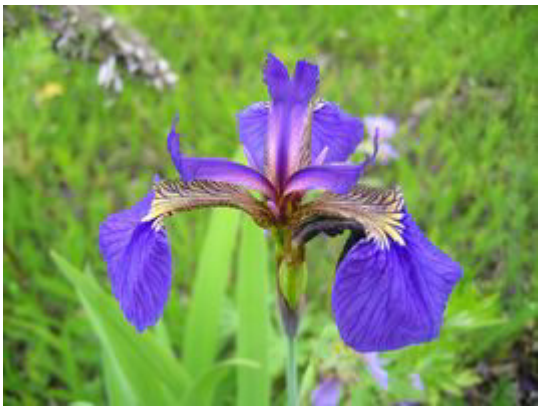
EDA Danych dotyczących Irysów

EDA to proces opisany w kilku krokach umozliwiajacy analize danych. Odkrywa sie roznego rodzaju prawidłowosci, wzorce w danych przy pomocy odpowiednich narzedzi. Jest to procedura opisana w kiku krokach. Pozwala zrozumiec dane, jakie mają tendencyjnosci, problemy, relacje.

1. Ogólny przeglad danych (kształt i opis danych)
2. Analiza brakujacych informacji
3. Analiza pojedynczych zmiennych. Eksplorujemy poszczegolne kolumny. Krok analizy pojedynczych kolumn (zmiennych)
4. Transformacje danych
5. Analiza relacji miedzy zmiennymi
6. Analiza wartosci odstajacych

1. Ogólny przegląd i opis danych

Irysy (Iris) to wieloletnie rośliny ozdobne z rodziny kosaćcowatych, znane z dużych, kolorowych kwiatów o charakterystycznym kształcie. Występują naturalnie w klimacie umiarkowanym półkuli północnej, a wiele gatunków uprawia się w ogrodach. Kwiaty irysów mogą mieć różne barwy – od fioletowej i niebieskiej po żółtą i białą. Rośliny te rosną z kłączy lub cebulek, w zależności od gatunku.



Iris setosa to gatunek irysa pochodzący głównie z północno-wschodniej Azji i północno-zachodniej Ameryki Północnej. Rośnie zwykle na wilgotnych łąkach i brzegach zbiorników wodnych. Ma charakterystyczne niebiesko-fioletowe kwiaty o delikatnym zapachu oraz wąskie, mieczowate liście. Iris setosa jest rośliną kłączową, dobrze przystosowaną do chłodniejszego klimatu. Często wykorzystywany jest w badaniach botaniki i ekologii ze względu na swoją odporność i zdolność do adaptacji.

Iris versicolor, znany też jako kosaćiec trójbarwny, to gatunek irysa występujący naturalnie w Ameryce Północnej, zwłaszcza na terenach podmokłych, takich jak brzegi rzek i jezior. Roślina ta ma efektowne, trójbarwne kwiaty – zazwyczaj w odcieniach niebieskiego, fioletu i białego. Iris versicolor rośnie z kłącza i osiąga zwykle wysokość od 50 do 90 cm. Jest popularny w ogrodnictwie ze względu na swoje dekoracyjne kwiaty oraz zdolność do oczyszczania wód, dzięki czemu bywa wykorzystywany w naturalnych systemach filtracji.

Iris virginica, zwany kosaćcem wirginijskim, to gatunek irysa występujący głównie na południowo-wschodnich terenach Stanów Zjednoczonych. Rośnie naturalnie na wilgotnych łąkach i w pobliżu zbiorników wodnych. Ma duże, efektowne kwiaty w odcieniach niebieskiego i fioletu oraz mieczowate liście. Roślina ta jest kłączowa i może osiągać wysokość do około metra. Iris virginica jest ceniony zarówno w botanice, jak i ogrodnictwie, gdzie używany jest jako roślina ozdobna i do rekultywacji terenów podmokłych.

Przykład danych

	Długość kielicha	Szerokość kielicha	Długość płatka	Szerokość płatka	Gatunek
0	5.1	3.5	1.4	0.2	Iris-setosa
1	4.9	3.0	1.4	0.2	Iris-setosa
2	4.7	3.2	1.3	0.2	Iris-setosa
3	4.6	3.1	1.5	0.2	Iris-setosa
4	5.0	3.6	1.4	0.2	Iris-setosa
...	...	...	...	...	...
145	6.7	3.0	5.2	2.3	Iris-virginica
146	6.3	2.5	5.0	1.9	Iris -virginica
147	6.5	3.0	5.2	2.0	Iris-virginica
148	6.2	3.4	5.4	2.3	Iris-virginica
149	5.9	3.0	5.1	1.8 I	ris-virginica

150 rows × 5 columns

Zbiór danych składa się z 150 wierszy i 5 kolumn

```
Liczba brakujących wartości w każdej kolumnie:
Długość kielicha      0
Szerokość kielicha    0
Długość płatka        0
Szerokość płatka      0
Gatunek               0
dtype: int64
```

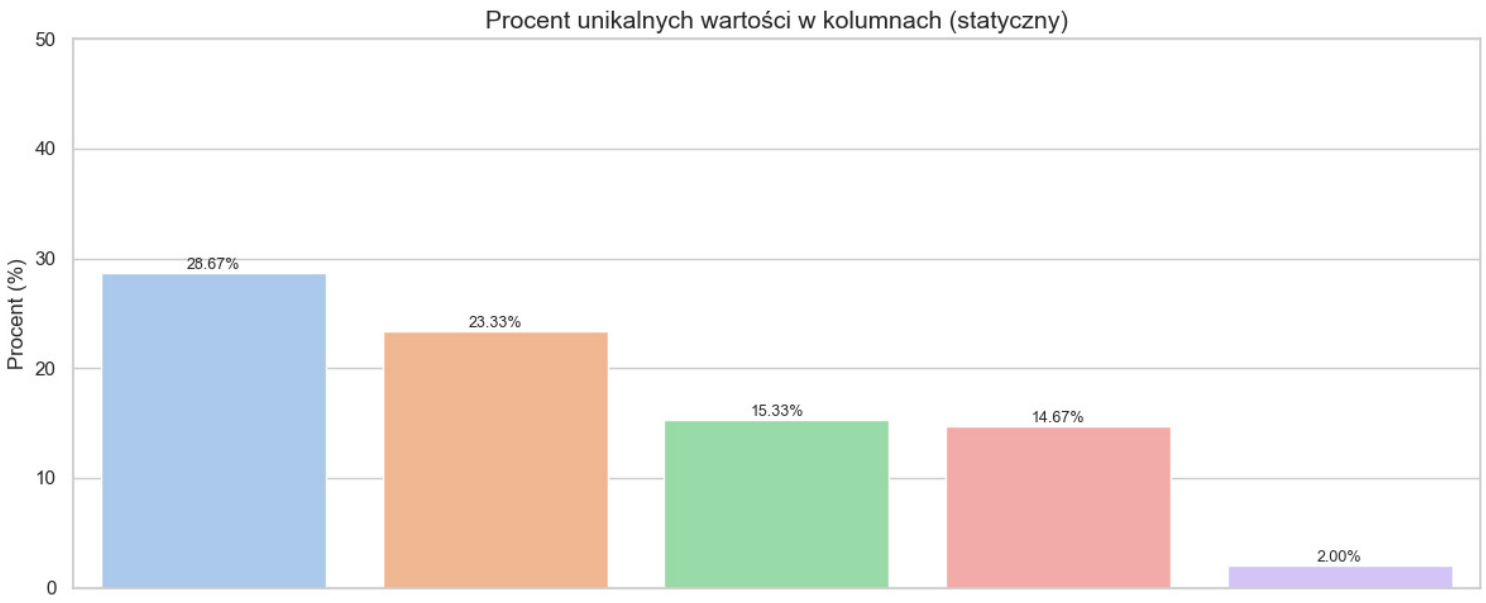
Brak wartości brakujących w kolumnach

	Długość kielicha	Szerokość kielicha	Długość płatka	Szerokość płatka	Gatunek
73	6.1	2.8	4.7	1.2	Iris-versicolor
18	5.7	3.8	1.7	0.3	Iris-setosa
118	7.7	2.6	6.9	2.3	Iris-virginica
78	6.0	2.9	4.5	1.5	Iris-versicolor
76	6.8	2.8	4.8	1.4	Iris-versicolor
31	5.4	3.4	1.5	0.4	Iris-setosa
64	5.6	2.9	3.6	1.3	Iris-versicolor
141	6.9	3.1	5.1	2.3	Iris-virginica
68	6.2	2.2	4.5	1.5	Iris-versicolor
82	5.8	2.7	3.9	1.2	Iris-versicolor

```
Unikalne wartości w kolumnach:
Długość kielicha       35
Szerokość kielicha     23
Długość płatka         43
Szerokość płatka       22
Gatunek                 3
dtype: int64
```

Procent unikalnych wartości w kolumnach

```
Długość kielicha: 23.33% unikalnych wartości
Szerokość kielicha: 15.33% unikalnych wartości
Długość płatka: 28.67% unikalnych wartości
Szerokość płatka: 14.67% unikalnych wartości
Gatunek: 2.0% unikalnych wartości
```



Z danych wynika ze mamy trzy irysy setosa i dwa irysy virginica o takich samych wymiarach

	Długość kielicha	Szerokość kielicha	Długość płatka	Szerokość płatka	Gatunek
34	4.9	3.1	1.5	0.1	Iris-setosa
37	4.9	3.1	1.5	0.1	Iris-setosa
142	5.8	2.7	5.1	1.9	Iris-virginica

Wykaz liczby obserwacji z obliczeniami średniej, wartosci maksymalnej i minimalnej, odchylenia standardowego i kwartylów 25%, 50% (mediana), 75%

Kolor czerwony maksymalna wartosc natomiast niebieski minimalna

	count	mean	std	min	25%	50%	75%	max
Długość kielicha	150.00	5.84	0.83	4.30	5.10	5.80	6.40	7.90
Szerokość kielicha	150.00	3.05	0.43	2.00	2.80	3.00	3.30	4.40
Długość płatka	150.00	3.76	1.76	1.00	1.60	4.35	5.10	6.90
Szerokość płatka	150.00	1.20	0.76	0.10	0.30	1.30	1.80	2.50

Wnioski:

1. Zbiór danych zawiera 150 obserwacji trzech gatunków irysów, opisanych czterema cechami numerycznymi: długością i szerokością kielicha (sepal) oraz długością i szerokością płatka (petal). Długość kielicha (sepal_length) ma średnią wartość 5.84 cm i waha się od 4.3 cm do 7.9 cm, natomiast jej szerokość (sepal_width) jest mniej zróżnicowana, z średnią 3.05 cm. Długość płatka (petal_length) jest najbardziej zróżnicowaną cechą – jej wartości rozciągają się od 1.0 cm do 6.9 cm, a średnia wynosi 3.76 cm, co odzwierciedla duże różnice między gatunkami. Szerokość płatka (petal_width) również wykazuje dużą zmienność, od 0.1 cm do 2.5 cm, ze średnią 1.20 cm. Dane nie zawierają wartości brakujących i każda cecha posiada dokładnie 150 pomiarów.
2. Std (odchylenie): większe wartości świadczą o większym zróżnicowaniu danych (np.długość płatka ma wyraźnie większe odchylenie niż szerokość kielicha).
3. W zbiorze danych nie ma zadnych brakujących wartosci. Dane sa przejrzyste i gotowe do analizy
4. Rozkład unikalnych wartosci wyrazonych procentowo:

- Długość kielicha: 23.33% unikalnych wartości
- Szerokość kielicha: 15.33% unikalnych wartości
- Długość płatka: 28.67% unikalnych wartości
- Szerokość płatka: 14.67% unikalnych wartości
- Gatunek: 2.0% unikalnych wartości

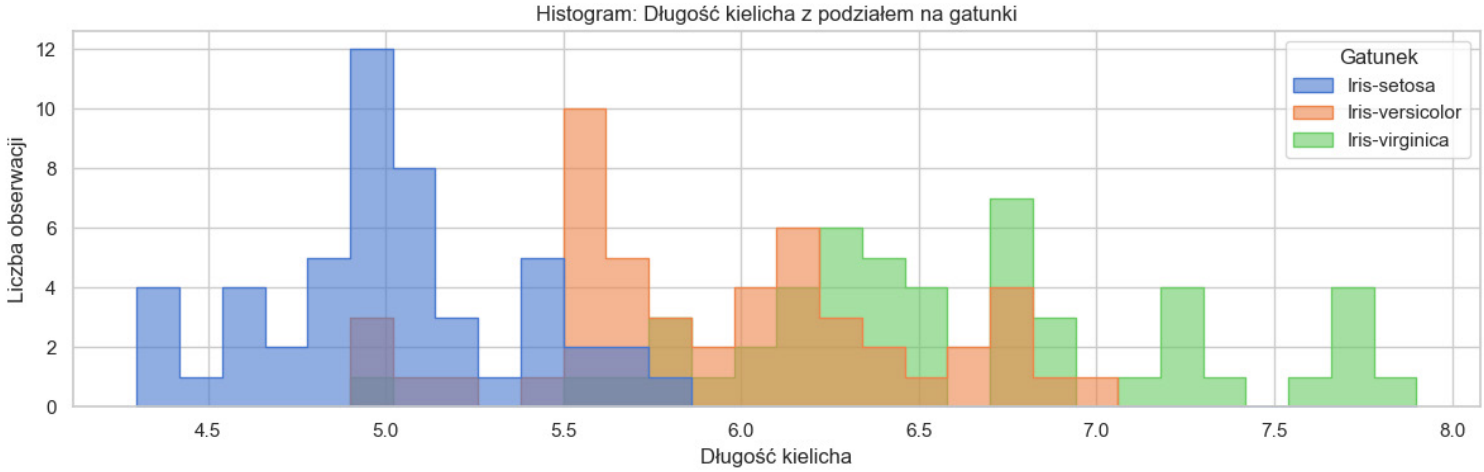
4. Z danych wynika ze mamy trzy irysy setosa i dwa irysy virginica o takich samych wymiarach

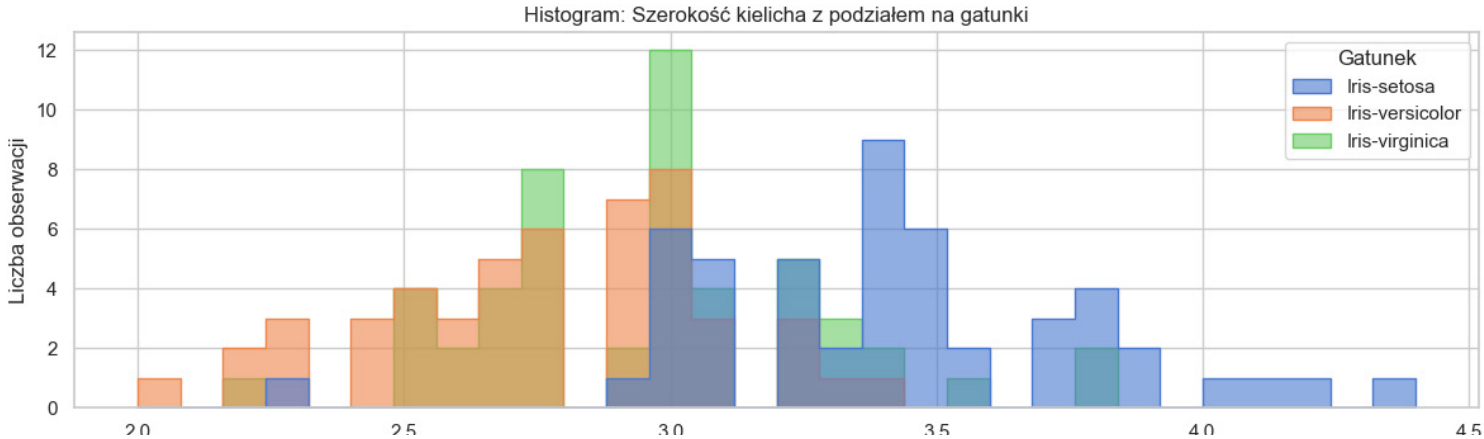
2 Analiza brakujących wartości

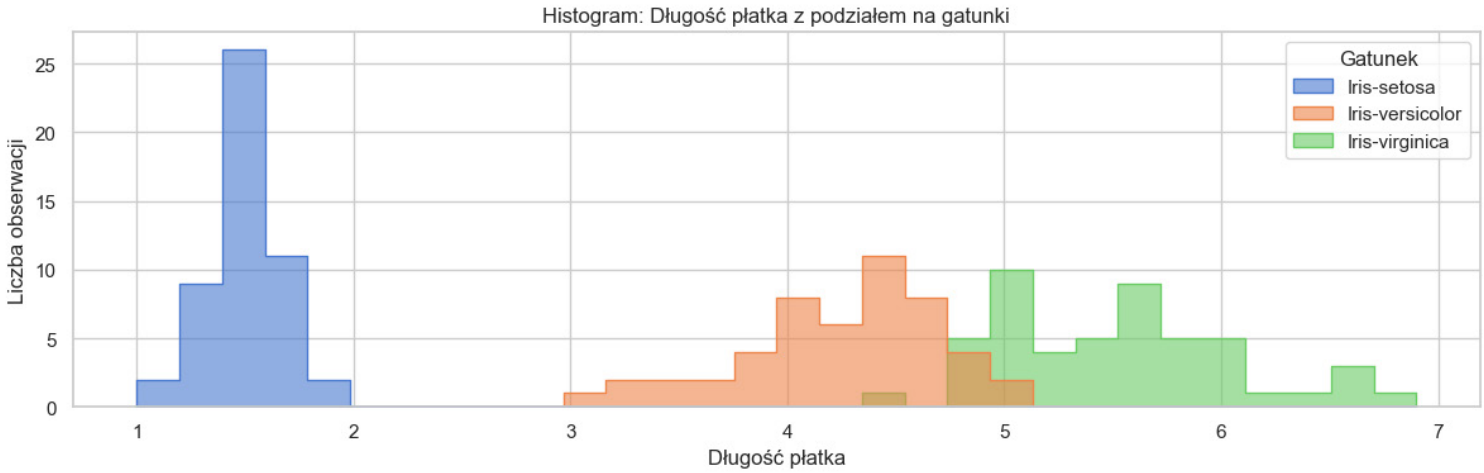
```
Długość kielicha      0
Szerokość kielicha    0
Długość płatka        0
Szerokość płatka      0
Gatunek               0
dtype: int64
```

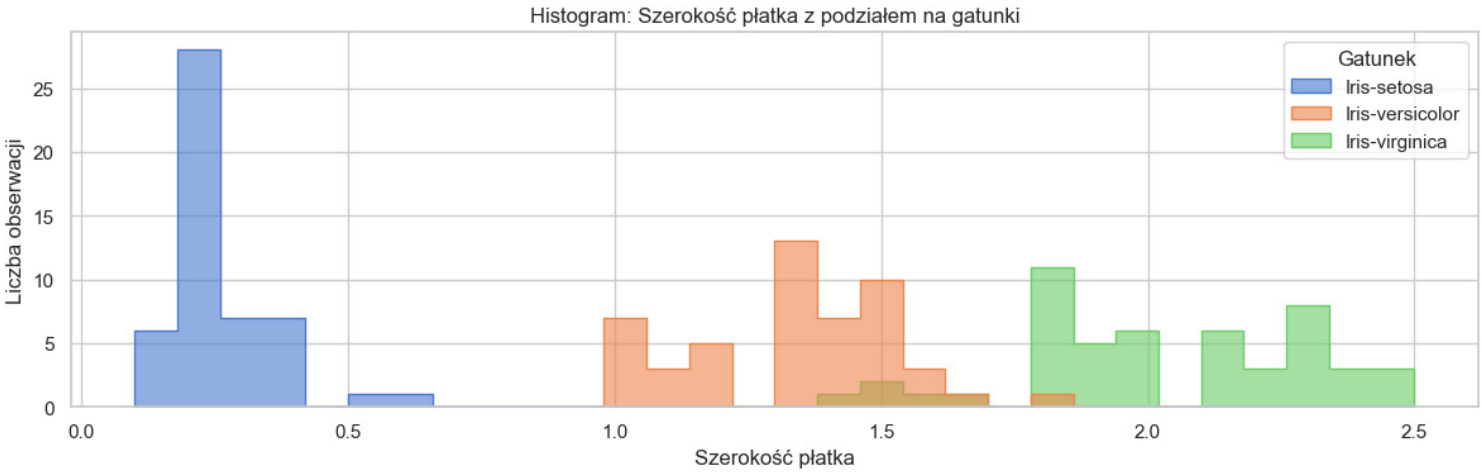
W zbiorze danych nie ma zadnych brakujących wartosci. Dane sa przejrzyste i gotowe do analizy

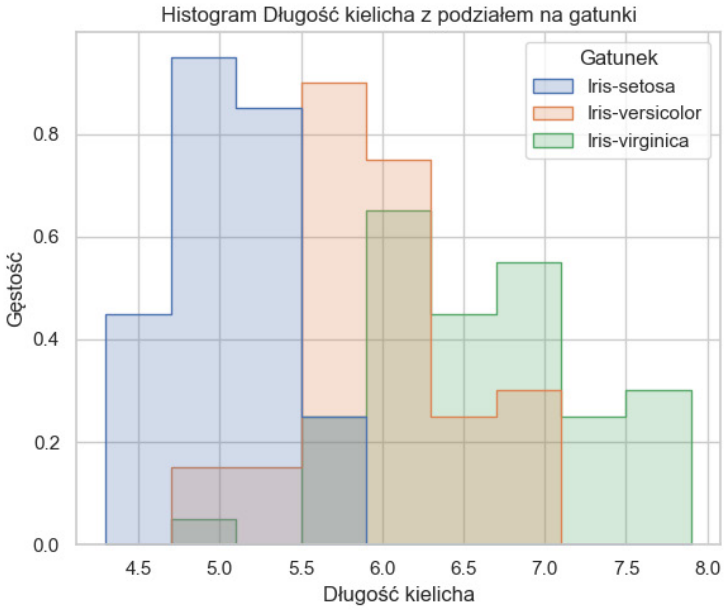
3 Analiza pojedynczych kolumn (zmiennych)

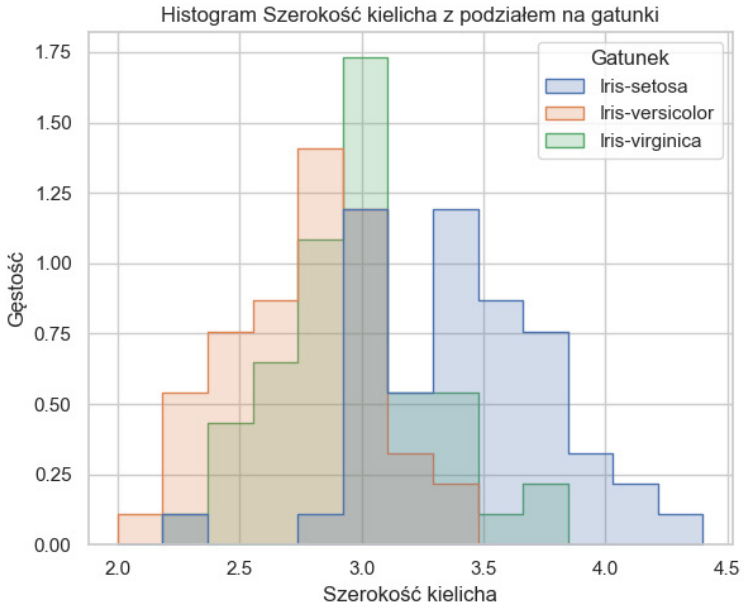


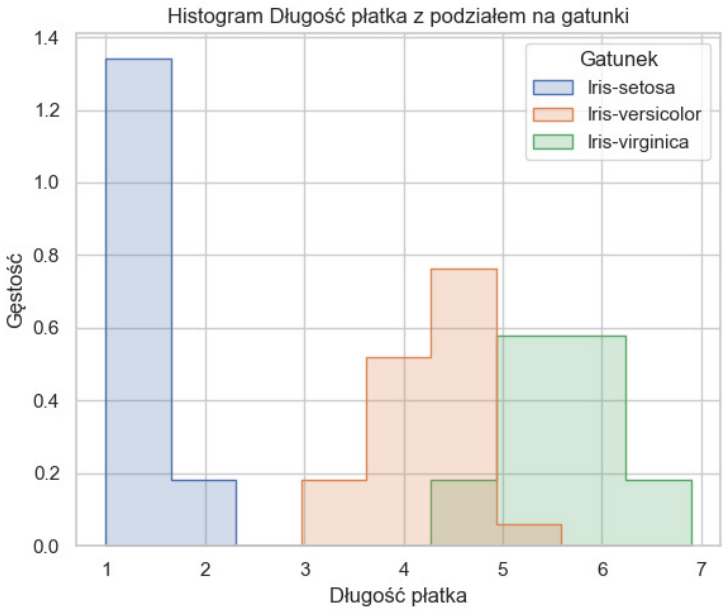


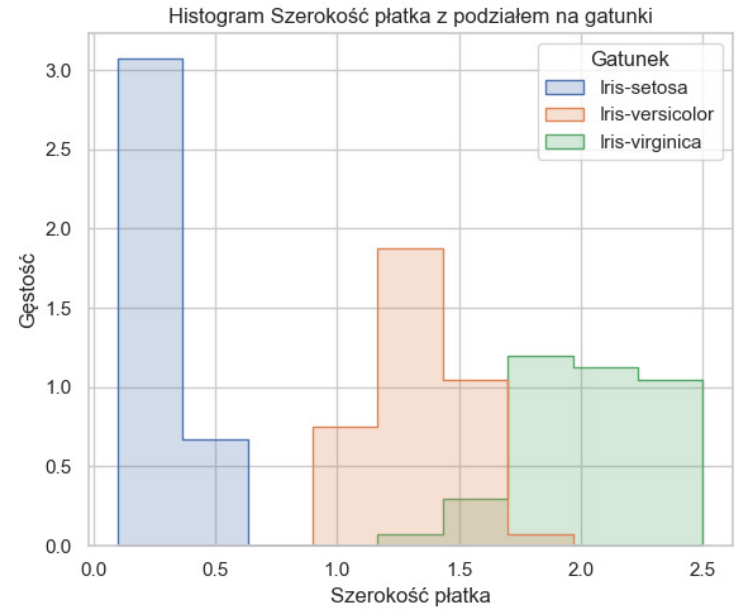



Histogram: Szerokość płatka z podziałem na gatunki
Gatunek
Iris-setosa
Iris-versicolor
Iris-virginica
Liczba obserwacji
Szerokość płatka


Histogram Długość kielicha z podziałem na gatunki
Gatunek
Iris-setosa
Iris-versicolor
Iris-virginica
Gęstość
Długość kielicha


Histogram Szerokość kielicha z podziałem na gatunki
Gatunek
Iris-setosa
Iris-versicolor
Iris-virginica
Gęstość
Szerokość kielicha


Histogram Długość płatka z podziałem na gatunki
Gatunek
Iris-setosa
Iris-versicolor
Iris-virginica
Gęstość
Długość płatka


Histogram Szerokość płatka z podziałem na gatunki
Gatunek
Iris-setosa
Iris-versicolor
Iris-virginica
Gęstość
Szerokość płatka

Wnioski z wykresów: Cechy płatka (długość i szerokość) są zdecydowanie najlepszymi wskaźnikami do rozróżniania gatunków irysów, natomiast cechy kielicha dostarczają mniej jednoznacznych informacji i mogą stanowić cechy uzupełniające w modelach klasyfikacyjnych.

Długość płatka (petal_length) i szerokość płatka (petal_width): Wykresy pokazują wyraźną trójmodalność, gdzie trzy różne grupy odpowiadają trzem gatunkom irysa. Iris-setosa ma najniższe wartości, Iris-versicolor wartości pośrednie, a Iris-virginica największe. Te cechy bardzo dobrze separują gatunki i są kluczowe w klasyfikacji.

Długość kielicha (sepal_length): Rozkład ma jeden wyraźny szczyt, choć lekko przesunięty w prawo. Gatunki są mniej wyraźnie rozdzielone, co sugeruje, że długość kielicha ma mniejszą zdolność separacji niż cechy płatka.

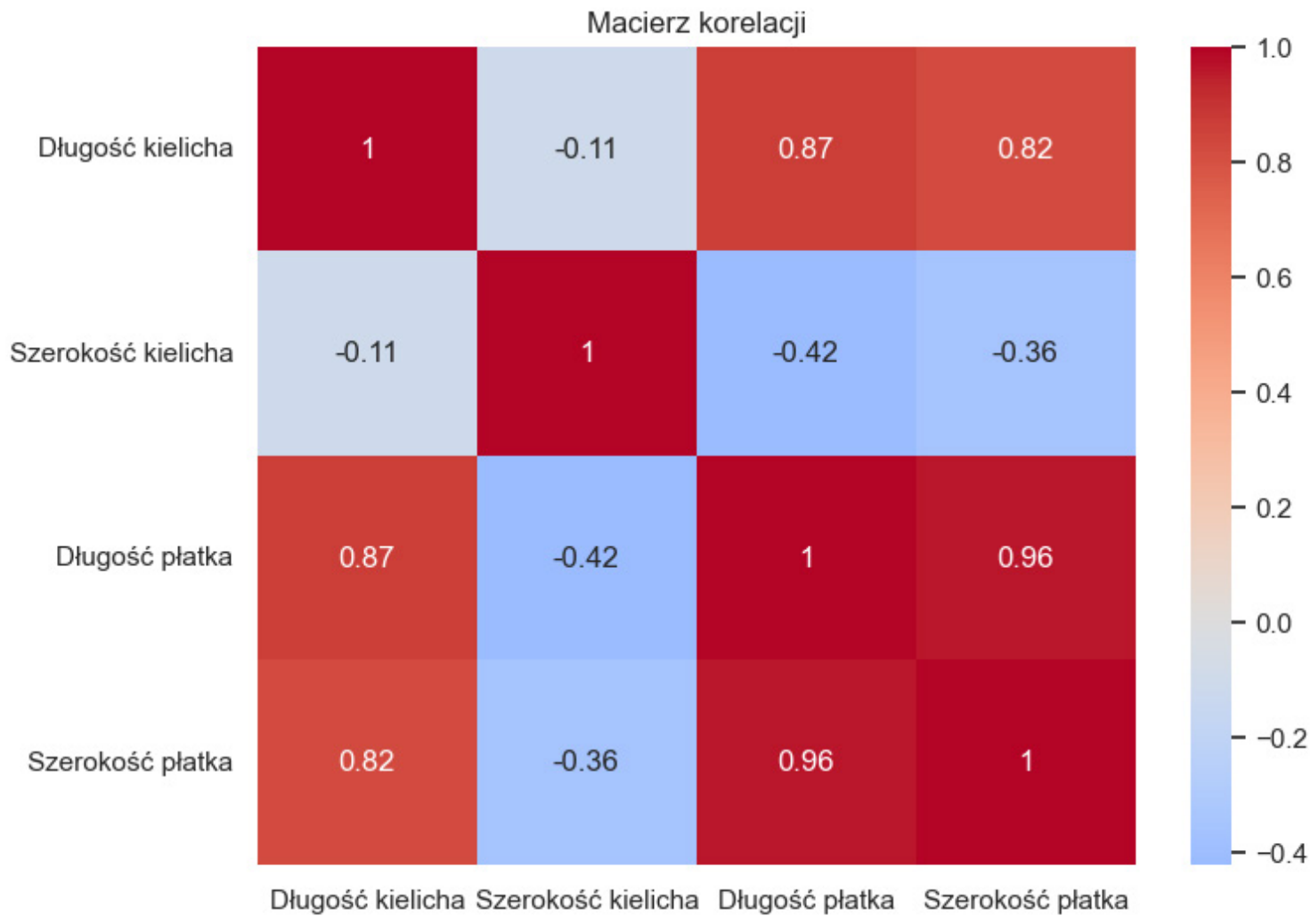
Szerokość kielicha (sepal_width): Rozkład jest bardziej rozproszony i mniej znormalizowany. Występują wartości odstające oraz niewielka dwumodalność, co wskazuje na subtelne różnice między gatunkami, jednak separacja jest słabsza niż przy płatkach.

4 Transformacje danych

Transformacje sa niepotrzebne w podanym zbiorze danych

5 Analiza relacji miedzy zmiennymi

Macierz Korelacji



Wnioski z macierzy korelacji

Silna dodatnia korelacja:

Długość płatka ↔ szerokość płatka

- Korelacja o wartosci 0.96.
- Oznacza, że im dłuższy płatek, tym jest on szerszy. Są to cechy silnie powiązane.

Długość kielicha ↔ długość płatka

- Korelacja również wysoka o wartości 0.87.
- Wskazuje, że im większy kielich, tym zwykle dłuższy płatek – może to wynikać z proporcjonalnego wzrostu kwiatu.

Umiarkowana korelacja: Długość kielicha ↔ szerokość płatka

- Korelacja o wartości 0.81
- Nadal znaczące powiązanie w ktorej większy kielich często współwystępuje z szerszym płatkiem.

Słaba lub ujemna korelacja:

Szerokość kielicha ↔ pozostałe cechy Korelacje niskie lub nawet ujemne:

- Z długością płatka: -0.42.
- Z szerokością płatka: -0.35 Wskazuje, że szerokość kielicha nie zmienia się proporcjonalnie do innych cech i może być najmniej przydatna do klasyfikacji.

Wyraźna separacja gatunku Iris-setosa: W większości wykresów rozrzutu (długość płatka, szerokość płatka) setosa tworzy osobny, wyraźnie oddzielony klaster, co sugeruje, że ten gatunek jest bardzo łatwy do odróżnienia od pozostałych. Jego płatki są znacznie krótsze i węższe niż u versicolor i virginica.

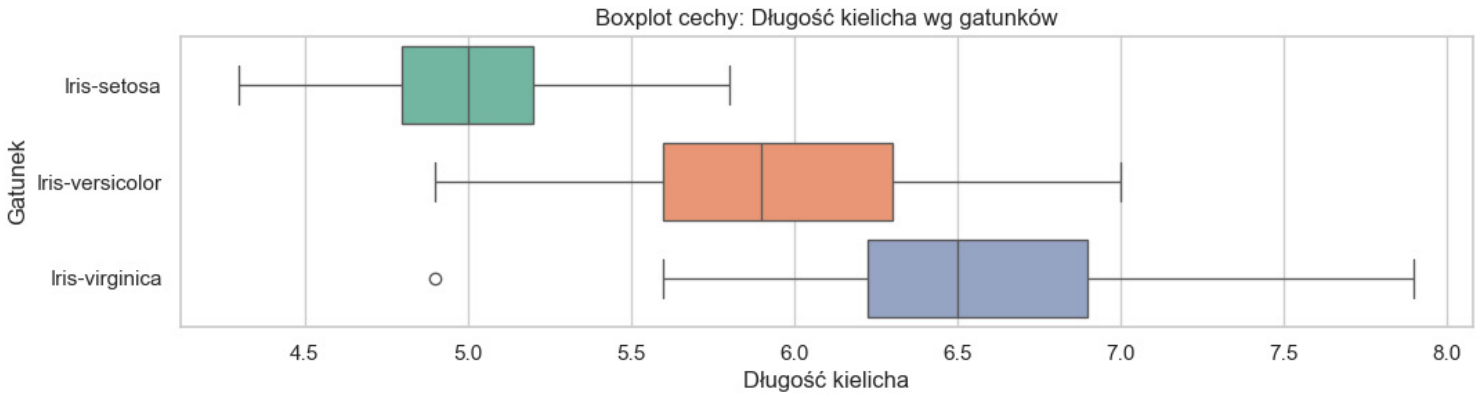
Relacja między długość płatka a szerokość płatka: Widoczna jest bardzo silna dodatnia zależność — im dłuższy płatek, tym jest on szerszy. Ten związek występuje w każdym gatunku, ale szczególnie mocno u virginica i versicolor.

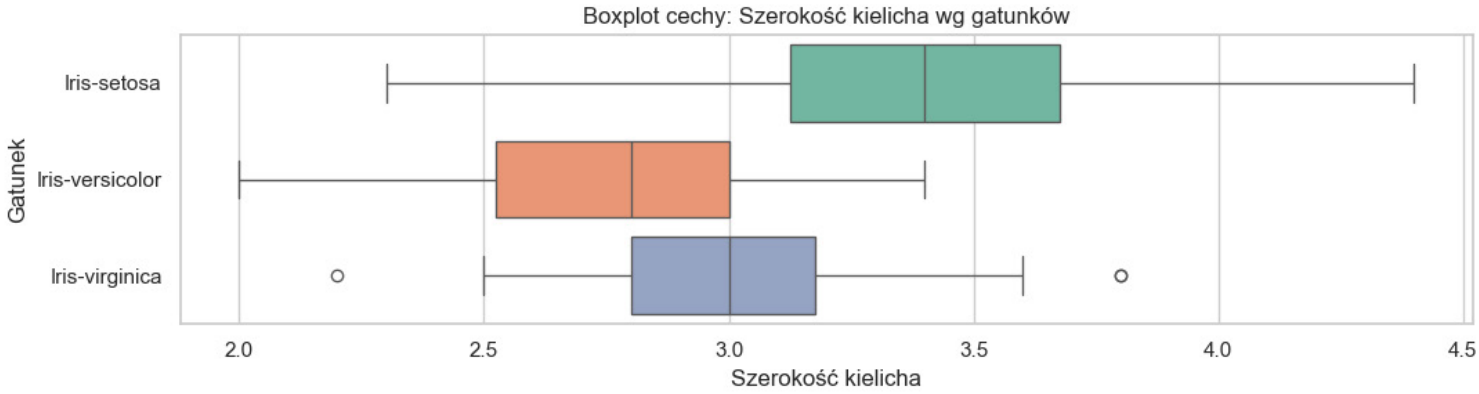
Nakładanie się (overlap) versicolor i virginica: Na wielu wykresach punkty tych dwóch gatunków się przenikają — szczególnie w parach cech związanych z kielichem (długość kielicha, szerokość kielicha). To sugeruje, że są trudniejsze do odróżnienia na podstawie tych cech.

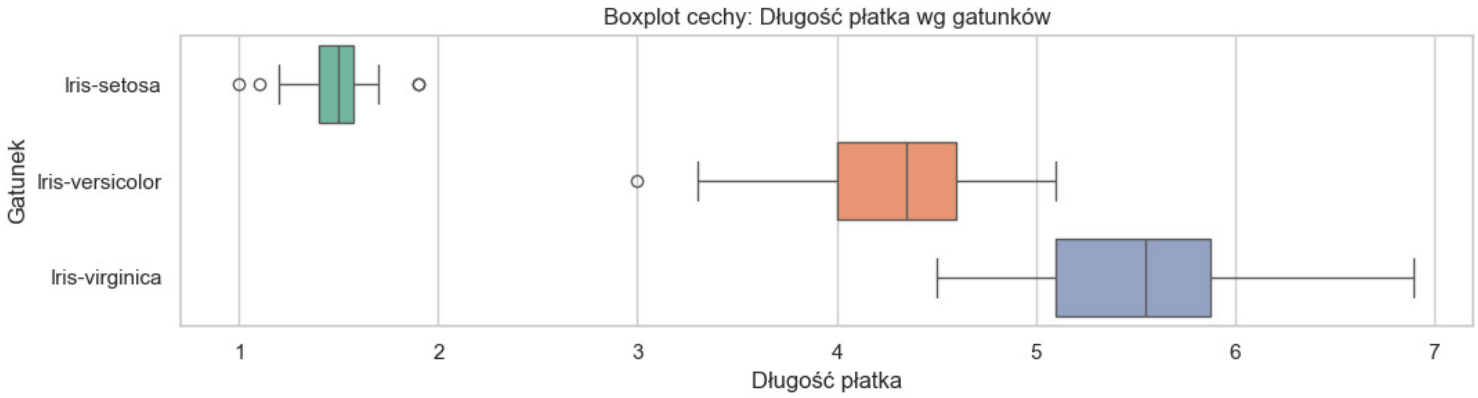
Słabsze relacje z szerokościa kielicha: Zmienna szerokość kielicha wykazuje słabe korelacje z innymi cechami i nie umożliwia wyraźnej separacji międzygatunkowej. Rozkład punktów jest bardziej rozproszony, co zmniejsza jej przydatność w klasyfikacji.

Wniosek: Cechy związane z płatkami są znacznie bardziej przydatne do rozróżniania gatunków niż cechy kielicha. Najbardziej charakterystyczny i łatwy do identyfikacji jest Iris-setosa, natomiast versicolor i virginica wymagają bardziej złożonych metod klasyfikacji ze względu na nakładające się wartości.

6 Analiza wartosci odstajacych







Wnioski z wartości odstających

Długość kielicha (sepal length)

- Brak wartości odstających dla Iris-setosa i Iris-versicolor.
- Jedna wartość odstająca dla Iris-virginica: 4.9 cm (znacznie mniejsza niż pozostałe wartości dla tego gatunku).
- Różnice między gatunkami są zauważalne, ale nie bardzo silne.

Szerokość kielicha (sepal width)

- Iris-virginica ma 2 wartości odstające. Wskazują na dużą zmienność tej cechy w tej grupie.
- Irys setosa i versicolor nie mają outlierów.
- Cecha ta ma słabsze właściwości separujące gatunki, szczególnie między versicolor a virginica.
- Największy rozrzut w pomiarach posiada iris-setosa.

Długość płatka (petal length)

- Iris-setosa: 3 wartości odstające — 1.0, 1.9, 1.9 cm. Prawdopodobnie wynika to z bardzo zwartego rozkładu danych dla tego gatunku.
- Iris-versicolor: 1 wartość odstająca — 3.0 cm.
- Iris-virginica: brak wartości odstających.
- Bardzo dobra cecha separująca dla iris-setosa wyraźnie niżej, virginica najwyżej.

Szerokość płatka (petal width)

- Iris-setosa posiada 2 outliery — 0.5 i 0.6 cm
- Pozostałe gatunki nie mają wartości odstających.
- Bardzo dobra cecha klasyfikacyjna dla iris-setosa odróżnia się wyraźnie od innych.
- Największy rozrzut w pomiarach posiada iris-virginica.

Podsumowanie:

Najbardziej istotne cechy to: długość i szerokość płatka — wyraźnie rozdzielają gatunki. Cecha o największej liczbie outlierów: długość płatka (Iris-setosa) Wartości odstające mogą wskazywać na zmienność wewnątrzgatunkową lub błędy pomiarowe, ale nie są licznie reprezentowane.

Wnioski końcowe

Zbiór danych zawiera trzy gatunki irysów: Iris-setosa, Iris-versicolor, Iris-virginica oraz cztery cechy: długości i szerokości płatków oraz szerokości i długości kielicha. Dane są kompletne, bez brakujących wartości, a rozkłady poszczególnych zmiennych wskazują na ich naturalne zróżnicowanie.

Iris-setosa wykazuje wyraźnie odmienny rozkład cech i tworzy osobną, zwartą grupę – jest najłatwiejsza do identyfikacji. Iris-versicolor i Iris-virginica częściowo się nakładają, zwłaszcza w zakresie długości kielicha, co może utrudniać ich jednoznaczne rozdzielenie na podstawie pojedynczej cechy.

Cechy związane z kielichem, szczególnie szerokość, wykazują słabsze korelacje i mniejszą wartość rozróżniającą. Między długość płatka a szerokość płatka występuje bardzo silna dodatnia korelacja (0.96). Długość płatka koreluje także z długością kielicha, co sugeruje, że większe kwiaty mają większe wymiary ogólne. Szerokość kielicha ma najsłabsze korelacje z pozostałymi cechami

Pomimo występowania kilku wartości odstających, zbiór cechuje się wysoką jakością i spójnością. Dane są kompletne, brak jest wartości pustych (null), co oznacza, że nie wymagają uzupełniania. ystępuje kilka wartości odstających: dla Iris-setosa: w długość płatka i szerokość płatka (np. bardzo krótkie płatki). Dla Iris-virginica: w szerokość kielicha (zarówno niskie, jak i wysokie wartości). Dla versicolor: pojedynczy outlier w długości płatka. Outliery nie są liczne i nie zaburzają znacząco struktury danych, ale warto je uwzględnić przy analizie modelu lub przy czyszczeniu danych.

Boxploty ujawniają, że długość płatka i szerokość płatka bardzo dobrze separują klasy. Pairplot pokazuje, że setosa jest wyraźnie oddzielona, natomiast versicolor i virginica tworzą bardziej rozmyte klastry. W wielu wykresach cechy szerokość kielicha i długość kielicha nie wykazują silnej separacji między gatunkami. Najbardziej istotne cechy to długość płatka, szerokość płatka są one silnie skorelowane i wyraźnie różnicujące klasy. Mniej przydatne cechy to szerokość kielicha, ktora słabo koreluje i nie rozróżnia dobrze gatunków.

Iris-setosa jest gatunkiem najłatwiejszym do odróżnienia i moze być poprawnie klasyfikowany nawet przez proste modele uczenia maszynowego.