📆 DAY 13 — Exploratory Data Analysis (EDA)

Goal : Understand your data before training models

1️⃣ What Is EDA? (Very Important Concept)

EDA = Exploring your data to understand:

- What values exist?
- Are there errors?
- Are there patterns?
- Are there outliers?

> 80% of ML problems are data problems, not model problems.

2️⃣ First things You ALWAYS Do

Step 1 — Look at the data

```
df.head()
df.tail()
```

Step 2 — Size & structure

```
df.shape()
df.info()
```

Why this matters

- Are there missing values?
- Are data types correct?
- Is dataset too small / too large?

3️⃣ Summary Statistics (know Your Numbers)

```
df.describe()
```

This gives:

- Mean
- Std
- Min / Max
- Quartiles

🧠 Important Insight

If `max` is **far bigger** than `75%` → likely outliers.

4️⃣ Distribution of Data (very IMPORTANT)

Why distributions matter

ML models assume:

- Reasonable ranges
- No extreme skew

Histogram (Most common EDA plot)

```
import matplotlib.pyplot as plt

df["salary"].hist(bins=20)
plt.show()
```

🧠 What to look for

- Is data skewed?
- Are there long tails?
- Multiple peaks?

5️⃣ Outliers (Critical Concept)

Simple check

```
df.boxplot(column="salary")
plt.show()
```

🧠 **Why outliers matter**

- Can dominate loss function
- Break linear models
- Affect scaling

6️⃣ Relationship Between Features

Scatter plot

```
df.plot.scatter(x="age", y="salary")
plt.show()
```

🧠 **What you learn**

- Linear relationship?
- Clusters?
- Noise?

7️⃣ Correlation (IMPORTANT FOR FEATURE SELECTION)

```
df.corr()
```

Heatmap (Visual)

```
import seaborn as sns

sns.heatmap(df.corr(), annot=True)
plt.show()
```

🧠 **Key Insight**

- High correlation → redundant features
- Target correlation → useful features

8️⃣ Categorical Feature Analysis

```
df["city"].value_counts()
```

plot:

```
df["city"].value_ciounts().plot(king="bar")
plt.show()
```

🧠 **Why this matters**

- Class imbalance
- Rare categories
- Encoding decisions

9️⃣ EDA in ML Workflow (Big Picture)

```
Raw Data
 ↓
EDA ← 🔥 Most IMPORTANT
 ↓
Cleaninig
 ↓
Feature Engineering
 ↓
Modeling
```

Skipping EDA = guessing.
