Data Loading

```
import pyspark

from pyspark.sql import SparkSession

spark = SparkSession.builder.appName("example").getOrCreate()

from google.colab import files
uploaded = files.upload()
```

⥽   [Choose Files] government_finance.csv
- **government_finance.csv**(text/csv) - 511593 bytes, last modified: 12/23/2024 - 100% done
  Saving government_finance.csv to government_finance (1).csv

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

Basic Data Inspection

```
df.show()
```

⥽
```
+---------------+-------+----------+-----------+-------------------+-------------------+-------------------+--------------+----
|Series_reference| Period|Data_value|     STATUS|              Group|            Subject|      Series_title_1|Series_title_2|Serie
+---------------+-------+----------+-----------+-------------------+-------------------+-------------------+--------------+----
|GFSA.SLS01G01Z90|2009.06|       715|      FINAL|Local Government,...|Government Financ...|Net operating bal...|          NULL|
|GFSA.SLS01G01Z90|2010.06|       907|      FINAL|Local Government,...|Government Financ...|Net operating bal...|          NULL|
|GFSA.SLS01G01Z90|2011.06|       350|      FINAL|Local Government,...|Government Financ...|Net operating bal...|          NULL|
|GFSA.SLS01G01Z90|2012.06|       512|      FINAL|Local Government,...|Government Financ...|Net operating bal...|          NULL|
|GFSA.SLS01G01Z90|2013.06|      1033|      FINAL|Local Government,...|Government Financ...|Net operating bal...|          NULL|
|GFSA.SLS01G01Z90|2014.06|       944|      FINAL|Local Government,...|Government Financ...|Net operating bal...|          NULL|
|GFSA.SLS01G01Z90|2015.06|      1033|      FINAL|Local Government,...|Government Financ...|Net operating bal...|          NULL|
|GFSA.SLS01G01Z90|2016.06|      1154|      FINAL|Local Government,...|Government Financ...|Net operating bal...|          NULL|
|GFSA.SLS01G01Z90|2017.06|      1106|      FINAL|Local Government,...|Government Financ...|Net operating bal...|          NULL|
|GFSA.SLS01G01Z90|2018.06|      1554|      FINAL|Local Government,...|Government Financ...|Net operating bal...|          NULL|
|GFSA.SLS01G01Z90|2019.06|      1758|      FINAL|Local Government,...|Government Financ...|Net operating bal...|          NULL|
|GFSA.SLS01G01Z90|2020.06|      1360|      FINAL|Local Government,...|Government Financ...|Net operating bal...|          NULL|
|GFSA.SLS01G01Z90|2021.06|      1790|      FINAL|Local Government,...|Government Financ...|Net operating bal...|          NULL|
|GFSA.SLS01G01Z90|2022.06|      2183|    REVISED|Local Government,...|Government Financ...|Net operating bal...|          NULL|
|GFSA.SLS01G01Z90|2023.06|      1960|    REVISED|Local Government,...|Government Financ...|Net operating bal...|          NULL|
|GFSA.SLS01G01Z90|2024.06|      2010|PROVISIONAL|Local Government,...|Government Financ...|Net operating bal...|          NULL|
|GFSA.SLS01G01Z91|2009.06|     -1092|      FINAL|Local Government,...|Government Financ...|Net lending/borro...|          NULL|
|GFSA.SLS01G01Z91|2010.06|      -924|      FINAL|Local Government,...|Government Financ...|Net lending/borro...|          NULL|
|GFSA.SLS01G01Z91|2011.06|      -849|      FINAL|Local Government,...|Government Financ...|Net lending/borro...|          NULL|
|GFSA.SLS01G01Z91|2012.06|      -545|      FINAL|Local Government,...|Government Financ...|Net lending/borro...|          NULL|
+---------------+-------+----------+-----------+-------------------+-------------------+-------------------+--------------+----
only showing top 20 rows
```

```
df.head(2)
```

⥽ [Row(Series_reference='GFSA.SLS01G01Z90', Period=2009.06, Data_value=715, STATUS='FINAL', Group='Local Government, Operating Statement, Net Balances', Subject='Government Financial Statistics - GFS', Series_title_1='Net operating balance', Series_title_2=None, Series_title_3=None, Series_title_4=None, Series_title_5=None, UNITS='Dollars', MAGNTUDE=6),
Row(Series_reference='GFSA.SLS01G01Z90', Period=2010.06, Data_value=907, STATUS='FINAL', Group='Local Government, Operating Statement, Net Balances', Subject='Government Financial Statistics - GFS', Series_title_1='Net operating balance', Series_title_2=None, Series_title_3=None, Series_title_4=None, Series_title_5=None, UNITS='Dollars', MAGNTUDE=6)]

```
df.describe()
```

⥽ DataFrame[summary: string, Series_reference: string, Period: string, Data_value: string, STATUS: string, Group: string, Subject: string, Series_title_1: string, Series_title_2: string, Series_title_3: string, Series_title_4: string, Series_title_5: string, UNITS: string, MAGNTUDE: string]

```
df.printSchema()
```

⥽
```
root
 |-- Series_reference: string (nullable = true)
 |-- Period: double (nullable = true)
 |-- Data_value: integer (nullable = true)
 |-- STATUS: string (nullable = true)
 |-- Group: string (nullable = true)
 |-- Subject: string (nullable = true)
 |-- Series_title_1: string (nullable = true)
 |-- Series_title_2: string (nullable = true)
 |-- Series_title_3: string (nullable = true)
 |-- Series_title_4: string (nullable = true)
```

```
     |-- Series_title_5: string (nullable = true)
     |-- UNITS: string (nullable = true)
     |-- MAGNTUDE: integer (nullable = true)
```

`df.show(4)`

```
+---------------+-------+----------+------+------------------+------------------+------------------+-------------+---------
|Series_reference| Period|Data_value|STATUS|             Group|           Subject|    Series_title_1|Series_title_2|Series_tit
+---------------+-------+----------+------+------------------+------------------+------------------+-------------+---------
|GFSA.SLS01G01Z90|2009.06|       715| FINAL|Local Government,...|Government Financ...|Net operating bal...|         NULL|
|GFSA.SLS01G01Z90|2010.06|       907| FINAL|Local Government,...|Government Financ...|Net operating bal...|         NULL|
|GFSA.SLS01G01Z90|2011.06|       350| FINAL|Local Government,...|Government Financ...|Net operating bal...|         NULL|
|GFSA.SLS01G01Z90|2012.06|       512| FINAL|Local Government,...|Government Financ...|Net operating bal...|         NULL|
+---------------+-------+----------+------+------------------+------------------+------------------+-------------+---------
only showing top 4 rows
```

`df.count()`

```
2664
```

Data Cleaning

`df.na.drop()`

```
DataFrame[Series_reference: string, Period: double, Data_value: int, STATUS: string, Group: string, Subject: string,
Series_title_1: string, Series_title_2: string, Series_title_3: string, Series_title_4: string, Series_title_5: string, UNITS:
string, MAGNTUDE: int]
```

`df.na.fill('NA')`

```
DataFrame[Series_reference: string, Period: double, Data_value: int, STATUS: string, Group: string, Subject: string,
Series_title_1: string, Series_title_2: string, Series_title_3: string, Series_title_4: string, Series_title_5: string, UNITS:
string, MAGNTUDE: int]
```

`df.show(4)`

```
+---------------+-------+----------+------+------------------+------------------+------------------+-------------+---------
|Series_reference| Period|Data_value|STATUS|             Group|           Subject|    Series_title_1|Series_title_2|Series_tit
+---------------+-------+----------+------+------------------+------------------+------------------+-------------+---------
|GFSA.SLS01G01Z90|2009.06|       715| FINAL|Local Government,...|Government Financ...|Net operating bal...|         NULL|
|GFSA.SLS01G01Z90|2010.06|       907| FINAL|Local Government,...|Government Financ...|Net operating bal...|         NULL|
|GFSA.SLS01G01Z90|2011.06|       350| FINAL|Local Government,...|Government Financ...|Net operating bal...|         NULL|
|GFSA.SLS01G01Z90|2012.06|       512| FINAL|Local Government,...|Government Financ...|Net operating bal...|         NULL|
+---------------+-------+----------+------+------------------+------------------+------------------+-------------+---------
only showing top 4 rows
```

`df.drop('Series_title_5','Series_title_4')`

```
DataFrame[Series_reference: string, Period: double, Data_value: int, STATUS: string, Group: string, Subject: string,
Series_title_1: string, Series_title_2: string, Series_title_3: string, UNITS: string, MAGNTUDE: int]
```

`df.show()`

```
+---------------+-------+----------+-----------+------------------+------------------+------------------+-------------+----
|Series_reference| Period|Data_value|     STATUS|             Group|           Subject|    Series_title_1|Series_title_2|Serie
+---------------+-------+----------+-----------+------------------+------------------+------------------+-------------+----
|GFSA.SLS01G01Z90|2009.06|       715|      FINAL|Local Government,...|Government Financ...|Net operating bal...|         NULL|
|GFSA.SLS01G01Z90|2010.06|       907|      FINAL|Local Government,...|Government Financ...|Net operating bal...|         NULL|
|GFSA.SLS01G01Z90|2011.06|       350|      FINAL|Local Government,...|Government Financ...|Net operating bal...|         NULL|
|GFSA.SLS01G01Z90|2012.06|       512|      FINAL|Local Government,...|Government Financ...|Net operating bal...|         NULL|
|GFSA.SLS01G01Z90|2013.06|      1033|      FINAL|Local Government,...|Government Financ...|Net operating bal...|         NULL|
|GFSA.SLS01G01Z90|2014.06|       944|      FINAL|Local Government,...|Government Financ...|Net operating bal...|         NULL|
|GFSA.SLS01G01Z90|2015.06|      1033|      FINAL|Local Government,...|Government Financ...|Net operating bal...|         NULL|
|GFSA.SLS01G01Z90|2016.06|      1154|      FINAL|Local Government,...|Government Financ...|Net operating bal...|         NULL|
|GFSA.SLS01G01Z90|2017.06|      1106|      FINAL|Local Government,...|Government Financ...|Net operating bal...|         NULL|
|GFSA.SLS01G01Z90|2018.06|      1554|      FINAL|Local Government,...|Government Financ...|Net operating bal...|         NULL|
|GFSA.SLS01G01Z90|2019.06|      1758|      FINAL|Local Government,...|Government Financ...|Net operating bal...|         NULL|
|GFSA.SLS01G01Z90|2020.06|      1360|      FINAL|Local Government,...|Government Financ...|Net operating bal...|         NULL|
|GFSA.SLS01G01Z90|2021.06|      1790|      FINAL|Local Government,...|Government Financ...|Net operating bal...|         NULL|
|GFSA.SLS01G01Z90|2022.06|      2183|    REVISED|Local Government,...|Government Financ...|Net operating bal...|         NULL|
|GFSA.SLS01G01Z90|2023.06|      1960|    REVISED|Local Government,...|Government Financ...|Net operating bal...|         NULL|
|GFSA.SLS01G01Z90|2024.06|      2010|PROVISIONAL|Local Government,...|Government Financ...|Net operating bal...|         NULL|
|GFSA.SLS01G01Z91|2009.06|     -1092|      FINAL|Local Government,...|Government Financ...|Net lending/borro...|         NULL|
|GFSA.SLS01G01Z91|2010.06|      -924|      FINAL|Local Government,...|Government Financ...|Net lending/borro...|         NULL|
|GFSA.SLS01G01Z91|2011.06|      -849|      FINAL|Local Government,...|Government Financ...|Net lending/borro...|         NULL|
|GFSA.SLS01G01Z91|2012.06|      -545|      FINAL|Local Government,...|Government Financ...|Net lending/borro...|         NULL|
+---------------+-------+----------+-----------+------------------+------------------+------------------+-------------+----
```

only showing top 20 rows

```
df.drop('Series_title_4').show()
```

```
+---------------+-------+----------+-----------+------------------+------------------+------------------+-------------+----
|Series_reference| Period|Data_value|     STATUS|             Group|           Subject|     Series_title_1|Series_title_2|Serie
+---------------+-------+----------+-----------+------------------+------------------+------------------+-------------+----
|GFSA.SLS01G01Z90|2009.06|       715|      FINAL|Local Government,...|Government Financ...|Net operating bal...|         NULL|
|GFSA.SLS01G01Z90|2010.06|       907|      FINAL|Local Government,...|Government Financ...|Net operating bal...|         NULL|
|GFSA.SLS01G01Z90|2011.06|       350|      FINAL|Local Government,...|Government Financ...|Net operating bal...|         NULL|
|GFSA.SLS01G01Z90|2012.06|       512|      FINAL|Local Government,...|Government Financ...|Net operating bal...|         NULL|
|GFSA.SLS01G01Z90|2013.06|      1033|      FINAL|Local Government,...|Government Financ...|Net operating bal...|         NULL|
|GFSA.SLS01G01Z90|2014.06|       944|      FINAL|Local Government,...|Government Financ...|Net operating bal...|         NULL|
|GFSA.SLS01G01Z90|2015.06|      1033|      FINAL|Local Government,...|Government Financ...|Net operating bal...|         NULL|
|GFSA.SLS01G01Z90|2016.06|      1154|      FINAL|Local Government,...|Government Financ...|Net operating bal...|         NULL|
|GFSA.SLS01G01Z90|2017.06|      1106|      FINAL|Local Government,...|Government Financ...|Net operating bal...|         NULL|
|GFSA.SLS01G01Z90|2018.06|      1554|      FINAL|Local Government,...|Government Financ...|Net operating bal...|         NULL|
|GFSA.SLS01G01Z90|2019.06|      1758|      FINAL|Local Government,...|Government Financ...|Net operating bal...|         NULL|
|GFSA.SLS01G01Z90|2020.06|      1360|      FINAL|Local Government,...|Government Financ...|Net operating bal...|         NULL|
|GFSA.SLS01G01Z90|2021.06|      1790|      FINAL|Local Government,...|Government Financ...|Net operating bal...|         NULL|
|GFSA.SLS01G01Z90|2022.06|      2183|    REVISED|Local Government,...|Government Financ...|Net operating bal...|         NULL|
|GFSA.SLS01G01Z90|2023.06|      1960|    REVISED|Local Government,...|Government Financ...|Net operating bal...|         NULL|
|GFSA.SLS01G01Z90|2024.06|      2010|PROVISIONAL|Local Government,...|Government Financ...|Net operating bal...|         NULL|
|GFSA.SLS01G01Z91|2009.06|     -1092|      FINAL|Local Government,...|Government Financ...|Net lending/borro...|         NULL|
|GFSA.SLS01G01Z91|2010.06|      -924|      FINAL|Local Government,...|Government Financ...|Net lending/borro...|         NULL|
|GFSA.SLS01G01Z91|2011.06|      -849|      FINAL|Local Government,...|Government Financ...|Net lending/borro...|         NULL|
|GFSA.SLS01G01Z91|2012.06|      -545|      FINAL|Local Government,...|Government Financ...|Net lending/borro...|         NULL|
+---------------+-------+----------+-----------+------------------+------------------+------------------+-------------+----
only showing top 20 rows
```

```
df1 = df.drop('Series_title_2')
```

```
df.count()
```

2664

```
df.head(3)
```

```
[Row(Series_reference='GFSA.SLS01G01Z90', Period=2009.06, Data_value=715, STATUS='FINAL', Group='Local Government, Operating
Statement, Net Balances', Subject='Government Financial Statistics - GFS', Series_title_1='Net operating balance',
Series_title_2=None, Series_title_3=None, Series_title_4=None, Series_title_5=None, UNITS='Dollars', MAGNTUDE=6),
 Row(Series_reference='GFSA.SLS01G01Z90', Period=2010.06, Data_value=907, STATUS='FINAL', Group='Local Government, Operating
Statement, Net Balances', Subject='Government Financial Statistics - GFS', Series_title_1='Net operating balance',
Series_title_2=None, Series_title_3=None, Series_title_4=None, Series_title_5=None, UNITS='Dollars', MAGNTUDE=6),
 Row(Series_reference='GFSA.SLS01G01Z90', Period=2011.06, Data_value=350, STATUS='FINAL', Group='Local Government, Operating
Statement, Net Balances', Subject='Government Financial Statistics - GFS', Series_title_1='Net operating balance',
Series_title_2=None, Series_title_3=None, Series_title_4=None, Series_title_5=None, UNITS='Dollars', MAGNTUDE=6)]
```

```
df1.show(2)
```

```
+---------------+-------+----------+------+------------------+------------------+------------------+-------------+--------
|Series_reference| Period|Data_value|STATUS|             Group|           Subject|     Series_title_1|Series_title_3|Series_tit
+---------------+-------+----------+------+------------------+------------------+------------------+-------------+--------
|GFSA.SLS01G01Z90|2009.06|       715| FINAL|Local Government,...|Government Financ...|Net operating bal...|         NULL|
|GFSA.SLS01G01Z90|2010.06|       907| FINAL|Local Government,...|Government Financ...|Net operating bal...|         NULL|
+---------------+-------+----------+------+------------------+------------------+------------------+-------------+--------
only showing top 2 rows
```

```
df1 = df.drop('Series_title_5')
```

```
df1.show(2)
```

```
+---------------+-------+----------+------+------------------+------------------+------------------+-------------+--------
|Series_reference| Period|Data_value|STATUS|             Group|           Subject|     Series_title_1|Series_title_2|Series_tit
+---------------+-------+----------+------+------------------+------------------+------------------+-------------+--------
|GFSA.SLS01G01Z90|2009.06|       715| FINAL|Local Government,...|Government Financ...|Net operating bal...|         NULL|
|GFSA.SLS01G01Z90|2010.06|       907| FINAL|Local Government,...|Government Financ...|Net operating bal...|         NULL|
+---------------+-------+----------+------+------------------+------------------+------------------+-------------+--------
only showing top 2 rows
```

```
df1 = df.drop('Series_title_2')
```

```
df1.show()
```

```
+---------------+-------+----------+----------+-------------------+-------------------+-------------------+---------------+--------------+----
|Series_reference| Period|Data_value|    STATUS|              Group|            Subject|     Series_title_1|Series_title_3|Serie
+---------------+-------+----------+----------+-------------------+-------------------+-------------------+---------------+--------------+----
|GFSA.SLS01G01Z90|2009.06|       715|     FINAL|Local Government,...|Government Financ...|Net operating bal...|          NULL|
|GFSA.SLS01G01Z90|2010.06|       907|     FINAL|Local Government,...|Government Financ...|Net operating bal...|          NULL|
|GFSA.SLS01G01Z90|2011.06|       350|     FINAL|Local Government,...|Government Financ...|Net operating bal...|          NULL|
|GFSA.SLS01G01Z90|2012.06|       512|     FINAL|Local Government,...|Government Financ...|Net operating bal...|          NULL|
|GFSA.SLS01G01Z90|2013.06|      1033|     FINAL|Local Government,...|Government Financ...|Net operating bal...|          NULL|
|GFSA.SLS01G01Z90|2014.06|       944|     FINAL|Local Government,...|Government Financ...|Net operating bal...|          NULL|
|GFSA.SLS01G01Z90|2015.06|      1033|     FINAL|Local Government,...|Government Financ...|Net operating bal...|          NULL|
|GFSA.SLS01G01Z90|2016.06|      1154|     FINAL|Local Government,...|Government Financ...|Net operating bal...|          NULL|
|GFSA.SLS01G01Z90|2017.06|      1106|     FINAL|Local GovernmenL,...|GovernmeN Financ...|Net operating bal...|          NULL|
|GFSA.SLS01G01Z90|2018.06|      1554|     FINAL|Local Government,...|Government Financ...|Net operating bal...|          NULL|
|GFSA.SLS01G01Z90|2019.06|      1758|     FINAL|Local Government,...|Government Financ...|Net operating bal...|          NULL|
|GFSA.SLS01G01Z90|2020.06|      1360|     FINAL|Local Government,...|Government Financ...|Net operating bal...|          NULL|
|GFSA.SLS01G01Z90|2021.06|      1790|     FINAL|Local Government,...|Government Financ...|Net operating bal...|          NULL|
|GFSA.SLS01G01Z90|2022.06|      2183|   REVISED|Local Government,...|Government Financ...|Net operating bal...|          NULL|
|GFSA.SLS01G01Z90|2023.06|      1960|   REVISED|Local Government,...|Government Financ...|Net operating bal...|          NULL|
|GFSA.SLS01G01Z90|2024.06|      2010|PROVISIONAL|Local Government,...|Government Financ...|Net operating bal...|          NULL|
|GFSA.SLS01G01Z91|2009.06|     -1092|     FINAL|Local Government,...|Government Financ...|Net lending/borro...|          NULL|
|GFSA.SLS01G01Z91|2010.06|      -924|     FINAL|Local Government,...|Government Financ...|Net lending/borro...|          NULL|
|GFSA.SLS01G01Z91|2011.06|      -849|     FINAL|Local Government,...|Government Financ...|Net lending/borro...|          NULL|
|GFSA.SLS01G01Z91|2012.06|      -545|     FINAL|Local Government,...|Government Financ...|Net lending/borro...|          NULL|
+---------------+-------+----------+----------+-------------------+-------------------+-------------------+---------------+--------------+----
only showing top 20 rows
```

Data Transformation

```python
df.select('Data_value', 'STATUS', 'Group').show()
```

```
+----------+----------+--------------------+
|Data_value|    STATUS|               Group|
+----------+----------+--------------------+
|       715|     FINAL|Local Government,...|
|       907|     FINAL|Local Government,...|
|       350|     FINAL|Local Government,...|
|       512|     FINAL|Local Government,...|
|      1033|     FINAL|Local Government,...|
|       944|     FINAL|Local Government,...|
|      1033|     FINAL|Local Government,...|
|      1154|     FINAL|Local Government,...|
|      1106|     FINAL|Local Government,...|
|      1554|     FINAL|Local Government,...|
|      1758|     FINAL|Local Government,...|
|      1360|     FINAL|Local Government,...|
|      1790|     FINAL|Local Government,...|
|      2183|   REVISED|Local Government,...|
|      1960|   REVISED|Local Government,...|
|      2010|PROVISIONAL|Local Government,...|
|     -1092|     FINAL|Local Government,...|
|      -924|     FINAL|Local Government,...|
|      -849|     FINAL|Local Government,...|
|      -545|     FINAL|Local Government,...|
+----------+----------+--------------------+
only showing top 20 rows
```

```python
df.withColumn('New_Group', df['Group']).show()
```

```
--+-------+----------+----------+-------------------+-------------------+-------------------+---------------+--------------+-----
ce| Period|Data_value|    STATUS|              Group|            Subject|     Series_title_1|Series_title_2|Series_title_3|Serie
--+-------+----------+----------+-------------------+-------------------+-------------------+---------------+--------------+-----
90|2009.06|       715|     FINAL|Local Government,...|Government Financ...|Net operating bal...|          NULL|          NULL|
90|2010.06|       907|     FINAL|Local Government,...|Government Financ...|Net operating bal...|          NULL|          NULL|
90|2011.06|       350|     FINAL|Local Government,...|Government Financ...|Net operating bal...|          NULL|          NULL|
90|2012.06|       512|     FINAL|Local Government,...|Government Financ...|Net operating bal...|          NULL|          NULL|
90|2013.06|      1033|     FINAL|Local Government,...|Government Financ...|Net operating bal...|          NULL|          NULL|
90|2014.06|       944|     FINAL|Local Government,...|Government Financ...|Net operating bal...|          NULL|          NULL|
90|2015.06|      1033|     FINAL|Local Government,...|Government Financ...|Net operating bal...|          NULL|          NULL|
90|2016.06|      1154|     FINAL|Local Government,...|Government Financ...|Net operating bal...|          NULL|          NULL|
90|2017.06|      1106|     FINAL|Local Government,...|Government Financ...|Net operating bal...|          NULL|          NULL|
90|2018.06|      1554|     FINAL|Local Government,...|Government Financ...|Net operating bal...|          NULL|          NULL|
90|2019.06|      1758|     FINAL|Local Government,...|Government Financ...|Net operating bal...|          NULL|          NULL|
90|2020.06|      1360|     FINAL|Local Government,...|Government Financ...|Net operating bal...|          NULL|          NULL|
90|2021.06|      1790|     FINAL|Local Government,...|Government Financ...|Net operating bal...|          NULL|          NULL|
90|2022.06|      2183|   REVISED|Local Government,...|Government Financ...|Net operating bal...|          NULL|          NULL|
90|2023.06|      1960|   REVISED|Local Government,...|Government Financ...|Net operating bal...|          NULL|          NULL|
90|2024.06|      2010|PROVISIONAL|Local Government,...|Government Financ...|Net operating bal...|          NULL|          NULL|
91|2009.06|     -1092|     FINAL|Local Government,...|Government Financ...|Net lending/borro...|          NULL|          NULL|
91|2010.06|      -924|     FINAL|Local Government,...|Government Financ...|Net lending/borro...|          NULL|          NULL|
91|2011.06|      -849|     FINAL|Local Government,...|Government Financ...|Net lending/borro...|          NULL|          NULL|
91|2012.06|      -545|     FINAL|Local Government,...|Government Financ...|Net lending/borro...|          NULL|          NULL|
--+-------+----------+----------+-------------------+-------------------+-------------------+---------------+--------------+-----
p 20 rows
```

```
df.withColumn('Status', df['STATUS']).show()
```

```
+----------------+-------+----------+-----------+------------------+------------------+------------------+-------------+----
|Series_reference| Period|Data_value|     Status|             Group|           Subject|    Series_title_1|Series_title_2|Serie
+----------------+-------+----------+-----------+------------------+------------------+------------------+-------------+----
|GFSA.SLS01G01Z90|2009.06|       715|      FINAL|Local Government,...|Government Financ...|Net operating bal...|         NULL|
|GFSA.SLS01G01Z90|2010.06|       907|      FINAL|Local Government,...|Government Financ...|Net operating bal...|         NULL|
|GFSA.SLS01G01Z90|2011.06|       350|      FINAL|Local Government,...|Government Financ...|Net operating bal...|         NULL|
|GFSA.SLS01G01Z90|2012.06|       512|      FINAL|Local Government,...|Government Financ...|Net operating bal...|         NULL|
|GFSA.SLS01G01Z90|2013.06|      1033|      FINAL|Local Government,...|Government Financ...|Net operating bal...|         NULL|
|GFSA.SLS01G01Z90|2014.06|       944|      FINAL|Local Government,...|Government Financ...|Net operating bal...|         NULL|
|GFSA.SLS01G01Z90|2015.06|      1033|      FINAL|Local Government,...|Government Financ...|Net operating bal...|         NULL|
|GFSA.SLS01G01Z90|2016.06|      1154|      FINAL|Local Government,...|Government Financ...|Net operating bal...|         NULL|
|GFSA.SLS01G01Z90|2017.06|      1106|      FINAL|Local Government,...|Government Financ...|Net operating bal...|         NULL|
|GFSA.SLS01G01Z90|2018.06|      1554|      FINAL|Local Government,...|Government Financ...|Net operating bal...|         NULL|
|GFSA.SLS01G01Z90|2019.06|      1758|      FINAL|Local Government,...|Government Financ...|Net operating bal...|         NULL|
|GFSA.SLS01G01Z90|2020.06|      1360|      FINAL|Local Government,...|Government Financ...|Net operating bal...|         NULL|
|GFSA.SLS01G01Z90|2021.06|      1790|      FINAL|Local Government,...|Government Financ...|Net operating bal...|         NULL|
|GFSA.SLS01G01Z90|2022.06|      2183|    REVISED|Local Government,...|Government Financ...|Net operating bal...|         NULL|
|GFSA.SLS01G01Z90|2023.06|      1960|    REVISED|Local Government,...|Government Financ...|Net operating bal...|         NULL|
|GFSA.SLS01G01Z90|2024.06|      2010|PROVISIONAL|Local Government,...|Government Financ...|Net operating bal...|         NULL|
|GFSA.SLS01G01Z91|2009.06|     -1092|      FINAL|Local Government,...|Government Financ...|Net lending/borro...|         NULL|
|GFSA.SLS01G01Z91|2010.06|      -924|      FINAL|Local Government,...|Government Financ...|Net lending/borro...|         NULL|
|GFSA.SLS01G01Z91|2011.06|      -849|      FINAL|Local Government,...|Government Financ...|Net lending/borro...|         NULL|
|GFSA.SLS01G01Z91|2012.06|      -545|      FINAL|Local Government,...|Government Financ...|Net lending/borro...|         NULL|
+----------------+-------+----------+-----------+------------------+------------------+------------------+-------------+----
only showing top 20 rows
```

```
df.filter(df['Status'] == 'FINAL').show()
```

```
+----------------+-------+----------+------+------------------+------------------+------------------+-------------+---------
|Series_reference| Period|Data_value|STATUS|             Group|           Subject|    Series_title_1|Series_title_2|Series_tit
+----------------+-------+----------+------+------------------+------------------+------------------+-------------+---------
|GFSA.SLS01G01Z90|2009.06|       715| FINAL|Local Government,...|Government Financ...|Net operating bal...|         NULL|
|GFSA.SLS01G01Z90|2010.06|       907| FINAL|Local Government,...|Government Financ...|Net operating bal...|         NULL|
|GFSA.SLS01G01Z90|2011.06|       350| FINAL|Local Government,...|Government Financ...|Net operating bal...|         NULL|
|GFSA.SLS01G01Z90|2012.06|       512| FINAL|Local Government,...|Government Financ...|Net operating bal...|         NULL|
|GFSA.SLS01G01Z90|2013.06|      1033| FINAL|Local Government,...|Government Financ...|Net operating bal...|         NULL|
|GFSA.SLS01G01Z90|2014.06|       944| FINAL|Local Government,...|Government Financ...|Net operating bal...|         NULL|
|GFSA.SLS01G01Z90|2015.06|      1033| FINAL|Local Government,...|Government Financ...|Net operating bal...|         NULL|
|GFSA.SLS01G01Z90|2016.06|      1154| FINAL|Local Government,...|Government Financ...|Net operating bal...|         NULL|
|GFSA.SLS01G01Z90|2017.06|      1106| FINAL|Local Government,...|Government Financ...|Net operating bal...|         NULL|
|GFSA.SLS01G01Z90|2018.06|      1554| FINAL|Local Government,...|Government Financ...|Net operating bal...|         NULL|
|GFSA.SLS01G01Z90|2019.06|      1758| FINAL|Local Government,...|Government Financ...|Net operating bal...|         NULL|
|GFSA.SLS01G01Z90|2020.06|      1360| FINAL|Local Government,...|Government Financ...|Net operating bal...|         NULL|
|GFSA.SLS01G01Z90|2021.06|      1790| FINAL|Local Government,...|Government Financ...|Net operating bal...|         NULL|
|GFSA.SLS01G01Z91|2009.06|     -1092| FINAL|Local Government,...|Government Financ...|Net lending/borro...|         NULL|
|GFSA.SLS01G01Z91|2010.06|      -924| FINAL|Local Government,...|Government Financ...|Net lending/borro...|         NULL|
|GFSA.SLS01G01Z91|2011.06|      -849| FINAL|Local Government,...|Government Financ...|Net lending/borro...|         NULL|
|GFSA.SLS01G01Z91|2012.06|      -545| FINAL|Local Government,...|Government Financ...|Net lending/borro...|         NULL|
|GFSA.SLS01G01Z91|2013.06|     -1196| FINAL|Local Government,...|Government Financ...|Net lending/borro...|         NULL|
|GFSA.SLS01G01Z91|2014.06|      -714| FINAL|Local Government,...|Government Financ...|Net lending/borro...|         NULL|
|GFSA.SLS01G01Z91|2015.06|      -778| FINAL|Local Government,...|Government Financ...|Net lending/borro...|         NULL|
+----------------+-------+----------+------+------------------+------------------+------------------+-------------+---------
only showing top 20 rows
```

```
df.filter(df['Status'] =='PROVISIONAL').show()
```

```
+----------------+-------+----------+-----------+------------------+------------------+------------------+-------------+----
|Series_reference| Period|Data_value|     STATUS|             Group|           Subject|    Series_title_1|Series_title_2|Serie
+----------------+-------+----------+-----------+------------------+------------------+------------------+-------------+----
|GFSA.SLS01G01Z90|2024.06|      2010|PROVISIONAL|Local Government,...|Government Financ...|Net operating bal...|         NULL|
|GFSA.SLS01G01Z91|2024.06|     -2667|PROVISIONAL|Local Government,...|Government Financ...|Net lending/borro...|         NULL|
|GFSA.SLS01G02121|2024.06|         0|PROVISIONAL|Local Government,...|Government Financ...|Social security c...|         NULL|
|GFSA.SLS01G02140|2024.06|         0|PROVISIONAL|Local Government,...|Government Financ...|      Other income|         NULL|
|GFSA.SLS01G02142|2024.06|      2890|PROVISIONAL|Local Government,...|Government Financ...|Sales of goods an...|         NULL|
|GFSA.SLS01G02143|2024.06|       324|PROVISIONAL|Local Government,...|Government Financ...|   Dividend income|         NULL|
|GFSA.SLS01G02144|2024.06|       482|PROVISIONAL|Local Government,...|Government Financ...|   Interest income|         NULL|
|GFSA.SLS01G02981|2024.06|      1960|PROVISIONAL|Local Government,...|Government Financ...|    Current grants|         NULL|
|GFSA.SLS01G02982|2024.06|       188|PROVISIONAL|Local Government,...|Government Financ...|Other current tra...|         NULL|
|GFSA.SLS01G02991|2024.06|      2082|PROVISIONAL|Local Government,...|Government Financ...|    Capital grants|         NULL|
|GFSA.SLS01G02992|2024.06|      1596|PROVISIONAL|Local Government,...|Government Financ...|Other capital tra...|         NULL|
|GFSA.SLS01G02998|2024.06|      9576|PROVISIONAL|Local Government,...|Government Financ...|  Taxation revenue|         NULL|
|GFSA.SLS01G02Z99|2024.06|     19097|PROVISIONAL|Local Government,...|Government Financ...|             Total|         NULL|
|GFSA.SLS01G03210|2024.06|      3641|PROVISIONAL|Local Government,...|Government Financ...| Employee expenses|         NULL|
|GFSA.SLS01G03220|2024.06|      6540|PROVISIONAL|Local Government,...|Government Financ...|Purchases of good...|         NULL|
|GFSA.SLS01G03240|2024.06|      1408|PROVISIONAL|Local Government,...|Government Financ...|  Interest expense|         NULL|
|GFSA.SLS01G03270|2024.06|         0|PROVISIONAL|Local Government,...|Government Financ...|    Social benefits|         NULL|
|GFSA.SLS01G03280|2024.06|         9|PROVISIONAL|Local Government,...|Government Financ...|    Other expenses|         NULL|
|GFSA.SLS01G03981|2024.06|         0|PROVISIONAL|Local Government,...|Government Financ...|    Current grants|         NULL|
```

```
|GFSA.SLS01G03982|2024.06|         802|PROVISIONAL|Local Government,...|Government Financ...|Other current tra...|          NULL|----
+---------------+-------+----------+-----------+--------------------+--------------------+--------------------+--------------+----
only showing top 20 rows
```

```
df.groupby('STATUS').count().show()
```

```
+-----------+-----+
|     STATUS|count|
+-----------+-----+
|    REVISED|  195|
|PROVISIONAL|  166|
|      FINAL| 2303|
+-----------+-----+
```

```
df.groupby(df['Series_title_2']).count().show()
```

```
+--------------------+-----+
|      Series_title_2|count|
+--------------------+-----+
|               Total|  128|
|           Education|   16|
|Customs and other...|   16|
|Recurrent taxes o...|   16|
|Equity accounted ...|   16|
|             Defence|   16|
|Public order and ...|   16|
|     Taxes on exports|   16|
|Insurance technic...|   16|
|                NULL| 1120|
|   Social protection|   16|
|Other taxes on in...|   16|
|               Loans|   16|
|Other non-recurre...|   16|
|              Health|   16|
|Environmental pro...|   16|
|             Excises|   16|
|     Economic affairs|   16|
|Taxes on use of g...|   16|
|         Unallocable|   16|
+--------------------+-----+
only showing top 20 rows
```

```
df.sort(df['Period'].desc()).show(5)
```

```
+---------------+-------+----------+-----------+--------------------+--------------------+--------------------+--------------+----
|Series_reference| Period|Data_value|     STATUS|               Group|             Subject|      Series_title_1|Series_title_2|Serie
+---------------+-------+----------+-----------+--------------------+--------------------+--------------------+--------------+----
|GFSA.SLS01G02143|2024.06|       324|PROVISIONAL|Local Government,...|Government Financ...|     Dividend income|          NULL|
|GFSA.SLS01G02992|2024.06|      1596|PROVISIONAL|Local Government,...|Government Financ...|Other capital tra...|          NULL|
|GFSA.SLS01G02144|2024.06|       482|PROVISIONAL|Local Government,...|Government Financ...|     Interest income|          NULL|
|GFSA.SLS01G01Z91|2024.06|     -2667|PROVISIONAL|Local Government,...|Government Financ...|Net lending/borro...|          NULL|
|GFSA.SLS01G02981|2024.06|      1960|PROVISIONAL|Local Government,...|Government Financ...|      Current grants|          NULL|
+---------------+-------+----------+-----------+--------------------+--------------------+--------------------+--------------+----
only showing top 5 rows
```

```
df.fillna('Series_title_2').show()
```

```
+---------------+-------+----------+-----------+--------------------+--------------------+--------------------+--------------+----
|Series_reference| Period|Data_value|     STATUS|               Group|             Subject|      Series_title_1|Series_title_2|Serie
+---------------+-------+----------+-----------+--------------------+--------------------+--------------------+--------------+----
|GFSA.SLS01G01Z90|2009.06|       715|      FINAL|Local Government,...|Government Financ...|Net operating bal...|Series_title_2|Serie
|GFSA.SLS01G01Z90|2010.06|       907|      FINAL|Local Government,...|Government Financ...|Net operating bal...|Series_title_2|Serie
|GFSA.SLS01G01Z90|2011.06|       350|      FINAL|Local Government,...|Government Financ...|Net operating bal...|Series_title_2|Serie
|GFSA.SLS01G01Z90|2012.06|       512|      FINAL|Local Government,...|Government Financ...|Net operating bal...|Series_title_2|Serie
|GFSA.SLS01G01Z90|2013.06|      1033|      FINAL|Local Government,...|Government Financ...|Net operating bal...|Series_title_2|Serie
|GFSA.SLS01G01Z90|2014.06|       944|      FINAL|Local Government,...|Government Financ...|Net operating bal...|Series_title_2|Serie
|GFSA.SLS01G01Z90|2015.06|      1033|      FINAL|Local Government,...|Government Financ...|Net operating bal...|Series_title_2|Serie
|GFSA.SLS01G01Z90|2016.06|      1154|      FINAL|Local Government,...|Government Financ...|Net operating bal...|Series_title_2|Serie
|GFSA.SLS01G01Z90|2017.06|      1106|      FINAL|Local Government,...|Government Financ...|Net operating bal...|Series_title_2|Serie
|GFSA.SLS01G01Z90|2018.06|      1554|      FINAL|Local Government,...|Government Financ...|Net operating bal...|Series_title_2|Serie
|GFSA.SLS01G01Z90|2019.06|      1758|      FINAL|Local Government,...|Government Financ...|Net operating bal...|Series_title_2|Serie
|GFSA.SLS01G01Z90|2020.06|      1360|      FINAL|Local Government,...|Government Financ...|Net operating bal...|Series_title_2|Serie
|GFSA.SLS01G01Z90|2021.06|      1790|      FINAL|Local Government,...|Government Financ...|Net operating bal...|Series_title_2|Serie
|GFSA.SLS01G01Z90|2022.06|      2183|    REVISED|Local Government,...|Government Financ...|Net operating bal...|Series_title_2|Serie
|GFSA.SLS01G01Z90|2023.06|      1960|    REVISED|Local Government,...|Government Financ...|Net operating bal...|Series_title_2|Serie
|GFSA.SLS01G01Z90|2024.06|      2010|PROVISIONAL|Local Government,...|Government Financ...|Net operating bal...|Series_title_2|Serie
|GFSA.SLS01G01Z91|2009.06|     -1092|      FINAL|Local Government,...|Government Financ...|Net lending/borro...|Series_title_2|Serie
|GFSA.SLS01G01Z91|2010.06|      -924|      FINAL|Local Government,...|Government Financ...|Net lending/borro...|Series_title_2|Serie
|GFSA.SLS01G01Z91|2011.06|      -849|      FINAL|Local Government,...|Government Financ...|Net lending/borro...|Series_title_2|Serie
|GFSA.SLS01G01Z91|2012.06|      -545|      FINAL|Local Government,...|Government Financ...|Net lending/borro...|Series_title_2|Serie
```

```
+----------------+------+----------+----------+-------------------+-------------------+-------------------+--------------+----
only showing top 20 rows
```

```python
df.dropDuplicates()
```

```
DataFrame[Series_reference: string, Period: double, Data_value: int, STATUS: string, Group: string, Subject: string,
Series_title_1: string, Series_title_2: string, Series_title_3: string, Series_title_4: string, Series_title_5: string, UNITS:
string, MAGNTUDE: int]
```

```python
df.show(2)
```

```
+----------------+-------+----------+------+-------------------+-------------------+-------------------+--------------+---------
|Series_reference| Period|Data_value|STATUS|              Group|            Subject|     Series_title_1|Series_title_2|Series_tit
+----------------+-------+----------+------+-------------------+-------------------+-------------------+--------------+---------
|GFSA.SLS01G01Z90|2009.06|       715| FINAL|Local Government,...|Government Financ...|Net operating bal...|          NULL|
|GFSA.SLS01G01Z90|2010.06|       907| FINAL|Local Government,...|Government Financ...|Net operating bal...|          NULL|
+----------------+-------+----------+------+-------------------+-------------------+-------------------+--------------+---------
only showing top 2 rows
```

```python
df.withColumn('new_col', df['Data_value'].cast('int')).show()
```

```
+----------------+-------+----------+-----------+-------------------+-------------------+-------------------+--------------+----
|Series_reference| Period|Data_value|     STATUS|              Group|            Subject|     Series_title_1|Series_title_2|Serie
+----------------+-------+----------+-----------+-------------------+-------------------+-------------------+--------------+----
|GFSA.SLS01G01Z90|2009.06|       715|      FINAL|Local Government,...|Government Financ...|Net operating bal...|          NULL|
|GFSA.SLS01G01Z90|2010.06|       907|      FINAL|Local Government,...|Government Financ...|Net operating bal...|          NULL|
|GFSA.SLS01G01Z90|2011.06|       350|      FINAL|Local Government,...|Government Financ...|Net operating bal...|          NULL|
|GFSA.SLS01G01Z90|2012.06|       512|      FINAL|Local Government,...|Government Financ...|Net operating bal...|          NULL|
|GFSA.SLS01G01Z90|2013.06|      1033|      FINAL|Local Government,...|Government Financ...|Net operating bal...|          NULL|
|GFSA.SLS01G01Z90|2014.06|       944|      FINAL|Local Government,...|Government Financ...|Net operating bal...|          NULL|
|GFSA.SLS01G01Z90|2015.06|      1033|      FINAL|Local Government,...|Government Financ...|Net operating bal...|          NULL|
|GFSA.SLS01G01Z90|2016.06|      1154|      FINAL|Local Government,...|Government Financ...|Net operating bal...|          NULL|
|GFSA.SLS01G01Z90|2017.06|      1106|      FINAL|Local Government,...|Government Financ...|Net operating bal...|          NULL|
|GFSA.SLS01G01Z90|2018.06|      1554|      FINAL|Local Government,...|Government Financ...|Net operating bal...|          NULL|
|GFSA.SLS01G01Z90|2019.06|      1758|      FINAL|Local Government,...|Government Financ...|Net operating bal...|          NULL|
|GFSA.SLS01G01Z90|2020.06|      1360|      FINAL|Local Government,...|Government Financ...|Net operating bal...|          NULL|
|GFSA.SLS01G01Z90|2021.06|      1790|      FINAL|Local Government,...|Government Financ...|Net operating bal...|          NULL|
|GFSA.SLS01G01Z90|2022.06|      2183|    REVISED|Local Government,...|Government Financ...|Net operating bal...|          NULL|
|GFSA.SLS01G01Z90|2023.06|      1960|    REVISED|Local Government,...|Government Financ...|Net operating bal...|          NULL|
|GFSA.SLS01G01Z90|2024.06|      2010|PROVISIONAL|Local Government,...|Government Financ...|Net operating bal...|          NULL|
|GFSA.SLS01G01Z91|2009.06|     -1092|      FINAL|Local Government,...|Government Financ...|Net lending/borro...|          NULL|
|GFSA.SLS01G01Z91|2010.06|      -924|      FINAL|Local Government,...|Government Financ...|Net lending/borro...|          NULL|
|GFSA.SLS01G01Z91|2011.06|      -849|      FINAL|Local Government,...|Government Financ...|Net lending/borro...|          NULL|
|GFSA.SLS01G01Z91|2012.06|      -545|      FINAL|Local Government,...|Government Financ...|Net lending/borro...|          NULL|
+----------------+-------+----------+-----------+-------------------+-------------------+-------------------+--------------+----
only showing top 20 rows
```

```python
df.sample(False, 0.1).show()
```

```
+----------------+-------+----------+-----------+-------------------+-------------------+-------------------+--------------+----
|Series_reference| Period|Data_value|     STATUS|              Group|            Subject|     Series_title_1|Series_title_2|Serie
+----------------+-------+----------+-----------+-------------------+-------------------+-------------------+--------------+----
|GFSA.SLS01G01Z90|2013.06|      1033|      FINAL|Local Government,...|Government Financ...|Net operating bal...|          NULL|
|GFSA.SLS01G01Z90|2018.06|      1554|      FINAL|Local Government,...|Government Financ...|Net operating bal...|          NULL|
|GFSA.SLS01G01Z90|2023.06|      1960|    REVISED|Local Government,...|Government Financ...|Net operating bal...|          NULL|
|GFSA.SLS01G01Z91|2023.06|     -2383|    REVISED|Local Government,...|Government Financ...|Net lending/borro...|          NULL|
|GFSA.SLS01G02121|2009.06|         0|      FINAL|Local Government,...|Government Financ...|Social security c...|          NULL|
|GFSA.SLS01G02121|2012.06|         0|      FINAL|Local Government,...|Government Financ...|Social security c...|          NULL|
|GFSA.SLS01G02121|2014.06|         0|      FINAL|Local Government,...|Government Financ...|Social security c...|          NULL|
|GFSA.SLS01G02121|2022.06|         0|      FINAL|Local Government,...|Government Financ...|Social security c...|          NULL|
|GFSA.SLS01G02140|2011.06|         0|      FINAL|Local Government,...|Government Financ...|       Other income|          NULL|
|GFSA.SLS01G02140|2012.06|         0|      FINAL|Local Government,...|Government Financ...|       Other income|          NULL|
|GFSA.SLS01G02140|2015.06|         0|      FINAL|Local Government,...|Government Financ...|       Other income|          NULL|
|GFSA.SLS01G02143|2013.06|       190|      FINAL|Local Government,...|Government Financ...|    Dividend income|          NULL|
|GFSA.SLS01G02143|2015.06|       256|      FINAL|Local Government,...|Government Financ...|    Dividend income|          NULL|
|GFSA.SLS01G02143|2018.06|       435|      FINAL|Local Government,...|Government Financ...|    Dividend income|          NULL|
|GFSA.SLS01G02144|2013.06|       143|      FINAL|Local Government,...|Government Financ...|    Interest income|          NULL|
|GFSA.SLS01G02144|2024.06|       482|PROVISIONAL|Local Government,...|Government Financ...|    Interest income|          NULL|
|GFSA.SLS01G02981|2016.06|       791|      FINAL|Local Government,...|Government Financ...|     Current grants|          NULL|
|GFSA.SLS01G02981|2017.06|       833|      FINAL|Local Government,...|Government Financ...|     Current grants|          NULL|
|GFSA.SLS01G02981|2018.06|       899|      FINAL|Local Government,...|Government Financ...|     Current grants|          NULL|
|GFSA.SLS01G02981|2022.06|      1450|      FINAL|Local Government,...|Government Financ...|     Current grants|          NULL|
+----------------+-------+----------+-----------+-------------------+-------------------+-------------------+--------------+----
only showing top 20 rows
```