Dataset: https://www.kaggle.com/competitions/cat-in-the-dat/data

Dataset Understanding & Exploration

1. After loading train.csv, determine:

- The number of rows and columns
- Dataset information using info()
- Basic statistics using describe()

2. How many features are categorical and how many are numerical?

3. Which columns contain missing values?

- Report the number of missing values per column.

4. Classify the categorical features into:

- Binary
- Nominal
- Ordinal

(Identify all **binary categorical columns**(columns with exactly two unique values),

identify **ordinal categorical columns** (ord_1 to ord_5),

identify **nominal columns** (nom_0 to ord_9).)

5. Which features have high cardinality (more than 10 unique values)?

6. Which categorical feature has the highest number of unique categories?

----

7. Identify the **target column**.

- Separate features (X) and target (y).
- Print the first 5 values of y.

8. Apply **Label Encoding** to the following columns: bin_0, bin_1, bin_2, bin_3, bin_4

9. Encode ord_1 using this order: Novice < Contributor < Expert < Master < Grandmaster

- Create a manual mapping dictionary

- Apply the mapping to the column

10. Apply **One-Hot Encoding** to: nom_0, nom_1, nom_2
    - Encode only these columns
    - Drop original columns

// code to unzip the folder:

```
import zipfile

zip_path = "cat-in-the-dat.zip"
extract_path = "cat_in_the_dat"
with zipfile.ZipFile(zip_path, 'r') as zip_ref:
   zip_ref.extractall(extract_path)
print("Unzipping completed.")
```