Final Project - Data Analysis - Imputation Methods

DORADO CERRATO CHRISTIAN, FORTUNY PARULL LAIA FORTUNY PARULL LAIA, KAUR N

sys.date()

Objective

Based on the information extracted from nhs.uk/conditions/kidney-disease the **Chronic Kidney Disease** (CKD) is a long-term condition where damaged kidneys can't effectively filter waste and extra fluid from the blood. It leading to buildup and potential health problems like heart disease, anemia, and high blood pressure, with diabetes and hypertension being leading causes. Often showing few symptoms until advanced stages, requiring lifestyle changes, medications, or, in severe cases, dialysis or transplant to manage its progression.

There are usually no **symptoms of CKD** in the early stages. It may only be diagnosed if you have a blood or urine test for another reason and the results show a possible problem with your kidneys. At a more advanced stage, symptoms can include:

- tiredness
- swollen ankles, feet or hands
- shortness of breath
- feeling sick
- blood in your pee (urine)

The Chronic kidney disease is **usually caused by other conditions** that put a strain on the kidneys. Often it's the result of a combination of different problems. CKD can be caused by:

- high blood pressure – over time, this can put strain on the small blood vessels in the kidneys and stop the kidneys working properly
- diabetes – too much glucose in your blood can damage the tiny filters in the kidneys
- high cholesterol – this can cause a build-up of fatty deposits in the blood vessels supplying your kidneys, which can make it harder for them to work properly
- kidney infections
- glomerulonephritis – kidney inflammation
- autosomal dominant polycystic kidney disease – an inherited condition where growths called cysts develop in the kidneys
- blockages in the flow of urine – for example, from kidney stones that keep coming back, or an enlarged prostate
- long-term, regular use of certain medicines – such as lithium and non-steroidal anti-inflammatory drugs (NSAIDs)

CKD **can be diagnosed** using blood and urine tests. These tests look for high levels of certain substances in the blood and urine that are signs your kidneys are not working properly.

If the person at a high risk of developing kidney disease (for example, it has a known risk factor such as high blood pressure or diabetes), he may be advised to have regular tests to check for CKD so it's found at an early stage.

The results of the blood and urine tests can be used to tell the stage of your kidney disease. This is a number that reflects how severe the damage to your kidneys is, with a higher number indicating more serious CKD.

There's **no cure for CKD**, but treatment can help relieve the symptoms and stop it getting worse. The treatment will depend on how severe the condition is. The main treatments are:

- lifestyle changes to help you remain as healthy as possible
- medicine to control associated problems such as high blood pressure and high cholesterol
- medicine that can help the kidneys keep working for longer
- dialysis – treatment to replicate some of the kidney's functions (this may be necessary in advanced CKD)
- kidney transplant – this may also be necessary in advanced CKD

CKD can range from a mild condition with no or few symptoms, to a very serious condition where the kidneys stop working, sometimes called kidney failure.

Most people with CKD will be able to control their condition with medicine and regular check-ups. CKD only progresses to kidney failure in around 2 in 100 people with the condition.

If someone has CKD, even if it's mild, he's at an increased risk of developing other serious problems, such as cardiovascular disease. This is a group of conditions affecting the heart and blood vessels, which includes heart attack and stroke.

Cardiovascular disease is one of the main causes of death in people with kidney disease, although healthy lifestyle changes and medicine can help reduce your risk of developing it.

---

1. Some research questions

1. **How do blood markers (creatinine, urea, hemoglobin) correlate with kidney disease progression?**
2. **Which combination of features best predicts CKD status?**
3. **Are there distinct patient clusters based on their biochemical profiles?**
4. **How do demographic factors (age, hypertension, diabetes) interact with biochemical markers?**

---

2. Dataset loading

The dataset has been downloaded from PubMed Central and this study has been replicated to study improvements[ˆ1].

[ˆ1]V. Kumar et al., "The Indian Chronic Kidney Disease (ICKD) study: baseline characteristics," Clin Kidney J, vol. 15, no. 1, pp. 60–69, Jan. 2022, doi: 10.1093/CKJ/SFAB149.

```r
library(readr)
setwd('/Volumes/HHD_iMac_Storage/URV/SCIENTIFIC_PROGRAMMING/FINAL/SP-Final-Project')
ckd_data <- read_csv("data/raw/chronic_kindey_disease.csv")
```

```r
summary(ckd_data)
```

```
##      age                 bp                 sg                 al
##  Length:400         Length:400         Length:400         Length:400
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##       su                rbc                 pc                pcc
##  Length:400         Length:400         Length:400         Length:400
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##       ba                bgr                 bu                 sc
##  Length:400         Length:400         Length:400         Length:400
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##      sod                pot                hemo               pcv
##  Length:400         Length:400         Length:400         Length:400
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##      wbcc               rbcc               htn                 dm
##  Length:400         Length:400         Length:400         Length:400
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##      cad               appet                pe                ane
##  Length:400         Length:400         Length:400         Length:400
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##     status
##  Length:400
##  Class :character
##  Mode  :character
```

```r
head(ckd_data)
```

```
## # A tibble: 6 x 25
##   age   bp    sg    al    su    rbc    pc    pcc   ba    bgr   bu    sc    sod
##   <chr> <chr> <chr> <chr> <chr> <chr>  <chr> <chr> <chr> <chr> <chr> <chr> <chr>
## 1 48.0  80.0  1.02  1.0   0.0   ?      norm~ notp~ notp~ 121.0 36.0  1.2   ?
## 2 7.0   50.0  1.02  4.0   0.0   ?      norm~ notp~ notp~ ?     18.0  0.8   ?
## 3 62.0  80.0  1.01  2.0   3.0   normal norm~ notp~ notp~ 423.0 53.0  1.8   ?
## 4 48.0  70.0  1.005 4.0   0.0   normal abno~ pres~ notp~ 117.0 56.0  3.8   111.0
## 5 51.0  80.0  1.01  2.0   0.0   normal norm~ notp~ notp~ 106.0 26.0  1.4   ?
## 6 60.0  90.0  1.015 3.0   0.0   ?      ?     notp~ notp~ 74.0  25.0  1.1   142.0
## # i 12 more variables: pot <chr>, hemo <chr>, pcv <chr>, wbcc <chr>,
## #   rbcc <chr>, htn <chr>, dm <chr>, cad <chr>, appet <chr>, pe <chr>,
## #   ane <chr>, status <chr>
```

Attribute Information

2.1 Data Set Information:

We use the following representation to collect the dataset:-

- age - age
- bp - blood pressure
- sg - specific gravity
- al - albumin
- su - sugar
- rbc - red blood cells
- pc - pus cell
- pcc - pus cell clumps
- ba - bacteria
- bgr - blood glucose random
- bu - blood urea
- sc - serum creatinine
- sod - sodium
- pot - potassium
- hemo - hemoglobin
- pcv - packed cell volume
- wc - white blood cell count
- rc - red blood cell count
- htn - hypertension
- dm - diabetes mellitus
- cad - coronary artery disease
- appet - appetite
- pe - pedal edema
- ane - anemia
- class - class

Additional Feature Details

2.2 Attribute Information:

We use 24 + class = 25 (11 numeric ,14 nominal)

```
Age(numerical) - age in years
Blood Pressure(numerical) - bp in mm/Hg
Specific Gravity(nominal) - sg - (1.005,1.010,1.015,1.020,1.025)
Albumin(nominal) - al - (0,1,2,3,4,5)
Sugar(nominal) - su - (0,1,2,3,4,5)
Red Blood Cells(nominal) - rbc - (normal,abnormal)
Pus Cell (nominal) - pc - (normal,abnormal)
Pus Cell clumps(nominal) - pcc - (present,notpresent)
Bacteria(nominal) - ba - (present,notpresent)
Blood Glucose Random(numerical) - bgr in mgs/dl
Blood Urea(numerical) -bu in mgs/dl
Serum Creatinine(numerical) - sc in mgs/dl
Sodium(numerical) - sod in mEq/L
Potassium(numerical) - pot in mEq/L
Hemoglobin(numerical) - hemo in gms
Packed Cell Volume(numerical)
```

```
White Blood Cell Count(numerical) - wc in cells/cumm
Red Blood Cell Count(numerical) - rc in millions/cmm
Hypertension(nominal) - htn - (yes,no)
Diabetes Mellitus(nominal) - dm - (yes,no)
Coronary Artery Disease(nominal) - cad - (yes,no)
Appetite(nominal) - appet - (good,poor)
Pedal Edema(nominal) - pe - (yes,no)
Anemia(nominal) - ane - (yes,no)
Class (nominal)- class - (ckd,notckd)
```

Acknowledgements

https://archive.ics.uci.edu/ml/datasets/Chronic_Kidney_Disease

3. Setup and Data Loading

```r
# Load required libraries
if (!require("tidyverse")) install.packages("tidyverse")
if (!require("corrplot")) install.packages("corrplot")
if (!require("ggplot2")) install.packages("ggplot2")
if (!require("gridExtra")) install.packages("gridExtra")
if (!require("caret")) install.packages("caret")
if (!require("mice")) install.packages("mice")
if (!require("dplyr")) install.packages("dplyr")
if (!require("tidyr")) install.packages("tidyr")
if (!require("lattice")) install.packages("lattice")
if (!require("GGally")) install.packages("GGally")
if (!require("VIM")) install.packages("VIM")
if (!require("psych")) install.packages("psych")
if (!require("car")) install.packages("car")
if (!require("factoextra")) install.packages("factoextra")
if (!require("pROC")) install.packages("pROC")
if (!require("ggpubr")) install.packages("ggpubr")
if (!require("rstatix")) install.packages("rstatix")
if (!require("missForest")) install.packages("missForest")

library("missForest")
library(tidyverse)
library(corrplot)
library(ggplot2)
library(gridExtra)
library(caret)
library(mice)
library(dplyr)
library(tidyr)
library(lattice)

library(GGally)
library(VIM)
library(psych)
library(lmtest)
library(car)
library(factoextra)
library(pROC)
library(ggpubr)
library(rstatix)
```

```r
# Set seed for reproducibility
set.seed(17)

# Read data
setwd('/Volumes/HHD_iMac_Storage/URV/SCIENTIFIC_PROGRAMMING/FINAL/SP-Final-Project')
ckd_data <- read.table("data/raw/chronic_kindey_disease.csv",
                       sep = ",",
                       header = TRUE,
                       na.strings = c("?", "NA", "", "NAN", "-"),
                       stringsAsFactors = FALSE)

# Set column names based on dataset description
col_names <- c("age", "bp", "sg", "al", "su", "rbc", "pc", "pcc", "ba", "bgr",
               "bu", "sc", "sod", "pot", "hemo", "pcv", "wbcc", "rbcc",
               "htn", "dm", "cad", "appet", "pe", "ane", "status")

names(ckd_data) <- col_names

# Check structure
str(ckd_data)
```

```
## 'data.frame':    400 obs. of  25 variables:
##  $ age   : num  48 7 62 48 51 60 68 24 52 53 ...
##  $ bp    : num  80 50 80 70 80 90 70 NA 100 90 ...
##  $ sg    : num  1.02 1.02 1.01 1 1.01 ...
##  $ al    : num  1 4 2 4 2 3 0 2 3 2 ...
##  $ su    : num  0 0 3 0 0 0 0 4 0 0 ...
##  $ rbc   : chr  NA NA "normal" "normal" ...
##  $ pc    : chr  "normal" "normal" "normal" "abnormal" ...
##  $ pcc   : chr  "notpresent" "notpresent" "notpresent" "present" ...
##  $ ba    : chr  "notpresent" "notpresent" "notpresent" "notpresent" ...
##  $ bgr   : num  121 NA 423 117 106 74 100 410 138 70 ...
##  $ bu    : num  36 18 53 56 26 25 54 31 60 107 ...
##  $ sc    : num  1.2 0.8 1.8 3.8 1.4 1.1 24 1.1 1.9 7.2 ...
##  $ sod   : num  NA NA NA 111 NA 142 104 NA NA 114 ...
##  $ pot   : num  NA NA NA 2.5 NA 3.2 4 NA NA 3.7 ...
##  $ hemo  : num  15.4 11.3 9.6 11.2 11.6 12.2 12.4 12.4 10.8 9.5 ...
##  $ pcv   : num  44 38 31 32 35 39 36 44 33 29 ...
##  $ wbcc  : num  7800 6000 7500 6700 7300 7800 NA 6900 9600 12100 ...
##  $ rbcc  : num  5.2 NA NA 3.9 4.6 4.4 NA 5 4 3.7 ...
##  $ htn   : chr  "yes" "no" "no" "yes" ...
##  $ dm    : chr  "yes" "no" "yes" "no" ...
##  $ cad   : chr  "no" "no" "no" "no" ...
##  $ appet : chr  "good" "good" "poor" "poor" ...
##  $ pe    : chr  "no" "no" "no" "yes" ...
##  $ ane   : chr  "no" "no" "yes" "yes" ...
##  $ status: chr  "ckd" "ckd" "ckd" "ckd" ...
```

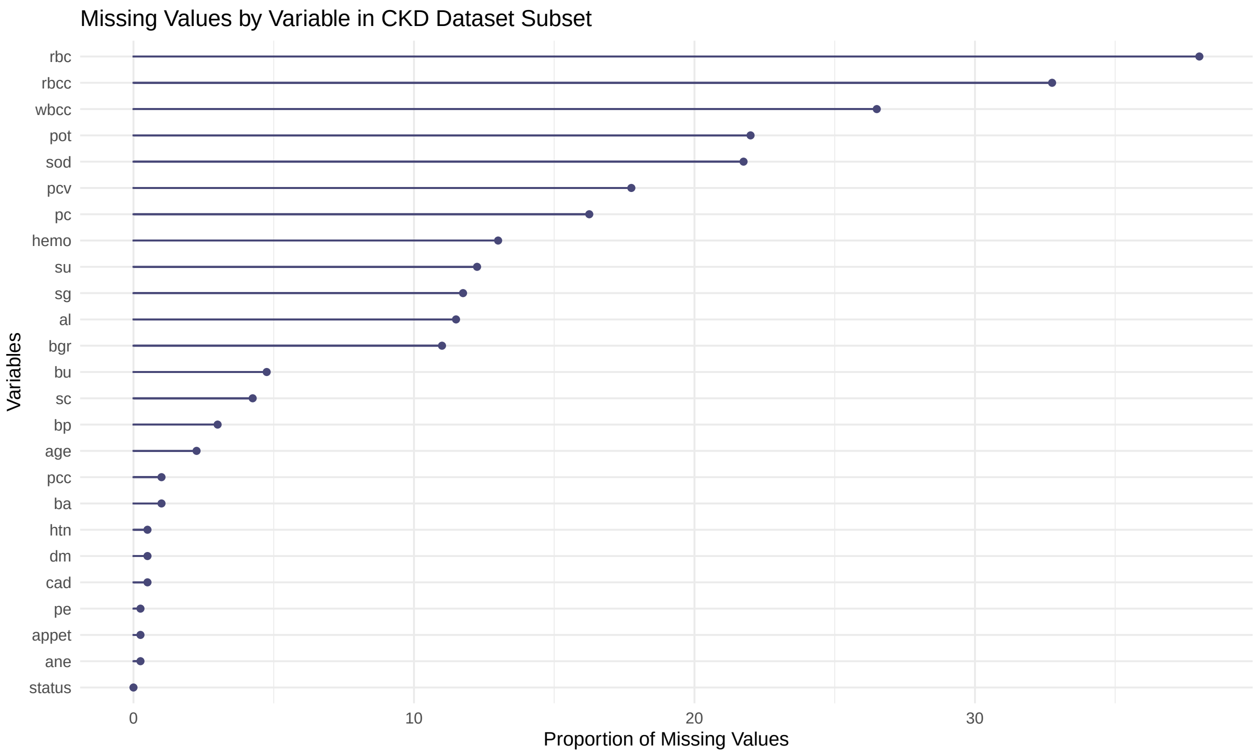
---

4. Data cleaning

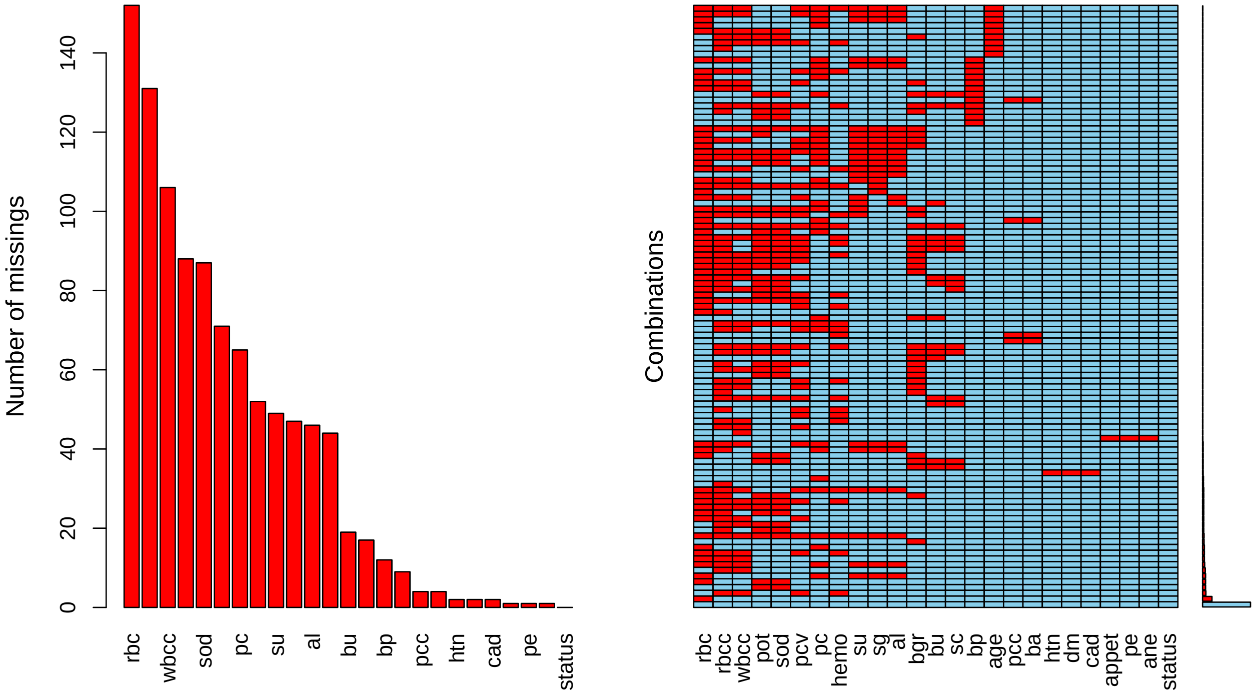
```
ckd_data %>%
  summarise(across(everything(), ~ sum(is.na(.))))
```

```
##   age bp sg al su rbc pc pcc ba bgr bu sc sod pot hemo pcv wbcc rbcc htn dm cad
## 1   9 12 47 46 49 152 65   4  4  44 19 17  87  88   52  71  106  131   2  2   2
##   appet pe ane status
## 1     1  1   1      0
```

```
library(naniar)
library(ggplot2)
# Visualize missing values by variable
gg_miss_var(ckd_data, show_pct = TRUE) +
  labs(title = "Missing Values by Variable in CKD Dataset Subset",
       x = "Variables",
       y = "Proportion of Missing Values")
```



```
library(VIM)

aggr(ckd_data, numbers = TRUE, prop = FALSE, sortVar = TRUE)
```



```
##
##  Variables sorted by number of missings:
##  Variable Count
##       rbc   152
##      rbcc   131
##      wbcc   106
##       pot    88
##       sod    87
##       pcv    71
##        pc    65
##      hemo    52
##        su    49
##        sg    47
##        al    46
##       bgr    44
##        bu    19
##        sc    17
##        bp    12
##       age     9
##       pcc     4
##        ba     4
##       htn     2
##        dm     2
##       cad     2
##     appet     1
##        pe     1
##       ane     1
##    status     0
```

```r
# Display first few rows with original character format
head(ckd_data)
```

```
##   age bp    sg al su    rbc       pc        pcc         ba bgr bu  sc sod pot
## 1  48 80 1.020  1  0   <NA>   normal notpresent notpresent 121 36 1.2  NA  NA
## 2   7 50 1.020  4  0   <NA>   normal notpresent notpresent  NA 18 0.8  NA  NA
## 3  62 80 1.010  2  3 normal   normal notpresent notpresent 423 53 1.8  NA  NA
## 4  48 70 1.005  4  0 normal abnormal    present notpresent 117 56 3.8 111 2.5
## 5  51 80 1.010  2  0 normal   normal notpresent notpresent 106 26 1.4  NA  NA
## 6  60 90 1.015  3  0   <NA>     <NA> notpresent notpresent  74 25 1.1 142 3.2
##   hemo pcv wbcc rbcc htn  dm cad appet  pe ane status
## 1 15.4  44 7800  5.2 yes yes  no  good  no  no    ckd
## 2 11.3  38 6000   NA  no  no  no  good  no  no    ckd
## 3  9.6  31 7500   NA  no yes  no  poor  no yes    ckd
## 4 11.2  32 6700  3.9 yes  no  no  poor yes yes    ckd
## 5 11.6  35 7300  4.6  no  no  no  good  no  no    ckd
## 6 12.2  39 7800  4.4 yes yes  no  good yes  no    ckd
```

```r
# Function to clean and convert numeric columns
clean_numeric <- function(x) {
  # Remove any non-numeric characters except decimal points and minus signs
  x_clean <- gsub("[^0-9.-]", "", x)
  # Convert to numeric
  as.numeric(x_clean)
}

# Function to clean factor columns
clean_factor <- function(x, levels = NULL) {
  # Trim whitespace
  x_clean <- trimws(x)
  # Convert to factor
  if(is.null(levels)) {
    factor(x_clean)
  } else {
    factor(x_clean, levels = levels)
  }
}

# Identify numeric and factor columns based on dataset description
numeric_cols <- c("age", "bp", "bgr", "bu", "sc", "sod", "pot",
                  "hemo", "pcv", "wbcc", "rbcc", "sg", "al", "su")

factor_cols <- c("sg", "al", "su", "rbc", "pc", "pcc", "ba",
                 "htn", "dm", "cad", "appet", "pe", "ane", "status")

# Apply cleaning
ckd_clean <- ckd_data

# Clean numeric columns
for(col in numeric_cols) {
  ckd_clean[[col]] <- clean_numeric(ckd_clean[[col]])
}
```

```
# Clean factor columns with appropriate levels
factor_levels <- list(
  sg = c(1.005, 1.010, 1.015, 1.020, 1.025),
  al = c(0.0, 1.0, 2.0, 3.0, 4.0, 5.0),
  su = c(0.0, 1.0, 2.0, 3.0, 4.0, 5.0),
  rbc = c("normal", "abnormal"),
  pc = c("normal", "abnormal"),
  pcc = c("present", "notpresent"),
  ba = c("present", "notpresent"),
  htn = c("yes", "no"),
  dm = c("yes", "no"),
  cad = c("yes", "no"),
  appet = c("good", "poor"),
  pe = c("yes", "no"),
  ane = c("yes", "no"),
  status = c("ckd", "notckd")
)

for(col in factor_cols) {
  ckd_clean[[col]] <- clean_factor(ckd_clean[[col]], factor_levels[[col]])
}

# Check the cleaned data structure
str(ckd_clean)
```

```
## ’data.frame':    400 obs. of  25 variables:
##  $ age   : num  48 7 62 48 51 60 68 24 52 53 ...
##  $ bp    : num  80 50 80 70 80 90 70 NA 100 90 ...
##  $ sg    : Factor w/ 5 levels "1.005","1.01",..: 4 4 2 1 2 3 2 3 3 4 ...
##  $ al    : Factor w/ 6 levels "0","1","2","3",..: 2 5 3 5 3 4 1 3 4 3 ...
##  $ su    : Factor w/ 6 levels "0","1","2","3",..: 1 1 4 1 1 1 1 5 1 1 ...
##  $ rbc   : Factor w/ 2 levels "normal","abnormal": NA NA 1 1 1 NA NA 1 1 2 ...
##  $ pc    : Factor w/ 2 levels "normal","abnormal": 1 1 1 2 1 NA 1 2 2 2 ...
##  $ pcc   : Factor w/ 2 levels "present","notpresent": 2 2 2 1 2 2 2 2 1 1 ...
##  $ ba    : Factor w/ 2 levels "present","notpresent": 2 2 2 2 2 2 2 2 2 2 ...
##  $ bgr   : num  121 NA 423 117 106 74 100 410 138 70 ...
##  $ bu    : num  36 18 53 56 26 25 54 31 60 107 ...
##  $ sc    : num  1.2 0.8 1.8 3.8 1.4 1.1 24 1.1 1.9 7.2 ...
##  $ sod   : num  NA NA NA 111 NA 142 104 NA NA 114 ...
##  $ pot   : num  NA NA NA 2.5 NA 3.2 4 NA NA 3.7 ...
##  $ hemo  : num  15.4 11.3 9.6 11.2 11.6 12.2 12.4 12.4 10.8 9.5 ...
##  $ pcv   : num  44 38 31 32 35 39 36 44 33 29 ...
##  $ wbcc  : num  7800 6000 7500 6700 7300 7800 NA 6900 9600 12100 ...
##  $ rbcc  : num  5.2 NA NA 3.9 4.6 4.4 NA 5 4 3.7 ...
##  $ htn   : Factor w/ 2 levels "yes","no": 1 2 2 1 2 1 2 2 1 1 ...
##  $ dm    : Factor w/ 2 levels "yes","no": 1 2 1 2 2 1 2 1 1 1 ...
##  $ cad   : Factor w/ 2 levels "yes","no": 2 2 2 2 2 2 2 2 2 2 ...
##  $ appet : Factor w/ 2 levels "good","poor": 1 1 2 2 1 1 1 1 1 2 ...
##  $ pe    : Factor w/ 2 levels "yes","no": 2 2 2 1 2 1 2 1 2 2 ...
##  $ ane   : Factor w/ 2 levels "yes","no": 2 2 1 1 2 2 2 2 1 1 ...
##  $ status: Factor w/ 2 levels "ckd","notckd": 1 1 1 1 1 1 1 1 1 1 ...
```

```r
# Summary statistics
summary(ckd_clean)
```

```
##       age              bp             sg         al         su     
##  Min.   : 2.00   Min.   : 50.00   1.005:  7   0   :199   0   :290  
##  1st Qu.:42.00   1st Qu.: 70.00   1.01 : 84   1   : 44   1   : 13  
##  Median :55.00   Median : 80.00   1.015: 75   2   : 43   2   : 18  
##  Mean   :51.48   Mean   : 76.47   1.02 :106   3   : 43   3   : 14  
##  3rd Qu.:64.50   3rd Qu.: 80.00   1.025: 81   4   : 24   4   : 13  
##  Max.   :90.00   Max.   :180.00   NA’s : 47   5   :  1   5   :  3  
##  NA’s   :9       NA’s   :12                   NA’s: 46   NA’s: 49  
##       rbc             pc              pcc               ba           bgr     
##  normal  :201   normal  :259   present   : 42   present   : 22   Min.   : 22  
##  abnormal: 47   abnormal: 76   notpresent:354   notpresent:374   1st Qu.: 99  
##  NA’s    :152   NA’s    : 65   NA’s      :  4   NA’s      :  4   Median :121  
##                                                                  Mean   :148  
##                                                                  3rd Qu.:163  
##                                                                  Max.   :490  
##                                                                  NA’s   :44   
##       bu               sc              sod              pot        
##  Min.   :  1.50   Min.   : 0.400   Min.   :  4.5   Min.   : 2.500  
##  1st Qu.: 27.00   1st Qu.: 0.900   1st Qu.:135.0   1st Qu.: 3.800  
##  Median : 42.00   Median : 1.300   Median :138.0   Median : 4.400  
##  Mean   : 57.43   Mean   : 3.072   Mean   :137.5   Mean   : 4.627  
##  3rd Qu.: 66.00   3rd Qu.: 2.800   3rd Qu.:142.0   3rd Qu.: 4.900  
##  Max.   :391.00   Max.   :76.000   Max.   :163.0   Max.   :47.000  
##  NA’s   :19       NA’s   :17       NA’s   :87      NA’s   :88      
##       hemo            pcv             wbcc            rbcc         htn     
##  Min.   : 3.10   Min.   : 9.00   Min.   : 2200   Min.   :2.100   yes :147  
##  1st Qu.:10.30   1st Qu.:32.00   1st Qu.: 6500   1st Qu.:3.900   no  :251  
##  Median :12.65   Median :40.00   Median : 8000   Median :4.800   NA’s:  2  
##  Mean   :12.53   Mean   :38.88   Mean   : 8406   Mean   :4.707             
##  3rd Qu.:15.00   3rd Qu.:45.00   3rd Qu.: 9800   3rd Qu.:5.400             
##  Max.   :17.80   Max.   :54.00   Max.   :26400   Max.   :8.000             
##  NA’s   :52      NA’s   :71      NA’s   :106     NA’s   :131               
##     dm         cad         appet        pe         ane         status   
##  yes :137   yes : 34   good:317   yes : 76   yes : 60   ckd   :250  
##  no  :261   no  :364   poor: 82   no  :323   no  :339   notckd:150  
##  NA’s:  2   NA’s:  2   NA’s:  1   NA’s:  1   NA’s:  1               
##                                                                     
##                                                                     
##                                                                     
## 
```

```r
# Check missing values
missing_summary <- sapply(ckd_clean, function(x) sum(is.na(x)))
missing_df <- data.frame(
  Column = names(missing_summary),
  Missing_Count = missing_summary,
  Missing_Percent = round(missing_summary/nrow(ckd_clean)*100, 2)
) %>%
  arrange(desc(Missing_Count))
```

```r
print("Missing Value Summary:")
```

```
## [1] "Missing Value Summary:"
```

```r
print(missing_df)
```

```
##        Column Missing_Count Missing_Percent
## rbc       rbc           152           38.00
## rbcc     rbcc           131           32.75
## wbcc     wbcc           106           26.50
## pot       pot            88           22.00
## sod       sod            87           21.75
## pcv       pcv            71           17.75
## pc         pc            65           16.25
## hemo     hemo            52           13.00
## su         su            49           12.25
## sg         sg            47           11.75
## al         al            46           11.50
## bgr       bgr            44           11.00
## bu         bu            19            4.75
## sc         sc            17            4.25
## bp         bp            12            3.00
## age       age             9            2.25
## pcc       pcc             4            1.00
## ba         ba             4            1.00
## htn       htn             2            0.50
## dm         dm             2            0.50
## cad       cad             2            0.50
## appet   appet             1            0.25
## pe         pe             1            0.25
## ane       ane             1            0.25
## status status             0            0.00
```

```r
# Add binary outcome variable
# ckd_clean$status_binary <- ifelse(ckd_clean$status == "ckd", 1, 0)

# Check cleaned structure
cat("\nCleaned Data Structure:\n")
```

```
##
## Cleaned Data Structure:
```

```r
str(ckd_clean)
```

```
## ’data.frame’:    400 obs. of  25 variables:
##  $ age   : num  48 7 62 48 51 60 68 24 52 53 ...
##  $ bp    : num  80 50 80 70 80 90 70 NA 100 90 ...
##  $ sg    : Factor w/ 5 levels "1.005","1.01",..: 4 4 2 1 2 3 2 3 3 4 ...
##  $ al    : Factor w/ 6 levels "0","1","2","3",..: 2 5 3 5 3 4 1 3 4 3 ...
##  $ su    : Factor w/ 6 levels "0","1","2","3",..: 1 1 4 1 1 1 1 5 1 1 ...
##  $ rbc   : Factor w/ 2 levels "normal","abnormal": NA NA 1 1 1 NA NA 1 1 2 ...
##  $ pc    : Factor w/ 2 levels "normal","abnormal": 1 1 1 2 1 NA 1 2 2 2 ...
##  $ pcc   : Factor w/ 2 levels "present","notpresent": 2 2 2 1 2 2 2 2 1 1 ...
##  $ ba    : Factor w/ 2 levels "present","notpresent": 2 2 2 2 2 2 2 2 2 2 ...
##  $ bgr   : num  121 NA 423 117 106 74 100 410 138 70 ...
##  $ bu    : num  36 18 53 56 26 25 54 31 60 107 ...
##  $ sc    : num  1.2 0.8 1.8 3.8 1.4 1.1 24 1.1 1.9 7.2 ...
##  $ sod   : num  NA NA NA 111 NA 142 104 NA NA 114 ...
##  $ pot   : num  NA NA NA 2.5 NA 3.2 4 NA NA 3.7 ...
##  $ hemo  : num  15.4 11.3 9.6 11.2 11.6 12.2 12.4 12.4 10.8 9.5 ...
##  $ pcv   : num  44 38 31 32 35 39 36 44 33 29 ...
##  $ wbcc  : num  7800 6000 7500 6700 7300 7800 NA 6900 9600 12100 ...
##  $ rbcc  : num  5.2 NA NA 3.9 4.6 4.4 NA 5 4 3.7 ...
##  $ htn   : Factor w/ 2 levels "yes","no": 1 2 2 1 2 1 2 2 1 1 ...
##  $ dm    : Factor w/ 2 levels "yes","no": 1 2 1 2 2 1 2 1 1 1 ...
##  $ cad   : Factor w/ 2 levels "yes","no": 2 2 2 2 2 2 2 2 2 2 ...
##  $ appet : Factor w/ 2 levels "good","poor": 1 1 2 2 1 1 1 1 1 2 ...
##  $ pe    : Factor w/ 2 levels "yes","no": 2 2 2 1 2 1 2 1 2 2 ...
##  $ ane   : Factor w/ 2 levels "yes","no": 2 2 1 1 2 2 2 2 1 1 ...
##  $ status: Factor w/ 2 levels "ckd","notckd": 1 1 1 1 1 1 1 1 1 1 ...
```

```
# Summary statistics
cat("\nSummary Statistics:\n")
```

```
## 
## Summary Statistics:
```

```
summary(ckd_clean)
```

```
##       age              bp              sg          al         su     
##  Min.   : 2.00   Min.   : 50.00   1.005:  7   0   :199   0   :290  
##  1st Qu.:42.00   1st Qu.: 70.00   1.01 : 84   1   : 44   1   : 13  
##  Median :55.00   Median : 80.00   1.015: 75   2   : 43   2   : 18  
##  Mean   :51.48   Mean   : 76.47   1.02 :106   3   : 43   3   : 14  
##  3rd Qu.:64.50   3rd Qu.: 80.00   1.025: 81   4   : 24   4   : 13  
##  Max.   :90.00   Max.   :180.00   NA's : 47   5   :  1   5   :  3  
##  NA's   :9       NA's   :12                   NA's: 46   NA's: 49  
##        rbc              pc              pcc              ba            bgr     
##  normal  :201   normal  :259   present   : 42   present   : 22   Min.   : 22  
##  abnormal: 47   abnormal: 76   notpresent:354   notpresent:374   1st Qu.: 99  
##  NA's    :152   NA's    : 65   NA's      :  4   NA's      :  4   Median :121  
##                                                                  Mean   :148  
##                                                                  3rd Qu.:163  
##                                                                  Max.   :490  
##                                                                  NA's   :44   
##        bu               sc              sod             pot        
##  Min.   :  1.50   Min.   : 0.400   Min.   :  4.5   Min.   : 2.500  
##  1st Qu.: 27.00   1st Qu.: 0.900   1st Qu.:135.0   1st Qu.: 3.800  
##  Median : 42.00   Median : 1.300   Median :138.0   Median : 4.400  
##  Mean   : 57.43   Mean   : 3.072   Mean   :137.5   Mean   : 4.627  
##  3rd Qu.: 66.00   3rd Qu.: 2.800   3rd Qu.:142.0   3rd Qu.: 4.900  
##  Max.   :391.00   Max.   :76.000   Max.   :163.0   Max.   :47.000  
##  NA's   :19       NA's   :17       NA's   :87      NA's   :88      
##       hemo            pcv             wbcc            rbcc         htn     
##  Min.   : 3.10   Min.   : 9.00   Min.   : 2200   Min.   :2.100   yes :147
##  1st Qu.:10.30   1st Qu.:32.00   1st Qu.: 6500   1st Qu.:3.900   no  :251
##  Median :12.65   Median :40.00   Median : 8000   Median :4.800   NA's:  2
##  Mean   :12.53   Mean   :38.88   Mean   : 8406   Mean   :4.707
##  3rd Qu.:15.00   3rd Qu.:45.00   3rd Qu.: 9800   3rd Qu.:5.400
##  Max.   :17.80   Max.   :54.00   Max.   :26400   Max.   :8.000
##  NA's   :52      NA's   :71      NA's   :106     NA's   :131
##     dm        cad       appet        pe        ane         status
##  yes :137   yes : 34   good:317   yes : 76   yes : 60   ckd   :250
##  no  :261   no  :364   poor: 82   no  :323   no  :339   notckd:150
##  NA's:  2   NA's:  2   NA's:  1   NA's:  1   NA's:  1
##
##
##
##
```

4.1 DATA IMPUTATION

The MICE Algorithm

Multiple Imputation by Chained Equations is a robust, informative method of dealing with missing data in datasets. The procedure 'fills in' (imputes) missing data in a dataset through an iterative series of predictive models. In each iteration, each specified variable in the dataset is imputed using the other variables in the dataset. These iterations should be run until it appears that convergence has been met.

Data Leakage:

MICE is particularly useful if missing values are associated with the target variable in a way that introduces leakage. For instance, let's say you wanted to model customer retention at the time of sign up. A certain variable is collected at sign up or 1 month after sign up. The absence of that variable is a data leak, since it tells you that the customer did not retain for 1 month.

Funnel Analysis:

Information is often collected at different stages of a 'funnel'. MICE can be used to make educated guesses about the characteristics of entities at different points in a funnel.

Confidence Intervals:

MICE can be used to impute missing values, however it is important to keep in mind that these imputed values are a prediction. Creating multiple datasets with different imputed values allows you to do two types of inference:

- Imputed Value Distribution: A profile can be built for each imputed value, allowing you to make statements about the likely distribution of that value.
- Model Prediction Distribution: With multiple datasets, you can build multiple models and create a distribution of predictions for each sample. Those samples with imputed values which were not able to be imputed with much confidence would have a larger variance in their predictions.

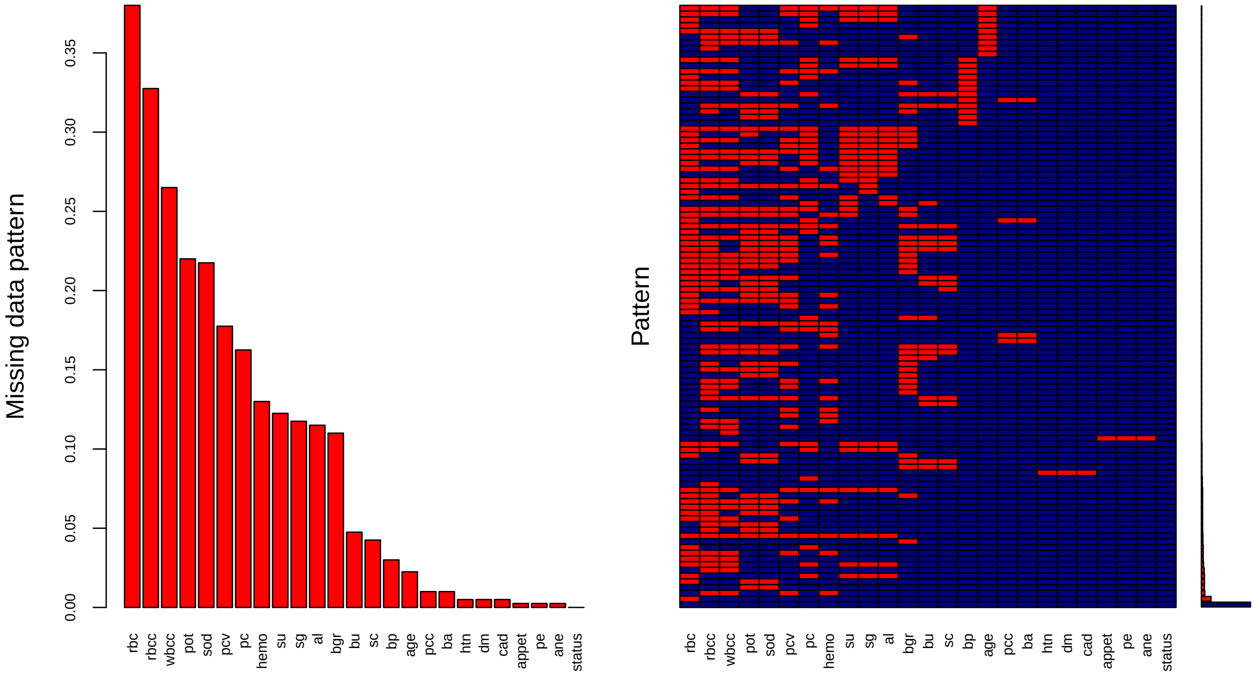
```r
## Missing Data Analysis and Imputation



# Calculate missingness
missing_summary <- data.frame(
  Variable = names(ckd_clean),
  Missing_Count = colSums(is.na(ckd_clean)),
  Missing_Percent = round(colSums(is.na(ckd_clean))/nrow(ckd_clean)*100, 2)
) %>%
  arrange(desc(Missing_Percent))

print("Missing Data Summary:")
```

```
## [1] "Missing Data Summary:"
```

```r
print(missing_summary)
```

```
##         Variable Missing_Count Missing_Percent
## rbc         rbc           152           38.00
## rbcc       rbcc           131           32.75
## wbcc       wbcc           106           26.50
## pot         pot            88           22.00
## sod         sod            87           21.75
## pcv         pcv            71           17.75
## pc           pc            65           16.25
## hemo       hemo            52           13.00
## su           su            49           12.25
## sg           sg            47           11.75
## al           al            46           11.50
## bgr         bgr            44           11.00
## bu           bu            19            4.75
## sc           sc            17            4.25
## bp           bp            12            3.00
## age         age             9            2.25
## pcc         pcc             4            1.00
## ba           ba             4            1.00
## htn         htn             2            0.50
## dm           dm             2            0.50
## cad         cad             2            0.50
## appet     appet             1            0.25
## pe           pe             1            0.25
## ane         ane             1            0.25
## status   status             0            0.00
```

```r
# Visualize missing data pattern
missing_plot <-  aggr(ckd_clean,
                     col = c('navyblue', 'red'),
                     numbers = TRUE,
                     sortVars = TRUE,
                     labels = names(ckd_clean),
                     cex.axis = .7,
                     gap = 3,
                     ylab = c("Missing data pattern", "Pattern"))
```



```
##
##  Variables sorted by number of missings:
##  Variable  Count
##       rbc 0.3800
##      rbcc 0.3275
##      wbcc 0.2650
##       pot 0.2200
##       sod 0.2175
##       pcv 0.1775
##        pc 0.1625
##      hemo 0.1300
##        su 0.1225
##        sg 0.1175
##        al 0.1150
##       bgr 0.1100
##        bu 0.0475
##        sc 0.0425
##        bp 0.0300
##       age 0.0225
##       pcc 0.0100
##        ba 0.0100
##       htn 0.0050
##        dm 0.0050
##       cad 0.0050
##     appet 0.0025
##        pe 0.0025
##       ane 0.0025
##    status 0.0000
```

defaultMethod = c("pmm", "logreg", "polyreg", "polr") A vector of length 4 containing the default imputation methods for 1) numeric data, 2) factor data with 2 levels, 3) factor data with > 2 unordered levels, and 4) factor data with > 2 ordered levels. By default, the method uses pmm, predictive mean matching (numeric data) logreg, logistic regression imputation (binary data, factor with 2 levels) polyreg, polytomous regression imputation for unordered categorical data (factor > 2 levels) polr, proportional odds model for (ordered, > 2 levels).

```
# Multiple imputation using MICE
cat("\nPerforming Multiple Imputation...\n")
```

```
##
## Performing Multiple Imputation...
```

```
imputed_data <- mice(ckd_clean,
                    m = 5,
                    maxit = 50,
                    method = 'pmm',
                    seed = 42)
```

```
##
##  iter imp variable
##   1   1  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn  d
##   1   2  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn  d
##   1   3  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn  d
##   1   4  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn  d
##   1   5  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn  d
##   2   1  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn  d
##   2   2  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn  d
##   2   3  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn  d
##   2   4  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn  d
##   2   5  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn  d
##   3   1  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn  d
##   3   2  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn  d
##   3   3  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn  d
##   3   4  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn  d
##   3   5  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn  d
##   4   1  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn  d
##   4   2  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn  d
##   4   3  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn  d
##   4   4  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn  d
##   4   5  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn  d
##   5   1  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn  d
##   5   2  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn  d
##   5   3  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn  d
##   5   4  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn  d
##   5   5  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn  d
##   6   1  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn  d
##   6   2  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn  d
##   6   3  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn  d
##   6   4  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn  d
##   6   5  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn  d
##   7   1  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn  d
##   7   2  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn  d
##   7   3  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn  d
##   7   4  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn  d
##   7   5  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn  d
##   8   1  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn  d
##   8   2  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn  d
##   8   3  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn  d
##   8   4  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn  d
##   8   5  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn  d
##   9   1  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn  d
##   9   2  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn  d
##   9   3  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn  d
##   9   4  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn  d
##   9   5  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn  d
##   10   1  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   10   2  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   10   3  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   10   4  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   10   5  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   11   1  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   11   2  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   11   3  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   11   4  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   11   5  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   12   1  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   12   2  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   12   3  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   12   4  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   12   5  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   13   1  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   13   2  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   13   3  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   13   4  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   13   5  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   14   1  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   14   2  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   14   3  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   14   4  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   14   5  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   15   1  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   15   2  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   15   3  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   15   4  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   15   5  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   16   1  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   16   2  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   16   3  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   16   4  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   16   5  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   17   1  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   17   2  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   17   3  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   17   4  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   17   5  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   18   1  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   18   2  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   18   3  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   18   4  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##    18   5  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##    19   1  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##    19   2  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##    19   3  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##    19   4  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##    19   5  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##    20   1  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##    20   2  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##    20   3  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##    20   4  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##    20   5  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##    21   1  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##    21   2  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##    21   3  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##    21   4  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##    21   5  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##    22   1  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##    22   2  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##    22   3  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##    22   4  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##    22   5  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##    23   1  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##    23   2  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##    23   3  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##    23   4  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##    23   5  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##    24   1  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##    24   2  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##    24   3  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##    24   4  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##    24   5  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##    25   1  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##    25   2  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##    25   3  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##    25   4  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##    25   5  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##    26   1  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##    26   2  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##    26   3  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##    26   4  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##    26   5  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##    27   1  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##    27   2  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##    27   3  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##    27   4  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##    27   5  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##    28   1  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##    28   2  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##    28   3  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##    28   4  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##    28   5  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##    29   1  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##    29   2  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##    29   3  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   29   4  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   29   5  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   30   1  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   30   2  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   30   3  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   30   4  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   30   5  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   31   1  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   31   2  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   31   3  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   31   4  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   31   5  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   32   1  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   32   2  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   32   3  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   32   4  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   32   5  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   33   1  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   33   2  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   33   3  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   33   4  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   33   5  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   34   1  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   34   2  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   34   3  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   34   4  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   34   5  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   35   1  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   35   2  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   35   3  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   35   4  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   35   5  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   36   1  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   36   2  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   36   3  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   36   4  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   36   5  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   37   1  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   37   2  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   37   3  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   37   4  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   37   5  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   38   1  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   38   2  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   38   3  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   38   4  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   38   5  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   39   1  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   39   2  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   39   3  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   39   4  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   39   5  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   40   1  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   40   2  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   40   3  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   40   4  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   40   5  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   41   1  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   41   2  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   41   3  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   41   4  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   41   5  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   42   1  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   42   2  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   42   3  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   42   4  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   42   5  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   43   1  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   43   2  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   43   3  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   43   4  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   43   5  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   44   1  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   44   2  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   44   3  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   44   4  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   44   5  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   45   1  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   45   2  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   45   3  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   45   4  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   45   5  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   46   1  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   46   2  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   46   3  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   46   4  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   46   5  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   47   1  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   47   2  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   47   3  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   47   4  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   47   5  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   48   1  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   48   2  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   48   3  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   48   4  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   48   5  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   49   1  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   49   2  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   49   3  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   49   4  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   49   5  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   50   1  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   50   2  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   50   3  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   50   4  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
##   50   5  age  bp  sg  al  su  rbc  pc  pcc  ba  bgr  bu  sc  sod  pot  hemo  pcv  wbcc  rbcc  htn
```

```
# Extract first imputed dataset
ckd_mice_imputed <- complete(imputed_data, 1)

# Check imputation quality
cat("\nMissing values after imputation:", sum(is.na(ckd_mice_imputed)), "\n")
```

```
##
## Missing values after imputation: 0
```

```
summary(ckd_clean)
```

```
##       age              bp             sg         al         su
##  Min.   : 2.00   Min.   : 50.00   1.005:  7   0   :199   0   :290
##  1st Qu.:42.00   1st Qu.: 70.00   1.01 : 84   1   : 44   1   : 13
##  Median :55.00   Median : 80.00   1.015: 75   2   : 43   2   : 18
##  Mean   :51.48   Mean   : 76.47   1.02 :106   3   : 43   3   : 14
##  3rd Qu.:64.50   3rd Qu.: 80.00   1.025: 81   4   : 24   4   : 13
##  Max.   :90.00   Max.   :180.00   NA’s : 47   5   :  1   5   :  3
##  NA’s   :9       NA’s   :12                   NA’s: 46   NA’s: 49
##        rbc             pc              pcc               ba           bgr
##  normal  :201   normal  :259   present   : 42   present   : 22   Min.   : 22
##  abnormal: 47   abnormal: 76   notpresent:354   notpresent:374   1st Qu.: 99
##  NA’s    :152   NA’s    : 65   NA’s      :  4   NA’s      :  4   Median :121
##                                                                  Mean   :148
##                                                                  3rd Qu.:163
##                                                                  Max.   :490
##                                                                  NA’s   :44
##        bu               sc              sod             pot
##  Min.   :  1.50   Min.   : 0.400   Min.   :  4.5   Min.   : 2.500
##  1st Qu.: 27.00   1st Qu.: 0.900   1st Qu.:135.0   1st Qu.: 3.800
##  Median : 42.00   Median : 1.300   Median :138.0   Median : 4.400
##  Mean   : 57.43   Mean   : 3.072   Mean   :137.5   Mean   : 4.627
##  3rd Qu.: 66.00   3rd Qu.: 2.800   3rd Qu.:142.0   3rd Qu.: 4.900
##  Max.   :391.00   Max.   :76.000   Max.   :163.0   Max.   :47.000
##  NA’s   :19       NA’s   :17       NA’s   :87      NA’s   :88
##       hemo            pcv             wbcc            rbcc         htn
##  Min.   : 3.10   Min.   : 9.00   Min.   : 2200   Min.   :2.100   yes :147
##  1st Qu.:10.30   1st Qu.:32.00   1st Qu.: 6500   1st Qu.:3.900   no  :251
##  Median :12.65   Median :40.00   Median : 8000   Median :4.800   NA’s:  2
##  Mean   :12.53   Mean   :38.88   Mean   : 8406   Mean   :4.707
##  3rd Qu.:15.00   3rd Qu.:45.00   3rd Qu.: 9800   3rd Qu.:5.400
##  Max.   :17.80   Max.   :54.00   Max.   :26400   Max.   :8.000
##  NA’s   :52      NA’s   :71      NA’s   :106     NA’s   :131
##     dm        cad      appet        pe        ane         status
##  yes :137   yes : 34   good:317   yes : 76   yes : 60   ckd   :250
##  no  :261   no  :364   poor: 82   no  :323   no  :339   notckd:150
##  NA’s:  2   NA’s:  2   NA’s:  1   NA’s:  1   NA’s:  1
##
##
##
##
```

```r
summary(ckd_mice_imputed)
```

```
##       age              bp            sg      al      su            rbc
##  Min.   : 2.0   Min.   : 50.0   1.005:  9   0:219   0:319   normal  :277
##  1st Qu.:42.0   1st Qu.: 70.0   1.01 :101   1: 53   1: 18   abnormal:123
##  Median :54.5   Median : 80.0   1.015: 89   2: 48   2: 23
##  Mean   :51.5   Mean   : 76.6   1.02 :115   3: 51   3: 17
##  3rd Qu.:65.0   3rd Qu.: 80.0   1.025: 86   4: 28   4: 19
##  Max.   :90.0   Max.   :180.0               5:  1   5:  4
##         pc              pcc               ba           bgr
##  normal  :313   present   : 42   present   : 22   Min.   : 22.0
##  abnormal: 87   notpresent:358   notpresent:378   1st Qu.: 99.0
##                                                   Median :121.0
##                                                   Mean   :149.2
##                                                   3rd Qu.:163.5
##                                                   Max.   :490.0
##         bu               sc              sod             pot
##  Min.   :  1.50   Min.   : 0.400   Min.   :  4.5   Min.   : 2.500
##  1st Qu.: 27.00   1st Qu.: 0.900   1st Qu.:135.0   1st Qu.: 3.800
##  Median : 41.00   Median : 1.250   Median :138.0   Median : 4.300
##  Mean   : 56.82   Mean   : 3.022   Mean   :137.5   Mean   : 4.671
##  3rd Qu.: 65.25   3rd Qu.: 2.800   3rd Qu.:141.0   3rd Qu.: 4.900
##  Max.   :391.00   Max.   :76.000   Max.   :163.0   Max.   :47.000
##       hemo            pcv             wbcc            rbcc        htn
##  Min.   : 3.10   Min.   : 9.00   Min.   : 2200   Min.   :2.100   yes:147
##  1st Qu.:10.38   1st Qu.:32.00   1st Qu.: 6675   1st Qu.:3.800   no :253
##  Median :12.50   Median :40.00   Median : 8100   Median :4.500
##  Mean   :12.45   Mean   :38.33   Mean   : 8562   Mean   :4.505
##  3rd Qu.:14.80   3rd Qu.:45.00   3rd Qu.: 9800   3rd Qu.:5.200
##  Max.   :17.80   Max.   :54.00   Max.   :26400   Max.   :8.000
##    dm       cad       appet       pe       ane         status
##  yes:137   yes: 34   good:318   yes: 76   yes: 60   ckd   :250
##  no :263   no :366   poor: 82   no :324   no :340   notckd:150
##
##
##
##
```

```r
setwd('/Volumes/HHD_iMac_Storage/URV/SCIENTIFIC_PROGRAMMING/FINAL/SP-Final-Project')
write_csv(ckd_mice_imputed, "data/processed/dataset_mice_imputed.csv",
          progress = show_progress())
```

```r
null_values <- colSums(is.na(ckd_mice_imputed))
print(null_values)
```

```
##    age     bp     sg     al     su    rbc     pc    pcc     ba    bgr     bu
##      0      0      0      0      0      0      0      0      0      0      0
##     sc    sod    pot   hemo    pcv   wbcc   rbcc    htn     dm    cad  appet
##      0      0      0      0      0      0      0      0      0      0      0
##     pe    ane status
##      0      0      0
```

Simple Imputation

Idea. Fill missing values with a single plausible value (one pass). Fast and convenient, but it underestimates uncertainty (standard errors too small) and can distort distributions.

Typical choices

- Mean/Median/Mode (baselines; median is more robust to skew)
- k-Nearest Neighbors (kNN) (borrows information from similar rows)
- Hot-deck (donor-based; similar spirit to kNN)

VIM:KNN Detecting missing values mechanisms is usually done by statistical tests or models. Visualization of missing and imputed values can support the test decision, but also reveals more details about the data structure. Most notably, statistical requirements for a test can be checked graphically, and problems like outliers or skewed data distributions can be discovered. Furthermore, the included plot methods may also be able to detect missing values mechanisms in the first place.

k-Nearest Neighbour Imputation based on a variation of the Gower Distance for numerical, categorical, ordered and semi-continous variables.

EXAMPLE

```
# kNN imputation with VIM::kNN (works on data frames; chooses donors by similarity)
# library(VIM)
#
# # We impute only BMI here; set k=5 as a reasonable starting point.
# ckd_knn <- ckd_clean |>
#   select(age, bp, ) |>
#   VIM::kNN(k = 5, imp_var = FALSE
#            ,trace = TRUE, )  # imp_var=FALSE avoids extra *_imp columns
#
# # Check imputation effect
# sum(is.na(ckd_knn$BMI))   # original missing BMI
```

```
null_values <- colSums(is.na(ckd_clean))
print(null_values)
```

```
##    age     bp     sg     al     su    rbc     pc    pcc     ba    bgr     bu
##      9     12     47     46     49    152     65      4      4     44     19
##     sc    sod    pot   hemo    pcv   wbcc   rbcc    htn     dm    cad  appet
##     17     87     88     52     71    106    131      2      2      2      1
##     pe    ane status
##      1      1      0
```

```
numeric_vars <- c("age", "bp", "bgr", "bu", "sc", "sod", "pot",
                  "hemo", "pcv", "wbcc", "rbcc")

categorical_vars <- c("sg", "al", "su", "rbc", "pc", "pcc", "ba",
                      "htn", "dm", "cad", "appet", "pe", "ane", "status")

binary_vars <- c("htn", "dm", "cad", "appet", "pe", "ane")
```

```r
# Basic kNN imputation using all variables
cat("Starting basic kNN imputation (k=5)...\n")
```

Basic kNN imputation using all variables

```
## Starting basic kNN imputation (k=5)...
```

```r
start_time <- Sys.time()

ckd_knn_basic <- VIM::kNN(
  data = ckd_clean,  # Remove derived variable
  k = 5,
  imp_var = FALSE,  # Don't create imputation indicator variables
  trace = TRUE,     # Show progress
  useImputedDist = TRUE  # Use imputed values for distance calculation
)
```

```
##      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc    rbcc 
##    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0     2.1 
##      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc    rbcc 
##   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0 26400.0     8.0 

##     age     bgr      bu      sc     sod     pot    hemo     pcv    wbcc    rbcc 
##     2.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0     2.1 
##     age     bgr      bu      sc     sod     pot    hemo     pcv    wbcc    rbcc 
##    90.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0 26400.0     8.0 

##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc 
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0 
##    rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv 
##     2.1    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0 
##    wbcc    rbcc 
## 26400.0     8.0 

##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc 
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0 
##    rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv 
##     2.1    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0 
##    wbcc    rbcc 
## 26400.0     8.0 

##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc 
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0 
##    rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv 
##     2.1    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0 
##    wbcc    rbcc 
## 26400.0     8.0 

##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc 
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0 
##    rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv 
##     2.1    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0 
##    wbcc    rbcc 
## 26400.0     8.0 
```

```
##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc 
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0 
##    rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv 
##     2.1    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0 
##    wbcc    rbcc 
## 26400.0     8.0 
```

```
##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc 
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0 
##    rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv 
##     2.1    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0 
##    wbcc    rbcc 
## 26400.0     8.0 
```

```
##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc 
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0 
##    rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv 
##     2.1    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0 
##    wbcc    rbcc 
## 26400.0     8.0 
```

```
##     age      bp      bu      sc     sod     pot    hemo     pcv    wbcc    rbcc 
##     2.0    50.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0     2.1 
##     age      bp      bu      sc     sod     pot    hemo     pcv    wbcc    rbcc 
##    90.0   180.0   391.0    76.0   163.0    47.0    17.8    54.0 26400.0     8.0 
```

```
##     age      bp     bgr      sc     sod     pot    hemo     pcv    wbcc    rbcc 
##     2.0    50.0    22.0     0.4     4.5     2.5     3.1     9.0  2200.0     2.1 
##     age      bp     bgr      sc     sod     pot    hemo     pcv    wbcc    rbcc 
##    90.0   180.0   490.0    76.0   163.0    47.0    17.8    54.0 26400.0     8.0 
```

```
##     age      bp     bgr      bu     sod     pot    hemo     pcv    wbcc    rbcc 
##     2.0    50.0    22.0     1.5     4.5     2.5     3.1     9.0  2200.0     2.1 
##     age      bp     bgr      bu     sod     pot    hemo     pcv    wbcc    rbcc 
##    90.0   180.0   490.0   391.0   163.0    47.0    17.8    54.0 26400.0     8.0 
```

```
##     age      bp     bgr      bu      sc     pot    hemo     pcv    wbcc    rbcc 
##     2.0    50.0    22.0     1.5     0.4     2.5     3.1     9.0  2200.0     2.1 
##     age      bp     bgr      bu      sc     pot    hemo     pcv    wbcc    rbcc 
##    90.0   180.0   490.0   391.0    76.0    47.0    17.8    54.0 26400.0     8.0 
```

```
##     age      bp     bgr      bu      sc     sod    hemo     pcv    wbcc    rbcc 
##     2.0    50.0    22.0     1.5     0.4     4.5     3.1     9.0  2200.0     2.1 
##     age      bp     bgr      bu      sc     sod    hemo     pcv    wbcc    rbcc 
##    90.0   180.0   490.0   391.0    76.0   163.0    17.8    54.0 26400.0     8.0 
##     age      bp     bgr      bu      sc     sod     pot     pcv    wbcc    rbcc 
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     9.0  2200.0     2.1 
##     age      bp     bgr      bu      sc     sod     pot     pcv    wbcc    rbcc 
##    90.0   180.0   490.0   391.0    76.0   163.0    47.0    54.0 26400.0     8.0 

##     age      bp     bgr      bu      sc     sod     pot    hemo    wbcc    rbcc 
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1  2200.0     2.1 
##     age      bp     bgr      bu      sc     sod     pot    hemo    wbcc    rbcc 
##    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8 26400.0     8.0 

##   age    bp   bgr    bu    sc   sod   pot  hemo   pcv  rbcc   age    bp   bgr 
##   2.0  50.0  22.0   1.5   0.4   4.5   2.5   3.1   9.0   2.1  90.0 180.0 490.0 
##    bu    sc   sod   pot  hemo   pcv  rbcc 
## 391.0  76.0 163.0  47.0  17.8  54.0   8.0 

##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc 
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0 
##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc 
##    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0 26400.0 

##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc 
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0 
##    rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv 
##     2.1    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0 
##    wbcc    rbcc 
## 26400.0     8.0 

##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc 
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0 
##    rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv 
##     2.1    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0 
##    wbcc    rbcc 
## 26400.0     8.0 

##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc 
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0 
##    rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv 
##     2.1    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0 
##    wbcc    rbcc 
## 26400.0     8.0 

##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc 
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0 
##    rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv 
##     2.1    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0 
##    wbcc    rbcc 
## 26400.0     8.0 

##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc 
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0 
##    rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv 
##     2.1    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0 
##    wbcc    rbcc 
## 26400.0     8.0
##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0
##    rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv
##     2.1    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0
##    wbcc    rbcc
## 26400.0     8.0
```

```
## Time difference of 1.043229 secs
```

```
end_time <- Sys.time()
cat(sprintf("Basic kNN imputation completed in %.2f seconds\n",
            end_time - start_time))
```

```
## Basic kNN imputation completed in 1.05 seconds
```

```
# Check missing values after imputation
missing_after <- colSums(is.na(ckd_knn_basic))
cat("\nMissing values after basic kNN imputation:\n")
```

```
##
## Missing values after basic kNN imputation:
```

```
print(missing_after[missing_after > 0])
```

```
## named numeric(0)
```

```
if(sum(missing_after) > 0) {
  cat("Warning: Not all missing values were imputed\n")
  cat("This is likely due to:")
  cat("1. Too many missing values in some observations\n")
  cat("2. Insufficient complete cases for some variables\n")
  cat("3. Need to increase k value\n")
}
```

```
null_values <- colSums(is.na(ckd_knn_basic))
print(null_values)
```

```
##    age     bp     sg     al     su    rbc     pc    pcc     ba    bgr     bu
##      0      0      0      0      0      0      0      0      0      0      0
##     sc    sod    pot   hemo    pcv   wbcc   rbcc    htn     dm    cad  appet
##      0      0      0      0      0      0      0      0      0      0      0
##     pe    ane status
##      0      0      0
```

```
summary(ckd_knn_basic)
```

```
##       age              bp              sg        al       su            rbc
##  Min.   : 2.00   Min.   : 50.00   1.005:  7   0:223   0:334   normal  :322
##  1st Qu.:42.00   1st Qu.: 70.00   1.01 :101   1: 53   1: 13   abnormal: 78
##  Median :54.50   Median : 80.00   1.015: 90   2: 46   2: 18
##  Mean   :51.50   Mean   : 76.45   1.02 :116   3: 53   3: 17
##  3rd Qu.:64.25   3rd Qu.: 80.00   1.025: 86   4: 24   4: 15
##  Max.   :90.00   Max.   :180.00               5:  1   5:  3
##         pc              pcc               ba            bgr
##  normal  :319   present   : 42   present   : 22   Min.   : 22.0
##  abnormal: 81   notpresent:358   notpresent:378   1st Qu.:100.0
##                                                   Median :121.0
##                                                   Mean   :146.2
##                                                   3rd Qu.:158.2
##                                                   Max.   :490.0
##        bu               sc               sod              pot
##  Min.   :  1.50   Min.   : 0.400   Min.   :  4.5   Min.   : 2.500
##  1st Qu.: 26.00   1st Qu.: 0.900   1st Qu.:135.0   1st Qu.: 3.900
##  Median : 40.00   Median : 1.200   Median :138.0   Median : 4.300
##  Mean   : 56.18   Mean   : 2.987   Mean   :137.5   Mean   : 4.561
##  3rd Qu.: 64.25   3rd Qu.: 2.725   3rd Qu.:141.0   3rd Qu.: 4.900
##  Max.   :391.00   Max.   :76.000   Max.   :163.0   Max.   :47.000
##       hemo             pcv             wbcc             rbcc         htn
##  Min.   : 3.10   Min.   : 9.00   Min.   : 2200   Min.   :2.100   yes:147
##  1st Qu.:10.38   1st Qu.:32.00   1st Qu.: 6775   1st Qu.:3.900   no :253
##  Median :12.25   Median :38.00   Median : 7900   Median :4.600
##  Mean   :12.39   Mean   :38.27   Mean   : 8289   Mean   :4.555
##  3rd Qu.:14.62   3rd Qu.:44.00   3rd Qu.: 9500   3rd Qu.:5.200
##  Max.   :17.80   Max.   :54.00   Max.   :26400   Max.   :8.000
##    dm       cad       appet       pe       ane        status
##  yes:137   yes: 34   good:318   yes: 76   yes: 60   ckd   :250
##  no :263   no :366   poor: 82   no :324   no :340   notckd:150
##
##
##
##
```

```r
# For datasets with high missingness, use multi-stage imputation
cat("\n=== MULTI-STAGE kNN IMPUTATION ===\n")
```

Multi-stage imputation

```
##
## === MULTI-STAGE kNN IMPUTATION ===
```

```r
# Stage 1: Impute variables with low missingness first
ckd_stage1 <- ckd_clean

# Identify variables by missingness level
missing_levels <- data.frame(
  Variable = names(ckd_stage1),
  Missing_Pct = colMeans(is.na(ckd_stage1)) * 100
) %>%
  mutate(
    Level = case_when(
      Missing_Pct < 10 ~ "Low",
      Missing_Pct >= 10 & Missing_Pct < 30 ~ "Medium",
      Missing_Pct >= 30 ~ "High"
    )
  )

print("Missingness Levels:")
```

```
## [1] "Missingness Levels:"
```

```r
print(table(missing_levels$Level))
```

```
##
##   High    Low Medium
##      2     13     10
```

```r
# Stage 1: Impute low missingness variables
low_missing_vars <- missing_levels$Variable[missing_levels$Level == "Low"]
cat("\nStage 1: Imputing low missingness variables (k=10)...\n")
```

```
##
## Stage 1: Imputing low missingness variables (k=10)...
```

```r
if(length(low_missing_vars) > 0) {
  ckd_stage1 <- VIM::kNN(
    data = ckd_stage1,
    variable = low_missing_vars,
    k = 10,
    imp_var = FALSE,
    trace = FALSE,
    useImputedDist = TRUE
  )
}
```

```
##      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc    rbcc
##    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0     2.1
##      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc    rbcc
##   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0 26400.0     8.0
##     age     bgr      bu      sc     sod     pot    hemo     pcv    wbcc    rbcc
##     2.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0     2.1
##     age     bgr      bu      sc     sod     pot    hemo     pcv    wbcc    rbcc
##    90.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0 26400.0     8.0
##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0
##    rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv
##     2.1    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0
##    wbcc    rbcc
## 26400.0     8.0
##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0
##    rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv
##     2.1    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0
##    wbcc    rbcc
## 26400.0     8.0
##     age      bp     bgr      sc     sod     pot    hemo     pcv    wbcc    rbcc
##     2.0    50.0    22.0     0.4     4.5     2.5     3.1     9.0  2200.0     2.1
##     age      bp     bgr      sc     sod     pot    hemo     pcv    wbcc    rbcc
##    90.0   180.0   490.0    76.0   163.0    47.0    17.8    54.0 26400.0     8.0
##     age      bp     bgr      bu     sod     pot    hemo     pcv    wbcc    rbcc
##     2.0    50.0    22.0     1.5     4.5     2.5     3.1     9.0  2200.0     2.1
##     age      bp     bgr      bu     sod     pot    hemo     pcv    wbcc    rbcc
##    90.0   180.0   490.0   391.0   163.0    47.0    17.8    54.0 26400.0     8.0
##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0
##    rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv
##     2.1    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0
##    wbcc    rbcc
## 26400.0     8.0
##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0
##    rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv
##     2.1    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0
##    wbcc    rbcc
## 26400.0     8.0
##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0
##    rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv
##     2.1    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0
##    wbcc    rbcc
## 26400.0     8.0
##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0
##    rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv
##     2.1    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0
##    wbcc    rbcc
## 26400.0     8.0
##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0
##    rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv
##     2.1    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0
##    wbcc    rbcc
## 26400.0     8.0
##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0
##    rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv
##     2.1    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0
##    wbcc    rbcc
## 26400.0     8.0
```

```r
# Stage 2: Impute medium missingness variables using imputed variables
medium_missing_vars <- missing_levels$Variable[missing_levels$Level == "Medium"]
cat("Stage 2: Imputing medium missingness variables (k=15)...\n")
```

```
## Stage 2: Imputing medium missingness variables (k=15)...
```

```r
if(length(medium_missing_vars) > 0) {
  ckd_stage1 <- VIM::kNN(
    data = ckd_stage1,
    variable = medium_missing_vars,
    k = 15,
    imp_var = FALSE,
    trace = FALSE,
    useImputedDist = TRUE
  )
}
```

```
##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0
##    rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv
##     2.1    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0
##    wbcc    rbcc
## 26400.0     8.0
##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0
##    rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv
##     2.1    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0
##    wbcc    rbcc
## 26400.0     8.0
##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0
##    rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv
##     2.1    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0
##    wbcc    rbcc
## 26400.0     8.0
##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0
##    rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv
##     2.1    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0
##    wbcc    rbcc
## 26400.0     8.0
##     age      bp      bu      sc     sod     pot    hemo     pcv    wbcc    rbcc
##     2.0    50.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0     2.1
##     age      bp      bu      sc     sod     pot    hemo     pcv    wbcc    rbcc
##    90.0   180.0   391.0    76.0   163.0    47.0    17.8    54.0 26400.0     8.0
##     age      bp     bgr      bu      sc     pot    hemo     pcv    wbcc    rbcc
##     2.0    50.0    22.0     1.5     0.4     2.5     3.1     9.0  2200.0     2.1
##     age      bp     bgr      bu      sc     pot    hemo     pcv    wbcc    rbcc
##    90.0   180.0   490.0   391.0    76.0    47.0    17.8    54.0 26400.0     8.0
##     age      bp     bgr      bu      sc     sod    hemo     pcv    wbcc    rbcc
##     2.0    50.0    22.0     1.5     0.4     4.5     3.1     9.0  2200.0     2.1
##     age      bp     bgr      bu      sc     sod    hemo     pcv    wbcc    rbcc
##    90.0   180.0   490.0   391.0    76.0   163.0    17.8    54.0 26400.0     8.0
##     age      bp     bgr      bu      sc     sod     pot     pcv    wbcc    rbcc
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     9.0  2200.0     2.1
##     age      bp     bgr      bu      sc     sod     pot     pcv    wbcc    rbcc
##    90.0   180.0   490.0   391.0    76.0   163.0    47.0    54.0 26400.0     8.0
##     age      bp     bgr      bu      sc     sod     pot    hemo    wbcc    rbcc
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1  2200.0     2.1
##      age       bp      bgr       bu       sc      sod      pot     hemo     wbcc     rbcc
##     90.0    180.0    490.0    391.0     76.0    163.0     47.0     17.8  26400.0      8.0
##    age     bp    bgr     bu     sc    sod    pot   hemo    pcv   rbcc    age     bp    bgr
##    2.0   50.0   22.0    1.5    0.4    4.5    2.5    3.1    9.0    2.1   90.0  180.0  490.0
##     bu     sc    sod    pot   hemo    pcv   rbcc
##  391.0   76.0  163.0   47.0   17.8   54.0    8.0
```

```
# Stage 3: Impute high missingness variables last
high_missing_vars <- missing_levels$Variable[missing_levels$Level == "High"]
cat("Stage 3: Imputing high missingness variables (k=20)...\n")
```

```
## Stage 3: Imputing high missingness variables (k=20)...
```

```
if(length(high_missing_vars) > 0) {
  ckd_multistage <- VIM::kNN(
    data = ckd_stage1,
    variable = high_missing_vars,
    k = 20,
    imp_var = FALSE,
    trace = FALSE,
    useImputedDist = TRUE
  )
} else {
  ckd_multistage <- ckd_stage1
}
```

```
##      age       bp      bgr       bu       sc      sod      pot     hemo      pcv     wbcc
##      2.0     50.0     22.0      1.5      0.4      4.5      2.5      3.1      9.0   2200.0
##     rbcc      age       bp      bgr       bu       sc      sod      pot     hemo      pcv
##      2.1     90.0    180.0    490.0    391.0     76.0    163.0     47.0     17.8     54.0
##     wbcc     rbcc
##  26400.0      8.0
##      age       bp      bgr       bu       sc      sod      pot     hemo      pcv     wbcc
##      2.0     50.0     22.0      1.5      0.4      4.5      2.5      3.1      9.0   2200.0
##      age       bp      bgr       bu       sc      sod      pot     hemo      pcv     wbcc
##     90.0    180.0    490.0    391.0     76.0    163.0     47.0     17.8     54.0  26400.0
```

```
# Check results
cat("\nMissing values after multi-stage kNN imputation:\n")
```

```
##
## Missing values after multi-stage kNN imputation:
```

```
print(colSums(is.na(ckd_multistage)))
```

```
##    age     bp     sg     al     su    rbc     pc    pcc     ba    bgr     bu
##      0      0      0      0      0      0      0      0      0      0      0
##     sc    sod    pot   hemo    pcv   wbcc   rbcc    htn     dm    cad  appet
##      0      0      0      0      0      0      0      0      0      0      0
##     pe    ane status
##      0      0      0
```

```r
# More sophisticated kNN imputation
cat("\n=== ADVANCED kNN IMPUTATION ===\n")
```

More sophisticated kNN imputation

```
##
## === ADVANCED kNN IMPUTATION ===
```

```r
# Calculate optimal k (rule of thumb: sqrt(n))
optimal_k <- round(sqrt(nrow(ckd_clean)))
cat(sprintf("Optimal k (sqrt(n)): %d\n", optimal_k))
```

```
## Optimal k (sqrt(n)): 20
```

```r
# Create a copy for imputation
ckd_for_imputation <- ckd_clean

# Step 1: Identify variables with too many missing values
missing_pct <- colMeans(is.na(ckd_for_imputation)) * 100
high_missing_vars <- names(missing_pct[missing_pct > 30])

cat("Variables with >30% missing:", paste(high_missing_vars, collapse = ", "), "\n")
```

```
## Variables with >30% missing: rbc, rbcc
```

```r
# Step 2: Create distance matrix using only complete-ish variables
# Select variables with <20% missing for distance calculation
good_distance_vars <- names(missing_pct[missing_pct < 20])

cat("Using these variables for distance calculation:",
    paste(good_distance_vars, collapse = ", "), "\n")
```

```
## Using these variables for distance calculation: age, bp, sg, al, su, pc, pcc, ba, bgr, bu, sc, hemo,
```

```r
# Step 3: Perform kNN imputation with custom parameters
cat("\nPerforming advanced kNN imputation...\n")
```

```
##
## Performing advanced kNN imputation...
```

```r
start_time <- Sys.time()

ckd_knn_advanced <- VIM::kNN(
  data = ckd_for_imputation,
  variable = colnames(ckd_for_imputation),  # Impute all variables
  dist_var = good_distance_vars,            # Use only good variables for distance
  k = optimal_k,                            # Optimal k
  numFun = median,                          # Use median for numeric (robust)
  catFun = function(x) {                      # Custom mode function
    tab <- table(x)
    names(tab)[which.max(tab)]
  },
  imp_var = FALSE,                             # Keep imputation indicators
  trace = TRUE,
  useImputedDist = TRUE
)
```

```
##    bp   bgr    bu    sc  hemo   pcv    bp   bgr    bu    sc  hemo   pcv 
##  50.0  22.0   1.5   0.4   3.1   9.0 180.0 490.0 391.0  76.0  17.8  54.0 

##   age   bgr    bu    sc  hemo   pcv   age   bgr    bu    sc  hemo   pcv 
##   2.0  22.0   1.5   0.4   3.1   9.0  90.0 490.0 391.0  76.0  17.8  54.0 

##   age    bp   bgr    bu    sc  hemo   pcv   age    bp   bgr    bu    sc  hemo 
##   2.0  50.0  22.0   1.5   0.4   3.1   9.0  90.0 180.0 490.0 391.0  76.0  17.8 
##   pcv 
##  54.0 

##   age    bp   bgr    bu    sc  hemo   pcv   age    bp   bgr    bu    sc  hemo 
##   2.0  50.0  22.0   1.5   0.4   3.1   9.0  90.0 180.0 490.0 391.0  76.0  17.8 
##   pcv 
##  54.0 

##   age    bp   bgr    bu    sc  hemo   pcv   age    bp   bgr    bu    sc  hemo 
##   2.0  50.0  22.0   1.5   0.4   3.1   9.0  90.0 180.0 490.0 391.0  76.0  17.8 
##   pcv 
##  54.0 

##   age    bp   bgr    bu    sc  hemo   pcv   age    bp   bgr    bu    sc  hemo 
##   2.0  50.0  22.0   1.5   0.4   3.1   9.0  90.0 180.0 490.0 391.0  76.0  17.8 
##   pcv 
##  54.0 

##   age    bp   bgr    bu    sc  hemo   pcv   age    bp   bgr    bu    sc  hemo 
##   2.0  50.0  22.0   1.5   0.4   3.1   9.0  90.0 180.0 490.0 391.0  76.0  17.8 
##   pcv 
##  54.0 

##   age    bp   bgr    bu    sc  hemo   pcv   age    bp   bgr    bu    sc  hemo 
##   2.0  50.0  22.0   1.5   0.4   3.1   9.0  90.0 180.0 490.0 391.0  76.0  17.8 
##   pcv 
##  54.0 

##   age    bp   bgr    bu    sc  hemo   pcv   age    bp   bgr    bu    sc  hemo 
##   2.0  50.0  22.0   1.5   0.4   3.1   9.0  90.0 180.0 490.0 391.0  76.0  17.8 
##   pcv 
##  54.0 

##   age    bp    bu    sc  hemo   pcv   age    bp    bu    sc  hemo   pcv 
##   2.0  50.0   1.5   0.4   3.1   9.0  90.0 180.0 391.0  76.0  17.8  54.0 
##   age    bp   bgr    sc  hemo   pcv   age    bp   bgr    sc  hemo   pcv
##   2.0  50.0  22.0   0.4   3.1   9.0  90.0 180.0 490.0  76.0  17.8  54.0

##   age    bp   bgr    bu  hemo   pcv   age    bp   bgr    bu  hemo   pcv
##   2.0  50.0  22.0   1.5   3.1   9.0  90.0 180.0 490.0 391.0  17.8  54.0

##   age    bp   bgr    bu    sc  hemo   pcv   age    bp   bgr    bu    sc  hemo
##   2.0  50.0  22.0   1.5   0.4   3.1   9.0  90.0 180.0 490.0 391.0  76.0  17.8
##   pcv
##  54.0

##   age    bp   bgr    bu    sc  hemo   pcv   age    bp   bgr    bu    sc  hemo
##   2.0  50.0  22.0   1.5   0.4   3.1   9.0  90.0 180.0 490.0 391.0  76.0  17.8
##   pcv
##  54.0

##   age    bp   bgr    bu    sc   pcv   age    bp   bgr    bu    sc   pcv
##   2.0  50.0  22.0   1.5   0.4   9.0  90.0 180.0 490.0 391.0  76.0  54.0

##   age    bp   bgr    bu    sc  hemo   age    bp   bgr    bu    sc  hemo
##   2.0  50.0  22.0   1.5   0.4   3.1  90.0 180.0 490.0 391.0  76.0  17.8

##   age    bp   bgr    bu    sc  hemo   pcv   age    bp   bgr    bu    sc  hemo
##   2.0  50.0  22.0   1.5   0.4   3.1   9.0  90.0 180.0 490.0 391.0  76.0  17.8
##   pcv
##  54.0

##   age    bp   bgr    bu    sc  hemo   pcv   age    bp   bgr    bu    sc  hemo
##   2.0  50.0  22.0   1.5   0.4   3.1   9.0  90.0 180.0 490.0 391.0  76.0  17.8
##   pcv
##  54.0

##   age    bp   bgr    bu    sc  hemo   pcv   age    bp   bgr    bu    sc  hemo
##   2.0  50.0  22.0   1.5   0.4   3.1   9.0  90.0 180.0 490.0 391.0  76.0  17.8
##   pcv
##  54.0

##   age    bp   bgr    bu    sc  hemo   pcv   age    bp   bgr    bu    sc  hemo
##   2.0  50.0  22.0   1.5   0.4   3.1   9.0  90.0 180.0 490.0 391.0  76.0  17.8
##   pcv
##  54.0

##   age    bp   bgr    bu    sc  hemo   pcv   age    bp   bgr    bu    sc  hemo
##   2.0  50.0  22.0   1.5   0.4   3.1   9.0  90.0 180.0 490.0 391.0  76.0  17.8
##   pcv
##  54.0

##   age    bp   bgr    bu    sc  hemo   pcv   age    bp   bgr    bu    sc  hemo
##   2.0  50.0  22.0   1.5   0.4   3.1   9.0  90.0 180.0 490.0 391.0  76.0  17.8
##   pcv
##  54.0
##   age    bp   bgr    bu    sc  hemo   pcv   age    bp   bgr    bu    sc  hemo 
##   2.0  50.0  22.0   1.5   0.4   3.1   9.0  90.0 180.0 490.0 391.0  76.0  17.8 
##   pcv 
##  54.0 
```

```
##   age    bp   bgr    bu    sc  hemo   pcv   age    bp   bgr    bu    sc  hemo 
##   2.0  50.0  22.0   1.5   0.4   3.1   9.0  90.0 180.0 490.0 391.0  76.0  17.8 
##   pcv 
##  54.0 
```

```
## Time difference of 0.963021 secs
```

```r
end_time <- Sys.time()
cat(sprintf("Advanced kNN imputation completed in %.2f seconds\n",
            end_time - start_time))
```

```
## Advanced kNN imputation completed in 0.96 seconds
```

```r
# Extract just the imputed data (without indicator columns)
imputed_cols <- !grepl("_imp$", colnames(ckd_knn_advanced))
ckd_imputed <- ckd_knn_advanced[, imputed_cols]

# Add back status feature
ckd_imputed$status <- ckd_clean$status

# Check results
cat("\nMissing values after advanced kNN imputation:\n")
```

```
## 
## Missing values after advanced kNN imputation:
```

```r
print(colSums(is.na(ckd_imputed)))
```

```
##    age     bp     sg     al     su    rbc     pc    pcc     ba    bgr     bu 
##      0      0      0      0      0      0      0      0      0      0      0 
##     sc    sod    pot   hemo    pcv   wbcc   rbcc    htn     dm    cad  appet 
##      0      0      0      0      0      0      0      0      0      0      0 
##     pe    ane status 
##      0      0      0 
```

```r
# Create imputation summary
imp_indicators <- ckd_knn_advanced[, grepl("_imp$", colnames(ckd_knn_advanced))]
imp_summary <- data.frame(
  Variable = gsub("_imp", "", colnames(imp_indicators)),
  Imputed_Count = colSums(imp_indicators),
  Imputed_Percent = round(colSums(imp_indicators)/nrow(ckd_imputed)*100, 2)
) %>%
  arrange(desc(Imputed_Percent))

print("Imputation Summary:")
```

```
## [1] "Imputation Summary:"
```

```
print(imp_summary)
```

```
## [1] Variable        Imputed_Count   Imputed_Percent
## <0 rows> (or 0-length row.names)
```

```


# Compare different k values
cat("\n=== COMPARING DIFFERENT k VALUES ===\n")
```

Comparison different k values

```
##
## === COMPARING DIFFERENT k VALUES ===
```

```
k_values <- c(3, 5, 10, 15, 20, 25)
imputation_results <- list()

for(k_val in k_values) {
  cat(sprintf("\nTesting k = %d...\n", k_val))

  # Impute with current k
  ckd_temp <- VIM::kNN(
    data = ckd_for_imputation,
    k = k_val,
    imp_var = FALSE,
    trace = TRUE,
    useImputedDist = TRUE
  )

  # Store results
  imputation_results[[as.character(k_val)]] <- list(
    k = k_val,
    missing_after = sum(is.na(ckd_temp)),
    variables_imputed = sum(colSums(is.na(ckd_temp)) == 0)
  )
}
```

```
##
## Testing k = 3...

##      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc    rbcc
##    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0     2.1
##      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc    rbcc
##   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0 26400.0     8.0

##     age     bgr      bu      sc     sod     pot    hemo     pcv    wbcc    rbcc
##     2.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0     2.1
##     age     bgr      bu      sc     sod     pot    hemo     pcv    wbcc    rbcc
##    90.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0 26400.0     8.0
##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc 
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0 
##    rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv 
##     2.1    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0 
##    wbcc    rbcc 
## 26400.0     8.0
```

```
##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc 
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0 
##    rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv 
##     2.1    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0 
##    wbcc    rbcc 
## 26400.0     8.0
```

```
##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc 
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0 
##    rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv 
##     2.1    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0 
##    wbcc    rbcc 
## 26400.0     8.0
```

```
##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc 
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0 
##    rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv 
##     2.1    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0 
##    wbcc    rbcc 
## 26400.0     8.0
```

```
##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc 
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0 
##    rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv 
##     2.1    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0 
##    wbcc    rbcc 
## 26400.0     8.0
```

```
##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc 
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0 
##    rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv 
##     2.1    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0 
##    wbcc    rbcc 
## 26400.0     8.0
```

```
##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc 
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0 
##    rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv 
##     2.1    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0 
##    wbcc    rbcc 
## 26400.0     8.0
```

```
##     age      bp      bu      sc     sod     pot    hemo     pcv    wbcc    rbcc 
##     2.0    50.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0     2.1 
##     age      bp      bu      sc     sod     pot    hemo     pcv    wbcc    rbcc 
##    90.0   180.0   391.0    76.0   163.0    47.0    17.8    54.0 26400.0     8.0
##     age      bp     bgr      sc     sod     pot    hemo     pcv    wbcc    rbcc
##     2.0    50.0    22.0     0.4     4.5     2.5     3.1     9.0  2200.0     2.1
##     age      bp     bgr      sc     sod     pot    hemo     pcv    wbcc    rbcc
##    90.0   180.0   490.0    76.0   163.0    47.0    17.8    54.0 26400.0     8.0

##     age      bp     bgr      bu     sod     pot    hemo     pcv    wbcc    rbcc
##     2.0    50.0    22.0     1.5     4.5     2.5     3.1     9.0  2200.0     2.1
##     age      bp     bgr      bu     sod     pot    hemo     pcv    wbcc    rbcc
##    90.0   180.0   490.0   391.0   163.0    47.0    17.8    54.0 26400.0     8.0

##     age      bp     bgr      bu      sc     pot    hemo     pcv    wbcc    rbcc
##     2.0    50.0    22.0     1.5     0.4     2.5     3.1     9.0  2200.0     2.1
##     age      bp     bgr      bu      sc     pot    hemo     pcv    wbcc    rbcc
##    90.0   180.0   490.0   391.0    76.0    47.0    17.8    54.0 26400.0     8.0

##     age      bp     bgr      bu      sc     sod    hemo     pcv    wbcc    rbcc
##     2.0    50.0    22.0     1.5     0.4     4.5     3.1     9.0  2200.0     2.1
##     age      bp     bgr      bu      sc     sod    hemo     pcv    wbcc    rbcc
##    90.0   180.0   490.0   391.0    76.0   163.0    17.8    54.0 26400.0     8.0

##     age      bp     bgr      bu      sc     sod     pot     pcv    wbcc    rbcc
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     9.0  2200.0     2.1
##     age      bp     bgr      bu      sc     sod     pot     pcv    wbcc    rbcc
##    90.0   180.0   490.0   391.0    76.0   163.0    47.0    54.0 26400.0     8.0

##     age      bp     bgr      bu      sc     sod     pot    hemo    wbcc    rbcc
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1  2200.0     2.1
##     age      bp     bgr      bu      sc     sod     pot    hemo    wbcc    rbcc
##    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8 26400.0     8.0

##   age    bp   bgr    bu    sc   sod   pot  hemo   pcv  rbcc   age    bp   bgr 
##   2.0  50.0  22.0   1.5   0.4   4.5   2.5   3.1   9.0   2.1  90.0 180.0 490.0 
##    bu    sc   sod   pot  hemo   pcv  rbcc 
## 391.0  76.0 163.0  47.0  17.8  54.0   8.0

##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0
##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc
##    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0 26400.0

##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0
##    rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv
##     2.1    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0
##    wbcc    rbcc 
## 26400.0     8.0

##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0
##    rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv
##     2.1    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0
##    wbcc    rbcc 
## 26400.0     8.0
##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0
##    rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv
##     2.1    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0
##    wbcc    rbcc
## 26400.0     8.0

##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0
##    rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv
##     2.1    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0
##    wbcc    rbcc
## 26400.0     8.0

##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0
##    rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv
##     2.1    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0
##    wbcc    rbcc
## 26400.0     8.0

##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0
##    rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv
##     2.1    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0
##    wbcc    rbcc
## 26400.0     8.0

## Time difference of 1.044969 secs
##
## Testing k = 5...

##      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc    rbcc
##    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0     2.1
##      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc    rbcc
##   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0 26400.0     8.0

##     age     bgr      bu      sc     sod     pot    hemo     pcv    wbcc    rbcc
##     2.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0     2.1
##     age     bgr      bu      sc     sod     pot    hemo     pcv    wbcc    rbcc
##    90.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0 26400.0     8.0

##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0
##    rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv
##     2.1    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0
##    wbcc    rbcc
## 26400.0     8.0

##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0
##    rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv
##     2.1    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0
##    wbcc    rbcc
## 26400.0     8.0

##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0
##    rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv
##     2.1    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0
##    wbcc    rbcc
## 26400.0     8.0

##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0
##    rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv
##     2.1    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0
##    wbcc    rbcc
## 26400.0     8.0

##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0
##    rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv
##     2.1    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0
##    wbcc    rbcc
## 26400.0     8.0

##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0
##    rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv
##     2.1    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0
##    wbcc    rbcc
## 26400.0     8.0

##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0
##    rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv
##     2.1    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0
##    wbcc    rbcc
## 26400.0     8.0

##     age      bp      bu      sc     sod     pot    hemo     pcv    wbcc    rbcc
##     2.0    50.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0     2.1
##     age      bp      bu      sc     sod     pot    hemo     pcv    wbcc    rbcc
##    90.0   180.0   391.0    76.0   163.0    47.0    17.8    54.0 26400.0     8.0

##     age      bp     bgr      sc     sod     pot    hemo     pcv    wbcc    rbcc
##     2.0    50.0    22.0     0.4     4.5     2.5     3.1     9.0  2200.0     2.1
##     age      bp     bgr      sc     sod     pot    hemo     pcv    wbcc    rbcc
##    90.0   180.0   490.0    76.0   163.0    47.0    17.8    54.0 26400.0     8.0

##     age      bp     bgr      bu     sod     pot    hemo     pcv    wbcc    rbcc
##     2.0    50.0    22.0     1.5     4.5     2.5     3.1     9.0  2200.0     2.1
##     age      bp     bgr      bu     sod     pot    hemo     pcv    wbcc    rbcc
##    90.0   180.0   490.0   391.0   163.0    47.0    17.8    54.0 26400.0     8.0
##     age      bp     bgr      bu      sc     pot    hemo     pcv    wbcc    rbcc
##     2.0    50.0    22.0     1.5     0.4     2.5     3.1     9.0  2200.0     2.1
##     age      bp     bgr      bu      sc     pot    hemo     pcv    wbcc    rbcc
##    90.0   180.0   490.0   391.0    76.0    47.0    17.8    54.0 26400.0     8.0

##     age      bp     bgr      bu      sc     sod    hemo     pcv    wbcc    rbcc
##     2.0    50.0    22.0     1.5     0.4     4.5     3.1     9.0  2200.0     2.1
##     age      bp     bgr      bu      sc     sod    hemo     pcv    wbcc    rbcc
##    90.0   180.0   490.0   391.0    76.0   163.0    17.8    54.0 26400.0     8.0

##     age      bp     bgr      bu      sc     sod     pot     pcv    wbcc    rbcc
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     9.0  2200.0     2.1
##     age      bp     bgr      bu      sc     sod     pot     pcv    wbcc    rbcc
##    90.0   180.0   490.0   391.0    76.0   163.0    47.0    54.0 26400.0     8.0

##     age      bp     bgr      bu      sc     sod     pot    hemo    wbcc    rbcc
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1  2200.0     2.1
##     age      bp     bgr      bu      sc     sod     pot    hemo    wbcc    rbcc
##    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8 26400.0     8.0

##   age    bp   bgr    bu    sc   sod   pot  hemo   pcv  rbcc   age    bp   bgr
##   2.0  50.0  22.0   1.5   0.4   4.5   2.5   3.1   9.0   2.1  90.0 180.0 490.0
##    bu    sc   sod   pot  hemo   pcv  rbcc
## 391.0  76.0 163.0  47.0  17.8  54.0   8.0

##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0
##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc
##    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0 26400.0

##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0
##    rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv
##     2.1    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0
##    wbcc    rbcc
## 26400.0     8.0

##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0
##    rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv
##     2.1    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0
##    wbcc    rbcc
## 26400.0     8.0

##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0
##    rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv
##     2.1    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0
##    wbcc    rbcc
## 26400.0     8.0
##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0
##    rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv
##     2.1    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0
##    wbcc    rbcc
## 26400.0     8.0

##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0
##    rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv
##     2.1    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0
##    wbcc    rbcc
## 26400.0     8.0

##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0
##    rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv
##     2.1    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0
##    wbcc    rbcc
## 26400.0     8.0

## Time difference of 1.037088 secs
## 
## Testing k = 10...

##      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc    rbcc
##    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0     2.1
##      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc    rbcc
##   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0 26400.0     8.0

##     age     bgr      bu      sc     sod     pot    hemo     pcv    wbcc    rbcc
##     2.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0     2.1
##     age     bgr      bu      sc     sod     pot    hemo     pcv    wbcc    rbcc
##    90.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0 26400.0     8.0

##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0
##    rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv
##     2.1    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0
##    wbcc    rbcc
## 26400.0     8.0

##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0
##    rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv
##     2.1    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0
##    wbcc    rbcc
## 26400.0     8.0

##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0
##    rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv
##     2.1    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0
##    wbcc    rbcc
## 26400.0     8.0

##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0
##    rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv
##     2.1    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0
##    wbcc    rbcc
## 26400.0     8.0

##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0
##    rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv
##     2.1    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0
##    wbcc    rbcc
## 26400.0     8.0

##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0
##    rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv
##     2.1    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0
##    wbcc    rbcc
## 26400.0     8.0

##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0
##    rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv
##     2.1    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0
##    wbcc    rbcc
## 26400.0     8.0

##     age      bp      bu      sc     sod     pot    hemo     pcv    wbcc    rbcc
##     2.0    50.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0     2.1
##     age      bp      bu      sc     sod     pot    hemo     pcv    wbcc    rbcc
##    90.0   180.0   391.0    76.0   163.0    47.0    17.8    54.0 26400.0     8.0

##     age      bp     bgr      sc     sod     pot    hemo     pcv    wbcc    rbcc
##     2.0    50.0    22.0     0.4     4.5     2.5     3.1     9.0  2200.0     2.1
##     age      bp     bgr      sc     sod     pot    hemo     pcv    wbcc    rbcc
##    90.0   180.0   490.0    76.0   163.0    47.0    17.8    54.0 26400.0     8.0

##     age      bp     bgr      bu     sod     pot    hemo     pcv    wbcc    rbcc
##     2.0    50.0    22.0     1.5     4.5     2.5     3.1     9.0  2200.0     2.1
##     age      bp     bgr      bu     sod     pot    hemo     pcv    wbcc    rbcc
##    90.0   180.0   490.0   391.0   163.0    47.0    17.8    54.0 26400.0     8.0

##     age      bp     bgr      bu      sc     pot    hemo     pcv    wbcc    rbcc
##     2.0    50.0    22.0     1.5     0.4     2.5     3.1     9.0  2200.0     2.1
##     age      bp     bgr      bu      sc     pot    hemo     pcv    wbcc    rbcc
##    90.0   180.0   490.0   391.0    76.0    47.0    17.8    54.0 26400.0     8.0
##     age      bp     bgr      bu      sc     sod    hemo     pcv    wbcc    rbcc
##     2.0    50.0    22.0     1.5     0.4     4.5     3.1     9.0  2200.0     2.1
##     age      bp     bgr      bu      sc     sod    hemo     pcv    wbcc    rbcc
##    90.0   180.0   490.0   391.0    76.0   163.0    17.8    54.0 26400.0     8.0
```

```
##     age      bp     bgr      bu      sc     sod     pot     pcv    wbcc    rbcc
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     9.0  2200.0     2.1
##     age      bp     bgr      bu      sc     sod     pot     pcv    wbcc    rbcc
##    90.0   180.0   490.0   391.0    76.0   163.0    47.0    54.0 26400.0     8.0
```

```
##     age      bp     bgr      bu      sc     sod     pot    hemo    wbcc    rbcc
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1  2200.0     2.1
##     age      bp     bgr      bu      sc     sod     pot    hemo    wbcc    rbcc
##    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8 26400.0     8.0
```

```
##   age    bp   bgr    bu    sc   sod   pot  hemo   pcv  rbcc   age    bp   bgr
##   2.0  50.0  22.0   1.5   0.4   4.5   2.5   3.1   9.0   2.1  90.0 180.0 490.0
##    bu    sc   sod   pot  hemo   pcv  rbcc
## 391.0  76.0 163.0  47.0  17.8  54.0   8.0
```

```
##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0
##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc
##    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0 26400.0
```

```
##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0
##    rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv
##     2.1    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0
##    wbcc    rbcc
## 26400.0     8.0
```

```
##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0
##    rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv
##     2.1    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0
##    wbcc    rbcc
## 26400.0     8.0
```

```
##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0
##    rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv
##     2.1    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0
##    wbcc    rbcc
## 26400.0     8.0
```

```
##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0
##    rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv
##     2.1    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0
##    wbcc    rbcc
## 26400.0     8.0
##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0
##    rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv
##     2.1    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0
##    wbcc    rbcc
## 26400.0     8.0

##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0
##    rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv
##     2.1    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0
##    wbcc    rbcc
## 26400.0     8.0

## Time difference of 1.016556 secs
##
## Testing k = 15...

##      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc    rbcc
##    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0     2.1
##      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc    rbcc
##   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0 26400.0     8.0

##     age     bgr      bu      sc     sod     pot    hemo     pcv    wbcc    rbcc
##     2.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0     2.1
##     age     bgr      bu      sc     sod     pot    hemo     pcv    wbcc    rbcc
##    90.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0 26400.0     8.0

##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0
##    rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv
##     2.1    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0
##    wbcc    rbcc
## 26400.0     8.0

##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0
##    rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv
##     2.1    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0
##    wbcc    rbcc
## 26400.0     8.0

##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0
##    rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv
##     2.1    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0
##    wbcc    rbcc
## 26400.0     8.0

##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0
##    rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv
##     2.1    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0
##    wbcc    rbcc
## 26400.0     8.0

##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0
##    rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv
##     2.1    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0
##    wbcc    rbcc
## 26400.0     8.0

##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0
##    rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv
##     2.1    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0
##    wbcc    rbcc
## 26400.0     8.0

##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0
##    rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv
##     2.1    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0
##    wbcc    rbcc
## 26400.0     8.0

##     age      bp      bu      sc     sod     pot    hemo     pcv    wbcc    rbcc
##     2.0    50.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0     2.1
##     age      bp      bu      sc     sod     pot    hemo     pcv    wbcc    rbcc
##    90.0   180.0   391.0    76.0   163.0    47.0    17.8    54.0 26400.0     8.0

##     age      bp     bgr      sc     sod     pot    hemo     pcv    wbcc    rbcc
##     2.0    50.0    22.0     0.4     4.5     2.5     3.1     9.0  2200.0     2.1
##     age      bp     bgr      sc     sod     pot    hemo     pcv    wbcc    rbcc
##    90.0   180.0   490.0    76.0   163.0    47.0    17.8    54.0 26400.0     8.0

##     age      bp     bgr      bu     sod     pot    hemo     pcv    wbcc    rbcc
##     2.0    50.0    22.0     1.5     4.5     2.5     3.1     9.0  2200.0     2.1
##     age      bp     bgr      bu     sod     pot    hemo     pcv    wbcc    rbcc
##    90.0   180.0   490.0   391.0   163.0    47.0    17.8    54.0 26400.0     8.0

##     age      bp     bgr      bu      sc     pot    hemo     pcv    wbcc    rbcc
##     2.0    50.0    22.0     1.5     0.4     2.5     3.1     9.0  2200.0     2.1
##     age      bp     bgr      bu      sc     pot    hemo     pcv    wbcc    rbcc
##    90.0   180.0   490.0   391.0    76.0    47.0    17.8    54.0 26400.0     8.0

##     age      bp     bgr      bu      sc     sod    hemo     pcv    wbcc    rbcc
##     2.0    50.0    22.0     1.5     0.4     4.5     3.1     9.0  2200.0     2.1
##     age      bp     bgr      bu      sc     sod    hemo     pcv    wbcc    rbcc
##    90.0   180.0   490.0   391.0    76.0   163.0    17.8    54.0 26400.0     8.0

##     age      bp     bgr      bu      sc     sod     pot     pcv    wbcc    rbcc
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     9.0  2200.0     2.1
##     age      bp     bgr      bu      sc     sod     pot     pcv    wbcc    rbcc
##    90.0   180.0   490.0   391.0    76.0   163.0    47.0    54.0 26400.0     8.0
##     age      bp     bgr      bu      sc     sod     pot    hemo    wbcc    rbcc
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1  2200.0     2.1
##     age      bp     bgr      bu      sc     sod     pot    hemo    wbcc    rbcc
##    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8 26400.0     8.0
```

```
##   age    bp   bgr    bu    sc   sod   pot  hemo   pcv  rbcc   age    bp   bgr
##   2.0  50.0  22.0   1.5   0.4   4.5   2.5   3.1   9.0   2.1  90.0 180.0 490.0
##    bu    sc   sod   pot  hemo   pcv  rbcc
## 391.0  76.0 163.0  47.0  17.8  54.0   8.0
```

```
##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0
##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc
##    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0 26400.0
```

```
##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0
##    rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv
##     2.1    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0
##    wbcc    rbcc
## 26400.0     8.0
```

```
##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0
##    rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv
##     2.1    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0
##    wbcc    rbcc
## 26400.0     8.0
```

```
##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0
##    rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv
##     2.1    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0
##    wbcc    rbcc
## 26400.0     8.0
```

```
##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0
##    rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv
##     2.1    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0
##    wbcc    rbcc
## 26400.0     8.0
```

```
##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0
##    rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv
##     2.1    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0
##    wbcc    rbcc
## 26400.0     8.0
```

```
##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0
##    rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv
##     2.1    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0
##    wbcc    rbcc
## 26400.0     8.0

## Time difference of 1.036865 secs
##
## Testing k = 20...

##      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc    rbcc
##    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0     2.1
##      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc    rbcc
##   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0 26400.0     8.0

##     age     bgr      bu      sc     sod     pot    hemo     pcv    wbcc    rbcc
##     2.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0     2.1
##     age     bgr      bu      sc     sod     pot    hemo     pcv    wbcc    rbcc
##    90.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0 26400.0     8.0

##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0
##    rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv
##     2.1    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0
##    wbcc    rbcc
## 26400.0     8.0

##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0
##    rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv
##     2.1    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0
##    wbcc    rbcc
## 26400.0     8.0

##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0
##    rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv
##     2.1    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0
##    wbcc    rbcc
## 26400.0     8.0

##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0
##    rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv
##     2.1    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0
##    wbcc    rbcc
## 26400.0     8.0

##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0
##    rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv
##     2.1    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0
##    wbcc    rbcc
## 26400.0     8.0
##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc 
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0 
##    rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv 
##     2.1    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0 
##    wbcc    rbcc 
## 26400.0     8.0 
```

```
##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc 
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0 
##    rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv 
##     2.1    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0 
##    wbcc    rbcc 
## 26400.0     8.0 
```

```
##     age      bp      bu      sc     sod     pot    hemo     pcv    wbcc    rbcc 
##     2.0    50.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0     2.1 
##     age      bp      bu      sc     sod     pot    hemo     pcv    wbcc    rbcc 
##    90.0   180.0   391.0    76.0   163.0    47.0    17.8    54.0 26400.0     8.0 
```

```
##     age      bp     bgr      sc     sod     pot    hemo     pcv    wbcc    rbcc 
##     2.0    50.0    22.0     0.4     4.5     2.5     3.1     9.0  2200.0     2.1 
##     age      bp     bgr      sc     sod     pot    hemo     pcv    wbcc    rbcc 
##    90.0   180.0   490.0    76.0   163.0    47.0    17.8    54.0 26400.0     8.0 
```

```
##     age      bp     bgr      bu     sod     pot    hemo     pcv    wbcc    rbcc 
##     2.0    50.0    22.0     1.5     4.5     2.5     3.1     9.0  2200.0     2.1 
##     age      bp     bgr      bu     sod     pot    hemo     pcv    wbcc    rbcc 
##    90.0   180.0   490.0   391.0   163.0    47.0    17.8    54.0 26400.0     8.0 
```

```
##     age      bp     bgr      bu      sc     pot    hemo     pcv    wbcc    rbcc 
##     2.0    50.0    22.0     1.5     0.4     2.5     3.1     9.0  2200.0     2.1 
##     age      bp     bgr      bu      sc     pot    hemo     pcv    wbcc    rbcc 
##    90.0   180.0   490.0   391.0    76.0    47.0    17.8    54.0 26400.0     8.0 
```

```
##     age      bp     bgr      bu      sc     sod    hemo     pcv    wbcc    rbcc 
##     2.0    50.0    22.0     1.5     0.4     4.5     3.1     9.0  2200.0     2.1 
##     age      bp     bgr      bu      sc     sod    hemo     pcv    wbcc    rbcc 
##    90.0   180.0   490.0   391.0    76.0   163.0    17.8    54.0 26400.0     8.0 
```

```
##     age      bp     bgr      bu      sc     sod     pot     pcv    wbcc    rbcc 
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     9.0  2200.0     2.1 
##     age      bp     bgr      bu      sc     sod     pot     pcv    wbcc    rbcc 
##    90.0   180.0   490.0   391.0    76.0   163.0    47.0    54.0 26400.0     8.0 
```

```
##     age      bp     bgr      bu      sc     sod     pot    hemo    wbcc    rbcc 
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1  2200.0     2.1 
##     age      bp     bgr      bu      sc     sod     pot    hemo    wbcc    rbcc 
##    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8 26400.0     8.0 
```

```
##   age    bp   bgr    bu    sc   sod   pot  hemo   pcv  rbcc   age    bp   bgr 
##   2.0  50.0  22.0   1.5   0.4   4.5   2.5   3.1   9.0   2.1  90.0 180.0 490.0 
##    bu    sc   sod   pot  hemo   pcv  rbcc 
## 391.0  76.0 163.0  47.0  17.8  54.0   8.0 
##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0
##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc
##    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0 26400.0

##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0
##    rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv
##     2.1    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0
##    wbcc    rbcc
## 26400.0     8.0

##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0
##    rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv
##     2.1    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0
##    wbcc    rbcc
## 26400.0     8.0

##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0
##    rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv
##     2.1    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0
##    wbcc    rbcc
## 26400.0     8.0

##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0
##    rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv
##     2.1    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0
##    wbcc    rbcc
## 26400.0     8.0

##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0
##    rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv
##     2.1    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0
##    wbcc    rbcc
## 26400.0     8.0

##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0
##    rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv
##     2.1    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0
##    wbcc    rbcc
## 26400.0     8.0

## Time difference of 1.011072 secs
##
## Testing k = 25...
##      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc    rbcc
##    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0     2.1
##      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc    rbcc
##   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0 26400.0     8.0
```

```
##     age     bgr      bu      sc     sod     pot    hemo     pcv    wbcc    rbcc
##     2.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0     2.1
##     age     bgr      bu      sc     sod     pot    hemo     pcv    wbcc    rbcc
##    90.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0 26400.0     8.0
```

```
##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0
##    rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv
##     2.1    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0
##    wbcc    rbcc
## 26400.0     8.0
```

```
##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0
##    rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv
##     2.1    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0
##    wbcc    rbcc
## 26400.0     8.0
```

```
##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0
##    rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv
##     2.1    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0
##    wbcc    rbcc
## 26400.0     8.0
```

```
##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0
##    rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv
##     2.1    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0
##    wbcc    rbcc
## 26400.0     8.0
```

```
##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0
##    rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv
##     2.1    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0
##    wbcc    rbcc
## 26400.0     8.0
```

```
##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0
##    rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv
##     2.1    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0
##    wbcc    rbcc
## 26400.0     8.0
##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc 
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0 
##    rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv 
##     2.1    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0 
##    wbcc    rbcc 
## 26400.0     8.0 

##     age      bp      bu      sc     sod     pot    hemo     pcv    wbcc    rbcc 
##     2.0    50.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0     2.1 
##     age      bp      bu      sc     sod     pot    hemo     pcv    wbcc    rbcc 
##    90.0   180.0   391.0    76.0   163.0    47.0    17.8    54.0 26400.0     8.0 

##     age      bp     bgr      sc     sod     pot    hemo     pcv    wbcc    rbcc 
##     2.0    50.0    22.0     0.4     4.5     2.5     3.1     9.0  2200.0     2.1 
##     age      bp     bgr      sc     sod     pot    hemo     pcv    wbcc    rbcc 
##    90.0   180.0   490.0    76.0   163.0    47.0    17.8    54.0 26400.0     8.0 

##     age      bp     bgr      bu     sod     pot    hemo     pcv    wbcc    rbcc 
##     2.0    50.0    22.0     1.5     4.5     2.5     3.1     9.0  2200.0     2.1 
##     age      bp     bgr      bu     sod     pot    hemo     pcv    wbcc    rbcc 
##    90.0   180.0   490.0   391.0   163.0    47.0    17.8    54.0 26400.0     8.0 

##     age      bp     bgr      bu      sc     pot    hemo     pcv    wbcc    rbcc 
##     2.0    50.0    22.0     1.5     0.4     2.5     3.1     9.0  2200.0     2.1 
##     age      bp     bgr      bu      sc     pot    hemo     pcv    wbcc    rbcc 
##    90.0   180.0   490.0   391.0    76.0    47.0    17.8    54.0 26400.0     8.0 

##     age      bp     bgr      bu      sc     sod    hemo     pcv    wbcc    rbcc 
##     2.0    50.0    22.0     1.5     0.4     4.5     3.1     9.0  2200.0     2.1 
##     age      bp     bgr      bu      sc     sod    hemo     pcv    wbcc    rbcc 
##    90.0   180.0   490.0   391.0    76.0   163.0    17.8    54.0 26400.0     8.0 

##     age      bp     bgr      bu      sc     sod     pot     pcv    wbcc    rbcc 
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     9.0  2200.0     2.1 
##     age      bp     bgr      bu      sc     sod     pot     pcv    wbcc    rbcc 
##    90.0   180.0   490.0   391.0    76.0   163.0    47.0    54.0 26400.0     8.0 

##     age      bp     bgr      bu      sc     sod     pot    hemo    wbcc    rbcc 
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1  2200.0     2.1 
##     age      bp     bgr      bu      sc     sod     pot    hemo    wbcc    rbcc 
##    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8 26400.0     8.0 

##   age    bp   bgr    bu    sc   sod   pot  hemo   pcv  rbcc   age    bp   bgr 
##   2.0  50.0  22.0   1.5   0.4   4.5   2.5   3.1   9.0   2.1  90.0 180.0 490.0 
##    bu    sc   sod   pot  hemo   pcv  rbcc 
## 391.0  76.0 163.0  47.0  17.8  54.0   8.0 

##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc 
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0 
##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc 
##    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0 26400.0
##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0
##    rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv
##     2.1    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0
##    wbcc    rbcc
## 26400.0     8.0

##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0
##    rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv
##     2.1    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0
##    wbcc    rbcc
## 26400.0     8.0

##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0
##    rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv
##     2.1    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0
##    wbcc    rbcc
## 26400.0     8.0

##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0
##    rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv
##     2.1    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0
##    wbcc    rbcc
## 26400.0     8.0

##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0
##    rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv
##     2.1    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0
##    wbcc    rbcc
## 26400.0     8.0

##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc
##     2.0    50.0    22.0     1.5     0.4     4.5     2.5     3.1     9.0  2200.0
##    rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv
##     2.1    90.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0
##    wbcc    rbcc
## 26400.0     8.0
```

```
## Time difference of 1.004764 secs
```

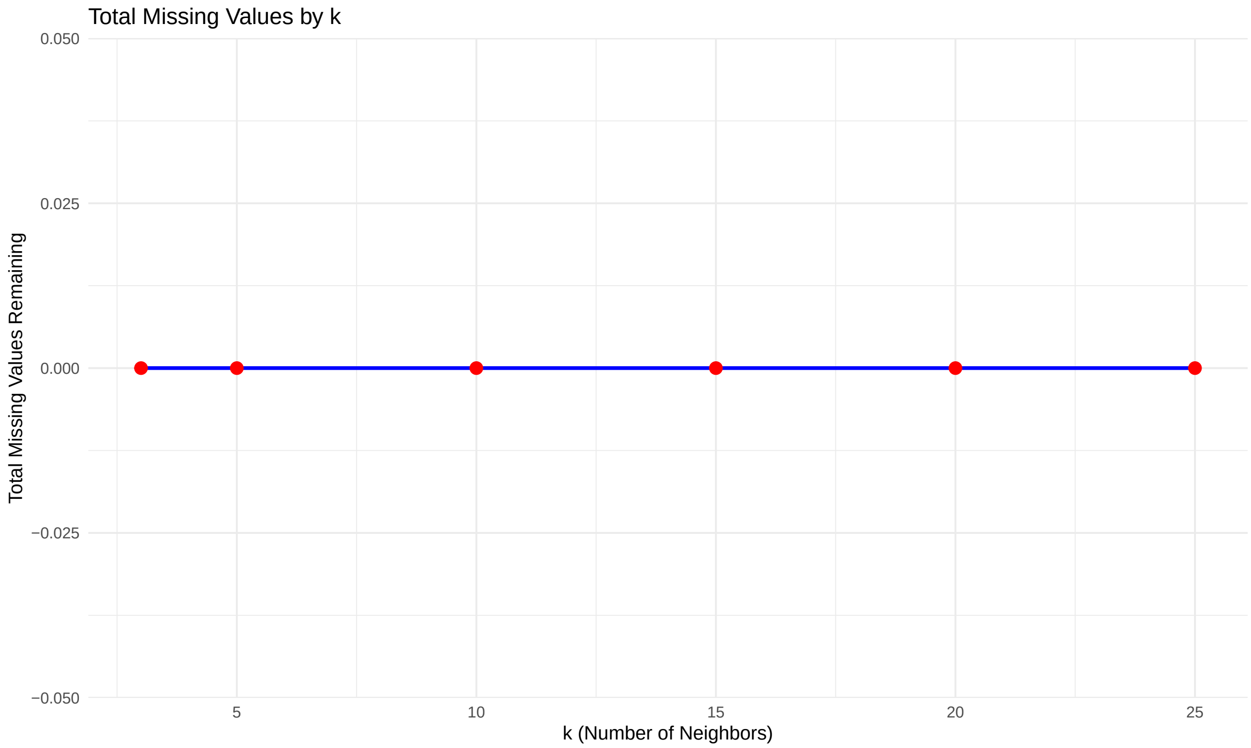
```r
# Create comparison table
comparison_table <- do.call(rbind, lapply(imputation_results, as.data.frame))
print("kNN Imputation Performance by k Value:")
```

```
## [1] "kNN Imputation Performance by k Value:"
```

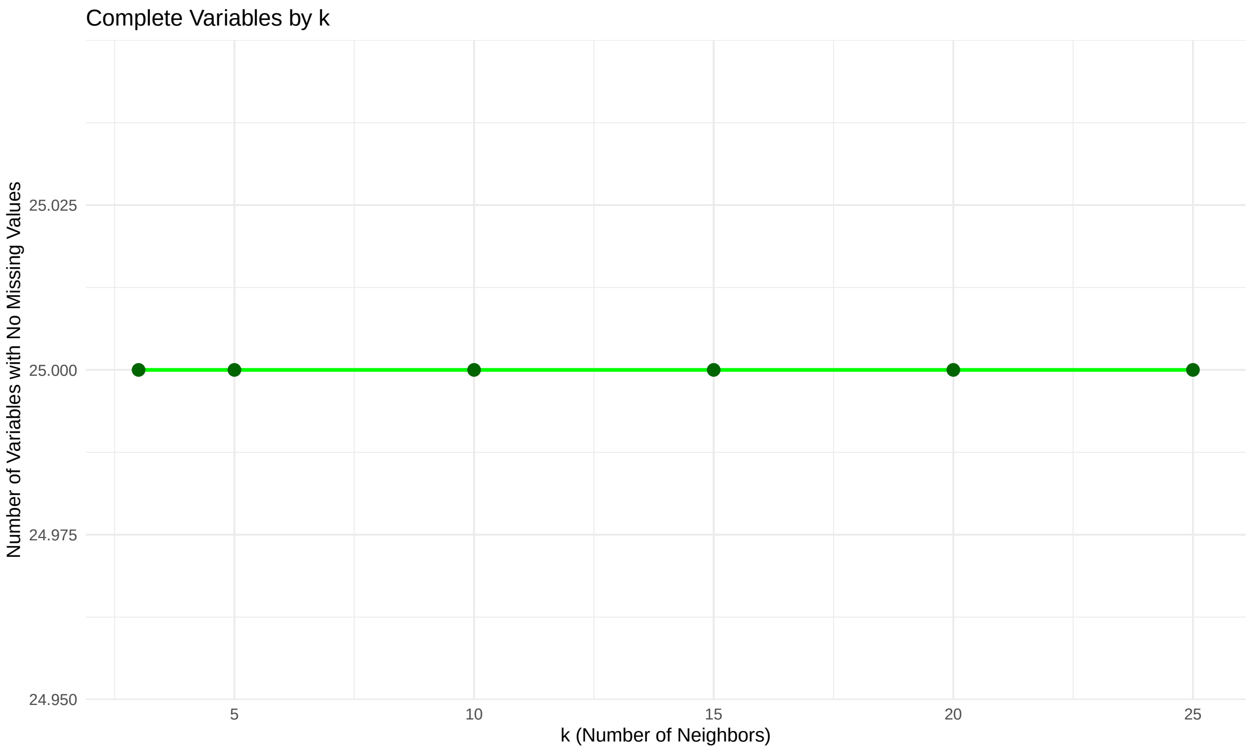
```r
print(comparison_table)
```

```
##      k missing_after variables_imputed
## 3   3             0                25
## 5   5             0                25
## 10 10             0                25
## 15 15             0                25
## 20 20             0                25
## 25 25             0                25
```

```r
# Visualize comparison
ggplot(comparison_table, aes(x = k, y = missing_after)) +
  geom_line(color = "blue", size = 1) +
  geom_point(color = "red", size = 3) +
  labs(title = "Total Missing Values by k",
       x = "k (Number of Neighbors)",
       y = "Total Missing Values Remaining") +
  theme_minimal()
```



```r
ggplot(comparison_table, aes(x = k, y = variables_imputed)) +
  geom_line(color = "green", size = 1) +
  geom_point(color = "darkgreen", size = 3) +
  labs(title = "Complete Variables by k",
       x = "k (Number of Neighbors)",
       y = "Number of Variables with No Missing Values") +
  theme_minimal()
```



```r
# Analyze distance metrics for mixed data
cat("\n=== DISTANCE METRIC ANALYSIS ===\n")
```

Distance Metric Analysis

```
##
## === DISTANCE METRIC ANALYSIS ===
```

```r
# Gower's distance calculation
calculate_gower <- function(data) {
  # Convert all to numeric for distance calculation
  # In practice, VIM uses specialized distance for mixed data

  # For demonstration, let's calculate on a subset
  # subset_data <- data %>%
  #   select(age, bp, sc, hemo, htn, dm) %>%
  #   mutate(
  #     htn_num = as.numeric(factor(htn)),
  #     dm_num = as.numeric(factor(dm))
  #   ) %>%
  #   select(-htn, -dm)

  subset_data <- data

  # Handle missing values by imputing with mean for this analysis
  subset_data_imputed <- subset_data
  for(col in names(subset_data)) {
    subset_data_imputed[[col]][is.na(subset_data[[col]])] <-
      mean(subset_data[[col]], na.rm = TRUE)
  }

  # Calculate Euclidean distance (simplified)
  dist_matrix <- dist(subset_data_imputed, method = "euclidean")

  return(dist_matrix)
}

# Calculate distances on original data
cat("Calculating distance matrix on key variables...\n")
```

```
## Calculating distance matrix on key variables...
```

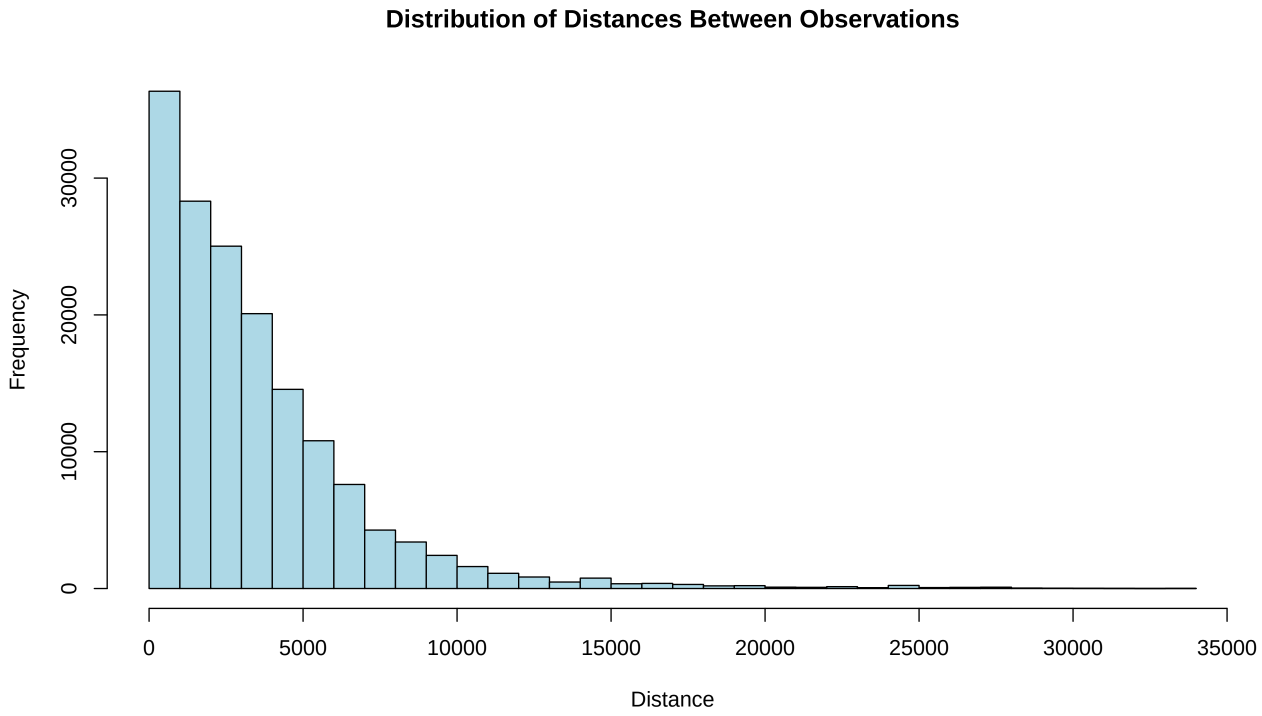
```r
distance_matrix <- calculate_gower(ckd_for_imputation)

# Analyze distance distribution
distance_values <- as.vector(as.matrix(distance_matrix))
distance_summary <- summary(distance_values)

cat("Distance Distribution Summary:\n")
```

```
## Distance Distribution Summary:
```

```r
print(distance_summary)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0    1103    2574    3444    4677   33318
```

```r
# Visualize distance distribution
hist(distance_values,
     main = "Distribution of Distances Between Observations",
     xlab = "Distance",
     ylab = "Frequency",
     col = "lightblue",
     breaks = 30)
```



#### Validation of kNN Imputation

```r
# Validate kNN imputation by creating artificial missing values
cat("\n=== VALIDATION OF kNN IMPUTATION ===\n")
```

```
##
## === VALIDATION OF kNN IMPUTATION ===
```

```r
# Create a validation dataset with known values
set.seed(42)

# Select 10% of values to make missing (MAR - Missing at Random)
n_to_missing <- round(nrow(ckd_for_imputation) * ncol(ckd_for_imputation) * 0.10)
missing_indices <- sample(1:(nrow(ckd_for_imputation) * ncol(ckd_for_imputation)),
                          n_to_missing)

# Store original values
original_values <- matrix(NA, nrow = nrow(ckd_for_imputation),
                          ncol = ncol(ckd_for_imputation))
for(idx in missing_indices) {
  row_idx <- ((idx - 1) %% nrow(ckd_for_imputation)) + 1
  col_idx <- floor((idx - 1) / nrow(ckd_for_imputation)) + 1
  original_values[row_idx, col_idx] <- ckd_for_imputation[row_idx, col_idx]
  ckd_for_imputation[row_idx, col_idx] <- NA
}

# Perform kNN imputation on validation data
cat("Imputing validation dataset...\n")
```

```
## Imputing validation dataset...
```

```r
validation_imputed <- VIM::kNN(
  data = ckd_for_imputation,
  k = 10,
  imp_var = FALSE,
  trace = FALSE,
  useImputedDist = TRUE
)
```

```
##       bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc    rbcc
##     50.0    22.0    10.0     0.4     4.5     2.5     3.1     9.0  2200.0     2.1
##       bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc    rbcc
##    180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0 26400.0     6.5
##      age     bgr      bu      sc     sod     pot    hemo     pcv    wbcc    rbcc
##      3.0    22.0    10.0     0.4     4.5     2.5     3.1     9.0  2200.0     2.1
##      age     bgr      bu      sc     sod     pot    hemo     pcv    wbcc    rbcc
##     83.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0 26400.0     6.5
##      age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc
##      3.0    50.0    22.0    10.0     0.4     4.5     2.5     3.1     9.0  2200.0
##     rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv
##      2.1    83.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0
##     wbcc    rbcc
## 26400.0     6.5
##      age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc
##      3.0    50.0    22.0    10.0     0.4     4.5     2.5     3.1     9.0  2200.0
##     rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv
##      2.1    83.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0
##     wbcc    rbcc
## 26400.0     6.5
##      age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc
##      3.0    50.0    22.0    10.0     0.4     4.5     2.5     3.1     9.0  2200.0
##     rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv
##      2.1    83.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0
##     wbcc    rbcc
## 26400.0     6.5
##      age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc
##      3.0    50.0    22.0    10.0     0.4     4.5     2.5     3.1     9.0  2200.0
##     rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv
##      2.1    83.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0
##     wbcc    rbcc
## 26400.0     6.5
##      age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc
##      3.0    50.0    22.0    10.0     0.4     4.5     2.5     3.1     9.0  2200.0
##     rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv
##      2.1    83.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0
##     wbcc    rbcc
## 26400.0     6.5
##      age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc
##      3.0    50.0    22.0    10.0     0.4     4.5     2.5     3.1     9.0  2200.0
##     rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv
##      2.1    83.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0
##     wbcc    rbcc
## 26400.0     6.5
##      age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc
##     3.0    50.0    22.0    10.0     0.4     4.5     2.5     3.1     9.0  2200.0
##    rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv
##     2.1    83.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0
##    wbcc    rbcc
## 26400.0     6.5
##     age      bp      bu      sc     sod     pot    hemo     pcv    wbcc    rbcc
##     3.0    50.0    10.0     0.4     4.5     2.5     3.1     9.0  2200.0     2.1
##     age      bp      bu      sc     sod     pot    hemo     pcv    wbcc    rbcc
##    83.0   180.0   391.0    76.0   163.0    47.0    17.8    54.0 26400.0     6.5
##     age      bp     bgr      sc     sod     pot    hemo     pcv    wbcc    rbcc
##     3.0    50.0    22.0     0.4     4.5     2.5     3.1     9.0  2200.0     2.1
##     age      bp     bgr      sc     sod     pot    hemo     pcv    wbcc    rbcc
##    83.0   180.0   490.0    76.0   163.0    47.0    17.8    54.0 26400.0     6.5
##     age      bp     bgr      bu     sod     pot    hemo     pcv    wbcc    rbcc
##     3.0    50.0    22.0    10.0     4.5     2.5     3.1     9.0  2200.0     2.1
##     age      bp     bgr      bu     sod     pot    hemo     pcv    wbcc    rbcc
##    83.0   180.0   490.0   391.0   163.0    47.0    17.8    54.0 26400.0     6.5
##     age      bp     bgr      bu      sc     pot    hemo     pcv    wbcc    rbcc
##     3.0    50.0    22.0    10.0     0.4     2.5     3.1     9.0  2200.0     2.1
##     age      bp     bgr      bu      sc     pot    hemo     pcv    wbcc    rbcc
##    83.0   180.0   490.0   391.0    76.0    47.0    17.8    54.0 26400.0     6.5
##     age      bp     bgr      bu      sc     sod    hemo     pcv    wbcc    rbcc
##     3.0    50.0    22.0    10.0     0.4     4.5     3.1     9.0  2200.0     2.1
##     age      bp     bgr      bu      sc     sod    hemo     pcv    wbcc    rbcc
##    83.0   180.0   490.0   391.0    76.0   163.0    17.8    54.0 26400.0     6.5
##     age      bp     bgr      bu      sc     sod     pot     pcv    wbcc    rbcc
##     3.0    50.0    22.0    10.0     0.4     4.5     2.5     9.0  2200.0     2.1
##     age      bp     bgr      bu      sc     sod     pot     pcv    wbcc    rbcc
##    83.0   180.0   490.0   391.0    76.0   163.0    47.0    54.0 26400.0     6.5
##     age      bp     bgr      bu      sc     sod     pot    hemo    wbcc    rbcc
##     3.0    50.0    22.0    10.0     0.4     4.5     2.5     3.1  2200.0     2.1
##     age      bp     bgr      bu      sc     sod     pot    hemo    wbcc    rbcc
##    83.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8 26400.0     6.5
##   age    bp   bgr    bu    sc   sod   pot  hemo   pcv  rbcc   age    bp   bgr
##   3.0  50.0  22.0  10.0   0.4   4.5   2.5   3.1   9.0   2.1  83.0 180.0 490.0
##    bu    sc   sod   pot  hemo   pcv  rbcc
## 391.0  76.0 163.0  47.0  17.8  54.0   6.5
##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc
##     3.0    50.0    22.0    10.0     0.4     4.5     2.5     3.1     9.0  2200.0
##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc
##    83.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0 26400.0
##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc
##     3.0    50.0    22.0    10.0     0.4     4.5     2.5     3.1     9.0  2200.0
##    rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv
##     2.1    83.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0
##    wbcc    rbcc
## 26400.0     6.5
##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc
##     3.0    50.0    22.0    10.0     0.4     4.5     2.5     3.1     9.0  2200.0
##    rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv
##     2.1    83.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0
##    wbcc    rbcc
## 26400.0     6.5
##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc
##     3.0    50.0    22.0    10.0     0.4     4.5     2.5     3.1     9.0  2200.0
##    rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv
##     2.1    83.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0
##    wbcc    rbcc
## 26400.0     6.5
##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc
##     3.0    50.0    22.0    10.0     0.4     4.5     2.5     3.1     9.0  2200.0
##    rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv
##     2.1    83.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0
##    wbcc    rbcc
## 26400.0     6.5
##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc
##     3.0    50.0    22.0    10.0     0.4     4.5     2.5     3.1     9.0  2200.0
##    rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv
##     2.1    83.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0
##    wbcc    rbcc
## 26400.0     6.5
##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc
##     3.0    50.0    22.0    10.0     0.4     4.5     2.5     3.1     9.0  2200.0
##    rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv
##     2.1    83.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0
##    wbcc    rbcc
## 26400.0     6.5
##     age      bp     bgr      bu      sc     sod     pot    hemo     pcv    wbcc
##     3.0    50.0    22.0    10.0     0.4     4.5     2.5     3.1     9.0  2200.0
##    rbcc     age      bp     bgr      bu      sc     sod     pot    hemo     pcv
##     2.1    83.0   180.0   490.0   391.0    76.0   163.0    47.0    17.8    54.0
##    wbcc    rbcc
## 26400.0     6.5
```

```r
# Calculate imputation accuracy
imputation_accuracy <- data.frame()
for(col in 1:ncol(ckd_for_imputation)) {
  col_name <- colnames(ckd_for_imputation)[col]
  original_col <- original_values[, col]
  imputed_col <- validation_imputed[[col_name]]

  # Find indices where we artificially created missing values
  missing_in_col <- which(!is.na(original_values[, col]))

  if(length(missing_in_col) > 0 && is.numeric(original_col)) {
    # For numeric variables: RMSE
    rmse <- sqrt(mean((original_col[missing_in_col] -
                       imputed_col[missing_in_col])^2, na.rm = TRUE))

    imputation_accuracy <- rbind(imputation_accuracy,
                                 data.frame(Variable = col_name,
                                            Type = "Numeric",
                                            RMSE = rmse))
  } else if(length(missing_in_col) > 0 && is.factor(original_col)) {
    # For categorical variables: Accuracy
    accuracy <- mean(original_col[missing_in_col] ==
                     imputed_col[missing_in_col], na.rm = TRUE)
    imputation_accuracy <- rbind(imputation_accuracy,
                                 data.frame(Variable = col_name,
                                            Type = "Categorical",
                                            RMSE = 1 - accuracy))
  }
}

print("Imputation Accuracy (Lower RMSE is better):")
```

```
## [1] "Imputation Accuracy (Lower RMSE is better):"
```

```r
print(imputation_accuracy %>% arrange(RMSE))
```

```
##    Variable    Type         RMSE
## 1       pot Numeric    0.8878937
## 2      rbcc Numeric    0.9187229
## 3      hemo Numeric    1.6020820
## 4       pcv Numeric    4.2436928
## 5       sod Numeric    5.5965551
## 6        sc Numeric    6.4052379
## 7        bp Numeric   12.2793017
## 8       age Numeric   16.1033123
## 9        bu Numeric   45.0362222
## 10      bgr Numeric   60.3632457
## 11     wbcc Numeric 2871.1309370
## 12       sg Numeric          NaN
## 13       al Numeric          NaN
## 14       su Numeric          NaN
## 15      rbc Numeric          NaN
## 16       pc Numeric          NaN
## 17      pcc Numeric          NaN
## 18       ba Numeric          NaN
## 19      htn Numeric          NaN
## 20       dm Numeric          NaN
## 21      cad Numeric          NaN
## 22    appet Numeric          NaN
## 23       pe Numeric          NaN
## 24      ane Numeric          NaN
## 25   status Numeric          NaN
```

```r
# Overall accuracy
overall_accuracy <- mean(imputation_accuracy$RMSE)
cat(sprintf("\nOverall Imputation Error (RMSE): %.4f\n", overall_accuracy))
```

```
##
## Overall Imputation Error (RMSE): NaN
```

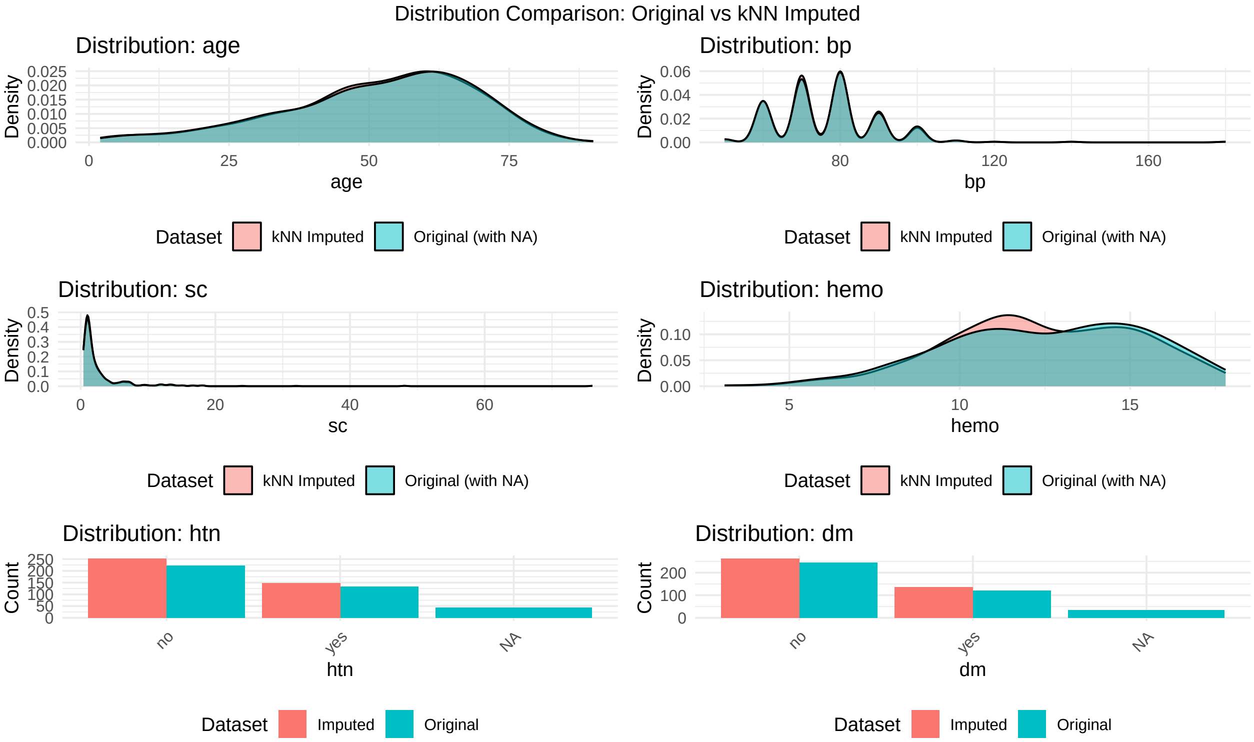
```r


# Compare distributions before and after imputation
cat("\n=== DISTRIBUTION PRESERVATION ANALYSIS ===\n")
```

Distribution Preservation Analysis

```
##
## === DISTRIBUTION PRESERVATION ANALYSIS ===
```

```
# Select key variables for distribution comparison
key_vars <- c("age", "bp", "sc", "hemo", "htn", "dm")

# Create comparison plots
plot_list <- list()

for(var in key_vars) {
  if(is.numeric(ckd_for_imputation[[var]])) {
    # For numeric variables: density plot
    df_compare <- data.frame(
      Value = c(ckd_for_imputation[[var]], ckd_imputed[[var]]),
      Dataset = rep(c("Original (with NA)", "kNN Imputed"),
                    each = nrow(ckd_for_imputation))
    )

    p <- ggplot(df_compare, aes(x = Value, fill = Dataset)) +
      geom_density(alpha = 0.5) +
      labs(title = paste("Distribution:", var),
           x = var,
           y = "Density") +
      theme_minimal() +
      theme(legend.position = "bottom")

  } else {
    # For categorical variables: bar plot
    orig_counts <- table(ckd_for_imputation[[var]], useNA = "always")
    imp_counts <- table(ckd_imputed[[var]])

    df_compare <- data.frame(
      Category = c(names(orig_counts), names(imp_counts)),
      Count = c(as.numeric(orig_counts), as.numeric(imp_counts)),
      Dataset = rep(c("Original", "Imputed"),
                    c(length(orig_counts), length(imp_counts)))
    )

    p <- ggplot(df_compare, aes(x = Category, y = Count, fill = Dataset)) +
      geom_bar(stat = "identity", position = "dodge") +
      labs(title = paste("Distribution:", var),
           x = var,
           y = "Count") +
      theme_minimal() +
      theme(axis.text.x = element_text(angle = 45, hjust = 1),
            legend.position = "bottom")
  }

  plot_list[[var]] <- p
}

# Arrange plots
```

```
grid.arrange(grobs = plot_list, ncol = 2,
             top = "Distribution Comparison: Original vs kNN Imputed")
```



```
# Statistical comparison of distributions
cat("\nStatistical Comparison of Distributions (Kolmogorov-Smirnov Test):\n")
```

```
##
## Statistical Comparison of Distributions (Kolmogorov-Smirnov Test):
```

```
for(var in key_vars) {
  if(is.numeric(ckd_for_imputation[[var]]) &&
     is.numeric(ckd_imputed[[var]])) {

    # Remove NAs for KS test
    orig_clean <- na.omit(ckd_for_imputation[[var]])
    imp_clean <- ckd_imputed[[var]]

    if(length(orig_clean) > 0 && length(imp_clean) > 0) {
      ks_test <- ks.test(orig_clean, imp_clean)
      cat(sprintf("%s: D = %.3f, p = %.4f\n",
                  var, ks_test$statistic, ks_test$p.value))
    }
  }
}
```

```
## age: D = 0.015, p = 1.0000
```

```
## bp: D = 0.010, p = 1.0000
## sc: D = 0.021, p = 1.0000
```

```
## hemo: D = 0.053, p = 0.6961
```

```r
cat("\n=== COMPARISON: kNN vs MICE IMPUTATION ===\n")
```

Comparison with MICE Imputation

```
##
## === COMPARISON: kNN vs MICE IMPUTATION ===
```

```r
library(dplyr)
library(ggplot2)

# Compare missing values
cat("\nMissing Values Comparison:\n")
```

```
##
## Missing Values Comparison:
```

```r
comparison <- data.frame(
  Variable = names(ckd_for_imputation %>% select(-status)),
  Original_NA = colSums(is.na(ckd_clean %>% select(-status))),
  kNN_NA = colSums(is.na(ckd_imputed %>% select(-status))),
  MICE_NA = colSums(is.na(ckd_mice_imputed %>% select(-status)))
)

  print(comparison %>% filter(Original_NA > 0))
```

```
##      Variable Original_NA kNN_NA MICE_NA
## age       age           9      0       0
## bp         bp          12      0       0
## sg         sg          47      0       0
## al         al          46      0       0
## su         su          49      0       0
## rbc       rbc         152      0       0
## pc         pc          65      0       0
## pcc       pcc           4      0       0
## ba         ba           4      0       0
## bgr       bgr          44      0       0
## bu         bu          19      0       0
## sc         sc          17      0       0
## sod       sod          87      0       0
## pot       pot          88      0       0
## hemo     hemo          52      0       0
## pcv       pcv          71      0       0
## wbcc     wbcc         106      0       0
## rbcc     rbcc         131      0       0
## htn       htn           2      0       0
## dm          dm             2      0       0
## cad        cad             2      0       0
## appet    appet             1      0       0
## pe          pe             1      0       0
## ane        ane             1      0       0
```

```
# Compare distributions for key variables
cat("\nDistribution Comparison (Key Variables):\n")
```

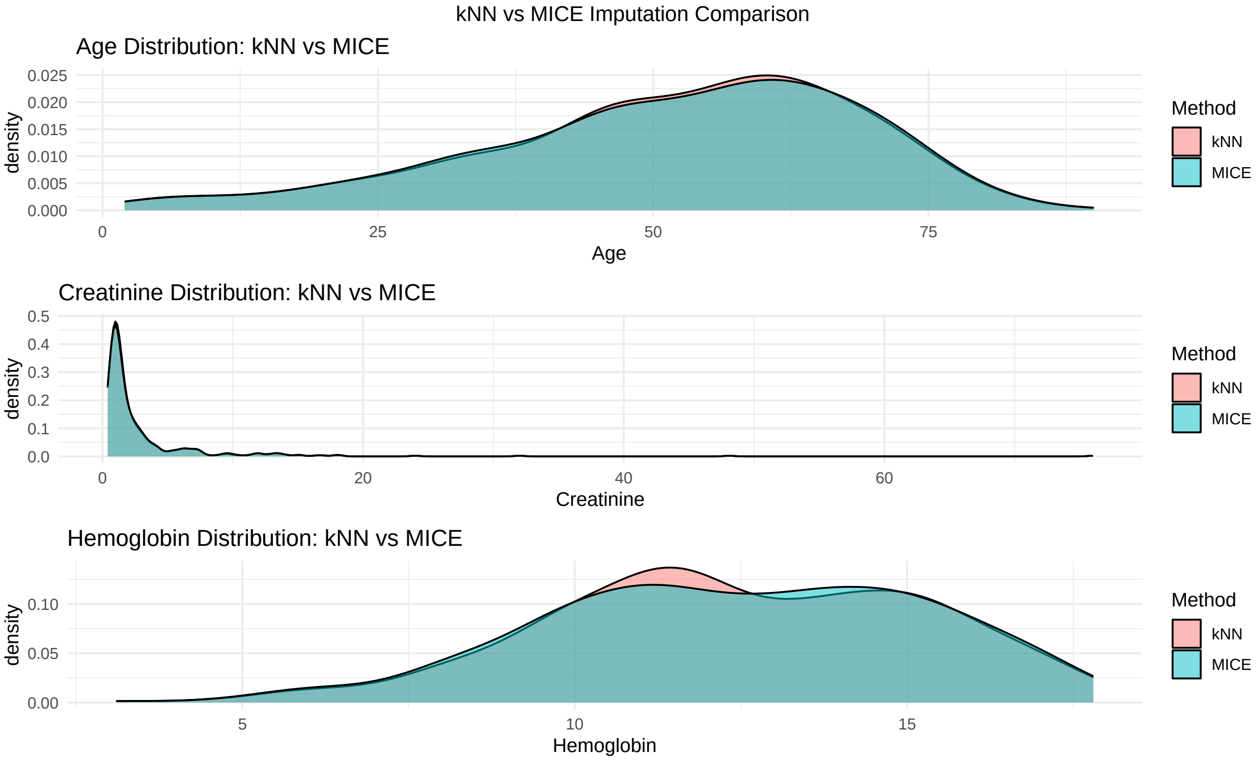
```
##
## Distribution Comparison (Key Variables):
```

```
for(var in c("age", "sc", "hemo")) {
  if(is.numeric(ckd_for_imputation[[var]])) {
    # Calculate means
    orig_mean <- mean(ckd_clean[[var]], na.rm = TRUE)
    knn_mean <- mean(ckd_imputed[[var]], na.rm = TRUE)
    mice_mean <- mean(ckd_mice_imputed[[var]], na.rm = TRUE)

    cat(sprintf("\n%s:\n", var))
    cat(sprintf("  Original (with NA): mean = %.2f\n", orig_mean))
    cat(sprintf("  kNN imputed:        mean = %.2f (diff: %.2f)\n",
                knn_mean, knn_mean - orig_mean))
    cat(sprintf("  MICE imputed:       mean = %.2f (diff: %.2f)\n",
                mice_mean, mice_mean - orig_mean))
  }
}
```

```
##
## age:
##   Original (with NA): mean = 51.48
##   kNN imputed:        mean = 51.53 (diff: 0.05)
##   MICE imputed:       mean = 51.50 (diff: 0.02)
##
## sc:
##   Original (with NA): mean = 3.07
##   kNN imputed:        mean = 3.00 (diff: -0.08)
##   MICE imputed:       mean = 3.02 (diff: -0.05)
##
## hemo:
##   Original (with NA): mean = 12.53
##   kNN imputed:        mean = 12.44 (diff: -0.09)
##   MICE imputed:       mean = 12.45 (diff: -0.08)
```

```
# Create comparison plot
comparison_plot_data <- data.frame(
  Method = rep(c("kNN", "MICE"), each = nrow(ckd_imputed)),
  Age = c(ckd_imputed$age, ckd_mice_imputed$age),
  Creatinine = c(ckd_imputed$sc, ckd_mice_imputed$sc),
  Hemoglobin = c(ckd_imputed$hemo, ckd_mice_imputed$hemo)
)
```

```
p1 <- ggplot(comparison_plot_data, aes(x = Age, fill = Method)) +
  geom_density(alpha = 0.5) +
  labs(title = "Age Distribution: kNN vs MICE") +
  theme_minimal()

p2 <- ggplot(comparison_plot_data, aes(x = Creatinine, fill = Method)) +
  geom_density(alpha = 0.5) +
  labs(title = "Creatinine Distribution: kNN vs MICE") +
  theme_minimal()

p3 <- ggplot(comparison_plot_data, aes(x = Hemoglobin, fill = Method)) +
  geom_density(alpha = 0.5) +
  labs(title = "Hemoglobin Distribution: kNN vs MICE") +
  theme_minimal()

grid.arrange(p1, p2, p3, ncol = 1,
             top = "kNN vs MICE Imputation Comparison")
```



Multiple imputation with MissForest `missForest` is a nonparametric imputation method for basically any kind of tabular data. It handles mixed types (numeric + categorical), nonlinear relations, interactions, and even high dimensionality ((p n)). For each variable with missingness, it fits a random forest on the observed part and predicts the missing part, iterating until a stopping rule is met (or maxiter says "enough").

By default, missForest() now uses the ranger backend for speed and multithreading. For legacy/compatibility, you can select the classic randomForest backend via backend = "randomForest". The out-of-bag (OOB) error from the backend is transformed into an imputation error estimate — one for numeric variables (NRMSE) and one for factors (PFC).

```
library(dplyr)
library(missForest)

# Start from the existing subset:
# nhanes_sub <- NHANES |> select(ID, Age, Gender, BMI, BPSysAve, Diabetes)

# 1) Keep only model-relevant columns (drop pure identifier)
# 2) Convert character variables to factors (missForest expects factors, not raw character)
# 3) Coerce to base data.frame to avoid tibble-related method dispatch issues

library(dplyr)
library(missForest)

# Prepare the data for missForest
# Select relevant predictor variables and the target 'status'
# Convert character/factor variables appropriately (missForest handles factors)

set.seed(17) # For reproducibility
mf_fit <- missForest(
  ckd_for_imputation,
  ntree   = 50,    # Number of trees in the random forest
  maxiter = 5,      # Outer iterations
  verbose = TRUE    # Set to TRUE to see progress
)
```

```
##   missForest iteration 1 in progress...done!
##     estimated error(s): 0.3617892 0.1919184
##     difference(s): 0.004145646 0.03946429
##     time: 0.351 seconds
##
##   missForest iteration 2 in progress...done!
##     estimated error(s): 0.3665411 0.1829868
##     difference(s): 0.0008872063 0.01964286
##     time: 0.334 seconds
##
##   missForest iteration 3 in progress...done!
##     estimated error(s): 0.3591449 0.1830853
##     difference(s): 0.0007512638 0.015
##     time: 0.333 seconds
##
##   missForest iteration 4 in progress...done!
##     estimated error(s): 0.3638142 0.1818486
##     difference(s): 0.0008043499 0.01535714
##     time: 0.333 seconds
```

```
# Extract the imputed data and error metrics
ckd_missForest_imputed <- mf_fit$ximp
mf_oob_error <- mf_fit$OOBerror

# Check the Out-of-Bag (OOB) imputation error
print(paste("Normalized Root Mean Squared Error (NRMSE) for numeric variables:", mf_fit[["OOBerror"]][[
```

```
## [1] "Normalized Root Mean Squared Error (NRMSE) for numeric variables: 0.35914485370588"
```

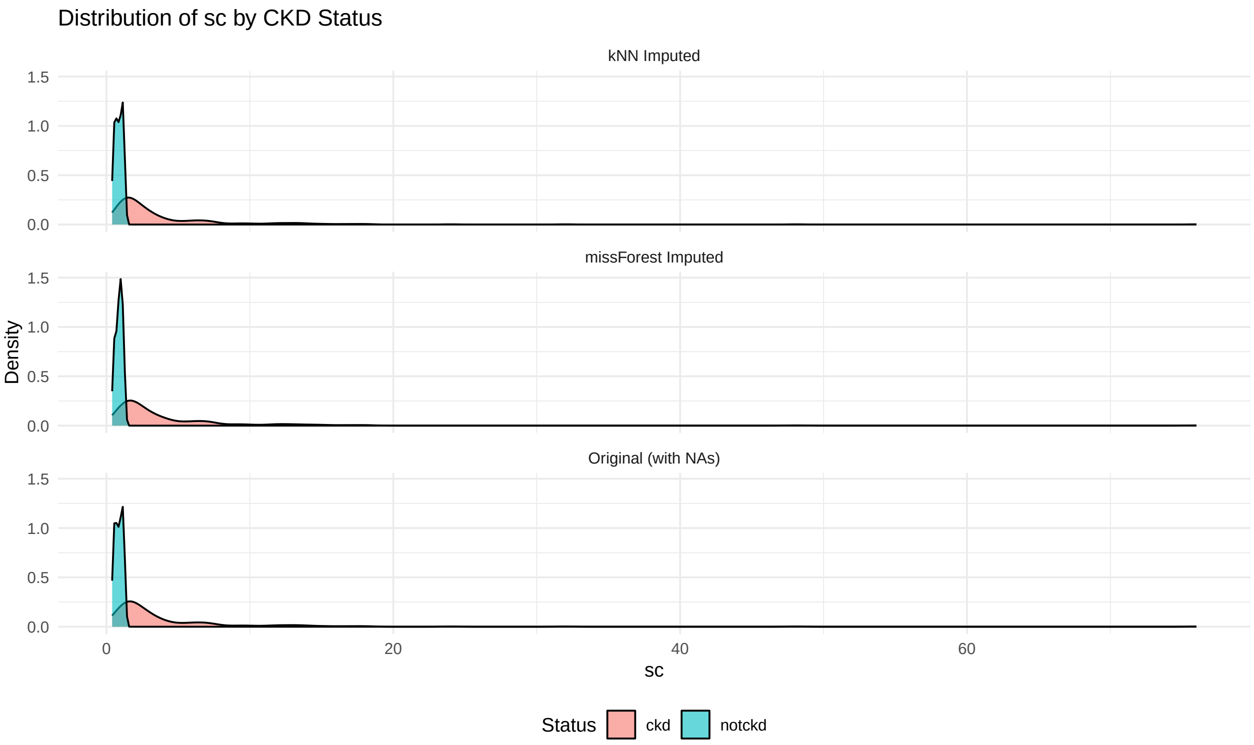
```r
print(paste("Proportion of Falsely Classified entries (PFC) for categorical variables:", mf_fit[["OOBer
```

```
## [1] "Proportion of Falsely Classified entries (PFC) for categorical variables: 0.183085325017773"
```

```r
# Verify no missing values remain
sum(is.na(ckd_missForest_imputed))
```

```
## [1] 0
```

```r
library(ggplot2)
library(patchwork)

# Function to plot distributions by class for a key variable
plot_dist_comparison <- function(orig_data,
                                 knn_data,
                                 mf_data,
                                 var_name,
                                 method_names = c("Original (with NAs)", "kNN Imputed", "missForest Impu

  # Combine data for plotting
  plot_data <- rbind(
    data.frame(Value = orig_data[[var_name]], Status = orig_data$status, Method = method_names[1]),
    data.frame(Value = knn_data[[var_name]], Status = knn_data$status, Method = method_names[2]),
    data.frame(Value = mf_data[[var_name]], Status = mf_data$status, Method = method_names[3])
  )



  # Create density plot for numeric variables
  p <- ggplot(plot_data, aes(x = Value, fill = Status)) +
    geom_density(alpha = 0.6) +
    facet_wrap(~ Method, ncol = 1) +
    labs(title = paste("Distribution of", var_name, "by CKD Status"),
         x = var_name,
         y = "Density") +
    theme_minimal() +
    theme(legend.position = "bottom")

  return(p)
}

# Example: Compare distributions for Serum Creatinine (sc)
# You will need your kNN-imputed dataset (ckd_knn_imputed) and missForest dataset
p_sc <- plot_dist_comparison(ckd_clean, ckd_knn_basic, ckd_missForest_imputed, "sc")
print(p_sc)
```



Correlation betweeen features

```
kidney_data <- ckd_clean

categorical_vars <- c("rbc", "pc", "pcc", "ba", "htn", "dm", "cad", "appet", "pe", "ane", "status")
kidney_data[categorical_vars] <- lapply(kidney_data[categorical_vars], as.factor)

# Convert ordinal/numeric-looking categorical variables to numeric
kidney_data$sg <- as.numeric(as.character(kidney_data$sg))
kidney_data$al <- as.numeric(as.character(kidney_data$al))
kidney_data$su <- as.numeric(as.character(kidney_data$su))

# Create binary target variable (1 for ckd, 0 for notckd)
kidney_data$target <- ifelse(kidney_data$status == "ckd", 1, 0)

# Remove original status column if needed
kidney_data$status <- NULL
```

```
# Select only numeric columns for correlation analysis
numeric_vars <- sapply(kidney_data, is.numeric)
numeric_data <- kidney_data[, numeric_vars]

# Calculate correlation matrix with target
cor_matrix <- cor(numeric_data, use = "complete.obs")
target_cor <- cor_matrix[,"target"]

# Sort correlations by absolute value
sorted_cor <- sort(abs(target_cor[names(target_cor) != "target"]), decreasing = TRUE)
```

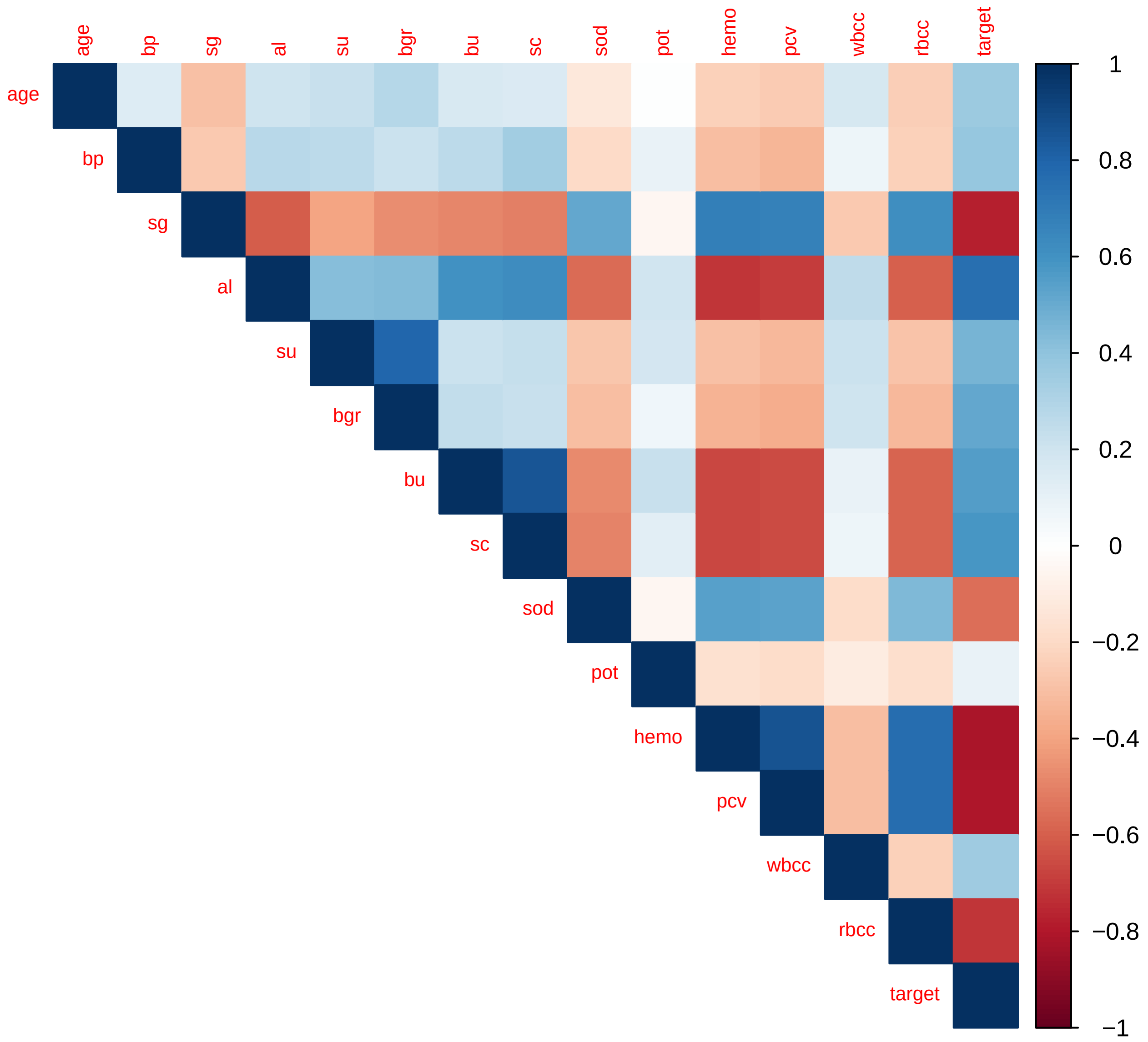
```r
# Print top correlated features
cat("Top features correlated with CKD:\n")
```

```
## Top features correlated with CKD:
```

```r
print(sorted_cor)
```

```
##      hemo       pcv        sg        al      rbcc        sc       sod        bu
## 0.8151133 0.8057846 0.7788644 0.7597193 0.7121427 0.5880926 0.5572077 0.5542620
##       bgr        su        bp       age      wbcc       pot
## 0.5155330 0.4647734 0.3871122 0.3608664 0.3549493 0.0989747
```

```r
# Visualize correlations
library(corrplot)
corrplot(cor_matrix, method = "color", type = "upper",
         tl.cex = 0.7, number.cex = 0.7)
```



```r
# Select features above threshold (e.g., |cor| > 0.3)
threshold <- 0.3
important_numeric <- names(sorted_cor[abs(sorted_cor) > threshold])
cat("\nImportant numeric features (|cor| >", threshold, "):\n")
```

```
##
## Important numeric features (|cor| > 0.3 ):
```

```r
print(important_numeric)
```

```
##  [1] "hemo" "pcv"  "sg"   "al"   "rbcc" "sc"   "sod"  "bu"   "bgr"  "su"
## [11] "bp"   "age"  "wbcc"
```

```r
# Function to calculate chi-square statistic between categorical variables
chi_square_test <- function(data, categorical_vars, target_var = "target") {
  results <- data.frame()

  for(var in categorical_vars) {
    if(var != target_var) {
      # Create contingency table
      contingency_table <- table(data[[var]], data[[target_var]])

      # Perform chi-square test
      chi_test <- chisq.test(contingency_table)

      # Calculate Cramér's V (effect size)
      n <- sum(contingency_table)
      k <- min(dim(contingency_table))
      cramers_v <- sqrt(chi_test$statistic / (n * (k - 1)))

      results <- rbind(results, data.frame(
        Feature = var,
        Chi_Square = chi_test$statistic,
        p_value = chi_test$p.value,
        Cramers_V = cramers_v
      ))
    }
  }

  return(results[order(-results$Cramers_V), ])
}

# Test categorical variables (excluding target)
cat_vars <- categorical_vars[categorical_vars != "status"]
chi_results <- chi_square_test(kidney_data, cat_vars)

# Print results
print(chi_results)
```

```
##            Feature Chi_Square      p_value Cramers_V
## X-squared4     htn  135.47743 2.596025e-31 0.5834343
## X-squared5      dm  121.26010 3.351777e-28 0.5519725
## X-squared      rbc   73.58383 9.645164e-18 0.5447100
## X-squared1      pc   69.22963 8.764136e-17 0.4545939
## X-squared7   appet   59.52152 1.209639e-14 0.3862341
## X-squared8      pe   53.99972 2.005182e-13 0.3678826
## X-squared9     ane   40.23169 2.255601e-10 0.3175394
## X-squared2     pcc   25.69478 3.999079e-07 0.2547269
## X-squared6     cad   20.30180 6.613818e-06 0.2258530
## X-squared3      ba   11.97828 5.382418e-04 0.1739201
```

```r
# Select significant features (p < 0.05 and Cramér's V > 0.1)
significant_cat <- chi_results$Feature[chi_results$p_value < 0.05 &
                                       chi_results$Cramers_V > 0.1]
cat("\nSignificant categorical features:\n")
```

```
##
## Significant categorical features:
```

```r
print(significant_cat)
```

```
##  [1] "htn"   "dm"    "rbc"   "pc"    "appet" "pe"    "ane"   "pcc"   "cad"
## [10] "ba"
```

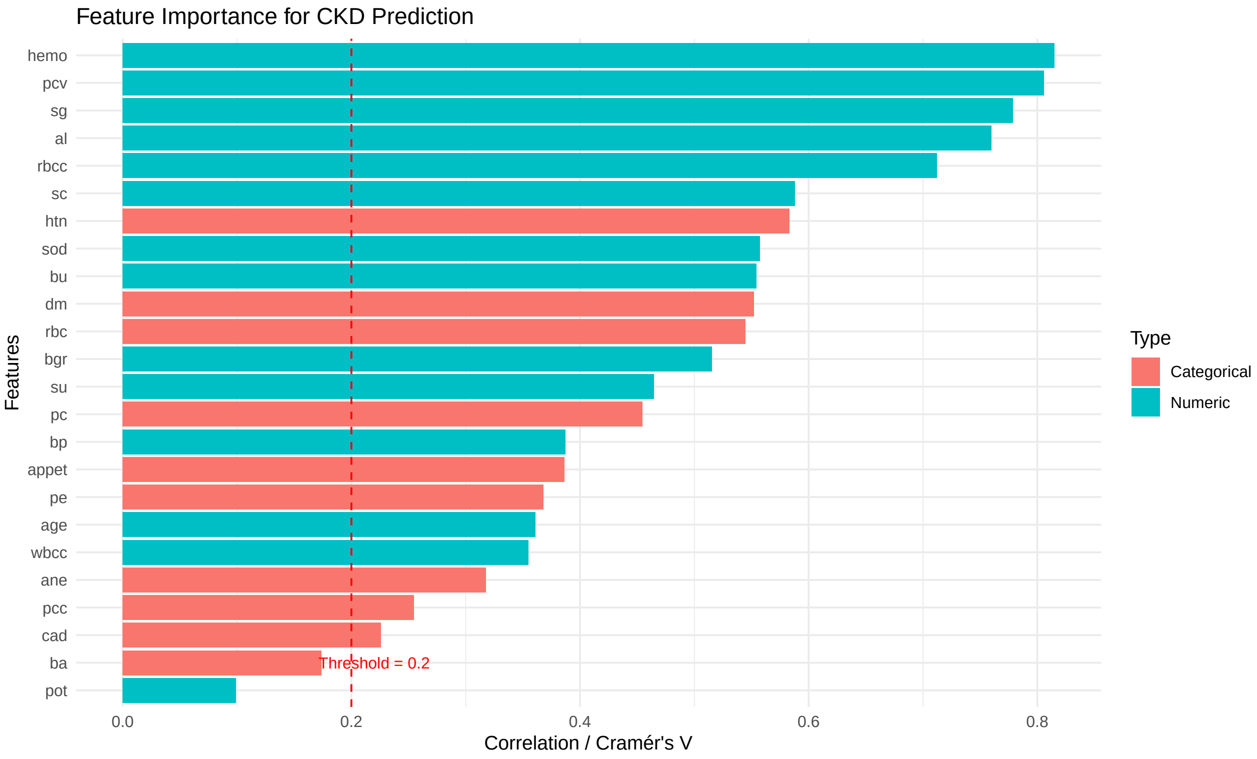
```r
library(caret)

feature_selection_pipeline <- function(data, target_name = "target") {

  # Separate numeric and categorical features
  numeric_features <- names(data)[sapply(data, is.numeric)]
  numeric_features <- numeric_features[numeric_features != target_name]

  categorical_features <- names(data)[sapply(data, is.factor)]
  categorical_features <- categorical_features[categorical_features != target_name]

  # 1. Numeric features: Correlation with target
  cor_values <- sapply(numeric_features, function(x) {
    cor(data[[x]], data[[target_name]], use = "complete.obs")
  })

  # Select numeric features with |cor| > 0.25
  selected_numeric <- names(cor_values)[abs(cor_values) > 0.25]

  # 2. Categorical features: Chi-square test
  chi_results <- data.frame()
  for(cat_var in categorical_features) {
    tbl <- table(data[[cat_var]], data[[target_name]])
    chi_test <- chisq.test(tbl)
    n <- sum(tbl)
    k <- min(dim(tbl))
    cramers_v <- sqrt(chi_test$statistic / (n * (k - 1)))

    chi_results <- rbind(chi_results,
                         data.frame(Feature = cat_var,
                                    p_value = chi_test$p.value,
                                    Cramers_V = cramers_v))
  }

  # Select categorical features with p < 0.05 and Cramér's V > 0.15
  selected_categorical <- chi_results$Feature[
    chi_results$p_value < 0.05 & chi_results$Cramers_V > 0.15
  ]

  # 3. Check for multicollinearity among selected numeric features
  if(length(selected_numeric) > 1) {
    numeric_cor <- cor(data[, selected_numeric], use = "complete.obs")
    high_cor <- findCorrelation(numeric_cor, cutoff = 0.8)
    if(length(high_cor) > 0) {
      selected_numeric <- selected_numeric[-high_cor]
    }
  }

  # Combine selected features
  all_selected <- c(selected_numeric, as.character(selected_categorical))

  cat("Selected Features:\n")
  cat("-----------------\n")
  cat("Numeric (", length(selected_numeric), "): ",
      paste(selected_numeric, collapse = ", "), "\n\n")
  cat("Categorical (", length(selected_categorical), "): ",
      paste(selected_categorical, collapse = ", "), "\n")

  return(list(
    numeric = selected_numeric,
    categorical = as.character(selected_categorical),
    all = all_selected
  ))
}

# Run the pipeline
selected_features <- feature_selection_pipeline(kidney_data)
```

```
## Selected Features:
## -----------------
## Numeric ( 9 ):  bp, sg, al, su, bgr, bu, sod, hemo, rbcc
##
## Categorical ( 10 ):  rbc, pc, pcc, ba, htn, dm, cad, appet, pe, ane
```

```
# Create dataset with only important features
important_vars <- selected_features$all
final_dataset <- kidney_data[, c(important_vars, "target")]
```

```
library(ggplot2)

# Create a feature importance plot
importance_df <- data.frame(
  Feature = names(sorted_cor),
```

```r
  Correlation = as.numeric(sorted_cor),
  Type = "Numeric"
)

# Add categorical features with Cramér's V
cat_importance <- data.frame(
  Feature = chi_results$Feature,
  Correlation = chi_results$Cramers_V,
  Type = "Categorical"
)

importance_df <- rbind(importance_df, cat_importance)

# Plot
ggplot(importance_df, aes(x = reorder(Feature, Correlation),
                          y = Correlation, fill = Type)) +
  geom_bar(stat = "identity") +
  coord_flip() +
  labs(title = "Feature Importance for CKD Prediction",
       x = "Features",
       y = "Correlation / Cramér's V") +
  theme_minimal() +
  geom_hline(yintercept = 0.2, linetype = "dashed", color = "red") +
  annotate("text", x = 2, y = 0.22, label = "Threshold = 0.2",
         color = "red", size = 3)
```



```r
names(importance_df)
```

```
## [1] "Feature"     "Correlation" "Type"
```

```r
selected_cols <- c(importance_df %>% filter(Correlation >= 0.20) %>% pull(Feature))
selected_cols
```

```
##  [1] "hemo"  "pcv"   "sg"    "al"    "rbcc"  "sc"    "sod"   "bu"    "bgr"  
## [10] "su"    "bp"    "age"   "wbcc"  "htn"   "dm"    "rbc"   "pc"    "appet"
## [19] "pe"    "ane"   "pcc"   "cad"
```

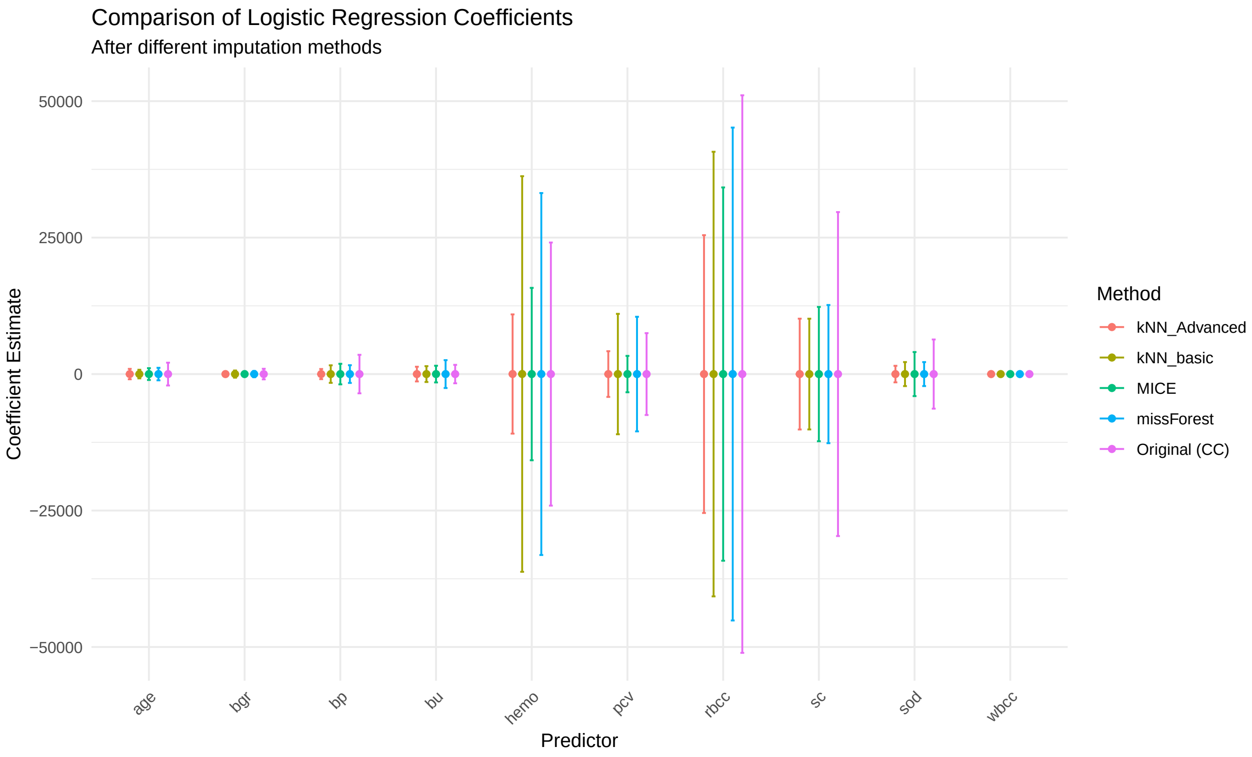
```r
typeof(selected_cols)
```

```
## [1] "character"
```

```r
model_formula <- reformulate(selected_cols, response = "status")
model_formula
```

```
## status ~ hemo + pcv + sg + al + rbcc + sc + sod + bu + bgr + 
##     su + bp + age + wbcc + htn + dm + rbc + pc + appet + pe + 
##     ane + pcc + cad
```

```r
# Build logistic regression models on datasets from different imputation methods
model_original <- glm(model_formula,
                      data = ckd_clean, family = binomial, na.action = na.omit)
model_knn_basic <- glm(model_formula,
                 data = ckd_knn_basic, family = binomial)
model_knn_advanced <- glm(model_formula,
                 data = ckd_knn_advanced, family = binomial)
model_mice <- glm(model_formula,
                 data = ckd_mice_imputed, family = binomial)
model_missForest <- glm(model_formula,
                        data = ckd_missForest_imputed, family = binomial)

# Compare model coefficients
library(broom)
model_summary <- bind_rows(
  tidy(model_original) %>% mutate(Method = "Original (CC)"),
  tidy(model_knn_basic) %>% mutate(Method = "kNN_basic"),
  tidy(model_knn_advanced) %>% mutate(Method = "kNN_Advanced"),
  tidy(model_mice) %>% mutate(Method = "MICE"),
  tidy(model_missForest) %>% mutate(Method = "missForest")
)

# Plot coefficient comparisons for key predictors
coef_plot <- model_summary %>%
  filter(term %in% selected_cols) %>%
  ggplot(aes(x = term, y = estimate, color = Method)) +
  geom_point(position = position_dodge(width = 0.5)) +
  geom_errorbar(aes(ymin = estimate - std.error, ymax = estimate + std.error),
                width = 0.2, position = position_dodge(width = 0.5)) +
  labs(title = "Comparison of Logistic Regression Coefficients",
       subtitle = "After different imputation methods",
       x = "Predictor",
       y = "Coefficient Estimate") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

print(coef_plot)
```



```
model_knn_advanced
```

```
##
## Call:  glm(formula = model_formula, family = binomial, data = ckd_knn_advanced)
##
## Coefficients:
##   (Intercept)           hemo            pcv         sg1.01         sg1.015
##    -2.290e+02      1.333e+01      2.524e+00     -5.368e+01      -6.050e+01
##       sg1.02        sg1.025            al1            al2             al3
##     2.811e+00      1.645e+01     -4.669e+01      4.604e+01       3.283e+00
##          al4            al5           rbcc             sc             sod
##     1.924e+01      2.272e+02     -8.700e-02      2.571e-01      -3.266e-01
##           bu            bgr            su1            su2             su3
##    -1.166e-01     -1.329e-01     -5.677e+01     -4.823e+01      -3.345e+01
##          su4            su5             bp            age            wbcc
##    -2.469e+01      5.396e+01      8.797e-02      5.768e-01       1.146e-05
##        htnno           dmno    rbcabnormal     pcabnormal       appetpoor
##     4.618e+01     -2.069e+01      3.375e+00     -2.334e+01      -5.651e+01
##         peno          aneno  pccnotpresent          cadno
##     4.516e+01     -4.717e+01      1.049e+00     -3.335e+01
##
## Degrees of Freedom: 399 Total (i.e. Null);  366 Residual
## Null Deviance:       529.3
## Residual Deviance: 3.701e-08     AIC: 68
```

```
model_original
```

```
##
## Call:  glm(formula = model_formula, family = binomial, data = ckd_clean,
##     na.action = na.omit)
##
## Coefficients:
##   (Intercept)           hemo            pcv         sg1.01        sg1.015
##    -2.563e+01      1.999e-02      1.358e-02      4.652e+01      4.687e+01
##        sg1.02        sg1.025            al1            al2            al3
##     4.807e+01      4.798e+01     -5.023e+01     -5.114e+01     -5.085e+01
##           al4           rbcc             sc            sod             bu
##    -5.060e+01      9.931e-02      6.602e-02      1.777e-02     -1.800e-03
##           bgr            su1            su2            su3            su4
##     1.540e-03      1.083e-01      2.222e-01      3.631e-01      4.268e-02
##           su5             bp            age           wbcc          htnno
##     1.837e+00      1.104e-03     -1.150e-03     -1.017e-04      1.542e+00
##          dmno    rbcabnormal     pcabnormal      appetpoor           peno
##     4.558e-01     -2.631e-01      3.188e-01      6.761e-01     -4.653e-01
##         aneno  pccnotpresent          cadno
##    -6.505e-02     -1.610e-01     -5.365e-01
##
## Degrees of Freedom: 157 Total (i.e. Null);  125 Residual
##   (242 observations deleted due to missingness)
## Null Deviance:       185
## Residual Deviance: 9.715e-10     AIC: 66
```

```
table(ckd_clean$status)
```

```
##
##    ckd notckd
##    250    150
```

```
prop.table(table(ckd_clean$status))
```

```
##
##    ckd notckd
##  0.625  0.375
```

```
table(na.omit(ckd_clean)$status)
```

```
##
##    ckd notckd
##     43    115
```

```
prop.table(table(na.omit(ckd_clean)$status))
```

```
##
##       ckd    notckd
## 0.2721519 0.7278481
```

## 5. Imputation Comparison table

```r
# Load required libraries
library(dplyr)
library(tidyr)
library(kableExtra)
library(ggplot2)

# Identify missing value positions in original data
# Get indices of all missing values
missing_indices <- which(is.na(ckd_clean), arr.ind = TRUE)

# Convert to a dataframe for easier handling
missing_df <- data.frame(
  row = missing_indices[, 1],
  column = colnames(ckd_clean)[missing_indices[, 2]],
  variable = colnames(ckd_clean)[missing_indices[, 2]]
)

# Remove duplicates and sort
missing_df <- missing_df %>%
  distinct(row, column, .keep_all = TRUE) %>%
  arrange(row, column)

# Extract original patient characteristics (non-missing values) for context
extract_context <- function(row_idx, col_names) {
  patient_data <- ckd_clean[row_idx, ]
  # Get non-missing values for context
  context_vars <- append(selected_cols, "status")
  context_vals <- sapply(context_vars, function(v) {
    if(v %in% names(patient_data)) {
      val <- patient_data[[v]]
      if(is.na(val)) "NA"
      else as.character(val)
    } else NA
  })
  return(context_vals)
}

# Get context for each missing value
context_data <- t(sapply(missing_df$row, extract_context,
                         col_names = names(ckd_clean)))
colnames(context_data) <- paste("Context", colnames(context_data))

# Extract imputed values from different methods
# Make sure you have these imputed datasets from previous code:
# ckd_knn_imputed (from kNN imputation)
# ckd_missForest_imputed (from missForest imputation)

# Function to extract values from imputed datasets
extract_imputed_values <- function(row_idx, col_name, dataset) {
  if(col_name %in% colnames(dataset)) {
    return(dataset[row_idx, col_name])
  }
  return(NA)
}

# Extract values for each missing position
missing_df$kNN_value <- mapply(extract_imputed_values,
                               missing_df$row, missing_df$column,
                               MoreArgs = list(dataset = ckd_knn_basic))

missing_df$missForest_value <- mapply(extract_imputed_values,
                                     missing_df$row, missing_df$column,
                                     MoreArgs = list(dataset = ckd_missForest_imputed))

# For MICE imputation (if you ran it earlier)
if(exists("ckd_mice_imputed")) {
  missing_df$MICE_value <- mapply(extract_imputed_values,
                                 missing_df$row, missing_df$column,
                                 MoreArgs = list(dataset = ckd_mice_imputed))
}

# Combine with context
comparison_table <- cbind(missing_df, context_data)

# Reorder columns for better readability
comparison_table <- comparison_table %>%
  select(row, variable, starts_with("Context"),
         kNN_value, missForest_value, everything())

# Display first 20 rows of the comparison
head(comparison_table, 20) %>%
  kable("latex", caption = "Imputation Comparison at Missing Positions") %>%
  kable_styling(bootstrap_options = c("striped", "hover", "condensed")) %>%
  landscape()
```

Table 1: Imputation Comparison at Missing Positions

row	variable	Context hemo	Context pcv	Context sg	Context al	Context rbcc	Context sc	Context sod	Context bu	Context bgr	Context su	C
1	pot	15.4	44	1.02	1	5.2	1.2	NA	36	121	0	8
1	rbc	15.4	44	1.02	1	5.2	1.2	NA	36	121	0	8
1	sod	15.4	44	1.02	1	5.2	1.2	NA	36	121	0	8
2	bgr	11.3	38	1.02	4	NA	0.8	NA	18	NA	0	5
2	pot	11.3	38	1.02	4	NA	0.8	NA	18	NA	0	5
2	rbc	11.3	38	1.02	4	NA	0.8	NA	18	NA	0	5
2	rbcc	11.3	38	1.02	4	NA	0.8	NA	18	NA	0	5
2	sod	11.3	38	1.02	4	NA	0.8	NA	18	NA	0	5
3	pot	9.6	31	1.01	2	NA	1.8	NA	53	423	3	8
3	rbcc	9.6	31	1.01	2	NA	1.8	NA	53	423	3	8
3	sod	9.6	31	1.01	2	NA	1.8	NA	53	423	3	8
5	pot	11.6	35	1.01	2	4.6	1.4	NA	26	106	0	8
5	sod	11.6	35	1.01	2	4.6	1.4	NA	26	106	0	8
6	pc	12.2	39	1.015	3	4.4	1.1	142	25	74	0	9
6	rbc	12.2	39	1.015	3	4.4	1.1	142	25	74	0	9
7	rbc	12.4	36	1.01	0	NA	24	104	54	100	0	7
7	rbcc	12.4	36	1.01	0	NA	24	104	54	100	0	7
7	wbcc	12.4	36	1.01	0	NA	24	104	54	100	0	7
8	bp	12.4	44	1.015	2	5	1.1	NA	31	410	4	N
8	pot	12.4	44	1.015	2	5	1.1	NA	31	410	4	N

```r
setwd('/Volumes/HHD_iMac_Storage/URV/SCIENTIFIC_PROGRAMMING/FINAL/SP-Final-Project')
write_csv(ckd_knn_advanced, "data/processed/dataset_knn_imputed.csv",
          progress = show_progress())
```

```r
setwd('/Volumes/HHD_iMac_Storage/URV/SCIENTIFIC_PROGRAMMING/FINAL/SP-Final-Project')
write_csv(ckd_missForest_imputed, "data/processed/dataset_missForest_imputed.csv",
          progress = show_progress())
```