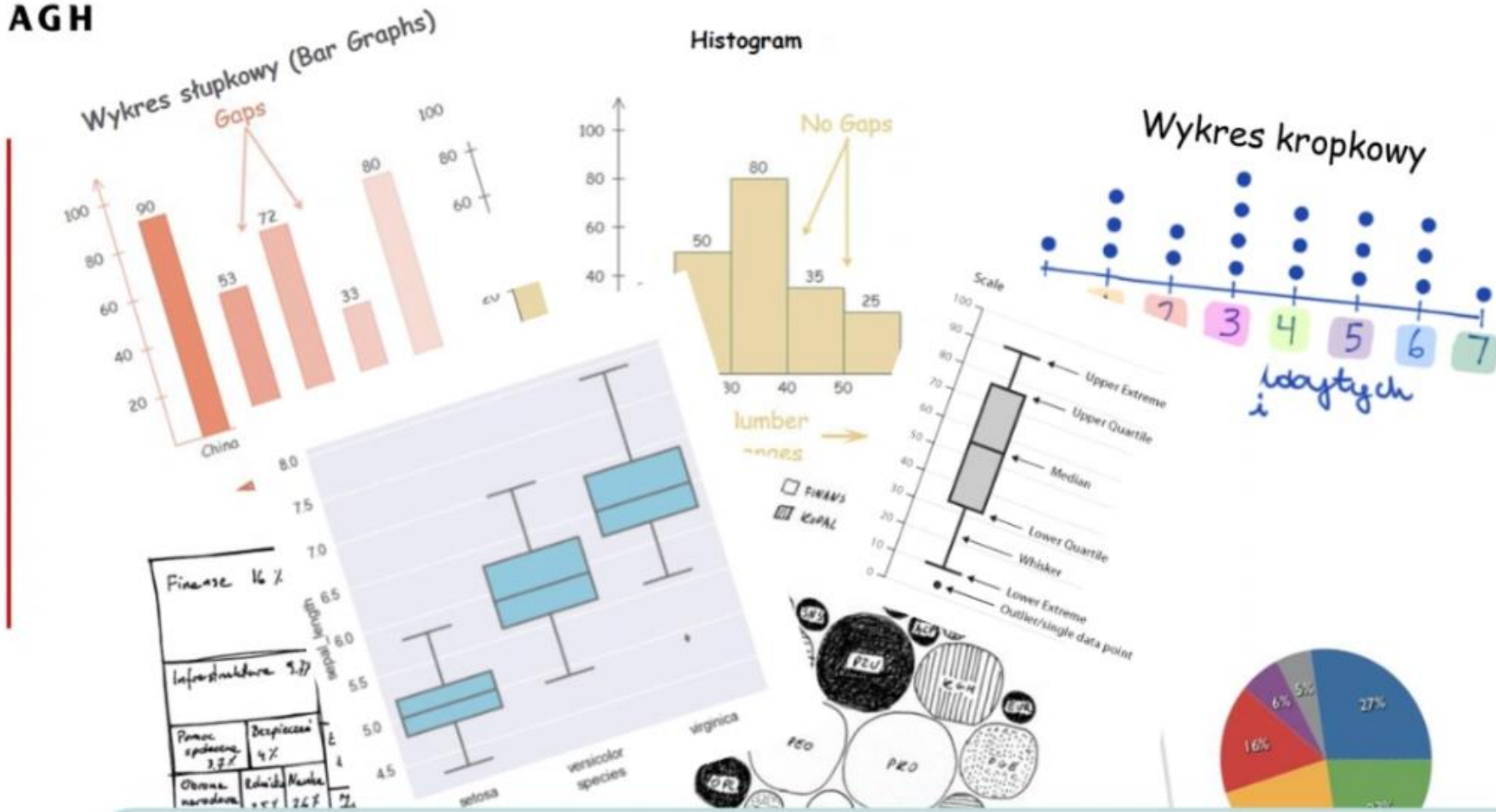
Statystyczna analiza danych W2

6.03.2023r.

Graficzna prezentacja danych



„… otwieracie sobie jakiś plik z danymi, wybieracie dostępne opcje wizualizacji i widzicie: wykres słupkowy, kołowy, liniowy, pudełkowy, histogram…. I wiem co sobie teraz myślicie. Myślicie sobie: o mój Boże, tu jest jak w monopolowym, tu wszystko jest pyszne!"
(*Janina Bąk „**Statystycznie rzecz biorąc**"*)

Graficzna prezentacja danych

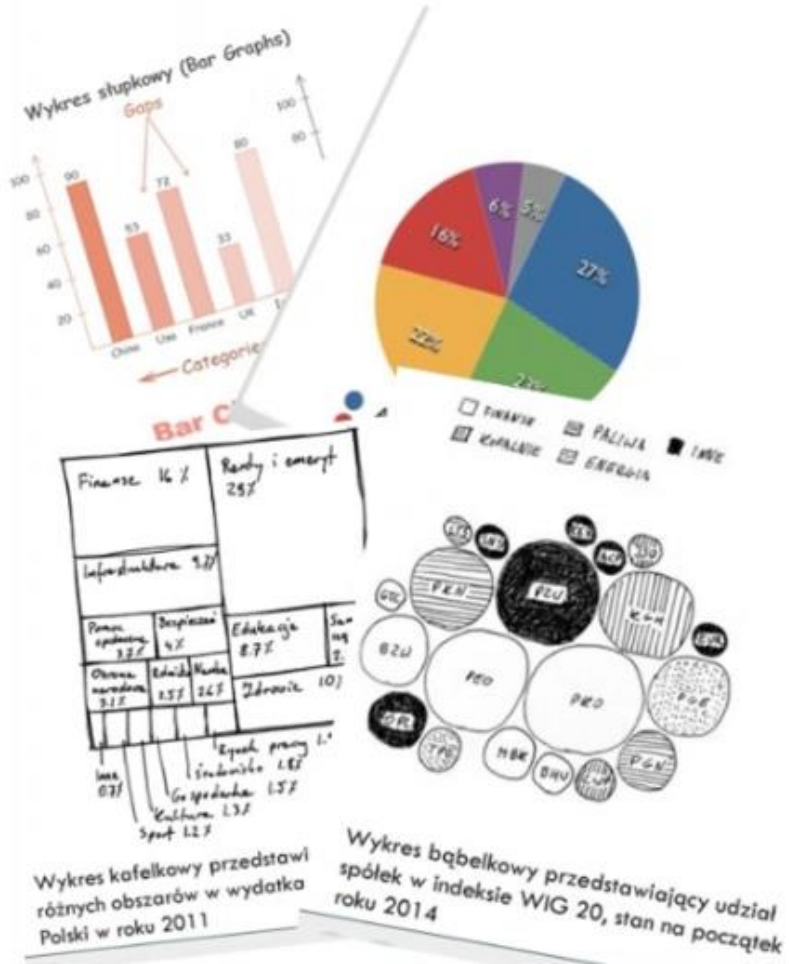
Pojawia się pytanie...

Jak wybrać odpowiedni wykres?

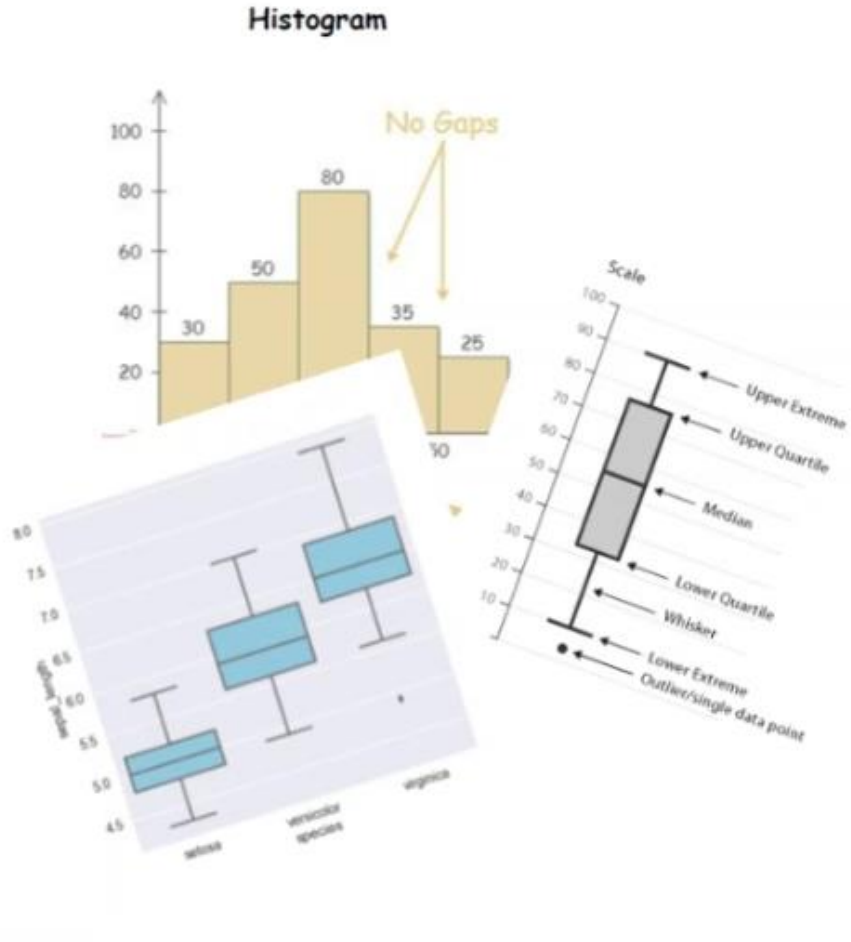
Musimy zastanowić się, jakim rodzajem **skali pomiarowej** jest przedstawiona zmienna na owym wykresie

Graficzna prezentacja danych

Dane jakościowe



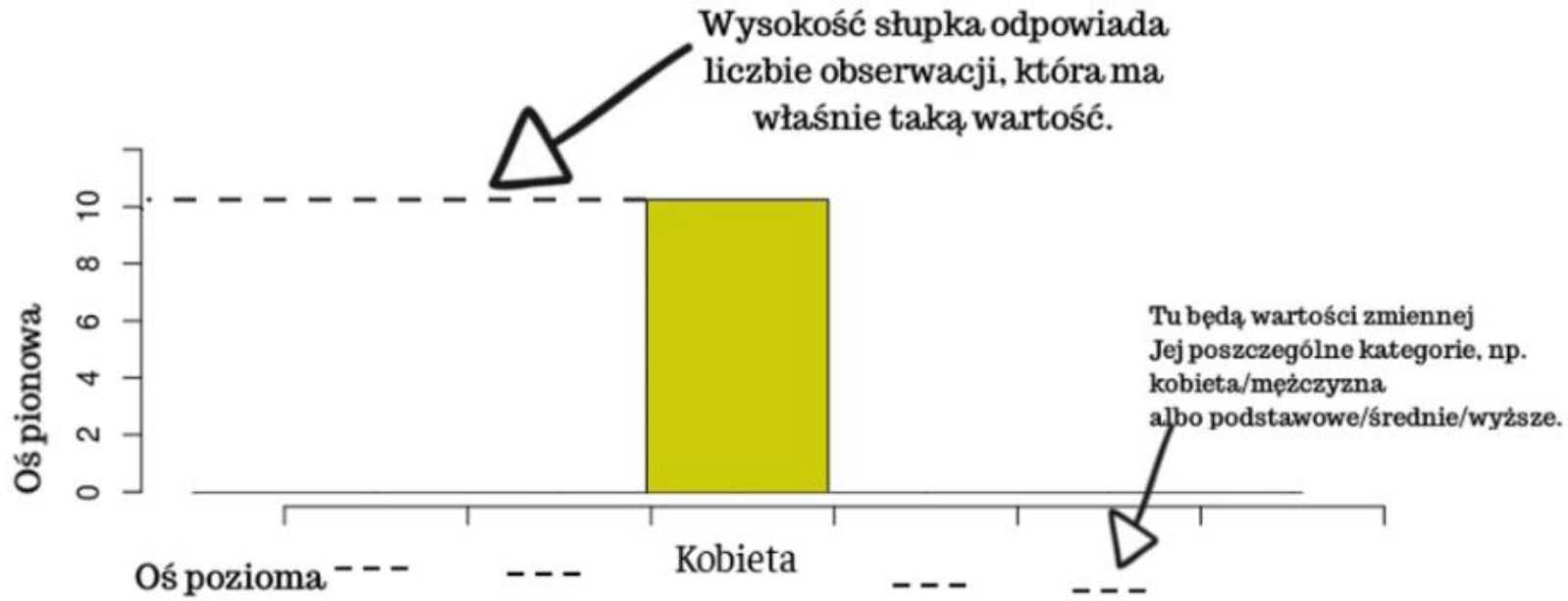
Dane ilościowe



Graficzna prezentacja danych

Wykres słupkowy

Wartości są wskazywane przez **długość słupków**, z których każdy odpowiada mierzonej grupie.
Oś Y pokazuje **częstość** czyli **jak wiele elementów** zostało zaliczonych do określonej kategorii/przedziału danych.



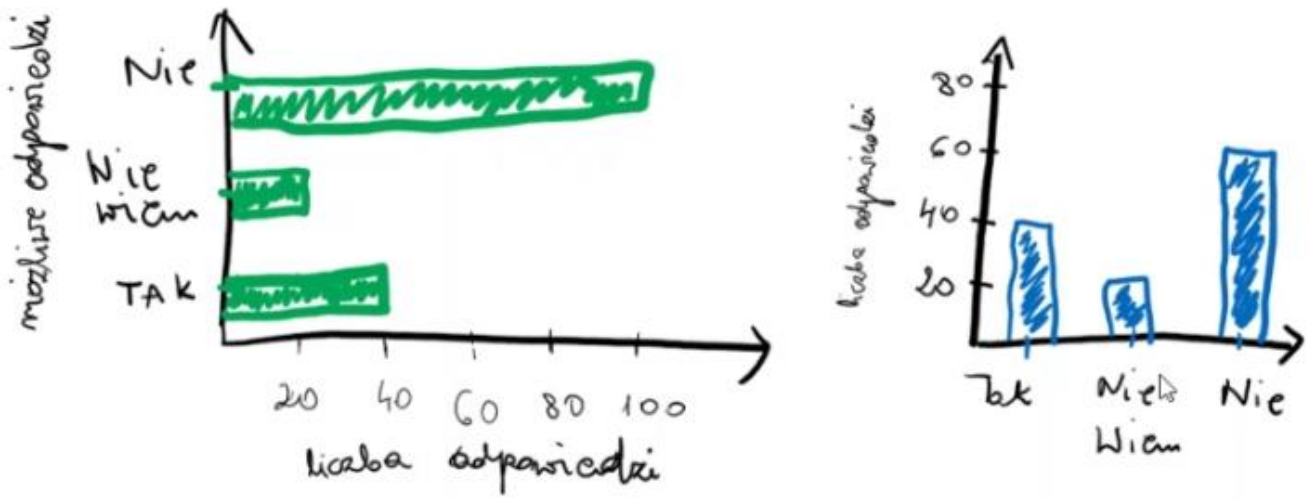
Źródło: Statystyka w psychologii

Graficzna prezentacja danych
Wykres słupkowy

Słupki mogą być zorientowane:

» pionowo (kolumny)

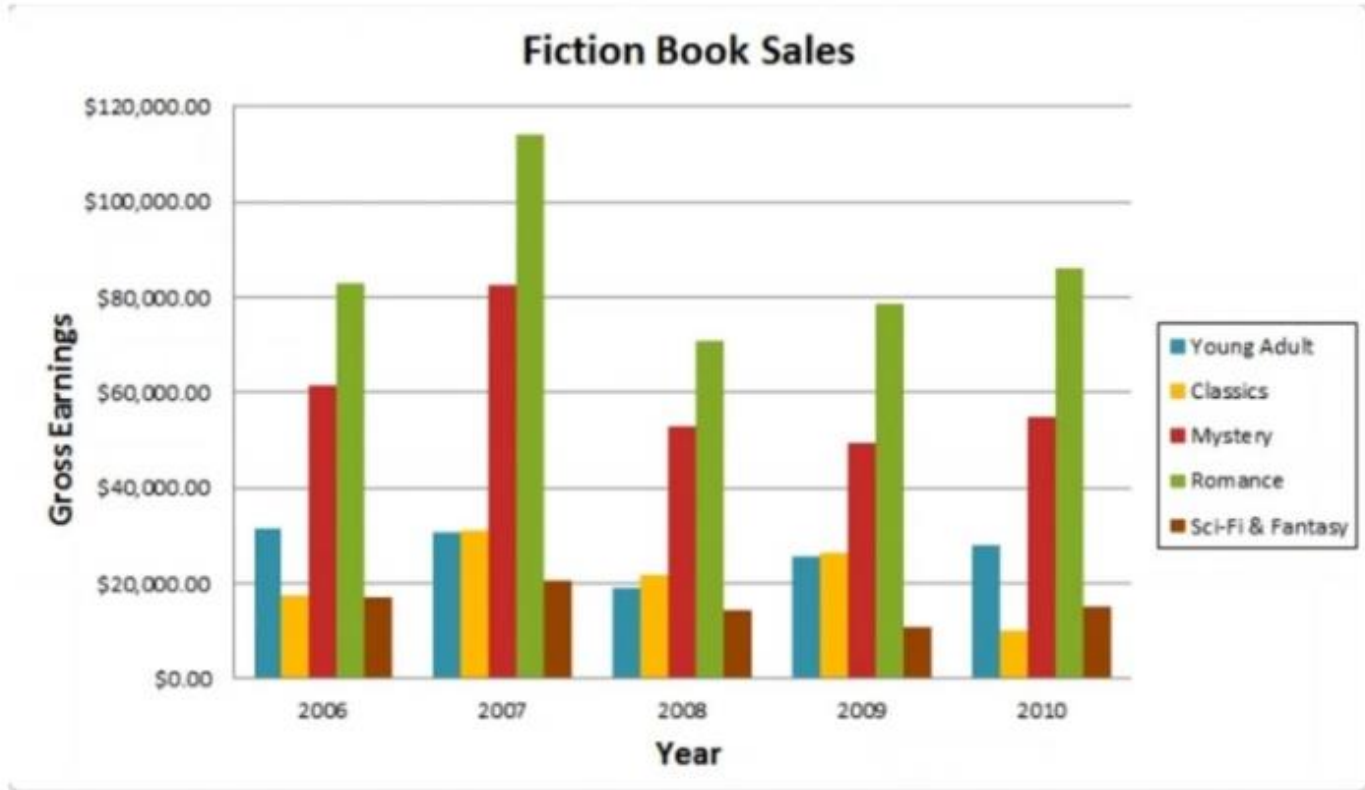
» lub poziomo (dużo kategorii, długie nazwy etykiet)



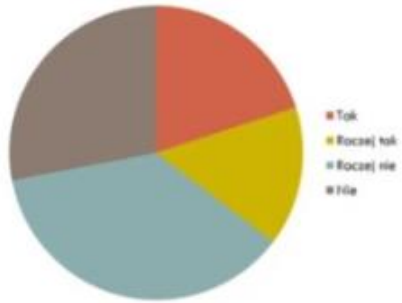
Graficzna prezentacja danych
Wykres słupkowy

Wykres słupkowy składa się z serii słupków ilustrujących rozwój zmiennej.

Graficzna prezentacja danych
Wykres słupkowy

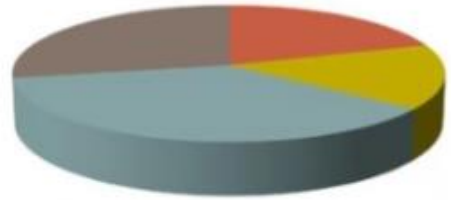


Graficzna prezentacja danych
Wykres kołowy

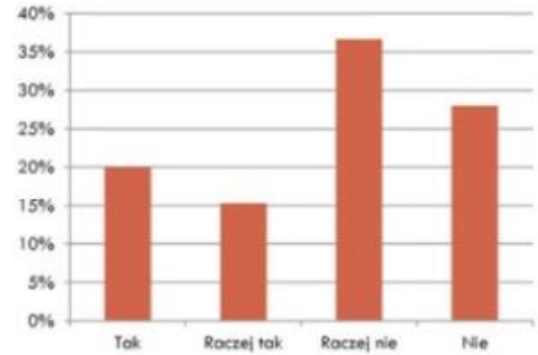


Dodajemy etykiety z wartościami liczbowymi lub procentowymi!

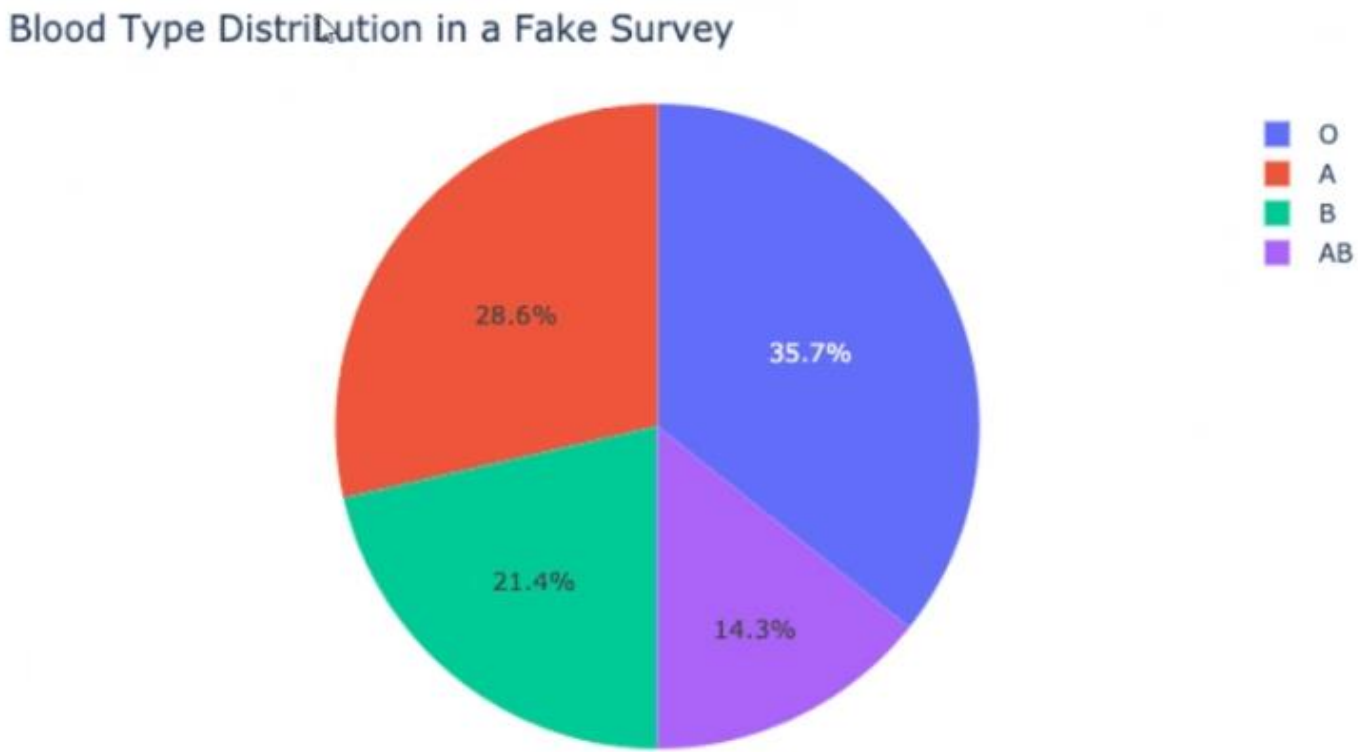
Wykres kołowy - sposób prezentacji danych kategorialnych/jakościowych wyrażonych w ujęciu względnym - procentowym (udział w całości).



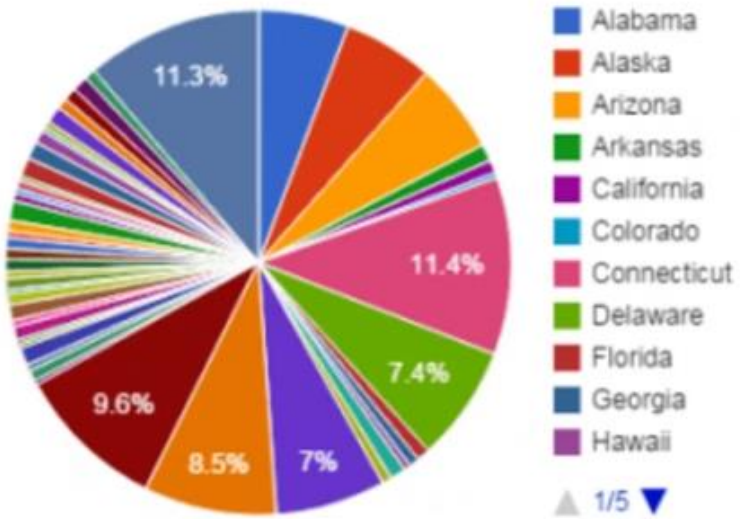
Wykres słupkowy - pozwala na **odczytanie liczebności** poszczególnych klas i **porównanie** ich między sobą .

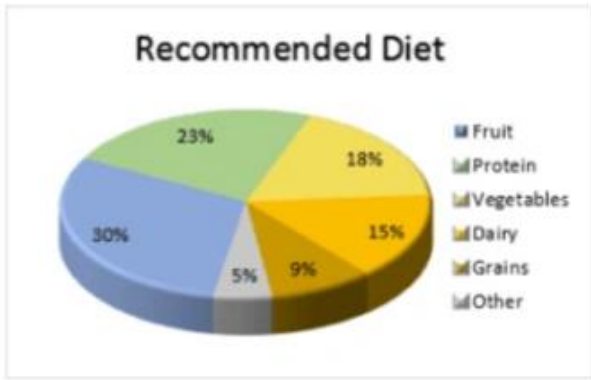


Graficzna prezentacja danych
Wykres kołowy



Pie Chart





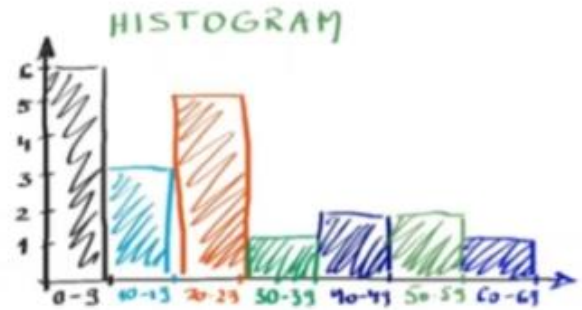
Zwykle **nie** jest właściwe używanie:

- Wykresów kołowych **dla więcej niż 5** lub 6 różnych kategorii. Wiele grup jest **trudnych do wizualizacji**.
- Wykresów **3D**

„Wykres kołowy to taki Fiat Multipla świata statystyki – niby ładnie wygląda, ale nikt go nie szanuje.
Janina Bąk „Statystycznie rzecz biorąc"

Graficzna prezentacja danych HISTOGRAM

„Histogram to jest graficzny sposób przedstawienia rozkładu empirycznego cechy."



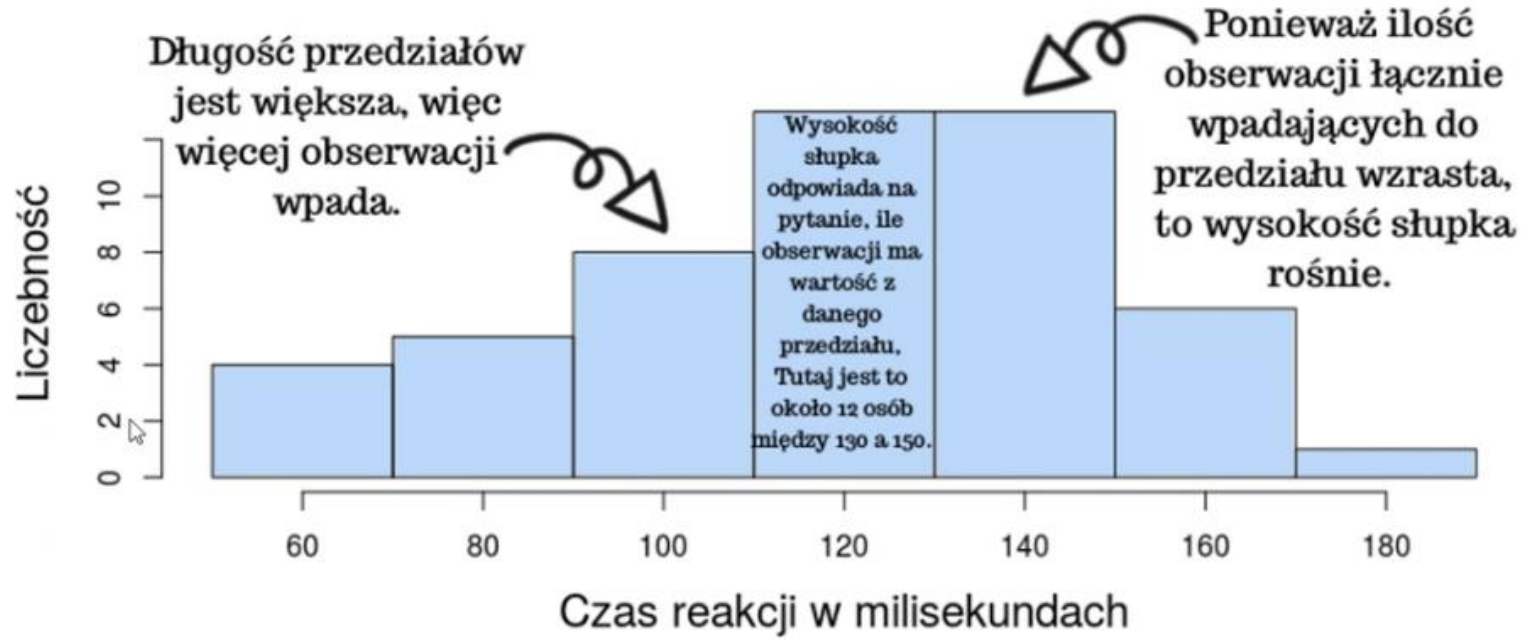
Rozkład empiryczny oznacza rozkład otrzymany na podstawie danych. **Empiryczny** znaczy **doświadczalny**.

Histogram służy do **analizy zbiorowość**, co oznacza, że na podstawie histogramu jesteśmy w stanie określić:

- ✓ który **przedział** jest **najliczniejszy**,
- ✓ czy rozkład jest **symetryczny** czy może **asymetryczny** / skośny,
- ✓ czy występują przedziały mocno **odstające** od reszty.

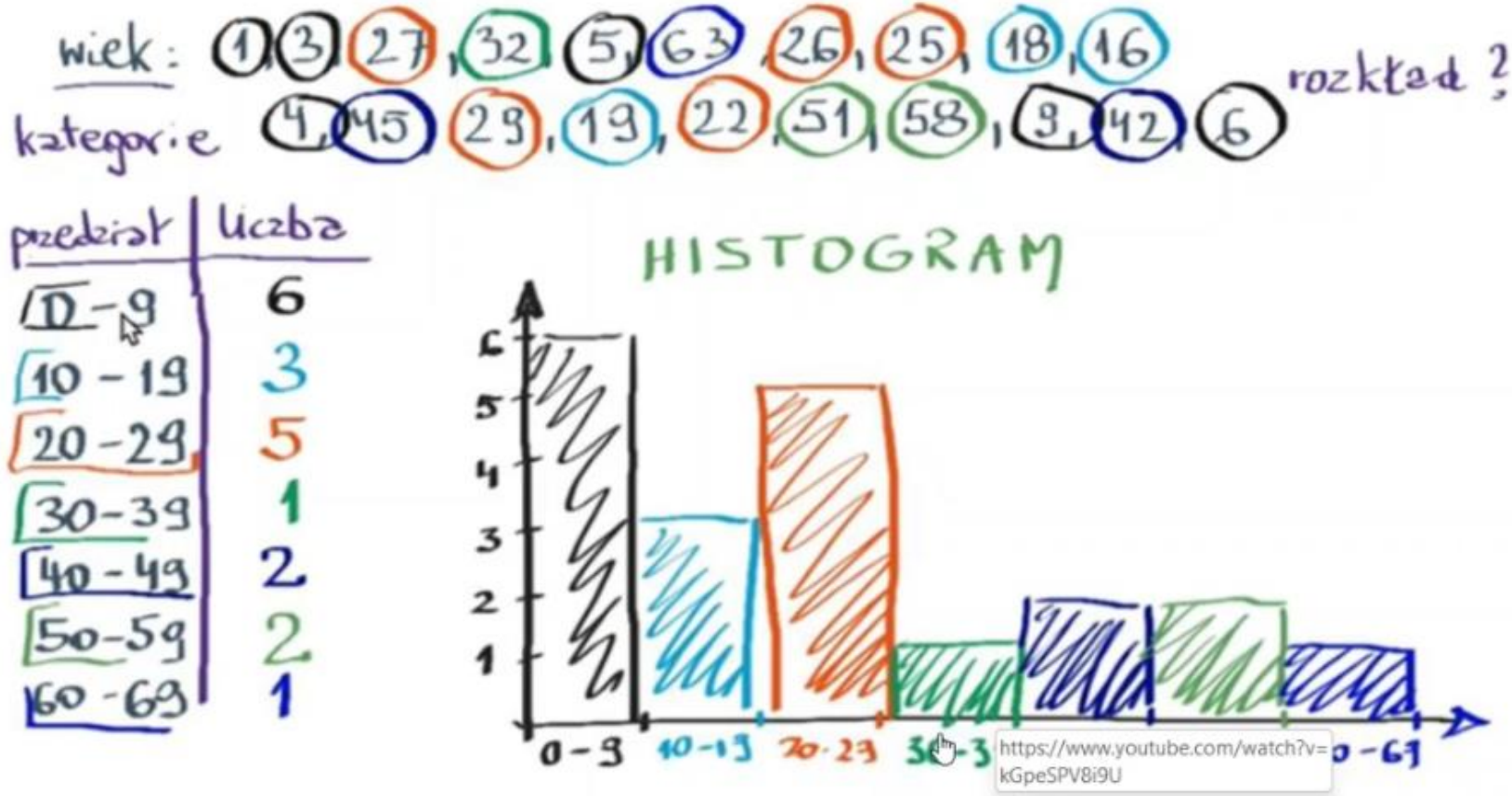
Graficzna prezentacja danych
HISTOGRAM

Wartości są wskazywane przez **długość słupków**,
z których każdy odpowiada mierzonej grupie.
Oś Y pokazuje **częstość** czyli **jak wiele elementów** zostało zaliczonych do określonej kategorii/przedziału danych.

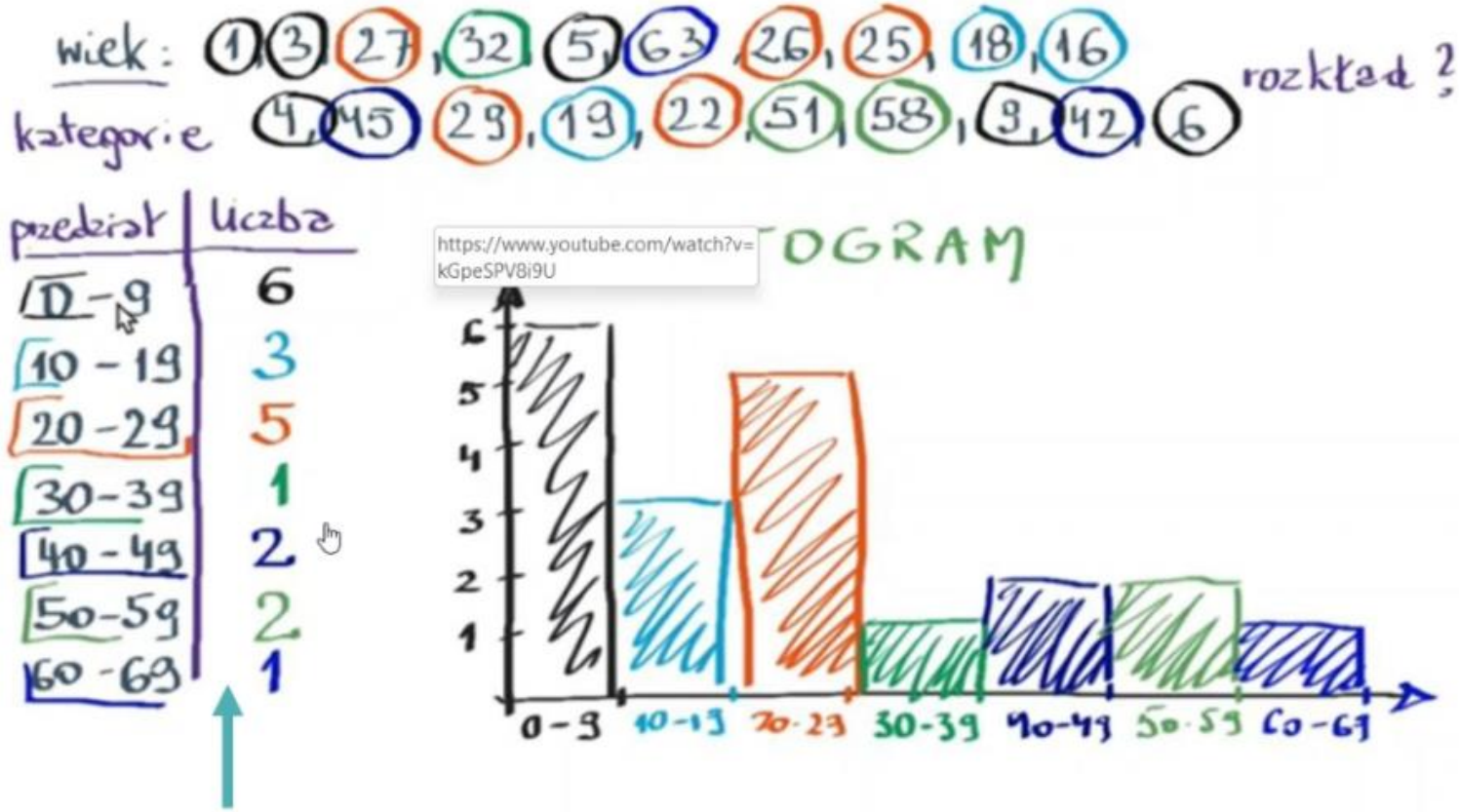


Źródło: Statystyka w psychologii

Graficzna prezentacja danych
HISTOGRAM



Graficzna prezentacja danych
HISTOGRAM



Szereg przedziałowy składa się z przedziałów oraz liczby obserwacji, które się w nich znajdują.

Histogram to graficznie przedstawiony szereg przedziałowy.

Graficzna prezentacja danych
HISTOGRAM – szereg rozdzielczy

Przy budowie szeregu rozdzielczego wyróżnia się następujące etapy:

1. Określenie liczby przedziałów histogramu (szeregu rozdzielczego)

Stosowane bywają następujące wzory pomocne do szacowania liczby przedziałów budowanego szeregu:

$$k \approx 1 + 3{,}322 \log n$$

lub

$$k \approx \sqrt{n}$$

2. Określenie wielkości przedziałów

$$\Delta x \approx \frac{x_{max} - x_{min}}{k} \approx \frac{R}{k}$$

3. Wyznacz przedziałów i przyporządkowanie danych do określonych przedziałów

Przydziały muszą być jednostronnie domknięte.

Graficzna prezentacja danych
HISTOGRAM

$n = 20$

wiek: 1, 3, 27, 32, 5, 63, 26, 25, 18, 16

kategorie 4, 45, 29, 19, 22, 51, 58, 3, 42, 6

rozkład ?

$k = 1 + 3{,}222 \log 20 = 5{,}19$

$k \approx 5$

$k = \sqrt{20} = 4{,}47$

$k \approx 4{,}5$

$$\Delta x = \frac{63 - 1}{k} \approx 12$$

Graficzna prezentacja danych
HISTOGRAM

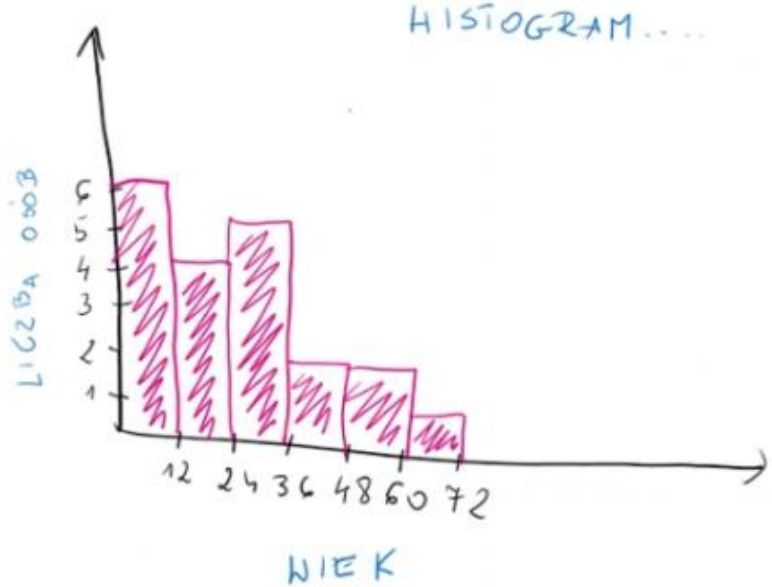
n = 20

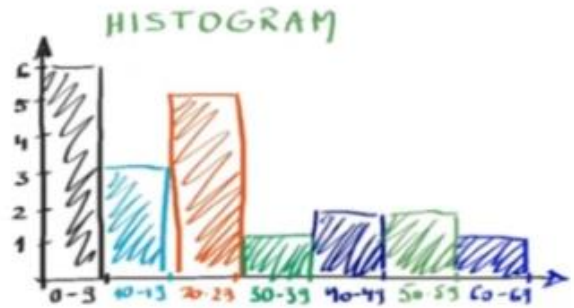
wiek: 1, 3, 27, 32, 5, 63, 26, 25, 18, 16

kategorie 4, 45, 29, 19, 22, 51, 58, 3, 42, 6

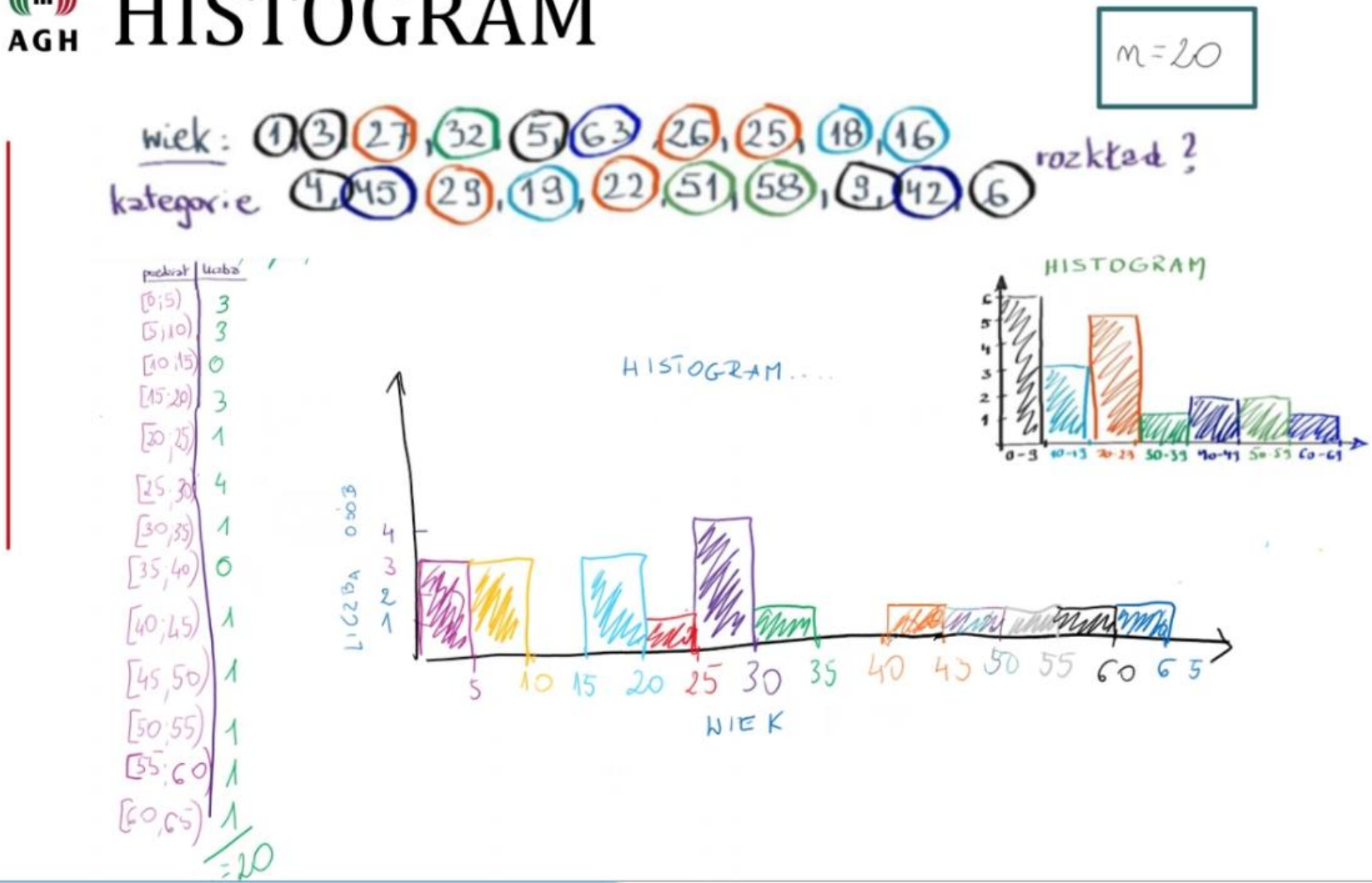
rozkład ?

podział	liczba
[0,12)	6
[12,24)	4
[24,36)	5
[36,48)	2
[48,60)	2
[60,72)	1

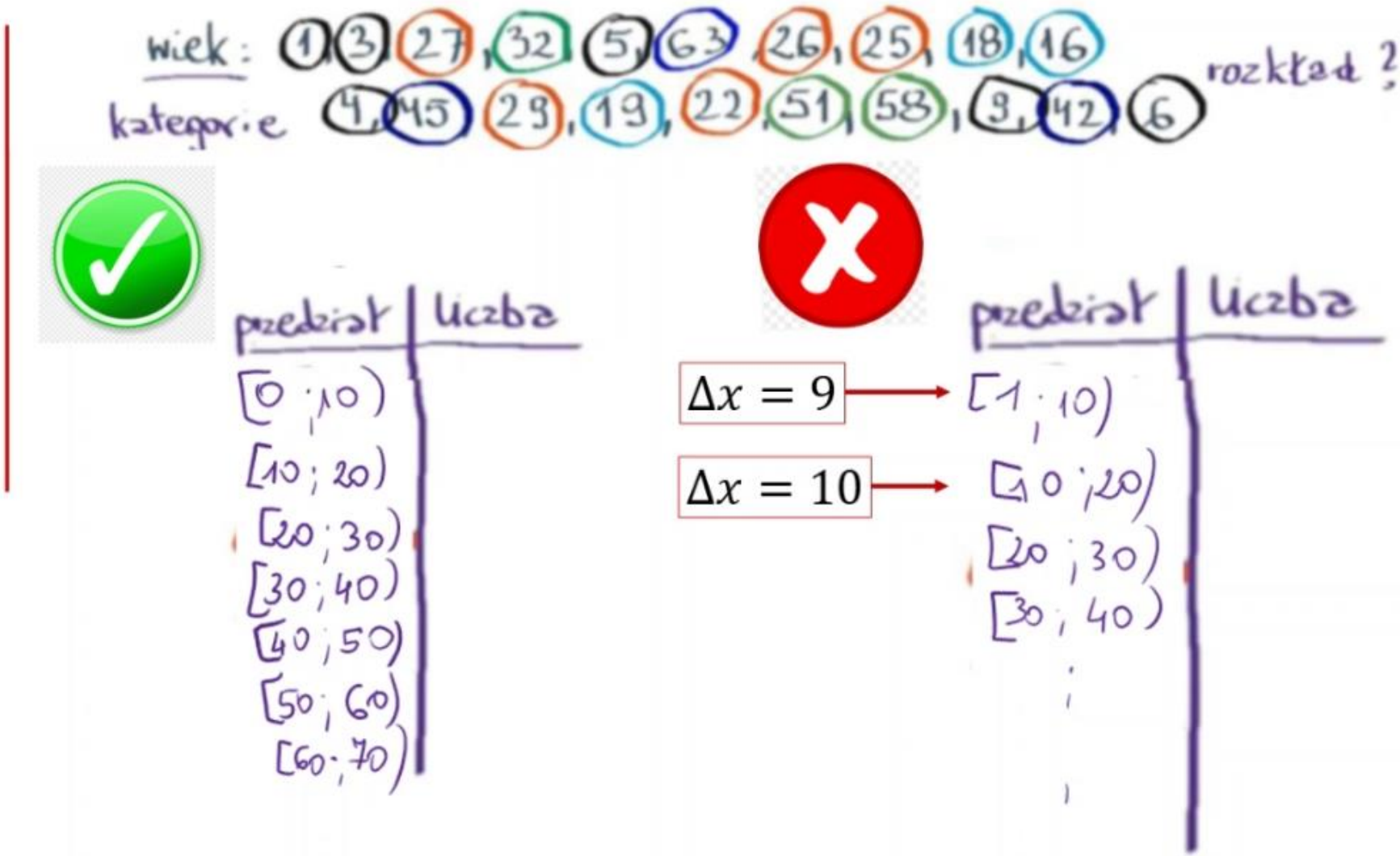




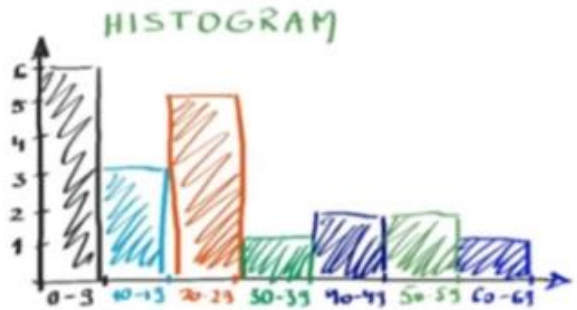
Graficzna prezentacja danych
HISTOGRAM



Graficzna prezentacja danych
HISTOGRAM



Graficzna prezentacja danych HISTOGRAM

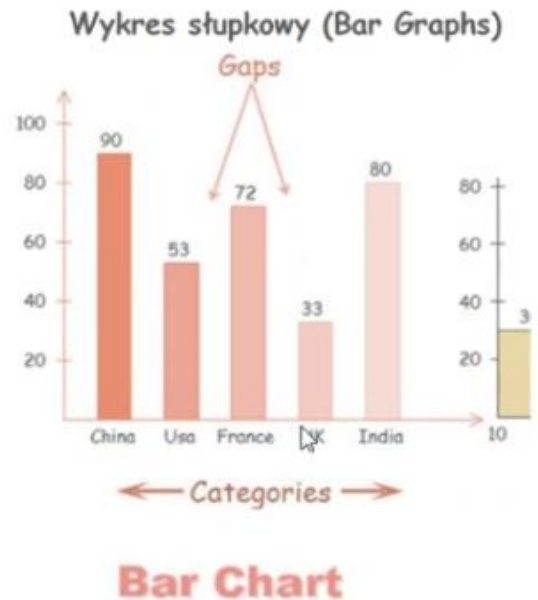


!!! NIE ISTNIEJĄ SZTYWNE ZASADY TWORZENIA HISTOGRAMÓW

1. Każdy **zbiór** danych wymaga **innego zakresu grupowań** i innych **szerokości** przedziałów
   a) **bardzo szerokie** – sensowne porównania są niemożliwe
   b) **bardzo wąskie** – uwydatnienie najmniejszych różnic
2. **Szerokość** poszczególnych przedziałów musi być **jednakowa**
3. **Konsekwentne** umieszczanie wartości granicznych (**domykanie**, albo na końcu, albo na początku przedziału)
4. **Precyzyjne** określanie **nazw osi**

Graficzna pre danych
Wykres słupkowy/Histogram

dane jakościowe



Bar Chart

dane ilościowe

Na obu typach wykresów **wartości** są wskazywane przez **długość słupków**, z których każdy odpowiada mierzonej grupie.

Oś Y pokazuje **częstość** czyli **jak wiele elementów** zostało zaliczonych do określonej kategorii/przedziału danych.

Graficzna prezentacja danych

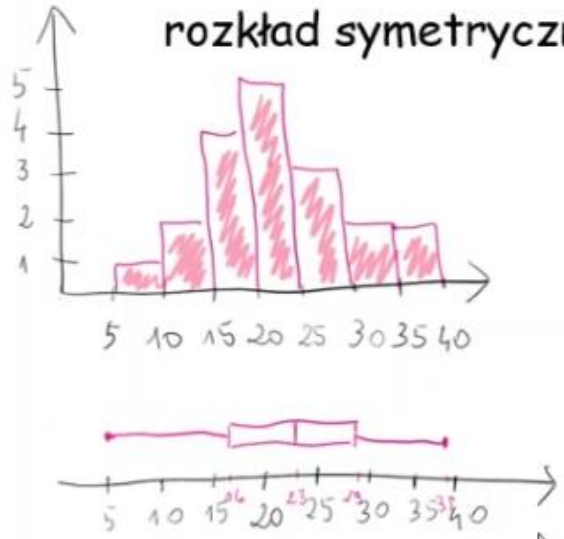
INTERPRETACJA HISTOGRAM

I KSZTAŁT DANYCH (ROZKŁAD DANYCH) – sprawdzenie w jaki sposób dane się grupują

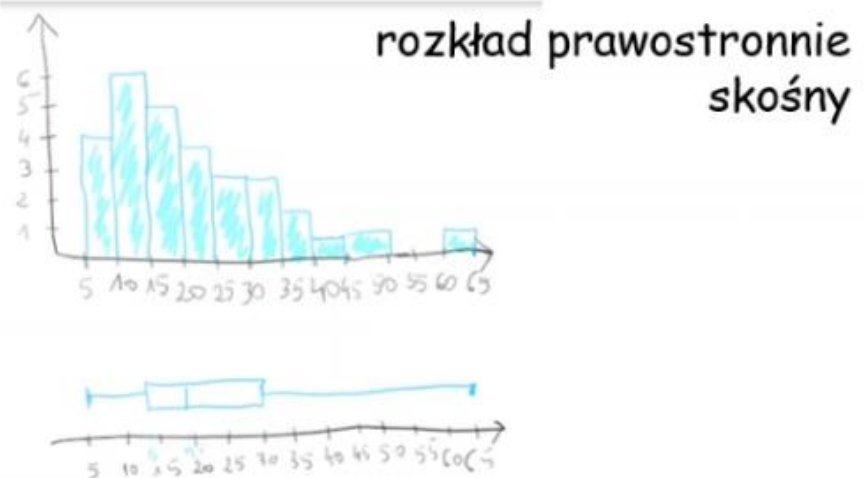
1. Szereg zbiorów danych ma **charakterystyczny** kształt:
   - **symetryczny** (normalny/płaski) (średnia i mediana znajdują się blisko siebie)
   - **Prawostronnie skośny** (średnia jest większa od mediany)
   - **lewostronnie skośny** (średnia jest mniejsza od mediany)
2. !!! Nie należy oczekiwać, że dane symetryczne będą miały identyczny kształt.
3. !!! Nie należy zakładać, że dane są skośne, jeżeli ich kształt nie jest symetryczny

Graficzna prezentacja danych

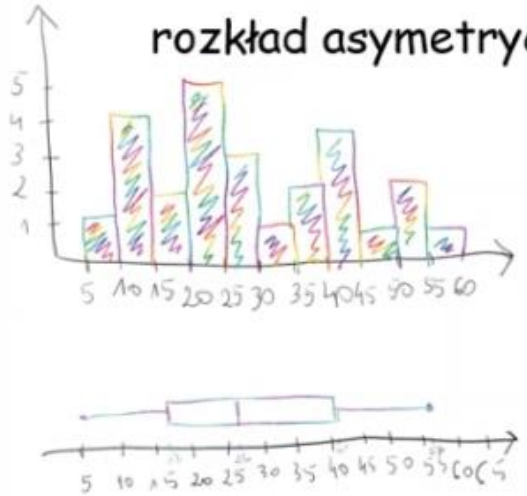
rozkład symetryczny



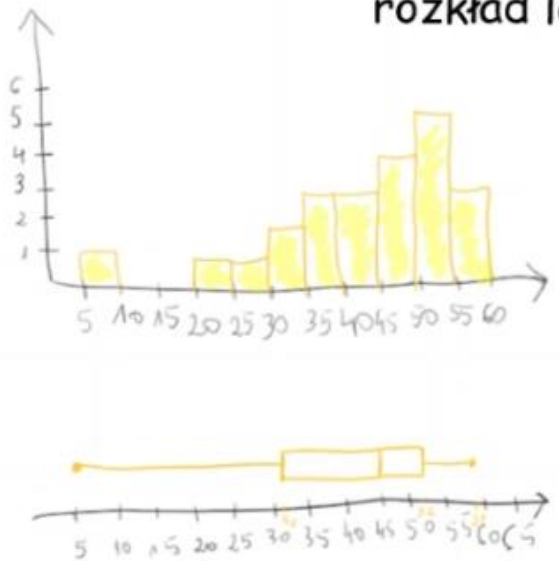
rozkład prawostronnie skośny



rozkład asymetryczny



rozkład lewostronnie skośny

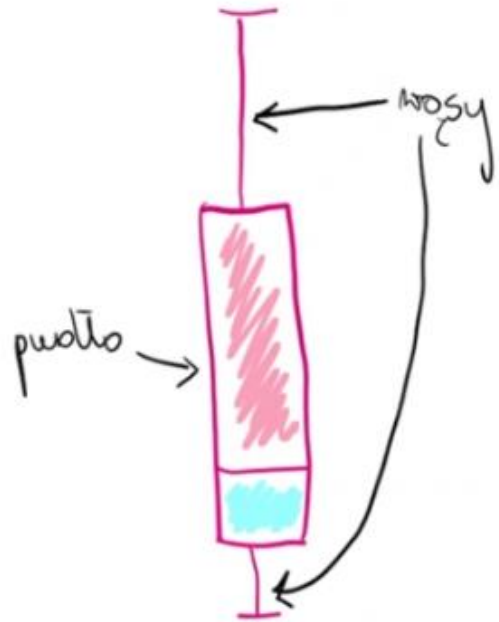


Graficzna prezentacja danych

WYKRES:
RAMKA-WĄSY / PUDEŁKOWY / SKRZYNKOWY

Wykres pudełkowy to również graficzny sposób przedstawienia rozkładu cechy statystycznej.

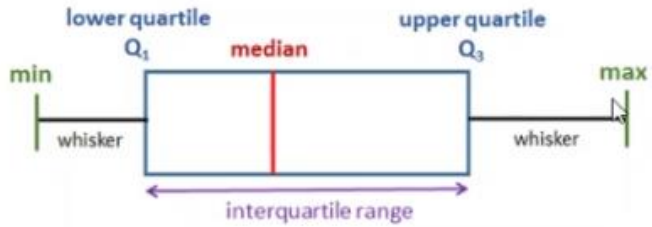
Zawiera informacje odnośnie **położenia**, **rozproszenia** i **kształtu** rozkładu danych. Ze względu na swoją kompleksowość (m.in. użycie ME, Q1, Q3) wykres pudełkowy jest często używany do **porównania** „różnych rzeczy".



Graficzna prezentacja danych

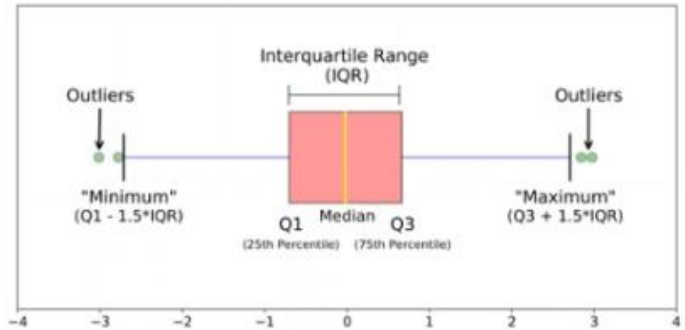
WYKRES:
RAMKA-WĄSY / PUDEŁKOWY / SKRZYNKOWY

wersja „MIN-MAX"



1. **Długość pudełka** jest równa rozstępowi ćwiartkowemu **Q3–Q1=IQR**
2. **Wąsy** mogą być **różnej długości**- wszystko zależy od tego ile wynosi wartość maksymalna i minimalna
3. **Mediana** nie musi leżeć na środku pudełka

wersja „1,5•IQR"



1. **Długość pudełka** jest równa rozstępowi ćwiartkowemu **Q3–Q1=IQR**
2. **Wąsy** są **tej samej** długości
3. **Mediana** nie musi leżeć na środku pudełka
4. Wykres służy również do **znajdowania** wartości **odstających**, które w sposób znaczący mogą zaburzać interpretację

Graficzna prezentacja danych

WYKRES:
PUDEŁKOWY

Wykres pudełkowy wykorzystano do graficznego przedstawienia zbioru informacji o cenach danego wyrobu na terenie województwa lubuskiego. Ceny wyrobu zawarto w tabeli.

Pomiar	1	2	3	4	5	6	7	8	9	10	11
Cena w (zł)	12	10	13	10	10	15	9	40	11	10	16

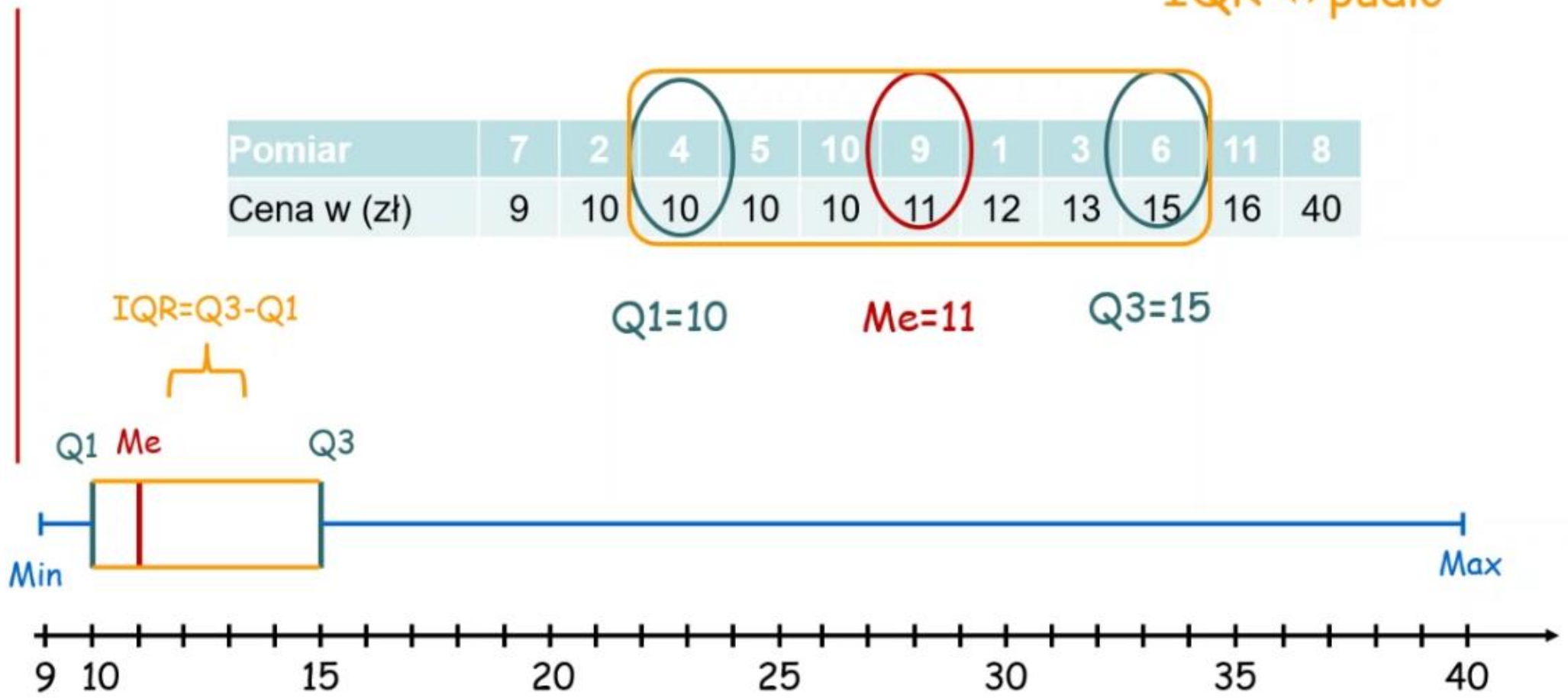
Pomiar	7	2	4	5	10	9	1	3	6	11	8
Cena w (zł)	9	10	10	10	10	11	12	13	15	16	40

Graficzna prezentacja danych

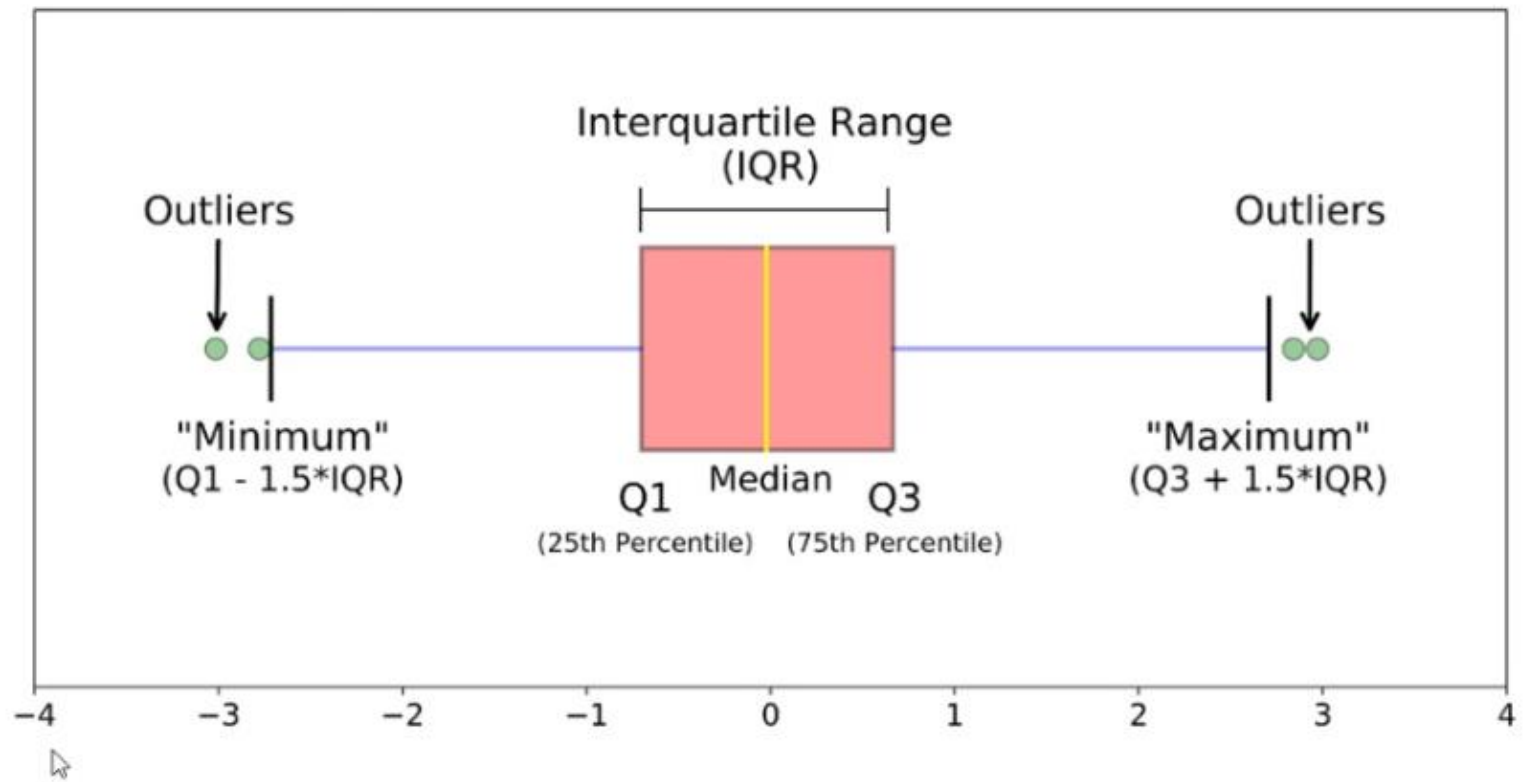
WYKRES:
PUDEŁKOWY – wersja MIN-MAX

IQR ⇔pudło

Pomiar	7	2	4	5	10	9	1	3	6	11	8
Cena w (zł)	9	10	10	10	10	11	12	13	15	16	40

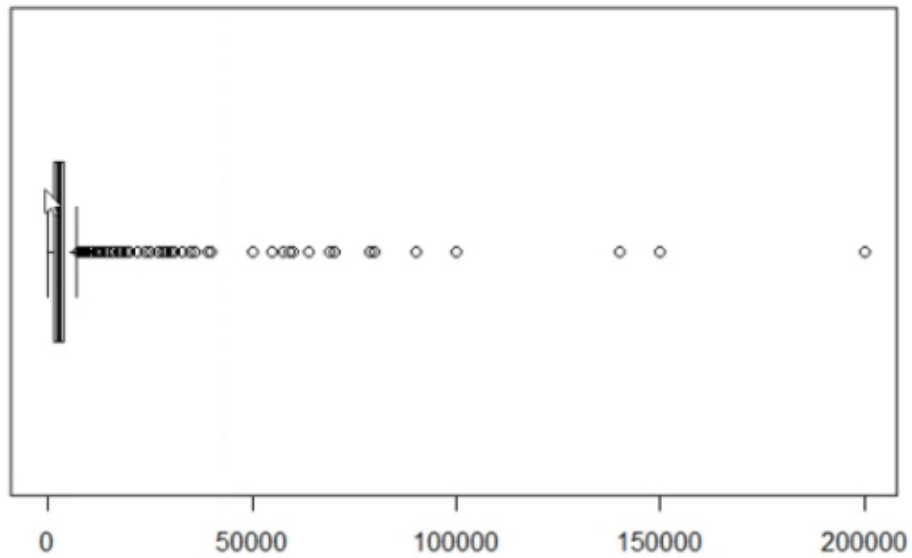


Boxplot



Wykres pudełkowy to ustandaryzowany sposób wyświetlania rozkładu danych na podstawie podsumowania pięciu liczb

Boxplot



Wykres ten nie jest rekomendowany dla dany silnie skośnych o znacznej ilości danych odstających

Graficzna prezentacja danych

INTERPRETACJA

WYKRES PUDEŁKOWY

I OSTROŻNE SPRAWDZENIE **KSZTAŁTU DANYCH**

» Jeżeli wykres **wygląda symetrycznie** względem mediany to możemy **podejrzewać** że jest symetryczny.

» W przypadku danych **skośnych** wykres jest niesymetryczny, jeśli **dłuższa część** pudełka znajduje się **powyżej mediany** to dane określa się jako **prawoskośne**, a jeśli **poniżej – lewostronnie skośne, ale…** trzeba **pamiętać** o **wąsach !!!**

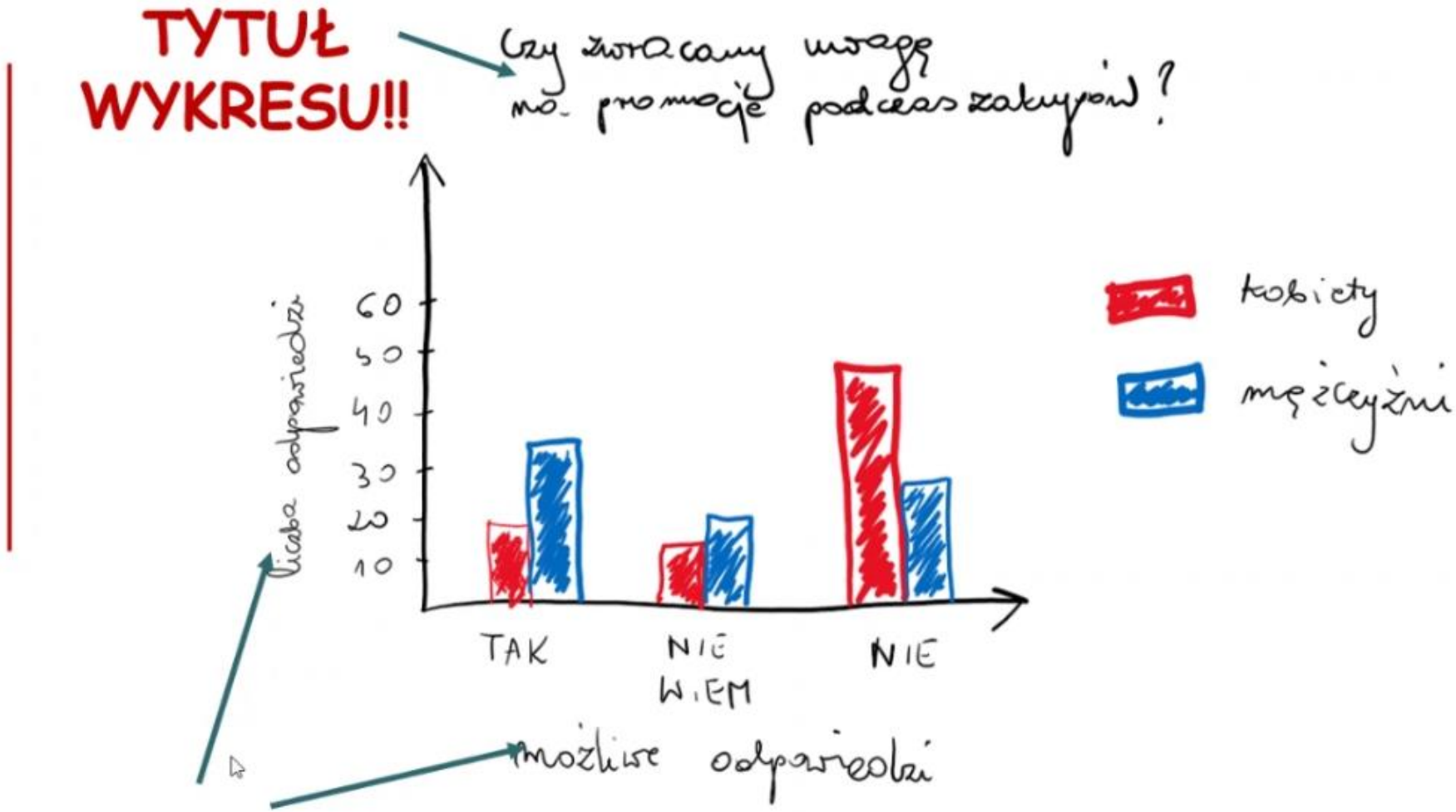
II OBRAZ POŁOŻENIA „CENTRUM" DANYCH – Mediana

III POZIOM ROZPROSZENIA - im **dłuższy** wykres/jego cześć tym dane są **bardziej rozproszone** tzn. mogą przyjmować bardziej różniące się wartości.

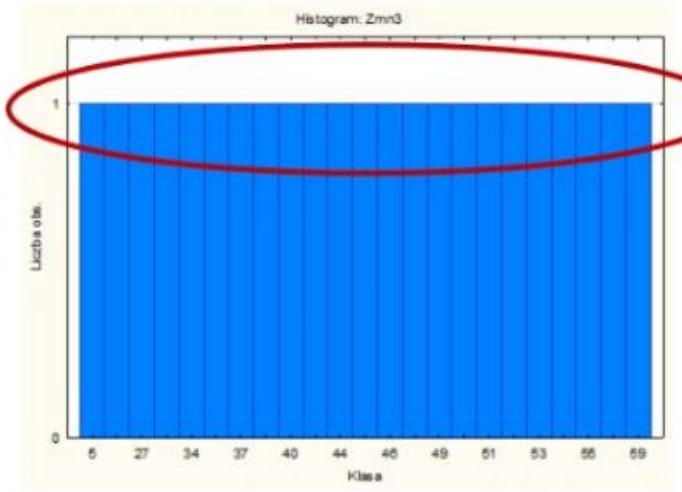
!!! Na podstawie wykresu pudełkowego **nie da się określić liczebności próby**, bo **opiera się** on na **wartościach procentowych** a nie zliczeniach. Każda sekcja wykresu pudełkowego zawiera 25% danych.

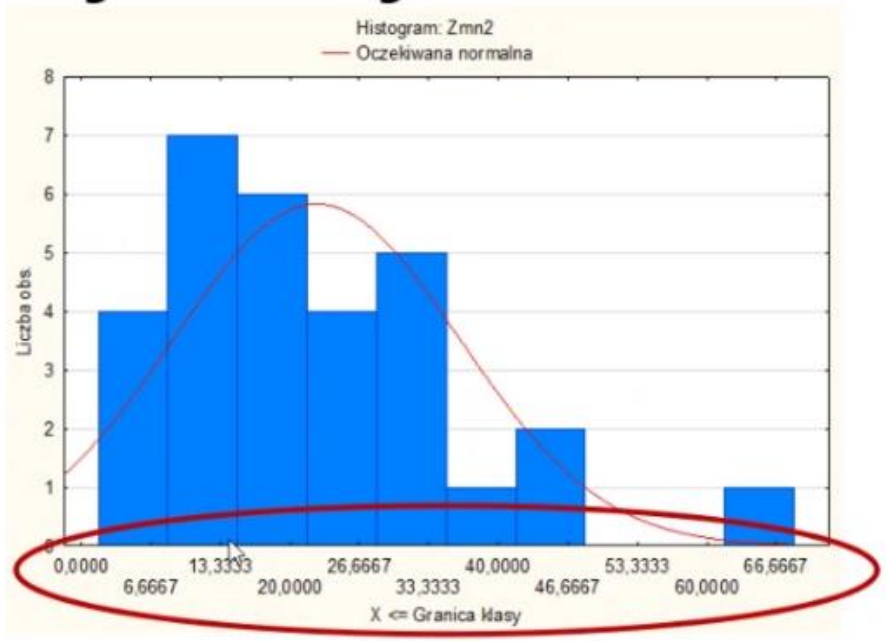
!!! Jeżeli jedna **część pudełka** jest **większa** od drugiej oznacza to **większy poziom rozproszenia** danych.

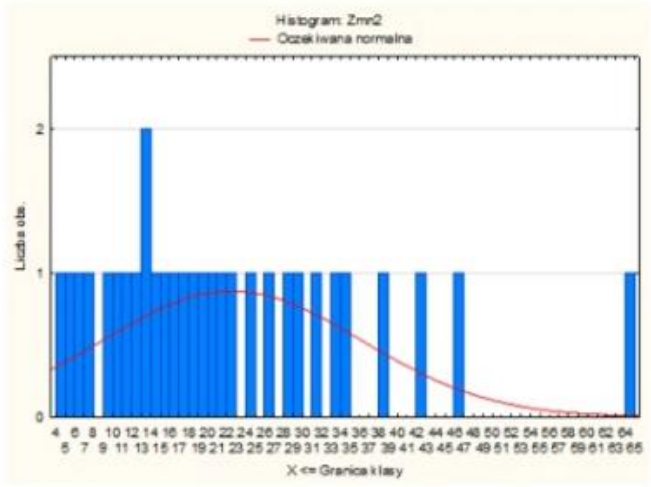
Graficzna prezentacja danych

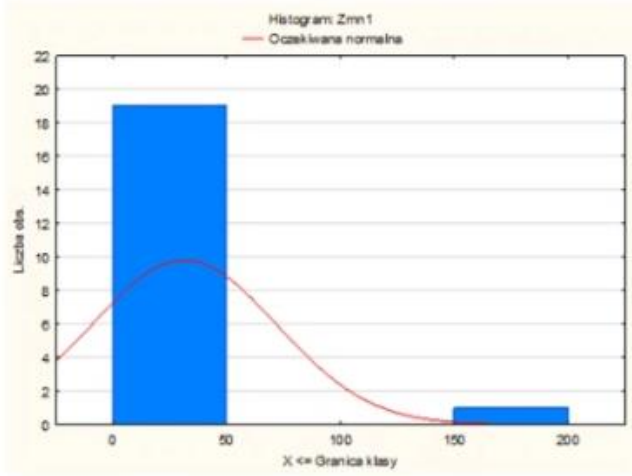


Graficzna pre danych
Nie najlepsze wykresy

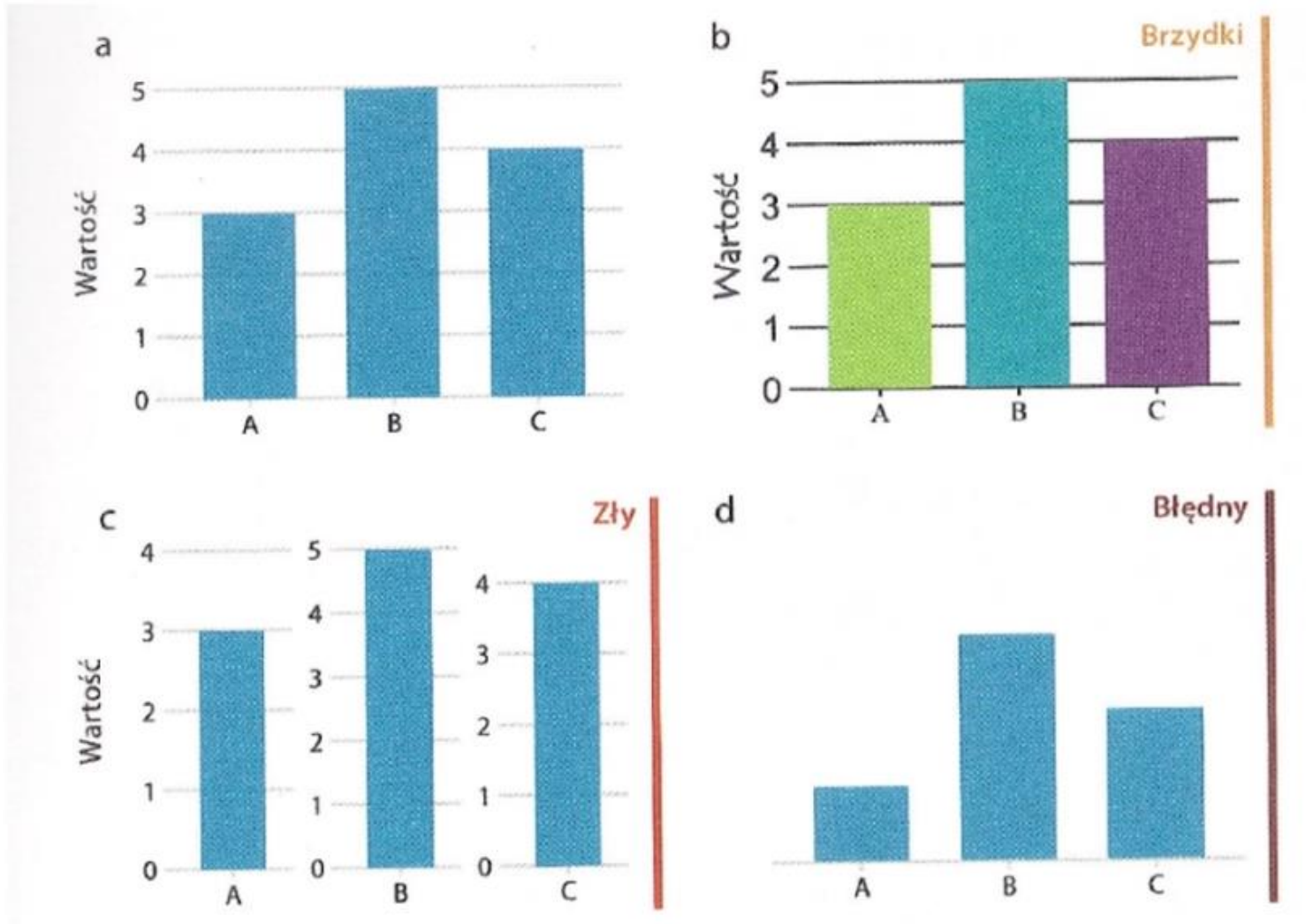








Graficzna prezentacja danych
Nie najlepsze wykresy



Savoir-vivre wykresów

cytat: Janina Bąk „Statystycznie rzecz biorąc”

1. Wykres to nie walentynka – musi być **PODPISANY**. Dotyczy to zarówno **tytułu**, jak i opisania odpowiednich **osi**.

2. Wszystkie **elementy** wykresy (np. kolory, kształty punktów, tekstury) są jak dzieci - muszą odpowiednio **NAZWANE**. Odpowiednio to znaczy, by nikt nie miał żadnych wątpliwości, czego dotyczą i co oznaczają. (…)

3. Pamiętajcie, że 25% dużej pizzy to nie to samo co 25% małej pizzy – zawsze **podawajcie LICZBĘ WSZYSTKICH OBSERWACJI N**, które obejmują wykres.

4. Wizualizacja to nie aktywność fizyczna – jej celem nie jest utrudnienie ludziom życia. Uproście swoim odbiorcom zrozumienie wykresu – np. dzięki dodaniu **ETYKIET Z WARTOŚCIAMI**.

5. Wykres to nie pisanka – **KOLORY** stosujcie **ostrożnie**.

6. (…) pokazujcie na wykresie dokładnie tyle, ile jest niezbędne, i **PODKREŚLAJCIE** to co **NAJWAŻNIEJSZE**. Wykres to nie studniówka – nie potrzebuje brokatu, udziwnień i innych fiu-bździu.

Kluczowe punkty

- Wybór wykresów zależy od skali pomiarowej/rodzaju zmiennej:
  - Zmienne ilościowe (histogram częstości, wykres ramka-wąsy)
  - Zmienne jakościowe (wykres słupkowy, wykres kołowy)
- Zwykle nie jest właściwe używanie wykresów kołowych dla więcej niż 5
- Na podstawie wykresu pudełkowego nie da się określić liczebności próby. Jeżeli jedna część pudełka jest większa od drugiej oznacza to większy poziom rozproszenia danych.
- Wykresy powinny ułatwić odbiorcy zrozumienie i interpretację danych – tak aby mógł to zrobić jak najszybciej i jak najtrafniej.