Exploratory data analysis (3)

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 9: Clustering and Segmentation

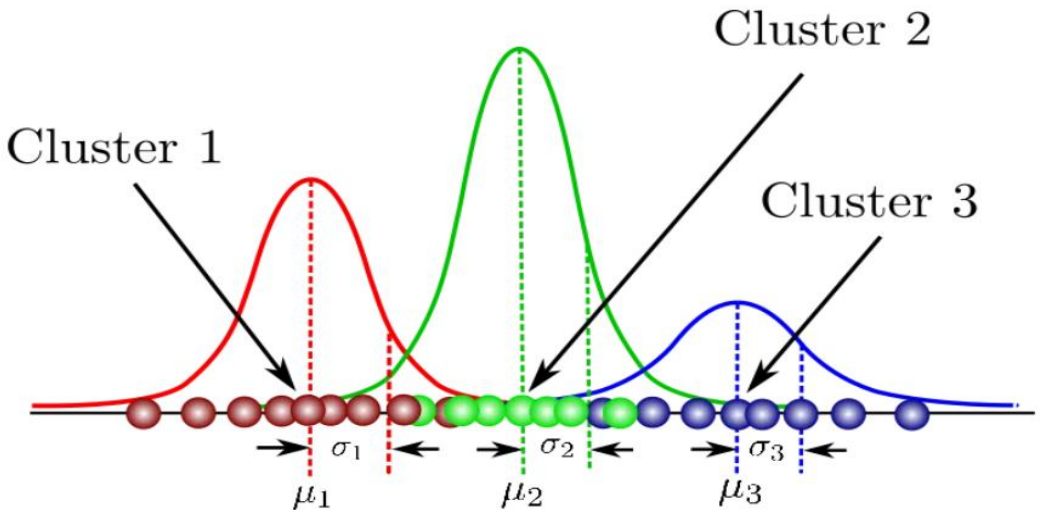
- **Clustering** is the task of **grouping data points** such that those in the same group (cluster) are more similar to each other than to those in other groups.
- **Segmentation** often refers to applying clustering to real-world use cases — like customer segmentation, image segmentation, or genetic subtype discovery.

Algorithm	Type	Notes
K-Means	Centroid-based	Assumes spherical clusters, fast, widely used
Agglomerative Clustering	Hierarchical	Builds a tree of clusters, no need to pre-specify k
DBSCAN	Density-based	Finds arbitrarily shaped clusters, detects noise
Spectral Clustering	Graph-based	Good for non-convex clusters using graph Laplacian
Gaussian Mixture Models (GMM)	Probabilistic	Allows soft assignments (probabilities instead of labels)

- **K-means:** https://www.naftaliharris.com/blog/visualizing-k-means-clustering/
- **Agglomerative Clustering:** https://online.stat.psu.edu/stat555/node/86/
- **DBSCAN (Density-Based Spatial Clustering of Applications with Noise):**

https://www.naftaliharris.com/blog/visualizing-dbscan-clustering/

- **Spectral clustering**: Dimension reduction+K-means
- **GMM:**



Example: Real data and simulated data

- Real data: Iris dataset
- Simulated mixture of Gaussian

Segmentation

Segmentation typically includes **3 steps**:

1. **Clustering**: Find groups using unsupervised learning
2. **Profiling**: Describe each segment using summary stats (e.g., mean income, age)
3. **Application**: Use segments to tailor decisions (e.g., personalized marketing)

Image segmentation

Image segmentation means dividing an image into **regions (segments)** based on similarity in pixel values, texture, or other features.

Each pixel is assigned to a **segment or object**, producing a simplified, structured version of the image.

- Load an image
- Convert it into a 2D array of RGB values
- Apply **K-Means clustering** to group pixels by color
- Reconstruct a **segmented version** using cluster labels

Task 10: Clustering Visualization

- Clustering algorithms assign data points to groups — but to **understand**, **trust**, and **refine** these results, **visualization is essential**.

Technique	Purpose
2D/3D scatter plots	Visualize clusters in reduced space (e.g., PCA, t-SNE)
Cluster heatmaps	Show patterns within clusters across features
Elbow/Gap methods	Help determine the optimal number of clusters
Dendrograms	Visualize hierarchical clustering

Task 11: Hypothesis generation

Hypothesis generation is the process of:

- **Formulating questions or testable ideas** based on patterns, trends, or anomalies observed during EDA.
- These hypotheses can then guide further **statistical testing**, **model development**, or **experimental design**.

It's the bridge between **exploration** and **explanation**.

Dataset	Observation	Generated Hypothesis
Iris	Petal length varies by species	"Petal length differs significantly across species"
Titanic	Higher survival among women/children	"Gender and age affect survival rate"
Customer data	One cluster has high spending	"This customer segment responds better to promos"