databricksDatabricks Project

(https://databricks.com)

Analysis on googleplaystore data

Notebook is created by Danish A GitHub link (https://github.com/danisha138/DataWarehousing)

Source Youtube Channel learnbydoingit - Source (https://www.youtube.com/@learnbydoingit)

Datasets Downloaded from Kaggle kaggle (https://www.kaggle.com/datasets/lava18/google-play-store-apps)

Import Library

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

Check the Loaded file on below location

```
%fs ls dbfs:/FileStore/tables/
```

Table

	path	name	size	modificationTime
1	dbfs:/FileStore/tables/googleplaystore.csv	googleplaystore.csv	1360155	1687692374000

1 row

Create Dataframe

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

Check the count of loaded file

```
df.count()
```

```
Out[4]: 10841
```

To check Top 5 records

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

```
+--------------------+--------------+------+-------+----+-----------+----+-----+--------------+--------------------+-----
-----------+------------------+------------+
|                 App|      Category|Rating|Reviews|Size|   Installs|Type|Price|Content Rating|              Genres|    L
ast Updated|       Current Ver| Android Ver|
+--------------------+--------------+------+-------+----+-----------+----+-----+--------------+--------------------+-----
-----------+------------------+------------+
|Photo Editor & Ca...|ART_AND_DESIGN|   4.1|    159| 19M|    10,000+|Free|    0|      Everyone|        Art & Design| Janu
ary 7, 2018|             1.0.0|4.0.3 and up|
| Coloring book moana|ART_AND_DESIGN|   3.9|    967| 14M|   500,000+|Free|    0|      Everyone|Art & Design;Pret...|Janua
ry 15, 2018|             2.0.0|4.0.3 and up|
|U Launcher Lite -...|ART_AND_DESIGN|   4.7|  87510|8.7M| 5,000,000+|Free|    0|      Everyone|        Art & Design|  Aug
```

```
ust 1, 2018|             1.2.4|4.0.3 and up|
|Sketch - Draw & P...|ART_AND_DESIGN|   4.5| 215644| 25M|50,000,000+|Free|    0|          Teen|        Art & Design|    J
une 8, 2018|Varies with device|  4.2 and up|
|Pixel Draw - Numb...|ART_AND_DESIGN|   4.3|    967|2.8M|   100,000+|Free|    0|      Everyone|Art & Design;Crea...|   Ju
ne 20, 2018|               1.1|  4.4 and up|
+--------------------+--------------+------+-------+----+-----------+----+-----+--------------+--------------------+-----
-----------+-----------------+------------+
only showing top 5 rows
```

Print the 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)
```

Drop the unnessary columns

```
df=df.drop("size","Content Rating","Last Updated","Android Ver","Current Ver")
```

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

```
+--------------------+--------------+------+-------+-----------+----+-----+--------------------+
|                 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...|
|U Launcher Lite -...|ART_AND_DESIGN|   4.7|  87510| 5,000,000+|Free|    0|        Art & Design|
|Sketch - Draw & P...|ART_AND_DESIGN|   4.5| 215644|50,000,000+|Free|    0|        Art & Design|
|Pixel Draw - Numb...|ART_AND_DESIGN|   4.3|    967|   100,000+|Free|    0|Art & Design;Crea...|
+--------------------+--------------+------+-------+-----------+----+-----+--------------------+
only showing top 5 rows
```

Change the Datatype and Cleaning of data

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

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()))
```

```
df.printSchema()
```

```
root
 |-- App: string (nullable = true)
 |-- Category: string (nullable = true)
```

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

```
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
```

Create the Temp view of dataframe

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

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

Table

	App	Category
1	Photo Editor & Candy Camera & Grid & ScrapBook	ART_AND_DESIGN
2	Coloring book moana	ART_AND_DESIGN
3	U Launcher Lite – FREE Live Cool Themes, Hide Apps	ART_AND_DESIGN
4	Sketch - Draw & Paint	ART_AND_DESIGN
5	Pixel Draw - Number Art Coloring Book	ART_AND_DESIGN
6	Paper flowers instructions	ART_AND_DESIGN
7	Smoke Effect Photo Maker - Smoke Editor	ART_AND_DESIGN

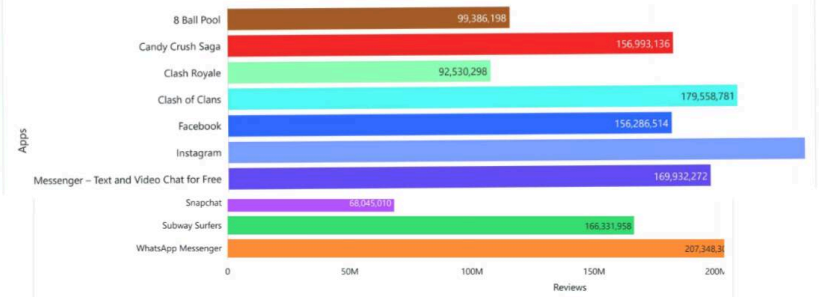
10,000 rows | Truncated data

Top 10 reviews given to apps

```
%sql

select app,sum(reviews) from apps
group by 1
order by 2 desc limit 10;
```

Table Visualization 1



10 rows

Top paid price

```
%sql

select app,sum(price) from apps
where type='Paid'
group by 1
order by 2 desc
```

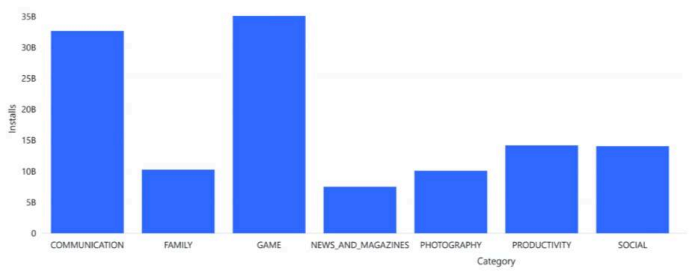
Table

	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

756 rows

Category wise distribution

Table Visualization 1



10 rows