EDA & Data Preprocessing on Google App Store Rating Dataset.

Domain: Mobile device apps

Context: The Play Store apps data has enormous potential to drive app-making businesses to success. However, many apps are being developed every single day and only a few of them become profitable. It is important for developers to be able to predict the success of their app and incorporate features which makes an app successful. Before any such predictive-study can be done, it is necessary to do EDA and data-preprocessing on the apps data available for google app store applications. From the collected apps data and user ratings from the app stores, let's try to extract insightful information.

Objective: The Goal is to explore the data and pre-process it for future use in any predictive analytics study.

Data set Information: Web scraped data of 10k Play Store apps for analyzing the Android market. Each app (row) has values for category, rating, size, and more.

SUBMITTED BY:

NITHISHWAR

1. Import required libraries and read the dataset.

In [134]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

In [135]:
```
df = pd.read_csv(r"C:\Users\Nithish\Desktop\GL\Python\week 4\week 4 Graded project\Datasets\Apps_data+(1).csv")
df
```

Out[135]:

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
0	Photo Editor & Candy Camera & Grid & ScrapBook	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
1	Coloring book moana	ART_AND_DESIGN	3.9	967	14M	500,000+	Free	0	Everyone	Art & Design;Pretend Play	January 15, 2018	2.0.0	4.0.3 and up
2	U Launcher Lite – FREE Live Cool Themes, Hide ...	ART_AND_DESIGN	4.7	87510	8.7M	5,000,000+	Free	0	Everyone	Art & Design	August 1, 2018	1.2.4	4.0.3 and up
3	Sketch - Draw & Paint	ART_AND_DESIGN	4.5	215644	25M	50,000,000+	Free	0	Teen	Art & Design	June 8, 2018	Varies with device	4.2 and up
4	Pixel Draw - Number Art Coloring Book	ART_AND_DESIGN	4.3	967	2.8M	100,000+	Free	0	Everyone	Art & Design;Creativity	June 20, 2018	1.1	4.4 and up
...	...	...	...	...	...	...	...	...	...	...	...	...	...
10836	Sya9a Maroc - FR	FAMILY	4.5	38	53M	5,000+	Free	0	Everyone	Education	July 25, 2017	1.48	4.1 and up
10837	Fr. Mike Schmitz Audio Teachings	FAMILY	5.0	4	3.6M	100+	Free	0	Everyone	Education	July 6, 2018	1.0	4.1 and up
10838	Parkinson Exercices FR	MEDICAL	NaN	3	9.5M	1,000+	Free	0	Everyone	Medical	January 20, 2017	1.0	2.2 and up
10839	The SCP Foundation DB fr nn5n	BOOKS_AND_REFERENCE	4.5	114	Varies with device	1,000+	Free	0	Mature 17+	Books & Reference	January 19, 2015	Varies with device	Varies with device
10840	iHoroscope - 2018 Daily Horoscope & Astrology	LIFESTYLE	4.5	398307	19M	10,000,000+	Free	0	Everyone	Lifestyle	July 25, 2018	Varies with device	Varies with device

10841 rows × 13 columns

2. Check the first few samples, shape, info of the data and try to familiarize yourself with different features.

In [136]:
```
1 df.head(10)  #---First few samples
```

Out[136]:

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
0	Photo Editor & Candy Camera & Grid & ScrapBook	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
1	Coloring book moana	ART_AND_DESIGN	3.9	967	14M	500,000+	Free	0	Everyone	Art & Design;Pretend Play	January 15, 2018	2.0.0	4.0.3 and up
2	U Launcher Lite – FREE Live Cool Themes, Hide ...	ART_AND_DESIGN	4.7	87510	8.7M	5,000,000+	Free	0	Everyone	Art & Design	August 1, 2018	1.2.4	4.0.3 and up
3	Sketch - Draw & Paint	ART_AND_DESIGN	4.5	215644	25M	50,000,000+	Free	0	Teen	Art & Design	June 8, 2018	Varies with device	4.2 and up
4	Pixel Draw - Number Art Coloring Book	ART_AND_DESIGN	4.3	967	2.8M	100,000+	Free	0	Everyone	Art & Design;Creativity	June 20, 2018	1.1	4.4 and up
5	Paper flowers instructions	ART_AND_DESIGN	4.4	167	5.6M	50,000+	Free	0	Everyone	Art & Design	March 26, 2017	1.0	2.3 and up
6	Smoke Effect Photo Maker - Smoke Editor	ART_AND_DESIGN	3.8	178	19M	50,000+	Free	0	Everyone	Art & Design	April 26, 2018	1.1	4.0.3 and up
7	Infinite Painter	ART_AND_DESIGN	4.1	36815	29M	1,000,000+	Free	0	Everyone	Art & Design	June 14, 2018	6.1.61.1	4.2 and up
8	Garden Coloring Book	ART_AND_DESIGN	4.4	13791	33M	1,000,000+	Free	0	Everyone	Art & Design	September 20, 2017	2.9.2	3.0 and up
9	Kids Paint Free - Drawing Fun	ART_AND_DESIGN	4.7	121	3.1M	10,000+	Free	0	Everyone	Art & Design;Creativity	July 3, 2018	2.8	4.0.3 and up

In [137]:
```
1 df.shape #---This dataframe has 10841 rows and 13 columns
```

Out[137]: (10841, 13)

In [138]:
```
df.info()    #---By using info() function, i got to know that this df has 12 object columns and one float columns
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10841 entries, 0 to 10840
Data columns (total 13 columns):
 #   Column          Non-Null Count  Dtype  
---  ------          --------------  -----  
 0   App             10841 non-null  object 
 1   Category        10841 non-null  object 
 2   Rating          9367 non-null   float64
 3   Reviews         10841 non-null  object 
 4   Size            10841 non-null  object 
 5   Installs        10841 non-null  object 
 6   Type            10840 non-null  object 
 7   Price           10841 non-null  object 
 8   Content Rating  10840 non-null  object 
 9   Genres          10841 non-null  object 
 10  Last Updated    10841 non-null  object 
 11  Current Ver     10833 non-null  object 
 12  Android Ver     10838 non-null  object 
dtypes: float64(1), object(12)
memory usage: 1.1+ MB
```

In [139]:
```
df.columns   #---This dataframe has various type of features(columns), will analyze each and everything
```

Out[139]:
```
Index(['App', 'Category', 'Rating', 'Reviews', 'Size', 'Installs', 'Type',
       'Price', 'Content Rating', 'Genres', 'Last Updated', 'Current Ver',
       'Android Ver'],
      dtype='object')
```

3. Check summary statistics of the dataset. List out the columns that need to be worked upon for model building.

In [140]:
```
df.describe()   #-- Since rating dtype is float, describe() function only applied to this column
```

Out[140]:

	Rating
count	9367.000000
mean	4.193338
std	0.537431
min	1.000000
25%	4.000000
50%	4.300000
75%	4.500000
max	19.000000

In [141]:
```
df['Rating'].skew()
```

Out[141]:
```
0.5956367473804342
```

In [142]:
```
df['Rating'].kurt()
```

Out[142]:
```
65.99478130826435
```

In [143]:
```
df['Rating'].corr
```

Out[143]:
```
<bound method Series.corr of 0        4.1
1        3.9
2        4.7
3        4.5
4        4.3
        ...
10836    4.5
10837    5.0
10838    NaN
10839    4.5
10840    4.5
Name: Rating, Length: 10841, dtype: float64>
```

In [144]:
```
# --These columns are need to be worked upon for model building.

catcols = df.select_dtypes(include='O')
catcols
```

Out[144]:

	App	Category	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
0	Photo Editor & Candy Camera & Grid & ScrapBook	ART_AND_DESIGN	159	19M	10,000+	Free	0	Everyone	Art & Design	January 7, 2018	1.0.0	4.0.3 and up
1	Coloring book moana	ART_AND_DESIGN	967	14M	500,000+	Free	0	Everyone	Art & Design;Pretend Play	January 15, 2018	2.0.0	4.0.3 and up
2	U Launcher Lite – FREE Live Cool Themes, Hide ...	ART_AND_DESIGN	87510	8.7M	5,000,000+	Free	0	Everyone	Art & Design	August 1, 2018	1.2.4	4.0.3 and up
3	Sketch - Draw & Paint	ART_AND_DESIGN	215644	25M	50,000,000+	Free	0	Teen	Art & Design	June 8, 2018	Varies with device	4.2 and up
4	Pixel Draw - Number Art Coloring Book	ART_AND_DESIGN	967	2.8M	100,000+	Free	0	Everyone	Art & Design;Creativity	June 20, 2018	1.1	4.4 and up
...	...	...	...	...	...	...	...	...	...	...	...	...
10836	Sya9a Maroc - FR	FAMILY	38	53M	5,000+	Free	0	Everyone	Education	July 25, 2017	1.48	4.1 and up

4. Check if there are any duplicate records in the dataset? if any drop them.

In [145]:
```
1 df[df.duplicated()]
```

Out[145]:

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
229	Quick PDF Scanner + OCR FREE	BUSINESS	4.2	80805	Varies with device	5,000,000+	Free	0	Everyone	Business	February 26, 2018	Varies with device	4.0.3 and up
236	Box	BUSINESS	4.2	159872	Varies with device	10,000,000+	Free	0	Everyone	Business	July 31, 2018	Varies with device	Varies with device
239	Google My Business	BUSINESS	4.4	70991	Varies with device	5,000,000+	Free	0	Everyone	Business	July 24, 2018	2.19.0.204537701	4.4 and up
256	ZOOM Cloud Meetings	BUSINESS	4.4	31614	37M	10,000,000+	Free	0	Everyone	Business	July 20, 2018	4.1.28165.0716	4.0 and up
261	join.me - Simple Meetings	BUSINESS	4.0	6989	Varies with device	1,000,000+	Free	0	Everyone	Business	July 16, 2018	4.3.0.508	4.4 and up
...	...	...	...	...	...	...	...	...	...	...	...	...	...
8643	Wunderlist: To-Do List & Tasks	PRODUCTIVITY	4.6	404610	Varies with device	10,000,000+	Free	0	Everyone	Productivity	April 6, 2018	Varies with device	Varies with device
8654	TickTick: To Do List with Reminder, Day Planner	PRODUCTIVITY	4.6	25370	Varies with device	1,000,000+	Free	0	Everyone	Productivity	August 6, 2018	Varies with device	Varies with device
8658	ColorNote Notepad Notes	PRODUCTIVITY	4.6	2401017	Varies with device	100,000,000+	Free	0	Everyone	Productivity	June 27, 2018	Varies with device	Varies with device
10049	Airway Ex - Intubate. Anesthetize. Train.	MEDICAL	4.3	123	86M	10,000+	Free	0	Everyone	Medical	June 1, 2018	0.6.88	5.0 and up
10768	AAFP	MEDICAL	3.8	63	24M	10,000+	Free	0	Everyone	Medical	June 22, 2018	2.3.1	5.0 and up

483 rows × 13 columns

In [146]:
```
1 10841 - 483 #--(duplicated columns before dropping)
```

Out[146]: 10358

In [147]:
```
#---Dropping duplicate values

df.drop_duplicates(inplace=True)
df
```

Out[147]:

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
0	Photo Editor & Candy Camera & Grid & ScrapBook	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
1	Coloring book moana	ART_AND_DESIGN	3.9	967	14M	500,000+	Free	0	Everyone	Art & Design;Pretend Play	January 15, 2018	2.0.0	4.0.3 and up
2	U Launcher Lite – FREE Live Cool Themes, Hide ...	ART_AND_DESIGN	4.7	87510	8.7M	5,000,000+	Free	0	Everyone	Art & Design	August 1, 2018	1.2.4	4.0.3 and up
3	Sketch - Draw & Paint	ART_AND_DESIGN	4.5	215644	25M	50,000,000+	Free	0	Teen	Art & Design	June 8, 2018	Varies with device	4.2 and up
4	Pixel Draw - Number Art Coloring	ART_AND_DESIGN	4.3	967	2.8M	100,000+	Free	0	Everyone	Art & Design;Creativity	June 20, 2018	1.1	4.4 and up

5. Check the unique categories of the column 'Category', Is there any invalid category? If yes, drop them.

In [148]:
```
len(df['Category'].unique())  #--length of the category column
```

Out[148]: 34

In [149]:
```
df['Category'].unique()
```

Out[149]:
```
array(['ART_AND_DESIGN', 'AUTO_AND_VEHICLES', 'BEAUTY',
       'BOOKS_AND_REFERENCE', 'BUSINESS', 'COMICS', 'COMMUNICATION',
       'DATING', 'EDUCATION', 'ENTERTAINMENT', 'EVENTS', 'FINANCE',
       'FOOD_AND_DRINK', 'HEALTH_AND_FITNESS', 'HOUSE_AND_HOME',
       'LIBRARIES_AND_DEMO', 'LIFESTYLE', 'GAME', 'FAMILY', 'MEDICAL',
       'SOCIAL', 'SHOPPING', 'PHOTOGRAPHY', 'SPORTS', 'TRAVEL_AND_LOCAL',
       'TOOLS', 'PERSONALIZATION', 'PRODUCTIVITY', 'PARENTING', 'WEATHER',
       'VIDEO_PLAYERS', 'NEWS_AND_MAGAZINES', 'MAPS_AND_NAVIGATION',
       '1.9'], dtype=object)
```

In [150]:
```
df['Category'].value_counts()  #---checking before dropping the invalid category(1.9)
```

Out[150]:
```
Category
FAMILY                 1943
GAME                   1121
TOOLS                   843
BUSINESS                427
MEDICAL                 408
PRODUCTIVITY            407
PERSONALIZATION         388
LIFESTYLE               373
COMMUNICATION           366
FINANCE                 360
SPORTS                  351
PHOTOGRAPHY             322
HEALTH_AND_FITNESS      306
SOCIAL                  280
NEWS_AND_MAGAZINES      264
TRAVEL_AND_LOCAL        237
BOOKS_AND_REFERENCE     230
SHOPPING                224
DATING                  196
VIDEO_PLAYERS           175
MAPS_AND_NAVIGATION     137
EDUCATION               130
FOOD_AND_DRINK          124
ENTERTAINMENT           111
AUTO_AND_VEHICLES        85
LIBRARIES_AND_DEMO       85
WEATHER                  82
HOUSE_AND_HOME           80
ART_AND_DESIGN           65
EVENTS                   64
PARENTING                60
COMICS                   60
BEAUTY                   53
1.9                       1
Name: count, dtype: int64
```

In [151]:
```
df['Category'][df['Category'] == '1.9']    #--pull the index of 1.9
```

Out[151]:
```
10472    1.9
Name: Category, dtype: object
```

In [152]:
```
#---Another method to pull index of specific values

# --df.iloc(10472)
```

In [153]:
```
df['Category'] = df['Category'].drop(index=10472)   #---Dropping the invalid category(1.9)
df['Category']
```

Out[153]:
```
0              ART_AND_DESIGN
1              ART_AND_DESIGN
2              ART_AND_DESIGN
3              ART_AND_DESIGN
4              ART_AND_DESIGN
                  ...
10836                  FAMILY
10837                  FAMILY
10838                 MEDICAL
10839    BOOKS_AND_REFERENCE
10840               LIFESTYLE
Name: Category, Length: 10358, dtype: object
```

In [154]:
```
df            #---df after dropping 1.9-invalid category
```

Out[154]:

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
0	Photo Editor & Candy Camera & Grid & ScrapBook	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
1	Coloring book moana	ART_AND_DESIGN	3.9	967	14M	500,000+	Free	0	Everyone	Art & Design;Pretend Play	January 15, 2018	2.0.0	4.0.3 and up
2	U Launcher Lite – FREE Live Cool Themes, Hide ...	ART_AND_DESIGN	4.7	87510	8.7M	5,000,000+	Free	0	Everyone	Art & Design	August 1, 2018	1.2.4	4.0.3 and up
3	Sketch - Draw & Paint	ART_AND_DESIGN	4.5	215644	25M	50,000,000+	Free	0	Teen	Art & Design	June 8, 2018	Varies with device	4.2 and up
4	Pixel Draw - Number Art Coloring Book	ART_AND_DESIGN	4.3	967	2.8M	100,000+	Free	0	Everyone	Art & Design;Creativity	June 20, 2018	1.1	4.4 and up
...	...	...	...	...	...	...	...	...	...	...	...	...	...
10836	Sya9a Maroc - FR	FAMILY	4.5	38	53M	5,000+	Free	0	Everyone	Education	July 25, 2017	1.48	4.1 and up
10837	Fr. Mike Schmitz Audio Teachings	FAMILY	5.0	4	3.6M	100+	Free	0	Everyone	Education	July 6, 2018	1.0	4.1 and up
10838	Parkinson Exercices FR	MEDICAL	NaN	3	9.5M	1,000+	Free	0	Everyone	Medical	January 20, 2017	1.0	2.2 and up
10839	The SCP Foundation DB fr nn5n	BOOKS_AND_REFERENCE	4.5	114	Varies with device	1,000+	Free	0	Mature 17+	Books & Reference	January 19, 2015	Varies with device	Varies with device

In [155]:
```
df['Category'].value_counts()  #---checking after dropping the invalid category(1.9)
```

Out[155]:
```
Category
FAMILY                1943
GAME                  1121
TOOLS                  843
BUSINESS               427
MEDICAL                408
PRODUCTIVITY           407
PERSONALIZATION        388
LIFESTYLE              373
COMMUNICATION          366
FINANCE                360
SPORTS                 351
PHOTOGRAPHY            322
HEALTH_AND_FITNESS     306
SOCIAL                 280
NEWS_AND_MAGAZINES     264
TRAVEL_AND_LOCAL       237
BOOKS_AND_REFERENCE    230
SHOPPING               224
DATING                 196
VIDEO_PLAYERS          175
MAPS_AND_NAVIGATION    137
EDUCATION              130
FOOD_AND_DRINK         124
ENTERTAINMENT          111
AUTO_AND_VEHICLES       85
LIBRARIES_AND_DEMO      85
WEATHER                 82
HOUSE_AND_HOME          80
ART_AND_DESIGN          65
EVENTS                  64
PARENTING               60
COMICS                  60
BEAUTY                  53
Name: count, dtype: int64
```

In [156]:
```
df['Category'].isnull().sum()
```

Out[156]:
```
1
```

In [157]:
```
df['Category'].mode()[0]
```

Out[157]:
```
'FAMILY'
```

6. Check if there are missing values present in the column Rating, If any? drop them and and create a new column as 'Rating_category' by converting ratings to high and low categories(>3.5 is high rest low)

In [159]:
```
df['Rating'].isnull().sum()     #---Checking before removing null values
```

Out[159]: 1465

In [160]:
```
df = df.dropna(subset='Rating')
df
```

Out[160]:

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
0	Photo Editor & Candy Camera & Grid & ScrapBook	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
1	Coloring book moana	ART_AND_DESIGN	3.9	967	14M	500,000+	Free	0	Everyone	Art & Design;Pretend Play	January 15, 2018	2.0.0	4.0.3 and up
2	U Launcher Lite – FREE Live Cool Themes, Hide ...	ART_AND_DESIGN	4.7	87510	8.7M	5,000,000+	Free	0	Everyone	Art & Design	August 1, 2018	1.2.4	4.0.3 and up
3	Sketch - Draw & Paint	ART_AND_DESIGN	4.5	215644	25M	50,000,000+	Free	0	Teen	Art & Design	June 8, 2018	Varies with device	4.2 and up
4	Pixel Draw - Number Art Coloring Book	ART_AND_DESIGN	4.3	967	2.8M	100,000+	Free	0	Everyone	Art & Design;Creativity	June 20, 2018	1.1	4.4 and up
...	...	...	...	...	...	...	...	...	...	...	...	...	...
10834	FR Calculator	FAMILY	4.0	7	2.6M	500+	Free	0	Everyone	Education	June 18, 2017	1.0.0	4.1 and up
10836	Sya9a Maroc - FR	FAMILY	4.5	38	53M	5,000+	Free	0	Everyone	Education	July 25, 2017	1.48	4.1 and up
10837	Fr. Mike Schmitz Audio Teachings	FAMILY	5.0	4	3.6M	100+	Free	0	Everyone	Education	July 6, 2018	1.0	4.1 and up
10839	The SCP Foundation DB fr nn5n	BOOKS_AND_REFERENCE	4.5	114	Varies with device	1,000+	Free	0	Mature 17+	Books & Reference	January 19, 2015	Varies with device	Varies with device
10840	iHoroscope - 2018 Daily Horoscope & Astrology	LIFESTYLE	4.5	398307	19M	10,000,000+	Free	0	Everyone	Lifestyle	July 25, 2018	Varies with device	Varies with device

8893 rows × 13 columns

In [161]:
```
df['Rating'].isnull().sum()   #---Checking after removing null values
```

Out[161]:
```
0
```

In [162]:
```
df['Rating_category'] = df['Rating'].apply(lambda x:'High' if x>3.5 else 'Low')
df['Rating_category']
```

```
C:\Users\Nithish\AppData\Local\Temp\ipykernel_12012\87577131.py:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  df['Rating_category'] = df['Rating'].apply(lambda x:'High' if x>3.5 else 'Low')
```

Out[162]:
```
0        High
1        High
2        High
3        High
4        High
         ...
10834    High
10836    High
10837    High
10839    High
10840    High
Name: Rating_category, Length: 8893, dtype: object
```

In [163]:
```
df
```

Out[163]:

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver	Rating_
0	Photo Editor & Candy Camera & Grid & ScrapBook	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	
1	Coloring book moana	ART_AND_DESIGN	3.9	967	14M	500,000+	Free	0	Everyone	Art & Design;Pretend Play	January 15, 2018	2.0.0	4.0.3 and up	
2	U Launcher Lite – FREE Live Cool Themes, Hide ...	ART_AND_DESIGN	4.7	87510	8.7M	5,000,000+	Free	0	Everyone	Art & Design	August 1, 2018	1.2.4	4.0.3 and up	
3	Sketch - Draw & Paint	ART_AND_DESIGN	4.5	215644	25M	50,000,000+	Free	0	Teen	Art & Design	June 8, 2018	Varies with device	4.2 and up	

	Paint											device	
4	Pixel Draw - Number Art Coloring Book	ART_AND_DESIGN	4.3	967	2.8M	100,000+	Free	0	Everyone	Art & Design;Creativity	June 20, 2018	1.1	4.4 and up
...	...	...	...	...	...	...	...	...	...	...	...	...	...
10834	FR Calculator	FAMILY	4.0	7	2.6M	500+	Free	0	Everyone	Education	June 18, 2017	1.0.0	4.1 and up
10836	Sya9a Maroc - FR	FAMILY	4.5	38	53M	5,000+	Free	0	Everyone	Education	July 25, 2017	1.48	4.1 and up
10837	Fr. Mike Schmitz Audio Teachings	FAMILY	5.0	4	3.6M	100+	Free	0	Everyone	Education	July 6, 2018	1.0	4.1 and up
10839	The SCP Foundation DB fr nn5n	BOOKS_AND_REFERENCE	4.5	114	Varies with device	1,000+	Free	0	Mature 17+	Books & Reference	January 19, 2015	Varies with device	Varies with device
10840	iHoroscope - 2018 Daily Horoscope & Astrology	LIFESTYLE	4.5	398307	19M	10,000,000+	Free	0	Everyone	Lifestyle	July 25, 2018	Varies with device	Varies with device

8893 rows × 14 columns

In [164]:
```
df['Rating_category'].isnull().sum()
```

Out[164]: 0

In [ ]:
```

```

7. Check the distribution of the newly created column 'Rating_category' and comment on the distribution.

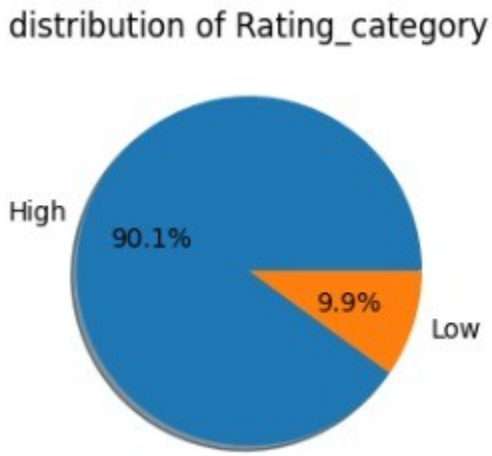
In [165]:
```
# --- To pull the distribution of Rating_category column

Rating_category_Distribution = df['Rating_category'].value_counts()
Rating_category_Distribution
```

Out[165]:
```
Rating_category
High    8013
Low      880
Name: count, dtype: int64
```

In [166]:
```
#---Distribution of 'Rating_category' using pie chart


plt.figure(figsize=(5,3))
plt.pie(df['Rating_category'].value_counts(),labels=df['Rating_category'].value_counts().index, autopct='%1.1f%%', shadow=Tr
plt.title('distribution of Rating_category')
plt.show()
```



8. Convert the column "Reviews" to numeric data type and check the presence of outliers in the column and handle the outliers using a transformation approach.(Hint: Use log transformation)

In [167]:
```
df['Reviews'].unique()
```

Out[167]:
```
array(['159', '967', '87510', ..., '603', '1195', '398307'], dtype=object)
```

In [168]:
```
df['Reviews'].dtypes
```

Out[168]:
```
dtype('O')
```

In [169]:
```
df['Reviews'] =  pd.to_numeric(df['Reviews'], errors='coerce')
df['Reviews']
```

```
C:\Users\Nithish\AppData\Local\Temp\ipykernel_12012\676789892.py:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead
```

In [169]:
```
df['Reviews'] =  pd.to_numeric(df['Reviews'], errors='coerce')
df['Reviews']
```

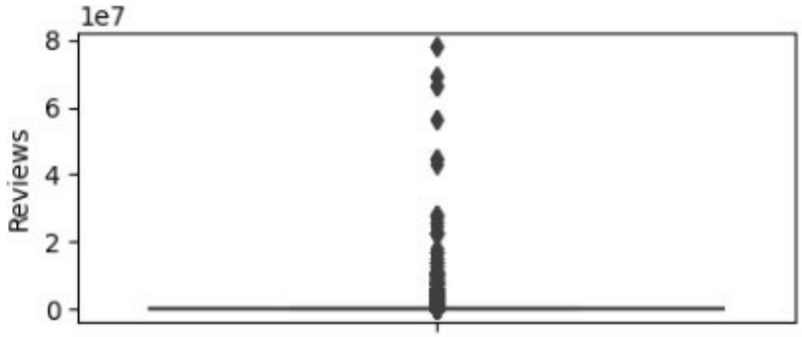
```
C:\Users\Nithish\AppData\Local\Temp\ipykernel_12012\676789892.py:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  df['Reviews'] =  pd.to_numeric(df['Reviews'], errors='coerce')
```

Out[169]:
```
0           159.0
1           967.0
2         87510.0
3        215644.0
4           967.0
           ...
10834         7.0
10836        38.0
10837         4.0
10839       114.0
10840    398307.0
Name: Reviews, Length: 8893, dtype: float64
```

In [170]:
```
# Check for the presence of outliers using box plot in 'Reviews' column

plt.figure(figsize=(5,2))
sns.boxplot(data=df, y=df['Reviews'])
plt.show()
```



In [171]:
```
# Handle outliers using log transformation
df['reviews'] = np.log(df['Reviews'])   # --Log transformation to supress the outliers
df['reviews']
```

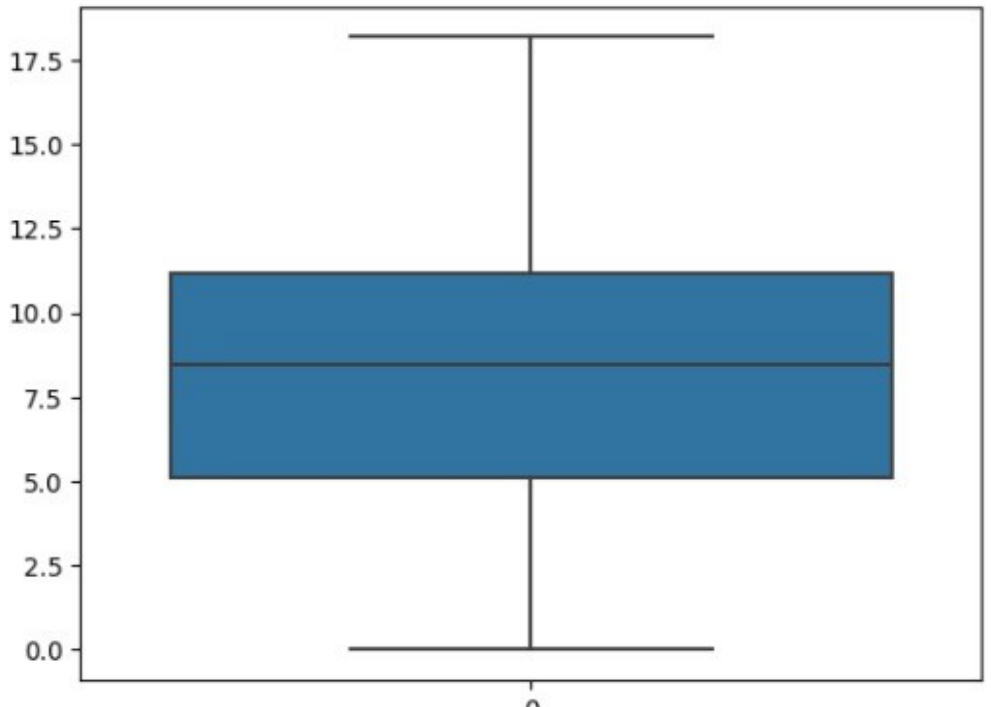
```
C:\Users\Nithish\AppData\Local\Temp\ipykernel_12012\2382583333.py:3: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  df['reviews'] = np.log(df['Reviews'])   # --Log transformation to supress the outliers
```

Out[171]:
```
0         5.068904
1         6.874198
2        11.379508
3        12.281384
4         6.874198
           ...
10834     1.945910
10836     3.637586
10837     1.386294
10839     4.736198
10840    12.894978
Name: reviews, Length: 8893, dtype: float64
```

In [172]:
```
sns.boxplot(df['reviews'])  #---Checking if Reviews column has outliers or not after supress the outliers
```

Out[172]: <Axes: >



In [173]:
```
df = df.drop(columns='Reviews')   #---Dropping Reviews column from df
```

In [174]:
```
df    #---Checking if Reviews column dropped or not
```

Out[174]:

	App	Category	Rating	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver	Rating_category
0	Photo Editor & Candy Camera & Grid & ScrapBook	ART_AND_DESIGN	4.1	19M	10,000+	Free	0	Everyone	Art & Design	January 7, 2018	1.0.0	4.0.3 and up	High
1	Coloring book moana	ART_AND_DESIGN	3.9	14M	500,000+	Free	0	Everyone	Art & Design;Pretend Play	January 15, 2018	2.0.0	4.0.3 and up	High
2	U Launcher Lite – FREE Live Cool Themes, Hide ...	ART_AND_DESIGN	4.7	8.7M	5,000,000+	Free	0	Everyone	Art & Design	August 1, 2018	1.2.4	4.0.3 and up	High
3	Sketch - Draw & Paint	ART_AND_DESIGN	4.5	25M	50,000,000+	Free	0	Teen	Art & Design	June 8, 2018	Varies with device	4.2 and up	High
4	Pixel Draw - Number Art Coloring Book	ART_AND_DESIGN	4.3	2.8M	100,000+	Free	0	Everyone	Art & Design;Creativity	June 20, 2018	1.1	4.4 and up	High
...	...	...	...	...	...	...	...	...	...	...	...	...	...
10834	FR Calculator	FAMILY	4.0	2.6M	500+	Free	0	Everyone	Education	June 18, 2017	1.0.0	4.1 and up	High
10836	Sya9a Maroc - FR	FAMILY	4.5	53M	5,000+	Free	0	Everyone	Education	July 25, 2017	1.48	4.1 and up	High
10837	Fr. Mike Schmitz Audio Teachings	FAMILY	5.0	3.6M	100+	Free	0	Everyone	Education	July 6, 2018	1.0	4.1 and up	High
10839	The SCP Foundation DB fr nn5n	BOOKS_AND_REFERENCE	4.5	Varies with device	1,000+	Free	0	Mature 17+	Books & Reference	January 19, 2015	Varies with device	Varies with device	High
10840	iHoroscope - 2018 Daily Horoscope & Astrology	LIFESTYLE	4.5	19M	10,000,000+	Free	0	Everyone	Lifestyle	July 25, 2018	Varies with device	Varies with device	High

8893 rows × 14 columns

In [175]:
```
df['reviews'].isnull().sum()  #---checking null values of newly created reviews column
```

Out[175]: 1

In [176]:
```
df['reviews'].mean()     #---pull mean value
```

Out[176]: 8.234892597330452

In [177]:
```
df['reviews'].median()   #---pull median values
```

Out[177]: 8.458397689368425

In [178]:
```
df['reviews'].mode()[0]  #--pull mode values
```

Out[178]: 0.6931471805599453

In [179]:
```
df['reviews'] = df['reviews'].fillna(df['reviews'].mode()[0])
df['reviews']

#---fill the nan using mode values
```

Out[179]:
```
0         5.068904
1         6.874198
2        11.379508
3        12.281384
4         6.874198
           ...
10834     1.945910
10836     3.637586
10837     1.386294
10839     4.736198
10840    12.894978
Name: reviews, Length: 8893, dtype: float64
```

In [180]:
```
df['reviews'].isnull().sum()    #---Now, nan is zero
```

Out[180]: 0

9. The column 'Size' contains alphanumeric values, treat the non numeric data and convert the column into suitable data type. (hint: Replace M with 1 million and K with 1 thousand, and drop the entries where size='Varies with device')

In [181]:
```
df['Size'].dtypes    #---Checking dtype
```

Out[181]:
```
dtype('O')
```

In [182]:
```
df['Size']
```

Out[182]:
```
0                     19M
1                     14M
2                    8.7M
3                     25M
4                    2.8M
                ...
10834                2.6M
10836                 53M
10837                3.6M
10839    Varies with device
10840                 19M
Name: Size, Length: 8893, dtype: object
```

In [183]:
```
df['Size'] = df['Size'].str.replace('M', '000000')
df['Size'] = df['Size'].str.replace('k', '000')


#---Replacing M with 000000 and K with 000
```

In [184]:
```
df['Size'].sample(15)    #---Checking that all the k and M got changed or not
```

Out[184]:
```
5962              32000000
4599              19000000
7466              38000000
9779              17000000
5535              91000000
4582                314000
4803             5.0000000
2759    Varies with device
5849    Varies with device
6661    Varies with device
4744    Varies with device
8279              11000000
10011             82000000
1358              24000000
```

In [185]:
```
df['Size'][df['Size'] == 'Varies with device']     #---Checking if 'Varies with device' is present or not
```

Out[185]:
```
37       Varies with device
42       Varies with device
52       Varies with device
67       Varies with device
68       Varies with device
               ...
10713    Varies with device
10725    Varies with device
10765    Varies with device
10826    Varies with device
10839    Varies with device
Name: Size, Length: 1468, dtype: object
```

In [186]:
```
len(df['Size'][df['Size'] == 'Varies with device'])   #---len of 'Varies with device'
```

Out[186]:
```
1468
```

In [187]:
```
len(df)
```

Out[187]:
```
8893
```

In [188]:
```
8893 - 1468
```

Out[188]:
```
7425
```

In [189]:
```
df['changed_size'] = df['Size'][df['Size'] != 'Varies with device']
df['changed_size']


#---Pull all the values except 'Varies with device' from 'changed_size' column
```

Out[189]:
```
0         19000000
1         14000000
2        8.7000000
3         25000000
4        2.8000000
            ...
10834    2.6000000
10836     53000000
10837    3.6000000
10839          NaN
10840     19000000
Name: changed_size, Length: 8893, dtype: object
```

In [190]:
```
df['changed_size'].dtypes   #---Checking dtypes
```

Out[190]:
```
dtype('O')
```

In [191]:
```
df['changed_size'] = df['changed_size'].str.replace(',','')
df['changed_size']

#---Replace , with empty
```

Out[191]:
```
0         19000000
1         14000000
2        8.7000000
3         25000000
4        2.8000000
           ...
10834    2.6000000
10836     53000000
10837    3.6000000
10839          NaN
10840     19000000
Name: changed_size, Length: 8893, dtype: object
```

In [192]:
```
df['changed_size'] = df['changed_size'].str.replace('+','')
df['changed_size']

#---Replace + with empty
```

Out[192]:
```
0         19000000
1         14000000
2        8.7000000
3         25000000
4        2.8000000
           ...
10834    2.6000000
10836     53000000
10837    3.6000000
10839          NaN
10840     19000000
Name: changed_size, Length: 8893, dtype: object
```

In [193]:
```
df['changed_size'] = df['changed_size'].str.replace('.','')
df['changed_size']

#---Replace . with empty
```

Out[193]:
```
0        19000000
1        14000000
2        87000000
3        25000000
4        28000000
           ...
10834    26000000
10836    53000000
10837    36000000
10839         NaN
10840    19000000
Name: changed_size, Length: 8893, dtype: object
```

In [194]:
```
df['changed_size'].isnull().sum()
```

Out[194]:
```
1468
```

In [195]:
```
df['changed_size'] = df['changed_size'].fillna(0)
df['changed_size']
```

Out[195]:
```
0        19000000
1        14000000
2        87000000
3        25000000
4        28000000
           ...
10834    26000000
10836    53000000
10837    36000000
10839           0
10840    19000000
Name: changed_size, Length: 8893, dtype: object
```

In [196]:
```
df['changed_size'] = df['changed_size'].astype(int)
df['changed_size']

#---Converting changed_size column to int
```

Out[196]:
```
0        19000000
1        14000000
2        87000000
3        25000000
4        28000000
           ...
10834    26000000
10836    53000000
10837    36000000
10839           0
10840    19000000
Name: changed_size, Length: 8893, dtype: int32
```

In [197]:
```
df = df.drop(columns='Size')   #--Dropping Size column
```

In [198]:
```
df
```

Out[198]:

	App	Category	Rating	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver	Rating_category	revie
0	Photo Editor & Candy Camera & Grid & ScrapBook	ART_AND_DESIGN	4.1	10,000+	Free	0	Everyone	Art & Design	January 7, 2018	1.0.0	4.0.3 and up	High	5.0689
1	Coloring book moana	ART_AND_DESIGN	3.9	500,000+	Free	0	Everyone	Art & Design;Pretend Play	January 15, 2018	2.0.0	4.0.3 and up	High	6.8741
2	U Launcher Lite – FREE Live Cool Themes, Hide ...	ART_AND_DESIGN	4.7	5,000,000+	Free	0	Everyone	Art & Design	August 1, 2018	1.2.4	4.0.3 and up	High	11.3795
3	Sketch - Draw & Paint	ART_AND_DESIGN	4.5	50,000,000+	Free	0	Teen	Art & Design	June 8, 2018	Varies with device	4.2 and up	High	12.2813
4	Pixel Draw - Number Art Coloring Book	ART_AND_DESIGN	4.3	100,000+	Free	0	Everyone	Art & Design;Creativity	June 20, 2018	1.1	4.4 and up	High	6.8741
...	...	...	...	...	...	...	...	...	...	...	...	...	...
10834	FR Calculator	FAMILY	4.0	500+	Free	0	Everyone	Education	June 18, 2017	1.0.0	4.1 and up	High	1.9459

10. Check the column 'Installs', treat the unwanted characters and convert the column into a suitable data type.

In [199]:
```
df['Installs']   #---Pull values of installs
```

Out[199]:
```
0            10,000+
1           500,000+
2         5,000,000+
3        50,000,000+
4           100,000+
             ...
10834           500+
10836         5,000+
10837           100+
10839         1,000+
10840    10,000,000+
Name: Installs, Length: 8893, dtype: object
```

In [200]:
```
df['Installs'].value_counts()   #---Checking the counts of values
```

Out[200]:
```
Installs
1,000,000+        1486
10,000,000+       1132
100,000+          1110
10,000+            989
1,000+             698
5,000,000+         683
500,000+           516
50,000+            462
5,000+             426
100,000,000+       369
100+               303
50,000,000+        272
500+               199
10+                 69
500,000,000+        61
50+                 56
1,000,000,000+      49
5+                   9
1+                   3
Free                 1
Name: count, dtype: int64
```

In [201]:
```
df['Installs'].unique()   #---Pull unique values
```

Out[201]:
```
array(['10,000+', '500,000+', '5,000,000+', '50,000,000+', '100,000+',
       '50,000+', '1,000,000+', '10,000,000+', '5,000+', '100,000,000+',
       '1,000,000,000+', '1,000+', '500,000,000+', '100+', '500+', '10+',
       '5+', '50+', '1+', 'Free'], dtype=object)
```

In [202]:
```
df['Installs'] = df['Installs'].str.replace(',', '')
df['Installs'] = df['Installs'].str.replace('+', '')

#---Replace , with empty
#---Replace + with empty
```

In [203]:
```
df['Installs'][df['Installs'] == 'Free']     #---Checking if 'Free' is present or not
```

Out[203]:
```
10472    Free
Name: Installs, dtype: object
```

In [204]:
```
df['Installs'] = df['Installs'].drop(index= 10472)
df['Installs']

#---Dropping the specific index value
```

Out[204]:
```
0          10000
1         500000
2        5000000
3       50000000
4         100000
           ...
10834        500
10836       5000
10837        100
10839       1000
10840    10000000
Name: Installs, Length: 8893, dtype: object
```

In [205]:
```
df['Installs'].dtypes   #---Checking dtypes
```

Out[205]:
```
dtype('O')
```

In [206]:
```
df['Installs'].isnull().sum()   #----Checking null values
```

Out[206]:
```
1
```

In [207]:
```
df['Installs'] = df['Installs'].fillna(0)
df['Installs']


#---Fill all the null values
```

Out[207]:
```
0          10000
1         500000
2        5000000
3       50000000
4         100000
          ...
10834        500
10836       5000
10837        100
10839       1000
10840    10000000
Name: Installs, Length: 8893, dtype: object
```

In [208]:
```
df['Installs'].isnull().sum()
```

Out[208]:
```
0
```

In [209]:
```
df['Installs'] = df['Installs'].astype(int)
df['Installs']

#---Changing dtypes
```

Out[209]:
```
0          10000
1         500000
2        5000000
3       50000000
4         100000
          ...
10834        500
10836       5000
10837        100
10839       1000
10840    10000000
Name: Installs, Length: 8893, dtype: int32
```

11. Check the column 'Price' , remove the unwanted characters and convert the column into a suitable data type.

In [210]:
```
df['Price']
```

Out[210]:
```
0        0
1        0
2        0
3        0
4        0
        ..
10834    0
10836    0
10837    0
10839    0
10840    0
Name: Price, Length: 8893, dtype: object
```

In [211]:
```
df['Price'].value_counts()
```

Out[211]:
```
Price
0          8279
$2.99       110
$0.99       105
$4.99        68
$1.99        59
           ...
$1.29         1
$299.99       1
$379.99       1
$37.99        1
$1.20         1
Name: count, Length: 74, dtype: int64
```

In [212]:
```
df['Price'].unique()
```

Out[212]:
```
array(['0', '$4.99', '$3.99', '$6.99', '$7.99', '$5.99', '$2.99', '$3.49',
       '$1.99', '$9.99', '$7.49', '$0.99', '$9.00', '$5.49', '$10.00',
       '$24.99', '$11.99', '$79.99', '$16.99', '$14.99', '$29.99',
       '$12.99', '$2.49', '$10.99', '$1.50', '$19.99', '$15.99', '$33.99',
       '$39.99', '$3.95', '$4.49', '$1.70', '$8.99', '$1.49', '$3.88',
       '$399.99', '$17.99', '$400.00', '$3.02', '$1.76', '$4.84', '$4.77',
       '$1.61', '$2.50', '$1.59', '$6.49', '$1.29', '$299.99', '$379.99',
       '$37.99', '$18.99', '$389.99', '$8.49', '$1.75', '$14.00', '$2.00',
       '$3.08', '$2.59', '$19.40', '$3.90', '$4.59', '$15.46', '$3.04',
       '$13.99', '$4.29', '$3.28', '$4.60', '$1.00', '$2.95', '$2.90',
       '$1.97', '$2.56', 'Everyone', '$1.20'], dtype=object)
```

In [213]:
```
df['Price'] = df['Price'].str.replace('$','')
df['Price'] = df['Price'].str.replace('.','')

#---Replace $ with empty
#---Replace . with empty
```

In [214]:
```
df['Price'][df['Price'] == 'Everyone']        #---Checking if 'Everyone' is present or not
```

Out[214]:
```
10472    Everyone
Name: Price, dtype: object
```

In [215]:
```
df['Price'] = df['Price'].drop(index=10472)
df['Price']

#---Dropping the specific index value
```

Out[215]:
```
0        0
1        0
2        0
3        0
4        0
        ..
10834    0
10836    0
10837    0
10839    0
10840    0
Name: Price, Length: 8893, dtype: object
```

In [216]:
```
df['Price'].isnull().sum()   #----Checking null values
```

Out[216]:
```
1
```

In [217]:
```
df['Price'] = df['Price'].fillna(0)
df['Price']

#---Fill all the null values
```

Out[217]:
```
0        0
1        0
2        0
3        0
4        0
        ..
10834    0
10836    0
10837    0
10839    0
10840    0
```

In [218]:
```
df['Price'].isnull().sum()
```

Out[218]:
```
0
```

In [219]:
```
df['Price'] = df['Price'].astype(int)
df['Price']

#---Changing dtypes
```

Out[219]:
```
0        0
1        0
2        0
3        0
4        0
        ..
10834    0
10836    0
10837    0
10839    0
10840    0
Name: Price, Length: 8893, dtype: int32
```

In [ ]:
```

```

Handling 'Content Rating' column

In [220]:
```
df['Content Rating'].isnull().sum()  #---Checking null values
```

Out[220]:
```
1
```

In [221]:
```
df['Content Rating'].mode()[0]       #--pull mode value
```

Out[221]:
```
'Everyone'
```

In [222]:
```
df['Content Rating'] = df['Content Rating'].fillna(df['Content Rating'].mode()[0])
df['Content Rating']

#---Fill all the null values
```

Out[222]:
```
0          Everyone
1          Everyone
2          Everyone
3              Teen
4          Everyone
            ...
10834      Everyone
10836      Everyone
10837      Everyone
10839    Mature 17+
10840      Everyone
Name: Content Rating, Length: 8893, dtype: object
```

In [223]:
```
df['Content Rating'].isnull().sum()
```

Out[223]:
```
0
```

Handling 'App' column

In [224]:
```
df['App'].unique()
```

Out[224]:
```
array(['Photo Editor & Candy Camera & Grid & ScrapBook',
       'Coloring book moana',
       'U Launcher Lite – FREE Live Cool Themes, Hide Apps', ...,
       'Fr. Mike Schmitz Audio Teachings',
       'The SCP Foundation DB fr nn5n',
       'iHoroscope - 2018 Daily Horoscope & Astrology'], dtype=object)
```

Handling 'Type' column

In [225]:
```
df['Type'][df['Type'] == '0']           #---Checking if '0' is present or not
```

Out[225]:
```
10472    0
Name: Type, dtype: object
```

In [226]:
```
df['Type'] = df['Type'].drop(index=10472)
df['Type']

#---Dropping the specific index value
```

Out[226]:
```
0        Free
1        Free
2        Free
3        Free
4        Free
         ...
10834    Free
10836    Free
10837    Free
10839    Free
10840    Free
Name: Type, Length: 8893, dtype: object
```

In [227]:
```
df['Type'] = df['Type'].str.replace('0','Free')
df['Type']

#---Replacing 0 with free
```

Out[227]:
```
0        Free
1        Free
2        Free
3        Free
4        Free
         ...
10834    Free
10836    Free
10837    Free
10839    Free
10840    Free
Name: Type, Length: 8893, dtype: object
```

In [228]:
```
df['Type'].unique()
```

Out[228]:
```
array(['Free', 'Paid', nan], dtype=object)
```

In [229]:
```
df['Type'].isnull().sum()    #---Checking null values
```

Out[229]:
```
1
```

In [230]:
```
df['Type'].mode()[0]         #--Pull mode value
```

Out[230]:
```
'Free'
```

In [230]:
```
df['Type'].mode()[0]          #--Pull mode value
```

Out[230]:
```
'Free'
```

In [231]:
```
df['Type'] = df['Type'].fillna(df['Type'].mode()[0])
df['Type']

#--Fill all the null values
```

Out[231]:
```
0        Free
1        Free
2        Free
3        Free
4        Free
         ...
10834    Free
10836    Free
10837    Free
10839    Free
10840    Free
Name: Type, Length: 8893, dtype: object
```

In [232]:
```
df['Type'].isnull().sum()
```

Out[232]:
```
0
```

Handling 'Genres' column

In [233]:
```
df['Genres'].unique()   #---Pull unique values
```

Out[233]:
```
array(['Art & Design', 'Art & Design;Pretend Play',
       'Art & Design;Creativity', 'Auto & Vehicles', 'Beauty',
       'Books & Reference', 'Business', 'Comics', 'Comics;Creativity',
       'Communication', 'Dating', 'Education;Education', 'Education',
       'Education;Creativity', 'Education;Music & Video',
       'Education;Action & Adventure', 'Education;Pretend Play',
       'Education;Brain Games', 'Entertainment',
       'Entertainment;Music & Video', 'Entertainment;Brain Games',
       'Entertainment;Creativity', 'Events', 'Finance', 'Food & Drink',
       'Health & Fitness', 'House & Home', 'Libraries & Demo',
       'Lifestyle', 'Lifestyle;Pretend Play',
       'Adventure;Action & Adventure', 'Arcade', 'Casual', 'Card',
       'Casual;Pretend Play', 'Action', 'Strategy', 'Puzzle', 'Sports',
       'Music', 'Word', 'Racing', 'Casual;Creativity',
       'Casual;Action & Adventure', 'Simulation', 'Adventure', 'Board',
       'Trivia', 'Role Playing', 'Simulation;Education',
       'Action;Action & Adventure', 'Casual;Brain Games',
       'Simulation;Action & Adventure', 'Educational;Creativity',
```

12. Drop the columns which you think redundant for the analysis.(suggestion: drop column 'rating', since we created a new feature from it (i.e. rating_category) and the columns 'App', 'Rating' ,'Genres','Last Updated', 'Current Ver','Android Ver' columns since which are redundant for our analysis)

In [234]:
```
1 df
```

Out[234]:

	App	Category	Rating	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver	Rating_category	reviews
0	Photo Editor & Candy Camera & Grid & ScrapBook	ART_AND_DESIGN	4.1	10000	Free	0	Everyone	Art & Design	January 7, 2018	1.0.0	4.0.3 and up	High	5.068904
1	Coloring book moana	ART_AND_DESIGN	3.9	500000	Free	0	Everyone	Art & Design;Pretend Play	January 15, 2018	2.0.0	4.0.3 and up	High	6.874198
2	U Launcher Lite – FREE Live Cool Themes, Hide ...	ART_AND_DESIGN	4.7	5000000	Free	0	Everyone	Art & Design	August 1, 2018	1.2.4	4.0.3 and up	High	11.379508
3	Sketch - Draw & Paint	ART_AND_DESIGN	4.5	50000000	Free	0	Teen	Art & Design	June 8, 2018	Varies with device	4.2 and up	High	12.281384
4	Pixel Draw - Number Art Coloring Book	ART_AND_DESIGN	4.3	100000	Free	0	Everyone	Art & Design;Creativity	June 20, 2018	1.1	4.4 and up	High	6.874198
...	...	...	...	...	...	...	...	...	...	...	...	...	...
10834	FR Calculator	FAMILY	4.0	500	Free	0	Everyone	Education	June 18, 2017	1.0.0	4.1 and up	High	1.945910
10836	Sya9a Maroc - FR	FAMILY	4.5	5000	Free	0	Everyone	Education	July 25, 2017	1.48	4.1 and up	High	3.637586
10837	Fr. Mike Schmitz Audio Teachings	FAMILY	5.0	100	Free	0	Everyone	Education	July 6, 2018	1.0	4.1 and up	High	1.386294
10839	The SCP Foundation DB fr nn5n	BOOKS_AND_REFERENCE	4.5	1000	Free	0	Mature 17+	Books & Reference	January 19, 2015	Varies with device	Varies with device	High	4.736198
	iHoroscope												

In [235]:
```
df = df.drop(columns=['Rating', 'Last Updated', 'Current Ver', 'Android Ver'])

#---Dropping all the redundant columns to do the proper analysis
```

In [236]:
```
df  #---Checking if all the columns are dropped or not
```

Out[236]:

	App	Category	Installs	Type	Price	Content Rating	Genres	Rating_category	reviews	changed_size
0	Photo Editor & Candy Camera & Grid & ScrapBook	ART_AND_DESIGN	10000	Free	0	Everyone	Art & Design	High	5.068904	19000000
1	Coloring book moana	ART_AND_DESIGN	500000	Free	0	Everyone	Art & Design;Pretend Play	High	6.874198	14000000
2	U Launcher Lite – FREE Live Cool Themes, Hide ...	ART_AND_DESIGN	5000000	Free	0	Everyone	Art & Design	High	11.379508	87000000
3	Sketch - Draw & Paint	ART_AND_DESIGN	50000000	Free	0	Teen	Art & Design	High	12.281384	25000000
4	Pixel Draw - Number Art Coloring Book	ART_AND_DESIGN	100000	Free	0	Everyone	Art & Design;Creativity	High	6.874198	28000000
...	...	...	...	...	...	...	...	...	...	...
10834	FR Calculator	FAMILY	500	Free	0	Everyone	Education	High	1.945910	26000000
10836	Sya9a Maroc - FR	FAMILY	5000	Free	0	Everyone	Education	High	3.637586	53000000
10837	Fr. Mike Schmitz Audio Teachings	FAMILY	100	Free	0	Everyone	Education	High	1.386294	36000000
10839	The SCP Foundation DB fr nn5n	BOOKS_AND_REFERENCE	1000	Free	0	Mature 17+	Books & Reference	High	4.736198	0
10840	iHoroscope - 2018 Daily Horoscope & Astrology	LIFESTYLE	10000000	Free	0	Everyone	Lifestyle	High	12.894978	19000000

8893 rows × 10 columns

13. Encode the categorical columns.

In [237]:
```
catcols = df.select_dtypes(include='O').columns
catcols

#---Pull all the categorical columns
```

Out[237]:
```
Index(['App', 'Category', 'Type', 'Content Rating', 'Genres',
       'Rating_category'],
      dtype='object')
```

In [238]:
```
from sklearn.preprocessing import LabelEncoder
le = LabelEncoder()

#---Importing LabelEncoder from sklearn - preprocessing
```

In [241]:
```
df['App'] = le.fit_transform(df['App'])
df['App']

#--fit and transform the values in App column
```

Out[241]:
```
0        5809
1        2134
2        7326
3        6577
4        5865
        ...
10834    3410
10836    6896
10837    3701
10839    7099
10840    8049
Name: App, Length: 8893, dtype: int64
```

In [244]:
```
df['Category'] = le.fit_transform(df['Category'])
df['Category']

#--fit and transform the values in category column
```

Out[244]:
```
0         0
1         0
2         0
3         0
4         0
        ..
10834    11
10836    11
10837    11
10839     3
10840    18
Name: Category, Length: 8893, dtype: int64
```

In [246]:
```
df['Type'] = le.fit_transform(df['Type'])
df['Type']

#--fit and transform the values in Type column
```

Out[246]:
```
0        0
1        0
2        0
3        0
4        0
        ..
10834    0
10836    0
10837    0
10839    0
10840    0
Name: Type, Length: 8893, dtype: int32
```

In [247]:
```
df['Content Rating'] = le.fit_transform(df['Content Rating'])
df['Content Rating']

#--fit and transform the values in content Rating column
```

Out[247]:
```
0        1
1        1
2        1
3        4
4        1
        ..
10834    1
10836    1
10837    1
10839    3
10840    1
Name: Content Rating, Length: 8893, dtype: int32
```

In [248]:
```
df['Genres'] = le.fit_transform(df['Genres'])
df['Genres']

#--fit and transform the values in Genres column
```

Out[248]:
```
0         9
1        11
2         9
3         9
4        10
        ..
10834    37
10836    37
10837    37
10839    18
10840    66
Name: Genres, Length: 8893, dtype: int32
```

In [249]:
```
df['Rating_category'] = le.fit_transform(df['Rating_category'])
df['Rating_category']

#--fit and transform the values in Rating_category column
```

Out[249]:
```
0        0
1        0
2        0
3        0
4        0
       ..
10834    0
10836    0
10837    0
10839    0
10840    0
Name: Rating_category, Length: 8893, dtype: int32
```

In [250]:
```
df  #---Checking if all the columns are got transformed or not in df
```

Out[250]:

	App	Category	Installs	Type	Price	Content Rating	Genres	Rating_category	reviews	changed_size
0	5809	0	10000	0	0	1	9	0	5.068904	19000000
1	2134	0	500000	0	0	1	11	0	6.874198	14000000
2	7326	0	5000000	0	0	1	9	0	11.379508	87000000
3	6577	0	50000000	0	0	4	9	0	12.281384	25000000
4	5865	0	100000	0	0	1	10	0	6.874198	28000000
...	...	...	...	...	...	...	...	...	...	...
10834	3410	11	500	0	0	1	37	0	1.945910	26000000
10836	6896	11	5000	0	0	1	37	0	3.637586	53000000
10837	3701	11	100	0	0	1	37	0	1.386294	36000000
10839	7099	3	1000	0	0	3	18	0	4.736198	0
10840	8049	18	10000000	0	0	1	66	0	12.894978	19000000

8893 rows × 10 columns

In [252]:
```
df.dtypes   #---Checking dtypes
```

Out[252]:
```
App                int64
Category           int64
Installs           int32
Type               int32
Price              int32
Content Rating     int32
Genres             int32
Rating_category    int32
reviews          float64
changed_size       int32
dtype: object
```

In [ ]:
```

```

14. Segregate the target and independent features (Hint: Use Rating_category as the target)

In [253]:
```
independent_features_x = df.drop(columns='Rating_category')
target_y = df[['Rating_category']]

#----Segregating the target(y) and independent features(x)
```

In [254]:
```
independent_features_x
```

Out[254]:

	App	Category	Installs	Type	Price	Content Rating	Genres	reviews	changed_size
0	5809	0	10000	0	0	1	9	5.068904	19000000
1	2134	0	500000	0	0	1	11	6.874198	14000000
2	7326	0	5000000	0	0	1	9	11.379508	87000000
3	6577	0	50000000	0	0	4	9	12.281384	25000000
4	5865	0	100000	0	0	1	10	6.874198	28000000
...	...	...	...	...	...	...	...	...	...
10834	3410	11	500	0	0	1	37	1.945910	26000000
10836	6896	11	5000	0	0	1	37	3.637586	53000000
10837	3701	11	100	0	0	1	37	1.386294	36000000
10839	7099	3	1000	0	0	3	18	4.736198	0
10840	8049	18	10000000	0	0	1	66	12.894978	19000000

8893 rows × 9 columns

In [255]:
```
target_y
```

Out[255]:

	Rating_category
0	0
1	0
2	0
3	0
4	0
...	...
10834	0
10836	0
10837	0

15. Split the dataset into train and test.

In [256]:
```
from sklearn.model_selection import train_test_split

#---Importing train_test_split from sklearn - model_selection
```

In [257]:
```
x_train, x_test, y_train, y_test = train_test_split(independent_features_x,target_y, test_size=0.25, random_state=555)


#----Spliting as x_train, x_test, y_train, y_test
```

In [258]:
```
x_train
```

Out[258]:

	App	Category	Installs	Type	Price	Content Rating	Genres	reviews	changed_size
3651	7624	32	100000	0	0	1	114	7.669962	0
4928	133	21	1000000	0	0	2	74	10.055822	14000000
9734	2608	11	100000	0	0	1	95	7.855545	53000000
168	8004	3	1000000	0	0	1	18	9.968151	23000000
922	839	9	1000000	0	0	4	50	9.602855	19000000
...	...	...	...	...	...	...	...	...	...
8638	2698	14	5000000	0	0	1	87	11.945409	78000000
247	2229	4	500000	0	0	1	20	8.333030	48000000
8026	4153	11	500000	0	0	1	50	8.822322	13000000
9088	6796	14	10000000	0	0	4	0	11.702322	21000000
5552	6822	11	10000	0	0	4	50	6.016157	28000000

6669 rows × 9 columns

In [259]:
```
x_test
```

Out[259]:

	App	Category	Installs	Type	Price	Content Rating	Genres	reviews	changed_size
4014	4711	11	10000	0	0	1	37	3.761200	37000000
4932	137	16	100000	0	0	1	64	7.279319	97000000
4950	416	29	1000000	0	0	1	106	9.581283	43000000
8064	1801	29	500	0	0	1	106	2.484907	25000000
4107	7486	29	50000000	0	0	4	106	13.703948	31000000
...	...	...	...	...	...	...	...	...	...

In [260]:
```
1 y_train
```

Out[260]:

	Rating_category
3651	0
4928	0
9734	1
168	0
922	0
...	...
8638	0
247	0
8026	0
9088	0
5552	0

6669 rows × 1 columns

In [261]:
```
1 y_test
```

Out[261]:

	Rating_category
4014	0
4932	1
4950	0
8064	0
4107	0
...	...
6207	0
7415	0
9601	0
8432	0
6087	0

2224 rows × 1 columns

16. Standardize the data, so that the values are within a particular range.

In [262]:
```
from sklearn.preprocessing import StandardScaler

#---Importing StandardScaler from sklearn - preprocessing
```

In [263]:
```
ss = StandardScaler()
```

In [267]:
```
df = ss.fit_transform(df)    #--fit and transform the values in df
```

In [269]:
```
df = pd.DataFrame(df)
```

In [270]:
```
df   #---Checking if all the columns are got transformed or not in df
```

Out[270]:

	0	1	2	3	4	5	6	7	8	9
0	0.708736	-2.033208	-0.190789	-0.272092	-0.059493	-0.469985	-1.635118	-0.331393	-0.815634	-0.520794
1	-0.846927	-2.033208	-0.185116	-0.272092	-0.059493	-0.469985	-1.574482	-0.331393	-0.350423	-0.703414
2	1.350897	-2.033208	-0.133012	-0.272092	-0.059493	-0.469985	-1.635118	-0.331393	0.810564	1.962849
3	1.033838	-2.033208	0.388023	-0.272092	-0.059493	2.481001	-1.635118	-0.331393	1.042971	-0.301649
4	0.732441	-2.033208	-0.189747	-0.272092	-0.059493	-0.469985	-1.604800	-0.331393	-0.350423	-0.192076
...	...	...	...	...	...	...	...	...	...	...
8888	-0.306784	-0.704090	-0.190899	-0.272092	-0.059493	-0.469985	-0.786217	-0.331393	-1.620408	-0.265124
8889	1.168874	-0.704090	-0.190847	-0.272092	-0.059493	-0.469985	-0.786217	-0.331393	-1.184475	0.721028
8890	-0.183601	-0.704090	-0.190904	-0.272092	-0.059493	-0.469985	-0.786217	-0.331393	-1.764617	0.100117
8891	1.254806	-1.670721	-0.190893	-0.272092	-0.059493	1.497339	-1.362257	-0.331393	-0.901370	-1.214753
8892	1.656950	0.141712	-0.075119	-0.272092	-0.059493	-0.469985	0.093003	-0.331393	1.201090	-0.520794

8893 rows × 10 columns

In [ ]:
```

```