



GHOST

Group of Horribly Optimistic Statisticians



Intro to ML

#2 Dane

kluczowy element w uczeniu maszynowym





GHOST

Group of Horribly Optimistic Statisticians





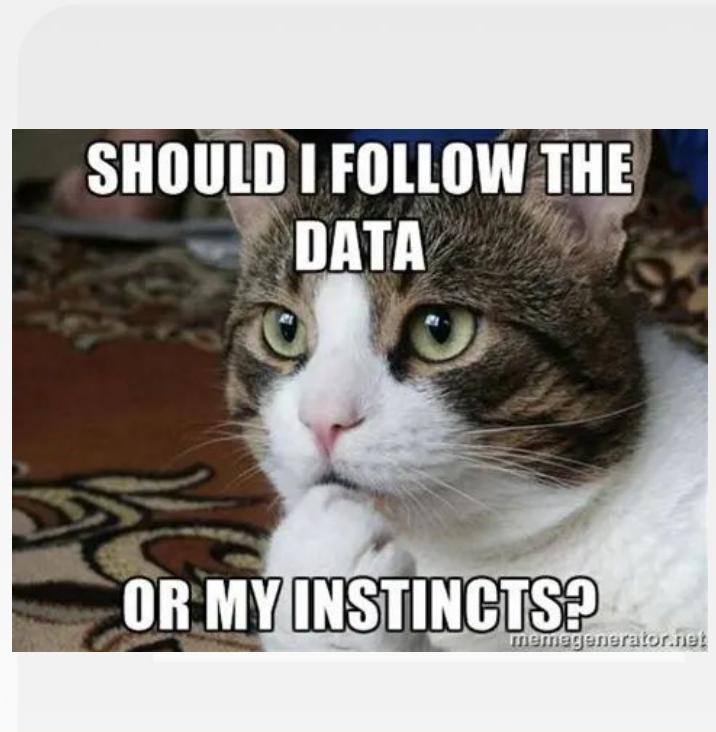
GHOST

Group of Horribly Optimistic Statisticians



Agenda

1. Czym są dane?
2. Typy danych
3. Skąd pochodzą dane?
4. Dane surowe a dane przetworzone
5. Przetwarzanie danych
6. Podział Danych
7. Rozkład Danych
8. Zależności między danymi





GHOST

Group of Horribly Optimistic Statisticians



Czym są dane?

Zbiory wartości, które przekazują informacje, opisując ilość, jakość, fakt, statystyki, inne znaczenia lub sekwencje symboli, które mogą być dalej interpretowane i przetwarzane



eurostat



GHOST

Group of Horribly Optimistic Statisticians



Maszyna uczy się na podstawie danych

Uczenie maszynowe polega na analizie danych i poszukiwaniu w nich zależności.

*Model nie ma wiedzy od początku,
wszystkiego musi się nauczyć.*





GHOST

Group of Horribly Optimistic Statisticians



Maszyna uczy się na podstawie danych

Bez danych model nie istnieje i nie jest w stanie wykonywać żadnych zadań.

Jakość danych ma bezpośredni wpływ na skuteczność algorytmu.

Im lepsze dane, tym bardziej trafne przewidywania modelu.



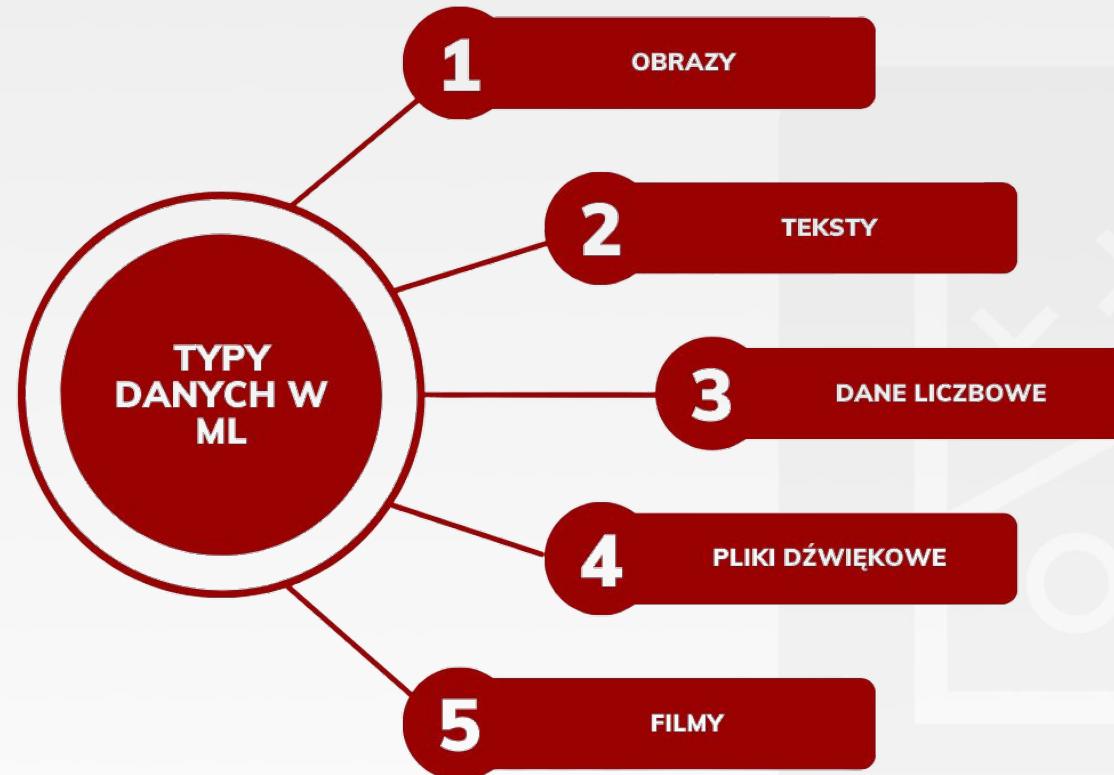


GHOST

Group of Horribly Optimistic Statisticians



Typy danych w uczeniu maszynowym





GHOSH

Group of Horribly Optimistic Statisticians



Skąd pochodzą dane?





GHOST

Group of Horribly Optimistic Statisticians



Dane surowe a dane przetworzone

Dane surowe – raw data – to informacje zebrane bez żadnej obróbki.

Mogą one zawierać błędy, braki, czy być niespójne.

Zanim użyjemy ich w uczeniu maszynowym, musimy je przetworzyć i oczyścić.

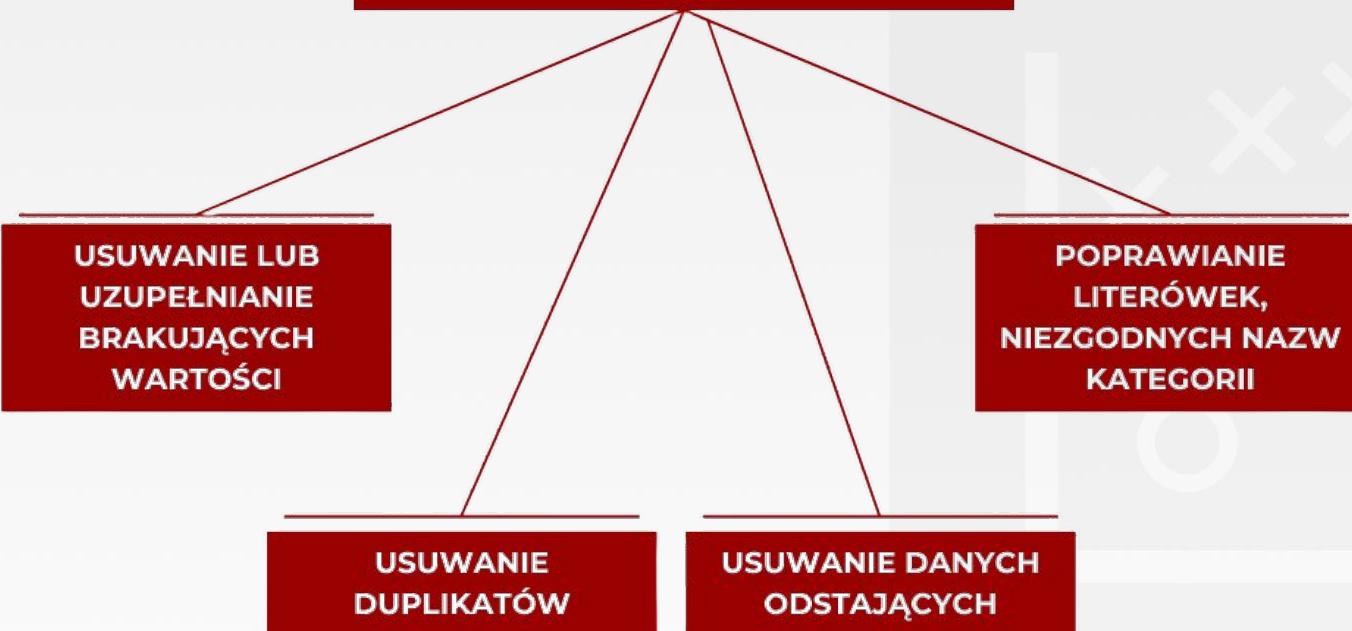


GHOSH

Group of Horribly Optimistic Statisticians



OCZYSZCZANIE DANYCH



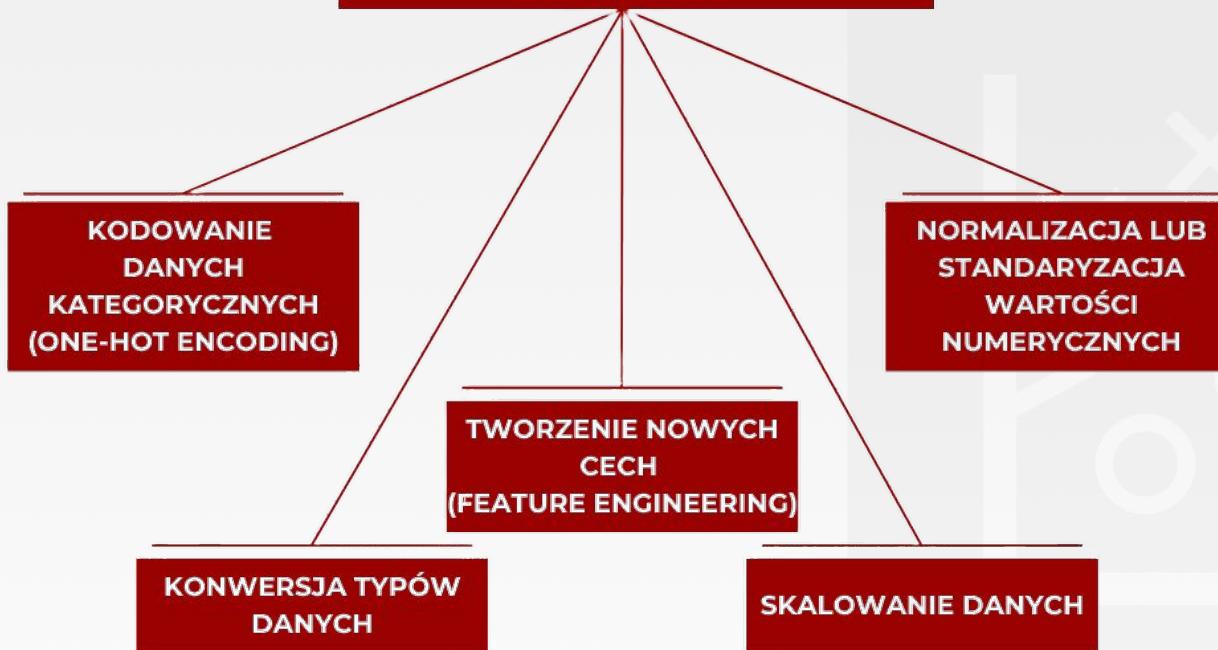


GHOST

Group of Horribly Optimistic Statisticians



PRZETWARZANIE DANYCH





GHOST

Group of Horribly Optimistic Statisticians



Dane mogą być błędne

Nie wszystkie dane są poprawne lub wiarygodne.

Modele uczone na takich danych mogą się mylić

lub podejmować nieadekwatne decyzje.

Im wcześniej wykryjemy błędy, tym łatwiej je naprawić.



GHOST

Group of Horribly Optimistic Statisticians



Jakie są dobre dane?

Dobre dane są:

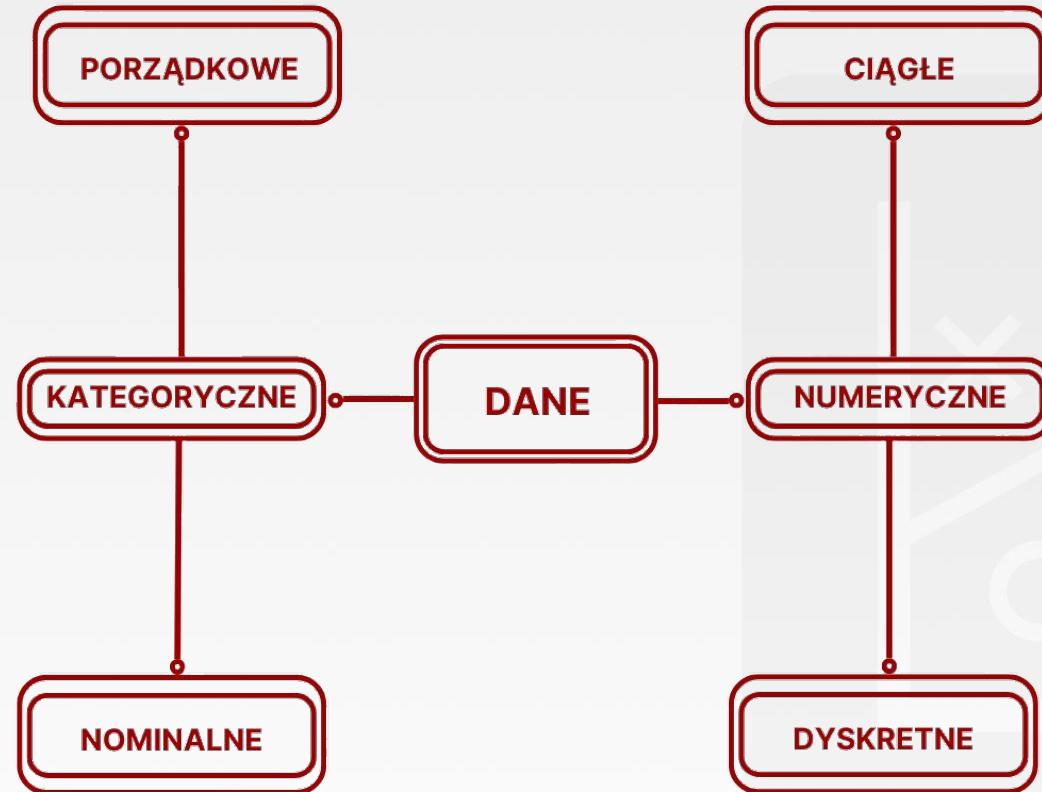
- *spójne,*
- *kompletne,*
- *dokładne,*
- *aktualne,*
- *reprezentatywne dla problemu, który chcemy rozwiązać.*

W praktyce jakość danych bywa ważniejsza niż sam wybór algorytmu.



GHOST

Group of Horribly Optimistic Statisticians





GHOST

Group of Horribly Optimistic Statisticians

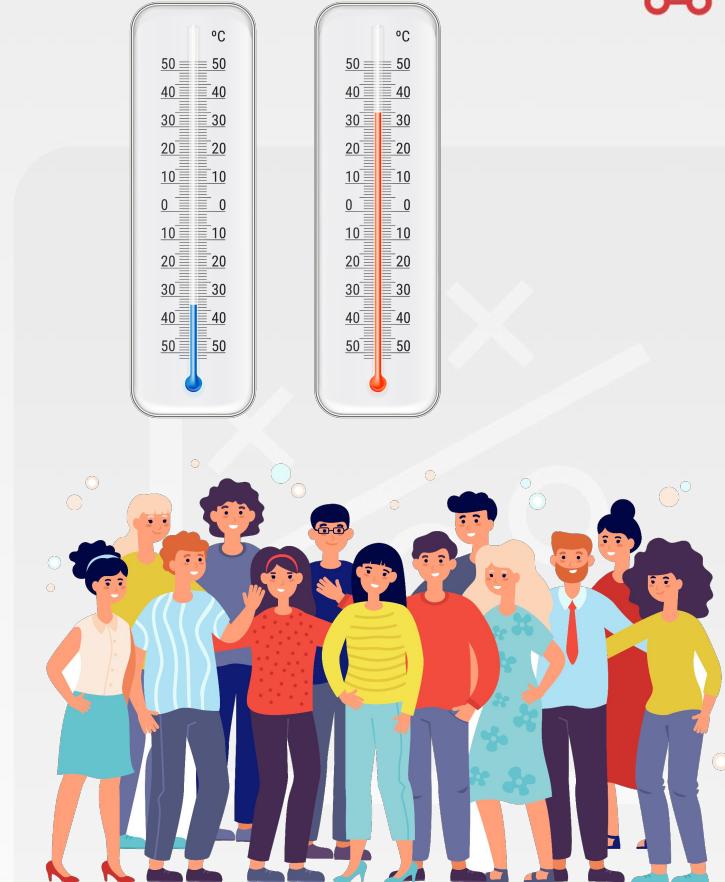


Dane numeryczne

Dane numeryczne są wartościami liczbowymi.

Dzielimy je na:

- ciągłe – mogą przyjmować dowolne wartości w pewnym zakresie, np. temperatura
- dyskretne – to liczby całkowite, np. ilość osób na dzisiejszym spotkaniu





GH^OST

Group of Horribly Optimistic Statisticians



Dane kategoryczne

Dane kategoryczne to takie, które dzielą elementy na grupy.

Dzielimy je na:

- nominalne – między grupami nie ma określonego porządku, np. kolor włosów
- porządkowe – mają one ustaloną kolejność, np. oceny w skali od 1 do 5





GHOST

Group of Horribly Optimistic Statisticians



Podział Danych





GHOST

Group of Horribly Optimistic Statisticians



Po co dzielimy dane?

W uczeniu maszynowym dzieli się dane na kilka części, aby uczciwie ocenić model. Model nie powinien być testowany na tych samych danych, na których się uczył. Dzięki podziałowi widzimy, jak dobrze zachowuje się dla nowych przypadków.



GHOST

Group of Horribly Optimistic Statisticians



Zbiór treningowy (training set)

Zbiór treningowy to dane, na których model się uczy.

Są to przykłady zawierające zarówno dane wejściowe, jak i poprawne odpowiedzi. Model analizuje je, by rozpoznać wzorce i zależności. Im bardziej zróżnicowany i kompletny jest zbiór treningowy, tym lepiej model się nauczy.

Trening to pierwszy etap budowania modelu.



GHOST

Group of Horribly Optimistic Statisticians



Zbiór walidacyjny (validation set)

Zbiór walidacyjny służy do wyboru najlepszych parametrów modelu. Nie jest używany do nauki, ale do sprawdzania, jak dobrze model radzi sobie z nieznanymi danymi. Pomaga to uniknąć przeuczenia (overfitting) lub niedouczenia (underfitting). Walidacja to kluczowy etap testowania modelu w trakcie jego tworzenia.



GHOST

Group of Horribly Optimistic Statisticians



Zbiór testowy (test set)

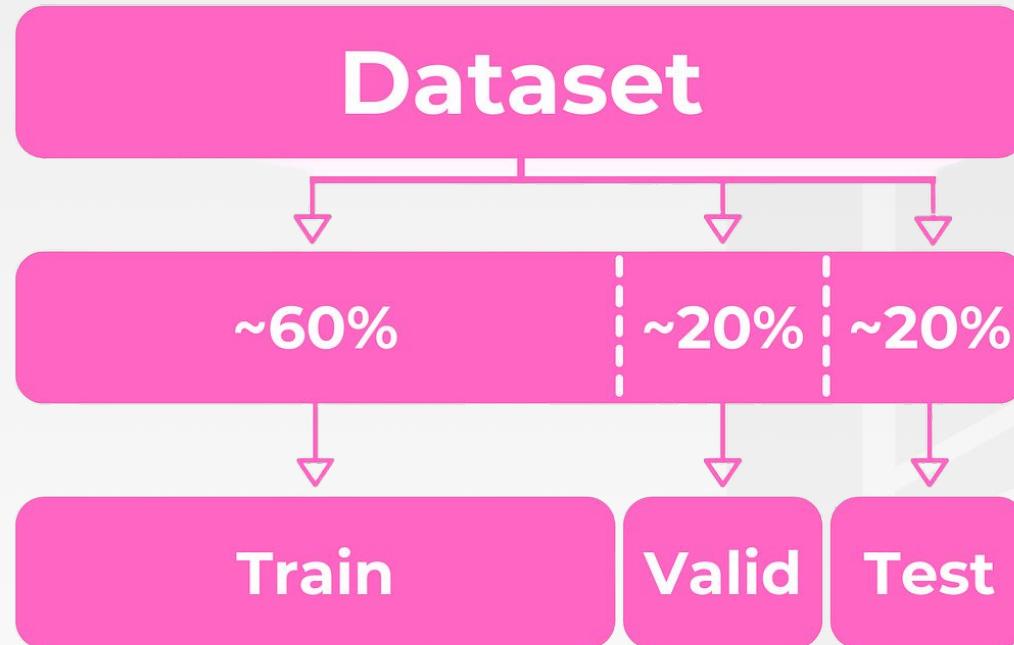
Zbiór testowy zawiera dane, których model wcześniej nie widział. Służy do końcowej oceny skuteczności modelu po jego wytrenowaniu i dostrojeniu. Wyniki uzyskane na tym zbiorze pozwalają oszacować, jak dobrze model poradzi sobie w rzeczywistych zastosowaniach.

To ostatni krok w procesie budowania modelu.



GHOST

Group of Horribly Optimistic Statisticians



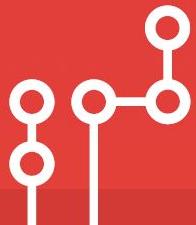


GHOST

Group of Horribly Optimistic Statisticians



Rozkład Danych





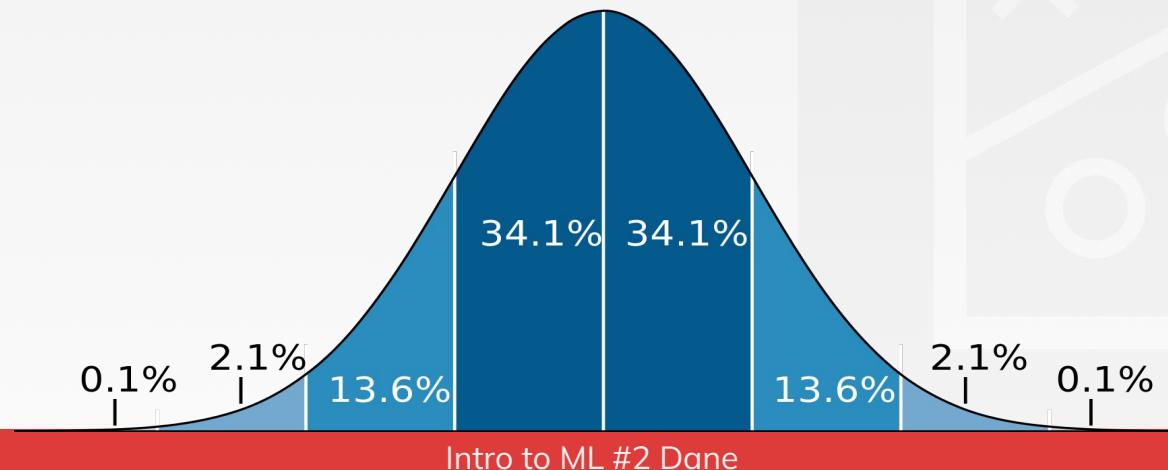
GHOST

Group of Horribly Optimistic Statisticians



Czym jest rozkład danych?

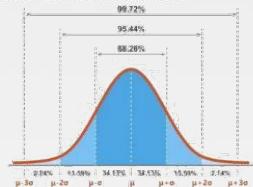
Rozkład danych to sposób przedstawienia, jak często występują poszczególne wartości. Dzięki niemu możemy lepiej zrozumieć strukturę zbioru i sprawdzić, czy dane są zrównoważone.





GHOST

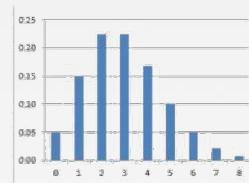
Group of Horribly Optimistic Statisticians



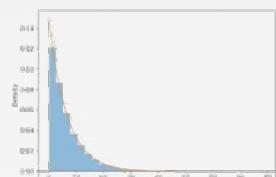
Normal
Distribution



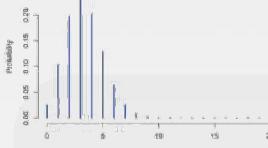
Bernoulli
Distribution



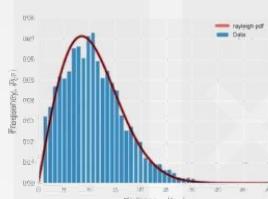
Poisson
Distribution



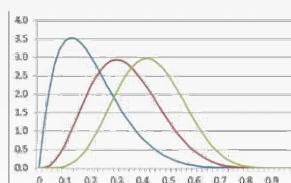
Exponential
Distribution



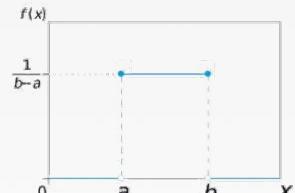
Binomial
Distribution



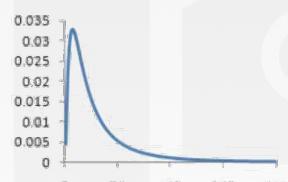
Gamma
Distribution



Beta
Distribution



Uniform
Distribution



Log Normal
Distribution



GHOST

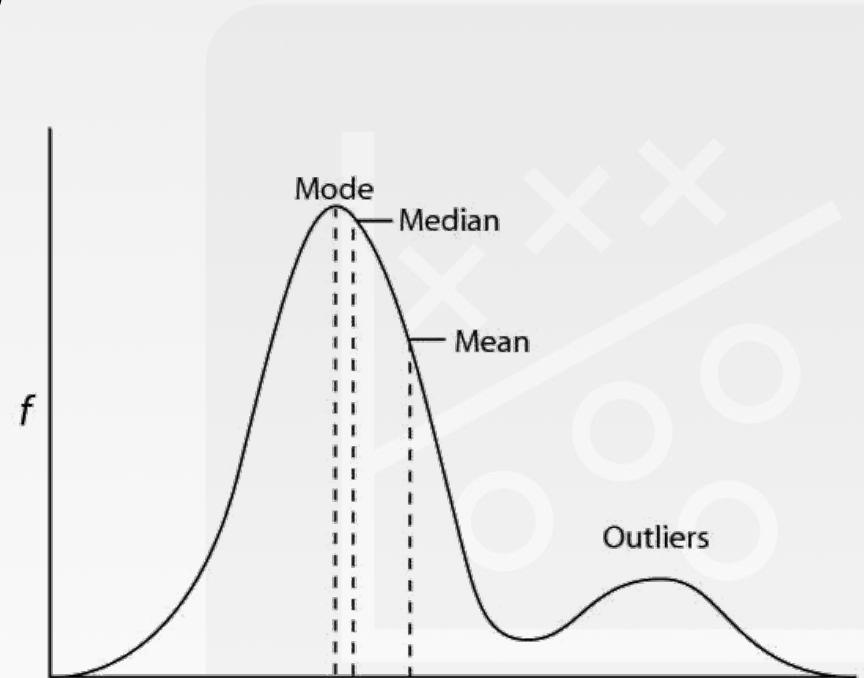
Group of Horribly Optimistic Statisticians



Wartości odstające (outliers)

Wartości odstające to takie, które znacznie różnią się od większości danych.

Mogą one wynikać z błędów pomiaru, literówek, ale też być poprawnymi wartościami.



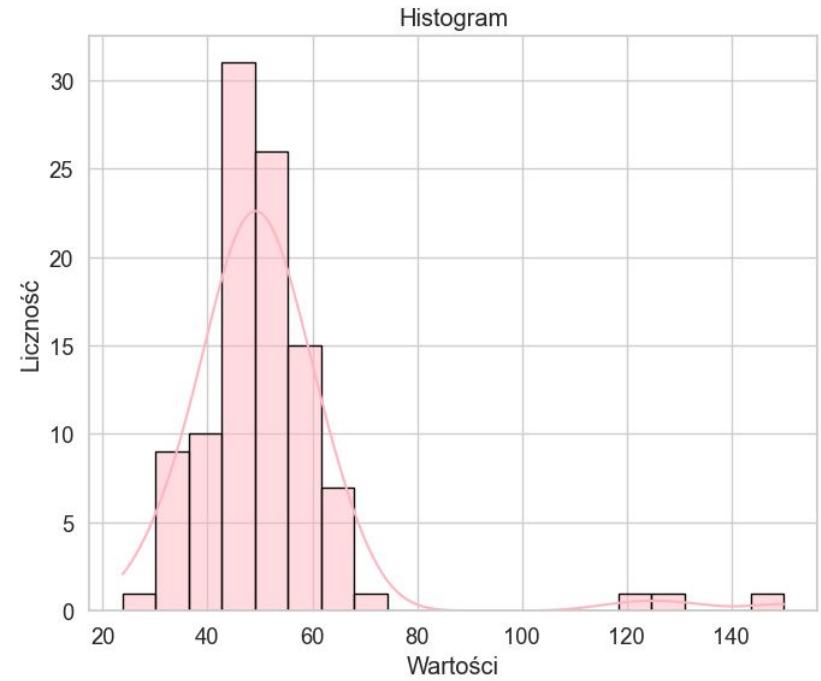


GHOST

Group of Horribly Optimistic Statisticians



Wykrywanie wartości odstających



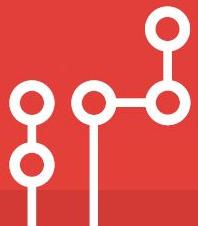


GHOST

Group of Horribly Optimistic Statisticians



Zależności między danymi





GHOST

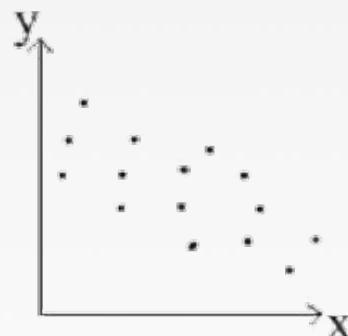
Group of Horribly Optimistic Statisticians



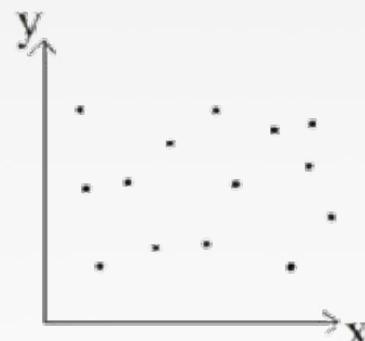
Relacje między zmiennymi

Relacje między zmiennymi to zależności, które mogą występować w danych.

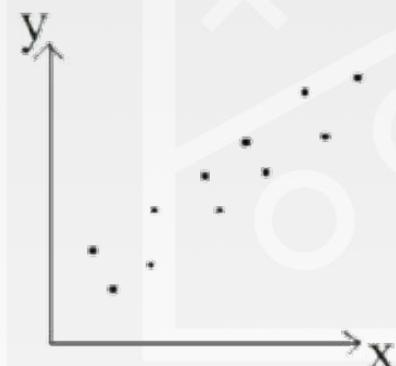
Często mogą być one skorelowane.



korelacja ujemna



brak korelacji



korelacja dodatnia



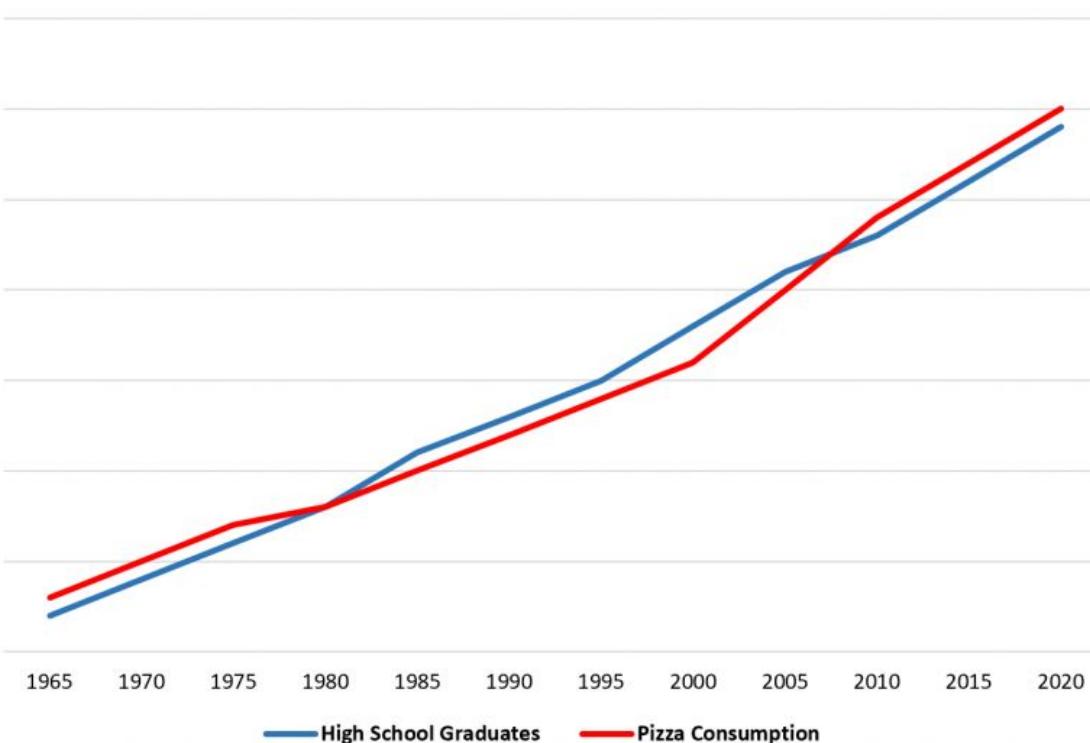
GHOST

Group of Horribly Optimistic Statisticians



Korelacja a przyczynowość

High School Graduates vs. Pizza Consumption





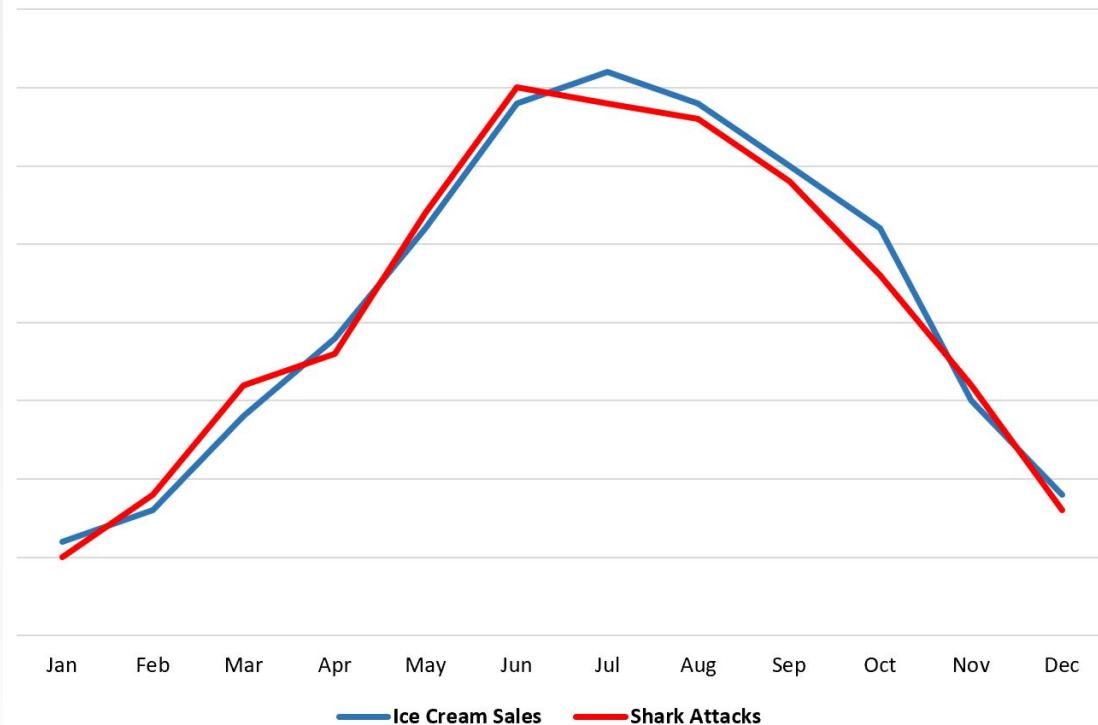
GHOST

Group of Horribly Optimistic Statisticians



Korelacja a przyczynowość

Ice Cream Sales vs. Shark Attacks





GHOST

Group of Horribly Optimistic Statisticians



Correlation does not mean causation

Korelacja nie implikuje związku przyczynowego.

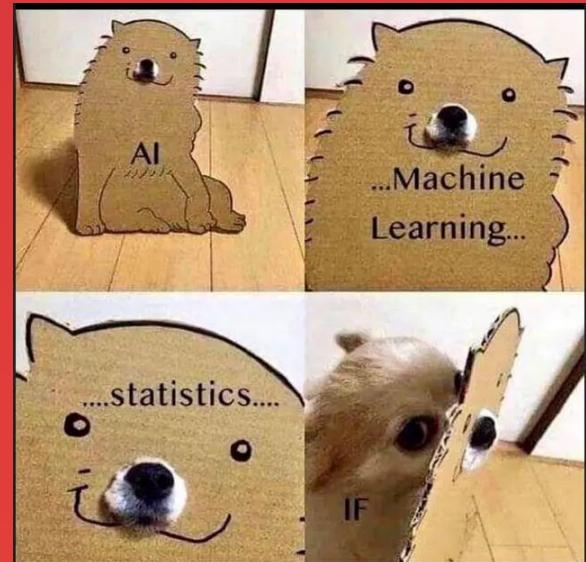


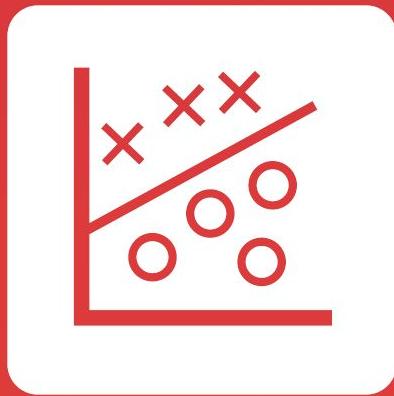


GHOST

Group of Horribly Optimistic Statisticians

Dziękuję za uwagę!





GHOST

Group of Horribly Optimistic Statisticians

