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

Laboratorium 327.11.2021  
Temat:​ "Użycie biblioteki „PySpark" dla dużych zbiorów danych.

**Wariant: 2**

Repozytorium z kodem:

<https://colab.research.google.com/drive/1rdOS3RLI3cXrvE-St82s8otl1ZQS6jM7?usp=sharing>

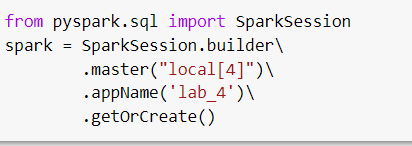
Przemysław Dziki  
Informatyka II stopień,  
niestacjonarne (zaoczne),  
III semestr,  
Gr.A

Zadanie 1. Instalacja pakietu „pyspark”

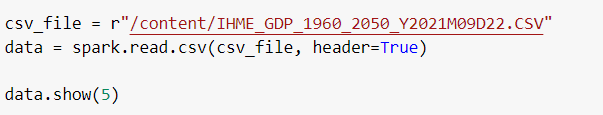
Instalacja biblioteki w notatniku Google Colab

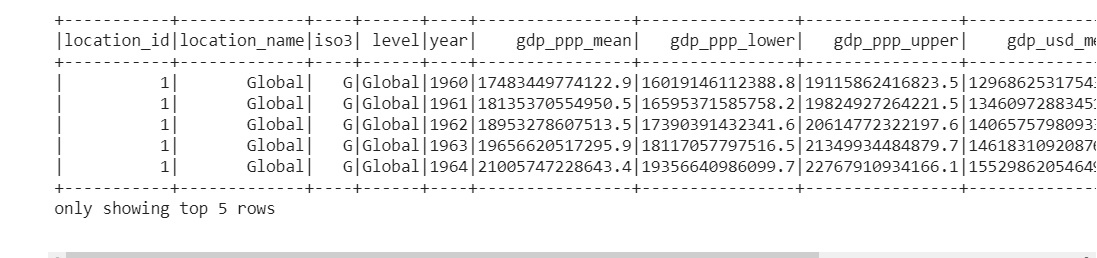


Zadanie 2. Tworzenie sesji „SparkSession”



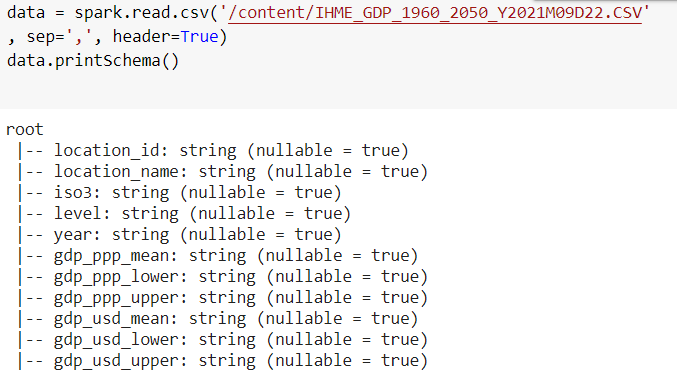
Zadanie 3. Wczytać dane z pliku CSV i wyświetlić pierwsze 5 wierszy

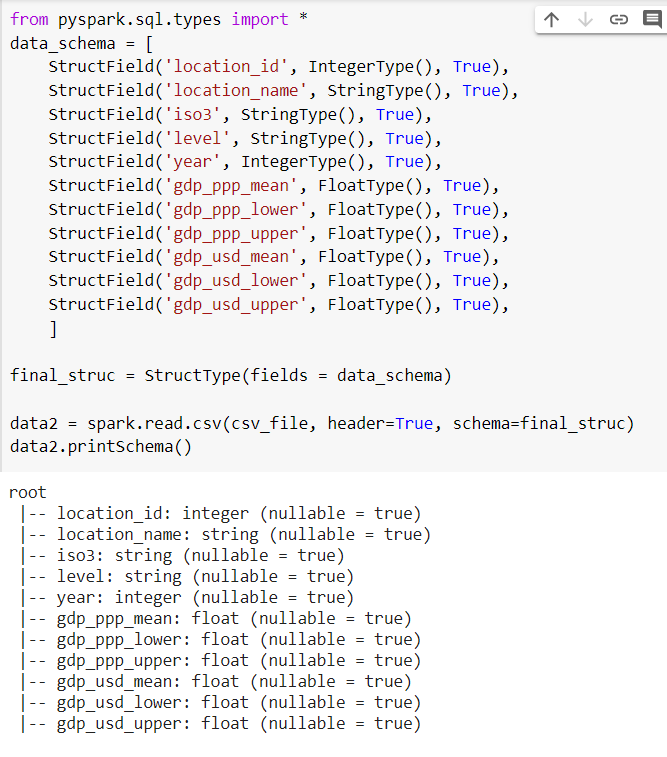




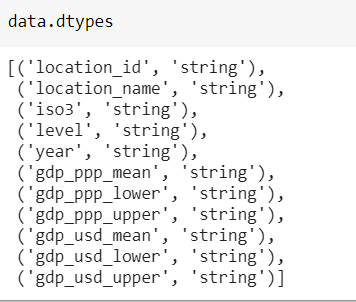
Zadanie 4. Utworzyć schemat danych i zastosować go na zbiorze danych

Wyświetlanie schematu danych

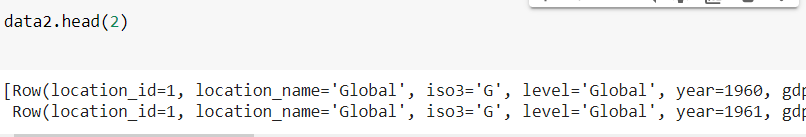


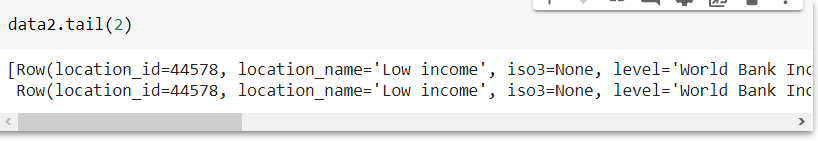


Zadanie 5. Wyświetlić typy danych w poszczególnych kolumnach

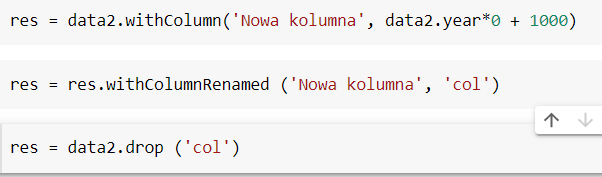


Zadanie 6. Pokazać 2 pierwsze i 2 ostatnie wiersze danych

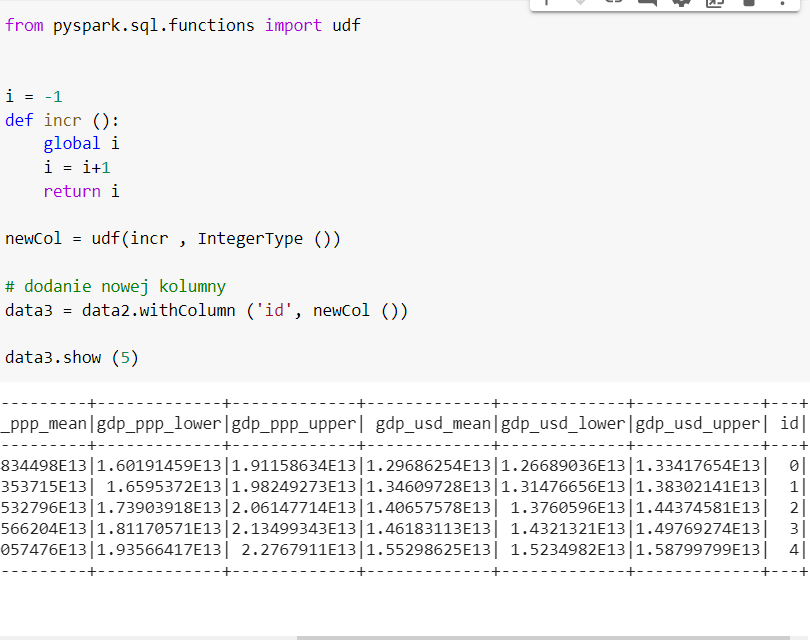




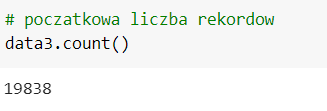
Zadanie 7. Dodać nową kolumnę zawierającą zera, zmienić jej nazwę a następnie usunąć

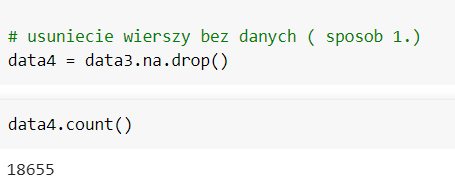


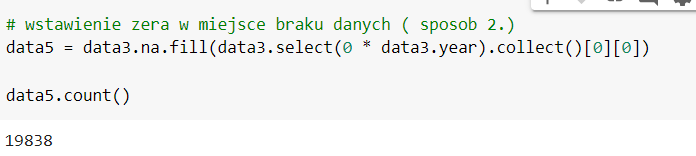
Zadanie 8. Dodać do danych kolumnę zawierającą numer wiersza



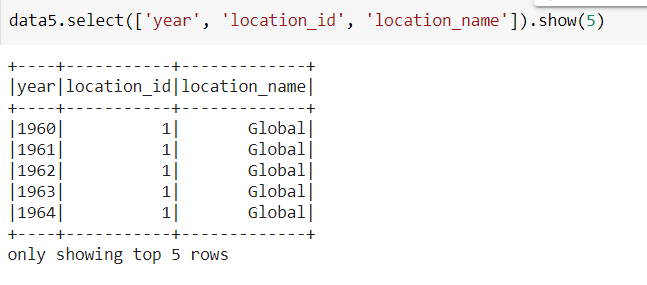
Zadanie 9. Sprawdzić liczbę wierszy danych przed i po usunięciu wierszy bez danych dwoma sposobami.



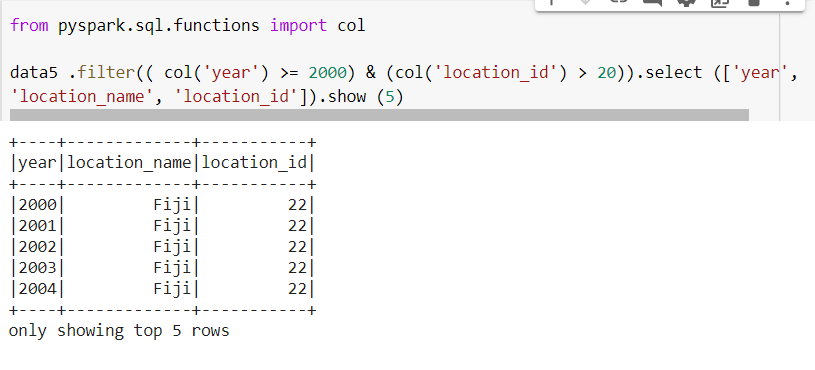




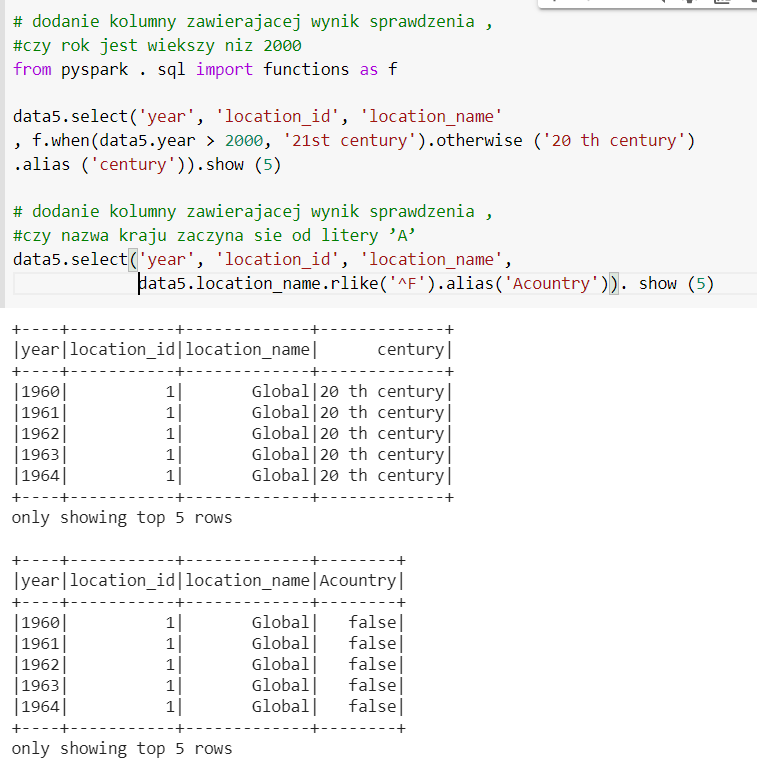
Zadanie 10. Wyświetlić wybrane kolumny danych



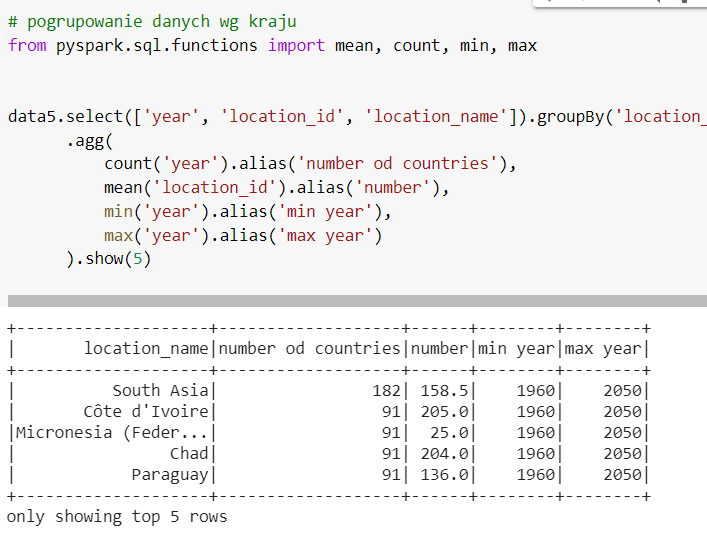
Zadanie 11. Odfiltrować dane zawierające dane od roku 2000



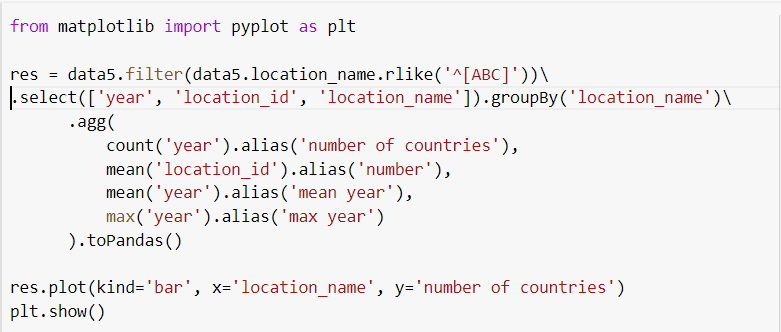
Zadanie 12. Dodać kolumnę zawierającą wynik sprawdzania warunku

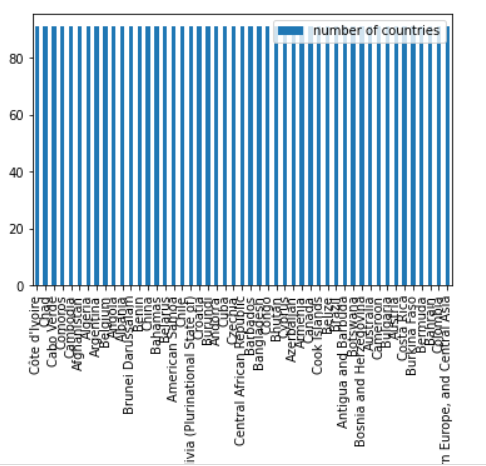


Zadanie 13. Pogrupować dane wg kraju i obliczyć liczbę danych, średnią wartość oraz wartości minimalne i maksymalne



Zadanie 14. Wygenerować wykres słupkowy na podstawie pogrupowanych danych.





**Wnioski**

1. Biblioteka „pyspark” jest wygodną alternatywą dla pakietu „pandas”, zwłaszcza jeśli analizowane są duże zbiory danych.  
Wszelkie operacje na danych wymagają utworzenia sesji „SparkSession” przy użyciu biblioteki builder.

2. Biblioteka „pyspark” pozwala wczytywać dane z różnych źródeł danych:

\* pliki csv

\* pliki json

\* pliki parquet

3. Biblioteka „pyspark” pozwala wczytywać dane z różnych źródeł danych:

# pliki CSV  
spark.write.csv(csv\_file\_path, sep=’,’, header=True)  
  
# pliki JSON i Parquet  
data = spark.write.save(json\_data\_file)  
data = spark.write.save(parquet\_data\_file)

4. Wczytane dane są zwykle w formacie „string”, ale można narzucić im typ przy użyciu schematu.

5. Wybieranie wierszy odbywa się analogicznie jak w bibliotece „pandas”.

6. Możliwe jest dodanie kolumny, zmiana jej nazwy oraz usunięcie.

7. Wartości w nowych kolumnach można też dodawać przy użyciu osobnej funkcji.

8. Wiersze niezawierające danych mogą być łatwo usunięte.