PROJECT REPORT

CSE-413

BIG DATA & IOT LAB

Submitted to:

Tahmina Sultana Priya

Lecturer,Department of CSE

Daffodil International University

Submitted by:

Md.Jannatul Ferdous (191-15-2497)

Mahin Ahmed (191-15-1029)

Saiful Islam Sagor (191-15-2678)

Md.Omer Faruk Tusher (192-15-13122)

Section: PC-A

Department of CSE

Daffodil International University

Submission Date: 06/12/2022

Abstract:

The purpose of our project is to predict the populations growth-rate of different countries over the world. Global population growth is a challenging factor for the human race.

Distributed data processing platforms for cloud computing are important tools for large-scale data analytics. Apache Hadoop MapReduce has become the de facto standard in this space, though its programming interface is relatively low-level, requiring many implementation steps even for simple analysis tasks.

The main aim of this project is to analyze and predict the massive amount of data (world population), with the help of various types of tools such as apache spark which is used for real-time processing and analysis of large amounts of data.

Introduction:

The annual average rate of change of population size, for a given country, territory, or geographic area, during a specified period. It expresses the ratio between the annual increase in the population size and the total population for that year, usually multiplied by 100.

Understanding population growth is important for predicting, managing, monitoring, and eradicating pest and disease outbreaks.

Objectives:

The main objectives of this project are:

- ❖ To predict the number of population and growth rate of 2023.

Tools has been used:

- Colab

Language has been used:

- Python

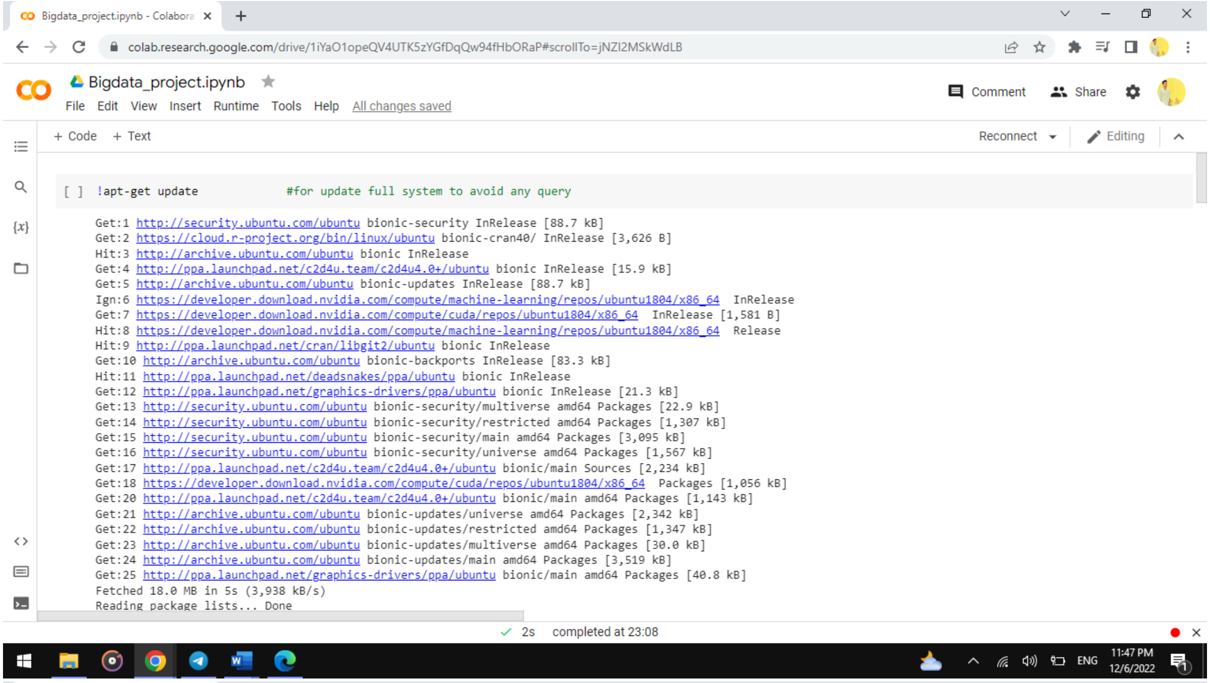
Framework has been used:

- PySpark

Dataset link: Click_here

Project github link: Click_here

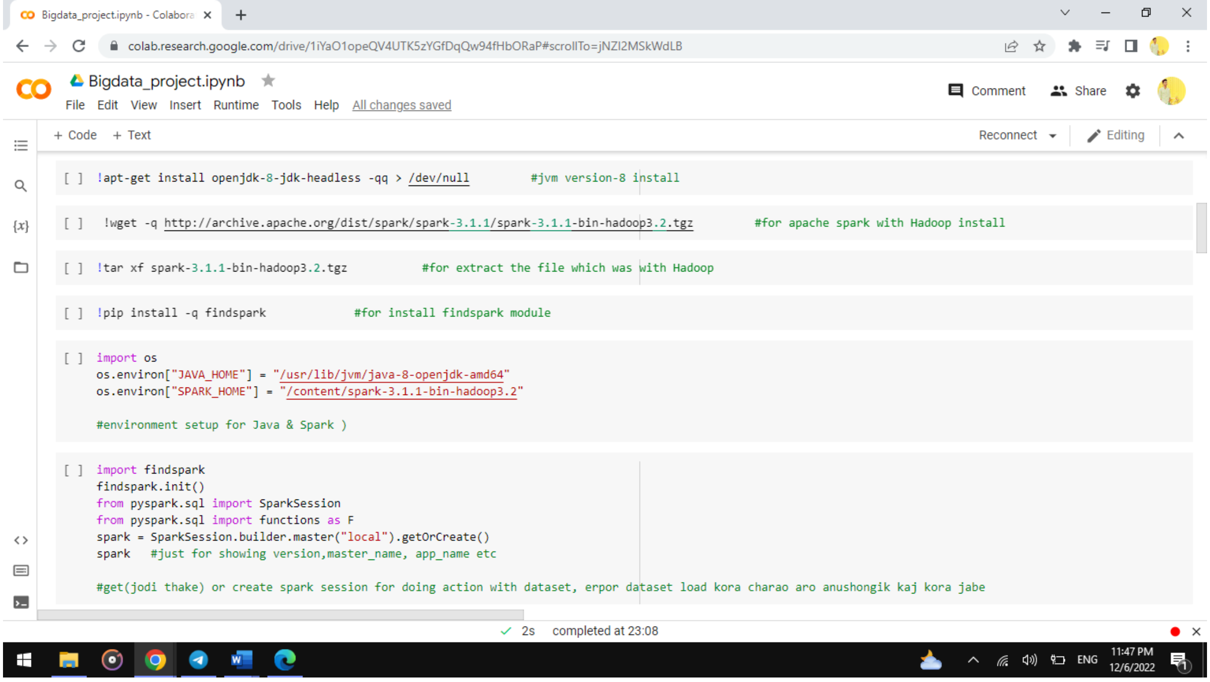
Code & Output:



```
!apt-get update            #for update full system to avoid any query
```

```
Get:1 http://security.ubuntu.com/ubuntu bionic-security InRelease [88.7 kB]
Get:2 https://cloud.r-project.org/bin/linux/ubuntu bionic-cran40/ InRelease [3,626 B]
Hit:3 http://archive.ubuntu.com/ubuntu bionic InRelease
Get:4 http://ppa.launchpad.net/c2d4u.team/c2d4u4.0+/ubuntu bionic InRelease [15.9 kB]
Get:5 http://archive.ubuntu.com/ubuntu bionic-updates InRelease [88.7 kB]
Ign:6 https://developer.download.nvidia.com/compute/machine-learning/repos/ubuntu1804/x86_64  InRelease
Get:7 https://developer.download.nvidia.com/compute/cuda/repos/ubuntu1804/x86_64  InRelease [1,581 B]
Hit:8 https://developer.download.nvidia.com/compute/machine-learning/repos/ubuntu1804/x86_64  Release
Hit:9 http://ppa.launchpad.net/cran/libgit2/ubuntu bionic InRelease
Get:10 http://archive.ubuntu.com/ubuntu bionic-backports InRelease [83.3 kB]
Hit:11 http://ppa.launchpad.net/deadsnakes/ppa/ubuntu bionic InRelease
Get:12 http://ppa.launchpad.net/graphics-drivers/ppa/ubuntu bionic InRelease [21.3 kB]
Get:13 http://security.ubuntu.com/ubuntu bionic-security/multiverse amd64 Packages [22.9 kB]
Get:14 http://security.ubuntu.com/ubuntu bionic-security/restricted amd64 Packages [1,307 kB]
Get:15 http://security.ubuntu.com/ubuntu bionic-security/main amd64 Packages [3,095 kB]
Get:16 http://security.ubuntu.com/ubuntu bionic-security/universe amd64 Packages [1,567 kB]
Get:17 http://ppa.launchpad.net/c2d4u.team/c2d4u4.0+/ubuntu bionic/main Sources [2,234 kB]
Get:18 https://developer.download.nvidia.com/compute/cuda/repos/ubuntu1804/x86_64  Packages [1,056 kB]
Get:20 http://ppa.launchpad.net/c2d4u.team/c2d4u4.0+/ubuntu bionic/main amd64 Packages [1,143 kB]
Get:21 http://archive.ubuntu.com/ubuntu bionic-updates/universe amd64 Packages [2,342 kB]
Get:22 http://archive.ubuntu.com/ubuntu bionic-updates/restricted amd64 Packages [1,347 kB]
Get:23 http://archive.ubuntu.com/ubuntu bionic-updates/multiverse amd64 Packages [30.0 kB]
Get:24 http://archive.ubuntu.com/ubuntu bionic-updates/main amd64 Packages [3,519 kB]
Get:25 http://ppa.launchpad.net/graphics-drivers/ppa/ubuntu bionic/main amd64 Packages [40.8 kB]
Fetched 18.0 MB in 5s (3,938 kB/s)
Reading package lists... Done
```

Bigdata_project.ipynb - Colabora
colab.research.google.com/drive/1iYaO1opeQV4UTK5zYGfDqQw94fHbORaP#scrollTo=jNZI2MSkWdLB
Bigdata_project.ipynb
File Edit View Insert Runtime Tools Help All changes saved
Comment
Share
+ Code + Text
Reconnect
Editing
✓ 2s completed at 23:08
ENG
11:47 PM
12/6/2022



```
!apt-get install openjdk-8-jdk-headless -qq > /dev/null        #jvm version-8 install
```

```
!wget -q http://archive.apache.org/dist/spark/spark-3.1.1/spark-3.1.1-bin-hadoop3.2.tgz        #for apache spark with Hadoop install
```

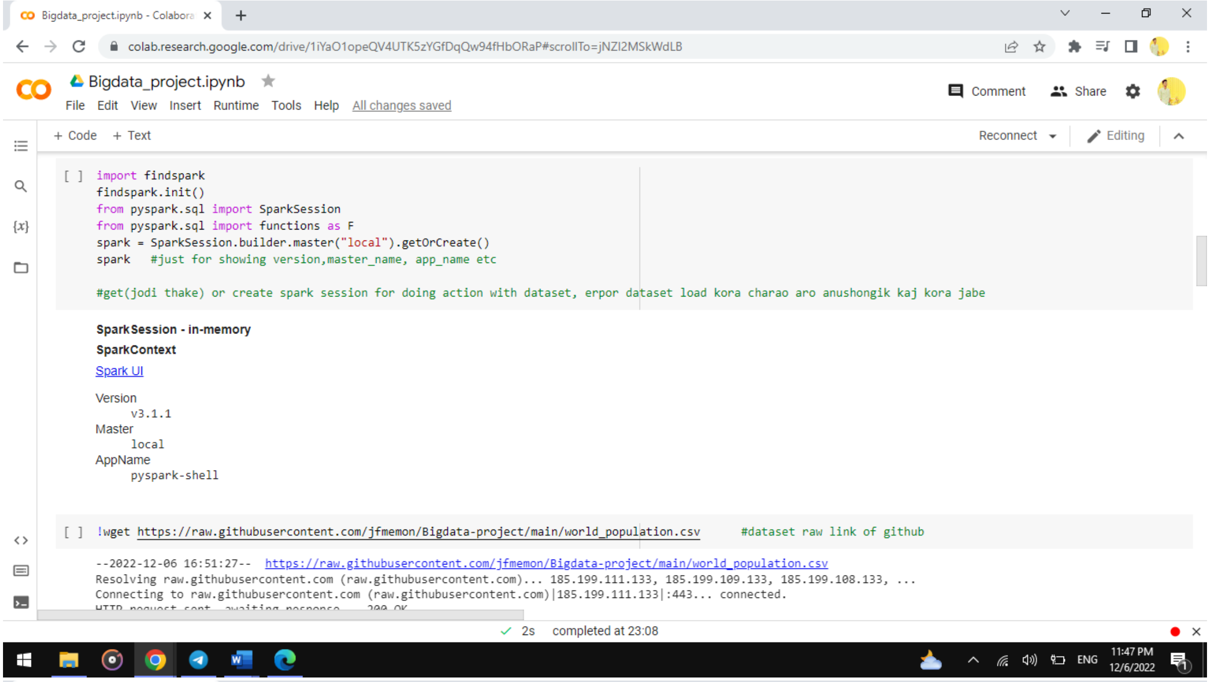
```
!tar xf spark-3.1.1-bin-hadoop3.2.tgz           #for extract the file which was with Hadoop
```

```
!pip install -q findspark            #for install findspark module
```

```
import os
os.environ["JAVA_HOME"] = "/usr/lib/jvm/java-8-openjdk-amd64"
os.environ["SPARK_HOME"] = "/content/spark-3.1.1-bin-hadoop3.2"

#environment setup for Java & Spark )
```

```
import findspark
findspark.init()
from pyspark.sql import SparkSession
from pyspark.sql import functions as F
spark = SparkSession.builder.master("local").getOrCreate()
spark   #just for showing version,master_name, app_name etc

#get(jodi thake) or create spark session for doing action with dataset, erpor dataset load kora charao aro anushongik kaj kora jabe
```

Bigdata_project.ipynb - Colabora
colab.research.google.com/drive/1iYaO1opeQV4UTK5zYGfDqQw94fHbORaP#scrollTo=jNZI2MSkWdLB
Bigdata_project.ipynb
File Edit View Insert Runtime Tools Help All changes saved
Comment
Share
+ Code + Text
Reconnect
Editing
✓ 2s completed at 23:08
ENG
11:47 PM
12/6/2022


Bigdata_project.ipynb - Colabora
colab.research.google.com/drive/1iYaO1opeQV4UTK5zYGfDqQw94fHbORaP#scrollTo=jNZl2MSkWdLB
Bigdata_project.ipynb
File Edit View Insert Runtime Tools Help All changes saved
Comment
Share
+ Code + Text
Reconnect
Editing
✓ 2s completed at 23:08
ENG
11:47 PM
12/6/2022

```
[ ] import findspark
    findspark.init()
    from pyspark.sql import SparkSession
    from pyspark.sql import functions as F
    spark = SparkSession.builder.master("local").getOrCreate()
    spark   #just for showing version,master_name, app_name etc

    #get(jodi thake) or create spark session for doing action with dataset, erpor dataset load kora charao aro anushongik kaj kora jabe
```

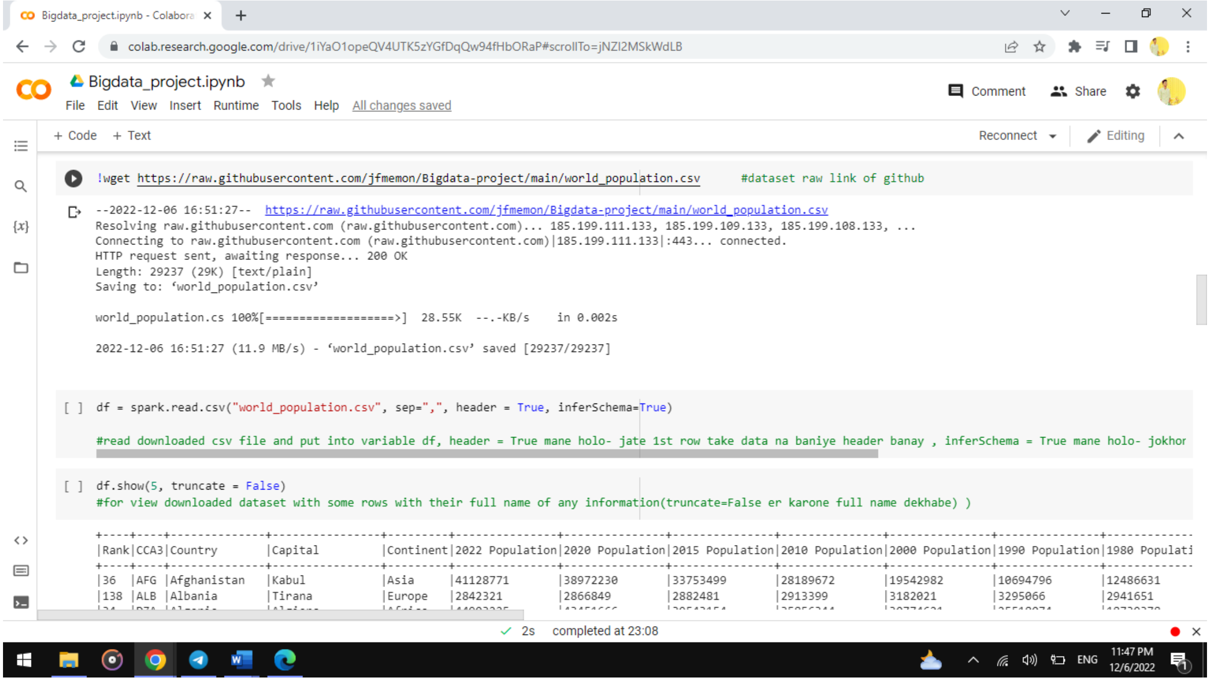
SparkSession - in-memory

SparkContext

Spark UI

Version
    v3.1.1
Master
    local
AppName
    pyspark-shell

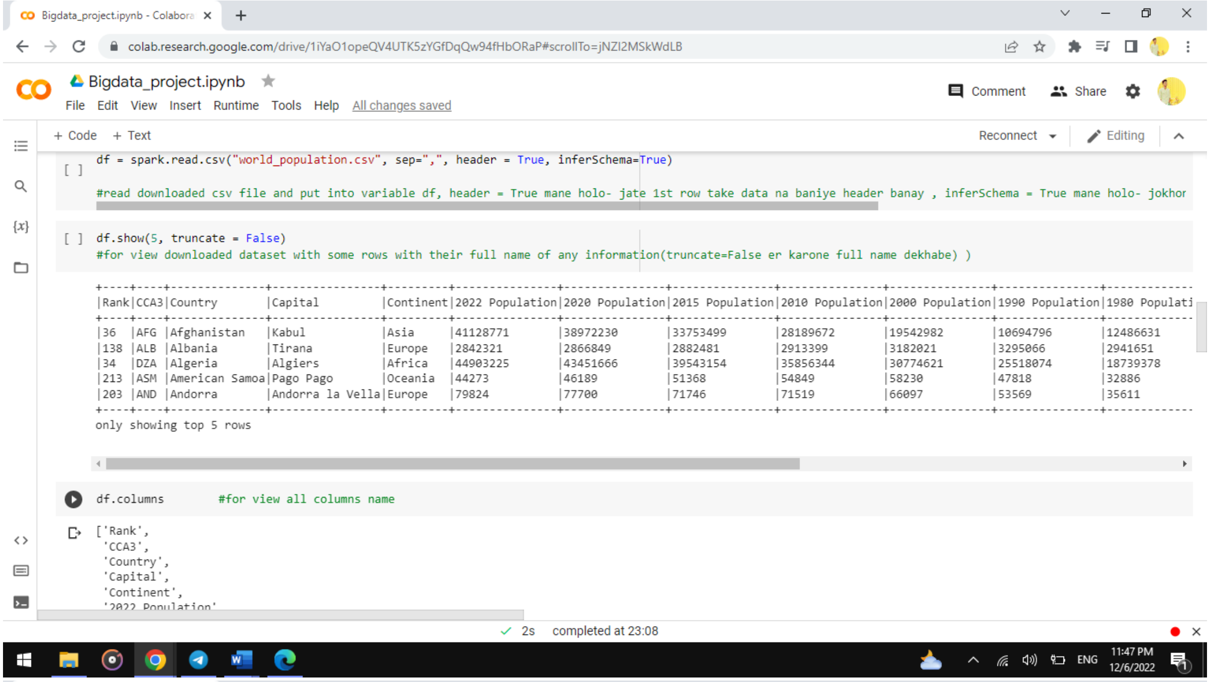
```
[ ] !wget https://raw.githubusercontent.com/jfmemon/Bigdata-project/main/world_population.csv      #dataset raw link of github

    --2022-12-06 16:51:27--  https://raw.githubusercontent.com/jfmemon/Bigdata-project/main/world_population.csv
    Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 185.199.111.133, 185.199.109.133, 185.199.108.133, ...
    Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|185.199.111.133|:443... connected.
    HTTP request sent, awaiting response... 200 OK
```


Bigdata_project.ipynb - Colabora
colab.research.google.com/drive/1iYaO1opeQV4UTK5zYGfDqQw94fHbORaP#scrollTo=jNZl2MSkWdLB
Bigdata_project.ipynb
File Edit View Insert Runtime Tools Help All changes saved
Comment
Share
+ Code + Text
Reconnect
Editing
✓ 2s completed at 23:08
ENG
11:47 PM
12/6/2022

```
!wget https://raw.githubusercontent.com/jfmemon/Bigdata-project/main/world_population.csv      #dataset raw link of github

--2022-12-06 16:51:27--  https://raw.githubusercontent.com/jfmemon/Bigdata-project/main/world_population.csv
Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 185.199.111.133, 185.199.109.133, 185.199.108.133, ...
Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|185.199.111.133|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 29237 (29K) [text/plain]
Saving to: ‘world_population.csv’

world_population.cs 100%[===================>]  28.55K  --.-KB/s    in 0.002s

2022-12-06 16:51:27 (11.9 MB/s) - ‘world_population.csv’ saved [29237/29237]
```

```
[ ] df = spark.read.csv("world_population.csv", sep=",", header = True, inferSchema=True)

    #read downloaded csv file and put into variable df, header = True mane holo- jate 1st row take data na baniye header banay , inferSchema = True mane holo- jokhor
```

```
[ ] df.show(5, truncate = False)
    #for view downloaded dataset with some rows with their full name of any information(truncate=False er karone full name dekhabe) )

    +----+----+---------------+----------------+---------+---------------+---------------+---------------+---------------+---------------+---------------+------------
    |Rank|CCA3|Country        |Capital         |Continent|2022 Population|2020 Population|2015 Population|2010 Population|2000 Population|1990 Population|1980 Populati
    +----+----+---------------+----------------+---------+---------------+---------------+---------------+---------------+---------------+---------------+------------
    |36  |AFG |Afghanistan    |Kabul           |Asia     |41128771       |38972230       |33753499       |28189672       |19542982       |10694796       |12486631
    |138 |ALB |Albania        |Tirana          |Europe   |2842321        |2866849        |2882481        |2913399        |3182021        |3295066        |2941651
```


Bigdata_project.ipynb - Colabora
colab.research.google.com/drive/1iYaO1opeQV4UTK5zYGfDqQw94fHbORaP#scrollTo=jNZI2MSkWdLB
Bigdata_project.ipynb
File Edit View Insert Runtime Tools Help All changes saved
Comment Share
+ Code + Text
Reconnect
Editing

```
df = spark.read.csv("world_population.csv", sep=",", header = True, inferSchema=True)

#read downloaded csv file and put into variable df, header = True mane holo- jate 1st row take data na baniye header banay , inferSchema = True mane holo- jokhor
```

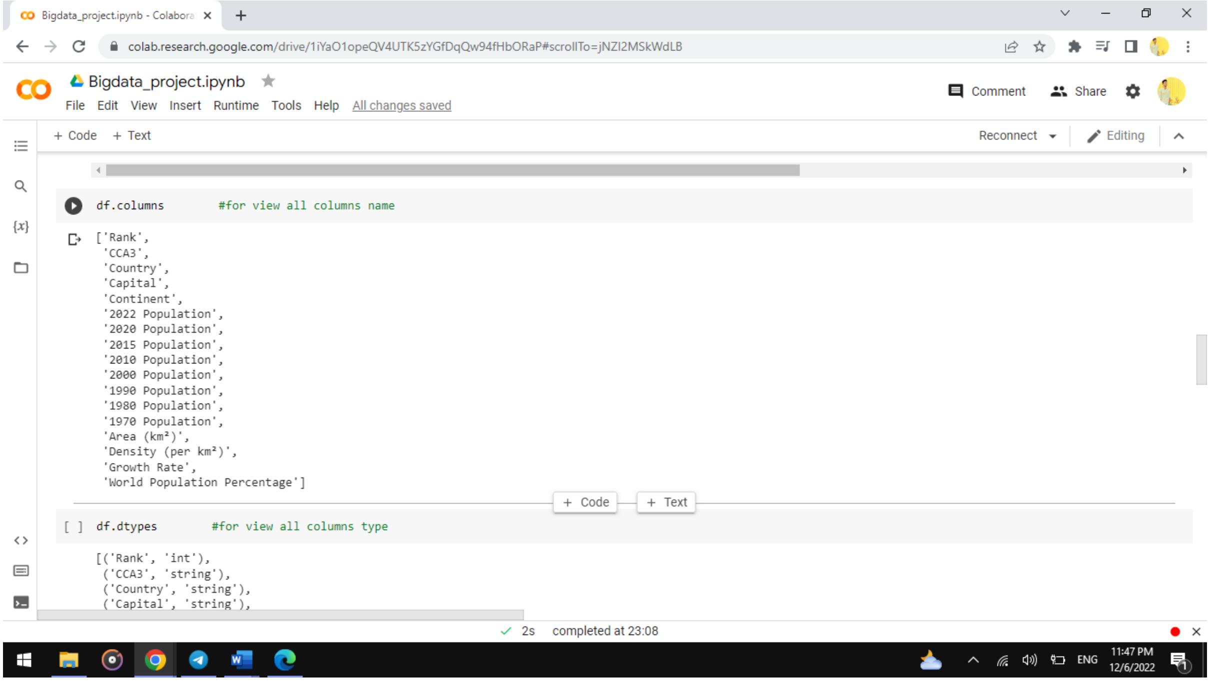
```
df.show(5, truncate = False)
#for view downloaded dataset with some rows with their full name of any information(truncate=False er karone full name dekhabe) )
```

```
+----+----+--------------+----------------+---------+---------------+---------------+---------------+---------------+---------------+---------------+-----------
|Rank|CCA3|Country       |Capital         |Continent|2022 Population|2020 Population|2015 Population|2010 Population|2000 Population|1990 Population|1980 Populati
+----+----+--------------+----------------+---------+---------------+---------------+---------------+---------------+---------------+---------------+-----------
|36  |AFG |Afghanistan   |Kabul           |Asia     |41128771       |38972230       |33753499       |28189672       |19542982       |10694796       |12486631
|138 |ALB |Albania       |Tirana          |Europe   |2842321        |2866849        |2882481        |2913399        |3182021        |3295066        |2941651
|34  |DZA |Algeria       |Algiers         |Africa   |44903225       |43451666       |39543154       |35856344       |30774621       |25518074       |18739378
|213 |ASM |American Samoa|Pago Pago       |Oceania  |44273          |46189          |51368          |54849          |58230          |47818          |32886
|203 |AND |Andorra       |Andorra la Vella|Europe   |79824          |77700          |71746          |71519          |66097          |53569          |35611
+----+----+--------------+----------------+---------+---------------+---------------+---------------+---------------+---------------+---------------+-----------
only showing top 5 rows
```

```
df.columns        #for view all columns name
```

```
['Rank',
 'CCA3',
 'Country',
 'Capital',
 'Continent',
 '2022 Population'
```

2s completed at 23:08
11:47 PM
12/6/2022
ENG


Bigdata_project.ipynb - Colabora
colab.research.google.com/drive/1iYaO1opeQV4UTK5zYGfDqQw94fHbORaP#scrollTo=jNZI2MSkWdLB
Bigdata_project.ipynb
File Edit View Insert Runtime Tools Help All changes saved
Comment Share
+ Code + Text
Reconnect
Editing

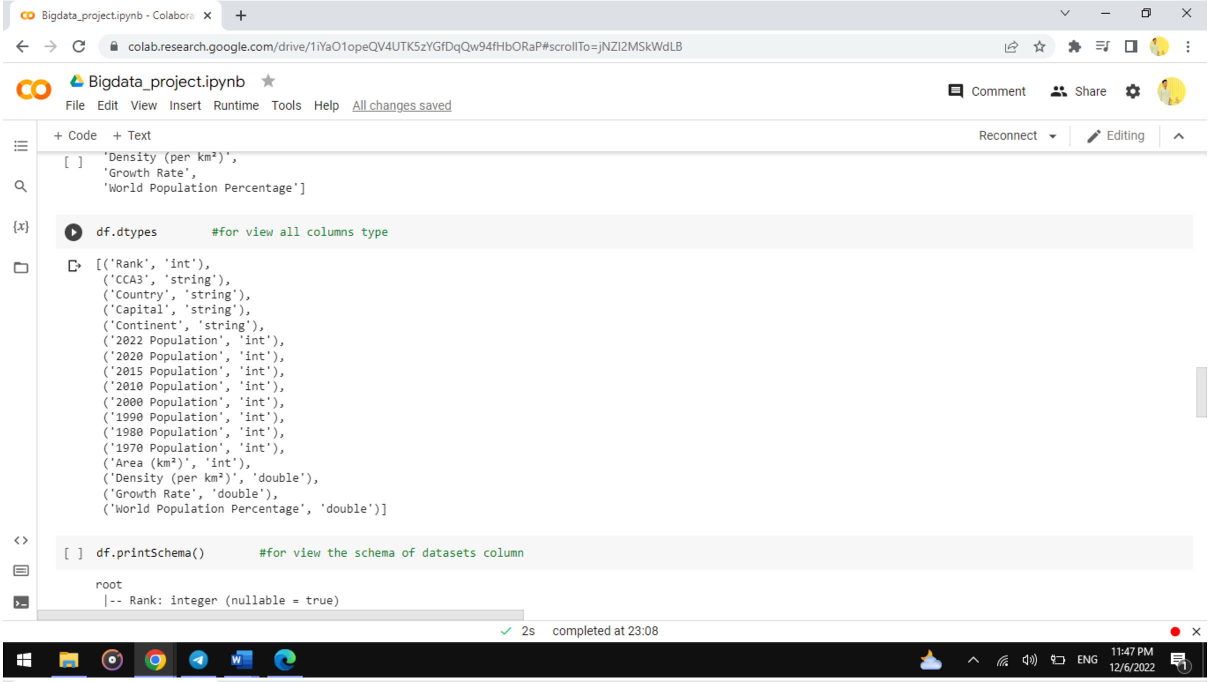
```
df.columns        #for view all columns name
```

```
['Rank',
 'CCA3',
 'Country',
 'Capital',
 'Continent',
 '2022 Population',
 '2020 Population',
 '2015 Population',
 '2010 Population',
 '2000 Population',
 '1990 Population',
 '1980 Population',
 '1970 Population',
 'Area (km²)',
 'Density (per km²)',
 'Growth Rate',
 'World Population Percentage']
```

```
df.dtypes        #for view all columns type
```

```
[('Rank', 'int'),
 ('CCA3', 'string'),
 ('Country', 'string'),
 ('Capital', 'string'),
```

2s completed at 23:08
11:47 PM
12/6/2022
ENG


Bigdata_project.ipynb - Colabora
colab.research.google.com/drive/1iYaO1opeQV4UTK5zYGfDqQw94fHbORaP#scrollTo=jNZl2MSkWdLB
Bigdata_project.ipynb
File Edit View Insert Runtime Tools Help All changes saved
Comment Share
+ Code + Text
Reconnect Editing
✓ 2s completed at 23:08
ENG 11:47 PM 12/6/2022

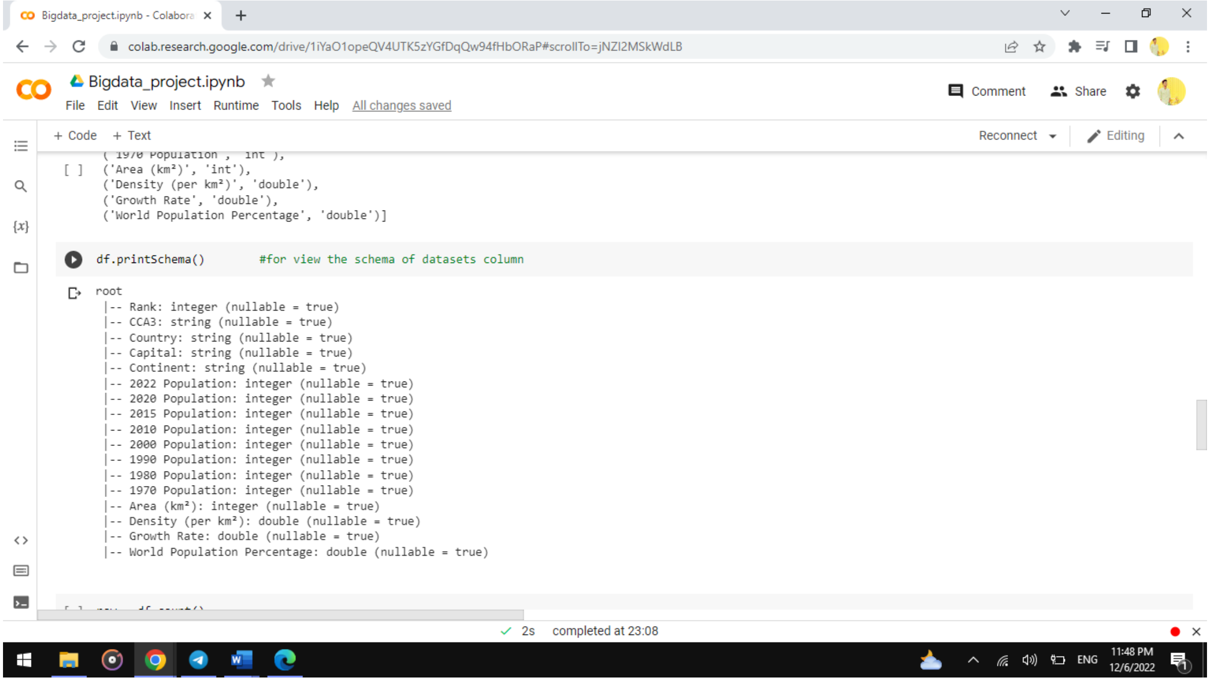
```
[ ]  'Density (per km²)',
     'Growth Rate',
     'World Population Percentage']
```

```
df.dtypes        #for view all columns type
```

```
[('Rank', 'int'),
 ('CCA3', 'string'),
 ('Country', 'string'),
 ('Capital', 'string'),
 ('Continent', 'string'),
 ('2022 Population', 'int'),
 ('2020 Population', 'int'),
 ('2015 Population', 'int'),
 ('2010 Population', 'int'),
 ('2000 Population', 'int'),
 ('1990 Population', 'int'),
 ('1980 Population', 'int'),
 ('1970 Population', 'int'),
 ('Area (km²)', 'int'),
 ('Density (per km²)', 'double'),
 ('Growth Rate', 'double'),
 ('World Population Percentage', 'double')]
```

```
[ ]  df.printSchema()        #for view the schema of datasets column
```

```
root
 |-- Rank: integer (nullable = true)
```

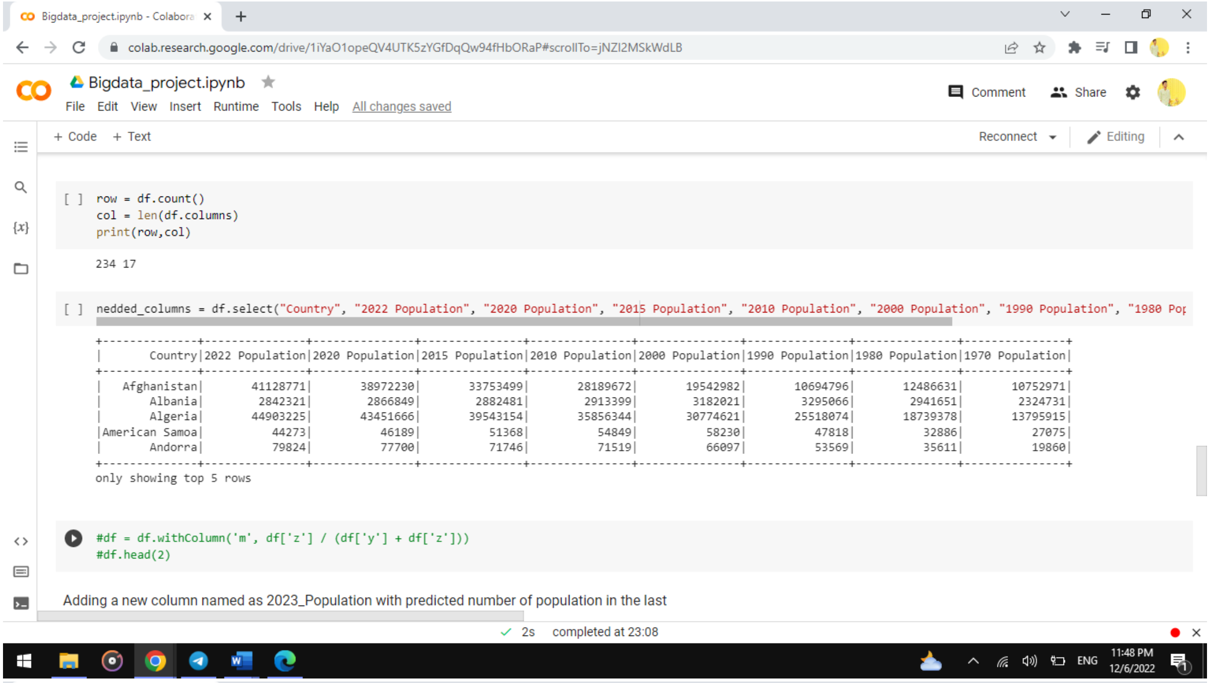

Bigdata_project.ipynb - Colabora
colab.research.google.com/drive/1iYaO1opeQV4UTK5zYGfDqQw94fHbORaP#scrollTo=jNZl2MSkWdLB
Bigdata_project.ipynb
File Edit View Insert Runtime Tools Help All changes saved
Comment Share
+ Code + Text
Reconnect Editing
✓ 2s completed at 23:08
ENG 11:48 PM 12/6/2022

```
[ ]  ('1970 Population', 'int'),
     ('Area (km²)', 'int'),
     ('Density (per km²)', 'double'),
     ('Growth Rate', 'double'),
     ('World Population Percentage', 'double')]
```

```
df.printSchema()        #for view the schema of datasets column
```

```
root
 |-- Rank: integer (nullable = true)
 |-- CCA3: string (nullable = true)
 |-- Country: string (nullable = true)
 |-- Capital: string (nullable = true)
 |-- Continent: string (nullable = true)
 |-- 2022 Population: integer (nullable = true)
 |-- 2020 Population: integer (nullable = true)
 |-- 2015 Population: integer (nullable = true)
 |-- 2010 Population: integer (nullable = true)
 |-- 2000 Population: integer (nullable = true)
 |-- 1990 Population: integer (nullable = true)
 |-- 1980 Population: integer (nullable = true)
 |-- 1970 Population: integer (nullable = true)
 |-- Area (km²): integer (nullable = true)
 |-- Density (per km²): double (nullable = true)
 |-- Growth Rate: double (nullable = true)
 |-- World Population Percentage: double (nullable = true)
```

```
[ ]  row = df.count()
```


Bigdata_project.ipynb - Colabora
colab.research.google.com/drive/1iYaO1opeQV4UTK5zYGfDqQw94fHbORaP#scrollTo=jNZI2MSkWdLB
Bigdata_project.ipynb
File Edit View Insert Runtime Tools Help All changes saved
Comment Share
+ Code + Text
Reconnect Editing

```
row = df.count()
col = len(df.columns)
print(row,col)
```

```
234 17
```

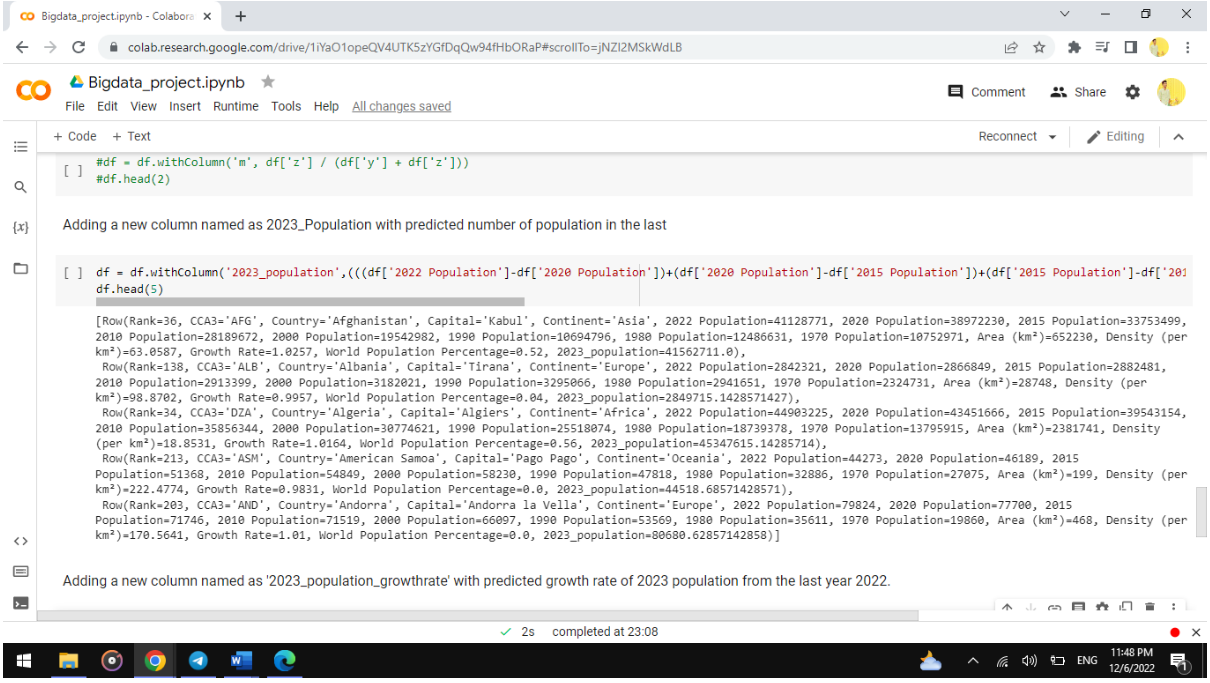
```
nedded_columns = df.select("Country", "2022 Population", "2020 Population", "2015 Population", "2010 Population", "2000 Population", "1990 Population", "1980 Pop
```

```
+--------------+---------------+---------------+---------------+---------------+---------------+---------------+---------------+---------------+
|       Country|2022 Population|2020 Population|2015 Population|2010 Population|2000 Population|1990 Population|1980 Population|1970 Population|
+--------------+---------------+---------------+---------------+---------------+---------------+---------------+---------------+---------------+
|   Afghanistan|       41128771|       38972230|       33753499|       28189672|       19542982|       10694796|       12486631|       10752971|
|       Albania|        2842321|        2866849|        2882481|        2913399|        3182021|        3295066|        2941651|        2324731|
|       Algeria|       44903225|       43451666|       39543154|       35856344|       30774621|       25518074|       18739378|       13795915|
|American Samoa|          44273|          46189|          51368|          54849|          58230|          47818|          32886|          27075|
|       Andorra|          79824|          77700|          71746|          71519|          66097|          53569|          35611|          19860|
+--------------+---------------+---------------+---------------+---------------+---------------+---------------+---------------+---------------+
only showing top 5 rows
```

```
#df = df.withColumn('m', df['z'] / (df['y'] + df['z']))
#df.head(2)
```

Adding a new column named as 2023_Population with predicted number of population in the last

2s completed at 23:08
ENG 11:48 PM 12/6/2022


Bigdata_project.ipynb - Colabora
colab.research.google.com/drive/1iYaO1opeQV4UTK5zYGfDqQw94fHbORaP#scrollTo=jNZI2MSkWdLB
Bigdata_project.ipynb
File Edit View Insert Runtime Tools Help All changes saved
Comment Share
+ Code + Text
Reconnect Editing

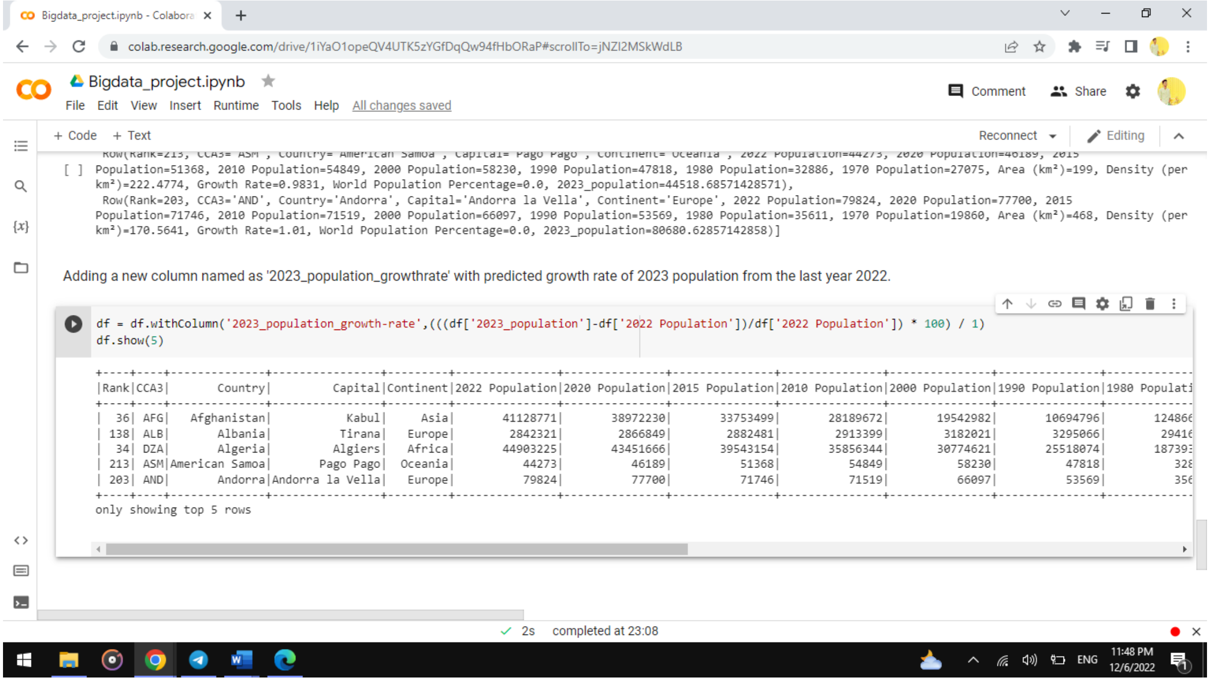
#df = df.withColumn('m', df['z'] / (df['y'] + df['z']))
#df.head(2)

Adding a new column named as 2023_Population with predicted number of population in the last

```
df = df.withColumn('2023_population',(((df['2022 Population']-df['2020 Population'])+(df['2020 Population']-df['2015 Population'])+(df['2015 Population']-df['201
df.head(5)
```

```
[Row(Rank=36, CCA3='AFG', Country='Afghanistan', Capital='Kabul', Continent='Asia', 2022 Population=41128771, 2020 Population=38972230, 2015 Population=33753499,
2010 Population=28189672, 2000 Population=19542982, 1990 Population=10694796, 1980 Population=12486631, 1970 Population=10752971, Area (km²)=652230, Density (per
km²)=63.0587, Growth Rate=1.0257, World Population Percentage=0.52, 2023_population=41562711.0),
 Row(Rank=138, CCA3='ALB', Country='Albania', Capital='Tirana', Continent='Europe', 2022 Population=2842321, 2020 Population=2866849, 2015 Population=2882481,
2010 Population=2913399, 2000 Population=3182021, 1990 Population=3295066, 1980 Population=2941651, 1970 Population=2324731, Area (km²)=28748, Density (per
km²)=98.8702, Growth Rate=0.9957, World Population Percentage=0.04, 2023_population=2849715.1428571427),
 Row(Rank=34, CCA3='DZA', Country='Algeria', Capital='Algiers', Continent='Africa', 2022 Population=44903225, 2020 Population=43451666, 2015 Population=39543154,
2010 Population=35856344, 2000 Population=30774621, 1990 Population=25518074, 1980 Population=18739378, 1970 Population=13795915, Area (km²)=2381741, Density
(per km²)=18.8531, Growth Rate=1.0164, World Population Percentage=0.56, 2023_population=45347615.14285714),
 Row(Rank=213, CCA3='ASM', Country='American Samoa', Capital='Pago Pago', Continent='Oceania', 2022 Population=44273, 2020 Population=46189, 2015
Population=51368, 2010 Population=54849, 2000 Population=58230, 1990 Population=47818, 1980 Population=32886, 1970 Population=27075, Area (km²)=199, Density (per
km²)=222.4774, Growth Rate=0.9831, World Population Percentage=0.0, 2023_population=44518.68571428571),
 Row(Rank=203, CCA3='AND', Country='Andorra', Capital='Andorra la Vella', Continent='Europe', 2022 Population=79824, 2020 Population=77700, 2015
Population=71746, 2010 Population=71519, 2000 Population=66097, 1990 Population=53569, 1980 Population=35611, 1970 Population=19860, Area (km²)=468, Density (per
km²)=170.5641, Growth Rate=1.01, World Population Percentage=0.0, 2023_population=80680.62857142858)]
```

Adding a new column named as '2023_population_growthrate' with predicted growth rate of 2023 population from the last year 2022.



colab.research.google.com/drive/1iYaO1opeQV4UTK5zYGfDqQw94fHbORaP#scrollTo=jNZI2MSkWdLB

Bigdata_project.ipynb

```
Row(Rank=213, CCA3='ASM', Country='American Samoa', Capital='Pago Pago', Continent='Oceania', 2022 Population=44273, 2020 Population=46189, 2015
Population=51368, 2010 Population=54849, 2000 Population=58230, 1990 Population=47818, 1980 Population=32886, 1970 Population=27075, Area (km²)=199, Density (per
km²)=222.4774, Growth Rate=0.9831, World Population Percentage=0.0, 2023_population=44518.68571428571),
 Row(Rank=203, CCA3='AND', Country='Andorra', Capital='Andorra la Vella', Continent='Europe', 2022 Population=79824, 2020 Population=77700, 2015
Population=71746, 2010 Population=71519, 2000 Population=66097, 1990 Population=53569, 1980 Population=35611, 1970 Population=19860, Area (km²)=468, Density (per
km²)=170.5641, Growth Rate=1.01, World Population Percentage=0.0, 2023_population=80680.62857142858)]
```

Adding a new column named as '2023_population_growthrate' with predicted growth rate of 2023 population from the last year 2022.

```
df = df.withColumn('2023_population_growth-rate',(((df['2023_population']-df['2022 Population'])/df['2022 Population']) * 100) / 1)
df.show(5)
```

```
+----+----+--------------+----------------+---------+---------------+---------------+---------------+---------------+---------------+---------------+-------------
|Rank|CCA3|       Country|         Capital|Continent|2022 Population|2020 Population|2015 Population|2010 Population|2000 Population|1990 Population|1980 Populati
+----+----+--------------+----------------+---------+---------------+---------------+---------------+---------------+---------------+---------------+-------------
|  36| AFG|   Afghanistan|           Kabul|     Asia|       41128771|       38972230|       33753499|       28189672|       19542982|       10694796|       124866
| 138| ALB|       Albania|          Tirana|   Europe|        2842321|        2866849|        2882481|        2913399|        3182021|        3295066|        29416
|  34| DZA|       Algeria|         Algiers|   Africa|       44903225|       43451666|       39543154|       35856344|       30774621|       25518074|       187393
| 213| ASM|American Samoa|       Pago Pago|  Oceania|          44273|          46189|          51368|          54849|          58230|          47818|          328
| 203| AND|       Andorra|Andorra la Vella|   Europe|          79824|          77700|          71746|          71519|          66097|          53569|          356
+----+----+--------------+----------------+---------+---------------+---------------+---------------+---------------+---------------+---------------+-------------
only showing top 5 rows
```

✓ 2s completed at 23:08

REFERENCES

- https://stackoverflow.com/questions/40728017/how-to-do-mathematical-operation-with-two-column-in-dataframe-using-pyspark
- https://pages.uoregon.edu/rgp/PPPM613/class8a.htm
- https://www.google.com/search?q=what+is+population+growth+rate&oq=what+is+population+growth+rate&aqs=chrome.0.0i20i263i512j0i512l8j0i390.8918j1j15&sourceid=chrome&ie=UTF-8
- https://research.google.com/colaboratory/faq.html#:~:text=Colaboratory%2C%20or%20%E2%80%9CColab%E2%80%9D%20for,learning%2C%20data%20analysis%20and%20education.
- https://en.wikipedia.org/wiki/Population_growth
- https://sparkbyexamples.com/pyspark/pyspark-lit-add-literal-constant/#:~:text=PySpark%20SQL%20functions%20lit(),by%20importing%20pyspark.sql.functions
- https://www.youtube.com/watch?v=uZqS6pJnosU&t=470s