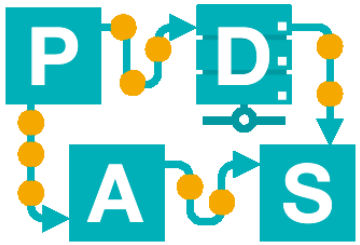
Data Visualization and Exploration

Lecture 3

IDS-L3



Chair of Process and Data Science

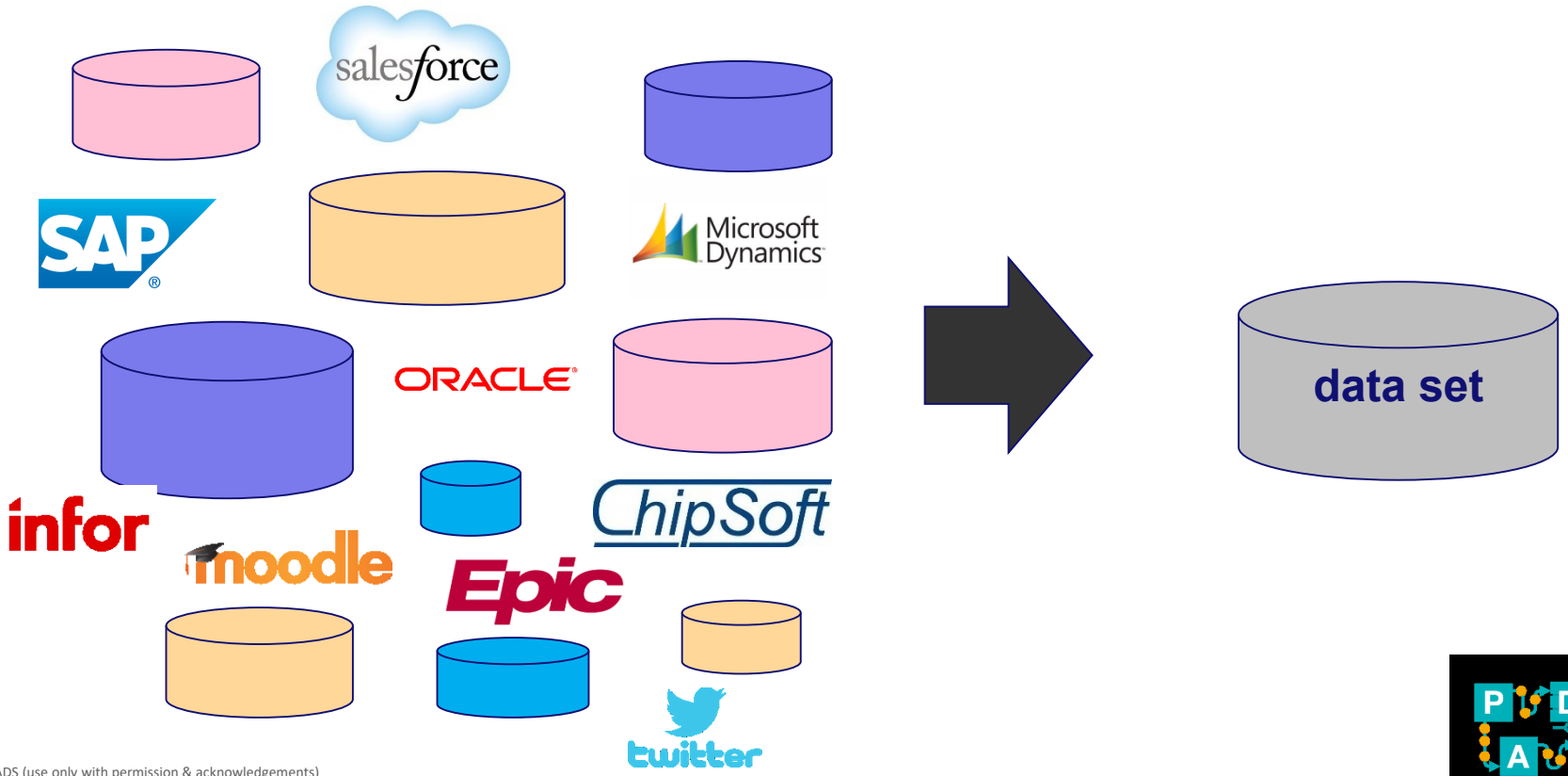
RWTH AACHEN UNIVERSITY

Outline of Today's Lecture

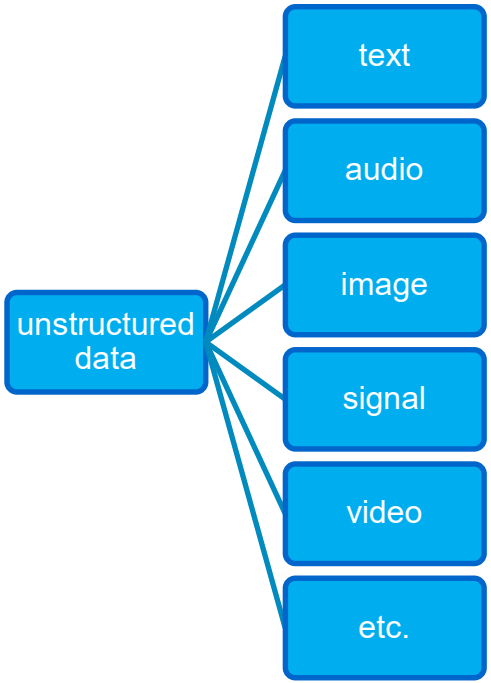
- Data extraction
- Tabular data
- Importance of visualization
- Characterizing individual features
- Data quality
- Showing relations among features
- Preparing for analysis
- Good and poor visualizations

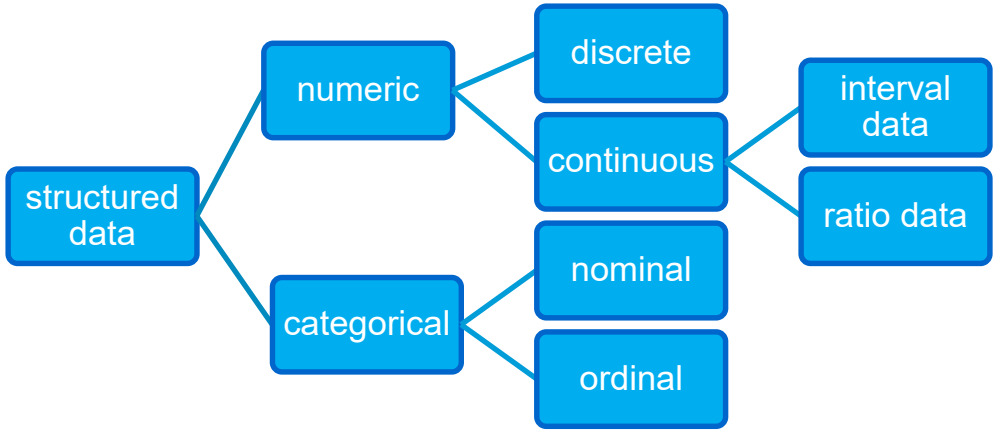
Data extraction

Extracting data



Types of data



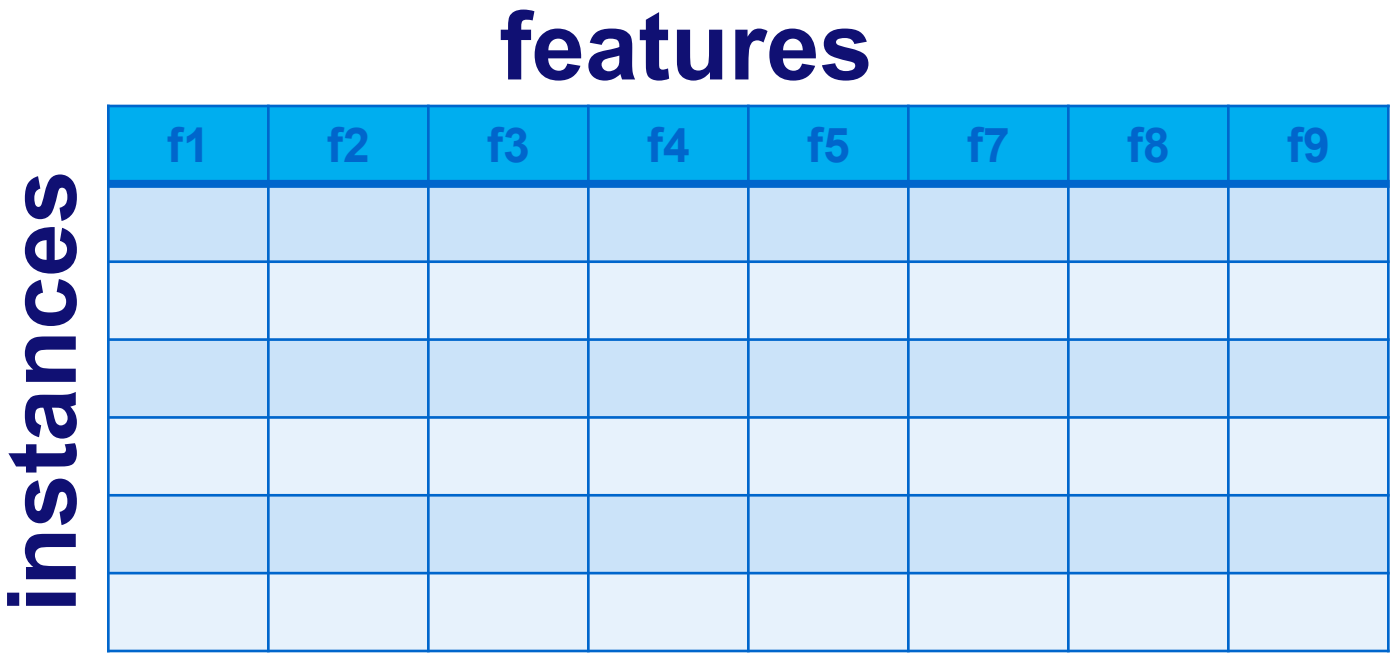


Feature extraction

CDDDCDCDCDCDCDCD

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



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

Importance of visualization

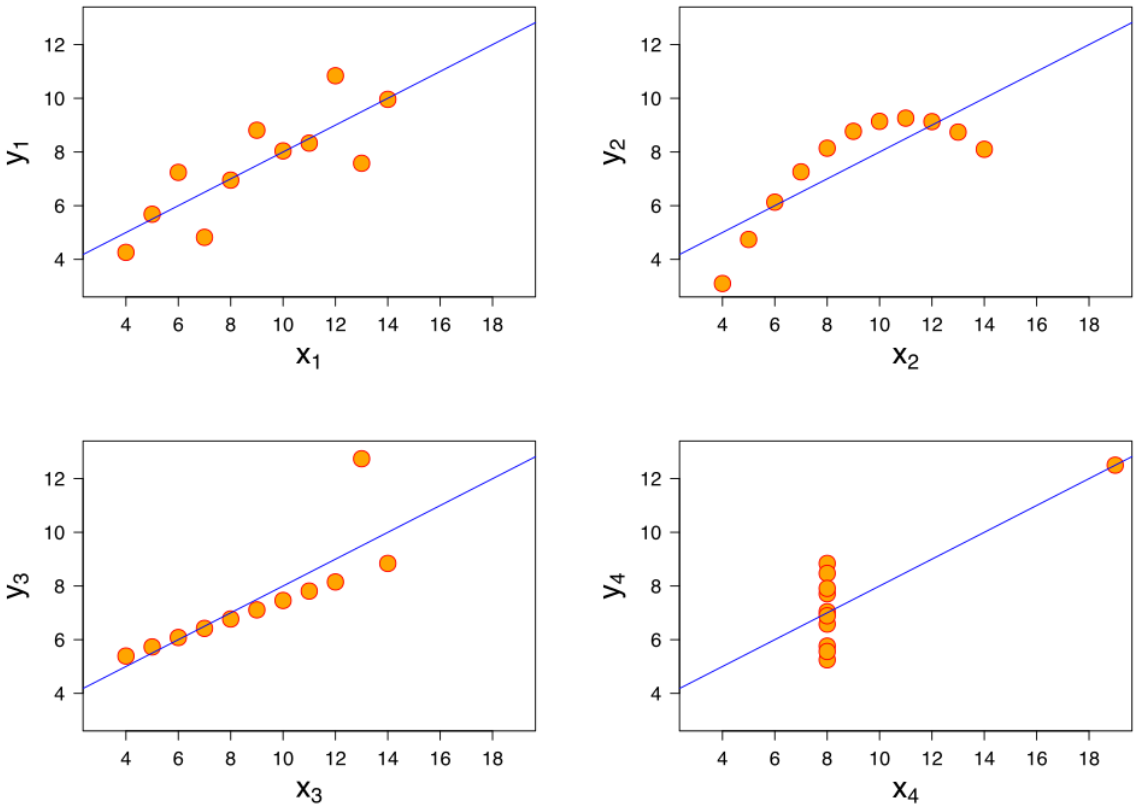
Anscombe's quartet

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

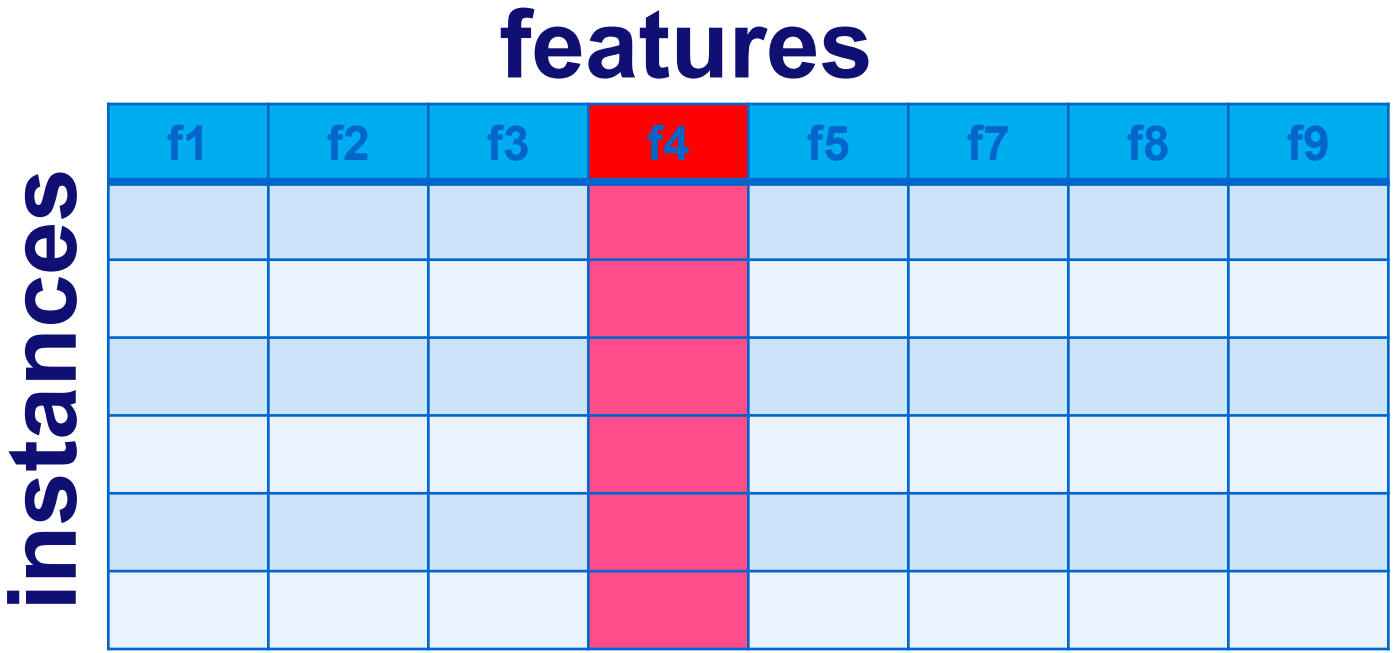
Data set 1		Data set 2		Data set 3		Data set 4	
x	y	x	y	x	y	x	y
10,0	8,04	10,0	9,14	10,0	7,46	8,0	6,58
8,0	6,95	8,0	8,14	8,0	6,77	8,0	5,76
13,0	7,58	13,0	8,74	13,0	12,74	8,0	7,71
9,0	8,81	9,0	8,77	9,0	7,11	8,0	8,84
11,0	8,33	11,0	9,26	11,0	7,81	8,0	8,47
14,0	9,96	14,0	8,10	14,0	8,84	8,0	7,04
6,0	7,24	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!

Characterizing individual features

Let’s focus on one feature



Investigating individual features

(a) Continuous Features

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

(b) Categorical Features

Feature	Count	% Miss.	Card.	Mode	Mode Freq.	Mode %	2nd Mode	2nd Mode Freq.	2nd 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.	1st Qrt.	Mean	Median	3rd 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.	1st Qrt.	Mean	Median	3rd 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 %	2nd Mode	2nd Mode Freq.	2nd 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)

ID	Type	Inc.	Marital Status	Num Clmnts.	Injury Type	Hospital Stay	Claim Amnt.	Total Claimed	Num Claims	Num Soft Tiss.	% Soft Tiss.	Claim Amt Rcvd.	Fraud Flag
1	CI	0		2	Soft Tissue	No	1,625	3250	2	2	1.0	0	1
2	CI	0		2	Back	Yes	15,028	60,112	1		0	15,028	0
3	CI	54,613	Married	1	Broken Limb	No	-99,999	0	0	0	0	572	0
4	CI	0		4	Broken Limb	Yes	5,097	11,661	1	1	1.0	7,864	0
5	CI	0		4	Soft Tissue	No	8869	0	0	0	0	0	1
6	CI	0		1	Broken Limb	Yes	17,480	0	0	0	0	17,480	0
7	CI	52,567	Single	3	Broken Limb	No	3,017	18,102	2	1	0.5	0	1
8	CI	0		2	Back	Yes	7463	0	0	0	0	7,463	0
9	CI	0		1	Soft Tissue	No	2,067	0	0	0	0	2,067	0
10	CI	42,300	Married	4	Back	No	2,260	0	0	0	0	2,260	0
		⋮				⋮					⋮		
300	CI	0		2	Broken Limb	No	2,244	0	0	0	0	2,244	0
301	CI	0		1	Broken Limb	No	1,627	92,283	3	0	0	1,627	0
302	CI	0		3	Serious	Yes	270,200	0	0	0	0	270,200	0
303	CI												
304	CI	46											
458	CI	48											
459	CI												
460	CI												
461	CI	47											
462	CI												
491	CI	40											
492	CI												
493	CI												
494	CI	31											
495	CI												
496	CI												
497	CI	29											
498	CI												
499	CI	46											
500	CI												

Feature	Count	% Miss.	Card.	Min	1st Qrt.	Mean	Median	3rd Qrt.	Max	Std. Dev.
Income	500	0.0	171	0.0	0.0	13,740.0	0.0	33,918.5	71,284.0	20,081.5
Num Claimants	500	0.0	4	1.0	1.0	1.9	2	3.0	4.0	1.0
Claim Amount	500	0.0	493	-99,999	3,322.3	16,373.2	5,663.0	12,245.5	270,200.0	29,426.3
Total Claimed	500	0.0	235	0.0	0.0	9,597.2	0.0	11,282.8	729,792.0	35,655.7
Num Claims	500	0.0	7	0.0	0.0	0.8	0.0	1.0	56.0	2.7
Num Soft Tissue	500	2.0	6	0.0	0.0	0.2	0.0	0.0	5.0	0.6
% Soft Tissue	500	0.0	9	0.0	0.0	0.2	0.0	0.0	2.0	0.4
Amount Received	500	0.0	329	0.0	0.0	13,051.9	3,253.5	8,191.8	295,303.0	30,547.2
Fraud Flag	500	0.0	2	0.0	0.0	0.3	0.0	1.0	1.0	0.5

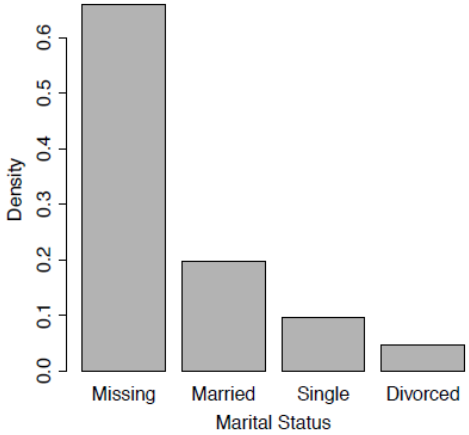
Example table (from book)

ID	Type	Inc.	Marital Status	Num Clmnts.	Injury Type	Hospital Stay	Claim Amnt.	Total Claimed	Num Claims	Num Soft Tiss.	% Soft Tiss.	Claim Amt Rcvd.	Fraud Flag
1	CI	0		2	Soft Tissue	No	1,625	3250	2	2	1.0	0	1
2	CI	0		2	Back	Yes	15,028	60,112	1		0	15,028	0
3	CI	54,613	Married	1	Broken Limb	No	-99,999	0	0	0	0	572	0
4	CI	0		4	Broken Limb	Yes	5,097	11,661	1	1	1.0	7,864	0
5	CI	0		4	Soft Tissue	No	8869	0	0	0	0	0	1
6	CI	0		1	Broken Limb	Yes	17,480	0	0	0	0	17,480	0
7	CI	52,567	Single	3	Broken Limb	No	3,017	18,102	2	1	0.5	0	1
8	CI	0		2	Back	Yes	7463	0	0	0	0	7,463	0
9	CI	0		1	Soft Tissue	No	2,067	0	0	0	0	2,067	0
10	CI	42,300	Married	4	Back	No	2,260	0	0	0	0	2,260	0
		⋮				⋮					⋮		
300	CI	0		2	Broken Limb	No	2,244	0	0	0	0	2,244	0
301	CI	0		1	Broken Limb	No	1,627	92,283	3	0	0	1,627	0
302	CI	0		3	Serious	Yes	270,200	0	0	0	0	270,200	0
303	CI												
304	CI												
458	CI												
459	CI												
460	CI												
461	CI												
462	CI												
491	CI												
492	CI												
493	CI												
494	CI												
495	CI												
496	CI	0		1	Soft Tissue	No	2,118	0	0	0	0	0	1
497	CI	29,280	Married	4	Broken Limb	Yes	3,199	0	0	0	0	0	1
498	CI	0		1	Broken Limb	Yes	32,469	0	0	0	0	16,763	0
499	CI	46,683	Married	1	Broken Limb	No	179,448	0	0		0	179,448	0
500	CI	0		1	Broken Limb	No	8,259	0	0	0	0	0	1

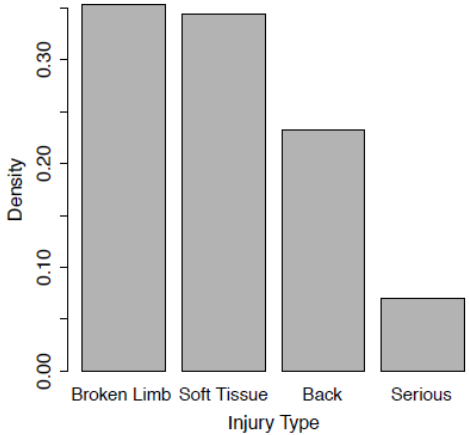
Feature	Count	% Miss.	Card.	Mode	Mode Freq.	Mode %	2nd Mode	2nd Mode Freq.	2nd Mode %
Insurance Type	500	0.0	1	CI	500	1.0	–	–	–
Marital Status	500	61.2	4	Married	99	51.0	Single	48	24.7
Injury Type	500	0.0	4	Broken Limb	177	35.4	Soft Tissue	172	34.4
Hospital Stay	500	0.0	2	No	354	70.8	Yes	146	29.2

Visualizations of distributions

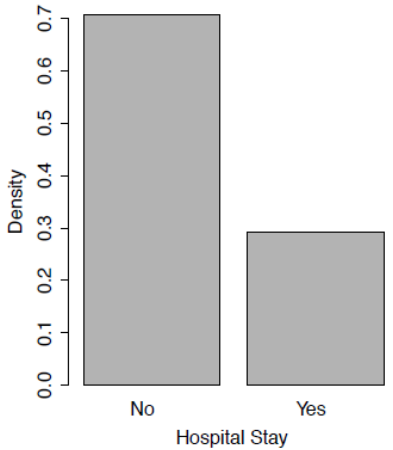
Y-axis can be frequency or a percentage



(a) MARITAL STATUS



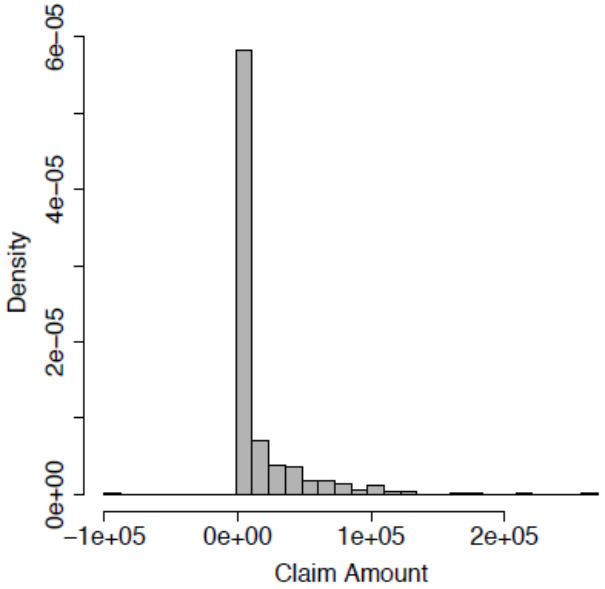
(b) INJURY TYPE

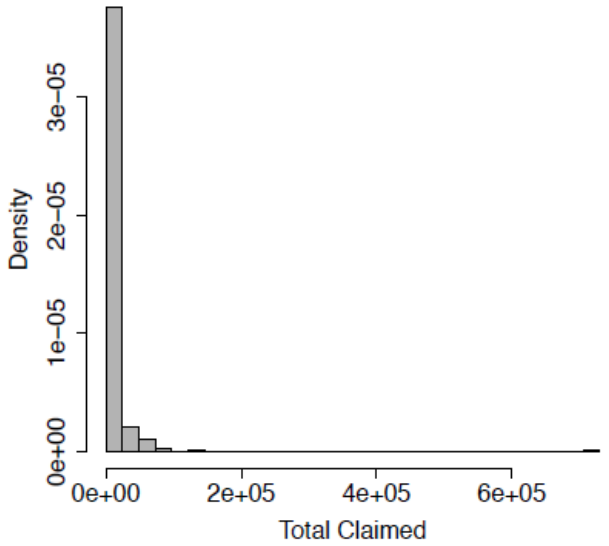


(c) HOSPITAL STAY

Visualizations of distributions

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

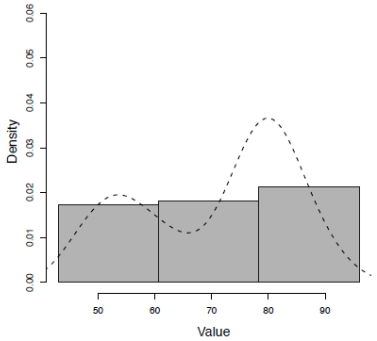




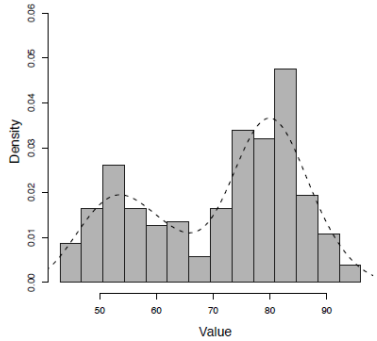
Chair of Process and Data Science

Visualizations of distributions

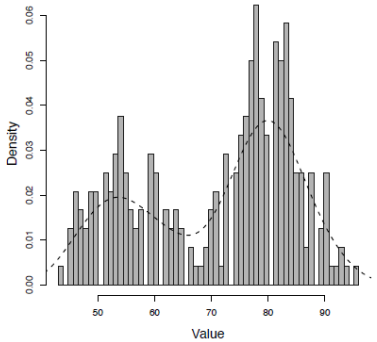
for continuous variables one may want to group items (binning)



(e) 3 bins



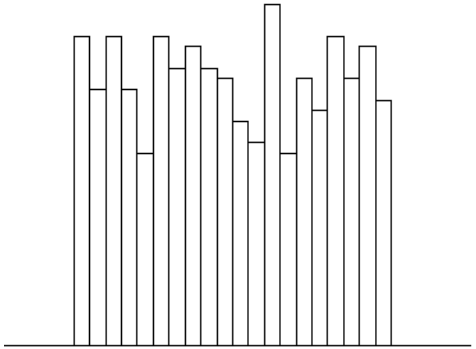
(f) 14 bins



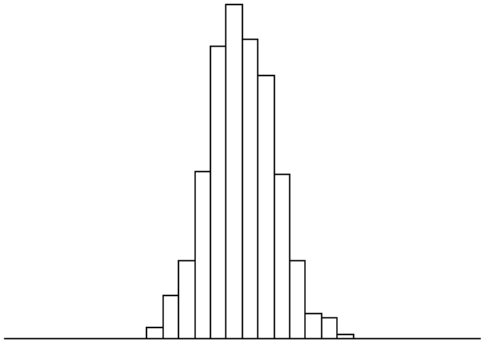
(g) 60 bins

Chair of Process and Data Science

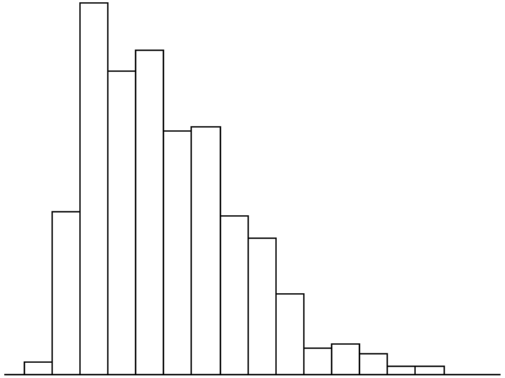
Different types of histograms



(a) Uniform



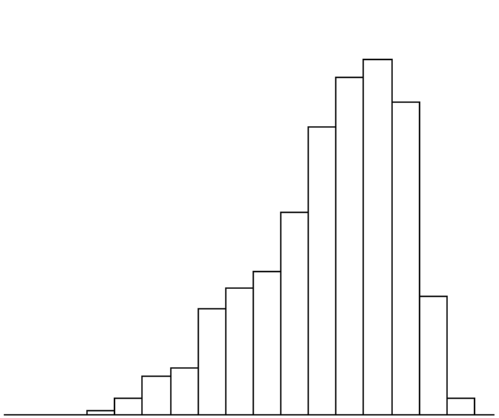
(b) Normal (Unimodal)



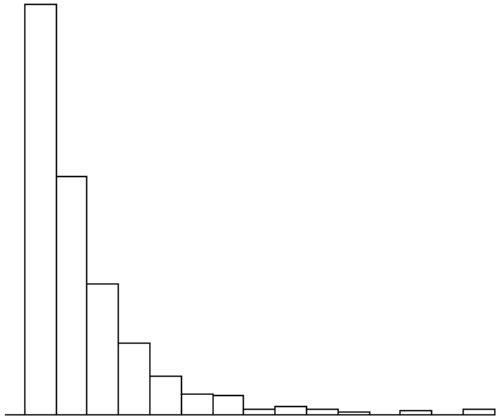
(c) Unimodal (skewed right)

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

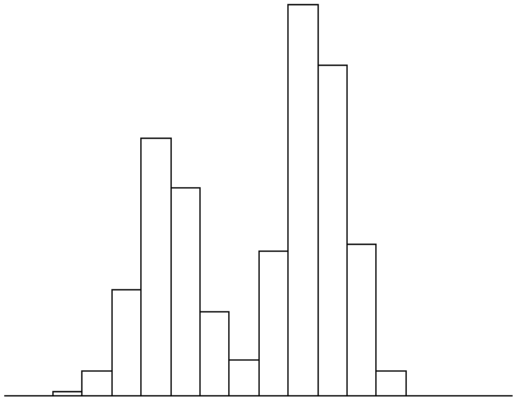
Different types of histograms



(a) Unimodal (skewed left)

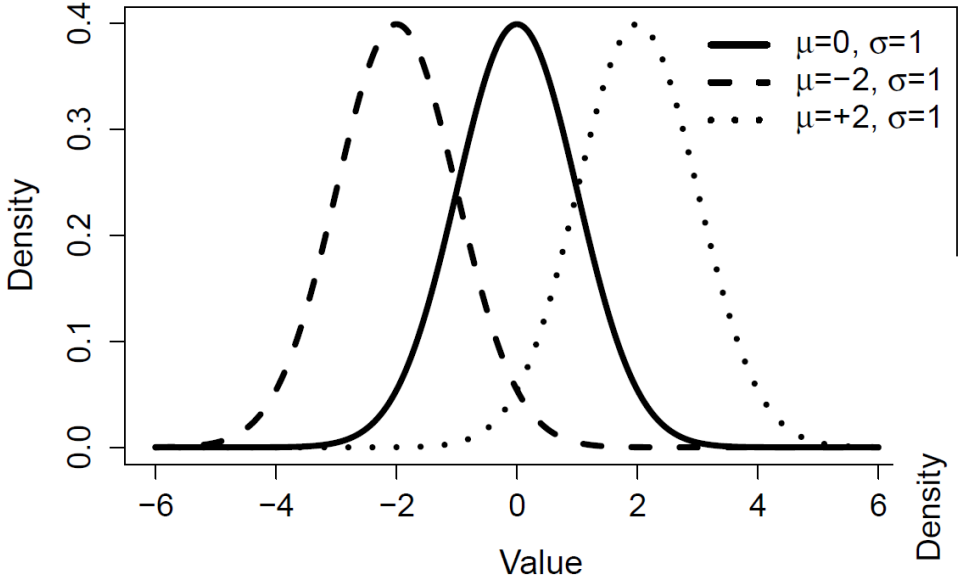


(b) Exponential



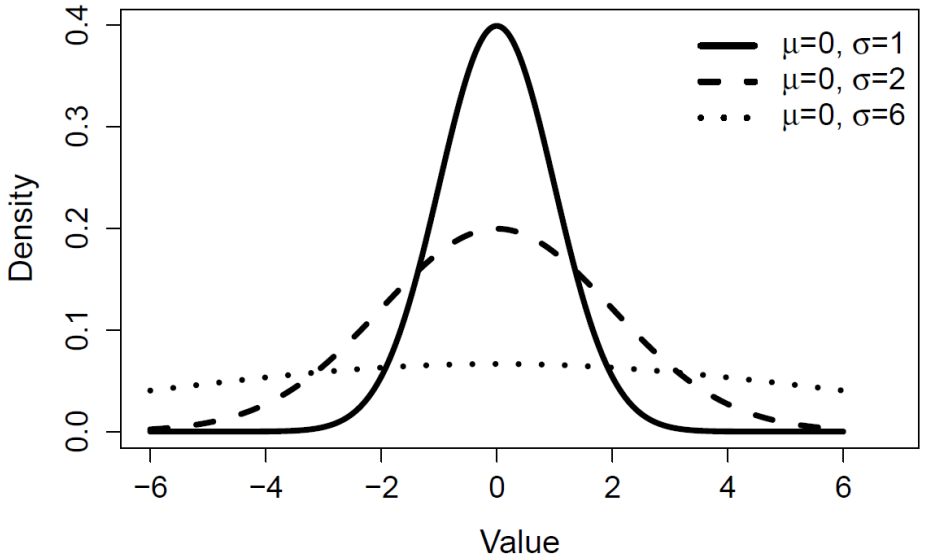
(c) Multimodal

Normal distribution

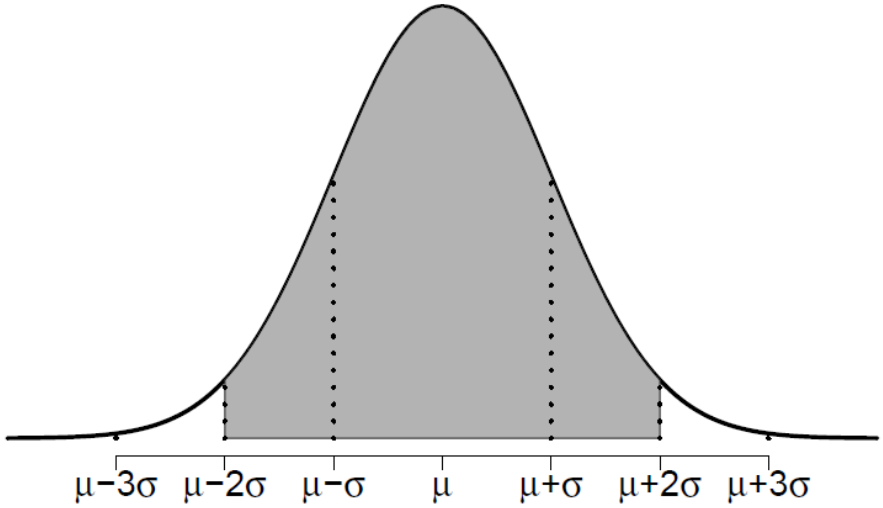


μ is the mean

σ is the standard deviation

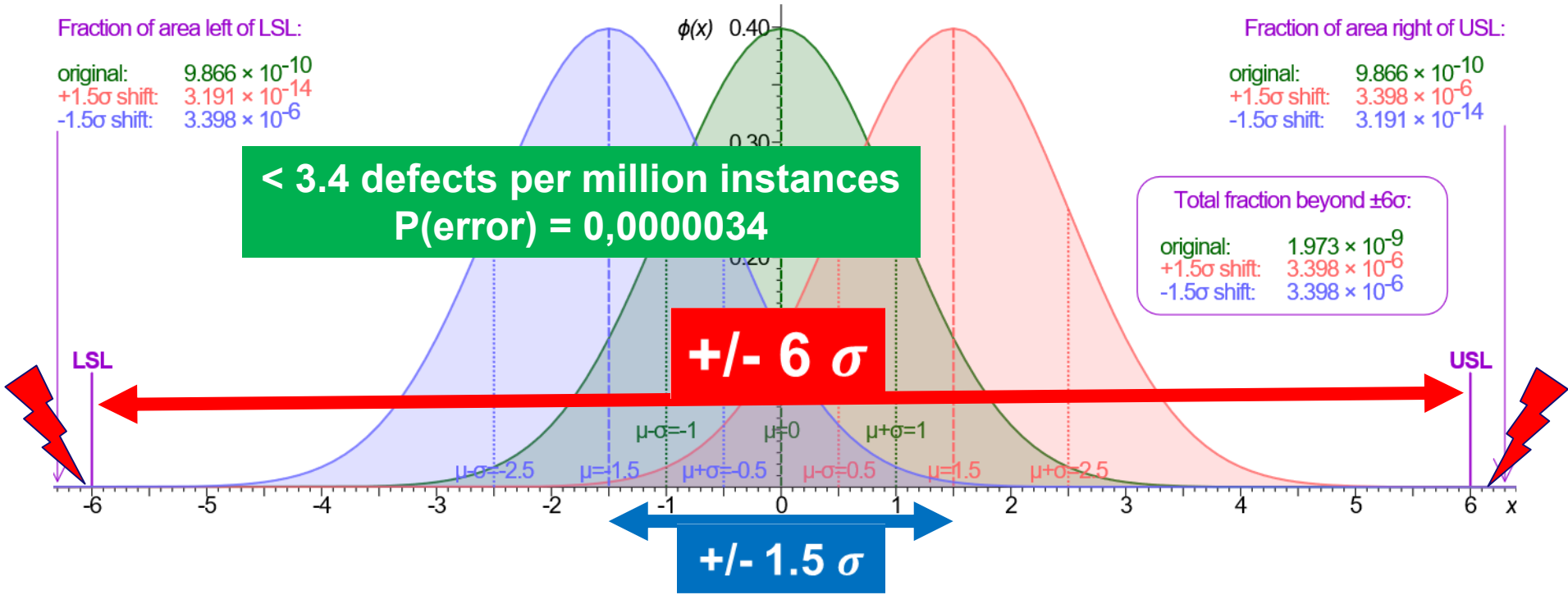


Normal distribution (68-95-99.7)



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.

Six Sigma



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.

Chair of Process and Data Science

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 feature (imputation).

Missing values

features

instances

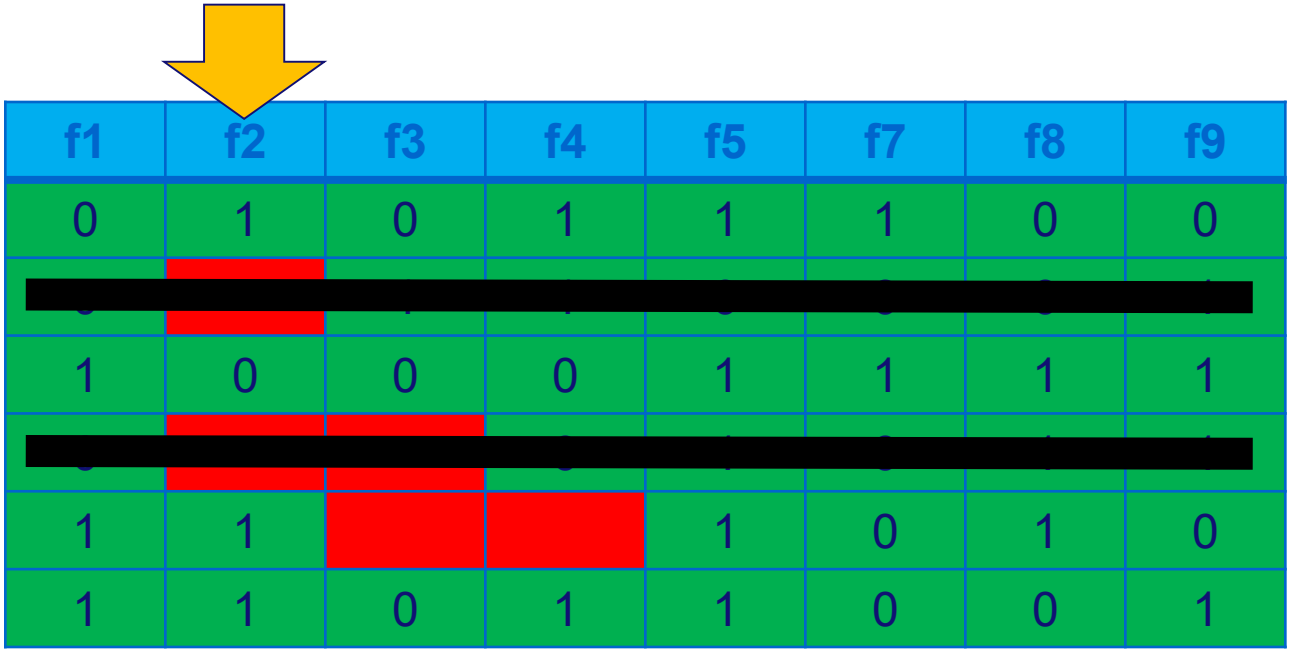
f1	f2	f3	f4	f5	f7	f8	f9
0	1	0	1	1	1	0	0
0		1	1	0	0	0	1
1	0	0	0	1	1	1	1
0			0	1	0	1	1
1	1			1	0	1	0
1	1	0	1	1	0	0	1

Handling missing values (1)

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	1	1	1	1
0			0	1	0	1	1
1	1			1	0	1	0
1	1	0	1	1	0	0	1

Remove feature completely

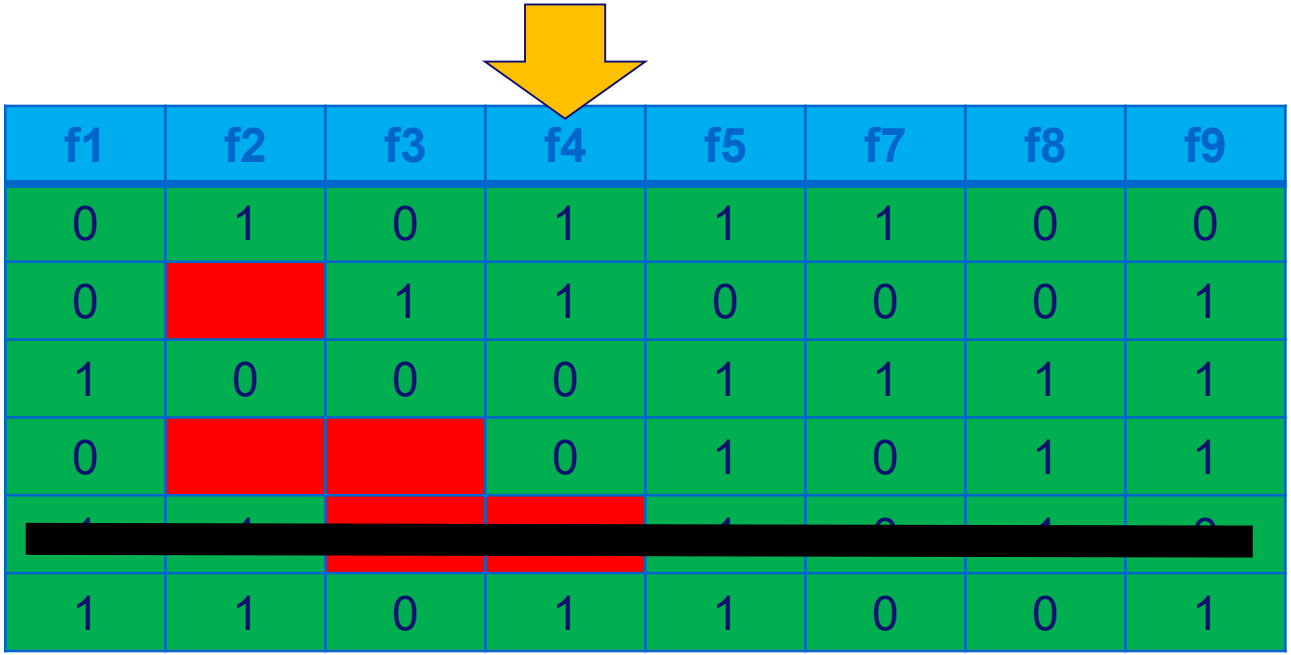
Handling missing values (2)



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

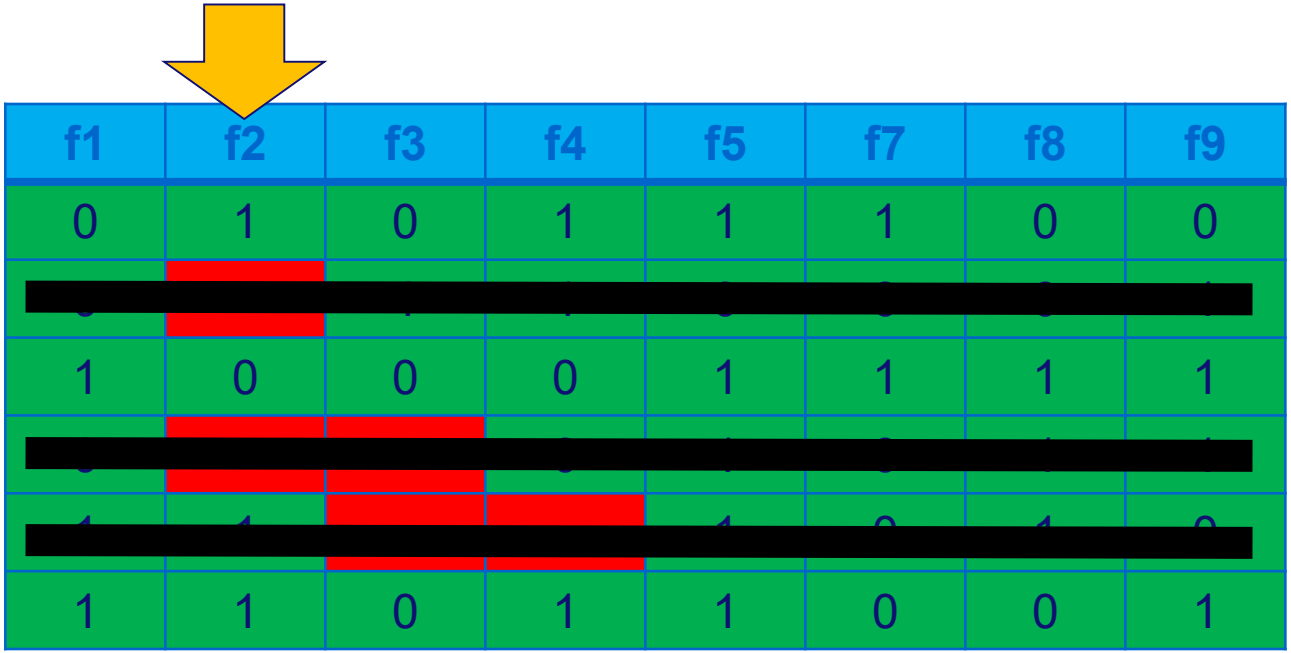
Only consider instances that have a value (per feature)

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)

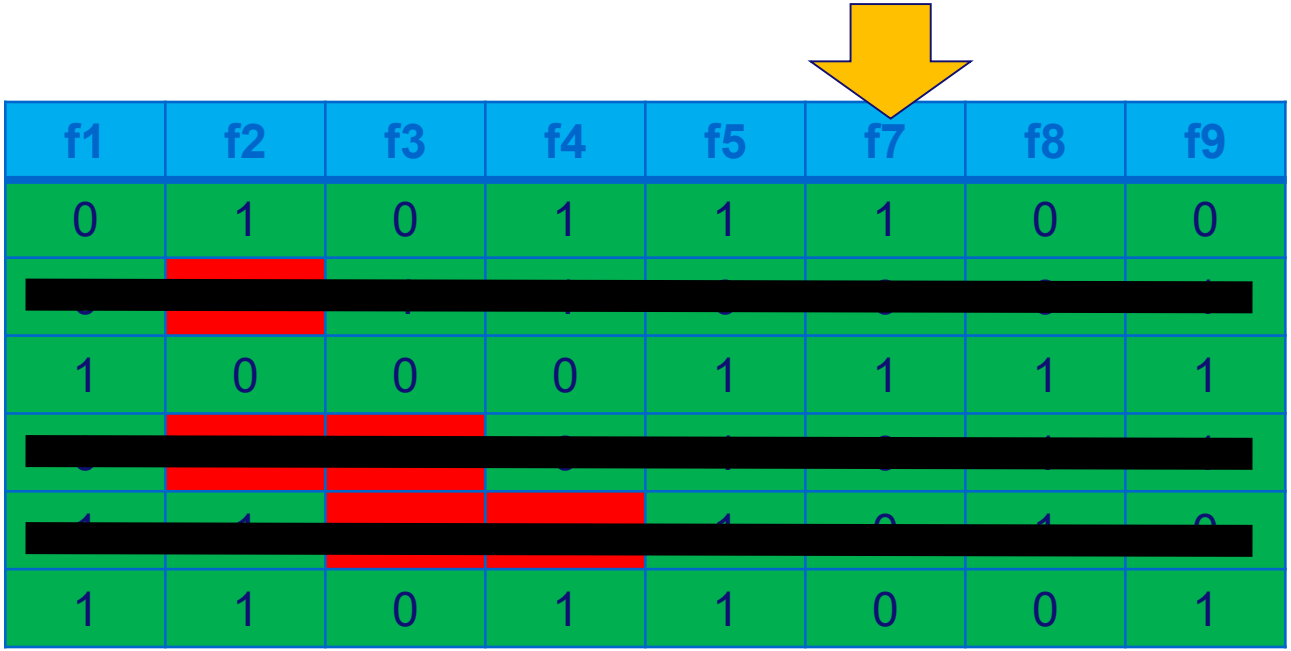
Handling missing values (3)



f1	f2	f3	f4	f5	f7	f8	f9
0	1	0	1	1	1	0	0
[redacted]							
1	0	0	0	1	1	1	1
[redacted]							
[redacted]							
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
1	0	0	0	1	1	1	1
1	1	0	1	1	0	0	1

Remove all instances that have one of the features missing

Chair of Process and Data Science

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 feature (imputation)

Impossible values

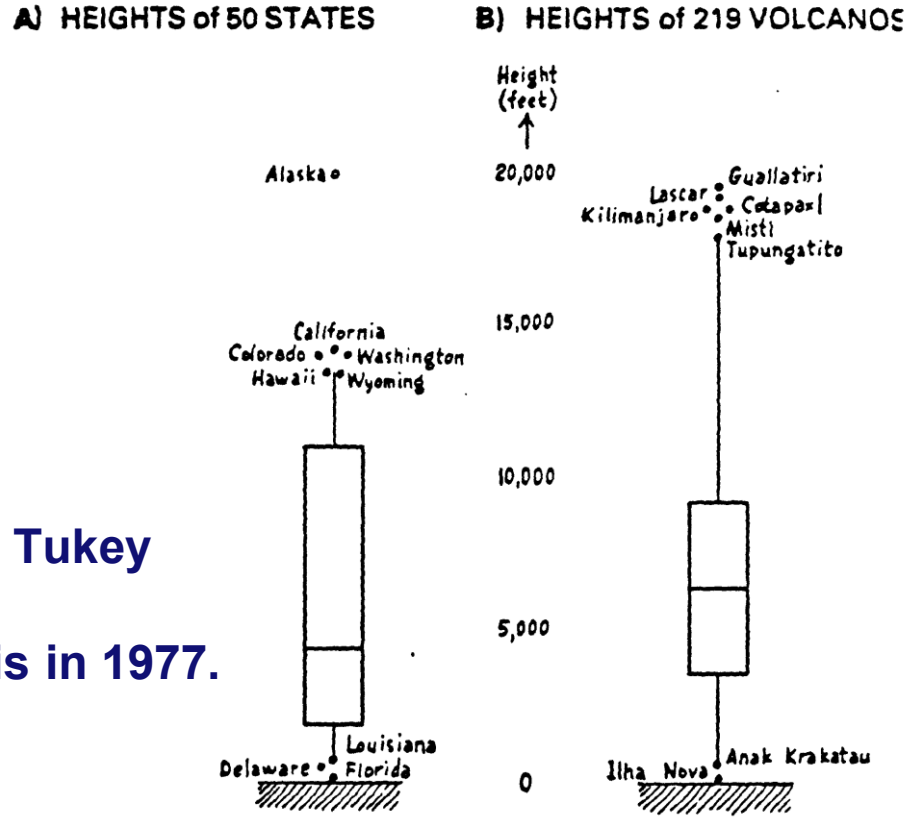
- Date: 30-2-2018
- Date: 13-13-2018
- Time: 23:61
- Color: Bllue
- Members: 6.5

Can be handled like missing values.

Unlikely values

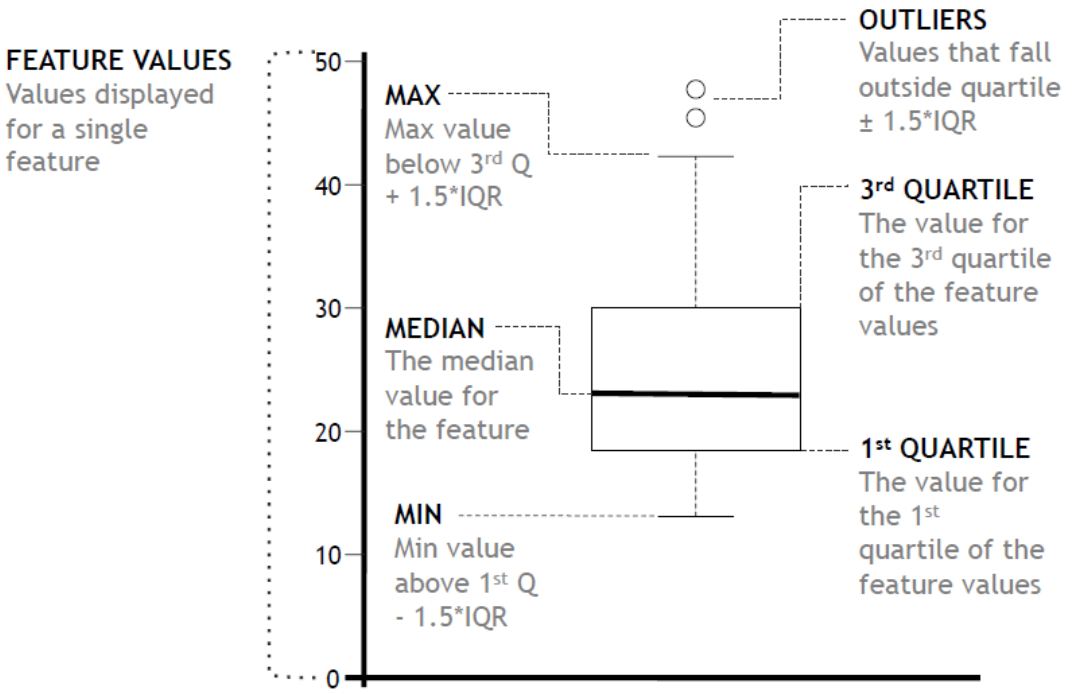
- Dates: One may expect a uniform distribution over months and days.
  - When formats DD-MM-YYYY and MM-DD-YYYY are mixed up this is no longer the case.
  - Days 1-12 are more frequent than 13-31
- Age: An age of 123 is not impossible, but unlikely.
- Price: An item priced 120.000 (rather than 120).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

Box plots



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

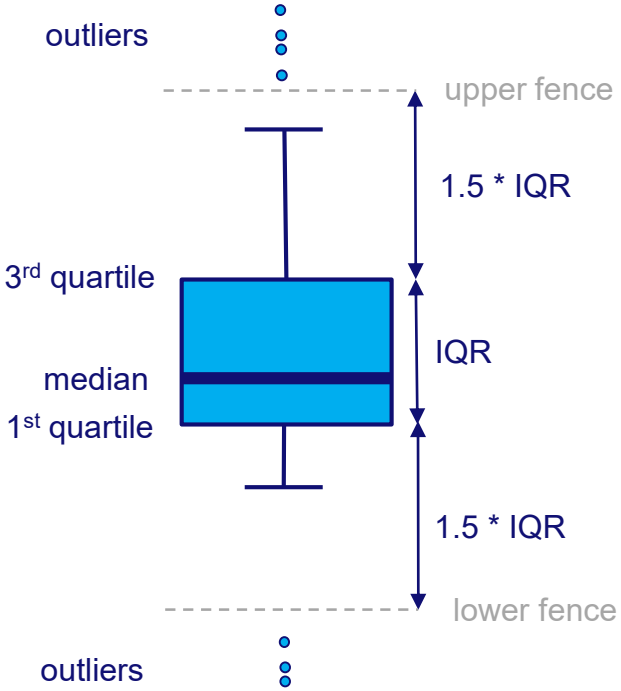
Box plots



- 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.

Chair of Process and Data Science

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.

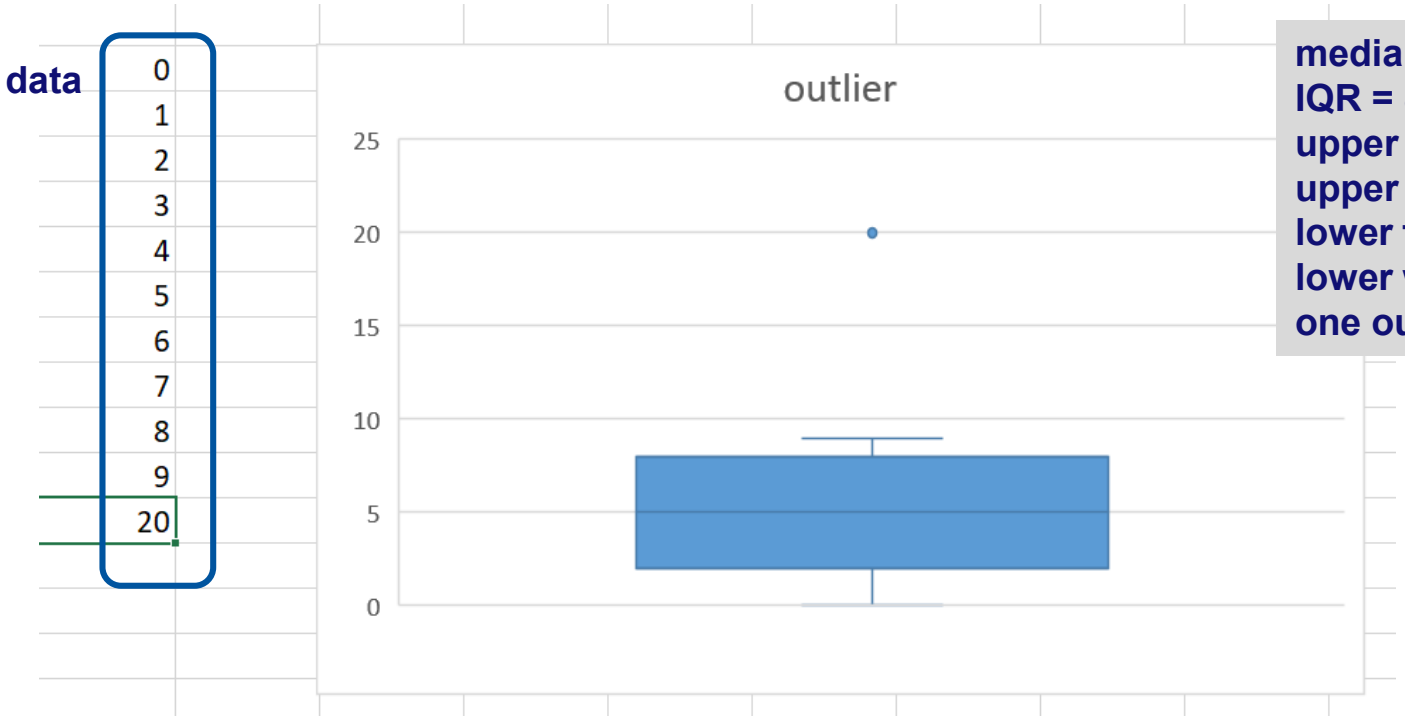
Chair of Process and Data Science

Box plot (Excel)

data
0
1
2
3
4
5
6
7
8
9
10

median = 5
IQR = 8-2 = 6
upper fence = 8 + 1.5*6 = 17
upper whisker = 10
lower fence = 2 - 1.5*6 = -7
lower whisker = 0
no outliers

Box plot (Excel)



median = 5
IQR = 8-2 = 6
upper fence = 8 + 1.5*6 = 17
upper whisker = 9
lower fence = 2 - 1.5*6 = -7
lower whisker = 0
one outlier

Box plot (Excel)

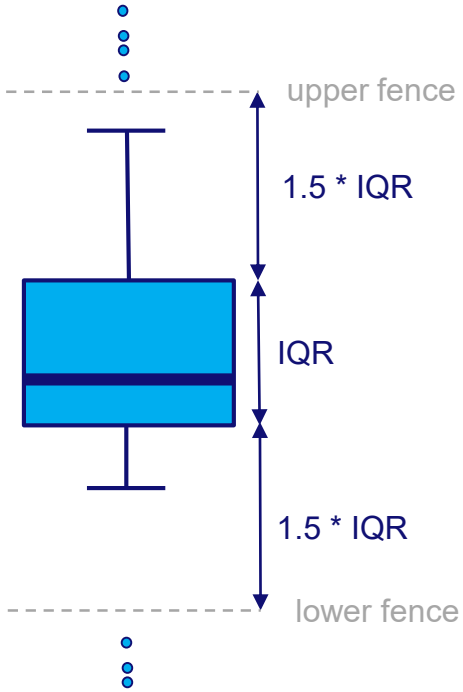
data
0
2
6
6
8
8
8
8
9
9
10

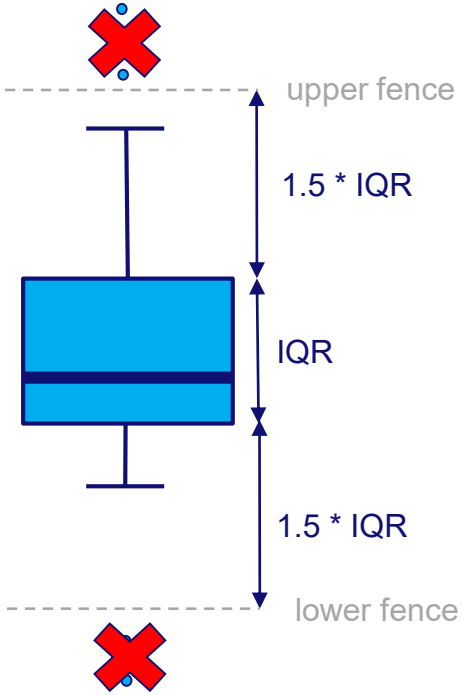
median = 8
IQR = 9-6 = 3
upper fence = 9 + 1.5*3 = 13.5
upper whisker = 10
lower fence = 6 - 1.5*3 = 1.5
lower whisker = 2
one outlier

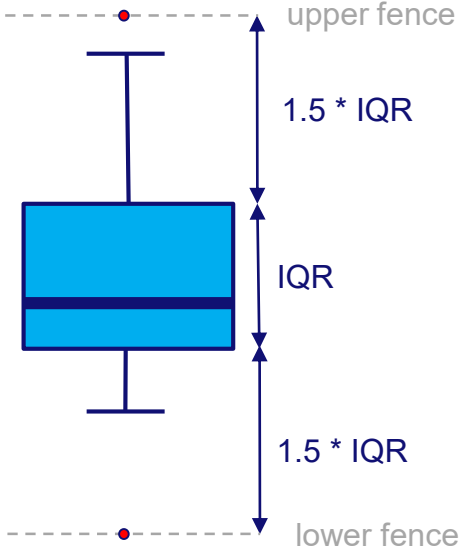
Handling outliers

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

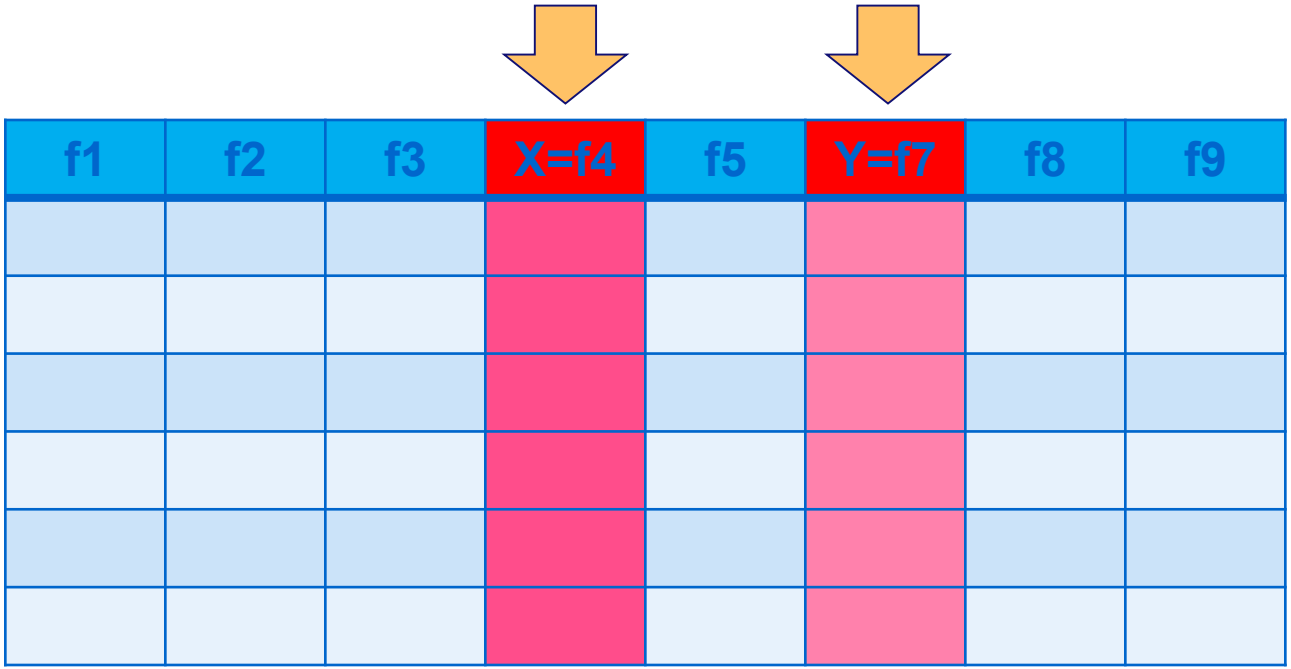






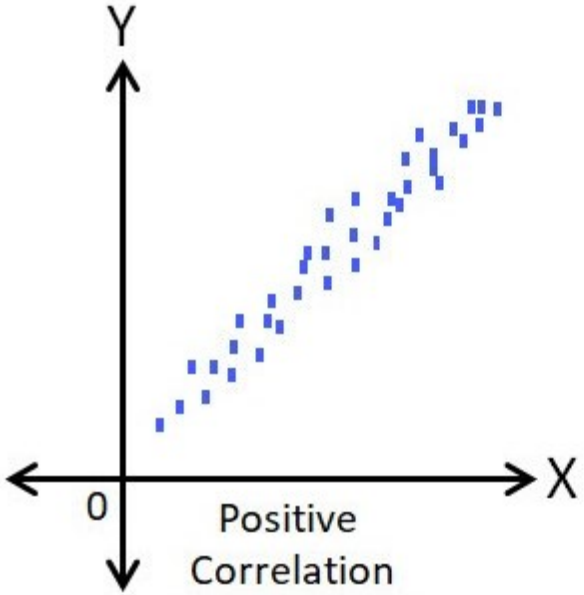
Showing relations among features

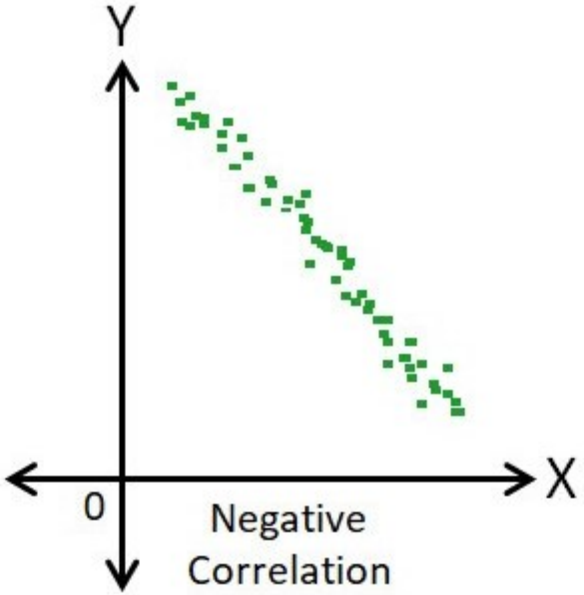
Let’s focus on two features

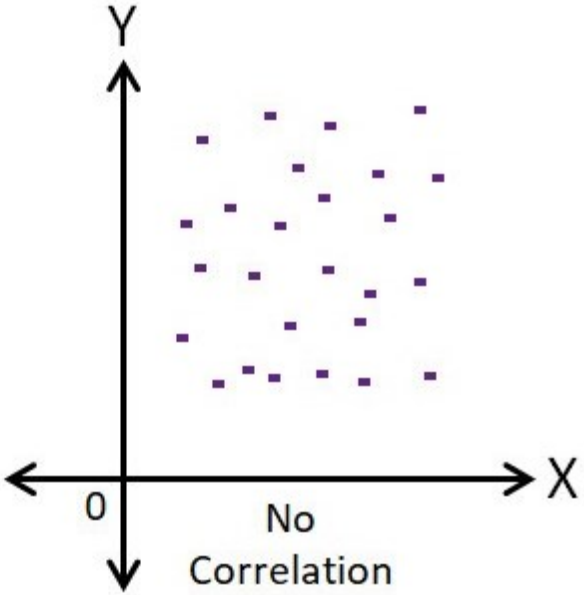


Chair of Process and Data Science

Scatter Plots & Correlation





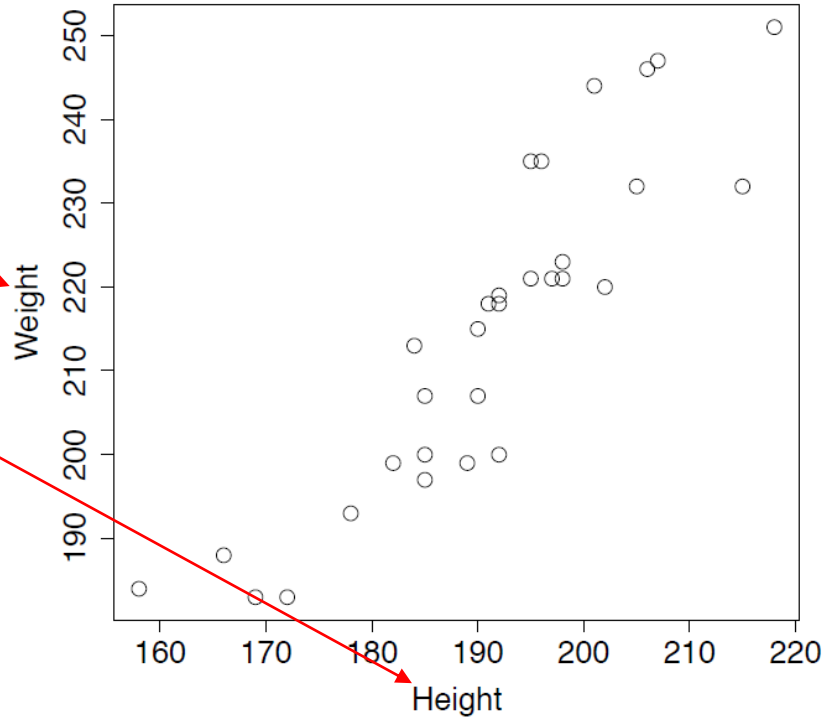


Chair of Process and Data Science

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

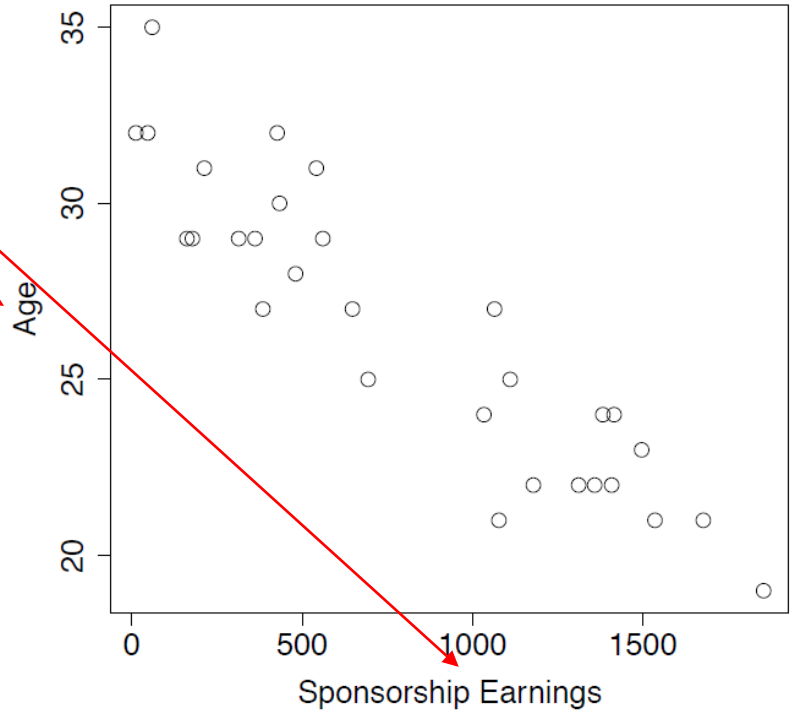
basketball players



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

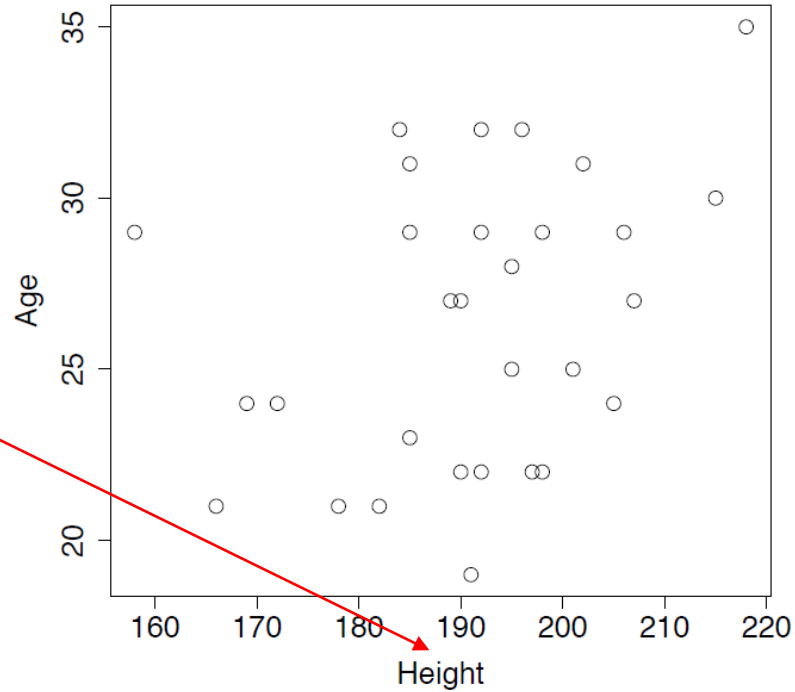
Data set in book (2/3)

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

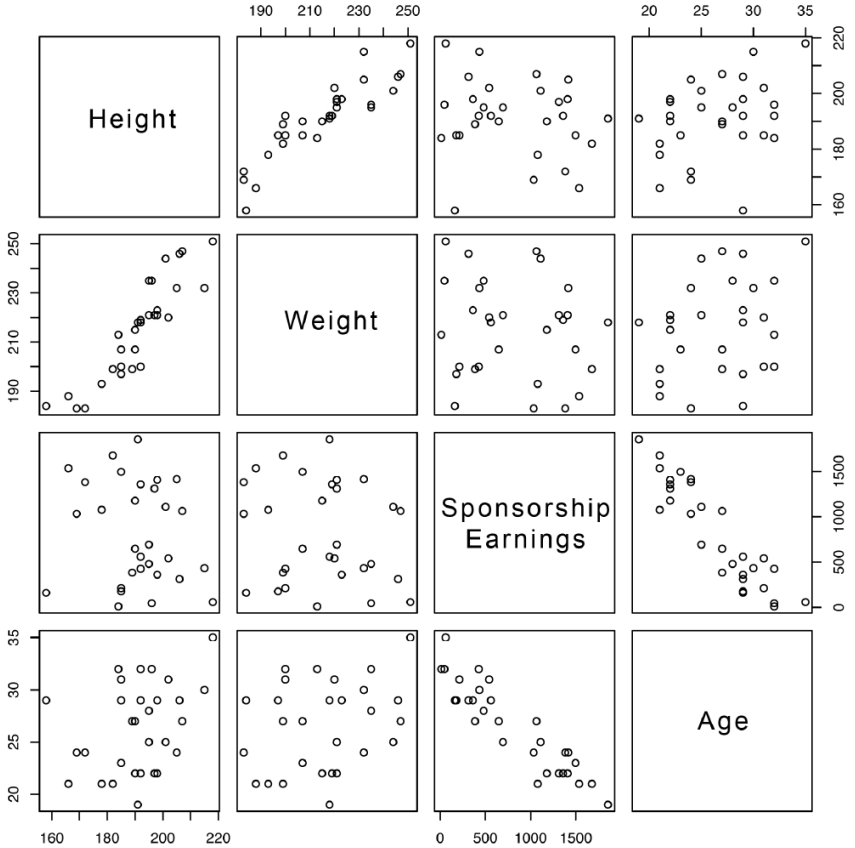


Data set in book (3/3)

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

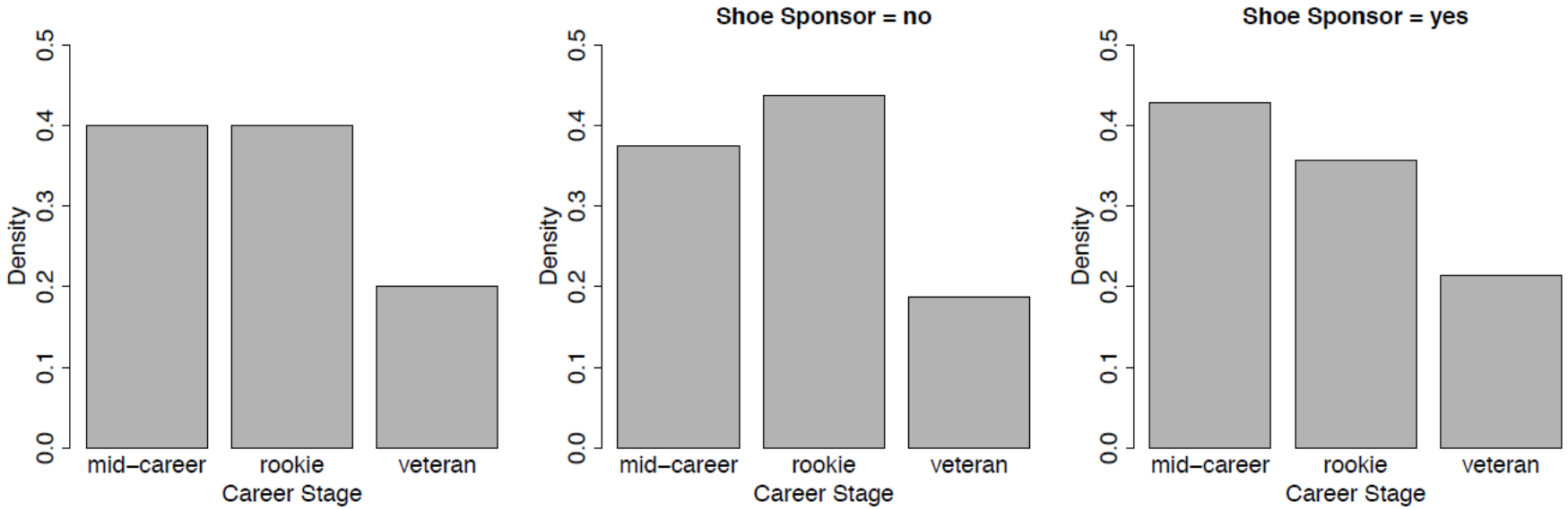


Scatter Plot Matrix (SPLOM)



Chair of Process and Data Science

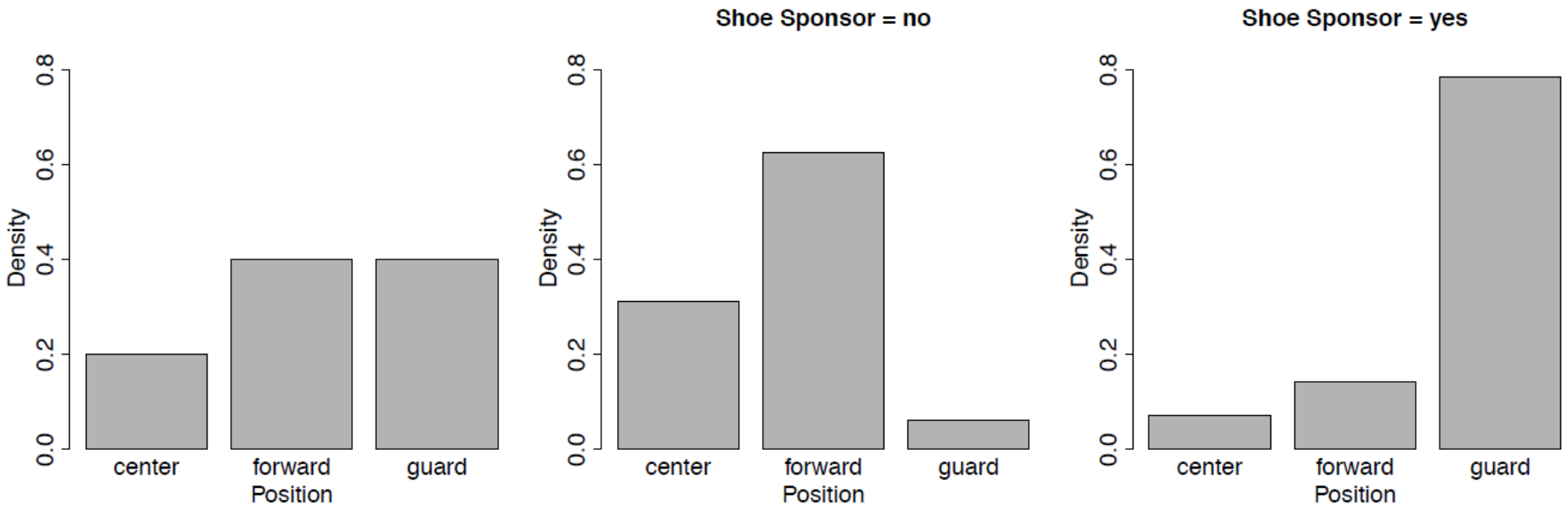
Collection of small multiple bar plots



No strong relation between shoe sponsor and career stage.

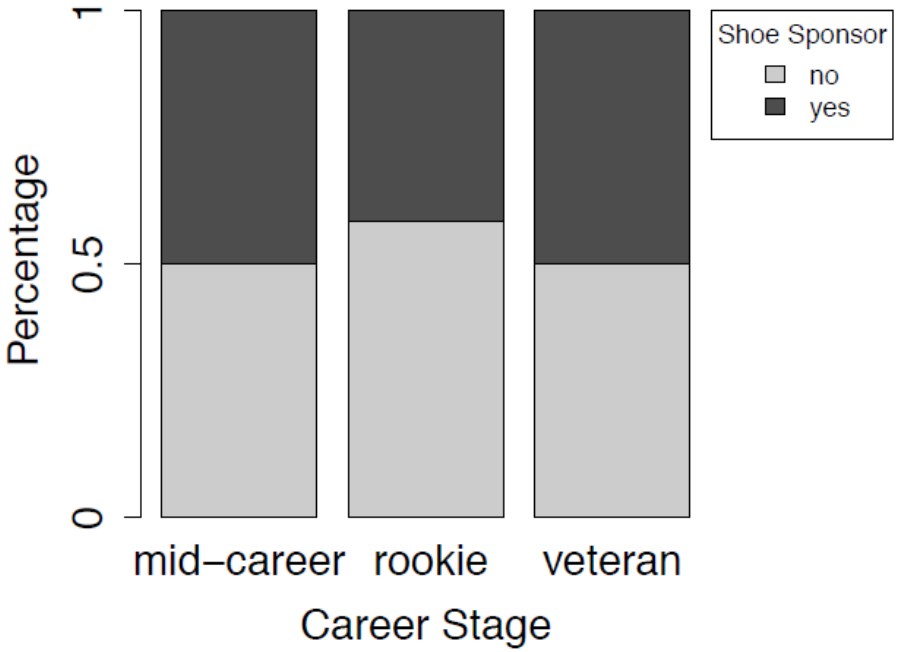
Chair of Process and Data Science

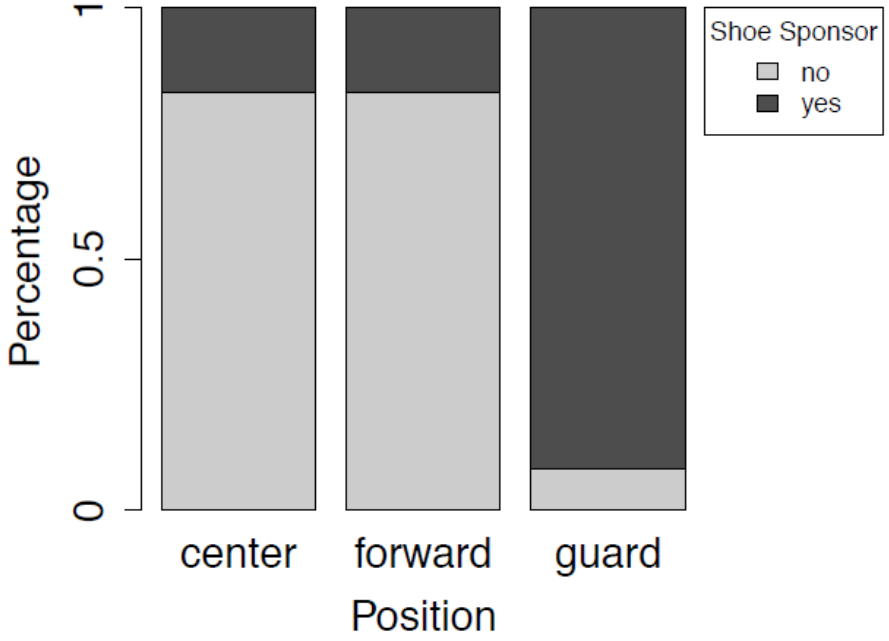
Collection of small multiple bar plots



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

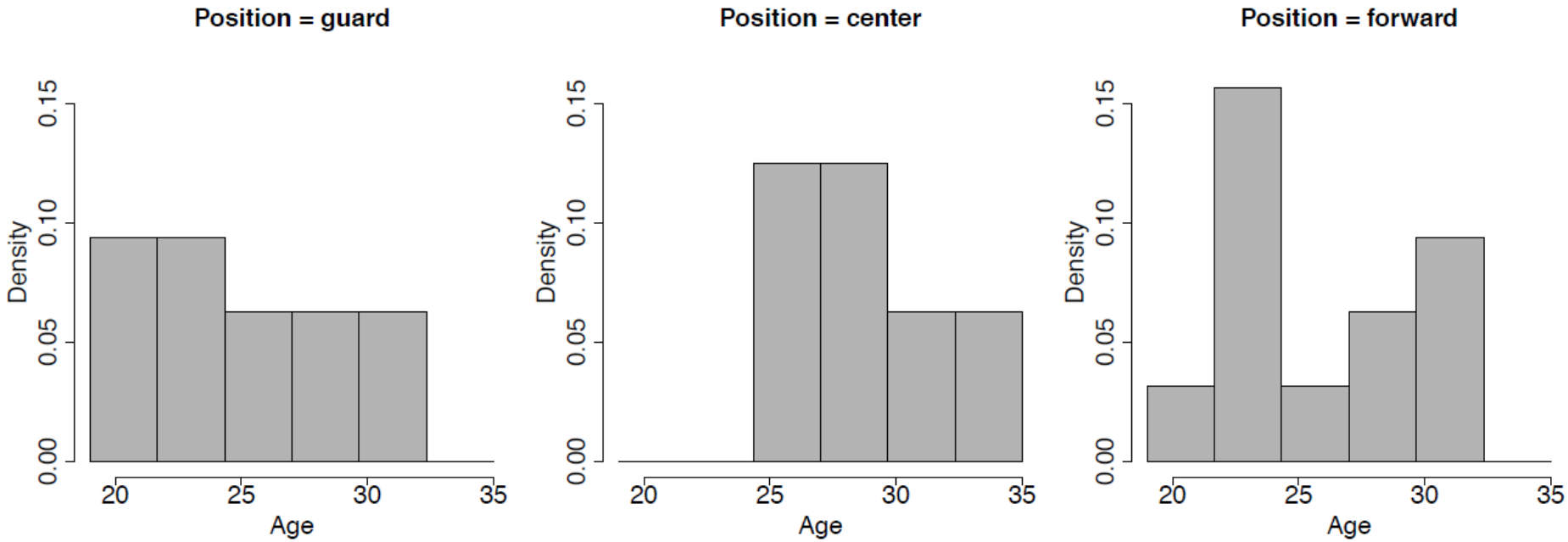
Stacked bar plot (showing percentages)





Chair of Process and Data Science

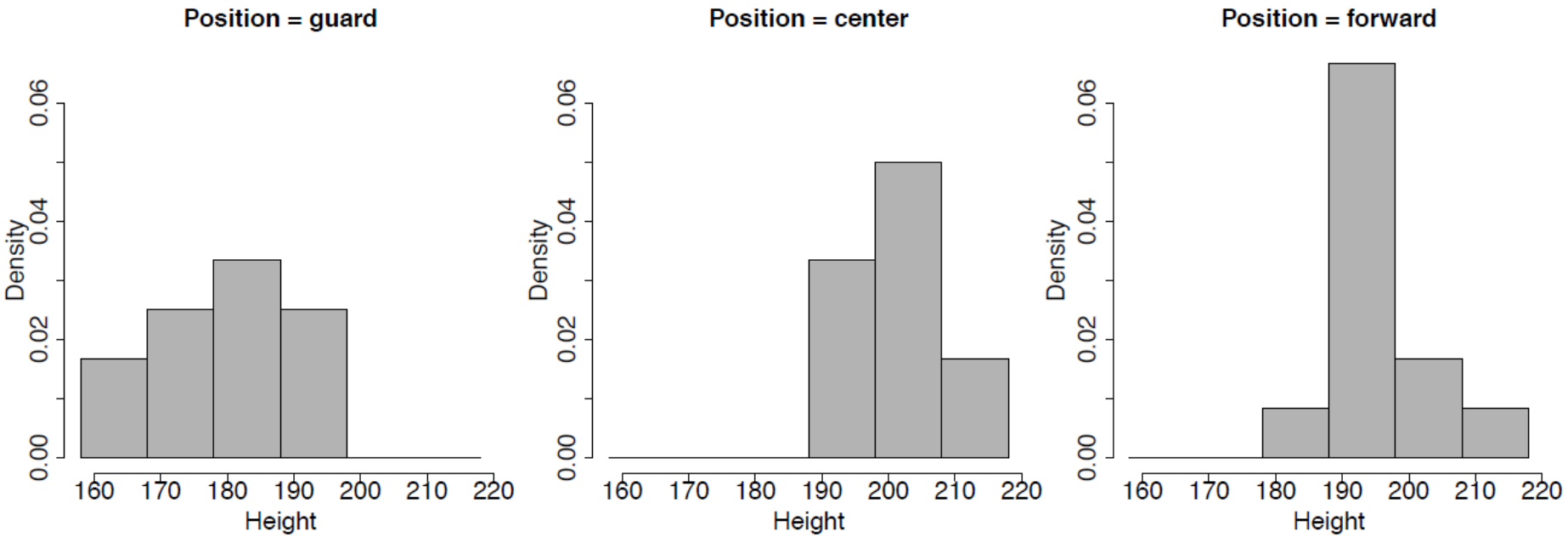
Collection of small multiple histograms



Age has now 6 classes.

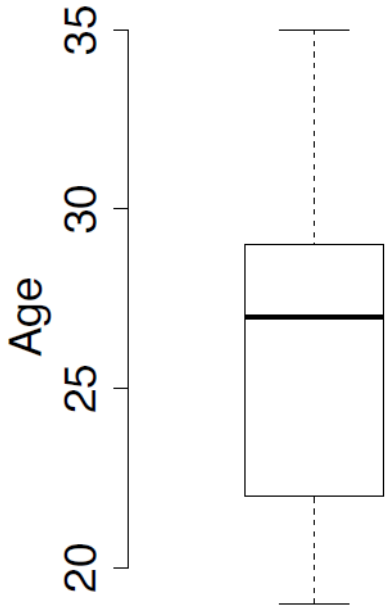
Chair of Process and Data Science

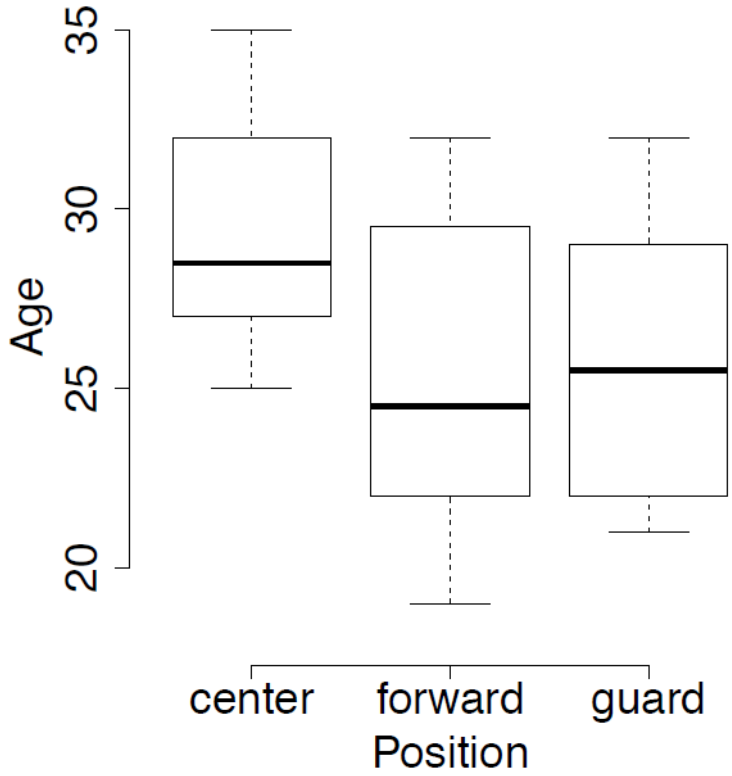
Collection of small multiple histograms



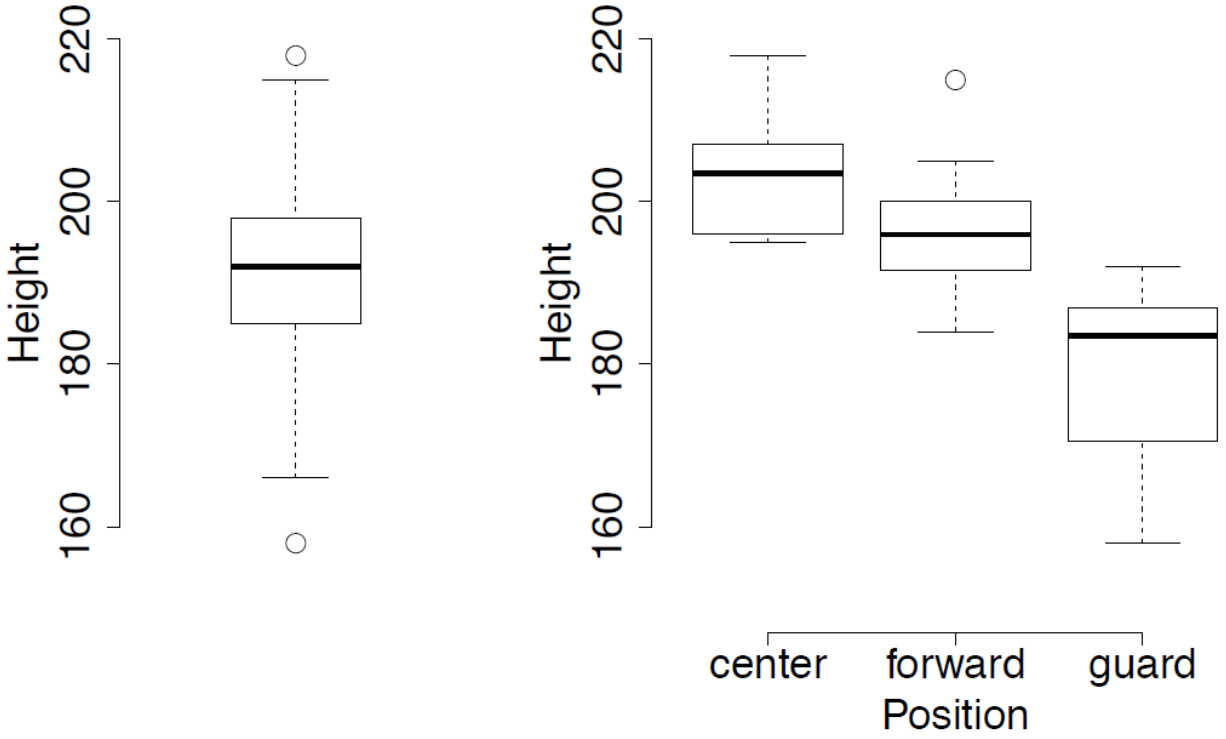
Height has now 6 classes.

Collections of box plots



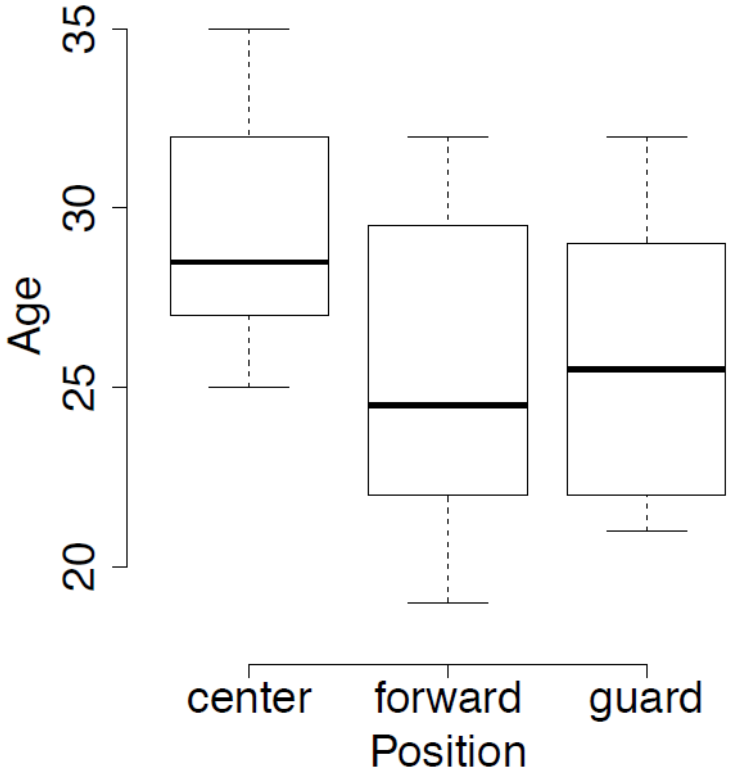


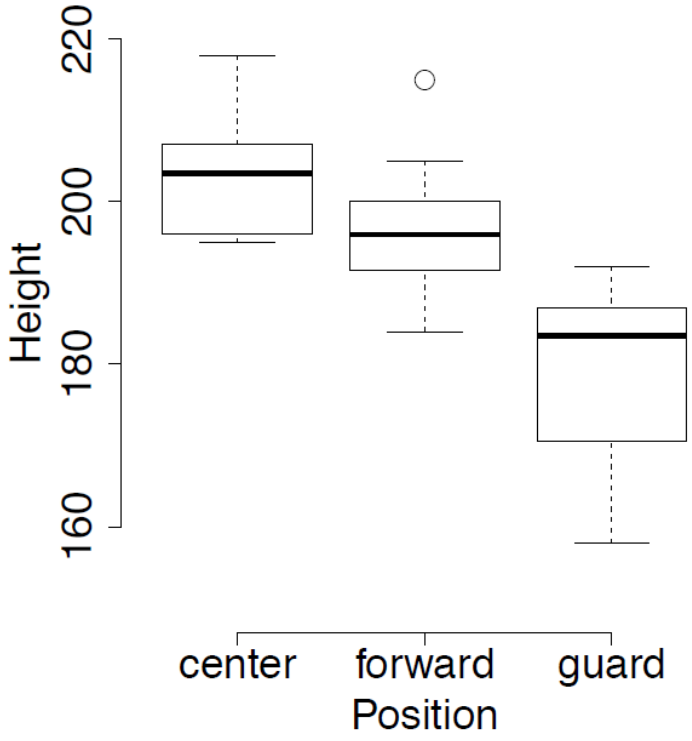
Collections of box plots



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, .. a_n)

Some basic descriptive statistics

standard deviation

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

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

(based on n values a_1, .. a_n)

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

Sample covariance

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

$$[-\infty, \infty]$$

+ & + ⇒ +

+ & - ⇒ -

- & + ⇒ -

- & - ⇒ +

(based on *n* pairs of values (a_1,b_1), .. (a_n,b_n))

Correlation

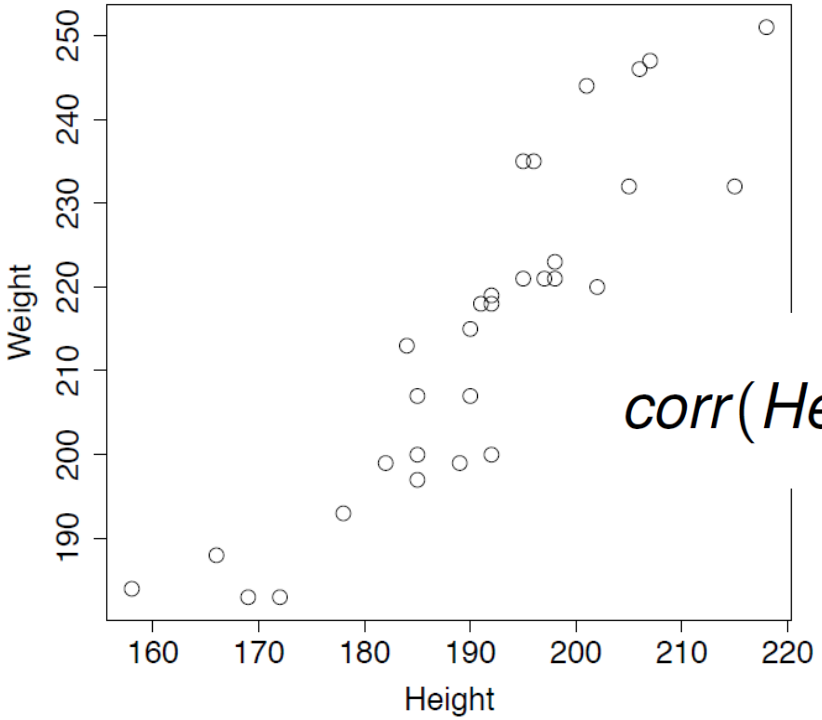
$$corr(a,b) = \frac{cov(a,b)}{sd(a) \times sd(b)} \qquad [-1,1]$$

> 0 positive
< 0 negative
≈ 0 independent

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

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

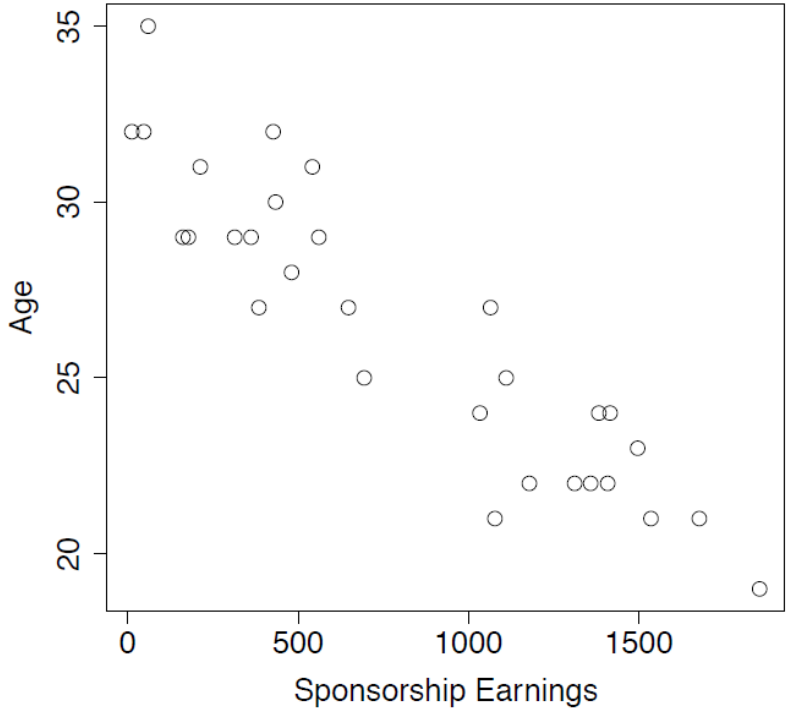
Positive correlation



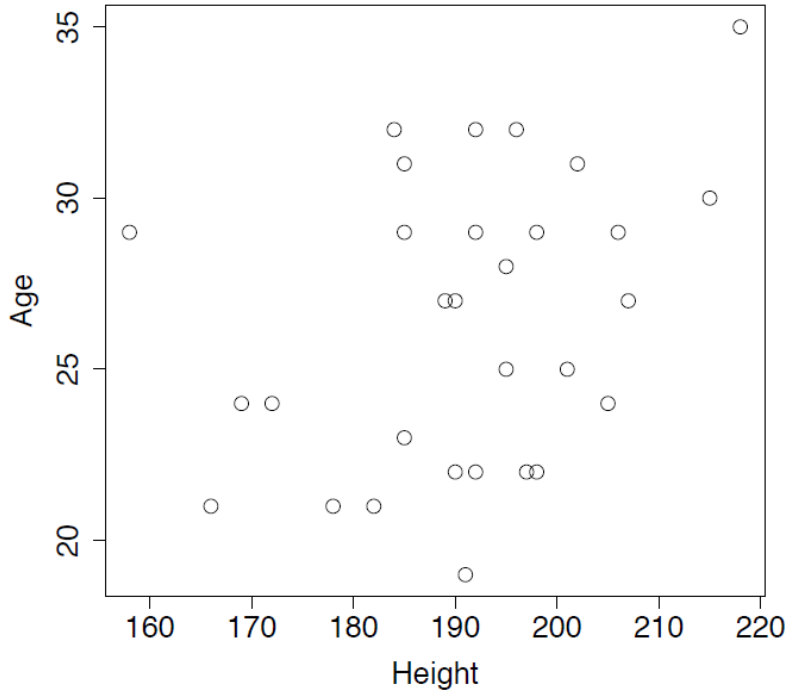
$$corr(Height, Weight) = \frac{241.72}{13.6 \times 19.8} = 0.898$$

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

Negative correlation



Weaker positive correlation



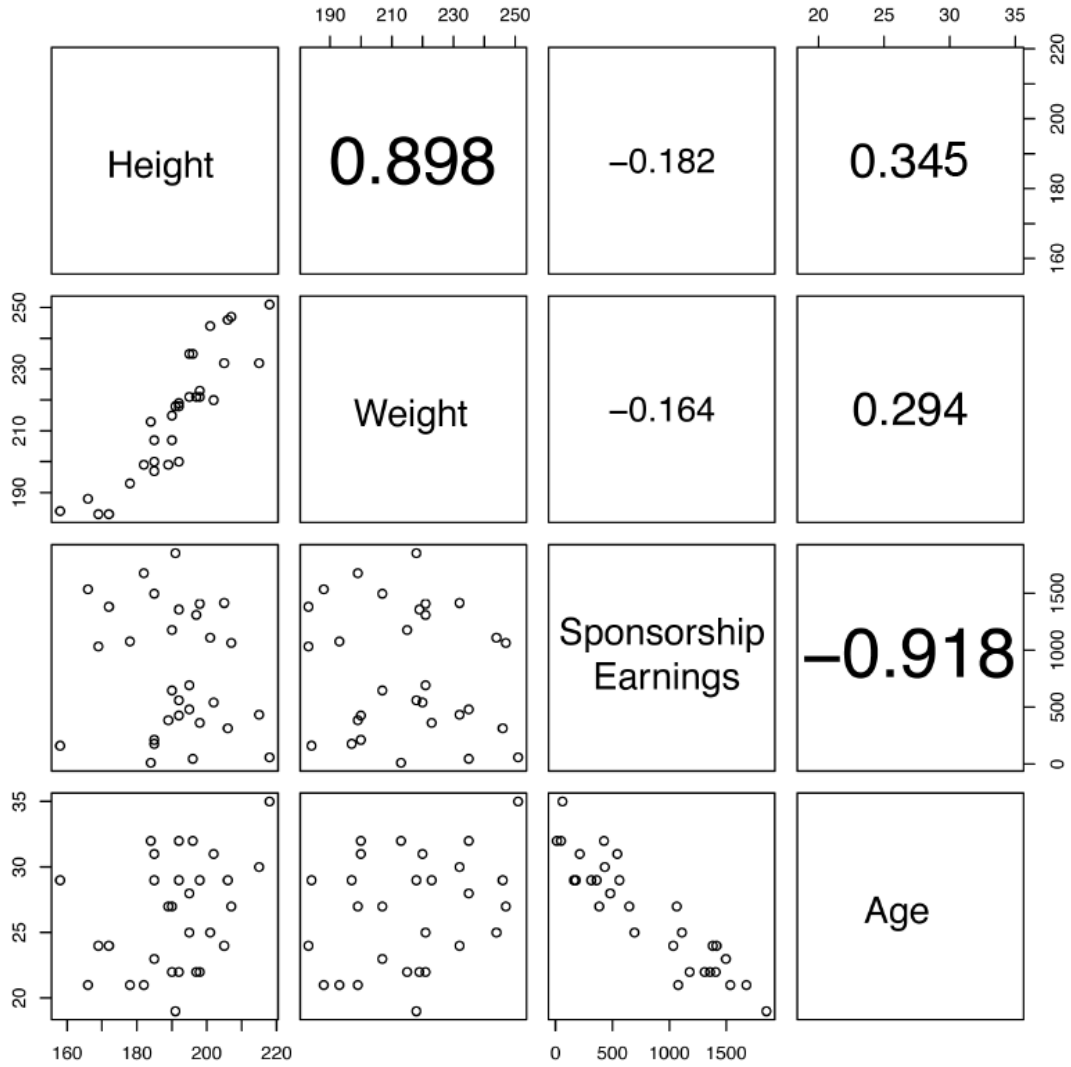
$$corr(Height, Age) = \frac{19.7}{13.6 \times 4.2} = 0.345$$

Correlation matrix

$$\underset{\{a,b,\ldots,z\}}{\textit{correlation matrix}} = \begin{bmatrix} corr(a,a) & corr(a,b) & \cdots & corr(a,z) \\ corr(b,a) & corr(b,b) & \cdots & corr(b,z) \\ \vdots & \vdots & \ddots & \vdots \\ corr(z,a) & corr(z,b) & \cdots & corr(z,z) \end{bmatrix}$$

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.**

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

maps values onto $[low, high]$

Standard score

- **Standard score uses the standard deviation to normalize.**

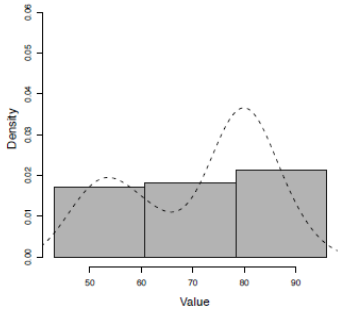
$$a_i' = \frac{a_i - \bar{a}}{sd(a)}$$

$$[-\infty, \infty]$$

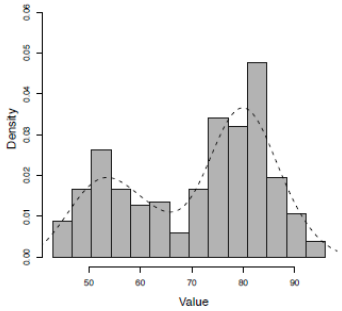
Binning

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

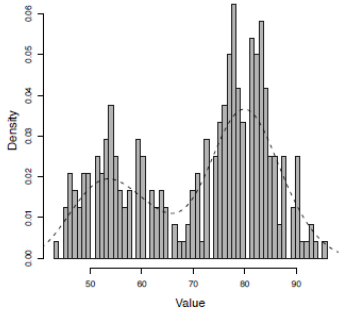
Number of bins



(e) 3 bins



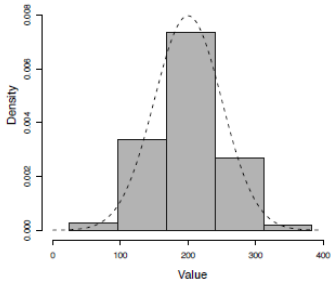
(f) 14 bins



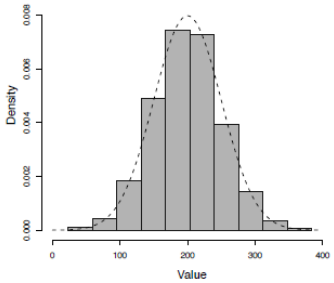
(g) 60 bins

- Too few bins may leads 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).

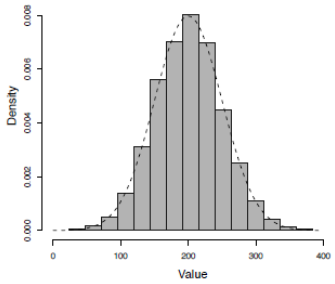
Equal-width binning



(h) 5 Equal-width bins



(i) 10 Equal-width bins

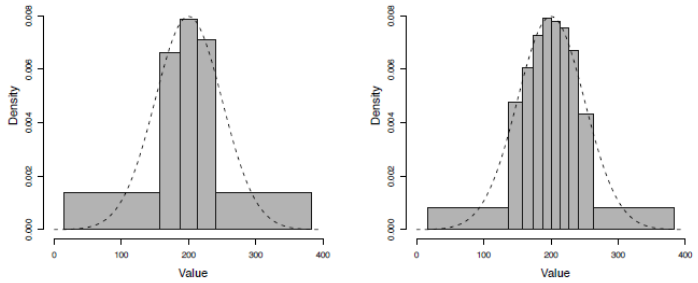


(j) 15 Equal-width bins

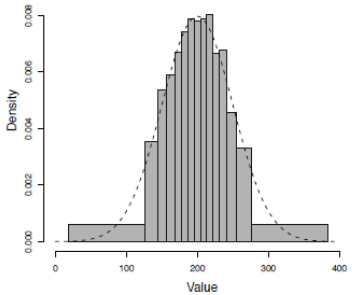
Bins have a fixed width, but the number of items per bin may vary greatly.

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

Equal-frequency binning



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

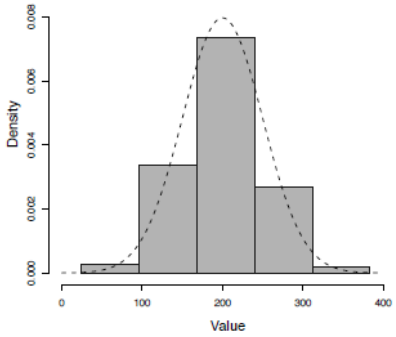


(m) 15 Equal-frequency bins

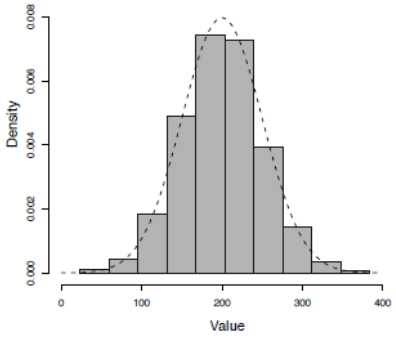
Bins have a variable width, but the number of items per bin is fixed.

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

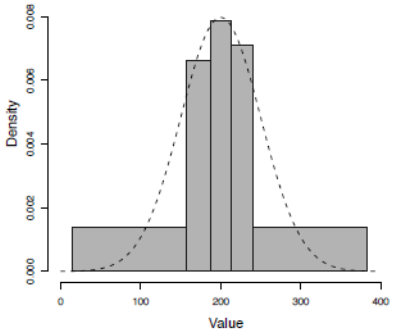
Chair of Process and Data Science



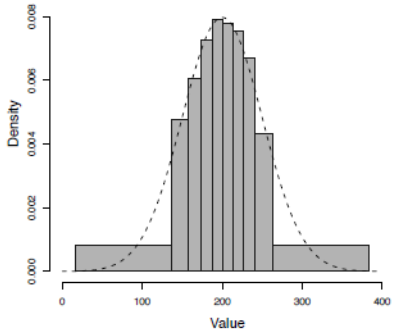
(h) 5 Equal-width bins



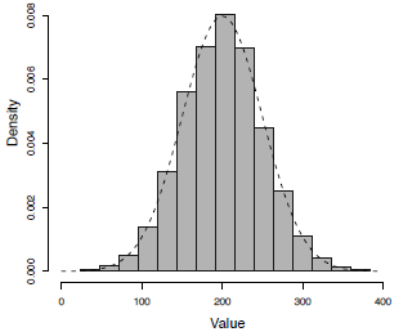
(i) 10 Equal-width bins



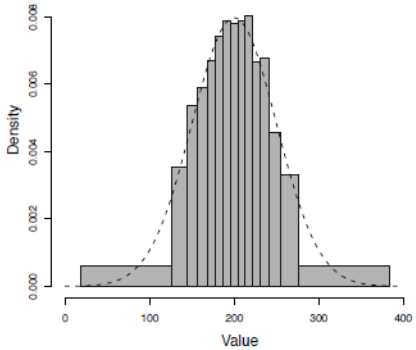
(k) 5 Equal-frequency bins



(l) 10 Equal-frequency bins



(j) 15 Equal-width bins

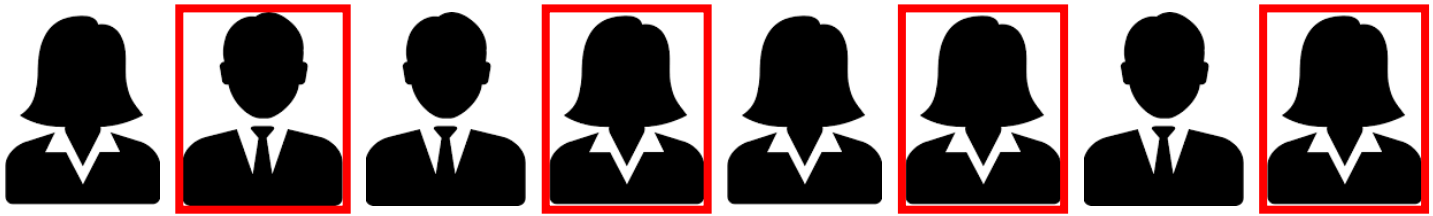


(m) 15 Equal-frequency bins

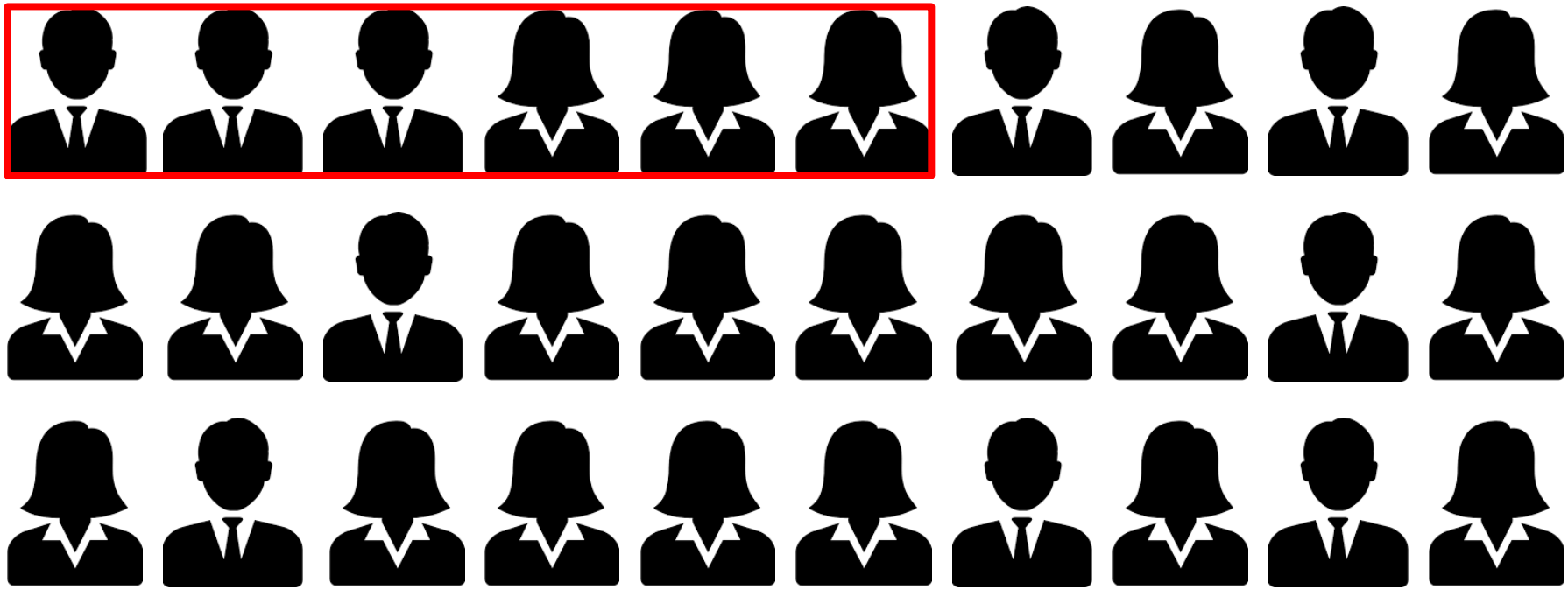
Surface of a bin reflects number of items in bin.

Sampling

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

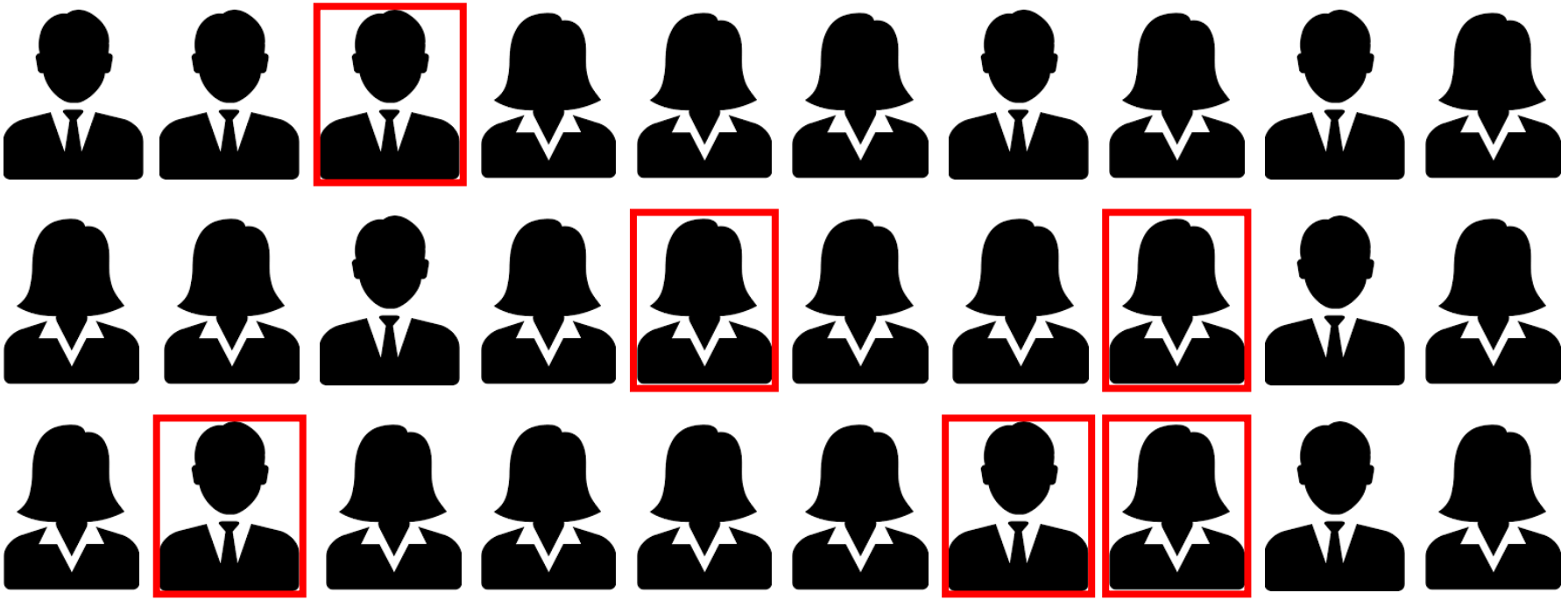


Top sampling (first n instances)



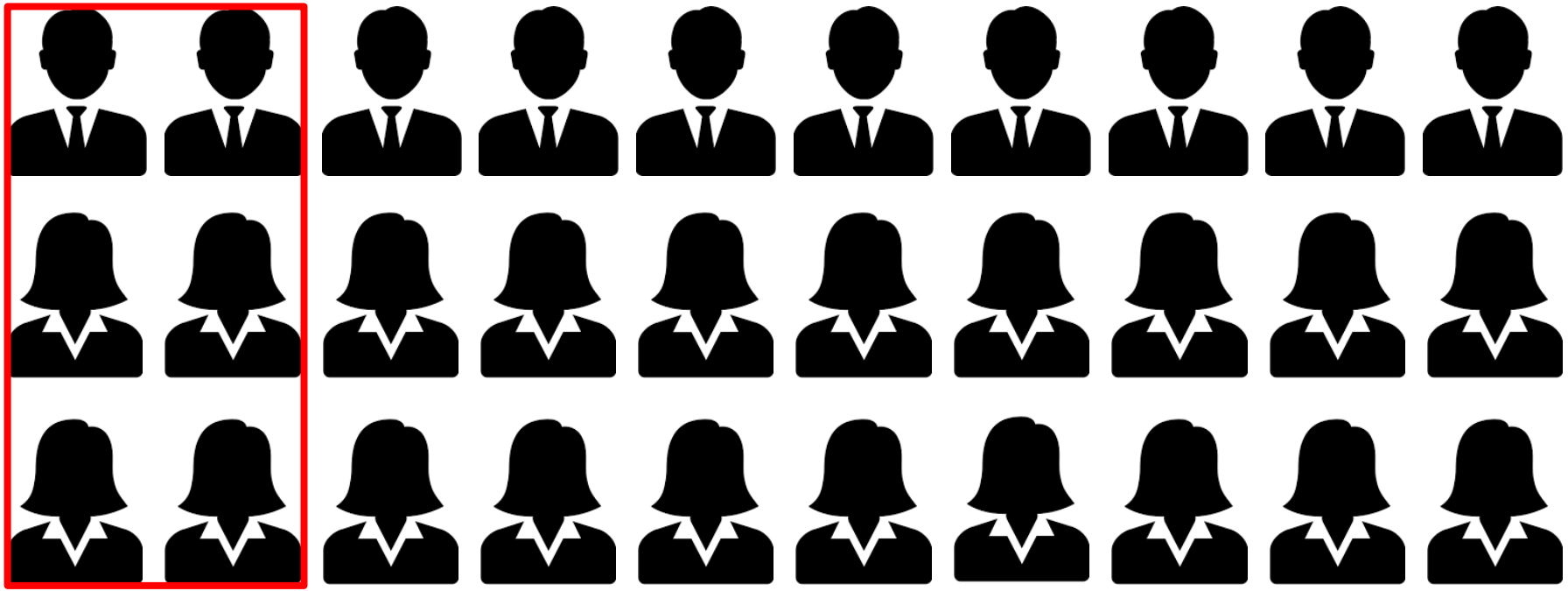
Chair of Process and Data Science

Random sampling

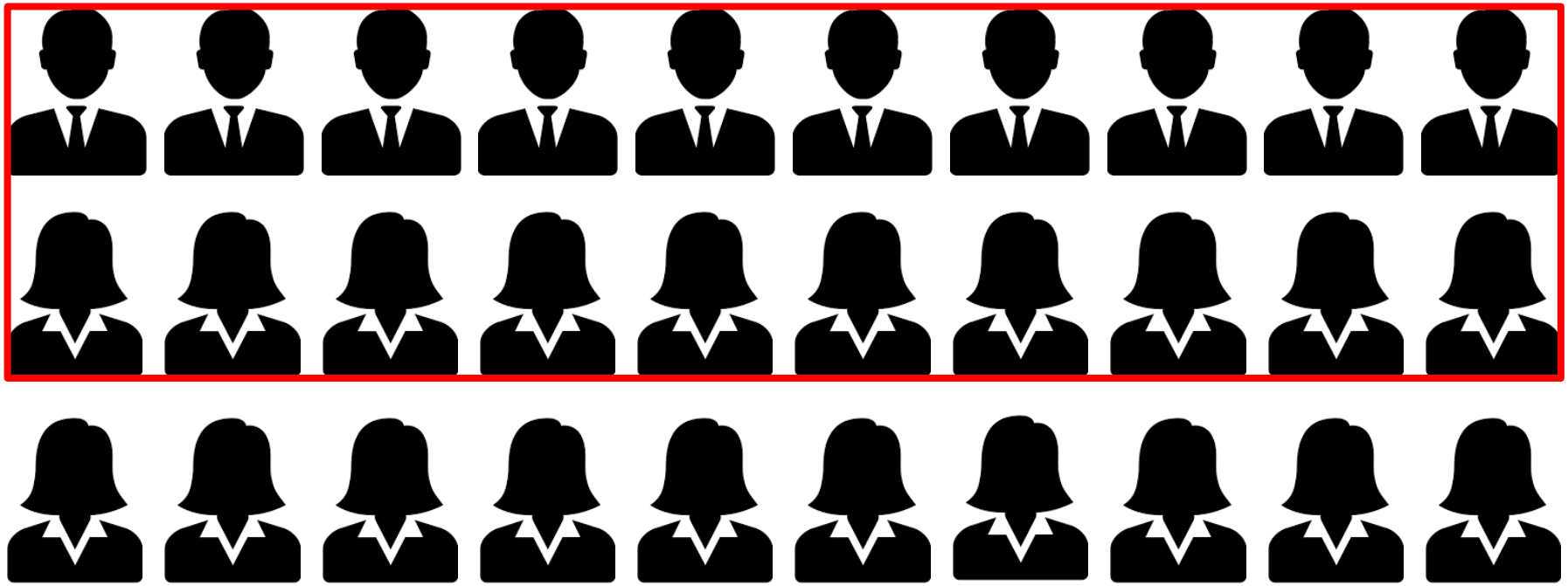


Chair of Process and Data Science

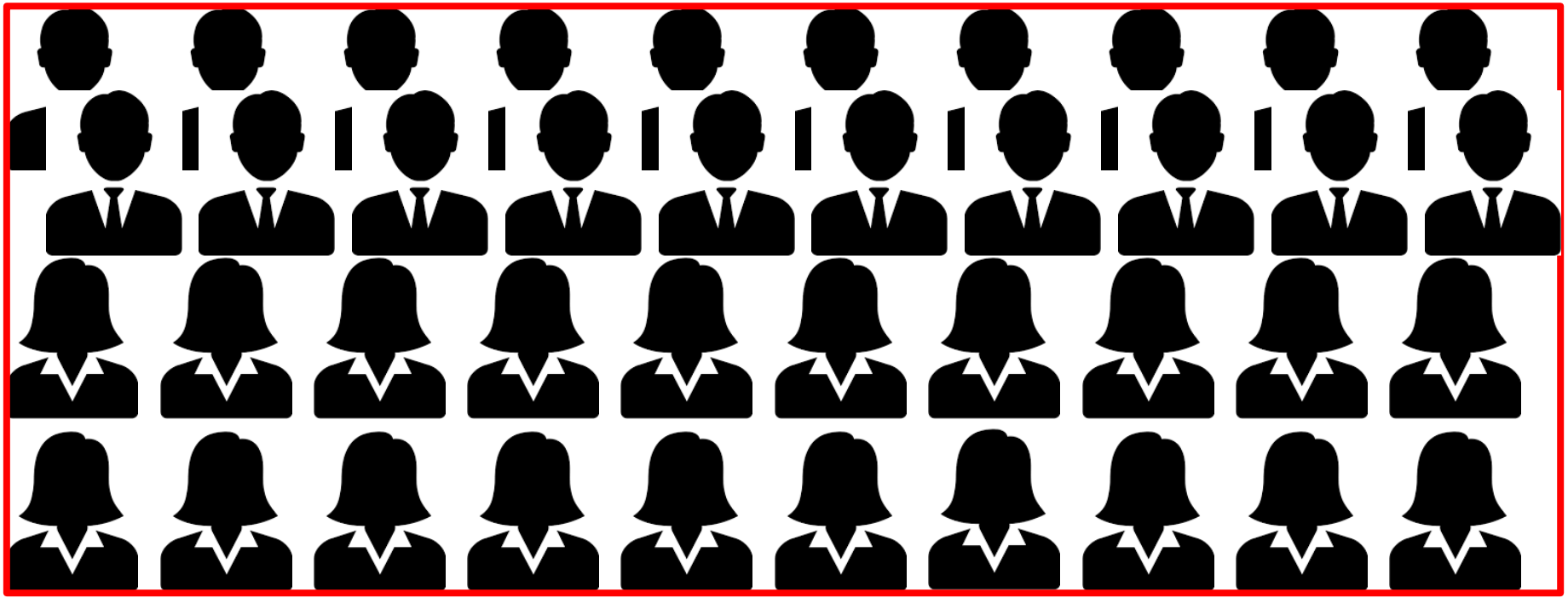
Stratified sampling (ensure that relative frequencies are maintained, e.g., by taking the same percentage from every group)



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



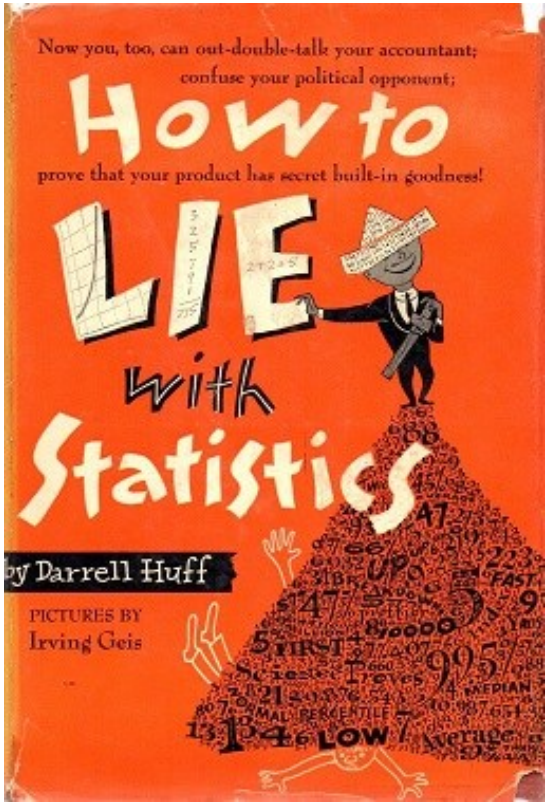
Over-sampling (ensure a balance by possible duplication of under-represented instances)



Chair of Process and Data Science

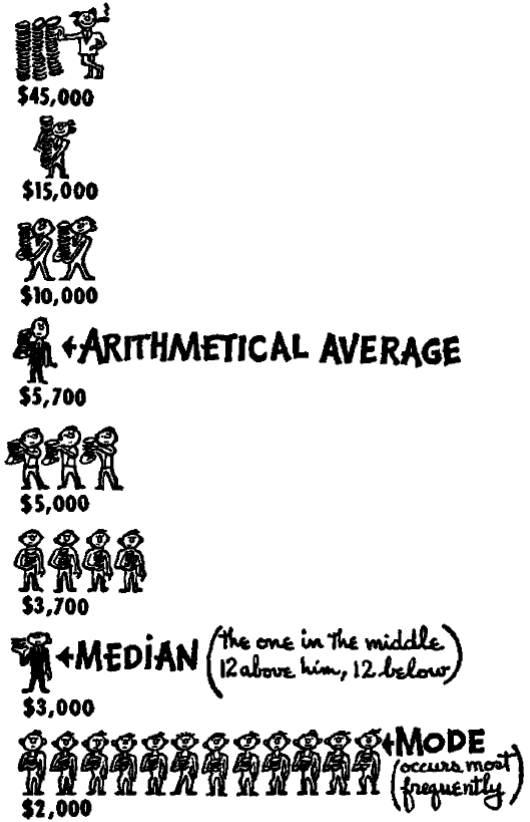
Good and poor visualizations

Examples from “How to Lie with Statistics”



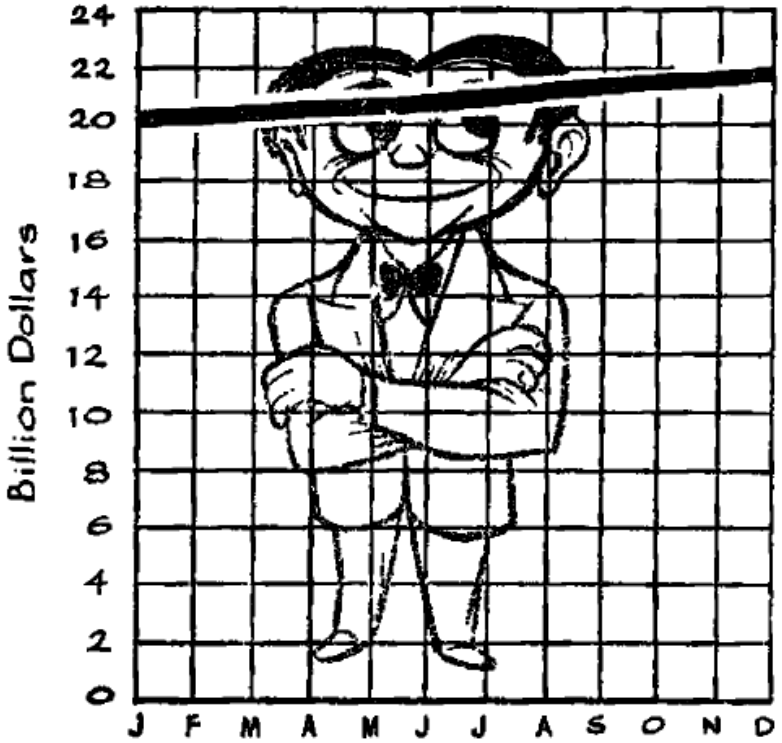
How to Lie with Statistics is a book written by Darrell Huff in 1954 with many examples that are still relevant after more than 60 years.

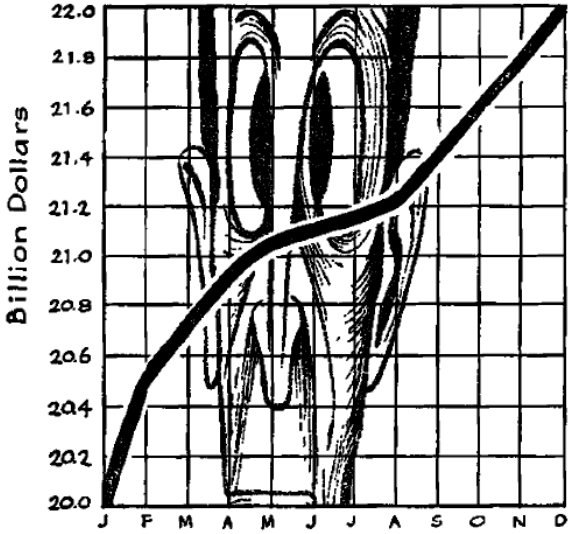
Choosing the “middle”



Chair of Process and Data Science

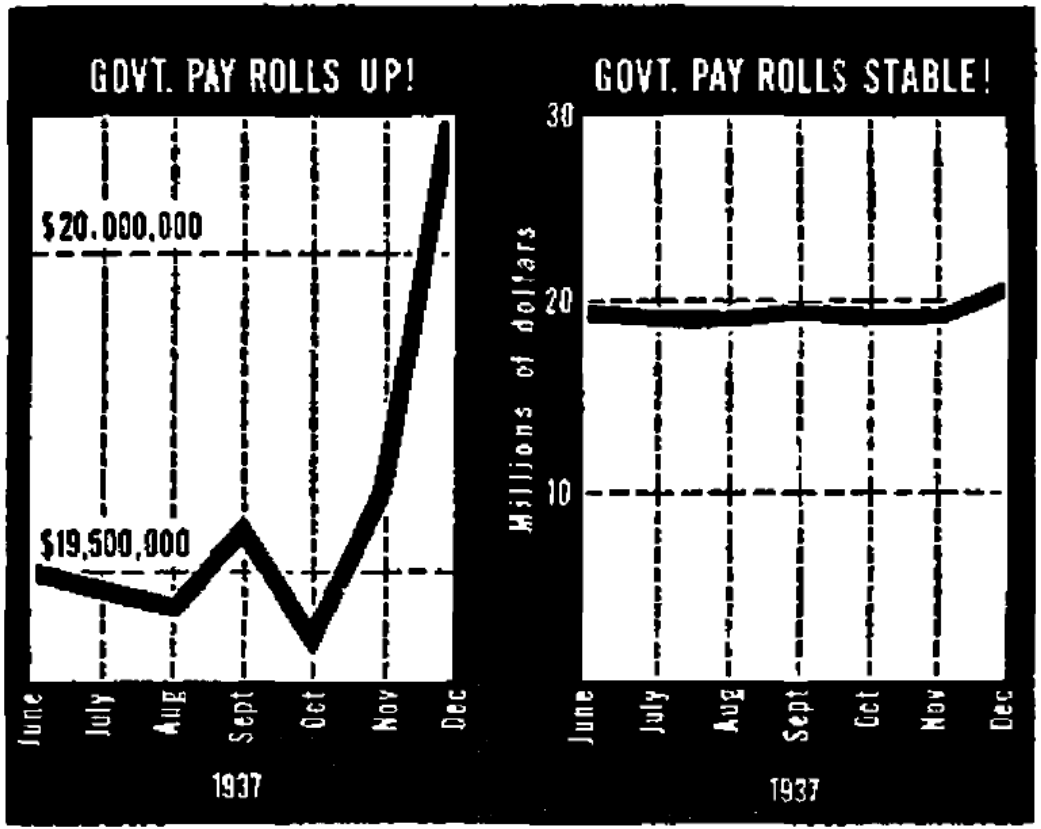
Tampering with the scales



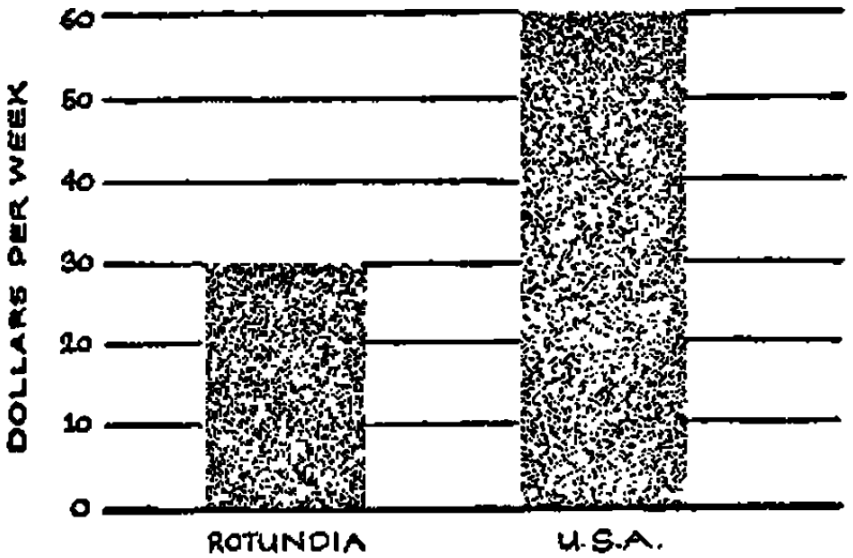


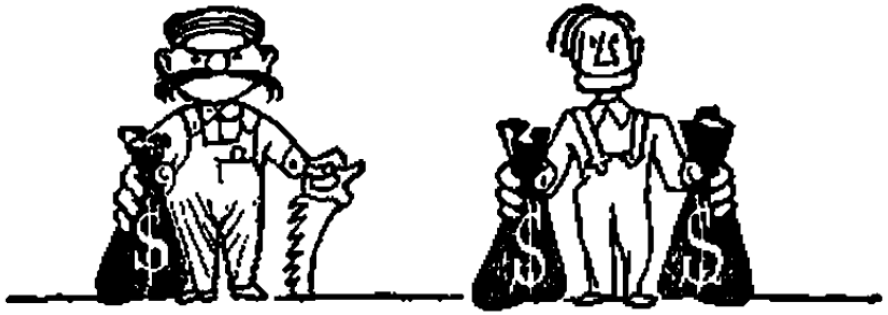
Chair of Process and Data Science

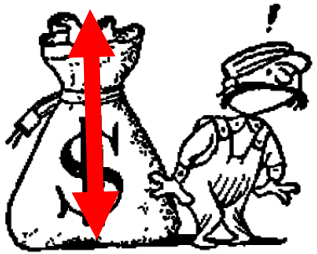
Tampering with the scales



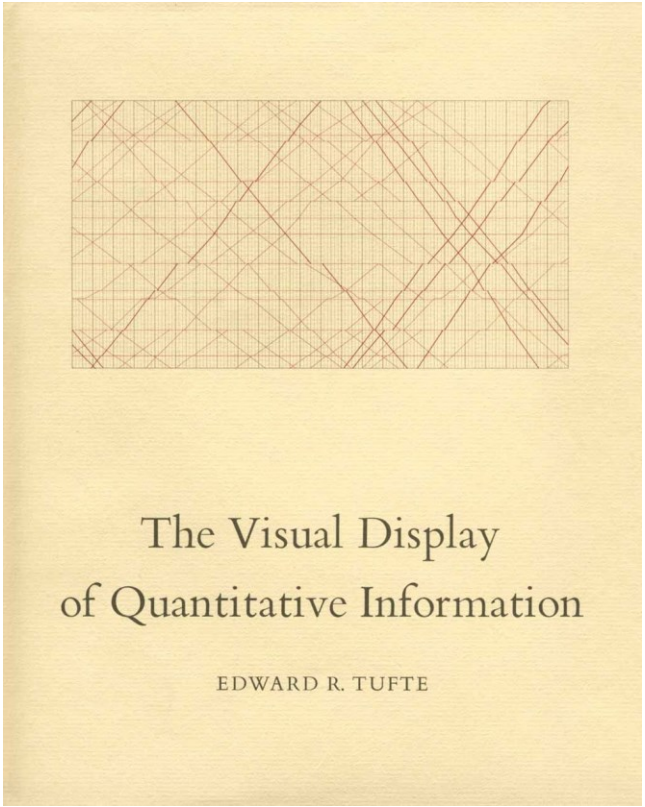
Visualizing a factor 2







Examples from “The Visual Display of Quantitative Information”



- 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

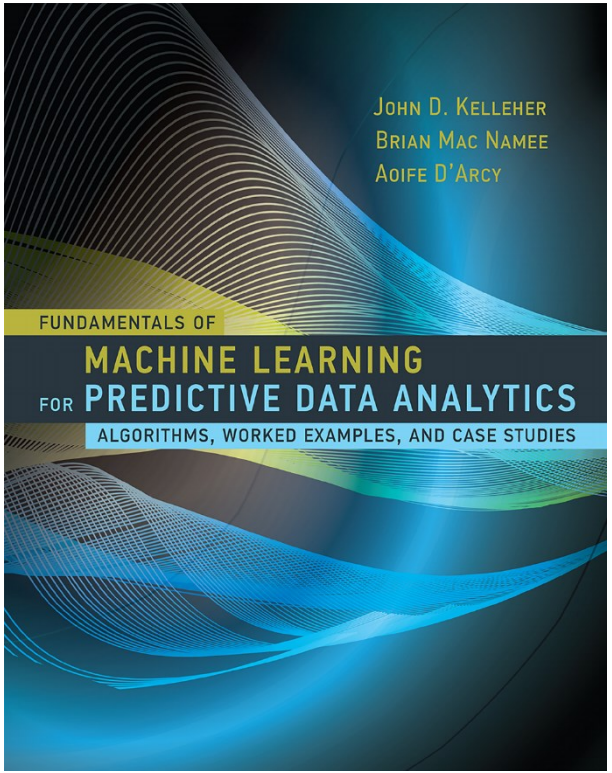

GREAT MARLBOROUGH STREET
PORTLAND STREET
WORK HOUSE
PORTLAND MEWS
BENTINCK STREET
WARDOUR MEWS
ST ANNS COURT
RICHMOND BUILDINGS
RICHMOND MEWS
BREWERY YARD
SHIP YARD
EDWARD STREET
DUKE STREET
DUKE LANE
HAM SQUARE
TYLERS CT
MEARDS COURT
PETER STREET
WALKERS COURT
OLD CROWN CT
OLD CROWN COURT
PULTENEY
PULTENEY COURT
GREAT PULTENEY STREET
LITTLE WINDMILL STREET
WILLIAM AND MARY YARD
HUSBAND ST
HAIRDRESSERS COURT
BREWERY
HOPKINS STREET
NEW STREET
COCK CT
BROAD STREET
PUMP
DUFOURS PLACE
CAMBRIDGE ST
MARSHALL STREET
MARSHALL ST.
TYLER COURT
WEST STREET
LOWNDES CT
MARLBOROUGH CT
SOUTH ROW
CROSS STREET
MARLBOROUGH ROW
CARNABY STREET
KING STREET
TYLER STREET
LITTLE MARLBOROUGH ST
GREEN DRAGON YARD
PUMP
CHAPEL PLACE
DAVIDGE STREET
UP JAMES ST
BRIDLE
ROUBERTS PL.
CYLL PLACE
STREET

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.

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	description
Lecture 1	Introduction	10/10/2018	Wednesday	
Lecture 2	Crash Course in Python	11/10/2018	Thursday	
Instruction 1	*Python*	*12/10/2018*	*Friday*	
Lecture 3	Basic data visualisation/exploration	17/10/2018	Wednesday	
Lecture 4	Decision trees	18/10/2018	Thursday	
Instruction 2	*Decision trees and data visualization/explora*	*19/10/2018*	*Friday*	
Lecture 5	Regression	24/10/2018	Wednesday	
Lecture 6	Support vector machines	25/10/2018	Thursday	
Instruction 3	*Regression and support vector machines*	*26/10/2018*	*Friday*	
Lecture 7	Neural networks (1/2)	31/10/2018	Wednesday	
Instruction 4	*Neural networks and supervised learning*	*02/11/2018*	*Friday*	
Lecture 8	Neural networks (2/2)	07/11/2018	Wednesday	
Lecture 9	Evaluation of supervised learning problems	08/11/2018	Thursday	
Instruction 5	*Neural networks and supervised learning*	*09/11/2018*	*Friday*	
Lecture 10	Clustering	14/11/2018	Wednesday	
Lecture 11	Frequent items sets	15/11/2018	Thursday	
Lecture 12	Association rules	21/11/2018	Wednesday	
Lecture 13	Sequence mining	22/11/2018	Thursday	
Instruction 6	*Clustering, frequent items sets, association ru*	*23/11/2018*	*Friday*	
Lecture 14	Process mining (unsupervised)	28/11/2018	Wednesday	
Lecture 15	Process mining (supervised)	29/11/2018	Thursday	
Instruction 7				
Lecture 16				
Instruction 8				
Lecture 17				
Lecture 18				
Lecture 19				
backup				
Instruction 9				
Lecture 20				
Lecture 21				
Instruction 10				
Lecture 22				
Lecture 23				
Instruction 11	*Big data*	*18/01/2019*	*Friday*	
Lecture 24	Closing	23/01/2019	Wednesday	
backup		24/01/2019	Thursday	
Instruction 12	*Example exam questions*	*25/01/2018*	*Friday*	
backup		30/01/2019	Wednesday	
backup		31/01/2019	Thursday	
extra	*Question hour*	*01/02/2019*	*Friday*	

Lecture 3	Basic data visualisation/exploration	17/10/2018	Wednesday	
Lecture 4	Decision trees	18/10/2018	Thursday	
Instruction 2	*Decision trees and data visualization/explorat*	*19/10/2018*	*Friday*	
Lecture 5	Regression	24/10/2018	Wednesday	
Lecture 6	Support vector machines	25/10/2018	Thursday	
Instruction 3	*Regression and support vector machines*	*26/10/2018*	*Friday*	