

In [1]: *#Importing Required Libraries*

```
import pandas as pd
import numpy as np
import math
import sys
import findspark
findspark.init()
```

In [2]:

```
from pyspark import StorageLevel
from pyspark.sql import SparkSession
from pyspark.sql.functions import isnan, lit, col, explode, initcap, regexp_replace
from pyspark.sql.types import StructType, StructField, StringType, IntegerType, FloatType
from pyspark.sql.dataframe import DataFrame
```

In [3]:

```
spark = SparkSession \
    .builder \
    .appName("PlaysoreAppsPySpark") \
    .getOrCreate()
```

```
In [4]: # Reading the raw dataset from Cloud Storage bucket in to the python dataframe psapps
psapps_df = spark.read.csv("gs://ca675-assignment2-bucket/Raw_Data/Google-Playstore")
# Checking the first 10 rows of the dataframe
psapps_df.show(10)
```

```
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
|          App Name|          App Id|          Category|Rating|Rating Cou
nt|Installs|Minimum Installs|Maximum Installs|Free|Price|Currency|Size|Minimu
m Android|          Developer Id|  Developer Website|  Developer Email|
Released|Last Updated|Content Rating|  Privacy Policy|Ad Supported|In App
Purchases|Editors Choice|  Scraped Time|
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
|          Gakondo| com.ishakwe.gakondo|          Adventure|  0.0|
0|  10+|  10|  15|true|  0.0|  USD| 10M|  7.
1 and up|Jean Confident Ir...|https://beniyizib...|jean21101999@gmai...|Feb 2
6, 2020|Feb 26, 2020|  Everyone|https://beniyizib...|  false|
false|  false|2021-06-15 20:19:35|
| Ampere Battery Info|com.webserveis.ba...|  Tools|  4.4|
64|  5,000+|  5000|  7662|true|  0.0|  USD|2.9M|
5.0 and up|  Webserveis|https://webservei...|webserveis@gmail.com|May
21, 2020|May 06, 2021|  Everyone|https://dev4phone...|  true|
false|  false|2021-06-15 20:19:35|
|          Vibook|com.doantiepvien.crm|  Productivity|  0.0|
0|  50+|  50|  58|true|  0.0|  USD|3.7M|  4.0.
3 and up|  Cabin Crew|  null|vnacrewit@gmail.com|Aug
9, 2019|Aug 19, 2019|  Everyone|https://www.vietn...|  false|
false|  false|2021-06-15 20:19:35|
|Smart City Trichy...|cst.stJoseph.ug17...|  Communication|  5.0|
5|  10+|  10|  19|true|  0.0|  USD|1.8M|  4.0.
3 and up|Climate Smart Tech2|http://www.climat...|climatesmarttech2...|Sep 1
0, 2018|Oct 13, 2018|  Everyone|  null|  true|
false|  false|2021-06-15 20:19:35|
|          GROW.me|com.horodyski.grower|  Tools|  0.0|
0|  100+|  100|  478|true|  0.0|  USD|6.2M|  4.
1 and up|Rafal Milek-Horod...|http://www.horody...|rmilekhorodyski@g...|Feb 2
1, 2020|Nov 12, 2018|  Everyone|http://www.horody...|  false|
false|  false|2021-06-15 20:19:35|
|          IMOCCI|  com.imocci|  Social|  0.0|
0|  50+|  50|  89|true|  0.0|  USD| 46M|  6.
0 and up|  Imocci GmbH|http://www.imocci...|  info@imocci.com|Dec 2
4, 2018|Dec 20, 2019|  Teen|https://www.imocc...|  false|
true|  false|2021-06-15 20:19:35|
|unlimited 4G data...|getfreedata.super...|Libraries & Demo|  4.5|
12|  1,000+|  1000|  2567|true|  0.0|  USD|2.5M|
```

```

4.1 and up|android developer779|          null|aitomgharfatimezz...|Sep
23, 2019|Sep 27, 2019|          Everyone|https://sites.goo...|          true|
false|          false|2021-06-15 20:19:35|
|The Everyday Cale...|com.mozaix.simone...|          Lifestyle|          2.0|
39|          500+|          500|          702|true|          0.0|          USD|16M|
5.0 and up|          Mozaix LLC|          null|elementuser03@gma...|Jun
21, 2019|Jun 21, 2019|          Everyone|https://www.freep...|          false|
false|          false|2021-06-15 20:19:35|
|          WhatsOpen|          com.whatsopen.app|          Communication|          0.0|
0|          10+|          10|          18|true|          0.0|          USD|1.3M|          4.
4 and up|Yilver Molina Hur...|http://yilvermoli...|yilver.mh1996@gma...|
null|Dec 07, 2018|          Teen|http://elcafedela...|          false|
false|          false|2021-06-15 20:19:35|
|Neon 3d Iron Tech...|com.ikeyboard.the...|          Personalization|          4.7|          8
20|          50,000+|          50000|          62433|true|          0.0|          USD|3.5M|
4.1 and up|Free 2021 Themes ...|https://trendytem...|trendyteme.888@gm...|Sep
22, 2019|Oct 07, 2020|          Everyone|http://bit.ly/Emo...|          true|
false|          false|2021-06-15 20:19:35|
+-----+-----+-----+-----+-----+-----+-----+-----+-----+
-+-----+-----+-----+-----+-----+-----+-----+-----+-----+
-----+-----+-----+-----+-----+-----+-----+-----+-----+
-----+-----+-----+-----+-----+-----+-----+-----+-----+
-----+-----+-----+-----+-----+-----+-----+-----+-----+
only showing top 10 rows

```

In [5]: *#Checking the structure of the numeric feature*
psapps_df.describe()

Out[5]: DataFrame[summary: string, App Name: string, App Id: string, Category: string, Rating: string, Rating Count: string, Installs: string, Minimum Installs: string, Maximum Installs: string, Price: string, Currency: string, Size: string, Minimum Android: string, Developer Id: string, Developer Website: string, Developer Email: string, Released: string, Last Updated: string, Content Rating: string, Privacy Policy: string, Scraped Time: string]

```
In [6]: #Checking the total count null values in each feature of the dataframe psapps_df
psapps_null_df=psapps_df.select([count(when(col(c).isNull(),c)) for c in psapps_df.columns])
psapps_null_df.show()
```

```
[Stage 6:=====> (5 + 1) / 6]
```

[illegible]

The diagram consists of a grid of points connected by dashed lines. The points are labeled with various symbols and numbers. The labels are as follows:

- Top row: 0|, 22883|, 107|, 0|, 0|, 135|, 6530|, 760835|, 71053|, 420953|, 196|
- Second row: 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|
- Third row: 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|
- Fourth row: 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|
- Fifth row: 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|
- Sixth row: 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|
- Seventh row: 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|
- Eighth row: 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|
- Ninth row: 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|
- Tenth row: 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|
- Eleventh row: 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|
- Twelfth row: 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|
- Thirteenth row: 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|
- Fourteenth row: 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|
- Fifteenth row: 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|
- Sixteenth row: 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|
- Seventeenth row: 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|
- Eighteenth row: 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|
- Nineteenth row: 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|
- Twentieth row: 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|
- Twenty-first row: 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|
- Twenty-second row: 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|
- Twenty-third row: 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|
- Twenty-fourth row: 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|
- Twenty-fifth row: 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|
- Twenty-sixth row: 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|
- Twenty-seventh row: 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|
- Twenty-eighth row: 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|
- Twenty-ninth row: 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|
- Thirtieth row: 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|
- Thirty-first row: 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|
- Thirty-second row: 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|
- Thirty-third row: 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|
- Thirty-fourth row: 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|
- Thirty-fifth row: 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|
- Thirty-sixth row: 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|
- Thirty-seventh row: 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|
- Thirty-eighth row: 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|
- Thirty-ninth row: 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|
- Fortieth row: 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|
- Forty-first row: 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|
- Forty-second row: 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|
- Forty-third row: 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|
- Forty-fourth row: 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|
- Forty-fifth row: 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|
- Forty-sixth row: 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|
- Forty-seventh row: 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|
- Forty-eighth row: 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|
- Forty-ninth row: 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|
- Fiftieth row: 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|, 0|

```
In [7]: #Cleaning-Step1
#Removing the unwanted features abd storing the resultant in new dataframe 'psapps'
psapps_selected_df=psapps_df.drop("App Id","Minimum Installs","Currency","Minimum
# Checking top 10 values after removal of unwanted features
psapps_selected_df.head(10)
```

```
Out[7]: [Row(App Name='Gakondo', Category='Adventure', Rating=0.0, Rating Count=0, Inst
alls='10+', Maximum Installs=15, Free=True, Price=0.0, Size='10M', Released='Fe
b 26, 2020', Content Rating='Everyone', Ad Supported=False, In App Purchases=Fa
lse),
Row(App Name='Ampere Battery Info', Category='Tools', Rating=4.4, Rating Count
=64, Installs='5,000+', Maximum Installs=7662, Free=True, Price=0.0, Size='2.9
M', Released='May 21, 2020', Content Rating='Everyone', Ad Supported=True, In A
pp Purchases=False),
Row(App Name='Vibook', Category='Productivity', Rating=0.0, Rating Count=0, In
stalls='50+', Maximum Installs=58, Free=True, Price=0.0, Size='3.7M', Released
='Aug 9, 2019', Content Rating='Everyone', Ad Supported=False, In App Purchases
=False),
Row(App Name='Smart City Trichy Public Service Vehicles 17UCS548', Category='C
ommunication', Rating=5.0, Rating Count=5, Installs='10+', Maximum Installs=19,
Free=True, Price=0.0, Size='1.8M', Released='Sep 10, 2018', Content Rating='Eve
ryone', Ad Supported=True, In App Purchases=False),
Row(App Name='GROW.me', Category='Tools', Rating=0.0, Rating Count=0, Installs
='100+', Maximum Installs=478, Free=True, Price=0.0, Size='6.2M', Released='Feb
21, 2020', Content Rating='Everyone', Ad Supported=False, In App Purchases=Fals
e),
Row(App Name='IMOCCI', Category='Social', Rating=0.0, Rating Count=0, Installs
='50+', Maximum Installs=89, Free=True, Price=0.0, Size='46M', Released='Dec 2
4, 2018', Content Rating='Teen', Ad Supported=False, In App Purchases=True),
Row(App Name='unlimited 4G data prank free app', Category='Libraries & Demo',
Rating=4.5, Rating Count=12, Installs='1,000+', Maximum Installs=2567, Free=Tru
e, Price=0.0, Size='2.5M', Released='Sep 23, 2019', Content Rating='Everyone',
Ad Supported=True, In App Purchases=False),
Row(App Name='The Everyday Calendar', Category='Lifestyle', Rating=2.0, Rating
Count=39, Installs='500+', Maximum Installs=702, Free=True, Price=0.0, Size='16
M', Released='Jun 21, 2019', Content Rating='Everyone', Ad Supported=False, In
App Purchases=False),
Row(App Name='WhatsOpen', Category='Communication', Rating=0.0, Rating Count=
0, Installs='10+', Maximum Installs=18, Free=True, Price=0.0, Size='1.3M', Rele
ased=None, Content Rating='Teen', Ad Supported=False, In App Purchases=False),
Row(App Name='Neon 3d Iron Tech Keyboard Theme', Category='Personalization', R
ating=4.7, Rating Count=820, Installs='50,000+', Maximum Installs=62433, Free=T
rue, Price=0.0, Size='3.5M', Released='Sep 22, 2019', Content Rating='Everyon
e', Ad Supported=True, In App Purchases=False)]
```

```
In [8]: #Cleaning-Step2
# Replacing the space in column names with '_' and storing the new fieldnames in
df = [col(column).alias(column.replace(' ', '_')) for column in psapps_selected_c
psapps_selected_df.select(*df).show()
df1=psapps_selected_df.select(*df)
```

```
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
+-----+
|
|           App_Name|           Category|Rating|Rating_Count|Insta
lls|Maximum_Installs|Free|Price|Size|       Released|Content_Rating|Ad_Supported|I
n_App_Purchases|
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
+-----+
|
|           Gakondo|           Adventure|  0.0|           0|
10+|           15|true|  0.0| 10M|Feb 26, 2020|           Everyone|           false|
false|
|
| Ampere Battery Info|           Tools|  4.4|           64|  5,0
00+|       7662|true|  0.0|2.9M|May 21, 2020|           Everyone|           true|
false|
|
|           Vibook|           Productivity|  0.0|           0|
50+|           58|true|  0.0|3.7M| Aug 9, 2019|           Everyone|           false|
false|
|
| Smart City Trichy...|           Communication|  5.0|           5|
10+|           19|true|  0.0|1.8M|Sep 10, 2018|           Everyone|           true|
false|
|
|           GROW.me|           Tools|  0.0|           0|  1
00+|           478|true|  0.0|6.2M|Feb 21, 2020|           Everyone|           false|
false|
|
|           IMOCCI|           Social|  0.0|           0|
50+|           89|true|  0.0| 46M|Dec 24, 2018|           Teen|           false|
true|
|
| unlimited 4G data...| Libraries & Demo|  4.5|           12|  1,0
00+|       2567|true|  0.0|2.5M|Sep 23, 2019|           Everyone|           true|
false|
|
| The Everyday Cale...|           Lifestyle|  2.0|           39|  5
00+|           702|true|  0.0| 16M|Jun 21, 2019|           Everyone|           false|
false|
|
|           WhatsOpen|           Communication|  0.0|           0|
10+|           18|true|  0.0|1.3M|           null|           Teen|           false|
false|
|
| Neon 3d Iron Tech...|           Personalization|  4.7|           820| 50,0
00+|      62433|true|  0.0|3.5M|Sep 22, 2019|           Everyone|           true|
false|
|
| Dodge The Cars!|           Racing|  4.9|           55|  1
00+|           329|true|  0.0| 51M|Jul 30, 2020|           Everyone|           false|
false|
|
|           Parents|Maps & Navigation|  0.0|           0|  1
00+|           330|true|  0.0|2.7M|Jan 10, 2018|           Everyone|           false|
false|
|桃園機場捷運時刻表 - 捷運轉乘路...| Travel & Local|  3.9|           118| 10,000+
|       37763|true|  0.0|7.6M| Apr 3, 2018|           Everyone|           true|
false|
|
|           be.MOBILISED|Maps & Navigation|  0.0|           0|  1
```

```

00+|          274|true|  0.0| 15M| Feb 9, 2020|          Everyone|          false|
false|
|          Caliway Conductor|  Travel & Local|  3.7|          1572| 10,0
00+|          42856|true|  0.0| 29M| Sep 5, 2018|          Everyone|          false|
false|
|          Readymade Grocery...|  Food & Drink|  4.2|          16| 1,0
00+|          4313|true|  0.0| 12M| Apr 5, 2020|          Everyone|          false|
false|
|          OTENTIK Discovery FR|  Travel & Local|  3.4|          5| 5
00+|          949|true|  0.0|2.9M|Nov 28, 2016|          Everyone|          false|
false|
|          All in one shoppi...|          Tools|  5.0|          6|
50+|          62|true|  0.0|2.0M|Apr 24, 2019|          Everyone|          false|
false|
|          REDMOND Robot|          Lifestyle|  3.7|          328| 10,0
00+|          31235|true|  0.0| 70M| Jul 1, 2020|          Everyone|          false|
false|
|          Contemporary Love...|Books & Reference|  0.0|          0|
1+|          2|true|  0.0| 11M|Dec 26, 2020|          Mature 17+|          true|
false|
+-----+-----+-----+-----+-----+-----+-----+-----+-----+
---+-----+-----+-----+-----+-----+-----+-----+-----+-----+
-----+
only showing top 20 rows

```

In [9]: *# Checking data structure of the dataframe df1*
df1.printSchema()

```

root
|-- App_Name: string (nullable = true)
|-- Category: string (nullable = true)
|-- Rating: double (nullable = true)
|-- Rating_Count: integer (nullable = true)
|-- Installs: string (nullable = true)
|-- Maximum_Installs: long (nullable = true)
|-- Free: boolean (nullable = true)
|-- Price: double (nullable = true)
|-- Size: string (nullable = true)
|-- Released: string (nullable = true)
|-- Content_Rating: string (nullable = true)
|-- Ad_Supported: boolean (nullable = true)
|-- In_App_Purchases: boolean (nullable = true)

```



```
In [10]: #Cleaning-Step3
# Changing the datatype of Rating column to Double, Installs to Integer, Price to Double
df1 = df1 \
    .withColumn("Rating", col("Rating").cast(DoubleType())) \
    .withColumn("Rating_Count", col("Rating_Count").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(DoubleType())) \
    .withColumn("Released", to_date('Released', 'MMM d, yyyy'))
```

```
In [11]: #Checking data structure of the dataframe df1
df1.printSchema()
```

```
root
|-- App_Name: string (nullable = true)
|-- Category: string (nullable = true)
|-- Rating: double (nullable = true)
|-- Rating_Count: integer (nullable = true)
|-- Installs: integer (nullable = true)
|-- Maximum_Installs: long (nullable = true)
|-- Free: boolean (nullable = true)
|-- Price: double (nullable = true)
|-- Size: string (nullable = true)
|-- Released: date (nullable = true)
|-- Content_Rating: string (nullable = true)
|-- Ad_Supported: boolean (nullable = true)
|-- In_App_Purchases: boolean (nullable = true)
```

```
In [23]: #Cleaning-Step4
# Dropping rows having null values in Release column and storing the result in df2
df2=df1.dropna(subset=["Released"])
df2.show(10)
```

```
+-----+-----+-----+-----+-----+-----+
+-----+-----+-----+-----+-----+-----+
|          App_Name|          Category|Rating|Rating_Count|Installs|Maximum_Ins
talls|Free|Price|Size|   Released|Content_Rating|Ad_Supported|In_App_Purchases|
+-----+-----+-----+-----+-----+-----+
+-----+-----+-----+-----+-----+-----+
|          Gakondo|          Adventure|  0.0|          0|      10|
15|true|  0.0| 10M|2020-02-26|          Everyone|          false|          false|
| Ampere Battery Info|          Tools|  4.4|          64|     5000|
7662|true|  0.0|2.9M|2020-05-21|          Everyone|          true|          false|
|          Vibook|          Productivity|  0.0|          0|      50|
58|true|  0.0|3.7M|2019-08-09|          Everyone|          false|          false|
|Smart City Trichy...|          Communication|  5.0|          5|      10|
19|true|  0.0|1.8M|2018-09-10|          Everyone|          true|          false|
|          GROW.me|          Tools|  0.0|          0|     100|
478|true|  0.0|6.2M|2020-02-21|          Everyone|          false|          false|
|          IMOCCI|          Social|  0.0|          0|      50|
89|true|  0.0| 46M|2018-12-24|          Teen|          false|          true|
|unlimited 4G data...|Libraries & Demo|  4.5|          12|     1000|
2567|true|  0.0|2.5M|2019-09-23|          Everyone|          true|          false|
|The Everyday Cale...|          Lifestyle|  2.0|          39|      500|
702|true|  0.0| 16M|2019-06-21|          Everyone|          false|          false|
|Neon 3d Iron Tech...|          Personalization|  4.7|          820|    50000|
62433|true|  0.0|3.5M|2019-09-22|          Everyone|          true|          false|
|          Dodge The Cars!|          Racing|  4.9|          55|      100|
329|true|  0.0| 51M|2020-07-30|          Everyone|          false|          false|
+-----+-----+-----+-----+-----+-----+
+-----+-----+-----+-----+-----+-----+
only showing top 10 rows
```

```
In [24]: #Checking the total count null values in eah feature of the dataframe df2
df2_null1 = df2.select([count(when(col(c).isNull(),c)) for c in df2.columns])
df2_null1.show()
```

```
[Stage 31:=====> (5 + 1) /
6]
```

```
|count(CASE WHEN (App_Name IS NULL) THEN App_Name END)|count(CASE WHEN (Category IS NULL) THEN Category END)|count(CASE WHEN (Rating IS NULL) THEN Rating END)|count(CASE WHEN (Rating_Count IS NULL) THEN Rating_Count END)|count(CASE WHEN (Installs IS NULL) THEN Installs END)|count(CASE WHEN (Maximum_Installs IS NULL) THEN Maximum_Installs END)|count(CASE WHEN (Free IS NULL) THEN Free END)|count(CASE WHEN (Price IS NULL) THEN Price END)|count(CASE WHEN (Size IS NULL) THEN Size END)|count(CASE WHEN (Released IS NULL) THEN Released END)|co
```

```
In [14]: #Cleaning-Step5
# Dropping rows having null values in Installs column and storing the result in df3
df3=df2.drop(subset=["Installs"])
df3.show(10)
```

```
+-----+-----+-----+-----+-----+-----+
+-----+-----+-----+-----+-----+-----+
|          App_Name|          Category|Rating|Rating_Count|Installs|Maximum_Installs|Free|Price|Size|  Released|Content_Rating|Ad_Supported|In_App_Purchases|
+-----+-----+-----+-----+-----+-----+
+-----+-----+-----+-----+-----+-----+
|          Gakondo|          Adventure|  0.0|          0|    10|          false|  false|
15|true|  0.0| 10M|2020-02-26|          Everyone|          false|  false|
| Ampere Battery Info|          Tools|  4.4|          64|   5000|          false|  false|
7662|true|  0.0|2.9M|2020-05-21|          Everyone|          true|  false|
|          Vibook|          Productivity|  0.0|          0|    50|          false|  false|
58|true|  0.0|3.7M|2019-08-09|          Everyone|          false|  false|
|Smart City Trichy...|          Communication|  5.0|          5|    10|          false|  false|
19|true|  0.0|1.8M|2018-09-10|          Everyone|          true|  false|
|          GROW.me|          Tools|  0.0|          0|   100|          false|  false|
478|true|  0.0|6.2M|2020-02-21|          Everyone|          false|  false|
|          IMOCCI|          Social|  0.0|          0|    50|          false|  true|
89|true|  0.0| 46M|2018-12-24|          Teen|          false|  true|
|unlimited 4G data...|Libraries & Demo|  4.5|          12|   1000|          false|  false|
2567|true|  0.0|2.5M|2019-09-23|          Everyone|          true|  false|
|The Everyday Cale...|          Lifestyle|  2.0|          39|    500|          false|  false|
702|true|  0.0| 16M|2019-06-21|          Everyone|          false|  false|
|Neon 3d Iron Tech...|          Personalization|  4.7|          820|  50000|          false|  false|
62433|true|  0.0|3.5M|2019-09-22|          Everyone|          true|  false|
|          Dodge The Cars!|          Racing|  4.9|          55|    100|          false|  false|
329|true|  0.0| 51M|2020-07-30|          Everyone|          false|  false|
+-----+-----+-----+-----+-----+-----+
+-----+-----+-----+-----+-----+-----+
only showing top 10 rows
```



```
In [16]: #Cleaning-Step6
# Modifying the values of Size Column in dataframe df3
df3 = df3 \
    .withColumn("Size", regexp_replace(col("Size"), 'k', 'e+3')) \
    .withColumn("Size", regexp_replace(col("Size"), 'M', 'e+6')) \
    .withColumn("Size", regexp_replace(col("Size"), 'G', 'e+9')) \
    .withColumn("Size", regexp_replace(col("Size"), "Varies with Device", 'nan')) \
    .withColumn("Size", regexp_replace(col("Size"), "[+]", "")) \
    .withColumn("Size", regexp_replace(col("Size"), "[^0-9]", "")) \
    .withColumn("Size", col("Size").cast(IntegerType()))
```

```
In [17]: #Checking the total count nan values in Size feature of the dataframe df3
df3.select([count(when(isnan('Size'), True))]).show()
```

```
[Stage 19:=====> (5 + 1) / 6]
```

```
+-----+
|count(CASE WHEN isnan(Size) THEN true END)|
+-----+
|                                0|
+-----+
```

```
In [18]: # Checking the data structure of dataframe df3
df3.printSchema()
```

```
root
|-- App_Name: string (nullable = true)
|-- Category: string (nullable = true)
|-- Rating: double (nullable = true)
|-- Rating_Count: integer (nullable = true)
|-- Installs: integer (nullable = true)
|-- Maximum_Installs: long (nullable = true)
|-- Free: boolean (nullable = true)
|-- Price: double (nullable = true)
|-- Size: integer (nullable = true)
|-- Released: date (nullable = true)
|-- Content_Rating: string (nullable = true)
|-- Ad_Supported: boolean (nullable = true)
|-- In_App_Purchases: boolean (nullable = true)
```

```
In [25]: #Cleaning-Step7
# Dropping rows having null values in columns Rating and Rating_Count
df4=df3.na.drop(subset=["Rating", "Rating_Count"])
df4.show(10)
```

```
+-----+-----+-----+-----+-----+-----+
+-----+-----+-----+-----+-----+-----+
|          App_Name|          Category|Rating|Rating_Count|Installs|Maximum_Ins
talls|Free|Price|Size|   Released|Content_Rating|Ad_Supported|In_App_Purchases|
+-----+-----+-----+-----+-----+-----+
+-----+-----+-----+-----+-----+-----+
|          Gakondo|          Adventure|  0.0|          0|    10|
15|true|  0.0| 106|2020-02-26|    Everyone|    false|    false|
| Ampere Battery Info|          Tools|  4.4|         64|   5000|
7662|true|  0.0| 296|2020-05-21|    Everyone|    true|    false|
|          Vibook|    Productivity|  0.0|          0|    50|
58|true|  0.0| 376|2019-08-09|    Everyone|    false|    false|
|Smart City Trichy...|    Communication|  5.0|          5|    10|
19|true|  0.0| 186|2018-09-10|    Everyone|    true|    false|
|          GROW.me|          Tools|  0.0|          0|   100|
478|true|  0.0| 626|2020-02-21|    Everyone|    false|    false|
|          IMOCCI|          Social|  0.0|          0|    50|
89|true|  0.0| 466|2018-12-24|    Teen|    false|    true|
|unlimited 4G data...|Libraries & Demo|  4.5|         12|   1000|
2567|true|  0.0| 256|2019-09-23|    Everyone|    true|    false|
|The Everyday Cale...|    Lifestyle|  2.0|         39|    500|
702|true|  0.0| 166|2019-06-21|    Everyone|    false|    false|
|Neon 3d Iron Tech...|    Personalization|  4.7|        820|  50000|
62433|true|  0.0| 356|2019-09-22|    Everyone|    true|    false|
|    Dodge The Cars!|          Racing|  4.9|         55|    100|
329|true|  0.0| 516|2020-07-30|    Everyone|    false|    false|
+-----+-----+-----+-----+-----+-----+
+-----+-----+-----+-----+-----+-----+
only showing top 10 rows
```

```
In [28]: #Checking the total count null values in eah feature of the dataframe df4
df4_null1 = df4.select([count(when(col(c).isNull(),c)) for c in df4.columns])
df4_null1.show()
```

```
[Stage 37:=====> (5 + 1) / 6]
```

[illegible]


```
In [29]: #Cleaning-Step8
# Dropping rows having null values in column Size
df5=df4.na.drop(subset=["Size"])
df5.show(10)
```

```
+-----+-----+-----+-----+-----+-----+
+-----+-----+-----+-----+-----+-----+
|          App_Name|          Category|Rating|Rating_Count|Installs|Maximum_Ins
talls|Free|Price|Size|   Released|Content_Rating|Ad_Supported|In_App_Purchases|
+-----+-----+-----+-----+-----+-----+
+-----+-----+-----+-----+-----+-----+
|          Gakondo|          Adventure|  0.0|          0|      10|
15|true|  0.0| 106|2020-02-26|      Everyone|      false|      false|
| Ampere Battery Info|          Tools|  4.4|          64|     5000|
7662|true|  0.0| 296|2020-05-21|      Everyone|      true|      false|
|          Vibook|          Productivity|  0.0|          0|      50|
58|true|  0.0| 376|2019-08-09|      Everyone|      false|      false|
|Smart City Trichy...|      Communication|  5.0|          5|      10|
19|true|  0.0| 186|2018-09-10|      Everyone|      true|      false|
|          GROW.me|          Tools|  0.0|          0|     100|
478|true|  0.0| 626|2020-02-21|      Everyone|      false|      false|
|          IMOCCI|          Social|  0.0|          0|      50|
89|true|  0.0| 466|2018-12-24|          Teen|      false|      true|
|unlimited 4G data...|Libraries & Demo|  4.5|         12|     1000|
2567|true|  0.0| 256|2019-09-23|      Everyone|      true|      false|
|The Everyday Cale...|          Lifestyle|  2.0|         39|      500|
702|true|  0.0| 166|2019-06-21|      Everyone|      false|      false|
|Neon 3d Iron Tech...|      Personalization|  4.7|        820|    50000|
62433|true|  0.0| 356|2019-09-22|      Everyone|      true|      false|
|   Dodge The Cars!|          Racing|  4.9|         55|      100|
329|true|  0.0| 516|2020-07-30|      Everyone|      false|      false|
+-----+-----+-----+-----+-----+-----+
+-----+-----+-----+-----+-----+-----+
only showing top 10 rows
```


In [31]: `df5.count()`

Out[31]: 2190445