**Exploratory Data Analysis (EDA)**

**Visualisation**

* **Definition**: The use of graphics to examine data. It's often better than text for conveying information.
* **William Cleveland's Philosophy**:
  + Graphics should maximize information while minimizing cognitive strain.
  + Key tips for clarity:
    - Avoid excessive elements.
    - Use appropriate aspect ratios and scaling.
    - Ensure data is centred and balanced in visuals.
  + Visualization is iterative and aims to answer data-related questions.
  + Different graphics suit different data questions.

**Exploratory Data Analysis (EDA)**

* **Definition**: EDA uses visualization and transformation to systematically explore data.
* **Process**:
  1. Generate questions about the data.
  2. Search for answers through visualization, transformation, and modelling.
  3. Refine questions based on insights gained.
* **Mindset**: EDA is about understanding data, not following strict rules.
* **Good Questions**:
  1. Variation within variables: What types of variation exist?
  2. Covariation between variables: How do variables change together?

**Key EDA Terms**

* **Variable**: A measurable quantity, quality, or property.
* **Value**: The state of a variable at measurement.
* **Observation**: A set of measurements under similar conditions, often called a data point.
* **Tabular Data**: Data in a table format with variables as columns and observations as rows.
* **Tidy Data**: Each value is in its own cell, each variable in its own column, and each observation in its own row.

**Analysing Graphs**

* **Histograms and Bar Charts**:
  + Tall bars indicate common values; short bars show rare values.
  + Gaps indicate absent values.
  + Use unexpected patterns to generate questions:
    - Why are some values common or rare?
    - Are there unusual patterns?
* **Subgroups and Clusters**:
  + Clusters suggest subgroups within data.
  + Questions to explore clusters:
    - Similarities within clusters?
    - Differences between clusters?
    - How to describe clusters?
    - Misleading appearances?
* **Covariation**:
  + Variation: Within a variable.
  + Covariation: Between variables.
  + Use visualization to identify covariation:
    - Continuous vs. categorical variables: Use categorical as a legend.
    - Two categorical variables: Use geom\_count or geom\_tile.
    - Two continuous variables: Use geom\_point, geom\_boxplot, geom\_bin2d, or geom\_hex.

**Questions for Covariation**

* **Patterns**:
  + Is the pattern due to chance?
  + Describe the relationship.
  + Strength of the relationship?
  + Other influencing variables?
  + Does the relationship change across subgroups?

**Histogram and Density Plot**

**Histogram**

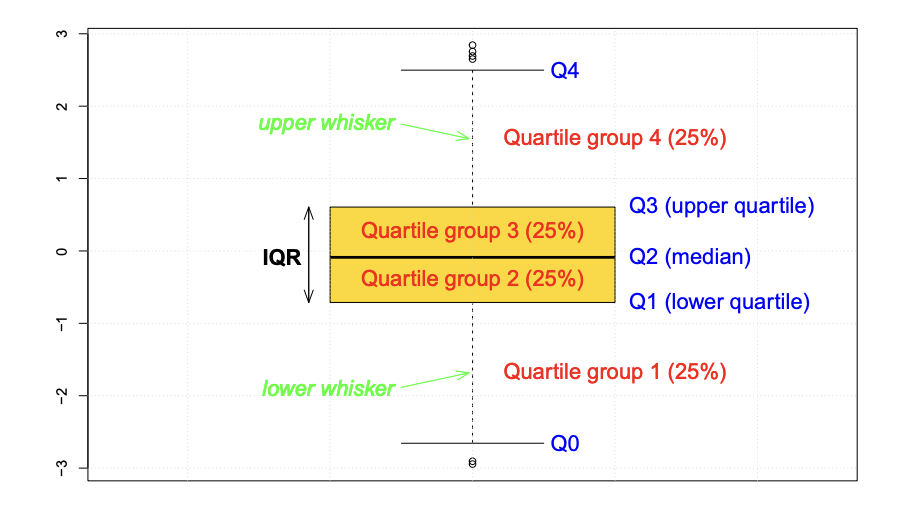
* **Definition**: A graphical representation of the distribution of numerical data using bars to show the frequency of data intervals.
* **Reading a Histogram**:
  + Taller bars indicate more frequent values.
  + No bars indicate values do not present in the data.
* **Common Issues**:
  + **Numerical and Sentinel Values**: Values can be mixed with symbolic codes encoded as numbers.
  + **Solutions**: Convert codes to NA and add Boolean variables to indicate cases.
* **Advantages**:
  + Visualize where data is concentrated.
  + Identify potential outliers and anomalies.
* **Disadvantages**:
  + **Bin width Sensitivity**:
    - Too wide: Lose information about distribution shape.
    - Too narrow: Histogram appears too noisy.

**Density Plot**

* **Definition**: A continuous histogram showing the distribution of data as a smooth curve.
* **Key Points**:
  + The area under the density plot equals one.
  + Focus on the shape of the curve rather than specific y-axis values.
* **Annotated Plot**:
  + Can add text annotations for better interpretation.
  + Example: Indicate where most data is concentrated or highlight subpopulations.
* **Logarithmic Scale**:
  + Used when percent changes or magnitude changes are more significant than absolute changes.
  + Effective for heavily skewed data, ensuring more accurate visualization.

**Box Plot**

* **Definition**: A graphical representation of data through quartiles, highlighting medians, and potential outliers.
* **Components**:
  + **Median**: The midpoint of data, dividing the box into two parts.
  + **Interquartile Range (IQR)**: The range from lower to upper quartile, representing the middle 50% of data.
  + **Whiskers**: Extend beyond the quartiles to indicate variability outside the middle 50%.
* **Interpretation**:
  + Short box plots suggest low variability in data.
  + Tall box plots indicate high variability.
  + Differences between plots warrant further investigation.
* **Advantages over Mean and Standard Deviation**:
  + Robust against outliers and non-normal data.
  + IQR helps identify outliers and skewness.



Why median and IQR are better than mean and standard deviation?A graph of value and value

Description automatically generated

**Bar Chart and Dot Plot**

* **Bar Chart**:
  + Used for discrete data, showing the frequency of categorical variables.
  + **ggplot2** simplifies bar chart creation in R.
  + Horizontal bar charts enhance readability when data is sorted.
* **Dot Plot**:
  + Preferred over bar charts for discrete counts.
  + Avoids misleading perceptions due to bar area differences.
  + Sorted plots support efficient insight extraction.
* **Data Type: Factors**:
  + Used to declare nominal variables in R.
  + Factors store nominal values as integers and map them to character strings.