SPRAWOZDANIE

Zajęcia: Eksploracja i wizualizacja danych
Prowadzący: prof. dr hab. Vasyl Martsenyuk

Laboratorium: 1
Data: 23.02.2023
Temat: "Wstęp do Python. Biblioteka Pandas"
Wariant: 7

Justyna Kowal
Informatyka II stopień,
stacjonarne,
semestr 3,
Gr. 1
https://github.com/amaix3/eiwd

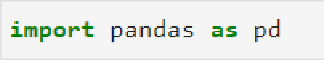
1. Polecenie

Celem zajęć jest nabycie podstawowej znajomości języka Python - rozwiązując zadanie

tworzenia i wyświetlenia ramki danych odpowiednio do określonego wariantu. Dane do zadania zostały pobrane ze strony https://ghdx.healthdata.org/ihme_data. Wariant wybrany w zadaniu jest wariant 7: Global Burden of Disease Study 2019 (GBD 2019) Smoking Tobacco Use Prevalence 1990-2019

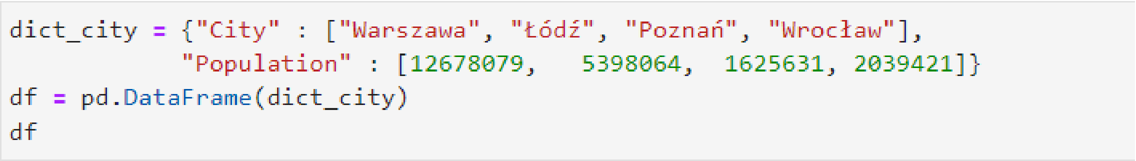
2. Zadania

1 - ładowanie biblioteki Pandas

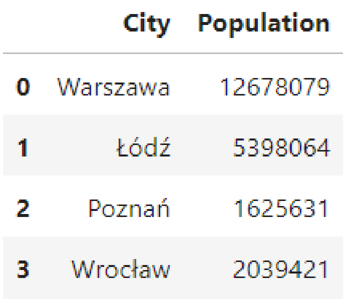


```
import pandas as pd
```

2 - tworzenie ramki danych ze słownika

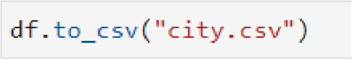


```
dict_city = {"City" : ["Warszawa", "Łódź", "Poznań", "Wrocław"],
             "Population" : [12678079,   5398064,   1625631, 2039421]}
df = pd.DataFrame(dict_city)
df
```



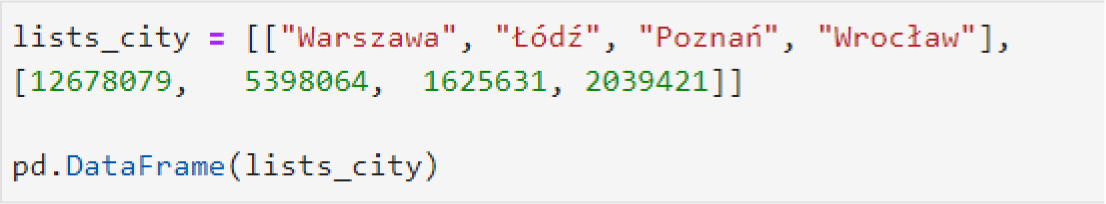
	City	Population
0	Warszawa	12678079
1	Łódź	5398064
2	Poznań	1625631
3	Wrocław	2039421

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



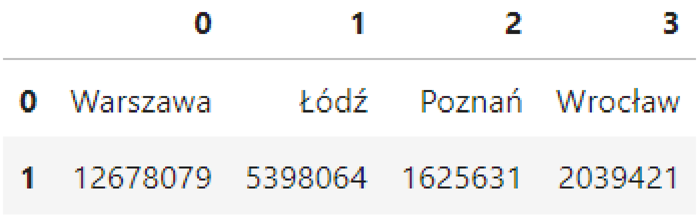
```
df.to_csv("city.csv")
```

4 - tworzenie ramki danych z listy list



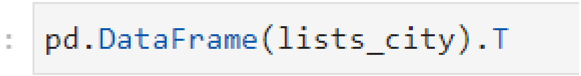
```
lists_city = [["Warszawa", "Łódź", "Poznań", "Wrocław"],
[12678079,   5398064,   1625631, 2039421]]

pd.DataFrame(lists_city)
```

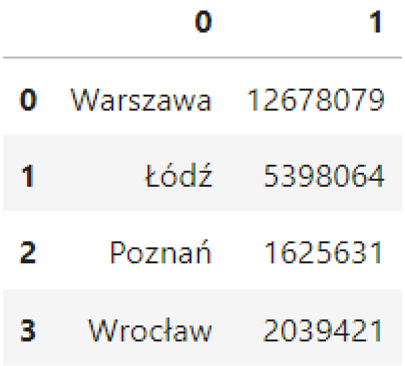


	0	1	2	3
0	Warszawa	Łódź	Poznań	Wrocław
1	12678079	5398064	1625631	2039421

5 - transponowanie (wymieniamy kolumny a wierszy)

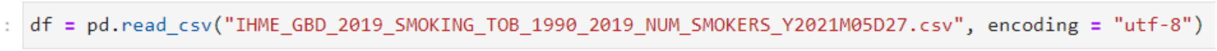


```
pd.DataFrame(lists_city).T
```

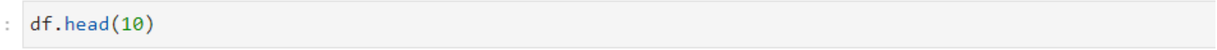


	0	1
0	Warszawa	12678079
1	Łódź	5398064
2	Poznań	1625631
3	Wrocław	2039421

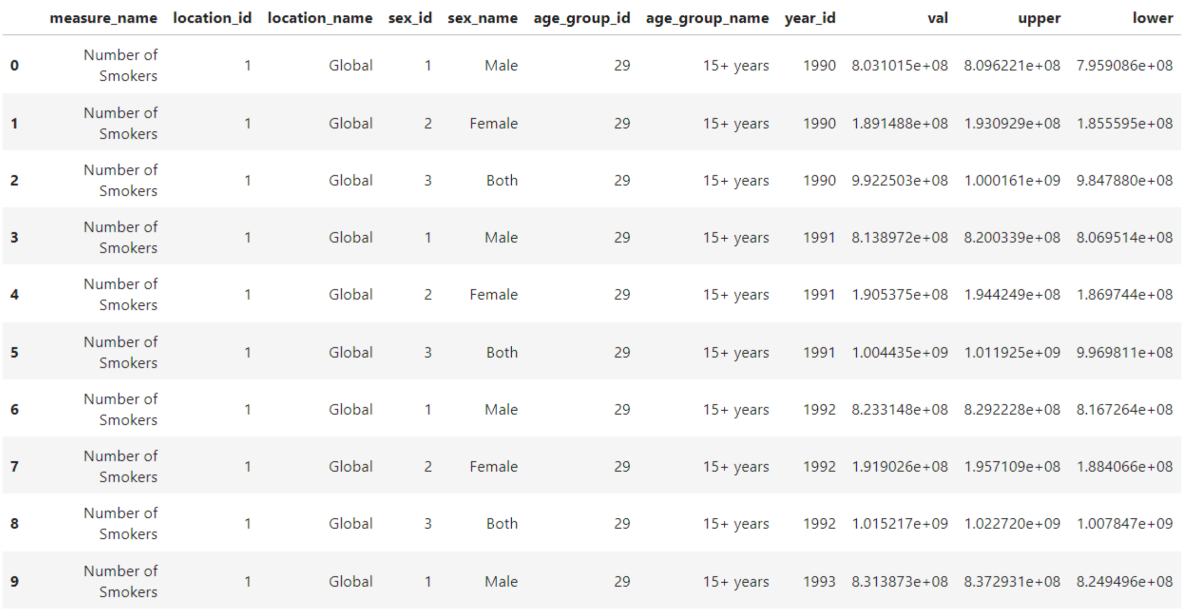
6 - wyświetlić pierwsze 10 wierszy ramki danych



```
df = pd.read_csv("IHME_GBD_2019_SMOKING_TOB_1990_2019_NUM_SMOKERS_Y2021M05D27.csv", encoding = "utf-8")
```

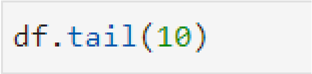


```
df.head(10)
```

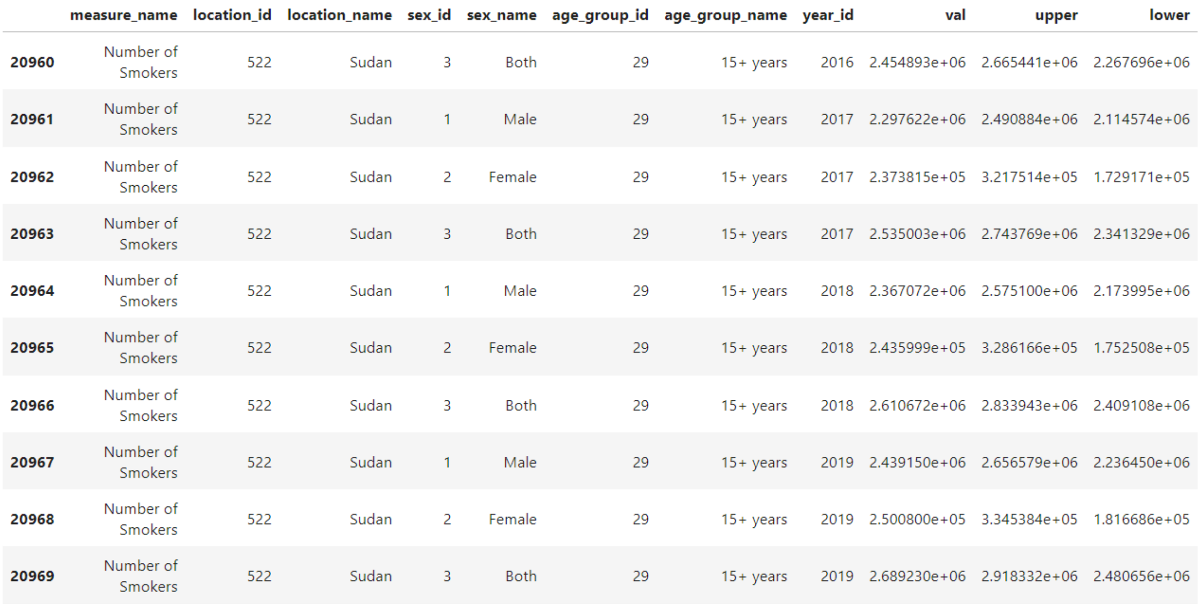


	measure_name	location_id	location_name	sex_id	sex_name	age_group_id	age_group_name	year_id	val	upper	lower
0	Number of Smokers	1	Global	1	Male	29	15+ years	1990	8.031015e+08	8.096221e+08	7.959086e+08
1	Number of Smokers	1	Global	2	Female	29	15+ years	1990	1.891488e+08	1.930929e+08	1.855595e+08
2	Number 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
5	Number of Smokers	1	Global	3	Both	29	15+ years	1991	1.004435e+09	1.011925e+09	9.969811e+08
6	Number of Smokers	1	Global	1	Male	29	15+ years	1992	8.233148e+08	8.292228e+08	8.167264e+08
7	Number of Smokers	1	Global	2	Female	29	15+ years	1992	1.919026e+08	1.957109e+08	1.884066e+08
8	Number of Smokers	1	Global	3	Both	29	15+ years	1992	1.015217e+09	1.022720e+09	1.007847e+09
9	Number of Smokers	1	Global	1	Male	29	15+ years	1993	8.313873e+08	8.372931e+08	8.249496e+08

7 - wyświetlić ostatnie 10 wierszy ramki danych

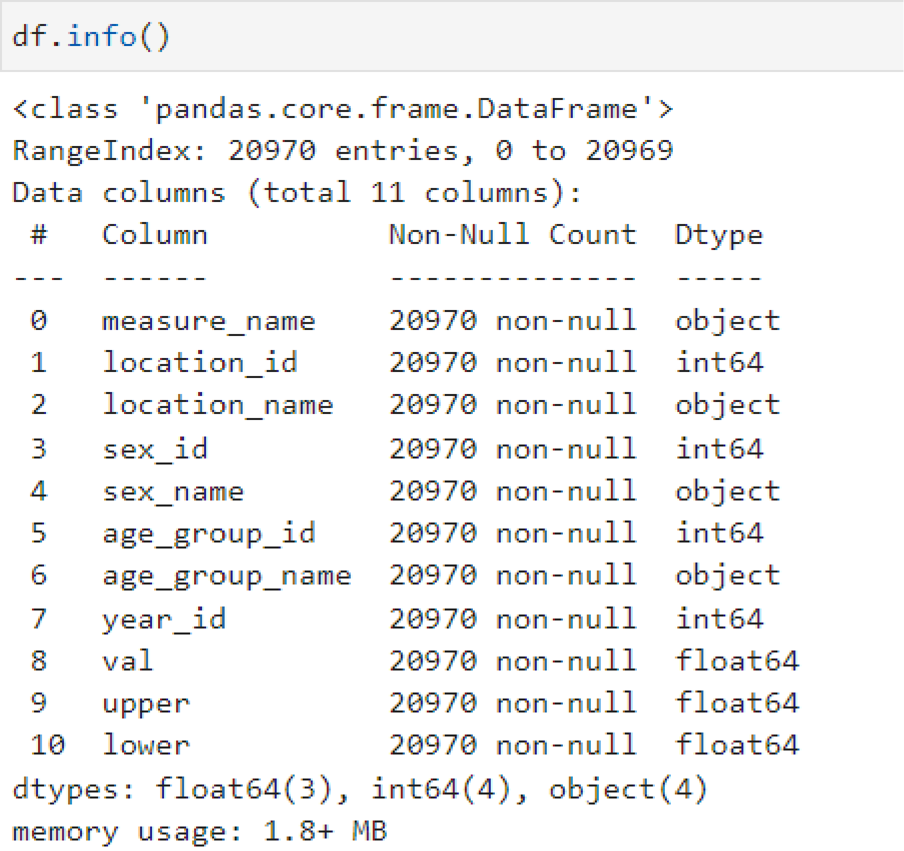


```
df.tail(10)
```



	measure_name	location_id	location_name	sex_id	sex_name	age_group_id	age_group_name	year_id	val	upper	lower
20960	Number of Smokers	522	Sudan	3	Both	29	15+ years	2016	2.454893e+06	2.665441e+06	2.267696e+06
20961	Number of Smokers	522	Sudan	1	Male	29	15+ years	2017	2.297622e+06	2.490884e+06	2.114574e+06
20962	Number of Smokers	522	Sudan	2	Female	29	15+ years	2017	2.373815e+05	3.217514e+05	1.729171e+05
20963	Number of Smokers	522	Sudan	3	Both	29	15+ years	2017	2.535003e+06	2.743769e+06	2.341329e+06
20964	Number of Smokers	522	Sudan	1	Male	29	15+ years	2018	2.367072e+06	2.575100e+06	2.173995e+06
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

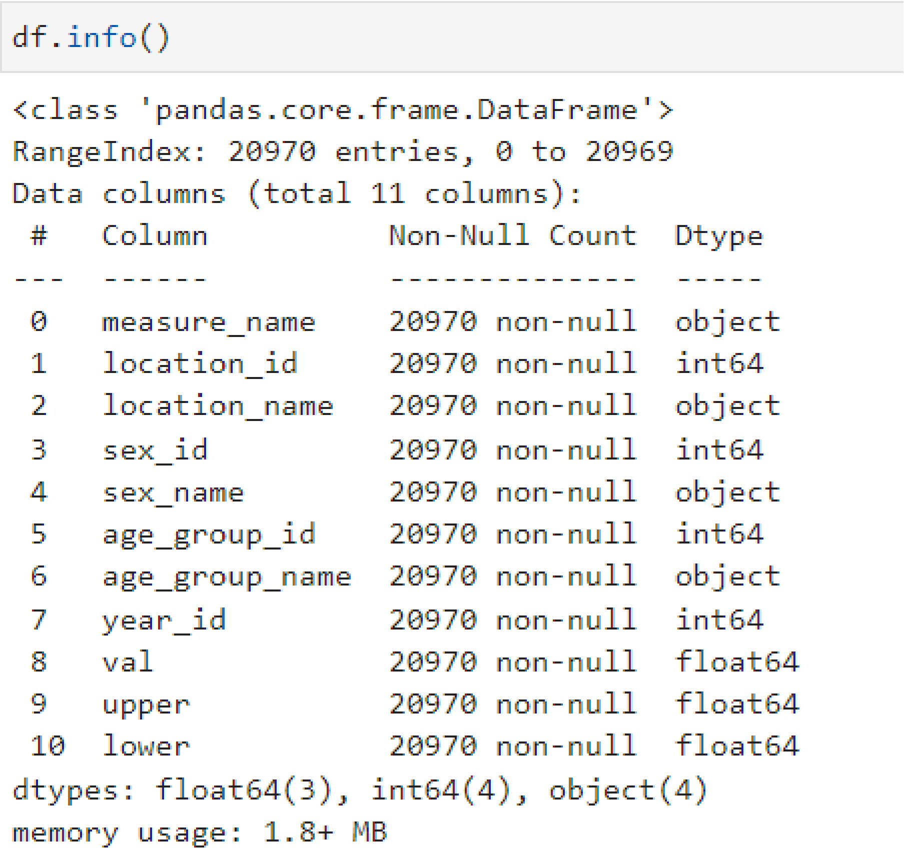
8 - wyświetlić informację o ramce danych



```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 20970 entries, 0 to 20969
Data columns (total 11 columns):
 #   Column          Non-Null Count  Dtype  
---  ------          --------------  -----  
 0   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
```

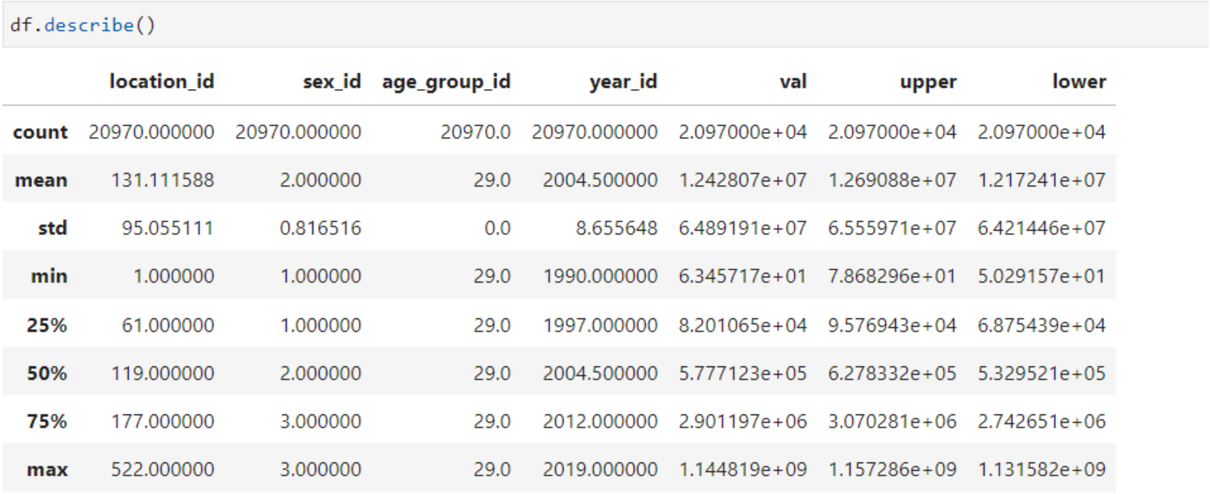
9 - wyświetlić, ile wierszy i kolumn znajduje się w ramce danych



```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 20970 entries, 0 to 20969
Data columns (total 11 columns):
 #   Column          Non-Null Count  Dtype  
---  ------          --------------  -----  
 0   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
```

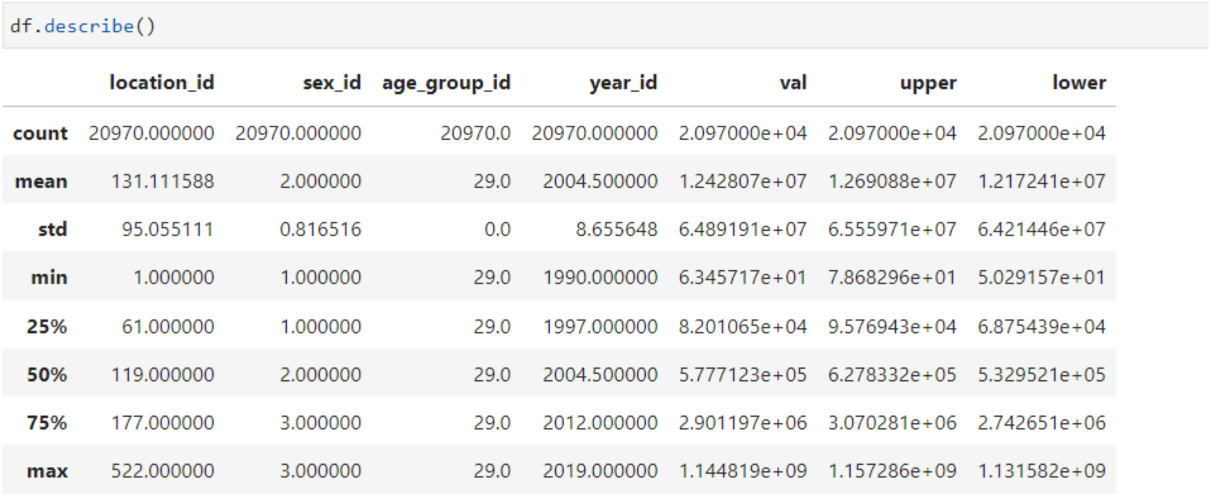
10 - wyświetlić informację statystyczną o kolumnach liczbowych (wartości niepowtarzalne, średnia, odchylenie standardowe, minimum, kwartale, maksimum)



```
df.describe()
```

	location_id	sex_id	age_group_id	year_id	val	upper	lower
count	20970.000000	20970.000000	20970.0	20970.000000	2.097000e+04	2.097000e+04	2.097000e+04
mean	131.111588	2.000000	29.0	2004.500000	1.242807e+07	1.269088e+07	1.217241e+07
std	95.055111	0.816516	0.0	8.655648	6.489191e+07	6.555971e+07	6.421446e+07
min	1.000000	1.000000	29.0	1990.000000	6.345717e+01	7.868296e+01	5.029157e+01
25%	61.000000	1.000000	29.0	1997.000000	8.201065e+04	9.576943e+04	6.875439e+04
50%	119.000000	2.000000	29.0	2004.500000	5.777123e+05	6.278332e+05	5.329521e+05
75%	177.000000	3.000000	29.0	2012.000000	2.901197e+06	3.070281e+06	2.742651e+06
max	522.000000	3.000000	29.0	2019.000000	1.144819e+09	1.157286e+09	1.131582e+09

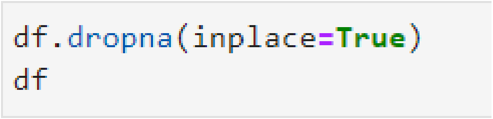
11 - wyświetlić informację statystyczną o kolumnach kategoryzowanych (ile unikalnych wartości, top - jaka jest najpopularniejsza wartość, freq - jak często najpopularniejsza)



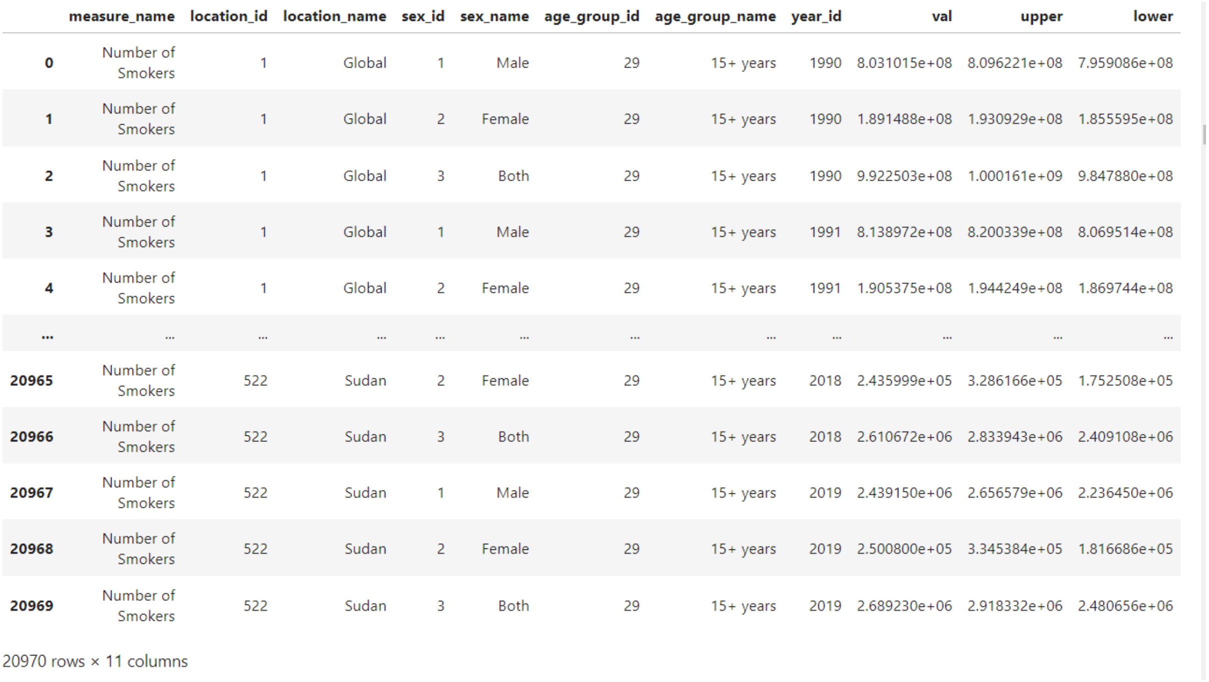
```
df.describe()
```

	location_id	sex_id	age_group_id	year_id	val	upper	lower
count	20970.000000	20970.000000	20970.0	20970.000000	2.097000e+04	2.097000e+04	2.097000e+04
mean	131.111588	2.000000	29.0	2004.500000	1.242807e+07	1.269088e+07	1.217241e+07
std	95.055111	0.816516	0.0	8.655648	6.489191e+07	6.555971e+07	6.421446e+07
min	1.000000	1.000000	29.0	1990.000000	6.345717e+01	7.868296e+01	5.029157e+01
25%	61.000000	1.000000	29.0	1997.000000	8.201065e+04	9.576943e+04	6.875439e+04
50%	119.000000	2.000000	29.0	2004.500000	5.777123e+05	6.278332e+05	5.329521e+05
75%	177.000000	3.000000	29.0	2012.000000	2.901197e+06	3.070281e+06	2.742651e+06
max	522.000000	3.000000	29.0	2019.000000	1.144819e+09	1.157286e+09	1.131582e+09

12 - usunąć brakujące wartości w ramce danych



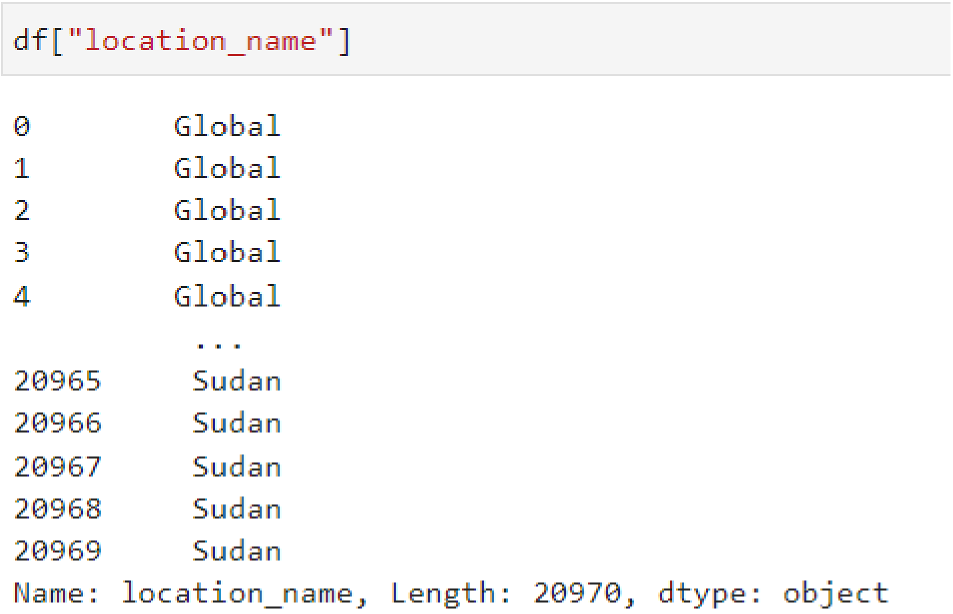
```
df.dropna(inplace=True)
df
```



	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

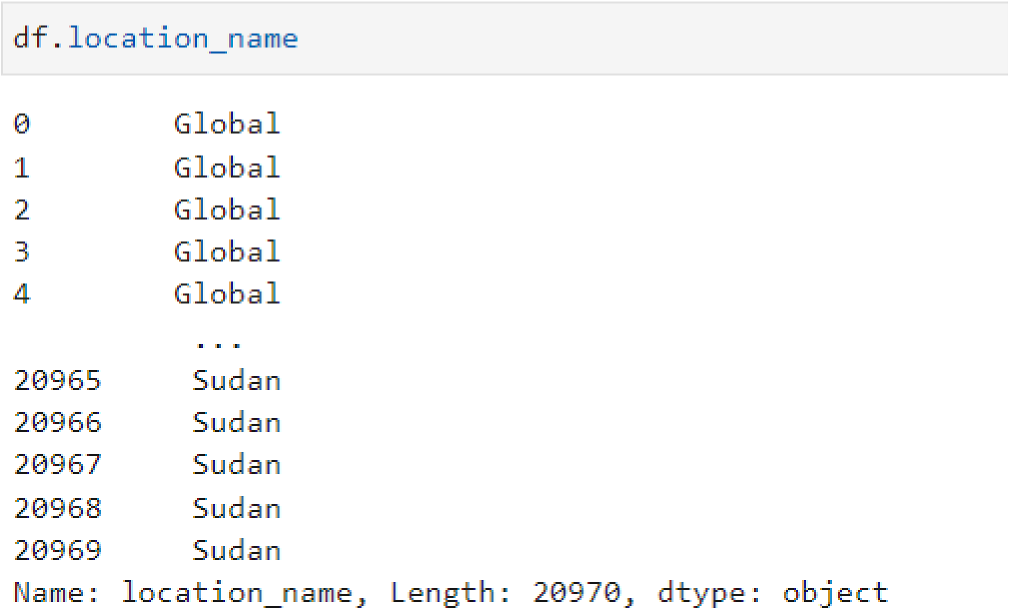
20970 rows × 11 columns

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



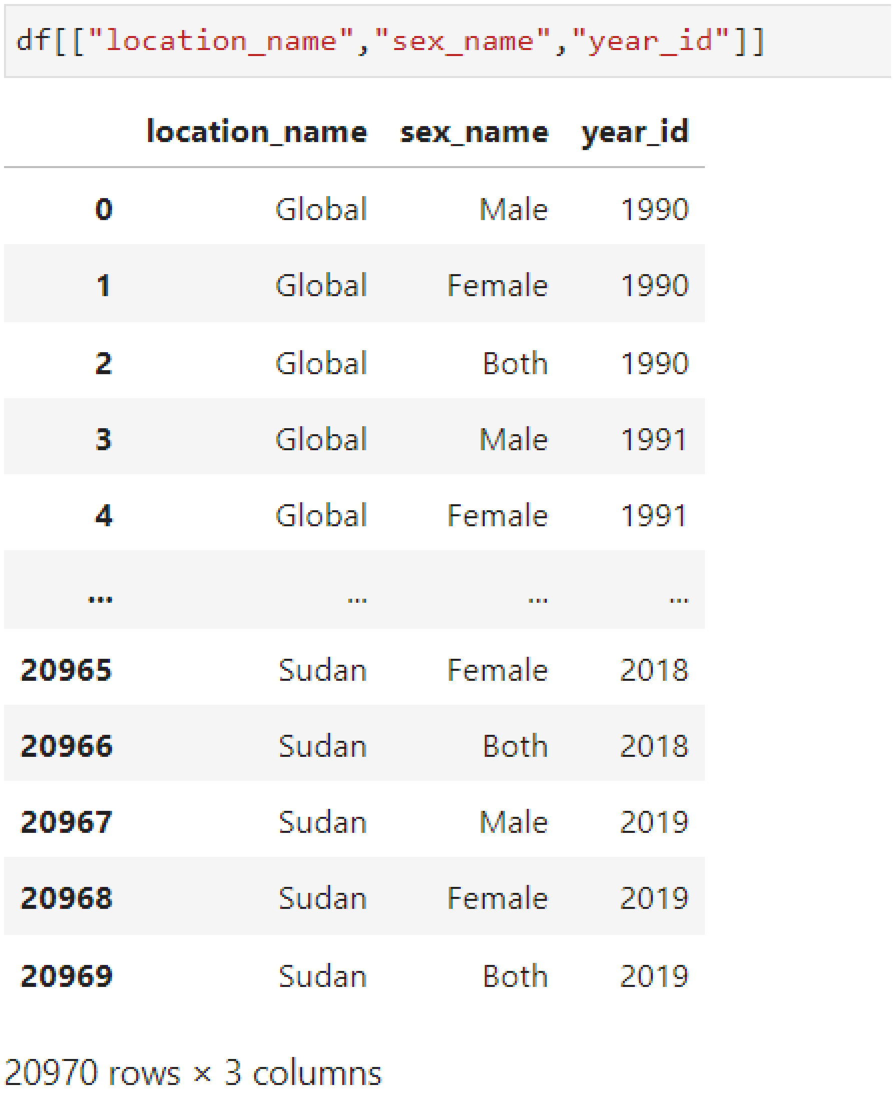
```
df["location_name"]
```

```
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
```



```
df.location_name
```

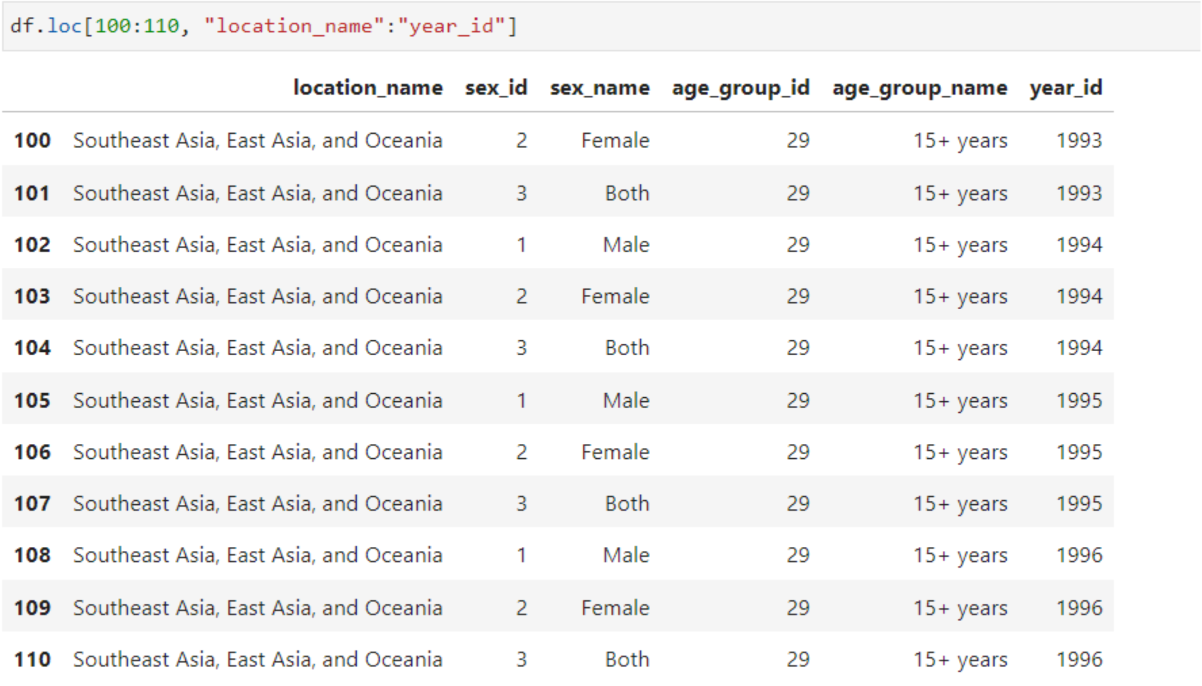
```
0        Global
1        Global
2        Global
3        Global
4        Global
          ...
20965     Sudan
20966     Sudan
20967     Sudan
20968     Sudan
20969     Sudan
Name: location_name, Length: 20970, dtype: object
```



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

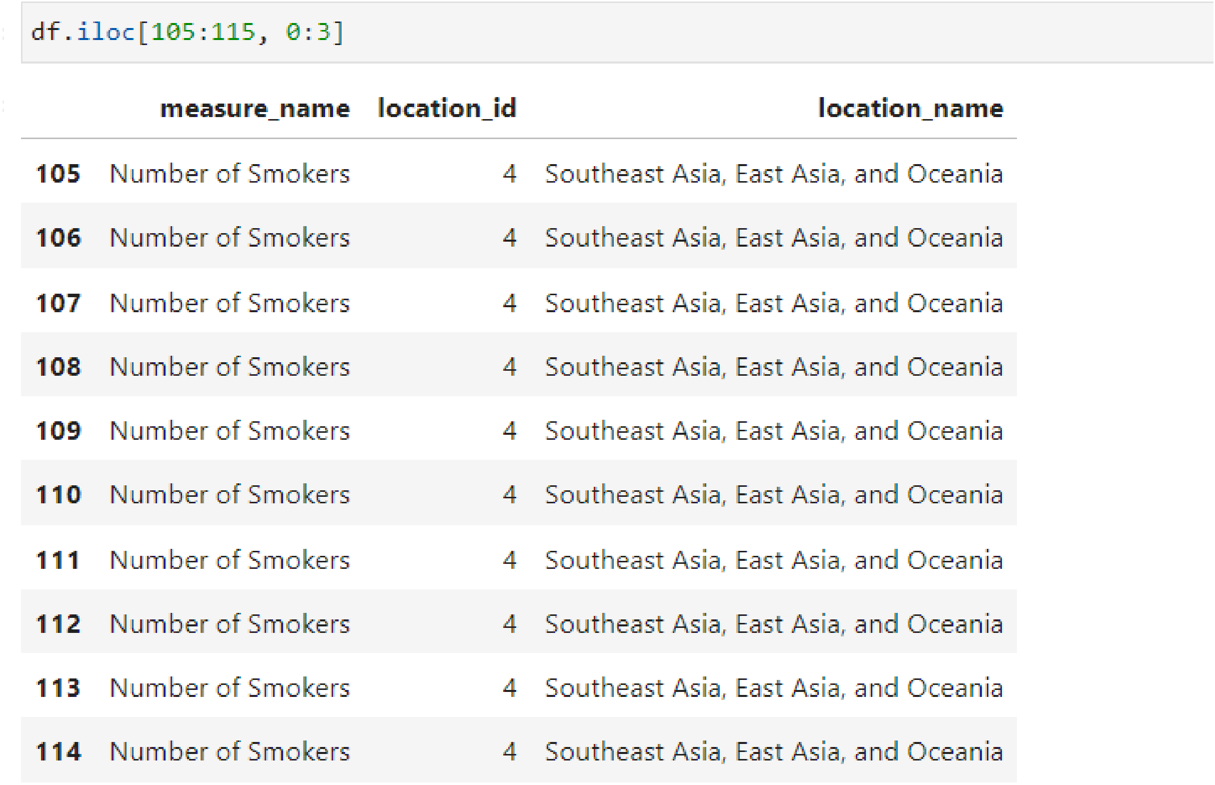
	location_name	sex_name	year_id
0	Global	Male	1990
1	Global	Female	1990
2	Global	Both	1990
3	Global	Male	1991
4	Global	Female	1991
...	...	...	...
20965	Sudan	Female	2018
20966	Sudan	Both	2018
20967	Sudan	Male	2019
20968	Sudan	Female	2019
20969	Sudan	Both	2019

20970 rows × 3 columns



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

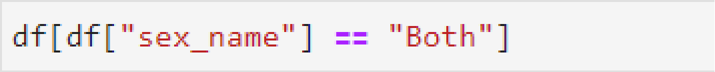
	location_name	sex_id	sex_name	age_group_id	age_group_name	year_id
100	Southeast Asia, East Asia, and Oceania	2	Female	29	15+ years	1993
101	Southeast Asia, East Asia, and Oceania	3	Both	29	15+ years	1993
102	Southeast Asia, East Asia, and Oceania	1	Male	29	15+ years	1994
103	Southeast Asia, East Asia, and Oceania	2	Female	29	15+ years	1994
104	Southeast Asia, East Asia, and Oceania	3	Both	29	15+ years	1994
105	Southeast Asia, East Asia, and Oceania	1	Male	29	15+ years	1995
106	Southeast Asia, East Asia, and Oceania	2	Female	29	15+ years	1995
107	Southeast Asia, East Asia, and Oceania	3	Both	29	15+ years	1995
108	Southeast Asia, East Asia, and Oceania	1	Male	29	15+ years	1996
109	Southeast Asia, East Asia, and Oceania	2	Female	29	15+ years	1996
110	Southeast Asia, East Asia, and Oceania	3	Both	29	15+ years	1996



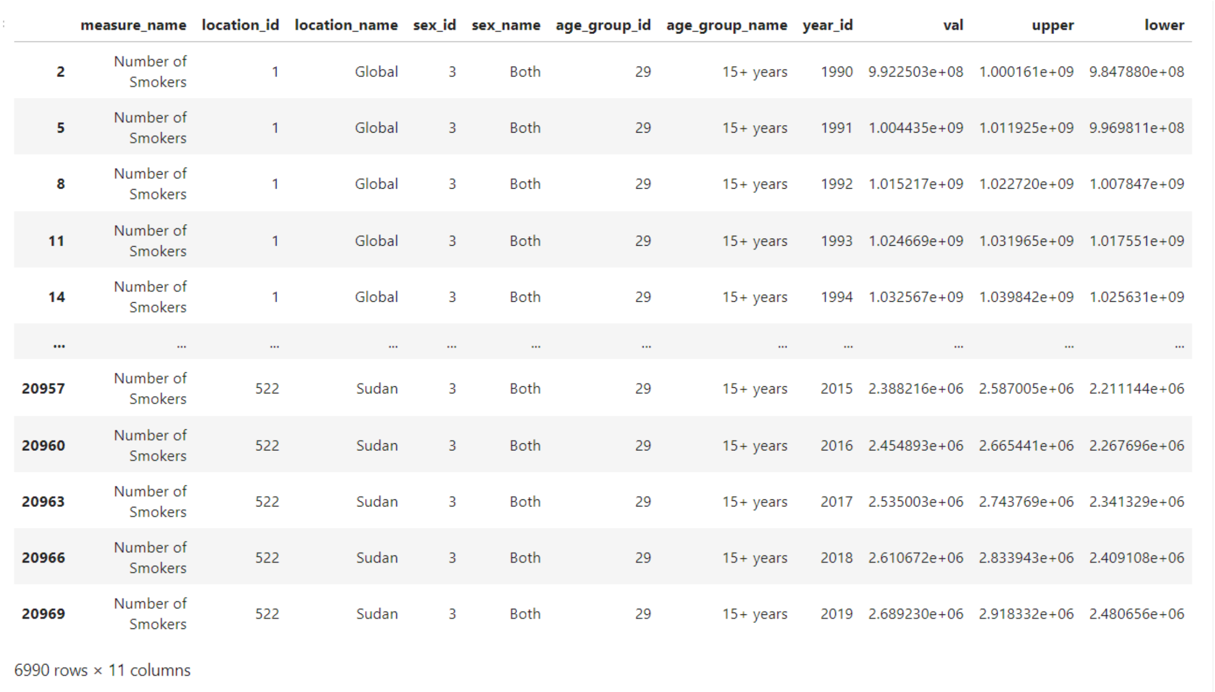
```
df.iloc[105:115, 0:3]
```

	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



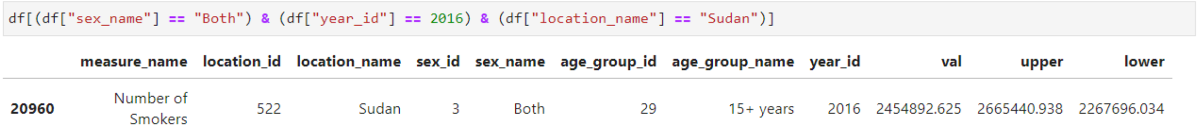
```
df[df["sex_name"] == "Both"]
```



	measure_name	location_id	location_name	sex_id	sex_name	age_group_id	age_group_name	year_id	val	upper	lower
2	Number of Smokers	1	Global	3	Both	29	15+ years	1990	9.922503e+08	1.000161e+09	9.847880e+08
5	Number of Smokers	1	Global	3	Both	29	15+ years	1991	1.004435e+09	1.011925e+09	9.969811e+08
8	Number of Smokers	1	Global	3	Both	29	15+ years	1992	1.015217e+09	1.022720e+09	1.007847e+09
11	Number of Smokers	1	Global	3	Both	29	15+ years	1993	1.024669e+09	1.031965e+09	1.017551e+09
14	Number of Smokers	1	Global	3	Both	29	15+ years	1994	1.032567e+09	1.039842e+09	1.025631e+09
...	...	...	...	...	...	...	...	...	...	...	...
20957	Number of Smokers	522	Sudan	3	Both	29	15+ years	2015	2.388216e+06	2.587005e+06	2.211144e+06
20960	Number of Smokers	522	Sudan	3	Both	29	15+ years	2016	2.454893e+06	2.665441e+06	2.267696e+06
20963	Number of Smokers	522	Sudan	3	Both	29	15+ years	2017	2.535003e+06	2.743769e+06	2.341329e+06
20966	Number of Smokers	522	Sudan	3	Both	29	15+ years	2018	2.610672e+06	2.833943e+06	2.409108e+06
20969	Number of Smokers	522	Sudan	3	Both	29	15+ years	2019	2.689230e+06	2.918332e+06	2.480656e+06

6990 rows × 11 columns

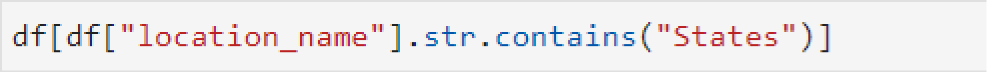
15 - przedstawić wybór wierszy z ramki danych pod warunkiem spełnienia kilku warunków jednocześnie



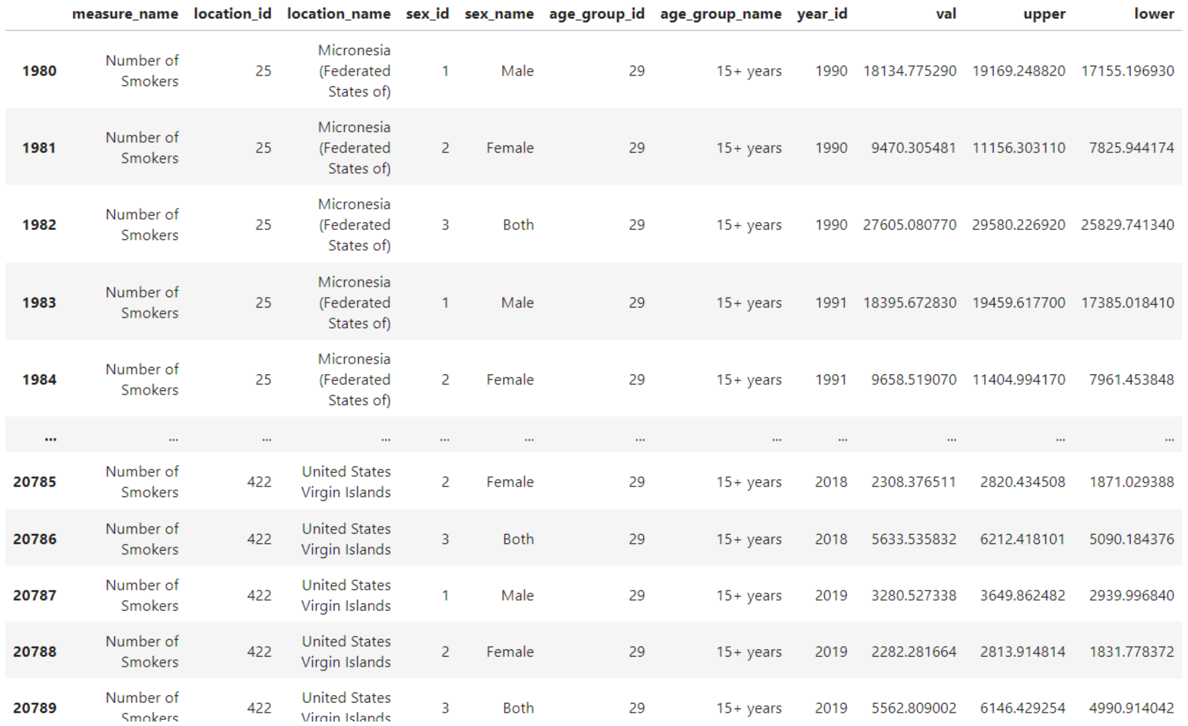
```
df[(df["sex_name"] == "Both") & (df["year_id"] == 2016) & (df["location_name"] == "Sudan")]
```

	measure_name	location_id	location_name	sex_id	sex_name	age_group_id	age_group_name	year_id	val	upper	lower
20960	Number of Smokers	522	Sudan	3	Both	29	15+ years	2016	2454892.625	2665440.938	2267696.034

16 - wybrać wiersze które zawierają w kolumnie kategoryzowanej określone słowo

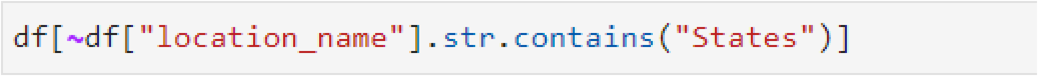


```
df[df["location_name"].str.contains("States")]
```

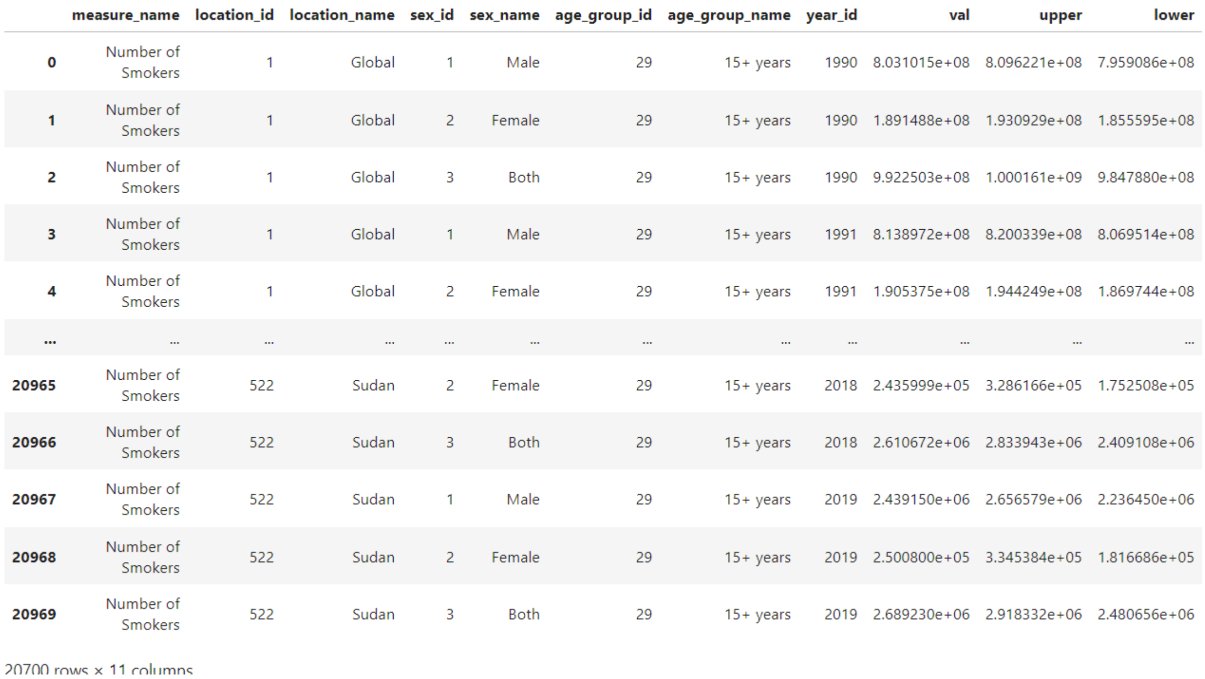


	measure_name	location_id	location_name	sex_id	sex_name	age_group_id	age_group_name	year_id	val	upper	lower
1980	Number of Smokers	25	Micronesia (Federated States of)	1	Male	29	15+ years	1990	18134.775290	19169.248820	17155.196930
1981	Number of Smokers	25	Micronesia (Federated States of)	2	Female	29	15+ years	1990	9470.305481	11156.303110	7825.944174
1982	Number of Smokers	25	Micronesia (Federated States of)	3	Both	29	15+ years	1990	27605.080770	29580.226920	25829.741340
1983	Number of Smokers	25	Micronesia (Federated States of)	1	Male	29	15+ years	1991	18395.672830	19459.617700	17385.018410
1984	Number of Smokers	25	Micronesia (Federated States of)	2	Female	29	15+ years	1991	9658.519070	11404.994170	7961.453848
...	...	...	...	...	...	...	...	...	...	...	...
20785	Number of Smokers	422	United States Virgin Islands	2	Female	29	15+ years	2018	2308.376511	2820.434508	1871.029388
20786	Number of Smokers	422	United States Virgin Islands	3	Both	29	15+ years	2018	5633.535832	6212.418101	5090.184376
20787	Number of Smokers	422	United States Virgin Islands	1	Male	29	15+ years	2019	3280.527338	3649.862482	2939.996840
20788	Number of Smokers	422	United States Virgin Islands	2	Female	29	15+ years	2019	2282.281664	2813.914814	1831.778372
20789	Number of Smokers	422	United States Virgin Islands	3	Both	29	15+ years	2019	5562.809002	6146.429254	4990.914042

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



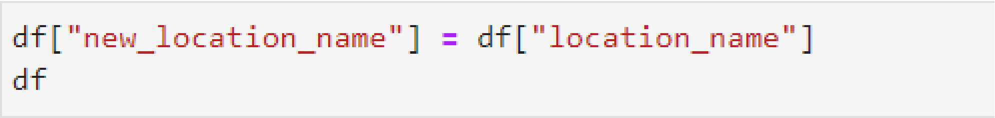
```
df[~df["location_name"].str.contains("States")]
```



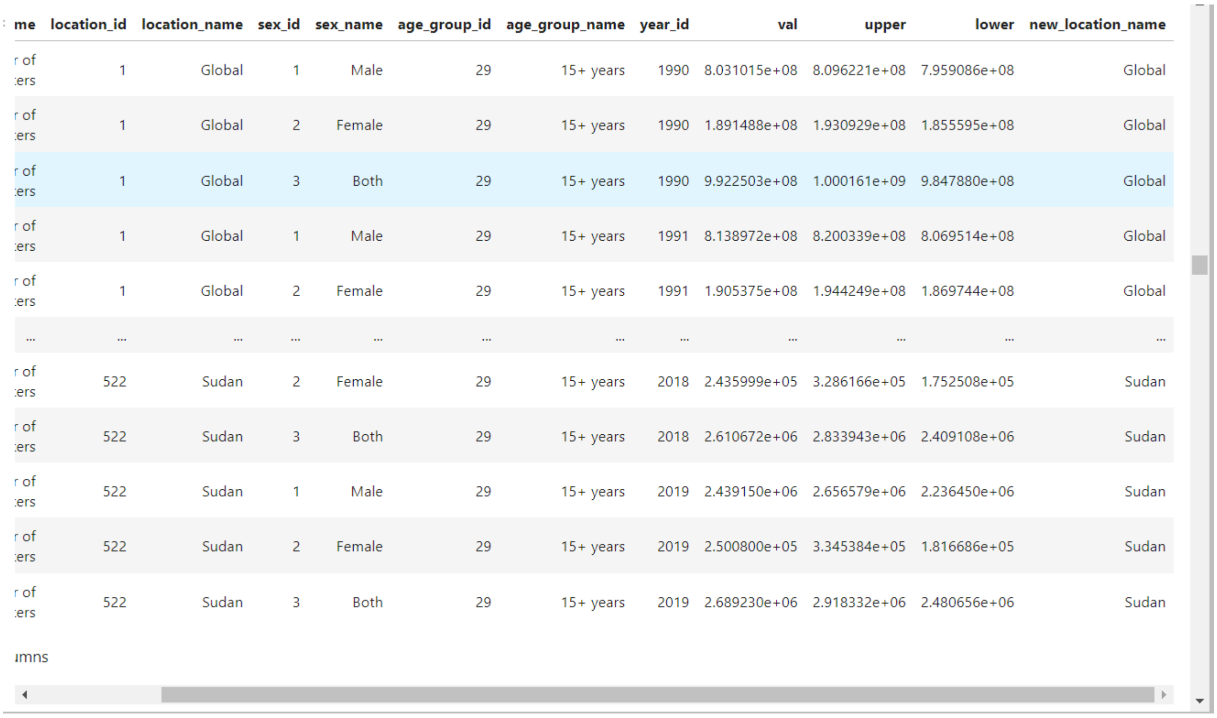
	measure_name	location_id	location_name	sex_id	sex_name	age_group_id	age_group_name	year_id	val	upper	lower
0	Number of Smokers	1	Global	1	Male	29	15+ years	1990	8.031015e+08	8.096221e+08	7.959086e+08
1	Number of Smokers	1	Global	2	Female	29	15+ years	1990	1.891488e+08	1.930929e+08	1.855595e+08
2	Number 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

20700 rows × 11 columns

18 - utwórz kolumnę na podstawie istniejącej



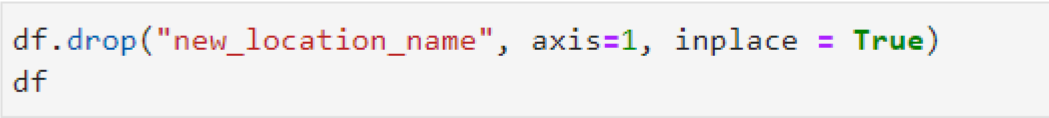
```
df["new_location_name"] = df["location_name"]
df
```



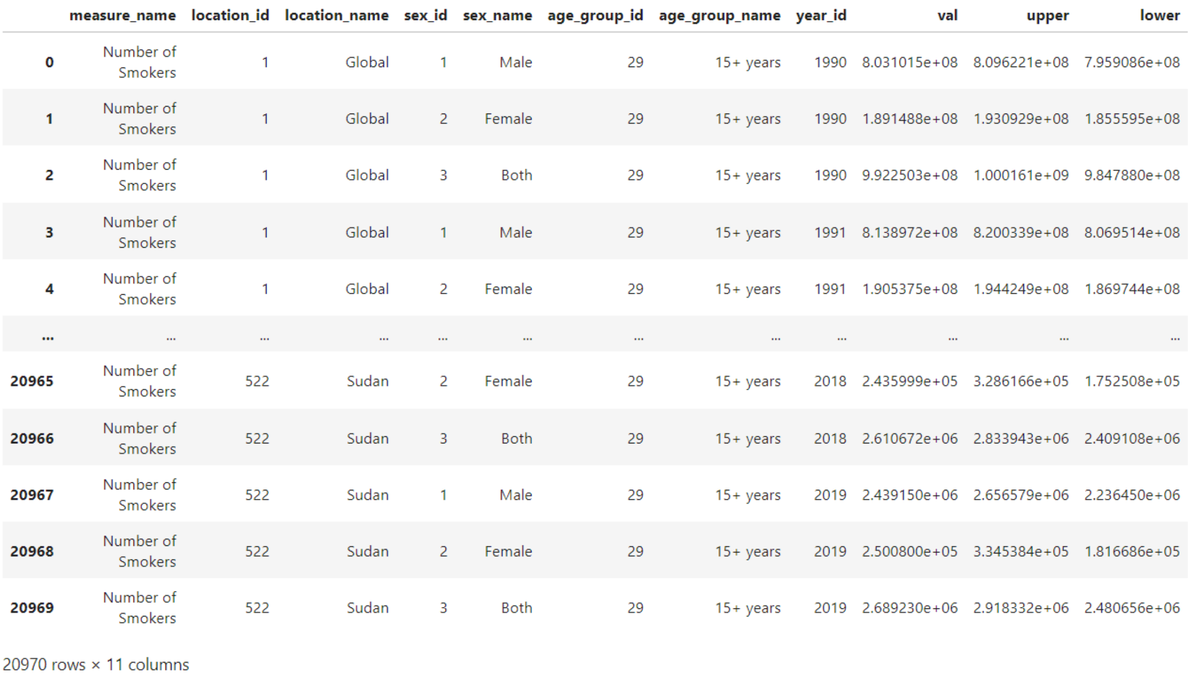
me	location_id	location_name	sex_id	sex_name	age_group_id	age_group_name	year_id	val	upper	lower	new_location_name
r of ers	1	Global	1	Male	29	15+ years	1990	8.031015e+08	8.096221e+08	7.959086e+08	Global
r of ers	1	Global	2	Female	29	15+ years	1990	1.891488e+08	1.930929e+08	1.855595e+08	Global
r of ers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

ımns

19 - usuń kolumnę



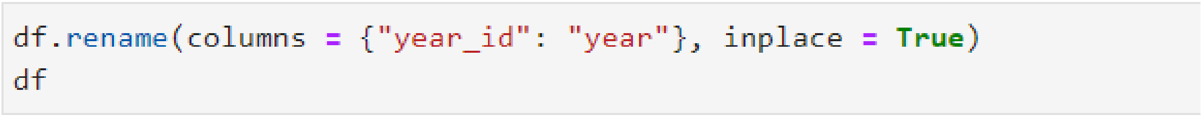
```
df.drop("new_location_name", axis=1, inplace = True)
df
```



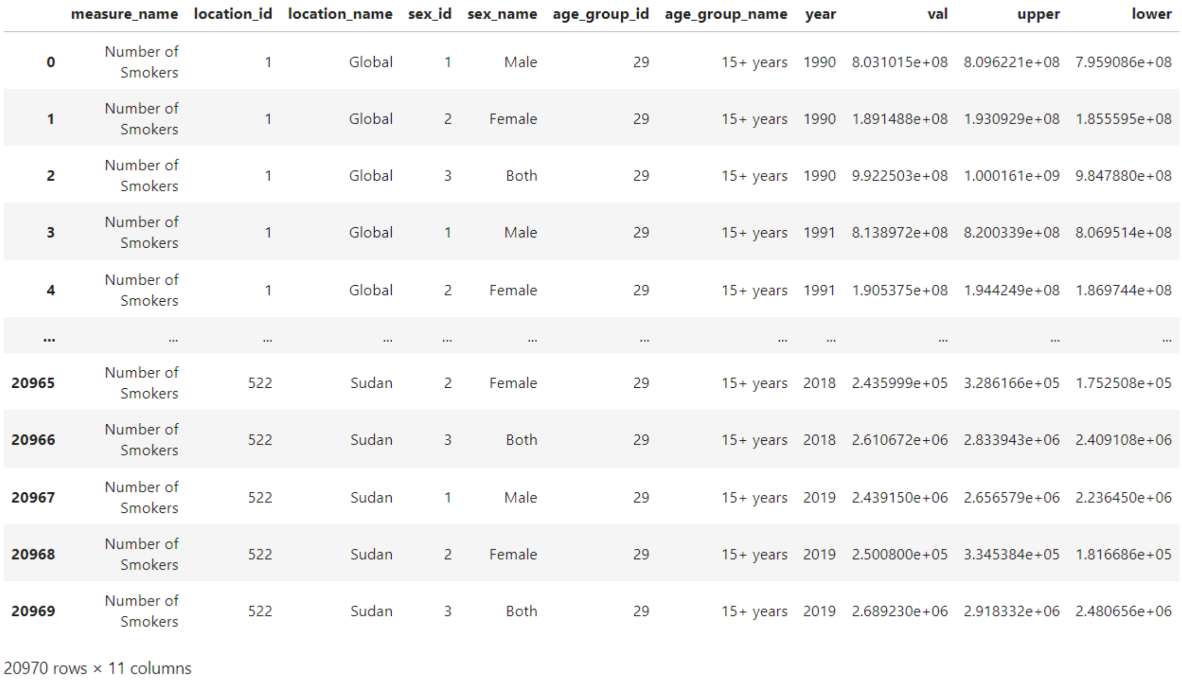
	measure_name	location_id	location_name	sex_id	sex_name	age_group_id	age_group_name	year_id	val	upper	lower
0	Number of Smokers	1	Global	1	Male	29	15+ years	1990	8.031015e+08	8.096221e+08	7.959086e+08
1	Number of Smokers	1	Global	2	Female	29	15+ years	1990	1.891488e+08	1.930929e+08	1.855595e+08
2	Number 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

20 - zmień nazwę kolumny



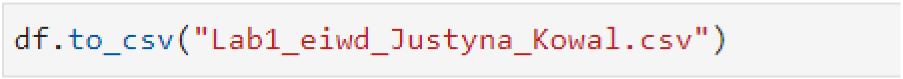
```
df.rename(columns = {"year_id": "year"}, inplace = True)
df
```



	measure_name	location_id	location_name	sex_id	sex_name	age_group_id	age_group_name	year	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

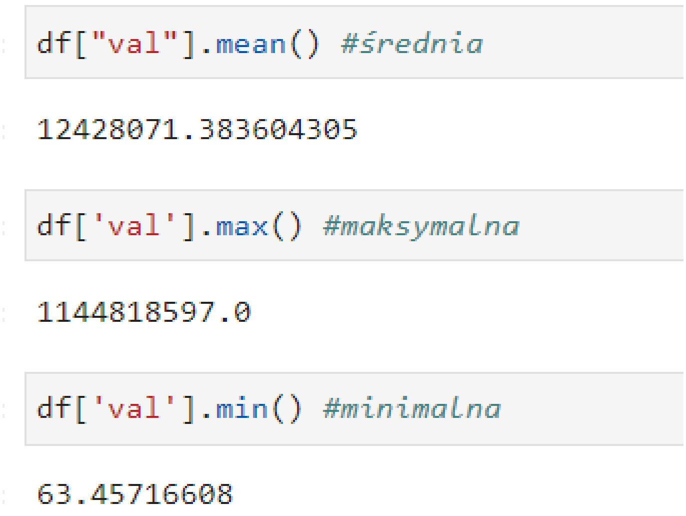
20970 rows × 11 columns

21 - zachowaj ramkę danych jako plik csv na komputerze



```
df.to_csv("Lab1_eiwd_Justyna_Kowal.csv")
```

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



```
df["val"].mean() #średnia
```

```
12428071.383604305
```

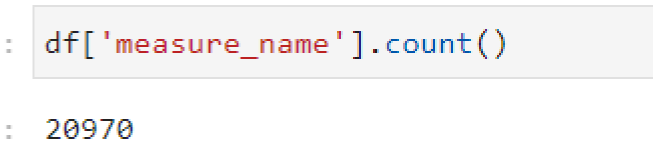
```
df['val'].max() #maksymalna
```

```
1144818597.0
```

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

```
63.45716608
```

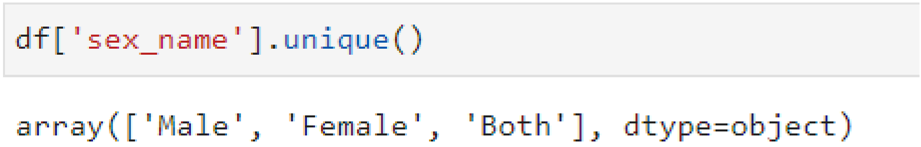
23 - wyświetlić liczbę wierszy



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

```
20970
```

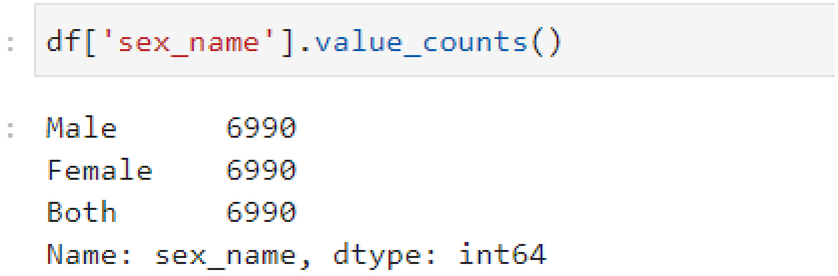
24 - wyświetlić wartości unikatowe w kolumnie



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

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

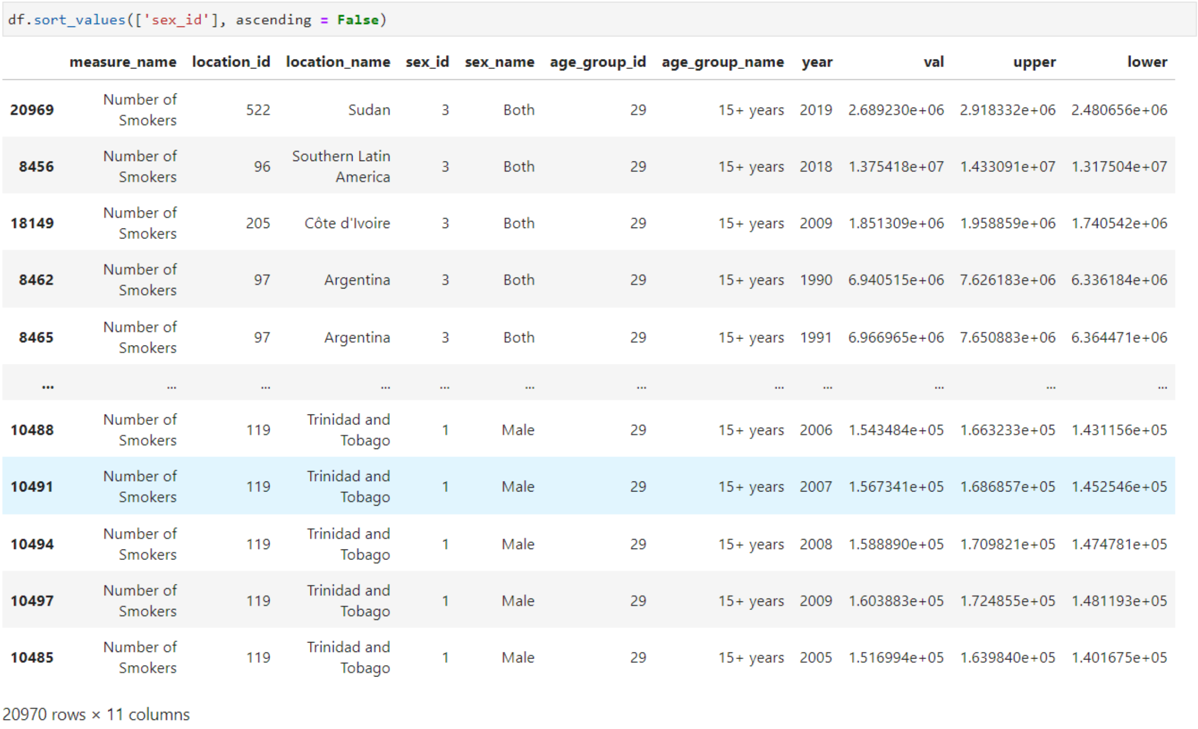
25 - wyświetlić liczby rekordów odpowiadających do wartości



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

```
Male      6990
Female    6990
Both      6990
Name: sex_name, dtype: int64
```

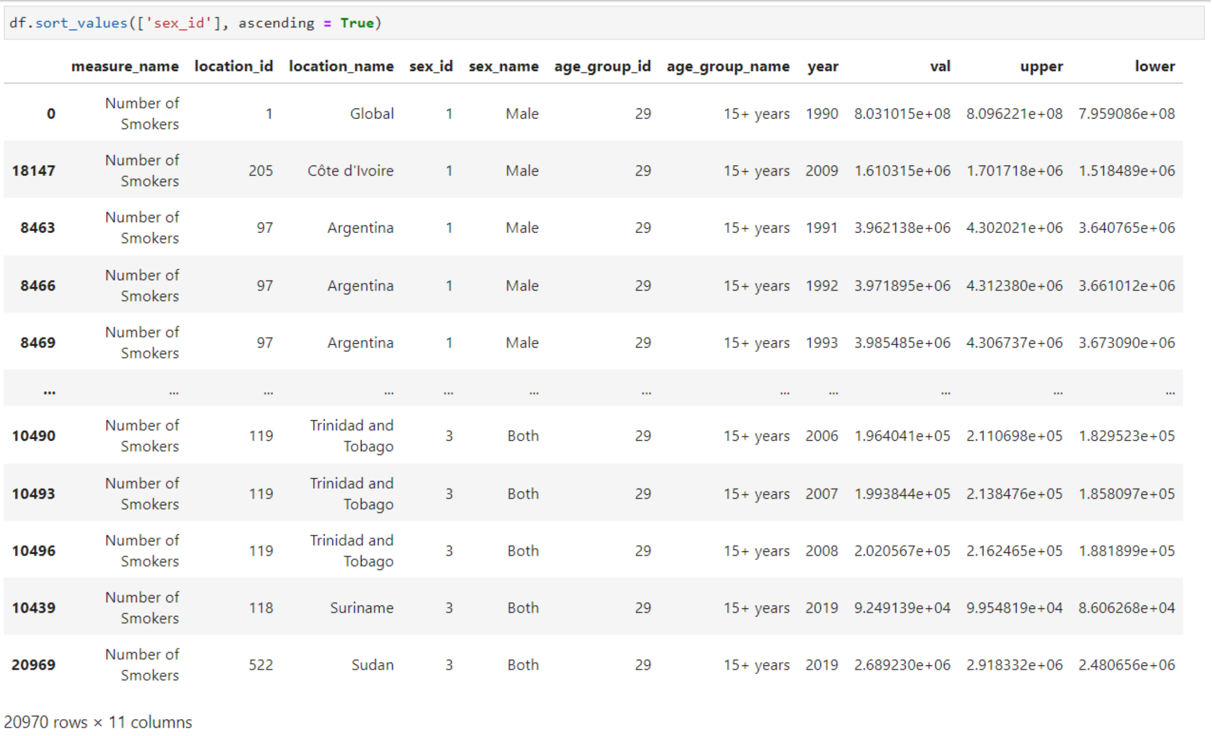
26 - sortowanie wierszy ramki danych według wartości określonej kolumny (malejąco, rosnąco)



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

	measure_name	location_id	location_name	sex_id	sex_name	age_group_id	age_group_name	year	val	upper	lower
20969	Number of Smokers	522	Sudan	3	Both	29	15+ years	2019	2.689230e+06	2.918332e+06	2.480656e+06
8456	Number of Smokers	96	Southern Latin America	3	Both	29	15+ years	2018	1.375418e+07	1.433091e+07	1.317504e+07
18149	Number of Smokers	205	Côte d'Ivoire	3	Both	29	15+ years	2009	1.851309e+06	1.958859e+06	1.740542e+06
8462	Number of Smokers	97	Argentina	3	Both	29	15+ years	1990	6.940515e+06	7.626183e+06	6.336184e+06
8465	Number of Smokers	97	Argentina	3	Both	29	15+ years	1991	6.966965e+06	7.650883e+06	6.364471e+06
...	...	...	...	...	...	...	...	...	...	...	...
10488	Number of Smokers	119	Trinidad and Tobago	1	Male	29	15+ years	2006	1.543484e+05	1.663233e+05	1.431156e+05
10491	Number of Smokers	119	Trinidad and Tobago	1	Male	29	15+ years	2007	1.567341e+05	1.686857e+05	1.452546e+05
10494	Number of Smokers	119	Trinidad and Tobago	1	Male	29	15+ years	2008	1.588890e+05	1.709821e+05	1.474781e+05
10497	Number of Smokers	119	Trinidad and Tobago	1	Male	29	15+ years	2009	1.603883e+05	1.724855e+05	1.481193e+05
10485	Number of Smokers	119	Trinidad and Tobago	1	Male	29	15+ years	2005	1.516994e+05	1.639840e+05	1.401675e+05

20970 rows × 11 columns

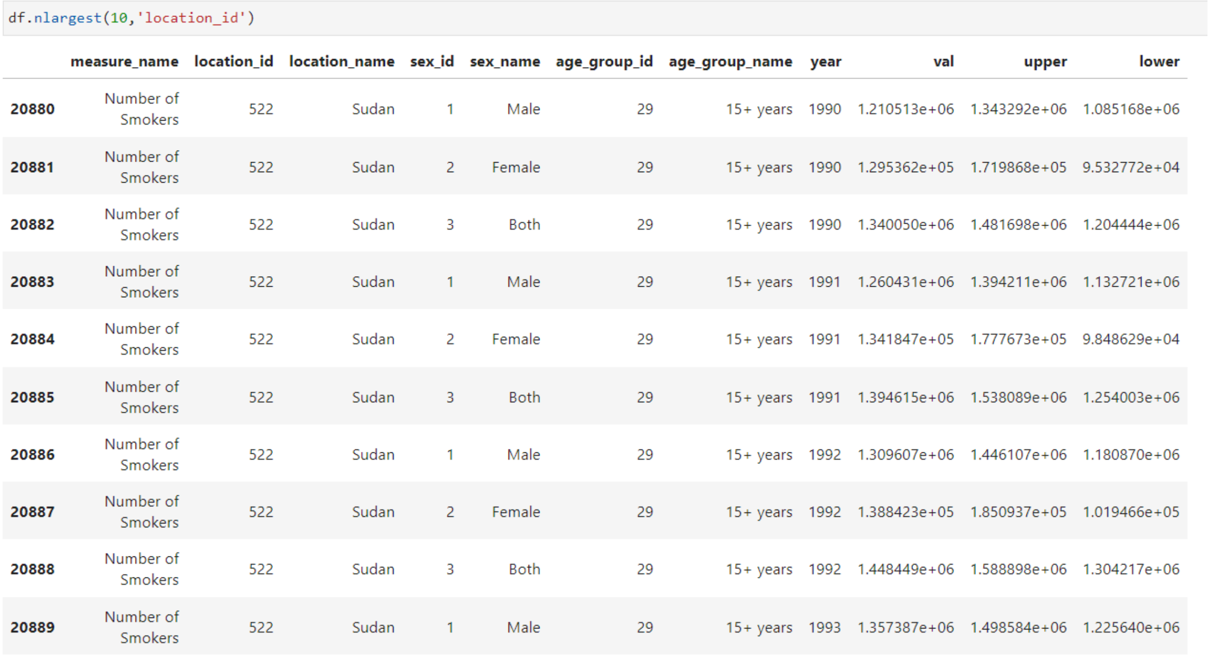


```
df.sort_values(['sex_id'], ascending = True)
```

	measure_name	location_id	location_name	sex_id	sex_name	age_group_id	age_group_name	year	val	upper	lower
0	Number of Smokers	1	Global	1	Male	29	15+ years	1990	8.031015e+08	8.096221e+08	7.959086e+08
18147	Number of Smokers	205	Côte d'Ivoire	1	Male	29	15+ years	2009	1.610315e+06	1.701718e+06	1.518489e+06
8463	Number of Smokers	97	Argentina	1	Male	29	15+ years	1991	3.962138e+06	4.302021e+06	3.640765e+06
8466	Number of Smokers	97	Argentina	1	Male	29	15+ years	1992	3.971895e+06	4.312380e+06	3.661012e+06
8469	Number of Smokers	97	Argentina	1	Male	29	15+ years	1993	3.985485e+06	4.306737e+06	3.673090e+06
...	...	...	...	...	...	...	...	...	...	...	...
10490	Number of Smokers	119	Trinidad and Tobago	3	Both	29	15+ years	2006	1.964041e+05	2.110698e+05	1.829523e+05
10493	Number of Smokers	119	Trinidad and Tobago	3	Both	29	15+ years	2007	1.993844e+05	2.138476e+05	1.858097e+05
10496	Number of Smokers	119	Trinidad and Tobago	3	Both	29	15+ years	2008	2.020567e+05	2.162465e+05	1.881899e+05
10439	Number of Smokers	118	Suriname	3	Both	29	15+ years	2019	9.249139e+04	9.954819e+04	8.606268e+04
20969	Number of Smokers	522	Sudan	3	Both	29	15+ years	2019	2.689230e+06	2.918332e+06	2.480656e+06

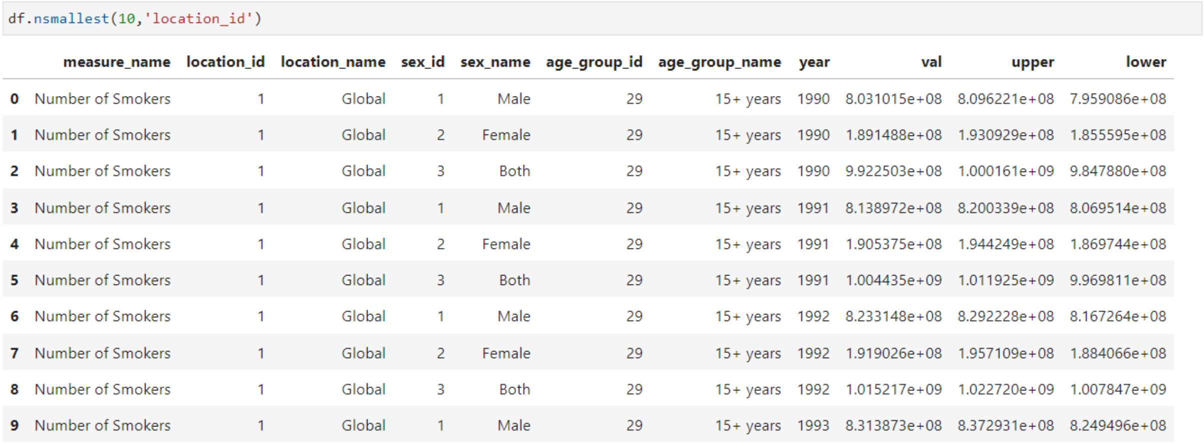
20970 rows × 11 columns

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



```
df.nlargest(10,'location_id')
```

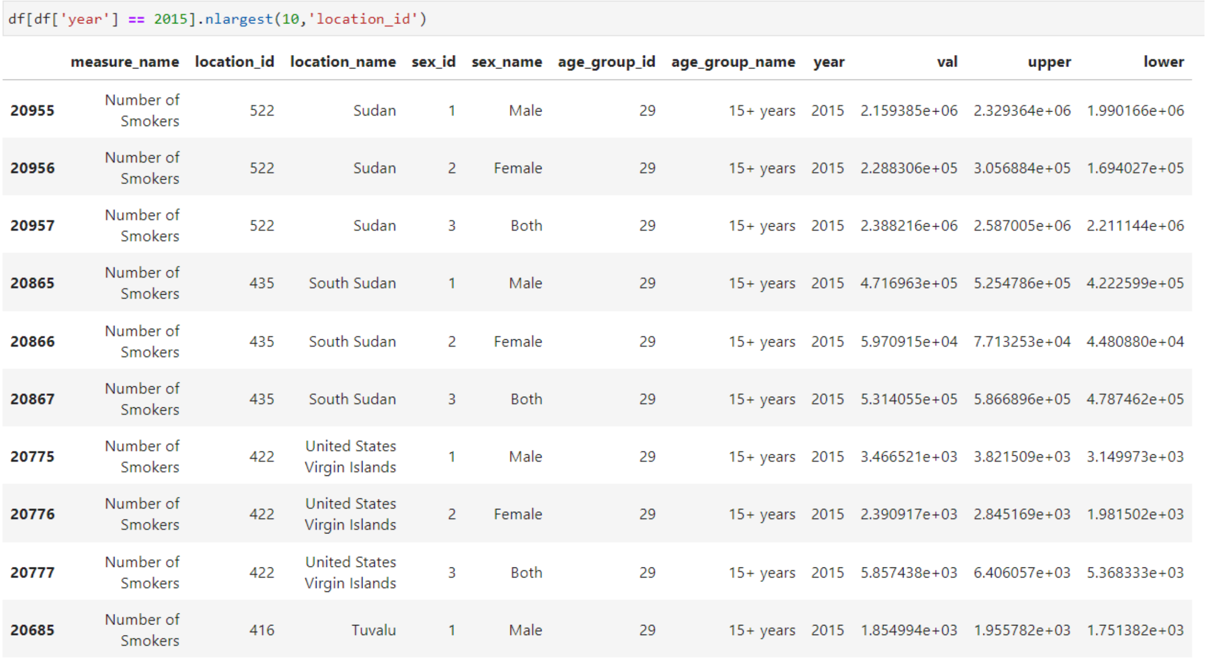
	measure_name	location_id	location_name	sex_id	sex_name	age_group_id	age_group_name	year	val	upper	lower
20880	Number of Smokers	522	Sudan	1	Male	29	15+ years	1990	1.210513e+06	1.343292e+06	1.085168e+06
20881	Number of Smokers	522	Sudan	2	Female	29	15+ years	1990	1.295362e+05	1.719868e+05	9.532772e+04
20882	Number of Smokers	522	Sudan	3	Both	29	15+ years	1990	1.340050e+06	1.481698e+06	1.204444e+06
20883	Number of Smokers	522	Sudan	1	Male	29	15+ years	1991	1.260431e+06	1.394211e+06	1.132721e+06
20884	Number of Smokers	522	Sudan	2	Female	29	15+ years	1991	1.341847e+05	1.777673e+05	9.848629e+04
20885	Number of Smokers	522	Sudan	3	Both	29	15+ years	1991	1.394615e+06	1.538089e+06	1.254003e+06
20886	Number of Smokers	522	Sudan	1	Male	29	15+ years	1992	1.309607e+06	1.446107e+06	1.180870e+06
20887	Number of Smokers	522	Sudan	2	Female	29	15+ years	1992	1.388423e+05	1.850937e+05	1.019466e+05
20888	Number of Smokers	522	Sudan	3	Both	29	15+ years	1992	1.448449e+06	1.588898e+06	1.304217e+06
20889	Number of Smokers	522	Sudan	1	Male	29	15+ years	1993	1.357387e+06	1.498584e+06	1.225640e+06



```
df.nsmallest(10,'location_id')
```

	measure_name	location_id	location_name	sex_id	sex_name	age_group_id	age_group_name	year	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
5	Number of Smokers	1	Global	3	Both	29	15+ years	1991	1.004435e+09	1.011925e+09	9.969811e+08
6	Number of Smokers	1	Global	1	Male	29	15+ years	1992	8.233148e+08	8.292228e+08	8.167264e+08
7	Number of Smokers	1	Global	2	Female	29	15+ years	1992	1.919026e+08	1.957109e+08	1.884066e+08
8	Number of Smokers	1	Global	3	Both	29	15+ years	1992	1.015217e+09	1.022720e+09	1.007847e+09
9	Number of Smokers	1	Global	1	Male	29	15+ years	1993	8.313873e+08	8.372931e+08	8.249496e+08

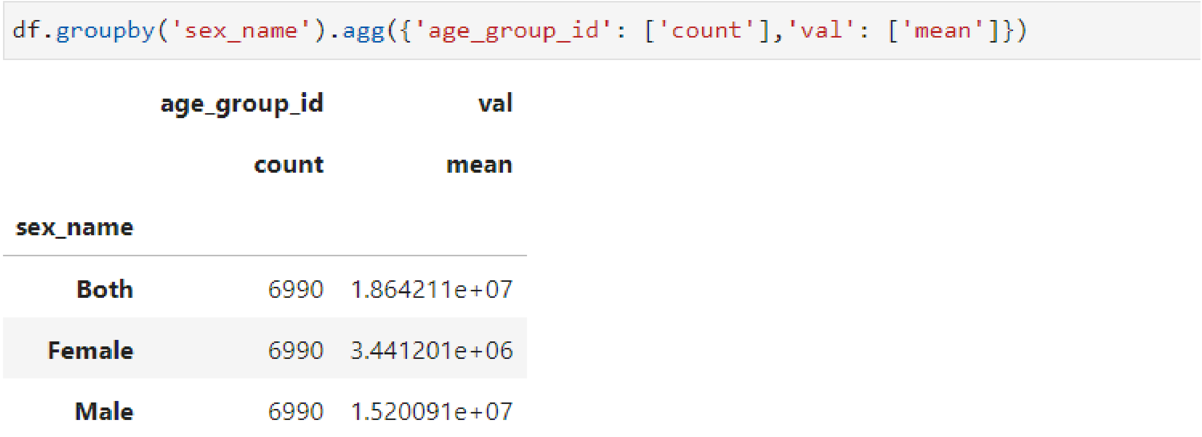
28 - wyświetlić wierszy dla 10 największych wartości określonej kolumny pod warunkiem określonych wartości innej kolumny



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

	measure_name	location_id	location_name	sex_id	sex_name	age_group_id	age_group_name	year	val	upper	lower
20955	Number of Smokers	522	Sudan	1	Male	29	15+ years	2015	2.159385e+06	2.329364e+06	1.990166e+06
20956	Number of Smokers	522	Sudan	2	Female	29	15+ years	2015	2.288306e+05	3.056884e+05	1.694027e+05
20957	Number of Smokers	522	Sudan	3	Both	29	15+ years	2015	2.388216e+06	2.587005e+06	2.211144e+06
20865	Number of Smokers	435	South Sudan	1	Male	29	15+ years	2015	4.716963e+05	5.254786e+05	4.222599e+05
20866	Number of Smokers	435	South Sudan	2	Female	29	15+ years	2015	5.970915e+04	7.713253e+04	4.480880e+04
20867	Number of Smokers	435	South Sudan	3	Both	29	15+ years	2015	5.314055e+05	5.866896e+05	4.787462e+05
20775	Number of Smokers	422	United States Virgin Islands	1	Male	29	15+ years	2015	3.466521e+03	3.821509e+03	3.149973e+03
20776	Number of Smokers	422	United States Virgin Islands	2	Female	29	15+ years	2015	2.390917e+03	2.845169e+03	1.981502e+03
20777	Number of Smokers	422	United States Virgin Islands	3	Both	29	15+ years	2015	5.857438e+03	6.406057e+03	5.368333e+03
20685	Number of Smokers	416	Tuvalu	1	Male	29	15+ years	2015	1.854994e+03	1.955782e+03	1.751382e+03

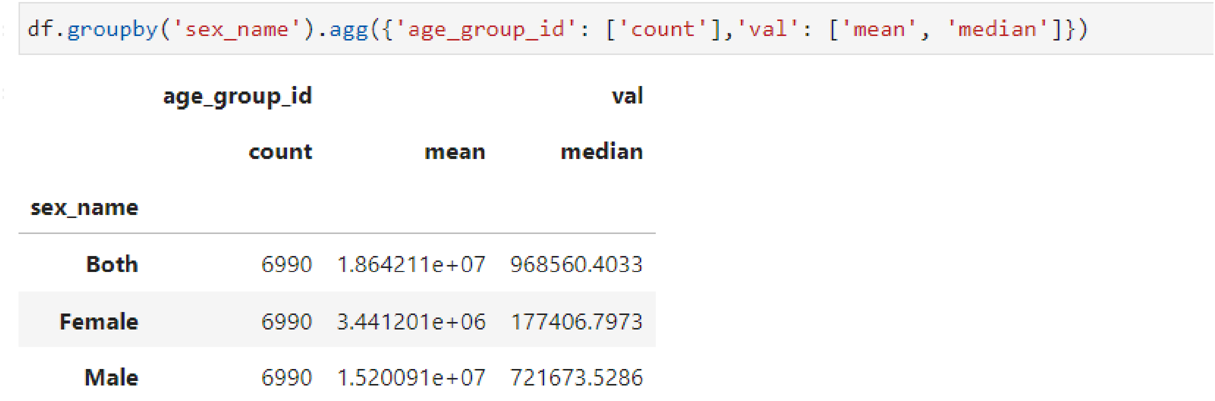
29 - grupowanie wierszy według wartości kolumny kategoryzowanej, potem - uśrednienie wartości wszystkich kolumn w grupie – MultiIndex



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

	age_group_id	val
	count	mean
sex_name		
Both	6990	1.864211e+07
Female	6990	3.441201e+06
Male	6990	1.520091e+07

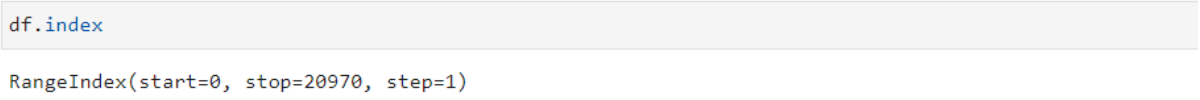
30 - grupowanie wierszy według wartości kolumny kategoryzowanej, potem - uśrednienie wartości dla pewnych kolumn, liczba wartości i mediana dla pozostałych kolumn w grupach



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

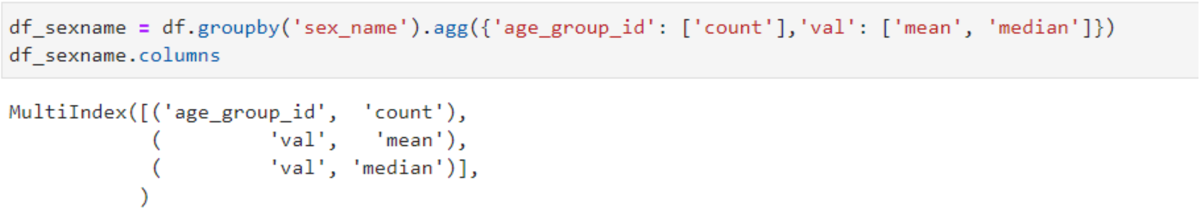
	age_group_id	val	
	count	mean	median
sex_name			
Both	6990	1.864211e+07	968560.4033
Female	6990	3.441201e+06	177406.7973
Male	6990	1.520091e+07	721673.5286

31 - wyświetlić nazwy kolumn indeksu złożonego



```
df.index
```

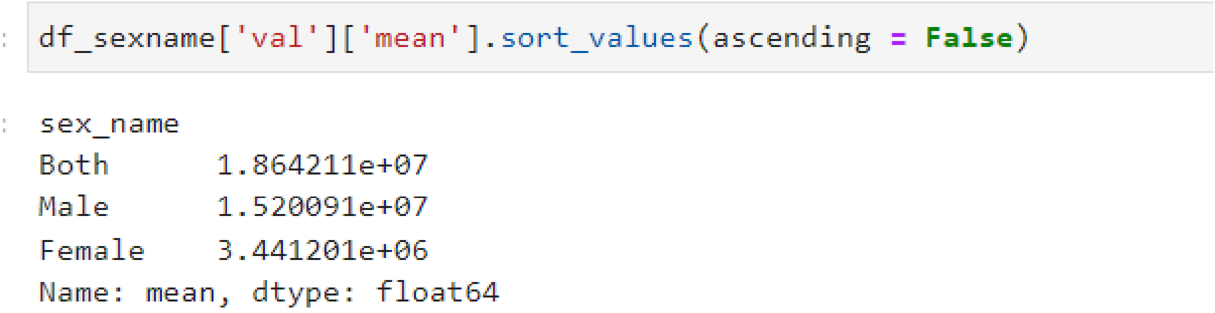
```
RangeIndex(start=0, stop=20970, step=1)
```



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

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

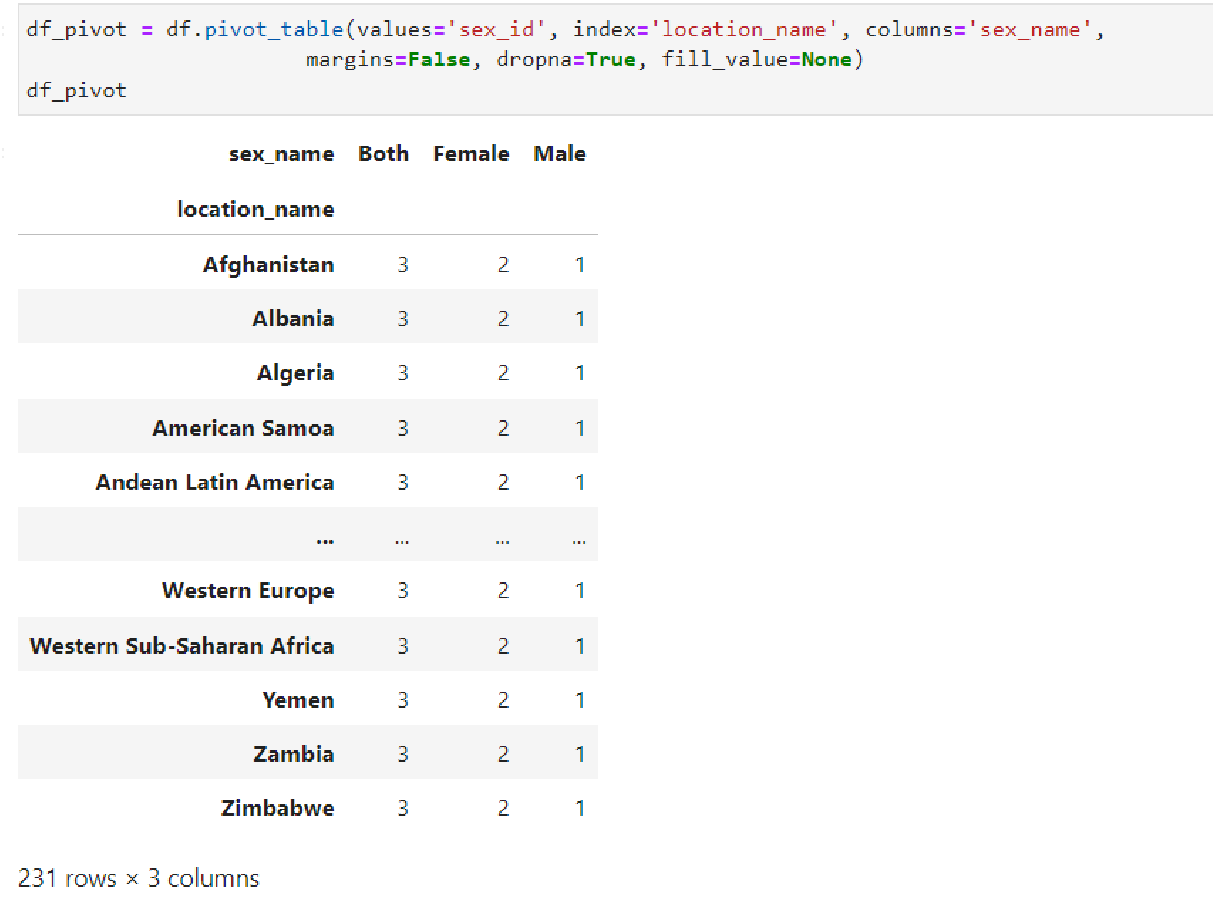
32 - sortować kolumnę indeksu złożonego



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

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

33 - stworzyć tabele przystawna (pivot table) na podstawie ramki danych

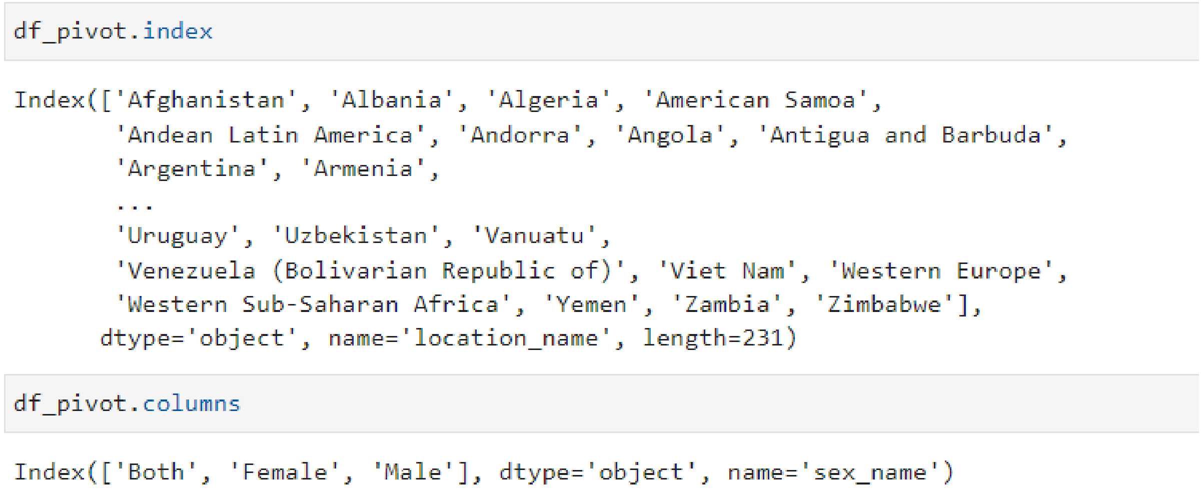


```
df_pivot = df.pivot_table(values='sex_id', index='location_name', columns='sex_name',
                          margins=False, dropna=True, fill_value=None)
df_pivot
```

sex_name	Both	Female	Male
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 × 3 columns

34 - wyświetlić indeksy i kolumny tabeli przystawnej



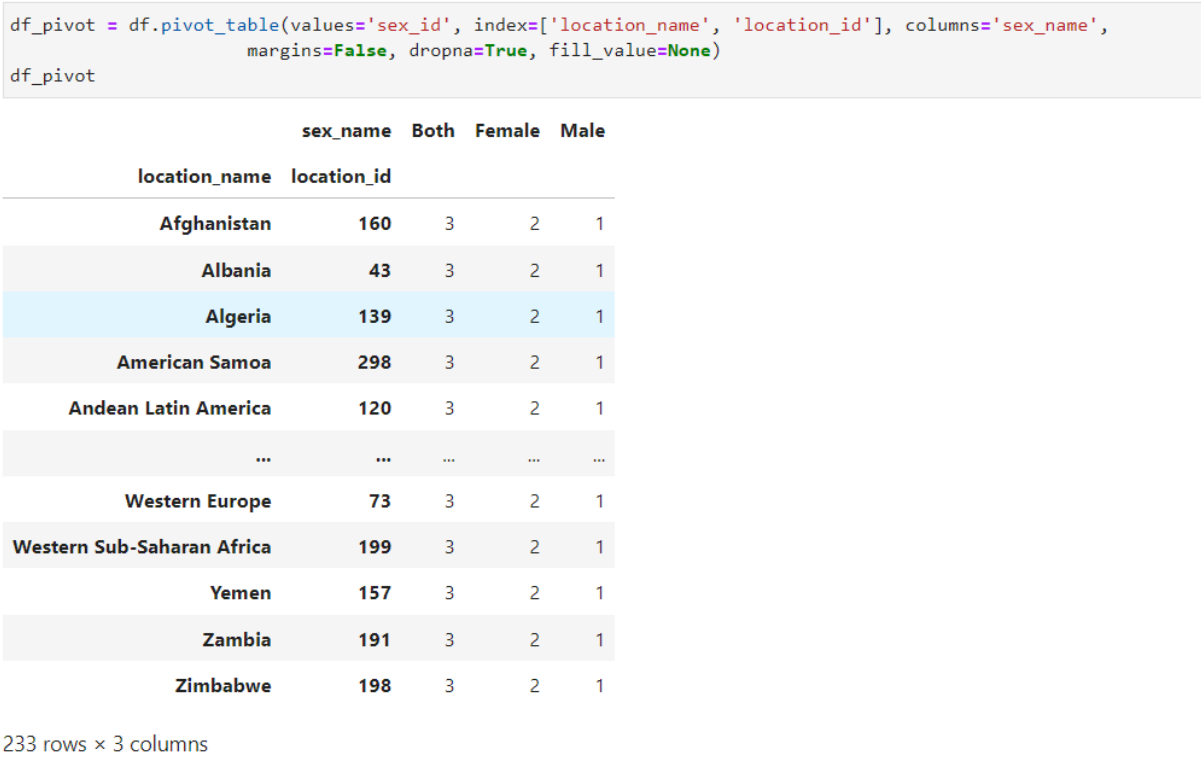
```
df_pivot.index
```

```
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)
```

```
df_pivot.columns
```

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

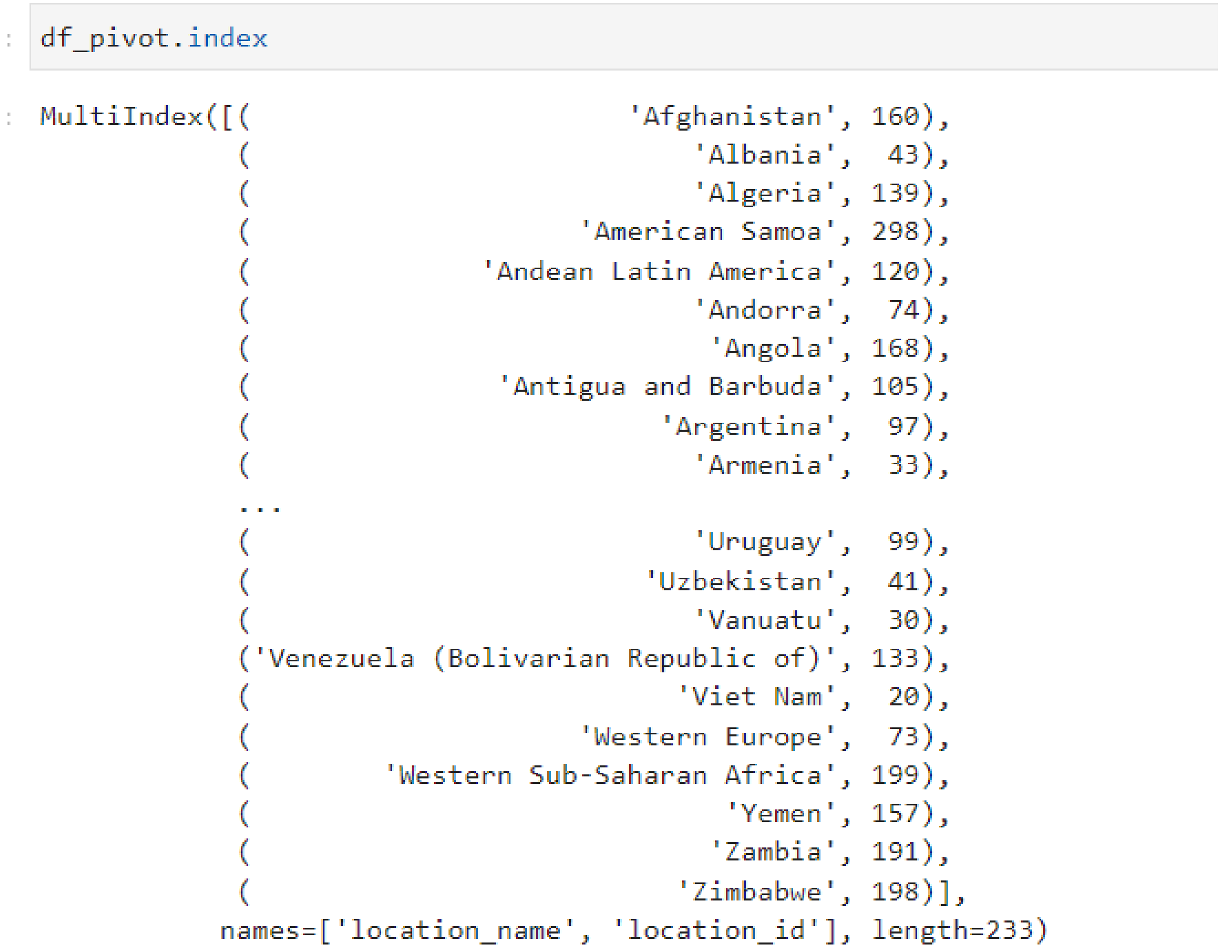
35 - utwórz indeks złożony tabeli przystawnej i wyświetl go



```
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
```

	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

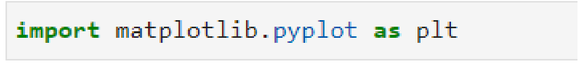
233 rows × 3 columns



```
df_pivot.index
```

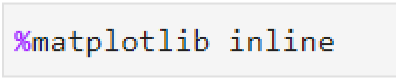
```
MultiIndex([(                       'Afghanistan', 160),
            (                           'Albania',  43),
            (                           'Algeria', 139),
            (                    'American Samoa', 298),
            (              'Andean Latin America', 120),
            (                           '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



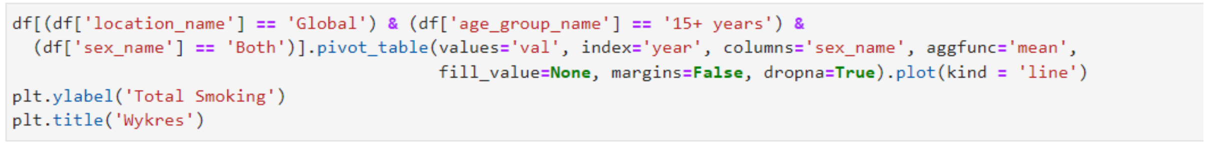
```
import matplotlib.pyplot as plt
```

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



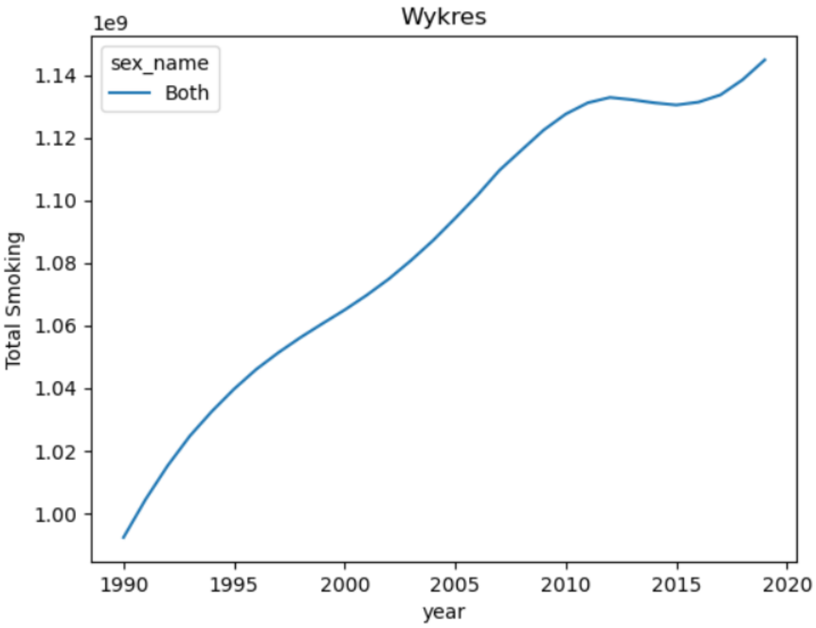
```
%matplotlib inline
```

38 - wyświetlić wykres na podstawie tabeli przystawnej

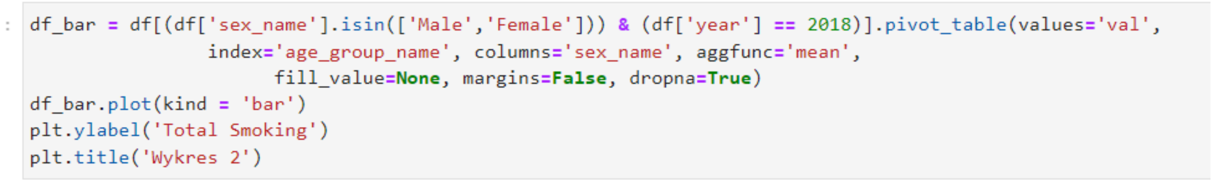


```
df[(df['location_name'] == 'Global') & (df['age_group_name'] == '15+ years') &
  (df['sex_name'] == 'Both')].pivot_table(values='val', index='year', columns='sex_name', aggfunc='mean',
                                          fill_value=None, margins=False, dropna=True).plot(kind = 'line')
plt.ylabel('Total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

```
Text(0.5, 1.0, 'Wykres')
```

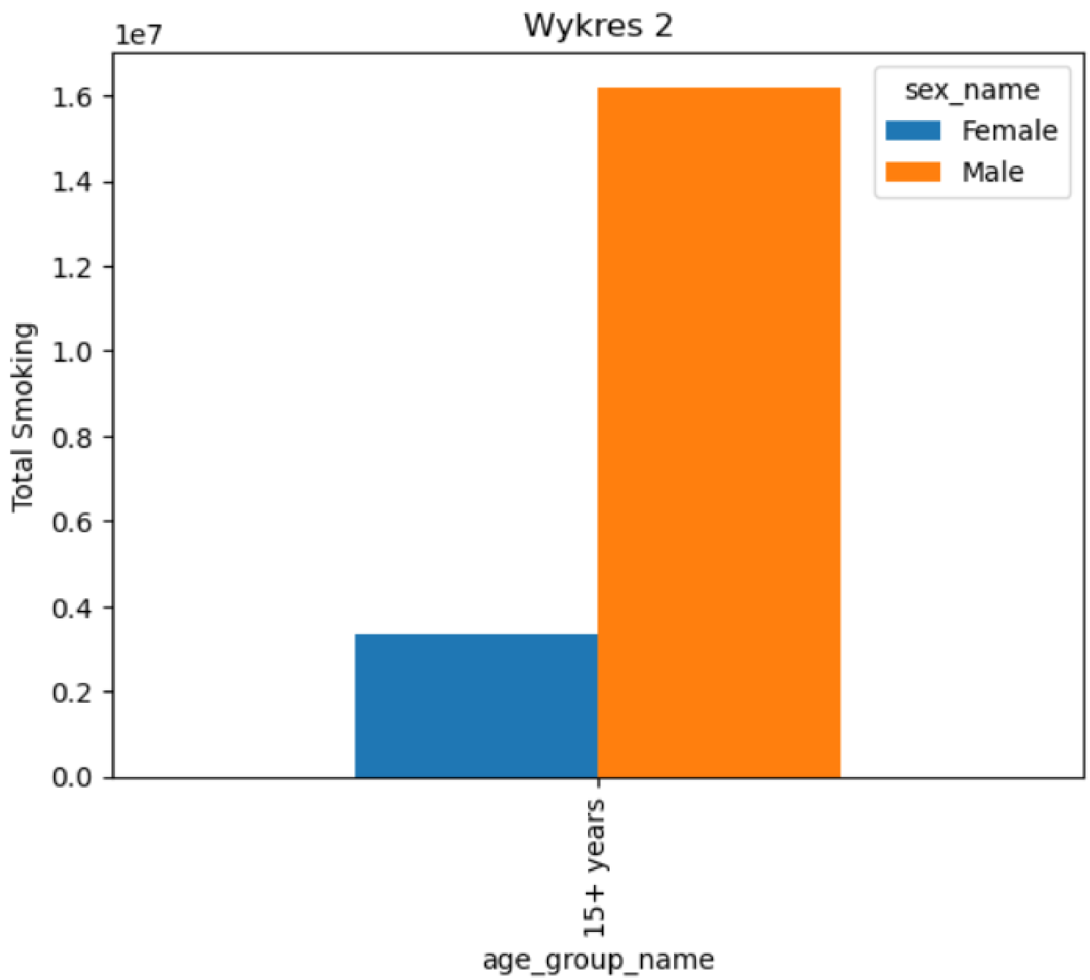


39 - narysować histogram na podstawie wartości kolumny

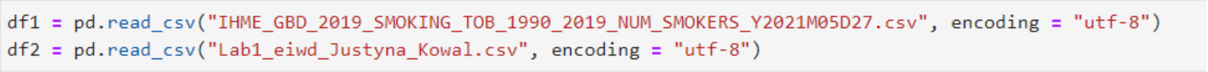


```
df_bar = df[(df['sex_name'].isin(['Male','Female'])) & (df['year'] == 2018)].pivot_table(values='val',
                 index='age_group_name', columns='sex_name', aggfunc='mean',
                        fill_value=None, margins=False, dropna=True)
df_bar.plot(kind = 'bar')
plt.ylabel('Total Smoking')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

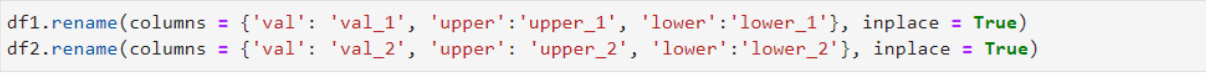
```
Text(0.5, 1.0, 'Wykres 2')
```



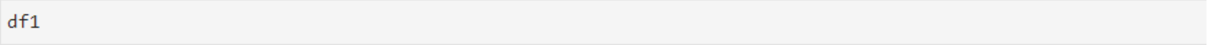
40 - przedstawić sposoby łączenia ramek danych za pomocą metod merge i concat



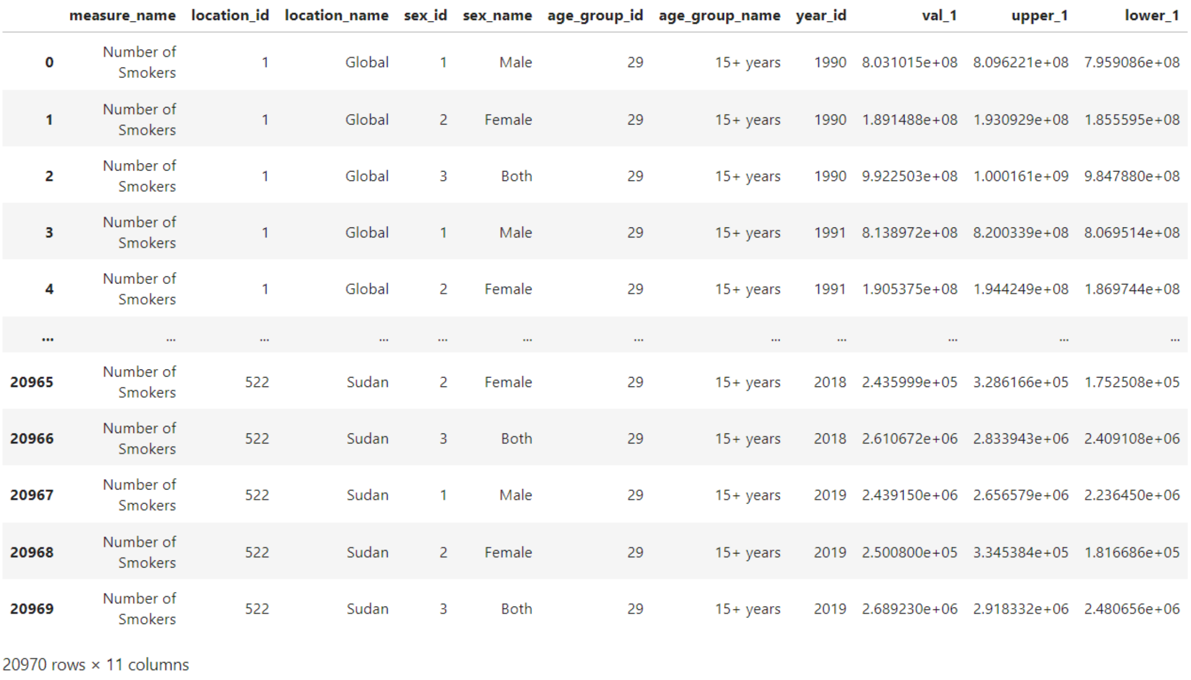
```
df1 = pd.read_csv("IHME_GBD_2019_SMOKING_TOB_1990_2019_NUM_SMOKERS_Y2021M05D27.csv", encoding = "utf-8")
df2 = pd.read_csv("Lab1_eiwd_Justyna_Kowal.csv", encoding = "utf-8")
```



```
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)
```

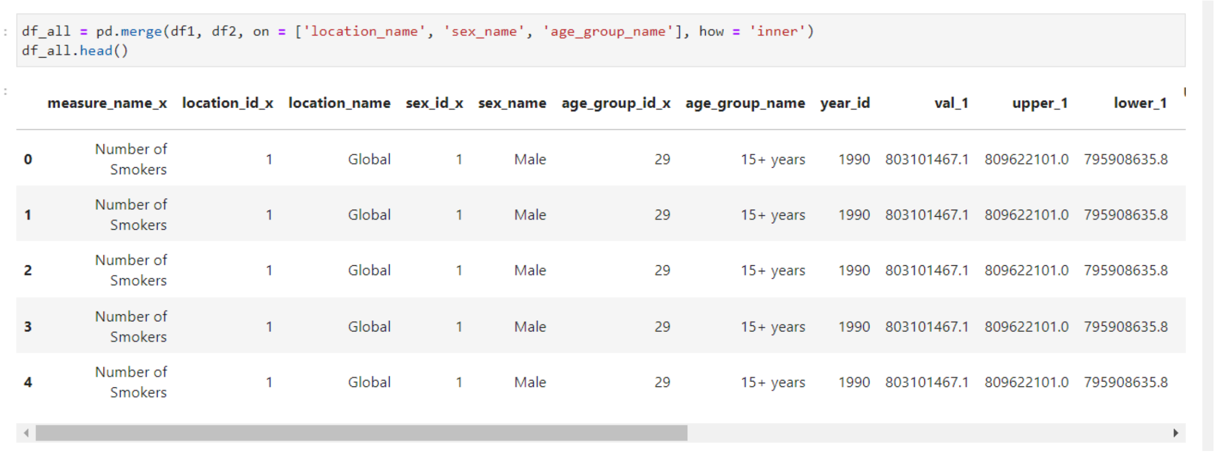


```
df1
```



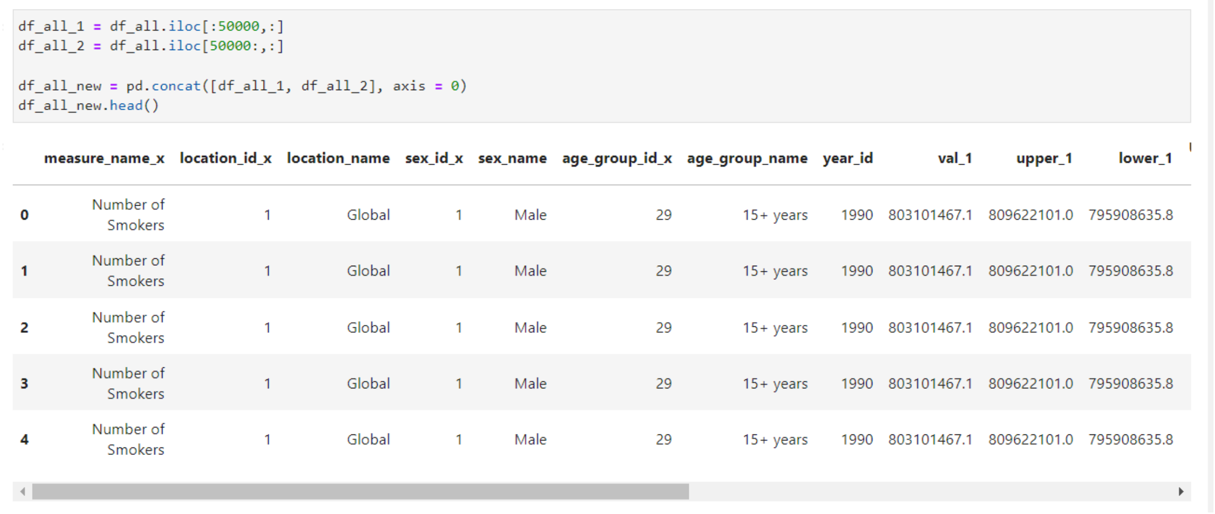
	measure_name	location_id	location_name	sex_id	sex_name	age_group_id	age_group_name	year_id	val_1	upper_1	lower_1
0	Number of Smokers	1	Global	1	Male	29	15+ years	1990	8.031015e+08	8.096221e+08	7.959086e+08
1	Number of Smokers	1	Global	2	Female	29	15+ years	1990	1.891488e+08	1.930929e+08	1.855595e+08
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



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

	measure_name_x	location_id_x	location_name	sex_id_x	sex_name	age_group_id_x	age_group_name	year_id	val_1	upper_1	lower_1
0	Number of Smokers	1	Global	1	Male	29	15+ years	1990	803101467.1	809622101.0	795908635.8
1	Number of Smokers	1	Global	1	Male	29	15+ years	1990	803101467.1	809622101.0	795908635.8
2	Number of Smokers	1	Global	1	Male	29	15+ years	1990	803101467.1	809622101.0	795908635.8
3	Number of Smokers	1	Global	1	Male	29	15+ years	1990	803101467.1	809622101.0	795908635.8
4	Number of Smokers	1	Global	1	Male	29	15+ years	1990	803101467.1	809622101.0	795908635.8

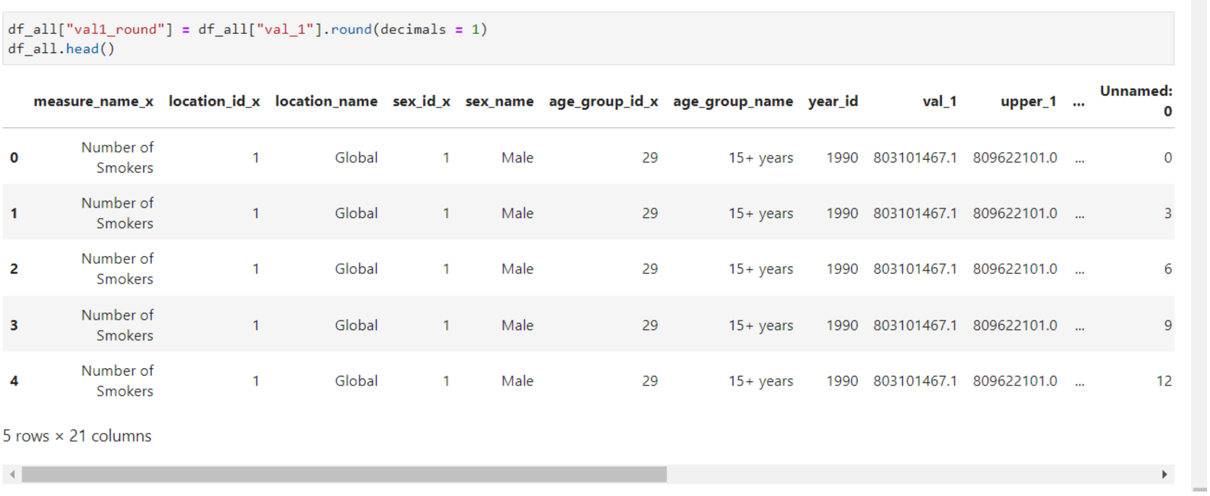


```
df_all_1 = df_all.iloc[:50000,:]
df_all_2 = df_all.iloc[50000:,:]

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

	measure_name_x	location_id_x	location_name	sex_id_x	sex_name	age_group_id_x	age_group_name	year_id	val_1	upper_1	lower_1
0	Number of Smokers	1	Global	1	Male	29	15+ years	1990	803101467.1	809622101.0	795908635.8
1	Number of Smokers	1	Global	1	Male	29	15+ years	1990	803101467.1	809622101.0	795908635.8
2	Number of Smokers	1	Global	1	Male	29	15+ years	1990	803101467.1	809622101.0	795908635.8
3	Number of Smokers	1	Global	1	Male	29	15+ years	1990	803101467.1	809622101.0	795908635.8
4	Number of Smokers	1	Global	1	Male	29	15+ years	1990	803101467.1	809622101.0	795908635.8

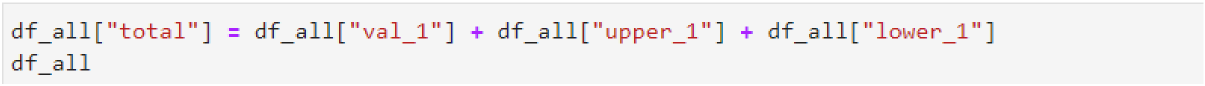
41 - pokazać dodawanie nowych kolumn za pomocą operacji matematycznych



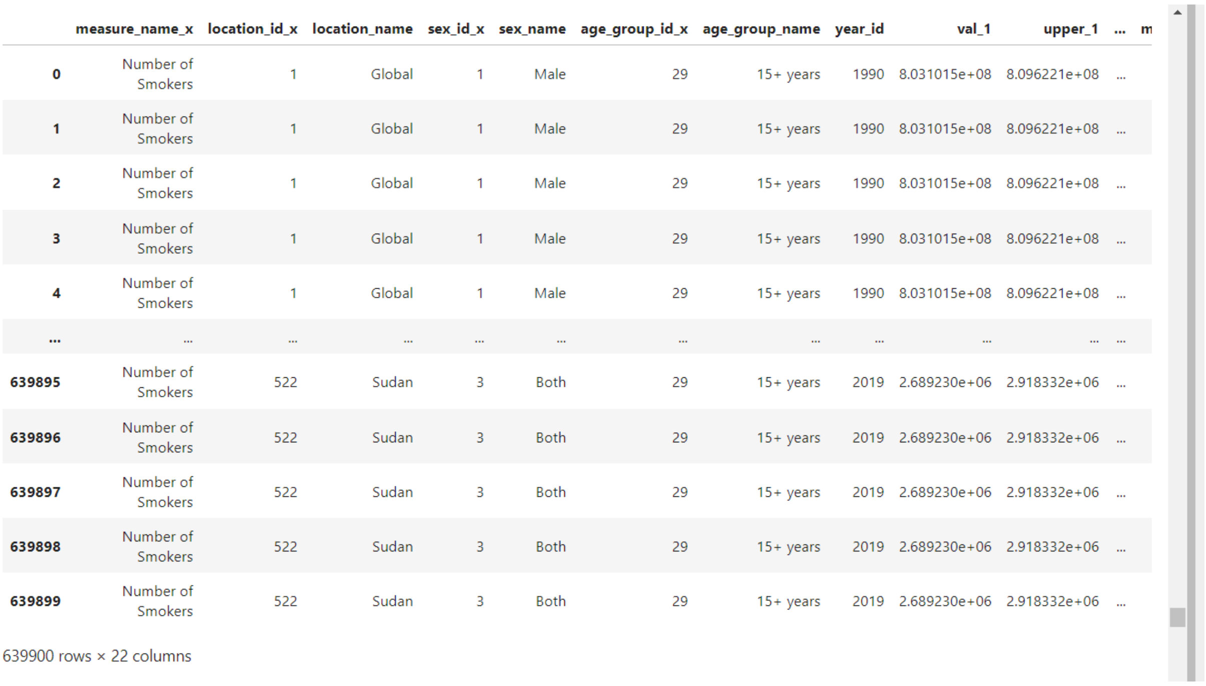
```
df_all["val1_round"] = df_all["val_1"].round(decimals = 1)
df_all.head()
```

	measure_name_x	location_id_x	location_name	sex_id_x	sex_name	age_group_id_x	age_group_name	year_id	val_1	upper_1	...	Unnamed: 0
0	Number of Smokers	1	Global	1	Male	29	15+ years	1990	803101467.1	809622101.0	...	0
1	Number of Smokers	1	Global	1	Male	29	15+ years	1990	803101467.1	809622101.0	...	3
2	Number of Smokers	1	Global	1	Male	29	15+ years	1990	803101467.1	809622101.0	...	6
3	Number of Smokers	1	Global	1	Male	29	15+ years	1990	803101467.1	809622101.0	...	9
4	Number of Smokers	1	Global	1	Male	29	15+ years	1990	803101467.1	809622101.0	...	12

5 rows × 21 columns



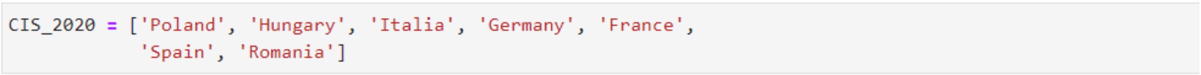
```
df_all["total"] = df_all["val_1"] + df_all["upper_1"] + df_all["lower_1"]
df_all
```



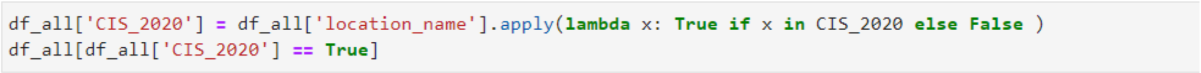
	measure_name_x	location_id_x	location_name	sex_id_x	sex_name	age_group_id_x	age_group_name	year_id	val_1	upper_1	...	m
0	Number of Smokers	1	Global	1	Male	29	15+ years	1990	8.031015e+08	8.096221e+08	...	
1	Number of Smokers	1	Global	1	Male	29	15+ years	1990	8.031015e+08	8.096221e+08	...	
2	Number of Smokers	1	Global	1	Male	29	15+ years	1990	8.031015e+08	8.096221e+08	...	
3	Number of Smokers	1	Global	1	Male	29	15+ years	1990	8.031015e+08	8.096221e+08	...	
4	Number of Smokers	1	Global	1	Male	29	15+ years	1990	8.031015e+08	8.096221e+08	...	
...	...	...	...	...	...	...	...	...	...	...	...	...
639895	Number of Smokers	522	Sudan	3	Both	29	15+ years	2019	2.689230e+06	2.918332e+06	...	
639896	Number of Smokers	522	Sudan	3	Both	29	15+ years	2019	2.689230e+06	2.918332e+06	...	
639897	Number of Smokers	522	Sudan	3	Both	29	15+ years	2019	2.689230e+06	2.918332e+06	...	
639898	Number of Smokers	522	Sudan	3	Both	29	15+ years	2019	2.689230e+06	2.918332e+06	...	
639899	Number of Smokers	522	Sudan	3	Both	29	15+ years	2019	2.689230e+06	2.918332e+06	...	

639900 rows × 22 columns

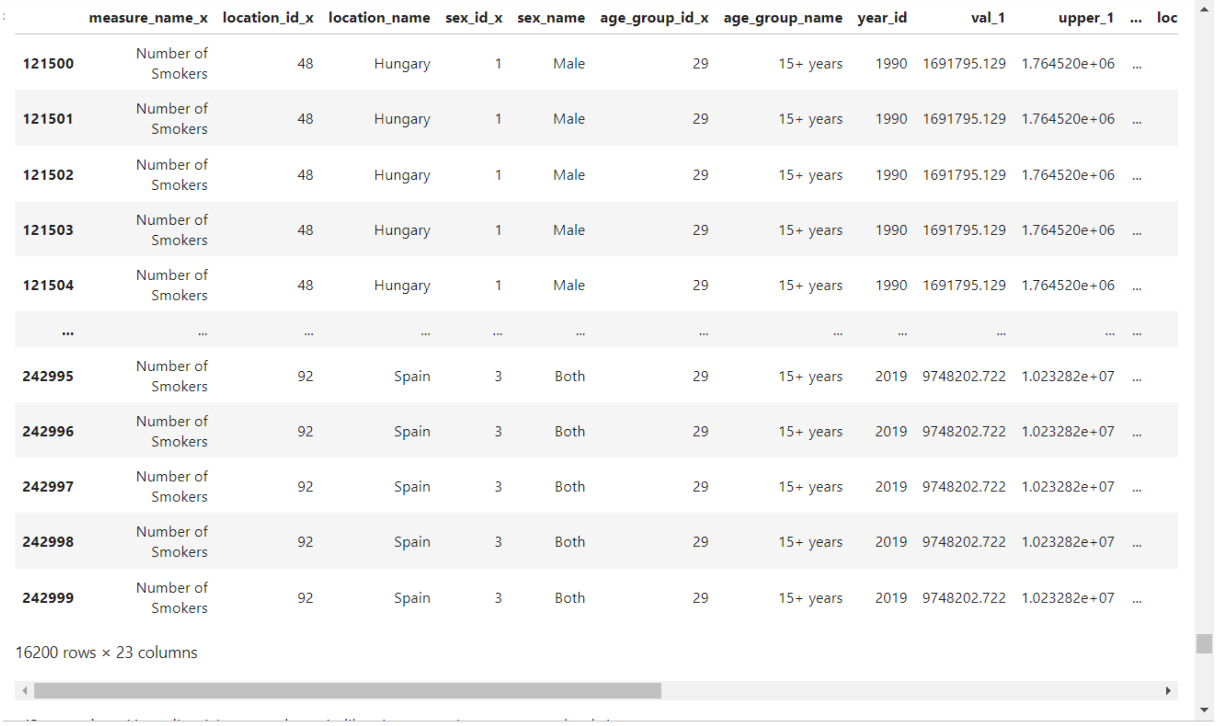
42 - przedstawić na przykładzie dodawanie nowych kolumn z pomocą funkcji lambda



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



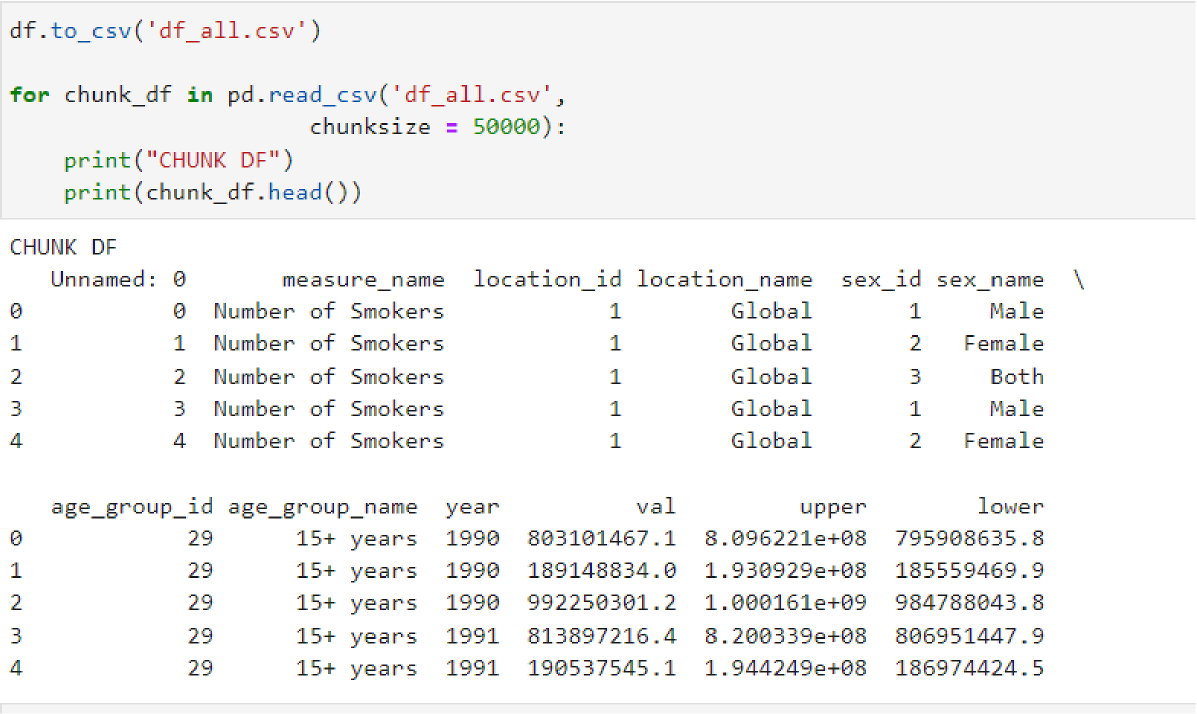
```
df_all['CIS_2020'] = df_all['location_name'].apply(lambda x: True if x in CIS_2020 else False )
df_all[df_all['CIS_2020'] == True]
```



	measure_name_x	location_id_x	location_name	sex_id_x	sex_name	age_group_id_x	age_group_name	year_id	val_1	upper_1	...	loc
121500	Number of Smokers	48	Hungary	1	Male	29	15+ years	1990	1691795.129	1.764520e+06	...	
121501	Number of Smokers	48	Hungary	1	Male	29	15+ years	1990	1691795.129	1.764520e+06	...	
121502	Number of Smokers	48	Hungary	1	Male	29	15+ years	1990	1691795.129	1.764520e+06	...	
121503	Number of Smokers	48	Hungary	1	Male	29	15+ years	1990	1691795.129	1.764520e+06	...	
121504	Number of Smokers	48	Hungary	1	Male	29	15+ years	1990	1691795.129	1.764520e+06	...	
...	...	...	...	...	...	...	...	...	...	...	...	
242995	Number of Smokers	92	Spain	3	Both	29	15+ years	2019	9748202.722	1.023282e+07	...	
242996	Number of Smokers	92	Spain	3	Both	29	15+ years	2019	9748202.722	1.023282e+07	...	
242997	Number of Smokers	92	Spain	3	Both	29	15+ years	2019	9748202.722	1.023282e+07	...	
242998	Number of Smokers	92	Spain	3	Both	29	15+ years	2019	9748202.722	1.023282e+07	...	
242999	Number of Smokers	92	Spain	3	Both	29	15+ years	2019	9748202.722	1.023282e+07	...	

16200 rows × 23 columns

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

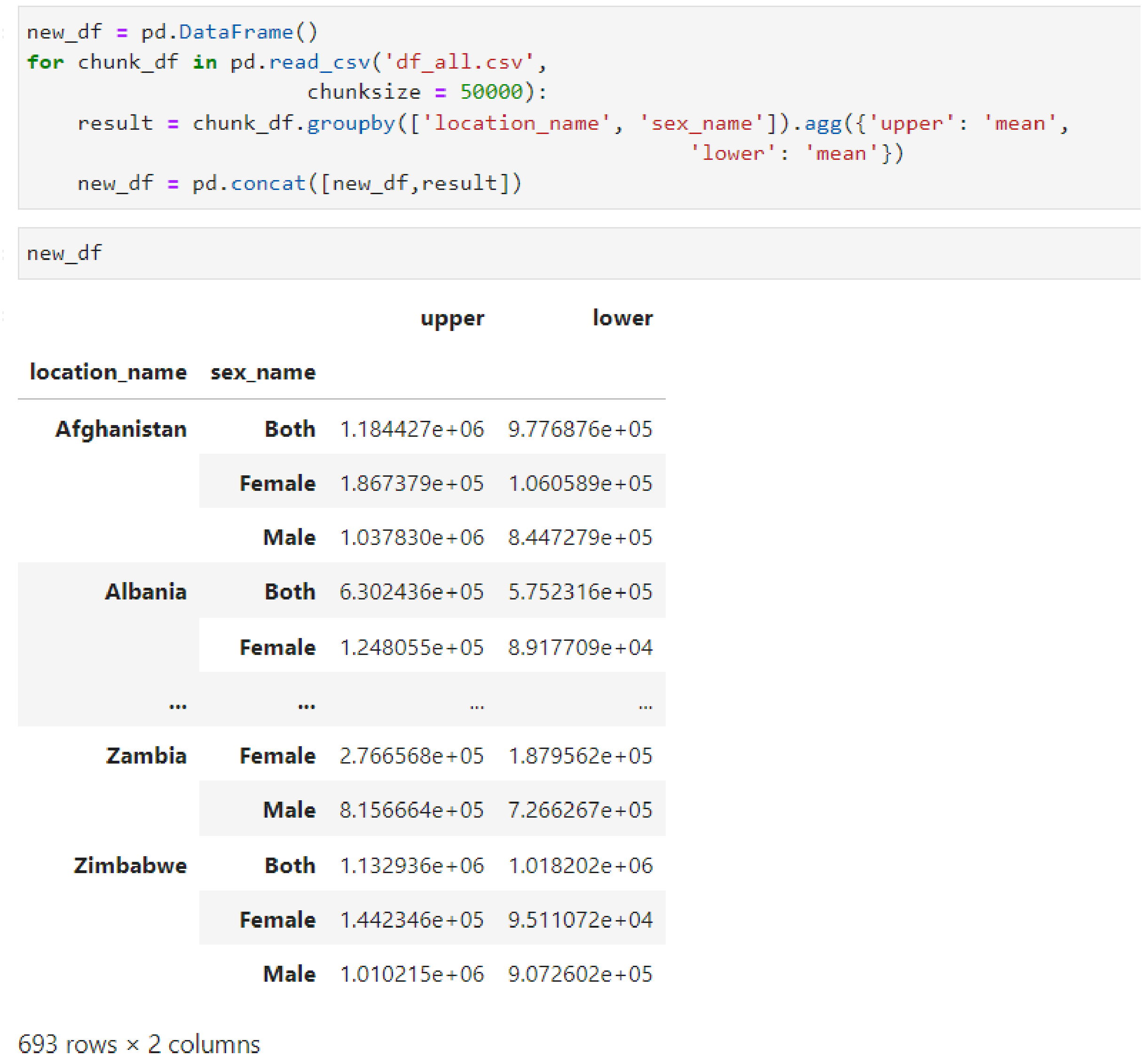


```
df.to_csv('df_all.csv')

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
```



```
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])
```

```
new_df
```

		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 × 2 columns

3. Wnioski

Na podstawie otrzymanego wyniku można stwierdzić, że biblioteka Pandas pozwala na analizę danych, oraz wczytywać, czyścić oraz modyfikować dane. Moduł pyplot umożliwia stworzyć różne wykresy.