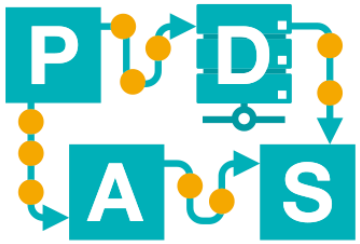
Data Quality & Preprocessing

(Tsung-Hao Huang)

IDS-I-L17



Chair of Process and Data Science

RWTH AACHEN UNIVERSITY

Outliers & Missing Values

Outlier detection techniques based on previous lectures:

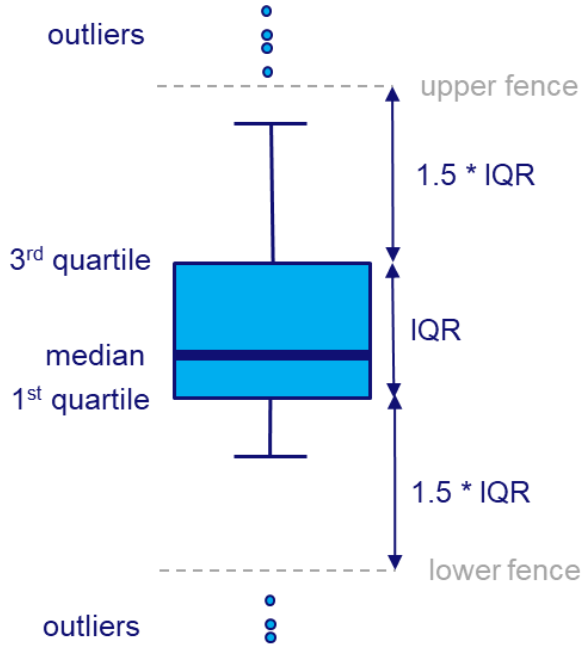
- Boxplots (Lecture 2)
- Decision Trees (Lecture 3)
- Regression (Lecture 4)
- SVMs (Lecture 5)
- Clustering (Lecture 9)

Outliers & Missing Values

We can handle outliers/missing values in several ways, depending on our needs, e.g.:

- **Ignore the feature or instance**
- **Fill in the correct value (domain knowledge)**
- **Fill in a value based on other data**
  - **Mean, median, min, max, most frequent... (possibly focus on similar data)**
  - **Prediction model (regression, SVM, decision tree, NN, …)**

Outliers & Missing Values - Boxplots



Instances outside the fences can be considered outliers.

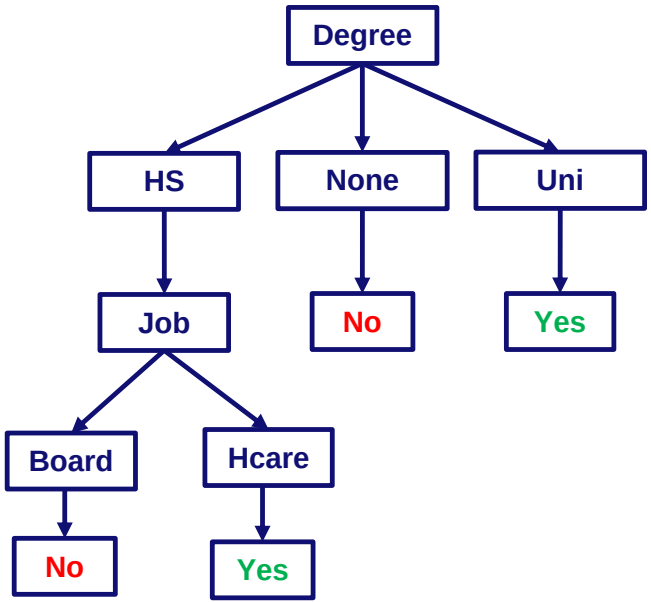
See Instruction 3 for exercises.

Chair of Process and Data Science

Outliers & Missing Values – Decision trees

	Experience	Degree	Job	Class
1	Exp >10	HS	Board	No
2	5< Exp <10	Uni	Board	Yes
3	Exp >10	HS	Board	No
4	5< Exp <10	HS	Hcare	Yes
5	Exp < 5	HS	Hcare	Yes
6	Exp < 5	HS	Board	No
7	Exp < 5	None	Edu	No
8	Exp >10	None	Hcare	No
9	Exp < 5	Uni	Edu	Yes
10	Exp >10	Uni	Board	Yes
11	Exp >10	HS	Board	Yes

Consider the following data about accepting or rejecting job applications based on “Experience”, “Degree”, and “Job”, as well as the corresponding decision tree (extension of Instruction 3).



Outliers & Missing Values – Decision trees

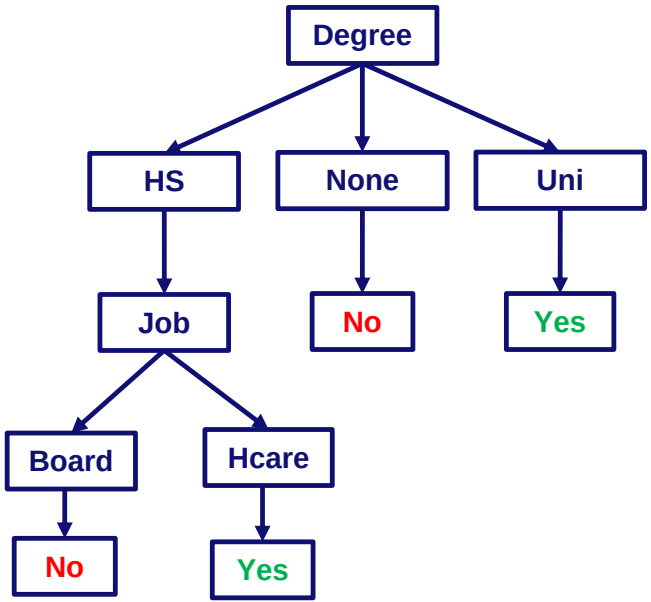
	Experience	Degree	Job	Class
1	Exp >10	HS	Board	No
2	5< Exp <10	Uni	Board	Yes
3	Exp >10	HS	Board	No
4	5< Exp <10	HS	Hcare	Yes
5	Exp < 5	HS	Hcare	Yes
6	Exp < 5	HS	Board	No
7	Exp < 5	None	Edu	No
8	Exp >10	None	Hcare	No
9	Exp < 5	Uni	Edu	Yes
10	Exp >10	Uni	Board	Yes
11	Exp >10	HS	Board	Yes

Exercise 1

Consider the given the data table and corresponding decision tree.

a) Which instances can be identified as outliers based on the given decision tree?

b) Using the decision tree as predictive model to replace the outliers target features values, what would be the resulting data table?



Outliers & Missing Values – Decision trees

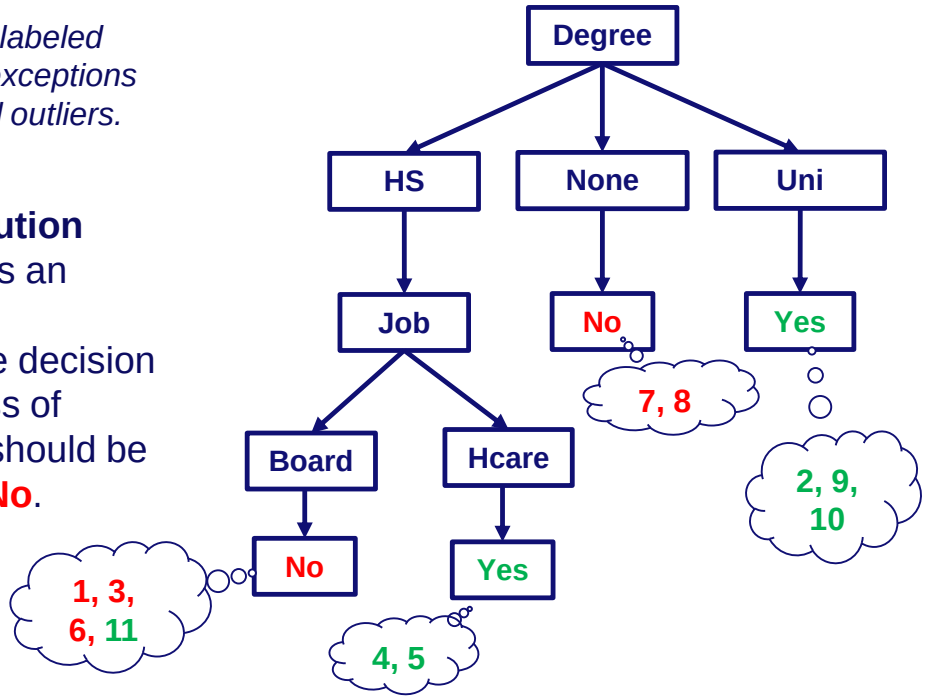
	Experience	Degree	Job	Class
1	Exp >10	HS	Board	No
2	5< Exp <10	Uni	Board	Yes
3	Exp >10	HS	Board	No
4	5< Exp <10	HS	Hcare	Yes
5	Exp < 5	HS	Hcare	Yes
6	Exp < 5	HS	Board	No
7	Exp < 5	None	Edu	No
8	Exp >10	None	Hcare	No
9	Exp < 5	Uni	Edu	Yes
10	Exp >10	Uni	Board	Yes
11	Exp >10	HS	Board	Yes

Leaves should be labeled homogenously – exceptions can be considered outliers.

Exercise 1 - Solution

a) Instance 11 is an outlier!

b) Based on the decision tree, the class of instance 11 should be changed to **No**.



Outliers & Missing Values – Regression

Age [months]	Weight [kg]	Size [cm]
0	2.5	46.4
0	3.8	52.9
1	3.2	50.4
3	6.7	63.4
9	7.5	68
9	10.2	75
12	8.4	71.7
12	11.2	79.1
18	9.6	77.5
18	12.8	86.1
24	10.5	82.3
24	14.1	92

$$\mathbb{M}_{\mathbf{w}}(d) = -8.4 + 1 \times \text{WEIGHT} + 0.12 \times \text{SIZE}$$

Exercise 2

Consider the given baby data and corresponding regression function $\mathbb{M}_{\mathbf{w}}(d)$, predicting *Age* based on *Weight* and *Size*.

a) For an error threshold of 3, (using the absolute error), which data points will be identified as outliers?

b) Use the regression function to replace the outliers target feature values by the predicted values.

Outliers & Missing Values – Regression

Age [months]	Weight [kg]	Size [cm]	Pred	Error
0	2.5	46.4	-0.33	0.33
0	3.8	52.9	1.75	1.75
1	3.2	50.4	0.85	0.15
3	6.7	63.4	5.91	2.91
9	7.5	68	7.26	1.74
9	10.2	75	10.80	1.80
12	8.4	71.7	8.60	3.40
12	11.2	79.1	12.29	0.29
18	9.6	77.5	10.50	7.50
18	12.8	86.1	14.73	3.27
24	10.5	82.3	11.98	12.02
24	14.1	92	16.74	7.26

$$\mathbb{M}_{\mathbf{w}}(d) = -8.4 + 1 \times \text{WEIGHT} + 0.12 \times \text{SIZE}$$

Exercise 2 - Solution

Consider the given baby data and corresponding regression function $\mathbb{M}_{\mathbf{w}}(d)$, predicting *Age* based on *Weight* and *Size*.

a) For an error threshold of 3, (using the absolute error), which data points will be identified as outliers?

b) Use the regression function to replace the outliers target feature values by the predicted values.

All instances with an error above a certain threshold are considered outliers.

Outliers & Missing Values – Regression

Age [months]	Weight [kg]	Size [cm]	Pred	Error
0	2.5	46.4	-0.33	0.33
0	3.8	52.9	1.75	1.75
1	3.2	50.4	0.85	0.15
3	6.7	63.4	5.91	2.91
9	7.5	68	7.26	1.74
9	10.2	75	10.80	1.80
8.60	8.4	71.7	8.60	0
12	11.2	79.1	12.29	0.29
10.50	9.6	77.5	10.50	0
14.73	12.8	86.1	14.73	0
11.98	10.5	82.3	11.98	0
16.74	14.1	92	16.74	0

$$\mathbb{M}_{\mathbf{w}}(d) = -8.4 + 1 \times \text{WEIGHT} + 0.12 \times \text{SIZE}$$

Exercise 2 - Solution

Consider the given baby data and corresponding regression function $\mathbb{M}_{\mathbf{w}}(d)$, predicting *Age* based on *Weight* and *Size*.

a) For an error threshold of 3, (using the absolute error), which data points will be identified as outliers?

b) Use the regression function to replace the outliers target feature values by the predicted values.

All instances with an error above a certain threshold are considered outliers.

Outlier Detection - SVMs

Weight [kg]	Age [months]	Class
2.5	0	underweight
4.2	0	overweight
2.4	0	underweight
3.8	0	overweight
3.2	1	underweight
5.4	1	overweight
3	1	underweight
4.9	1	overweight
4.4	3	underweight
7.4	3	overweight
4.2	3	underweight
6.7	3	overweight
6.2	6	underweight
9.5	6	overweight
5.8	6	underweight
8.7	6	overweight

Exercise 3

Consider the given data table and corresponding separating hyperplane.

a) Based on this hyperplane, which instances can be considered outliers?

b) Replace the outliers target feature values by the values predicted by the SVM.

$$0 = 0.4 \cdot \text{AGE} - \text{WEIGHT} + 4.0$$
$$\Leftrightarrow (w_1, w_2) = (-1.0, 0.4), b = 4.0$$

Outlier Detection - SVMs

Weight [kg]	Age [months]	Class	Classification
2.5	0	underweight	1.5
4.2	0	overweight	-0.2
2.4	0	underweight	1.6
3.8	0	overweight	0.2
3.2	1	underweight	1.2
5.4	1	overweight	-1
3	1	underweight	1.4
4.9	1	overweight	-0.5
4.4	3	underweight	0.8
7.4	3	overweight	-2.2
4.2	3	underweight	1
6.7	3	overweight	-1.5
6.2	6	underweight	0.2
9.5	6	overweight	-3.1
5.8	6	underweight	0.6
8.7	6	overweight	-2.3

Exercise 3 - Solution

Consider the given data table and corresponding separating hyperplane.

a) Based on this hyperplane, which instances can be considered outliers?

b) Replace the outliers target feature values by the values predicted by the SVM.

$$0 = 0.4 \cdot \text{AGE} - \text{WEIGHT} + 4.0$$
$$\Leftrightarrow (w_1, w_2) = (-1.0, 0.4), b = 4.0$$

Instances positioned (far) on the wrong side of the separating hyperplane can be considered outliers.

Here, the outliers are marked in orange.

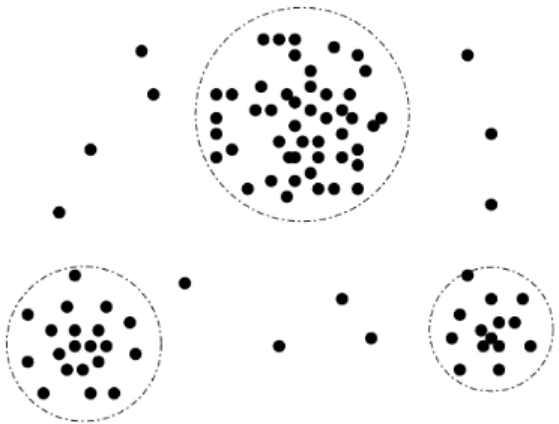
Outlier Detection - Clustering

Weight [kg]	Age [months]	Class
2.5	0	underweight
4.2	0	overweight
2.4	0	underweight
2.7	0	overweight
3.2	1	underweight
5.4	1	overweight
3	1	underweight
4.9	1	overweight
4.4	3	underweight
7.4	3	overweight
4.2	3	underweight
6.7	3	overweight
6.2	6	underweight
9.5	6	overweight
5.8	6	underweight
8.7	6	overweight

Exercise 4

Consider the given data clustered into the blue, red, yellow and purple clusters. Which instances can be considered outliers?

Outlier Detection - Clustering



All data points outside of clusters can be identified as outliers.

Outlier Detection - Clustering

Weight [kg]	Age [months]	Class
2.5	0	underweight
4.2	0	overweight
2.4	0	underweight
2.7	0	overweight
3.2	1	underweight
5.4	1	overweight
3	1	underweight
4.9	1	overweight
4.4	3	underweight
7.4	3	overweight
4.2	3	underweight
6.7	3	overweight
6.2	6	underweight
9.5	6	overweight
5.8	6	underweight
8.7	6	overweight

Exercise 4 - Solution

Consider the given data clustered into the blue, red, yellow and purple clusters. Which instances can be considered outliers?

The grey instance is not part of any cluster and can therefore be considered an outlier.

Transformation

Getting the right data type, e.g.

- **One-hot encoding: categorical to numerical**
- **Binning: numerical to categorical (ordinal)**

Known from earlier lectures!

Transformation - Binning

Introduced in Lecture 2!

Many details can vary in real applications:

- What to do, if there is not enough data to ‘fill the last bin’?
- What to do, if the data is not dividable into bins according to requested parameters?
- Do we define interval limits closed or open ?
- …

Transformation - Binning

Weight [kg]
2.5
3.8
3.2
4.9
4.4
6.7
6.2
7.5
8.4
9.6
12.8
10.5
14.1

Exercise 5

Apply equal-width binning to the given data. Your lowest bin interval should have 0 as open lower limit, and your highest interval should have 16 as closed upper limit.

1) Give a mapping of the data to two bins. Indicate the interval limits as well as the bins boundary values.
2) Give a mapping of the data to four bins. Indicate the interval limits as well as the bins boundary values.

Transformation - Binning

Weight [kg]
2.5
3.2
3.8
4.4
4.9
6.2
6.7
7.5
8.4
9.6
10.5
12.8
14.1

Weight [kg]
2.5
3.8
3.2
4.9
4.4
6.7
6.2
7.5
8.4
9.6
12.8
10.5
14.1

Exercise 5 - Solution

Apply equal-width binning to the given data. Your lowest bin interval should have 0 as open lower limit, and your highest interval should have 16 as closed upper limit.

1) Give a mapping of the data to two bins. Indicate the interval limits as well as the bins boundary values.
2) Give a mapping of the data to four bins. Indicate the interval limits as well as the bins boundary values.

Transformation - Binning

Weight [kg]
2.5
3.2
3.8
4.4
4.9
6.2
6.7
7.5
8.4
9.6
10.5
12.8
14.1

	Bin 1	Bin 2
Interval limits	(0,8]	(8,16]
Data mapping	[2.5, 3.2, 3.8, 4.4, 4.9, 6.2, 6.7, 7.5]	[8.4, 9.6,10.5,12.8, 14.1]
Bin boundary values	2.5, 7.5	8.4, 14.1

Exercise 5 - Solution

Apply equal-width binning to the given data. Your lowest bin interval should have 0 as open lower limit, and your highest interval should have 16 as closed upper limit.

1) **Give a mapping of the data to two bins. Indicate the interval limits as well as the bins boundary values.**
2) Give a mapping of the data to four bins. Indicate the interval limits as well as the bins boundary values.

Transformation - Binning

Weight [kg]
2.5
3.2
3.8
4.4
4.9
6.2
6.7
7.5
8.4
9.6
10.5
12.8
14.1

	Bin 1	Bin 2	Bin 3	Bin 4
Interval limits	Solution next slide			
Data mapping				
Bin boundary values				

Exercise 5 - Solution

Apply equal-width binning to the given data. Your lowest bin interval should have 0 as open lower limit, and your highest interval should have 16 as closed upper limit.

1) Give a mapping of the data to two bins. Indicate the interval limits as well as the bins boundary values.
2) **Give a mapping of the data to four bins. Indicate the interval limits as well as the bins boundary values.**

Transformation - Binning

Exercise 5 – Solution 2)

Weight [kg]
2.5
3.2
3.8
4.4
4.9
6.2
6.7
7.5
8.4
9.6
10.5
12.8
14.1

	Bin 1	Bin 2	Bin 3	Bin 4
Interval limits	(0,4]	(4,8]	(8,12]	(12, 16]
Data mapping	[2.5, 3.2, 3.8]	[4.4, 4.9, 6.2, 6.7, 7.5]	[8.4, 9.6, 10.5]	[12.8, 14.1]
Bin boundary values	2.5, 3.8	4.4, 7.5	8.4, 10.5	12.8, 14.1

Transformation - Binning

Weight [kg]
2.5
3.2
3.8
4.4
4.9
6.2
6.7
7.5
8.4
10.5
12.8
14.1

(sorted)

Exercise 6

Apply equal-frequency binning to the given data.

1) Give a mapping of the data to two bins. Indicate the interval limits as well as the bins boundary values.
2) Give a mapping of the data such that the bin size is four. Indicate the interval limits as well as the bins boundary values.

Transformation - Binning

Weight [kg]
2.5
3.2
3.8
4.4
4.9
6.2
6.7
7.5
8.4
10.5
12.8
14.1

(sorted)

	Bin 1	Bin 2
Interval limits	[2, 6.5)	[6.5, 14.5)
Data mapping	[2.5, 3.2, 3.8, 4.4, 4.9, 6.2]	[6.7, 7.5, 8.4, 10.5, 12.8, 14.1]
Bin boundary values	2.5, 6.2	6.7, 14.1

Exercise 6 - Solution

Apply equal-frequency binning to the given data.

1) **Give a mapping of the data to two bins. Indicate the interval limits as well as the bins boundary values.**
2) Give a mapping of the data such that the bin size is four. Indicate the interval limits as well as the bins boundary values.

Transformation - Binning

Weight [kg]
2.5
3.2
3.8
4.4
4.9
6.2
6.7
7.5
8.4
10.5
12.8
14.1

(sorted)

	Bin 1	Bin 2	Bin 3
Interval limits	Solution next slide		
Data mapping			
Bin boundary values			

Exercise 6 - Solution

Apply equal-frequency binning to the given data.

1) Give a mapping of the data to two bins. Indicate the interval limits as well as the bins boundary values.
2) **Give a mapping of the data such that the bin size is four. Indicate the interval limits as well as the bins boundary values.**

Transformation - Binning

Exercise 6 – Solution 2)

Weight [kg]
2.5
3.2
3.8
4.4
4.9
6.2
6.7
7.5
8.4
10.5
12.8
14.1

(sorted)

	Bin 1	Bin 2	Bin 3
Interval limits	(2, 4.5]	(4.5, 7.5]	(7.5, 15]
Data mapping	[2.5, 3.2, 3.8, 4.4]	[4.9, 6.2, 6.7, 7.5]	[8.4, 9.6, 10.5, 14.1]
Bin boundary values	2.5, 4.4	4.9, 7.5	8.4, 14.1

Normalization

Adjusting the influence of features.

Introduced in the lecture:

- **Min-Max Normalization**
- **Standard Score Normalization**
- **Decimal Scaling**

Normalization – Min-Max

Age [months]	Weight [kg]	Size [cm]
0	2.5	46.4
0	3.8	52.9
1	3.2	50.4
1	4.9	56.9
3	4.4	56.7
3	6.7	63.4
6	6.2	63.4
6	8.7	70.2
9	7.5	68
9	10.2	75
12	8.4	71.7
12	11.2	79.1
18	9.6	77.5
18	12.8	86.1
24	10.5	82.3
24	14.1	92

Exercise 7

Apply Min-Max Normalization for each feature of the given data with a target interval of [1,10].

Normalization – Min-Max

Age [months]	Weight [kg]	Size [cm]
0	2.5	46.4
0	3.8	52.9
1	3.2	50.4
1	4.9	56.9
3	4.4	56.7
3	6.7	63.4
6	6.2	63.4
6	8.7	70.2
9	7.5	68
9	10.2	75
12	8.4	71.7
12	11.2	79.1
18	9.6	77.5
18	12.8	86.1
24	10.5	82.3
24	14.1	92

Age scaled	Weight scaled	Size scaled
1.00		
1.00		
1.38		
1.38		
2.13		
2.13		
3.25		
3.25		
4.38		
4.38		
5.50		
5.50		
7.75		
7.75		
10.00		
10.00		

Exercise 7 - Solution

Apply Min-Max Normalization for each feature of the given data with a target interval of [1,10].

General Formula:

$$a'_i = \frac{(a_i - a_{\min})}{(a_{\max} - a_{\min})} \cdot (\text{high} - \text{low}) + \text{low}$$

For *Age*:

$$a'_i = \frac{(a_i - 0)}{(24 - 0)} \cdot (10 - 1) + 1$$

$$= \frac{9a_i}{24} + 1$$

Normalization – Min-Max

Age [months]	Weight [kg]	Size [cm]
0	2.5	46.4
0	3.8	52.9
1	3.2	50.4
1	4.9	56.9
3	4.4	56.7
3	6.7	63.4
6	6.2	63.4
6	8.7	70.2
9	7.5	68
9	10.2	75
12	8.4	71.7
12	11.2	79.1
18	9.6	77.5
18	12.8	86.1
24	10.5	82.3
24	14.1	92

Age scaled	Weight scaled	Size scaled
1.0	1.0	
1.0	2.0	
1.4	1.5	
1.4	2.9	
2.1	2.5	
2.1	4.3	
3.3	3.9	
3.3	5.8	
4.4	4.9	
4.4	7.0	
5.5	5.6	
5.5	7.8	
7.8	6.5	
7.8	9.0	
10.0	7.2	
10.0	10.0	

Exercise 7 - Solution

Apply Min-Max Normalization for each feature of the given data with a target interval of [1,10].

General Formula:

$$a'_i = \frac{(a_i - a_{\min})}{(a_{\max} - a_{\min})} \cdot (\text{high} - \text{low}) + \text{low}$$

For *Weight*:

$$a'_i = \frac{(a_i - 2.5)}{(14.1 - 2.5)} \cdot (10 - 1) + 1$$

$$= \frac{9(a_i - 2.5)}{11.16} + 1$$

Normalization – Min-Max

Age [months]	Weight [kg]	Size [cm]
0	2.5	46.4
0	3.8	52.9
1	3.2	50.4
1	4.9	56.9
3	4.4	56.7
3	6.7	63.4
6	6.2	63.4
6	8.7	70.2
9	7.5	68
9	10.2	75
12	8.4	71.7
12	11.2	79.1
18	9.6	77.5
18	12.8	86.1
24	10.5	82.3
24	14.1	92

Age scaled	Weight scaled	Size scaled
1.0	1.0	1.00
1.0	2.0	2.28
1.4	1.5	1.79
1.4	2.9	3.07
2.1	2.5	3.03
2.1	4.3	4.36
3.3	3.9	4.36
3.3	5.8	5.70
4.4	4.9	5.26
4.4	7.0	6.64
5.5	5.6	5.99
5.5	7.8	7.45
7.8	6.5	7.14
7.8	9.0	8.84
10.0	7.2	8.09
10.0	10.0	10.00

Exercise 7 - Solution

Apply Min-Max Normalization for each feature of the given data with a target interval of [1,10].

General Formula:

$$a'_i = \frac{(a_i - a_{\min})}{(a_{\max} - a_{\min})} \cdot (\text{high} - \text{low}) + \text{low}$$

For *Size*:

$$a'_i = \frac{(a_i - 46.4)}{(92 - 46.4)} \cdot (10 - 1) + 1$$

$$= \frac{9(a_i - 46.4)}{45.6} + 1$$

Normalization – Standard Score

Age [months]	Weight [kg]	Size [cm]
0	2.5	46.4
0	3.8	52.9
1	3.2	50.4
1	4.9	56.9
3	4.4	56.7
3	6.7	63.4
6	6.2	63.4
6	8.7	70.2
9	7.5	68
9	10.2	75
12	8.4	71.7
12	11.2	79.1
18	9.6	77.5
18	12.8	86.1
24	10.5	82.3
24	14.1	92

Exercise 8

Scale each feature of the given data points using Standard Score Normalization.

Normalization – Standard Score

Age [months]	Weight [kg]	Size [cm]
0	2.5	46.4
0	3.8	52.9
1	3.2	50.4
1	4.9	56.9
3	4.4	56.7
3	6.7	63.4
6	6.2	63.4
6	8.7	70.2
9	7.5	68
9	10.2	75
12	8.4	71.7
12	11.2	79.1
18	9.6	77.5
18	12.8	86.1
24	10.5	82.3
24	14.1	92

Age scaled	Weight scaled	Size scaled

Exercise 8 - Solution

Scale each feature of the given data points using Standard Score Normalization.

Formulas:

$$a'_i = \frac{a_i - \text{mean}(a)}{\text{sd}(a)}$$

$$\text{sd}(a) = \sqrt{\frac{\Sigma_{i=1}^{n}(a_i - \text{mean}(a))^2}{n-1}}$$

Normalization – Standard Score

Age [months]	Weight [kg]	Size [cm]
0	2.5	46.4
0	3.8	52.9
1	3.2	50.4
1	4.9	56.9
3	4.4	56.7
3	6.7	63.4
6	6.2	63.4
6	8.7	70.2
9	7.5	68
9	10.2	75
12	8.4	71.7
12	11.2	79.1
18	9.6	77.5
18	12.8	86.1
24	10.5	82.3
24	14.1	92

Age scaled	Weight scaled	Size scaled
-1.11		
-1.11		
-0.99		
-0.99		
-0.75		
-0.75		
-0.38		
-0.38		
-0.02		
-0.02		
0.35		
0.35		
1.08		
1.08		
1.81		
1.81		

Exercise 8 - Solution

Scale each feature of the given data points using Standard Score Normalization.

Formulas:

$$a'_i = \frac{a_i - \text{mean}(a)}{\text{sd}(a)}$$

$$\text{sd}(a) = \sqrt{\frac{\sum_{i=1}^{n}(a_i - \text{mean}(a))^2}{n-1}}$$

For *Age*:

$$a'_i = \frac{a_i - 9.13}{8.2}$$

$$\text{sd}(a) = \sqrt{\frac{\sum_{i=1}^{16}(a_i - 9.13)^2}{16-1}} = 8.2$$

$$\text{mean(a)} = 9.13$$

Normalization – Standard Score

Age [months]	Weight [kg]	Size [cm]
0	2.5	46.4
0	3.8	52.9
1	3.2	50.4
1	4.9	56.9
3	4.4	56.7
3	6.7	63.4
6	6.2	63.4
6	8.7	70.2
9	7.5	68
9	10.2	75
12	8.4	71.7
12	11.2	79.1
18	9.6	77.5
18	12.8	86.1
24	10.5	82.3
24	14.1	92

Age scaled	Weight scaled	Size scaled
-1.11	-1.52	
-1.11	-1.14	
-0.99	-1.32	
-0.99	-0.83	
-0.75	-0.97	
-0.75	-0.31	
-0.38	-0.46	
-0.38	0.26	
-0.02	-0.08	
-0.02	0.69	
0.35	0.17	
0.35	0.98	
1.08	0.52	
1.08	1.43	
1.81	0.78	
1.81	1.81	

Exercise 8 - Solution

Scale each feature of the given data points using Standard Score Normalization.

Formulas:

$$a'_i = \frac{a_i - \text{mean}(a)}{\text{sd}(a)}$$

$$\text{sd}(a) = \sqrt{\frac{\Sigma_{i=1}^{n}(a_i - \text{mean}(a))^2}{n-1}}$$

For *Weight*:

$$a'_i = \frac{a_i - 7.79}{3.49}$$

$$\text{sd}(a) = \sqrt{\frac{\Sigma_{i=1}^{16}(a_i - 7.79)^2}{16-1}} = 3.49$$

$$\text{mean(a)} = 7.79$$

Normalization – Standard Score

Age [months]	Weight [kg]	Size [cm]
0	2.5	46.4
0	3.8	52.9
1	3.2	50.4
1	4.9	56.9
3	4.4	56.7
3	6.7	63.4
6	6.2	63.4
6	8.7	70.2
9	7.5	68
9	10.2	75
12	8.4	71.7
12	11.2	79.1
18	9.6	77.5
18	12.8	86.1
24	10.5	82.3
24	14.1	92

Age scaled	Weight scaled	Size scaled
-1.11	-1.52	-1.63
-1.11	-1.14	-1.15
-0.99	-1.32	-1.33
-0.99	-0.83	-0.85
-0.75	-0.97	-0.86
-0.75	-0.31	-0.36
-0.38	-0.46	-0.36
-0.38	0.26	0.15
-0.02	-0.08	-0.02
-0.02	0.69	0.50
0.35	0.17	0.26
0.35	0.98	0.81
1.08	0.52	0.69
1.08	1.43	1.33
1.81	0.78	1.05
1.81	1.81	1.77

Exercise 8 - Solution

Scale each feature of the given data points using Standard Score Normalization.

Formulas:

$$a'_i = \frac{a_i - \text{mean}(a)}{\text{sd}(a)}$$

$$\text{sd}(a) = \sqrt{\frac{\Sigma_{i=1}^{n}(a_i - \text{mean}(a))^2}{n-1}}$$

For *Size*:

$$a'_i = \frac{a_i - 68.25}{13.38}$$

$$\text{sd}(a) = \sqrt{\frac{\Sigma_{i=1}^{16}(a_i - 68.25)^2}{16-1}} = 13.38$$

$$\text{mean(a)} = 68.25$$

Normalization – Decimal Scaling

Age [months]	Weight [kg]	Size [cm]
0	2.5	46.4
0	3.8	52.9
1	3.2	50.4
1	4.9	56.9
3	4.4	56.7
3	6.7	63.4
6	6.2	63.4
6	8.7	70.2
9	7.5	68
9	10.2	75
12	8.4	71.7
12	11.2	79.1
18	9.6	77.5
18	12.8	86.1
24	10.5	82.3
30	17.1	102

Exercise 9

Scale each feature of the given data points using Decimal Scaling.

Normalization – Decimal Scaling

Age [months]	Weight [kg]	Size [cm]
0	2.5	46.4
0	3.8	52.9
1	3.2	50.4
1	4.9	56.9
3	4.4	56.7
3	6.7	63.4
6	6.2	63.4
6	8.7	70.2
9	7.5	68
9	10.2	75
12	8.4	71.7
12	11.2	79.1
18	9.6	77.5
18	12.8	86.1
24	10.5	82.3
30	17.1	102

Age scaled	Weight scaled	Size scaled

Exercise 9 - Solution

Scale each feature of the given data points using Decimal Scaling.

Formula: $a'_i = \frac{a_i}{10^j}$

with j being the minimal integer such that $\frac{a_{\max}}{10^j} < 1$

Normalization – Decimal Scaling

Age [months]	Weight [kg]	Size [cm]
0	2.5	46.4
0	3.8	52.9
1	3.2	50.4
1	4.9	56.9
3	4.4	56.7
3	6.7	63.4
6	6.2	63.4
6	8.7	70.2
9	7.5	68
9	10.2	75
12	8.4	71.7
12	11.2	79.1
18	9.6	77.5
18	12.8	86.1
24	10.5	82.3
30	17.1	102

Age scaled	Weight scaled	Size scaled
0		
0		
0.01		
0.01		
0.03		
0.03		
0.06		
0.06		
0.09		
0.09		
0.12		
0.12		
0.18		
0.18		
0.24		
0.3		

Exercise 9 - Solution

Scale each feature of the given data points using Decimal Scaling.

Formula: $a'_i = \frac{a_i}{10^j}$

with j being the minimal integer such that $\frac{a_{\max}}{10^j} < 1$

For *Age*: $a_{\max} = 30 \implies j = 2$

$$a'_i = \frac{a_i}{100}$$

Normalization – Decimal Scaling

Age [months]	Weight [kg]	Size [cm]
0	2.5	46.4
0	3.8	52.9
1	3.2	50.4
1	4.9	56.9
3	4.4	56.7
3	6.7	63.4
6	6.2	63.4
6	8.7	70.2
9	7.5	68
9	10.2	75
12	8.4	71.7
12	11.2	79.1
18	9.6	77.5
18	12.8	86.1
24	10.5	82.3
30	17.1	102

Age scaled	Weight scaled	Size scaled
0	0.025	
0	0.038	
0.01	0.032	
0.01	0.049	
0.03	0.044	
0.03	0.067	
0.06	0.062	
0.06	0.087	
0.09	0.075	
0.09	0.102	
0.12	0.084	
0.12	0.112	
0.18	0.096	
0.18	0.128	
0.24	0.105	
0.3	0.171	

Exercise 9 - Solution

Scale each feature of the given data points using Decimal Scaling.

Formula: $a'_i = \frac{a_i}{10^j}$

with j being the minimal integer such that $\frac{a_{\max}}{10^j} < 1$

For *Weight*: $a_{\max} = 17.1 \implies j = 2$

$$a'_i = \frac{a_i}{100}$$

Normalization – Decimal Scaling

Age [months]	Weight [kg]	Size [cm]
0	2.5	46.4
0	3.8	52.9
1	3.2	50.4
1	4.9	56.9
3	4.4	56.7
3	6.7	63.4
6	6.2	63.4
6	8.7	70.2
9	7.5	68
9	10.2	75
12	8.4	71.7
12	11.2	79.1
18	9.6	77.5
18	12.8	86.1
24	10.5	82.3
30	17.1	102

Age scaled	Weight scaled	Size scaled
0.0000	0.0250	0.0464
0.0000	0.0380	0.0529
0.0100	0.0320	0.0504
0.0100	0.0490	0.0569
0.0300	0.0440	0.0567
0.0300	0.0670	0.0634
0.0600	0.0620	0.0634
0.0600	0.0870	0.0702
0.0900	0.0750	0.0680
0.0900	0.1020	0.0750
0.1200	0.0840	0.0717
0.1200	0.1120	0.0791
0.1800	0.0960	0.0775
0.1800	0.1280	0.0861
0.2400	0.1050	0.0823
0.3000	0.1710	0.1020

Exercise 9 - Solution

Scale each feature of the given data points using Decimal Scaling.

Formula: $a'_i = \frac{a_i}{10^j}$

with j being the minimal integer such that $\frac{a_{\max}}{10^j} < 1$

For *Size*: $a_{\max} = 102 \implies j = 3$

$$a'_i = \frac{a_i}{1000}$$

Reduction

Making the data smaller for analysis, but produce the same or similar analysis results.

- **Two types of reduction:**
  - **Feature reduction (keep all the rows, but reduce the number of columns in dataset)**
  - **Instance reduction (store summary, remove or reduce the number of rows)**

Reduction - Aggregation

Weight [kg]	Age [months]	Class
2.5	0	underweight
4.2	0	overweight
2.4	0	underweight
2.7	0	overweight
3.2	1	underweight
5.4	1	overweight
3	1	underweight
4.9	1	overweight
4.4	3	underweight
7.4	3	overweight
4.2	3	underweight
6.7	3	overweight
6.2	6	underweight
9.5	6	overweight
5.8	6	underweight
8.7	6	overweight

Exercise 10

Consider the data table on the left. Reduce the data by dropping features and applying aggregation to

a) Show the minimum weight of overweight babies, for each age class.

b) For each age class, show the average weight.

c) For age class, show the summarized weight of all underweight babies and of all overweight babies.

Reduction - Aggregation

Weight [kg]	Age [months]	Class
2.5	0	underweight
4.2	0	overweight
2.4	0	underweight
2.7	0	overweight
3.2	1	underweight
5.4	1	overweight
3	1	underweight
4.9	1	overweight
4.4	3	underweight
7.4	3	overweight
4.2	3	underweight
6.7	3	overweight
6.2	6	underweight
9.5	6	overweight
5.8	6	underweight
8.7	6	overweight

Exercise 10 - Solution

Consider the data table on the left. Reduce the data by dropping features and applying aggregation to

a) **Show the minimum weight of overweight babies, for each age class.**

b) For each age class, show the average weight.

c) For age class, show the summarized weight of all underweight babies and of all overweight babies.

Reduction - Aggregation

Weight [kg]	Age [months]
2.7	0
4.9	1
6.7	3
8.7	6

Exercise 10 - Solution

Consider the data table on the left. Reduce the data by dropping features and applying aggregation to

a) Show the minimum weight of overweight babies, for each age class.

b) For each age class, show the average weight.

c) For age class, show the summarized weight of all underweight babies and of all overweight babies.

Reduction - Aggregation

Weight [kg]	Age [months]	Class
2.5	0	underweight
4.2	0	overweight
2.4	0	underweight
2.7	0	overweight
3.2	1	underweight
5.4	1	overweight
3	1	underweight
4.9	1	overweight
4.4	3	underweight
7.4	3	overweight
4.2	3	underweight
6.7	3	overweight
6.2	6	underweight
9.5	6	overweight
5.8	6	underweight
8.7	6	overweight

Exercise 10 - Solution

Consider the data table on the left. Reduce the data by dropping features and applying aggregation to

a) Show the minimum weight of overweight babies, for each age class.

b) For each age class, show the average weight.

c) For age class, show the summarized weight of all underweight babies and of all overweight babies.

Reduction - Aggregation

Weight [kg]	Age [months]
2.950	0
4.125	1
5.675	3
7.55	6

Exercise 10 - Solution

Consider the data table on the left. Reduce the data by dropping features and applying aggregation to

a) Show the minimum weight of overweight babies, for each age class.

b) For each age class, show the average weight.

c) For age class, show the summarized weight of all underweight babies and of all overweight babies.

Reduction - Aggregation

Weight [kg]	Age [months]	Class
4.9	0	underweight
6.9	0	overweight
6.2	1	underweight
10.3	1	overweight
8.6	3	underweight
14.1	3	overweight
12.0	6	underweight
18.2	6	overweight

Exercise 10 - Solution

Consider the data table on the left. Reduce the data by dropping features and applying aggregation to

a) Show the minimum weight of overweight babies, for each age class.

b) For each age class, show the average weight.

c) For age class, show the summarized weight of all underweight babies and of all overweight babies.