# corVis: An R Package for Visualising Associations and Conditional Associations

*by Amit Chinwan and Catherine Hurley*

**Abstract** Correlation matrix displays are important tools to explore multivariate datasets. These displays with other measures of association can summarize interesting patterns to an analyst and assist them in framing questions while performing exploratory data analysis. In this paper, we present new visualisation techniques to visualise association between all the variable pairs in a dataset in a single plot, which is something existing displays lack. Also, we propse new methods to visualise relationship among variable pairs using conditioning. We use different layouts like matrix or linear for our displays. We use seriation in our displays which helps in highlighting interesting patterns easily. The R package `corVis` provides an implementation.

## Section 1: Introduction

Correlation matrix display is a popular tool to visually explore correlations among variables while performing Exploratory Data Analysis (EDA) on a multivariate dataset. Popularized by Friendly (2002) as corrgram, these displays are produced by first calculating the correlation among the variables and then plotting these calculated values in a matrix display. With effective ordering techniques, these displays quickly highlight variables which are highly correlated and an analyst interested in building a predictive model could use these displays to remove correlated variables and avoid multicollinearity.

The correlation displays are generally used with one of the Pearson's, Spearman's or Kendall's correlation coefficient and are therefore limited to quantitative variables. An analyst can use one-hot encoding of the qualitative variables in order to use these displays but will need to deal with the high dimensions as a result of the encoding. In addition to the dimensionality problem, it is not easy to assess the overall correlation when using the one-hot encoding. The existing methods to quickly explore association among qualitative variables in a dataset include using proportions or counts with different graphical displays like boxplots or barplots. Using association measures for qualitative pairs similar to correlation for quantitative pairs will help in summarizing the relationship, which then can be displayed like the correlation displays.

Tukey and Tukey introduced scagnostics which are measures for scatterplots (Tukey and Tukey, 1985). Along with scagnostics, they proposed a scagnostics scatterplot matrix which is a visual display to explore and compare these measures for all the variable pairs in a dataset. By comparing multiple measures at once, the unusual variable pairs could be identified and looked at in more detail. In a similar manner, a display comparing association measures will help in finding interesting variable pairs. Many association measures have been proposed to summarize different types of relationships. The most commonly used measure is Pearson's correlation coefficient which captures any linear trend present between the variables. Other popular measures include Kendall's or Spearman's rank correlation coefficient which are non-parametric measures and looks for monotonic relationship. Distance correlation (Székely et al., 2007) is an important measure useful in exploring non-linear relationships. The information theory measure maximal information coefficient (MIC) (Reshef et al., 2011) is capable of summarizing complex relationships. With effective displaying techniques, the multiple measures of association provide a comparison tool that assist an analyst to reveal structure present in the data.

Small multiples (or Trellis display) is a simple yet powerful approach to compare partitions of data and understand multidimensional datasets (Tufte, 1986). The display is produced by splitting the data into groups by a conditioning variable and then plotting the data for each group. Such displays allow analysts to quickly infer about the impact of the conditioning variable. A similar idea applied to displays of association measures (correlation plot) will help uncover underlying patterns in the data. One such pattern is Simpson's paradox which can be detected by comparing Pearson's correlation for data at overall level versus individual levels of the conditioning variable.

In this paper, we propose extensions of the correlation plot and new visualizations which look at variables of mixed type, multiple association measures and conditional associations. These displays are implemented in the R package **corVis**. The next section provides a review of existing packages which deal with correlation displays and a quick background on association measures and the packages used for calculating them. Then we describe our approach to calculate the association measures, followed by visualizations of associations and conditional associations. We conclude with a summary and future work.

**Table 1:** List of the R packages dealing with correlation or correlation displays with information on whether the plots are seriated, interactive and are useful for high dimensions.

| package | display | displayType | seriatedPlot | interactivePlot | highDimension |
|---|---|---|---|---|---|
| coreheat | heatmap | overlay | No | No | No |
| corrplot | heatmap | overlay | Yes | No | No |
| corrr | heatmap/network | facet/overlay | Yes | No | Yes |
| corrgrapher | network | overlay | No | No | Yes |
| corrarray | no display | none | No | No | No |
| correlationfunnel | funnel | overlay | No | No | Yes |
| linkspotter | network | overlay | No | Yes | Yes |
| correlation | heatmap/network | overlay | No | No | Yes |

## Section 2: Background

In this section we provide a brief review of the existing packages used for correlation displays and the association measures. The correlation matrix display is an important tool to explore association among variables in a multivariate analysis. The display was made popular by Friendly (2002) who called them corrgrams, wherein he rendered the correlation values with shaded squares, bars, ellipses, or circular 'pac-man' symbols. The main goal of these displays is to describe the bivariate patterns in a dataset.

### Section 2.1: Literature Review on Correlation Displays

There are various packages available in R which can be used to create correlation matrix displays. The R package **corrplot** (Wei and Simko, 2021) provides an implementation of the Friendly (2002) paper. It serves as a visual exploratory tool for correlation matrices and includes various variable ordering methods which assist in detecting patterns of relations among the variables. The package **corrr** (Kuhn et al., 2020) is useful in calculating as well as visualising correlations. It calculates a correlation `dataframe` for a dataset and hence is easier to focus on the correlations of certain variables. The package provide different ways to produce correlation displays. In addition to matrix display, the package also plots the correlation values in a network display which is useful when dealing with high-dimensional datasets. Table @ref(tab:corrdisplay_packages) provides a list of the packages available in R for either calculating correlations, visualising correlations or both.

The majority of the packages in Table @ref(tab:corrdisplay_packages) focus only on quantitative variables of the dataset. The packages **correlationfunnel** (Dancho, 2020) and **linkspotter** (Samba, 2020) look at both the quantitative as well as the qualitative variables during the correlation analysis. **correlationfunnel** speeds up the feature selection step by looking at the relationship of predictors to a response (or a target). It converts numeric predictors into factors, applies one-hot encoding to all the factors and then calculates and visualise the Pearson's correlation coefficient among the response and the predictors. On the other hand, **linkspotter** calculates different association measures for the quantitative and qualitative variables and then uses a network plot to visualise these association. We believe that using different association measures is a better approach than using one-hot encoding for exploring all the variables, as the encoding increases the dimensions very quickly and it is not easy to assess the overall correlation.

There have been other extensions to correlation displays which are useful when dealing with high dimensional datasets. Buja et al. (2016) proposed Association Navigator which is an interactive visualization tool for large correlation matrices with upto 2000 variables. They also provide different functionalities including highlighting variables, plotting scatterplot for variable pairs. The R package **scorrplot2022** (McKenna et al., 2016) produces an interactive scatterplot for exploring pairwise correlations in a large dataset by projecting variables as points with respect to some user-selected variables on a scatterplot, driven by geometric interpretation of correlation. A user can update variables of interest and can create tours of the correlation space between different projections of the data using this tool. In this paper, we do not focus on the high dimensionality aspect of the correlation displays like these tools but introduce conditional association displays which are useful in uncovering conditional structure present in the data. Conditional displays (also called Small multiple displays or Trellis displays) are visualisations for the subsets of data produced by dividing the data by a partitioning variable and then plotting them. Popularised by Tufte (1986) and Becker et al. (1996), these displays are efficient for discovering interesting patterns in the data.

**Table 2:** List of the functions available in the package for calculating different association measures along with the packages used for calculation.

| funName | typeX | typeY | from | symmetric | min | max |
|---|---|---|---|---|---|---|
| tbl_cor | numeric | numeric | stats::cor | TRUE | -1 | 1 |
| tbl_dcor | numeric | numeric | energy::dcor2d | TRUE | 0 | 1 |
| tbl_mine | numeric | numeric | minerva::mine | TRUE | 0 | 1 |
| tbl_polycor | ordinal | ordinal | polycor::polychor | TRUE | -1 | 1 |
| tbl_tau | ordinal | ordinal | DescTools::KendalTauA,B,C,W | TRUE | -1 | 1 |
| tbl_gkTau | nominal | nominal | DescTools::GoodmanKruskalTau | FALSE | 0 | 1 |
| tbl_gkLambda | nominal | nominal | DescTools::GoodmanKruskalTau | TRUE | 0 | 1 |
| tbl_gkGamma | nominal | nominal | DescTools::GoodmanKruskalTau | TRUE | 0 | 1 |
| tbl_uncertainty | nominal | nominal | DescTools::UncertCoef | TRUE | 0 | 1 |
| tbl_chi | nominal | nominal | DescTools::ContCoef | TRUE | 0 | 1 |
| tbl_cancor | nominal | nominal | corVis | TRUE | 0 | 1 |
| tbl_cancor | nominal | numerical | corVis | TRUE | 0 | 1 |
| tbl_nmi | any | any | corVis | TRUE | 0 | 1 |
| tbl_easy | any | any | correlation::correlation | TRUE | -1 | 1 |

### Section 2.2: Literature Review on Association Measures

An association measure can be defined as a numerical summary quantifying relationship between two or more variables. For example, Pearson's correlation coefficient summarizes the strength and direction of the linear relationship present between two numeric variables and is in the range $[-1, 1]$. Kendall's or Spearman's rank correlation coefficient are other popular measures which look for the presence of montonic relationship among two numeric variables and are in the range $[-1, 1]$. As these measures are limited to linear or monotonic relationships, there's a need to use association measures which can capture complex relationships (like non-linear or periodic). The distance correlation coefficient (Székely et al., 2007) is one such measure which looks for the non-linear association between two numeric variables and summarizes it in $[0, 1]$. Similarly, MIC (Reshef et al., 2011) is capable of summarizing non-linear as well as periodic relationships between numeric variables. In addition to association measures for numeric variables, association measures for ordinal or nominal or mixed variables will help an analyst in exploring a multivariate dataset. Taha and Hadi (2016) provides an overview of the association measures used for categorical or mixed data.

Tukey and Tukey (1985) proposed scatterplot matrix of the scagnostics measures, which are measures summarizing a scatterplot. They suggested that scatterplot matrix of the measures can be used to identify unusual scatterplots or variable pairs. Wilkinson et al. (2005) used this idea with their graph-theoretic scagnostic measures to highlight unusual scatterplot. Similarly, Kuhn et al. (2013) have used this idea in a predictive modeling context. They have produced a scatterplot matrix of the measures between the response and continuous predictors such as Pearson's correlation coefficient, pseudo-$R^2$ from the locally weighted regression model, MIC and Spearman's rank correlation coefficient to explore the predictor importance during feature selection step. These displays show the importance of comparing multiple association measures at once for different variable pairs. In this paper, we propose different visualization techniques to compare multiple association measures for all the variable pairs in a dataset which can assist a user in finding interesting patterns.

## Section 3: corVis: Calculating Association

This section provides an implementation for the calculation of association measures in our package **corVis**. The package provides a collection of various measures of association which is used to quantify the relationship between two variables and is used to explore patterns during EDA. The measures available in the package are not limited to numeric variables and are used with categorical and ordinal variables as well. Table @ref(tab:association_measures) lists the different measures of association provided in the package with the variable types they can be used with, the package used for calculation, the information on whether the measure is symmetric, and the minimum and maximum value of the measure.

### Calculating association for relevant variables

We introduce a method which creates a tibble structure for the variable pairs in a dataset along with calculated association measure. The package contains various functions (shown in Table 1) for different association measures in the form `tbl_*` to calculate them. For example, a user might be interested in calculating distance correlation for numeric pair of variables in a dataset. This can be done by using `tbl_dcor`.

```
df <- penguins
distance <- tbl_dcor(df)
head(distance)
```

```
#> # A tibble: 6 x 4
#>   x                 y              measure measure_type
#>   <chr>             <chr>            <dbl> <chr>
#> 1 bill_depth_mm     bill_length_mm  0.387  dcor
#> 2 flipper_length_mm bill_length_mm  0.666  dcor
#> 3 body_mass_g       bill_length_mm  0.587  dcor
#> 4 year              bill_length_mm  0.0784 dcor
#> 5 flipper_length_mm bill_depth_mm   0.704  dcor
#> 6 body_mass_g       bill_depth_mm   0.614  dcor
```

Similarly, one can use `tbl_nmi` to calculate normalised mutual information for numeric, nominal and mixed pair of variables.

```
nmi <- tbl_nmi(df)
head(nmi)
```

```
#> # A tibble: 6 x 4
#>   x                 y         measure measure_type
#>   <chr>             <chr>       <dbl> <chr>
#> 1 island            species 0.507     nmi
#> 2 bill_length_mm    species 0.353     nmi
#> 3 bill_depth_mm     species 0.315     nmi
#> 4 flipper_length_mm species 0.343     nmi
#> 5 body_mass_g       species 0.300     nmi
#> 6 sex               species 0.0000854 nmi
```

These functions return a tibble with the variable pairs and calculated measure, and also with additional classes `pairwise` and `data.frame`. With the pairwise measures of association in a tibble or dataframe structure, the output of these functions can then be used with packages like `dplyr` , `ggplot2` for further exploration of association measures.

```
class(distance)
```

```
#> [1] "pairwise"   "tbl_df"     "tbl"        "data.frame"
```

The function `matrix_assoc` helps in converting the tibble of association measure to matrix structure. The function takes a tibble or dataframe of the variable pairs of the dataset along with the calculated association measures and returns a symmetric matrix of the variables.

```
head(matrix_assoc(distance))
```

```
#>                   bill_length_mm bill_depth_mm flipper_length_mm body_mass_g
#> bill_length_mm                NA     0.3872021         0.6664558   0.5871319
#> bill_depth_mm         0.38720211            NA         0.7039636   0.6141631
#> flipper_length_mm     0.66645577     0.7039636                NA   0.8674122
#> body_mass_g           0.58713186     0.6141631         0.8674122          NA
#> year                  0.07842516     0.1117057         0.1643876   0.0790560
#>                         year
#> bill_length_mm    0.07842516
#> bill_depth_mm     0.11170568
#> flipper_length_mm 0.16438763
#> body_mass_g       0.07905600
#> year                      NA
```

The function outputs a matrix even if any variable pair is missing in the input tibble with `NA` for corresponding variable pair cell in the matrix output.

```
distance <- distance[-1,]
matrix_assoc(distance)
```

```
#>                   bill_length_mm bill_depth_mm flipper_length_mm body_mass_g
#> bill_length_mm                NA            NA         0.6664558   0.5871319
#> bill_depth_mm                 NA            NA         0.7039636   0.6141631
#> flipper_length_mm     0.66645577     0.7039636                NA   0.8674122
#> body_mass_g           0.58713186     0.6141631         0.8674122          NA
#> year                  0.07842516     0.1117057         0.1643876   0.0790560
#>                         year
#> bill_length_mm    0.07842516
#> bill_depth_mm     0.11170568
#> flipper_length_mm 0.16438763
#> body_mass_g       0.07905600
#> year                      NA
```

The function has an additional argument called group which represents the level of the grouping categorical variable for which the matrix output needs to be calculated and is set to overall as default.

### Calculating association measures for whole dataset

calc_assoc can be used to calculate association measures for all the variable pairs in the dataset at once in a tibble structure. In addition to tibble structure, the output also has paiwise and data.frame class which are important class attributes for producing visual summaries in this package.

```
complete_assoc <- calc_assoc(df)
glimpse(complete_assoc)
```

```
#> Rows: 28
#> Columns: 4
#> $ x            <chr> "island", "bill_length_mm", "bill_depth_mm", "flipper_len~
#> $ y            <chr> "species", "species", "species", "species", "species", "s~
#> $ measure      <dbl> 0.81328762, 0.84131393, 0.82447508, 0.88217284, 0.8183348~
#> $ measure_type <chr> "cancor", "cancor", "cancor", "cancor", "cancor", "cancor~
```

```
class(complete_assoc)
```

```
#> [1] "pairwise"   "tbl_df"     "tbl"        "data.frame"
```

The function has a *types* argument which is basically a tibble of the association measure to be calculated for different variable pairs. The default tibble of measures is default_assoc() which calculates Pearson's correlation if both the variables are numeric, Kendall's tau-b if both the variables are ordinal, canonical correlation if one is factor and other is numeric and canonical correlation for the rest of the variable pairs.

```
default_measures <- update_assoc()
default_measures
```

```
#> # A tibble: 4 x 4
#>   funName    typeX   typeY   argList
#>   <chr>      <chr>   <chr>   <list>
#> 1 tbl_cor    numeric numeric <NULL>
#> 2 tbl_tau    ordered ordered <NULL>
#> 3 tbl_cancor factor  numeric <NULL>
#> 4 tbl_cancor other   other   <NULL>
```

An analyst can update these measures using the update_assoc function where one can specify a tbl_* function to calculate association measure depending on the variable pair in the dataset and a method if it calculates more than one measure.

```
updated_assoc <- update_assoc(num_pair = "tbl_cor",
                              num_pair_argList = "spearman",
                              mixed_pair = "tbl_cancor",
                              other_pair = "tbl_nmi")
updated_assoc
```

```
#> # A tibble: 4 x 4
#>   funName    typeX   typeY   argList
#>   <chr>      <chr>   <chr>   <list>
#> 1 tbl_cor    numeric numeric <chr [1]>
#> 2 tbl_tau    ordered ordered <NULL>
#> 3 tbl_cancor factor  numeric <NULL>
#> 4 tbl_nmi    other   other   <NULL>
```

```
updated_complete_assoc <- calc_assoc(df, types = updated_assoc)
head(updated_complete_assoc)
```

```
#> # A tibble: 6 x 4
#>   x                 y         measure measure_type
#>   <chr>             <chr>       <dbl> <chr>
#> 1 island            species 0.507     nmi
#> 2 bill_length_mm    species 0.841     cancor
#> 3 bill_depth_mm     species 0.824     cancor
#> 4 flipper_length_mm species 0.882     cancor
#> 5 body_mass_g       species 0.818     cancor
#> 6 sex               species 0.0000854 nmi
```

The tibble output for calc_assoc has the following structure:

- x and y representing a pair of variables
- measure representing the calculated value for association measure
- measure_type representing the association measure calculated for x and y pair.

The variable pairs in the output are unique pairs and a subset of all the variable pairs of a dataset where x $\neq$ y. As explained earlier, the measure_type represents the association measure calculated for a specific type of variable pair. A user can be interested in calculating multiple association measures for a type of variable pair. This can be done by using the calc_assoc and update_assoc together for calculating different association measures and then merging the output tibbles.

### Calculating conditional association

calc_assoc_by can be used to calculate association measures for all the variable pairs at different levels of a categorical variable. This can help in exploring the conditional associations and find out interesting patterns in the data prior to modeling. The output of this function is a tibble structure with pairwise and data.frame as additional class attributes. The by argument is used for the grouping variable which needs to be categorical.

```
complete_assoc_by <- calc_assoc_by(df,by = "sex")
```

The function also has a types argument which can be updated similarly to calc_assoc.

```
updated_assoc <- update_assoc(num_pair = "tbl_cor",
                              num_pair_argList = "spearman",
                              mixed_pair = "tbl_cancor",
                              other_pair = "tbl_nmi")
updated_complete_assoc_by <- calc_assoc_by(df,by = "sex", types = updated_assoc)
head(updated_complete_assoc_by)
```

```
#> # A tibble: 6 x 5
#>   x                 y       measure measure_type by
#>   <chr>             <chr>     <dbl> <chr>        <fct>
#> 1 island            species  0.502  nmi          female
#> 2 bill_length_mm    species  0.885  cancor       female
#> 3 bill_depth_mm     species  0.900  cancor       female
#> 4 flipper_length_mm species  0.914  cancor       female
#> 5 body_mass_g       species  0.911  cancor       female
#> 6 year              species  0.0457 cancor       female
```

By default, the function calculates the association measures for all the variable pairs at different levels of the grouping variable and the pairwise association measures for the ungrouped data (*overall*). This behavior can be changed by setting include.overall to *FALSE*.

```
complete_assoc_by <- calc_assoc_by(df,by = "sex",include.overall = FALSE)
```

The tibble output for `calc_assoc_by` has the similar structure as `calc_assoc` with an additional `by` column representing the levels of the categorical variable used in the function. The variable pairs in the output are repeated for every level of `by` variable. At present the function doesn't allow multiple `by` variables to be used for conditioning but is something which can be done by using the `calc_assoc_by` function multiple times and then merging the multiple outputs.For calculating multiple measures for a specific variable type, one can use `update_assoc` with `calc_assoc_by` and then can merge these multiple tibble outputs.

## Section 4: corVis: Visualising Association

We propose novel visualisations to display association for every variable pair in a dataset in a single plot and show multiple bivariate measures of association simultaneously to find out interesting patterns. Efficient seriation techniques have been included to order and highlight interesting relationships. These ordered association and conditional association displays can help find interesting patterns in the dataset. While designing these displays we considered matrix-type, linear and network-based layouts. A matrix-type layout simplifies lookup, and different measures may be displayed on the upper and lower diagonal. Linear layouts are more space-efficient than matrix plots, but lookup is more challenging. Variable pairs can be ordered by relevance (usually difference in measures of association or across the factor levels), and less relevant pairs can be omitted. Linear displays are also suitable to display associations between the response and predictors only. Our selection criteria for a better display were based on :

- Number of variables
- Easier pixel-variable or variable-pixel look up
- Number of levels of a factor for conditional association displays

Figure 1 shows this display for every variable pair in the *penguins* dataset from the *palmerpenguins* package. It shows a high positive Pearson's correlation among flipper_length_mm and body_mass_g, flipper_length_mm and bill_length_mm, and bill_length_mm and bodymass_g. There seems to be a strong negative Pearson's correlation between flipper_length_mm and bill_depth_mm, and bill_depth_mm and body_mass_g. The plot also shows that there is a high canonical correlation between species and other variables except year and sex, and a high canonical correlation between island and species, which traditional correlation matrix display would omit as they are limited to numeric variable pairs only. The variables in the display are ordered using average linkage clustering method to find out highly associated variables quickly.

We can also calculate multiple association measures for all the variable pairs in the dataset and compare them. This will help in finding out pairs of variables with a high difference among different measures and one can investigate these bivariate relationships in more detail. The `pairwise_summary_plot` function can be used to compare various measures using the matrix layout. It plots multiple measures among the variable pairs as bars, where each bar represents one measure of association. Figure 2 shows a matrix layout comparing Pearson's and Spearman's correlation coefficient for the numeric variable pairs in *penguins* data.

In addition to matrix layout, we can also use linear layouts for comparing multiple measures. Figure 3 shows a linear layout comparing multiple association measures for all the variable pairs in the penguins data. Linear layouts seems to be more suitable when comparing high number of association measures.

### Visualising Conditional Association

The package includes a function `calc_assoc_by` which calculates the pairwise association at different levels of a categorical conditioning variable. This helps in finding out interesting variable triples which can be explored further prior to modeling. Figure 4 shows a conditional association plot for the *penguins* data. Each cell corresponding to a variable pair shows three bars which correspond to the association measure (Pearson's correlation for numeric pair and Normalized mutual information for other combination of variables) calculated at the levels of conditioning variable *island*. The dashed line represents the overall association measure. The plot shows that there is a high value for normalised mutual information between bill_length_mm and species for the penguins which lived in *Biscoe* island compared to the penguins which lived in *Dream* island. It can also be seen that the cell corresponding to variable pair flipper_length_mm and bill_depth_mm has a high negative overall Pearson's correlation and for the penguins which lived in *Biscoe* island but positive correlation for penguins which lived in

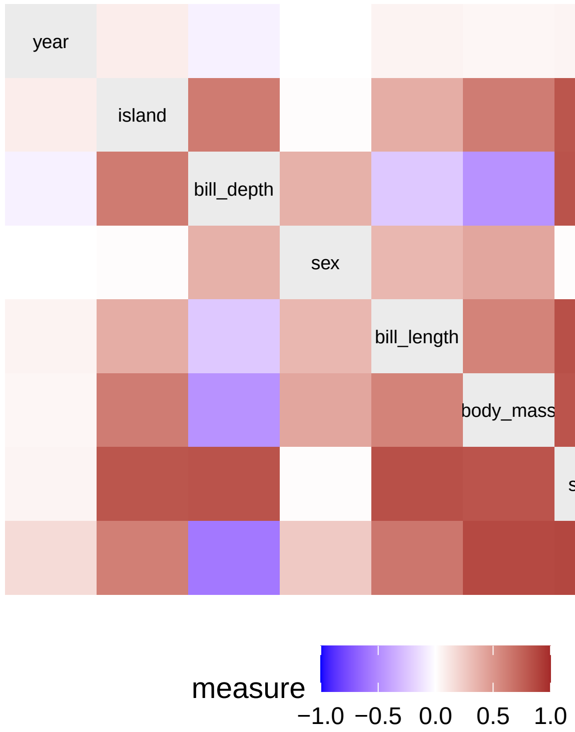

**Figure 1:** Association matrix display for penguins data showing Pearson's correlation for numeric variable pairs, canonical correlation for mixed variable pairs and categorical variable pairs.

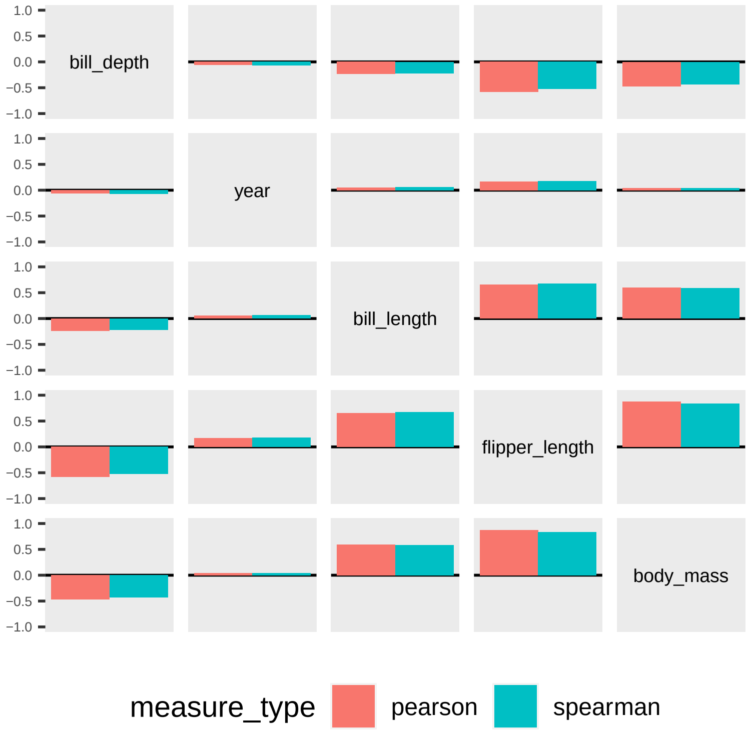

**Figure 2:** Matrix display comparing Pearson's and Spearman's correlation coefficient. All the variable pairs have similar values for both correlations.

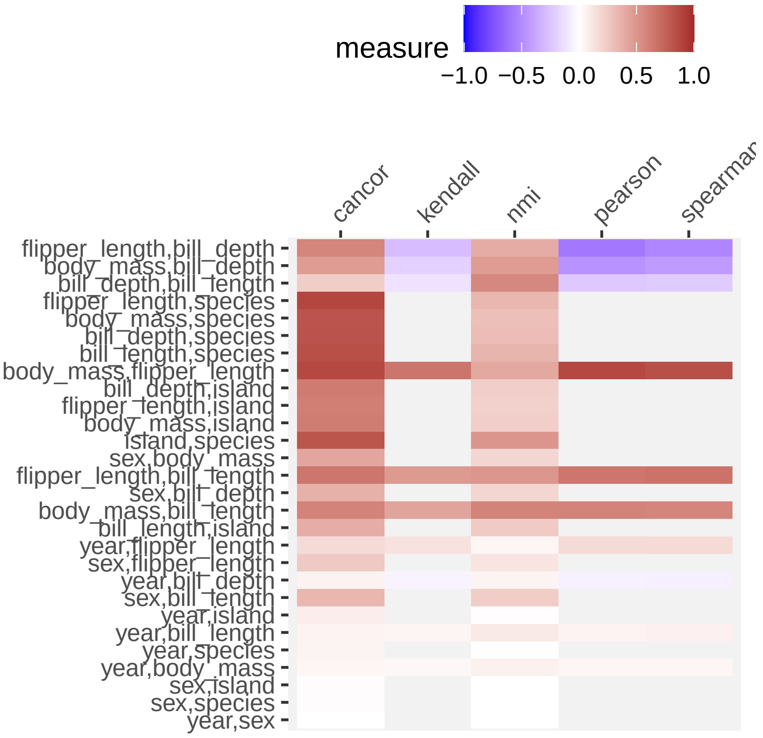

**Figure 3:** Comparing multiple association measures using a linear layout. The display has variable pairs on the Y-axis and association measures on the X-axis. The cell corresponding to a variable pair and an association measure has been colored grey showing that the measure is not defined for corresponding pair.

*Dream* and *Torgersen* island. This is an instance of Simpson's paradox which can be taken into account during the modeling step.

We also provide a functionality for highlighting interesting patterns like Simpson's paradox. Figure 5 shows the matrix plot with highlighted cells for the variable pairs where Simpson's paradox is present.

The cells can also be highlighted on the basis of a score calculated by the user. This can be done by providing a dataframe with pairs of variables to highlight and a score for highlighting variable pairs. The cells with high score will have a thicker border compared to cells with low score. Figure 6 shows highlighted cells on the basis of a score provided for a subset of variable pairs.

We can also use linear layouts for displaying conditional association. Figure 7 shows a funnel-like linear display for conditional association measures with all the variable pairs on the y-axis, the value of association measure on x-axis and color of the points representing the level of the grouping variable. The linear layout becomes more useful over the matrix layout when the number of variables and number of levels of grouping variable are high.

## Bibliography

R. A. Becker, W. S. Cleveland, and M.-J. Shyu. The visual design and control of trellis display. *Journal of Computational and Graphical Statistics*, 5(2):123–155, 1996. doi: 10.1080/10618600.1996.10474701. URL https://www.tandfonline.com/doi/abs/10.1080/10618600.1996.10474701. [p2]

A. Buja, A. M. Krieger, and E. I. George. A visualization tool for mining large correlation tables: The association navigator., 2016. [p2]

M. Dancho. *correlationfunnel: Speed Up Exploratory Data Analysis (EDA) with the Correlation Funnel*, 2020. URL https://CRAN.R-project.org/package=correlationfunnel. R package version 0.2.0. [p2]

M. Friendly. Corrgrams: Exploratory displays for correlation matrices. *The American Statistician*, 56(4): 316–324, 2002. [p1, 2]

M. Kuhn, K. Johnson, et al. *Applied predictive modeling*, volume 26. Springer, 2013. [p3]

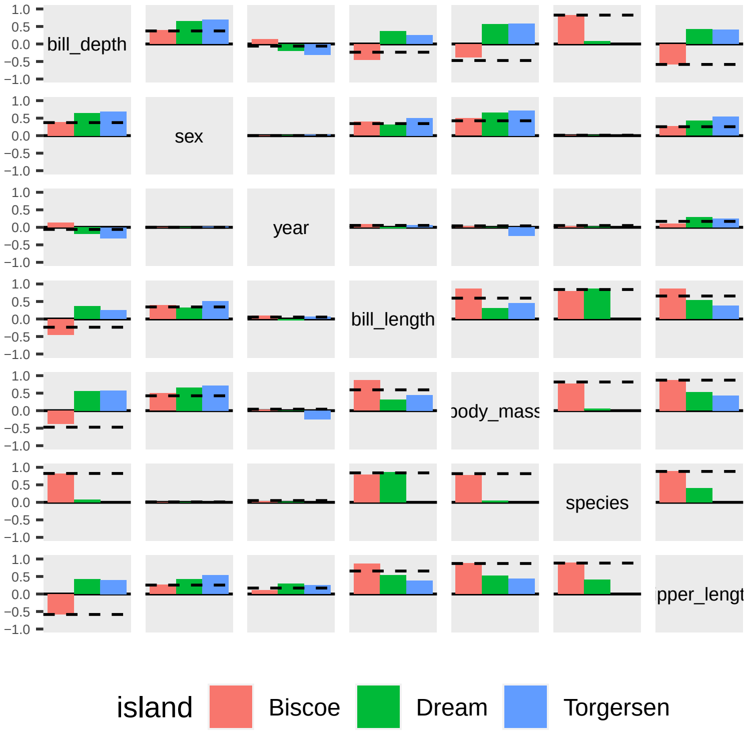

**Figure 4:** Conditional Association plot for penguins data showing Pearson's correlation for numeric pairs and normalised mutual information for categorical or mixed pairs. The bars in each cell represent the value for asssociation measure colored by the conditioning variable 'island'. The dashed line in each cell represents overall value of the association measure.

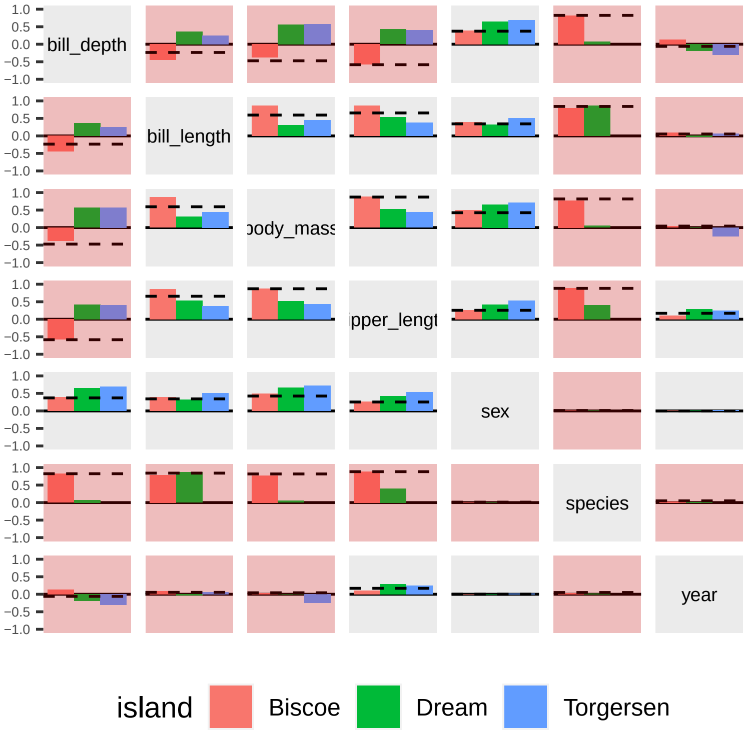

**Figure 5:** Conditional Association plot with examples of Simpson's paradox

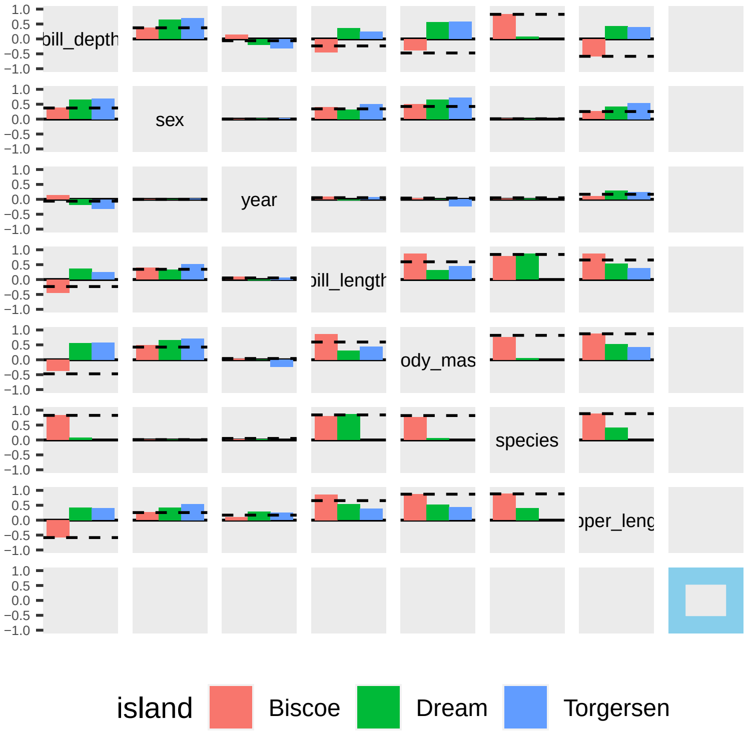

**Figure 6:** Conditional Association plot with manual highlighting

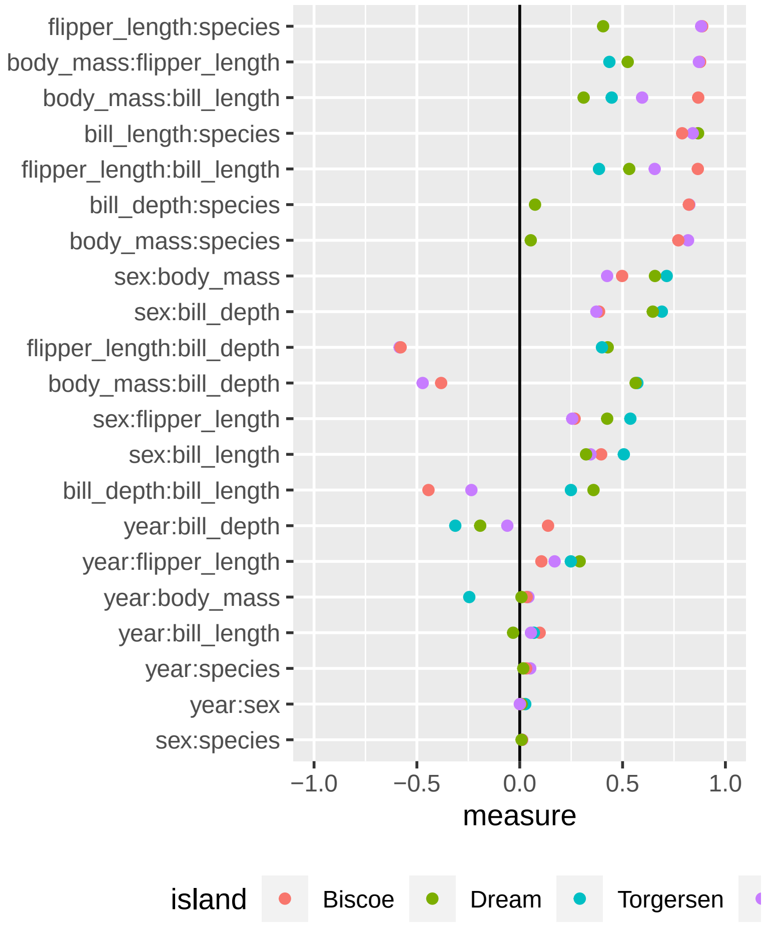

**Figure 7:** Conditional Association plot using linear layout.The display has variable pairs on the Y-axis and the value of association measures on the X-axis. The points corresponding to every variable pair represents the value of association measure for different levels of the conditioning variable and the overall value of association measure.

M. Kuhn, S. Jackson, and J. Cimentada. *corrr: Correlations in R*, 2020. URL https://CRAN.R-project.org/package=corrr. R package version 0.4.3. [p2]

S. McKenna, M. Meyer, C. Gregg, and S. Gerber. s-corrplot: An interactive scatterplot for exploring correlation. *Journal of Computational and Graphical Statistics*, 25(2):445–463, 2016. doi: 10.1080/10618600.2015.1021926. URL https://doi.org/10.1080/10618600.2015.1021926. [p2]

D. N. Reshef, Y. A. Reshef, H. K. Finucane, S. R. Grossman, G. McVean, P. J. Turnbaugh, E. S. Lander, M. Mitzenmacher, and P. C. Sabeti. Detecting novel associations in large data sets. *science*, 334(6062): 1518–1524, 2011. [p1, 3]

A. Samba. *linkspotter: Bivariate Correlations Calculation and Visualization*, 2020. URL https://CRAN.R-project.org/package=linkspotter. R package version 1.3.0. [p2]

G. J. Székely, M. L. Rizzo, and N. K. Bakirov. Measuring and testing dependence by correlation of distances. *The annals of statistics*, 35(6):2769–2794, 2007. [p1, 3]

A. Taha and A. S. Hadi. Pair-wise association measures for categorical and mixed data. *Information Sciences*, 346-347:73–89, 2016. ISSN 0020-0255. doi: https://doi.org/10.1016/j.ins.2016.01.022. URL https://www.sciencedirect.com/science/article/pii/S0020025516000335. [p3]

E. R. Tufte. *The Visual Display of Quantitative Information*. Graphics Press, USA, 1986. ISBN 096139210X. [p1, 2]

J. W. Tukey and P. A. Tukey. Computer graphics and exploratory data analysis: An introduction. In *Proceedings of the sixth annual conference and exposition: computer graphics*, volume 85, pages 773–785, 1985. [p1, 3]

T. Wei and V. Simko. *R package 'corrplot': Visualization of a Correlation Matrix*, 2021. URL https://github.com/taiyun/corrplot. (Version 0.92). [p2]

L. Wilkinson, A. Anand, and R. Grossman. Graph-theoretic scagnostics. In *Information Visualization, IEEE Symposium on*, pages 21–21. IEEE Computer Society, 2005. [p3]

*Amit Chinwan*
*Maynooth University*
*Hamilton Institute*
*Maynooth, Ireland*
amit.chinwan.2019@mumail.ie

*Catherine Hurley*
*Maynooth University*
*Department of Mathematics and Statistics*
*Maynooth, Ireland*
catherine.hurley@mu.ie