ASSIGNMENT-11

NAME – NAGA ISWARYA LAKSHMI BHOGADI COMPANY – DXC TECHNOLOGY

BATCH – DXC-262-ANALYTICS-B12-AZURE ROLLNO – DXC-262AB-1205

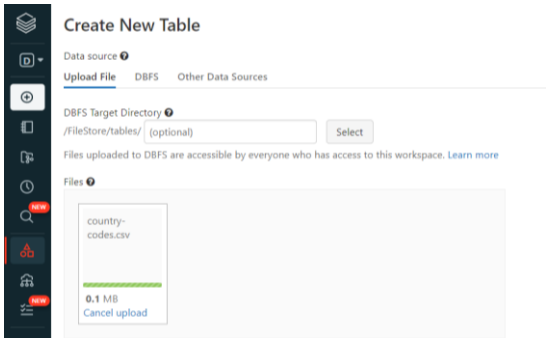
DATE OF SUBMISSION – 15th JUNE 2022 TRAINER NAME – MR. AJAY KUMAR

---

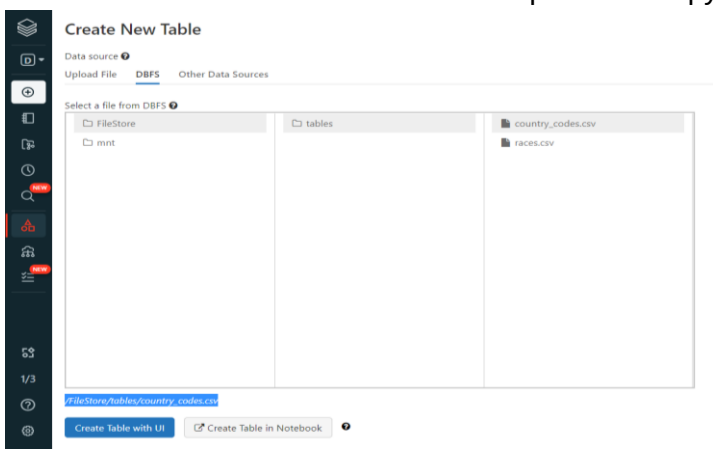
1. Using archive1.zip file - please ingest data into Databricks DBFS path & query the data, redesign columns accordingly using Dataframe commands - display with notebooks accordingly.

File Being used: country.csv

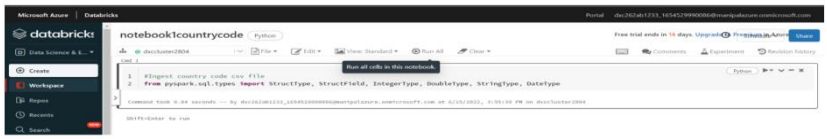
First, login to your Azure Portal and create a Databricks workspace. After, Open the Databricks workspace and create cluster. Now, create a notebook by clicking on the create Notebook option from the side panel. After creating the notebook, upload the data into the Databricks by dragging and dropping the required file.



Now, click on DBFS and select the file that we have uploaded. Copy the file path.



Import the required fields and features from pyspark.



Now, we need to ingest the schema

```python
#include the schema
country_schema = StructType(fields=[StructField("marc",StringType(),True),
                                    StructField("Capital",StringType(),True),
                                    StructField("M49",IntegerType(),True),
                                    StructField("Regioncode",IntegerType(),True),

                                   ])
```

Command took 0.03 seconds -- by dxc262ab1233_1654529990086@manipalazure.onmicrosoft.com at 6/15/2022, 4:04:19 PM on dxccluster2804

Shift+Enter to run

Then we need to create a data frame

Command took 0.03 seconds -- by dxc262ab1233_1654529990086@manipalazure.onmicrosoft.com at 6/15/2022, 4:04:19 PM on dxccluster2804

Cmd 3

```python
#creating a data frame
country_df = spark.read \
.option("header",True) \
.schema(country_schema) \
.csv("/FileStore/tables/country_codes.csv")
```

country_df: pyspark.sql.dataframe.DataFrame
- marc: string
- Capital: string
- M49: integer
- Regioncode: integer

Command took 0.14 seconds -- by dxc262ab1233_1654529990086@manipalazure.onmicrosoft.com at 6/15/2022, 4:09:57 PM on dxccluster2804

Shift+Enter to run

We need to add the ingestion date

Cmd 5

```python
#add ingestion date to the data frame
country_final_df = country_df.withColumn("ingestion_date",current_timestamp())
```

country_final_df: pyspark.sql.dataframe.DataFrame
- marc: string
- Capital: string
- M49: integer
- Regioncode: integer
- ingestion_date: timestamp

Command took 0.04 seconds -- by dxc262ab1233_1654529990086@manipalazure.onmicrosoft.com at 6/15/2022, 4:18:51 PM on dxccluster2804

Shift+Enter to run

Processed container in parquet format

Cmd 7

```python
country_renamed_final_df.write.mode('overwrite').partitionBy('MARC').parquet('/mnt/formulaldl/processed/country')
```

Cancel Running command...

(1) Spark Jobs

To display the data we have been included.

```python
display(spark.read.parquet("/mnt/formulaldl/processed/country"))
```

(5) Spark Jobs
- Job 2 View (Stages: 1/1)
- Job 3 View (Stages: 1/1)
- Job 4 View (Stages: 1/1)
- Job 5 View (Stages: 1/1)
- Job 6 View (Stages: 1/1)

Table Data Profile

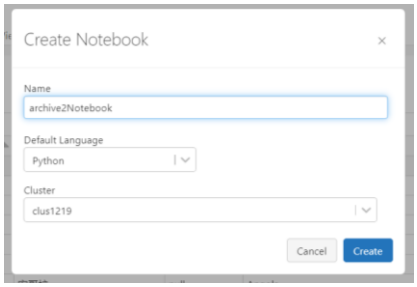
	CAPITAL	m49	Regioncode	INGESTION_DATE	MARC
1	1	null	null	2022-06-15T10:58:27.936+0000	PUR
2	7	null	null	2022-06-15T10:58:27.936+0000	KAZ
3	7	null	null	2022-06-15T10:58:27.936+0000	RUS
4	1	null	null	2022-06-15T10:58:27.936+0000	CAN
5	1	null	null	2022-06-15T10:58:27.936+0000	USA
6	95	null	null	2022-06-15T10:58:27.936+0000	MYA
7	31	null	null	2022-06-15T10:58:27.936+0000	NED

Showing all 250 rows.

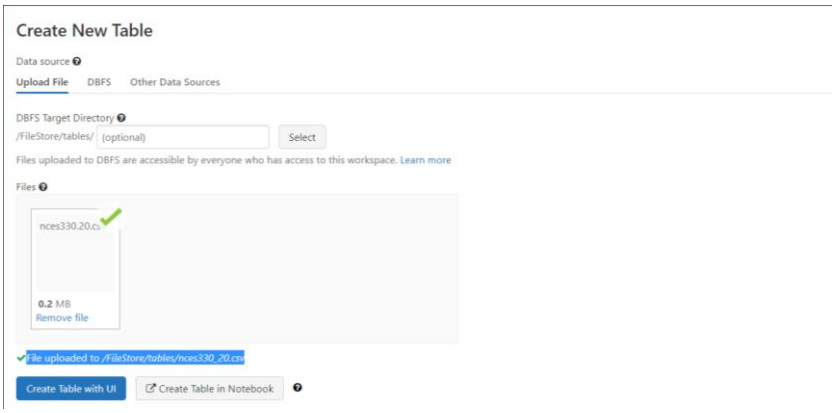
2. Using archive2.zip file - please ingest data into Databricks DBFS path & query the data, redesign columns accordingly using Dataframe commands - display with notebooks accordingly.

File Being Used: nces330.20.csv

Follow steps in 1 and create a new notebook by clicking on the create Notebook option from the side panel.



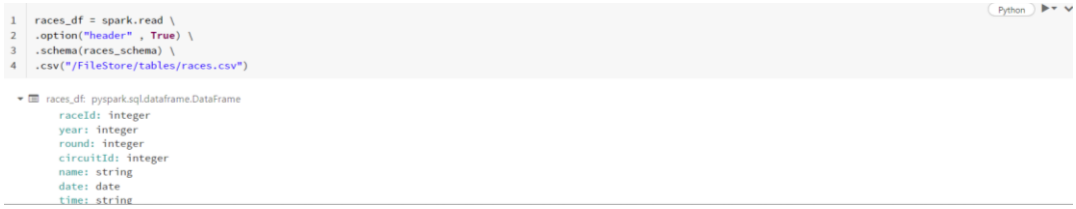
After creating the notebook, upload the data into the Databricks by dragging and dropping the required file.



Import required fields and features from pyspark

Create a schema

Ingest the data



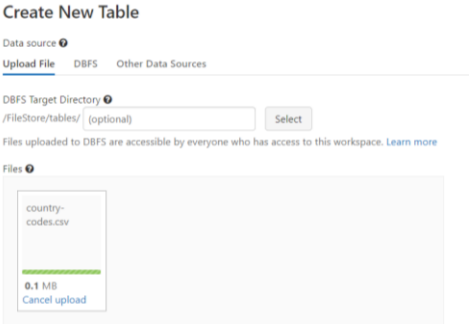
Import col from sql functions

Displays the data we have been included.

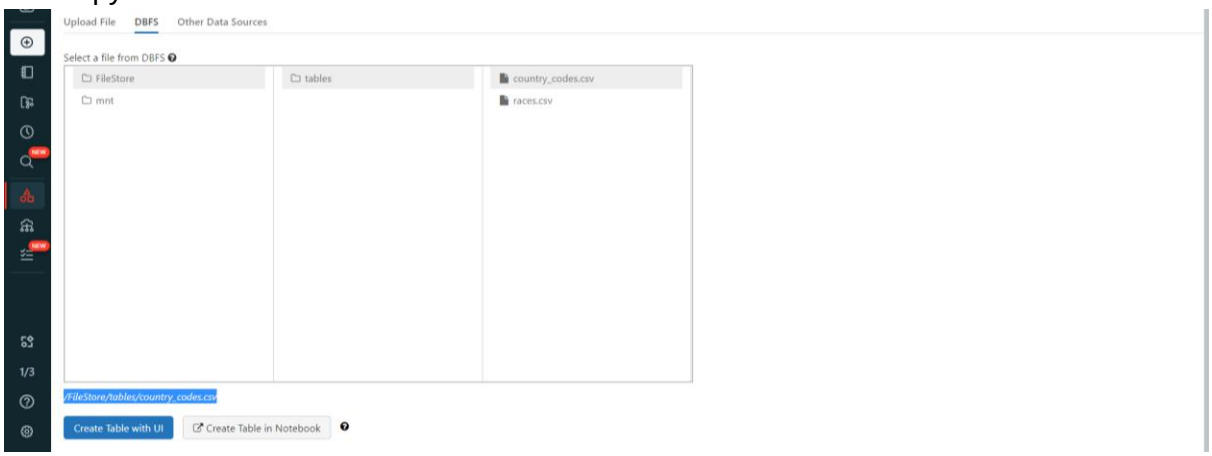
3. Using archive3.zip file - please ingest data into Databricks DBFS path & query the data, redesign columns accordingly using Dataframe commands - display with notebooks accordingly.

File Being used: final_data.csv

create a new notebook by clicking on the create Notebook option from the side panel.After creating the notebook, ingest the data into the Databricks by dragging and dropping the required file in the **drag & drop** region.



Later, click on DBFS and select the file that you have dropped. This will give you the file path and copy that.



Import the required fields and features from pyspark.

```
from pyspark.sql.types import StructType, StructField, IntegerType, StringType
```

Command took 0.84 seconds -- by dxc262ab1225_1654530016505@manipalazure.onmicrosoft.com at 6/15/2022, 5:24:11 PM on clus1219

Cmd 2

```
final_data_schema = StructType(fields=[StructField("tweet_text",StringType(),False),
                                       StructField("emotion_in_tweet_is_directed_at",StringType(),True),
                                       StructField("is_there_an_emotion_directed_at_a_brand_or_product",StringType(),True),
                                        ])
```

Command took 0.05 seconds -- by dxc262ab1225_1654530016505@manipalazure.onmicrosoft.com at 6/15/2022, 5:47:53 PM on clus1219

Cmd 3

```
final_data_df = spark.read \
.option("header" , True) \
.schema(final_data_schema) \
.csv("/FileStore/tables/final_data.csv")
```

final_data_df: pyspark.sql.dataframe.DataFrame
tweet_text: string
emotion_in_tweet_is_directed_at: string
is_there_an_emotion_directed_at_a_brand_or_product: string

Command took 0.19 seconds -- by dxc262ab1225_1654530016505@manipalazure.onmicrosoft.com at 6/15/2022, 5:51:51 PM on clus1219

Cmd 4

Cmd 4

```
from pyspark.sql.functions import col,lit
```

Command took 0.03 seconds -- by dxc262ab1225_1654530016505@manipalazure.onmicrosoft.com at 6/15/2022, 5:34:02 PM on clus1219

Cmd 5

```
final_data_selected_df = final_data_df.select(col('tweet_text'),

col('emotion_in_tweet_is_directed_at').alias('emotion_towards'),col('is_there_an_emotion_directed_at_a_brand_or_product').alias('is_there_a_brand'))
```

final_data_selected_df: pyspark.sql.dataframe.DataFrame
tweet_text: string
emotion_towards: string
is_there_a_brand: string

Command took 0.06 seconds -- by dxc262ab1225_1654530016505@manipalazure.onmicrosoft.com at 6/15/2022, 5:54:25 PM on clus1219

Cmd 6

Display the data.

```
display(final_data_selected_df)
```

(1) Spark Jobs

Table | Data Profile

	tweet_text	emotion_towards	is_there_a_brand
1	.@wesley83 I have a 3G iPhone. After 3 hrs tweeting at #RISE_Austin, it was dead! I need to upgrade. Plugin stations at #SXSW.	iPhone	Negative emotion
2	@jessedee Know about @fludapp ? Awesome iPad/iPhone app that you'll likely appreciate for its design. Also, they're giving free Ts at #SXSW	iPad or iPhone App	Positive emotion
3	@swonderlin Can not wait for #iPad 2 also. They should sale them down at #SXSW.	iPad	Positive emotion
4	@sxsw I hope this year's festival isn't as crashy as this year's iPhone app. #sxsw	iPad or iPhone App	Negative emotion
5	@sxtxstate great stuff on Fri #SXSW: Marissa Mayer (Google), Tim O'Reilly (tech books/conferences) & Matt Mullenweg (Wordpress)	Google	Positive emotion
6	@teachntech00 New iPad Apps For #SpeechTherapy And Communication Are Showcased At The #SXSW Conference http://ht.ly/49n4M #iear #edchat #asd	null	No emotion toward brand or product
7	null	null	No emotion toward brand or product

Truncated results, showing first 1000 rows.
Click to re-execute with maximum result limits.

Command took 0.56 seconds -- by dxc262ab1225_1654530016505@manipalazure.onmicrosoft.com at 6/15/2022, 5:55:03 PM on clus1219

Cmd 7

4. Using archive4.zip file - please ingest data into Databricks DBFS path & query the data, redesign columns accordingly using Dataframe commands - display with notebooks accordingly.

create a notebook by clicking on the create Notebook option from the side panel. After creating the notebook, ingest the data into the Databricks by dragging and dropping the required file.

Create New Table

Data source

Upload File | DBFS | Other Data Sources

DBFS Target Directory

/FileStore/tables/ (optional) Select

Files uploaded to DBFS are accessible by everyone who has access to this workspace. Learn more

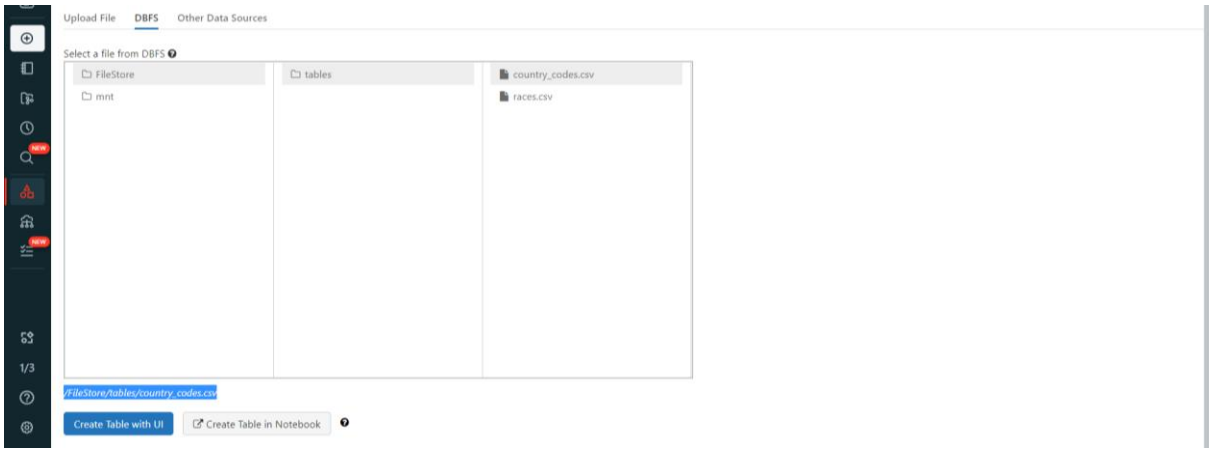
Files

country-codes.csv

0.1 MB

Cancel upload

Later, click on DBFS and select the file that you have dropped. This will give you the file path and copy that.



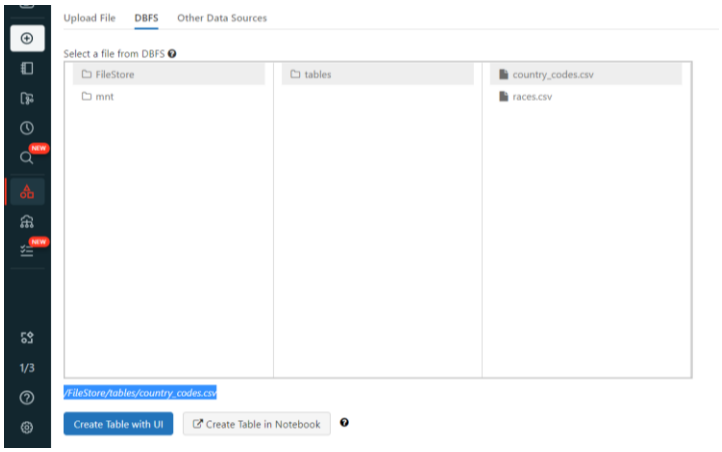
Import the required fields and features from pyspark.

```
from pyspark.sql.types import StructType, StructField, IntegerType, StringType

from pyspark.sql.types import StructType, StructField, IntegerType, StringType

SEntFiN-v1_1_schema = StructType(fields=[StructField("S No.",IntegerType(),False),
                                StructField("Title",StringType(),True),
                                StructField("Decisions",StringType(),True),
                                StructField("Words",IntegerType(),True),

])


SEntFiN-v1_1_df = spark.read \
.option("header" , True) \
.schema(SEntFiN-v1_1_schema) \
.csv("/FileStore/tables/ SEntFiN-v1_1.csv")

from pyspark.sql.functions import col,lit

SEntFiN-v1_1_selected_df = SEntFiN-v1_1_df.select(col('S No'),
                                        col('Title'),col('Words'))

display(SEntFiN-v1_1_df)
```

5:Using archive5.zip file - please ingest data into Databricks DBFS path & query the data, redesign columns accordingly using Dataframe commands - display with notebooks accordingly.

create a notebook by clicking on the create Notebook option from the side panel.
After creating the notebook, ingest the data into the Databricks by dragging and dropping the required file.

Later, click on DBFS and select the file that you have dropped. This will give you the file path and copy that.

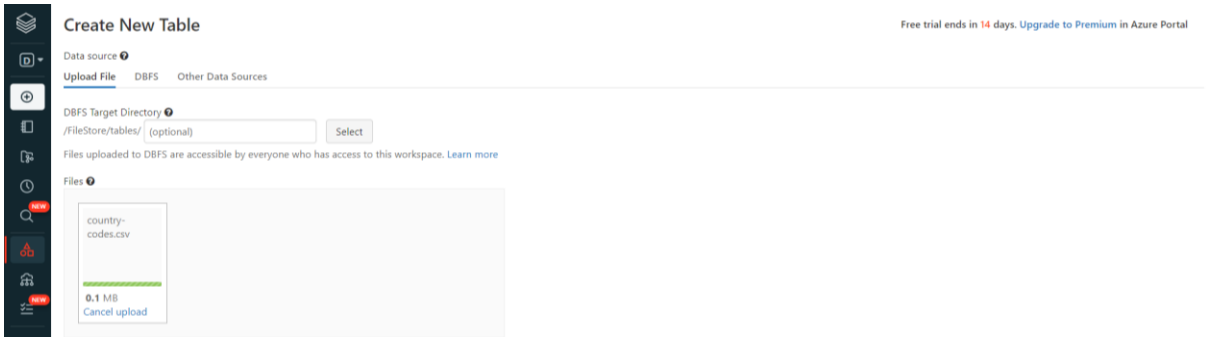


Import the required fields and features from pyspark.

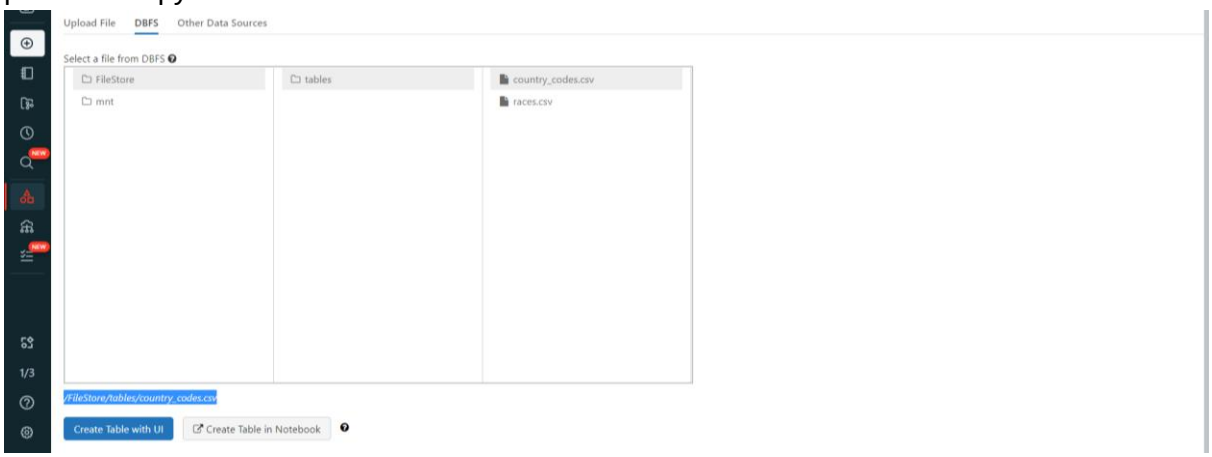
```
from pyspark.sql.types import StructType, StructField, IntegerType, StringType, FloatType

cancer_death_rates_schema = StructType(fields=[StructField("Entity",StringType(),False),
                                StructField("Code",StringType(),True),
                                StructField("Year",IntegerType(),True),
                                StructField("Deaths - Neoplasms - Sex: Both - Age: Age-standardized
(Rate)",FloatType(),True),

])


Cancer_death_rates_df = spark.read \
.option("header" , True) \
.schema(cancer_death_rates_schema) \
.csv("/FileStore/tables/ cancer_death_rates.csv")

from pyspark.sql.functions import col,lit

cancer_death_rates_selected_df = cancer_death_rates_df.select(col(' Entity'),
                                        col(' Year'),col(' Deaths - Neoplasms - Sex: Both - Age: Age-
standardized (Rate)').alias('Deaths'))

display(cancer_death_rates_df)
```

6. Using archive6.zip file - please ingest data into Databricks DBFS path & query the data, redesign columns accordingly using Dataframe commands - display with notebooks accordingly.

create a notebook by clicking on the create Notebook option from the side panel. After creating the notebook, ingest the data into the Databricks by dragging and dropping the required file.



Later, click on DBFS and select the file that you have dropped. This will give you the file path and copy that.



Import the required fields and features from pyspark.

```
from pyspark.sql.types import StructType, StructField, IntegerType, StringType, FloatType

inflation_gdp_schema = StructType(fields=[StructField("Country ",StringType(),False),
                                StructField("Country Code ",StringType(),True),
                                StructField("Year ",IntegerType(),True),
                                StructField("Inflation ",FloatType(),True),

])


inflation_gdp _df = spark.read \
.option("header" , True) \
.schema(inflation_gdp _schema) \
.csv("/FileStore/tables/ inflation_gdp.csv")

from pyspark.sql.functions import col,lit

inflation_gdp _selected_df = inflation_gdp _df.select(col('S No'),
                                        col('Title'),col('Words'))

display(inflation_gdp _df)
```