Imputation Analysis Report

Dataset Information

Data file used: blood_storage_example
Overall missing rate: 15.54%

Variables used for imputation:

- Median.RBC.Age
- Age
- PVol
- PreopPSA
- Units
- TimeToRecurrence
- RBC.Age.Group
- TVol
- T.Stage
- bGS
- sGS
- AA
- FamHx
- BN+
- OrganConfined
- PreopTherapy
- AnyAdjTherapy
- AdjRadTherapy
- Recurrence

- Censor

Best model selected: RandomForest
Imputed dataset saved as: final_imputed_data_rf.csv

Correlation Heatmaps

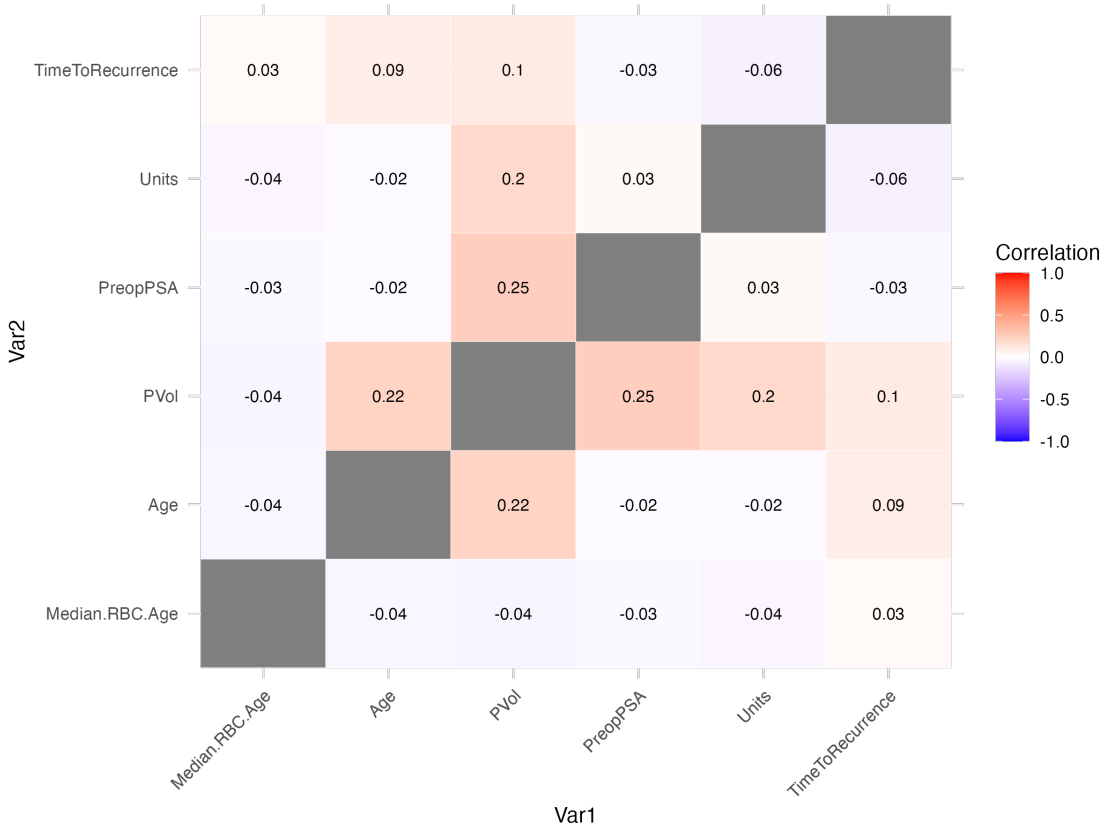

Figure 1: Correlation Heatmap - Continuous Variables

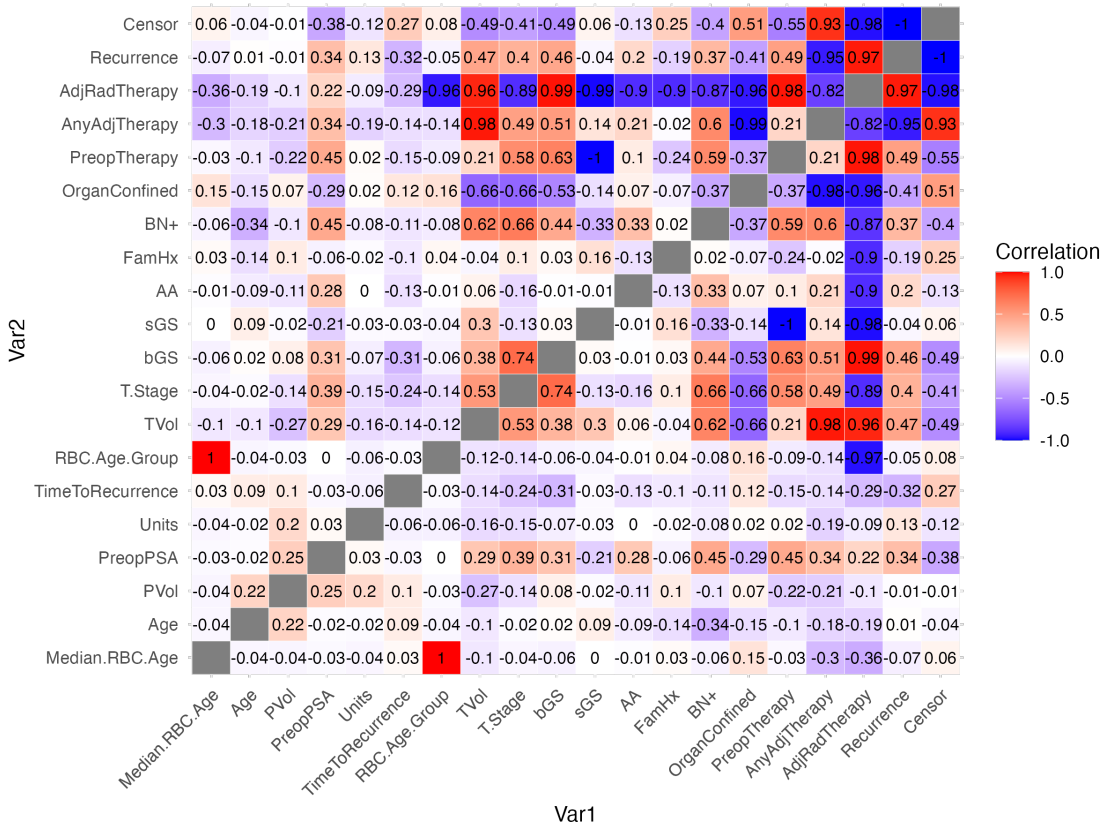

Figure 2: Correlation Heatmap - All Variables

Imputation Settings

Imputation Configuration:
- Missing rate levels (`list_noNA`): 0.05, 0.1, 0.15, 0.2
- Number of Trees (`ntree`): 50
- Maximum Number of Iterations in RandomForest (`maxiter`): 5
- K values for KNN (`k_values`): 3, 5, 7
- MICE iterations (`mice_m`): 3
- MICE maxit (`mice_maxit`): 5
- Number of decision trees in MiceRanger (`micer_num_trees`): 50
- Random seed (`seed`): 123
- Number of iterations (`niter`): 2
- Methods used to simulate imputation (`methods`): rf, mice

Model Performance Summary

Table 1: Model Performance Across Missing Rates

Method	Missing_Rate	Average_MSE	Average_Accuracy	Best_K_Value
MICE	0.05	0.0928285	0.9873810	NA
RandomForest	0.05	0.0251190	0.9935714	NA
MICE	0.10	0.1780190	0.9750000	NA
RandomForest	0.10	0.1133017	0.9845238	NA
MICE	0.15	0.1727021	0.9540476	NA
RandomForest	0.15	0.1167390	0.9726190	NA
MICE	0.20	0.2953958	0.9419048	NA
RandomForest	0.20	0.1684802	0.9642857	NA

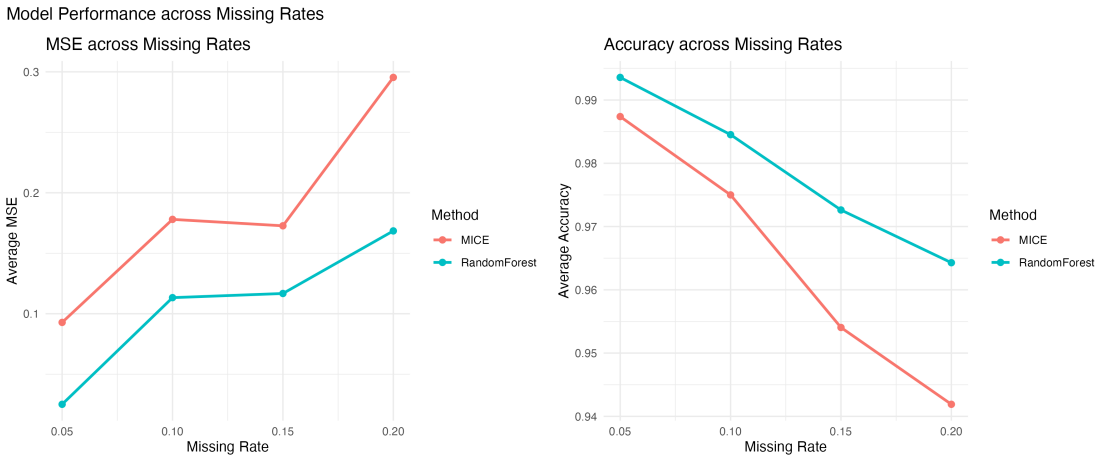

Figure 3: Performance Performance Comparison Plot Across Missing Rates

Imputation Comparison: Before vs After

Table 2: Missing Values Before vs. After Imputation

Variable	Missing_Before	Missing_After
Median.RBC.Age	43	0
Age	49	0
PVol	55	0
PreopPSA	45	0
Units	43	0
TimeToRecurrence	50	0
RBC.Age.Group	49	0
TVol	60	0
T.Stage	54	0
bGS	49	0
sGS	51	0
AA	47	0

Variable	Missing_Before	Missing_After
FamHx	43	0
BN+	50	0
OrganConfined	50	0
PreopTherapy	50	0
AnyAdjTherapy	43	0
AdjRadTherapy	43	0
Recurrence	54	0
Censor	54	0

Table 3: Summary Statistics Before vs. After Imputation

Variable_Statistic	Before	After
Median.RBC.Age_mean	16.648352	17.089385
Median.RBC.Age_sd	6.259769	6.210462
Median.RBC.Age_min	10.000000	10.000000
Median.RBC.Age_max	25.000000	25.000000
Age_mean	61.124345	61.219344
Age_sd	7.208566	6.715862
Age_min	38.400000	38.400000
Age_max	79.000000	79.000000
PVol_mean	56.052874	56.575945
PVol_sd	31.260544	28.786488
PVol_min	19.400000	19.400000
PVol_max	274.000000	274.000000
PreopPSA_mean	8.216052	8.167973
PreopPSA_sd	6.011284	5.659542
PreopPSA_min	1.300000	1.300000
PreopPSA_max	39.000000	39.000000
Units_mean	2.454213	2.446114
Units_sd	1.832787	1.712699
Units_min	1.000000	1.000000
Units_max	19.000000	19.000000
TimeToRecurrence_mean	32.500489	33.122535
TimeToRecurrence_sd	28.111005	26.241702
TimeToRecurrence_min	0.270000	0.270000
TimeToRecurrence_max	103.600000	103.600000

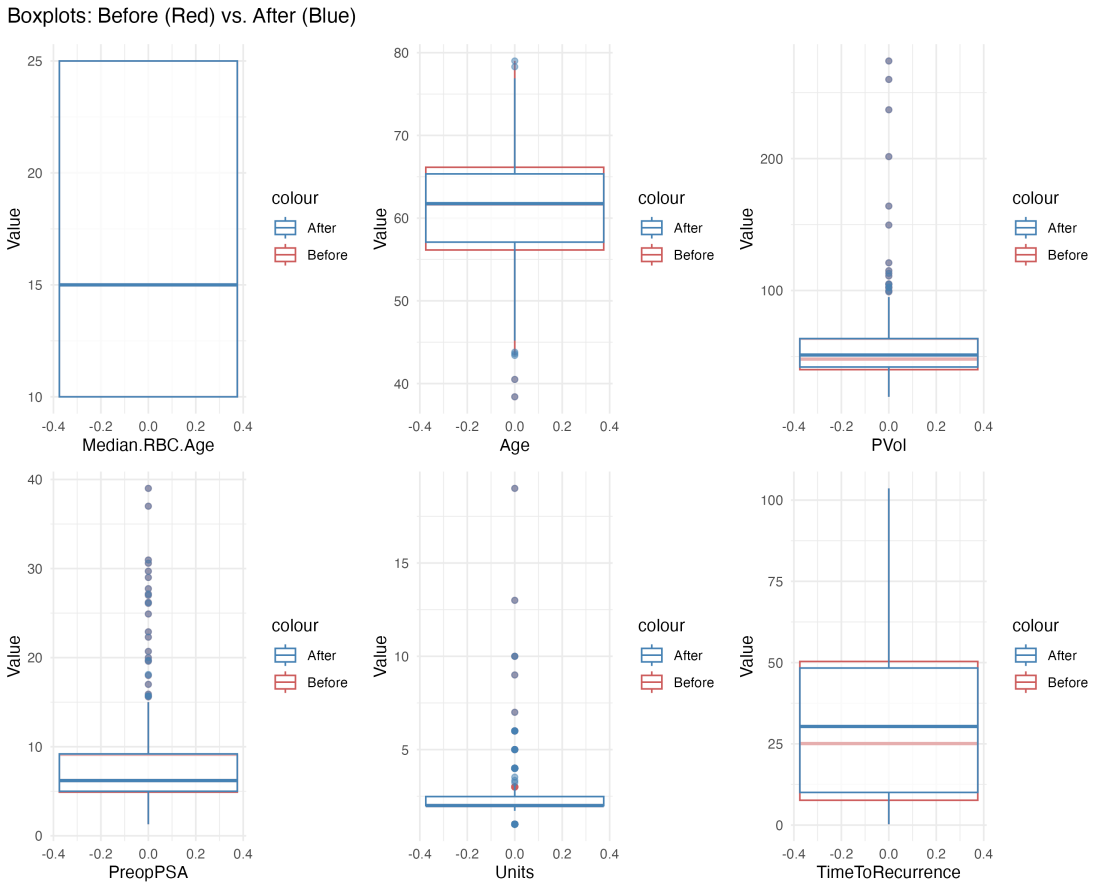

Figure 4: Boxplots (Before vs After)

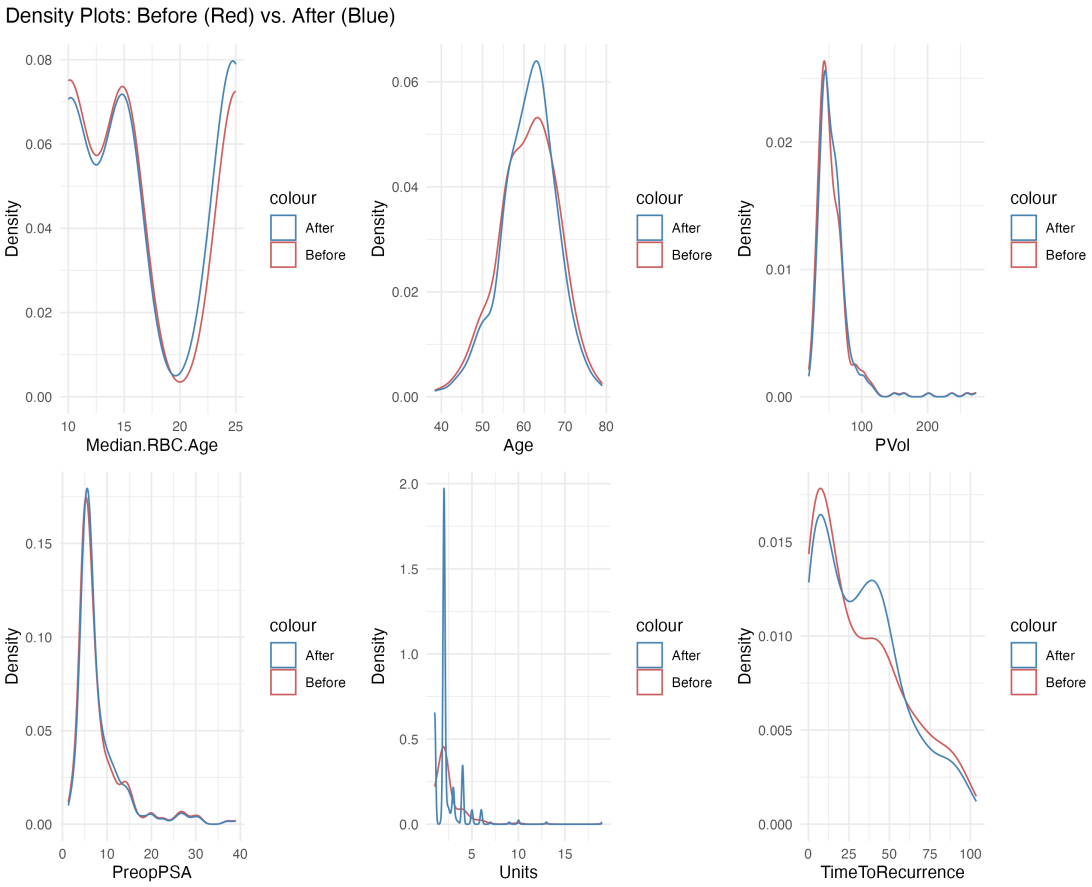

Figure 5: Density Plots (Before vs After)

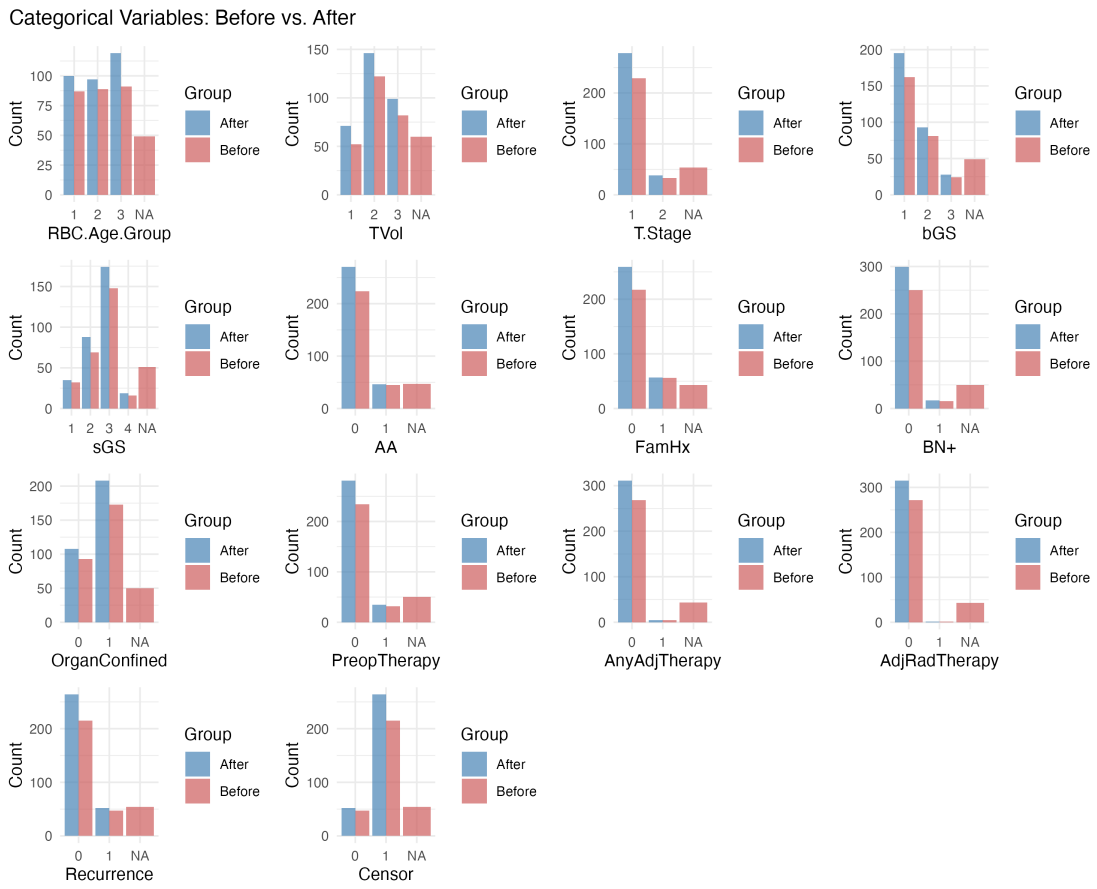

Figure 6: Bar Plots (Before vs After)

Interpretation Guide

- **Correlation Heatmaps**: Red/blue tiles indicate strong positive/negative correlations.
- **Model Performance Table**: Compare methods (Random Forest, KNN, MICE, MiceForest) across missing rates using:
  - **Average MSE**: Lower is better for numeric variables.
  - **Average Accuracy**: Higher is better for categorical variables.
  - **Best K (KNN only)**: The value of `k` yielding best performance.
- **Imputation Comparison**: Boxplots and density plots illustrate that the distribution of imputed values aligns well with original patterns.
- The imputation method with lowest MSE and high accuracy is applied to generate the final complete dataset.