Data Science Hackathon

Missing Values Treatment

Rianne Schouten

1. University Utrecht, Department of Methodology and Statistics
2. DPA Professionals, Data Science Excellence Program

January 18, 2017

Welcome

Introduction

- Rianne Schouten
- Missing Data Specialist

Welcome

Introduction

- Rianne Schouten
- Missing Data Specialist

What do you expect of today?

Welcome

Introduction

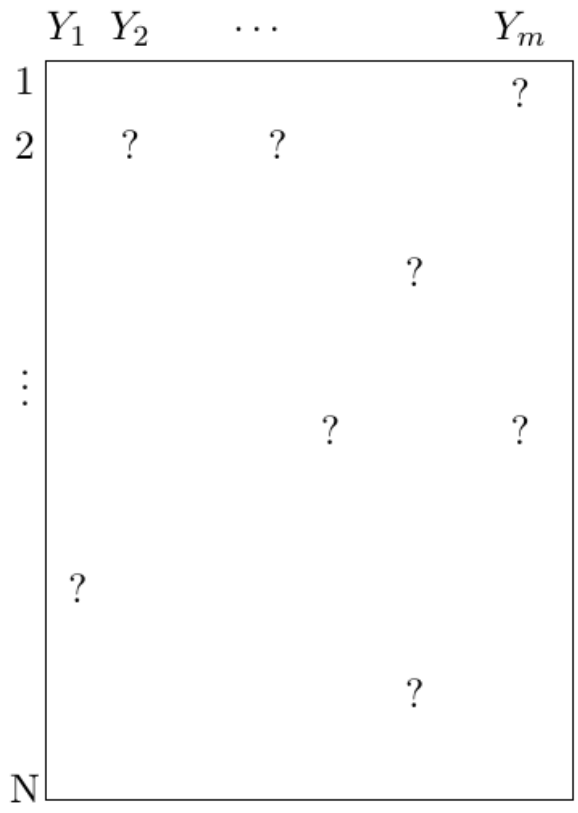
- Rianne Schouten
- Missing Data Specialist

What do you expect of today?

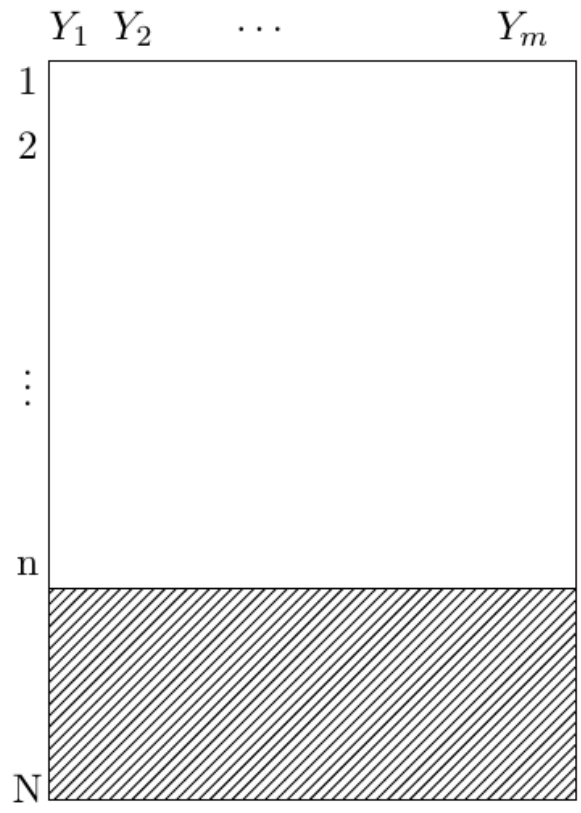
In this presentation:

- What is missing data?
- How to deal with missing data?
- Today's challenge

What is missing data?



Item nonresponse



Unit nonresponse

What is missing data?

```
head(inc_data)
```

```
##    outcome  feature
## 1 5.963154       NA
## 2 5.646671       NA
## 3 4.350638 10.23729
## 4 4.846355 10.25432
## 5 6.287034       NA
## 6 4.964498       NA
```

```
require(mice)
md.pattern(inc_data)
```

```
##     outcome feature
## 295       1       1   0
## 705       1       0   1
##           0     705 705
```

What is missing data?

- MCAR: Missingness is fixed, not related to any variable
- MAR: Missingness is related to an observed variable
- MNAR: Missingness is related to the missingness itself or to an unobserved variable

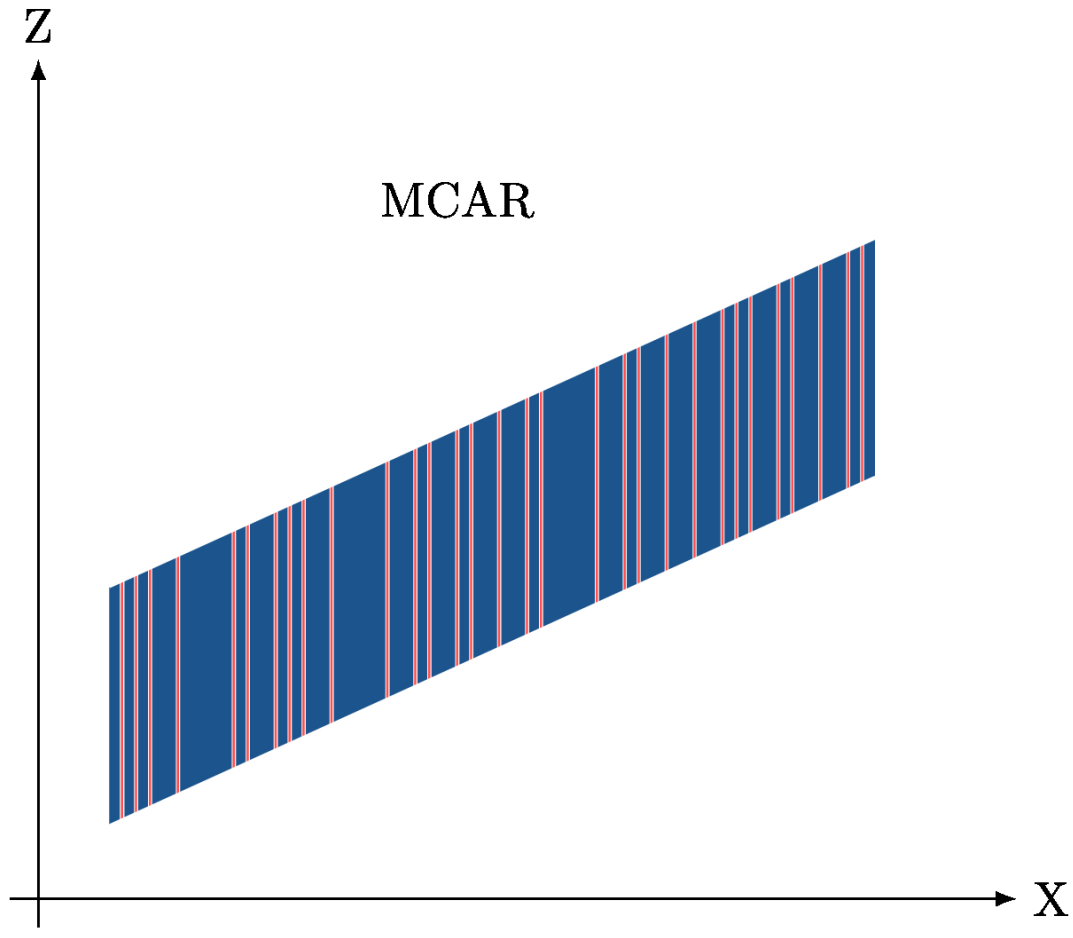
Example:

Consider outcome variable ‘income’ and feature ‘age’

- MCAR: Some age values are missing, both older and younger ages
- MAR: Age values are missing, especially for people with a high income
- MNAR: Age values are missing, especially for older people

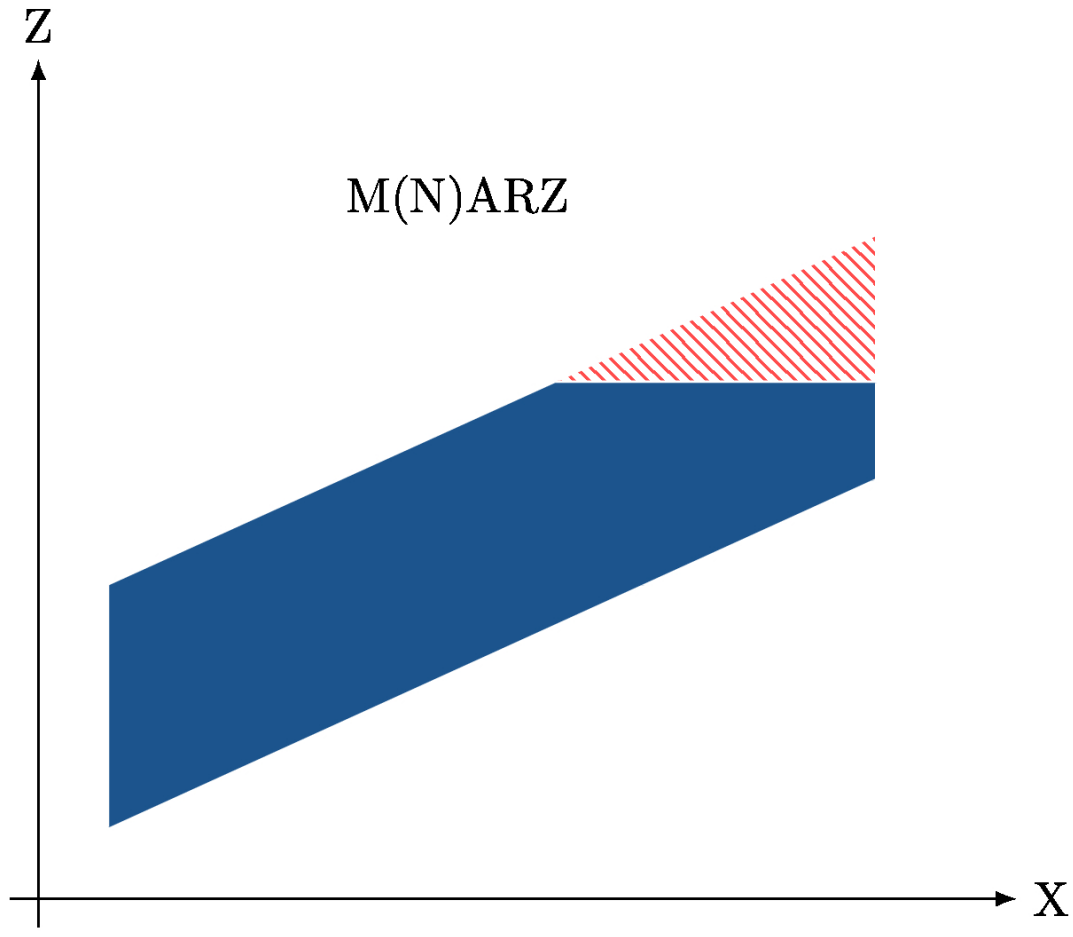
What is missing data: MCAR

Independent of value size, values on 'feature X' are missing



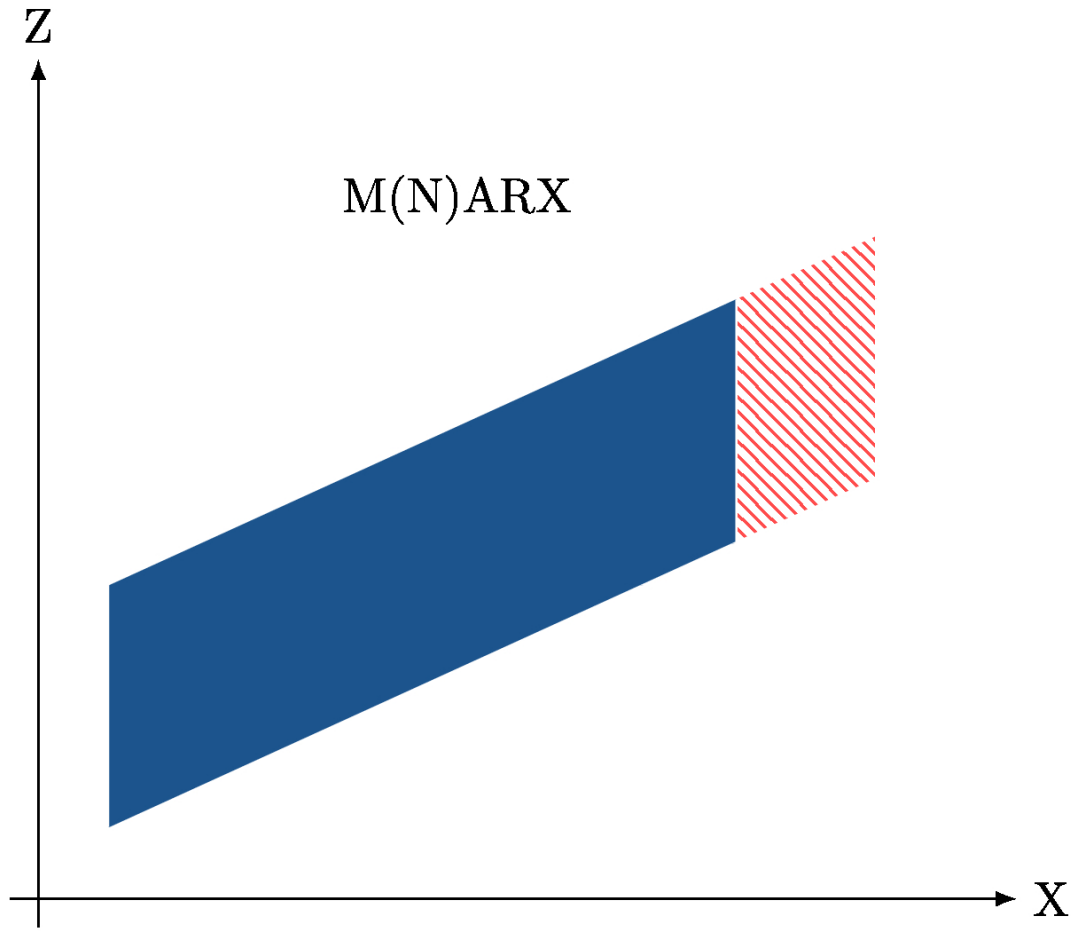
What is missing data: MAR and MNAR based on Z

Records with a large value on 'z' are missing on 'feature X'

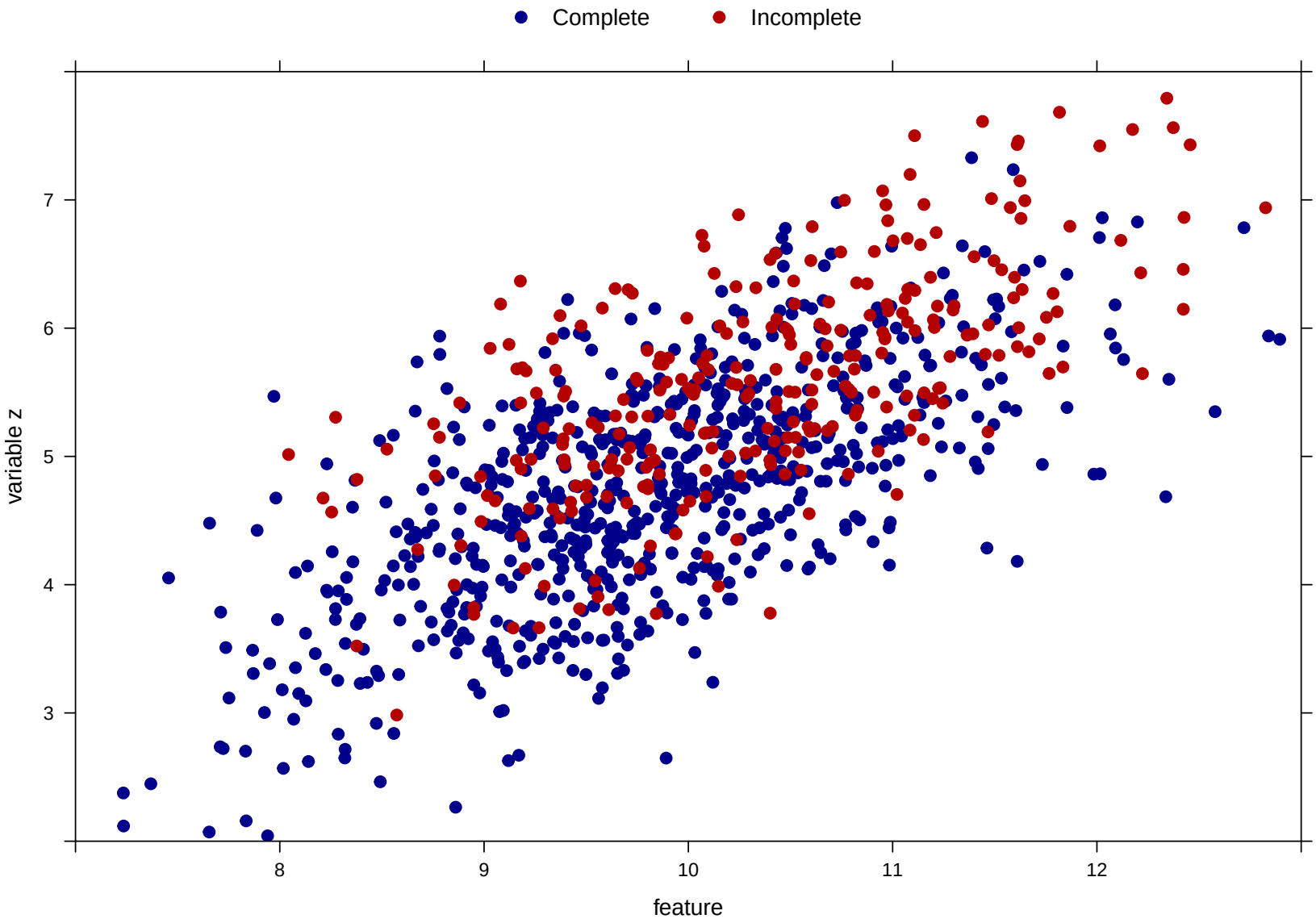


What is missing data: MAR and MNAR based on X

Records with a large value on 'feature X' are missing on 'feature X'



What is missing data?



How to deal with missing data?

1. Drop incomplete rows/columns
2. Imputation

- random imputation
- mean/median imputation
- regression imputation
- random forest imputation
- multiple imputation
- and more...

3. Other methods such as

- weighting procedures
- likelihood based methods
- and more...

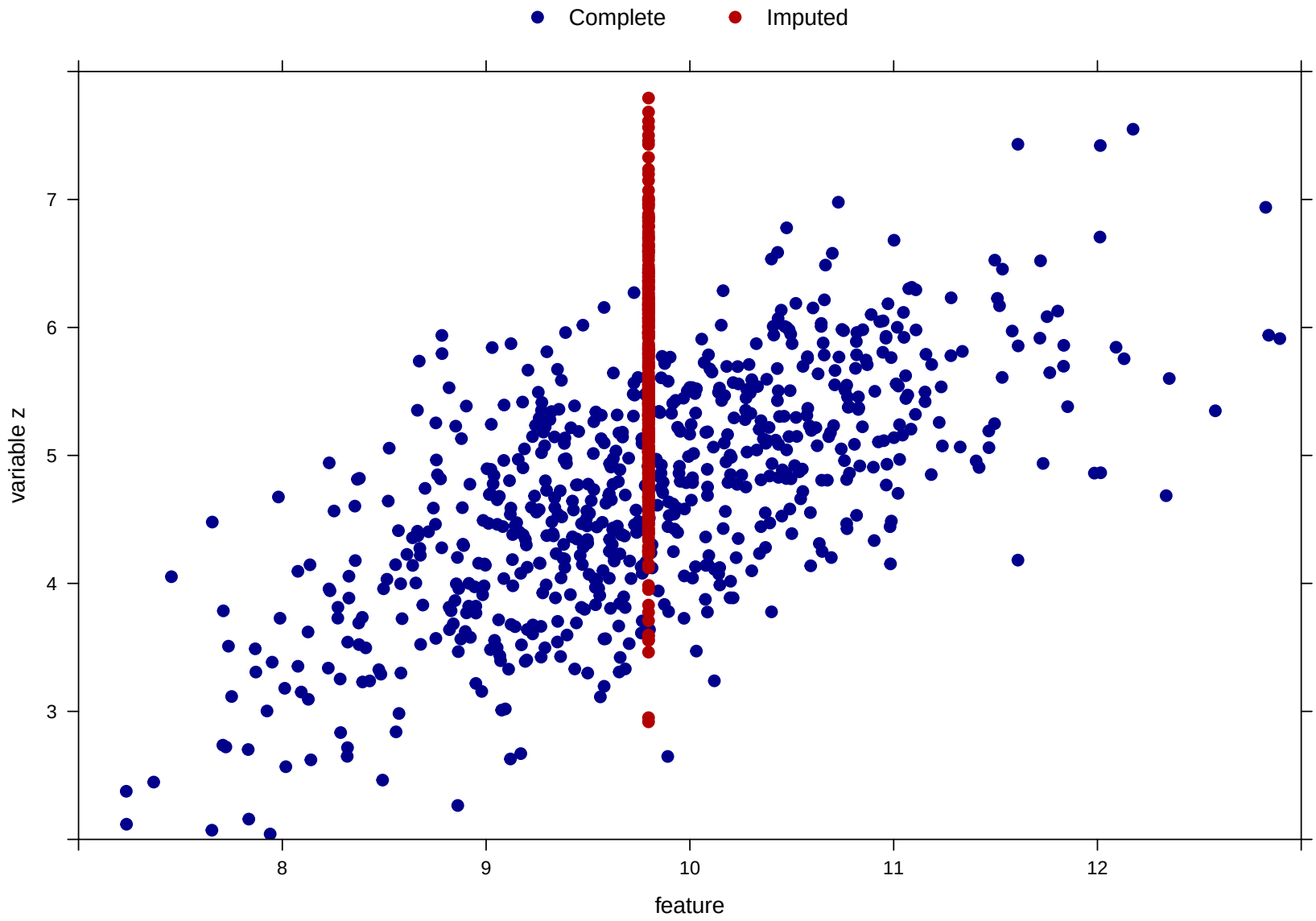
How to deal with missing data: mean imputation

```
head(inc_data)
```

```
##    outcome feature
## 1 5.963154      NA
## 2 5.646671      NA
## 3 4.350638      NA
## 4 4.846355      NA
## 5 6.287034      NA
## 6 4.964498      NA
```

```
com_data <- inc_data
com_data[is.na(inc_data$feature), 'feature'] <-
  mean(inc_data$feature, na.rm = TRUE)
```

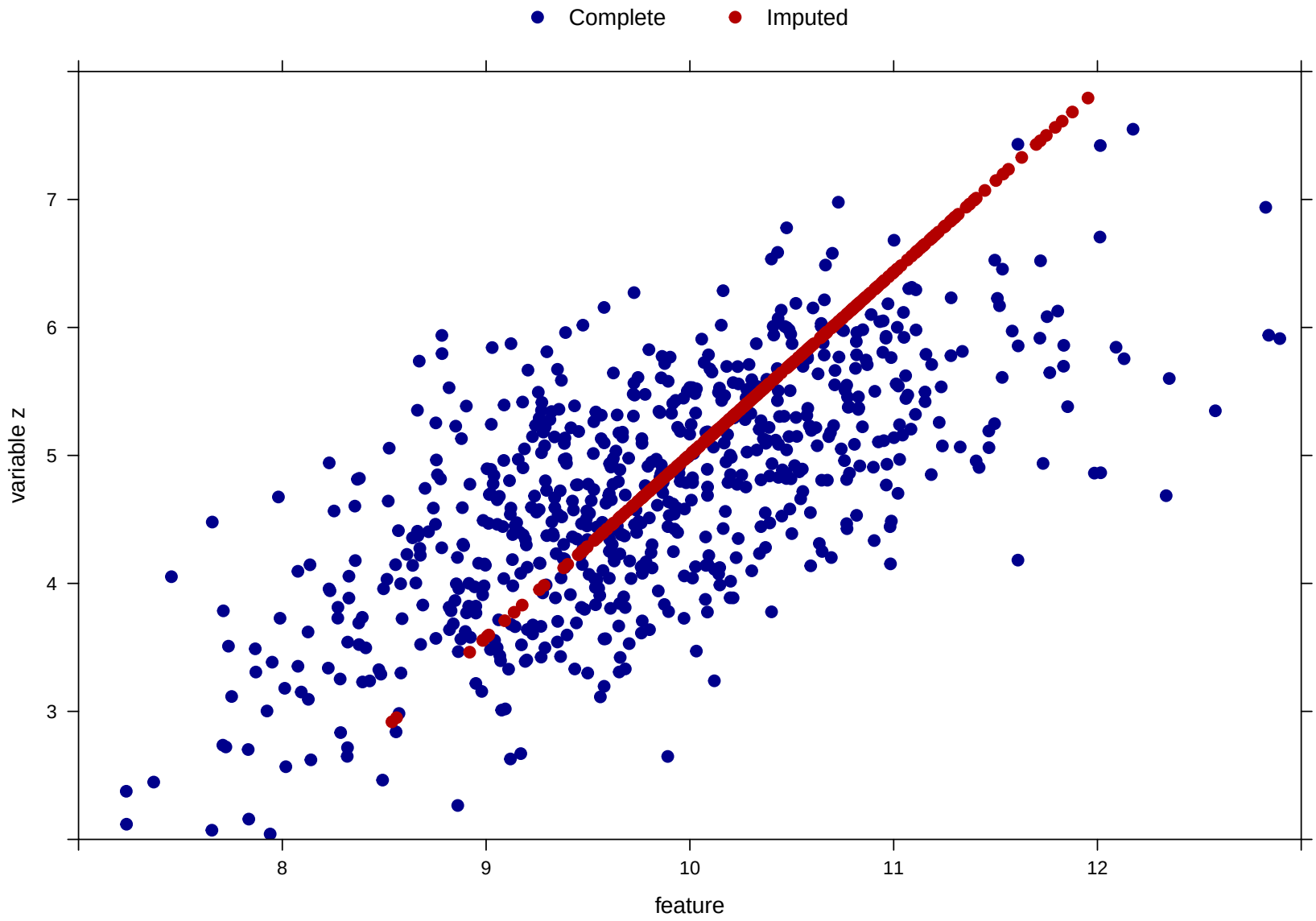
How to deal with missing data: mean imputation



How to deal with missing data: regression imputation

```
fit <- lm(feature ~ z, data = inc_data)
pred <- predict(fit, newdata = ic(inc_data))
com_data <- inc_data
com_data[is.na(inc_data$feature), 'feature'] <- pred
```

How to deal with missing data: regression imputation



Today's challenge

You receive:

- an incomplete training dataset
- outcome variable (dummy) is complete
- missingness in training and testset is comparable
- combination of MCAR, MARX and MNARX

Make sure:

- you fit your imputation method on your trainingset
- and transform/apply on testset

Question:

- How would you evaluate whether your imputation method is okay?

Contact information

Ask me anything, always:
Rianne Schouten, r.m.schouten@uu.nl, rianne.schouten@dpa.nl

Follow my work: rianneschouten.github.io

Work in progress