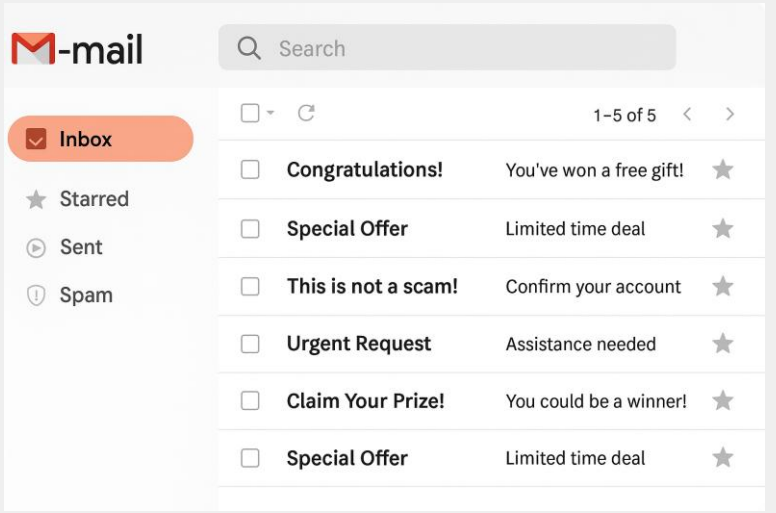
Przetwarzanie i wizualizacja danych

06.08.2025

Do czego możemy użyć AI?


-mail
Search
Inbox
Starred
Sent
Spam
1–5 of 5
Congratulations!
You've won a free gift!
Special Offer
Limited time deal
This is not a scam!
Confirm your account
Urgent Request
Assistance needed
Claim Your Prize!
You could be a winner!
Special Offer
Limited time deal

Scenariusz: manualna anotacja

Amazon <unexercisable@d505.fzpaunso.us>
to me

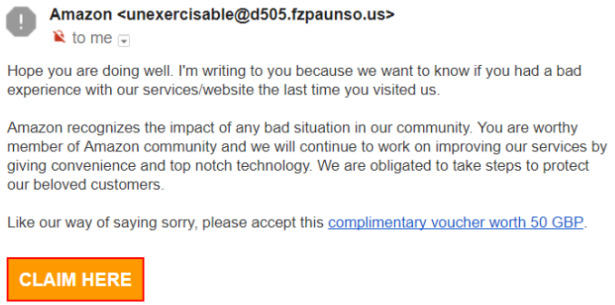
Hope you are doing well. I'm writing to you because we want to know if you had a bad experience with our services/website the last time you visited us.

Amazon recognizes the impact of any bad situation in our community. You are worthy member of Amazon community and we will continue to work on improving our services by giving convenience and top notch technology. We are obligated to take steps to protect our beloved customers.

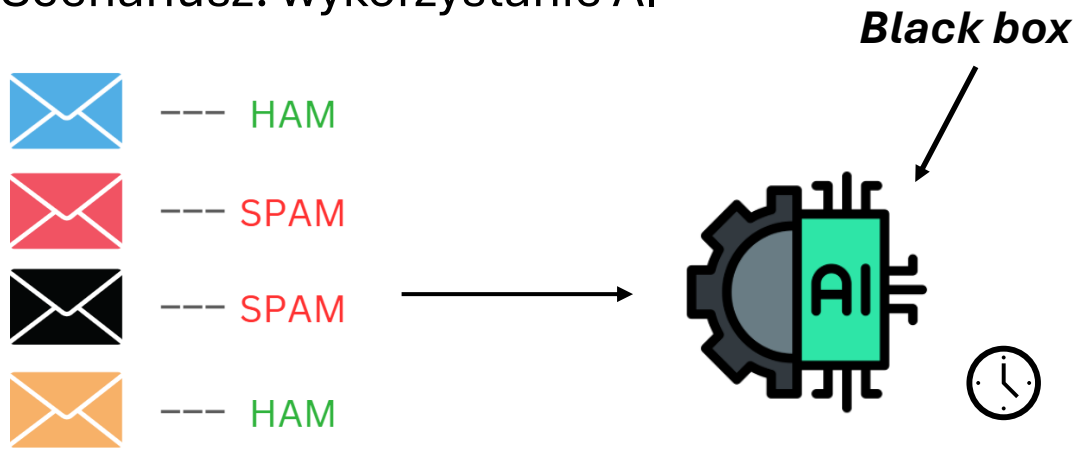
Like our way of saying sorry, please accept this complimentary voucher worth 50 GBP.

CLAIM HERE

Scenariusz: manualna anotacja

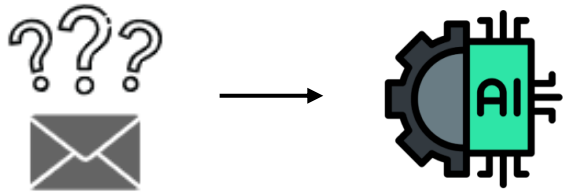


Scenariusz: wykorzystanie AI



Etykietowane dane

Trenowanie klasyfikatora

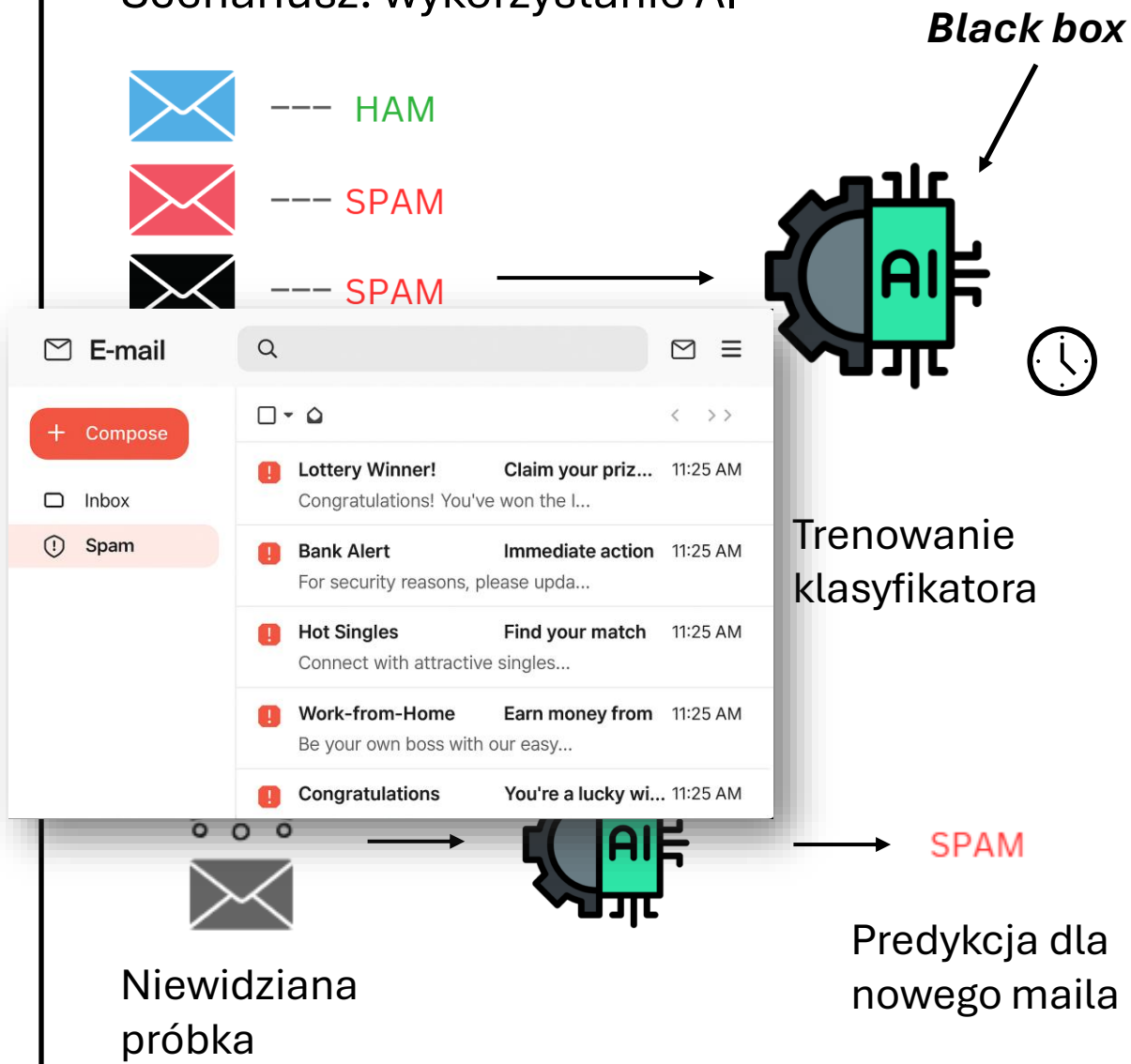


Niewidziana próbka

Predykcja dla nowego maila

Scenariusz: manualna anotacja

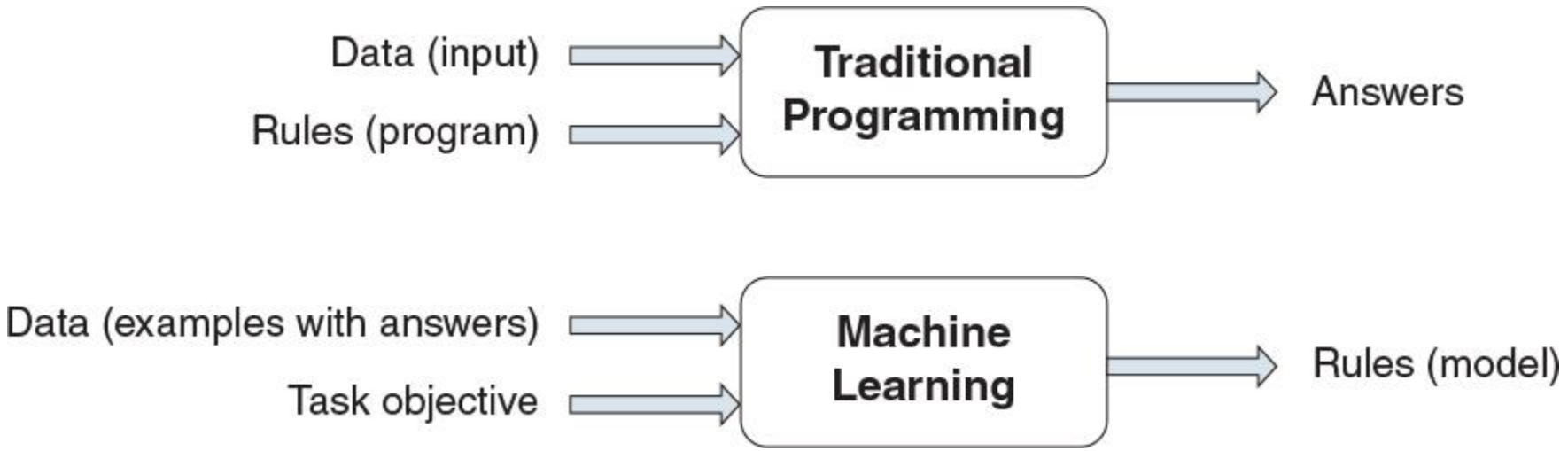
Scenariusz: wykorzystanie AI




Czym różni się machine learning od tradycyjnego programowania?

?

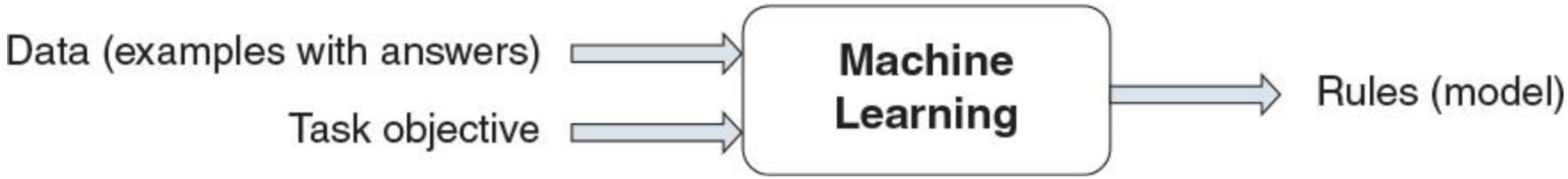
ML vs Tradycyjne programowanie



Tradycyjne programowanie


```
funkcja klasyfikuj_email(tresc_emaila):
    jeśli "darmowy" w tresc_emaila:
        zwróć "spam"
    jeśli "wygraj nagrodę" w tresc_emaila:
        zwróć "spam"
    jeśli "kliknij tutaj" w tresc_emaila:
        zwróć "spam"
    w przeciwnym razie:
        zwróć "nie-spam"
```

Uczenie maszynowe




```
# Trening modelu
DANE_TRENINGOWE = zbiór e-maili oznaczonych jako SPAM lub NIE_SPAM
MODEL = NAUCZ(NaiwnyBayes, DANE_TRENINGOWE)

# Predykcja dla nowej wiadomości
NOWY_EMAIL = "Otrzymałeś darmowy bilet, kliknij tutaj!"
KLASA = MODEL.PREDYKCJA(NOWY_EMAIL)

jeśli KLASA == SPAM:
    oznacz jako spam
inaczej:
    zostaw w skrzynce odbiorczej
```

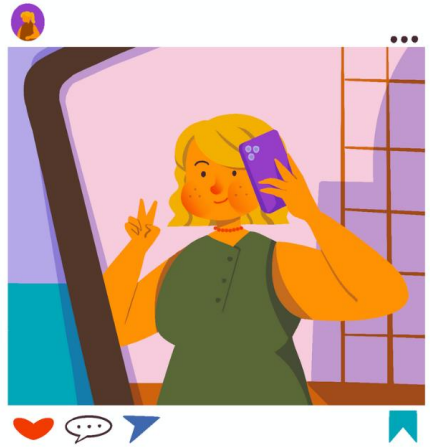
A co to są dane?

Dana to zapisana w określonej formie informacja o jakimś zjawisku, obiekcie lub procesie

A co to są dane?

Dana to zapisana w określonej formie informacja o jakimś zjawisku, obiekcie lub procesie

Wypożyczenie książki
- tytuł, data, kod biblioteczny

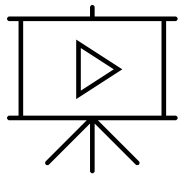


Post na social media
- treść, tagi, nazwa użytkownika

Modalności AI

„surowe” dane

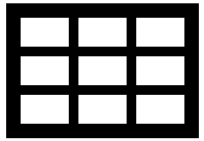
audio



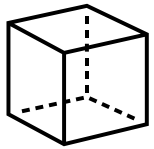
video

tekst

obrazki



tabela



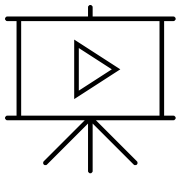
3D

Modalności AI

Więcej struktury ~ mniej wysiłku w przygotowanie danych

audio

Próbkowanie, kwantyzacja, transformata Fouriera



video

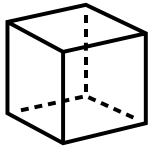
Zmiana częstości ramek, filtrowanie

tekst

Zamiana na lowercase, usuwanie „stop words”, tokenizacja

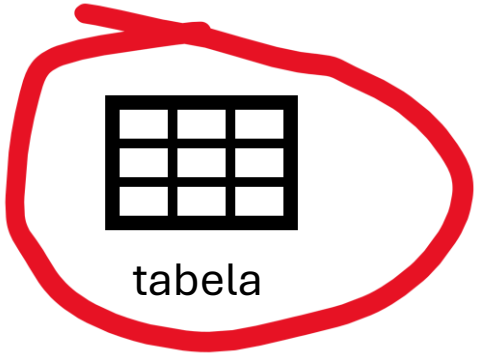
obrazki

Przycięcie, zmiana skali wartości



3D

Próbkowanie, zamiana siatki na chmurę punktów



tabela

ustrukturyzowane

Pierwsze zapoznanie z danymi

Wymiarowość : Każda obserwacja opisana jest przez 6 **zmiennych**

Wielkość: Mamy 10 **obserwacji (próbek)**

	species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g
0	Adelie	Torgersen	39.1	18.7	181.0	3750.0
1	Adelie	Torgersen	39.5	17.4	186.0	3800.0
2	Adelie	Torgersen	40.3	18.0	195.0	3250.0
3	Adelie	Torgersen	NaN	NaN	NaN	NaN
4	Adelie	Torgersen	36.7	19.3	193.0	3450.0
5	Adelie	Torgersen	39.3	20.6	190.0	3650.0
6	Adelie	Torgersen	38.9	17.8	181.0	3625.0
7	Adelie	Torgersen	39.2	19.6	195.0	4675.0
8	Adelie	Torgersen	34.1	18.1	193.0	3475.0
9	Adelie	Torgersen	42.0	20.2	190.0	4250.0

Pierwsze zapoznanie z danymi

Wymiarowość : Każda obserwacja opisana jest przez 6 **zmiennych**

Wielkość: Mamy 10 **obserwacji (próbek)**

	species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g
0	Adelie	Torgersen	39.1	18.7	181.0	3750.0
1	Adelie	Torgersen	39.5	17.4	186.0	3800.0
2	Adelie	Torgersen	40.3	18.0	195.0	3250.0
3	Adelie	Torgersen	NaN	NaN	NaN	NaN
4	Adelie	Torgersen	36.7	19.3	193.0	3450.0
5	Adelie	Torgersen	39.3	20.6	190.0	3650.0
6	Adelie	Torgersen	38.9	17.8	181.0	3625.0
7	Adelie	Torgersen	39.2	19.6	195.0	4675.0
8	Adelie	Torgersen	34.1	18.1	193.0	3475.0
9	Adelie	Torgersen	42.0	20.2	190.0	4250.0

species i island to zmienne **kategorialne**

bill_length_mm, bill_depth_mm, flipper_length_mm, body_mass_g to zmienne **numeryczne**

Pierwsze zapoznanie z danymi

Wymiarowość : Każda obserwacja opisana jest przez 6 **zmiennych**

	species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g
0	Adelie	Torgersen	39.1	18.7	181.0	3750.0
1	Adelie	Torgersen	39.5	17.4	186.0	3800.0
2	Adelie	Torgersen	40.3	18.0	195.0	3250.0
3	Adelie	Torgersen	NaN	NaN	NaN	NaN
4	Adelie	Torgersen	36.7	19.3	193.0	3450.0
5	Adelie	Torgersen	39.3	20.6	190.0	3650.0
6	Adelie	Torgersen	38.9	17.8	181.0	3625.0
7	Adelie	Torgersen	39.2	19.6	195.0	4675.0
8	Adelie	Torgersen	34.1	18.1	193.0	3475.0
9	Adelie	Torgersen	42.0	20.2	190.0	4250.0

Wielkość: Mamy 10 **obserwacji (próbek)**

species	
Adelie	152
Chinstrap	68
Gentoo	124

island	
Biscoe	168
Dream	124
Torgersen	52

species i island to zmienne **kategorialne**

bill_length_mm, bill_depth_mm, flipper_length_mm, body_mass_g to zmienne **numeryczne**

	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g
mean	43.921930	17.151170	200.915205	4201.754386
std	5.459584	1.974793	14.061714	801.954536

Pierwsze zapoznanie z danymi

Wymiarowość : Każda obserwacja opisana jest przez 6 **zmiennych**

Wielkość: Mamy 10 **obserwacji (próbek)**

	species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g
0	Adelie	Torgersen	39.1	18.7	181.0	3750.0
1	Adelie	Torgersen	39.5	17.4	186.0	3800.0
2	Adelie	Torgersen	40.3	18.0	195.0	3250.0
3	Adelie	Torgersen	NaN	NaN	NaN	NaN
4	Adelie	Torgersen	36.7	19.3	193.0	3450.0
5	Adelie	Torgersen	39.3	20.6	190.0	3650.0
6	Adelie	Torgersen	38.9	17.8	181.0	3625.0
7	Adelie	Torgersen	39.2	19.6	195.0	4675.0
8	Adelie	Torgersen	34.1	18.1	193.0	3475.0
9	Adelie	Torgersen	42.0	20.2	190.0	4250.0

Występują wartości brakujące

species i island to zmienne **kategorialne**

bill_length_mm, bill_depth_mm, flipper_length_mm, body_mass_g to zmienne **numeryczne**

Dane mają różny zakres wartości

	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g
min	32.1	13.1	172.0	2700.0
max	59.6	21.5	231.0	6300.0

Pierwsze zapoznanie z danymi

Predyktory, zmienne objaśniające

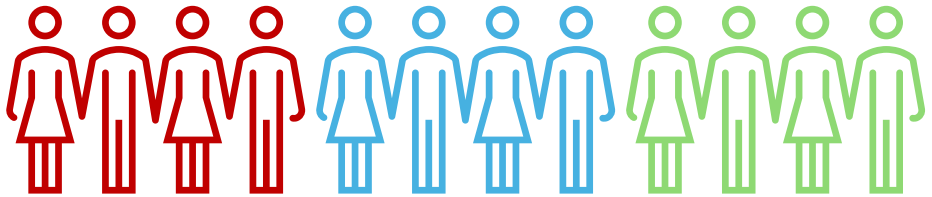
	species
0	Adelie
1	Adelie
2	Adelie
3	Adelie
4	Adelie
5	Adelie
6	Adelie
7	Adelie
8	Adelie
9	Adelie

island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g
Torgersen	39.1	18.7	181.0	3750.0
Torgersen	39.5	17.4	186.0	3800.0
Torgersen	40.3	18.0	195.0	3250.0
Torgersen	NaN	NaN	NaN	NaN
Torgersen	36.7	19.3	193.0	3450.0
Torgersen	39.3	20.6	190.0	3650.0
Torgersen	38.9	17.8	181.0	3625.0
Torgersen	39.2	19.6	195.0	4675.0
Torgersen	34.1	18.1	193.0	3475.0
Torgersen	42.0	20.2	190.0	4250.0

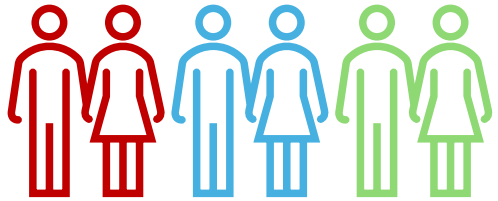
Klasa (zadanie klasyfikacji)

Ile danych potrzeba?

Populacja



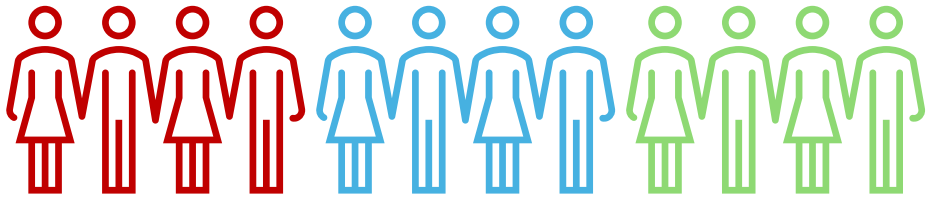
Próbka



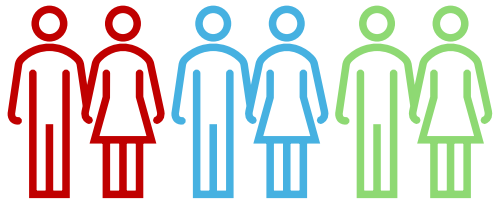
Próbka powinna być reprezentatywna!

Ile danych potrzeba?

Populacja

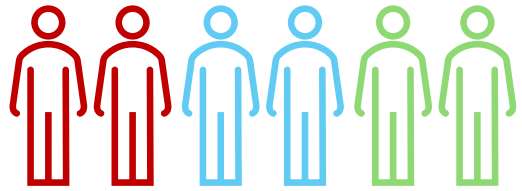


Próbka



Próbka powinna być reprezentatywna!

Próbka



?

Predykcja wyników studentów
~
1,000
(obserwacji)

Klasyfikacja obrazów
~
10,000

Analiza sentymentu modelem językowym
~
1 000 000 000
(słów)

Dane bez anotacji (klas) są szeroko dostępne

Eksploracyjna analiza danych – po co i na co?

EDA (*Exploratory Data Analysis*)

„Let the data speak” – nie szukamy potwierdzenia konkretnego zjawiska na tym etapie

Celem jest zastosowanie technik statystycznych, grupowania i wizualizacji w celu odkrycia struktury i modelu w danych

According to Howard Seltman (Carnegie Mellon University), “loosely speaking, any method of looking at data that does not include formal statistical modeling and inference falls under the term exploratory data analysis”

Eksploracyjna analiza danych – po co i na co?

EDA (*Exploratory Data Analysis*)

„Let the data speak” – nie szukamy potwierdzenia konkretnego zjawiska na tym etapie

Celem jest zastosowanie technik statystycznych, grupowania i wizualizacji w celu odkrycia struktury i modelu w danych

Jak zmienne są ze sobą związane?

Czy w danych są pewne anomalie?

Które zmienne mogą być istotne?

According to Howard Seltman (Carnegie Mellon University), “loosely speaking, any method of looking at data that does not include formal statistical modeling and inference falls under the term exploratory data analysis”

Eksploracyjna analiza danych – po co i na co?

EDA

≠

Wizualizacja danych!

Mimo, że EDA w większości używa techniki wizualizacji.

Praktyczny przykład

Dane na temat użytkowników

Id	Miasto	Płeć	Wiek	Subskrybuje_aktualnie
101	Chicago, U.S	K	23	Tak
102	CHICAGO, US	M	68	Nie
103	New York	NaN	19	Tak
104	Warsaw, Poland	K	125	Nie
105	chicago, us	M	24	Tak

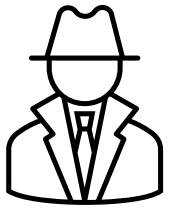
Załóżmy, że pracujemy w jakiejś firmie,
która streamuje filmy/seriale

Jak możemy zwiększyć liczbę subskrybentów?

Dane z wyników ankiety

Id	Strona	Kliknięcie
101	A	1
102	B	0
104	C	1
106	B	NaN
999	A	0

Praktyczny przykład



Dane na temat użytkowników

Id	Miasto	Płeć	Wiek	Subskrybuje_aktualnie
101	Chicago, U.S	K	23	Tak
102	CHICAGO, US	M	68	Nie
103	New York	NaN	19	Tak
104	Warsaw, Poland	K	125	Nie
105	chicago, us	M	24	Tak

Załóżmy, że pracujemy w jakiejś firmie,
która streamuje filmy/seriale

?

Dane z wyników ankiety

Id	Strona	Kliknięcie
101	A	1
102	B	0
104	C	1
106	B	NaN
999	A	0

Praktyczny przykład

Ta wartość to prawdopodobnie jakiś błąd

Id	*Miasto*	*Płeć*	Wiek	*Strona*	Kliknięcie	Subskrybuje_aktualnie
101	Chicago, U.S	K	23	A	1	Tak
102	CHICAGO, US	M	68	B	0	Nie
103	New York	NaN	19	C	1	Tak
104	Warsaw, Poland	K	125	NaN	NaN	Nie
105	chicago, us	M	24	NaN	NaN	Tak

Wymaga jednolitego formatu

Praktyczny przykład

Ktoś nie podał swojej płci – może to jakaś informacja?

Ta wartość to prawdopodobnie jakiś błąd

Id	*Miasto*	*Płeć*	Wiek	*Strona*	Kliknięcie	Subskrybuje_aktualnie
101	Chicago, U.S	K	23	A	1	Tak
102	CHICAGO, US	M	68	B	0	Nie
103	New York	NaN	19	C	1	Tak
104	Warsaw, Poland	K	125	NaN	NaN	Nie
105	chicago, us	M	24	NaN	NaN	Tak

Wymaga jednolitego formatu

Braki wynikają z połączenia po indeksach tabeli „po lewej”

Praktyczny przykład

Ktoś nie podał swojej płci – może to jakaś informacja?

Ta wartość to prawdopodobnie jakiś błąd

Id	*Miasto*	*Płeć*	Wiek	*Strona*	Kliknięcie	Subskrybuje_aktualnie
101	Chicago, U.S	K	23	A	1	Tak
102	CHICAGO, US	M	68	B	0	Nie
103	New York	NaN	19	C	1	Tak
104	Warsaw, Poland	K	125	NaN	NaN	Nie
105	chicago, us	M	24	NaN	NaN	Tak

Wymaga jednolitego formatu

Braki wynikają z połączenia po indeksach tabeli „po lewej”

Kolumnę docelową możemy zamienić na 0/1

Praktyczny przykład

Id	Miasto	Płeć	Wiek	Strona	Kliknięcie	Subskrybuje_aktualnie
101	Chicago	K	23	A	1	1
102	Chicago	M	68	B	0	0
103	New York	O	19	C	1	1
104	Warsaw	K	35	X	0	0
105	Chicago	M	24	X	0	1

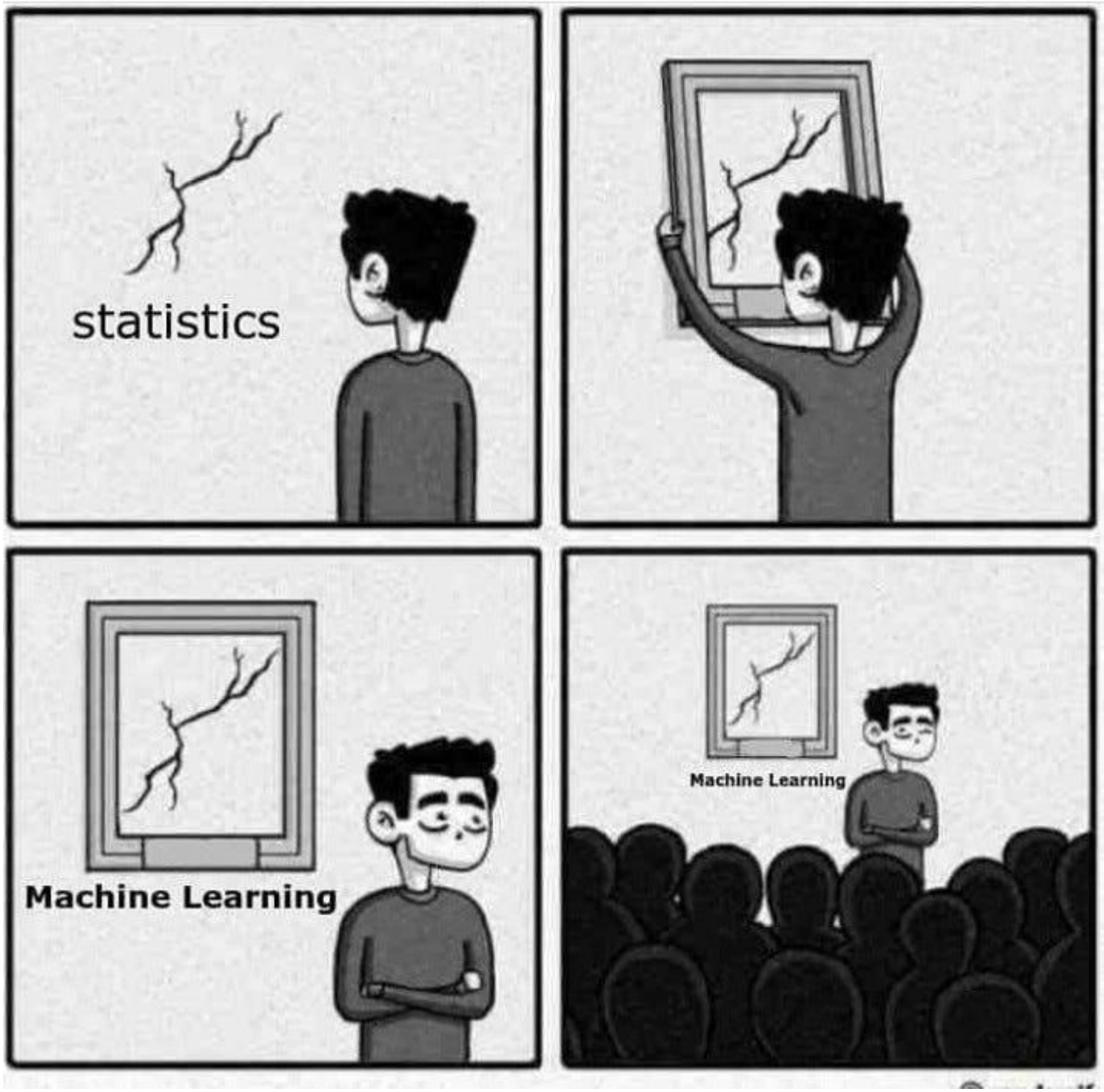
Praktyczny przykład

Id	Miasto	Płeć	Wiek	Strona	Kliknięcie	Subskrybuje_aktualnie
101	Chicago	K	23	A	1	1
102	Chicago	M	68	B	0	0
103	New York	O	19	C	1	1
104	Warsaw	K	35	X	0	0
105	Chicago	M	24	X	0	1

Jak możemy zwiększyć liczbę subskrybentów?

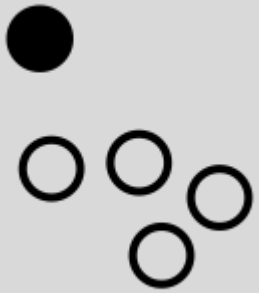
- Jaki jest profil typowego subskrybenta? (np. czy to kobieta w średnim wieku z New York?)
- Jakie są cechy nie-subskrybentów, którzy wzięli udział w ankiecie?
- Czy potrzebujemy więcej danych z jakiejś konkretnej grupy, aby wyciągnąć sensowny wniosek?
- Braki w danych i błędy występują zarówno dla subskrybujących jak i nie?

Przerwa

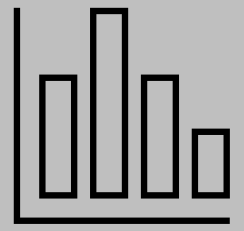


Wizualizacja w danych

Detekcja outlierów



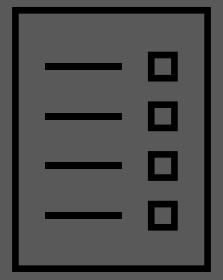
Badanie równowagi klas



Wykrywanie błędów

Rozumienie zależności i wzorców

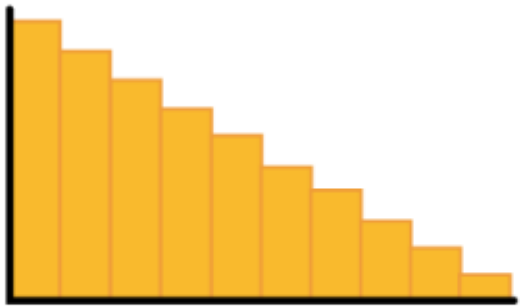
Analiza wyników



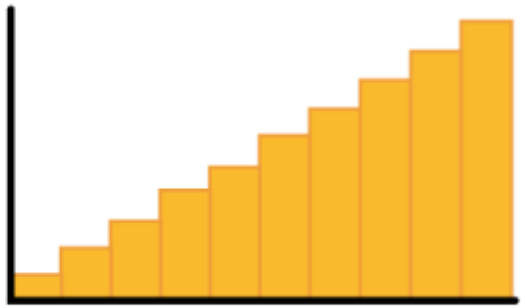
Podział rozkładów danych

Symmetric (normal) vs skewed and uniform distriutions

Normal distribution
(unimodal, symmetric,
the “bell curve”)

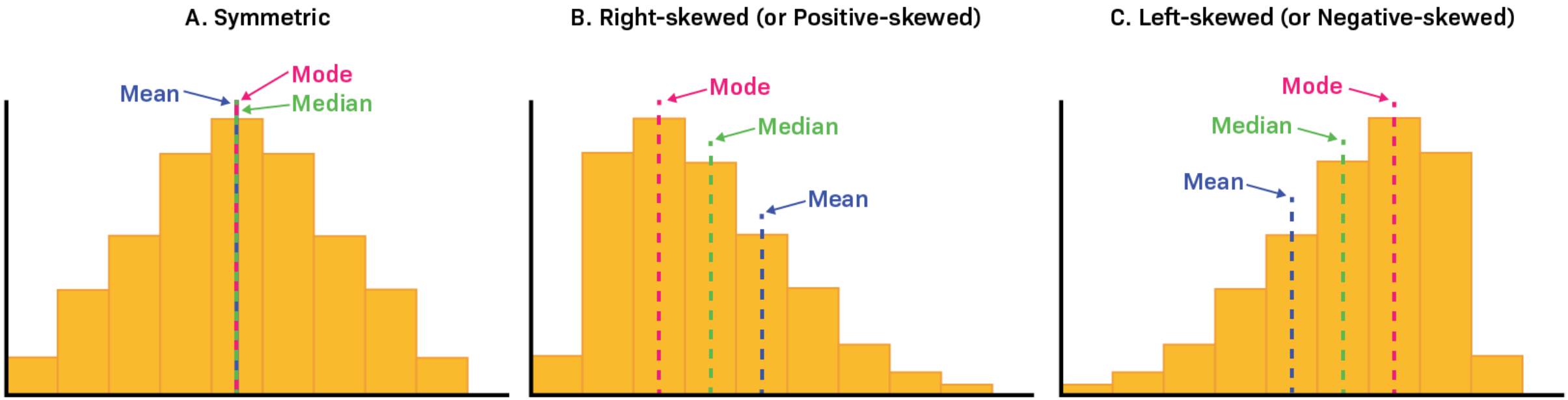


Right-skewed distribution
(Positively-skewed)



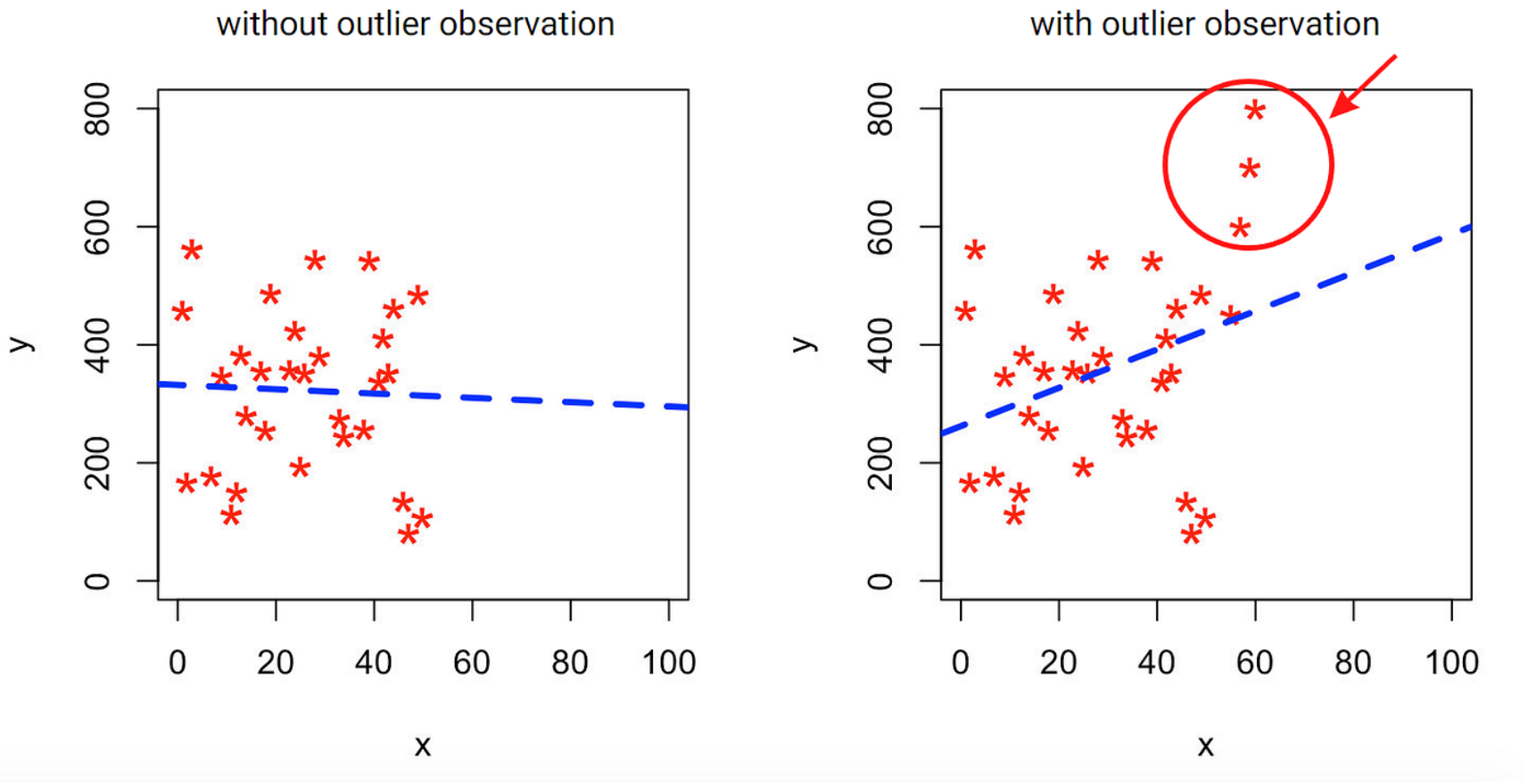
Left-skewed distribution
(Negatively-skewed)

Uniform distribution
(equal spread,
no peaks)

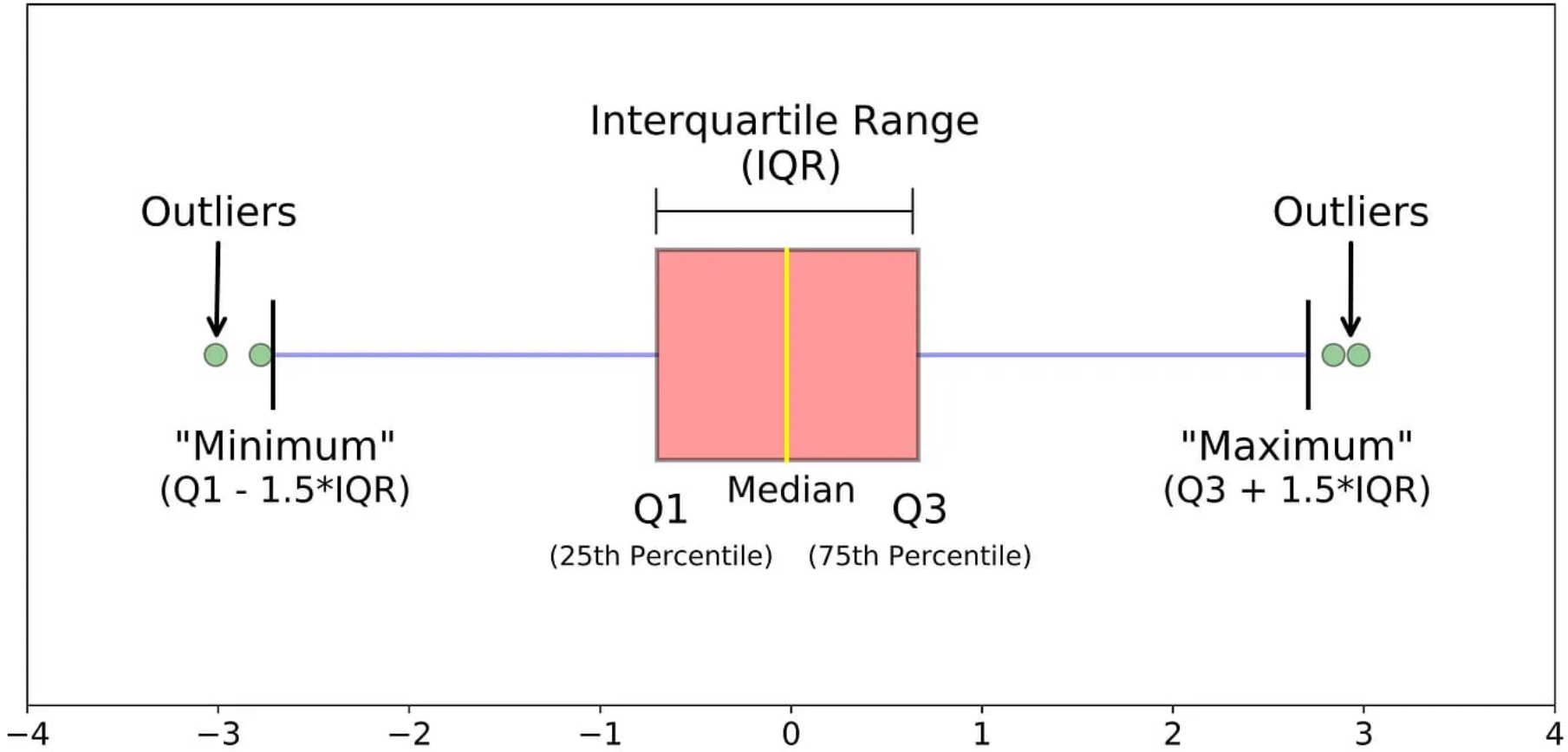


Moda – najczęściej występująca wartość
Mediana – wartość środkowa w uporządkowanym zbiorze danych
Średnia - suma wszystkich wartości podzielona przez ich liczbę

Wartości odstające



outlier - obserwacja, która znacząco odstaje od reszty danych



1.5 * IQR – to standardowy próg, najczęściej używany

3.0 * IQR - to próg używany do wykrywania ekstremalnych outlierów

Co zrobić z outlierami?

zostawić - jeśli są prawdziwe i istotne

flooring – ograniczyć wartość poniżej progu

capping – ograniczyć wartość powyżej górnego progu

przekształcić – aby lepiej dopasować dane do skali

zastąpić – medianą lub średnią, kiedy dane są błędne

Uwaga! Niezbalansowane dane!

accuracy
staje się mniej istotne

ignorowanie mniej licznej klasy

faworyzacja bardziej licznej klasy

niedokładne testowanie

Czy brak balansu zawsze jest zły?

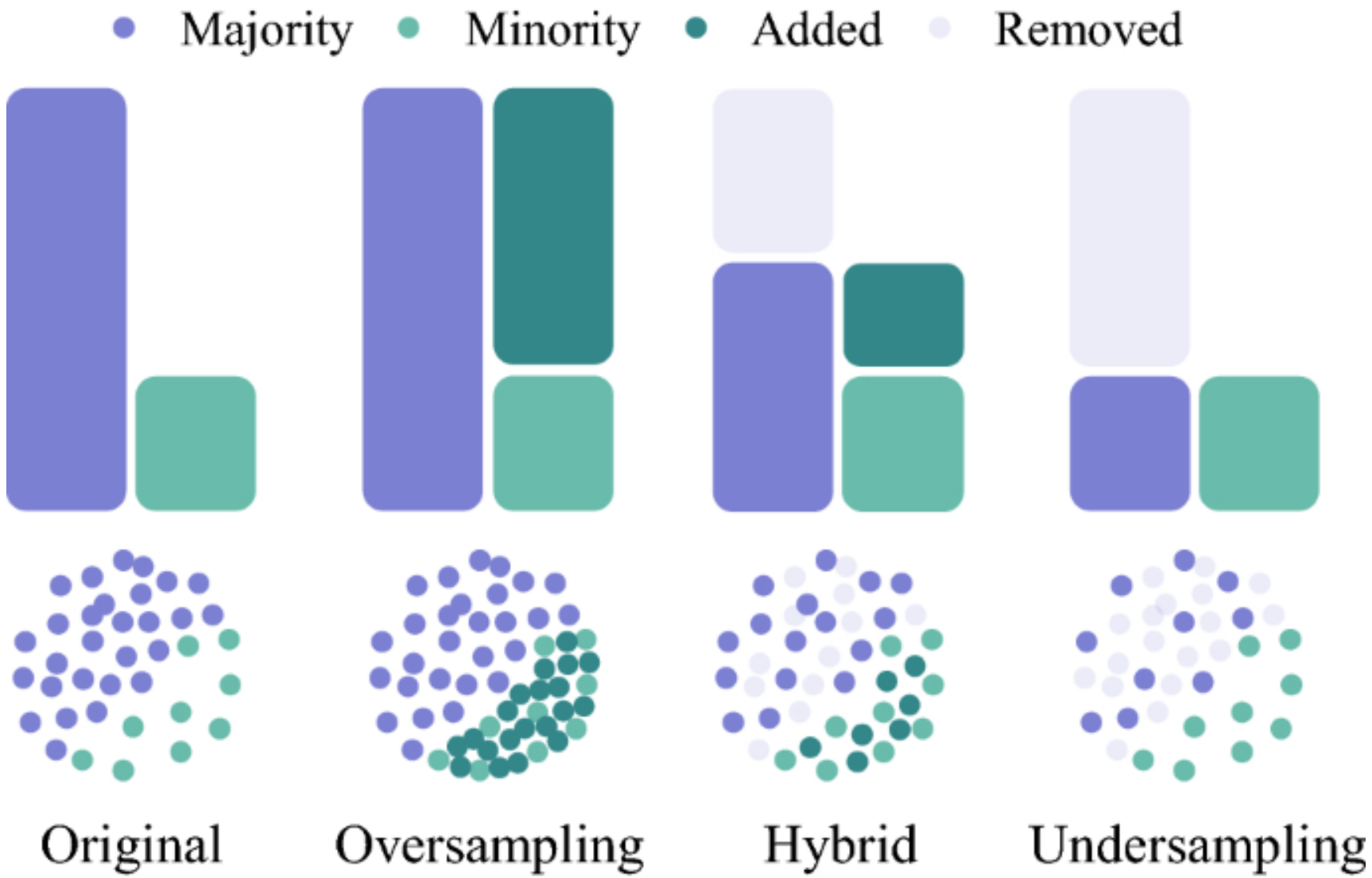
950 kotków

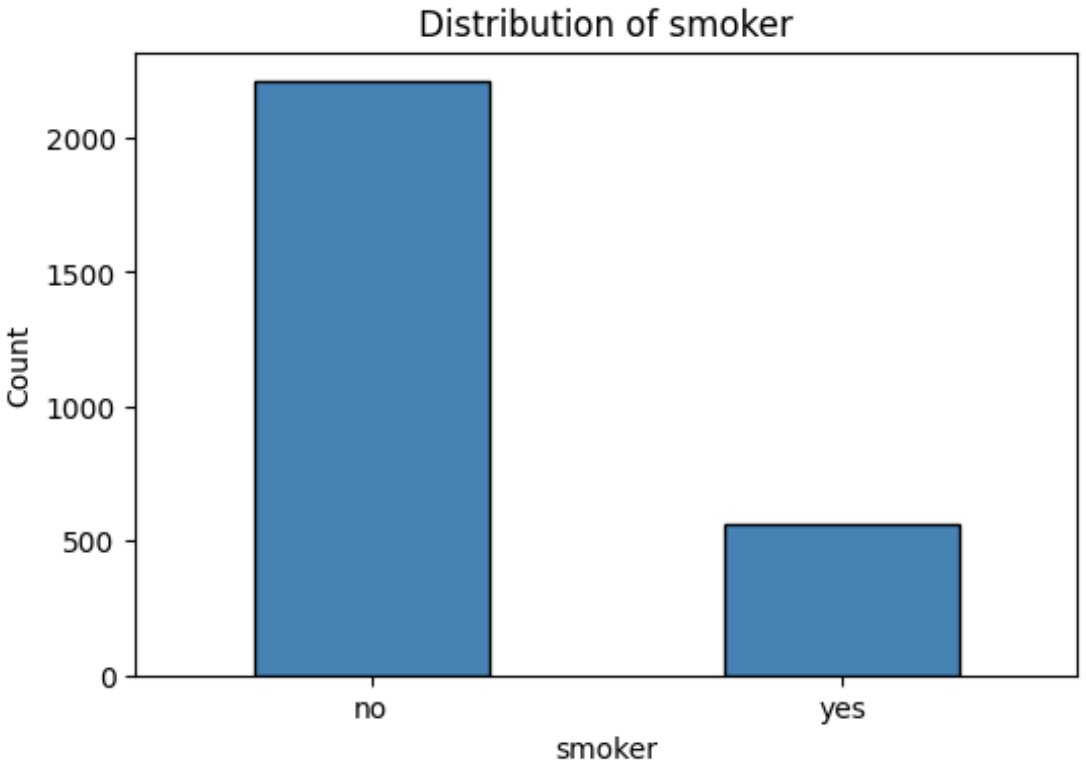
Model dostaje 95 zdjęć kotków i 5 piesków, ale...

Przewiduje, że **wszystkie to kotki!**

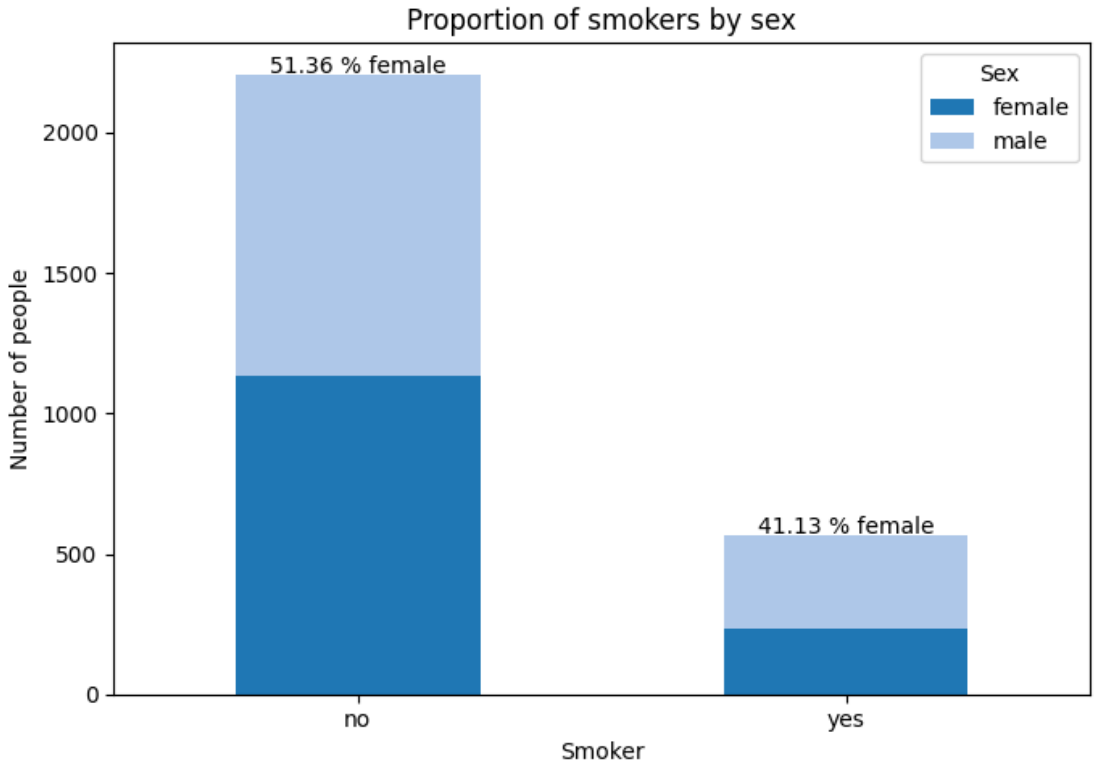
Accuracy = 95%, ale... **model nie rozpoznaje żadnego pieska!**

50 piesków


Majority
Minority
Added
Removed
Original
Oversampling
Hybrid
Undersampling

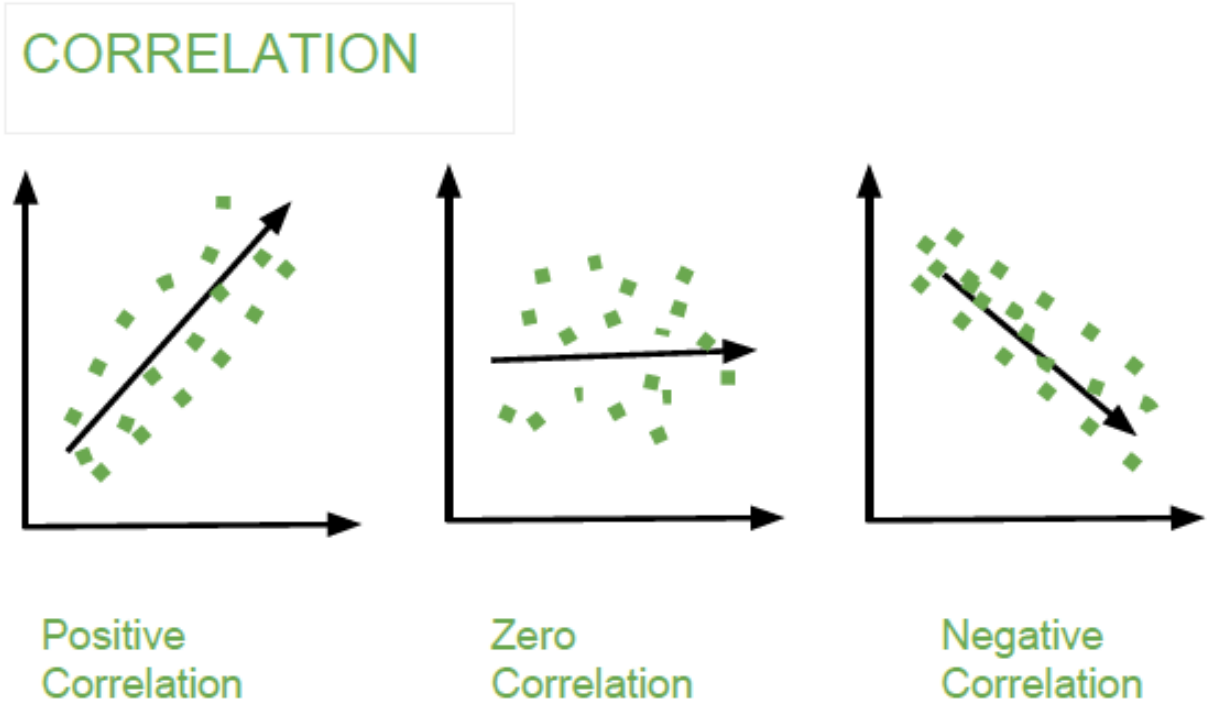


1 zmienna kategoryczna



Kilka zmiennych na jednym wykresie słupkowym!

Zmienna nie jest sama – czyli o zależnościach



$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Where,

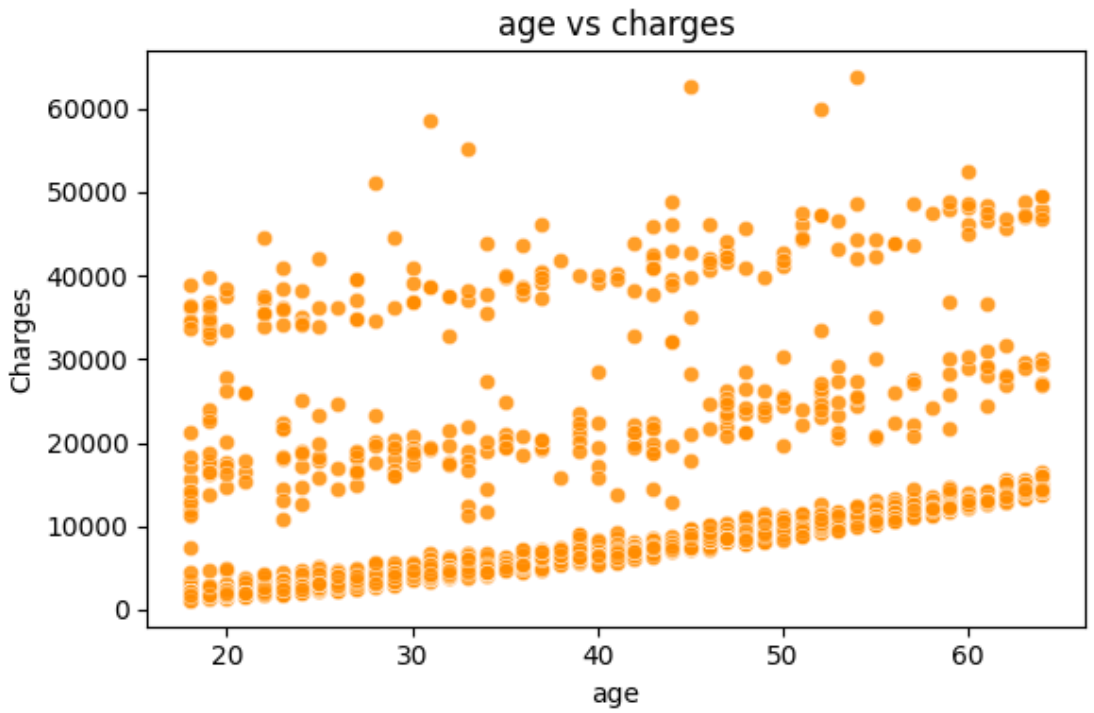
r = Pearson Correlation Coefficient

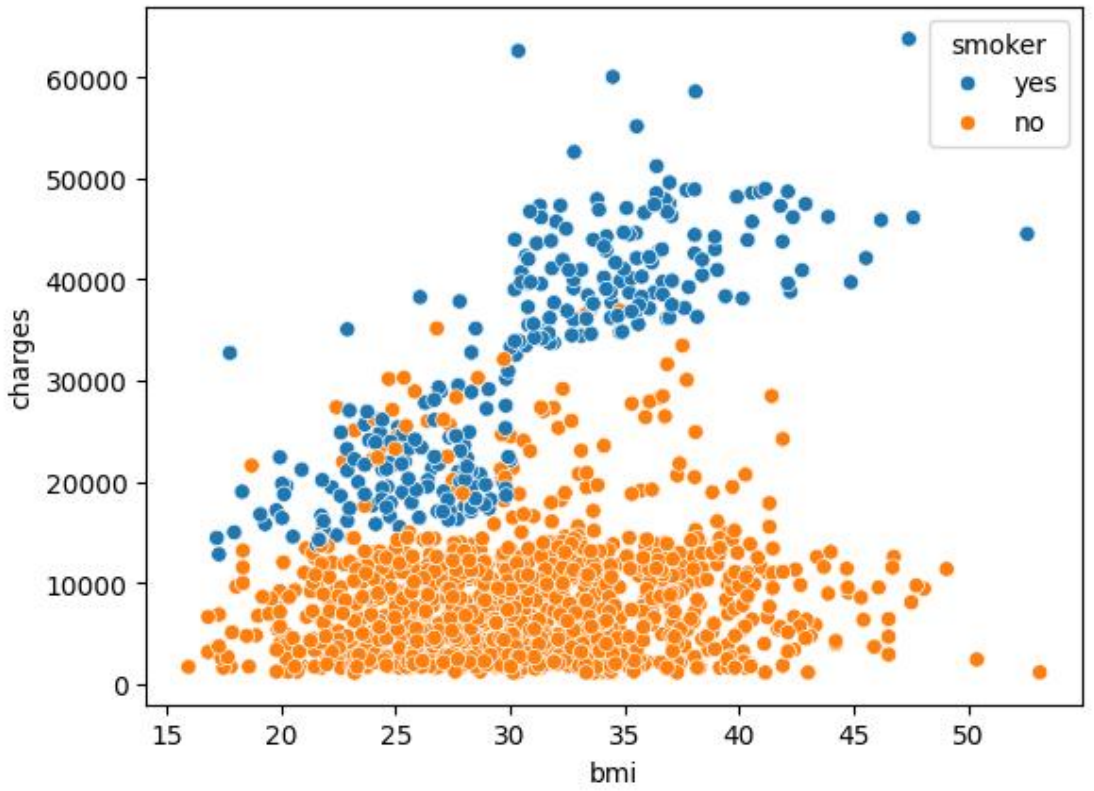
x_i = x variable samples

y_i = y variable sample

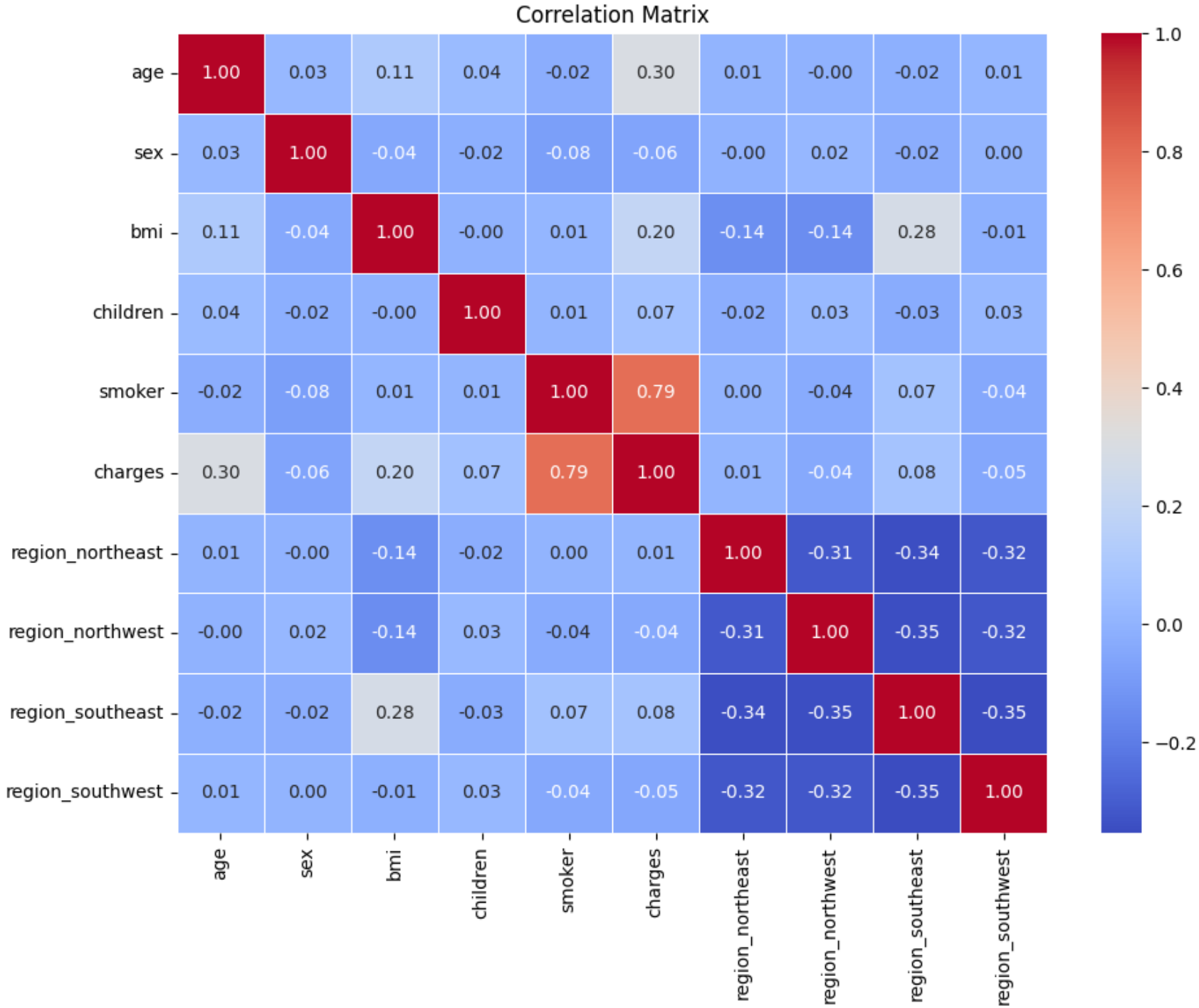
$\bar{x}$ = mean of values in x variable

$\bar{y}$ =mean of values in y variable





Dawanie kilku zmiennych na wykresie daje ciekawe rezultaty :)


Correlation Matrix
age 1.00 0.03 0.11 0.04 -0.02 0.30 0.01 -0.00 -0.02 0.01
sex 0.03 1.00 -0.04 -0.02 -0.08 -0.06 -0.00 0.02 -0.02 0.00
bmi 0.11 -0.04 1.00 -0.00 0.01 0.20 -0.14 -0.14 0.28 -0.01
children 0.04 -0.02 -0.00 1.00 0.01 0.07 -0.02 0.03 -0.03 0.03
smoker -0.02 -0.08 0.01 0.01 1.00 0.79 0.00 -0.04 0.07 -0.04
charges 0.30 -0.06 0.20 0.07 0.79 1.00 0.01 -0.04 0.08 -0.05
region_northeast 0.01 -0.00 -0.14 -0.02 0.00 0.01 1.00 -0.31 -0.34 -0.32
region_northwest -0.00 0.02 -0.14 0.03 -0.04 -0.04 -0.31 1.00 -0.35 -0.32
region_southeast -0.02 -0.02 0.28 -0.03 0.07 0.08 -0.34 -0.35 1.00 -0.35
region_southwest 0.01 0.00 -0.01 0.03 -0.04 -0.05 -0.32 -0.32 -0.35 1.00
age
sex
bmi
children
smoker
charges
region_northeast
region_northwest
region_southeast
region_southwest
1.0
0.8
0.6
0.4
0.2
0.0
−0.2

To my :))

- Julia Farganus – 266564@student.pwr.edu.pl
- Julia Słowińska – 268313@student.pwr.edu.pl

I oczywiście pv na discordzie