Tutorial for World Development Indicators

Made by:

Nachiket Parab(CIN:305079923)

Saket Wadhera(CIN:305086930)

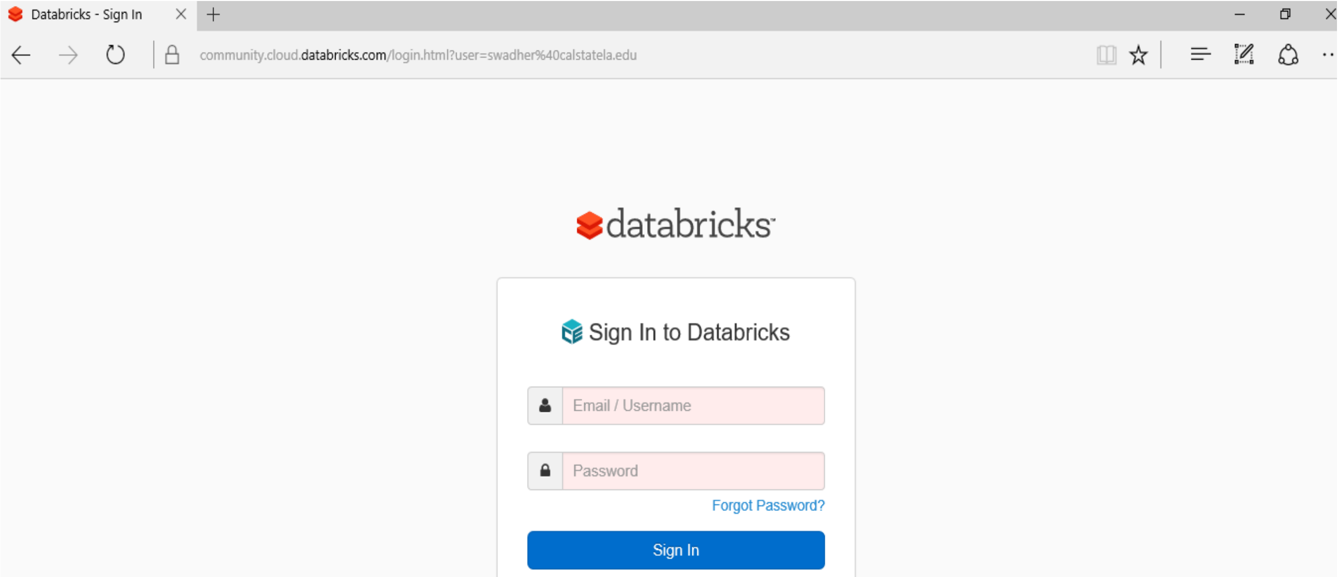
Chanpreet Khanuja(CIN:305073189)

Ishan Fafadia(CIN:305058603)

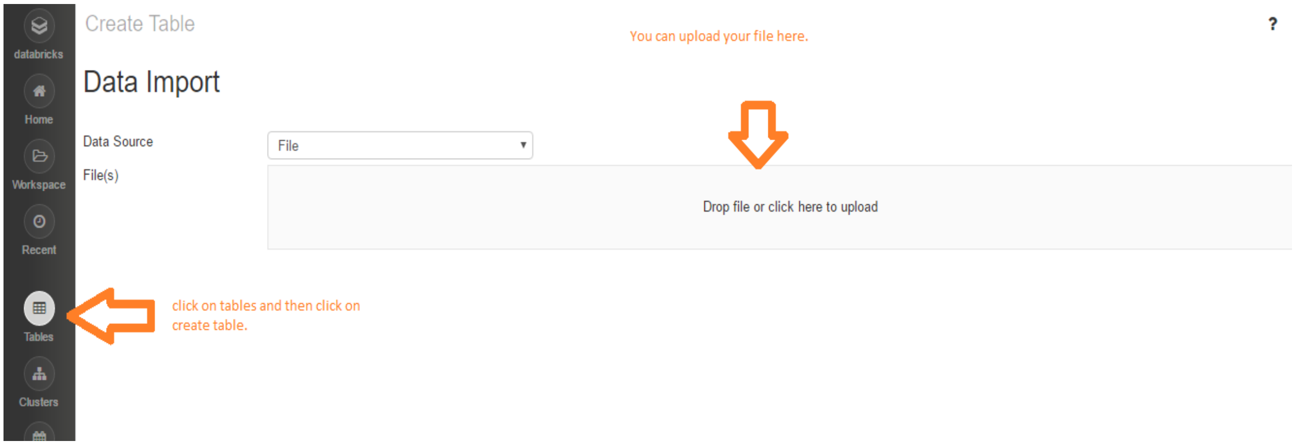
Prerequisite:

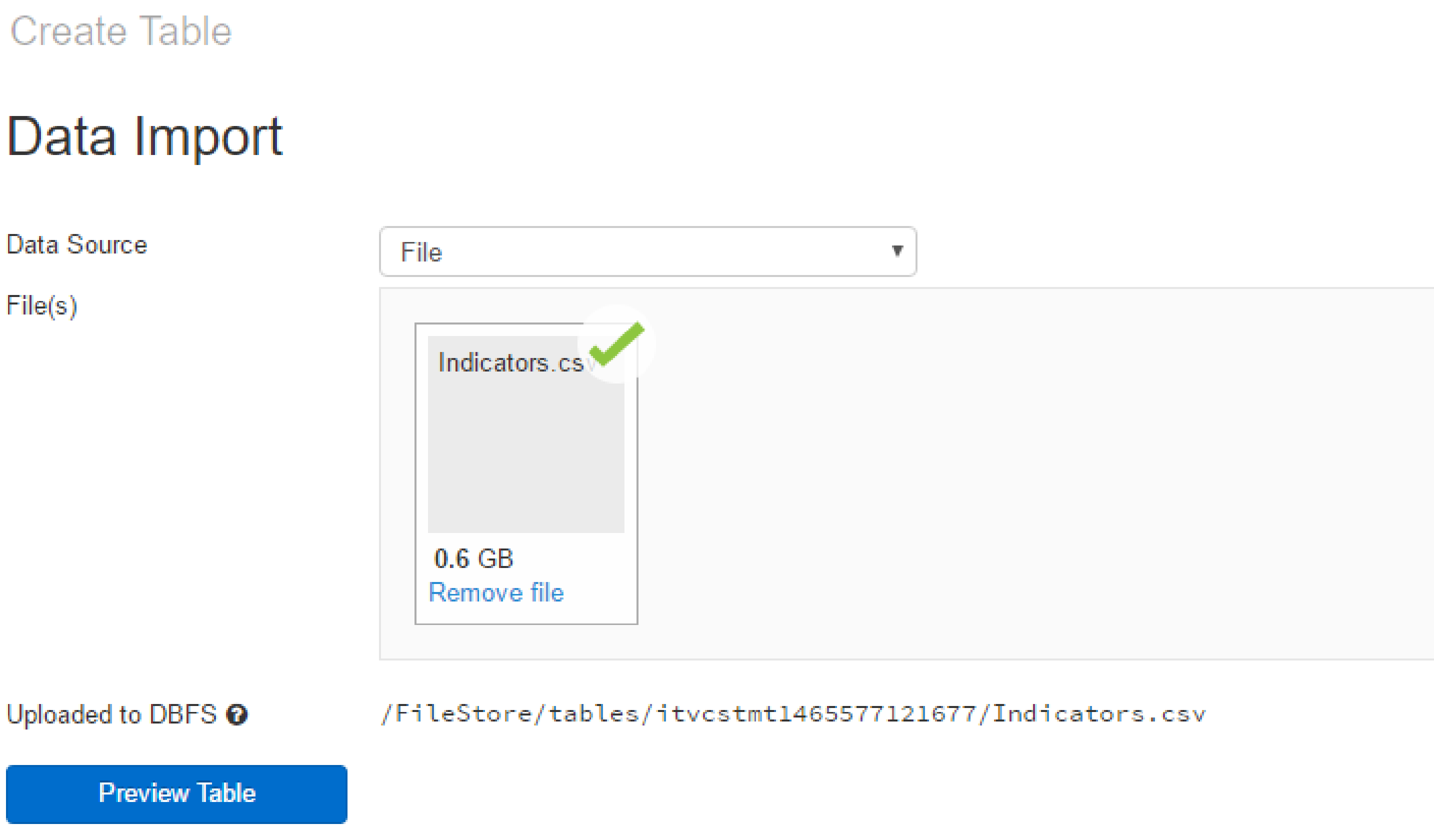
- Get data from Kaggle https://www.kaggle.com/worldbank/world-development-indicators
- You need community databricks account.
- You also need Microsoft Excel, Power BI.

1. Community Databricks Website: https://community.cloud.databricks.com/login.html. Sign in using your credentials. After sign in you will be on the dashboard of your account.



2. Upload the table or file into databricks. In this case the file is of 0.6 GB. Click Tables tab>create table>Select Data Source> Click drag file to upload> Give Table name as Indicators and change the data type of some columns if its required.


Create Table
You can upload your file here.
?
databricks
Home
Workspace
Recent
Tables
Clusters
Data Import
Data Source
File
File(s)
Drop file or click here to upload
click on tables and then click on create table.



Create Table

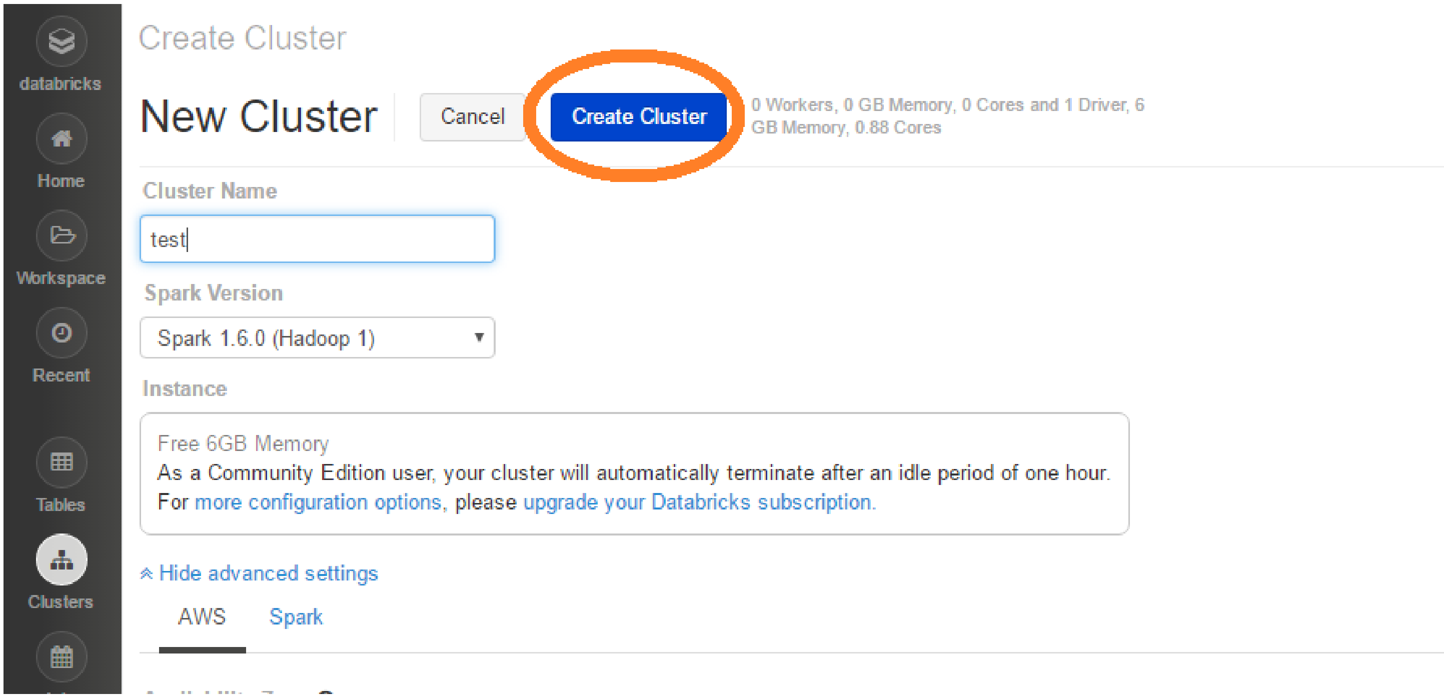
Data Import

Data Source: File

File(s): Indicators.csv 0.6 GB

Uploaded to DBFS: /FileStore/tables/itvcstmt1465577121677/Indicators.csv

3. On the dashboard of your account, click New > Cluster. Select Spark version 1.6.0(Hadoop 1) and click on create cluster. This process will take few minutes to create cluster.



Create Cluster

New Cluster

0 Workers, 0 GB Memory, 0 Cores and 1 Driver, 6 GB Memory, 0.88 Cores

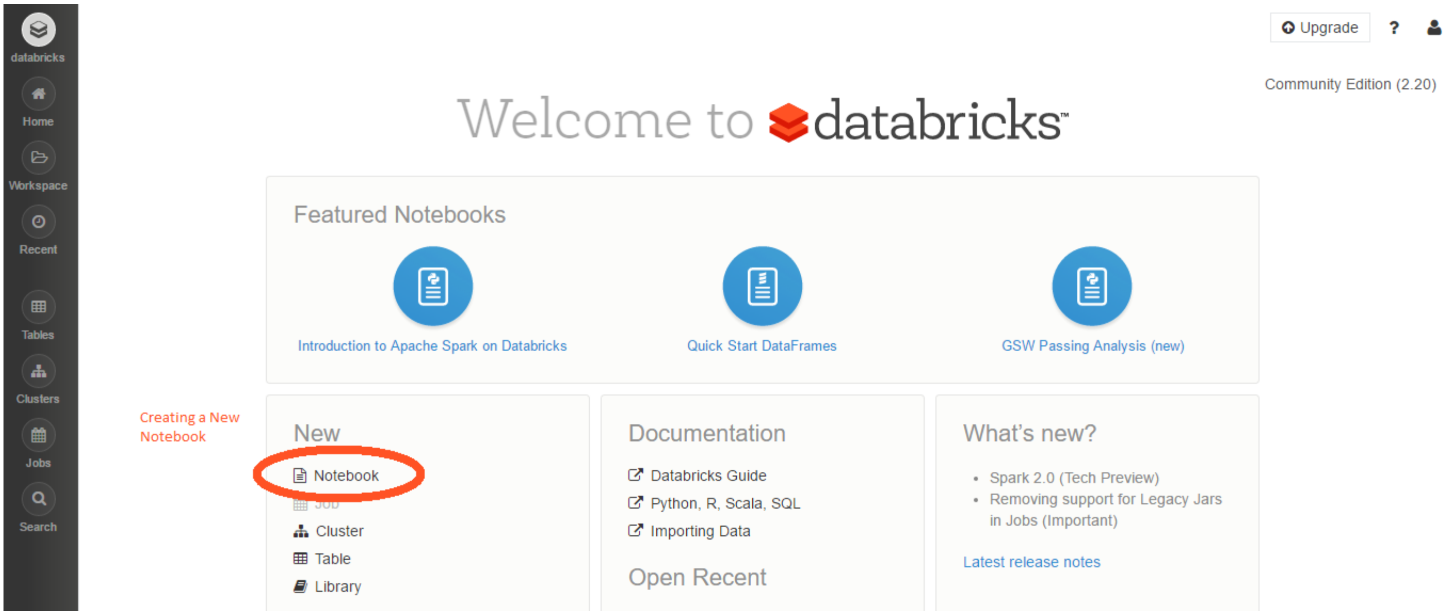
Cluster Name: test

Spark Version: Spark 1.6.0 (Hadoop 1)

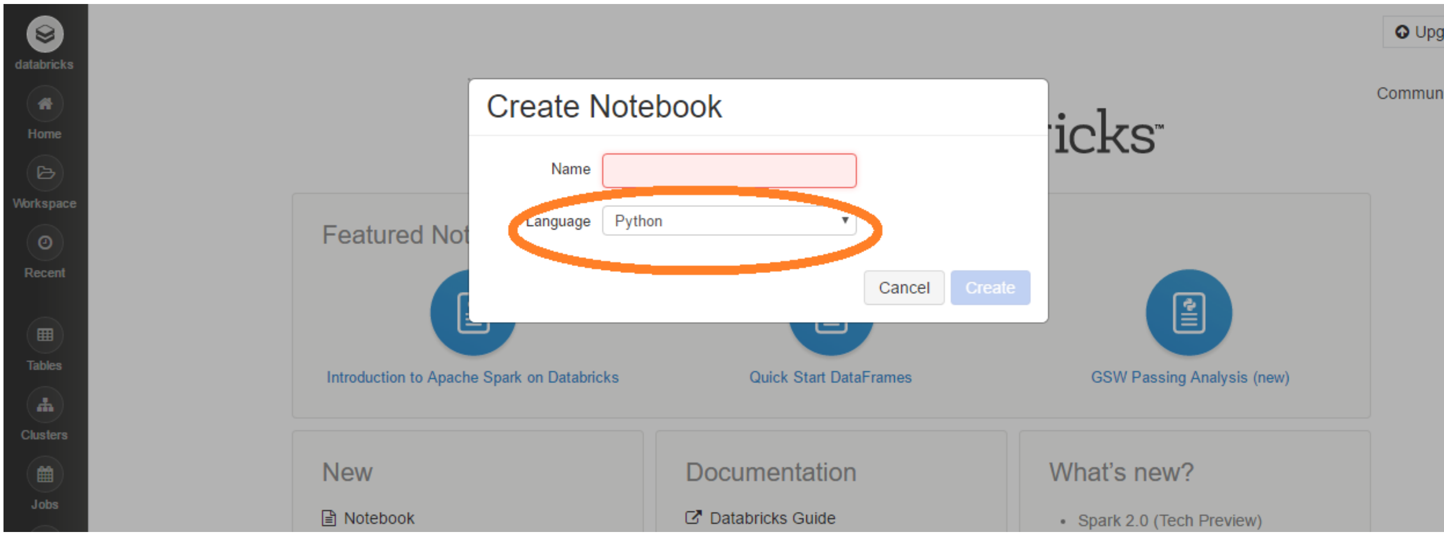
Instance

Free 6GB Memory
As a Community Edition user, your cluster will automatically terminate after an idle period of one hour. For more configuration options, please upgrade your Databricks subscription.

4. Once cluster is created ,On the dashboard of your account, click New > Notebook


Creating a New Notebook

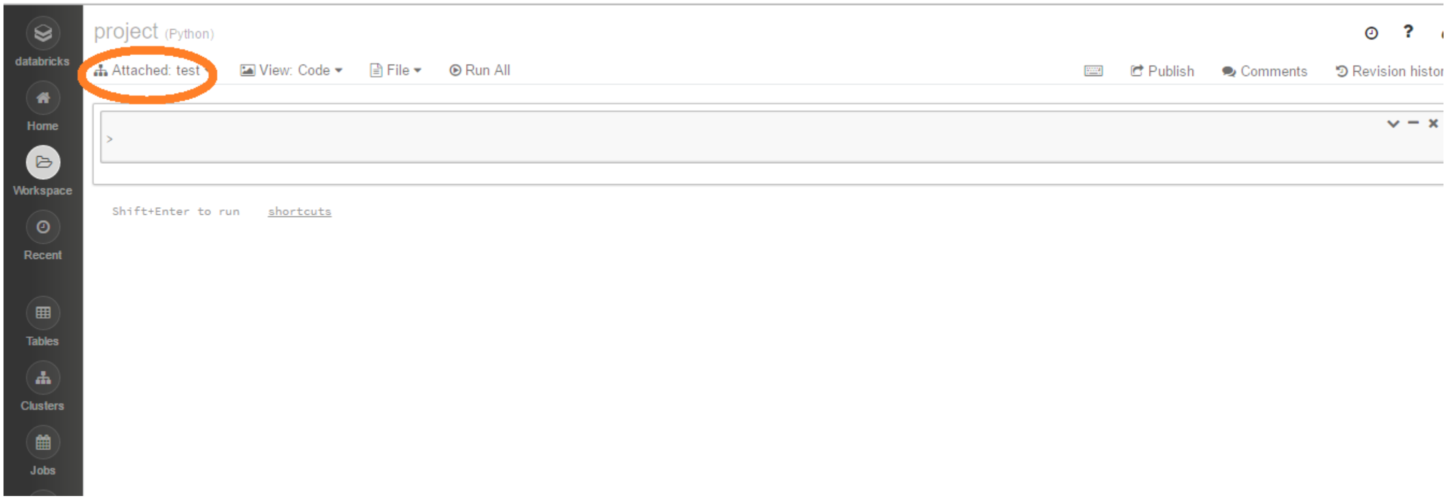
5. New dialog box will appear, fill the required information. Select Language as 'Python' Then Click on Create.


Create Notebook

Name

Language Python

6. Make sure that your notebook is attached to a cluster, which you have created.



7. Create a RDD using SQL context and run the below 4 queries and then display(results) results is the RDD name. You can use a different name also. You can see the below screenshot for the database queries.

Queries:

results = sqlContext.sql(‘SELECT year,(country_name),indicator_name,percentile(cast(value as bigint),0.5) from Indicators where (year between 2012 and 2014) and country_name IN ('United States','China','United Kingdom','India','Japan') and indicator_name IN ('Urban population (% of total)','Population ages 65 and above (% of total)')group by year,country_name,indicator_name order by year’)

> display(results)

results = sqlContext.sql (‘SELECT country_name,indicator_name,percentile(cast(value as bigint),0.5) from INDICATORS_1 where country_name IN ('Kenya','Sudan','European Union','Uganda') and indicator_name IN ('GDP growth (annual %)','Alternative and nuclear energy (% of total energy use)','Deposit interest rate (%)','Trade (% of GDP)')group by country_name,indicator_name’)

-> display(results)

results = sqlContext.sql (‘select country_name,max(value) as value from INDICATORS_1 where indicator_code ="NY.GDP.PCAP.CD" group by country_name order by value desc limit 10’)

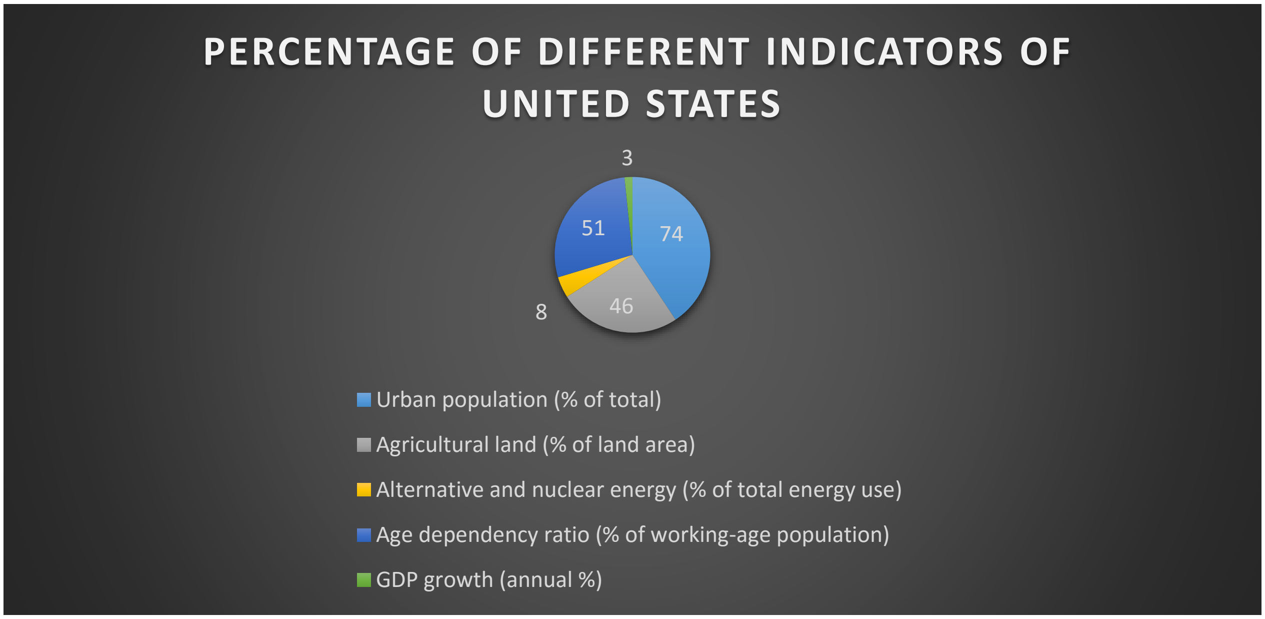
-> display(results)

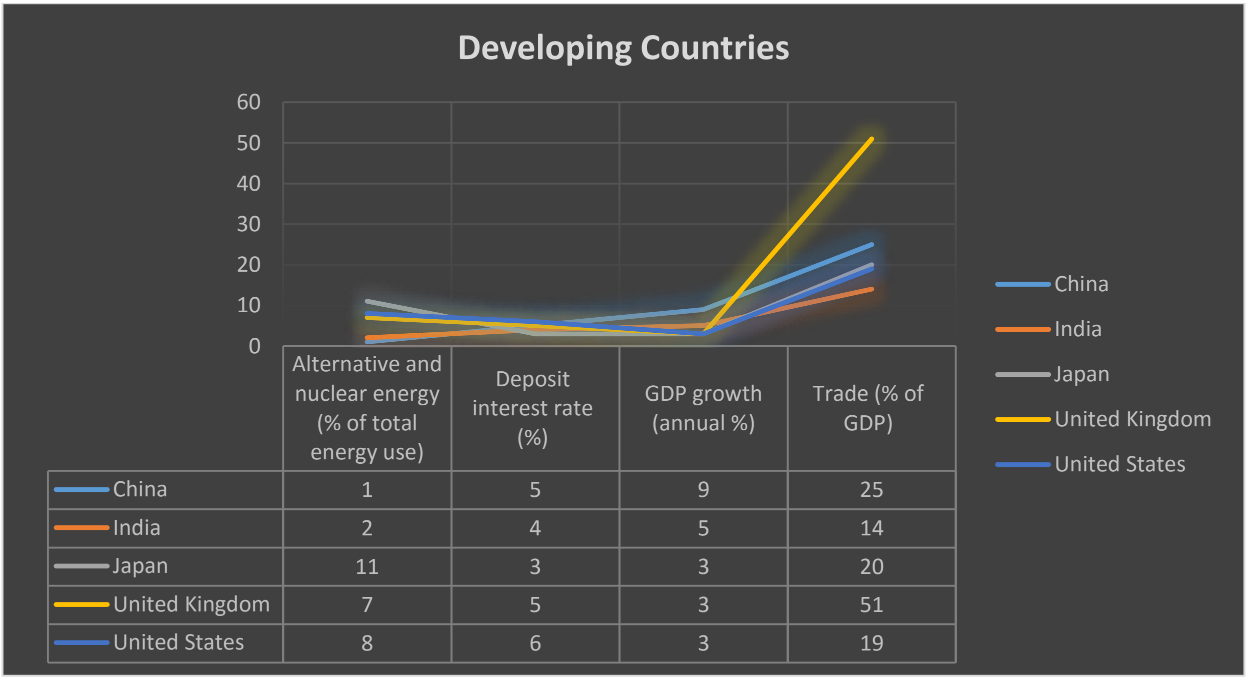
results = sqlContext.sql(‘ SELECT FIRST(country_name),indicator_name,percentile(cast(value as bigint),0.5) from INDICATORS_1 where country_name='United States' and indicator_name IN ('Urban population (% of total)','Age dependency ratio (% of working-age population)','GDP growth (annual %)','Agricultural land (% of land area)', 'Life expectancy at birth, female (years)','Alternative and nuclear energy (% of total energy use)')group by indicator_name')

-> display(results)

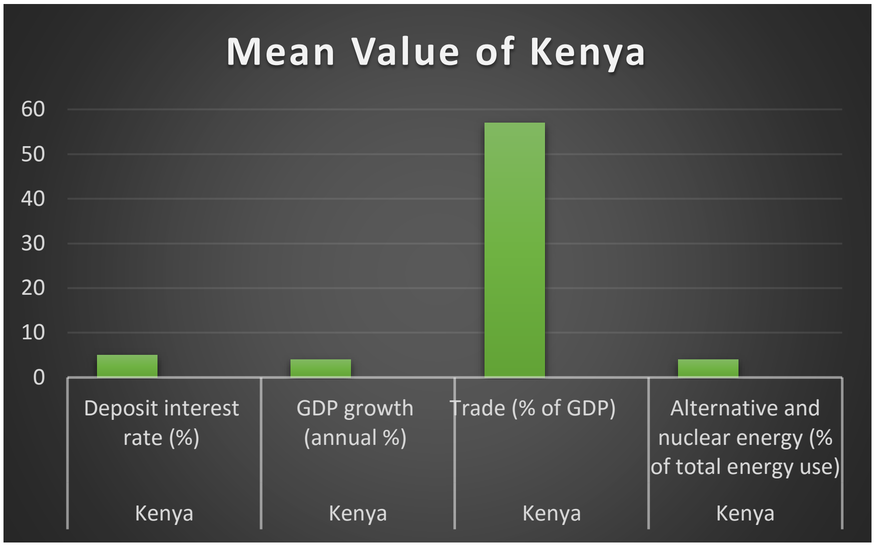
8. Download the analyzed result in local machine.

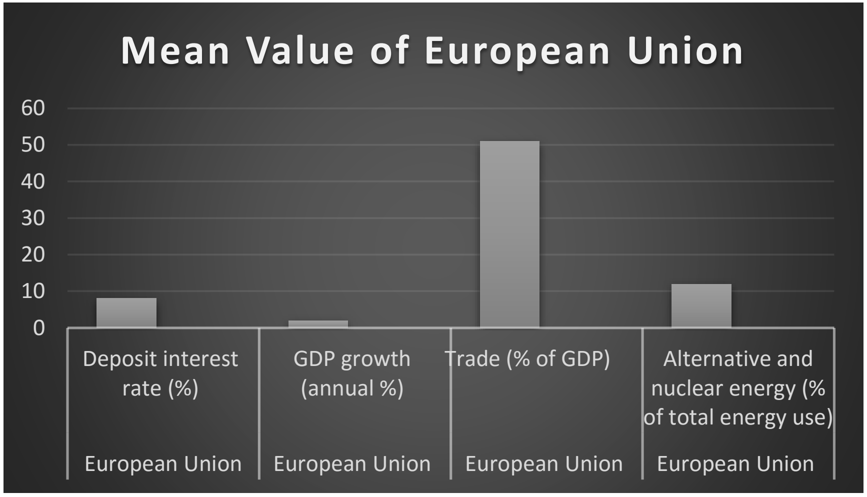
9. Using Microsoft Power BI get the following graph:

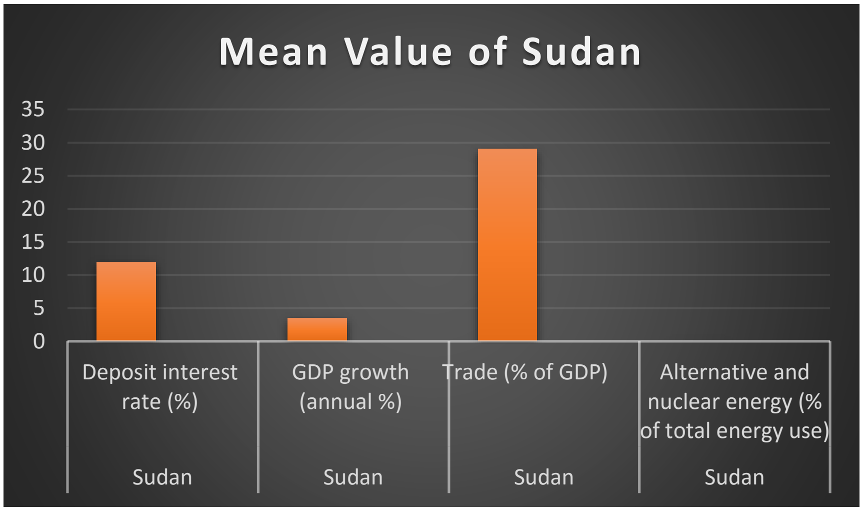


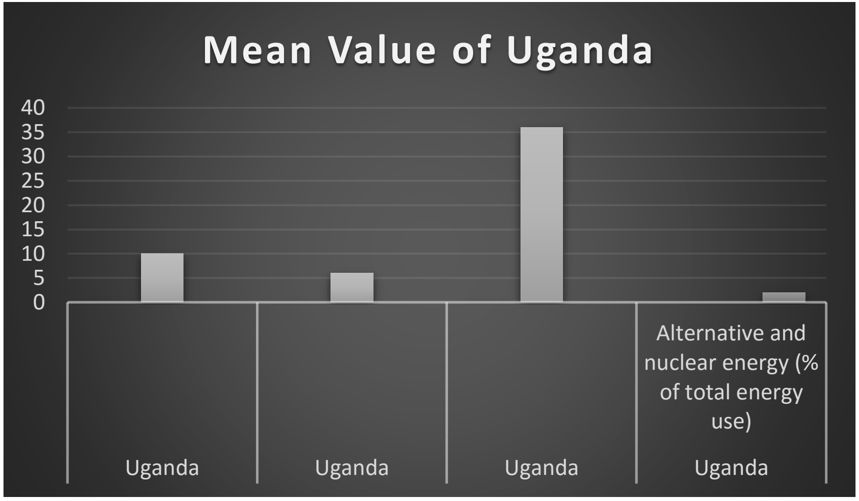


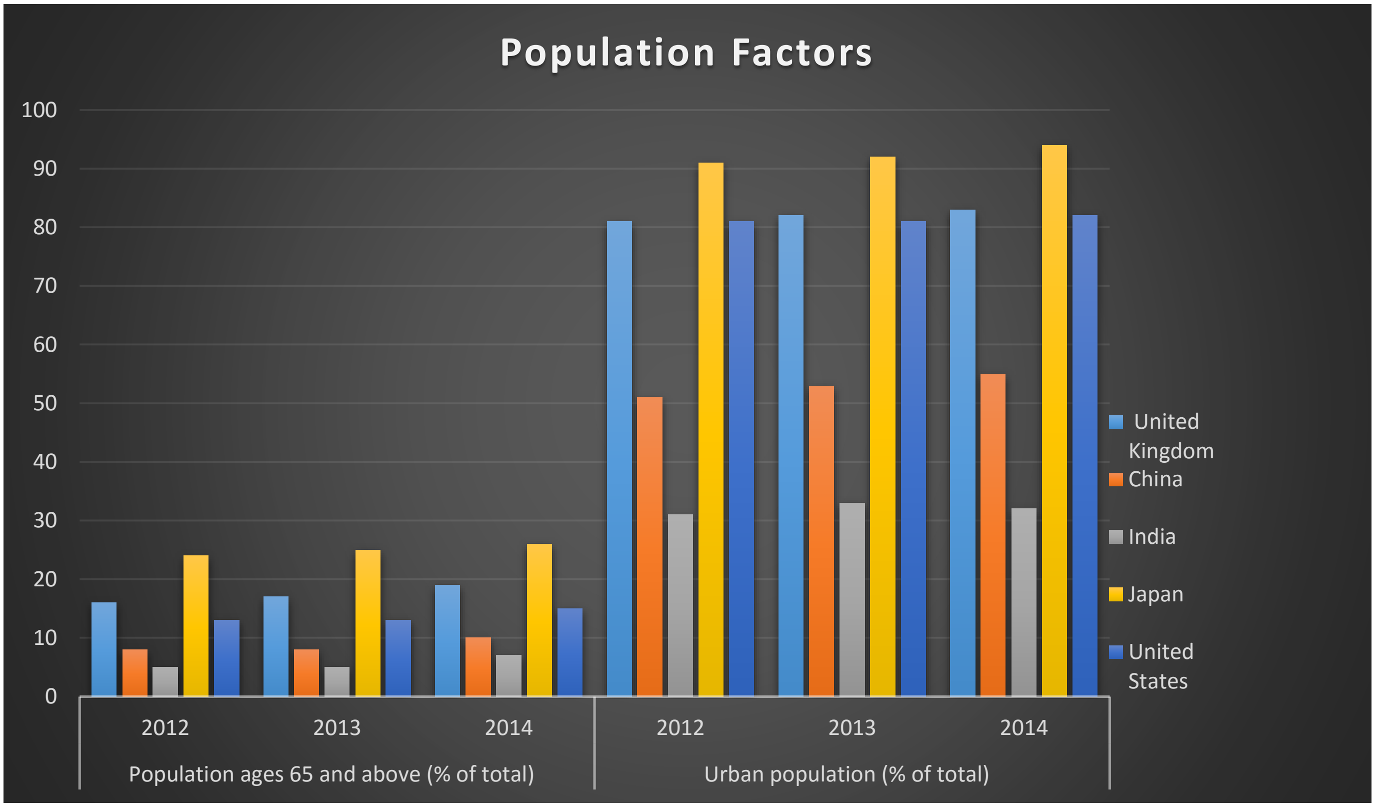
	Alternative and nuclear energy (% of total energy use)	Deposit interest rate (%)	GDP growth (annual %)	Trade (% of GDP)
China	1	5	9	25
India	2	4	5	14
Japan	11	3	3	20
United Kingdom	7	5	3	51
United States	8	6	3	19


Mean Value of Kenya
60
50
40
30
20
10
0
Deposit interest rate (%)
GDP growth (annual %)
Trade (% of GDP)
Alternative and nuclear energy (% of total energy use)
Kenya
Kenya
Kenya
Kenya


Mean Value of European Union
60
50
40
30
20
10
0
Deposit interest rate (%)
GDP growth (annual %)
Trade (% of GDP)
Alternative and nuclear energy (% of total energy use)
European Union
European Union
European Union
European Union


Mean Value of Sudan
35
30
25
20
15
10
5
0
Deposit interest rate (%)
GDP growth (annual %)
Trade (% of GDP)
Alternative and nuclear energy (% of total energy use)
Sudan
Sudan
Sudan
Sudan


Mean Value of Uganda
40
35
30
25
20
15
10
5
0
Alternative and nuclear energy (% of total energy use)
Uganda
Uganda
Uganda
Uganda


Population Factors
100
90
80
70
60
50
40
30
20
10
0
2012
2013
2014
2012
2013
2014
Population ages 65 and above (% of total)
Urban population (% of total)
United Kingdom
China
India
Japan
United States