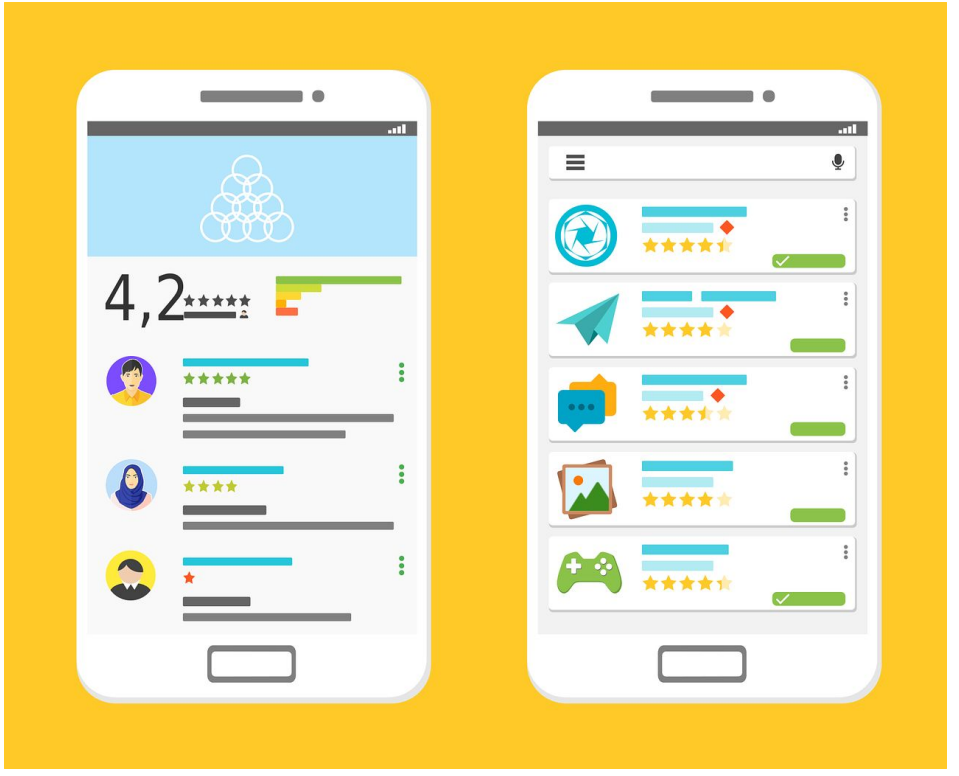
CAPSTONE PROJECT-1

Exploratory Data Analysis on Google Play Store Apps

RADHIKA R MENON

CONTENT

1. **Problem Statement**
2. **Data description**
   a. Play store data
   b. User Reviews data
3. **Data preparation**
4. **Exploratory Data Analysis**
   - Distribution of ratings and installs
   - Type-wise Analysis
   - Content-rating Analysis
   - Price Analysis
   - Analysis of effect of last update year
   - Category-wise Analysis
   - User-review analysis
5. **Conclusion**
6. **Suggestions**



PROBLEM STATEMENT

- Google Play Store is the biggest digital distribution platform for Android apps worldwide, offering users close to 3.3 million mobile apps to choose from.
- Historical data about the Apps has enormous potential to drive app-making businesses to success. Actionable insights can be drawn for developers to work on and capture the Android market.
- *Play Store apps data* -details about the apps,*User reviews data*- customer reviews for the apps.
- Objective : Explore and analyze the data to discover key factors responsible for app engagement and success.

ATTRIBUTE INFORMATION

PLAY STORE DATA - PS_df

➢ **10841 entries,13 columns**

	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

- **App: App name**
- **Category: App category**
- **Rating: Average user rating of the app**
- **Reviews: Number of reviews**
- **Size: App size**
- **Installs: The number of installations of the app**
- **Type: Whether it is free or paid**
- **Price: App price**
- **Content rating: The targeted audience of the app**
- **Genres: The genre of content offered by the app**
- **Last updated: Date of last update**
- **Current Ver: The current version of the app**
- **Android Ver: The Android version(s) supported by the app**

ATTRIBUTE INFORMATION

USER REVIEW DATA- UR_df

➢ **64295 reviews,5 columns**

	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

- **App: Name of the app.**
- **Translated_Review: English translated version of user review**
- **Sentiment: The nature of the review,whether it is positive , negative or neutral**
- **Sentiment_Polarity: numerical value given to the sentiment of the user by analyzing the translated review.Its value ranges from[-1,1].**
- **Sentiment_Subjectivity: quantifies the amount of personal opinion and factual information contained in the translated reviews.Its value ranges from [0,1].**

DATA PREPARATION

Play store data(PS_df) & User Reviews data (UR_df)

- **Duplicates removal: In case of PS_df,removed duplicates in the 'App' column.**
- **Correcting data types and formats :Removed characters like '$','+','M','k' from 'Price','Installs' and 'Size' columns and corrected their data types**
- **Handling missing values:Imputed the missing values in Rating with the median value and those in 'Android Ver' with mode,appropriate value through manual search.In UR_df, rows with NaN values were dropped.**

```
dataframe_info(PS_df)
```

	Column_name	NaN_count	% of NaN	dtype	Unique_count
0	App	0	0.00	object	9660
1	Category	0	0.00	object	34
2	Rating	1474	13.60	float64	40
3	Reviews	0	0.00	object	6002
4	Size	0	0.00	object	462
5	Installs	0	0.00	object	22
6	Type	1	0.01	object	3
7	Price	0	0.00	object	93
8	Content Rating	1	0.01	object	6
9	Genres	0	0.00	object	120
10	Last Updated	0	0.00	object	1378
11	Current Ver	8	0.07	object	2832
12	Android Ver	3	0.03	object	33

```
dataframe_info(UR_df)
```

	Column_name	NaN_count	% of NaN	dtype	Unique_count
0	App	0	0.00	object	1074
1	Translated_Review	26868	41.79	object	27994
2	Sentiment	26863	41.78	object	3
3	Sentiment_Polarity	26863	41.78	float64	5410
4	Sentiment_Subjectivity	26863	41.78	float64	4474

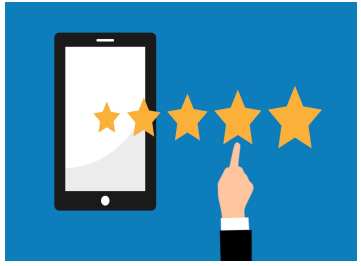
DATA PREPARATION (ctd..)

- Sanity checks:Removed few apps which were rated but have no installs.
- Outlier treatment:The extremely high priced apps where found to junk,so we eliminated them.
- Based on the main genre, apps placed in the "FAMILY" category were moved to more suitable categories.
- Created new columns such as Last_update_year,Revenue and Size_group for more meaningful insights.

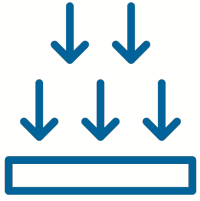
Outliers in 'Price' column

App	Price	Installs
BP Fitness Lead Scanner	109.99	1
I AM RICH PRO PLUS	399.99	1000
I Am Rich	389.99	10000
I Am Rich Premium	399.99	50000
I Am Rich Pro	399.99	5000
I am Rich	399.99	5000
I am Rich Plus	399.99	10000
I am Rich!	399.99	1000
I am extremely Rich	379.99	1000
I am rich	399.99	100000
I am rich (Most expensive app)	399.99	1000
I am rich VIP	299.99	10000
I am rich(premium)	399.99	5000
I'm Rich - Trump Edition	400.00	10000
most expensive app (H)	399.99	100

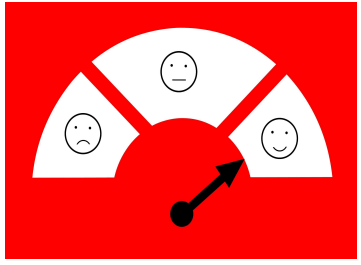
KEY PERFORMANCE INDICATORS



App Ratings



Installs

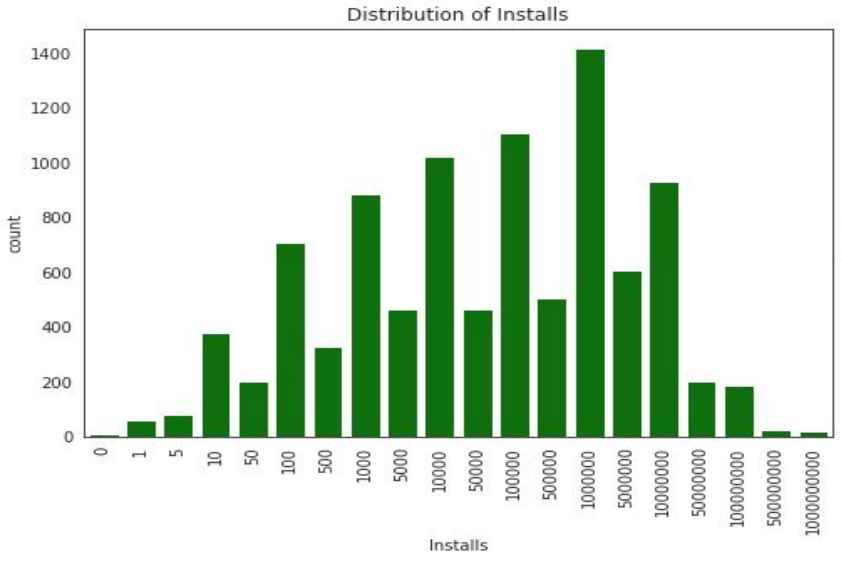


User Reviews

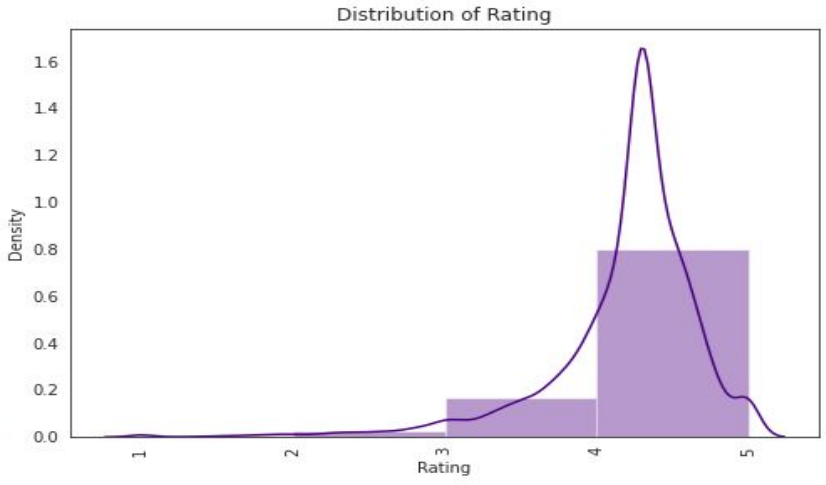
Revenue

EXPLORATORY DATA ANALYSIS

DISTRIBUTION OF INSTALLS



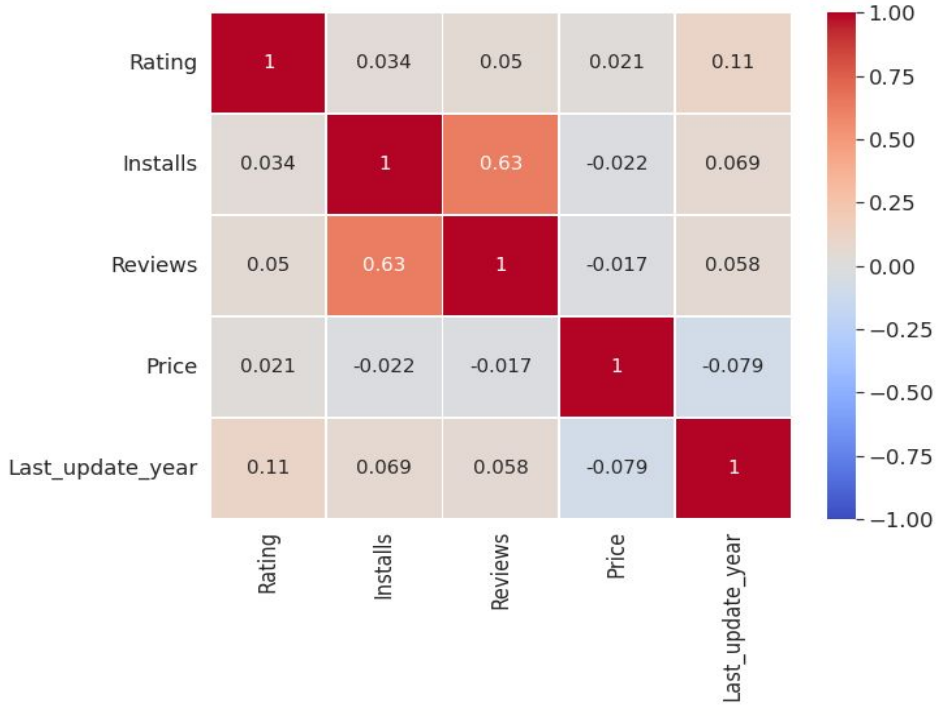
APP RATING DISTRIBUTION



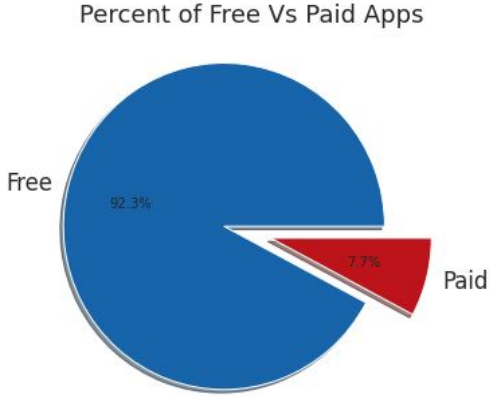
- The majority of apps have less than or equal to one million installations
- Most of the apps have app rating in the range [3.8 - 4.6] .

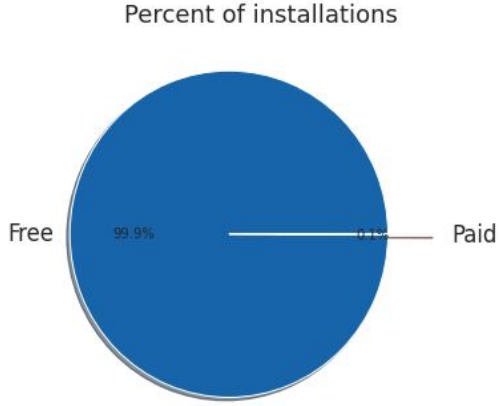
CORRELATION HEATMAP -PS_DF

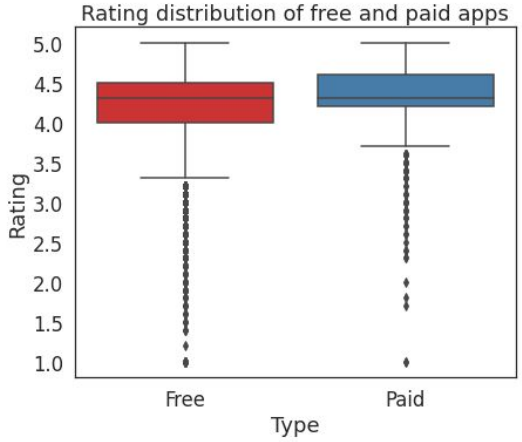
- App rating is positively correlated with year of last update.
- Slight positive correlation between Rating and Installs.
- Strong positive correlation between Reviews and Installs
- Last update year can also be seen to be positively correlated with the installs and number of reviews.

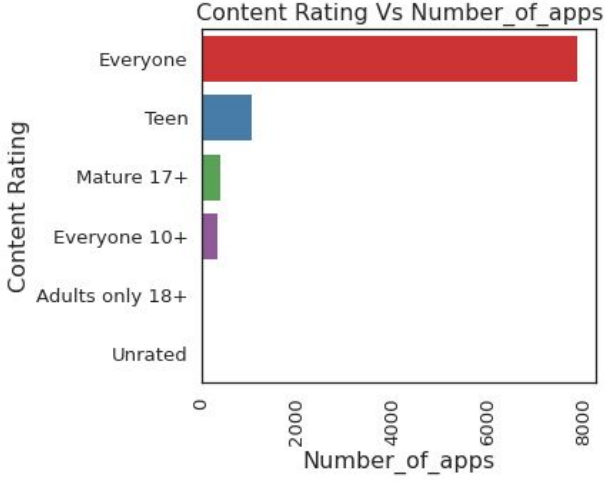


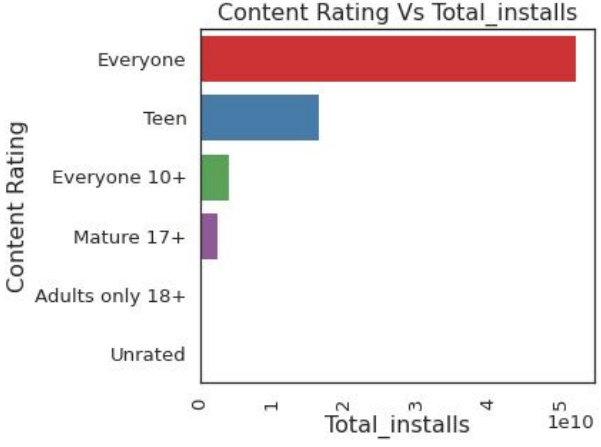
TYPE & CONTENT RATING

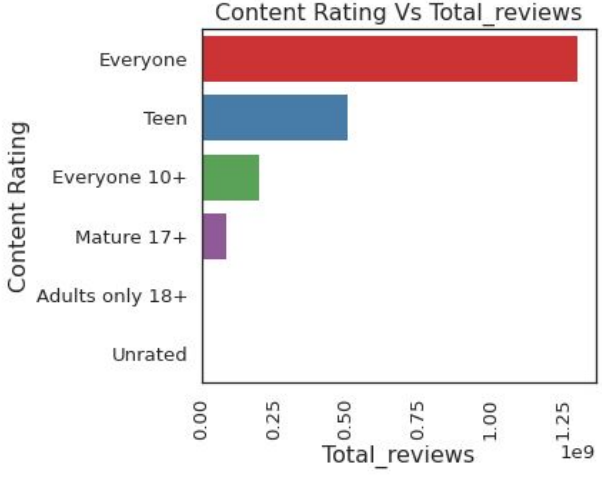




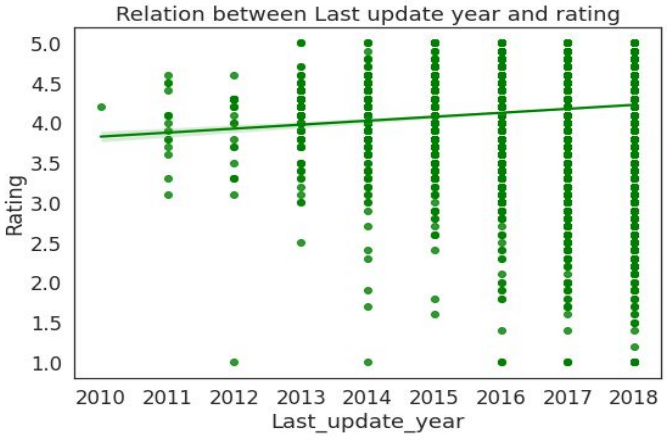


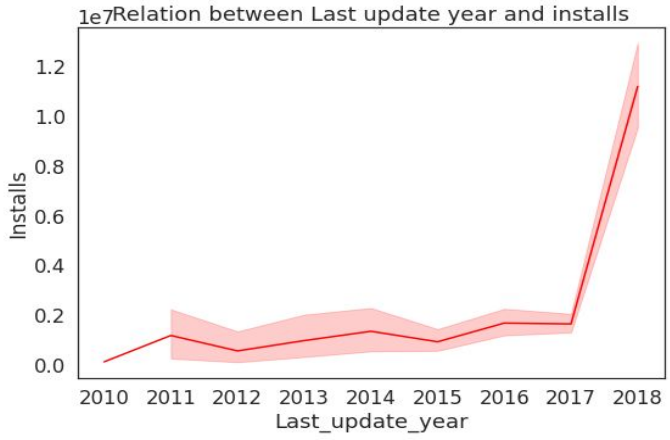




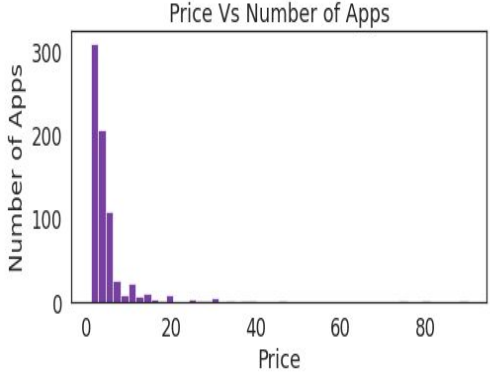


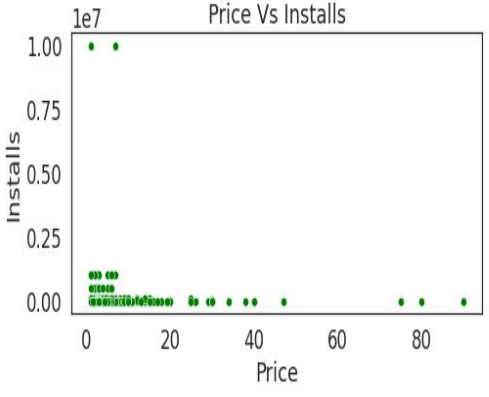
LAST UPDATE YEAR & PRICE





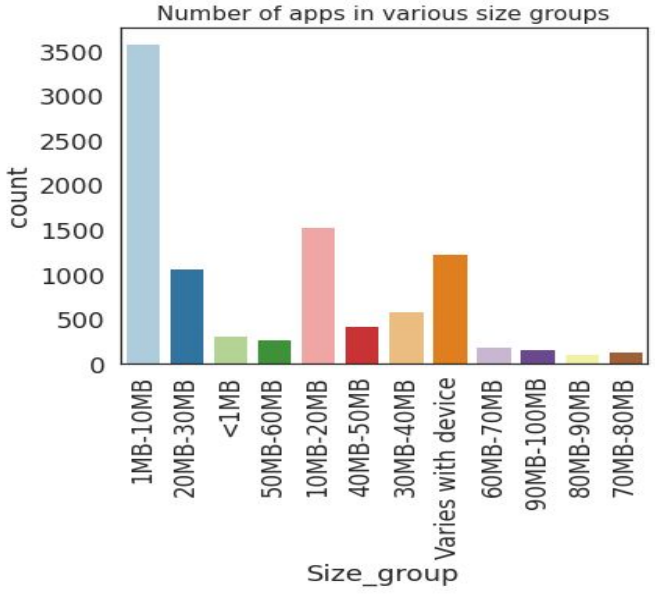
From these plots,we can conclude that Apps which has been updated recently have higher chances of receiving better ratings and higher number of installations.

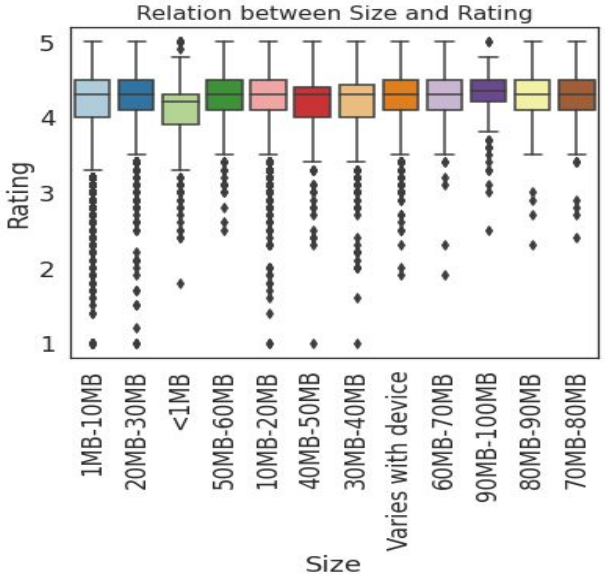


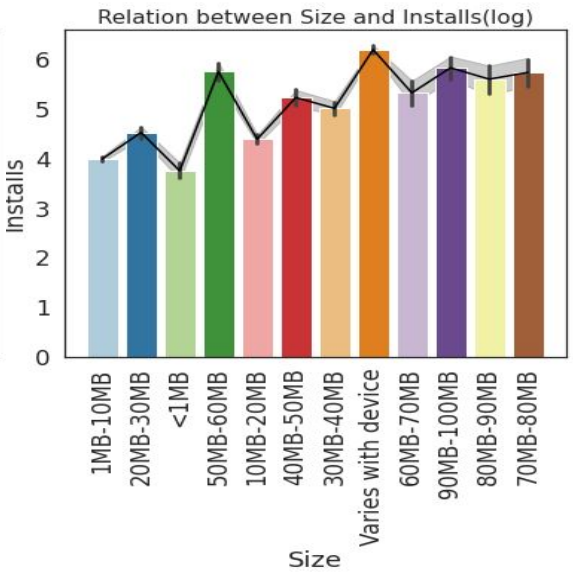


- Majority of apps are priced under $20.
- Both Rating and Number of downloads are inversely proportional to Price.

APP SIZE



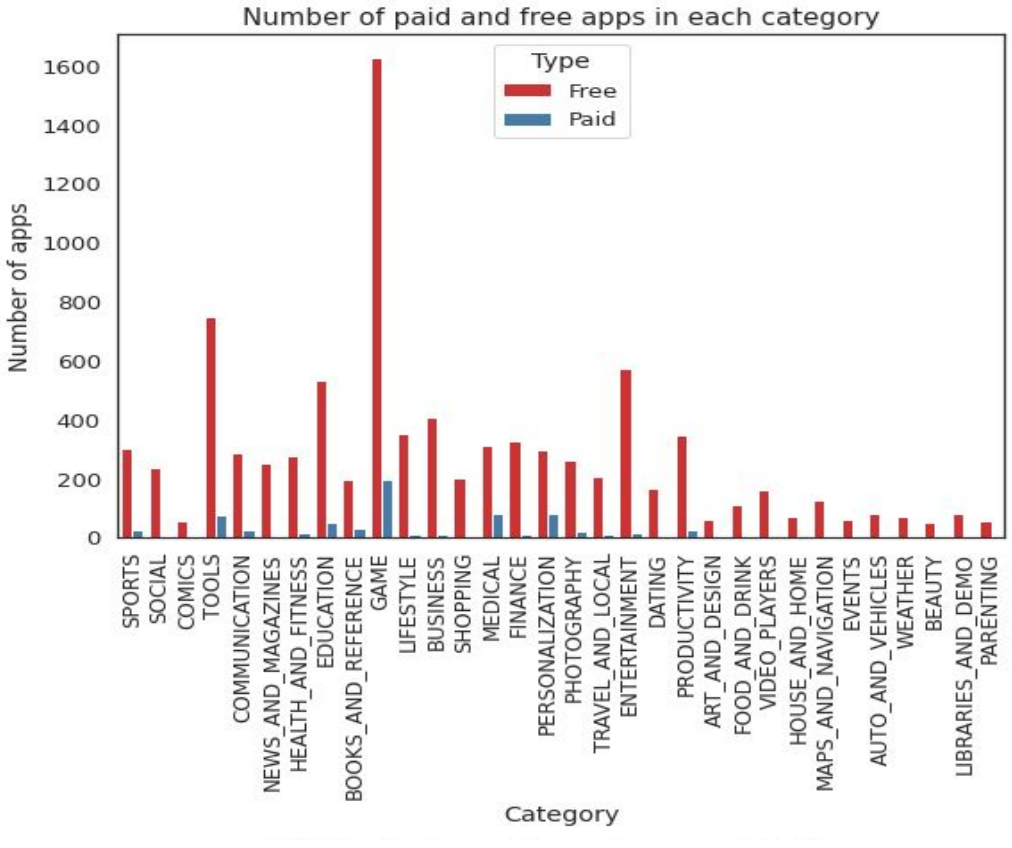




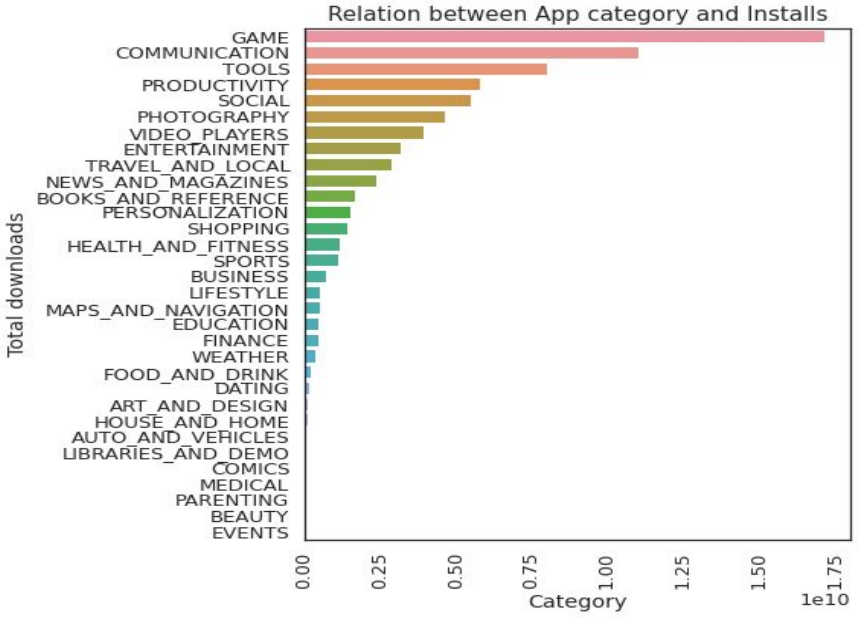
- Lighter apps of size <20 MB dominate the app market.
- Most of the apps in all size ranges have managed to receive ratings above 4 .
- As size of app increases,it is less likely to receive poor ratings(below 3).
- Apps with size that varies with device are downloaded the most.

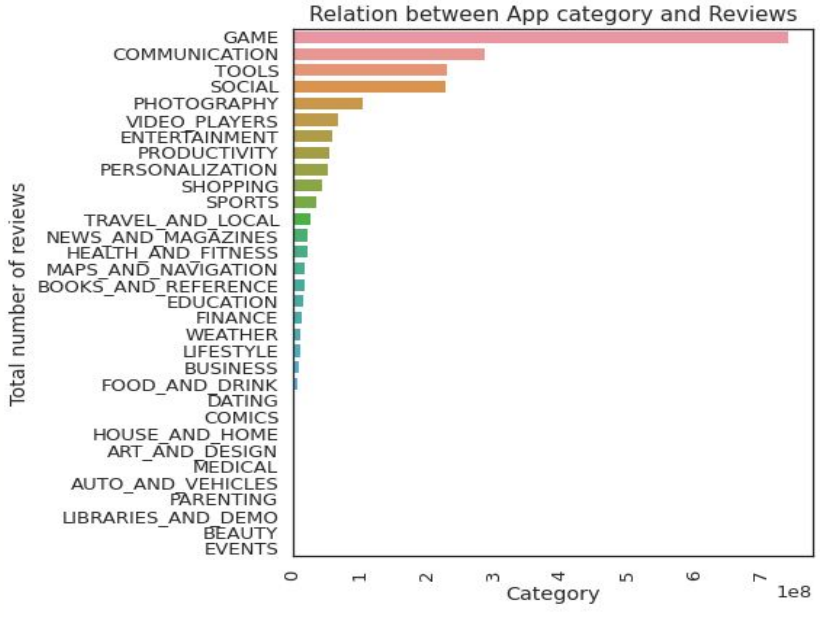
CATEGORY

- Top 5 categories in terms of market prevalence for free apps
  1. Games (~1600 apps)
  2. Tools (~750 apps)
  3. Entertainment (~600 apps)
  4. Education
  5. Business.
- Top 5 categories in terms of market prevalence for paid apps
  1. Games (~200 apps)
  2. Tools (~80 apps)
  3. Medical (~80 apps)
  4. Personalization
  5. Productivity &Communication
- The categories like Comics,Beauty,Art and Design,Parenting and Events have least number of apps.



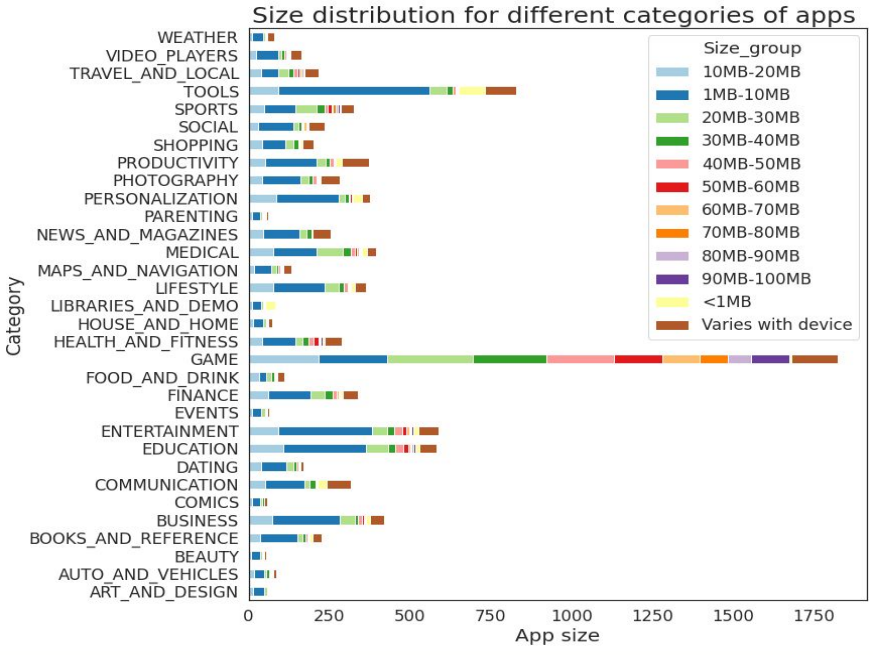
CATEGORY(ctd..)

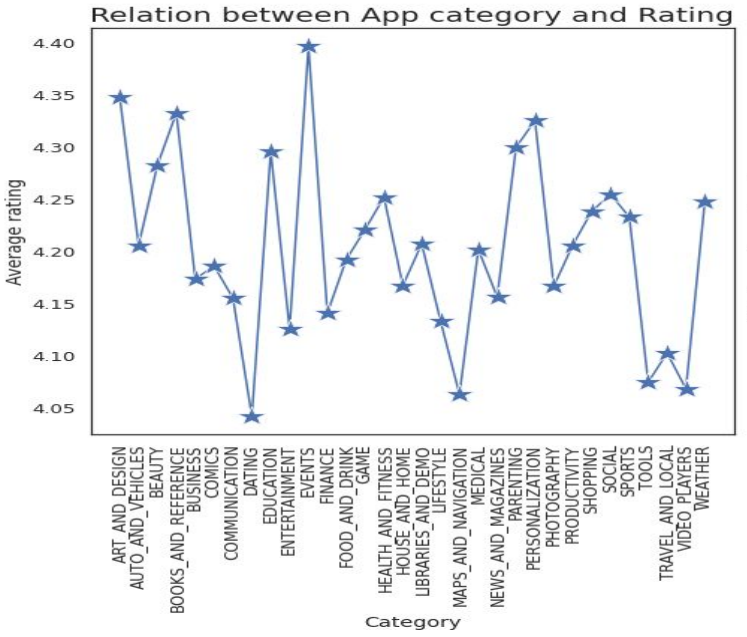




- Games has the most number of downloads,followed by Communication and Tools.
- Category with most number of reviews are Games,Communication and Tools

CATEGORY(ctd.)

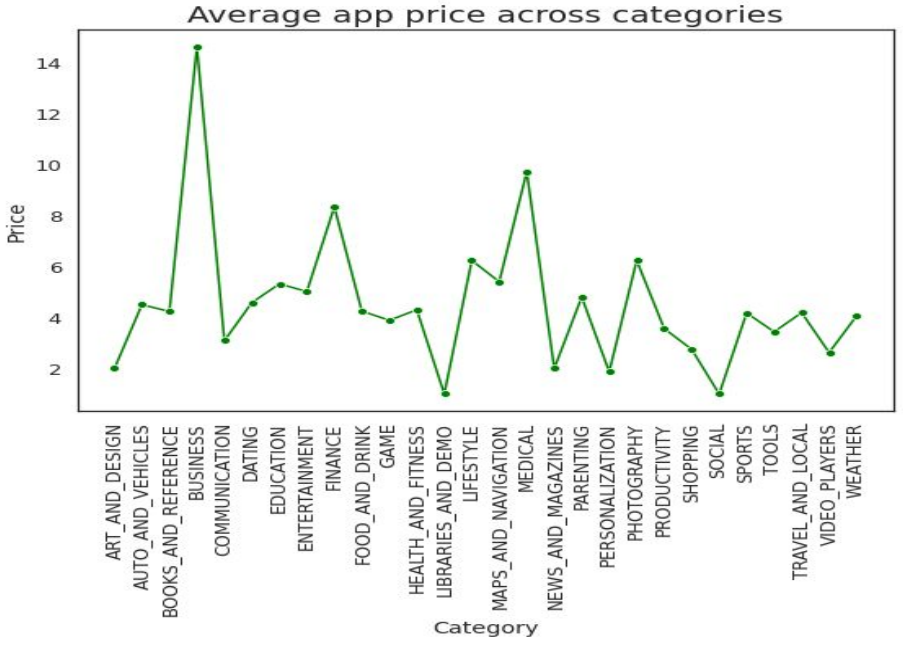




- Games has a wider distribution of sizes.
- There are least number of apps with a size of less than 1 MB.

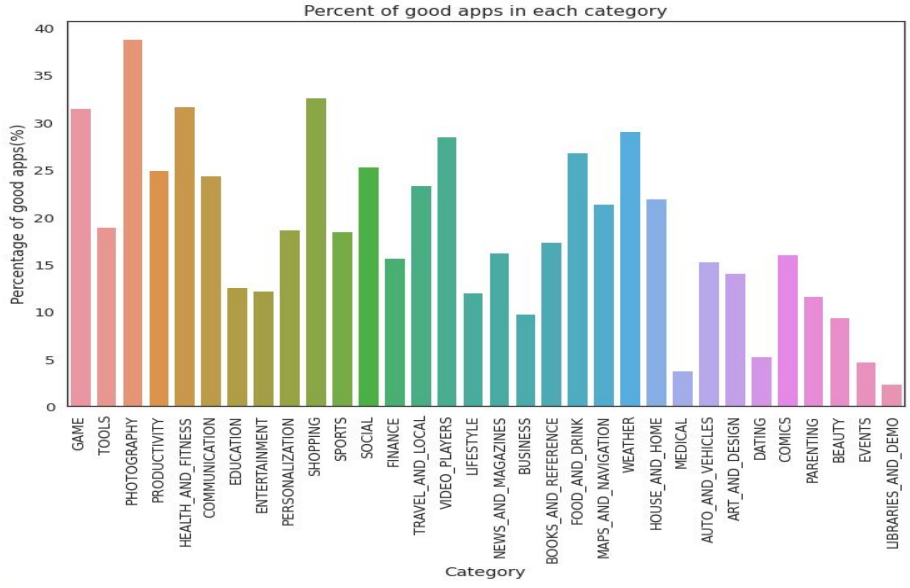
- 'Events' category has the highest ratings

AVERAGE PRICE



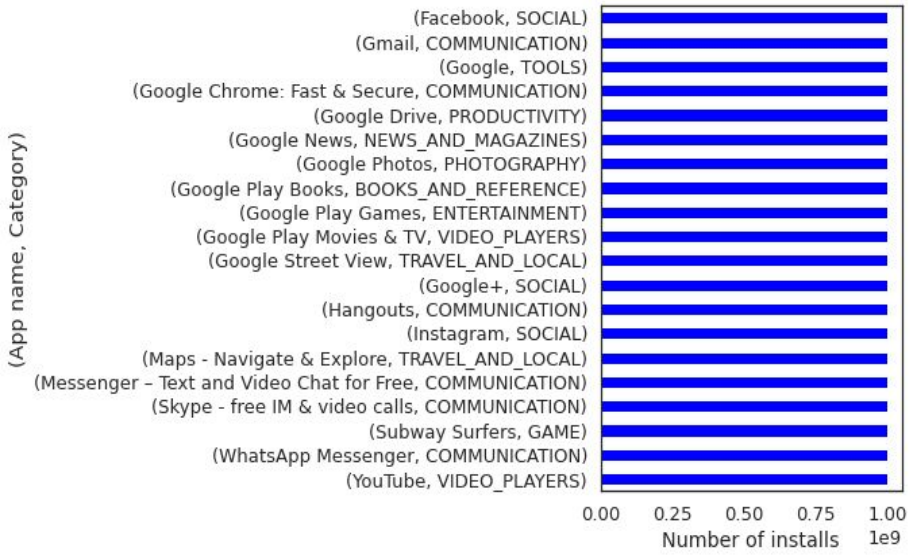
- Highest average price -:Business
- Least average price : Libraries and demo & Social

BEST PERFORMING CATEGORY



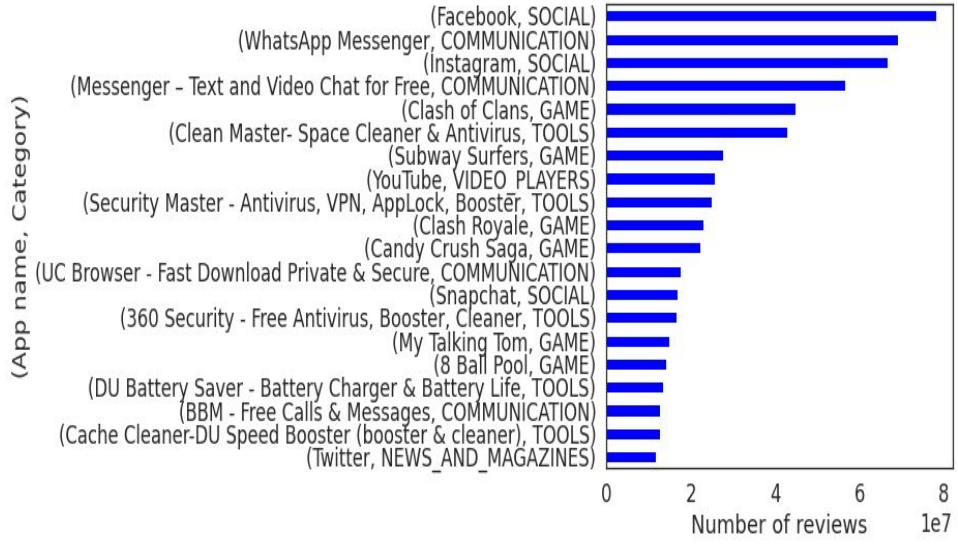
- Percentage of apps with Rating>= 4.3 & Installs >= 1Million
- Photography-highest percent of good quality apps

APPS WITH 1 BILLION DOWNLOADS



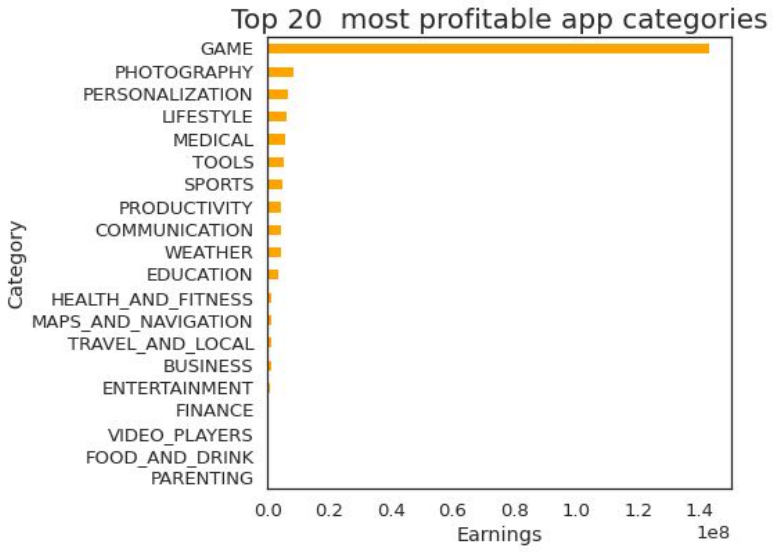
- There are 20 apps in Play store that have crossed 1B downloads.
- Communication(6 apps),Social (3 apps)

MOST REVIEWED APPS

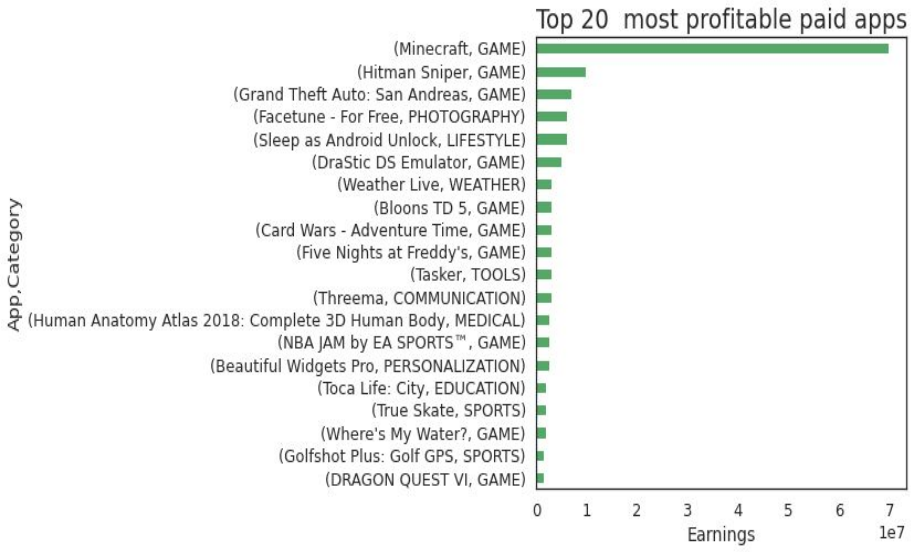


- Number of Reviews can be considered as measure of User Engagement.
- Most reviews : Facebook (80M).
- Games(5 apps),Tools(5 apps), Communication(4 apps), Social(3 apps)

MOST PROFITABLE CATEGORY

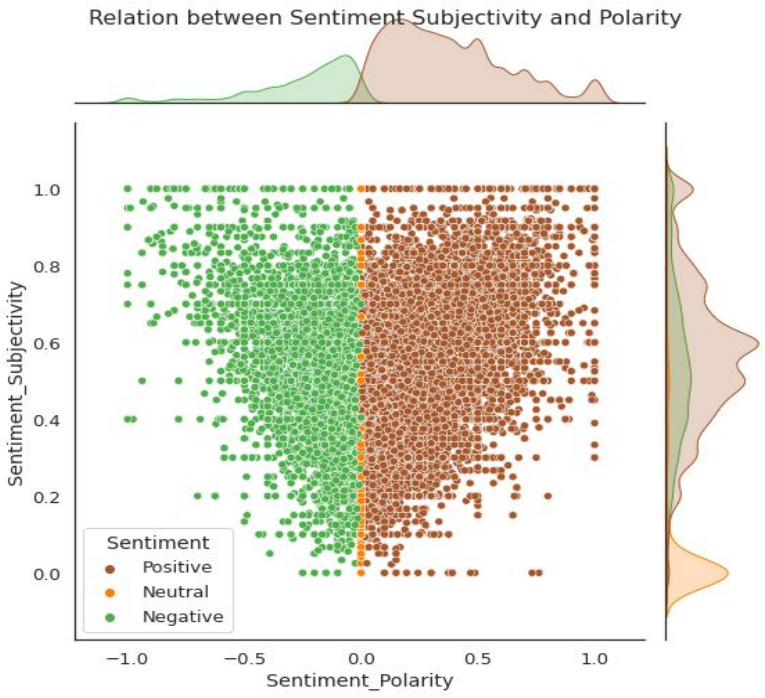


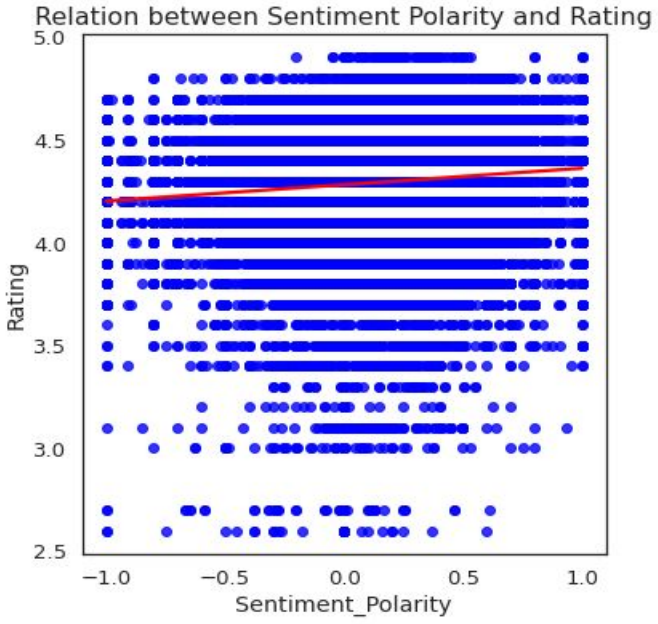
PAID APPS WITH HIGHEST EARNINGS



- Earnings = Price × Installs
- Games is the most profitable category
- Top earning app : Minecraft($ 70 M)

USER REVIEWS

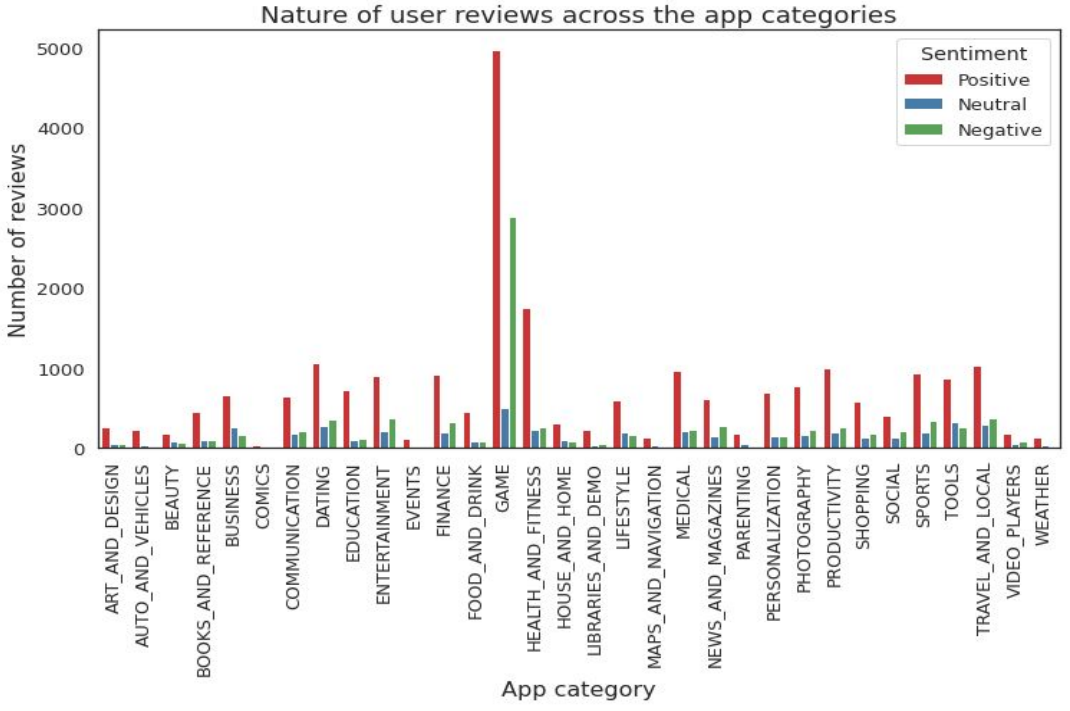


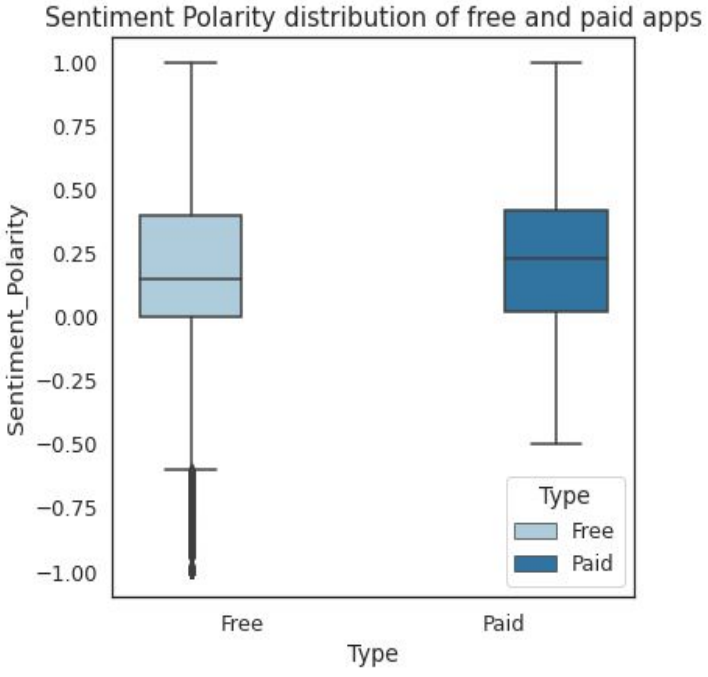


- Higher number of positive reviews
- Most user reviews have a sentiment polarity of [-0.5,0,5].
- The majority of positive and negative reviews are personal (subjectivity scores between [0.4,0.8]), whereas neutral user reviews are factual.

- Higher rated apps,more positive reviews

USER REVIEWS

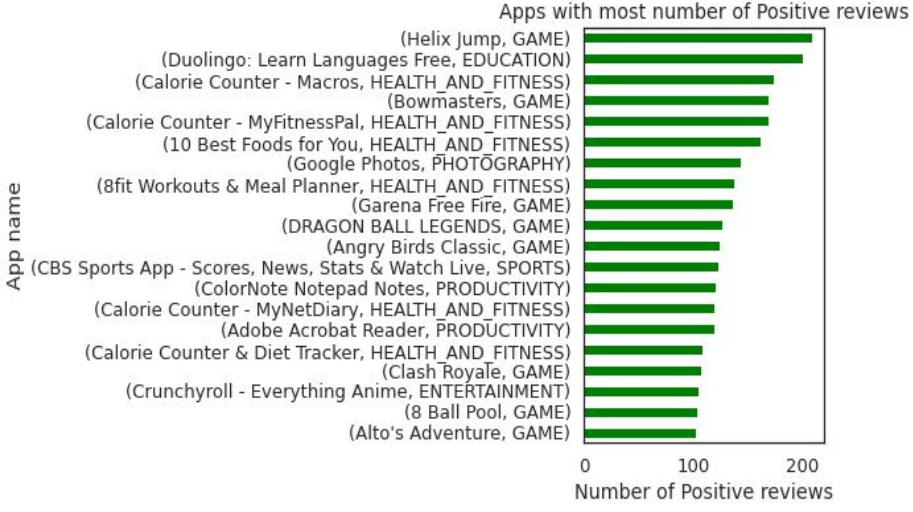




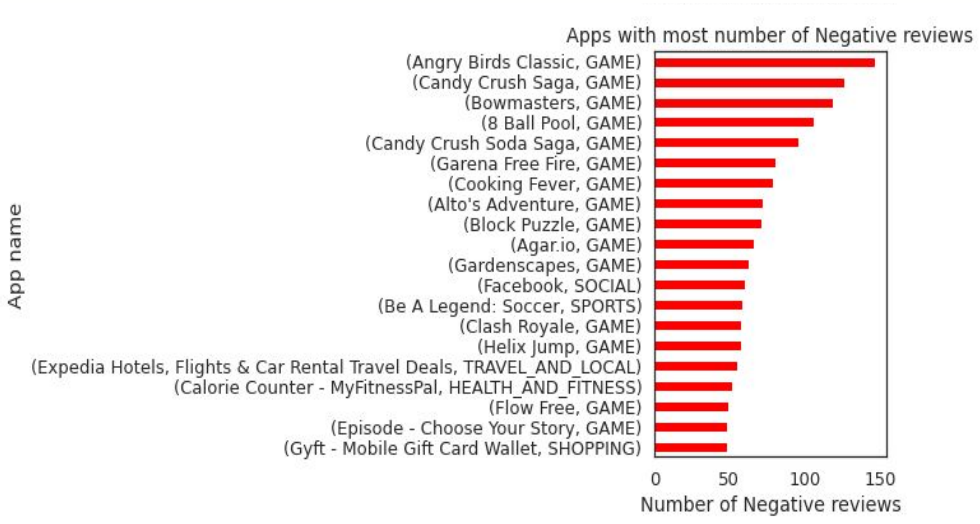
- Apps in the Health & Fitness category do very well since they have higher positive-to-negative reviews ratio.

- Paid app reviews are less harsh than free app reviews.

MOST POSITIVE REVIEWS



MOST NEGATIVE REVIEWS



- Highest number of positive reviews-Helix Jump
- Highest number of negative reviews-Angry Birds Classic

WORD CLOUD

Word cloud of all positive reviews

Word cloud of all negative reviews

- Most of the positive reviews for all apps have been related to food ,health and fitness.
- ‘Time’, ‘unable’ ,’support’, and ‘response’ are frequently used words in the negative reviews for all apps.

CONCLUSIONS

- Free apps dominate the App market(92.2%) and have most installations(99.9%).Paid apps have better ratings and reviews.
- Apps with no age restrictions(Everyone) dominate the App market and have most installations.
- Recently updated apps have higher chances of better rating.
- Although most of apps in Play store are sized under 20 MB,apps with size that Varies with device are downloaded the most.
- Category with the highest number of apps : Game
- Category with the highest number of downloads: Game
- Category with most number of apps that have crossed 1 billion downloads: Communication
- Best performing category:Photography
- Category with highest Positive-to-Negative reviews ratio: Health and Fitness
- Most profitable paid app:Minecraft
- App with the highest user engagement (number of reviews) : Facebook
- App with the highest number of positive reviews : Helix Jump
- App with the highest number of negative reviews: Angry Birds Classic

SUGGESTIONS

- It is preferable for a company attempting to build a brand image to start by offering free apps.
- App price should be decided based on Category. Business,Medical and Finance apps have relatively higher priced apps.Majority of apps in other categories such as Games,Tools,Events etc are priced under $20.
- Categories like Shopping,Beauty are relatively less explored,but have huge potential.
- Apps that are updated often tend to receive better ratings and downloads.
- Paying attention to users' feedbacks- indicate brand's commitment and reliability.

Thank You