```python
import pandas as pd
import numpy as np
from matplotlib import pyplot as plt
import seaborn as sns
from warnings import filterwarnings
filterwarnings("ignore")
!pip3 install ppscore
import ppscore as pps
#Import Library RobustScaler
from sklearn.preprocessing import RobustScaler
#Cluster Model
from sklearn.cluster import KMeans
from sklearn.metrics import silhouette_score
```

```
Collecting ppscore
  Downloading ppscore-1.2.0.tar.gz (47 kB)
     -------------------------------------- 47.1/47.1 kB 472.0 kB/s
eta 0:00:00
  Preparing metadata (setup.py): started
  Preparing metadata (setup.py): finished with status 'done'
Requirement already satisfied: pandas<2.0.0,>=1.0.0 in d:\anacondaa\
lib\site-packages (from ppscore) (1.3.4)
Requirement already satisfied: scikit-learn<1.0.0,>=0.20.2 in d:\
anacondaa\lib\site-packages (from ppscore) (0.24.2)
Requirement already satisfied: python-dateutil>=2.7.3 in d:\anacondaa\
lib\site-packages (from pandas<2.0.0,>=1.0.0->ppscore) (2.8.2)
Requirement already satisfied: numpy>=1.17.3 in d:\anacondaa\lib\site-
packages (from pandas<2.0.0,>=1.0.0->ppscore) (1.20.3)
Requirement already satisfied: pytz>=2017.3 in d:\anacondaa\lib\site-
packages (from pandas<2.0.0,>=1.0.0->ppscore) (2021.3)
Requirement already satisfied: scipy>=0.19.1 in d:\anacondaa\lib\site-
packages (from scikit-learn<1.0.0,>=0.20.2->ppscore) (1.7.1)
Requirement already satisfied: threadpoolctl>=2.0.0 in d:\anacondaa\
lib\site-packages (from scikit-learn<1.0.0,>=0.20.2->ppscore) (2.2.0)
Requirement already satisfied: joblib>=0.11 in d:\anacondaa\lib\site-
packages (from scikit-learn<1.0.0,>=0.20.2->ppscore) (1.1.0)
Requirement already satisfied: six>=1.5 in d:\anacondaa\lib\site-
packages (from python-dateutil>=2.7.3->pandas<2.0.0,>=1.0.0->ppscore)
(1.16.0)
Building wheels for collected packages: ppscore
  Building wheel for ppscore (setup.py): started
  Building wheel for ppscore (setup.py): finished with status 'done'
  Created wheel for ppscore: filename=ppscore-1.2.0-py2.py3-none-
any.whl size=13068
sha256=4cee543679fbcf1ee7258a6c5332b6b37f3c4b3120c8e827f996cf34ae2278c
2
  Stored in directory: c:\users\casper\appdata\local\pip\cache\wheels\
66\5f\af\a0de66f8359588661c0b1239580f4788dba33a4a1e504ef682
Successfully built ppscore
```

```
Installing collected packages: ppscore
Successfully installed ppscore-1.2.0
```

```
WARNING: Ignoring invalid distribution -ip (d:\anacondaa\lib\site-
packages)
WARNING: Ignoring invalid distribution -ip (d:\anacondaa\lib\site-
packages)
WARNING: Ignoring invalid distribution -ip (d:\anacondaa\lib\site-
packages)
WARNING: Ignoring invalid distribution -ip (d:\anacondaa\lib\site-
packages)
WARNING: Ignoring invalid distribution -ip (d:\anacondaa\lib\site-
packages)
WARNING: Ignoring invalid distribution -ip (d:\anacondaa\lib\site-
packages)
WARNING: Ignoring invalid distribution -ip (d:\anacondaa\lib\site-
packages)
```

```
#load_data
data = pd.read_csv('D:/G-PYTHON/Python 42/Data science/Data Science
Projects/App_Store_Data_Analysis/Dataset/AppleStore.csv' ,sep =',' ,
encoding = 'utf8' )
data.head()
```

```
   Unnamed: 0         id
track_name  \
0           1  281656475                                     PAC-MAN
Premium
1           2  281796108                          Evernote - stay
organized
2           3  281940292    WeatherBug - Local Weather, Radar, Maps,
Alerts
3           4  282614216  eBay: Best App to Buy, Sell, Save! Online
Shop...
4           5  282935706
Bible

   size_bytes currency  price  rating_count_tot  rating_count_ver  \
0   100788224      USD   3.99             21292                26
1   158578688      USD   0.00            161065                26
2   100524032      USD   0.00            188583              2822
3   128512000      USD   0.00            262241               649
4    92774400      USD   0.00            985920              5320

   user_rating  user_rating_ver     ver cont_rating   prime_genre  \
0          4.0              4.5   6.3.5          4+         Games
1          4.0              3.5   8.2.2          4+  Productivity
2          3.5              4.5   5.0.0          4+       Weather
3          4.0              4.5  5.10.0         12+      Shopping
4          4.5              5.0   7.5.1          4+     Reference
```

```
   sup_devices.num  ipadSc_urls.num  lang.num  vpp_lic
0               38                5        10        1
1               37                5        23        1
2               37                5         3        1
3               37                5         9        1
4               37                5        45        1
```

```
#drop column (Unnamed) as semiler ID column
data.drop(['Unnamed: 0'], axis=1 ,inplace=True)
#show data after drop
data.head(2)
```

```
          id                 track_name  size_bytes currency  price  \
0  281656475            PAC-MAN Premium   100788224      USD   3.99   
1  281796108  Evernote - stay organized   158578688      USD   0.00   

   rating_count_tot  rating_count_ver  user_rating  user_rating_ver
ver  \
0             21292                26          4.0              4.5
6.3.5   
1            161065                26          4.0              3.5
8.2.2   

  cont_rating   prime_genre  sup_devices.num  ipadSc_urls.num
lang.num  \
0          4+         Games               38                5
10   
1          4+  Productivity               37                5
23   

   vpp_lic  
0        1  
1        1  
```

```
#data about data
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7197 entries, 0 to 7196
Data columns (total 16 columns):
 #   Column            Non-Null Count  Dtype  
---  ------            --------------  -----  
 0   id                7197 non-null   int64  
 1   track_name        7197 non-null   object 
 2   size_bytes        7197 non-null   int64  
 3   currency          7197 non-null   object 
 4   price             7197 non-null   float64
 5   rating_count_tot  7197 non-null   int64  
 6   rating_count_ver  7197 non-null   int64  
 7   user_rating       7197 non-null   float64
 8   user_rating_ver   7197 non-null   float64
 9   ver               7197 non-null   object 
 10  cont_rating       7197 non-null   object 
 11  prime_genre       7197 non-null   object 
 12  sup_devices.num   7197 non-null   int64  
 13  ipadSc_urls.num   7197 non-null   int64  
 14  lang.num          7197 non-null   int64  
 15  vpp_lic           7197 non-null   int64  
dtypes: float64(3), int64(8), object(5)
memory usage: 899.8+ KB
```

```
#show shape of data 7197 Row and 16 columns
data.shape
```

```
(7197, 16)
```

```
data.isnull().sum().sum()
#not found null data
```

```
0
```

```
data.currency.value_counts()
#All of Apps has same currency paid
```

```
USD    7197
Name: currency, dtype: int64
```

```
data.nunique()
#target maybe vpp_lic
```

```
id                  7197
track_name          7195
size_bytes          7107
currency               1
price                 36
rating_count_tot    3185
rating_count_ver    1138
user_rating           10
user_rating_ver       10
ver                 1590
cont_rating            4
prime_genre           23
sup_devices.num       20
ipadSc_urls.num        6
lang.num              57
vpp_lic                2
dtype: int64
```

Exploratory Data Analaysis

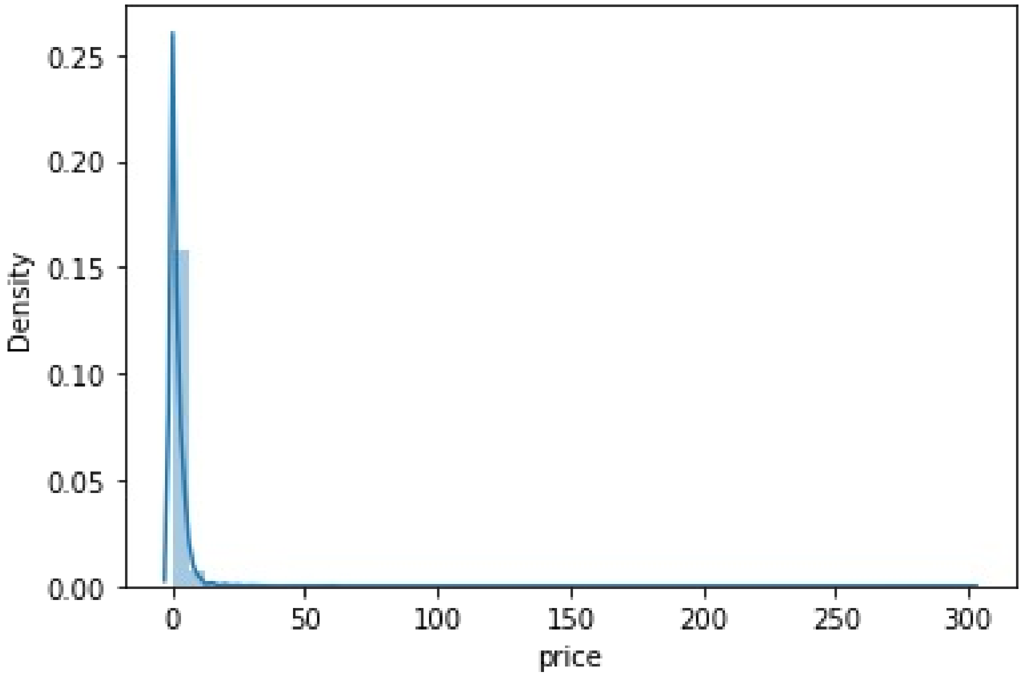
How do you visualize price distribution of paid apps ?

```
data.price.value_counts()
#4056 free apps
#another apps is paid
```

```
0.00      4056
0.99       728
2.99       683
1.99       621
4.99       394
3.99       277
6.99       166
9.99        81
5.99        52
7.99        33
14.99       21
19.99       13
8.99         9
24.99        8
29.99        6
13.99        6
11.99        6
12.99        5
15.99        4
17.99        3
59.99        3
39.99        2
20.99        2
23.99        2
49.99        2
22.99        2
27.99        2
16.99        2
299.99       1
21.99        1
47.99        1
99.99        1
74.99        1
34.99        1
18.99        1
249.99       1
Name: price, dtype: int64
```

```
sns.distplot(data.price)
```

```
<AxesSubplot:xlabel='price', ylabel='Density'>
```



```
free_apps = data[(data.price==0.00)]

paid_apps  = data[(data.price>0)]

free_apps.head(10)
```

```
           id                                         track_name  
size_bytes  \
1   281796108                          Evernote - stay organized  
158578688
2   281940292    WeatherBug - Local Weather, Radar, Maps, Alerts  
100524032
3   282614216  eBay: Best App to Buy, Sell, Save! Online Shop...  
128512000
4   282935706                                              Bible  
92774400
6   283646709             PayPal - Send and request money safely  
227795968
7   284035177                            Pandora - Music & Radio  
130242560
12  284815942               Google – Search made just for mobile  
179979264
13  284847138                   Bank of America - Mobile Banking  
160925696
15  284876795             TripAdvisor Hotels Flights Restaurants  
207907840
16  284882215                                           Facebook  
389879808
```

```
   currency  price  rating_count_tot  rating_count_ver  user_rating  \
1       USD    0.0            161065                26          4.0   
2       USD    0.0            188583              2822          3.5   
3       USD    0.0            262241               649          4.0   
4       USD    0.0            985920              5320          4.5   
6       USD    0.0            119487               879          4.0   
7       USD    0.0           1126879              3594          4.0   
12      USD    0.0            479440               203          3.5   
13      USD    0.0            119773              2336          3.5   
15      USD    0.0             56194                87          4.0   
16      USD    0.0           2974676               212          3.5   

    user_rating_ver     ver cont_rating        prime_genre  
sup_devices.num  \
1               3.5   8.2.2          4+       Productivity   
37   
2               4.5   5.0.0          4+            Weather   
37   
3               4.5  5.10.0         12+           Shopping   
37   
4               5.0   7.5.1          4+          Reference   
37   
6               4.5  6.12.0          4+            Finance   
37   
7               4.5   8.4.1         12+              Music   
37   
12              4.0    27.0         17+          Utilities   
37   
13              4.5   7.3.8          4+            Finance   
37   
15              3.5    21.1          4+             Travel   
37   
16              3.5    95.0          4+  Social Networking   
37   

    ipadSc_urls.num  lang.num  vpp_lic  
1                 5        23        1  
2                 5         3        1  
3                 5         9        1  
4                 5        45        1  
6                 0        19        1  
7                 4         1        1  
12                4        33        1  
13                0         2        1  
15                1        26        1  
16                1        29        1  
```

```
paid_apps.head(10)
```

```
           id                     track_name  size_bytes currency  
price  \
0   281656475                PAC-MAN Premium   100788224      USD  
3.99   
5   283619399               Shanghai Mahjong    10485713      USD  
0.99   
8   284666222    PCalc - The Best Calculator    49250304      USD  
9.99   
9   284736660                    Ms. PAC-MAN    70023168      USD  
3.99   
10  284791396      Solitaire by MobilityWare    49618944      USD  
4.99   
11  284815117               SCRABBLE Premium   227547136      USD  
7.99   
14  284862767                       FreeCell    55153664      USD  
4.99   
19  285005463  Crash Bandicoot Nitro Kart 3D    10735026      USD  
2.99   
20  285946052                         iQuran    70707916      USD  
1.99   
21  285994151                    :) Sudoku +     6169600      USD  
2.99   

    rating_count_tot  rating_count_ver  user_rating  user_rating_ver  
ver  \
0              21292                26          4.0              4.5  
6.3.5   
5               8253              5516          4.0              4.0  
1.8   
8               1117                 4          4.5              5.0  
3.6.6   
9               7885                40          4.0              4.0  
4.0.4   
10             76720              4017          4.5              4.5  
4.10.1   
11            105776               166          3.5              2.5  
5.19.0   
14              6340               668          4.5              4.5  
4.0.3   
19             31456              4178          4.0              3.5  
1.0.0   
20              2929               966          4.5              4.5  
3.3   
21             11447               781          5.0              5.0  
5.2.6   

   cont_rating prime_genre  sup_devices.num  ipadSc_urls.num  lang.num  
\
0           4+       Games               38                5        10  
```

```
5           4+       Games               47                5         1

8           4+   Utilities               37                5         1

9           4+       Games               38                0        10

10          4+       Games               38                4        11

11          4+       Games               37                0         6

14          4+       Games               38                5         2

19          4+       Games               47                0         1

20          4+   Reference               43                0         2

21          4+       Games               40                5         1
```

```
    vpp_lic
0         1
5         1
8         1
9         1
10        1
11        1
14        1
19        1
20        1
21        1
```

```
paid_apps.price.value_counts()
```

```
0.99      728
2.99      683
1.99      621
4.99      394
3.99      277
6.99      166
9.99       81
5.99       52
7.99       33
14.99      21
19.99      13
8.99        9
24.99       8
29.99       6
13.99       6
11.99       6
12.99       5
```

```
15.99       4
17.99       3
59.99       3
39.99       2
20.99       2
23.99       2
49.99       2
22.99       2
27.99       2
16.99       2
299.99      1
21.99       1
47.99       1
99.99       1
74.99       1
34.99       1
18.99       1
249.99      1
Name: price, dtype: int64
```

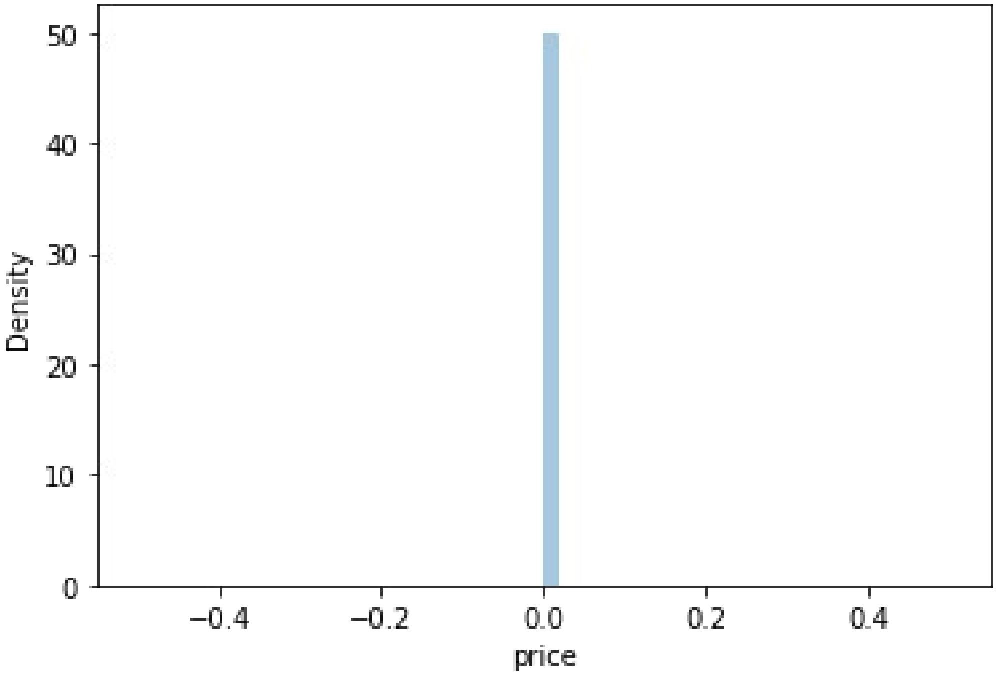
The number of apps decreases with increasing his price

```
free_apps.price.value_counts()
```

```
0.0    4056
Name: price, dtype: int64
```

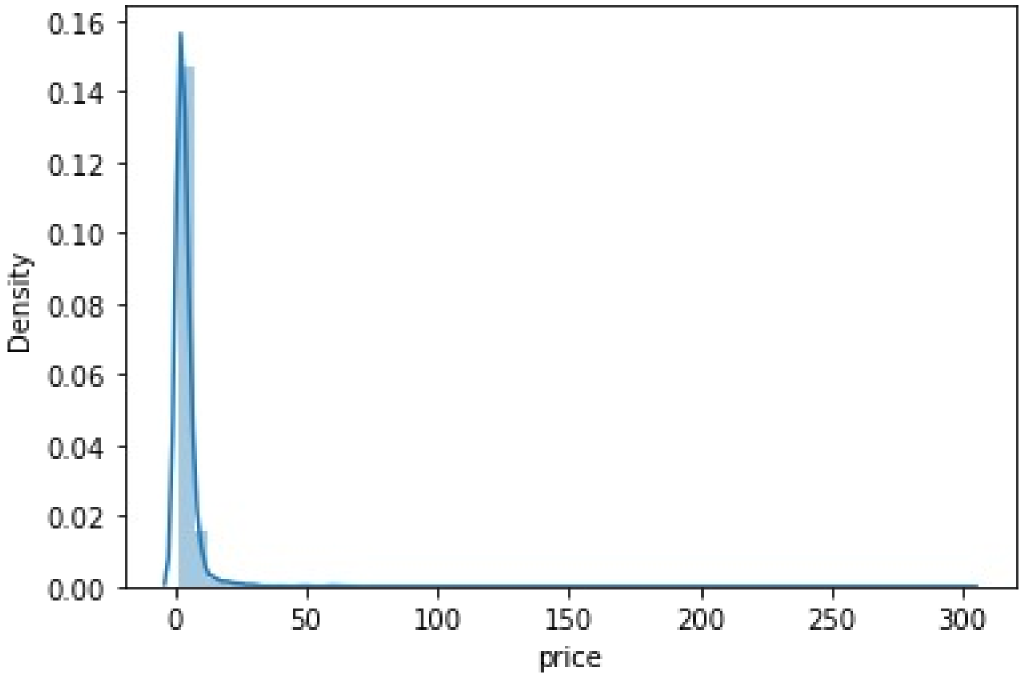
```
sns.distplot(free_apps['price'])
```

```
<AxesSubplot:xlabel='price', ylabel='Density'>
```



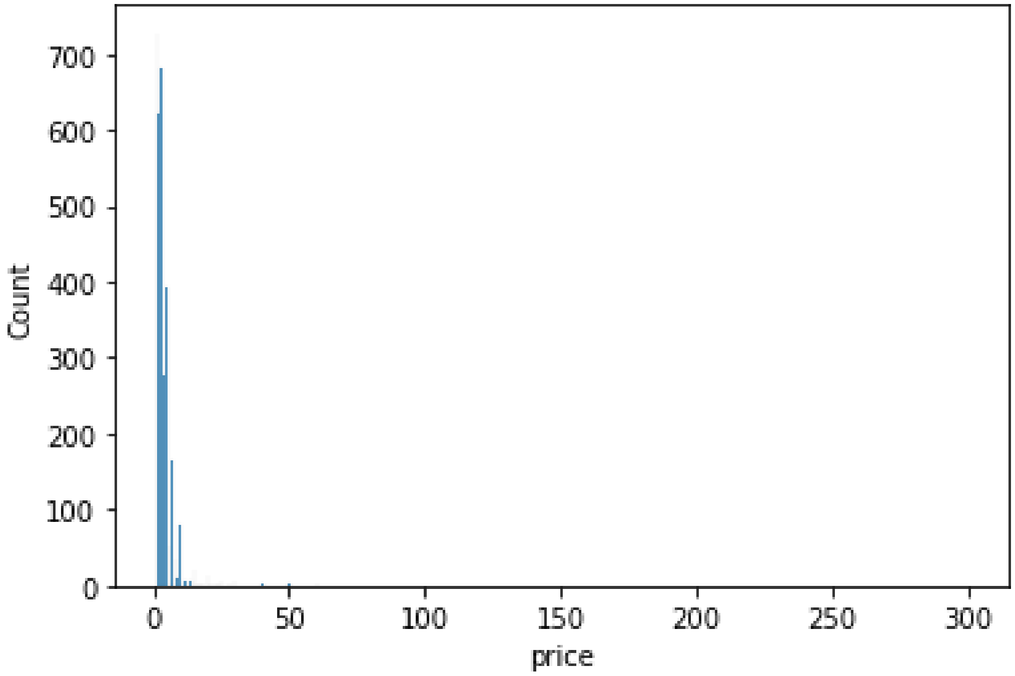
```
sns.distplot(paid_apps['price'])
```

```
<AxesSubplot:xlabel='price', ylabel='Density'>
```

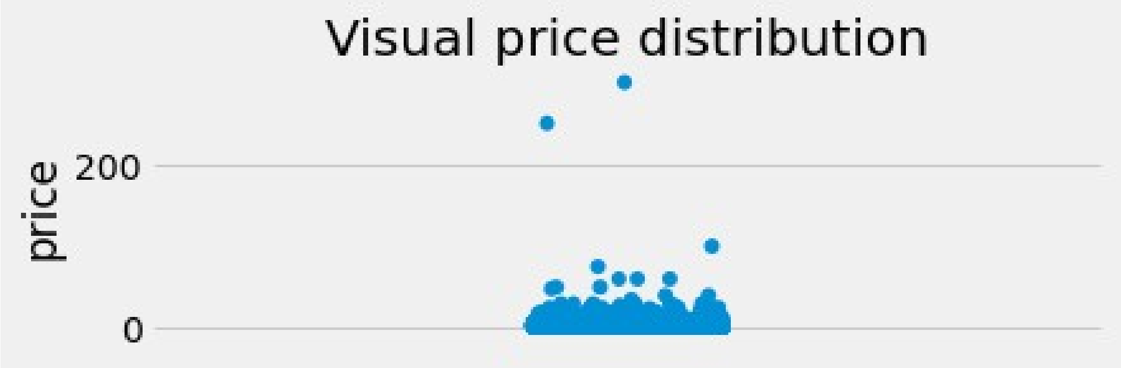


```
sns.histplot(paid_apps['price'])
```

```
<AxesSubplot:xlabel='price', ylabel='Count'>
```



```
plt.style.use('fivethirtyeight')
plt.figure(figsize=(6,4))

plt.subplot(2,1,2)
plt.title('Visual price distribution')
sns.stripplot(data=paid_apps,y='price',jitter= True,orient =
'h' ,size=6)
plt.show()
```



from this graph The number of apps that have a price greater than 50 is few compared to before 50 USD

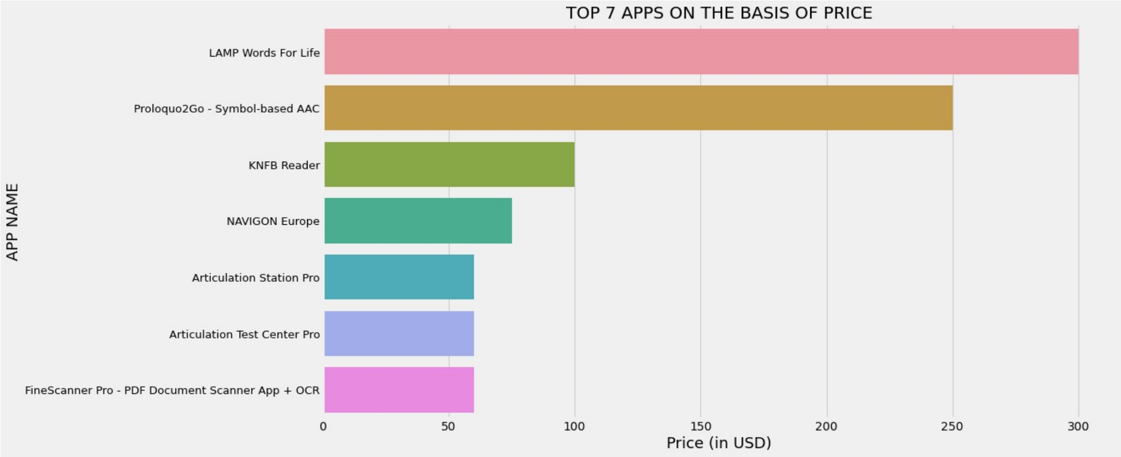
```
Top_Apps=paid_apps[paid_apps.price>50]
[['track_name','price','prime_genre','user_rating']]
Top_Apps
#7 Top apps with price, prime_genre and user rating
```

```
                                          track_name   price
prime_genre  \
115                    Proloquo2Go - Symbol-based AAC  249.99
Education
162                                   NAVIGON Europe   74.99
1136                         Articulation Station Pro   59.99
Education
1479                               LAMP Words For Life  299.99
Education
2181                     Articulation Test Center Pro   59.99
Education
2568                                       KNFB Reader   99.99
Productivity
3238  FineScanner Pro - PDF Document Scanner App + OCR   59.99
Business

      user_rating
115           4.0
162           3.5
1136          4.5
1479          4.0
```

```
2181         4.5
2568         4.5
3238         4.0
```

Top 7 apps on the basis of price

```
#Function for visualizaiton
def visualizer(x, y, plot_type, title, xlabel, ylabel, rotation=False, 
rotation_value=60, figsize=(15,8)):
    plt.figure(figsize=figsize)
    
    if plot_type == "bar":
        sns.barplot(x=x, y=y)
    elif plot_type == "count":
        sns.countplot(x)
        
    plt.title(title, fontsize=20)
    plt.xlabel(xlabel, fontsize=18)
    plt.ylabel(ylabel, fontsize=18)
    plt.yticks(fontsize=13)
    if rotation == True:
        plt.xticks(fontsize=13,rotation=rotation_value)
    plt.show()

Top_Apps = Top_Apps.sort_values('price', ascending=False)

visualizer(Top_Apps.price,Top_Apps.track_name, "bar", "TOP 7 APPS ON 
THE BASIS OF PRICE","Price (in USD)","APP NAME")
#names of track in y axis to be readable
```



```
paid_apps.head(5)

          id                     track_name  size_bytes currency  price 
\
0  281656475                PAC-MAN Premium   100788224      USD   3.99 

5  283619399               Shanghai Mahjong    10485713      USD   0.99 
```

```
8   284666222  PCalc - The Best Calculator    49250304      USD   9.99

9   284736660                  Ms. PAC-MAN    70023168      USD   3.99

10  284791396    Solitaire by MobilityWare    49618944      USD   4.99


    rating_count_tot  rating_count_ver  user_rating  user_rating_ver
ver  \
0              21292                26          4.0              4.5
6.3.5
5               8253              5516          4.0              4.0
1.8
8               1117                 4          4.5              5.0
3.6.6
9               7885                40          4.0              4.0
4.0.4
10             76720              4017          4.5              4.5
4.10.1

   cont_rating prime_genre  sup_devices.num  ipadSc_urls.num  lang.num
\
0           4+      Games               38                5        10

5           4+      Games               47                5         1

8           4+  Utilities               37                5         1

9           4+      Games               38                0        10

10          4+      Games               38                4        11


    vpp_lic
0         1
5         1
8         1
9         1
10        1
```

```
#sum of all paid apps
sum_paid = paid_apps.price.value_counts().sum()
sum_paid
```

```
3141
```

```
#sum of all free apps
sum_free = free_apps.price.value_counts().sum()
sum_free
```

```
4056
```

How does the price distribution get affected by category ?

```
data.prime_genre.value_counts()
```

```
Games                3862
Entertainment         535
Education             453
Photo & Video         349
Utilities             248
Health & Fitness      180
Productivity          178
Social Networking     167
Lifestyle             144
Music                 138
Shopping              122
Sports                114
Book                  112
Finance               104
Travel                 81
News                   75
Weather                72
Reference              64
Food & Drink           63
Business               57
Navigation             46
Medical                23
Catalogs               10
Name: prime_genre, dtype: int64
```

Top app category is Games Games # is 3862 and Entertainment # is 535

```
data.head()
```

```
          id                                         track_name
size_bytes  \
0  281656475                                    PAC-MAN Premium
100788224
1  281796108                          Evernote - stay organized
158578688
2  281940292    WeatherBug - Local Weather, Radar, Maps, Alerts
100524032
3  282614216  eBay: Best App to Buy, Sell, Save! Online Shop...
128512000
4  282935706                                              Bible
92774400

  currency  price  rating_count_tot  rating_count_ver  user_rating  \
0      USD   3.99             21292                26          4.0
1      USD   0.00            161065                26          4.0
```

```
2      USD   0.00            188583              2822          3.5
3      USD   0.00            262241               649          4.0
4      USD   0.00            985920              5320          4.5

   user_rating_ver     ver cont_rating   prime_genre  sup_devices.num
\
0              4.5   6.3.5          4+         Games               38

1              3.5   8.2.2          4+  Productivity               37

2              4.5   5.0.0          4+       Weather               37

3              4.5  5.10.0         12+      Shopping               37

4              5.0   7.5.1          4+     Reference               37


   ipadSc_urls.num  lang.num  vpp_lic
0                5        10        1
1                5        23        1
2                5         3        1
3                5         9        1
4                5        45        1
```

```
new_data_cate = data.groupby([data.prime_genre])
[['id']].count().reset_index().sort_values('id' ,ascending = False)
new_data_cate.columns = ['prime_genre','# of Apps']
new_data_cate.head()
#Categories and number of apps in each category
```

```
      prime_genre  # of Apps
7           Games       3862
4   Entertainment        535
3       Education        453
14  Photo & Video        349
21      Utilities        248
```

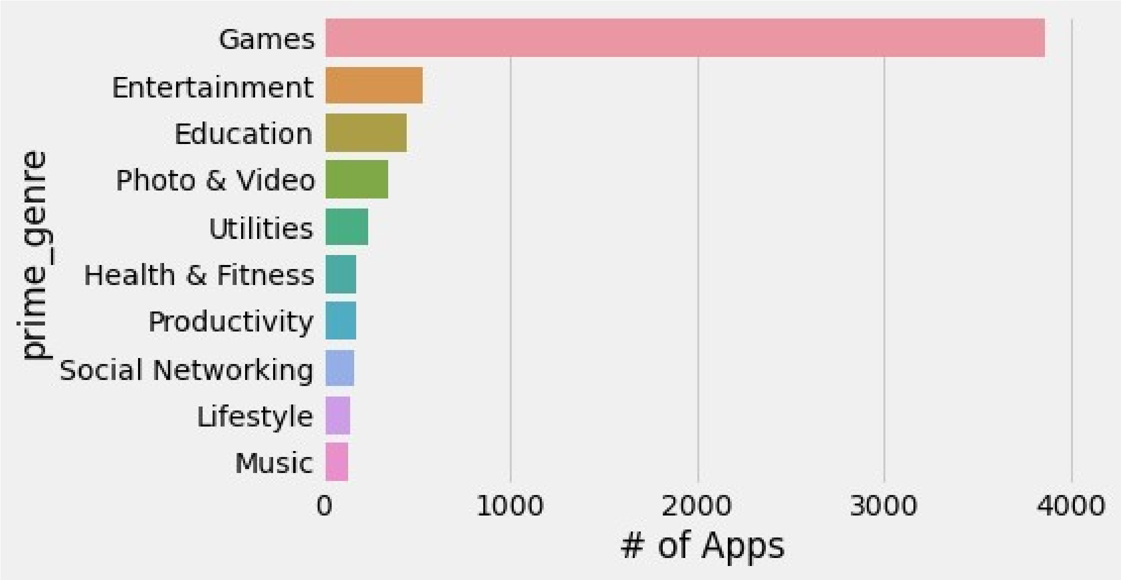
```
#Top_Categories accorrding number of apps
new_data_cate.head(10)
```

```
          prime_genre  # of Apps
7               Games       3862
4       Entertainment        535
3           Education        453
14      Photo & Video        349
21          Utilities        248
8    Health & Fitness        180
15       Productivity        178
18  Social Networking        167
```

```
9           Lifestyle        144
11              Music        138
```

```
sns.barplot(y = 'prime_genre',x = '# of Apps',
data=new_data_cate.head(10))
```

```
<AxesSubplot:xlabel='# of Apps', ylabel='prime_genre'>
```

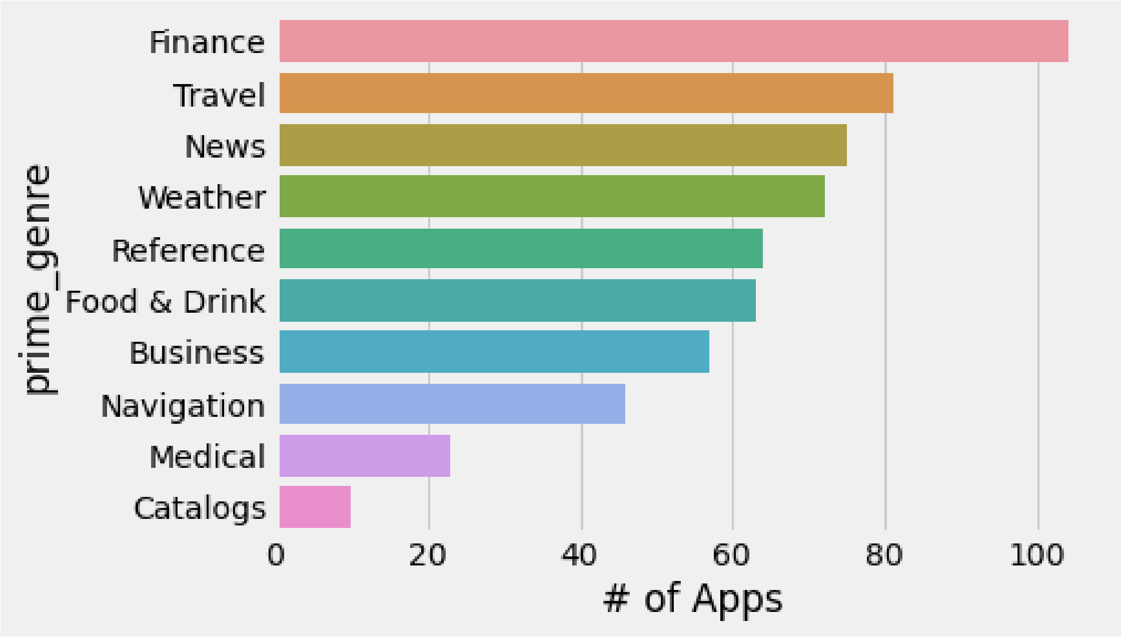


```
#Lower Categories according number of apps Categories unpopular
new_data_cate.tail(10)
```

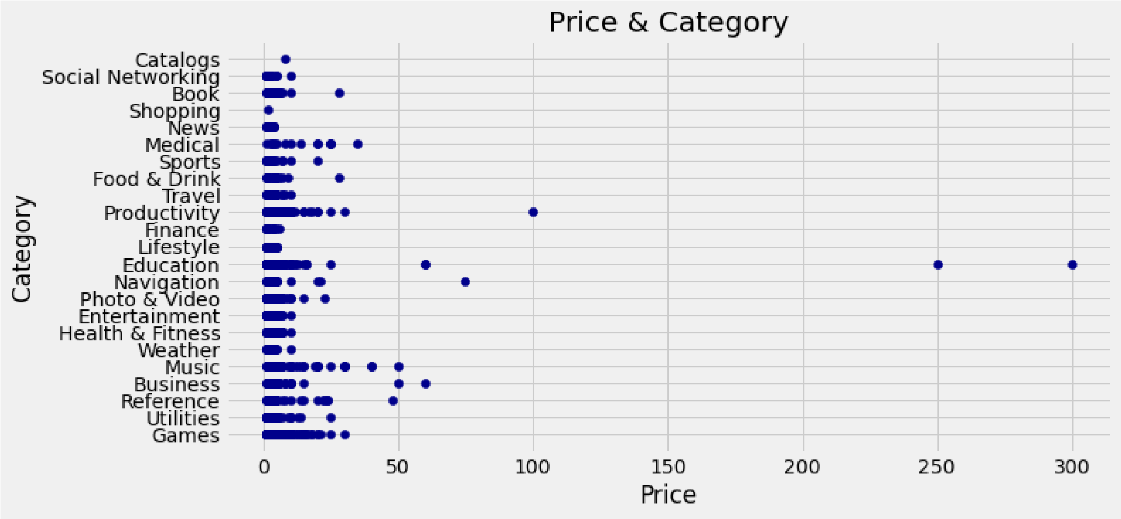
```
     prime_genre  # of Apps
5        Finance        104
20        Travel         81
13          News         75
22       Weather         72
16     Reference         64
6   Food & Drink         63
1       Business         57
12    Navigation         46
10       Medical         23
2       Catalogs         10
```

```
sns.barplot(x= '# of Apps' , y = 'prime_genre' , data =
new_data_cate.tail(10))
```

```
<AxesSubplot:xlabel='# of Apps', ylabel='prime_genre'>
```



```
plt.figure(figsize=(10,5))
plt.scatter(y=paid_apps.prime_genre ,x=paid_apps.price,c='DarkBlue')
plt.title('Price & Category')
plt.xlabel('Price')
plt.ylabel('Category')
plt.show()
```



Top Price in important Category (Business , Navigation , Education , Productivity)

in another side price for all of apps less than 50 USD

Education Apps has a higher price

Shopping Apps has a lower price

What about paid apps Vs Free apps ?

```
free_apps.head(3)
```

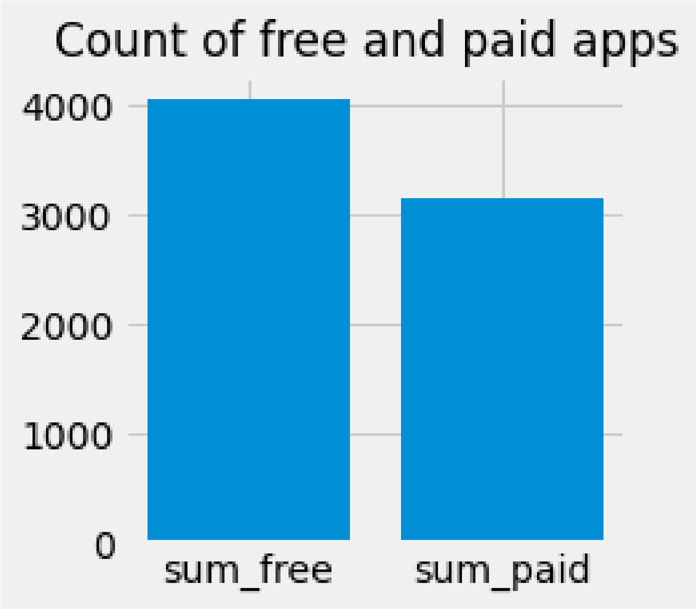
```
          id                                         track_name
size_bytes  \
1  281796108                          Evernote - stay organized
158578688
2  281940292    WeatherBug - Local Weather, Radar, Maps, Alerts
100524032
3  282614216  eBay: Best App to Buy, Sell, Save! Online Shop...
128512000

  currency  price  rating_count_tot  rating_count_ver  user_rating  \
1      USD    0.0            161065                26          4.0
2      USD    0.0            188583              2822          3.5
3      USD    0.0            262241               649          4.0

   user_rating_ver     ver cont_rating   prime_genre  sup_devices.num
\
1              3.5   8.2.2          4+  Productivity               37

2              4.5   5.0.0          4+       Weather               37

3              4.5  5.10.0         12+      Shopping               37


   ipadSc_urls.num  lang.num  vpp_lic
1                5        23        1
2                5         3        1
3                5         9        1
```

```
paid_apps.head(3)
```

```
          id                    track_name  size_bytes currency  price
\
0  281656475              PAC-MAN Premium   100788224      USD   3.99

5  283619399             Shanghai Mahjong    10485713      USD   0.99

8  284666222  PCalc - The Best Calculator    49250304      USD   9.99


   rating_count_tot  rating_count_ver  user_rating  user_rating_ver
ver  \
0             21292                26          4.0              4.5
6.3.5
5              8253              5516          4.0              4.0
1.8
8              1117                 4          4.5              5.0
3.6.6
```

```
  cont_rating prime_genre  sup_devices.num  ipadSc_urls.num  lang.num
vpp_lic
0          4+       Games               38                5        10
1
5          4+       Games               47                5         1
1
8          4+   Utilities               37                5         1
1
```

```
names = ['sum_free', 'sum_paid']
values = [sum_free, sum_paid]
plt.figure(figsize=(3, 3))
plt.suptitle('Count of free and paid apps')
plt.bar(names, values)
plt.show()
```



```
print('number of Catigories in free apps is' ,
len(free_apps.prime_genre.value_counts().index))
print('number of Catigories in paid apps is' ,
len(paid_apps.prime_genre.value_counts().index))
#all categories has free & paid apps
```

```
number of Catigories in free apps is 23
number of Catigories in paid apps is 23
```

```
free_apps.head()
```

```
          id                                         track_name
size_bytes  \
1  281796108                          Evernote - stay organized
158578688
2  281940292    WeatherBug - Local Weather, Radar, Maps, Alerts
```

```
100524032
3  282614216  eBay: Best App to Buy, Sell, Save! Online Shop...
128512000
4  282935706                                              Bible
92774400
6  283646709             PayPal - Send and request money safely
227795968

  currency  price  rating_count_tot  rating_count_ver  user_rating  \
1      USD    0.0            161065                26          4.0
2      USD    0.0            188583              2822          3.5
3      USD    0.0            262241               649          4.0
4      USD    0.0            985920              5320          4.5
6      USD    0.0            119487               879          4.0

   user_rating_ver     ver cont_rating   prime_genre  sup_devices.num
\
1              3.5   8.2.2          4+  Productivity               37

2              4.5   5.0.0          4+       Weather               37

3              4.5  5.10.0         12+      Shopping               37

4              5.0   7.5.1          4+     Reference               37

6              4.5  6.12.0          4+       Finance               37


   ipadSc_urls.num  lang.num  vpp_lic
1                5        23        1
2                5         3        1
3                5         9        1
4                5        45        1
6                0        19        1
```

```
free = free_apps.prime_genre.value_counts().sort_index().to_frame()
paid = paid_apps.prime_genre.value_counts().sort_index().to_frame()
total = data.prime_genre.value_counts().sort_index().to_frame()
free.columns=['free']
paid.columns=['paid']
total.columns=['total']
fig  =free.join(paid).join(total)
fig['%paid'] = fig.paid*100 /fig.total
fig['%free'] = fig.free*100/ fig.total
fig
```

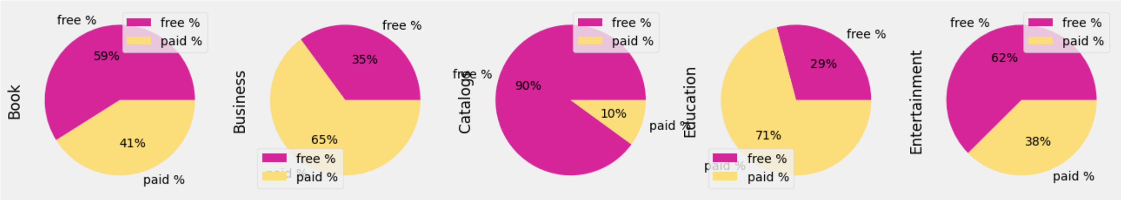
```
                   free  paid  total      %paid      %free
Book                 66    46    112  41.071429  58.928571
Business             20    37     57  64.912281  35.087719
Catalogs              9     1     10  10.000000  90.000000
```

```
Education            132   321    453  70.860927  29.139073
Entertainment        334   201    535  37.570093  62.429907
Finance               84    20    104  19.230769  80.769231
Food & Drink          43    20     63  31.746032  68.253968
Games               2257  1605   3862  41.558778  58.441222
Health & Fitness      76   104    180  57.777778  42.222222
Lifestyle             94    50    144  34.722222  65.277778
Medical                8    15     23  65.217391  34.782609
Music                 67    71    138  51.449275  48.550725
Navigation            20    26     46  56.521739  43.478261
News                  58    17     75  22.666667  77.333333
Photo & Video        167   182    349  52.148997  47.851003
Productivity          62   116    178  65.168539  34.831461
Reference             20    44     64  68.750000  31.250000
Shopping             121     1    122   0.819672  99.180328
Social Networking    143    24    167  14.371257  85.628743
Sports                79    35    114  30.701754  69.298246
Travel                56    25     81  30.864198  69.135802
Utilities            109   139    248  56.048387  43.951613
Weather               31    41     72  56.944444  43.055556
```

of paid apps greater than # of free apps

```
# for pie chart
pies = fig[['%free','%paid']].head()
pies.columns=['free %','paid %']

plt.figure(figsize=(15,10))
pies.T.plot.pie(subplots=True,figsize=(20,4),colors=['#D62598','#FBDD7A'],autopct = '%1.0f%%')
plt.show()
```

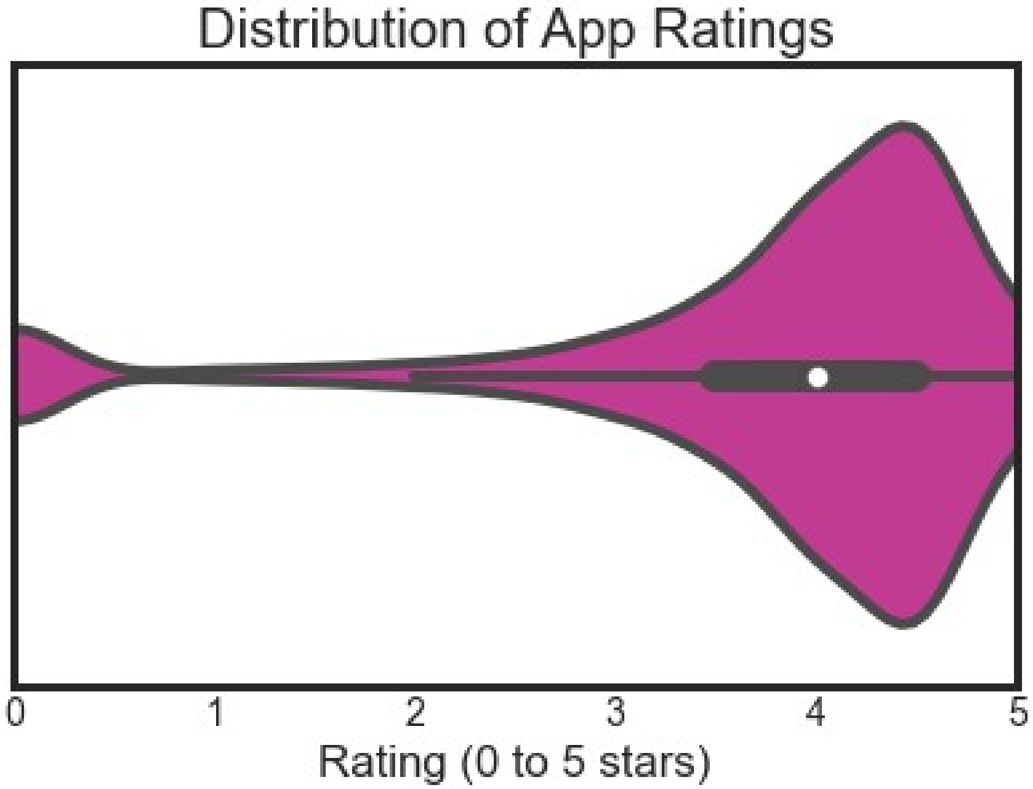
```
<Figure size 1080x720 with 0 Axes>
```



```
data[data['rating_count_tot']==data['rating_count_tot'].max()]
#Most rated & highest total rating for all version app:
```

```
           id track_name  size_bytes currency  price  rating_count_tot  \
16  284882215   Facebook   389879808      USD    0.0           2974676


    rating_count_ver  user_rating  user_rating_ver   ver cont_rating  \
16               212          3.5              3.5  95.0          4+
          prime_genre  sup_devices.num  ipadSc_urls.num  lang.num
vpp_lic
16  Social Networking               37                1        29
1
```

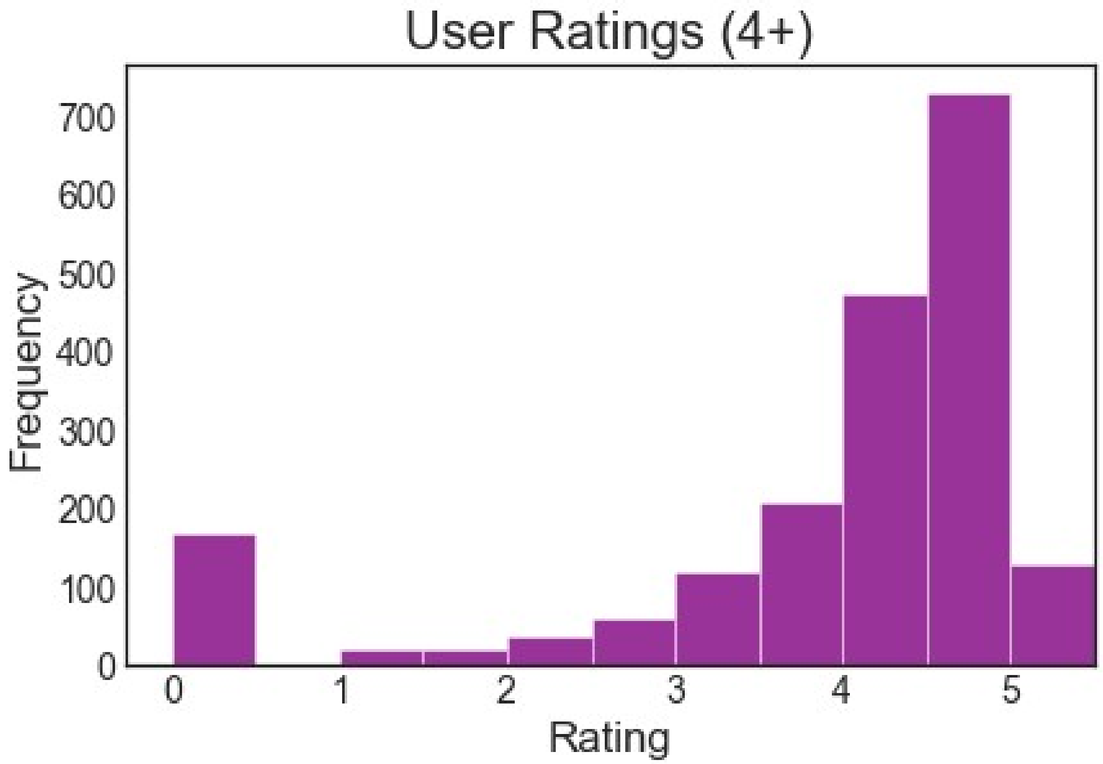
```
sns.set_style('white')
sns.violinplot(x=paid_apps['user_rating'],color='#D62598')
plt.xlim(0,5)
plt.xlabel('Rating (0 to 5 stars)')
_ = plt.title('Distribution of App Ratings')
```



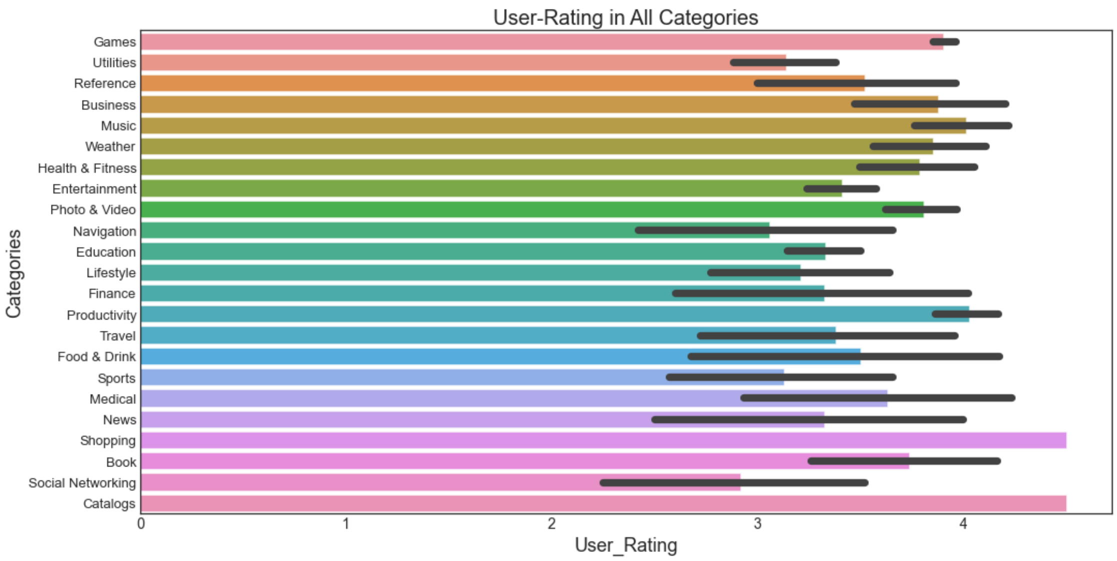
```
paid_apps.cont_rating.value_counts()
```

```
4+     1967
9+      549
12+     450
17+     175
Name: cont_rating, dtype: int64
```

```
bins = (0,0.5,1,1.5,2,2.5,3,3.5,4,4.5,5,5.5)
plt.style.use('seaborn-white')
plt.hist(paid_apps[paid_apps['cont_rating']=='4+']
['user_rating'],alpha=.8,bins=bins,color='purple')
plt.xticks((0,1,2,3,4,5))
plt.title('User Ratings (4+)')
plt.xlabel('Rating')
plt.ylabel('Frequency')
_ = plt.xlim(right=5.5)
```

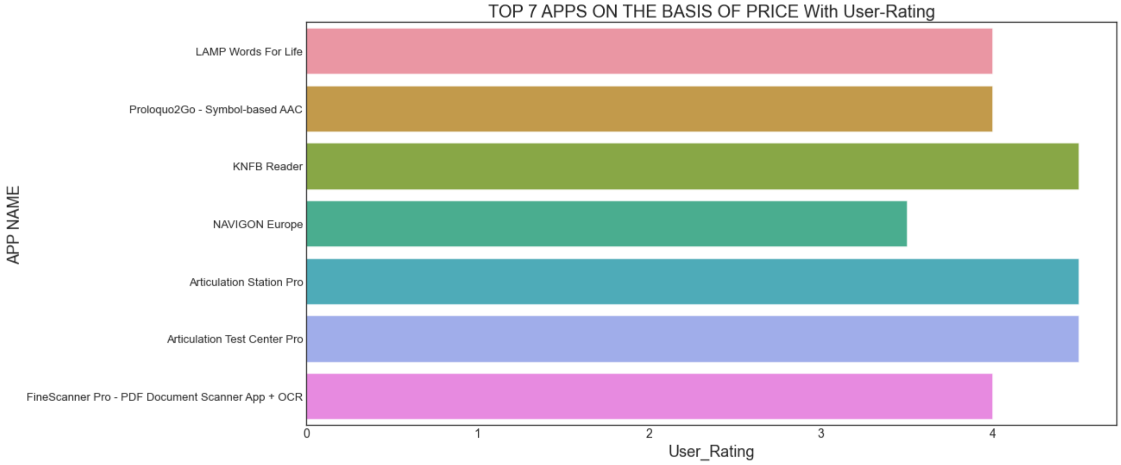


```
visualizer(paid_apps['user_rating'],paid_apps.prime_genre, "bar",
"User-Rating in All Categories","User_Rating","Categories")
```



```
Top_Apps = Top_Apps.sort_values('price', ascending=False)
```

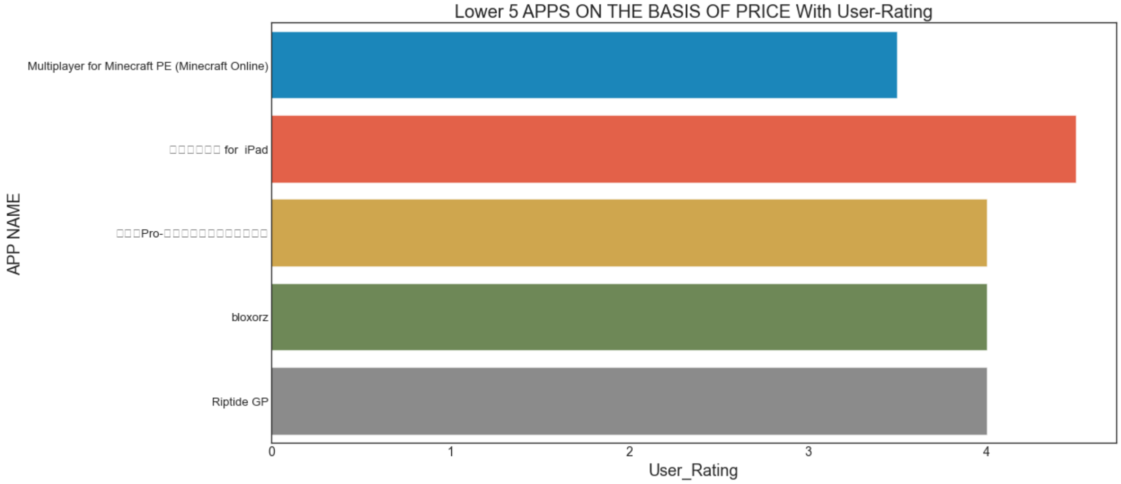
```
visualizer(Top_Apps.user_rating,Top_Apps.track_name, "bar", "TOP 7
APPS ON THE BASIS OF PRICE With User-Rating","User_Rating","APP NAME")
#names of track in y axis to be readable
```



```
Lower_Apps=paid_apps[paid_apps.price<=50]
[['track_name','price','prime_genre','user_rating']]
Lower_Apps.head()
```

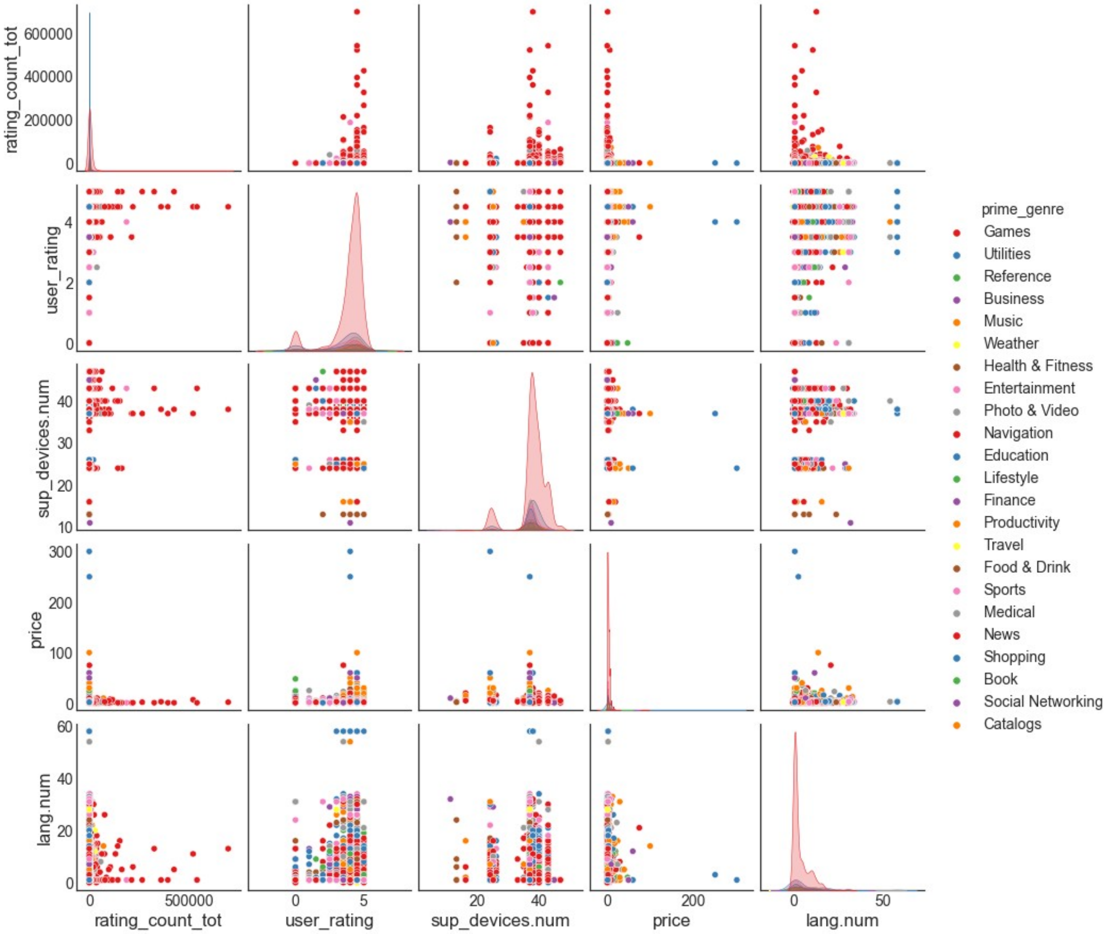
```
                     track_name  price prime_genre  user_rating
0                PAC-MAN Premium   3.99       Games          4.0
5               Shanghai Mahjong   0.99       Games          4.0
8    PCalc - The Best Calculator   9.99   Utilities          4.5
9                    Ms. PAC-MAN   3.99       Games          4.0
10    Solitaire by MobilityWare   4.99       Games          4.5
```

```
Lower_Apps = Lower_Apps.sort_values('price', ascending=True)
lower = Lower_Apps.head()
visualizer(lower.user_rating,lower.track_name, "bar", "Lower 5 APPS ON
THE BASIS OF PRICE With User-Rating","User_Rating","APP NAME")
```



```
numCol = paid_apps[['rating_count_tot', 'user_rating',
'sup_devices.num', 'price','lang.num', 'prime_genre']]
sns.pairplot(data = numCol, hue='prime_genre',palette='Set1')
```

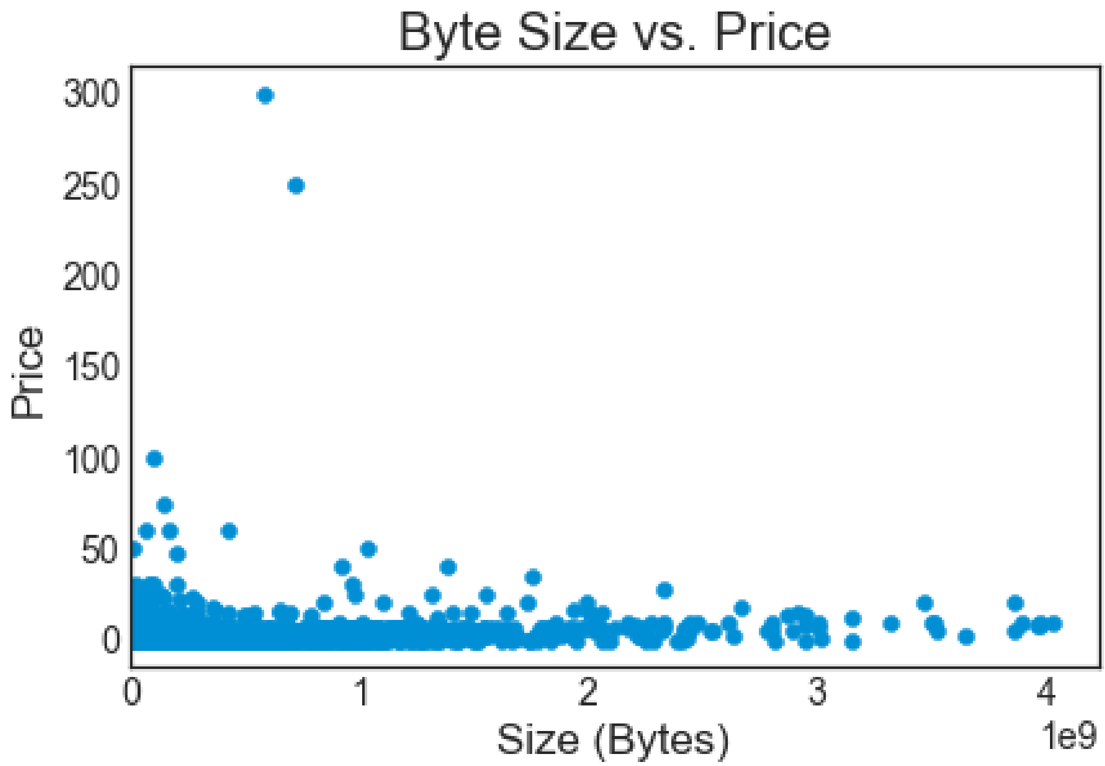
```
<seaborn.axisgrid.PairGrid at 0x1cf866dd340>
```



As the size of the app increases do they get pricier ?

```
plt.style.use('seaborn-white')
plt.scatter(data['size_bytes'],data['price'])
plt.title('Byte Size vs. Price')
plt.xlabel('Size (Bytes)')
plt.ylabel('Price')
plt.xlim(0)
```

```
(0.0, 4227238656.0)
```



size of App not corelated with price

we show that if size is big ,price is low

the value of an app to the user isn't necessarily related to its size.

How are the apps distributed category wise ? Can we split by paid category ?

```
grp = paid_apps.groupby('prime_genre')
x = grp['user_rating'].agg(np.mean)

y = grp['price'].agg(np.sum)
z = grp['user_rating_ver'].agg(np.mean)
print(x)
print(y)
print(z)
```

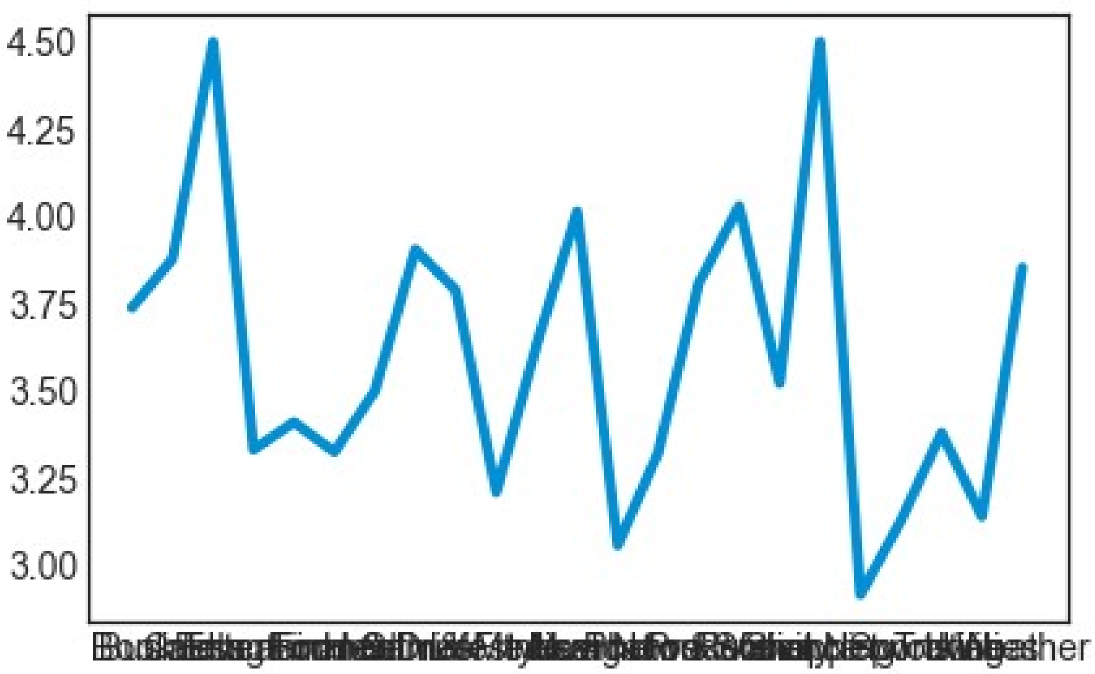
```
prime_genre
Book                    3.739130
Business                3.878378
Catalogs                4.500000
Education               3.331776
Entertainment           3.410448
Finance                 3.325000
Food & Drink            3.500000
Games                   3.904984
Health & Fitness        3.788462
Lifestyle               3.210000
```

```
Medical              3.633333
Music                4.014085
Navigation           3.057692
News                 3.323529
Photo & Video        3.807692
Productivity         4.030172
Reference            3.522727
Shopping             4.500000
Social Networking    2.916667
Sports               3.128571
Travel               3.380000
Utilities            3.140288
Weather              3.853659
Name: user_rating, dtype: float64
prime_genre
Book                  200.54
Business              291.63
Catalogs                7.99
Education            1824.79
Entertainment         475.99
Finance                43.80
Food & Drink           97.80
Games                5533.95
Health & Fitness      344.96
Lifestyle             127.50
Medical               201.85
Music                 667.29
Navigation            189.74
News                   38.83
Photo & Video         514.18
Productivity          770.84
Reference             309.56
Shopping                1.99
Social Networking      56.76
Sports                108.65
Travel                 90.75
Utilities             408.61
Weather               115.59
Name: price, dtype: float64
prime_genre
Book                 3.163043
Business             3.729730
Catalogs             5.000000
Education            2.992212
Entertainment        3.129353
Finance              2.000000
Food & Drink         2.575000
Games                3.777882
Health & Fitness     3.485577
Lifestyle            2.960000
```

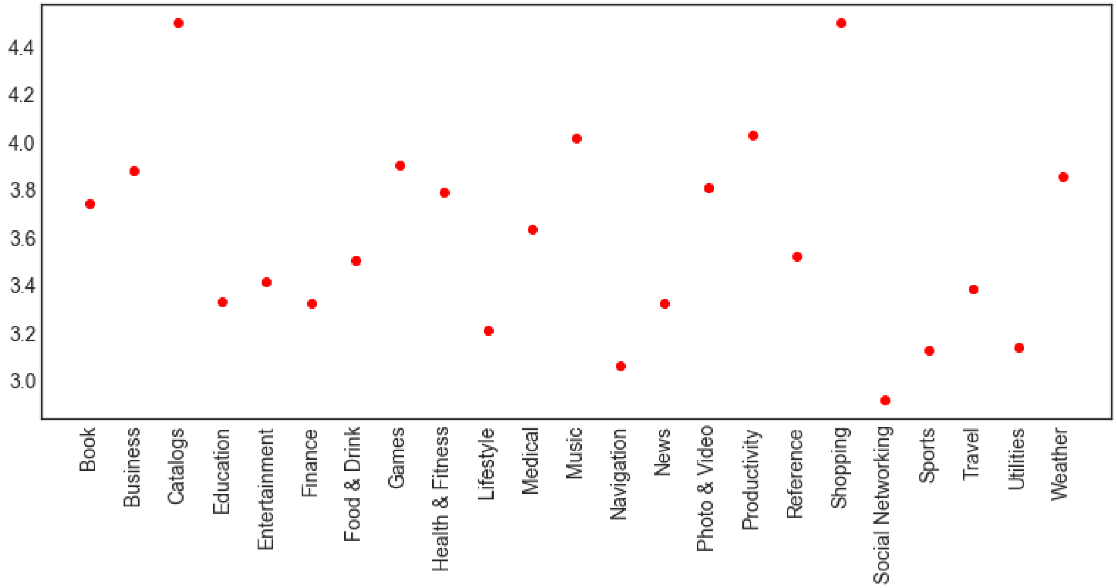
```
Medical                3.366667
Music                  3.683099
Navigation             2.500000
News                   2.647059
Photo & Video          3.681319
Productivity           3.689655
Reference              2.920455
Shopping               5.000000
Social Networking      2.729167
Sports                 2.885714
Travel                 3.640000
Utilities              2.899281
Weather                3.597561
Name: user_rating_ver, dtype: float64
```

```
# lets plot
plt.plot(x)
```

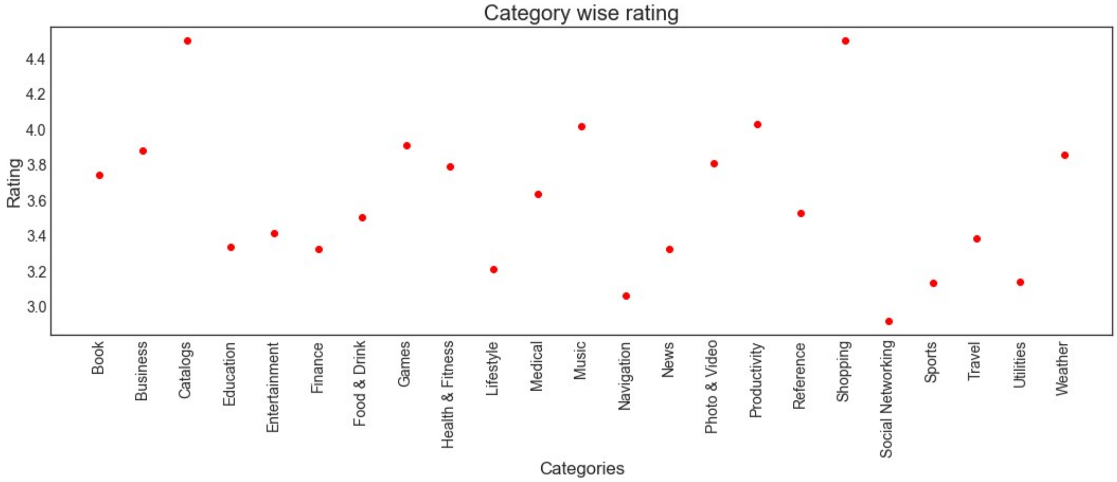
```
[<matplotlib.lines.Line2D at 0x1cf88a679d0>]
```



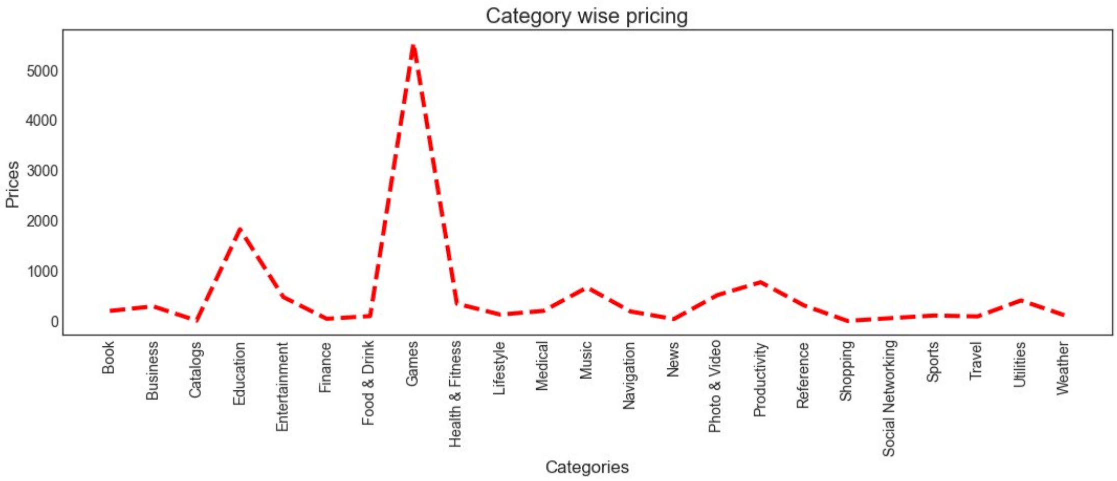
```
#again need to expand
plt.figure(figsize=(12,5))
plt.plot(x, 'ro')
plt.xticks(rotation=90)
plt.show()
```



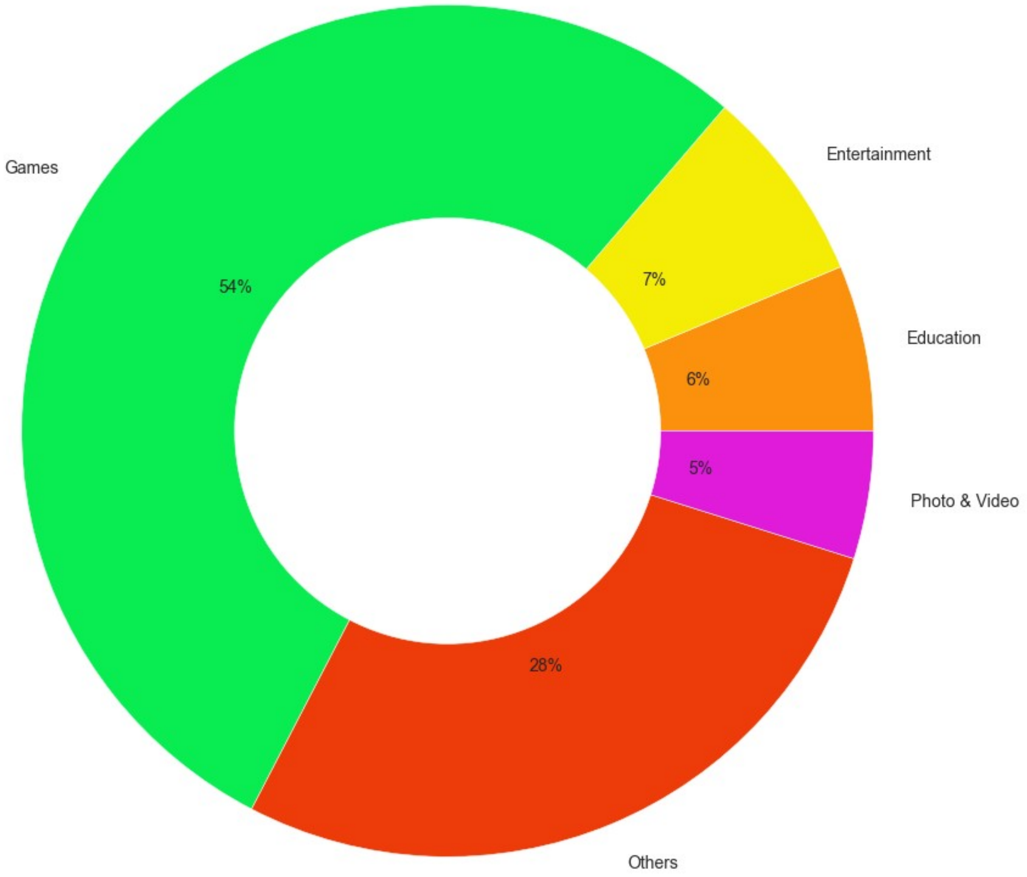
```
# for x
plt.figure(figsize=(16,5))
plt.plot(x, 'ro')
plt.xticks(rotation=90)
plt.title('Category wise rating')
plt.xlabel('Categories')
plt.ylabel('Rating')
plt.show()
```



```
# for Y
plt.figure(figsize=(16,5))
plt.plot(y, 'r--')
plt.xticks(rotation=90)
plt.title('Category wise pricing')
plt.xlabel('Categories')
plt.ylabel('Prices')
plt.show()
```



```
# reducing the number of categories to 5 categories

s = data.prime_genre.value_counts().index[:4]
def categ(x):
    if x in s:
        return x
    else :
        return "Others"

data['broad_genre']= data.prime_genre.apply(lambda x : categ(x))

data['broad_genre'].value_counts()

Games            3862
Others           1998
Entertainment     535
Education         453
Photo & Video     349
Name: broad_genre, dtype: int64

BlueOrangeWapang = ['#fc910d','#f5ed05','#09ed52','#ed3b09','#e01bda']
plt.figure(figsize=(15,15))
label_names=data.broad_genre.value_counts().sort_index().index
size = data.broad_genre.value_counts().sort_index().tolist()

my_circle=plt.Circle( (0,0), 0.5, color='white')
plt.pie(size, labels=label_names, colors=BlueOrangeWapang ,autopct =
'%1.0f%%',)
p=plt.gcf()
p.gca().add_artist(my_circle)
plt.show()
```



```
free =
data[data.price==0].broad_genre.value_counts().sort_index().to_frame()
paid =
data[data.price>0].broad_genre.value_counts().sort_index().to_frame()
total = data.broad_genre.value_counts().sort_index().to_frame()
free.columns=['free']
paid.columns=['paid']
total.columns=['total']
five_ca  =free.join(paid).join(total)
five_ca['Paid_per'] = five_ca.paid*100 /five_ca.total
five_ca['Free_per'] = five_ca.free*100/ five_ca.total
five_ca
```

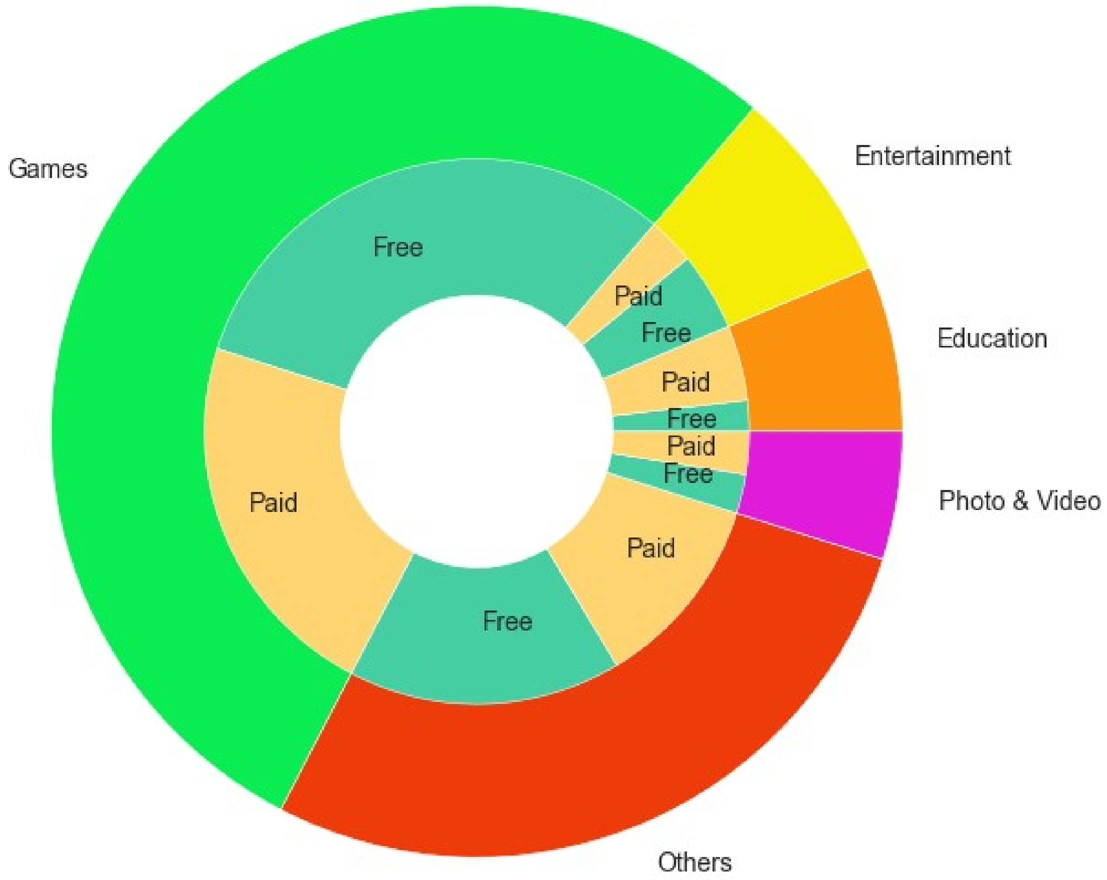
```
               free  paid  total   Paid_per   Free_per
Education       132   321    453  70.860927  29.139073
Entertainment   334   201    535  37.570093  62.429907
Games          2257  1605   3862  41.558778  58.441222
```

```
Others          1166   832   1998  41.641642  58.358358
Photo & Video    167   182    349  52.148997  47.851003
```

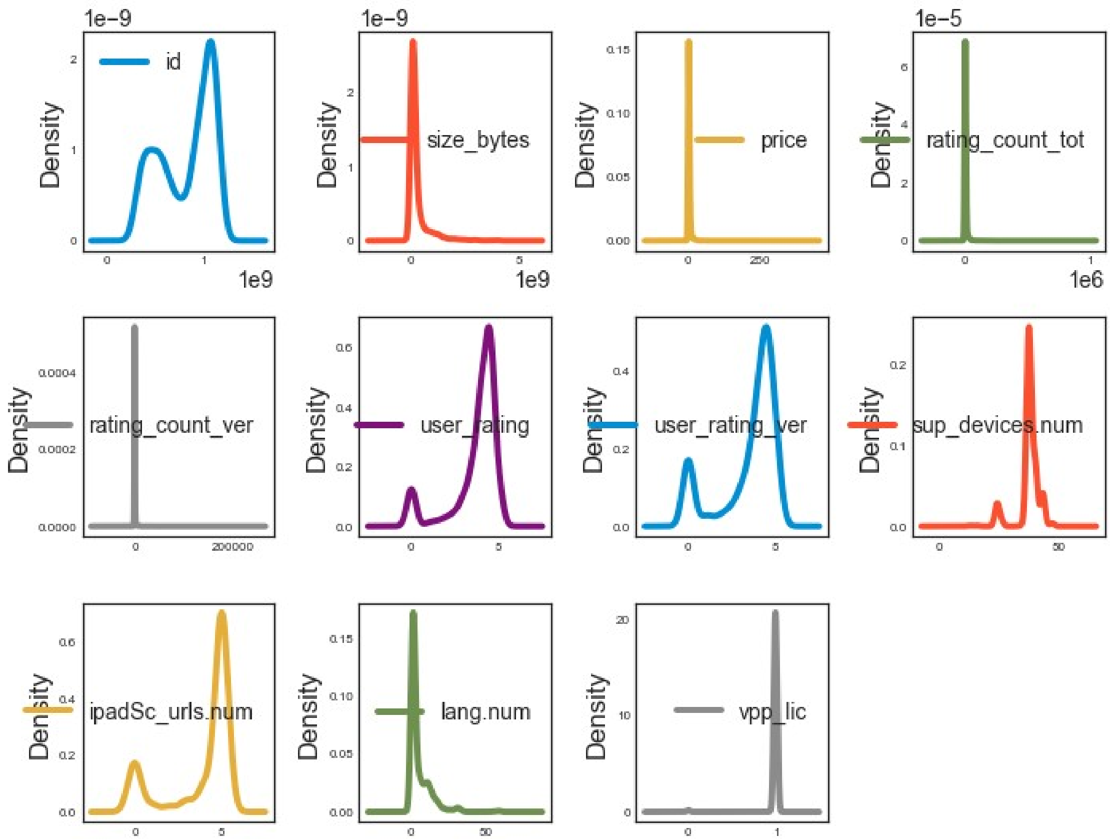
```
plt.figure(figsize=(15,15))
f=pd.DataFrame(index=np.arange(0,10,2),data=five_ca['free'].values,col
umns=['num'])
p=pd.DataFrame(index=np.arange(1,11,2),data=five_ca['paid'].values,col
umns=['num'])
final = pd.concat([f,p],names=['labels']).sort_index()
final.num.tolist()

plt.figure(figsize=(25,25))
group_names=data.broad_genre.value_counts().sort_index().index
group_size=data.broad_genre.value_counts().sort_index().tolist()
h = ['Free', 'Paid']
subgroup_names= 5*h
sub= ['#45cea2','#fdd470']
subcolors= 5*sub
subgroup_size=final.num.tolist()


# First Ring (outside)
fig, ax = plt.subplots()
ax.axis('equal')
mypie, _ = ax.pie(group_size, radius=2.5, labels=group_names,
colors=BlueOrangeWapang)
plt.setp( mypie, width=1.2, edgecolor='white')

# Second Ring (Inside)
mypie2, _ = ax.pie(subgroup_size, radius=1.6, labels=subgroup_names,
labeldistance=0.7, colors=subcolors)
plt.setp( mypie2, width=0.8, edgecolor='white')
plt.margins(0,0)

# show it
plt.show()
```

```
<Figure size 1080x1080 with 0 Axes>
```

```
<Figure size 1800x1800 with 0 Axes>
```



```
paid_apps.plot(kind='density' , subplots=True , layout=(4,4) ,
sharex=False ,
          fontsize=8 , figsize=(10,10))
plt.tight_layout()
```



Feature Engineering

```
from sklearn.preprocessing import LabelEncoder
USD_LABEL =  LabelEncoder()
data['currency']= USD_LABEL.fit_transform(data['currency'])

data.drop(['broad_genre'] ,
          #['currency'],
          axis = 1, inplace = True)

data.drop(['currency'],
          axis = 1, inplace = True)

data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7197 entries, 0 to 7196
Data columns (total 15 columns):
 #   Column             Non-Null Count  Dtype  
---  ------             --------------  -----  
 0   id                 7197 non-null   int64  
 1   track_name         7197 non-null   object 
 2   size_bytes         7197 non-null   int64  
 3   price              7197 non-null   float64
 4   rating_count_tot   7197 non-null   int64  
 5   rating_count_ver   7197 non-null   int64  
 6   user_rating       7197 non-null   float64
 7   user_rating_ver   7197 non-null   float64
 8   ver               7197 non-null   object
 9   cont_rating       7197 non-null   object
 10  prime_genre       7197 non-null   object
 11  sup_devices.num   7197 non-null   int64
 12  ipadSc_urls.num   7197 non-null   int64
 13  lang.num          7197 non-null   int64
 14  vpp_lic           7197 non-null   int64
dtypes: float64(3), int64(8), object(4)
memory usage: 843.5+ KB
```

```
#encoding object columns int
track_name_LABEL =  LabelEncoder()
data['track_name']= track_name_LABEL.fit_transform(data['track_name'])

ver_LABEL =  LabelEncoder()
data['ver']= ver_LABEL.fit_transform(data['ver'])

prime_genre_LABEL =  LabelEncoder()
data['prime_genre']=
prime_genre_LABEL.fit_transform(data['prime_genre'])


cont_rating_LABEL =  LabelEncoder()
data['cont_rating']=
cont_rating_LABEL.fit_transform(data['cont_rating'])

data.head()
```

```
          id  track_name  size_bytes  price  rating_count_tot  \
0  281656475        3676   100788224   3.99             21292
1  281796108        1664   158578688   0.00            161065
2  281940292        5870   100524032   0.00            188583
3  282614216        6132   128512000   0.00            262241
4  282935706         527    92774400   0.00            985920

   rating_count_ver  user_rating  user_rating_ver   ver
cont_rating  \
0                 26          4.0              4.5  1379            2

1                 26          4.0              3.5  1514            2

2               2822          3.5              4.5  1210            2

3                649          4.0              4.5  1236            0

4               5320          4.5              5.0  1472            2
```

```
   prime_genre  sup_devices.num  ipadSc_urls.num  lang.num  vpp_lic
0            7               38                5        10        1
1           15               37                5        23        1
2           22               37                5         3        1
3           17               37                5         9        1
4           16               37                5        45        1
```

```
data.drop(['id'] , axis =1 , inplace = True)
```

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7197 entries, 0 to 7196
Data columns (total 14 columns):
 #   Column             Non-Null Count  Dtype  
---  ------             --------------  -----  
 0   track_name         7197 non-null   int32  
 1   size_bytes         7197 non-null   int64  
 2   price              7197 non-null   float64
 3   rating_count_tot   7197 non-null   int64  
 4   rating_count_ver   7197 non-null   int64  
 5   user_rating        7197 non-null   float64
 6   user_rating_ver    7197 non-null   float64
 7   ver                7197 non-null   int32  
 8   cont_rating        7197 non-null   int32  
 9   prime_genre        7197 non-null   int32  
 10  sup_devices.num    7197 non-null   int64  
 11  ipadSc_urls.num    7197 non-null   int64  
 12  lang.num           7197 non-null   int64  
 13  vpp_lic            7197 non-null   int64  
dtypes: float64(3), int32(4), int64(7)
memory usage: 674.8 KB
```

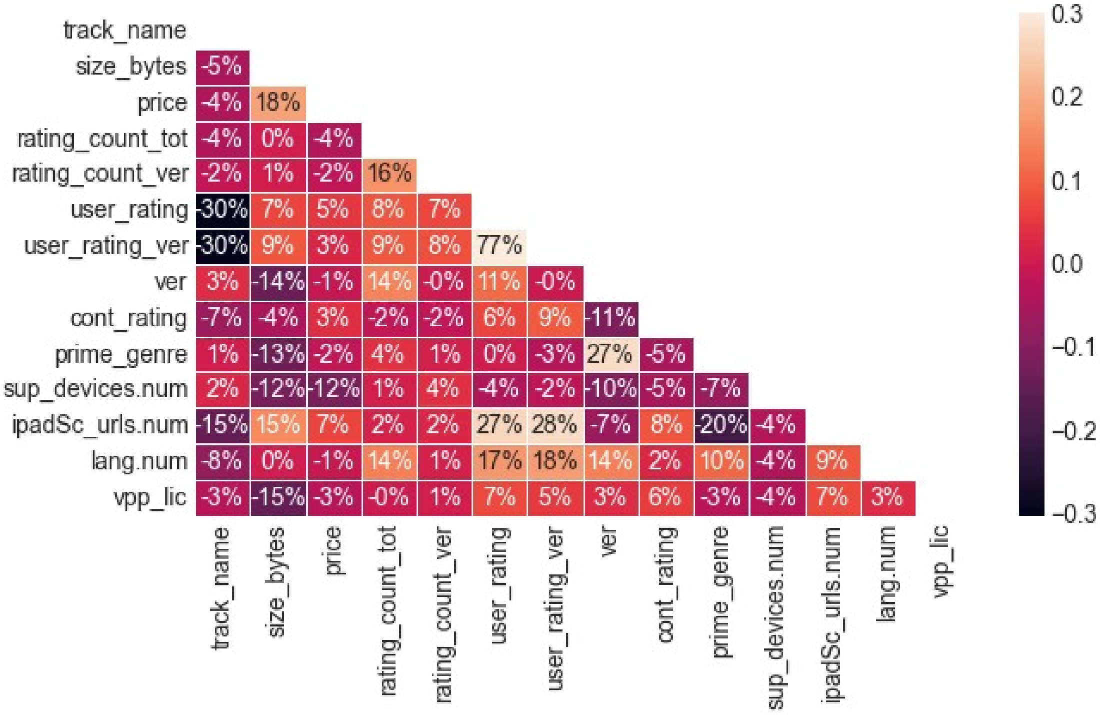
```
#Data about Data
data.describe().style.background_gradient(cmap='Purples')
```

```
<pandas.io.formats.style.Styler at 0x1cf887b0250>
```

```
D_corr = data.corr()
D_corr.style.background_gradient()
```

```
<pandas.io.formats.style.Styler at 0x1cf891cfee0>
```

```
mask = np.zeros_like(data.corr())
mask[np.triu_indices_from(mask)] = True
with sns.axes_style("ticks"):
    f, ax = plt.subplots(figsize=(9, 5))
    ax = sns.heatmap(data.corr(), mask=mask,
vmax=.3,annot=True,fmt=".0%",linewidth=0.5,square=False)
```



PPS(Predictive Power Score)

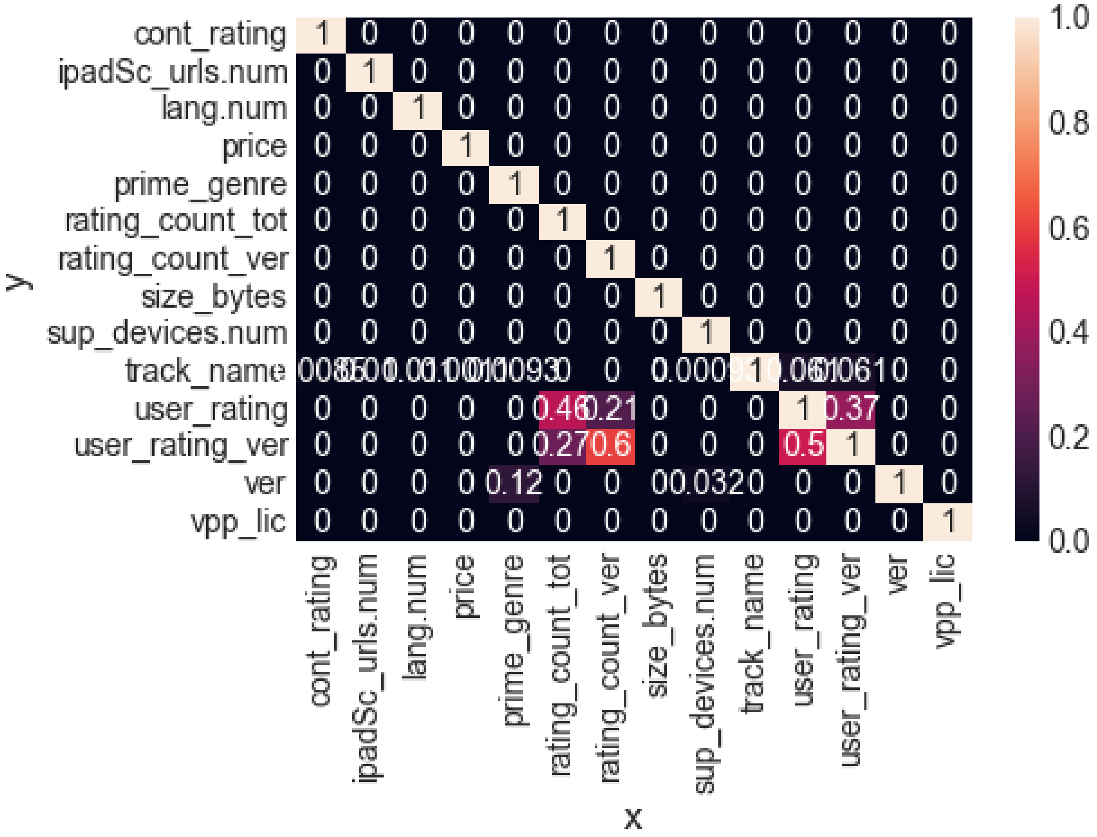
```
#Calculating ppscore
import ppscore
c=pps.matrix(data)
c
```

```
                  x                 y  ppscore            case  
is_valid_score  \
0    track_name        track_name      1.0  predict_itself  
True   
1    track_name        size_bytes      0.0      regression  
True   
2    track_name             price      0.0      regression  
True   
3    track_name  rating_count_tot      0.0      regression  
True   
4    track_name  rating_count_ver      0.0      regression  
True   
..          ...               ...      ...             ...  
...   
191     vpp_lic       prime_genre      0.0      regression  
True   
192     vpp_lic   sup_devices.num      0.0      regression  
True   
193     vpp_lic   ipadSc_urls.num      0.0      regression  
True   
194     vpp_lic          lang.num      0.0      regression  
True   
```

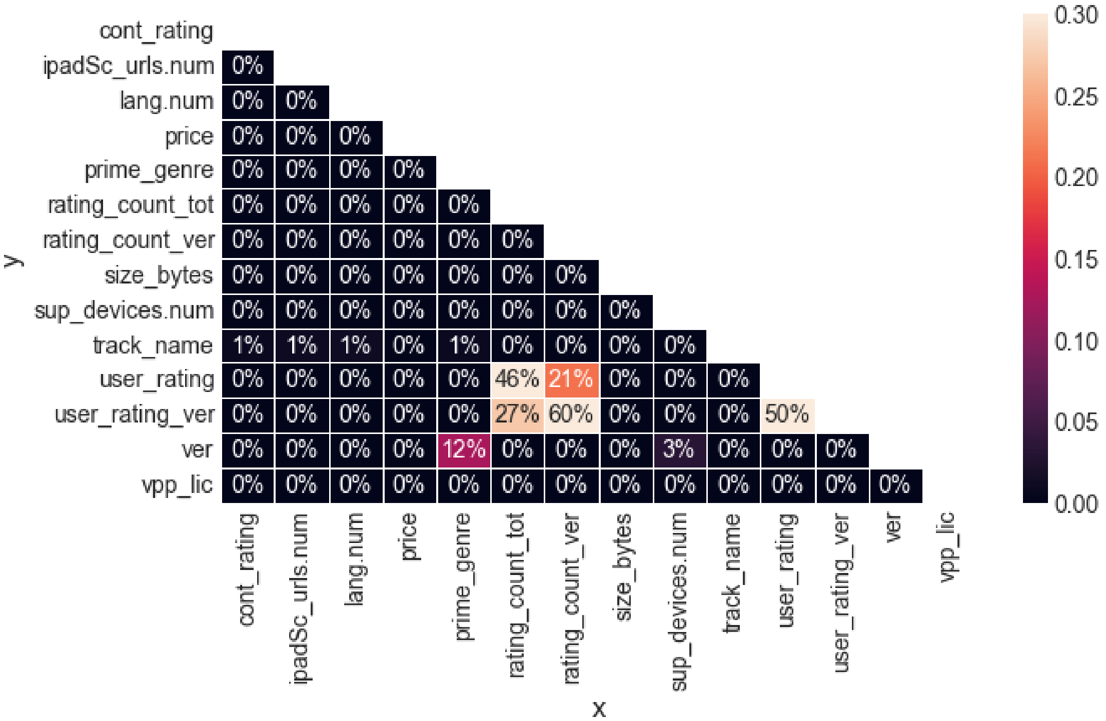
```
195     vpp_lic           vpp_lic      1.0  predict_itself
True

                  metric  baseline_score   model_score  \
0                   None    0.000000e+00  1.000000e+00
1    mean absolute error    1.507944e+08  2.034129e+08
2    mean absolute error    1.771204e+00  2.240976e+00
3    mean absolute error    1.216113e+04  2.047588e+04
4    mean absolute error    4.637924e+02  7.655062e+02
..                   ...             ...           ...
191  mean absolute error    2.912000e+00  3.623269e+00
192  mean absolute error    1.827400e+00  1.836597e+00
193  mean absolute error    1.280600e+00  1.656586e+00
194  mean absolute error    4.398600e+00  5.489964e+00
195                 None    0.000000e+00  1.000000e+00

                       model
0                       None
1    DecisionTreeRegressor()
2    DecisionTreeRegressor()
3    DecisionTreeRegressor()
4    DecisionTreeRegressor()
..                       ...
191  DecisionTreeRegressor()
192  DecisionTreeRegressor()
193  DecisionTreeRegressor()
194  DecisionTreeRegressor()
195                     None

[196 rows x 9 columns]
```

```
figsize=(20,20)
a = pps.matrix(data).pivot(columns='x', index='y', values='ppscore')
sns.heatmap(a, annot=True)
```

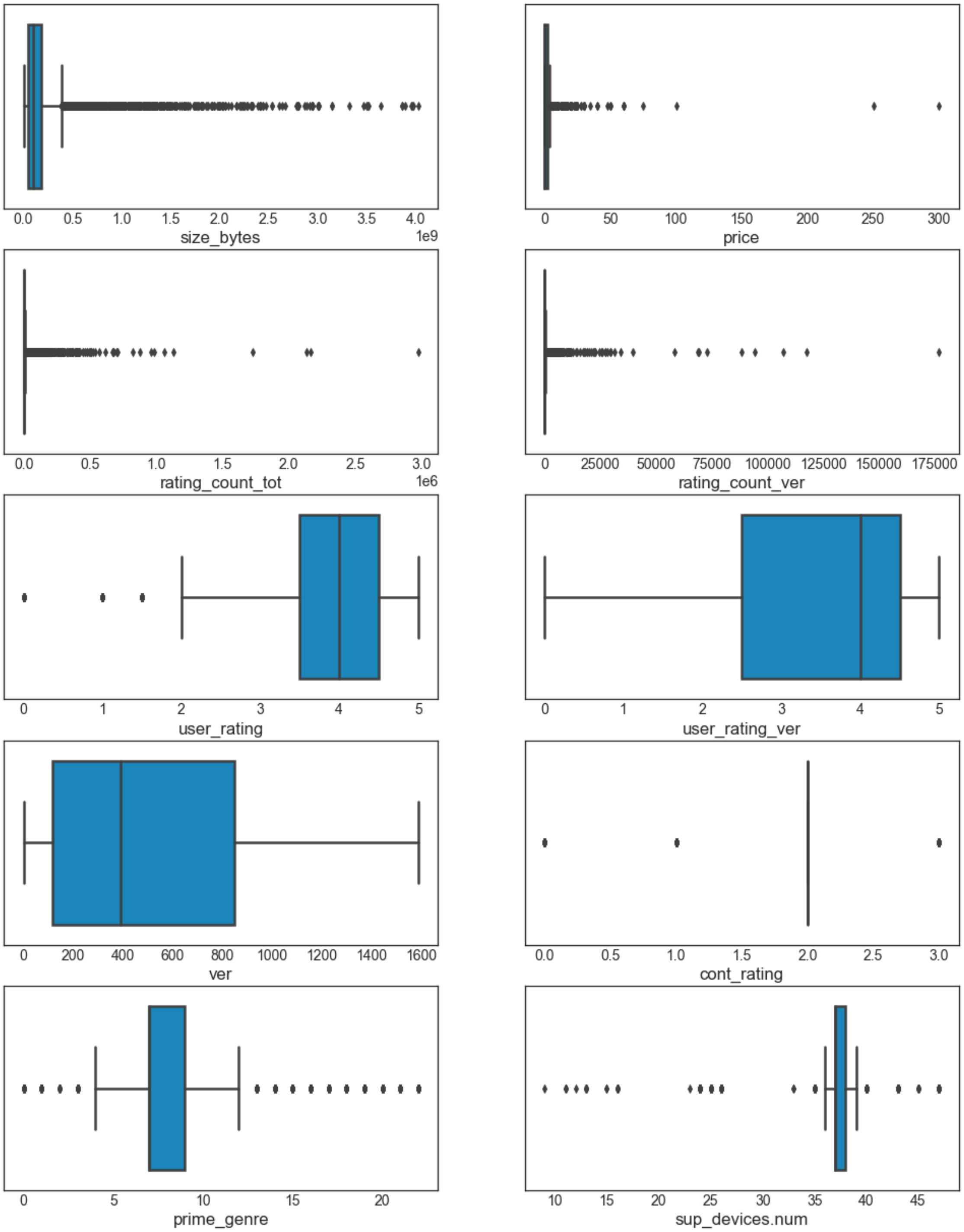
```
<AxesSubplot:xlabel='x', ylabel='y'>
```



```
mask = np.zeros_like(a)
mask[np.triu_indices_from(mask)] = True
with sns.axes_style("ticks"):
    f, ax = plt.subplots(figsize=(9, 5))
    ax = sns.heatmap(a, mask=mask,
vmax=.3,annot=True,fmt=".0%",linewidth=0.5,square=False)
```



```
#Show outliers with boxplot
plt.figure(figsize = (15,20))
col_names = [ 'size_bytes', 'price', 'rating_count_tot',
        'rating_count_ver', 'user_rating', 'user_rating_ver', 'ver',
        'cont_rating', 'prime_genre', 'sup_devices.num']
for i in range(10):
    plt.subplot(5,2,i+1)#3 number of row #2 number of columns
    sns.boxplot(x=data[col_names[i]], linewidth=2.5)
plt.show()
```



```
data.shape
```

```
(7197, 14)
```

```
 outliers_list = []
# For each feature find the data points with extreme high or low values
for feature in data.keys():
    # Calculate Q1 (25th percentile of the data) for the given feature
    Q1 = np.percentile(data[feature], 25)

    # Calculate Q3 (75th percentile of the data) for the given feature
    Q3 = np.percentile(data[feature], 75)

    # Use the interquartile range to calculate an outlier step (1.5
times the interquartile range)
    step = (Q3 - Q1) * 1.5

    # Display the outliers
    print("Data points considered outliers for the feature
'{}':".format(feature))
    outliers = list(data[~((data[feature] >= Q1 - step) &
(data[feature] <= Q3 + step))].index.values)

    display(data[~((data[feature] >= Q1 - step) & (data[feature] <= Q3
+ step))])
    print('-=-=-=-=-=-=-=-=-=-=-=-=------------------------------=-=-
=-=-=-=-=-=-=-=-=-=')
    outliers_list.extend(outliers)

#print("List of Outliers -> \n :{}".format(outliers_list))
```

```
Data points considered outliers for the feature 'track_name':

Empty DataFrame
Columns: [track_name, size_bytes, price, rating_count_tot,
rating_count_ver, user_rating, user_rating_ver, ver, cont_rating,
prime_genre, sup_devices.num, ipadSc_urls.num, lang.num, vpp_lic]
Index: []

-=-=-=-=-=-=-=-=-=-=-=-=------------------------------=-=-=-=-=-=-
=-=-=-=-=
Data points considered outliers for the feature 'size_bytes':

      track_name  size_bytes   price  rating_count_tot
rating_count_ver  \
16          1743   389879808    0.00           2974676
212
103         4440   431771648    2.99             35074
403
115         4054   723764224  249.99               773
10
152         5327   430128128    6.99             54408
65
281         2232   878883840    4.99             15142
73
...          ...         ...     ...               ...               .
..
```

```
7076        2342   479346688    0.00                60
10
7133        6103  3148421120    0.00                 0
0
7162        2606   628180992    2.99                26
0
7164        7192  3503480832    9.99                 0
0
7189        7163   537462784    0.99                 0
0

      user_rating  user_rating_ver   ver  cont_rating  prime_genre  \
16            3.5              3.5  1577            2           18
103           4.5              4.0   962            2            7
115           4.0              3.5  1211            2            3
152           3.5              2.0   350            0            7
281           4.0              4.0  1152            1            7
...           ...              ...   ...          ...          ...
7076          4.5              3.5   609            0            7
7133          0.0              0.0    31            0            7
7162          4.5              0.0   254            3            7
7164          0.0              0.0    45            0            7
7189          0.0              0.0   648            3            7

      sup_devices.num  ipadSc_urls.num  lang.num  vpp_lic
16                 37                1        29        1
103                43                0         6        1
115                37                5         3        1
152                37                5         8        1
281                38                5         6        1
...               ...              ...       ...      ...
7076               40                5         1        1
7133               40                0         1        0
7162               37                5         1        1
7164               40                0         1        0
7189               38                5         1        1

[778 rows x 14 columns]

-=-=-=-=-=-=-=-=-=-=-=-=--------------------------------=-=-=-=-=-=-=-
=-=-=-=-=
Data points considered outliers for the feature 'price':

      track_name  size_bytes  price  rating_count_tot
rating_count_ver  \
8           3688    49250304   9.99              1117
4
10          4768    49618944   4.99             76720
4017
11          4372   227547136   7.99            105776
```

```
166
14          2031    55153664   4.99              6340
668
23          3191    71203840   5.99                 8
0
...          ...         ...    ...              ...               ..
.
7042        6506   196380672   4.99                 1
1
7073        6464    51174400  12.99                 0
0
7105         796    94401536   9.99                39
7
7164        7192  3503480832   9.99                 0
0
7165        7134   192621568   4.99                 0
0

      user_rating  user_rating_ver   ver  cont_rating  prime_genre  \
8             4.5              5.0  1008            2           21
10            4.5              4.5  1096            2            7
11            3.5              2.5  1247            2            7
14            4.5              4.5  1065            2            7
23            4.5              0.0  1541            2            1
...           ...              ...   ...          ...         ...
7042          3.0              3.0    92            2            7
7073          0.0              0.0    31            2            3
7105          2.5              2.5   430            2            1
7164          0.0              0.0    45            0            7
7165          0.0              0.0    75            2            7

      sup_devices.num  ipadSc_urls.num  lang.num  vpp_lic
8                  37                5         1        1
10                 38                4        11        1
11                 37                0         6        1
14                 38                5         2        1
23                 37                5         3        1
...               ...              ...       ...      ...
7042               38                5         1        1
7073               37                5         1        1
7105               37                0         7        1
7164               40                0         1        0
7165               38                5         1        1

[832 rows x 14 columns]

-=-=-=-=-=-=-=-=-=-=-=-=--------------------------------=-=-=-=-=-=-=-
=-=-=-=-=
Data points considered outliers for the feature 'rating_count_tot':
      track_name  size_bytes  price  rating_count_tot
rating_count_ver  \
0           3676   100788224   3.99             21292
26
1           1664   158578688   0.00            161065
26
2           5870   100524032   0.00            188583
2822
3           6132   128512000   0.00            262241
649
4            527    92774400   0.00            985920
5320
...          ...         ...    ...               ...               ..
.
6897        1971    35136512   0.00             13914
493
6914        1944   143591424   0.00             42435
2957
6935        3228   187168768   0.00             11602
191
6969        4391   137533440   0.00             25859
839
7068        3759   281393152   0.00             24097
4469

      user_rating  user_rating_ver   ver  cont_rating  prime_genre  \
0             4.0              4.5  1379            2            7
1             4.0              3.5  1514            2           15
2             3.5              4.5  1210            2           22
3             4.0              4.5  1236            0           17
4             4.5              5.0  1472            2           16
...           ...              ...   ...          ...          ...
6897          4.5              4.0   725            2           18
6914          4.5              4.5   430            2            7
6935          4.5              4.5   148            0            7
6969          5.0              4.5   162            0            7
7068          4.0              4.5   815            2            7

      sup_devices.num  ipadSc_urls.num  lang.num  vpp_lic
0                  38                5        10        1
1                  37                5        23        1
2                  37                5         3        1
3                  37                5         9        1
4                  37                5        45        1
...               ...              ...       ...      ...
6897               37                0         8        1
6914               40                5         1        1
6935               38                5         9        1
6969               37                5         1        1
7068               37                4         1        1
```

```
[1231 rows x 14 columns]

-=-=-=-=-=-=-=-=-=-=-=-=-------------------------------=-=-=-=-=-=-=-
=-=-=-=-=
Data points considered outliers for the feature 'rating_count_ver':

      track_name  size_bytes  price  rating_count_tot
rating_count_ver  \
2           5870   100524032   0.00            188583
2822
3           6132   128512000   0.00            262241
649
4            527    92774400   0.00            985920
5320
5           4528    10485713   0.99              8253
5516
6           3784   227795968   0.00            119487
879
...          ...         ...    ...               ...               ..
.
7044        2661    34719744   0.00              2772
2495
7068        3759   281393152   0.00             24097
4469
7125        4932   323822592   0.00              1362
1142
7129         648   124506112   0.00              3384
3124
7166          56   278811648   0.00              1441
1441

      user_rating  user_rating_ver   ver  cont_rating  prime_genre  \
2             3.5              4.5  1210            2           22
3             4.0              4.5  1236            0           17
4             4.5              5.0  1472            2           16
5             4.0              4.0   454            2            7
6             4.0              4.5  1359            2            5
...           ...              ...   ...          ...          ...
7044          3.5              3.5   345            2            7
7068          4.0              4.5   815            2            7
7125          5.0              5.0    31            2            7
7129          4.5              4.5   118            2            7
7166          5.0              5.0    31            2            7

      sup_devices.num  ipadSc_urls.num  lang.num  vpp_lic
2                  37                5         3        1
3                  37                5         9        1
4                  37                5        45        1
5                  47                5         1        1
```

```
6                  37                0        19        1
...               ...              ...       ...      ...
7044               37                3         1        1
7068               37                4         1        1
7125               38                5         1        1
7129               40                5         1        1
7166               37                5         1        1

[1061 rows x 14 columns]

-=-=-=-=-=-=-=-=-=-=-=-=-------------------------------=-=-=-=-=-=-=-
=-=-=-=-=
Data points considered outliers for the feature 'user_rating':

      track_name  size_bytes  price  rating_count_tot
rating_count_ver  \
199         6172     3375104   3.99                 0
0
301         6133     8039424   3.99                 0
0
330         2653   147066880   7.99                 0
0
402         1761     7689537   0.00               354
215
441         6782    41207059   0.00                 0
0
...          ...         ...    ...              ...               ..
.
7181        6621   178160640   0.99                 0
0
7182        6120     9362432   0.00                 0
0
7184        6622   171944960   0.00                 0
0
7185        7137   208026624   0.99                 0
0
7189        7163   537462784   0.99                 0
0

      user_rating  user_rating_ver   ver  cont_rating  prime_genre  \
199           0.0              0.0  1227            2            5
301           0.0              0.0  1435            3            0
330           0.0              0.0   815            2           20
402           1.5              1.0    91            2           18
441           0.0              0.0   648            2           12
...           ...              ...   ...          ...          ...
7181          0.0              0.0    29            3            7
7182          0.0              0.0    62            2           14
7184          0.0              0.0    29            3            7
7185          0.0              0.0    29            3            7
```

```
7189          0.0              0.0   648            3            7

      sup_devices.num  ipadSc_urls.num  lang.num  vpp_lic
199                37                5         3        1
301                37                5         1        1
330                40                5         1        1
402                43                0         1        1
441                39                5         1        1
...               ...              ...       ...      ...
7181               40                5         0        1
7182               37                0         1        1
7184               40                5         0        1
7185               38                5         1        1
7189               38                5         1        1

[1029 rows x 14 columns]

-=-=-=-=-=-=-=-=-=-=-=-=-------------------------------=-=-=-=-=-=-=-
=-=-=-=-=
Data points considered outliers for the feature 'user_rating_ver':

Empty DataFrame
Columns: [track_name, size_bytes, price, rating_count_tot,
rating_count_ver, user_rating, user_rating_ver, ver, cont_rating,
prime_genre, sup_devices.num, ipadSc_urls.num, lang.num, vpp_lic]
Index: []

-=-=-=-=-=-=-=-=-=-=-=-=-------------------------------=-=-=-=-=-=-=-
=-=-=-=-=
Data points considered outliers for the feature 'ver':

Empty DataFrame
Columns: [track_name, size_bytes, price, rating_count_tot,
rating_count_ver, user_rating, user_rating_ver, ver, cont_rating,
prime_genre, sup_devices.num, ipadSc_urls.num, lang.num, vpp_lic]
Index: []

-=-=-=-=-=-=-=-=-=-=-=-=-------------------------------=-=-=-=-=-=-=-
=-=-=-=-=
Data points considered outliers for the feature 'cont_rating':

      track_name  size_bytes  price  rating_count_tot
rating_count_ver  \
3           6132   128512000   0.00            262241
649
7           3740   130242560   0.00           1126879
3594
12          2226   179979264   0.00            479440
203
17          6030   167407616   0.00            223885
3726
```

```
18          4531   147093504   0.00            402925
136
...          ...         ...    ...               ...               ..
.
7188        1280   168774656   0.00                18
18
7189        7163   537462784   0.99                 0
0
7191        3939    27853824   2.99                11
0
7194         674   111322112   1.99                15
0
7195        5692    97235968   0.00                85
32

      user_rating  user_rating_ver   ver  cont_rating  prime_genre  \
3             4.0              4.5  1236            0           17
7             4.0              4.5  1522            0           11
12            3.5              4.0   858            1           21
17            4.0              4.5   550            0           20
18            4.0              4.5   548            0           11
...           ...              ...   ...          ...          ...
7188          4.0              4.0    30            0            7
7189          0.0              0.0   648            3            7
7191          4.0              0.0   118            1            7
7194          4.5              0.0    45            3           21
7195          4.5              4.5    40            0            7

      sup_devices.num  ipadSc_urls.num  lang.num  vpp_lic
3                  37                5         9        1
7                  37                4         1        1
12                 37                4        33        1
17                 37                5        18        1
18                 37                3        16        1
...               ...              ...       ...      ...
7188               38                4         1        1
7189               38                5         1        1
7191               37                0         1        1
7194               37                1         1        1
7195               38                0         2        1

[2764 rows x 14 columns]

-=-=-=-=-=-=-=-=-=-=-=-=-------------------------------=-=-=-=-=-=-=-
=-=-=-=-=
Data points considered outliers for the feature 'prime_genre':

      track_name  size_bytes  price  rating_count_tot
rating_count_ver  \
1           1664   158578688   0.00            161065
```

```
26
2          5870   100524032   0.00           188583
2822
3          6132   128512000   0.00           262241
649
4           527    92774400   0.00           985920
5320
8          3688    49250304   9.99             1117
4
...          ...         ...    ...              ...               ..
.
7173        3356    18164736   0.99                0
0
7179        1766   113382400   0.00              279
5
7180        2846    94008320   0.00               26
3
7182        6120     9362432   0.00                0
0
7194         674   111322112   1.99               15
0

      user_rating  user_rating_ver   ver  cont_rating  prime_genre  \
1             4.0              3.5  1514            2           15
2             3.5              4.5  1210            2           22
3             4.0              4.5  1236            0           17
4             4.5              5.0  1472            2           16
8             4.5              5.0  1008            2           21
...           ...              ...   ...          ...          ...
7173          0.0              0.0   114            2           21
7179          3.5              3.0    14            2           18
7180          5.0              5.0    70            3           21
7182          0.0              0.0    62            2           14
7194          4.5              0.0    45            3           21

      sup_devices.num  ipadSc_urls.num  lang.num  vpp_lic
1                  37                5        23        1
2                  37                5         3        1
3                  37                5         9        1
4                  37                5        45        1
8                  37                5         1        1
...               ...              ...       ...      ...
7173               37                0         1        1
7179               37                4         1        1
7180               37                4         1        1
7182               37                0         1        1
7194               37                1         1        1

[2102 rows x 14 columns]
```

```
-=-=-=-=-=-=-=-=-=-=-=-=--------------------------------=-=-=-=-=-=-
=-=-=-=-=
Data points considered outliers for the feature 'sup_devices.num':

      track_name  size_bytes  price  rating_count_tot
rating_count_ver  \
5           4528    10485713   0.99              8253
5516
19          1104    10735026   2.99             31456
4178
20          6181    70707916   1.99              2929
966
21            82     6169600   2.99             11447
781
24           980    11423008   3.99              3241
297
...          ...         ...    ...               ...               ..
.
7176        2477   184324096   0.00                 0
0
7181        6621   178160640   0.99                 0
0
7183        5039    57349120   3.99               292
292
7184        6622   171944960   0.00                 0
0
7196        1653    90898432   0.00                 3
3

      user_rating  user_rating_ver   ver  cont_rating  prime_genre  \
5             4.0              4.0   454            2            7
19            4.0              3.5    30            2            7
20            4.5              4.5   961            2           16
21            5.0              5.0  1257            2            7
24            4.0              4.0   656            2           11
...           ...              ...   ...          ...          ...
7176          0.0              0.0    31            2            7
7181          0.0              0.0    29            3            7
7183          4.0              4.0   621            3            7
7184          0.0              0.0    29            3            7
7196          5.0              5.0    29            2            7

      sup_devices.num  ipadSc_urls.num  lang.num  vpp_lic
5                  47                5         1        1
19                 47                0         1        1
20                 43                0         2        1
21                 40                5         1        1
24                 43                2        10        1
...               ...              ...       ...      ...
7176               40                4         1        1
```

```
7181               40                5         0        1
7183               40                5         1        1
7184               40                5         0        1
7196               40                0         2        1

[1975 rows x 14 columns]

-=-=-=-=-=-=-=-=-=-=-=-=-------------------------------=-=-=-=-=-=-
=-=-=-=-=
Data points considered outliers for the feature 'ipadSc_urls.num':

Empty DataFrame
Columns: [track_name, size_bytes, price, rating_count_tot,
rating_count_ver, user_rating, user_rating_ver, ver, cont_rating,
prime_genre, sup_devices.num, ipadSc_urls.num, lang.num, vpp_lic]
Index: []

-=-=-=-=-=-=-=-=-=-=-=-=-------------------------------=-=-=-=-=-=-
=-=-=-=-=
Data points considered outliers for the feature 'lang.num':

      track_name  size_bytes  price  rating_count_tot
rating_count_ver  \
1           1664   158578688   0.00            161065
26
4            527    92774400   0.00            985920
5320
6           3784   227795968   0.00            119487
879
12          2226   179979264   0.00            479440
203
15          5555   207907840   0.00             56194
87
...          ...         ...    ...               ...               ..
.
7109        4758    55877632   0.00               103
1
7112        2355    45356032   1.99                28
5
7148         934    62359552   0.00                96
96
7169        5120    32685056   2.99                 9
3
7175        4453    64244736   0.00                41
19

      user_rating  user_rating_ver   ver  cont_rating  prime_genre  \
1             4.0              3.5  1514            2           15
4             4.5              5.0  1472            2           16
6             4.0              4.5  1359            2            5
12            3.5              4.0   858            1           21
```

```
15            4.0              3.5   849           2          20
...           ...              ...   ...         ...         ...
7109          4.0              5.0  1348           0          19
7112          3.5              4.0   126           2           8
7148          4.5              4.5    29           2           7
7169          3.0              3.5    31           2           4
7175          4.5              4.5   298           2           7

      sup_devices.num  ipadSc_urls.num  lang.num  vpp_lic
1                  37                5        23        1
4                  37                5        45        1
6                  37                0        19        1
12                 37                4        33        1
15                 37                1        26        1
...               ...              ...       ...      ...
7109               37                4        26        1
7112               13                0        24        1
7148               40                3        25        1
7169               37                0        31        1
7175               37                5        25        1

[428 rows x 14 columns]

-=-=-=-=-=-=-=-=-=-=-=-=-------------------------------=-=-=-=-=-=-=-
=-=-=-=-=
Data points considered outliers for the feature 'vpp_lic':

      track_name  size_bytes  price  rating_count_tot
rating_count_ver  \
42          6193    44241920   3.99                30
0
48          1718    72748032   0.00             57500
103
74           100    98108416   0.00             48407
20
122         4930    17010688   0.99             53821
165
180         3698    64147456   0.00             10472
58
181          380   100779008   0.00               115
70
311         4538     7344128   1.99               348
33
334         2736   151864320   0.00            260965
228
366          102    85724160   0.00             78890
449
391         2006    64705536   0.00            132703
394
392         2931    37161984   4.99               286
```

```
25
534          4364   116247552   0.00                3818
69
605          1523    11911065   0.99                4401
2927
818          5133    16048128   0.99               14361
918
1019          393   134715392   0.00                1628
0
1153         5920     4303262   0.99                 118
118
1200         5134    23957504   0.99                1418
138
1550         1298   321376256   2.99                1474
125
2203         1717    99644416   0.00               13898
336
2333         6803  1406185472  14.99                   0
0
2445         4541    27975680   7.99                 147
147
2585         5876    43324416   0.00                 912
298
2608         6472  1997754368  14.99                   0
0
3055         6804  1338712064  11.99                   0
0
3175         5332   105910272   1.99                   3
3
3309         6473  1637367808  14.99                   0
0
3454         3893    38805504   0.00                 199
199
4013         7185  2170589184   9.99                   0
0
4082         2500    69058560   1.99                  79
35
4300         6471  2057388032  14.99                   0
0
4673         6452  3956326400   7.99                   0
0
5178         1850   111337472   1.99                   3
3
5207          287   202067968   1.99                   3
3
5208         4997   109820928   1.99                1246
936
5384         6101   199159808   0.00                   0
0
6162         1772   281761792   0.00                 966
```

```
2
6196        1565   148648960   0.00               823
74
6303        6104  1239953408   0.00                 0
0
6443        6436  2002585600  14.99                 0
0
6494        5923    30086144   0.00                 0
0
6594        6633   418452480   5.99                 0
0
6643        6100  2808324096   0.00                 0
0
6649        6102    62540800   0.00                 0
0
6677        6802  3968637952   7.99                 0
0
6709        6470  3148132352  11.99                 0
0
6772        6469  3975609344   9.99                 0
0
6797        1719   405783552   0.00                81
0
7008        5894   167348224   2.99                 3
3
7133        6103  3148421120   0.00                 0
0
7164        7192  3503480832   9.99                 0
0

      user_rating  user_rating_ver   ver  cont_rating  prime_genre  \
42            3.5              0.0   829            2           12
48            3.0              4.0   990            2           19
74            3.0              3.5  1245            0           13
122           3.5              4.0   468            0            7
180           3.5              2.0  1505            1           19
181           4.0              5.0  1218            1           13
311           4.0              4.0   390            2           15
334           4.0              3.0   553            0           18
366           3.0              3.5  1213            0            4
391           4.0              3.0   783            2           13
392           2.5              5.0   613            2           20
534           2.5              1.5  1084            2           19
605           5.0              5.0   314            3            7
818           4.5              4.5   134            3            7
1019          2.0              0.0  1401            2           19
1153          1.5              1.5    30            2           18
1200          4.5              4.0   298            3            7
1550          4.5              4.5    67            0            7
2203          3.0              4.0   988            0           19
```

```
2333          0.0              0.0    67            1            7
2445          5.0              5.0    30            0            7
2585          4.0              4.5   264            2           20
2608          0.0              0.0    45            0            7
3055          0.0              0.0    30            1            7
3175          5.0              5.0   114            2            7
3309          0.0              0.0    30            0            7
3454          4.0              4.0    30            3            7
4013          0.0              0.0   115            0            7
4082          4.5              4.5    62            3            7
4300          0.0              0.0    31            0            7
4673          0.0              0.0    62            0            7
5178          3.5              3.5    92            3            7
5207          2.5              2.5    29            3            7
5208          4.0              4.5   638            2            7
5384          0.0              0.0    45            0            7
6162          4.5              5.0   803            2            7
6196          4.0              4.5   298            2            7
6303          0.0              0.0    30            0            7
6443          0.0              0.0    30            0            7
6494          0.0              0.0   118            2           21
6594          0.0              0.0    45            0            7
6643          0.0              0.0    30            0            7
6649          0.0              0.0    45            0            7
6677          0.0              0.0    62            0            7
6709          0.0              0.0    30            0            7
6772          0.0              0.0    45            0            7
6797          3.0              0.0   475            2           19
7008          3.5              3.5    62            2            7
7133          0.0              0.0    31            0            7
7164          0.0              0.0    45            0            7

      sup_devices.num  ipadSc_urls.num  lang.num  vpp_lic
42                 37                0         2        0
48                 37                0         1        0
74                 37                0         1        0
122                43                1         1        0
180                37                5         1        0
181                37                4         1        0
311                40                0         1        0
334                37                0        14        0
366                37                5         1        0
391                37                5         1        0
392                24                5         1        0
534                37                1         1        0
605                47                0         1        0
818                43                0         1        0
1019               37                5         1        0
1153               45                1         1        0
1200               43                5         1        0
```

```
1550                38                  5          1         0
2203                37                  4          1         0
2333                43                  0          2         0
2445                43                  0          1         0
2585                37                  5          1         0
2608                43                  0          1         0
3055                40                  0          2         0
3175                40                  5         10         0
3309                40                  0          1         0
3454                43                  3         16         0
4013                40                  0          2         0
4082                40                  4          1         0
4300                40                  0          1         0
4673                40                  0          1         0
5178                37                  5          1         0
5207                38                  5          8         0
5208                40                  5          1         0
5384                40                  0          1         0
6162                37                  5          1         0
6196                37                  5          9         0
6303                40                  0          1         0
6443                40                  0          1         0
6494                37                  3          1         0
6594                38                  0          1         0
6643                38                  0          1         0
6649                40                  0          1         0
6677                38                  0          1         0
6709                40                  0          1         0
6772                38                  0          1         0
6797                37                  4          1         0
7008                37                  5         13         0
7133                40                  0          1         0
7164                40                  0          1         0
```

```
-=-=-=-=-=-=-=-=-=-=-=-=-------------------------------=-=-=-=-=-=-
=-=-=-=-=
```

With this project, we have analyzed the app_store data. Hope to see you in another project...