

# Data Exploration



JENS BAETENS

1 t test 7017 pts

6 L Lukaas 3100 pts

11 ? Anonymou  
s 1200 pts

2 B Bean 5200 pts

7 b bruh 2125 pts

12 F Friend 1175 pts

3 M Mike 3925 pts

8 R ightto 2100 pts

13 M Mike 1167 pts



Click on the projected screen to start the question

5

4 J Jesus 3192 pts

9 g gustavo  
fring 1225 pts

14 J Joske  
Vermeulen 1050 pts

10

15

5 7 Zak 3150 pts

10 I Jose 1208 pts

15 N Nohhier 925 pts

wooclap



100 %



18



# Top players

1

t

test

7017 pts

6

L

Lukaas

3100 pts

11

?

Anonymous

1200 pts

2

B

Bean

5200 pts

7

b

bruh

2125 pts

12

F

Friend

1175 pts

3

M

Mike

3925 pts

8

R

Rightho

2100 pts

13

M

Mike

1167 pts

Click on the projected screen to start the question

5

4

J

Jesus

3192 pts

9

g

gustavo  
fring

1225 pts

14

J

Joske  
Vermeulen

1050 pts

10

15

5

Z

Zak

3150 pts

10

J

Jose

1208 pts

15

N

Nohhier

925 pts

# Top players

1

t

test

7017 pts

6

L

Lukaas

3100 pts

11

?

Anonymous

1200 pts

2

B

Bean

5200 pts

7

b

bruh

2125 pts

12

F

Friend

1175 pts

3

M

Mike

3925 pts

8

R

Rightho

2100 pts

13

M

Mike

1167 pts

Click on the projected screen to start the question

5

4

J

Jesus

3192 pts

9

g

gustavo  
fring

1225 pts

14

J

Joske  
Vermeulen

1050 pts

10

15

5

Z

Zak

3150 pts

10

J

Jose

1208 pts

15

N

Nohhier

925 pts

# Top players

1

t

test

7017 pts

6

L

Lukaas

3100 pts

11

?

Anonymous

1200 pts

2

B

Bean

5200 pts

7

b

bruh

2125 pts

12

F

Friend

1175 pts

3

M

Mike

3925 pts

8

R

Rightho

2100 pts

13

M

Mike

1167 pts

Click on the projected screen to start the question

5

4

J

Jesus

3192 pts

9

g

gustavo  
fring

1225 pts

14

J

Joske  
Vermeulen

1050 pts

10

15

5

Z

Zak

3150 pts

10

J

Jose

1208 pts

15

N

Nohhier

925 pts



# Top players

1

t

test

7017 pts

6

L

Lukaas

3100 pts

11

?

Anonymous

1200 pts

2

B

Bean

5200 pts

7

b

bruh

2125 pts

12

F

Friend

1175 pts

3

M

Mike

3925 pts

8

R

Righto

2100 pts

13

M

Mike

1167 pts

Click on the projected screen to start the question

5

4

J

Jesus

3192 pts

9

g

gustavo  
fring

1225 pts

14

J

Joske  
Vermeulen

1050 pts

10

15

5

Z

Zak

3150 pts

10

J

Jose

1208 pts

15

N

Nohhier

925 pts

# Top players

1

t

test

7017 pts

6

L

Lukaas

3100 pts

11

?

Anonymous

1200 pts

2

B

Bean

5200 pts

7

b

bruh

2125 pts

12

F

Friend

1175 pts

3

M

Mike

3925 pts

8

R

Righto

2100 pts

13

M

Mike

1167 pts

Click on the projected screen to start the question

5

4

J

Jesus

3192 pts

9

g

gustavo  
fring

1225 pts

14

J

Joske  
Vermeulen

1050 pts

10

15

5

Z

Zak

3150 pts

10

J

Jose

1208 pts

15

N

Nohhier

925 pts

# Wat is het?

---


Wordt ook Exploratory Data Analysis of EDA genoemd

Beter begrip van de karakteristieken van de dataset

Waarom?

- Welk model is het best geschikt?
- Herkennen van patronen die niet door tools herkend worden
- Welke kolommen / features kunnen gebruikt worden
- Hoe kan de data beter bewerkt worden om betere resultaten te bekomen





# Wat zit er in de dataset?

---

*algemeen*

①

Hoeveel rijen (observations) en kolommen (features) zijn er?

Wat voor data zit er in elke kolom

- Categorieke data of numerieke?
- Discrete of Continue?

② Per kolom

*↳ per verdeling per kolom gebalanceerd?*

③ Naar de kolommen

*↳ Verbonden tussen kolommen*

df.info() → numerische kolommen  
df.describe() → numerische kolommen

# Technieken – Unieke Waarden

↳ categorische kolommen

Het aantal verschillende waarden per kolom

→ df.unique() → ##

Kan gebruikt worden voor kolommen die een categorie bevatten

→ df.unique() → alle categorieën

- Geeft het aantal elementen in elke categorie weer

→ df.value\_counts()

Kan voorgesteld worden in een barplot

- 1 bar per kolom met categorische data

Kan gebruikt worden om te gebalanceerdheid van een dataset te controleren

60% → klasse 1  
40% → klasse 2

14%  
86%

↑  
niet gebalanceerd

undersampling



oversampling



# Technieken – Frequentie

---

Geef weer hoe frequent een waarde voorkomt in  
een kolom *→ `value_counts()`*

Kan gebruikt worden voor kolommen die een  
categorie bevatten

Kan voorgesteld worden in een barplot

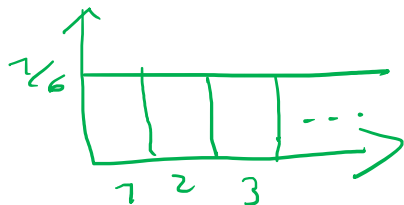
- 1 plot per kolom
- 1 bar per unieke waarde

# Kansen

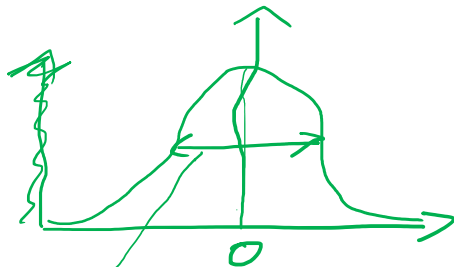
De kans  
 probability  $\Pr[X=1] = 1/6$   
 $2 = 2/6$   
 $\vdots$   
 $6 = 1/6$

Event  $X$

$\rightarrow$  (kansverdeling)



uniforme  
verdeling



Std  
 - Standard deviation  
 - standaardafwijking  
 gemiddelde  
 mean

Normal verdeling = mean = 0  
 std = 1

$$\text{mean} = E[X] = \sum_{i=1}^6 i \times \Pr[X=i]$$

- expected value  
 - verwachte waarde

$$= 3,5$$

$$\text{variance} = E[(X - E[X])^2]$$

$\approx \text{std}^2$

$\rightarrow$  spreiding t.o.v.  
 gemiddelde

# Technieken – Statistische waarden

Een aantal interessante waarden berekenen en vergelijken:

- Gemiddelden ( $E[X]$ )
- Minimum / Maximum
- Variantie (Informatie over de spreiding)  
 $= E[(X - E[X])^2]$
- Mediaan / IQR beter als er veel outliers/extreme waarden zijn

- Outlier als waarde kleiner is dan 25% kwartiel – 1.5 IQR
- Extreem als waarde kleiner is dan 25% kwartiel – 3.5 IQR

$\rightarrow 75\% + 1.5$

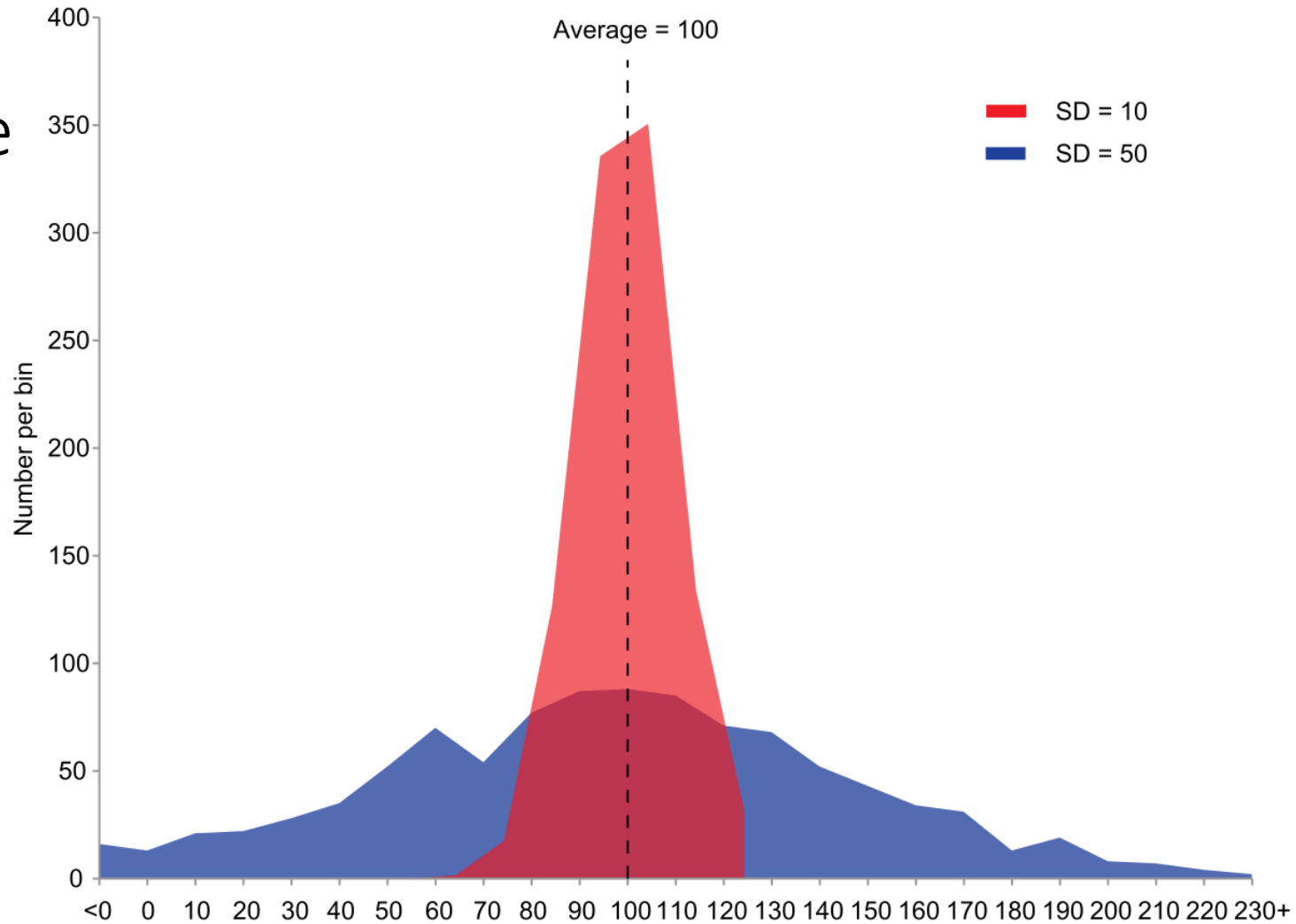
Toepasbaar op numerieke kolommen

$\rightarrow$  outliers opsporen

$\rightarrow$  fouten gaan zoeken (onverwacht grote/onmogelijke waarden)

~~Bereken~~ opp :  $1000 \text{ m}^2$  Breedte = 20  
lengte = 20

# Variance



# Technieken – Histogram

Geeft informatie over in welk bereik de meeste waarden vallen.

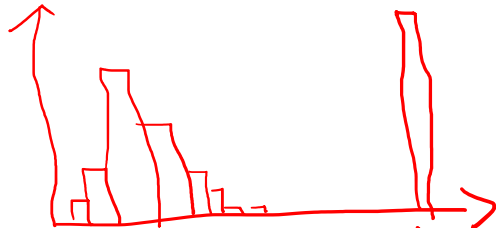
Aantal bins heeft een grote impact

↳ de intervallen waarin data  
samengevoerd wordt

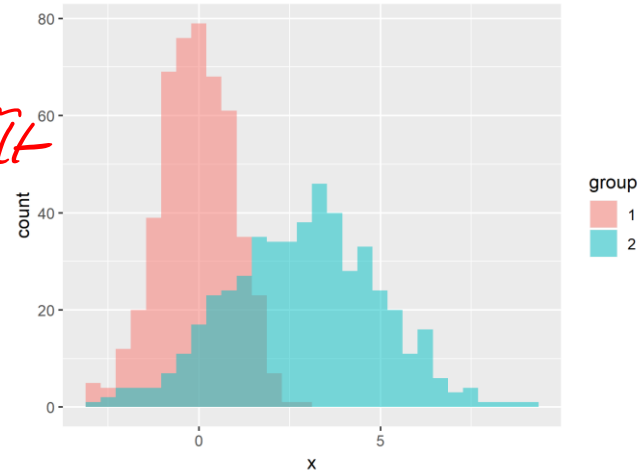
Een histogram per kolom met numerieke data

Komt gedeeltelijk overeen met een kansverdeling

Typisch Gaussiaanse verdeling / normale verdeling



↳ echte waarden of ontbrekend?

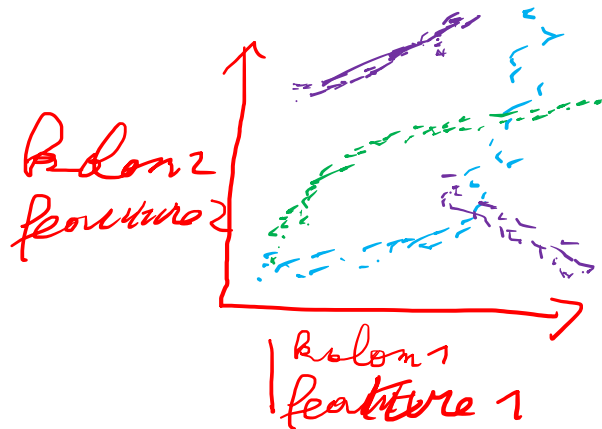


# Zoeken naar verbanden tussen features

Met behulp van een scatterplot zoeken naar features die met elkaar verband kunnen houden.

Voor numerieke waarden

Alle combinaties afzoeken kan veel werk zijn



1) kwadratisch

2) logaritmisch / vierkantswortel  
↳  $\sqrt{\text{pow}(0,5)}$

3) niet lineair

↳ afgeleide bestaat niet overal



# Technieken – correlation Heat Map

Geeft de samenhang tussen twee elementen weer

als x stijgt gaat y  
- stijgen?  
- dalen?  
- weet het niet?

Kans als A hoog is dat dan ook B hoog is: Positieve Correlatie

Kans als A hoog is dat B dan laag is: Negatieve Correlatie

Wordt berekend als:  $\frac{E[(X-E[X])(Y-E[Y])]}{\sqrt{\text{Var}(X)}\sqrt{\text{Var}(Y)}}$

$$\text{Var}(X) = E[(X-E[X])^2]$$

(1) Heatmap met zowel op X als Y als de numerieke kolommen

Bekijk de correlatie van 2 kolommen in meer detail met een scatterplot

(2)

	K <sub>1</sub>	K <sub>2</sub>	K <sub>3</sub>
K <sub>1</sub>	1	0,1	-0,2
K <sub>2</sub>	0,1	1	0,3
K <sub>3</sub>	-0,2	0,3	1

correlatiematrix

↳ diagonaal steeds 1

↳ symmetrisch

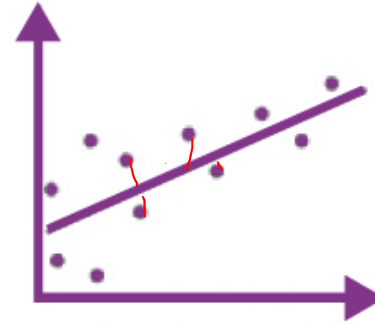
Sterke correlatie  
sterk "verband  
tussen 2 features

# Correlation

*weak = verband  
minder  
duidelijk*



**Strong positive correlation**



**Weak positive correlation**



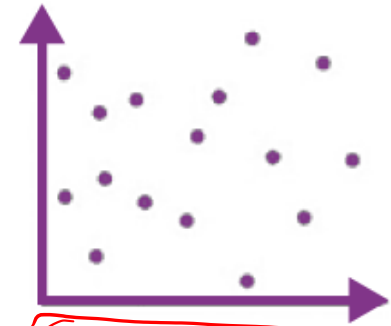
**Strong negative correlation**



**Weak negative correlation**



**Moderate negative correlation**



**No correlation**

# Technieken – Pearson correlation and Trend

---

Plot een aantal interessante combinaties uit de heatmap als scatterplot.

Bijvoorbeeld de combinaties met een sterke negatieve of positieve correlatie

↳ zie code

# Technieken – Cramer-v correlation

---

Correlation heat map voor kolommen met categorieke data.

Cramer's V correlatie =  $\sqrt{\text{phi} / \min(r-1, k-1)}$

Waar  $\text{phi} = \sum_{i,j} \frac{\text{Pr}[A=i, B=j]^2}{\text{Pr}[A=i] * \text{Pr}[B=j]} - 1$

De correlatie is

- 0 als de kolommen onafhankelijk zijn
- 1 als de kolommen volledig samenhangen

*→ hoe dichtbij 1  
hoe sterkere correlatie*

Correlatie kan in meer detail bekeken worden met een bubble plot

- size bubble is het aantal keer het voorkomt

# Phi and Cramer's V

## Interpretation

> 0.25

Very strong

> 0.15

Strong

> 0.10

Moderate

> 0.05

Weak

> 0

No or very weak

Orukkaan

minderheit

Orukkaan

# Technieken – Important Features

*kolommen*

Important features zijn de features die een grote impact hebben op de gewenste feature.

Kan uit de correlation heatmaps gehaald worden

Getoond als een bar-plot met op de x-as de kolommen en op de y-as de correlatie coëfficiënt

*⇒ enkel kolommen met meeste informatie gaan overhouden*

*↓  
zit in correlatie*

# Technieken – Outlier Detection

---

Wordt ook anomaly detection genoemd

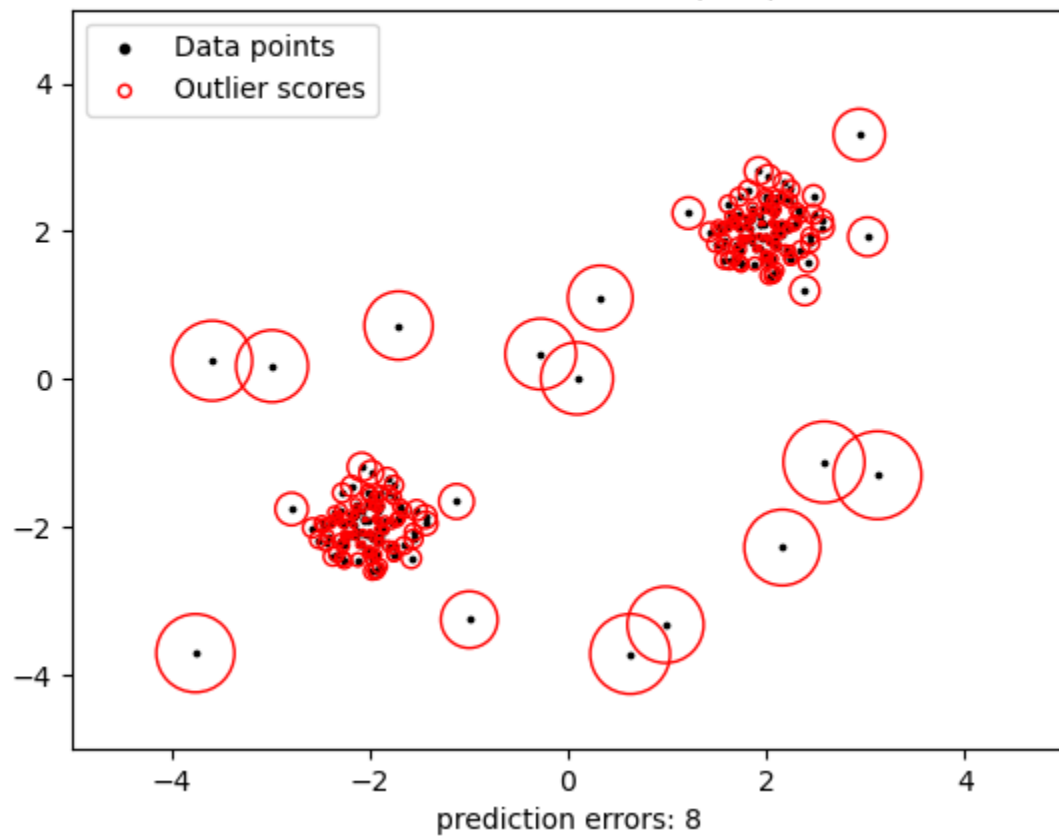
Outliers komen overeen met zeldzame gevallen (positief of negatief)

Kan gedaan worden door

- standard deviation analysis
- Isolation forest (Machine learning techniek)

Bubble chart met op de x-as alle numerieke kolommen

Local Outlier Factor (LOF)





# Technieken – Outlier analysis

---

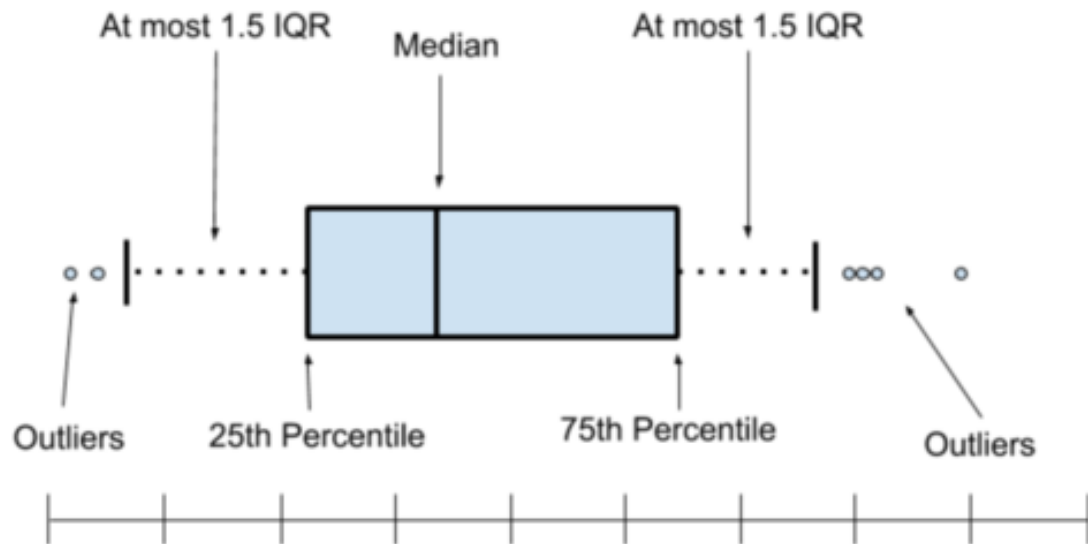
Meer gedetailleerde overzicht van outliers en statistische waarden

Enkele kolom

- Box plot

Meerdere kolommen

- Scatter plot en outliers in aparte kleur
- Outliers moeten eerst gedetecteerd worden (op basis van statistische gegevens of ML-technieken zoals Standard Deviation Analysis of Isolation Forest)



X Axis

Shows data range and labels  
the values you are graphing.

# Technieken – Pareto Analysis

---

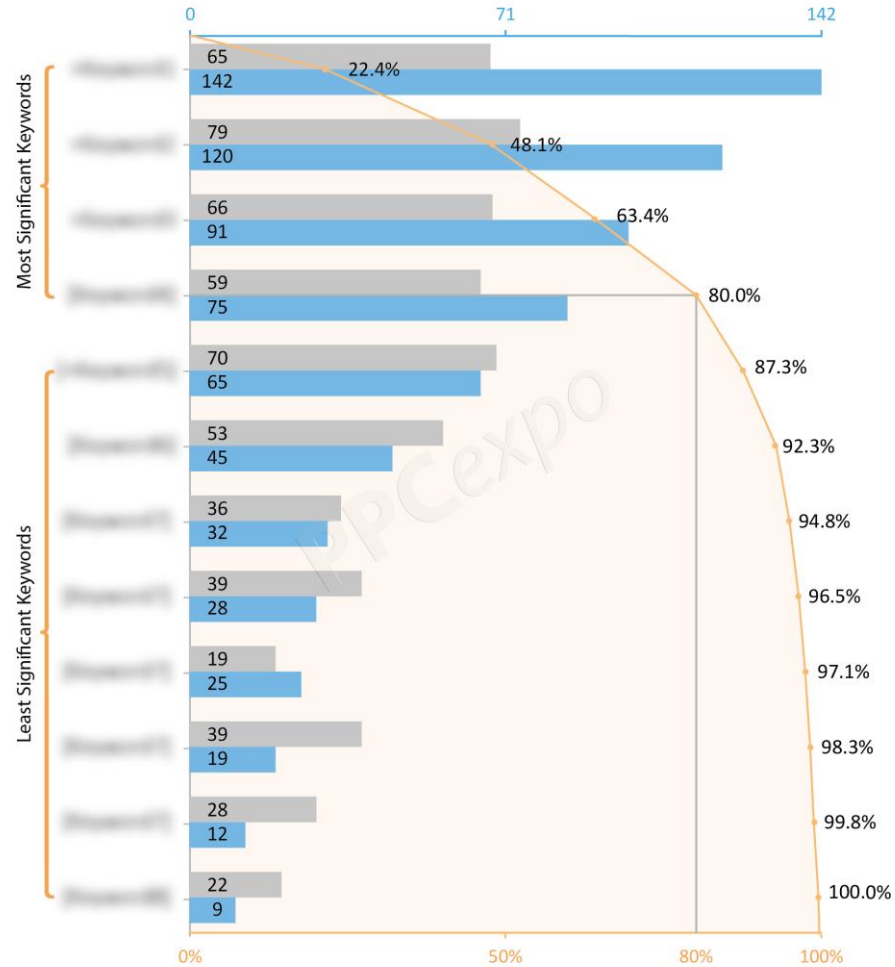
Om te onderzoeken welke data belangrijk kan zijn.

Pareto 80-20 vaak gebruikt:

- De waarden kleiner dan 20% van het maximum zijn klein
- De waarden groter dan 80% van het maximum zijn groot

Afhankelijk van je vraag kan 1 of beide groepen genegeerd worden.

# Pareto Chart



1 t test 7017 pts

6 L Lukaas 3100 pts

11 ? Anonymou  
s 1200 pts

2 B Bean 5200 pts

7 b bruh 2125 pts

12 F Friend 1175 pts

3 M Mike 3925 pts

8 R ightto 2100 pts

13 M Mike 1167 pts



Click on the projected screen to start the question

5

4 J Jesus 3192 pts

9 g gustavo  
fring 1225 pts

14 J Joske  
Vermeulen 1050 pts

10

15

5 Z Zak 3150 pts

10 J Jose 1208 pts

15 N Nohhier 925 pts

wooclap



100 %



40% correct

5



1 t test 7017 pts

6 L Lukaas 3100 pts

11 ? Anonymou  
s 1200 pts

2 B Bean 5200 pts

7 b bruh 2125 pts

12 F Friend 1175 pts

3 M Mike 3925 pts

8 R ightto 2100 pts

13 M Mike 1167 pts

4 J Jesus 3192 pts

9 g gustavo  
fring 1225 pts

14 J Joske  
Vermeulen 1050 pts

5 Z Zak 3150 pts

10 J Jose 1208 pts

15 N Nohhier 925 pts



Click on the projected screen to start the question



wooclap



100 %



4



1 t test 7017 pts

6 L Lukaas 3100 pts

11 ? Anonymous 1200 pts

2 B Bean 5200 pts

7 b bruh 2125 pts

12 F Friend 1175 pts

3 M Mike 3925 pts

8 R ightto 2100 pts

13 M Mike 1167 pts

4 J Jesus 3192 pts

9 g gustavo  
fring 1225 pts

14 J Joske  
Vermeulen 1050 pts

5 Z Zak 3150 pts

10 J Jose 1208 pts

15 N Nohhier 925 pts

Click on the projected screen to start the question

wooclap



100 %



40% correct

5 / 18



1 t test 7017 pts

6 L Lukaas 3100 pts

11 ? Anonymous 1200 pts

2 B Bean 5200 pts

7 b bruh 2125 pts

12 F Friend 1175 pts

3 M Mike 3925 pts

8 R ightto 2100 pts

13 M Mike 1167 pts



Click on the projected screen to start the question

5

4 J Jesus 3192 pts

9 g gustavo fring 1225 pts

14 J Joske Vermeulen 1050 pts

10

15 5 Z Zak 3150 pts

10 J Jose 1208 pts

15 N Nohhier 925 pts

wooclap



100 %



50% correct

8 / 18





1 t test 7017 pts

6 L Lukaas 3100 pts

11 ? Anonymous 1200 pts

2 B Bean 5200 pts

7 b bruh 2125 pts

12 F Friend 1175 pts

3 M Mike 3925 pts

8 R ightto 2100 pts

13 M Mike 1167 pts

4 J Jesus 3192 pts

9 g gustavo fring 1225 pts

14 J Joske Vermeulen 1050 pts

5 Z Zak 3150 pts

10 J Jose 1208 pts

15 N Nohhier 925 pts

Click on the projected screen to start the question

wooclap



100 %



67% correct

9 / 18



1 t test 7017 pts

6 L Lukaas 3100 pts

11 ? Anonymous 1200 pts

2 B Bean 5200 pts

7 b bruh 2125 pts

12 F Friend 1175 pts

3 M Mike 3925 pts

8 R Righto 2100 pts

13 M Mike 1167 pts

4 J Jesus 3192 pts

9 g gustavo fring 1225 pts

14 J Joske Vermeulen 1050 pts

5 Z Zak 3150 pts

10 J Jose 1208 pts

15 N Nohhier 925 pts

wooclap



100 %



43% correct

7 / 18

