# Keynote
## ANALYSIS AND IMPUTATION OF MISSING VALUES

Matthias Templ

Institute for Data Analysis and Process Design
Zurich University of Applied Sciences, Switzerland &
Vienna University of Technology

iCMS2021 (04.08.2021)

Zürcher Hochschule
für Angewandte Wissenschaften

zh aw School of Engineering
IDP Institut für Datenanalyse
und Prozessdesign

# Should we impute?

# Should we impute?

- Mr. Sere won $500 as Zimbabwe's ugliest man, mainly because of his lack of values in his mouth:
- Mr. Goldtoothfinger's dentures are worth more after imputation even done in a non-realistic manner.

# Motivation


(artist: Johannes Templ, 4-5 years, commissioned work for the conference)

# Motivation

Corrected and imputed versus non-imputed

# Motivation

Corrected and imputed versus non-imputed

# Missing values are everywhere

Missing values are present in almost all real-world data sets and disciplines of research, not only in

- clinical studies,
- official statistics,
- sociology,
- omics sciences,
- geochemistry,
- microarray research,
- psychology,
- educational research,
- . . . ,
- for COVID-19 data,
- and even in rather exotic fields of research such as sport sciences, crystallography, just to name a few.

# Reasons for missing values in data sets

Reasons for missing values (selection):

- failed measurement units for measurements of groundwater quality or temperature
- lost soil samples in geochemistry, or soil samples that to analyze anew, but are exhausted.
- respondents who **don't want** to provide information, **don't know** information or **skip** questions because of a too long or too complicated questionnaire.
- A patient dies and is therefore no longer in a medical study.
- Measurements are implausible and thus set to be missing.
- ...

# Influence of missing values

- Standard statistical methods are typically designed for complete data sets.
- Missing values can have a strong influence on resulting figures and analysis.
- Estimators can be **bias**ed and their **variance** will be underestimated if missing values and their structure are ignored.

To ensure the quality of the estimates, missing values should be imputed using advanced methods.

# Questions

Planning to analyze data that include missing values involves:

- Which **kind of missing data** are present in the data?

# Questions

Planning to analyze data that include missing values involves:

- Which **kind of missing data** are present in the data?
- What **mechanism** does this missing data follow?

# Questions

Planning to analyze data that include missing values involves:

- Which **kind of missing data** are present in the data?
- What **mechanism** does this missing data follow?
- Which **imputation** strategy is appropriate for these data?

# Questions

Planning to analyze data that include missing values involves:

- Which **kind of missing data** are present in the data?
- What **mechanism** does this missing data follow?
- Which **imputation** strategy is appropriate for these data?
- Is there an accepted model that these data follow?

# Questions

Planning to analyze data that include missing values involves:

- Which **kind of missing data** are present in the data?
- What **mechanism** does this missing data follow?
- Which **imputation** strategy is appropriate for these data?
- Is there an accepted model that these data follow?
- **How many** missing values are present in the data set?

# Questions

Planning to analyze data that include missing values involves:

- Which **kind of missing data** are present in the data?
- What **mechanism** does this missing data follow?
- Which **imputation** strategy is appropriate for these data?
- Is there an accepted model that these data follow?
- **How many** missing values are present in the data set?
- What are the consequences of a **sensitivity analysis**, e.g. by comparing the study results with and without imputed data?

# Questions

Planning to analyze data that include missing values involves:

- Which **kind of missing data** are present in the data?
- What **mechanism** does this missing data follow?
- Which **imputation** strategy is appropriate for these data?
- Is there an accepted model that these data follow?
- **How many** missing values are present in the data set?
- What are the consequences of a **sensitivity analysis**, e.g. by comparing the study results with and without imputed data?
- Is the aim to train and use a predictive model with high predictive power? Or is the statistical uncertainty in main focus? Is there a need for **multiple imputation** due to these facts, or is a **single imputation** the better, more practical way?

# Questions

Planning to analyze data that include missing values involves:

- Which **kind of missing data** are present in the data?
- What **mechanism** does this missing data follow?
- Which **imputation** strategy is appropriate for these data?
- Is there an accepted model that these data follow?
- **How many** missing values are present in the data set?
- What are the consequences of a **sensitivity analysis**, e.g. by comparing the study results with and without imputed data?
- Is the aim to train and use a predictive model with high predictive power? Or is the statistical uncertainty in main focus? Is there a need for **multiple imputation** due to these facts, or is a **single imputation** the better, more practical way?
- Are there project-related time constraints for the analysis and imputation of missing values?

# Difficulties

- outliers may influence a (classical) imputation method so that the imputations became arbitrary.

# Difficulties

- outliers may influence a (classical) imputation method so that the imputations became arbitrary.
- imputation of data sets with mixed scaled variables. For example, a data set contains a mix of continuous, semi-continuous, binary, nominal, ordinal or count variables.

# Difficulties

- outliers may influence a (classical) imputation method so that the imputations became arbitrary.
- imputation of data sets with mixed scaled variables. For example, a data set contains a mix of continuous, semi-continuous, binary, nominal, ordinal or count variables.
- special kind of data sets, e.g., compositional data, whereby other methods are applied than for *standard* data sets.

# Difficulties

- outliers may influence a (classical) imputation method so that the imputations became arbitrary.
- imputation of data sets with mixed scaled variables. For example, a data set contains a mix of continuous, semi-continuous, binary, nominal, ordinal or count variables.
- special kind of data sets, e.g., compositional data, whereby other methods are applied than for *standard* data sets.
- selection of the imputation method

# Difficulties

- outliers may influence a (classical) imputation method so that the imputations became arbitrary.
- imputation of data sets with mixed scaled variables. For example, a data set contains a mix of continuous, semi-continuous, binary, nominal, ordinal or count variables.
- special kind of data sets, e.g., compositional data, whereby other methods are applied than for *standard* data sets.
- selection of the imputation method

Analyse the structure of missing values before impute missing values

...

# Properties of missing values

- **Missing At Random (MAR)**: the reason why a random value is missing from a variable can be explained from the other variables in a data set.
  - Example: The more dangerous pretator, the more likely that the measurement of the length of the dreaming phase per day of a mammal fails.

# Properties of missing values

- **Missing At Random (MAR)**: the reason why a random value is missing from a variable can be explained from the other variables in a data set.
  - Example: The more dangerous pretator, the more likely that the measurement of the length of the dreaming phase per day of a mammal fails.
- **Missing Completely At Random (MCAR)**: there is nothing fishy, the missing is absolutely random.
  - Example: Respondents may decide to omit a question due to loss of interest.

# Properties of missing values

- **Missing At Random (MAR)**: the reason why a random value is missing from a variable can be explained from the other variables in a data set.
  - Example: The more dangerous pretator, the more likely that the measurement of the length of the dreaming phase per day of a mammal fails.
- **Missing Completely At Random (MCAR)**: there is nothing fishy, the missing is absolutely random.
  - Example: Respondents may decide to omit a question due to loss of interest.
- **Missing Not At Random (MNAR)**: reasons of missingness can be not explained with the help of your data set.
  - Example: The weight of heavy weighted mammals cannot be measured.

# Properties of missing values

The detection of non-MCAR situations for a variable of interest or to check for any variable with missing values in a dataset has a long history, starting with Little (1988) who formulated the first test for MCAR. Various extensions have been made, for example, by Li (2013), Li (2010) and Bojinov, Pillai, and Rubin (2017).

These tests are very sensitive to non-normal data and outliers.

- M. Templ, Alfons, and Filzmoser (2012) :

> *Much better is to use explanatory data analysis to assess the structure of missing values.*

# Visualization of missing values

Learn relationships: learn about the **relationships** of the variables in your data set(s) even they include missing values.

# Visualization of missing values

Learn relationships: learn about the **relationships** of the variables in your data set(s) even they include missing values.

Observe particularities: explore data with missing values for the **dectection of possible problems**, such as outliers or skewness of variables, and other peculiarities.

# Visualization of missing values

Learn relationships: learn about the **relationships** of the variables in your data set(s) even they include missing values.

Observe particularities: explore data with missing values for the **dectection of possible problems**, such as outliers or skewness of variables, and other peculiarities.

Learn special structures of missing values: learn about the incomplete information in the data and to **identify possible structures of the missing values**.

# Visualization of missing values

Learn relationships: learn about the **relationships** of the variables in your data set(s) even they include missing values.

Observe particularities: explore data with missing values for the **dectection of possible problems**, such as outliers or skewness of variables, and other peculiarities.

Learn special structures of missing values: learn about the incomplete information in the data and to **identify possible structures of the missing values**.

Support decisions on data pre-preprocessing: decide how to handle the data, either whether to contact some respondents again or perform measurements again, calibrate parts of the data for missing values, or perform imputation.

# Visualization of missing values

Learn relationships: learn about the **relationships** of the variables in your data set(s) even they include missing values.

Observe particularities: explore data with missing values for the **dectection of possible problems**, such as outliers or skewness of variables, and other peculiarities.

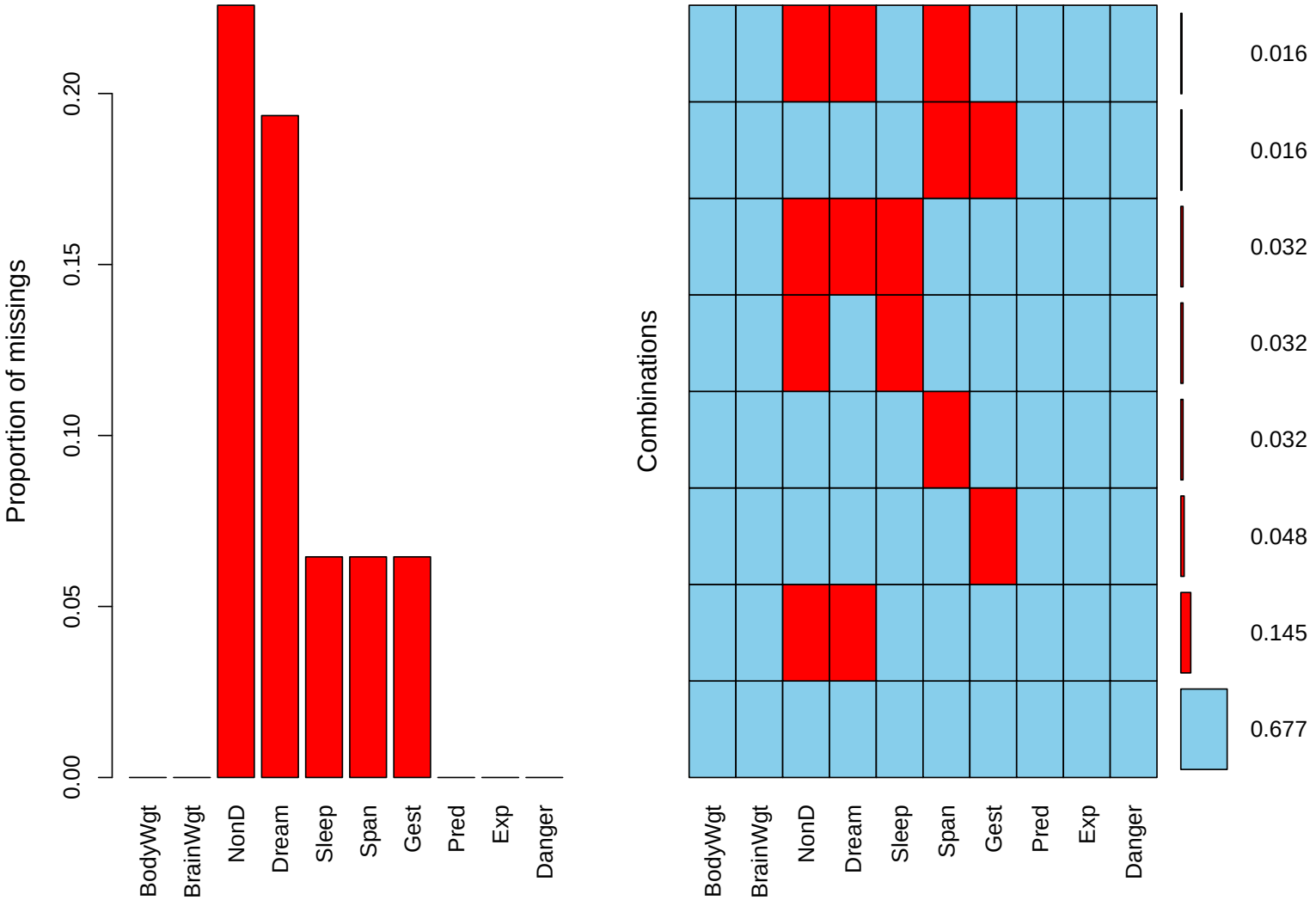
Learn special structures of missing values: learn about the incomplete information in the data and to **identify possible structures of the missing values**.

Support decisions on data pre-preprocessing: decide how to handle the data, either whether to contact some respondents again or perform measurements again, calibrate parts of the data for missing values, or perform imputation.

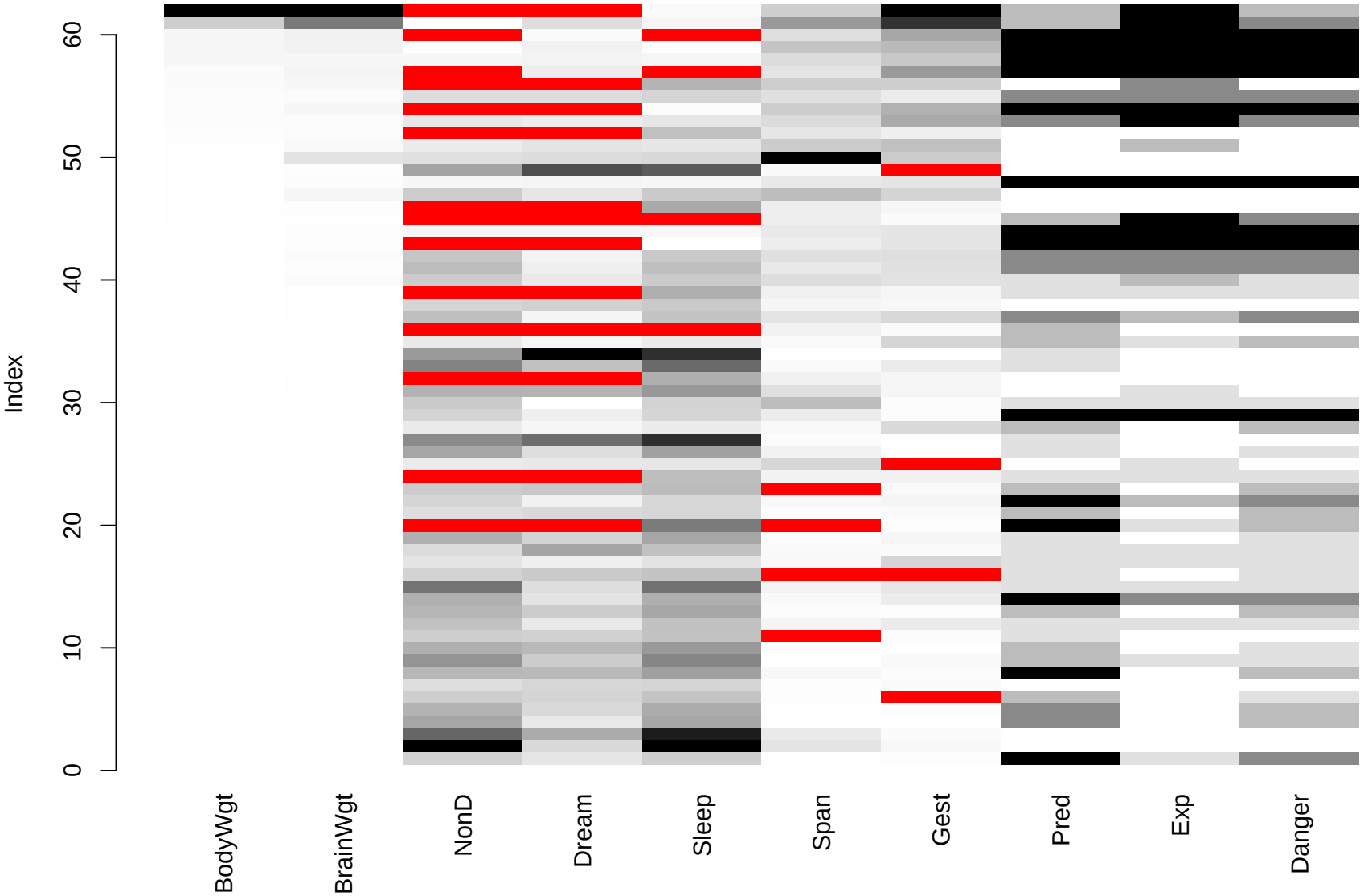
Selection of an imputation method: Choosing an inappropriate imputation method can destroy the multivariate relationships in the data and biased results may result.

# Visualization: summaries

```
library("VIM"); data("sleep")
a <- aggr(sleep, plot = FALSE); plot(a, numbers = TRUE)
```

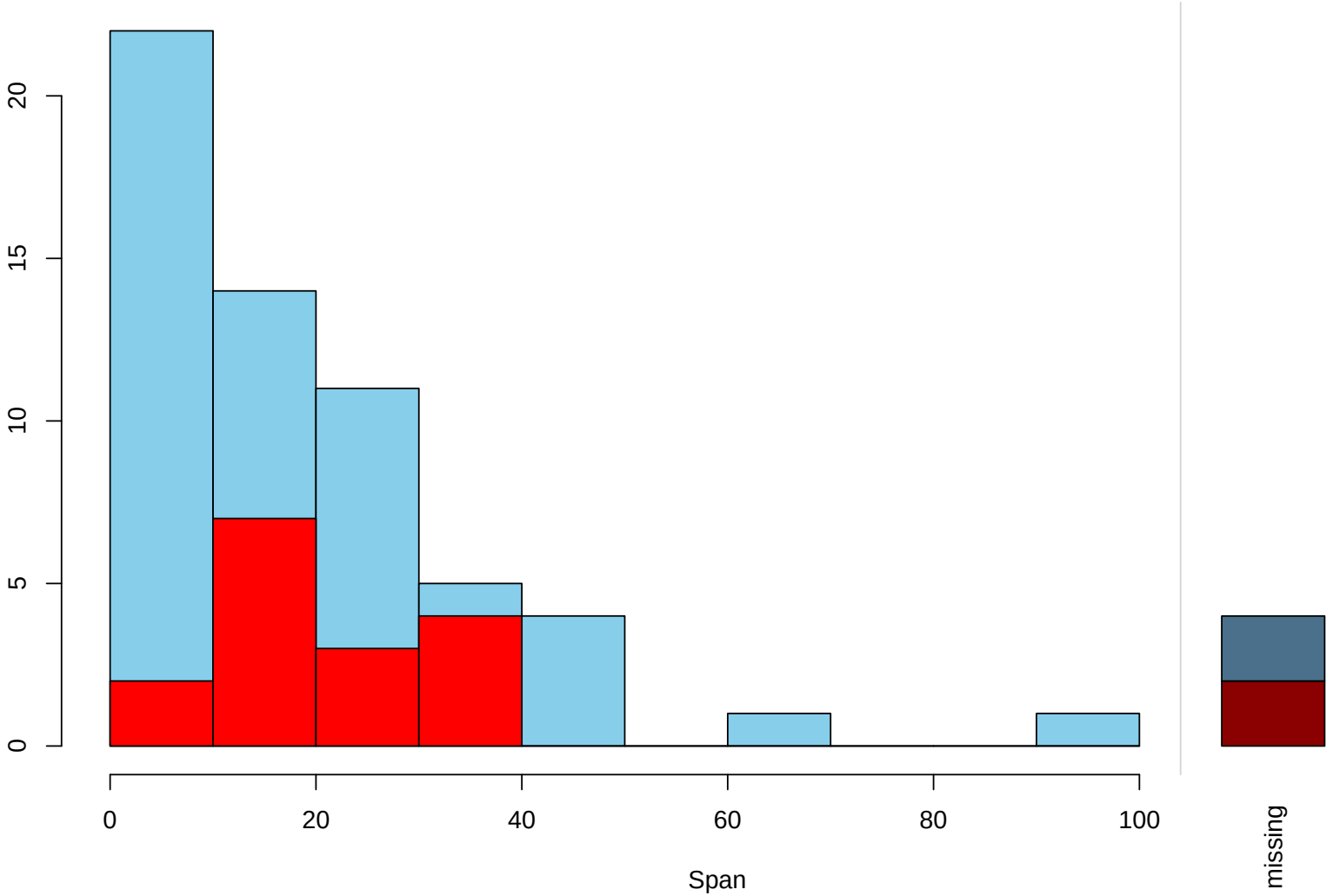


# Visualization: detecting MAR situations using the matrix plot



# Visualization: detecting MAR situations using histograms

- Distribution of variable `Span` (life span of mammals)
- Subset in red: Distribution of variable `Span` that includes missing values in covariates



# Visualization: detecting MAR situations using spinnograms

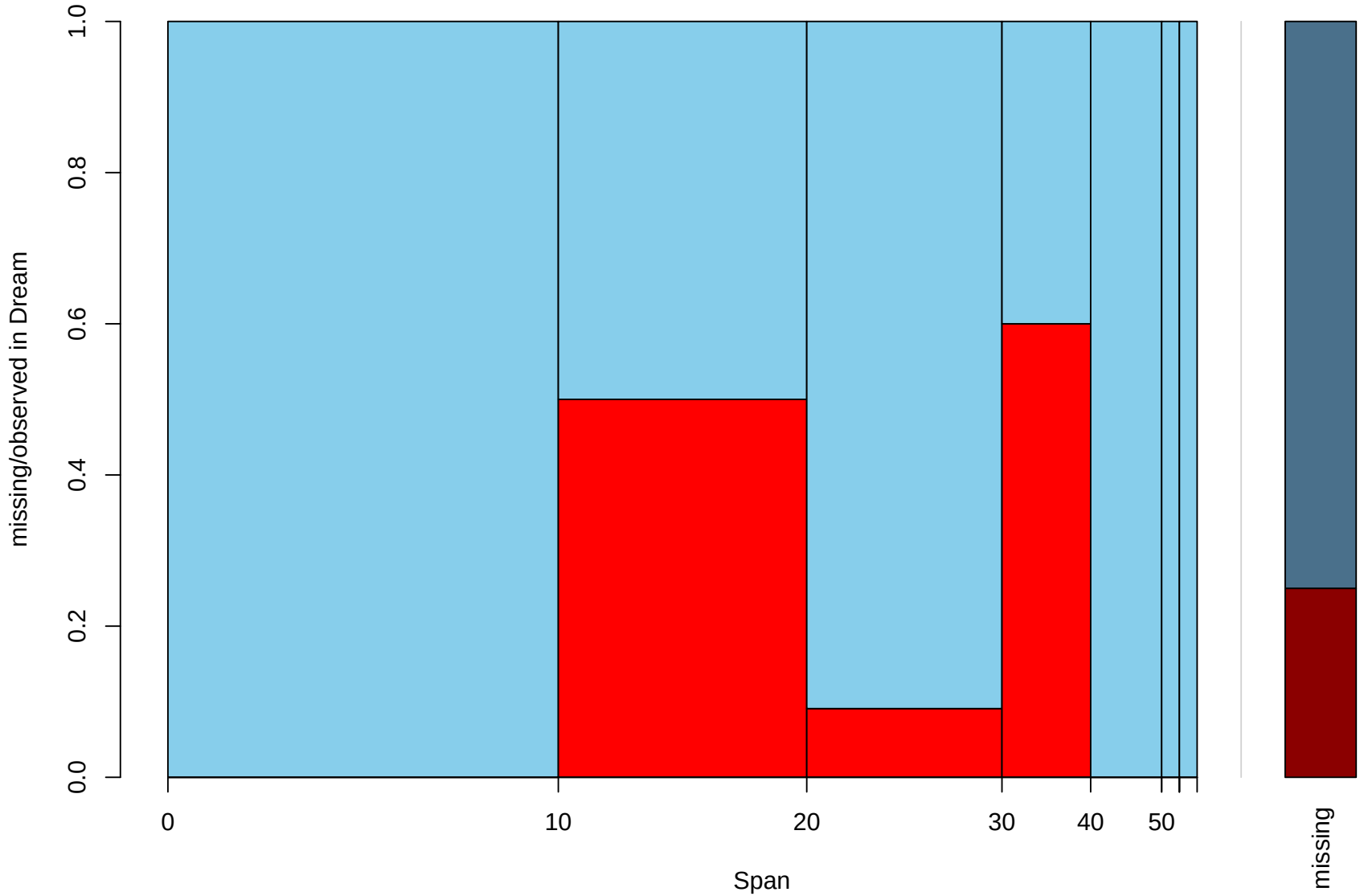



Figure 3: Spinogram of Span with highlighting based on variable Dream.

# Visualization: detecting MAR situations using barcharts

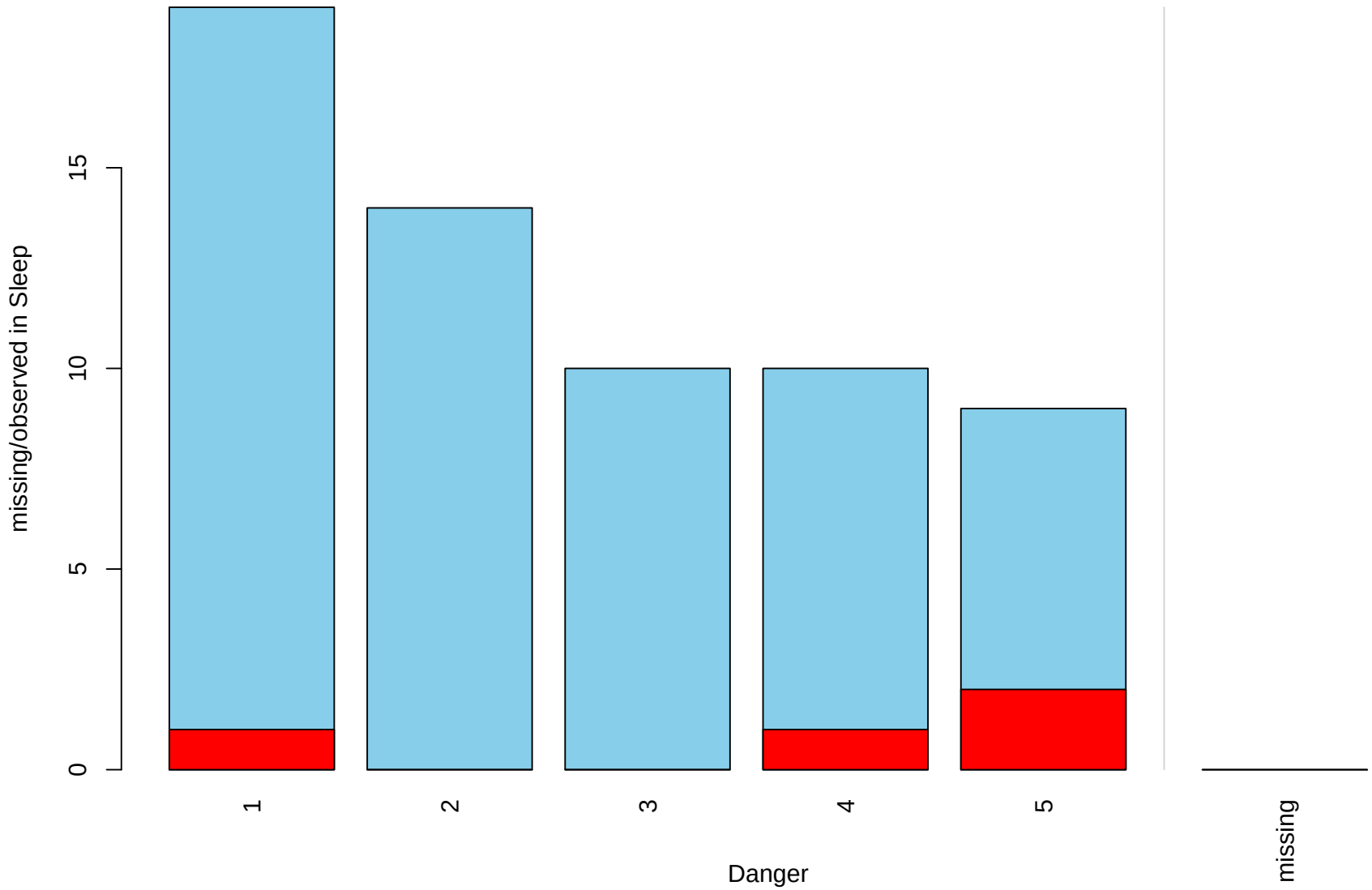



Figure 4: Barplot of variable Danger with highlighting based on variable

# Visualization: detecting MAR situations using spineplots

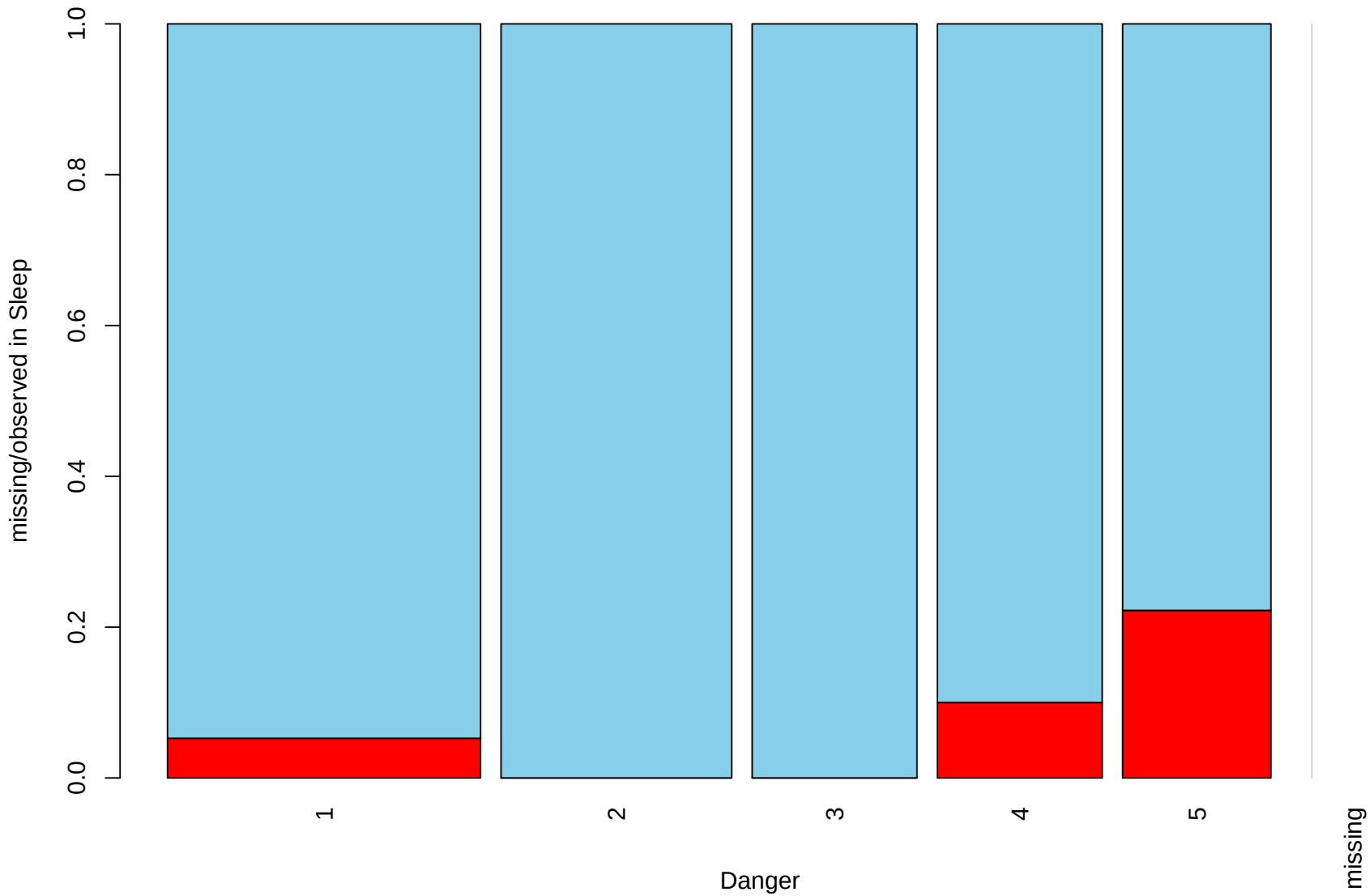



Figure 5: Barplot of variable Danger with highlighting based on variable

# Visualization: detecting MAR situations using boxplots

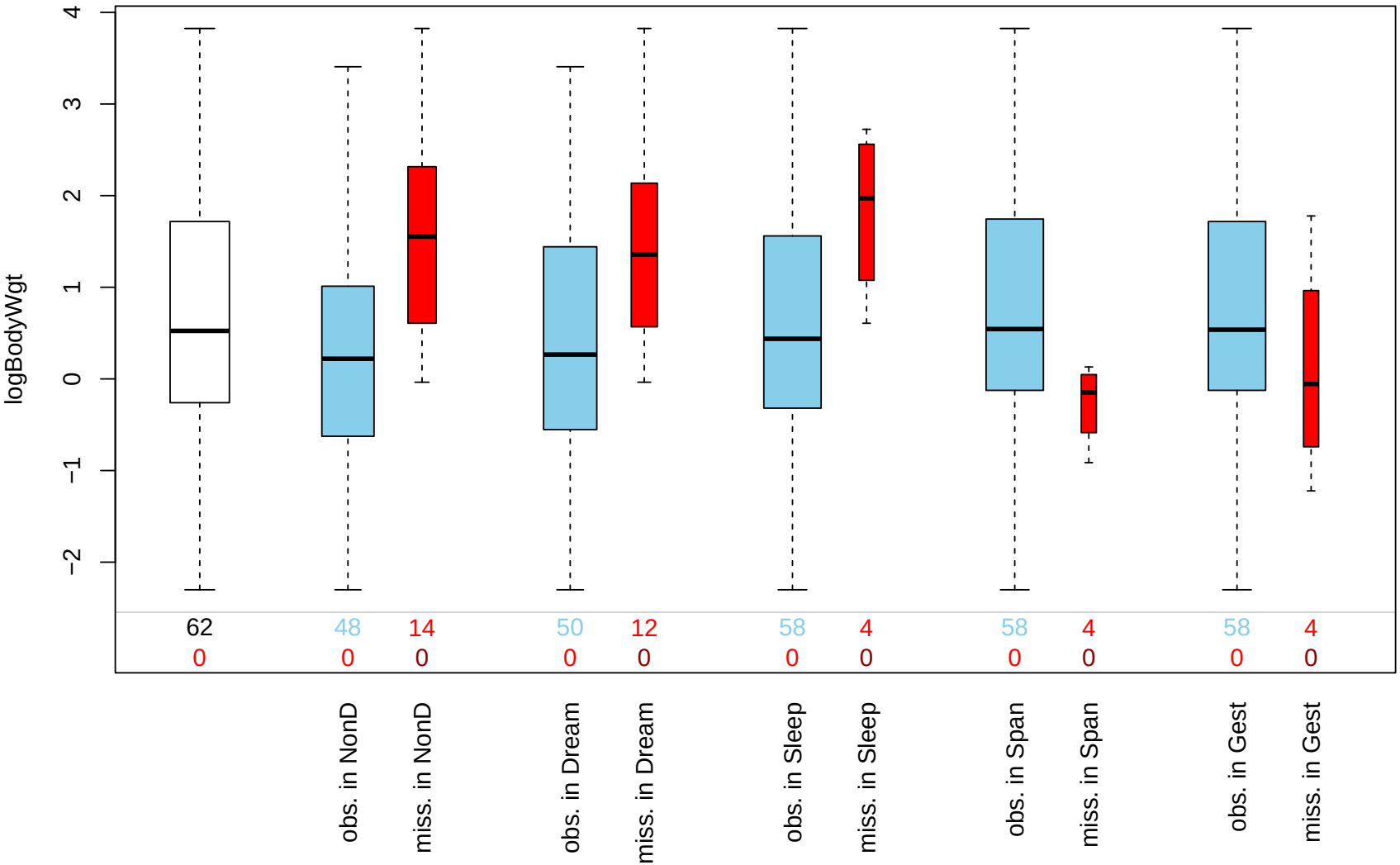
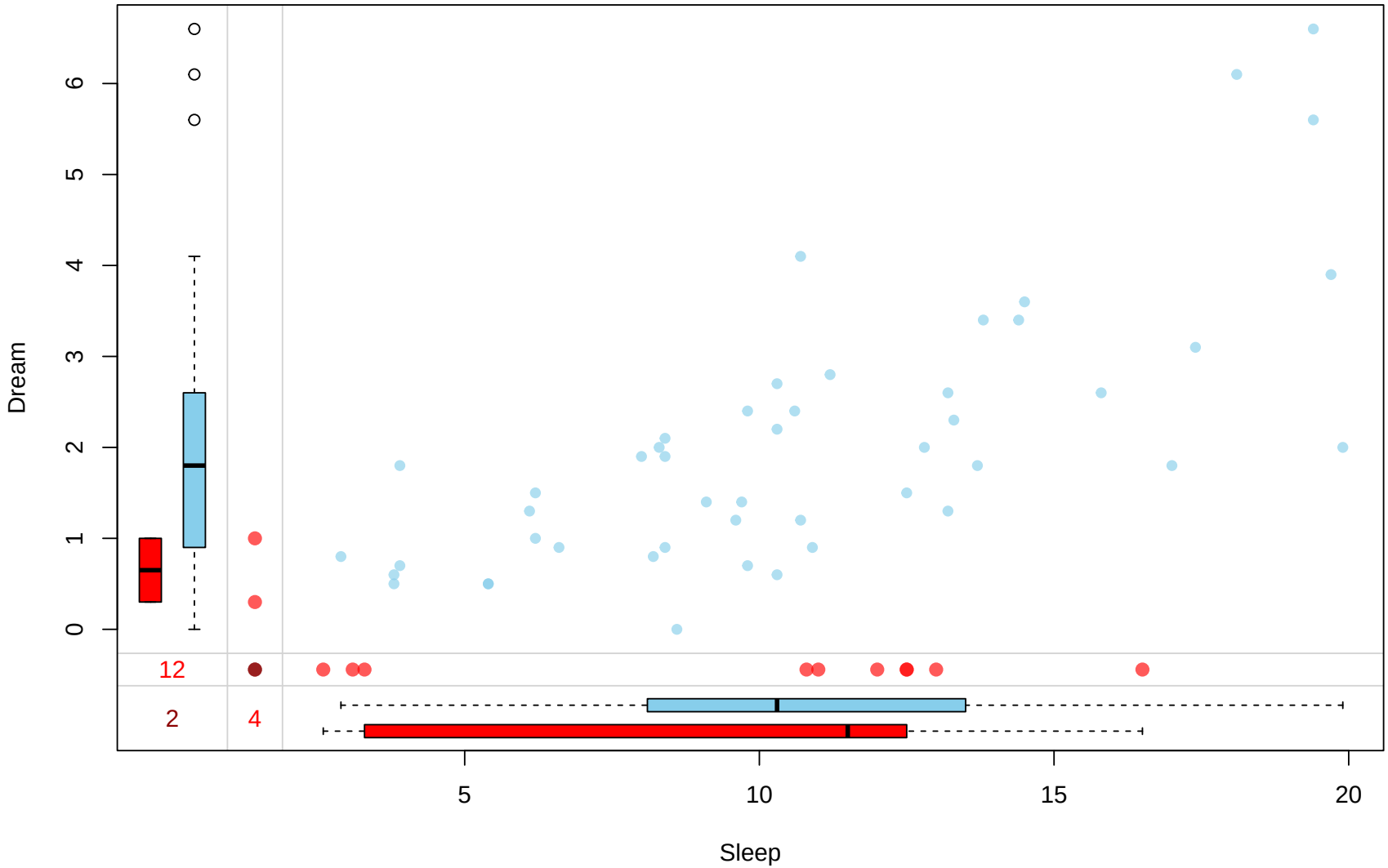



Figure 6: Parallel boxplots to observe MAR situations for a selected continuous variable.

# Visualization: detecting MAR situations using scatterplots



# Visualization

- Multiple scatterplots with missing values
- Parallel coordinate plots with missing information
- Mosaic plots with missing values
- Maps with missing values
- ...

$\longrightarrow$ many more visualization tools included in R package VIM (M. Templ et al. 2019)

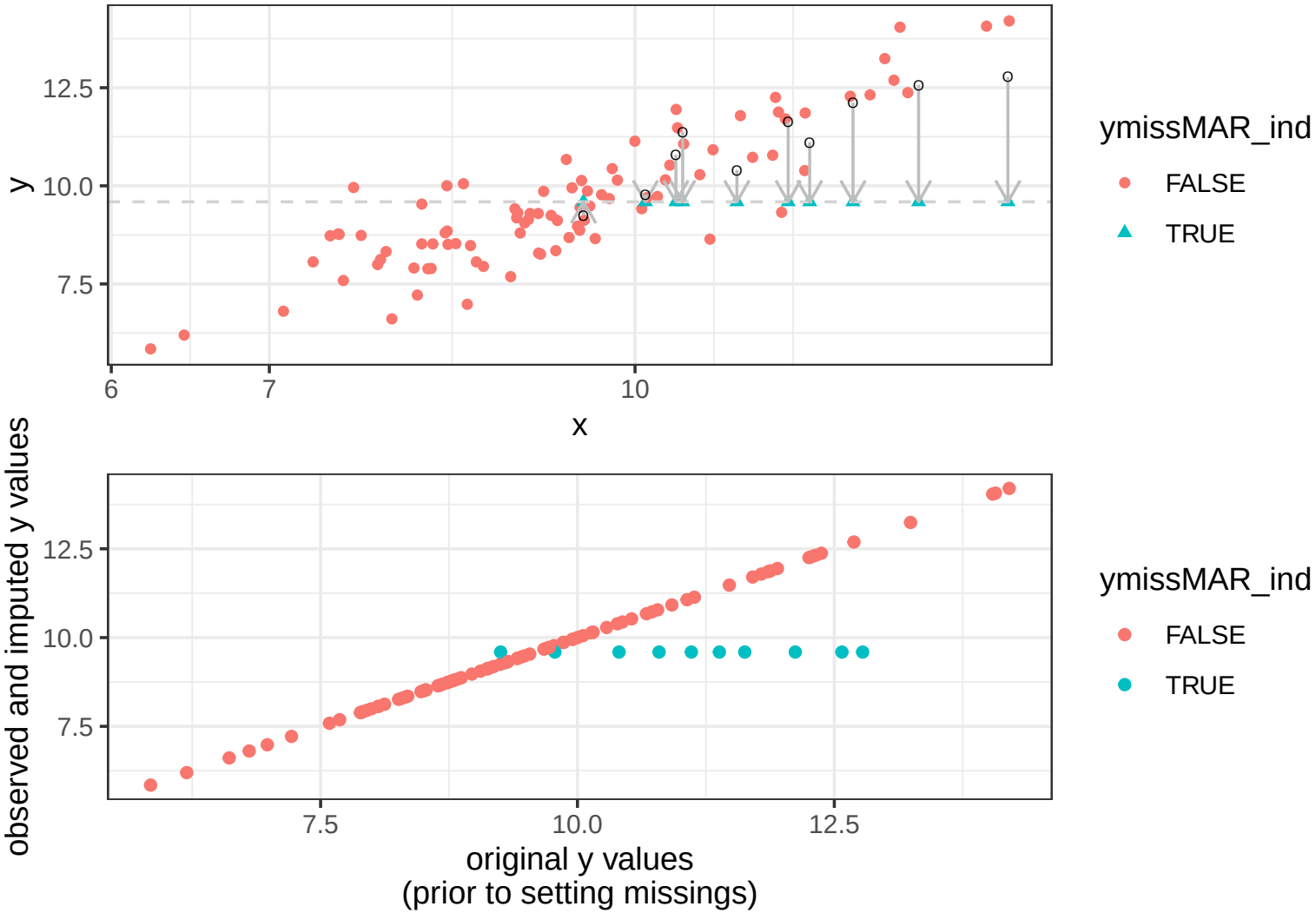
# Imputation

- Deductive methods using logical rules
- Univariate methods
- Nearest-neighbor methods
- Distributional and covariance-based methods
- Model-based methods
- Tree-based methods
- Neural networks
- Methods for time series
- Methods for compositional data

and always a decision for single or multiple imputation

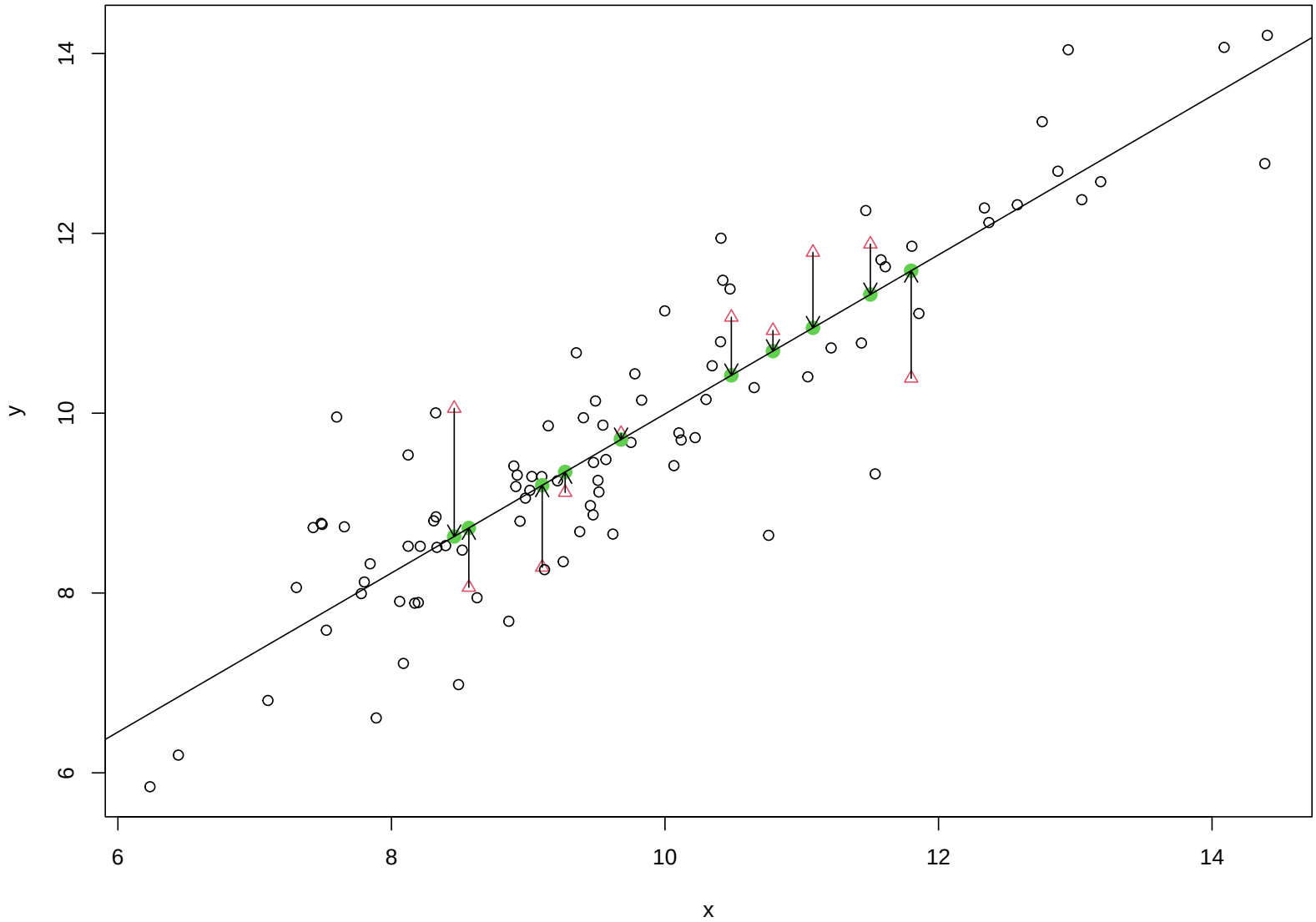
# Univariate methods

Don't apply them, see e.g. imputation with the mean:



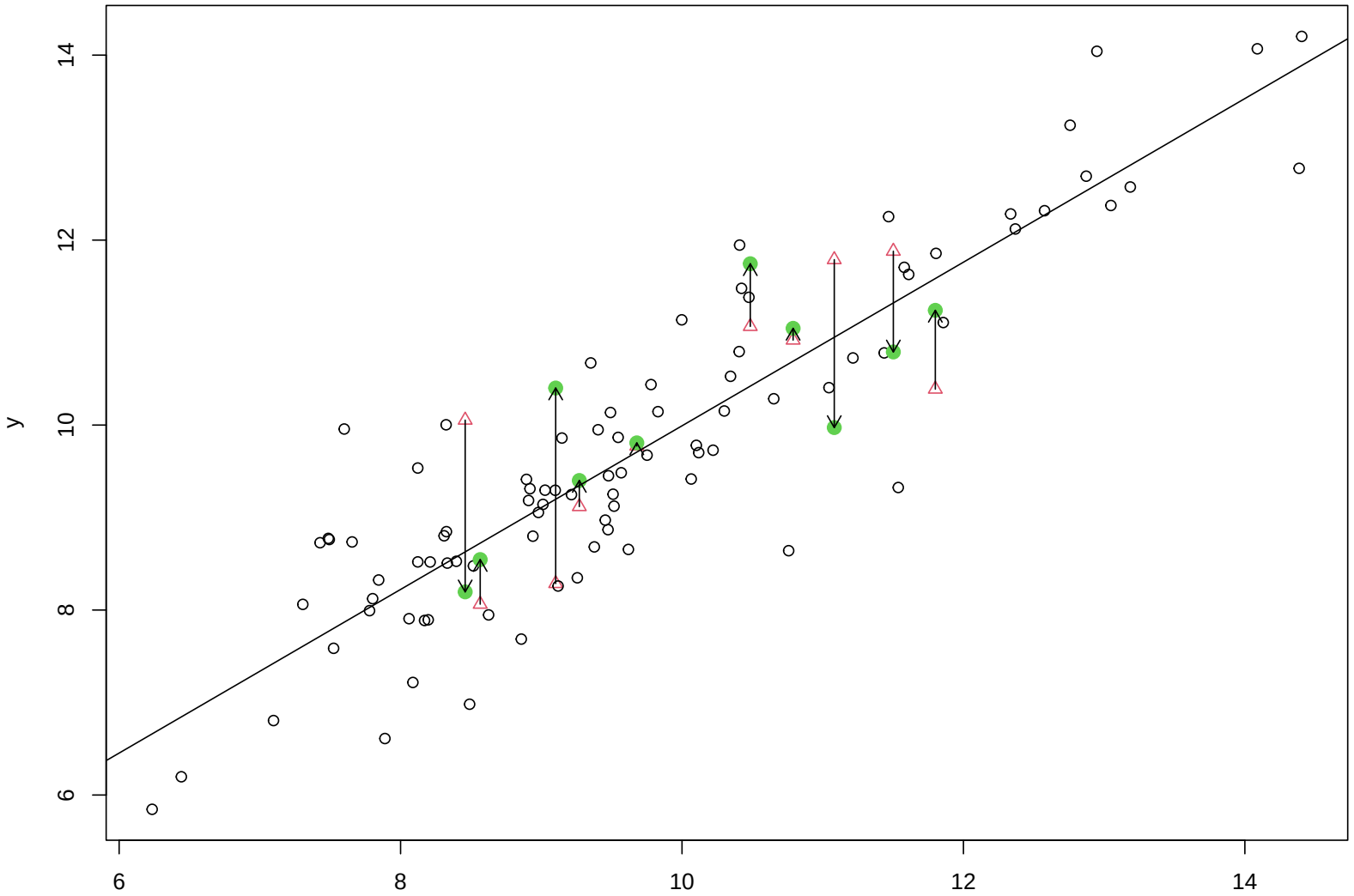
# Regression-based methods

Do not use the expected values for imputation:



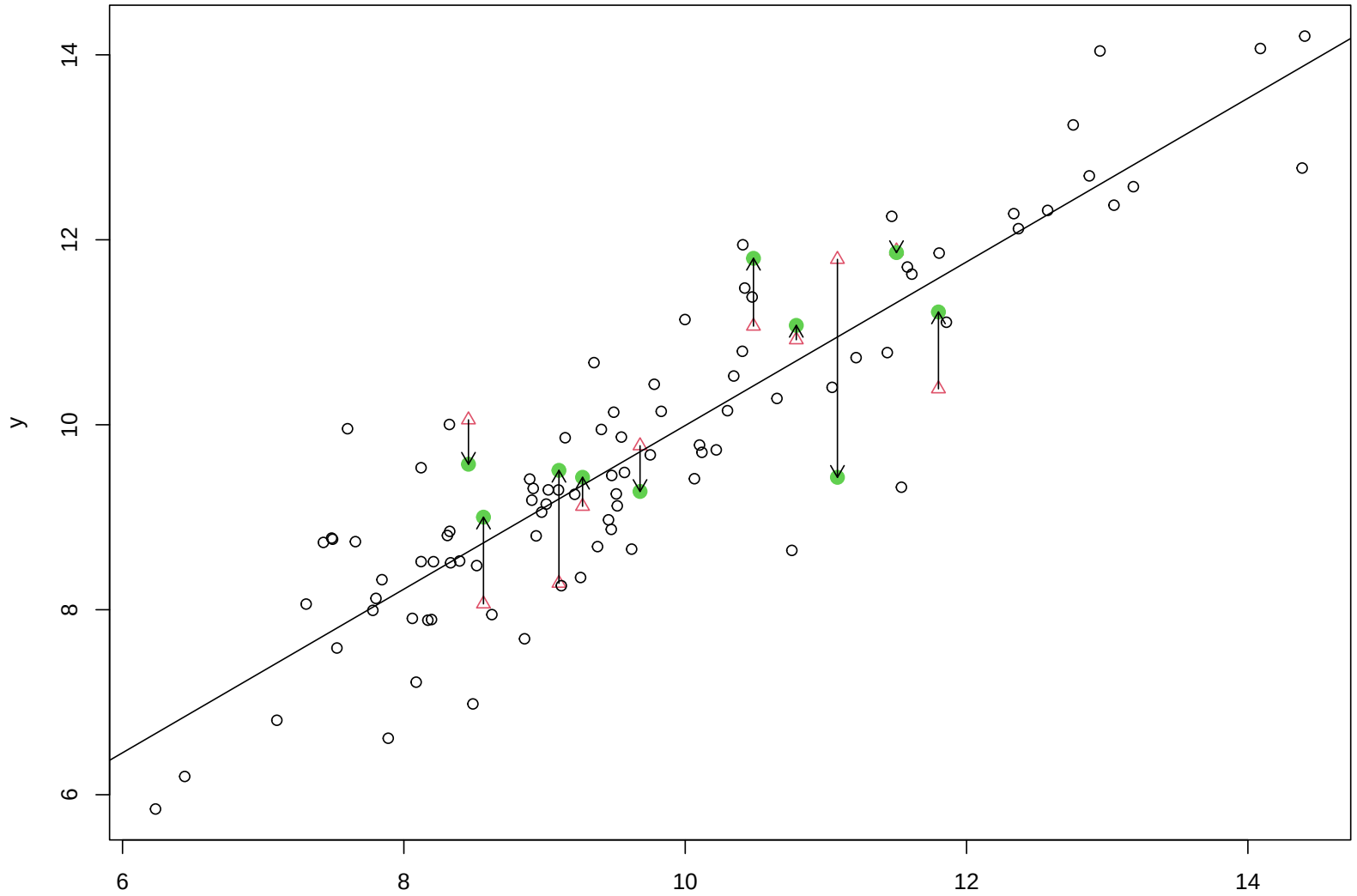
# Regression-based methods

Add stochastic noise (draw from predictive distribution, add normal noise to expected values, or draw from residuals)



# Regression-based methods

Add stochastic noise more than once (**multiple imputation**). Do this, e.g. 10 times to result in 10 imputed data sets.



# Selected imputation methods in R

Imputation methods in VIM (M. Templ et al. 2019)

- `hotdeck()` for **huge** data sets
- `kNN()` of mixed scaled variables
- `regressionImp()`
- `irmi()` for robust ("multiple'') EM-based imputation and potentially mixed scaled variables (M. Templ, Kowarik, and Filzmoser 2011)

# Selected imputation methods in R

Imputation methods in VIM (M. Templ et al. 2019)

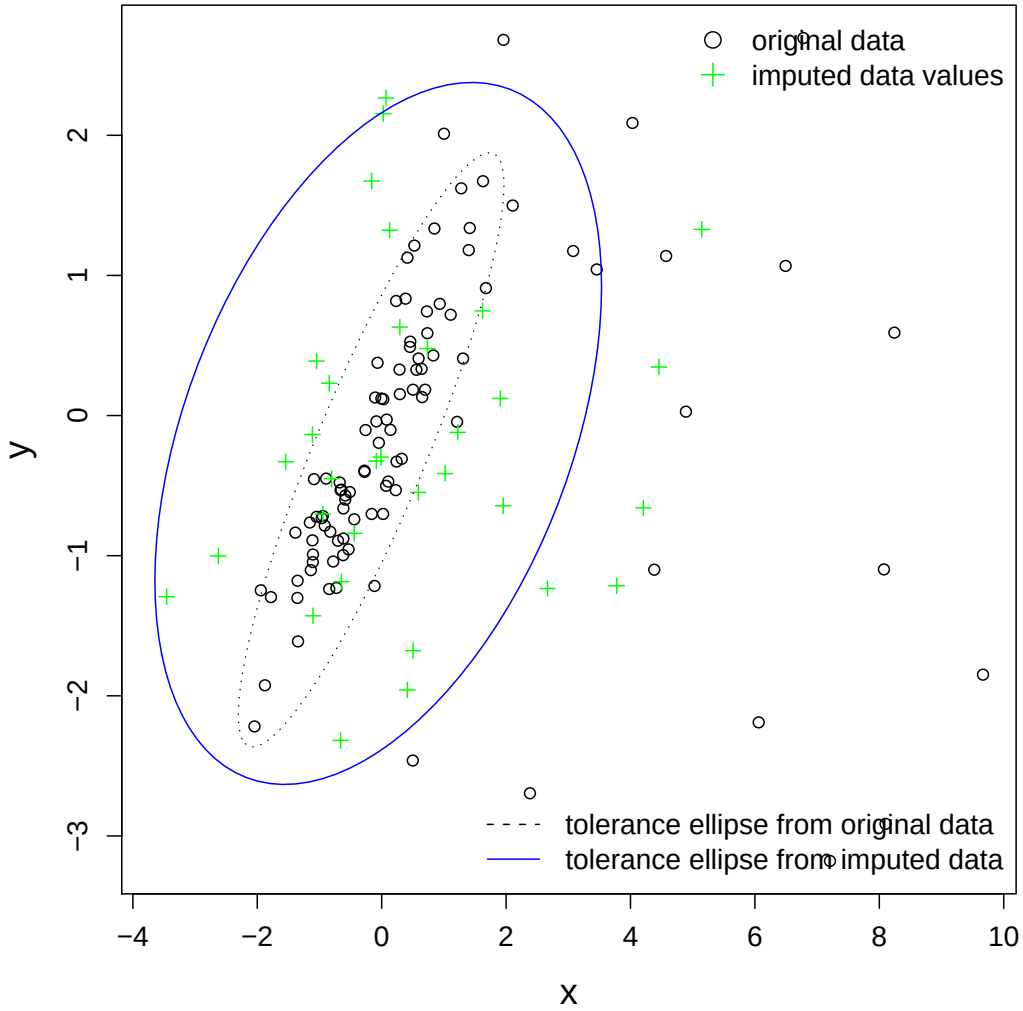
- `hotdeck()` for **huge** data sets
- `kNN()` of mixed scaled variables
- `regressionImp()`
- `irmi()` for robust ("multiple'') EM-based imputation and potentially mixed scaled variables (M. Templ, Kowarik, and Filzmoser 2011)

(Multiple) Imputation methods in other packages (selection)

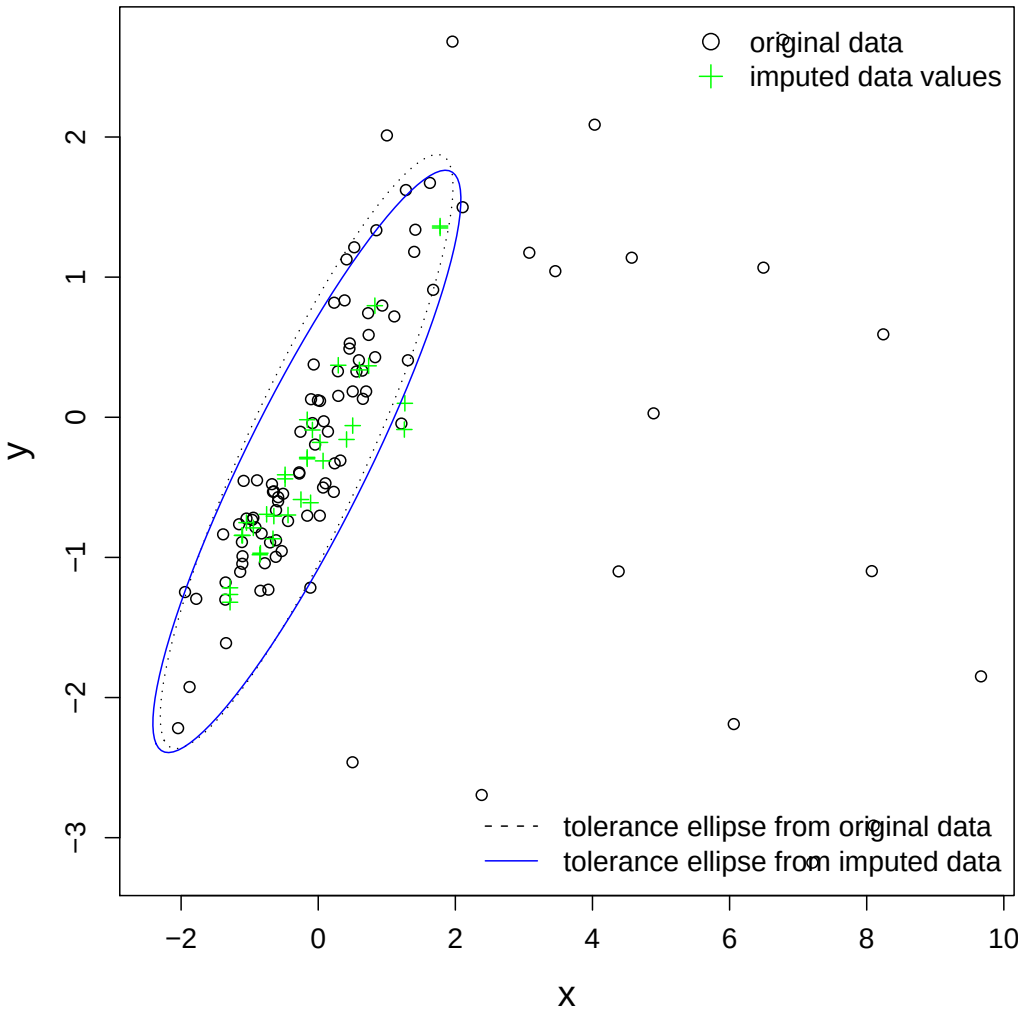
- `mi` in package `mi` (Gelman and Hill 2011)
- `mice` in package `mice` (van Buuren and Groothuis-Oudshoorn 2011), especially the pmm (predictive mean matching) method
- `missRanger` (Mayer 2019) for imputation with random forests
- `missMDA` (Josse and Husson 2016) for PCA-methods
- `deepImp` for using ANN's (Matthias Templ 2021)
- `impCoda` in `robCompositions` (Filzmoser, Hron, and Templ 2018) for imputing compositional data

# Non-robust Imputation with mi, mice, iveware, . . .

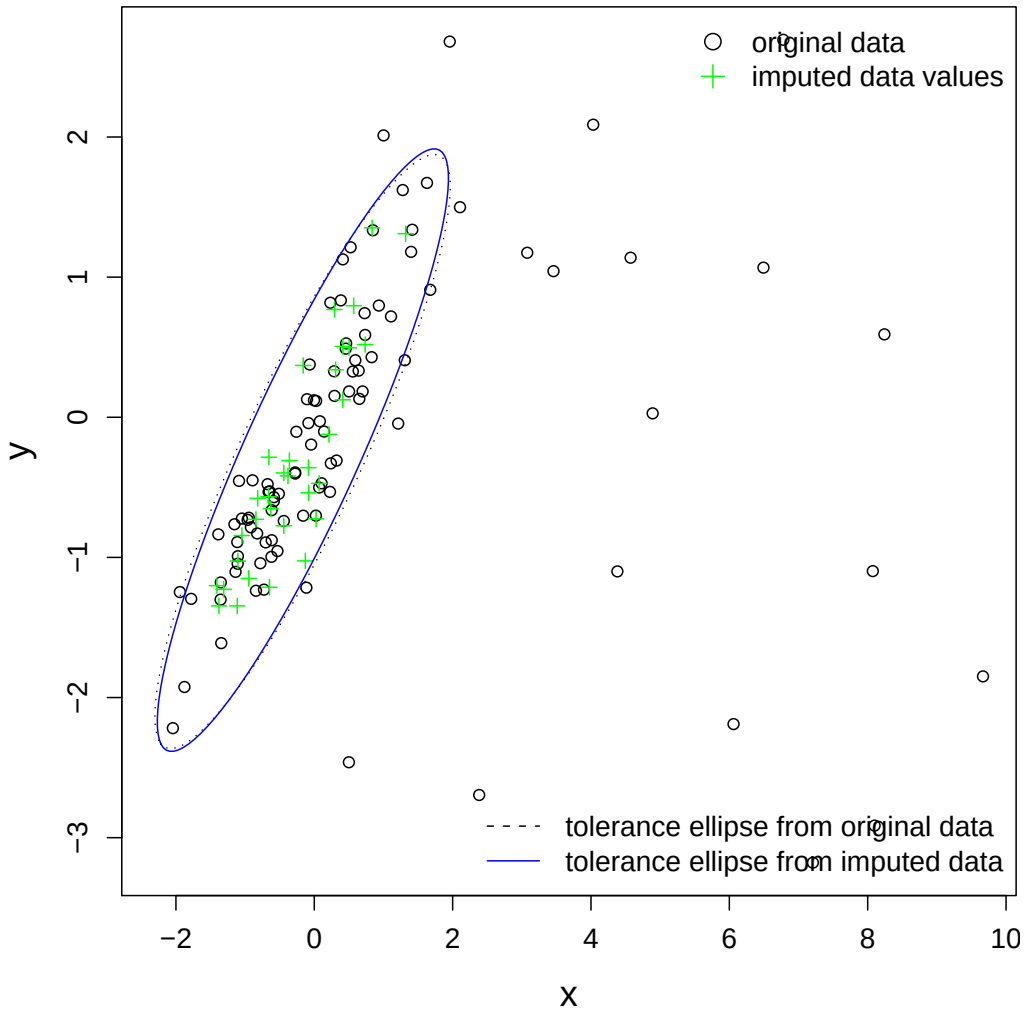
Storytelling based on a 2-dim toy data set . . .



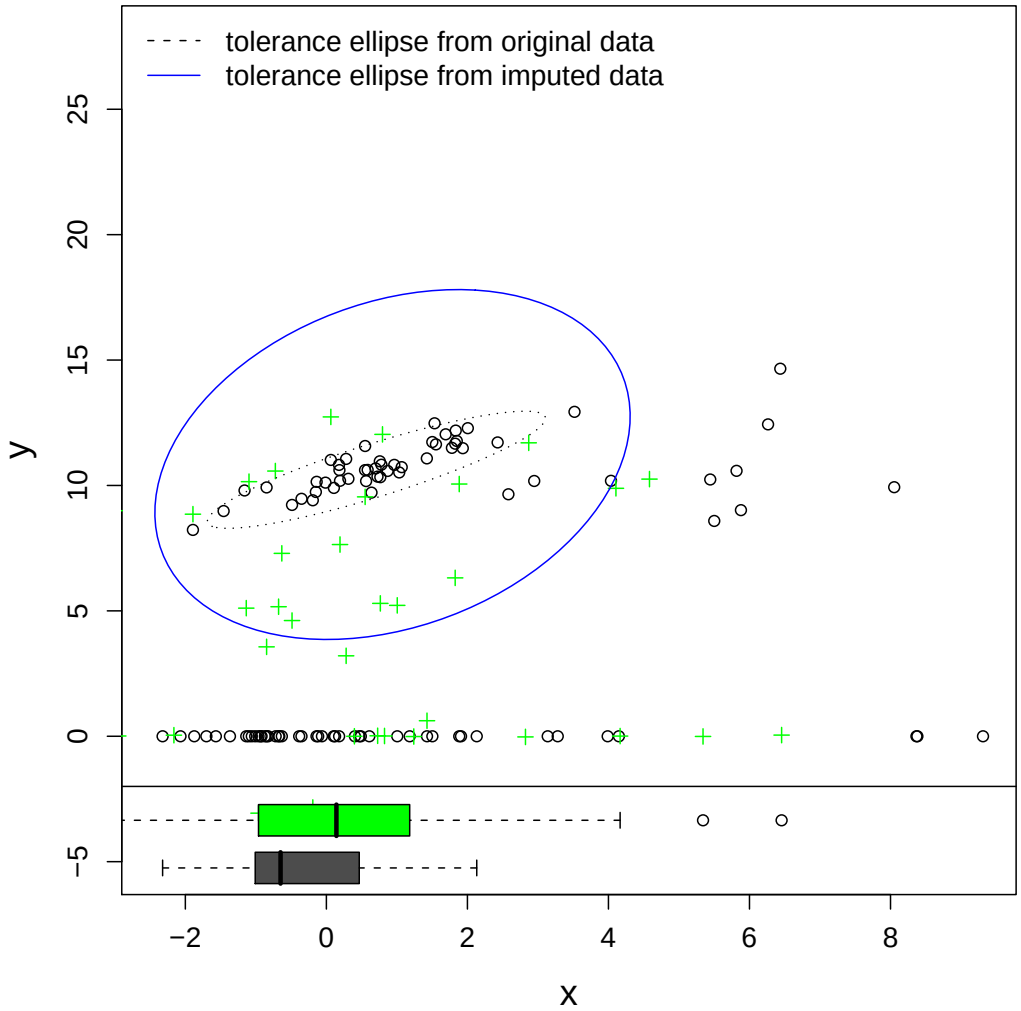
# Imputation with VIM (method kNN)



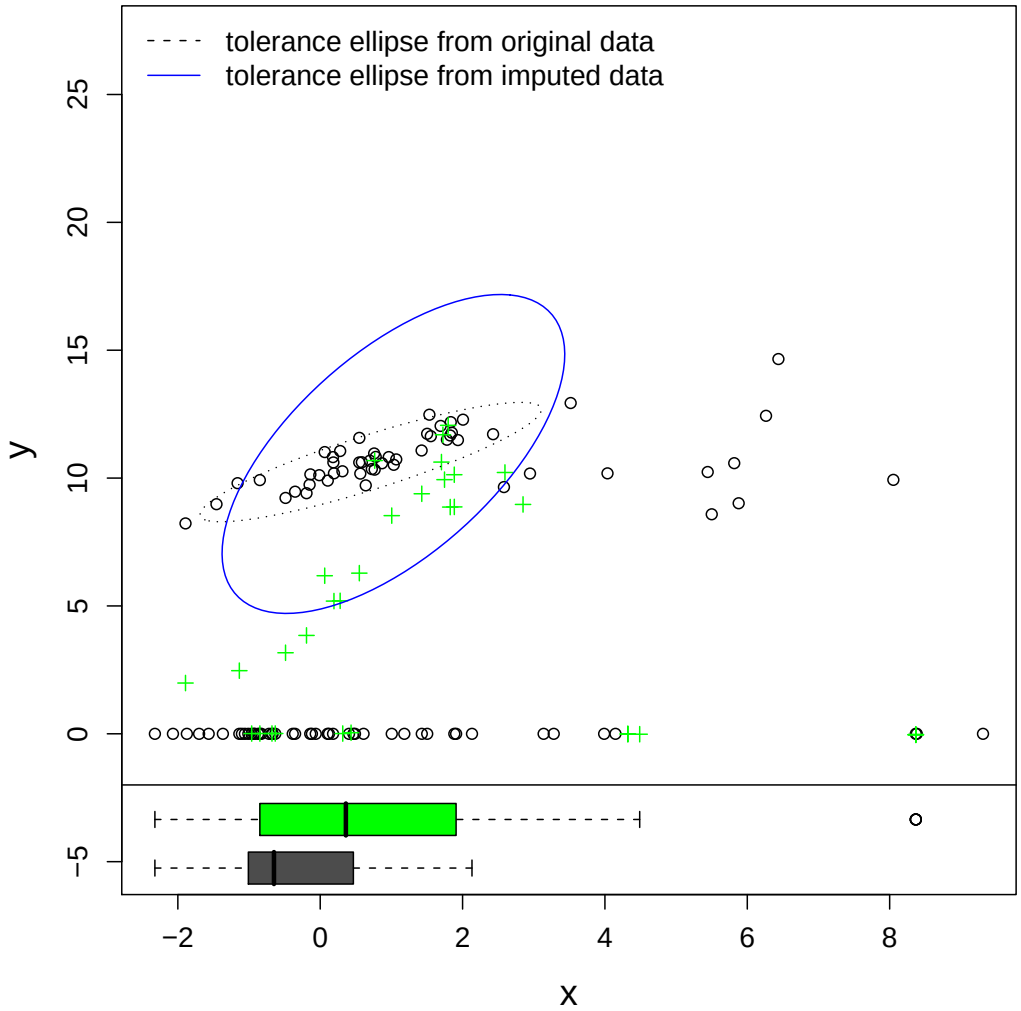
# Robust Imputation with VIM (method irmi)



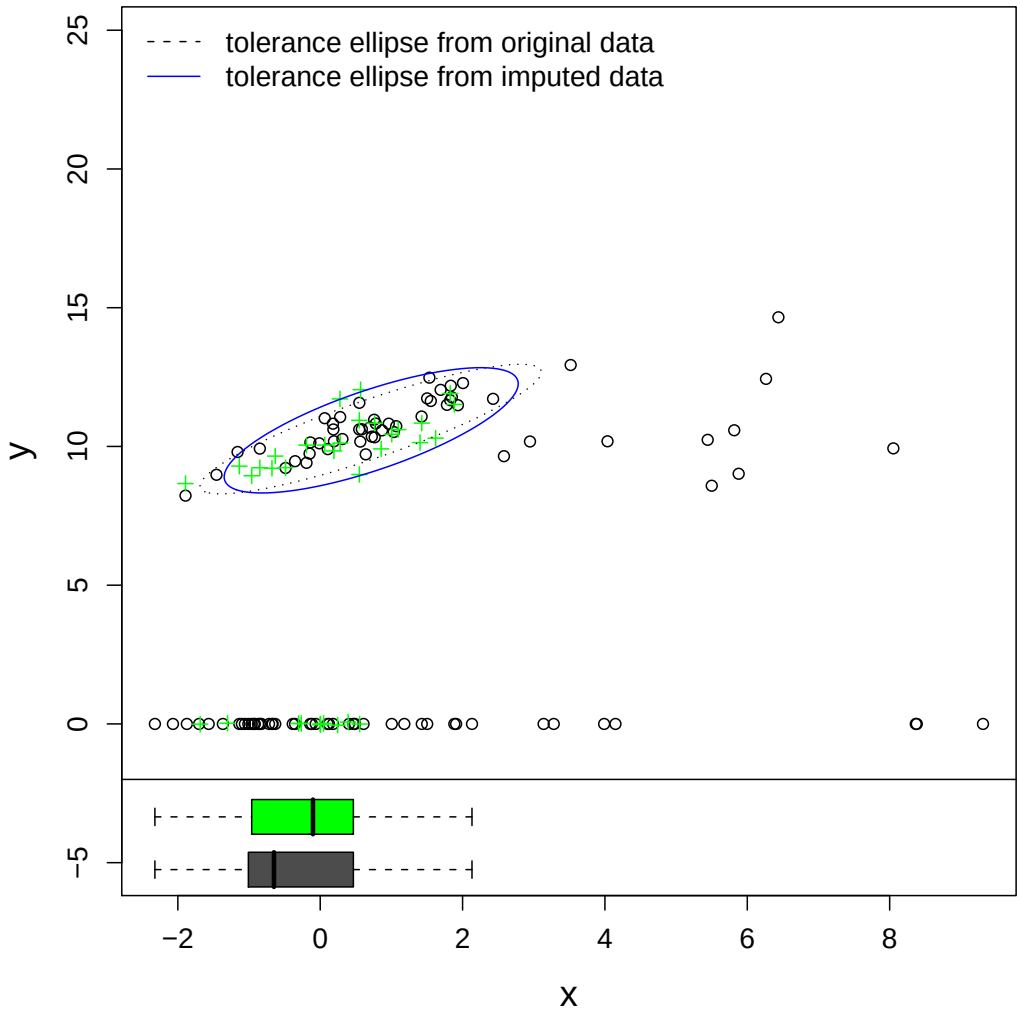
# Non-robust Imputation with mi, mice, iveware, . . .



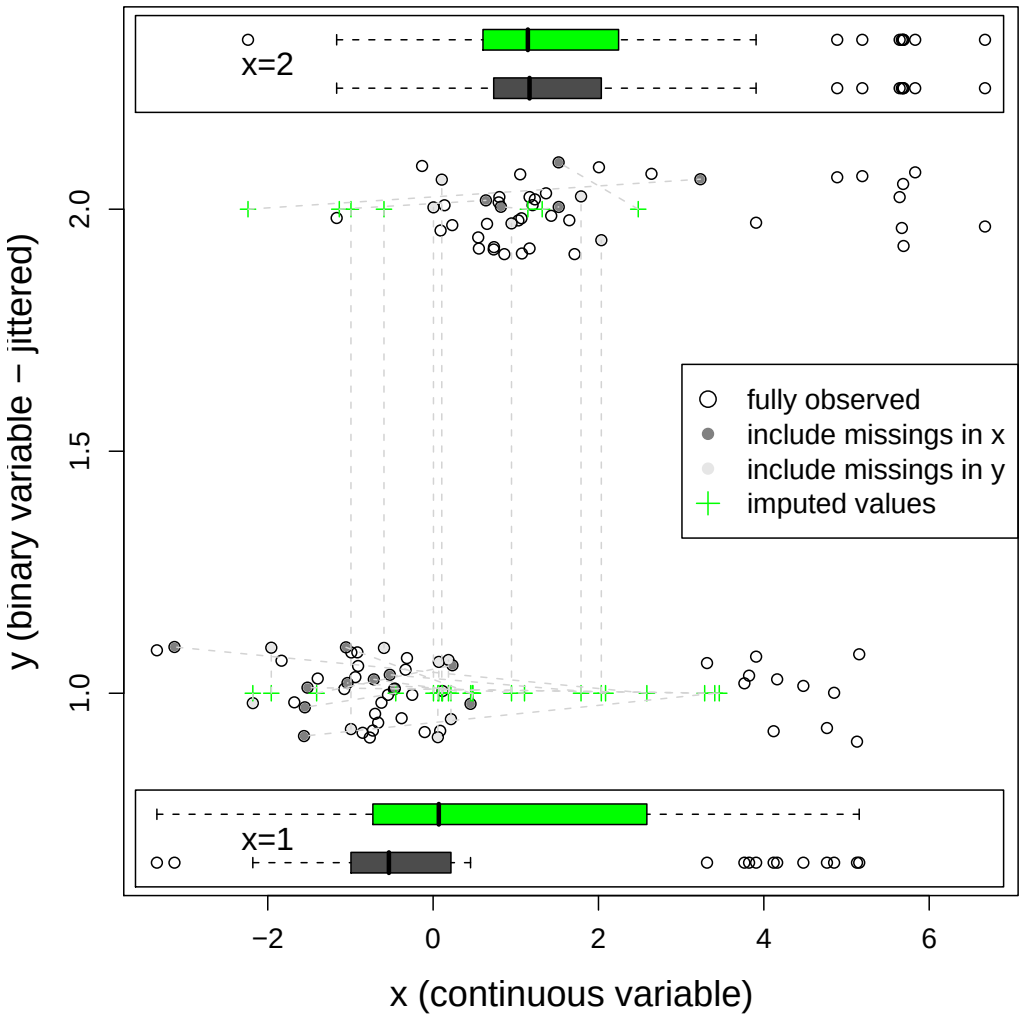
# Imputation with VIM (method kNN)



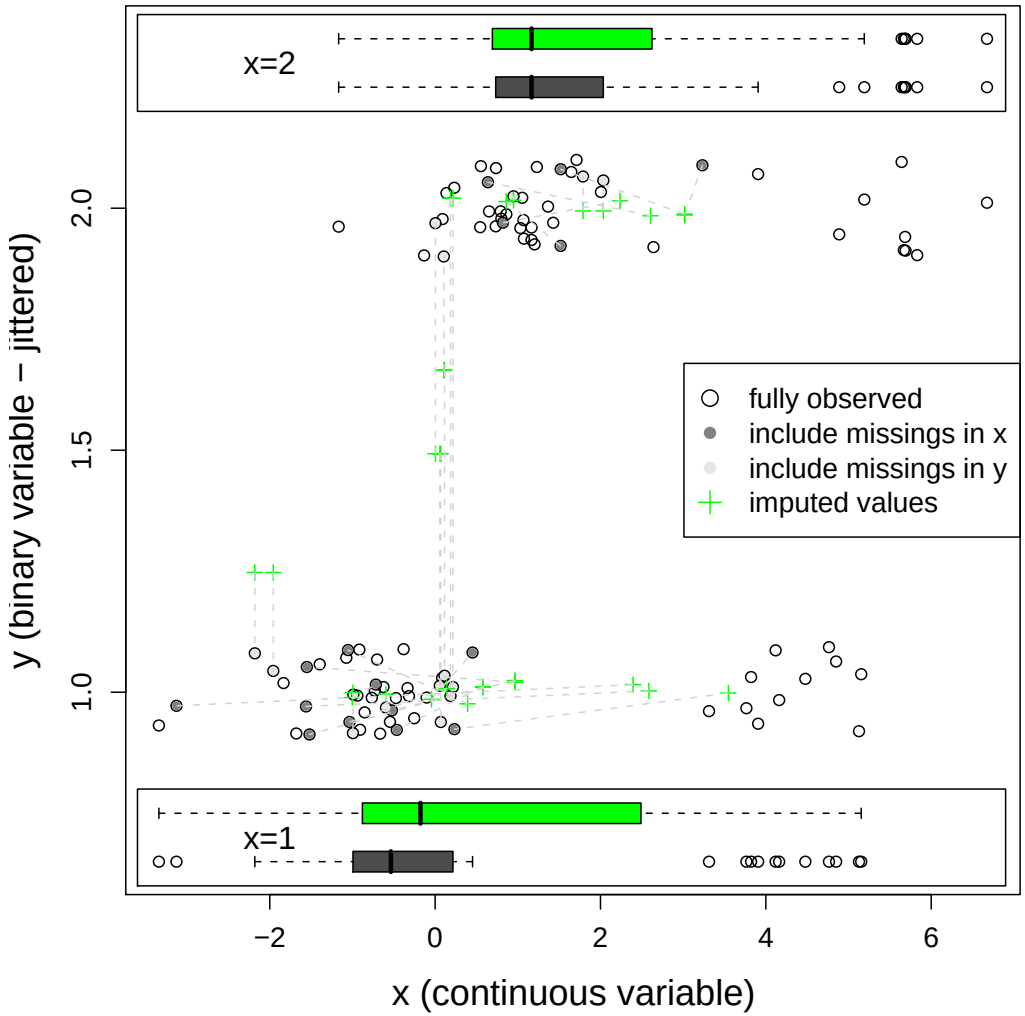
# Robust Imputation with VIM (method irmi)



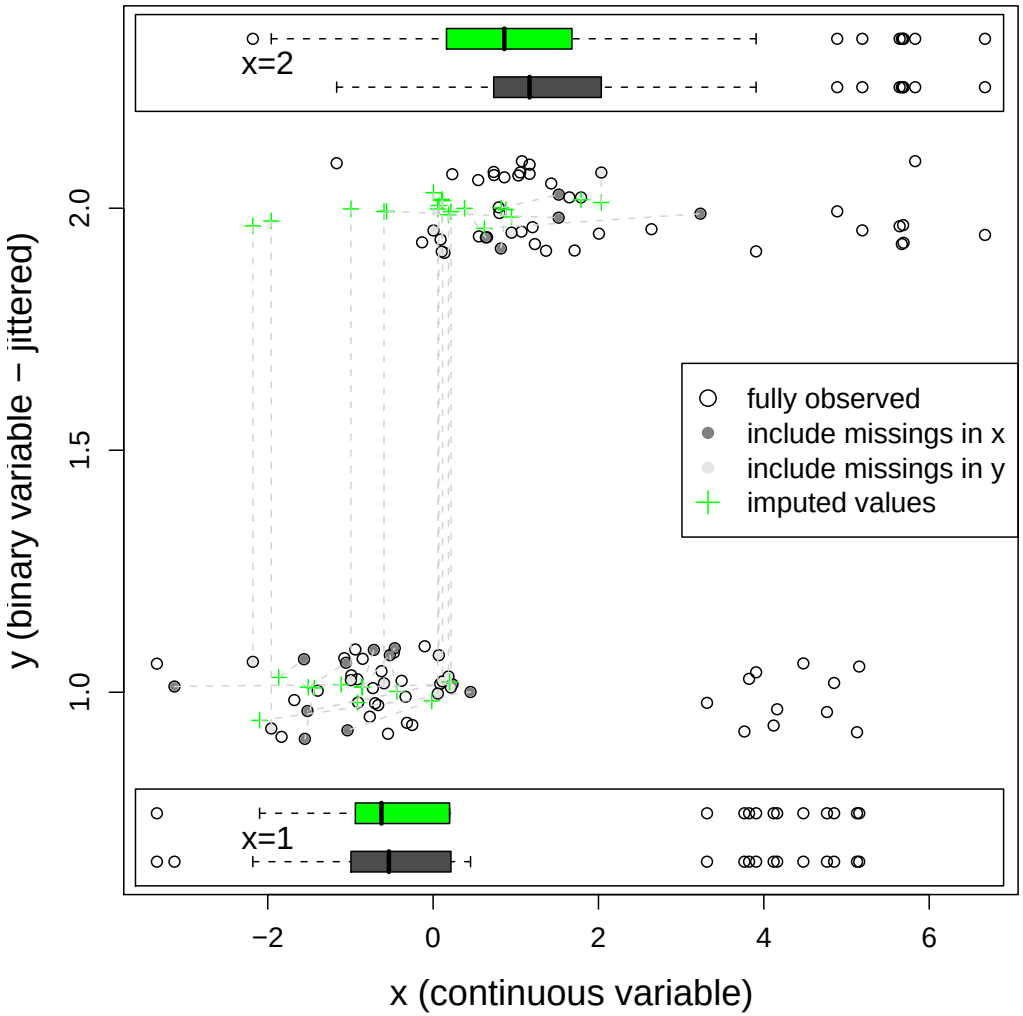
# Non-robust Imputation with mi, mice, iveware, . . .



# Imputation with VIM (method kNN)



# Robust Imputation with VIM (method irmi)



# MSE

- 100 obs. from a bivariate normal and a bivariate normal + a few moderate outliers.
- Moderate MAR situation.

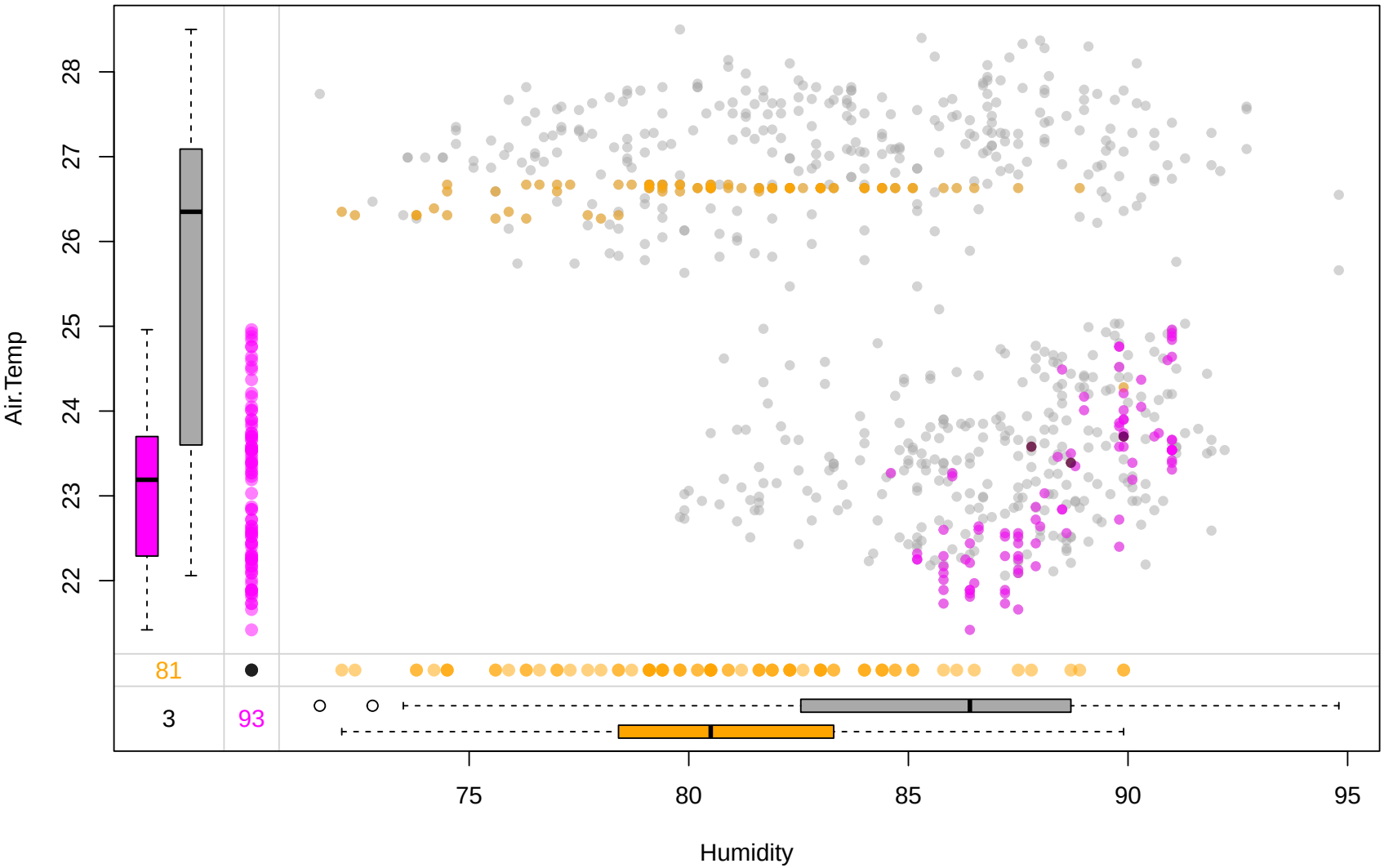
| method | mvn | mvn out |
|---|---|---|
| compl. case anal. | 0.01220 | 0.01230 |
| knn | 0.01072 | 0.01104 |
| imi | 0.01071 | 0.01397 |
| irmi robust | 0.01002 | 0.01024 |
| irmi mi robust | **0.00010** | **0.00010** |
| missForest | 0.00999 | 0.01171 |
| mice | 0.02366 | 0.14096 |

# So what is IRMI doing?

- Missing values are intitialized (using `kNN`)
- Inner loop imputes (updates) one variable after each other using robust regression methods and (robust) stochastic noise, depending on the scale of the variable to impute (binary, categorical, continuous, semi-continuous, count, . . . )
- Outer loop: repeat the imputations until convergence
- Multiple imputation

# After imputation

Use the same plots as before to check the quality of imputations, e.g.



# Conclusions

- Imputation is an important data pre-processing step
- Essential to have knowledge about imputation of missing values
- Missing values can change your results of your study!
  - if not treated: potential bias in estimates
  - if treated not well: potential bias in estimates and bias in the variance of the estimates
- Make a picture of the distribution of missing values in your data set first
- Use sophisticated imputation methods
  - there are much more, for special cases (e.g. time series, compositional data, . . . )

Software is available, e.g. R package `VIM` (M. Templ, Alfons, and Filzmoser 2012; M. Templ et al. 2019; Kowarik and Templ 2016)

# References I


Bojinov, I., N. Pillai, and D. Rubin. 2017. "Diagnosing Missing Always at Random in Multivariate Data." https://arxiv.org/abs/1710.06891.

Filzmoser, P., K. Hron, and M. Templ. 2018. *Applied Compositional Data Analysis. With Worked Examples in R*. Springer Series in Statistics. Springer Publishing, Cham, Switzerland.

Gelman, Andrew, and Jennifer Hill. 2011. "Opening Windows to the Black Box." *Journal of Statistical Software* 40.

Josse, J., and F. Husson. 2016. "missMDA: A Package for Handling Missing Values in Multivariate Data Analysis." *Journal of Statistical Software, Articles* 70 (1): 1–31. https://doi.org/10.18637/jss.v070.i01.

Kowarik, A., and M. Templ. 2016. "Imputation with the R Package VIM." *Journal of Statistical Software* 74 (7): 1–16. https://doi.org/10.18637/jss.v074.i07.

Li, C. 2010. "Tests of Homoscedasticity, Normality, and Missing Completely at Random for Incomplete Multivariate Data." *Psychometrika* 75 (4): 649–74.

———. 2013. "Little's Test of Missing Completely at Random." *The Stata Journal* 13 (4): 795–809.

Little, R. J. A. 1988. "A Test of Missing Completely at Random for Multivariate Data with Missing Values." *Journal of the American Statistical Association* 83 (404): 1198–1202.

# References II


Mayer, M. 2019. *MissRanger: Fast Imputation of Missing Values*. https://CRAN.R-project.org/package=missRanger.

Templ, M., A. Alfons, and P. Filzmoser. 2012. "Exploring Incomplete Data Using Visualization Techniques." *Advances in Data Analysis and Classification* 6 (1): 29–47. https://doi.org/DOI:%2010.1007/s11634-011-0102-y.

Templ, Matthias. 2021. "Artificial Neural Networks to Impute Rounded Zeros in Compositional Data." In *Festschrift*, edited by P. Filzmoser, K. Hron, J. A. Martin-Fernandez, and J. Palarea-Albaladejo. New York: Springer.

Templ, M., A. Kowarik, A. Alfons, and B. Prantner. 2019. *Visualization and Imputation of Missing Values*. http://CRAN.R-project.org/package=VIM.

Templ, M., A. Kowarik, and P. Filzmoser. 2011. "Iterative Stepwise Regression Imputation Using Standard and Robust Methods." *Computational Statistics & Data Analysis* 55 (10): 2793–2806.

van Buuren, S., and K. Groothuis-Oudshoorn. 2011. "mice: Multivariate Imputation by Chained Equations in R." *Journal of Statistical Software* 45 (3): 1–67. http://www.jstatsoft.org/v45/i03/.