Data preprocessing and sampling

Part 2: Correlation, covariance, heatmaps and scatter plots

By: Noureddin Sadawi, PhD

University of London

Correlation

- Quantifies the strength of a relationship between two numeric variables.
- Usually we are interested in the linear correlation between two numeric variables.
- ‘Variables X and Y (each with measured data) are said to be positively correlated if high values of X go with high values of Y, and low values of X go with low values of Y.
- If high values of X go with low values of Y, and vice versa, the variables are negatively correlated.’

(Bruce and Bruce *Practical statistics for data scientists*, second edition, 2020).

Key terms for correlation

‘**Correlation coefficient:**

A metric that measures the extent to which numeric variables are associated with one another (ranges from –1 to +1).

Correlation matrix:

A table where the variables are shown on both rows and columns, and the cell values are the correlations between the variables.

Scatterplot:

A plot in which the x-axis is the value of one variable, and the y-axis the value of another.’

(Bruce and Bruce *Practical statistics for data scientists*, second edition, 2020).

Variance and covariance

- The **variance** of a variable is the average of the squared deviations of values from the mean.

$$Var\,(x) = \frac{\sum(x-\overline{x})^2}{n} \qquad Var(y) = \frac{\sum(y-\overline{y})^2}{n}$$

Variance and covariance

- ‘**Covariance:** A measure of the extent to which one variable varies in concert with another (i.e., similar magnitude and direction).’ (Bruce and Bruce *Practical statistics for data scientists*, second edition, 2020).

$$Covar\ (x, y) = \frac{\sum(x-\overline{x})(y-\overline{y})}{n}$$

Correlation coefficient

- ‘The correlation coefficient always lies between +1 (perfect positive correlation) and –1 (perfect negative correlation); 0 indicates no correlation.
- Variables can have an association that is not linear, in which case the correlation coefficient may not be a useful metric.’

(Bruce and Bruce *Practical statistics for data scientists*, second edition, 2020).

$$r = \frac{Covar(x, y)}{\sqrt{Var(x)Var(y)}}$$

Correlation matrix

- This is a table that displays the correlation coefficients between numeric variables.
- Each entry shows the correlation coefficients between two variables.
- Usually used to summarise data.
- Symmetric around the diagonal.
- Diagonal values are 1.

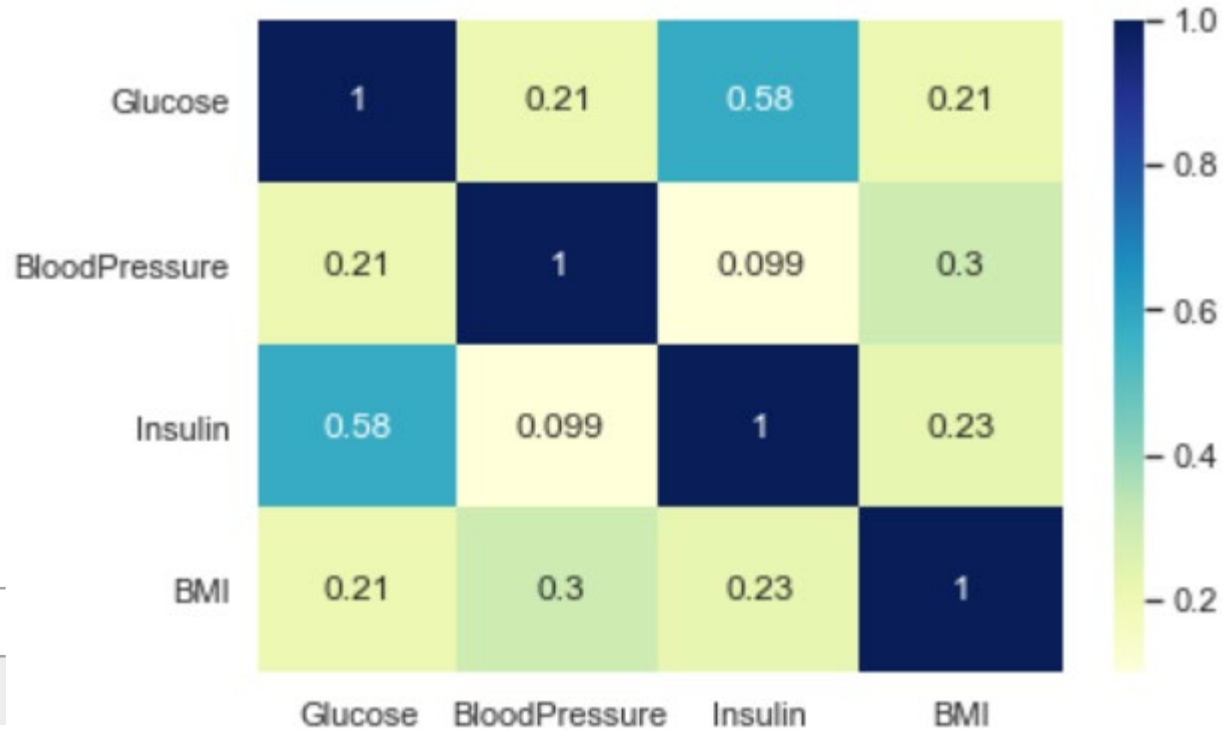
	T	CTL	FTR	VZ	LVLT
T	1.000	0.475	0.328	0.678	0.279
CTL	0.475	1.000	0.420	0.417	0.287
FTR	0.328	0.420	1.000	0.287	0.260
VZ	0.678	0.417	0.287	1.000	0.242
LVLT	0.279	0.287	0.260	0.242	1.000

Example from (Bruce and Bruce *Practical statistics for data scientists*, second edition, 2020) p.32.

Correlation heatmap

Glucose	BloodPressure	Insulin	BMI
89.0	66	94.0	28.1
137.0	40	168.0	43.1
78.0	50	88.0	31.0
197.0	70	543.0	30.5
189.0	60	846.0	30.1
…	…	…	…
181.0	88	510.0	43.3
128.0	88	110.0	36.5

	Glucose	BloodPressure	Insulin	BMI
Glucose	1.000000	0.210027	0.581223	0.209516
BloodPressure	0.210027	1.000000	0.098512	0.304403
Insulin	0.581223	0.098512	1.000000	0.226397
BMI	0.209516	0.304403	0.226397	1.000000



Correlation heatmap

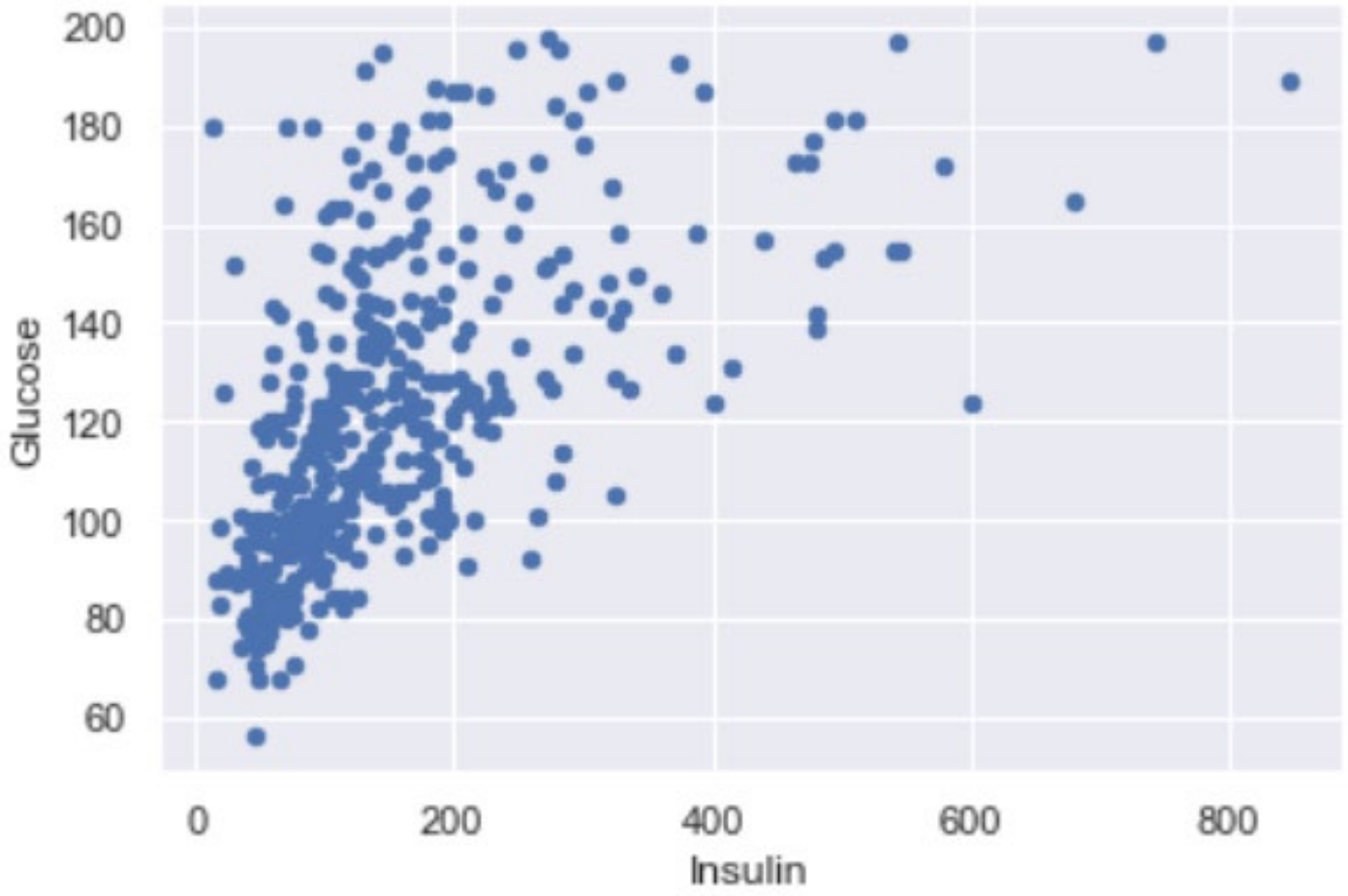
Scatter plot

- ‘The standard way to visualize the relationship between two measured data variables is with a scatterplot.
- The x-axis represents one variable and the y-axis another, and each point on the graph is a record.’

(Bruce and Bruce *Practical statistics for data scientists*, second edition, 2020).

- From a scatter plot, it is possible to spot if the two variables are correlated.

Scatter plot



Key ideas

- ‘The correlation coefficient measures the extent to which two paired variables (e.g., height and weight for individuals) are associated with one another.
- When high values of v1 go with high values of v2, v1 and v2 are positively associated.
- When high values of v1 go with low values of v2, v1 and v2 are negatively associated.
- The correlation coefficient is a standardized metric, so that it always ranges from –1 (perfect negative correlation) to +1 (perfect positive correlation).
- A correlation coefficient of zero indicates no correlation, but be aware that random arrangements of data will produce both positive and negative values for the correlation coefficient just by chance.’

(Bruce and Bruce *Practical statistics for data scientists*, second edition, 2020).

Correlation vs causation

- Causation is about a statement such as:
    - Action *x* causes outcome *y*
- Correlation is just a relationship.
- It is important to bear in mind that if *x* and *y* are highly correlated, it does not necessarily mean *x* causes *y* or *y* causes *x*.
- Correlation and causation can exist at the same time but that is not always the case.
- Example: there might be a high correlation between the sale of cheese in Italy with the number of car accidents in Mexico.