Missing values and outliers

Luigi Borzì

Overview

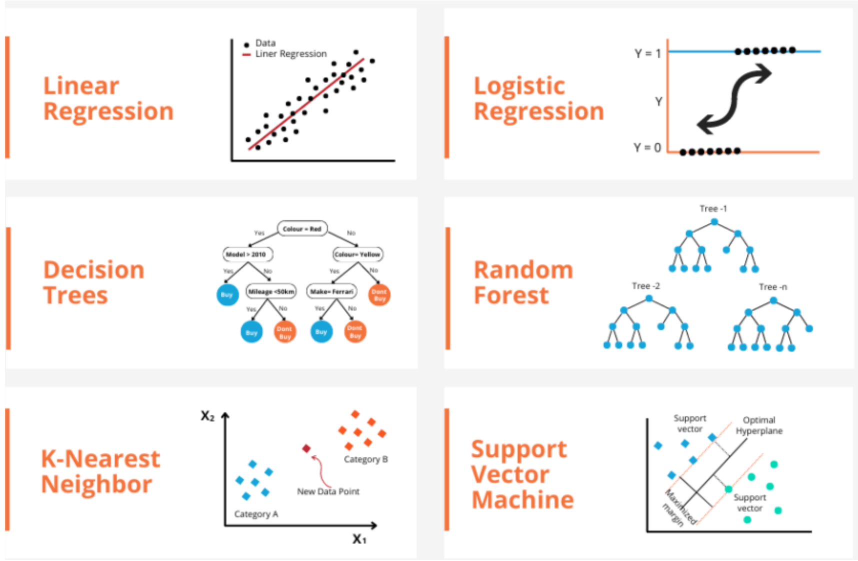
Preprocessing	Outliers
	Missing values
	Resampling
	Filtering
	Segmentation
	Data representation and transformation
	Feature extraction
	Feature selection/dimensionality reduction
	Normalization/standardization

Why preprocessing?

Tabular data

Sample	Feature a	Feature b	Label
1	0.4	12	0
2	0.3	24	1
3	0.2	13	0
4	0.3	25	1

Machine learning algorithms



Why preprocessing?

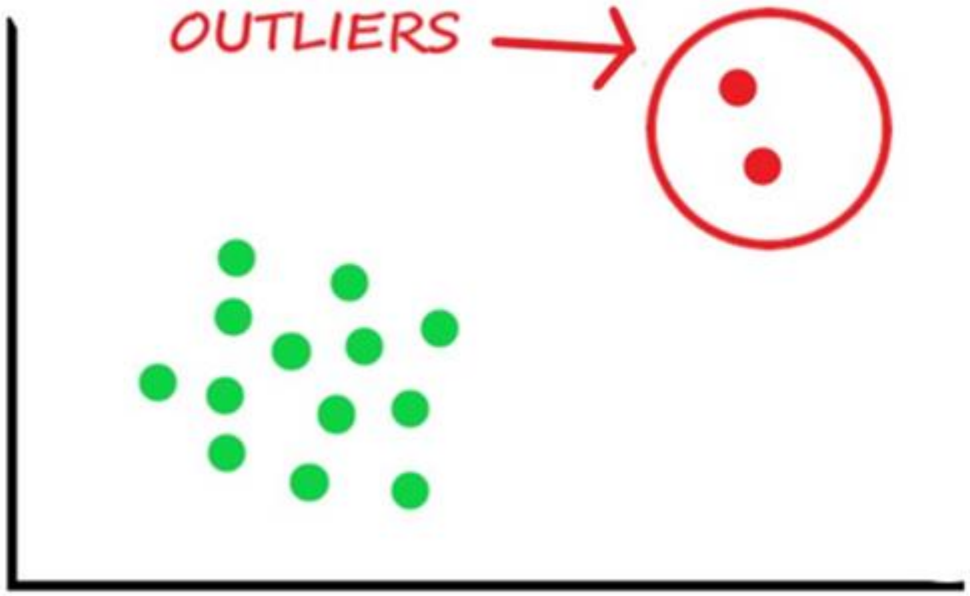
Semi-structured data

Patient ID	Diagnosis	Medication	Vital Signs	Laboratory Results
001	Hypertension	Lisinopril 10mg	Blood Pressure: 140/90	Cholesterol: 210 mg/dL
			Heart Rate: 72 bpm	Glucose: 110 mg/dL
			Respiratory Rate: 18	Hemoglobin A1c: 6.0%
			Temperature: 98.6°F	
			Oxygen Saturation: 97%	
002	Type 2 Diabetes	Metformin 500mg BID	Blood Pressure: 130/80	Hemoglobin A1c: 8.2%
		Sitagliptin 100mg QD	Heart Rate: 80 bpm	Glucose: 180 mg/dL
			Respiratory Rate: 16	Cholesterol: 220 mg/dL
			Temperature: 98.4°F	Microalbuminuria: Positive
			Oxygen Saturation: 98%	

Tabular data

ID	Diagnosis	Medication	Blood pressure	Heart rate	Respiratory rate	Oxygen saturation	Glucose
001	Hypertension	Lisinopril 10mg	140/90	72 bpm	18	97%	110 mg/dL
002	Type 2 Diabetes	Metformin 500mg BID	130/80	80 bpm	16	98%	NaN

Outliers



- Abnormal data-points (values)
- Data-points distant from all the other data-points

Outliers: why remove them?

- Data Quality: Improve the overall quality of the dataset and reduce the likelihood of training a model on incorrect or misleading information.

- Model Performance: Outliers can have a significant impact on the performance of machine learning models, especially those sensitive to the scale and distribution of the data. Removing outliers can help prevent models from being skewed or biased by extreme values.

- Robustness: Removing outliers can make models more robust and resistant to noise in the data. Models trained on clean datasets are generally better able to generalize to unseen data and perform well in real-world scenarios.

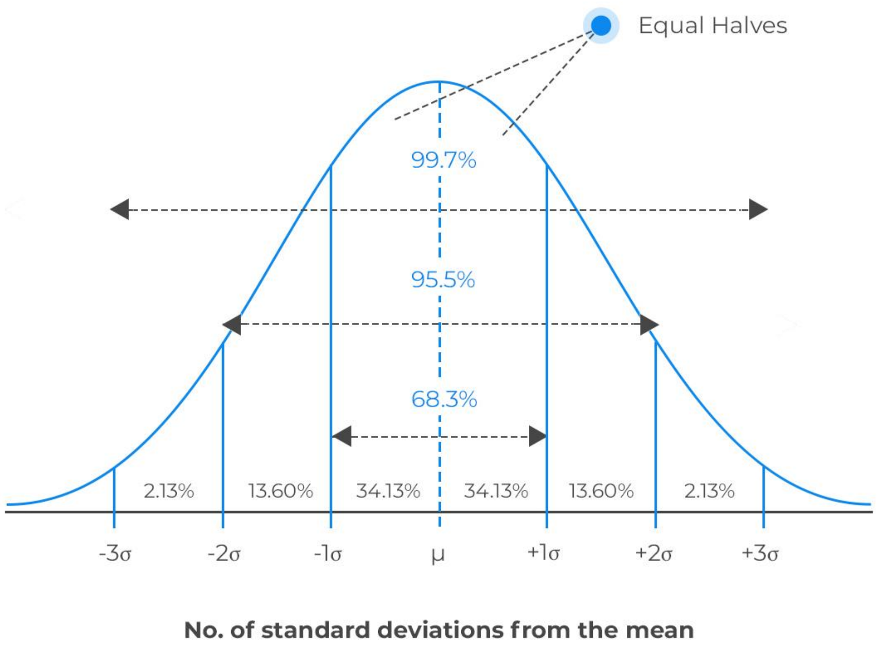
- Interpretability: Outliers can distort the interpretation of results and make it difficult to draw meaningful insights from the data. Removing them can lead to more accurate and interpretable models.

- Assumption Violation: Many machine learning algorithms make certain assumptions about the distribution of the data, such as normality or homoscedasticity. Outliers can violate these assumptions and lead to biased estimates and unreliable predictions.

Outliers: distribution

To define outliers, we should first define what is “normal”. Then we can identify what is not “normal”

Normal distribution



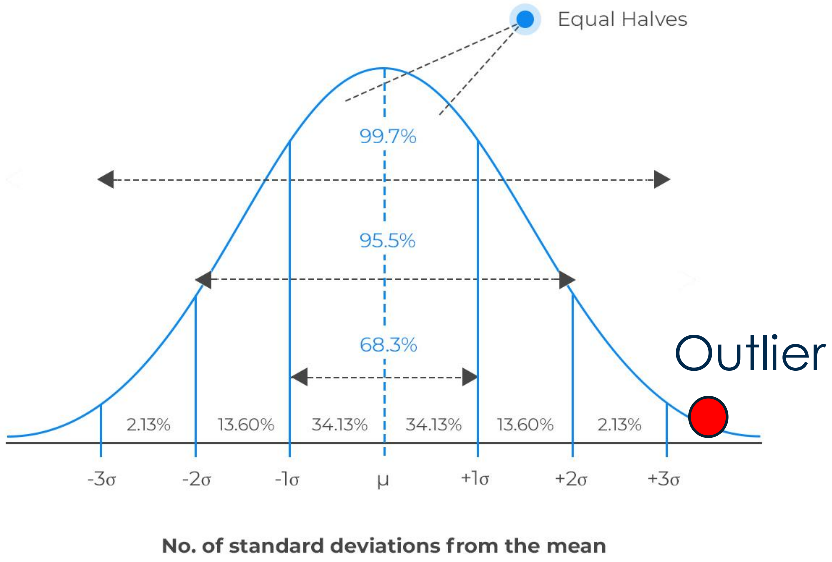
$$f(x|\mu,\sigma) = \frac{1}{\sigma \cdot 2\pi} \cdot \mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$\text{mean} = \frac{\sum_{\mathrm{i}=1}^{N} x_{\mathrm{i}}}{N} \qquad \text{std} = \frac{1}{N}\sqrt{\sum_{i=1}^{N}(x_i - \mu)^2}$$

Outliers: distribution

To define outliers, we should first define what is “normal”. Then we can identify what is not “normal”

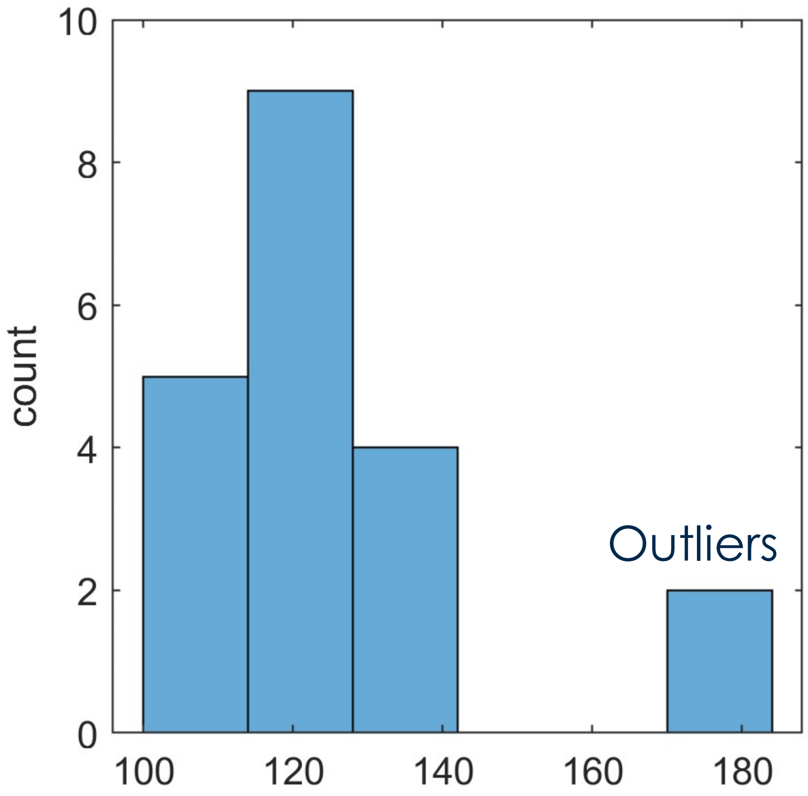
Normal distribution



Outliers: distribution

To define outliers, we should first define what is "normal". Then we can identify what is not "normal"

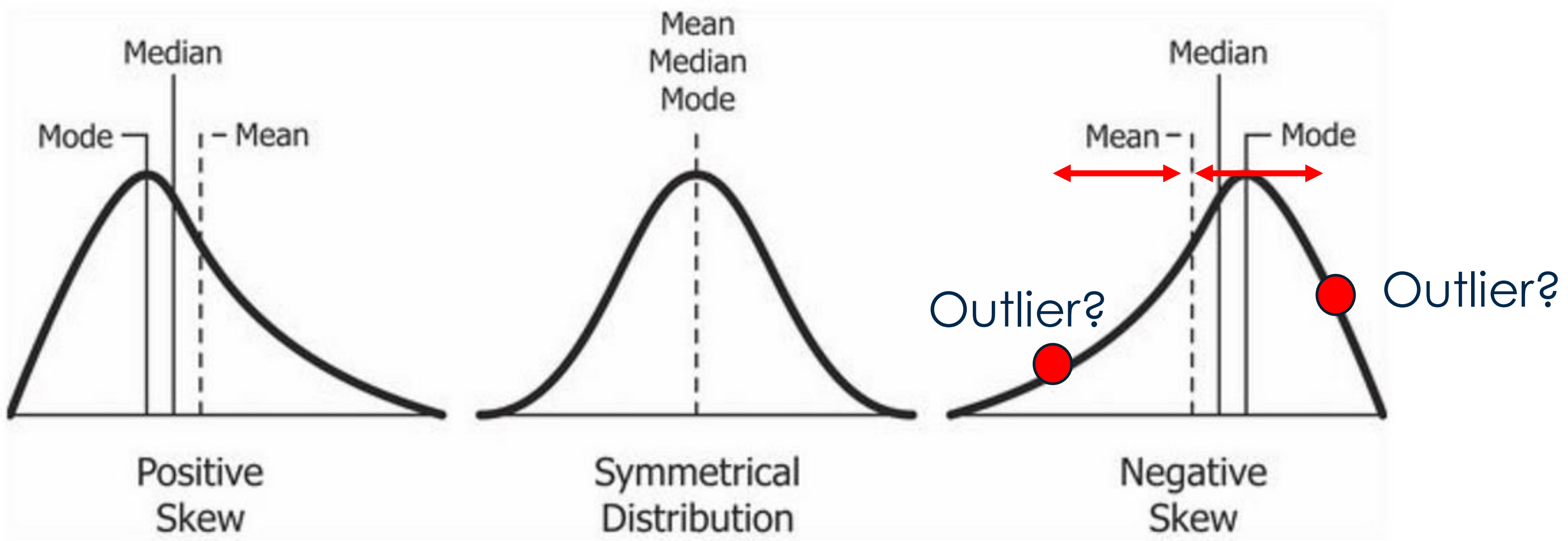
Histogram



You can consider outliers data-points that are 2σ or 3σ far from the mean

Outliers: distribution

To define outliers, we should first define what is “normal”. Then we can identify what is not “normal”

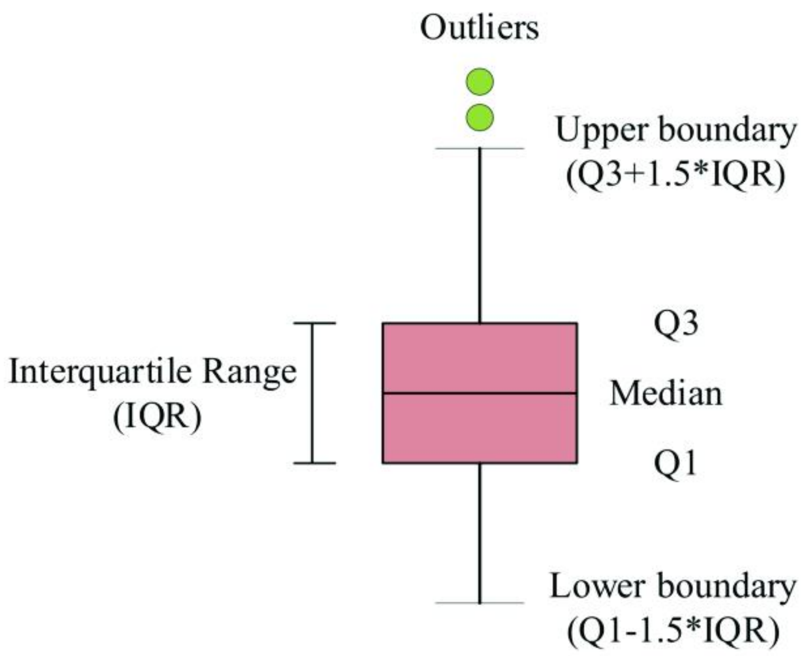


Be careful with skewed distribution. Not all distances from the mean are the same!

Outliers: distribution

To define outliers, we should first define what is “normal”. Then we can identify what is not “normal”

Boxplot



First Quartile (Q1): value below which 25% of the data fall. Q1 is also known as the 25th percentile.

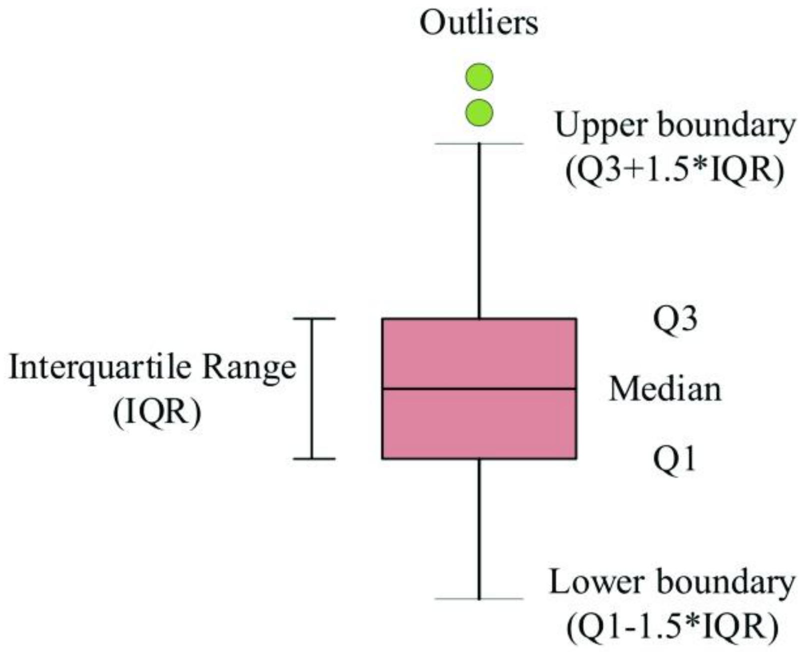
Third Quartile (Q3): value below which 75% of the data fall. Q3 is also known as the 75th percentile.

Second Quartile: value below which 50% of the data fall. Q2 is also known as the 50th percentile or median.

Outliers: distribution

To define outliers, we should first define what is “normal”. Then we can identify what is not “normal”

Boxplot



- Interquartile range (IQR) method: data points that fall below Q1 - 1.5 * IQR or above Q3 + 1.5 * IQR are considered outliers.
- Mean Method: data points far more than three standard deviations from the mean are considered outliers.

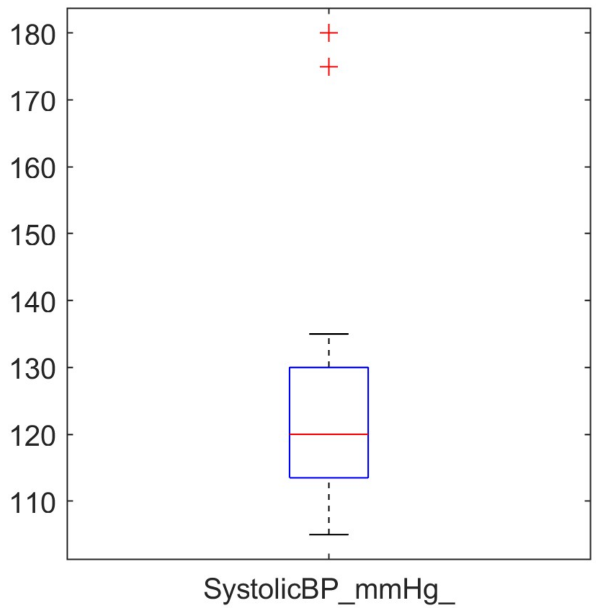
Outliers: sample size

Patient ID	Age (years)	Gender	Systolic BP (mmHg)	Diastolic BP (mmHg)	Heart Rate (bpm)	Glucose Level (mg/dL)	Cholesterol Level (mg/dL)
1	45	Male	110	80	72	90	180
2	35	Female	110	70	65	95	200
3	50	Male	100	80	80	105	220
4	28	Female	105	75	68	88	190
5	50	Male	160	100	85	120	250
6	32	Female	108	78	70	98	210
8	40	Female	112	72	60	92	195
9	48	Male	125	85	75	102	215

- Ensure your sample is representative of the population.
- A too small sample (few subjects/patients) is not.
- A biased sample (young subjects, healthy subjects) is not

Outliers: context

Patient ID	Age (years)	Gender	Systolic BP (mmHg)	Diastolic BP (mmHg)	Heart Rate (bpm)	Glucose Level (mg/dL)	Cholesterol Level (mg/dL)
1	45	Male	120	80	72	90	180
2	35	Female	110	70	65	95	200
3	50	Male	130	80	80	105	220
4	28	Female	115	75	68	88	190
5	50	Male	130	90	85	120	250
6	32	Female	118	78	70	98	210
7	55	Male	135	95	85	95	230
8	40	Female	112	72	60	92	195
9	48	Male	125	85	75	102	215
10	38	Female	120	80	70	100	200
11	67	Male	175	115	100	210	280
12	30	Female	105	68	62	85	180
13	75	Male	180	120	95	190	260
14	42	Female	122	78	72	94	205
15	55	Male	130	85	82	160	200
16	36	Female	118	75	68	100	190
17	58	Male	120	80	85	98	225
18	45	Female	110	70	65	93	198
19	50	Male	120	75	80	110	240
20	40	Female	112	72	60	95	200

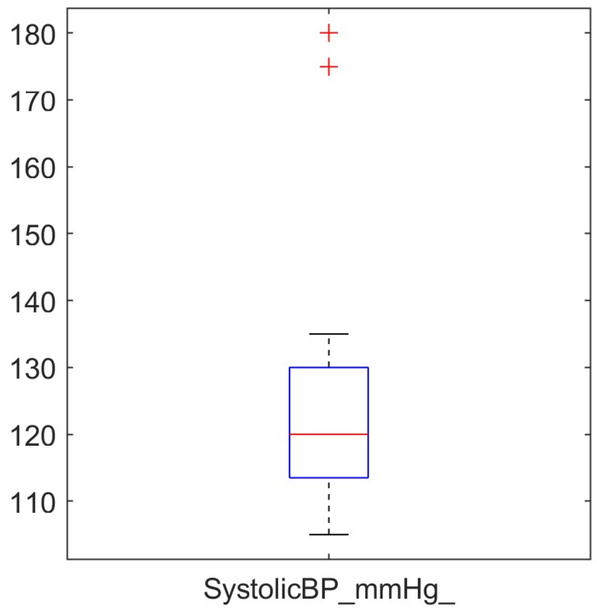


The two outliers are:

- Elderly (age > 65): their age is significantly > than the average age in the sample
- They present high values of systolic and diastolic blood pressure, as well as hear rate and glucose level.
- They probably suffer from hypertension and diabetes.
- So they are not outliers! The sample is small and biased

Outliers: context

Patient ID	Age (years)	Gender	Systolic BP (mmHg)	Diastolic BP (mmHg)	Heart Rate (bpm)	Glucose Level (mg/dL)	Cholesterol Level (mg/dL)
1	45	Male	120	80	72	90	180
2	35	Female	110	70	65	95	200
3	50	Male	130	80	80	105	220
4	25	Female	140	100	90	88	190
5	50	Male	130	90	85	120	250
6	32	Female	118	78	70	98	210
7	55	Male	135	95	85	95	230
8	40	Female	112	72	60	92	195
9	48	Male	125	85	75	102	215
10	38	Female	120	80	70	100	200
11	67	Male	175	115	100	210	280
12	30	Female	105	68	62	85	180
13	75	Male	180	120	95	190	260
14	42	Female	122	78	72	94	205
15	55	Male	130	85	82	160	200
16	36	Female	118	75	68	100	190
17	58	Male	120	80	85	98	225
18	45	Female	110	70	65	93	198
19	50	Male	120	75	80	110	240
20	40	Female	112	72	60	95	200



By looking at the boxplot, this does not seem an outlier.
However, considering the age, it can be.
If she is healthy, the values for blood pressure are very high!

Outliers: physiological range

Patient ID	Age (years)	Gender	Systolic BP (mmHg)	Diastolic BP (mmHg)	Heart Rate (bpm)	Glucose Level (mg/dL)	Cholesterol Level (mg/dL)
1	45	Male	120	80	72	90	180
2	35	Female	110	70	65	95	200
3	50	Male	130	80	80	105	220
4	28	Female	115	75	68	88	190
5	50	Male	130	90	85	120	250
6	32	Female	118	78	70	98	210
7	55	Male	135	95	85	95	230
8	40	Female	112	72	60	92	195
9	48	Male	125	85	75	102	215
10	38	Female	120	80	70	100	200
11	67	Male	175	115	100	210	280
12	30	Female	105	68	62	85	180
13	75	Male	180	120	95	190	260
14	42	Female	122	78	72	94	205
15	55	Male	130	85	82	160	200
16	36	Female	118	75	68	100	190
17	58	Male	120	80	85	98	225
18	45	Female	110	70	65	93	198
19	50	Male	120	75	80	110	240
20	40	Female	112	72	60	95	200

Blood Pressure: 80-120 mm Hg (hypertension if > 90-130)

Heart Rate: 60-100 bpm at rest

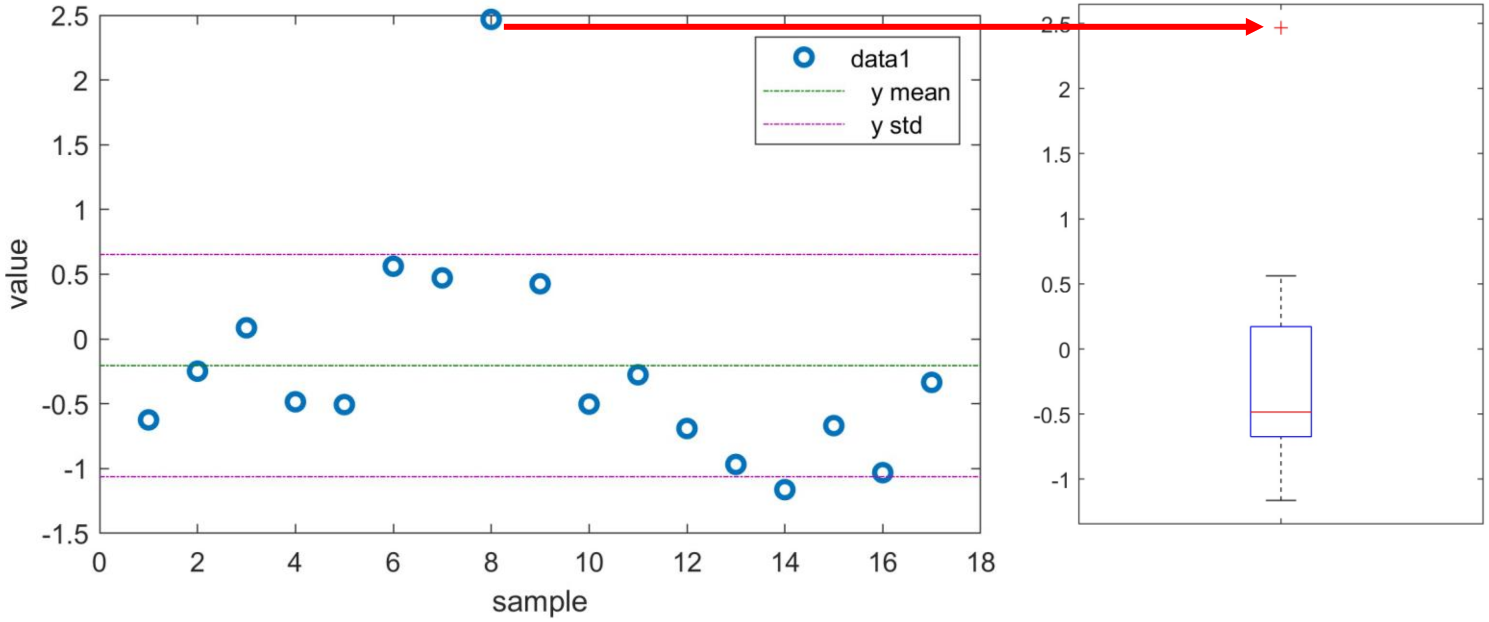
Glucose Level: < 100 mg/dL (diabetes if >130)

Cholesterol Level: < 200 mg/dL (high if > 240)

Oxygen saturation: > 95% (hypoxemia if < 90%)

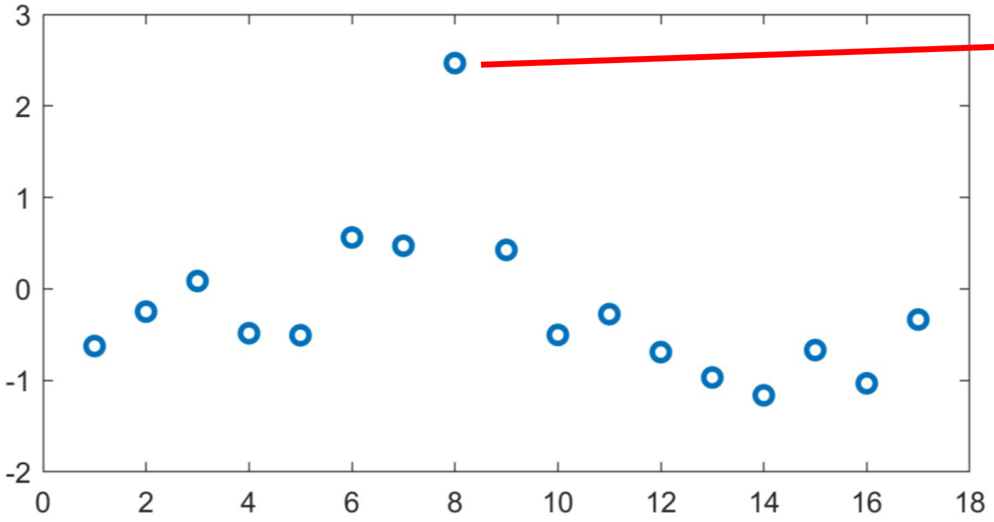
You can discard impossible values (oxygen <60%, glucose > 600, heart rate > 200, ….

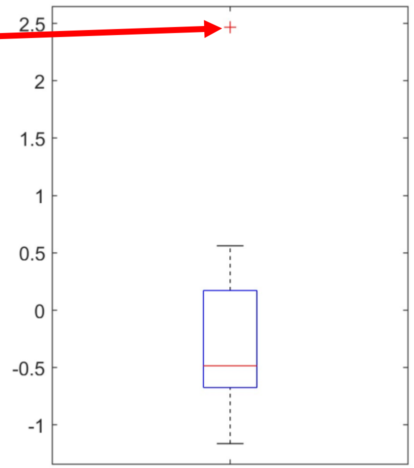
Outliers: time series

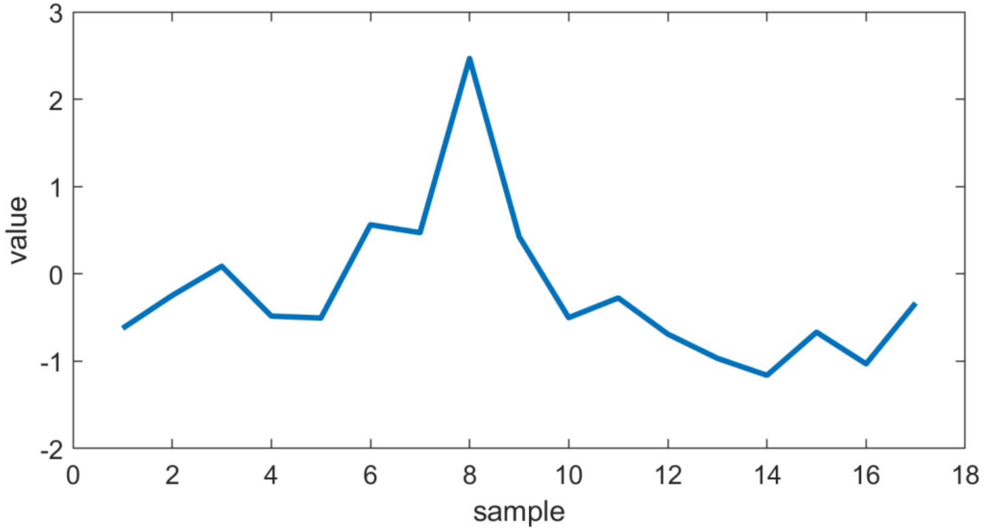


Considering the overall mean and std, this seems an outlier

Outliers: time series





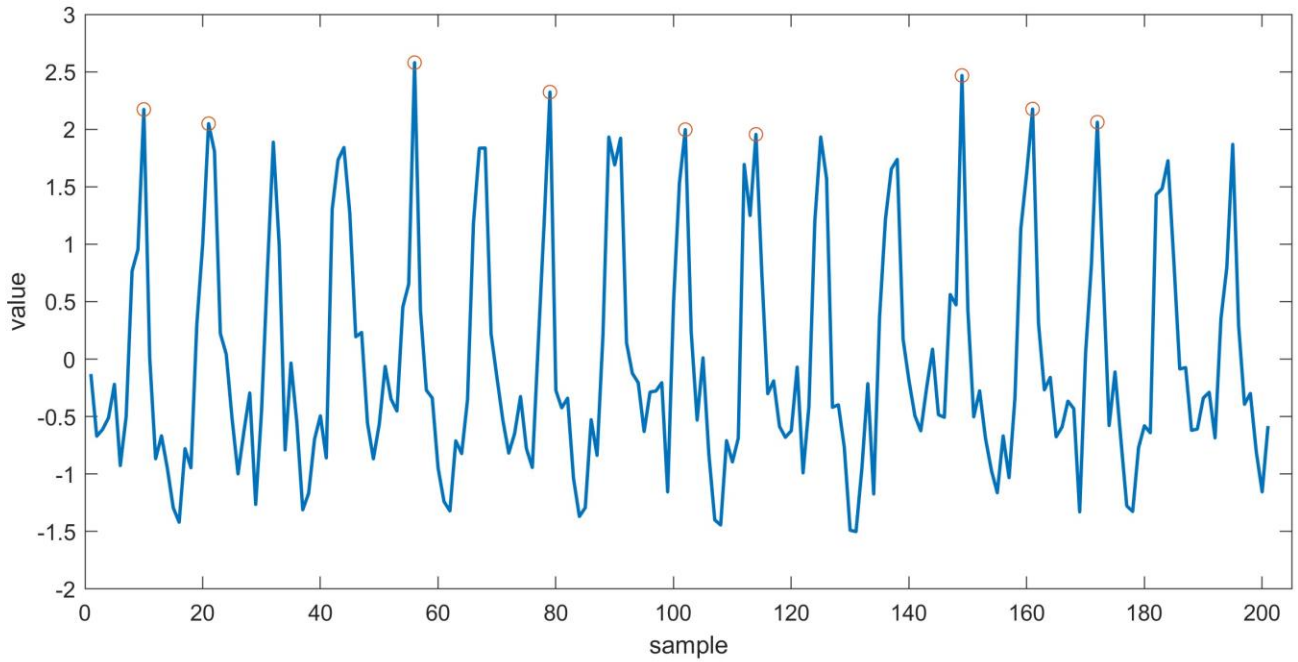


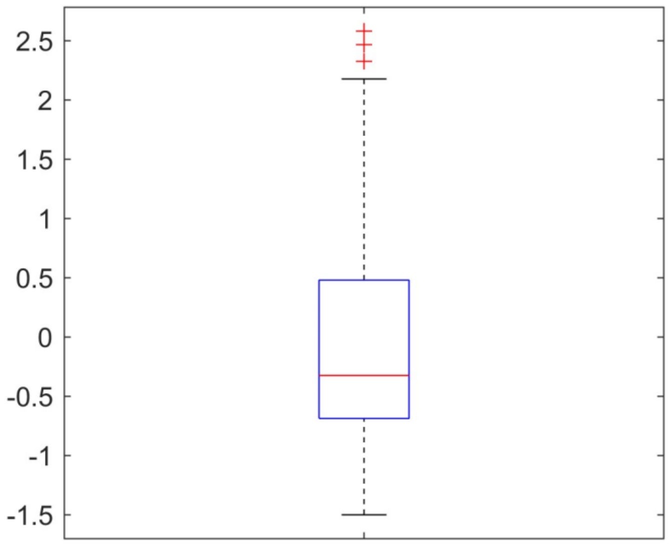
But it is not!

It is an acceleration signal recorded during walking.
The data-point corresponds to the contact of the heel to the ground.

It is of utmost importance!

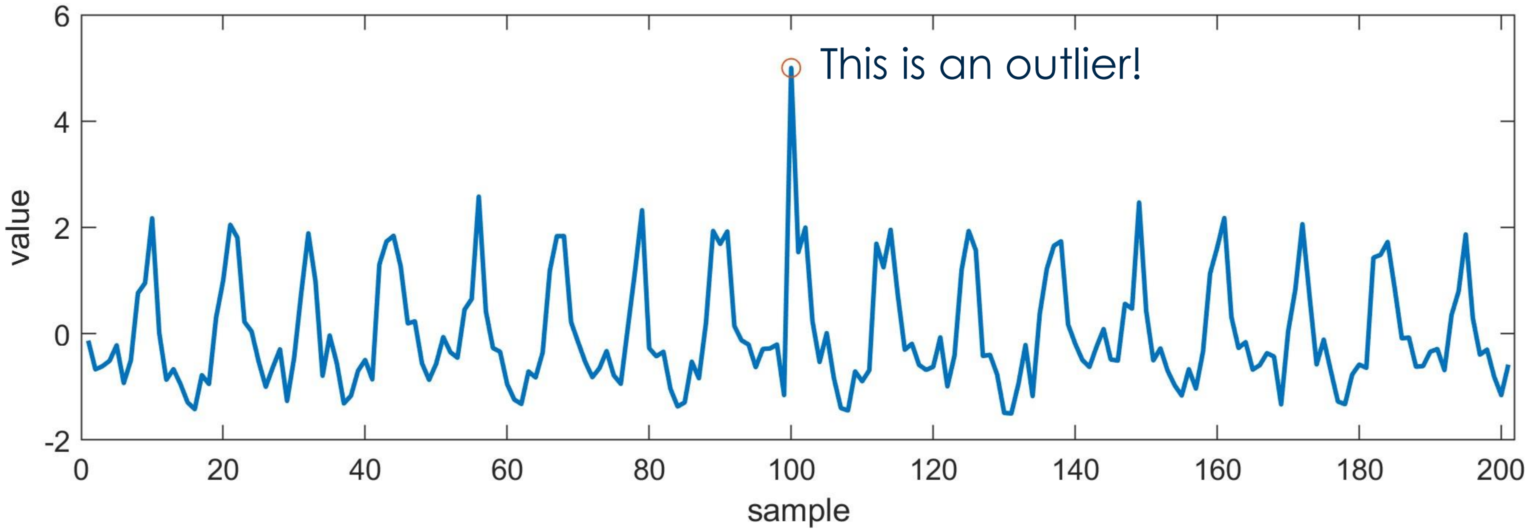
Outliers: time series



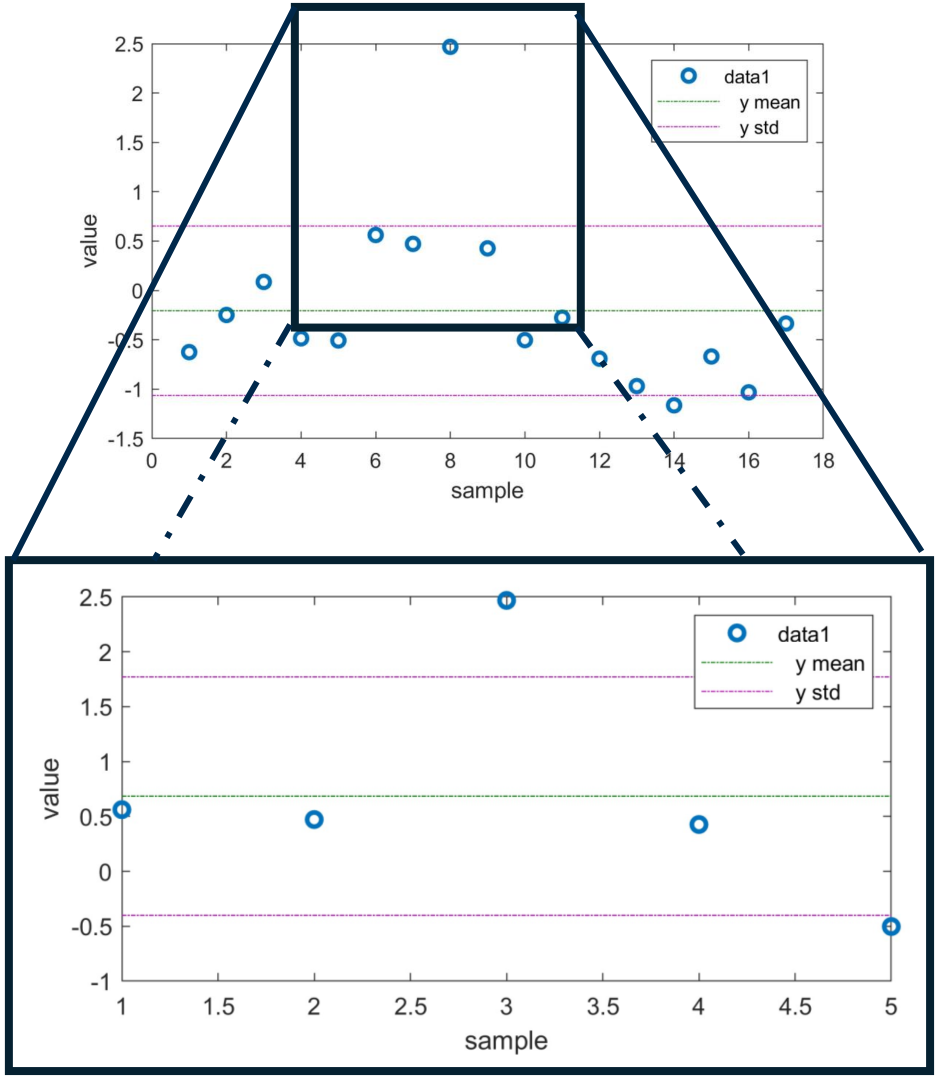


These seem outliers, but they are not!

Outliers: time series



Outliers: time series

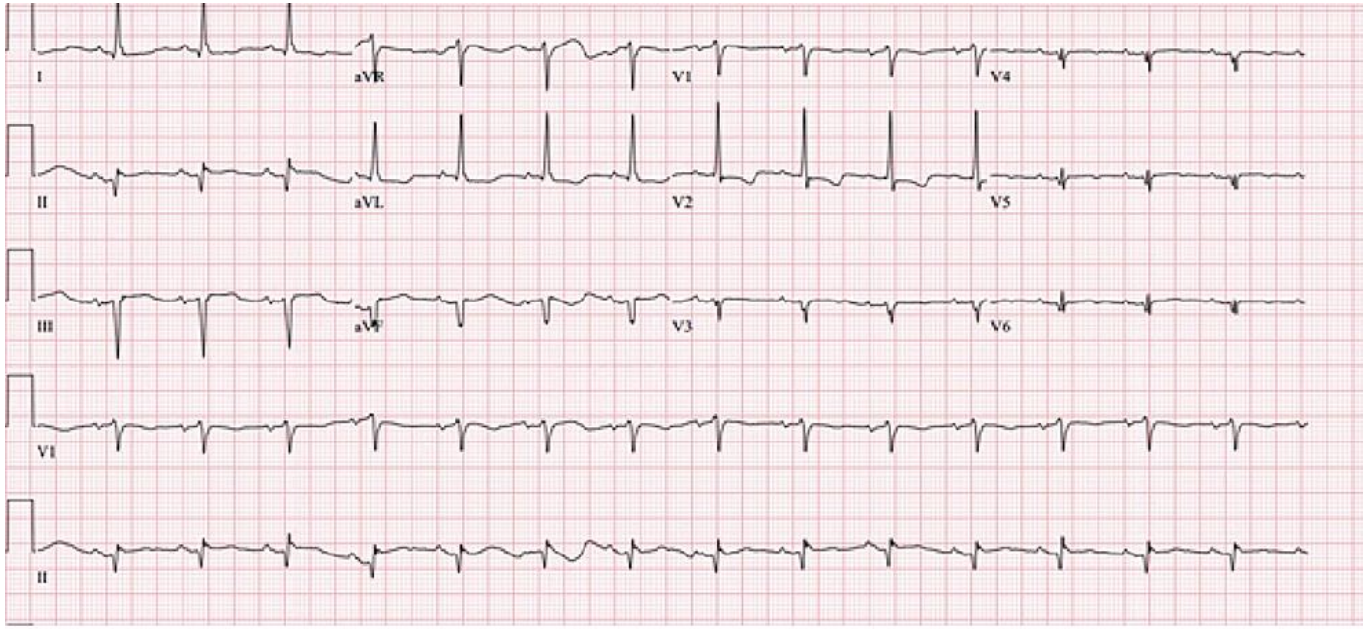


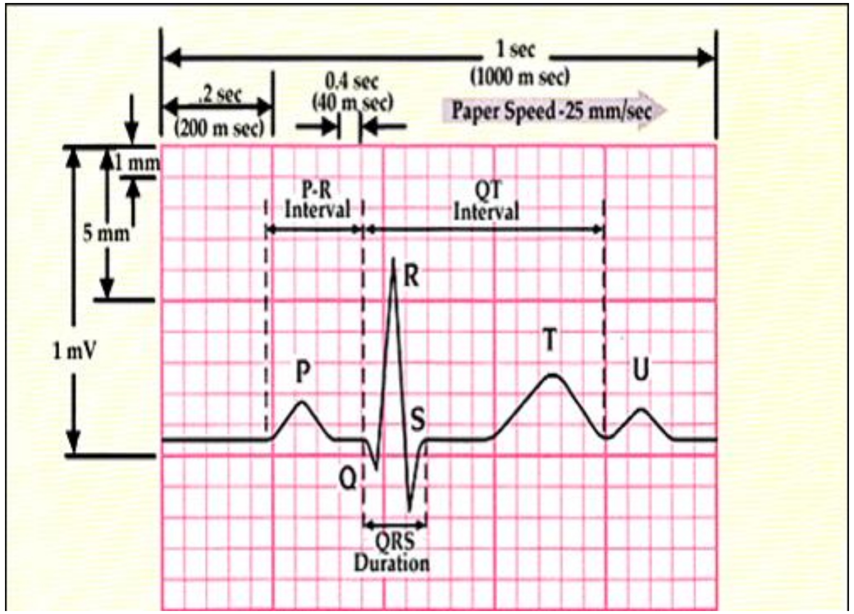
You can use a moving window approach

Define outliers based on a neighborhood, not on the entire signal

If you consider a moving window of 5 samples,
Then the point is not an outlier!

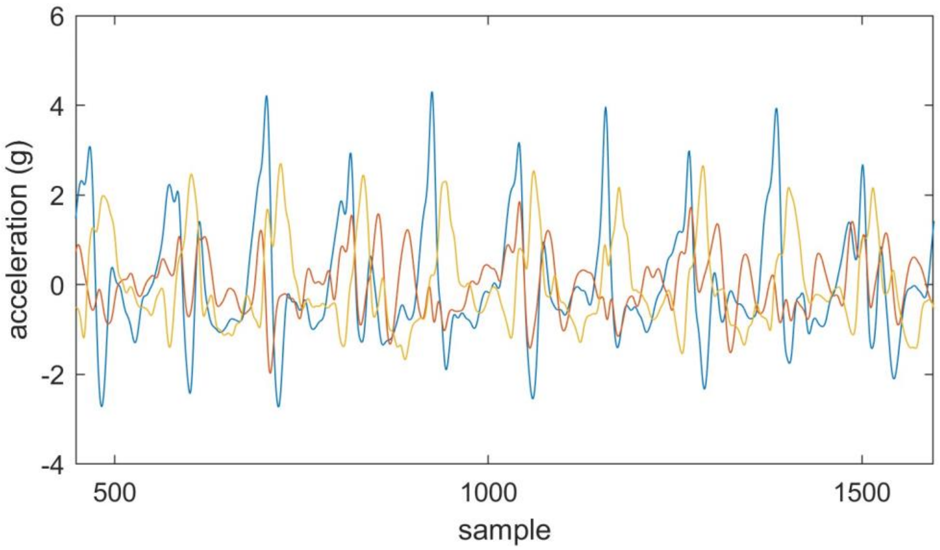
Outliers: instruments





If the recorded ECG data is usually in the range of few mV, values of 10-100 mV are outliers!

Outliers: instruments



If the recorded human acceleration data is usually in the range of 2g, values of 8 g are outliers!

Outliers: summary

- Ensure your sample is representative of the population
- Contextualize to the measured variable
- Consider instrumentation range
- Consider domain knowledge
- Always look at the signals!

- Work with large and heterogeneous datasets
- Consider what you are measuring
- If the value is outside the instrument range, it is an outlier
- A certain value can be normal in a population and abnormal in another.
- Visualize data for a comprehensive understanding

Outliers: implementation

- Matlab
  - rmoutliers(data, “method”); method: “mean”, “median”, “quartiles”
  - rmoutliers(data, “movmethod”, window); movmethod: “movmean”, “movmedian”

- Python
  - You should do it manually (compute mean and std, then distance of points from the mean) or use some libraries.

Suggestion: do it manually, considering all aspects discussed previously

Outliers: examples

- Outlier in a variable
- Normal value in a variable but outlier considering other variables
- Normal value in a variable but outlier if dividing by class
- Plot differences before/after imputation
  - With mean
  - With mean +- std
  - With median
  - With median +- std
  - Considering classes

Missing values

Missing values

Column 0	age	years_seniority	income	parking_space	attending_party	entree	pets	emergency_contact
Tony	48	27		1	5	shrimp		Pepper
Donald	67	25	86	10	2	beef		Jane
Henry	69	21	95	6	1	chicken	62	Janet
Janet	62	21	110	3	1	beef		Henry
Nick		17		4				
Bruce	37	14	63		1	veggie		NA
Steve	83		77	7	1	chicken		n/a
Clint	27	9	118	9		shrimp	3	None
Wanda	19	7	52	2	2	shrimp		empty
Natasha	26	4	162	5	3			_
Carol		3	127	11	1	veggie	1	""""
Mandy	44	2	68	8	1	chicken		null

Missing information in the dataset:

- Not collected
- Not transcribed
- Errors when saving
- Errors when loading
- Merge multiple datasets

Missing values: why treat them?

- Model Performance: Many machine learning algorithms cannot handle missing values directly and may produce errors or suboptimal results if missing values are present in the dataset.

- Data Integrity: Missing values can distort the statistical properties of the dataset, such as the mean, variance, and covariance.

- Interpretability: Missing values can complicate the interpretation of model results and make it difficult to draw meaningful insights from the data.

Missing values: scenario 1-variables

Patient ID	Age (years)	Gender	Systolic BP (mmHg)	Diastolic BP (mmHg)	Heart Rate (bpm)	Glucose Level (mg/dL)	Cholesterol Level (mg/dL)
1	45	Male	120	80	72	90	180
2	35	Female	110	70	65	95	NaN
3	50	Male	130	80	80	105	220
4	28	Female	115	75	68	88	NaN
5	50	Male	130	90	85	120	250
6	32	Female	118	78	70	98	NaN
7	55	Male	135	95	85	95	NaN
8	40	Female	112	72	60	92	195
9	48	Male	125	85	75	102	NaN
10	38	Female	120	80	70	100	200
11	67	Male	175	115	100	210	NaN
12	30	Female	105	68	62	85	180
13	75	Male	180	120	95	190	NaN
14	42	Female	122	78	72	94	205
15	55	Male	130	85	82	160	NaN
16	36	Female	118	75	68	100	NaN
17	58	Male	120	80	85	98	225
18	45	Female	110	70	65	93	198
19	50	Male	120	75	80	110	240
20	40	Female	112	72	60	95	200

If you have a well-organized dataset, and several entries for some variables are not available:
You may think to discard that variable(s)

You lose some information, but there are not alternatives.

It depends on the percentage of missing values on that specific variable.

Let’s say, 10% missing values can be solved, 50% can not.

Missing values: scenario 2-subjects

Patient ID	Age (years)	Gender	Systolic BP (mmHg)	Diastolic BP (mmHg)	Heart Rate (bpm)	Glucose Level (mg/dL)	Cholesterol Level (mg/dL)
1	45	Male	120	80	72	90	180
2	35	Female	110	70	65	95	200
3	50	Male	130	80	80	105	220
4	28	Female	115	75	68	88	190
5	50	Male	130	90	85	120	250
6	32	Female	118	78	70	98	210
7	55	Male	135	95	85	95	230
8	40	Female	NaN	NaN	60	NaN	195
9	48	Male	125	85	75	102	215
10	38	Female	120	80	70	100	200
11	67	Male	175	115	100	210	280
12	30	Female	105	68	62	85	180
13	75	Male	180	120	95	190	260
14	42	Female	122	78	72	94	205
15	55	Male	130	85	82	160	200
16	36	Female	118	75	68	100	190
17	58	Male	120	80	85	98	225
18	45	Female	110	70	65	93	198
19	50	Male	120	75	80	110	240
20	40	Female	112	72	60	95	200

If you have a well-organized dataset, and several entries for some subjects are not available:
You may think to discard that subject(s)

You lose some information, but there are not alternatives.

It depends on the percentage of missing values on that specific subject.

Let’s say, 10% missing values can be solved, 50% can not.

Missing values: scenario 3-fuck

Patient ID	Age (years)	Gender	Systolic BP (mmHg)	Diastolic BP (mmHg)	Heart Rate (bpm)	Glucose Level (mg/dL)	Cholesterol Level (mg/dL)
1	45	Male	120	80	72	90	180
2	35	Female	110	70	65	NaN	200
3	50	Male	130	80	80	105	220
4	28	Female	115	75	68	88	190
5	50	Male	130	90	NaN	120	250
6	32	Female	118	78	70	98	210
7	55	Male	135	95	85	95	230
8	40	Female	NaN	NaN	60	NaN	195
9	48	Male	125	85	75	102	215
10	38	Female	120	80	NaN	100	200
11	67	Male	175	115	100	210	280
12	30	Female	105	68	62	85	180
13	75	Male	180	120	95	NaN	260
14	42	Female	122	78	72	94	205
15	55	Male	NaN	NaN	82	160	200
16	36	Female	118	75	68	100	NaN
17	58	Male	120	80	85	98	225
18	45	Female	110	70	65	NaN	198
19	50	Male	120	75	80	110	240
20	40	Female	112	72	60	95	200

You can not remove subjects or variables only because there are missing values.

Otherwise, you will lose most of your dataset

↓

We should find a solution!

Missing values: imputation

Patient ID	Age (years)	Gender	Systolic BP (mmHg)	Diastolic BP (mmHg)	Heart Rate (bpm)	Glucose Level (mg/dL)	Cholesterol Level (mg/dL)
1	45	Male	120	80	72	90	180
2	35	Female	110	70	65	NaN	200
3	50	Male	130	80	80	105	220
4	28	Female	115	75	68	88	190
5	50	Male	130	90	NaN	120	250
6	32	Female	118	78	70	98	210
7	55	Male	135	95	85	95	230
8	40	Female	NaN	NaN	60	NaN	195
9	48	Male	125	85	75	102	215
10	38	Female	120	80	NaN	100	200
11	67	Male	175	115	100	210	280
12	30	Female	105	68	62	85	180
13	75	Male	180	120	95	NaN	260
14	42	Female	122	78	72	94	205
15	55	Male	NaN	NaN	82	160	200
16	36	Female	118	75	68	100	NaN
17	58	Male	120	80	85	98	225
18	45	Female	110	70	65	NaN	198
19	50	Male	120	75	80	110	240
20	40	Female	112	72	60	95	200

Several methods are available:

- **Statistical**
- Classification
- **Distance-based** (Clustering)

Missing values: statistical imputation

Patient ID	Age (years)	Gender	Systolic BP (mmHg)	Diastolic BP (mmHg)	Heart Rate (bpm)	Glucose Level (mg/dL)	Cholesterol Level (mg/dL)
1	45	Male	120	80	72	**90**	180
2	35	Female	110	70	65	**95**	200
3	50	Male	130	80	80	**105**	220
4	28	Female	115	75	68	**88**	190
5	50	Male	130	90	85	**120**	250
6	32	Female	118	78	70	**98**	210
7	55	Male	135	95	85	**95**	230
8	40	Female	120	80	60	**NaN**	195
9	48	Male	125	85	75	**102**	215
10	38	Female	120	80	70	**100**	200
11	67	Male	175	115	100	**210**	280
12	30	Female	105	68	62	**85**	180
13	75	Male	180	120	95	**190**	260
14	42	Female	122	78	72	**NaN**	205
15	55	Male	130	85	82	**160**	200
16	36	Female	118	75	68	**100**	190
17	58	Male	120	80	85	**98**	225
18	45	Female	110	70	65	**93**	198
19	50	Male	120	75	80	**110**	240
20	40	Female	112	72	60	**95**	200

For glucose level, 2/20 (10%) of values are missing.

You can easily assign to those entries:

- The mean value of that column (if normal distribution)
- The median value of that column (if not normal)
- The mean + **noise**
- The median + **noise**

Noise:

- Random values in the range [mean-std, mean+std]
- Random values in the range [median-std, median+std]
- Random values in the range [Q1, Q3]

Missing values: statistical imputation

Patient ID	Age (years)	Gender	Systolic BP (mmHg)	Diastolic BP (mmHg)	Heart Rate (bpm)	Glucose Level (mg/dL)	Cholesterol Level (mg/dL)	Hypertension
1	45	Male	120	80	72	**90**	180	Yes
2	35	Female	110	70	65	**95**	200	No
3	50	Male	130	80	80	**105**	220	Yes
4	28	Female	115	75	68	**88**	190	No
5	50	Male	130	90	85	**120**	250	Yes
6	32	Female	118	78	70	**98**	210	Yes
7	55	Male	135	95	85	**95**	230	Yes
8	40	Female	110	80	60	**NaN**	195	No
9	48	Male	125	85	75	**102**	215	Yes
10	38	Female	120	80	70	**100**	200	No
11	67	Male	175	115	100	**210**	280	No
12	30	Female	105	68	62	**85**	180	No
13	75	Male	180	120	95	**190**	260	Yes
14	42	Female	142	98	72	**NaN**	205	Yes
15	55	Male	130	85	82	**160**	200	Yes
16	36	Female	118	75	68	**100**	190	No
17	58	Male	120	80	85	**98**	225	No
18	45	Female	110	70	65	**93**	198	Yes
19	50	Male	120	75	80	**110**	240	Yes
20	40	Female	112	72	60	**95**	200	No

If you have multiple classes (e.g., heathy subject and subjects with hypertension)

You should impute missing values based on the specific class.

E.g.

- subject 8 is healthy. Replace the NaN with the mean/median (+ noise) of healthy subjects
- Subject 14 suffer from hypertension. Replace the NaN with the mean/median (+ noise) of subjects with hypertension

Missing values: distance-based imputation

Patient ID	Age (years)	Gender	Systolic BP (mmHg)	Diastolic BP (mmHg)	Heart Rate (bpm)	Glucose Level (mg/dL)	Cholesterol Level (mg/dL)	Hypertension
1	45	Male	120	80	72	**90**	180	Yes
2	35	Female	110	70	65	**95**	200	No
3	50	Male	130	80	80	**105**	220	Yes
4	28	Female	115	75	68	**88**	190	No
5	50	Male	130	90	85	**120**	250	Yes
6	32	Female	118	78	70	**98**	210	Yes
7	55	Male	135	95	85	**95**	230	Yes
8	40	Female	110	80	60	**NaN**	195	No
9	48	Male	125	85	75	**102**	215	Yes
10	38	Female	120	80	70	**100**	200	No
11	67	Male	175	115	100	**210**	280	No
12	30	Female	105	68	62	**85**	180	No
13	75	Male	180	120	95	**190**	260	Yes
14	42	Female	142	98	72	**NaN**	205	Yes
15	55	Male	130	85	82	**160**	200	Yes
16	36	Female	118	75	68	**100**	190	No
17	58	Male	120	80	85	**98**	225	No
18	45	Female	110	70	65	**93**	198	Yes
19	50	Male	120	75	80	**110**	240	Yes
20	40	Female	112	72	60	**95**	200	No

Statistical approaches do not consider the specific subject demographic and clinical information.

Distance-based approaches aim to use data from similar subjects for assigning the missing values to a specific subject.

For a specific subject with one or more missing values, you can select the most **K** similar subjects, and use the mean over this subset of **similar subjects.** This is a reasonable approach.

Missing values: distance-based imputation

Patient ID	Age (years)	Gender	Systolic BP (mmHg)	Diastolic BP (mmHg)	Heart Rate (bpm)	Glucose Level (mg/dL)	Cholesterol Level (mg/dL)	Hypertension
1	45	Male	120	80	72	**90**	180	Yes
2	35	Female	110	70	65	**95**	200	No
3	50	Male	130	80	80	**105**	220	Yes
4	28	Female	115	75	68	**88**	190	No
5	50	Male	130	90	85	**120**	250	Yes
6	32	Female	118	78	70	**98**	210	Yes
7	55	Male	135	95	85	**95**	230	Yes
8	40	Female	110	80	70	**NaN**	195	No
9	48	Male	125	85	75	**102**	215	Yes
10	38	Female	120	80	70	**100**	200	No
11	67	Male	175	115	100	**210**	280	No
12	30	Female	105	68	62	**85**	180	No
13	75	Male	180	120	95	**190**	260	Yes
14	42	Female	142	98	72	**NaN**	205	Yes
15	55	Male	130	85	82	**160**	200	Yes
16	36	Female	118	75	68	**100**	190	No
17	58	Male	120	80	85	**98**	225	No
18	45	Female	110	70	65	**93**	198	Yes
19	50	Male	120	75	80	**110**	240	Yes
20	40	Female	112	72	60	**95**	200	No

Let's consider subject 8. A 40 years old female subject, with no particular health problems.

1. Select only subjects belonging to her class (no hypertension).

Missing values: distance-based imputation

Patient ID	Age (years)	Gender	Systolic BP (mmHg)	Diastolic BP (mmHg)	Heart Rate (bpm)	Glucose Level (mg/dL)	Cholesterol Level (mg/dL)	Hypertension
2	35	Female	110	70	65	**95**	200	No
4	28	Female	115	75	68	**88**	190	No
8	40	Female	110	80	70	**NaN**	195	No
10	38	Female	120	80	70	**100**	200	No
11	67	Male	175	115	100	**210**	280	No
12	30	Female	105	68	62	**85**	180	No
16	36	Female	118	75	68	**100**	190	No
17	58	Male	120	80	85	**98**	225	No
20	40	Female	112	72	60	**95**	200	No

Let's consider subject 8. A 40 years old female subject, with no particular health problems.

1. Select only subjects belonging to her class (no hypertension).
2. Better to consider the same gender, thus selecting only females. Male and females can have different baseline values.

Missing values: distance-based imputation

Patient ID	Age (years)	Gender	Systolic BP (mmHg)	Diastolic BP (mmHg)	Heart Rate (bpm)	Glucose Level (mg/dL)	Cholesterol Level (mg/dL)	Hypertension
2	35	Female	110	70	65	**95**	200	No
4	28	Female	115	75	68	**88**	190	No
8	40	Female	110	80	70	**NaN**	195	No
10	38	Female	120	80	70	**100**	200	No
12	30	Female	105	68	62	**85**	180	No
16	36	Female	118	75	68	**100**	190	No
20	40	Female	112	72	60	**95**	200	No

Age (years)	Systolic BP (mmHg)	Diastolic BP (mmHg)	Heart Rate (bpm)	Cholesterol Level (mg/dL)
35	110	70	65	200
28	115	75	68	190
40	110	80	70	195
38	120	80	70	200
30	105	68	62	180
36	118	75	68	190
40	112	72	60	200

Let’s consider subject 8. A 40 years old female subject, with no particular health problems.

1. Select only subjects belonging to her class (no hypertension).
2. Better to consider the same gender, thus selecting only females. Male and females can have different baseline values.
3. Select the features/characteristics (demographic/clinical information) from which you want to define the similarity. Gender can now be discarded, they are all females. Glucose level is not available for subject 8. All the classes are the same (no hypertension).

Missing values: distance-based imputation

Age (years)	Systolic BP (mmHg)	Diastolic BP (mmHg)	Heart Rate (bpm)	Cholesterol Level (mg/dL)
35	110	70	65	200
28	115	75	68	190
40	110	80	70	195
38	120	80	70	200
30	105	68	62	180
36	118	75	68	190
40	112	72	60	200

$$f' = \frac{f - f^{min}}{f^{max} - f^{min}}$$

Age (years)	Systolic BP (mmHg)	Diastolic BP (mmHg)	Heart Rate (bpm)	Cholesterol Level (mg/dL)
0.58	0.33	0.17	0.50	1.00
0.00	0.67	0.58	0.80	0.50
1.00	0.33	1.00	1.00	0.75
0.83	1.00	1.00	1.00	1.00
0.17	0.00	0.00	0.20	0.00
0.67	0.87	0.58	0.80	0.50
1.00	0.47	0.33	0.00	1.00

Let's consider subject 8. A 40 years old female subject, with no particular health problems.

1. Select only subjects belonging to her class (no hypertension).
2. Better to consider the same gender, thus selecting only females. Male and females can have different baseline values.
3. Select the features/characteristics (demographic/clinical information) from which you want to define the similarity (Gender can now be discarded, they are all females!).
4. Normalize the dataset (and weight each feature if you prefer).

Missing values: distance-based imputation

Age (years)	Systolic BP (mmHg)	Diastolic BP (mmHg)	Heart Rate (bpm)	Cholesterol Level (mg/dL)
0.58	0.33	0.17	0.50	1.00
0.00	0.67	0.58	0.80	0.50
1.00	0.33	1.00	1.00	0.75
0.83	1.00	1.00	1.00	1.00
0.17	0.00	0.00	0.20	0.00
0.67	0.87	0.58	0.80	0.50
1.00	0.47	0.33	0.00	1.00

$$d = \sqrt{\sum_{n=1}^{N} (f_s - f_n)^2}$$

Age (years)	Systolic BP (mmHg)	Diastolic BP (mmHg)	Heart Rate (bpm)	Cholesterol Level (mg/dL)	Distance
0.58	0.33	0.17	0.50	1.00	1.09
0.00	0.67	0.58	0.80	0.50	1.18
0.83	1.00	1.00	1.00	1.00	0.74
0.17	0.00	0.00	0.20	0.00	1.73
0.67	0.87	0.58	0.80	0.50	0.82
1.00	0.47	0.33	0.00	1.00	1.24

Let’s consider subject 8. A 40 years old female subject, with no particular health problems.

1. Select only subjects belonging to her class (no hypertension).
2. Better to consider the same gender, thus selecting only females. Male and females can have different baseline values.
3. Select the features/characteristics (demographic/clinical information) from which you want to define the similarity (Gender can now be discarded, they are all females!).
4. Normalize the dataset (and weight each feature if you prefer).
5. Sort subjects based on the overall distance from the subject 8 (n=number of features, s=subject 8).

Missing values: distance-based imputation

Age (years)	Systolic BP (mmHg)	Diastolic BP (mmHg)	Heart Rate (bpm)	Cholesterol Level (mg/dL)
0.58	0.33	0.17	0.50	1.00
0.00	0.67	0.58	0.80	0.50
1.00	0.33	1.00	1.00	0.75
0.83	1.00	1.00	1.00	1.00
0.17	0.00	0.00	0.20	0.00
0.67	0.87	0.58	0.80	0.50
1.00	0.47	0.33	0.00	1.00

$$d = \sqrt{\sum_{n=1}^{N} (f_s - f_n)^2}$$

ID	Age (years)	Systolic BP (mmHg)	Diastolic BP (mmHg)	Heart Rate (bpm)	Cholesterol Level (mg/dL)	Distance
2	0.58	0.33	0.17	0.50	1.00	1.09
4	0.00	0.67	0.58	0.80	0.50	1.18
10	0.83	1.00	1.00	1.00	1.00	0.74
12	0.17	0.00	0.00	0.20	0.00	0.73
16	0.67	0.87	0.58	0.80	0.50	0.82
20	1.00	0.47	0.33	0.00	1.00	1.24

Let’s consider subject 8. A 40 years old female subject, with no particular health problems.

1. Select only subjects belonging to her class (no hypertension).
2. Better to consider the same gender, thus selecting only females. Male and females can have different baseline values.
3. Select the features/characteristics (demographic/clinical information) from which you want to define the similarity (Gender can now be discarded, they are all females!).
4. Normalize the dataset (and weight each feature if you prefer).
5. Sort subjects based on the overall distance from the subject 8 (n=number of features, s=subject 8).
6. Select K (let’s say K=3)

Missing values: distance-based imputation

Patient ID	Age (years)	Gender	Systolic BP (mmHg)	Diastolic BP (mmHg)	Heart Rate (bpm)	Glucose Level (mg/dL)	Cholesterol Level (mg/dL)	Hypertension
2	35	Female	110	70	65	**95**	200	No
4	28	Female	115	75	68	**88**	190	No
8	40	Female	110	80	70	**NaN**	195	No
10	38	Female	120	80	70	**100**	200	No
12	30	Female	105	68	62	**85**	180	No
16	36	Female	118	75	68	**100**	190	No
20	40	Female	112	72	60	**95**	200	No

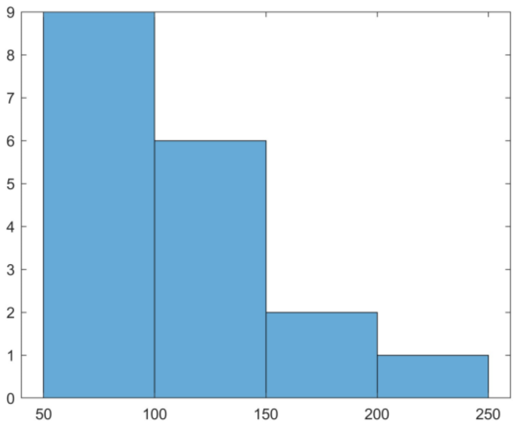
$$\text{mean} = \frac{\sum_{i=1}^{K} x_i}{K} = \frac{100+85+100}{3} = 95$$

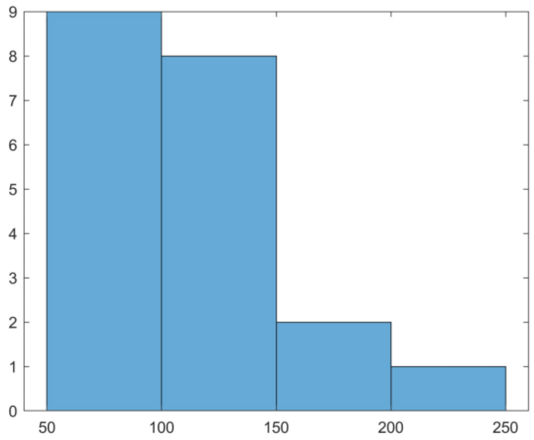
Let's consider subject 8. A 40 years old female subject, with no particular health problems.

1. Select only subjects belonging to her class (no hypertension).
2. Better to consider the same gender, thus selecting only females. Male and females can have different baseline values.
3. Select the features/characteristics (demographic/clinical information) from which you want to define the similarity (Gender can now be discarded, they are all females!).
4. Normalize the dataset (and weight each feature if you prefer).
5. Sort subjects based on the overall distance from the subject 8 (n=number of features, s=subject 8).
6. Select K (let's say K=3)
7. Average the values from the K similar subjects

Missing values: check

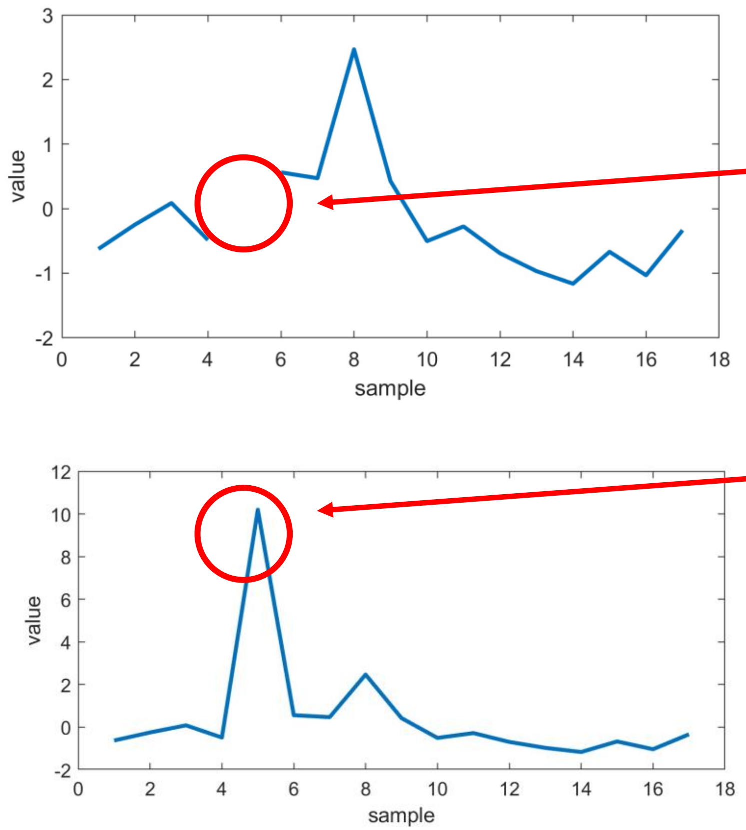
Patient ID	Age (years)	Gender	Systolic BP (mmHg)	Diastolic BP (mmHg)	Heart Rate (bpm)	Glucose Level (mg/dL)	Cholesterol Level (mg/dL)	Hypertension
1	45	Male	120	80	72	**90**	180	Yes
2	35	Female	110	70	65	**95**	200	No
3	50	Male	130	80	80	**105**	220	Yes
4	28	Female	115	75	68	**88**	190	No
5	50	Male	130	90	85	**120**	250	Yes
6	32	Female	118	78	70	**98**	210	Yes
7	55	Male	135	95	85	**95**	230	Yes
8	40	Female	110	80	60	**95**	195	No
9	48	Male	125	85	75	**102**	215	Yes
10	38	Female	120	80	70	**100**	200	No
11	67	Male	175	115	100	**210**	280	No
12	30	Female	105	68	62	**85**	180	No
13	75	Male	180	120	95	**190**	260	Yes
14	42	Female	142	98	72	**93**	205	Yes
15	55	Male	130	85	82	**160**	200	Yes
16	36	Female	118	75	68	**100**	190	No
17	58	Male	120	80	85	**98**	225	No
18	45	Female	110	70	65	**93**	198	Yes
19	50	Male	120	75	80	**110**	240	Yes
20	40	Female	112	72	60	**95**	200	No





Verify that you have not significantly altered the distribution!

Missing values and outliers imputation: time series



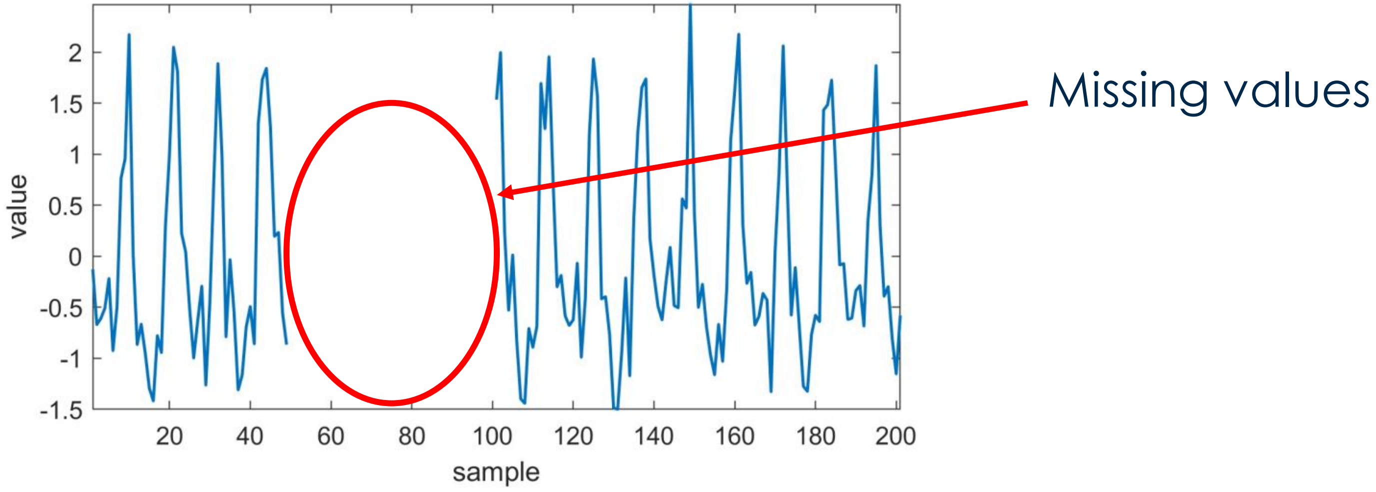
Missing value

Outlier

If they are isolated (single data-points), interpolation (e.g., linear) solves the problem.

You can assign to that point the average value of the preceding and following.

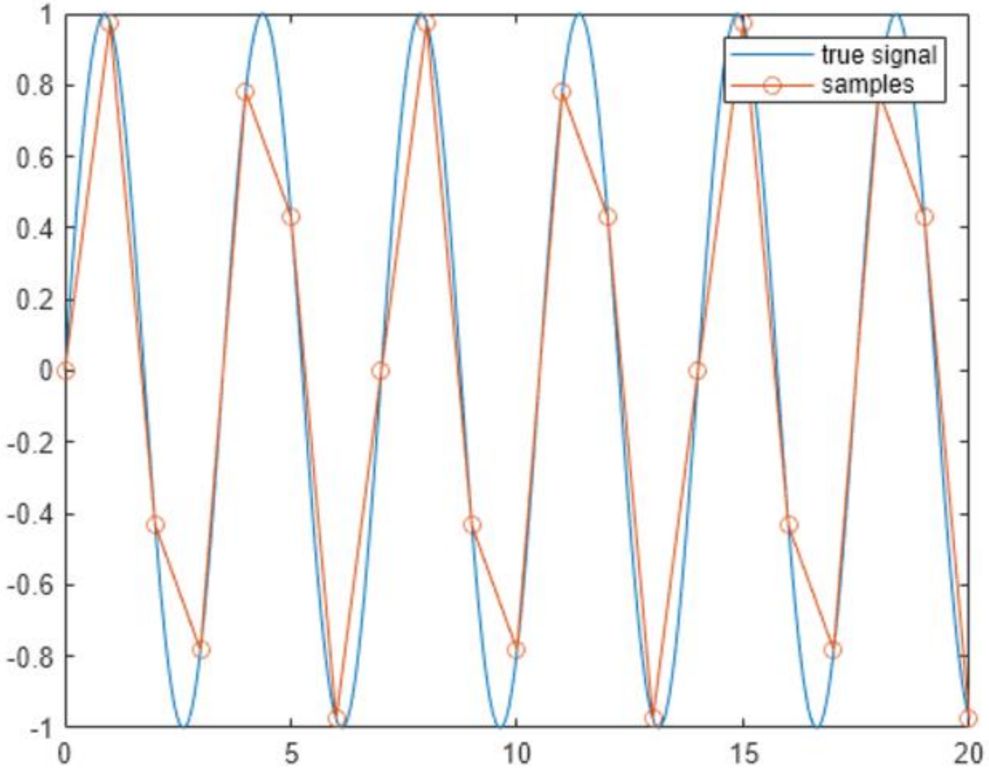
Missing values and outliers imputation: time series

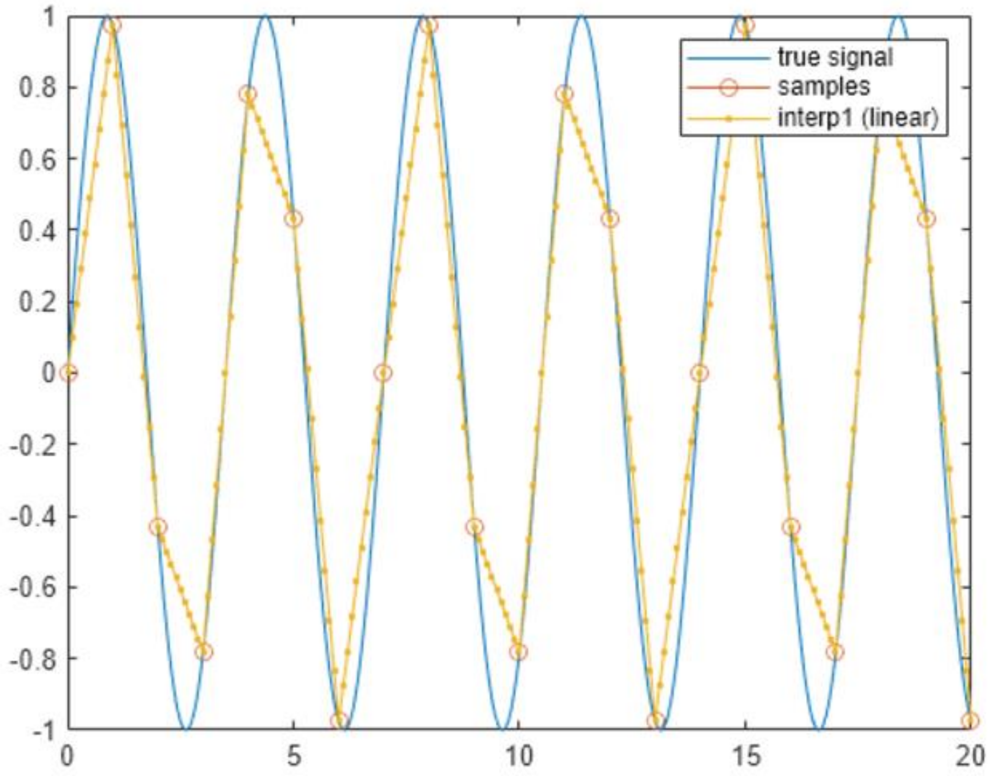


If a significant number of data-points is missing, then it is better not to consider that portion (or the entire time-series).

Missing values and outliers imputation: interpolation

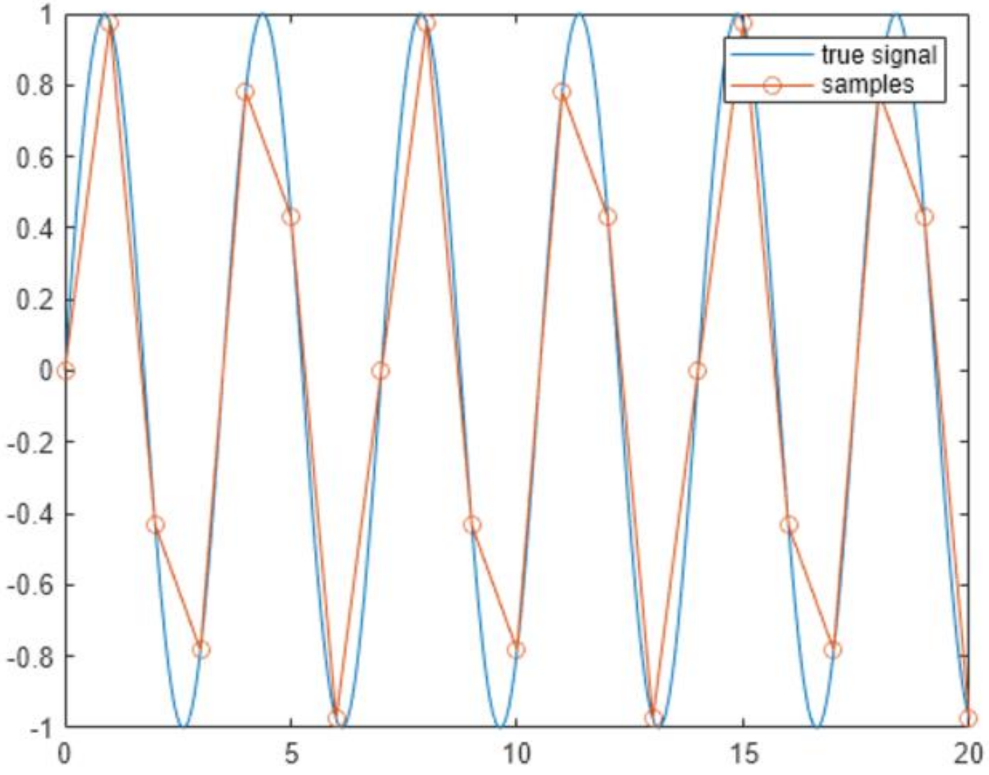
Linear interpolation is by far the most common method of inferring values between sampled points. However, not always is the best choice.

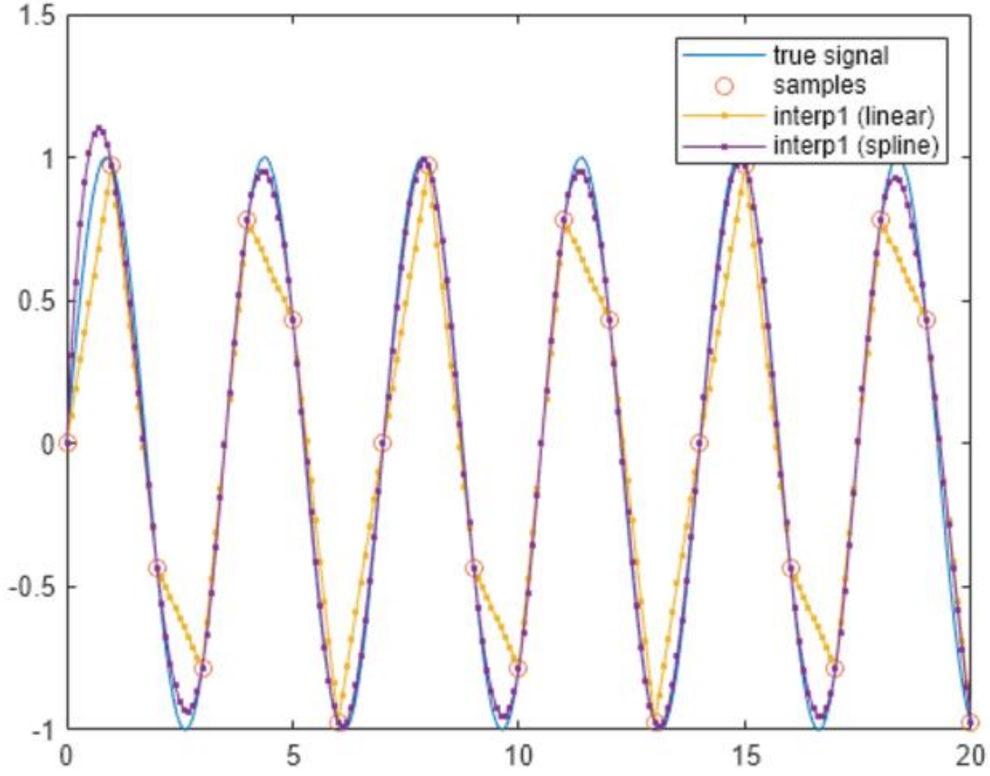




Missing values and outliers imputation: interpolation

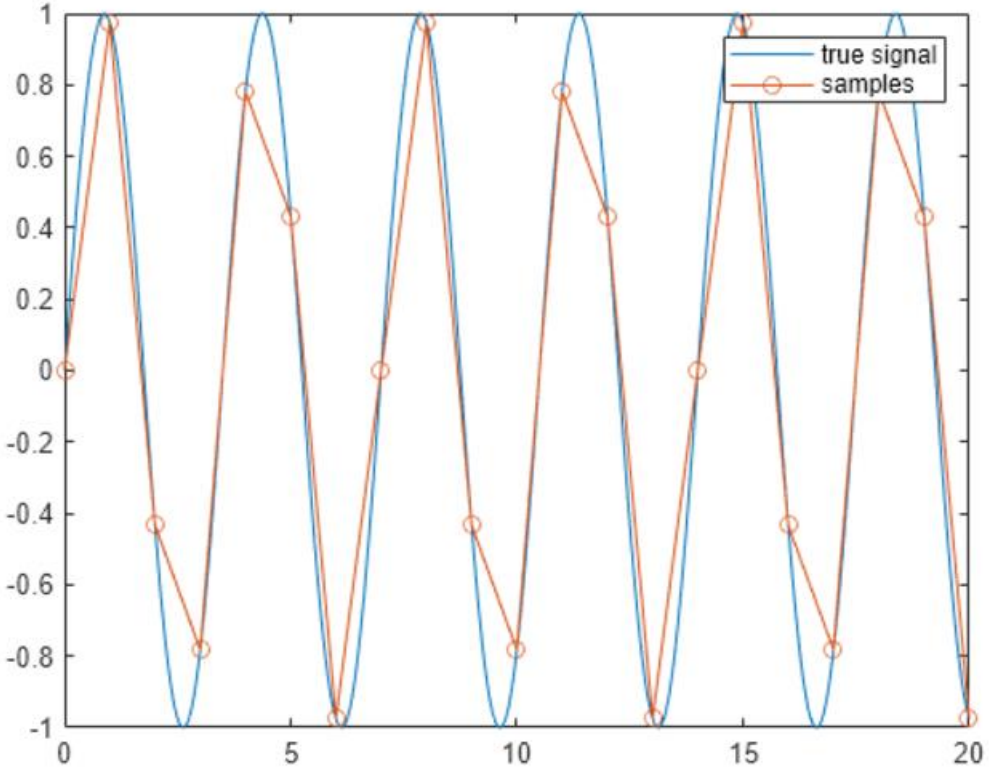
Spline interpolation is often preferred over linear interpolation because the interpolation error can be much smaller.

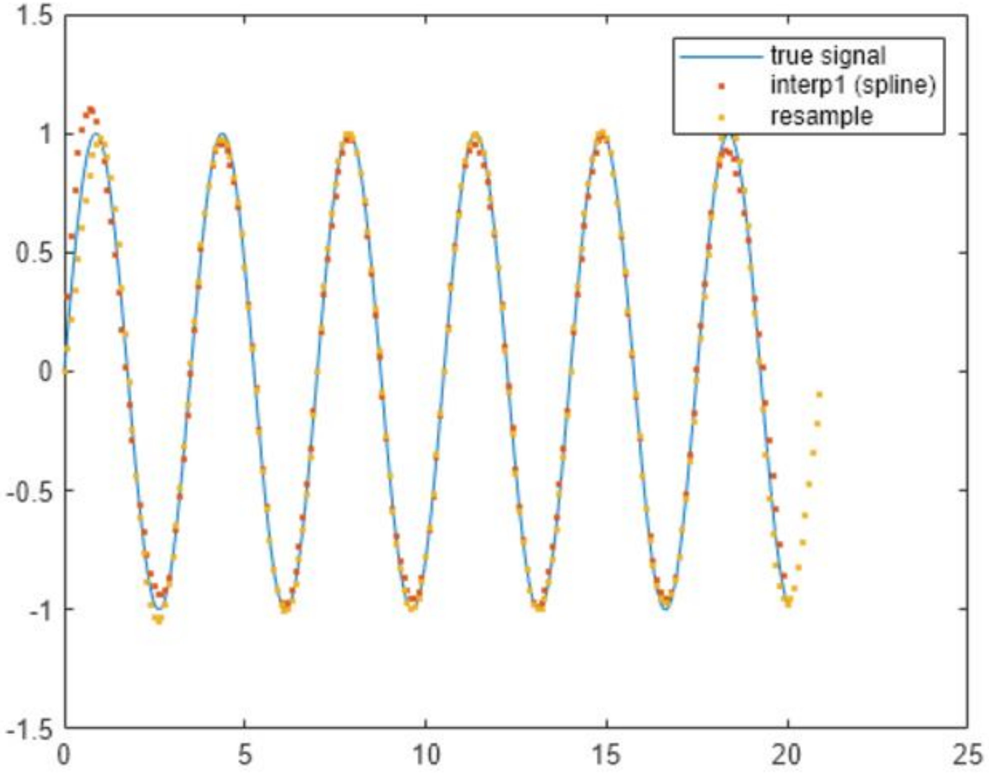




Missing values and outliers imputation: interpolation

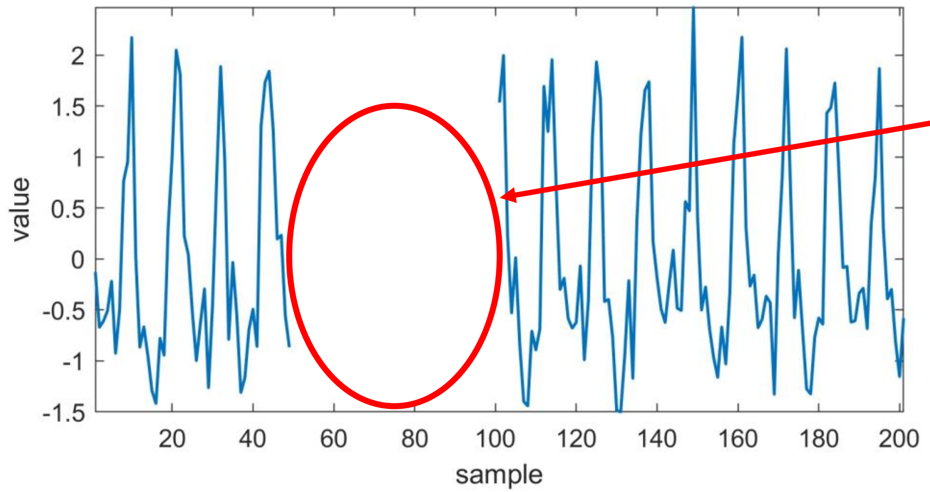
Resampling can be a valid alternative.

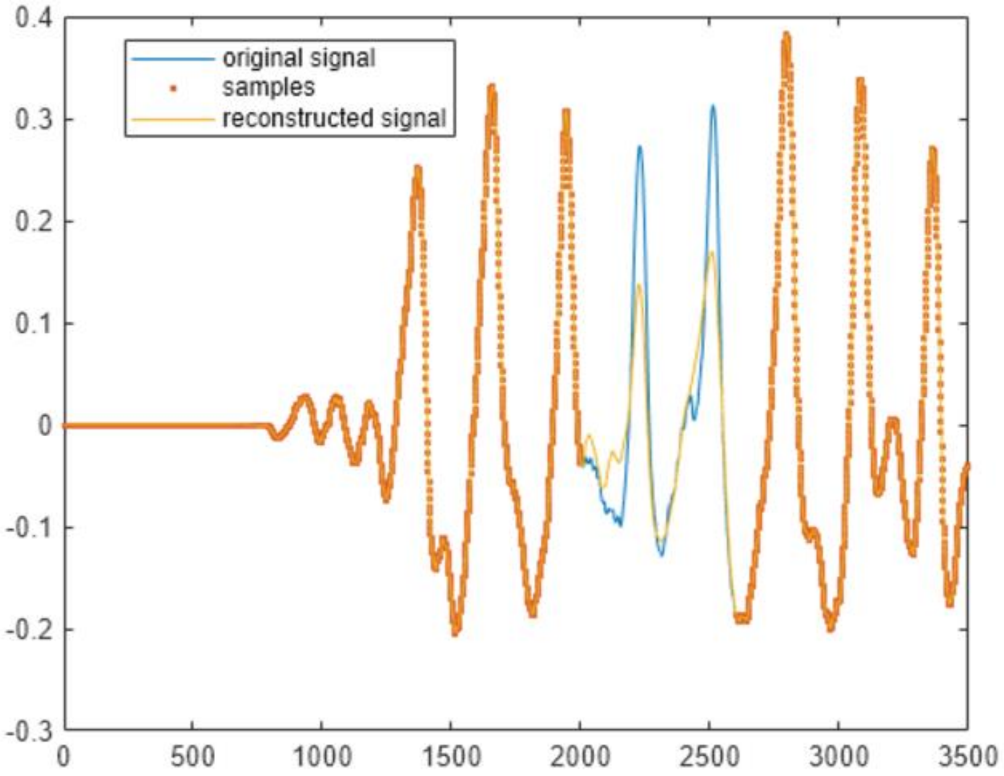




Missing values and outliers imputation: interpolation

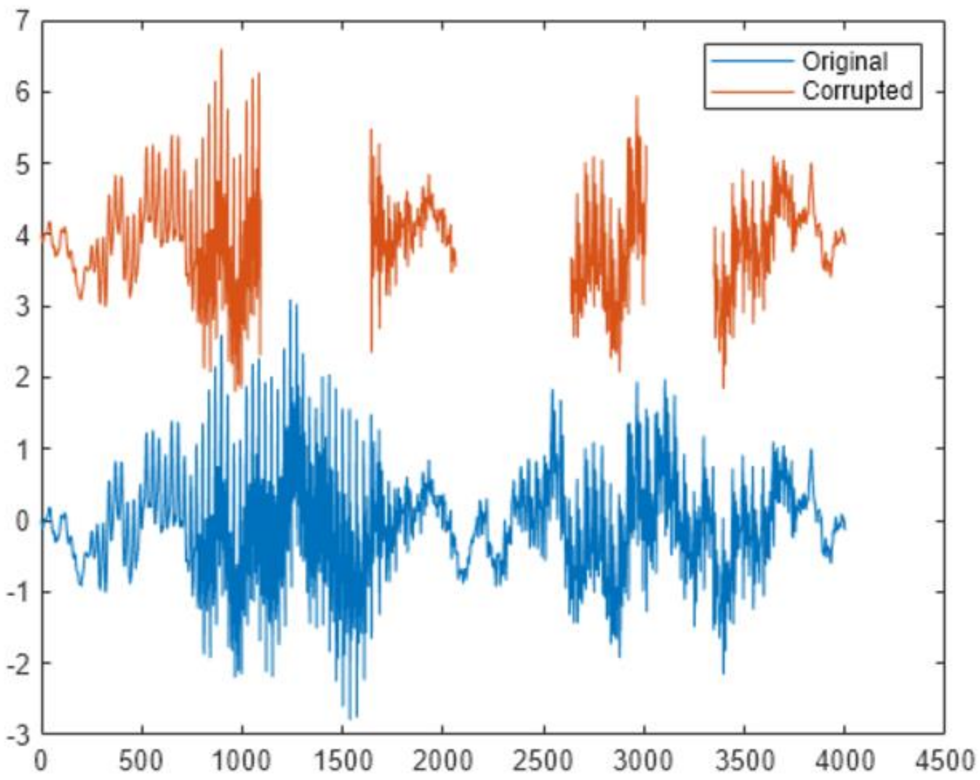
Some methods (fill gaps) can even reconstruct a portion of signal that was completely missing.

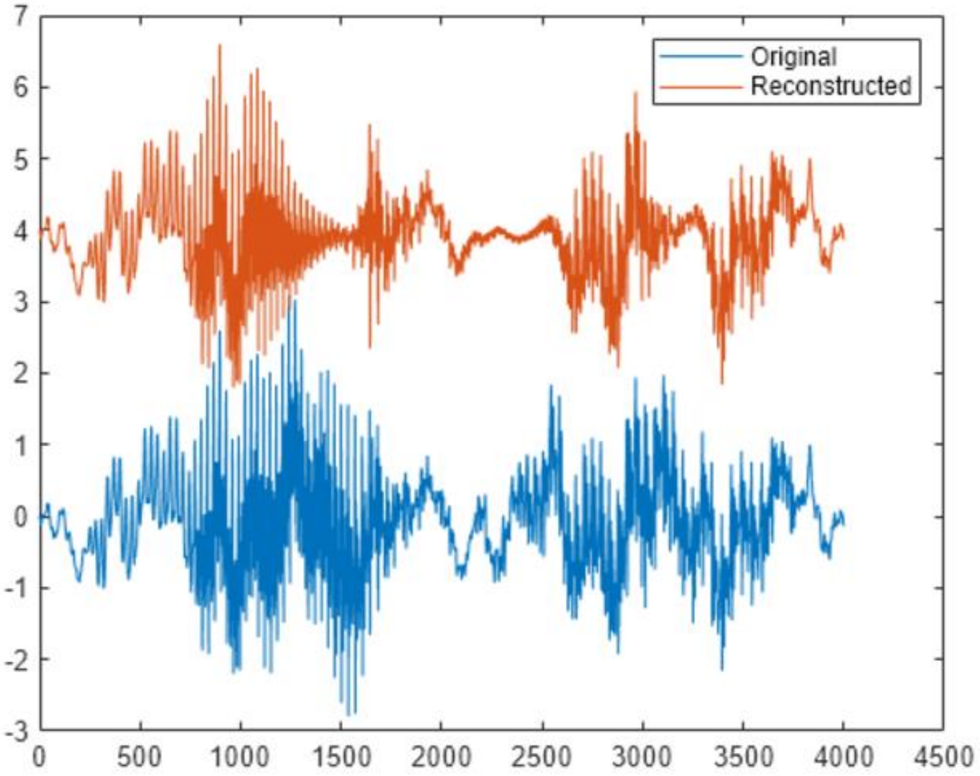




Missing values and outliers imputation: interpolation

Some methods (fill gaps) can even reconstruct a portion of signal that was completely missing.





Missing values and outliers detection

Colab notebooks:

Outliers detection:
https://colab.research.google.com/drive/12mUM5Xbn5pkDuOjNz5y437IZ3UF_CcWf?usp=sharing

Missing values detection:
https://colab.research.google.com/drive/1PSORtHr7SoVHxk07ybW4e9wI3cj_ATCJ?usp=sharing

Missing values and outliers imputation

Colab notebook

Load the 5a.iris_mvs.csv file

Colab notebook:
https://colab.research.google.com/drive/1PSORtHr7SoVHxk07ybW4e9wI3cj_ATCJ?usp=sharing

MIMIC Dataset

MIMIC-III is a large, freely-available database comprising deidentified health-related data associated with over forty thousand patients who stayed in critical care units.

The database includes information such as demographics, vital sign measurements made at the bedside (~1 data point per hour), laboratory test results, procedures, medications, caregiver notes, imaging reports, and mortality (including post-hospital discharge).

13691 subjects
93 variables
118903 missing values!

MIMIC Dataset

1. Remove variables having a large number of MVs (%MV > 30%). How many?

2. Remove patients having a large number of MVs (%MV > 10%). How many?

3. Recalculate the number of MVs for each variable. How many?

4. For variables with a low number of MVs (5%):
    1. Divide patients into two groups according to their class (last column)
    2. For each group, impute the MVs for a given variable with the mean ± noise (or median + noise)

5. Compare the value distributions before and after imputation using boxplots (keep the class division)

Contacts

Politecnico di Torino
1859

SYSBIO GROUP
SYSTEMS BIOLOGY AND BIOINFORMATICS

ANTHEA LAB
Analytics and technology for health

E-mail: luigi.borzi@polito.it

Website: https://www.sysbio.polito.it/analytics-technologies-health/