Exploratory data analysis (EDA)

Part 2: Types of data

By: Noureddin Sadawi, PhD

University of London

Types of data

- Data may be classified as either categorical or numerical.
- Categorical data (also known as Qualitative data) uses names or labels to identify the characteristics of an individual.
- Categorical data uses either the nominal or ordinal scales of measurement and may be non-numeric or numeric in nature.

Numerical (or quantitative) data

- Numerical data (also known as quantitative data) have numeric values which indicate how much or how many of a quantity.
- Numeric data use either the interval or ratio scales of measurement.

Categorical variables

- A variable with categorical data is known as a categorical variable.
- A few examples of categorical variables are gender, ethnicity, types of cars, marital status, types of crops.

Quantitative variables

- A variable with numeric values is known as a quantitative variable.
- A few examples of quantitative variables are age, height, weight, number of cars, survival rate, temperature.
- Quantitative variables can be discrete or continuous.

Discrete variables

- Quantitative data that measure how many are **discrete**.
- They usually take integer values (that is, positive or negative whole numbers) but may also take non-integer values.

Examples of discrete variables

- Number of children and number of cars are examples of discrete variables that take integer values.
- Scores in a test where ‘half mark’ is allowed is an example of a discrete variable that takes non-integer values.

Continuous variables

- Quantitative data that measure how much are **continuous**.
- They can take any value on a continuous scale.
- For example, height and time taken to complete a task are continuous variables.

Continuous variables

- In many situations, continuous variables often have values rounded to the nearest integer but are still considered continuous provided there is an underlying continuous scale.
- A good example of this is age of a person.

Statistical analysis suitable for categorical variables

- For categorical variables, data may be summarised by counting the number of observations in each category or by evaluating the proportion of the observations in each category.
- However, in situations where a numeric code is used for categorical data, arithmetic operations such as addition, subtraction, multiplication and division provide no meaningful results.

Statistical analysis suitable for quantitative variables

- Arithmetic operations provide meaningful results for quantitative variables.
- For example, given a quantitative variable such as age, we may compute the average by adding all the data and divide by the total number of observations in order to obtain the average value.
- The average value usually provides meaningful information about the variable and is easy to interpret.

Time-series and cross-sectional data

- Time series data refers to data taken over a range of time periods.
- Cross-sectional data refers to data collected from a number of subjects during a single time period.
- Pictorial example of cross-sectional and time series data. (Use UK election results).

Presentation of data

- Tabular presentation of data.
- In this section, we will look at different ways to summarise data in a table, specifically, Frequency, percentage frequency, relative frequency and cumulative frequency distributions.

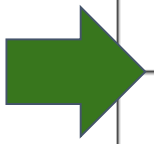
Frequency distribution

- Frequency distribution is a way of summarising data in a table, which shows the possible categories or classes as well as the corresponding number of observations (frequency) in each category.
- The category or classes in a frequency distribution are nonoverlapping.
- Frequency distribution can be used to summarise both qualitative and quantitative data.

Frequency distribution

- Suppose the five most popular brands of cars at a dealership in London are Audi, Ford, Jaguar, Tesla and Vauxhall. The data below are for 20 new purchases of these five brands.

Vauxhall	Ford	Vauxhall	Jaguar
Vauxhall	Ford	Vauxhall	Tesla
Vauxhall	Audi	Ford	Audi
Audi	Tesla	Audi	Jaguar
Tesla	Ford	Vauxhall	Jaguar



Car Brand	Frequency
Audi	4
Ford	4
Jaguar	3
Tesla	3
Vauxhall	6