Warsztaty modelowania

03 – wykresy i statystyki

opracowała

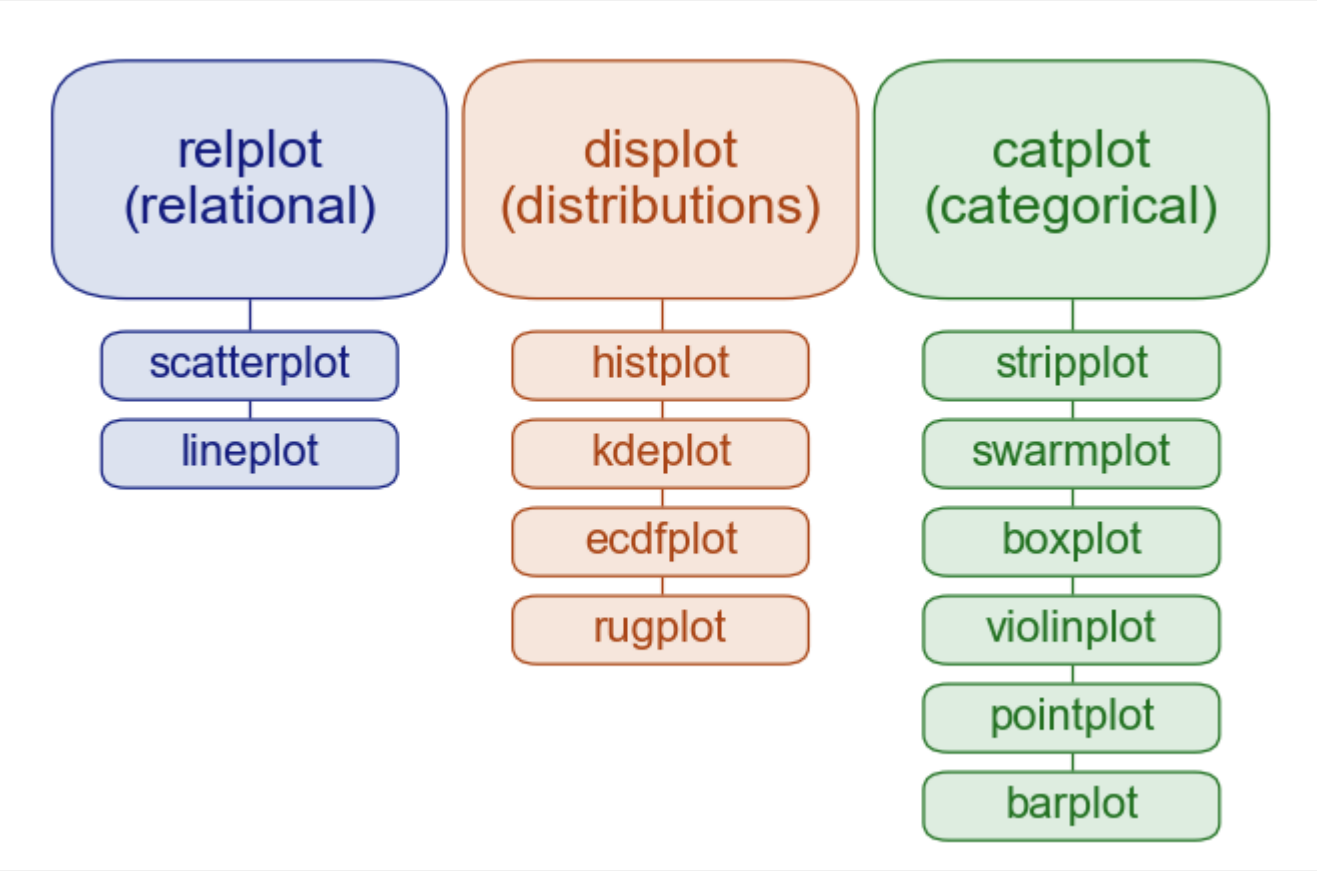
Patrycja Naumczyk

O czym będzie?

1. Seaborn – Matplotlib inaczej
    a) Struktura wykresów
    b) Format danych
2. Identyfikacja outlierów
3. Statystyki w Pythonie
    a) Biblioteki godne uwagi
    b) Kiedy który test?

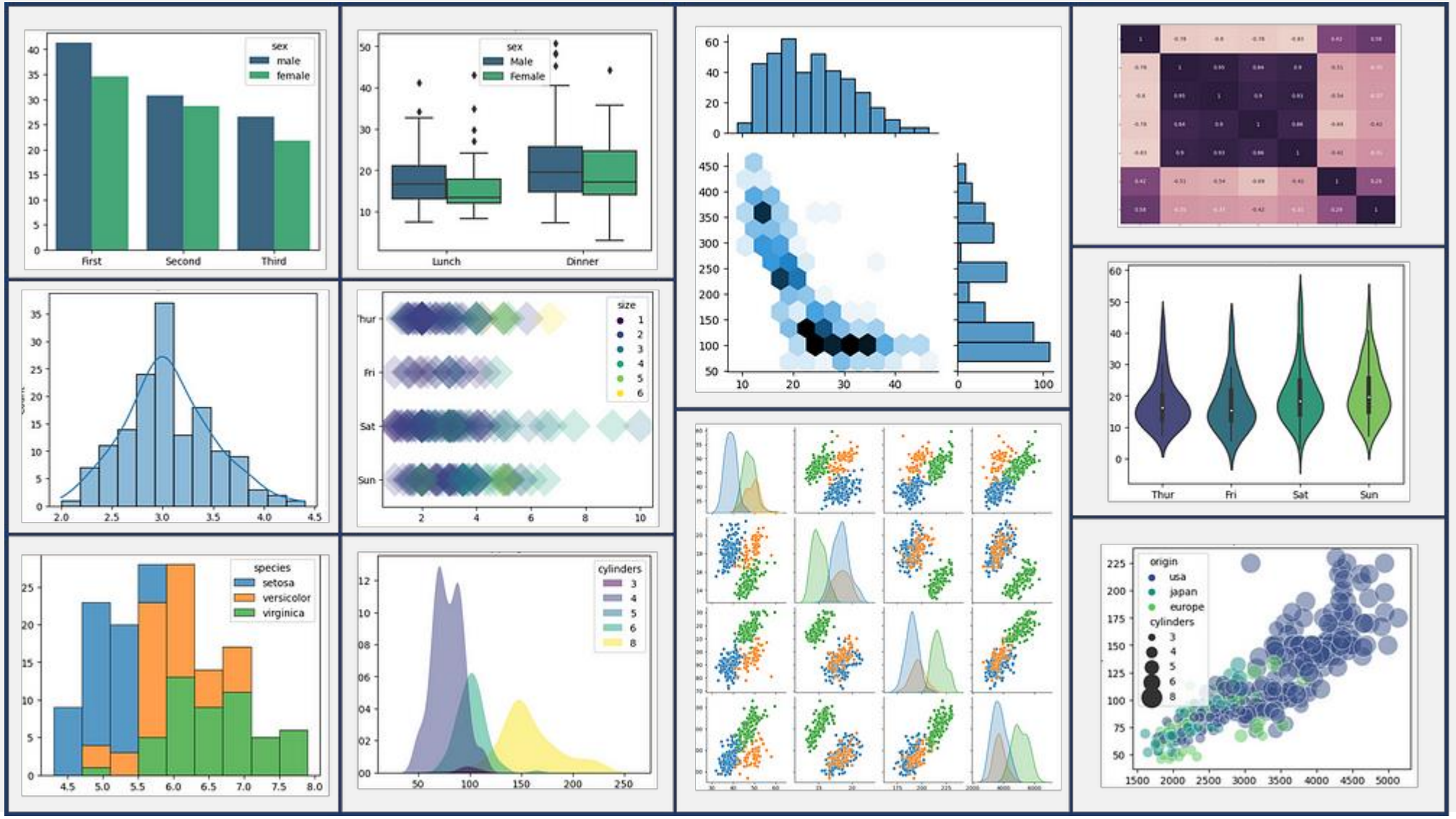
Seaborn – przyjazny Matplotlib

1. Struktura wykresów: figure-level vs. axes-level

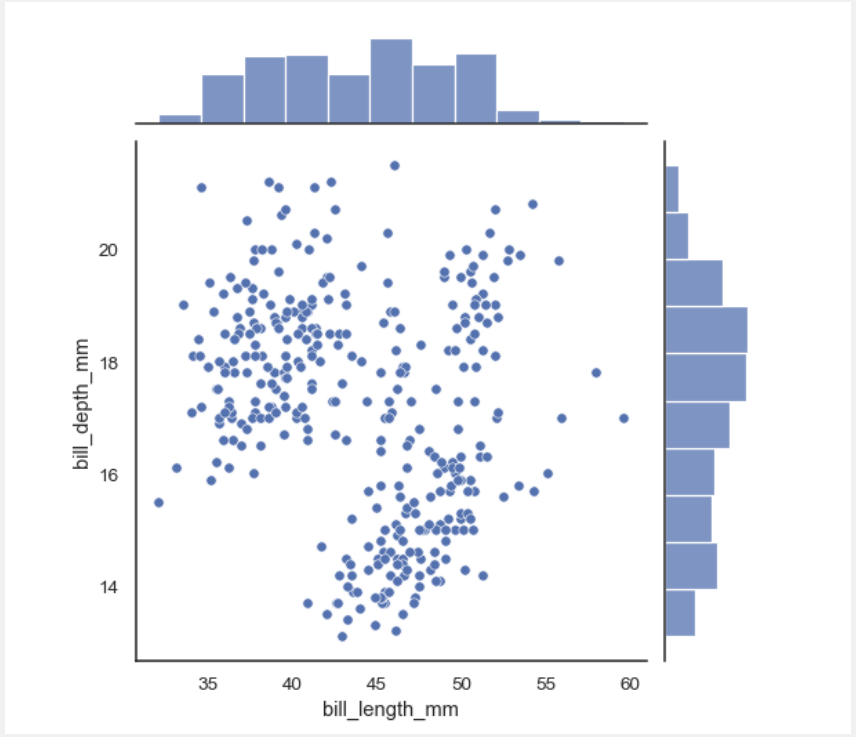


https://seaborn.pydata.org/tutorial.html

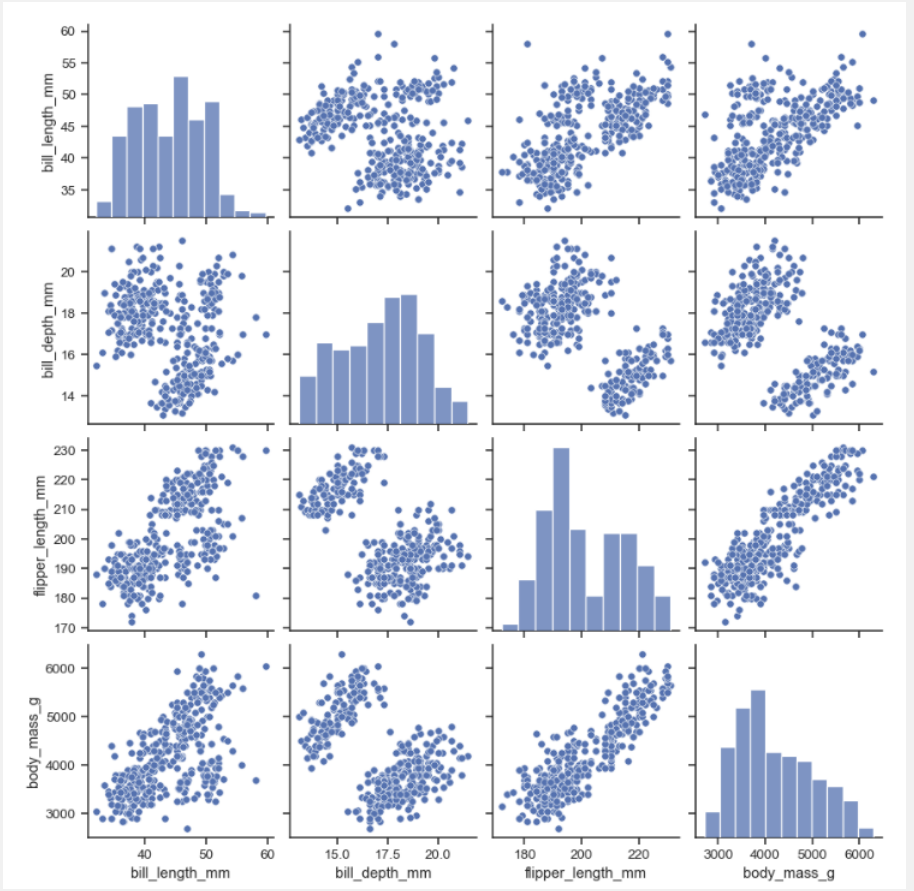
Seaborn – rodzaje wykresów cd.



Specyficznie Seaborn’owe wykresy



sns.jointplot()



sns.pairplot()

Seaborn – przyjazny Matplotlib

1. Struktura danych : wide i long format

Wide Format

Team	Points	Assists	Rebounds
A	88	12	22
B	91	17	28
C	99	24	30
D	94	28	31

Long Format

Team	Variable	Value
A	Points	88
A	Assists	12
A	Rebounds	22
B	Points	91
B	Assists	17
B	Rebounds	28
C	Points	99
C	Assists	24
C	Rebounds	30
D	Points	94
D	Assists	28
D	Rebounds	31

Melt

`df3`

	first	last	height	weight
0	John	Doe	5.5	130
1	Mary	Bo	6.0	150

`df3.melt(id_vars=['first', 'last'])`

	first	last	variable	value
0	John	Doe	height	5.5
1	Mary	Bo	height	6.0
2	John	Doe	weight	130
3	Mary	Bo	weight	150

Pandas melt

Wide DataFrame

	Person	House	Age	Books	Movies
0	Alan	A	32	100	10
1	Berta	B	46	30	20
2	Charlie	A	35	20	80
3	Danielle	C	28	40	60

	Person	House	Age	Books	Movies
	Name	Flat	Old	Text	Video
0	Alan	A	32	100	10
1	Berta	B	46	30	20
2	Charlie	A	35	20	80
3	Danielle	C	28	40	60

Long DataFrame

	Person	Info	Numerical
0	Alan	Books	100
1	Berta	Books	30
2	Charlie	Books	20
3	Danielle	Books	40
4	Alan	Movies	10
5	Berta	Movies	20
6	Charlie	Movies	80
7	Danielle	Movies	60

Long DataFrame

	Person	House	variable	value
0	Alan	A	Age	32
1	Berta	B	Age	46
2	Charlie	A	Age	35
3	Danielle	C	Age	28
4	Alan	A	Books	100
5	Berta	B	Books	30
6	Charlie	A	Books	20
7	Danielle	C	Books	40
8	Alan	A	Movies	10
9	Berta	B	Movies	20
10	Charlie	A	Movies	80
11	Danielle	C	Movies	60

Long DataFrame

	Person	House	variable	value
0	Alan	A	Age	32
1	Berta	B	Age	46
2	Charlie	A	Age	35
3	Danielle	C	Age	28
0	Alan	A	Books	100
1	Berta	B	Books	30
2	Charlie	A	Books	20
3	Danielle	C	Books	40
0	Alan	A	Movies	10
1	Berta	B	Movies	20
2	Charlie	A	Movies	80
3	Danielle	C	Movies	60

Long DataFrame

	variable	value
0	Person	Alan
1	Person	Berta
2	Person	Charlie
3	Person	Danielle
4	House	A
5	House	B
6	House	A
7	House	C
8	Age	32
9	Age	46
10	Age	35
11	Age	28
12	Books	100
13	Books	30
14	Books	20
15	Books	40
16	Movies	10
17	Movies	20
18	Movies	80
19	Movies	60

Pandas.melt( ) Function Examples

```
data_wide.melt()
```

```
data_wide.melt(id_vars=["Person","House"])
```

```
data_wide.melt(
id_vars=["Person","House"],
value_vars=["Age","Books","Movies"])
```

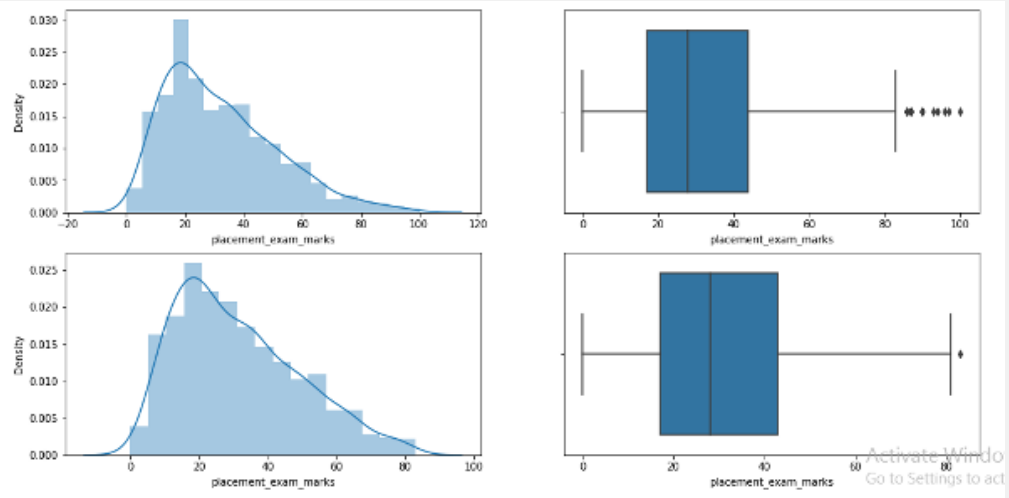
```
data_wide.melt(
id_vars=["Person"],
value_vars=["Books","Movies"],
var_name="Info",
value_name="Numerical")
```

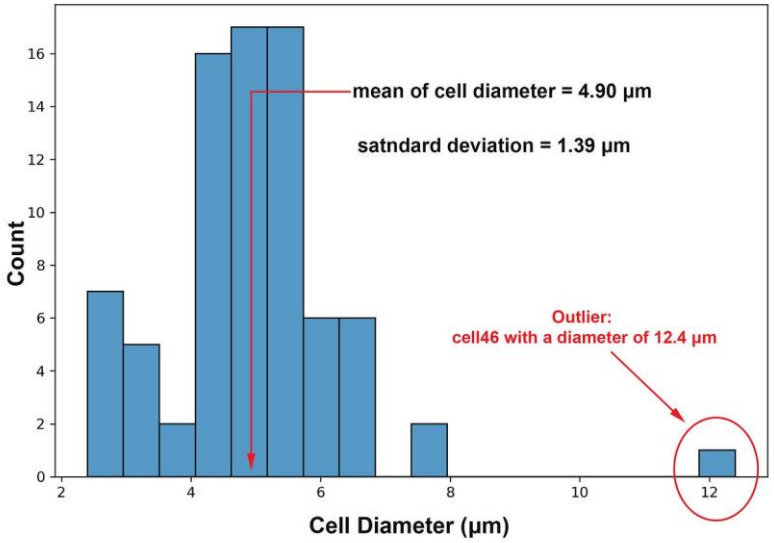
```
data_wide.melt(
id_vars=["Name","Flat"],
value_vars=["Old","Text","Video"],
var_name="Info",
value_name="Numerical",
col_level=1)
```

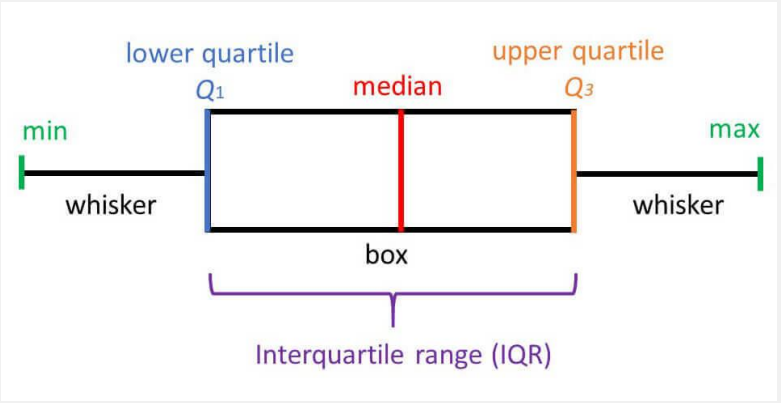
```
data_wide.melt(
id_vars=["Person","House"],
value_vars=["Age","Books","Movies"],
var_name="Info",
value_name="Numerical",
ignore_index=False)
```

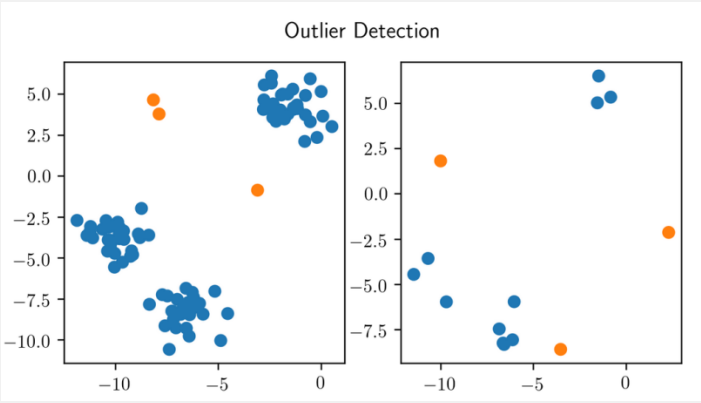
Identyfikacja outlierów

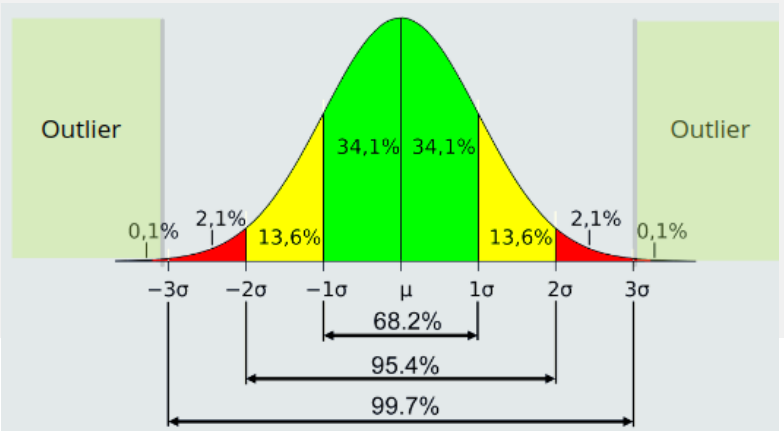
1. Wizualizacja
   a) Histogram
   b) Wykres rozrzutu
   c) Wykresy pudełkowe
2. Rozstęp ćwiartkowy (metoda .quantile())
3. Wartość Z
4. Modelowanie





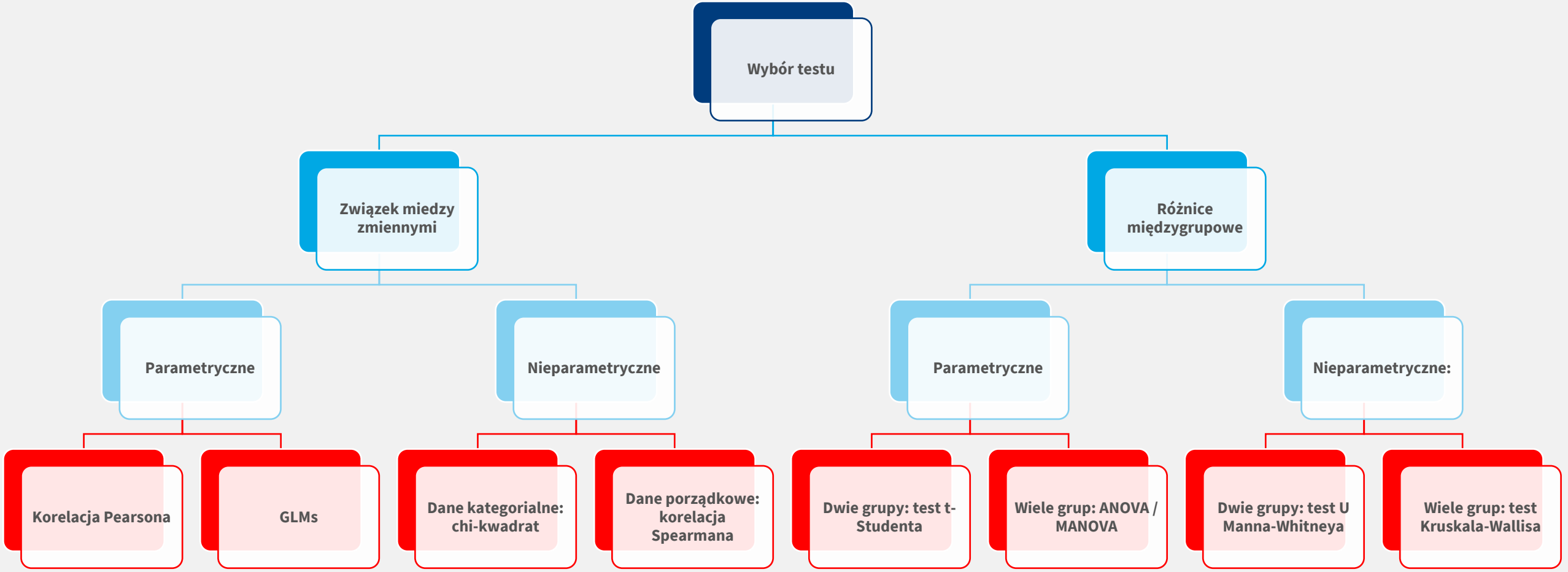






Podstawowa analiza statystyczna w Pythonie - biblioteki

Kiedy który test?



Podstawowa analiza statystyczna – by me

Test	Biblioteka	Miara effect size
GLMs	Statsmodels, jak się nie da to scikit-learn	Zależy od modelu
Korelacja Pearsona / Spearmana	pandas (metoda .corr()), scipy jeśli potrzebuję wartości *p*	r^2
Chi-kwadrat	researchpy	ϕ-Cramera
t-Studenta	researchpy	d-Cohena
U Manna-Whitneya	scipy	Δ-Cliffa
ANOVA	statsmodels (model OLS)	η^2
Kruskala-Wallisa	scipy	η^2

https://www.pythonfordatascience.org/home
https://en.wikipedia.org/wiki/Effect_size

Tomczak, Maciej & Tomczak-Łukaszewska, Ewa. (2014). The need to report effect size estimates revisited. An overview of some recommended measures of effect size. 21. 19-25.