Continuous x Continous data (1 of 2)

Aug 6, 2023.

Exploring bivariate Continuous x Continuous data

THIS CHAPTER explores how to summarize and visualize the interaction between *bivariate continuous data* using correlation analysis, scatter plots, scatter plot matrices and other such techniques.

Data: Let us work with the same `mtcars` data from the previous chapter. Suppose we run the following code to prepare the data for subsequent analysis. The data is now in a tibble called `tb`:

```
# Load the required libraries, suppressing annoying startup messages
library(tibble)
suppressPackageStartupMessages(library(dplyr))
# Read the mtcars dataset into a tibble called tb
data(mtcars)
tb <- as_tibble(mtcars)
# Convert several numeric columns into factor variables
tb$cyl <- as.factor(tb$cyl)
tb$vs <- as.factor(tb$vs)
tb$am <- as.factor(tb$am)
tb$gear <- as.factor(tb$gear)
# Directly access the data columns of tb, without tb$mpg
attach(tb)
```

Scatterplots

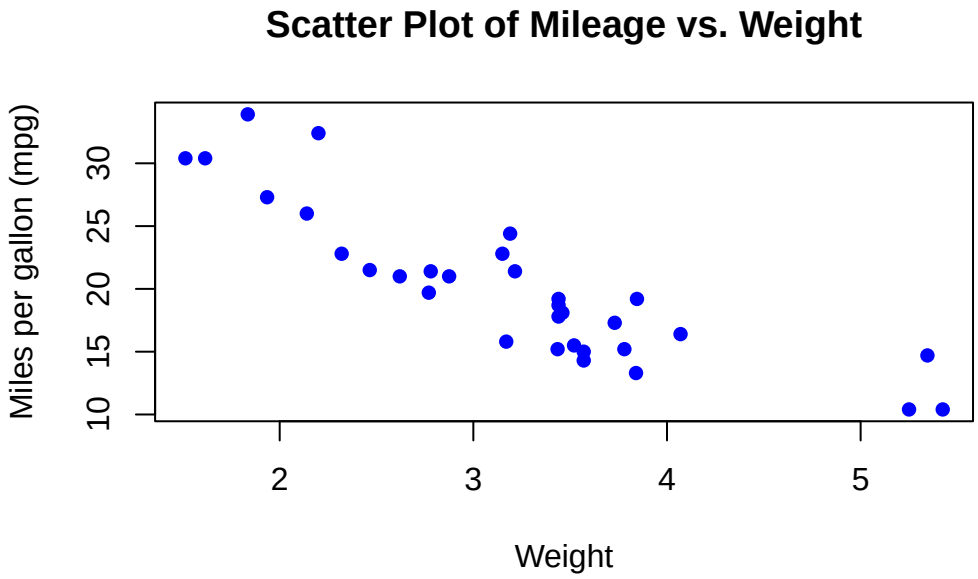
1. A scatter plot is a type of graph used to display the relationship between two continuous variables. It is a graphical representation of a bivariate distribution, where the values of two variables are plotted as points on a two-dimensional coordinate system.

2. A scatter plot can be used to identify trends, clusters, outliers, and other patterns in the data. It is also useful for detecting the presence of any outliers or influential observations that may affect the analysis.

3. To create a scatter plot of mpg (miles per gallon) against wt (weight) in the mtcars data set, we can use the following code:

Scatterplot using plot()

```
plot(tb$wt,
     tb$mpg,
     main = "Scatter Plot of Mileage vs. Weight",
     xlab = "Weight", ylab = "Miles per gallon (mpg)",
     pch = 16,
     col="blue")
```



4. Discussion:

- This code will first load the mtcars data set, then create a scatter plot of `mpg` against `wt` using the `plot()` function.
- The `main` argument adds a title to the plot, the `xlab` and `ylab` arguments add axis labels
- the `pch` argument sets the shape of the points to a solid circle. pch = 15 gives a filled square. Here are popular values: pch = 16 gives a filled circle, pch = 17 gives a filled

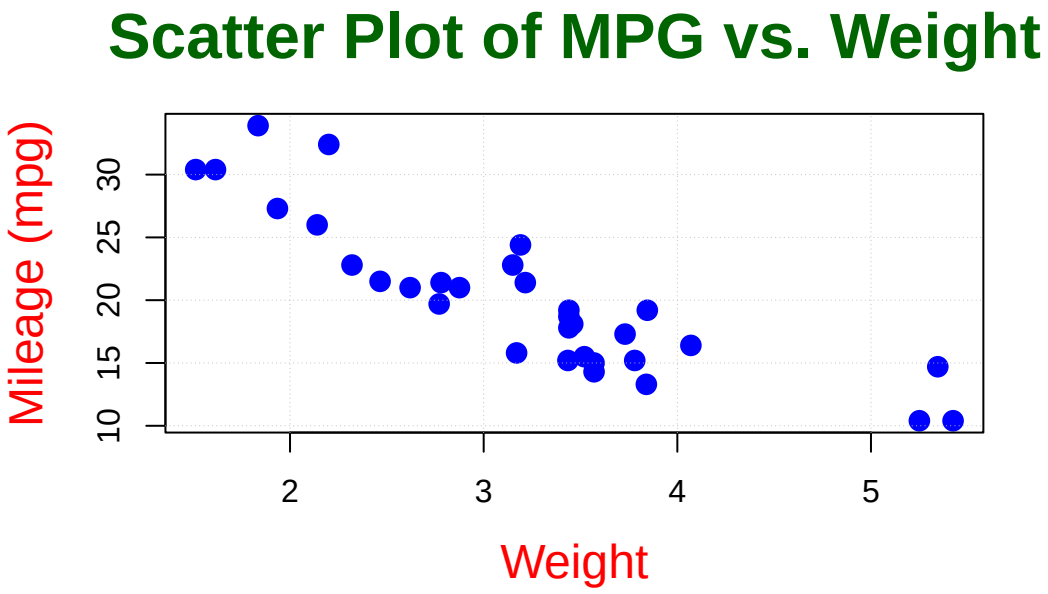
triangle (pointing upwards), pch = 18 gives a filled diamond, pch = 19 gives a solid circle, pch = 20 gives a filled bullet (smaller than pch = 19)

- the `col` argument specifies the color of the data points. We can use any named color in R, or we can use hexadecimal color codes. For instance, `col = "#FF0000"` would give us red points.

5. We can personalize the appearance of the scatterplot in a variety of additional ways.

```
# Create the scatterplot
plot(tb$wt,
     tb$mpg,
     main = "Scatter Plot of MPG vs. Weight",
     xlab = "Weight", ylab = "Mileage (mpg)",
     pch = 16, cex = 1.5, col="blue",
     col.lab="red", cex.lab=1.5,
     col.main="darkgreen", cex.main=2,
     bg = "gray")

# Add a grid
grid(col = "gray", lty = "dotted", lwd = 0.5)
```



6. Discussion

- Point Size: In the second plot, the size of the points is 1.5 times the default size (`cex = 1.5`), while in the first plot, the size of the points is the default size as `cex` is not

specified.

- Axis Labels' Color and Size: The second plot has red-colored, larger size axis labels (`col.lab="red", cex.lab=1.5`), while the first plot uses the default color and size as these parameters are not specified.
- Title's Color and Size: The second plot has a dark green title that is twice the default size (`col.main="darkgreen", cex.main=2`), while the first plot uses the default color and size for the title as these parameters are not specified.
- Background Color: The second plot has a light gray background (`bg = "lightgray"`), while the first plot uses the default background color as the bg parameter is not specified.
- Grid: The second plot includes a grid with gray dotted lines (`grid(col = "gray", lty = "dotted", lwd = 0.5)`), while the first plot does not have a grid as the grid() function is not called.

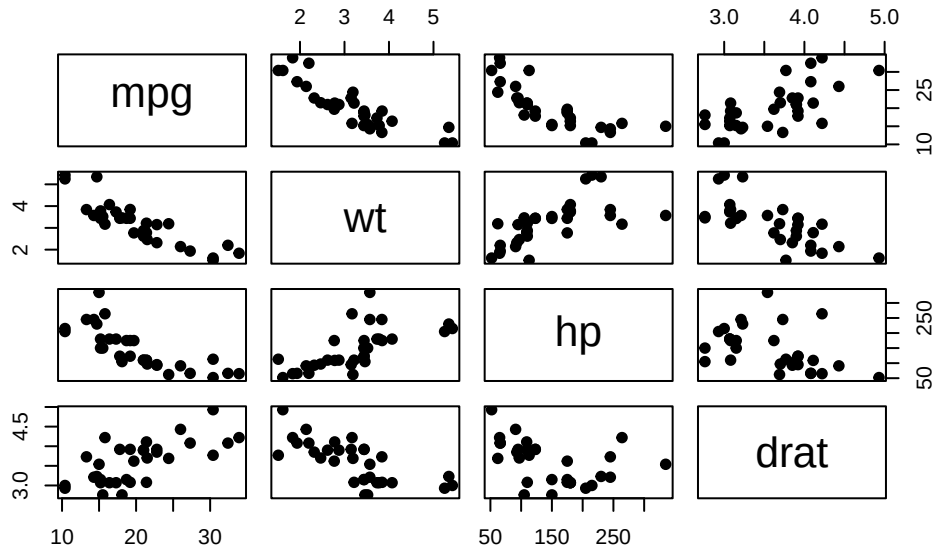
Scatterplot Matrix

A scatter plot matrix (also called a pairs plot or a SPLOM) is a graphical display of pairwise scatter plots of a set of variables. In a scatter plot matrix, each variable in the dataset is plotted against every other variable in a matrix format. This allows us to visualize the relationships between pairs of variables and explore potential patterns or trends in the data.

A scatter plot matrix is particularly useful for exploring multivariate datasets, as it allows us to quickly identify which pairs of variables may be strongly correlated, which may have weak or no correlation, and which may exhibit nonlinear relationships. It can also be used to identify outliers or unusual observations, and to visualize clusters or groups of observations based on patterns in the scatter plots.

Scatterplot Matrix Using pairs()

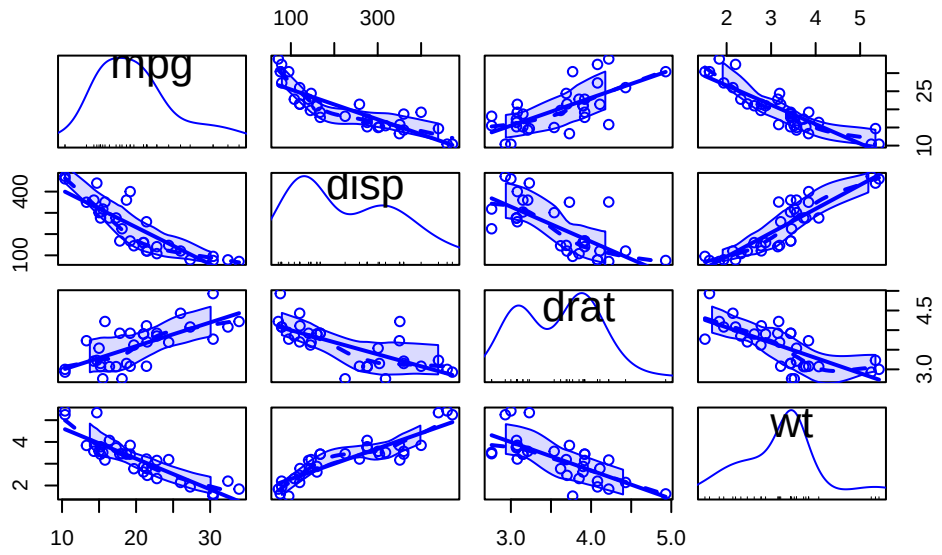
```
# scatter plot matrix for mpg, wt, hp, drat
pairs(tb[,c("mpg","wt","hp","drat")], pch = 19)
```



Scatterplot Matrix Using scatterplotMatrix()

```
# Load the car package
library(car)

# Create a scatterplot matrix using scatterplotMatrix()
scatterplotMatrix(~ mpg + disp +drat +wt,
                  data = tb, col = c("blue", "red"))
```



Scatterplot Matrix Using pairs.panels()

```
# Load the psych package
library(psych)

# Create a scatterplot matrix using pairs.panels()
pairs.panels(tb[,c("mpg","wt","hp","drat")],
             main = "Scatterplot Matrix")
```