DM3-IMPUTE EPCV

Mohamed Moukhtar / KHALIL

2023-04-22

Exploring the Data

```
library(tidyverse)
```

```
## Warning in Sys.timezone(): unable to identify current timezone 'C':
## please set environment variable 'TZ'
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✓ dplyr     1.1.1     ✓ readr     2.1.4
## ✓ forcats   1.0.0     ✓ stringr   1.5.0
## ✓ ggplot2   3.4.2     ✓ tibble    3.2.1
## ✓ lubridate 1.9.2     ✓ tidyr     1.3.0
## ✓ purrr     1.0.1
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to
become errors
```

```
library(haven)
library(VIM)
```

```
## Loading required package: colorspace
## Loading required package: grid
## VIM is ready to use.
##
## Suggestions and bug-reports can be submitted at: https://github.com/statistikat/VIM/iss
ues
##
## Attaching package: 'VIM'
##
## The following object is masked from 'package:datasets':
##
##     sleep
```

```
library(mice)
```

```
##
## Attaching package: 'mice'
##
## The following object is masked from 'package:stats':
##
##     filter
##
## The following objects are masked from 'package:base':
##
##     cbind, rbind
```

```
df = read_sav("C:/Users/USER/Desktop/Base_EPCV2019-2020/Base_EPCV2019-2020/menage_2019.sa
v")
head(df)
```

```
## # A tibble: 6 × 150
##   US_ORDRE    A7   A41   A42 A1            A1_1  A2       A2_1  A3         A3_1
##      <dbl> <dbl> <dbl> <dbl> <dbl+lbl>     <chr> <dbl+lb> <chr> <dbl+lbl>  <chr>
## 1        1     1     3     1 1 [Hodh char… "11 …الح [Amo… "1102 …أمر [Ade… "عدل…
## 2        1     2     3     1 1 [Hodh char… "11 …الح [Amo… "1102 …أمر [Ade… "عدل…
## 3        1     3     3     1 1 [Hodh char… "11 …الح [Amo… "1102 …أمر [Ade… "عدل…
## 4        1     4     3     1 1 [Hodh char… "11 …الح [Amo… "1102 …أمر [Ade… "عدل…
## 5        1     5     3     1 1 [Hodh char… "11 …الح [Amo… "1102 …أمر [Ade… "عدل…
## 6        1     6     3     1 1 [Hodh char… "11 …الح [Amo… "1102 …أمر [Ade… "عدل…
## # ℹ 140 more variables: A5 <dbl+lbl>, A6_1 <chr>, A6_2 <dbl>, A8 <chr>,
## #   A8A <dbl+lbl>, A8B <dbl+lbl>, A9 <dbl+lbl>, A91 <dbl+lbl>, A10J <dbl>,
## #   A10M <dbl>, A10A <dbl>, A11H <dbl>, A11M <dbl>, A12 <dbl>, A14 <dbl>,
## #   A15 <dbl+lbl>, A16H <dbl>, A16M <dbl>, A17H <dbl>, A17M <dbl>,
## #   EL_TR <dbl+lbl>, EL_TE <dbl+lbl>, F1 <dbl+lbl>, F2 <dbl>, F2_1 <dbl+lbl>,
## #   F2_2 <dbl>, F3 <dbl>, F4 <dbl+lbl>, F5 <dbl+lbl>, F6 <dbl>, F7 <dbl+lbl>,
## #   F8 <dbl>, F9 <dbl+lbl>, F10 <dbl>, F11 <dbl+lbl>, F12 <dbl>, …
```

The dataframe contains variables with missing values.

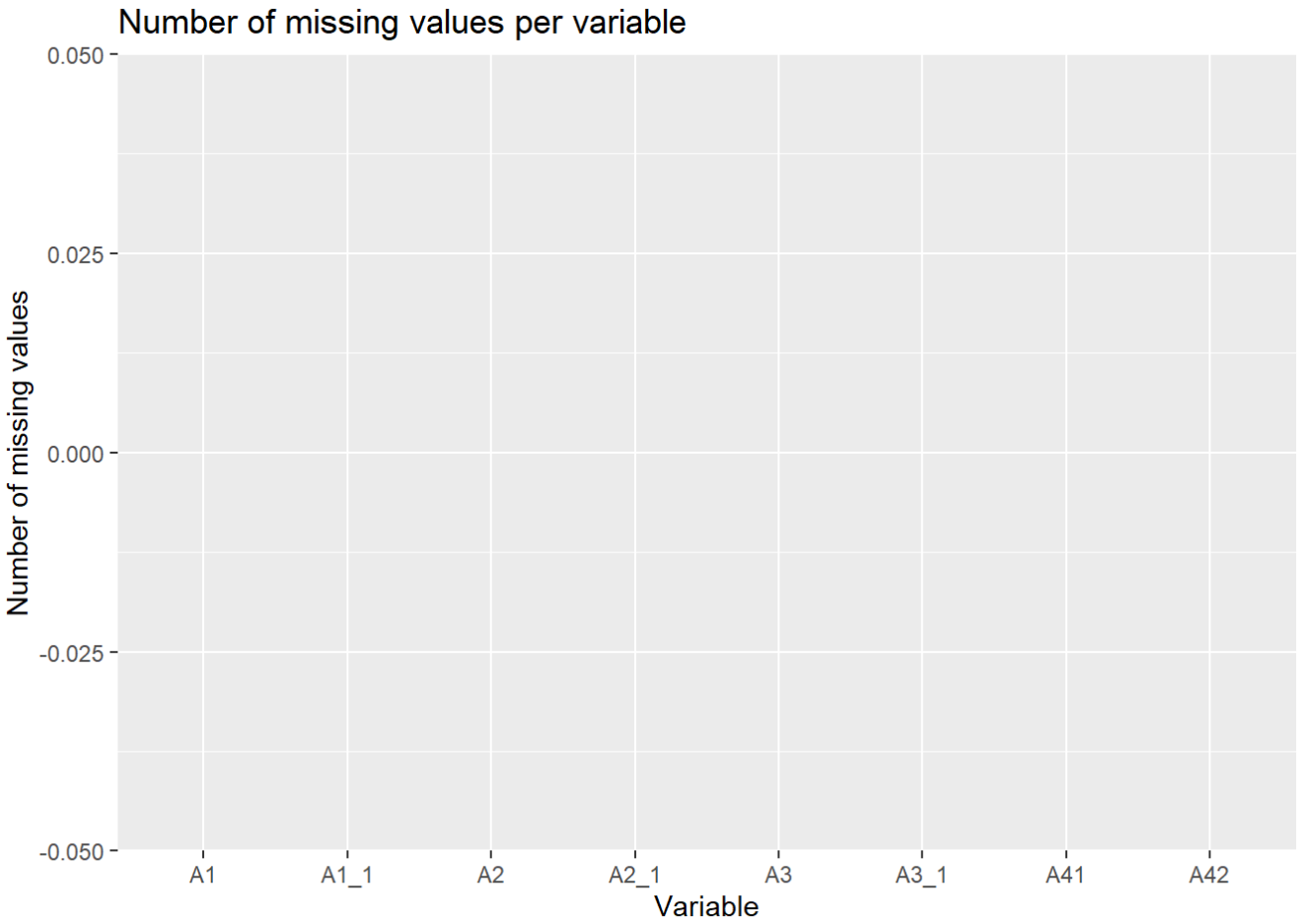
```
na_df = df %>%
  summarise_all(~sum(is.na(.))) %>% # count NAs for each column
  gather() %>% # convert to long format
  arrange(value)
#drop(na_df[,c('G11B1','G11B2')])
na_df = as.data.frame(na_df)
na_df
```

```
##           key value
## 1    US_ORDRE     0
## 2          A7     0
## 3         A41     0
## 4         A42     0
## 5          A1     0
## 6        A1_1     0
## 7          A2     0
## 8        A2_1     0
## 9          A3     0
## 10       A3_1     0
## 11         A5     0
## 12       A6_1     0
## 13       A6_2     0
## 14         A8     0
## 15        A8A     0
## 16        A8B     0
## 17         A9     0
## 18        A91     0
## 19       A10J     0
## 20       A10M     0
## 21       A10A     0
## 22       A11H     0
## 23       A11M     0
## 24        A15     0
## 25       G14A     0
## 26        I1A     0
## 27        I1B     0
## 28        I1C     0
## 29        I1D     0
## 30        I1E     0
## 31        I1F     0
## 32        I1G     0
## 33        I1H     0
## 34        I1I     0
## 35        I1J     0
## 36       I3BA     0
## 37       I3BB     0
## 38       I3BC     0
## 39       I3BD     0
## 40       I3BE     0
## 41       I3BF     0
## 42       I3BG     0
## 43       I3BH     0
## 44      idmen     0
## 45     wilaya     0
## 46  moughataa     0
## 47    commune     0
## 48     milieu     0
## 49        hid     0
## 50       A17H    56
## 51       A17M    56
## 52        A12   253
## 53        A14   253
```

```
## 54      EL_TR   257
## 55      EL_TE   259
## 56        SA1   261
## 57      SA2_A   261
## 58      SA2_B   261
## 59      SA2_C   261
## 60      SA2_D   261
## 61      SA2_E   261
## 62      SA3_A   261
## 63         F1   262
## 64        SA4   262
## 65        SA7   262
## 66       SA_8   262
## 67        SA9   262
## 68       SA10   262
## 69       SA11   262
## 70       SA12   262
## 71       SA13   262
## 72       SA15   262
## 73         F2   263
## 74       F2_1   263
## 75       F2_2   263
## 76         F3   263
## 77         F4   265
## 78         F5   265
## 79         F8   266
## 80         F9   266
## 81        F10   267
## 82        F11   267
## 83        F12   267
## 84        F13   267
## 85        F14   267
## 86        F15   267
## 87        F17   267
## 88        F18   267
## 89        F19   267
## 90        F20   267
## 91        F21   267
## 92         G0   274
## 93         G1   274
## 94         G2   274
## 95         G3   274
## 96        G10   274
## 97         G5   274
## 98         G4   274
## 99       G4_1   274
## 100     G12_Q   274
## 101     G12_U   274
## 102     G12_F   274
## 103       G13   274
## 104       G15   274
## 105       G16   274
## 106        G6   274
## 107      A16H   275
## 108      A16M   275
```

```
## 109        G7   275
## 110      G11A   275
## 111        I2   275
## 112        I3   275
## 113        I4   275
## 114       I11   275
## 115       I12   275
## 116       I13   275
## 117      G12A  1431
## 118      I3A1  2998
## 119      I3A2  2998
## 120      I3A3  3634
## 121       G5A  3982
## 122       G5B  3982
## 123       G5C  3982
## 124       G5D  3982
## 125     SA3_B  4803
## 126     SA5_1  5924
## 127     SA5_2  5924
## 128     SA5_3  5924
## 129     SA5_4  5924
## 130     SA5_5  5924
## 131     SA5_6  5924
## 132     SA5_7  5924
## 133     SA5_8  5924
## 134     SA5_9  5924
## 135    SA5_10  5924
## 136    SA5_11  5924
## 137    SA5_12  5924
## 138     SA6_A  5924
## 139     SA6_B  5924
## 140       G14  6206
## 141     SA6_C  6503
## 142      SA14  8304
## 143      G12B  9064
## 144      SA16  9298
## 145       G5E  9317
## 146       G5F  9317
## 147        F7  9532
## 148        F6  9534
## 149     G11B1 10109
## 150     G11B2 10109
```

```
ggplot(na_df[3:10,], aes(x = key, y = value)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  xlab("Variable") +
  ylab("Number of missing values") +
  ggtitle("Number of missing values per variable")
```



We can see that the number of missing values for each variable is significant.

Begin Imputation We have selected 5 quantitative variables: F2, F6, I12, I13, and SA6_C.

Impute F2 - How many rooms are there in your accommodation?

Create a sub-dataframe without missing values of F2 and create new arbitrary missing values.

```
df_notna <- df[!is.na(df$F2), ]
set.seed(123)
sample_indices <- sample(1:nrow(df_notna), round(0.4*nrow(df_notna)), replace=FALSE)
F2_na = df_notna['F2']
F2_na = data.frame(F2_na)
F2_na[sample_indices,] = NA
df_notna = data.frame(df_notna)
df_notna['F2'] = F2_na
```

Imputation evaluation using mice:

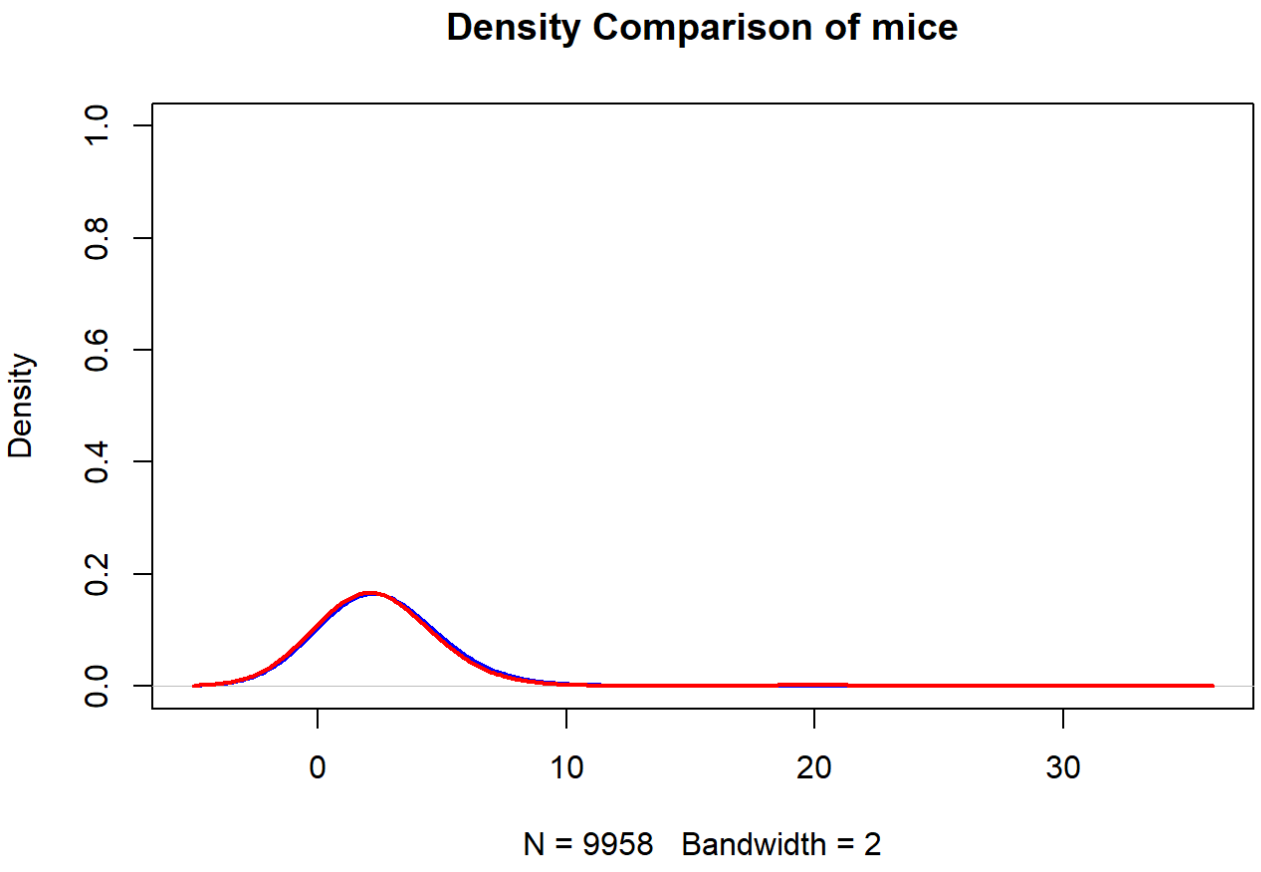
```
df_mice <- mice(df_notna[,c(34,35,36,37)], m=3,method = "pmm")
```

```
##
##  iter imp variable
##   1   1  F2
##   1   2  F2
##   1   3  F2
##   2   1  F2
##   2   2  F2
##   2   3  F2
##   3   1  F2
##   3   2  F2
##   3   3  F2
##   4   1  F2
##   4   2  F2
##   4   3  F2
##   5   1  F2
##   5   2  F2
##   5   3  F2
```

Evaluate:

```
plot(density(df[!is.na(df$F2), ]$F2,bw=2), main = "Density Comparison of mice", col = "blu
e", lwd = 2, ylim = c(0, 1))

# Add a density plot for F6_na
lines(density(as.numeric(unlist(df_mice$imp$F2)),bw=2), col = "red", lwd = 2)
```



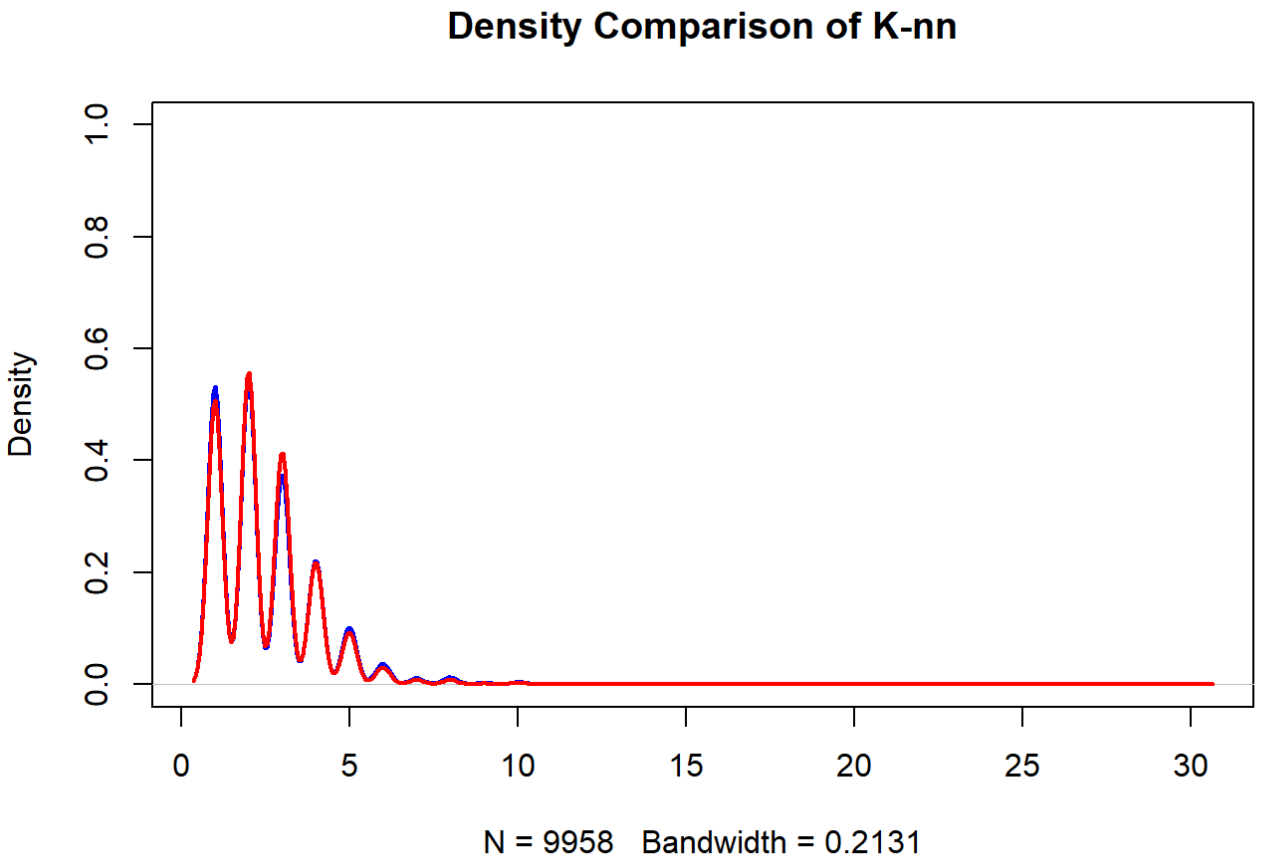
Imputation d'evaluation par : k-nn

```
df_knn=kNN(df_notna,variable = 'F2',k=3)
```

Evaluate

```
plot(density(df[!is.na(df$F2), ]$F2), main = "Density Comparison of K-nn", col = "blue", l
wd = 2, ylim = c(0, 1))

# Add a density plot for F6_na
lines(density(df_knn$F2), col = "red", lwd = 2)
```



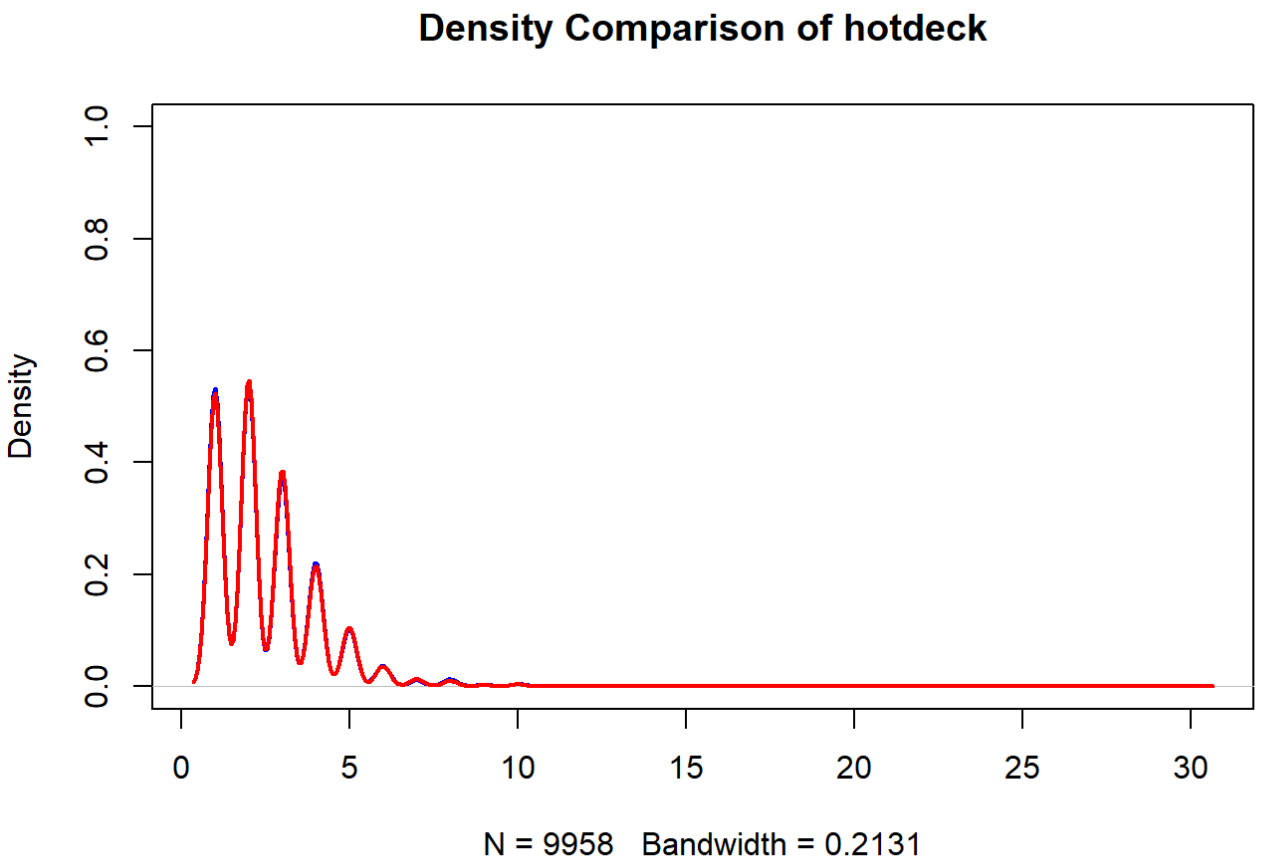
Imputation evaluation using hotdeck:

```
df_hd = hotdeck(df_notna)
```

Evaluate:

```
plot(density(df[!is.na(df$F2), ]$F2), main = "Density Comparison of hotdeck", col = "blu
e", lwd = 2, ylim = c(0, 1))

# Add a density plot for F6_na
lines(density(df_hd$F2), col = "red", lwd = 2)
```



Imputer F6 - Combien d'hectares de terres agricoles le ménage utilise qui ne lui appartiennent pas ?

Creer un sous-dataframe sans les valeurs manquantes de F6 et creer des nouvelles valeurs manquantes arbitraire de F6

```
df_notna <- df[!is.na(df$F6), ]
set.seed(123)
sample_indices <- sample(1:nrow(df_notna), round(0.4*nrow(df_notna)), replace=FALSE)
F6_na = df_notna['F6']
F6_na = data.frame(F6_na)
F6_na[sample_indices,] = NA
df_notna = data.frame(df_notna)
df_notna['F6'] = F6_na
```

Imputation d'evaluation par : mice

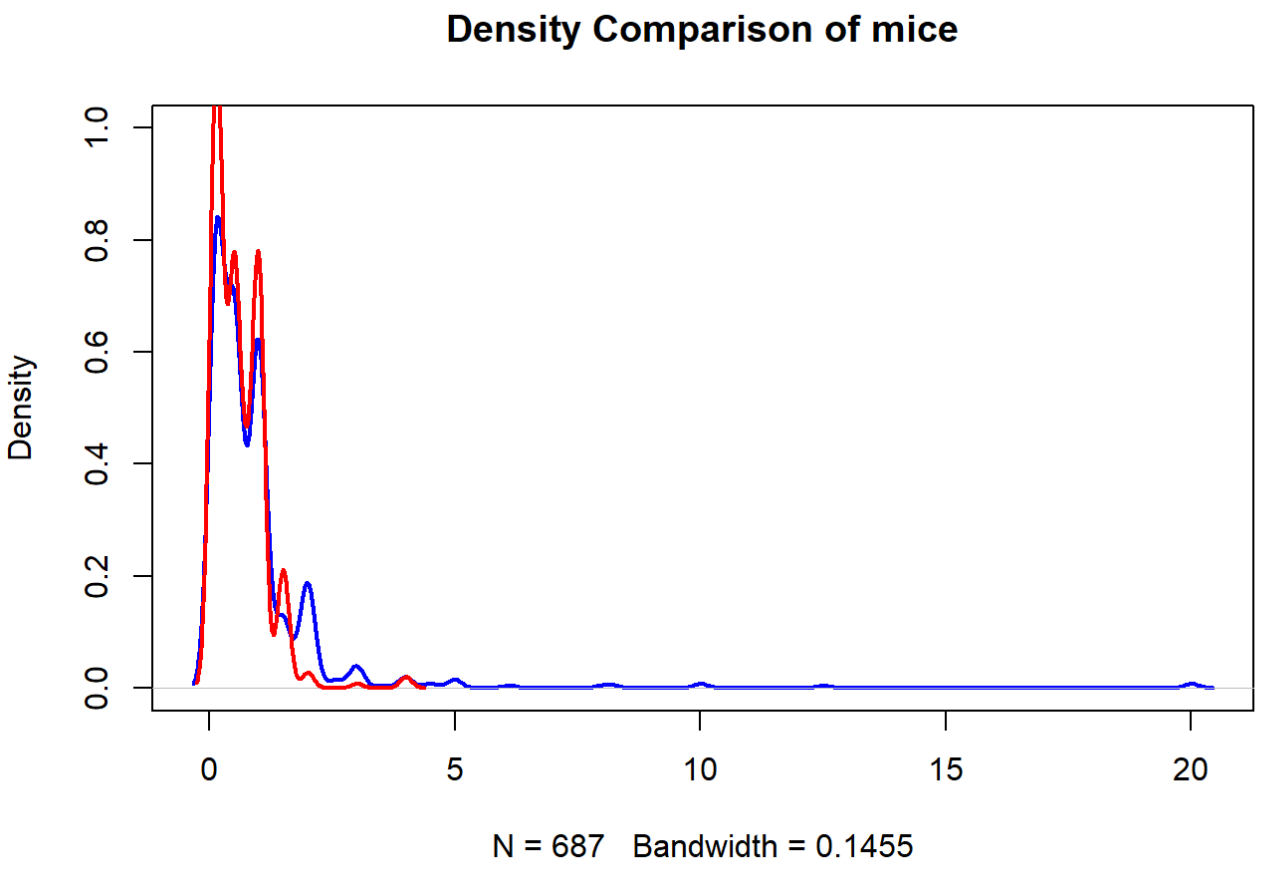
```
df_mice <- mice(df_notna[,c(39,40)], m=3,method = "pmm")
```

```
##
##  iter imp variable
##   1   1  F6
##   1   2  F6
##   1   3  F6
##   2   1  F6
##   2   2  F6
##   2   3  F6
##   3   1  F6
##   3   2  F6
##   3   3  F6
##   4   1  F6
##   4   2  F6
##   4   3  F6
##   5   1  F6
##   5   2  F6
##   5   3  F6
```

Evaluer :

```
plot(density(df[!is.na(df$F6), ]$F6), main = "Density Comparison of mice", col = "blue", l
wd = 2, ylim = c(0, 1))

# Add a density plot for F6_na
lines(density(as.numeric(unlist(df_mice$imp$F6))), col = "red", lwd = 2)
```



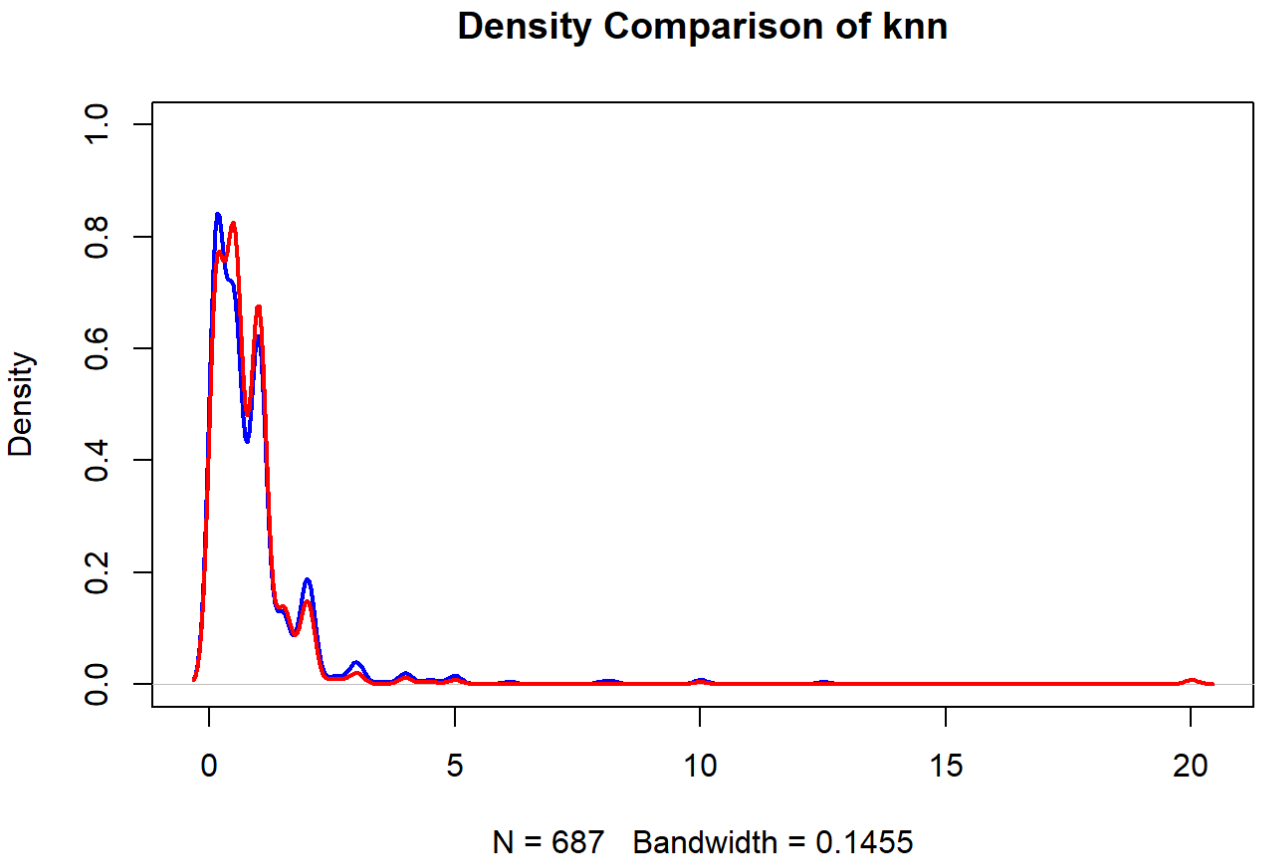
Imputation d'evaluation par : k-nn

```
df_knn=kNN(df_notna,variable = 'F6',k=3)
```

Evaluer

```
plot(density(df[!is.na(df$F6), ]$F6), main = "Density Comparison of knn", col = "blue", lw
d = 2, ylim = c(0, 1))

# Add a density plot for F6_na
lines(density(df_knn$F6), col = "red", lwd = 2)
```



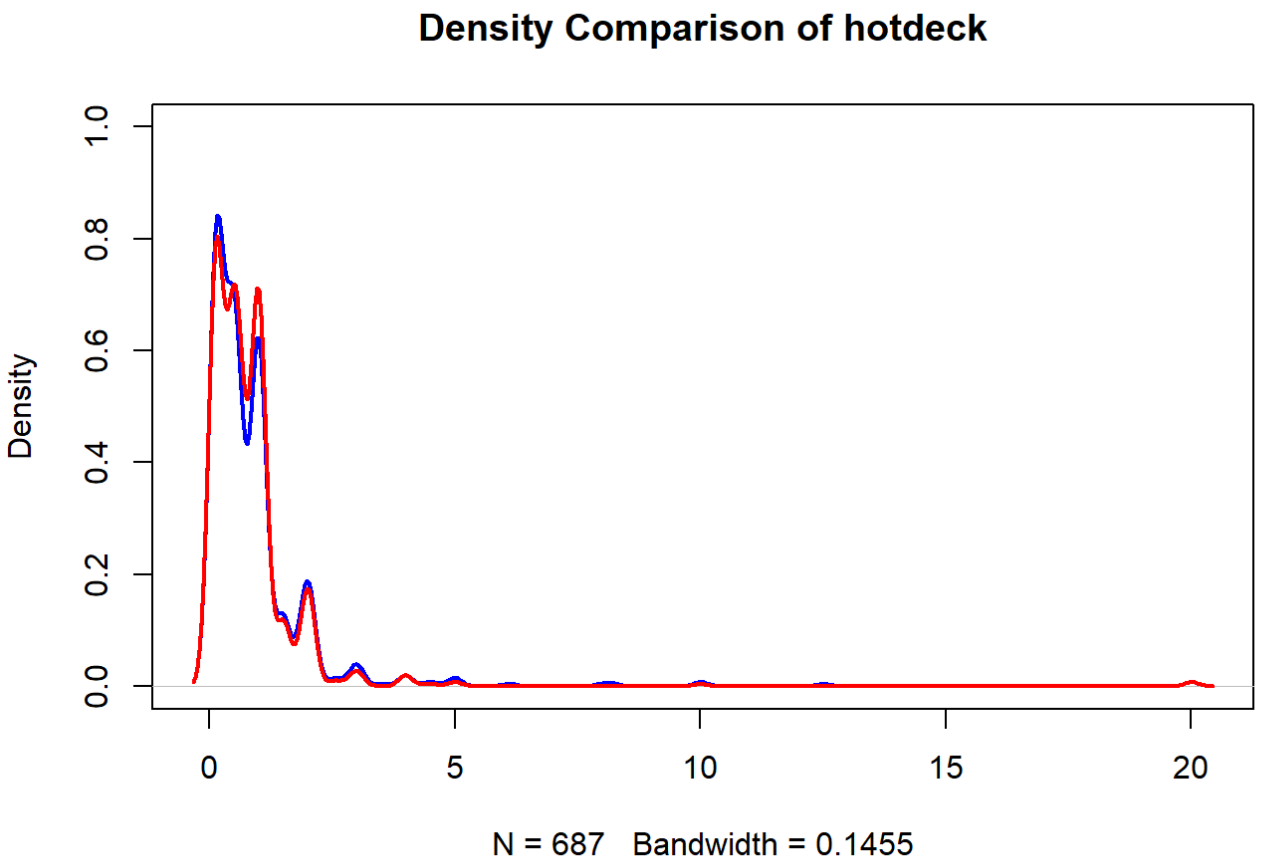
Imputation d'evaluation par : hotdeck

```
df_hd = hotdeck(df_notna)
```

Evaluer

```
plot(density(df[!is.na(df$F6), ]$F6), main = "Density Comparison of hotdeck", col = "blu
e", lwd = 2, ylim = c(0, 1))

# Add a density plot for F6_na
lines(density(df_hd$F6), col = "red", lwd = 2)
```


Density Comparison of hotdeck
Density
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
N = 687 Bandwidth = 0.1455