Read CSV from URL

Read comma separated data tables from a URL into Modeler

Product: IBM® SPSS® Modeler

Extension type: Utility

Table of Contents

Description:

This node allows you to read data in a csv format on a web page directly into IBM SPSS Modeler. Many dataset repositories make data available in this format and they often provide a URL containing the data. Copy a URL into this node and it can read the file for you without manually downloading it.

Requirements:

- SPSS Modeler v16.0 or later
- R: http://www.r-project.org/
- 'R Essentials for SPSS Modeler' plugin: https://developer.ibm.com/predictiveanalytics/downloads/predictive-extensions

Installation:

Close SPSS Modeler. Save the .cfe file in the CDB folder of the IBM SPSS Modeler installation directory for Windows and Linux. The copy should reside in that same folder and not in a sub-folder.

For example, for Windows 7 and Modeler 17 the default location is "C:\ProgramData\IBM\SPSS\Modeler\17\CDB". If the ProgramData folder is hidden type the path manually.

Restart SPSS Modeler: the node will now appear in the Field Ops palette.

R Package used:

- 'RCurl' package created by Duncan Temple Lang and the CRAN team https://cran.r-project.org/web/packages/RCurl/

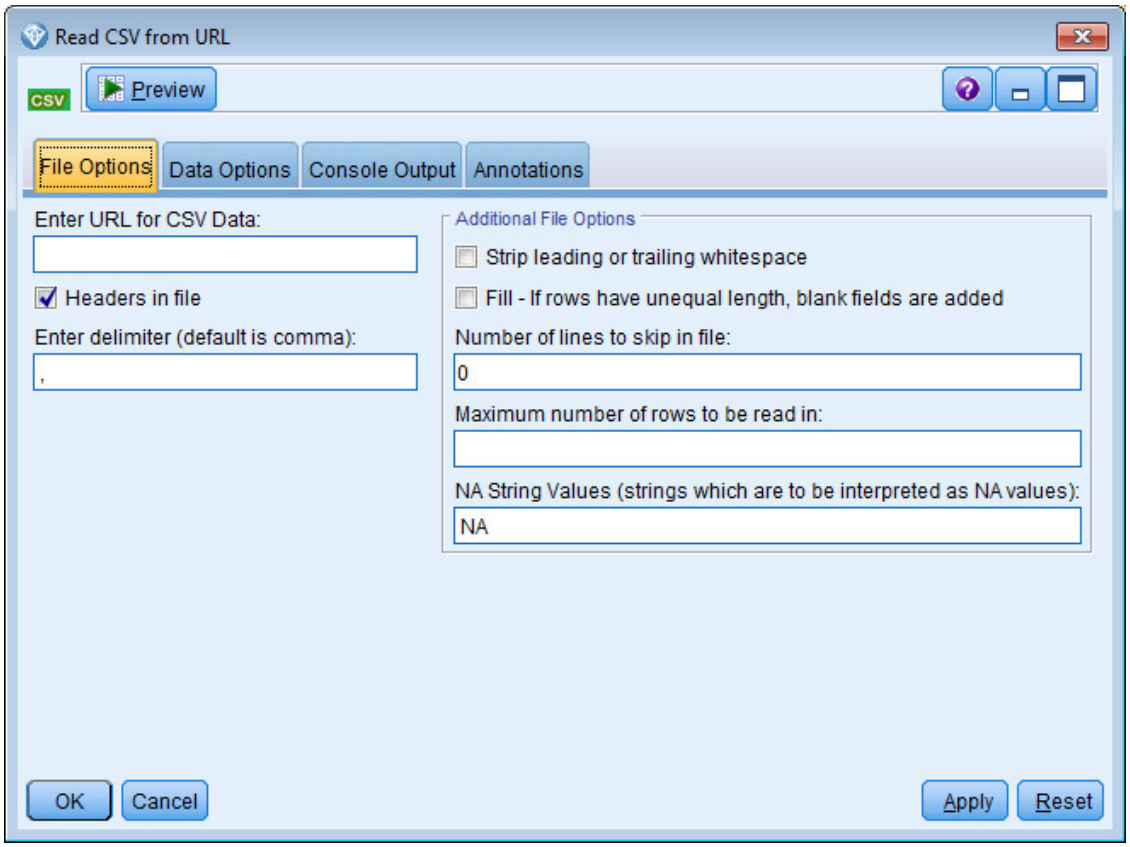
User Interface

This node has one tab with details on reading the data.

The most commonly adjusted parameters are found on the left side of this tab. Here you can paste the URL containing the data. The headers in file checkbox allows you to check the box if your data has headers, or column names, in the first row. The delimiter can be changed to another value, rather than a comma, by typing it in the Enter delimiter box.

On the right side of this tab, there are more advanced settings for reading in files.

- Strip leading or trailing whitespace – if checked will remove white space from the start and end of unquoted characters
- Fill – if rows have unequal length, blank fields are added
- Number of lines to skip in file – node will skip this number of lines before reading the data
- Maximum number of rows to be read in
- NA String Values (strings which are to be interpreted as NA values) – these values will be read into SPSS Modeler as Null values

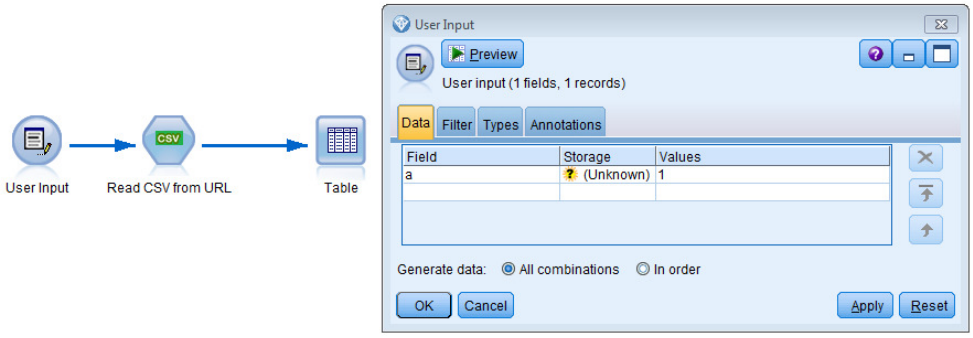


Example

This example will demonstrate how to read CSV data into SPSS Modeler from an URL. The sample data we will read for demonstration purposes only is provided by the University of California Irvine Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets.html). For the example, we will read in the 'Wine' dataset.

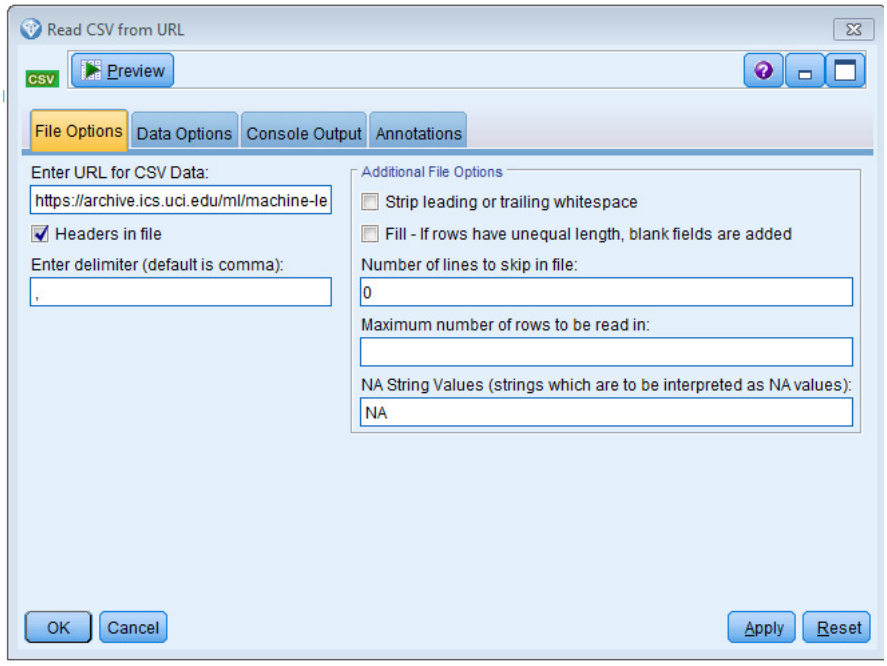
User Input

It is currently not possible to create proper 'source' nodes with R code. However, it is made possible if you trick the software a bit: you just have to put a user input node with fake fields before the Read CSV from URL node in order to make the stream start. In this example we put a user input with: Field a, Storage Unknown, Values 1 (see the screen shot below).

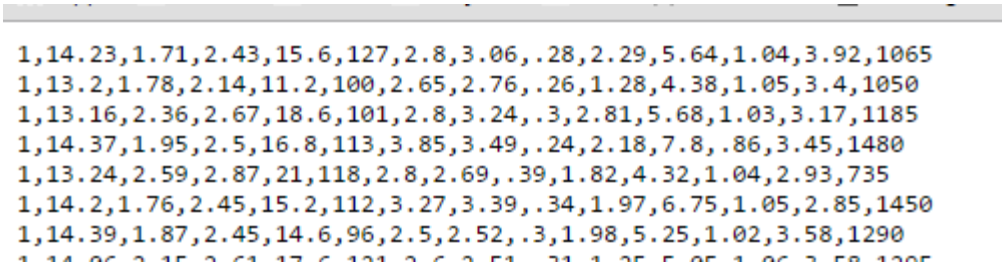


Double click the 'Read CSV from URL' node. Copy this URL and paste it into the 'Enter URL for CSV Data' field:

https://archive.ics.uci.edu/ml/machine-learning-databases/wine/wine.data

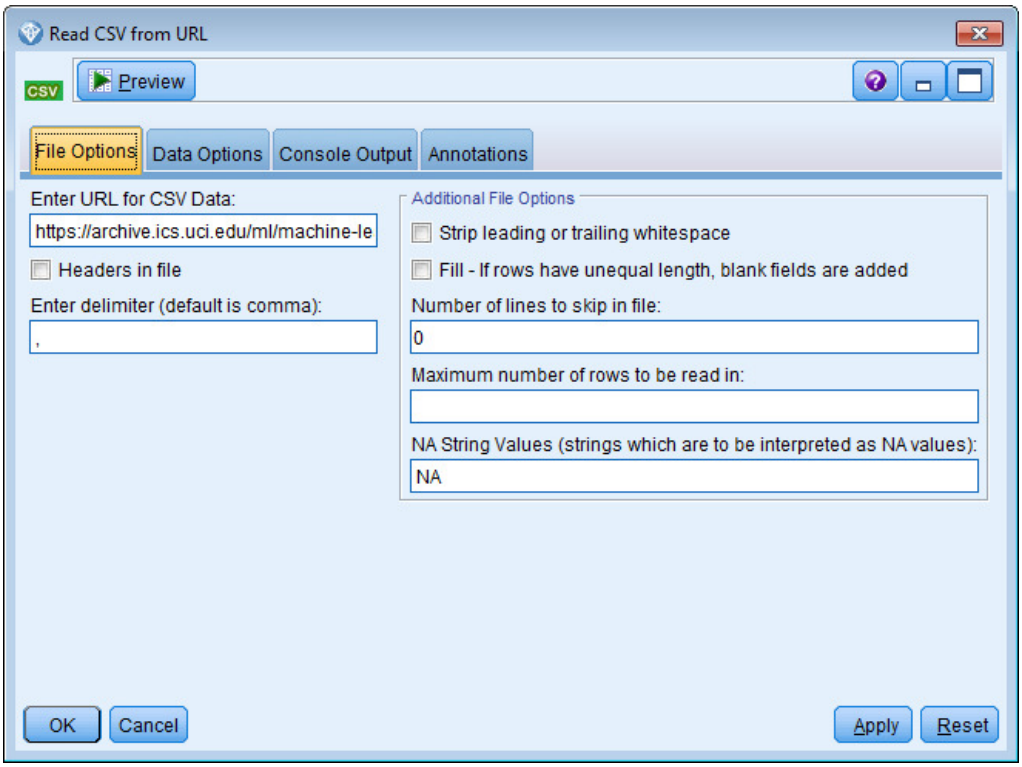


In order to determine how the rest of the dialog should be completed, we'll need to take a look at the raw data in the web page. See the screen shot below:

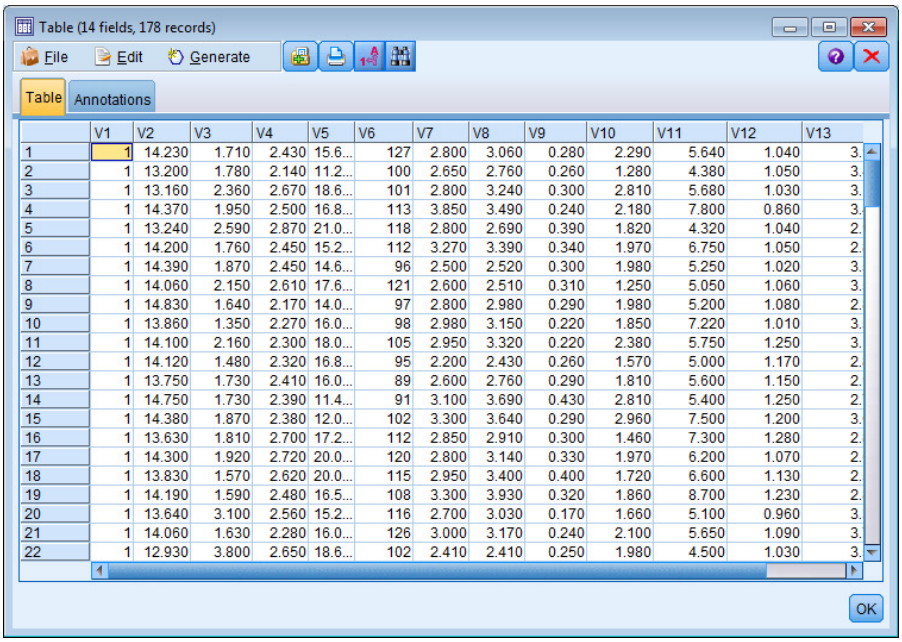


```
1,14.23,1.71,2.43,15.6,127,2.8,3.06,.28,2.29,5.64,1.04,3.92,1065
1,13.2,1.78,2.14,11.2,100,2.65,2.76,.26,1.28,4.38,1.05,3.4,1050
1,13.16,2.36,2.67,18.6,101,2.8,3.24,.3,2.81,5.68,1.03,3.17,1185
1,14.37,1.95,2.5,16.8,113,3.85,3.49,.24,2.18,7.8,.86,3.45,1480
1,13.24,2.59,2.87,21,118,2.8,2.69,.39,1.82,4.32,1.04,2.93,735
1,14.2,1.76,2.45,15.2,112,3.27,3.39,.34,1.97,6.75,1.05,2.85,1450
1,14.39,1.87,2.45,14.6,96,2.5,2.52,.3,1.98,5.25,1.02,3.58,1290
```

We see that this data has no header, meaning no column names are provided in the first row. This data is also separated with a comma, so we will want to use a comma for a delimiter. Other additional file options can be tweaked if desired, but knowing these two pieces of information will allow us to read the data in. The screenshot below shows the dialog setup for reading this data followed by a table output.



Results



	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13
1	1	14.230	1.710	2.430	15.6...	127	2.800	3.060	0.280	2.290	5.640	1.040	3.
2	1	13.200	1.780	2.140	11.2...	100	2.650	2.760	0.260	1.280	4.380	1.050	3.
3	1	13.160	2.360	2.670	18.6...	101	2.800	3.240	0.300	2.810	5.680	1.030	3.
4	1	14.370	1.950	2.500	16.8...	113	3.850	3.490	0.240	2.180	7.800	0.860	3.
5	1	13.240	2.590	2.870	21.0...	118	2.800	2.690	0.390	1.820	4.320	1.040	2.
6	1	14.200	1.760	2.450	15.2...	112	3.270	3.390	0.340	1.970	6.750	1.050	2.
7	1	14.390	1.870	2.450	14.6...	96	2.500	2.520	0.300	1.980	5.250	1.020	3.
8	1	14.060	2.150	2.610	17.6...	121	2.600	2.510	0.310	1.250	5.050	1.060	3.
9	1	14.830	1.640	2.170	14.0...	97	2.800	2.980	0.290	1.980	5.200	1.080	2.
10	1	13.860	1.350	2.270	16.0...	98	2.980	3.150	0.220	1.850	7.220	1.010	3.
11	1	14.100	2.160	2.300	18.0...	105	2.950	3.320	0.220	2.380	5.750	1.250	3.
12	1	14.120	1.480	2.320	16.8...	95	2.200	2.430	0.260	1.570	5.000	1.170	2.
13	1	13.750	1.730	2.410	16.0...	89	2.600	2.760	0.290	1.810	5.600	1.150	2.
14	1	14.750	1.730	2.390	11.4...	91	3.100	3.690	0.430	2.810	5.400	1.250	2.
15	1	14.380	1.870	2.380	12.0...	102	3.300	3.640	0.290	2.960	7.500	1.200	3.
16	1	13.630	1.810	2.700	17.2...	112	2.850	2.910	0.300	1.460	7.300	1.280	2.
17	1	14.300	1.920	2.720	20.0...	120	2.800	3.140	0.330	1.970	6.200	1.070	2.
18	1	13.830	1.570	2.620	20.0...	115	2.950	3.400	0.400	1.720	6.600	1.130	2.
19	1	14.190	1.590	2.480	16.5...	108	3.300	3.930	0.320	1.860	8.700	1.230	2.
20	1	13.640	3.100	2.560	15.2...	116	2.700	3.030	0.170	1.660	5.100	0.960	3.
21	1	14.060	1.630	2.280	16.0...	126	3.000	3.170	0.240	2.100	5.650	1.090	3.
22	1	12.930	3.800	2.650	18.6...	102	2.410	2.410	0.250	1.980	4.500	1.030	3.

Your data is now ready to use in SPSS Modeler with only one step!

Important Links

Learn

- Learn more about SPSS software.
- Visit developerWorks Business analytics for more technical analytics resources for developers.
- Create your Bluemix account here: https://console.ng.bluemix.net/solutions/watson
- The Comprehensive R Archive Network is the main site for the R project and each R package. The help pages and manuals that are associated with optimx, nlmrt, and Rcgmin are detailed. Numerous references are provided.
- Read "Do I need to learn R?" (Catherine Dalzell, developerWorks, September 2013) to learn why R is a valuable tool for data analytics that was expressly designed to reflect the way that statisticians think and work.
- "Calling R from SPSS" describes how to use R code inside IBM SPSS Modeler 16.
- Read "Create new nodes for IBM SPSS Modeler 16 using R" to learn how to create new extensions easily.

Discuss

- Visit the IBM SPSS Community to share tips and experiences with other IBM SPSS developers.
- Follow developerWorks on Twitter to be among the first to hear about new resources.