Exploratory Data Analysis (2)

Task	Tools & Techniques
~~Understand structure~~	~~.info(), .shape, .columns, .dtypes, .head()~~
~~Summary statistics~~	~~.describe(), mean, median, std, IQR, min/max~~
~~Univariate distributions~~	~~Histograms, boxplots, density plots~~
~~Categorical analysis~~	~~Value counts, bar plots, pie charts~~
Bivariate relationships	Scatter plots, grouped boxplots, violin plots, faceting
Correlation analysis	.corr(), heatmaps, pairplots
Group-wise summaries/Multivariate Visualization	groupby, aggregation by category
Dimensionality reduction	PCA, t-SNE, UMAP (for visualization and structure discovery)
Clustering and segmentation	KMeans, DBSCAN, hierarchical clustering, Gaussian Mixture Models (GMM)
Clustering visualization	PCA/t-SNE scatter plots, cluster heatmaps, dendrograms, parallel coordinates
Hypothesis generation	Observations and questions based on patterns or subgroup differences

Task 5: Bivariate relationship

- A **bivariate relationship** examines how **two variables** are related — whether one tends to change when the other does, or whether group membership affects outcomes.

Type 1	Type 2	Example	Questions to Ask
Numeric	Numeric	Age vs. Fare	Is there correlation? Linear or not?
Categorical	Numeric	Class vs. Fare	Do different classes pay different fares?
Categorical	Categorical	Sex vs. Survived	Are survival rates different by gender?

Pair Type	Recommended Plot(s)
Numeric vs. Numeric	Scatter plot, hexbin, jointplot
Categorical vs. Numeric	Boxplot, violin plot, strip plot
Categorical vs. Categorical	Grouped bar chart, heatmap, mosaic plot

- Scatter Plot: Shows the **relationship** between two numeric variables

Pattern Seen	**Interpretation**
Upward trend	Older passengers tend to pay higher fares
Flat/no pattern	No clear relationship between age and fare
Clusters	Possible subgroups (e.g., children, seniors)
Outliers	A few paid unusually high fares (e.g., wealthy passengers)

- Joint Plot: Combines a **scatter plot** with **histograms or KDEs** (kernel density estimates) for both variables.
- LM Plot (Linear Model Plot): Plots a **scatter plot** with a **fitted regression line**.
- Hexbin:Divides the 2D space into **hexagonal bins** and **counts** how many points fall into each.

Feature	Meaning
Darker hexagons	More data points in that age–income range
Linear alignment	Suggests a trend or correlation
Scattered brightness	Indicates noise or spread
Empty regions	No observations in those ranges

- Box Plot (Categorical vs. Numeric):Summarizes the distribution of a **numeric variable** grouped by a **categorical variable**.
- Violin Plot (Categorical vs. Numeric)
- Strip Plot (Categorical vs. Numeric): Plots **each individual observation** as a **dot** along a categorical axis; Helps you **see all values**, detect **clusters**, and spot **overlaps**.
- Swarm Plot (Categorical vs. Numeric): **neatly arranged strip plot, avoids overlap**
- Grouped Bar & Stacked Bar (Categorical vs. Categorical)**:** Visualizes the **frequency or proportion** of combinations of two categorical variables.
- Heatmap of Crosstab (Categorical vs. Categorical)**:** Summarizes **counts or proportions** of two categorical variables in a grid**;** Each cell shows how **often a combination occurs;** Heatmap coloring makes **differences easy to spot visually**.

More examples

- In practice, one does not have to use ALL methods
- Often one tries a few and pick the ones making more senses
- https://www.kaggle.com/datasets/maharshipandya/-spotify-tracks-dataset/data

Task 6: Correlation analysis

- **Correlation analysis** measures the **strength and direction** of a linear relationship between pairs of **numeric variables**.
- Pearson Correlation Coefficient

Correlation Value	**Interpretation**
> 0.8 or < -0.8	Very strong
0.6 – 0.8 or -0.6 – -0.8	Strong
0.4 – 0.6 or -0.4 – -0.6	Moderate
0.2 – 0.4 or -0.2 – -0.4	Weak
≈ 0	Very weak or no linear relationship

Example

- Not much linear relationship

- A lot of linear relationship

Task 7: multivariate visualization

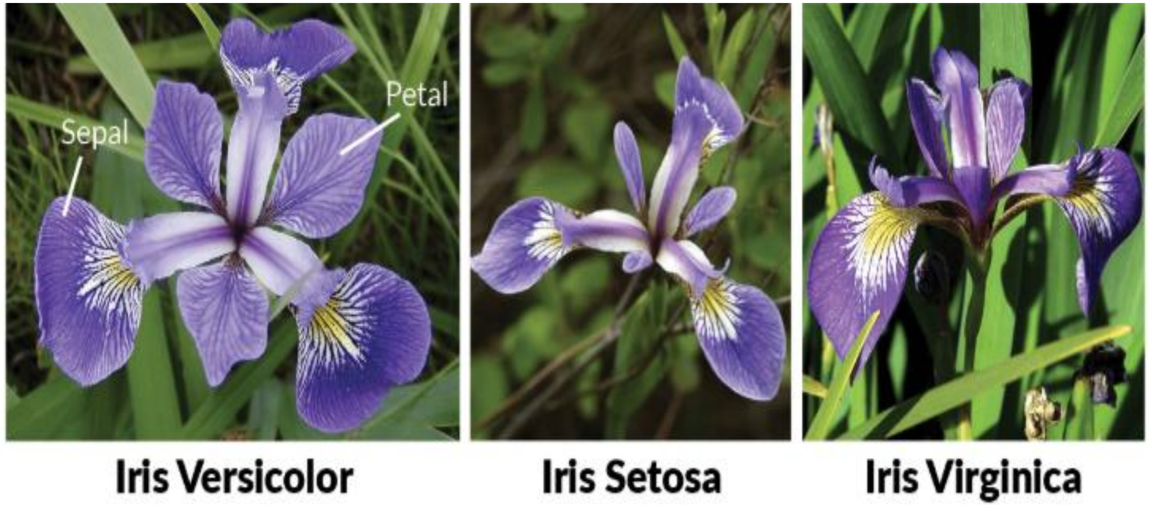
Go beyond pairwise comparisons to **simultaneously explore multiple variables** (numeric and/or categorical), revealing:

- **Interactions between features**
- **Group patterns** across multiple dimensions
- Hidden **subgroup structures** that wouldn't show up in bivariate plots

Method	**Type of Variables**	**Highlights**
Enhanced Scatterplot	numeric + categorical	Quick overview of multiple variable influences
Pairplot with Hue	numeric + categorical	Class-wise exploration across many variables
FacetGrid / Catplot	numeric × categorical	Plots conditioned by group
Parallel Coordinates	numeric + class label	Feature profile comparison
3D Scatterplot (optional)	numeric × 3	Visualize tri-variable structure (interactively)

- Enhanced Scatterplot: A 2D scatterplot enriched with **hue**, **size**, and **style** encodings; Encode up to **4 variables** (x, y, color, size/shape); Ideal for continuous + categorical variable combinations
- Pairplot (with Hue)： All pairwise scatterplots with **grouping information** overlaid; hue to highlight **subgroup differences**
- FacetGrid & Catplot: Grid of plots split by one or two **categorical variables**
- Parallel Coordinates**:** Each observation is a line plotted across multiple numeric axes.
- 3D Scatterplot

Example: Iris dataset



- **150 observations** of iris flowers
- **3 species** of iris:
- *setosa*
- *versicolor*
- *virginica*
- **4 numeric features** measured in centimeters:
- sepal length (cm)
- sepal width (cm)
- petal length (cm)
- petal width (cm)

Task 8: Dimensionality Reduction & Structure Discovery

- **Dimensionality reduction** is the process of **transforming high-dimensional data into a lower-dimensional representation**, typically for downstream tasks
- To project high-dimensional data into **lower-dimensional space (2D or 3D)** in order to:
- **Visualize structure** that's hard to see in raw features
- **Identify clusters, patterns, or anomalies**
- **Prepare data** for further modeling (e.g. classification, clustering)

Type	Examples	Preserves…
Linear	PCA, MDS	Global structure (variance/distance)
Manifold-based	Isomap, Laplacian Eigenmap, *Diffusion Maps, *LLE	Geometry or topology of curved spaces
Probabilistic	t-SNE, *UMAP	Local neighborhood structure
Neural	Autoencoders	Learned compressed representation

- PCA, MDS: looks for **orthogonal directions (axes)** in the data where the **variance is largest**
- Isomap, Laplacian Eigenmap, Diffusion Maps, Locally Linear Embedding: All four are **manifold learning** techniques that aim to **unfold or unroll** a nonlinear manifold embedded in high-dimensional space.
- t-SNE, U-map: Minimizes **Kullback-Leibler (KL) divergence** between high-D and low-D similarity distributions.
- Autoencoders: An **autoencoder** is a type of artificial neural network used to learn **efficient representations** (i.e., encodings) of data, typically for the purpose of **dimensionality reduction** or **denoising**.

Example

Dataset one: https://scikit-learn.org/1.5/auto_examples/datasets/plot_digits_last_image.html
Dataset two: https://rasbt.github.io/mlxtend/user_guide/data/wine_data/