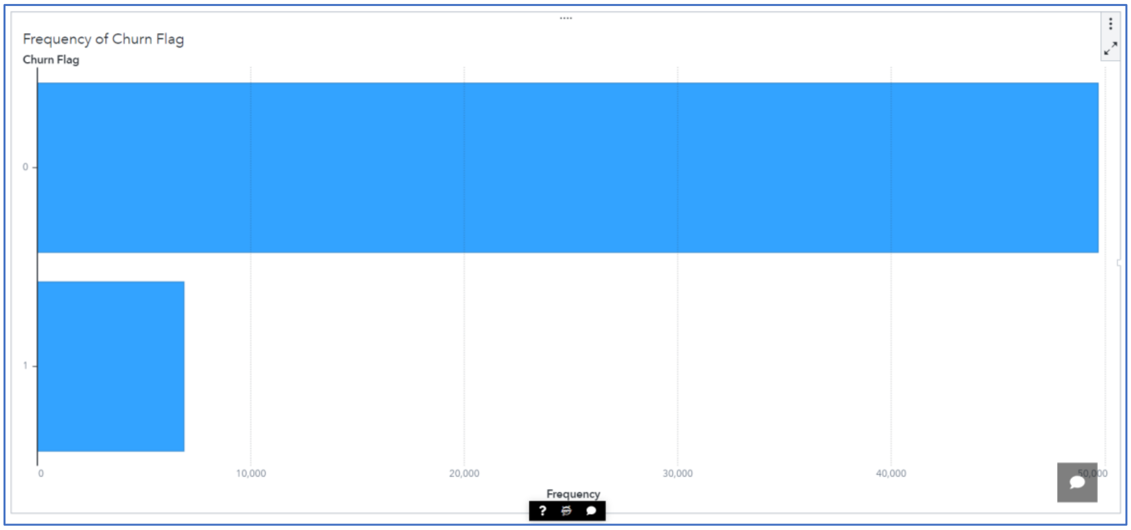


  - I do! Churn is a yes/no, or 1/0, variable. So, we don't want it as a regular ole **Measure** in SAS VA.

- How can we change it?
  - Delete the Frequency of Churn Flag chart.
  - **Data ➔ Churn Flag ➔ Convert to category**

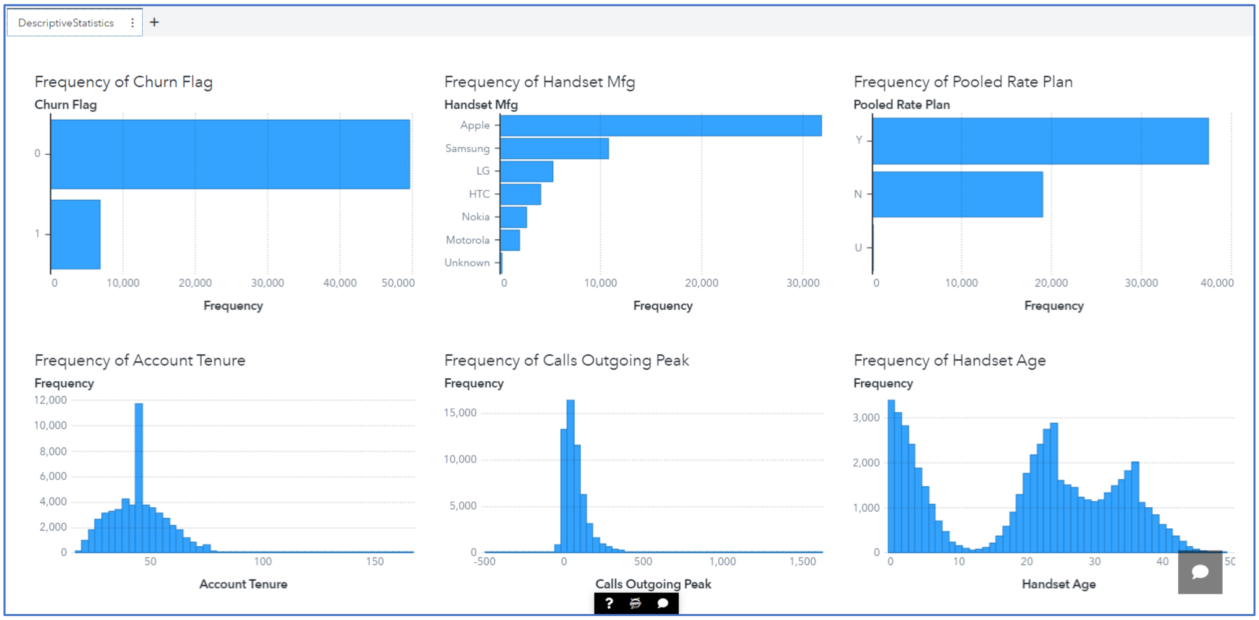


        - Now drag-and-drop Churn to the canvas:



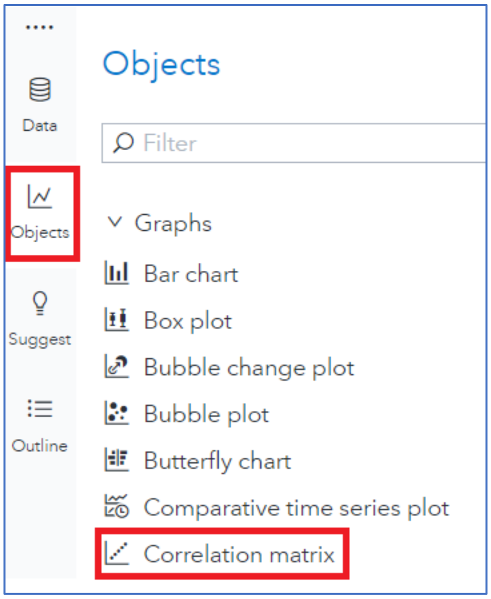
            - So much better!

- Let’s add more **Auto charts** to the canvas. Plotting the data can help us better understand our data type and whether statistical challenges, such as skewness, are an issue.
    - Follow my lead and produce the following:

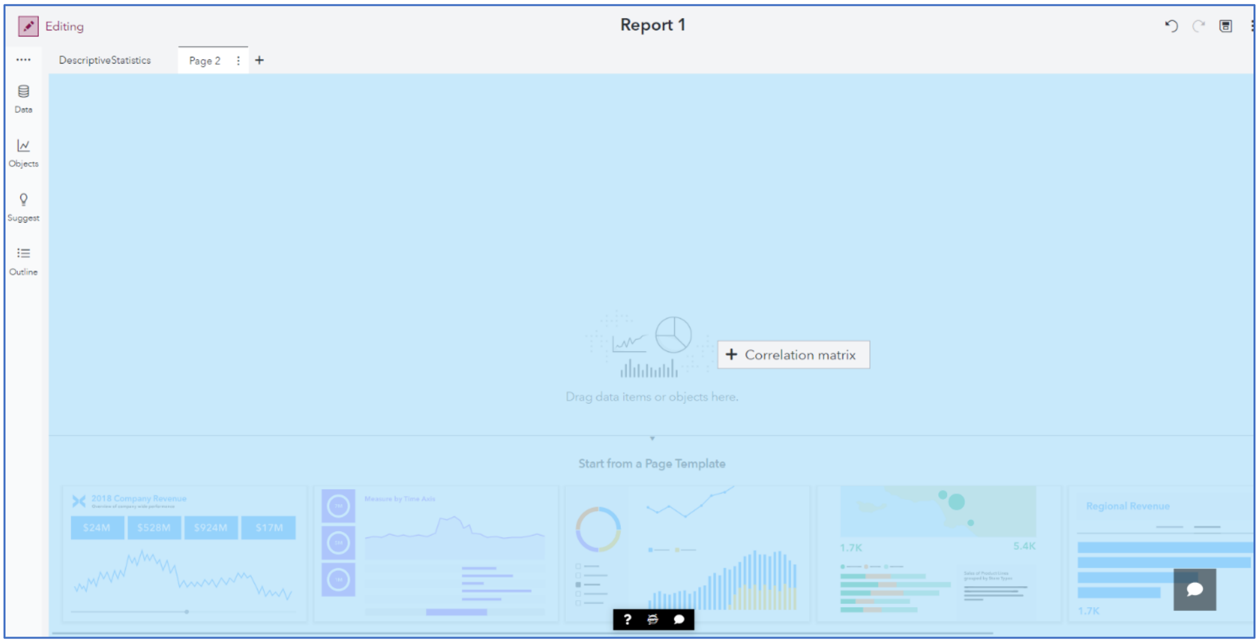


    - Rename the page to “DescriptiveStatistics”
- Want to see how your variables are related?
    - How about a correlation matrix?
        - Yes!
        - Start by creating a new page

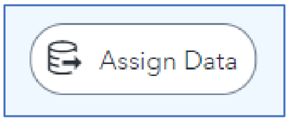
- Then navigate to **Objects → Correlation matrix**



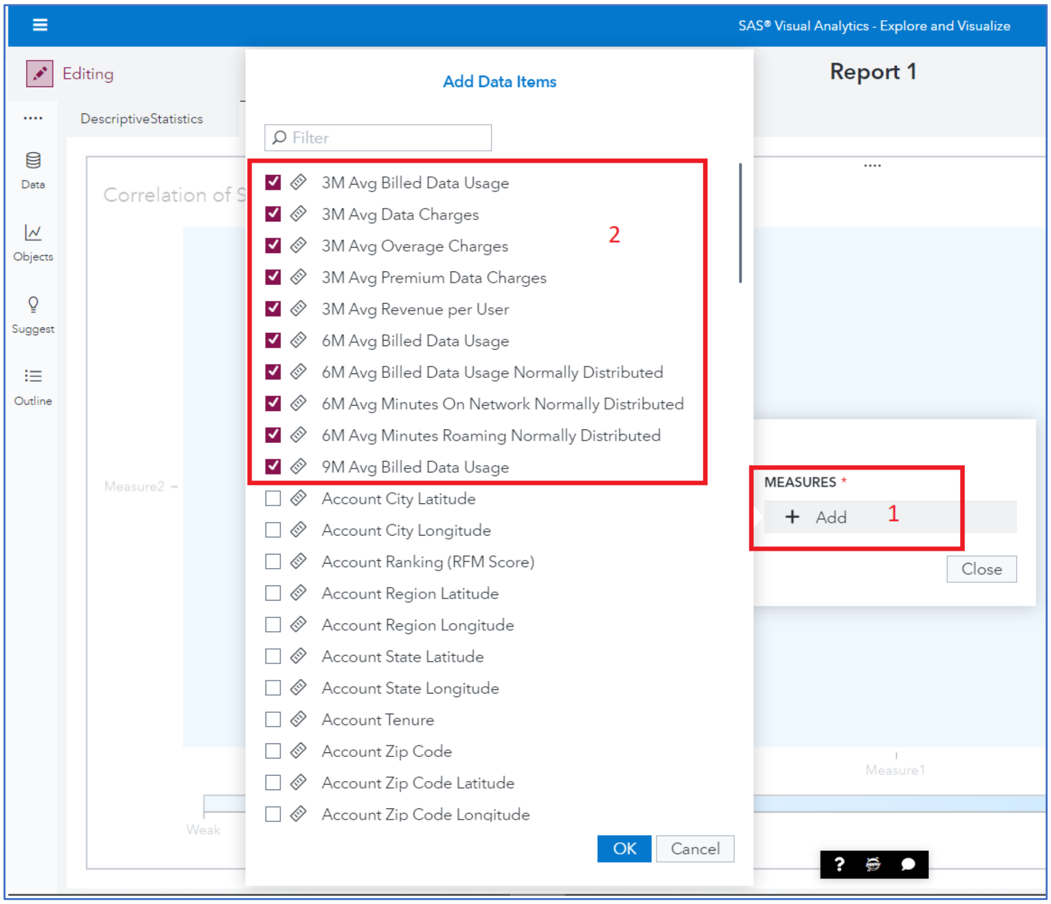
- Pull it onto the Page 2 canvas:



- Click on **Assign Data**



- There are a TON of variables to choose from. Let's not make it too overwhelming and just select the first 10 variables – all of which are some sort of over time variable:



**Add Data Items**

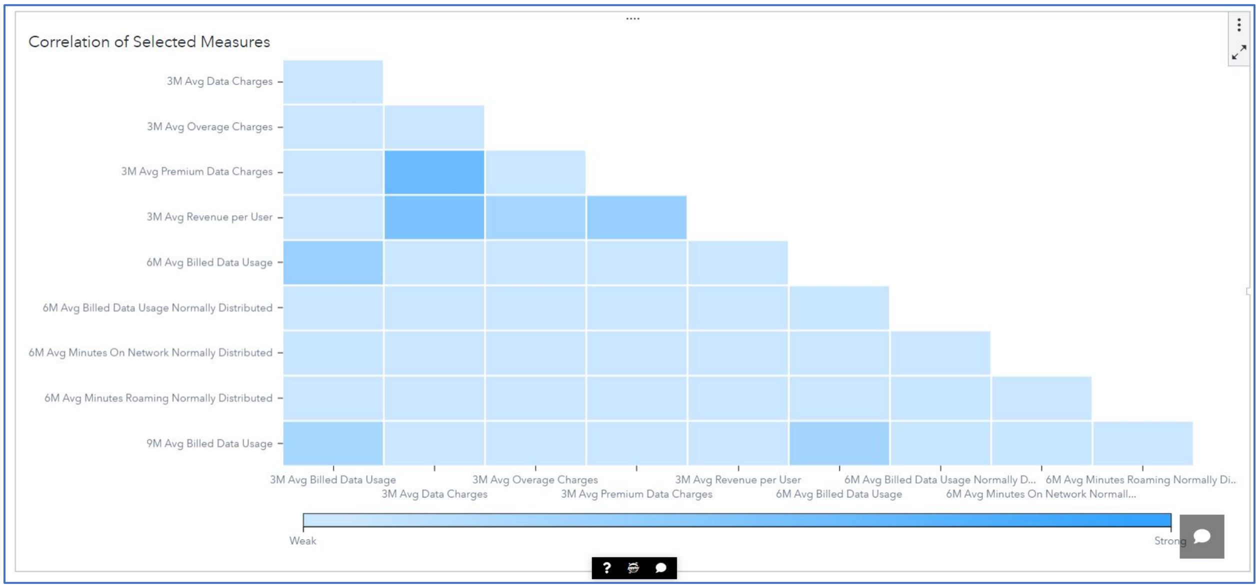
Filter

- ☑ 3M Avg Billed Data Usage
- ☑ 3M Avg Data Charges
- ☑ 3M Avg Overage Charges
- ☑ 3M Avg Premium Data Charges
- ☑ 3M Avg Revenue per User
- ☑ 6M Avg Billed Data Usage
- ☑ 6M Avg Billed Data Usage Normally Distributed
- ☑ 6M Avg Minutes On Network Normally Distributed
- ☑ 6M Avg Minutes Roaming Normally Distributed
- ☑ 9M Avg Billed Data Usage
- □ Account City Latitude
- □ Account City Longitude
- □ Account Ranking (RFM Score)
- □ Account Region Latitude
- □ Account Region Longitude
- □ Account State Latitude
- □ Account State Longitude
- □ Account Tenure
- □ Account Zip Code
- □ Account Zip Code Latitude
- □ Account Zip Code Longitude

OK Cancel

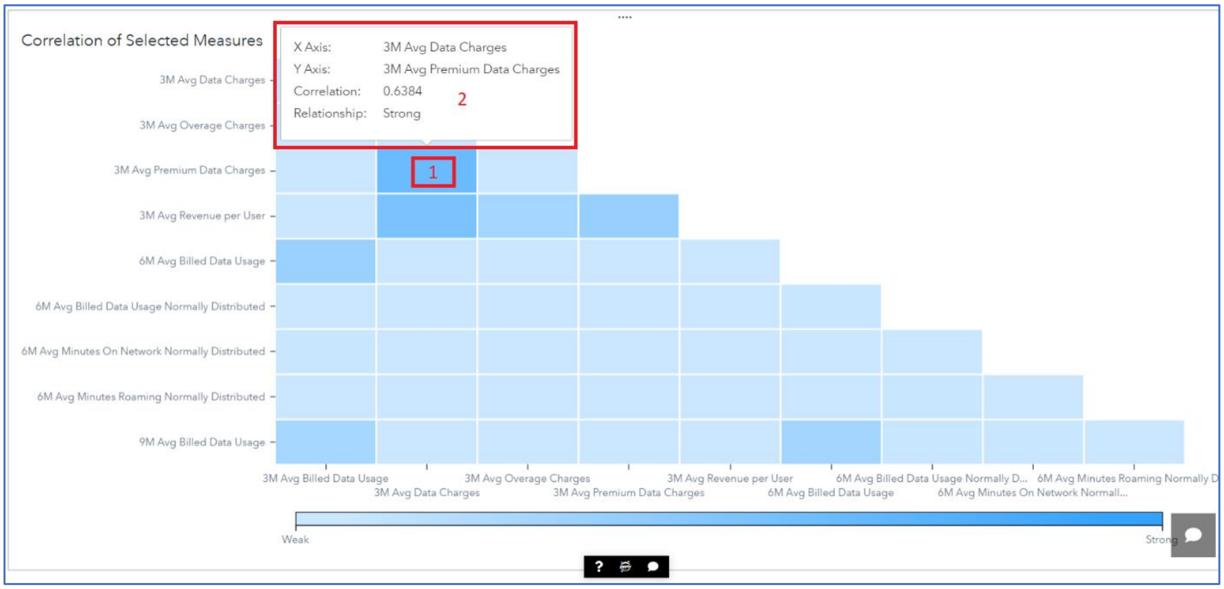
MEASURES * + Add

- Click **OK** to produce the output:



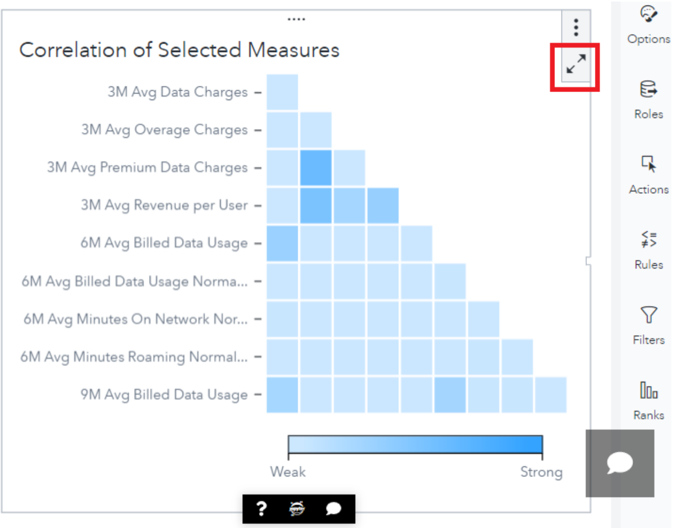
    - How do we interpret this chart? Well, the darker the blue, the stronger the correlation

- Do you want to know individual correlation values? Simply hover over a relationship:

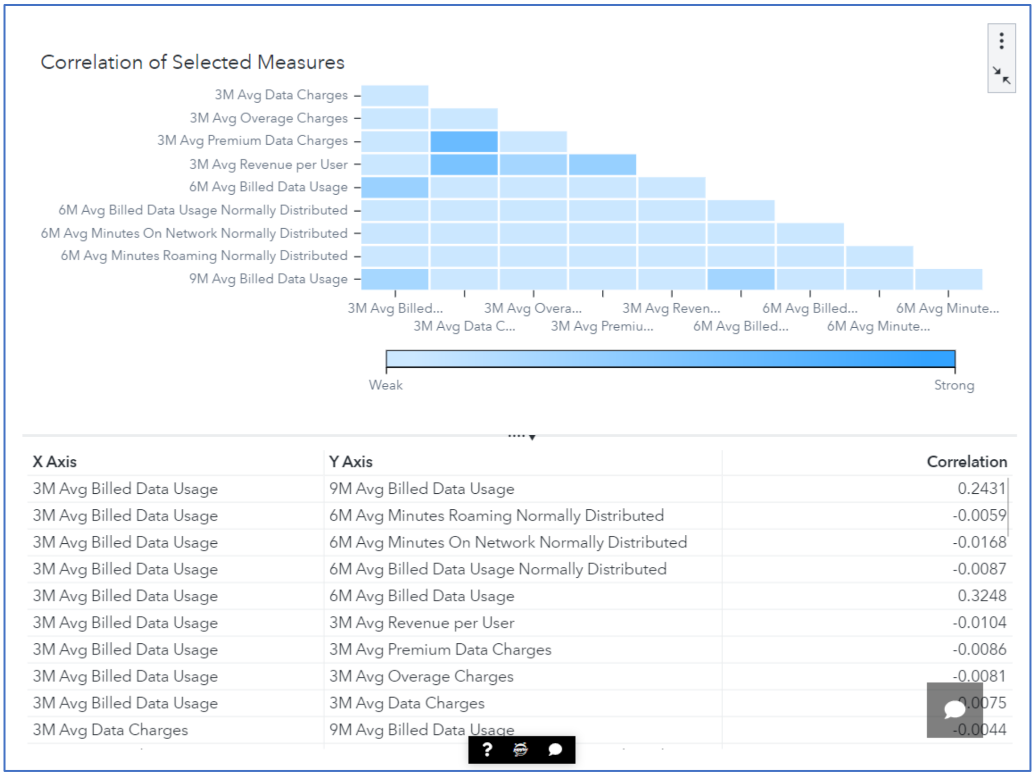

Correlation of Selected Measures
3M Avg Data Charges
3M Avg Overage Charges
3M Avg Premium Data Charges
3M Avg Revenue per User
6M Avg Billed Data Usage
6M Avg Billed Data Usage Normally Distributed
6M Avg Minutes On Network Normally Distributed
6M Avg Minutes Roaming Normally Distributed
9M Avg Billed Data Usage
3M Avg Billed Data Usage
3M Avg Data Charges
3M Avg Overage Charges
3M Avg Premium Data Charges
3M Avg Revenue per User
6M Avg Billed Data Usage
6M Avg Billed Data Usage Normally D...
6M Avg Minutes On Network Normall...
6M Avg Minutes Roaming Normally D
1
2
Weak
Strong

X Axis: 3M Avg Data Charges
Y Axis: 3M Avg Premium Data Charges
Correlation: 0.6384
Relationship: Strong

- Want all the correlational statistics you can handle?
    - Find the Maximize button:

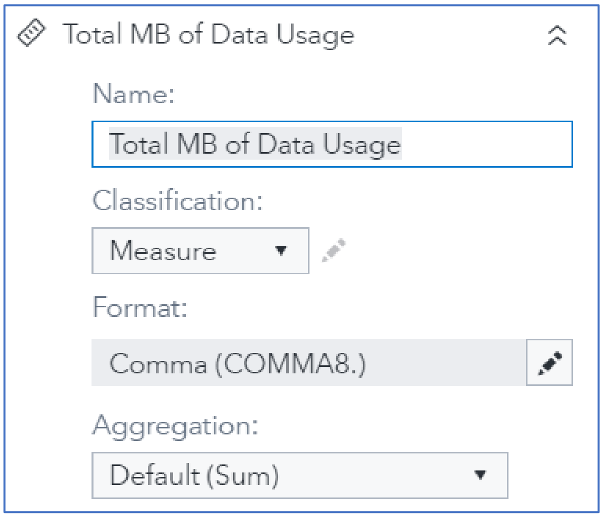


- Click it and explore all the underlying statistics:



| X Axis | Y Axis | Correlation |
|---|---|---|
| 3M Avg Billed Data Usage | 9M Avg Billed Data Usage | 0.2431 |
| 3M Avg Billed Data Usage | 6M Avg Minutes Roaming Normally Distributed | -0.0059 |
| 3M Avg Billed Data Usage | 6M Avg Minutes On Network Normally Distributed | -0.0168 |
| 3M Avg Billed Data Usage | 6M Avg Billed Data Usage Normally Distributed | -0.0087 |
| 3M Avg Billed Data Usage | 6M Avg Billed Data Usage | 0.3248 |
| 3M Avg Billed Data Usage | 3M Avg Revenue per User | -0.0104 |
| 3M Avg Billed Data Usage | 3M Avg Premium Data Charges | -0.0086 |
| 3M Avg Billed Data Usage | 3M Avg Overage Charges | -0.0081 |
| 3M Avg Billed Data Usage | 3M Avg Data Charges | 0.0075 |
| 3M Avg Data Charges | 9M Avg Billed Data Usage | -0.0044 |

- **Great – I see that there are issues with the data. What can I do in SAS Visual Analytics?**
  - **Option 1**: change the variable **Classification**
    - We saw this with Churn above
    - You can also change some of the other data attributes under **Edit Properties.** An example:



Total MB of Data Usage

Name:

Total MB of Data Usage

Classification:

Measure ▼

Format:

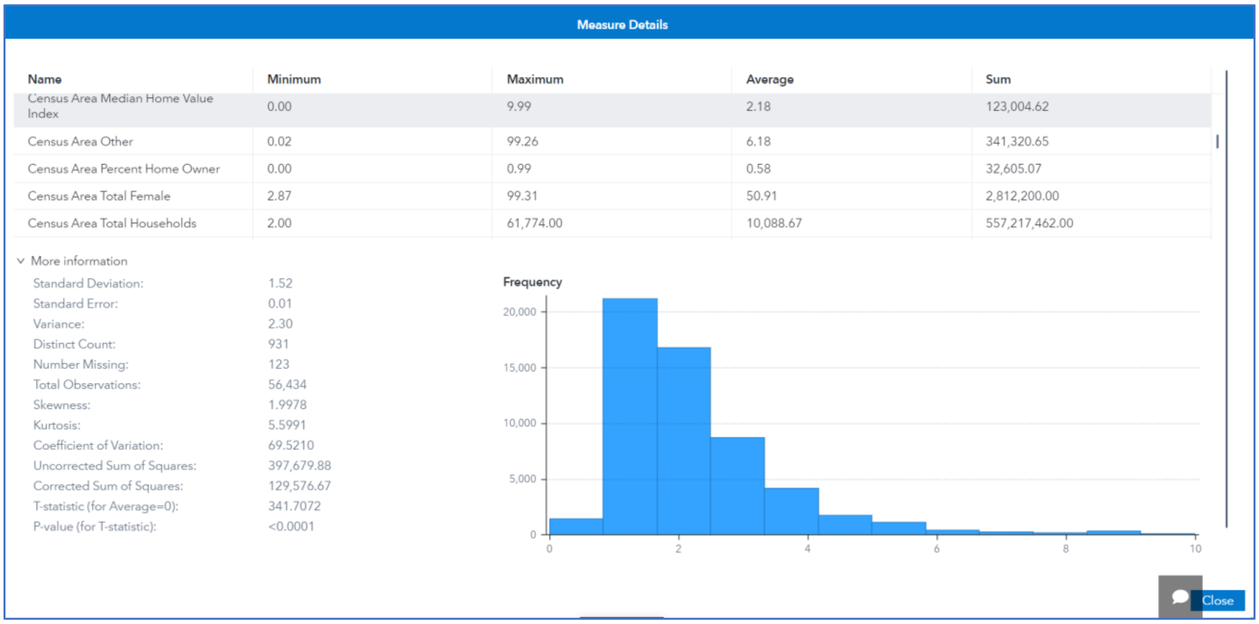
Comma (COMMA8.)

Aggregation:

Default (Sum) ▼

  - **Option 2**: create a **New data item**
    - This is likely the most helpful tool to create new variables

- For our example, let’s suppose that we want to change the 123 missing values of Census Area Median Home Value Index to the average for the whole data set.
    - From the Measure Details, we can see that the average value for this variable is 2.18.
    - Additionally, you can confirm that 123 observations have missing values:



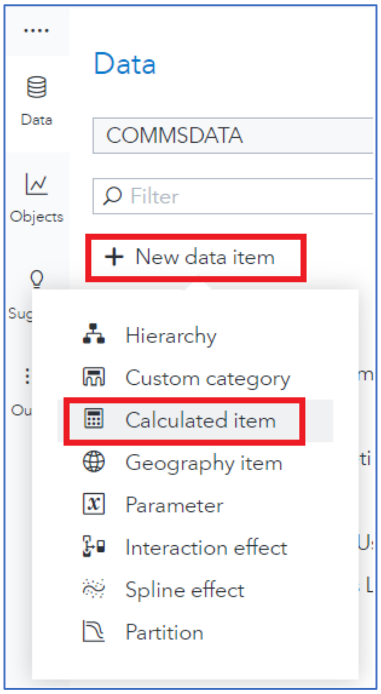
**Measure Details**

| Name | Minimum | Maximum | Average | Sum |
|---|---|---|---|---|
| Census Area Median Home Value Index | 0.00 | 9.99 | 2.18 | 123,004.62 |
| Census Area Other | 0.02 | 99.26 | 6.18 | 341,320.65 |
| Census Area Percent Home Owner | 0.00 | 0.99 | 0.58 | 32,605.07 |
| Census Area Total Female | 2.87 | 99.31 | 50.91 | 2,812,200.00 |
| Census Area Total Households | 2.00 | 61,774.00 | 10,088.67 | 557,217,462.00 |

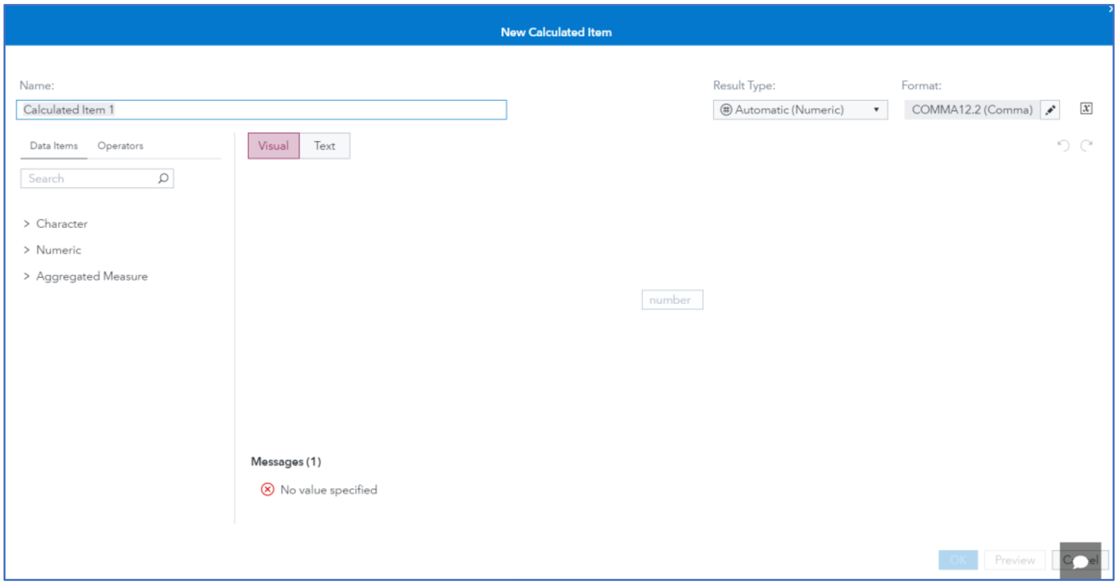
More information

| | |
|---|---|
| Standard Deviation: | 1.52 |
| Standard Error: | 0.01 |
| Variance: | 2.30 |
| Distinct Count: | 931 |
| Number Missing: | 123 |
| Total Observations: | 56,434 |
| Skewness: | 1.9978 |
| Kurtosis: | 5.5991 |
| Coefficient of Variation: | 69.5210 |
| Uncorrected Sum of Squares: | 397,679.88 |
| Corrected Sum of Squares: | 129,576.67 |
| T-statistic (for Average=0): | 341.7072 |
| P-value (for T-statistic): | <0.0001 |

    - So, the goal is to create a new variable that has 2.18 where the previous value for Census Area Median Home Value Index was missing
- To do this, navigate to **New data item ➔ Calculated item**:



- Which yields the following **New Calculated Item** window:



Name: Calculated Item 1

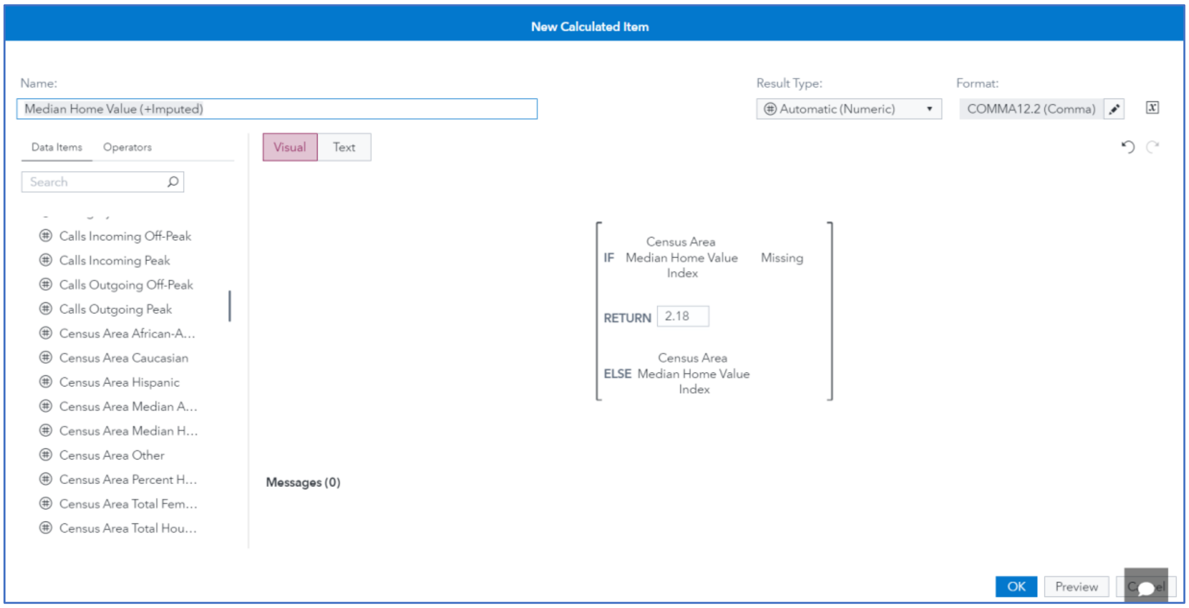
Result Type: Automatic (Numeric)

Format: COMMA12.2 (Comma)

Messages (1)

No value specified

- There is a LOT to explore here. Let's just keep it simple and make the following changes:



Name: Median Home Value (+Imputed)

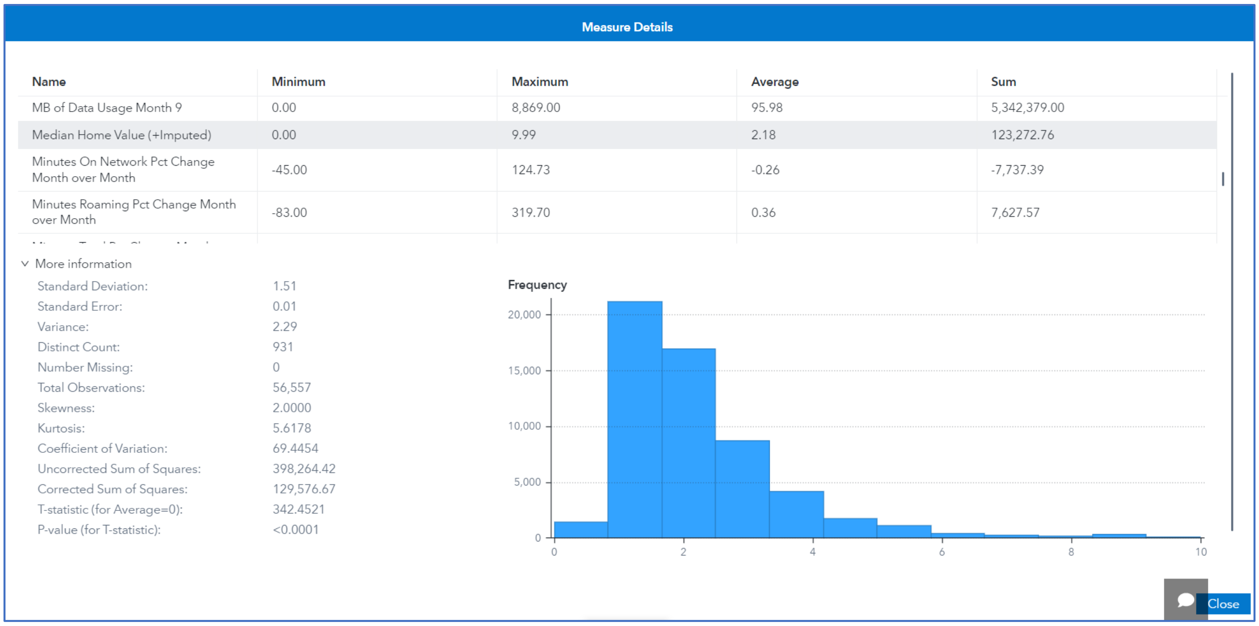
Result Type: Automatic (Numeric)

Format: COMMA12.2 (Comma)

```
IF Census Area Median Home Value Index Missing
RETURN 2.18
ELSE Census Area Median Home Value Index
```

Messages (0)

- Click **OK**
- Navigate back to the **Measure Details** for our new variable and ensure that there are no missing values:



| Name | Minimum | Maximum | Average | Sum |
|---|---|---|---|---|
| MB of Data Usage Month 9 | 0.00 | 8,869.00 | 95.98 | 5,342,379.00 |
| Median Home Value (+Imputed) | 0.00 | 9.99 | 2.18 | 123,272.76 |
| Minutes On Network Pct Change Month over Month | -45.00 | 124.73 | -0.26 | -7,737.39 |
| Minutes Roaming Pct Change Month over Month | -83.00 | 319.70 | 0.36 | 7,627.57 |

More information

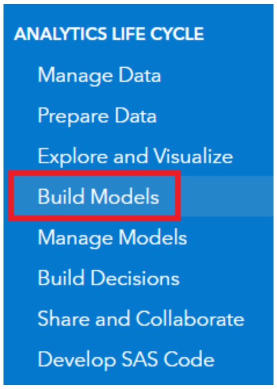
| | |
|---|---|
| Standard Deviation: | 1.51 |
| Standard Error: | 0.01 |
| Variance: | 2.29 |
| Distinct Count: | 931 |
| Number Missing: | 0 |
| Total Observations: | 56,557 |
| Skewness: | 2.0000 |
| Kurtosis: | 5.6178 |
| Coefficient of Variation: | 69.4454 |
| Uncorrected Sum of Squares: | 398,264.42 |
| Corrected Sum of Squares: | 129,576.67 |
| T-statistic (for Average=0): | 342.4521 |
| P-value (for T-statistic): | <0.0001 |

  - Bingo!

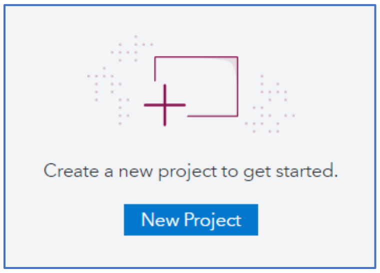
- I definitely did not do the **New Data items** justice. To learn more, I recommend the following SAS courses:
  - SAS Visual Analytics 1 for SAS Viya: Basics
  - SAS Visual Analytics 2 for SAS Viya: Advanced

## Part III: SAS Model Studio

- **Objective**
    - Introduce the Data Exploration node
    - Expose users to Data Mining Preprocessing tools via Automated Pipelines
    - Show how Feature Machine can help address certain data challenges
- **Start a new SAS Model Studio Project**
    - Navigate to SAS Model Studio

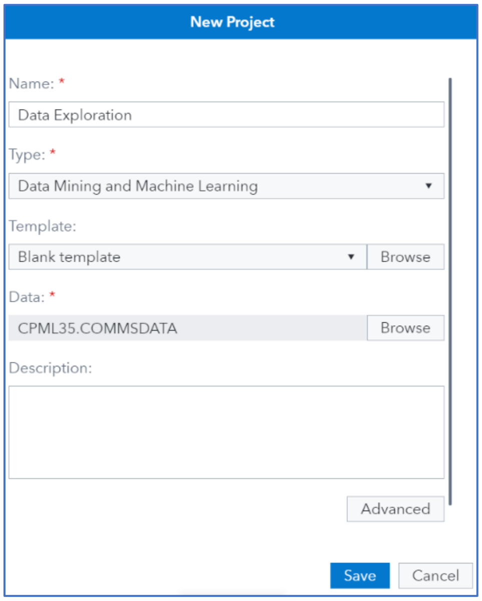


- Start a New Project with 1 of 2 buttons:



Or

- In the **New Project** window, match the following settings:



Name: * Data Exploration

Type: * Data Mining and Machine Learning

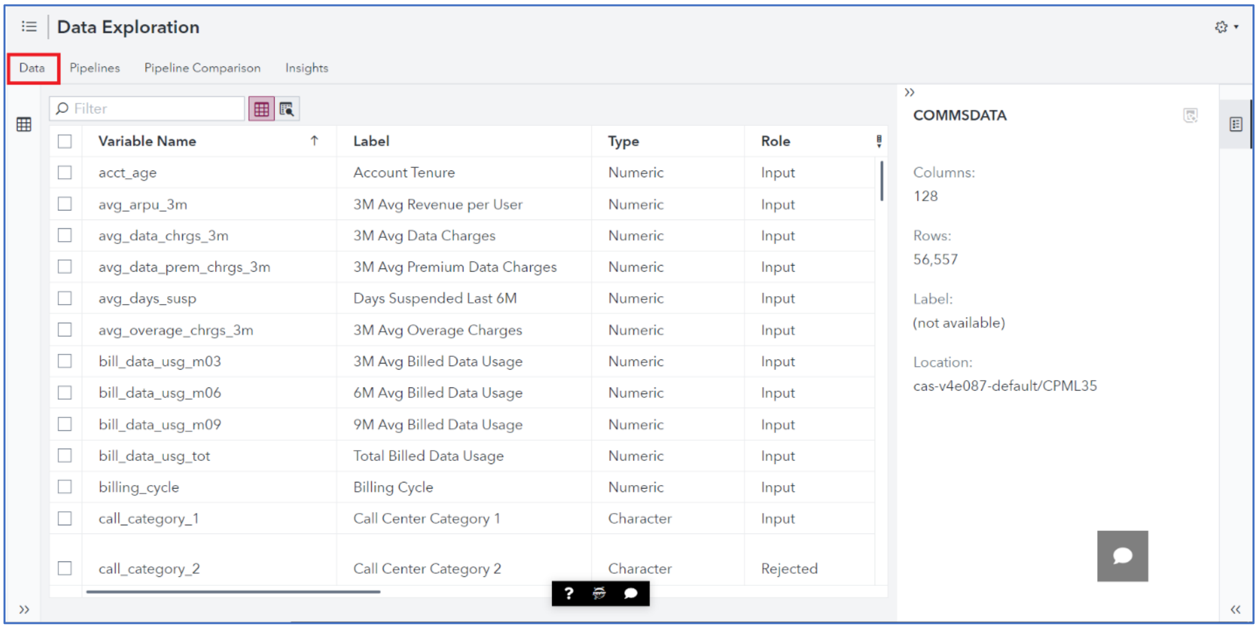
Template: Blank template

Data: * CPML35.COMMSDATA

Description:

- Click **Save** to create the project.

- **Explore Metadata**
  - The next step is to ensure that the metadata are set up properly. Start on the **Data** tab:



|  | Variable Name | Label | Type | Role |
|---|---|---|---|---|
| □ | acct_age | Account Tenure | Numeric | Input |
| □ | avg_arpu_3m | 3M Avg Revenue per User | Numeric | Input |
| □ | avg_data_chrgs_3m | 3M Avg Data Charges | Numeric | Input |
| □ | avg_data_prem_chrgs_3m | 3M Avg Premium Data Charges | Numeric | Input |
| □ | avg_days_susp | Days Suspended Last 6M | Numeric | Input |
| □ | avg_overage_chrgs_3m | 3M Avg Overage Charges | Numeric | Input |
| □ | bill_data_usg_m03 | 3M Avg Billed Data Usage | Numeric | Input |
| □ | bill_data_usg_m06 | 6M Avg Billed Data Usage | Numeric | Input |
| □ | bill_data_usg_m09 | 9M Avg Billed Data Usage | Numeric | Input |
| □ | bill_data_usg_tot | Total Billed Data Usage | Numeric | Input |
| □ | billing_cycle | Billing Cycle | Numeric | Input |
| □ | call_category_1 | Call Center Category 1 | Character | Input |
| □ | call_category_2 | Call Center Category 2 | Character | Rejected |

COMMSDATA

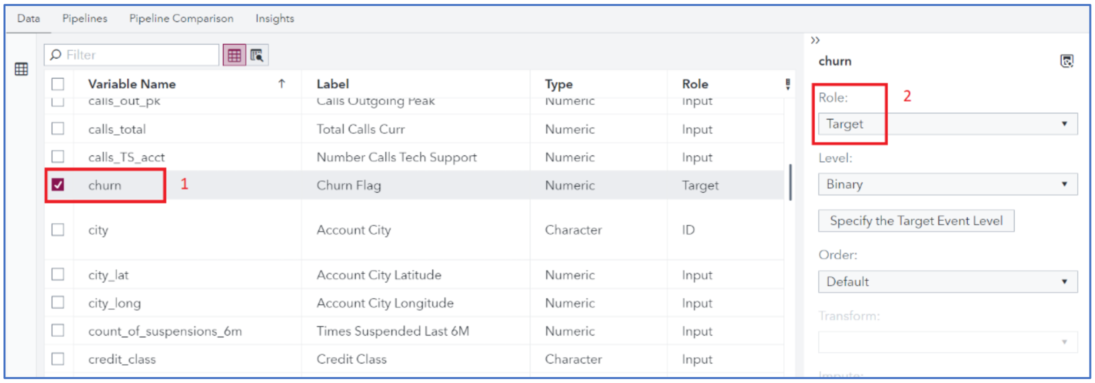
Columns:
128

Rows:
56,557

Label:
(not available)

Location:
cas-v4e087-default/CPML35

  - Find our outcome variable, *Churn*. Ensure that the **Role** is set to **Target**.



|  | Variable Name | Label | Type | Role |
|---|---|---|---|---|
| □ | calls_out_pk | Calls Outgoing Peak | Numeric | Input |
| □ | calls_total | Total Calls Curr | Numeric | Input |
| □ | calls_TS_acct | Number Calls Tech Support | Numeric | Input |
| ☑ | churn | Churn Flag | Numeric | Target |
| □ | city | Account City | Character | ID |
| □ | city_lat | Account City Latitude | Numeric | Input |
| □ | city_long | Account City Longitude | Numeric | Input |
| □ | count_of_suspensions_6m | Times Suspended Last 6M | Numeric | Input |
| □ | credit_class | Credit Class | Character | Input |

churn

Role:
Target

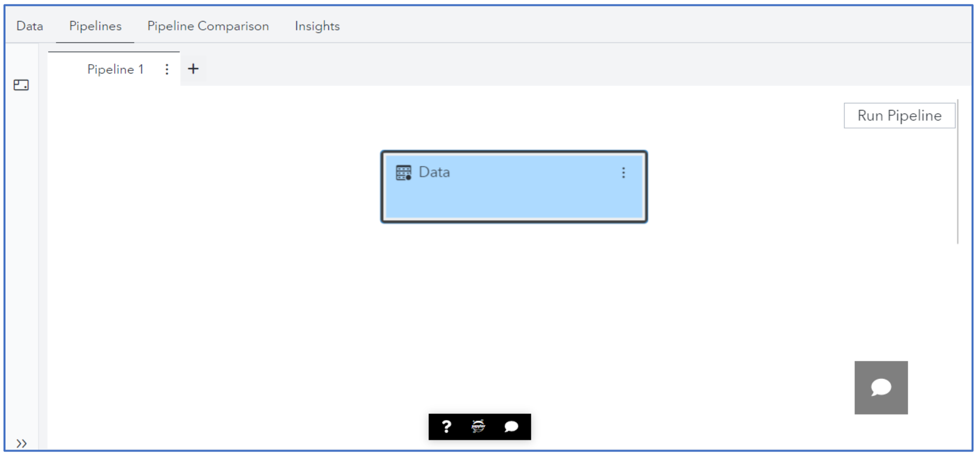
Level:
Binary

Order:
Default

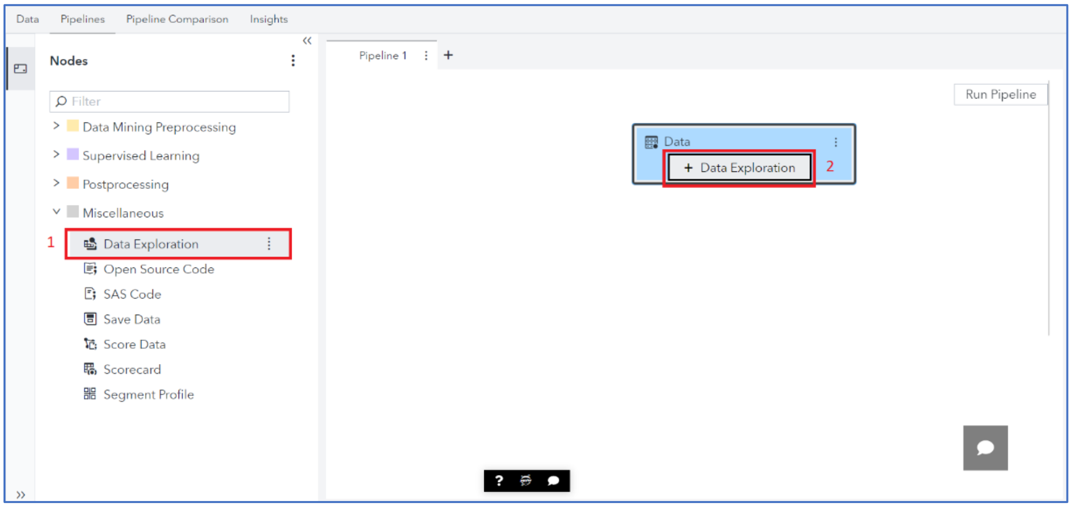
Transform:

  - Accept all other **Roles** as defined, as we're here to learn – not find the perfect model. Perfection is for another day 😊

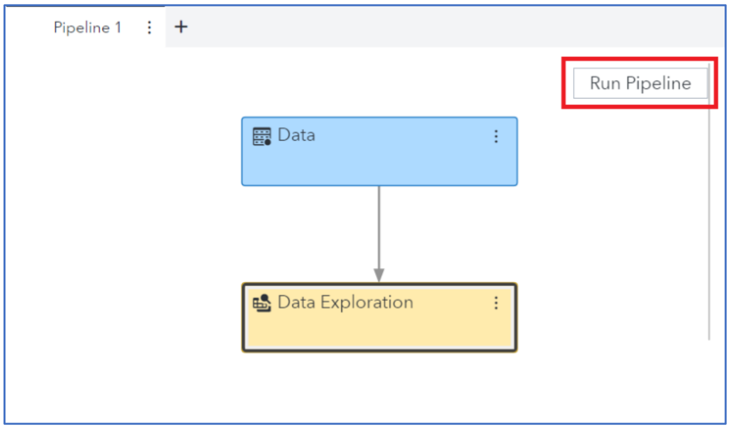
- **Expand the Pipeline 1 | Part 1: Add Data Exploration**
  - Click on the **Pipelines** tab. You should be taken to *Pipeline 1* by default:



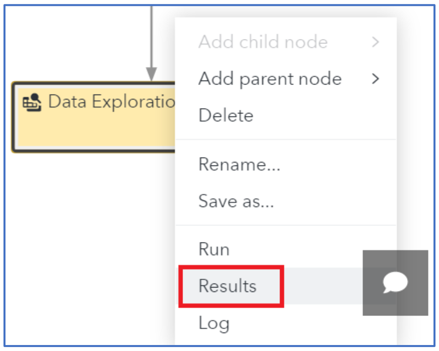
  - Let's start by adding a **Data Exploration** node – so we can see how SAS Model Studio handles exploratory data analysis. Data Exploration is found under the **Miscellaneous** tab. And you can drag-and-drop that node on top of the data node, as follows:



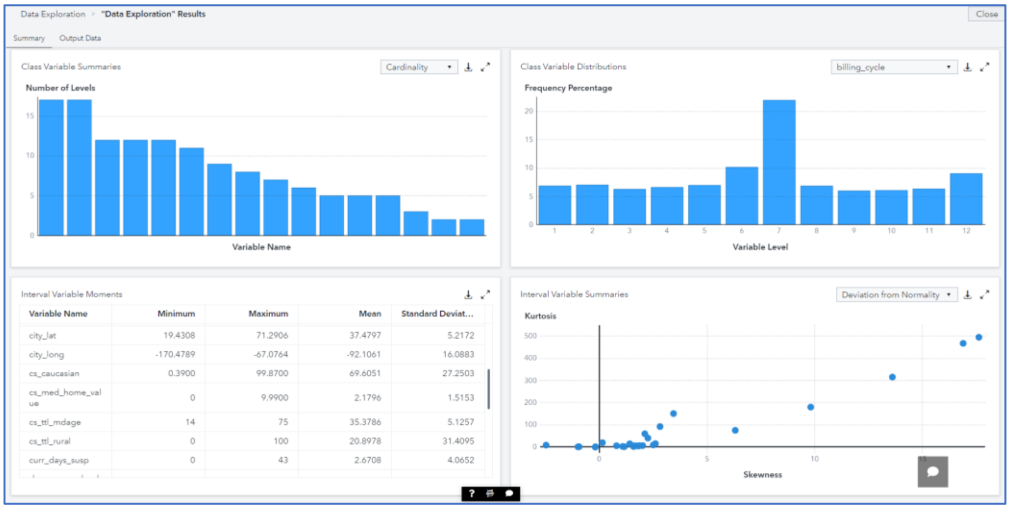
  - Your new pipeline appears as follows. Click on **Run Pipeline** to run the two nodes:



- Let’s examine some output!  Right click on the **Data Exploration** node and then select **Results**:

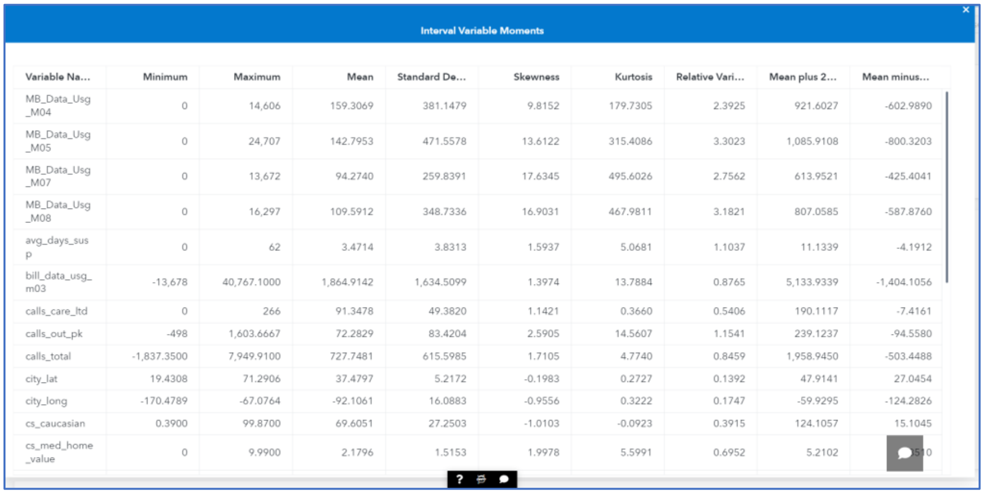


- The following output should appear:



| Variable Name | Minimum | Maximum | Mean | Standard Deviat... |
|---|---|---|---|---|
| city_lat | 19.4308 | 71.2906 | 37.4797 | 5.2172 |
| city_long | -170.4789 | -67.0764 | -92.1061 | 16.0883 |
| cs_caucasian | 0.3900 | 99.8700 | 69.6051 | 27.2503 |
| cs_med_home_value | 0 | 9.9900 | 2.1796 | 1.5153 |
| cs_ttl_mdage | 14 | 75 | 35.3786 | 5.1257 |
| cs_ttl_rural | 0 | 100 | 20.8978 | 31.4095 |
| curr_days_susp | 0 | 43 | 2.6708 | 4.0652 |

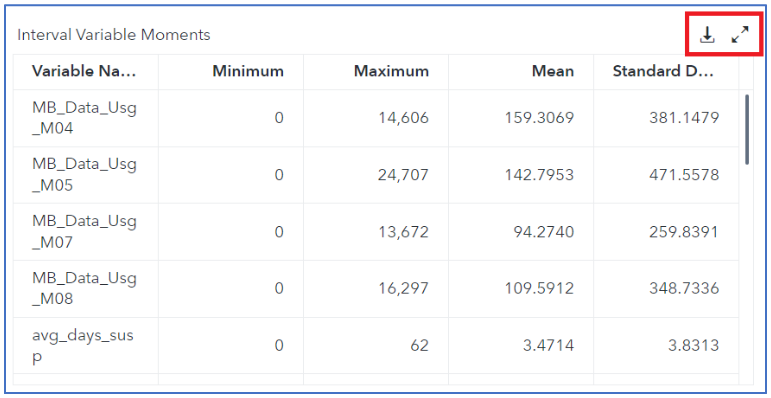
  - Whoa – that’s a lot of statistics. And a fantastic way to get to know your data a bit better.
  - Use the **Expand** button within individual windows to do a deeper dive on the statistics. Here is the **Interval Variable Moments**:



Interval Variable Moments

| Variable Na... | Minimum | Maximum | Mean | Standard De... | Skewness | Kurtosis | Relative Vari... | Mean plus 2... | Mean minus... |
|---|---|---|---|---|---|---|---|---|---|
| MB_Data_Usg_M04 | 0 | 14,606 | 159.3069 | 381.1479 | 9.8152 | 179.7305 | 2.3925 | 921.6027 | -602.9890 |
| MB_Data_Usg_M05 | 0 | 24,707 | 142.7953 | 471.5578 | 13.6122 | 315.4086 | 3.3023 | 1,085.9108 | -800.3203 |
| MB_Data_Usg_M07 | 0 | 13,672 | 94.2740 | 259.8391 | 17.6345 | 495.6026 | 2.7562 | 613.9521 | -425.4041 |
| MB_Data_Usg_M08 | 0 | 16,297 | 109.5912 | 348.7336 | 16.9031 | 467.9811 | 3.1821 | 807.0585 | -587.8760 |
| avg_days_susp | 0 | 62 | 3.4714 | 3.8313 | 1.5937 | 5.0681 | 1.1037 | 11.1339 | -4.1912 |
| bill_data_usg_m03 | -13,678 | 40,767.1000 | 1,864.9142 | 1,634.5099 | 1.3974 | 13.7884 | 0.8765 | 5,133.9339 | -1,404.1056 |
| calls_care_ltd | 0 | 266 | 91.3478 | 49.3820 | 1.1421 | 0.3660 | 0.5406 | 190.1117 | -7.4161 |
| calls_out_pk | -498 | 1,603.6667 | 72.2829 | 83.4204 | 2.5905 | 14.5607 | 1.1541 | 239.1237 | -94.5580 |
| calls_total | -1,837.3500 | 7,949.9100 | 727.7481 | 615.5985 | 1.7105 | 4.7740 | 0.8459 | 1,958.9450 | -503.4488 |
| city_lat | 19.4308 | 71.2906 | 37.4797 | 5.2172 | -0.1983 | 0.2727 | 0.1392 | 47.9141 | 27.0454 |
| city_long | -170.4789 | -67.0764 | -92.1061 | 16.0883 | -0.9556 | 0.3222 | 0.1747 | -59.9295 | -124.2826 |
| cs_caucasian | 0.3900 | 99.8700 | 69.6051 | 27.2503 | -1.0103 | -0.0923 | 0.3915 | 124.1057 | 15.1045 |
| cs_med_home_value | 0 | 9.9900 | 2.1796 | 1.5153 | 1.9978 | 5.5991 | 0.6952 | 5.2102 | [illegible] |

- You can also easily download data tables. The download and expand buttons are located here:

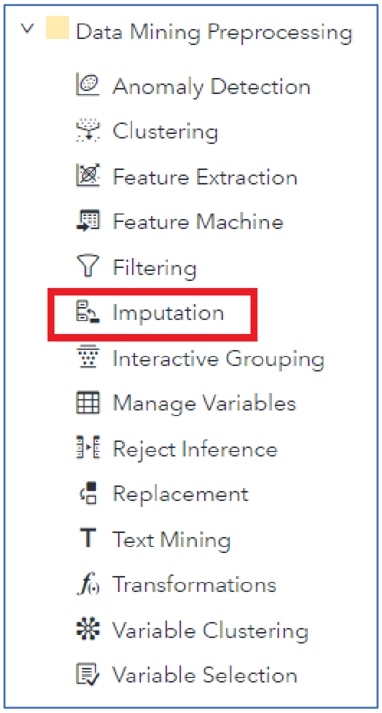


Interval Variable Moments

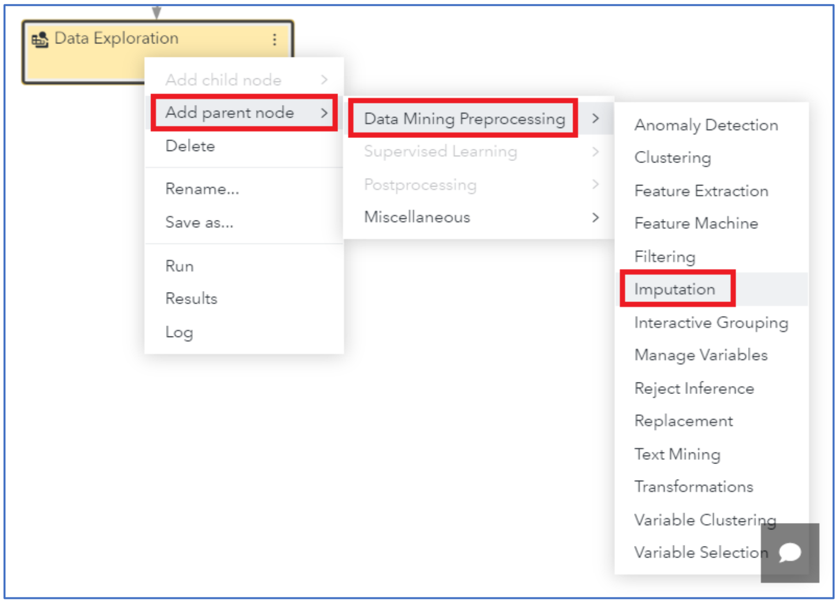
| Variable Na... | Minimum | Maximum | Mean | Standard D... |
|---|---|---|---|---|
| MB_Data_Usg_M04 | 0 | 14,606 | 159.3069 | 381.1479 |
| MB_Data_Usg_M05 | 0 | 24,707 | 142.7953 | 471.5578 |
| MB_Data_Usg_M07 | 0 | 13,672 | 94.2740 | 259.8391 |
| MB_Data_Usg_M08 | 0 | 16,297 | 109.5912 | 348.7336 |
| avg_days_susp | 0 | 62 | 3.4714 | 3.8313 |

- Explore the data until you're satiated! And click **Close** when you're done.

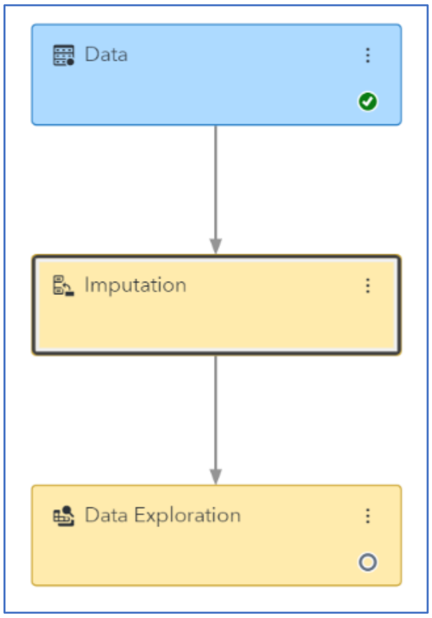
- **Expand the Pipeline 1 | Part 2: Add an Imputation Node**
  - Let's suppose we have some issues with our data. Which, well, we do.
  - The first issue we'd like to address is missing values. This can be done with an **Imputation** node, which is found under **Data Mining Preprocessing**:



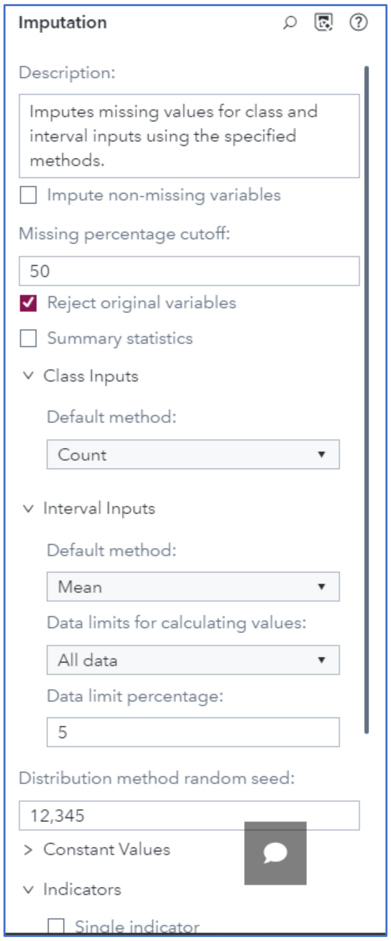
- Let's try a different way to add nodes to the pipeline.  Right click on the **Data Exploration** node. Select **Add parent node > Data Mining Preprocessing > Imputation**:



- Our new pipeline appears as:



- Click on the Imputation node. Open the **Node Options** pane and explore the options available:



**Imputation**

Description:

Imputes missing values for class and interval inputs using the specified methods.

☐ Impute non-missing variables

Missing percentage cutoff:

50

☑ Reject original variables

☐ Summary statistics

∨ Class Inputs

Default method:

Count ▾

∨ Interval Inputs

Default method:

Mean ▾

Data limits for calculating values:

All data ▾

Data limit percentage:

5

Distribution method random seed:

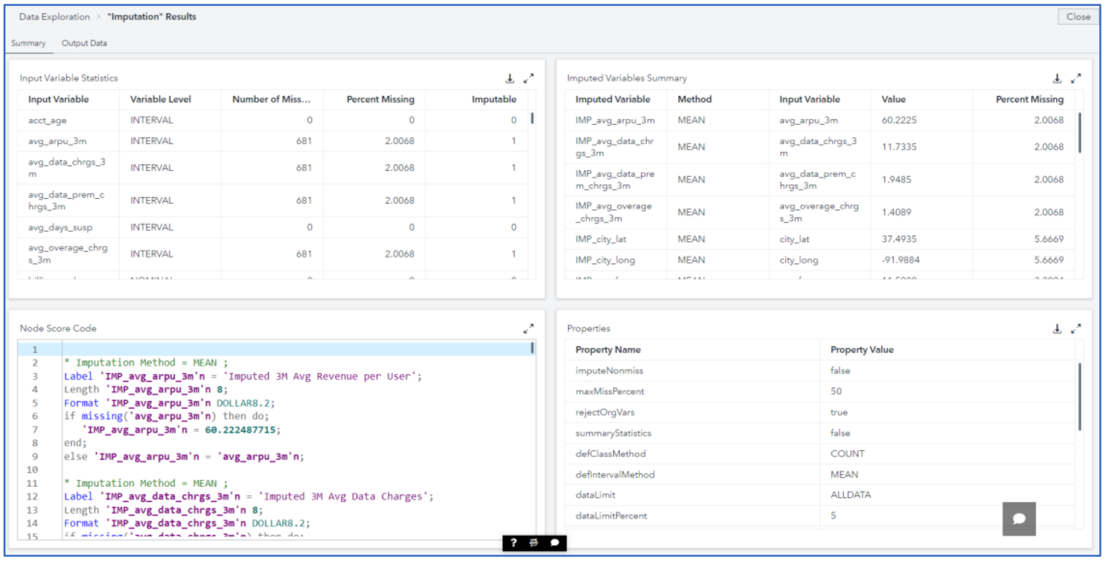
12,345

> Constant Values

∨ Indicators

☐ Single indicator

- Let’s just accept the defaults for now and **Run** our new pipeline. When it’s finished, open the **Results** from the **Imputation** node:



Input Variable Statistics

| Input Variable | Variable Level | Number of Miss... | Percent Missing | Imputable |
|---|---|---|---|---|
| acct_age | INTERVAL | 0 | 0 | 0 |
| avg_arpu_3m | INTERVAL | 681 | 2.0068 | 1 |
| avg_data_chrgs_3m | INTERVAL | 681 | 2.0068 | 1 |
| avg_data_prem_chrgs_3m | INTERVAL | 681 | 2.0068 | 1 |
| avg_days_susp | INTERVAL | 0 | 0 | 0 |
| avg_overage_chrgs_3m | INTERVAL | 681 | 2.0068 | 1 |

Imputed Variables Summary

| Imputed Variable | Method | Input Variable | Value | Percent Missing |
|---|---|---|---|---|
| IMP_avg_arpu_3m | MEAN | avg_arpu_3m | 60.2225 | 2.0068 |
| IMP_avg_data_chrgs_3m | MEAN | avg_data_chrgs_3m | 11.7335 | 2.0068 |
| IMP_avg_data_prem_chrgs_3m | MEAN | avg_data_prem_chrgs_3m | 1.9485 | 2.0068 |
| IMP_avg_overage_chrgs_3m | MEAN | avg_overage_chrgs_3m | 1.4089 | 2.0068 |
| IMP_city_lat | MEAN | city_lat | 37.4935 | 5.6669 |
| IMP_city_long | MEAN | city_long | -91.9884 | 5.6669 |

Node Score Code

```

* Imputation Method = MEAN ;
Label 'IMP_avg_arpu_3m'n = 'Imputed 3M Avg Revenue per User';
Length 'IMP_avg_arpu_3m'n 8;
Format 'IMP_avg_arpu_3m'n DOLLAR8.2;
if missing('avg_arpu_3m'n) then do;
   'IMP_avg_arpu_3m'n = 60.222487715;
end;
else 'IMP_avg_arpu_3m'n = 'avg_arpu_3m'n;

* Imputation Method = MEAN ;
Label 'IMP_avg_data_chrgs_3m'n = 'Imputed 3M Avg Data Charges';
Length 'IMP_avg_data_chrgs_3m'n 8;
Format 'IMP_avg_data_chrgs_3m'n DOLLAR8.2;
if missing('avg_data_chrgs_3m'n) then do;
```

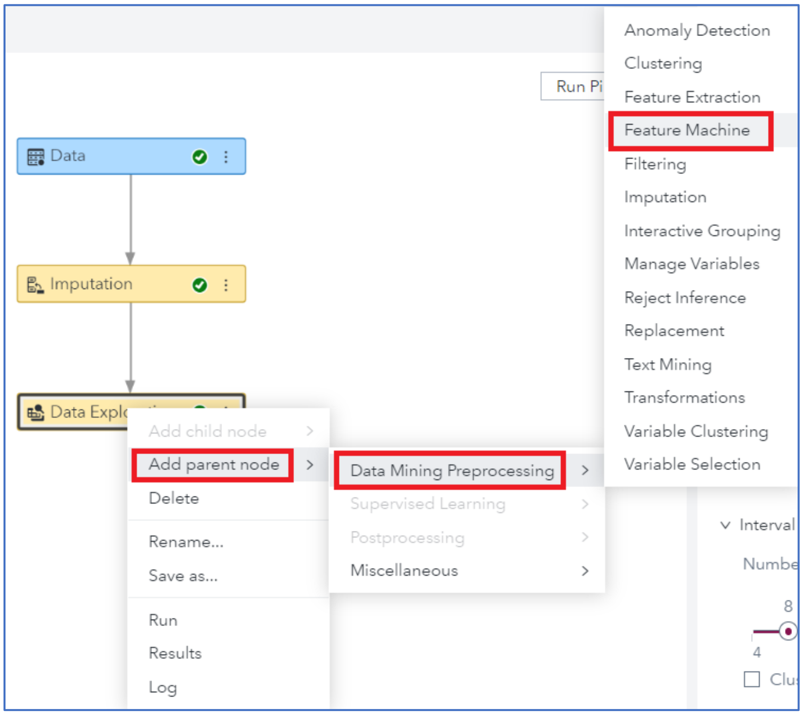
Properties

| Property Name | Property Value |
|---|---|
| imputeNonmiss | false |
| maxMissPercent | 50 |
| rejectOrgVars | true |
| summaryStatistics | false |
| defClassMethod | COUNT |
| defIntervalMethod | MEAN |
| dataLimit | ALLDATA |
| dataLimitPercent | 5 |

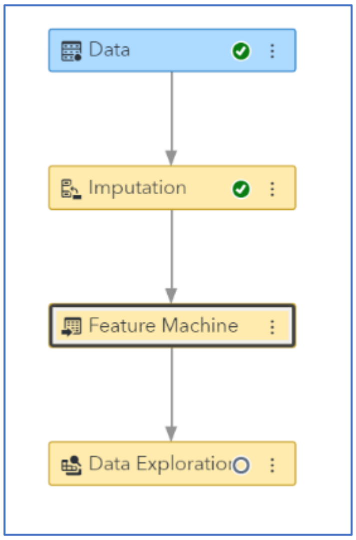
  - The top two windows are likely the most interesting. Expand and explore, separately, the **Input Variable Statistics** and the **Imputed Variables Summary**. We've automatically created several new variables – starting with *IMP_* – that now contain imputed values. For models dependent upon complete-case analysis, we are now able to retain those observations. Hurray!

- **Expand the Pipeline 1 | Part 3: Add a Feature Machine**
  - We've addressed the issue of missing values. But what about skewness and other data transformations? Well, there is an automated tool for that too!
    - It's called a **Feature Machine**
    - Feature what?
  - Fun facts about **Feature Machines**, courtesy of ChatGPT:
    - *Feature engineering involves transforming raw data into a format that is suitable for machine learning algorithms, with the aim of improving the model's predictive performance.*
    - *Some key capabilities of the Feature Machine in SAS Model Studio include:*
      - **Variable Selection***: The Feature Machine provides options to analyze the relevance and importance of variables in the dataset, allowing users to select the most influential features for modeling.*
      - **Variable Creation***: Users can create new variables or derived features based on existing variables using mathematical operations, aggregations, or other transformations.*
      - **Missing Value Handling***: The Feature Machine offers methods for dealing with missing values, such as imputation techniques that estimate missing values based on available information.*
      - **Categorical Variable Encoding***: Categorical variables can be transformed into numeric representations using techniques like one-hot encoding, target encoding, or ordinal encoding.*
      - **Variable Transformation***: Users can apply various transformations to variables, such as log transformations, square roots, scaling, or normalizing, to make them more suitable for modeling.*
      - **Outlier Treatment***: The Feature Machine provides options to detect and handle outliers, allowing users to address extreme values that may affect the model's performance.*

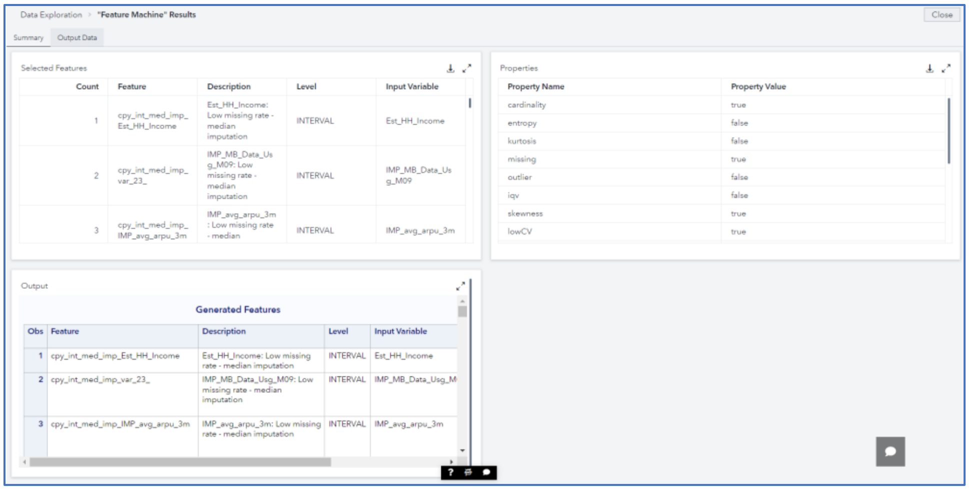
- Let's add the **Feature Machine** between the **Imputation** and **Data Exploration** node. There are many ways to do this, but I'll simply use the approach from last time:



- Double-check that the Feature Machine is here:



- And **Run the Pipeline**. Open the **Results** from the **Feature Machine** node:



Selected Features

| Count | Feature | Description | Level | Input Variable |
|---|---|---|---|---|
| 1 | cpy_int_med_imp_Est_HH_Income | Est_HH_Income: Low missing rate - median imputation | INTERVAL | Est_HH_Income |
| 2 | cpy_int_med_imp_var_23_ | IMP_MB_Data_Usg_M09: Low missing rate - median imputation | INTERVAL | IMP_MB_Data_Usg_M09 |
| 3 | cpy_int_med_imp_IMP_avg_arpu_3m | IMP_avg_arpu_3m : Low missing rate - median | INTERVAL | IMP_avg_arpu_3m |

Properties

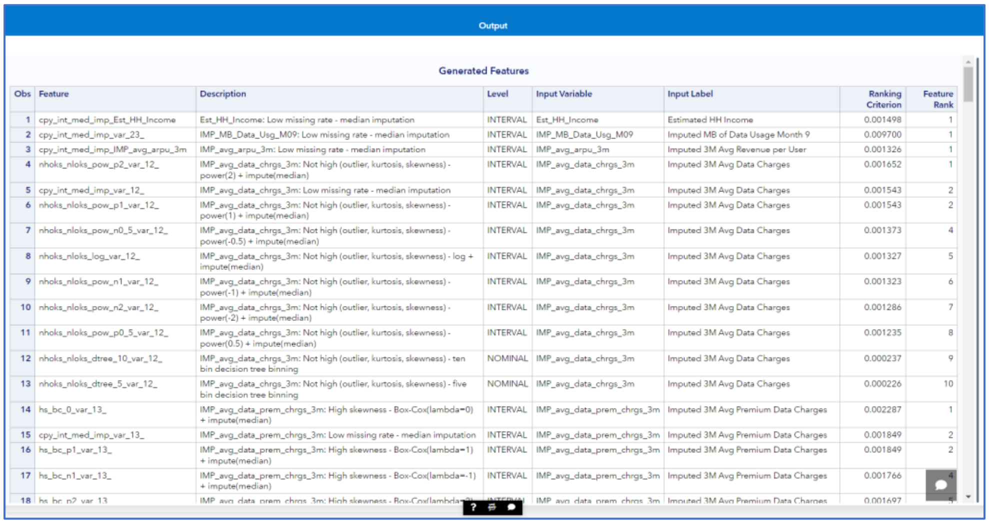
| Property Name | Property Value |
|---|---|
| cardinality | true |
| entropy | false |
| kurtosis | false |
| missing | true |
| outlier | false |
| iqv | false |
| skewness | true |
| lowCV | true |

Output

Generated Features

| Obs | Feature | Description | Level | Input Variable |
|---|---|---|---|---|
| 1 | cpy_int_med_imp_Est_HH_Income | Est_HH_Income: Low missing rate - median imputation | INTERVAL | Est_HH_Income |
| 2 | cpy_int_med_imp_var_23_ | IMP_MB_Data_Usg_M09: Low missing rate - median imputation | INTERVAL | IMP_MB_Data_Usg_M |
| 3 | cpy_int_med_imp_IMP_avg_arpu_3m | IMP_avg_arpu_3m: Low missing rate - median imputation | INTERVAL | IMP_avg_arpu_3m |

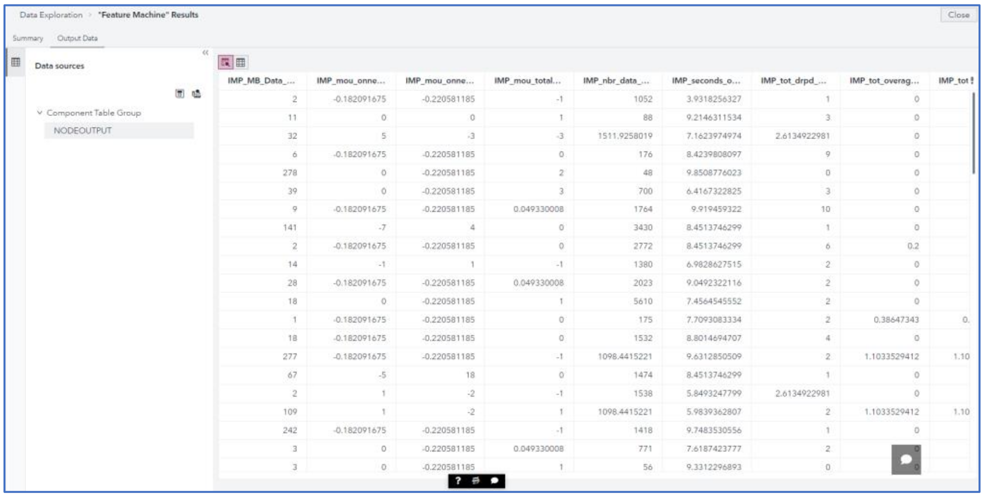
- More good stuff to explore here! Do a deep dive into the **Generated Features**:


Output

Generated Features

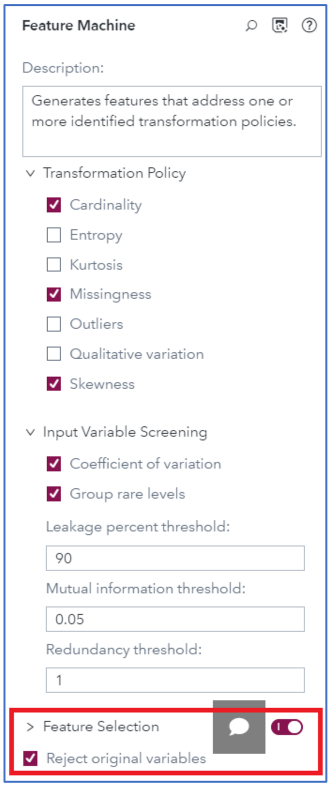
| Obs | Feature | Description | Level | Input Variable | Input Label | Ranking Criterion | Feature Rank |
|---|---|---|---|---|---|---|---|
| 1 | cpy_int_med_imp_Est_HH_Income | Est_HH_Income: Low missing rate - median imputation | INTERVAL | Est_HH_Income | Estimated HH Income | 0.001498 | 1 |
| 2 | cpy_int_med_imp_var_23_ | IMP_MB_Data_Usg_M09: Low missing rate - median imputation | INTERVAL | IMP_MB_Data_Usg_M09 | Imputed MB of Data Usage Month 9 | 0.009700 | 1 |
| 3 | cpy_int_med_imp_IMP_avg_arpu_3m | IMP_avg_arpu_3m: Low missing rate - median imputation | INTERVAL | IMP_avg_arpu_3m | Imputed 3M Avg Revenue per User | 0.001326 | 1 |
| 4 | nhoks_nloks_pow_p2_var_12_ | IMP_avg_data_chrgs_3m: Not high (outlier, kurtosis, skewness) - power(2) + impute(median) | INTERVAL | IMP_avg_data_chrgs_3m | Imputed 3M Avg Data Charges | 0.001652 | 1 |
| 5 | cpy_int_med_imp_var_12_ | IMP_avg_data_chrgs_3m: Low missing rate - median imputation | INTERVAL | IMP_avg_data_chrgs_3m | Imputed 3M Avg Data Charges | 0.001543 | 2 |
| 6 | nhoks_nloks_pow_p1_var_12_ | IMP_avg_data_chrgs_3m: Not high (outlier, kurtosis, skewness) - power(1) + impute(median) | INTERVAL | IMP_avg_data_chrgs_3m | Imputed 3M Avg Data Charges | 0.001543 | 2 |
| 7 | nhoks_nloks_pow_n0_5_var_12_ | IMP_avg_data_chrgs_3m: Not high (outlier, kurtosis, skewness) - power(-0.5) + impute(median) | INTERVAL | IMP_avg_data_chrgs_3m | Imputed 3M Avg Data Charges | 0.001373 | 4 |
| 8 | nhoks_nloks_log_var_12_ | IMP_avg_data_chrgs_3m: Not high (outlier, kurtosis, skewness) - log + impute(median) | INTERVAL | IMP_avg_data_chrgs_3m | Imputed 3M Avg Data Charges | 0.001327 | 5 |
| 9 | nhoks_nloks_pow_n1_var_12_ | IMP_avg_data_chrgs_3m: Not high (outlier, kurtosis, skewness) - power(-1) + impute(median) | INTERVAL | IMP_avg_data_chrgs_3m | Imputed 3M Avg Data Charges | 0.001323 | 6 |
| 10 | nhoks_nloks_pow_n2_var_12_ | IMP_avg_data_chrgs_3m: Not high (outlier, kurtosis, skewness) - power(-2) + impute(median) | INTERVAL | IMP_avg_data_chrgs_3m | Imputed 3M Avg Data Charges | 0.001286 | 7 |
| 11 | nhoks_nloks_pow_p0_5_var_12_ | IMP_avg_data_chrgs_3m: Not high (outlier, kurtosis, skewness) - power(0.5) + impute(median) | INTERVAL | IMP_avg_data_chrgs_3m | Imputed 3M Avg Data Charges | 0.001235 | 8 |
| 12 | nhoks_nloks_dtree_10_var_12_ | IMP_avg_data_chrgs_3m: Not high (outlier, kurtosis, skewness) - ten bin decision tree binning | NOMINAL | IMP_avg_data_chrgs_3m | Imputed 3M Avg Data Charges | 0.000237 | 9 |
| 13 | nhoks_nloks_dtree_5_var_12_ | IMP_avg_data_chrgs_3m: Not high (outlier, kurtosis, skewness) - five bin decision tree binning | NOMINAL | IMP_avg_data_chrgs_3m | Imputed 3M Avg Data Charges | 0.000226 | 10 |
| 14 | hs_bc_0_var_13_ | IMP_avg_data_prem_chrgs_3m: High skewness - Box-Cox(lambda=0) + impute(median) | INTERVAL | IMP_avg_data_prem_chrgs_3m | Imputed 3M Avg Premium Data Charges | 0.002287 | 1 |
| 15 | cpy_int_med_imp_var_13_ | IMP_avg_data_prem_chrgs_3m: Low missing rate - median imputation | INTERVAL | IMP_avg_data_prem_chrgs_3m | Imputed 3M Avg Premium Data Charges | 0.001849 | 2 |
| 16 | hs_bc_p1_var_13_ | IMP_avg_data_prem_chrgs_3m: High skewness - Box-Cox(lambda=1) + impute(median) | INTERVAL | IMP_avg_data_prem_chrgs_3m | Imputed 3M Avg Premium Data Charges | 0.001849 | 2 |
| 17 | hs_bc_n1_var_13_ | IMP_avg_data_prem_chrgs_3m: High skewness - Box-Cox(lambda=-1) + impute(median) | INTERVAL | IMP_avg_data_prem_chrgs_3m | Imputed 3M Avg Premium Data Charges | 0.001766 | 4 |
| 18 | hs_bc_p2_var_13 | IMP_avg_data_prem_chrgs_3m: High skewness - Box-Cox(lambda | INTERVAL | IMP_avg_data_prem_chrgs_3m | Imputed 3M Avg Premium Data Charges | 0.001697 | 5 |

- Under **Output Data**, you can see our newly created variables in action:



| IMP_MB_Data_... | IMP_mou_onne... | IMP_mou_onne... | IMP_mou_total... | IMP_nbr_data_... | IMP_seconds_o... | IMP_tot_drpd_... | IMP_tot_overag... | IMP_tot |
|---|---|---|---|---|---|---|---|---|
| 2 | -0.182091675 | -0.220581185 | -1 | 1052 | 3.9318256327 | 1 | 0 | |
| 11 | 0 | 0 | 1 | 88 | 9.2146311534 | 3 | 0 | |
| 32 | 5 | -3 | -3 | 1511.9258019 | 7.1623974974 | 2.6134922981 | 0 | |
| 6 | -0.182091675 | -0.220581185 | 0 | 176 | 8.4239808097 | 9 | 0 | |
| 278 | 0 | -0.220581185 | 2 | 48 | 9.8508776023 | 0 | 0 | |
| 39 | 0 | -0.220581185 | 3 | 700 | 6.4167322825 | 3 | 0 | |
| 9 | -0.182091675 | -0.220581185 | 0.049330008 | 1764 | 9.919459322 | 10 | 0 | |
| 141 | -7 | 4 | 0 | 3430 | 8.4513746299 | 1 | 0 | |
| 2 | -0.182091675 | -0.220581185 | 0 | 2772 | 8.4513746299 | 6 | 0.2 | |
| 14 | -1 | 1 | -1 | 1380 | 6.9828627515 | 2 | 0 | |
| 28 | -0.182091675 | -0.220581185 | 0.049330008 | 2023 | 9.0492322116 | 2 | 0 | |
| 18 | 0 | -0.220581185 | 1 | 5610 | 7.4564545552 | 2 | 0 | |
| 1 | -0.182091675 | -0.220581185 | 0 | 175 | 7.7093083334 | 2 | 0.38647343 | 0. |
| 18 | -0.182091675 | -0.220581185 | 0 | 1532 | 8.8014694707 | 4 | 0 | |
| 277 | -0.182091675 | -0.220581185 | -1 | 1098.4415221 | 9.6312850509 | 2 | 1.1033529412 | 1.10 |
| 67 | -5 | 18 | 0 | 1474 | 8.4513746299 | 1 | 0 | |
| 2 | 1 | -2 | -1 | 1538 | 5.8493247799 | 2.6134922981 | 0 | |
| 109 | 1 | -2 | 1 | 1098.4415221 | 5.9839362807 | 2 | 1.1033529412 | 1.10 |
| 242 | -0.182091675 | -0.220581185 | -1 | 1418 | 9.7483530556 | 1 | 0 | |
| 3 | 0 | -0.220581185 | 0.049330008 | 771 | 7.6187423777 | 2 | 0 | |
| 3 | 0 | -0.220581185 | 1 | 56 | 9.3312296893 | 0 | 0 | |

- You might think: what happened to our old variables? Well, many of them have been dropped, given this setting:



Feature Machine

Description:

Generates features that address one or more identified transformation policies.

Transformation Policy

☑ Cardinality
☐ Entropy
☐ Kurtosis
☑ Missingness
☐ Outliers
☐ Qualitative variation
☑ Skewness

Input Variable Screening

☑ Coefficient of variation
☑ Group rare levels

Leakage percent threshold: 90

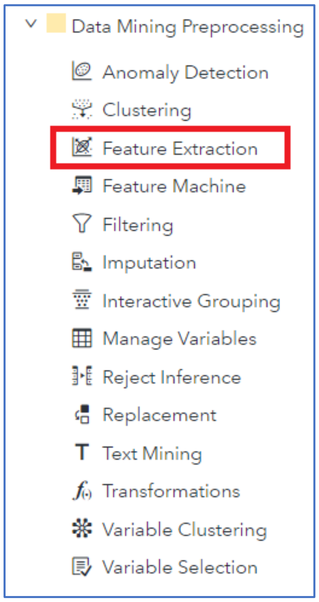
Mutual information threshold: 0.05

Redundancy threshold: 1

Feature Selection

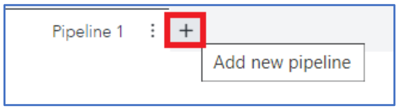
☑ Reject original variables

- However, we could choose to keep the original variables – and let the machine learning models sort them out. We could also use the **Feature Extraction** node, to use statistics to help us choose which variables to keep – and which to drop. **Feature Extraction** is found here:

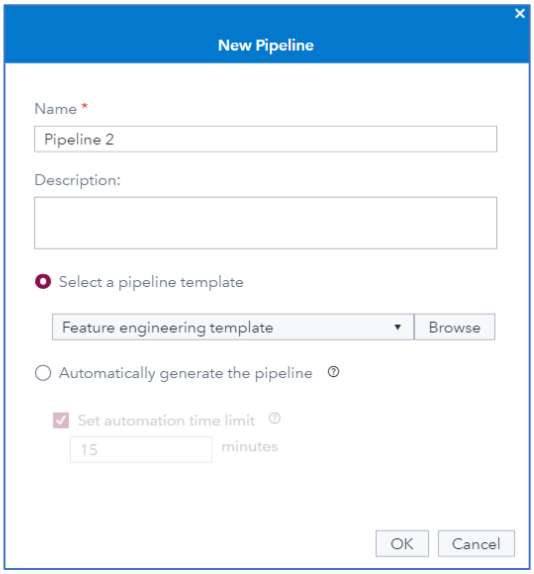


- But we'll save that tool for another time

- **Add a New Pipeline**
  - For the last trick in this section, I'd just like to show how many of the pre-built pipelines in SAS Model Studio already come with Data Mining Preprocessing built in.  So, you don't even have to worry about it…
  - To get started, click the **Add new pipeline** button in SAS Model Studio



  - For Pipeline 2, let's add the **Feature engineering template**, as follows:



New Pipeline

Name *
Pipeline 2

Description:

● Select a pipeline template

Feature engineering template ▾ Browse

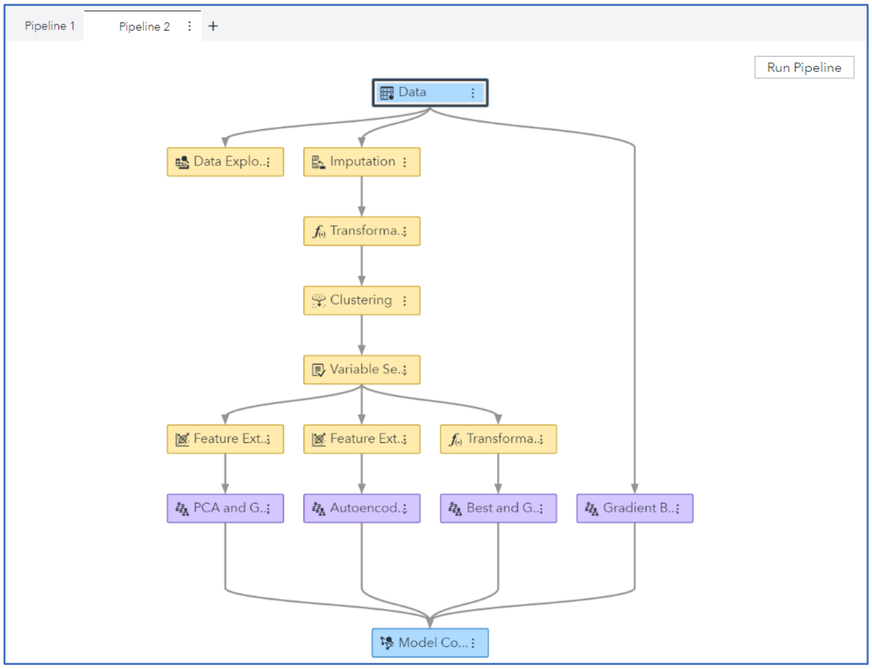
○ Automatically generate the pipeline

☑ Set automation time limit

15 minutes

OK Cancel

  - Pipeline 2 should appear as follows:



*SAS Institute, Inc.*

- What in the what?  That's a proper pipeline.  Just marvel at it for now. Learn all about more advanced modeling pipelines in another workshop.

- **Other resources to help you with SAS Model Studio:**
  - Machine Learning Using SAS Viya 3.5

**Appendix**

- **Data = COMMSDATA**

| • **Name** | **Label** | **Description** |
|---|---|---|
| Churn | Churn | Indicates whether customers churned. |
| Upsell_xsell | Xsell Upsell Flag | Indicates customer's flag for cross-sell or up-sell.<br>(You do not use this variable in this course.) |

***Categorical-valued inputs***

| **Name** | **Label** | **Description** |
|---|---|---|
| credit_class | Credit Class | Credit category for an account or customer. It summarizes the overall credit worthiness of a customer or account. |
| sales_channel | Acquisition Channel | The way in which the consumer was persuaded to purchase company's services. |
| region | Account Region | Customer account region. |
| state | Account State | Customer state location. |
| city | Account City | City designation for customer address. |
| zipcode_primary | Account Code | Primary customer ZIP code. |
| product_plan_desc | Plan Name | Customer's product plan. |
| handset_age_grp | Handset Age Group | Customer's handset age in days. |
| handset | Handset Mfg | Handset manufacturer. Values include *Apple*, *HTC*, *LG*, *Motorola*, *Nokia*, *Samsung*, and *Unknown*. |
| lifestage | Plan Life Stage | Type of contract. |
| rp_pooled_ind | Pooled Rate Plan | Indicates whether customer has pooled rate. |
| call_center | Last Call Center Used | Location of the last call center used. |
| issue_level1 | Call Center Issue Level 1 | Level 1 reason of the call. |
| issue_level2 | Call Center Issue Level 2 | Level 2 reason of the call. |

| Name | Label | Description |
|---|---|---|
| call_category_1 | Call Center Category 1 | Category 1 for the call. |
| call_category_2 | Call Center Category 2 | Category 2 for the call. |
| resolution | Final Resolution | Resolution action taken by call center. |
| verbatims | Survey Verbatim | Feedback from customers via call centers. |

***Interval-valued inputs***

| Name | Label | Description |
|---|---|---|
| lifetime_value | Lifetime Value | Customer's value. |
| avg_arpu_3m | 3M Avg Revenue per User | Average revenue for the past three months. |
| acct_age | Account Tenure | Number of months that the account has been active. |
| billing_cycle | Billing Cycle | Customer's billing cycle (period of the month). |
| nbr_contract_ltd | Total Number Contract lifetime | Number of contracts during life cycle. |
| rfm_score | Account Ranking (RFM Score) | Customer's account score. |
| Est_HH_Income | Estimated HH Income | Household income. |
| region_lat | Account Region Latitude | Customer region latitude. |
| region_long | Account Region Longitude | Customer region longitude. |
| state_lat | Account State Latitude | State latitude. |
| state_long | Account State Longitude | State longitude. |
| city_lat | Account City Latitude | Customer city latitude. |
| city_long | Account City Longitude | Customer city longitude. |

| Name | Label | Description |
|---|---|---|
| zip_lat | Account ZIP Code Latitude | ZIP code latitude. |
| zip_long | Account ZIP Code Longitude | ZIP code longitude. |
| cs_med_home_value | Census Area Median Home Value Index | Median home value in customer’s area. |
| cs_pct_home_owner | Census Area Percent Home Owner | Percentage home owner in customer’s area. |
| cs_ttl_pop | Census Area Total Population | Population in customer’s area. |
| cs_hispanic | Census Area Hispanic | Hispanic population in customer’s area. |
| cs_caucasian | Census Area Caucasian | Caucasian population in customer’s area. |
| cs_afr_amer | Census Area African-American | African-American population in customer’s area. |
| cs_other | Census Area Other | Other population in customer’s area. |
| cs_ttl_urban | Census Area Total Urban | Urban population in customer’s area. |
| cs_ttl_rural | Census Area Total Rural | Rural population in customer’s area. |
| cs_ttl_male | Census Area Total Males | Male population in customer’s area. |
| cs_ttl_female | Census Area Total Females | Female population in customer’s area. |
| cs_ttl_hhlds | Census Area Total Households | Households in customer’s area. |
| cs_ttl_mdage | Census Area Median Age | Median age in customer’s area. |
| mb_inclplan | Plan Data MB | MB included in the plan. |
| ever_days_over_plan | Total Days Over Plan | Total days over the plan. |
| ever_times_over_plan | Total Times Over Plan | Total times over the plan. |

| Name | Label | Description |
|---|---|---|
| data_device_age | Avg Age of Devices on Plan | Average age of devices on the plan. |
| equip_age | Handset Age | Age of equipment history, whether mobile device, smartphone, or another handset type. |
| mfg_apple | Own Apple | Apple manufactured device. 1 is *Yes*, 0 means *No*. |
| mfg_samsung | Own Samsung | Samsung manufactured device. 1 is *Yes*, 0 means *No*. |
| mfg_htc | Own HTC | HTC manufactured device. 1 is *Yes*, 0 means *No*. |
| mfg_motorola | Own Motorola | Motorola manufactured device. 1 is *Yes*, 0 means *No*. |
| mfg_lg | Own LG | LG manufactured device. 1 is *Yes*, 0 means *No*. |
| mfg_nokia | Own Nokia | Nokia manufactured device. 1 is *Yes*, 0 means *No*. |
| delinq_indicator | Delinquent Indicator | Delinquency indicator. Scale varies from -2 to +4, depending on customer history. |
| times_delinq | Consecutive Mths Delinquent | Consecutive months in default. |
| count_of_suspensions_6m | Times Suspended Last 6M | Times suspended in the past six months. |
| avg_days_susp | Days Suspended Last 6M | Days suspended in the past six months. |
| calls_total | Total Calls Curr | Current number of calls. |
| calls_in_pk | Calls Incoming Peak | Number of calls received in peak time. |
| calls_in_offpk | Calls Incoming Off-Peak | Number of call received off peak time. |
| calls_out_offpk | Calls Outgoing Off-Peak | Number of calls made in peak time. |

| Name | Label | Description |
|---|---|---|
| calls_out_pk | Calls Outgoing Peak | Number of calls made off peak time. |
| mou_total_pct_MOM | Minutes Total Pct Change Month over Month | Percentage of minutes change month over month. |
| mou_onnet_pct_MOM | Minutes on Network Pct Change Month over Month | Percentage of minutes on network change month over month. |
| mou_roam_pct_MOM | Minutes Roaming Pct Change Month over Month | Percentage of minutes on roaming change month over month. |
| mou_onnet_6m_normal | 6M Avg Minutes on Network Normally Distributed | Minutes of use on network over six months normally distributed. |
| mou_roam_6m_normal | 6M Avg Minutes Roaming Normally Distributed | Minutes of use in roaming over six months normally distributed. |
| voice_total_bill_mou_curr | Total Voice Billed Minutes of Use | Current minutes of voice billed. |
| tot_voice_chrgs_curr | Total Voice Charges | Current minutes of voice charged. |
| tot_drpd_pr1 | Number of Dropped Calls 1 Mth Prior | Number of dropped calls on the previous month. |
| bill_data_usg_m03 | 3M Avg Billed Data Usage | Average data billed over the past three months. |
| bill_data_usg_m06 | 6M Avg Billed Data Usage | Average data billed over the past six months. |
| bill_data_usg_m09 | 9M Avg Billed Data Usage | Average data billed over the past nine months. |
| mb_data_usg_m01 | MB Data Usage 1 Mth Prior | MB data used on the previous month. |
| mb_data_usg_m02 | MB Data Usage 2 Mths Prior | MB data used prior two months. |

| Name | Label | Description |
|---|---|---|
| mb_data_usg_m03 | MB Data Usage 3 Mths Prior | MB data used prior three months. |
| mb_data_ndist_mo6m | 6M Avg Billed Data Usage Normally Distributed | Data used on network over six months normally distributed. |
| mb_data_usg_roamm01 | MB Data Usage Roam 1 Mth Prior | Data used in roaming in the previous month. |
| mb_data_usg_ roamm02 | MB Data Usage Roam 2 Mths Prior | Data used in roaming prior two months. |
| mb_data_usg_ roamm03 | MB Data Usage Roam 3 Mths Prior | Data used in roaming prior three months. |
| data_usage_amt | Data Usage Amount | Total data usage amount over last month. |
| tweedie_adjusted | Data Usage Amt Tweedie Distributed | Data used in Twitter. |
| tot_mb_data_curr | Total MB of Data Usage | Current MB data used. |
| tot_mb_data_roam_curr | Total MB of Roam Data Usage | Current MB data used in roaming. |
| bill_data_usg_total | Total Billed Data usage | Total billed data. |
| tot_overage_chgs | Total Overage Charges | Total overage charged. |
| data_prem_chrgs_curr | Premium Data Charges | Premium data charged. |
| nbr_data_cdrs | Number of Data Records | Number of call detail records. |
| avg_data_chrgs_3m | 3M Avg Data Charges | Average data charged in the past three months. |
| avg_data_prem_chrgs_3m | 3M Avg Premium Data Charges | Average premium data charged in the past three months. |
| avg_overage_chrgs_3m | 3M Avg Overage Charges | Average overage data charged in the past three months. |

| Name | Label | Description |
|---|---|---|
| nbr_contacts | Number Times Customer Contacted | Number of contacts customer made to the company. |
| calls_TS_acct | Number Calls Tech Support | Number of tech support calls. |
| open_tsupcomplnts | Open Tech Support Complaints | Number of tech support complains opened. |
| num_tsupcomplnts | Tech Support Complaints - LTD | Number of tech support complains. |
| unsolv_tsupcomplnts | Unresolved Tech Support Complaints - LTD | Number of unsolved tech support complaints. |
| wrk_orders | Open Work Orders | Number of open work. |
| days_openwrkorders | Days of Open Work Orders | Days of open work. |
| resolved_complnts | Resolved Complaints | Number of complaints resolved. |
| calls_care_acct | Number Calls Care Center | Call center care account assignment, which takes values between 0-9. |
| calls_care_3mavg_acct | Number Calls Care Center 3 Month Avg | Call center care account score over past three months averaged. |
| calls_care_6mavg_acct | Number Calls Care Center 6 Month Avg | Call center care account score over past six months averaged. |
| res_calls_3mavg_acct | Resolved Calls – 3Mo Average | Average number of resolved customer service calls over past three months for the customer account. |
| res_calls_6mavg_acct | Resolved Calls – 6Mo Average | Average number of resolved customer service calls over past six months for the customer account. |
| last_rep_sat_score | Last Call Satisfaction Rating Given | Latest customer service representative satisfaction score (given by past customers). |
| network_mention | Network Issues Discussed | Number of network issues discussed. |

| Name | Label | Description |
|---|---|---|
| service_mention | Service Issues Discussed | Number of service issues discussed. |
| price_mention | Price Issues Discussed | Number of prices issues discussed. |
| times_susp | Number of Times Suspended | Number of times suspended. |
| curr_days_susp | Number of Days Suspended | Number of days suspended. |
| pymts_late_ltd | Total Late Payments Lifetime | Number of late payments. |
| calls_care_ltd | Total Calls to Care Lifetime | Number of calls to call center. |
| MB_Data_Usg_M04 | MB of Data Usage Month 4 | MB data used in past four months. |
| MB_Data_Usg_M05 | MB of Data Usage Month 5 | MB data used in past five months. |
| MB_Data_Usg_M06 | MB of Data Usage Month 6 | MB data used in past six months. |
| MB_Data_Usg_M07 | MB of Data Usage Month 7 | MB data used in past seven months. |
| MB_Data_Usg_M08 | MB of Data Usage Month 8 | MB data used in past eight months. |
| MB_Data_Usg_M09 | MB of Data Usage Month 9 | MB data used in past nine months. |
| seconds_of_data_norm | Seconds of Data - Normalized | Number of seconds of data normalized. |
| seconds_of_data_log | Seconds of Data - Natural Log | Number of seconds of data transformed by log. |