

# SPRAWOZDANIE

Zajęcia: Nauka o danych I

Prowadzący: prof. dr hab. Vasyl Martsenyuk

Laboratorium Nr 2 Data 05.10.2024 Temat: Praktyczne zastosowanie podstawowych funkcji statystycznych w analizie danych Wariant drugi (2)	Bartosz Bieniek Informatyka II stopień, stacjonarne, 1 semestr, gr.A
---	---

## 1. Polecenie: wariant drugi zadania

<http://ghdx.healthdata.org/record/ihme-data/global-gdp-per-capita-1960-2050>

Oblicz podstawowe funkcje statystyczne zbioru danych z poprzednich zajęć:

1. Wczytywanie danych i wyświetlanie podstawowych informacji
2. Obliczanie podstawowych statystyk
3. Identyfikacja i obsługa brakujących danych
4. Wykrywanie wartości odstających
5. Analiza zależności między kolumnami
6. Przekształcanie danych

## 2. Opis programu opracowanego

<https://github.com/mindgoner/Studia/tree/master/Nauka%20o%20Danych/Laboratorium%202>

```
Nauka o Danych > Laboratorium 2 > Laboratorium2.py:nb > # 3. Identyfikacja i obsługa brakujących danych
Generate + Code + Markdown ▶ Run All ⌂ Restart ⌂ Clear All Outputs | Jupyter Variables Outline ... Python 3.11.9

# 1. Wczytywanie danych i wyświetlanie podstawowych informacji
import pandas as pd
df = pd.read_csv("IHME_GDP_1960_2050_V2021M09D22.CSV")
print(df.head())
print(df.info())
print(df.describe())

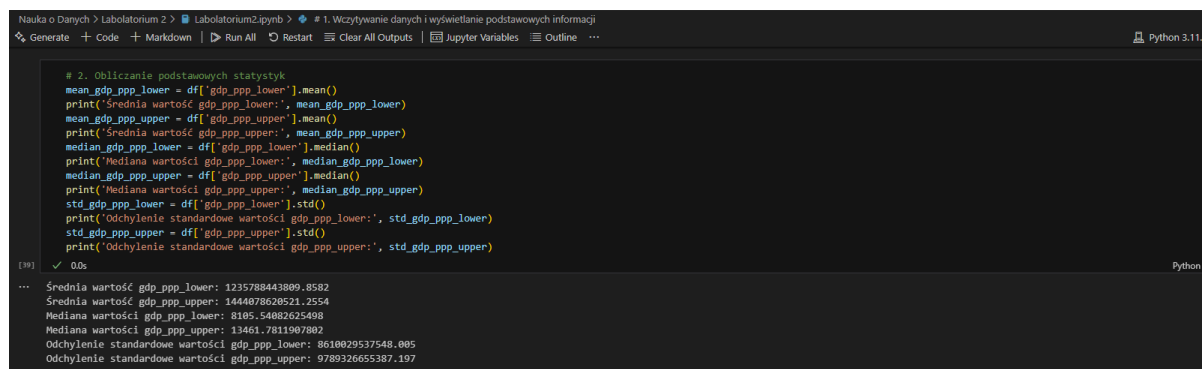
[38] ✓ 0.0s Python

location_id location_name iso3 level year gdp_ppp_mean gdp_ppp_lower \
0 1 Global G Global 1960 1.748345e+13 1.601915e+13
1 1 Global G Global 1961 1.813537e+13 1.659537e+13
2 1 Global G Global 1962 1.895328e+13 1.739039e+13
3 1 Global G Global 1963 1.965662e+13 1.811786e+13
4 1 Global G Global 1964 2.108575e+13 1.935664e+13

gdp_ppp_upper gdp_usd_mean gdp_usd_lower gdp_usd_upper
0 1.911586e+13 1.296863e+13 1.266890e+13 1.334177e+13
1 1.982493e+13 1.346097e+13 1.314767e+13 1.383021e+13
2 2.061477e+13 1.406576e+13 1.376060e+13 1.443746e+13
3 2.134993e+13 1.461831e+13 1.432132e+13 1.497693e+13
4 2.276791e+13 1.552986e+13 1.523498e+13 1.587998e+13
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 19838 entries, 0 to 19837
Data columns (total 11 columns):
# Column Non-Null Count Dtype
---
0 location_id 19838 non-null int64
1 location_name 19838 non-null object
2 iso3 18655 non-null object
3 level 19838 non-null object
4 year 19838 non-null int64
5 gdp_ppp_mean 19838 non-null float64
6 gdp_ppp_lower 19838 non-null float64
...
25% 1.624411e+03 1.395430e+03 1.828575e+03
50% 4.863298e+03 4.279291e+03 5.465731e+03
75% 1.997525e+04 1.795003e+04 2.223434e+04
max 1.19468e+14 1.017185e+14 1.239708e+14
Output is truncated. View as a scrollable element or open in a text editor. Adjust cell output settings...
```

Rys. 1. Fragment kodu Źródłowego

Wczytanie danych z pliku CSV pozwala poznać się z ich strukturą, typami kolumn oraz ogólną zawartością. Wyświetlenie pierwszych kilku wierszy i informacji o danych (`head()` i `info()`) pomaga zrozumieć, z czym pracujemy.

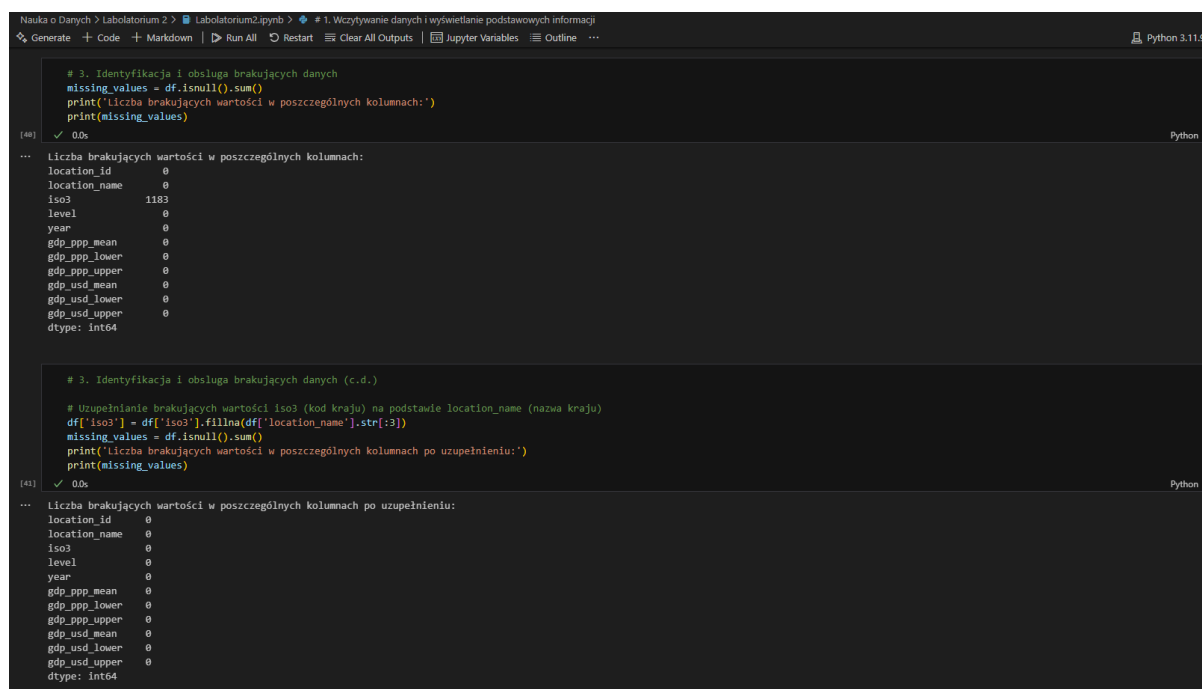


```
# 2. Obliczanie podstawowych statystyk
mean_gdp_ppp_lower = df['gdp_ppp_lower'].mean()
print('Średnia wartość gdp_ppp_lower:', mean_gdp_ppp_lower)
mean_gdp_ppp_upper = df['gdp_ppp_upper'].mean()
print('Średnia wartość gdp_ppp_upper:', mean_gdp_ppp_upper)
median_gdp_ppp_lower = df['gdp_ppp_lower'].median()
print('Mediana wartości gdp_ppp_lower:', median_gdp_ppp_lower)
median_gdp_ppp_upper = df['gdp_ppp_upper'].median()
print('Mediana wartości gdp_ppp_upper:', median_gdp_ppp_upper)
std_gdp_ppp_lower = df['gdp_ppp_lower'].std()
print('Odchylenie standardowe wartości gdp_ppp_lower:', std_gdp_ppp_lower)
std_gdp_ppp_upper = df['gdp_ppp_upper'].std()
print('Odchylenie standardowe wartości gdp_ppp_upper:', std_gdp_ppp_upper)
```

```
Średnia wartość gdp_ppp_lower: 1235788443809.8582
Średnia wartość gdp_ppp_upper: 1444878628521.2554
Mediana wartości gdp_ppp_lower: 8185.54882625498
Mediana wartości gdp_ppp_upper: 13461.7811907802
Odchylenie standardowe wartości gdp_ppp_lower: 8618029537548.805
Odchylenie standardowe wartości gdp_ppp_upper: 9789326655387.197
```

*Rys. 2. Fragment kodu Źródłowego*

Podstawowe statystyki opisowe, takie jak średnią, medianą, odchylenie standardowe czy zakres, dostarczają kluczowych informacji o rozkładzie danych. Umożliwiają szybkie wychwycenie potencjalnych anomalii i zrozumienie zakresu wartości w każdej kolumnie.



```
# 3. Identyfikacja i obsługa brakujących danych
missing_values = df.isnull().sum()
print('Liczba brakujących wartości w poszczególnych kolumnach:')
print(missing_values)
```

```
Liczba brakujących wartości w poszczególnych kolumnach:
location_id      0
location_name    0
iso3             1183
level            0
year             0
gdp_ppp_mean     0
gdp_ppp_lower    0
gdp_ppp_upper    0
gdp_usd_mean     0
gdp_usd_lower    0
gdp_usd_upper    0
dtype: int64
```

```
# 3. Identyfikacja i obsługa brakujących danych (c.d.)

# Uzupełnianie brakujących wartości iso3 (kod kraju) na podstawie location_name (nazwa kraju)
df['iso3'] = df['iso3'].fillna(df['location_name'].str[:3])
missing_values = df.isnull().sum()
print('Liczba brakujących wartości w poszczególnych kolumnach po uzupełnieniu:')
print(missing_values)
```

```
Liczba brakujących wartości w poszczególnych kolumnach po uzupełnieniu:
location_id      0
location_name    0
iso3             0
level            0
year             0
gdp_ppp_mean     0
gdp_ppp_lower    0
gdp_ppp_upper    0
gdp_usd_mean     0
gdp_usd_lower    0
gdp_usd_upper    0
dtype: int64
```

*Rys. 3. Fragment kodu Źródłowego*

Analiza brakujących wartości (`isnull().sum()`) pozwala określić, w których kolumnach występują luki w danych. Możemy uzupełniać brakujące dane (np. średnią, medianą) lub usuwać odpowiednie wiersze, aby zachować

integralność analizy. Jak można zaobserwować na powyższym fragmencie kodu, uzupełnianie danych bazując na różnych kolumnach pozwala uzupełnić braki danych w prosty sposób.

```
Nauka o Danych > Laboratorium 2 > Laboratorium2.ipynb > # 1. Wczytywanie danych i wyświetlanie podstawowych informacji
Generate + Code + Markdown Run All Restart Clear All Outputs Jupyter Variables Outline Python 3.11.9

# 4. Wykrywanie wartości odstających (używając metody IQR):
Q1 = df['gdp_ppp_upper'].quantile(0.25)
Q3 = df['gdp_ppp_upper'].quantile(0.75)
IQR = Q3 - Q1
lower_bound = Q1 - 1.5 * IQR
upper_bound = Q3 + 1.5 * IQR
outliers = df[(df['gdp_ppp_upper'] < lower_bound) | (df['gdp_ppp_upper'] > upper_bound)]
print('Wartości odstające w kolumnie gdp_ppp_upper:')
print(outliers)

(42) ✓ 0.0s Python

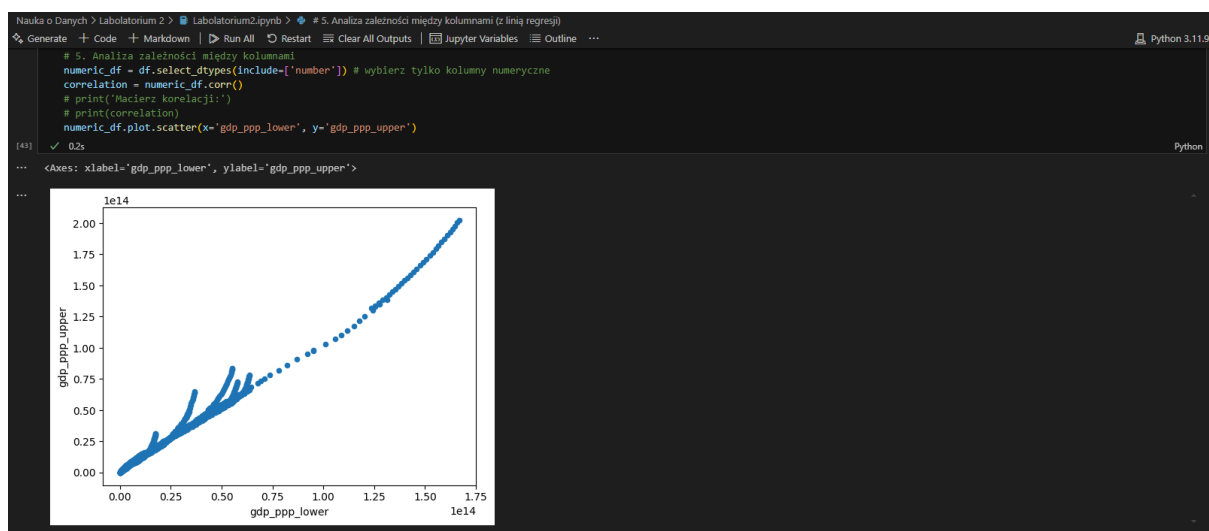
...
Wartości odstające w kolumnie gdp_ppp_upper:
   location_id location_name iso3 level year \
0            1         Global    G      Global  1990
1            1         Global    G      Global  1991
2            1         Global    G      Global  1992
3            1         Global    G      Global  1993
4            1         Global    G      Global  1994
...
19833      44578    Low income    Low World Bank Income Group  2846
19834      44578    Low income    Low World Bank Income Group  2847
19835      44578    Low income    Low World Bank Income Group  2848
19836      44578    Low income    Low World Bank Income Group  2849
19837      44578    Low income    Low World Bank Income Group  2850

   gdp_ppp_mean gdp_ppp_lower gdp_ppp_upper gdp_usd_mean \
0  1.748345e+13  1.601915e+13  1.911586e+13  1.296863e+13
1  1.813537e+13  1.659537e+13  1.982493e+13  1.346097e+13
2  1.895328e+13  1.739039e+13  2.061477e+13  1.406576e+13
3  1.965662e+13  1.811706e+13  2.134093e+13  1.461831e+13
4  2.106575e+13  1.935664e+13  2.276791e+13  1.552986e+13
...
19833  3.617310e+12  3.140835e+12  4.166469e+12  1.149318e+12
19834  3.724063e+12  3.225849e+12  4.292403e+12  1.186597e+12
19835  3.831942e+12  3.307609e+12  4.424674e+12  1.224062e+12
19836  3.941856e+12  3.398884e+12  4.560961e+12  1.262129e+12
...
19836  1.122895e+12  1.413991e+12
19837  1.151548e+12  1.457362e+12

[1901 rows x 11 columns]
Output is truncated. View as a scrollable element or open in a text editor. Adjust cell output settings...
```

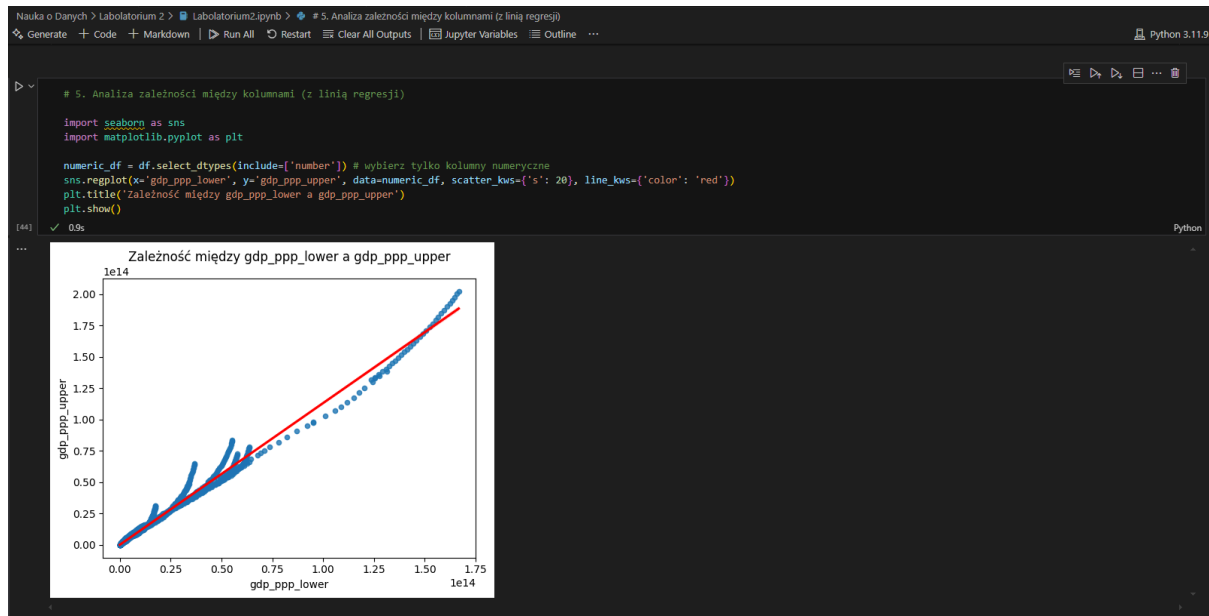
Rys. 4. Fragment kodu Źródłowego

Identyfikacja wartości odstających za pomocą wykresów pudełkowych (boxplot) lub analizy statystycznej (np. IQR) pomaga znaleźć dane, które mogą zakłócać wyniki. Odstające wartości można usunąć lub zbadać ich znaczenie w kontekście.



Rys. 5. Fragment kodu Źródłowego

Obliczanie współczynników korelacji pozwala zrozumieć, jak silnie powiązane są różne zmienne. Wizualizacja relacji na wykresach rozrzutu z liniami regresji ułatwia dostrzeżenie wzorców i trendów. Linię regresji pokazano modyfikując kod:



Rys. 6. Fragment kodu Źródłowego

Alternatywna wizualizacja z liniami regresji przy użyciu numpy:



Rys. 7. Fragment kodu Źródłowego

Niezależnie od wybranej metody wyniki są takie same, a wykresy podobne.

```
Nauka o Danych > Laboratorium 2 > Laboratorium2.ipynb > # 6. Przekształcanie danych
Generate + Code + Markdown | Run All | Restart | Clear All Outputs | Jupyter Variables | Outline ...

# 6. Przekształcanie danych
df['gdp_ppp_diff'] = df['gdp_ppp_upper'] - df['gdp_ppp_lower'] # Dodanie nowej kolumny

mean_diff = df.groupby('location_name')['gdp_ppp_diff'].mean() # Grupowanie wg nazwy kraju i obliczanie średniej wartości gdp_ppp_diff:
print('Średnia różnica między gdp_ppp_upper a gdp_ppp_lower dla poszczególnych krajów:')
print(mean_diff)

# Sortowanie po kolumnie year:
df = df.sort_values(by='year', ascending=True)
print(df.head())

[46] ✓ 0.0s Python

... Średnia różnica między gdp_ppp_upper a gdp_ppp_lower dla poszczególnych krajów:
location_name
Afghanistan      1413.412350
Albania           3310.023453
Algeria           4607.599093
American Samoa   3479.864825
Andorra           11830.056188
...
Venezuela (Bolivarian Republic of)  8096.820512
Viet Nam                     4157.516836
Yemen                       2963.275206
Zambia                      1817.071337
Zimbabwe                   1829.976392
Name: gdp_ppp_diff, Length: 216, dtype: float64

location_id  location_name iso3  level  year \
13832      171  Democratic Republic of the Congo  COD  Country  1968
6097         70                Finland  FIN  Country  1968
10465       132                Panama  PAN  Country  1968
4459         58                Estonia  EST  Country  1968
15015       185                Rwanda  RWA  Country  1968

...
gdp_ppp_mean  gdp_ppp_lower  gdp_ppp_upper  gdp_usd_mean \
13832    2529.408870      1497.917422    4030.682914    1192.389928
6097    13070.254728    10573.992294    15651.094527    13228.211834
10465    5315.615945      3772.621329      7078.008722     2825.271221
...
6097    12068.494529    14689.938429    5077.102233
10465    2570.171286     3120.499562     3305.387392
4459    5340.938458     5856.943423     8494.642253
15015    221.465258      334.067442       713.288674

Output is truncated. View as a scrollable element or open in a text editor. Adjust cell output settings...
```

Rys. 8. Fragment kodu Źródłowego

Przekształcenia, takie jak tworzenie nowych kolumn, sortowanie czy tworzenie nowych grup, pozwalają lepiej dostosować dane do analizy. Umożliwiają również poprawę efektywności algorytmów uczenia maszynowego i wykrycie ukrytych relacji.

### 3. Wnioski

Używanie dataframe'ów pozwala na szybkie zarządzanie i przetwarzanie danych. Brakujące dane mogą być uzupełnianie w oparciu o średnie wartości lub w oparciu o dane z innej kolumny, gdzie wartości w tej kolumnie są uzupełnione, jak przedstawiono na rysunku drugim.

Polecenia znajdujące się w treści zadania można realizować wskazanymi metodami lub alternatywnymi rozwiązaniami, korzystając z innych bibliotek, jak przedstawiono w piątym podpunkcie. Skorzystanie z biblioteki *seaborn* pozwalało dodać linię regresji na wykresie, która nie występowała w standardowej funkcji `plot.scatter(...)`. Alternatywnym rozwiązaniem było użycie plottera z wykorzystaniem biblioteki *numpy*.

Biblioteka *pandas* jest potężnym narzędziem do przetwarzania danych.

