SPRAWOZDANIE

Zajęcia: Nauka o danych I

Prowadzący: prof. dr hab. Vasyl Martsenyuk

Laboratorium Nr 2 Data 19.10.2024 Temat: "Praktyczne zastosowanie podstawowych funkcji statystycznych w analizie danych" Wariant 6	Dawid Klimek Informatyka II stopień, niestacjonarne, 1semestr, gr.1A

1. Polecenie: wariant 6 zadania

Oblicz podstawowe funkcje statystyczne używając zbiory danych z poprzedniego zajęcia

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

[9]:
```
import pandas as pd
import numpy as np
```

[10]:
```
df = pd.read_csv('IHME_GBD_2019_CHEWING_TOB_1990_2019_DATA_Y2021M05D27.CSV',encoding='latin1')
df
```

[10]:

	measure_id	measure_name	location_id	location_name	sex_id	sex_name	age_group_id	age_group_name	rei_id	rei_name	metric_id	metric_name	year_id	val
0	5	Prevalence	1	Global	1	Male	8	15 to 19	332	Chewing tobacco	3	Rate	1990	0.038740
1	5	Prevalence	1	Global	2	Female	8	15 to 19	332	Chewing tobacco	3	Rate	1990	0.011356
2	5	Prevalence	1	Global	1	Male	8	15 to 19	332	Chewing tobacco	3	Rate	1991	0.039253
3	5	Prevalence	1	Global	2	Female	8	15 to 19	332	Chewing tobacco	3	Rate	1991	0.011516
4	5	Prevalence	1	Global	1	Male	8	15 to 19	332	Chewing tobacco	3	Rate	1992	0.039863
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
350545	5	Prevalence	522	Sudan	3	Both	27	Age standardized	332	Chewing tobacco	3	Rate	2015	0.029518
350546	5	Prevalence	522	Sudan	3	Both	27	Age standardized	332	Chewing tobacco	3	Rate	2016	0.029855
350547	5	Prevalence	522	Sudan	3	Both	27	Age standardized	332	Chewing tobacco	3	Rate	2017	0.029768
350548	5	Prevalence	522	Sudan	3	Both	27	Age standardized	332	Chewing tobacco	3	Rate	2018	0.029760
350549	5	Prevalence	522	Sudan	3	Both	27	Age standardized	332	Chewing tobacco	3	Rate	2019	0.029752

350550 rows × 16 columns

[11]:
```
df.dropna()
```

[11]:

	measure_id	measure_name	location_id	location_name	sex_id	sex_name	age_group_id	age_group_name	rei_id	rei_name	metric_id	metric_name	year_id	val
0	5	Prevalence	1	Global	1	Male	8	15 to 19	332	Chewing tobacco	3	Rate	1990	0.038740
1	5	Prevalence	1	Global	2	Female	8	15 to 19	332	Chewing tobacco	3	Rate	1990	0.011356
2	5	Prevalence	1	Global	1	Male	8	15 to 19	332	Chewing tobacco	3	Rate	1991	0.039253
3	5	Prevalence	1	Global	2	Female	8	15 to 19	332	Chewing tobacco	3	Rate	1991	0.011516
4	5	Prevalence	1	Global	1	Male	8	15 to 19	332	Chewing tobacco	3	Rate	1992	0.039863
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
350545	5	Prevalence	522	Sudan	3	Both	27	Age standardized	332	Chewing tobacco	3	Rate	2015	0.029518
350546	5	Prevalence	522	Sudan	3	Both	27	Age standardized	332	Chewing tobacco	3	Rate	2016	0.029855
350547	5	Prevalence	522	Sudan	3	Both	27	Age standardized	332	Chewing tobacco	3	Rate	2017	0.029768
350548	5	Prevalence	522	Sudan	3	Both	27	Age standardized	332	Chewing tobacco	3	Rate	2018	0.029760
350549	5	Prevalence	522	Sudan	3	Both	27	Age standardized	332	Chewing tobacco	3	Rate	2019	0.029752

350550 rows × 16 columns

[12]:
```
ser = df['val']
ser
```

[12]:
```
0         0.038740
1         0.011356
2         0.039253
3         0.011516
4         0.039863
            ...
350545    0.029518
350546    0.029855
350547    0.029768
350548    0.029760
350549    0.029752
Name: val, Length: 350550, dtype: float64
```

[13]:
```
# obliczanie średniej

np.mean(ser)
```

[13]:
```
0.020178816871399226
```

[14]:
```
#obliczanie mediany

np.median(ser)
```

[14]:
```
0.0048691325
```

[15]:
```
#Odchylenie standardowe

np.std(ser)
```

[15]:
```
0.049593893684802264
```

[16]:
```
#wariancja

np.var(ser)
```

[16]:
```
0.0024595429081947
```

[17]:
```
ser2 = df['lower']

ser2
```

[17]:
```
0         0.027147
1         0.007779
2         0.027608
3         0.007906
4         0.027800
            ...
350545    0.024255
350546    0.024500
350547    0.024428
350548    0.024462
350549    0.024318
Name: lower, Length: 350550, dtype: float64
```

[18]:
```
#Koleracja

correlation = np.corrcoef(ser,ser2) [0,1]
print(correlation)
```

```
0.9780038085580645
```

[19]:
```
#Kowariancja

covariance = np.cov(ser,ser2) [0,1]
covariance
```

[19]:
```
0.0016552467290792375
```

3. Wnioski

Wartości są bardzo zbliżone, operując na niskich wartościach, korelacja pomiędzy ser i ser 2 jest bardzo zbliżona do 1 co oznaczy że korelacja jest dodatnia i bardzo silna. Niska wartość wariancji świadczy o niskim rozrzucie danych.