Post 01: Correlograms

Exploratory Displays For Correlation Matrices

Xia Liu

10/28/2017

> ## Learning Objectives
>
> - Learn to how to plot a correlogram with R “corrplot” package
> - Familiar youself with correlation analysis
> - Compute correlation matrix with correct information
> - Customize the correlogram by layout, color, order and visualization methods
> - Open to more possibilities (R “corrgram” package)
> - Stay interested in R and more

Introduction

In this post, I will introduce to you R **corrplot** package and how to plot a **correlogram** in R. Correlogram is a graphical display of a correlation matrix, confidence interval or general matrix. It is very useful to highlight the most correlated variables in a data table. In addition, corrplot is good at details, including choosing color, text labels, color labels, layout, etc. In the following tutorial, **correlation coefficients** is colored according to the value. **Correlation matrix** can be also reordered according to the degree of association between variables.

MORE(mathy background): Correlation and covariance matrices provide the basis for all classical multivariate techniques, because (together with mean vectors) they provide sufficient statistics under multivariate normal linear models. Many statistical tools exist for analyzing multivariate structure: principal component analysis, factor analysis, canonical correlation analysis, and so forth. All of these have the goal of reducing high-dimensional multivariate structure to a smaller number of dimensions, so that the relationships among the variables may be more readily apprehended.

So let’s start it!

Install R corrplot package

corrplot package is required to execute the R code in this article.

```
install.packages("corrplot")
```

Data for correlation analysis

We are using **mtcars** data(Motor Trend Car Road Tests) to compute correlation matrix. a built-in data frame with 32 observations on 11 variables, extracted from the 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973–74 models).

mtcars is a data frame with 32 observations on 11 variables. Here’s a data dictionary below.

[, 1] mpg Miles/(US) gallon

[, 2] cyl Number of cylinders

[, 3] disp Displacement (cu.in.)

[, 4] hp Gross horsepower

[, 5] drat Rear axle ratio

[, 6] wt Weight (1000 lbs)

[, 7] qsec 1/4 mile time

[, 8] vs V/S

[, 9] am Transmission (0 = automatic, 1 = manual)

[,10] gear Number of forward gears

[,11] carb Number of carburetors

```
head(mtcars)
```

```
##                    mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

Computing correlation matrix

```
M<-cor(mtcars)
head(round(M,2))
```

```
##        mpg   cyl  disp    hp  drat    wt  qsec    vs    am  gear  carb
## mpg   1.00 -0.85 -0.85 -0.78  0.68 -0.87  0.42  0.66  0.60  0.48 -0.55
## cyl  -0.85  1.00  0.90  0.83 -0.70  0.78 -0.59 -0.81 -0.52 -0.49  0.53
## disp -0.85  0.90  1.00  0.79 -0.71  0.89 -0.43 -0.71 -0.59 -0.56  0.39
## hp   -0.78  0.83  0.79  1.00 -0.45  0.66 -0.71 -0.72 -0.24 -0.13  0.75
## drat  0.68 -0.70 -0.71 -0.45  1.00 -0.71  0.09  0.44  0.71  0.70 -0.09
## wt   -0.87  0.78  0.89  0.66 -0.71  1.00 -0.17 -0.55 -0.69 -0.58  0.43
```

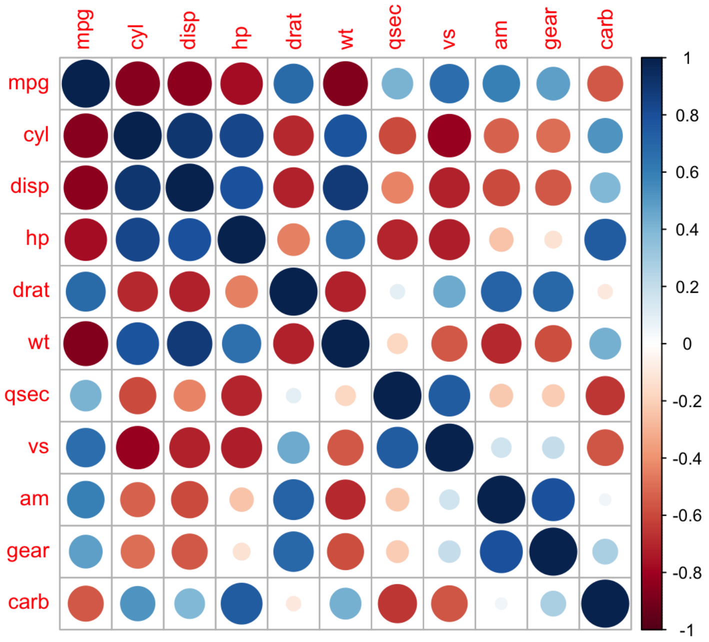
> ## Visualization Method
>
> There are seven visualization methods (parameter method) in corrplot package, named `"circle"`, `"square"`, `"ellipse"`, `"number"`, `"shade"`, `"color"`, `"pie"`.

```
library(corrplot)
```

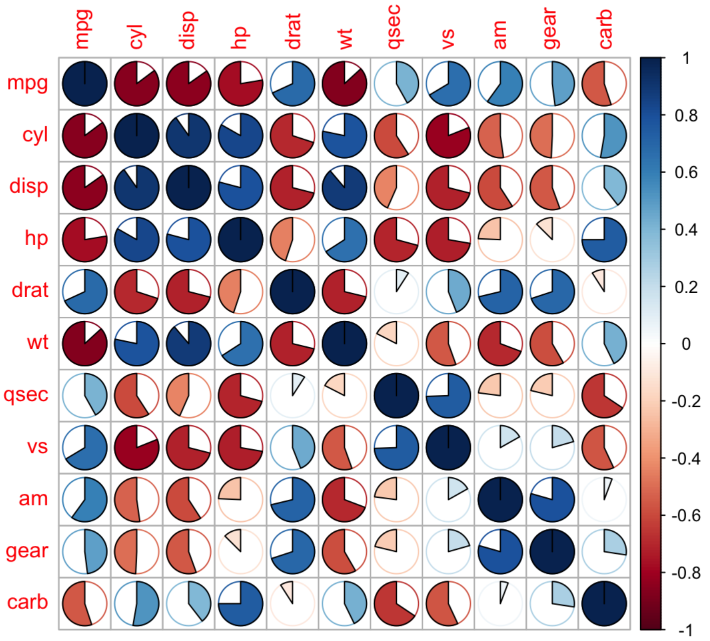
```
## Warning: package 'corrplot' was built under R version 3.4.2
```

```
## corrplot 0.84 loaded
```

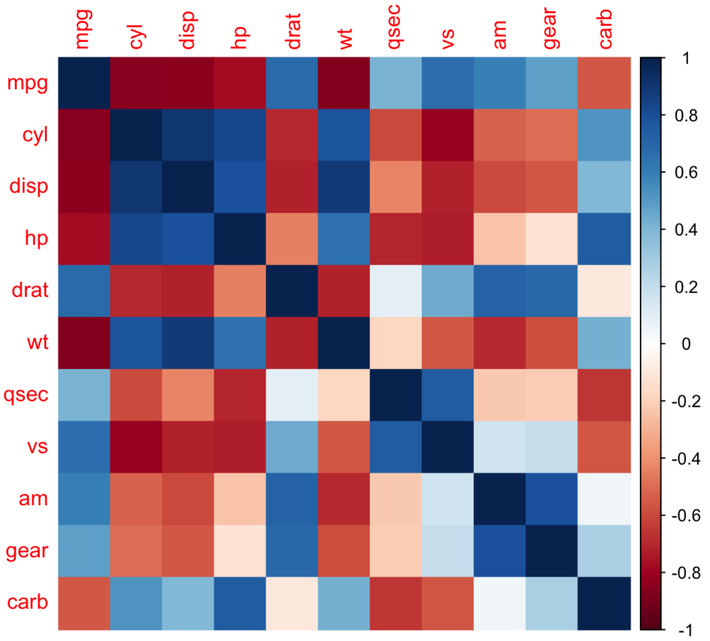
```
corrplot(M, method="circle")
```



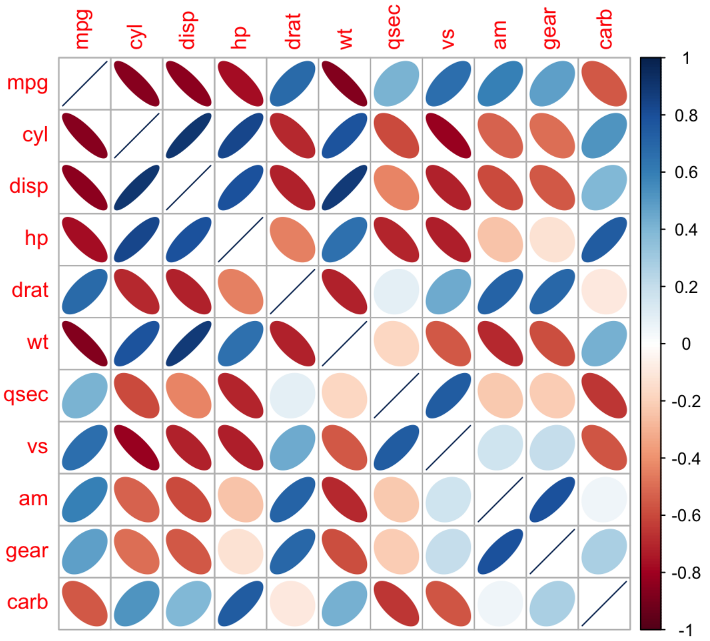
```
corrplot(M, method="pie")
```



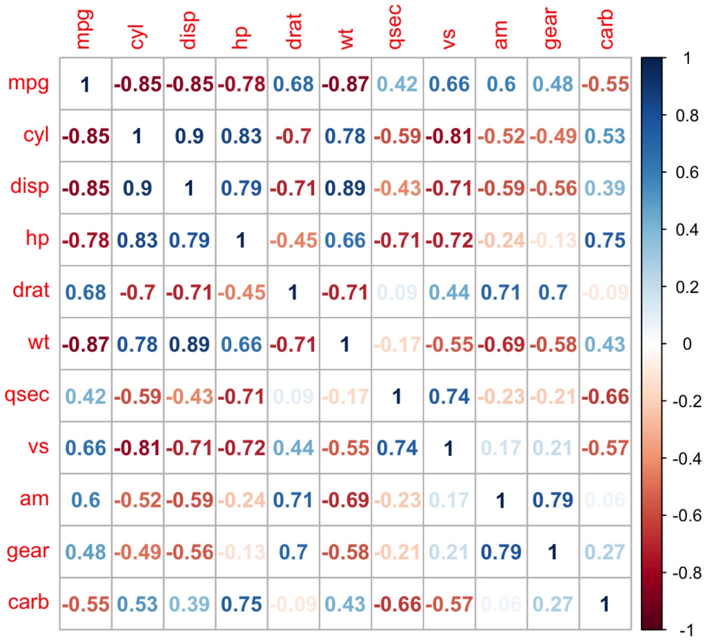
```
corrplot(M, method="color")
```



```
corrplot(M, method="ellipse")
```



```
#Display the correlation coefficient:
corrplot(M, method="number")
```

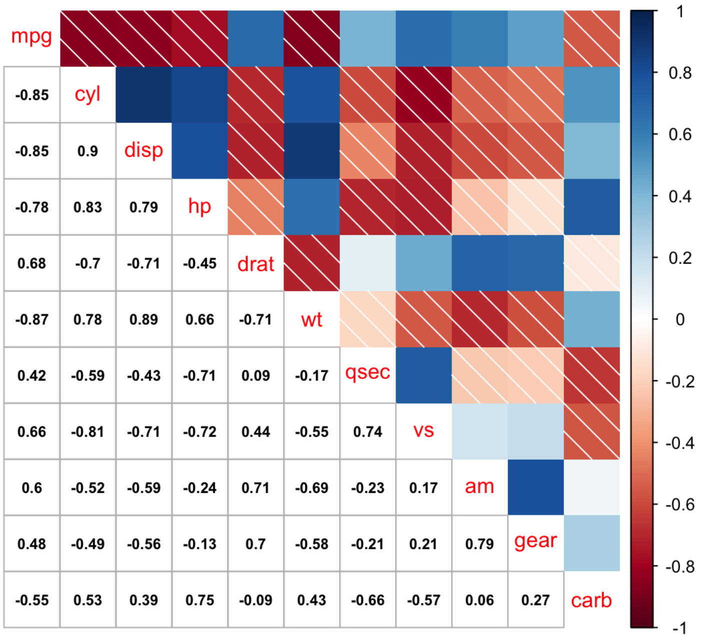


	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
mpg	1	-0.85	-0.85	-0.78	0.68	-0.87	0.42	0.66	0.6	0.48	-0.55
cyl	-0.85	1	0.9	0.83	-0.7	0.78	-0.59	-0.81	-0.52	-0.49	0.53
disp	-0.85	0.9	1	0.79	-0.71	0.89	-0.43	-0.71	-0.59	-0.56	0.39
hp	-0.78	0.83	0.79	1	-0.45	0.66	-0.71	-0.72	-0.24	-0.13	0.75
drat	0.68	-0.7	-0.71	-0.45	1	-0.71	0.09	0.44	0.71	0.7	-0.09
wt	-0.87	0.78	0.89	0.66	-0.71	1	-0.17	-0.55	-0.69	-0.58	0.43
qsec	0.42	-0.59	-0.43	-0.71	0.09	-0.17	1	0.74	-0.23	-0.21	-0.66
vs	0.66	-0.81	-0.71	-0.72	0.44	-0.55	0.74	1	0.17	0.21	-0.57
am	0.6	-0.52	-0.59	-0.24	0.71	-0.69	-0.23	0.17	1	0.79	0.06
gear	0.48	-0.49	-0.56	-0.13	0.7	-0.58	-0.21	0.21	0.79	1	0.27
carb	-0.55	0.53	0.39	0.75	-0.09	0.43	-0.66	-0.57	0.06	0.27	1

Here I have displayed four examples of visualizaiton methods and you are encouraged to try on your own the rest of them for the best presention.

> ## Types of correlogram layout
>
> There are essentially three types of layout:
>
> - “full” (default) : display full correlation matrix
> - “upper”: display upper triangular of the correlation matrix
> - “lower”: display lower triangular of the correlation matrix
>
> This is significant because the displayed matrix is essentially symmetric with respect to its top left to bottom right diagonal. By laying out the visulization differently could help to achiev the most effective way to understand the correlation matrices.
>
> In the following content, I will also demonstrate how the functions and options can be mixed to achieve effective customization of displays. `corrplot.mixed()` is a wrapped function for mixed visualization style.

```
# Full correlation matrix with a mixed visualization style using "corrplot.mixed()"
corrplot.mixed(M, lower = "number", upper = "shade", lower.col = "black", number.cex = .7)
```

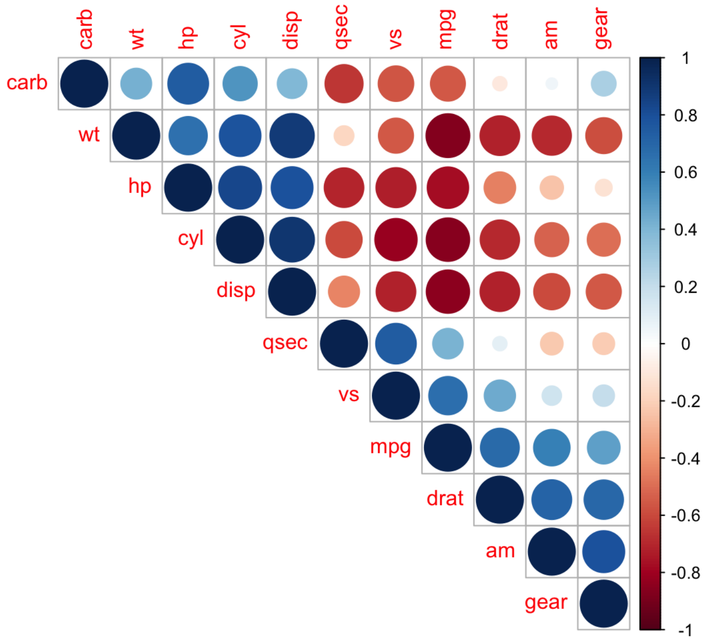


mpg										
-0.85	cyl									
-0.85	0.9	disp								
-0.78	0.83	0.79	hp							
0.68	-0.7	-0.71	-0.45	drat						
-0.87	0.78	0.89	0.66	-0.71	wt					
0.42	-0.59	-0.43	-0.71	0.09	-0.17	qsec				
0.66	-0.81	-0.71	-0.72	0.44	-0.55	0.74	vs			
0.6	-0.52	-0.59	-0.24	0.71	-0.69	-0.23	0.17	am		
0.48	-0.49	-0.56	-0.13	0.7	-0.58	-0.21	0.21	0.79	gear	
-0.55	0.53	0.39	0.75	-0.09	0.43	-0.66	-0.57	0.06	0.27	carb

Reordering the correlation matrix

> The correlation matrix can be reordered according to the correlation coefficient. This is important to identify the hidden structure and pattern in the matrix. There are four methods in corrplot (parameter order), named `"AOE"`, `"FPC"`, `"hclust"`, `"alphabet"`. More algorithms can be found in seriation package.
>
> `“hclust”` for hierarchical clustering order is used in the following examples.

```r
# Upper triangular correlogram with hclust reordering
corrplot(M, type="upper", order="hclust")
```

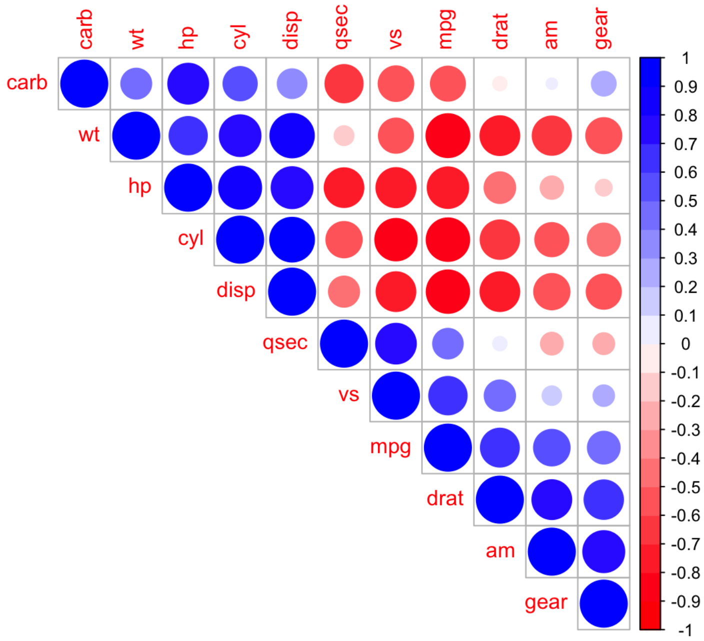


Changing the color and the rotation of text labels

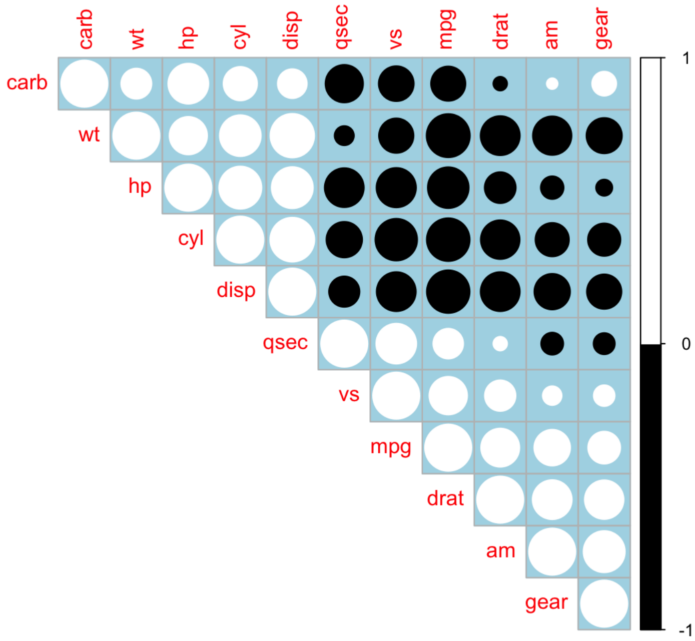
As shown in the section below, the color of the correlogram can be customized.

tl.col (for text label color) and tl.srt (for text label string rotation) are used to change text colors and rotations.

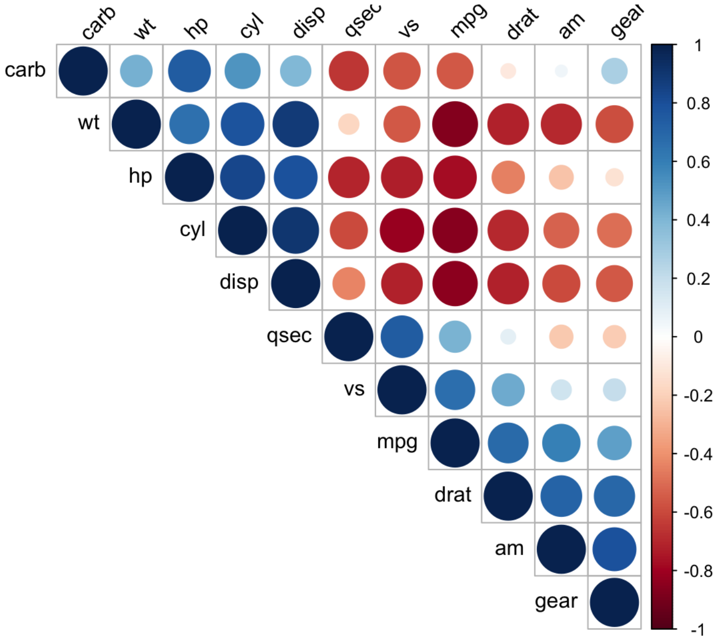
```r
# Using different color spectrum
col<- colorRampPalette(c("red", "white", "blue"))(20)
corrplot(M, type="upper", order="hclust", col=col)
```



```r
# Change background color to lightblue
corrplot(M, type="upper", order="hclust", col=c("black", "white"),
         bg="lightblue")
```



```
# Rotate the text lebel for 45 degrees,
corrplot(M, type="upper", order="hclust", tl.col="black", tl.srt=45)
```



Combining correlogram with the significance test

Computing the p-value of correlations

To determine whether the correlation between variables is significant, compare the p-value to your significance level. Usually, a significance level (denoted as α or alpha) of 0.05 works well.

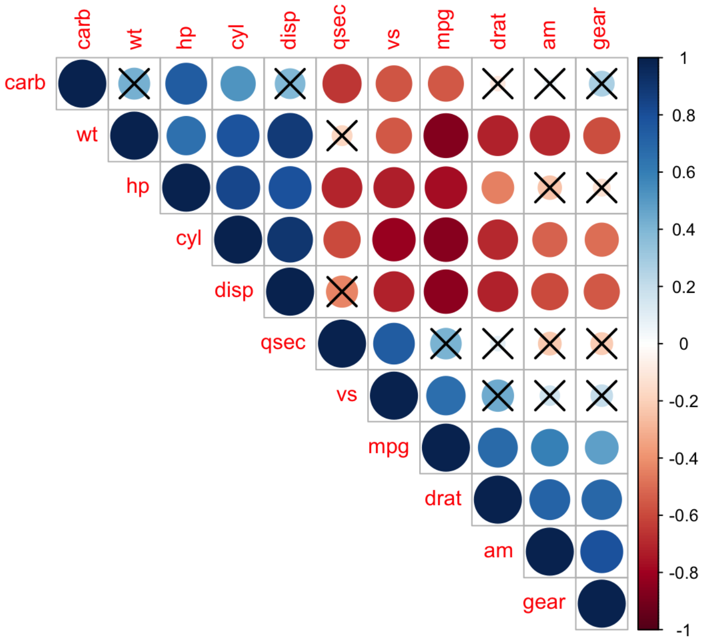
To compute the matrix of p-value, a custom R function is used:

```
# mat : is a matrix of data
# ... : further arguments to pass to the native R cor.test function
cor.mtest <- function(mat, ...) {
    mat <- as.matrix(mat)
    n <- ncol(mat)
    p.mat<- matrix(NA, n, n)
    diag(p.mat) <- 0
    for (i in 1:(n - 1)) {
        for (j in (i + 1):n) {
            tmp <- cor.test(mat[, i], mat[, j], ...)
            p.mat[i, j] <- p.mat[j, i] <- tmp$p.value
        }
    }
  colnames(p.mat) <- rownames(p.mat) <- colnames(mat)
  p.mat
}
# matrix of the p-value of the correlation
p.mat <- cor.mtest(mtcars)
head(p.mat[, 1:5])
```

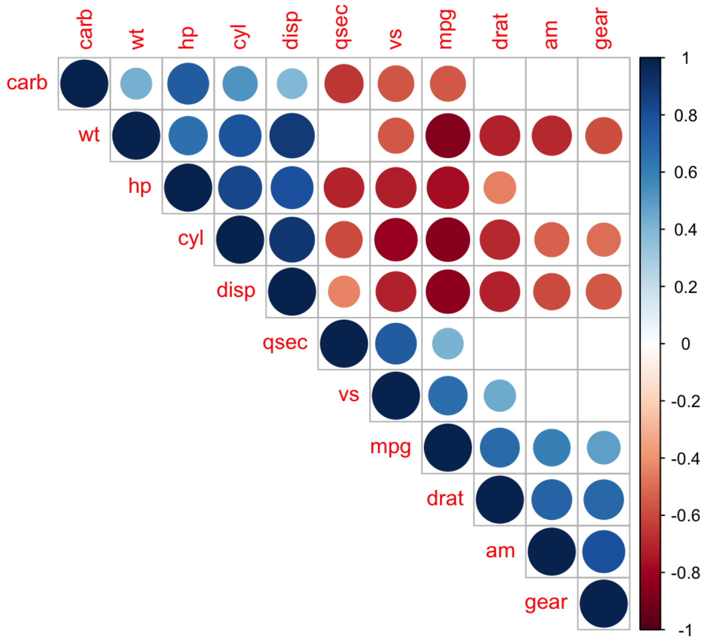
```
##               mpg          cyl         disp           hp         drat
## mpg  0.000000e+00 6.112687e-10 9.380327e-10 1.787835e-07 1.776240e-05
## cyl  6.112687e-10 0.000000e+00 1.802838e-12 3.477861e-09 8.244636e-06
## disp 9.380327e-10 1.802838e-12 0.000000e+00 7.142679e-08 5.282022e-06
## hp   1.787835e-07 3.477861e-09 7.142679e-08 0.000000e+00 9.988772e-03
## drat 1.776240e-05 8.244636e-06 5.282022e-06 9.988772e-03 0.000000e+00
## wt   1.293959e-10 1.217567e-07 1.222320e-11 4.145827e-05 4.784260e-06
```

Add significance level to the correlogram

```
# Specialized the insignificant value according to the significant level
# In the below figure, correlations with p-value > 0.01 are considered as insignificant. In this case crosses are added to the correlation coefficient values crosses.
corrplot(M, type="upper", order="hclust",
         p.mat = p.mat, sig.level = 0.01)
```



```
# Leave blank on no significant coefficient
#In the below figure, correlations with p-value > 0.03 are considered as insignificant. In this case the correlation coefficient values are leaved blank.
corrplot(M, type="upper", order="hclust",
         p.mat = p.mat, sig.level = 0.03, insig = "blank")
```



Another R package: “corrgram”

This is another option to display correlation matrices. Michael Friendly from Yale Univeristy in his article describes **corrgrams** as **“exploratory displays for correlation matrices”**. Moreover, many of its functions and options correspond to those in the “corrplot” package.

In R, correlograms are implimented through the **corrgram(x, order = , panel=, lower.panel=, upper.panel=, text.panel=, diag.panel=)** function in the corrgram package.

- **x** is a data frame with one observation per row.
- **order=TRUE** will cause the variables to be ordered using principal component analysis of the correlation matrix.

panel= refers to the off-diagonal panels.

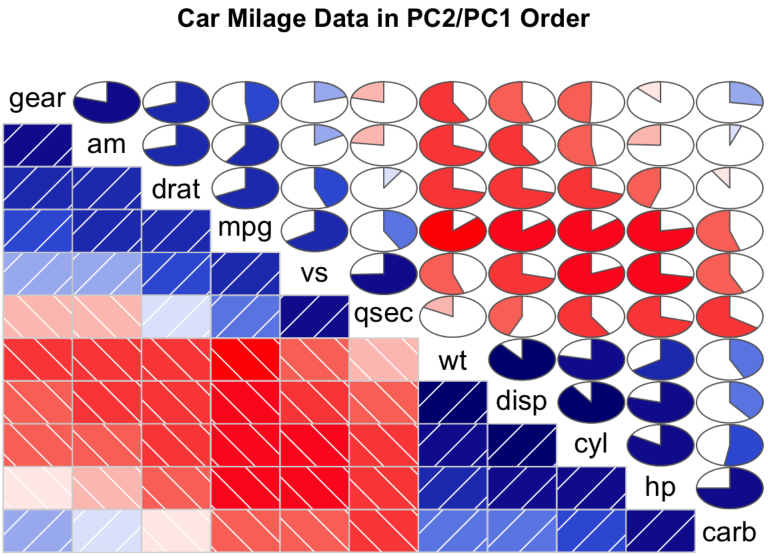
- You can use **lower.panel=** and **upper.panel=** to choose different options below and above the main diagonal respectively.
- **text.panel=** and **diag.panel=** refer to the main diagnonal. Allowable parameters are given below.

off diagonal panels - **panel.pie** (the filled portion of the pie indicates the magnitude of the correlation) - **panel.shade** (the depth of the shading indicates the magnitude of the correlation) - **panel.ellipse** (confidence ellipse and smoothed line) - **panel.pts** (scatterplot)

- **main diagonal panels**
- **panel.minmax** (min and max values of the variable)
- **panel.txt** (variable name).

```
# Install R Package "corregram"
install.packages("corrgram")
```

```
# A Correlogram Example
library(corrgram)
corrgram(mtcars, order=TRUE, lower.panel=panel.shade,
  upper.panel=panel.pie, text.panel=panel.txt,
  main="Car Milage Data in PC2/PC1 Order")
```



> # Recap and Take Home Message
>
> The post’s main objective is to offer a comprehensive guide of visualizing correlation matrix using correlogram with the help of R **corrplot** package.
>
> The covered topics ranges from installment, computing correlation matrix of a given data frame to customizing visualization methods via layout, color, order and significance level. The organization of the post serves to provide logical structure with identifiable sections and subsections of the “corrplot” package, for a variety of audience. In addition, I included the possibility of using another R package “corrgram” for an open leanring expereience.

References and Acknowledgements

- Quick-R https://www.statmethods.net/advgraphs/correlograms.html
- Statistical tools for high-throughput data analysis http://www.sthda.com/english/wiki/visualize-correlation-matrix-using-correlogram
- Friendly, Michael. 2002. Corrgrams: Exploratory Displays for Correlation Matrices. The American Statistician, 56, 316–324. http://datavis.ca/papers/corrgram.pdf
- Source of mtcars data: Henderson and Velleman (1981), Building multiple regression models interactively. Biometrics, 37, 391–411.
- D. J. Murdoch and E. D. Chow. 1996. A Graphical Display of Large Correlation Matrices. The American Statistician, 50, 178-180.
- R Help Page: An Introduction To Correplot Package
- R Help Page: Draw a correlogram
- R Help Page: Visualization of a correlation matrix
- Ripley, B. D. (1981) Spatial Statistics. Wiley.
- Venables, W. N. and Ripley, B. D. (2002) Modern Applied Statistics with S. Fourth edition. Springer.
- Minitab Express Support http://support.minitab.com/en-us/minitab-express/1/help-and-how-to/modeling-statistics/regression/how-to/correlation/interpret-the-results/