Table of Contents

Abstract

The following data is downloaded from Kaggle about Google Play Store Apps. The contributor scraped 10k PlayStore apps to analyze the Android apps market. In this report, we will look for the characteristics of apps that are popular, relatively higher number of downloads, the most reviews and mid to high rating. We will see if there are any relationships or correlations between each category and how they would contribute or relate to each other. How big of an impact does one category have on each other. Another thing to consider is whether the app is free or paid as this would probably be a major contributor to the success or popularity of the app.
We will start off by taking a quick look at the data to draw some initial thoughts about the data. Pandas package will come in handy to summarize the data. Data cleaning such as missing, duplicate and outliers will be check and necessary measures will be taken to prepare the data for analysis.

In the analysis phase, we will try to see if there are any correlations in the data that could lead to interesting findings and insights. Some potential visualizations we will use are pie charts, bar graphs or scatter plots to map out the number of app categories, count of apps, total downloads/installations and number of reviews for each category. Box plots to look at the relationships between categories, ratings or downloads. Histograms to have a better understanding of the download and rating distribution. In the end, we were hoping to be able to generate insights and recommendations to figure out how app developers can maximize their chance of success.

Dataset Details:
Google Play Store App data, it has 13 columns and 10,841 rows, features include app category, ratings, reviews, app sizes, number of installs, free/paid, pricing and genres. Link: https://www.kaggle.com/datasets/lava18/google-play-store-apps

1. Introduction and Background

Mobile apps have shaped the world and our way of living in many ways. From transport, groceries, productivity, banking, there is almost an app for everything we do, and this has fundamentally changed the way human get things done. Some of the most impactful apps in today's world are arguably social network apps, productivity apps and gaming apps. We can literally 'google' almost every information we need and all on our fingertips which has transformed our work productivity and simplified our daily tasks.

The widespread use of mobile apps has also created new opportunities for researchers and businesses. It is now possible to gather data in real time since people have their phones on them almost the entire time. Additionally, sensors on smartphones and other wearable devices offer even detailed and comprehensive data collection (Lemmens et al, 2021). For volunteers taking part in scientific observations, deployment of apps proved to be a game changer, not only volunteers can contribute observations in real time but also the data quality has improved significantly (Lemmens et al, 2021).

According to Statista (Statista, 2022), there are around 2.65 million apps on Google Play Store in June 2022. The number of apps surpassed the 1 million marks in 2013 and has been growing rapidly ever since. Play Store offers users numerous ranges of applications from games, music, movies, books and many more. But the top grossing apps are dominated by the games category some of which are PUBG and Candy Crush Saga (Statista, 2022). We can see similar trend in App Store where in 2021, bulk of revenue share, around 61%, are also coming from games apps (Statista, 2021). In 2021, most popular app categories were gaming followed by education and business apps. By 2024, it estimated that more than 70 percent of Play Store revenue will be coming from games, a slight decrease from 83 percent in 2020 (Statista, 2021).

In this exercise, we will be conducting data analysis on Google Play Store dataset to draw insights on what are the characteristics of successful app and what developers should consider when building new apps. We will start off by:

-Exploring the dataset
-Familiarising ourselves with relevant attributes
-Cleaning and preparing the data for analysis
-Visualizing the data to understand the data and identify patterns
-Summary of the findings and insights

1.1 Problem

With millions of apps out there, it can be a daunting task to make sure your app stands out among the crowd, here we will try to identify how to maximize the chances of succeeding. Depending on the objective or purpose of the developers, if they are looking to maximize utilization of their apps, they would need to know the current state of the Play Store market.

1.2 Business Question

Every developer might have different objectives or goals in mind when developing apps. But most probably, most developers would want to increase their chances of success in terms of app popularity, what type of apps should they be focusing on?

1.3 Dataset

Brief summary table:

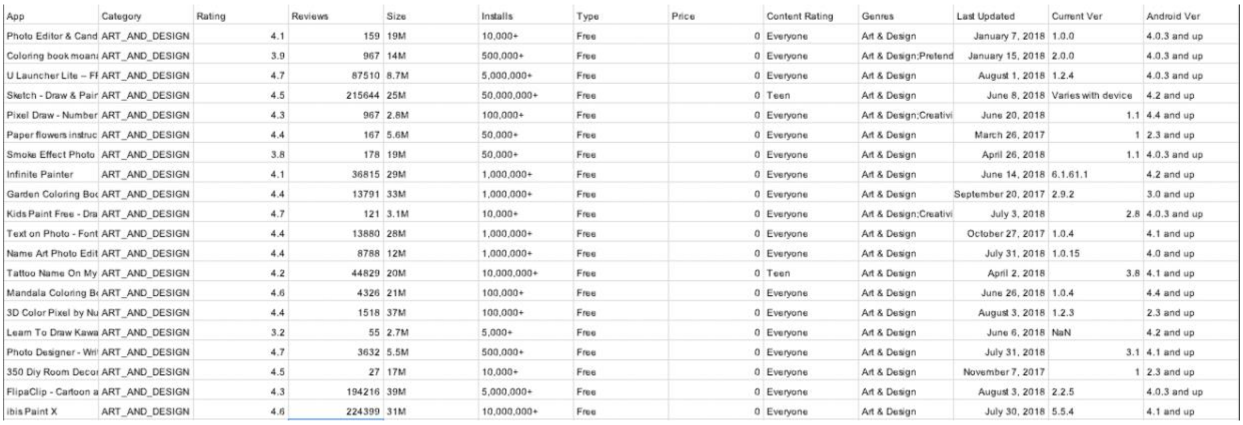
Item	Description
Dataset	Google Play Store apps
Number of rows	10,841
Number of columns	13
Column names	App, Category, Rating, Reviews, Size, Installs, Type, Price, Content Rating, Genres, Last Update, Current Ver, Android Ver

We will be exploring certain key attributes including Category, Rating, Reviews, Installs, Type and Price in our analysis to identify the correlation between them

Link:

https://www.kaggle.com/datasets/lava18/google-play-store-apps

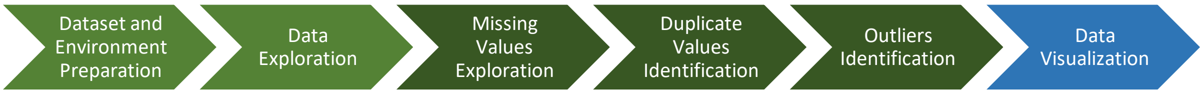
Screenshot of the dataset:



App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
Photo Editor & Cand	ART_AND_DESIGN	4.1	159	19M	10,000+	Free	0	Everyone	Art & Design	January 7, 2018	1.0.0	4.0.3 and up
Coloring book moan	ART_AND_DESIGN	3.9	967	14M	500,000+	Free	0	Everyone	Art & Design;Pretend	January 15, 2018	2.0.0	4.0.3 and up
U Launcher Lite – FI	ART_AND_DESIGN	4.7	87510	8.7M	5,000,000+	Free	0	Everyone	Art & Design	August 1, 2018	1.2.4	4.0.3 and up
Sketch - Draw & Pair	ART_AND_DESIGN	4.5	215644	25M	50,000,000+	Free	0	Teen	Art & Design	June 8, 2018	Varies with device	4.2 and up
Pixel Draw - Number	ART_AND_DESIGN	4.3	967	2.8M	100,000+	Free	0	Everyone	Art & Design;Creativi	June 20, 2018	1.1	4.4 and up
Paper flowers instruc	ART_AND_DESIGN	4.4	167	5.6M	50,000+	Free	0	Everyone	Art & Design	March 26, 2017	1	2.3 and up
Smoke Effect Photo	ART_AND_DESIGN	3.8	178	19M	50,000+	Free	0	Everyone	Art & Design	April 26, 2018	1.1	4.0.3 and up
Infinite Painter	ART_AND_DESIGN	4.1	36815	29M	1,000,000+	Free	0	Everyone	Art & Design	June 14, 2018	6.1.61.1	4.2 and up
Garden Coloring Boc	ART_AND_DESIGN	4.4	13791	33M	1,000,000+	Free	0	Everyone	Art & Design	September 20, 2017	2.9.2	3.0 and up
Kids Paint Free - Dra	ART_AND_DESIGN	4.7	121	3.1M	10,000+	Free	0	Everyone	Art & Design;Creativi	July 3, 2018	2.8	4.0.3 and up
Text on Photo - Font	ART_AND_DESIGN	4.4	13880	28M	1,000,000+	Free	0	Everyone	Art & Design	October 27, 2017	1.0.4	4.1 and up
Name Art Photo Edit	ART_AND_DESIGN	4.4	8788	12M	1,000,000+	Free	0	Everyone	Art & Design	July 31, 2018	1.0.15	4.0 and up
Tattoo Name On My	ART_AND_DESIGN	4.2	44829	20M	10,000,000+	Free	0	Teen	Art & Design	April 2, 2018	3.8	4.1 and up
Mandala Coloring Bc	ART_AND_DESIGN	4.6	4326	21M	100,000+	Free	0	Everyone	Art & Design	June 26, 2018	1.0.4	4.4 and up
3D Color Pixel by Nu	ART_AND_DESIGN	4.4	1518	37M	100,000+	Free	0	Everyone	Art & Design	August 3, 2018	1.2.3	2.3 and up
Learn To Draw Kawa	ART_AND_DESIGN	3.2	55	2.7M	5,000+	Free	0	Everyone	Art & Design	June 6, 2018	NaN	4.2 and up
Photo Designer - Wri	ART_AND_DESIGN	4.7	3632	5.5M	500,000+	Free	0	Everyone	Art & Design	July 31, 2018	3.1	4.1 and up
350 Diy Room Decor	ART_AND_DESIGN	4.5	27	17M	10,000+	Free	0	Everyone	Art & Design	November 7, 2017	1	2.3 and up
FlipaClip - Cartoon a	ART_AND_DESIGN	4.3	194216	39M	5,000,000+	Free	0	Everyone	Art & Design	August 3, 2018	2.2.5	4.0.3 and up
ibis Paint X	ART_AND_DESIGN	4.6	224399	31M	10,000,000+	Free	0	Everyone	Art & Design	July 30, 2018	5.5.4	4.1 and up

2 Overview of the Data Analysis Pipeline

2.1 Flow Diagram/Flowchart/Work Flow



Dataset and Environment Preparation

Data Exploration

Missing Values Exploration

Duplicate Values Identification

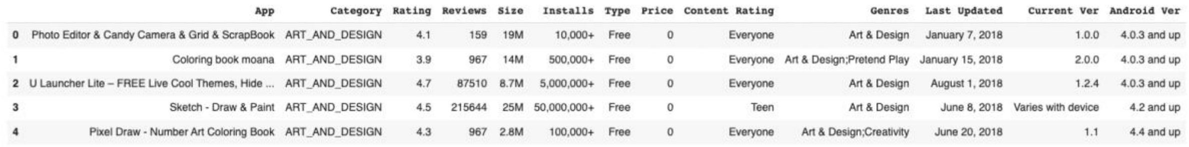
Outliers Identification

Data Visualization

2.2 Data Preparation

After mounting google drive to google colab, we started reading the csv file and saving it. First, we try to get a better sense of the dataset and see if there is any data cleaning needed to be done prior to analysis.

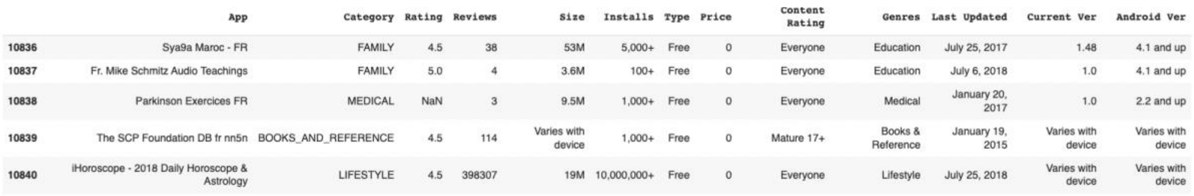
A. First 5 rows of the dataset:



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

B. Last 5 rows of the dataset:

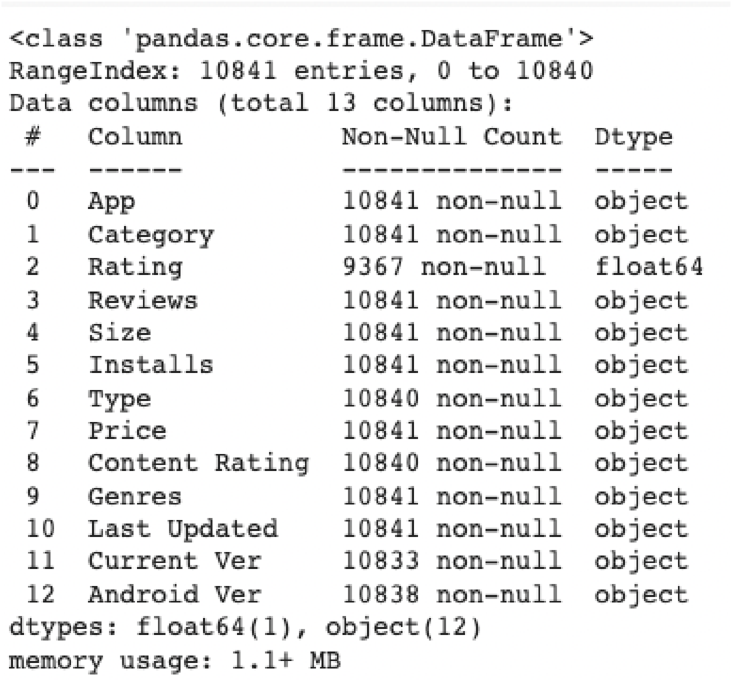
These gave us a quick glimpse into the dataset, to understand the data in its raw state to kick off the exploration phase. Next, we dive deeper into the dataset by looking at the data types and statistical summary



	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
10836	Sya9a Maroc - FR	FAMILY	4.5	38	53M	5,000+	Free	0	Everyone	Education	July 25, 2017	1.48	4.1 and up
10837	Fr. Mike Schmitz Audio Teachings	FAMILY	5.0	4	3.6M	100+	Free	0	Everyone	Education	July 6, 2018	1.0	4.1 and up
10838	Parkinson Exercices FR	MEDICAL	NaN	3	9.5M	1,000+	Free	0	Everyone	Medical	January 20, 2017	1.0	2.2 and up
10839	The SCP Foundation DB fr nn5n	BOOKS_AND_REFERENCE	4.5	114	Varies with device	1,000+	Free	0	Mature 17+	Books & Reference	January 19, 2015	Varies with device	Varies with device
10840	iHoroscope - 2018 Daily Horoscope & Astrology	LIFESTYLE	4.5	398307	19M	10,000,000+	Free	0	Everyone	Lifestyle	July 25, 2018	Varies with device	Varies with device

C. Looking at each column/attribute:

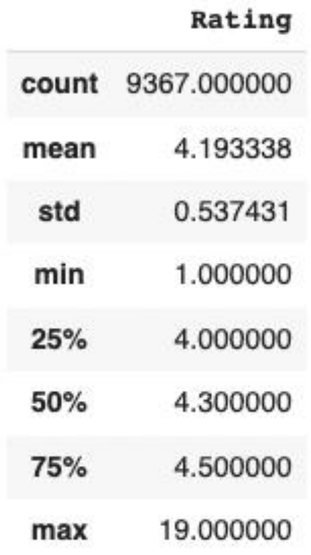
According to the result, looks like there is only 1 numeric data types out of the 13 columns, we might need to modify the data type later in the analysis



```
<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
```

D. Statistical summary of the dataset:

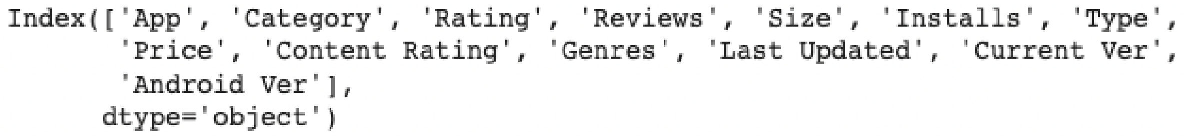
-We know apps rating only go as high as 5/5, here we see the max value is 19, this could be an outlier
-Min value is 1 which makes sense considering the lowest rating we can give an app is 1 out of 5 stars.
-The row count is 9,367 but we know there are 10,841 rows in total, this could mean there are null values
-Overall, it looks like the apps in the data set are fairly well rated apps, considering the mean and median ratings are 4.193 and 4.3 respectively



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

E. Describing the name of the columns in the dataset

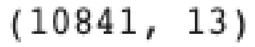
Here we can see the list of the column names in the dataset, this will be our guidance when we do further analysis



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

F. Total number of rows and columns

In total there are 10,841 rows and 13 columns in the dataset



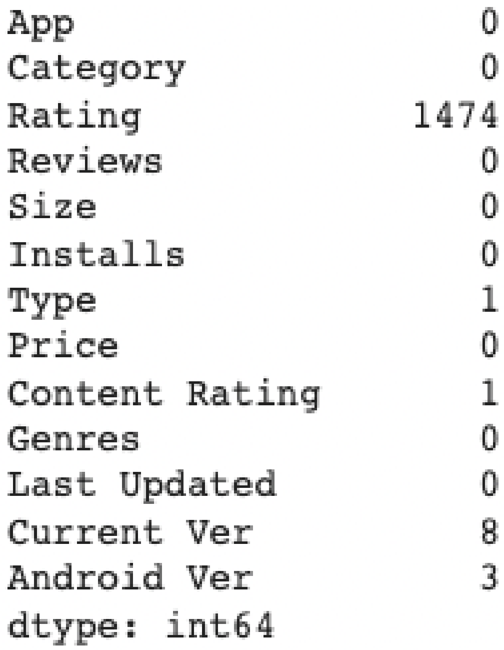
```
(10841, 13)
```

2.3 Missing value exploration

In the next phase, we will attempt to explore the missing values in the dataset, several techniques are used to identify missing values and remove them to prevent skewing the final analysis results, they are as follow:

A. Checking the total missing values in every column

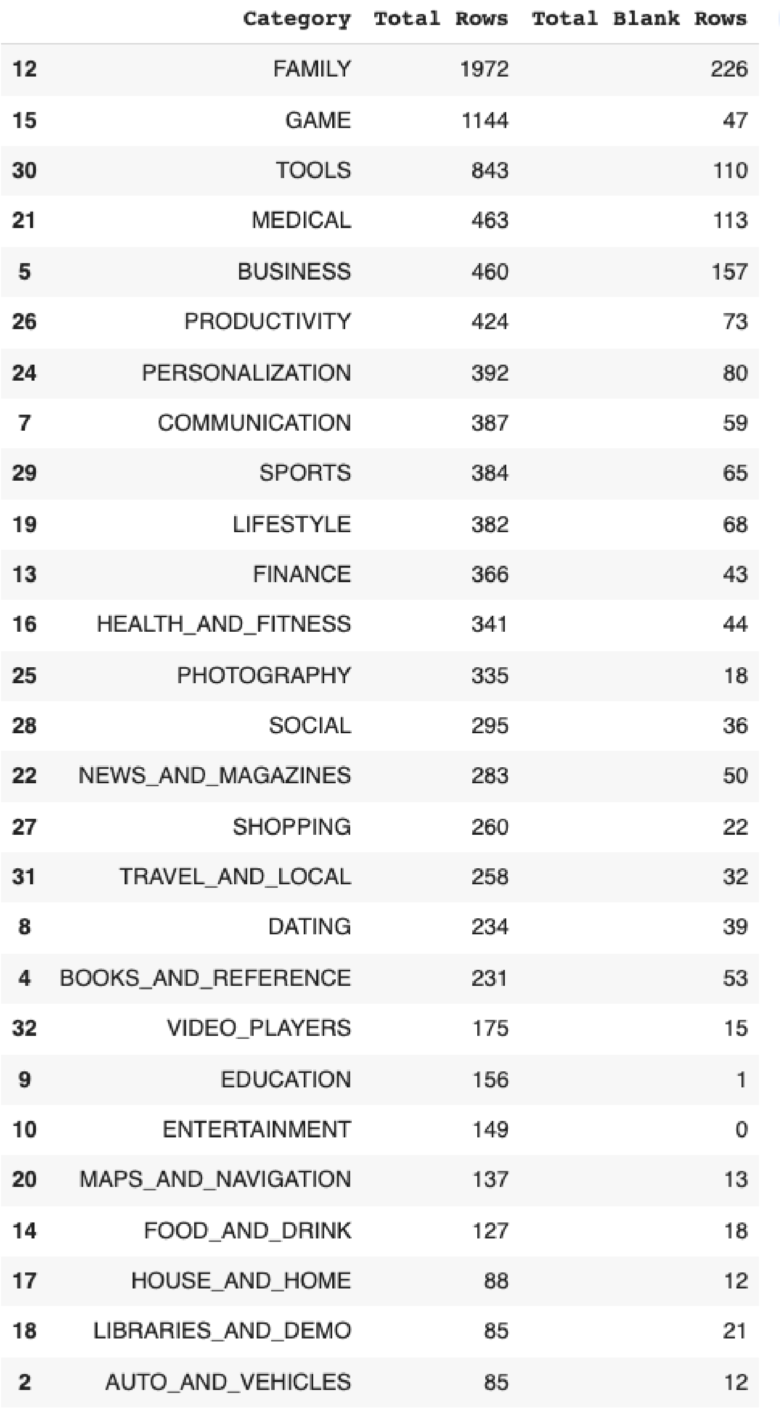
Evidently there are 1,474 missing values in 'Rating' columns and also some missing values in other columns, this proved our presumption in **section 2.2 D**



```
App                  0
Category             0
Rating            1474
Reviews              0
Size                 0
Installs             0
Type                 1
Price                0
Content Rating       1
Genres               0
Last Updated         0
Current Ver          8
Android Ver          3
dtype: int64
```

B. Missing values per category

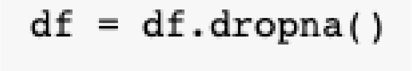
Since category is one of the main attributes we are going to use for the further analysis, we want to make sure that all the missing values are not concentrated in just a handful of categories, which might have unintended consequences to our analysis if remove them. We know the majority of the missing values are in the ‘rating’ column, here we can see which categories has missing ‘rating’ values and how many. Turns out the missing values are distributed among all categories not concentrated in just a few categories, which is a good thing.



	Category	Total Rows	Total Blank Rows
12	FAMILY	1972	226
15	GAME	1144	47
30	TOOLS	843	110
21	MEDICAL	463	113
5	BUSINESS	460	157
26	PRODUCTIVITY	424	73
24	PERSONALIZATION	392	80
7	COMMUNICATION	387	59
29	SPORTS	384	65
19	LIFESTYLE	382	68
13	FINANCE	366	43
16	HEALTH_AND_FITNESS	341	44
25	PHOTOGRAPHY	335	18
28	SOCIAL	295	36
22	NEWS_AND_MAGAZINES	283	50
27	SHOPPING	260	22
31	TRAVEL_AND_LOCAL	258	32
8	DATING	234	39
4	BOOKS_AND_REFERENCE	231	53
32	VIDEO_PLAYERS	175	15
9	EDUCATION	156	1
10	ENTERTAINMENT	149	0
20	MAPS_AND_NAVIGATION	137	13
14	FOOD_AND_DRINK	127	18
17	HOUSE_AND_HOME	88	12
18	LIBRARIES_AND_DEMO	85	21
2	AUTO_AND_VEHICLES	85	12

C. Removing all missing values from the dataset

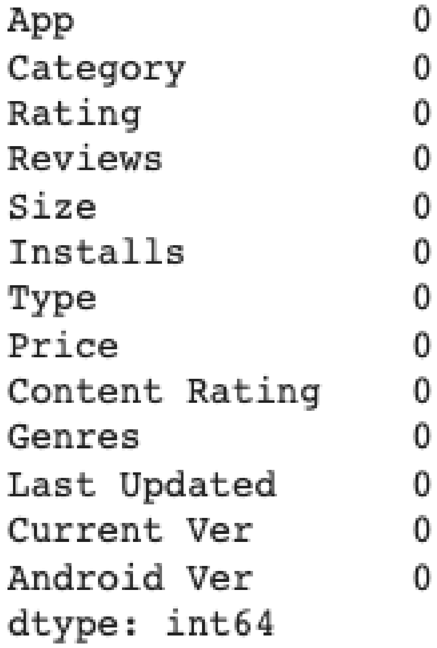
We checked and verified that the missing values are distributed across the categories so we decided better to remove the missing values to prevent the missing values skewing the end result of our analysis.



```
df = df.dropna()
```

D. Checking for missing values in the dataset

We performed the final check to confirm there are no more missing values in the dataset, we can see in this result that there are no more missing values, we can move on to the next phase which we will check for duplicate values and outliers in the dataset.



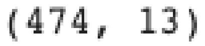
```
App               0
Category          0
Rating            0
Reviews           0
Size              0
Installs          0
Type              0
Price             0
Content Rating    0
Genres            0
Last Updated      0
Current Ver       0
Android Ver       0
dtype: int64
```

2.4 Duplicate values identification

In this phase, we will check if there are any duplicate values in the dataset and also to modify the data types to prepare for analysis

A. Duplicate value

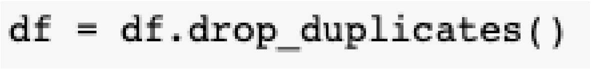
Looks like there are total of 474 duplicate values in the dataset which might have negative effect in the analysis



```
(474, 13)
```

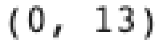
B. Duplicate removal

We removed all the duplicate values before proceeding



```
df = df.drop_duplicates()
```

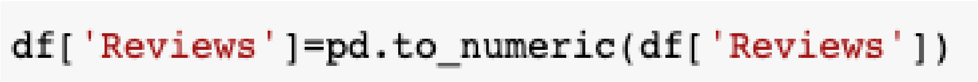
C. Making sure there are no more duplicate values after removal



```
(0, 13)
```

D. We noticed that there is a column with inaccurate data type which is the 'Reviews' column, we will modify its data type to the appropriate one.

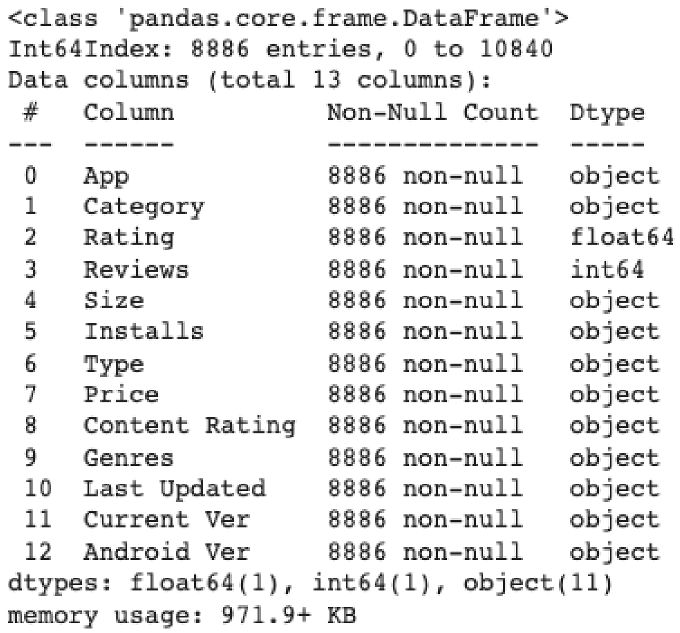
Converted 'Reviews' to numeric instead of object data type



```
df['Reviews']=pd.to_numeric(df['Reviews'])
```

E. Taking another look at the dataset columns/attributes

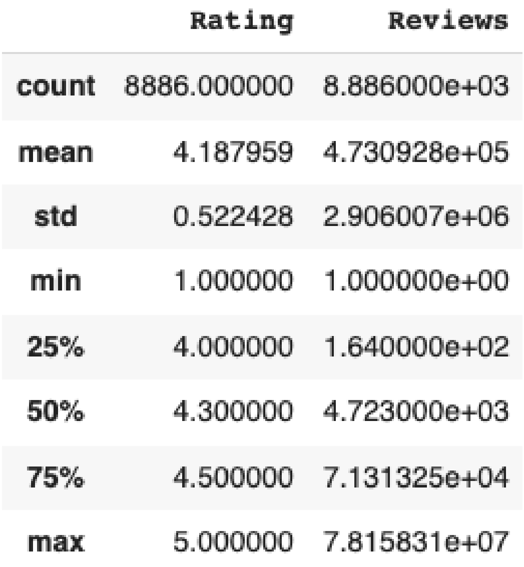
'Reviews' column has been successfully converted to numeric



```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 8886 entries, 0 to 10840
Data columns (total 13 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   App             8886 non-null   object
 1   Category        8886 non-null   object
 2   Rating          8886 non-null   float64
 3   Reviews         8886 non-null   int64
 4   Size            8886 non-null   object
 5   Installs        8886 non-null   object
 6   Type            8886 non-null   object
 7   Price           8886 non-null   object
 8   Content Rating  8886 non-null   object
 9   Genres          8886 non-null   object
 10  Last Updated    8886 non-null   object
 11  Current Ver     8886 non-null   object
 12  Android Ver     8886 non-null   object
dtypes: float64(1), int64(1), object(11)
memory usage: 971.9+ KB
```

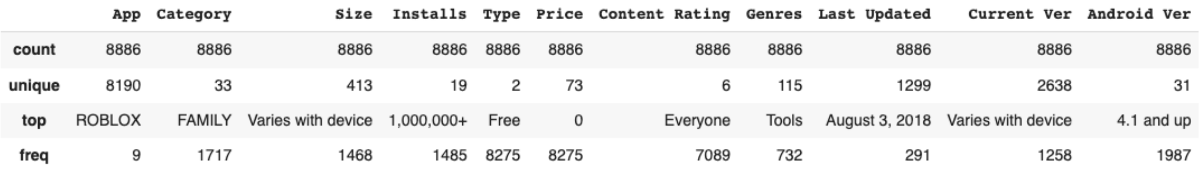
F. Statistical summary of the dataset

Now we have statistic summary of two columns since ‘Reviews’ column has been converted to numerical data type.



	Rating	Reviews
count	8886.000000	8.886000e+03
mean	4.187959	4.730928e+05
std	0.522428	2.906007e+06
min	1.000000	1.000000e+00
25%	4.000000	1.640000e+02
50%	4.300000	4.723000e+03
75%	4.500000	7.131325e+04
max	5.000000	7.815831e+07

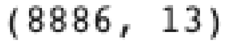
G. Statistical summary of columns other than numerical data type



	App	Category	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
count	8886	8886	8886	8886	8886	8886	8886	8886	8886	8886	8886
unique	8190	33	413	19	2	73	6	115	1299	2638	31
top	ROBLOX	FAMILY	Varies with device	1,000,000+	Free	0	Everyone	Tools	August 3, 2018	Varies with device	4.1 and up
freq	9	1717	1468	1485	8275	8275	7089	732	291	1258	1987

H. Showing the total rows and columns after removing missing and duplicate values

Now we have 8,886 rows and 13 columns in the dataset

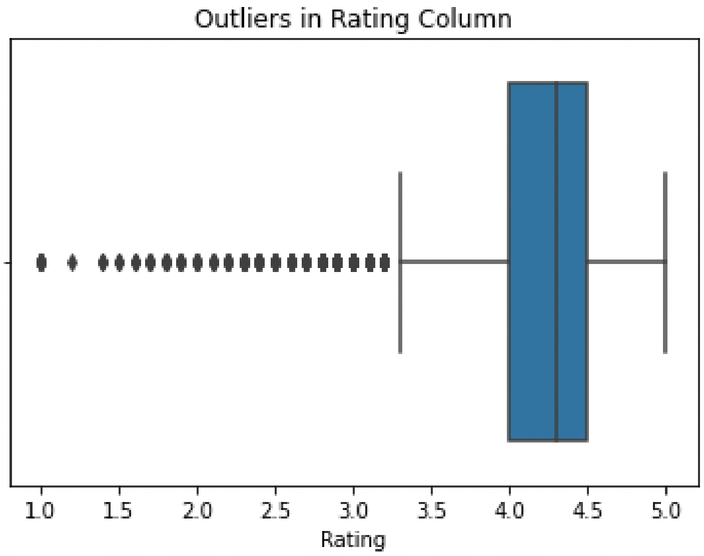


```
(8886, 13)
```

2.5 Outliers identification

In this section, we will identify outliers in our dataset:

A. Checking for outliers in the 'Rating' column



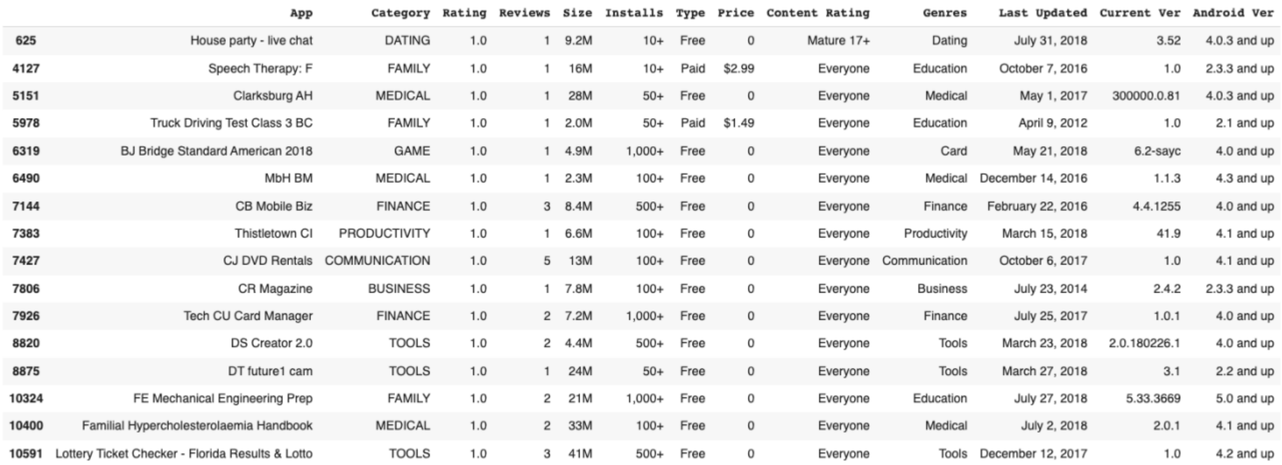
B. Calculating the number of outliers in 'Rating' column

There are 494 outliers according to the boxplot, it looks like apps with ratings 3.3 and below are considered outliers, but I think it is very possible and natural that certain apps will have 3.3 rating or less in the Play Store, even a 1/5 rating app exists in the Play Store since some apps might have very poor performance. There are also many cases of harmful or fraudulent app in the Play Store. It's possible these apps might receive the lowest rating possible.

```
There are total 494 outliers in Rating column
```

C. Showing apps with the lowest rating

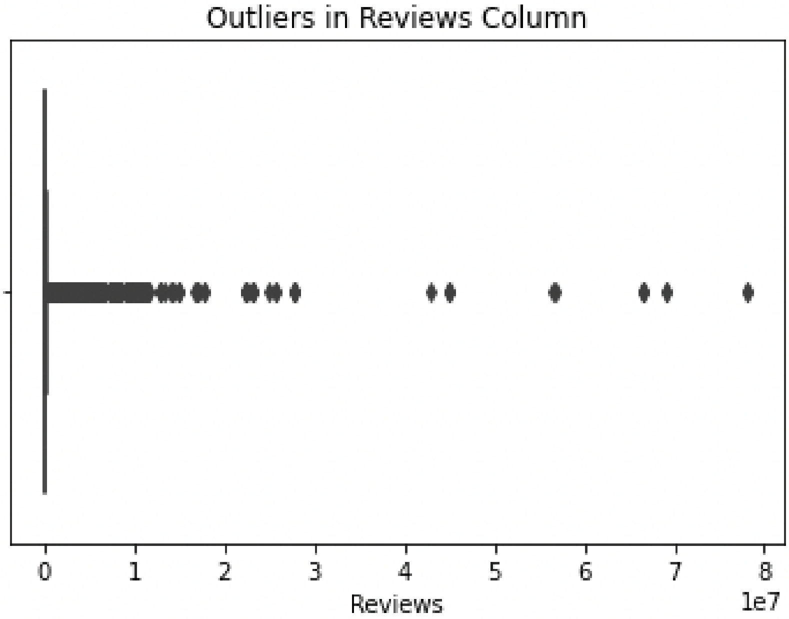
The bottom lowest ratings apps have less than 5 reviews on them and most of them have less than 1000 installs. It could be due the apps quality being on the poorer side which makes them less popular.



	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
625	House party - live chat	DATING	1.0	1	9.2M	10+	Free	0	Mature 17+	Dating	July 31, 2018	3.52	4.0.3 and up
4127	Speech Therapy: F	FAMILY	1.0	1	16M	10+	Paid	$2.99	Everyone	Education	October 7, 2016	1.0	2.3.3 and up
5151	Clarksburg AH	MEDICAL	1.0	1	28M	50+	Free	0	Everyone	Medical	May 1, 2017	300000.0.81	4.0.3 and up
5978	Truck Driving Test Class 3 BC	FAMILY	1.0	1	2.0M	50+	Paid	$1.49	Everyone	Education	April 9, 2012	1.0	2.1 and up
6319	BJ Bridge Standard American 2018	GAME	1.0	1	4.9M	1,000+	Free	0	Everyone	Card	May 21, 2018	6.2-sayc	4.0 and up
6490	MbH BM	MEDICAL	1.0	1	2.3M	100+	Free	0	Everyone	Medical	December 14, 2016	1.1.3	4.3 and up
7144	CB Mobile Biz	FINANCE	1.0	3	8.4M	500+	Free	0	Everyone	Finance	February 22, 2016	4.4.1255	4.0 and up
7383	Thistletown CI	PRODUCTIVITY	1.0	1	6.6M	100+	Free	0	Everyone	Productivity	March 15, 2018	41.9	4.1 and up
7427	CJ DVD Rentals	COMMUNICATION	1.0	5	13M	100+	Free	0	Everyone	Communication	October 6, 2017	1.0	4.1 and up
7806	CR Magazine	BUSINESS	1.0	1	7.8M	100+	Free	0	Everyone	Business	July 23, 2014	2.4.2	2.3.3 and up
7926	Tech CU Card Manager	FINANCE	1.0	2	7.2M	1,000+	Free	0	Everyone	Finance	July 25, 2017	1.0.1	4.0 and up
8820	DS Creator 2.0	TOOLS	1.0	2	4.4M	500+	Free	0	Everyone	Tools	March 23, 2018	2.0.180226.1	4.0 and up
8875	DT future1 cam	TOOLS	1.0	1	24M	50+	Free	0	Everyone	Tools	March 27, 2018	3.1	2.2 and up
10324	FE Mechanical Engineering Prep	FAMILY	1.0	2	21M	1,000+	Free	0	Everyone	Education	July 27, 2018	5.33.3669	5.0 and up
10400	Familial Hypercholesterolaemia Handbook	MEDICAL	1.0	2	33M	100+	Free	0	Everyone	Medical	July 2, 2018	2.0.1	4.1 and up
10591	Lottery Ticket Checker - Florida Results & Lotto	TOOLS	1.0	3	41M	500+	Free	0	Everyone	Tools	December 12, 2017	1.0	4.2 and up

D. Checking for outliers in ‘Reviews’ column

This figure looks odd, it could be due to huge range of review count from hundreds to millions, making the boxplot difficult to interpret.



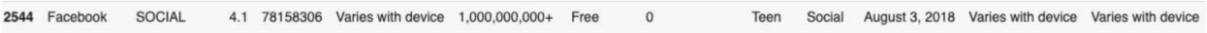
E. Calculating total number of outliers in ‘Reviews’ column

Looks like there a whopping 1,555 outliers in the ‘Reviews’ column, I doubt these are errors or incorrect input. Since apps popularity vary so much, it is very likely some apps have considerably larger count of reviews compare to the majority of less popular apps, we will try to pull the app with the most reviews in the next figure

```
There are total 1555 in Reviews column
```

F. App with the most reviews

Facebook has the highest review among the apps in the dataset but is considered an outlier. Considering the popularity of Facebook which has 2.93 billion monthly active users in the second quarter of 2022 (Statista, 2022), it makes sense Facebook has more than 78 million reviews in the Play Store.



2544	Facebook	SOCIAL	4.1	78158306	Varies with device	1,000,000,000+	Free	0	Teen	Social	August 3, 2018	Varies with device	Varies with device

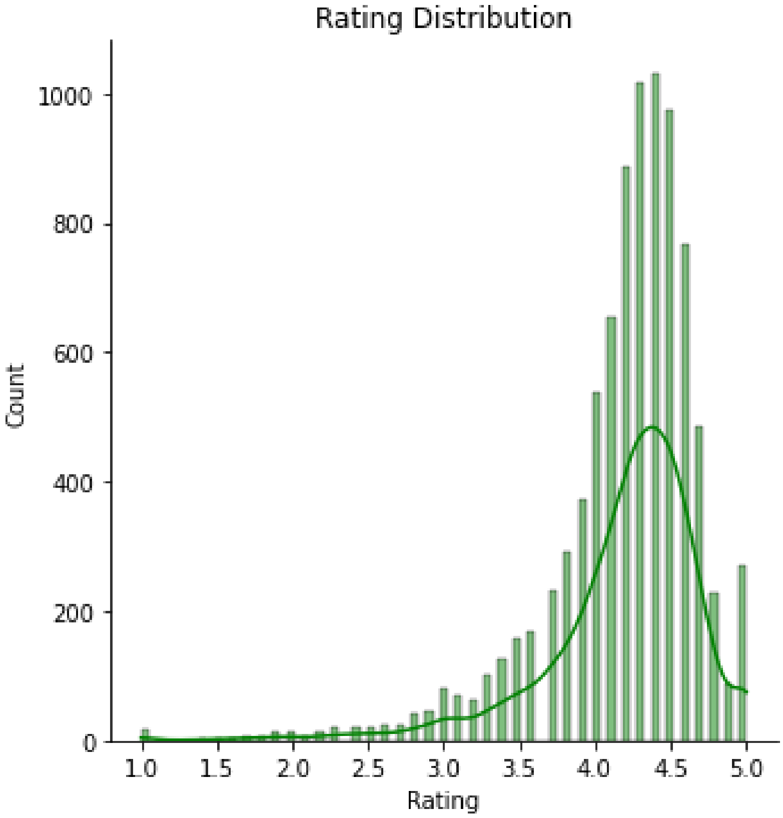
We will not be removing the outliers in this case since chances are, they are legitimate ratings from the users.

2.6 Data Visualization

In this section, we will explore the attributes of the dataset, distributions and correlation between the attributes:

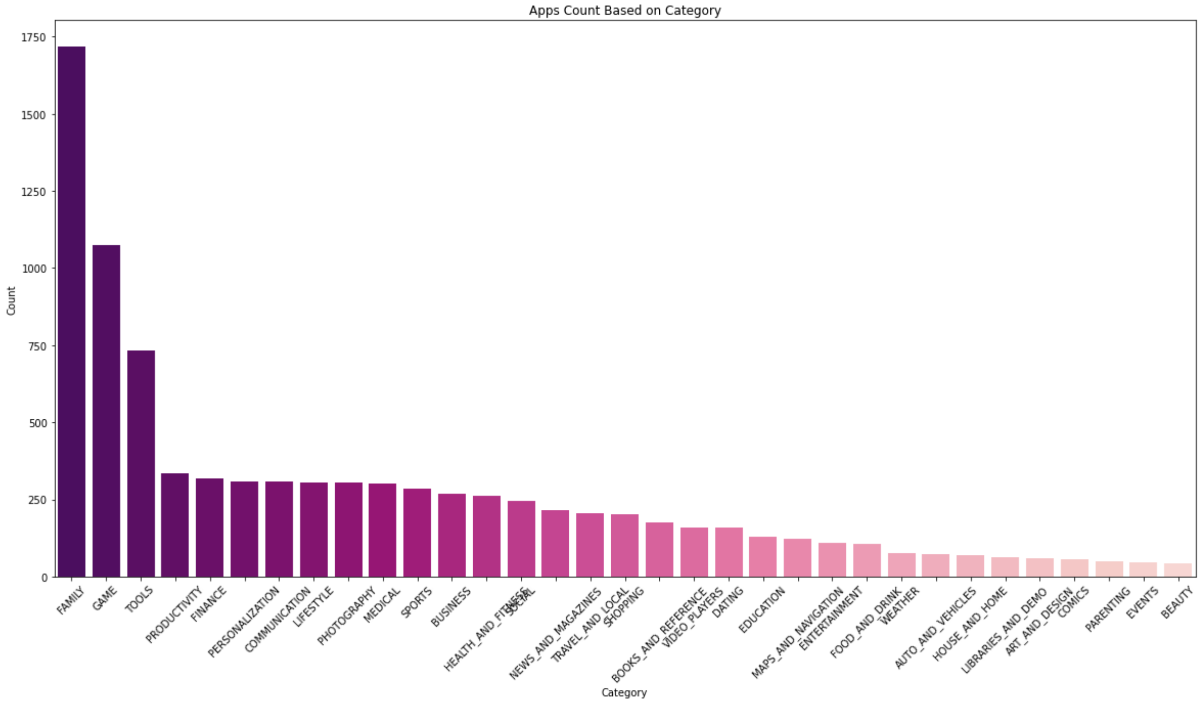
A. Distribution of ratings in the dataset

This figure shows that majority of the apps falls between 4.0 and 4.7 rating, which align with the mean and median of 4.19 and 4.3 respectively from section **2.4 F**



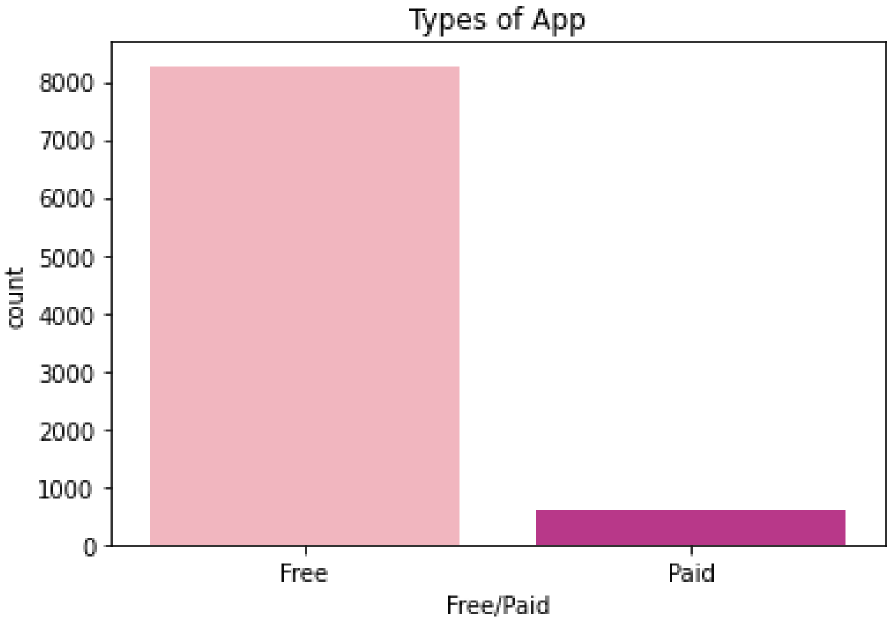
B. Apps distribution based on category

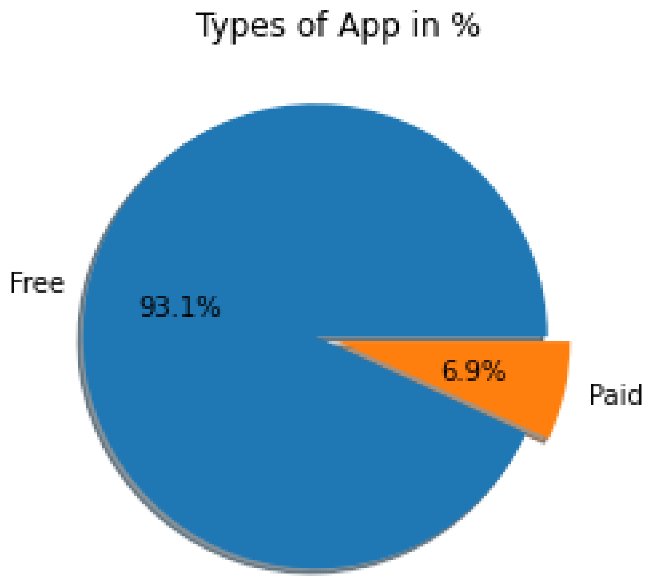
The bar graph shows the distribution of the apps based on category with family, games and tools being the top categories in terms of numbers



C. Count of apps based on type (Free/Paid)

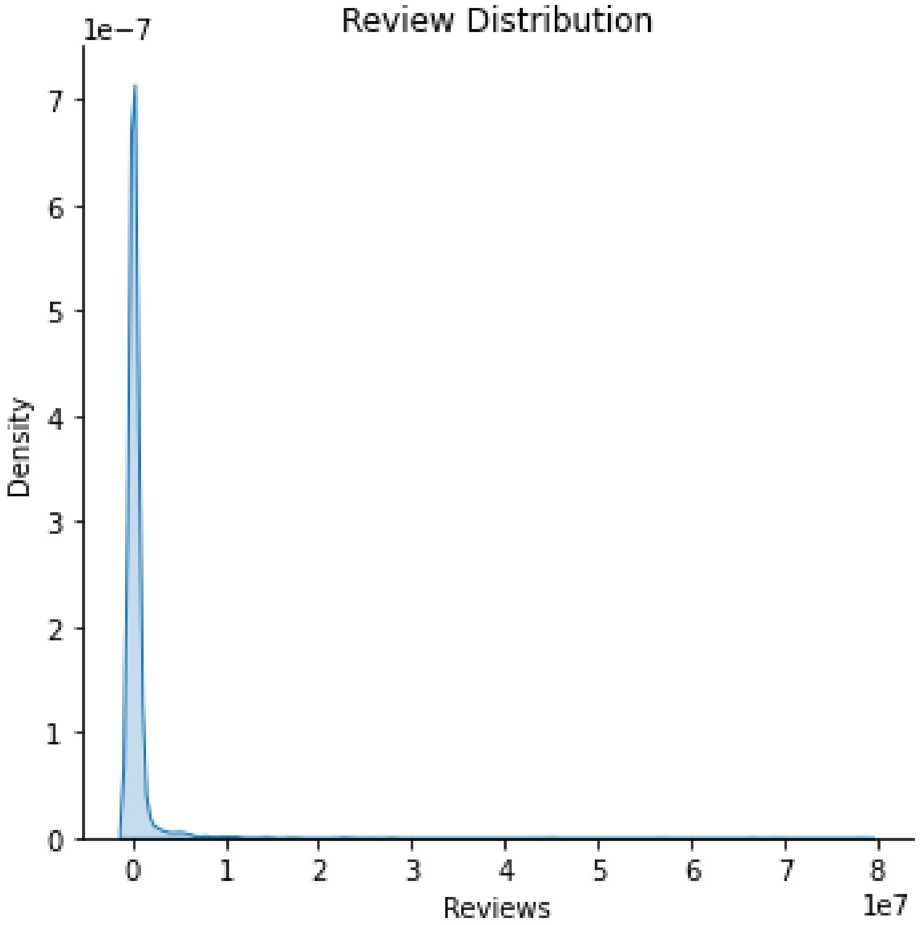
There are 2 types of apps, free and paid apps. In this graph we can see the majority of the apps are free, more than 8,000 or 93.1% apps are free to download compared to paid apps at around 500 or 6.9%.





D. Distribution of Apps based on Review

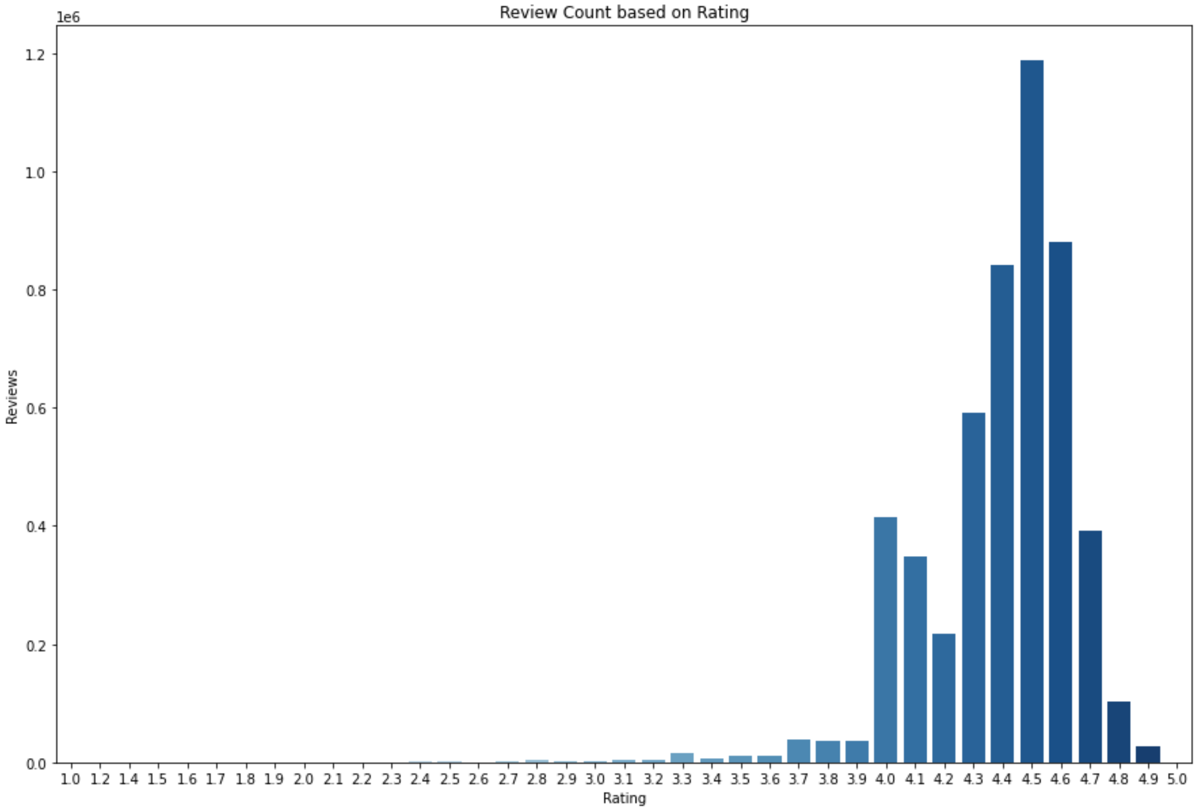
The following graph shows most apps has less than 300,000 reviews, with the overwhelming minority with more than 1 million reviews. This indicates that only a handful of apps make it above the 1M+ review mark.

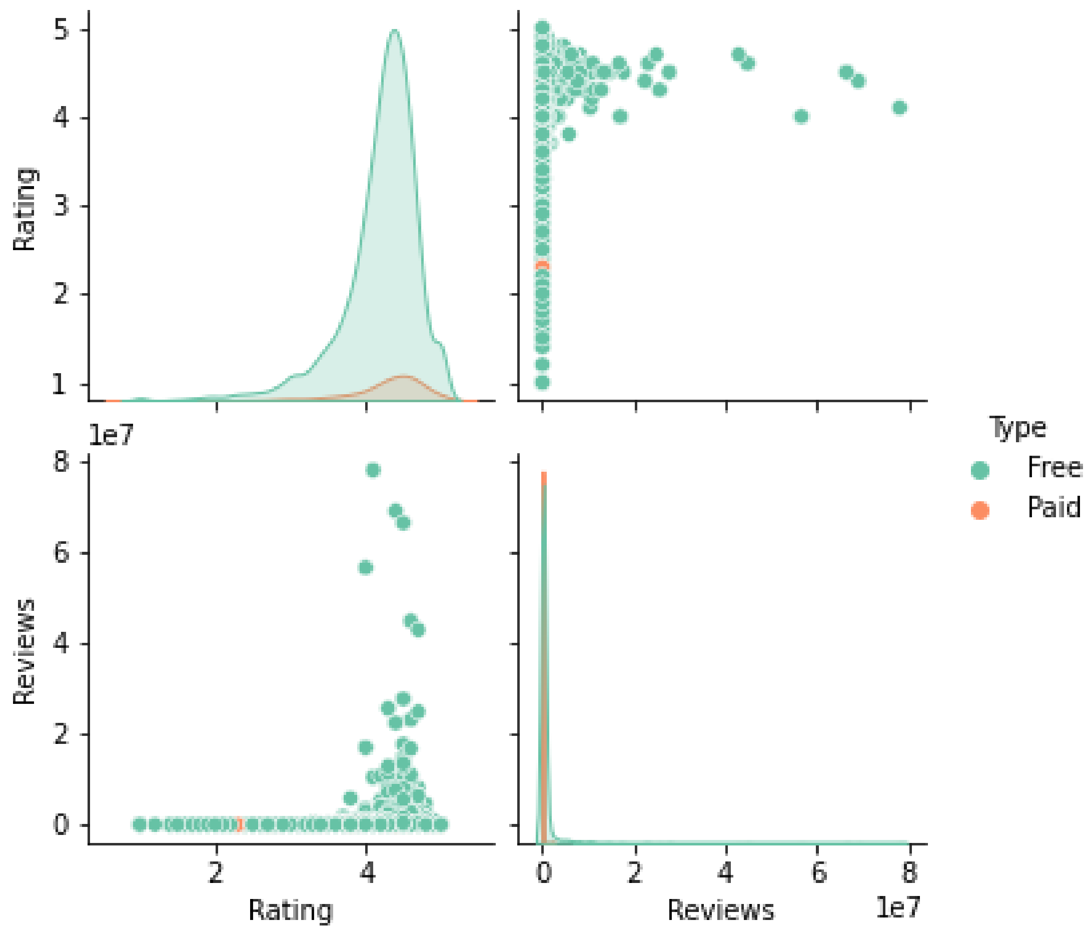


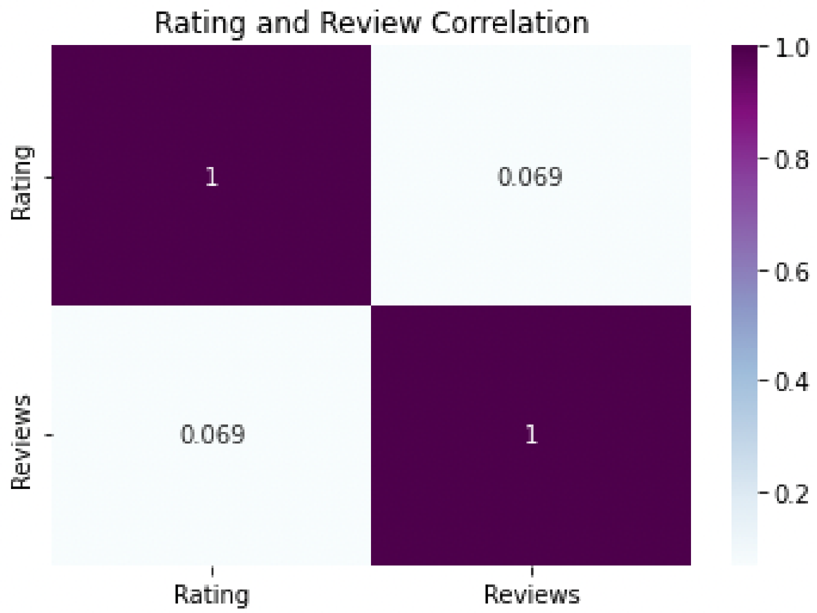
E. Correlation between Rating and Review

The following graphs show apps with higher ratings received significantly more reviews than the apps with rating under 4.0. The initial hypothesis is that the higher the apps rating, the higher the popularity, which translates to higher number of reviews.

But if we look closer, the correlation between reviews and ratings seems to be low at just 0.069, meaning just because the apps are top rated, that don't necessarily mean they will receive significantly more reviews than lower rated apps. But according to **figure 2.6 A**, most of the apps in the dataset falls between 4.0 and 4.7 rating, which might better explain why majority of the reviews are concentrated in apps with 4.0 rating and higher.

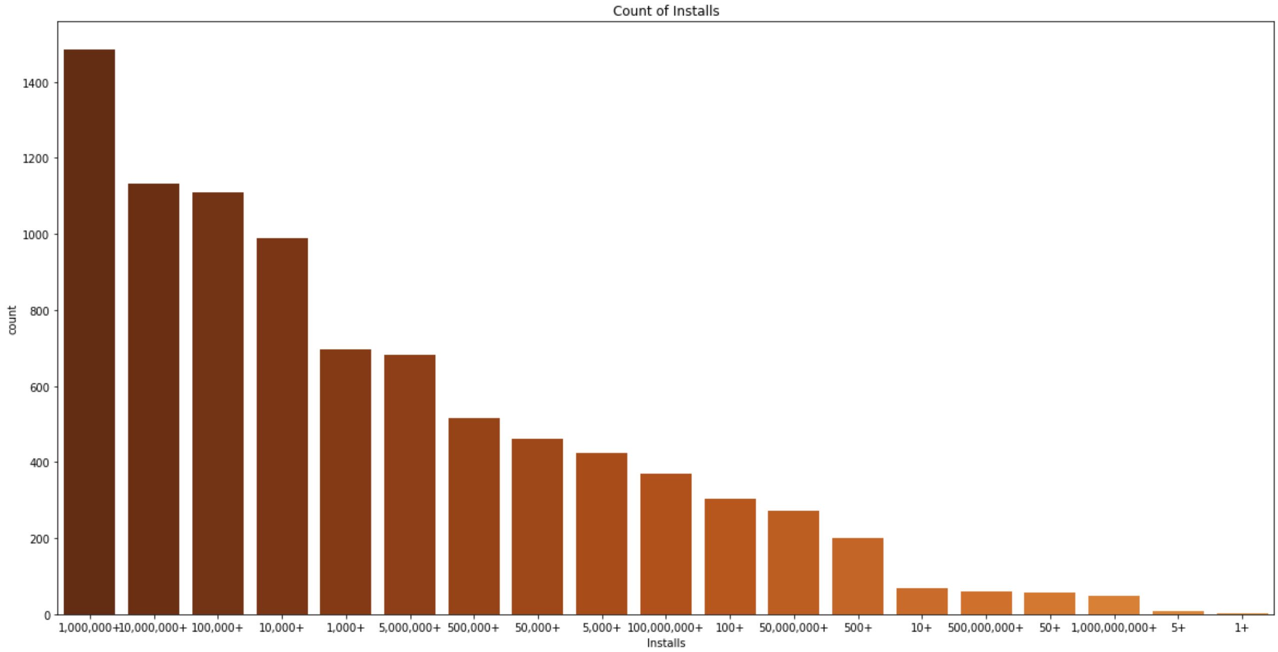



Rating
5
4
3
2
1
Reviews
1e7
8
6
4
2
0
2
4
Rating
0
2
4
6
8
Reviews
1e7
Type
Free
Paid


Rating and Review Correlation
Rating
Reviews
1
0.069
0.069
1
Rating
Reviews
1.0
0.8
0.6
0.4
0.2

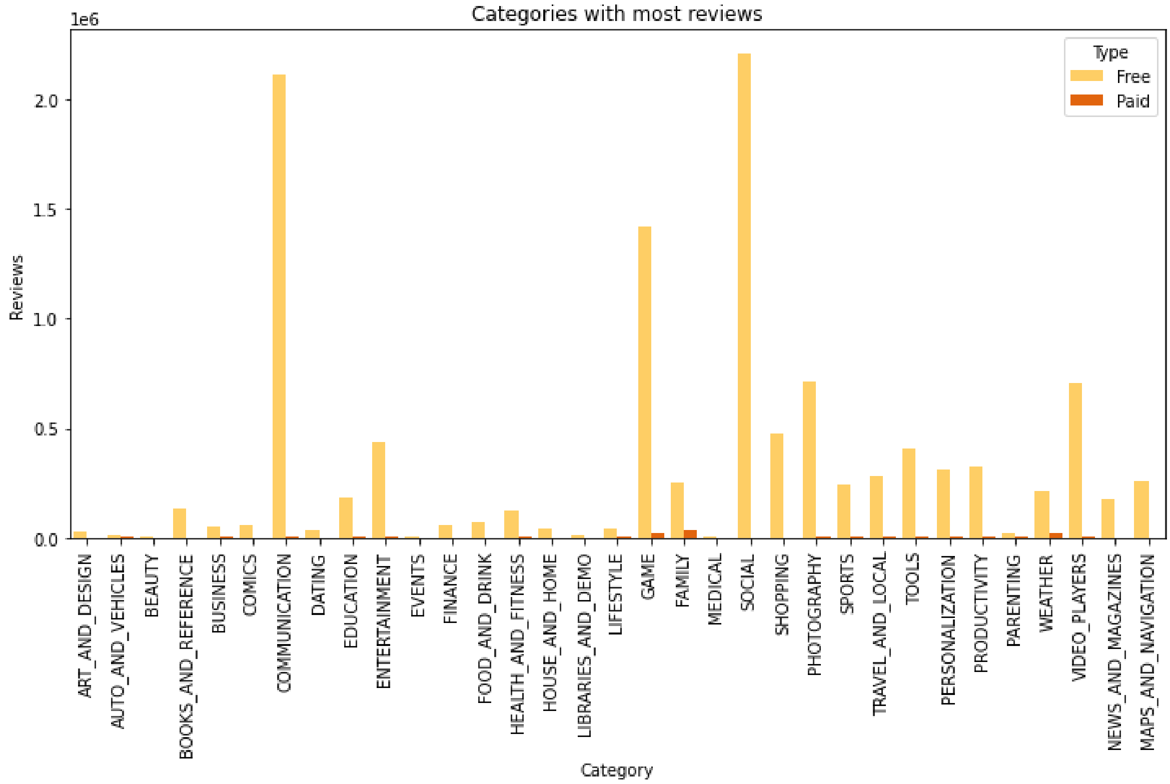
F. Apps Distribution Based on Number of Installation

Most of the apps in the dataset have less than 50 million downloads, the top 3 are apps with 1M+ downloads followed by 10M+ and 100k+



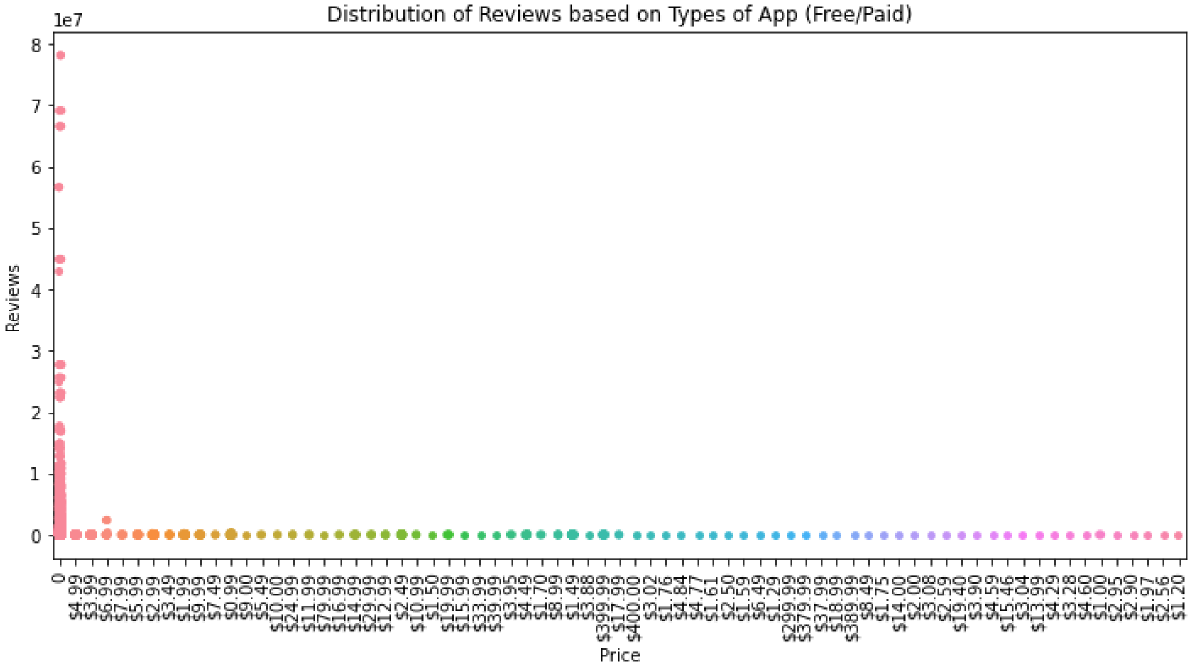
G. Categories with most reviews

Apps that were able to gather the most reviews fall into these 3 categories: social, communication and game



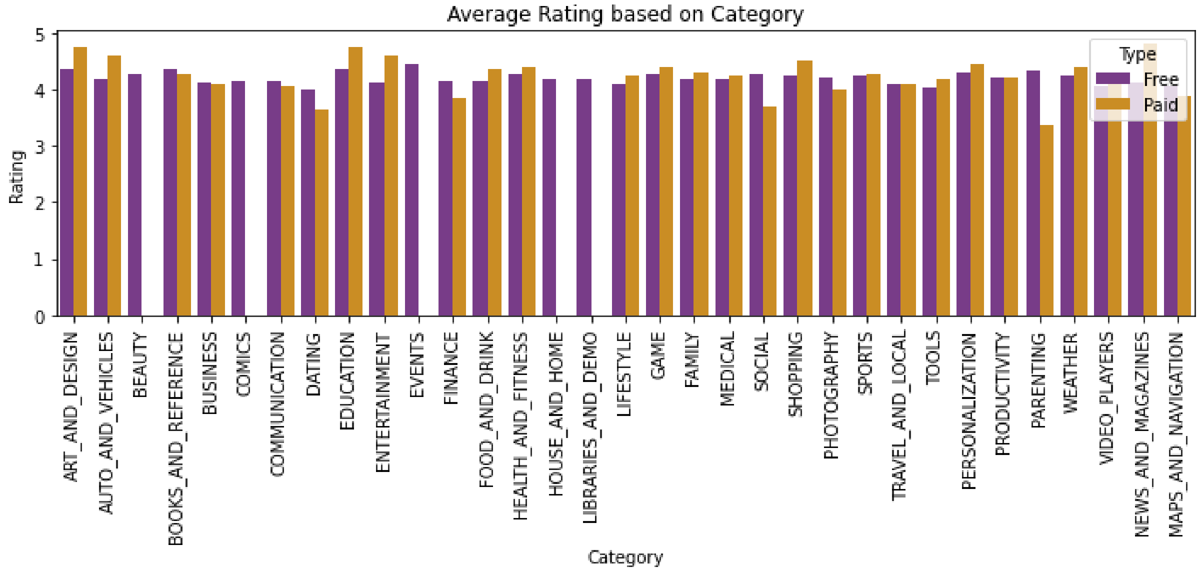
H. Distribution of Reviews Based on Pricing

Apps that are free download received overwhelmingly more reviews compared to their paid counterparts, which makes sense since if we refer to figure **2.6 C,** 93.1% of the apps are free to download.



I. Average rating based on category

We breakdown the average rating for each category, it looks like majority of them have higher than 3.5 rating and paid apps have the tendency to accumulate higher rating compared to free ones.



3 Discussion and Conclusions

The definition of successful apps can be different between developers, some might focus on higher return of investment, some might focus on a more niche market, others might build apps dedicated to improving the state of the world. In this report, we breakdown the characteristics of successful apps (apps with high rating, reviews and installs) based on their category and types whether its free or paid.

Figuring out which type of apps to build can be very difficult especially in this era where competition is stiff, the ever-evolving technological landscape and high user expectation. A lot of time needs to be devoted to figure out what problems our apps are trying to solve, what kind of technology are we going to utilize and are available, which market segments are we targeting and how to reach them, how to scale rapidly and so many other things to consider.

We emphasized more on reviews rather than installs count since anyone could be installing apps and use it only once or twice or not at all. Review can be a more reliable indicator since users would have to download and use the app before providing review. From the above analysis, we can conclude that the top 3 most popular categories are:
-Social
-Communication
-Games

Higher rated apps might play a role in attracting more reviews and installs, but not significantly, but free apps are generally more popular than paid ones, it minimizes the barrier for users to download and use the app. Developers can utilize other ways to monetize their apps, some of the most popular monetization methods in 2021 were video ads, in-app purchases and subscription (Statista, 2022). The goal is to develop useful app and get as many users as they can, the revenue can be generated later.

Another interesting finding is that paid apps generally receive higher rating compared to free ones, the reason could be free apps can range from low to high quality but paid ones, in order to get users agree to shell out money, they have to have at least certain level of quality.

Overall, depending on the main objectives of the developers, if they want to maximize the popularity of their apps, they might want to focus on the top 3 categories and removing the paywall. Once they have enough users, they can start monetizing their apps using methods such as ads, subscriptions or in-app purchases.

4 References

[1] Lemmens, R.,Antoniou, V., Hummer, P., & Potsiou, C. (2021). Citizen science in the digital world of apps. In: ,et al. The Science of Citizen Science. Springer, Cham. https://doi.org/10.1007/978-3-030-58278-4_23

[2] Statista. (2022). Number of available applications in the Google Play Store from December 2009 to March 2022. https://www.statista.com/statistics/266210/number-of-available-applications-in-the-google-play-store/

[3] Statista. (2022). Leading Android apps in the Google Play Store worldwide in September 2022, by revenue. https://www.statista.com/statistics/271674/top-apps-in-google-play-by-revenue/

[4] Statista. (2021). Distribution of worldwide mobile app revenues in the Apple App Store from 2019 to 2025, by category. https://www.statista.com/statistics/1010701/apple-app-store-revenue-share-by-category-worldwide/

[5] Statista. (2021). Distribution of worldwide mobile app revenues in the Google Play store from 2019 to 2025, by category. https://www.statista.com/statistics/1010710/google-play-app-revenue-share-by-category-worldwide/

[6] Statista. (2022). Number of monthly active Facebook users worldwide as of 2nd quarter 2022. https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/

[7] Statista. (2022). Most used monetization methods for mobile gaming and non-gaming apps according to mobile publishers worldwide in 2021. https://www.statista.com/statistics/384215/app-developer-monetization-mix/