Pairwise Correlation of Variables

Correlation measures how two variables behave with respect to each other. The key statistic is the *correlation coefficient,* a number that captures both the *direction* and *magnitude* of the relationship as a number that ranges from -1 to +1.

The direction describes the behavior of the two variables. If the variables increase or decrease in unison, then the variables are "directly" correlated and the correlation coefficient is a positive number. For example, energy consumption during hot summer days is directly correlated to temperature because, as things get warmer, air conditioners have to use more power to keep things cool.

Conversely, when the values of the variables move in opposite directions, then the variables are "inversely" correlated. Variables that are inversely correlated have a negative correlation coefficient. For example, the law of supply and demand is based on the inverse relationship between the price of something and how many of those things people can actually afford to buy.

In addition to the direction, there's also the magnitude of the correlation, which ranges from:
+1 = The variables are perfectly correlated.
0 = The variables are uncorrelated.
-1 = The variables are in perfect inverse correlation.

This analytic correlates two or more numeric vectors, along with an optional column for labeling each observation (or row). After processing those inputs, the analytic generates two types of outputs:

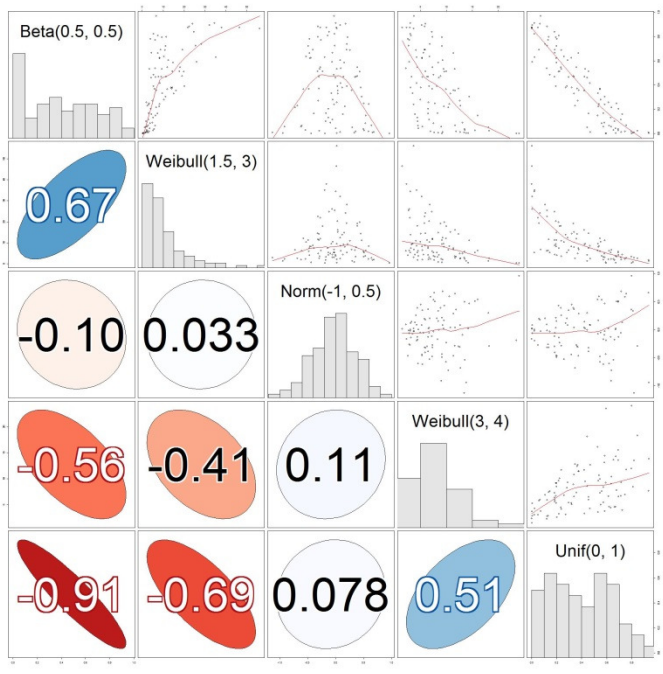
- Correlation Plot
- Correlation Table

Correlation Plot:

The graph to the right is a correlation plot of five variables. This 5x5 matrix contains ten correlations, each corresponding one of the possible pairs of five variables when taken two at a time. This plot has the following features:

- **Main diagonal**: A histogram showing the distribution of values for each variable. The variable names are also shown and indicate that this variable is paired in that row and column.
- **Lower triangle:** Value for the correlation coefficient of the pair of variables, and a color-coded oval as a visual cue to the direction and magnitude of the correlation.
- **Upper triangle:** Scatterplot of the values for pair of variables with the best fit curve shown in red.

The variables are ordered using the first principal component. For more details, see [3].



Correlation Table:

A table containing the correlation coefficient for each pair of variables. For n variables, the number of pairs is n*(n-1)/2. This table includes a "windowing" feature which allows for a moving correlation of values. This table is persisted in two different types of files:

- **CSV:** This file contains the table of pairwise correlations as comma-separated-value text that can be imported to MicroStrategy using the Data Import feature.
- **XML:** This file contains the table of pairwise correlations in XML format that can be sourced to MicroStrategy using the xQuery code provided below.

Note: Only numeric variables are allowed; variables with string values are ignored.

This is the Correlation Table for the Correlation Plot on the previous page:

Index	Label	A	B	A-B	Corr
101	ALL	Unif(0, 1)	Norm(-1, 0.5)	Unif(0, 1)-Norm(-1, 0.5)	0.03
101	ALL	Unif(0, 1)	Weibull(1.5, 3)	Unif(0, 1)-Weibull(1.5, 3)	-0.69
101	ALL	Unif(0, 1)	Weibull(3, 4)	Unif(0, 1)-Weibull(3, 4)	0.65
101	ALL	Unif(0, 1)	Beta(0.5, 0.5)	Unif(0, 1)-Beta(0.5, 0.5)	-0.89
101	ALL	Norm(-1, 0.5)	Weibull(1.5, 3)	Norm(-1, 0.5)-Weibull(1.5, 3)	-0.06
101	ALL	Norm(-1, 0.5)	Weibull(3, 4)	Norm(-1, 0.5)-Weibull(3, 4)	0.04
101	ALL	Norm(-1, 0.5)	Beta(0.5, 0.5)	Norm(-1, 0.5)-Beta(0.5, 0.5)	0.04
101	ALL	Weibull(1.5, 3)	Weibull(3, 4)	Weibull(1.5, 3)-Weibull(3, 4)	-0.52
101	ALL	Weibull(1.5, 3)	Beta(0.5, 0.5)	Weibull(1.5, 3)-Beta(0.5, 0.5)	0.68
101	ALL	Weibull(3, 4)	Beta(0.5, 0.5)	Weibull(3, 4)-Beta(0.5, 0.5)	-0.62

How to Deploy to MicroStrategy:

Prerequisite: Please follow the instructions in the R Integration Pack User Guide [1] for configuring your MicroStrategy environment with R and that the R Script functions have been installed in your MicroStrategy project(s).

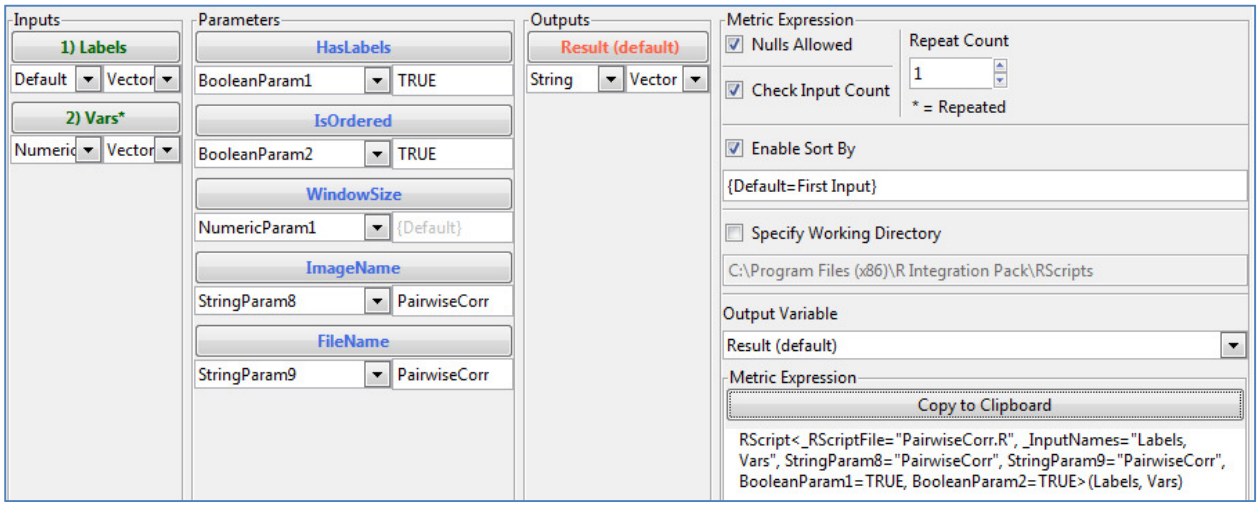
1) Download the PairwiseCorr.R file from the R Script Shelf located at http://RIntegrationPack.CodePlex.com [2] and place it in the Centralized RScripts Repository, the RScripts folder where the R Integration Pack is installed (for example, the Windows folder is usually found at C:\Program Files (x86)\R Integration Pack\RScripts).
2) From the R console, run the PairwiseCorr.R script to verify the script runs correctly. For details, see the "**Running from the R Console**" section below.
3) Cut-and-paste the metric expression below in any metric editor.
4) Use the new metric in reports, dashboards and documents. This metric returns "Ok "if the script executed without issue. If null (empty) values are returned, that's an indication of an error and, as described in the Troubleshooting section below, you should check the error log for details.

Metric Expression:

1) **Result**: Returns Ok if the script executed without problem.

```
RScript<_RScriptFile="PairwiseCorr.R", _InputNames="Labels, Vars",
StringParam8="PairwiseCorr", StringParam9="PairwiseCorr", BooleanParam1=TRUE>(Labels,
Vars)
```

Analytic Signature:



Inputs:

Labels: (optional) String or numeric value to be associated with each observation (i.e., row labels). NOTE: If the parameter HasLabels is "FALSE" then this input will be treated as one of the variables to be included in the pairwise correlations.

Vars: One or more numeric metrics representing the variables to be correlated pairwise.

Parameters:

HasLabels: Uses BooleanParam1 with a default of "TRUE". This parameter indicates whether the first input contains labels and therefore this variable will not be included in the pairwise correlations. If "FALSE" then the first input will be includes with the other inputs for pairwise correlation.

WindowSize: Uses NumericParam1 with a default of 0. This parameter specifies the number of observations (or rows) to be included in the moving window. For more details, see the "**Correlation Table**" section below.

ImageName: Uses StringParam9 with a default of "PairwiseCorr". This parameter specifies the file name (and optionally the path) to use for the image files saved by the script. Please note the R Script automatically appends the appropriate file extensions to this name.

FileName: Uses StringParam9 with a default of "PairwiseCorr". This parameter specifies the file name (and optionally the path) to use for the non-image files saved by the script. Please note the R Script automatically appends the appropriate file extensions to this name.

Output:

Result: A string of "Ok" when the analytic completes successfully. Anything else is an indicator that an error has occurred, see the "**Troubleshooting**" section below for more details.

Additional Results Generated by the R Script:

Four files are stored in the RScripts folder, or the working directory if one is specified by the _WorkingDir function parameter:

.Rdata File: This file persists the state of several objects from the R environment for later inspection and analysis, including df (the data frame of the variables and their values), matCorr (the correlation matrix for all the data), t (the table of pairwise correlations) and timerXeq (the elapsed time of the analytic execution).

.jpg File: This is the image containing the correlation plot.

.csv File: This is a CSV representation of the pairwise correlation table.

.xml File: This is an XML representation of the pairwise correlation table.

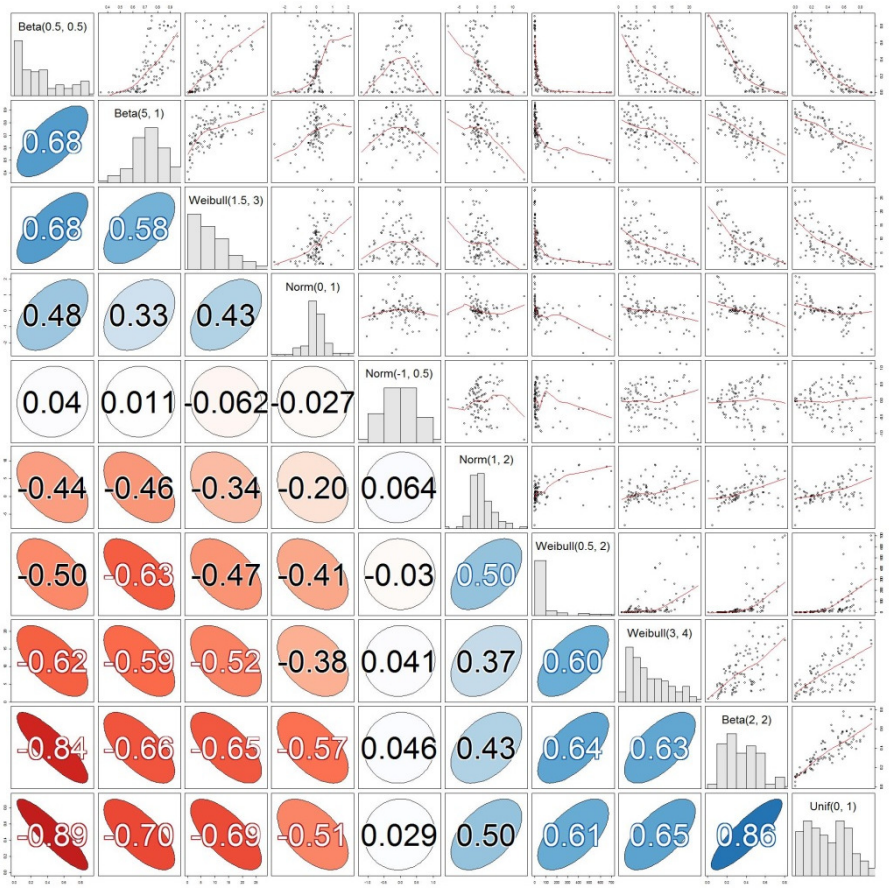
Running from the R Console:

In addition to processing data from MicroStrategy during execution of a report or dashboard, the R script is also configured to run from the R console. Running the script for the R Console verifies that the script is functioning as expected, a good practice when initially deploying this analytic to a new system (for more details, see "Configuring dual execution modes" in [1]).

When executed from the R Console, ten variables with 100 values each are created, each taking random values from these probability distributions: Uniform (1 variable), Normal (3), Weibull (3) and Beta (3):

row.names	Label	Unif(0, 1)	Norm(0, 1)	Norm(1, 2)	Norm(-1, 0.5)	Weibull(0.5, 2)	Weibull(1.5, 3)	Weibull(3, 4)	Beta(0.5, 0.5)	Beta(2, 2)	Beta(5, 1)
1	1	0.800694819	0.9287042	13.3169337	-1.194813798	5.992018e+02	0.6225527	22.3825248	6.494639e-06	0.48583866	0.3476878
2	2	0.597028263	-1.2225723	4.9134824	1.110459283	5.326004e+02	0.1678571	19.1534058	6.344624e-04	0.73301706	0.4097118
3	3	0.593954036	-1.6249641	-3.0247186	-1.028450990	6.859846e+02	1.8294931	21.3353268	9.858570e-04	0.77872882	0.4766642
4	4	0.674064677	-0.7611568	7.5269197	-0.237097866	1.893652e+02	1.1958936	7.6048777	1.307449e-03	0.71894282	0.4405798
5	5	0.937953933	-2.4291243	8.3882464	-0.121156757	6.612266e+02	0.7904429	16.3553780	2.209612e-03	0.69758049	0.4985203
96	96	0.004150391	-0.1140044	0.6386488	-0.016726109	3.357015e-02	23.1632386	4.9161504	5.749226e-01	0.10287223	0.8664376
97	97	0.005596240	2.1655071	-0.8499076	-0.020448332	1.091829e-01	22.5767614	4.4143736	7.678799e-01	0.10367584	0.7218555
98	98	0.002193841	0.3790224	2.5614458	-0.004947477	1.901753e-02	27.5234042	0.9569582	7.855690e-01	0.10750414	0.9081694
99	99	0.001362451	-0.2232857	-8.2335770	0.030730421	2.732656e-02	17.5504205	1.6782930	8.017898e-01	0.05053995	0.8822036
100	100	0.000205840	-1.1747865	4.2919827	0.072397724	5.902744e-04	13.7948414	0.8101073	9.206790e-01	0.01094106	0.7534951

When run from the R Console, if the script is executing properly, a "Success!" message with the elapsed execution time will appear in the console and the image shown below will be plotted. If a "Success!" message does not appear, please consult the **Troubleshooting** section below. Below is the expected Correlation Plot.



In addition to the Correlation Plot, the Correlation Table of pairwise correlations is generated.

For example, here's the correlation table for when the script is run from the R Console:

Index	Label	A	B	A-B	Corr
10	10	Unif(0, 1)	Norm(0, 1)	Unif(0, 1)-Norm(0, 1)	-0.006276817
10	10	Unif(0, 1)	Norm(1, 2)	Unif(0, 1)-Norm(1, 2)	0.256921767
10	10	Unif(0, 1)	Norm(-1, 0.5)	Unif(0, 1)-Norm(-1, 0.5)	0.239490359
10	10	Unif(0, 1)	Weibull(0.5, 2)	Unif(0, 1)-Weibull(0.5, 2)	0.340534167
10	10	Unif(0, 1)	Weibull(1.5, 3)	Unif(0, 1)-Weibull(1.5, 3)	-0.115722762
101	ALL	Weibull(3, 4)	Beta(2, 2)	Weibull(3, 4)-Beta(2, 2)	0.634697026
101	ALL	Weibull(3, 4)	Beta(5, 1)	Weibull(3, 4)-Beta(5, 1)	-0.585596968
101	ALL	Beta(0.5, 0.5)	Beta(2, 2)	Beta(0.5, 0.5)-Beta(2, 2)	-0.843307885
101	ALL	Beta(0.5, 0.5)	Beta(5, 1)	Beta(0.5, 0.5)-Beta(5, 1)	0.678703043
101	ALL	Beta(2, 2)	Beta(5, 1)	Beta(2, 2)-Beta(5, 1)	-0.660820965

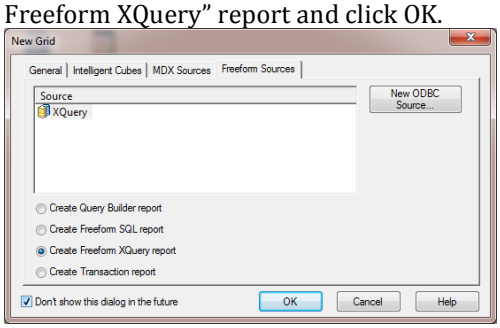
There are 4,140 rows in this table, which can be explained as follows:

- With 10 variables, there are 45 pairs ((n*(n-1)/2, where n=10).
- There were 100 rows and the "WindowSize" for this analysis was 10. So, there are 91 "windows" where the first contains rows 1 – 10, the second rows 2 – 11, etc. The last window is 91 – 100.
- There's a final "ALL" set of 45 pairs with the correlation using all the data.

So, 91 "windows" plus 1 "ALL" times 45 pairs = (91+1) * 45 = 4140.

The table above is persisted in two formats:

- **Comma-separated-value (CSV):** The CSV can be used with Data Import or saved as a table in the warehouse and modeled in metadata.
- **XML:** The XML can be read into a MicroStrategy report using the Free Form Report feature:
  - From MicroStrategy Developer, create a new report, and click on the tab Freeform Sources.
  - Then, choose the radio button "Create Freeform XQuery" report and click OK.



  - When in design mode, select the menu "Data – Free Form SQL Definition" and use the query shown here (NOTE: When viewing this document electronically, the XQuery text shown here can be copied and pasted directly into the Freeform Editor).

Database Instance: XQuery

```
let $xml := fn:doc('file:///C:/PairwiseCorr_console.xml')
return
<Table>
 <ColumnHeaders>
  <ColumnHeader name='A' type='xsd:string' />
  <ColumnHeader name='B' type='xsd:string' />
  <ColumnHeader name='A-B' type='xsd:string' />
  <ColumnHeader name='Index' type='xsd:decimal' />
  <ColumnHeader name='Label' type='xsd:string' />
  <ColumnHeader name='Correlation' type='xsd:decimal' />
 </ColumnHeaders>
 <Data>
 {for $pair in $xml/Pairs/Pair
  return
  <Row>
   <A>{fn:data($pair/@A)}</A>
   <B>{fn:data($pair/@B)}</B>
   <A-B>{fn:data($pair/@A-B)}</A-B>
   <Index>{fn:data($pair/@Index)}</Index>
   <Label>{fn:data($pair/@Label)}</Label>
   <Correlation>{fn:data($pair/@Corr)}</Correlation>
  </Row>}
 </Data>
</Table>
```

Column	Object	Form	Type
1	A	ID	Text
2	B	ID	Text
3	A-B	ID	Text
4	Index	ID	Number
5	Label	ID	Text
6	Correlation		Number

Troubleshooting:

This section covers certain situations you might encounter but it's not intended as a comprehensive list of possible errors.

1) If an error occurs, the report may fail with an error message, or nulls returned as the output. In these cases, please refer to the RScriptErrors.log file generated for further guidance and the DSSErrors.log. Please consult the User Guide [1] and the R documentation for additional guidance.

2) The script will attempt to install two required R packages:
   XML: This R package is used to persist the Correlation Table as XML.
   ellipse: This R package is used to draw ellipses in the Correlation Plot.
   If a package is not successfully installed, you can install using the R console using these commands:
   ```
   install.pacakges("XML", repos="http://cran.rstudio.com/")
   install.pacakges("ellipse", repos="http://cran.rstudio.com/")
   ```

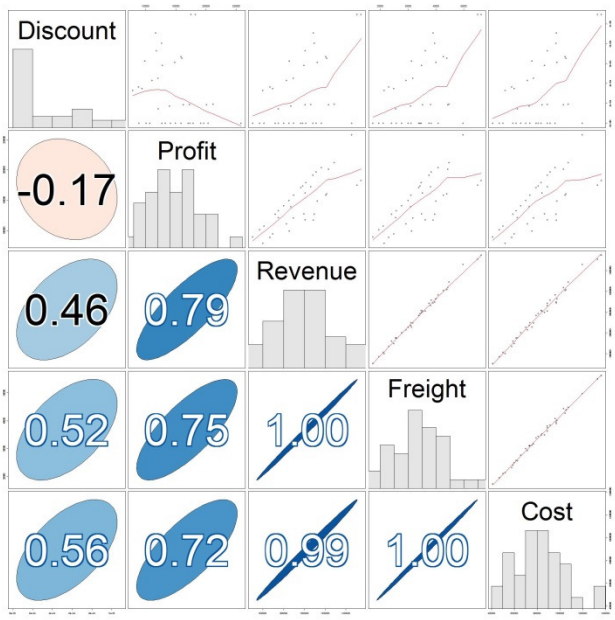
Example:

Using MicroStrategy Tutorial, create a report with Month on Rows and the following metrics on Columns (from `\Public Objects\Metrics\Sales Metrics`): Cost, Discount, Freight, Revenue and Profit.

Next, Insert-New Metric with this expression:

```
RScript<[BooleanParam1]=True,
[NumericParam1]=24,
[StringParam8]="PairwiseCorr",
[StringParam9]="PairwiseCorr",
[_RScriptFile]="PairwiseCorr.R",
[_InputNames]="Month, Cost, Discount,
Freight, Revenue, Profit",
SortBy=[Month]>(Max(Month@DESC) {~},
Cost, Discount, Freight, Revenue, Profit)
```

If successful, "Ok" should appear in the new metric's column, and this plot will be in the R Script's folder.



References:

1) MicroStrategy R Integration Pack User Guide:
   https://rintegrationpack.codeplex.com/documentation
2) R Integration Pack Wiki:
   https://home.microstrategy.com/wiki/Pages/Analytics_RIntegrationPack_SE.aspx
3) Michael Friendly (2002). *Corrgrams: Exploratory displays for correlation matrices*. The American Statistician, 56, 316–324.
   http://www.datavis.ca/papers/corrgram.pdf

