Summary

Exploratory Data Analysis

The purpose of exploratory data analysis (EDA) is to convert the available data from their raw form to an informative one, in which the main features of the data are illuminated.

- Data can be collected from two main types of studies:
    - Observational study:
        - The treatment variable's values are allowed to occur naturally.
        - Because of the possibility of confounding variables, it is difficult to establish causation.
        - Some confounding variables are difficult to control for; others may not be identified.
    - Experimental study
        - The treatment variable's values are controlled by researchers.
        - Random assignment to treatments automatically controls for confounding variables.
        - A randomized controlled double-blind experiment is generally optimal for establishing causation.
        - In real life, it is sometimes impossible, impractical or unethical to impose some treatments.
- When performing EDA, we should always:
    - use **visual displays** (graphs or tables) plus **numerical summaries**.
    - describe the **overall pattern** and mention any **obvious deviations** from that pattern.
    - **interpret** the results **in context of the data**.
- When examining the **distribution** of a single variable, we distinguish between a **categorical** variable and a **numerical** variable.

- The distribution of a categorical variable is summarized using:
    - Display: pie-chart or bar-chart
    - Numerical summaries: category (group) percentages.
- The distribution of a quantitative variable is summarized using:
    - Display: histogram (or dotplot, mainly for small data sets). When describing the distribution as displayed by the histogram, we should include the:
        - Overall pattern → shape, center, spread.
        - Deviations from the pattern → outliers.
    - Numerical summaries: descriptive statistics (measure of center plus measure of spread):
        - If distribution is symmetric with no outliers, use mean and standard deviation.
        - Otherwise, use the five-number summary, in particular, median and IQR (inter-quartile range).
- Graphs → can be misleading—beware!
- The five-number summary and the 1.5(IQR) Criterion for detecting outliers are the ingredients we need to build the **boxplot**. Boxplots are most effective when used side-by-side for comparing distributions.
- When a distribution has a unimodal symmetric shape, the ***Empirical Rule*** applies. This rule tells us approximately what percent of the observations fall within 1,2, or 3 standard deviations away from the mean.

- When examining the relationship between two variables, the first step is to classify the two relevant variables according to their role and type.

- For two numerical variables, we examine the relationship using:

    - Display: scatterplot. When describing the relationship as displayed by the scatterplot, be sure to consider:

        - Overall pattern → trend/direction, shape, strength.

        - Deviations from the pattern → outliers.

In the special case that the scatterplot displays a linear relationship (and only then), we supplement the scatterplot with:

- Numerical summaries: the correlation coefficient (r) measures the direction and, more importantly, the strength of the linear relationship. The closer r is to 1 (or -1), the stronger the positive (or negative) linear relationship. r is unitless, influenced by outliers, and should be used only as a supplement to the scatterplot.

- When the relationship is linear, we can summarize the linear pattern using the least squares regression line.

    - The slope of the regression line tells us the average change in the response variable that results from a 1-unit increase in the explanatory variable.

    - When using the regression line for predictions, you should beware of extrapolation.

- When examining the relationship between two variables (regardless of the case), any observed relationship (association) does not imply causation, due to the possible presence of confounding variables.

- We can evaluate the regression model with

    - The coefficient of determination (r^2)

    - The residual plot