MATH 205: Statistical methods

Lecture 4: Correlation

Tentative schedule

Date	Theme/Topic	Labs	Assignments
Aug 31	Syllabus		
Sep 2—9	Chapter 1: Describing dataset	Section 2: Handling data	
Sep 12—16	Chapter 2: Looking at Relationships	Section 3: Univariate data	
Sep 19—23	Chapter 3: Basic Ideas in Probability	Section 4: Bivariate Data	Homework 1 (due 09/23)
Sep 26—30	Chapters 3-4	Section 4: Correlation	
Oct 3—7	Chapter 4: Random variables and expectations	Section 6: Random data	Homework 2 (due 10/07)
Oct 10—14	Chapter 5: Useful distributions	Section 7: The central limit theorem	
Oct 17—21	Chapter 6: Samples and populations	Section 9: Confidence interval estimation	Homework 3 (due 10/21)
Oct 24—28	Review Midterm exam		Midterm: Oct 28 (lecture) Oct 24-26 (labs)
Oct 31—Nov 4	Chapter 7: The significance of evidence	Section 10: Hypothesis testing	
Nov 7—11	Goodness of Fit	Section 12: Goodness of Fit	Homework 4 (due 11/11)
Nov 14—18	Linear Regression	Section 13: Linear regression	
Nov 21–25	Thanksgiving break		
Nov 28 —Dec 2	One-Way Analysis of Variance	Section 15: Analysis of variance	Homework 5 (due 12/02)
Dec 5—7	Selected topics + Review		
Exam week			

Lectures 1–3: Describing univariate data

Summarizing univariate data:

- Mean
- Median
- Standard deviation
- Interquartile Range

Visualizing univariate data:

- Bar chart
- Pie chart
- Histogram
- Box plot

Chapter 2: Looking at relationship

Plotting 2D data

- We take a dataset, choose two different features, and extract the corresponding elements from each tuple
- The result is a dataset consisting of 2-tuples, and we think of this as a two dimensional dataset
- Goal: to plot this dataset in a way that reveals relationships

Example: CO2 dataset

The CO2 uptake of six plants from Quebec and six plants from Mississippi was measured at several levels of ambient CO2 concentration. Half the plants of each type were chilled overnight before the experiment was conducted.

	Plant	Type	Treatment	conc	uptake
1	Qn1	Quebec	nonchilled	95	16.0
2	Qn1	Quebec	nonchilled	175	30.4
3	Qn1	Quebec	nonchilled	250	34.8
4	Qn1	Quebec	nonchilled	350	37.2
5	Qn1	Quebec	nonchilled	500	35.3
6	Qn1	Quebec	nonchilled	675	39.2
7	Qn1	Quebec	nonchilled	1000	39.7
8	Qn2	Quebec	nonchilled	95	13.6
9	Qn2	Quebec	nonchilled	175	27.3
10	Qn2	Quebec	nonchilled	250	37.1
11	Qn2	Quebec	nonchilled	350	41.8
12	Qn2	Quebec	nonchilled	500	40.6

Example: Animals {MASS} dataset

	body	brain
Arctic fox	3.385	44.50
Owl monkey	0.480	15.50
Mountain beaver	1.350	8.10
Cow	465.000	423.00
Grey wolf	36.330	119.50
Goat	27.660	115.00
Roe deer	14.830	98.20
Guinea pig	1.040	5.50
Verbet	4.190	58.00
Chinchilla	0.425	6.40
Ground squirrel	0.101	4.00
Arctic ground squirrel	0.920	5.70
African giant pouched rat	1.000	6.60

Plotting 2D data

- categorical vs categorical: create a richer set of categories
- categorical vs continuous: comparative box plots
- continuous vs continuous: scatter plots

Categorial vs categorical

- Common approach: create a richer set of categories
- Example: Relationship between
  - automobile class (2seater, compact, midsize, minivan, pickup, subcompact, suv)
  - drive type (front-wheel, rear-wheel, or 4-wheel drive)

Example: mpg {ggplot2} dataset

	manufacturer	model	displ	year	cyl	trans	drv	cty	hwy	fl	class
7	audi	a4	3.1	2008	6	auto(av)	f	18	27	p	compact
8	audi	a4 quattro	1.8	1999	4	manual(m5)	4	18	26	p	compact
9	audi	a4 quattro	1.8	1999	4	auto(l5)	4	16	25	p	compact
10	audi	a4 quattro	2.0	2008	4	manual(m6)	4	20	28	p	compact
11	audi	a4 quattro	2.0	2008	4	auto(s6)	4	19	27	p	compact
12	audi	a4 quattro	2.8	1999	6	auto(l5)	4	15	25	p	compact
13	audi	a4 quattro	2.8	1999	6	manual(m5)	4	17	25	p	compact
14	audi	a4 quattro	3.1	2008	6	auto(s6)	4	17	25	p	compact
15	audi	a4 quattro	3.1	2008	6	manual(m6)	4	15	25	p	compact
16	audi	a6 quattro	2.8	1999	6	auto(l5)	4	15	24	p	midsize
17	audi	a6 quattro	3.1	2008	6	auto(s6)	4	17	25	p	midsize
18	audi	a6 quattro	4.2	2008	8	auto(s6)	4	16	23	p	midsize
19	chevrolet	c1500 suburban 2wd	5.3	2008	8	auto(l4)	r	14	20	r	suv
20	chevrolet	c1500 suburban 2wd	5.3	2008	8	auto(l4)	r	11	15	e	suv

Grouped bar charts

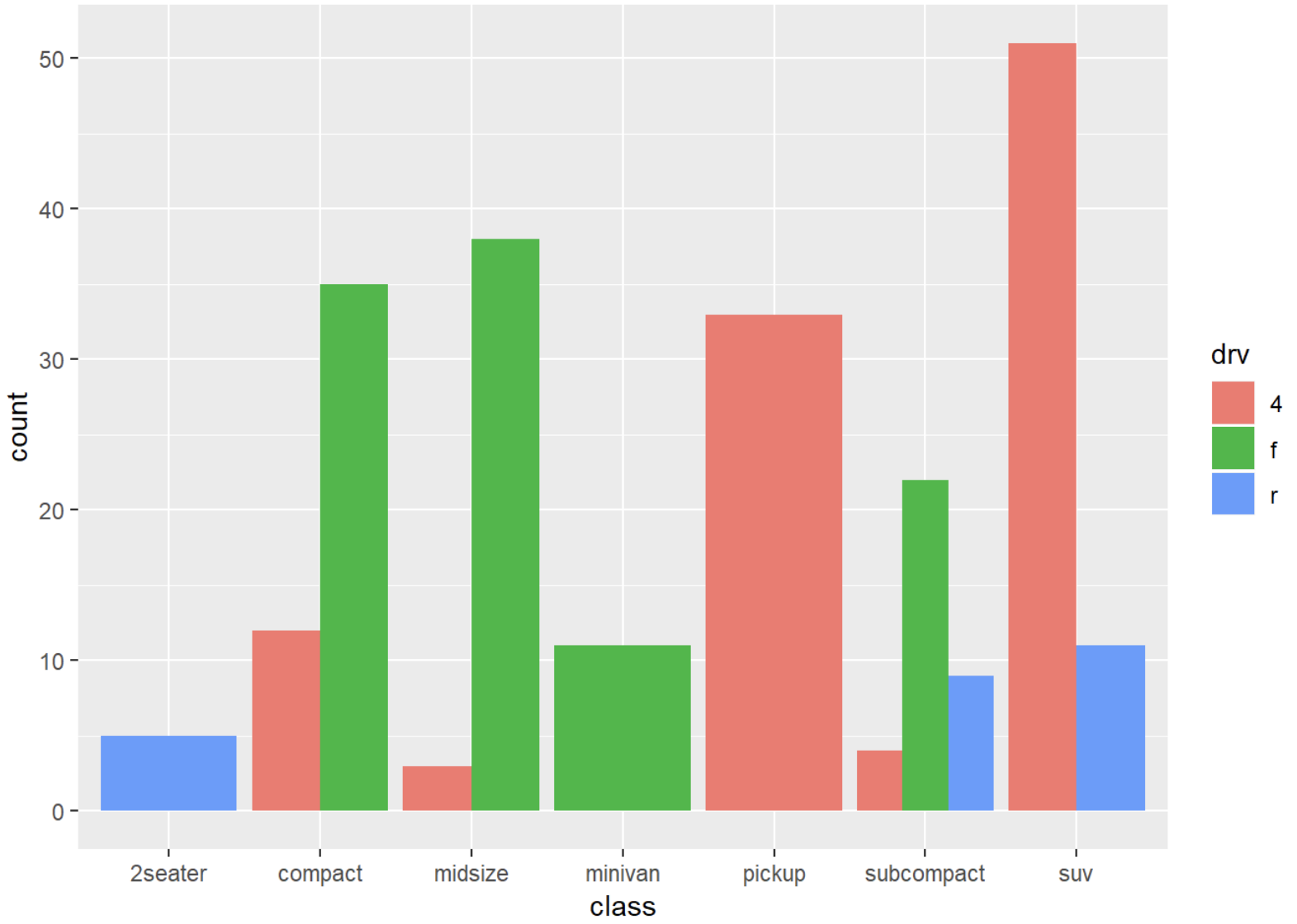
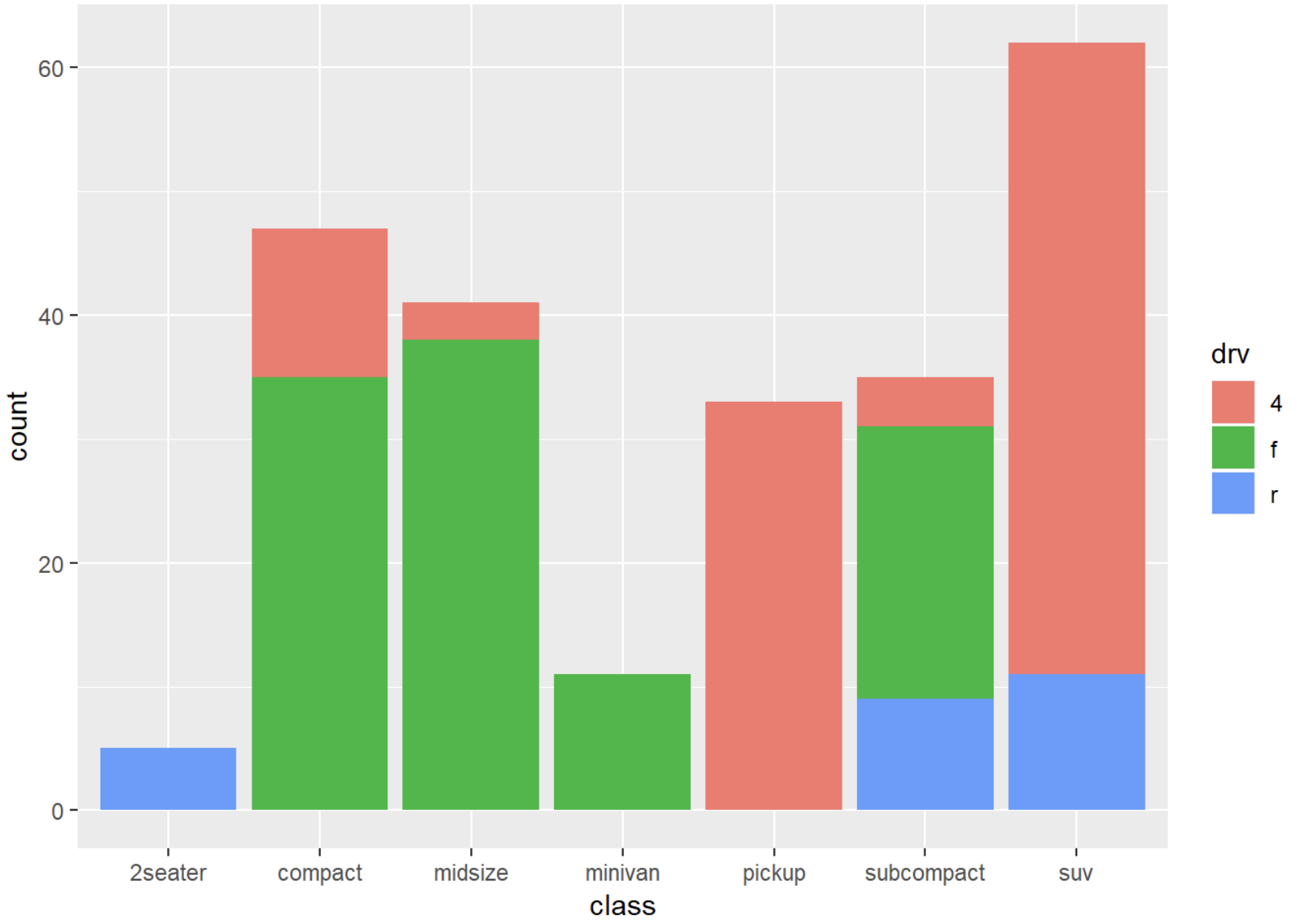
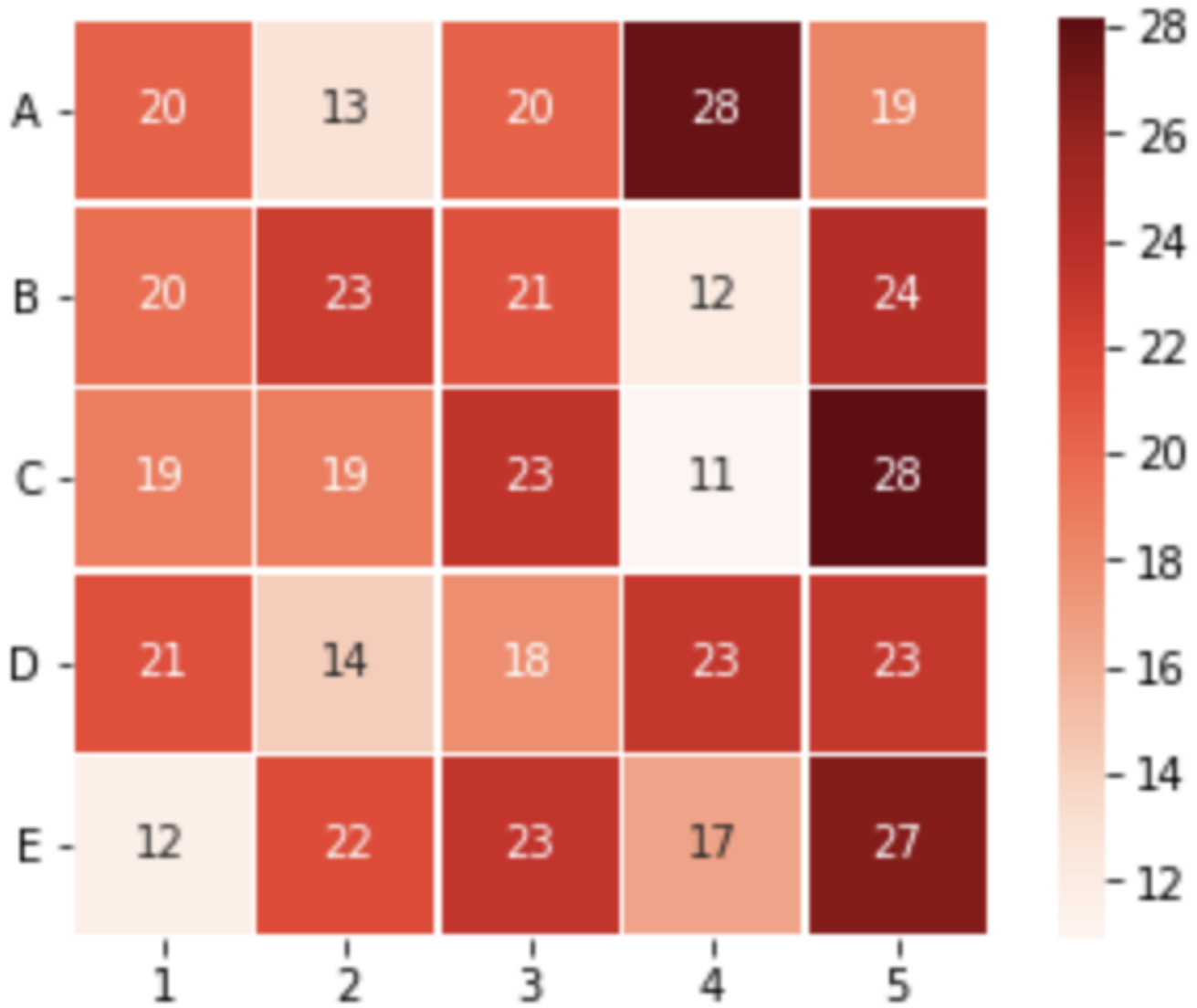


Figure 4.2: Side-by-side bar chart

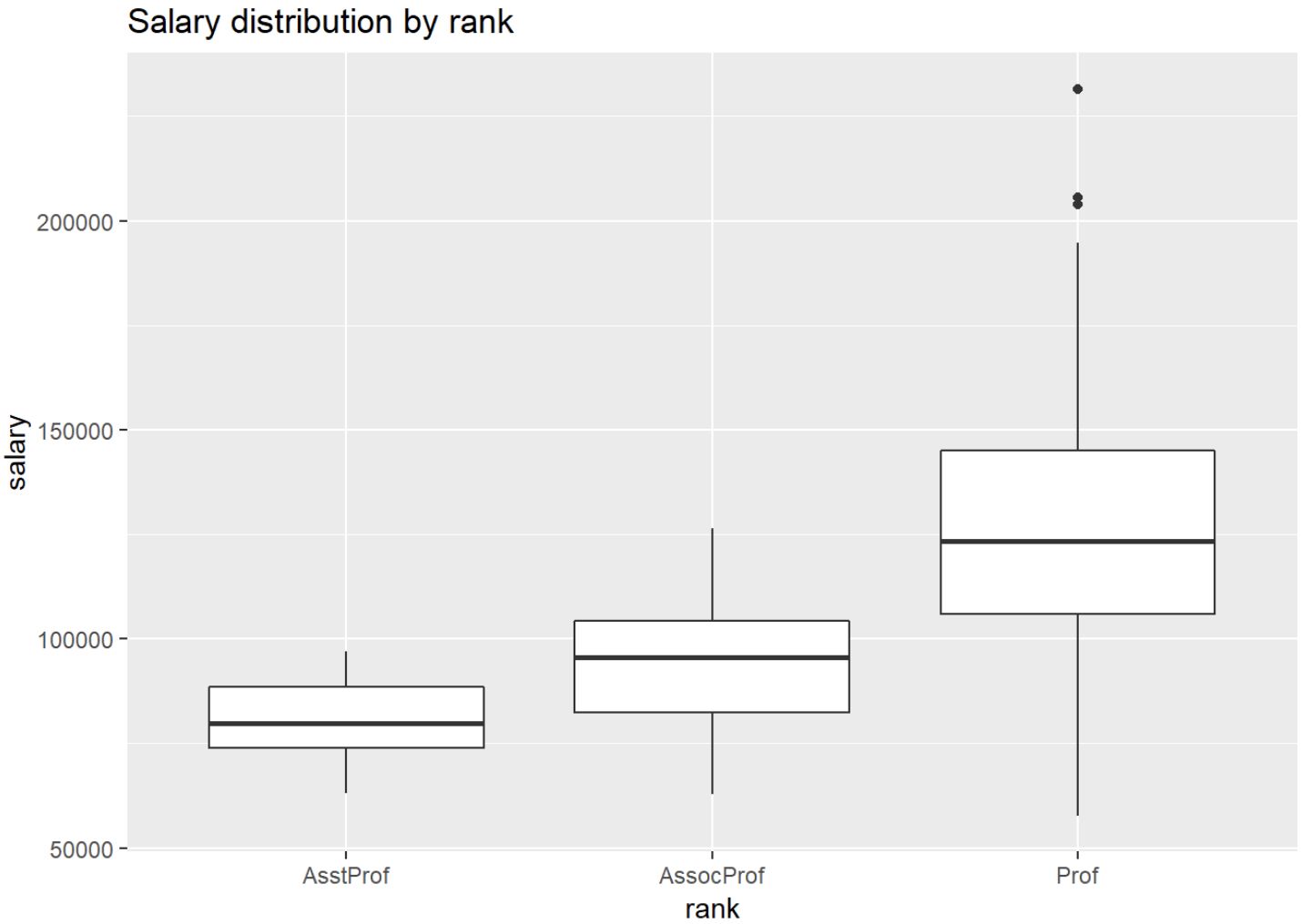
Stacked bar charts



Heatmap


A
B
C
D
E
20 13 20 28 19
20 23 21 12 24
19 19 23 11 28
21 14 18 23 23
12 22 23 17 27
1
2
3
4
5
28
26
24
22
20
18
16
14
12

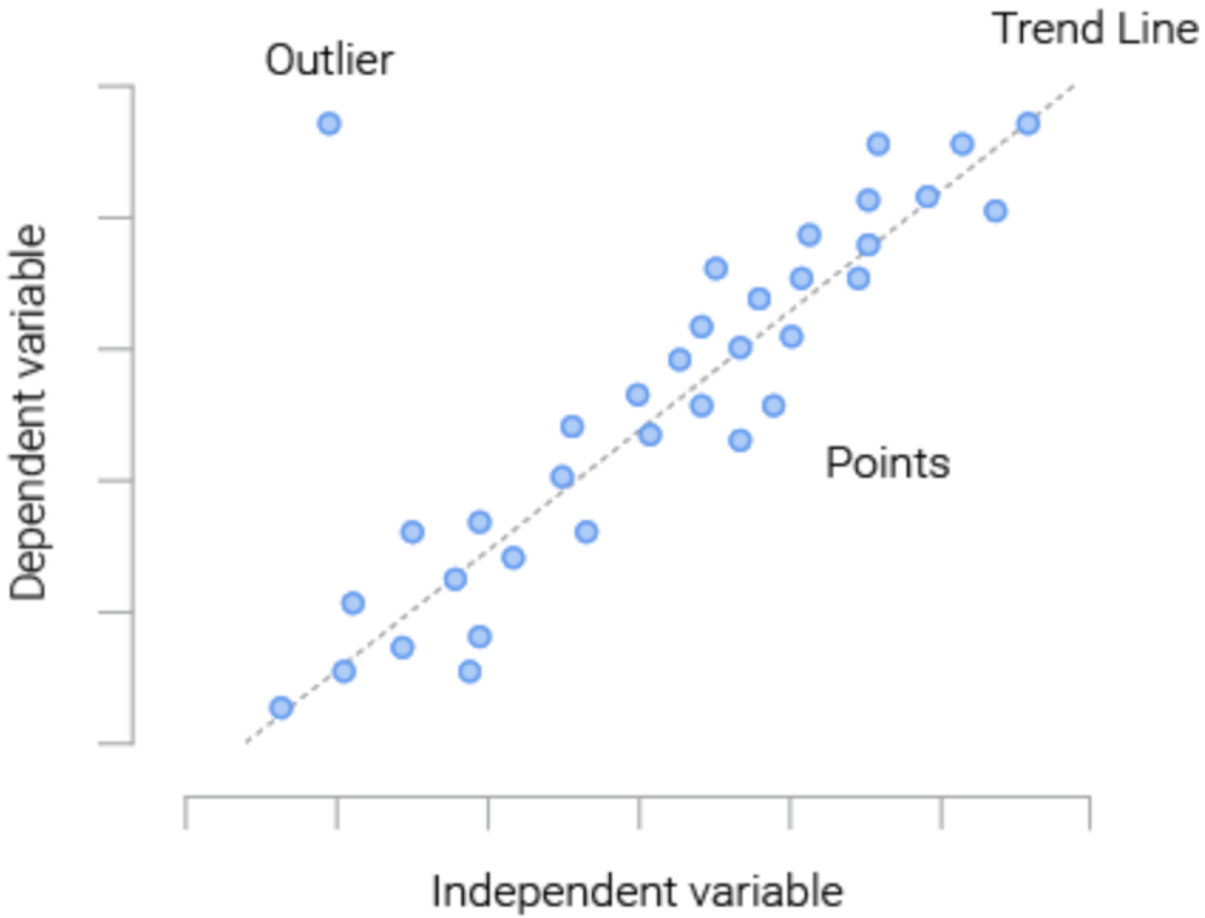
Categorial vs continuous



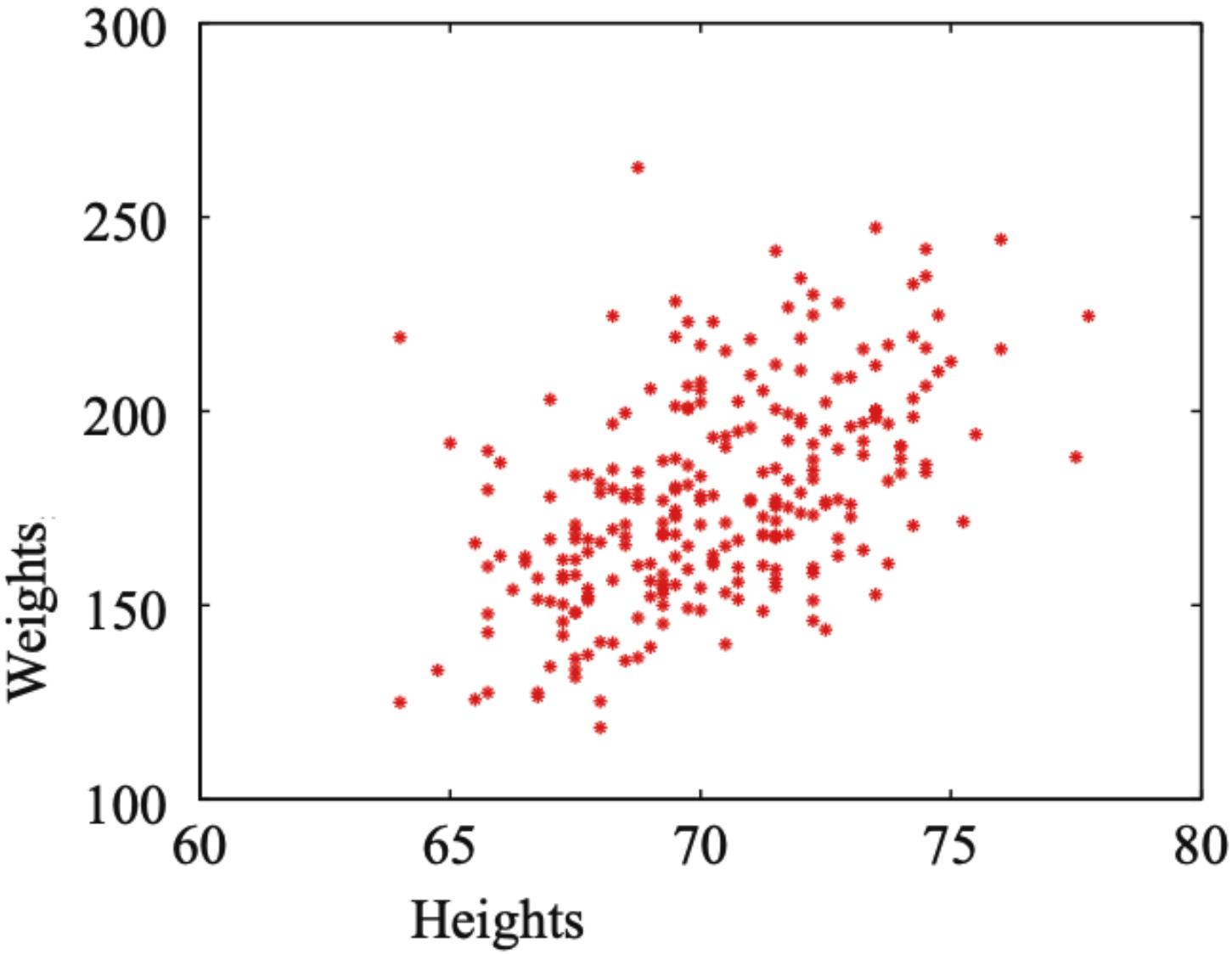
continuous vs continuous: scatterplot

- Use Cartesian coordinates to display values for two variables for a set of data
- The data are displayed as a collection of points, each having the value of one variable determining the position on the horizontal axis and the value of the other variable determining the position on the vertical axis

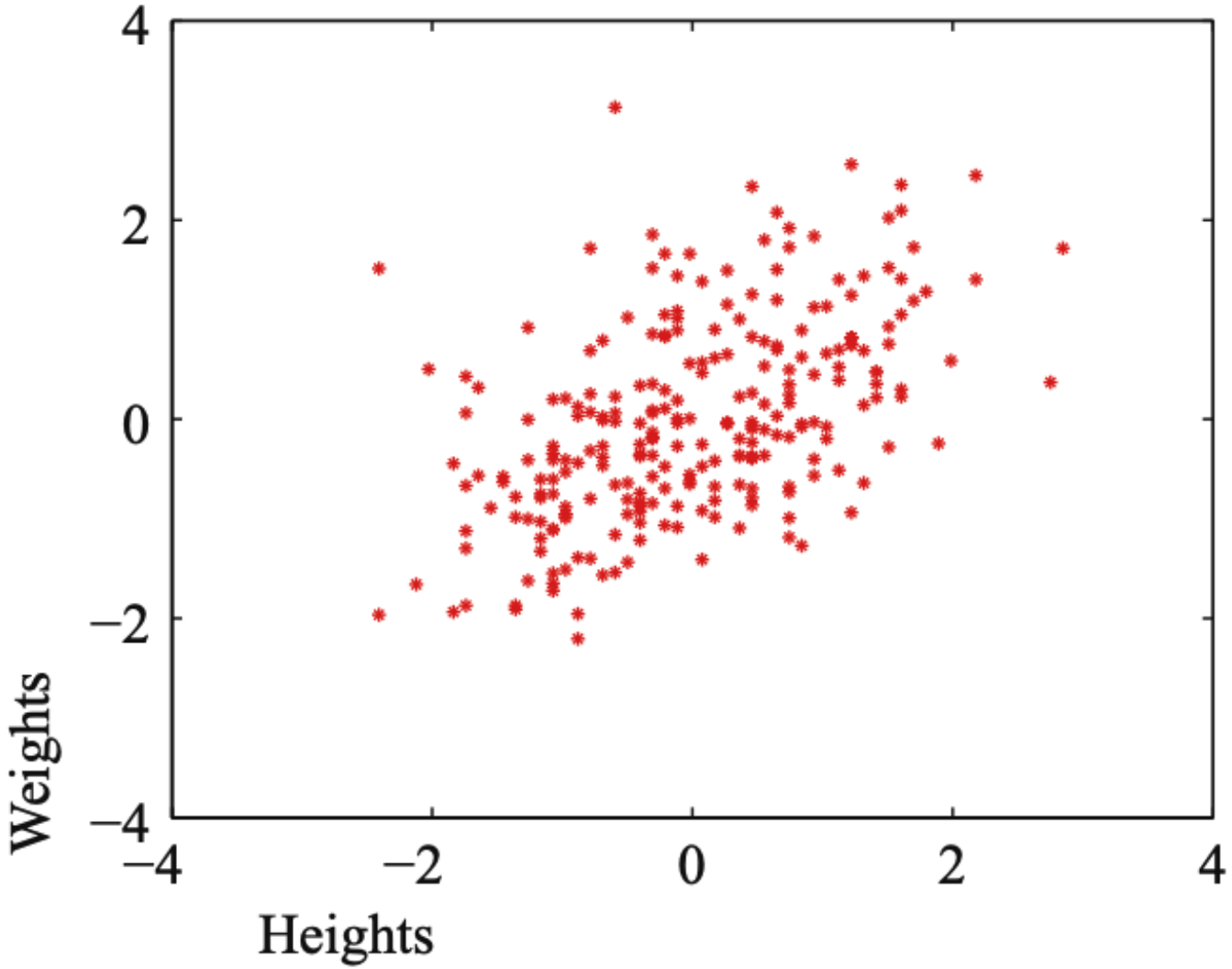
Scatterplot



Scatterplot: example



Standard coordinates



Problem

Definition 1.8 (Standard Coordinates) Assume we have a dataset $\{x\}$ of N data items, $x_1, \ldots, x_N$. We represent these data items in standard coordinates by computing

$$\hat{x}_i = \frac{(x_i - \text{mean}(\{x\}))}{\text{std}(\{x\})}.$$

We write $\{\hat{x}\}$ for a dataset that happens to be in standard coordinates.

Prove that:

- $mean(\{\hat{x}\}) = 0$
- $std(\{\hat{x}\}) = 1$

Correlation

Correlations

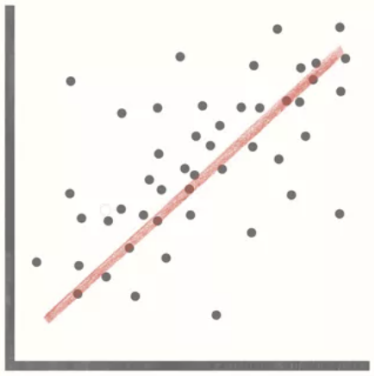
- From the figure: someone who is taller than the mean will tend to be heavier than the mean too
- This relationship is not always true for specific cases (and can not be represented by a function): some people are quite a lot taller than the mean, and quite a lot lighter

Correlations

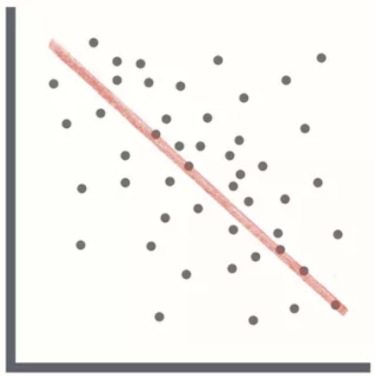
Question: when $\hat{x}$ increases, does $\hat{y}$ tend to increase, decrease, or stay the same?

- Positive correlation: larger $\hat{x}$ values tend to appear with larger $\hat{y}$ values
- Negative correlation: larger $\hat{x}$ values tend to appear with smaller $\hat{y}$ values
- Zero correlation: no relationship

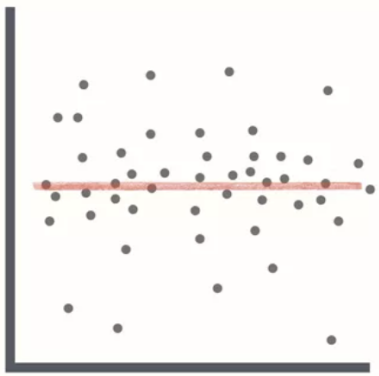
Correlations



Positive Correlation



Negative Correlation



No Correlation

Correlation coefficient

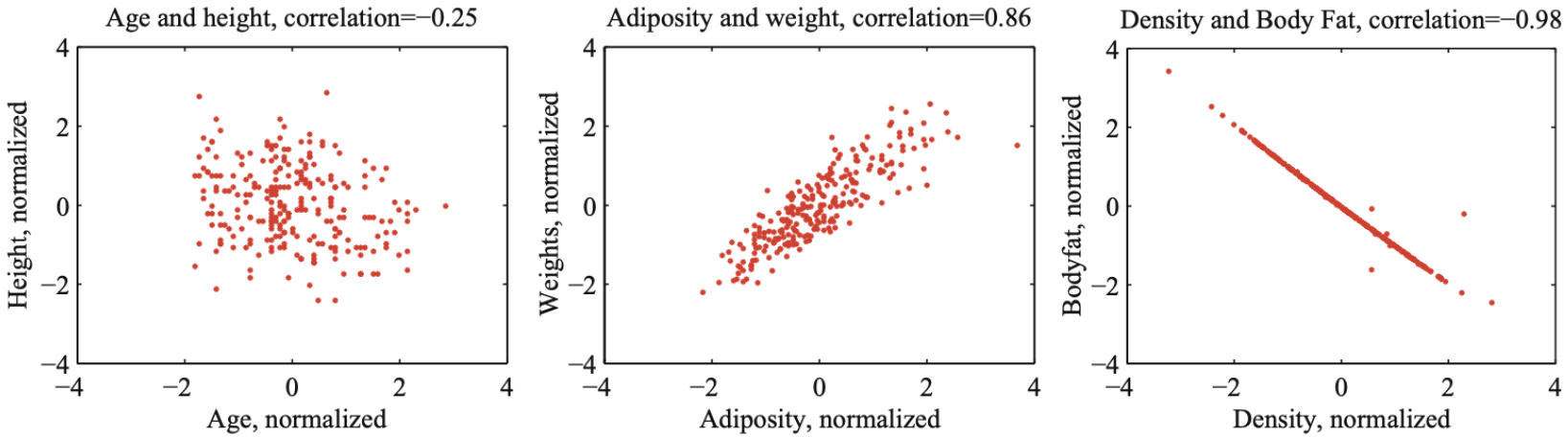
Definition 2.1 (Correlation Coefficient) Assume we have N data items which are 2-vectors $(x_1, y_1), \ldots, (x_N, y_N)$, where $N > 1$. These could be obtained, for example, by extracting components from larger vectors. We compute the correlation coefficient by first normalizing the x and y coordinates to obtain $\hat{x}_i = \frac{(x_i - \text{mean}(\{x\}))}{\text{std}(x)}$, $\hat{y}_i = \frac{(y_i - \text{mean}(\{y\}))}{\text{std}(y)}$. The correlation coefficient is the mean value of $\hat{x}\hat{y}$, and can be computed as:

$$\text{corr}(\{(x, y)\}) = \frac{\sum_i \hat{x}_i \hat{y}_i}{N}$$

Correlation coefficient

- correlation is a measure of our ability to predict one value from another
- correlation coefficient takes values between -1 and 1
- If the correlation coefficient is close to 1 or -1, then we are likely to predict very well.

Correlation coefficient



Correlation coefficient: properties

Useful Facts 2.1 (Properties of the Correlation Coefficient)

- The correlation coefficient is symmetric (it doesn't depend on the order of its arguments), so

$$\text{corr}(\{(x, y)\}) = \text{corr}(\{(y, x)\})$$

- The value of the correlation coefficient is not changed by translating the data. Scaling the data can change the sign, but not the absolute value. For constants $a \neq 0$, b, $c \neq 0$, d we have

$$\text{corr}(\{(ax + b, cx + d)\}) = \text{sign}(ab)\text{corr}(\{(x, y)\})$$

- If $\hat{y}$ tends to be large (resp. small) for large (resp. small) values of $\hat{x}$, then the correlation coefficient will be positive.
- If $\hat{y}$ tends to be small (resp. large) for large (resp. small) values of $\hat{x}$, then the correlation coefficient will be negative.
- If $\hat{y}$ doesn't depend on $\hat{x}$, then the correlation coefficient is zero (or close to zero).
- The largest possible value is 1, which happens when $\hat{x} = \hat{y}$.
- The smallest possible value is -1, which happens when $\hat{x} = -\hat{y}$.