Summer of Code
Artificial Intelligence
(Machine Learning & Deep Learning)

Instructor
Wajahat Ullah
- *Research Assistant* (DIP Lab)

Duration
03 Months
(September – November)

Day 05 – Exploratory Data Analysis
(Data Visualization)

Objectives:

❖ Data Visualization

❖ Plots for Quantitative Data

❖ Plots for Qualitative Data

❖ Python Libraries

Data Visualization

- The graphical representation of information and data using visual elements such as charts, graphs, and maps.

- It is a critical skill in data science, helping transform raw data into understandable and actionable insights.
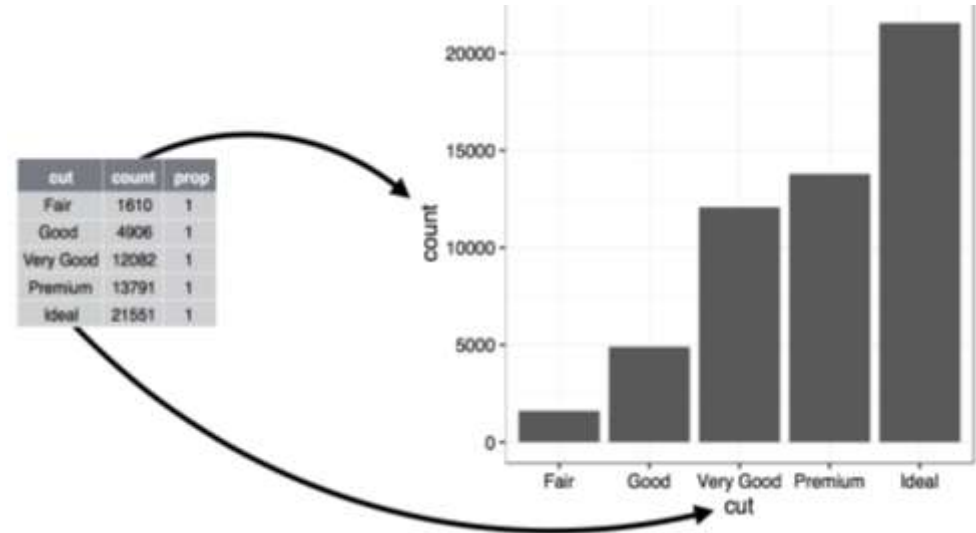
Purpose:

- Helps in understanding complex data through visual context.

- Visualizing data allows people to see relationships, patterns, and trends in the information you're trying to communicate.

# Sources of Data

Primary sources:

- Surveys and questionnaires experiments and observations

Secondary sources:

- Public or private data collections

- Academic research publications and web scraping

Modern Data sources:

- Big Data, Sensors and IoT devices (smart cities etc.)

- Surveillance systems (security cameras)

- Medical Data (electronic health records, medical imaging, etc.)
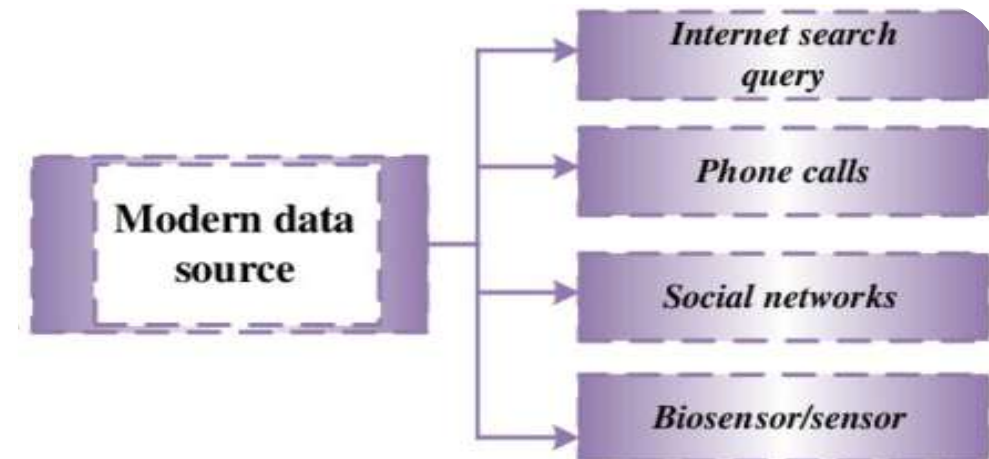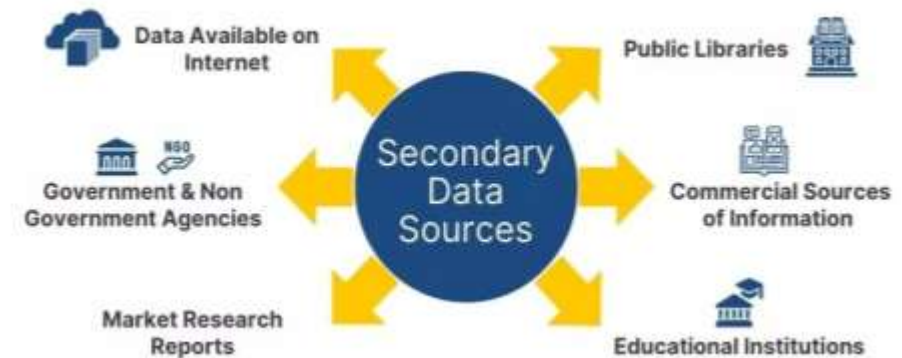
Data Visualization Libraries

Python has become a dominant language in data science, offering a rich ecosystem of visualization libraries.

Matplotlib

- Matplotlib is the grandfather of python visualization packages.
- Foundation for many other libraries.
- It is extremely powerful but with that power comes complexity.
- Best for creating publication-quality static visualizations.

Seaborn

- Leverages matplotlib for beautiful, minimal-code charts.
- It's default styles and color palettes are more modern and visually appealing. Specialized in statistical visualizations.

# Anatomy of Matplotlib Figure
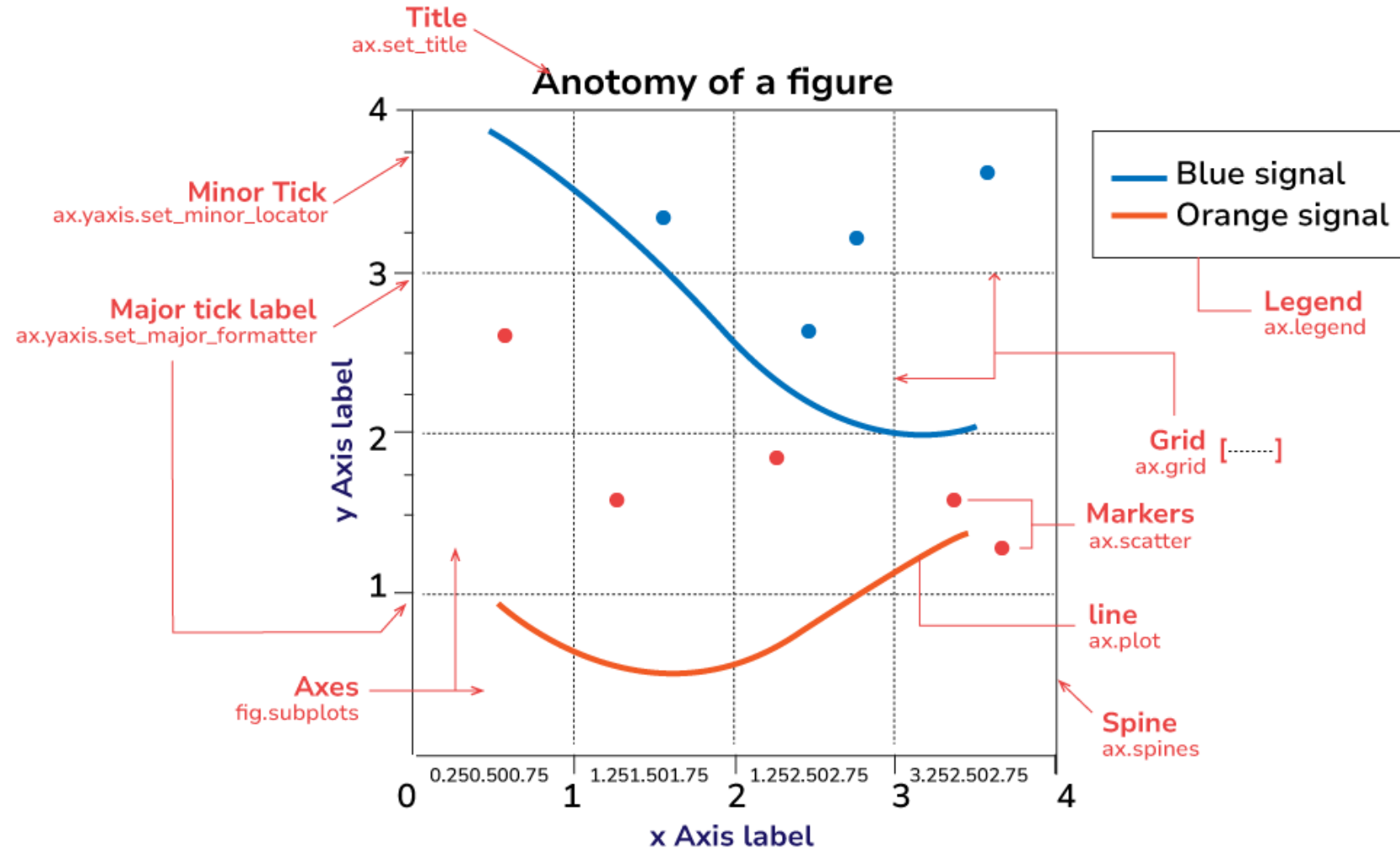
Quantitative Data Plots

Histogram

- A bar chart showing the frequency of a numerical variable within specific intervals.

- **Interpretation:** Look at the shape to understand the distribution (e.g., symmetric, skewed left/right).
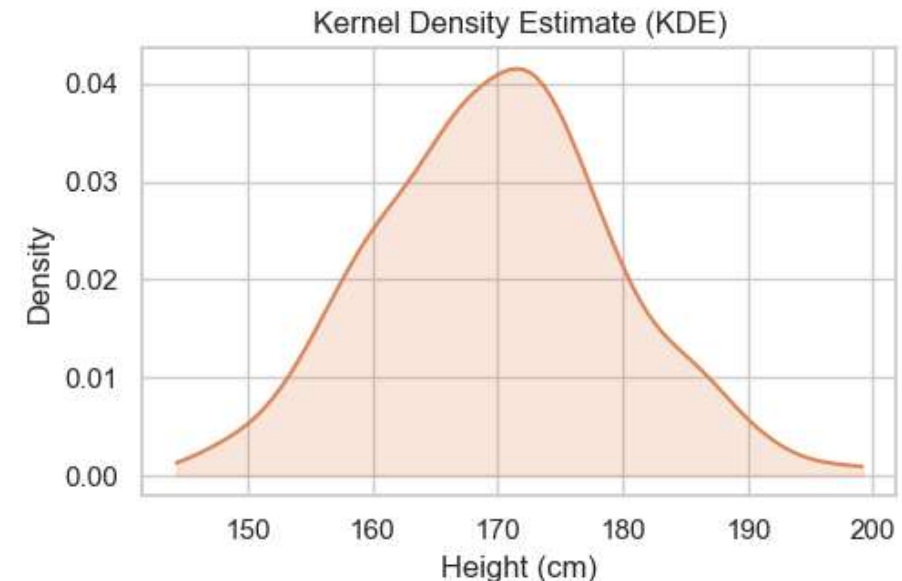
KDE Plot

- A Kernel Density Estimate plot; a smoothed version of a histogram showing a probability curve.

- **Interpretation:** Peaks in the curve represent values with the highest concentration of data. It helps visualize the underlying distribution shape more clearly than a histogram.

Histogram
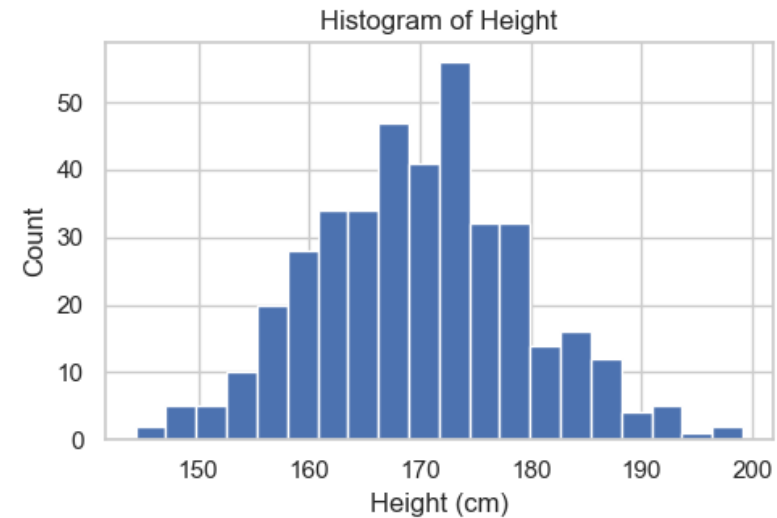
KDE Plot
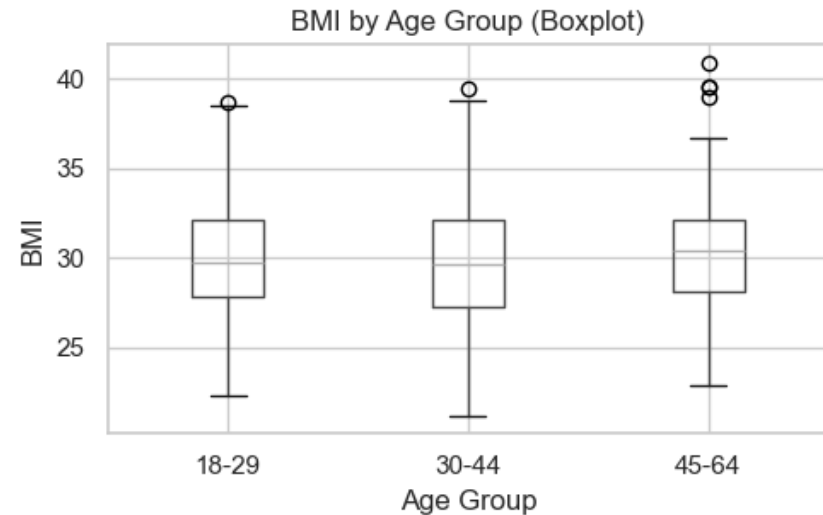
Quantitative Data Plots

Box Plot

- A plot displaying the five-number summary: minimum, first quartile (Q1), median, third quartile (Q3), and maximum.

- **Interpretation:** The central box represents the middle 50% of your data. The line inside is the median (50th percentile). "Whiskers" show the data's range.
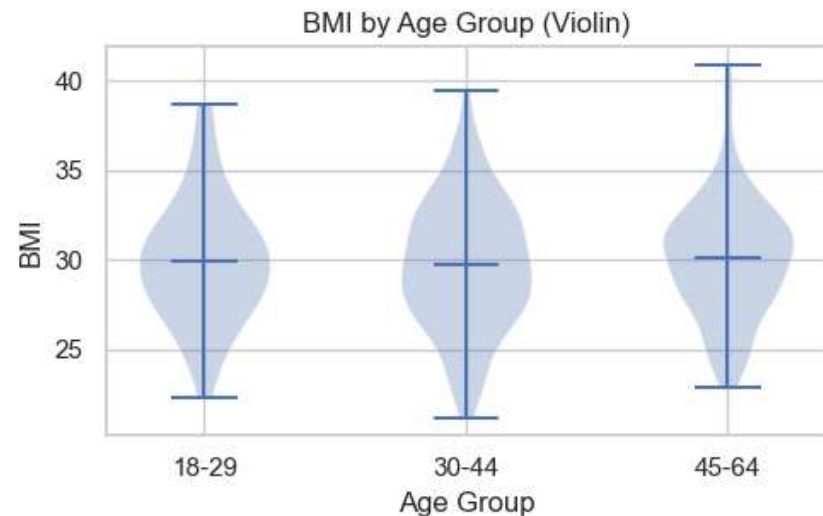
Violin Plot

- A combination of a Box Plot and a KDE plot.

- **Interpretation:** Use the inner box plot for summary statistics. The outer shape's width shows the data's density. Wider sections mean more data points are at that value, which is useful for seeing multiple peaks (bimodal distribution).

Box Plot

BMI by Age Group (Boxplot)

Violin Plot
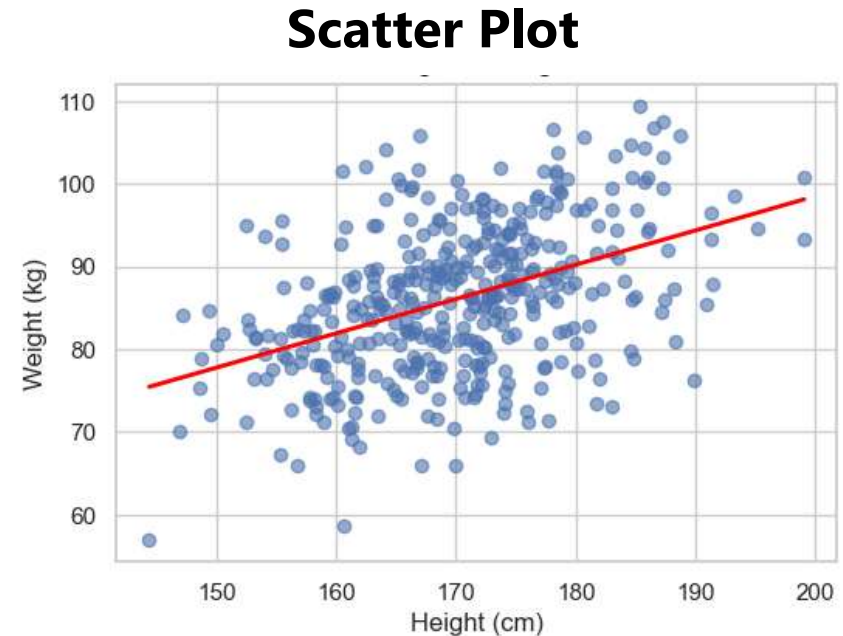

BMI by Age Group (Violin)
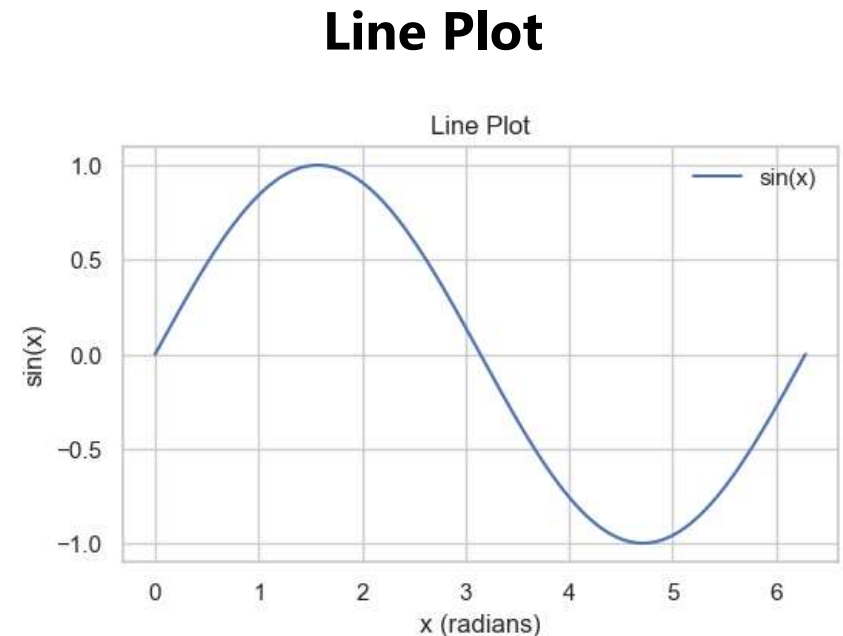
Quantitative Data Plots

Scatter Plot

- A plot that places a dot for each data point based on its value for two different numerical variables.

- **Interpretation:** Look for a pattern. Do the dots trend upwards (positive correlation), downwards (negative correlation), or form no pattern?

Line Plot

- A plot that connects data points with a line, typically used to show a trend over time.

- **Interpretation:** Observe the slope of the line to see if the value is increasing, decreasing, or staying constant.

Scatter Plot

**Line Plot**

Quantitative Data Plots

Heatmap

- A grid where data values in a matrix are represented by different colors.

- **Interpretation:** Use the color bar as a key. Bright or dark colors (depending on the palette) indicate high values or strong correlations. It's perfect for finding patterns in large matrices, like feature correlation.
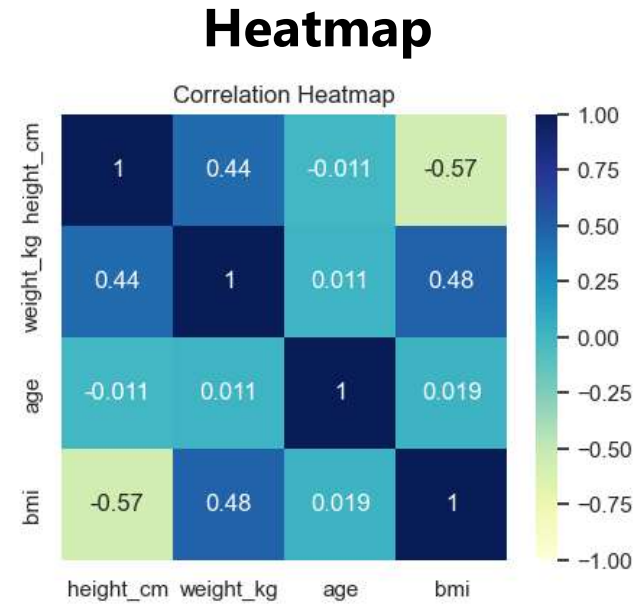
Joint Plot

- Combines a bivariate plot (like a scatter plot) with univariate plots (histograms) on the margins.

- **Interpretation:** Analyze the central plot for the relationship between the two variables. Use the plots on the top and right to understand the distribution of each individual variable separately.

Heatmap

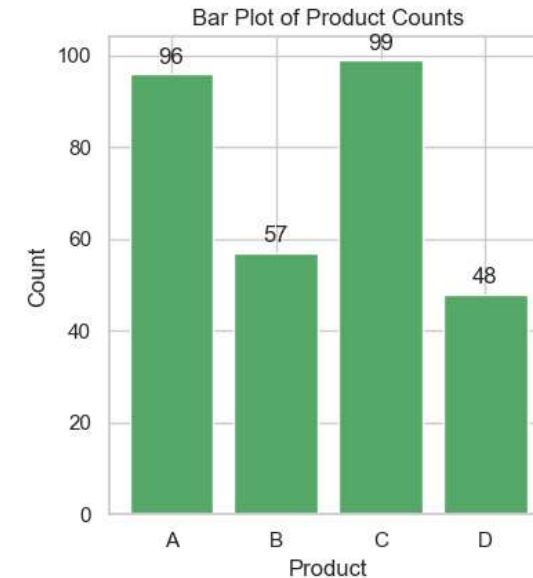**Joint Plot**
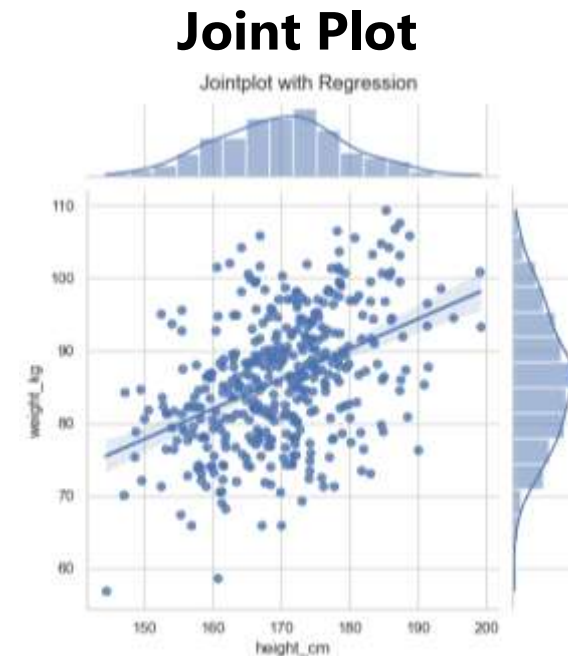
Qualitative Data Plots

Bar Plot

- A plot that shows an aggregate statistic (like the mean) for a numerical variable across different categories.

- **Interpretation:** Compare the bar heights to see how the numerical value differs between categories.
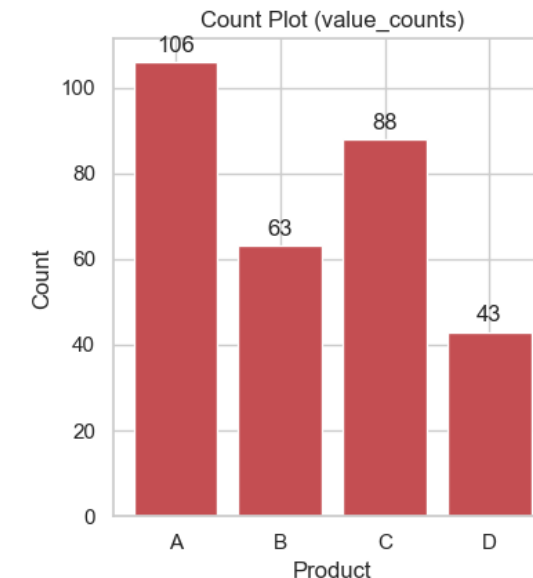
Count Plot

- A bar chart showing the number of observations in each category.

- **Interpretation:** Compare the heights of the bars. A taller bar means that category appears more frequently in the data. It's used to understand the balance of categorical data.

Bar plot

Count Plot

Thank You