```
# instalacja pakietu pyspark

!pip install pyspark==3.0.1 py4j==0.10.9
```

```
Collecting pyspark==3.0.1
  Downloading pyspark-3.0.1.tar.gz (204.2 MB)
     |████████████████████████████████| 204.2 MB 33 kB/s
Collecting py4j==0.10.9
  Downloading py4j-0.10.9-py2.py3-none-any.whl (198 kB)
     |████████████████████████████████| 198 kB 18.7 MB/s
Building wheels for collected packages: pyspark
  Building wheel for pyspark (setup.py) ... done
  Created wheel for pyspark: filename=pyspark-3.0.1-py2.py3-none-any.whl size=204612
  Stored in directory: /root/.cache/pip/wheels/5e/34/fa/b37b5cef503fc5148b478b249504
Successfully built pyspark
Installing collected packages: py4j, pyspark
Successfully installed py4j-0.10.9 pyspark-3.0.1
```

```
from pyspark.sql import SparkSession

# utworzenie sesji
spark = SparkSession\
        .builder\
        .master('local[4]')\
        .appName('zadanie_4')\
        .getOrCreate()
```

```
csv_file = r"/content/IHME_DAH_DATABASE_1990_2020_Y2021M09D22.CSV"

# pobranie danych z pliku CSV
data = spark.read.csv(csv_file, header=True)
```

```
# wyświetlenie 5 wierszy danych
data.show(5)
```

```
+----+---------+-------+----------------+----------------+---------------+--------
|year|   source|channel|recipient_isocode|recipient_country|gbd_location_id|wb_regio
+----+---------+-------+----------------+----------------+---------------+--------
|1990|Australia|BIL_AUS|             AGO|          Angola|            168|
|1990|Australia|BIL_AUS|             BDI|         Burundi|            175|
|1990|Australia|BIL_AUS|             BEN|           Benin|            200|
|1990|Australia|BIL_AUS|             BFA|    Burkina Faso|            201|
|1990|Australia|BIL_AUS|             BWA|        Botswana|            193|
+----+---------+-------+----------------+----------------+---------------+--------
only showing top 5 rows
```

```
# wyświetlenie schematu danych
data.printSchema()
```

```
 |-- rmh_other_dah_20: string (nullable = true)
```

```
 |-- nch_cnn_dah_20: string (nullable = true)
 |-- nch_cnv_dah_20: string (nullable = true)
 |-- nch_other_dah_20: string (nullable = true)
 |-- nch_hss_other_dah_20: string (nullable = true)
 |-- nch_hss_hrh_dah_20: string (nullable = true)
 |-- hiv_treat_dah_20: string (nullable = true)
 |-- hiv_prev_dah_20: string (nullable = true)
 |-- hiv_pmtct_dah_20: string (nullable = true)
 |-- hiv_other_dah_20: string (nullable = true)
 |-- hiv_ct_dah_20: string (nullable = true)
 |-- hiv_ovc_dah_20: string (nullable = true)
 |-- hiv_care_dah_20: string (nullable = true)
 |-- hiv_hss_other_dah_20: string (nullable = true)
 |-- hiv_hss_hrh_dah_20: string (nullable = true)
 |-- hiv_amr_dah_20: string (nullable = true)
 |-- mal_diag_dah_20: string (nullable = true)
 |-- mal_hss_other_dah_20: string (nullable = true)
 |-- mal_hss_hrh_dah_20: string (nullable = true)
 |-- mal_con_nets_dah_20: string (nullable = true)
 |-- mal_con_irs_dah_20: string (nullable = true)
 |-- mal_con_oth_dah_20: string (nullable = true)
 |-- mal_treat_dah_20: string (nullable = true)
 |-- mal_comm_con_dah_20: string (nullable = true)
 |-- mal_other_dah_20: string (nullable = true)
 |-- mal_amr_dah_20: string (nullable = true)
 |-- tb_other_dah_20: string (nullable = true)
 |-- tb_treat_dah_20: string (nullable = true)
 |-- tb_diag_dah_20: string (nullable = true)
 |-- tb_hss_other_dah_20: string (nullable = true)
 |-- tb_hss_hrh_dah_20: string (nullable = true)
 |-- tb_amr_dah_20: string (nullable = true)
 |-- oid_hss_other_dah_20: string (nullable = true)
 |-- oid_hss_hrh_dah_20: string (nullable = true)
 |-- oid_ebz_dah_20: string (nullable = true)
 |-- oid_zika_dah_20: string (nullable = true)
 |-- oid_covid_dah_20: string (nullable = true)
 |-- oid_other_dah_20: string (nullable = true)
 |-- oid_amr_dah_20: string (nullable = true)
 |-- ncd_hss_other_dah_20: string (nullable = true)
 |-- ncd_hss_hrh_dah_20: string (nullable = true)
 |-- ncd_tobac_dah_20: string (nullable = true)
 |-- ncd_mental_dah_20: string (nullable = true)
 |-- ncd_other_dah_20: string (nullable = true)
 |-- swap_hss_other_dah_20: string (nullable = true)
 |-- swap_hss_hrh_dah_20: string (nullable = true)
 |-- swap_hss_pp_dah_20: string (nullable = true)
 |-- other_dah_20: string (nullable = true)
 |-- rmh_dah_20: string (nullable = true)
 |-- nch_dah_20: string (nullable = true)
 |-- ncd_dah_20: string (nullable = true)
 |-- hiv_dah_20: string (nullable = true)
 |-- mal_dah_20: string (nullable = true)
 |-- tb_dah_20: string (nullable = true)
 |-- swap_hss_total_dah_20: string (nullable = true)
 |-- oid_dah_20: string (nullable = true)
 |-- unalloc_dah_20: string (nullable = true)
```

```
1 from pyspark.sql.types import *
```

```
 2
 3 # utworzenie schematu danych
 4 data_schema = [
 5       StructField('year', IntegerType(), True),
 6       StructField('source', StringType(), True),
 7       StructField('channel', StringType(), True),
 8       StructField('recipient_isocode', StringType(), True),
 9       StructField('recipient_country', StringType(), True),
10       StructField('gbd_location_id', IntegerType(), True),
11       StructField('wb_regioncode', StringType(), True),
12       StructField('wb_location_id', IntegerType(), True),
13       StructField('gbd_region', StringType(), True),
14       StructField('gbd_region_id', IntegerType(), True),
15       StructField('gbd_superregion', StringType(), True),
16       StructField('gbd_superregion_id', IntegerType(), True),
17       StructField('elim_ch', IntegerType(), True),
18       StructField('prelim_est', IntegerType(), True),
19       StructField('dah_20', IntegerType(), True),
20       StructField('rmh_fp_dah_20', IntegerType(), True),
21       StructField('rmh_mh_dah_20', IntegerType(), True),
22       StructField('rmh_hss_other_dah_20', IntegerType(), True),
23       StructField('rmh_hss_hrh_dah_20', IntegerType(), True),
24       StructField('rmh_other_dah_20', IntegerType(), True),
25       StructField('nch_cnn_dah_20', IntegerType(), True),
26       StructField('nch_cnv_dah_20', IntegerType(), True),
27       StructField('nch_other_dah_20', IntegerType(), True),
28       StructField('nch_hss_other_dah_20', IntegerType(), True),
29       StructField('nch_hss_hrh_dah_20', IntegerType(), True),
30       StructField('hiv_treat_dah_20', IntegerType(), True),
31       StructField('hiv_prev_dah_20', IntegerType(), True),
32       StructField('hiv_pmtct_dah_20', IntegerType(), True),
33       StructField('hiv_other_dah_20', IntegerType(), True),
34       StructField('hiv_ct_dah_20', IntegerType(), True),
35       StructField('hiv_ovc_dah_20', IntegerType(), True),
36       StructField('hiv_care_dah_20', IntegerType(), True),
37       StructField('hiv_hss_other_dah_20', IntegerType(), True),
38       StructField('hiv_hss_hrh_dah_20', IntegerType(), True),
39       StructField('hiv_amr_dah_20', IntegerType(), True),
40       StructField('mal_diag_dah_20', IntegerType(), True),
41       StructField('mal_hss_other_dah_20', IntegerType(), True),
42       StructField('mal_hss_hrh_dah_20', IntegerType(), True),
43       StructField('mal_con_nets_dah_20', IntegerType(), True),
44       StructField('mal_con_irs_dah_20', IntegerType(), True),
45       StructField('mal_con_oth_dah_20', IntegerType(), True),
46       StructField('mal_treat_dah_20', IntegerType(), True),
47       StructField('mal_comm_con_dah_20', IntegerType(), True),
48       StructField('mal_other_dah_20', IntegerType(), True),
49       StructField('mal_amr_dah_20', IntegerType(), True),
50       StructField('tb_other_dah_20', IntegerType(), True),
51       StructField('tb_treat_dah_20', IntegerType(), True),
52       StructField('tb_diag_dah_20', IntegerType(), True),
53       StructField('tb_hss_other_dah_20', IntegerType(), True),
54       StructField('tb_hss_hrh_dah_20', IntegerType(), True),
55       StructField('tb_amr_dah_20', IntegerType(), True),
56       StructField('oid_hss_other_dah_20', IntegerType(), True),
```

```
      StructField('oid_hss_hrh_dah_20', IntegerType(), True),
      StructField('oid_ebz_dah_20', IntegerType(), True),
      StructField('oid_zika_dah_20', IntegerType(), True),
      StructField('oid_covid_dah_20', IntegerType(), True),
      StructField('oid_other_dah_20', IntegerType(), True),
      StructField('oid_amr_dah_20', IntegerType(), True),
      StructField('ncd_hss_other_dah_20', IntegerType(), True),
      StructField('ncd_hss_hrh_dah_20', IntegerType(), True),
      StructField('ncd_tobac_dah_20', IntegerType(), True),
      StructField('ncd_mental_dah_20', IntegerType(), True),
      StructField('ncd_other_dah_20', IntegerType(), True),
      StructField('swap_hss_other_dah_20', IntegerType(), True),
      StructField('swap_hss_hrh_dah_20', IntegerType(), True),
      StructField('swap_hss_pp_dah_20', IntegerType(), True),
      StructField('other_dah_20', IntegerType(), True),
      StructField('rmh_dah_20', IntegerType(), True),
      StructField('nch_dah_20', IntegerType(), True),
      StructField('ncd_dah_20', IntegerType(), True),
      StructField('hiv_dah_20', IntegerType(), True),
      StructField('mal_dah_20', IntegerType(), True),
      StructField('tb_dah_20', IntegerType(), True),
      StructField('swap_hss_total_dah_20', IntegerType(), True),
      StructField('oid_dah_20', IntegerType(), True),
      StructField('unalloc_dah_20', IntegerType(), True),
]

data_struc = StructType(fields = data_schema)
```

```
# dodanie schematu do danych
data2 = spark.read.csv(csv_file, header=True, schema=data_struc)

data2.printSchema()
```

```
    |-- nch_cnn_dah_20: integer (nullable = true)

    |-- nch_cnv_dah_20: integer (nullable = true)
    |-- nch_other_dah_20: integer (nullable = true)
    |-- nch_hss_other_dah_20: integer (nullable = true)
    |-- nch_hss_hrh_dah_20: integer (nullable = true)
    |-- hiv_treat_dah_20: integer (nullable = true)
    |-- hiv_prev_dah_20: integer (nullable = true)
    |-- hiv_pmtct_dah_20: integer (nullable = true)
    |-- hiv_other_dah_20: integer (nullable = true)
    |-- hiv_ct_dah_20: integer (nullable = true)
    |-- hiv_ovc_dah_20: integer (nullable = true)
    |-- hiv_care_dah_20: integer (nullable = true)
    |-- hiv_hss_other_dah_20: integer (nullable = true)
    |-- hiv_hss_hrh_dah_20: integer (nullable = true)
    |-- hiv_amr_dah_20: integer (nullable = true)
    |-- mal_diag_dah_20: integer (nullable = true)
    |-- mal_hss_other_dah_20: integer (nullable = true)
    |-- mal_hss_hrh_dah_20: integer (nullable = true)
    |-- mal_con_nets_dah_20: integer (nullable = true)
    |-- mal_con_irs_dah_20: integer (nullable = true)
    |-- mal_con_oth_dah_20: integer (nullable = true)
    |-- mal_treat_dah_20: integer (nullable = true)
```

```
 |-- mal_comm_con_dah_20: integer (nullable = true)
 |-- mal_other_dah_20: integer (nullable = true)
 |-- mal_amr_dah_20: integer (nullable = true)
 |-- tb_other_dah_20: integer (nullable = true)
 |-- tb_treat_dah_20: integer (nullable = true)
 |-- tb_diag_dah_20: integer (nullable = true)
 |-- tb_hss_other_dah_20: integer (nullable = true)
 |-- tb_hss_hrh_dah_20: integer (nullable = true)
 |-- tb_amr_dah_20: integer (nullable = true)
 |-- oid_hss_other_dah_20: integer (nullable = true)
 |-- oid_hss_hrh_dah_20: integer (nullable = true)
 |-- oid_ebz_dah_20: integer (nullable = true)
 |-- oid_zika_dah_20: integer (nullable = true)
 |-- oid_covid_dah_20: integer (nullable = true)
 |-- oid_other_dah_20: integer (nullable = true)
 |-- oid_amr_dah_20: integer (nullable = true)
 |-- ncd_hss_other_dah_20: integer (nullable = true)
 |-- ncd_hss_hrh_dah_20: integer (nullable = true)
 |-- ncd_tobac_dah_20: integer (nullable = true)
 |-- ncd_mental_dah_20: integer (nullable = true)
 |-- ncd_other_dah_20: integer (nullable = true)
 |-- swap_hss_other_dah_20: integer (nullable = true)
 |-- swap_hss_hrh_dah_20: integer (nullable = true)
 |-- swap_hss_pp_dah_20: integer (nullable = true)
 |-- other_dah_20: integer (nullable = true)
 |-- rmh_dah_20: integer (nullable = true)
 |-- nch_dah_20: integer (nullable = true)
 |-- ncd_dah_20: integer (nullable = true)
 |-- hiv_dah_20: integer (nullable = true)
 |-- mal_dah_20: integer (nullable = true)
 |-- tb_dah_20: integer (nullable = true)
 |-- swap_hss_total_dah_20: integer (nullable = true)
 |-- oid_dah_20: integer (nullable = true)
 |-- unalloc_dah_20: integer (nullable = true)
```

```
1 data2.show(5)
```

```
+----+---------+-------+-----------------+-----------------+---------------+--------
|year|   source|channel|recipient_isocode|recipient_country|gbd_location_id|wb_regio
+----+---------+-------+-----------------+-----------------+---------------+--------
|1990|Australia|BIL_AUS|              AGO|           Angola|            168|
|1990|Australia|BIL_AUS|              BDI|          Burundi|            175|
|1990|Australia|BIL_AUS|              BEN|            Benin|            200|
|1990|Australia|BIL_AUS|              BFA|     Burkina Faso|            201|
|1990|Australia|BIL_AUS|              BWA|         Botswana|            193|
+----+---------+-------+-----------------+-----------------+---------------+--------
only showing top 5 rows
```

```
1 # wyświetlenie typów danych w poszczególnych kolumnach
2 data.dtypes
```

```
 ('rmh_hss_hrh_dah_20', 'string'),
 ('rmh_other_dah_20', 'string'),
 ('nch_cnn_dah_20', 'string'),
 ('nch_cnv_dah_20', 'string'),
 ('nch_other_dah_20', 'string'),
```

```
('nch_hss_other_dah_20', 'string'),
('nch_hss_hrh_dah_20', 'string'),
('hiv_treat_dah_20', 'string'),
('hiv_prev_dah_20', 'string'),
('hiv_pmtct_dah_20', 'string'),
('hiv_other_dah_20', 'string'),
('hiv_ct_dah_20', 'string'),
('hiv_ovc_dah_20', 'string'),
('hiv_care_dah_20', 'string'),
('hiv_hss_other_dah_20', 'string'),
('hiv_hss_hrh_dah_20', 'string'),
('hiv_amr_dah_20', 'string'),
('mal_diag_dah_20', 'string'),
('mal_hss_other_dah_20', 'string'),
('mal_hss_hrh_dah_20', 'string'),
('mal_con_nets_dah_20', 'string'),
('mal_con_irs_dah_20', 'string'),
('mal_con_oth_dah_20', 'string'),
('mal_treat_dah_20', 'string'),
('mal_comm_con_dah_20', 'string'),
('mal_other_dah_20', 'string'),
('mal_amr_dah_20', 'string'),
('tb_other_dah_20', 'string'),
('tb_treat_dah_20', 'string'),
('tb_diag_dah_20', 'string'),
('tb_hss_other_dah_20', 'string'),
('tb_hss_hrh_dah_20', 'string'),
('tb_amr_dah_20', 'string'),
('oid_hss_other_dah_20', 'string'),
('oid_hss_hrh_dah_20', 'string'),
('oid_ebz_dah_20', 'string'),
('oid_zika_dah_20', 'string'),
('oid_covid_dah_20', 'string'),
('oid_other_dah_20', 'string'),
('oid_amr_dah_20', 'string'),
('ncd_hss_other_dah_20', 'string'),
('ncd_hss_hrh_dah_20', 'string'),
('ncd_tobac_dah_20', 'string'),
('ncd_mental_dah_20', 'string'),
('ncd_other_dah_20', 'string'),
('swap_hss_other_dah_20', 'string'),
('swap_hss_hrh_dah_20', 'string'),
('swap_hss_pp_dah_20', 'string'),
('other_dah_20', 'string'),
('rmh_dah_20', 'string'),
('nch_dah_20', 'string'),
('ncd_dah_20', 'string'),
('hiv_dah_20', 'string'),
('mal_dah_20', 'string'),
('tb_dah_20', 'string'),
('swap_hss_total_dah_20', 'string'),
('oid_dah_20', 'string'),
('unalloc_dah_20', 'string')]
```

```
1 data2.dtypes
```

```
('rmh_hss_hrh_dah_20', 'int'),
('rmh_other_dah_20', 'int'),
('nch_cnn_dah_20', 'int'),
('nch_cnv_dah_20', 'int')
```

```
 ('nch_chv_dah_20', 'int'),
 ('nch_other_dah_20', 'int'),
 ('nch_hss_other_dah_20', 'int'),
 ('nch_hss_hrh_dah_20', 'int'),
 ('hiv_treat_dah_20', 'int'),
 ('hiv_prev_dah_20', 'int'),
 ('hiv_pmtct_dah_20', 'int'),
 ('hiv_other_dah_20', 'int'),
 ('hiv_ct_dah_20', 'int'),
 ('hiv_ovc_dah_20', 'int'),
 ('hiv_care_dah_20', 'int'),
 ('hiv_hss_other_dah_20', 'int'),
 ('hiv_hss_hrh_dah_20', 'int'),
 ('hiv_amr_dah_20', 'int'),
 ('mal_diag_dah_20', 'int'),
 ('mal_hss_other_dah_20', 'int'),
 ('mal_hss_hrh_dah_20', 'int'),
 ('mal_con_nets_dah_20', 'int'),
 ('mal_con_irs_dah_20', 'int'),
 ('mal_con_oth_dah_20', 'int'),
 ('mal_treat_dah_20', 'int'),
 ('mal_comm_con_dah_20', 'int'),
 ('mal_other_dah_20', 'int'),
 ('mal_amr_dah_20', 'int'),
 ('tb_other_dah_20', 'int'),
 ('tb_treat_dah_20', 'int'),
 ('tb_diag_dah_20', 'int'),
 ('tb_hss_other_dah_20', 'int'),
 ('tb_hss_hrh_dah_20', 'int'),
 ('tb_amr_dah_20', 'int'),
 ('oid_hss_other_dah_20', 'int'),
 ('oid_hss_hrh_dah_20', 'int'),
 ('oid_ebz_dah_20', 'int'),
 ('oid_zika_dah_20', 'int'),
 ('oid_covid_dah_20', 'int'),
 ('oid_other_dah_20', 'int'),
 ('oid_amr_dah_20', 'int'),
 ('ncd_hss_other_dah_20', 'int'),
 ('ncd_hss_hrh_dah_20', 'int'),
 ('ncd_tobac_dah_20', 'int'),
 ('ncd_mental_dah_20', 'int'),
 ('ncd_other_dah_20', 'int'),
 ('swap_hss_other_dah_20', 'int'),
 ('swap_hss_hrh_dah_20', 'int'),
 ('swap_hss_pp_dah_20', 'int'),
 ('other_dah_20', 'int'),
 ('rmh_dah_20', 'int'),
 ('nch_dah_20', 'int'),
 ('ncd_dah_20', 'int'),
 ('hiv_dah_20', 'int'),
 ('mal_dah_20', 'int'),
 ('tb_dah_20', 'int'),
 ('swap_hss_total_dah_20', 'int'),
 ('oid_dah_20', 'int'),
 ('unalloc_dah_20', 'int')]
```

```
1 # wyświetlenie pierwszych 2 wierszy
2 data2.head(5)
```

```
[Row(year=1990, source='Australia', channel='BIL_AUS', recipient_isocode='AGO', reci
 Row(year=1990, source='Australia', channel='BIL_AUS', recipient_isocode='BDI', reci
 Row(year=1990, source='Australia', channel='BIL_AUS', recipient_isocode='BEN', reci
 Row(year=1990, source='Australia', channel='BIL_AUS', recipient_isocode='BFA', reci
 Row(year=1990, source='Australia', channel='BIL_AUS', recipient_isocode='BWA', reci
```

```
# wyświetlenie ostatnich 2 wierszy
data2.tail(5)
```

```
[Row(year=2020, source='United_States', channel='UNICEF', recipient_isocode='QZA', r
 Row(year=2020, source='United_States', channel='UNITAID', recipient_isocode='QZA',
 Row(year=2020, source='United_States', channel='UNITAID', recipient_isocode='QZA',
 Row(year=2020, source='United_States', channel='WB_IDA', recipient_isocode='QZA', r
 Row(year=2020, source='United_States', channel='WHO', recipient_isocode='QZA', reci
```

```
# dodanie kolumny
res = data2.withColumn('Nowa kolumna', data2.year*0 + 1000)

res.show(5)
```

```
+----+---------+-------+-----------------+-----------------+---------------+--------
|year|   source|channel|recipient_isocode|recipient_country|gbd_location_id|wb_regio
+----+---------+-------+-----------------+-----------------+---------------+--------
|1990|Australia|BIL_AUS|              AGO|           Angola|            168|
|1990|Australia|BIL_AUS|              BDI|          Burundi|            175|
|1990|Australia|BIL_AUS|              BEN|            Benin|            200|
|1990|Australia|BIL_AUS|              BFA|     Burkina Faso|            201|
|1990|Australia|BIL_AUS|              BWA|         Botswana|            193|
+----+---------+-------+-----------------+-----------------+---------------+--------
only showing top 5 rows
```

```
# zmiana nazwy kolumny
res = res.withColumnRenamed('Nowa kolumna', 'col')

res.show(5)
```

```
+----+---------+-------+-----------------+-----------------+---------------+--------
|year|   source|channel|recipient_isocode|recipient_country|gbd_location_id|wb_regio
+----+---------+-------+-----------------+-----------------+---------------+--------
|1990|Australia|BIL_AUS|              AGO|           Angola|            168|
|1990|Australia|BIL_AUS|              BDI|          Burundi|            175|
|1990|Australia|BIL_AUS|              BEN|            Benin|            200|
|1990|Australia|BIL_AUS|              BFA|     Burkina Faso|            201|
|1990|Australia|BIL_AUS|              BWA|         Botswana|            193|
+----+---------+-------+-----------------+-----------------+---------------+--------
only showing top 5 rows
```

```
# usunięcie kolumny
res = data2.drop('col')
res.show(5)
```

```
+----+---------+-------+----------------+----------------+---------------+--------
|year|   source|channel|recipient_isocode|recipient_country|gbd_location_id|wb_regio
+----+---------+-------+----------------+----------------+---------------+--------
|1990|Australia|BIL_AUS|             AGO|          Angola|            168|
|1990|Australia|BIL_AUS|             BDI|         Burundi|            175|
|1990|Australia|BIL_AUS|             BEN|           Benin|            200|
|1990|Australia|BIL_AUS|             BFA|    Burkina Faso|            201|
|1990|Australia|BIL_AUS|             BWA|        Botswana|            193|
+----+---------+-------+----------------+----------------+---------------+--------
only showing top 5 rows
```

```
from pyspark.sql.functions import udf

# funkcja zwracająca kolejne liczby naturalne od 0
i = -1
def incr():
  global i
  i = i+1
  return i

# utworzenie nowej kolumny
newCol = udf(incr, IntegerType())

# dodanie nowej kolumny
data3 = data2.withColumn('id', newCol())

data3.show(5)
```

```
+----+---------+-------+----------------+----------------+---------------+--------
|year|   source|channel|recipient_isocode|recipient_country|gbd_location_id|wb_regio
+----+---------+-------+----------------+----------------+---------------+--------
|1990|Australia|BIL_AUS|             AGO|          Angola|            168|
|1990|Australia|BIL_AUS|             BDI|         Burundi|            175|
|1990|Australia|BIL_AUS|             BEN|           Benin|            200|
|1990|Australia|BIL_AUS|             BFA|    Burkina Faso|            201|
|1990|Australia|BIL_AUS|             BWA|        Botswana|            193|
+----+---------+-------+----------------+----------------+---------------+--------
only showing top 5 rows
```

```
# liczba rekordów

data3.count()
```

```
384306
```

```
# usunięcie wierszy bez danych
```

```
data4 = data3.na.drop()

data4.count()
```

```
136560
```

```
# wstawienie zera w miejsce braku danych
data5 = data3.na.fill(data3.select(0 * data3.year).collect()[0][0])

data5.count()
```

```
384306
```

```
# wybranie kolumn
data5.select(['year', 'source', 'dah_20']).show(5)
```

```
+----+---------+------+
|year|   source|dah_20|
+----+---------+------+
|1990|Australia|    14|
|1990|Australia|    12|
|1990|Australia|    12|
|1990|Australia|    13|
|1990|Australia|    25|
+----+---------+------+
only showing top 5 rows
```

```
# odfiltrowanie wierszy
from pyspark.sql.functions import col

data5.filter((col('year') >= 2000) & (col('dah_20') > 20)).select(['year', 'source', 'd
```

```
+----+---------+------+
|year|   source|dah_20|
+----+---------+------+
|2000|Australia|   333|
|2000|Australia|    25|
|2000|Australia|    46|
|2000|Australia|   106|
|2000|Australia|    22|
+----+---------+------+
only showing top 5 rows
```

```
# dodanie kolumny zawierającej wynik sprawdzenia, czy rok jest większy niż 2000
from pyspark.sql import functions as f

data5.select('year', 'source', 'dah_20', f.when(data5.year > 2000, '21st century').othe
```

```
+----+---------+------+------------+
|year|   source|dah_20|     century|
+----+---------+------+------------+
```

```
|1990|Australia|    14|20th century|
|1990|Australia|    12|20th century|
|1990|Australia|    12|20th century|
|1990|Australia|    13|20th century|
|1990|Australia|    25|20th century|
+----+---------+------+------------+
only showing top 5 rows
```

```
# dodanie kolumny zawierającej wynik sprawdzenai, czy nazwa kraju zaczyna się od litery
data5.select('year', 'source', 'dah_20', data5.source.rlike('^A').alias('A-country')).s
```

```
+----+---------+------+---------+
|year|   source|dah_20|A-country|
+----+---------+------+---------+
|1990|Australia|    14|     true|
|1990|Australia|    12|     true|
|1990|Australia|    12|     true|
|1990|Australia|    13|     true|
|1990|Australia|    25|     true|
+----+---------+------+---------+
only showing top 5 rows
```

```
# pogrupowanie danych wg kraju
from pyspark.sql.functions import mean, count, min, max

data5\
  .select(['year', 'source', 'dah_20'])\
  .groupBy('source')\
  .agg(
      count('year').alias('number of countries'),
      mean('dah_20').alias('meah dah_20'),
      min('year').alias('min year'),
      max('year').alias('max year'),
  ).show(5)
```

```
+---------------+------------------+------------------+--------+--------+
|         source|number of countries|       meah dah_20|min year|max year|
+---------------+------------------+------------------+--------+--------+
|         Sweden|             19691| 945.3851505764054|    1990|    2020|
|Debt_repayments|              4893| 8464.654199877376|    1990|    2020|
|        Germany|             17489|1851.7729429927383|    1990|    2020|
|         France|             17156|1562.2891116810445|    1990|    2020|
|         Greece|              8178| 80.97407679139154|    1990|    2020|
+---------------+------------------+------------------+--------+--------+
only showing top 5 rows
```

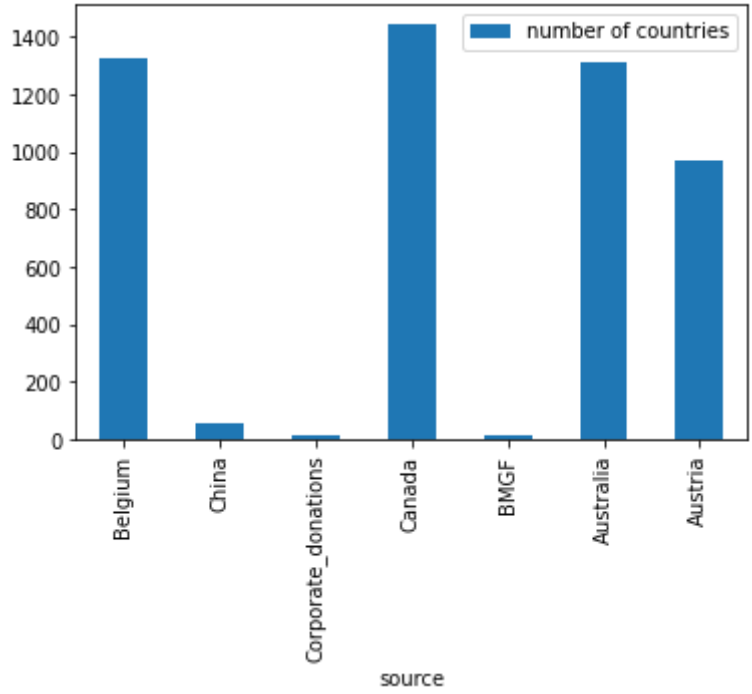
```
# wyświetlenie wykresu słupkowego na podstawie danych
from matplotlib import pyplot as plt

res = data5\
  .filter(data5.source.rlike('^[ABC]'))\
  .select(['year', 'source', 'dah_20'])\
  .groupBy('source')\
```

```
  .agg(
      count('year').alias('number of countries'),
      mean('dah_20').alias('mean dah_20'),
      mean('year').alias('mean year'),
      max('year').alias('max year'))\
  .toPandas()

res\
  .plot(kind='bar', x='source', y='number of countries')

plt.show()
```




```
# zapis danych do plików CSV, JSON i Parquet
data5.select(['year', 'source', 'dah_20']).write.csv(r'/content/csv_res_file.csv')
data5.select(['year', 'source', 'dah_20']).write.save(r'/content/json_res_file.json')
data5.select(['year', 'source', 'dah_20']).write.save(r'/content/parquet_res_file.parqu
```

```
1
```

✓ 2 s ukończono o 20:07