

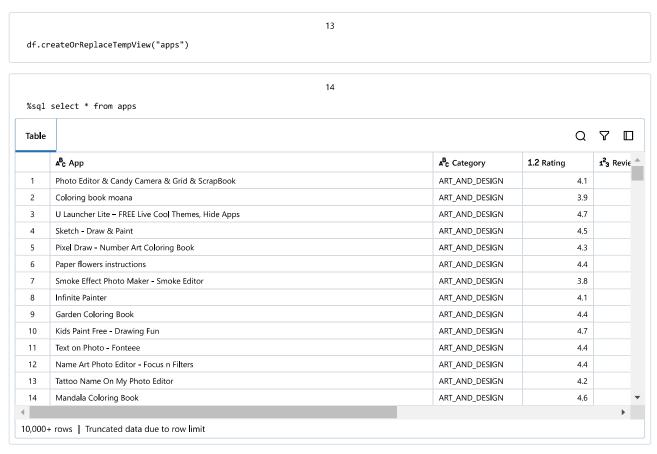
```
(https://databricks.com)
                                                            1
   import pyspark
   from pyspark.sql import SparkSession
   from pyspark.sql.types import StructType, StructField, StringType, IntegerType
   from pyspark.sql.functions import *
   df=spark.read.load('/FileStore/tables/googleplaystore-2.csv',format='csv',sep=',',header='true',escape='"',inferschema='true')
                                                            3
   df.count()
 Out[5]: 10841
   df.show(1)
 +------
                        Category|Rating|Reviews|Size|Installs|Type|Price|Content Rating| Genres| Last Updated|Current Ve
 r Android Ver
 |Photo Editor & Ca...|ART_AND_DESIGN| 4.1| 159| 19M| 10,000+|Free| 0|
                                                                         Everyone|Art & Design|January 7, 2018|
 0 4.0.3 and up
 only showing top 1 row
                                                     5: CHECK SCHEMA
   df.printSchema()
 root
  |-- App: string (nullable = true)
  |-- Category: string (nullable = true)
  |-- Rating: double (nullable = true)
  |-- Reviews: string (nullable = true)
  |-- Size: string (nullable = true)
  |-- Installs: string (nullable = true)
  |-- Type: string (nullable = true)
  |-- Price: string (nullable = true)
  |-- Content Rating: string (nullable = true)
  |-- Genres: string (nullable = true)
  |-- Last Updated: string (nullable = true)
  |-- Current Ver: string (nullable = true)
  |-- Android Ver: string (nullable = true)
                                                     6: DATA CLEANING
   df=df.drop("size","Content Rating","Last Updated","Android Ver")
   df.show(2)
```

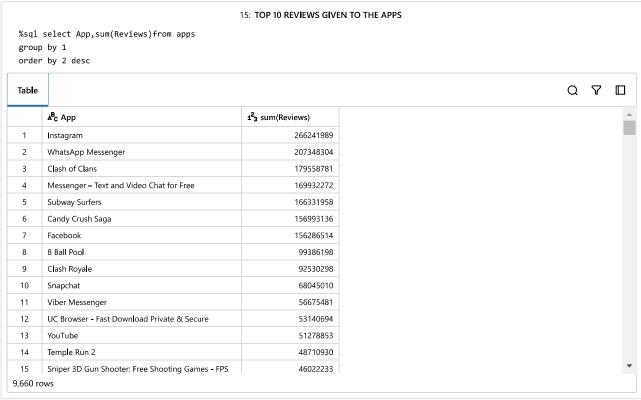
```
8
df=df.drop('Current Ver')
```

```
df.printSchema()

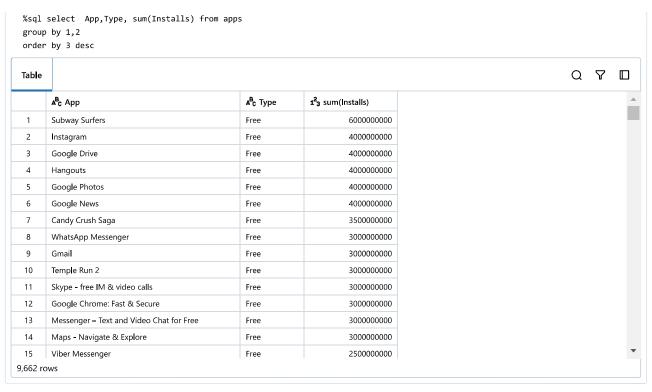
root
    |-- App: string (nullable = true)
    |-- Category: string (nullable = true)
    |-- Rating: double (nullable = true)
    |-- Reviews: string (nullable = true)
    |-- Installs: string (nullable = true)
    |-- Type: string (nullable = true)
    |-- Price: string (nullable = true)
    |-- Genres: string (nullable = true)
```

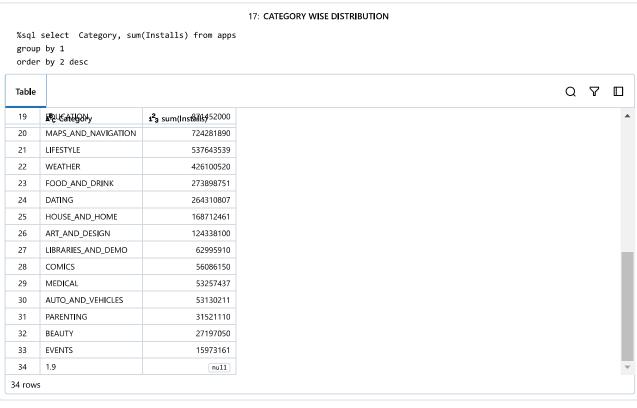
```
from pyspark.sql.functions import regexp_replace,col
from pyspark.sql.types import IntegerType
df = df.withColumn("Reviews", col("Reviews").cast(IntegerType())) \
    .withColumn("Installs", regexp_replace(col("Installs"), "[^0-9]", "")) \
    .withColumn("Installs", col("Installs").cast(IntegerType())) \
    .withColumn("Price", regexp_replace(col("Price"), "[$]", "")) \
    .withColumn("Price", col("Price").cast(IntegerType()))
```

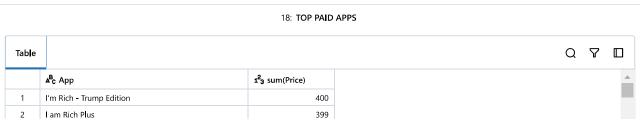




	16: TOP 10 INST A	LLED APPS	







3	I AM RICH PRO PLUS	399
4	l'm Rich/Eu sou Rico/أنا غني/我很有錢	399
5	I Am Rich Premium	399
6	most expensive app (H)	399
7	I Am Rich Pro	399
8	I am rich(premium)	399
9	I am Rich	399
10	I am Rich!	399
11	✓ I'm rich	399
12	I am rich (Most expensive app)	399
13	I am rich	399
14	Eu Sou Rico	394
15	I Am Rich	389
756 ro	ws	