Summary (EDA)

This summary provides a quick recap of the material you've learned in the Exploratory Data Analysis section. Please note that this summary **does not provide complete coverage** of the material, but just lists the main points. We therefore recommend that you use this summary only as a checklist or a review before going on to the next section, or before an exam.

- 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.
- When performing EDA, we should always:
  - use **visual displays** (graphs or tables) plus **numerical summaries**.
  - describe the **overall pattern** and mention any **striking deviations** from that pattern.
  - **interpret** the results we got **in context**.
- When examining the **distribution** of a single variable, we distinguish between a **categorical** variable and a **quantitative** variable.
- The distribution of a **categorical** variable is summarized using:
  - Display: pie-chart or bar-chart (variation: pictogram → can be misleading—beware!)
  - Numerical summaries: category (group) percentages.
- The distribution of a **quantitative** variable is summarized using:
  - Display: histogram (or stemplot, mainly for small data sets). When describing the distribution as displayed by the histogram, we should describe 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).

- 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 (see also case C→Q in examining relationships).

- In the special case of a distribution having the normal shape, the *Standard Deviation Rule* applies. This rule tells us approximately what percent of the observations fall within 1,2, or 3 standard deviations away from the mean. In particular, when a distribution is approximately normal, almost all the observations (99.7%) fall within 3 standard deviations of 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:

		Response	
		Categorical	**Quantitative**
Explanatory	**Categorical**	C→C	C→Q
	Quantitative	Q→C	Q→Q

  and only then to determine the appropriate tools for summarizing the data. (We don't deal with case Q→C in this course).

- Case C→Q:

  Exploring the relationship amounts to **comparing the distributions** of the quantitative response variable for each category of the explanatory variable. To do this, we use:

  - Display: side-by-side boxplots.
  - Numerical summaries: descriptive statistics of the response variable, for each value (category) of the explanatory variable separately.

- Case C→C:
  Exploring the relationship amounts to **comparing the distributions** of the categorical response variable, for each category of the explanatory variable. To do this, we use:

  - Display: two-way table.

- Numerical summaries: conditional percentages (of the response variable for each value (category) of the explanatory variable separately).

- Case Q→Q:
  We examine the relationship using:

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

    - Overall pattern → direction, form, strength.
    - Deviations from the pattern → outliers.

    Labeling the scatterplot (including a relevant third categorical variable in our analysis), might add some insight into the nature of the relationship.

  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 (as displayed by the scatterplot, and supported by the correlation r), we can summarize the linear pattern using the **least squares regression line**. Remember that:

    - 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 lurking variables.
- When we include a lurking variable in our analysis, we might need to rethink the direction of the relationship → **Simpson's paradox**.
