sparksessionpractice

```
%pyspark
# Numero de casos positivos en Colombia, de acuerdo a 'ins.gov.co'
```

FINISHED

Took 0 sec. Last updated by anonymous at May 16 2020, 9:07:29 PM. (outdated)

```
%pyspark
from pyspark.sql import SparkSession
```

FINISHED

Took 0 sec. Last updated by anonymous at May 16 2020, 9:07:32 PM.

```
%pyspark
spark = SparkSession.builder.appName('Basics').getOrCreate()
```

FINISHED

Took 0 sec. Last updated by anonymous at May 16 2020, 9:07:34 PM.

```
%pyspark

SensorDataRDD = sc.textFile('s3://aws-logs-061895363576-us-east-1/colombia/Casos_positivos_de_COVI
SensorDataCount = SensorDataRDD.count()
SensorDataCount
```

FINISHED

```
14217
```

Took 7 sec. Last updated by anonymous at May 16 2020, 9:07:45 PM.

```
%pyspark
df = spark.read.csv('s3://aws-logs-061895363576-us-east-1/colombia/Casos_positivos_de_COVID-19_en_
```

FINISHED

Took 1 sec. Last updated by anonymous at May 16 2020, 9:07:53 PM.

```
%pyspark
df.columns
```

FINISHED

```
['ID de caso', 'Fecha de notificación', 'Codigo DIVIPOLA', 'Ciudad de ubicación', 'Departamento o D
istrito ', 'atención', 'Edad', 'Sexo', 'Tipo', 'Estado', 'País de procedencia', 'FIS', 'Fecha de mu
erte', 'Fecha diagnostico', 'Fecha recuperado', 'fecha reporte web']
```

Took 0 sec. Last updated by anonymous at May 16 2020, 9:07:56 PM.

```
%pyspark
df.printSchema()
```

FINISHED

```
root
 |-- ID de caso: string (nullable = true)
 |-- Fecha de notificación: string (nullable = true)
 |-- Codigo DIVIPOLA: string (nullable = true)
 |-- Ciudad de ubicación: string (nullable = true)
 |-- Departamento o Distrito : string (nullable = true)
 |-- atención: string (nullable = true)
 |-- Edad: string (nullable = true)
```

```
 |-- Sexo: string (nullable = true)
 |-- Tipo: string (nullable = true)
 |-- Estado: string (nullable = true)
 |-- País de procedencia: string (nullable = true)
 |-- FIS: string (nullable = true)
 |-- Fecha de muerte: string (nullable = true)
 |-- Fecha diagnostico: string (nullable = true)
 |-- Fecha recuperado: string (nullable = true)
 |-- fecha reporte web: string (nullable = true)
```

sparksessionpractice

Took 0 sec. Last updated by anonymous at May 16 2020, 9:08:03 PM.

```
%pyspark

df.head(5)
```

FINISHED

```
[Row(ID de caso='1', Fecha de notificación='2020-03-02T00:00:00.000', Codigo DIVIPOLA='11001', Ciudad de ubicación='Bogotá D.C.', Departamento o Distrito ='Bogotá D.C.', atención='Recuperado', Edad='19', Sexo='F', Tipo='Importado', Estado='Leve', País de procedencia='Italia', FIS='2020-02-27T00:00:00.000', Fecha de muerte='-   -', Fecha diagnostico='2020-03-06T00:00:00.000', Fecha recuperado='2020-03-13T00:00:00.000', fecha reporte web='2020-03-06T00:00:00.000'), Row(ID de caso='2', Fecha de notificación='2020-03-06T00:00:00.000', Codigo DIVIPOLA='76111', Ciudad de ubicación='Guadalajara de Buga', Departamento o Distrito ='Valle del Cauca', atención='Recuperado', Edad='34', Sexo='M', Tipo='Importado', Estado='Leve', País de procedencia='España', FIS='2020-03-04T00:00:00.000', Fecha de muerte='-   -', Fecha diagnostico='2020-03-09T00:00:00.000', Fecha recuperado='2020-03-19T00:00:00.000', fecha reporte web='2020-03-09T00:00:00.000'), Row(ID de caso='3', Fecha de notificación='2020-03-07T00:00:00.000', Codigo DIVIPOLA='5001', Ciudad de ubicación='Medellín', Departamento o Distrito ='Antioquia', atención='Recuperado', Edad='50', Sexo='F', Tipo='Importado', Estado='Leve', País de procedencia='España', FIS='2020-02-29T00:00:00.000', Fecha de muerte='-   -', Fecha diagnostico='2020-03-09T00:00:00.000', Fecha recuperado='2020-03-15T00:00:00.000', fecha reporte web='2020-03-09T00:00:00.000'), Row(ID de caso='4', Fecha de notificación='2020-03-09T00:00:00.000', Codigo DIVIPOLA='5001', Ciudad de ubicación='Medellín', Departamento o Distrito ='Antioquia', atención='Recuperado', Edad='55', Sexo='M', Tipo='Relacionado', Estado='Leve', País de procedencia='Colombia', FIS='2020-03-06T00:00:00.000', Fecha de muerte='-
```

Took 1 sec. Last updated by anonymous at May 16 2020, 9:08:36 PM.

```
%pyspark

df.describe().show()
```

FINISHED

```
---------------+----------+-----------------+-----+----------+-----------+-----------------+--------------------+-------------------+-------------------+----------+
|summary|       ID de caso|Fecha de notificación|   Codigo DIVIPOLA|Ciudad de ubicación|Departamento o Distrito |  atención|              Edad| Sexo|       Tipo|      Estado|País de procedencia|               FIS|     Fecha de muerte|   Fecha diagnostico|    Fecha recuperado|   fecha reporte web|
+-------+-----------------+--------------------+-----------------+-------------------+--------------------+----------+------------------+-----+-----------+------------+-------------------+-------------------+--------------------+--------------------+--------------------+--------------------+----------+
|  count|            14216|                14216|            14216|              14216|               14216|     14216|             14216|14216|      14201|       14216|              14216|              14216|               14216|               14216|               14216|               14216|
|   mean|7136.179727068093|                null| 32188.87239729882|               null|                null|      null| 39.78601575689364| null|       null|        null|               null|               null|                null|                null|                null|                null|
| stddev|4118.718201906753|                null|29136.690713820568|               null|                null|      null|                 | null|       null|        null|               null|               null|                null|                null|                null|                null|
```

Took 3 sec. Last updated by anonymous at May 16 2020, 9:08:43 PM.

```
%pyspark

df.orderBy(df['Edad'].desc()).show()
```

FINISHED

sparksessionpractice

```
+----------+--------------------+--------------+------------------+------------------------+-
---------+----+----+-----------+-----------+------------------+--------------------+---------
---------+--------------------+--------------------+--------------------+
|ID de caso|Fecha de notificación|Codigo DIVIPOLA|Ciudad de ubicación|Departamento o Distrito |
atención|Edad|Sexo|       Tipo|      Estado|País de procedencia|                 FIS|     Fecha
de muerte|   Fecha diagnostico|    Fecha recuperado|   fecha reporte web|
+----------+--------------------+--------------+------------------+------------------------+-
---------+----+----+-----------+-----------+------------------+--------------------+---------
---------+--------------------+--------------------+--------------------+
|      2172| 2020-03-28T00:00:...|          54498|             Ocaña|      Norte de Santander|
Fallecido|  98|   M| En estudio|   Fallecido|          Colombia|2020-03-24T00:00:...|2020-03-28
T00:00:...|2020-04-09T00:00:...|                -   -|2020-04-09T00:00:...|
|     13814| 2020-05-08T00:00:...|           8001|       Barranquilla|       Barranquilla D.E.|
Fallecido|  98|   F| En estudio|   Fallecido|          Colombia|2020-05-08T00:00:...|2020-05-11
T00:00:...|2020-05-15T00:00:...|                -   -|2020-05-15T00:00:...|
|      1483| 2020-04-03T00:00:...|          11001|        Bogotá D.C.|             Bogotá D.C.|
Fallecido|  97|   F| En estudio|   Fallecido|          Colombia|2020-03-30T00:00:...|2020-04-14
T00:00:...|2020-04-05T00:00:...|                -   -|2020-04-05T00:00:...|
```

Took 0 sec. Last updated by anonymous at May 16 2020, 9:09:32 PM.

```
%pyspark
from pyspark.sql.functions import mean
```

FINISHED

Took 0 sec. Last updated by anonymous at May 16 2020, 9:09:52 PM.

```
%pyspark
df.select(mean("Edad")).show()
# Media de edad de los casos confirmados
```

FINISHED

```
+-----------------+
|        avg(Edad)|
+-----------------+
|39.78601575689364|
+-----------------+
```

Took 0 sec. Last updated by anonymous at May 16 2020, 9:10:16 PM.

```
%pyspark
```

READY