ENERGY DATA SCIENCE

Data processing: Cleaning

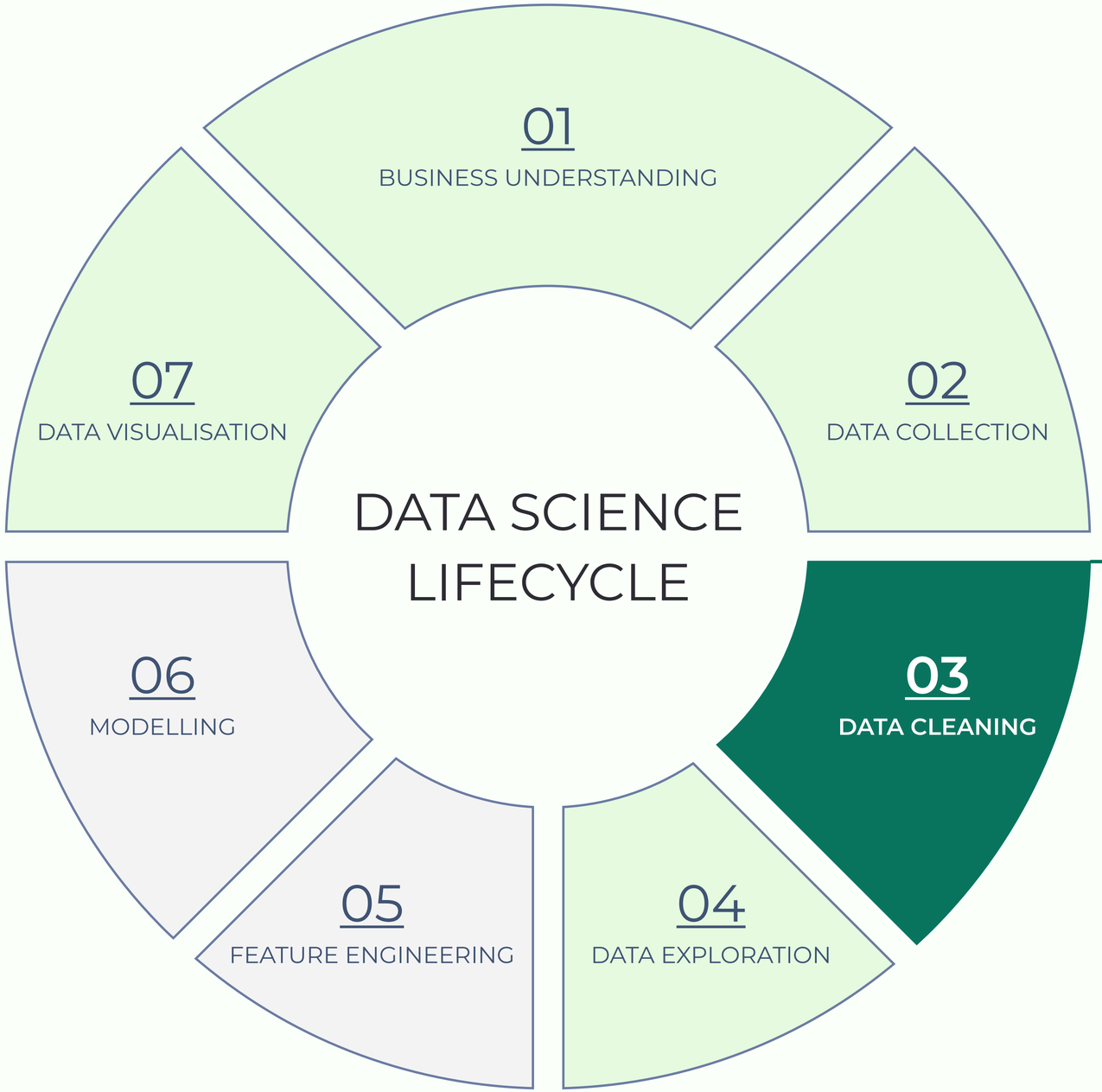
Prof. Juri Belikov

Department of Software Science
Tallinn University of Technology
juri.belikov@taltech.ee

PREVIOUSLY IN ITS8080 ...

Key takeaways:

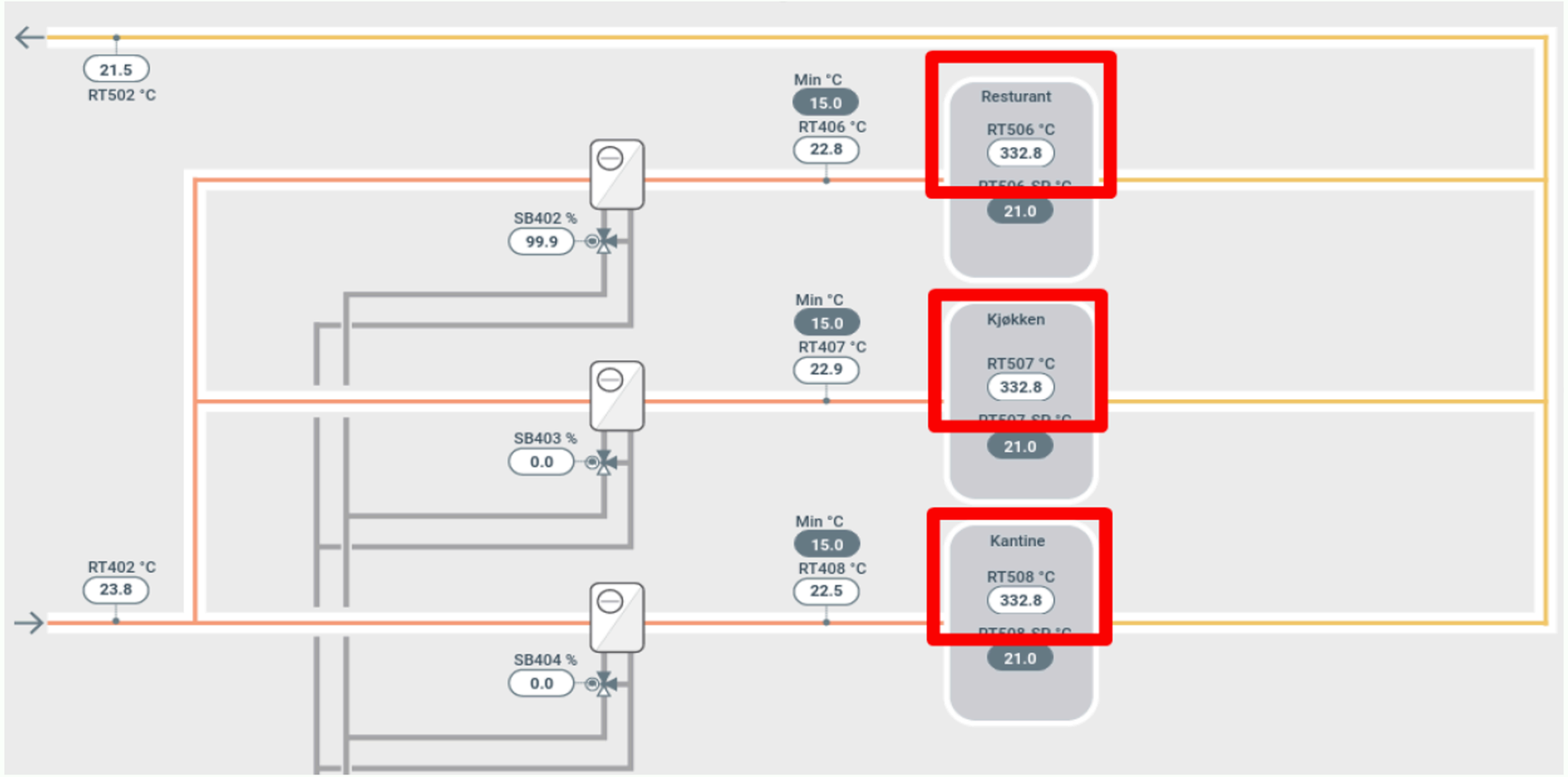
- Basics of typography
- Data visualisation
- Graphic forms and their uses
- Ethics in data visualisation



Remove data that does not belong in your dataset. Fix inconsistencies within the data and handle missing values.

Illustrative examples

ROOM TEMPERATURE SENSORS ARE STATIC

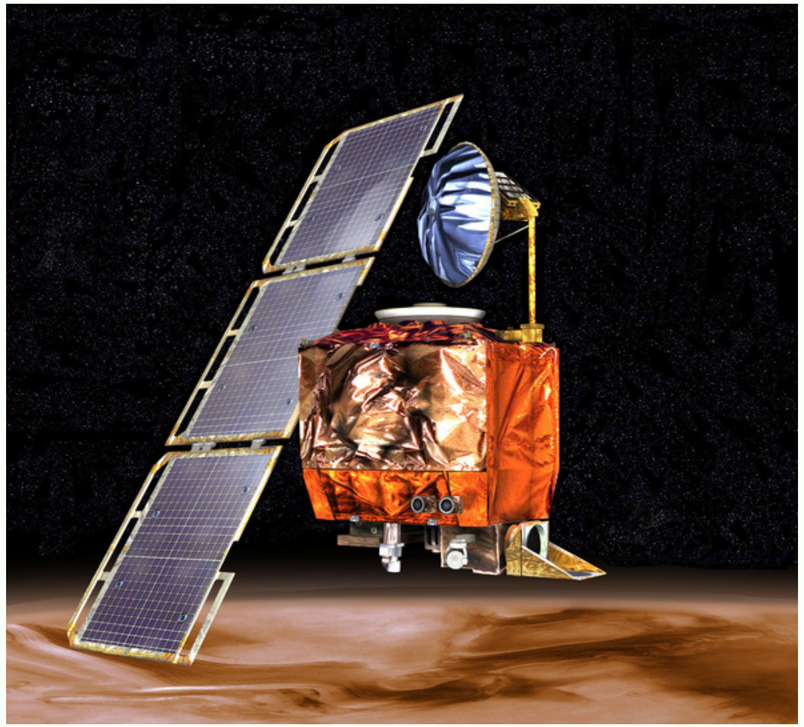


MARS CLIMATE ORBITER

On September 23, 1999, NASA's Mars Climate Orbiter entered Mars' atmosphere too low and was destroyed.

A navigation error occurred because one engineering team used imperial units (pound-seconds), while another expected metric units (newton-seconds).

The $327 million spacecraft was lost shortly before it was supposed to begin orbiting Mars.



[Accessed April 4, 2025], https://llis.nasa.gov/llis_lib/pdf/1009464main1_0641-mr.pdf

BOEING 737 MAX

Two Boeing 737 MAX airplanes crashed.

- Indonesia, October 2018
- Ethiopia, March 2019

A faulty sensor fed bad data to MCAS (Maneuvering Characteristics Augmentation System), making it falsely detect a stall.

Over $20 billion in costs and settlements.

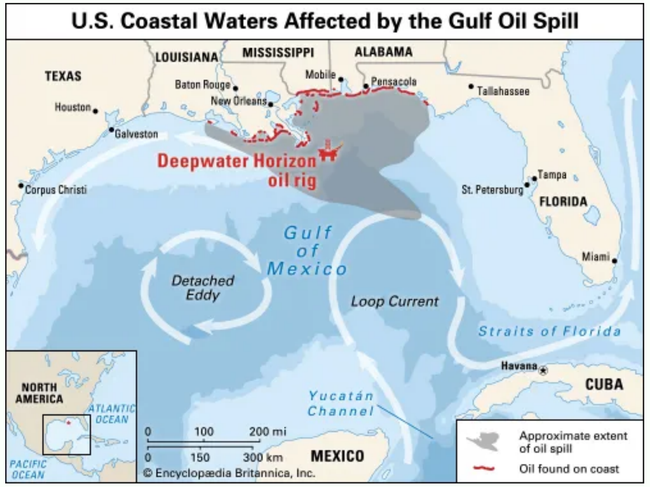
[Accessed April 4, 2025], https://eu.usatoday.com/story/opinion/2019/03/12/ethiopian-boeing-737-max-8-crash-safety-cause-indonesia-column/3133013002/

DEEPWATER HORIZON

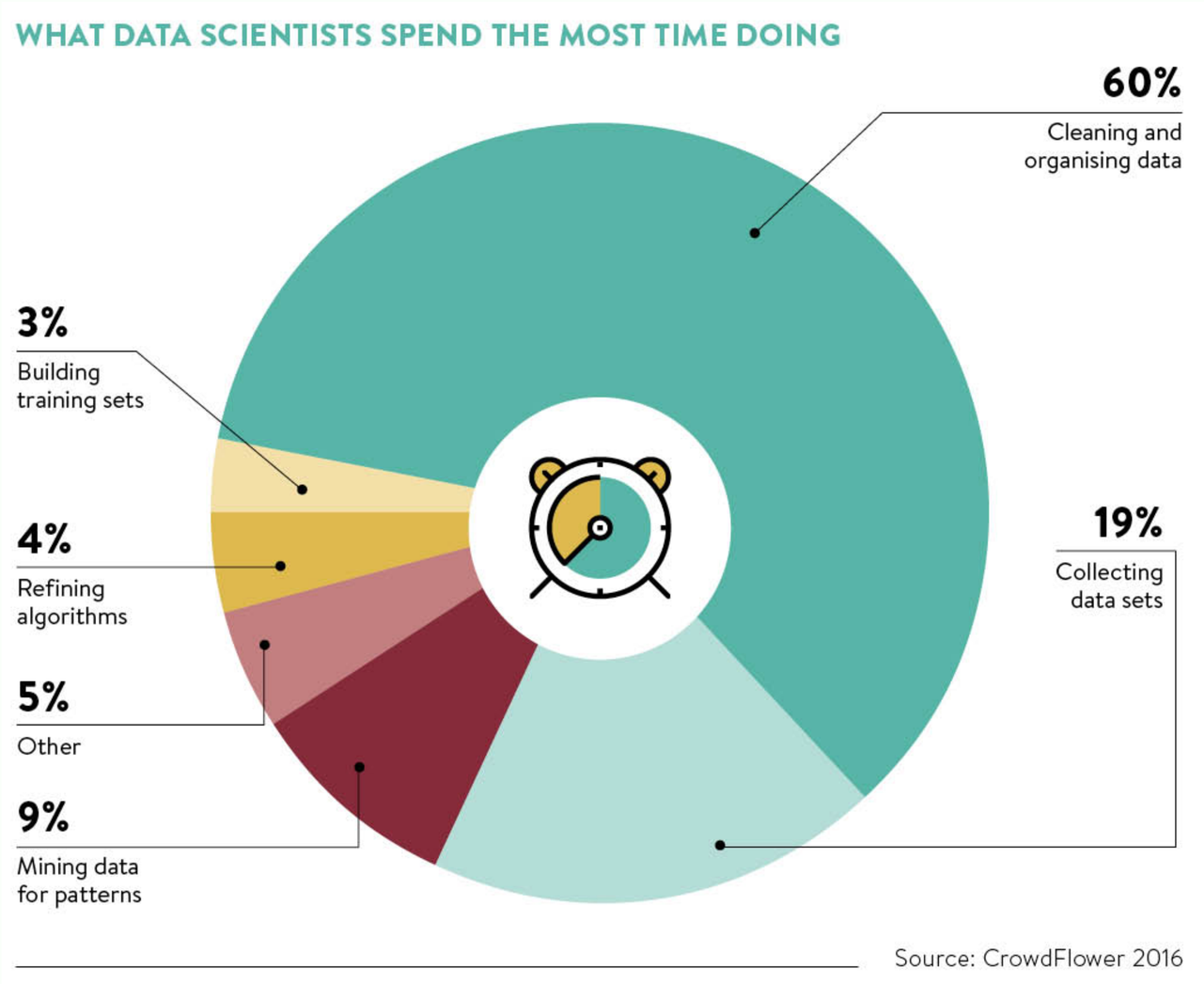
On April 20, 2010, a blowout occurred during drilling, causing an explosion on the platform.

Pressure test misinterpretation.

BP paid over $60 billion in fines, cleanup costs, etc.



[Accessed April 4, 2025], https://www.britannica.com/event/Deepwater-Horizon-oil-spill



[Accessed September 13], https://lewtun.github.io/dslectures/lesson03_data-cleaning/

DATA TYPES

Categorical (qualitative):

- Nominal: no quantitative value, no ordering, used to label variables.
  Example - Fuel type: {oil, diesel, gas, …}.

- Ordinal: discrete and ordered units.
  Example – Efficiency rating: {1-very poor, 2-poor, 3-average, 4-good, 5-excellent}.

DATA TYPES (2)

Numerical (quantitative):

- Discrete: countable, but not measurable.

  Example: coin flips, number of cars.

- Continuous: not countable, but measurable.

  Example: temperature, height.

- Interval scale: ordered, can go below zero.

  Example: temperature in C or F.

- Ratio: ordered, cannot go below zero.

  Example: height, weight, pressure.

VISUALISATION TYPES: MEMO

Data type	Bar chart	Pie chart	Histogram	Scatter plot		Heatmap		Box/violin plot		Line plot	
				Dim 1	Dim 2	Dim 1	Dim 2	Dim 1	Dim 2	Dim 1	Dim 2
Nominal	✓	✗	✗	✗	✗	✗	✗	✓	✗	✗	✗
Ordinal	✓	✗	✗	✗	✗	✗	✗	✓	✗	✗	✗
Interval	✗	✗	✓	✓	✓	✓	✓	✗	✓	✓	✓
Ratio	✗	✗	✓	✓	✓	✓	✓	✗	✓	✓	✓

DATA CLEANING

DATA CLEANING

Remove duplicates

When you combine data sets from multiple sources, there is a chance to create duplicate data.

Total electricity production [GWh], 2022

Fuel type	Production
Oil shale, thousand t	5,078
Natural gas, million m3	29
Hydro energy	23
Wind energy	668
Solar energy	596
Solar energy	596

Sample of electricity production/demand data [MWh]

Date (Estonia time)	Consumption
30.10.2022 00:00	768
30.10.2022 01:00	748.9
30.10.2022 02:00	734.1
30.10.2022 03:00	724.2
30.10.2022 03:00	715.4
30.10.2022 04:00	718.8

DATA CLEANING

Remove duplicates

When you combine data sets from multiple sources, there is a chance to create duplicate data.

Total electricity production [GWh], 2022

Fuel type	Production
Oil shale, thousand t	5,078
Natural gas, million m3	29
Hydro energy	23
Wind energy	668
Solar energy	596
Solar energy	596

Sample of electricity production/demand data [MWh]

Date (Estonia time)	Consumption
30.10.2022 00:00	768
30.10.2022 01:00	748.9
30.10.2022 02:00	734.1
30.10.2022 03:00	724.2
30.10.2022 03:00	715.4
30.10.2022 04:00	718.8

DATA CLEANING

Remove duplicates

When you combine data sets from multiple sources, there is a chance to create duplicate data.

Total electricity production [GWh], 2022

Fuel type	Production
Oil shale, thousand t	5,078
Natural gas, million m3	29
Hydro energy	23
Wind energy	668
Solar energy	596
Solar energy	596

Sample of electricity production/demand data [MWh]

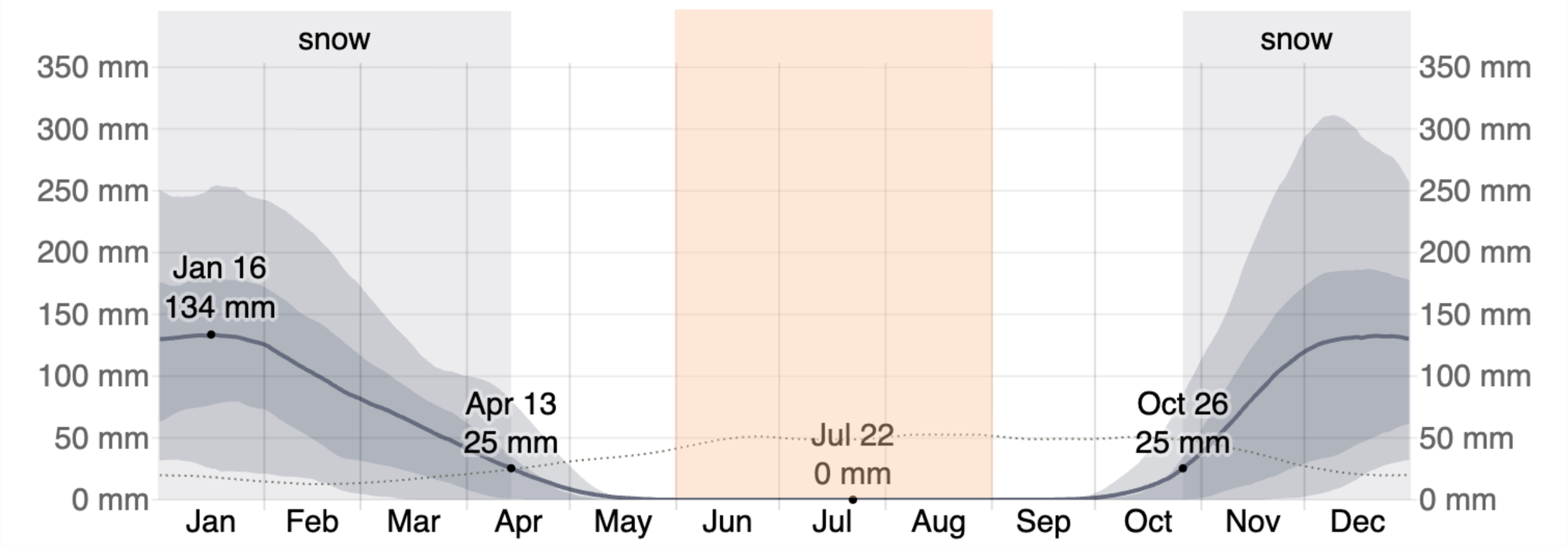
Date (Estonia time)	Consumption
30.10.2022 00:00	768
30.10.2022 01:00	748.9
30.10.2022 02:00	734.1
30.10.2022 03:00	724.2
30.10.2022 03:00	715.4
30.10.2022 04:00	718.8

DATA CLEANING (2)

Remove irrelevant observations

You probably do not need data related to snow while analysing summer periods.

Average monthly snowfall in Tallinn



[Accessed September 9, 2025], https://weatherspark.com/y/91604/Average-Weather-in-Tallinn-Estonia-Year-Round

DATA CLEANING (3)

Fix structural errors

Unify naming conventions (typos, incorrect capitalisation, etc).

Building 1	
Electricity	100
Gas	80
Water	N/A

Building 2	
Elctricity	50
GaS	50
Water	Not Available

Building 3	
ElectricitY	50
gas	-
Water	-

DATA CLEANING (3)

Fix structural errors

Unify naming conventions (typos, incorrect capitalisation, etc).

Building 1

Electricity	100
Gas	80
Water	N/A

Building 2

Elctricity	50
GaS	50
Water	Not Available

Building 3

ElectricitY	50
gas	-
Water	-

DATA CLEANING (3)

Fix structural errors

Unify naming conventions (typos, incorrect capitalisation, etc).

Building 1

Electricity	100
Gas	80
Water	N/A

Building 2

Elctricity	50
GaS	50
Water	Not Available

Building 3

ElectricitY	50
gas	-
Water	-

Utility bills

Electricity	200
Gas	130
Water	N/A

DATA CLEANING (3)

Fix structural errors

Unify naming conventions (typos, incorrect capitalisation, etc).

Building 1

Electricity	100
Gas	80
Water	N/A

Building 2

Elctricity	50
GaS	50
Water	Not Available

Building 3

ElectricitY	50
gas	-
Water	-

Utility bills

Electricity	200
Gas	130
Water	N/A

DATA CLEANING (3)

Fix structural errors

Unify naming conventions (typos, incorrect capitalisation, etc).

Building 1

Electricity	100
Gas	80
Water	N/A

Building 2

Elctricity	50
GaS	50
Water	Not Available

Building 3

ElectricitY	50
gas	-
Water	-

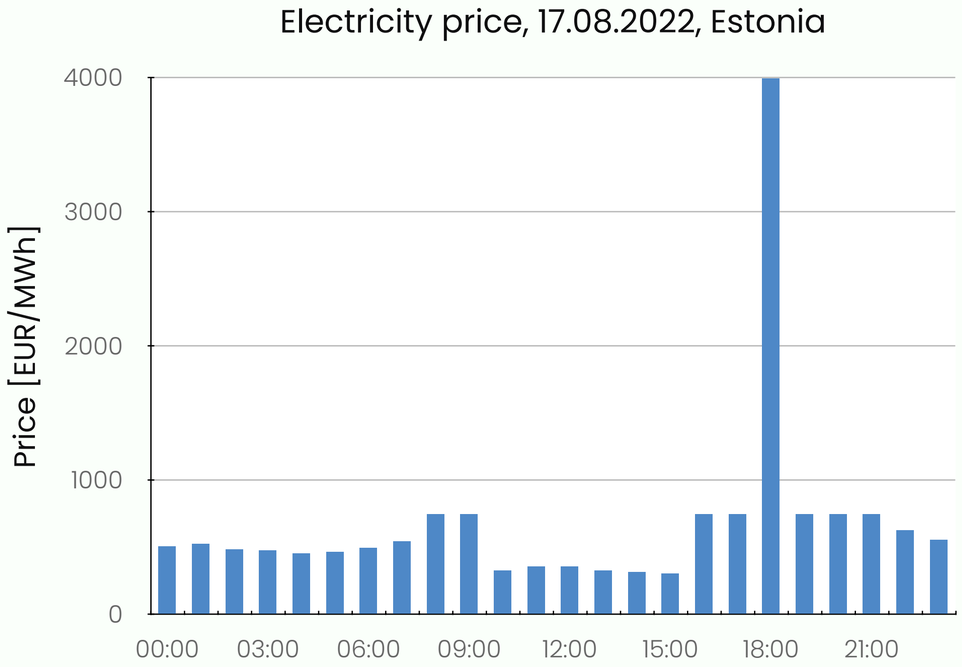
Utility bills

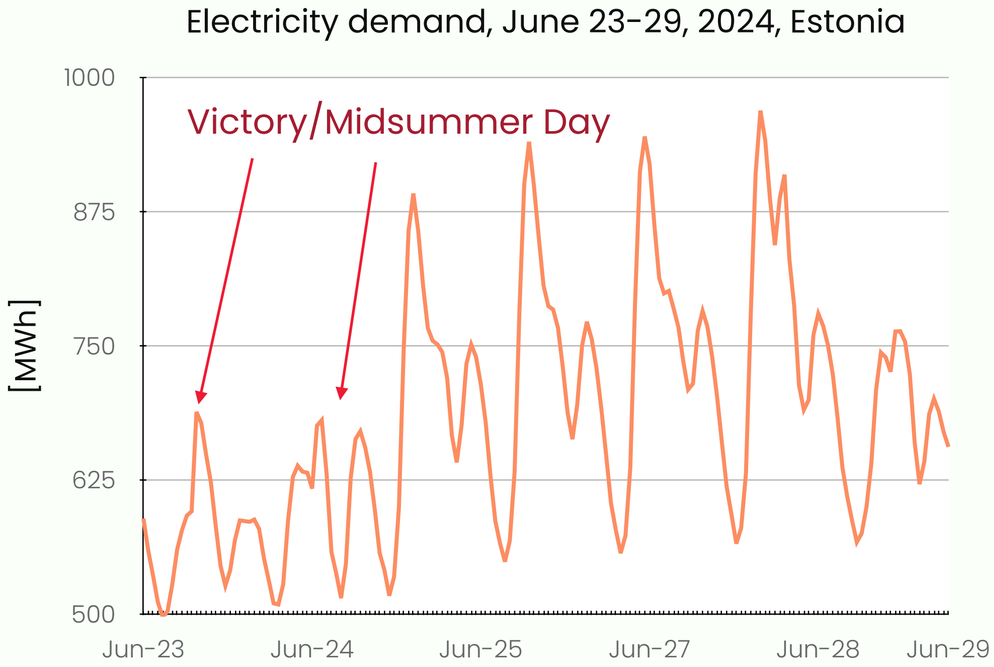
Electricity	200
Gas	130
Water	N/A

DATA CLEANING (4)

Handle unwanted outliers

Determine the validity of an outlier and remove it if needed.





DATA CLEANING (5)

Convert and unify Data Types

Convert numbers, imputed as text, to numerals. Unify units eg choose one currency (EUR vs USD) or measurement system (km vs mi).

Electricity prices, NordPool, July 2024

Estonia	97.97	EUR
Norway (NO1)	286.65	NOK
Denmark (DK1)	462.79	DKK
Sweden (SE1)	205.50	SEK

DATA CLEANING (5)

Convert and unify Data Types

Convert numbers, imputed as text, to numerals. Unify units eg choose one currency (EUR vs USD) or measurement system (km vs mi).

Electricity prices, NordPool, July 2024

Estonia	97.97	EUR
Norway (NO1)	286.65	NOK
Denmark (DK1)	462.79	DKK
Sweden (SE1)	205.50	SEK

Electricity prices, NordPool, July 2024

Estonia	97.97	EUR
Norway (NO1)	24.53	
Denmark (DK1)	62.03	
Sweden (SE1)	17.85	

DATA CLEANING (6)

Handle missing data

Many algorithms operate on integral datasets.

MISSING DATA MECHANISMS: MCAR

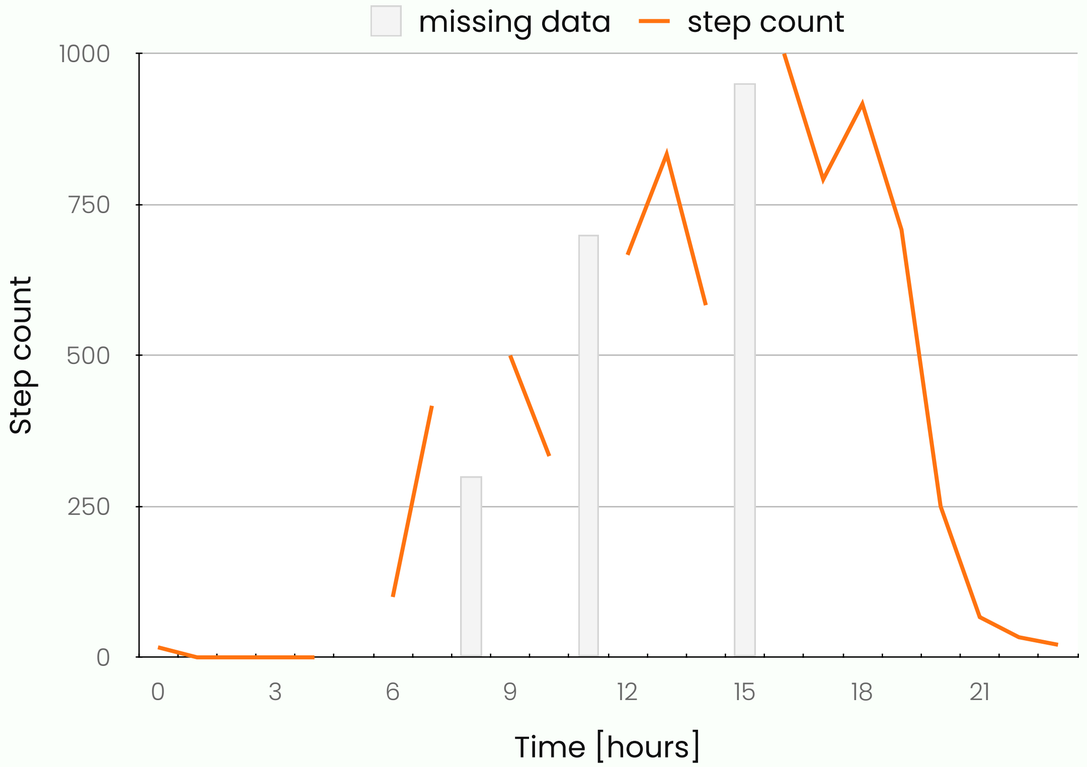
Missing Completely At Random (MCAR):

The probability of missingness does not depend on the observed data values nor on the missing data values, i.e.,

$$P(\text{missing} \mid \text{complete data}) = P(\text{missing}).$$

MCAR (2)

Example: missing randomly at some hours due to **random** tracker glitches.



MISSING DATA MECHANISMS: MAR

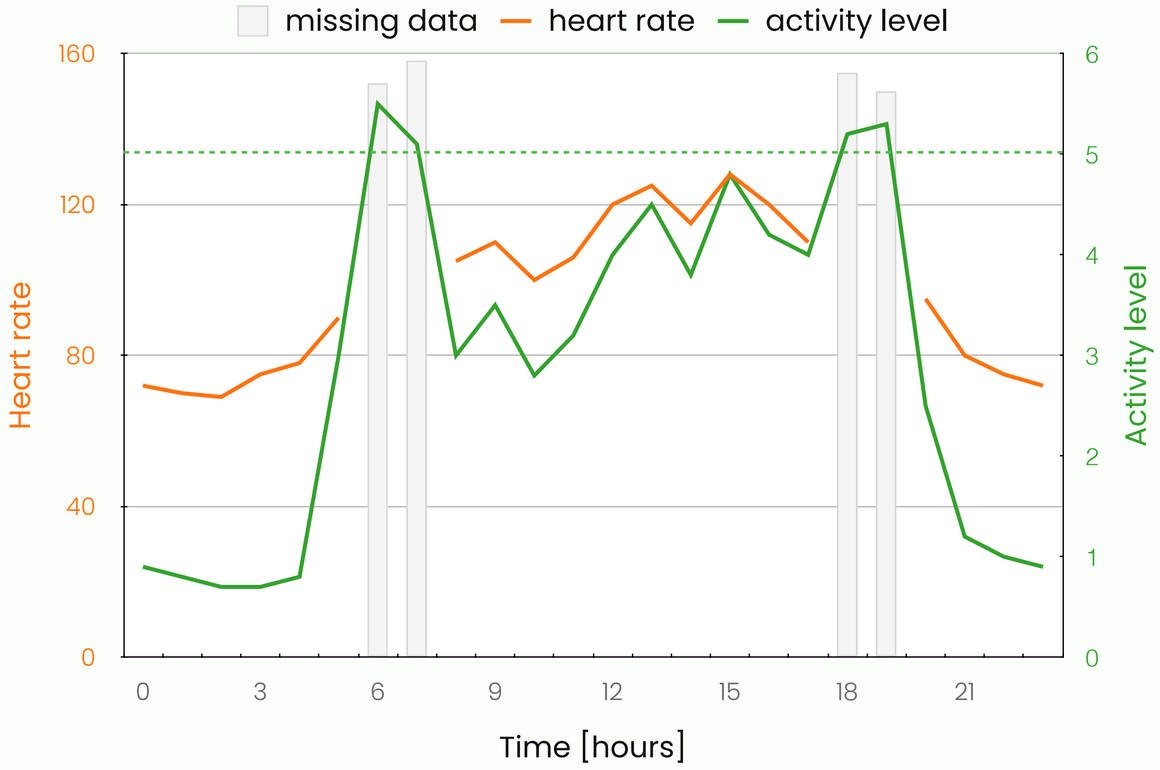
Missing At Random (MAR):

The probability of missingness is partly depends on other observed data, but does not depend on any of the values that are missing, i.e.,

P(missing|complete data) = P(missing|observed data).

MAR (2)

Example: missing only when **Activity level** ≥ 5 (e.g., smartwatch struggles to record during very intense movement).



MISSING DATA MECHANISMS: MNAR

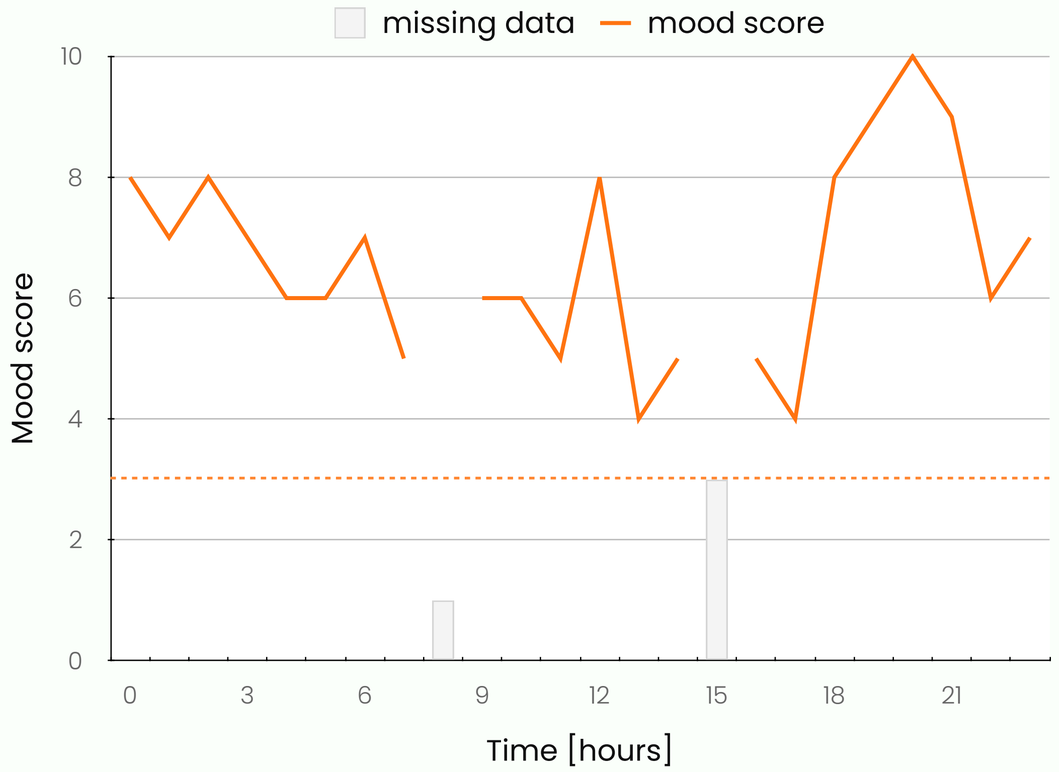
Missing Not At Random (MNAR):

The probability of missingness depends on the missing values themselves, i.e.,

$$P(\text{missing}|\text{complete data}) \neq P(\text{missing}|\text{observed data}).$$

MNAR (2)

Example: missing when the mood score is very low (people often skip reporting mood when feeling bad, e.g., mood score ≤ 3).



FORMALISATION

$y_c = (y_m, y_o)$ are complete data with m standing to missing and o to observed data

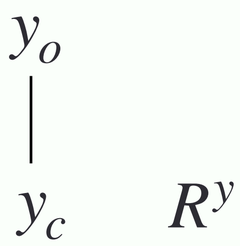
R^y denotes the missingness indicator: 0 for y_o and 1 for y_m

MCAR: R^y is not related to y_o or y_c

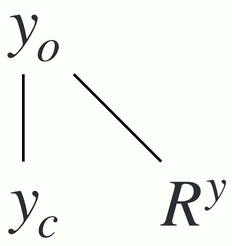
MAR: R^y is related to y_o, but is not related to y_c after controlling for y_o

MNAR: R^y is related to y_c

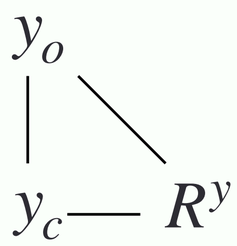
MCAR



MAR



MNAR



EXAMPLE

Given a dataset with monthly energy demand, income, number of residents, gender of the person reporting electricity usage, etc.

Missing **demand** data occur due to:

- data are not stored due to communication problems

EXAMPLE

Given a dataset with monthly energy demand, income, number of residents, gender of the person reporting electricity usage, etc.

Missing **demand** data occur due to:

MCAR
- data are not stored due to communication problems

EXAMPLE

Given a dataset with monthly energy demand, income, number of residents, gender of the person reporting electricity usage, etc.

Missing **demand** data occur due to:

MCAR

- data are not stored due to communication problems
- metering device malfunctions due to unusual high temperatures

EXAMPLE

Given a dataset with monthly energy demand, income, number of residents, gender of the person reporting electricity usage, etc.

Missing **demand** data occur due to:

MCAR
- data are not stored due to communication problems

MAR
- metering device malfunctions due to unusual high temperatures

EXAMPLE

Given a dataset with monthly energy demand, income, number of residents, gender of the person reporting electricity usage, etc.

Missing **demand** data occur due to:

MCAR
- data are not stored due to communication problems

MAR
- metering device malfunctions due to unusual high temperatures
- gender of a reporter (e.g., men tend to forget to report in X% cases)

EXAMPLE

Given a dataset with monthly energy demand, income, number of residents, gender of the person reporting electricity usage, etc.

Missing **demand** data occur due to:

MCAR
- data are not stored due to communication problems

MAR
- metering device malfunctions due to unusual high temperatures

MAR
- gender of a reporter (e.g., men tend to forget to report in X% cases)

EXAMPLE

Given a dataset with monthly energy demand, income, number of residents, gender of the person reporting electricity usage, etc.

Missing **demand** data occur due to:

MCAR
- data are not stored due to communication problems

MAR
- metering device malfunctions due to unusual high temperatures

MAR
- gender of a reporter (e.g., men tend to forget to report in X% cases)
- magnitude of demand (e.g., fear of penalties for high demand)

EXAMPLE

Given a dataset with monthly energy demand, income, number of residents, gender of the person reporting electricity usage, etc.

Missing **demand** data occur due to:

MCAR ▸ data are not stored due to communication problems

MAR ▸ metering device malfunctions due to unusual high temperatures

MAR ▸ gender of a reporter (e.g., men tend to forget to report in X% cases)

MNAR ▸ magnitude of demand (e.g., fear of penalties for high demand)

EXAMPLE

Given a dataset with monthly energy demand, income, number of residents, gender of the person reporting electricity usage, etc.

Missing **demand** data occur due to:

MCAR
- data are not stored due to communication problems

MAR
- metering device malfunctions due to unusual high temperatures

MAR
- gender of a reporter (e.g., men tend to forget to report in X% cases)

MNAR
- magnitude of demand (e.g., fear of penalties for high demand)
- income of the household (low income households report less)

EXAMPLE

Given a dataset with monthly energy demand, income, number of residents, gender of the person reporting electricity usage, etc.

Missing **demand** data occur due to:

MCAR
- data are not stored due to communication problems

MAR
- metering device malfunctions due to unusual high temperatures

MAR
- gender of a reporter (e.g., men tend to forget to report in X% cases)

MNAR
- magnitude of demand (e.g., fear of penalties for high demand)

MAR
- income of the household (low income households report less)

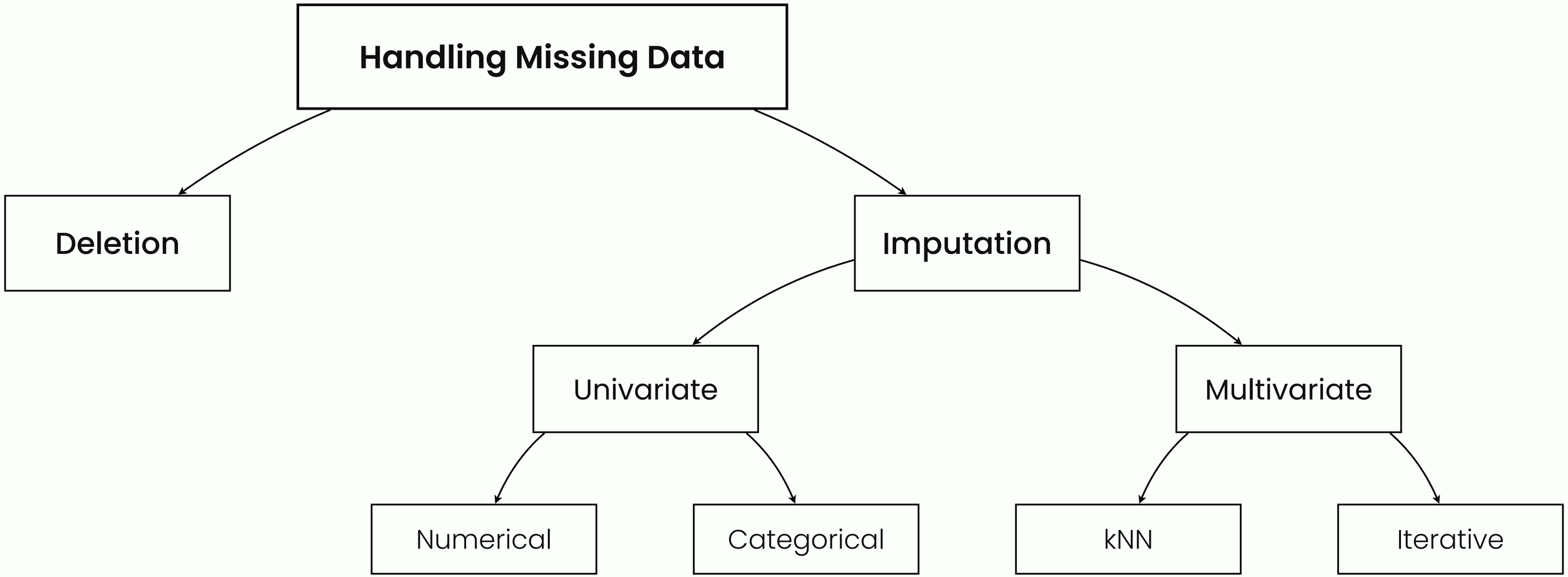
DETECTING MISSING VALUES

Can we be sure about missing mechanism?

Missing data can be detected with

- numerical methods (statistical analysis),
- visual methods (data visualisation), or
- conceptual methods (patterns in the data).

MISSING DATA HANDLING METHODS



Deletion

LIST (ROW)-WISE

Excludes the entire variable with missing values.

- Assumes that the missing data are MCAR.
- It can create a bias in dataset.

Electricity tariffs [cent/kWh]

Client ID	Package type	Day	Night
1	Fixed	16.24	12.76
2	Green	N/A	12.85
3	Fixed	14.5	14.5
4	Green	14.5	N/A
5	Green	13.83	10.87
6	Exchange	N/A	N/A

LIST (ROW)-WISE

Excludes the entire variable with missing values.

- Assumes that the missing data are MCAR.
- It can create a bias in dataset.

Electricity tariffs [cent/kWh]

Client ID	Package type	Day	Night
1	Fixed	16.24	12.76
~~2~~	~~Green~~	~~N/A~~	~~12.85~~
3	Fixed	14.5	14.5
~~4~~	~~Green~~	~~14.5~~	~~N/A~~
5	Green	13.83	10.87
~~6~~	~~Exchange~~	~~N/A~~	~~N/A~~

3 out of 6 records are deleted!

PAIR-WISE

Excludes variables with missing values but still including those cases in the analysis of other variables with non-missing values.

- Assumes that the missing data are MCAR.
- Allows for using more data.

Electricity tariffs [cent/kWh]

Client ID	Package type	Day	Night
1	Fixed	16.24	12.76
2	Green	N/A	12.85
3	Fixed	14.5	14.5
4	Green	14.5	N/A
5	Green	13.83	10.87
6	Exchange	N/A	N/A

PAIR-WISE

Excludes variables with missing values but still including those cases in the analysis of other variables with non-missing values.

- Assumes that the missing data are MCAR.
- Allows for using more data.

Electricity tariffs [cent/kWh]

Client ID	Package type	Day	Night
1	Fixed	16.24	12.76
2	Green	N/A	12.85
3	Fixed	14.5	14.5
4	Green	14.5	N/A
5	Green	13.83	10.87
~~6~~	~~Exchange~~	~~N/A~~	~~N/A~~

PAIR-WISE

Excludes variables with missing values but still including those cases in the analysis of other variables with non-missing values.

- Assumes that the missing data are MCAR.
- Allows for using more data.

Electricity tariffs [cent/kWh]

Client ID	Package type	Day	Night
1	Fixed	16.24	12.76
2	Green	N/A	12.85
3	Fixed	14.5	14.5
4	Green	14.5	N/A
5	Green	13.83	10.87
6	Exchange	N/A	N/A

PAIR-WISE

Excludes variables with missing values but still including those cases in the analysis of other variables with non-missing values.

- Assumes that the missing data are MCAR.
- Allows for using more data.

Keep it

Electricity tariffs [cent/kWh]

Client ID	Package type	Day	Night
1	Fixed	16.24	12.76
~~2~~	~~Green~~	~~N/A~~	~~12.85~~
3	Fixed	14.5	14.5
4	Green	14.5	N/A
5	Green	13.83	10.87
~~6~~	~~Exchange~~	~~N/A~~	~~N/A~~

PAIR-WISE

Excludes variables with missing values but still including those cases in the analysis of other variables with non-missing values.

- Assumes that the missing data are MCAR.
- Allows for using more data.

Electricity tariffs [cent/kWh]

Client ID	Package type	Day	Night
1	Fixed	16.24	12.76
2	Green	N/A	12.85
3	Fixed	14.5	14.5
4	Green	14.5	N/A
5	Green	13.83	10.87
~~6~~	~~Exchange~~	~~N/A~~	~~N/A~~

PAIR-WISE

Excludes variables with missing values but still including those cases in the analysis of other variables with non-missing values.

- Assumes that the missing data are MCAR.
- Allows for using more data.

Electricity tariffs [cent/kWh]

Client ID	Package type	Day	Night
1	Fixed	16.24	12.76
2	Green	N/A	12.85
3	Fixed	14.5	14.5
4	Green	14.5	N/A
5	Green	13.83	10.87
6	Exchange	N/A	N/A

PAIR-WISE

Excludes variables with missing values but still including those cases in the analysis of other variables with non-missing values.

- Assumes that the missing data are MCAR.
- Allows for using more data.

Keep it

Electricity tariffs [cent/kWh]

Client ID	Package type	Day	Night
1	Fixed	16.24	12.76
2	Green	N/A	12.85
3	Fixed	14.5	14.5
~~4~~	~~Green~~	~~14.5~~	~~N/A~~
5	Green	13.83	10.87
~~6~~	~~Exchange~~	~~N/A~~	~~N/A~~

Pair-wise 2 out of 6 records are deleted!

COLUMN-WISE

Discard entire variable if it has missing values.

- If a column has a large percentage of missing values (>80%).
- If the missing data cannot be accurately imputed.

Electricity tariffs [cent/kWh]

Client ID	Package type	Day	Night
1	Fixed	16.24	N/A
2	Green	N/A	N/A
3	Fixed	14.5	14.5
4	Green	14.5	N/A
5	Green	13.83	10.87
6	Exchange	N/A	N/A

COLUMN-WISE

Discard entire variable if it has missing values.

- If a column has a large percentage of missing values (>80%).
- If the missing data cannot be accurately imputed.

Electricity tariffs [cent/kWh]

Client ID	Package type	Day	Night
1	Fixed	16.24	N/A
2	Green	N/A	N/A
3	Fixed	14.5	14.5
4	Green	14.5	N/A
5	Green	13.83	10.87
6	Exchange	N/A	N/A

Univariate Imputation

CONSTANT

Constant imputation assigns the same fixed value to all missing entries instead of removing the observations.

Pros:

- Easy to implement.
- Preserves the structure and size of the dataset.

Cons:

- Can distort statistical analysis.
- Does not provide any additional information about the missing values.

When: missing data have a specific meaning.

MEAN

Missing values of the variable are replaced with average of it’s observed values.

Pros:

- Simple and preserves the sample size.
- If the data are MCAR, mean of the variable remains unbiased.

Cons:

- Affects relationships among variables.
- If the missing data is MAR or MNAR, mean of the variable is biased.

When: If less than 10% of the data is missing and correlations between the variables are low.

EXAMPLE

Temp	Demand
-2.8	810.5
2.8	721
8.8	667.8
10.2	598.5
19.9	616
17.9	566.7
16.9	594.3
9.5	656.1
3.5	687.1
0.9	749.7
-0.3	707.1
-1.7	806.7

$$\bar{x}_{before} = 681.8$$

EXAMPLE

Temp	Demand
-2.8	810.5
2.8	721
8.8	667.8
10.2	598.5
19.9	616
17.9	566.7
16.9	594.3
9.5	656.1
3.5	687.1
0.9	749.7
-0.3	707.1
-1.7	806.7

Temp	Demand
-2.8	810.5
2.8	N/A
8.8	667.8
10.2	598.5
19.9	N/A
17.9	566.7
16.9	594.3
9.5	N/A
3.5	687.1
0.9	N/A
-0.3	707.1
-1.7	N/A

$$\bar{x}_{before} = 681.8$$

EXAMPLE

Temp	Demand
-2.8	810.5
2.8	721
8.8	667.8
10.2	598.5
19.9	616
17.9	566.7
16.9	594.3
9.5	656.1
3.5	687.1
0.9	749.7
-0.3	707.1
-1.7	806.7

Temp	Demand
-2.8	810.5
2.8	N/A
8.8	667.8
10.2	598.5
19.9	N/A
17.9	566.7
16.9	594.3
9.5	N/A
3.5	687.1
0.9	N/A
-0.3	707.1
-1.7	N/A

Temp	Demand with mean
-2.8	810.5
2.8	661.7
8.8	667.8
10.2	598.5
19.9	661.7
17.9	566.7
16.9	594.3
9.5	661.7
3.5	687.1
0.9	661.7
-0.3	707.1
-1.7	661.7

$$\bar{x}_{before} = 681.8$$

EXAMPLE

Temp	Demand
-2.8	810.5
2.8	721
8.8	667.8
10.2	598.5
19.9	616
17.9	566.7
16.9	594.3
9.5	656.1
3.5	687.1
0.9	749.7
-0.3	707.1
-1.7	806.7

$\bar{x}_{before} = 681.8$

Temp	Demand
-2.8	810.5
2.8	N/A
8.8	667.8
10.2	598.5
19.9	N/A
17.9	566.7
16.9	594.3
9.5	N/A
3.5	687.1
0.9	N/A
-0.3	707.1
-1.7	N/A

Temp	Demand with mean
-2.8	810.5
2.8	661.7
8.8	667.8
10.2	598.5
19.9	661.7
17.9	566.7
16.9	594.3
9.5	661.7
3.5	687.1
0.9	661.7
-0.3	707.1
-1.7	661.7

$\bar{x}_{after} = 661.7$

EXAMPLE

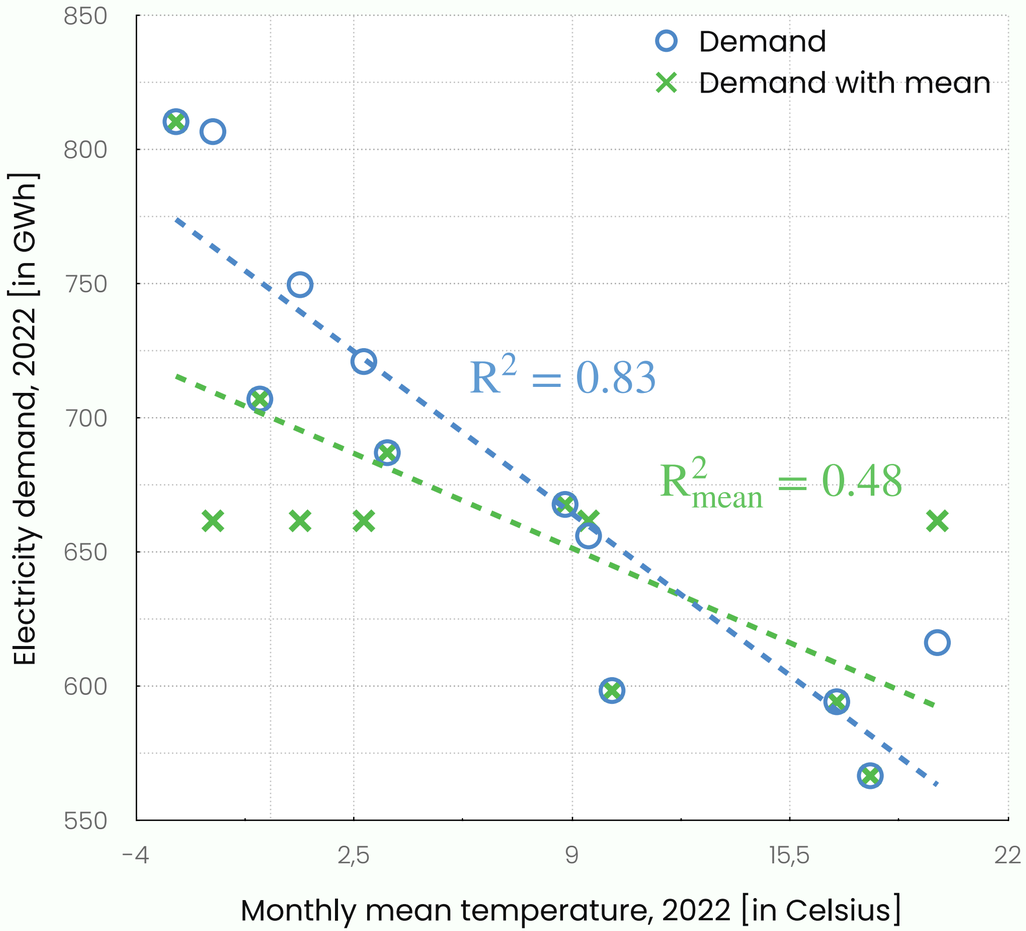
Temp	Demand
-2.8	810.5
2.8	721
8.8	667.8
10.2	598.5
19.9	616
17.9	566.7
16.9	594.3
9.5	656.1
3.5	687.1
0.9	749.7
-0.3	707.1
-1.7	806.7

$\bar{x}_{before} = 681.8$

Temp	Demand
-2.8	810.5
2.8	N/A
8.8	667.8
10.2	598.5
19.9	N/A
17.9	566.7
16.9	594.3
9.5	N/A
3.5	687.1
0.9	N/A
-0.3	707.1
-1.7	N/A

Temp	Demand with mean
-2.8	810.5
2.8	661.7
8.8	667.8
10.2	598.5
19.9	661.7
17.9	566.7
16.9	594.3
9.5	661.7
3.5	687.1
0.9	661.7
-0.3	707.1
-1.7	661.7

$\bar{x}_{after} = 661.7$



MEDIAN

Missing values are replaced with median of the present data.

Pros:

- Simple and preserves the sample size.
- Robust to outliers.

Cons:

- Introduce bias in the data.
- Assumes same distribution for the missing and observed values.

MODE

Missing values are replaced with value with the highest frequency.

Pros:

- Simple and preserves the sample size.
- Suitable for MAR or MCAR data.

Cons:

- Introduce bias in the data.
- Underestimates the variance of the data.
- Not suitable for MNAR data.

HOT DECK IMPUTATION

For each respondent (recipient) with a missing y, find a respondent (donor) with similar values of x in the observed data and take its y value.
Donor: e.g., last/next observation or nearest neighbour.

Pros:

- Allows more educated guess on the missing data.

Cons:

- Can lead to bias in the data if the data is not stationary.
- MNAR data can lead to over/underestimation of the model parameters.
- Not suitable for data with a continuous gap.

EXAMPLE

Electricity tariffs [cent/kWh]

Client ID	Package type	Day	Night
1	Fixed	16.24	12.76
2	Green	N/A	12.85
3	Fixed	14.5	14.5
4	Green	14.5	N/A
5	Green	13.83	10.87
6	Exchange	N/A	N/A

Last observation carried forward

Client ID	Package type	Day	Night
1	Fixed	16.24	12.76
2	Green	N/A	12.85
3	Fixed	14.5	14.5
4	Green	14.5	N/A
5	Green	13.83	10.87
6	Exchange	N/A	N/A

Next observation carried backward

EXAMPLE

Electricity tariffs [cent/kWh]

Client ID	Package type	Day	Night
1	Fixed	16.24	12.76
2	Green	N/A	12.85
3	Fixed	14.5	14.5
4	Green	14.5	N/A
5	Green	13.83	10.87
6	Exchange	N/A	N/A

Last observation carried forward

Electricity tariffs [cent/kWh]

Client ID	Package type	Day	Night
1	Fixed	16.24	12.76
2	Green	N/A	12.85
3	Fixed	14.5	14.5
4	Green	14.5	12.85
5	Green	13.83	10.87
6	Exchange	N/A	N/A

Client ID	Package type	Day	Night
1	Fixed	16.24	12.76
2	Green	N/A	12.85
3	Fixed	14.5	14.5
4	Green	14.5	N/A
5	Green	13.83	10.87
6	Exchange	N/A	N/A

Next observation carried backward

Client ID	Package type	Day	Night
1	Fixed	16.24	12.76
2	Green	N/A	12.85
3	Fixed	14.5	14.5
4	Green	14.5	10.87
5	Green	13.83	10.87
6	Exchange	N/A	N/A

COLD DECK IMPUTATION

The missing value is replaced with one constant from an external source (e.g., previous analysis).

Pros:

- Effective when the data points are expected to have standard or normative values, such as demographic data.

Cons:

- Potential for bias if the external dataset is not perfect.
- Outdated information.

LINEAR INTERPOLATION

Replace a missing value by fitting a line between two known values.

Pros:

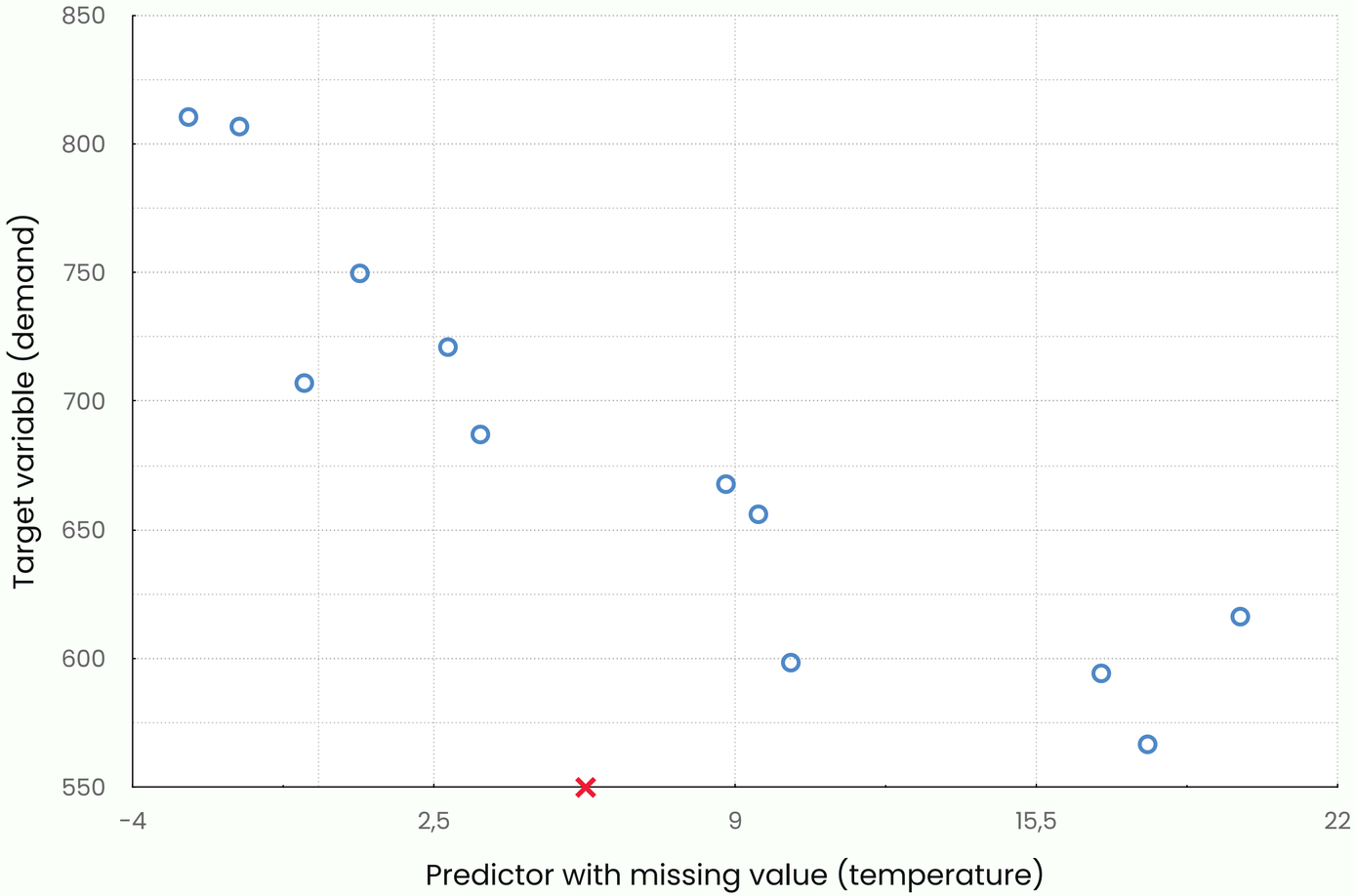
- Simple and light.
- Almost no effect to the structure of the data.

Cons:

- Assumes linear trend.
- Sensitive to outliers.

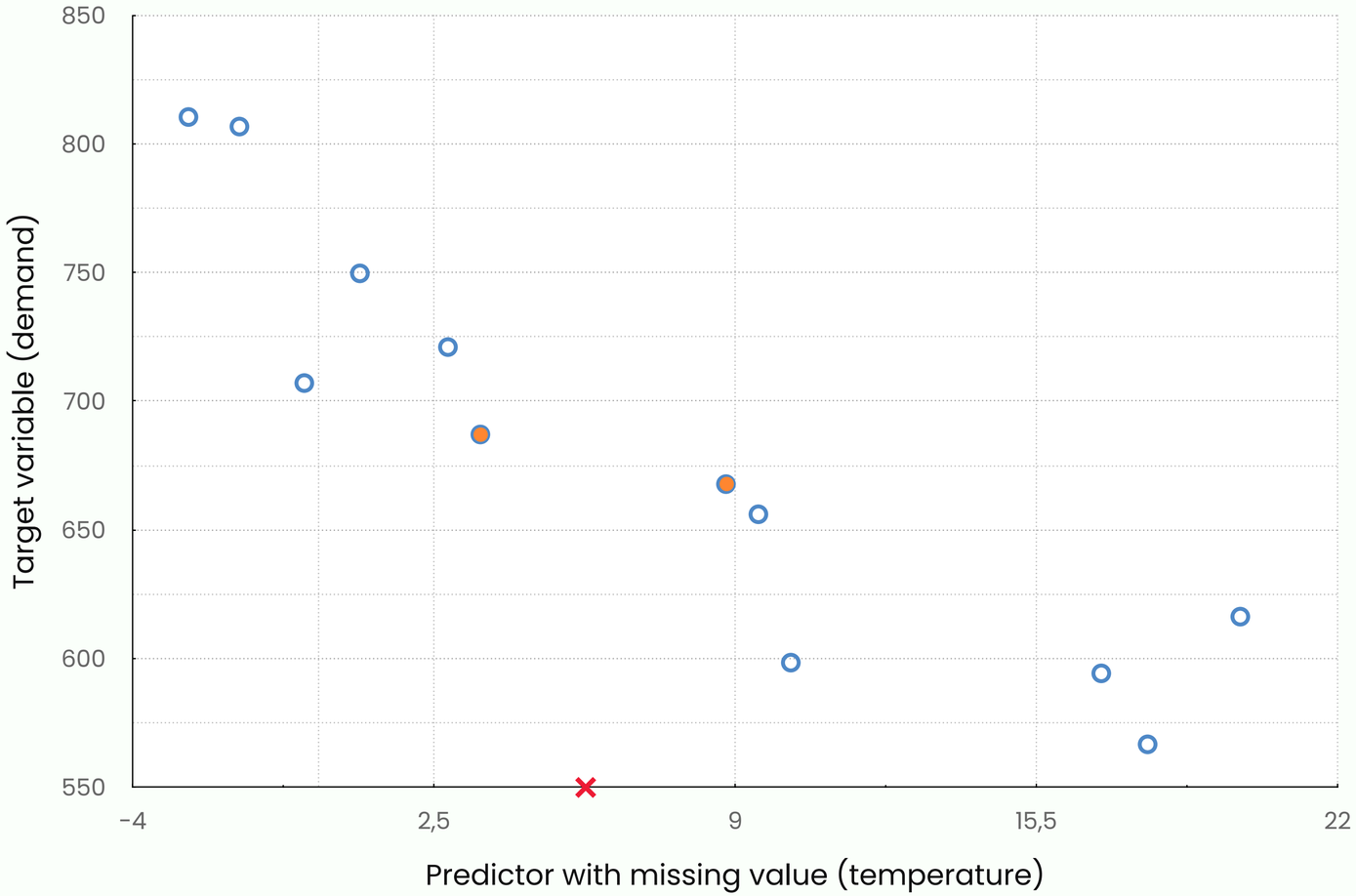
EXAMPLE

Temp	Demand
-2.8	810.5
2.8	721
5.75	N/A
8.8	667.8
10.2	598.5
19.9	616
17.9	566.7
16.9	594.3
9.5	656.1
3.5	687.1
0.9	749.7
-0.3	707.1
-1.7	806.7



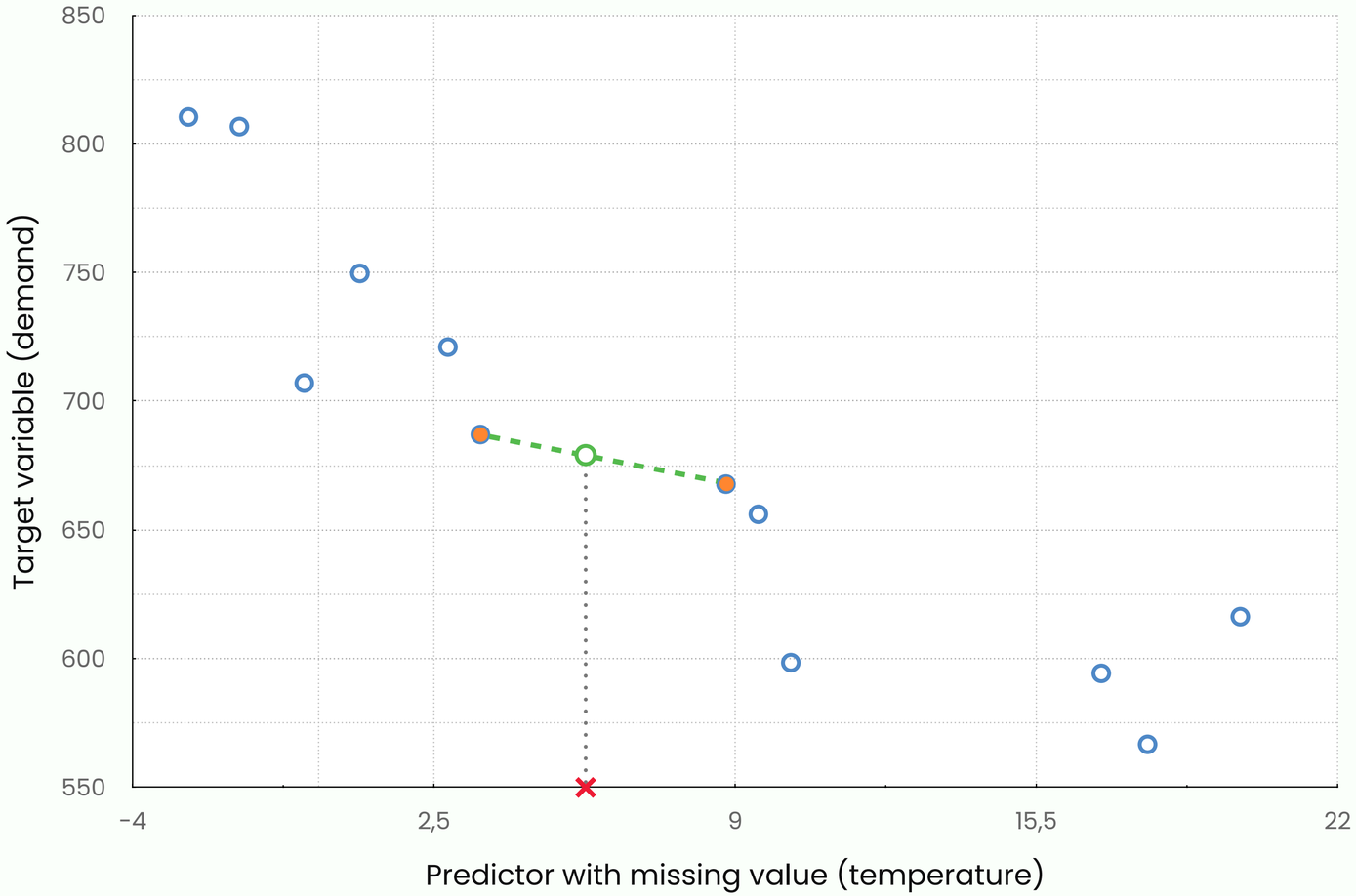
EXAMPLE

Temp	Demand
-2.8	810.5
2.8	721
5.75	N/A
8.8	667.8
10.2	598.5
19.9	616
17.9	566.7
16.9	594.3
9.5	656.1
3.5	687.1
0.9	749.7
-0.3	707.1
-1.7	806.7



EXAMPLE

Temp	Demand
-2.8	810.5
2.8	721
5.75	N/A
8.8	667.8
10.2	598.5
19.9	616
17.9	566.7
16.9	594.3
9.5	656.1
3.5	687.1
0.9	749.7
-0.3	707.1
-1.7	806.7

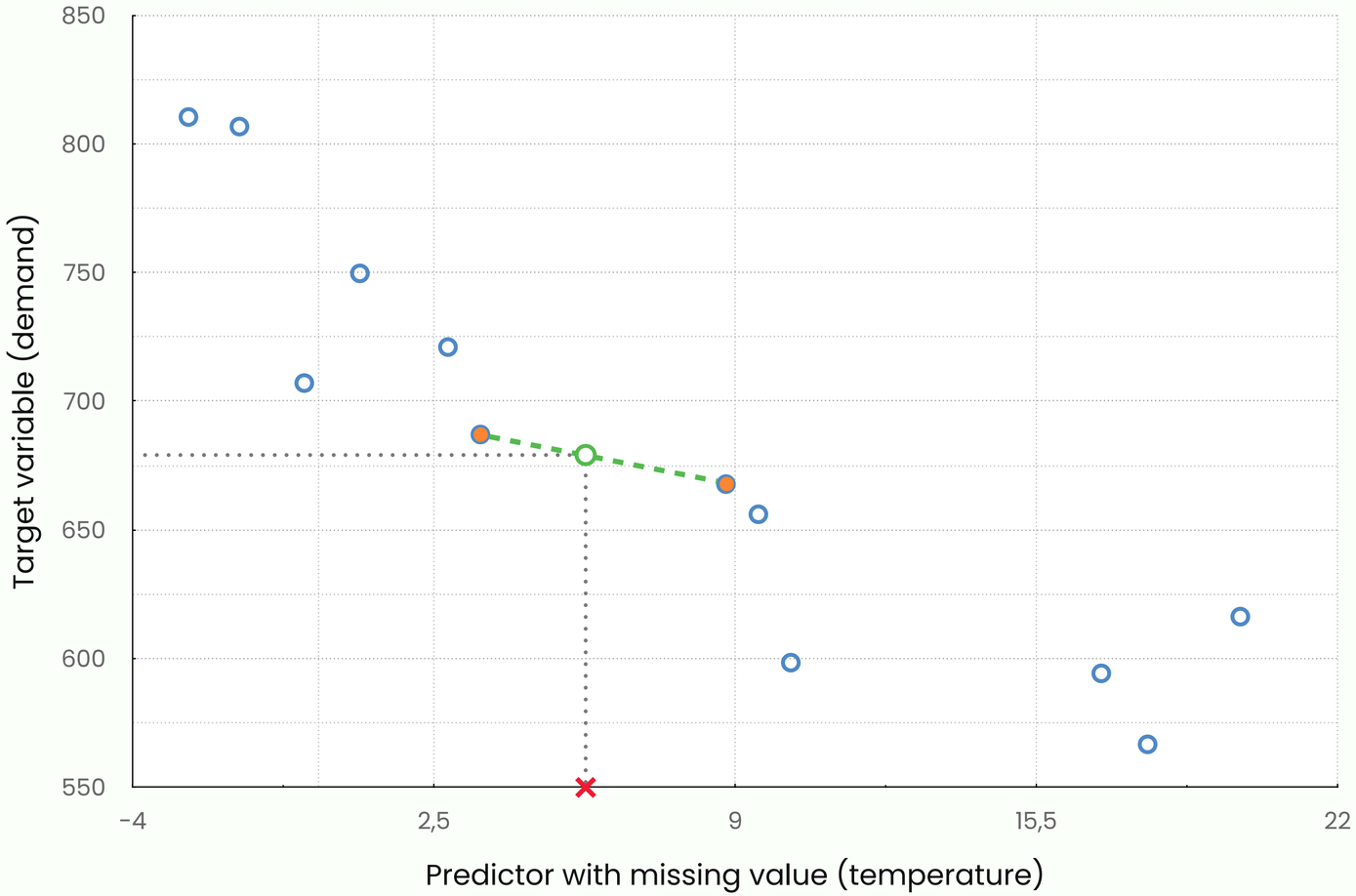


$$\frac{x - 8.8}{3.5 - 8.8} = \frac{y - 667.8}{687.1 - 667.8} \quad \rightarrow \quad y = -3.64x + 699.85$$

EXAMPLE

Temp	Demand
-2.8	810.5
2.8	721
5.75	N/A
8.8	667.8
10.2	598.5
19.9	616
17.9	566.7
16.9	594.3
9.5	656.1
3.5	687.1
0.9	749.7
-0.3	707.1
-1.7	806.7

Temp	Demand
-2.8	810.5
2.8	721
5.75	678.92
8.8	667.8
10.2	598.5
19.9	616
17.9	566.7
16.9	594.3
9.5	656.1
3.5	687.1
0.9	749.7
-0.3	707.1
-1.7	806.7



$$\frac{x - 8.8}{3.5 - 8.8} = \frac{y - 667.8}{687.1 - 667.8} \quad \rightarrow \quad y = -3.64x + 699.85$$

REGRESSION

A regression model is used to predict the missing values based on the entire dataset.

Pros:

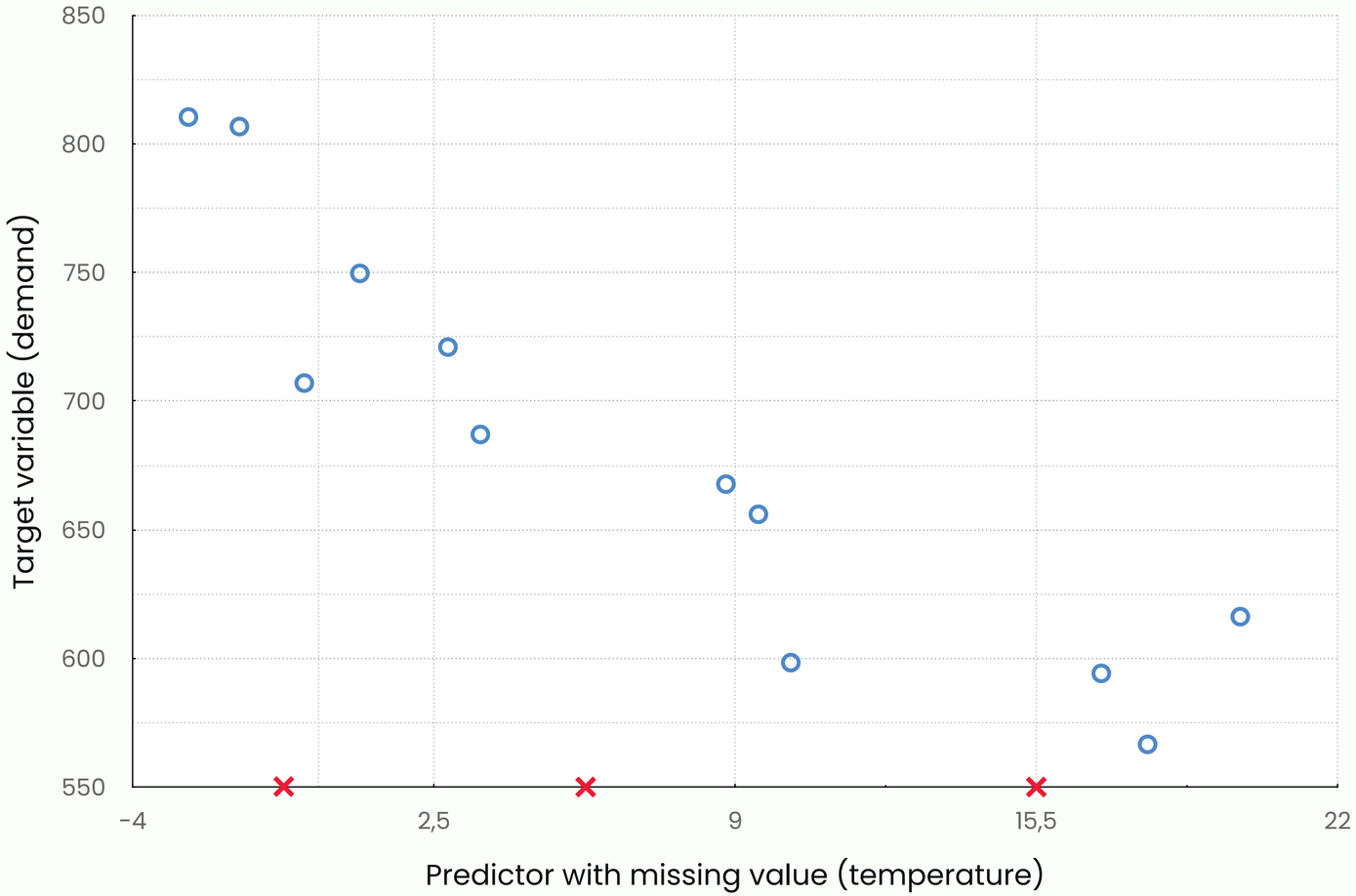
- Uses relationship between variables.

Cons:

- Imputed values are estimates.

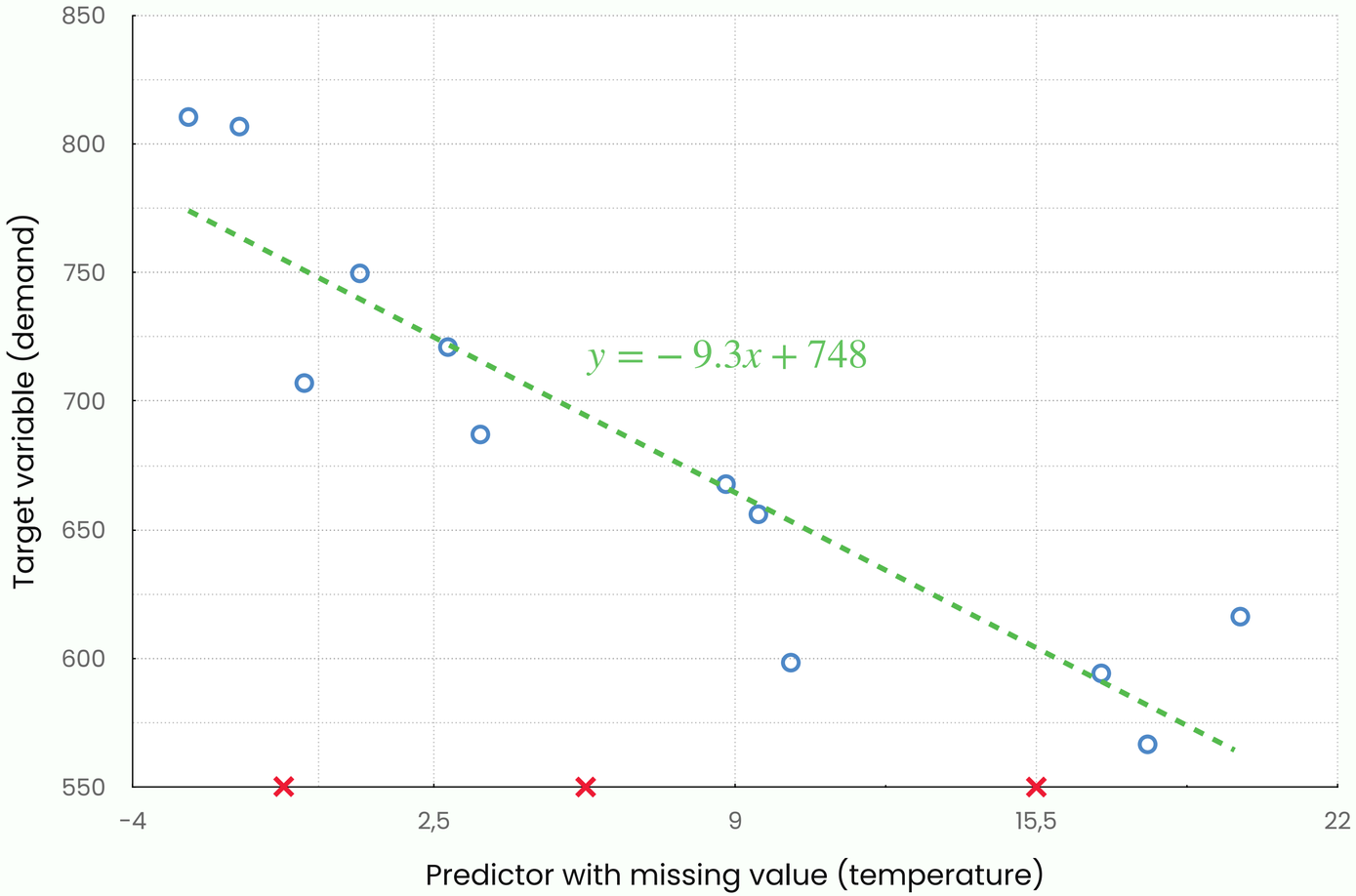
EXAMPLE

Temp	Demand
-2.8	810.5
-0.75	N/A
2.8	721
5.75	N/A
8.8	667.8
10.2	598.5
15.5	N/A
19.9	616
17.9	566.7
16.9	594.3
9.5	656.1
3.5	687.1
0.9	749.7
-0.3	707.1
-1.7	806.7



EXAMPLE

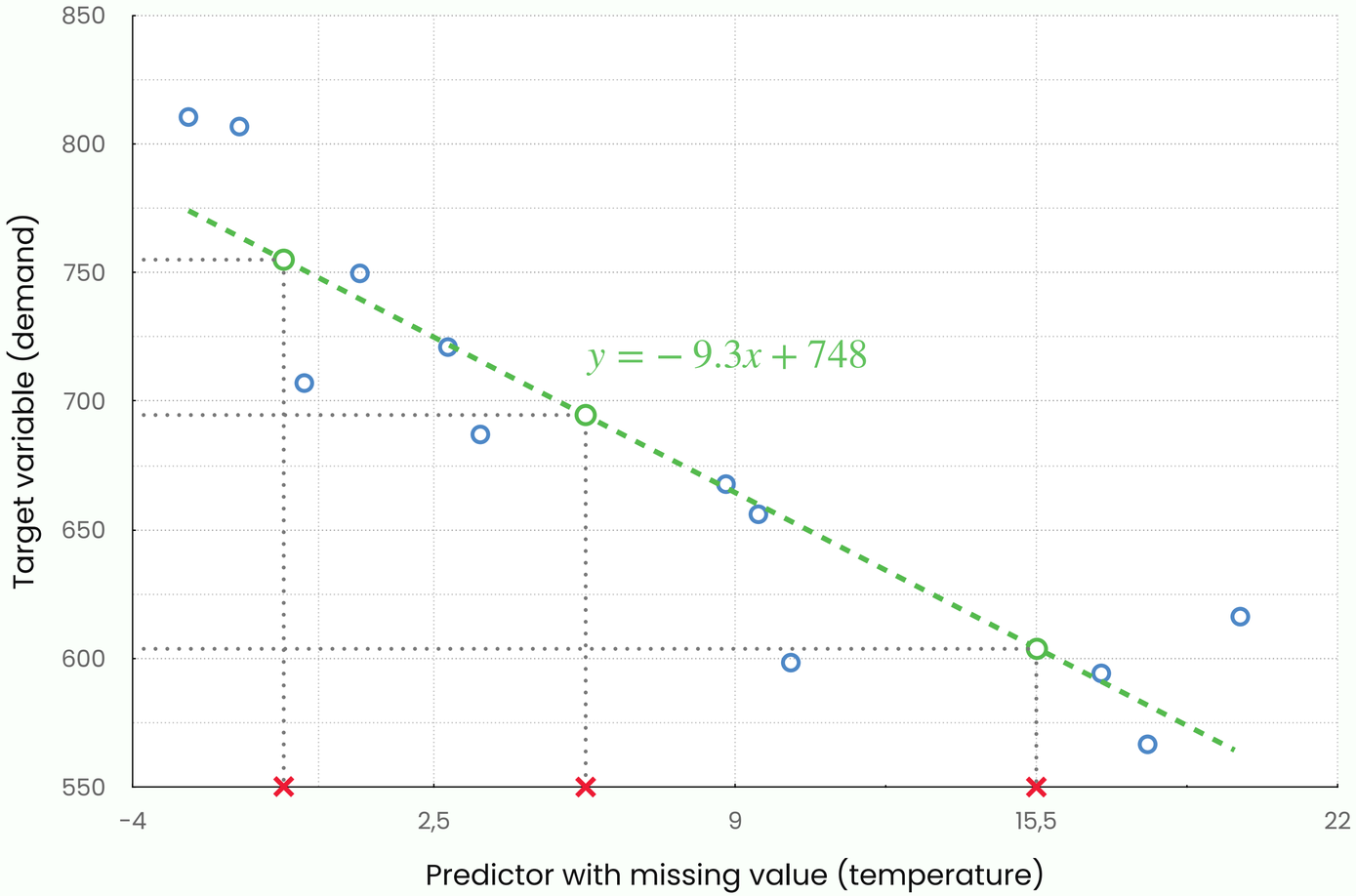
Temp	Demand
-2.8	810.5
-0.75	N/A
2.8	721
5.75	N/A
8.8	667.8
10.2	598.5
15.5	N/A
19.9	616
17.9	566.7
16.9	594.3
9.5	656.1
3.5	687.1
0.9	749.7
-0.3	707.1
-1.7	806.7



EXAMPLE

Temp	Demand
-2.8	810.5
-0.75	N/A
2.8	721
5.75	N/A
8.8	667.8
10.2	598.5
15.5	N/A
19.9	616
17.9	566.7
16.9	594.3
9.5	656.1
3.5	687.1
0.9	749.7
-0.3	707.1
-1.7	806.7

Temp	Demand
-2.8	810.5
-0.75	755
2.8	721
5.75	694.5
8.8	667.8
10.2	598.5
15.5	603.9
19.9	616
17.9	566.7
16.9	594.3
9.5	656.1
3.5	687.1
0.9	749.7
-0.3	707.1
-1.7	806.7



Multivariate Imputation

GENERAL IDEA

Multiple imputation generates several different sets of plausible values for missing data to reflect the uncertainty about what the missing values could be.

MI methods try to **averaging** the outcomes across multiple imputed data sets.

All multiple imputation methods follow three steps.

THREE MAIN STEPS

1. **Imputation**: Generate m datasets by imputing missing values with values randomly drawn from some distributions.
2. **Analysis**: Perform statistical analysis on each of the m datasets.
3. **Pooling**: Pool the results into a single set of results.

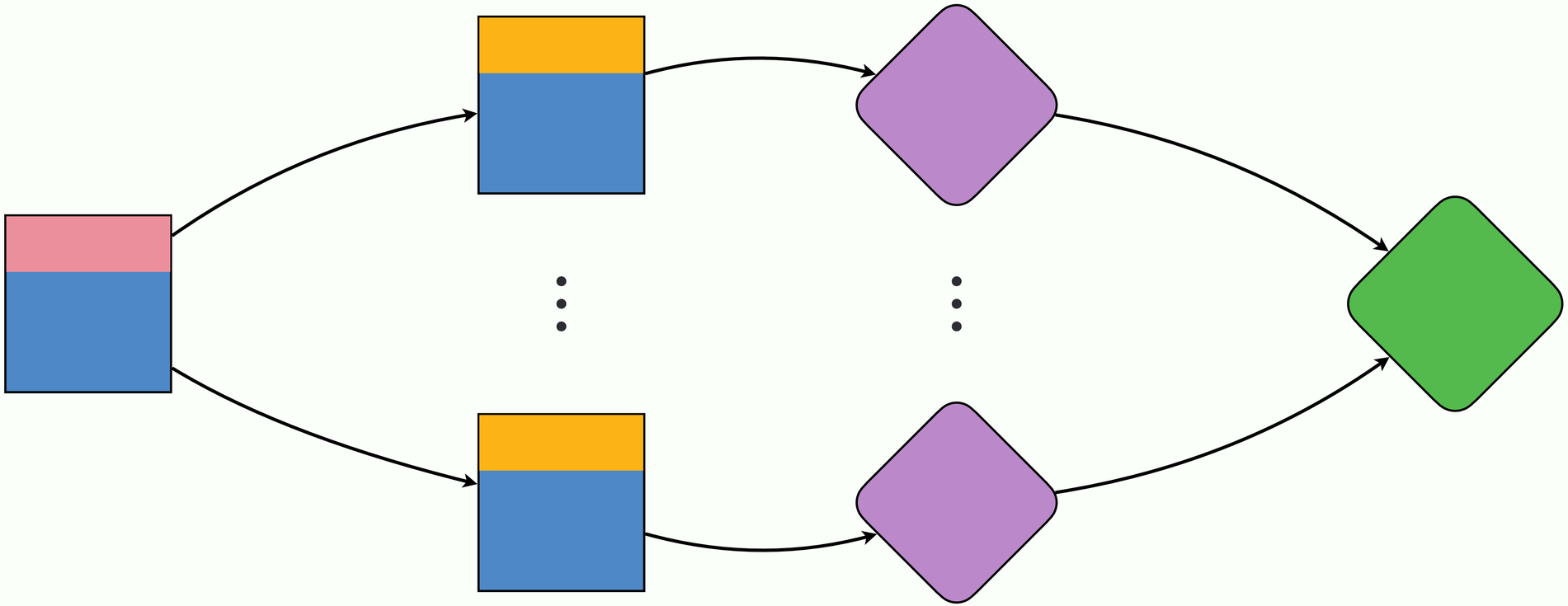
Pros:

- Improved accuracy
- Accounts for uncertainty
- Handles complex data structures

Cons:

- Complexity and computational intensity
- Model specification and convergence issues

WORKFLOW



Incomplete Data

Imputed Data

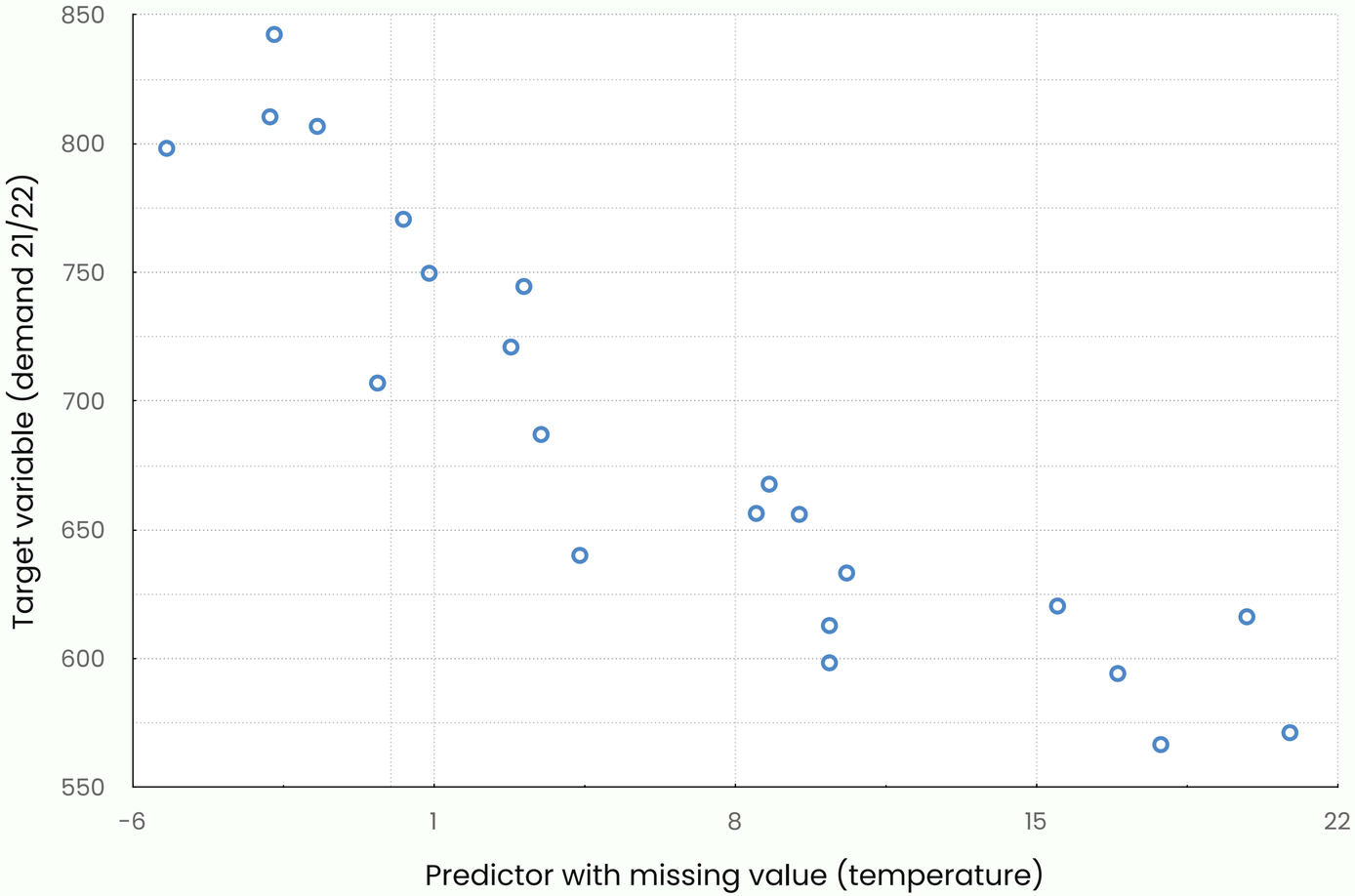
Analysis Results

Pooled Results

EXAMPLE: INCOMPLETE DATA

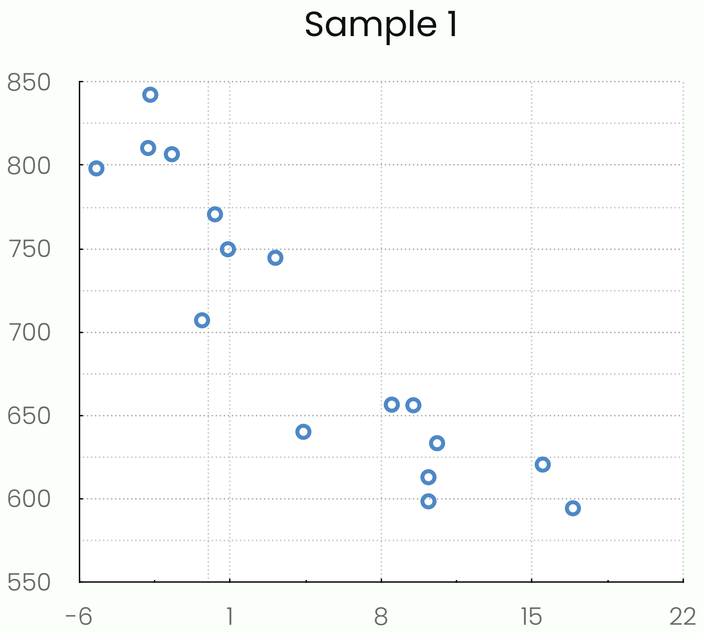
Temp	Demand	...
-2.8	810.5	
-0.75	N/A	
2.8	721	
5.75	N/A	
8.8	667.8	
10.2	598.5	
16.1	N/A	
19.9	616	
...	...	

Two years (21/22) with total of 24 records



IMPUTATION

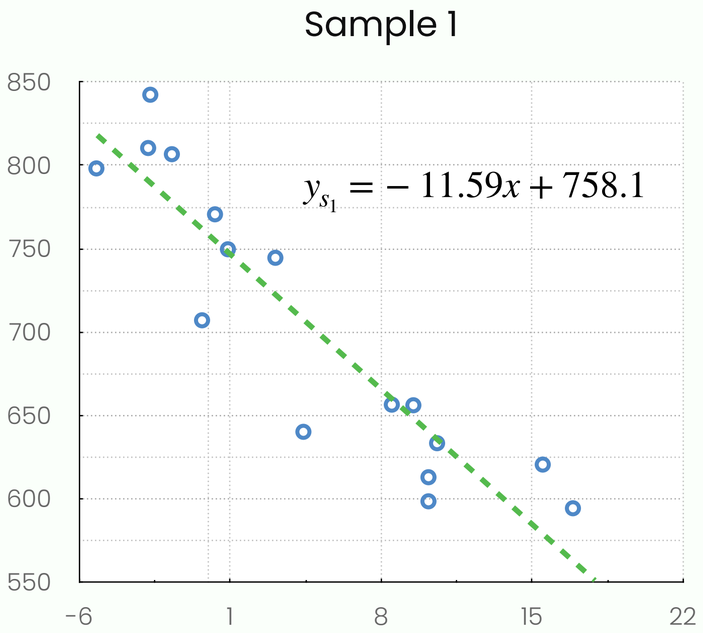
Randomly select 16 datapoints from the raw data.



NB! Subsample may be of any size, but not less than 50% of the size of the raw data.

IMPUTATION

Randomly select 16 datapoints from the raw data.



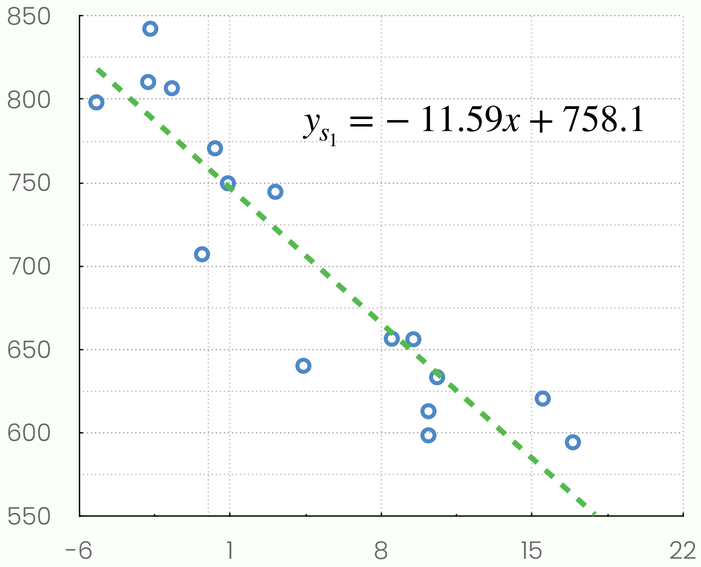
Temp	Demand
-0.75	766.8
5.75	691.5
16.1	571,5

NB! Subsample may be of any size, but not less than 50% of the size of the raw data.

IMPUTATION

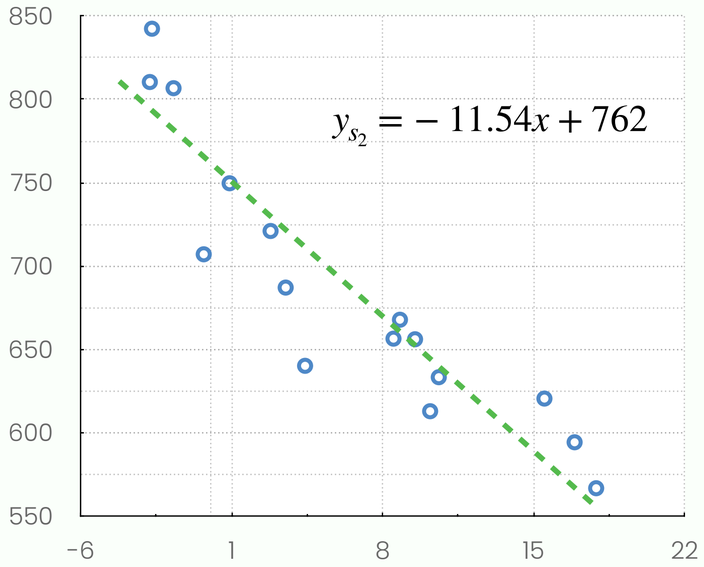
Randomly select 16 datapoints from the raw data.

Sample 1



Temp	Demand
-0.75	766.8
5.75	691.5
16.1	571,5

Sample 2



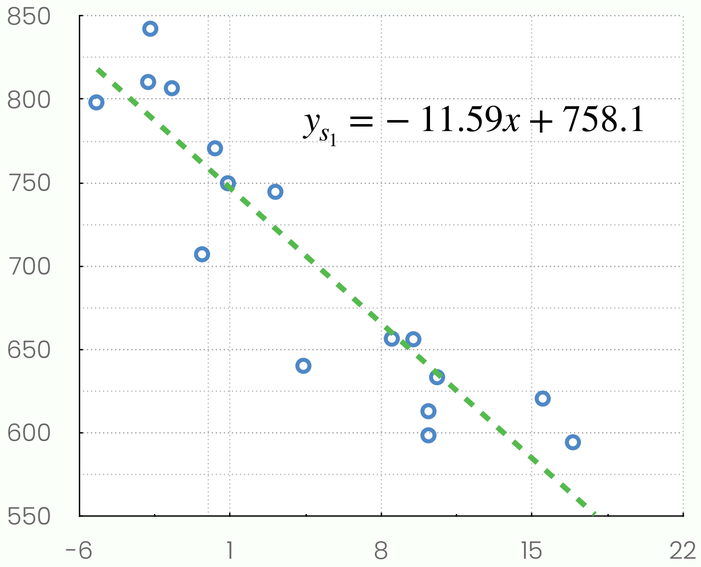
Temp	Demand
-0.75	770.7
5.75	695.6
16.1	576.2

NB! Subsample may be of any size, but not less than 50% of the size of the raw data.

IMPUTATION

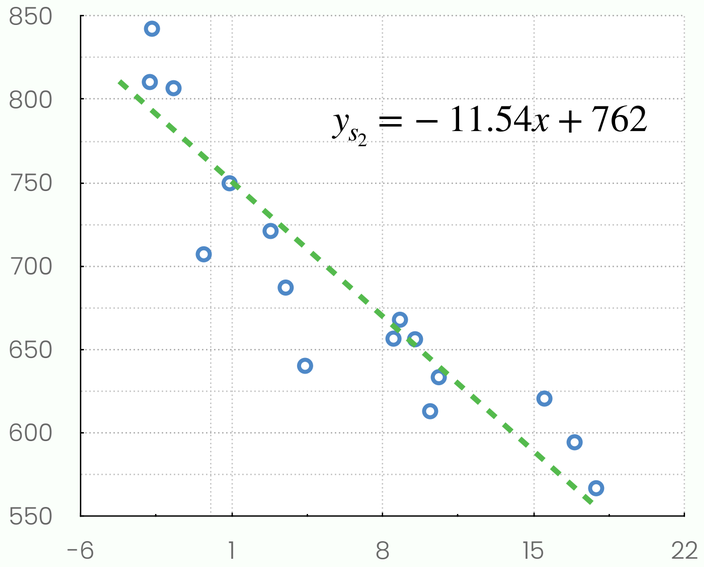
Randomly select 16 datapoints from the raw data.

Sample 1



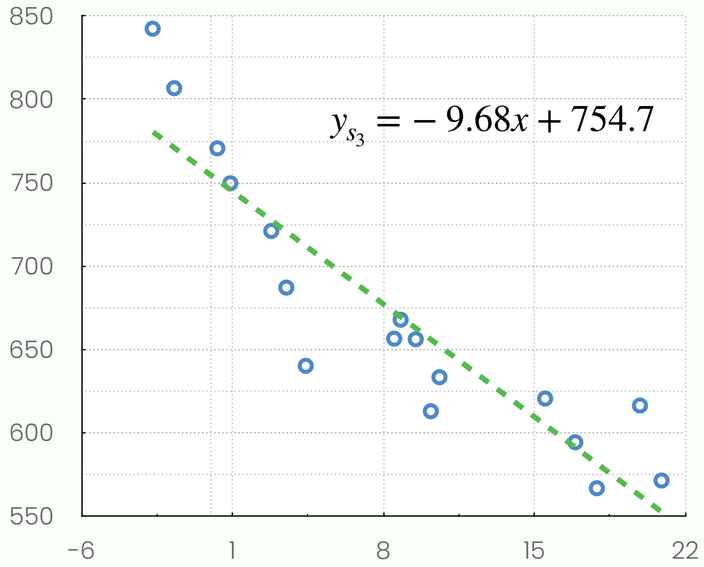
Temp	Demand
-0.75	766.8
5.75	691.5
16.1	571,5

Sample 2



Temp	Demand
-0.75	770.7
5.75	695.6
16.1	576.2

Sample 3



Temp	Demand
-0.75	762
5.75	699
16.1	598.9

NB! Subsample may be of any size, but not less than 50% of the size of the raw data.

POOLING

From developed estimators $(y_{s_i} = \beta_1^{s_i} x + \beta_0^{s_i})$, collect parameters:

$$\begin{cases} \beta_1^{s_1} = -11.59, & \beta_0^{s_1} = 758.1, \\ \beta_1^{s_2} = -11.54, & \beta_0^{s_2} = 762, \\ \beta_1^{s_3} = -9.68, & \beta_0^{s_3} = 754.7. \end{cases}$$

$$\begin{cases} y_{s_1} = -11.59x + 758.1, \\ y_{s_2} = -11.54x + 762, \\ y_{s_3} = -9.68x + 754.7 \end{cases}$$

NB! The same result can be obtained by directly pooling (averaging) estimates from all samples. In other words, average values for the same datapoint across all samples.
A. Miles. Obtaining predictions from models fit to multiply imputed data, *Sociological Methods & Research* 45(1), pp. 1–11, 2015.

POOLING

From developed estimators $\left(y_{s_i} = \beta_1^{s_i} x + \beta_0^{s_i}\right)$, collect parameters:

$$\begin{cases} \beta_1^{s_1} = -11.59, & \beta_0^{s_1} = 758.1, \\ \beta_1^{s_2} = -11.54, & \beta_0^{s_2} = 762, \\ \beta_1^{s_3} = -9.68, & \beta_0^{s_3} = 754.7. \end{cases}$$

$$\begin{cases} y_{s_1} = -11.59x + 758.1, \\ y_{s_2} = -11.54x + 762, \\ y_{s_3} = -9.68x + 754.7 \end{cases}$$

Using Rubin's Rule, results can be pooled as:

$$\begin{cases} \bar{\beta}_1^{\mathrm{MI}} = \left(\beta_1^{s_1} + \beta_1^{s_2} + \beta_1^{s_3}\right) = \frac{1}{3}\left(-11.59 - 11.54 + -9.68\right) \approx -10.94, \\ \bar{\beta}_0^{\mathrm{MI}} = \frac{1}{3}\left(\beta_0^{s_1} + \beta_0^{s_2} + \beta_0^{s_3}\right) = \frac{1}{3}\left(758.1 + 762 + 754.7\right) \approx 758.3. \end{cases}$$

NB! The same result can be obtained by directly pooling (averaging) estimates from all samples. In other words, average values for the same datapoint across all samples.
A. Miles. Obtaining predictions from models fit to multiply imputed data, *Sociological Methods & Research* 45(1), pp. 1–11, 2015.

POOLING

From developed estimators $(y_{s_i} = \beta_1^{s_i} x + \beta_0^{s_i})$, collect parameters:

$$\begin{cases} \beta_1^{s_1} = -11.59, & \beta_0^{s_1} = 758.1, \\ \beta_1^{s_2} = -11.54, & \beta_0^{s_2} = 762, \\ \beta_1^{s_3} = -9.68, & \beta_0^{s_3} = 754.7. \end{cases}$$

$$\begin{cases} y_{s_1} = -11.59x + 758.1, \\ y_{s_2} = -11.54x + 762, \\ y_{s_3} = -9.68x + 754.7 \end{cases}$$

Using Rubin's Rule, results can be pooled as:

$$\begin{cases} \bar{\beta}_1^{\text{MI}} = \left(\beta_1^{s_1} + \beta_1^{s_2} + \beta_1^{s_3}\right) = \frac{1}{3}\left(-11.59 - 11.54 + -9.68\right) \approx -10.94, \\ \bar{\beta}_0^{\text{MI}} = \frac{1}{3}\left(\beta_0^{s_1} + \beta_0^{s_2} + \beta_0^{s_3}\right) = \frac{1}{3}\left(758.1 + 762 + 754.7\right) \approx 758.3. \end{cases}$$

And the final model reads as:

$$y^{\text{MI}} = \bar{\beta}_1^{\text{MI}} * \text{temp} + \bar{\beta}_0^{\text{MI}} = -10.94x + 758.3.$$

NB! The same result can be obtained by directly pooling (averaging) estimates from all samples. In other words, average values for the same datapoint across all samples.
A. Miles. Obtaining predictions from models fit to multiply imputed data, *Sociological Methods & Research* 45(1), pp. 1–11, 2015.

POOLING

From developed estimators $(y_{s_i} = \beta_1^{s_i} x + \beta_0^{s_i})$, collect parameters:

$$\begin{cases} \beta_1^{s_1} = -11.59, & \beta_0^{s_1} = 758.1, \\ \beta_1^{s_2} = -11.54, & \beta_0^{s_2} = 762, \\ \beta_1^{s_3} = -9.68, & \beta_0^{s_3} = 754.7. \end{cases}$$

$$\begin{cases} y_{s_1} = -11.59x + 758.1, \\ y_{s_2} = -11.54x + 762, \\ y_{s_3} = -9.68x + 754.7 \end{cases}$$

Using Rubin's Rule, results can be pooled as:

$$\begin{cases} \bar{\beta}_1^{\text{MI}} = \left(\beta_1^{s_1} + \beta_1^{s_2} + \beta_1^{s_3}\right) = \frac{1}{3}\left(-11.59 - 11.54 + -9.68\right) \approx -10.94, \\ \bar{\beta}_0^{\text{MI}} = \frac{1}{3}\left(\beta_0^{s_1} + \beta_0^{s_2} + \beta_0^{s_3}\right) = \frac{1}{3}\left(758.1 + 762 + 754.7\right) \approx 758.3. \end{cases}$$

And the final model reads as:

$$y^{\text{MI}} = \bar{\beta}_1^{\text{MI}} * \text{temp} + \bar{\beta}_0^{\text{MI}} = -10.94x + 758.3.$$

Temp	Demand
-0.75	766.5
5.75	695.4
16.1	582.2

NB! The same result can be obtained by directly pooling (averaging) estimates from all samples. In other words, average values for the same datapoint across all samples.
A. Miles. Obtaining predictions from models fit to multiply imputed data, *Sociological Methods & Research* 45(1), pp. 1–11, 2015.

MULTIVARIATE IMPUTATION BY CHAINED EQUATIONS (MICE) ALGORITHM

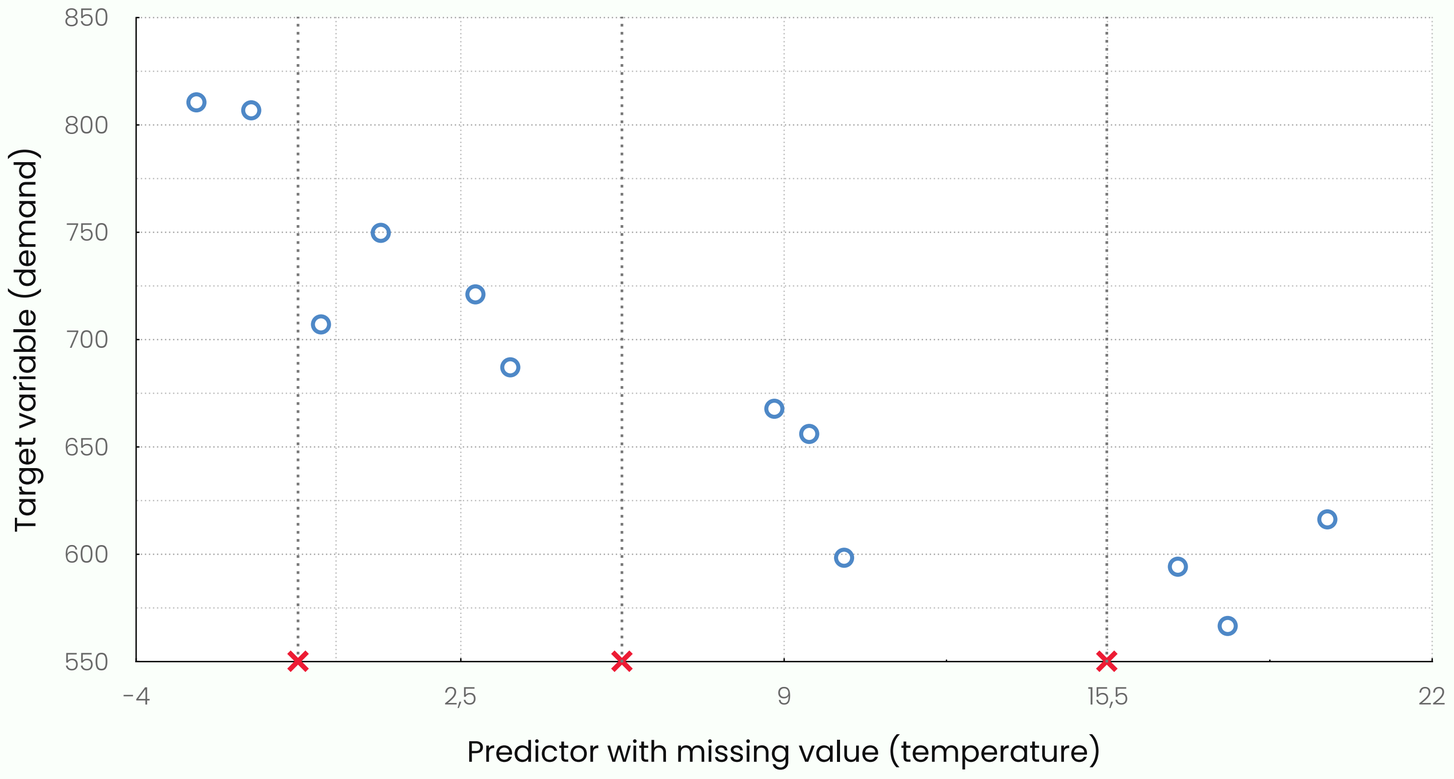
Table 1. Multivariate imputation by chained equations (MICE) algorithm for multiple imputation

1. Specify an imputation model for each of the k variables that are subject to missing data.
2. For each of the k variables that are subject to missing data, fill in the missing values with random draws from those subjects with observed values for the variable in question. Note that these initial imputed values do not respect the multivariate relations in the data and will be overwritten by better imputed values in later stages of the algorithm.
3. For the first variable that is subject to missing data:
   a. Regress this first variable on all the other variables using those subjects with complete data on the first variable and observed or currently imputed values of the other variables.
   b. The estimated regression coefficients and their variance-covariance matrix (and the estimated variance of the residual distribution if a linear regression model was fit for a continuous variable) are extracted from the regression model estimated in (a).
   c. Using the quantities obtained in (b), randomly perturb the estimated regression coefficients in a way that reflects the degree of uncertainty arising from the data.
   d. Using the set of perturbed regression coefficients obtained in (c), the conditional distribution of the first variable is determined for each subject with missing data on that variable.
   e. A value of the variable is drawn from this conditional distribution for each subject with missing data on the first variable.
4. Repeat step 3 for each of the variables that is subject to missing data. Steps 3 and 4 form 1 cycle of the imputation process for creating 1 imputed data set.
5. Repeat steps 3 and 4 the desired number of times (suggested 5 to 20 cycles). The final imputed values are used as the imputed values in first imputed data set.
6. Repeat steps 2-5 M times to produce M imputed data sets (the choice of M, the number of imputed data sets, is discussed in the section How Many Imputations: How Large Should M Be?).

P. C. Austin, I. R. White, D. S Lee, and S. van Buuren. Missing data in clinical research: A tutorial on multiple imputation, *Canadian Journal of Cardiology* 37(9), pp. 1322–1331, 2021.
S. van Buuren and K. Groothuis-Oudshoorn. mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, *45*(3), pp. 1–67, 2011.

EXAMPLE MICE: SIMPLIFIED

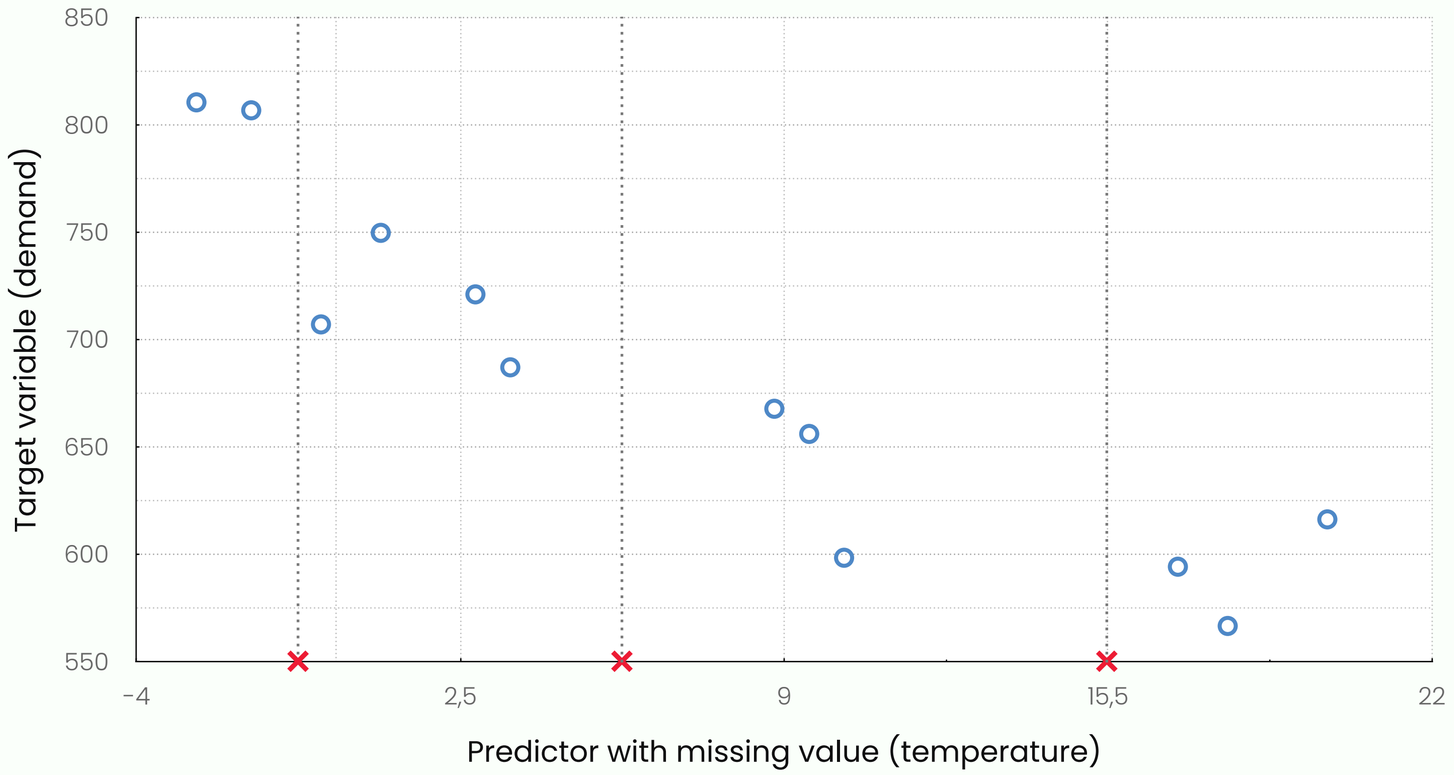
Temp	Demand
-2.8	810.5
-0.75	N/A
2.8	721
5.75	N/A
8.8	667.8
10.2	598.5
15.5	N/A
19.9	616
17.9	566.7
16.9	594.3
9.5	656.1
3.5	687.1
0.9	749.7
-0.3	707.1
-1.7	806.7



EXAMPLE MICE: SIMPLIFIED

Temp	Demand
-2.8	810.5
-0.75	N/A
2.8	721
5.75	N/A
8.8	667.8
10.2	598.5
15.5	N/A
19.9	616
17.9	566.7
16.9	594.3
9.5	656.1
3.5	687.1
0.9	749.7
-0.3	707.1
-1.7	806.7

$\bar{x} = 681.8$



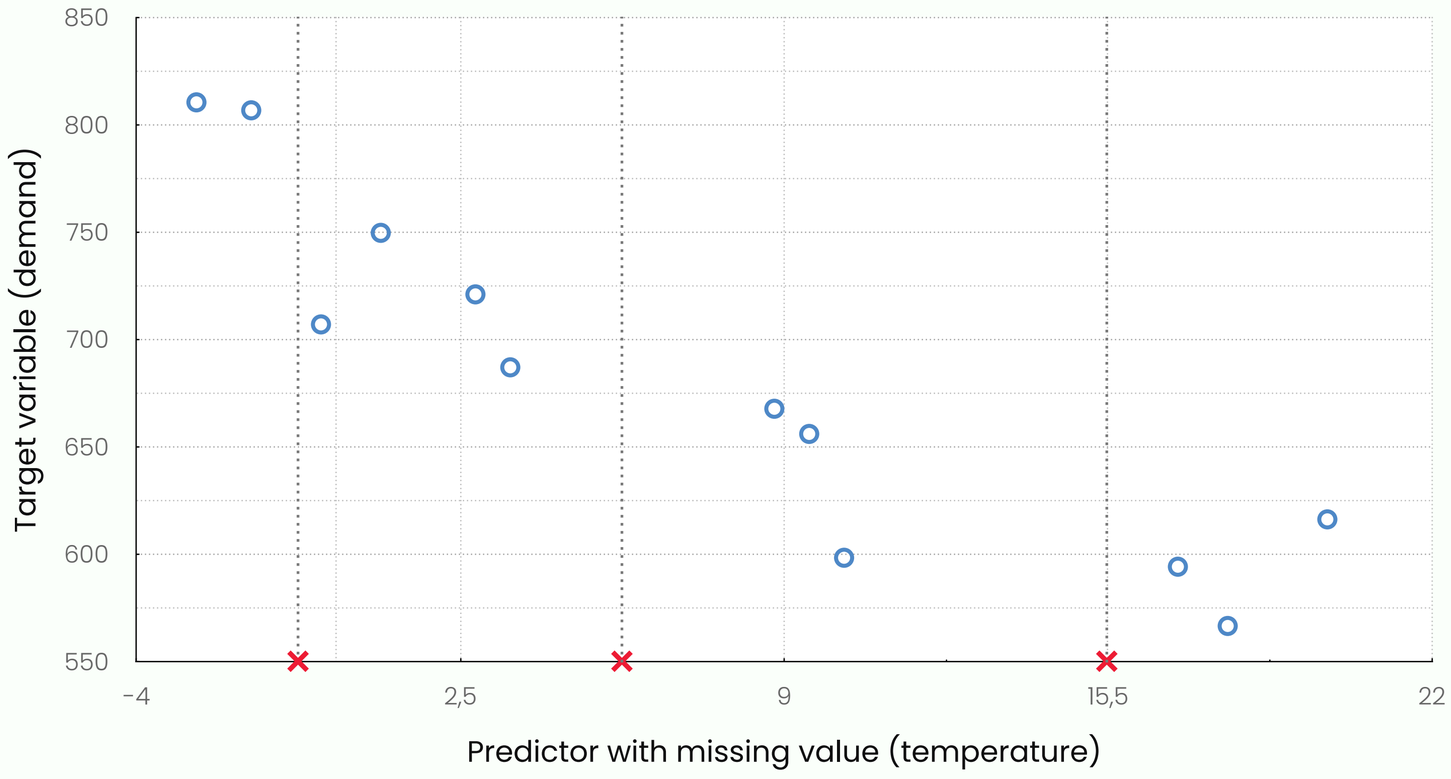
EXAMPLE MICE: SIMPLIFIED

Temp	Demand
-2.8	810.5
-0.75	N/A
2.8	721
5.75	N/A
8.8	667.8
10.2	598.5
15.5	N/A
19.9	616
17.9	566.7
16.9	594.3
9.5	656.1
3.5	687.1
0.9	749.7
-0.3	707.1
-1.7	806.7

Iter0

$\bar{x} = 681.8$

T	D
-0.75	681.8
5.75	681.8
15.5	681.8



EXAMPLE MICE: SIMPLIFIED

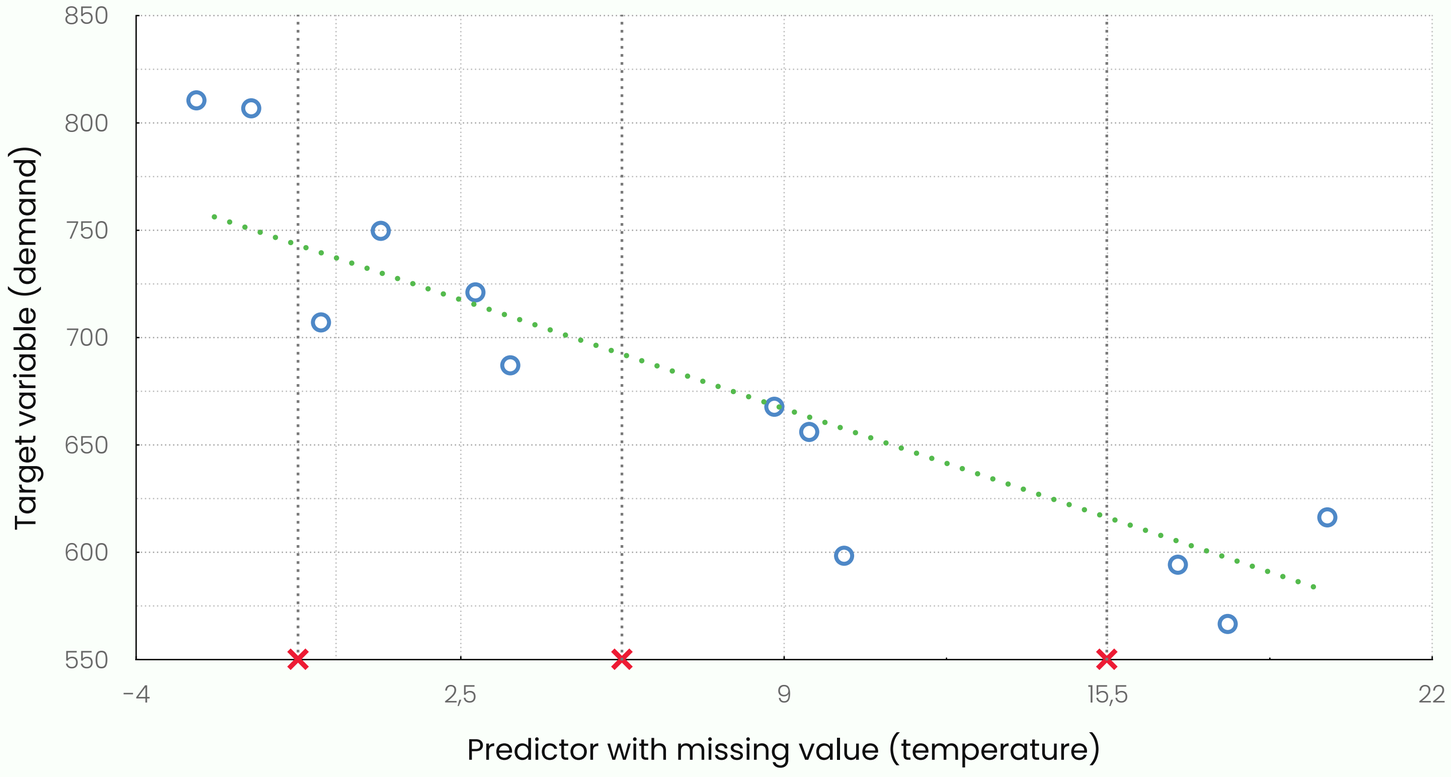
Temp	Demand
-2.8	810.5
-0.75	N/A
2.8	721
5.75	N/A
8.8	667.8
10.2	598.5
15.5	N/A
19.9	616
17.9	566.7
16.9	594.3
9.5	656.1
3.5	687.1
0.9	749.7
-0.3	707.1
-1.7	806.7

Iter0

$\bar{x} = 681.8$

$y = -7.78x + 736.87$

T	D
-0.75	681.8
5.75	681.8
15.5	681.8



EXAMPLE MICE: SIMPLIFIED

Temp	Demand
-2.8	810.5
-0.75	N/A
2.8	721
5.75	N/A
8.8	667.8
10.2	598.5
15.5	N/A
19.9	616
17.9	566.7
16.9	594.3
9.5	656.1
3.5	687.1
0.9	749.7
-0.3	707.1
-1.7	806.7

Iter0

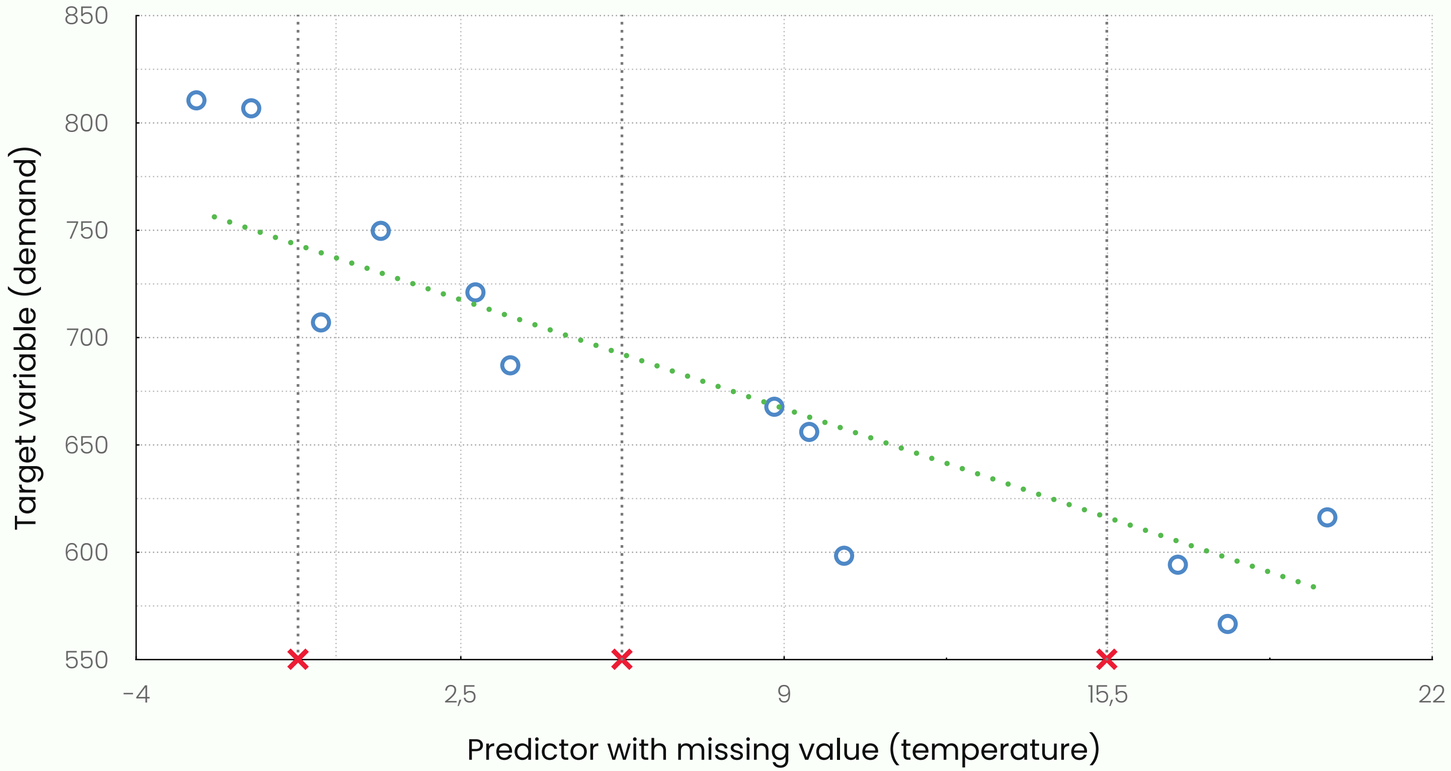
$\bar{x} = 681.8$

T	D
-0.75	681.8
5.75	681.8
15.5	681.8

$y = -7.78x + 736.87$

Iter1

T	D
-0.75	742.7
5.75	692.1
15.5	616.3



EXAMPLE MICE: SIMPLIFIED

Temp	Demand
-2.8	810.5
-0.75	N/A
2.8	721
5.75	N/A
8.8	667.8
10.2	598.5
15.5	N/A
19.9	616
17.9	566.7
16.9	594.3
9.5	656.1
3.5	687.1
0.9	749.7
-0.3	707.1
-1.7	806.7

Iter0

$\bar{x} = 681.8$

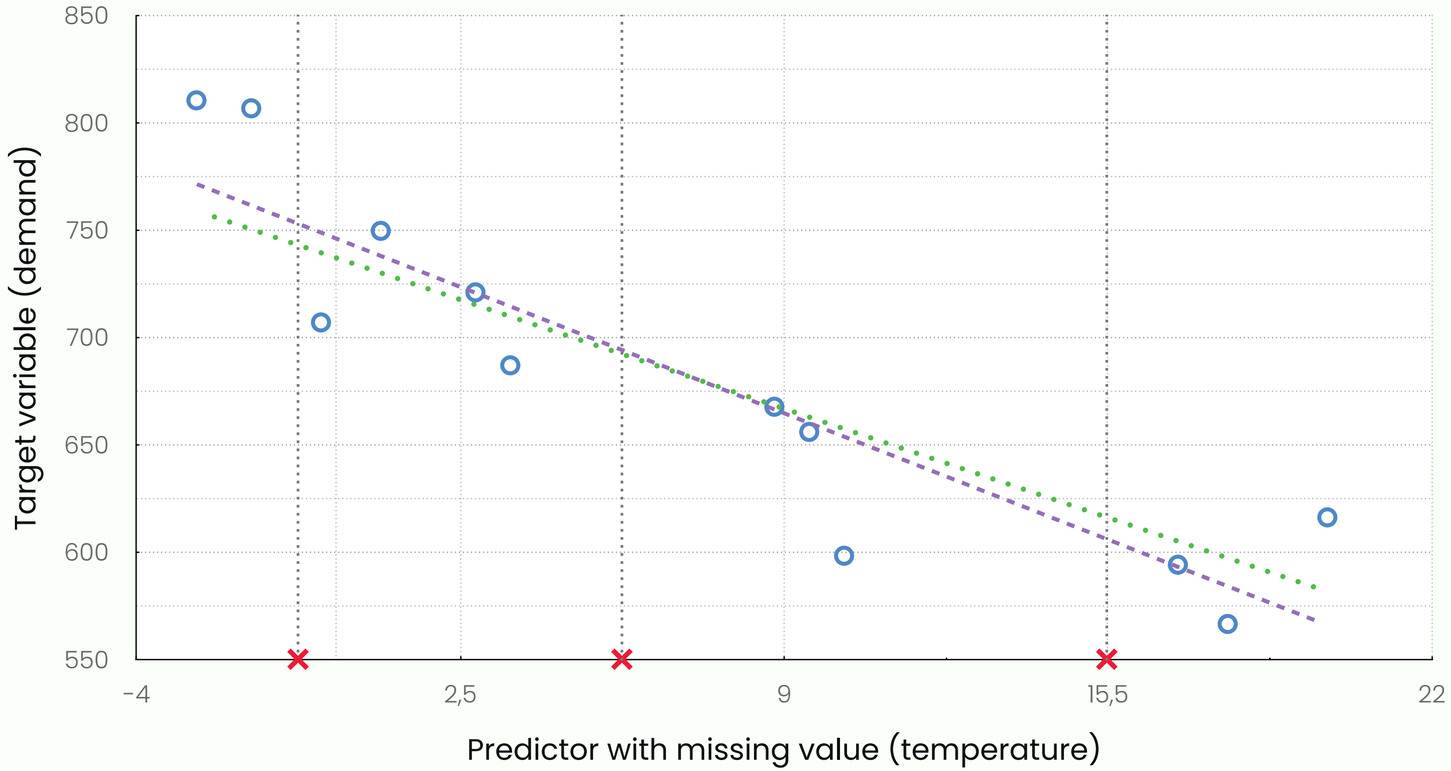
T	D
-0.75	681.8
5.75	681.8
15.5	681.8

$y = -7.78x + 736.87$

Iter1

T	D
-0.75	742.7
5.75	692.1
15.5	616.3

$y = -9.04x + 746.14$



EXAMPLE MICE: SIMPLIFIED

Temp	Demand
-2.8	810.5
-0.75	N/A
2.8	721
5.75	N/A
8.8	667.8
10.2	598.5
15.5	N/A
19.9	616
17.9	566.7
16.9	594.3
9.5	656.1
3.5	687.1
0.9	749.7
-0.3	707.1
-1.7	806.7

Iter0

$\bar{x} = 681.8$

T	D
-0.75	681.8
5.75	681.8
15.5	681.8

$y = -7.78x + 736.87$

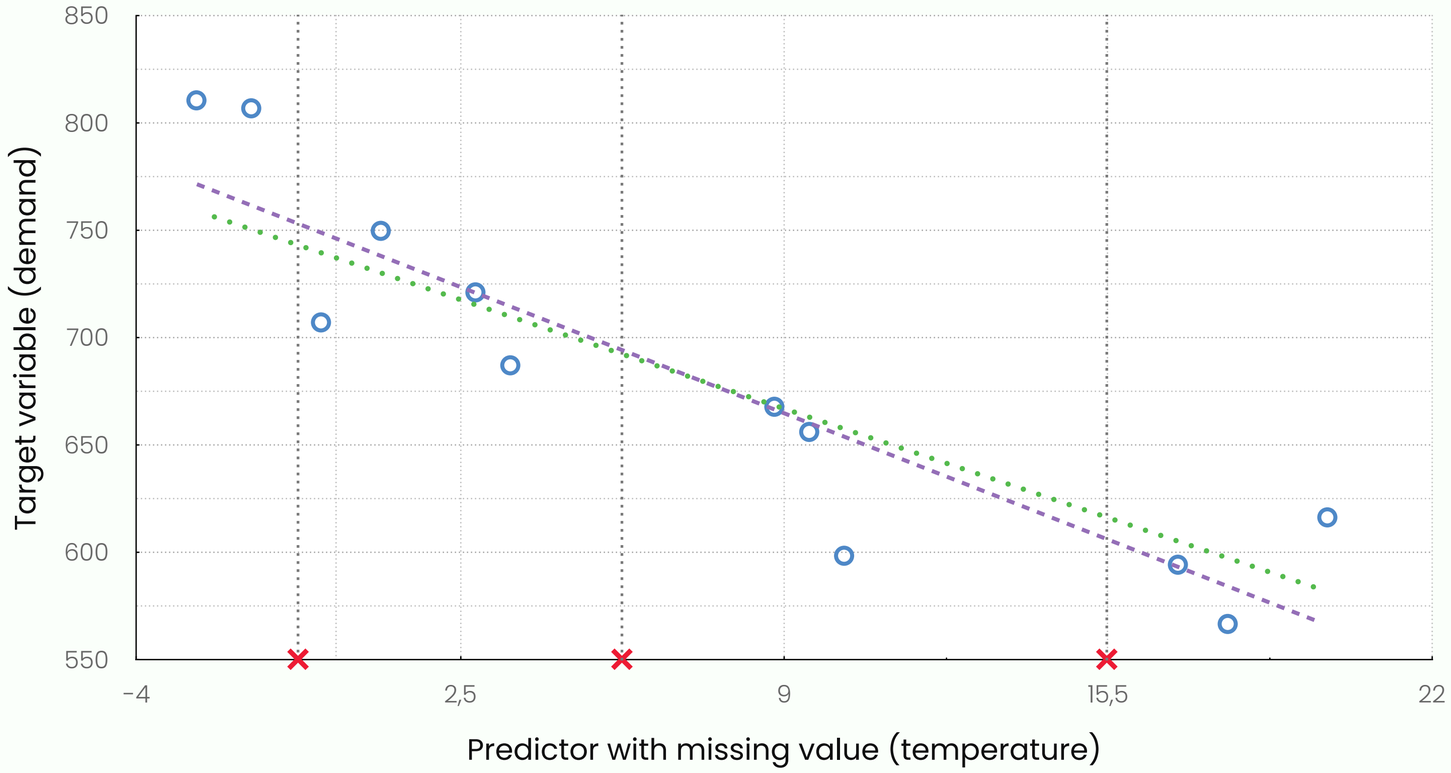
Iter1

T	D
-0.75	742.7
5.75	692.1
15.5	616.3

$y = -9.04x + 746.14$

Iter2

T	D
-0.75	752.9
5.75	694.2
15.5	606



EXAMPLE MICE: SIMPLIFIED

Temp	Demand
-2.8	810.5
-0.75	N/A
2.8	721
5.75	N/A
8.8	667.8
10.2	598.5
15.5	N/A
19.9	616
17.9	566.7
16.9	594.3
9.5	656.1
3.5	687.1
0.9	749.7
-0.3	707.1
-1.7	806.7

Iter0

$\bar{x} = 681.8$

T	D
-0.75	681.8
5.75	681.8
15.5	681.8

$y = -7.78x + 736.87$

Iter1

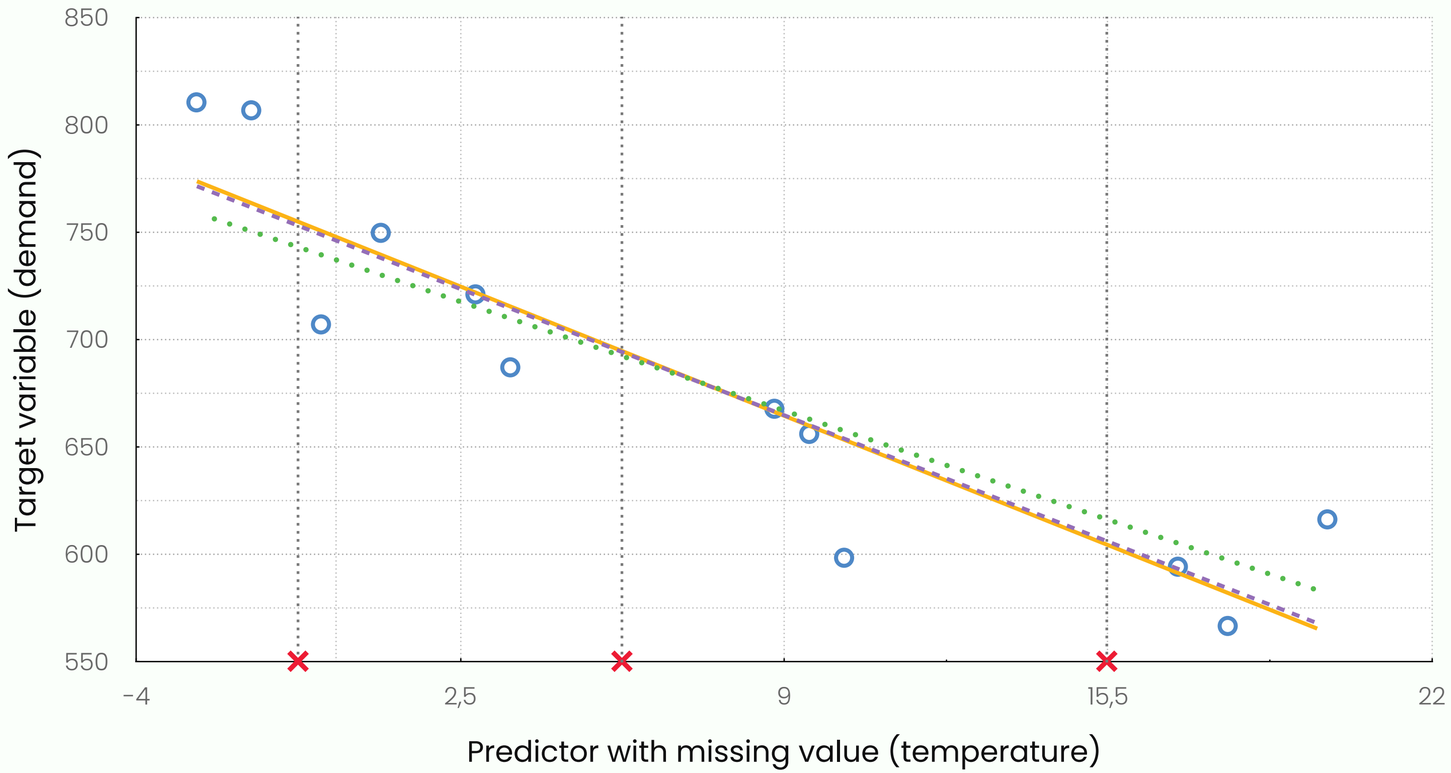
T	D
-0.75	742.7
5.75	692.1
15.5	616.3

$y = -9.04x + 746.14$

Iter2

T	D
-0.75	752.9
5.75	694.2
15.5	606

$y = -9.25x + 747.72$



EXAMPLE MICE: SIMPLIFIED

Temp	Demand
-2.8	810.5
-0.75	N/A
2.8	721
5.75	N/A
8.8	667.8
10.2	598.5
15.5	N/A
19.9	616
17.9	566.7
16.9	594.3
9.5	656.1
3.5	687.1
0.9	749.7
-0.3	707.1
-1.7	806.7

Iter0

$\bar{x} = 681.8$

T	D
-0.75	681.8
5.75	681.8
15.5	681.8

$y = -7.78x + 736.87$

Iter1

T	D
-0.75	742.7
5.75	692.1
15.5	616.3

$y = -9.04x + 746.14$

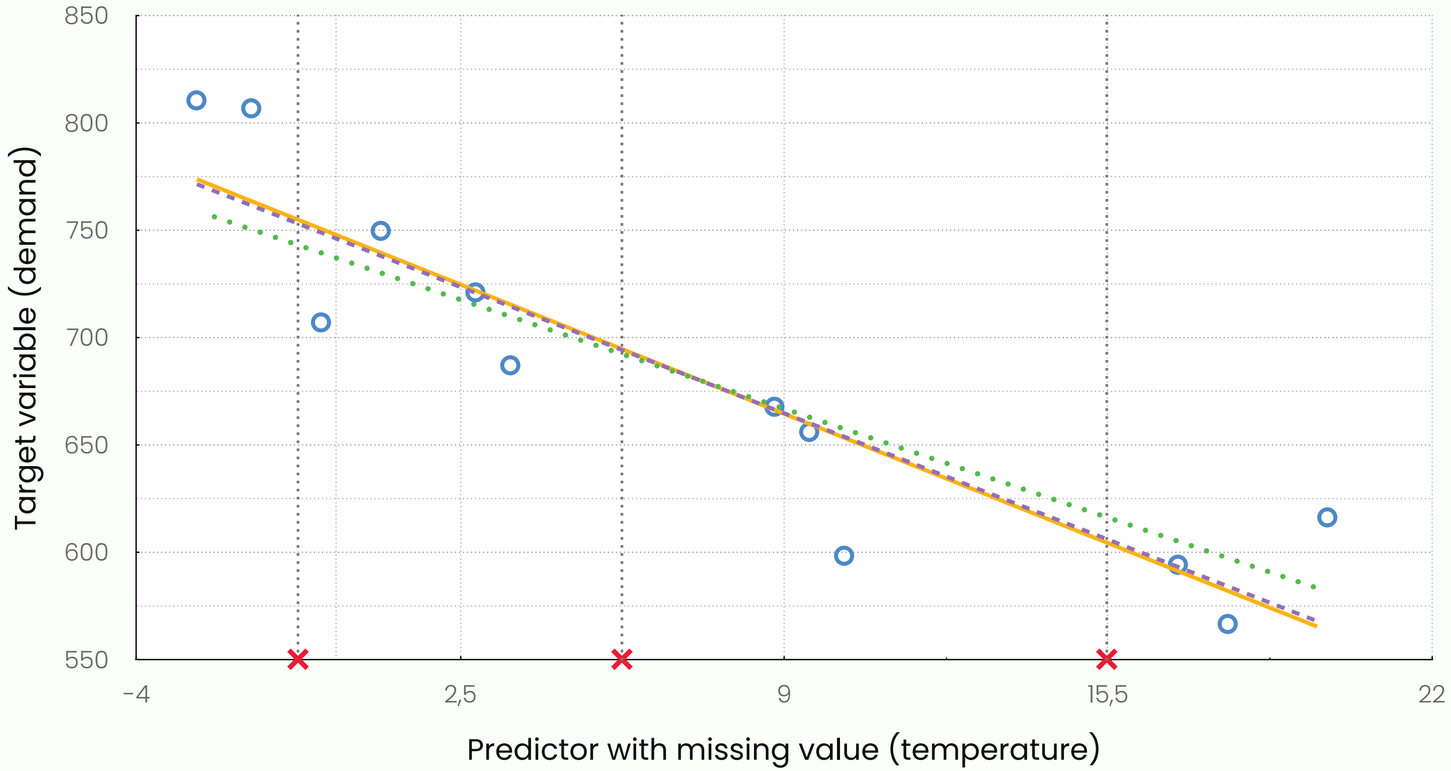
Iter2

T	D
-0.75	752.9
5.75	694.2
15.5	606

$y = -9.25x + 747.72$

Iter3

T	D
-0.75	754.7
5.75	694.6
15.5	604.4



SOME MORE MI METHODS

- Predictive mean matching imputation
- Linear discriminant analysis
- Seasonal trend decomposition using LOESS imputation
- ML models (eg LR, RFR, NN, SVM) can be trained on complete records to predict incomplete data.
- Some algorithms like XGBoost and LightGBM can handle missing values without any preprocessing, by supplying relevant parameters.

SUMMARY: PRACTICAL ADVICES

Technique	MCAR	MAR	MNAR
List-wise deletion	unbiased large std. errors simple	biased large std. errors simple	biased large std. errors simple
Pair-wise deletion	unbiased inaccurate std. errors	biased inaccurate std. errors	biased inaccurate std. errors
Single imputation	often biased inaccurate std. errors	often biased inaccurate std. errors	biased inaccurate std. errors
Multiple Imputation	unbiased accurate std. errors	unbiased accurate std. errors	biased accurate std. errors

HOME ACTIVITIES & BRAIN EXERCISE

✳ Use file “Dataset_for_Home_Activities.xlsx”, and data for Demand (NaNs).

✳ Univariate Imputation: use *Mean* and *Linear regression* methods.

✳ Multivariate Imputation: complete $M = 3$ statistical analyses on created samples 1-3. Calculate variances ($\sigma_0^{2,m}$ and $\sigma_1^{2,m}$) for each completed dataset.

- Using Rubin’s Rule, calculate the within- and between-imputation variances:

$$\mathrm{Var}(\beta_i) = \frac{1}{M}\sum_{m=1}^{M}\sigma_i^{2,m} + \frac{1}{M-1}\sum_{m=1}^{M}\left(\beta_i^m - \beta_i^{\mathrm{MI}}\right)^2.$$

✳ For which case the standard errors ($\mathrm{SE}(\beta_i) = \sqrt{\mathrm{Var}(\beta_i)}$) are smaller?

✳ Compare all three imputations to the true values. Which method performed better?

Thank you!

Questions?