databricks PlayStore_Analysis

(https://databricks.com)

Importing Libs:

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

Loading the dataset:

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

Cleaning the data:

```
df.count()
```

```
Out[7]: 10840
```

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

```
+------+--------------------+--------------+-------+----+--------+----+-----+--------------+------------+------------+-----------+------------+
|Rating|                 App|      Category|Reviews|Size|Installs|Type|Price|Content Rating|      Genres|Last Updated|Current Ver| Android Ver|
+------+--------------------+--------------+-------+----+--------+----+-----+--------------+------------+------------+-----------+------------+
|   4.1|Photo Editor & Ca...|ART_AND_DESIGN|    159| 19M| 10,000+|Free|    0|      Everyone|Art & Design|   07-Jan-18|      1.0.0|4.0.3 and up|
+------+--------------------+--------------+-------+----+--------+----+-----+--------------+------------+------------+-----------+------------+
only showing top 1 row
```

```
df.printSchema()
```

```
root
 |-- Rating: double (nullable = true)
 |-- App: string (nullable = true)
 |-- Category: string (nullable = true)
 |-- Reviews: integer (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)
```

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

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

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

```
df.printSchema()
```

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

```
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.show()
```

```
+------+--------------------+--------------+-------+--------+----+-----+--------------------+------------------+
|Rating|                 App|      Category|Reviews|Installs|Type|Price|              Genres|       Current Ver|
+------+--------------------+--------------+-------+--------+----+-----+--------------------+------------------+
|   4.1|Photo Editor & Ca...|ART_AND_DESIGN|    159|    null|Free|    0|        Art & Design|             1.0.0|
|   3.9| Coloring book moana|ART_AND_DESIGN|    967|    null|Free|    0|Art & Design;Pret...|             2.0.0|
|   4.7|U Launcher Lite –...|ART_AND_DESIGN|  87510|    null|Free|    0|        Art & Design|             1.2.4|
|   4.5|Sketch - Draw & P...|ART_AND_DESIGN| 215644|    null|Free|    0|        Art & Design|Varies with device|
|    4.3|Pixel Draw - Numb...|ART_AND_DESIGN|    967|    null|Free|    0|Art & Design;Crea...|                1.1|
|    4.4|Paper flowers ins...|ART_AND_DESIGN|    167|    null|Free|    0|        Art & Design|                  1|
|    3.8|Smoke Effect Phot...|ART_AND_DESIGN|    178|    null|Free|    0|        Art & Design|                1.1|
|    4.1|    Infinite Painter|ART_AND_DESIGN|  36815|    null|Free|    0|        Art & Design|           6.1.61.1|
|    4.4|Garden Coloring Book|ART_AND_DESIGN|  13791|    null|Free|    0|        Art & Design|              2.9.2|
|    4.7|Kids Paint Free -...|ART_AND_DESIGN|    121|    null|Free|    0|Art & Design;Crea...|                2.8|
|    4.4|Text on Photo - F...|ART_AND_DESIGN|  13880|    null|Free|    0|        Art & Design|              1.0.4|
|    4.4|Name Art Photo Ed...|ART_AND_DESIGN|   8788|    null|Free|    0|        Art & Design|             1.0.15|
|    4.2|Tattoo Name On My...|ART_AND_DESIGN|  44829|    null|Free|    0|        Art & Design|                3.8|
|    4.6|Mandala Coloring ...|ART_AND_DESIGN|   4326|    null|Free|    0|        Art & Design|              1.0.4|
|    4.4|3D Color Pixel by...|ART_AND_DESIGN|   1518|    null|Free|    0|        Art & Design|              1.2.3|
|    3.2|Learn To Draw Kaw...|ART_AND_DESIGN|     55|    null|Free|    0|        Art & Design|                NaN|
|    4.7|Photo Designer - ...|ART_AND_DESIGN|   3632|    null|Free|    0|        Art & Design|                3.1|
|    4.5|350 Diy Room Deco...|ART_AND_DESIGN|     27|    null|Free|    0|        Art & Design|                  1|
```

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

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

Table

	Rating	App	Category	Reviews	Installs	Typ
1	4.1	Photo Editor & Candy Camera & Grid & ScrapBook	ART_AND_DESIGN	159	null	Free
2	3.9	Coloring book moana	ART_AND_DESIGN	967	null	Free
3	4.7	U Launcher Lite – FREE Live Cool Themes, Hide Apps	ART_AND_DESIGN	87510	null	Free
4	4.5	Sketch - Draw & Paint	ART_AND_DESIGN	215644	null	Free
5	4.3	Pixel Draw - Number Art Coloring Book	ART_AND_DESIGN	967	null	Free
6	4.4	Paper flowers instructions	ART_AND_DESIGN	167	null	Free
7	3.8	Smoke Effect Photo Maker - Smoke Editor	ART_AND_DESIGN	178	null	Free

10,000 rows | Truncated data

Checking top 10 reviews given to a app:-

```
%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

9,659 rows

Which are top 10 installed App:

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

Table

	App	sum(Installs)
1	Google Chrome: Fast & Secure	null
2	free video calls and chat	null
3	Toddler Learning Games - Little Kids Games	null
4	MyChart	null
5	Davis's Drug Guide for Nurses	null
6	Diabetes Testing	null
7	Mercari: The Selling App	null

9,659 rows

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

Table

	App	Type	sum(Installs)
1	Mail.Ru - Email App	Free	null
2	RandoChat - Chat roulette	Free	null
3	BZWBK24 mobile	Free	null
4	The Cube	Free	null
5	Sago Mini Friends	Free	null
6	Manage My Pain Pro	Paid	null
7	Ada - Your Health Guide	Free	null

9,661 rows

Category wise apps:

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

Table

	Category	sum(Installs)
1	EVENTS	null
2	COMICS	null
3	SPORTS	null
4	WEATHER	null
5	VIDEO_PLAYERS	null
6	AUTO_AND_VEHICLES	null
7	PARENTING	null

33 rows

Top Paid Apps:

```
%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

Apps with High Ratings:

```
%sql select * from apps
WHERE Rating > 4.5
```

Table

	Rating	App	Category	Reviews	Installs	Typ
1	4.7	U Launcher Lite – FREE Live Cool Themes, Hide Apps	ART_AND_DESIGN	87510	null	Free
2	4.7	Kids Paint Free - Drawing Fun	ART_AND_DESIGN	121	null	Free
3	4.6	Mandala Coloring Book	ART_AND_DESIGN	4326	null	Free
4	4.7	Photo Designer - Write your name with shapes	ART_AND_DESIGN	3632	null	Free
5	4.6	ibis Paint X	ART_AND_DESIGN	224399	null	Free
6	4.7	Superheroes Wallpapers | 4K Backgrounds	ART_AND_DESIGN	7699	null	Free
7	NaN	Mcqueen Coloring pages	ART_AND_DESIGN	61	null	Free

3,391 rows

Count Apps in Each Category:

```
%sql select Category, COUNT(*) AS Count
FROM apps GROUP BY Category
```

Table

	Category	Count
1	EVENTS	64
2	COMICS	60
3	SPORTS	384
4	WEATHER	82
5	VIDEO_PLAYERS	175
6	AUTO_AND_VEHICLES	85
7	PARENTING	60

33 rows

Top Categories by Average Rating:

```
%sql select Category, AVG(Rating) AS AvgRating
FROM apps
GROUP BY Category
ORDER BY AvgRating DESC
```

Table

	Category	AvgRating
1	EVENTS	NaN
2	COMICS	NaN
3	SPORTS	NaN
4	WEATHER	NaN
5	VIDEO_PLAYERS	NaN
6	AUTO_AND_VEHICLES	NaN
7	PARENTING	NaN

33 rows

Apps with the Highest Price:

```
%sql select * from apps
ORDER BY Price DESC
```

Table

	Rating	App	Category	Reviews	Installs	Typ
1	3.6	I'm Rich - Trump Edition	LIFESTYLE	275	null	Paid
2	4.3	most expensive app (H)	FAMILY	6	null	Paid
3	3.8	💎 I'm rich	LIFESTYLE	718	null	Paid
4	3.8	I am rich	LIFESTYLE	3547	null	Paid
5	4	I am Rich Plus	FAMILY	856	null	Paid
6	4.1	I Am Rich Premium	FINANCE	1867	null	Paid
7	3.8	I am Rich!	FINANCE	93	null	Paid

10,000 rows | Truncated data

Apps with the Most Installs:

```
%sql select * FROM apps ORDER BY Installs DESC
```

Table

	Rating	App	Category	Reviews	Installs	Typ
1	4.1	Photo Editor & Candy Camera & Grid & ScrapBook	ART_AND_DESIGN	159	null	Free
2	3.9	Coloring book moana	ART_AND_DESIGN	967	null	Free
3	4.7	U Launcher Lite – FREE Live Cool Themes, Hide Apps	ART_AND_DESIGN	87510	null	Free
4	4.5	Sketch - Draw & Paint	ART_AND_DESIGN	215644	null	Free

5	4.3	Pixel Draw - Number Art Coloring Book	ART_AND_DESIGN	967	null	Free
6	4.4	Paper flowers instructions	ART_AND_DESIGN	167	null	Free
7	3.8	Smoke Effect Photo Maker - Smoke Editor	ART_AND_DESIGN	178	null	Free

10,000 rows | Truncated data

Free vs. Paid Apps:

```
%sql select Type, COUNT(*) AS Count FROM apps GROUP BY Type
```

Table

	Type	Count
1	NaN	1
2	Free	10039
3	Paid	800

3 rows

Apps with the Highest Price in Each Category:

```
%sql WITH RankedApps AS (
  SELECT
    *,
    ROW_NUMBER() OVER (PARTITION BY Category ORDER BY Price DESC) AS rank
  FROM apps
)
SELECT * FROM RankedApps WHERE rank = 1
```

Table

	Rating	App	Category	Reviews	Installs	Type	Price	Genres	Curre
1	4.7	X Launcher Pro: PhoneX Theme, OS11 Control Center	ART_AND_DESIGN	801	null	Paid	1	Art & Design	2.1.2
2	NaN	FORD V SERIES CALC - NO LIMIT	AUTO_AND_VEHICLES	2	null	Paid	9	Auto & Vehicles	3.0.0
3	4.7	Hush - Beauty for Everyone	BEAUTY	18900	null	Free	0	Beauty	6.10.1
4	NaN	FN pistol model 1903 explained	BOOKS_AND_REFERENCE	1	null	Paid	6	Books & Reference	Andro
5	NaN	Lean EQ	BUSINESS	6	null	Paid	89	Business	1
6	4.5	LINE WEBTOON - Free Comics	COMICS	1013635	null	Free	0	Comics	Varies
7	NaN	Z PIVOT	COMMUNICATION	0	null	Paid	19	Communication	1.3

33 rows

Top Apps in Each Category by Average Rating:

```
%sql WITH AvgRatings AS (
  SELECT
    Category,
    App,
    Rating,
    ROW_NUMBER() OVER (PARTITION BY Category ORDER BY Rating DESC) AS rank
  FROM apps
)
SELECT * FROM AvgRatings WHERE rank = 1
```

Table

	Category	App	Rating	rank
1	ART_AND_DESIGN	Mcqueen Coloring pages	NaN	1
2	AUTO_AND_VEHICLES	FORD V SERIES CALC - NO LIMIT	NaN	1
3	BEAUTY	Wrinkles and rejuvenation	NaN	1
4	BOOKS_AND_REFERENCE	Anonymous caller detection	NaN	1
5	BUSINESS	Y! Mobile menu	NaN	1
6	COMICS	【Ranobbe complete free】 Novelba - Free app that you can read and write novels	NaN	1
7	COMMUNICATION	J Alvarei Moii	NaN	1

33 rows

Categories with the Most Expensive Apps on Average:

```
%sql SELECT
  Category,
  AVG(Price) AS AvgPrice
FROM apps
GROUP BY Category
HAVING AVG(Price) = (SELECT MAX(AvgPrice) FROM (SELECT Category, AVG(Price) AS AvgPrice FROM apps GROUP BY Category) AS Averages)
```

Table

	Category	AvgPrice
1	FINANCE	7.879781420765028

1 row

Apps with the Highest Price-to-Rating Ratio:

```
%sql SELECT
  App,
  Category,
  (Price / Rating) AS PriceToRatingRatio
FROM apps
WHERE Rating IS NOT NULL AND Price IS NOT NULL
ORDER BY PriceToRatingRatio DESC
```

Table

	App	Category	PriceToRatingRatio
1	Mcqueen Coloring pages	ART_AND_DESIGN	NaN
2	Wrinkles and rejuvenation	BEAUTY	NaN
3	Manicure - nail design	BEAUTY	NaN
4	Skin Care and Natural Beauty	BEAUTY	NaN
5	Secrets of beauty, youth and health	BEAUTY	NaN
6	Recipes and tips for losing weight	BEAUTY	NaN
7	Lady adviser (beauty, health)	BEAUTY	NaN

10,000 rows | Truncated data

Categories with the Most Free Apps:

```
%sql SELECT
  Category,
  COUNT(*) AS FreeAppCount
FROM apps
WHERE Type = 'Free'
GROUP BY Category
ORDER BY FreeAppCount DESC
```

Table

	Category	FreeAppCount
1	FAMILY	1780
2	GAME	1061
3	TOOLS	765
4	BUSINESS	446
5	PRODUCTIVITY	396
6	LIFESTYLE	363

7	SPORTS	360

33 rows

Categories with the Most Paid Apps:

```
%sql SELECT
  Category,
  COUNT(*) AS PaidAppCount
FROM apps
WHERE Type = 'Paid'
GROUP BY Category
ORDER BY PaidAppCount DESC
```

Table

	Category	PaidAppCount
1	FAMILY	191
2	MEDICAL	109
3	PERSONALIZATION	83
4	GAME	83
5	TOOLS	78
6	BOOKS_AND_REFERENCE	28
7	PRODUCTIVITY	28

30 rows

Average Price of Paid Apps by Category:

Table

	Category	AvgPrice
1	EVENTS	109
2	SPORTS	3.3333333333333335
3	WEATHER	3.125
4	VIDEO_PLAYERS	1.75
5	AUTO_AND_VEHICLES	3.6666666666666665
6	PARENTING	4
7	ENTERTAINMENT	3

30 rows