

✓ Zbiór danych

Krótką charakterystyka zbioru

Zbiór danych *Student_Performance.csv* zawiera informacje o 10 000 uczniach. Każdy rekord zawiera dane o różnych czynnikach oraz wskaźniku wydajności, który reprezentuje ogólne wyniki ucznia.

Temat projektu i cel analizy badawczej

Tematem projektu jest **Analiza czynników wpływających na wyniki uczniów**

Celem analizy badawczej jest stworzenie modelu, który pozwoli przewidzieć wskaźnik wydajności ucznia na podstawie różnych czynników.

Zmienna objaśniana (zależna):


Performance Index (zmienna ilościowa): Miara ogólnych wyników każdego ucznia

Zmienne objaśniające:


- Hours Studied (zmienna ilościowa): Całkowita liczba godzin spędzonych na nauce przez każdego studenta
- Previous Scores (zmienna ilościowa): Wyniki uzyskane przez uczniów w poprzednich testach
- Extracurricular Activities (zmienna jakościowa): Czy uczeń uczestniczy w zajęciach pozalekcyjnych (Tak/Nie)
- Sleep Hours (zmienna ilościowa): Średnia liczba godzin snu ucznia w ciągu dnia
- Sample Question Papers Practiced (zmienna ilościowa): Liczba przykładowych arkuszy pytań przećwiczonych przez ucznia

```
import numpy as np
import pandas as pd
```

```
data = pd.read_csv('Student_Performance.csv')
data.head(4)
```



	Hours Studied	Previous Scores	Extracurricular Activities	Sleep Hours	Sample Question Papers Practiced	Performance Index
0	7	99	Yes	9	1	91.0
1	4	82	No	4	2	65.0
2	8	51	Yes	7	2	45.0




✓ Statystyczny opis struktury analizowanych cech

✓ Statystyki opisowe dla zmiennej objaśnianej - Performance Index

```
statystyki_zmiennej_objasnianej = data['Performance Index'].describe()
```

```
print(statystyki_zmiennej_objasnianej)
```



count	10000.000000
mean	55.224800
std	19.212558
min	10.000000
25%	40.000000
50%	55.000000
75%	71.000000
max	100.000000

Name: Performance Index, dtype: float64

- count - 10 000 uczniów, dla których dostępne są dane o wynikach
- mean - średnia arytmetyczna wyników uzyskanych przez uczniów

- std - odchylenie standardowe wynosi około 19.21, co wskazuje na przeciętne zróżnicowanie wyników uczniów wokół średniej
- min - najniższy uzyskany wynik to 10
- 25% - pierwszy kwartył wynosi 40, co oznacza, że 25% uczniów otrzymało wyniki mniejsze lub równe 40
- 50% - mediana wynosi 55, co oznacza, że 50% uczniów ma wskaźnik wydajności mniejszy lub równy 55, a pozostałe 50% ma wskaźnik wydajności większy lub równy 55
- 75% - trzeci kwartył wynosi 71, co oznacza, że 75% uczniów otrzymało wyniki mniejsze lub równe 71
- max - najwyższy uzyskany wynik to 100

Z otrzymanych wyników możemy dojść do następujących wniosków:

- średnia arytmetyczna jest nieco powyżej mediany, co sugeruje, że rozkład może być prawostronnie skośny
- wartość odchylenia standardowego informuje o przeciętnym rozproszeniu wyników
- wyniki uczniów mieszczą się w przedziale $<10, 100>$

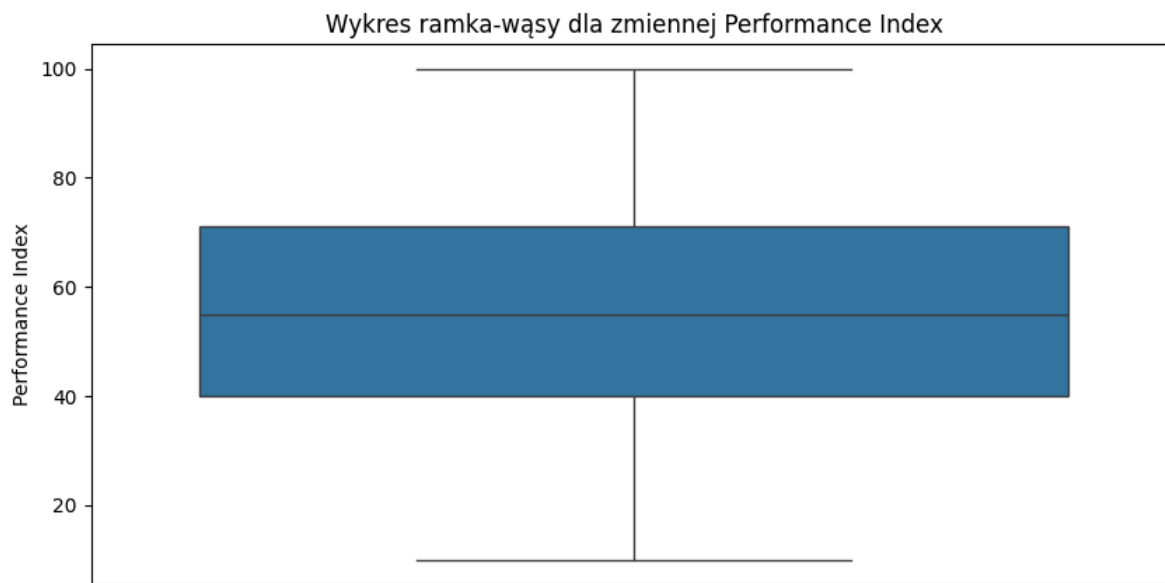
✓ Wykresy dla zmiennej objaśnianej - Performance Index

```
#wczytanie biblioteki potrzebnych do stworzenia wykresów w Python
import matplotlib.pyplot as plt
import seaborn as sns
```

✓ Wykres ramka wąsy dla mediany

```
plt.figure(figsize=(10, 5))
sns.boxplot(y=data['Performance Index'])
plt.title('Wykres ramka-wąsy dla zmiennej Performance Index')
plt.ylabel('Performance Index')

plt.show()
```



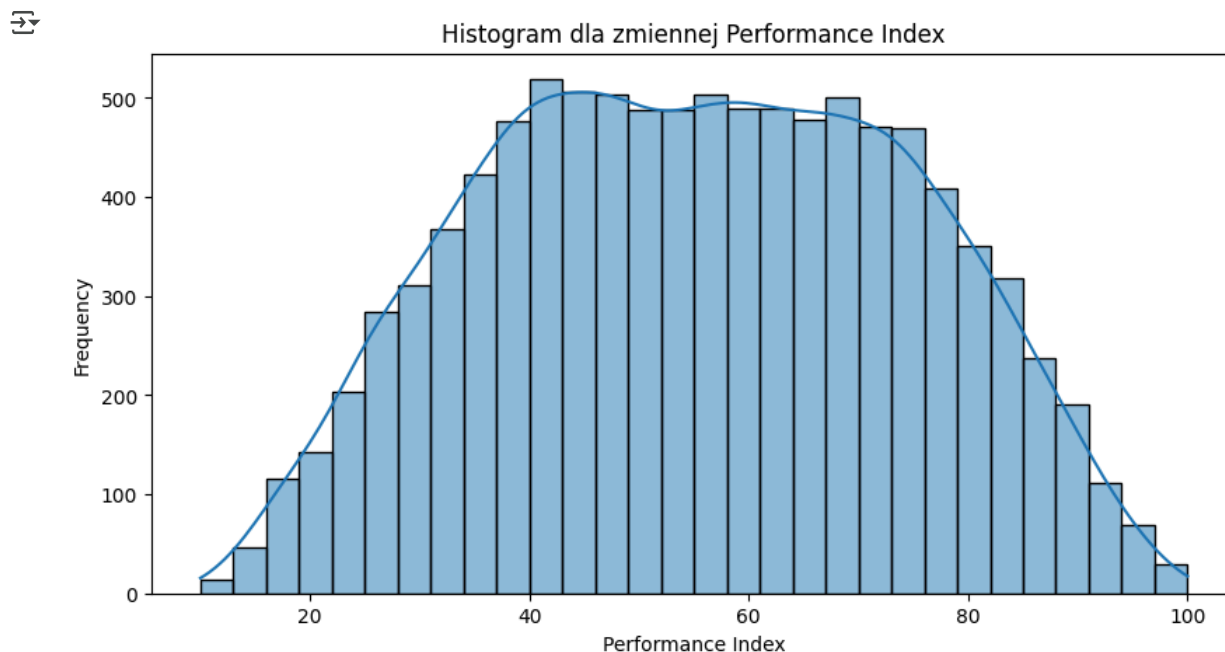
Z wykresu odczytujemy następujące własności:

- max wynik jest równy 100
- min wynik wynosi około 10
- pierwszy kwartył - 25% jest równy 40
- trzeci kwartył - 75% wynosi około 71
- pomiędzy pierwszym a trzecim kwartylem jest mediana - 50%. Wynosi ona 55.

Kształt wykresu na pierwszy rzut wydaje się symetryczny względem kreski z medianą, jednak jest delikatna asymetria prawostronna.

✓ Histogram

```
plt.figure(figsize=(10, 5))
sns.histplot(data['Performance Index'], bins=30, kde=True)
plt.title('Histogram dla zmiennej Performance Index')
plt.xlabel('Performance Index')
plt.ylabel('Frequency')
plt.show()
```



Histogram przedstawia liczbę uczniów, którzy uzyskali taki sam wynik. Oś pozioma (x) reprezentuje wartości wyników uzyskanych, natomiast oś pionowa (y) pokazuje liczbę uczniów, którzy osiągnęli daną wartość.

Dzięki analizie tego wykresu możemy dojść do następujących wniosków:

- Najczęściej występujące wyniki zawierają się w przedziale od 40 do 70.
- Jest to histogram zmiennej prawostronnie skośnej - skośność jest dodatnia.

✓ Skategoryzowany wykres ramka – wąsy oraz skategoryzowany histogram i/lub inne wykresy adekwatne do postawionego problemu badawczego

```
# Scatter plot dla "Hours Studied" vs. "Performance Index"
plt.figure(figsize=(10, 6))
sns.scatterplot(x='Hours Studied', y='Performance Index', data=data)
plt.title('Performance Index vs. Hours Studied')
plt.xlabel('Hours Studied')
plt.ylabel('Performance Index')
plt.show()
```

```
# Scatter plot dla "Previous Scores" vs. "Performance Index"
plt.figure(figsize=(10, 6))
sns.scatterplot(x='Previous Scores', y='Performance Index', data=data)
plt.title('Performance Index vs. Previous Scores')
plt.xlabel('Previous Scores')
plt.ylabel('Performance Index')
plt.show()
```

```
# Scatter plot dla "Sleep Hours" vs. "Performance Index"
plt.figure(figsize=(10, 6))
sns.scatterplot(x='Sleep Hours', y='Performance Index', data=data)
plt.title('Performance Index vs. Sleep Hours')
plt.xlabel('Sleep Hours')
```

```
plt.ylabel('Performance Index')
plt.show()

# Scatter plot dla "Sample Question Papers Practiced" vs. "Performance Index"
plt.figure(figsize=(10, 6))
sns.scatterplot(x='Sample Question Papers Practiced', y='Performance Index', data=data)
plt.title('Performance Index vs. Sample Question Papers Practiced')
plt.xlabel('Sample Question Papers Practiced')
plt.ylabel('Performance Index')
plt.show()
```

