

SPRAWOZDANIE

Zajęcia: Nauka o danych I

Prowadzący: prof. dr hab. Vasyl Martsenyuk

| | |
|---|---|
| Laboratorium Nr 2 Data 07.11.2024 Temat: Praktyczne Zastosowanie Podstawowych Funkcji Statystycznych w Analizie Danych Wariant 8 | Imię Nazwisko Hubert Mentel Informatyka II stopień, niestacjonarne, 1 semestr, gr.1a |
|---|---|

1. Załóżmy, że mamy dane dotyczące sprzedaży i cen w dwóch sklepach i chcemy zbadać, czy istnieje między nimi zależność oraz ocenić zmienność sprzedaży: wariant 8 zadania

ORB General Population COVID-19 Health Services Disruption Survey
2020 <http://ghdx.healthdata.org/record/ihme-data/orb-general-population-covid-19-health-services-disruption-survey-2020>

Pliki dostępne są na GitHubie pod linkiem:
<https://github.com/HubiPX/NOD/tree/master/Zadanie%202>

2. Opis programu opracowanego (kody Źródłowe, zrzuty ekranu)

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

```
[3]: df = pd.read_csv('IHME_ORB_C19HSDS_2020_Y2020M12D03.CSV')
df
```

[3]:

| | SbjNum | NetDuration | InterviewTimeVStart | InterviewTimeVEnd | Date | Srvyr | Country | LANG | R1 | R1_5 | ... | G11_Other | G11_99 | FinalOutcome | NumOfV |
|------|-----------|-------------|---------------------|-------------------|----------------|-------|---------|------|-----|------|-----|-----------|--------|--------------|--------|
| 0 | 133476254 | 0:10:14 | 7/17/2020 13:53 | 7/17/2020 14:26 | 7/17/2020 8:53 | 3232 | 2 | 1 | 9 | 15.0 | ... | NaN | NaN | 1 | |
| 1 | 133281846 | 0:22:16 | 7/10/2020 12:53 | 7/10/2020 14:47 | 7/10/2020 7:53 | 3206 | 2 | 4 | 12 | 22.0 | ... | NaN | NaN | 1 | |
| 2 | 133280780 | 0:19:23 | 7/10/2020 12:35 | 7/10/2020 12:54 | 7/10/2020 7:35 | 3202 | 2 | 3 | 10 | 13.0 | ... | NaN | NaN | 1 | |
| 3 | 133281834 | 0:10:11 | 7/10/2020 10:21 | 7/10/2020 10:32 | 7/10/2020 5:21 | 3212 | 2 | 1 | 12 | 9.0 | ... | NaN | NaN | 1 | |
| 4 | 133491249 | 0:09:59 | 7/18/2020 8:27 | 7/18/2020 8:39 | 7/18/2020 3:27 | 3225 | 2 | 3 | 11 | 28.0 | ... | NaN | NaN | 1 | |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 3053 | 133323839 | 0:09:03 | 7/11/2020 12:44 | 7/11/2020 12:53 | 7/11/2020 5:44 | 3012 | 1 | 7 | 8 | NaN | ... | NaN | NaN | 1 | |
| 3054 | 133305818 | 0:06:57 | 7/11/2020 16:18 | 7/11/2020 16:25 | 7/11/2020 9:18 | 3008 | 1 | 1 | 3 | NaN | ... | NaN | NaN | 1 | |
| 3055 | 133260048 | 0:21:46 | 7/9/2020 11:49 | 7/9/2020 12:12 | 7/9/2020 4:49 | 3004 | 1 | 1 | 7 | NaN | ... | NaN | NaN | 1 | |
| 3056 | 133305807 | 0:06:50 | 7/11/2020 9:05 | 7/11/2020 9:12 | 7/11/2020 2:05 | 3008 | 1 | 1 | 3 | NaN | ... | NaN | NaN | 1 | |
| 3057 | 133352713 | 0:09:20 | 7/13/2020 9:56 | 7/13/2020 14:44 | 7/13/2020 2:56 | 3003 | 1 | 1 | 2 | NaN | ... | NaN | NaN | 1 | |

3058 rows × 247 columns

```
[5]: # srednia
series1 = df["weight_combined"]
series1
np.mean(series1)

[5]: 1.0009869195359713

[7]: # mediana
series2 = df["gk_weight"]
series2
np.median(series2)

[7]: 2.046237019

[9]: # odchylenie standardowe
np.std(series1)

[9]: 0.4030393741519619

[11]: # wariancja
np.var(series1)

[11]: 0.16244073711680515

[13]: # korelacja
np.corrcoef(series1, series2)[0, 1]

[13]: 0.22123768551326237

[15]: # kowariancja
np.cov(series1, series2)[0, 1]

[15]: 0.07658173104603606
```

3. Wnioski

Dzięki funkcjom `mean`, `median`, `std`, `var`, `corrcoef` oraz `cov()` w Jupyterze możemy wszechstronnie badać dane, analizując trendy, zmienność i zależności. Średnia (`mean`) i mediana (`median`) pozwalają określić przeciętny i typowy poziom zmiennej, przy czym mediana jest odporna na wartości odstające. Wariancja (`var`) i odchylenie standardowe (`std`) pokazują rozproszenie danych wokół średniej, co jest kluczowe przy ocenie zmienności i stabilności wyników. Funkcja `corrcoef` pozwala zrozumieć siłę i kierunek zależności między zmiennymi, a kowariancja (`cov`) – ich wspólne zmiany, choć bez jednoznacznego określania ich siły. Razem te narzędzia pomagają w odkrywaniu wzorców, trendów i zależności w zbiorze danych, umożliwiając bardziej dogłębną analizę.