📅 DAY 14 — Feature Engineering

Goal : Transform raw data into useful ML Features

1️⃣ What Is Feature Engineering? (IMPORTANT)

Feature Engineering = Creating new, better features from raw data.

Model **do not think.**

They only learn from **what you give them**.

> Better feature > Better model

2️⃣ Why Feature Engineering Matters More Than Models

you can:

- Use a simple model + good features → ✅ great results
- Use a complex model + bad features → ❌ bad results

This is why **senior ML engineers focus here**.

3️⃣ Feature Types (Know This)

Feature Type	Example
Numeric	age, salary
Categorical	city, gender
Ordinal	rating (low, medium, high)
Datetime	date, time
Binary	yes/ no

Each type needs **different handling**.

4️⃣ Creating New Features (MOST COMMON)

Example : Experience from dates

```
df["experience_year"] = 2025 - df["start_year"]
```

🧠 Why this helps

Raw year ≠ meaningful.

Experience = Useful signal.

5️⃣ Binning (Discretization)

Convert numeric → categories

```
df["age_group"] = pd.cut(
    df["age"],
    bins=[0, 25, 40, 60],
    labels=["young", "adult", "senior"]
)
```

🧠 Why this helps

- Reduces noise
- Captures non-linear patterns

6️⃣ Encoding Categorical Features (Review + Insight)

One-Hot Encoding

```
pd.get_dummies(df, columns=["city"]
```

Ordinal Encoding (Order Matters)

```
df["education_level"] = df["education"].map({
    "High School" : 1,
    "Bachelor" : 2,
    "Master" : 3
})
```

🧠 Important

Never use ordinal encoding unless **order is real**.

7️⃣ Feature Interaction (POWERFUL)

Combine features

```
df["salary_per_year"] = df["salary"] / df["experience"]
```

🧠 Why this matters

Models often cannot discover interactions by themselves.

8️⃣ Log Transformation (Outliers Fix)

```
df["salary_log"] = np.log1p(df["salary"])
```

🧠 Why

- Reduces skew
- Makes distributions more normal
- Helps linear models

9️⃣ Scaling (Quick Reminder)

```
from sklearn.preprocessing import StandardScaler

scaler = StandardScaler()
df[["age", "salary"]] = scaler.fit_transform(df[["age", "salary"]])
```

🧠 Why

Distance-based models need scaled features.

1️⃣0️⃣ Feature Engineering in ML Workflow

```
Raw Data
↓
EDA
↓
Feature Engineering ← 🔥 MOST IMPORTANT
```

↓

Model Training