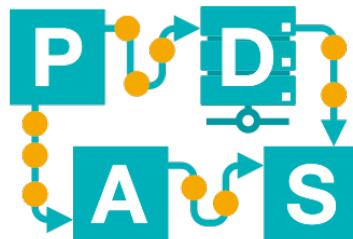


Data Visualization and Exploration

Lecture 2

IDS-L2



Chair of Process
and Data Science

RWTH AACHEN
UNIVERSITY

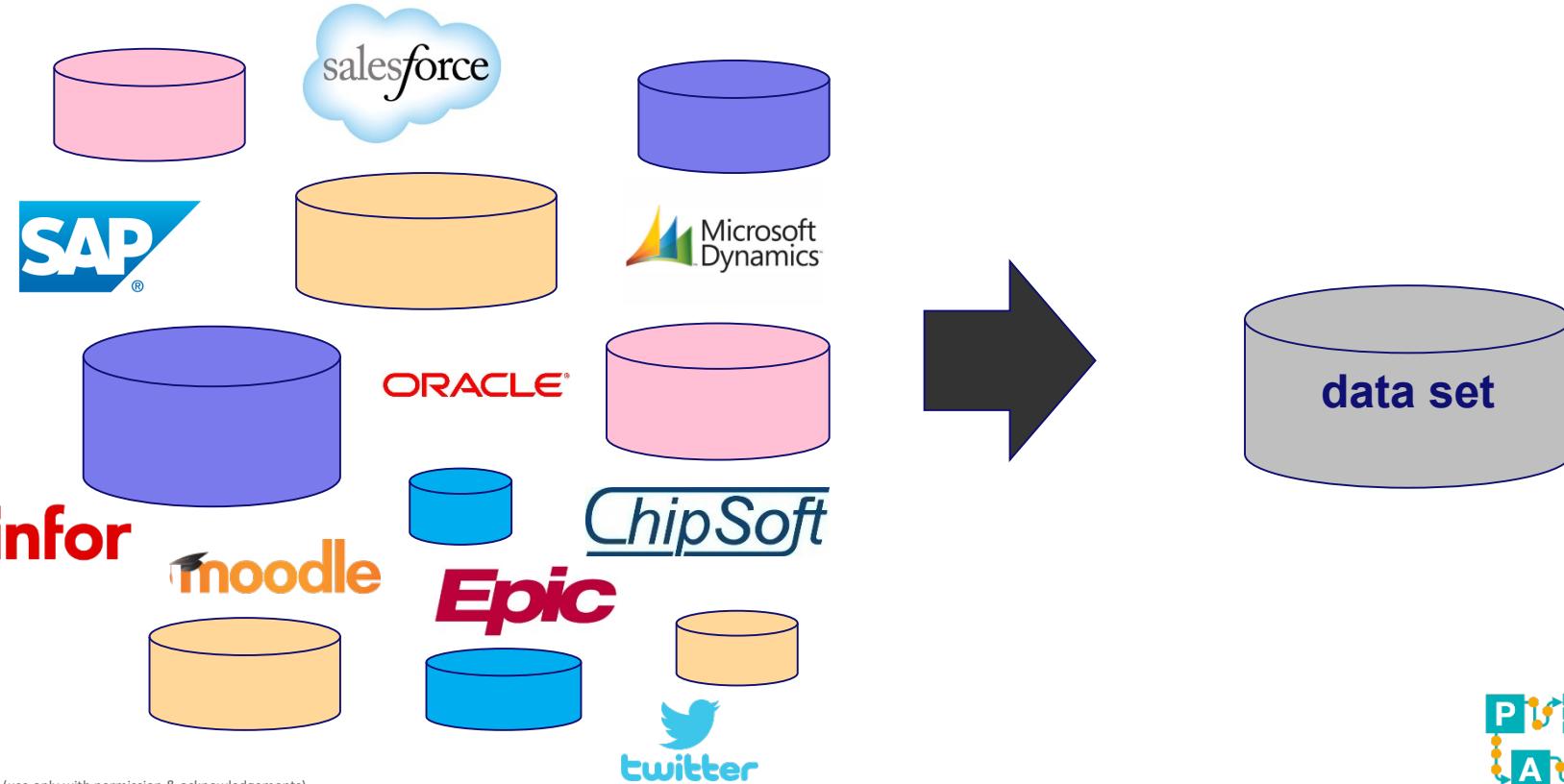
Outline of Today's Lecture

- Data extraction
- Tabular data
- Importance of visualization
- Characterizing individual features
- Data quality
- Showing relations among features
- Preparing for analysis
- Good and poor visualizations

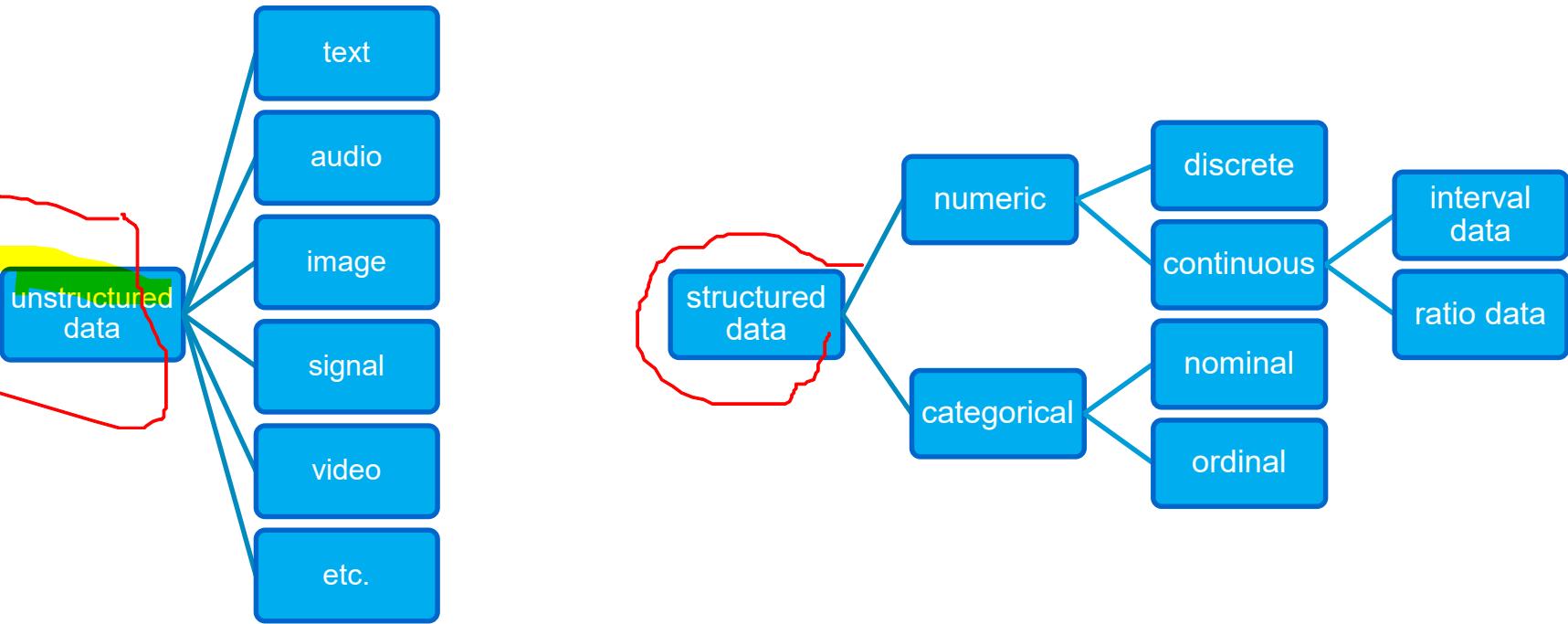
Data extraction



Extracting data



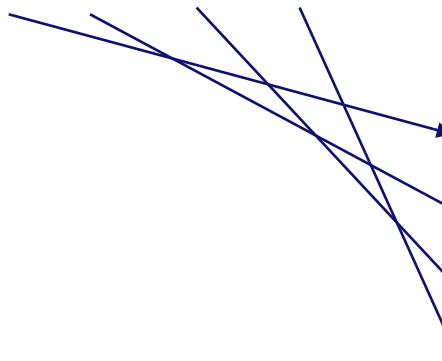
Types of data



Feature extraction



CDDDDCDCDCDCDCDCDCD



type	weight	age	...
Cat	3.1 kg	10 years	...
Dog	1.4 kg	7 years	...
Dog	2.6 kg	8 years	...
Dog	1.6 kg	3 years	...

Focus will be on tabular data

features

instances

f1	f2	f3	f4	f5	f7	f8	f9



Chair of Process
and Data Science

Special features: Time and Target

features

instances

f1	f2	f3	f4	f5	f7	time	target
						16:32	accept
						16:33	reject
						16:41	accept
						16:55	reject
						17:01	reject
						17:03	reject

- Alternative names for **descriptive features**

- predictor variables

- independent variables

- Alternative names for **target feature**

- response variable

- dependent variable

- Alternative names for **instances**: individuals, entities, cases, objects, or records.



Importance of visualization



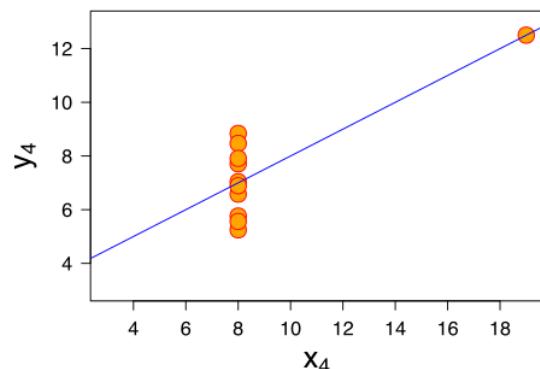
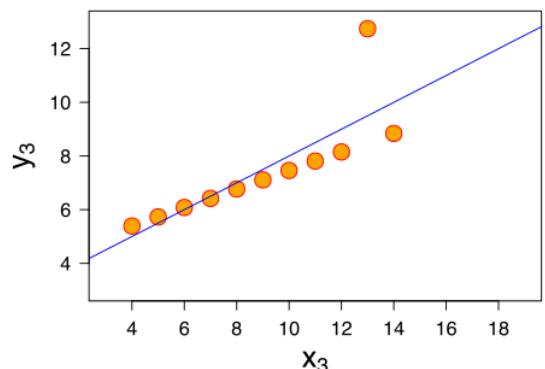
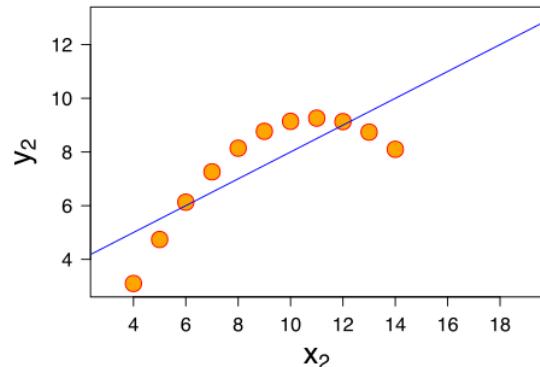
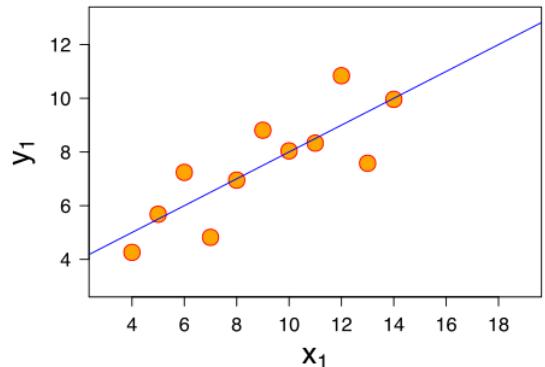
Anscombe's quartet

Data set 1		Data set 2		Data set 3		Data set 4	
x	y	x	y	x	y	x	y
10,0	8,04	10,0	9,14	10,0	7,46	8,0	6,58
8,0	6,95	8,0	8,14	8,0	6,77	8,0	5,76
13,0	7,58	13,0	8,74	13,0	12,74	8,0	7,71
9,0	8,81	9,0	8,77	9,0	7,11	8,0	8,84
11,0	8,33	11,0	9,26	11,0	7,81	8,0	8,47
14,0	9,96	14,0	8,10	14,0	8,84	8,0	7,04
6,0	7,24	6,0	6,13	6,0	6,08	8,0	5,25
4,0	4,26	4,0	3,10	4,0	5,39	19,0	12,50
12,0	10,84	12,0	9,13	12,0	8,15	8,0	5,56
7,0	4,82	7,0	7,26	7,0	6,42	8,0	7,91
5,0	5,68	5,0	4,74	5,0	5,73	8,0	6,89

- Mean of x = 9
- Sample variance of x = 11
- Mean of y ≈ 7.50
- Sample variance of y ≈ 4.125
- Correlation between x and y ≈ 0.816
- Linear regression line: y = 3.00 + 0.500x

Anscombe's quartet

Data set 1		Data set 2		Data set 3		Data set 4	
x	y	x	y	x	y	x	y
10,0	8,04	10,0	9,14	10,0	7,46	8,0	6,58
8,0	6,95	8,0	8,14	8,0	6,77	8,0	5,76
13,0	7,58	13,0	8,74	13,0	12,74	8,0	7,71
9,0	8,81	9,0	8,77	9,0	7,11	8,0	8,84
11,0	8,33	11,0	9,26	11,0	7,81	8,0	8,47
14,0	9,96	14,0	8,10	14,0	8,84	8,0	7,04
6,0	7,24	6,0	6,13	6,0	6,08	8,0	5,25
4,0	4,26	4,0	3,10	4,0	5,39	19,0	12,50
12,0	10,84	12,0	9,13	12,0	8,15	8,0	5,56
7,0	4,82	7,0	7,26	7,0	6,42	8,0	7,91
5,0	5,68	5,0	4,74	5,0	5,73	8,0	6,89



Explore your data first!

Charles Minard's 1869 chart showing the number of men in Napoleon's 1812 Russian campaign army, their movements, as well as the temperature they encountered on the return path.

Carte Figurative des pertes successives en hommes de l'Armée Française dans la Campagne de Russie 1812-1813.
Dessinée par M. Minard, Inspecteur Général des Ponts et Chaussées en retraite Paris, le 20 Novembre 1869.

Les nombres d'hommes présents sont représentés par les largeurs des zones colorées à raison d'un millimètre pour dix mille hommes; ils sont de plus écrits en travers des zones. Le rouge désigne les hommes qui entrent en Russie; le noir ceux qui en sortent. — Les renseignements qui ont servi à dresser la carte ont été puisés dans les ouvrages de M. Chiers, de Clément, de Fezensac, de Chambray et le journal inédit de Jacob, pharmacien de l'Armée depuis le 28 Octobre.

Pour mieux faire juger à l'œil la diminution de l'armée, j'ai supposé que les corps du Prince Jérôme et du Maréchal Davout, qui avaient été détachés sur Minsk en Mohilow et se rejoignent vers Orscha en Witlob, avaient toujours marché avec l'armée.

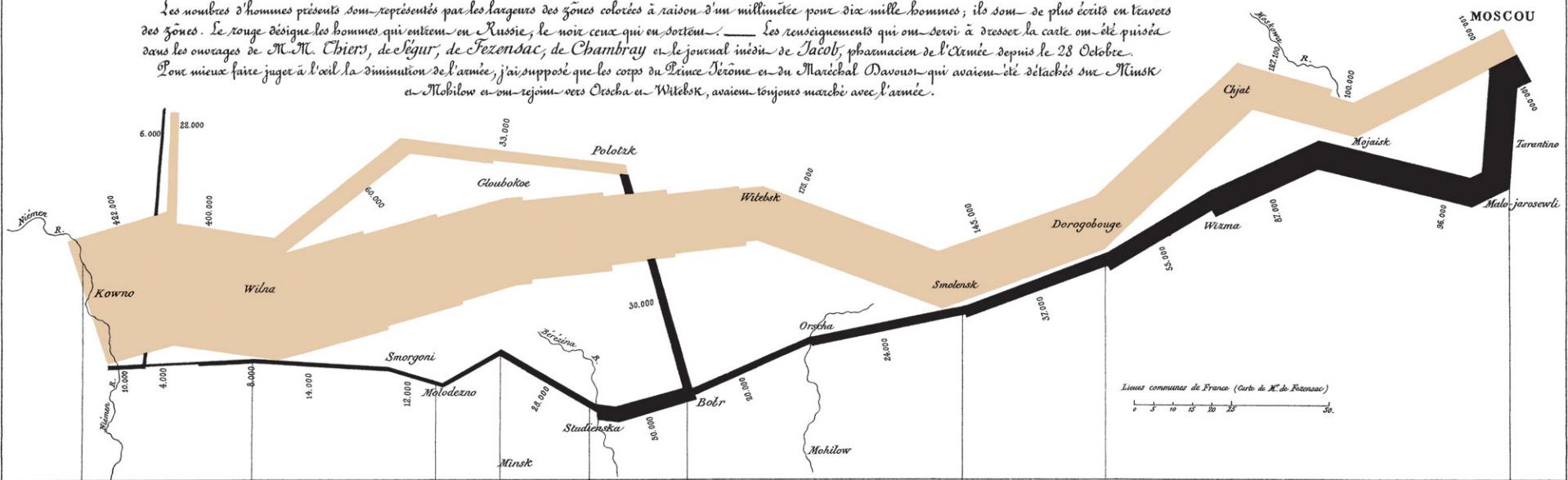
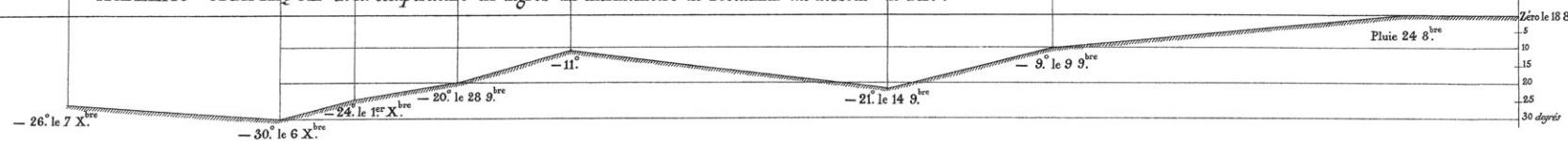
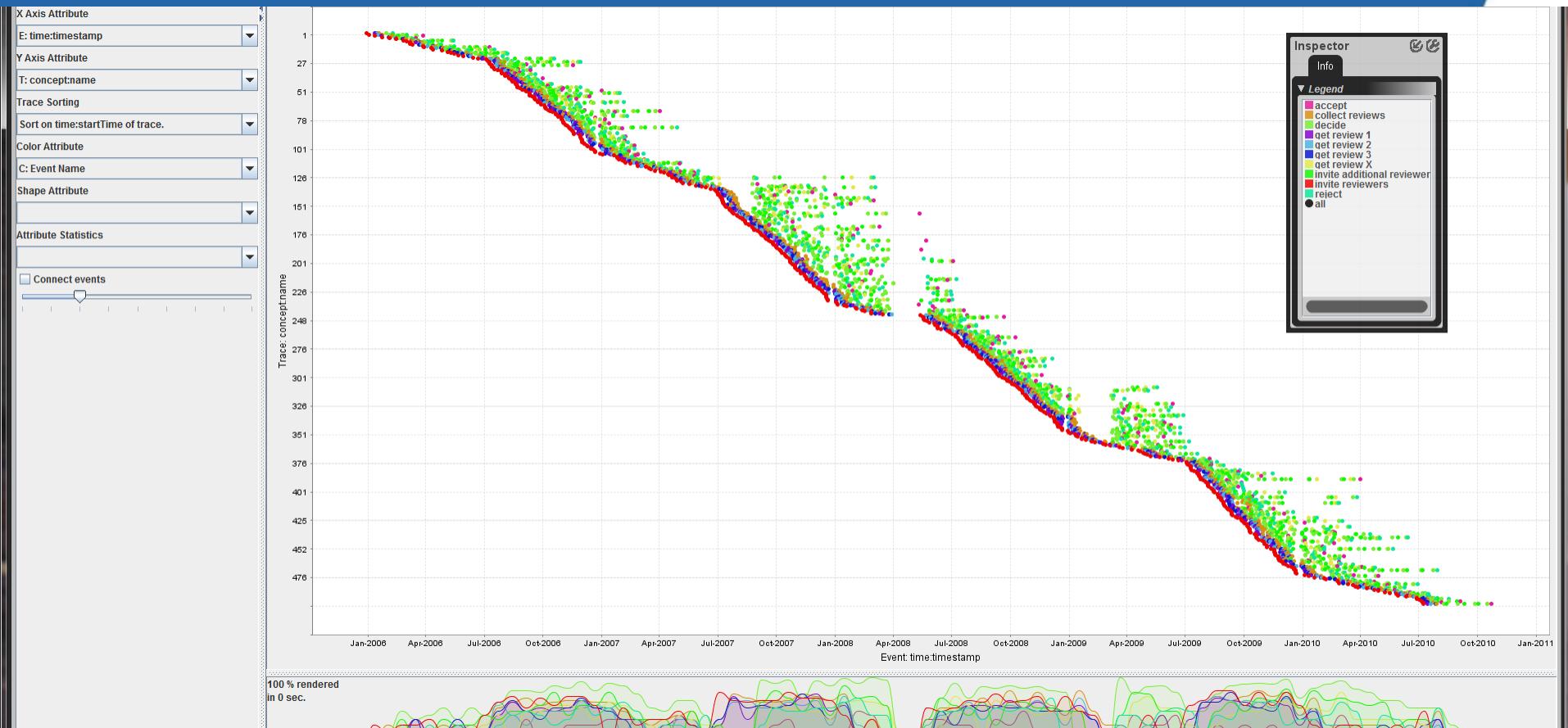


TABLEAU GRAPHIQUE de la température en degrés du thermomètre de Réaumur au dessous de zéro.

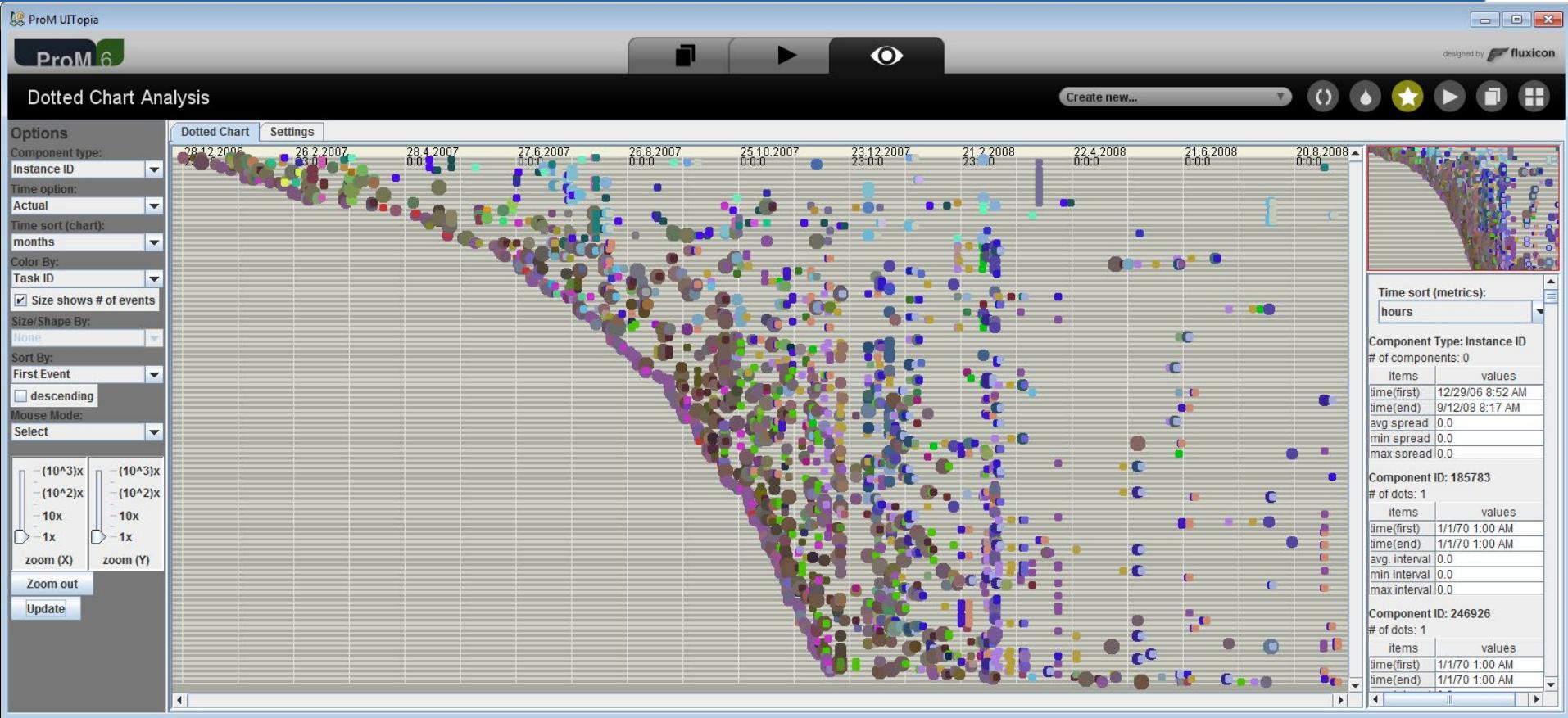
Les cosaques passent au galop
le Niemen gelé.



Event data



Event data



Characterizing individual features



Let's focus on one feature

features

instances	f1	f2	f3	f4	f5	f7	f8	f9

Investigating individual features

(a) Continuous Features

Feature	Count	% Miss.	Card.	Min.	1 st Qrt.	Mean	Median	3 rd Qrt.	Max.	Std. Dev.
—	—	—	—	—	—	—	—	—	—	—
—	—	—	—	—	—	—	—	—	—	—
—	—	—	—	—	—	—	—	—	—	—

(b) Categorical Features

Feature	Count	% Miss.	Card.	Mode	Mode Freq.	Mode %	2 nd Mode	2 nd Mode Freq.	2 nd Mode %
—	—	—	—	—	—	—	—	—	—
—	—	—	—	—	—	—	—	—	—
—	—	—	—	—	—	—	—	—	—

Tables taken from Fundamentals of Machine Learning for Predictive Data Analytics by J. Kelleher, B. Mac Namee and A. D'Arcy.

Investigating individual features

(a) Continuous Features

Feature	Count	% Miss.	Card.	Min.	1 st Qrt.	Mean	Median	3 rd Qrt.	Max.	Std. Dev.

number of instances

percentage missing

cardinality:
number of unique values

minimum

mean

median
(middle value)

maximum

standard deviation

Investigating individual features

(a) Continuous Features

Feature	Count	% Miss.	Card.	Min.	1 st Qrt.	Mean	Median	3 rd Qrt.	Max.	Std. Dev.
—	—	—	—	—	—	—	—	—	—	—
—	—	—	—	—	—	—	—	—	—	—
—	—	—	—	—	—	—	—	—	—	—

25th percentile: largest value for the quarter of instances having the lowest values

75th percentile: smallest value for the quarter of instances having the highest values

*Slightly more complex, see page 530 of Fundamentals of Machine Learning for Predictive Data Analytics by J. Kelleher, B. Mac Namee and A. D'Arcy.

Investigating individual features

(b) Categorical Features

Feature	Count	% Miss.	Card.	Mode	Mode Freq.	Mode %	2 nd Mode	2 nd Mode Freq.	2 nd Mode %
—	—	—	—	—	—	—	—	—	—
—	—	—	—	—	—	—	—	—	—
—	—	—	—	—	—	—	—	—	—

number of instances

percentage missing

cardinality:
number of unique values

mode: most common value

frequency of mode

percentage of mode

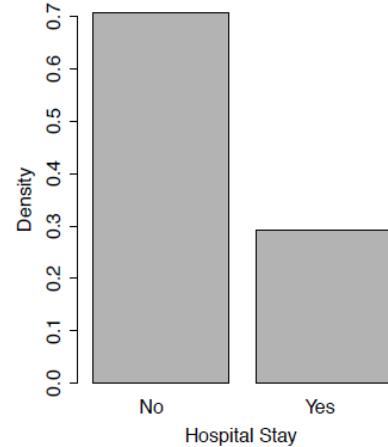
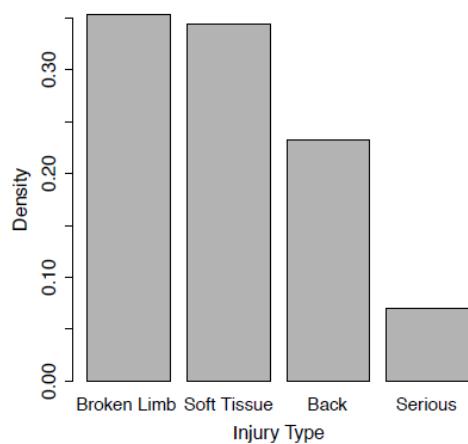
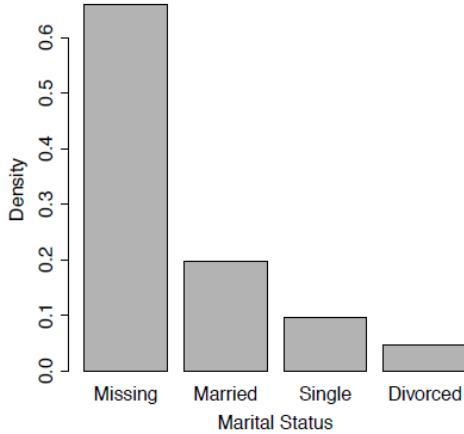
similar values for second most common value

Example table (from book)

Example table (from book)

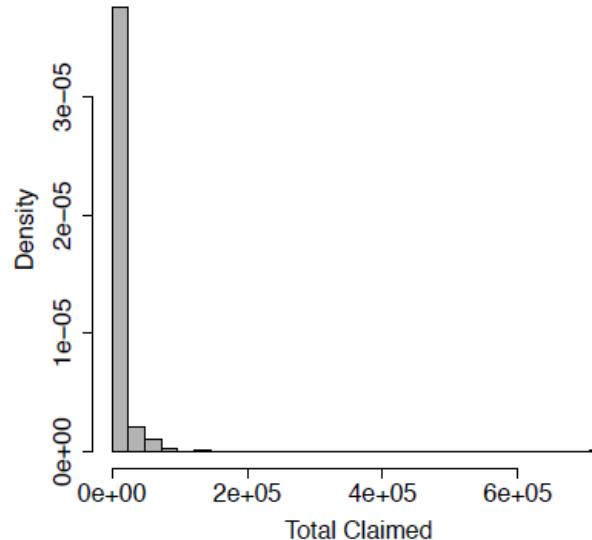
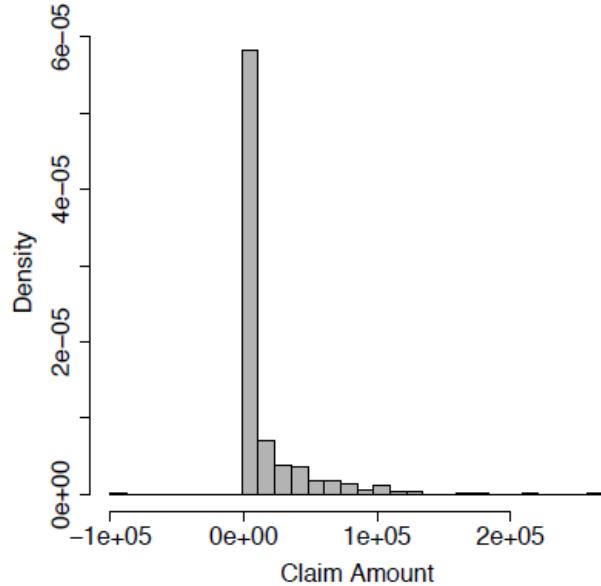
Visualizations of distributions

Y-axis can be frequency or a percentage



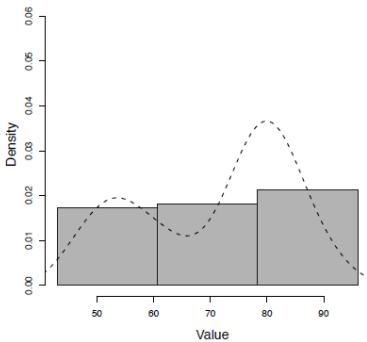
Visualizations of distributions

for continuous variables one may want to group items (binning)

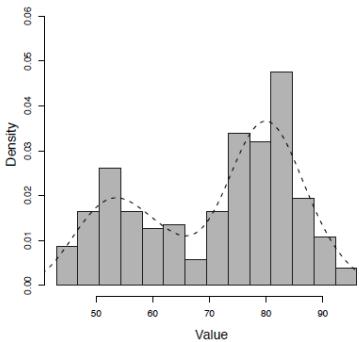


Visualizations of distributions

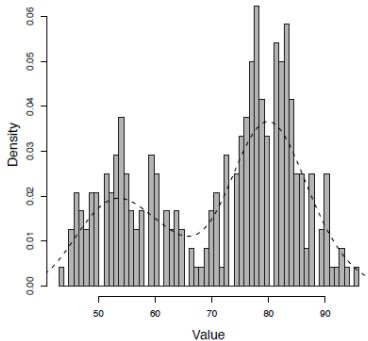
for continuous variables one may want to group items (binning)



(e) 3 bins

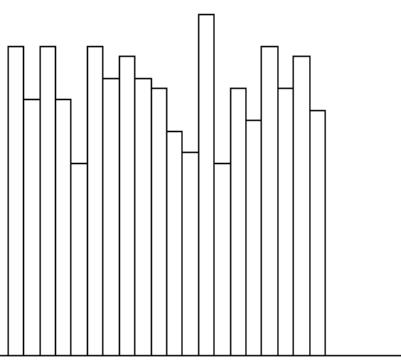


(f) 14 bins

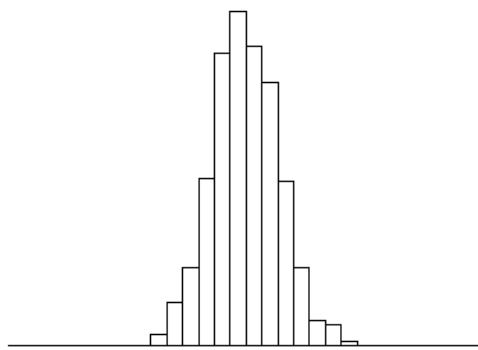


(g) 60 bins

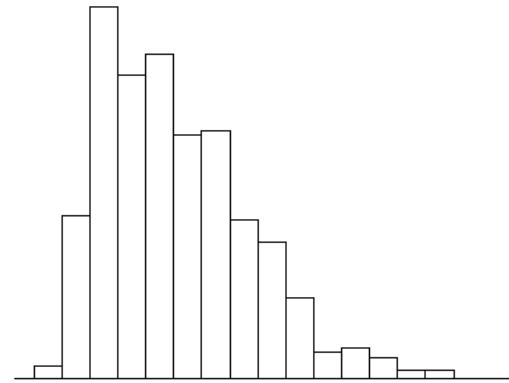
Different types of histograms



(a) Uniform

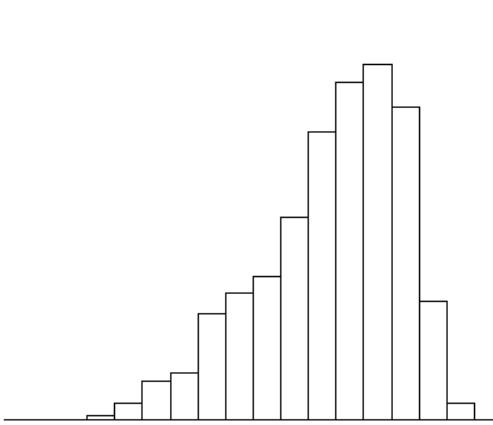


(b) Normal (Unimodal)

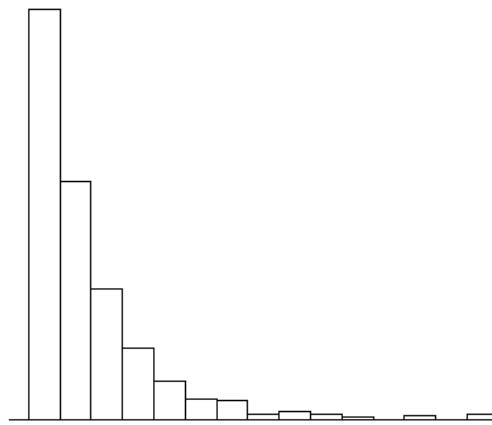


(c) Unimodal (skewed right)

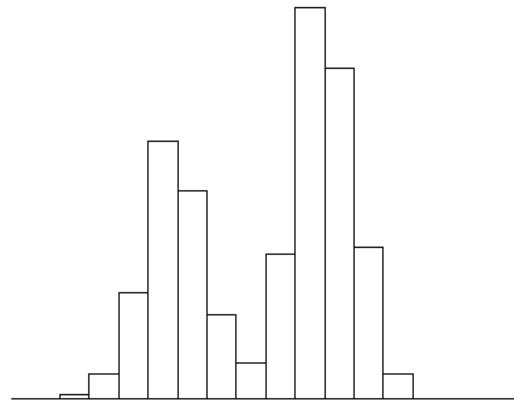
Different types of histograms



(a) Unimodal (skewed left)

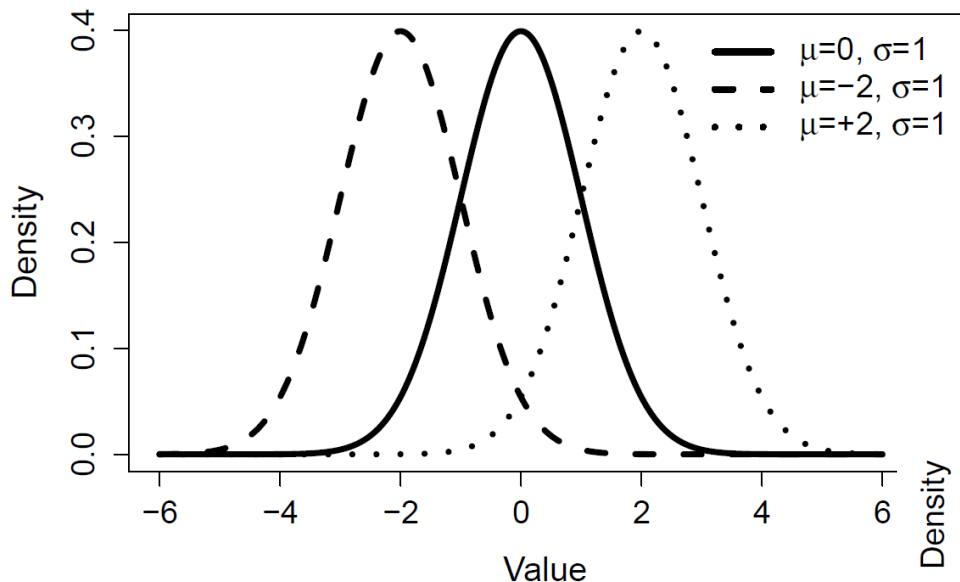


(b) Exponential



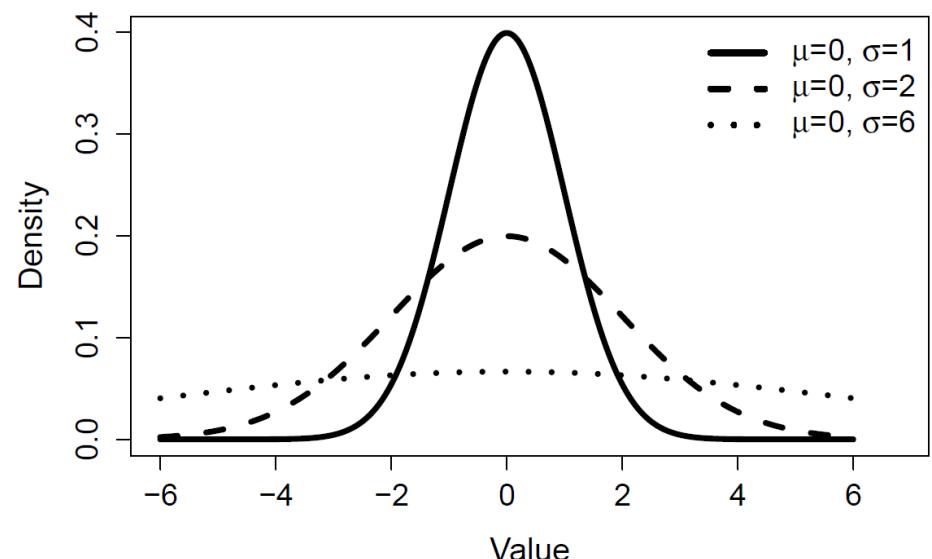
(c) Multimodal

Normal distribution

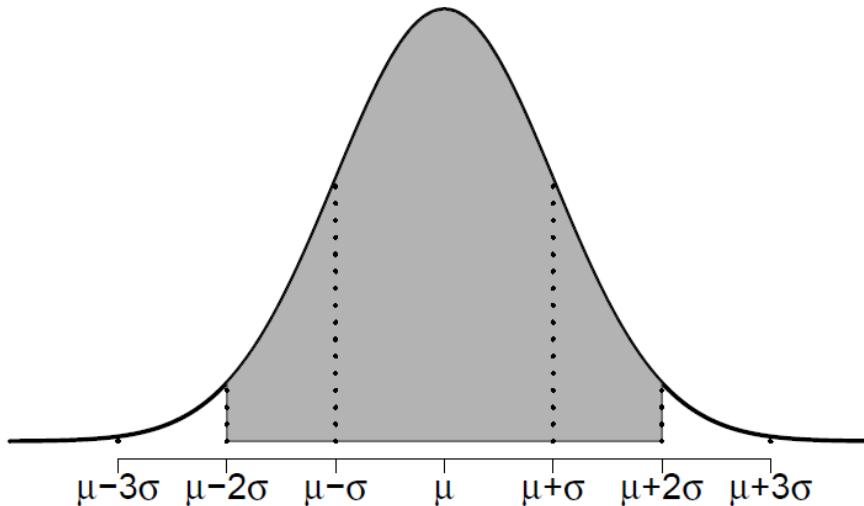


μ is the mean

σ is the standard deviation



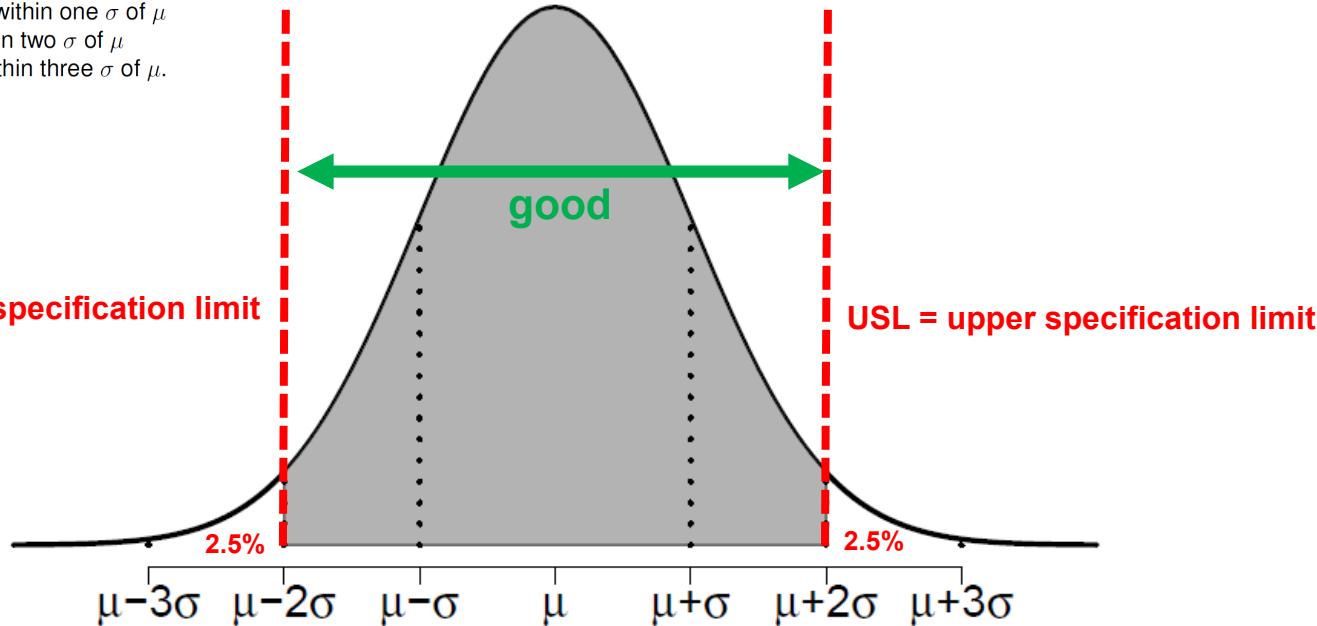
Normal distribution (68-95-99.7)



68% of the observations will be within one σ of μ
95% of observations will be within two σ of μ
99.7% of observations will be within three σ of μ .

Tolerances: How many defects?

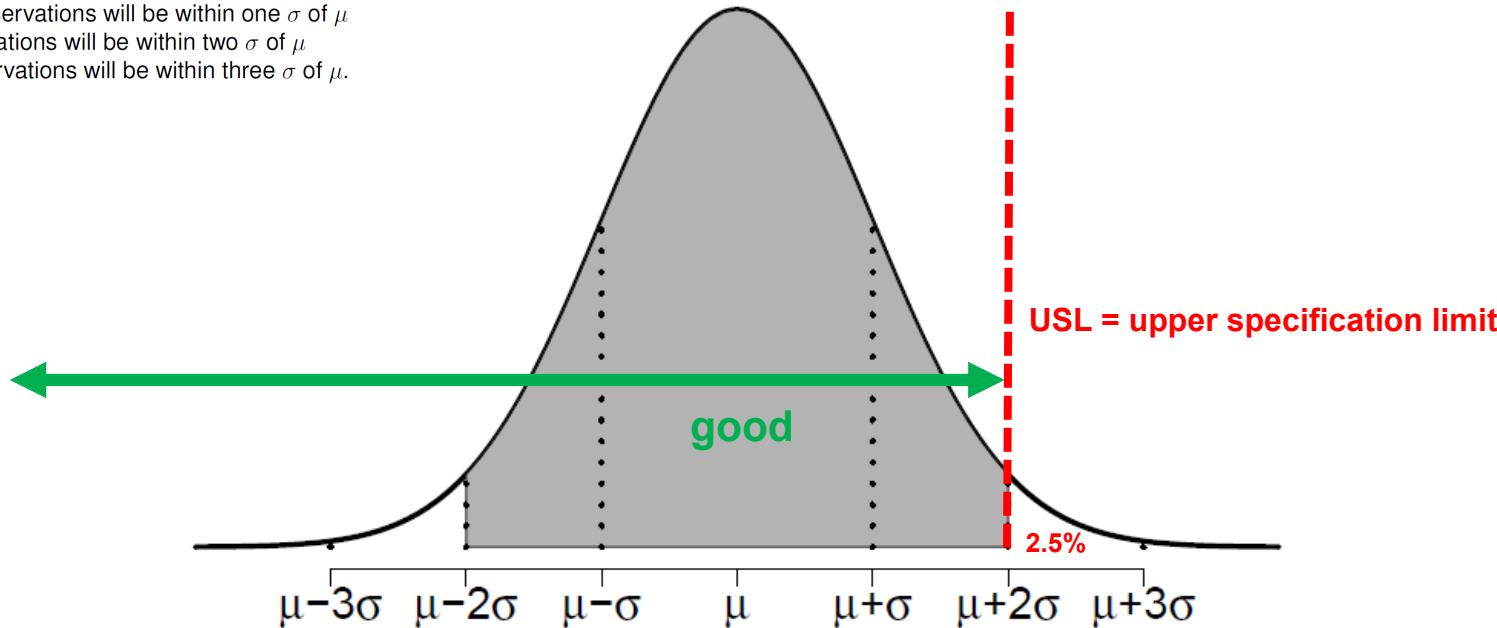
68% of the observations will be within one σ of μ
95% of observations will be within two σ of μ
99.7% of observations will be within three σ of μ .



100-95% = 5%

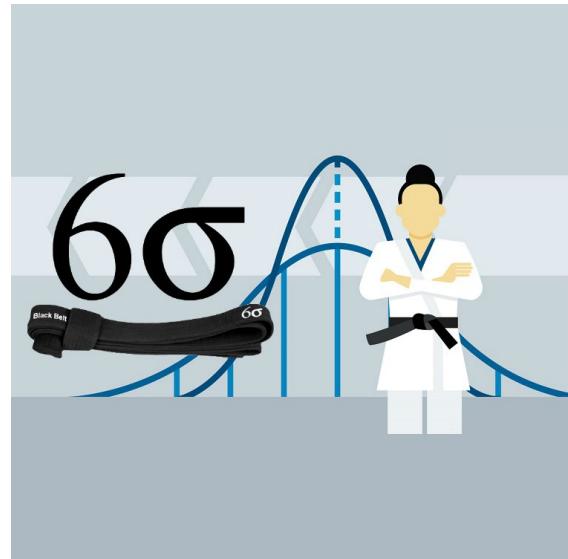
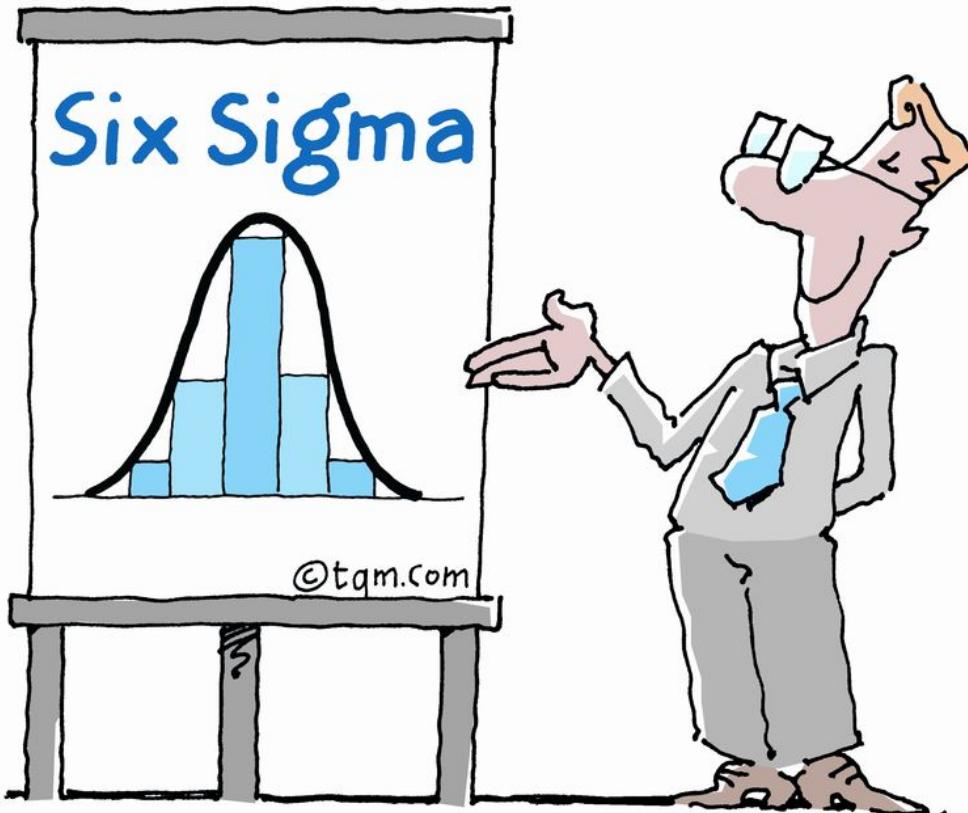
Delivery times: How many will be too late?

68% of the observations will be within one σ of μ
95% of observations will be within two σ of μ
99.7% of observations will be within three σ of μ .



$$100 - 97.5\% = 2.5\%$$

What is lean six sigma?



Six Sigma

(less than 3.4 defects/issues per million instances)

Fraction of area left of LSL:

original: 9.866×10^{-10}
+1.5 σ shift: 3.191×10^{-14}
-1.5 σ shift: 3.398×10^{-6}

Fraction of area right of USL:

original: 9.866×10^{-10}
+1.5 σ shift: 3.398×10^{-6}
-1.5 σ shift: 3.191×10^{-14}

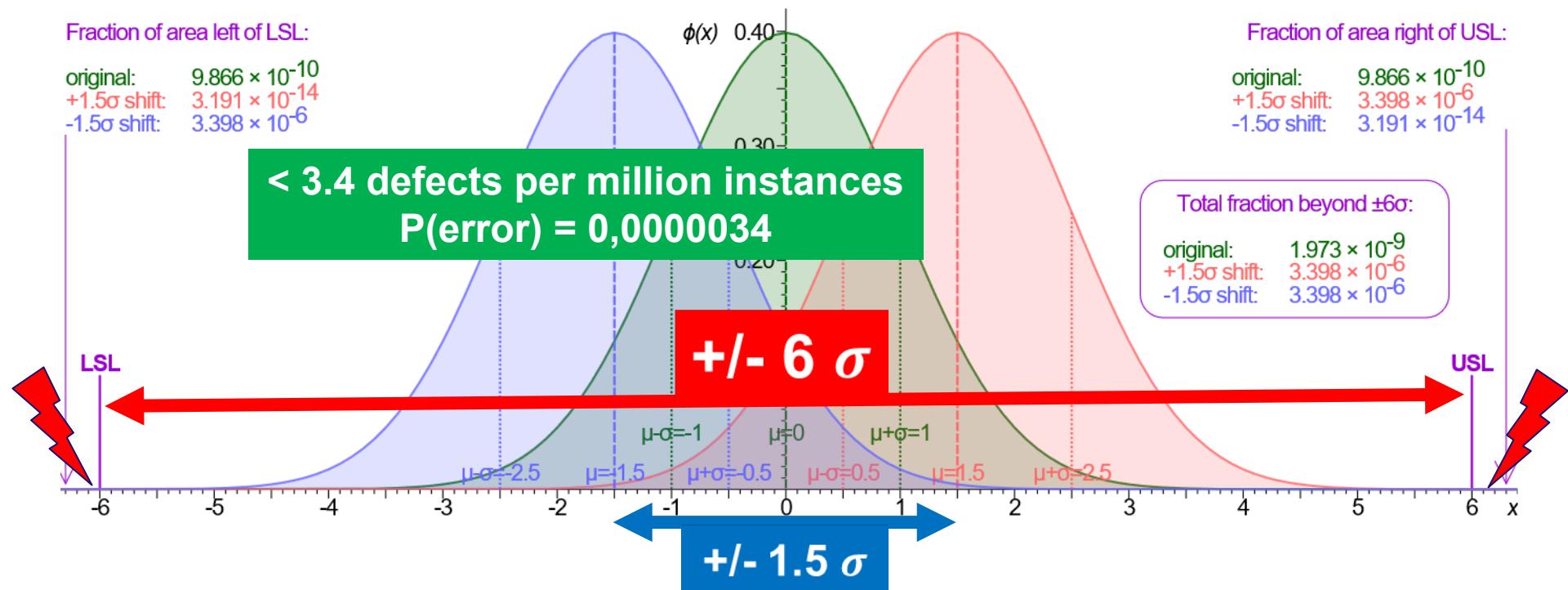
< 3.4 defects per million instances
 $P(\text{error}) = 0,0000034$

+/- 6 σ

+/- 1.5 σ

Total fraction beyond $\pm 6\sigma$:

original: 1.973×10^{-9}
+1.5 σ shift: 3.398×10^{-6}
-1.5 σ shift: 3.398×10^{-6}



Processes that operate with "six sigma quality" are assumed to have less than 3.4 defects per million cases. This is based on six sigma with a "drift" of +/- 1.5 sigma.

Data quality



Typical problems

- Data may be ...
 - **incomplete** (missing instances/attributes),
 - **invalid** (impossible values),
 - **inconsistent** (conflicting values),
 - **imprecise** (approximated or rounded), and/or
 - **outdated** (based on old observations).

Missing values

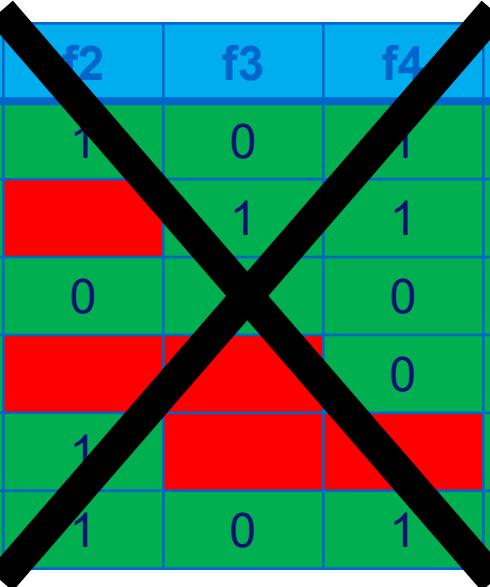
- Feature is missing for some instances.
- Options:
 1. Remove feature completely.
 2. Only consider instances that have a value (per feature).
 3. Remove all instances that have one of the features missing.
 4. Repair missing features (imputation).

Missing values

features

instances	f1	f2	f3	f4	f5	f7	f8	f9
	0	1	0	1	1	1	0	0
	0	1	1	1	0	0	0	1
	1	0	0	0	1	1	1	1
	0	1	1	0	1	0	1	1
	1	1	1	1	1	0	1	0
	1	1	0	1	1	0	0	1

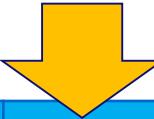
Handling missing values (1)



f1	f2	f3	f4	f5	f7	f8	f9
0	1	0	1	1	1	0	0
0	1	1	1	0	0	0	1
1	0	0	1	1	1	1	1
0	1	0	1	1	0	1	1
1	1	1	1	1	0	1	0
1	1	0	1	1	0	0	1

Remove features completely

Handling missing values (2)



f1	f2	f3	f4	f5	f7	f8	f9
0	1	0	1	1	1	0	0
0	1	0	1	1	1	0	0
1	0	0	0	1	1	1	1
0	1	0	1	1	1	0	0
1	1	1	1	1	0	1	0
1	1	0	1	1	0	0	1

Only consider instances that
have a value (per feature)

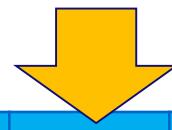
Handling missing values (2)



f1	f2	f3	f4	f5	f7	f8	f9
0	1	0	1	1	1	0	0
0		1	1	0	0	0	1
1	0	0	0	1	1	1	1
0			0	1	0	1	1
1				1	0	1	0
1	1	0	1	1	0	0	1

Only consider instances that
have a value (per feature)

Handling missing values (2)



f1	f2	f3	f4	f5	f7	f8	f9
0	1	0	1	1	1	0	0
0		1	1	0	0	0	1
1	0	0	0	1	1	1	1
0			0	1	0	1	1
1	1			1	0	1	0
1	1	0	1	1	0	0	1

Only consider instances that
have a value (per feature)

Handling missing values (3)



f1	f2	f3	f4	f5	f7	f8	f9
0	1	0	1	1	1	0	0
0	1	0	1	1	1	0	0
1	0	0	0	1	1	1	1
0	1	0	1	1	1	1	1
1	1	0	1	1	0	1	0
1	1	0	1	1	0	0	1

**Remove all instances that have
one of the features missing**

Handling missing values (3)



f1	f2	f3	f4	f5	f7	f8	f9
0	1	0	1	1	1	0	0
0	1	0	1	1	1	0	0
1	0	0	0	1	1	1	1
0	1	0	1	1	1	1	1
0	1	0	1	1	1	1	1
1	1	0	1	1	0	1	1
1	1	0	1	1	0	0	1

**Remove all instances that have
one of the features missing**

Handling missing values (4)

f1	f2	f3	f4	f5	f7	f8	f9
0	1	0	1	1	1	0	0
0	1	1	1	0	0	0	1
1	0	0	0	1	1	1	1
0	1	0	0	1	0	1	1
1	1	0	1	1	0	1	0
1	1	0	1	1	0	0	1

Repair missing features (imputation)

Impossible values

- Date: 30-2-2018
- Date: 13-13-2018
- Time: 23:61
- Color: Bllue
- Members: 6.5

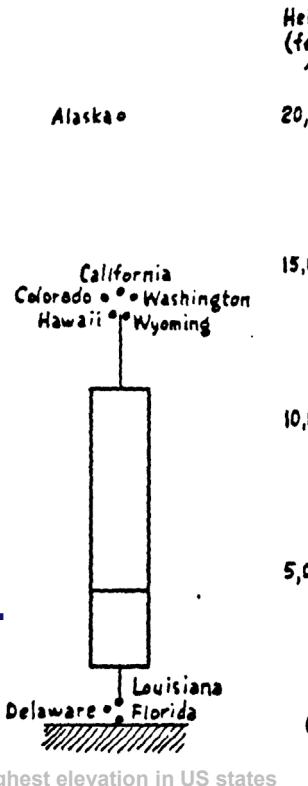
Can be handled like missing values.

Unlikely values

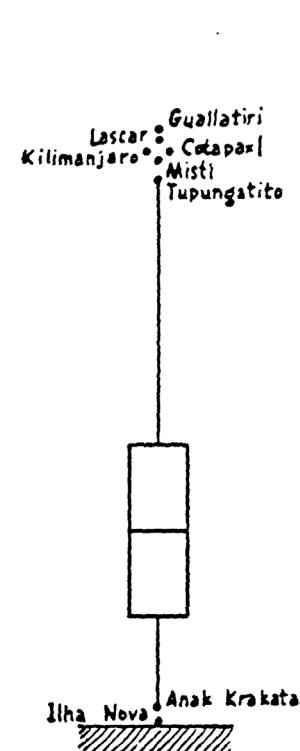
- Dates: One may expect a uniform distribution over months and days.
 - When formats DD-MM-YYYY and MM-DD-YYYY are mixed up this is no longer the case.
 - Days 1-12 are more frequent than 13-31
- Age: An age of 123 is not impossible, but unlikely.
- Price: An item priced 120.000 (rather than 120).
- Unlikely values are identified based on domain knowledge.
- Outlier values are identified based on the distribution.

Box plots

A) HEIGHTS of 50 STATES



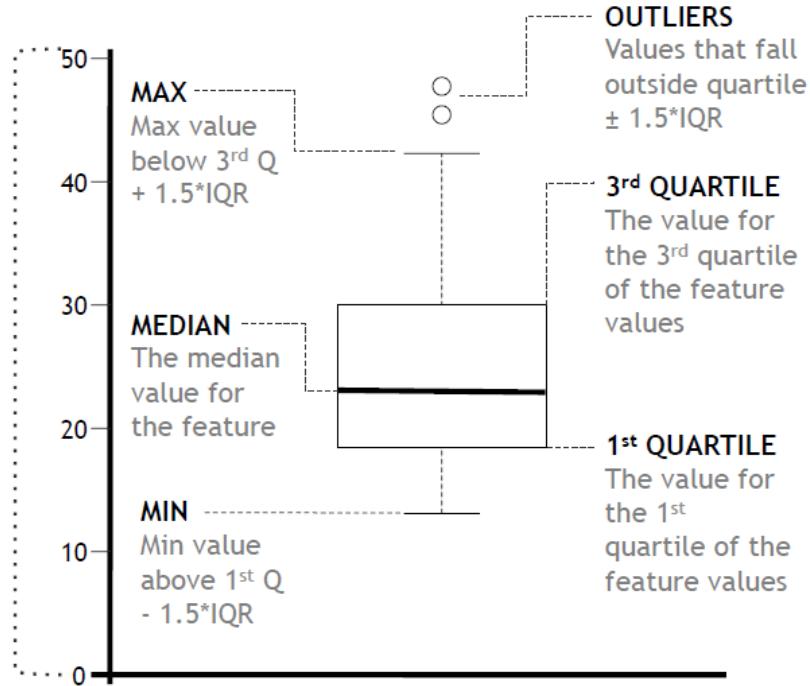
B) HEIGHTS of 219 VOLCANOS



First introduced by John Tukey
(1915-2000) in the book
Exploratory Data Analysis in 1977.

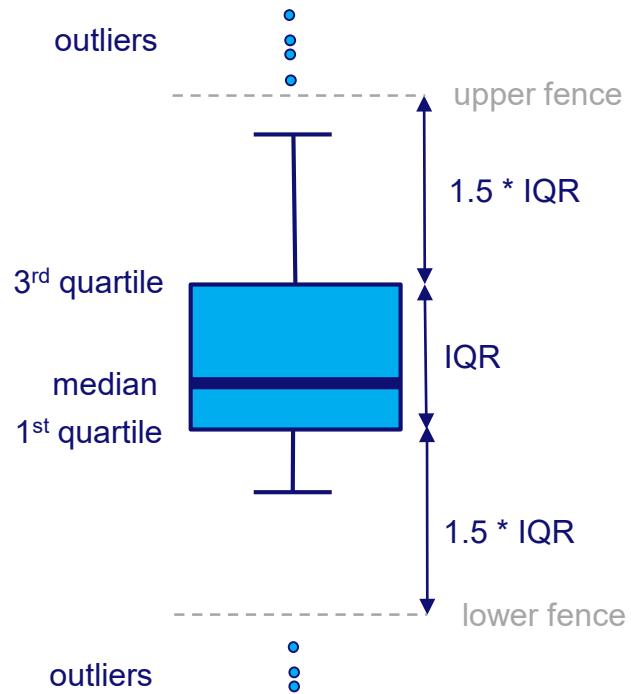
Box plots

FEATURE VALUES
Values displayed
for a single
feature



- Median value (middle) depicted by “Bar”
- IQR = Interquartile Range (covers 50% of “middle” instances) depicted by “Box”.
- Upper whisker: maximal value below 3rd quartile + 1.5 * IQR.
- Lower whisker: minimal value above 1st quartile - 1.5 * IQR.
- Outliers are drawn separately.

Box plots



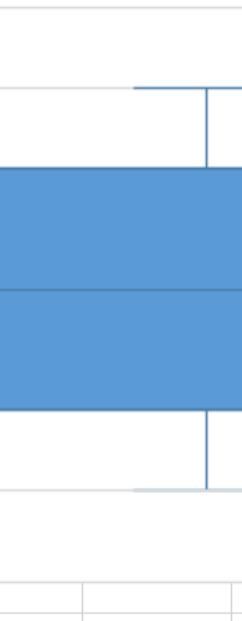
- Median value (middle) depicted by “Bar”
- IQR = Interquartile Range (covers 50% of “middle” instances) depicted by “Box”.
- Upper whisker: maximal value below 3rd quartile + 1.5 * IQR.
- Lower whisker: minimal value above 1st quartile - 1.5 * IQR.
- Outliers are drawn separately.

Box plot (Excel)

data

0
1
2
3
4
5
6
7
8
9
10

evenly distributed



$$\text{median} = 5$$

$$\text{IQR} = 8 - 2 = 6$$

$$\text{upper fence} = 8 + 1.5 \cdot 6 = 17$$

$$\text{upper whisker} = 10$$

$$\text{lower fence} = 2 - 1.5 \cdot 6 = -7$$

$$\text{lower whisker} = 0$$

no outliers

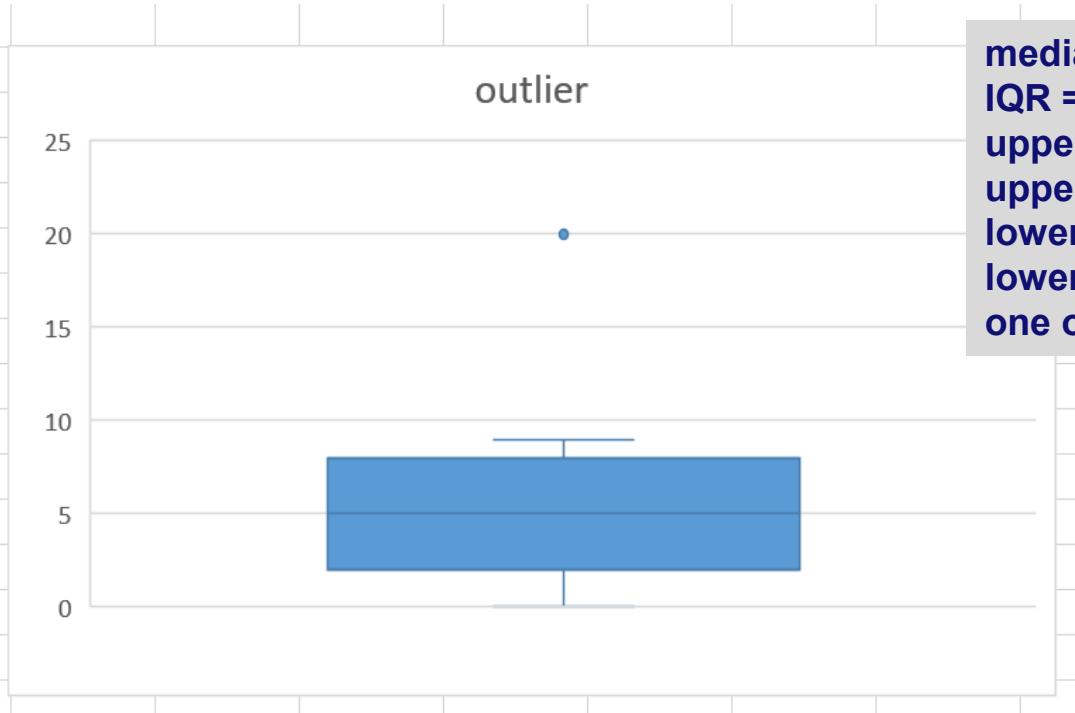


Chair of Process
and Data Science

Box plot (Excel)

data

0
1
2
3
4
5
6
7
8
9
20



$$\text{median} = 5$$

$$\text{IQR} = 8 - 2 = 6$$

$$\text{upper fence} = 8 + 1.5 \cdot 6 = 17$$

$$\text{upper whisker} = 9$$

$$\text{lower fence} = 2 - 1.5 \cdot 6 = -7$$

$$\text{lower whisker} = 0$$

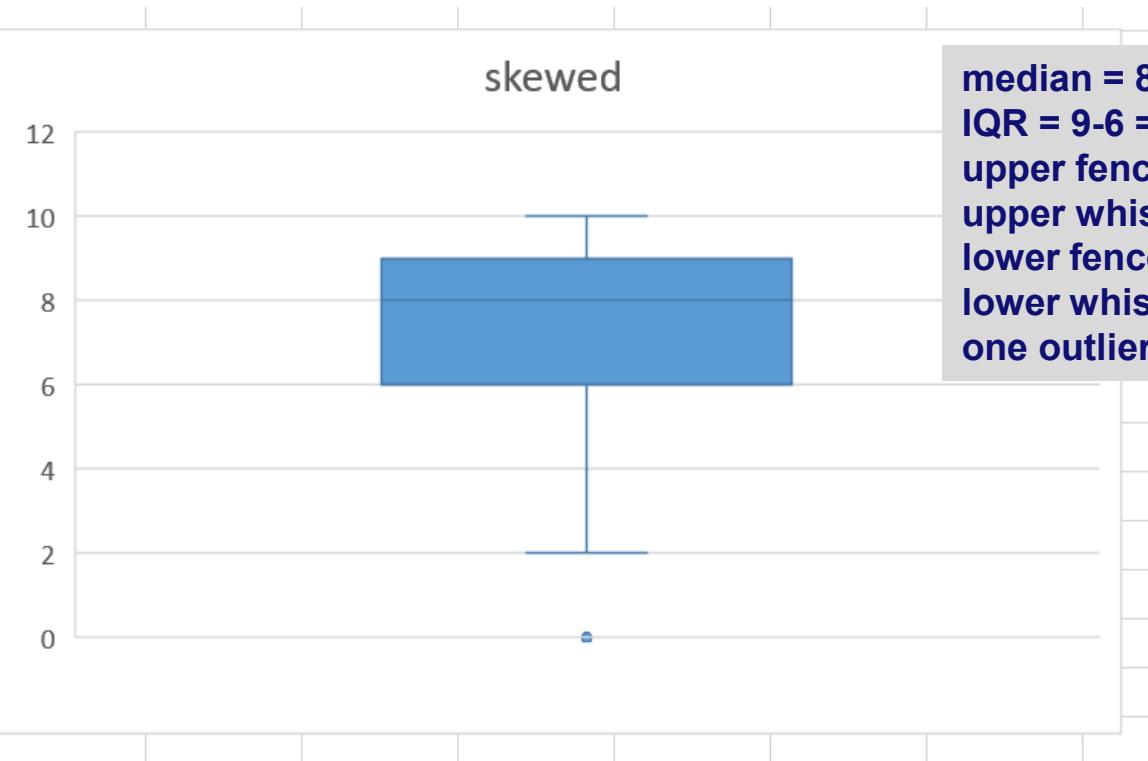
one outlier



Box plot (Excel)

data

0
2
6
6
8
8
8
8
9
9
10



median = 8

IQR = 9-6 = 3

upper fence = $9 + 1.5 \cdot 3 = 13.5$

upper whisker = 10

lower fence = $6 - 1.5 \cdot 3 = 1.5$

lower whisker = 2

one outlier

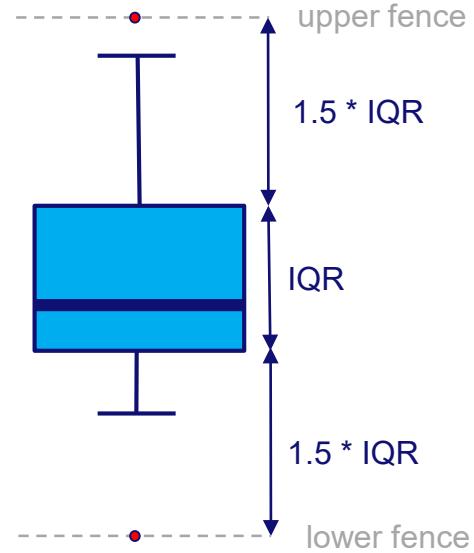
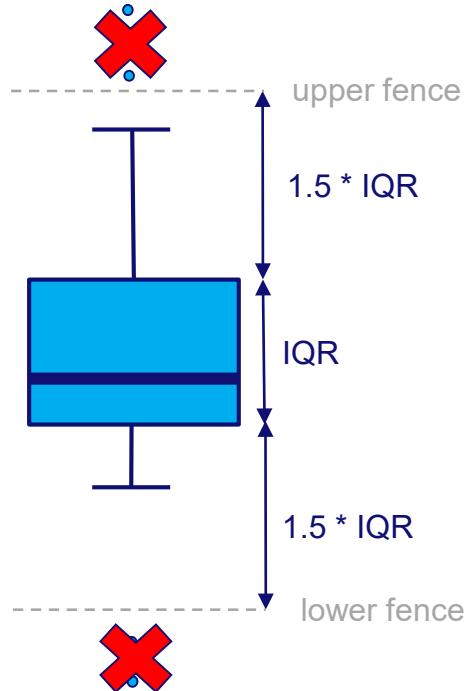
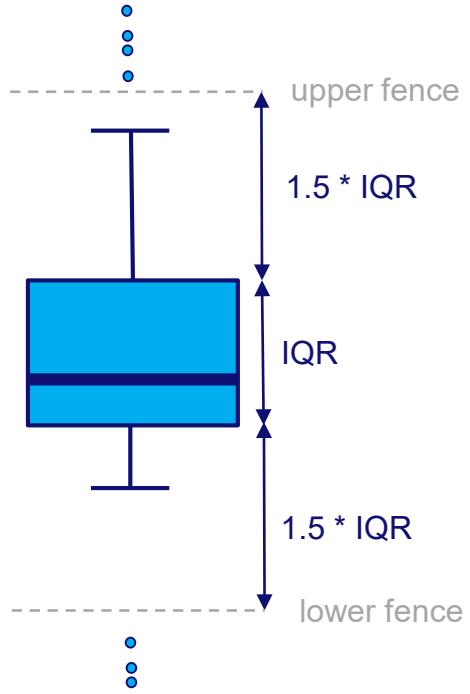


Chair of Process
and Data Science

Handling outliers

- 1. Remove values above and below thresholds (e.g., upper and lower fences).**
- 2. Clamp values above and below thresholds to these thresholds.**

Handling outliers

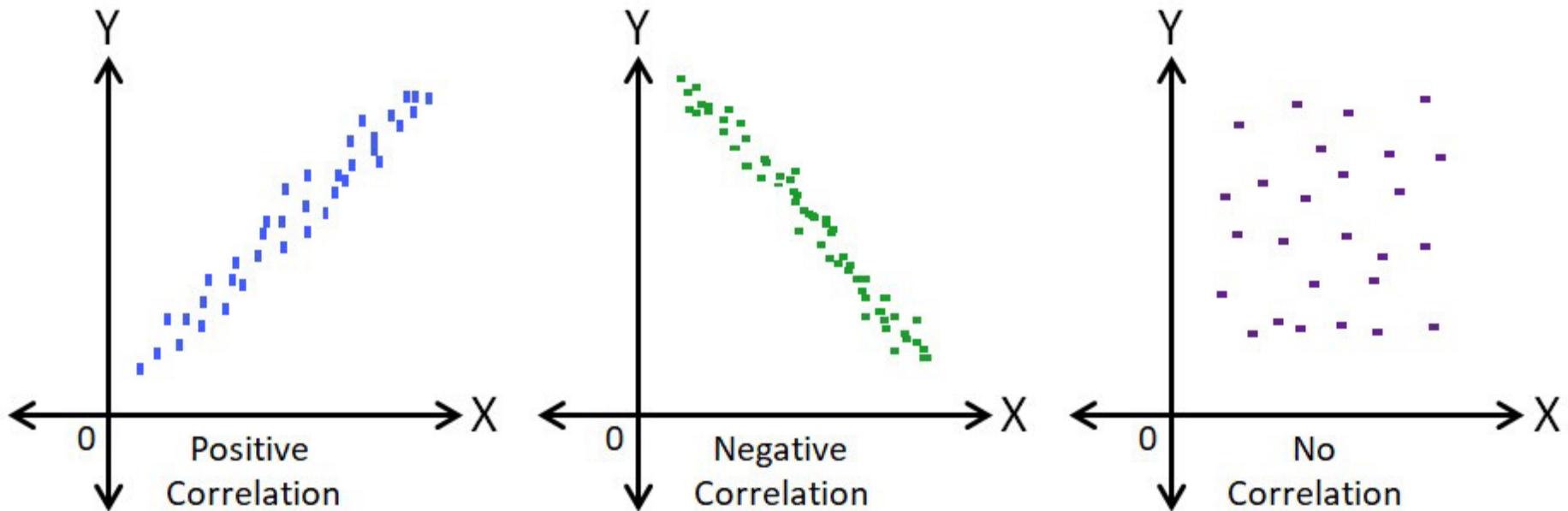


Showing relations among features

Let's focus on two features

f1	f2	f3	X=f4	f5	Y=f7	f8	f9

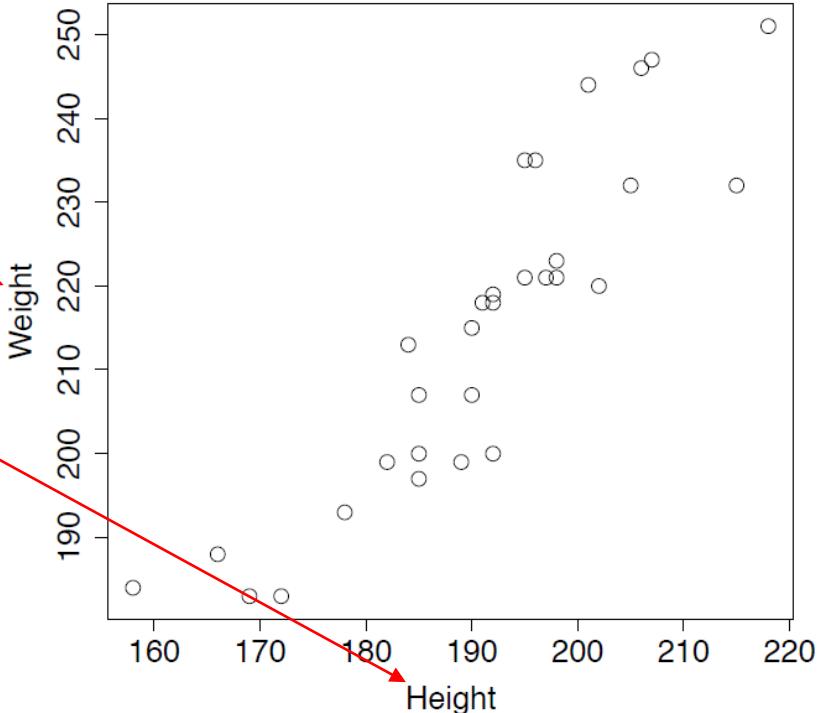
Scatter Plots & Correlation



Data set in book (1/3)

ID	POSITION	HEIGHT	WEIGHT	CAREER STAGE	AGE	SPONSORSHIP EARNINGS	SHOE SPONSOR
1	forward	192	218	veteran	29	561	yes
2	center	218	251	mid-career	35	60	no
3	forward	197	221	rookie	22	1,312	no
4	forward	192	219	rookie	22	1,359	no
5	forward	198	223	veteran	29	362	yes
6	guard	166	188	rookie	21	1,536	yes
7	forward	195	221	veteran	25	694	no
8	guard	182	199	rookie	21	1,678	yes
9	guard	189	199	mid-career	27	385	yes
10	forward	205	232	rookie	24	1,416	no
11	center	206	246	mid-career	29	314	no
12	guard	185	207	rookie	23	1,497	yes
13	guard	172	183	rookie	24	1,383	yes
14	guard	169	183	rookie	24	1,034	yes
15	guard	185	197	mid-career	29	178	yes
16	forward	215	232	mid-career	30	434	no
17	guard	158	184	veteran	29	162	yes
18	guard	190	207	mid-career	27	648	yes
19	center	195	235	mid-career	28	481	no
20	guard	192	200	mid-career	32	427	yes
21	forward	202	220	mid-career	31	542	no
22	forward	184	213	mid-career	32	12	no
23	forward	190	215	rookie	22	1,179	no
24	guard	178	193	rookie	21	1,078	no
25	guard	185	200	mid-career	31	213	yes
26	forward	191	218	rookie	19	1,855	no
27	center	196	235	veteran	32	47	no
28	forward	198	221	rookie	22	1,409	no
29	center	207	247	veteran	27	1,065	no
30	center	201	244	mid-career	25	1,111	yes

basketball players



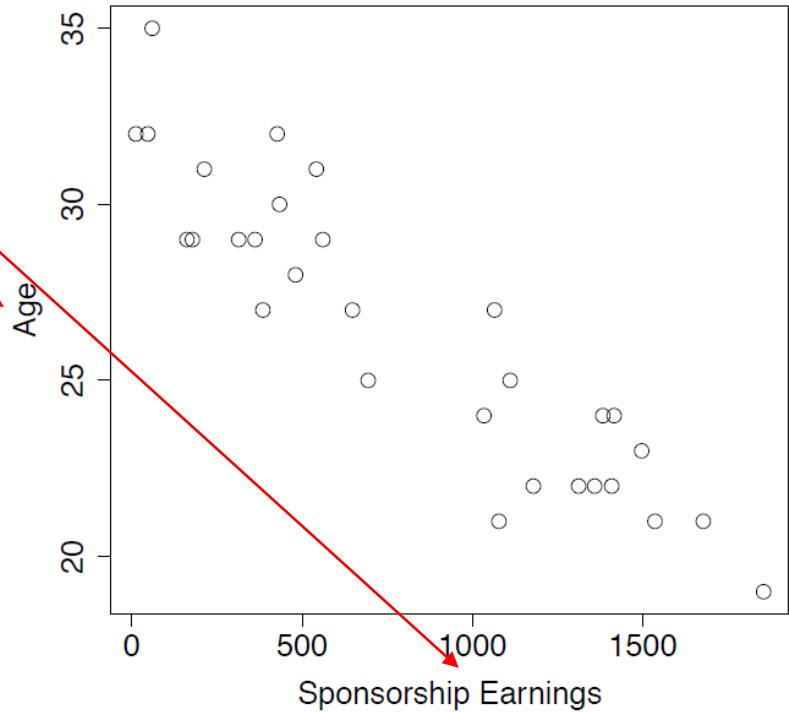
Tables and charts taken from Fundamentals of Machine Learning for Predictive Data Analytics by J. Kelleher, B. Mac Namee and A. D'Arcy.



Chair of Process and Data Science

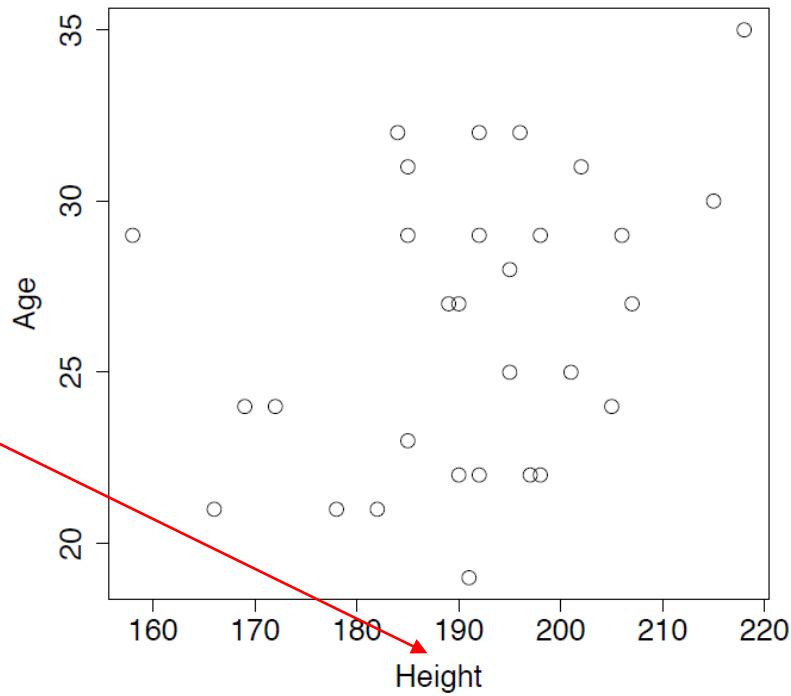
Data set in book (2/3)

ID	POSITION	HEIGHT	WEIGHT	CAREER STAGE	AGE	SPONSORSHIP EARNINGS	SHOE SPONSOR
1	forward	192	218	veteran	29	561	yes
2	center	218	251	mid-career	35	60	no
3	forward	197	221	rookie	22	1,312	no
4	forward	192	219	rookie	22	1,359	no
5	forward	198	223	veteran	29	362	yes
6	guard	166	188	rookie	21	1,536	yes
7	forward	195	221	veteran	25	694	no
8	guard	182	199	rookie	21	1,678	yes
9	guard	189	199	mid-career	27	385	yes
10	forward	205	232	rookie	24	1,416	no
11	center	206	246	mid-career	29	314	no
12	guard	185	207	rookie	23	1,497	yes
13	guard	172	183	rookie	24	1,383	yes
14	guard	169	183	rookie	24	1,034	yes
15	guard	185	197	mid-career	29	178	yes
16	forward	215	232	mid-career	30	434	no
17	guard	158	184	veteran	29	162	yes
18	guard	190	207	mid-career	27	648	yes
19	center	195	235	mid-career	28	481	no
20	guard	192	200	mid-career	32	427	yes
21	forward	202	220	mid-career	31	542	no
22	forward	184	213	mid-career	32	12	no
23	forward	190	215	rookie	22	1,179	no
24	guard	178	193	rookie	21	1,078	no
25	guard	185	200	mid-career	31	213	yes
26	forward	191	218	rookie	19	1,855	no
27	center	196	235	veteran	32	47	no
28	forward	198	221	rookie	22	1,409	no
29	center	207	247	veteran	27	1,065	no
30	center	201	244	mid-career	25	1,111	yes

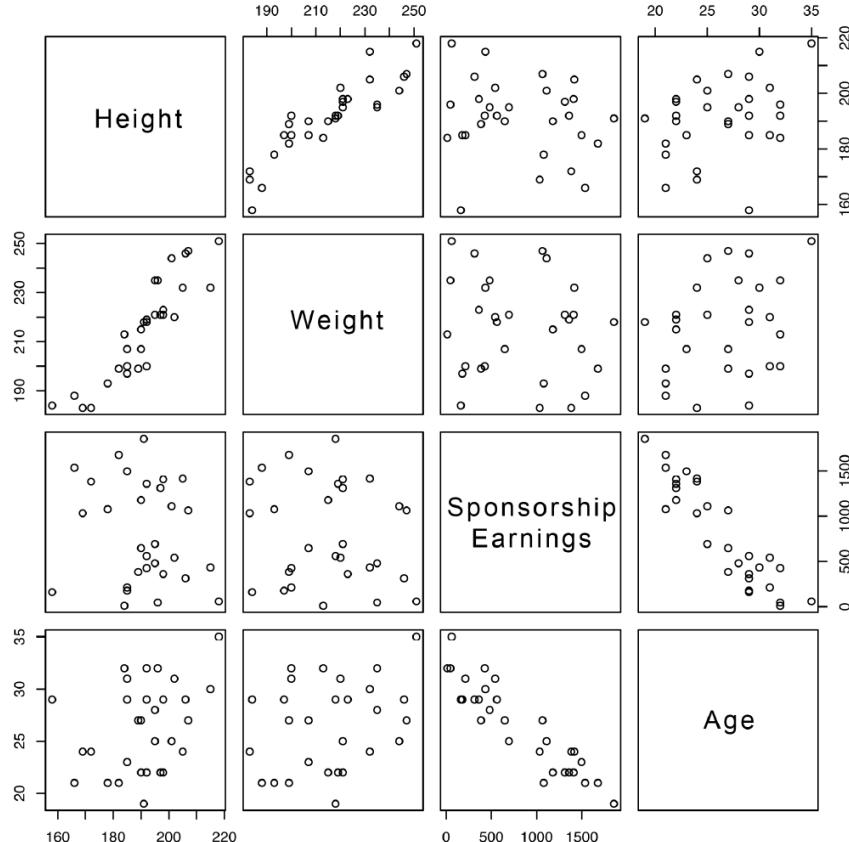


Data set in book (3/3)

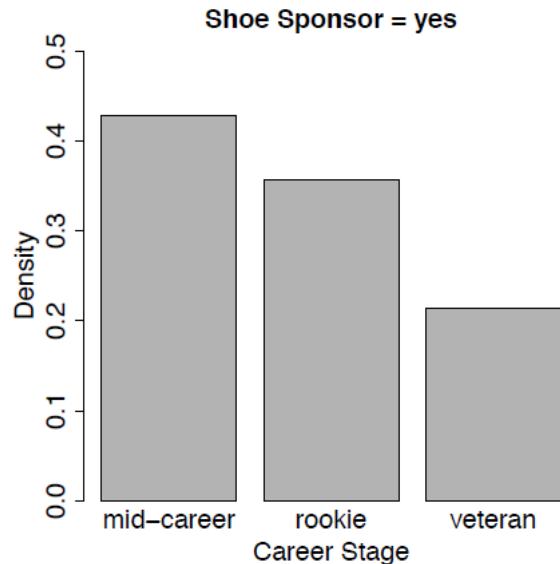
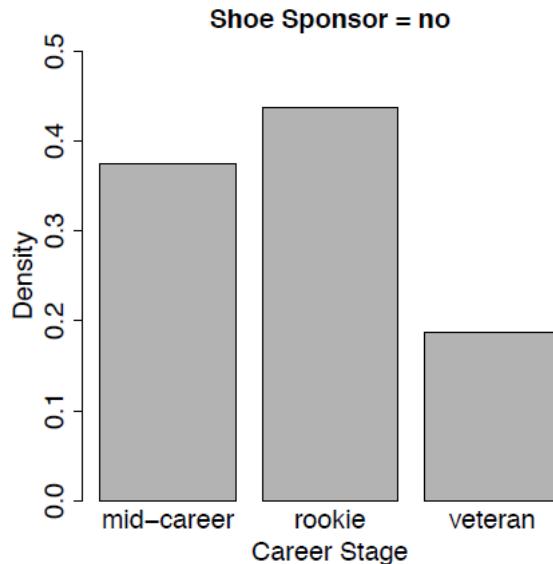
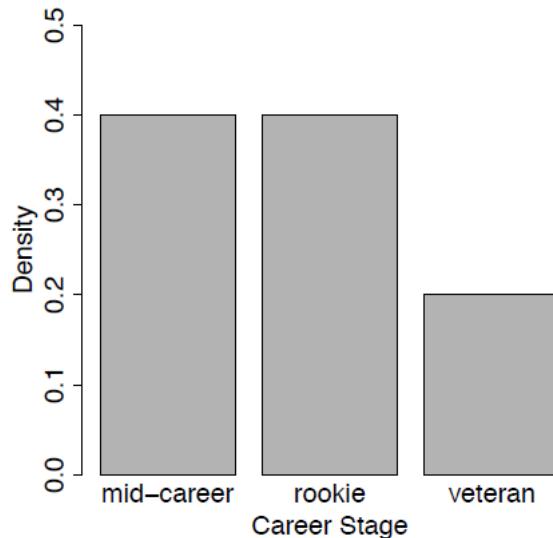
ID	POSITION	HEIGHT	WEIGHT	CAREER STAGE	AGE	SPONSORSHIP EARNINGS	SHOE SPONSOR
1	forward	192	218	veteran	29	561	yes
2	center	218	251	mid-career	35	60	no
3	forward	197	221	rookie	22	1,312	no
4	forward	192	219	rookie	22	1,359	no
5	forward	198	223	veteran	29	362	yes
6	guard	166	188	rookie	21	1,536	yes
7	forward	195	221	veteran	25	694	no
8	guard	182	199	rookie	21	1,678	yes
9	guard	189	199	mid-career	27	385	yes
10	forward	205	232	rookie	24	1,416	no
11	center	206	246	mid-career	29	314	no
12	guard	185	207	rookie	23	1,497	yes
13	guard	172	183	rookie	24	1,383	yes
14	guard	169	183	rookie	24	1,034	yes
15	guard	185	197	mid-career	29	178	yes
16	forward	215	232	mid-career	30	434	no
17	guard	158	184	veteran	29	162	yes
18	guard	190	207	mid-career	27	648	yes
19	center	195	235	mid-career	28	481	no
20	guard	192	200	mid-career	32	427	yes
21	forward	202	220	mid-career	31	542	no
22	forward	184	213	mid-career	32	12	no
23	forward	190	215	rookie	22	1,179	no
24	guard	178	193	rookie	21	1,078	no
25	guard	185	200	mid-career	31	213	yes
26	forward	191	218	rookie	19	1,855	no
27	center	196	235	veteran	32	47	no
28	forward	198	221	rookie	22	1,409	no
29	center	207	247	veteran	27	1,065	no
30	center	201	244	mid-career	25	1,111	yes



Scatter Plot Matrix (SPLOM)

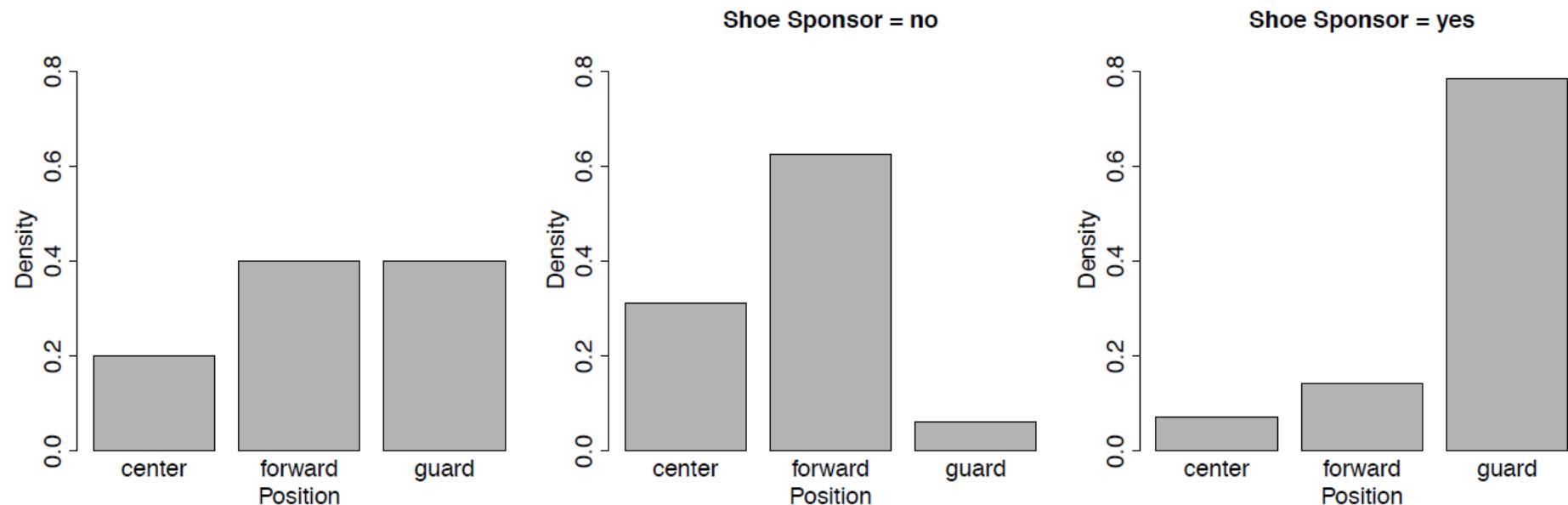


Collection of small multiple bar plots



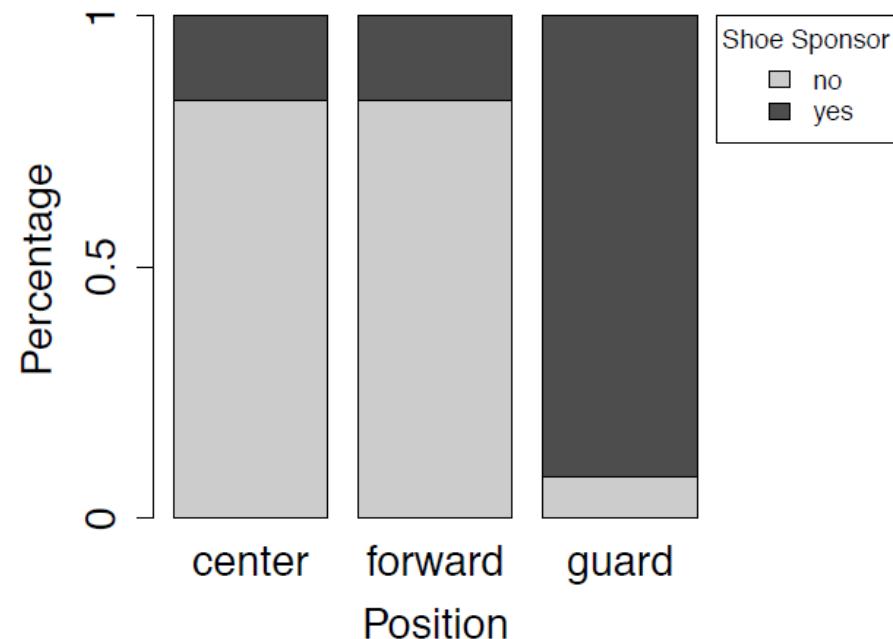
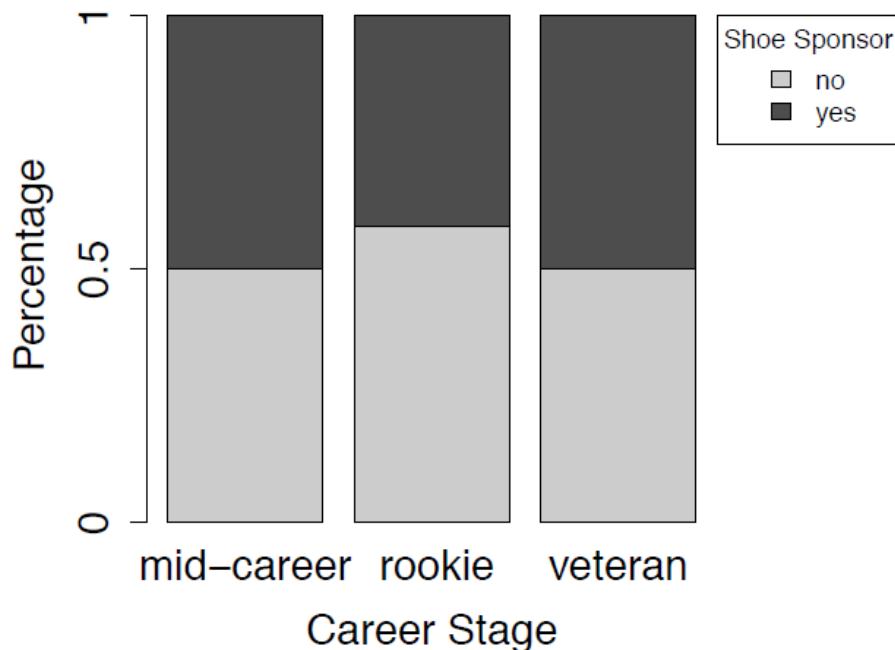
No strong relation between shoe sponsor and career stage.

Collection of small multiple bar plots



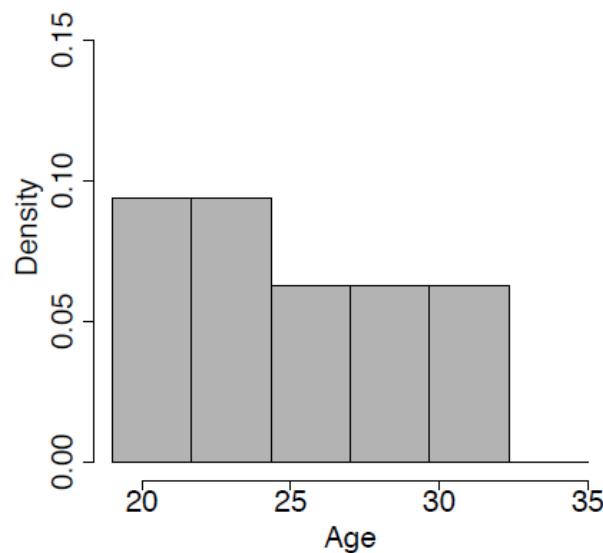
**Strong relation between shoe sponsor and position
(guards are more likely to have a shoe sponsor).**

Stacked bar plot (showing percentages)

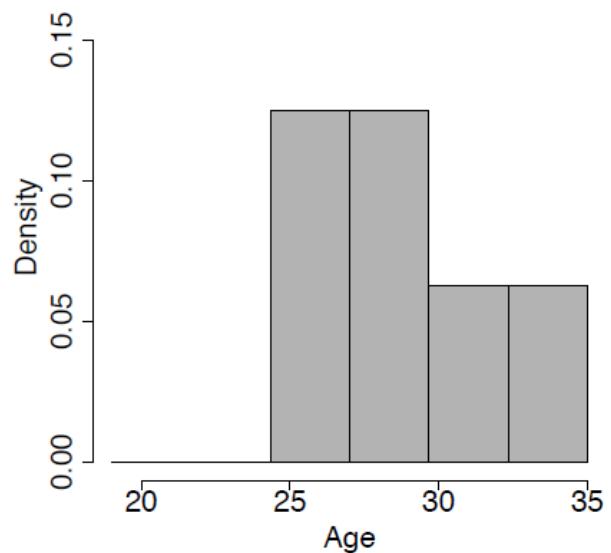


Collection of small multiple histograms

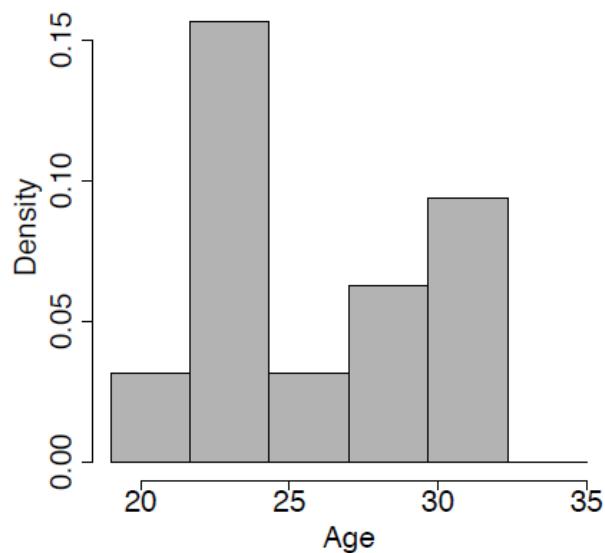
Position = guard



Position = center



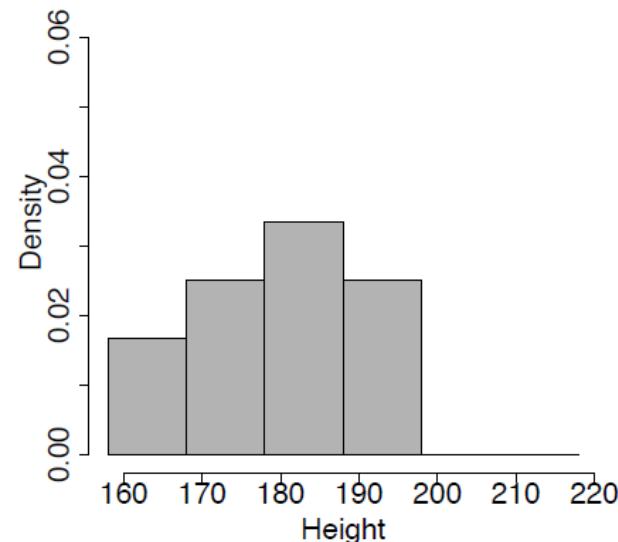
Position = forward



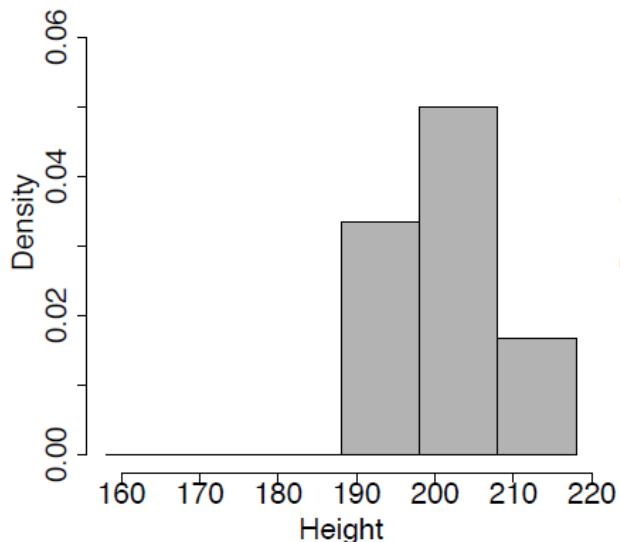
Age has now 6 classes.

Collection of small multiple histograms

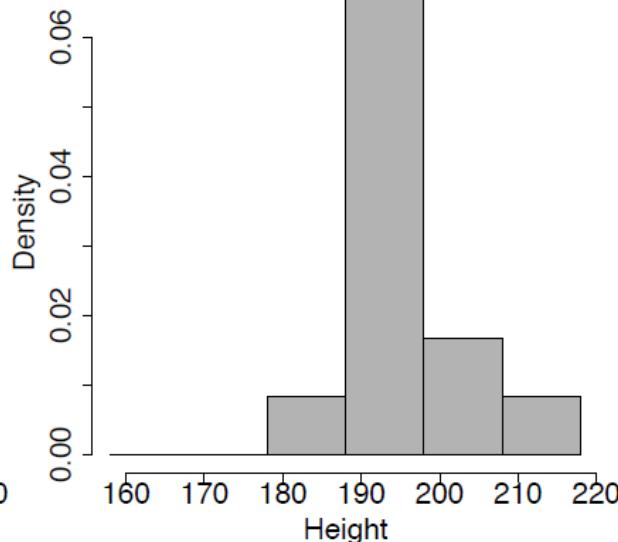
Position = guard



Position = center

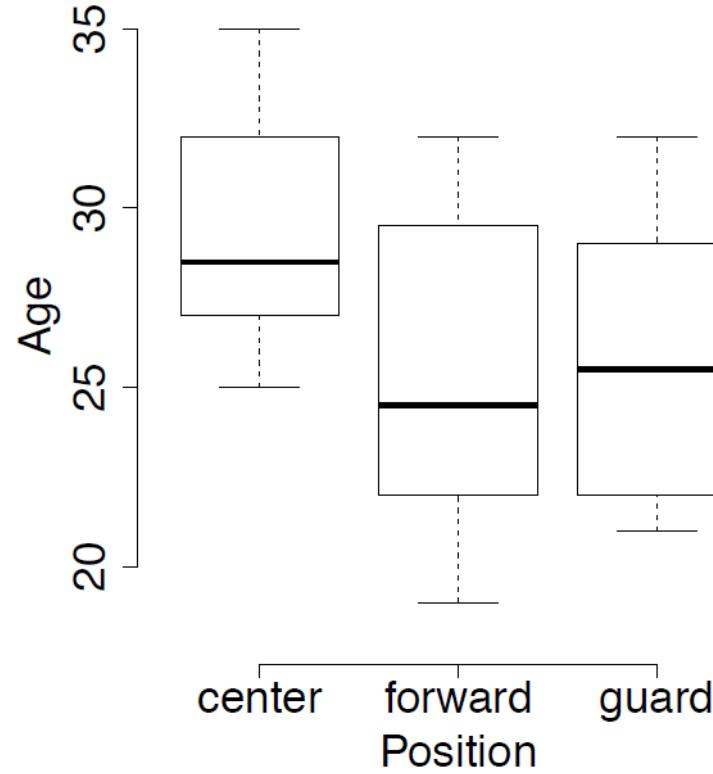
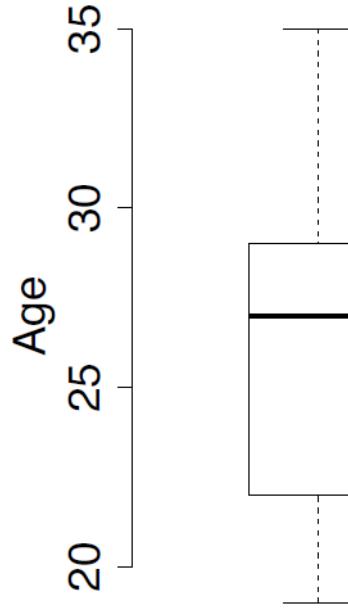


Position = forward

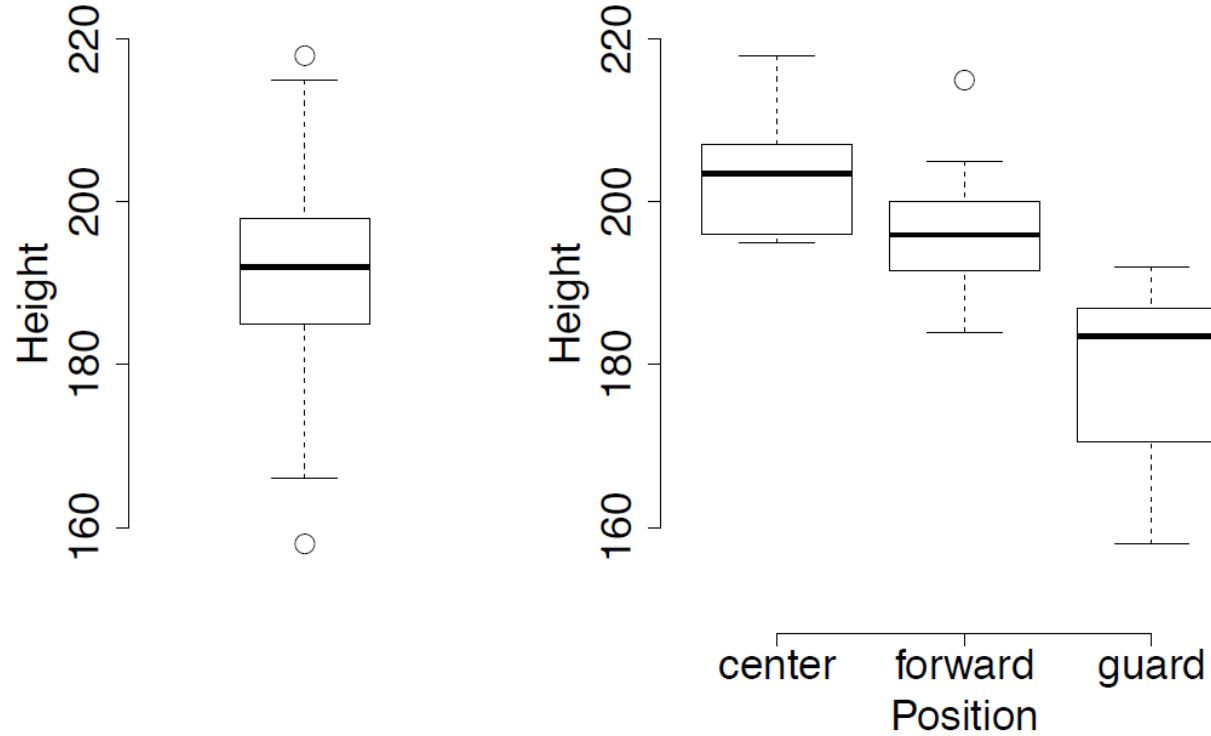


Height has now 6 classes.

Collections of box plots

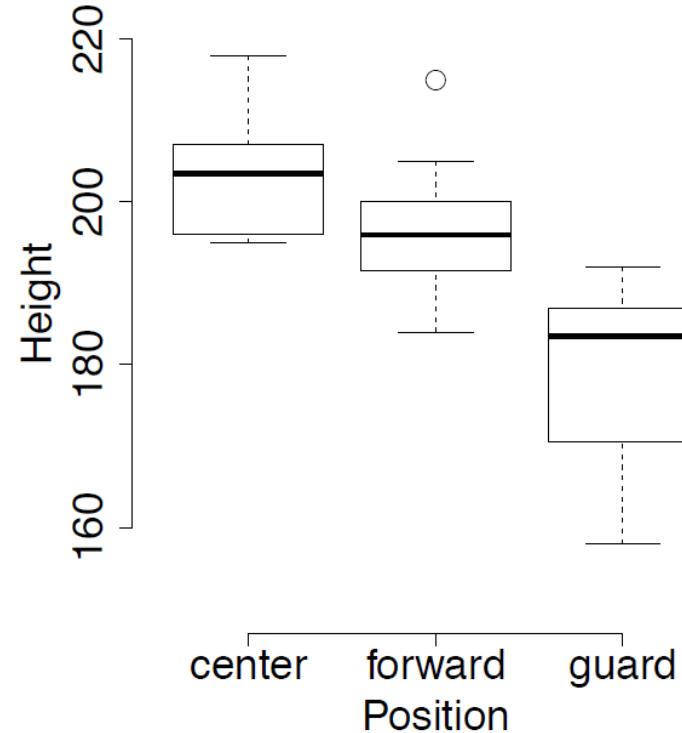
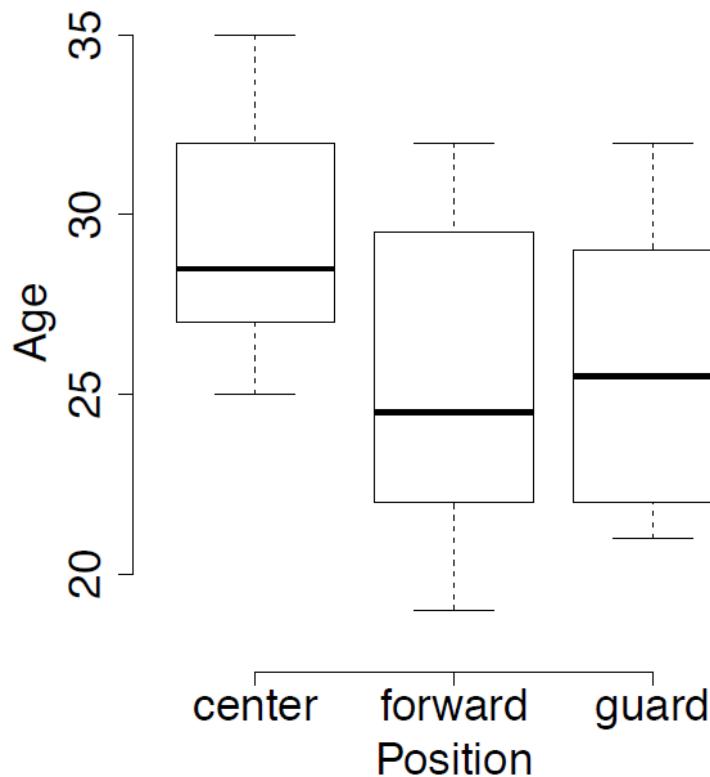


Collections of box plots



Strong relation between height and position (centers are taller than guards).

What is most significant ?



Some basic descriptive statistics

$$\bar{a} = \frac{1}{n} \sum_{i=1}^n a_i$$

sample mean

$$var(a) = \frac{\sum_{i=1}^n (a_i - \bar{a})^2}{n - 1}$$

sample variance

(based on n values a_1, \dots, a_n)



Chair of Process
and Data Science

Some basic descriptive statistics

$$sd(a) = \sqrt{var(a)}$$

standard deviation

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

(based on n values a_1, \dots, a_n)

$$\bar{a} = \frac{1}{n} \sum_{i=1}^n a_i$$



Chair of Process
and Data Science

Sample covariance

$$\text{cov}(a, b) = \frac{1}{n-1} \sum_{i=1}^n ((a_i - \bar{a}) \times (b_i - \bar{b}))$$

$[-\infty, \infty]$

+ & + \Rightarrow +
+ & - \Rightarrow -
- & + \Rightarrow -
- & - \Rightarrow +

(based on n pairs of values $(a_1, b_1), \dots, (a_n, b_n)$)



Chair of Process
and Data Science

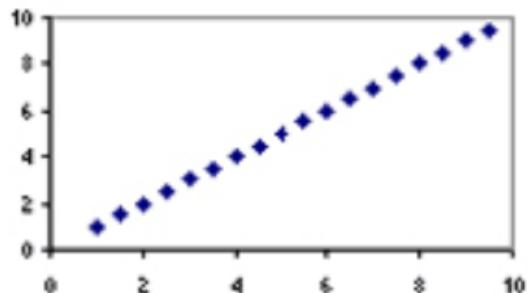
Correlation

$$\text{corr}(a, b) = \frac{\text{cov}(a, b)}{\text{sd}(a) \times \text{sd}(b)} \quad [-1, 1]$$

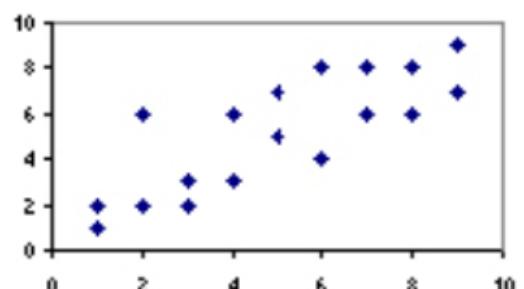
> 0 positive
< 0 negative
≈ 0 independent

$$\begin{aligned}\text{sd}(a) &= \sqrt{\text{var}(a)} \\ &= \sqrt{\frac{\sum_{i=1}^n (a_i - \bar{a})^2}{n-1}}\end{aligned}$$

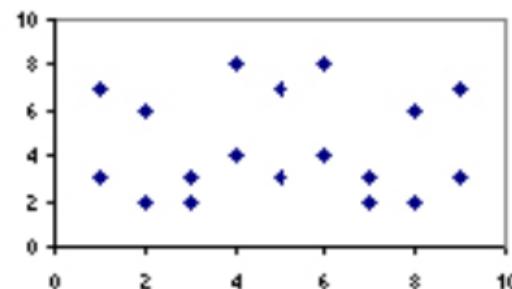
$$\text{cov}(a, b) = \frac{1}{n-1} \sum_{i=1}^n ((a_i - \bar{a}) \times (b_i - \bar{b}))$$



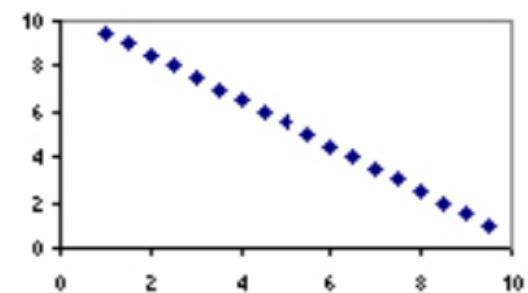
Maximum positive correlation
($r = 1.0$)



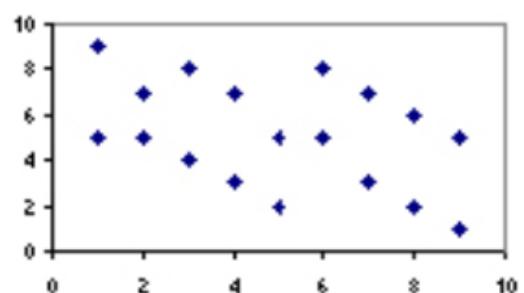
Strong positive correlation
($r = 0.80$)



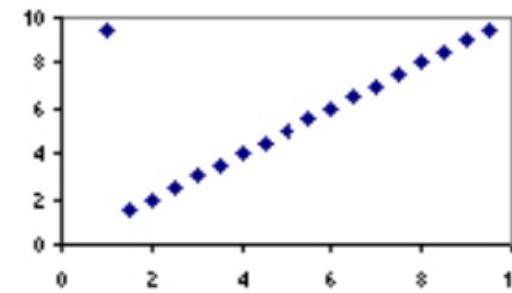
Zero correlation
($r = 0$)



Maximum negative correlation
($r = -1.0$)

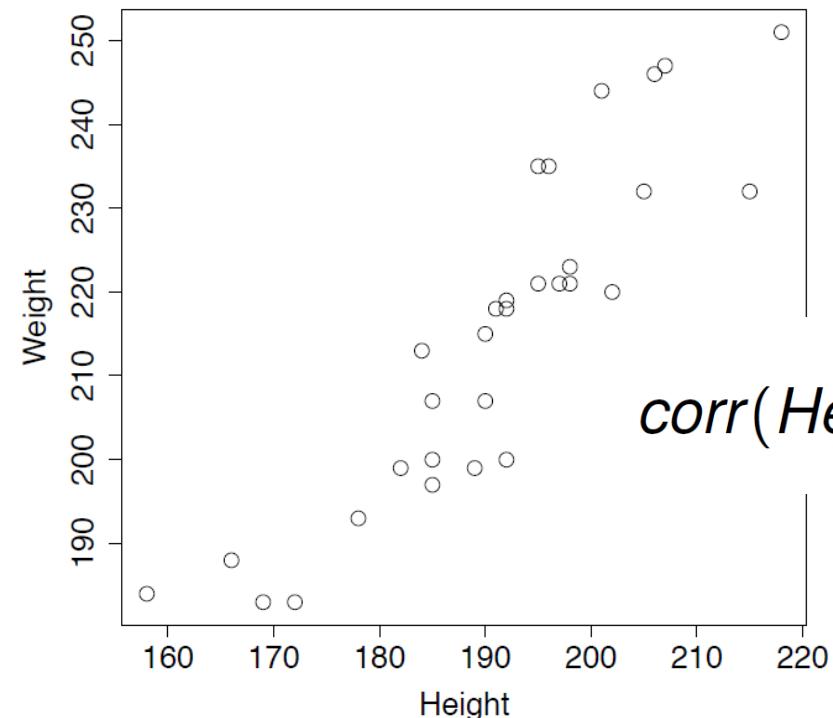


Moderate negative correlation
($r = -0.43$)



Strong correlation & outlier
($r = 0.71$)

Positive correlation



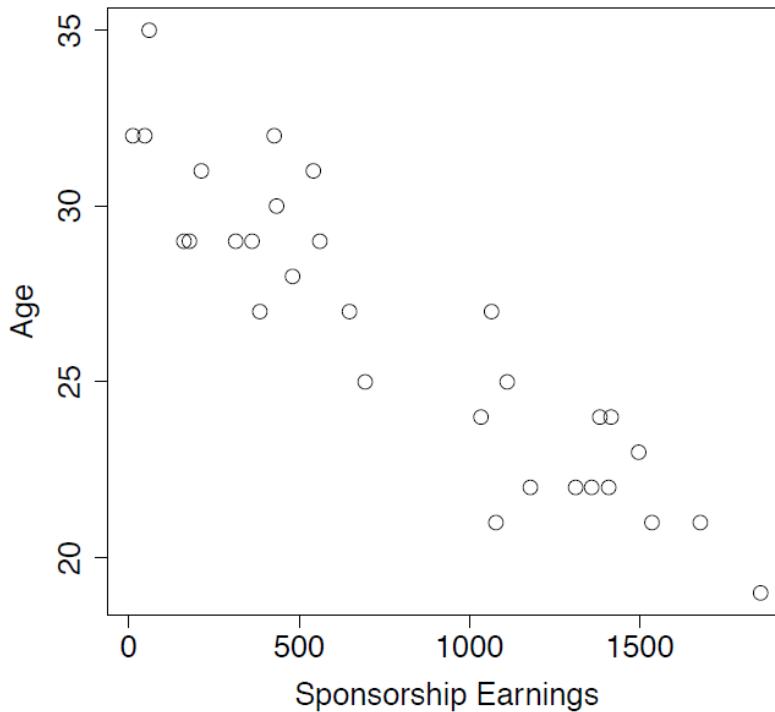
$$\text{corr}(\text{Height}, \text{Weight}) = \frac{241.72}{13.6 \times 19.8} = 0.898$$

Tables and charts taken from Fundamentals of Machine Learning for Predictive Data Analytics by J. Kelleher, B. Mac Namee and A. D'Arcy.

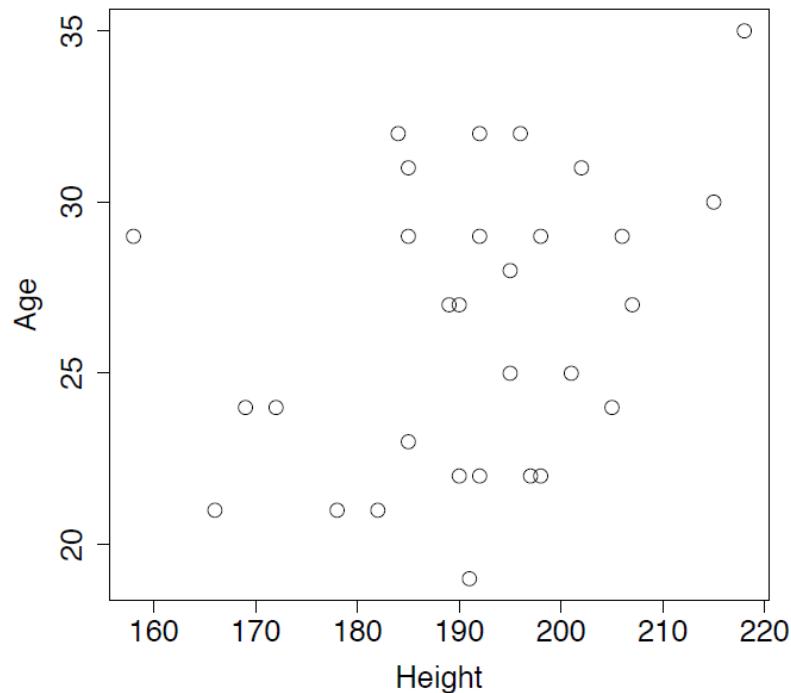


Chair of Process
and Data Science

Negative correlation



Weaker positive correlation



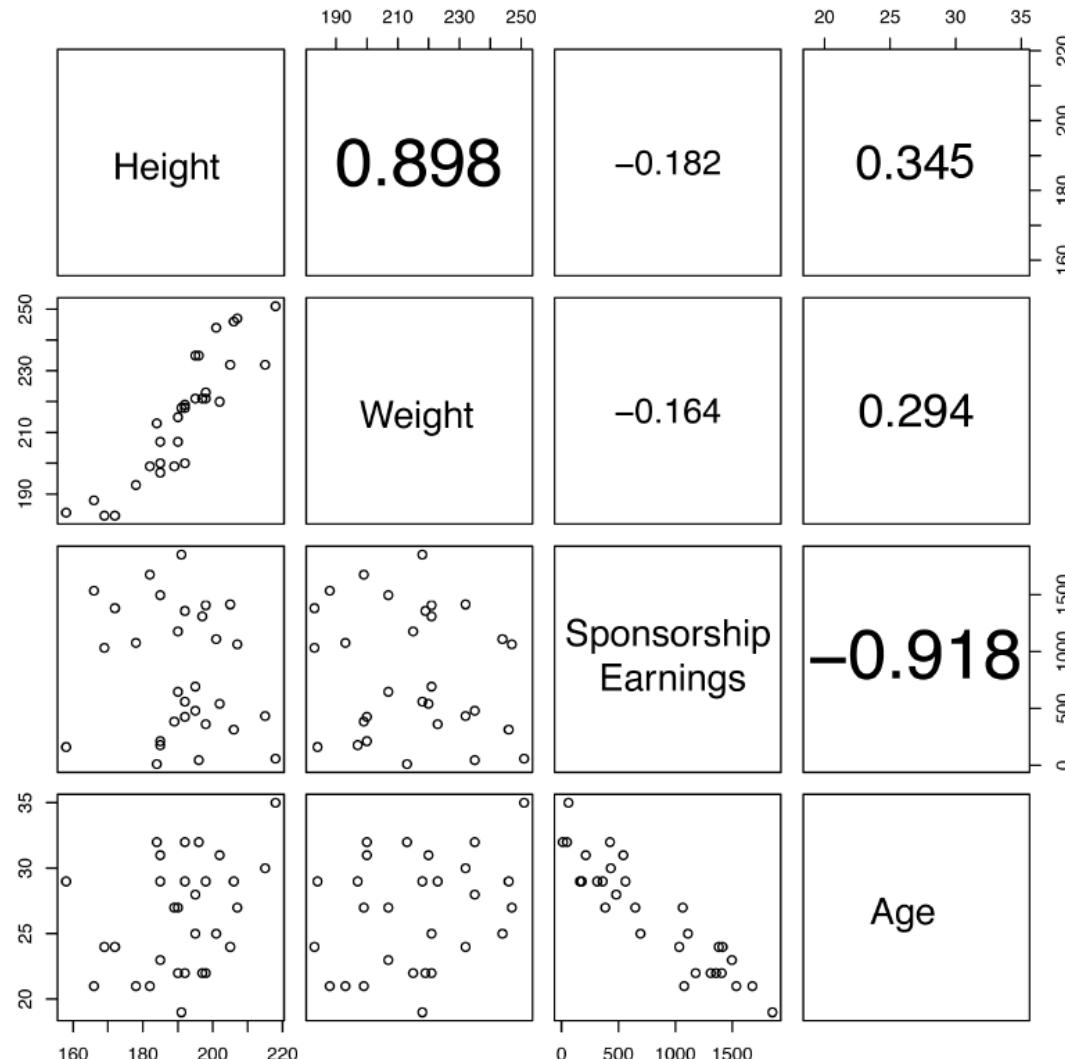
$$\text{corr}(\text{Height}, \text{Age}) = \frac{19.7}{13.6 \times 4.2} = 0.345$$

Correlation matrix

$$\text{correlation matrix}_{\{a,b,\dots,z\}} = \begin{bmatrix} \text{corr}(a, a) & \text{corr}(a, b) & \cdots & \text{corr}(a, z) \\ \text{corr}(b, a) & \text{corr}(b, b) & \cdots & \text{corr}(b, z) \\ \vdots & \vdots & \ddots & \vdots \\ \text{corr}(z, a) & \text{corr}(z, b) & \cdots & \text{corr}(z, z) \end{bmatrix}$$

Example correlation matrix

	height	weight	age
height	1.0	0.898	0.345
weight	0.898	1.0	0.294
age	0.345	0.294	1.0



Preparing for analysis



Preparing for analysis

- **Normalization** (to make things comparable).
- **Binning** (to make things categorical).
- **Sampling** (to make data smaller or to change the bias).

Normalization

- Normalization typically maps values onto a predefined range (e.g. [0,1], [-1,1]) while maintaining relative differences.

$$a'_i = \frac{a_i - \min(a)}{\max(a) - \min(a)} \times (\textit{high} - \textit{low}) + \textit{low}$$

maps values onto $[low, high]$



Standard score

- Standard score uses the standard deviation to normalize.

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

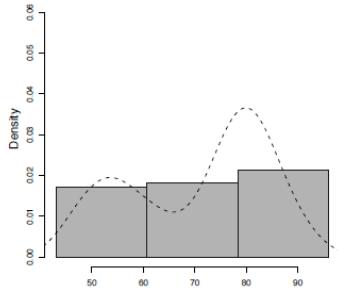
$[-\infty, \infty]$

Binning

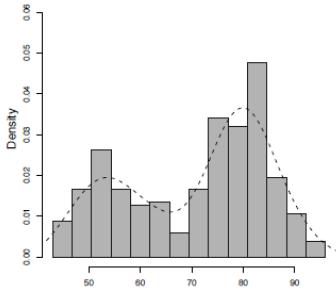
- **Binning is used to make continuous features categorical.**
- **Bins = a series of ranges.**
- **Equal-width binning versus equal-frequency binning.**



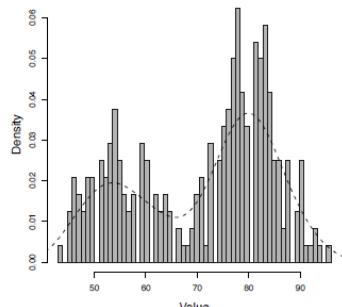
Number of bins



(e) 3 bins



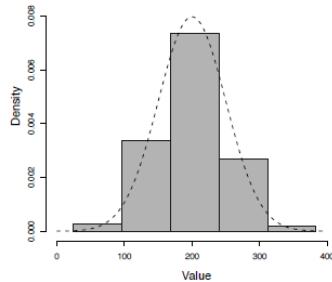
(f) 14 bins



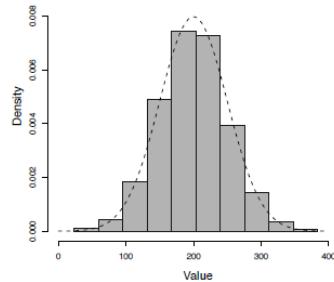
(g) 60 bins

- Too few bins may lead to loss of information (underfitting)
- Too many bins may lead to sparseness, i.e., bins that are empty or just have a few instances (overfitting).

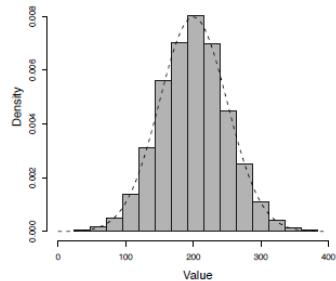
Equal-width binning



(h) 5 Equal-width bins



(i) 10 Equal-width bins

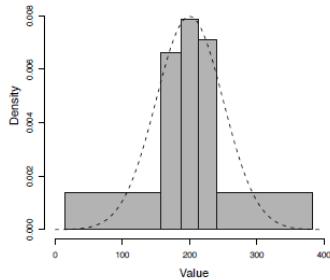


(j) 15 Equal-width bins

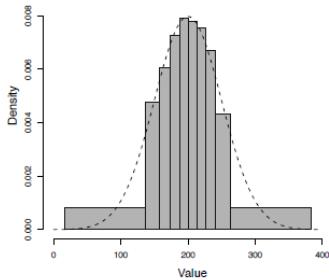
Bins have a **fixed width**,
but the number of items
per bin may vary greatly.

Dashed line shows the original distribution (used to generate the sample).

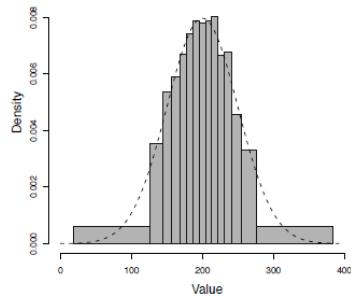
Equal-frequency binning



(k) 5 Equal-frequency bins (l) 10 Equal-frequency bins

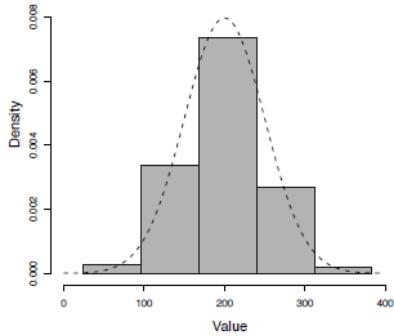


Bins have a **variable width**,
but the number of items
per bin is fixed.

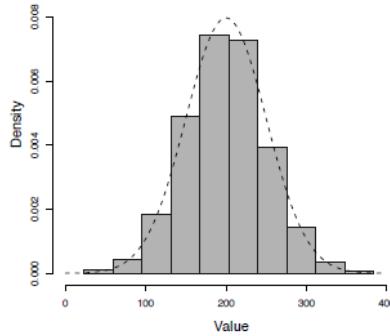


(m) 15 Equal-frequency bins

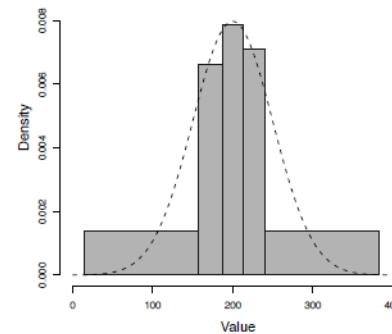
Dashed line shows the original distribution (used to generate the sample).



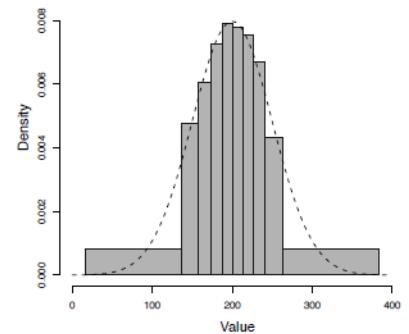
(h) 5 Equal-width bins



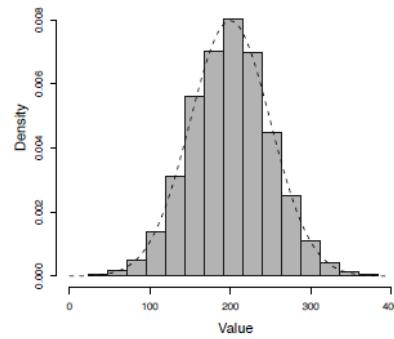
(i) 10 Equal-width bins



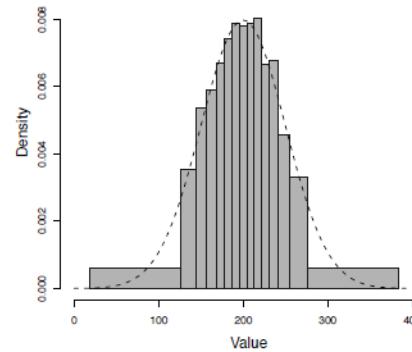
(k) 5 Equal-frequency bins



(l) 10 Equal-frequency bins



(j) 15 Equal-width bins

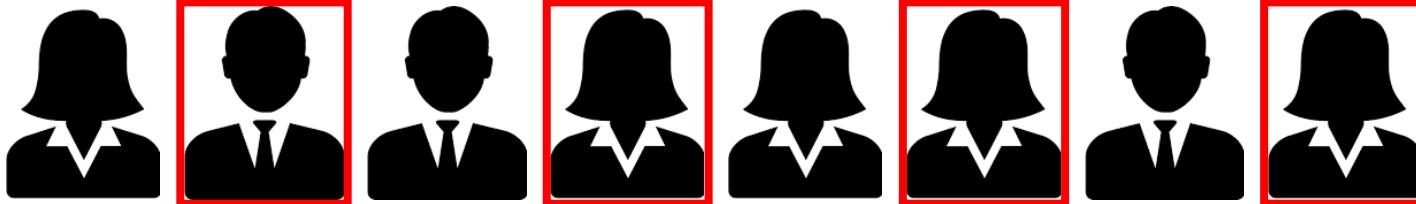


(m) 15 Equal-frequency bins

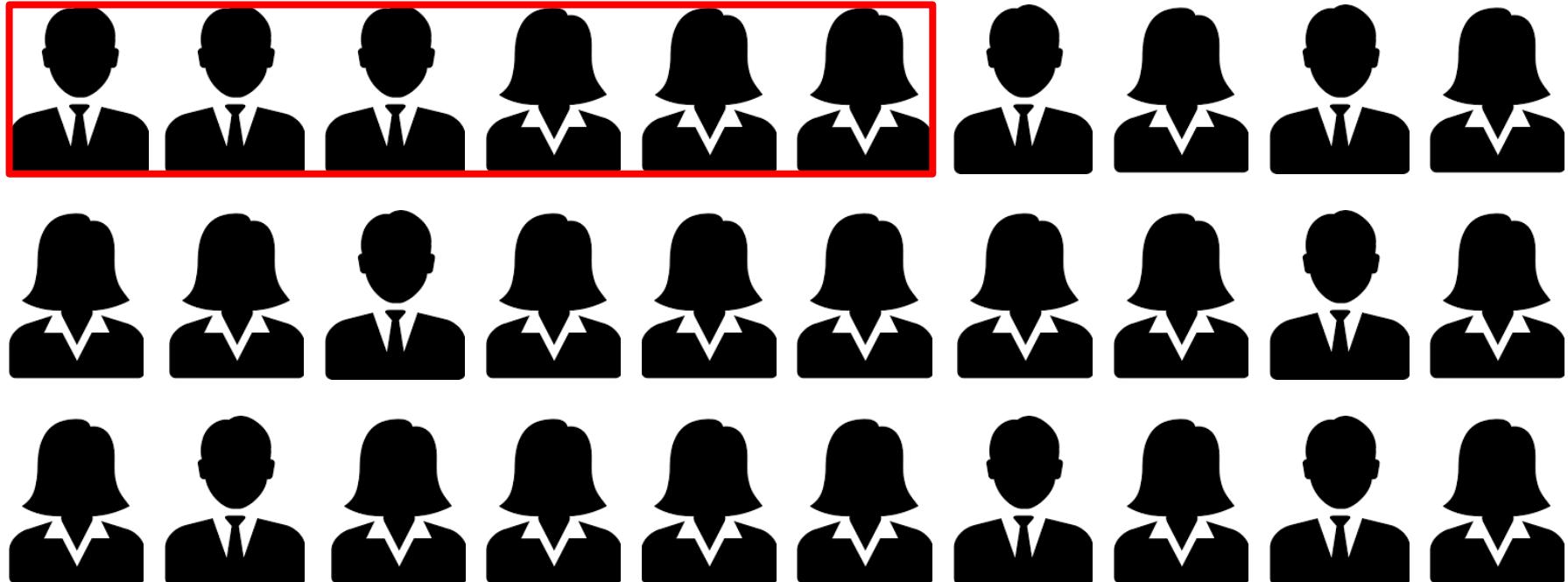
Surface of a bin reflects number of items in bin.

Sampling

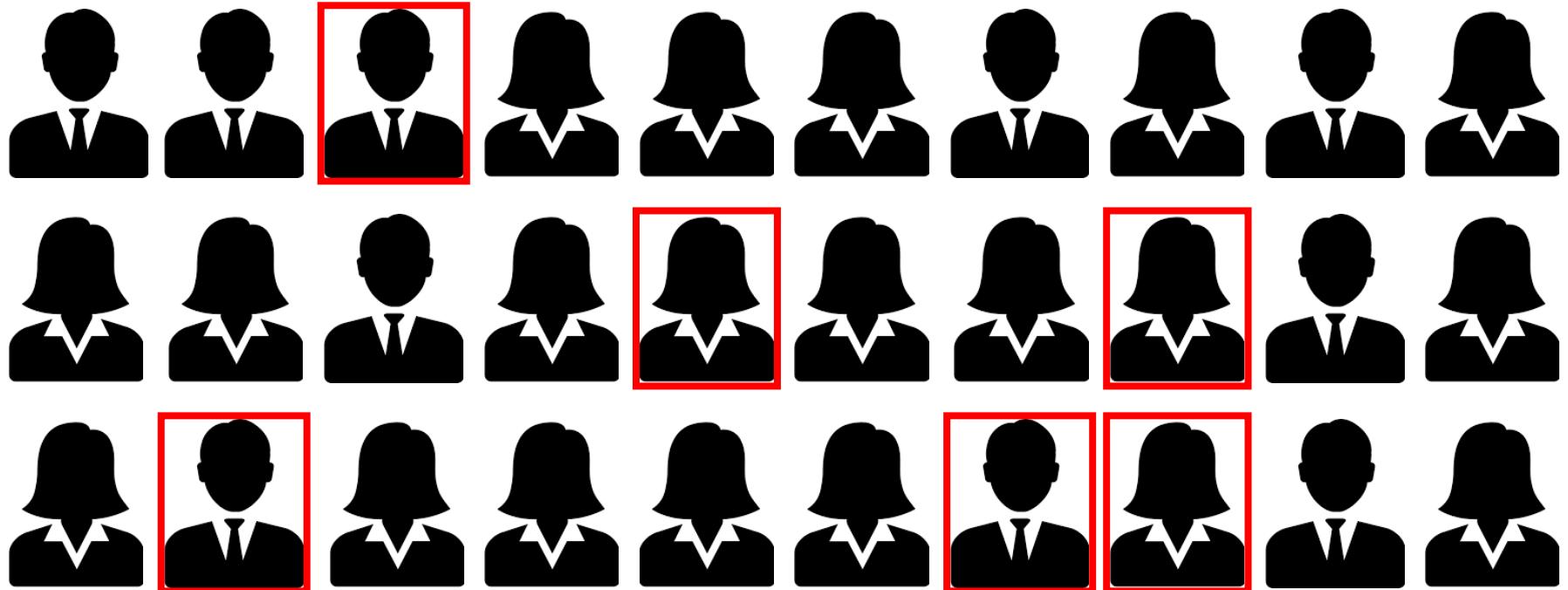
- Different types of sampling (top, random, stratified, and under/over sampling).
- To make data smaller or to remove/introduce a sample bias.



Top sampling (first n instances)

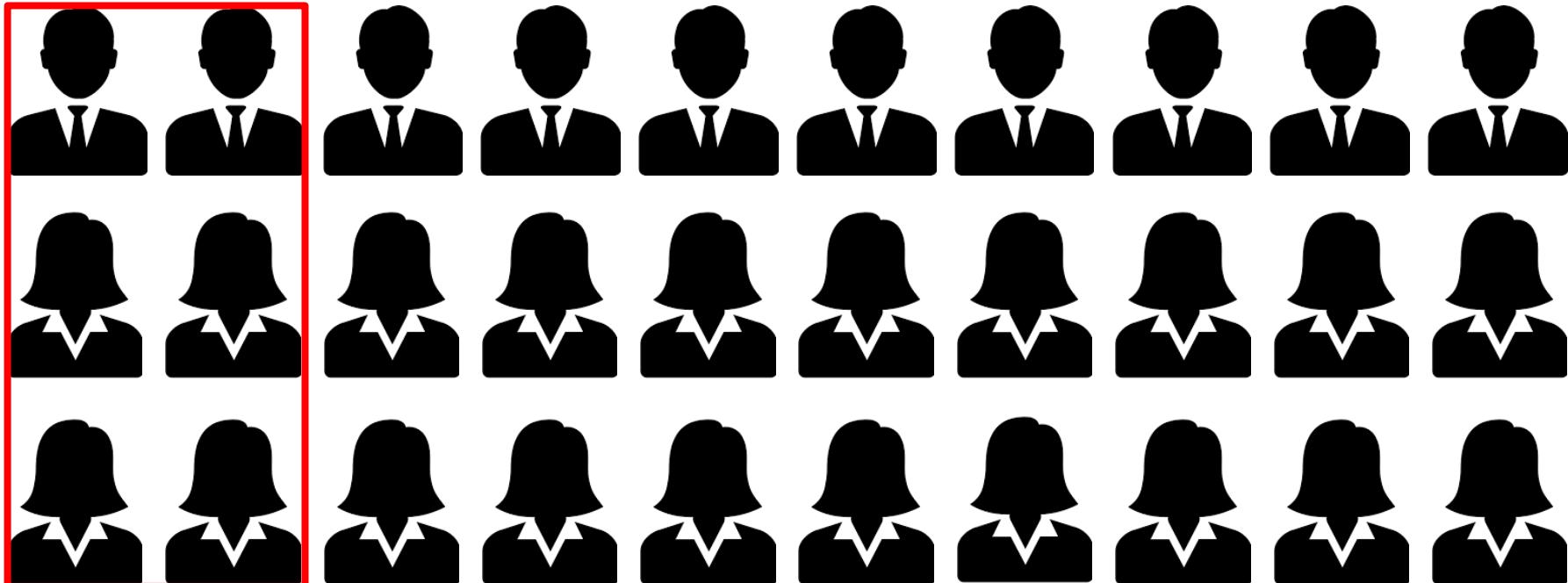


Random sampling

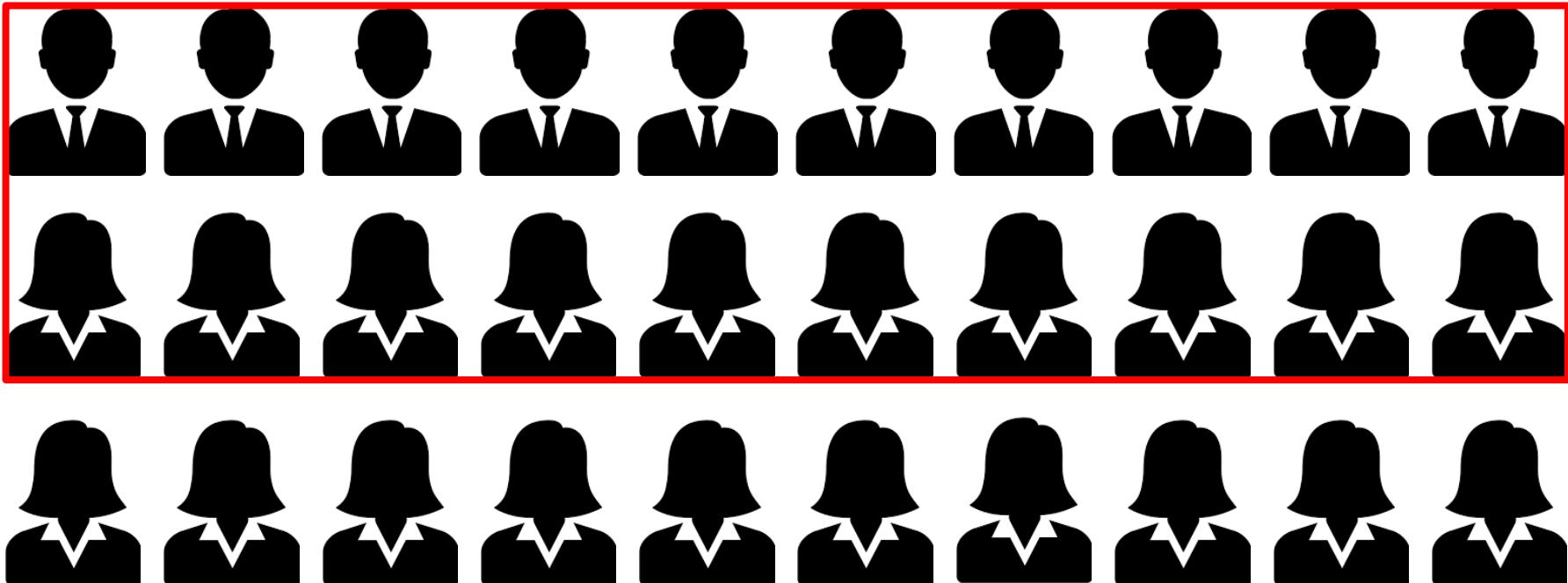


Stratified sampling

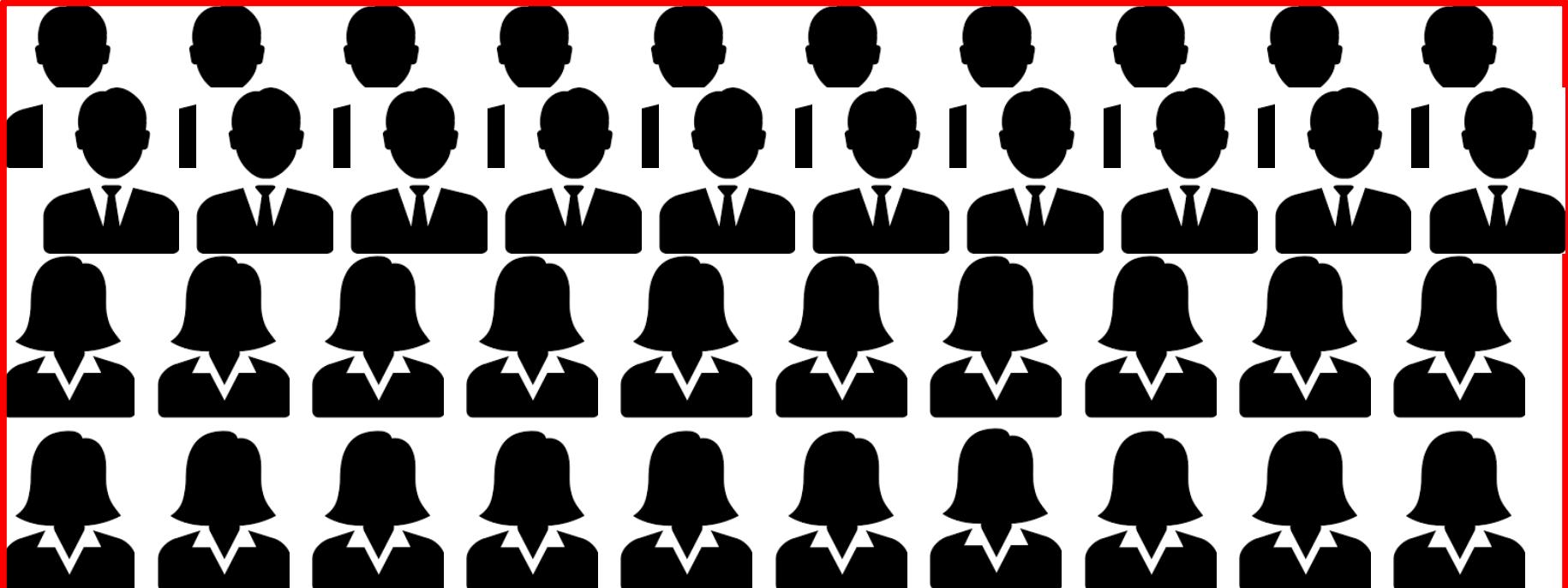
(ensure that relative frequencies are maintained, e.g., by taking the same percentage from every group)



Under-sampling (ensure a balance by leaving out instances of the over-represented group)



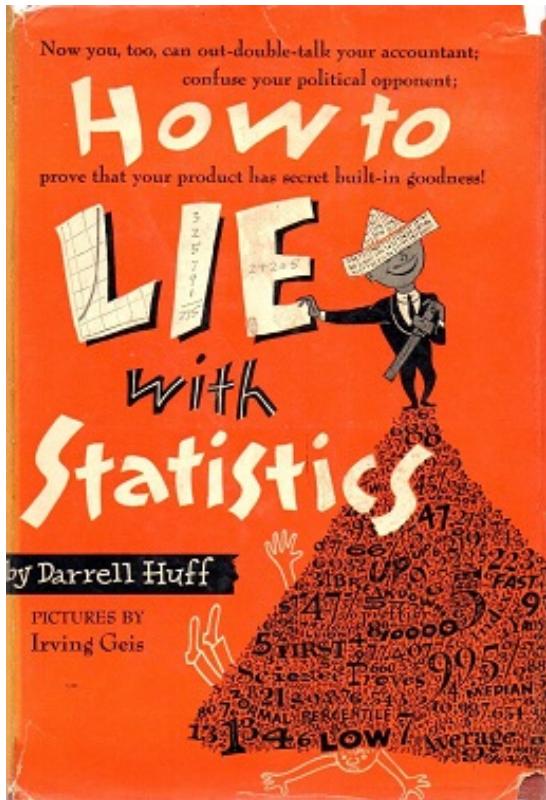
Over-sampling (ensure a balance by possible duplication of under-represented instances)



Good and poor visualizations

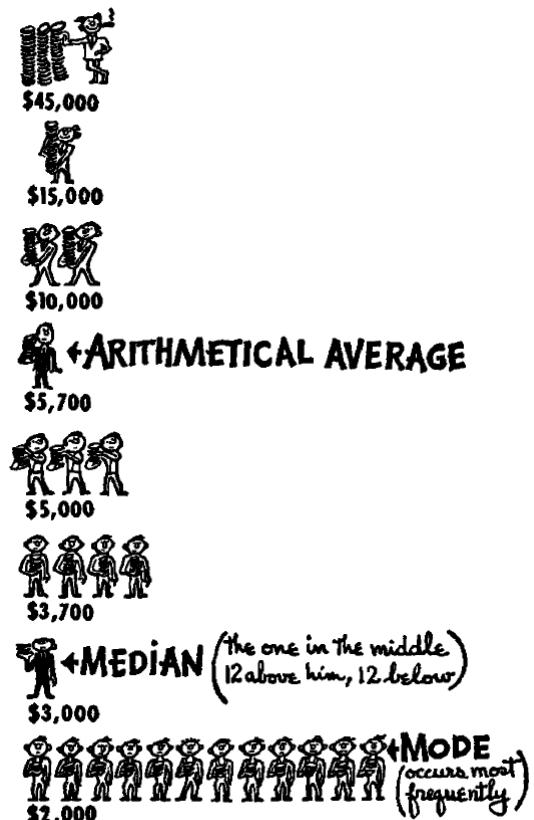


Examples from “How to Lie with Statistics”

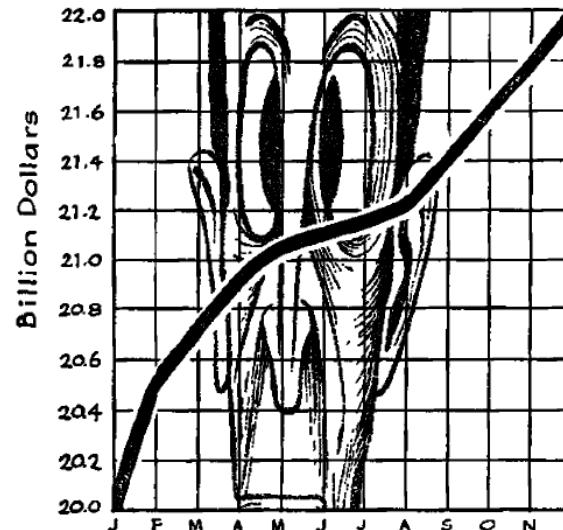
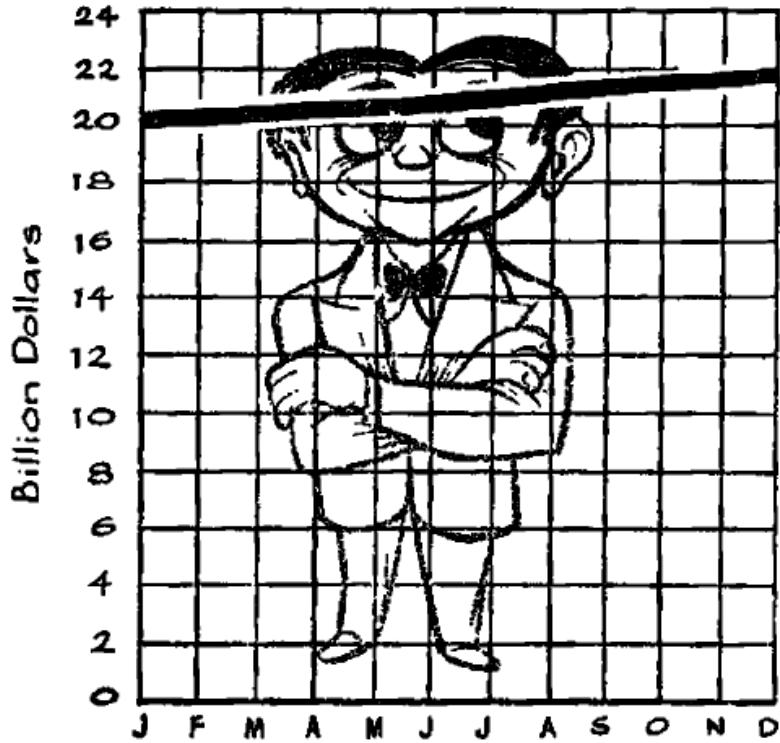


How to Lie with Statistics is a book written by Darrell Huff in 1954 with many examples that are still relevant after more than 60 years.

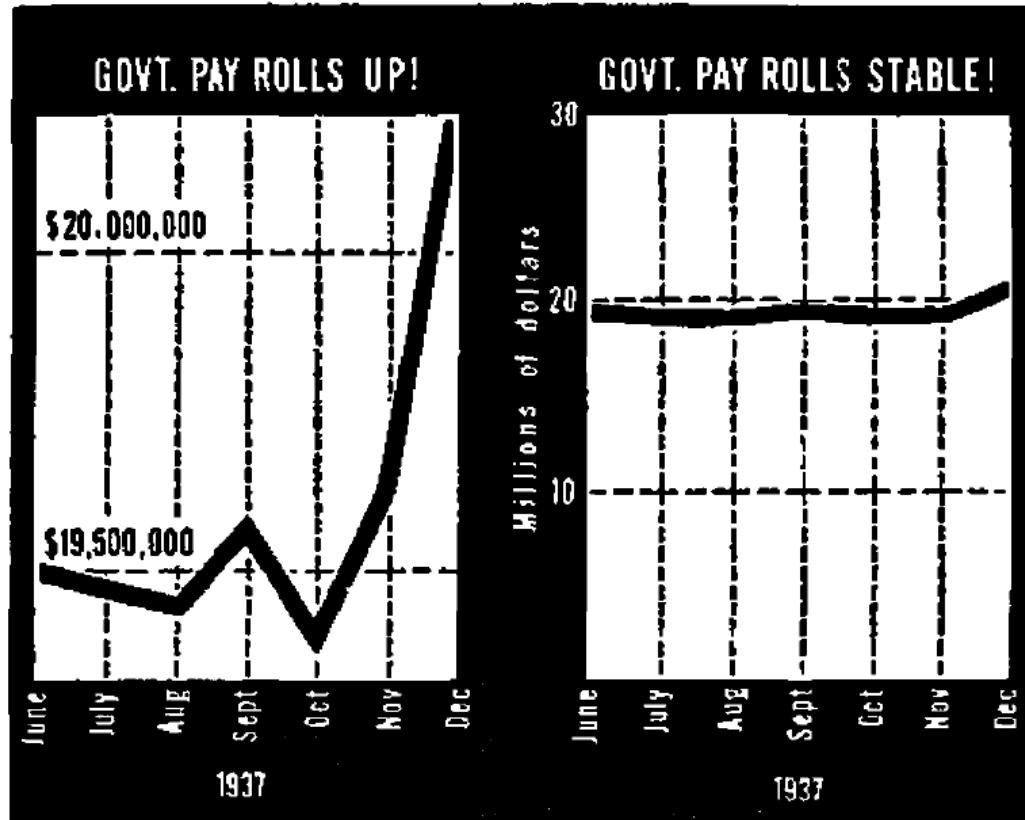
Choosing the “middle”



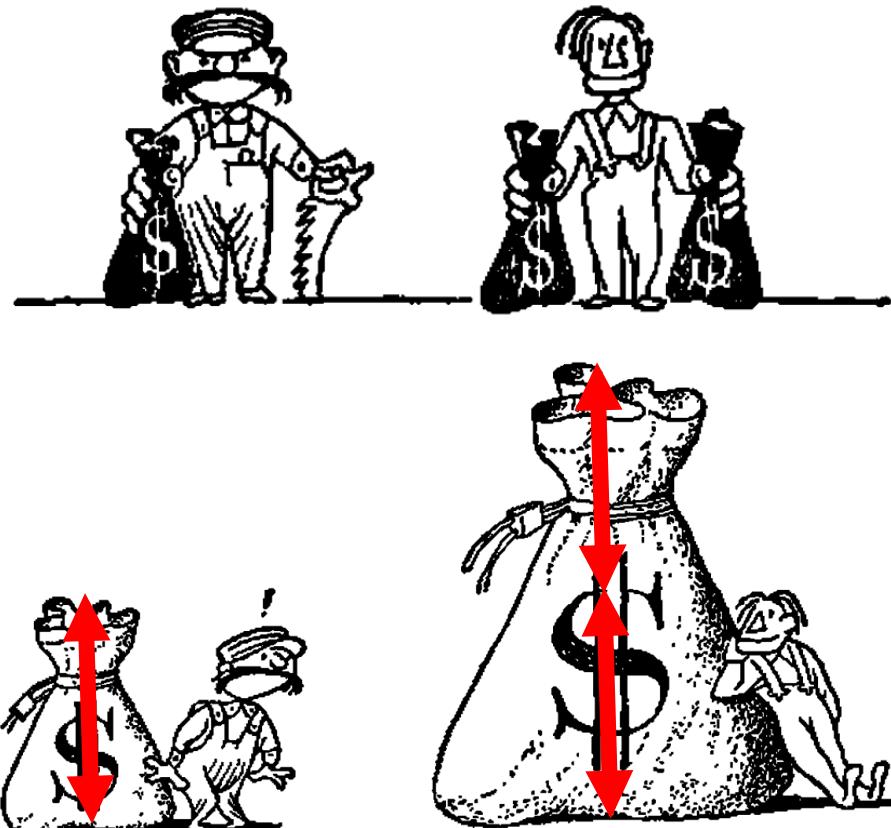
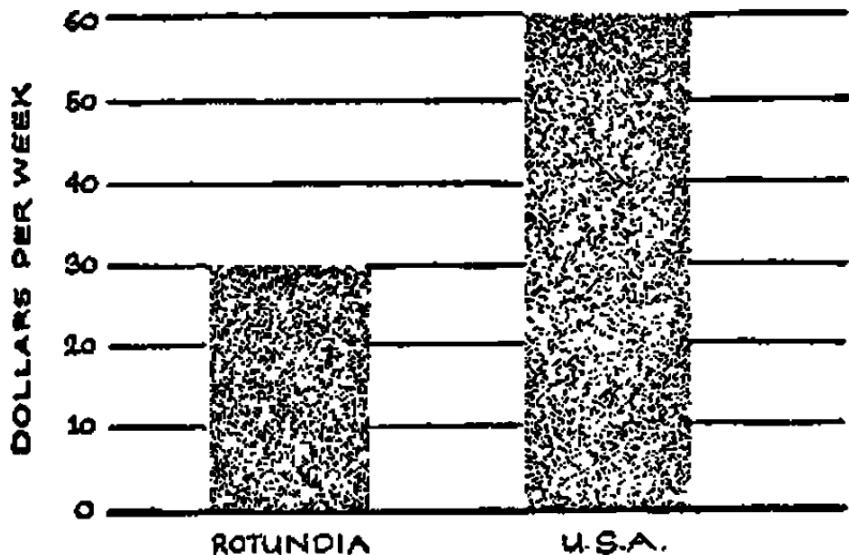
Tampering with the scales



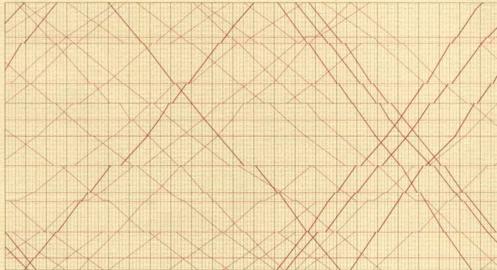
Tampering with the scales



Visualizing a factor 2



Examples from “The Visual Display of Quantitative Information”



The Visual Display
of Quantitative Information

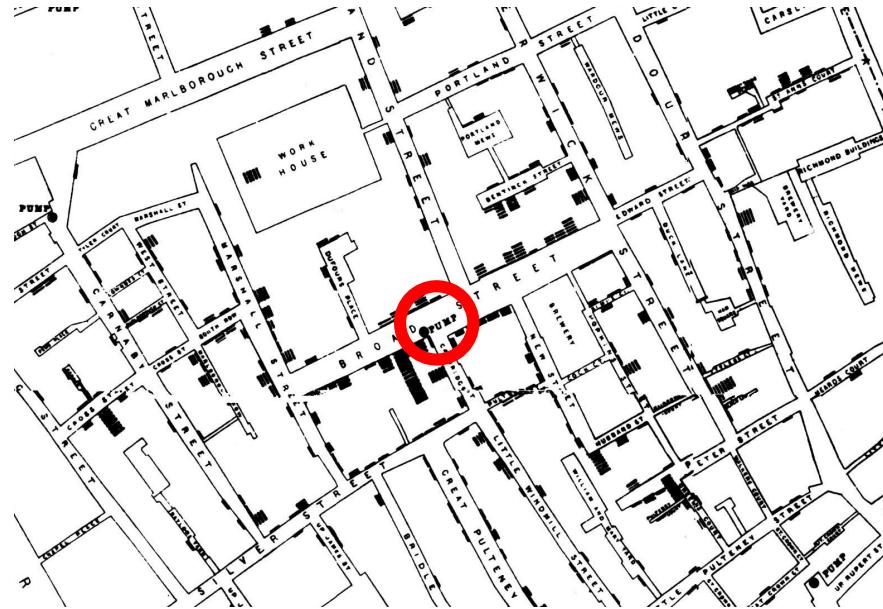
EDWARD R. TUFTE

- **Edward Rolf Tufte is an information visualization expert that wrote the influential book “The Visual Display of Quantitative Information” first published in 1983.**
- **Tufte encourages the use of data-rich illustrations that present all available data such that:**
 - **one can check individual values, and**
 - **see trends and patterns when looking at the whole.**

An early example



John Snow (1813-1858) was an English physician and considered one of the fathers of modern epidemiology. He traces the source of a cholera outbreak in Soho, London, in 1854. His findings inspired fundamental changes in the water and waste systems of London.



John Snow's map of the 1854 cholera outbreak in Soho, London



Chair of Process
and Data Science



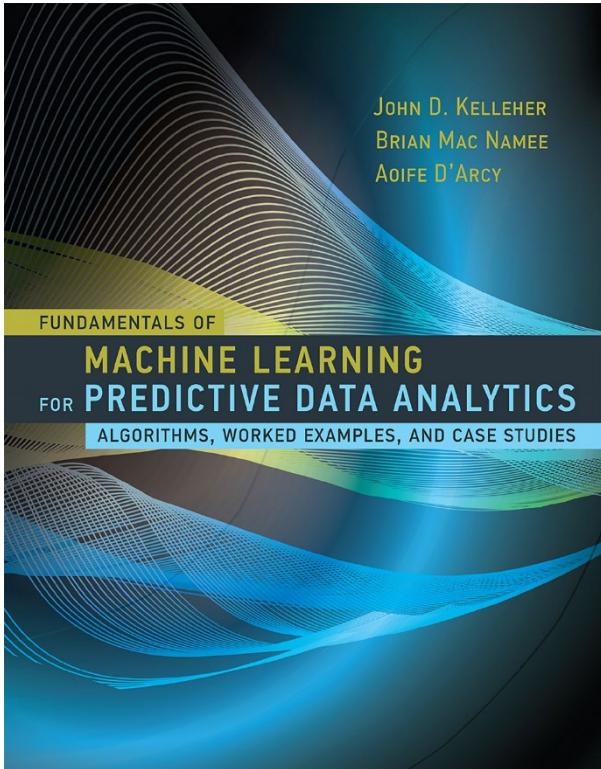
Conclusion



Short summary of the lecture

- Data extraction
- Tabular data
- Importance of visualization
- Characterizing individual features
- Data quality
- Showing relations among features
- Preparing for analysis
- Good and poor visualizations

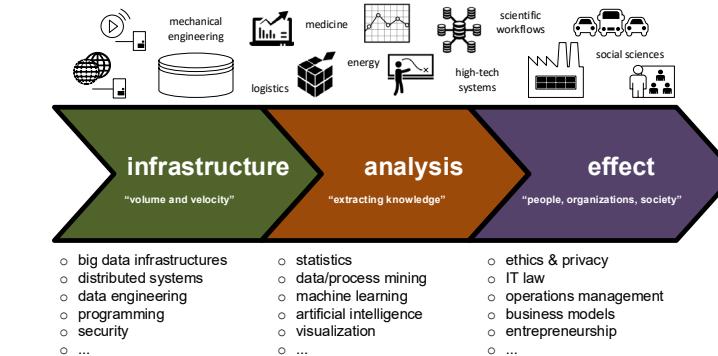
Relevant Literature



Chapters 2 & 3 of Fundamentals of Machine Learning for Predictive Data Analytics by J. Kelleher, B. Mac Namee and A. D'Arcy.

Lecture		date	day
Lecture 1	Introduction	13/10/2021	Wednesday
<i>Instruction 1</i>	Python	14/10/2021	Thursday
<i>Instruction 2</i>	Crash Course in Python	15/10/2021	Friday
Lecture 2	Basic data visualization/exploration	20/10/2021	Wednesday
Lecture 3	Decision trees	21/10/2021	Thursday
<i>Instruction 3</i>	Decision trees and data visualization/exploration	22/10/2021	Friday
Lecture 4	Regression	27/10/2021	Wednesday
Lecture 5	Support vector machines	28/10/2021	Thursday
<i>Instruction 4</i>	Regression and support vector machines	29/10/2021	Friday
Lecture 6	Neural networks (1/2)	03/11/2021	Wednesday
Lecture 7	Neural networks (2/2)	04/11/2021	Thursday
Lecture 8	Evaluation of supervised learning problems	10/11/2021	Wednesday
<i>Instruction 5</i>	Neural networks	11/11/2021	Thursday
<i>Instruction 6</i>	Neural networks and evaluation	12/11/2021	Friday
Lecture 9	Clustering	17/11/2021	Wednesday
Lecture 10	Frequent item sets	18/11/2021	Thursday
<i>Instruction 7</i>	Clustering and frequent item sets	19/11/2021	Friday
Lecture 11	Association rules	24/11/2021	Wednesday
Lecture 12	Sequence mining	25/11/2021	Thursday
<i>Instruction 8</i>	Association rules and sequence mining	26/11/2021	Friday
Lecture 13	Process mining (unsupervised)	01/12/2021	Wednesday
Lecture 14	Process mining (supervised)	02/12/2021	Thursday
<i>Instruction 9</i>	Process Mining	03/12/2021	Friday
Lecture 15	Text Mining (1/2)	08/12/2021	Wednesday
Lecture 16	Text Mining (2/2)	09/12/2021	Thursday
<i>Instruction 10</i>	Q&A Assignment 1	10/12/2021	Friday
Lecture 17	Data preprocessing, data quality, binning, etc.	15/12/2021	Wednesday
Lecture 18	Visual analytics & information visualization	16/12/2021	Thursday
<i>Instruction 11</i>	Text Mining	17/12/2021	Friday
Lecture 19	Responsible data science (1/2)	22/12/2021	Wednesday
Lecture 20	Responsible data science (2/2)	23/12/2021	Thursday
Lecture 21	Big data	12/01/2022	Wednesday
<i>Instruction 12</i>	Preprocessing and visualization	13/01/2022	Thursday
<i>Instruction 13</i>	Q&A Assignment 2	14/01/2022	Friday
Lecture 22	Closing	19/01/2022	Wednesday
<i>Instruction 14</i>	Big Data (1/2)	20/01/2022	Thursday
<i>Instruction 15</i>	Responsible data science	21/01/2022	Friday
<i>Instruction 16</i>	Big Data (2/2)	27/01/2022	Thursday
<i>Instruction 17</i>	Example Exam Questions	28/01/2022	Friday
<i>Instruction 18</i>	Questions	02/02/2022	Wednesday

Disclaimer: always check RWTH Moodle for last minute changes



Lecture		date	day
Lecture 1	Introduction	20/10/2021	Wednesday
Lecture 2	Basic data visualization/exploration		
Lecture 3	Decision trees	27/10/2021	Wednesday
Lecture 4	Regression		
Lecture 5	Support vector machines		
Lecture 6	Neural networks (1/2)		
Lecture 7	Neural networks (2/2)	10/11/2021	Wednesday
Lecture 8	Evaluation of supervised learning problems		
Lecture 9	Clustering	17/11/2021	Wednesday
Lecture 10	Frequent item sets		
Lecture 11	Association rules	24/11/2021	Wednesday
Lecture 12	Sequence mining		
Lecture 13	Process mining (unsupervised)		
Lecture 14	Process mining (supervised)	02/12/2021	Thursday
Lecture 15	Text Mining (1/2)		
Lecture 16	Text Mining (2/2)	09/12/2021	Thursday
Lecture 17	Data preprocessing, data quality, binning, etc.		
Lecture 18	Visual analytics & information visualization	16/12/2021	Thursday
Lecture 19	Responsible data science (1/2)		
Lecture 20	Responsible data science (2/2)	23/12/2021	Thursday
Lecture 21	Big data		
Lecture 22	Closing	19/01/2022	Wednesday