Wprowadzenie do analizy danych i uczenia maszynowego

Statystyka opisowa, pandas

Franciszek Saliński

Koło Naukowe Data Science, Wydział MiNI PW

8 listopada 2025

Czym jest uczenie maszynowe?

- **Uczenie maszynowe (Machine Learning)** to dziedzina z pogranicza informatyki i statystyki poświęcona algorytmom, które automatycznie "uczą się" wzorców w oparciu o dane
- Model uczy się zależności między zmiennymi wejściowymi (cechami) a wyjściem (etykietą, zmienną celu)
- Przykłady zastosowań:
  - przewidywanie cen mieszkań,
  - klasyfikacja obrazów,
  - analiza sentymentu w tekstach

Dlaczego analizować dane przed modelowaniem?

- Dane często zawierają **błędy, braki, wartości odstające**
- Zrozumienie struktury danych pozwala dobrać odpowiedni model
- Wstępna analiza zapobiega **błędnym wnioskom i nadmiernemu dopasowaniu (overfittingowi)**
- Cały ten proces nazywamy **Exploratory Data Analysis (EDA)** – *eksploracyjną analizą danych*.

Co to jest EDA?

- EDA polega na:
  - podsumowaniu danych za pomocą statystyk opisowych,
  - wizualizacji rozkładów i zależności,
  - wykrywaniu wartości odstających, anomalii
- Przydatne biblioteki w Pythonie:
  - `pandas`
  - `numpy`
  - `matplotlib`
  - `seaborn`

Typy danych

- **Dane ilościowe** – przyjmują wartości liczbowe.
  - *ciągłe* (np. wiek, zarobki),
  - *dyskretne* (np. ocena)
- **Dane jakościowe** – opisują cechy, kategorie.
  - *nominalne* (np. kolor, płeć),
  - *porządkowe* (np. poziom satysfakcji: niski/średni/wysoki)

Miary tendencji centralnej

- **Średnia arytmetyczna:**

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

opisuje „środek ciężkości" danych, lecz jest wrażliwa na obserwacje odstające.

- **Mediana:** środkowa obserwacja w uporządkowanych danych
- **Statystyki porządkowe:**

$$x_{1:n} \leqslant x_{2:n} \leqslant \cdots \leqslant x_{n:n}$$

to uporządkowane wartości próby od najmniejszej do największej.

- **Średnia ucięta:**

$$\bar{x}_{(\alpha)} = \frac{1}{n-2k} \sum_{i=k+1}^{n-k} x_{i:n}, \quad k = \lfloor \alpha n \rfloor$$

Miary tendencji centralnej

- **Średnia winsorowska:** wartości skrajne zastępuje się najbliższymi nieodciętymi obserwacjami:

$$x'_i = \begin{cases} x_{k+1:n}, & i \leqslant k, \\ x_{i:n}, & k < i \leqslant n-k, \\ x_{n-k:n}, & i > n-k. \end{cases}$$

Następnie liczymy

$$\bar{x}_W = \frac{1}{n} \sum_{i=1}^{n} x'_i.$$

Kwartyle

- **Pierwszy kwartyl (dolny):**

$$Q_1 = x_{(0.25 \cdot n):n}$$

dzieli próbę tak, że 25% obserwacji jest mniejszych lub równych Q_1.

- **Mediana (drugi kwartyl):**

$$Q_2 = Med(x) = \begin{cases} x_{(\frac{n+1}{2}):n}, & n \text{ nieparzyste}, \\ \dfrac{x_{(n/2):n} + x_{(n/2+1):n}}{2}, & n \text{ parzyste} \end{cases}$$

Dzieli dane na dwie równe części.

- **Trzeci kwartyl (górny):**

$$Q_3 = x_{(0.75 \cdot n):n}$$

dzieli próbę tak, że 75% obserwacji jest mniejszych lub równych Q_3.

Miary rozproszenia

- **Wariancja:**

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

- **Odchylenie standardowe:**

$$s = \sqrt{s^2}$$

- **Rozstęp:**

$$R = x_{\text{max}} - x_{\text{min}}$$

- **Kwartyle i IQR:**

$$IQR = Q_3 - Q_1$$

Miary kształtu rozkładu

- **Skośność (skewness):** miara asymetrii rozkładu

$$\text{Skew}(X) = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s} \right)^3$$

- **Kurtoza:** miara „spiczastości” rozkładu (w porównaniu do normalnego)

$$\text{Kurt}(X) = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s} \right)^4$$

Współczynniki korelacji

- **Korelacja liniowa Pearsona:**

$$r_{XY} = \frac{\mathrm{Cov}(X, Y)}{s_X s_Y}$$

mierzy siłę i kierunek liniowej zależności między zmiennymi X i Y.

- **Korelacja rang Spearmana:**

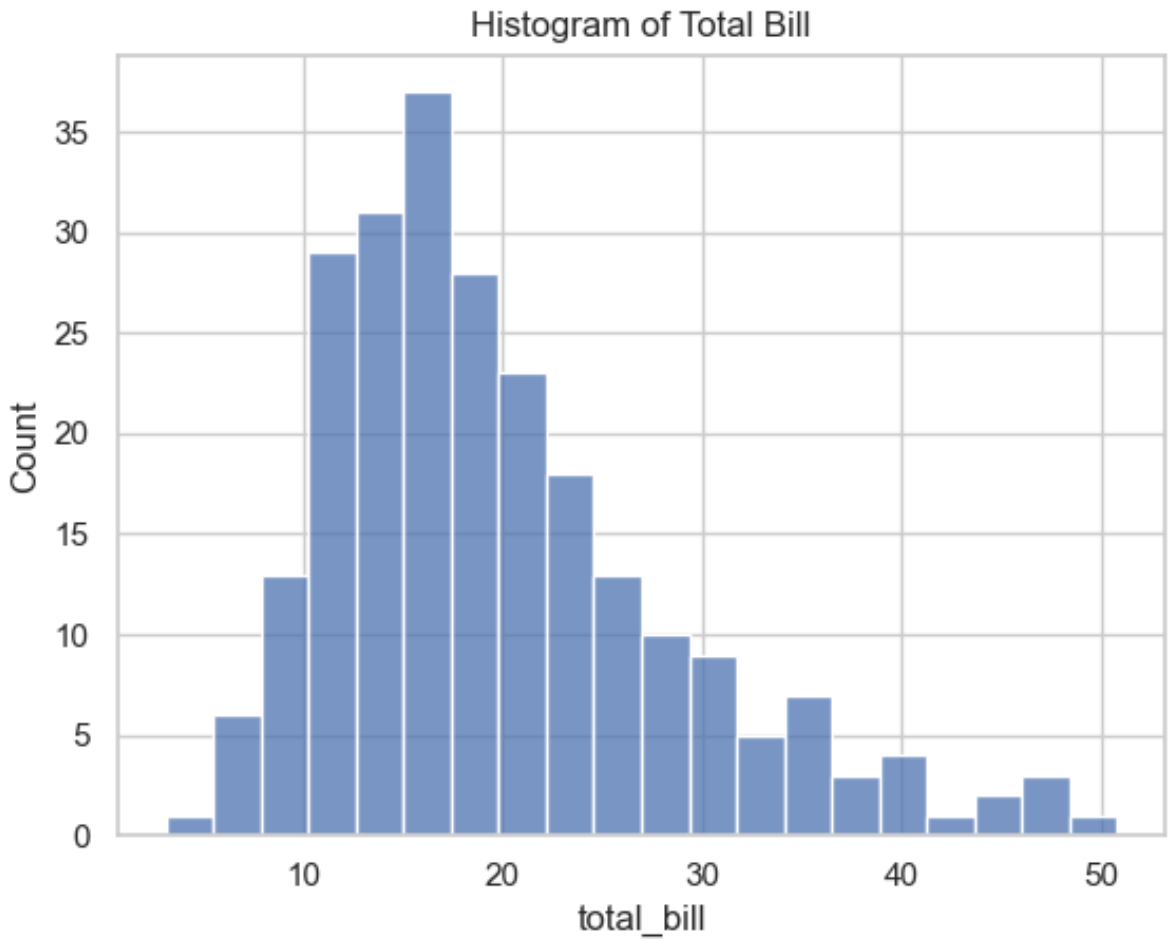
$$\rho = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2 - 1)}$$

gdzie d_i to różnice między rangami x_i i y_i. Opiera się na rangach wartości (kolejności obserwacji).

- Wartości obu współczynników należą do przedziału $[-1, 1]$:
  - $\rho \approx 1$ — silna zależność dodatnia,
  - $\rho \approx -1$ — silna zależność ujemna,
  - $\rho \approx 0$ — brak monotonicznej zależności.

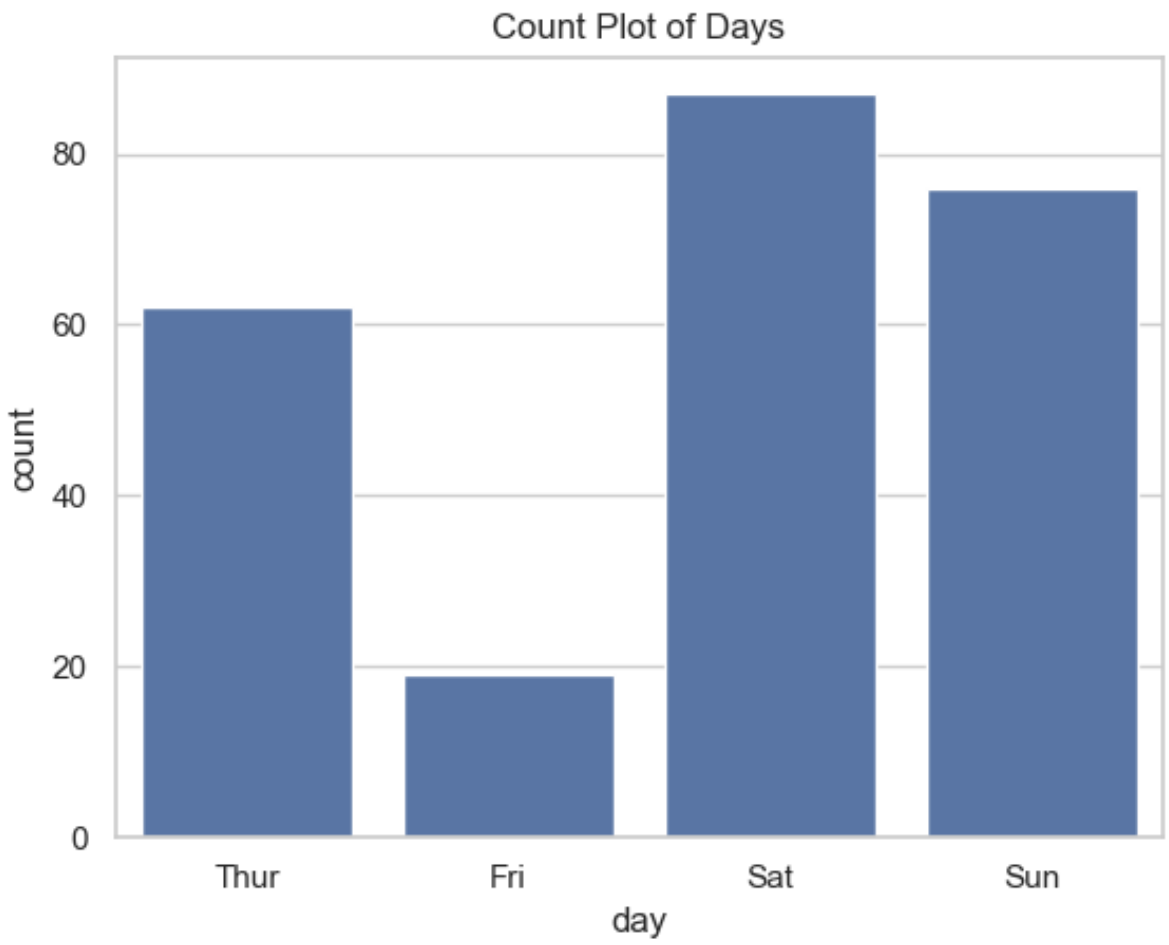
Podstawowe wykresy

Histogram - wizualizuje rozkład zmiennej ciągłej



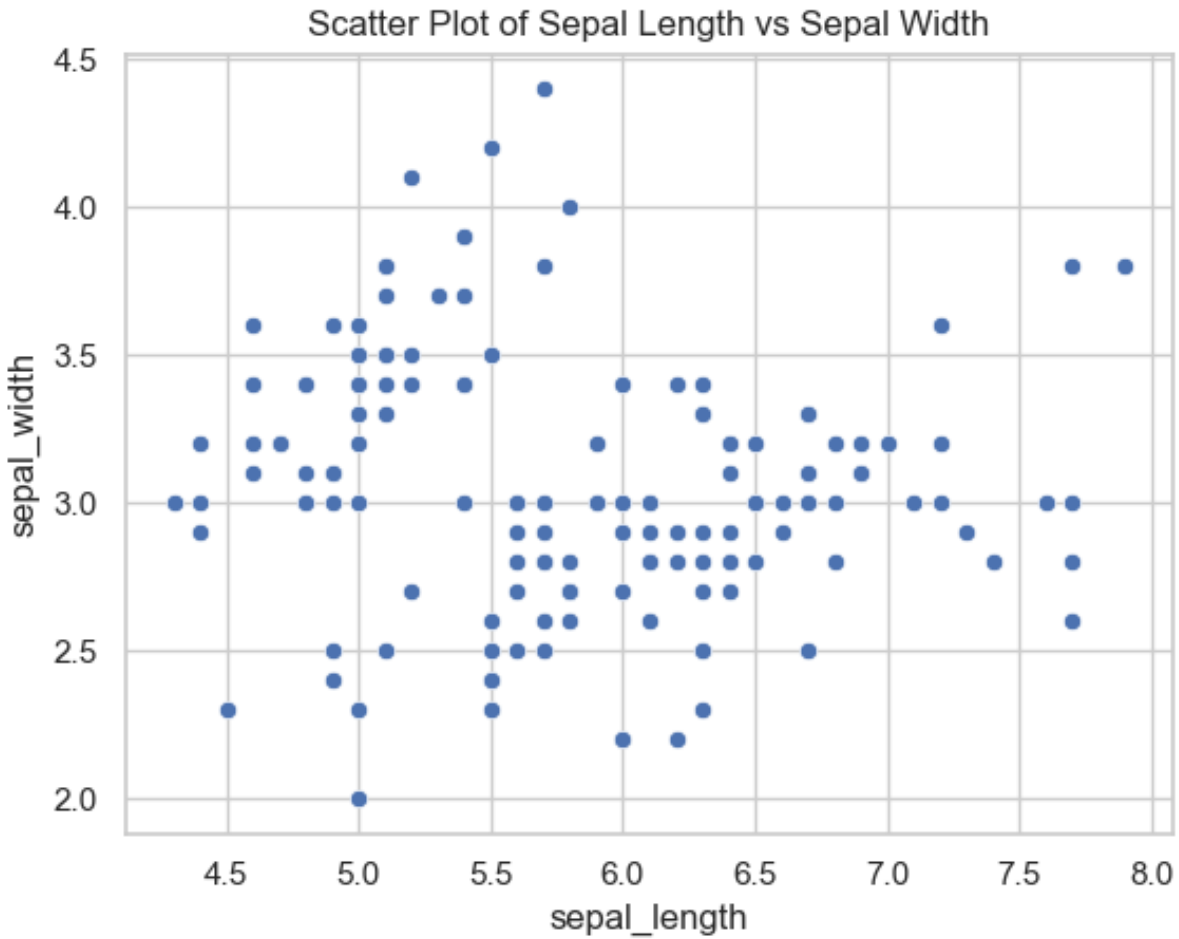
Podstawowe wykresy

Wykres słupkowy (bar) - wizualizuje rozkład zmiennej dyskretnej



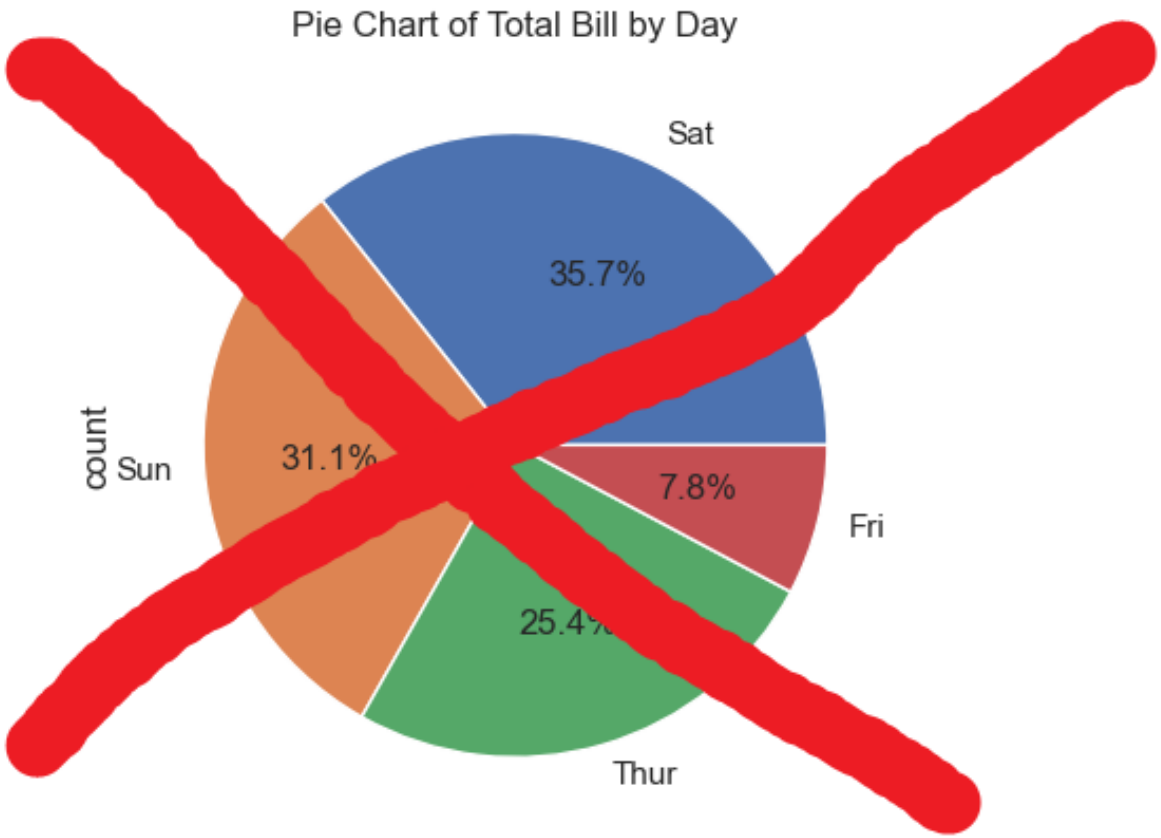
Podstawowe wykresy

Wykres punktowy (scatter) - wizualizuje zależność między dwoma zmiennymi ciągłymi



Podstawowe wykresy

Wykres kołowy - nie robimy



Podstawowe obiekty pandas

- **Series** – jednowymiarowa tablica danych z etykietami.
- **DataFrame** – dwuwymiarowa tabela z kolumnami różnego typu.

Polecane źródła

- **Minimalist Data Wrangling with Python:**
  https://datawranglingpy.gagolewski.com
- **An Introduction to Statistical Learning:**
  www.statlearning.com