Import library

```
import pyspark
from pyspark.sql import *
from pyspark.sql.functions import *
from pyspark.sql.types import *
```

create dataframe

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

```
df_google_app_raw.count()
```

```
10841
```

```
df_google_app_raw.show(2)
```

```
+--------------------+--------------+------+-------+----+--------+----+-----+--------------+--------------------+----------------+-----------+------------+
|                 App|      Category|Rating|Reviews|Size|Installs|Type|Price|Content Rating|              Genres|    Last Updated|Current Ver| 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|
| Coloring book moana|ART_AND_DESIGN|   3.9|    967| 14M|500,000+|Free|    0|      Everyone|Art & Design;Pret...|January 15, 2018|      2.0.0|4.0.3 and up|
+--------------------+--------------+------+-------+----+--------+----+-----+--------------+--------------------+----------------+-----------+------------+
only showing top 2 rows
```

check schema

```
df_google_app_raw.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)
```

Data cleaning Steps

Drop extra columns which is not required for our Analysis and storing in a new dataFrame

```
df_google_app_1 = df_google_app_raw.drop("Size","Content Rating","Last Updated","Current Ver","Android Ver")
```

Validate all the required column and how values are stored

```
df_google_app_1.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
```

Check Data type of every column

```
df_google_app_1.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)
```

Changing Datatype of columns and removing extra symbols/characters (+,$) from the value and

```
df_google_app_2 = df_google_app_1.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()))
```

Validate all the Datatypes of every columns.

```
df_google_app_2.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)
```

Validate values are correctly showing in column, extra symbols/special characters (+,$)are remov

```
df_google_app_2.show(10)
```

```
+--------------------+--------------+------+-------+--------+----+-----+--------------------+
|                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...|
|Paper flowers ins...|ART_AND_DESIGN|   4.4|    167|   50000|Free|    0|        Art & Design|
|Smoke Effect Phot...|ART_AND_DESIGN|   3.8|    178|   50000|Free|    0|        Art & Design|
|    Infinite Painter|ART_AND_DESIGN|   4.1|  36815| 1000000|Free|    0|        Art & Design|
|Garden Coloring Book|ART_AND_DESIGN|   4.4|  13791| 1000000|Free|    0|        Art & Design|
|Kids Paint Free -...|ART_AND_DESIGN|   4.7|    121|   10000|Free|    0|Art & Design;Crea...|
+--------------------+--------------+------+-------+--------+----+-----+--------------------+
only showing top 10 rows
```

Creating views of our final DataFrame, using this view we can perform SQL operation ()sql query

```
df_google_app_2.createOrReplaceTempView("apps")
```

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

Table

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

10,000 rows | Truncated data

1. Top 10 reviews given to the app

```
%sql
select  App, sum(Reviews) from apps
group by 1
order by 2 desc
LIMIT 10
```

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

10 rows

2. Top 10 installs apps and distribution of type (Free/paid)

```
%sql
select  App,Type, sum(Installs) from apps
group by 1,2
order by 3 desc
LIMIT 10
```

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

10 rows

3. Category wise distribution of installed app

```
%sql
select   Category,sum(Installs) from apps
group by 1
order by 2 desc
--LIMIT 10
```

Table

	Category	sum(Installs)
1	GAME	35086024415
2	COMMUNICATION	32647276251
3	PRODUCTIVITY	14176091369
4	SOCIAL	14069867902
5	TOOLS	11452771915
6	FAMILY	10258263505
7	PHOTOGRAPHY	10088247655

34 rows

3.2 . Which Category's app installed most name of the app which is under this most popular cate

```
%sql
select  App, Category,sum(Installs) from apps
group by 1,2
order by 3 desc
--LIMIT 10
```

Table

	App	Category	sum(Installs)
1	Subway Surfers	GAME	6000000000
2	Hangouts	COMMUNICATION	4000000000
3	Google News	NEWS_AND_MAGAZINES	4000000000
4	Instagram	SOCIAL	4000000000
5	Google Drive	PRODUCTIVITY	4000000000
6	Google Photos	PHOTOGRAPHY	4000000000
7	Gmail	COMMUNICATION	3000000000

9,745 rows

4. Top paid app

```
%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

4.2 Most installed Paid App

```
%sql
select * from apps
where Type="Paid"
order by Installs desc
```

Table

	App	Category	Rating	Reviews	Installs	Type	Price	Genr
1	Minecraft	FAMILY	4.5	2376564	10000000	Paid	6	Arcad
2	Hitman Sniper	GAME	4.6	408292	10000000	Paid	0	Actio
3	Minecraft	FAMILY	4.5	2375336	10000000	Paid	6	Arcad
4	Card Wars - Adventure Time	FAMILY	4.3	129603	1000000	Paid	2	Card;
5	Facetune - For Free	PHOTOGRAPHY	4.4	49553	1000000	Paid	5	Photo
6	Facetune - For Free	PHOTOGRAPHY	4.4	49553	1000000	Paid	5	Photo
7	Facetune - For Free	PHOTOGRAPHY	4.4	49553	1000000	Paid	5	Photo

800 rows

5. Top paid rating app

Table

	App	sum(Rating)
1	AF-STROKE	NaN
2	Servidor Privado CR y CoC - Royale Servers PRO	NaN
3	Language Therapy: Aphasia	NaN
4	Eu Sou Rico	NaN
5	Dz kayas	NaN
6	FJ Toolkit	NaN
7	Be the Expert in Phlebotomy - Professional Nursing	NaN

756 rows