SAS® GLOBAL FORUM 2020 #SASGF

Paper SAS4617-2020

Smarter and Faster Self-Service Data Preparation

Atrin Assa, SAS Institute Inc.

ABSTRACT

In this paper, you'll learn about the latest and greatest self-service data preparation capabilities of SAS® Visual Analytics. You will understand how smart suggestions can help you improve the quality of your data, how the new interface can help you work faster, and how better-prepared data can help you build better visualizations, better reports, and tell a more compelling data story.

INTRODUCTION

Everything in the world of analytics lives at the mercy of our ability to turn data into something that we can use. It's not just a challenge for the data scientist and the IT admin. It's a hurdle for everyone who has data and wants to glean insights from it.

With modern techniques, SAS Visual Analytics can help you improve your data more quickly. Smart algorithms woven throughout the experience can analyze your data and help improve your data with minimal effort. These capabilities are available throughout the SAS® experience, particularly as part of working with data and discovery.

Automatic profiles can help you get a quick understanding of the structure and quality of your data. Data prep suggestions can assess your data and suggest transformations that will make it ready for analysis. Finally, AI-powered exploration techniques will automatically prep your data to help you get the best results.

THREE WAYS TO SMARTER DATA PREP

There are three ways to make your data prep smarter: data profile, data prep suggestions, and automated data prep. The first of these, the data profile, helps you understand the structure of your data. This is the cornerstone of useful data analytics. No amount of automation will ever be a substitute for a robust understanding of the data that an analyst can bring.

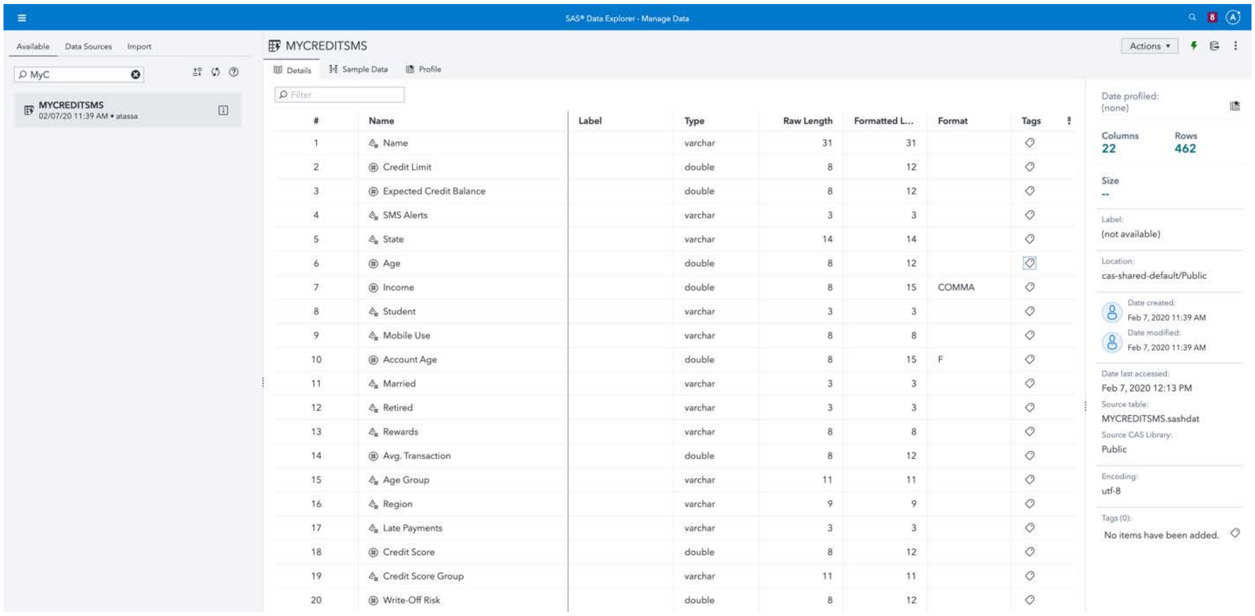
Once you understand your data, data prep suggestions can help you quickly identify areas to improve within your data. These suggestions are presented based on machine learning methods that analyze your data and determine ways you can improve the data for analysis. You don't have to take any of the suggestions, but if you're struggling to figure out how to improve the data set, suggestions can help you get started quickly.

No matter how well you understand the data or how well you prepare your data, there will always be situations where some additional data prep can improve the results of your analysis. Individual discovery techniques might require data to be in a particular arrangement that you might not have predicted at the outset when preparing the data. In those cases, SAS Visual Analytics automatically does the necessary prep for you without any intervention required.

USING THE DATA PROFILE

When you work with any table in SAS® Viya®, you'll always be able to see three tabs from the Data Explorer interface. These tabs are Details, Sample Data, and Profile. If you want to understand the structure of and content in your data, it's valuable to study these tabs.

The Details tab gives you the broadest overview of your data. It shows you the column, whether those columns are characters or numeric, and the size of your data. Size includes information like the number of rows in your table and the number of columns in your table. These values are important for working with your data later on. The more columns you have, the more variables you can analyze in your data, but also the greater the computational cost of analysis. All else being equal, data with more variables tend to be more challenging to analyze than data with more rows. Imagine you had a table with only two variables, height and weight, recorded for five million humans. There's plenty of data, but it would be easy to visualize and easy to analyze. Plot height and weight together and see if there is any pattern.



#	Name	Label	Type	Raw Length	Formatted L...	Format	Tags
1	Name		varchar	31	31		
2	Credit Limit		double	8	12		
3	Expected Credit Balance		double	8	12		
4	SMS Alerts		varchar	3	3		
5	State		varchar	14	14		
6	Age		double	8	12		
7	Income		double	8	15	COMMA	
8	Student		varchar	3	3		
9	Mobile Use		varchar	8	8		
10	Account Age		double	8	15	F	
11	Married		varchar	3	3		
12	Retired		varchar	3	3		
13	Rewards		varchar	8	8		
14	Avg. Transaction		double	8	12		
15	Age Group		varchar	11	11		
16	Region		varchar	9	9		
17	Late Payments		varchar	3	3		
18	Credit Score		double	8	12		
19	Credit Score Group		varchar	11	11		
20	Write-Off Risk		double	8	12		

Date profiled:
(none)

Columns 22
Rows 462

Size
--

Label:
(not available)

Location:
cas-shared-default/Public

Date created:
Feb 7, 2020 11:39 AM

Date modified:
Feb 7, 2020 11:39 AM

Date last accessed:
Feb 7, 2020 12:13 PM

Source table:
MYCREDITSMS.sashdat

Source CAS Library:
Public

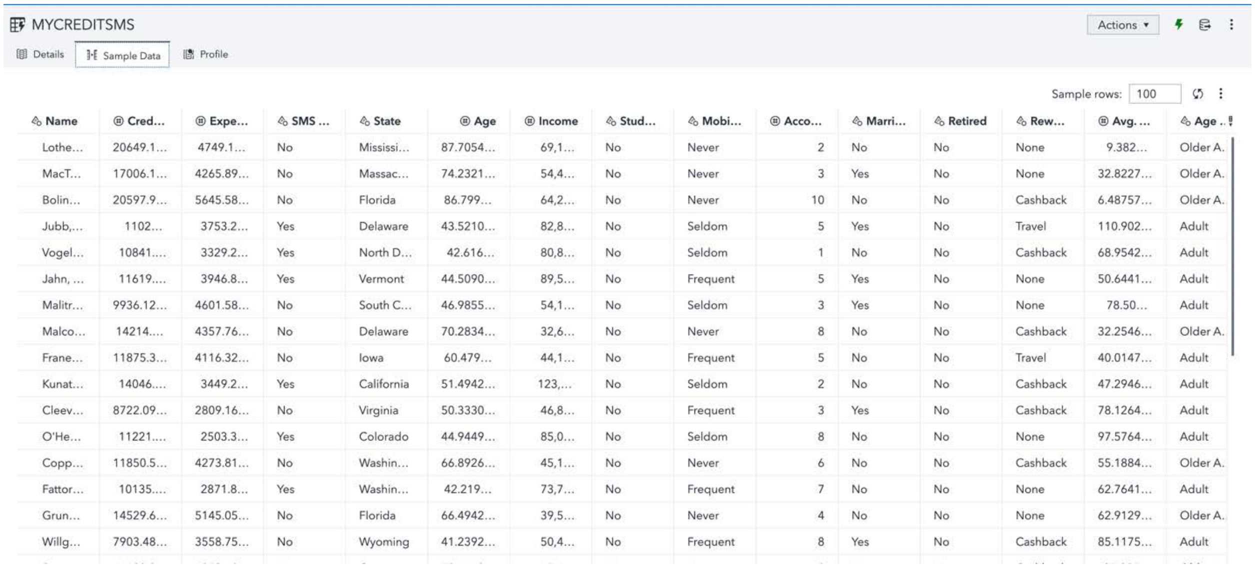
Encoding:
utf-8

Tags (0):
No items have been added.

Display 1. The Details Pane Provides You an Overview of the Data

On the other hand, imagine you have 2,000 variables recorded for about just 20 humans. Your table is smaller, you might not have as much data, but which two variables are you going to plot together? Which ones are important? It's an altogether more challenging task.

The Sample Data tab shows you a sample of your table so that you can understand its structure and the kind of data that it might contain. Through the sample, you might even be able to spot problem areas that you'll need to clean up.



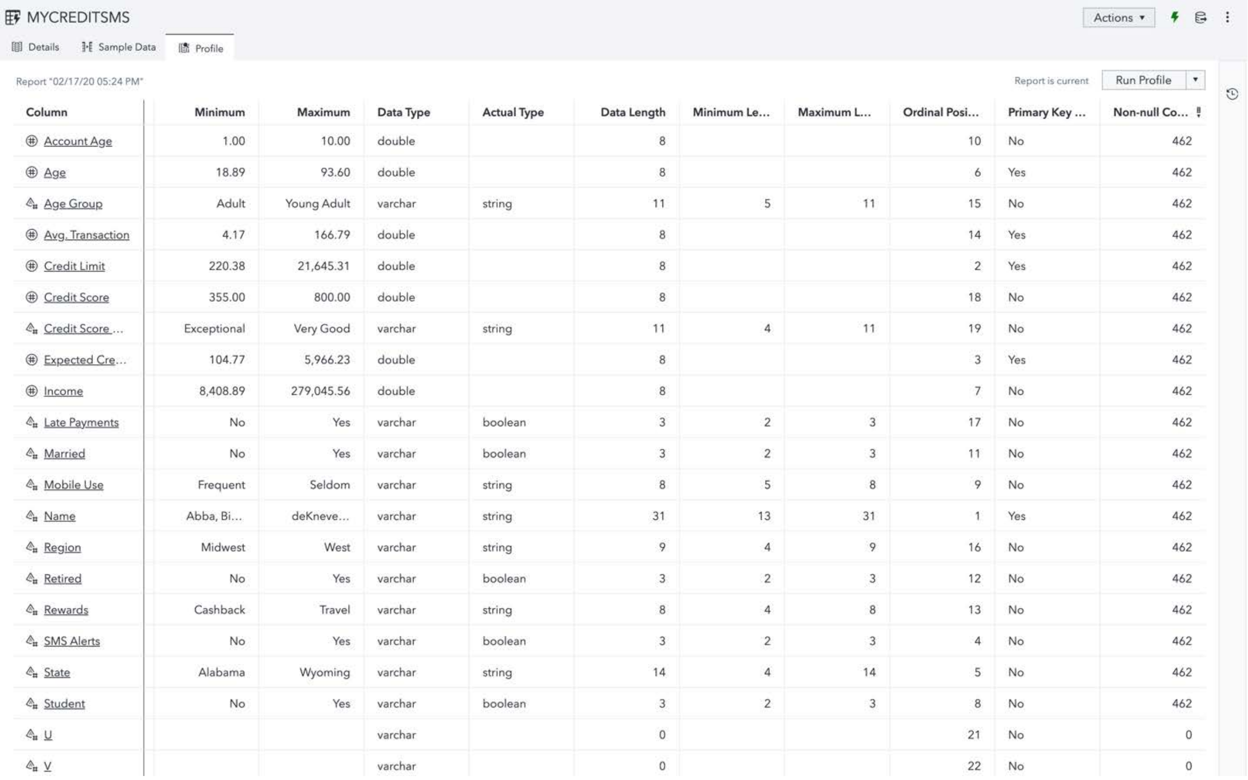
Name	Cred...	Expe...	SMS ...	State	Age	Income	Stud...	Mobi...	Acco...	Marri...	Retired	Rew...	Avg. ...	Age ..
Lothe...	20649.1...	4749.1...	No	Mississi...	87.7054...	69,1...	No	Never	2	No	No	None	9.382...	Older A.
MacT...	17006.1...	4265.89...	No	Massac...	74.2321...	54,4...	No	Never	3	Yes	No	None	32.8227...	Older A.
Bolin...	20597.9...	5645.58...	No	Florida	86.799...	64,2...	No	Never	10	No	No	Cashback	6.48757...	Older A.
Jubb,...	1102...	3753.2...	Yes	Delaware	43.5210...	82,8...	No	Seldom	5	Yes	No	Travel	110.902...	Adult
Vogel...	10841....	3329.2...	Yes	North D...	42.616...	80,8...	No	Seldom	1	No	No	Cashback	68.9542...	Adult
Jahn, ...	11619....	3946.8...	Yes	Vermont	44.5090...	89,5...	No	Frequent	5	Yes	No	None	50.6441...	Adult
Malitr...	9936.12...	4601.58...	No	South C...	46.9855...	54,1...	No	Seldom	3	Yes	No	None	78.50...	Adult
Malco...	14214....	4357.76...	No	Delaware	70.2834...	32,6...	No	Never	8	No	No	Cashback	32.2546...	Older A.
Frane...	11875.3...	4116.32...	No	Iowa	60.479...	44,1...	No	Frequent	5	No	No	Travel	40.0147...	Adult
Kunat...	14046....	3449.2...	Yes	California	51.4942...	123,...	No	Seldom	2	No	No	Cashback	47.2946...	Adult
Cleev...	8722.09...	2809.16...	No	Virginia	50.3330...	46,8...	No	Frequent	3	Yes	No	Cashback	78.1264...	Adult
O'He...	11221....	2503.3...	Yes	Colorado	44.9449...	85,0...	No	Seldom	8	No	No	None	97.5764...	Adult
Copp...	11850.5...	4273.81...	No	Washin...	66.8926...	45,1...	No	Never	6	No	No	Cashback	55.1884...	Older A.
Fattor...	10135....	2871.8...	Yes	Washin...	42.219...	73,7...	No	Frequent	7	No	No	None	62.7641...	Adult
Grun...	14529.6...	5145.05...	No	Florida	66.4942...	39,5...	No	Never	4	No	No	None	62.9129...	Older A.
Willg...	7903.48...	3558.75...	No	Wyoming	41.2392...	50,4...	No	Frequent	8	Yes	No	Cashback	85.1175...	Adult

Display 2. The Sample Data Tab Lets You See Your Data

Finally, the Profile tab can give you a deeper understanding of the structure of your idea and potential problem areas. It can show you many pieces of information including the following:

- the number of unique values in each column
- the number of missing values in each column
- descriptive statistics for numeric columns
- minimum and maximum lengths for strings
- and more

This information can be crucial for how you prepare the data. For example, you might want to deal with columns that have a high number of missing values.



Report "02/17/20 05:24 PM"

Report is current

Column	Minimum	Maximum	Data Type	Actual Type	Data Length	Minimum Le...	Maximum L...	Ordinal Posi...	Primary Key ...	Non-null Co...
Account Age	1.00	10.00	double		8			10	No	462
Age	18.89	93.60	double		8			6	Yes	462
Age Group	Adult	Young Adult	varchar	string	11	5	11	15	No	462
Avg. Transaction	4.17	166.79	double		8			14	Yes	462
Credit Limit	220.38	21,645.31	double		8			2	Yes	462
Credit Score	355.00	800.00	double		8			18	No	462
Credit Score ...	Exceptional	Very Good	varchar	string	11	4	11	19	No	462
Expected Cre...	104.77	5,966.23	double		8			3	Yes	462
Income	8,408.89	279,045.56	double		8			7	No	462
Late Payments	No	Yes	varchar	boolean	3	2	3	17	No	462
Married	No	Yes	varchar	boolean	3	2	3	11	No	462
Mobile Use	Frequent	Seldom	varchar	string	8	5	8	9	No	462
Name	Abba, Bi...	deKneve...	varchar	string	31	13	31	1	Yes	462
Region	Midwest	West	varchar	string	9	4	9	16	No	462
Retired	No	Yes	varchar	boolean	3	2	3	12	No	462
Rewards	Cashback	Travel	varchar	string	8	4	8	13	No	462
SMS Alerts	No	Yes	varchar	boolean	3	2	3	4	No	462
State	Alabama	Wyoming	varchar	string	14	4	14	5	No	462
Student	No	Yes	varchar	boolean	3	2	3	8	No	462
U			varchar		0			21	No	0
V			varchar		0			22	No	0

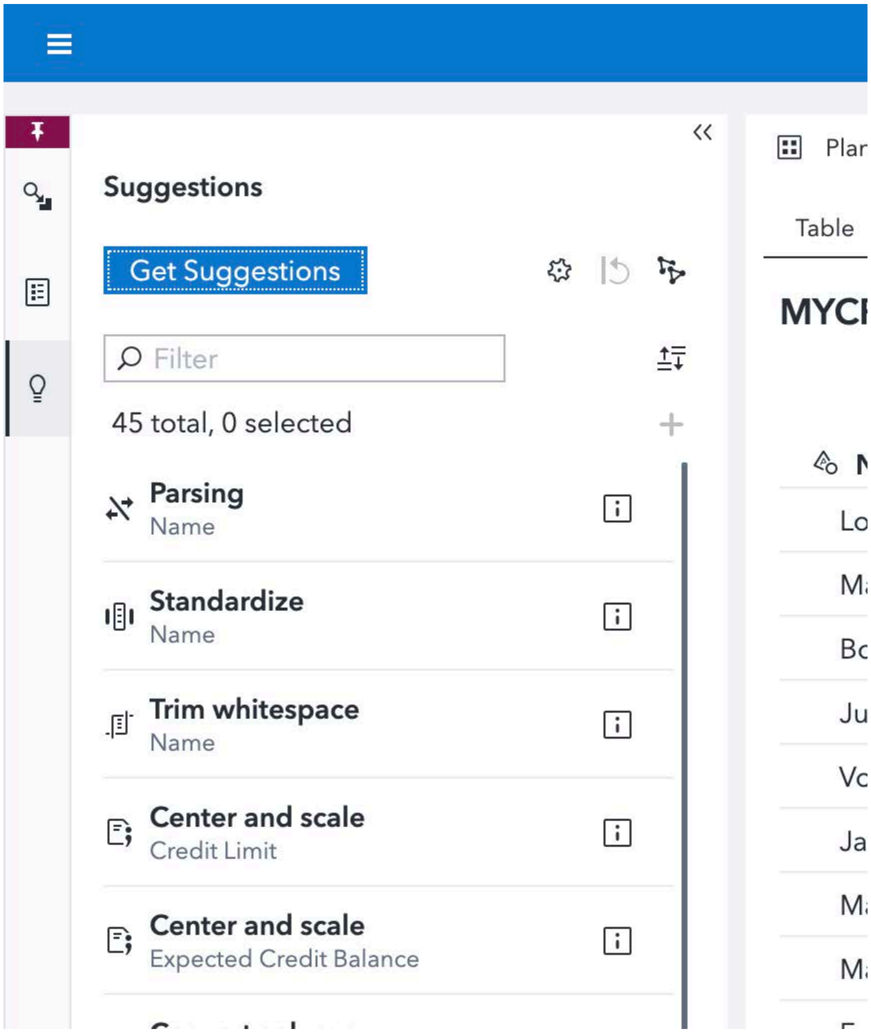
Display 3. The Profile Tab Shows You Important Information About Your Table

Understanding these simple statistics about your table and connecting them with how users will consume the data can help you improve the preparation of your data.

DATA PREP SUGGESTIONS

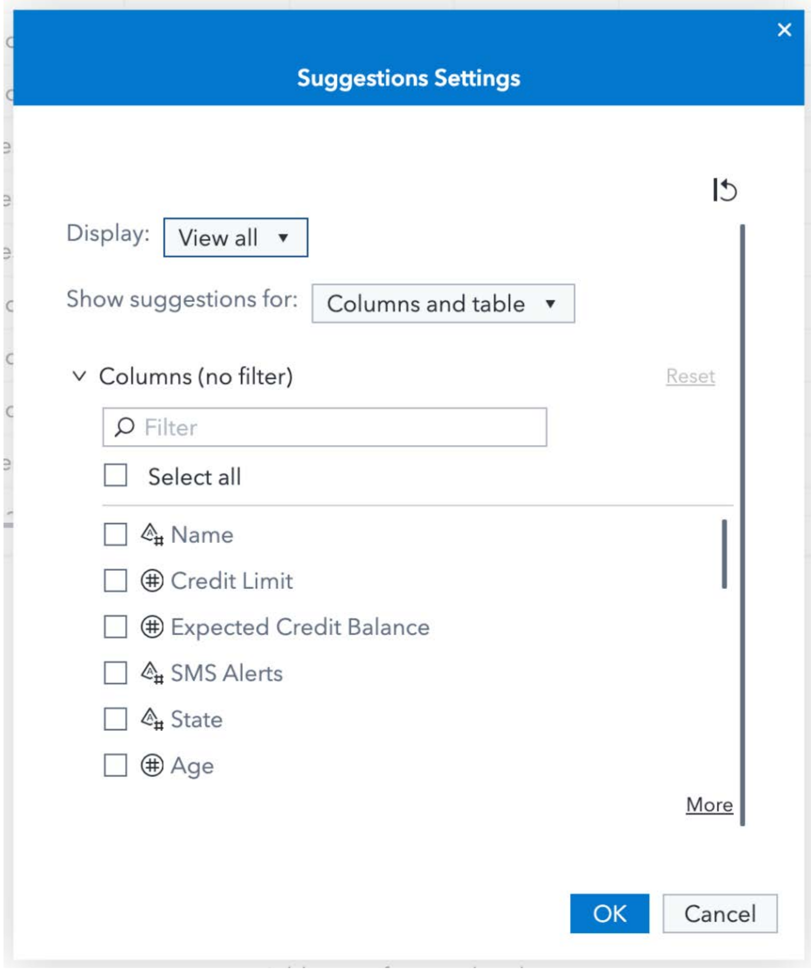
Once you're ready to prepare your data, SAS can help speed you up with data prep suggestions. The suggestions feature machine learning to analyze your data and suggest transforms and actions that you can add to your SAS® Data Studio plans. Suggestions analyze your data with models that have been registered in your SAS environment.

While preparing data, click the lightbulb icon on the left side of the screen to open the Suggestions pane and then click the blue Get Suggestions button.



Display 4. The Get Suggestions Button in the Suggestions Pane

You can set up your suggestions to only provide suggestions that apply to the whole table or particular columns. You can even choose the columns that you'd like to get suggestions for by using the suggestion settings (the gear icon in the Suggestions pane).



Display: View all ▾

Show suggestions for: Columns and table ▾

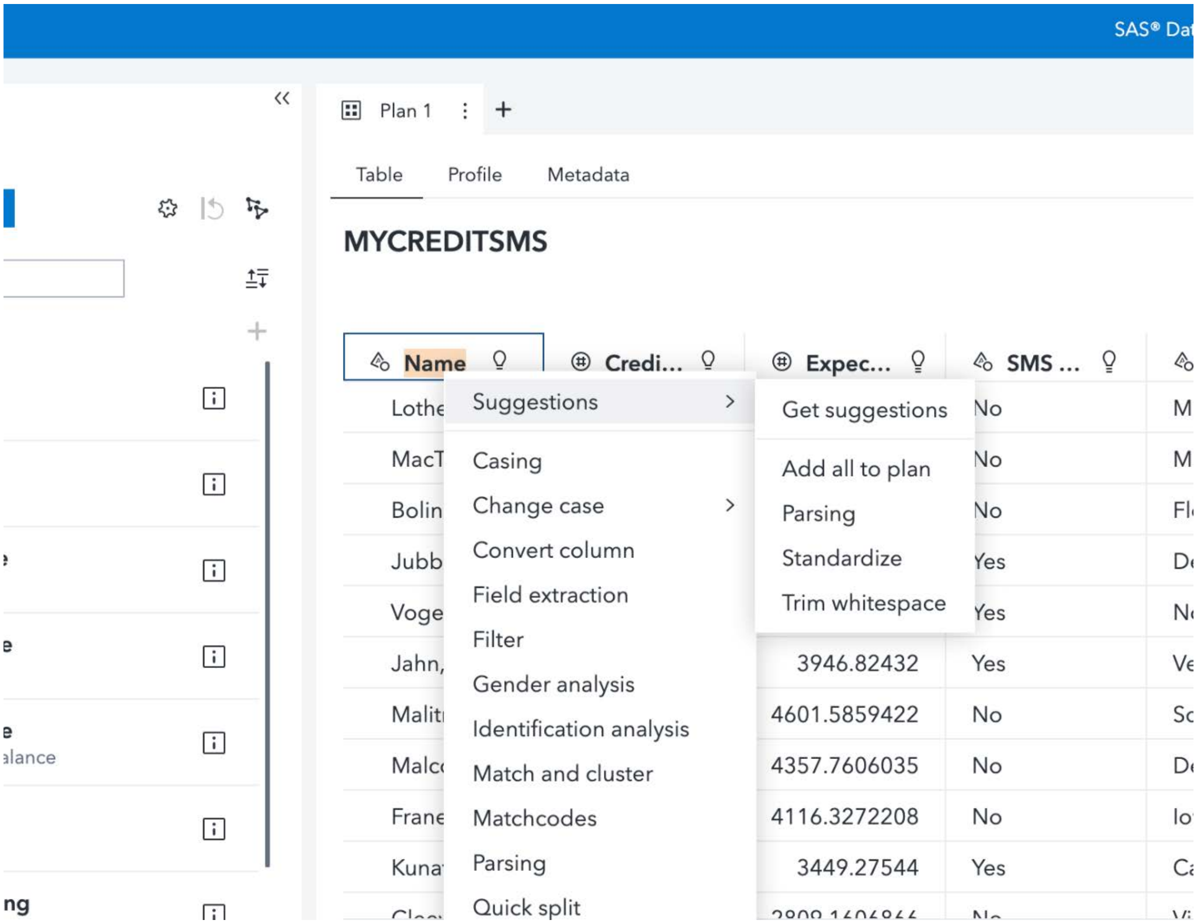
Columns (no filter)

☐ Select all

☐ Name
☐ Credit Limit
☐ Expected Credit Balance
☐ SMS Alerts
☐ State
☐ Age

Display 5. Suggestions Settings Let You Decide How You’d Like Your Data Preparation Suggestions

You can also get suggestions for an individual column. Right-click the column and then select Suggestions ->Get suggestions from the context menu that pops up.

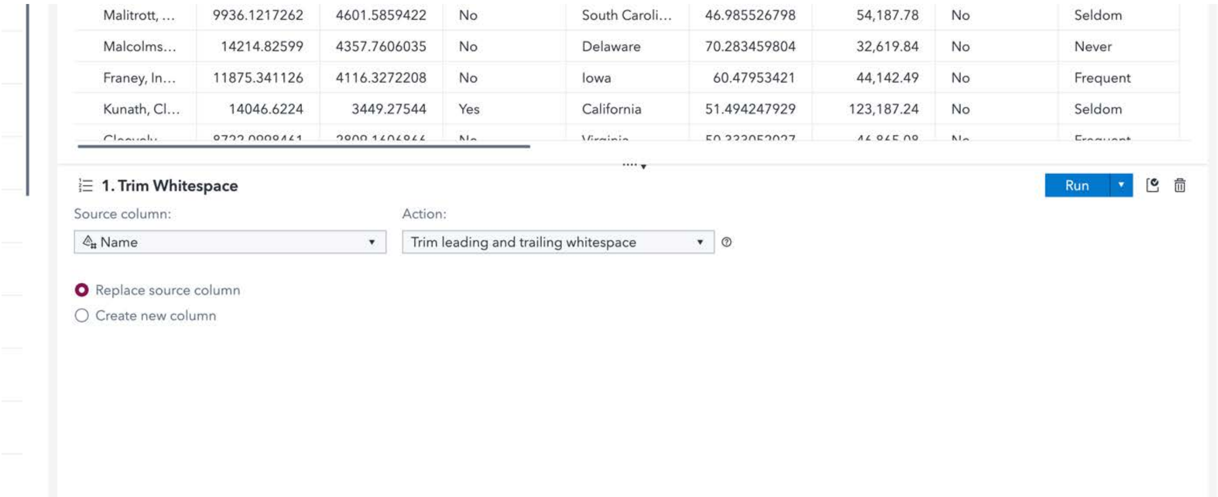


Name	Credi...	Expec...	SMS ...	
Lothe			No	M
MacT			No	M
Bolin			No	Fl
Jubb			Yes	D
Voge			Yes	N
Jahn,		3946.82432	Yes	Ve
Malit		4601.5859422	No	Sc
Malc		4357.7606035	No	D
Frane		4116.3272208	No	lo
Kuna		3449.27544	Yes	C

Display 6. Suggestions Available for the Column You're Interested in Improving

You can add suggestions from the suggestions pane on the left either by double-clicking on them or by dragging them to your data plan on the right. Or, you can right-click the column and select the suggested transformation for the column.

In the bottom center of the interface, you'll be able to change any settings the suggested transformation might have. When you are ready, click the blue Run button to run the transformation and see the results.



Malitrott, ...	9936.1217262	4601.5859422	No	South Caroli...	46.985526798	54,187.78	No	Seldom
Malcolms...	14214.82599	4357.7606035	No	Delaware	70.283459804	32,619.84	No	Never
Franey, In...	11875.341126	4116.3272208	No	Iowa	60.47953421	44,142.49	No	Frequent
Kunath, Cl...	14046.6224	3449.27544	Yes	California	51.494247929	123,187.24	No	Seldom

1. Trim Whitespace

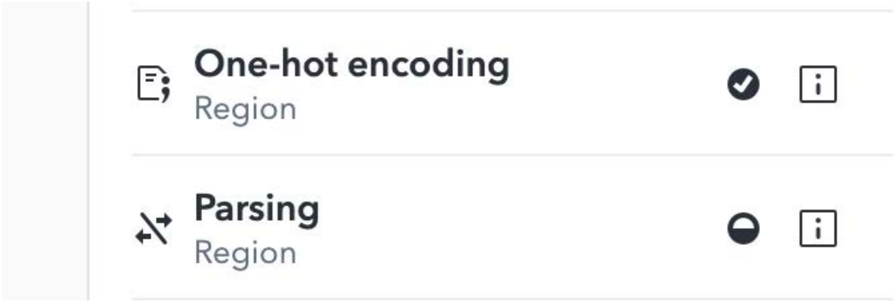
Source column: Name

Action: Trim leading and trailing whitespace

● Replace source column
○ Create new column

Display 7. You Can Adjust the Settings on a Suggested Transformation in the Bottom Center of the Window

At any time as you work to prepare your data, you can click the blue Get Suggestions button in the Suggestions pane to get an updated set of suggestions. Icons in the Suggestions pane show you which suggestions you've added to your plan (half-filled circle) and which ones you've already run (check mark icon).

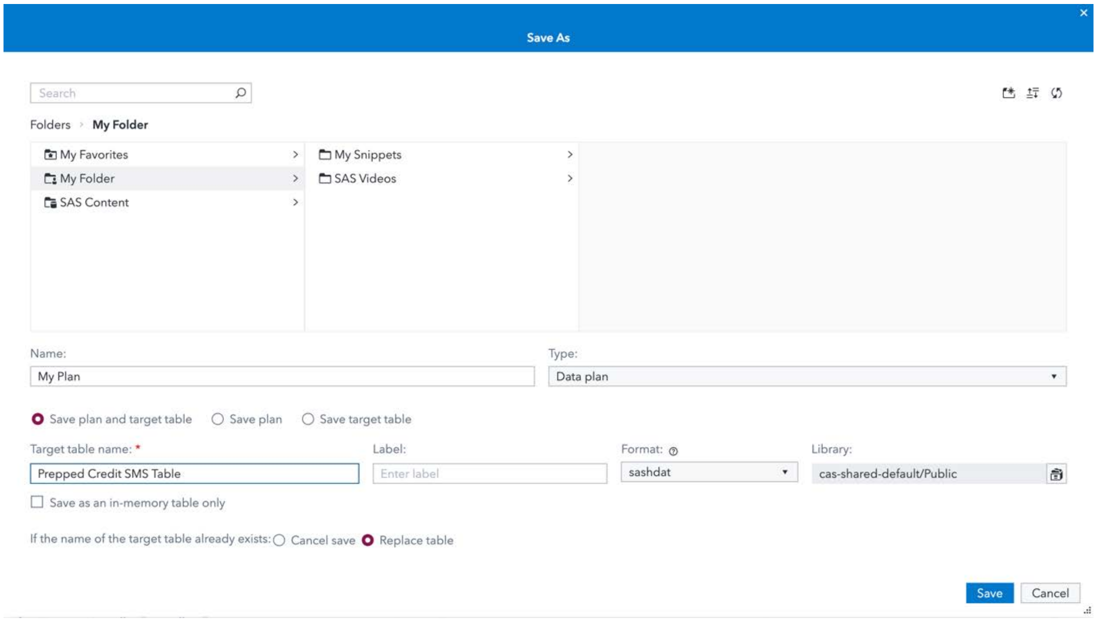


One-hot encoding
Region ✓ ⓘ

Parsing
Region ◓ ⓘ

Display 8. Icons Indicating Whether You've Added the Suggestion or Finished Running It

Working with suggested data prep is that easy. When you're done, you can save your plan and the table you've prepared and get to work analyzing your data and discovering insights.



Name: My Plan

Type: Data plan

● Save plan and target table ○ Save plan ○ Save target table

Target table name: * Prepped Credit SMS Table

Label: Enter label

Format: sashdat

Library: cas-shared-default/Public

□ Save as an in-memory table only

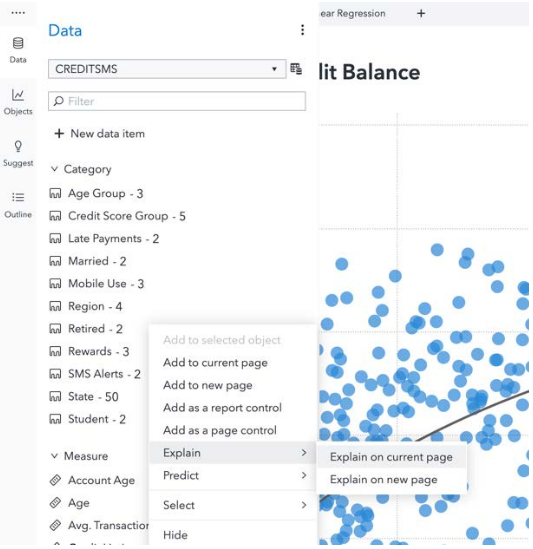
If the name of the target table already exists: ○ Cancel save ● Replace table

Display 9. Saving the Table and Plan for Analysis

AUTOMATIC DATA SCREENING IN AUTOMATED EXPLANATION

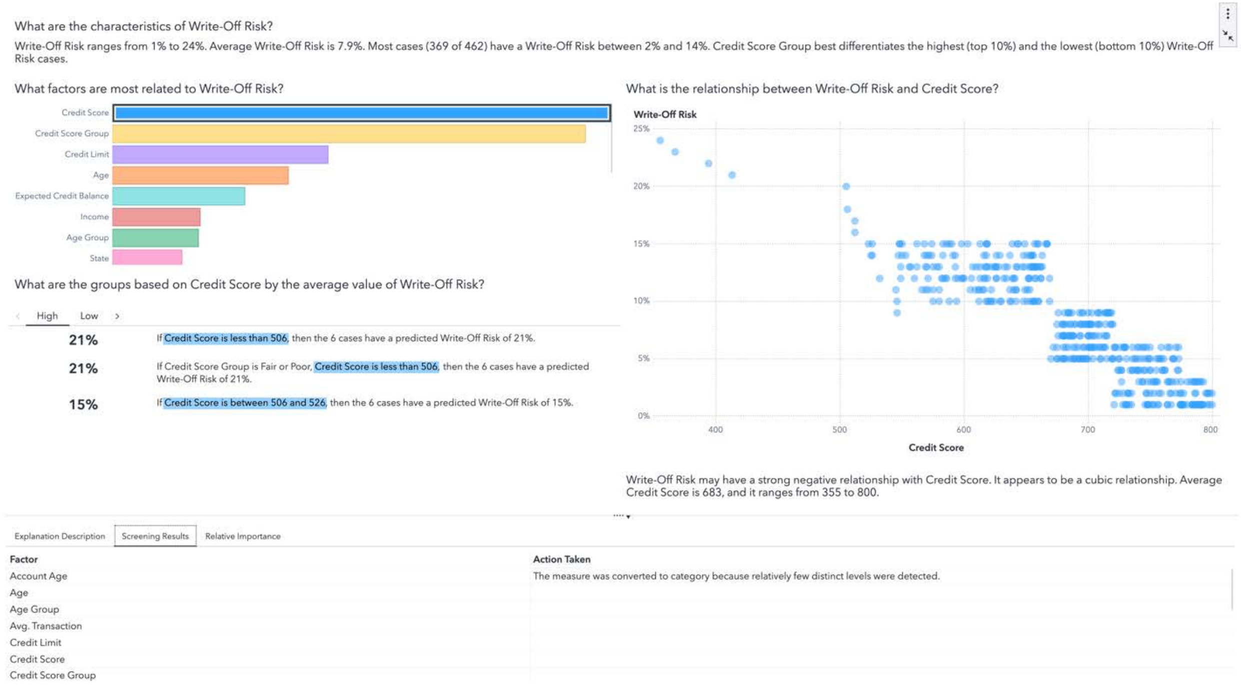
As you explore your data with SAS Visual Analytics, you might decide to use Automated Explanation to help speed you up. Automated Explanation analyzes your entire data set and creates an interactive analytical story with dynamic visualizations. In some cases, Automated Explanation can help you do hours of work of manual exploration in just a few seconds.

In SAS Visual Analytics' data pane, on the leftside, right-click the variable you'd like explained, and select Explain->Explain on new page or Explain->Explain on current page. Automated Explanation will take care of the rest.



Display 10. Explaining a Column in Your Data with SAS Visual Analytics

As part of taking care of the work for you, Automated Explanation does some additional screening on your data. To see what data screening occurred, expand the Automated Explanation and select the Screening Results tab.



What are the characteristics of Write-Off Risk?

Write-Off Risk ranges from 1% to 24%. Average Write-Off Risk is 7.9%. Most cases (369 of 462) have a Write-Off Risk between 2% and 14%. Credit Score Group best differentiates the highest (top 10%) and the lowest (bottom 10%) Write-Off Risk cases.

What factors are most related to Write-Off Risk?

What are the groups based on Credit Score by the average value of Write-Off Risk?

21% If Credit Score is less than 506, then the 6 cases have a predicted Write-Off Risk of 21%.

21% If Credit Score Group is Fair or Poor, Credit Score is less than 506, then the 6 cases have a predicted Write-Off Risk of 21%.

15% If Credit Score is between 506 and 526, then the 6 cases have a predicted Write-Off Risk of 15%.

What is the relationship between Write-Off Risk and Credit Score?

Write-Off Risk may have a strong negative relationship with Credit Score. It appears to be a cubic relationship. Average Credit Score is 683, and it ranges from 355 to 800.

Factor	Action Taken
Account Age	The measure was converted to category because relatively few distinct levels were detected.
Age	
Age Group	
Avg. Transaction	
Credit Limit	
Credit Score	
Credit Score Group	

Display 11. The Screening Results Tab Shows Which Actions Automated Explanation Made on the Data

The Screening Results tab shows you steps that Automated Explanation took to prepare your data for its analysis. It might remove some variables (for example, those that have a high number of missing values). It converts numeric columns with few distinct values to categories. It might even remove geographic variables like latitudes and longitudes from the analysis.

Data is something that might need constant preparation. Smart data prep algorithms in SAS Viya, like Automated Explanation's data screening, help you even after you finish your data prep.

CONCLUSION

With more AI-powered automation SAS Viya helps you speed up your data preparation process at every step of your analytical life cycle. It helps you learn about your data with automated statistics and profiles, it helps you quickly prepare that data with machine-learning powered suggestions, and it even continues to improve your data after you've finished preparing it with features like data screening in Automated Explanation.

RESOURCES

- SAS® Data Preparation Learn & Support web page
- SAS® Visual Analytics Learn & Support web page

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Atrin Assa
SAS Institute Inc.
Atrin.Assa@sas.com