NATIONAL RESEARCH UNIVERSITY

School of Data Analysis and Artificial Intelligence Department of Computer Science

DATA SCIENCE FOR BUSINESS

Lecture 2. Exploratory Data Analysis

Moscow, April 17th, 2020.

lzhukov@hse.ru

EXPLORATORY DATA ANALYSIS (EDA)

Exploratory Data Analysis (EDA) is an approach for data analysis without statistical model or formulated prior hypothesis

EDA goals

- Maximize insight into a data set
- Uncover underlying structure
- Detect missing data
- Detect outliers and anomalies
- Rank important factors
- Perform sanity check

Approaches

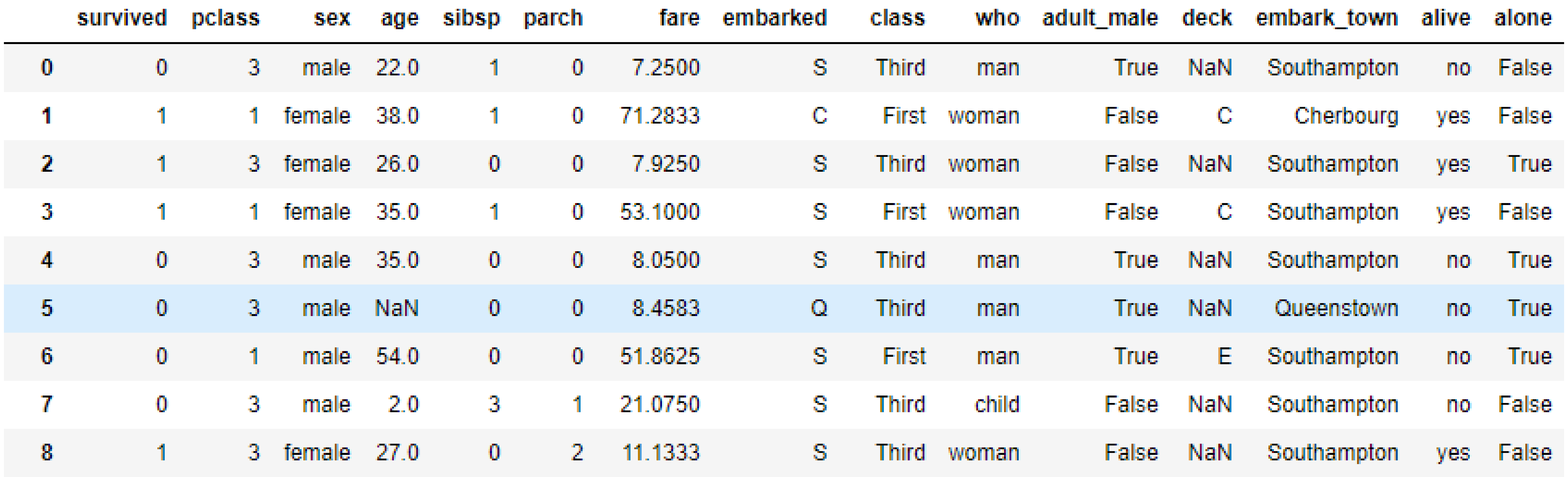
1. **Descriptive statistics:** computing simple summary statistics such as mean, median, standard deviation, plotting box plots, histograms

2. **Visualization:** plotting the raw data - data traces, scatter plots, frequency plots, probability plots, multivariate plots

John Tukey, “Exploratory Data Analysis”, 1977

DATA TYPES

- **Categorical data (= labels, nominal, ordinal [ordered], binary)**
- **Quantitative data (= numbers, discrete [integer], continues [real])**



	survived	pclass	sex	age	sibsp	parch	fare	embarked	class	who	adult_male	deck	embark_town	alive	alone
0	0	3	male	22.0	1	0	7.2500	S	Third	man	True	NaN	Southampton	no	False
1	1	1	female	38.0	1	0	71.2833	C	First	woman	False	C	Cherbourg	yes	False
2	1	3	female	26.0	0	0	7.9250	S	Third	woman	False	NaN	Southampton	yes	True
3	1	1	female	35.0	1	0	53.1000	S	First	woman	False	C	Southampton	yes	False
4	0	3	male	35.0	0	0	8.0500	S	Third	man	True	NaN	Southampton	no	True
5	0	3	male	NaN	0	0	8.4583	Q	Third	man	True	NaN	Queenstown	no	True
6	0	1	male	54.0	0	0	51.8625	S	First	man	True	E	Southampton	no	True
7	0	3	male	2.0	3	1	21.0750	S	Third	child	False	NaN	Southampton	no	False
8	1	3	female	27.0	0	2	11.1333	S	Third	woman	False	NaN	Southampton	yes	False

TABLE ROWS = instances, examples, data points, observations, samples
TABLE COLUMNS = attributes, features, variables

EDA TOOLS

Methods and approaches

Summary statistics

- min, max (range)
- mean, median (location)
- variance, standard deviation (dispersion)
- skewness (asymmetry)
- kurtosis (peakedness)

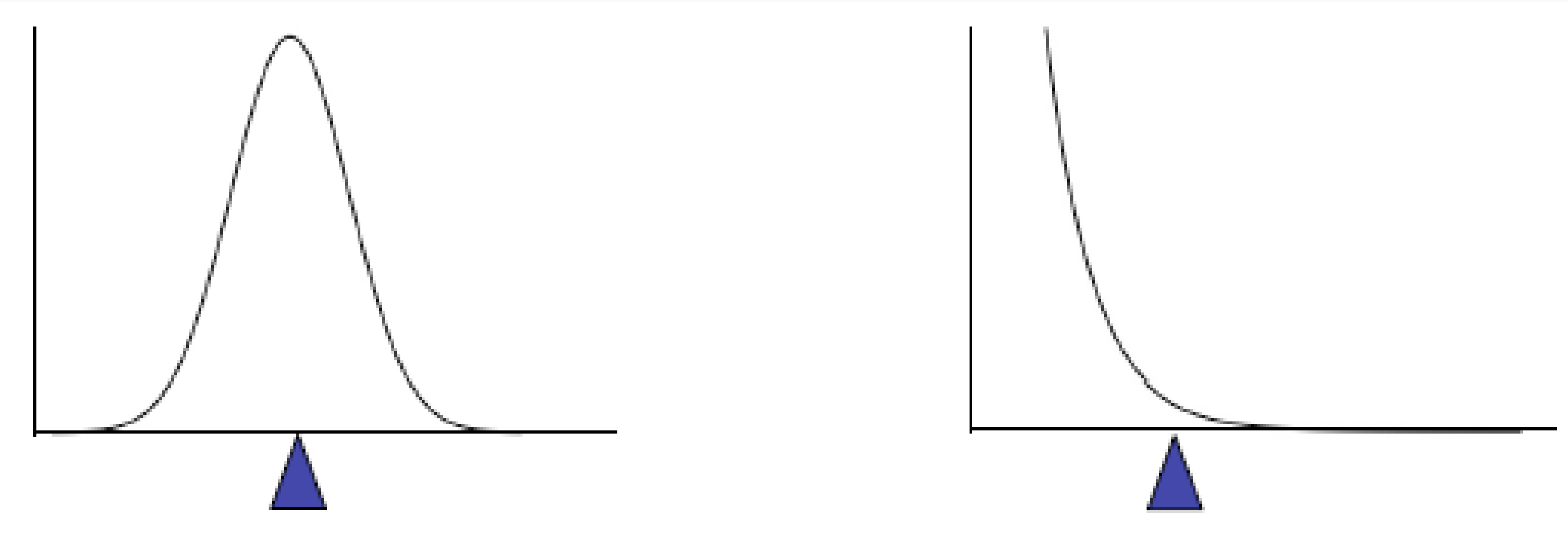
Visualization

- Bar plots
- Scatter plots
- Histograms
- Box plots
- Pairwise correlation matrix

CENTRALITY: MEAN AND MEDIAN

mean (average):

$$\overline{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$



The "average" number

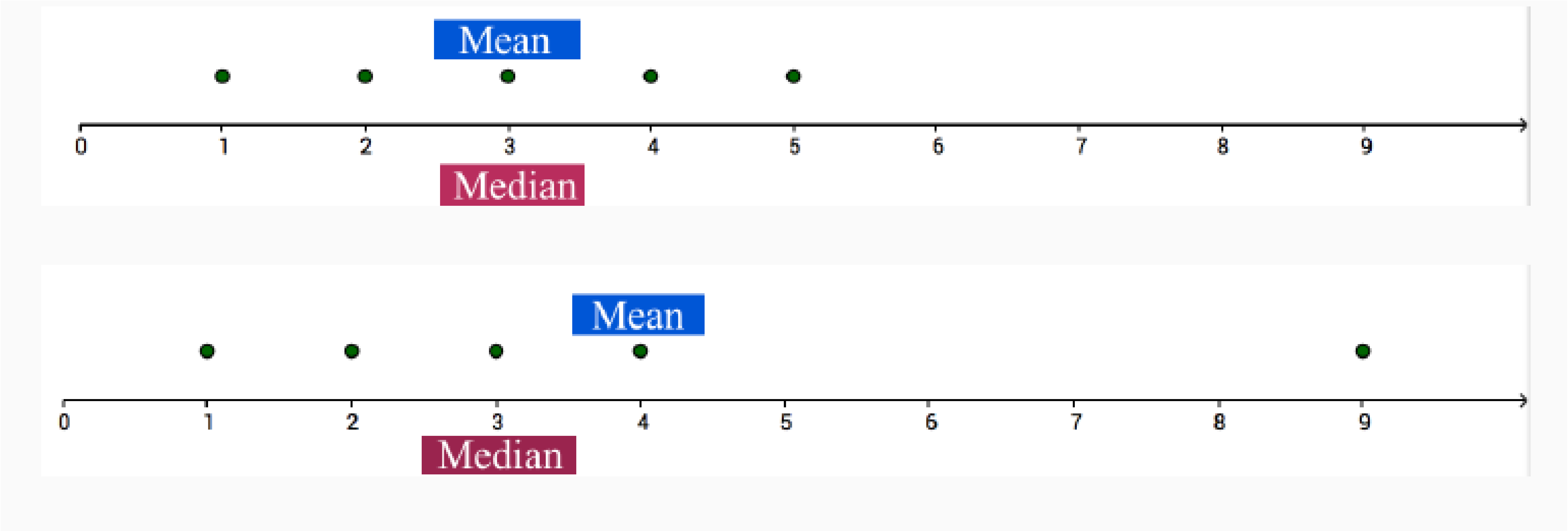
median:

$$\text{Median} = \begin{cases} x_{\lfloor n/2 \rfloor + 1}, & \text{if } n \text{ is odd} \\ \dfrac{x_{n/2} + x_{n/2+1}}{2}, & \text{if } n \text{ is even} \end{cases}$$

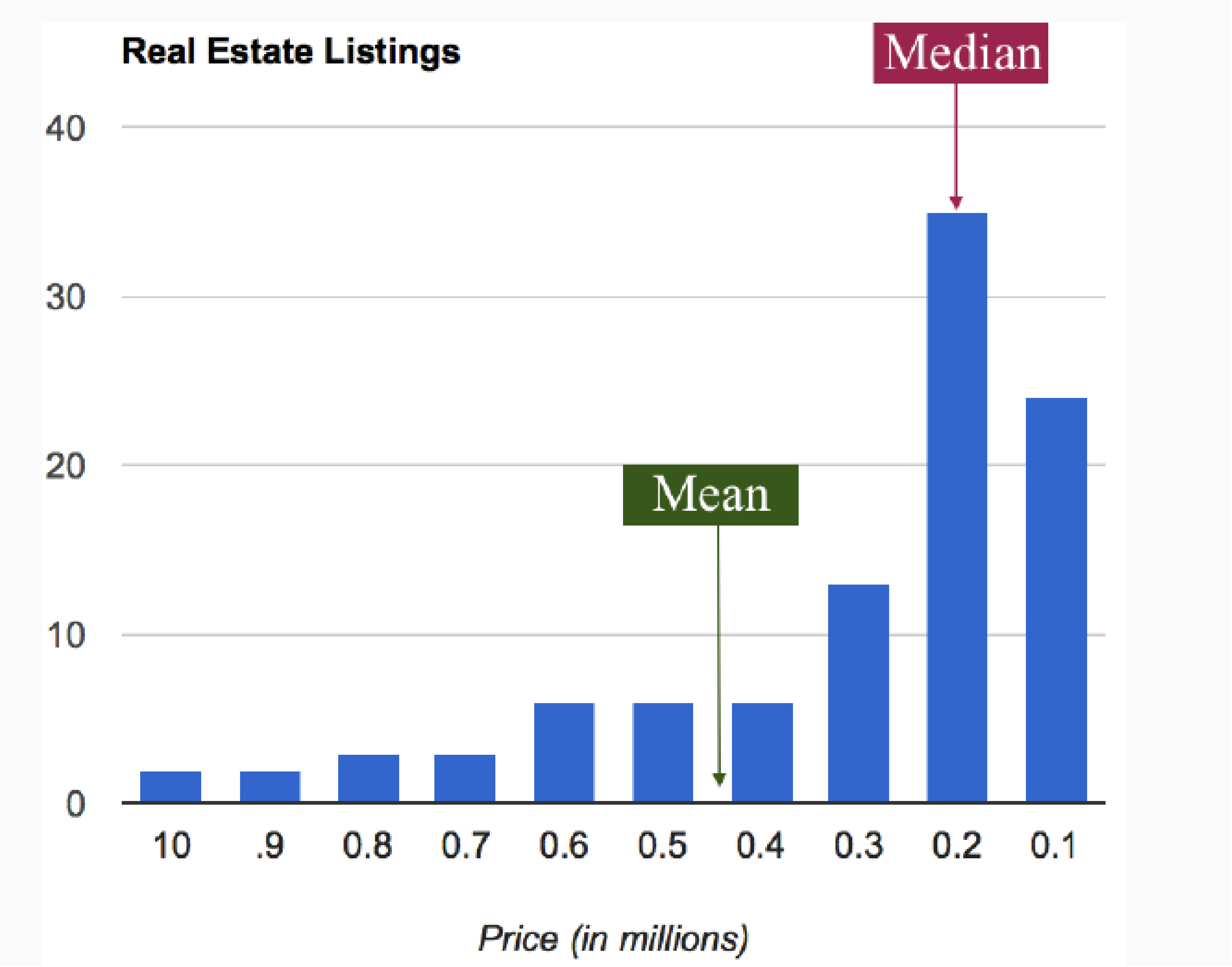
The middle number

MEAN VS MEDIAN

The mean is sensitive to outliers



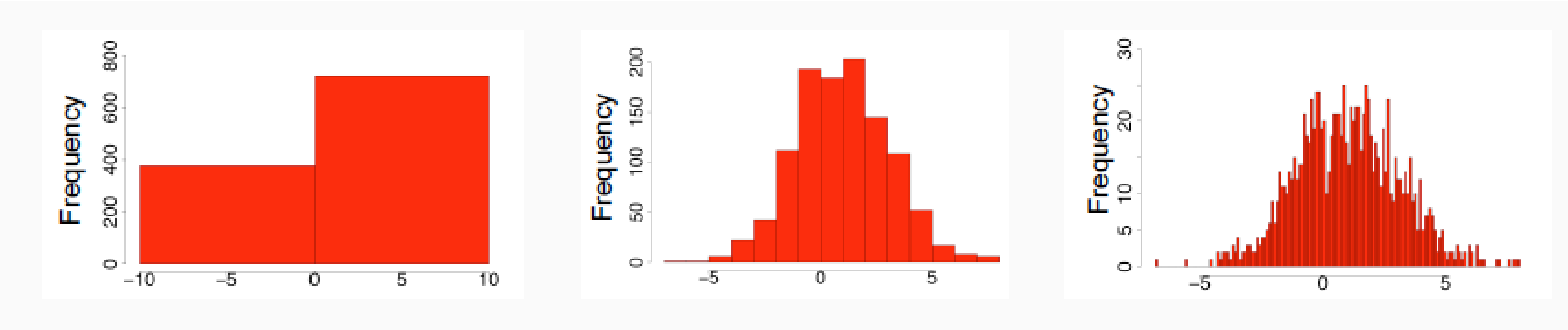
MEAN VS MEDIAN



The mean is sensitive to asymmetry (skewiness) in distribution!

HISTOGRAM

Visualizing how 1-dimensional data is distributed



Trends in histograms are sensitive to number of bins

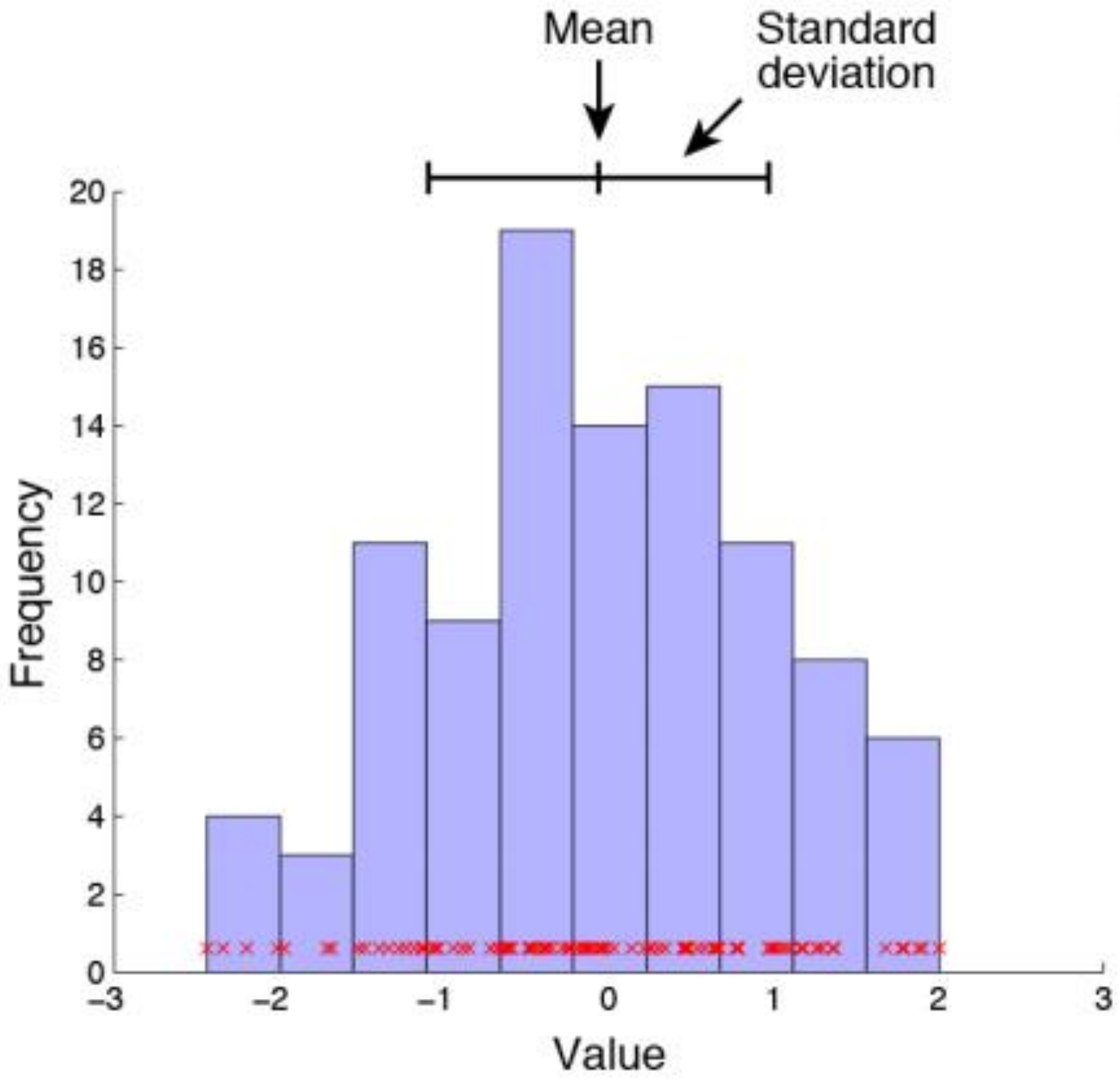
VARIANCE AND STANDARD DEVIATION

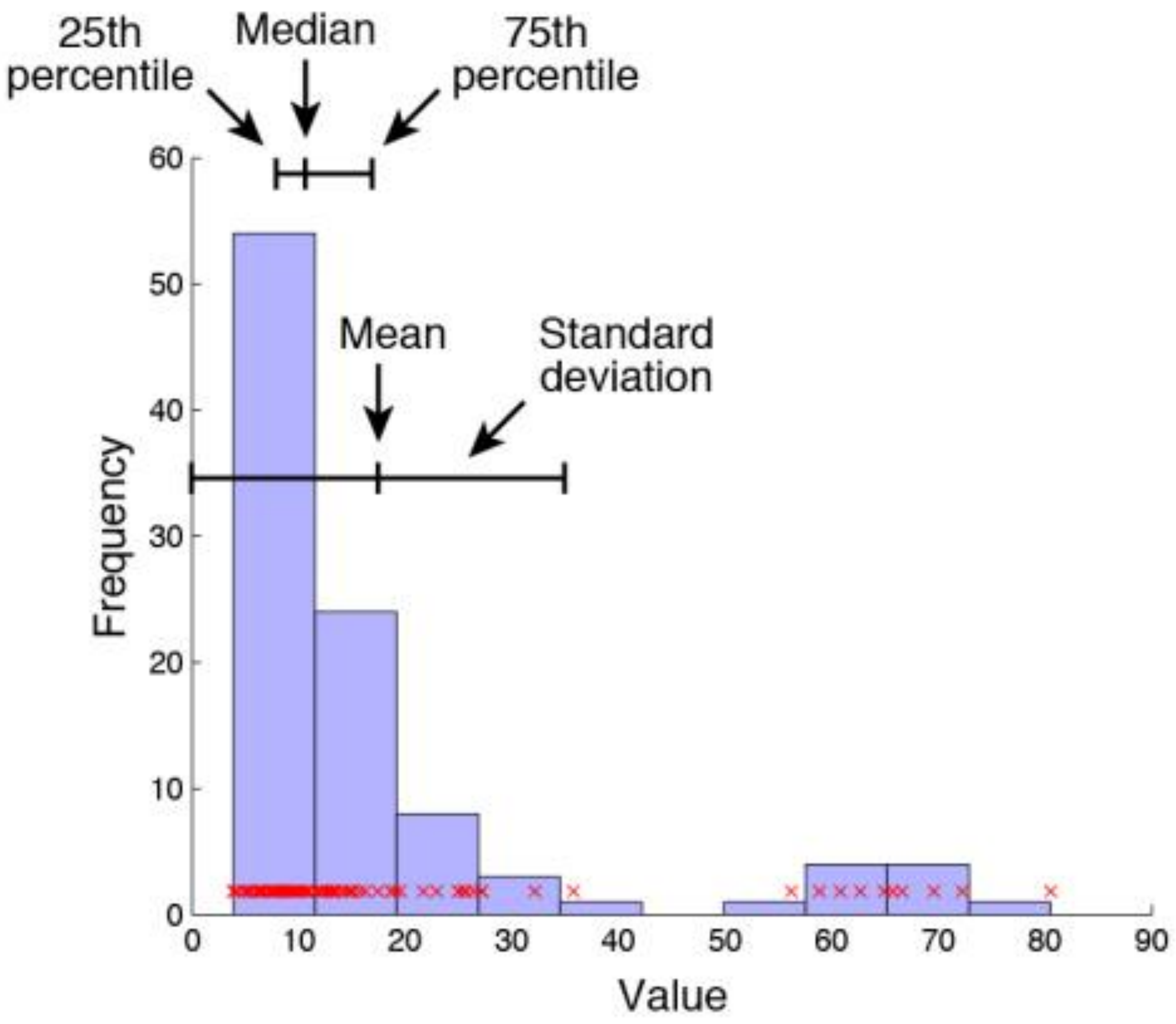
Variance

$$s_N^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2,$$

Standard deviation

$$s_N = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2},$$





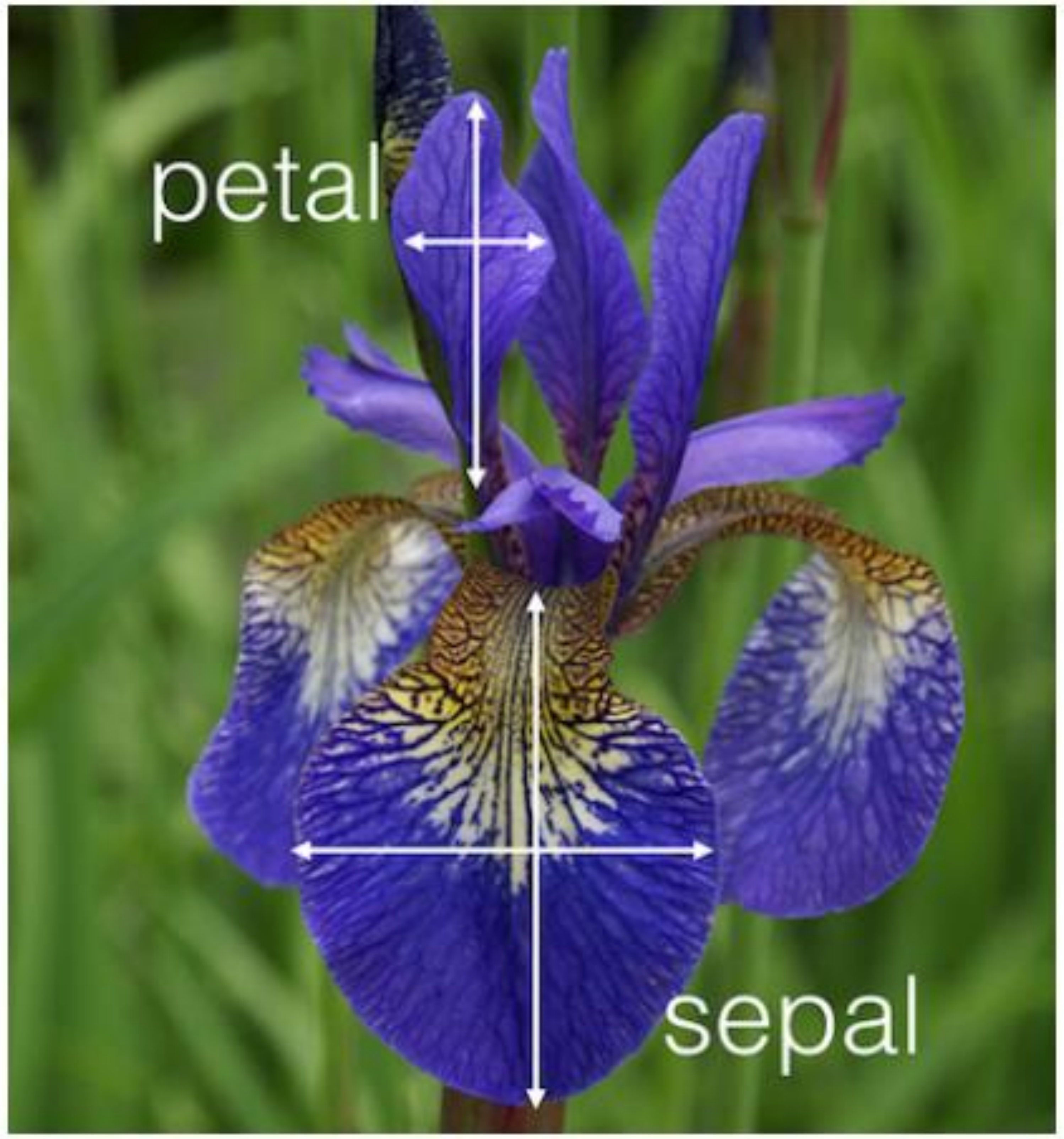
EXAMPLE: IRIS DATASET

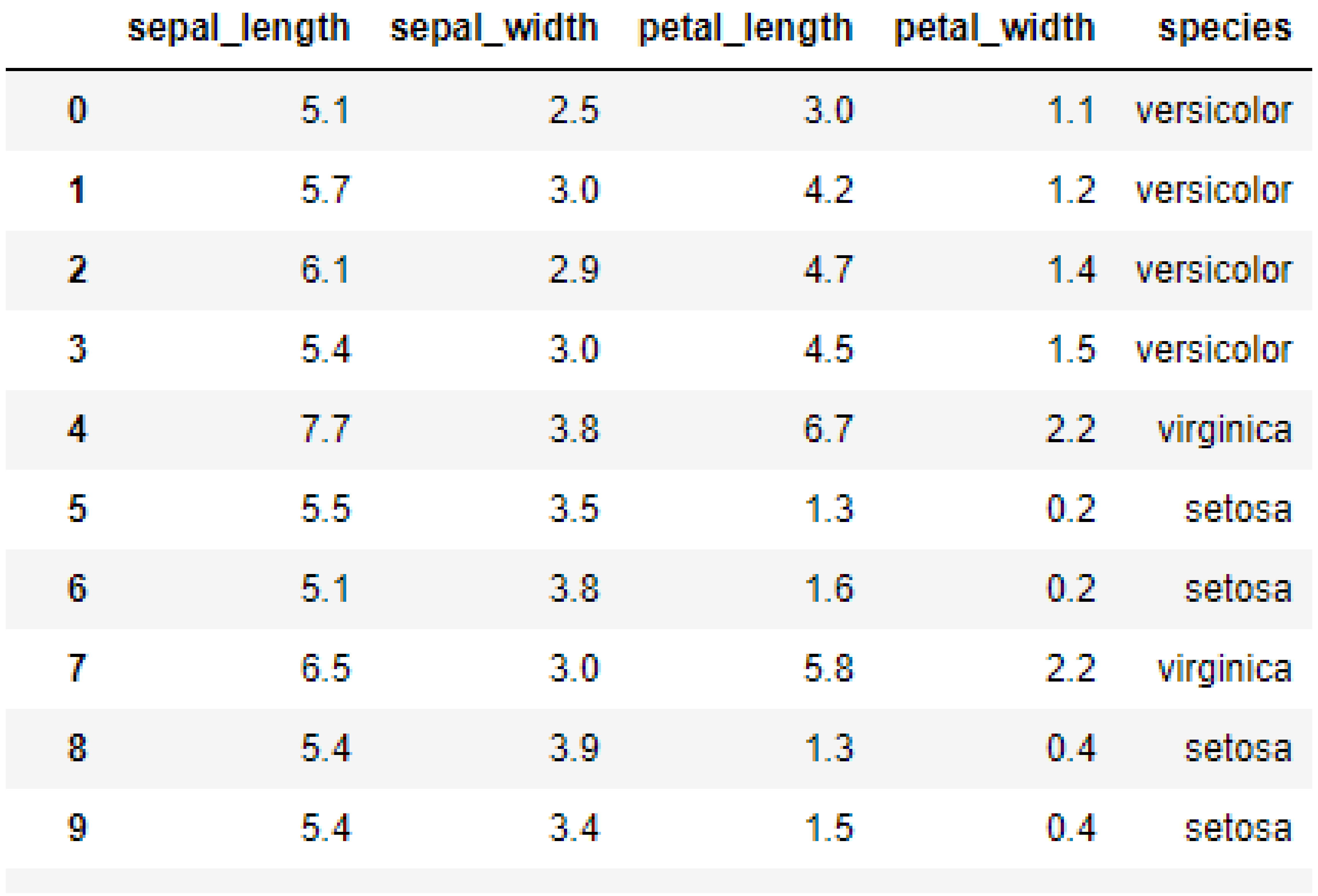
Iris Setosa

Iris Versicolor

Iris VIrginica

EXAMPLE: IRIS DATASET

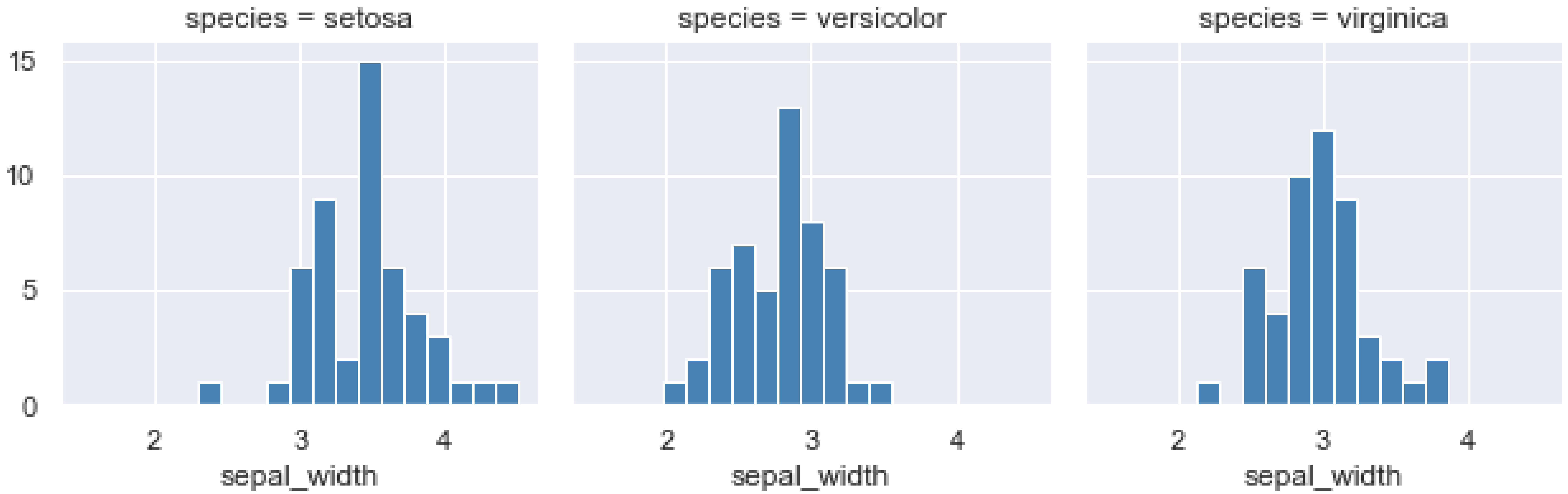




	sepal_length	sepal_width	petal_length	petal_width	species
0	5.1	2.5	3.0	1.1	versicolor
1	5.7	3.0	4.2	1.2	versicolor
2	6.1	2.9	4.7	1.4	versicolor
3	5.4	3.0	4.5	1.5	versicolor
4	7.7	3.8	6.7	2.2	virginica
5	5.5	3.5	1.3	0.2	setosa
6	5.1	3.8	1.6	0.2	setosa
7	6.5	3.0	5.8	2.2	virginica
8	5.4	3.9	1.3	0.4	setosa
9	5.4	3.4	1.5	0.4	setosa

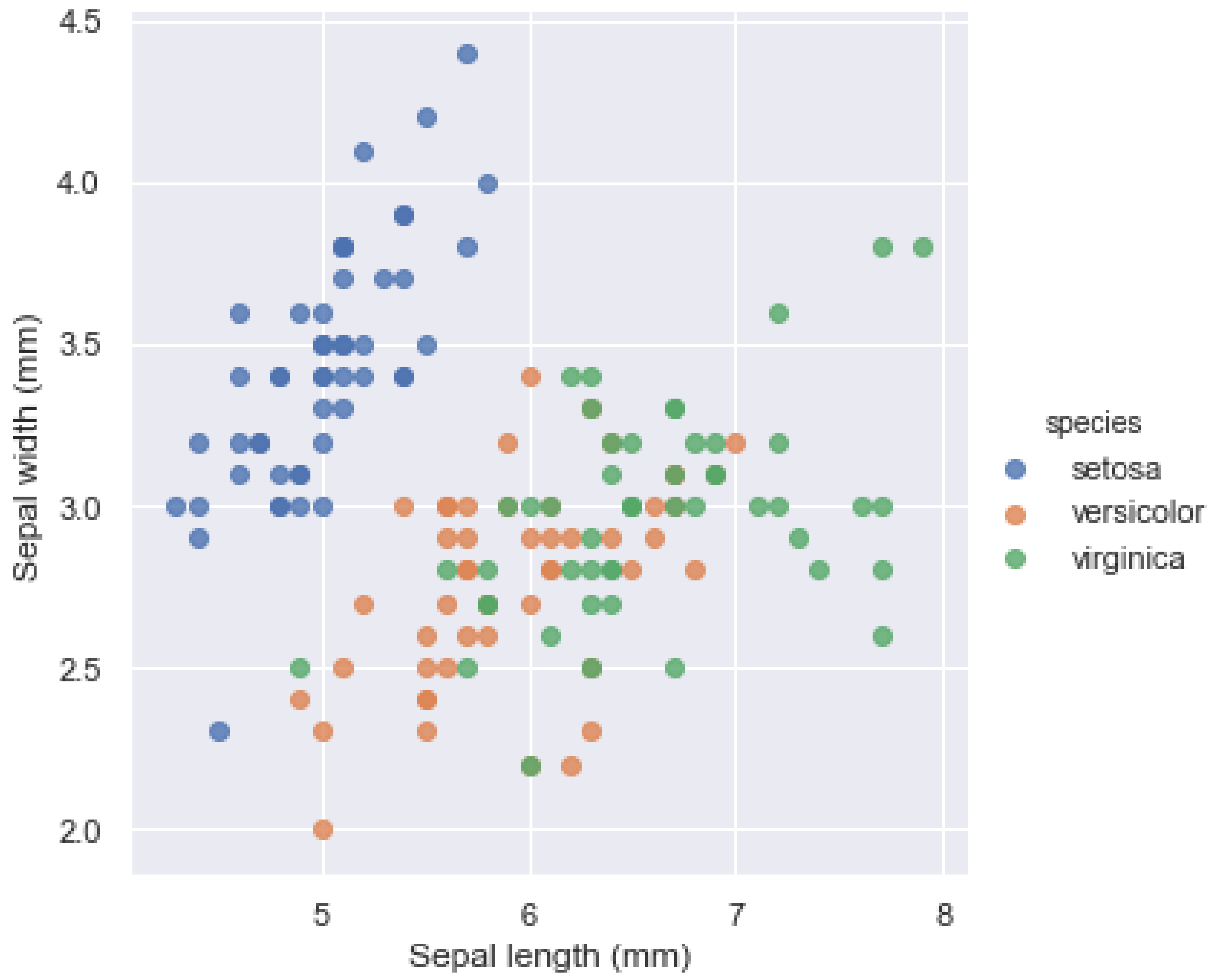
EXAMPLE: IRIS DATASET

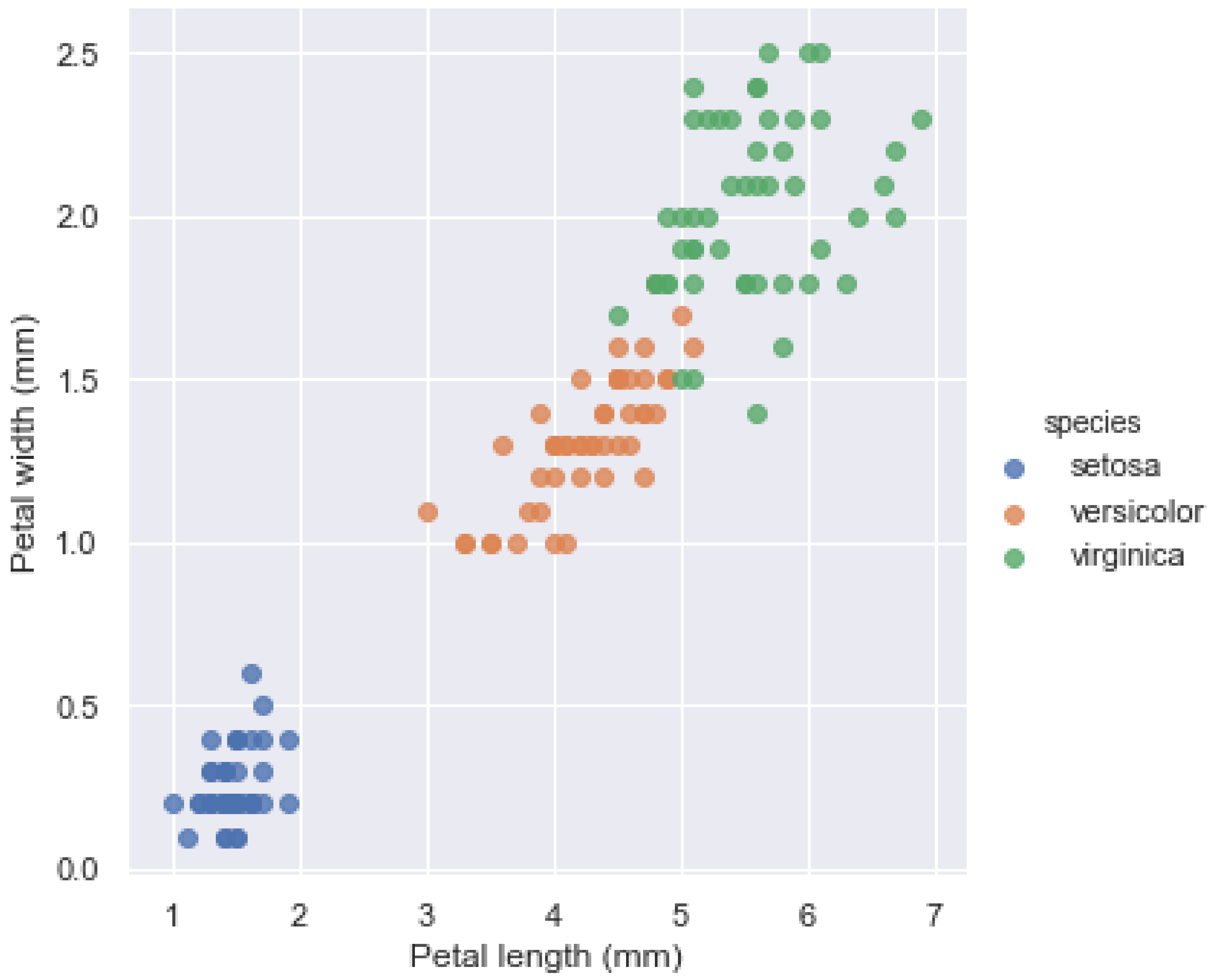
Histogram – each bin shows counts of samples within the bin range



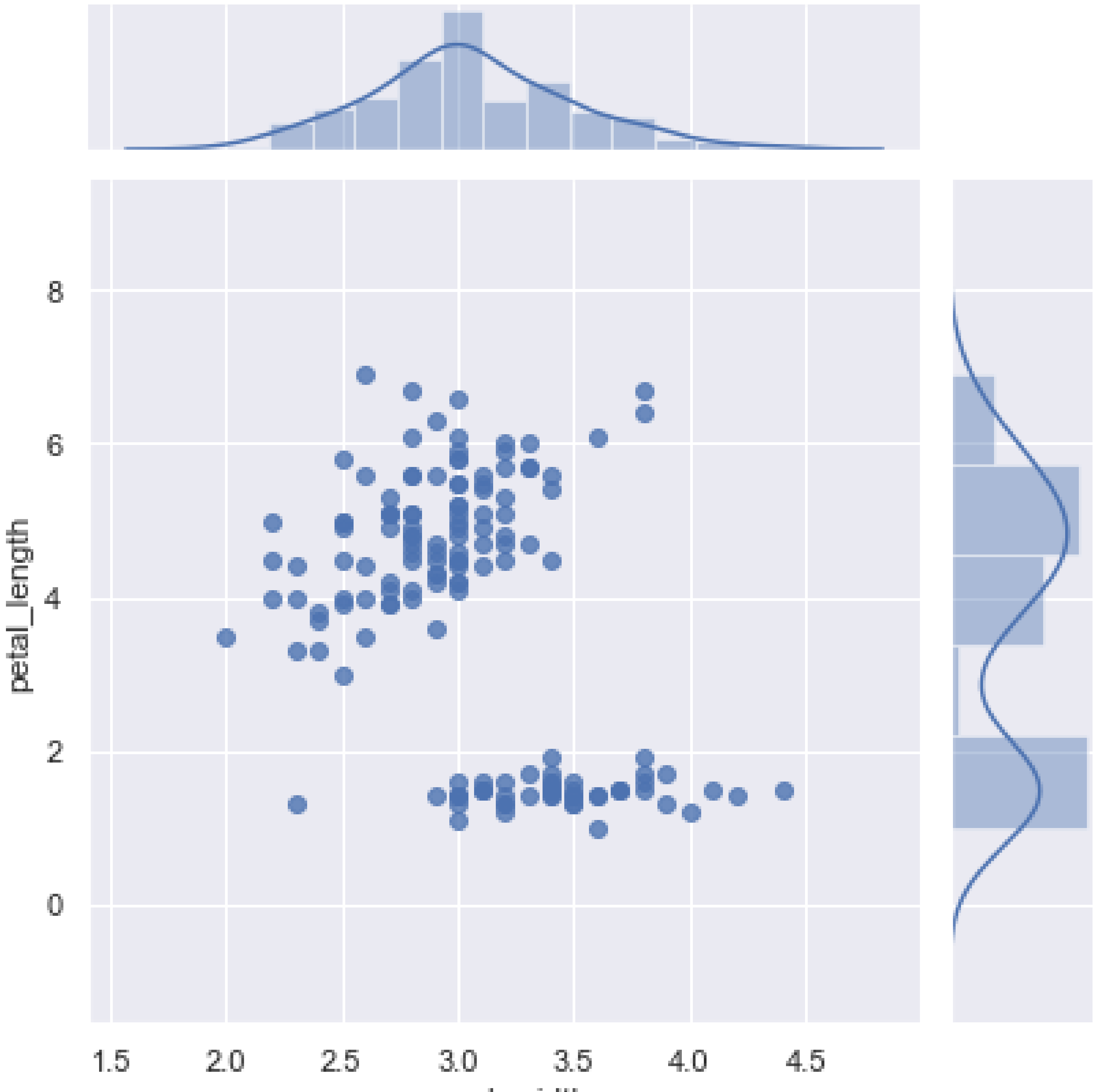
EXAMPLE: IRIS DATASET

Scatterplot – location of the points represents relationships between variables, colors - classes

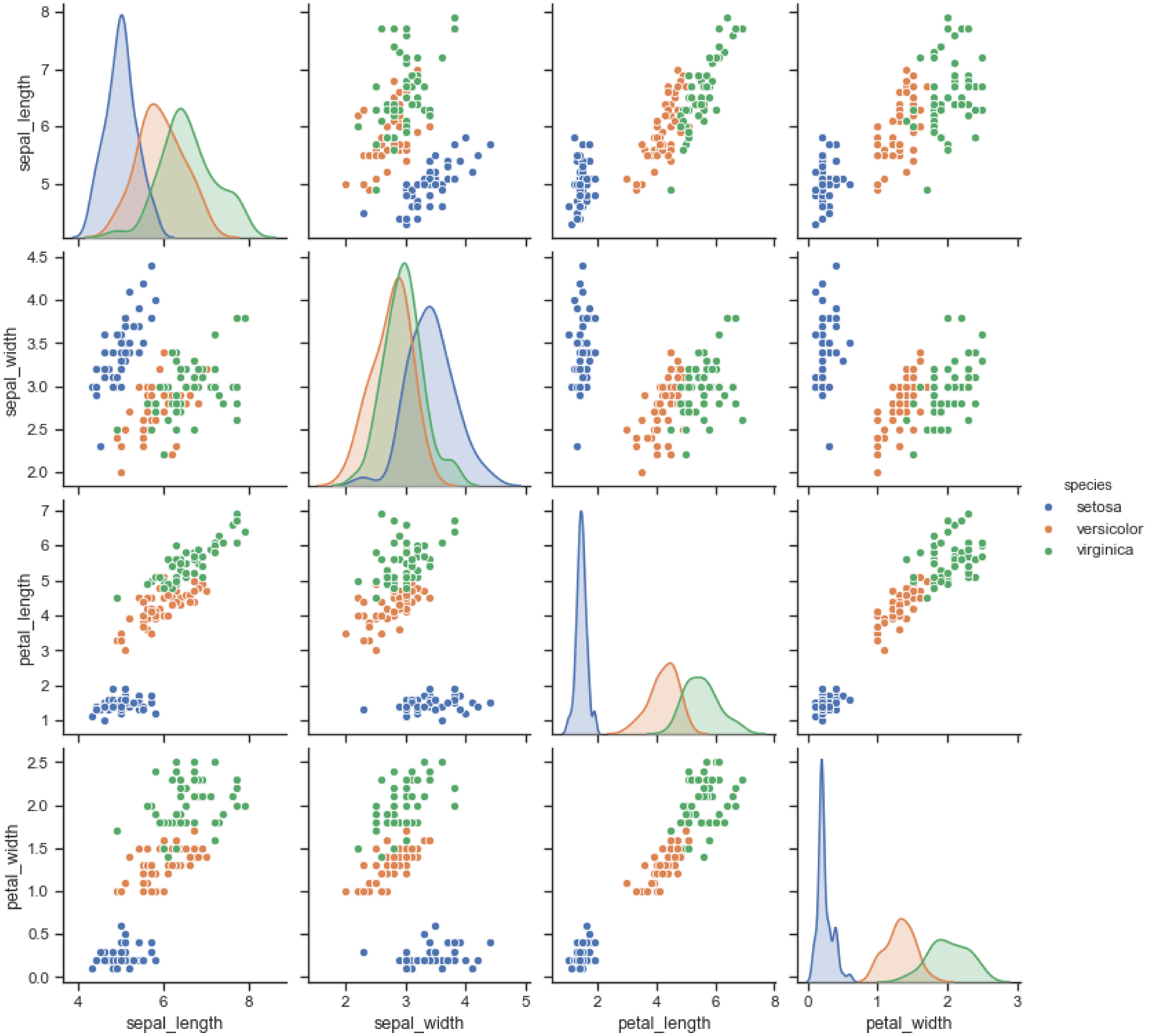




EXAMPLE: IRIS DATASET

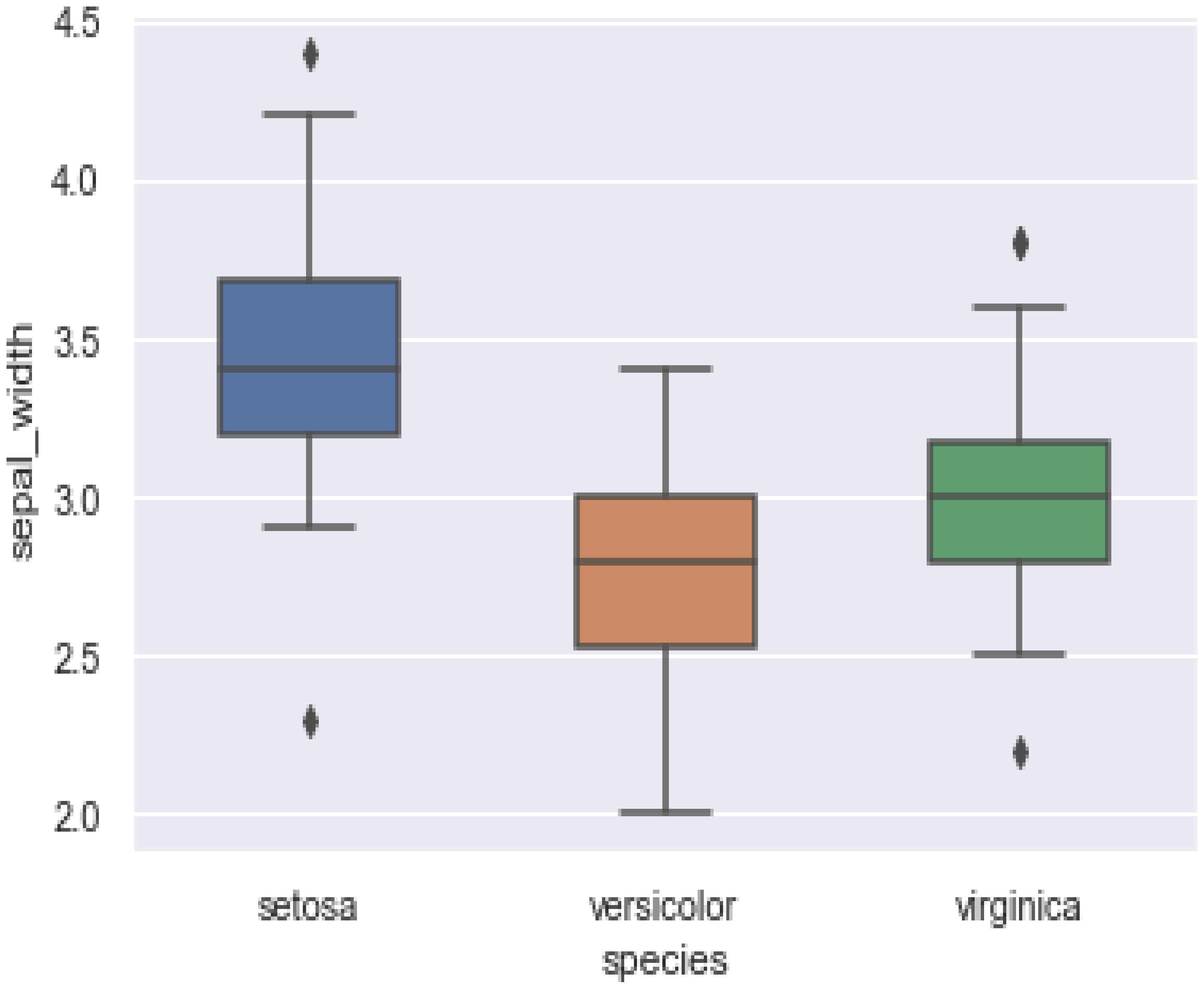


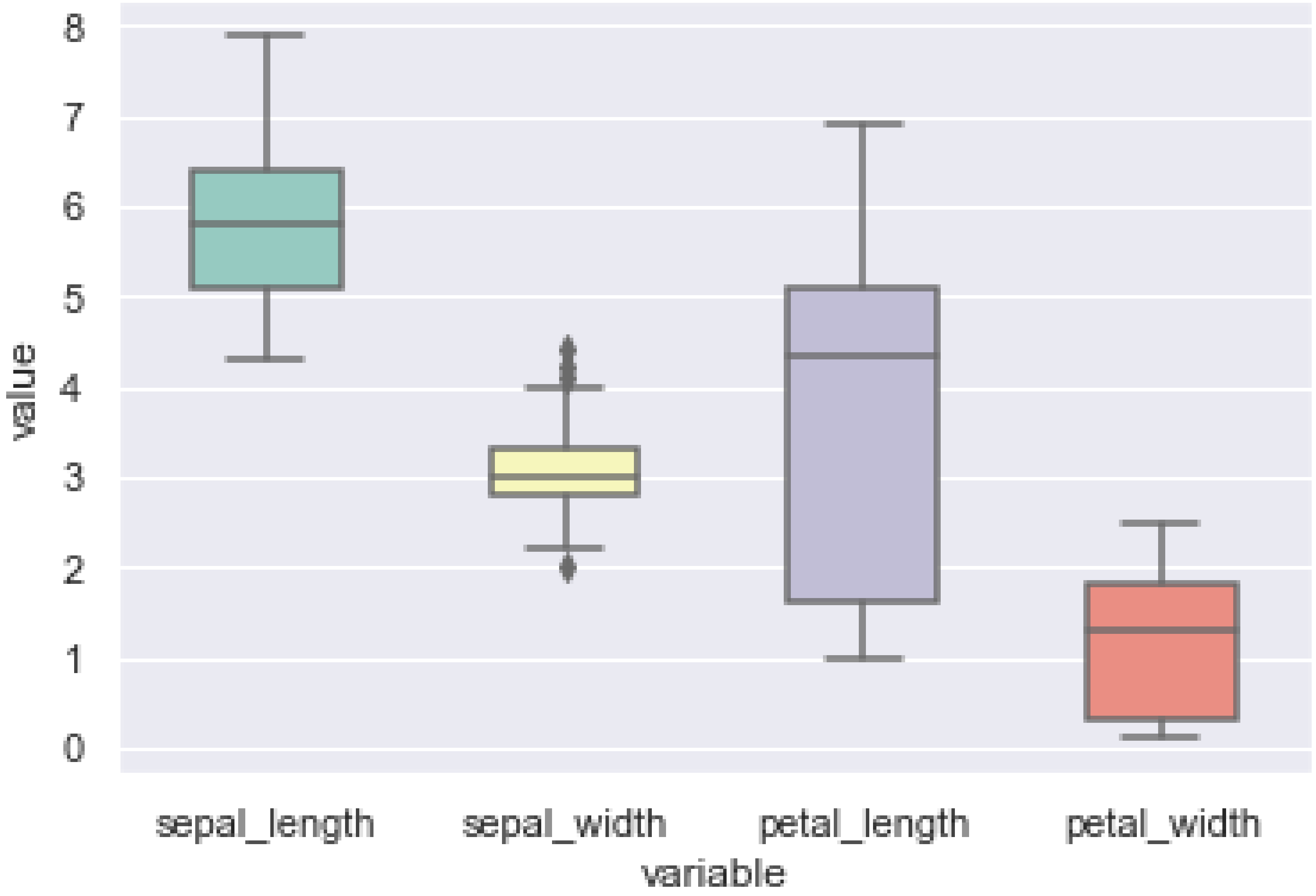
Pair plot
scatterplot matrix



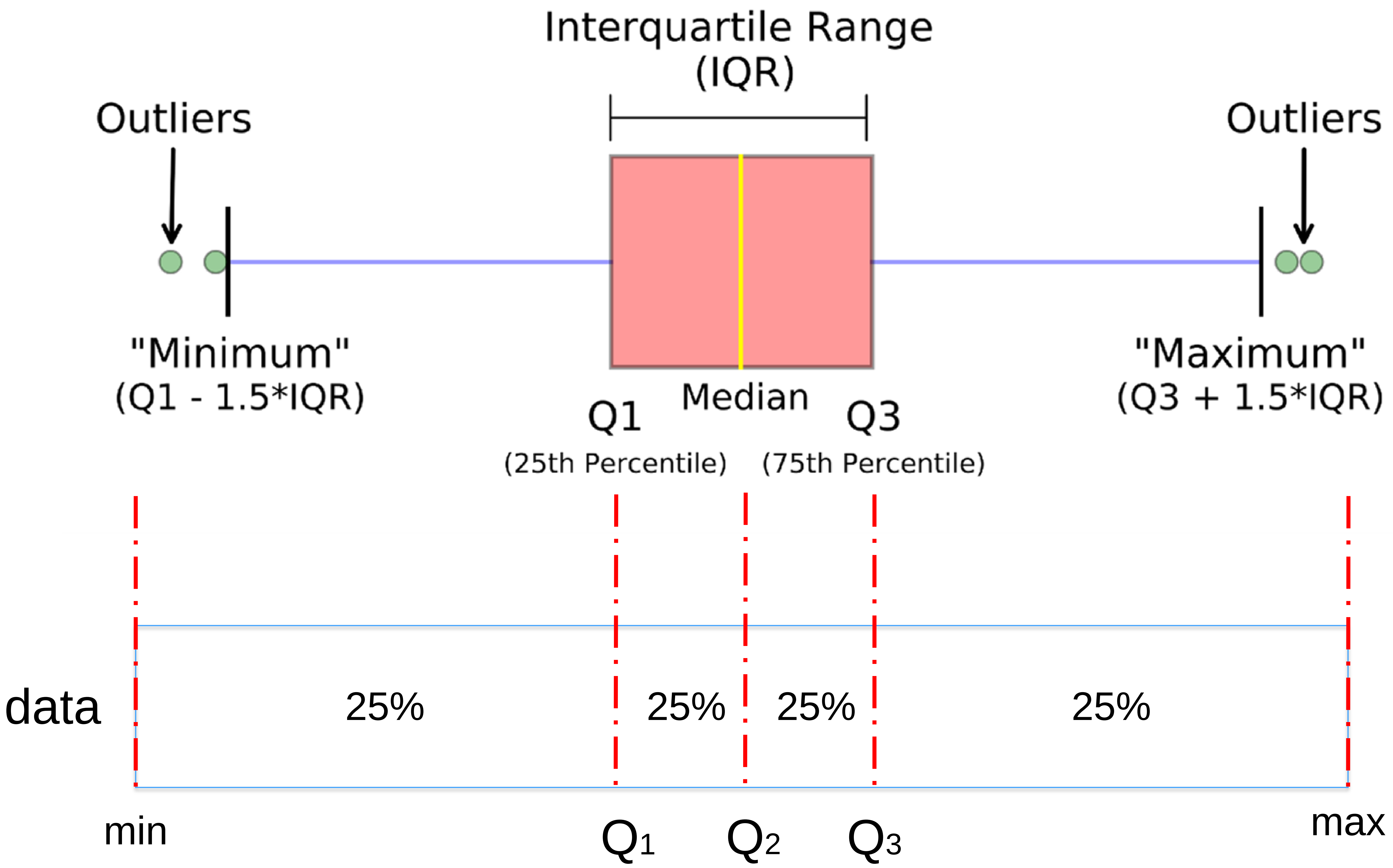
EXAMPLE: IRIS DATASET

Boxplot – location of the points represents relationships between variables, colors - classes

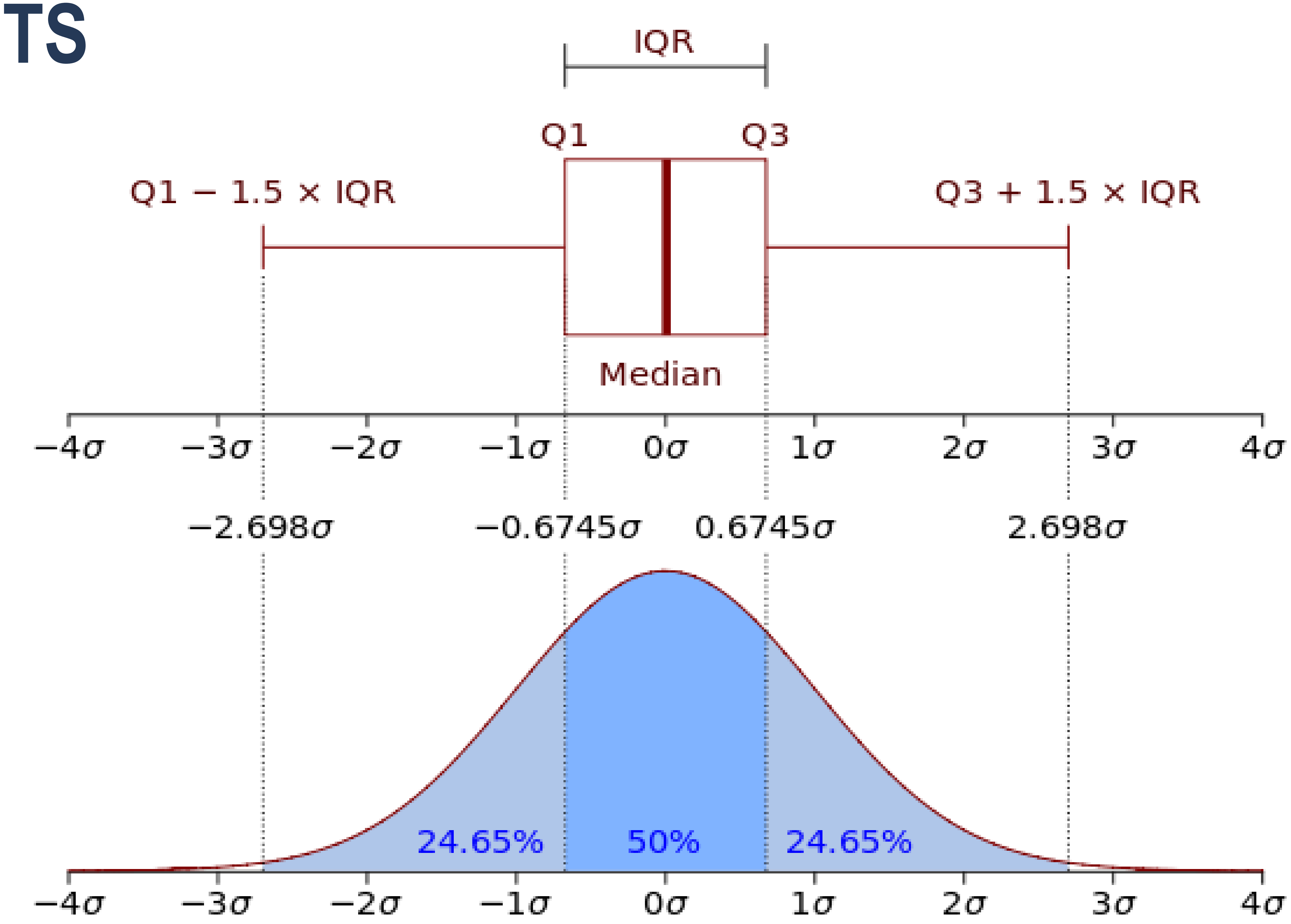




BOXPLOTS


Interquartile Range
(IQR)
Outliers
Outliers
"Minimum"
(Q1 - 1.5*IQR)
Q1
(25th Percentile)
Median
Q3
(75th Percentile)
"Maximum"
(Q3 + 1.5*IQR)
data
25%
25%
25%
25%
min
Q_1
Q_2
Q_3
max

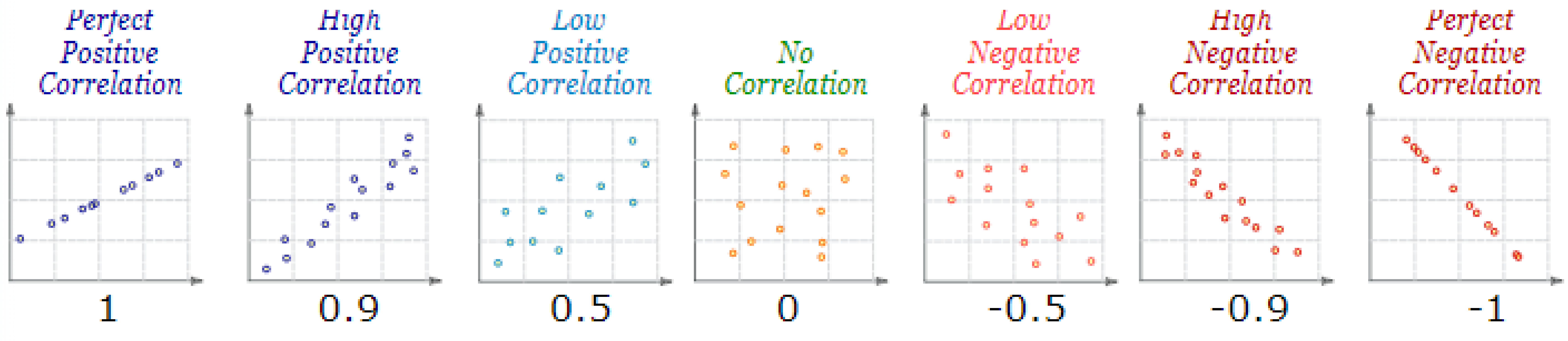
BOXPLOTS


IQR
Q1
Q3
Q1 − 1.5 × IQR
Q3 + 1.5 × IQR
Median
−4σ
−3σ
−2σ
−1σ
0σ
1σ
2σ
3σ
4σ
−2.698σ
−0.6745σ
0.6745σ
2.698σ
24.65%
50%
24.65%
−4σ
−3σ
−2σ
−1σ
0σ
1σ
2σ
3σ
4σ

DATA DEPENDENCE - CORRELATIONS

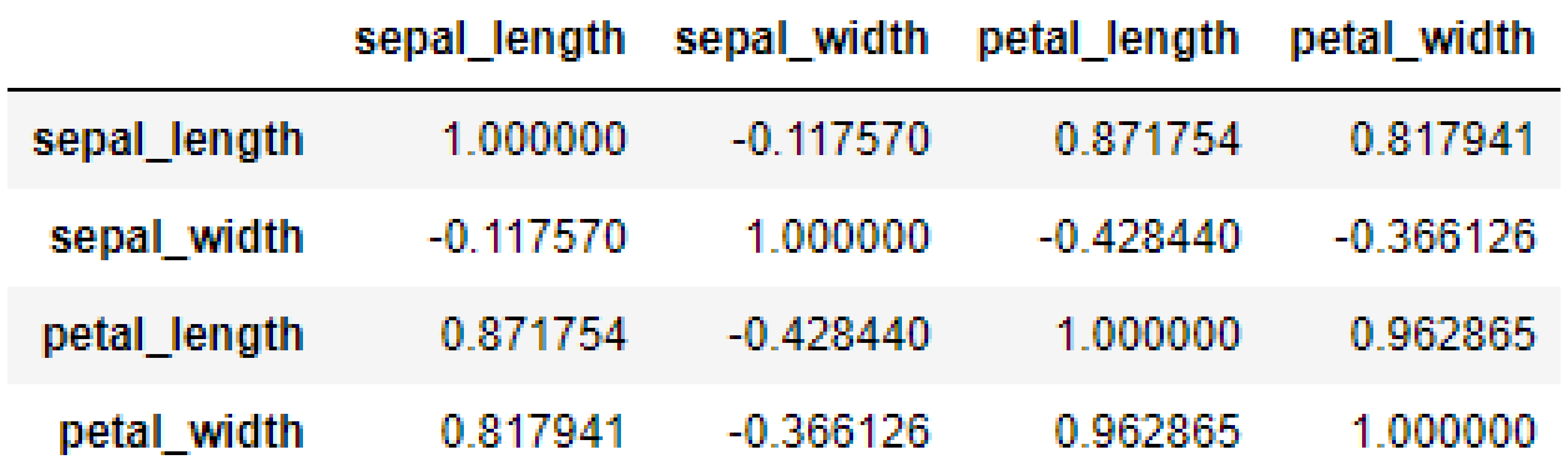
Correlation coefficient – measure of dependency between two variables (how much they change together)

Person correlation coefficient
$$r_{XY} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}}$$

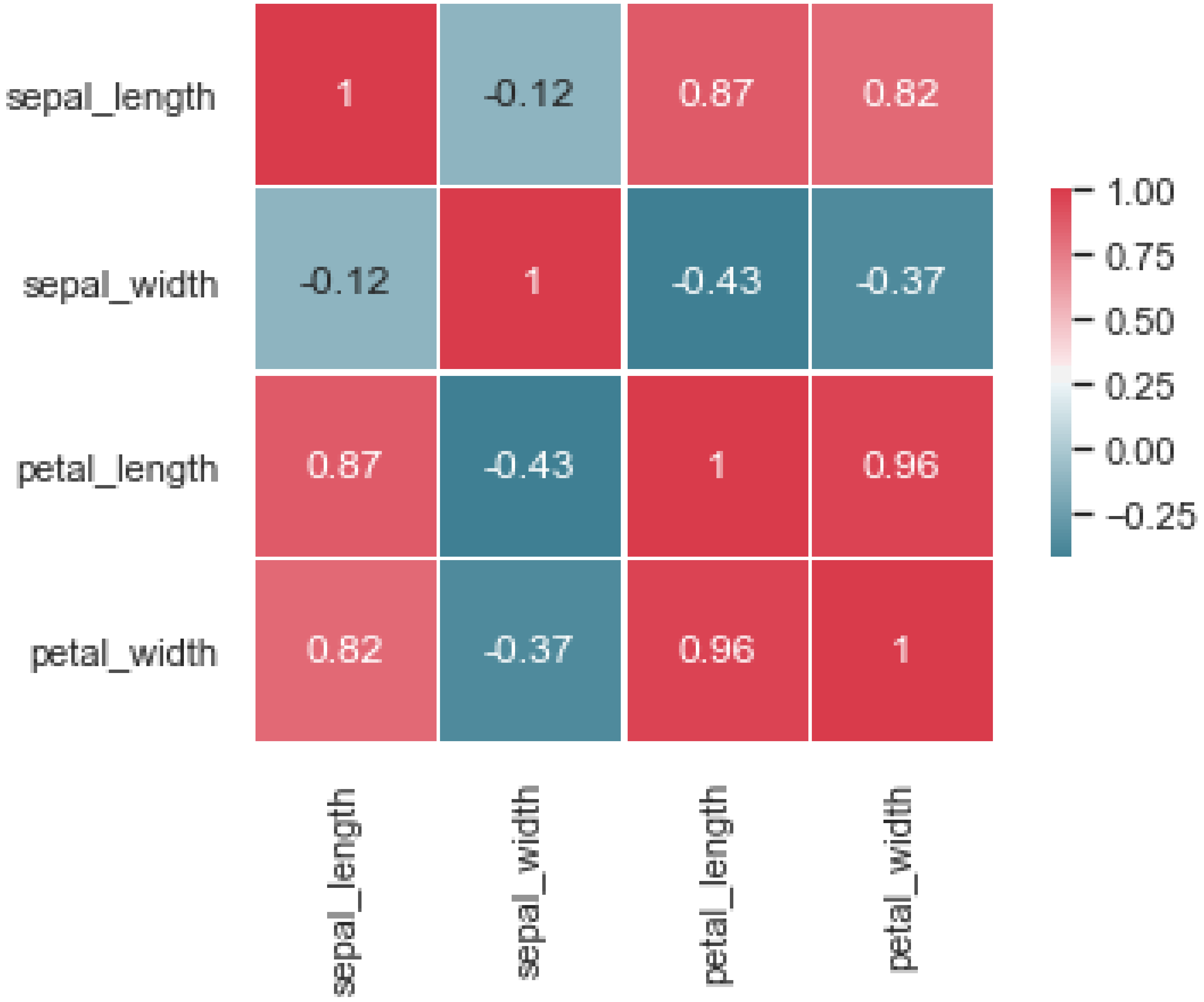


EXAMPLE: IRIS DATASET

Correlation matrix

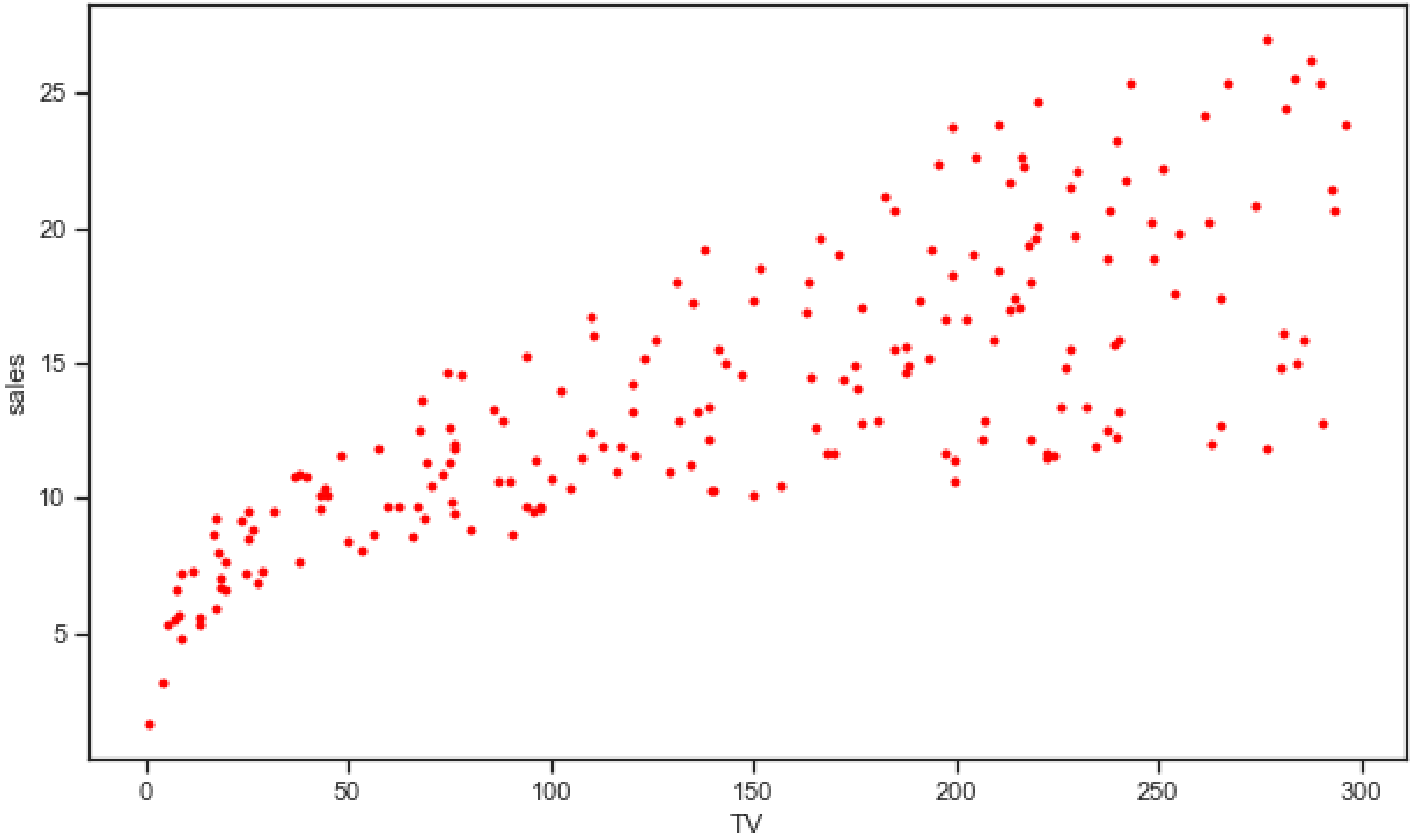


	sepal_length	sepal_width	petal_length	petal_width
sepal_length	1.000000	-0.117570	0.871754	0.817941
sepal_width	-0.117570	1.000000	-0.428440	-0.366126
petal_length	0.871754	-0.428440	1.000000	0.962865
petal_width	0.817941	-0.366126	0.962865	1.000000



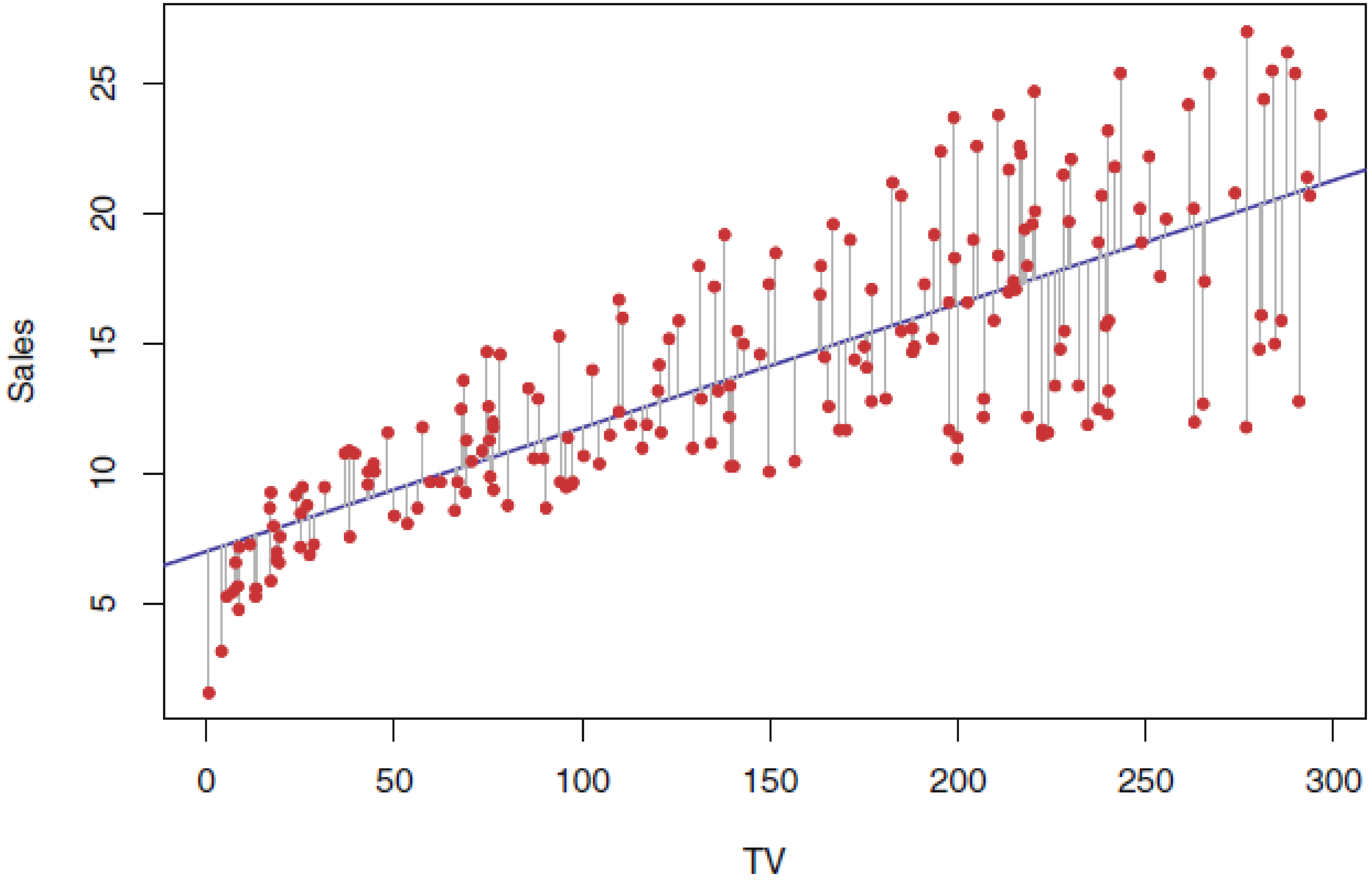
SIMPLE LINEAR REGRESSION

Linear trend in the data



SIMPLE LINEAR REGRESSION

Linear trend in the data

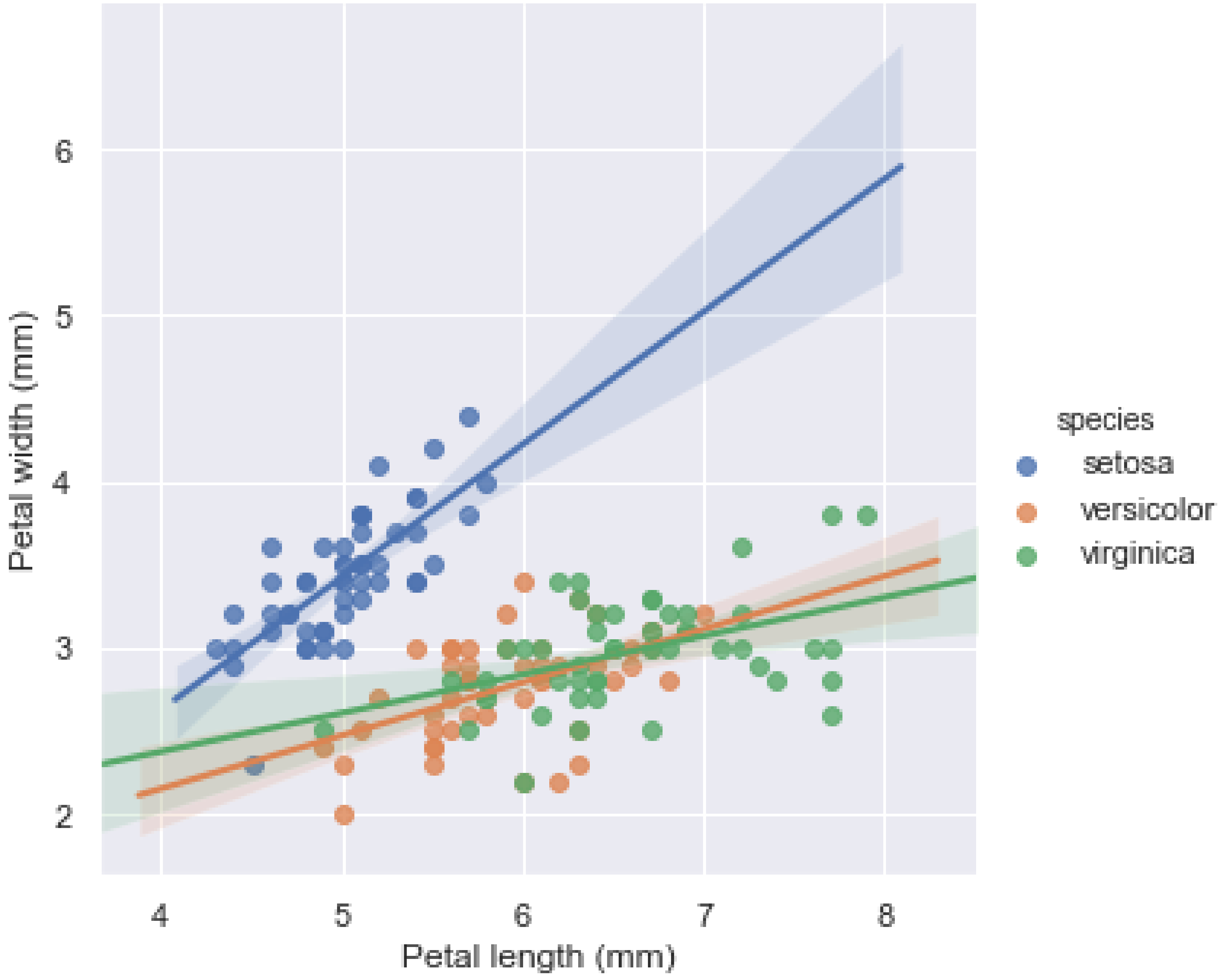


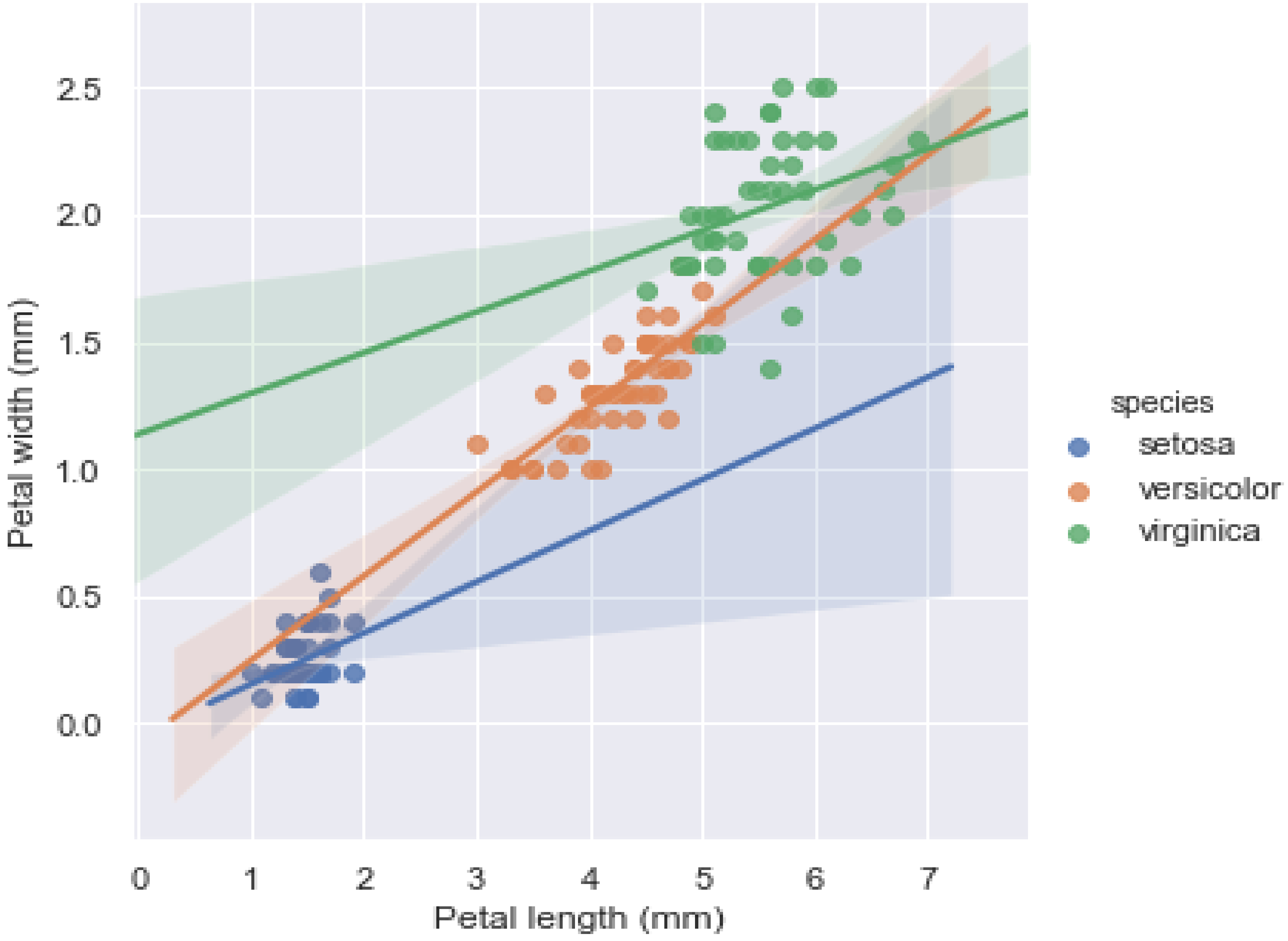
Sales = a * TV + b

y = a *x + b

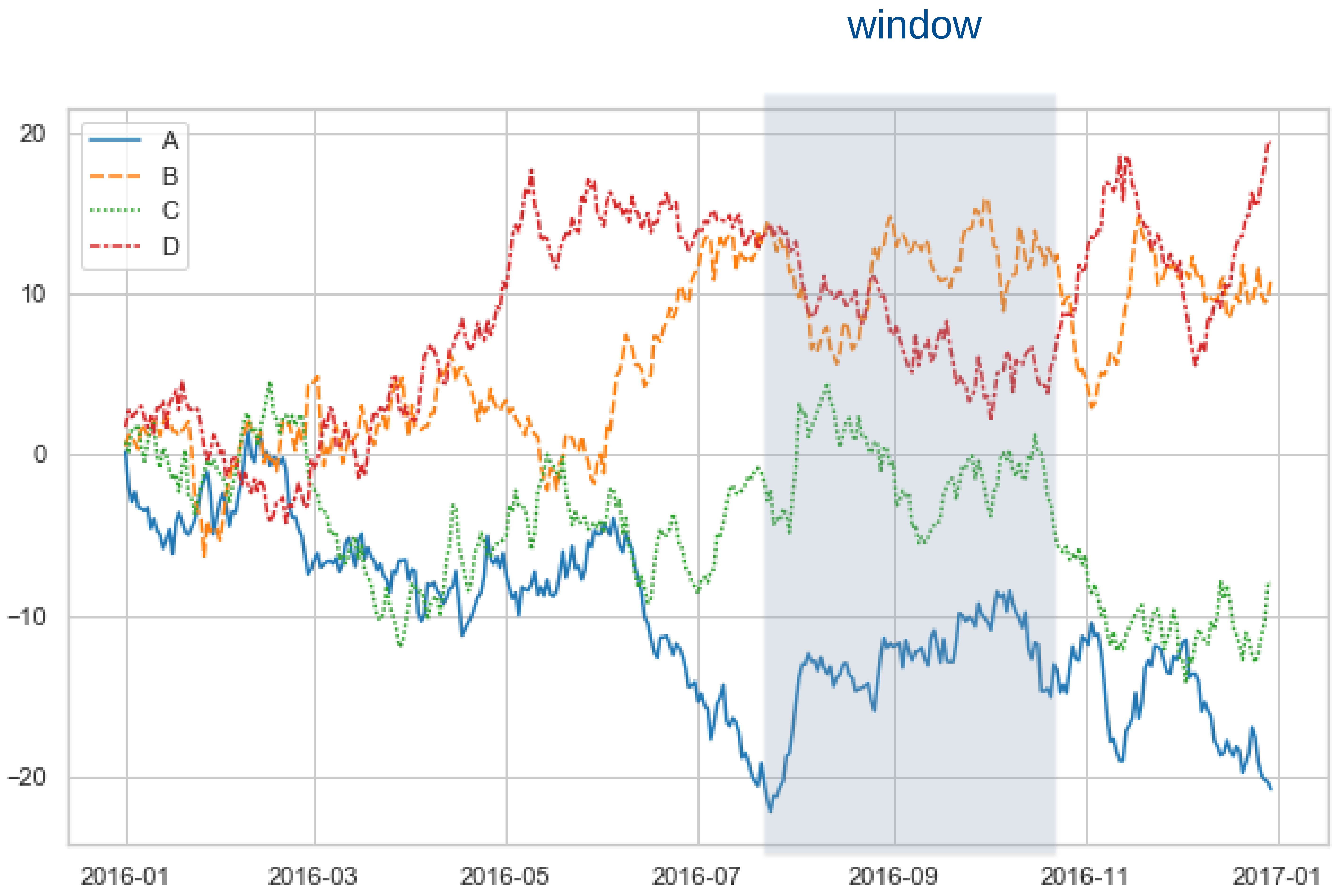
EXAMPLE: IRIS DATASET

Scatterplot + trend lines (linear regression)



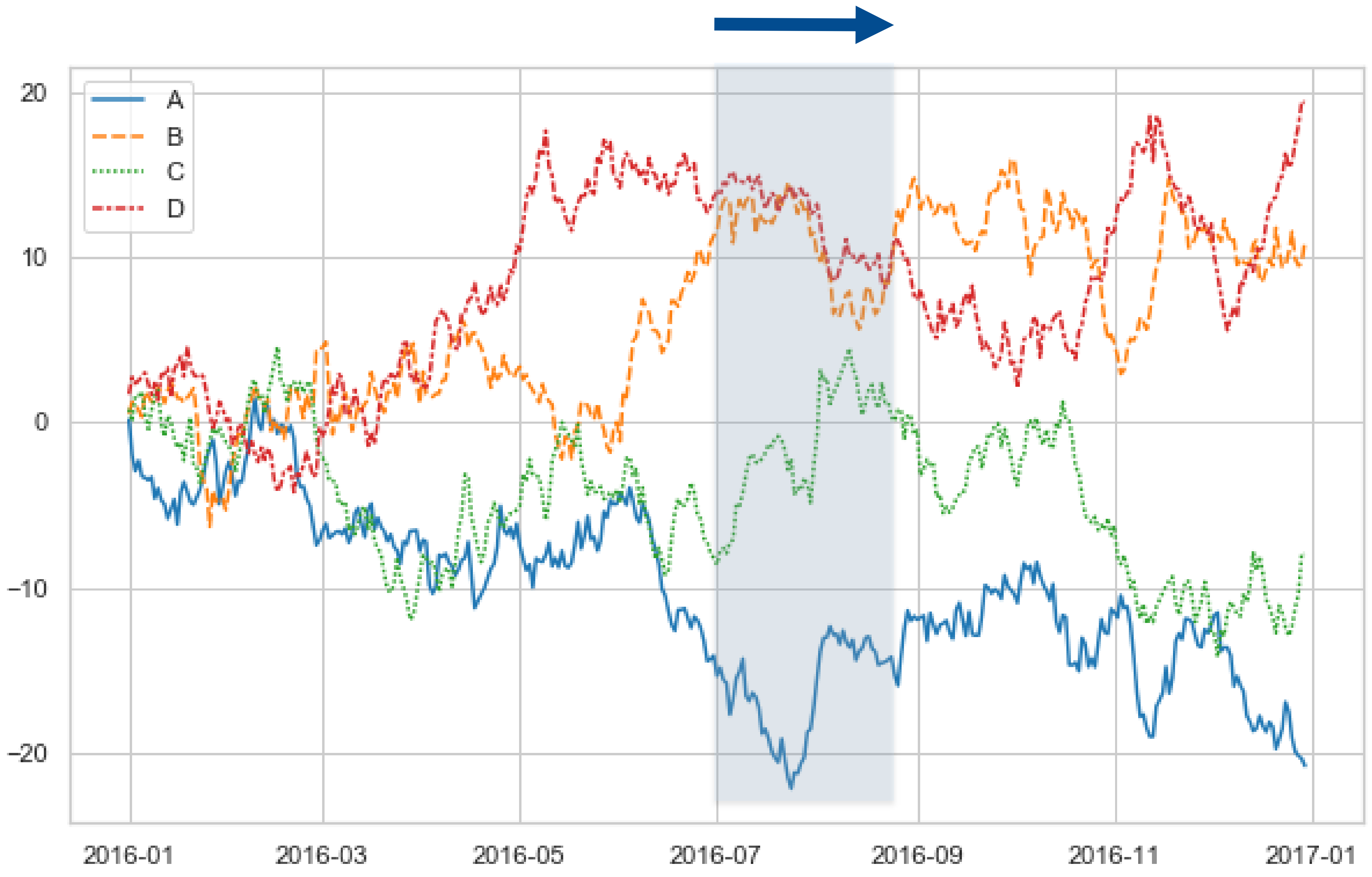


TIME DEPENDENT DATA

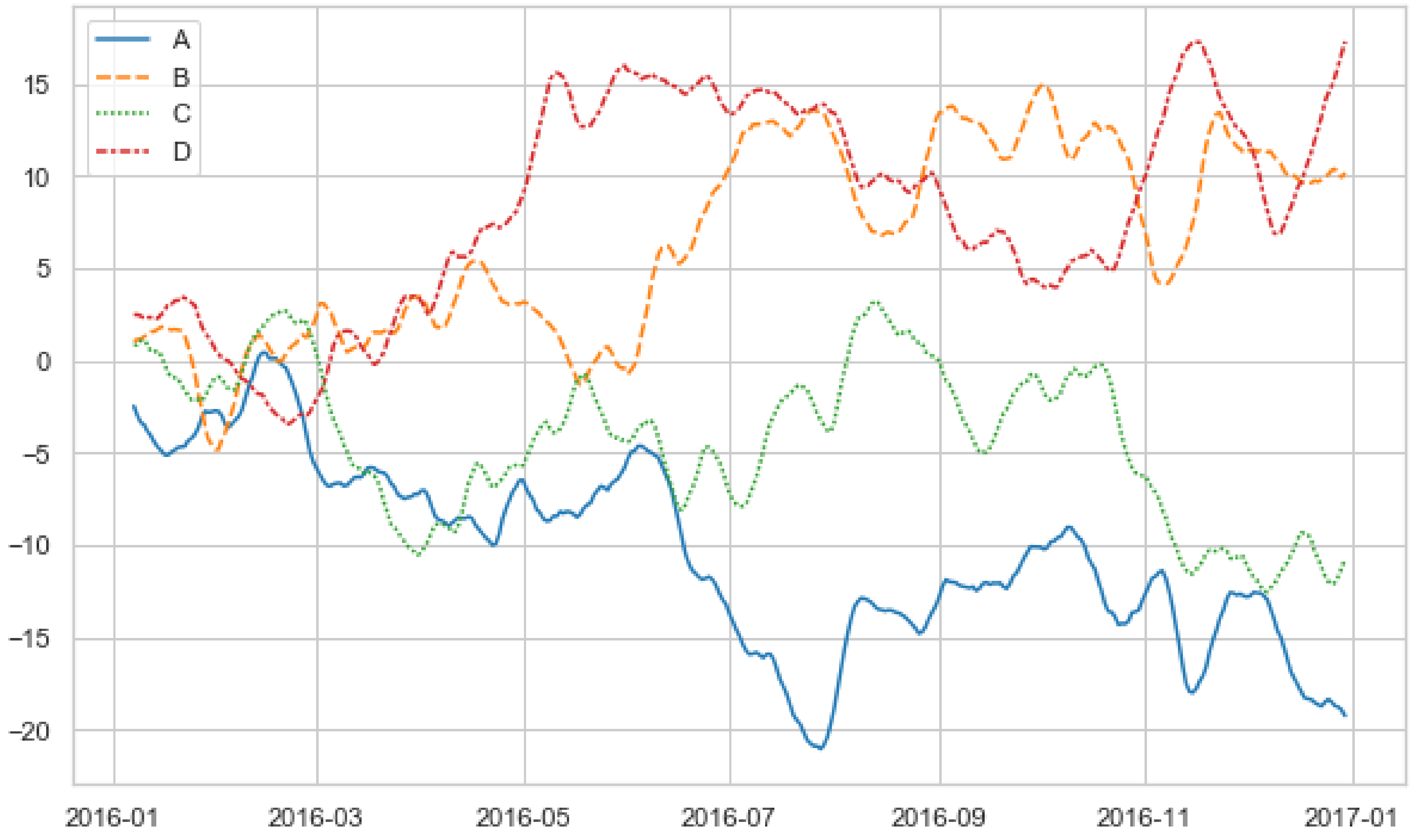


TIME DEPENDENT DATA

Filtering data: smoothing

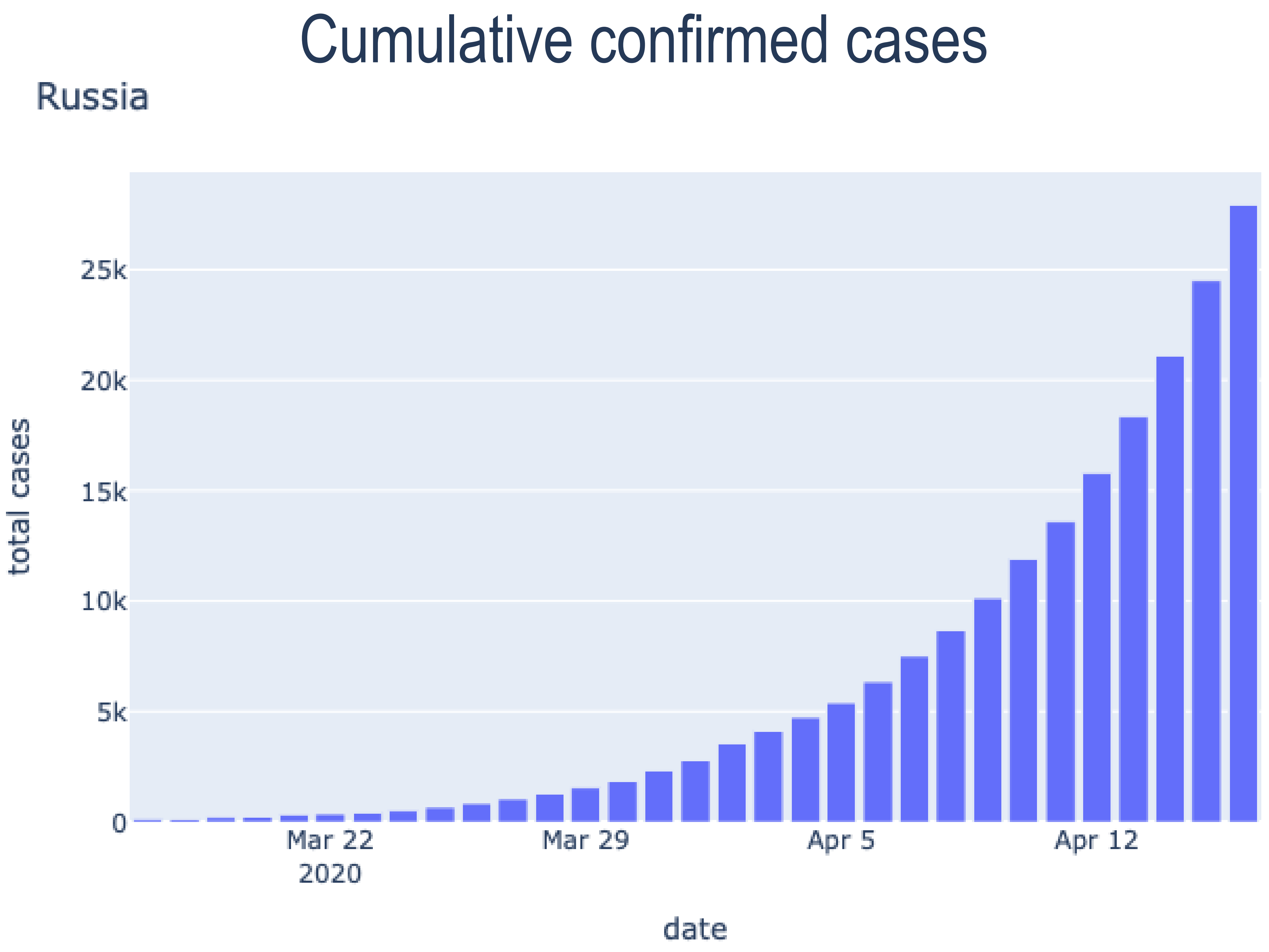


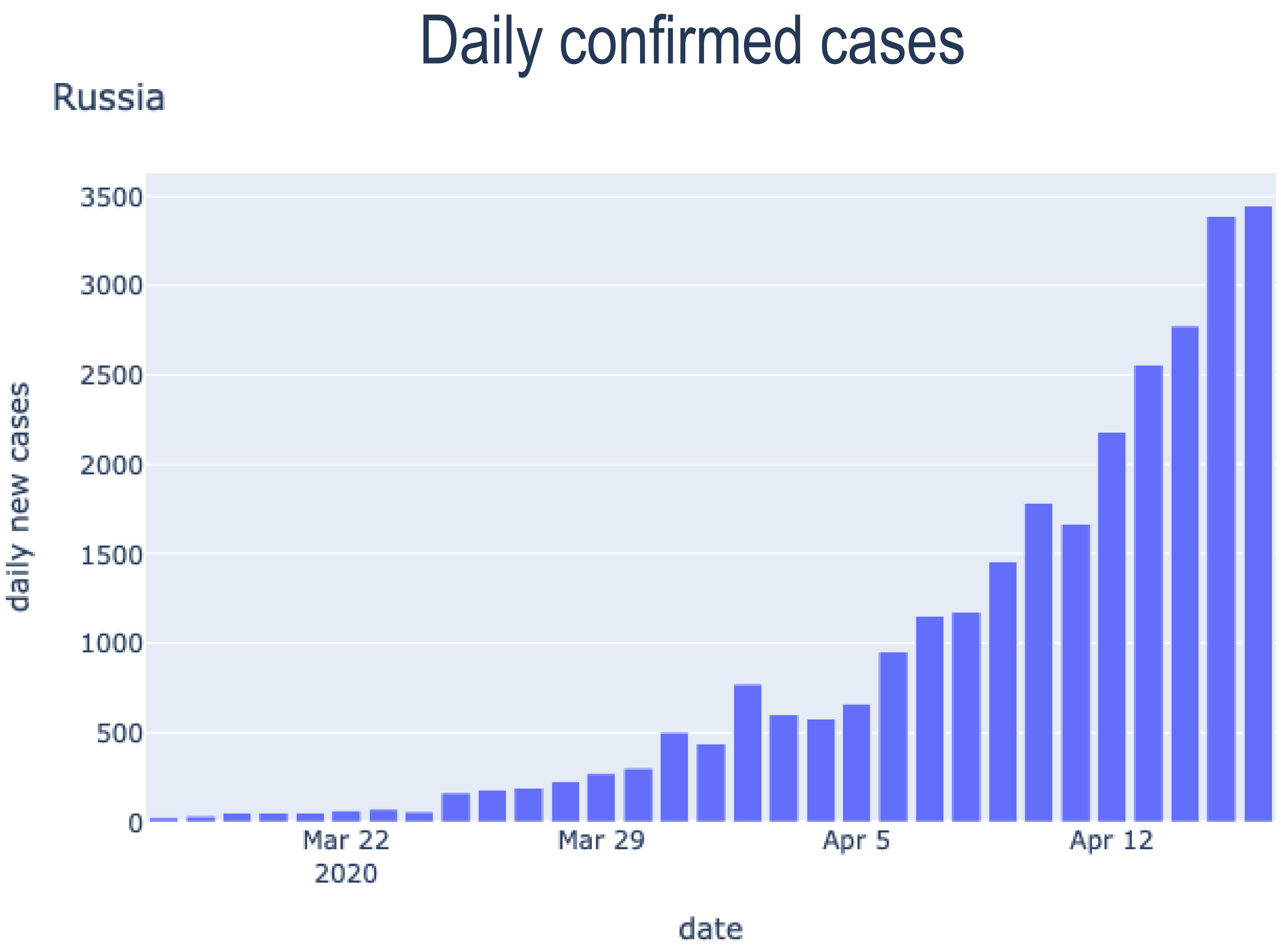
Sliding window "mean" filter



COVID-19 DATA

Russia

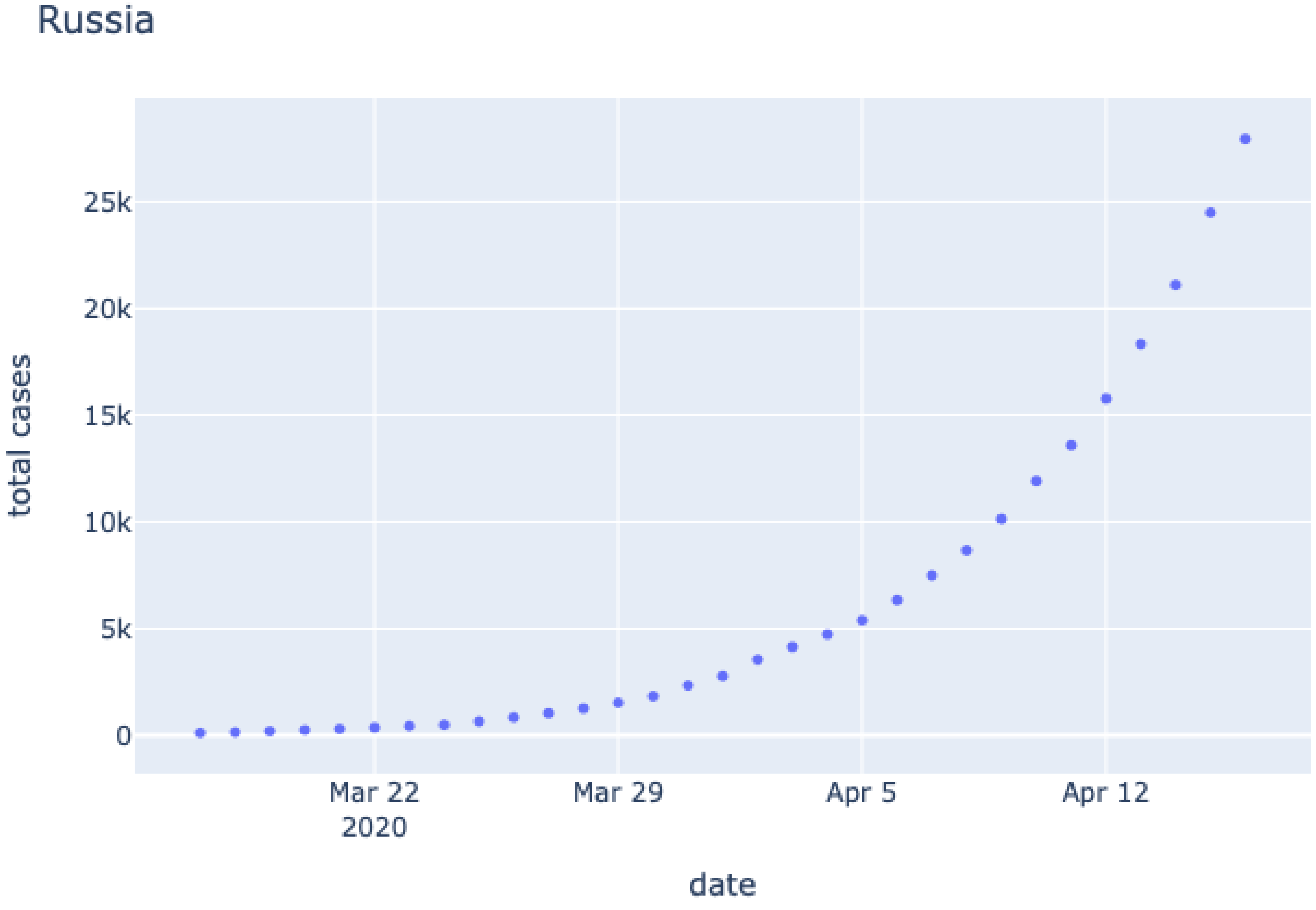




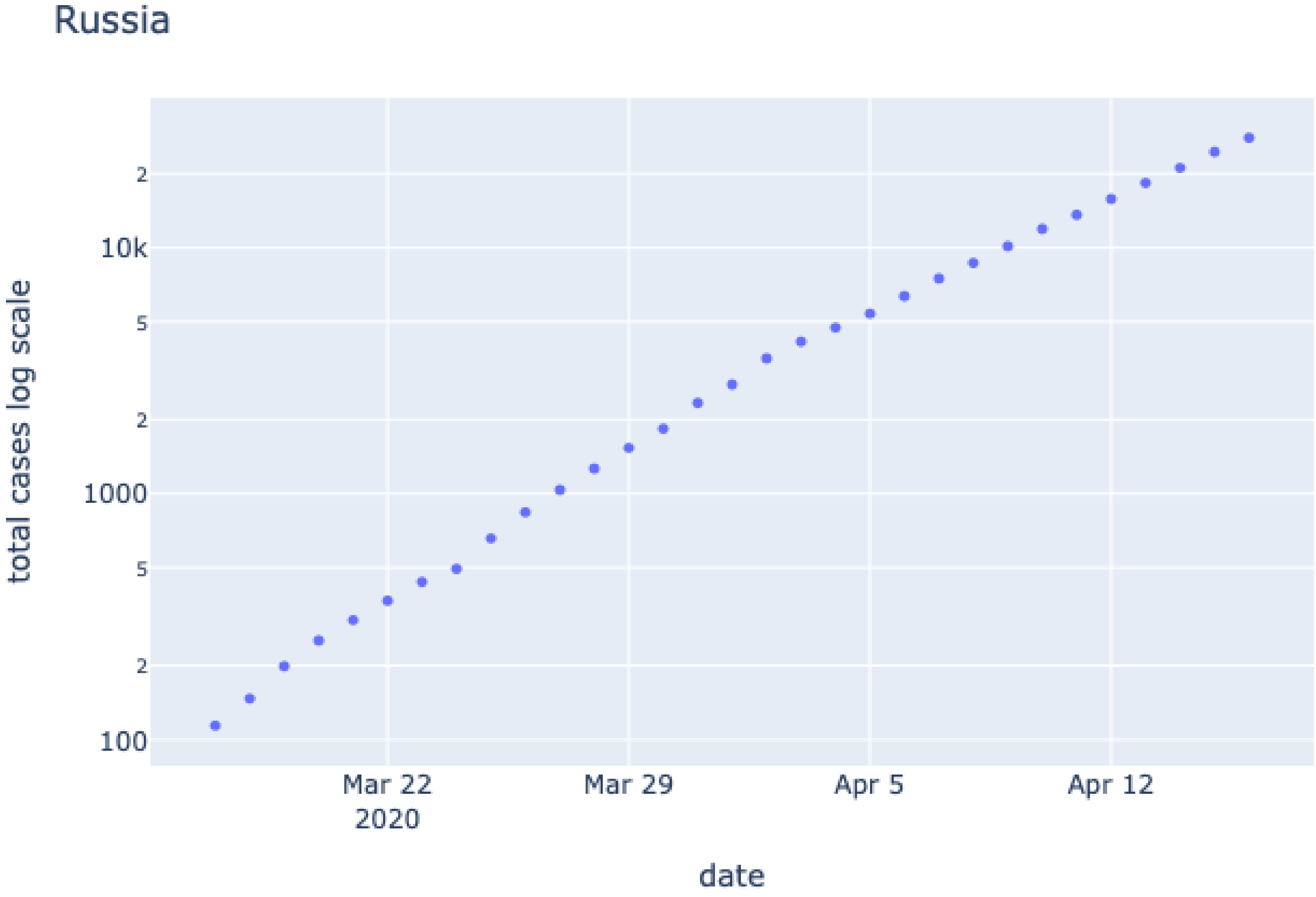
COVID-19 DATA

Russia

Cumulative confirmed cases



Cumulative confirmed cases, log scale



COVID-19 DATA

Russia

Naive growth model:

$N(t+1) = R*N(t)$

$N(1) = R*N(0)$

$N(2) = R*N(1) = R^2*N(0)$

$N(t) = R^t * N(0)$

$\text{daily\_N}(t) = N(t+1) - N(t)$

$\text{daily\_N}(t) = (R-1)*N(t)$

$\text{daily\_N}(t+1) = R*\text{daily\_N}(t)$

How to find R?

$\log N(t) = \log(R^t * N(0))$

$\log N(t) = t*\log R + \log N(0)$

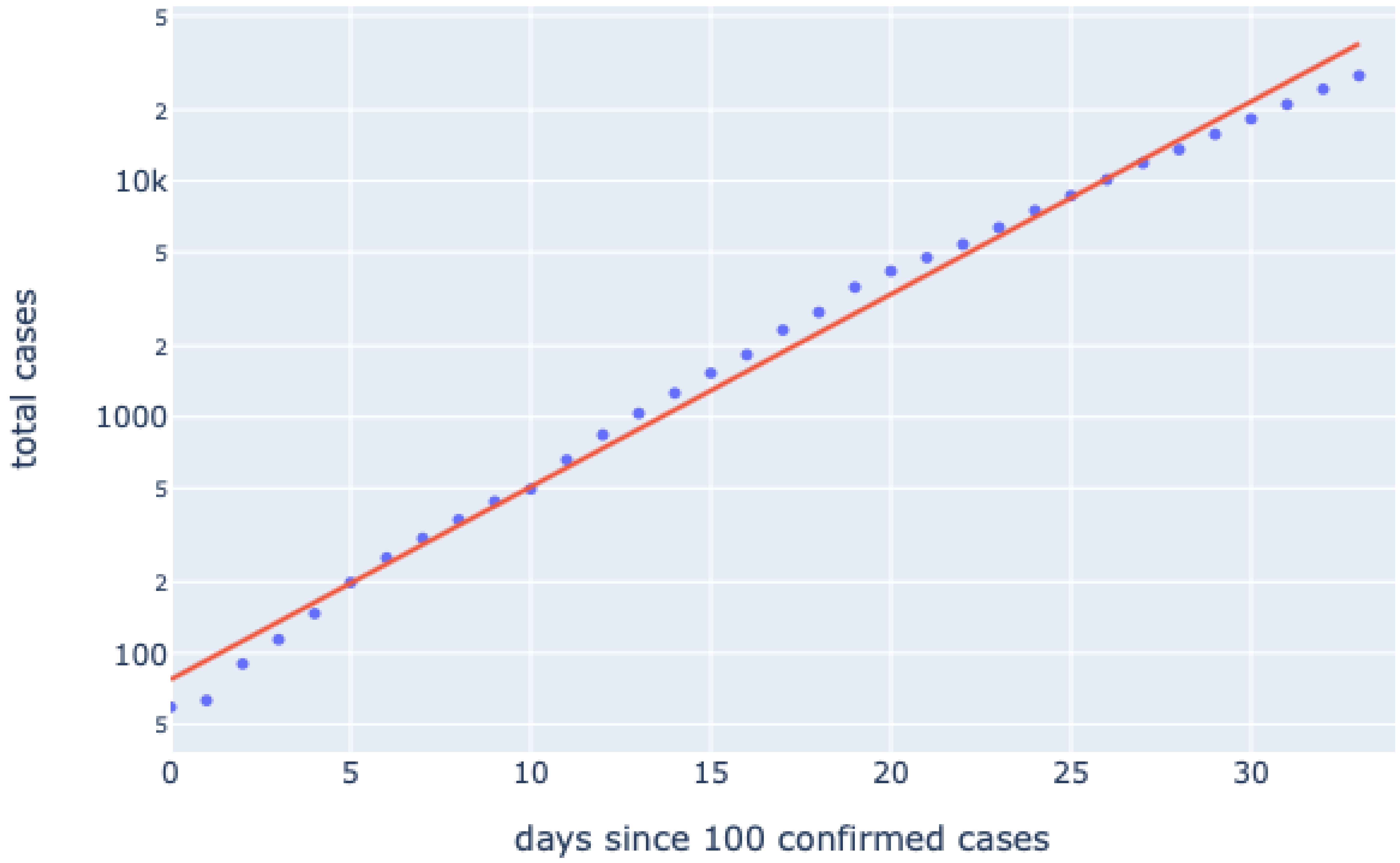
$y = a*t + b$

$a = \log R$

$R = \exp(a)$

COVID-19 DATA

Russia

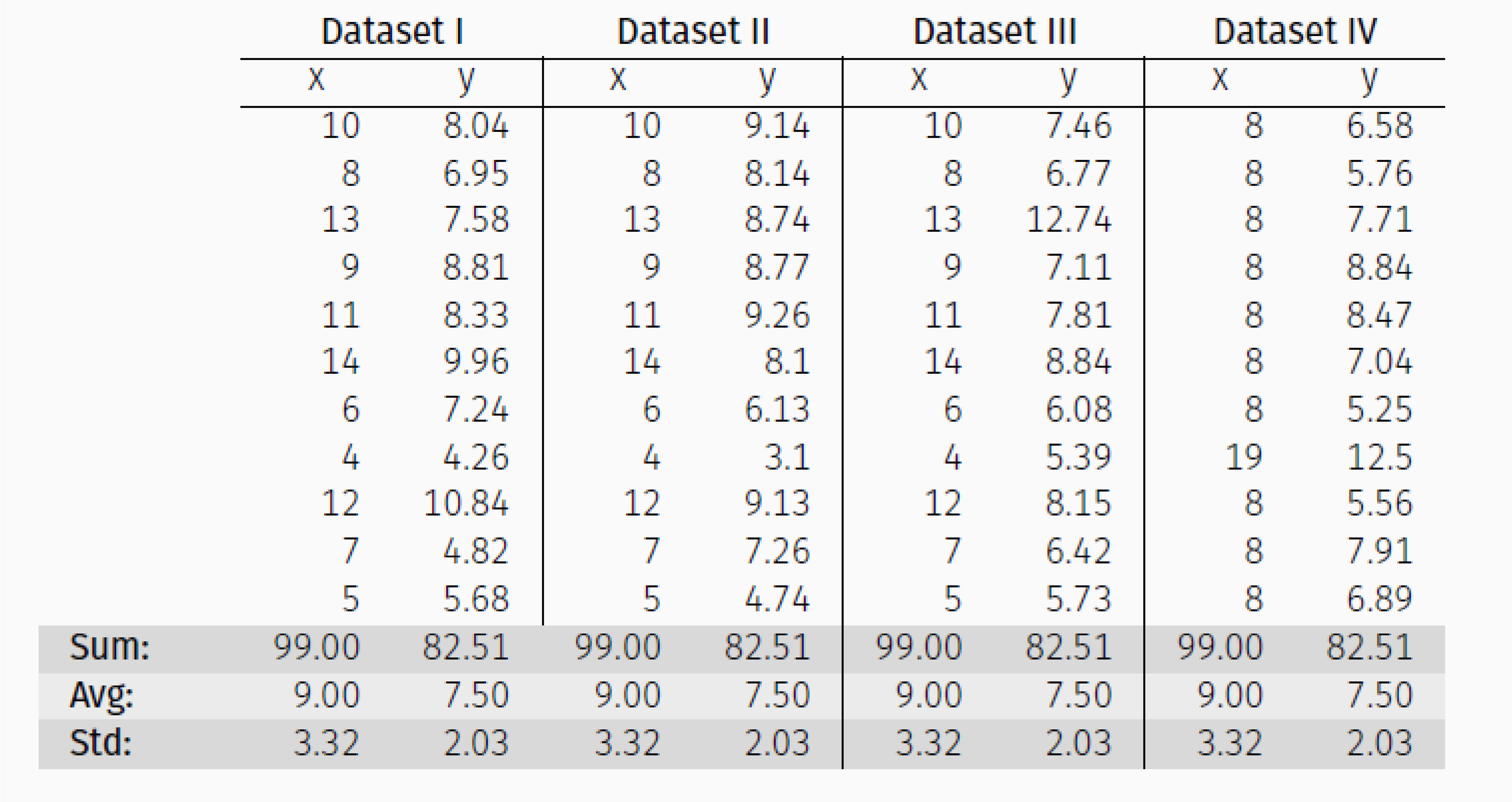


y = 0.18*t + 4.34
R = 1.20

Average 20% daily growth since March 17th

WHY PLOTTING YOUR DATA?

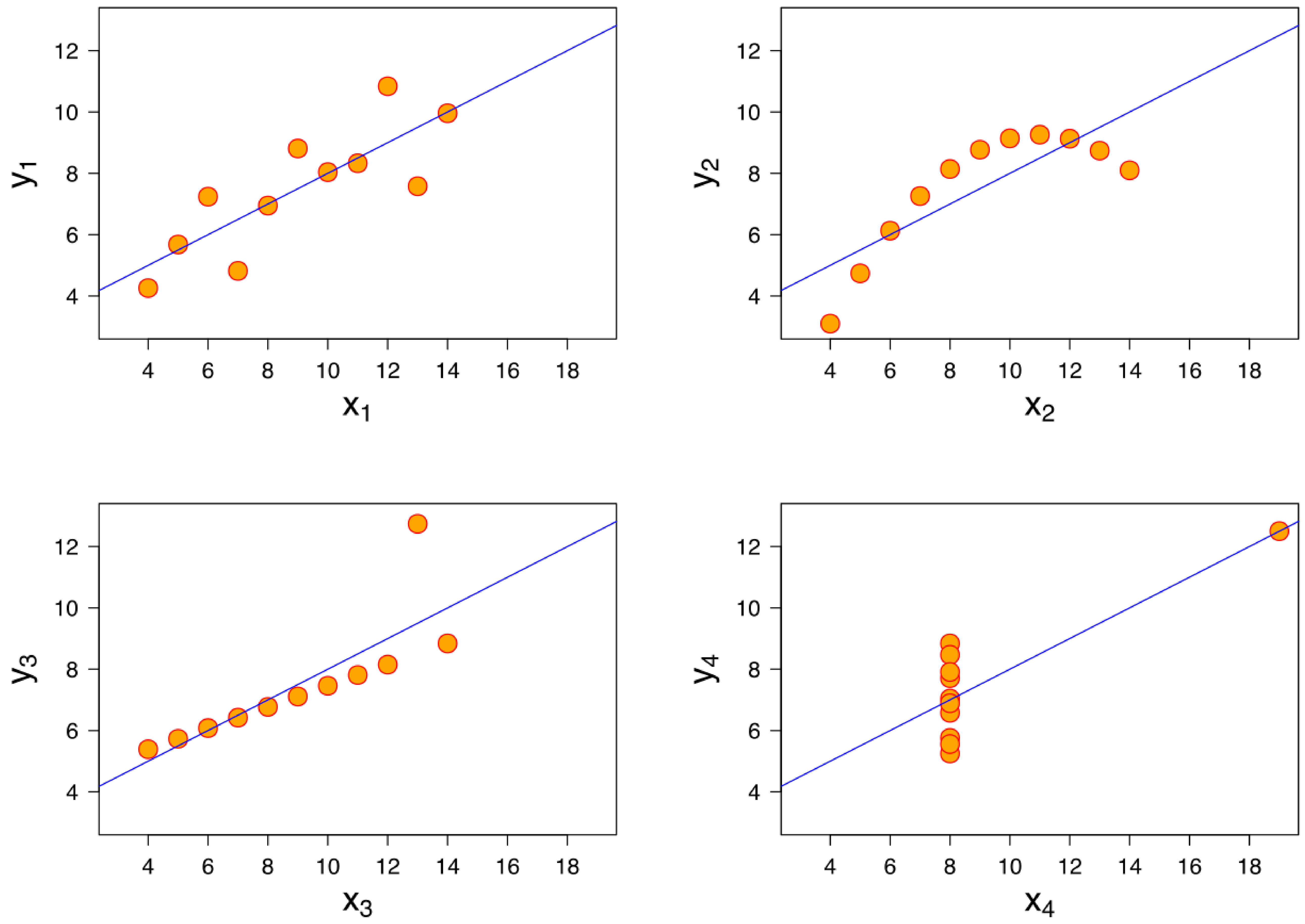
Anscombe's quartet



	Dataset I		Dataset II		Dataset III		Dataset IV	
	x	y	x	y	x	y	x	y
	10	8.04	10	9.14	10	7.46	8	6.58
	8	6.95	8	8.14	8	6.77	8	5.76
	13	7.58	13	8.74	13	12.74	8	7.71
	9	8.81	9	8.77	9	7.11	8	8.84
	11	8.33	11	9.26	11	7.81	8	8.47
	14	9.96	14	8.1	14	8.84	8	7.04
	6	7.24	6	6.13	6	6.08	8	5.25
	4	4.26	4	3.1	4	5.39	19	12.5
	12	10.84	12	9.13	12	8.15	8	5.56
	7	4.82	7	7.26	7	6.42	8	7.91
	5	5.68	5	4.74	5	5.73	8	6.89
Sum:	99.00	82.51	99.00	82.51	99.00	82.51	99.00	82.51
Avg:	9.00	7.50	9.00	7.50	9.00	7.50	9.00	7.50
Std:	3.32	2.03	3.32	2.03	3.32	2.03	3.32	2.03

WHY PLOTTING YOUR DATA?

Anscombe's quartet



HIGHER • SCHOOL • OF • ECONOMICS •

NATIONAL RESEARCH
UNIVERSITY