databricksgooglepaystore project

(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 *
```

2

```
df=spark.read.load('/FileStore/tables/googleplaystore-2.csv',format='csv',sep=',',header='true',escape='"',inferschema='true')
```

3

```
df.count()
```

```
Out[5]: 10841
```

4

```
df.show(1)
```

```
+--------------------+--------------+------+-------+----+--------+----+-----+--------------+------------+---------------+----------
-+------------+
|                 App|      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|      1.0.
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")
```

7

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

```
+--------------------+--------------+------+-------+--------+----+-----+--------------------+-----------+
|                 App|      Category|Rating|Reviews|Installs|Type|Price|              Genres|Current Ver|
+--------------------+--------------+------+-------+--------+----+-----+--------------------+-----------+
|Photo Editor & Ca...|ART_AND_DESIGN|   4.1|    159| 10,000+|Free|    0|        Art & Design|      1.0.0|
| Coloring book moana|ART_AND_DESIGN|   3.9|    967|500,000+|Free|    0|Art & Design;Pret...|      2.0.0|
+--------------------+--------------+------+-------+--------+----+-----+--------------------+-----------+
only showing top 2 rows
```

8

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

9

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

```
+--------------------+--------------+------+-------+--------+----+-----+--------------------+
|                 App|      Category|Rating|Reviews|Installs|Type|Price|              Genres|
+--------------------+--------------+------+-------+--------+----+-----+--------------------+
|Photo Editor & Ca...|ART_AND_DESIGN|   4.1|    159| 10,000+|Free|    0|        Art & Design|
| Coloring book moana|ART_AND_DESIGN|   3.9|    967|500,000+|Free|    0|Art & Design;Pret...|
+--------------------+--------------+------+-------+--------+----+-----+--------------------+
only showing top 2 rows
```

10

```
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)
```

11

```
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()))
```

12

```
df.show(5)
```

```
+--------------------+--------------+------+-------+--------+----+-----+--------------------+
|                 App|      Category|Rating|Reviews|Installs|Type|Price|              Genres|
+--------------------+--------------+------+-------+--------+----+-----+--------------------+
|Photo Editor & Ca...|ART_AND_DESIGN|   4.1|    159|   10000|Free|    0|        Art & Design|
| Coloring book moana|ART_AND_DESIGN|   3.9|    967|  500000|Free|    0|Art & Design;Pret...|
|U Launcher Lite –...|ART_AND_DESIGN|   4.7|  87510| 5000000|Free|    0|        Art & Design|
|Sketch - Draw & P...|ART_AND_DESIGN|   4.5| 215644|50000000|Free|    0|        Art & Design|
|Pixel Draw - Numb...|ART_AND_DESIGN|   4.3|    967|  100000|Free|    0|Art & Design;Crea...|
+--------------------+--------------+------+-------+--------+----+-----+--------------------+
only showing top 5 rows
```

13

```
df.createOrReplaceTempView("apps")
```

14

```
%sql select * from apps
```

Table

	App	Category	1.2 Rating	Revie
1	Photo Editor & Candy Camera & Grid & ScrapBook	ART_AND_DESIGN	4.1	
2	Coloring book moana	ART_AND_DESIGN	3.9	
3	U Launcher Lite – FREE Live Cool Themes, Hide Apps	ART_AND_DESIGN	4.7	
4	Sketch - Draw & Paint	ART_AND_DESIGN	4.5	
5	Pixel Draw - Number Art Coloring Book	ART_AND_DESIGN	4.3	
6	Paper flowers instructions	ART_AND_DESIGN	4.4	
7	Smoke Effect Photo Maker - Smoke Editor	ART_AND_DESIGN	3.8	
8	Infinite Painter	ART_AND_DESIGN	4.1	
9	Garden Coloring Book	ART_AND_DESIGN	4.4	
10	Kids Paint Free - Drawing Fun	ART_AND_DESIGN	4.7	
11	Text on Photo - Fonteee	ART_AND_DESIGN	4.4	
12	Name Art Photo Editor - Focus n Filters	ART_AND_DESIGN	4.4	
13	Tattoo Name On My Photo Editor	ART_AND_DESIGN	4.2	
14	Mandala Coloring Book	ART_AND_DESIGN	4.6	

10,000+ rows | Truncated data due to row limit

15: **TOP 10 REVIEWS GIVEN TO THE APPS**

```
%sql select App,sum(Reviews)from apps
group by 1
order by 2 desc
```

Table

	App	sum(Reviews)
1	Instagram	266241989
2	WhatsApp Messenger	207348304
3	Clash of Clans	179558781
4	Messenger – Text and Video Chat for Free	169932272
5	Subway Surfers	166331958
6	Candy Crush Saga	156993136
7	Facebook	156286514
8	8 Ball Pool	99386198
9	Clash Royale	92530298
10	Snapchat	68045010
11	Viber Messenger	56675481
12	UC Browser - Fast Download Private & Secure	53140694
13	YouTube	51278853
14	Temple Run 2	48710930
15	Sniper 3D Gun Shooter: Free Shooting Games - FPS	46022233

9,660 rows

16: **TOP 10 INSTALLED APPS**

```
%sql select  App,Type, sum(Installs) from apps
group by 1,2
order by 3 desc
```

Table

	App	Type	sum(Installs)
1	Subway Surfers	Free	6000000000
2	Instagram	Free	4000000000
3	Google Drive	Free	4000000000
4	Hangouts	Free	4000000000
5	Google Photos	Free	4000000000
6	Google News	Free	4000000000
7	Candy Crush Saga	Free	3500000000
8	WhatsApp Messenger	Free	3000000000
9	Gmail	Free	3000000000
10	Temple Run 2	Free	3000000000
11	Skype - free IM & video calls	Free	3000000000
12	Google Chrome: Fast & Secure	Free	3000000000
13	Messenger – Text and Video Chat for Free	Free	3000000000
14	Maps - Navigate & Explore	Free	3000000000
15	Viber Messenger	Free	2500000000

9,662 rows

17: CATEGORY WISE DISTRIBUTION

```
%sql select  Category, sum(Installs) from apps
group by 1
order by 2 desc
```

Table

	Category	sum(Installs)
19	EDUCATION	871452000
20	MAPS_AND_NAVIGATION	724281890
21	LIFESTYLE	537643539
22	WEATHER	426100520
23	FOOD_AND_DRINK	273898751
24	DATING	264310807
25	HOUSE_AND_HOME	168712461
26	ART_AND_DESIGN	124338100
27	LIBRARIES_AND_DEMO	62995910
28	COMICS	56086150
29	MEDICAL	53257437
30	AUTO_AND_VEHICLES	53130211
31	PARENTING	31521110
32	BEAUTY	27197050
33	EVENTS	15973161
34	1.9	null

34 rows

18: TOP PAID APPS

Table

	App	sum(Price)
1	I'm Rich - Trump Edition	400
2	I am Rich Plus	399

3	I AM RICH PRO PLUS	399
4	I'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 rows