# Task 3 : Binning

## Task 3a: Equal-Width Binning

number of bins : 3

Width: $\frac{70-13}{3} = \underline{\underline{19}}$

Daten in Bins hinzufügen: 1. [min + width]
2. [min + 2 · width]
⋮
[min + N · width]

immer exklusive!
ausser beim letzten Bin

Bin 1

Daten von 13 bis und ohne 32

d.h

[13, 15, 16, 18, 19, 20, 20, 21, 22, 22, 25, 25, 26, 26, 30]

Avg = 21.2

Bin 2

Daten von 32 (inkl.) bis 51 exkl.

d.h.

[33, 34, 35, 35, 35, 36, 37, 40, 42, 46]

Avg = 37.3

Bin 3

Daten von 51 bis 70 inkl.

[53, 70]

Avg = 61.5

Bins mit Avg füllen

Bin 1 = [ 21.2, 21.2, 21.2, 21.2, 21.2, 21.2, 21.2, 21.2, 21.2, 21.2, 21.2,
21.2, 21.2, 21.2, 21.2 ]

Size = 15

Bin 2 = [ 37.3, 37.3, 37.3, 37.3, 37.3, 37.3, 37.3, 37.3, 37.3, 37.3 ]

Size = 10

Bin 3 = [ 61.5, 61.5 ]

Size = 2

## Task 3b : Equal - Depth Binning

3 bins equal-depth
27 data points → 27 : 9 = 3

9 Elements per bin

Bin 1 = [ 13, 15, 16, 18, 19, 20, 20, 21, 22 ] mean = $18.\overline{22}$

Bin 2 = [ 22, 25, 25, 26, 26, 30, 33, 34, 35 ] mean = $28.\overline{44}$

Bin 3 = [ 35, 35, 36, 37, 40, 42, 46, 53, 70 ] mean = $43.\overline{77}$

Replacing with average

bin 1 = [ $18.\overline{22}$, $18.\overline{22}$, $18.\overline{22}$, $18.\overline{22}$, $18.\overline{22}$, $18.\overline{22}$, $18.\overline{22}$, $18.\overline{22}$, $18.\overline{22}$ ]

bin 2 = [ $28.\overline{44}$, $28.\overline{44}$, $28.\overline{44}$, $28.\overline{44}$, $28.\overline{44}$, $28.\overline{44}$, $28.\overline{44}$, $28.\overline{44}$, $28.\overline{44}$ ]

bin 3 = [ $43.\overline{77}$, $43.\overline{77}$, $43.\overline{77}$, $43.\overline{77}$, $43.\overline{77}$, $43.\overline{77}$, $43.\overline{77}$, $43.\overline{77}$, $43.\overline{77}$ ]

## Unterschiede 3a und 3b

It can happen that in the Equal-Width binning the outliers dominate the result while the Equal-Depth binning handles skewed data well. The Equal-Width binning divides the the range of data into N intervals of equal size while the Equal-Depth binning divides the range into N intervals where each interval contains approximately the same number of records.

## Wann sollte Data binning verwendet werden und wann nicht

Data binning should be used when you want to categorize data e.g. you want to show in which age the people are the most intelligent so you categorize the data age 1-10, 20-30, 40-50 and so one. The goal of data binning is to make the patterns more noticeable, to eliminate disturbances and outliers from the given data. Data binning is used when you want to transform numerical or continuous variable in categorial feature. There are also circumstances in which binning should not be applied when you wan't to see the real details of the data e.g. you wanna see the outliers too but you have then also meaningless (noisy) data.