Red wine dataset EDA:

Colab:

https://colab.research.google.com/drive/1wQamTcPpqRcGyYLbL9K7W9B0Qnw8jFyx#scrollTo=G0n54bJFw4Mc

What is EDA in Machine Learning?

Exploratory Data Analysis (EDA) is the process of understanding, summarizing, and visualizing data before applying ML models.

👉 Goal:

- Discover patterns
- Detect anomalies
- Test assumptions
- Understand relationships
- Prepare data for modeling

EDA is **not optional** — most ML performance gains come from good EDA, not complex models.

Where EDA Fits in ML Pipeline

Data Collection

↓

Exploratory Data Analysis (EDA) ← YOU ARE HERE

↓

Feature Engineering

↓

Model Training

↓

Evaluation & Deployment

Core Components of EDA (Basics → Advanced)

1️⃣ Understand the Dataset

Key Questions

- How many rows and columns?
- What does each column represent?
- What is the target variable?
- What are data types?

Data Types Analysis

Type	Examples	Handling
Numerical	Age, Salary	Scaling, outlier handling
Categorical	Gender, City	Encoding
Ordinal	Ratings (Low–High)	Label encoding
Datetime	Date	Feature extraction

Missing Values Analysis

Why Important?

- Missing data can bias models
- Some algorithms cannot handle NaNs

Techniques

- Count missing values
- Percentage missing
- Patterns of missingness

Statistical Summary

Helps understand **distribution & spread**.

df.describe()

Key statistics:

- Mean
- Median
- Standard Deviation
- Min / Max

- Quartiles

Univariate Analysis (One Variable)

Numerical Features

- Histogram
- Boxplot
- KDE

Purpose:

- Distribution shape
- Skewness
- Outliers

Categorical Features

- Count plots
- Value counts

Bivariate Analysis (Two Variables)

Used to understand **relationship between variables**.

Examples:

- Numerical vs Numerical → Scatter plot, correlation
- Categorical vs Numerical → Boxplot
- Categorical vs Categorical → Crosstab

7️⃣ Multivariate Analysis

Understand **interactions among multiple variables**.

Tools:

- Correlation matrix
- Pair plots
- Heatmaps

8️⃣ Outlier Detection

Outliers can:

- Skew model performance
- Affect distance-based models

Detection:

- Boxplots

- IQR method
- Z-score

9️⃣ Feature Relationships & Correlation

Correlation values:

- +1 → Strong positive
- -1 → Strong negative
- 0 → No linear relation

⚠️ Correlation ≠ Causation

🔟 Insights & Hypothesis Generation

EDA ends with:

- Observations
- Hypotheses
- Decisions for preprocessing & modeling