Capstone Project

On

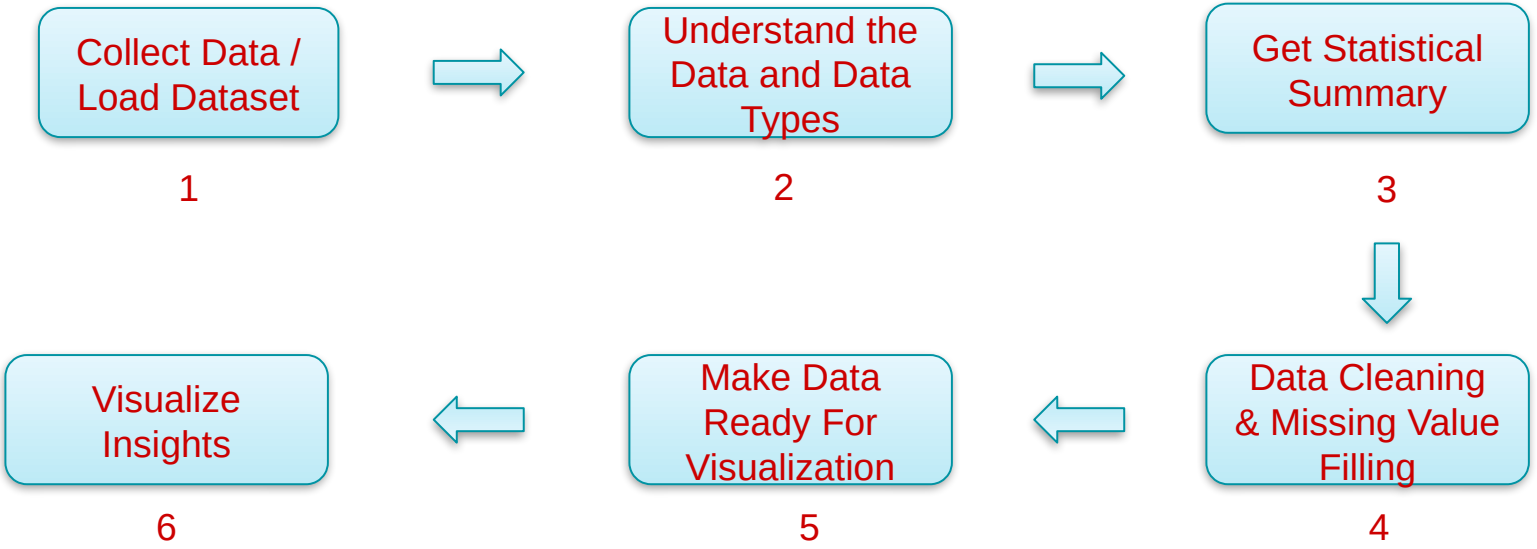
Play Store App Review Analysis(EDA)

Project Contributor : Kuldip Parmar

Table Of Content

What is Exploratory Data Analysis ?

- **Exploratory Data Analysis (EDA)** is an approach to analyze the data using visual techniques. It is used to discover trends, patterns, or to check assumptions with the help of statistical summary and graphical representations.



Problem Statement

1. How to know which Category’s or Genres’s App user preferred the most, that could leads the growth of Business or Organization or successive App launch for Developer or Company.
2. How to know the psychology of user preference or usage of App without Analysis of Data of that Area ?
3. How to know the key factors responsible for app engagement and success ?
4. How can someone Predict the App success ratio and user preferable App before launching the app by only seeing the raw data or just by seeing play store ?

Process of EDA

1. Understand Row & Column :

   - First step after collecting data or by using dataset is to understand Rows and Columns And their datatypes
   - As we know Data can be in structured form or unstructured also , so we need to change the datatypes.
   - If any numerical column is in string form we have to convert it into numeric.
   - Mostly analysis done on numerical values so have to keep in mind.

2. Handling Missing and Null values :

- There is no clean data in dataset so we have to fill out missing values as per requirements or have to drop it.
- To Fill a numeric value(e.g. 0) or string value(e.g. ‘Unknown’) at missing cell is not great deal.
- We have to use statistical methods like mean, median, mode etc. to put the accurate or nearest values to get more accurate results.

Libraries Used

- NumPy (Numerical Python)
- Pandas (Data reading & Data cleaning)
- Matplotlib(Data Visualization)
- Seaborn (Data Visualization)

Play Store Dataset

Column Overviews :

° Install column - total installed of application.

° Price column - price value is in dollar($).

° Reviews column - total reviews made by reviewer of application

° Size column - size of application which is in 'MB'

° Android_Ver column - compatible version of application for android device

° Current_Ver column - Latest version of application

- Genres column - Category and area of that application
- Content_Rating column - Age restriction for application installs
- Category column - main area of application
- Last_Update column - when last time application updated or review given
- Rating column - rating given by user to an application
- Type column - it describe wether application is free or paid
- App column - App contains name of application and what about the application

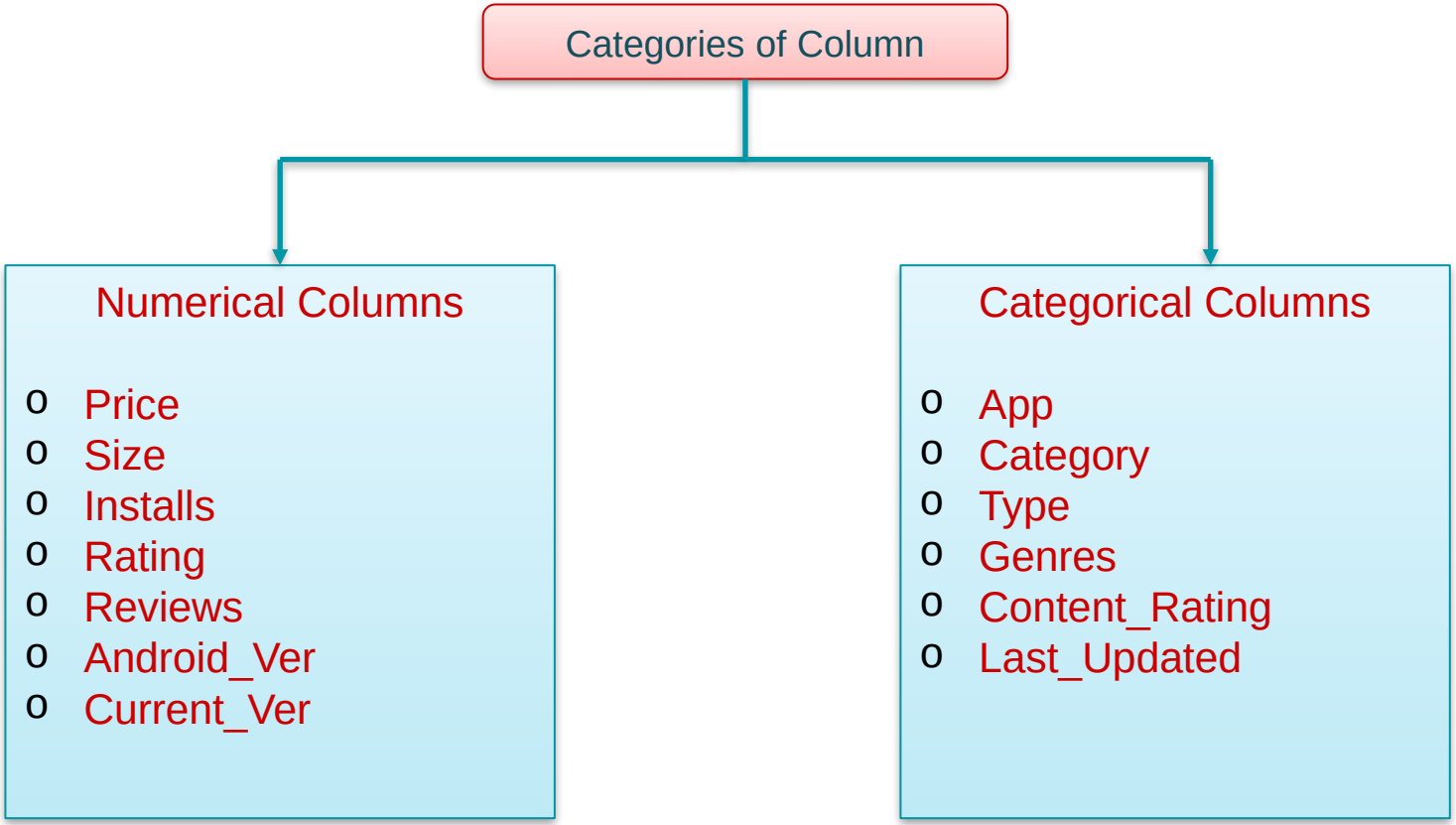
Data Exploration

First Look of Data

App	Photo Editor & Candy Camera & Grid & ScrapBook	Coloring book moana	U Launcher Lite – FREE Live Cool Themes, Hide ...	Sketch - Draw & Paint	Pixel Draw - Number Art Coloring Book
Category	ART_AND_DESIGN	ART_AND_DESIGN	ART_AND_DESIGN	ART_AND_DESIGN	ART_AND_DESIGN
Rating	4.1	3.9	4.7	4.5	4.3
Reviews	159	967	87510	215644	967
Size	19M	14M	8.7M	25M	2.8M
Installs	10,000+	500,000+	5,000,000+	50,000,000+	100,000+
Type	Free	Free	Free	Free	Free
Price	0	0	0	0	0
Content Rating	Everyone	Everyone	Everyone	Teen	Everyone
Genres	Art & Design	Art & Design;Pretend Play	Art & Design	Art & Design	Art & Design;Creativity
Last Updated	January 7, 2018	January 15, 2018	August 1, 2018	June 8, 2018	June 20, 2018
Current Ver	1.0.0	2.0.0	1.2.4	Varies with device	1.1
Android Ver	4.0.3 and up	4.0.3 and up	4.0.3 and up	4.2 and up	4.4 and up

Overall Info

```
<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
```


Categories of Column
Numerical Columns
Price
Size
Installs
Rating
Reviews
Android_Ver
Current_Ver
Categorical Columns
App
Category
Type
Genres
Content_Rating
Last_Updated

Data Cleaning

Nan Value Cleaning

	Total Values	No of NaN values
Rating	10841	1474
Current Ver	10841	8
Android Ver	10841	3
Type	10841	1
Content Rating	10841	1
App	10841	0
Category	10841	0
Reviews	10841	0
Size	10841	0
Installs	10841	0
Price	10841	0
Genres	10841	0
Last Updated	10841	0

Before Cleaning

Nan Values removed. also rows dropped from 10841 to 7021.

	Total Values	No of NaN values
App	7021	0
Category	7021	0
Rating	7021	0
Reviews	7021	0
Size	7021	0
Installs	7021	0
Type	7021	0
Price	7021	0
Content_Rating	7021	0
Genres	7021	0
Last_Updated	7021	0
Current_Ver	7021	0
Android_Ver	7021	0

After Cleaning

Install Column Cleaning

```
0         10,000+
1        500,000+
2      5,000,000+
3     50,000,000+
4        100,000+
Name: Installs, dtype: object
```

+ sign removed

```
0       10000
1      500000
2     5000000
3    50000000
4      100000
Name: Installs, dtype: int64
```

Before Cleaning

After Cleaning

Price Column Cleaning

```
array(['0', '$4.99', '$3.99', '$6.99', '$1.49', '$2.99', '$7.99', '$5.99',
       '$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',
       '$1.00', '$29.99', '$12.99', '$2.49', '$10.99', '$1.50', '$19.99',
       '$15.99', '$33.99', '$74.99', '$39.99', '$3.95', '$4.49', '$1.70',
       '$8.99', '$2.00', '$3.88', '$25.99', '$399.99', '$17.99',
       '$400.00', '$3.02', '$1.76', '$4.84', '$4.77', '$1.61', '$2.50',
       '$1.59', '$6.49', '$1.29', '$5.00', '$13.99', '$299.99', '$379.99',
       '$37.99', '$18.99', '$389.99', '$19.90', '$8.49', '$1.75',
       '$14.00', '$4.85', '$46.99', '$109.99', '$154.99', '$3.08',
       '$2.59', '$4.80', '$1.96', '$19.40', '$3.90', '$4.59', '$15.46',
       '$3.04', '$4.29', '$2.60', '$3.28', '$4.60', '$28.99', '$2.95',
       '$2.90', '$1.97', '$200.00', '$89.99', '$2.56', '$30.99', '$3.61',
       '$394.99', '$1.26', '$1.20', '$1.04'], dtype=object)
```

$ sign removed

```
array([  0.  ,   4.99,   7.99,   3.99,   2.99,   1.99,   5.99,   6.99,
         9.99,   0.99,   3.49,  10.99,   7.49,   1.5 ,  15.99,  79.99,
         9.  ,  10.  ,  16.99,  11.99,  29.99,  14.99,   5.49,  33.99,
        24.99,  39.99,  19.99,   4.49,   1.7 ,   1.49,   3.88, 399.99,
        17.99, 400.  ,   2.49,   3.02,   1.76,   4.84,   4.77,   1.61,
         1.59,   6.49,   1.29, 299.99, 379.99,  37.99,  18.99, 389.99,
         8.49,   1.75,  14.  ,   2.  ,   3.08,   2.59,  19.4 ,  15.46,
         8.99,   3.04,  12.99,  13.99,   4.29,   3.28,   4.6 ,   1.  ,
         2.9 ,   1.97,   2.56,   1.2 ])
```

Before Cleaning

After Cleaning

Size Column Cleaning

	index	Size
0	0	19M
1	1	14M
2	2	8.7M
3	3	25M
4	4	2.8M
5	5	5.6M
6	6	19M
7	7	29M
8	8	33M
9	9	3.1M
10	10	28M
11	11	12M
12	12	20M
13	13	21M
14	14	37M

Before Cleaning

M removed & All size converted into MB.

	index	Size
0	0	19.0
1	2	8.7
2	3	25.0
3	4	2.8
4	5	5.6
5	6	19.0
6	7	29.0
7	8	33.0
8	9	3.1
9	10	28.0
10	11	12.0
11	12	20.0
12	13	21.0
13	14	37.0
14	16	5.5

After Cleaning

Android_Ver Column Cleaning

```
array(['4.0.3 and up', '4.2 and up', '4.4 and up', '2.3 and up',
       '3.0 and up', '4.1 and up', '4.0 and up', '2.3.3 and up',
       'Varies with device', '2.2 and up', '5.0 and up', '6.0 and up',
       '1.6 and up', '1.5 and up', '2.1 and up', '7.0 and up',
       '5.1 and up', '4.3 and up', '4.0.3 - 7.1.1', '2.0 and up',
       '3.2 and up', '4.4W and up', '7.1 and up', '7.0 - 7.1.1',
       '8.0 and up', '5.0 - 8.0', '3.1 and up', '2.0.1 and up',
       '4.1 - 7.1.1', nan, '5.0 - 6.0', '1.0 and up', '2.2 - 7.1.1',
       '5.0 - 7.1.1'], dtype=object)
```

Before Cleaning

```
array(['4.0.3 and up', '4.2 and up', '4.4 and up', '2.3 and up',
       '3.0 and up', '4.1 and up', '4.0 and up', '2.3.3 and up',
       'Varies with device', '2.2 and up', '5.0 and up', '6.0 and up',
       '1.6 and up', '1.5 and up', '2.1 and up', '7.0 and up',
       '5.1 and up', '4.3 and up', '4.0.3 - 7.1.1', '2.0 and up',
       '3.2 and up', '7.1 and up', '7.0 - 7.1.1', '8.0 and up',
       '5.0 - 8.0', '3.1 and up', '2.0.1 and up', '4.1 - 7.1.1', nan,
       '5.0 - 6.0', '1.0 and up', '2.2 - 7.1.1', '5.0 - 7.1.1'],
      dtype=object)
```

After Cleaning

App Column Cleaning

```
ROBLOX                                                 9
CBS Sports App - Scores, News, Stats & Watch Live      8
ESPN                                                   7
Duolingo: Learn Languages Free                         7
Candy Crush Saga                                       7
                                                      ..
Meet U - Get Friends for Snapchat, Kik & Instagram     1
U-Report                                               1
U of I Community Credit Union                          1
Waiting For U Launcher Theme                           1
iHoroscope - 2018 Daily Horoscope & Astrology          1
Name: App, Length: 9659, dtype: int64
```

Before Cleaning

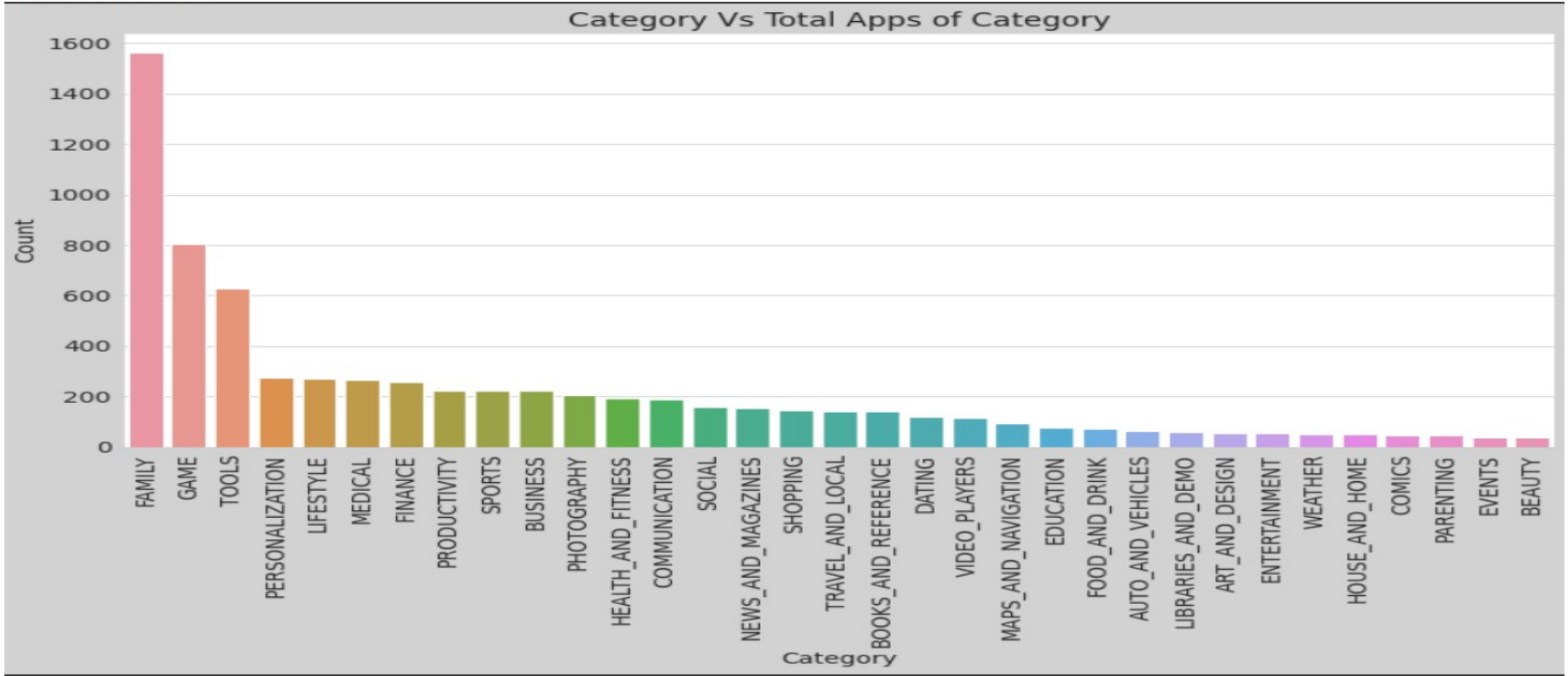
Duplicates removed

```
Photo Editor & Candy Camera & Grid & ScrapBook         1
Rockstars of Ooo                                       1
Angelo Rules - Crazy day                               1
Flipped Out! - Powerpuff Girls                         1
Adventure Time Game Wizard                             1
                                                      ..
Mopar Drag N Brag                                      1
Read Unlimitedly! Kids'n Books                         1
Dark Infusion Substratum Theme for Android N & O       1
Fantastic Chefs: Match 'n Cook                         1
iHoroscope - 2018 Daily Horoscope & Astrology          1
Name: App, Length: 9659, dtype: int64
```

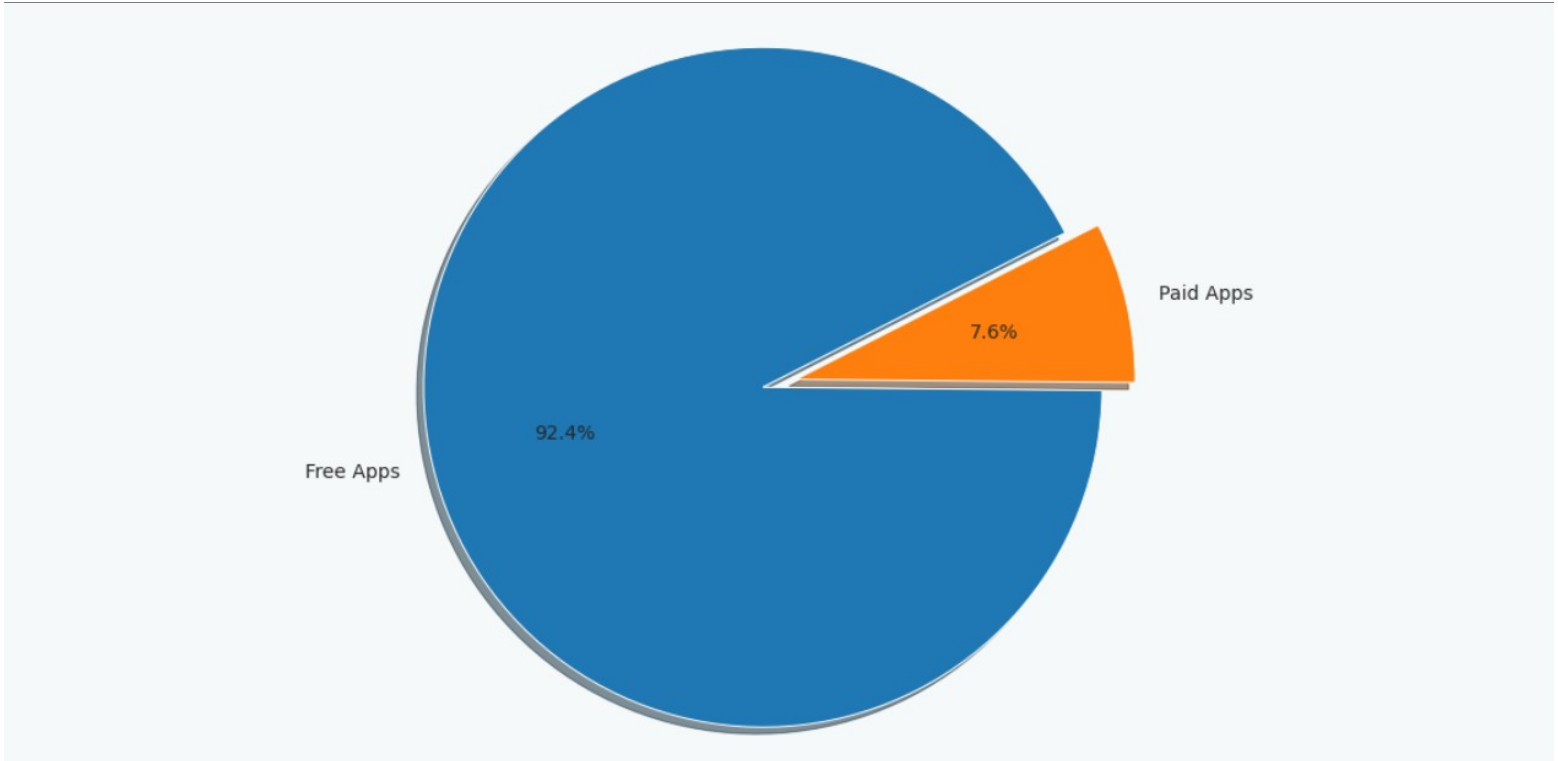
After Cleaning

Data Visualization

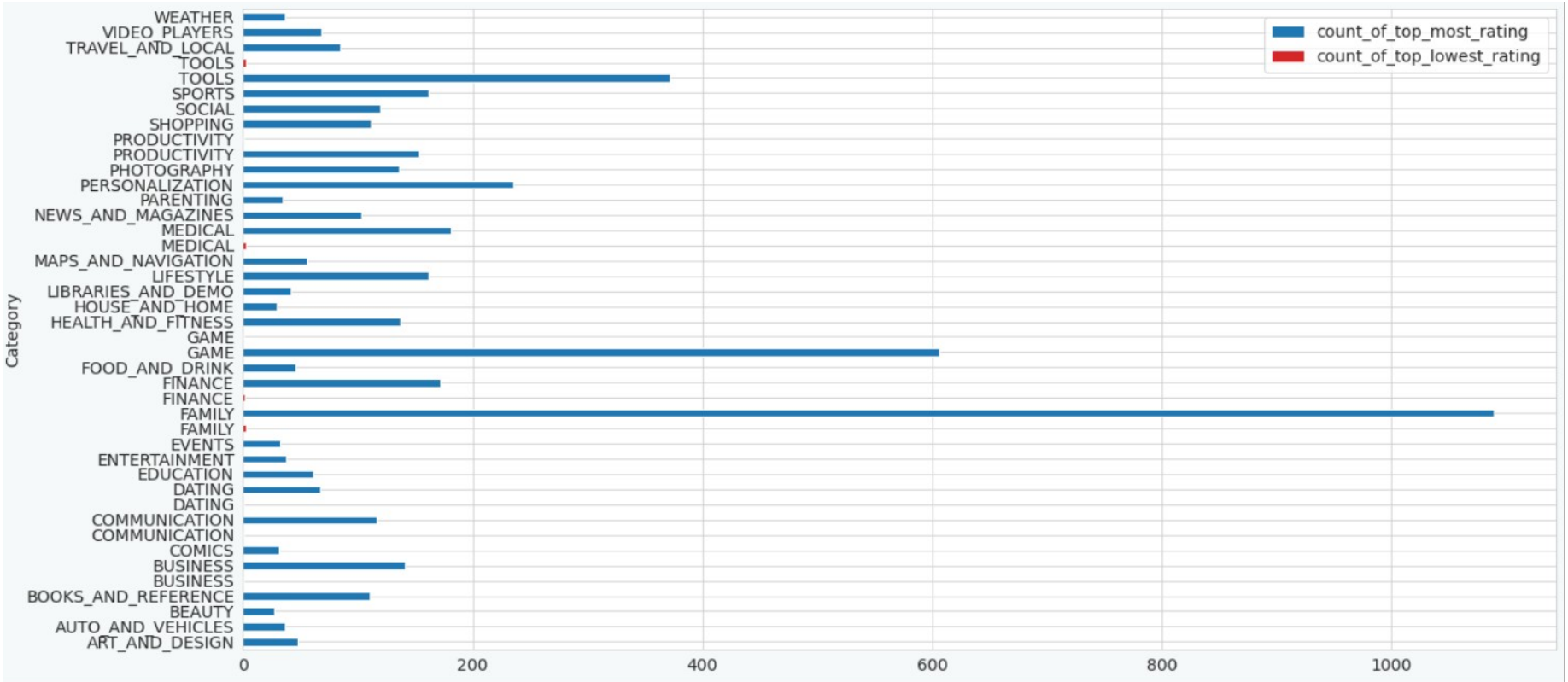
Which Category contains how many Apps



Free App and Paid App in percentage

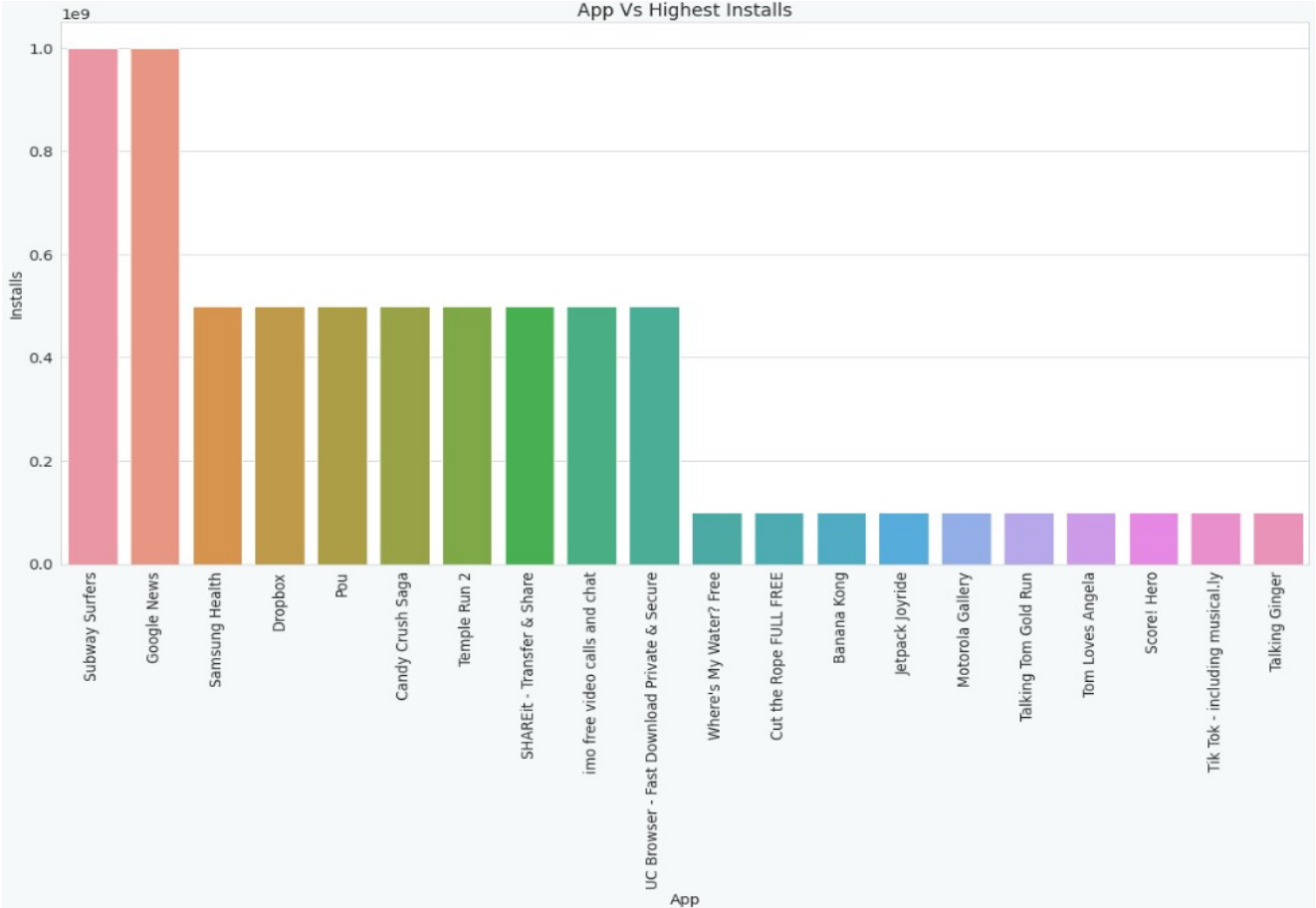


Rating over 4 (Blue) & Rating equals 1(Red)

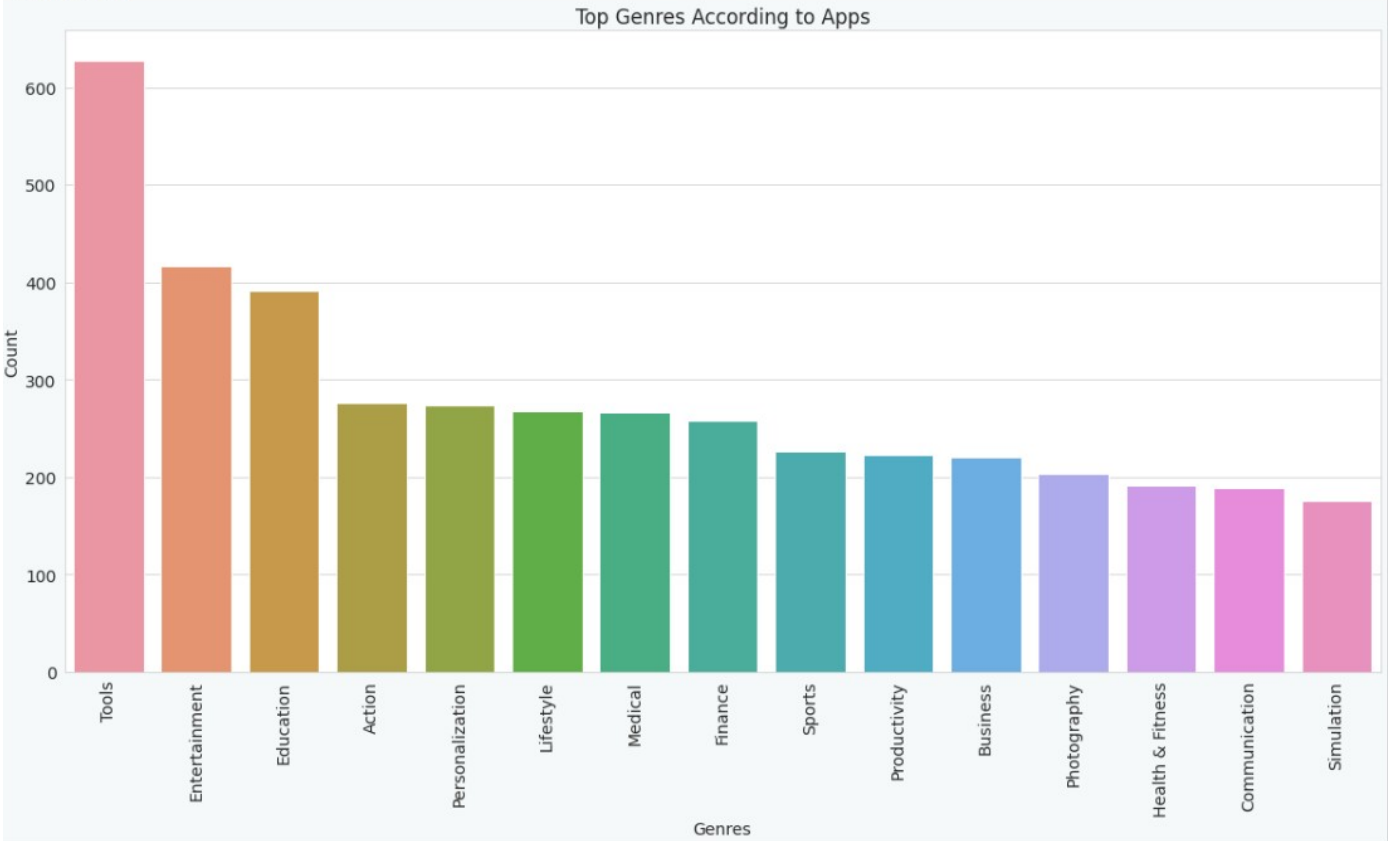


Top Most Reviewed Apps

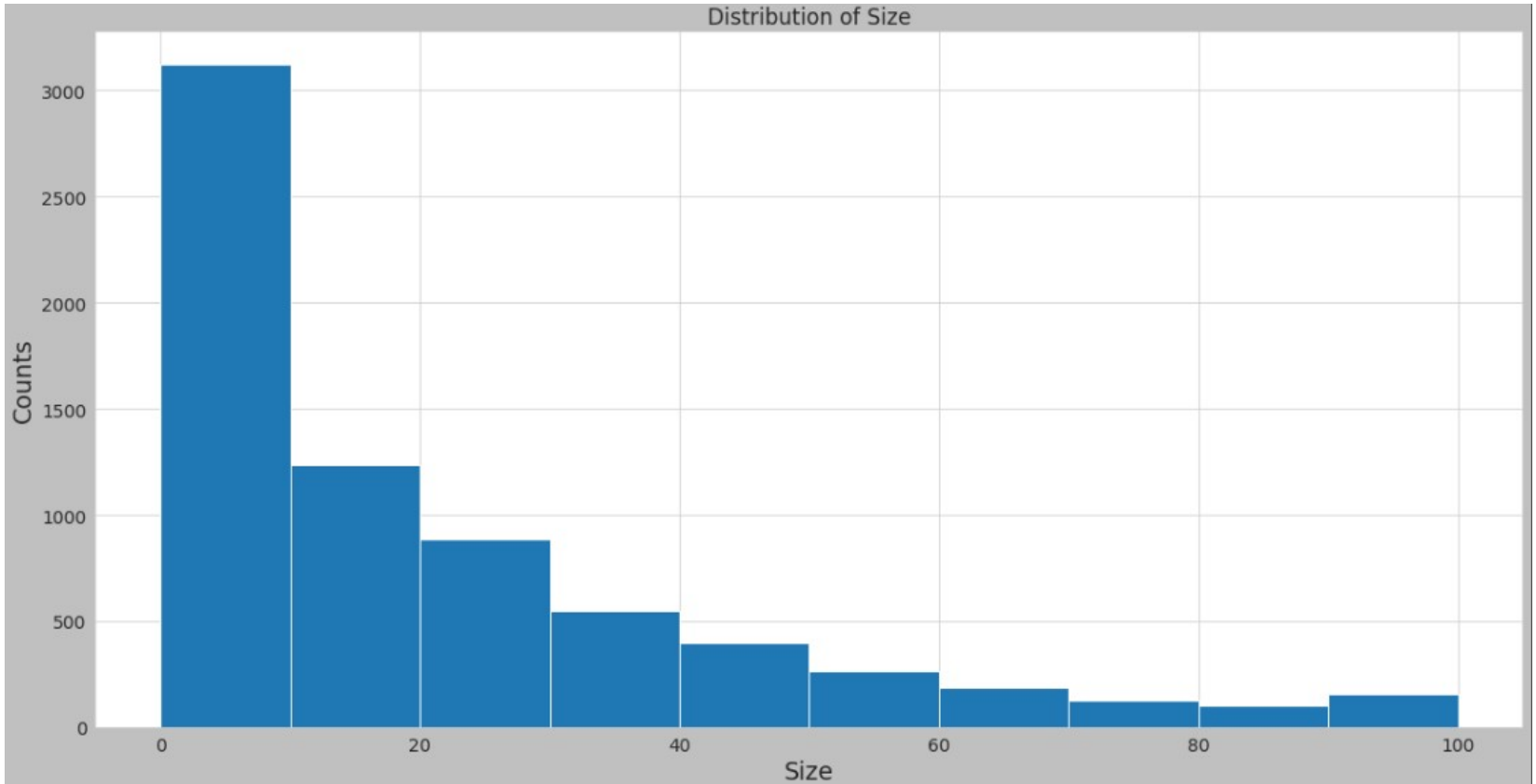
Most Downloaded App



Top Genres based on Apps



Size of Application over Number of App



User Reviews Dataset

Column Overviews :

- Sentiment basically determines the attitude or the emotion of the user. e.g. (whether it is positive or negative or neutral).

- Sentiment Polarity is float which lies in the range of [-1,1]
    - " 1 -> Positive Statement
    - " -1 -> Negative Statement
- Sentiment Subjectivity generally refer to personal opinion, emotion or judgment, which lies in the range of [0,1].

- Translated_Review is short explanation of feedback on App by user after using App.

Data Exploration

First Look of Data

	App	Translated_Review	Sentiment	Sentiment_Polarity	Sentiment_Subjectivity
0	10 Best Foods for You	I like eat delicious food. That's I'm cooking ...	Positive	1.00	0.533333
1	10 Best Foods for You	This help eating healthy exercise regular basis	Positive	0.25	0.288462
2	10 Best Foods for You	NaN	NaN	NaN	NaN
3	10 Best Foods for You	Works great especially going grocery store	Positive	0.40	0.875000
4	10 Best Foods for You	Best idea us	Positive	1.00	0.300000
5	10 Best Foods for You	Best way	Positive	1.00	0.300000
6	10 Best Foods for You	Amazing	Positive	0.60	0.900000
7	10 Best Foods for You	NaN	NaN	NaN	NaN
8	10 Best Foods for You	Looking forward app,	Neutral	0.00	0.000000
9	10 Best Foods for You	It helpful site ! It help foods get !	Neutral	0.00	0.000000

All the column is categorical so we have to convert Sentiment_Polarity , Sentiment_Subjectivity to Numeric type.

Also we have to remove Nan value.

Nan Value Removed

Shape (64295 , 5)

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 64295 entries, 0 to 64294
Data columns (total 5 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   App                     64295 non-null  object
 1   Translated_Review       37427 non-null  object
 2   Sentiment               37432 non-null  object
 3   Sentiment_Polarity      37432 non-null  float64
 4   Sentiment_Subjectivity  37432 non-null  float64
dtypes: float64(2), object(3)
memory usage: 2.5+ MB
```

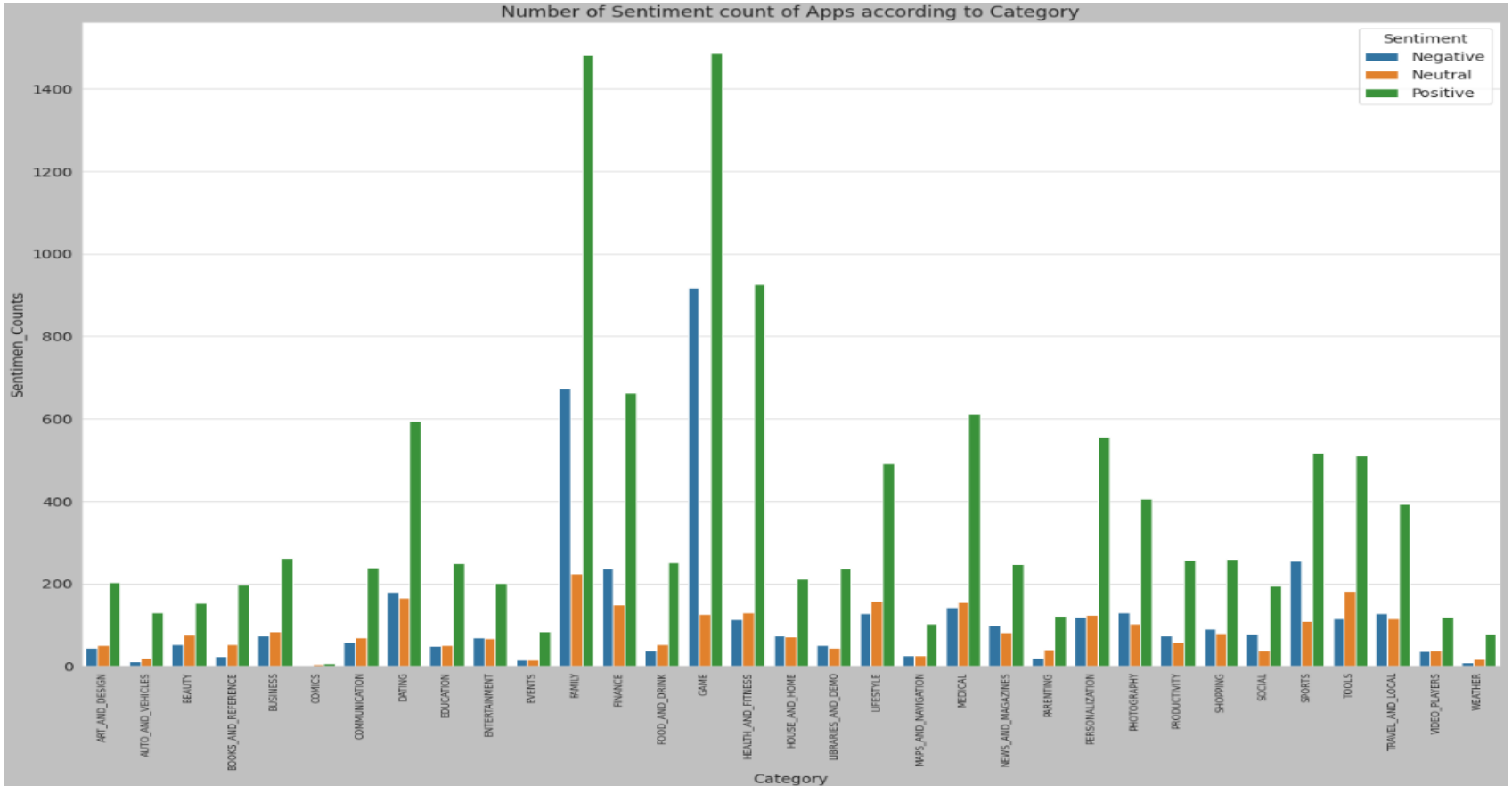
Before Cleaning

Shape (19349 , 17)

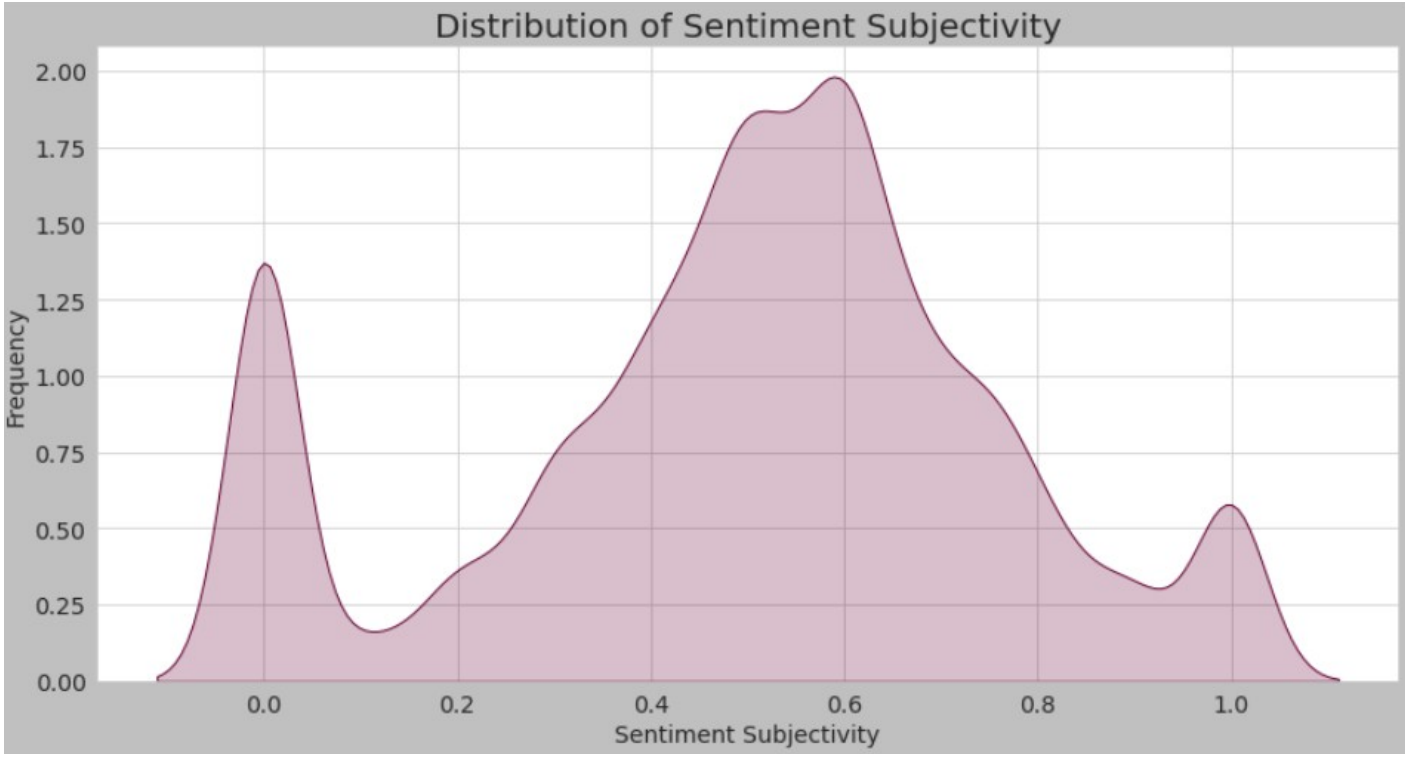
```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 19349 entries, 0 to 24100
Data columns (total 17 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   App                     19349 non-null  object
 1   Category                19349 non-null  object
 2   Rating                  19349 non-null  float64
 3   Reviews                 19349 non-null  int64
 4   Size                    19349 non-null  float32
 5   Installs                19349 non-null  int64
 6   Type                    19349 non-null  object
 7   Price                   19349 non-null  float64
 8   Content_Rating          19349 non-null  object
 9   Genres                  19349 non-null  object
 10  Last_Updated            19349 non-null  object
 11  Current_Ver             19349 non-null  object
 12  Android_Ver             19349 non-null  object
 13  Translated_Review       19349 non-null  object
 14  Sentiment               19349 non-null  object
 15  Sentiment_Polarity      19349 non-null  float64
 16  Sentiment_Subjectivity  19349 non-null  float64
dtypes: float32(1), float64(4), int64(2), object(10)
memory usage: 2.6+ MB
```

After Cleaning

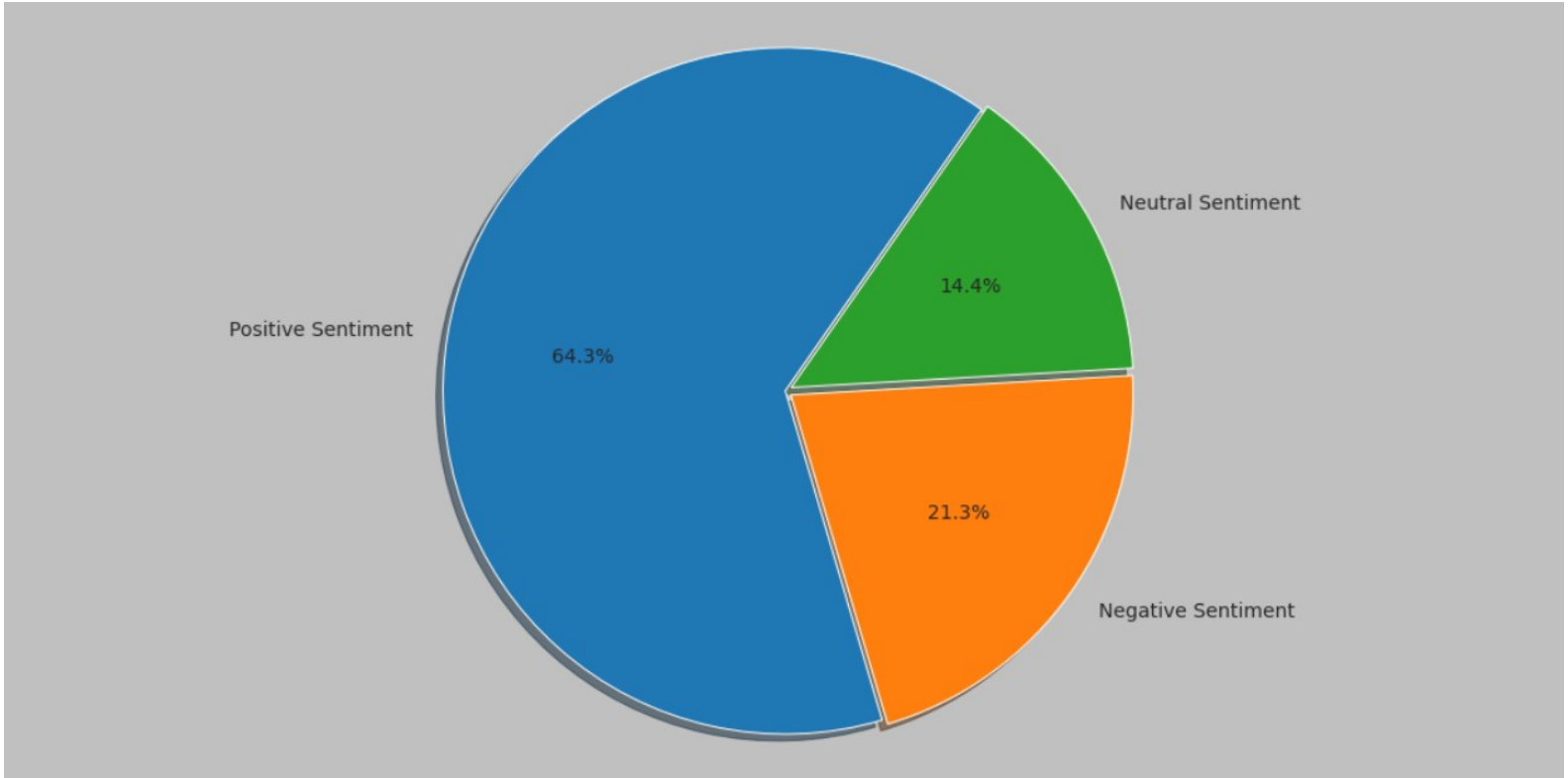
Number of Sentiments Counts of App according to Category



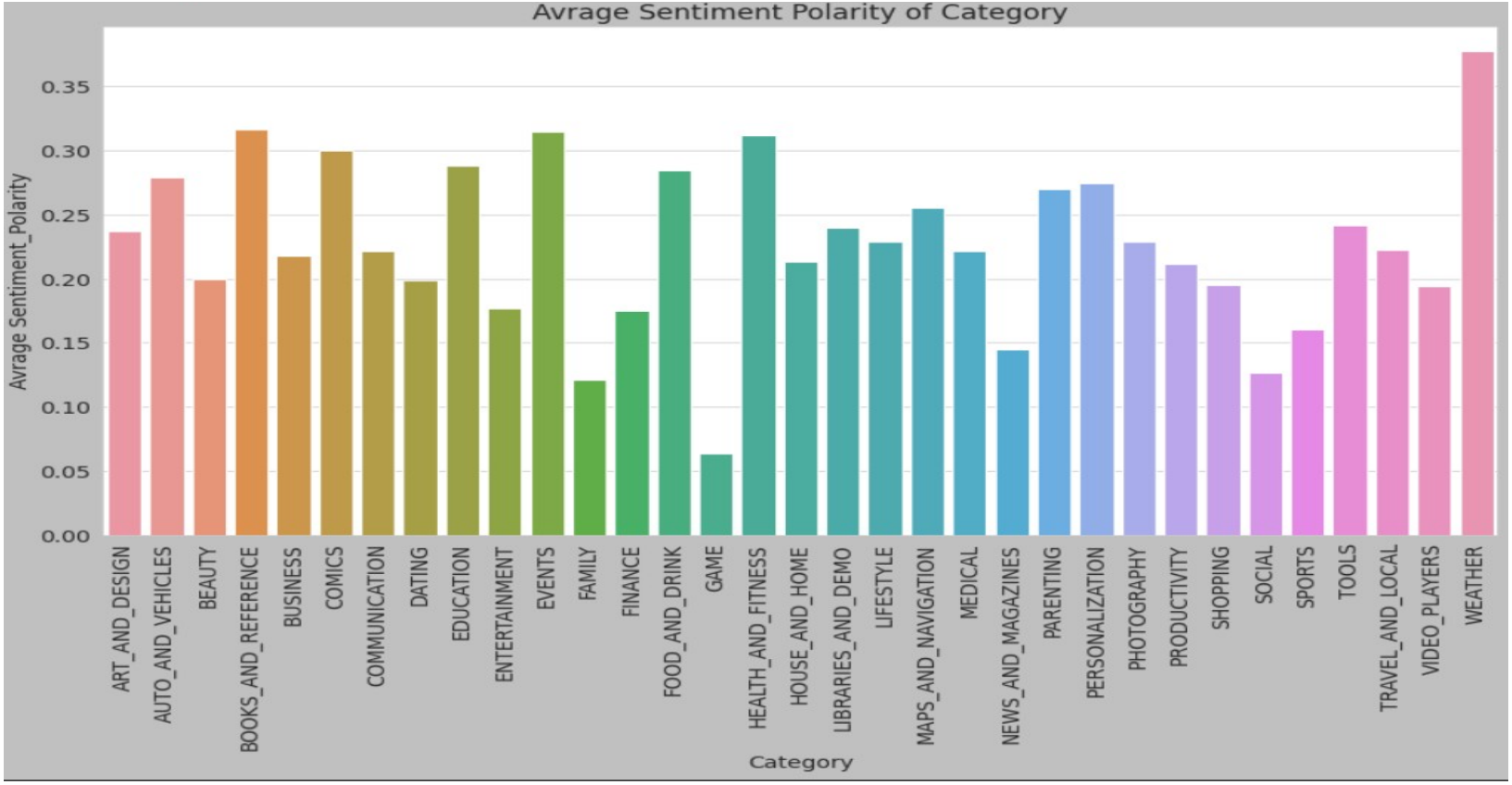
Distribution of Sentiment Subjectivity



Percentage of Different Sentiments



Average Sentiment Polarity of Category



Conclusion

1. Maximum Apps Belongs From Family and Game Category.
2. Play Store Contains 92.6% Free Apps and 7.4% Paid Apps.
3. (Family , Game , Tools) Category's App Got Maximum Rating where (Dating , Productivity , Business) Category's App Got Lower Rating(Equals 1).
4. Following Apps Got Maximum Reviews.
    1. Sketch - Draw & Paints.
    2. U Launcher Lite- Free Live Cool Themes,Hide Apps.
    3. Infinite Painter.
5. Subway Surferes , Google News , Samsung Health is Top Most Downloaded Apps on Play Store.
6. There are 3000+ Applications has Size between 0-10 MB, so User Prefers Small Size Apps Compared to Large one.
7. Installs & Reviews are highly positive correlated. The reason those application got Highest no of Installs they also got Highest no of Review.
8. Family and Game Category gains Maximum Postive Sentiment.
9. Most Apps emotion,personal opinion, judgment lays Between 0.4 to 0.7
10. App Feedback(Sentiment) Given by User As Follow.
    1. 64.3% - Positive Sentiment
    2. 21.3% - Negative Sentiment
    3. 14.4% - Neutral Sentiment

Thank You
For Your Attention