SPRAWOZDANIE

Zajęcia: Nauka o danych I

Prowadzący: prof. dr hab. Vasyl Martsenyuk

Laboratorium Nr 3 Data 19.10.2024 Temat: "Wykorzystanie pakietu Pandas do manipulacji i przetwarzania danych w Pythonie" Wariant 6	Dawid Klimek Informatyka II stopień, niestacjonarne, 1semestr, gr.1A

1. Polecenie: wariant 6 zadania

Zadanie 1: Wczytywanie danych i wyświetlanie

podstawowych informacji

Zadanie 2: Obliczanie podstawowych statystyk

Zadanie 3: Identyfikacja i obsługa brakujących danych

Zadanie 4: Wykrywanie wartości odstających

Zadanie 5: Analiza zależności między kolumnami

Zadanie 6: Przekształcanie danych

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

[1]:
```
import pandas as pd

df = pd.read_csv('IHME_GBD_2019_CHEWING_TOB_1990_2019_DATA_Y2021M05D27.CSV', encoding='latin1')

print(df.head())
```

```
   measure_id measure_name  location_id location_name  sex_id sex_name  \
0           5   Prevalence            1        Global       1     Male
1           5   Prevalence            1        Global       2   Female
2           5   Prevalence            1        Global       1     Male
3           5   Prevalence            1        Global       2   Female
4           5   Prevalence            1        Global       1     Male

   age_group_id age_group_name  rei_id        rei_name  metric_id  \
0             8       15 to 19     332  Chewing tobacco          3
1             8       15 to 19     332  Chewing tobacco          3
2             8       15 to 19     332  Chewing tobacco          3
3             8       15 to 19     332  Chewing tobacco          3
4             8       15 to 19     332  Chewing tobacco          3

  metric_name  year_id       val     upper     lower
0        Rate     1990  0.038740  0.055586  0.027147
1        Rate     1990  0.011356  0.017594  0.007779
2        Rate     1991  0.039253  0.055838  0.027608
3        Rate     1991  0.011516  0.017807  0.007906
4        Rate     1992  0.039863  0.056448  0.027800
```

[2]:
```
print(df.info())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 350550 entries, 0 to 350549
Data columns (total 16 columns):
 #   Column          Non-Null Count   Dtype
---  ------          --------------   -----
 0   measure_id      350550 non-null  int64
 1   measure_name    350550 non-null  object
 2   location_id     350550 non-null  int64
 3   location_name   350550 non-null  object
 4   sex_id          350550 non-null  int64
 5   sex_name        350550 non-null  object
 6   age_group_id    350550 non-null  int64
 7   age_group_name  350550 non-null  object
 8   rei_id          350550 non-null  int64
 9   rei_name        350550 non-null  object
 10  metric_id       350550 non-null  int64
 11  metric_name     350550 non-null  object
 12  year_id         350550 non-null  int64
 13  val             350550 non-null  float64
 14  upper           350550 non-null  float64
 15  lower           350550 non-null  float64
dtypes: float64(3), int64(7), object(6)
memory usage: 42.8+ MB
None
```

[3]:
```
print(df.describe())
```

```
       measure_id    location_id         sex_id   age_group_id    rei_id  \
count    350550.0  350550.000000  350550.000000  350550.000000  350550.0
mean          5.0     135.639024       2.000000      29.421053     332.0
std           0.0      98.136414       0.816498      48.993427       0.0
min           5.0       1.000000       1.000000       8.000000     332.0
25%           5.0      62.000000       1.000000      12.000000     332.0
50%           5.0     122.000000       2.000000      17.000000     332.0
75%           5.0     182.000000       3.000000      27.000000     332.0
max           5.0     522.000000       3.000000     235.000000     332.0

       metric_id       year_id            val          upper          lower
count   350550.0  350550.000000  350550.000000  350550.000000  350550.000000
mean         3.0    2004.500000       0.020179       0.032878       0.011704
std          0.0       8.655454       0.049594       0.070631       0.034127
min          3.0    1990.000000       0.001051       0.001597       0.000370
25%          3.0    1997.000000       0.002300       0.004438       0.001063
50%          3.0    2004.500000       0.004869       0.009064       0.002401
75%          3.0    2012.000000       0.014729       0.027427       0.007094
max          3.0    2019.000000       0.610180       0.773804       0.501187
```

[4]:
```
mean_val = df['val'].mean()
print(f'średnia ilość rzujących to {mean_val}')

mediane_upper = df['upper'].median()
print(f'mediana górnej graicyto {mediane_upper}')

std_lower = df['lower'].std()
print(f'odchylenie standardowe dolnej granicy to {std_lower}')
```

```
średnia ilość rzujących to 0.020178816871399226
mediana górnej graicyto 0.009064496
odchylenie standardowe dolnej granicy to 0.034126627038329235
```

[5]:
```
missing_values = df.isnull().sum()
print("Brakujące wartości w kazdej kolumnie:")
print(missing_values)
```

```
Brakujące wartości w kazdej kolumnie:
measure_id       0
measure_name     0
location_id      0
location_name    0
sex_id           0
sex_name         0
age_group_id     0
age_group_name   0
rei_id           0
rei_name         0
metric_id        0
metric_name      0
year_id          0
val              0
upper            0
lower            0
dtype: int64
```

[6]:
```
df['val'] = df['val'].fillna(df['val'].mean())
```

[7]:
```
df.dropna(subset=['val'], inplace = True)
```

[8]:
```
Q1 = df['val'].quantile(0.25)
Q3 = df['val'].quantile(0.75)
IQR = Q3 - Q1

outliers = df[(df['val'] < (Q1 - 1.5 * IQR)) | (df['val'] < (Q3 + 1.5 * IQR))]
print("wartości odstające:")
print(outliers)
```

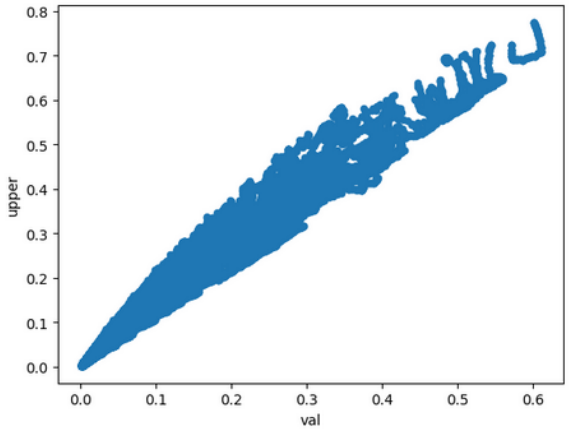
```
wartości odstające:
        measure_id measure_name  location_id location_name  sex_id sex_name  \
1                5   Prevalence            1        Global       2   Female
3                5   Prevalence            1        Global       2   Female
5                5   Prevalence            1        Global       2   Female
7                5   Prevalence            1        Global       2   Female
9                5   Prevalence            1        Global       2   Female
...            ...          ...          ...           ...     ...      ...
350545           5   Prevalence          522         Sudan       3     Both
350546           5   Prevalence          522         Sudan       3     Both
350547           5   Prevalence          522         Sudan       3     Both
350548           5   Prevalence          522         Sudan       3     Both
350549           5   Prevalence          522         Sudan       3     Both

        age_group_id    age_group_name  rei_id         rei_name  metric_id  \
1                  8          15 to 19     332  Chewing tobacco          3
3                  8          15 to 19     332  Chewing tobacco          3
5                  8          15 to 19     332  Chewing tobacco          3
7                  8          15 to 19     332  Chewing tobacco          3
9                  8          15 to 19     332  Chewing tobacco          3
...              ...               ...     ...              ...        ...
350545            27  Age standardized     332  Chewing tobacco          3
350546            27  Age standardized     332  Chewing tobacco          3
350547            27  Age standardized     332  Chewing tobacco          3
350548            27  Age standardized     332  Chewing tobacco          3
350549            27  Age standardized     332  Chewing tobacco          3

       metric_name  year_id       val     upper     lower
1             Rate     1990  0.011356  0.017594  0.007779
3             Rate     1991  0.011516  0.017807  0.007906
5             Rate     1992  0.011685  0.018425  0.007953
7             Rate     1993  0.011853  0.018675  0.008068
9             Rate     1994  0.012010  0.018950  0.008129
...            ...      ...       ...       ...       ...
350545        Rate     2015  0.029518  0.035685  0.024255
350546        Rate     2016  0.029855  0.036004  0.024500
350547        Rate     2017  0.029768  0.035934  0.024428
350548        Rate     2018  0.029760  0.035796  0.024462
350549        Rate     2019  0.029752  0.035857  0.024318

[308068 rows x 16 columns]
```

[9]:
```
correlation_matrix = df.corr(numeric_only = True)
print("macierz korelacji:")
print(correlation_matrix)

df.plot.scatter(x='val',y='upper')
```

```
macierz korelacji:
              measure_id   location_id       sex_id  age_group_id  rei_id  \
measure_id           NaN           NaN          NaN           NaN     NaN
location_id          NaN  1.000000e+00 -1.915736e-16 -9.954440e-16     NaN
sex_id               NaN -1.915736e-16  1.000000e+00  4.882346e-19     NaN
age_group_id         NaN -9.954440e-16  4.882346e-19  1.000000e+00     NaN
rei_id               NaN           NaN          NaN           NaN     NaN
metric_id            NaN           NaN          NaN           NaN     NaN
year_id              NaN  6.041010e-13  1.317749e-16 -4.971165e-16     NaN
val                  NaN  4.085158e-02 -1.796110e-02  1.440051e-02     NaN
upper                NaN  4.891981e-02 -2.994131e-02  1.937728e-02     NaN
lower                NaN  3.261103e-02 -3.448089e-03  8.555796e-03     NaN

              metric_id      year_id       val     upper     lower
measure_id          NaN          NaN       NaN       NaN       NaN
location_id         NaN  6.041010e-13  0.040652  0.048920  0.032611
sex_id              NaN  1.317749e-16 -0.017961 -0.029941 -0.003448
age_group_id        NaN -4.971165e-16  0.014401  0.019377  0.008556
rei_id              NaN          NaN       NaN       NaN       NaN
metric_id           NaN          NaN       NaN       NaN       NaN
year_id             NaN  1.000000e+00 -0.000865 -0.007611  0.008169
val                 NaN -8.648624e-04  1.000000  0.984881  0.978004
upper               NaN -7.610632e-03  0.984881  1.000000  0.928421
lower               NaN  8.168796e-03  0.978004  0.928421  1.000000
```

[9]:
```
<Axes: xlabel='val', ylabel='upper'>
```

[9]:
```
<Axes: xlabel='val', ylabel='upper'>
```



[10]:
```
df['LowerTolerance']=df['lower'] - df['val']
print(df)
```

```
        measure_id measure_name  location_id location_name  sex_id sex_name  \
0                5   Prevalence            1        Global       1     Male
1                5   Prevalence            1        Global       2   Female
2                5   Prevalence            1        Global       1     Male
3                5   Prevalence            1        Global       2   Female
4                5   Prevalence            1        Global       1     Male
...            ...          ...          ...           ...     ...      ...
350545           5   Prevalence          522         Sudan       3     Both
350546           5   Prevalence          522         Sudan       3     Both
350547           5   Prevalence          522         Sudan       3     Both
350548           5   Prevalence          522         Sudan       3     Both
350549           5   Prevalence          522         Sudan       3     Both

        age_group_id    age_group_name  rei_id         rei_name  metric_id  \
0                  8          15 to 19     332  Chewing tobacco          3
1                  8          15 to 19     332  Chewing tobacco          3
2                  8          15 to 19     332  Chewing tobacco          3
3                  8          15 to 19     332  Chewing tobacco          3
4                  8          15 to 19     332  Chewing tobacco          3
...              ...               ...     ...              ...        ...
350545            27  Age standardized     332  Chewing tobacco          3
350546            27  Age standardized     332  Chewing tobacco          3
350547            27  Age standardized     332  Chewing tobacco          3
350548            27  Age standardized     332  Chewing tobacco          3
350549            27  Age standardized     332  Chewing tobacco          3

       metric_name  year_id       val     upper     lower  LowerTolerance
0             Rate     1990  0.038740  0.055586  0.027147       -0.011593
1             Rate     1990  0.011356  0.017594  0.007779       -0.003578
2             Rate     1991  0.039253  0.055838  0.027608       -0.011645
3             Rate     1991  0.011516  0.017807  0.007906       -0.003610
4             Rate     1992  0.039863  0.056448  0.027800       -0.012063
...            ...      ...       ...       ...       ...             ...
350545        Rate     2015  0.029518  0.035685  0.024255       -0.005263
350546        Rate     2016  0.029855  0.036004  0.024500       -0.005355
350547        Rate     2017  0.029768  0.035934  0.024428       -0.005340
350548        Rate     2018  0.029760  0.035796  0.024462       -0.005298
350549        Rate     2019  0.029752  0.035857  0.024318       -0.005434

[350550 rows x 17 columns]
```

[11]:
```
grouped = df.groupby('location_name')['val'].mean()
grouped
```

[11]:
```
location_name
Afghanistan                          0.036774
Albania                              0.005881
Algeria                              0.042593
American Samoa                       0.023810
Andorra                              0.001858
                                       ...
Venezuela (Bolivarian Republic of)   0.012230
Viet Nam                             0.023155
Yemen                                0.104453
Zambia                               0.011334
Zimbabwe                             0.007761
Name: val, Length: 205, dtype: float64
```

[12]:
```
df_sorted = df.sort_values(by='val')
df_sorted.head(20)
```

[12]:

	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
340439	5	Prevalence	396	San Marino	2	Female	9	20 to 24	332	Chewing tobacco	3	Rate	2019	0.001051
340437	5	Prevalence	396	San Marino	2	Female	9	20 to 24	332	Chewing tobacco	3	Rate	2018	0.001051
340429	5	Prevalence	396	San Marino	2	Female	9	20 to 24	332	Chewing tobacco	3	Rate	2014	0.001051
340431	5	Prevalence	396	San Marino	2	Female	9	20 to 24	332	Chewing tobacco	3	Rate	2015	0.001053
340435	5	Prevalence	396	San Marino	2	Female	9	20 to 24	332	Chewing tobacco	3	Rate	2017	0.001054
340433	5	Prevalence	396	San Marino	2	Female	9	20 to 24	332	Chewing tobacco	3	Rate	2016	0.001054
340427	5	Prevalence	396	San Marino	2	Female	9	20 to 24	332	Chewing tobacco	3	Rate	2013	0.001055
340425	5	Prevalence	396	San Marino	2	Female	9	20 to 24	332	Chewing tobacco	3	Rate	2012	0.001057
340423	5	Prevalence	396	San Marino	2	Female	9	20 to 24	332	Chewing tobacco	3	Rate	2011	0.001060
340421	5	Prevalence	396	San Marino	2	Female	9	20 to 24	332	Chewing tobacco	3	Rate	2010	0.001063
340419	5	Prevalence	396	San Marino	2	Female	9	20 to 24	332	Chewing tobacco	3	Rate	2009	0.001066
340417	5	Prevalence	396	San Marino	2	Female	9	20 to 24	332	Chewing tobacco	3	Rate	2008	0.001068
340415	5	Prevalence	396	San Marino	2	Female	9	20 to 24	332	Chewing tobacco	3	Rate	2007	0.001072
340413	5	Prevalence	396	San Marino	2	Female	9	20 to 24	332	Chewing tobacco	3	Rate	2006	0.001076
340411	5	Prevalence	396	San Marino	2	Female	9	20 to 24	332	Chewing tobacco	3	Rate	2005	0.001080
340409	5	Prevalence	396	San Marino	2	Female	9	20 to 24	332	Chewing tobacco	3	Rate	2004	0.001084
111255	5	Prevalence	80	France	2	Female	9	20 to 24	332	Chewing tobacco	3	Rate	1997	0.001084
111257	5	Prevalence	80	France	2	Female	9	20 to 24	332	Chewing tobacco	3	Rate	1998	0.001084
111253	5	Prevalence	80	France	2	Female	9	20 to 24	332	Chewing tobacco	3	Rate	1996	0.001085
111259	5	Prevalence	80	France	2	Female	9	20 to 24	332	Chewing tobacco	3	Rate	1999	0.001086

3. Wnioski

Pakiet Pandas pozwala w łatwy sposób manipulować i przetwarzać dane.