1. Data Loading and Initial Inspection

- **Libraries Imported:** The code begins by importing essential Python libraries for data manipulation and visualization, including **pandas**, **numpy**, **seaborn**, and **matplotlib**.
- **Dataset Loading:** It loads the `roo_data.csv` dataset into a pandas DataFrame.
- **Initial Analysis:** The notebook displays the first few rows, column information (data types and null counts), and a statistical summary of the original dataset to get a preliminary understanding.

2. Simulating a Messy Dataset

To demonstrate data cleaning techniques, the notebook intentionally introduces noise into the data:

- A copy of the original dataset is created.
- **10%** of the cells in this new dataset are randomly selected.
- These selected cells are replaced with either empty strings (`''`) or `NaN` values to simulate missing or improperly recorded data.

3. Data Cleaning and Imputation

The noisy dataset is then systematically cleaned:

- **Standardization:** All empty strings are converted to `NaN` values to ensure missing data is represented consistently.
- **Imputation Strategy:**
  - For **numerical columns**, missing values are filled using the **median** of each respective column.
  - For **categorical columns**, missing values are filled using the **mode** (the most frequently occurring value) of each column.
- **Verification:** After cleaning, the notebook confirms that there are no more missing values in the dataset.

4. Exploratory Data Analysis (EDA)

Using the cleaned dataset, the notebook performs EDA by generating four distinct visualizations to uncover patterns and insights:

1. **Distribution of Logical Quotient Rating:** A **histogram** is created to visualize the frequency distribution of the 'Logical quotient rating'.
2. **Count of Suggested Job Roles:** A horizontal **bar chart** is used to display the count of each 'Suggested Job Role', ordered from most to least frequent.

3. **Correlation of Numerical Features:** A **heatmap** is generated to show the correlation matrix between all the numerical columns, helping to identify relationships between variables.
4. **Outlier Detection in Work Hours:** A **box plot** for 'Hours working per day' is created to analyze its distribution and identify any potential outliers.