

Inference DataFrames QuickStart 1

Assemble data to access from R code

Inference DataFrames QuickStart 3 shows you how to:

- Create a DataFrame container and specify attributes of the data set.
- Specify data vectors and add data.
- Specify the data type and additional attributes for data vectors.

What is a DataFrame?

Inference lets you save your data and R code in the same document. This not only makes it easy for your R code to access your data, but it also lets you keep track of which data you used each time you run the code.

To store data, Inference uses a DataFrame. A DataFrame is a rectangular data structure similar to spreadsheets, database tables, matrices and rectangular lists. DataFrames are useful because real-life data often is often formatted as a column representing a variable and each row representing a case for all the variables.

Inference DataFrames have the following properties:

- Each column corresponds to vector of values associated with a variable.
- All columns must be the same length.
- All values in a column must be of the same data type.
- Different columns may have different data types.

Here is a sample Data Frame:

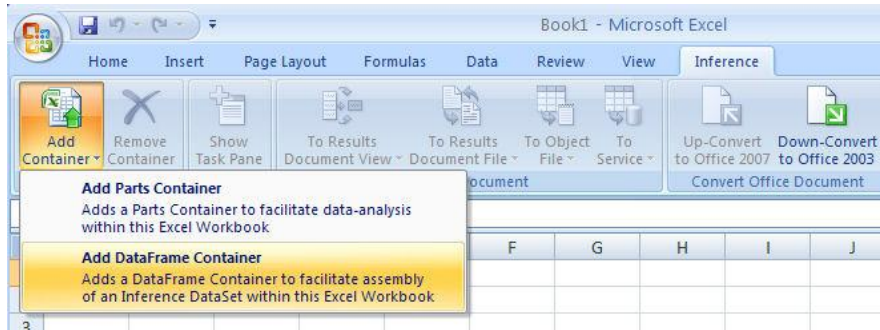
| | Temp | Conc | Cat | Yield |
|---|------|------|-----|-------|
| 1 | 160 | 20 | A | 60.1 |
| 2 | 180 | 20 | A | 72.5 |
| 3 | 160 | 40 | A | 54.2 |
| 4 | 180 | 40 | A | 68.7 |
| 5 | 160 | 20 | B | 52.4 |
| 6 | 180 | 20 | B | 83.1 |
| 7 | 160 | 40 | B | 45.7 |
| 8 | 180 | 40 | B | 80.3 |

In addition to the data vector of values (as shown above), DataFrames can also include attributes (metadata) about the DataFrame and the data vectors. Examples are included in this tutorial.

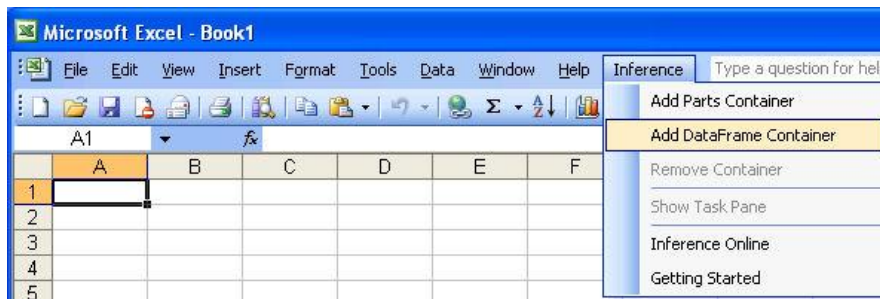
1. Add a DataFrame Container to an Excel Document

To add a DataFrame container to an Excel document:

1. Create or open an Excel document.
2. In Excel 2007: Click the **Inference** tab on the Excel Ribbon. Click **Add Container**, then select **Add DataFrame Container**.

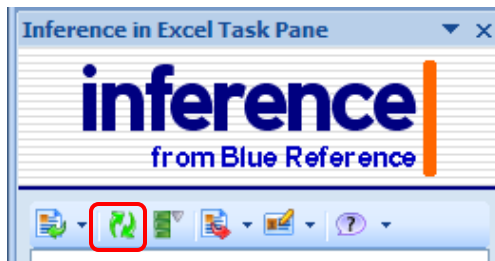


In Excel 2003: On the **Inference** menu, select **Add DataFrame Container**.

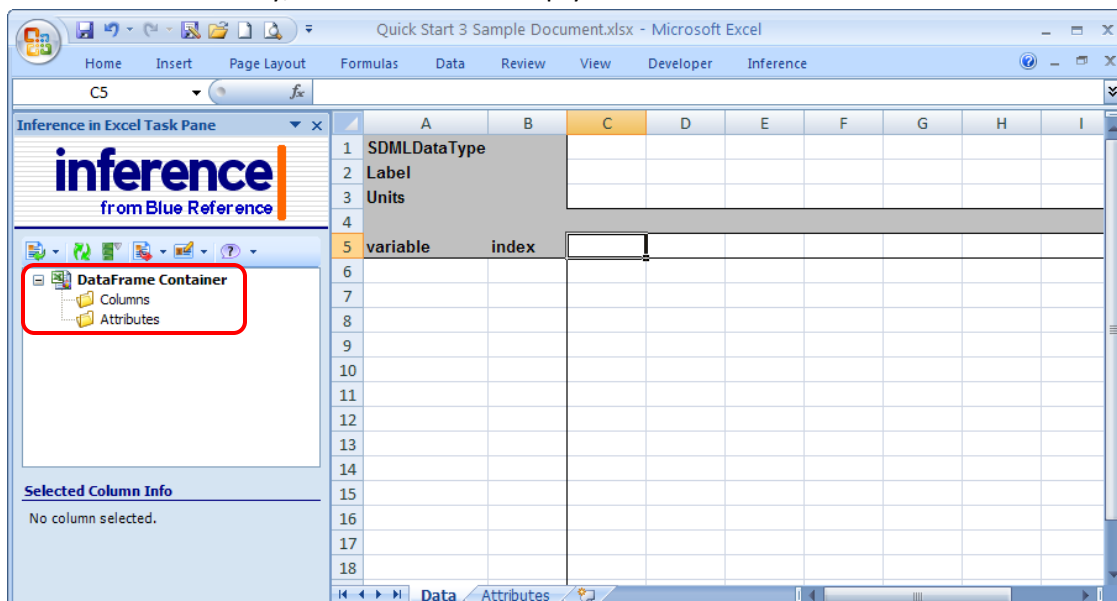


This creates a DataFrame container in the Excel document. You'll see the **Inference in Excel Task Pane** and two worksheets: **Data** and **Attributes**.

3. Save the document.
4. On the **DataFrame Container** toolbar, click the **Refresh** button.



5. Inference will inspect the contents of the DataFrame container and return a DataFrame Container tree. Initially, the tree contains empty Columns and Attributes.



2. Specify Data Frame Attributes

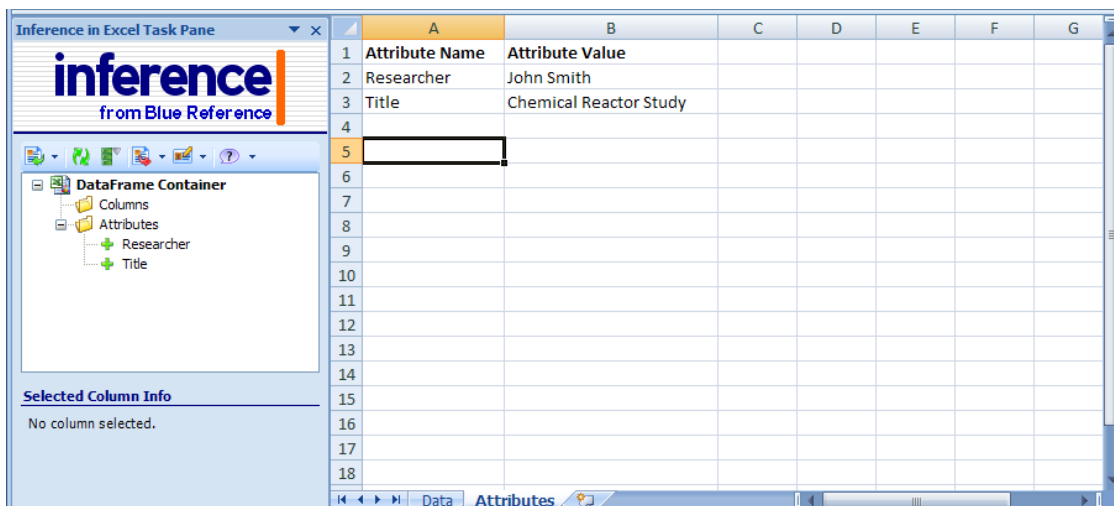
One of the benefits of the Inference DataFrame is the ability to specify attributes (metadata) of the DataFrame. DataFrame attributes are specified as parameter:value pairs.

When the DataFrame is embedded in Inference documents, your R code can access these attributes. This will be illustrated in other QuickStarts.

To add attributes to your DataFrame:

1. Select the **Attributes** worksheet.
2. In column A, type the name of the attribute. In column B, type the attribute value. For this example, add the following two attributes:
 - Researcher: John Smith
 - Title: Chemical Reactor Study

- On the **DataFrame Container** toolbar, click the **Refresh** button. Notice the new Attribute entries in the tree for **Researcher** and **Title**.



3. Specify Data Vectors and Add Data

Data vectors are specified on the **Data** sheet. To illustrate the procedure, do the following:

- Select the **Data** worksheet.
- Select and copy the four columns from the following table:

| Temp | Conc | Cat | Yield |
|------|------|-----|-------|
| 160 | 20 | A | 60.1 |
| 180 | 20 | A | 72.5 |
| 160 | 40 | A | 54.2 |
| 180 | 40 | A | 68.7 |
| 160 | 20 | B | 52.4 |
| 180 | 20 | B | 83.1 |
| 160 | 40 | B | 45.7 |
| 180 | 40 | B | 80.3 |

- Click in cell C:5 of the **Data** worksheet and press Ctrl-V to paste the contents of the above table.
- Click the **Update Indices** button on the **DataFrame Container** toolbar.
- Click the **Refresh** button on the **DataFrame Container** toolbar.
- Click the '+' next to **Temp** in the **DataFrame Container** tree.
- Compare your results to the following:

| variable | index | Temp | Conc | Cat | Yield |
|----------|-------|------|------|-----|-------|
| | 1 | 160 | 20 | A | 60.1 |
| | 2 | 180 | 20 | A | 72.5 |
| | 3 | 160 | 40 | A | 54.2 |
| | 4 | 180 | 40 | A | 68.7 |
| | 5 | 160 | 20 | B | 52.4 |
| | 6 | 180 | 20 | B | 83.1 |
| | 7 | 160 | 40 | B | 45.7 |
| | 8 | 180 | 40 | B | 80.3 |

Notice that:

- The indices for the rows in the DataFrame are automatically calculated and displayed.
- The new data vectors are displayed in the **DataFrame Container** tree.
- Inference makes a “best guess” as to the data type for each data vector. These are displayed in the tree including associated properties.
- Selecting one of the Data Vectors in the **DataFrame Container** tree tests the contents of the vector against the specifications of the data type and displays the results of the test under **Selected Column Info** on the Task Pane.

4. Make the Data Type for Each Data Vector Explicit

Excel has limited capabilities for inferring the data type of a cell entry. The Inference DataFrame container extends Excel’s capabilities by allowing a more diverse set of data types which can be inferred dynamically or explicitly specified. Inference can also validate the data against the specifications of the data types.

For ultimate control over data types, an Inference DataFrame allows the user to statically specify the data type. This eliminates all ambiguity regarding the intent when specifying the data vector and avoids the unintended consequences that can arise from a mismatched data type.

To assign explicit data types:

1. Select the **Data** worksheet.
2. In cells C:1 through F:1, click to select the following from the drop-down boxes:
 - Temperature: numeric-integer
 - Concentration: numeric-integer
 - Category: categorical

- Yield: numeric-real
3. Click **Refresh** on the **DataFrame Container** toolbar. Notice that the data type for **Category** has been updated to **categorical** with two labels.

The screenshot shows the 'Inference in Excel Task Pane' window. On the left, the 'DataFrame Container' tree view shows a 'Columns' folder containing 'Temp (numeric-integer)', 'Conc (numeric-integer)', 'Cat (categorical) (2 Labels)', and 'Yield (numeric-real)'. The 'Cat' column is selected. The main area displays a table with the following data:

| | A | B | C | D | E | F |
|----|--------------|-------|-----------------|-----------------|-------------|--------------|
| 1 | SDMLDataType | | numeric-integer | numeric-integer | categorical | numeric-real |
| 2 | Label | | | | | |
| 3 | Units | | | | | |
| 4 | | | | | | |
| 5 | variable | index | Temp | Conc | Cat | Yield |
| 6 | | 1 | 160 | 20 | A | 60.1 |
| 7 | | 2 | 180 | 20 | A | 72.5 |
| 8 | | 3 | 160 | 40 | A | 54.2 |
| 9 | | 4 | 180 | 40 | A | 68.7 |
| 10 | | 5 | 160 | 20 | B | 52.4 |
| 11 | | 6 | 180 | 20 | B | 83.1 |
| 12 | | 7 | 160 | 40 | B | 45.7 |
| 13 | | 8 | 180 | 40 | B | 80.3 |
| 14 | | | | | | |

5. Specify Additional Data Vector Attributes

An Inference DataFrame allows you to specify an unlimited set of additional attributes for the collection of Data Vectors. By default, an Inference DataFrame includes an attribute for **Label** and an attribute for **Units**.

To specify the **Label** and **Units** attributes:

1. Select the **Data** worksheet.
2. For each Data Vector, specify the **Label** as follows:
 - Temp: Temperature
 - Conc: Concentration
 - Cat: Catalyst
 - Yield: Yield
3. For the Data Vectors which have associated units, specify the **Units** as follows:
 - Temperature: deg-C
 - Concentration: grams/L
 - Yield: %
4. Compare your **Data** worksheet to the following:

| variable | index | Temp | Conc | Cat | Yield |
|----------|-------|------|------|-----|-------|
| | 1 | 160 | 20 | A | 60.1 |
| | 2 | 180 | 20 | A | 72.5 |
| | 3 | 160 | 40 | A | 54.2 |
| | 4 | 180 | 40 | A | 68.7 |
| | 5 | 160 | 20 | B | 52.4 |
| | 6 | 180 | 20 | B | 83.1 |
| | 7 | 160 | 40 | B | 45.7 |
| | 8 | 180 | 40 | B | 80.3 |

Since these Data Vector attributes are programmatically accessible from the scripting platform, the attributes and the corresponding values can be used in calculations, validation and as labels in graphics. This is illustrated in subsequent QuickStarts.

You can also add additional Data Vector attributes by inserting additional rows after the last attribute (row 3 in this case).

NOTE: It is important to leave a blank row (row 4 by default) above the variable names so that Inference can unambiguously determine where the value content of the DataFrame starts.

6. Save the Data Frame Document

Save this Data Frame document, which you will use again in other QuickStart tutorials.