

```
In [1]: #Ładowanie biblioteki Pandas
```

```
In [2]: #tworzenie ramki danych ze słownika
```

```
data = {'col_1': [3, 2, 1, 0], 'col_2': ['a', 'b', 'c', 'd']}  
pd.DataFrame.from_dict(data)
```

```
Out[2]:
```

	col_1	col_2
0	3	a
1	2	b
2	1	c
3	0	d

In [3]: *#zachowanie ramki danych pobranych z pliku w formacie csv (xlsx)*

```
df = pd.read_csv('IHME_GBD_2019_SMOKING_TOB_1990_2019_NUM_SMOKERS_Y2021M05D'
```

	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]

In [4]: *#tworzenie ramki danych z listy list*

```
lists_income = [ ["Adam", "Kuba", "Robert"],
[4500, 5500, 6500]]
```

Out[4]:

	0	1	2
0	Adam	Kuba	Robert
1	4500	5500	6500

In [5]: *#transponowanie (wymieniamy kolumny a wierszy)*

```
df1 = pd.DataFrame.transpose(pd.DataFrame(lists_income))
```

```
      0      1
0  Adam  4500
1   Kuba  5500
2 Robert  6500
```

In [ ]:

In [6]: *#wyświetlić pierwsze 10 wierszy ramki danych*

Out[6]:

	measure_name	location_id	location_name	sex_id	sex_name	age_group_id	age_group_n
0	Number of Smokers	1	Global	1	Male	29	15+ y
1	Number of Smokers	1	Global	2	Female	29	15+ y
2	Number of Smokers	1	Global	3	Both	29	15+ y
3	Number of Smokers	1	Global	1	Male	29	15+ y
4	Number of Smokers	1	Global	2	Female	29	15+ y
5	Number of Smokers	1	Global	3	Both	29	15+ y
6	Number of Smokers	1	Global	1	Male	29	15+ y
7	Number of Smokers	1	Global	2	Female	29	15+ y
8	Number of Smokers	1	Global	3	Both	29	15+ y
9	Number of Smokers	1	Global	1	Male	29	15+ y

In [7]: *#wyświetlić ostatnie 10 wierszy ramki danych*

Out[7]:

	measure_name	location_id	location_name	sex_id	sex_name	age_group_id	age_gro
20960	Number of Smokers	522	Sudan	3	Both	29	
20961	Number of Smokers	522	Sudan	1	Male	29	
20962	Number of Smokers	522	Sudan	2	Female	29	
20963	Number of Smokers	522	Sudan	3	Both	29	
20964	Number of Smokers	522	Sudan	1	Male	29	
20965	Number of Smokers	522	Sudan	2	Female	29	
20966	Number of Smokers	522	Sudan	3	Both	29	
20967	Number of Smokers	522	Sudan	1	Male	29	
20968	Number of Smokers	522	Sudan	2	Female	29	
20969	Number of Smokers	522	Sudan	3	Both	29	

In [8]: *#wyświetlić informacje o ramce danych*

```
<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
```

In [9]: *#wyświetlić, ile wierszy i kolumn znajduje się w ramce danych*

Out[9]: (20970, 11)

```
In [10]: #wyświetlić informacje statystyczna o kolumnach liczbowych (wartości
#niepowtarzalne, średnia, odchylenie standardowe, minimum, kwartyle,
#maksimum)
```

```
Out[10]:
```

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

```
In [11]: #wyświetlić informacje statystyczna o kolumnach kategoryzowanych (ile
#unikalnych wartości, top - jaka jest najpopularniejsza wartość, freq -
#jak często najpopularniejsza
```

```
Out[11]:
```

	measure_name	location_id	location_name	sex_id	sex_name	age_group_id
count	20970	20970.000000	20970	20970.000000	20970	20970.0
unique	1	NaN	231	NaN	3	NaN
top	Number of Smokers	NaN	South Asia	NaN	Male	NaN
freq	20970	NaN	180	NaN	6990	NaN
mean	NaN	131.111588	NaN	2.000000	NaN	29.0
std	NaN	95.055111	NaN	0.816516	NaN	0.0
min	NaN	1.000000	NaN	1.000000	NaN	29.0
25%	NaN	61.000000	NaN	1.000000	NaN	29.0
50%	NaN	119.000000	NaN	2.000000	NaN	29.0
75%	NaN	177.000000	NaN	3.000000	NaN	29.0
max	NaN	522.000000	NaN	3.000000	NaN	29.0

In [12]: *#usuna ,c brakuja ,ce warto sci w ramce danych*

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

	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]

```
In [13]: #przedstawić wybór wierszy i kolumny używając nazw oraz indeksów na
#różne sposoby
```

```
Out[13]: 0      Number of Smokers
1      Number of Smokers
2      Number of Smokers
3      Number of Smokers
4      Number of Smokers
...
20965  Number of Smokers
20966  Number of Smokers
20967  Number of Smokers
20968  Number of Smokers
20969  Number of Smokers
Name: measure_name, Length: 20970, dtype: object
```

```
In [14]:
```

```
Out[14]: 0      Number of Smokers
1      Number of Smokers
2      Number of Smokers
3      Number of Smokers
4      Number of Smokers
...
20965  Number of Smokers
20966  Number of Smokers
20967  Number of Smokers
20968  Number of Smokers
20969  Number of Smokers
Name: measure_name, Length: 20970, dtype: object
```

```
In [15]:
```

```
Out[15]:
```

	measure_name	age_group_name	year_id
0	Number of Smokers	15+ years	1990
1	Number of Smokers	15+ years	1990
2	Number of Smokers	15+ years	1990
3	Number of Smokers	15+ years	1991
4	Number of Smokers	15+ years	1991
...	...	...	...
20965	Number of Smokers	15+ years	2018
20966	Number of Smokers	15+ years	2018
20967	Number of Smokers	15+ years	2019
20968	Number of Smokers	15+ years	2019
20969	Number of Smokers	15+ years	2019

20970 rows × 3 columns

In [16]:

Out[16]:

	location_name	sex_id	sex_name	age_group_id	age_group_name	year_id	
0	Global	1	Male	29	15+ years	1990	8.031015e
1	Global	2	Female	29	15+ years	1990	1.891488e
2	Global	3	Both	29	15+ years	1990	9.922503e
3	Global	1	Male	29	15+ years	1991	8.138972e
4	Global	2	Female	29	15+ years	1991	1.905375e
...	...	...	...	...	...	...	...
20965	Sudan	2	Female	29	15+ years	2018	2.435999e
20966	Sudan	3	Both	29	15+ years	2018	2.610672e
20967	Sudan	1	Male	29	15+ years	2019	2.439150e
20968	Sudan	2	Female	29	15+ years	2019	2.500800e
20969	Sudan	3	Both	29	15+ years	2019	2.689230e

20970 rows × 7 columns



In [17]:

Out[17]:

	location_name	sex_id	sex_name	age_group_id	age_group_name	year_id	value
100	Southeast Asia, East Asia, and Oceania	2	Female	29	15+ years	1993	2.843611e+0
101	Southeast Asia, East Asia, and Oceania	3	Both	29	15+ years	1993	3.984578e+0
102	Southeast Asia, East Asia, and Oceania	1	Male	29	15+ years	1994	3.744452e+0
103	Southeast Asia, East Asia, and Oceania	2	Female	29	15+ years	1994	2.908238e+0
104	Southeast Asia, East Asia, and Oceania	3	Both	29	15+ years	1994	4.035276e+0
105	Southeast Asia, East Asia, and Oceania	1	Male	29	15+ years	1995	3.779077e+0
106	Southeast Asia, East Asia, and Oceania	2	Female	29	15+ years	1995	2.964857e+0
107	Southeast Asia, East Asia, and Oceania	3	Both	29	15+ years	1995	4.075563e+0
108	Southeast Asia, East Asia, and Oceania	1	Male	29	15+ years	1996	3.803069e+0
109	Southeast Asia, East Asia, and Oceania	2	Female	29	15+ years	1996	3.015084e+0
110	Southeast Asia, East Asia, and Oceania	3	Both	29	15+ years	1996	4.104577e+0

In [18]:

Out[18]:

	measure_name	location_id	location_name
100	Number of Smokers	4	Southeast Asia, East Asia, and Oceania
101	Number of Smokers	4	Southeast Asia, East Asia, and Oceania
102	Number of Smokers	4	Southeast Asia, East Asia, and Oceania
103	Number of Smokers	4	Southeast Asia, East Asia, and Oceania
104	Number of Smokers	4	Southeast Asia, East Asia, and Oceania
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

```
In [19]: #przedstawić wybór wierszy z ramki danych pod warunkiem odnośnie
#określonej wartości kolumny
```

Out[19]:

	measure_name	location_id	location_name	sex_id	sex_name	age_group_id	age_gro
2	Number of Smokers	1	Global	3	Both	29	
5	Number of Smokers	1	Global	3	Both	29	
8	Number of Smokers	1	Global	3	Both	29	
11	Number of Smokers	1	Global	3	Both	29	
14	Number of Smokers	1	Global	3	Both	29	
...	...	...	...	...	...	...	...
20957	Number of Smokers	522	Sudan	3	Both	29	
20960	Number of Smokers	522	Sudan	3	Both	29	
20963	Number of Smokers	522	Sudan	3	Both	29	
20966	Number of Smokers	522	Sudan	3	Both	29	
20969	Number of Smokers	522	Sudan	3	Both	29	

6990 rows × 11 columns

In [20]: *#przedstawić wybór wierszy z ramki danych pod warunkiem spełnienia  
#kilku warunków jednocześnie*

```
cardio = df[(df["sex_name"] == "Both") & (df["year_id"] == 2018) & (df["age_g
```

Out[20]:

	measure_name	location_id	location_name	sex_id	sex_name	age_group_id	age_gro
86	Number of Smokers	1	Global	3	Both	29	
176	Number of Smokers	4	Southeast Asia, East Asia, and Oceania	3	Both	29	
266	Number of Smokers	5	East Asia	3	Both	29	
356	Number of Smokers	6	China	3	Both	29	
446	Number of Smokers	7	Democratic People's Republic of Korea	3	Both	29	
...	...	...	...	...	...	...	
20606	Number of Smokers	413	Tokelau	3	Both	29	
20696	Number of Smokers	416	Tuvalu	3	Both	29	
20786	Number of Smokers	422	United States Virgin Islands	3	Both	29	
20876	Number of Smokers	435	South Sudan	3	Both	29	
20966	Number of Smokers	522	Sudan	3	Both	29	

233 rows × 11 columns

In [21]: *# wybrać wiersze które zawierają w kolumnie kategoryzowanej określone s*

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

Out[21]:

	measure_name	location_id	location_name	sex_id	sex_name	age_group_id	age_gro
1980	Number of Smokers	25	Micronesia (Federated States of)	1	Male	29	
1981	Number of Smokers	25	Micronesia (Federated States of)	2	Female	29	
1982	Number of Smokers	25	Micronesia (Federated States of)	3	Both	29	
1983	Number of Smokers	25	Micronesia (Federated States of)	1	Male	29	
1984	Number of Smokers	25	Micronesia (Federated States of)	2	Female	29	
...	...	...	...	...	...	...	
20785	Number of Smokers	422	United States Virgin Islands	2	Female	29	
20786	Number of Smokers	422	United States Virgin Islands	3	Both	29	
20787	Number of Smokers	422	United States Virgin Islands	1	Male	29	
20788	Number of Smokers	422	United States Virgin Islands	2	Female	29	
20789	Number of Smokers	422	United States Virgin Islands	3	Both	29	

270 rows × 11 columns

In [22]: *# wybrać wiersze które nie zawierają w kolumnie kategoryzowanej określenia słowa*

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

Out[22]:

	measure_name	location_id	location_name	sex_id	sex_name	age_group_id	age_group_name
0	Number of Smokers	1	Global	1	Male	29	
1	Number of Smokers	1	Global	2	Female	29	
2	Number of Smokers	1	Global	3	Both	29	
3	Number of Smokers	1	Global	1	Male	29	
4	Number of Smokers	1	Global	2	Female	29	
...	...	...	...	...	...	...	...
20965	Number of Smokers	522	Sudan	2	Female	29	
20966	Number of Smokers	522	Sudan	3	Both	29	
20967	Number of Smokers	522	Sudan	1	Male	29	
20968	Number of Smokers	522	Sudan	2	Female	29	
20969	Number of Smokers	522	Sudan	3	Both	29	

20700 rows × 11 columns

In [23]: *#utwórz kolumnę, na podstawie istnieją.cyn*

```
df["Tolerance_range"] = df["upper"] - df["lower"]
```

Out[23]:

	measure_name	location_id	location_name	sex_id	sex_name	age_group_id	age_gro
0	Number of Smokers	1	Global	1	Male	29	
1	Number of Smokers	1	Global	2	Female	29	
2	Number of Smokers	1	Global	3	Both	29	
3	Number of Smokers	1	Global	1	Male	29	
4	Number of Smokers	1	Global	2	Female	29	
...	...	...	...	...	...	...	
20965	Number of Smokers	522	Sudan	2	Female	29	
20966	Number of Smokers	522	Sudan	3	Both	29	
20967	Number of Smokers	522	Sudan	1	Male	29	
20968	Number of Smokers	522	Sudan	2	Female	29	
20969	Number of Smokers	522	Sudan	3	Both	29	

20970 rows × 12 columns

In [24]: *#usuń kolumnę*

```
df = df.drop("age_group_id", axis = 1)
```

Out[24]:

	measure_name	location_id	location_name	sex_id	sex_name	age_group_name	year
0	Number of Smokers	1	Global	1	Male	15+ years	19
1	Number of Smokers	1	Global	2	Female	15+ years	19
2	Number of Smokers	1	Global	3	Both	15+ years	19
3	Number of Smokers	1	Global	1	Male	15+ years	19
4	Number of Smokers	1	Global	2	Female	15+ years	19
...	...	...	...	...	...	...	...
20965	Number of Smokers	522	Sudan	2	Female	15+ years	20
20966	Number of Smokers	522	Sudan	3	Both	15+ years	20
20967	Number of Smokers	522	Sudan	1	Male	15+ years	20
20968	Number of Smokers	522	Sudan	2	Female	15+ years	20
20969	Number of Smokers	522	Sudan	3	Both	15+ years	20

20970 rows × 11 columns

In [25]: *#zmień nazwę kolumny*

```
df = df.rename(columns = {"sex_name": "sex"})
```

Out[25]:

	measure_name	location_id	location_name	sex_id	sex	age_group_name	year_id
0	Number of Smokers	1	Global	1	Male	15+ years	1990
1	Number of Smokers	1	Global	2	Female	15+ years	1990
2	Number of Smokers	1	Global	3	Both	15+ years	1990
3	Number of Smokers	1	Global	1	Male	15+ years	1991
4	Number of Smokers	1	Global	2	Female	15+ years	1991
...	...	...	...	...	...	...	...
20965	Number of Smokers	522	Sudan	2	Female	15+ years	2018
20966	Number of Smokers	522	Sudan	3	Both	15+ years	2018
20967	Number of Smokers	522	Sudan	1	Male	15+ years	2019
20968	Number of Smokers	522	Sudan	2	Female	15+ years	2019
20969	Number of Smokers	522	Sudan	3	Both	15+ years	2019

20970 rows × 11 columns



In [26]:

Out[26]:

	measure_name	location_id	location_name	sex_id	sex	age_group_name	year_id
0	Number of Smokers	1	Global	1	Male	15+ years	1990
1	Number of Smokers	1	Global	2	Female	15+ years	1990
2	Number of Smokers	1	Global	3	Both	15+ years	1990
3	Number of Smokers	1	Global	1	Male	15+ years	1991
4	Number of Smokers	1	Global	2	Female	15+ years	1991
...	...	...	...	...	...	...	...
20965	Number of Smokers	522	Sudan	2	Female	15+ years	2018
20966	Number of Smokers	522	Sudan	3	Both	15+ years	2018
20967	Number of Smokers	522	Sudan	1	Male	15+ years	2019
20968	Number of Smokers	522	Sudan	2	Female	15+ years	2019
20969	Number of Smokers	522	Sudan	3	Both	15+ years	2019

20970 rows × 11 columns

In [27]: *#zachowaj ramke , danych jako plik csv na komputerze*

In [28]:

Out[28]:

	measure_name	location_id	location_name	sex_id	sex	age_group_name	year_id
0	Number of Smokers	1	Global	1	Male	15+ years	1990
1	Number of Smokers	1	Global	2	Female	15+ years	1990
2	Number of Smokers	1	Global	3	Both	15+ years	1990
3	Number of Smokers	1	Global	1	Male	15+ years	1991
4	Number of Smokers	1	Global	2	Female	15+ years	1991
...	...	...	...	...	...	...	...
20965	Number of Smokers	522	Sudan	2	Female	15+ years	2018
20966	Number of Smokers	522	Sudan	3	Both	15+ years	2018
20967	Number of Smokers	522	Sudan	1	Male	15+ years	2019
20968	Number of Smokers	522	Sudan	2	Female	15+ years	2019
20969	Number of Smokers	522	Sudan	3	Both	15+ years	2019

20970 rows × 11 columns

In [29]: *#wyświetlić średnia (maksymalna, minimalna) wartość z jednej kolumny*

```
print(df["year_id"].mean())  
print(df["year_id"].max())
```

```
2004.5  
2019  
1990
```

In [30]: *#wyświetlić liczbę wierszy*

```
rows = len(df.axes[0])
```

Out[30]: 20970

In [31]: *#wyświetlić wartości unikatowe w kolumnie*

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

In [32]: *#wyświetlić liczby rekordów odpowiadających do wartości*

```
Out[32]: sex  
Male      6990  
Female    6990  
Both      6990  
Name: count, dtype: int64
```

In [33]:

#sortowanie wierszy ramki danych według warto´sci okre´slonej kolumny  
#(maleja.co, rosna.co)

Out[33]:

	measure_name	location_id	location_name	sex_id	sex	age_group_name	year_id
20572	Number of Smokers	413	Tokelau	2	Female	15+ years	2007
20575	Number of Smokers	413	Tokelau	2	Female	15+ years	2008
20569	Number of Smokers	413	Tokelau	2	Female	15+ years	2006
20578	Number of Smokers	413	Tokelau	2	Female	15+ years	2009
20581	Number of Smokers	413	Tokelau	2	Female	15+ years	2010
...	...	...	...	...	...	...	...
71	Number of Smokers	1	Global	3	Both	15+ years	2013
68	Number of Smokers	1	Global	3	Both	15+ years	2012
83	Number of Smokers	1	Global	3	Both	15+ years	2017
86	Number of Smokers	1	Global	3	Both	15+ years	2018
89	Number of Smokers	1	Global	3	Both	15+ years	2019

20970 rows × 11 columns

In [34]:

Out[34]:

	measure_name	location_id	location_name	sex_id	sex	age_group_name	year_id
89	Number of Smokers	1	Global	3	Both	15+ years	2019
86	Number of Smokers	1	Global	3	Both	15+ years	2018
83	Number of Smokers	1	Global	3	Both	15+ years	2017
68	Number of Smokers	1	Global	3	Both	15+ years	2012
71	Number of Smokers	1	Global	3	Both	15+ years	2013
...	...	...	...	...	...	...	...
20581	Number of Smokers	413	Tokelau	2	Female	15+ years	2010
20578	Number of Smokers	413	Tokelau	2	Female	15+ years	2009
20569	Number of Smokers	413	Tokelau	2	Female	15+ years	2006
20575	Number of Smokers	413	Tokelau	2	Female	15+ years	2008
20572	Number of Smokers	413	Tokelau	2	Female	15+ years	2007

20970 rows × 11 columns

In [35]:

Out[35]:

	measure_name	location_id	location_name	sex_id	sex	age_group_name	year_id
20572	Number of Smokers	413	Tokelau	2	Female	15+ years	2007
20575	Number of Smokers	413	Tokelau	2	Female	15+ years	2008
20569	Number of Smokers	413	Tokelau	2	Female	15+ years	2006
20578	Number of Smokers	413	Tokelau	2	Female	15+ years	2009
20581	Number of Smokers	413	Tokelau	2	Female	15+ years	2010
20566	Number of Smokers	413	Tokelau	2	Female	15+ years	2005
20584	Number of Smokers	413	Tokelau	2	Female	15+ years	2011
20587	Number of Smokers	413	Tokelau	2	Female	15+ years	2012
20590	Number of Smokers	413	Tokelau	2	Female	15+ years	2013
20593	Number of Smokers	413	Tokelau	2	Female	15+ years	2014

In [36]:

Out[36]:

	measure_name	location_id	location_name	sex_id	sex	age_group_name	year_id	
89	Number of Smokers	1	Global	3	Both	15+ years	2019	1.14
86	Number of Smokers	1	Global	3	Both	15+ years	2018	1.13
83	Number of Smokers	1	Global	3	Both	15+ years	2017	1.13
68	Number of Smokers	1	Global	3	Both	15+ years	2012	1.13
71	Number of Smokers	1	Global	3	Both	15+ years	2013	1.13
80	Number of Smokers	1	Global	3	Both	15+ years	2016	1.13
65	Number of Smokers	1	Global	3	Both	15+ years	2011	1.13
74	Number of Smokers	1	Global	3	Both	15+ years	2014	1.13
77	Number of Smokers	1	Global	3	Both	15+ years	2015	1.13
62	Number of Smokers	1	Global	3	Both	15+ years	2010	1.12

In [37]:

```
#wyświetlić wierszy dla 10 największych wartości określonej kolumny  
#pod warunkiem określonych wartości innej kolumny
```

Out[37]:

	measure_name	location_id	location_name	sex_id	sex	age_group_name	year_id	
4322	Number of Smokers	51	Poland	3	Both	15+ years	1990	11
4325	Number of Smokers	51	Poland	3	Both	15+ years	1991	11
4328	Number of Smokers	51	Poland	3	Both	15+ years	1992	11
4331	Number of Smokers	51	Poland	3	Both	15+ years	1993	11
4334	Number of Smokers	51	Poland	3	Both	15+ years	1994	11
4337	Number of Smokers	51	Poland	3	Both	15+ years	1995	10
4340	Number of Smokers	51	Poland	3	Both	15+ years	1996	10
4343	Number of Smokers	51	Poland	3	Both	15+ years	1997	10
4346	Number of Smokers	51	Poland	3	Both	15+ years	1998	10
4349	Number of Smokers	51	Poland	3	Both	15+ years	1999	10

```
In [38]: #grupowanie wierszy według wartości kolumny kategoryzowanej, potem
#- uśrednienie wartości wszystkich kolumn w grupie - MultiIndex

df_new = df.groupby(['location_name', 'sex']).agg({'val': 'mean',
                                                  'upper': 'mean',
                                                  'lower': 'mean'})

df_new
```

```
Out[38]:
```

		val	upper	lower
location_name	sex			
	Both	1.076844e+06	1.184427e+06	9.776876e+05
Afghanistan	Female	1.408633e+05	1.867379e+05	1.060589e+05
	Male	9.359803e+05	1.037830e+06	8.447279e+05
	Both	6.016696e+05	6.302436e+05	5.752316e+05
Albania	Female	1.060032e+05	1.248055e+05	8.917709e+04
...	...	...	...	...
	Female	2.285319e+05	2.766568e+05	1.879562e+05
Zambia	Male	7.708210e+05	8.156664e+05	7.266267e+05
	Both	1.075152e+06	1.132936e+06	1.018202e+06
Zimbabwe	Female	1.164704e+05	1.442346e+05	9.511072e+04
	Male	9.586813e+05	1.010215e+06	9.072602e+05

693 rows × 5 columns

In [39]: *#grupowanie wierszy według wartości kolumny kategoryzowanej, potem  
#- uśrednienie wartości dla pewnych kolumn, liczba wartości i mediana  
#dla pozostałych kolumn w grupach*

```
df_new = df.groupby(['location_name', 'sex']).agg({  
    'val': 'mean',  
    'upper': ['median', 'count'],  
    'sex_id' : ['median', 'count']})
```

Out[39]:

		val	upper		sex_id	
		mean	median	count	median	count
location_name	sex					
Afghanistan	Both	1.076844e+06	1.051483e+06	30	3.0	30
	Female	1.408633e+05	1.583590e+05	30	2.0	30
	Male	9.359803e+05	9.289502e+05	30	1.0	30
Albania	Both	6.016696e+05	5.955954e+05	30	3.0	30
	Female	1.060032e+05	1.133635e+05	30	2.0	30
...	...	...	...	...	...	...
Zambia	Female	2.285319e+05	2.926822e+05	30	2.0	30
	Male	7.708210e+05	7.427456e+05	30	1.0	30
	Both	1.075152e+06	1.065280e+06	30	3.0	30
Zimbabwe	Female	1.164704e+05	1.521148e+05	30	2.0	30
	Male	9.586813e+05	9.324646e+05	30	1.0	30

693 rows × 5 columns

In [40]: *#wyświetlić nazwy kolumn indeksu złożonego*

Out[40]: MultiIndex([( 'val', 'mean'),  
( 'upper', 'median'),  
( 'upper', 'count'),  
( 'sex\_id', 'median'),  
( 'sex\_id', 'count')],  
)

In [41]: *#sortować kolumnę indeksu złożonego*

Out[41]:

location_name	sex	
Tokelau	Female	9.568048e+01
Niue	Female	1.109018e+02
	Male	1.823211e+02
Tokelau	Male	1.978756e+02
Niue	Both	2.831822e+02
	...	
East Asia	Both	3.135026e+08
Southeast Asia, East Asia, and Oceania	Male	3.981023e+08
	Both	4.314121e+08
Global	Male	8.892060e+08
	Both	1.097092e+09

Name: median, Length: 693, dtype: float64

In [42]: *#stworzy´c tabele, przystawna, (pivot table) na podstawie ramki danych*

```
df_pivot = df.pivot_table(values='val', index='location_name', columns='sex',
                           margins=False, dropna=True, fill_value=None) # tabela
```

Out[42]:

	sex	Both	Female	Male
location_name				
Afghanistan		1.076844e+06	1.408633e+05	9.359803e+05
Albania		6.016696e+05	1.060032e+05	4.956664e+05
Algeria		3.873312e+06	2.192907e+05	3.654021e+06
American Samoa		1.225455e+04	4.320555e+03	7.933993e+03
Andean Latin America		3.566739e+06	9.573733e+05	2.609366e+06
...		...	...	...
Western Europe		9.714860e+07	4.233137e+07	5.481723e+07
Western Sub-Saharan Africa		1.377671e+07	1.868140e+06	1.190857e+07
Yemen		2.233432e+06	4.807306e+05	1.752701e+06
Zambia		9.993529e+05	2.285319e+05	7.708210e+05
Zimbabwe		1.075152e+06	1.164704e+05	9.586813e+05

231 rows × 3 columns

In [43]: *#wy´swietli´c indeksy i kolumny tabeli przystawnej*

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



In [44]:

#utwór indeks złożony tabeli przystawnej i wyświetl go

```
df_pivot = df.pivot_table(values='val', index=['location_name','year_id'],
                           margins=False, dropna=True, fill_value=None)
```

Out[44]:

	sex	Both	Female	Male
location_name	year_id			
Afghanistan	1990	1.952669e+05	33006.33880	1.622605e+05
	1991	2.296694e+05	37383.30729	1.922861e+05
	1992	2.895567e+05	44365.25615	2.451914e+05
	1993	3.435076e+05	50238.47594	2.932691e+05
	1994	3.806563e+05	54130.87718	3.265254e+05
...	...	...	...	...
Zimbabwe	2015	1.327739e+06	141442.90390	1.186297e+06
	2016	1.361949e+06	141928.43040	1.220020e+06
	2017	1.396007e+06	141905.15710	1.254102e+06
	2018	1.430277e+06	141451.47270	1.288825e+06
	2019	1.465099e+06	140772.04550	1.324327e+06

6930 rows × 3 columns

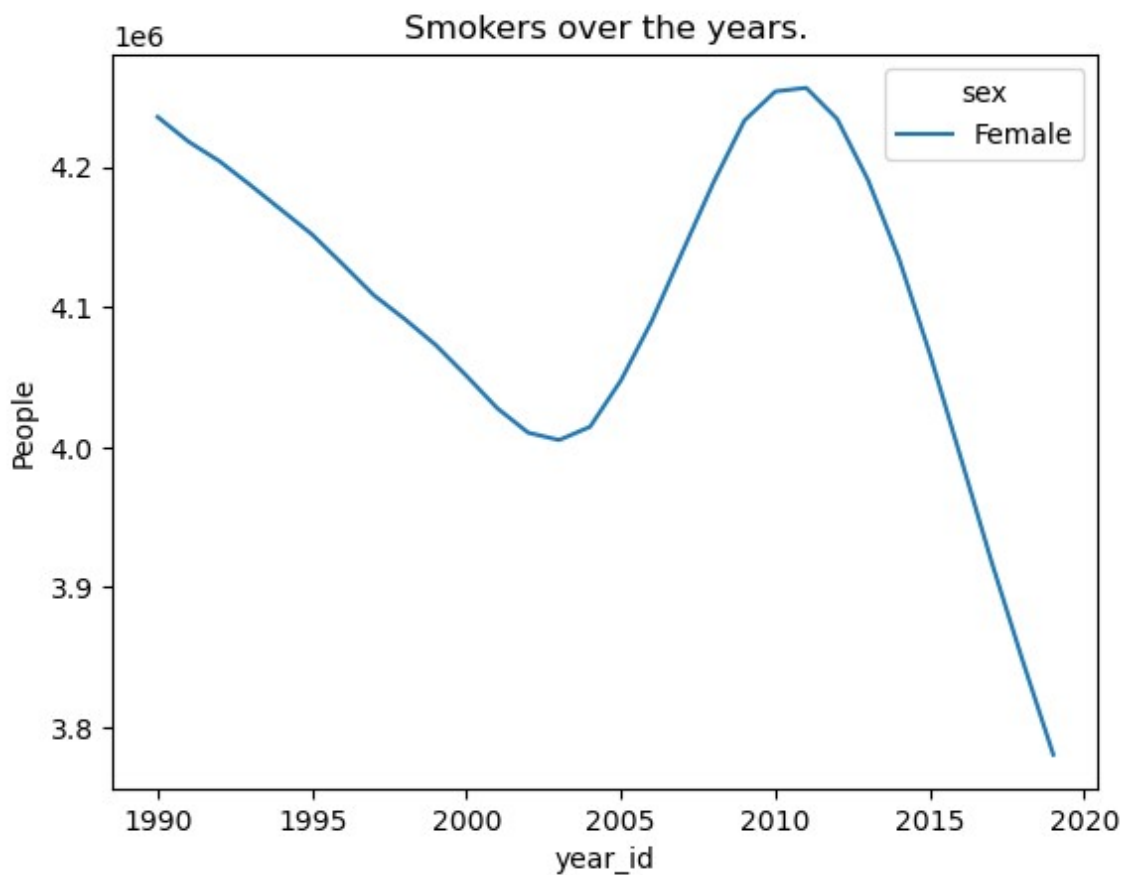
In [45]:

#zaimportuj modul pyplot z biblioteki matplotlib

In [46]: *#wyświeślić wykres na podstawie tabeli przystawnej*

```
df[(df['location_name'] == 'Poland') & (df['sex'] == 'Female')].pivot_table  
    fill_value=None, margins=False, d  
plt.ylabel('People') # etykieta osi y
```

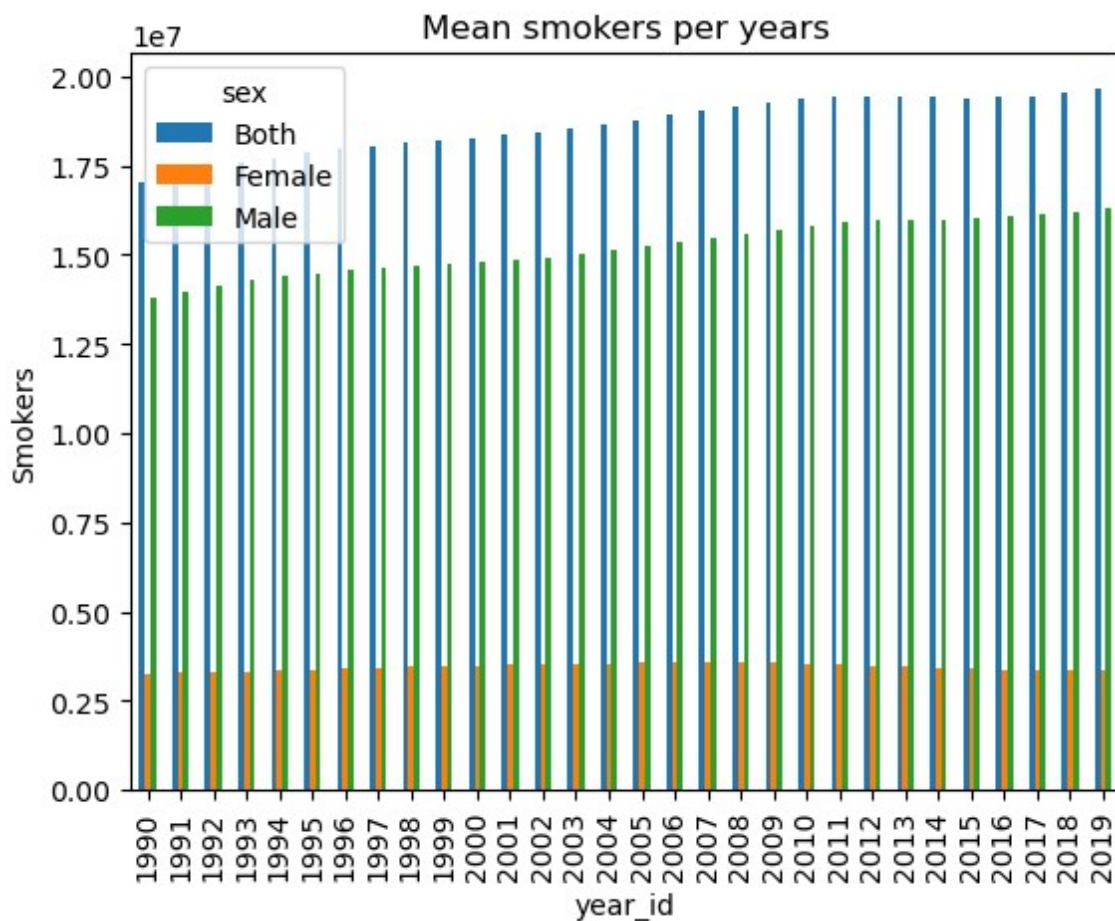
Out[46]: Text(0.5, 1.0, 'Smokers over the years.')



In [47]: *#narysować histogram na podstawie wartości kolumny*

```
df_bar = df[(df['sex'].isin(['Male', 'Female', 'Both']))].pivot_table(values=
    index='year_id', columns='sex', aggfunc='mean',
    fill_value=None, margins=False, dropna=True)
df_bar.plot(kind = 'bar')
plt.ylabel('Smokers')
```

Out[47]: Text(0.5, 1.0, 'Mean smokers per years')



```
In [48]: #przedstawić sposoby łączenia ramek danych za pomocą metod merge i
#concat

df2 = pd.read_csv('IHME_GBD_2019_SMOKING_TOB_1990_2019_CIG_PC_Y2021M05D27.CS'
```

Out[48]:

	measure_name	location_id	location_name	sex_id	sex_name	age_group_id	age_group
0	Cigarette-Equivalents Per Capita	1	Global	3	Both	29	15
1	Cigarette-Equivalents Per Capita	1	Global	3	Both	29	15
2	Cigarette-Equivalents Per Capita	4	Southeast Asia, East Asia, and Oceania	3	Both	29	15
3	Cigarette-Equivalents Per Capita	4	Southeast Asia, East Asia, and Oceania	3	Both	29	15
4	Cigarette-Equivalents Per Capita	5	East Asia	3	Both	29	15
...	...	...	...	...	...	...	...
461	Cigarette-Equivalents Per Capita	422	United States Virgin Islands	3	Both	29	15
462	Cigarette-Equivalents Per Capita	435	South Sudan	3	Both	29	15
463	Cigarette-Equivalents Per Capita	435	South Sudan	3	Both	29	15
464	Cigarette-Equivalents Per Capita	522	Sudan	3	Both	29	15
465	Cigarette-Equivalents Per Capita	522	Sudan	3	Both	29	15

466 rows × 11 columns

In [49]:

Out[49]:

	measure_name	location_id	location_name	sex_id	sex	age_group_name	year_id
0	Number of Smokers	1	Global	1	Male	15+ years	1990
1	Number of Smokers	1	Global	2	Female	15+ years	1990
2	Number of Smokers	1	Global	3	Both	15+ years	1990
3	Number of Smokers	1	Global	1	Male	15+ years	1991
4	Number of Smokers	1	Global	2	Female	15+ years	1991
...	...	...	...	...	...	...	...
20965	Number of Smokers	522	Sudan	2	Female	15+ years	2018
20966	Number of Smokers	522	Sudan	3	Both	15+ years	2018
20967	Number of Smokers	522	Sudan	1	Male	15+ years	2019
20968	Number of Smokers	522	Sudan	2	Female	15+ years	2019
20969	Number of Smokers	522	Sudan	3	Both	15+ years	2019

20970 rows × 11 columns

```
In [50]: df2.rename(columns = {'val': 'val_Cigarette-Equivalents Per Capita', 'upper':  
df.drop('sex_id', axis = 1)
```

```
Out[50]:
```

	measure_name	location_id	location_name	sex	age_group_name	year_id	
0	Number of Smokers	1	Global	Male	15+ years	1990	8.03107
1	Number of Smokers	1	Global	Female	15+ years	1990	1.89148
2	Number of Smokers	1	Global	Both	15+ years	1990	9.92250
3	Number of Smokers	1	Global	Male	15+ years	1991	8.13897
4	Number of Smokers	1	Global	Female	15+ years	1991	1.90537
...	...	...	...	...	...	...	...
20965	Number of Smokers	522	Sudan	Female	15+ years	2018	2.43596
20966	Number of Smokers	522	Sudan	Both	15+ years	2018	2.61067
20967	Number of Smokers	522	Sudan	Male	15+ years	2019	2.43915
20968	Number of Smokers	522	Sudan	Female	15+ years	2019	2.50080
20969	Number of Smokers	522	Sudan	Both	15+ years	2019	2.68923

20970 rows × 10 columns

```
In [51]:
```

In [52]:

Out[52]:

	measure_name_x	location_id	location_name	sex_id_x	sex	age_group_name	year
0	Number of Smokers	1	Global	1	Male	15+ years	19
1	Number of Smokers	1	Global	2	Female	15+ years	19
2	Number of Smokers	1	Global	3	Both	15+ years	19
3	Number of Smokers	1	Global	1	Male	15+ years	20
4	Number of Smokers	1	Global	2	Female	15+ years	20
...	...	...	...	...	...	...	...
1393	Number of Smokers	522	Sudan	2	Female	15+ years	19
1394	Number of Smokers	522	Sudan	3	Both	15+ years	19
1395	Number of Smokers	522	Sudan	1	Male	15+ years	20
1396	Number of Smokers	522	Sudan	2	Female	15+ years	20
1397	Number of Smokers	522	Sudan	3	Both	15+ years	20

1398 rows × 18 columns

```
In [53]: df_all_1 = df.iloc[1:15,:]  
df_all_1
```

```
Out[53]:
```

	measure_name	location_id	location_name	sex_id	sex	age_group_name	year_id	
1	Number of Smokers	1	Global	2	Female	15+ years	1990	1.
2	Number of Smokers	1	Global	3	Both	15+ years	1990	9.
3	Number of Smokers	1	Global	1	Male	15+ years	1991	8.
4	Number of Smokers	1	Global	2	Female	15+ years	1991	1.
5	Number of Smokers	1	Global	3	Both	15+ years	1991	1.
6	Number of Smokers	1	Global	1	Male	15+ years	1992	8.
7	Number of Smokers	1	Global	2	Female	15+ years	1992	1.
8	Number of Smokers	1	Global	3	Both	15+ years	1992	1.
9	Number of Smokers	1	Global	1	Male	15+ years	1993	8.
10	Number of Smokers	1	Global	2	Female	15+ years	1993	1.
11	Number of Smokers	1	Global	3	Both	15+ years	1993	1.
12	Number of Smokers	1	Global	1	Male	15+ years	1994	8.
13	Number of Smokers	1	Global	2	Female	15+ years	1994	1.
14	Number of Smokers	1	Global	3	Both	15+ years	1994	1.



```
In [54]: df_all_2 = df.iloc[-15::]
```

```
Out[54]:
```

	measure_name	location_id	location_name	sex_id	sex	age_group_name	year_id
<b>20955</b>	Number of Smokers	522	Sudan	1	Male	15+ years	2015
<b>20956</b>	Number of Smokers	522	Sudan	2	Female	15+ years	2015
<b>20957</b>	Number of Smokers	522	Sudan	3	Both	15+ years	2015
<b>20958</b>	Number of Smokers	522	Sudan	1	Male	15+ years	2016
<b>20959</b>	Number of Smokers	522	Sudan	2	Female	15+ years	2016
<b>20960</b>	Number of Smokers	522	Sudan	3	Both	15+ years	2016
<b>20961</b>	Number of Smokers	522	Sudan	1	Male	15+ years	2017
<b>20962</b>	Number of Smokers	522	Sudan	2	Female	15+ years	2017
<b>20963</b>	Number of Smokers	522	Sudan	3	Both	15+ years	2017
<b>20964</b>	Number of Smokers	522	Sudan	1	Male	15+ years	2018
<b>20965</b>	Number of Smokers	522	Sudan	2	Female	15+ years	2018
<b>20966</b>	Number of Smokers	522	Sudan	3	Both	15+ years	2018
<b>20967</b>	Number of Smokers	522	Sudan	1	Male	15+ years	2019
<b>20968</b>	Number of Smokers	522	Sudan	2	Female	15+ years	2019
<b>20969</b>	Number of Smokers	522	Sudan	3	Both	15+ years	2019

```
In [55]: df_all_new = pd.concat([df_all_1, df_all_2], axis = 0) # połącz ramki danych
# axis = 1, potem według kolumn
print(df_all_new.shape)

(29, 11)
```

Out[55]:

	measure_name	location_id	location_name	sex_id	sex	age_group_name	year_id
1	Number of Smokers	1	Global	2	Female	15+ years	1990
2	Number of Smokers	1	Global	3	Both	15+ years	1990
3	Number of Smokers	1	Global	1	Male	15+ years	1991
4	Number of Smokers	1	Global	2	Female	15+ years	1991
5	Number of Smokers	1	Global	3	Both	15+ years	1991
6	Number of Smokers	1	Global	1	Male	15+ years	1992
7	Number of Smokers	1	Global	2	Female	15+ years	1992
8	Number of Smokers	1	Global	3	Both	15+ years	1992
9	Number of Smokers	1	Global	1	Male	15+ years	1993
10	Number of Smokers	1	Global	2	Female	15+ years	1993
11	Number of Smokers	1	Global	3	Both	15+ years	1993
12	Number of Smokers	1	Global	1	Male	15+ years	1994
13	Number of Smokers	1	Global	2	Female	15+ years	1994
14	Number of Smokers	1	Global	3	Both	15+ years	1994
20955	Number of Smokers	522	Sudan	1	Male	15+ years	2015
20956	Number of Smokers	522	Sudan	2	Female	15+ years	2015
20957	Number of Smokers	522	Sudan	3	Both	15+ years	2015
20958	Number of Smokers	522	Sudan	1	Male	15+ years	2016
20959	Number of Smokers	522	Sudan	2	Female	15+ years	2016
20960	Number of Smokers	522	Sudan	3	Both	15+ years	2016
20961	Number of Smokers	522	Sudan	1	Male	15+ years	2017
20962	Number of Smokers	522	Sudan	2	Female	15+ years	2017
20963	Number of Smokers	522	Sudan	3	Both	15+ years	2017
20964	Number of Smokers	522	Sudan	1	Male	15+ years	2018

	measure_name	location_id	location_name	sex_id	sex	age_group_name	year_id
20965	Number of Smokers	522	Sudan	2	Female	15+ years	2018
20966	Number of Smokers	522	Sudan	3	Both	15+ years	2018
20967	Number of Smokers	522	Sudan	1	Male	15+ years	2019
20968	Number of Smokers	522	Sudan	2	Female	15+ years	2019
20969	Number of Smokers	522	Sudan	3	Both	15+ years	2019

```
In [56]: #pokazać dodawanie nowych kolumn za pomocą operacji matematycznych
df_all_new["smokers"] = df_all_new["val"] + df_all_new["upper"] + df_all_new["lower"]
df_all_new["%ValFromUpper"] = df_all_new["val"] / df_all_new["upper"]*100
```

In [57]:

---

Out[57]:

	measure_name	location_id	location_name	sex_id	sex	age_group_name	year_id
1	Number of Smokers	1	Global	2	Female	15+ years	1990
2	Number of Smokers	1	Global	3	Both	15+ years	1990
3	Number of Smokers	1	Global	1	Male	15+ years	1991
4	Number of Smokers	1	Global	2	Female	15+ years	1991
5	Number of Smokers	1	Global	3	Both	15+ years	1991
6	Number of Smokers	1	Global	1	Male	15+ years	1992
7	Number of Smokers	1	Global	2	Female	15+ years	1992
8	Number of Smokers	1	Global	3	Both	15+ years	1992
9	Number of Smokers	1	Global	1	Male	15+ years	1993
10	Number of Smokers	1	Global	2	Female	15+ years	1993
11	Number of Smokers	1	Global	3	Both	15+ years	1993
12	Number of Smokers	1	Global	1	Male	15+ years	1994
13	Number of Smokers	1	Global	2	Female	15+ years	1994
14	Number of Smokers	1	Global	3	Both	15+ years	1994
20955	Number of Smokers	522	Sudan	1	Male	15+ years	2015
20956	Number of Smokers	522	Sudan	2	Female	15+ years	2015
20957	Number of Smokers	522	Sudan	3	Both	15+ years	2015
20958	Number of Smokers	522	Sudan	1	Male	15+ years	2016
20959	Number of Smokers	522	Sudan	2	Female	15+ years	2016
20960	Number of Smokers	522	Sudan	3	Both	15+ years	2016
20961	Number of Smokers	522	Sudan	1	Male	15+ years	2017
20962	Number of Smokers	522	Sudan	2	Female	15+ years	2017
20963	Number of Smokers	522	Sudan	3	Both	15+ years	2017
20964	Number of Smokers	522	Sudan	1	Male	15+ years	2018

	measure_name	location_id	location_name	sex_id	sex	age_group_name	year_id
<b>20965</b>	Number of Smokers	522	Sudan	2	Female	15+ years	2018
<b>20966</b>	Number of Smokers	522	Sudan	3	Both	15+ years	2018
<b>20967</b>	Number of Smokers	522	Sudan	1	Male	15+ years	2019
<b>20968</b>	Number of Smokers	522	Sudan	2	Female	15+ years	2019
<b>20969</b>	Number of Smokers	522	Sudan	3	Both	15+ years	2019

```
In [58]: #przedstawić na przykładzie dodawanie nowych kolumn z pomocą funkcji  
#lambda  
year_id = [2017,2018,2019]  
df_all_new = df_all_new.reset_index()  
df_all_new['Is2017-2019'] = df_all['year_id'].apply(lambda x: True if x in ,
```



Out[58]:

	index	measure_name	location_id	location_name	sex_id	sex	age_group_name	yea
0	1	Number of Smokers	1	Global	2	Female	15+ years	1
1	2	Number of Smokers	1	Global	3	Both	15+ years	1
2	3	Number of Smokers	1	Global	1	Male	15+ years	1
3	4	Number of Smokers	1	Global	2	Female	15+ years	1
4	5	Number of Smokers	1	Global	3	Both	15+ years	1
5	6	Number of Smokers	1	Global	1	Male	15+ years	1
6	7	Number of Smokers	1	Global	2	Female	15+ years	1
7	8	Number of Smokers	1	Global	3	Both	15+ years	1
8	9	Number of Smokers	1	Global	1	Male	15+ years	1
9	10	Number of Smokers	1	Global	2	Female	15+ years	1
10	11	Number of Smokers	1	Global	3	Both	15+ years	1
11	12	Number of Smokers	1	Global	1	Male	15+ years	1
12	13	Number of Smokers	1	Global	2	Female	15+ years	1
13	14	Number of Smokers	1	Global	3	Both	15+ years	1
14	20955	Number of Smokers	522	Sudan	1	Male	15+ years	2
15	20956	Number of Smokers	522	Sudan	2	Female	15+ years	2
16	20957	Number of Smokers	522	Sudan	3	Both	15+ years	2
17	20958	Number of Smokers	522	Sudan	1	Male	15+ years	2
18	20959	Number of Smokers	522	Sudan	2	Female	15+ years	2
19	20960	Number of Smokers	522	Sudan	3	Both	15+ years	2
20	20961	Number of Smokers	522	Sudan	1	Male	15+ years	2
21	20962	Number of Smokers	522	Sudan	2	Female	15+ years	2
22	20963	Number of Smokers	522	Sudan	3	Both	15+ years	2
23	20964	Number of Smokers	522	Sudan	1	Male	15+ years	2

	index	measure_name	location_id	location_name	sex_id	sex	age_group_name	year_id
24	20965	Number of Smokers	522	Sudan	2	Female	15+ years	2019
25	20966	Number of Smokers	522	Sudan	3	Both	15+ years	2019
26	20967	Number of Smokers	522	Sudan	1	Male	15+ years	2019
27	20968	Number of Smokers	522	Sudan	2	Female	15+ years	2019
28	20969	Number of Smokers	522	Sudan	3	Both	15+ years	2019

```
In [59]: #przedstawić możliwości pracy z dużymi plikami przy użyciu argumentu
#chunksize

for chunk_df in pd.read_csv('IHME_GBD_2019_SMOKING_TOB_1990_2019_CIG_PC_Y2019.csv',
                             chunksize = 50000):
    print("CHUNK DF")
```

CHUNK DF

measure_name			location_id	\		
0	Cigarette-Equivalents	Per Capita	1			
1	Cigarette-Equivalents	Per Capita	1			
2	Cigarette-Equivalents	Per Capita	4			
3	Cigarette-Equivalents	Per Capita	4			
4	Cigarette-Equivalents	Per Capita	5			
			location_name	sex_id	sex_name	age_group_id
\						
0			Global	3	Both	29
1			Global	3	Both	29
2	Southeast Asia, East Asia, and Oceania			3	Both	29
3	Southeast Asia, East Asia, and Oceania			3	Both	29
4	East Asia			3	Both	29
age_group_name		year_id	val	upper		lower
0	15+ years	1990	1484.256502	1531.563739	1436.151878	
1	15+ years	2019	1113.754663	1161.263946	1069.765828	
2	15+ years	1990	1827.374739	1959.359086	1692.900863	
3	15+ years	2019	1778.846098	1927.560165	1640.645875	
4	15+ years	1990	2089.743405	2267.199999	1908.301510	

```
In [60]: new_df = pd.DataFrame() # pusta ramka danych
for chunk_df in pd.read_csv('IHME_GBD_2019_SMOKING_TOB_1990_2019_CIG_PC_Y20
                                chunksize = 50000):
    result = chunk_df.groupby(['location_name', 'year_id']).agg({'val': 'me
                                'upper': 'max'})

    new_df = pd.concat([new_df,result])
```

Out[60]:

		val	upper
location_name	year_id		
Afghanistan	1990	274.126957	320.558021
	2019	444.334632	546.171500
Albania	1990	1894.040861	2224.731864
	2019	1941.384044	2305.846372
Algeria	1990	1259.079364	1381.657971
...	...	...	...
Yemen	2019	1391.887788	1712.648491
Zambia	1990	308.165288	343.536927
	2019	296.250416	366.554416
Zimbabwe	1990	931.803728	1130.361142
	2019	898.367226	1132.376771

462 rows × 4 columns

In [ ]: