SPRAWOZDANIE

Zajęcia: Nauka o danych I

Prowadzący: prof. dr hab. Vasyl Martsenyuk

Laboratorium Nr 1 Data 28.09.2024 Temat: "Wprowadzenie do narzędzi i środowiska pracy w analizie danych" Wariant 6	Imię Nawisko Informatyka II stopień, stacjonarne, ?semestr, gr.???

1. Polecenie: wariant 1 zadania

Celem jest nabycie podstawowej znajomo´sci je¸zyka Python rozwia¸zuja¸c zadanie tworzenia i wy´swietlenia ramki danych odpowiednio do okre´slonego wariantu

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

[ ]:

[128]:
```
#ladowanie biblioteki Pandas
import pandas as pd
#zaimportuj modul pyplot z biblioteki matplotlib
import matplotlib.pyplot as plt
```

[129]:
```
#tworzenie ramki danych ze slownika

data = {'col_1': [3, 2, 1, 0], 'col_2': ['a', 'b', 'c', 'd']}

pd.DataFrame.from_dict(data)
```

[129]:

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

[130]:
```
#zachowanie ramki danych pobranych z pliku w formacie csv (xlsx)

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

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

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

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

[131]:
```
#tworzenie ramki danych z listy list

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

pd.DataFrame(lists_income)
```

[131]:

	0	1	2
0	Adam	Kuba	Robert
1	4500	5500	6500

[132]:
```
#transponowanie (wymieniamy kolumny a wierszy)

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

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

[ ]:

[133]:
```
#wy´swietli´c pierwsze 10 wierszy ramki danych

df.head(10)
```

[133]:

	measure_id	measure_name	location_id	location_name	sex_id	sex_name	age_group_id	age_group_name	rei_id	rei_name	metric_id	metric_name	year_id	
0	5	Prevalence	1	Global	1	Male	8	15 to 19	332	Chewing tobacco	3	Rate	1990	0.
1	5	Prevalence	1	Global	2	Female	8	15 to 19	332	Chewing tobacco	3	Rate	1990	0.
2	5	Prevalence	1	Global	1	Male	8	15 to 19	332	Chewing tobacco	3	Rate	1991	0.
3	5	Prevalence	1	Global	2	Female	8	15 to 19	332	Chewing tobacco	3	Rate	1991	0.
4	5	Prevalence	1	Global	1	Male	8	15 to 19	332	Chewing tobacco	3	Rate	1992	0.
5	5	Prevalence	1	Global	2	Female	8	15 to 19	332	Chewing tobacco	3	Rate	1992	0.
6	5	Prevalence	1	Global	1	Male	8	15 to 19	332	Chewing tobacco	3	Rate	1993	0.
7	5	Prevalence	1	Global	2	Female	8	15 to 19	332	Chewing tobacco	3	Rate	1993	0.
8	5	Prevalence	1	Global	1	Male	8	15 to 19	332	Chewing tobacco	3	Rate	1994	0.
9	5	Prevalence	1	Global	2	Female	8	15 to 19	332	Chewing tobacco	3	Rate	1994	0.

[134]:
```
#wy´swietli´c ostatnie 10 wierszy ramki danych

df.tail(10)
```

[134]:

	measure_id	measure_name	location_id	location_name	sex_id	sex_name	age_group_id	age_group_name	rei_id	rei_name	metric_id	metric_name	year_
350540	5	Prevalence	522	Sudan	3	Both	27	Age standardized	332	Chewing tobacco	3	Rate	20
350541	5	Prevalence	522	Sudan	3	Both	27	Age standardized	332	Chewing tobacco	3	Rate	20
350542	5	Prevalence	522	Sudan	3	Both	27	Age standardized	332	Chewing tobacco	3	Rate	20
350543	5	Prevalence	522	Sudan	3	Both	27	Age standardized	332	Chewing tobacco	3	Rate	20
350544	5	Prevalence	522	Sudan	3	Both	27	Age standardized	332	Chewing tobacco	3	Rate	20
350545	5	Prevalence	522	Sudan	3	Both	27	Age standardized	332	Chewing tobacco	3	Rate	20
350546	5	Prevalence	522	Sudan	3	Both	27	Age standardized	332	Chewing tobacco	3	Rate	20
350547	5	Prevalence	522	Sudan	3	Both	27	Age standardized	332	Chewing tobacco	3	Rate	20
350548	5	Prevalence	522	Sudan	3	Both	27	Age standardized	332	Chewing tobacco	3	Rate	20
350549	5	Prevalence	522	Sudan	3	Both	27	Age standardized	332	Chewing tobacco	3	Rate	20

[135]:
```
#wy´swietli´c informacje, o ramce danych

df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 350550 entries, 0 to 350549
Data columns (total 16 columns):
 #   Column           Non-Null Count   Dtype  
---  ------           --------------   -----  
 0   measure_id       350550 non-null  int64  
 1   measure_name     350550 non-null  object 
 2   location_id      350550 non-null  int64  
 3   location_name    350550 non-null  object 
 4   sex_id           350550 non-null  int64  
 5   sex_name         350550 non-null  object 
 6   age_group_id     350550 non-null  int64  
 7   age_group_name   350550 non-null  object 
 8   rei_id           350550 non-null  int64  
 9   rei_name         350550 non-null  object 
 10  metric_id        350550 non-null  int64  
 11  metric_name      350550 non-null  object 
 12  year_id          350550 non-null  int64  
 13  val              350550 non-null  float64
 14  upper            350550 non-null  float64
 15  lower            350550 non-null  float64
dtypes: float64(3), int64(7), object(6)
memory usage: 42.8+ MB
```

[136]:
```
#wy´swietli´c, ile wierszy i kolumn znajduje sie, w ramce danych

df.shape
```

[136]:
```
(350550, 16)
```

[137]:
```
#wy´swietli´c informacje, statystyczna, o kolumnach liczbowych (warto´sci
#niepowtarzalne, ´srednia, odchylenie standardowe, minimum, kwartyle,
#maksimum)

df.describe()
```

[137]:

	measure_id	location_id	sex_id	age_group_id	rei_id	metric_id	year_id	val	upper	lower
count	350550.0	350550.000000	350550.000000	350550.000000	350550.0	350550.0	350550.000000	350550.000000	350550.000000	350550.000000
mean	5.0	135.639024	2.000000	29.421053	332.0	3.0	2004.500000	0.020179	0.032878	0.011704
std	0.0	98.136414	0.816498	48.993427	0.0	0.0	8.655454	0.049594	0.070631	0.034127
min	5.0	1.000000	1.000000	8.000000	332.0	3.0	1990.000000	0.001051	0.001597	0.000370
25%	5.0	62.000000	1.000000	12.000000	332.0	3.0	1997.000000	0.002300	0.004438	0.001063
50%	5.0	122.000000	2.000000	17.000000	332.0	3.0	2004.500000	0.004869	0.009064	0.002401
75%	5.0	182.000000	3.000000	27.000000	332.0	3.0	2012.000000	0.014729	0.027427	0.007094
max	5.0	522.000000	3.000000	235.000000	332.0	3.0	2019.000000	0.610180	0.773804	0.501187

[138]:
```
#wy´swietli´c informacje, statystyczna, o kolumnach kategoryzowanych (ile
#unikalnych warto´sci, top - jaka jest najpopularniejsza warto´s´c, freq -
#jak cze,sto najpopularniejsza


df.describe(include = 'all')
```

[138]:

	measure_id	measure_name	location_id	location_name	sex_id	sex_name	age_group_id	age_group_name	rei_id	rei_name	metric_id	metric
count	350550.0	350550	350550.000000	350550	350550.000000	350550	350550.000000	350550	350550.0	350550	350550.0	
unique	NaN	1	NaN	205	NaN	3	NaN	19	NaN	1	NaN	
top	NaN	Prevalence	NaN	Global	NaN	Male	NaN	15 to 19	NaN	Chewing tobacco	NaN	
freq	NaN	350550	NaN	1710	NaN	116850	NaN	18450	NaN	350550	NaN	
mean	5.0	NaN	135.639024	NaN	2.000000	NaN	29.421053	NaN	332.0	NaN	3.0	
std	0.0	NaN	98.136414	NaN	0.816498	NaN	48.993427	NaN	0.0	NaN	0.0	
min	5.0	NaN	1.000000	NaN	1.000000	NaN	8.000000	NaN	332.0	NaN	3.0	
25%	5.0	NaN	62.000000	NaN	1.000000	NaN	12.000000	NaN	332.0	NaN	3.0	
50%	5.0	NaN	122.000000	NaN	2.000000	NaN	17.000000	NaN	332.0	NaN	3.0	
75%	5.0	NaN	182.000000	NaN	3.000000	NaN	27.000000	NaN	332.0	NaN	3.0	
max	5.0	NaN	522.000000	NaN	3.000000	NaN	235.000000	NaN	332.0	NaN	3.0	

[139]:
```
#usuna,´c brakuja,ce warto´sci w ramce danych

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

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

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

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

[350550 rows x 16 columns]
```

[140]:
```
#przedstawi´c wyb´or wierszy i kolumny u´zywaja,c nazw oraz indeks´ow na
#r´o´zne sposoby

df["measure_name"] # zmienic nazwe po swoj zbior
```

[140]:
```
0         Prevalence
1         Prevalence
2         Prevalence
3         Prevalence
4         Prevalence
             ...
350545    Prevalence
350546    Prevalence
350547    Prevalence
350548    Prevalence
350549    Prevalence
Name: measure_name, Length: 350550, dtype: object
```

[141]:
```
df.measure_name # zmienic nazwe pod swoj zbior
```

[141]:
```
0         Prevalence
1         Prevalence
2         Prevalence
3         Prevalence
4         Prevalence
             ...
350545    Prevalence
350546    Prevalence
350547    Prevalence
350548    Prevalence
350549    Prevalence
Name: measure_name, Length: 350550, dtype: object
```

[142]:
```
df[["measure_name","age_group_name","year_id"]] # wybór kilku kolumn jednocześnie zmienic pod swoj zbior
```

[142]:

	measure_name	age_group_name	year_id
0	Prevalence	15 to 19	1990
1	Prevalence	15 to 19	1990
2	Prevalence	15 to 19	1991
3	Prevalence	15 to 19	1991
4	Prevalence	15 to 19	1992
...	...	...	...
350545	Prevalence	Age standardized	2015
350546	Prevalence	Age standardized	2016
350547	Prevalence	Age standardized	2017
350548	Prevalence	Age standardized	2018
350549	Prevalence	Age standardized	2019

[143]:
```
df.loc[:, "location_name":"val"] # wybierz wszystkie wiersze i kolumny od „Location" do „val" zmienic pod swoj zbior
```

[143]:

	location_name	sex_id	sex_name	age_group_id	age_group_name	rei_id	rei_name	metric_id	metric_name	year_id	val
0	Global	1	Male	8	15 to 19	332	Chewing tobacco	3	Rate	1990	0.038740
1	Global	2	Female	8	15 to 19	332	Chewing tobacco	3	Rate	1990	0.011356
2	Global	1	Male	8	15 to 19	332	Chewing tobacco	3	Rate	1991	0.039253
3	Global	2	Female	8	15 to 19	332	Chewing tobacco	3	Rate	1991	0.011516
4	Global	1	Male	8	15 to 19	332	Chewing tobacco	3	Rate	1992	0.039863
...	...	...	...	...	...	...	...	...	...	...	...
350545	Sudan	3	Both	27	Age standardized	332	Chewing tobacco	3	Rate	2015	0.029518
350546	Sudan	3	Both	27	Age standardized	332	Chewing tobacco	3	Rate	2016	0.029855
350547	Sudan	3	Both	27	Age standardized	332	Chewing tobacco	3	Rate	2017	0.029768
350548	Sudan	3	Both	27	Age standardized	332	Chewing tobacco	3	Rate	2018	0.029760
350549	Sudan	3	Both	27	Age standardized	332	Chewing tobacco	3	Rate	2019	0.029752

350550 rows × 11 columns

[144]:
```
df.loc[100:110, "location_name":"val"] #wybierz wiersze od 100-110 i kolumny od locationa_name do val
```

[144]:

	location_name	sex_id	sex_name	age_group_id	age_group_name	rei_id	rei_name	metric_id	metric_name	year_id	val
100	Global	1	Male	9	20 to 24	332	Chewing tobacco	3	Rate	1995	0.055556
101	Global	2	Female	9	20 to 24	332	Chewing tobacco	3	Rate	1995	0.012084
102	Global	1	Male	9	20 to 24	332	Chewing tobacco	3	Rate	1996	0.056900
103	Global	2	Female	9	20 to 24	332	Chewing tobacco	3	Rate	1996	0.012324
104	Global	1	Male	9	20 to 24	332	Chewing tobacco	3	Rate	1997	0.058382
105	Global	2	Female	9	20 to 24	332	Chewing tobacco	3	Rate	1997	0.012566
106	Global	1	Male	9	20 to 24	332	Chewing tobacco	3	Rate	1998	0.059939
107	Global	2	Female	9	20 to 24	332	Chewing tobacco	3	Rate	1998	0.012805
108	Global	1	Male	9	20 to 24	332	Chewing tobacco	3	Rate	1999	0.061506
109	Global	2	Female	9	20 to 24	332	Chewing tobacco	3	Rate	1999	0.013047
110	Global	1	Male	9	20 to 24	332	Chewing tobacco	3	Rate	2000	0.062994

[145]:
```
df.iloc[100:110, 0:3] #wybierz wiersze od 100-110 i kolumny od 0-2
```

[145]:

	measure_id	measure_name	location_id
100	5	Prevalence	1
101	5	Prevalence	1
102	5	Prevalence	1
103	5	Prevalence	1
104	5	Prevalence	1
105	5	Prevalence	1
106	5	Prevalence	1
107	5	Prevalence	1
108	5	Prevalence	1
109	5	Prevalence	1

[146]:
```
#przedstawi´c wyb´or wierszy z ramki danych pod warunkiem odno´snie
#okre´slonej warto´sci kolumny

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

[146]:

	measure_id	measure_name	location_id	location_name	sex_id	sex_name	age_group_id	age_group_name	rei_id	rei_name	metric_id	metric_name	year_i
60	5	Prevalence	1	Global	3	Both	8	15 to 19	332	Chewing tobacco	3	Rate	19
61	5	Prevalence	1	Global	3	Both	8	15 to 19	332	Chewing tobacco	3	Rate	19
62	5	Prevalence	1	Global	3	Both	8	15 to 19	332	Chewing tobacco	3	Rate	19
63	5	Prevalence	1	Global	3	Both	8	15 to 19	332	Chewing tobacco	3	Rate	19
64	5	Prevalence	1	Global	3	Both	8	15 to 19	332	Chewing tobacco	3	Rate	19
...	...	...	...	...	...	...	...	...	...	...	...	...	
350545	5	Prevalence	522	Sudan	3	Both	27	Age standardized	332	Chewing tobacco	3	Rate	20
350546	5	Prevalence	522	Sudan	3	Both	27	Age standardized	332	Chewing tobacco	3	Rate	20

[147]:
```
#przedstawi´c wyb´or wierszy z ramki danych pod warunkiem spełnienia
#kilku warunk´ow jednocze´snie

cardio = df[(df["sex_name"] == "Both")&(df["year_id"] == 2018) & (df["age_group_id"] <= 29)]
cardio
```

[147]:

	measure_id	measure_name	location_id	location_name	sex_id	sex_name	age_group_id	age_group_name	rei_id	rei_name	metric_id	metric_name	year_
88	5	Prevalence	1	Global	3	Both	8	15 to 19	332	Chewing tobacco	3	Rate	20
178	5	Prevalence	1	Global	3	Both	9	20 to 24	332	Chewing tobacco	3	Rate	20
268	5	Prevalence	1	Global	3	Both	10	25 to 29	332	Chewing tobacco	3	Rate	20
358	5	Prevalence	1	Global	3	Both	11	30 to 34	332	Chewing tobacco	3	Rate	20
448	5	Prevalence	1	Global	3	Both	12	35 to 39	332	Chewing tobacco	3	Rate	20
…	…	…	…	…	…	…	…	…	…	…	…	…	
349828	5	Prevalence	522	Sudan	3	Both	18	65 to 69	332	Chewing tobacco	3	Rate	20
349918	5	Prevalence	522	Sudan	3	Both	19	70 to 74	332	Chewing tobacco	3	Rate	20
350008	5	Prevalence	522	Sudan	3	Both	20	75 to 79	332	Chewing tobacco	3	Rate	20
350098	5	Prevalence	522	Sudan	3	Both	22	All Ages	332	Chewing tobacco	3	Rate	20
350548	5	Prevalence	522	Sudan	3	Both	27	Age standardized	332	Chewing tobacco	3	Rate	20

3075 rows × 16 columns

[148]:
```
# wybra´c wiersze kt´ore zawieraja, w kolumnie kategoryzowanej okre´slone slowo

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

[148]:

	measure_id	measure_name	location_id	location_name	sex_id	sex_name	age_group_id	age_group_name	rei_id	rei_name	metric_id	metric_name	year_
30780	5	Prevalence	25	Micronesia (Federated States of)	1	Male	8	15 to 19	332	Chewing tobacco	3	Rate	19
30781	5	Prevalence	25	Micronesia (Federated States of)	2	Female	8	15 to 19	332	Chewing tobacco	3	Rate	19
30782	5	Prevalence	25	Micronesia (Federated States of)	1	Male	8	15 to 19	332	Chewing tobacco	3	Rate	19
30783	5	Prevalence	25	Micronesia (Federated States of)	2	Female	8	15 to 19	332	Chewing tobacco	3	Rate	19
30784	5	Prevalence	25	Micronesia (Federated States of)	1	Male	8	15 to 19	332	Chewing tobacco	3	Rate	19
…	…	…	…	…	…	…	…	…	…	…	…	…	
347125	5	Prevalence	422	United States Virgin Islands	3	Both	27	Age standardized	332	Chewing tobacco	3	Rate	20
347126	5	Prevalence	422	United States Virgin Islands	3	Both	27	Age standardized	332	Chewing tobacco	3	Rate	20
347127	5	Prevalence	422	United States Virgin Islands	3	Both	27	Age standardized	332	Chewing tobacco	3	Rate	20
347128	5	Prevalence	422	United States Virgin Islands	3	Both	27	Age standardized	332	Chewing tobacco	3	Rate	20
347129	5	Prevalence	422	United States Virgin Islands	3	Both	27	Age standardized	332	Chewing tobacco	3	Rate	20

5130 rows × 16 columns

[149]:
```
# wybra´c wiersze kt´ore nie zawieraja, w kolumnie kategoryzowanej okre´slone
# slowo

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

[149]:

	measure_id	measure_name	location_id	location_name	sex_id	sex_name	age_group_id	age_group_name	rei_id	rei_name	metric_id	metric_name	year_
0	5	Prevalence	1	Global	1	Male	8	15 to 19	332	Chewing tobacco	3	Rate	19
1	5	Prevalence	1	Global	2	Female	8	15 to 19	332	Chewing tobacco	3	Rate	19
2	5	Prevalence	1	Global	1	Male	8	15 to 19	332	Chewing tobacco	3	Rate	19
3	5	Prevalence	1	Global	2	Female	8	15 to 19	332	Chewing tobacco	3	Rate	19
4	5	Prevalence	1	Global	1	Male	8	15 to 19	332	Chewing tobacco	3	Rate	19
...	...	...	...	...	...	...	...	...	...	...	...	...	
350545	5	Prevalence	522	Sudan	3	Both	27	Age standardized	332	Chewing tobacco	3	Rate	20
350546	5	Prevalence	522	Sudan	3	Both	27	Age standardized	332	Chewing tobacco	3	Rate	20
350547	5	Prevalence	522	Sudan	3	Both	27	Age standardized	332	Chewing tobacco	3	Rate	20
350548	5	Prevalence	522	Sudan	3	Both	27	Age standardized	332	Chewing tobacco	3	Rate	20
350549	5	Prevalence	522	Sudan	3	Both	27	Age standardized	332	Chewing tobacco	3	Rate	20

345420 rows × 16 columns

[150]:

```
#utwórz kolumne, na podstawie istnieja,cyn

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

[150]:

	measure_id	measure_name	location_id	location_name	sex_id	sex_name	age_group_id	age_group_name	rei_id	rei_name	metric_id	metric_name	year_
0	5	Prevalence	1	Global	1	Male	8	15 to 19	332	Chewing tobacco	3	Rate	19
1	5	Prevalence	1	Global	2	Female	8	15 to 19	332	Chewing tobacco	3	Rate	19
2	5	Prevalence	1	Global	1	Male	8	15 to 19	332	Chewing tobacco	3	Rate	19
3	5	Prevalence	1	Global	2	Female	8	15 to 19	332	Chewing tobacco	3	Rate	19
4	5	Prevalence	1	Global	1	Male	8	15 to 19	332	Chewing tobacco	3	Rate	19
...	...	...	...	...	...	...	...	...	...	...	...	...	
350545	5	Prevalence	522	Sudan	3	Both	27	Age standardized	332	Chewing tobacco	3	Rate	20
350546	5	Prevalence	522	Sudan	3	Both	27	Age standardized	332	Chewing tobacco	3	Rate	20
350547	5	Prevalence	522	Sudan	3	Both	27	Age standardized	332	Chewing tobacco	3	Rate	20
350548	5	Prevalence	522	Sudan	3	Both	27	Age standardized	332	Chewing tobacco	3	Rate	20
350549	5	Prevalence	522	Sudan	3	Both	27	Age standardized	332	Chewing tobacco	3	Rate	20

350550 rows × 17 columns

[151]:

```
#usuń kolumne

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

[151]:

	measure_id	measure_name	location_id	location_name	sex_id	sex_name	age_group_name	rei_id	rei_name	metric_id	metric_name	year_id	val
0	5	Prevalence	1	Global	1	Male	15 to 19	332	Chewing tobacco	3	Rate	1990	0.038740
1	5	Prevalence	1	Global	2	Female	15 to 19	332	Chewing tobacco	3	Rate	1990	0.011356
2	5	Prevalence	1	Global	1	Male	15 to 19	332	Chewing tobacco	3	Rate	1991	0.039253
3	5	Prevalence	1	Global	2	Female	15 to 19	332	Chewing tobacco	3	Rate	1991	0.011516
4	5	Prevalence	1	Global	1	Male	15 to 19	332	Chewing tobacco	3	Rate	1992	0.039863
...	...	...	...	...	...	...	...	...	...	...	...	...	...
350545	5	Prevalence	522	Sudan	3	Both	Age standardized	332	Chewing tobacco	3	Rate	2015	0.029518
									Chewing				

[152]:
```
#zmie´n nazwe , kolumny

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

[152]:

	measure_id	measure_name	location_id	location_name	sex_id	sex	age_group_name	rei_id	rei_name	metric_id	metric_name	year_id	val	u
0	5	Prevalence	1	Global	1	Male	15 to 19	332	Chewing tobacco	3	Rate	1990	0.038740	0.05
1	5	Prevalence	1	Global	2	Female	15 to 19	332	Chewing tobacco	3	Rate	1990	0.011356	0.01
2	5	Prevalence	1	Global	1	Male	15 to 19	332	Chewing tobacco	3	Rate	1991	0.039253	0.05
3	5	Prevalence	1	Global	2	Female	15 to 19	332	Chewing tobacco	3	Rate	1991	0.011516	0.01
4	5	Prevalence	1	Global	1	Male	15 to 19	332	Chewing tobacco	3	Rate	1992	0.039863	0.05
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
350545	5	Prevalence	522	Sudan	3	Both	Age standardized	332	Chewing tobacco	3	Rate	2015	0.029518	0.03
350546	5	Prevalence	522	Sudan	3	Both	Age standardized	332	Chewing tobacco	3	Rate	2016	0.029855	0.03
350547	5	Prevalence	522	Sudan	3	Both	Age standardized	332	Chewing tobacco	3	Rate	2017	0.029768	0.03
350548	5	Prevalence	522	Sudan	3	Both	Age standardized	332	Chewing tobacco	3	Rate	2018	0.029760	0.03
350549	5	Prevalence	522	Sudan	3	Both	Age standardized	332	Chewing tobacco	3	Rate	2019	0.029752	0.03

350550 rows × 16 columns

[153]:
```
df
```

[153]:

	measure_id	measure_name	location_id	location_name	sex_id	sex	age_group_name	rei_id	rei_name	metric_id	metric_name	year_id	val	u
0	5	Prevalence	1	Global	1	Male	15 to 19	332	Chewing tobacco	3	Rate	1990	0.038740	0.05
1	5	Prevalence	1	Global	2	Female	15 to 19	332	Chewing tobacco	3	Rate	1990	0.011356	0.01
2	5	Prevalence	1	Global	1	Male	15 to 19	332	Chewing tobacco	3	Rate	1991	0.039253	0.05
3	5	Prevalence	1	Global	2	Female	15 to 19	332	Chewing tobacco	3	Rate	1991	0.011516	0.01
4	5	Prevalence	1	Global	1	Male	15 to 19	332	Chewing tobacco	3	Rate	1992	0.039863	0.05
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
350545	5	Prevalence	522	Sudan	3	Both	Age standardized	332	Chewing tobacco	3	Rate	2015	0.029518	0.03
350546	5	Prevalence	522	Sudan	3	Both	Age standardized	332	Chewing tobacco	3	Rate	2016	0.029855	0.03
350547	5	Prevalence	522	Sudan	3	Both	Age standardized	332	Chewing tobacco	3	Rate	2017	0.029768	0.03
350548	5	Prevalence	522	Sudan	3	Both	Age standardized	332	Chewing tobacco	3	Rate	2018	0.029760	0.03
350549	5	Prevalence	522	Sudan	3	Both	Age standardized	332	Chewing tobacco	3	Rate	2019	0.029752	0.03

350550 rows × 16 columns

[154]:
```
#zachowaj ramke, danych jako plik csv na komputerze

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

[155]:
```
df
```

[155]:

	measure_id	measure_name	location_id	location_name	sex_id	sex	age_group_name	rei_id	rei_name	metric_id	metric_name	year_id	val	u
0	5	Prevalence	1	Global	1	Male	15 to 19	332	Chewing tobacco	3	Rate	1990	0.038740	0.05
1	5	Prevalence	1	Global	2	Female	15 to 19	332	Chewing tobacco	3	Rate	1990	0.011356	0.01
2	5	Prevalence	1	Global	1	Male	15 to 19	332	Chewing tobacco	3	Rate	1991	0.039253	0.05
3	5	Prevalence	1	Global	2	Female	15 to 19	332	Chewing tobacco	3	Rate	1991	0.011516	0.01
4	5	Prevalence	1	Global	1	Male	15 to 19	332	Chewing tobacco	3	Rate	1992	0.039863	0.05
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
350545	5	Prevalence	522	Sudan	3	Both	Age standardized	332	Chewing tobacco	3	Rate	2015	0.029518	0.03
350546	5	Prevalence	522	Sudan	3	Both	Age standardized	332	Chewing tobacco	3	Rate	2016	0.029855	0.03
350547	5	Prevalence	522	Sudan	3	Both	Age standardized	332	Chewing tobacco	3	Rate	2017	0.029768	0.03
350548	5	Prevalence	522	Sudan	3	Both	Age standardized	332	Chewing tobacco	3	Rate	2018	0.029760	0.03
350549	5	Prevalence	522	Sudan	3	Both	Age standardized	332	Chewing tobacco	3	Rate	2019	0.029752	0.03

350550 rows × 16 columns

[156]:
```
#wy´swietli´c ´srednia (maksymalna,, minimalna,) warto´s´c z jednej kolumny

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

```
2004.5
2019
1990
```

[157]:
```
#wy´swietli´c liczbe, wierszy

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

[157]:
```
350550
```

[158]:
```
#wy´swietli´c warto´sci unikatowe w kolumnie

df['sex'].unique() # wartości unikatowe
```

[158]:
```
array(['Male', 'Female', 'Both'], dtype=object)
```

[159]:
```
#wy´swietli´c liczby rekord´ow odpowiadaja,cych do warto´sci

df['sex'].value_counts() # liczba rekordów pasujących do unikalnych wartości
```

[159]:
```
sex
Male      116850
Female    116850
Both      116850
Name: count, dtype: int64
```

[160]:
```
#sortowanie wierszy ramki danych wedlug warto´sci okre´slonej kolumny
#(maleja,co, rosna,co)

df.sort_values(['val'], ascending = True) # sortowanie rosnąco
```

[160]:

	measure_id	measure_name	location_id	location_name	sex_id	sex	age_group_name	rei_id	rei_name	metric_id	metric_name	year_id	val	u
340439	5	Prevalence	396	San Marino	2	Female	20 to 24	332	Chewing tobacco	3	Rate	2019	0.001051	0.00
340437	5	Prevalence	396	San Marino	2	Female	20 to 24	332	Chewing tobacco	3	Rate	2018	0.001051	0.00
340429	5	Prevalence	396	San Marino	2	Female	20 to 24	332	Chewing tobacco	3	Rate	2014	0.001051	0.00
340431	5	Prevalence	396	San Marino	2	Female	20 to 24	332	Chewing tobacco	3	Rate	2015	0.001053	0.00
340435	5	Prevalence	396	San Marino	2	Female	20 to 24	332	Chewing tobacco	3	Rate	2017	0.001054	0.00
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
8305	5	Prevalence	10	Cambodia	2	Female	90 to 94	332	Chewing tobacco	3	Rate	2002	0.610154	0.71
8217	5	Prevalence	10	Cambodia	2	Female	85 to 89	332	Chewing tobacco	3	Rate	2003	0.610180	0.70
8397	5	Prevalence	10	Cambodia	2	Female	95 plus	332	Chewing tobacco	3	Rate	2003	0.610180	0.70
8127	5	Prevalence	10	Cambodia	2	Female	80 to 84	332	Chewing tobacco	3	Rate	2003	0.610180	0.70
8307	5	Prevalence	10	Cambodia	2	Female	90 to 94	332	Chewing tobacco	3	Rate	2003	0.610180	0.70

350550 rows × 16 columns

[161]:
```
df.sort_values(['val'], ascending =  False) # sortowanie malejąco
```

[161]:

	measure_id	measure_name	location_id	location_name	sex_id	sex	age_group_name	rei_id	rei_name	metric_id	metric_name	year_id	val	u
8217	5	Prevalence	10	Cambodia	2	Female	85 to 89	332	Chewing tobacco	3	Rate	2003	0.610180	0.70
8397	5	Prevalence	10	Cambodia	2	Female	95 plus	332	Chewing tobacco	3	Rate	2003	0.610180	0.70
8127	5	Prevalence	10	Cambodia	2	Female	80 to 84	332	Chewing tobacco	3	Rate	2003	0.610180	0.70
8307	5	Prevalence	10	Cambodia	2	Female	90 to 94	332	Chewing tobacco	3	Rate	2003	0.610180	0.70
8305	5	Prevalence	10	Cambodia	2	Female	90 to 94	332	Chewing tobacco	3	Rate	2002	0.610154	0.71
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
340435	5	Prevalence	396	San Marino	2	Female	20 to 24	332	Chewing tobacco	3	Rate	2017	0.001054	0.00
340431	5	Prevalence	396	San Marino	2	Female	20 to 24	332	Chewing tobacco	3	Rate	2015	0.001053	0.00
340429	5	Prevalence	396	San Marino	2	Female	20 to 24	332	Chewing tobacco	3	Rate	2014	0.001051	0.00
340437	5	Prevalence	396	San Marino	2	Female	20 to 24	332	Chewing tobacco	3	Rate	2018	0.001051	0.00
340439	5	Prevalence	396	San Marino	2	Female	20 to 24	332	Chewing tobacco	3	Rate	2019	0.001051	0.00

350550 rows × 16 columns

[162]:
```
df.sort_values(['val'], ascending =  True).head(10) # 10 Największych
```

[162]:

	measure_id	measure_name	location_id	location_name	sex_id	sex	age_group_name	rei_id	rei_name	metric_id	metric_name	year_id	val	u
340439	5	Prevalence	396	San Marino	2	Female	20 to 24	332	Chewing tobacco	3	Rate	2019	0.001051	0.00
340437	5	Prevalence	396	San Marino	2	Female	20 to 24	332	Chewing tobacco	3	Rate	2018	0.001051	0.00
340429	5	Prevalence	396	San Marino	2	Female	20 to 24	332	Chewing tobacco	3	Rate	2014	0.001051	0.00
340431	5	Prevalence	396	San Marino	2	Female	20 to 24	332	Chewing tobacco	3	Rate	2015	0.001053	0.00
340435	5	Prevalence	396	San Marino	2	Female	20 to 24	332	Chewing tobacco	3	Rate	2017	0.001054	0.00
340433	5	Prevalence	396	San Marino	2	Female	20 to 24	332	Chewing tobacco	3	Rate	2016	0.001054	0.00
340427	5	Prevalence	396	San Marino	2	Female	20 to 24	332	Chewing tobacco	3	Rate	2013	0.001055	0.00
340425	5	Prevalence	396	San Marino	2	Female	20 to 24	332	Chewing tobacco	3	Rate	2012	0.001057	0.00
340423	5	Prevalence	396	San Marino	2	Female	20 to 24	332	Chewing tobacco	3	Rate	2011	0.001060	0.00
340421	5	Prevalence	396	San Marino	2	Female	20 to 24	332	Chewing tobacco	3	Rate	2010	0.001063	0.00

[163]:
```
df.sort_values(['val'], ascending =  False).head(10) # 10 Najmniejszych
```

[163]:

	measure_id	measure_name	location_id	location_name	sex_id	sex	age_group_name	rei_id	rei_name	metric_id	metric_name	year_id	val	upp
8217	5	Prevalence	10	Cambodia	2	Female	85 to 89	332	Chewing tobacco	3	Rate	2003	0.610180	0.7088
8397	5	Prevalence	10	Cambodia	2	Female	95 plus	332	Chewing tobacco	3	Rate	2003	0.610180	0.7088
8127	5	Prevalence	10	Cambodia	2	Female	80 to 84	332	Chewing tobacco	3	Rate	2003	0.610180	0.7088
8307	5	Prevalence	10	Cambodia	2	Female	90 to 94	332	Chewing tobacco	3	Rate	2003	0.610180	0.7088
8305	5	Prevalence	10	Cambodia	2	Female	90 to 94	332	Chewing tobacco	3	Rate	2002	0.610154	0.7164
8395	5	Prevalence	10	Cambodia	2	Female	95 plus	332	Chewing tobacco	3	Rate	2002	0.610154	0.7164
8215	5	Prevalence	10	Cambodia	2	Female	85 to 89	332	Chewing tobacco	3	Rate	2002	0.610154	0.7164
8125	5	Prevalence	10	Cambodia	2	Female	80 to 84	332	Chewing tobacco	3	Rate	2002	0.610154	0.7164
8129	5	Prevalence	10	Cambodia	2	Female	80 to 84	332	Chewing tobacco	3	Rate	2004	0.609981	0.7061
8309	5	Prevalence	10	Cambodia	2	Female	90 to 94	332	Chewing tobacco	3	Rate	2004	0.609981	0.7061

[164]:
```
#wy´swietli´c wierszy dla 10 najwie¸kszych warto´sci okre´slonej kolumny
#pod warunkiem okre´slonych warto´sci innej kolumny

df[(df['location_name'].isin(['Sudan','Poland','Hungary']))].nlargest(10,'val')
```

[164]:

	measure_id	measure_name	location_id	location_name	sex_id	sex	age_group_name	rei_id	rei_name	metric_id	metric_name	year_id	val	upp
349432	5	Prevalence	522	Sudan	1	Male	45 to 49	332	Chewing tobacco	3	Rate	2016	0.070387	0.1214
349438	5	Prevalence	522	Sudan	1	Male	45 to 49	332	Chewing tobacco	3	Rate	2019	0.070178	0.1200
349434	5	Prevalence	522	Sudan	1	Male	45 to 49	332	Chewing tobacco	3	Rate	2017	0.070171	0.1212
349436	5	Prevalence	522	Sudan	1	Male	45 to 49	332	Chewing tobacco	3	Rate	2018	0.070170	0.1208
349430	5	Prevalence	522	Sudan	1	Male	45 to 49	332	Chewing tobacco	3	Rate	2015	0.069704	0.1221
349428	5	Prevalence	522	Sudan	1	Male	45 to 49	332	Chewing tobacco	3	Rate	2014	0.069080	0.1214
349342	5	Prevalence	522	Sudan	1	Male	40 to 44	332	Chewing tobacco	3	Rate	2016	0.068875	0.1193
349348	5	Prevalence	522	Sudan	1	Male	40 to 44	332	Chewing tobacco	3	Rate	2019	0.068630	0.1205
349344	5	Prevalence	522	Sudan	1	Male	40 to 44	332	Chewing tobacco	3	Rate	2017	0.068616	0.1212
349346	5	Prevalence	522	Sudan	1	Male	40 to 44	332	Chewing tobacco	3	Rate	2018	0.068606	0.1210

[165]:
```
#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
```

[165]:

location_name	sex	val	upper	lower
Afghanistan	Both	0.036491	0.057880	0.021572
	Female	0.005077	0.010820	0.002041
	Male	0.068753	0.111568	0.039311
Albania	Both	0.005864	0.009785	0.003375
	Female	0.003130	0.006646	0.001277
...	...	...	...	...
Zambia	Female	0.014409	0.030814	0.005678
	Male	0.008172	0.015062	0.003965
Zimbabwe	Both	0.008677	0.017079	0.003970
	Female	0.010854	0.023365	0.004162
	Male	0.003754	0.006850	0.001870

615 rows × 3 columns

[166]:
```
#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']})
df_new
```

[166]:

		val	upper		sex_id	
		mean	median	count	median	count
location_name	sex					
Afghanistan	Both	0.036491	0.055316	570	3.0	570
	Female	0.005077	0.010201	570	2.0	570
	Male	0.068753	0.108540	570	1.0	570
Albania	Both	0.005864	0.009446	570	3.0	570
	Female	0.003130	0.005227	570	2.0	570
...	...	...	...	...	...	...
Zambia	Female	0.014409	0.027697	570	2.0	570
	Male	0.008172	0.011570	570	1.0	570
Zimbabwe	Both	0.008677	0.007834	570	3.0	570
	Female	0.010854	0.011297	570	2.0	570
	Male	0.003754	0.006515	570	1.0	570

615 rows × 5 columns

[167]:
```
#wyświetlić nazwy kolumn indeksu złożonego

df_new.columns
```

[167]:
```
MultiIndex([(   'val',    'mean'),
            ( 'upper',  'median'),
            ( 'upper',   'count'),
            ('sex_id',  'median'),
            ('sex_id',   'count')],
           )
```

[168]:
```
#sortowa´c kolumne, indeksu zlo´zonego

df_new['upper']['median'].sort_values(ascending = True)
```

[168]:
```
location_name  sex
France         Both      0.002428
Spain          Both      0.002435
Netherlands    Both      0.002438
Portugal       Both      0.002450
Luxembourg     Both      0.002460
                            ...
India          Male      0.386419
Bhutan         Male      0.419482
Bangladesh     Both      0.425999
Nepal          Male      0.503675
Bangladesh     Female    0.546821
Name: median, Length: 615, dtype: float64
```

[169]:
```
#stworzy´c tabele, przystawna, (pivot table) na podstawie ramki danych

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

[169]:

location_name	sex	Both	Female	Male
Afghanistan		0.036491	0.005077	0.068753
Albania		0.005864	0.003130	0.008647
Algeria		0.043335	0.006325	0.078118
American Samoa		0.023841	0.014986	0.032602
Andorra		0.001846	0.001504	0.002224
…		…	…	…
Venezuela (Bolivarian Republic of)		0.011944	0.004963	0.019783
Viet Nam		0.026296	0.034622	0.008547
Yemen		0.105273	0.075236	0.132851
Zambia		0.011422	0.014409	0.008172
Zimbabwe		0.008677	0.010854	0.003754

205 rows × 3 columns

[170]:
```
#wy´swietli´c indeksy i kolumny tabeli przystawnej

print(df_pivot.index)
print(df_pivot.columns)
```

```
Index(['Afghanistan', 'Albania', 'Algeria', 'American Samoa', 'Andorra',
       'Angola', 'Antigua and Barbuda', 'Argentina', 'Armenia', 'Australia',
       ...
       'United States Virgin Islands', 'United States of America', 'Uruguay',
       'Uzbekistan', 'Vanuatu', 'Venezuela (Bolivarian Republic of)',
       'Viet Nam', 'Yemen', 'Zambia', 'Zimbabwe'],
      dtype='object', name='location_name', length=205)
Index(['Both', 'Female', 'Male'], dtype='object', name='sex')
```

[171]:
```
#utw´orz indeks zlo´zony tabeli przystawnej i wy´swietl go

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

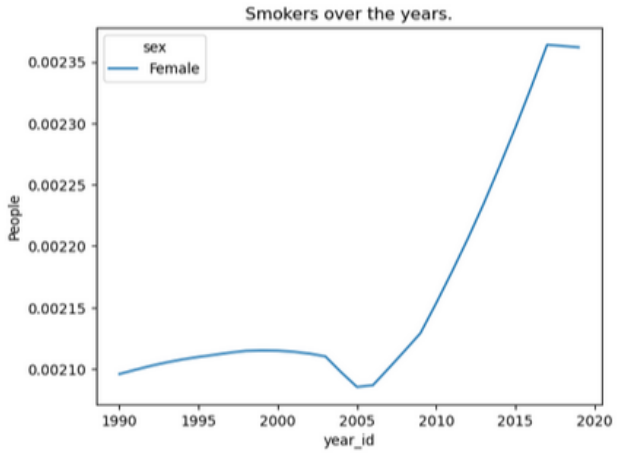
[171]:

location_name	year_id	sex	Both	Female	Male
Afghanistan	1990		0.032164	0.005649	0.062808
	1991		0.032468	0.005646	0.062973
	1992		0.032946	0.005642	0.063175
	1993		0.033264	0.005640	0.063388
	1994		0.033423	0.005639	0.063608
…	…		…	…	…
Zimbabwe	2015		0.009535	0.012373	0.003522
	2016		0.009718	0.012664	0.003525

```
df[(df['location_name'] == 'Poland') & (df['sex'] == 'Female')].pivot_table(values='val', index='year_id', columns='sex', aggfunc='mean',
                                    fill_value=None, margins=False, dropna=True).plot(kind = 'line')

plt.ylabel('People') # etykieta osi y
plt.title('Smokers over the years.') # tytuł wykresu
```

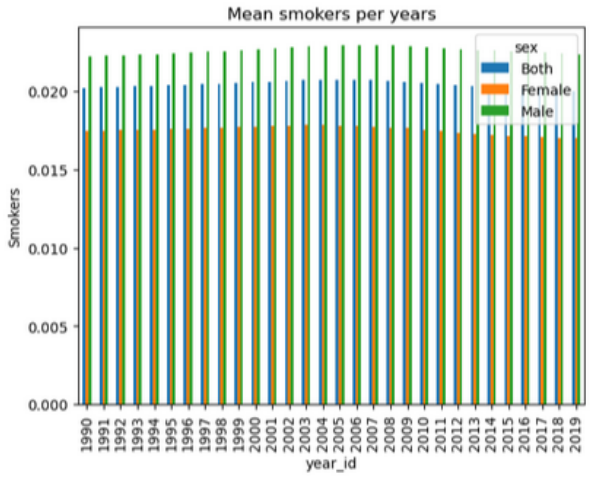
[172]:
```
Text(0.5, 1.0, 'Smokers over the years.')
```



[173]:
```
#narysować histogram na podstawie wartości kolumny

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

[173]:
```
Text(0.5, 1.0, 'Mean smokers per years')
```



[175]:
```
#przedstawić sposoby łączenia ramek danych za pomocą, metod merge i
#concat

df2 = pd.read_csv('IHME_GBD_2019_CHEWING_TOB_1990_2019_CIG_PC_Y2021M05D27.CSV')
df2
```

[175]:

	measure_name	location_id	location_name	sex_id	sex_name	age_group_id	age_group_name	year_id	val	upper	lower
0	Cigarette-Equivalents Per Capita	1	Global	3	Both	29	15+ years	1990	1484.256502	1531.563739	1436.151878
1	Cigarette-Equivalents Per Capita	1	Global	3	Both	29	15+ years	2019	1113.754663	1161.263946	1069.765828
2	Cigarette-Equivalents Per Capita	4	Southeast Asia, East Asia, and Oceania	3	Both	29	15+ years	1990	1827.374739	1959.359086	1692.900863
3	Cigarette-Equivalents Per Capita	4	Southeast Asia, East Asia, and Oceania	3	Both	29	15+ years	2019	1778.846098	1927.560165	1640.645875
4	Cigarette-Equivalents Per Capita	5	East Asia	3	Both	29	15+ years	1990	2089.743405	2267.199999	1908.301510
...	...	...	...	...	...	...	...	...	...	...	...
461	Cigarette-Equivalents Per Capita	422	United States Virgin Islands	3	Both	29	15+ years	2019	648.023999	821.503370	497.645622
462	Cigarette-Equivalents Per Capita	435	South Sudan	3	Both	29	15+ years	1990	337.087376	431.023296	256.275436
463	Cigarette-Equivalents Per Capita	435	South Sudan	3	Both	29	15+ years	2019	317.318526	410.888023	241.957923
464	Cigarette-Equivalents Per Capita	522	Sudan	3	Both	29	15+ years	1990	170.613365	200.058577	142.371142
465	Cigarette-Equivalents Per Capita	522	Sudan	3	Both	29	15+ years	2019	227.024745	279.919615	180.773129

466 rows × 11 columns

[176]:
```
df
```

[176]:

	measure_id	measure_name	location_id	location_name	sex_id	sex	age_group_name	rei_id	rei_name	metric_id	metric_name	year_id	val	u
0	5	Prevalence	1	Global	1	Male	15 to 19	332	Chewing tobacco	3	Rate	1990	0.038740	0.05
1	5	Prevalence	1	Global	2	Female	15 to 19	332	Chewing tobacco	3	Rate	1990	0.011356	0.01
2	5	Prevalence	1	Global	1	Male	15 to 19	332	Chewing tobacco	3	Rate	1991	0.039253	0.05
3	5	Prevalence	1	Global	2	Female	15 to 19	332	Chewing tobacco	3	Rate	1991	0.011516	0.01
4	5	Prevalence	1	Global	1	Male	15 to 19	332	Chewing tobacco	3	Rate	1992	0.039863	0.05
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
350545	5	Prevalence	522	Sudan	3	Both	Age standardized	332	Chewing tobacco	3	Rate	2015	0.029518	0.03
350546	5	Prevalence	522	Sudan	3	Both	Age standardized	332	Chewing tobacco	3	Rate	2016	0.029855	0.03
350547	5	Prevalence	522	Sudan	3	Both	Age standardized	332	Chewing tobacco	3	Rate	2017	0.029768	0.03
350548	5	Prevalence	522	Sudan	3	Both	Age standardized	332	Chewing tobacco	3	Rate	2018	0.029760	0.03
350549	5	Prevalence	522	Sudan	3	Both	Age standardized	332	Chewing tobacco	3	Rate	2019	0.029752	0.03

350550 rows × 16 columns

[177]:
```
df2.rename(columns = {'val': 'val_Cigarette-Equivalents Per Capita', 'upper': 'upper_Cigarette-Equivalents Per Capita', 'lower':'deaths_lower_Cigarette
df.drop('sex_id', axis = 1)
```

[177]:

	measure_id	measure_name	location_id	location_name	sex	age_group_name	rei_id	rei_name	metric_id	metric_name	year_id	val	upper	
0	5	Prevalence	1	Global	Male	15 to 19	332	Chewing tobacco	3	Rate	1990	0.038740	0.055586	0.
1	5	Prevalence	1	Global	Female	15 to 19	332	Chewing tobacco	3	Rate	1990	0.011356	0.017594	0.
2	5	Prevalence	1	Global	Male	15 to 19	332	Chewing tobacco	3	Rate	1991	0.039253	0.055838	0.
3	5	Prevalence	1	Global	Female	15 to 19	332	Chewing tobacco	3	Rate	1991	0.011516	0.017807	0.
4	5	Prevalence	1	Global	Male	15 to 19	332	Chewing tobacco	3	Rate	1992	0.039863	0.056448	0.
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
350545	5	Prevalence	522	Sudan	Both	Age standardized	332	Chewing tobacco	3	Rate	2015	0.029518	0.035685	0.
350546	5	Prevalence	522	Sudan	Both	Age standardized	332	Chewing tobacco	3	Rate	2016	0.029855	0.036004	0.
350547	5	Prevalence	522	Sudan	Both	Age standardized	332	Chewing tobacco	3	Rate	2017	0.029768	0.035934	0.
350548	5	Prevalence	522	Sudan	Both	Age standardized	332	Chewing tobacco	3	Rate	2018	0.029760	0.035796	0.
350549	5	Prevalence	522	Sudan	Both	Age standardized	332	Chewing tobacco	3	Rate	2019	0.029752	0.035857	0.

350550 rows × 15 columns

[178]:
```
df_all = pd.merge(df, df2, on = ['location_id', 'location_name', 'age_group_name', 'year_id'], how = 'inner')
```

[179]:
```
df_all
```

[179]:

	measure_id	measure_name_x	location_id	location_name	sex_id_x	sex	age_group_name	rei_id	rei_name	metric_id	...	upper	lower	Tolerance_range	measur

0 rows × 23 columns

[180]:
```
df_all_1 = df.iloc[1:15,:]
df_all_1
```

[180]:

	measure_id	measure_name	location_id	location_name	sex_id	sex	age_group_name	rei_id	rei_name	metric_id	metric_name	year_id	val	upper
1	5	Prevalence	1	Global	2	Female	15 to 19	332	Chewing tobacco	3	Rate	1990	0.011356	0.017594
2	5	Prevalence	1	Global	1	Male	15 to 19	332	Chewing tobacco	3	Rate	1991	0.039253	0.055838
3	5	Prevalence	1	Global	2	Female	15 to 19	332	Chewing tobacco	3	Rate	1991	0.011516	0.017807
4	5	Prevalence	1	Global	1	Male	15 to 19	332	Chewing tobacco	3	Rate	1992	0.039863	0.056448
5	5	Prevalence	1	Global	2	Female	15 to 19	332	Chewing tobacco	3	Rate	1992	0.011685	0.018425
6	5	Prevalence	1	Global	1	Male	15 to 19	332	Chewing tobacco	3	Rate	1993	0.040508	0.057025
7	5	Prevalence	1	Global	2	Female	15 to 19	332	Chewing tobacco	3	Rate	1993	0.011853	0.018675
8	5	Prevalence	1	Global	1	Male	15 to 19	332	Chewing tobacco	3	Rate	1994	0.041153	0.057284
9	5	Prevalence	1	Global	2	Female	15 to 19	332	Chewing tobacco	3	Rate	1994	0.012010	0.018950
10	5	Prevalence	1	Global	1	Male	15 to 19	332	Chewing tobacco	3	Rate	1995	0.041763	0.057306
11	5	Prevalence	1	Global	2	Female	15 to 19	332	Chewing tobacco	3	Rate	1995	0.012146	0.018994
12	5	Prevalence	1	Global	1	Male	15 to 19	332	Chewing tobacco	3	Rate	1996	0.042341	0.057583
13	5	Prevalence	1	Global	2	Female	15 to 19	332	Chewing tobacco	3	Rate	1996	0.012256	0.019243
14	5	Prevalence	1	Global	1	Male	15 to 19	332	Chewing tobacco	3	Rate	1997	0.042917	0.058115

[181]:
```
df_all_2 = df.iloc[-15::]
df_all_2
```

[181]:

	measure_id	measure_name	location_id	location_name	sex_id	sex	age_group_name	rei_id	rei_name	metric_id	metric_name	year_id	val	upp
350535	5	Prevalence	522	Sudan	3	Both	Age standardized	332	Chewing tobacco	3	Rate	2005	0.028086	0.0340
350536	5	Prevalence	522	Sudan	3	Both	Age standardized	332	Chewing tobacco	3	Rate	2006	0.028140	0.0341
350537	5	Prevalence	522	Sudan	3	Both	Age standardized	332	Chewing tobacco	3	Rate	2007	0.028231	0.0342
350538	5	Prevalence	522	Sudan	3	Both	Age standardized	332	Chewing tobacco	3	Rate	2008	0.028319	0.0343
350539	5	Prevalence	522	Sudan	3	Both	Age standardized	332	Chewing tobacco	3	Rate	2009	0.028395	0.0344
350540	5	Prevalence	522	Sudan	3	Both	Age standardized	332	Chewing tobacco	3	Rate	2010	0.028506	0.0344
350541	5	Prevalence	522	Sudan	3	Both	Age standardized	332	Chewing tobacco	3	Rate	2011	0.028617	0.0344
350542	5	Prevalence	522	Sudan	3	Both	Age standardized	332	Chewing tobacco	3	Rate	2012	0.028790	0.0348
350543	5	Prevalence	522	Sudan	3	Both	Age standardized	332	Chewing tobacco	3	Rate	2013	0.028989	0.0351
350544	5	Prevalence	522	Sudan	3	Both	Age standardized	332	Chewing tobacco	3	Rate	2014	0.029232	0.0353
350545	5	Prevalence	522	Sudan	3	Both	Age standardized	332	Chewing tobacco	3	Rate	2015	0.029518	0.0356
350546	5	Prevalence	522	Sudan	3	Both	Age standardized	332	Chewing tobacco	3	Rate	2016	0.029855	0.0360
350547	5	Prevalence	522	Sudan	3	Both	Age standardized	332	Chewing tobacco	3	Rate	2017	0.029768	0.0359
350548	5	Prevalence	522	Sudan	3	Both	Age standardized	332	Chewing tobacco	3	Rate	2018	0.029760	0.0357
350549	5	Prevalence	522	Sudan	3	Both	Age standardized	332	Chewing tobacco	3	Rate	2019	0.029752	0.0358

[182]:

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

```
(29, 16)
```

[182]:

	measure_id	measure_name	location_id	location_name	sex_id	sex	age_group_name	rei_id	rei_name	metric_id	metric_name	year_id	val	u
1	5	Prevalence	1	Global	2	Female	15 to 19	332	Chewing tobacco	3	Rate	1990	0.011356	0.01
2	5	Prevalence	1	Global	1	Male	15 to 19	332	Chewing tobacco	3	Rate	1991	0.039253	0.05
3	5	Prevalence	1	Global	2	Female	15 to 19	332	Chewing tobacco	3	Rate	1991	0.011516	0.01
4	5	Prevalence	1	Global	1	Male	15 to 19	332	Chewing tobacco	3	Rate	1992	0.039863	0.05
5	5	Prevalence	1	Global	2	Female	15 to 19	332	Chewing tobacco	3	Rate	1992	0.011685	0.01
6	5	Prevalence	1	Global	1	Male	15 to 19	332	Chewing tobacco	3	Rate	1993	0.040508	0.05
7	5	Prevalence	1	Global	2	Female	15 to 19	332	Chewing tobacco	3	Rate	1993	0.011853	0.01
8	5	Prevalence	1	Global	1	Male	15 to 19	332	Chewing tobacco	3	Rate	1994	0.041153	0.05
9	5	Prevalence	1	Global	2	Female	15 to 19	332	Chewing tobacco	3	Rate	1994	0.012010	0.01
10	5	Prevalence	1	Global	1	Male	15 to 19	332	Chewing tobacco	3	Rate	1995	0.041763	0.05
11	5	Prevalence	1	Global	2	Female	15 to 19	332	Chewing tobacco	3	Rate	1995	0.012146	0.01
12	5	Prevalence	1	Global	1	Male	15 to 19	332	Chewing tobacco	3	Rate	1996	0.042341	0.05
13	5	Prevalence	1	Global	2	Female	15 to 19	332	Chewing tobacco	3	Rate	1996	0.012256	0.01
14	5	Prevalence	1	Global	1	Male	15 to 19	332	Chewing tobacco	3	Rate	1997	0.042917	0.05
350535	5	Prevalence	522	Sudan	3	Both	Age standardized	332	Chewing tobacco	3	Rate	2005	0.028086	0.03
350536	5	Prevalence	522	Sudan	3	Both	Age standardized	332	Chewing tobacco	3	Rate	2006	0.028140	0.03
350537	5	Prevalence	522	Sudan	3	Both	Age standardized	332	Chewing tobacco	3	Rate	2007	0.028231	0.03
350538	5	Prevalence	522	Sudan	3	Both	Age standardized	332	Chewing tobacco	3	Rate	2008	0.028319	0.03

[183]:
```
#pokaza´c dodawanie nowych kolumn za pomoca, operacji matematy-
#cznych

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
df_all_new["%ValFromLower"] = df_all_new["val"] / df_all_new["lower"]*100
```

[184]:
```
df_all_new
```

[184]:

	measure_id	measure_name	location_id	location_name	sex_id	sex	age_group_name	rei_id	rei_name	metric_id	metric_name	year_id	val	u
1	5	Prevalence	1	Global	2	Female	15 to 19	332	Chewing tobacco	3	Rate	1990	0.011356	0.01
2	5	Prevalence	1	Global	1	Male	15 to 19	332	Chewing tobacco	3	Rate	1991	0.039253	0.05
3	5	Prevalence	1	Global	2	Female	15 to 19	332	Chewing tobacco	3	Rate	1991	0.011516	0.01
4	5	Prevalence	1	Global	1	Male	15 to 19	332	Chewing tobacco	3	Rate	1992	0.039863	0.05
5	5	Prevalence	1	Global	2	Female	15 to 19	332	Chewing tobacco	3	Rate	1992	0.011685	0.01
6	5	Prevalence	1	Global	1	Male	15 to 19	332	Chewing tobacco	3	Rate	1993	0.040508	0.05
7	5	Prevalence	1	Global	2	Female	15 to 19	332	Chewing tobacco	3	Rate	1993	0.011853	0.01
8	5	Prevalence	1	Global	1	Male	15 to 19	332	Chewing tobacco	3	Rate	1994	0.041153	0.05
9	5	Prevalence	1	Global	2	Female	15 to 19	332	Chewing tobacco	3	Rate	1994	0.012010	0.01
10	5	Prevalence	1	Global	1	Male	15 to 19	332	Chewing tobacco	3	Rate	1995	0.041763	0.05
11	5	Prevalence	1	Global	2	Female	15 to 19	332	Chewing tobacco	3	Rate	1995	0.012146	0.01
12	5	Prevalence	1	Global	1	Male	15 to 19	332	Chewing tobacco	3	Rate	1996	0.042341	0.05
13	5	Prevalence	1	Global	2	Female	15 to 19	332	Chewing tobacco	3	Rate	1996	0.012256	0.01
14	5	Prevalence	1	Global	1	Male	15 to 19	332	Chewing tobacco	3	Rate	1997	0.042917	0.05
350535	5	Prevalence	522	Sudan	3	Both	Age standardized	332	Chewing tobacco	3	Rate	2005	0.028086	0.03
350536	5	Prevalence	522	Sudan	3	Both	Age standardized	332	Chewing tobacco	3	Rate	2006	0.028140	0.03
350537	5	Prevalence	522	Sudan	3	Both	Age standardized	332	Chewing tobacco	3	Rate	2007	0.028231	0.03
350538	5	Prevalence	522	Sudan	3	Both	Age standardized	332	Chewing tobacco	3	Rate	2008	0.028319	0.03
350539	5	Prevalence	522	Sudan	3	Both	Age standardized	332	Chewing tobacco	3	Rate	2009	0.028395	0.03
350540	5	Prevalence	522	Sudan	3	Both	Age standardized	332	Chewing tobacco	3	Rate	2010	0.028506	0.03
350541	5	Prevalence	522	Sudan	3	Both	Age standardized	332	Chewing tobacco	3	Rate	2011	0.028617	0.03
350542	5	Prevalence	522	Sudan	3	Both	Age standardized	332	Chewing tobacco	3	Rate	2012	0.028790	0.03
350543	5	Prevalence	522	Sudan	3	Both	Age standardized	332	Chewing tobacco	3	Rate	2013	0.028989	0.03
350544	5	Prevalence	522	Sudan	3	Both	Age standardized	332	Chewing tobacco	3	Rate	2014	0.029232	0.03
350545	5	Prevalence	522	Sudan	3	Both	Age standardized	332	Chewing tobacco	3	Rate	2015	0.029518	0.03
350546	5	Prevalence	522	Sudan	3	Both	Age standardized	332	Chewing tobacco	3	Rate	2016	0.029855	0.03
350547	5	Prevalence	522	Sudan	3	Both	Age standardized	332	Chewing tobacco	3	Rate	2017	0.029768	0.03
350548	5	Prevalence	522	Sudan	3	Both	Age standardized	332	Chewing tobacco	3	Rate	2018	0.029760	0.03
350549	5	Prevalence	522	Sudan	3	Both	Age standardized	332	Chewing tobacco	3	Rate	2019	0.029752	0.03

[185]:
```
#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 year_id else False )
df_all_new
```

[185]:

	index	measure_id	measure_name	location_id	location_name	sex_id	sex	age_group_name	rei_id	rei_name	...	metric_name	year_id	val	uppe
0	1	5	Prevalence	1	Global	2	Female	15 to 19	332	Chewing tobacco	...	Rate	1990	0.011356	0.01759
1	2	5	Prevalence	1	Global	1	Male	15 to 19	332	Chewing tobacco	...	Rate	1991	0.039253	0.05583
2	3	5	Prevalence	1	Global	2	Female	15 to 19	332	Chewing tobacco	...	Rate	1991	0.011516	0.01780
3	4	5	Prevalence	1	Global	1	Male	15 to 19	332	Chewing tobacco	...	Rate	1992	0.039863	0.05644
4	5	5	Prevalence	1	Global	2	Female	15 to 19	332	Chewing tobacco	...	Rate	1992	0.011685	0.01842
5	6	5	Prevalence	1	Global	1	Male	15 to 19	332	Chewing tobacco	...	Rate	1993	0.040508	0.05702
6	7	5	Prevalence	1	Global	2	Female	15 to 19	332	Chewing tobacco	...	Rate	1993	0.011853	0.01867
7	8	5	Prevalence	1	Global	1	Male	15 to 19	332	Chewing tobacco	...	Rate	1994	0.041153	0.05728
8	9	5	Prevalence	1	Global	2	Female	15 to 19	332	Chewing tobacco	...	Rate	1994	0.012010	0.01895
9	10	5	Prevalence	1	Global	1	Male	15 to 19	332	Chewing tobacco	...	Rate	1995	0.041763	0.05730
10	11	5	Prevalence	1	Global	2	Female	15 to 19	332	Chewing tobacco	...	Rate	1995	0.012146	0.01899
11	12	5	Prevalence	1	Global	1	Male	15 to 19	332	Chewing tobacco	...	Rate	1996	0.042341	0.05758
12	13	5	Prevalence	1	Global	2	Female	15 to 19	332	Chewing tobacco	...	Rate	1996	0.012256	0.01924
13	14	5	Prevalence	1	Global	1	Male	15 to 19	332	Chewing tobacco	...	Rate	1997	0.042917	0.05811
14	350535	5	Prevalence	522	Sudan	3	Both	Age standardized	332	Chewing tobacco	...	Rate	2005	0.028086	0.03407
15	350536	5	Prevalence	522	Sudan	3	Both	Age standardized	332	Chewing tobacco	...	Rate	2006	0.028140	0.03412
16	350537	5	Prevalence	522	Sudan	3	Both	Age standardized	332	Chewing tobacco	...	Rate	2007	0.028231	0.03420
17	350538	5	Prevalence	522	Sudan	3	Both	Age standardized	332	Chewing tobacco	...	Rate	2008	0.028319	0.03430
18	350539	5	Prevalence	522	Sudan	3	Both	Age standardized	332	Chewing tobacco	...	Rate	2009	0.028395	0.03444
19	350540	5	Prevalence	522	Sudan	3	Both	Age standardized	332	Chewing tobacco	...	Rate	2010	0.028506	0.03448
20	350541	5	Prevalence	522	Sudan	3	Both	Age standardized	332	Chewing tobacco	...	Rate	2011	0.028617	0.03446
21	350542	5	Prevalence	522	Sudan	3	Both	Age standardized	332	Chewing tobacco	...	Rate	2012	0.028790	0.03482
22	350543	5	Prevalence	522	Sudan	3	Both	Age standardized	332	Chewing tobacco	...	Rate	2013	0.028989	0.03513
23	350544	5	Prevalence	522	Sudan	3	Both	Age standardized	332	Chewing tobacco	...	Rate	2014	0.029232	0.03539
24	350545	5	Prevalence	522	Sudan	3	Both	Age standardized	332	Chewing tobacco	...	Rate	2015	0.029518	0.03568
25	350546	5	Prevalence	522	Sudan	3	Both	Age standardized	332	Chewing tobacco	...	Rate	2016	0.029855	0.03600
26	350547	5	Prevalence	522	Sudan	3	Both	Age standardized	332	Chewing tobacco	...	Rate	2017	0.029768	0.03593
27	350548	5	Prevalence	522	Sudan	3	Both	Age standardized	332	Chewing tobacco	...	Rate	2018	0.029760	0.03579
28	350549	5	Prevalence	522	Sudan	3	Both	Age standardized	332	Chewing tobacco	...	Rate	2019	0.029752	0.03585

29 rows × 21 columns

[187]:
```
#przedstawić możliwości pracy z dużymi plikami przy użyciu argumentu
#chunksize

for chunk_df in pd.read_csv('IHME_GBD_2019_CHEWING_TOB_1990_2019_CIG_PC_Y2021M05D27.CSV',
                            chunksize = 50000):
    print("CHUNK DF")
    print(chunk_df.head())
```

```
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
```

[189]:
```
new_df = pd.DataFrame() # pusta ramka danych
for chunk_df in pd.read_csv('IHME_GBD_2019_CHEWING_TOB_1990_2019_CIG_PC_Y2021M05D27.CSV',
                            chunksize = 50000):
    result = chunk_df.groupby(['location_name', 'year_id']).agg({'val': 'mean',
                                                                 'upper': 'max'})
    new_df = pd.concat([new_df,result])

new_df
```

[189]:

location_name	year_id	val	upper
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 × 2 columns

[ ]:

3. Wnioski

Biblioteka Pandas w łątwy sposób pozwala manipulować danymi. Biblioteka Matplotlib pozwala na proste wizualizacje danych.