Exploratory Data Analysis (1)

Introduction

- Exploratory Data Analysis (EDA) is the process of investigating datasets to **summarize their main characteristics** — typically through **visualization** and **statistics** — before formal modeling or inference.
- It's like being a detective: you don't jump to conclusions; instead, you explore patterns, spot anomalies, and generate hypotheses.
- Understand your dataset's structure and contents
- Detect **missing values**, **outliers**, and **errors**
- Identify **patterns**, **trends**, and **relationships**
- Make decisions about **data cleaning** and **feature selection**
- Choose the right modeling techniques

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 analysis	groupby, aggregation by category
Dimensionality reduction	PCA, t-SNE, DNN (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 1: Understand Structure

Get a quick overview of the dataset's structure:

- Dimensions (rows × columns)
- Column names and types
- Glimpse at the data values

Feature	Why it matters
Number of rows	Affects computational choices
Number of columns	Helps judge complexity
Data types	Guides visualization and modeling
Column names	Watch for typos, ambiguous or duplicated names
Glimpse of values	Spot strange formats, early signs of dirty data

Example: real data and simulated data

- Dataset one: Titanic - Machine Learning from Disaster
- https://www.kaggle.com/c/titanic/data
- Dataset two: simulated dataset
- Customer behavior on some product

Figure out:

- How many rows and columns are there?
- Which variables are numerical? Which are categorical?
- Are there columns you'd ignore (e.g., ID, ticket)?
- Do any values look unusual or unexpected

Task 2: Summary Statistics

Use descriptive statistics to **summarize each variable** in the dataset. This provides insight into:

- Central tendency (mean, median)
- Spread (standard deviation, IQR)
- Min/max and range
- Possible skewness or outliers (visually or numerically)

Statistic	Interpretation Example
Mean/Median	Is the variable skewed? Mean >> median?
Min/Max	Are there suspiciously high/low values?
Std/IQR	How spread out is the data?
Counts	Are some categories rare? Balanced?
Unique Values	Detect ID columns or categorical variables

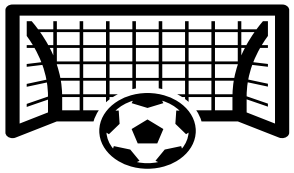
Why they are helpful?

- Understand Central Tendency
- Measure Spread / Dispersion
- Identify Outliers

Feature	Mean	Std Dev	Min	Max
Price ($)	450K	120K	150K	950K
Sqft	1900	550	800	5000
Bedrooms	3.2	0.9	1	6

- Houses vary a lot in price and size (→ diverse market).
- Bedrooms are less variable (most homes are 3BR).
- Price per sqft can be computed to identify high-value neighborhoods.

Example



Age, fare, sex, pclass, embarked

Column	What to Look For	Possible Insight
age	Is median lower than mean? Are many children onboard?	Skewed age distribution; prioritize young passengers in rescue?
fare	High std? Large difference between min and max?	Wide fare range → socioeconomic diversity
sex	More males or females?	Gender distribution might relate to survival rate
pclass	Which class had most passengers?	Most were in 3rd class → may link to survival/access differences
embarked	Are most passengers from one location?	Could suggest boarding preference/location demographics

Task 3: Univariate Distributions

A **univariate distribution** describes how the values of a **single variable** are distributed across the dataset.

1. The Typical Values: Are most observations centered around a mean or median?
• Example: Most Titanic passengers were in their 20s–40s.

2. Spread and Variability: Is the variable tightly concentrated or widely spread? This affects model performance, normalization choices, and more.
• Example: If income varies from $10K to $200K, it's highly spread and may need transformation.

3. Skewness and Shape Is the distribution symmetric or skewed (long tail on one side)?
• Example: fare on the Titanic is **right-skewed —** most people paid little, but a few paid a lot.
• Skewed variables may need **log or power transformations**.

4. Presence of Outliers: Outliers may reflect data entry issues or interesting rare cases.
• Visual tools like **boxplots** help identify these.

5. Modality: Does the distribution have one peak (unimodal) or several (bimodal/multimodal)?
• Multimodal distributions suggest **subgroups** in the data (e.g., male/female heights).

6. Categorical Balance: For categorical variables, univariate analysis reveals **class balance**.
• Useful for classification problems (e.g., is purchased 90% zeros and only 10% ones?)

Variable Type	Plot Type	Description
Numeric	**Histogram**	Shows frequency distribution; use bins to control resolution
Numeric	**KDE plot**	Smooth estimate of the distribution; useful for seeing shape
Numeric	**Boxplot**	Shows median, IQR, and potential outliers
Numeric	**Violin plot**	Combines boxplot and KDE; great for comparing groups
Numeric	**Rug plot**	Small ticks to show actual data points (usually combined with KDE)
Numeric	**ECDF plot**	Empirical cumulative distribution function — shows quantile behavior
Categorical	**Bar chart**	Frequency of each category (count-based)
Categorical	**Pie chart** (limited)	Proportional view, best with few categories
Categorical	**Waffle plot**	Grid-style proportion visualization (requires pywaffle or Plotly)
Categorical	**Treemap**	Area-based representation for many categories (squarify or Plotly)

- A **histogram** shows the **frequency** (count) of data points falling within specified ranges (“bins”). It's a bar chart for continuous/numeric variables.

- Shape of the distribution, Skewness (left/right), Gaps, spikes, and multiple peaks (modality), Outliers (if bins go far into tails)
- Choosing bin width is important (too few = oversimplified; too many = noisy)
- A **KDE plot** estimates the **probability density function (PDF)** of a continuous variable. Think of it as a smooth version of a histogram — it shows where values are concentrated without binning.
- A **boxplot** summarizes the **distribution of a numeric variable** using five key statistics: Minimum,1st quartile (Q1), Median (Q2), 3rd quartile (Q3), Maximum (excluding outliers)

- A **violin plot** is a hybrid of a **boxplot** and a **KDE plot: detailed distribution info + summary**
- An **ECDF plot** shows the proportion of observations **less than or equal to** each data point. It's a cumulative version of a histogram or KDE: **percentile thresholds** (e.g., what value corresponds to the 90th percentile?); compare **two or more distributions**
- A **bar chart** displays the **frequency (or proportion)** of each category in a categorical variable: Class imbalance in classification tasks
- A **treemap** is a space-filling visualization where each category is represented as a **rectangle**, and its area is proportional to the **frequency or size** of that category.

Task 4 Categorical Analysis

- Categorical analysis is the process of exploring and summarizing variables that represent **discrete groups or labels**, rather than numeric values. These variables tell you *what type or class* an observation belongs to.

Insight	Example from Titanic
Class balance	How many passengers are male vs. female
Dominant groups	Most passengers embarked from Southampton
Group sizes	value_counts() shows how common each is
Relationships to other variables	Survival rate by class, sex, or embarked
Feature usefulness	A column with 1 unique value is not useful

Task	Tools / Code Example
Count categories	df['var'].value_counts()
Proportions	df['var'].value_counts(normalize=True)
Number of unique categories	df['var'].nunique()
List categories	df['var'].unique()
Bar chart	sns.countplot(x='var', data=df)
Pie chart	df['var'].value_counts().plot.pie()
Group comparison	df.groupby('category')['other_var'].mean()

- Frequency Counts
- Proportions
- Number of Unique Categories
- Unique Values
- Grouped Summaries

Example

		Survival Rate by Sex	Survival Rate by Class
1	female	1.0	
2	male	0.2	
3	First		1.0
4	Second		1.0
5	Third		0.3333333333333333