DA Data Visualization

Data Visualization Exercise: Exploring and Communicating Data with Plots

Objective

This exercise is designed to help students explore **Data Visualization** techniques in depth using Python libraries such as **Seaborn** and **Matplotlib**. Students will gain hands-on experience in:

- Understanding and visualizing **univariate**, **bivariate**, and **multivariate** data
- Using **categorical and numerical** variables for analysis
- Applying visual encoding such as **hue**, **size**, and **faceting**
- Building storytelling plots to **communicate data insights**

Dataset

Use a dataset that includes a mix of **categorical** and **numerical** variables. You can choose from:

- Built-in Seaborn datasets (`sns.load_dataset()`)
- Public datasets e.g. Kaggle or UCI Machine Learning Repository
- A synthetic dataset created with Python libraries like `numpy`, `pandas`, or `faker`

Example variables to include:

Variable	Type	Example Values
Gender	Categorical	Male, Female
Region	Categorical	North, South, East, West
Age	Quantitative	22, 45, 36, 29
Income	Quantitative	32000, 45000, 88000, 120000
Purchase Amount	Quantitative	19.99, 45.00, 100.50, 250.75
Satisfaction	Ordinal	1 (Low) - 5 (High)

Tasks

Load and Inspect the Dataset

- Load the dataset using Pandas
- Use `.head()`, `.info()`, and `.describe()` to explore the structure and summary statistics
- Identify key variables to visualize

Univariate Visualizations

Categorical Variables

- `countplot()` to show frequency of categories (e.g., Gender, Region)
- Pie chart using `matplotlib` (optional)

Numerical Variables

- `histplot()` for distribution
- `kdeplot()` for density curves
- `boxplot()` for spread and outliers
- `violinplot()` for distribution and symmetry

Bivariate Visualizations

Categorical vs Numerical

- `boxplot()`, `violinplot()`, and `stripplot()` for comparisons by category
- Compare variables like Purchase Amount by Gender or Region

Numerical vs Numerical

- `scatterplot()` for relationship (e.g., Age vs PurchaseAmount)
- Add `hue` to introduce a categorical variable (e.g., Gender)
- `regplot()` to display a regression line
- Calculate the **correlation coefficient** and discuss

Multivariate Visualizations with Faceting

- Use `FacetGrid` or `catplot()` to split plots by additional variables

- Examples:
  - Histogram of Age by Gender
  - Boxplot of Income by Region and Gender
  - Scatterplot of Age vs PurchaseAmount by Region

Deliverables

- A **Jupyter Notebook** or **Python Script** (.ipynb or .py)
- At least:
  - 2 univariate plots
  - 2 bivariate plots
  - 2 faceted or multivariate visualizations
- Visualizations must include **titles**, **labels**, and use clear **color encoding**
- Short commentary for each plot explaining:
  - What is shown
  - Key observations
  - Any patterns or anomalies

Bonus (Optional)

- Use `plotly` or `altair` for interactive plots
- Create a `pairplot()` for a quick multivariate overview
- Try customizing themes and color palettes (`sns.set_palette()`)

Example Questions to Explore

- Do people from different regions have different purchase behavior?
- Is there a relationship between income and satisfaction?
- Do older customers tend to spend more?
- How do purchase amounts vary by gender and satisfaction levels?