Statystyczna analiza danych

dr Marcin Woźniak

woz@amu.edu.pl

- wtorek 13-15

(po wcześniejszym umówieniu!)

Literatura

- *Statystyczna analiza danych z wykorzystaniem programu R*, red. M. Walesiak, E. Gatnar, PWN 2012.
- B. Everitt, T. Hothorn, *A Handbook of Statistical Analyses Using R*, Taylor & Francis 2010.
- Robert Nisbet, Gary Miner and Ken Yale, *Handbook of Statistical Analysis and Data Mining Applications*, Elsevier, 2018.
- W. N. Venables i B. D. Ripley, *Modern Applied Statistics with S*, Springer 2002.

Warunki zaliczenia

- Egzamin pisemny z treści wykładów

  (druga połowa stycznia 2023)

Analiza statystyczna danych

Plan wykładów

- Czym jest statystyka i kontekst historyczny
- Dane statystyczne
- Metoda statystyczna
- Podstawowe miary opisu danych (miary rozrzutu i środka, miary zależności i bliskości, współczynnik Giniego i krzywa Lorenza)

Plan wykładów

- Eksploracja danych (podstawowe algorytmy klasyfikacji danych: k-means, DB-scan, metody hierarchiczne)
- Modele obliczeniowe
- Predykcja statystyczna:

  - prawdopodobieństwo; typy błędów wnioskowania; rozkłady danych

  - metody parametryczne i nieparametryczne (modele liniowe i nieliniowe; modele uczenia maszynowego, modele szeregów czasowych)

Analiza statystyczna danych

Czym jest statystyka?

https://www.youtube.com/watch?v=4DruxASC1kM

Czym jest statystyka?

"*Statistics is the branch of scientific method which deals with the data obtained by counting or measuring the properties of populations of natural phenomena. In this definition 'natural phenomena' includes all the happenings of the external world, whether human or not*."

Professor Maurice Kendall, 1943

Analiza statystyczna danych

"*Statistics is: the fun of finding patterns in data; the pleasure of making discoveries; the import of deep philosophical questions; the power to shed light on important decisions, and the ability to guide decisions..... in business, science, government, medicine, industry…*"

Professor David Hand, 2009

Czym jest statystyka?

- Dziedzina matematyki zajmująca się zbieraniem, analizą, wizualizacją i interpretacją danych
- Dwa (główne) działy: statystyka opisowa i wnioskowanie statystyczne
- Dostarcza metod do ilościowej analizy otaczającej nas rzeczywistości

Czym jest statystyka?

- Dziedzina matematyki zajmująca się zbieraniem, analizą, wizualizacją i interpretacją danych
- Dwa (główne) działy: statystyka opisowa i wnioskowanie statystyczne
- Dostarcza metod do ilościowej analizy otaczającej nas rzeczywistości
- Termin “statystyka” może też odnosić się do konkretnych wskaźników
- Rozwój wielu nowych technik analitycznych opartych na statystyce

Kontekst historyczny

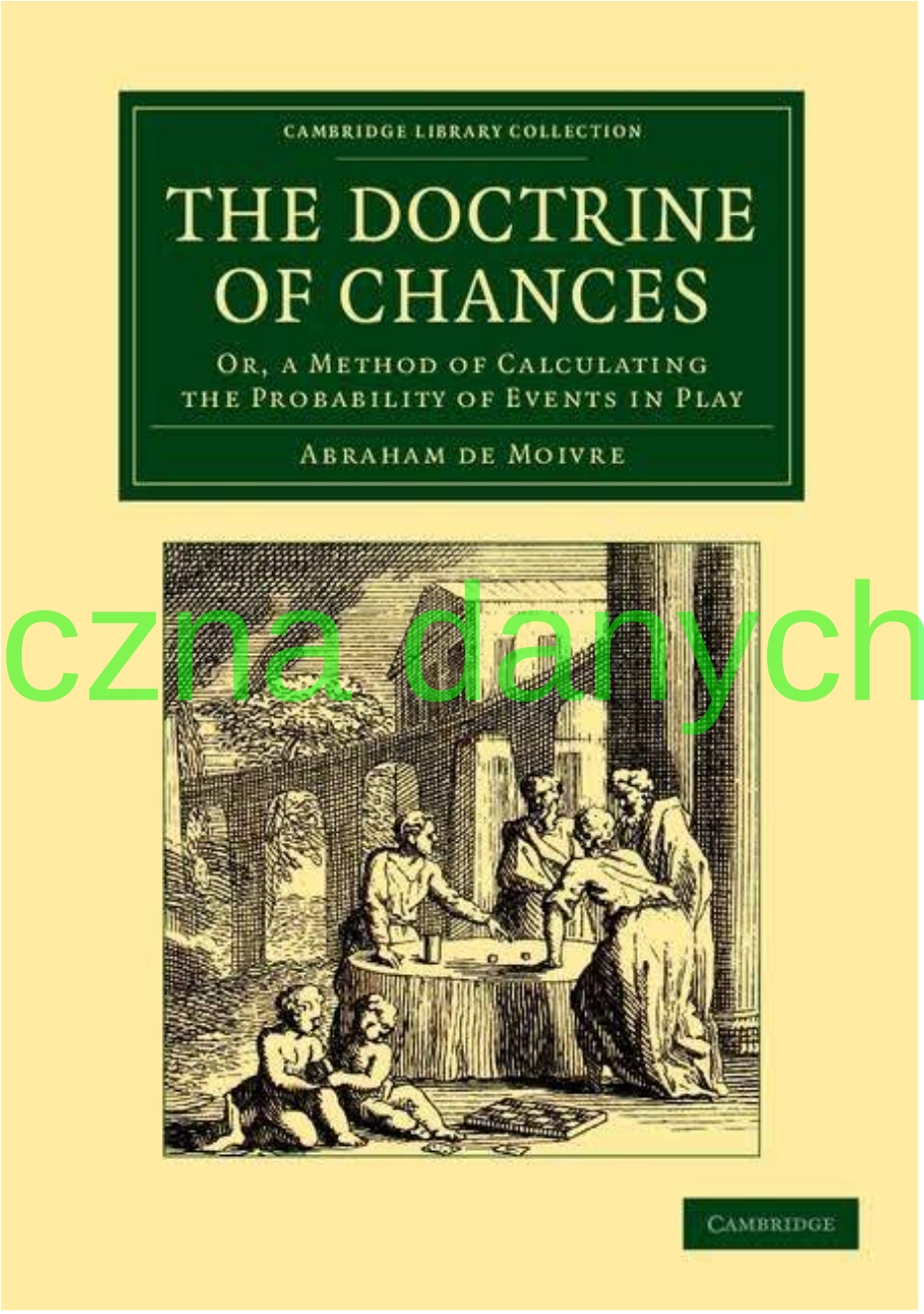
- Relatywnie nowa dyscyplina nauki
- Większość dorobku powstała w ostatnich 150 latach
- Początkowo głównym czynnikiem rozwoju był hazard
- Dualność współczesnej analizy statystycznej: ujęcie klasyczne (Fisher) vs ujęcie Bayesowskie (Bayes)

Początki...

Przypuśćmy, że mamy stos 13 kart w jednym kolorze i inny stos 13 kart w innym kolorze.

Załóżmy, że każdy stos zawiera po jednym asie.

Jakie jest prawdopodobieństwo, że biorąc po jednej karcie z każdego stosu, wyciągnę dwa asy?



Początki...

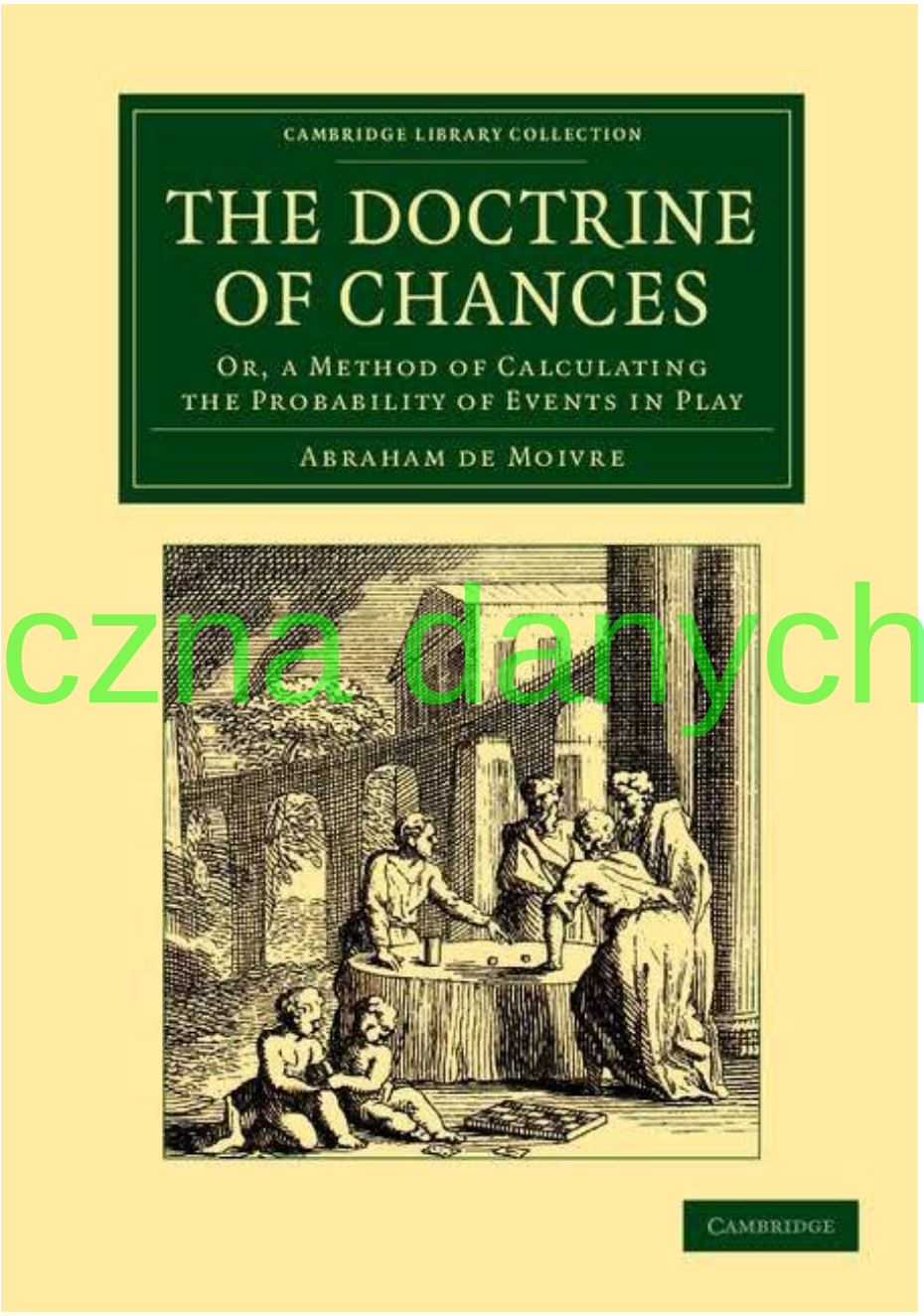
Przypuśćmy, że mamy stos 13 kart w jednym kolorze i inny stos 13 kart w innym kolorze.

Załóżmy, że każdy stos zawiera po jednym asie.

¿

Jakie jest prawdopodobieństwo, że biorąc po jednej karcie z każdego stosu, wyciągnę dwa asy?

$$1/13 * 1/13 = 1/169$$



Ujęcie Bayesa

Thomas Bayes
1702-1761

- Prawdopodobieństwo wystąpienia zdarzenia w **przyszłości** jest równe prawdopodobieństwu jego wystąpienia w **przeszłości** podzielonemu przez prawdopodobieństwo wystąpienia wszystkich konkurencyjnych zdarzeń.

Analiza statystyczna danych

Ujęcie Bayesa

Thomas Bayes
1702-1761

- Prawdopodobieństwo wystąpienia zdarzenia w **przyszłości** jest równe prawdopodobieństwu jego wystąpienia w **przeszłości** podzielonemu przez prawdopodobieństwo wystąpienia wszystkich konkurencyjnych zdarzeń.

- Analiza przebiega w oparciu o pojęcie **prawdopodobieństwa warunkowego:** prawdopodobieństwa wystąpienia zdarzenia przy założeniu, że inne zdarzenie już wystąpiło.

- Rozpoczyna się od kwantyfikacji istniejącego stanu wiedzy badacza, przekonań i założeń na temat przeszłych zdarzeń.

Twierdzenie Bayesa

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Jakie jest prawdopodobieństwo zachorowania na raka w wieku 65 lat?

Załóżmy, że ogólna częstość występowania raka wynosi 2% (wcześniejsze prawdopodobieństwo zachorowania na raka – *apriori*).

Następnie, załóżmy, że prawdopodobieństwo bycia w wieku 65 lat wynosi 0,3% i że prawdopodobieństwo, że ktoś, u kogo zdiagnozowano raka ma 65 lat, wynosi 0,4%.

Mając te dane, możemy obliczyć prawdopodobieństwo zachorowania na raka jako 65-latek.

Twierdzenie Bayesa

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Jakie jest prawdopodobieństwo zachorowania na raka w wieku 65 lat?

Załóżmy, że ogólna częstość występowania raka wynosi 2% (wcześniejsze prawdopodobieństwo zachorowania na raka – *apriori*).

Następnie, załóżmy, że prawdopodobieństwo bycia w wieku 65 lat wynosi 0,3% i że prawdopodobieństwo, że ktoś, u kogo zdiagnozowano raka ma 65 lat, wynosi 0,4%.

Mając te dane, możemy obliczyć prawdopodobieństwo zachorowania na raka jako 65-latek.

0.004*0.02/0.003 = 0.026 (2.6%)

Model parametryczny

Ronald Fisher
1890-1962

- Eliminacja pojęcia prawdopodobieństwa *apriori* na rzecz prawdopodobieństwa *rzeczywistego*
- Szereg założeń, m.in.:

  - dane wpasowują się w jeden ze znanych rozkładów prawdopodobieństwa (np. rozkład normalny (Gaussa))

  - niezależność danych objaśniających

Model parametryczny

Ronald Fisher
1890-1962

- Eliminacja pojęcia prawdopodobieństwa *apriori* na rzecz prawdopodobieństwa *rzeczywistego*
- Szereg założeń, m.in.:

  - dane wpasowują się w jeden ze znanych rozkładów prawdopodobieństwa (np. rozkład normalny (Gaussa))

  - niezależność danych objaśniających

  - efekty oddziaływania jednej zmiennej na drugą mają charakter liniowy

  - dane muszą być numeryczne i ciągłe
- Dylemat: warunki laboratoryjne a świat rzeczywisty? (Fisher czy Bayes?)

Dalszy rozwój analizy statystycznej

- II generacja (lata > 80 XX w.) – metody skupione na modelach nieliniowych i zmiennych dyskretnych (np. modele logitowe i logistyczne)
- III generacja (lata > 2010 XXI w.) – metody oparte na uczeniu maszynowym (np. sztuczne sieci neuronowe (ANN), drzewa decyzyjne)

Dane statystyczne

Analiza statystyczna danych

Dane statystyczne

- Statystyka wymaga wykorzystania danych
- Pozyskanie danych (zazwyczaj) wymaga pomiaru
- Różne rodzaje i typy danych
- Klasyczny podział na dane ilościowe i jakościowe

Rodzaje i typy danych

- **Dane ustrukturyzowane (np. dane tabelaryczne)**
- Dane nieustrukturyzowane (np. tekst, video, zdjęcia, audio)
- Dane częściowo ustrukturyzowane (np. logi, xml)

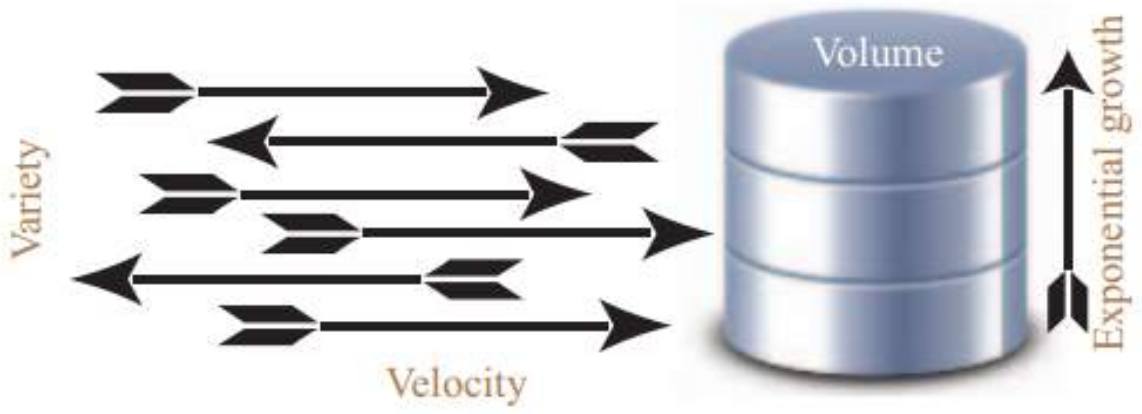
Rodzaje i typy danych

- **Dane ustrukturyzowane (np. dane tabelaryczne)**
- Dane nieustrukturyzowane (np. tekst, video, zdjęcia, audio)
- Dane częściowo ustrukturyzowane (np. logi, xml)
- Dane generowane w czasie rzeczywistym (np. streaming, transakcje bankowe)
- Data at Rest (dane w bibliotekach cyfrowych, np. dane sprzedażowe, dane z urządzeń mobilnych)
- Metadane (dane o danych)

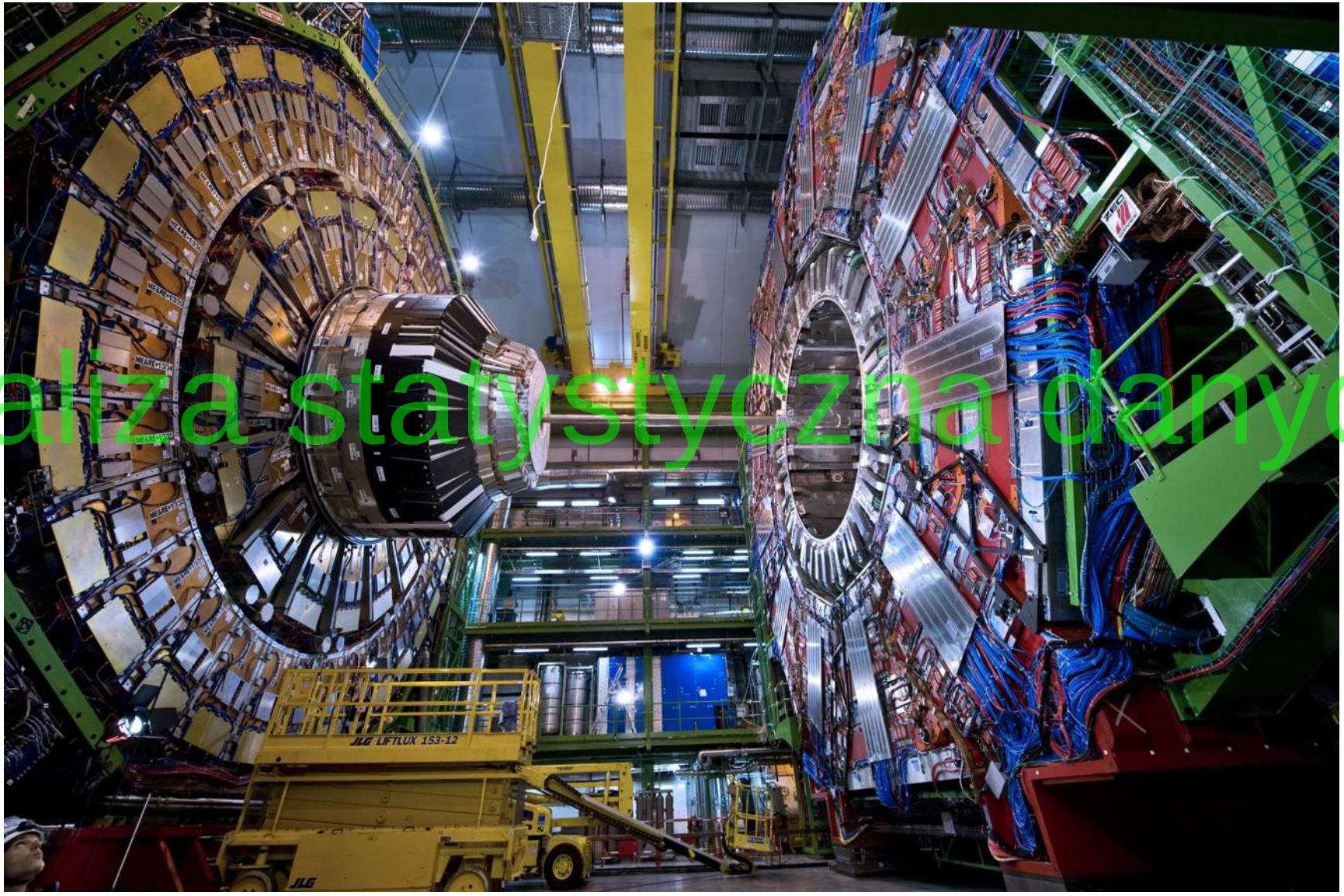
Big Data

Wg Doug'a Laney (2001) – 3V:

- **Volume** - duża (relatywnie) ilość n-wymiarowych danych
- **Velocity** – wysoka zmienność, dynamika
- **Variety** - różnorodność źródeł, typów, formatów



Volume



9 Petabajtów/ miesiąc

Volume

10 TB / silnik / 30 minut

Etapy pracy z danymi

1) Identyfikacja źródeł danych. Dane mogą być zebrane w jednej bazie albo rozproszone.

2) Pozyskanie danych (interfejs API, tablice: .csv, .xlsx, .txt GIS: geopackage, .shp).

Analiza statystyczna danych

Etapy pracy z danymi

1) Identyfikacja źródeł danych. Dane mogą być zebrane w jednej bazie albo rozproszone.

2) Pozyskanie danych (interfejs API, tablice: .csv, .xlsx, .txt GIS: geopackage, .shp).

3) Integracja (harmonizacja) danych: dane mogą być w różnych formatach lub występować na różnych poziomach agregacji lub są wyrażone w różnych jednostkach.

Skale pomiarowe

- **Nominalna:** przypisanie nazw do klas (np. czerwony, żółty, zielony). Skala jest nominalna, jeśli rozróżnia grupy, ale bez żadnego rankingu lub potencjału arytmetycznego. *Np. kolor może być użytecznym atrybutem, ale sam nie ma znaczenia numerycznego*.
- Inne przykłady?

Skale pomiarowe

- **Porządkowa:** kategorie danych, które można uporządkować.
- Może obejmować liczby ujemne i 0.
- Zbiór pozornie uporządkowanych kategorii nie tworzy skali porządkowej. Cecha jest porządkowa, jeśli implikuje ranking (stopniowanie). Operacje arytmetyczne nie mają sensu.
- Przykłady?

Skale pomiarowe

- **Interwałowa:** dane liczbowe, które wykazują porządek, a także możliwość zmierzenia interwału (odległości) pomiędzy dowolną parą obiektów na skali.
- Dane są typu interwałowego, jeśli różnice mają sens, np. daty.
- Inne przykłady?

Skale pomiarowe

- **Ilorazowa:** interwałowa + naturalne pochodzenie (np. stopa bezrobocia, waga ludzi).
- Pozwala określić rozmiar różnic pomiędzy analizowanymi cechami
- Stosunki między dwiema jej wartościami mają interpretację w świecie rzeczywistym.
- Nie nakłada ograniczeń w stosowaniu operacji matematycznych i metod statystycznych.
- Przykłady?

Skale pomiarowe

- **Skala cykliczna:** kąty i czas zegarowy. Pomiary atrybutów, które reprezentują kierunki lub zjawiska cykliczne mają tę własność, że dwa odrębne punkty na skali mogą być równe – np. 0 i 360 stopni.

Skale pomiarowe - podsumowanie

Nazwa uniwersytetu (cecha nominalna)	**Poziom edukacji (cecha porządkowa)**	Data rozpoczęcia sesji letniej (cecha interwałowa)	Ilość studentów (cecha ilorazowa)
Szkoła Handlowa Główna	**Wysoki**	14. czerwca 2019	32300
Uniwersytet Ekonomiczny	**Bardzo wysoki**	21. Czerwca 2019	12760
Politechnika Techniczna	**Bardzo wysoki**	13. Czerwca 2019	18710
SGWG	**Dostateczny**	28. Czerwca 2019	21290

Statystyki opisowe

Analiza statystyczna danych

Statystyki opisowe - narzędzia

- Ilustracja graficzna (np. histogram, wykres pudełkowy, słupkowy, liniowy, kartogramy, kartodiagramy)
- Tabele częstości + tabele krzyżowe
- Miary środka
- Miary rozproszenia
- Miary koncentracji i asymetrii
- Miary zależności i bliskości

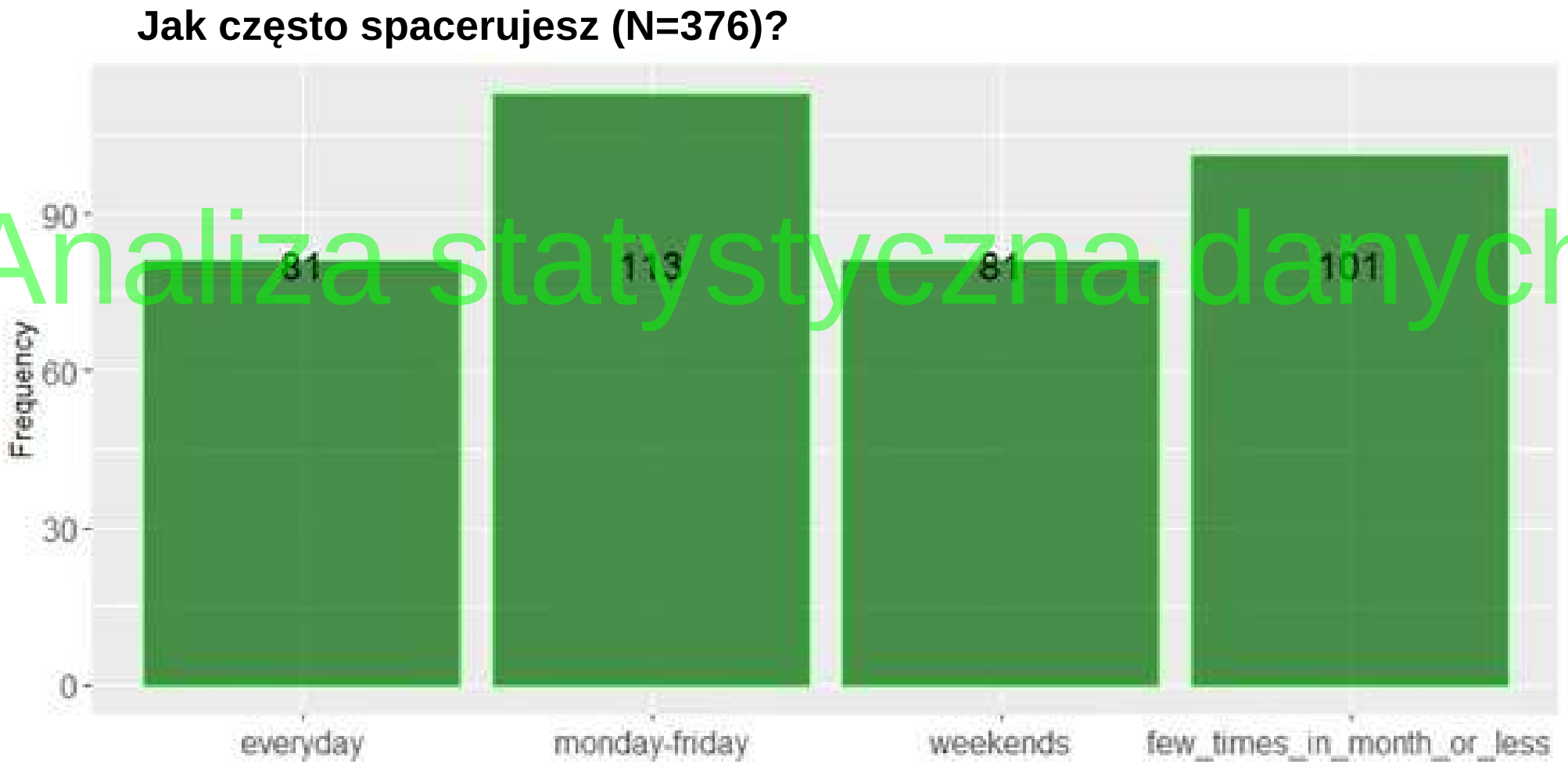
Podstawowe narzędzia graficznej prezentacji danych

Analiza statystyczna danych

Wykres słupkowy

- to wykres, który przedstawia dane za pomocą prostokątnych słupków o wysokości lub długości proporcjonalnej do wartości, które reprezentują
- pokazuje porównania pomiędzy kategoriami danych. Jedna oś wykresu pokazuje konkretne kategorie, a druga oś przedstawia mierzoną wartość.

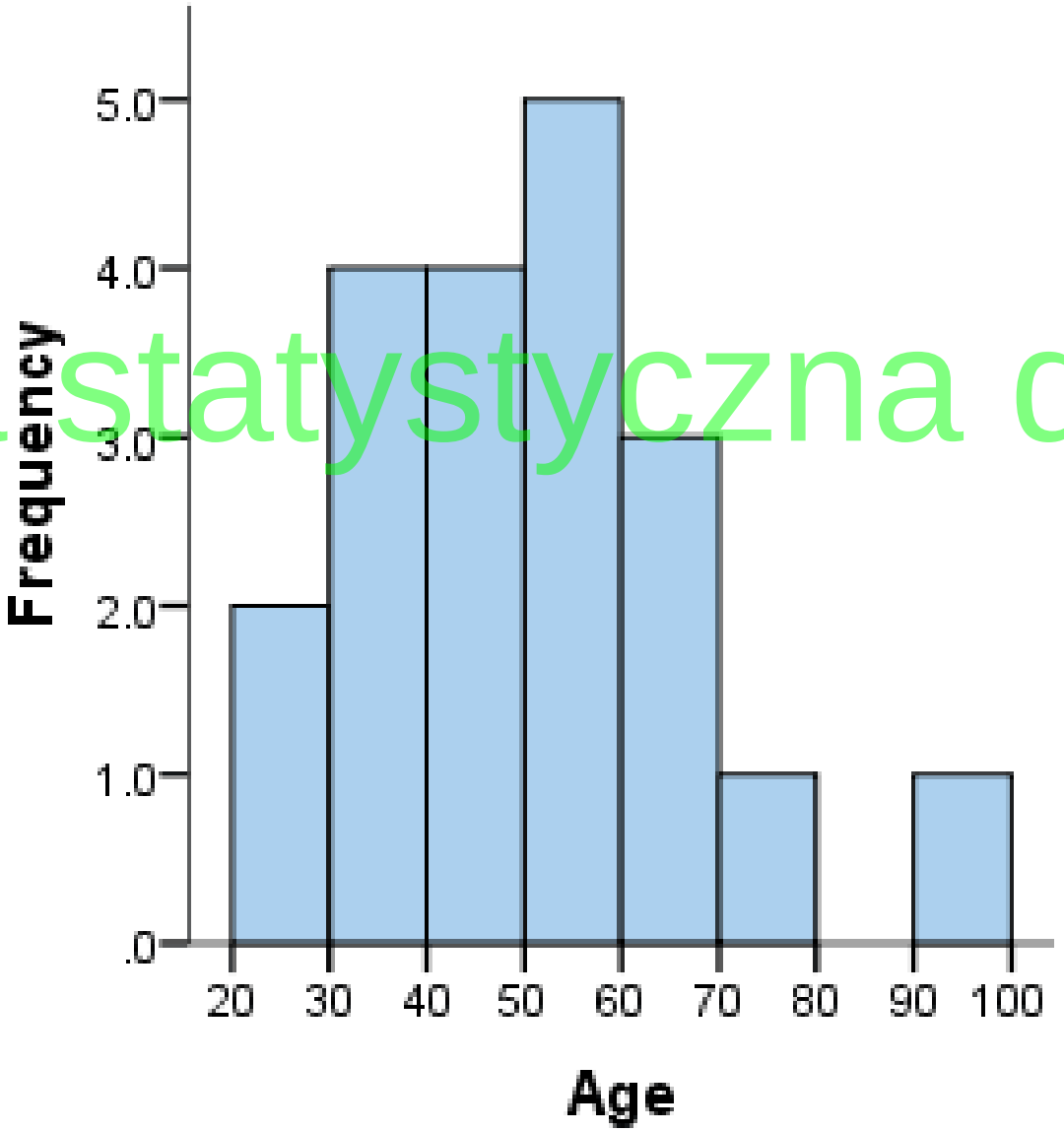
Wykres słupkowy



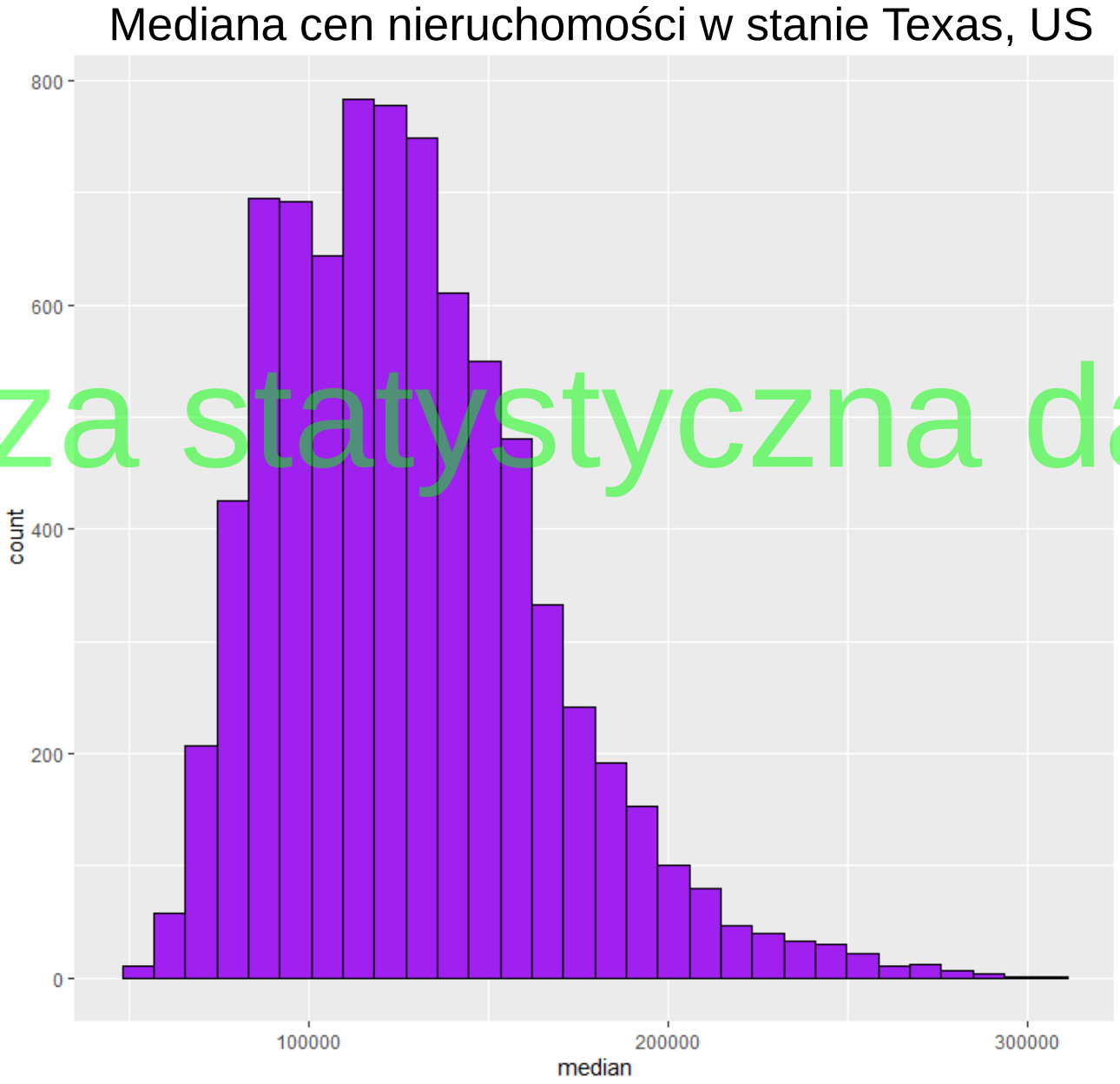
Histogram

- Jeśli pomiary mają wartości liczbowe w skali interwałowej lub ilorazowej, mogą one być pogrupowane w klasy (przedziały)
- Liczby w każdej klasie można zwizualizować jako wykres słupkowy, w którym ważna jest kolejność na osi x.
- Wykres słupkowy tego typu nazywany jest histogramem,

Histogram



Histogram



Przedziały klasowe

- Jeśli dane są pomiarami zmiennej ciągłej, wtedy standardową procedurą w tworzeniu histogramu jest utworzenie przedziałów klasowych i policzenie częstotliwości występujących w każdym przedziale obserwacji.

Przedziały klasowe

- Jeśli dane są pomiarami zmiennej ciągłej, wtedy standardową procedurą w tworzeniu histogramu jest utworzenie przedziałów klasowych i policzenie częstotliwości występujących w każdym przedziale obserwacji.
- Wartości, determinujące przedziały są określane jako punkty odcięcia.
- Kluczowe jest ustalenie liczby klas/przedziałów.

Przedziały klasowe

$$k = \frac{max - min}{h} \quad lub \quad h = \frac{max - min}{k}$$

Gdzie k to liczba klas, a h szerokość przedziału

$$k = 3.5 * sd / n^{(1/3)} \quad (formuła\ Scotta, 1979)$$

Zgodnie z formułą Scotta, ile przedziałów utworzymy dla zbioru liczącego 1000 elementów, w którym odchylenie standardowe wynosi 25?

Wykres pudełkowy (boxplot)

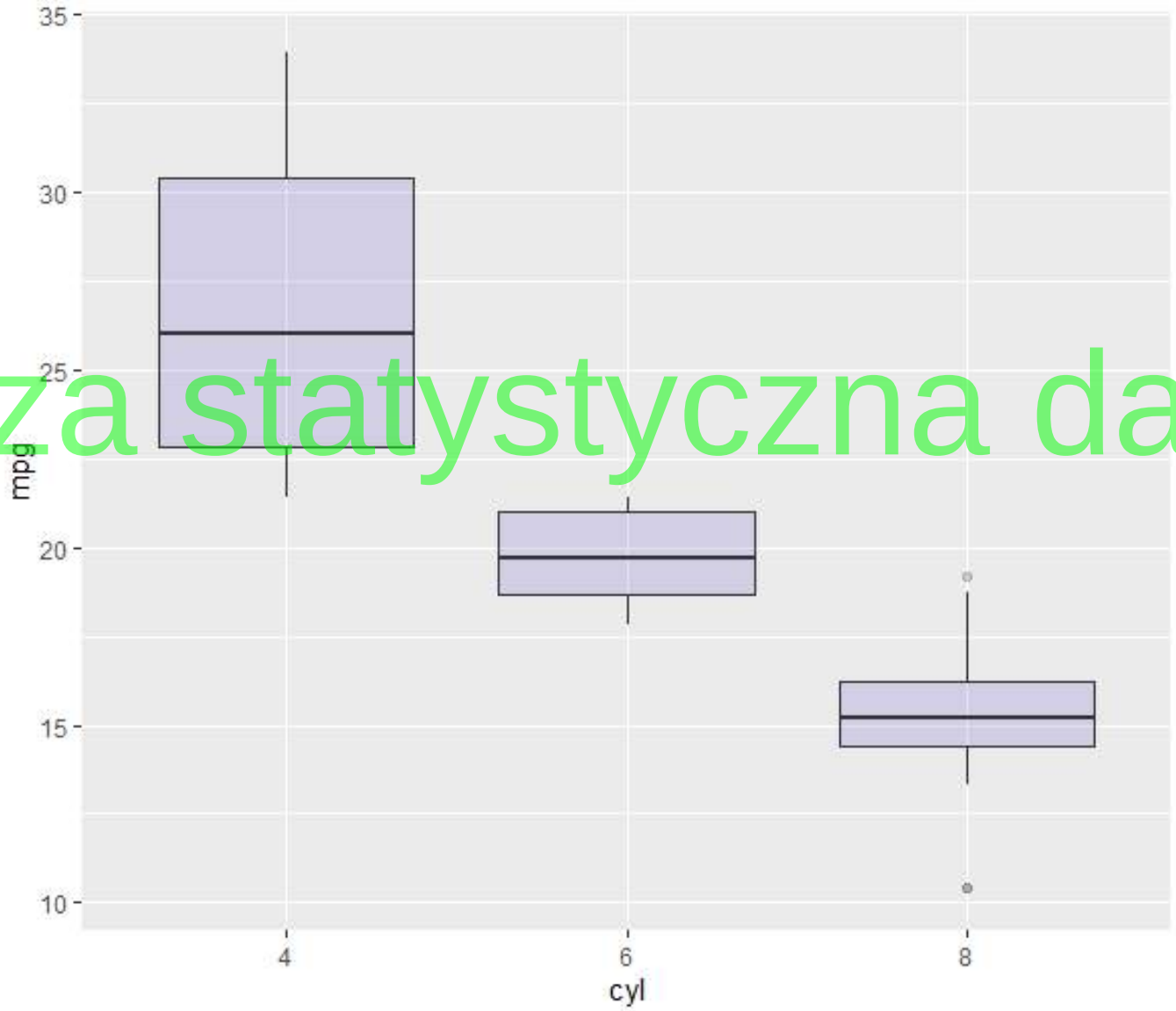
- Wykres pudełkowy (przedstawiony pionowo) tworzy się odkładając na osi y wartości kluczowych parametrów rozkładu

Analiza statystyczna danych

Wykres pudełkowy (boxplot)

- Wykres pudełkowy (przedstawiony pionowo) tworzy się odkładając na osi y wartości kluczowych parametrów rozkładu
- To sposób wyświetlania zbioru danych oparty na podsumowaniu pięciu statystyk: *minimum, maksimum, mediana oraz pierwszy i trzeci kwartyl.*
- Składa się z prostokąta i przylegających do niego “wąsów”

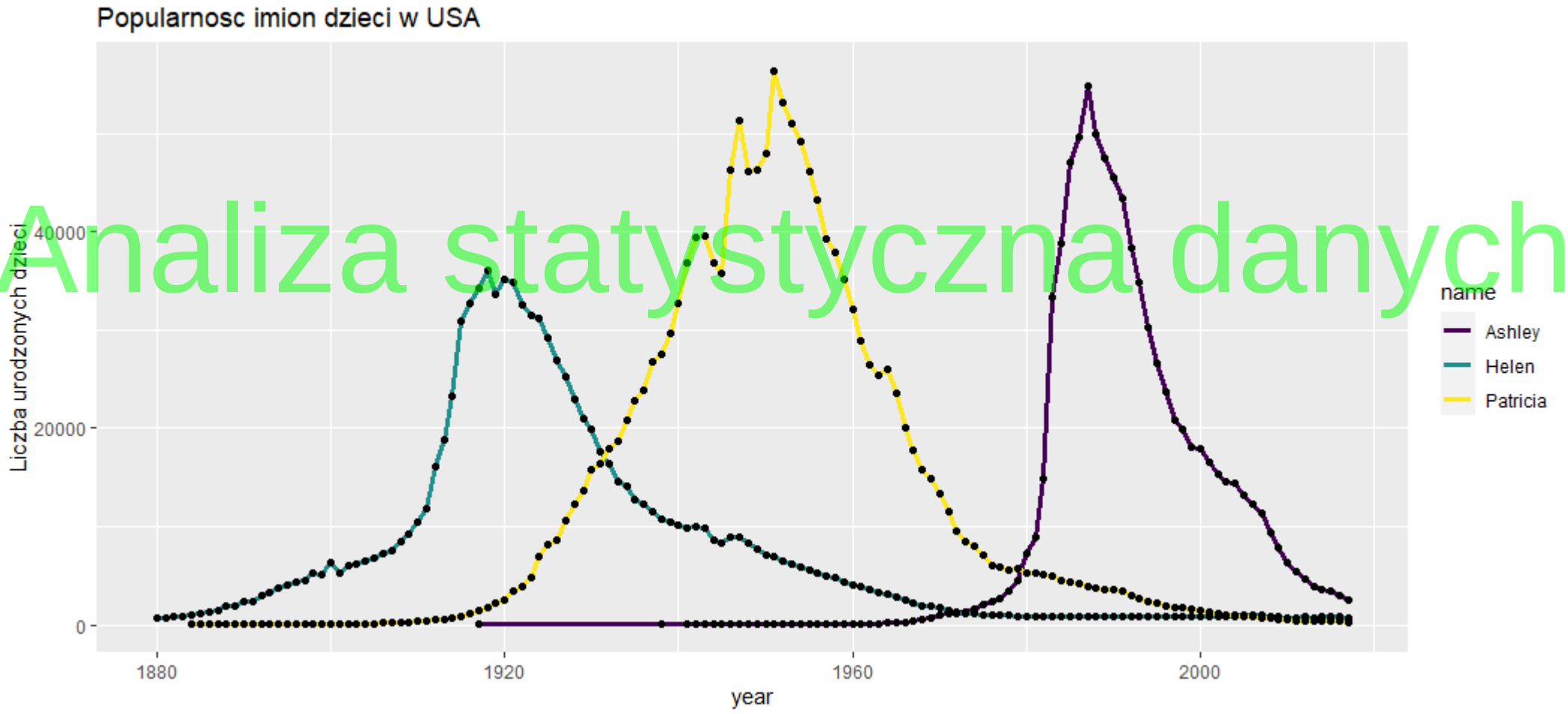
Wykres pudełkowy (boxplot)



Wykres liniowy

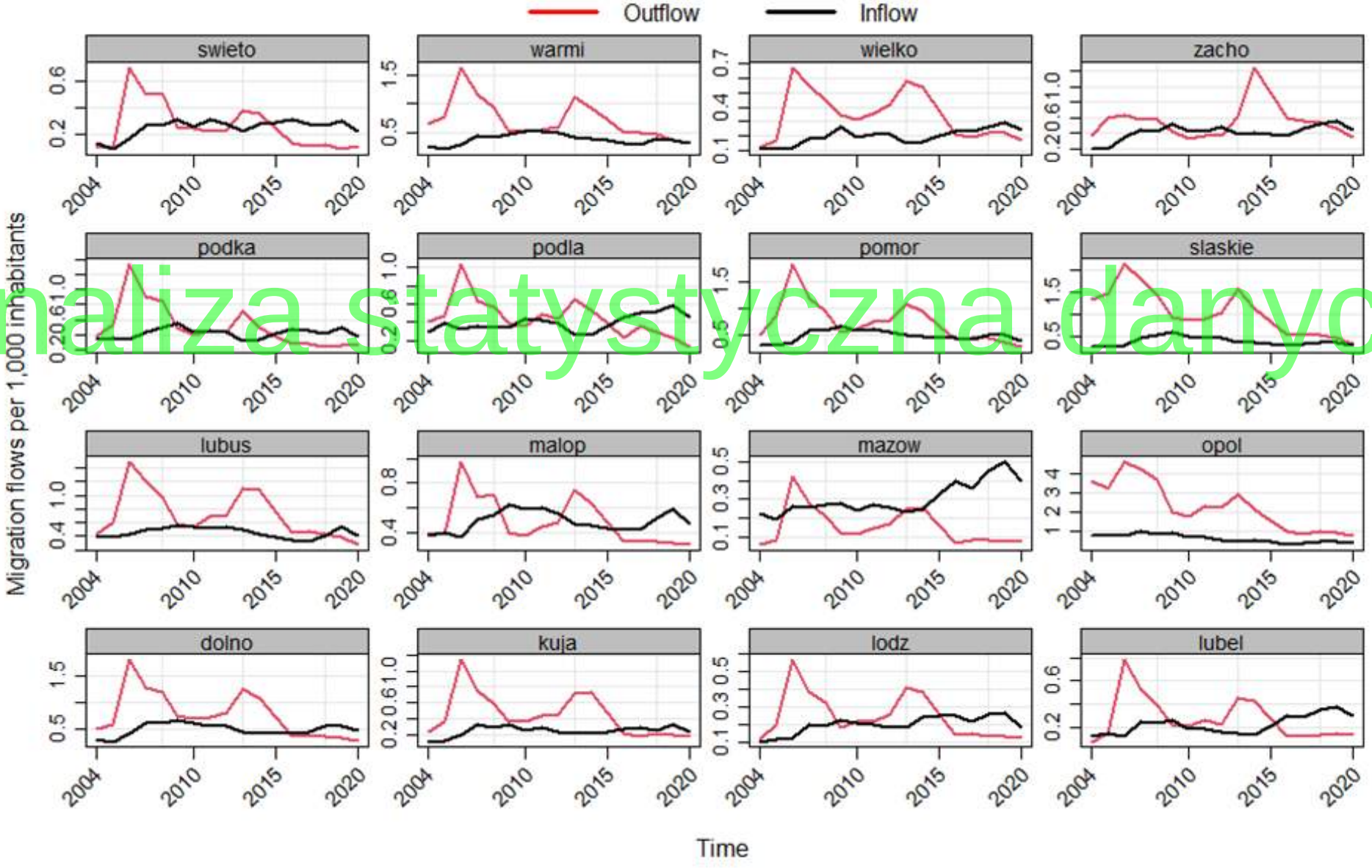
- To rodzaj wykresu, który wyświetla informacje w postaci serii punktów danych zwanych "markerami"
- Punkty te są połączone odcinakmi linii prostej
- Jest często używany do wizualizacji trendu w danych w odstępach czasu – tzw. szeregu czasowego.

Wykres liniowy



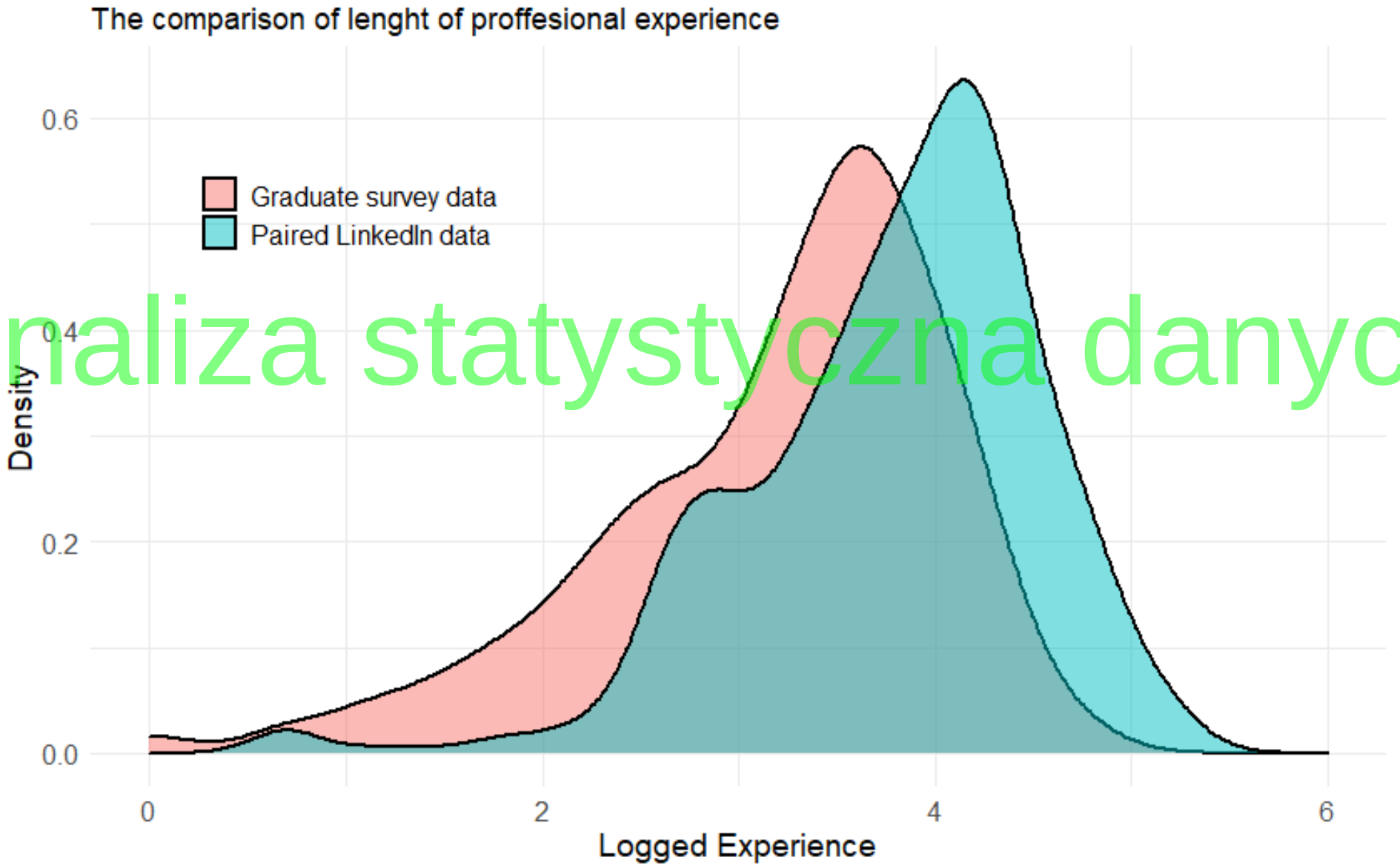
Wykres panelowy liniowy

Wykres. Odpływy i napływy migracyjne do polskich regionów 2004-2020

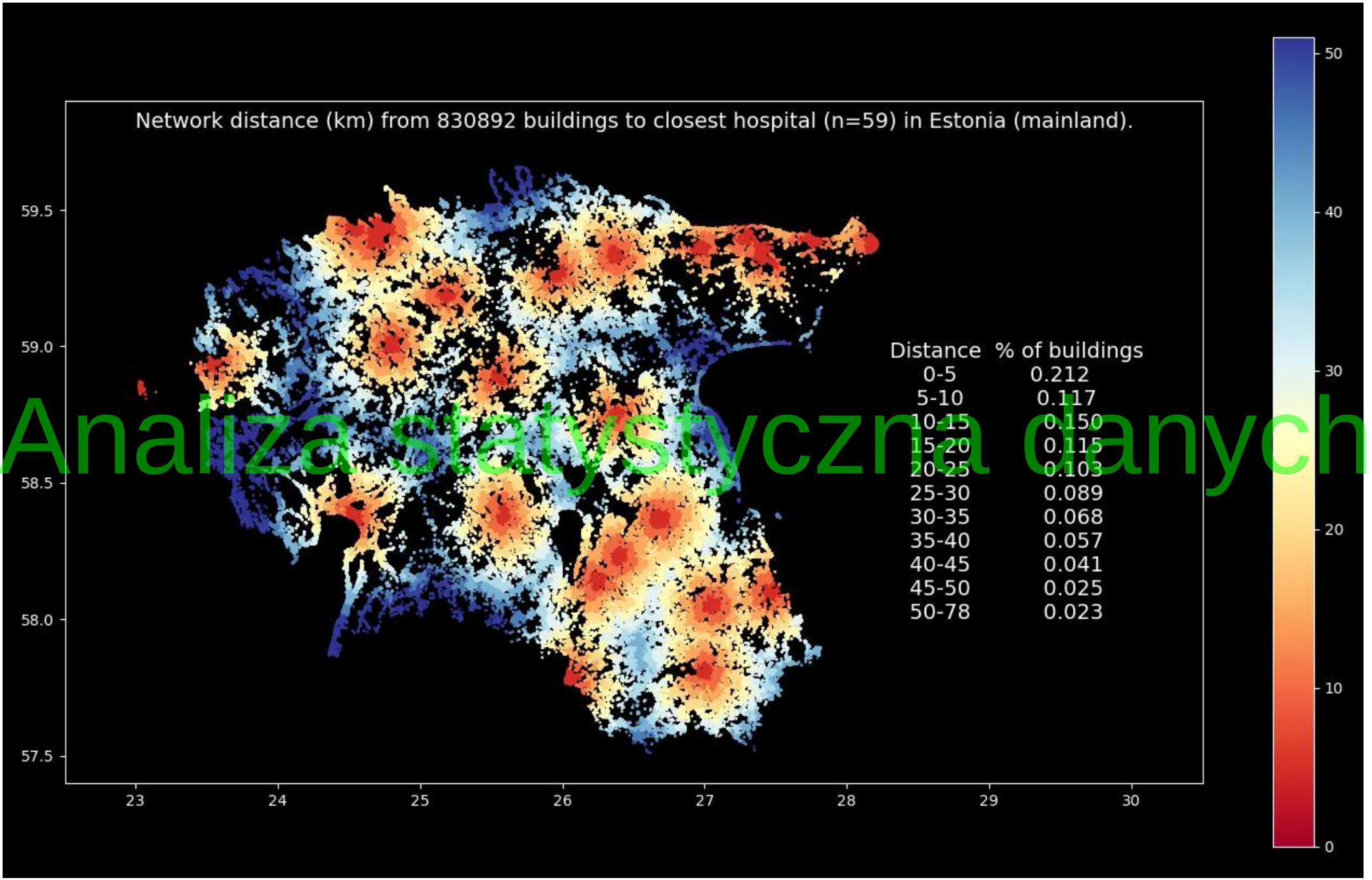


Inne przykłady

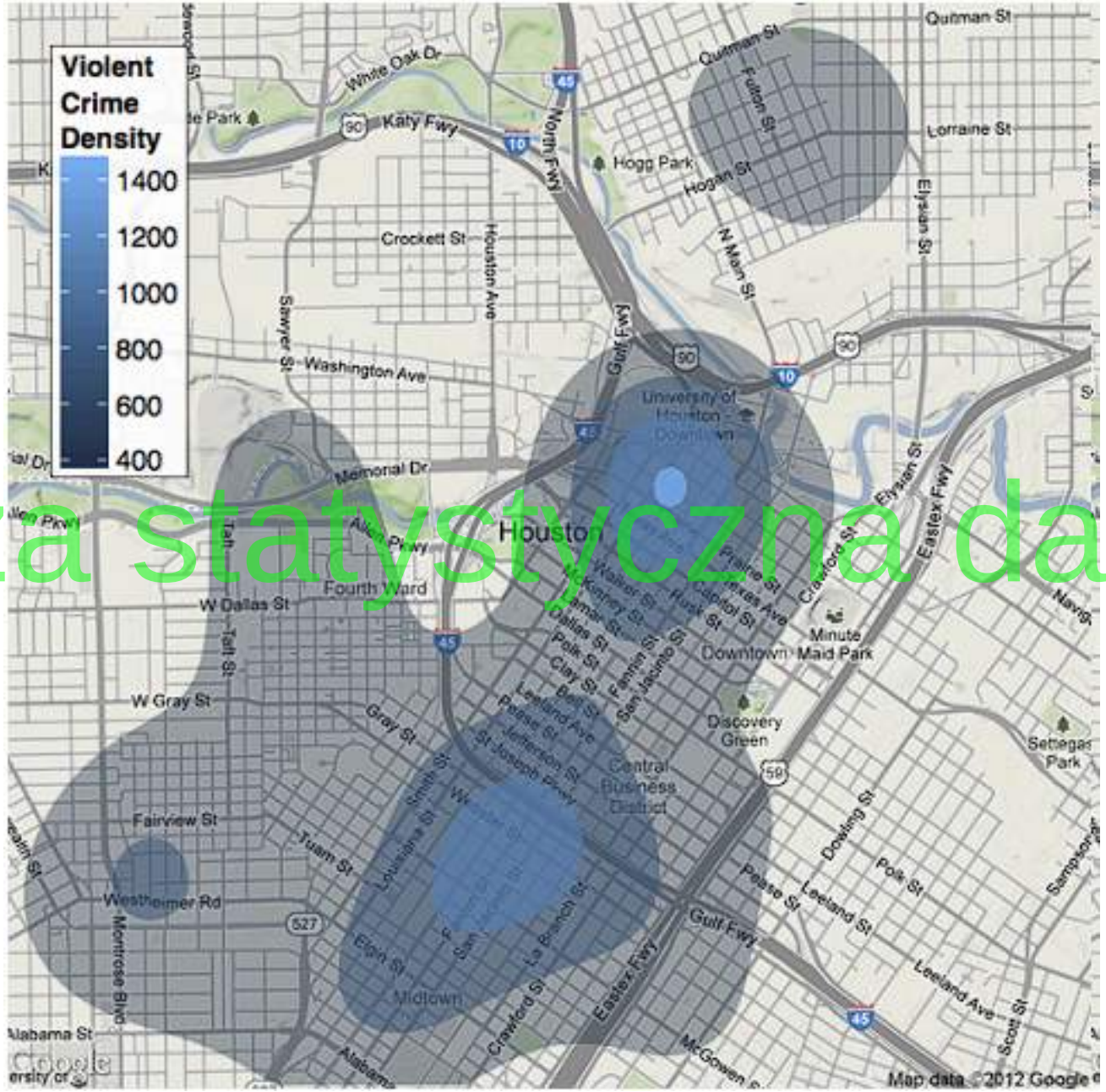
ggplot2 R library – wykres gętości



OSMNX Python library – kartogram - czas dostępności

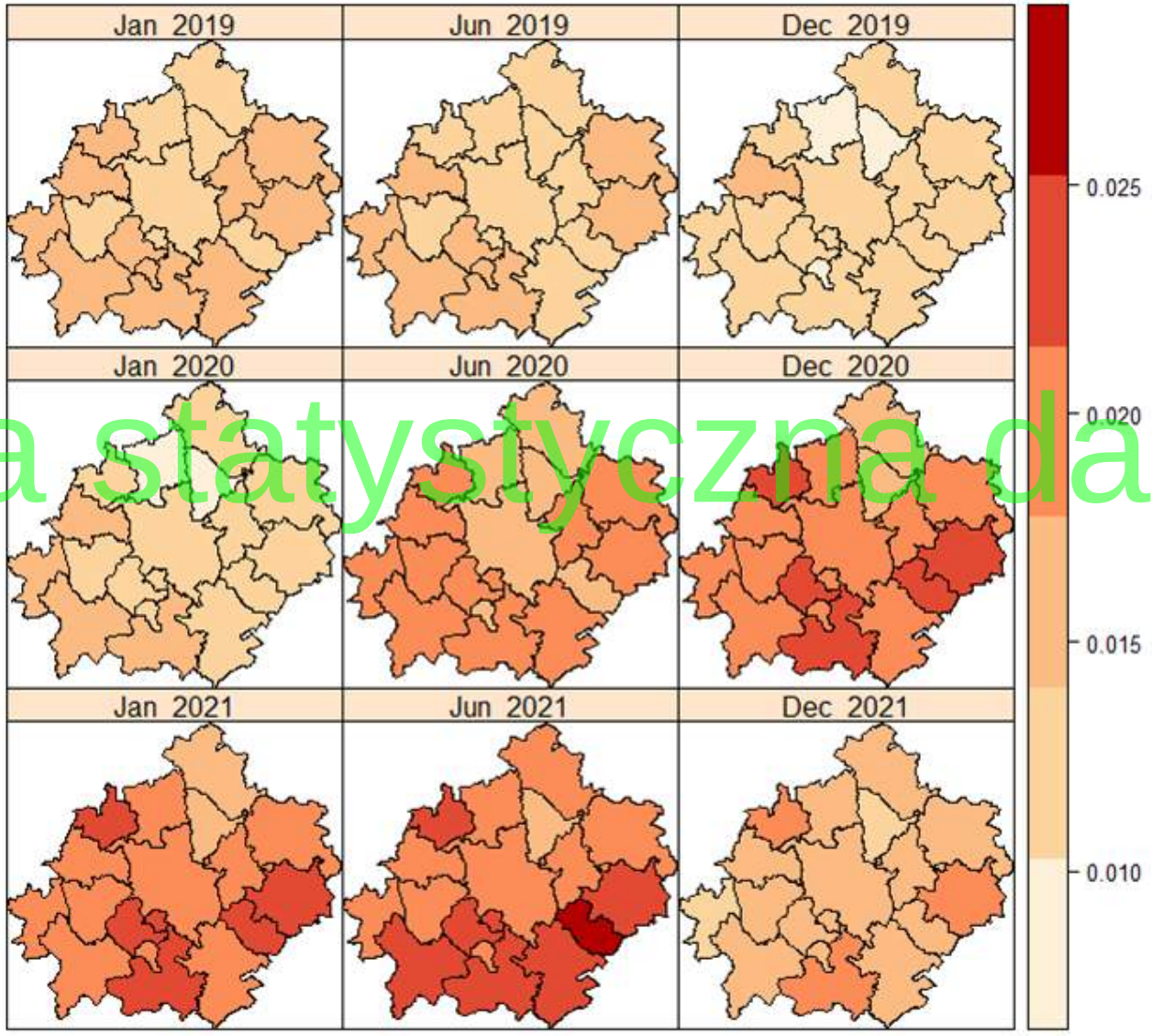


ggmap R library – kartogram gęstości

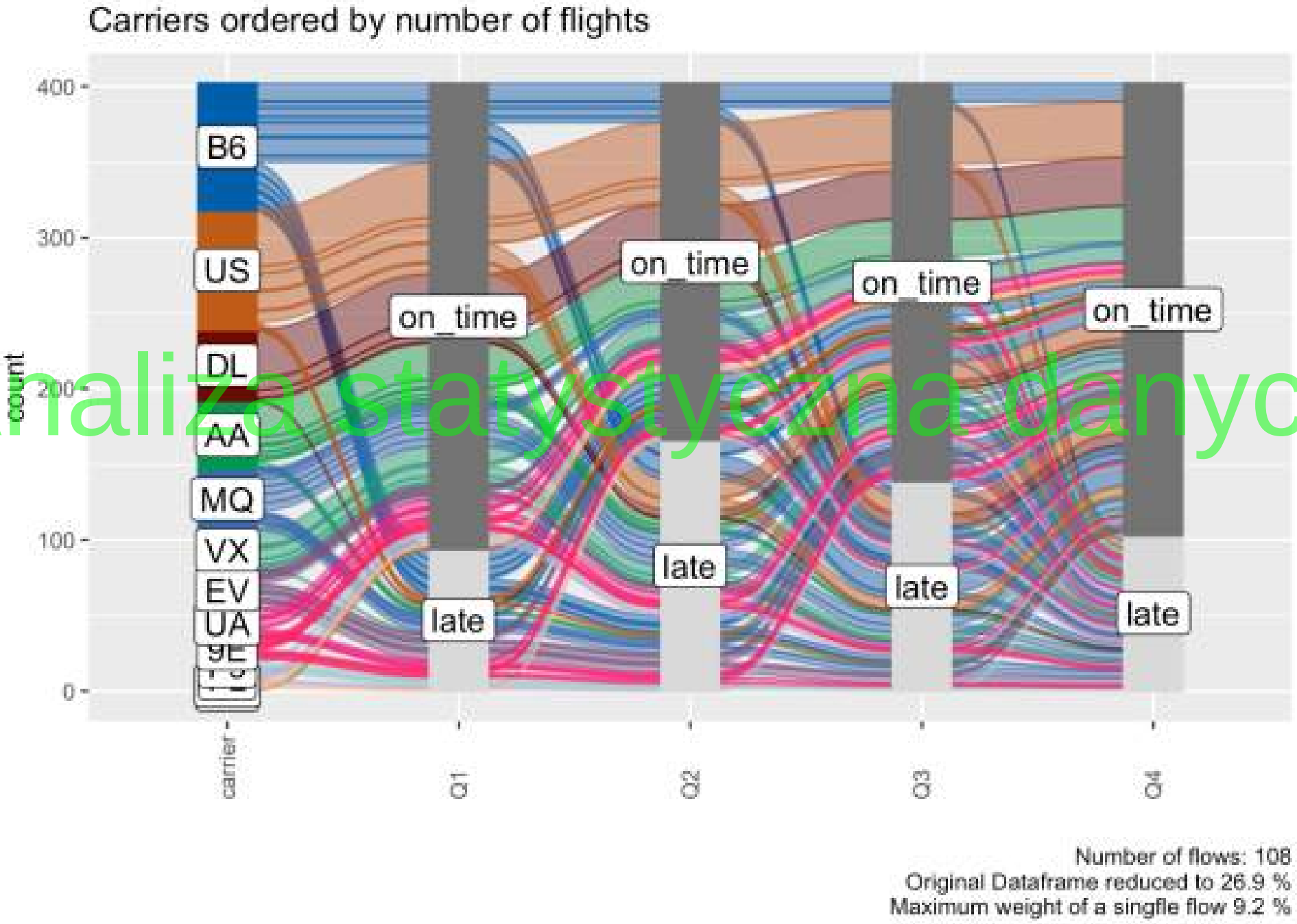


sp R library – kartogramy

Stopa bezrobocia w 17 gminach aglomeracji poznańskiej



Alluvial R library – wykres przepływów



Zalety (dobrej) grafiki:

- W porównaniu z innymi rodzajami prezentacji, dobrze zaprojektowane wykresy są bardziej efektywne i interesujące
- Relacje wizualne przedstawione przez wykresy są łatwiejsze do zapamiętania.
- Wykorzystanie wykresów oszczędza czas - istotne rezultaty mogą być widoczne “na pierwszy rzut oka”.
- Dają całościowy obraz problemu

Podstawowe narzędzia prezentacji danych

- **Tablice częstości:** to tabelaryczny zbiór danych, pokazujący liczbę wystąpień poszczególnych obserwacji.
- Jest to wygodny sposób uniknięcia konieczności wymieniania każdej obserwacji osobno.
- Rozkłady częstości mogą często zapewnić lepszy wgląd we wzorce pojawiające się w zbiorach.

Tablice częstości

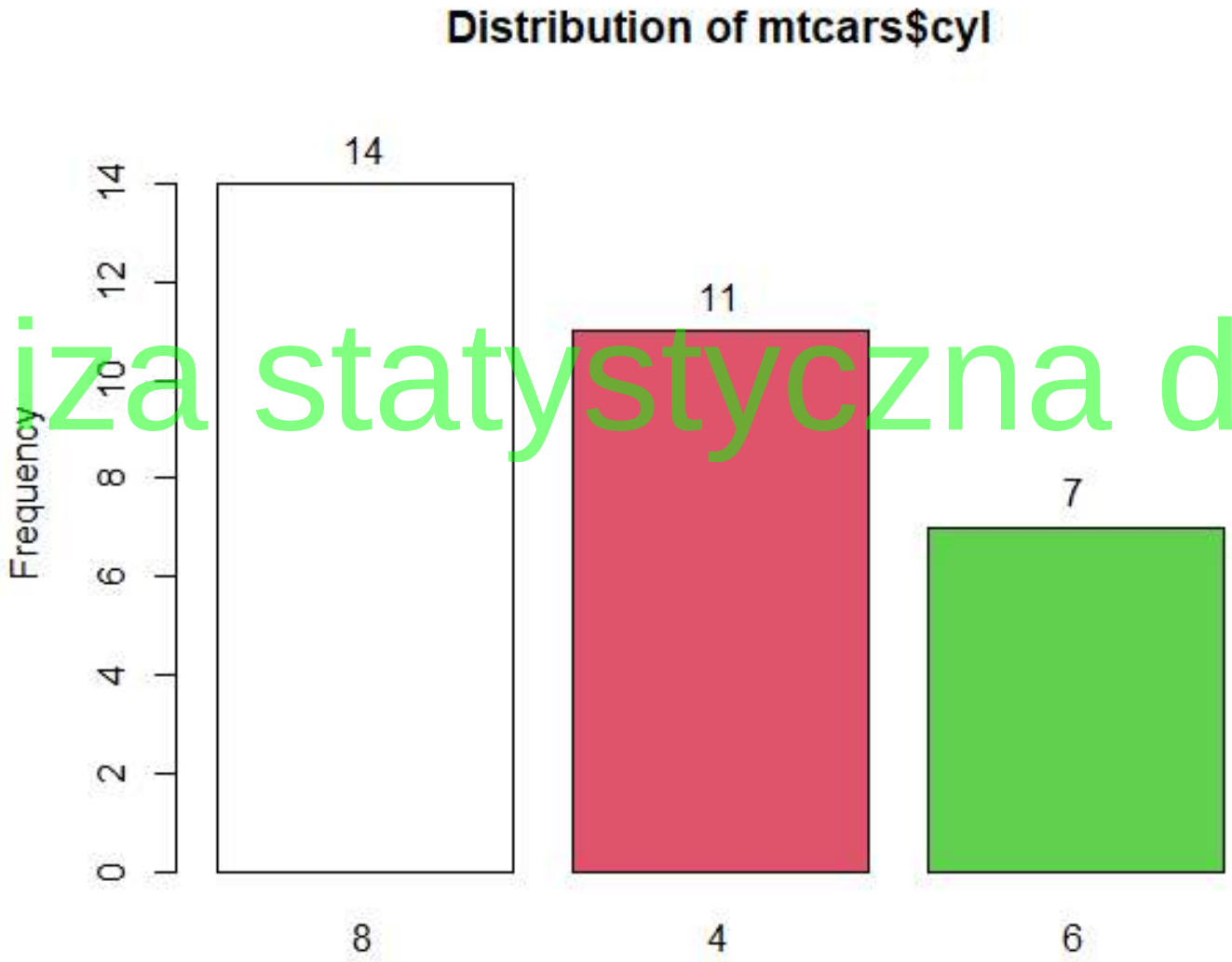
Number of Pets	Frequency
1-2	7
3-4	3
5-6	3
7-8	2

Jak mógłby wyglądać zbiór surowych danych na podstawie których powstała powyższa tablica?

Tablice częstości

```
> library(epiDisplay)
> head(mtcars)
                   mpg cyl disp  hp drat    wt  qsec vs am gear carb
Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
> tab1(mtcars$cyl, sort.group = "decreasing", cum.percent = TRUE)
mtcars$cyl :
        Frequency Percent Cum. percent
8              14    43.8         43.8
4              11    34.4         78.1
6               7    21.9        100.0
  Total        32   100.0        100.0
```

Tablice częstości



Tabele krzyżowe

- Tabele krzyżowe (dwudzielcze i wielodzielcze): rozszerzenie tabeli częstości na przypadki wielowymiarowe.
- W przypadku dwuwymiarowym dane mogą być oddzielnymi miarami zastosowanymi do tych samych klas, lub mogą to być wspólne miary.

Tabele krzyżowe

Age	Laptop	Phone	Tablet	Digital Camera
20-25	38%	29%	31%	12%
25-30	19%	15%	24%	17%
30-35	23%	19%	11%	27%
35-40	19%	12%	9%	30%
above 40	12%	17%	5%	31%

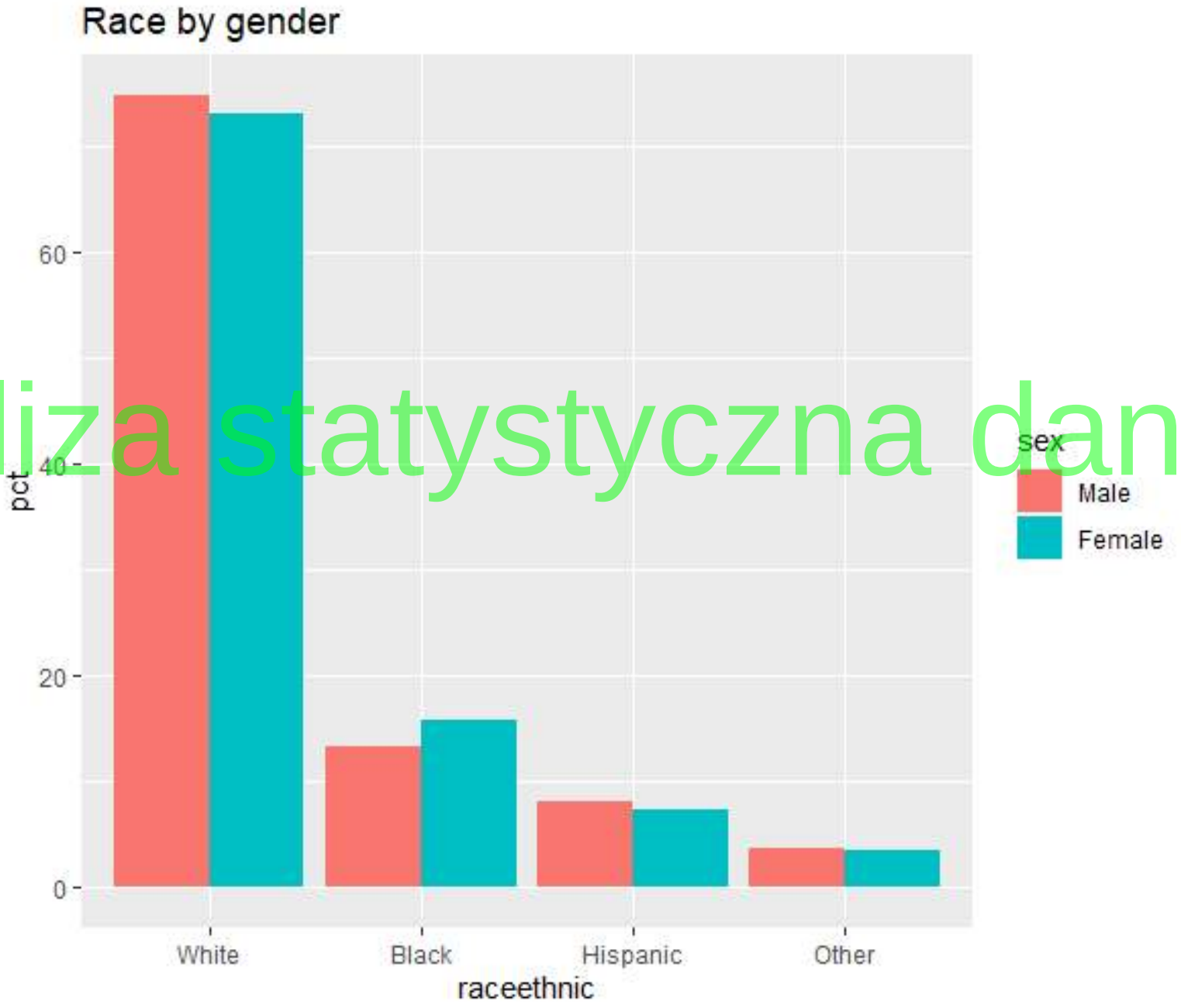
Jak mogłoby brzmieć pytanie, które zadano respondentom w poszczególnych grupach wiekowych?

Tabele krzyżowe

```
> ## ------------------------------------------------
> ## Tablice krzyżowe
> ## ------------------------------------------------
>
> # wczytuje potrzebne biblioteki
> library(pollster)
> library(dplyr)
> library(knitr)
>
> # Zbiór danych Illinois - tablica krzyżowa dla zmiennych płeć i rasa
> crosstab(df = illinois, x = sex, y = raceethnic, weight = weight) %>%
+   kable()
```


sex	White	Black	Hispanic	Other	n
Male	74.92684	13.26136	8.068462	3.743340	49108796
Female	73.15616	15.76848	7.457473	3.617883	53569718

Tabele krzyżowe



Miary środka (tendecji centralnej)

- Średnia arytmetyczna
- Mediana
- Dominanta

Analiza statystyczna danych

Średnia arytmetyczna

- Suma wszystkich liczb w zbiorze podzielona przez ich liczbę (długość zbioru)

$$m_n = \frac{a_1 + a_2 + \ldots + a_n}{n}$$

- Jaka jest średnia arytmetyczna zbioru?

  2,2,3,3,4,4,5,5,6,6

Mediana

- wartość środkowa zbioru
- mediana wskazuje, że połowa naszych wyników ma wartość poniżej, a druga połowa ma wartość powyżej wartości mediany

Mediana

- wartość środkowa zbioru
- mediana wskazuje, że połowa naszych wyników ma wartość poniżej, a druga połowa ma wartość powyżej wartości mediany
- mediana jest odporna na przypadki odstające znajdujące się w zbiorze (np. w Polsce mediana wynagrodzeń jest o ok. 1000 PLN niższa niż średnia:

Mediana

- wartość środkowa zbioru
- mediana wskazuje, że połowa naszych wyników ma wartość poniżej, a druga połowa ma wartość powyżej wartości mediany
- mediana jest odporna na przypadki odstające znajdujące się w zbiorze (np. w Polsce mediana wynagrodzeń jest o ok. 1000 PLN niższa niż średnia: 6800 vs 5800 brutto)

Dominanta

- Tzw. wartość modalna, wartość najczęstsza zbioru
- Jeżeli dwie wartości pojawiają się z równą i największą częstością, obie są dominantami
- W przypadku wynagrodzeń w 2021 roku dominantą była kwota ~2900 PLN brutto

Miary rozproszenia (rozrzutu)

- Rozstęp (różnica między największą a najmniejszą wartością w zbiorze)
- Wariancja
- Odchylenie standardowe

Wariancja

- Jest obliczana przez zsumowanie średniej kwadratów odchyleń od średniej zbioru:

$$\sigma^2 = \frac{\Sigma\left(X - \bar{X}\right)^2}{n-1}$$

- Wariancja mówi o stopniu rozrzutu danych. Im bardziej rozłożone są dane, tym większa jest wariancja w stosunku do średniej.

Odchylenie standardowe

- Podobnie jak wariancja mierzy rozproszenie w zbiorze w odniesieniu do średniej
- Jest to pierwiastek kwadratowy z wariancji:

$$\sigma = \sqrt{\frac{\sum (X - \bar{X})^2}{n-1}}$$

- Jest to również najczęściej pojawiająca się miara opisowa (oprócz średniej arytmetycznej)

Odchylenie standardowe

Analiza statystyczna danych

Miary koncentracji i asymetrii

- Są sytuacje, w których badanie średniego poziomu zmiennej i rozproszenia jej wartości nie wskazuje na istnienie różnic między badanymi zbiorowościami.
- Pomocne wtedy mogą być dodatkowe wskaźniki pozwalające określić kształt rozkładu:

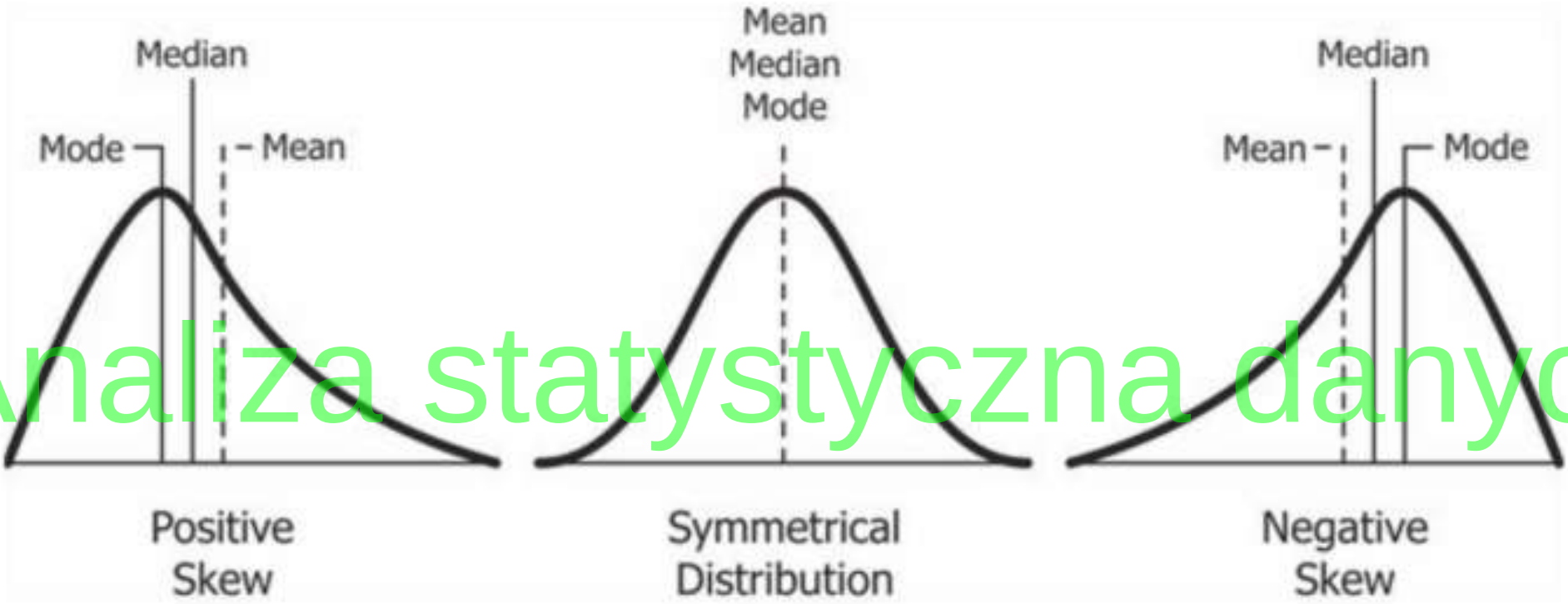
  - Skośność

  - Kurtoza

Skośność

- Jest miarą asymetrii rozkładu. Wartość ta może być dodatnia lub ujemna.
- Negatywna skośność wskazuje, że ogon znajduje się po lewej stronie rozkładu
- Pozytywna skośność wskazuje, że ogon znajduje się po prawej stronie rozkładu
- Wartość zero oznacza, że w rozkładzie nie ma skośności, co oznacza, że rozkład jest idealnie symetryczny.

Skośność



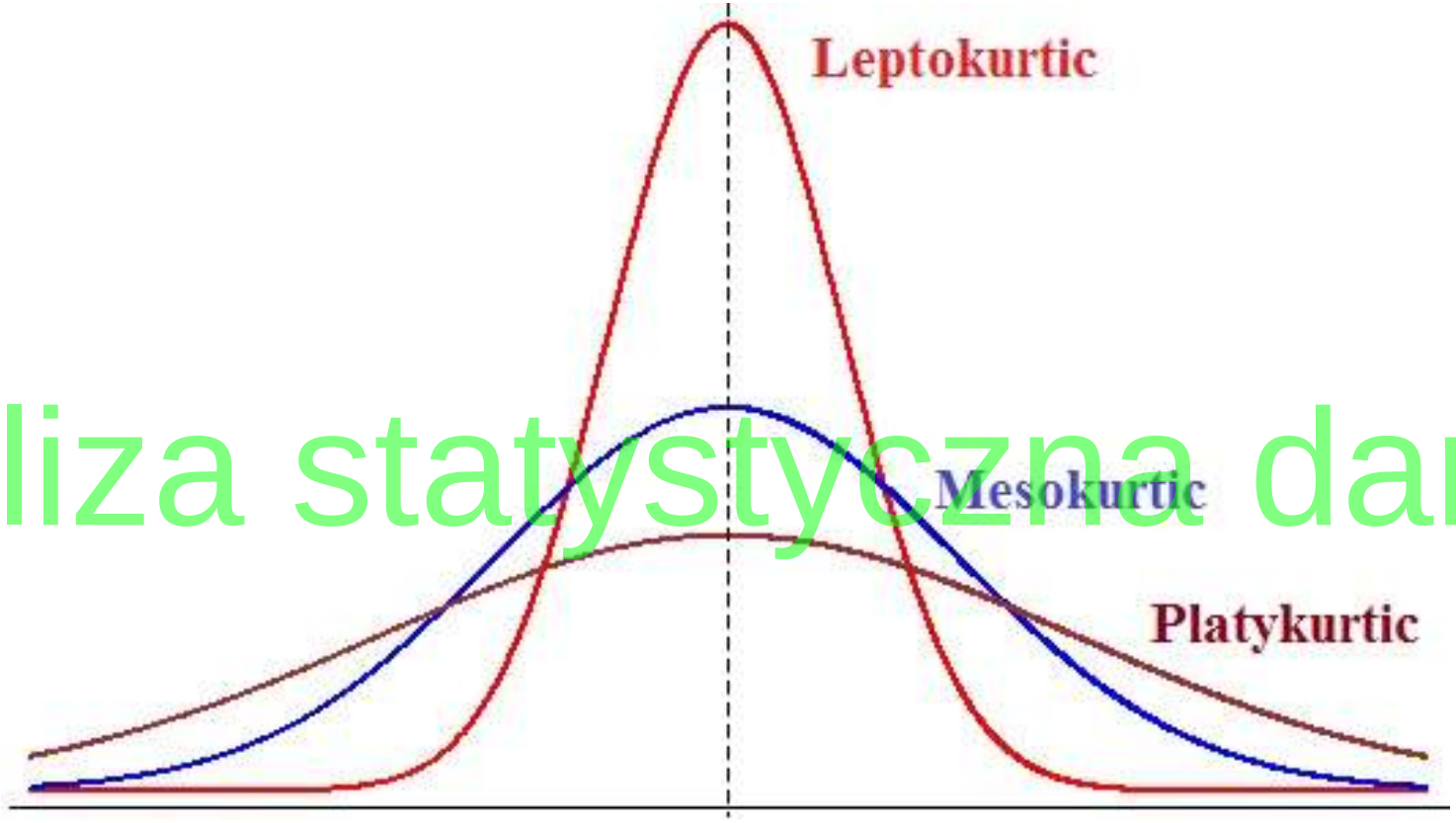
Współczynnik skośności:

- Symetryczne: Wartości od -0,5 do 0,5
- Skośność średnia: Wartości pomiędzy -1 i -0,5 lub pomiędzy 0,5 i 1
- Skośność wysoka: Wartości mniejsze niż -1 lub większe niż 1

Kurtoza

- Najpopularniejszą miarą skupienia obserwacji wokół średniej jest kurtoza
- Im wyższa jest jej wartość, tym bardziej wysmukła jest krzywa liczebności, a zatem większa koncentracja cechy wokół średniej

Kurtoza

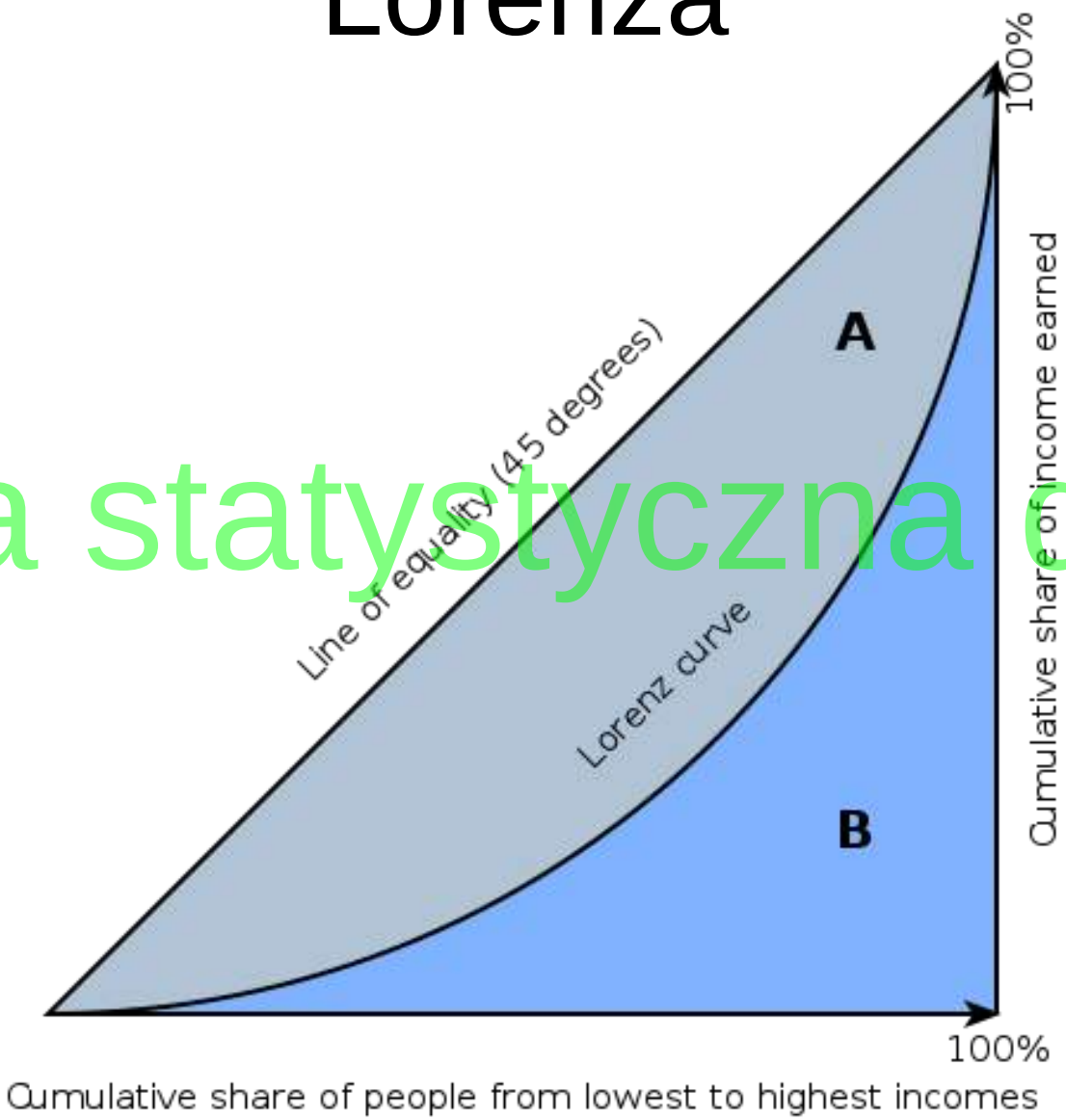


Kurt = 3 - mezokurtyczny
Kurt > 3 - leptokurtyczny
Kurt < 3 - platykurtyczny

Współczynnik (indeks) Giniego

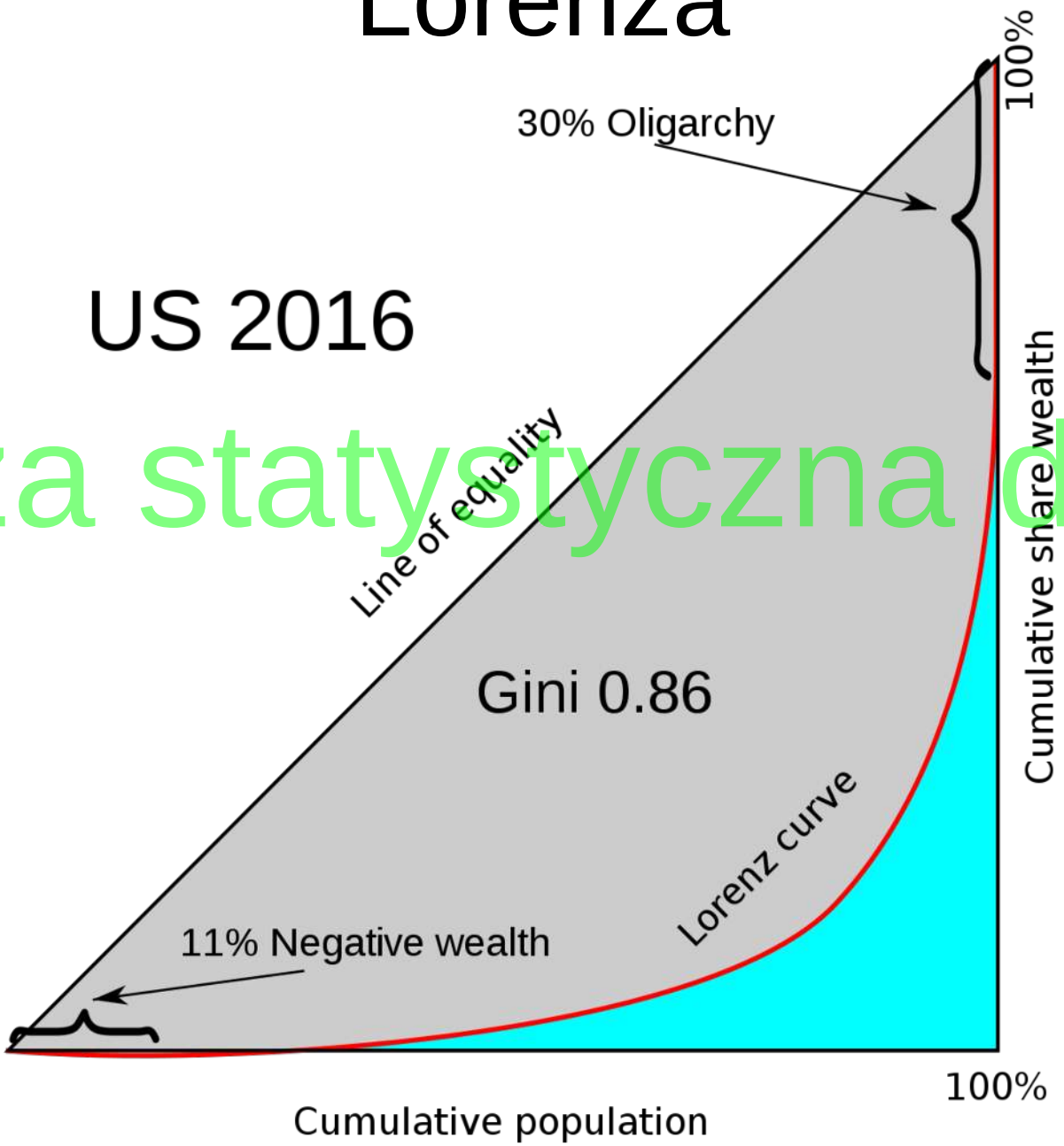
- Miara koncentracji rozkładu zmiennej
- Zaprojektowana, aby mierzyć nierówności dochodowe (ekonomia)
- Może mieścić się w przedziale od 0 (całkowita równość) do 1 (całkowita nierówność)
- Czasami jest wyrażany w procentach

Współczynnik Giniego i krzywa Lorenza

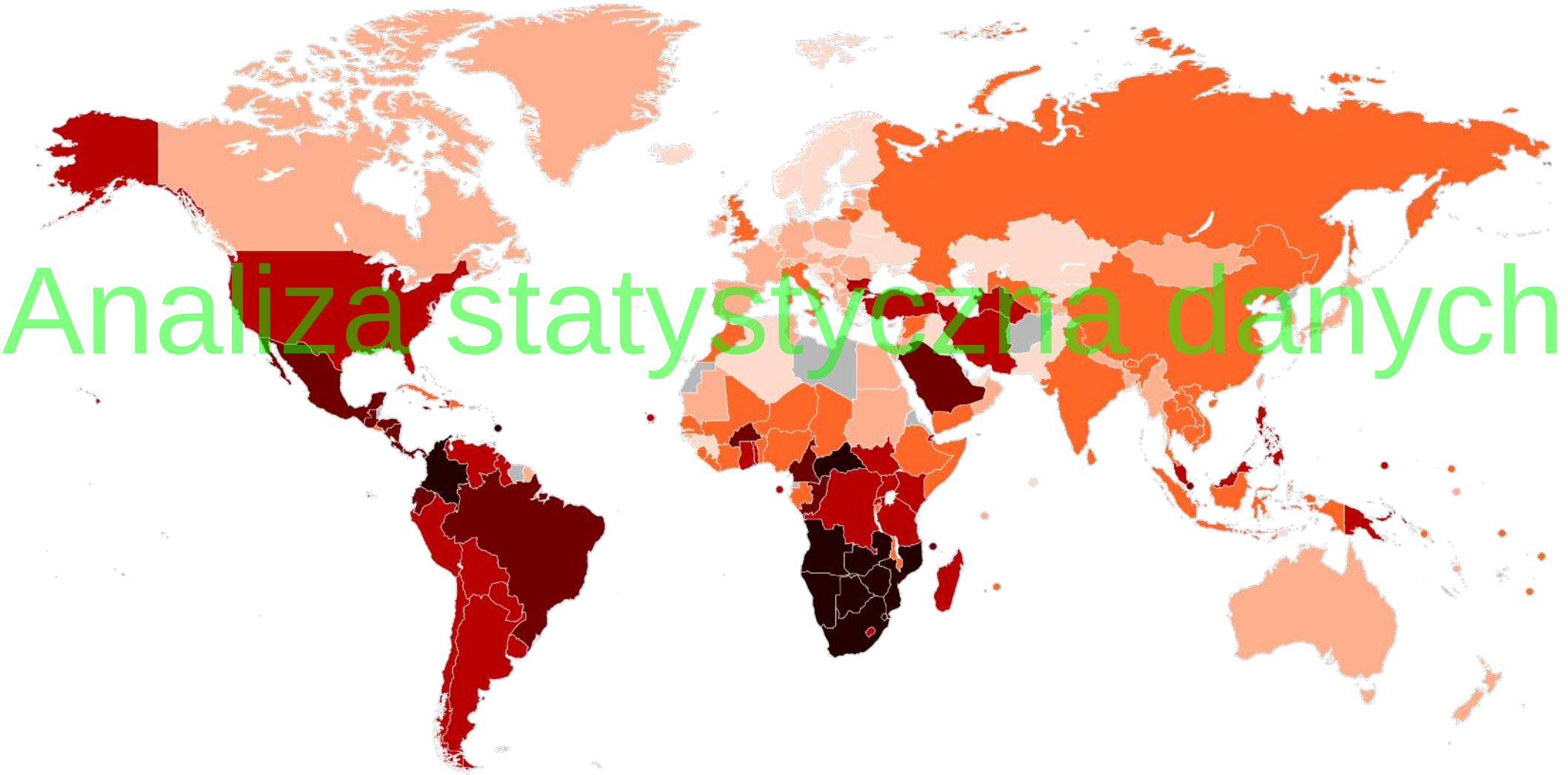


Gini = A/(A + B)

Współczynnik Giniego i krzywa Lorenza



Współczynnik Giniego



Miary zależności i bliskości

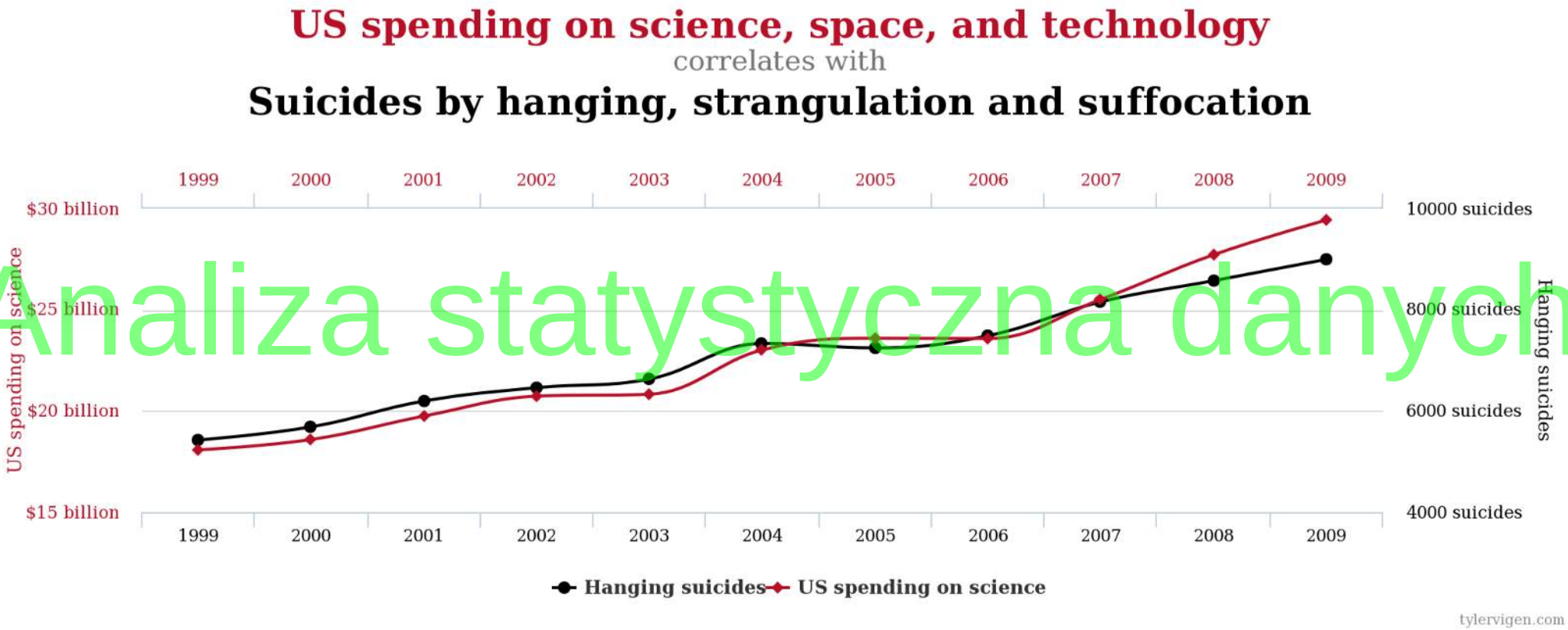
- Współczynniki korelacji Pearsona
- Test chi-kwadrat
- Współczynnik korelacji rang Speramana
- Autokorelacja
- Korelacja przestrzenna

Korelacja (ogólnie)

- Mierzy (potencjalną) zależność pomiędzy dwoma zmiennymi: jak jedna zmienna ewoluuje wraz ze zmianą innej
- To, że zmiene są skorelowane **nie oznacza**, że jedna wpływa/wywołuje drugą

Korelacja (ogólnie)

- Mierzy (potencjalną) zależność pomiędzy dwoma zmiennymi: jak jedna zmienna ewoluuje wraz ze zmianą innej
- To, że zmiene są skorelowane **nie oznacza**, że jedna wpływa/wywołuje drugą
- Pozwala stwierdzić, które zmienne ewoluują w tym samym kierunku, które w przeciwnym, a które są niezależne.
- Metody parametryczne i nieparametryczne obliczania korelacji



Współczynnik korelacji Pearsona

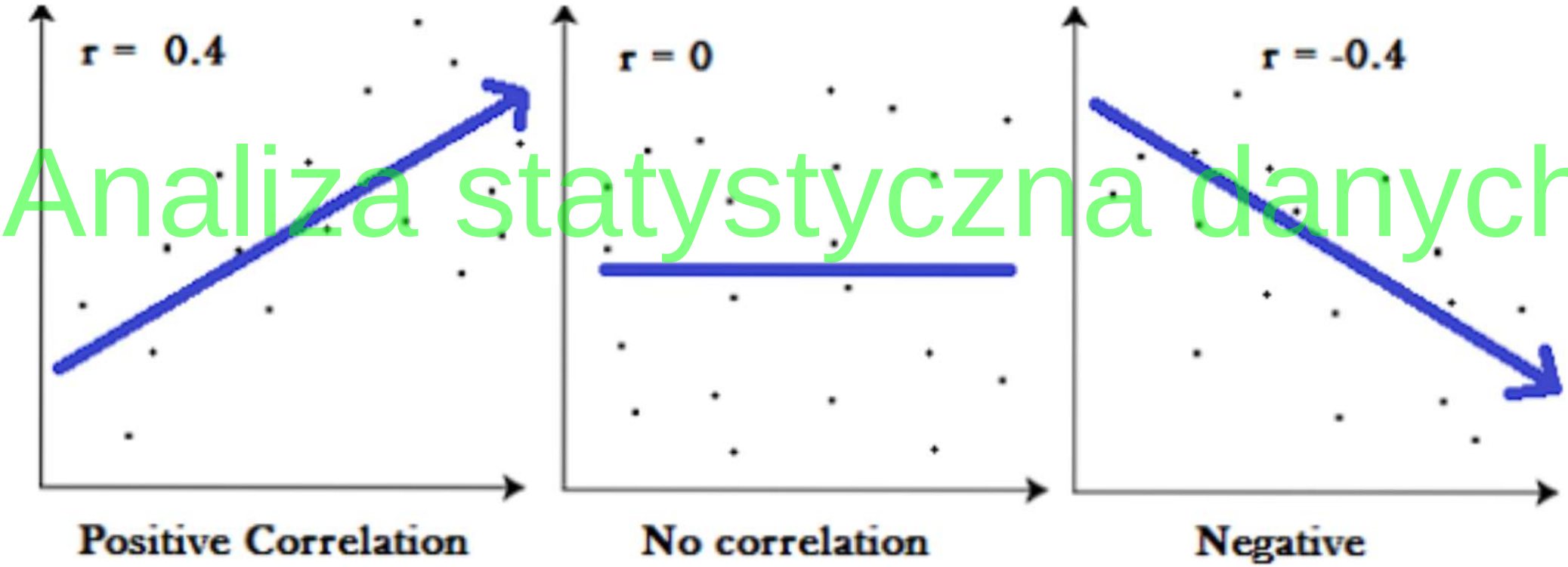
- Tzw. *r Pearsona*
- Mierzy korelację liniową pomiędzy dwoma seriami danych
- Ma zastosowanie tylko w przypadku zmiennych numerycznych
- Przyjmuje zawsze wartości z przedziału 0-1

Współczynnik korelacji Pearsona

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2]\,[n\sum y^2 - (\sum y)^2]}}$$

- r = 0 – brak zależności
- r <= 0.3 – korelacja słaba
- r = 0.4 - 0.6 – korelacja umiarkowana
- r = 0.7 - 0.9 – korelacja silna
- r = 1 – korelacja idealna

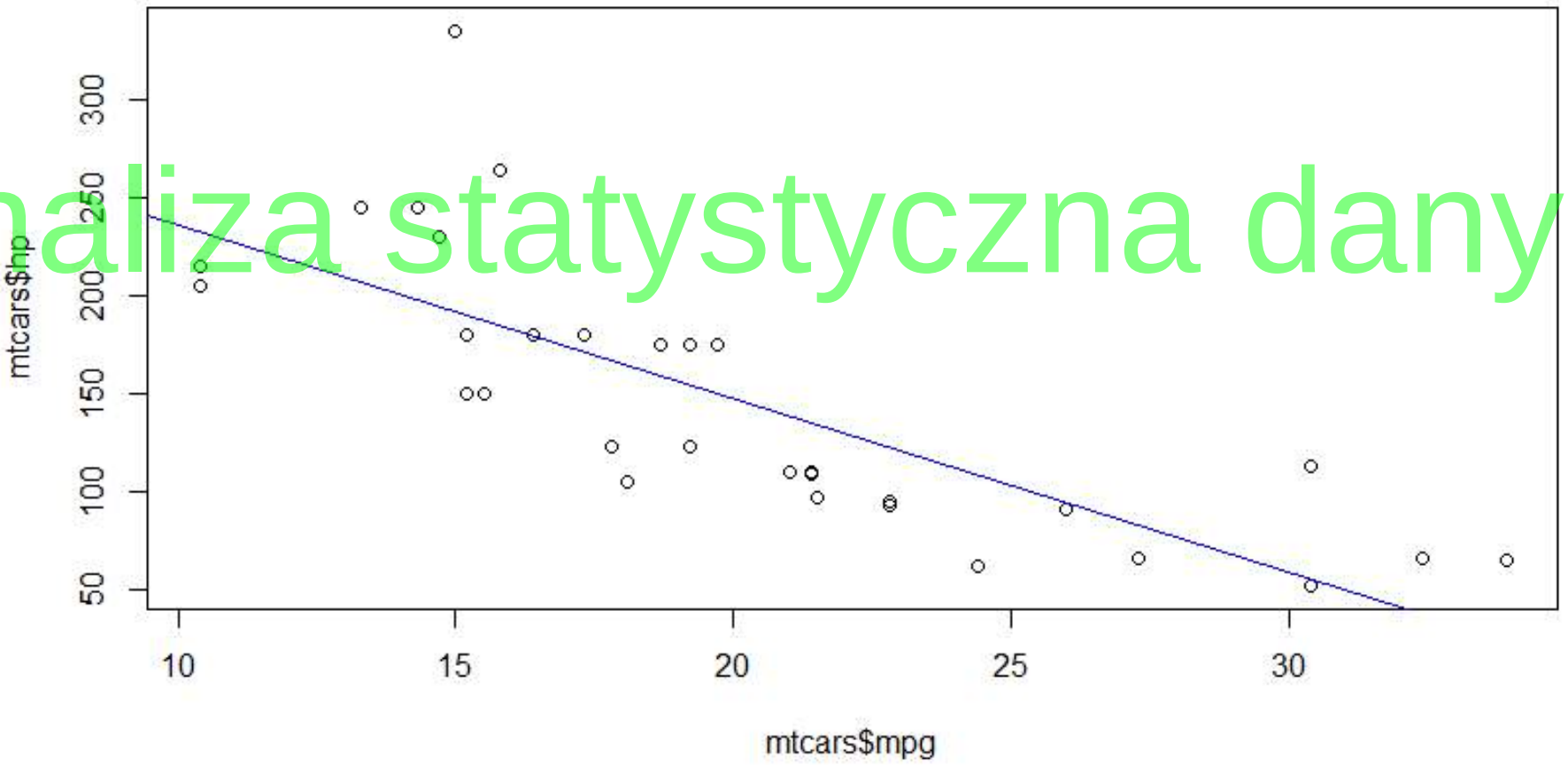
Współczynnik korelacji Pearsona



Przykład – korelacja Pearsona

```
> head(mtcars)
                   mpg cyl disp  hp drat    wt  qsec vs am gear carb
Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
> # Pearson correlation between 2 variables
> cor(mtcars$hp, mtcars$mpg)
[1] -0.7761684
```

moc auta a mile na galon



Test chi-kwadrat

- Jest metodą statystyczną, która służy do określenia, czy dwie zmienne kategoryczne są ze sobą powiązane (tablice częstości)
- Jak wszystkie testy statystyczne, tak i ten test zakłada hipotezę zerową i hipotezę alternatywną:

  H0: Zmienne są niezależne.

  H1: Zmienne są ze sobą powiązane (skorelowane).

Test chi-kwadrat

- Jest metodą statystyczną, która służy do określenia, czy dwie zmienne kategoryczne są ze sobą powiązane (tablice częstości)
- Jak wszystkie testy statystyczne, tak i ten test zakłada hipotezę zerową i hipotezę alternatywną:

  H0: Zmienne są niezależne.

  H1: Zmienne są ze sobą powiązane (skorelowane).
- Odrzucamy hipotezę zerową, jeśli tzw. ***wartość p***, która pojawia się w wyniku jest mniejsza od ustalonego wcześniej poziomu istotności (zazwyczaj 0.05)
- Brak informacji o sile związku

Współczynnik korelacji rang Spearmana (*rho*)

- Podobnie jak współczynnik korelacji Pearsona, tak i tzw. *rho* Spearmana pozwala określić siłę związku pomiędzy zmiennymi
- Jest to metoda nieparametryczna, którą można stosować dla danych porządkowych
- Wartości *rho* Speramana intepretujemy podobnie jak w przypadku r Pearsona
- Dodatkowo otrzymujemy ***wartość p***

rho Spearmana

```
> ## 3) Współczynnik korelacji rang Spearmana
>
> cor.test(x=as.numeric(Cars93$AirBags), y=as.numeric(Cars93$Type), method = 'spearman')

        Spearman's rank correlation rho

data:  as.numeric(Cars93$AirBags) and as.numeric(Cars93$Type)
S = 96894, p-value = 0.007157
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.2771473
```

Autokorelacja

- Jest to podobieństwo pomiędzy poszczególnymi obserwacjami zmiennej losowej w jej opóźnieniach czasowych.
- Gdy autokorelacja zmiennej jest wysoka, łatwe staje się przewidywanie jej przyszłych wartości poprzez odniesienie do wartości przeszłych.
- Wyniki i interpretacja → współczynnik korelacji liniowej Pearsona

Korelacja i zależność przyczynowo skutkowa

https://www.youtube.com/watch?v=Nre4cjz3U4A&ab_channel=KhanAcademyPoPolsku

Analiza statystyczna danych

Autokorelacja przestrzenna

- Pierwsze Prawo Geografii:

  "everything is related to everything else, but near things are more related than distant things."

- Jest to fundamentalny koncept analizy przestrzennej (nie tylko) w geografii

Waldo Tobler

Autokorelacja przestrzenna

- Pomaga zrozumieć, w jakim stopniu jeden obiekt jest podobny do innych pobliskich obiektów
- W przypadku gdy sąsiadujące ze sobą w przestrzeni obiekty mają podobne wartości danych, mamy dodatnią (pozytywną) autokorelację przestrzenną
- Wskaźnik “**I Morana”** jest najczęściej stosowaną miarą autokorelacji przestrzennej

Autokorelacja przestrzenna

- I Morana przyjmuje wartości od -1 do 1
- Interpretacja jest nieco inna, niż w przypadku klasycznych miar korelacji:

  -1 to idealne skupienie niepodobnych wartości (doskonałe rozproszenie)

  0 to idealna losowość

  1 to idelana pozytywna autokorelacja przestrzenna (idealna klasteryzacja)

Autokorelacja przestrzenna

Idealna pozytywna korelacja przestrzenna; Moran I = 1

Autokorelacja przestrzenna

Analiza statystyczna danych

Losowe rozproszenie, Moran I = 0

Autokorelacja przestrzenna

Idealna dyspersja, Moran I = -1

Autokorelacja przestrzenna

- Moran I procedura:

  1) dane geoprzestrzenne
  (.shp, .geopackage, .PostGIS, etc.)

  2) zdefiniowanie “sąsiadów” za pomocą tzw. matrycy wag przestrzennych

Autokorelacja przestrzenna

- Moran I procedura:

  1) dane geoprzestrzenne (.shp, .geopackage, .PostGIS, etc.)

  2) zdefiniowanie “sąsiadów” za pomocą tzw. matrycy wag przestrzennych

  3) przydzielenie wag poszczególnym sąsiadom

  4) obliczenie wartości statystyki testowej i weryfikacja hipotezy 0

Autokorelacja przestrzenna

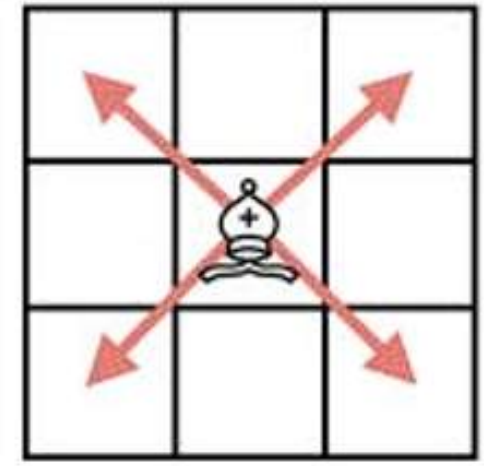
Macierze wag przestrzennych:

- Macierze wag przestrzennych bazują na relacjach przestrzennych obiektów.
- Istnieje wiele rodzajów macierzy wag przestrzennych (np. oparte na dystansie, na gęstości sieci drogowej, etc.)

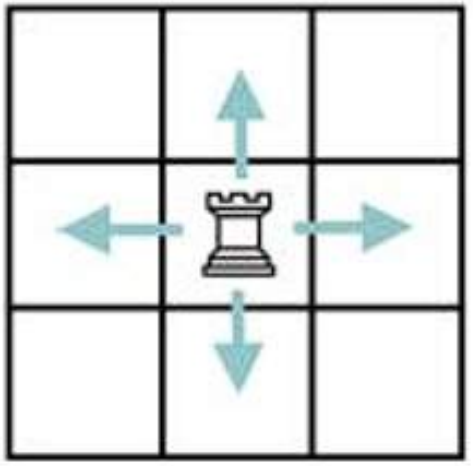
Autokorelacja przestrzenna

Macierze wag przestrzennych:

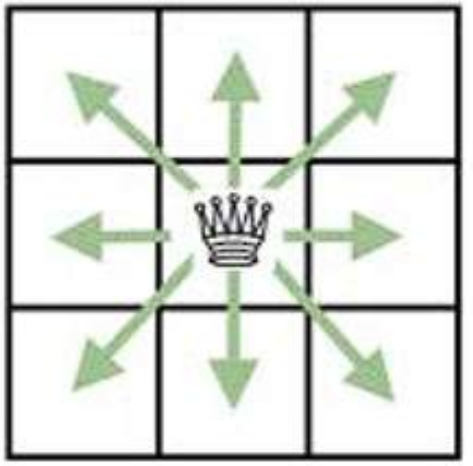
- Macierze wag przestrzennych bazują na relacjach przestrzennych obiektów.
- Istnieje wiele rodzajów macierzy wag przestrzennych (np. oparte na dystansie, na gęstości sieci drogowej, etc.)
- Podstawowe typy sąsiedztwa:



Sąsiedztwo typu Bishop



Sąsiedztwo typu Rook



Sąsiedztwo typu Quenn

Autokorelacja przestrzenna

$$I = \frac{N}{\sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j w_{ij}(X_i - \bar{X})(X_j - \bar{X})}{\sum_i (X_i - \bar{X})^2}$$

N – liczba obserwacji
$\mathbf{X_i}$ – wartość obserwacji w obiekcie i
$\mathbf{X_j}$ – wartość obserwacji w obiekcie j
$\mathbf{w_{ij}}$ – matryca wag przestrzennych dla połączeń obiektów i oraz j

Autokorelacja przestrzenna

- Przykład

Analiza statystyczna danych

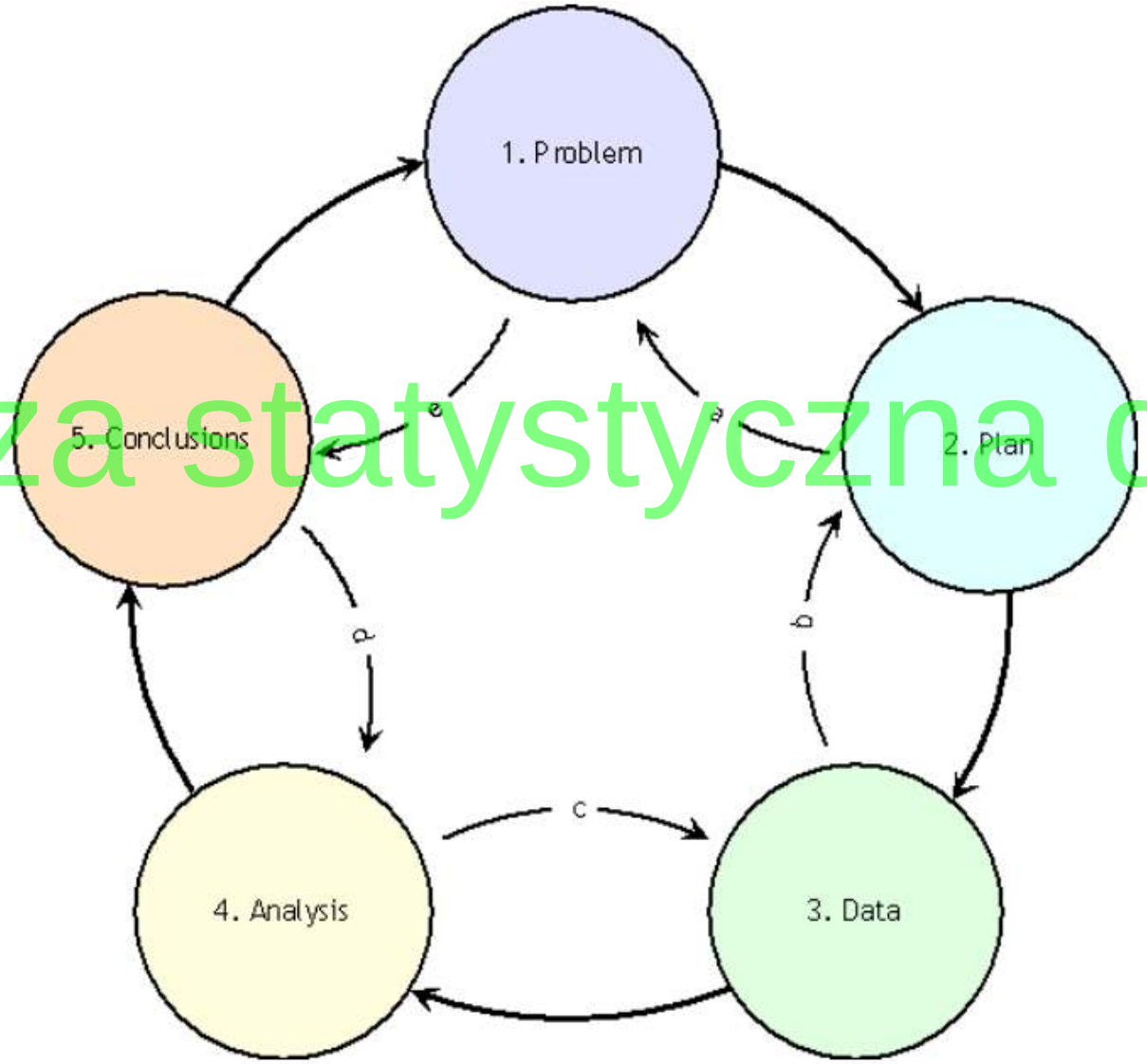
Metoda statystyczna

Analiza statystyczna danych

Metoda statystyczna

- Analiza statystyczna nie jest czysto technicznym ćwiczeniem
- Powinna być realizowana w szerokim kontekście zarówno metodologicznym jak i teoretycznym
- Wymaga więc zarówno wiedzy tematycznej jak i technicznej
- Powinna opierać się na modelu: PPDAK

Metoda statystyczna (PPDAK)



Metoda statystyczna

Problem:

- Zrozumienie i zdefiniowanie problemu jest istotną częścią całego procesu analitycznego (*o czym są te badania?*)
- Problem powinien wyczerpywać zakres badania i uwzględniać zależności pomiędzy zmiennymi

Metoda statystyczna

Problem:

- Zrozumienie i zdefiniowanie problemu jest istotną częścią całego procesu analitycznego (*o czym są te badania?*)
- Problem powinien wyczerpywać zakres badania i uwzględniać zależności pomiędzy zmiennymi
- Im więcej interakcji i zmiennych, tym bardziej skomplikowane wnioskowanie
- Sformułowanie problemu powinna porzedzać faza "desk research"

Metoda statystyczna

Problem – przykłady (obszar edukacji):

1) Odwrotna dyskryminacja w zatrudnieniu po studiach

2) Czy koncepcje edukacji wielokulturowej powinny być wdrażane w większym stopniu?

3) Nadużywanie narkotyków i alkoholu na kampusach uniwersyteckich

4) Czy osoby z ADHD i Autyzmem powinny być oddzielone od pozostałych studentów

Metoda statystyczna

Plan:

- Następnym etapem jest sformułowanie podejścia, które ma największe szanse na rozwiązanie problemu i uzyskanie odpowiedzi
- W przypadku projektów, które mają charakter bardziej eksperymentalny wymaga opracowania szczegółowych kroków
- Produktem etapu jest szczegółowy plan badań zawierający czas, zasoby, zaangażowane osoby, sprzęt, etc.

Metoda statystyczna

Dane:

- Dane pierwotne, dane wtórne, mix
- Dylematy: jakość danych, koszt, uzgodnienia licencyjne, dostępność, kompletność, format, szczegółowość

Analiza statystyczna danych

Metoda statystyczna

Dane:

- Dane pierwotne, dane wtórne, mix
- Dylematy: jakość danych, koszt, uzgodnienia licencyjne, dostępność, kompletność, format, szczegółowość
- Jeżeli dane są nieodpowiednie/niemożliwe do zdobycia --> reformulacja problemu badawczego
- Nie ma idealnego zbioru danych

Metoda statystyczna

Analiza:

- Jest zazwyczaj czynnością wieloetapową
- Zaczyna się od przeglądu i przekształcania danych, aby otrzymać spójny zbiór

Analiza statystyczna danych

Metoda statystyczna

Analiza:

- Jest zazwyczaj czynnością wieloetapową
- Zaczyna się od przeglądu i przekształcania danych, aby otrzymać spójny zbiór
- Kolejne kroki to np.: analiza opisowa, eksploracja danych, modelowanie statystyczne
- Należy unikać stosowania pojedynczej techniki analitycznej

Metoda statystyczna

Analiza:

- Jest zazwyczaj czynnością wieloetapową
- Zaczyna się od przeglądu i przekształcania danych, aby otrzymać spójny zbiór
- Kolejne kroki to np.: analiza opisowa, eksploracja danych, modelowanie statystyczne
- Należy unikać stosowania pojedynczej techniki analitycznej

"It is as well to remember the following truths about models: all models are wrong; some models are better than others; the correct model can never be known with certainty; and the simpler a model the better it is"!

George Box

Metoda statystyczna

Konkluzje

- Etap ma na celu opracowanie wniosków w "języku problemu" w celu ich upowszechnienia
- Powinny zawierać ścisłe podsumowanie badań i prezentację graficzną
- Nie powinny zawierać szczegółowych detali technicznych
- Wskazują mocne i słabe strony badań

Rodzaje badań statystycznych

https://www.youtube.com/watch?v=H5pLC05ycoo&ab_channel=KhanAcademyPoPolsku

Analiza statystyczna danych

Nadużycia, nadinterpretacje, błędy

Analiza statystyczna danych

Nadużycia, nadinterpretacje, błędy

1) Nieadekwatne lub niereprezentatywne dane
2) Myląca wizualizacja rezultatów
3) Błędy we wnioskowaniu
4) Celowe fałszowanie danych

1) Nieadekwatne lub niereprezentatywne dane

- Problem doboru i liczebności próby:

  - zbyt mała liczebność próby

  - niedoreprezentowanie/ nadreprezentowanie warstw

  - wykluczenie pewnych grup społecznych

1) Nieadekwatne lub niereprezentatywne dane

- Problem doboru i liczebności próby:

  - zbyt mała liczebność próby

  - niedoreprezentowanie/ nadreprezentowanie warstw

  - wykluczenie pewnych grup społecznych

  - niepoprawne wykorzystanie technik badawczych

  - efekty czasowe i przestrzenne

  - efekt społecznych oczekiwań (przeszacowanie lub niedoszacowanie)

2) Myląca wizualizacja rezultatów

- Brak skali i opisu osi
- Brak informacji o początku skali (0 lub inna wartość)
- Wybiórcze punkty danych (*cherry picking*)
- Nieodpowiednia forma wizualizacji
- Brak porównywalności pomiędzy wykresami

2) Myląca wizualizacja rezultatów

NEWS Magazine

News Front Page

Africa
Americas
Asia-Pacific
Europe
Middle East
South Asia
UK
England
Northern Ireland
Scotland
Wales
UK Politics
Education
Magazine
Business
Health
Science & Environment
Technology
Entertainment
Also in the news

Video and Audio

Programmes
Have Your Say
In Pictures
Country Profiles
Special Reports

Related BBC sites
Sport
Weather
On This Day
Editors' Blog
BBC World Service

Page last updated at 01:40 GMT, Friday, 29 January 2010

E-mail this to a friend Printable version

The bumps in a falling teenage pregnancy rate

GO FIGURE
Different ways of seeing stats

When teenage pregnancy rates in one community fell drastically, it looked like a policy of sex education had paid off. But, as Michael Blastland explains in his regular column, the bumps were still there... you just had to know where to look.

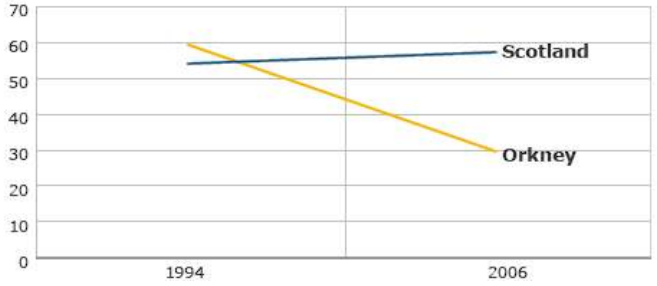
Can stories be bad for you? Let me tell you a true one.

Once upon a time, Orkney had one of the highest teenage pregnancy rates in Scotland. Scotland itself is high in the international league.

So health workers in Orkney tried something new. They began talking to young people about sex in terms of relationships, not only mechanics. They also made condoms easily available because in a small community the shopkeeper might just be your auntie.

Then came data showing that Orkney's teenage pregnancy rate had dramatically halved.

Teenage pregnancies
Per thousand women



COMPLETE MICHAEL BLASTLAND ARCHIVE
Measuring niceness GO

In today's Magazine

Big beasts
How elephants helped to shape human history, by David Cannadine

Change a-coming
Justin Webb on America's love affair with progress

Audience of one
Would you watch a play all on your own?

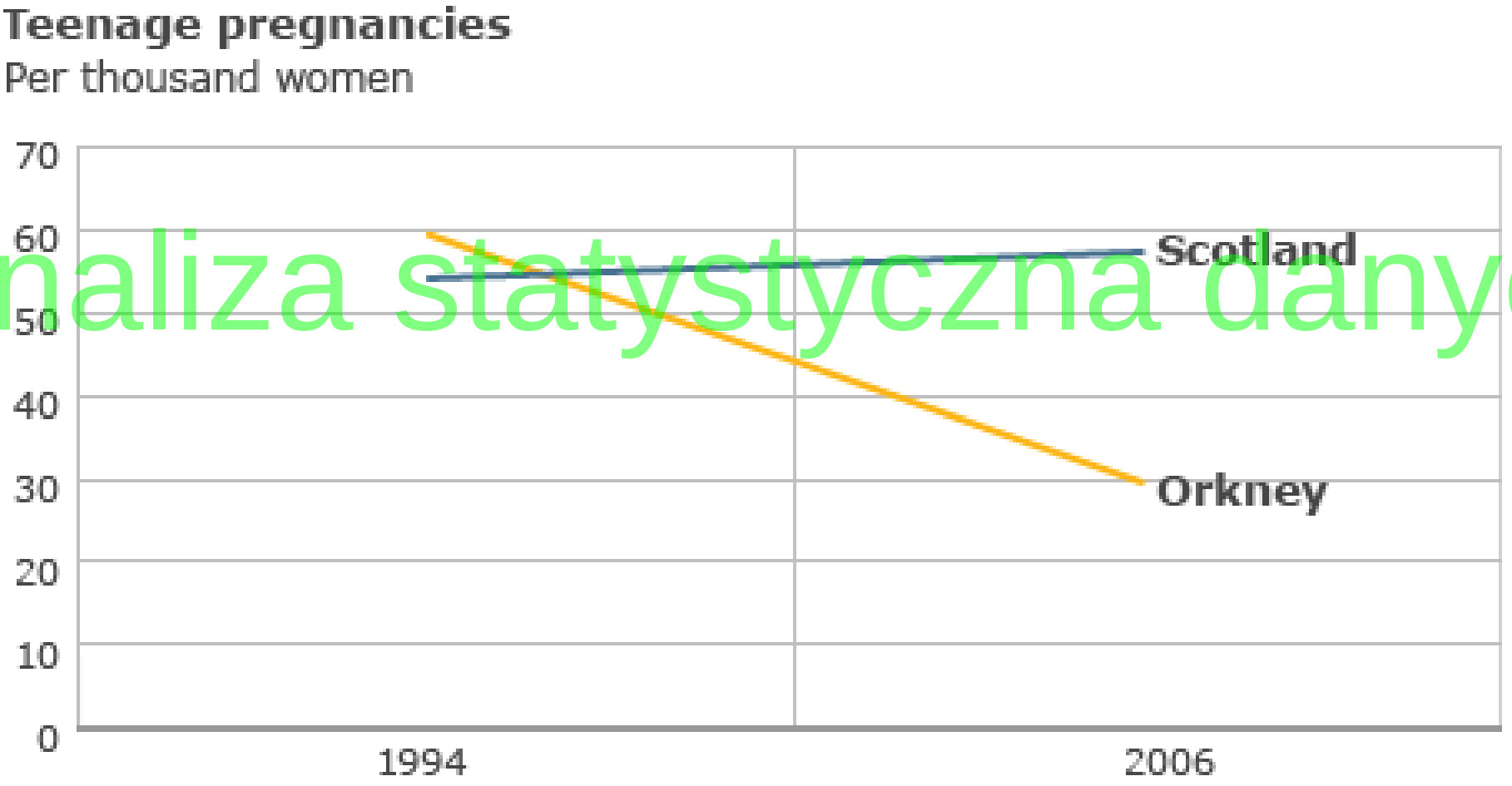
7 days quiz
What now for Paul the eight-limbed oracle?

Magazine regulars

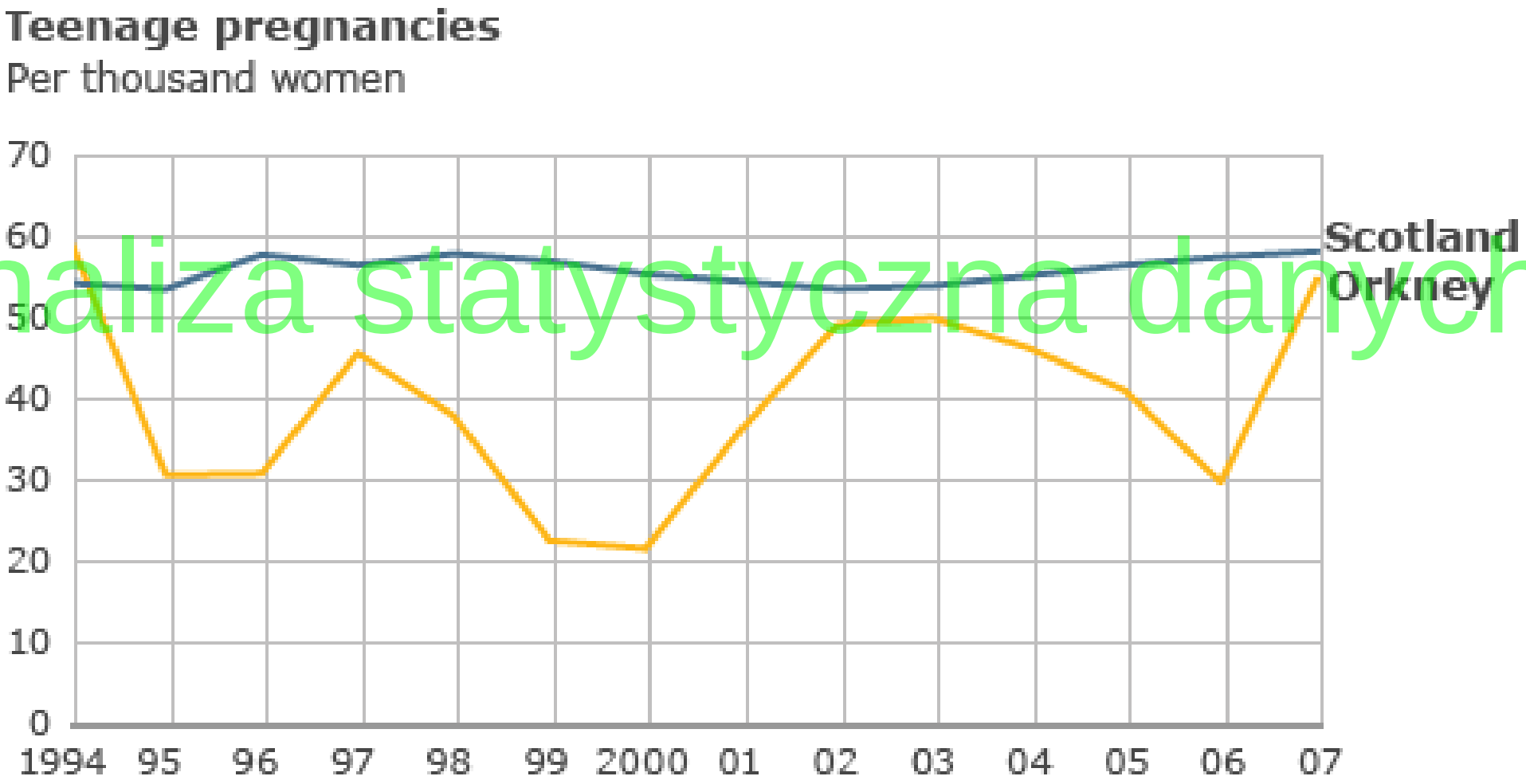
Tweetbook
Say goodbye to worktime boredom. Follow us on Facebook or Twitter

Magazine Monitor
Paper Monitor, Your Letters, Quote of the Day, Caption Competition and more

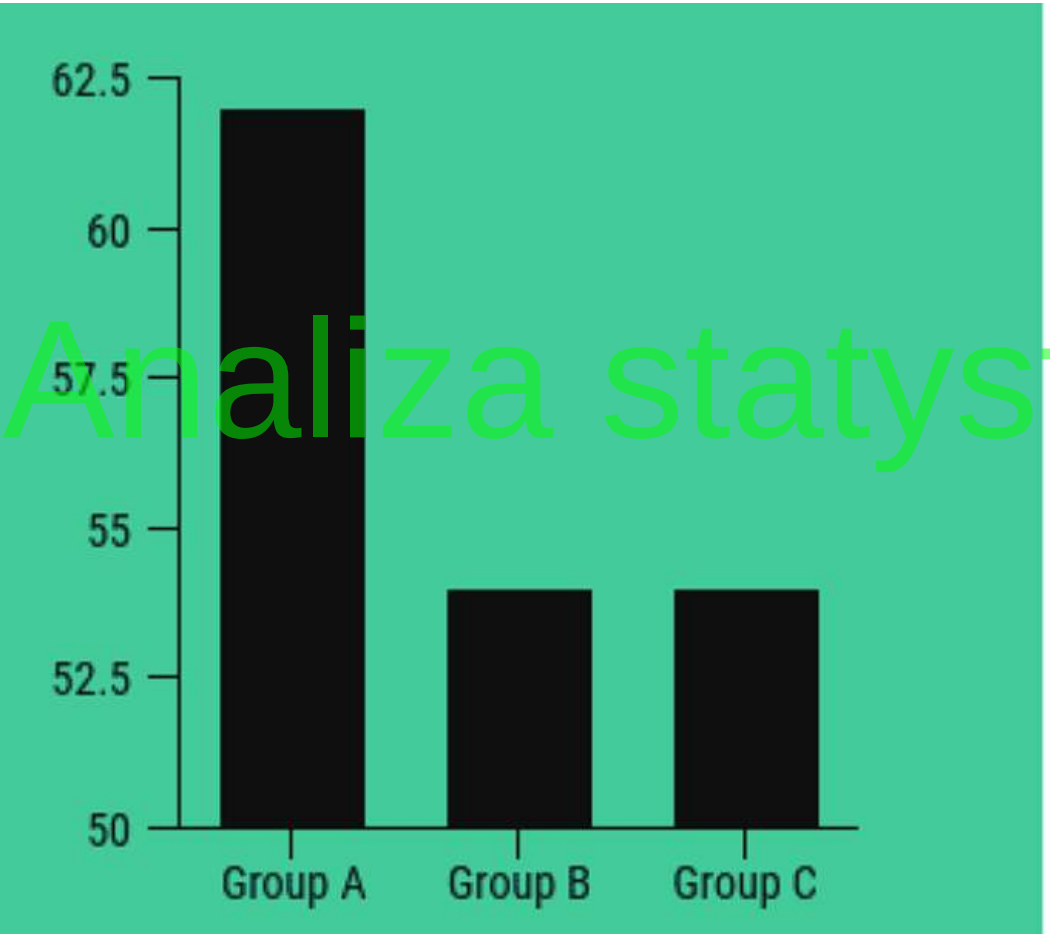
2) Myląca wizualizacja rezultatów



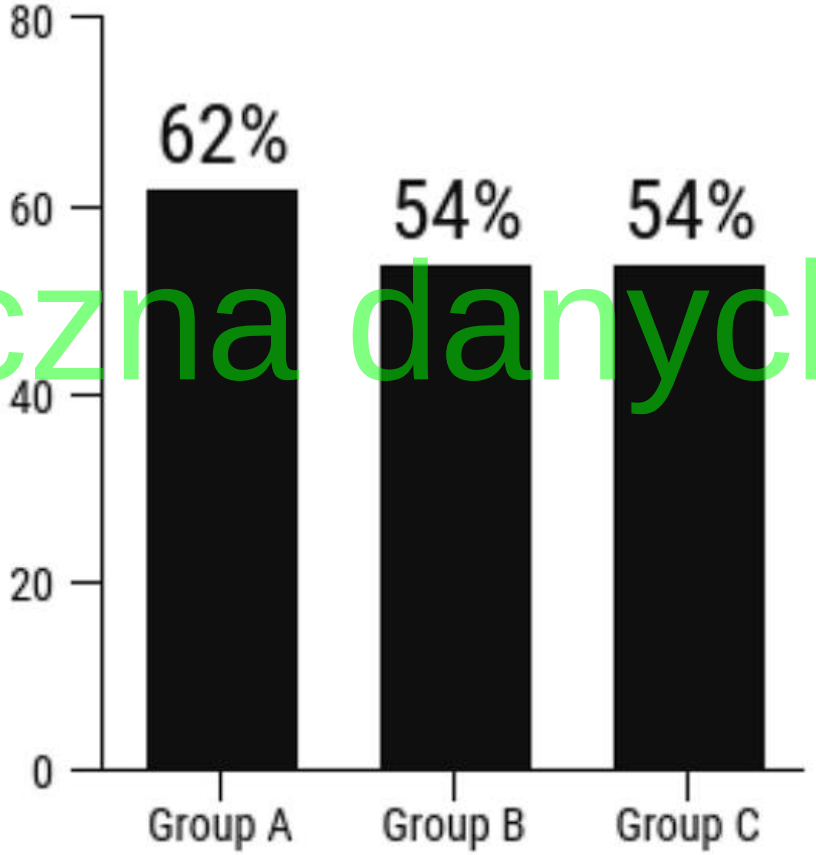
2) Myląca wizualizacja rezultatów



2) Myląca wizualizacja rezultatów

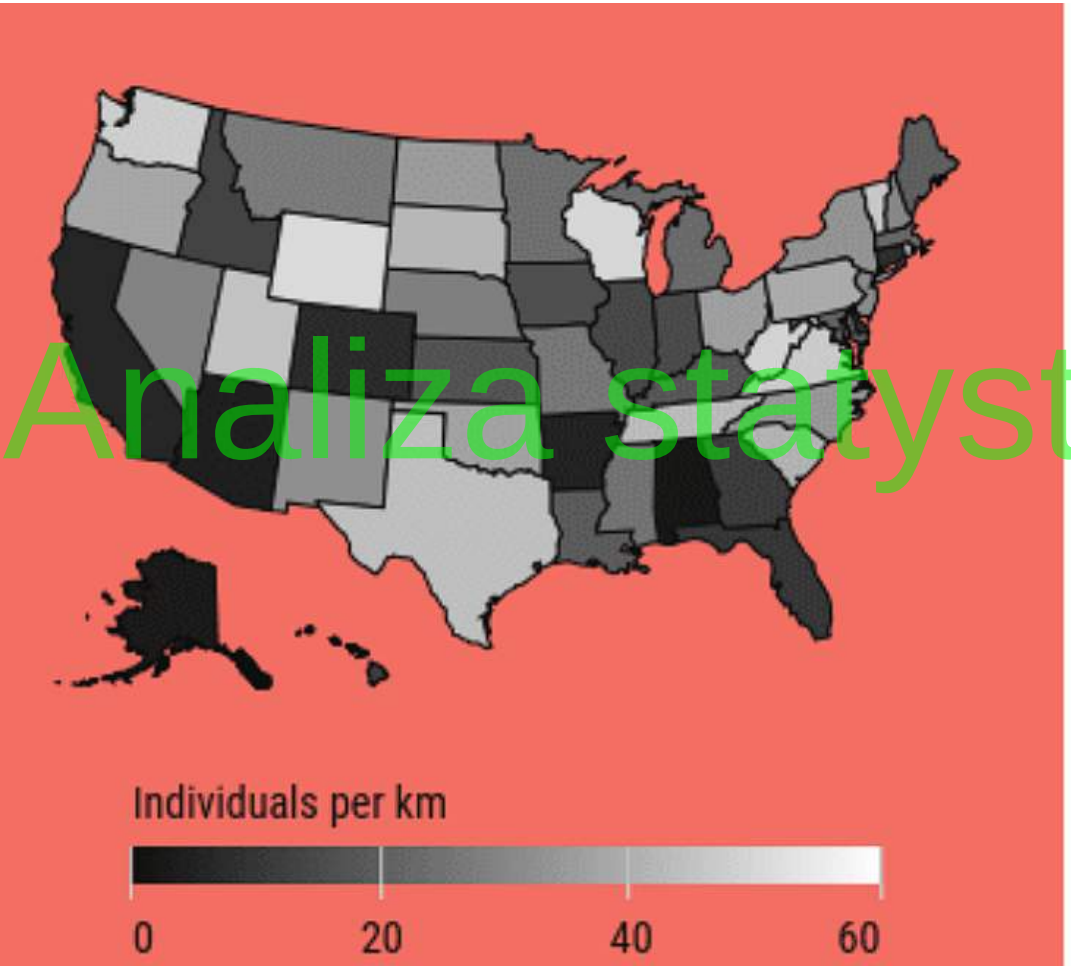


Źle

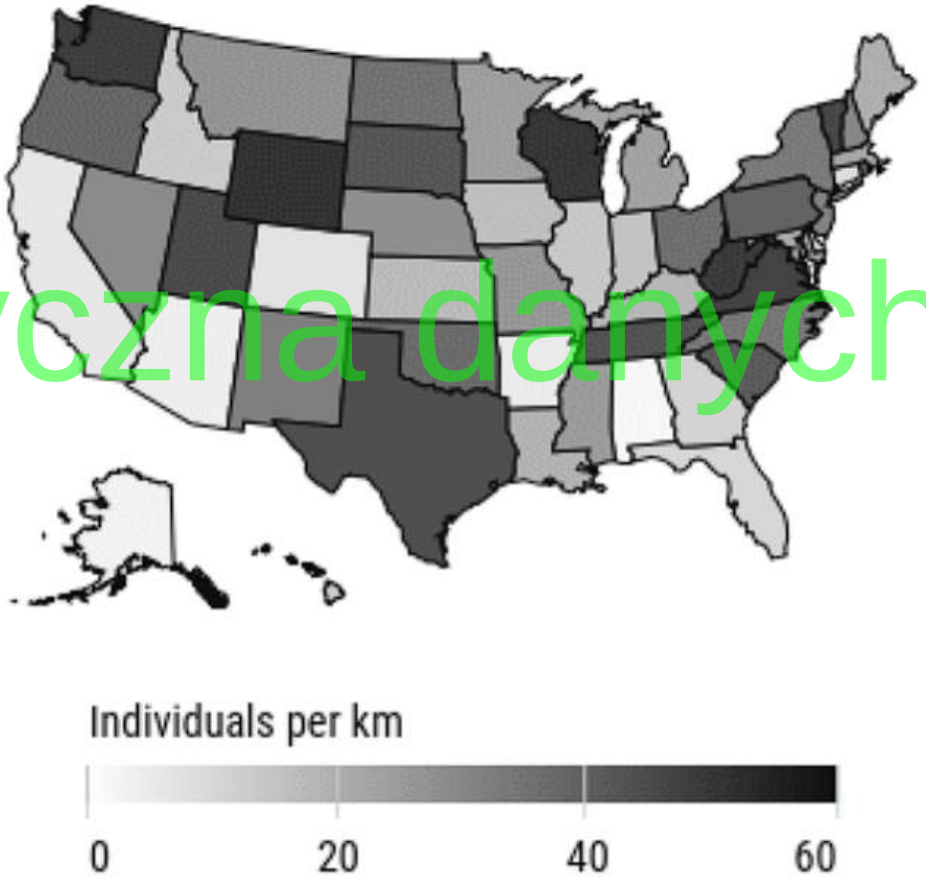


Dobrze

2) Myląca wizualizacja rezultatów



Źle



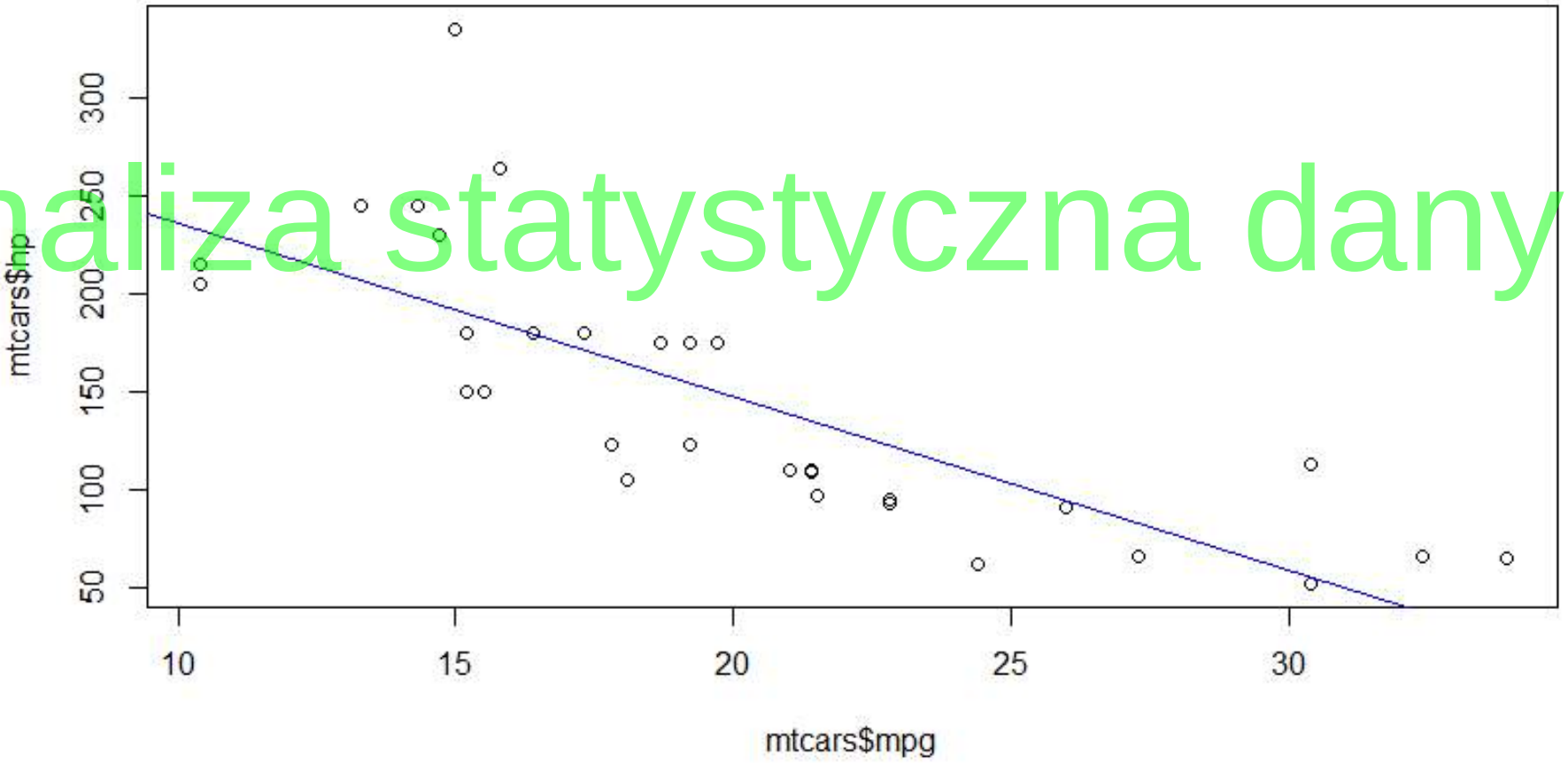
Dobrze

3) Nieadekwatne wnioskowanie

- Korelacja *versus* wywoływanie
- Niezrozumienie losowości i prawdopodobieństwa zajścia zdarzeń
- Błąd atomistyczny
- Błędne wnioskowanie z wizualizacji

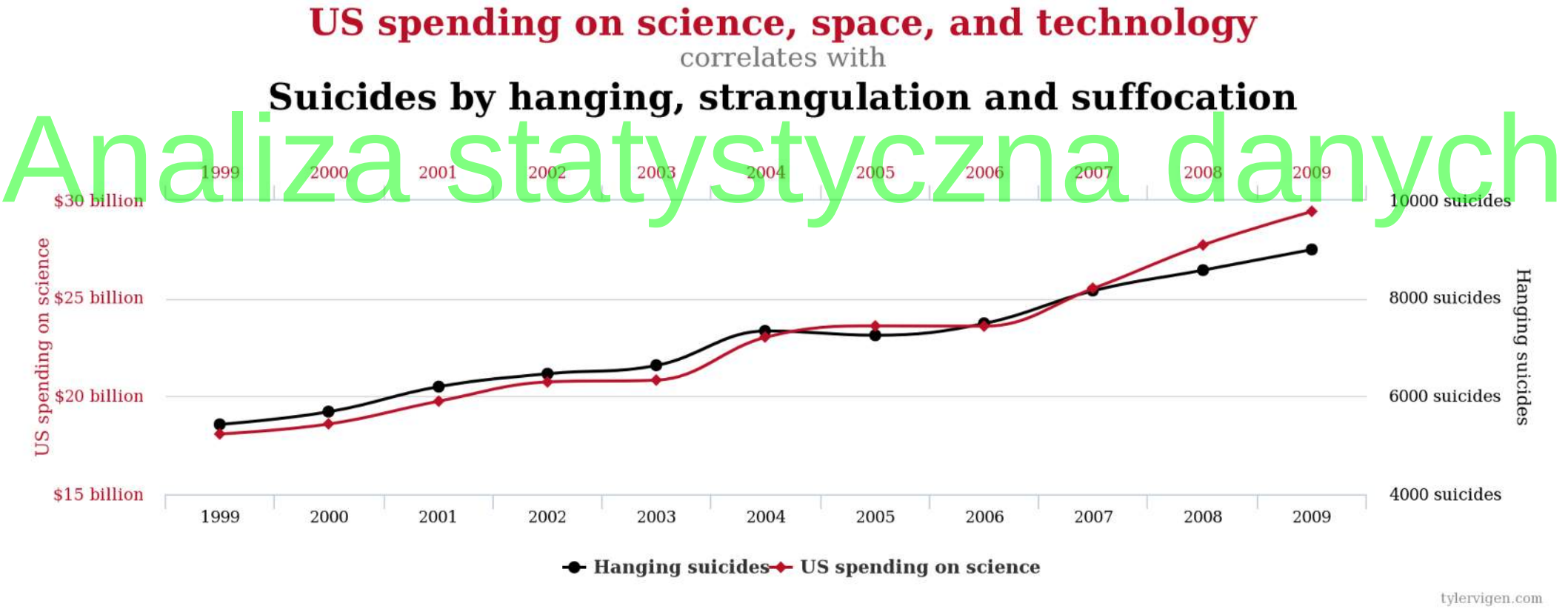
3) Nieadekwatne wnioskowanie

Korelacja *versus* wywoływanie



3) Nieadekwatne wnioskowanie

Korelacja *versus* wywoływanie



4) Celowe fałszowanie danych

10/26/2009, 00.00 | SOUTH KOREA

Send to a friend

Sentenced for fraud Hwang Woo-suk, pioneer of (false) "human cloning"

56 year old scientist also charged with embezzlement and breach of laws on bioethics. He falsified stem cell research, claiming to have cloned cells from healthy patients. The prosecutor asks for four years in prison; sentence due in the next few hours.

Seoul (AsiaNews / Agencies) - A court in Seoul sentenced for fraud the controversial South Korean scientist Hwang Woo-suk, famous for experiments on stem cells and human cloning. He is also charged with fraud, embezzlement and violation of laws on bioethics. He was celebrated and revered as a national hero for having led South Korea at the forefront of scientific research; revelations about the falsification of his experiments shocked the entire nation.

Pomiar (zebranie) danych

Analiza statystyczna danych

Pomiar (zebranie) danych

Typy zmiennych:

- **stymulanty** - zmienne, których wysokie wartości są pożądane
- **destymulanty** - zmienne, których wysokie wartości są niepożądane
- **nominanty** - zmienne, których odchylenia od poziomu najkorzystniejszego (optymalnego poziomu nasycenia) są niepożądane

Pomiar (zebranie) danych

- Populacja celowa – to czym się będziemy zajmować jako całość
- Populacja badana (operat losowania) – formalny zbiór jednostek, który potencjelnie możemy zbadać
- **Próba badawcza** – grupa jednostek wylosowanych do badania

Losowe schematy doboru prób badawczych

- Opierają się na losowości i wykorzystaniu rachunku prawdopodobieństwa, aby zmniejszyć ryzyko błędu
- Wykorzystywane są generatory liczb losowych dostępne w pakietach do obliczeń statystycznych
- Najczęściej wykorzystywane w badaniach ilościowych

Losowe schematy doboru prób badawczych

- Dobór losowy prosty
- Dobór losowy warstwowy
- Dobór losowy systematyczny
- Dobór zespołowy

Nielosowe schematy doboru prób badawczych

- Nie wykorzystuje się losowania i rachunku prawdopodobieństwa
- Badacz sam dokonuje wyboru konkretnych jednostek do badania
- Najczęściej wykorzystywane w badaniach jakościowych

Nielosowe schematy doboru prób badawczych

- Dobór celowy
- Dobór kwotowy
- Dobór oparty na dostępności danych
- Dobór metodą kuli śnieżnej

Wielkość próby badawczej

- Zarówno zbyt mała, jak i zbyt duża próba badawcza niesie ze sobą określone problemy
- Przyjmuje się zazwyczaj założenie o 95% **poziomie ufności** i przedziale **błędu losowego** +/- 3%
- Wielkość próby badawczej zależy od np. konkretnego problemu, rozproszenia danych, dostępnego czasu, *response rate*

Eksploracja danych – bardziej zaawansowane metody

- Transformacja i skalowanie zmiennych:
  - Transformacja logarytmiczna
  - Normalizacja
  - Standaryzacja
- Metody grupowania (klasteryzacji):
  - metody niehierachiczne (k-means)
  - metody hierarchiczne
  - metody oparte na gęstości (dbscan)
- Uczenie maszynowe (ML)

Transformacja i skalowanie zmiennych

- Poprawa interpretowalności danych
- Uporządkowanie prezentacji graficznej
- Głębszy wgląd w dane
- Spełnienie założeń do wnioskowania statystycznego

Transformacja logarytmiczna

- Dane często są mocno skrzywione, lub skoncentrowane wokół jednej wartości
- Wnioskowanie z takich danych jest utrudnione/niemożliwe
- Transformacja logarytmiczna ma na celu upodobnienie zbioru do rozkładu normalnego
- $x \rightarrow \log(x)$

Transformacja logarytmiczna

- Przykład

Analiza statystyczna danych

Normalizacja zmiennych

- Ma na celu przekształcenie zmiennych tak, aby zawierały się one w przedziale 0 – 1
- Zmniejsza złożoność danych
- Redukuje anomalie w zbiorze
- Normalizacja *min – max*:

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Normalizacja zmiennych

- Przykład

Analiza statystyczna danych

Standaryzacja zmiennych

- Przekształcenie zmiennych tak, aby miały średnią = 0 i odchylenie standardowe = 1
- Podobnie jak normalizacja *min-max,* standaryzacja ma na celu ułatwienie porównywania zmiennych o różnych skalach numerycznych

$$X_{stand} = \frac{X - \bar{X}}{\sigma}$$

Standaryzacja zmiennych

- Przykład

Analiza statystyczna danych

Eksploracja danych – bardziej zaawansoane metody

- Transformacja i skalowanie zmiennych:
  - Transformacja logarytmiczna
  - Normalizacja
  - Standaryzacja
- Metody grupowania (klasteryzacji):
  - metody niehierachiczne (k-means)
  - metody hierarchiczne
  - metody oparte na gęstości (dbscan)
- Uczenie maszynowe (ML)

Metody klasteryzacji (grupowania)

- Wykorzystywane w celu identyfikacji grup skupiających podobne do siebie jednostki
- Składa się na nie wiele algorytmów różniących się zarówno sposobem wykrywania grup jak i różnicami w ich definicji

Analiza statystyczna danych

Metody klasteryzacji (grupowania)

- Wykorzystywane w celu identyfikacji grup skupiających podobne do siebie jednostki
- Składa się na nie wiele algorytmów różniących się zarówno sposobem wykrywania grup jak i różnicami w ich definicji
- Algorytmy grupowania zmiennych dzielimy na:
  - niehierarchiczne (oparte na centroidach, np. *kmeans*),
  - hierarchiczne (np. *drzewo klasyfikacyjne*),
  - oparte na gęstości (łączą obszary o wysokiej gęstości, np. *dbscan*)

Algorytm k-średnich (k-means)

- Klasyczny algorytm *k-means* został wprowadzony przez Hartigana i Wonga (1979).
- Mając ustaloną liczbę skupień (*k*), przyporządkowuje obserwacje do klastrów tak, aby średnie w klastrach (dla wszystkich zmiennych) były jak najbardziej różne od siebie.
- Różnice między obserwacjami są mierzone w kategoriach jednej z kilku miar odległości (np. *euklidesową, Chebysheva, Manhattan*)

Algorytm k-średnich (k-means)

1) Określenie liczby klastrów (k) do utworzenia

2) Wybierz losowo k obiektów z zestawu danych jako początkowe centra klastrów

3) Przypisz każdą obserwację do najbliższego centroida, w oparciu o odległość euklidesową pomiędzy obiektem a centroidem

Algorytm k-średnich (k-means)

1) Określenie liczby klastrów (*k*) do utworzenia

2) Wybierz losowo k obiektów z zestawu danych jako początkowe centra klastrów

3) Przypisz każdą obserwację do najbliższego centroida, w oparciu o odległość euklidesową pomiędzy obiektem a centroidem

4) Dla każdego z *k* klastrów aktualizuj centroid poprzez obliczenie nowych wartości średnich dla wszystkich punktów danych w klastrze.

5) Iteruj kroki 3 i 4 do momentu, gdy przypisania klastrów przestaną się zmieniać lub gdy zostanie osiągnięta maksymalna liczba iteracji.

Algorytm k-średnich (k-means)

- Przykład

Analiza statystyczna danych

Algorytm k-średnich (k-means)

Problemy:

- Wymaga wybrania z góry odpowiedniej liczby klastrów
- Uzyskane wyniki końcowe są wrażliwe na początkowy losowy wybór centrów klastrów

Algorytm k-średnich (k-means)

Problemy:

- Wymaga wybrania z góry odpowiedniej liczby klastrów
- Uzyskane wyniki końcowe są wrażliwe na początkowy losowy wybór centrów klastrów
- Jest wrażliwy na wartości odstające.
- Działa dobrze dla wyraźnie odesparowanych klastrów
- Zmiena kolejności danych może prowadzić do innych wyników klasteryzacji

Metody hierarchiczne (drzewo klasyfikacyjne)

- Nie wymaga wcześniejszego określenia liczby skupień; wymaga jednak wskazania metody obliczania podobieństwa pomiędzy obserwacjami
- Efektem działania są obiekty pogrupowane w klastry według ich hierarchii.
- Algorytm tworzy drzewopodobny obiekt graficzny o nazwie dendrogram
- Poprzez odcinanie gałęzi dendrogramu użytkownik formuje porządaną liczbę grup

Drzewo klasyfikacyjne - algorytm

1) Umieść każdy punkt danych w jego własnym klastrze.

2) Zidentyfikuj najbliższe (najpodobniejsze) dwa klastry i połącz je w jeden klaster.

3) Powtarzaj krok 2, aż wszystkie punkty danych znajdą się w jednym klastrze.

Drzewo klasyfikacyjne - algorytm

Przykład

Analiza statystyczna danych

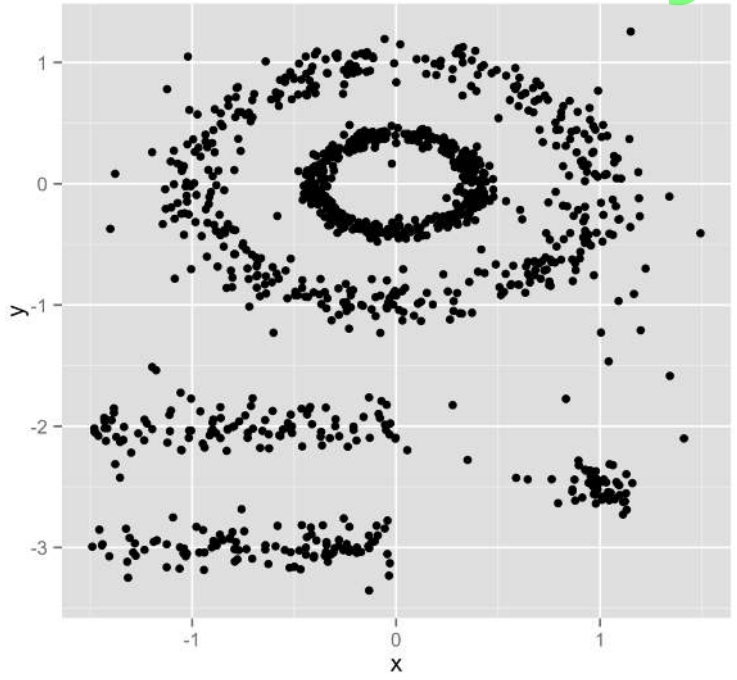
Drzewo klasyfikacyjne

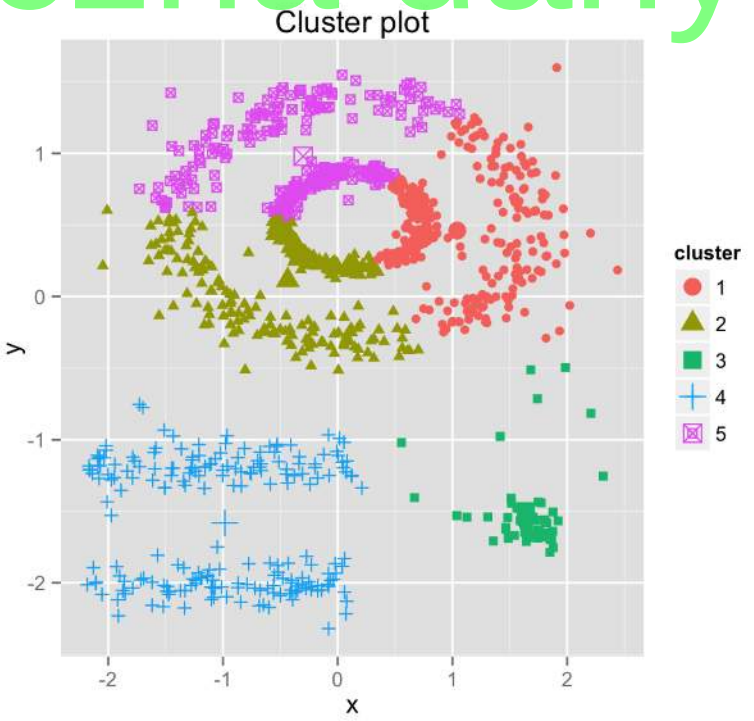
Problemy:

- Określenie miary podobieństwa/ niepodobieństwa pomiędzy obserwacjami
- Określenie miejsc odcięcia dendrogramu
- Wrażliwy na obserwacje odstające
- Działa dobrze dla wyraźnie odesparowanych klastrów

Klasyfikcja oparta na gęstości - dbscan

- W praktyce dane (zwłaszcza przestrzenne) znajdują się często w z góry określonych grupach i zawierają tzw. szum oraz wartości odstające
- Klasyczne algorytmy miałyby problem w klasyfikacji danych tego typu



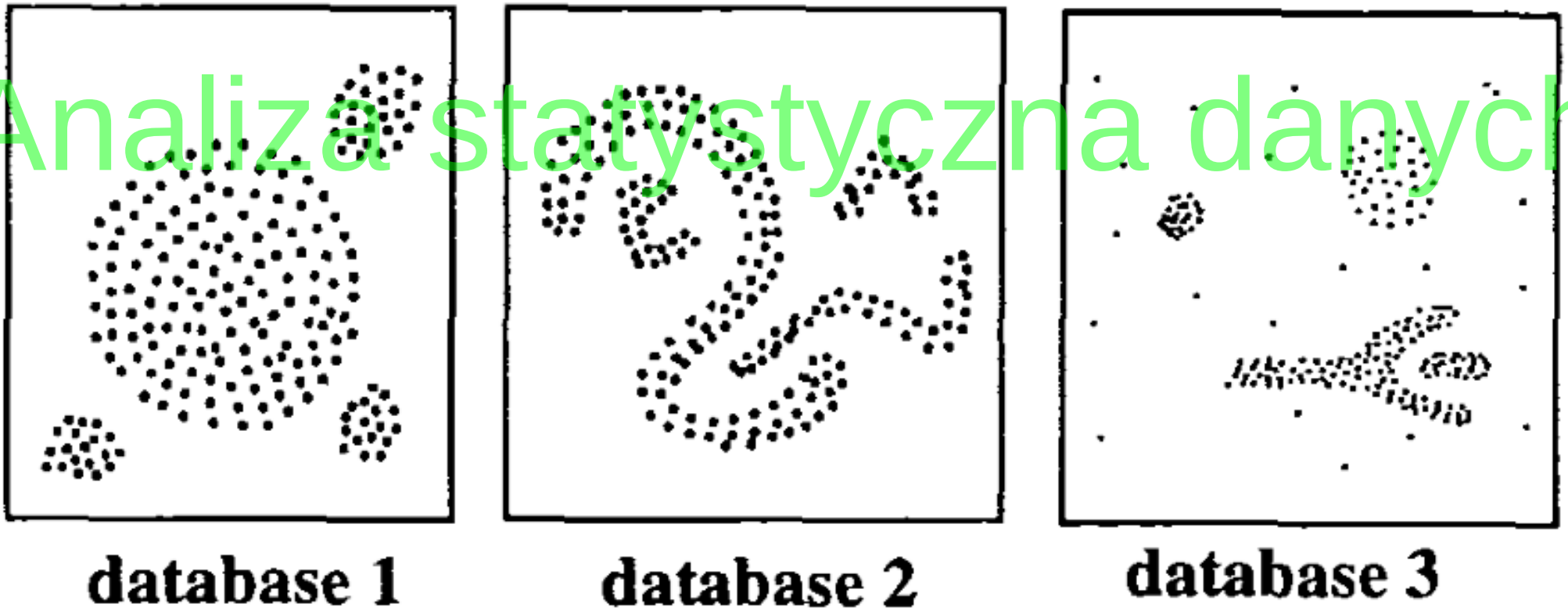


Klasyfikcja oparta na gęstości - dbscan

- Density-Based Spatial Clustering and Application with Noise
- DBSCAN (Ester et al. 1996) jest odporny na wskazane problemy
- Nie wymaga wskazania liczby skupień
- Wykrywa klastry o dowolnym kształcie
- Podstawowa idea wywodzi się z intuicyjnej dla człowieka metody klastrowania

Klasyfikcja oparta na gęstości - dbscan

Klastry to obszary o zwiększonej gęstości, oddzielone obszarami o małej gęstości



Ester et al, 1996

Klasyfikcja oparta na gęstości - dbscan

- Algorytm wymaga wskazania dwóch parametrów: *epsilon i minimum points*

  1) Dla każdego punktu x oblicz odległość między x a innymi punktami.

  2) Znajdź wszystkie punkty sąsiednie w odległości *epsilon* od punktu początkowego.

Klasyfikcja oparta na gęstości - dbscan

- Algorytm wymaga wskazania dwóch parametrów: *epsilon i minimum points*

  1) Dla każdego punktu x oblicz odległość między x a innymi punktami.

  2) Znajdź wszystkie punkty sąsiednie w odległości *epsilon* od punktu początkowego.

  3) Dla każdego punktu, jeśli nie jest on jeszcze przypisany do klastra, utwórz nowy klaster (jeżeli liczba sąsiadów >= *minimum points*).

  4) Znajdź wszystkie jego gęsto połączone punkty (*epsilon*) i przypisz je do tego samego klastra co punkt główny.

  5) Iteruj przez pozostałe nieodwiedzone punkty w zbiorze danych.

Klasyfikcja oparta na gęstości - dbscan

- Przykład

Analiza statystyczna danych

Eksploracja danych – bardziej zaawansoane metody

- Transformacja i skalowanie zmiennych:
  - Transformacja logarytmiczna
  - Normalizacja
  - Standaryzacja
- Metody grupowania (klasteryzacji):
  - metody niehierachiczne (k-means)
  - metody hierarchiczne
  - metody oparte na gęstości (dbscan)
- Uczenie maszynowe (ML)

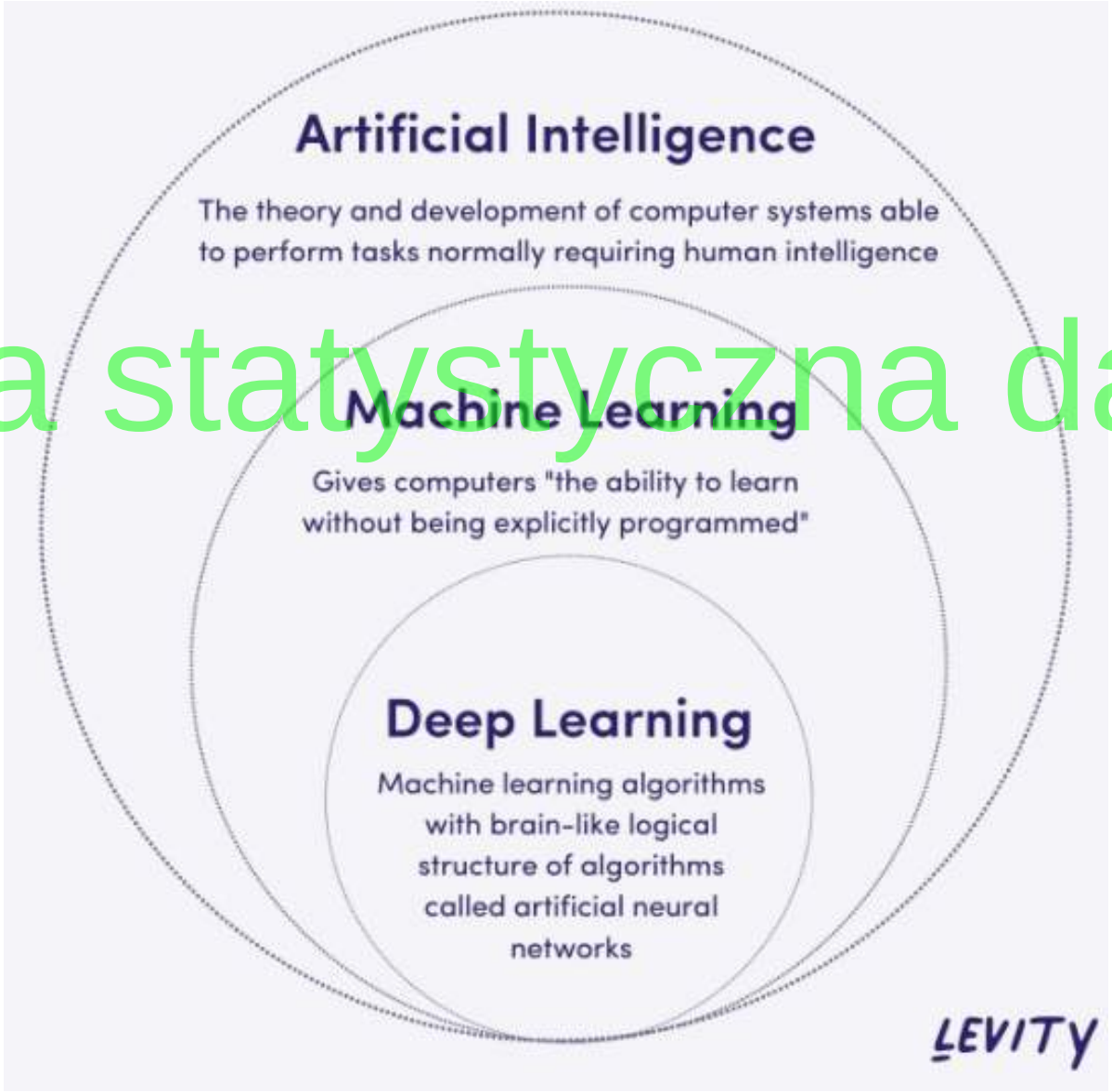
Uczenie maszynowe jako część AI

- Badania rozpoczęły się w Darthmouth College (USA) w 1956 r.
- Pierwsze implementacje obejmowały strategie gry w szachy, rozwiązywanie problemów matematycznych
- W tym czasie naukowcy wierzyli w szybki postęp i opracowanie uogólnionej sztucznej inteligencji (AGI)

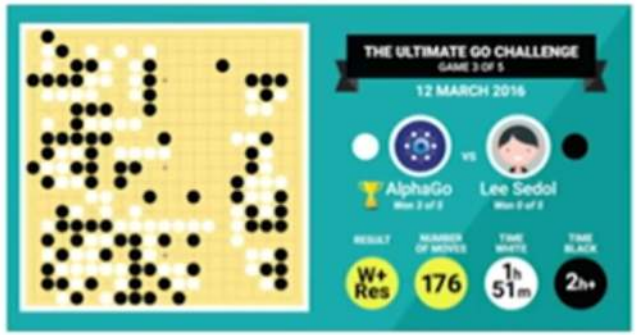
Uczenie maszynowe jako część AI

- Badania rozpoczęły się w Darthmouth College (USA) w 1956 r.
- Pierwsze implementacje obejmowały strategie gry w szachy, rozwiązywanie problemów matematycznych
- W tym czasie naukowcy wierzyli w szybki postęp i opracowanie uogólnionej sztucznej inteligencji (AGI)
- Okres zastoju od lat 70 do końca lat 90 XX w.
- Powolny rozwój nastąpił pod koniec lat 90-tych i na początku XXI w.
- Skokowy wzrost aplikacji od roku 2015
- Zmiana paradygmatu AI!

Sztuczna inteligencja i uczenie maszynowe

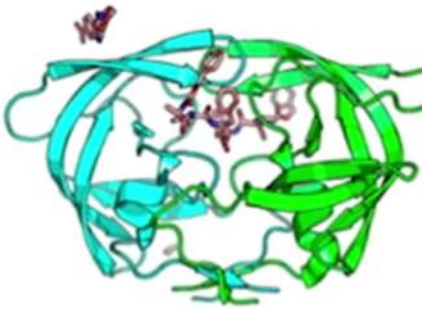


Uczenie maszynowe jest wszędzie

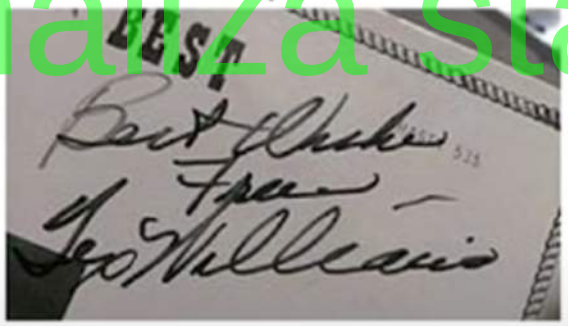


AlphaGo

Recommendation systems



Drug discovery



Character recognition

2σ TWO SIGMA

Hedge fund stock predictions

Voice assistants

Assisted driving

Face detection/recognition

Cancer diagnosis

Uczenie maszynowe

- Dziedzina badań, która skupia się na systemach komputerowych, które mogą uczyć się na podstawie danych.

Analiza statystyczna danych

Uczenie maszynowe

- Dziedzina badań, która skupia się na systemach komputerowych, które mogą uczyć się na podstawie danych.
- Systemy ML (modele) potrafią uczyć się konkretnych zadań na podstawie analizy dużej liczby przykładów, np. model ML może nauczyć się jak rozpoznać samochód na podstawie obserwacji dużej liczb aut.

Analiza statystyczna danych

Uczenie maszynowe

- Brak programowania reguł wprost (przez programistę) – model może nauczyć się rozwiązywać konkretny problem bez predefiniowanych konkretnych reguł
- Model uczy się sam, jakie charakterystyki są istotne, aby rozpoznać dany obiekt

Uczenie maszynowe

- Brak programowania reguł wprost (przez programistę) – model może nauczyć się rozwiązywać konkretny problem bez predefiniowanych konkretnych reguł
- Model uczy się sam, jakie charakterystyki są istotne, aby rozpoznać dany obiekt
- Istotna jest ilość i jakość danych
- Modele ML potrafią wykrywać wzorce, schematy w danych
- ML wspiera podejmowanie decyzji w oparciu o dane (data-driven decisions)

Uczenie maszynowe

- Nienadzorowane (*unsupervised learning*)
- Nadzorowane (*supervised learning*)
- Posiłkowane (*reinforcement learning*)

Analiza statystyczna danych

Uczenie maszynowe

Etapy budowy modelu ML:

1) Zgromadzenie/pozyskanie danych

2) Przygotowanie danych do dalszej analizy (porządkowanie, usuwanie obserwacji odstających)

3) Wybór metody i modelu

4) Trenowanie modelu (w przypadku metod nadzorowanych)

5) Ewaluacja, określenie i pomiar błędów

6) Dopasowanie parametrów

Uczenie maszynowe

Analiza statystyczna danych

https://www.youtube.com/watch?v=nKW8Ndu7Mjw&ab_channel=GoogleCloudTech

Uczenie maszynowe

- Przykład – uczenie nadzorowane:

  1) wykorzystamy zdjęcia satelitarne powiatu śremskiego, a także informacje o klasach pokryciach terenu dostępne na www.s2glc.cbk.waw.pl

  2) zdjęcia i dane muszą zostać przygotowane do pracy pod kątem modelu ML (np. przycinanie, przekształcanie, wartości odstające)

  3) losowo wybierzemy część danych do nauki wykrywania klas zagospodarowania terenu przez model ML

  4) użyjemy wytrenowanego modelu do klasyfikacji zagospodarowania terenu powiatu śremskiego (zdjęcia satelitarne)

  5) ocenimy jakość klasyfikacji przeprowadzonej przez model

DEVS (2020) – HBO MAX

Wnioskowanie statystyczne

Analiza parametryczna:

- Analiza wariancji (t-test i ANOVA)

- Regresja prosta i wieloraka

Analiza nieparametryczna:

- Regresja logistyczna

Analiza szeregów czasowych:

- modele autoregresji (AR) i średniej ruchomej (MA)

- modele zintegrowane (ARIMA)

Analiza parametryczna (model liniowy)

- Celem jest określenie wpływu zestawu zmiennych niezależnych na zmienną zależną.
- W modelu liniowym wielkość zmiany zmiennej zależnej ilustruje się jako wielokrotność zmiany zmiennych niezależnych.

Analiza parametryczna (model liniowy)

- Celem jest określenie wpływu zestawu zmiennych niezależnych na zmienną zależną.
- W modelu liniowym wielkość zmiany zmiennej zależnej ilustruje się jako wielokrotność zmiany zmiennych niezależnych.
- R. Fisher i założenia modelu parametrycznego (niezależność, normalność rozkładu, linowa zależność)
- Analiza wariancji i regresja jako podstawa do wnioskowania na temat wpływu jednej zmiennej na drugą

Analiza parametryczna – założenie o niezależności zmiennych

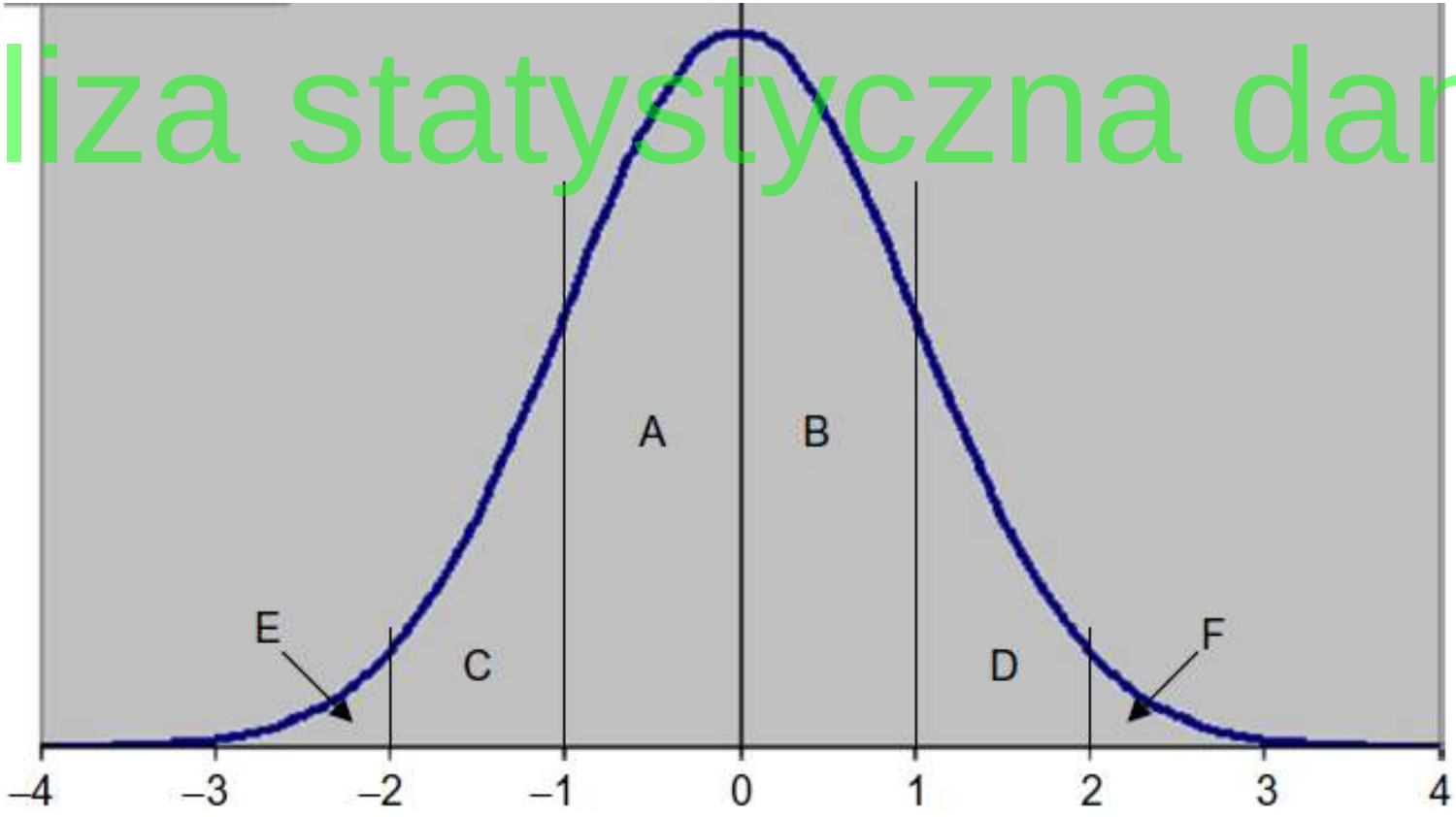
- Zależność oznacza istnienie (jakiegoś) połączenia pomiędzy zmiennymi
- Założenie niezależności oznacza, że dane nie są w żaden sposób powiązane

Analiza parametryczna – założenie o niezależności zmiennych

- Zależność oznacza istnienie (jakiegoś) połączenia pomiędzy zmiennymi
- Założenie niezależności oznacza, że dane nie są w żaden sposób powiązane
- Obserwacje pomiędzy grupami powinny być niezależne (grupy składają się z różnych osób).
- Obserwacje w obrębie każdej grupy muszą być niezależne (brak związku pomiędzy uczestnikami badania)
- Kluczowy więc jest etap gromadzenia/wyboru danych

Analiza parametryczna – założenie o normalności rozkładu

Rozkład wartości każdej zmiennej w zbiorze danych jest rozkładem normalnym skupionym wokół wartości średniej

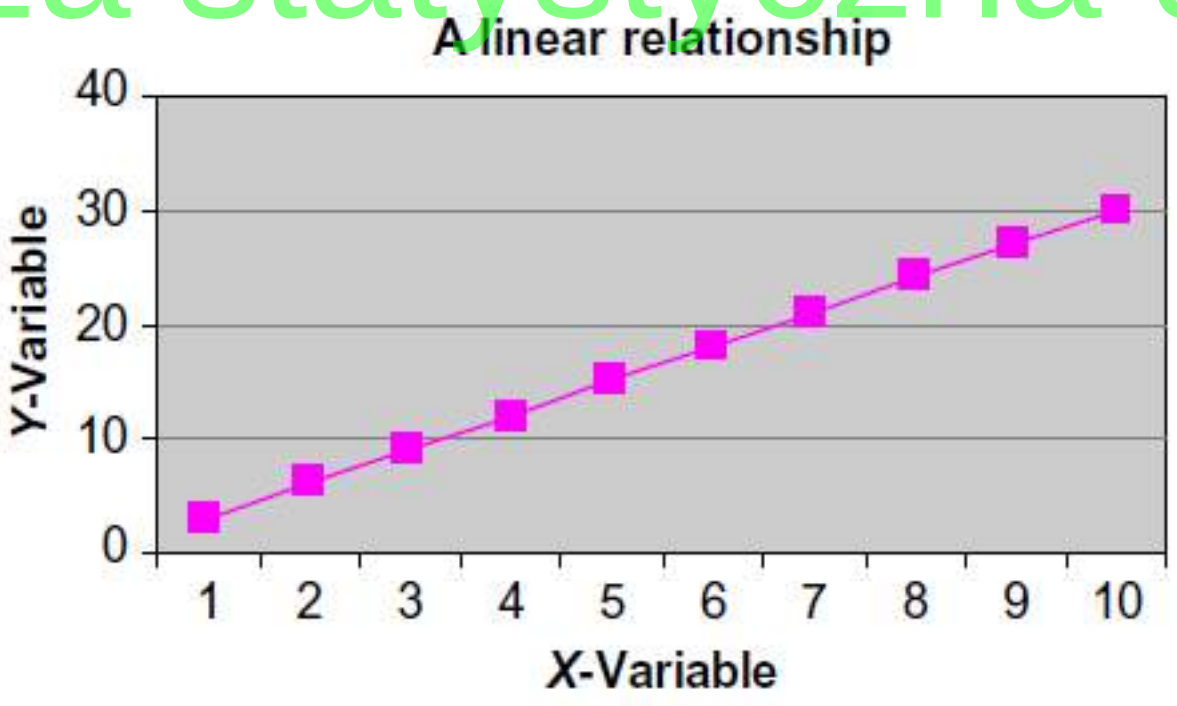


Założenie o normalności rozkładu i Centralne Twierdzenie Graniczne

- Grupa średnich z N próbek wylosowanych z rozkładu niebędącego rozkładem normalnym zbliża się do normalności, gdy N zbliża się do nieskończoności
- Im większa liczba próbek, tym bardziej rozkład średnich w wylosowanych próbkach (dla danej zmiennej) zbliża się do rozkładu normalnego
- Uwaga na błędy we wnioskowaniu!

Analiza parametryczna – założenie o linowej zależności pomiędzy zmiennymi

- Zmienna niezależna wywiera liniowy efekt na zmienną zależną
- Efekt ten może być zilustrowany linią prostą



Wnioskowanie statystyczne

Analiza parametryczna:

- Analiza wariancji (t-test i ANOVA)

- Regresja prosta i wieloraka

Analiza statystyczna danych

Analiza wariancji (test t)

- Test t-Studenta jest metodą pozwalającą określić, czy **dwie** populacje różnią się od siebie statystycznie
- Analizuje różnice w średnich i rozpiętości rozkładów (tj. wariancji) w poszczególnych grupach
- Dotyczy zmiennych numerycznych i wymaga spełnienia założeń analizy parametrycznej

William Sealy Gosset

Analiza wariancji – test t

- Typowy problem badawczy: czy średnia z grupy 1 jest równa średniej z grupy 2?
- Hipotezy:

  H0: Średnia w grupie 1 nie różni się istotnie od średniej w grupie 2: μ1=μ2

  H1: istnieją istotne różnice w średnich
- Gdy wartość p < od 0.05 odrzucamy H0 i przyjmujemy H1

Test-t

- Przykład:

  czy istnieją istotne różnice w średniej wartości zmiennej waga dla kobiet i mężczyzn?

Analiza statystyczna danych

Analiza wariancji (ANOVA)

- **An**alysis **o**f **va**riance (ANOVA)
- Służy do porównywania średnich trzech lub więcej grup

Analiza statystyczna danych

Analiza wariancji (ANOVA)

- **An**alysis **o**f **va**riance (ANOVA)
- Służy do porównywania średnich trzech lub więcej grup
- Pozwala stwierdzić, czy różnice średnich są istotne statystycznie (czy średnie w jednej grupie różnią się od innych)
- Jest wiele wersji testu (np. jednoczynnikowa, dwuczynnikowa)

Analiza wariancji (ANOVA)

Założenia:

- Wymaga spełnienia podstawowych założeń analizy parametrycznej
- Wymaga połączenia zmiennych nominalnych (o przynajmniej dwóch poziomach) ze zmienną numeryczną lub porządkową
- Brak (dużej liczby) wartości odstających

Analiza wariancji (ANOVA)

- Aplikacja testu wymaga zbadania dwóch hipotez:

  H0: wszystkie wartości średnie w grupach są takie same: μ1=μ2…=μn

  H1: istnieje przynajmniej jedna para wartości średnich różna od siebie

- Gdy wartość p < od 0.05 odrzucamy H0 i przyjmujemy H1

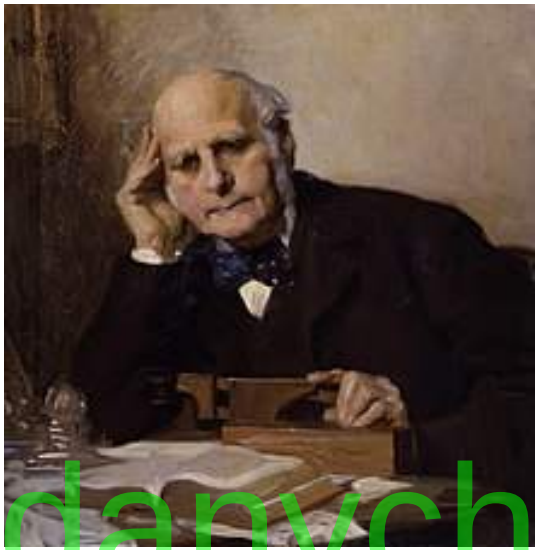
Jednoczynnikowa analiza wariancji

- Jednoczynnikowy model ANOVA służy do oceny wpływu jednej zmiennej grupującej na zmienną odpowiedzi
- Przykład:
  jedna zmienna grupująca: warunki uprawy i ich wpływ na wagę rośliny

Dwuczynnikowa analiza wariancji

- Dwuczynnikowy model ANOVA służy do jednoczesnej oceny wpływu dwóch zmiennych grupujących na zmienną odpowiedzi.
- Przykład:
  dwie zmienne grupujące: płeć i poziom wykształcenia i ich wpływ na poziom zadowolenia z wykonywanej pracy

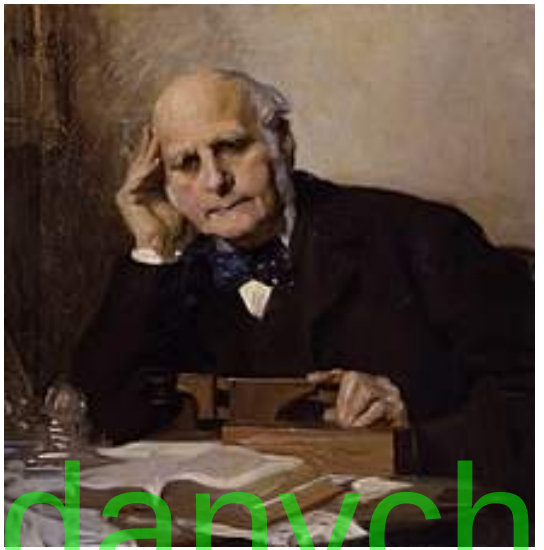
Regresja liniowa



1822-1911

- Ojcem był sir Francis Galton (1875)
- Stworzył fundament teoretyczny i empiryczny regresji analizując wagę nasion
- Koncepcja ta została potem rowinięta przez Pearsona (1896, 1930)

Regresja liniowa



1822-1911

- 'Ojcem' był sir Francis Galton (1875)
- Stworzył fundament teoretyczny i empiryczny regresji analizując wagę nasion
- Koncepcja ta została potem rowinięta przez Pearsona (1896, 1930)
- Dwa rodzaje modeli: regresja prosta i regresja wieloraka
- Jest szeroko stosowanym narzędziem statystycznym służącym do ustalenia zależności pomiędzy zmiennymi.

Regresja prosta

- Służy do przewidywania wartości zmiennej y na podstawie jednej zmiennej przewidującej x
- Celem jest zbudowanie modelu matematycznego (wzoru, formuły), który określa y jako funkcję zmiennej x.

Regresja prosta

- Służy do przewidywania wartości zmiennej y na podstawie jednej zmiennej przewidującej x
- Celem jest zbudowanie modelu matematycznego (wzoru, formuły), który określa y jako funkcję zmiennej x.
- Po zbudowaniu statystycznie istotnego modelu, można go wykorzystać do przewidywania przyszłych wyników zmiennej y na podstawie nowych wartości x
- Wymaga spełnienia założeń analizy parametrycznej

Regresja prosta

- Ogólny model regresji prostej to wzór na linię (y=ax+b):

$$Y = B_0 + B_1 X + u$$

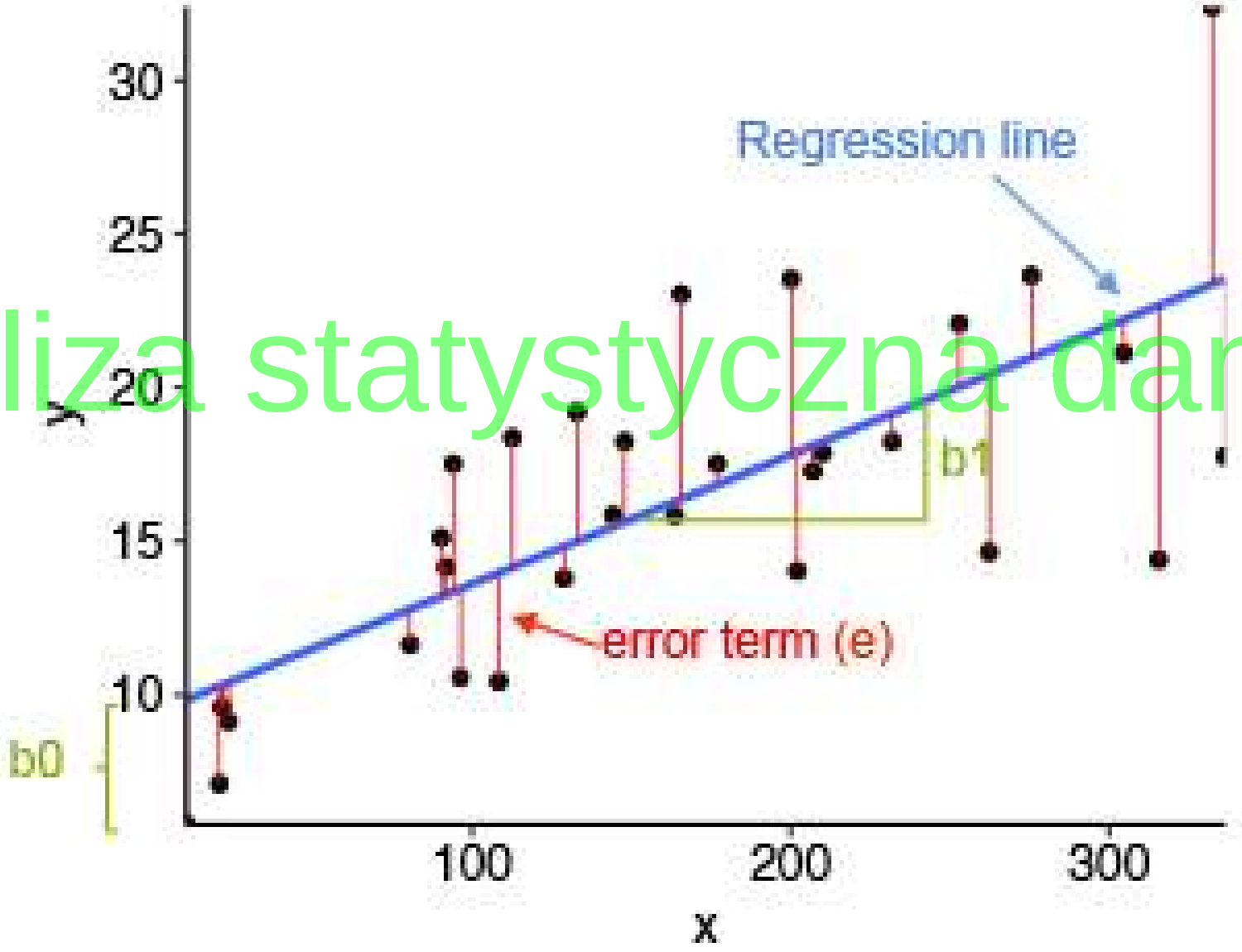
Gdzie:

Y - zmienna zależna
X – zmienna niezależna
B_0 – wyraz wolny
B_1 – współczynnik modelu regresji
u – składnik losowy modelu (błąd)

Regresja prosta

- Najprostszą metodą identyfikacji parametrów modelu jest metoda najmniejszych kwadratów (ordinary least squares (OLS))
- Bazuje ona na minimalizacji odległości punktów (obserwacji) od linii regresji

Regresja prosta



Regresja prosta

Przykład: Wpły budżetu reklamowego youtube na wielkosć sprzedaży

Analiza statystyczna danych

Regresja wieloraka

- To rozszerzenie regresji prostej na przypadki z wieloma zmiennymi niezależnymi
- Wykorzystywana do predykcji wartości zmiennej Y podstawie wartości (wielu) zmiennych objaśniających X

Regresja wieloraka

- Ogólny model regresji wielorakiej to wzór:

$$Y = \mathrm{B}_0 + \mathrm{B}_1 X_1 + \mathrm{B}_2 X_2 + \ldots + \mathrm{B}_n X_n + u$$

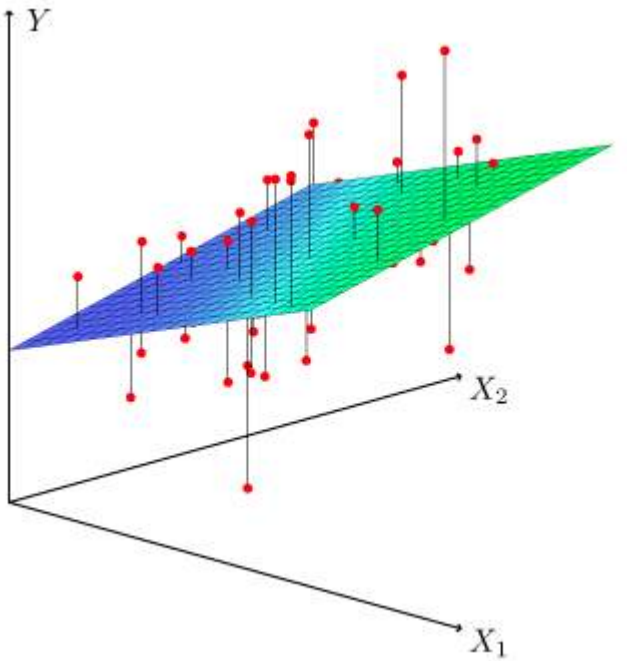
Analiza statystyczna danych

Gdzie:

Y - zmienna zależna
X_1 X_2 X_n – zmienne niezależne
B_0 – wyraz wolny
B_1 ,B_2,B_n współczynniki modelu regresji
u – składnik losowy modelu (błąd)

Regresja wieloraka

- Szacowanie wartości parametrów odbywa się tak samo jak w przypadku regresji prostej (OLS), jednak dla wielu wymiarów
- Chodzi o znalezienie n-wymiarowej płaszczyzny, która przechodzi najbliżej punktów współrzędnych (danych)

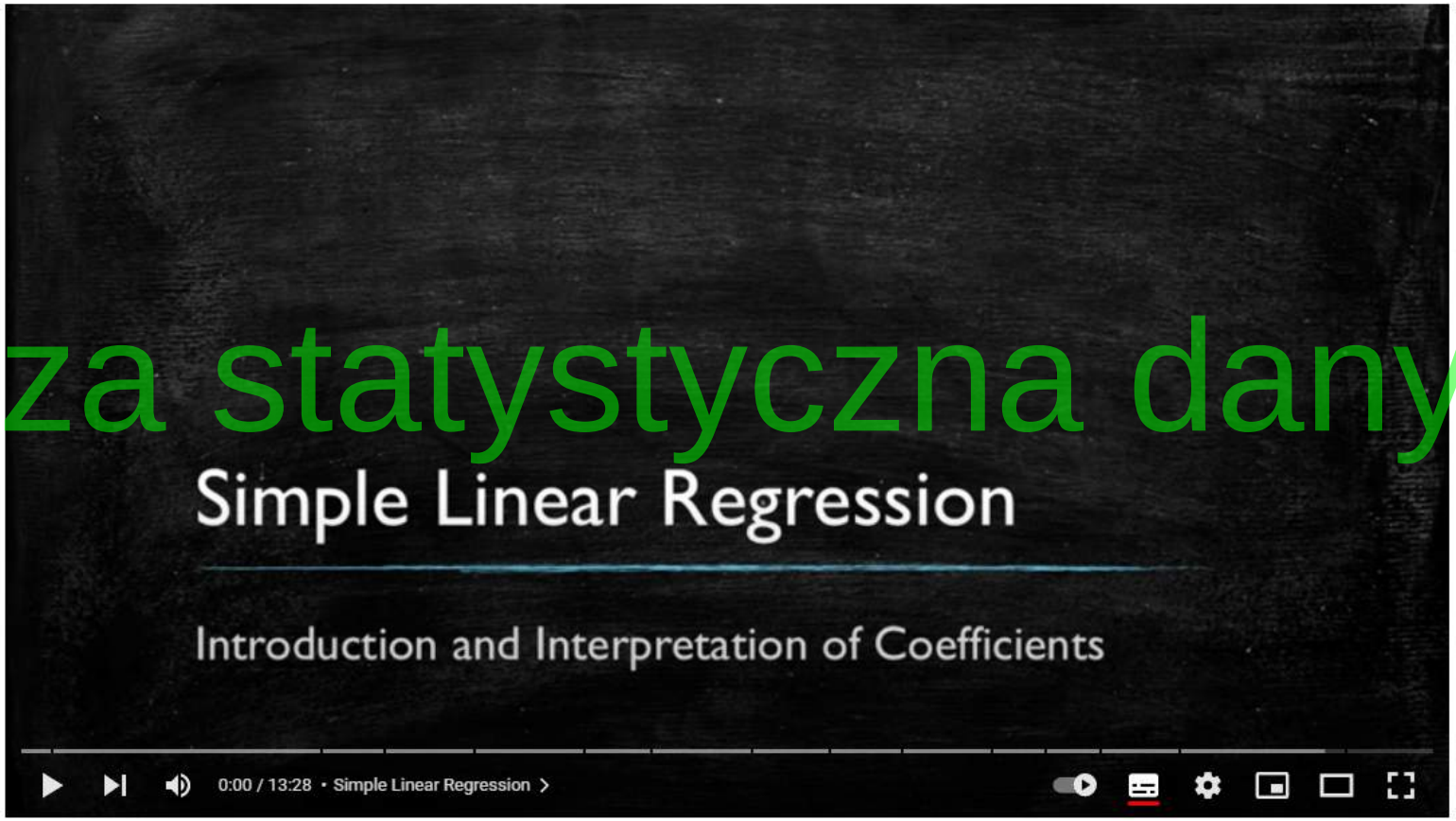


Regresja wieloraka

- Przykład: Wpły budżetu reklamowego youtube, facebook i gazeta na wielkosć sprzedaży

Analiza statystyczna danych

Regresja - podsumowanie



https://www.youtube.com/watch?v=owI7zxCqNY0&ab_channel=dataminingincae

Wnioskowanie statystyczne

Analiza nieparametryczna:

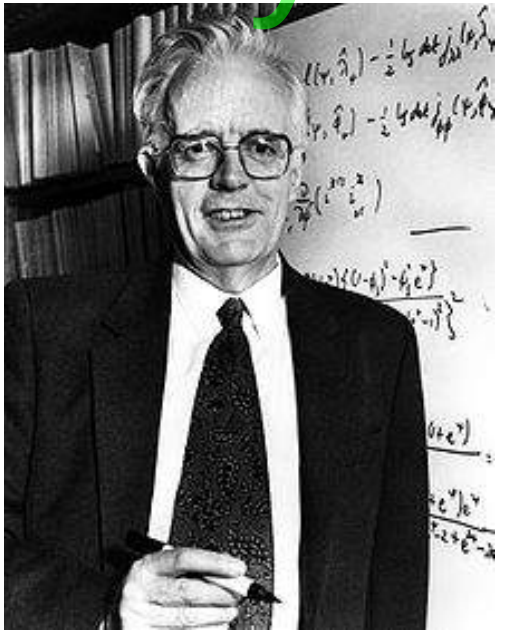
- Regresja logistyczna

Analiza szeregów czasowych:

- modele autoregresji i średniej ruchomej
- modele zintegrowane

Regresja logistyczna

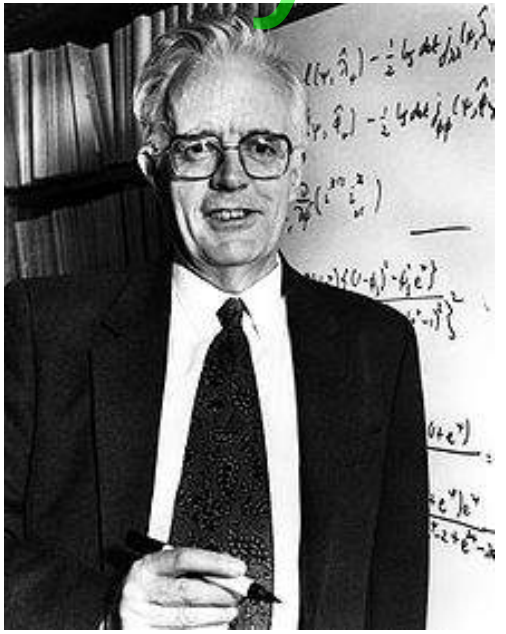
- Regresja logistyczna została wprowadzona przez Davida Coxa (1954)
- Należy do klasy tzw. uogólnionych modeli liniowych
- Modeluje prawdopodobieństwo zajścia zdarzenia jako liniową kombinacją jednej lub więcej zmiennych niezależnych.



1924-2022

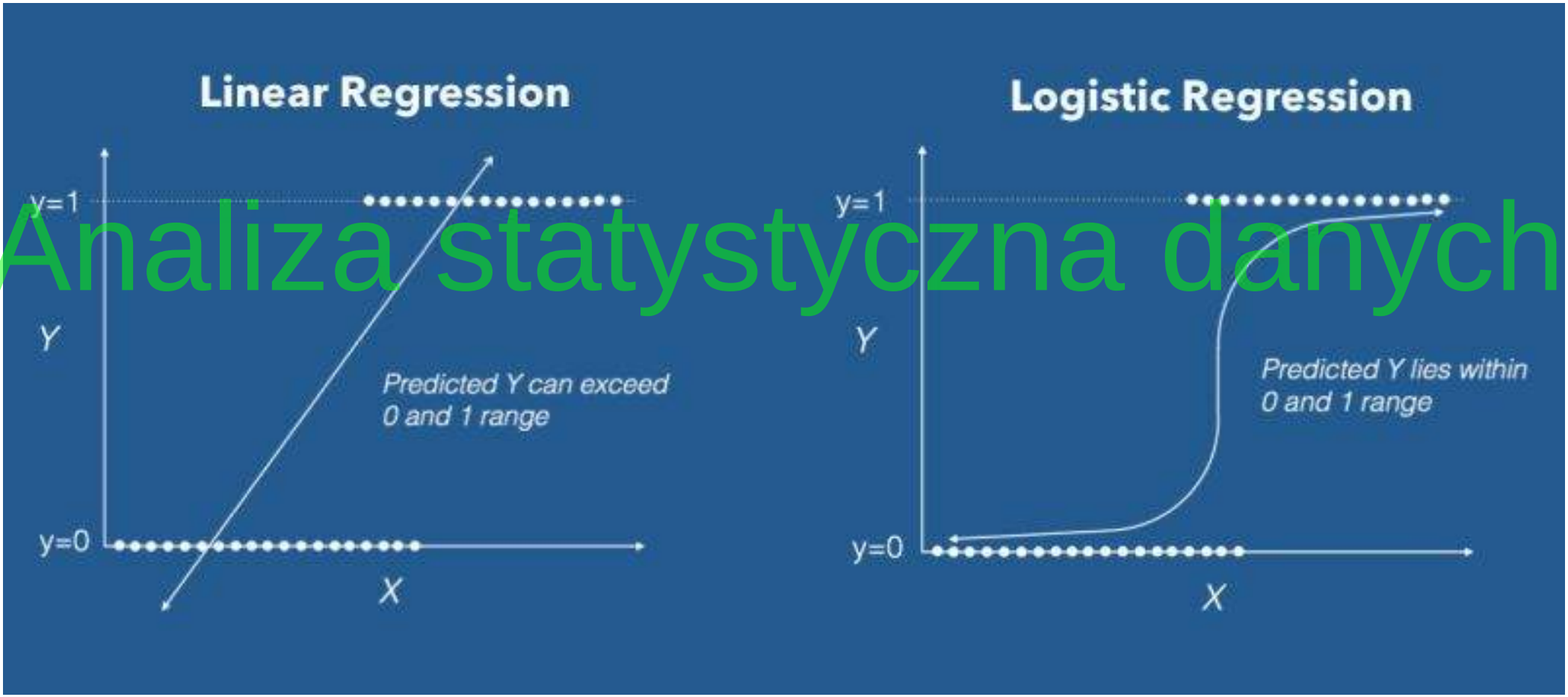
Regresja logistyczna

- Regresja logistyczna została wprowadzona przez Davida Coxa (1954)
- Należy do klasy tzw. uogólnionych modeli liniowych
- Modeluje prawdopodobieństwo zajścia zdarzenia jako liniową kombinacją jednej lub więcej zmiennych niezależnych.
- Zmienną zależną jest zmienna nominalna (binarna) lub porządkowa
- Model mierzy prawdopodobieństwo przynależności (nowej) obserwacji do określonej kategorii



1924-2022

Regresja logistyczna



Regresja logistyczna

Funkcja logistyczna, opisuje prawdopodobieństwo zajścia zdarzenia na podstawie zmiennej przewidującej (X):

$$p(X) = \frac{1}{1 + e^{-(X)}}$$

Gdzie:
e - to stała Eulera
p - jest prawdopodobieństwem wystąpienia zdarzenia (danego X)

Regresja logistyczna

Dla jednego predyktora:

$$\log\left(p/\left(1-p\right)\right)=\mathrm{B}_0+\mathrm{B}_1 X$$

Dla wielu predyktorów:

$$\log\left(p/\left(1-p\right)\right)=\mathrm{B}_0+\mathrm{B}_1 X+\mathrm{B}_2 X_2+\ldots+\mathrm{B}_n X_n$$

Gdzie:
log[p/(1-p)] to tzw. logit (logarytm szans na daną kategorię zmiennej Y);
X1, X2, Xn to zmienne objaśniające;
B0 i B1 to współczynniki beta regresji

Regresja logistyczna

- Szacowanie parametrów modelu odbywa się za pomocą metody największej wiarygodności (Maximum Likelihood)
- Rozluźnia większość założeń klasycznego modelu parametrycznego (oprócz założenia o niezależności zmiennych)

Regresja logistyczna

- Szacowanie parametrów modelu odbywa się za pomocą metody największej wiarygodności (Maximum Likelihood)
- Rozluźnia większość założeń klasycznego modelu parametrycznego (oprócz założenia o niezależności zmiennych)
- Diagnostyka obejmuje przede wszystkim dopasowanie do danych (pseudo-R2), porównanie reszt oraz ocenę możliwości predykcyjnych

Regresja logistyczna - podsumowanie

- Model regresji liniowej próbuje zminimalizować resztę. Model regresji logistycznej stara się przewidzieć wynik z jak największą dokładnością po uwzględnieniu wszystkich zmiennych.
- Oblicza prawdopodobieństwo dla każdej obserwacji w zbiorze danych (przewiduje, czy coś się zdarzy lub nie zdarzy)

Regresja logistyczna - podsumowanie

- Model regresji liniowej próbuje zminimalizować resztę. Model regresji logistycznej stara się przewidzieć wynik z jak największą dokładnością po uwzględnieniu wszystkich zmiennych.
- Oblicza prawdopodobieństwo dla każdej obserwacji w zbiorze danych (przewiduje, czy coś się zdarzy lub nie zdarzy)
- Współczynniki modelu, mówią nam, jak bardzo zmienne objaśniające przyczyniają się do prawdopodobieństwa tego, że coś się wydarzy bądź nie

Regresja logistyczna - przykład

- Model dla ryzyka bycia dłużnikiem (Tak/ Nie) na podstawie wybranych cech społeczno-ekonomicznych

Analiza statystyczna danych

Analiza szeregów czasowych

Analiza statystyczna danych

Analiza szeregów czasowych

- Seria czasowa to seria punktów danych, w której każdy punkt danych jest związany ze znacznikiem czasu.
- Szereg czasowy można rozłożyć na jego części składowe, tak aby go zrozumieć, analizować, modelować i prognozować.
- Komponenty szeregów czasowych to: **trend, fluktuacje sezonowe i cykliczne, losowe wahania**

Analiza szeregów czasowych

- Trend to ogólna tendencja do wzrostu lub spadku w długim czasie
- Fluktuacje sezonowe to regularne wahania w okresie 12 miesięcy związane z cyklicznością pór roku bądź działalnością człowieka

Analiza szeregów czasowych

- Trend to ogólna tendencja do wzrostu lub spadku w długim czasie
- Fluktuacje sezonowe to regularne wahania w okresie 12 miesięcy związane z cyklicznością pór roku bądź działalnością człowieka
- Fluktuacje cykliczne są związame z tzw. cyklem gospodarczym
- Losowe wahania są związane z nieprzewidywalnymi i niekontrolowanymi wydarzeniami

Analiza szeregów czasowych

- Szereg czasowy można zapisać jako:

$$y_t = f(t)$$

Gdzie y_t to wartość zmiennej y w czasie t

- Na tej podstawie możemy zapisać model addytywny:

$$y_t = T_t + S_t + C_t + R_t$$

Gdzie T_t, S_t, C_t, R_t to poszczególne komponenty szeregów czasowych

Analiza szeregów czasowych

- Na tej podstawie możemy również zapisać model multiplikatywny:

$$y_t = T_t * S_t * C_t * R_t$$

Gdzie T_t, S_t, C_t, R_t to poszczególne komponenty szeregów czasowych

Analiza szeregów czasowych

- Przykład: dekompozycja szeregu

Analiza statystyczna danych

Stacjonarność szeregów czasowych

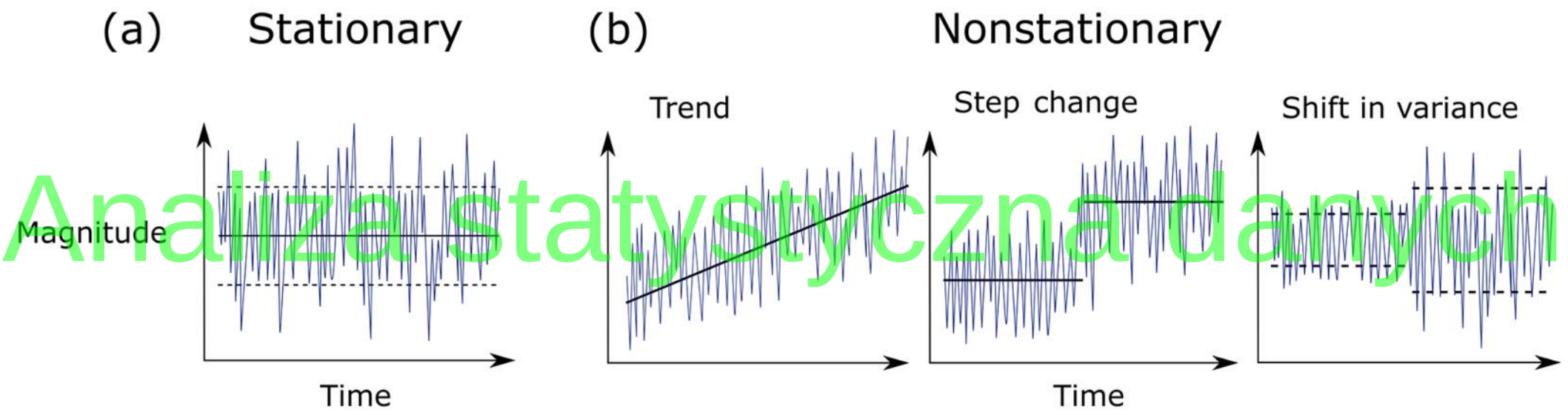
- Stacjonarność to istotne założenie dot. szeregów czasowych związane z ich dalszą analizą
- Szereg czasowy jest uważany za stacjonarny, jeśli spełnia następujące warunki:

  - średnia wartość szeregu czasowego jest stała w czasie (nieważność trendu).

  - wariancja nie wzrasta w czasie.

  - efekt sezonowości jest minimalny.

Stacjonarność szeregów czasowych



Prognozowanie szeregów czasowych

Analiza statystyczna danych

Prognozowanie szeregów czasowych

Główne klasy modeli szeregów czasowych:

- Modele autoregresyjne (AR)

- Modele średniej ruchomej (MA)

- Zintegrowane modele autoregresyjne ze średnią ruchomą (ARIMA)

- Zintegrowane sezonowe modele autoregresyjne ze średnią ruchomą (SARIMA)

Model ARIMA

- model określają 3 parametry: p,d,q

  - *d* to różnice zmiennych (-1,-2,-3, …)

  - *p* to rząd procesu autoregrsyjnego (AR)

  - *q* to rząd średniej ruchomej (MA)

Model ARIMA

- model określają 3 parametry: p,d,q

  - *d* to różnice zmiennych (-1,-2,-3, …)

  - *p* to rząd procesu autoregresyjnego (AR)

  - *q* to rząd średniej ruchomej (MA)

- Po ustaleniu specyfikacji modelu, szacujemy jego parametry
- W estymacji wykorzystuje się najczęściej metodę największej wiarygodności (ML)

Egzamin końcowy

- I termin: 31.01.2023, g.10, sala 21
- II termin: 14.02.2023, g.10, sala 21
- Obowiązuje materiał z wykładów (prezentacja, skrypty, dodatkowe istotne omówienia)

Około 20 pytań (pytania otwarte i zamknięte)

Pytania otwarte będą zawierać konkretne problemy, a rozwiązanie będzie polegać na znalezieniu najlepszego rozwiązania i jego uzasadnieniu.