Chapter 3 Dealing with Numerical Data

3.1 Univariate EDA

Key Concepts

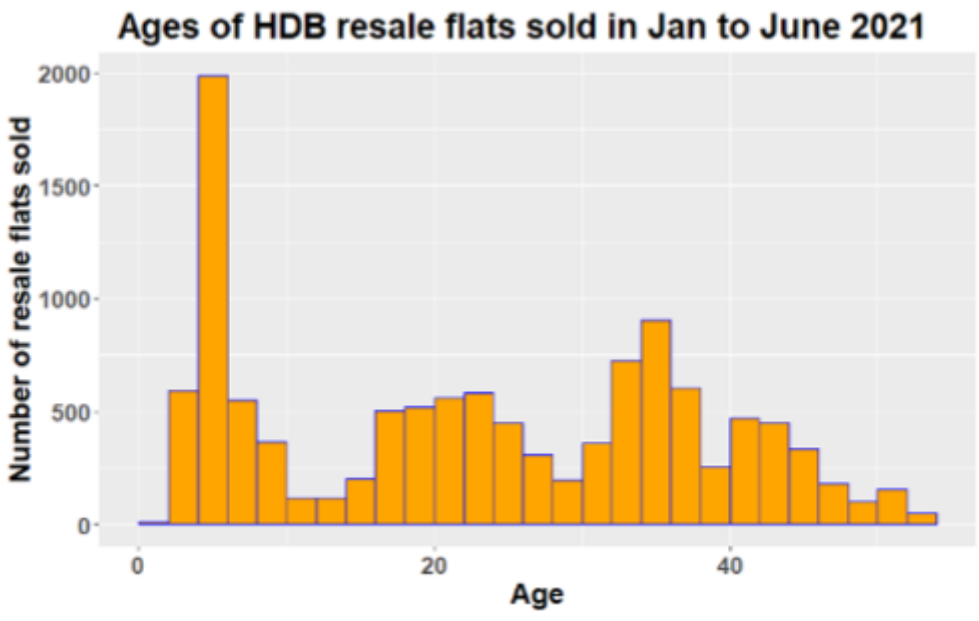
- Numerical variables can be analyzed through **Exploratory Data Analysis (EDA)** to summarize and understand data.
- Focus: Distribution of numerical variables (e.g., Age, Resale Price).
- **Distribution**: A breakdown of data points by their observed number or frequency.

Example: HDB Resale Prices

- Variables:
  - "Month" (time of transaction)
  - "Floor area sqm" (size of flat)
  - "Resale price" (sale value of flat)

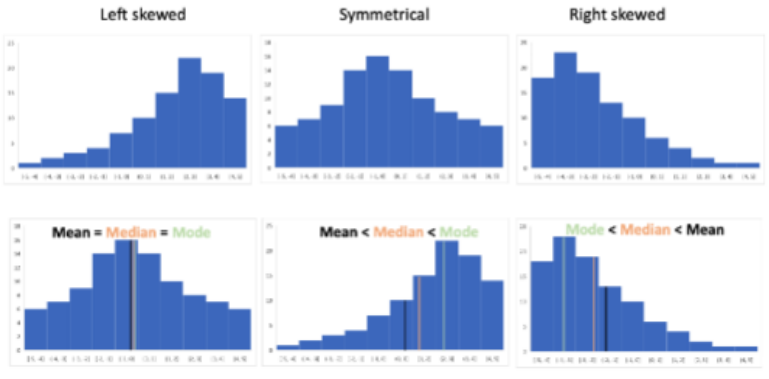
Visualizing Distributions

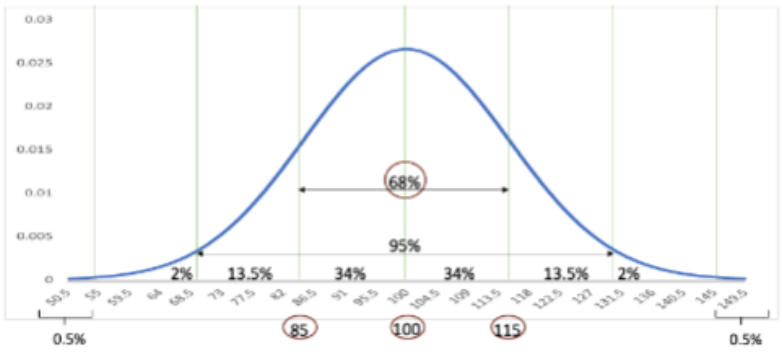
1. **Frequency Tables**: Tabulate counts of each value.
2. **Histograms**: Group values into ranges (bins) for visualization.
   - Example: HDB flat "Age" distribution using bin width = 2 years.
   - Key takeaway: The bin size affects insights drawn from histograms. Experiment with different sizes.

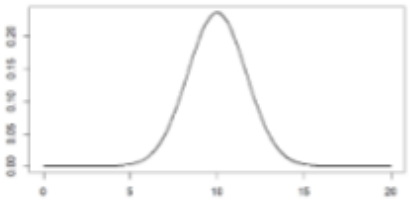


Describing Distributions

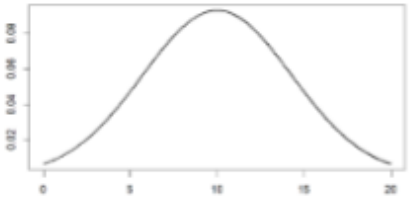
1. **Shape**:
   - **Peaks**: Unimodal, Bimodal, Multimodal.
   - **Skewness**:
     - **Left-skewed**: Long tail on the left.
     - **Right-skewed**: Long tail on the right.
     - **Symmetrical**: Bell curve (e.g., IQ scores).
2. **Center**:
   - **Mean**: Average.
   - **Median**: Middle value.
   - **Mode**: Most frequent value.
   - Relationships:
     - Right-skewed: $\text{mode} < \text{median} < \text{mean}$.
     - Left-skewed: $\text{mean} < \text{median} < \text{mode}$.
3. **Spread**:
   - **Range**: Difference between max and min.
   - **Standard Deviation**: Measure of variability.
   - **IQR**: $IQR = Q3 - Q1$.
   - **Outliers**:
     - Rule:
     $Q1 - 1.5 \times IQR \, (\text{lower bound})$
     $Q3 + 1.5 \times IQR \, (\text{upper bound})$







Low variability
s = 1.69



High variability
s = 4.30

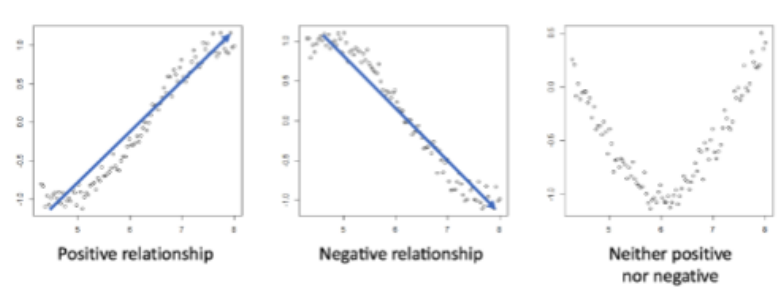
3.2 Bivariate EDA

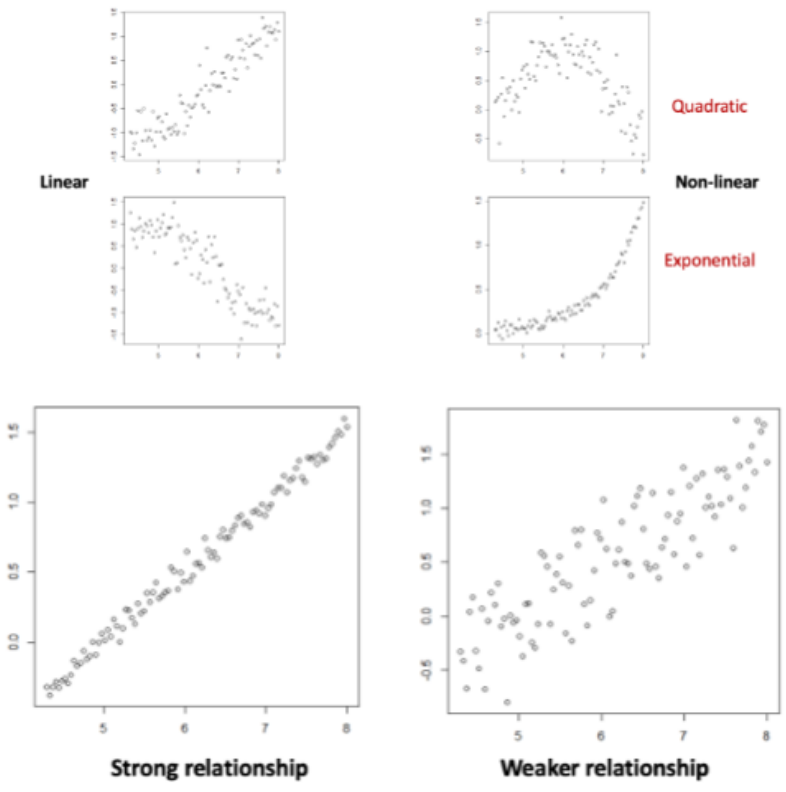
Key Concepts

- Analyze relationships between two numerical variables.
- **Scatter Plots**: Visualize relationships (e.g., Age vs. Resale Price).

Describing Bivariate Data

1. **Direction**:
   - Positive: Both variables increase together.
   - Negative: One increases as the other decreases.
   - None: No clear relationship.
2. **Form**:
   - Linear: Points scatter about a straight line.
   - Non-linear: Points follow a curve (e.g., exponential).
3. **Strength**:
   - Strong: Points closely follow the trend.
   - Weak: Points are widely scattered.





3.3 Correlation Coefficient

Definition

- **Correlation Coefficient (r)**: Measures strength and direction of linear association between two variables.
- Range: $-1 \leq r \leq 1$.
  - $r > 0$: Positive association.
  - $r < 0$: Negative association.
  - $r = 0$: No linear association.

Rules for r:

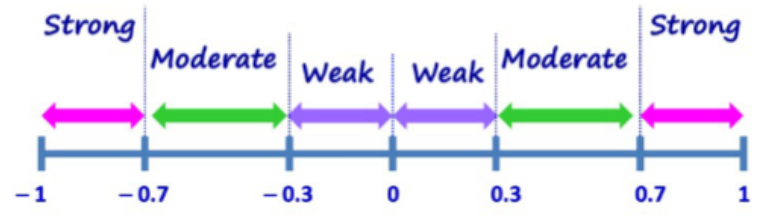
1. Interchanging x and y does not change r.
2. Adding/multiplying a constant to all values of x or y does not change r.

Strength Interpretation

- $|r| \in [0.7, 1]$: Strong.
- $|r| \in [0.3, 0.7]$: Moderate.
- $|r| \in [0, 0.3]$: Weak.

Example:

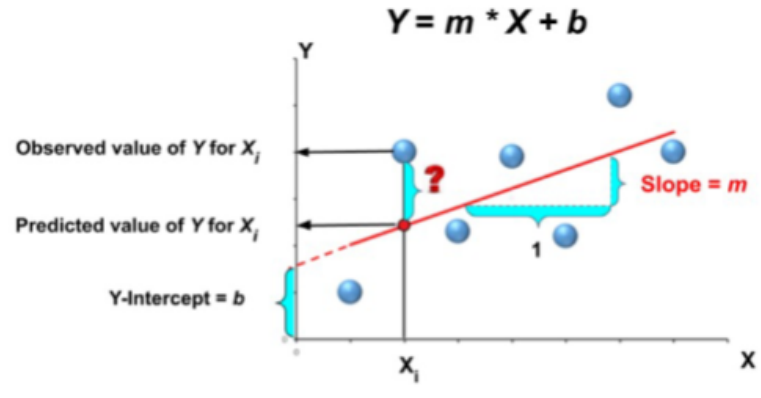
- HDB resale prices and floor area: $r = 0.626$ (strong positive correlation).



3.4 Linear Regression

Definition

- Fit a straight line to describe the relationship between two variables.
- Equation: $Y = mX + b$.
  - m: Slope (rate of change).
  - b: Y-intercept (value of Y when $X = 0$).



Prediction

- Use regression line to estimate Y for a given X.
- Example: Predict resale price of a 40-year-old flat.

$$Y = -4007X + 591857$$

$$Y = -4007 \times 40 + 591857 = 431577$$

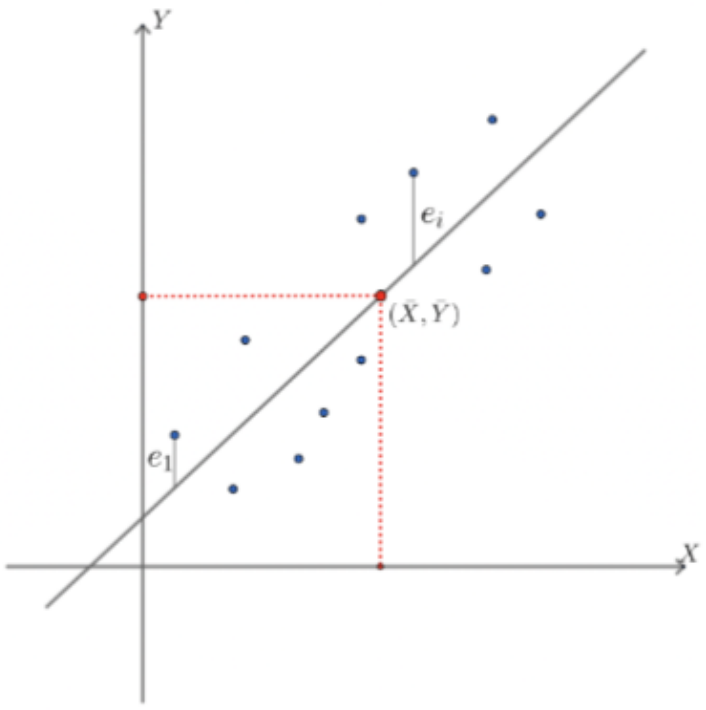
- Interpretation: The average resale price is $431,577 for a 40-year-old flat.

Method: Least Squares

- Minimizes the sum of squared errors:

$$e^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$e_1^2 + e_2^2 + \ldots + e_n^2$$



Limitations

1. Valid only within the observed range of X.
2. Sensitive to outliers.

Summary

1. **Histograms vs. Boxplots**:
   - Histograms: Shape, frequency distribution.
   - Boxplots: Outliers, comparisons.
2. **Bivariate Analysis**:
   - Use scatter plots and correlation coefficient for linear relationships.
   - Check for non-linear associations.
3. **Regression**:
   - Predict values within range of data.
   - Avoid extrapolation.