

lab1

May 18, 2023

1 - ładowanie biblioteki Pandas

```
[1]:
```

```
import pandas as pd
```

2 - tworzenie ramki danych ze słownika

```
[2]: dict_city = {"City" : "Warszawa", "Łódź", "Poznań", "Wrocław"},  
          "Population" : 12678079, 5398064, 1625631, 2039421}  
df = pd.DataFrame(dict_city)  
df
```

```
[2]:      City  Population  
0  Warszawa  12678079  
1    Łódź    5398064  
2   Poznań   1625631  
3   Wrocław  2039421
```

3 - zachowanie ramki danych pobranych z pliku w formacie csv (xlsx)

```
[3]: df.to_csv("city.csv")
```

4 - tworzenie ramki danych z listy list

```
[4]: lists_city = [{"Warszawa", "Łódź", "Poznań", "Wrocław"},  
                  [12678079, 5398064, 1625631, 2039421]]  
  
pd.DataFrame(lists_city)
```

```
[4]:      0      1      2      3  
0  Warszawa Łódź Poznań Wrocław  
1  12678079 5398064 1625631 2039421
```

5 - transponowanie (wymieniamy kolumny a wierszy)

```
[5]: pd.DataFrame(lists_city).T
```

```
[5]:      0      1  
0  Warszawa 12678079  
1    Łódź 5398064  
2  Poznań 1625631  
3  Wrocław 2039421
```

6 - wyświetlić pierwsze 10 wierszy ramki danych

```
[6]: df = pd.read_csv("IHME_GBD_2019_SMOKING_TOB_1990_2019_NUM_SMOKERS_Y2021M05D27.
↳csv", encoding = "utf-8")
```

```
[7]: df.head(10)
```

```
[7]:      measure_name location_id location_name sex_id sex_name \
0 Number of Smokers          1      Global      1      Male
1 Number of Smokers          1      Global      2  Female
2 Number of Smokers          1      Global      3      Both
3 Number of Smokers          1      Global      1      Male
4 Number of Smokers          1      Global      2  Female
5 Number of Smokers          1      Global      3      Both
6 Number of Smokers          1      Global      1      Male
7 Number of Smokers          1      Global      2  Female
8 Number of Smokers          1      Global      3      Both
9 Number of Smokers          1      Global      1      Male

      age_group_id age_group_name
      year_id
0          29      15+ years      1990      8.031015e+08
1          29      15+ years      1990      8.096221e+08
2          29      15+ years      1990      1.891488e+08
3          29      15+ years      1990      1.930929e+08
4          29      15+ years      1990      9.922503e+08
5          29      15+ years      1990      1.000161e+09
6          29      15+ years      1991      8.138972e+08
7          29      15+ years      1991      8.200339e+08
8          29      15+ years      1991      1.905375e+08
9          29      15+ years      1991      1.944249e+08
10         29      15+ years      1991      1.004435e+09
11         29      15+ years      1991      1.011925e+09
12         29      15+ years      1992      8.233148e+08
13         29      15+ years      1992      8.292228e+08
14         29      15+ years      1992      1.919026e+08
15         29      15+ years      1992      1.957109e+08
16         29      15+ years      1992      1.015217e+09
17         29      15+ years      1992      1.022720e+09
18         29      15+ years      1993      8.313873e+08
19         29      15+ years      1993      8.372931e+08

      lower
0
7.959086e+08
1
1.855595e+08
2
9.847880e+08
```

```

3
8.069514e+08
4
1.869744e+08
5
9.969811e+08
6
8.167264e+08
7
1.884066e+08
8
1.007847e+09
9 8.249496e+08

```

7 - wyświetlić ostatnie 10 wierszy ramki danych

```
[8]: df.tail(10)
```

```

[8]:      measure_name location_id location_name sex_id sex_name \
20960 Number of Smokers      522      Sudan      3      Both
20961 Number of Smokers      522      Sudan      1      Male
20962 Number of Smokers      522      Sudan      2  Female
20963 Number of Smokers      522      Sudan      3      Both
20964 Number of Smokers      522      Sudan      1      Male
20965 Number of Smokers      522      Sudan      2  Female
20966 Number of Smokers      522      Sudan      3      Both
20967 Number of Smokers      522      Sudan      1      Male
20968 Number of Smokers      522      Sudan      2  Female
20969 Number of Smokers      522      Sudan      3      Both

      age_group_id age_group_name
year_id
20960      29      15+ years      2016      2.454893e+06
      2.665441e+06
20961      29      15+ years      2017      2.297622e+06
      2.490884e+06
20962      29      15+ years      2017      2.373815e+05
      3.217514e+05
20963      29      15+ years      2017      2.535003e+06
      2.743769e+06
20964      29      15+ years      2018      2.367072e+06
      2.575100e+06
20965      29      15+ years      2018      2.435999e+05
      3.286166e+05
20966      29      15+ years      2018      2.610672e+06
      2.833943e+06
20967      29      15+ years      2019      2.439150e+06
      2.656579e+06

```

20968	29	15+ years	2019	2.500800e+05
			3.345384e+05	
20969	29	15+ years	2019	2.689230e+06
			2.918332e+06	
		lower		
20960				2.267696e+06
20961				2.114574e+06
20962				1.729171e+05
20963				2.341329e+06
20964				2.173995e+06
20965				1.752508e+05
20966				2.409108e+06
20967				2.236450e+06
20968				1.816686e+05
20969				2.480656e+06

8 - wyświetlić informację o ramce danych

```
[9]: df.info()
```

```
<class
'pandas.core.frame.DataFrame'>
RangeIndex: 20970 entries, 0 to
20969 Data columns (total 11
columns):
#   Column                Non-Null Count  Dtype
---  -
measure_name 20970 non-null object
1  location_id    20970 non-null
      int64
2  location_name 20970 non-null
      object
3  sex_id        20970 non-null
      int64
4  sex_name      20970 non-null
      object
5  age_group_id  20970 non-null
      int64
6  age_group_name 20970 non-null object
7  year_id       20970 non-null int64
8  val           20970 non-null float64
9  upper 20970 non-null float64 10 lower 20970 non-null float64
dtypes: float64(3), int64(4), object(4)
memory usage: 1.8+ MB
```

9 - wyświetlić, ile wierszy i kolumn znajduje się w ramce danych

```
[10]: df.shape
```

```
[10]: (20970, 11)
```

10 - wyświetlić informację statystyczną o kolumnach liczbowych (wartości niepowtarzalne, średnia, odchylenie standardowe, minimum, kwartyle, maksimum)

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

```
[11]:      location_id      sex_id age_group_id      year_id      val \
count  20970.000000  20970.000000    20970.0    20970.000000
      2.097000e+04
mean    131.111588    2.000000    29.0  2004.500000  1.242807e+07
std     95.055111    0.816516     0.0    8.655648  6.489191e+07
min      1.000000    1.000000    29.0  1990.000000  6.345717e+01
25%     61.000000    1.000000    29.0  1997.000000  8.201065e+04
50%    119.000000    2.000000    29.0  2004.500000  5.777123e+05
75%    177.000000    3.000000    29.0  2012.000000  2.901197e+06
max     522.000000    3.000000    29.0  2019.000000  1.144819e+09

      upper      lower
count      2.097000e+04
2.097000e+04      mean
1.269088e+07  1.217241e+07 std
6.555971e+07  6.421446e+07 min
7.868296e+01  5.029157e+01 25%
9.576943e+04  6.875439e+04
50%  6.278332e+05  5.329521e+05
75%  3.070281e+06  2.742651e+06
max  1.157286e+09  1.131582e+09
```

11 - wyświetlić informację statystyczną o kolumnach kategoryzowanych (ile unikalnych wartości, top - jaka jest najpopularniejsza wartość, freq - jak często najpopularniejsza)

```
[12]: df.describe(include = 'all')
```

```
[12]:      measure_name location_id location_name      sex_id sex_name \
count      20970  20970.000000      20970  20970.000000    20970
unique          1          NaN        231          NaN         3
top    Number of Smokers          NaN    South Asia          NaN    Male
freq      20970          NaN        180          NaN       6990
mean          NaN    131.111588          NaN    2.000000          NaN
std          NaN    95.055111          NaN    0.816516          NaN
min          NaN     1.000000          NaN    1.000000          NaN
25%          NaN     61.000000          NaN    1.000000          NaN
50%          NaN    119.000000          NaN    2.000000          NaN
75%          NaN    177.000000          NaN    3.000000          NaN
max          NaN    522.000000          NaN    3.000000          NaN

      age_group_id age_group_name      year_id      val      upper \
count      20970.0      20970  20970.000000  2.097000e+04  2.097000e+04
unique          NaN          1          NaN          NaN          NaN
```

top	NaN	15+ years	NaN	NaN	NaN
freq	NaN	20970	NaN	NaN	NaN
mean	29.0	NaN	2004.500000	1.242807e+07	1.269088e+07
std	0.0	NaN	8.655648	6.489191e+07	6.555971e+07
min	29.0	NaN	1990.000000	6.345717e+01	7.868296e+01
25%	29.0	NaN	1997.000000	8.201065e+04	9.576943e+04
50%	29.0	NaN	2004.500000	5.777123e+05	6.278332e+05
75%	29.0	NaN	2012.000000	2.901197e+06	3.070281e+06
max	29.0	NaN	2019.000000	1.144819e+09	1.157286e+09

count 2.097000e+04

unique NaN

top NaN

freq NaN

mean 1.217241e+07

std 6.421446e+07

min 5.029157e+01

25% 6.875439e+04

50% 5.329521e+05

75% 2.742651e+06

max 1.131582e+09

12 - usunąć brakujące wartości w ramce danych

```
[13]: df.dropna(inplace=True)
df
```

```
[13]:      measure_name location_id location_name sex_id sex_name \
0      Number of Smokers      1      Global      1      Male
1      Number of Smokers      1      Global      2  Female
2      Number of Smokers      1      Global      3      Both
3      Number of Smokers      1      Global      1      Male
4      Number of Smokers      1      Global      2  Female
...      ...      ...      ...      ...      ...
20965  Number of Smokers     522      Sudan      2  Female
20966  Number of Smokers     522      Sudan      3      Both
20967  Number of Smokers     522      Sudan      1      Male
```

20968	Number of Smokers	522	Sudan	2	Female
20969	Number of Smokers	522	Sudan	3	Both

	age_group_id	age_group_name	year_id	val	upper \
0	29	15+ years	1990	8.031015e+08	
				8.096221e+08	
1	29	15+ years	1990	1.891488e+08	
				1.930929e+08	
2	29	15+ years	1990	9.922503e+08	
				1.000161e+09	
3	29	15+ years	1991	8.138972e+08	
				8.200339e+08	
4	29	15+ years	1991	1.905375e+08	
				1.944249e+08	
...	
20965	29	15+ years	2018	2.435999e+05	
				3.286166e+05	
20966	29	15+ years	2018	2.610672e+06	
				2.833943e+06	
20967	29	15+ years	2019	2.439150e+06	
				2.656579e+06	
20968	29	15+ years	2019	2.500800e+05	
				3.345384e+05	
20969	29	15+ years	2019	2.689230e+06	
				2.918332e+06	

	lower
0	7.959086e+08
1	1.855595e+08
2	9.847880e+08
3	8.069514e+08
4	1.869744e+08
...	...
20965	1.752508e+05
20966	2.409108e+06
20967	2.236450e+06
20968	1.816686e+05
20969	2.480656e+06

[20970 rows x 11 columns]

13 - przedstawić wybór wierszy i kolumny używając nazw oraz indeksów na różne sposoby

```
[14]: df["location_name"]
```

```
[14]: 0      Global
```

```

1      Global
2      Global
3      Global
4      Global
...
20965   Sudan
20966   Sudan
20967   Sudan
20968   Sudan
20969   Sudan
Name: location_name, Length: 20970, dtype: object

```

```
[15]: df.location_name
```

```

[15]: 0      Global 1
Global
2      Global
3      Global
4      Global
...
20965   Sudan
20966   Sudan
20967   Sudan
20968   Sudan
20969   Sudan
Name: location_name, Length: 20970, dtype: object

```

```
[16]: df[["location_name", "sex_name", "year_id"]]
```

```

[16]:   location_name sex_name year_id
0      Global    Male    1990
1      Global  Female    1990
2      Global    Both    1990
3      Global    Male    1991
4      Global  Female    1991
...
20965   Sudan  Female    2018
20966   Sudan    Both    2018
20967   Sudan    Male    2019
20968   Sudan  Female    2019
20969   Sudan    Both    2019
[20970 rows x 3 columns]

```

```
[17]: df.loc[100:110, "location_name":"year_id"]
```

```

[17]:   location_name sex_id sex_name age_group_id \
100 Southeast Asia, East Asia, and Oceania 2 Female      29
101 Southeast Asia, East Asia, and Oceania 3    Both      29

```


102	Southeast Asia, East Asia, and Oceania	1	Male	29
103	Southeast Asia, East Asia, and Oceania	2	Female	29
104	Southeast Asia, East Asia, and Oceania	3	Both	29
105	Southeast Asia, East Asia, and Oceania	1	Male	29
106	Southeast Asia, East Asia, and Oceania	2	Female	29
107	Southeast Asia, East Asia, and Oceania	3	Both	29
108	Southeast Asia, East Asia, and Oceania	1	Male	29
109	Southeast Asia, East Asia, and Oceania	2	Female	29
110	Southeast Asia, East Asia, and Oceania	3	Both	29

	age_group_name	year_id
100	15+ years	1993
101	15+ years	1993
102	15+ years	1994
103	15+ years	1994
104	15+ years	1994
105	15+ years	1995
106	15+ years	1995
107	15+ years	1995
108	15+ years	1996
109	15+ years	1996
110	15+ years	1996

```
[18]: df.iloc[105:115, 0:3]
```

```
[18]:
```

	measure_name	location_id	location_name
105	Number of Smokers	4	Southeast Asia, East Asia, and Oceania
106	Number of Smokers	4	Southeast Asia, East Asia, and Oceania
107	Number of Smokers	4	Southeast Asia, East Asia, and Oceania
108	Number of Smokers	4	Southeast Asia, East Asia, and Oceania
109	Number of Smokers	4	Southeast Asia, East Asia, and Oceania
110	Number of Smokers	4	Southeast Asia, East Asia, and Oceania
111	Number of Smokers	4	Southeast Asia, East Asia, and Oceania
112	Number of Smokers	4	Southeast Asia, East Asia, and Oceania
113	Number of Smokers	4	Southeast Asia, East Asia, and Oceania
114	Number of Smokers	4	Southeast Asia, East Asia, and Oceania

14 - przedstawić wybór wierszy z ramki danych pod warunkiem odnośnie określonej wartości kolumny

```
[19]: df[df["sex_name"] == "Both"]
```

```
[19]:      measure_name location_id location_name sex_id sex_name \
2      Number of Smokers      1      Global      3      Both
5      Number of Smokers      1      Global      3      Both
8      Number of Smokers      1      Global      3      Both
11     Number of Smokers      1      Global      3      Both
14     Number of Smokers      1      Global      3      Both
...
20957 Number of Smokers      522      Sudan      3      Both
20960 Number of Smokers      522      Sudan      3      Both
20963 Number of Smokers      522      Sudan      3      Both
20966 Number of Smokers      522      Sudan      3      Both
20969 Number of Smokers      522      Sudan      3      Both

      age_group_id age_group_name
year_id
2      29      15+ years  1990      9.922503e+08
      1.000161e+09
5      29      15+ years  1991      1.004435e+09
      1.011925e+09
8      29      15+ years  1992      1.015217e+09
      1.022720e+09
11     29      15+ years  1993      1.024669e+09
      1.031965e+09
14     29      15+ years  1994      1.032567e+09
      1.039842e+09
...
20957     29      15+ years  2015      2.388216e+06
      2.587005e+06
20960     29      15+ years  2016      2.454893e+06
      2.665441e+06
20963     29      15+ years  2017      2.535003e+06
      2.743769e+06
20966     29      15+ years  2018      2.610672e+06
      2.833943e+06
20969     29      15+ years  2019      2.689230e+06
      2.918332e+06

      lower
2
9.847880e+08 5
```

```

9.969811e+08 8
1.007847e+09
11 1.017551e+09
14 1.025631e+09
...
20957 2.211144e+06
20960 2.267696e+06
20963 2.341329e+06
20966 2.409108e+06
20969 2.480656e+06

```

```
[6990 rows x 11 columns]
```

15 - przedstawić wybór wierszy z ramki danych pod warunkiem spełnienia kilku warunków jednocześnie

```
[20]: df[(df["sex_name"] == "Both") & (df["year_id"] == 2016) &
(df["location_name"]_ 4== "Sudan")]
```

```
[20]:
measure_name location_id location_name sex_id sex_name \
20960 Number of Smokers      522      Sudan      3      Both

age_group_id age_group_name year_id val upper \ 20960 29
15+ years 2016 2454892.625 2665440.938

lower
20960 2267696.034
```

16 - wybrać wiersze które zawierają w kolumnie kategoryzowanej określone słowo

```
[21]: df[df["location_name"].str.contains("States")]
```

```
[21]:
measure_name location_id location_name \
1980 Number of Smokers      25 Micronesia (Federated States
of)
1981 Number of Smokers      25 Micronesia (Federated States
of)
1982 Number of Smokers      25 Micronesia (Federated States
of)
1983 Number of Smokers      25 Micronesia (Federated States
of)
1984 Number of Smokers      25 Micronesia (Federated States
of)
...
20785 Number of Smokers      422 United States Virgin Islands
20786 Number of Smokers      422 United States Virgin Islands
20787 Number of Smokers      422 United States Virgin Islands
```

```

20788 Number of Smokers      422   United States Virgin Islands
20789 Number of Smokers      422   United States Virgin Islands

      sex_id sex_name age_group_id age_group_name year_id      val \
1980      1      Male 29      15+ years 1990 18134.775290
1981      2      Female 29      15+ years 1990 9470.305481
1982      3      Both 29      15+ years 1990 27605.080770
1983      1      Male 29      15+ years 1991 18395.672830
1984      2      Female 29      15+ years 1991 9658.519070
...      ...      ...      ...      ...      ...
20785 2      Female 29      15+ years 2018 2308.376511 20786
      3      Both 29      15+ years 2018 5633.535832
20787      1      Male 29      15+ years 2019 3280.527338
20788      2      Female 29      15+ years 2019 2282.281664
20789      3      Both 29      15+ years 2019 5562.809002

```

```

      upper      lower
1980 19169.248820 17155.196930
1981 11156.303110 7825.944174
1982 29580.226920 25829.741340
1983 19459.617700 17385.018410
1984 11404.994170 7961.453848
...      ...      ...
20785 2820.434508 1871.029388
20786 6212.418101 5090.184376
20787 3649.862482 2939.996840
20788 2813.914814 1831.778372
20789 6146.429254 4990.914042

```

[270 rows x 11 columns]

17 - wybrać wiersze które nie zawierają w kolumnie kategoryzowanej określone słowo

```
[22]: df[~df["location_name"].str.contains("States")]
```

```

[22]:      measure_name location_id location_name sex_id sex_name \
0      Number of Smokers      1      Global      1      Male
1      Number of Smokers      1      Global      2 Female
2      Number of Smokers      1      Global      3      Both
3      Number of Smokers      1      Global      1      Male
4      Number of Smokers      1      Global      2 Female
...      ...      ...      ...      ...
20965 Number of Smokers      522      Sudan      2 Female

```

20966	Number of Smokers	522	Sudan	3	Both
20967	Number of Smokers	522	Sudan	1	Male
20968	Number of Smokers	522	Sudan	2	Female
20969	Number of Smokers	522	Sudan	3	Both

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

[20700 rows x 11 columns]

18 - utwórz kolumnę na podstawie istniejącyn

```
[23]: df["new_location_name"] = df["location_name"]  
df
```

	measure_name	location_id	location_name	sex_id	sex_name \
0	Number of Smokers	1	Global	1	Male
1	Number of Smokers	1	Global	2	Female
2	Number of Smokers	1	Global	3	Both
3	Number of Smokers	1	Global	1	Male
4	Number of Smokers	1	Global	2	Female
...
20965	Number of Smokers	522	Sudan	2	Female

[23]:

20966	Number of Smokers	522	Sudan	3	Both
20967	Number of Smokers	522	Sudan	1	Male
20968	Number of Smokers	522	Sudan	2	Female
20969	Number of Smokers	522	Sudan	3	Both

	age_group_id	age_group_name	year_id	val	upper \
0	29	15+ years	1990	8.031015e+08	
				8.096221e+08	
1	29	15+ years	1990	1.891488e+08	
				1.930929e+08	
2	29	15+ years	1990	9.922503e+08	
				1.000161e+09	
3	29	15+ years	1991	8.138972e+08	
				8.200339e+08	
4	29	15+ years	1991	1.905375e+08	
				1.944249e+08	
...
20965	29	15+ years	2018	2.435999e+05	
				3.286166e+05	
20966	29	15+ years	2018	2.610672e+06	
				2.833943e+06	
20967	29	15+ years	2019	2.439150e+06	
				2.656579e+06	
20968	29	15+ years	2019	2.500800e+05	
				3.345384e+05	

```

      measure_name location_id location_name sex_id
      sex_name \
0    Number of Smokers      1      Global      1    Male
1    Number of Smokers      1      Global      2 Female
2    Number of Smokers      1      Global      3    Both
3    Number of Smokers      1      Global      1    Male
4    Number of Smokers      1      Global      2 Female
...
20965 Number of Smokers      522      Sudan      2 Female
20969      29      15+ years      2019      2.689230e+06
      2.918332e+06

      lower new_location_name
0    7.959086e+08      Global 1
      1.855595e+08      Global 2
      9.847880e+08      Global 3
      8.069514e+08      Global
4    1.869744e+08      Global
...
20965 1.752508e+05      Sudan
20966 2.409108e+06      Sudan
20967 2.236450e+06      Sudan
20968 1.816686e+05      Sudan
20969 2.480656e+06      Sudan
[20970 rows x 12 columns]

```

19 - usuń kolumnę

```
[24]: df.drop("new_location_name", axis=1, inplace = True)
df
```

[24]:

```

20966 Number of Smokers      522      Sudan      3    Both
20967 Number of Smokers      522      Sudan      1    Male
20968 Number of Smokers      522      Sudan      2 Female
20969 Number of Smokers      522      Sudan      3    Both

      age_group_id age_group_name
      year_id
0      29      15+ years      1990      8.031015e+08
      8.096221e+08

```


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

[20970 rows x 11 columns]

	measure_name	location_id	location_name	sex_id	sex_name \
0	Number of Smokers	1	Global	1	Male
1	Number of Smokers	1	Global	2	Female
2	Number of Smokers	1	Global	3	Both
3	Number of Smokers	1	Global	1	Male
4	Number of Smokers	1	Global	2	Female
...
20965	Number of Smokers	522	Sudan	2	Female

20 - zmień nazwę kolumny

```
[25]: df.rename(columns = {"year_id": "year"}, inplace = True)
df
```

[25]:

20966	Number of Smokers	522	Sudan	3	Both
20967	Number of Smokers	522	Sudan	1	Male
20968	Number of Smokers	522	Sudan	2	Female
20969	Number of Smokers	522	Sudan	3	Both

	age_group_id	age_group_name	year	val	upper \
0	29	15+ years	1990	8.031015e+08	
				8.096221e+08	
1	29	15+ years	1990	1.891488e+08	
				1.930929e+08	
2	29	15+ years	1990	9.922503e+08	
				1.000161e+09	
3	29	15+ years	1991	8.138972e+08	
				8.200339e+08	
4	29	15+ years	1991	1.905375e+08	
				1.944249e+08	
...
20965	29	15+ years	2018	2.435999e+05	
				3.286166e+05	
20966	29	15+ years	2018	2.610672e+06	
				2.833943e+06	
20967	29	15+ years	2019	2.439150e+06	
				2.656579e+06	
20968	29	15+ years	2019	2.500800e+05	
				3.345384e+05	

```
[27]: df["val"].mean() #średnia
```

```

      measure_name location_id location_name sex_id
sex_name \
0    Number of Smokers      1      Global      1    Male
1    Number of Smokers      1      Global      2 Female
2    Number of Smokers      1      Global      3    Both
3    Number of Smokers      1      Global      1    Male
4    Number of Smokers      1      Global      2 Female
...
20965 Number of Smokers      522      Sudan      2 Female
20969      29      15+      years      2019      2.689230e+06
      2.918332e+06

      lower
0
7.959086e+08 1
1.855595e+08 2
9.847880e+08 3
8.069514e+08
4      1.869744e+08
...
20965 1.752508e+05
20966 2.409108e+06
20967 2.236450e+06
20968 1.816686e+05
20969 2.480656e+06

```

[20970 rows x 11 columns]

21 - zachowaj ramkę danych jako plik csv na komputerze

```
[26]: df.to_csv("Lab1_eiwd_Michał_Stajerski.csv")
```

22 - wyświetlić średnia (maksymalną, minimalną) wartość z jednej kolumny

```
[27]: 12428071.383604305
```

```
[28]: df['val'].max() #maksymalna
```

```
[28]: 1144818597.0
```

```
[29]: df['val'].min() #minimalna
```

```
[29]: 63.45716608
```

23 - wyświetlić liczbę wierszy

```
[30]: df['measure_name'].count()
```

```
[30]: 20970
```

24 - wyświetlić wartości unikatowe w kolumnie

```
[31]: df['sex_name'].unique()
```

```
[31]: array(['Male', 'Female', 'Both'], dtype=object)
```

25 - wyświetlić liczby rekordów odpowiadających do wartości

```
[32]: df['sex_name'].value_counts()
```

```
[32]: Male    6990
      Female  6990
      Both   6990
      Name: sex_name, dtype: int64
```

26 - sortowanie wierszy ramki danych według wartości określonej kolumny (malejąco, rosnąco)

```
[33]: df.sort_values(['sex_id'], ascending = False)
```

```
[33]:
```

	measure_name	location_id	location_name	sex_id \
20969	Number of Smokers	522	Sudan	3
8456	Number of Smokers	96	Southern Latin America	3
18149	Number of Smokers	205	Côte d'Ivoire	3
8462	Number of Smokers	97	Argentina	3
8465	Number of Smokers	97	Argentina	3
...
10488	Number of Smokers	119	Trinidad and Tobago	1
10491	Number of Smokers	119	Trinidad and Tobago	1
10494	Number of Smokers	119	Trinidad and Tobago	1

10497	Number of Smokers	119	Trinidad and Tobago	1
10485	Number of Smokers	119	Trinidad and Tobago	1

	sex_name	age_group_id	age_group_name	year	val	upper \
20969	Both	29	15+ years	2019	2.689230e+06	2.918332e+06
8456	Both	29	15+ years	2018	1.375418e+07	1.433091e+07
18149	Both	29	15+ years	2009	1.851309e+06	1.958859e+06
8462	Both	29	15+ years	1990	6.940515e+06	7.626183e+06
8465	Both	29	15+ years	1991	6.966965e+06	7.650883e+06
...
10488	Male	29	15+ years	2006	1.543484e+05	1.663233e+05
10491	Male	29	15+ years	2007	1.567341e+05	1.686857e+05
10494	Male	29	15+ years	2008	1.588890e+05	1.709821e+05
10497	Male	29	15+ years	2009	1.603883e+05	1.724855e+05
10485	Male	29	15+ years	2005	1.516994e+05	1.639840e+05

	lower
20969	2.480656e+06
8456	1.317504e+07
18149	1.740542e+06
8462	6.336184e+06
8465	6.364471e+06
...	...
10488	1.431156e+05
10491	1.452546e+05
10494	1.474781e+05
10497	1.481193e+05
10485	1.401675e+05

[20970 rows x 11 columns]

```
[34]: df.sort_values(['sex_id'], ascending = True)
```

```
[34]:
```

	measure_name	location_id	location_name	sex_id	sex_name \
0	Number of Smokers	1	Global	1	Male
18147	Number of Smokers	205	Côte d'Ivoire	1	Male
8463	Number of Smokers	97	Argentina	1	Male
8466	Number of Smokers	97	Argentina	1	Male
8469	Number of Smokers	97	Argentina	1	Male

...
10490	Number of Smokers	119	Trinidad and Tobago	3	Both
10493	Number of Smokers	119	Trinidad and Tobago	3	Both
10496	Number of Smokers	119	Trinidad and Tobago	3	Both
10439	Number of Smokers	118	Suriname	3	Both
20969	Number of Smokers	522	Sudan	3	Both

	age_group_id	age_group_name	year	val	upper \
0	29	15+ years	1990	8.031015e+08	8.096221e+08
18147	29	15+ years	2009	1.610315e+06	1.701718e+06
8463	29	15+ years	1991	3.962138e+06	4.302021e+06
8466	29	15+ years	1992	3.971895e+06	4.312380e+06
8469	29	15+ years	1993	3.985485e+06	4.306737e+06
...
10490	29	15+ years	2006	1.964041e+05	2.110698e+05
10493	29	15+ years	2007	1.993844e+05	2.138476e+05
10496	29	15+ years	2008	2.020567e+05	2.162465e+05
10439	29	15+ years	2019	9.249139e+04	9.954819e+04
20969	29	15+ years	2019	2.689230e+06	2.918332e+06
		lower			
0		7.959086e+08			
18147		1.518489e+06			
8463		3.640765e+06			
8466		3.661012e+06			
8469		3.673090e+06			
...		...			
10490		1.829523e+05			
10493		1.858097e+05			
10496		1.881899e+05			
10439		8.606268e+04			
20969		2.480656e+06			

[20970 rows x 11 columns]

27 - wyświetlić wierszy dla 10 największych (najmniejszych) wartości określonej kolumny

```
[35]: df.nlargest(10, 'location_id')
```

```
[35]:      measure_name location_id location_name sex_id sex_name \
20880 Number of Smokers      522      Sudan      1      Male
20881 Number of Smokers      522      Sudan      2  Female
20882 Number of Smokers      522      Sudan      3      Both
20883 Number of Smokers      522      Sudan      1      Male
20884 Number of Smokers      522      Sudan      2  Female
20885 Number of Smokers      522      Sudan      3      Both
20886 Number of Smokers      522      Sudan      1      Male
20887 Number of Smokers      522      Sudan      2  Female
20888 Number of Smokers      522      Sudan      3      Both
20889 Number of Smokers      522      Sudan      1      Male
```

```
      age_group_id age_group_name year      val      upper \
20880          29      15+      years  1990  1.210513e+06
      1.343292e+06
20881          29      15+      years  1990  1.295362e+05
      1.719868e+05
20882          29      15+      years  1990  1.340050e+06
      1.481698e+06
20883          29      15+      years  1991  1.260431e+06
      1.394211e+06
20884          29      15+      years  1991  1.341847e+05
      1.777673e+05
20885          29      15+      years  1991  1.394615e+06
      1.538089e+06
20886          29      15+      years  1992  1.309607e+06
      1.446107e+06
20887          29      15+      years  1992  1.388423e+05
      1.850937e+05
20888          29      15+      years  1992  1.448449e+06
      1.588898e+06
20889          29      15+      years  1993  1.357387e+06
      1.498584e+06

      lower
20880  1.085168e+06
20881  9.532772e+04
20882  1.204444e+06
20883  1.132721e+06
20884  9.848629e+04
20885  1.254003e+06
20886  1.180870e+06
20887  1.019466e+05
20888  1.304217e+06
20889  1.225640e+06
```

```
[36]: df.nsmallest(10, 'location_id')
```

```
[36]:   measure_name location_id location_name sex_id sex_name \
0 Number of Smokers          1      Global      1    Male
1 Number of Smokers          1      Global      2  Female
2 Number of Smokers          1      Global      3    Both
3 Number of Smokers          1      Global      1    Male
4 Number of Smokers          1      Global      2  Female
5 Number of Smokers          1      Global      3    Both
6 Number of Smokers          1      Global      1    Male
7 Number of Smokers          1      Global      2  Female
8 Number of Smokers          1      Global      3    Both
9 Number of Smokers          1      Global      1    Male

   age_group_id age_group_name year      val      upper      lower
0           29    15+ years 1990 8.031015e+08 8.096221e+08 7.959086e+08
1           29    15+ years 1990 1.891488e+08 1.930929e+08 1.855595e+08
2           29    15+ years 1990 9.922503e+08 1.000161e+09 9.847880e+08
3           29    15+ years 1991 8.138972e+08 8.200339e+08 8.069514e+08
4           29    15+ years 1991 1.905375e+08 1.944249e+08 1.869744e+08
5           29    15+ years 1991 1.004435e+09 1.011925e+09 9.969811e+08
6           29    15+ years 1992 8.233148e+08 8.292228e+08 8.167264e+08
7           29    15+ years 1992 1.919026e+08 1.957109e+08 1.884066e+08
8           29    15+ years 1992 1.015217e+09 1.022720e+09 1.007847e+09
9           29    15+ years 1993 8.313873e+08 8.372931e+08 8.249496e+08
```

28 - wyświetlić wierszy dla 10 największych wartości określonej kolumny pod warunkiem określonych wartości innej kolumny

```
[37]: df[df['year'] == 2015].nlargest(10, 'location_id')
```

```
[37]:   measure_name location_id location_name sex_id \
20955 Number of Smokers    522      Sudan      1
20956 Number of Smokers    522      Sudan      2
20957 Number of Smokers    522      Sudan      3
20865 Number of Smokers    435  South Sudan      1
20866 Number of Smokers    435  South Sudan      2
20867 Number of Smokers    435  South Sudan      3
```



```

20775 Number of Smokers      422 United States Virgin Islands1
20776 Number of Smokers      422 United States Virgin Islands2
20777 Number of Smokers      422 United States Virgin Islands3

```

```

20685 Number of Smokers      416 Tuvalu      1
      sex_name age_group_id age_group_name year      val      upper \
20955   Male      29      15+      years      2015      2.159385e+06
      2.329364e+06
20956  Female      29      15+      years      2015      2.288306e+05
      3.056884e+05
20957   Both      29      15+      years      2015      2.388216e+06
      2.587005e+06
20865   Male      29      15+      years      2015      4.716963e+05
      5.254786e+05
20866  Female      29      15+      years      2015      5.970915e+04
      7.713253e+04
20867   Both      29      15+      years      2015      5.314055e+05
      5.866896e+05
20775   Male      29      15+      years      2015      3.466521e+03
      3.821509e+03
20776  Female      29      15+      years      2015      2.390917e+03
      2.845169e+03
20777   Both      29      15+      years      2015      5.857438e+03
      6.406057e+03
20685   Male      29      15+      years      2015      1.854994e+03
      1.955782e+03
      lower
20955 1.990166e+06
20956 1.694027e+05
20957 2.211144e+06
20865 4.222599e+05
20866 4.480880e+04
20867 4.787462e+05
20775 3.149973e+03
20776 1.981502e+03
20777 5.368333e+03
20685 1.751382e+03

```

29 - grupowanie wierszy według wartości kolumny kategoryzowanej, potem - uśrednienie wartości wszystkich kolumn w grupie - MultiIndex

```

[38]: df.groupby('sex_name').agg({'age_group_id': ['count'], 'val':
['mean']})

```

```

[38]: age_group_id val count      mean

```

```
sex_name
Both      6990 1.864211e+07
Female    6990 3.441201e+06
Male      6990 1.520091e+07
```

30 - grupowanie wierszy według wartości kolumny kategoryzowanej, potem - uśrednienie wartości dla pewnych kolumn, liczba wartości i mediana dla pozostałych kolumn w grupach

```
[39]: df.groupby('sex_name').agg({'age_group_id': ['count'], 'val':
    ['mean', '_'] + ['median']})
```

```
[39]:      age_group_id      val
      count      mean      median
sex_name
Both      6990 1.864211e+07 968560.4033
Female    6990 3.441201e+06 177406.7973
Male      6990 1.520091e+07 721673.5286
```

31 - wyświetlić nazwy kolumn indeksu złożonego

```
[40]: df.index
```

```
[40]: RangeIndex(start=0, stop=20970, step=1)
```

```
[41]: df_sexname = df.groupby('sex_name').agg({'age_group_id':
['count'], 'val':['_']
    + ['mean', 'median']})
df_sexname.columns
```

```
[41]: MultiIndex([('age_group_id', 'count'),
                ('val', 'mean'),
                ('val', 'median')],
                )
```

32 - sortować kolumnę indeksu złożonego

```
[42]: df_sexname['val']['mean'].sort_values(ascending = False)
```

```
[42]: sex_name
Both      1.864211e+07
Male      1.520091e+07
Female    3.441201e+06
Name: mean, dtype: float64
```

33 - stworzyć tabelę przystawną (pivot table) na podstawie ramki danych

```
[43]: df_pivot = df.pivot_table(values='sex_id', index='location_name',  
                                columns='sex_name',  
                                margins=False, dropna=True, fill_value=None)  
df_pivot
```

```
[43]: sex_name  
location_name  
Afghanistan      3      2      1  
Albania           3      2      1  
Algeria           3      2      1  
American Samoa   3      2      1  
Andean Latin America 3      2      1  
...              ...      ...      ...  
Western Europe   3      2      1  
Western Sub-Saharan Africa 3      2      1  
Yemen            3      2      1  
Zambia           3      2      1  
Zimbabwe         3      2      1  
[231 rows x 3 columns]
```

34 - wyświetlić indeksy i kolumny tabeli przystawnej

```
[44]: df_pivot.index
```

```
[44]: Index(['Afghanistan', 'Albania', 'Algeria', 'American Samoa',  
        'Andean Latin America', 'Andorra', 'Angola', 'Antigua and  
        Barbuda',  
        'Argentina', 'Armenia',  
        ...  
        'Uruguay', 'Uzbekistan', 'Vanuatu',  
        'Venezuela (Bolivarian Republic of)', 'Viet Nam', 'Western  
        Europe',  
        'Western Sub-Saharan Africa', 'Yemen', 'Zambia', 'Zimbabwe'],  
        dtype='object', name='location_name', length=231)
```

```
[45]: df_pivot.columns
```

```
[45]: Index(['Both', 'Female', 'Male'], dtype='object', name='sex_name')
```

35 - utwórz indeks złożony tabeli przystawnej i wyświetl go

```
[46]: df_pivot = df.pivot_table(values='sex_id',
    index=['location_name', 'location_id'],
    columns='sex_name', margins=False, dropna=True,
    fill_value=None)
df_pivot
```

```
[46]: sex_name  Both Female Male location_name  location_id
Afghanistan                160                3         2         1
Albania                    43                 3         2         1
Algeria                   139                3         2         1
American Samoa            298                3         2         1
Andean Latin America      120                3         2         1
...
Western Europe              73                3         2         1
Western Sub-Saharan Africa 199                3         2         1
Yemen                     157                3         2         1
Zambia                    191                3         2         1
Zimbabwe                  198                3         2         1
[233 rows x 3 columns]
```

```
[47]: df_pivot.index
```

```
[47]: MultiIndex([(
    'Afghanistan', 160),
    (
    'Albania', 43),
    (
    'Algeria', 139),
    (
    'American Samoa', 298),
    (
    'Andean Latin America',
    120),
    (
    'Andorra', 74),
    (
    'Angola', 168),
    (
    'Antigua and Barbuda',
    105),
    (
    'Argentina', 97),
    (
    'Armenia', 33),
    ...
    (
    'Uruguay', 99),
    (
    'Uzbekistan', 41),
    (
    'Vanuatu', 30),
    ('Venezuela (Bolivarian Republic of)', 133),
    (
    'Viet Nam', 20),
```

```
(
    ('Western Europe', 73),
    ('Western Sub-Saharan Africa', 199),
    ('Yemen', 157), ('Zambia', 191),
    ('Zimbabwe', 198)],
names=['location_name', 'location_id'], length=233)
```

36 - zaimportuj moduł pyplot z biblioteki matplotlib

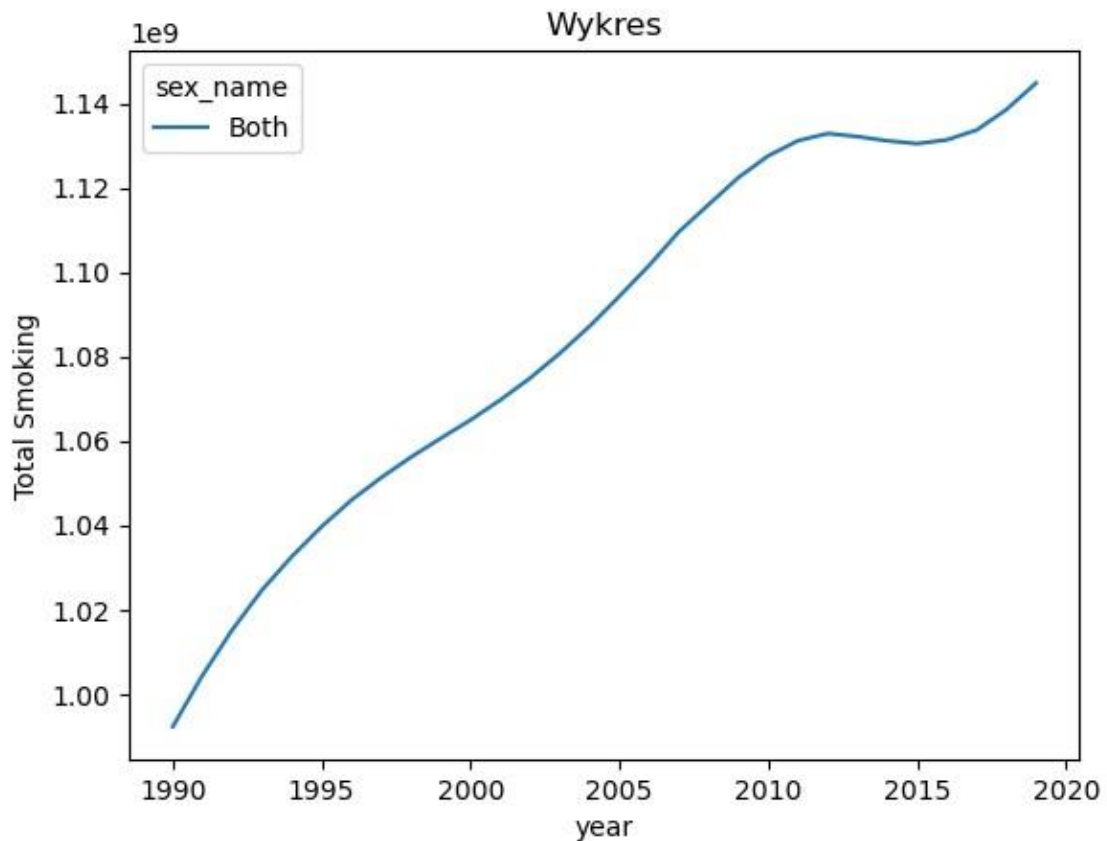
```
[48]: import matplotlib.pyplot as plt
```

37 - wskazać, że wykresy należy rysować bezpośrednio w zeszycie, a nie w osobnej zakładce

```
[49]: %matplotlib inline
```

38 - wyświetlić wykres na podstawie tabeli przystawnej

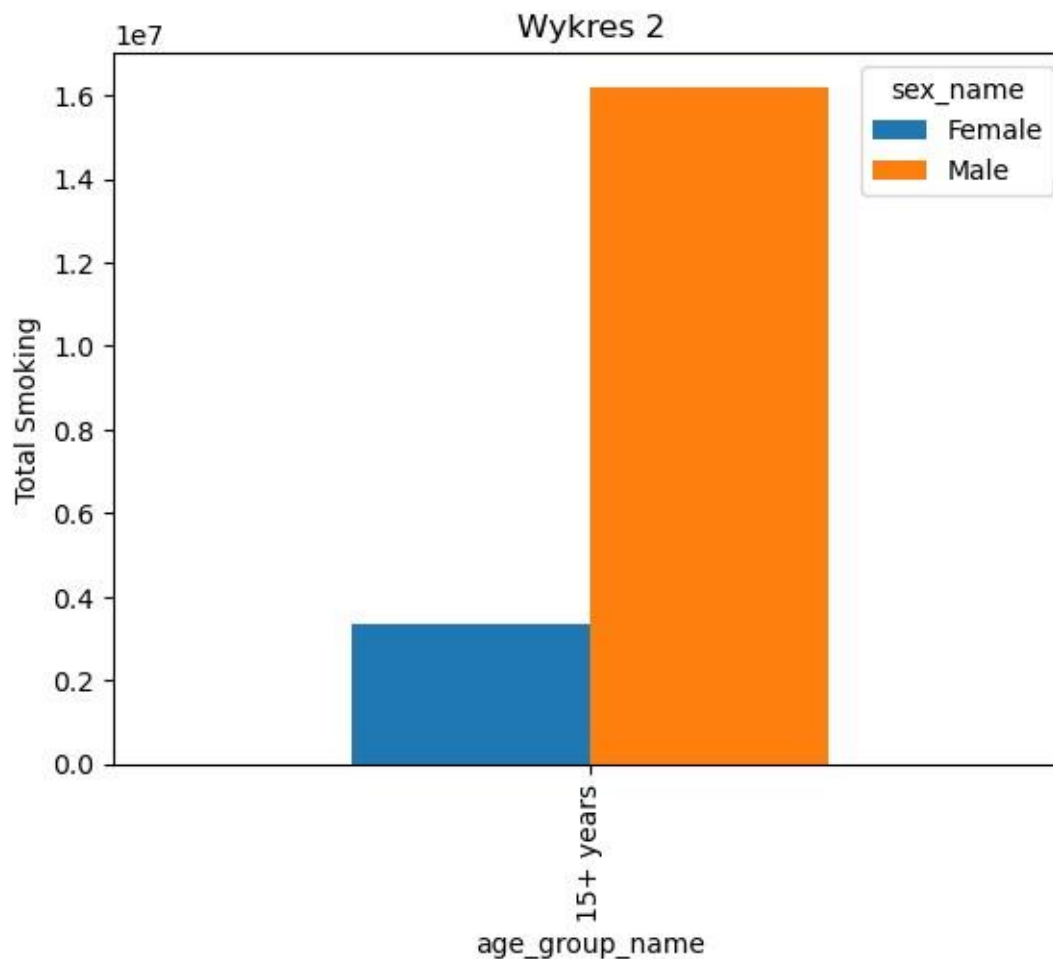
```
[50]: df[(df['location_name'] == 'Global') & (df['age_group_name'] == '15+
years')] &
(df['sex_name'] == 'Both')].pivot_table(values='val',
index='year', columns='sex_name', aggfunc='mean',
fill_value=None,
margins=False, dropna=True).plot(kind='line')
plt.ylabel('Total Smoking') plt.title('Wykres')
[50]: Text(0.5, 1.0, 'Wykres')
```



39 - narysować histogram na podstawie wartości kolumny

```
[51]: df_bar = df[(df['sex_name'].isin(['Male', 'Female'])) & (df['year'] ==
    2018)].pivot_table(values='val',
        index='age_group_name', columns='sex_name',
        aggfunc='mean', fill_value=None, margins=False,
        dropna=True)
df_bar.plot(kind = 'bar')
plt.ylabel('Total Smoking')
plt.title('Wykres 2')
```

```
[51]: Text(0.5, 1.0, 'Wykres 2')
```



40 - przedstawić sposoby łączenia ramek danych za pomocą metod merge i concat

```
[52]: df1 =
pd.read_csv("IHME_GBD_2019_SMOKING_TOB_1990_2019_NUM_SMOKERS_Y2021M05D27.
    csv", encoding = "utf-8") df2 =
```

```
pd.read_csv("Lab1_eiwd_Michał_Stajerski.csv", encoding =
    "utf-8")
```

```
[53]: df1.rename(columns = {'val': 'val_1', 'upper': 'upper_1',
    'lower': 'lower_1'}, inplace = True) df2.rename(columns = {'val': 'val_2', 'upper':
    'upper_2', 'lower': 'lower_2'}, inplace = True)
```

```
[54]: df1
```

```
[54]:      measure_name location_id location_name sex_id sex_name \
```

0	Number of Smokers	1	Global	1	Male
1	Number of Smokers	1	Global	2	Female
2	Number of Smokers	1	Global	3	Both
3	Number of Smokers	1	Global	1	Male
4	Number of Smokers	1	Global	2	Female
...
20965	Number of Smokers	522	Sudan	2	Female
20966	Number of Smokers	522	Sudan	3	Both
20967	Number of Smokers	522	Sudan	1	Male
20968	Number of Smokers	522	Sudan	2	Female
20969	Number of Smokers	522	Sudan	3	Both

	age_group_id	age_group_name	year_id	val_1	upper_1 \
0	29	15+ years	1990	8.031015e+08	
				8.096221e+08	
1	29	15+ years	1990	1.891488e+08	
				1.930929e+08	
2	29	15+ years	1990	9.922503e+08	
				1.000161e+09	
3	29	15+ years	1991	8.138972e+08	
				8.200339e+08	
4	29	15+ years	1991	1.905375e+08	
				1.944249e+08	
...
20965	29	15+ years	2018	2.435999e+05	
				3.286166e+05	
20966	29	15+ years	2018	2.610672e+06	
				2.833943e+06	
20967	29	15+ years	2019	2.439150e+06	
				2.656579e+06	
20968	29	15+ years	2019	2.500800e+05	
				3.345384e+05	
20969	29	15+ years	2019	2.689230e+06	
				2.918332e+06	

	lower_1
0	7.959086e+08
1	1.855595e+08
2	9.847880e+08
3	8.069514e+08
4	1.869744e+08
...	...
20965	1.752508e+05
20966	2.409108e+06
20967	2.236450e+06
20968	1.816686e+05

20969 2.480656e+06

[20970 rows x 11 columns]

```
[55]: df_all = pd.merge(df1, df2, on = ['location_name', 'sex_name', 'age_group_name'], how = 'inner')
df_all.head()
```

```
[55]:  measure_name_x location_id_x location_name sex_id_x sex_name \
0  Number of Smokers      1      Global      1      Male
1  Number of Smokers      1      Global      1      Male
2  Number of Smokers      1      Global      1      Male
3  Number of Smokers      1      Global      1      Male
4  Number of Smokers      1      Global      1      Male
```

```
   age_group_id_x age_group_name year_id      val_1      upper_1 \
0                29 15+ years  1990 803101467.1 809622101.0
1                29 15+ years  1990 803101467.1 809622101.0
2                29 15+ years  1990 803101467.1 809622101.0
3                29 15+ years  1990 803101467.1 809622101.0
4                29 15+ years  1990 803101467.1 809622101.0
```

```
   lower_1 Unnamed: 0  measure_name_y location_id_y sex_id_y \
0  795908635.8      0  Number of Smokers      1      1
1  795908635.8      3  Number of Smokers      1      1
2  795908635.8      6  Number of Smokers      1      1
3  795908635.8      9  Number of Smokers      1      1
4  795908635.8     12  Number of Smokers      1      1
```

```
   age_group_id_y year val_2 upper_2 lower_2
0                29 1990 803101467.1 809622101.0 795908635.8
1                29 1991 813897216.4 820033926.0 806951447.9
2                29 1992 823314827.8 829222821.2 816726365.2
3                29 1993 831387254.4 837293112.8 824949648.0
4                29 1994 837820449.8 843723308.3 831634003.9
```

```
[56]: df_all_1 = df_all.iloc[:50000,:]
df_all_2 = df_all.iloc[50000:,:]

df_all_new = pd.concat([df_all_1, df_all_2], axis = 0)
df_all_new.head()
```

```
[56]:  measure_name_x location_id_x location_name sex_id_x sex_name \
0  Number of Smokers      1      Global      1      Male
1  Number of Smokers      1      Global      1      Male
```

```

2   Number of Smokers      1      Global      1      Male
3   Number of Smokers      1      Global      1      Male 4 Number of Smokers
   1      Global      1      Male

```

```

age_group_id_x age_group_name year_id      val_1      upper_1 \
0              29 15+ years  1990 803101467.1 809622101.0
1              29 15+ years  1990 803101467.1 809622101.0
2              29 15+ years  1990 803101467.1 809622101.0
3              29 15+ years  1990 803101467.1 809622101.0
4              29 15+ years  1990 803101467.1 809622101.0

```

```

lower_1 Unnamed: 0 measure_name_y location_id_y sex_id_y \
0 795908635.8      0 Number of Smokers      1      1
1 795908635.8      3 Number of Smokers      1      1
2 795908635.8      6 Number of Smokers      1      1
3 795908635.8      9 Number of Smokers      1      1
4 795908635.8     12 Number of Smokers      1      1

```

```

age_group_id_y year val_2 upper_2 lower_2
0              29 1990 803101467.1 809622101.0 795908635.8
1              29 1991 813897216.4 820033926.0 806951447.9
2              29 1992 823314827.8 829222821.2 816726365.2
3              29 1993 831387254.4 837293112.8 824949648.0
4              29 1994 837820449.8 843723308.3 831634003.9

```

41 - pokazać dodawanie nowych kolumn za pomocą operacji matematycznych

```

[57]: df_all["val1_round"] = df_all["val_1"].round(decimals = 1)
      df_all.head()

```

```

[57]: measure_name_x location_id_x location_name sex_id_x sex_name \
0   Number of Smokers      1      Global      1      Male
1   Number of Smokers      1      Global      1      Male
2   Number of Smokers      1      Global      1      Male
3   Number of Smokers      1      Global      1      Male 4 Number of Smokers
   1      Global      1      Male

```

```

age_group_id_x age_group_name year_id      val_1      upper_1 ... \
0              29 15+ years  1990 803101467.1 809622101.0 ...
1              29 15+ years  1990 803101467.1 809622101.0 ...
2              29 15+ years  1990 803101467.1 809622101.0 ...
3              29 15+ years  1990 803101467.1 809622101.0 ...
4              29 15+ years  1990 803101467.1 809622101.0 ...

```

```

    Unnamed: 0  measure_name_y location_id_y sex_id_y age_group_id_y \
0           0 0 Number of Smokers 1      1      29
1           1 3 Number of Smokers 1      1      29
2           2 6 Number of Smokers 1      1      29
3           3 9 Number of Smokers 1      1      29 4 12 Number of Smokers
           1      1      29

```

```

    year      val_2      upper_2      lower_2  vall_round
0  1990  803101467.1  809622101.0  795908635.8  803101467.1
1  1991  813897216.4  820033926.0  806951447.9  803101467.1
2  1992  823314827.8  829222821.2  816726365.2  803101467.1
3  1993  831387254.4  837293112.8  824949648.0  803101467.1
4  1994  837820449.8  843723308.3  831634003.9  803101467.1

```

[5 rows x 21 columns]

```

[58]: df_all["total"] = df_all["val_1"] + df_all["upper_1"] + df_all["lower_1"]
      df_all

```

```

[58]: measure_name_x location_id_x location_name sex_id_x sex_name \
0      Number of Smokers 1      Global 1      Male
1      Number of Smokers 1      Global 1      Male
2      Number of Smokers 1      Global 1      Male
3      Number of Smokers 1      Global 1      Male
4      Number of Smokers 1      Global 1      Male
...
639895 Number of Smokers 522  Sudan 3      Both
639896 Number of Smokers 522  Sudan 3      Both
639897 Number of Smokers 522  Sudan 3      Both
639898 Number of Smokers 522  Sudan 3      Both
639899 Number of Smokers 522  Sudan 3      Both

```

```

    age_group_id_x age_group_name year_id      val_1      upper_1 \
0           29  15+ years  1990  8.031015e+08  8.096221e+08
1           29  15+ years  1990  8.031015e+08  8.096221e+08
2           29  15+ years  1990  8.031015e+08  8.096221e+08
3           29  15+ years  1990  8.031015e+08  8.096221e+08
4           29  15+ years  1990  8.031015e+08  8.096221e+08
...
639895      29  15+ years  2019  2.689230e+06  2.918332e+06
639896      29  15+ years  2019  2.689230e+06  2.918332e+06
639897      29  15+ years  2019  2.689230e+06  2.918332e+06
639898      29  15+ years  2019  2.689230e+06  2.918332e+06
639899      29  15+ years  2019  2.689230e+06  2.918332e+06

```

```

    ... measure_name_y location_id_y sex_id_y age_group_id_y year \
0      ... Number of Smokers 1      1      29  1990

```

```

1      ... Number of Smokers 1      1      29 1991
2      ... Number of Smokers 1      1      29 1992
3      ... Number of Smokers 1      1      29 1993
4      ... Number of Smokers 1      1      29 1994
...
639895 ... Number of Smokers 522    3      29 2015
639896 ... Number of Smokers 522    3      29 2016
639897 ... Number of Smokers 522    3      29 2017
639898 ... Number of Smokers 522    3      29 2018
639899 ... Number of Smokers 522    3      29 2019

      val_2      upper_2      lower_2  vall_round      total
0      8.031015e+08 8.096221e+08 7.959086e+08 803101467.1
      2.408632e+09 1      8.138972e+08 8.200339e+08 8.069514e+08
803101467.1 2.408632e+09 2      8.233148e+08 8.292228e+08 8.167264e+08
803101467.1 2.408632e+09 3      8.313873e+08 8.372931e+08 8.249496e+08
      803101467.1 2.408632e+09
4      8.378204e+08 8.437233e+08 8.316340e+08 803101467.1 2.408632e+09
...
639895 2.388216e+06 2.587005e+06 2.211144e+06 2689229.6 8.088217e+06
639896 2.454893e+06 2.665441e+06 2.267696e+06 2689229.6
      8.088217e+06
639897 2.535003e+06 2.743769e+06 2.341329e+06 2689229.6
      8.088217e+06
639898 2.610672e+06 2.833943e+06 2.409108e+06 2689229.6
      8.088217e+06
639899 2.689230e+06 2.918332e+06 2.480656e+06 2689229.6
      8.088217e+06
[639900 rows x 22 columns]

```

42 - przedstawić na przykładzie dodawanie nowych kolumn z pomocą funkcji lambda

```
[59]: CIS_2020 = ['Poland', 'Hungary', 'Italia', 'Germany', 'France',
                 'Spain', 'Romania']
```

```
[60]: df_all['CIS_2020'] = df_all['location_name'].apply(lambda x: True if x in_
    ↪ CIS_2020 else False )
df_all[df_all['CIS_2020'] == True]
```

```
[60]:      measure_name_x location_id_x location_name sex_id_x sex_name \
121500 Number of Smokers      48      Hungary      1      Male
121501 Number of Smokers      48      Hungary      1      Male
121502 Number of Smokers      48      Hungary      1      Male
121503 Number of Smokers      48      Hungary      1      Male
121504 Number of Smokers      48      Hungary      1      Male
...

```

242995	Number of Smokers	92	Spain 3	Both
242996	Number of Smokers	92	Spain 3	Both
242997	Number of Smokers	92	Spain 3	Both
242998	Number of Smokers	92	Spain 3	Both
242999	Number of Smokers	92	Spain 3	Both

	age_group_id_x	age_group_name	year_id	val_1	upper_1 \
121500	29	15+ years	1990	1691795.129	1.764520e+06
121501	29	15+ years	1990	1691795.129	1.764520e+06
121502	29	15+ years	1990	1691795.129	1.764520e+06
121503	29	15+ years	1990	1691795.129	1.764520e+06
121504	29	15+ years	1990	1691795.129	1.764520e+06
...
242995	29	15+ years	2019	9748202.722	1.023282e+07
242996	29	15+ years	2019	9748202.722	1.023282e+07
242997	29	15+ years	2019	9748202.722	1.023282e+07
242998	29	15+ years	2019	9748202.722	1.023282e+07
242999	29	15+ years	2019	9748202.722	1.023282e+07

...	location_id_y	sex_id_y	age_group_id_y	year	val_2 \
121500	...	48	1	29 1990	1.691795e+06
121501	...	48	1	29 1991	1.683045e+06
121502	...	48	1	29 1992	1.674390e+06
121503	...	48	1	29 1993	1.665224e+06
121504	...	48	1	29 1994	1.658081e+06
...
242995	...	92	3	29 2015	1.117824e+07
242996	...	92	3	29 2016	1.070556e+07
242997	...	92	3	29 2017	1.031479e+07
242998	...	92	3	29 2018	1.001427e+07
242999	...	92	3	29 2019	9.748203e+06

	upper_2	lower_2	val1_round	total	CIS_2020
121500	1.764520e+06		1691795.1	5.075818e+06	True
	1.619503e+06				
121501	1.753812e+06		1691795.1	5.075818e+06	True
	1.611946e+06				
121502	1.742527e+06		1691795.1	5.075818e+06	True
	1.607019e+06				
121503	1.730905e+06		1691795.1	5.075818e+06	True
	1.601383e+06				
121504	1.722809e+06		1691795.1	5.075818e+06	True
	1.596858e+06				
...

```

242995 1.152098e+07          9748202.7 2.928604e+07    True
1.086683e+07
242996 1.107828e+07          9748202.7 2.928604e+07    True
1.037339e+07
242997 1.071837e+07          9748202.7 2.928604e+07    True
9.948682e+06
242998 1.046173e+07          9748202.7 2.928604e+07    True
9.610001e+06
242999 1.023282e+07          9748202.7 2.928604e+07    True
9.305015e+06
[16200 rows x 23 columns]

```

43 - przedstawić możliwości pracy z dużymi plikami przy użyciu argumentu chunksize

```
[61]: df.to_csv('df_all.csv')
```

```

for chunk_df in pd.read_csv('df_all.csv',
                             chunksize = 50000):
    print("CHUNK DF")
    print(chunk_df.head())

```

CHUNK DF

```

Unnamed: 0  measure_name location_id location_name sex_id sex_name \
0          0 Number of Smokers      1      Global      1      Male 1      1
Number of Smokers      1      Global      2 Female 2 2 Number of
Smokers      1      Global      3      Both
3          3 Number of Smokers      1      Global      1      Male
4          4 Number of Smokers      1      Global      2 Female

```

```

age_group_id age_group_name year val upper lower 0 29 15+ years
1990 803101467.1 8.096221e+08 795908635.8
1          29      15+ years 1990 189148834.0 1.930929e+08 185559469.9
2          29      15+ years 1990 992250301.2 1.000161e+09 984788043.8
3          29      15+ years 1991 813897216.4 8.200339e+08 806951447.9
4          29      15+ years 1991 190537545.1 1.944249e+08 186974424.5

```

```

[62]: new_df = pd.DataFrame()
for chunk_df in pd.read_csv('df_all.csv',
                             chunksize = 50000):
    result = chunk_df.groupby(['location_name', 'sex_name']).agg({'upper': '_
    ↪mean',
                                                                    'lower': 'mean'})
    new_df = pd.concat([new_df, result])

```

```
[63]: new_df
```

```
[63]:          upper          lower
```

```

location_name
sex_name
Afghanistan Both 1.184427e+06
               9.776876e+05
               Female 1.867379e+05
               1.060589e+05
               Male 1.037830e+06
               8.447279e+05
Albania Both 6.302436e+05
           5.752316e+05
           Female 1.248055e+05
           8.917709e+04
...
Zambia Female 2.766568e+05
         1.879562e+05
         Male 8.156664e+05
         7.266267e+05
Zimbabwe Both 1.132936e+06
              1.018202e+06
              Female 1.442346e+05
              9.511072e+04
              Male 1.010215e+06
              9.072602e+05
[693 rows x 2 columns] [ ]:

```

