EDA Checklist for AI/ML Projects

Exploratory Data Analysis (EDA) Checklist

1. Basic Data Understanding

- [ ] Load dataset (df.head(), df.info(), df.shape)

- [ ] Check data types (df.dtypes)

- [ ] Check for missing values (df.isnull().sum())

- [ ] Get statistical summary (df.describe())

2. Data Cleaning

- [ ] Handle missing values (drop, impute, or fill)

- [ ] Remove duplicates (df.duplicated().sum())

- [ ] Fix data types (e.g., convert dates to datetime)

3. Univariate Analysis (Single Feature)

- [ ] For numerical columns: Use histogram, boxplot

- [ ] For categorical columns: Use countplot, value_counts()

- [ ] Detect outliers (e.g., using IQR or Z-score)

4. Bivariate/Multivariate Analysis

- [ ] Analyze correlation (df.corr(), heatmap)

- [ ] Scatter plots between features and target

- [ ] Group by categorical features to check averages

5. Target Variable Analysis

- [ ] Check distribution of target

- [ ] Class imbalance check (for classification)

- [ ] Skewness of the target (for regression)

6. Feature Engineering

- [ ] Encode categorical variables (Label/OneHot encoding)

EDA Checklist for AI/ML Projects

- [ ] Create new features (ratios, bins, interactions)

- [ ] Normalize/scale features if needed (StandardScaler, MinMaxScaler)

7. Dimensionality Reduction (optional)

- [ ] Use PCA, t-SNE, or correlation thresholding

8. Data Splitting

- [ ] Split into training and testing sets

- [ ] Optionally create a validation set

9. Automated EDA Tools (Optional)

- [ ] Use ydata-profiling, sweetviz, or dtale for quick reports