Imputation Analysis Report

Dataset Information

Data file used: blood
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

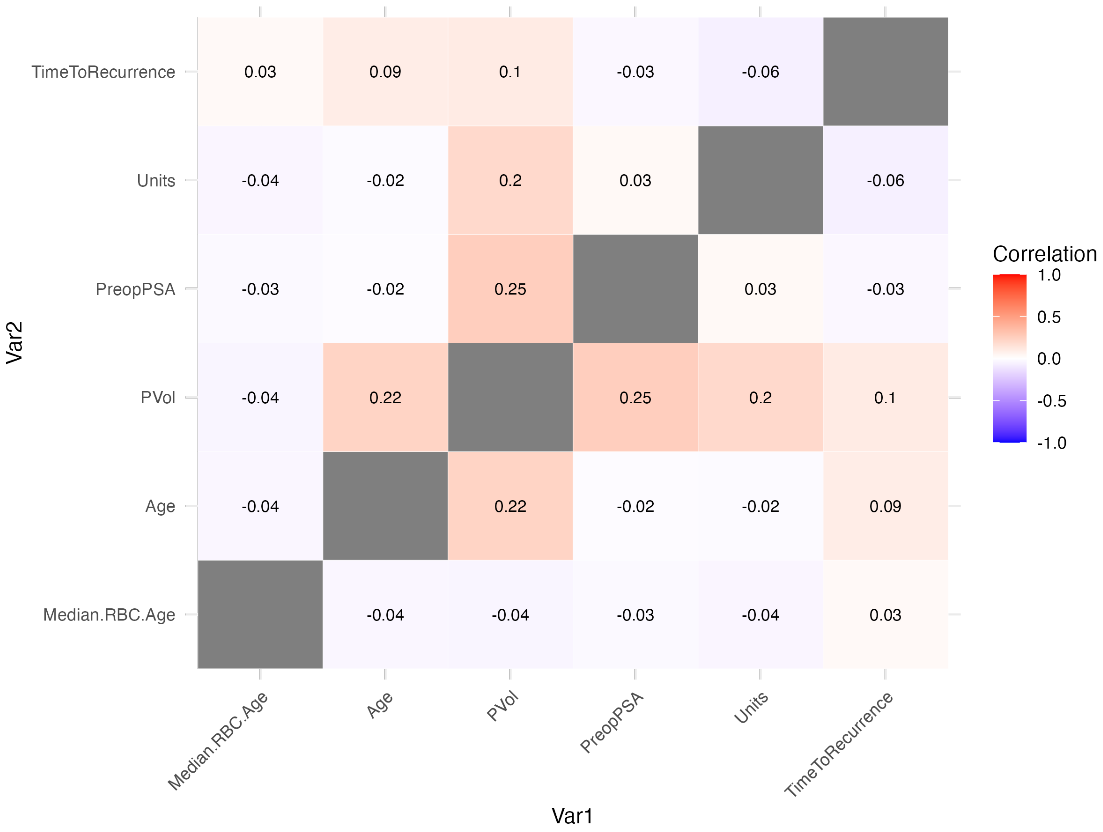

Figure 1: Correlation Heatmap - Continuous Variables

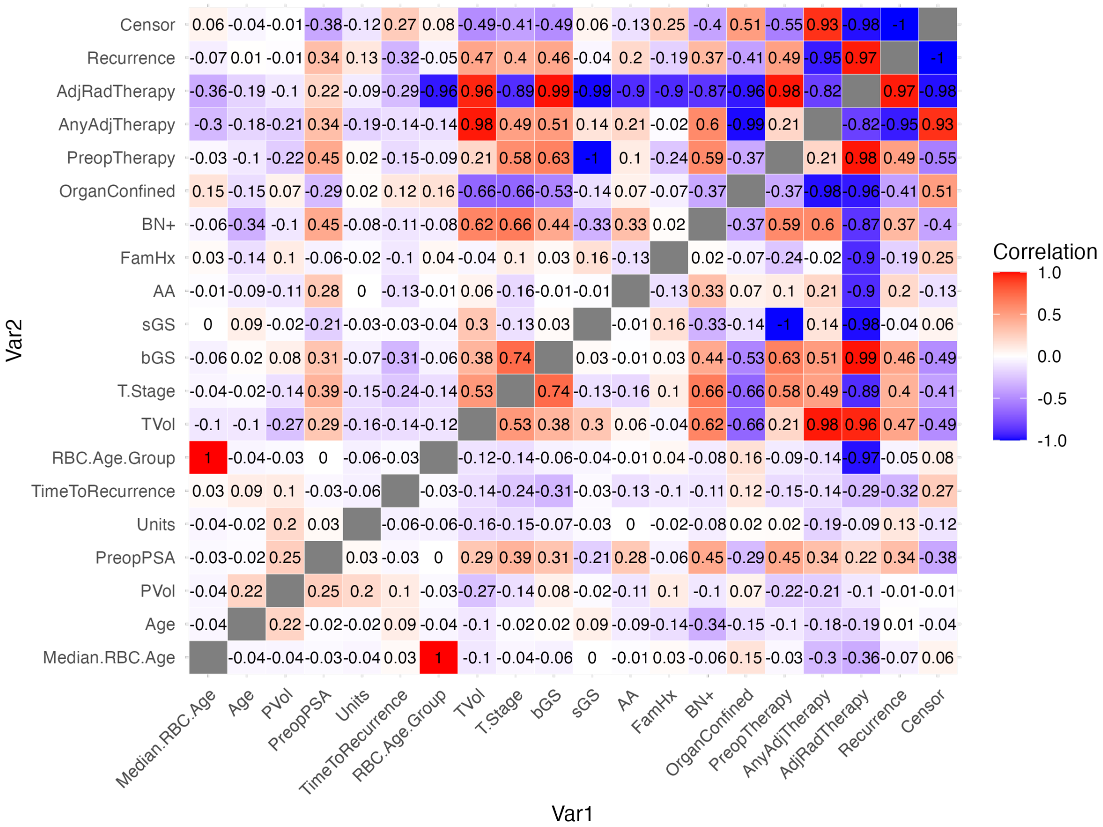

Figure 2: Correlation Heatmap - All Variables

Imputation Settings

Imputation Configuration:
- Missing rate levels (`list_noNA`): 0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4, 0.45, 0.5
- Number of Trees (`ntree`): 100
- Maximum Number of Iterations in RandomForest (`maxiter`): 10
- K values for KNN (`k_values`): 3, 5, 7
- MICE iterations (`mice_m`): 5
- MICE maxit (`mice_maxit`): 10
- Number of decision trees in MiceRanger (`micer_num_trees`): 100
- Random seed (`seed`): 123
- Number of iterations (`niter`): 10
- Methods used to simulate imputation (`methods`): rf, knn, mice, mice_rf

Model Performance Summary

Table 1: Model Performance Across Missing Rates

Method	Missing_Rate	Average_MSE	Average_Accuracy	Best_K_Value
KNN	0.05	0.0450746	0.9699487	7
MICE	0.05	0.0826358	0.9872381	NA
RandomForest	0.05	0.0415254	0.9926190	NA
miceRanger	0.05	0.0558511	0.9909524	NA
KNN	0.10	0.1126354	0.9370769	7
MICE	0.10	0.1708231	0.9722857	NA
RandomForest	0.10	0.1046712	0.9850952	NA
miceRanger	0.10	0.1268556	0.9814762	NA
KNN	0.15	0.2503742	0.8875000	5
MICE	0.15	0.3145225	0.9530000	NA
RandomForest	0.15	0.1852562	0.9735238	NA
miceRanger	0.15	0.2274537	0.9709048	NA
KNN	0.20	0.3293205	0.8610256	3
MICE	0.20	0.3977212	0.9382857	NA
RandomForest	0.20	0.2269373	0.9641429	NA
miceRanger	0.20	0.2808599	0.9587143	NA
MICE	0.25	0.4324824	0.9197619	NA
RandomForest	0.25	0.2476453	0.9542381	NA
miceRanger	0.25	0.3001071	0.9458095	NA
MICE	0.30	0.5376886	0.8925714	NA
RandomForest	0.30	0.2923537	0.9382381	NA
miceRanger	0.30	0.3593772	0.9310952	NA
MICE	0.35	0.6107660	0.8695238	NA
RandomForest	0.35	0.3500510	0.9325714	NA
miceRanger	0.35	0.4172728	0.9180952	NA
MICE	0.40	0.7630913	0.8475238	NA
RandomForest	0.40	0.4156270	0.9158571	NA
miceRanger	0.40	0.4854530	0.9032857	NA
MICE	0.45	0.7825379	0.8156190	NA
RandomForest	0.45	0.4683567	0.8985238	NA
miceRanger	0.45	0.5744656	0.8859524	NA
MICE	0.50	0.9614987	0.7902381	NA
RandomForest	0.50	0.5967739	0.8858571	NA
miceRanger	0.50	0.6857311	0.8720000	NA

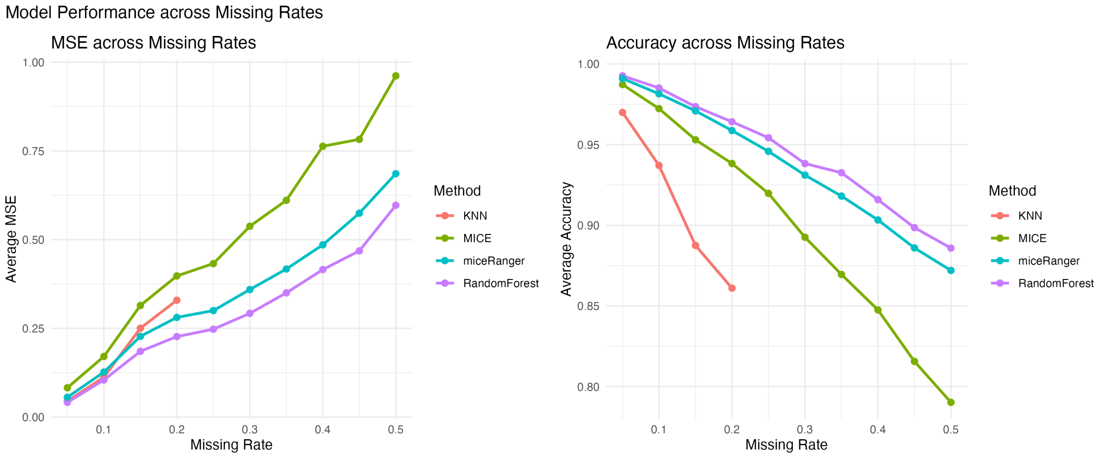

Figure 3: Performance Performance Comparison Plot Across Missing Rates

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
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	17.122665
Median.RBC.Age_sd	6.259769	6.237394
Median.RBC.Age_min	10.000000	10.000000

Median.RBC.Age_max	25.000000	25.000000
Age_mean	61.124345	61.188646
Age_sd	7.208566	6.741040
Age_min	38.400000	38.400000
Age_max	79.000000	79.000000
PVol_mean	56.052874	56.555322
PVol_sd	31.260544	28.800732
PVol_min	19.400000	19.400000
PVol_max	274.000000	274.000000
PreopPSA_mean	8.216052	8.131881
PreopPSA_sd	6.011284	5.640548
PreopPSA_min	1.300000	1.300000
PreopPSA_max	39.000000	39.000000
Units_mean	2.454213	2.462065
Units_sd	1.832787	1.714869
Units_min	1.000000	1.000000
Units_max	19.000000	19.000000
TimeToRecurrence_mean	32.500489	32.938703
TimeToRecurrence_sd	28.111005	26.138715
TimeToRecurrence_min	0.270000	0.270000
TimeToRecurrence_max	103.600000	103.600000

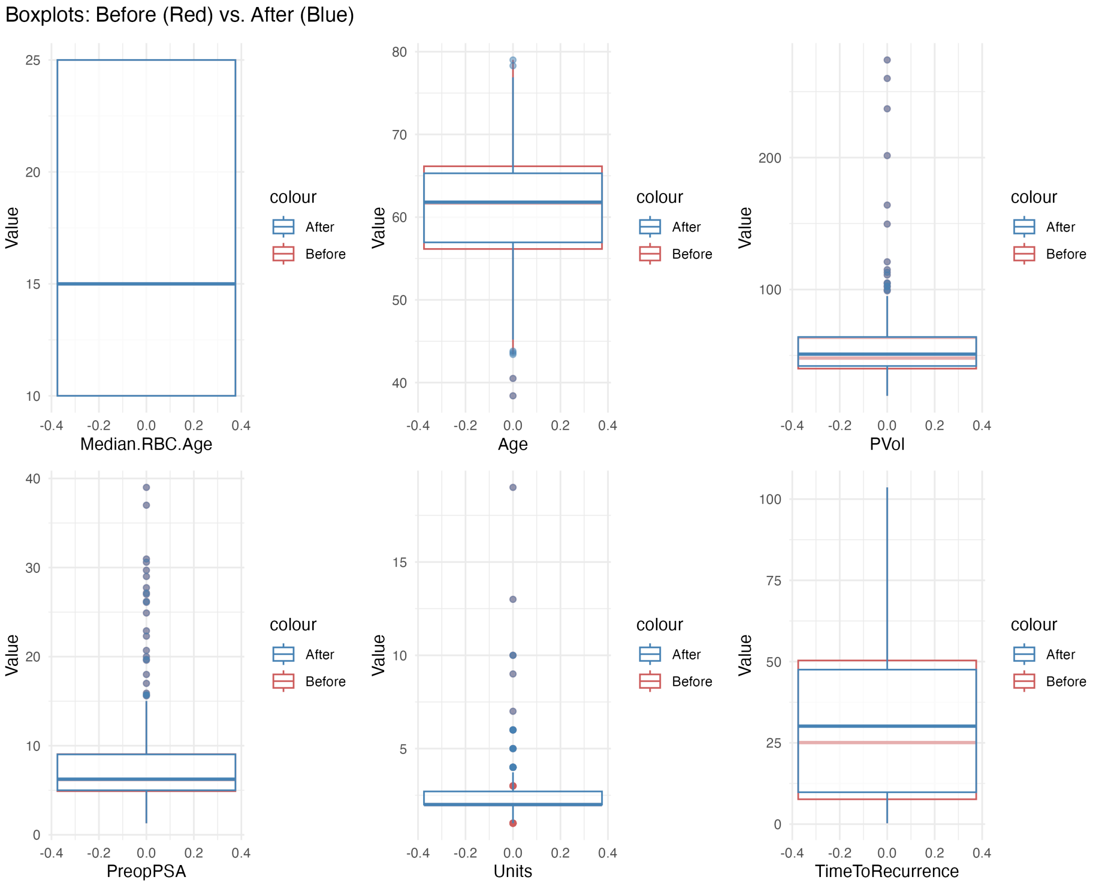

Figure 4: Boxplots (Before vs After)

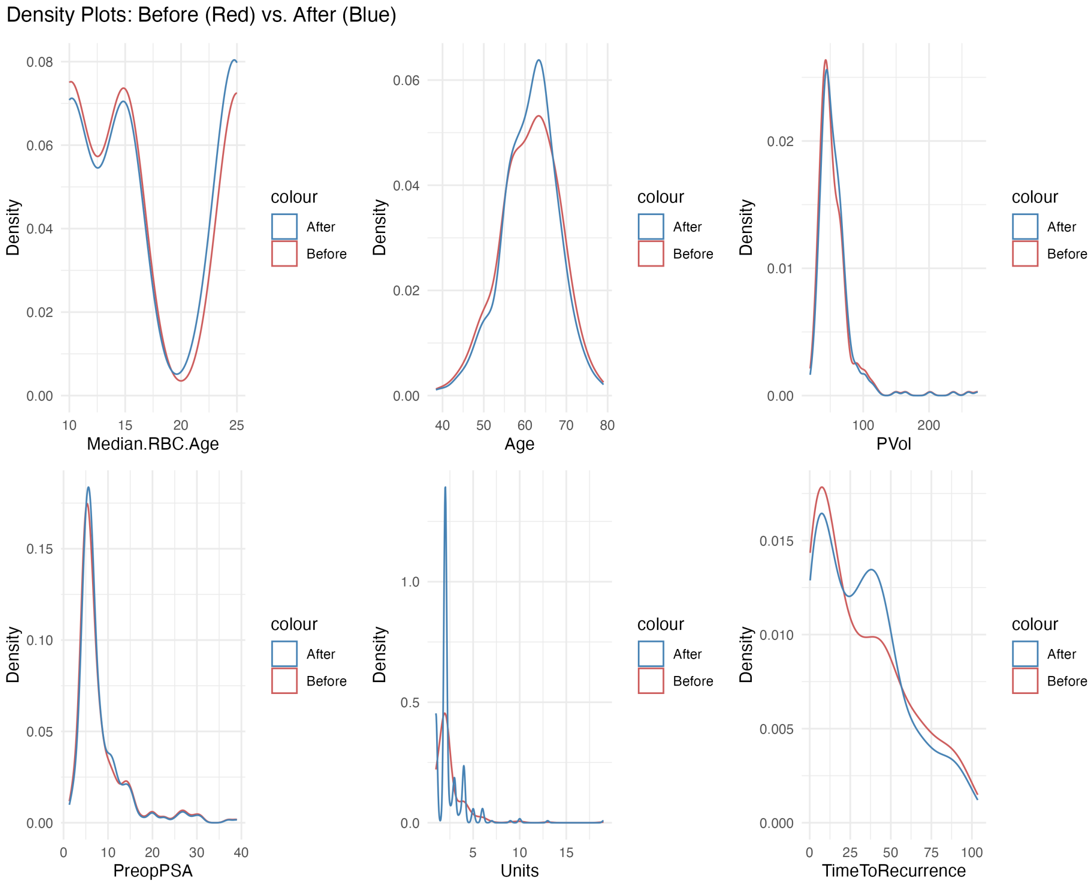

Figure 5: Density Plots (Before vs After)

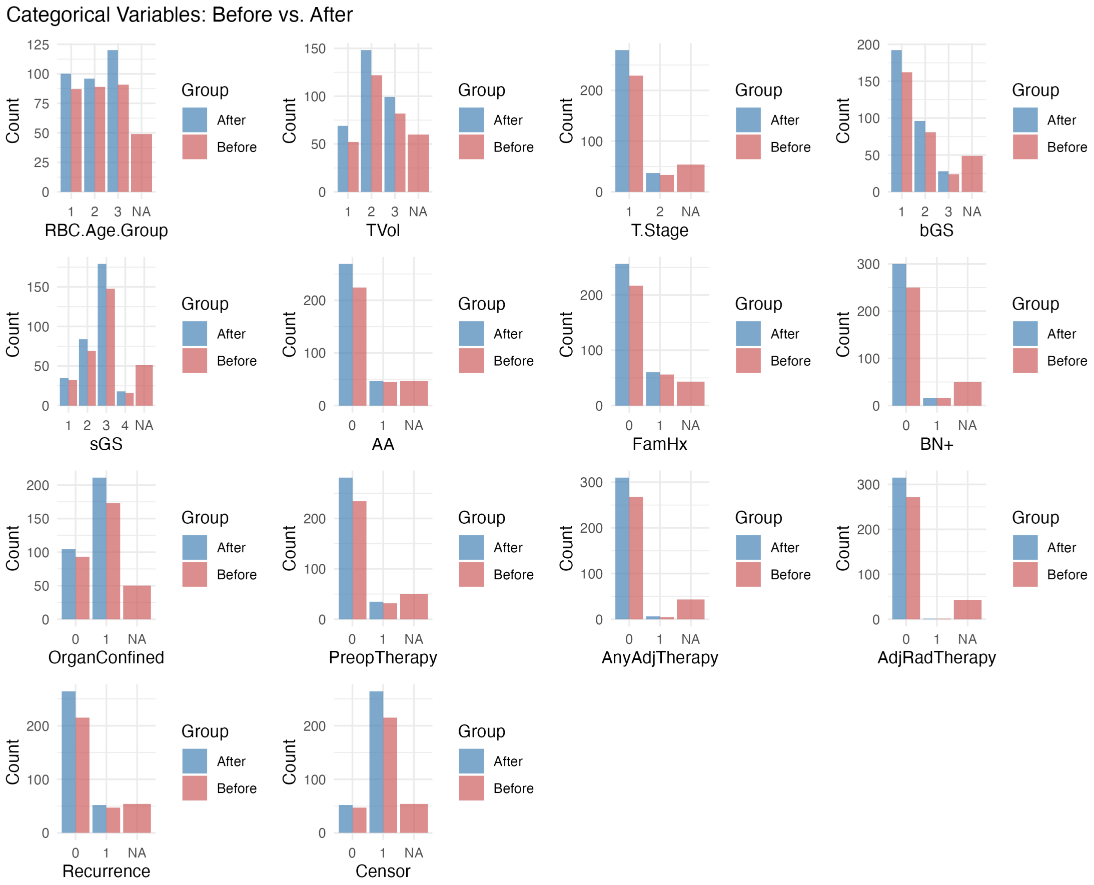

Figure 6: Bar Plots (Before vs After)

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.