**🧹 Data Cleaning and Exploratory Data Analysis (EDA) Summary**

**This script performs a systematic data cleaning and variable analysis process on a dataset (df), likely from a real estate or housing domain. Here's a breakdown of the logic and operations step-by-step:**

**🔍 1. Initial Setup**

* **Libraries imported: numpy, pandas, matplotlib.pyplot, seaborn.**
* **Pandas display settings adjusted to show all columns and rows for deeper inspection.**

**🧼 2. Missing Value Analysis**

* **Checks and prints the total missing values in each column.**
* **A Seaborn heatmap visually highlights missing data across the DataFrame.**
* **Calculates the percentage of missing values for each column.**
* **Identifies columns with more than 17% null values and drops them.**
* **A new heatmap shows the null status after dropping columns.**
* **Remaining rows with null values are also dropped, leaving a fully clean DataFrame (df3\_drop\_rows).**
* **Another heatmap confirms there are no missing values left.**

**🔢 3. Numerical Variable Analysis**

* **Identifies all numeric columns (int64 and float64).**
* **Stores a relevant subset of those in a list num\_var.**
* **For each variable in num\_var, it creates two distribution plots (displot):**
  + **One from the original DataFrame (df).**
  + **One from the cleaned DataFrame (df3\_drop\_rows).**
* **This helps compare how distributions of numeric variables changed due to the cleaning process.**

**🔠 4. Categorical Variable Analysis**

* **Extracts all categorical (object-type) columns.**
* **Stores relevant ones in the list obj\_var.**
* **For the MSZoning column, calculates and compares category percentages in the original vs. cleaned dataset.**
* **Combines and displays these percentages side-by-side for easy comparison.**

**🧰 5. Function for Categorical Comparison**

* **Defines a reusable function cat\_var\_def(var) that:**
  + **Returns a side-by-side comparison of category percentage distribution for any categorical variable.**
  + **Helps identify whether cleaning affected the frequency of each category.**

**🔁 6. Loop Through All Categorical Variables**

* **Loops over all variables in obj\_var list.**
* **For each, the function is called and printed.**
* **This gives a detailed printout of category distribution shifts across all key object-type columns in the dataset.**

**✅ Purpose of the Code**

**This pipeline:**

* **Handles missing data by dropping columns with high null values and removing remaining null rows.**
* **Compares distributions of numeric and categorical variables before and after cleaning.**
* **Ensures no significant data distortion occurs during preprocessing.**
* **Lays a solid foundation for modeling or deeper EDA, especially in structured datasets like housing prices.**