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 7	Tomasz Pietrzyk Informatyka II stopień, niestacjonarne, 1semestr, gr.1a

1. Polecenie: wariant 7 zadania

Zadanie dotyczy pobrania danych z pliku, tworzenia ramki danych, wykonania poszczególnych zadań poniżej na podstawie odpowiedniego zbioru danych:

7. Global Burden of Disease Study 2019 (GBD 2019) Smoking Tobacco Use Prevalence 1990-2019 http://ghdx.healthdata.org/record/ihme-data/gbd-2019-smoking-tobacco-use-prevalence-1990-2019

2. Opis programu opracowanego (kody źródłowe, rzuty ekranu)

GitHub: https://github.com/TomekPietrzyk/NOD_I_2024_NS.git

In [1]:
```
import pandas as pd
import numpy as np
```

In [2]:
```
df = pd.read_csv('IHME_GBD_2019_SMOKING_TOB_1990_2019_NUM_SMOKERS_Y2021M05D27.CSV')
df
```

Out[2]:

	measure_name	location_id	location_name	sex_id	sex_name	age_group_id	age_group_name	year_id	val	upper	lower
0	Number of Smokers	1	Global	1	Male	29	15+ years	1990	8.031015e+08	8.096221e+08	7.959086e+08
1	Number of Smokers	1	Global	2	Female	29	15+ years	1990	1.891488e+08	1.930929e+08	1.855595e+08
2	Number of Smokers	1	Global	3	Both	29	15+ years	1990	9.922503e+08	1.000161e+09	9.847880e+08
3	Number of Smokers	1	Global	1	Male	29	15+ years	1991	8.138972e+08	8.200339e+08	8.069514e+08
4	Number of Smokers	1	Global	2	Female	29	15+ years	1991	1.905375e+08	1.944249e+08	1.869744e+08
...	...	...	...	...	...	...	...	...	...	...	...
20965	Number of Smokers	522	Sudan	2	Female	29	15+ years	2018	2.435999e+05	3.286166e+05	1.752508e+05
20966	Number of Smokers	522	Sudan	3	Both	29	15+ years	2018	2.610672e+06	2.833943e+06	2.409108e+06
20967	Number of Smokers	522	Sudan	1	Male	29	15+ years	2019	2.439150e+06	2.656579e+06	2.236450e+06
20968	Number of Smokers	522	Sudan	2	Female	29	15+ years	2019	2.500800e+05	3.345384e+05	1.816686e+05
20969	Number of Smokers	522	Sudan	3	Both	29	15+ years	2019	2.689230e+06	2.918332e+06	2.480656e+06

20970 rows × 11 columns

In [3]:
```
df.dropna()
```

Out[3]:

	measure_name	location_id	location_name	sex_id	sex_name	age_group_id	age_group_name	year_id	val	upper	lower
0	Number of Smokers	1	Global	1	Male	29	15+ years	1990	8.031015e+08	8.096221e+08	7.959086e+08
1	Number of Smokers	1	Global	2	Female	29	15+ years	1990	1.891488e+08	1.930929e+08	1.855595e+08
2	Number of Smokers	1	Global	3	Both	29	15+ years	1990	9.922503e+08	1.000161e+09	9.847880e+08
3	Number of Smokers	1	Global	1	Male	29	15+ years	1991	8.138972e+08	8.200339e+08	8.069514e+08
4	Number of Smokers	1	Global	2	Female	29	15+ years	1991	1.905375e+08	1.944249e+08	1.869744e+08
...	...	...	...	...	...	...	...	...	...	...	...
20965	Number of Smokers	522	Sudan	2	Female	29	15+ years	2018	2.435999e+05	3.286166e+05	1.752508e+05
20966	Number of Smokers	522	Sudan	3	Both	29	15+ years	2018	2.610672e+06	2.833943e+06	2.409108e+06
20967	Number of Smokers	522	Sudan	1	Male	29	15+ years	2019	2.439150e+06	2.656579e+06	2.236450e+06
20968	Number of Smokers	522	Sudan	2	Female	29	15+ years	2019	2.500800e+05	3.345384e+05	1.816686e+05
20969	Number of Smokers	522	Sudan	3	Both	29	15+ years	2019	2.689230e+06	2.918332e+06	2.480656e+06

20970 rows × 11 columns

In [4]:
```
series1 = df['val']
series1
```

Out[4]:
```
0        8.031015e+08
1        1.891488e+08
2        9.922503e+08
3        8.138972e+08
4        1.905375e+08
             ...
20965    2.435999e+05
20966    2.610672e+06
20967    2.439150e+06
20968    2.500800e+05
20969    2.689230e+06
Name: val, Length: 20970, dtype: float64
```

In [5]:
```
# obliczanie średniej

np.mean(series1)
```

Out[5]:
```
12428071.383604305
```

In [6]:
```
#obliczanie mediany

np.median(series1)
```

Out[6]:
```
577712.25205
```

In [7]:
```
#Odchylenie standardowe

np.std(series1)
```

Out[7]:
```
64890362.21057887
```

In [8]:
```
#Wariancja

np.var(series1)
```

Out[8]:
```
4210759107820122.0
```

```
In [9]: series2 = df['upper']

        series2

Out[9]: 0        8.096221e+08
        1        1.930929e+08
        2        1.000161e+09
        3        8.200339e+08
        4        1.944249e+08
                     ...
        20965    3.286166e+05
        20966    2.833943e+06
        20967    2.656579e+06
        20968    3.345384e+05
        20969    2.918332e+06
        Name: upper, Length: 20970, dtype: float64

In [10]: #Koleracja

         correlation = np.corrcoef(series1,series2) [0,1]
         print(correlation)

         0.9999762487761583

In [11]: #Kowariancja

         covariance = np.cov(series1,series2) [0,1]
         covariance

Out[11]: 4254193895188547.0
```

3. Wnioski

Średnia równa 12428071.383604305 przy odchyleniu standardowym równym 64890362.21057887 sugeruje, że rozrzut danych jest bardzo wysoki i rekordy nie są bliskie mediany. Występują skrajne wartości, odbiegające w sposób bardzo znaczny od średniej. Potwierdzenie powyższej tezy powiązane jest z wysoką wartością wariancji równą 4210759107820122.0. Wynik korelacji dwóch analizowanych kolumn na poziomie 0.9999762487761583 wskazuje na silną zależność liniową kolumn i ich wzajemnym powiązaniu.