Visualizing association among two or more quantitative variables -

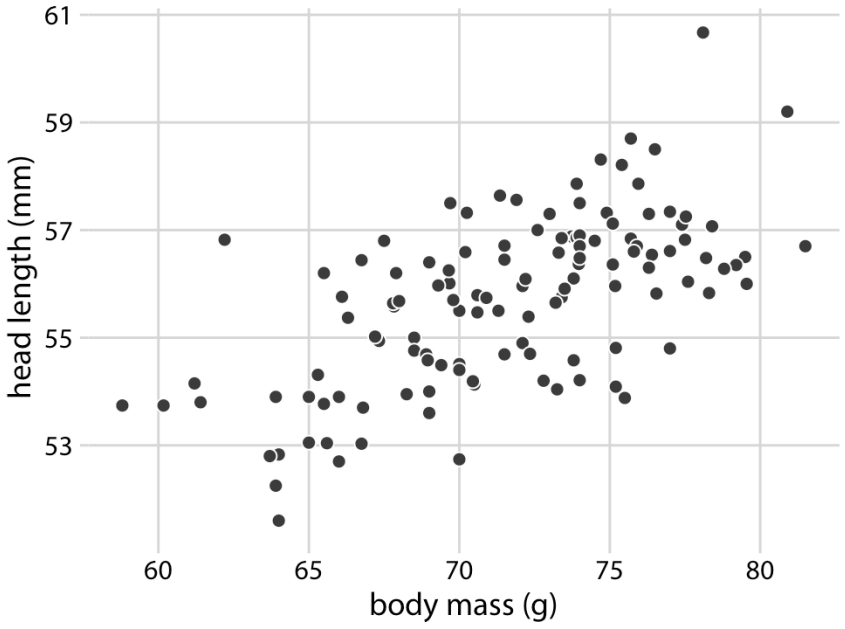
When we want to show more than two variables at once, we may use a bubble chart, a scatter plot or a correlogram. For example, measurements of different animals, such as the height, weight, length, and daily energy demands.
In this paper we also perform dimensionality reduction, in the form of principal component analysis.

Scatter plot-

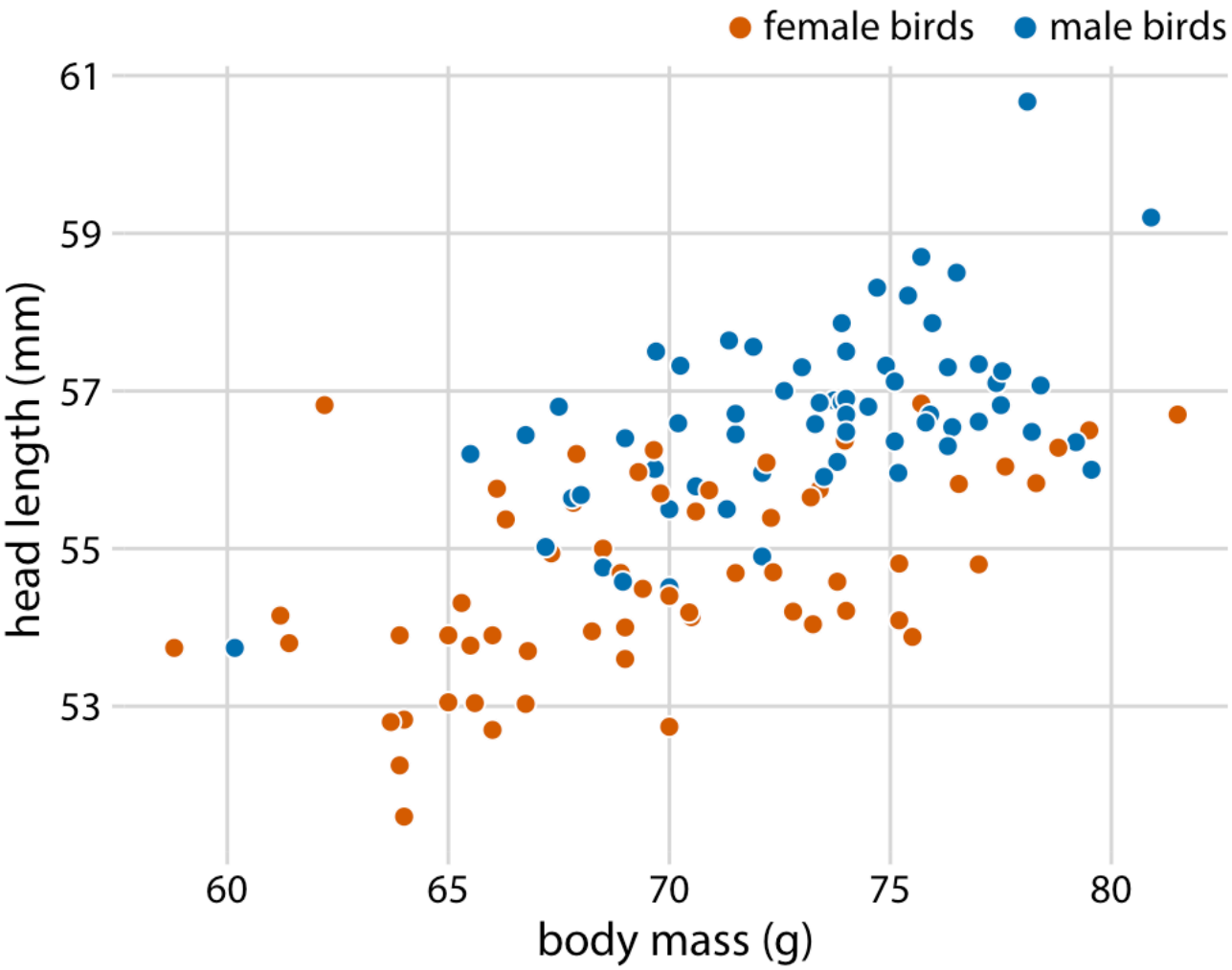
Let's start with example, I'm taking a dataset of measurements performed on 123 blue jay birds.
The dataset contains information such as head length, the skull size, and the body mass of each bird. We expect that there are relationship between these variables. For example, birds with longer bills would be expected have larger skull size, and birds with higher body mass should have larger bits and skulls than birds with lower body mass.

For plotting this above relations -
Relation - head length against body mass.
X-axis: head length
Y-axis: body mass
Dot: each bird is shown by one dot

There is a trend for birds, the birds with longest head falls close to the maximum body mass observed, and the bird with shortest head falls close to the minimum mass observed.

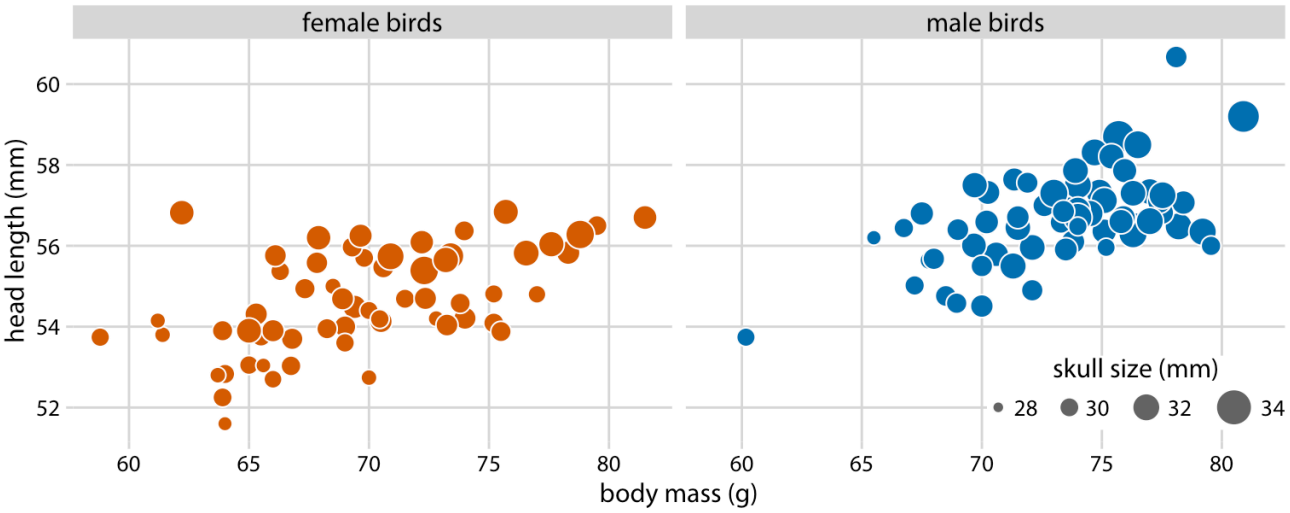


The dataset contains both male and female birds we can show the same relation separately for each sex by using colour of dots. Now, we can see the new trend over a trend that females tends to have shorter heads than males. At the same time, females tend to be lighter than males on average.



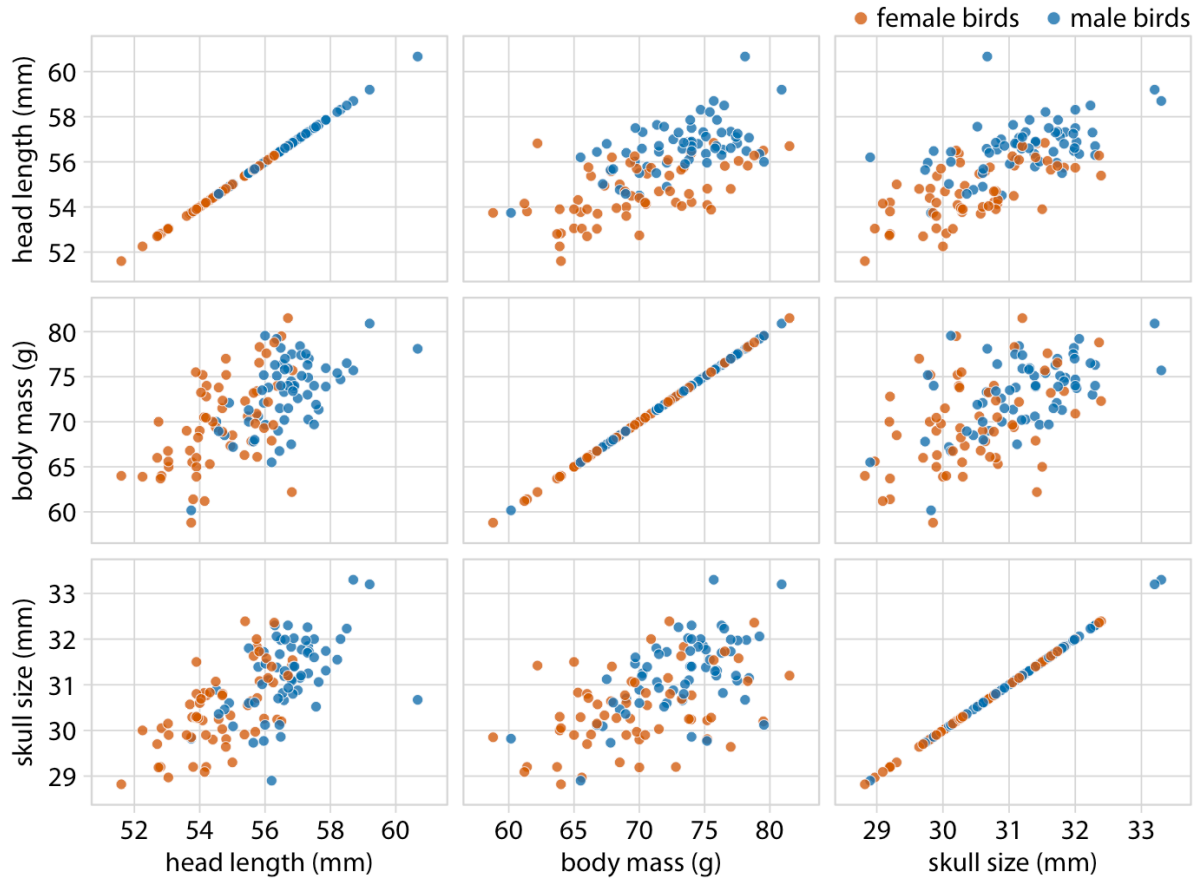
Bubble chart -

Bubble charts have disadvantage that they show the same types of variables, quantitative variable with two different types of scales, position and size. This makes it difficult to visually ascertain the strength of association between variables.



Because the head length is defined as the distance from the tip of the bill to the back of the head, a larger head length could imply a longer bill, as larger skull, or both. For showing more then two variables(body mass, head length, and skull size) we can use size of dots(referred as bubbles) for skull size variable that shows size of skull size based on size of bubble(dot) as we already using x-axis for body mass and y-axis for head length.

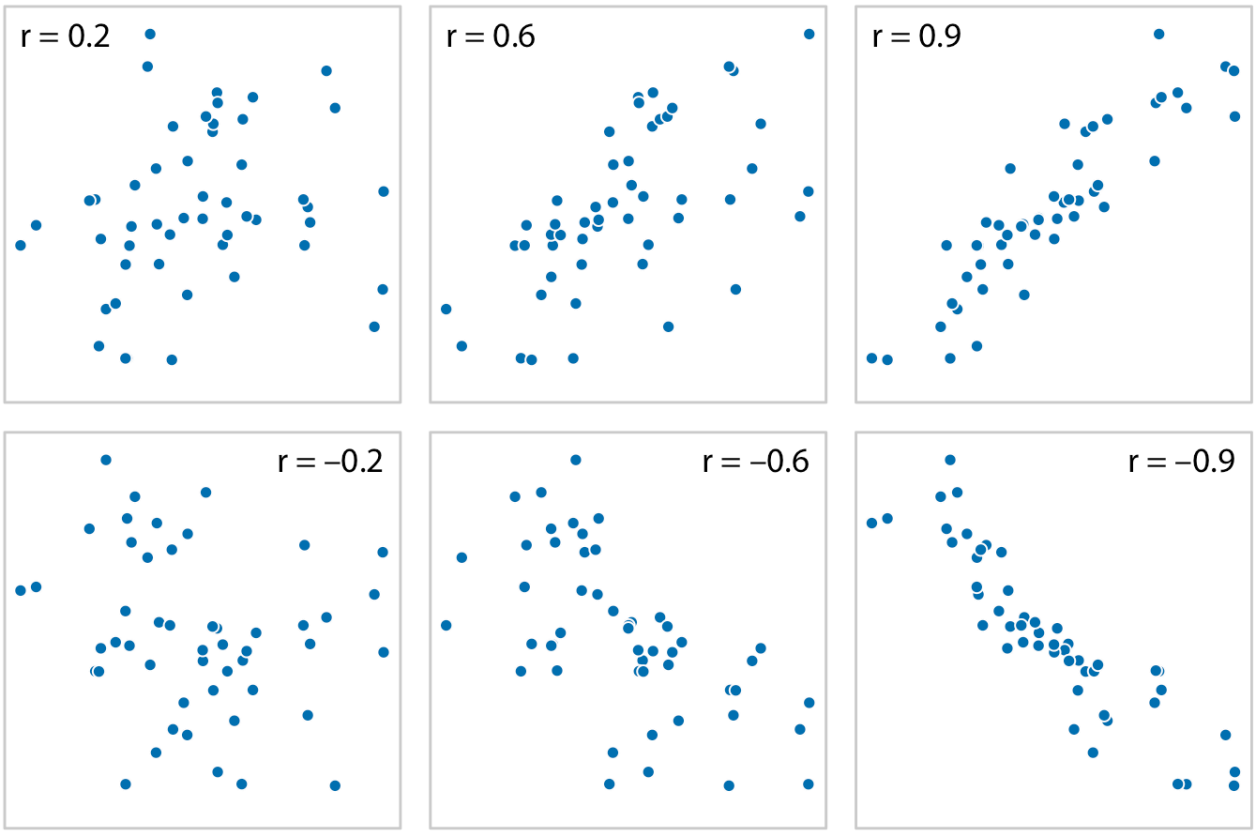
As an alternative to bubble chart it may be preferable to show an **all-against-all matrix of scatter plot** where each individual plot shows two data dimension. This figure shows clearly that the relationship between skull size



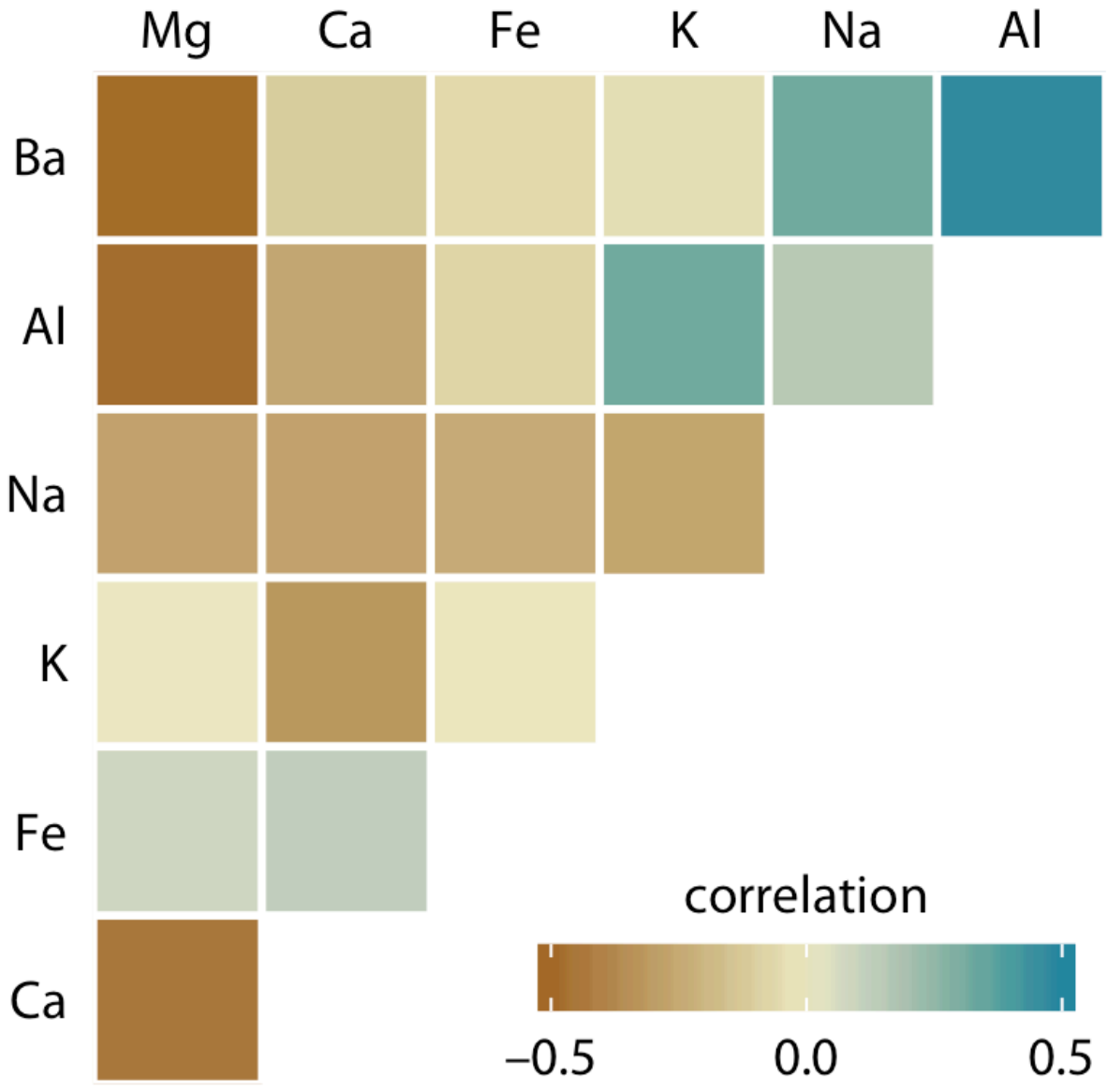
and body mass is comparable for female and male except that the female birds tend to be smaller.

Correlogram -

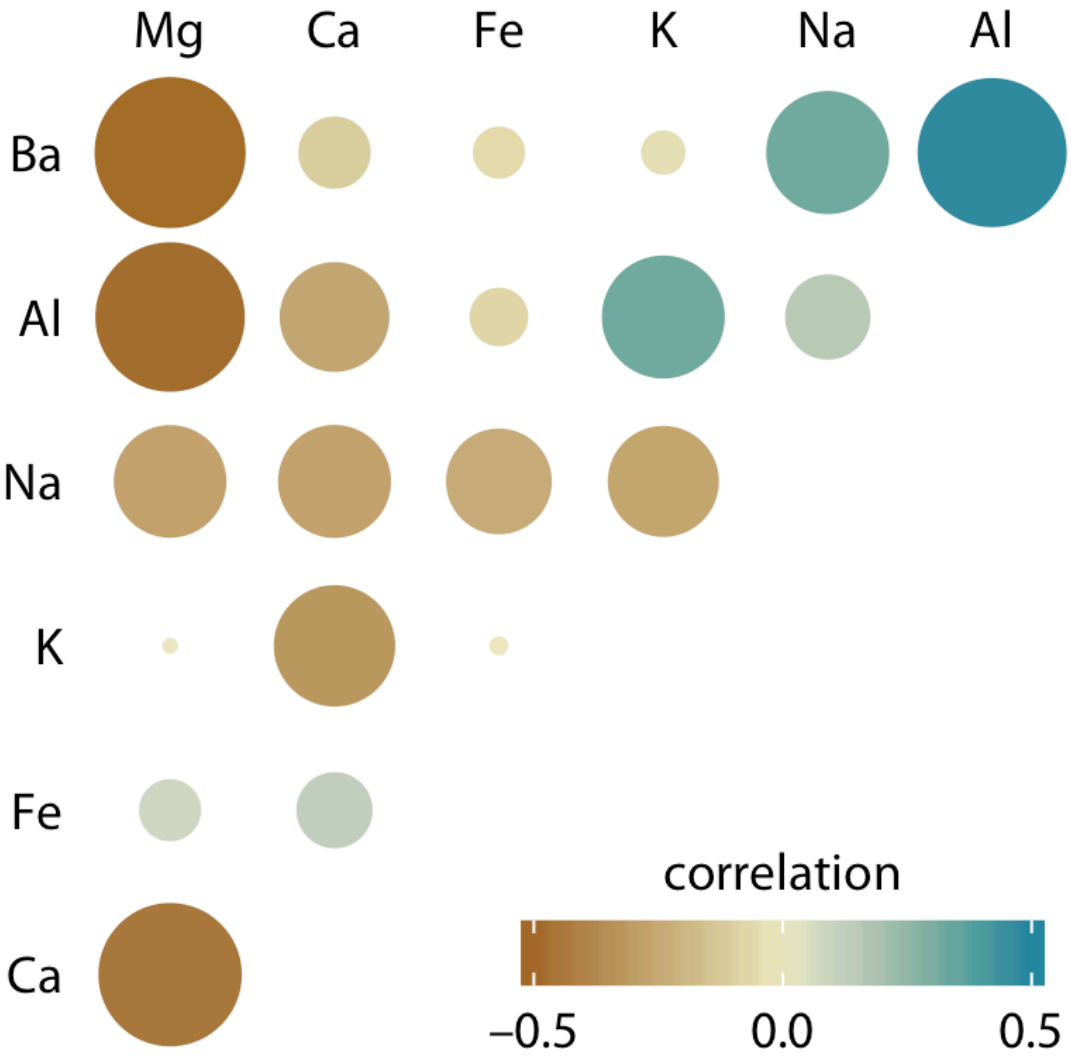
Visualisation of correlation coefficient is called correlograms. When we have more than three to four quantiative variables we correlograms. It is more useful to quantify the amount of association between pairs of variables and visualize the quantity rather then data. Calculate the correlation coefficient r(-1 < r < 1) that measure relation between two variables.
Value, r = 0 means there is no relation.
r = -1, 1 indicate a perfect relation
the sign of r shows whether the variable are correlated or anti-correlated .
Correlated - larger value in one variable coincide with larger values in other variable.
Anti-correlated - larger values in one variable coincide with smaller values in other variable.



We will consider a dataset of over 200 glass fragments. For each glass fragments, we have measurements about its composition, expressed as the percentage in weight of various mineral oxide. There are seven different oxides for which we have measurements, yielding a total of 6 + 5 + 4 + 3 + 2 + 1 = 21 pairwise correlations. We can display these 21 correlation at once as matrix of coloured tiles, where each tile represents one correlation coefficient. By correlograms we see that magnesium is negatively correlated with nearly all. Other oxides, and that aluminium and barium have string positive correlation.



One weakness of the correlation is that low correlation. I.e. correlation with absolute value near zero, are not as visually suppressed as they should be. For example magnesium(Mg) ans potassium(k) have correlation coefficient nearly zero. To overcome this limitation, we can display correlation as as coloured circle and scale the circle size based on their correlation coefficient value. Value near absolute zero have smaller circle.



Remember -

scatter plot - show relation between 2 variables.

Bubble chart - show relation between 3 variables.

Correlogram - show relation between more than 3 or 4 variables.

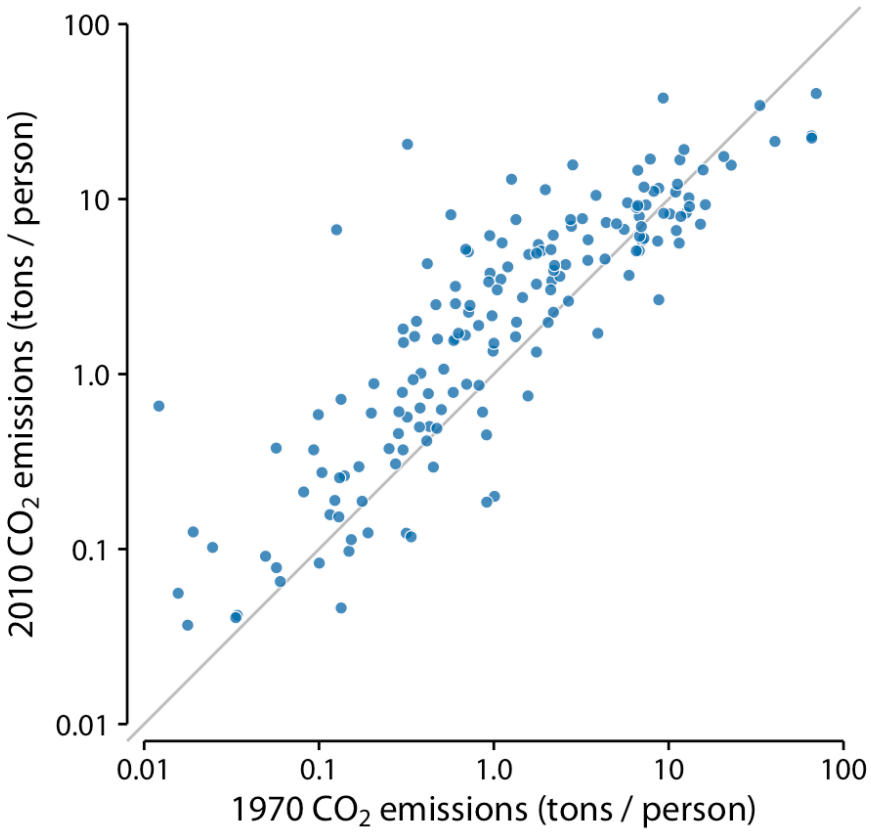
Paired data-

A special case of multivariate quantitative data is paired data: Data where there are two or more measurements of the same quantity under slightly different conditions.

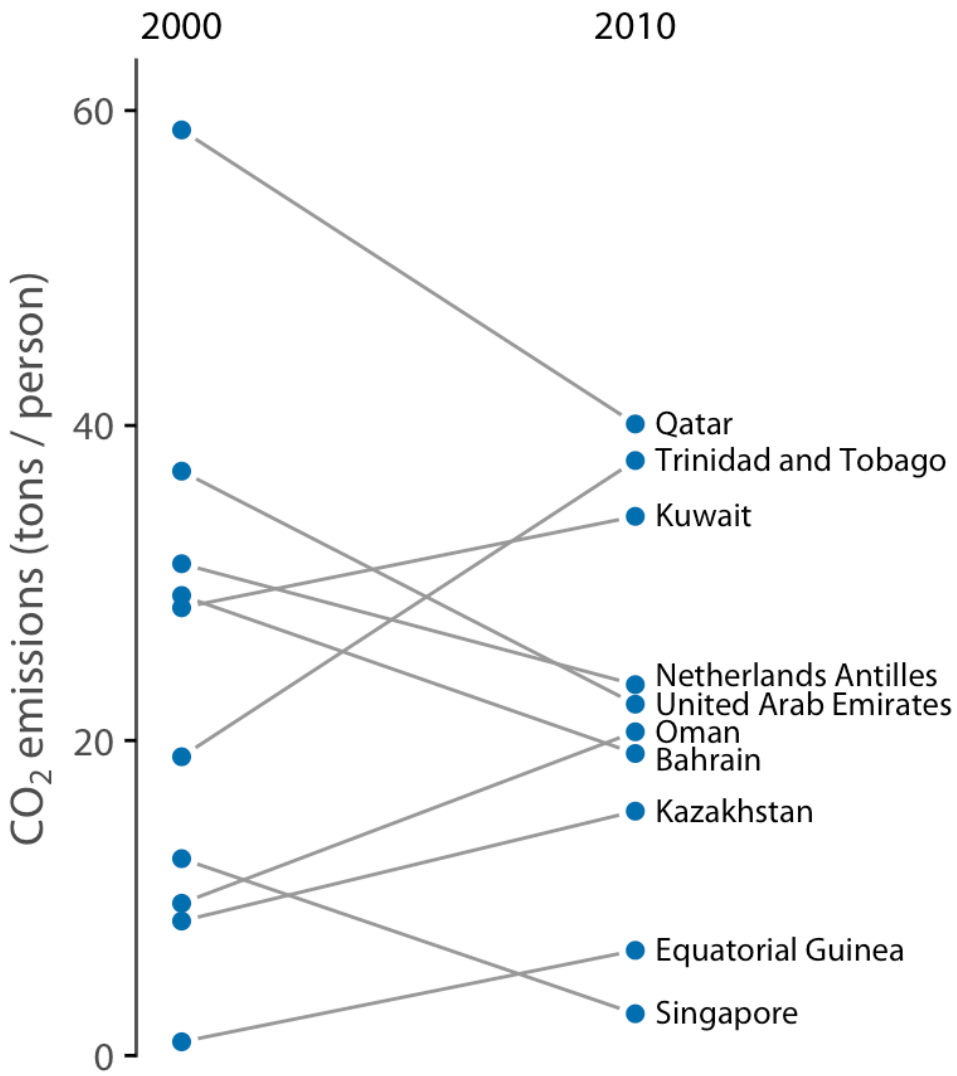
Examples- comparable measurements on each subject(the length of the right and the left arm of a person, repeat measurements on the subject at different time points(a person's weight at two different time during the year), or measurements on two closely related subjects(the eight of two identical twins).

For paired data we need to choose visualization that highlight any difference between the paired measurements.

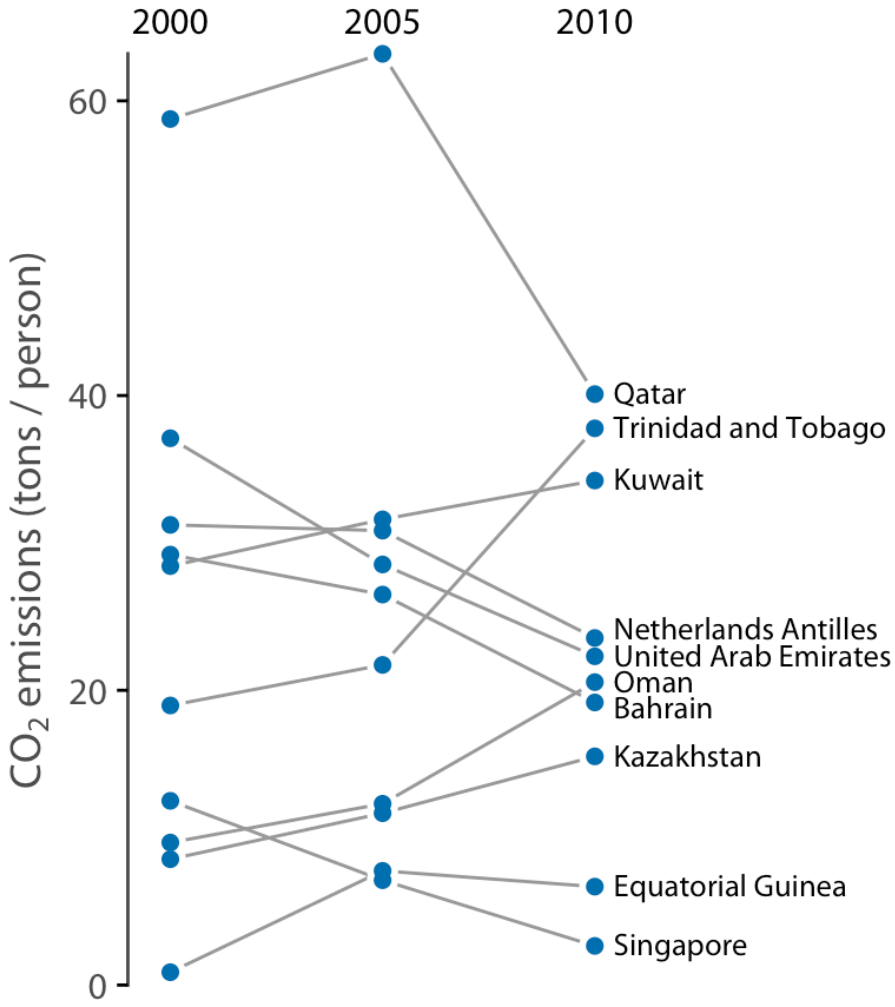
As an example, consider the carbon dioxide(CO2) emissions per person, measured for 166 countries Bothe in 1970 and 2010. This an example of repeat measurement of a subject at different time points. It highlights two common feature of paired data. First, most points are relatively close to the diagonal line. Second, the points are systematically shifted upwards relative to the diagonal line. The majority of countries has seen an increase in CO2 emission over the 40 years considered.



Scatter plot work well when we have large number of data points. If we have only small number of observation and primary interested in the identity of each individual case, a *slopegraph* may be a better choice. In a slopegraph, we draw individual measurements of dots arranged into two columns and indicate paring by connecting the paired data with a line the slope of each highlights the magnitude and direction of change.



Slopegraphs have one advantage over scatter plots that they can be used to compare more than two measurements at a time. For example, we can modify the above figure to show CO2 emission at three time points 2000, 2005, and 2010. Here we can see the large difference in the trend in 2005 to 2010 negative for the country Qatar and positive for Trinidad and Tobago.


2000
2005
2010
60
40
20
0
CO_2 emissions (tons / person)
Qatar
Trinidad and Tobago
Kuwait
Netherlands Antilles
United Arab Emirates
Oman
Bahrain
Kazakhstan
Equatorial Guinea
Singapore