EDA Guide: Categorical and Numerical Variables

# 1. Encoding vs. EDA: Which Comes First?

- Always perform \*\*EDA before encoding\*\* categorical variables.  
- Encoding too early can mask data quality issues (e.g., missing values, imbalanced categories).  
- Do EDA to understand:  
 • Value counts  
 • Cardinality (number of unique values)  
 • Relationship with target  
- Once you're done analyzing, encode only what you will use for modeling.

# 2. EDA for Categorical Variables

🔍 Questions to Ask:

- How many unique categories exist?  
- Are some categories dominant?  
- Are there typos or inconsistent labels?  
- Does the category distribution change across the target?  
- Are there rare labels that may need grouping?

🛠️ Useful Python Functions:

- `df['col'].value\_counts()`  
- `df['col'].nunique()`  
- `pd.crosstab(df['col'], df['target'])`  
- `df.groupby('col')['target'].mean()`

📊 Useful Visualizations:

- Bar plots (`sns.countplot(x='col', data=df)`)  
- Pie charts (`df['col'].value\_counts().plot.pie()`)  
- Boxplot or Violin plots if numeric target: `sns.boxplot(x='col', y='target')`

# 3. EDA for Numerical Variables

🔍 Questions to Ask:

- What is the distribution like (normal, skewed)?  
- Are there outliers?  
- What are the min/max/median/mean values?  
- Are there missing or zero values?  
- How does the variable correlate with the target?

🛠️ Useful Python Functions:

- `df.describe()`  
- `df['col'].isnull().sum()`  
- `df['col'].skew()`, `df['col'].kurt()`  
- `df.corr()`

📊 Useful Visualizations:

- Histograms (`sns.histplot(df['col'])`)  
- Boxplots (`sns.boxplot(x=df['col'])`)  
- Scatter plots (`sns.scatterplot(x='col1', y='col2')`)  
- KDE plots (`sns.kdeplot(df['col'])`)

# 4. Common Advice and Mistakes to Avoid

- Always \*\*visualize\*\* missing data using `sns.heatmap(df.isnull())` or `missingno.matrix()`  
- Avoid treating ID-like categorical variables as real features (e.g., Zip Codes, IDs)  
- Look out for \*\*data leakage\*\* — variables that contain future info  
- Use `.info()` to detect wrong data types (e.g., numeric as object)  
- Label categorical variables clearly before encoding to avoid confusion