Missing Data: Impact of Mean and Median Imputation Technique

Bharat Jambhulkar

Department of Statistics,
Savitribai Phule Pune University

Introduction

This exercise demonstrates how missing values affect the univariate distribution of variables in a dataset. The changes in relationships between variables are examined. The next section introduces the workflow of the exercise. Finally, charts are generated to better visualize the impact of missing value imputation.

Workflow

- Select a dataset with no missing values, ensuring that it contains continuous, categorical, and count variables.
- Decide the proportion of missing values for each variable in the dataset.
- For each column in the dataset, generate a random sample of size equal to the number of observations in the variable from a Bernoulli distribution.
- For each column, replace the observations where the generated Bernoulli random number is 1 with NA.
- On the generated dataset with missing values, apply imputation methods as follows: use mean imputation for continuous data and median imputation for categorical and count data.
- Construct density plots, bar plots, and scatter plots to visualize the impact of imputation on the univariate distribution of variables.

Dataset and Implementation

A dataset named Auto from the ISLR library in R is used for this exercise. The documentation can be found here (link). The dataset contains the following columns: mpg, cylinders, displacement, horsepower, weight, acceleration, year, origin, and name. Among these, cylinders is a count variable, and origin is a categorical variable. The workflow described above is applied to this dataset, and the resulting charts are included in the report. The missing proportions are set as follows: 0.17, 0.14, 0.14, 0.11, 0.19, 0.21, 0.22, and 0.12, corresponding to the sequence of the column names. The "name" column is excluded from the study.

Visualizations

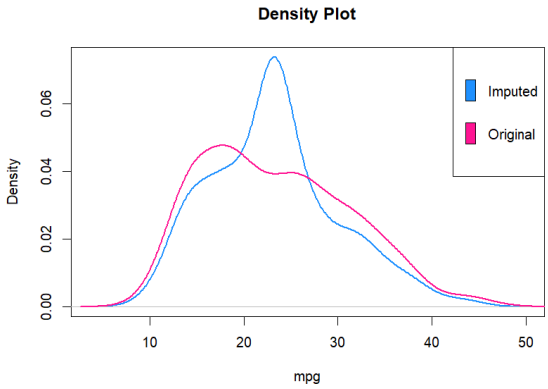

Figure 1: mpg

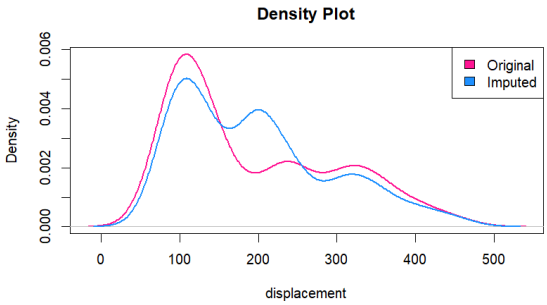

Figure 2: displacement

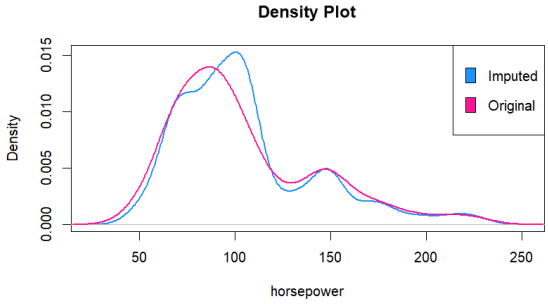

Figure 3: horsepower

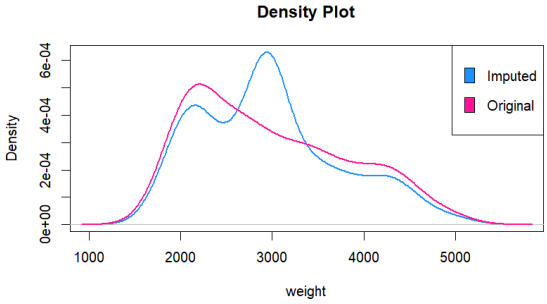

Figure 4: weight

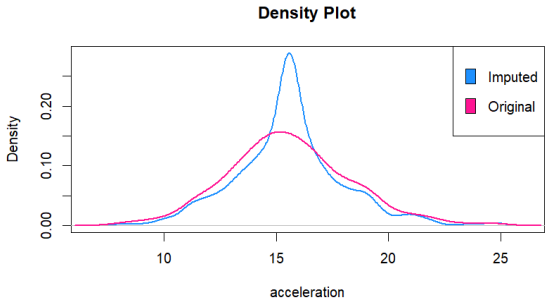

Figure 5: acceleration

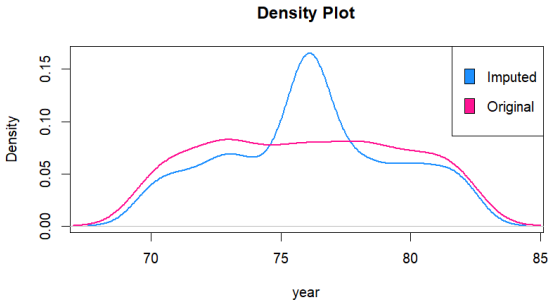

Figure 6: year

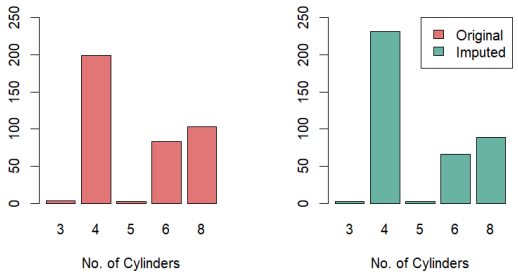

Figure 7: Cylinders

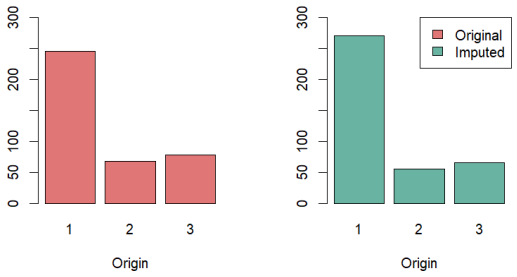

Figure 8: Origin

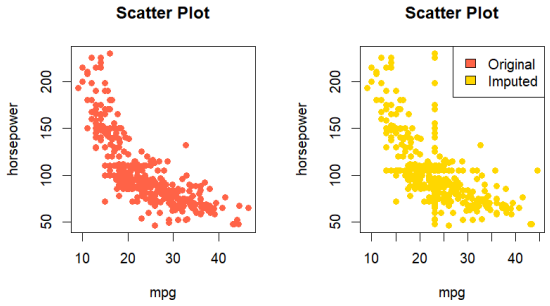

Figure 9: mpg vs horsepower

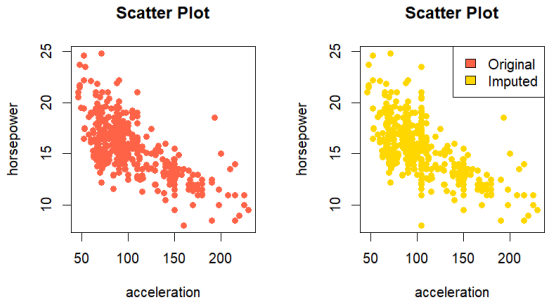

Figure 10: horsepower vs acceleration

Challenges With Mean/Median Imputation

- As evident in the density plot and bar plot, the mean and median values are inflated. This also results in a decrease in the overall dispersion of the data. The univariate distribution of the variable changes due to imputation.
- Moreover, the imputation overlooks the relationship between variables. As visible in Figures 9 and 10, there is a slight change in the scatter plot.

Link to the R Code

Link: github/bharatjambhulkar