

EDA Danych dotyczących Irysów

O Danych

Zbiór danych zawiera informacje o trzech gatunkach irysów: *Iris setosa*, *Iris versicolor*, i *Iris virginica*.

Dane obejmują pomiary czterech cech: długość i szerokość działki kielicha oraz długość i szerokość płatk.

Każdy wiersz w zbiorze danych reprezentuje pojedynczy kwiat, a wartości pomiarów są podane w centymetrach.

Zbiór składa się z 150 próbek, po 50 dla każdego gatunku, i jest szeroko stosowany jako podstawowy zbiór danych do testowania algorytmów klasyfikacji oraz w nauce o danych i uczeniu maszynowym.

Kolumny:

- **długość kielicha (sepal length)** - Długość kielicha w cm
- **szerokość kielicha (sepal width)** - Szerokość kielicha w cm
- **długość płatka (petal length)** - Długość płatka w cm
- **szerokość płatka (petal width)** - Szerokość płatka w cm
- **klasa (class)** - Klasa irysa (*setosa*, *versicolor*, *virginica*)

Ogólny przegląd danych

	długość kielicha (sepal length)	szerokość kielicha (sepal width)	długość płatka (petal length)	szerokość płatka (petal width)	klasa (class)
132	6.4	2.8	5.6	2.2	Iris- virginica
39	5.1	3.4	1.5	0.2	Iris- setosa
133	6.3	2.8	5.1	1.5	Iris- virginica
36	5.5	3.5	1.3	0.2	Iris- setosa
60	5.0	2.0	3.5	1.0	Iris- versicolor

Analiza danych brakujących i duplikatów

```
długość kielicha (sepal length)    0
szerokość kielicha (sepal width)   0
długość płatka (petal length)      0
szerokość płatka (petal width)     0
klasa (class)                      0
dtype: int64
```

Suma duplikatów

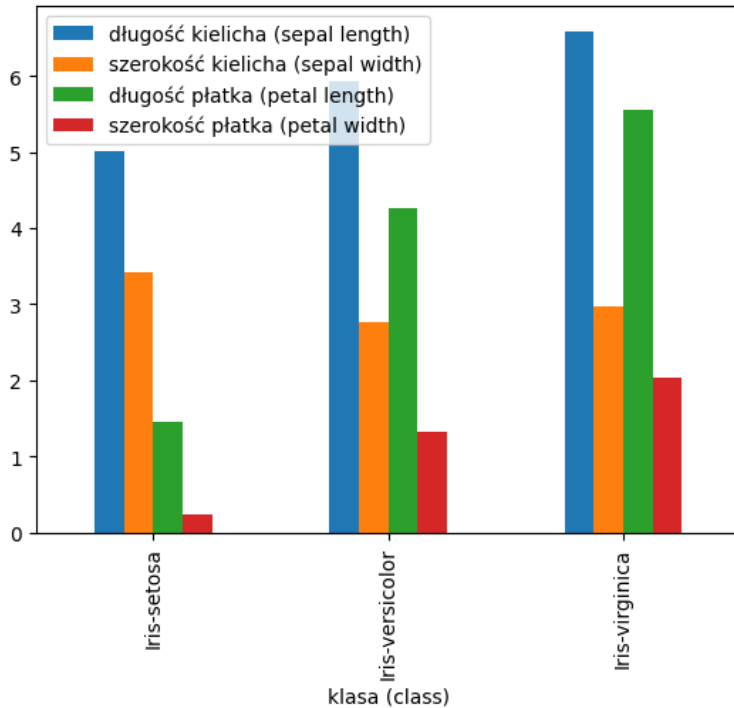
3

średnie wartości dla elementów kwiatów

	klasa (class)	długość kielicha (sepal length)		szerokość kielicha (sepal width)		długość płatka (petal length)		szerokość płatka (petal width)	
		mean	std	mean	std	mean	std	mean	std
0	Iris-setosa	5.006	0.352490	3.418	0.381024	1.464	0.173511	0.244	0.107210
1	Iris-versicolor	5.936	0.516171	2.770	0.313798	4.260	0.469911	1.326	0.197753
2	Iris-virginica	6.588	0.635880	2.974	0.322497	5.552	0.551895	2.026	0.274650

średnie wartości dla elementów kwiatów wykres

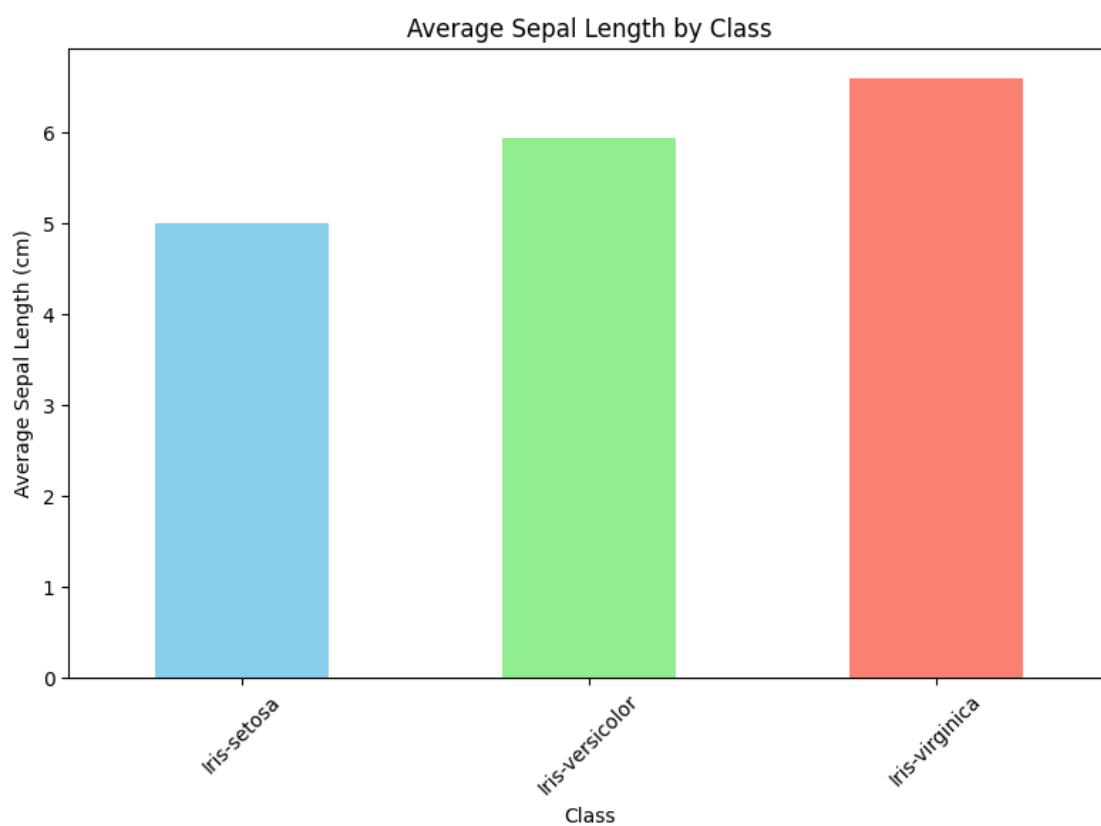
<Axes: xlabel='klasa (class) '>



Obliczenie pola powierzchni płatków kwiatów w grupach klas

	klasa (class)	mean	std
0	Iris-setosa	0.3628	0.183248
1	Iris-versicolor	5.7204	1.368403
2	Iris-virginica	11.2962	2.157412

Wykres przedstawiający różnice w średnich długości kielicha w podziale na klasy



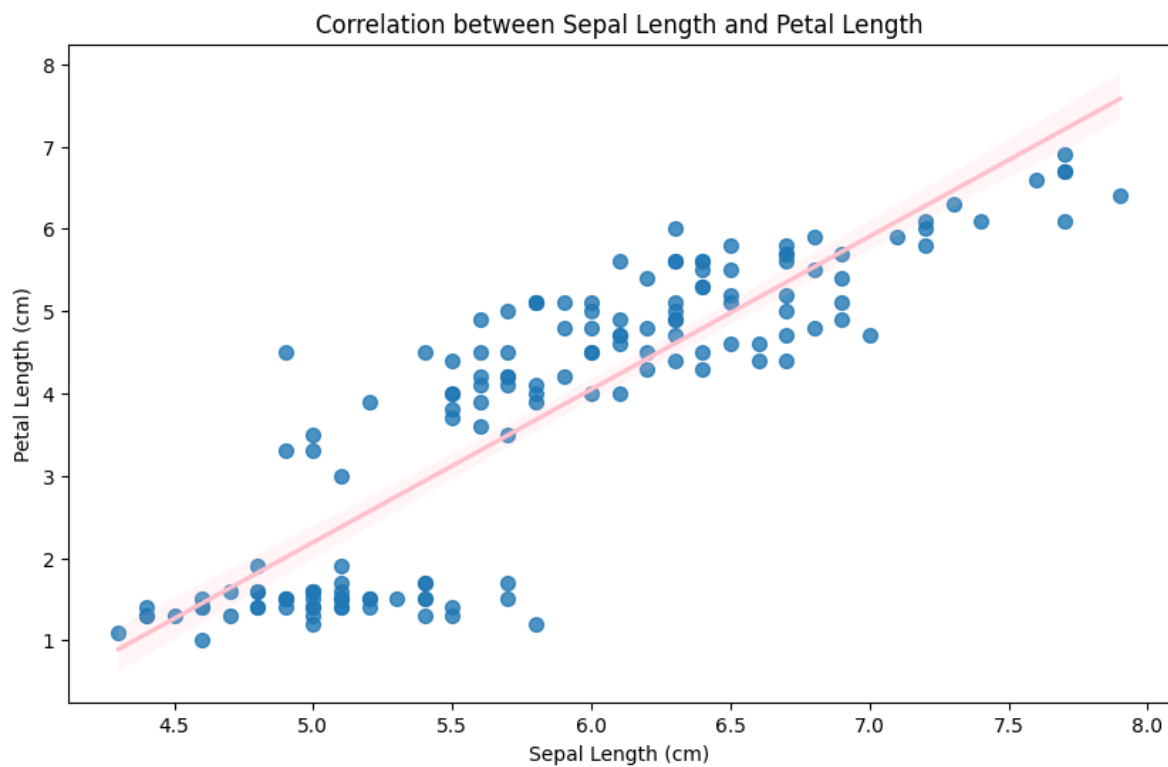
Analiza korelacji

wynik korelacji

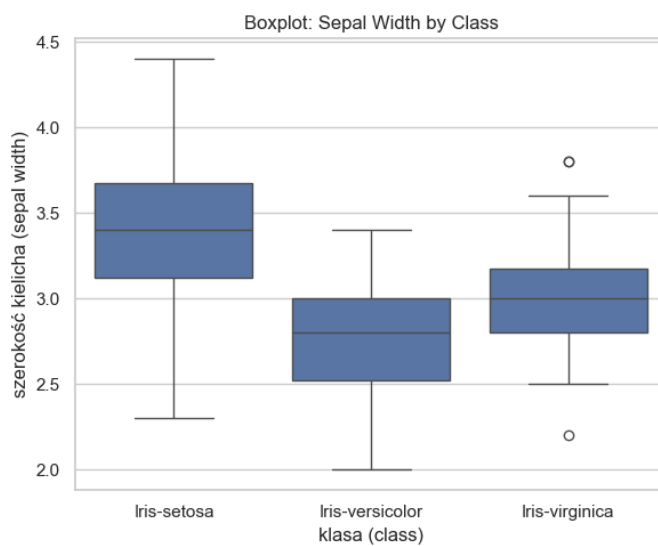
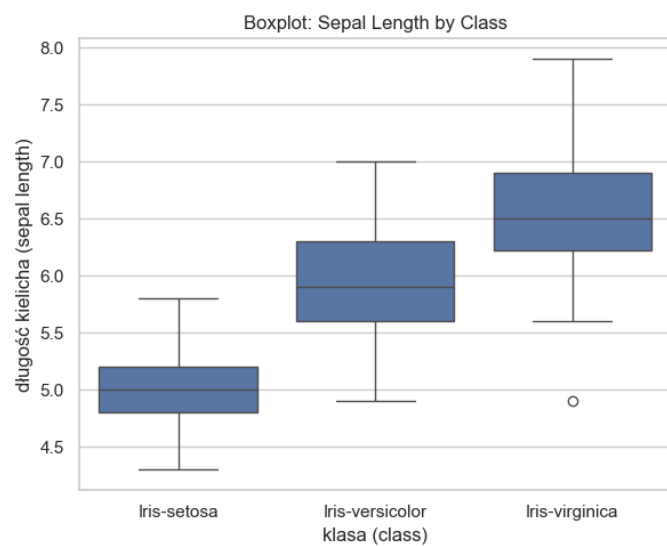
0.8717541573048712

Wykres korelacji

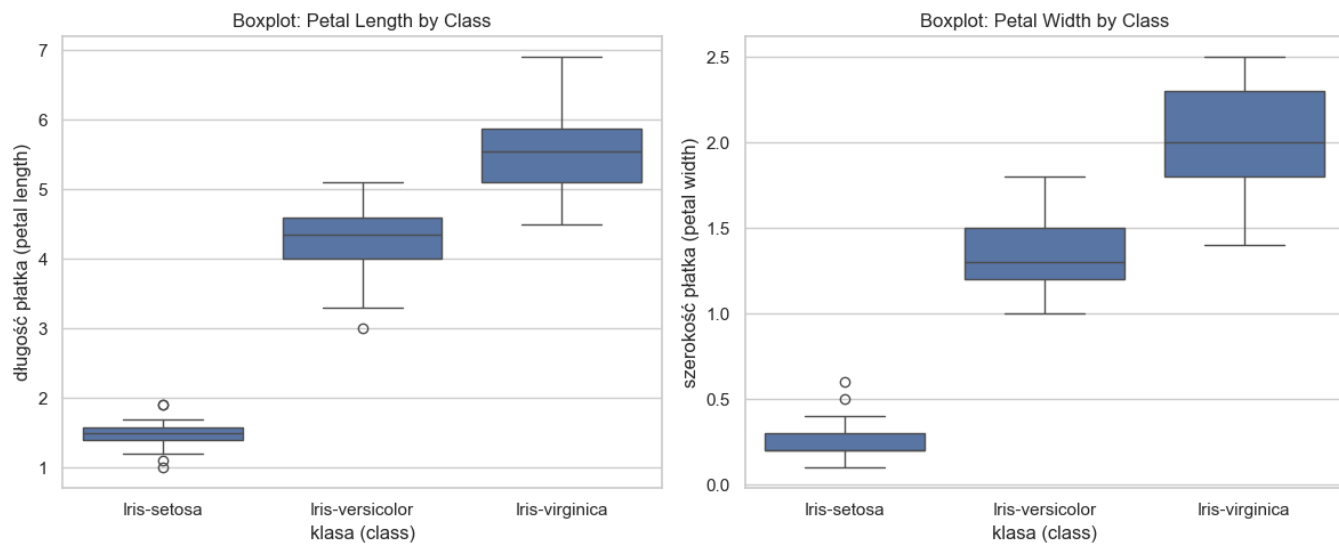
```
Text(0, 0.5, 'Petal Length (cm)')
```



Analiza danych odstajacych



Analiza danych odstających – ciąg dalszy



Wnioski analizy

Baza danych była kompletna, nie zawierała braków, a wykryte 3 duplikaty prawdopodobnie wynikają z powtórzeń pomiarowych i nie miały istotnego wpływu na dalszą analizę.

Aby porównać rozmiary kwiatów między klasami, obliczono pole powierzchni płatków (petal_area). Największą średnią powierzchnię miał Iris-virginica (11.30 cm², SD = 2.16), natomiast najmniejszą Iris-setosa (0.36 cm², SD = 0.18). Podobny trend zaobserwowano w długości kielicha, najdłuższe miał Iris-virginica (6.59 cm, SD = 0.64).

Analiza korelacji wykazała silne zależności między zmiennymi: między długością kielicha a długością płatka ($r = 0.87$), między szerokością a długością płatka ($r = 0.96$) oraz długością kielicha a szerokością płatka ($r = 0.82$).

Boxplot ujawnił obecność wartości odstających, głównie w grupie Iris-setosa, w zmiennych opisujących płatki. Może to wskazywać na większą zmienność wewnątrz tej klasy lub obecność pojedynczych, nietypowych okazów.

Podsumowując, dane okazały się kompletne i nie wymagały większego czyszczenia. Analiza pokazała wyraźne różnice między gatunkami kwiatów, zwłaszcza jeśli chodzi o długość i szerokość płatków. Niektóre cechy były ze sobą mocno powiązane, co może być przydatne przy budowaniu modelu. W kolejnym kroku warto spróbować stworzyć model klasyfikacyjny, który rozpozna gatunek na podstawie tych cech, gdzie szczególnie pomocne mogą być cechy płatków.