SAS3183-2019

World Wide Wrangler: How to Create SAS® Data Sets from HTML Files in Seconds

Ed Summers, Brice Smith, and Sean Mealin, SAS Institute Inc., Cary, NC

ABSTRACT

Have you ever found a table of data on a web page and wanted to scrape that data into a SAS® data set? This paper shows you how to do it in seconds using SAS® Graphics Accelerator for Google Chrome, which is a free extension for Google Chrome. First, you open the web page in Chrome. Then, you extract data from the table into your Laboratory within SAS Graphics Accelerator for Google Chrome. SAS Graphics Accelerator for Google Chrome automatically derives variable labels from column headings and variable lengths from observations. It also derives variable "types" if all observations in a column match patterns for a single type. If they don't, you can choose the type of a variable manually. Finally, you generate a SAS program that creates a data set in your WORK library with correct variable labels, lengths, and formats. With SAS Graphics Accelerator for Google Chrome, you'll be the fastest data wrangler on the web!

INTRODUCTION

Data are frequently presented within tables on the web. Sometimes the author of the website includes a facility that enables users to download the data in a convenient format such as a comma-separated values (CSV) file. Unfortunately, many times they don't.

This paper shows you how to extract data from HTML tables on the web using SAS Graphics Accelerator for Google Chrome. It shows you how to quickly assign variable labels and types. It shows you how to filter the data, if needed. Finally, it shows you how to save the data as a CSV file or a SAS program that, when executed, loads the data into a SAS data set in your WORK library.

ABOUT SAS GRAPHICS ACCELERATOR FOR GOOGLE CHROME

SAS Graphics Accelerator for Google Chrome is designed for users with visual impairments or blindness (VIB). The primary purpose of SAS Graphics Accelerator for Google Chrome is to enable users with VIB to perceive data, charts, and graphs using alternative, that is, non-visual or visually enhanced, methods. These methods include the ability to explore charts and graphs interactively using sound. In other words, SAS Graphics Accelerator for Google Chrome converts the data in the graphs that you see visually into sound that totally blind users can hear.

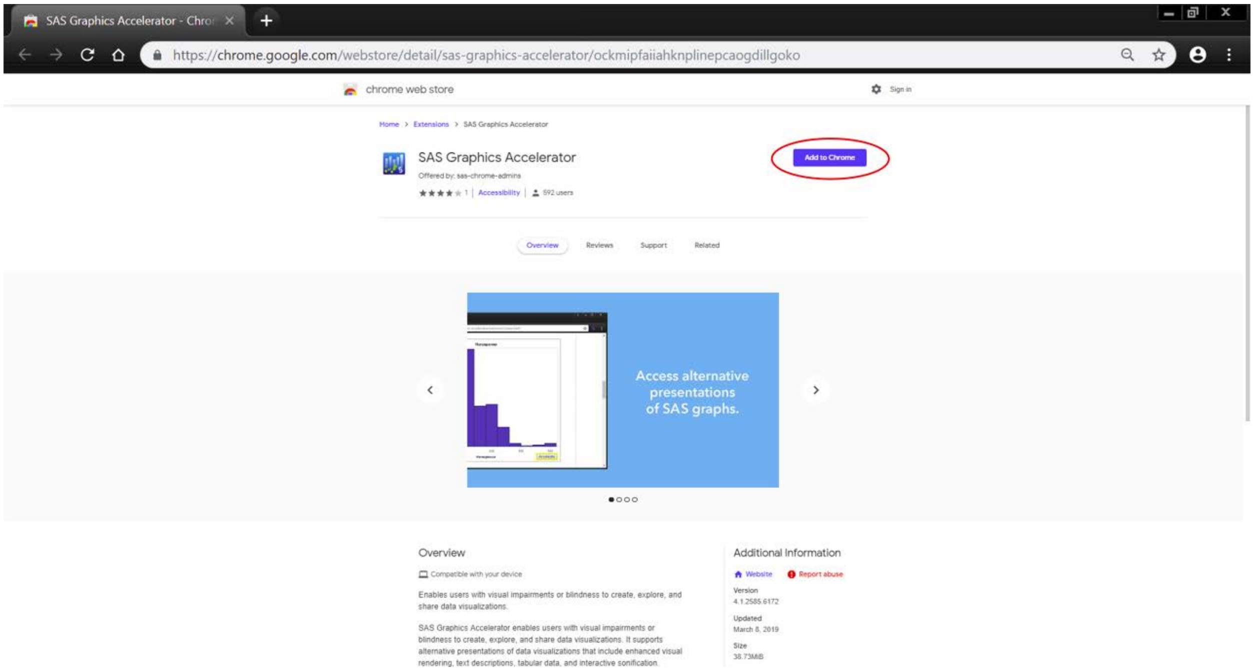
As a SAS programmer, you can create graphs that support this behavior by specifying the ACCESSIBLE_GRAPH option in the ODS HTML5 statement and then creating your graphs using the Statistical Graphics (SG) procedures in 9.4M6 or a later version of SAS. For more information, see Creating Accessible SAS 9.4 Output Using ODS and ODS Graphics.

When SAS programmers create accessible graphs as described above, users with VIB can easily perceive the graphs in approximately the same amount of time as their sighted peers. Unfortunately, most graphs on the web were not created in that way. However, the authors of those graphs might have followed a best practice for accessibility, which is to display the data that were used to create the graphs in a tabular format. We enhanced SAS Graphics Accelerator for Google Chrome to enable users with VIB to extract data from those tables and then re-create graphs from that data. We think that behavior might be useful to a wider audience. So, we wrote this paper.

INSTALLING SAS GRAPHICS ACCELERATOR FOR GOOGLE CHROME

Follow these steps to install SAS Graphics Accelerator for Google Chrome:

1. Launch Google Chrome and open the [SAS Graphics Accelerator for Google Chrome product page in the Chrome Web Store](#).
2. Press the "Add to Chrome" button.


SAS Graphics Accelerator - Chro
https://chrome.google.com/webstore/detail/sas-graphics-accelerator/ockmipfaiiahknplinepcaogdillgoko
chrome web store
Sign in
Home > Extensions > SAS Graphics Accelerator
SAS Graphics Accelerator
Offered by: sas-chrome-admins
1 | Accessibility | 592 users
Add to Chrome
Overview
Reviews
Support
Related
Access alternative presentations of SAS graphs.
Overview
Compatible with your device
Additional Information
Website
Report abuse
Version
4.1.2585.6172
Updated
March 8, 2019
Size
38.73MiB

Enables users with visual impairments or blindness to create, explore, and share data visualizations.

SAS Graphics Accelerator enables users with visual impairments or blindness to create, explore, and share data visualizations. It supports alternative presentations of data visualizations that include enhanced visual rendering, text descriptions, tabular data, and interactive sonification.

EXTRACTING A TABLE

SAS Graphics Accelerator for Google Chrome can extract a table if it meets all of the following criteria:

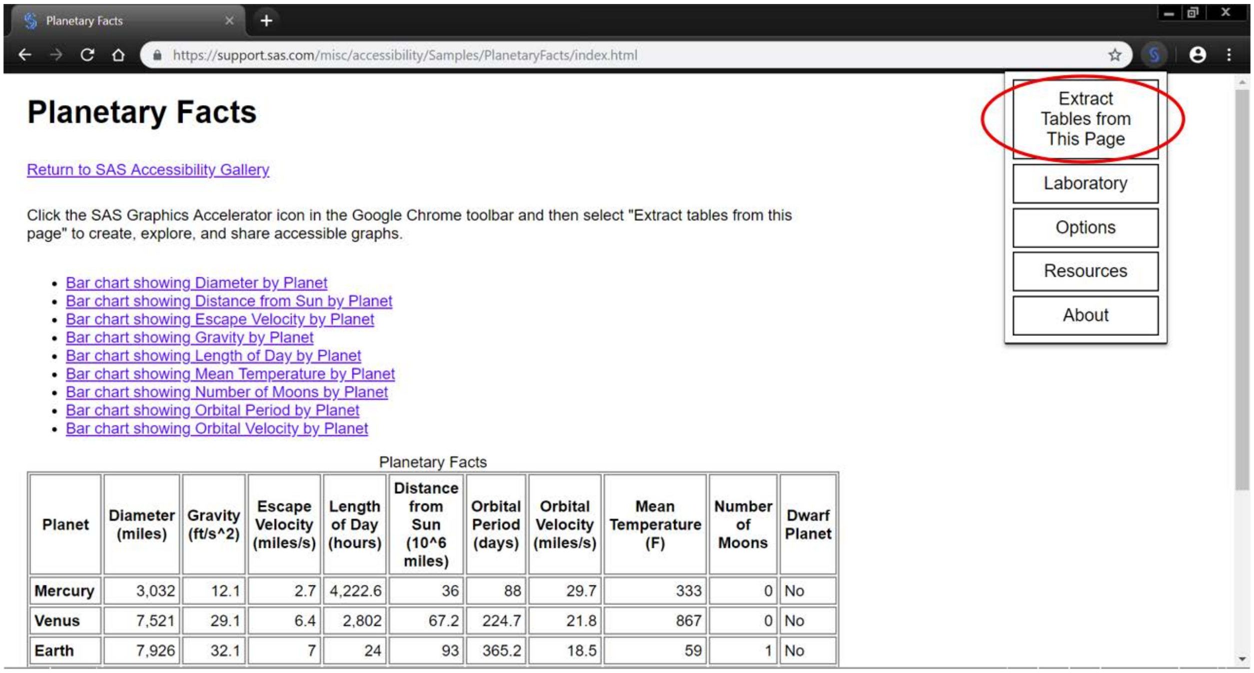
1. The page that contains the table is in English.
2. The table is rectangular, that is, each row contains the same number of cells.
3. The table is visible, that is, it has not been deliberately hidden by the web developer that created the web page.
4. The table is equal to or larger than 2 by 3 or 3 by 2.
5. The table's role attribute has not been set to "presentation". Note that web developers sometimes use a table with a role of "presentation" to position content spatially within a web page.
6. The table does not contain nested tables.
7. The size of the table is less than 1MB.

For example, suppose you want to extract the table on this page:

http://support.sas.com/misc/accessibility/Samples/PlanetaryFacts/index.html

Follow these steps to extract data from the table within that page into your Laboratory:

1. Open the web page in Google Chrome.
2. Click the SAS Graphics Accelerator for Google Chrome icon in the Chrome toolbar to open the pop-up menu.
3. In the pop-up menu, click the "Extract Tables from This Page" button. That should open the Prepare Table page in a new browser tab.
4. On the Prepare Table page, select the table(s) you wish to extract.
5. If the default values are not correct, provide a name for the table and select the appropriate values for row headers and column headers.
6. Click the "Save to Laboratory" button.


Planetary Facts
https://support.sas.com/misc/accessibility/Samples/PlanetaryFacts/index.html
Extract Tables from This Page
Laboratory
Options
Resources
About

Planetary Facts

Return to SAS Accessibility Gallery

Click the SAS Graphics Accelerator icon in the Google Chrome toolbar and then select "Extract tables from this page" to create, explore, and share accessible graphs.

- Bar chart showing Diameter by Planet
- Bar chart showing Distance from Sun by Planet
- Bar chart showing Escape Velocity by Planet
- Bar chart showing Gravity by Planet
- Bar chart showing Length of Day by Planet
- Bar chart showing Mean Temperature by Planet
- Bar chart showing Number of Moons by Planet
- Bar chart showing Orbital Period by Planet
- Bar chart showing Orbital Velocity by Planet

Planetary Facts

Planet	Diameter (miles)	Gravity (ft/s^2)	Escape Velocity (miles/s)	Length of Day (hours)	Distance from Sun (10^6 miles)	Orbital Period (days)	Orbital Velocity (miles/s)	Mean Temperature (F)	Number of Moons	Dwarf Planet
Mercury	3,032	12.1	2.7	4,222.6	36	88	29.7	333	0	No
Venus	7,521	29.1	6.4	2,802	67.2	224.7	21.8	867	0	No
Earth	7,926	32.1	7	24	93	365.2	18.5	59	1	No

ASSIGNING VARIABLE LABELS AND TYPES

When SAS Graphics Accelerator for Google Chrome extracts a table from a web page, the accelerator attempts to automatically detect the label and type of each variable (column). The label for each variable and icons that represent the type of each variable are displayed in the first row of the associated column.

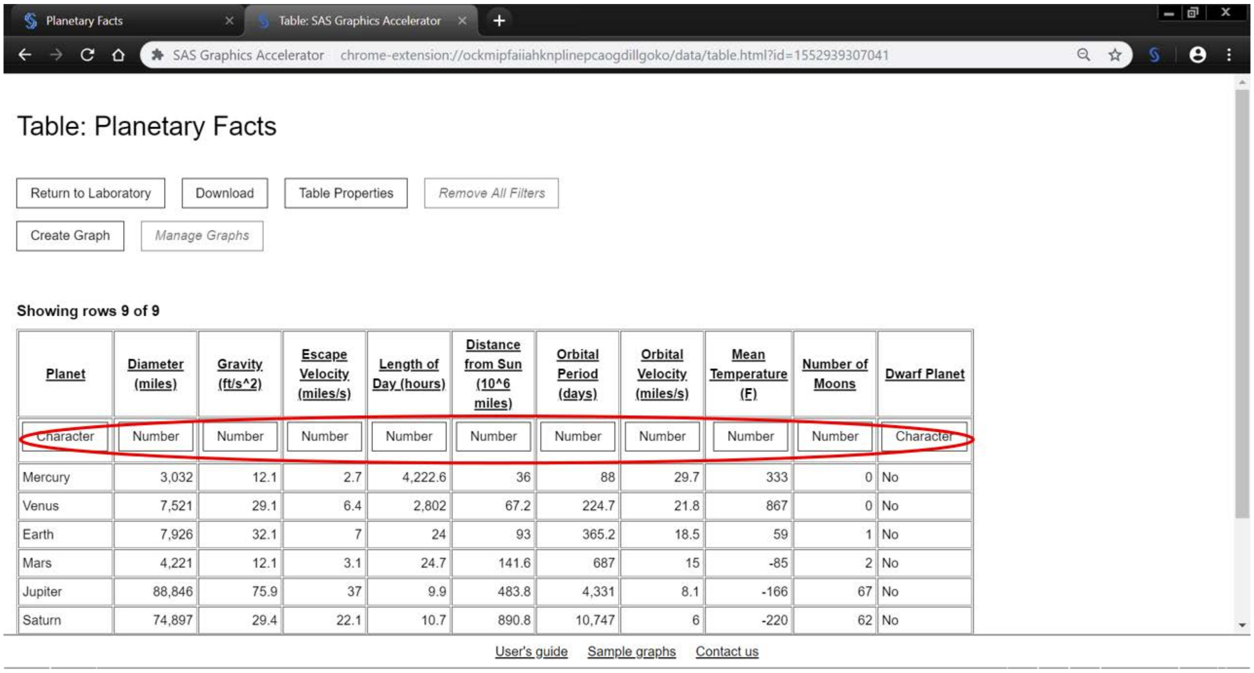
The label for each variable is derived from the column headings within the web page where the table was extracted.

By default, every variable is assigned the type of Character. However, if every cell in a column matches a heuristic for a non-Character type, then that type will be assigned to the corresponding variable. The following types are supported as of the time this paper was published:

- Character
- Number
- Currency
- Date

In addition to the types listed above, any variable can also be categorical. The Categorical attribute is a Boolean attribute. The default value of the attribute is determined by a heuristic.

To change these values for any variable, click the button in the second row of the corresponding column. Note that the "missing" value will be assigned to any cell that does not match a non-character type that is manually assigned. For example, if a cell contains "foo", and you assign the number type to the column that contains that cell, the value of that cell will be changed to "missing". The "missing" value is represented using a period (".").



Planetary Facts
Table: SAS Graphics Accelerator
SAS Graphics Accelerator chrome-extension://ockmipfaiiahknplinepcaogdillgoko/data/table.html?id=1552939307041

Table: Planetary Facts

Return to Laboratory
Download
Table Properties
Remove All Filters
Create Graph
Manage Graphs

Showing rows 9 of 9

Planet	Diameter (miles)	Gravity (ft/s^2)	Escape Velocity (miles/s)	Length of Day (hours)	Distance from Sun (10^6 miles)	Orbital Period (days)	Orbital Velocity (miles/s)	Mean Temperature (F)	Number of Moons	Dwarf Planet
Character	Number	Number	Number	Number	Number	Number	Number	Number	Number	Character
Mercury	3,032	12.1	2.7	4,222.6	36	88	29.7	333	0	No
Venus	7,521	29.1	6.4	2,802	67.2	224.7	21.8	867	0	No
Earth	7,926	32.1	7	24	93	365.2	18.5	59	1	No
Mars	4,221	12.1	3.1	24.7	141.6	687	15	-85	2	No
Jupiter	88,846	75.9	37	9.9	483.8	4,331	8.1	-166	67	No
Saturn	74,897	29.4	22.1	10.7	890.8	10,747	6	-220	62	No

FILTERING THE DATA

After you extract a table from a web page, you can filter the table so that it contains only the rows that are of interest. SAS Graphics Accelerator for Google Chrome supports the following filters as of the time this paper was published:

- Filter categorical variables by a category.
- Filter numeric and currency variables by comparison against a specified number.
- Filter character variables by comparison against a specified string.
- Filter any variable by missing values.

To apply or remove a filter for a specific variable, click on the button in the second row of the corresponding column. Note that if filters are applied to two or more variables, the table will display only the rows that match all the filters.

To remove all filters from a table, click the "Remove All Filters" button.

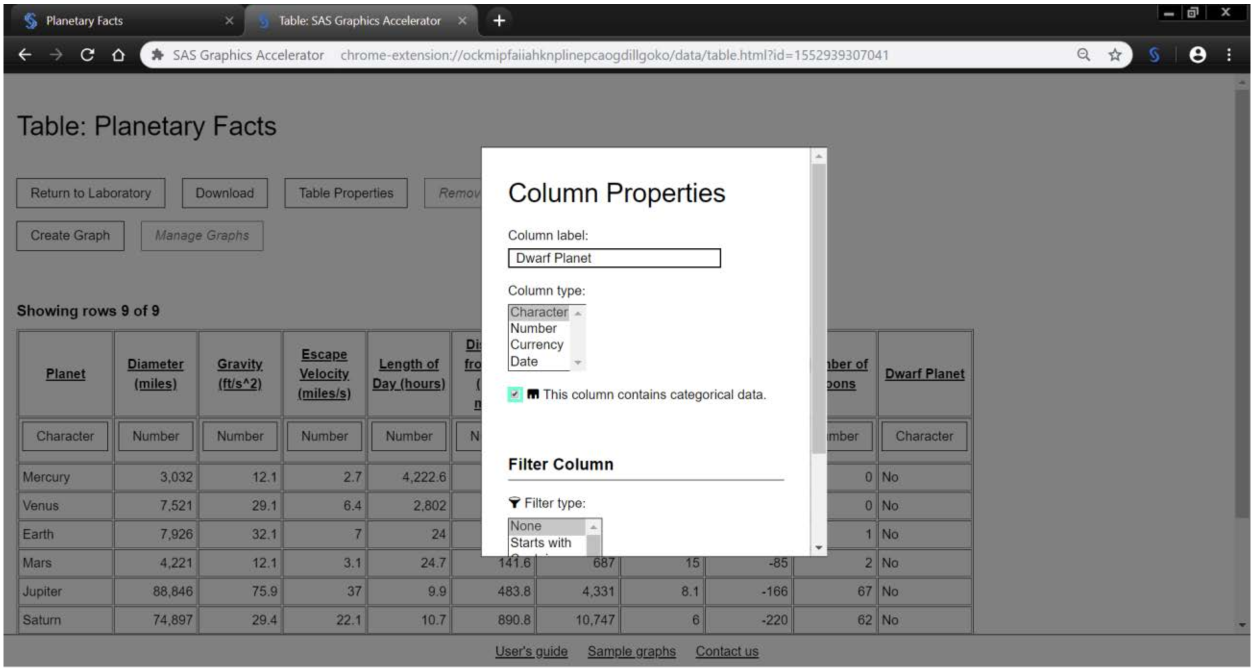

Planetary Facts
Table: SAS Graphics Accelerator
SAS Graphics Accelerator chrome-extension://ockmipfaiiahknplinepcaogdillgoko/data/table.html?id=1552939307041
Return to Laboratory
Download
Table Properties
Remov
Create Graph
Manage Graphs
User's guide
Sample graphs
Contact us

Table: Planetary Facts

Showing rows 9 of 9

Planet	Diameter (miles)	Gravity (ft/s^2)	Escape Velocity (miles/s)	Length of Day (hours)	Di fro (n				ber of oons	Dwarf Planet
Character	Number	Number	Number	Number	N				mber	Character
Mercury	3,032	12.1	2.7	4,222.6					0	No
Venus	7,521	29.1	6.4	2,802					0	No
Earth	7,926	32.1	7	24					1	No
Mars	4,221	12.1	3.1	24.7	141.6	687	15	-85	2	No
Jupiter	88,846	75.9	37	9.9	483.8	4,331	8.1	-166	67	No
Saturn	74,897	29.4	22.1	10.7	890.8	10,747	6	-220	62	No

Column Properties

Column label: Dwarf Planet

Column type: Character

☑ This column contains categorical data.

Filter Column

Filter type: None

CREATING A SAS DATA SET

After you extract a table from a web page, you can quickly create a data set in SAS using the following steps:

1. Click the "Download" button.
2. In the Download dialog box, select "SAS program" as the file type, enter a filename and then press the "Download" button.

SAS Graphics Accelerator for Google Chrome generates a SAS program and saves it to the downloads folder you have configured within Google Chrome. When you run the SAS program, it generates a SAS data set in your WORK library. Note that if you have applied one or more filters to a table, only the matching rows will be included in the resulting data set.

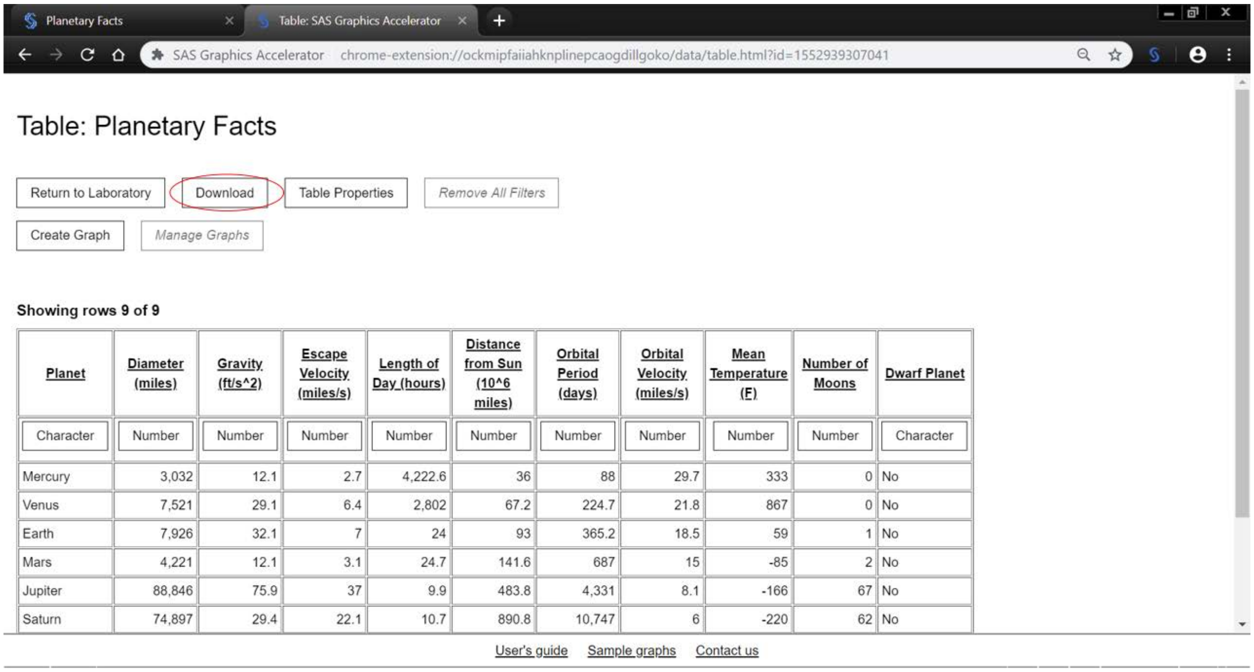

Planetary Facts
Table: SAS Graphics Accelerator
SAS Graphics Accelerator
chrome-extension://ockmipfaiiahknplinepcaogdillgoko/data/table.html?id=1552939307041

Table: Planetary Facts

Showing rows 9 of 9

Planet	Diameter (miles)	Gravity (ft/s^2)	Escape Velocity (miles/s)	Length of Day (hours)	Distance from Sun (10^6 miles)	Orbital Period (days)	Orbital Velocity (miles/s)	Mean Temperature (F)	Number of Moons	Dwarf Planet
Character	Number	Number	Number	Number	Number	Number	Number	Number	Number	Character
Mercury	3,032	12.1	2.7	4,222.6	36	88	29.7	333	0	No
Venus	7,521	29.1	6.4	2,802	67.2	224.7	21.8	867	0	No
Earth	7,926	32.1	7	24	93	365.2	18.5	59	1	No
Mars	4,221	12.1	3.1	24.7	141.6	687	15	-85	2	No
Jupiter	88,846	75.9	37	9.9	483.8	4,331	8.1	-166	67	No
Saturn	74,897	29.4	22.1	10.7	890.8	10,747	6	-220	62	No

CONCLUSION

SAS Graphics Accelerator for Google Chrome enables you to extract data from tables within web pages and create a SAS data set. This feature was originally created for users with visual impairments or blindness including all authors of this paper. Sometimes technology created specifically for people with disabilities can also be useful to a wider audience. We hope you enjoy this feature as much as we do.

ACKNOWLEDGMENTS

The authors would like to acknowledge the contributions of our team members including Jeanette Bottitta, Greg Kraus, Julianna Langston, Lisa Morton, Jesse Sookne, and Tyler Williamson.

RECOMMENDED READING

- *SAS Graphics Accelerator User’s Guide*
- *Creating Accessible SAS 9.4 Output Using ODS and ODS Graphics*
- *Creating Accessible Reports Using SAS Visual Analytics 7.4*

CONTACT INFORMATION

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

Ed Summers
SAS
ed.summers@sas.com
Brice Smith
SAS
brice.smith@sas.com
Sean Mealin
SAS

sean.mealin@sas.com