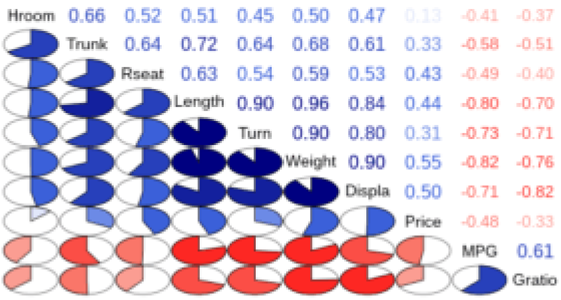
Correlation and Correlogram Exercises



Correlation analysis is one of the most popular techniques for data exploration. This set of exercises is intended to help you to extend, speed up, and validate your correlation analysis. It allows to practice in:
– calculating linear and nonlinear correlation coefficients,
– testing those coefficients for statistical significance,
– creating correlation matrices to study interdependence between variables in dataframes,
– drawing graphical representations of those matrices (correlograms),
– calculating coefficients for partial correlation between two variables (controlling for their correlation with other variables).
The exercises make use of functions from the packages Hmisc, corrgram, and ggm. Please install these packages, but do not load them before starting the exercises in which they are needed (to avoid a namespace conflict) (the ggm package contains the function called rcorr which masks the rcorr function from the Hmisc package, and vice versa. If you want to return to the rcorr function from the Hmisc after loading the ggm package run detach(package:ggm)).
Exercises are based on a reduced version of the auto dataset from the corrgram package (download here). The dataset contains characteristics of 1979 automobile models.
Answers to the exercises are available here

Exercise 1
Calculate simple (linear) correlation between car price and its fuel economy (measured in miles per gallon, or mpg).

Exercise 2
Use the cor.test function to check whether the obtained coefficient is statistically significant at 5% level.

Exercise 3
Simple correlation assumes a linear relationship between variables, but it may be useful to relax this assumption. Calculate Spearman's correlation coefficient for the same variables, and find its statistical significance.

Exercise 4
In R, it is possible to calculate correlation for all pairs of numeric variables in a dataframe at once. However, this requires excluding non-numeric variables first.
Create a new dataframe, auto_num, that contains only columns with numeric values from the auto dataframe. You can do this using the Filter function.

Exercise 5
Use the cor function to create a matrix of correlation coefficients for variables in the auto_num dataframe.

Exercise 6
The standard cor.test function does not work with dataframes. However, statistical significance of correlation coefficients for a dataframe can be verified using the rcorr function from the Hmisc package.
Transform the auto_num dataframe into a matrix (auto_mat), and use it to check significance of the correlation coefficients with the rcorr function.

Exercise 7
Use the corrgram function from the corrgram package to create a default correlogram to visualize correlations between variables in the auto dataframe.

Exercise 8
Create another correlogram that (1) includes only the lower panel, (2) uses pie diagrams to represent correlation coefficients, and (3) orders the variables using the default order.

Exercise 9
Create a new dataframe, auto_subset, by subsetting the auto dataframe to include only the Price, MPG, Hroom, and Rseat variables. Use the new dataframe to create a correlogram that (1) shows correlation coefficients on the lower panel, and (2) shows scatter plots (points) on the upper panel.

Exercise 10
Use the the correlations function from the ggm package to create a correlation matrix with both full and partial correlation coefficients for the auto_subset dataframe. Find the partial correlation between car price and its fuel economy.