Capstone Project : 1

Play Store App Review Analysis

Team Members

i) Bhargav Krishna

ii) Saurabh Yadav

iii) Ajaykumar Gujja

Content

Problem Statement:

1. The Objective of the project to explore and analyze the data to discover key factor responsible for app engagement and success.

Data Summary:

There are two dataset: 1) Play Store Data 2) User Review

1. Play Store Data: The dataset contains 13 features with more than 10k observations.

Important Features :

- App – App name
- Category – Which category the app belongs
- Rating – Rating given to an App from 1.0 to 5.0
- Reviews – Number of reviews given to an App
- Size – Size of an App
- Installs – Number of users installed the particular app
- Type – Is an app free or paid type
- Price – Price of a given app in $
- Content Rating- An app belongs which age group

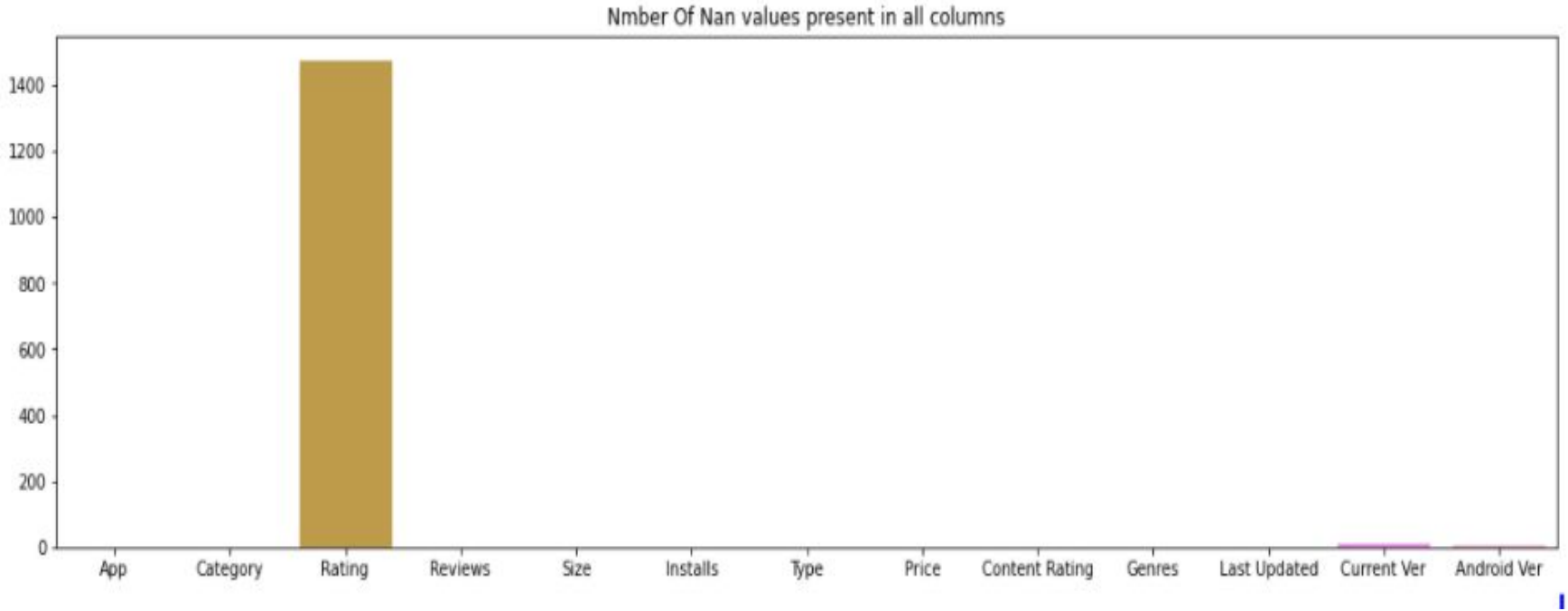
Contd…

2. User Review Data: The dataset contains 5 features with more than 60k observations.

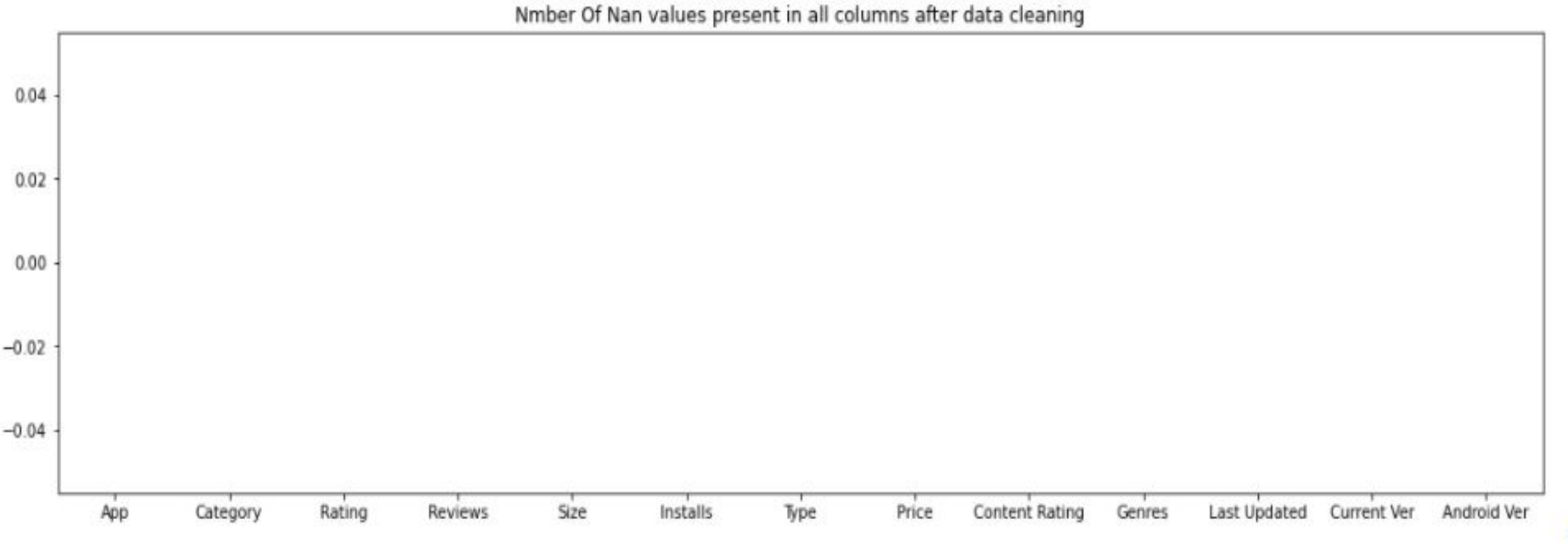
Important Features :

- App – An app name
- Sentiment – Sentiment given to an app by users (i.e Positive, Neutral, Negative).
- Sentiment Polarity – The polarity of sentiment measures how negative or positive the context is. In the data we have, the polarity ranges from +1(Positive) to -1(Negative).
- Sentiment Subjectivity - The subjectivity of a sentiment is how likely that sentiment is to be based on data or factual information, versus personal opinions or public notions.

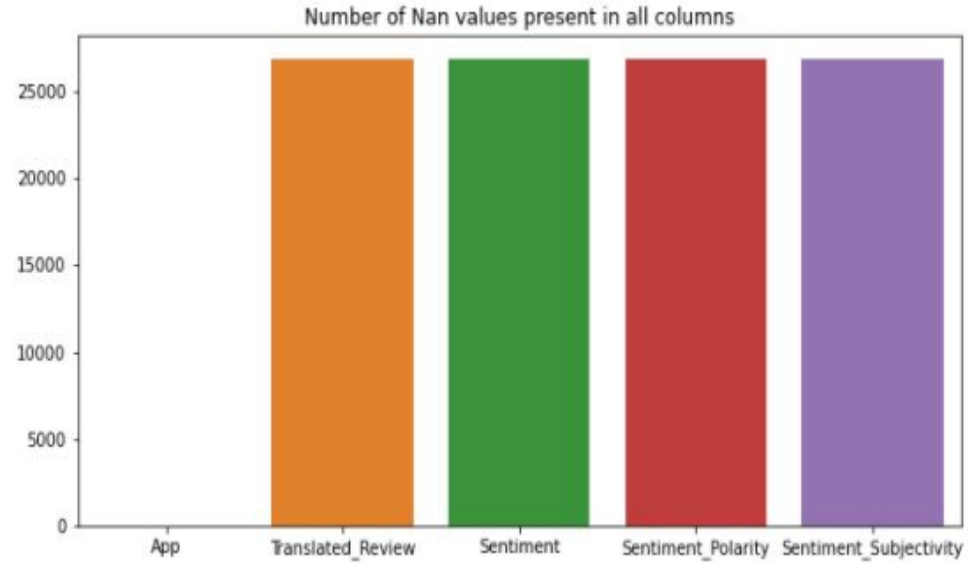
Where are the missing values in Play Store Data?

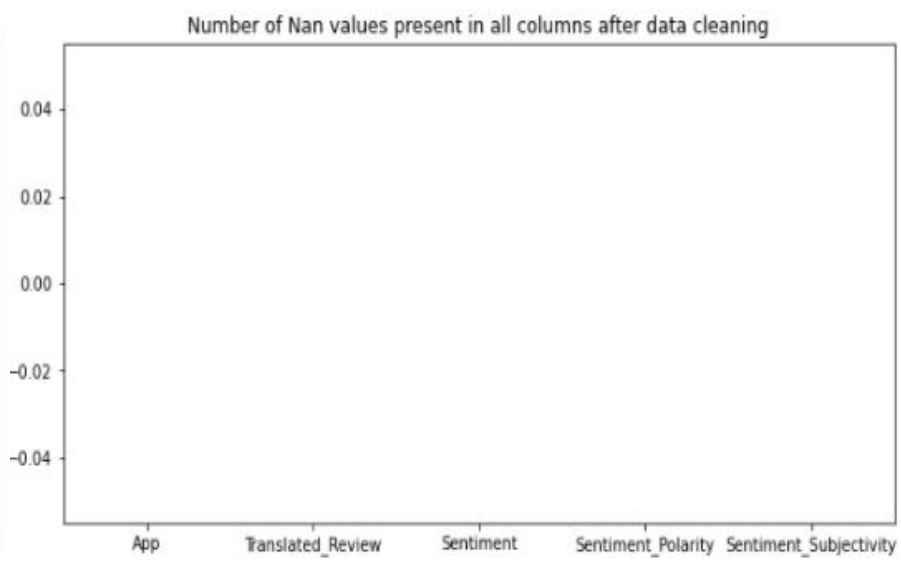


After doing the missing value treatment in “Rating”, “Current-Ver” & “Android-Ver”



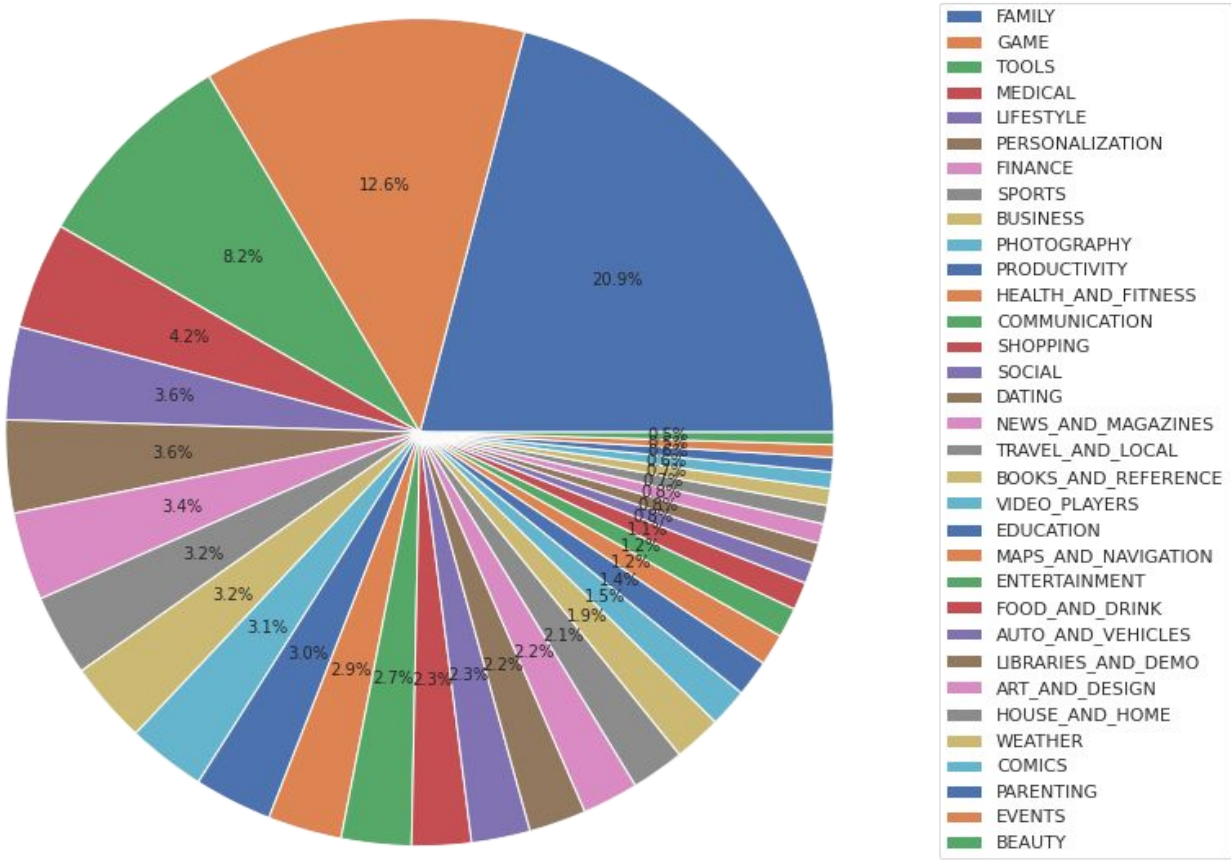
Missing values in User review Data & After cleaning the Data





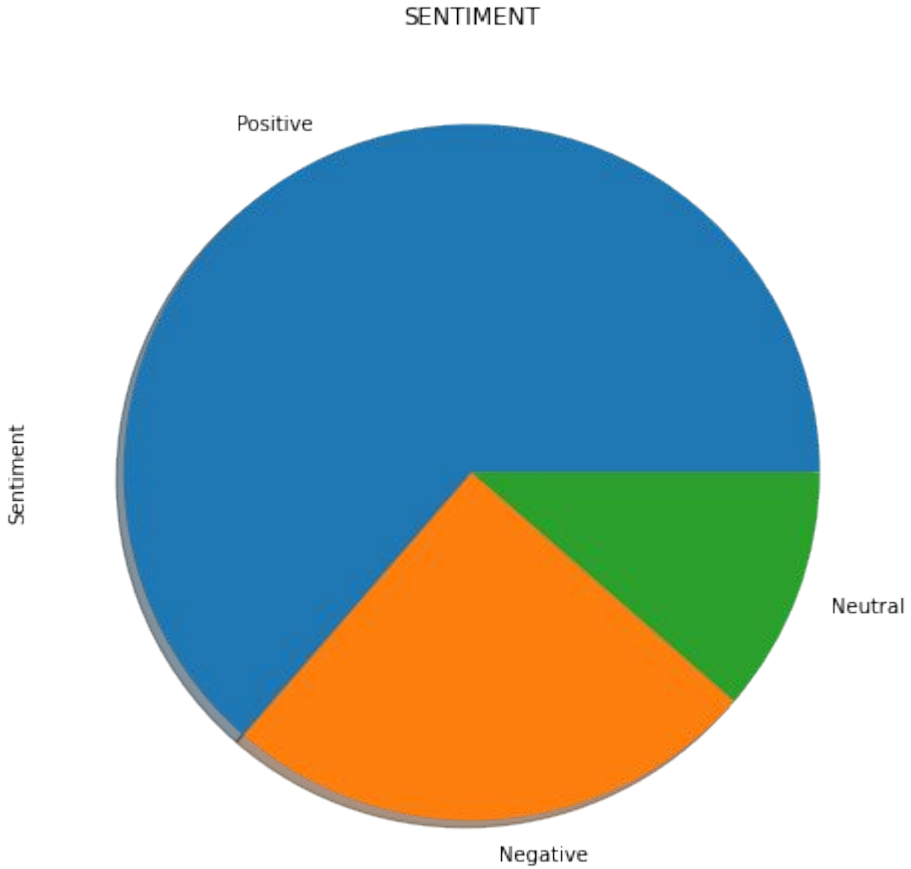
Exploratory Analysis

Overall Analysis of Category



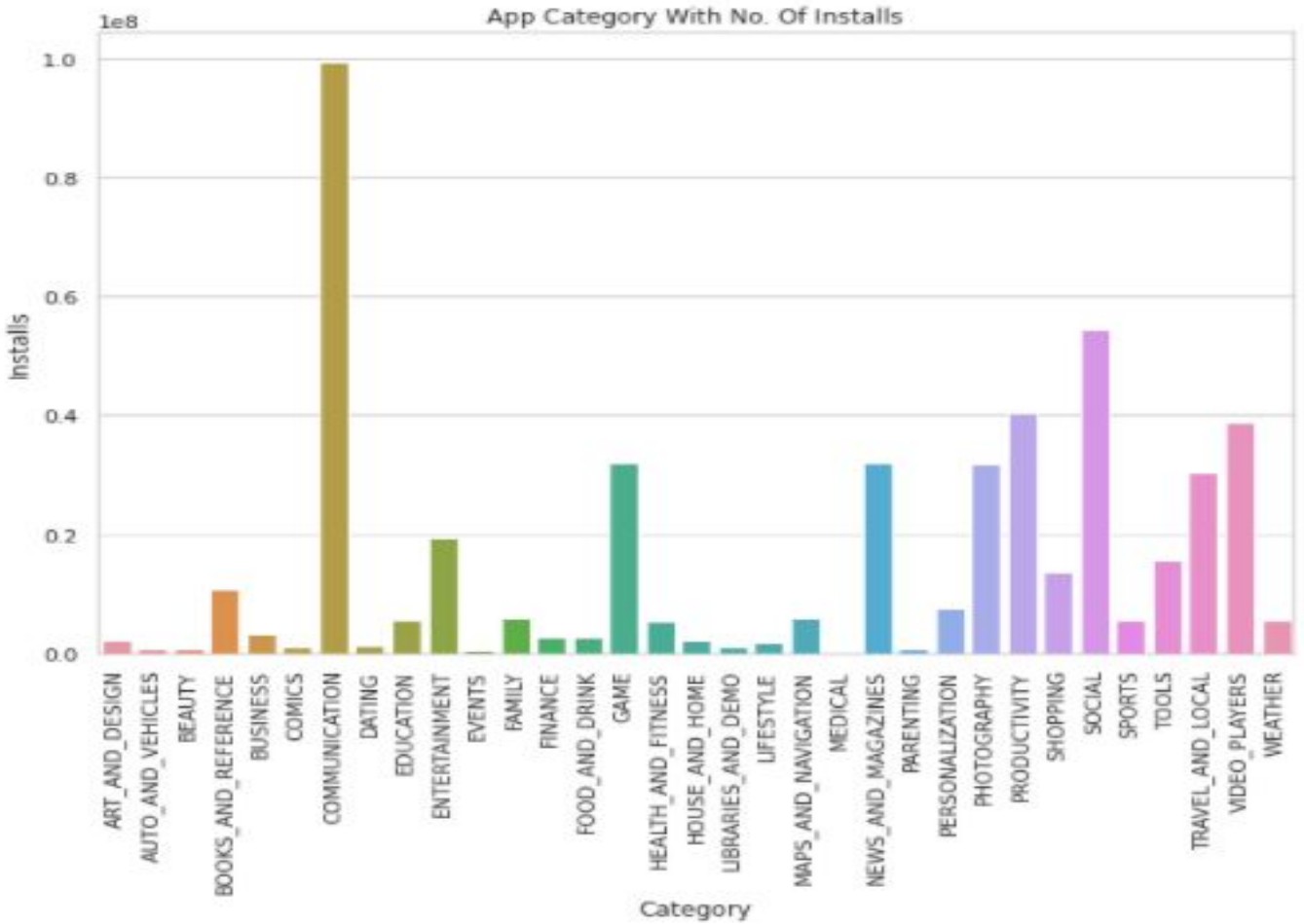
Analysis of Content Rating

Analysis of Sentiment

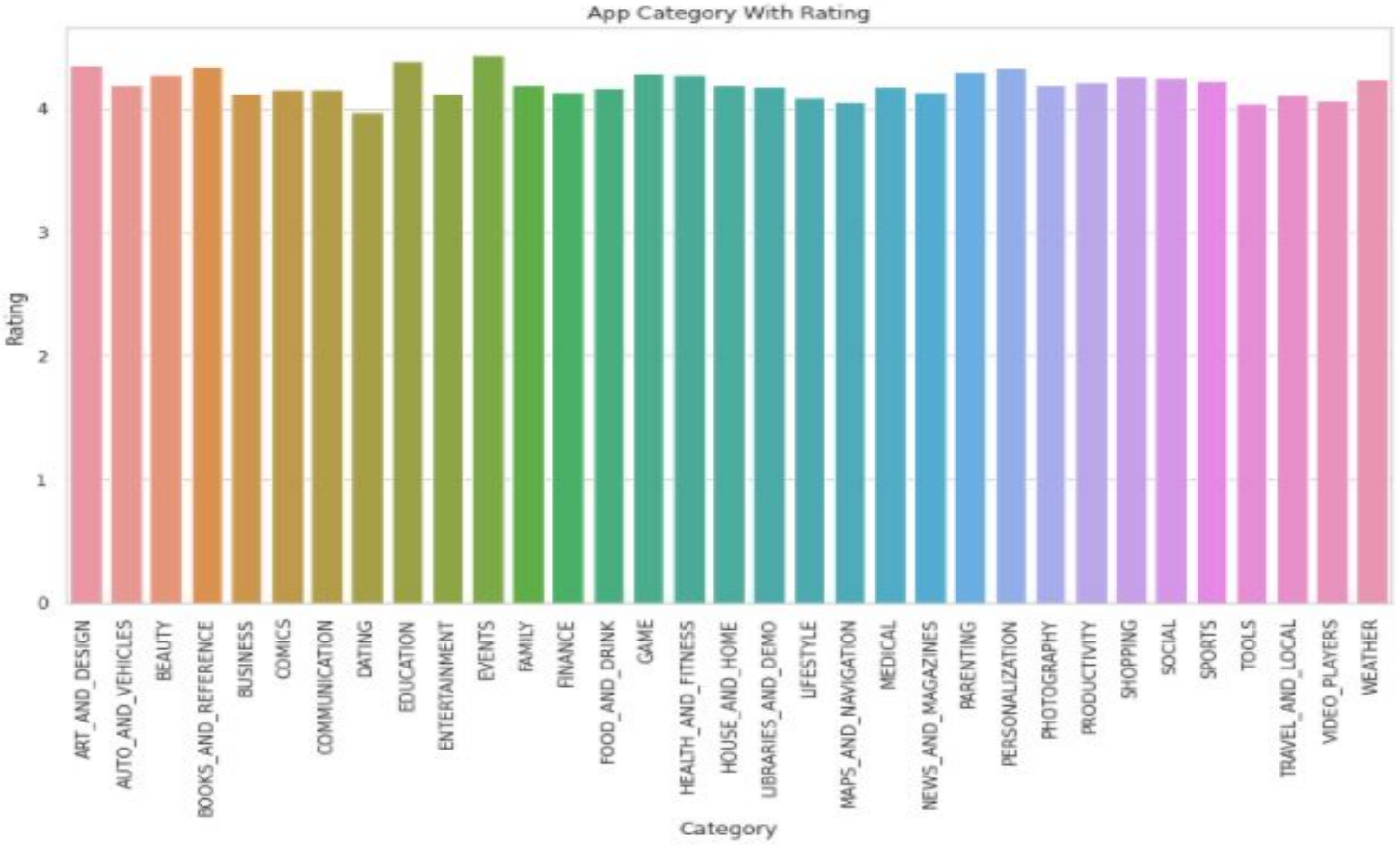


Best Performing Categories

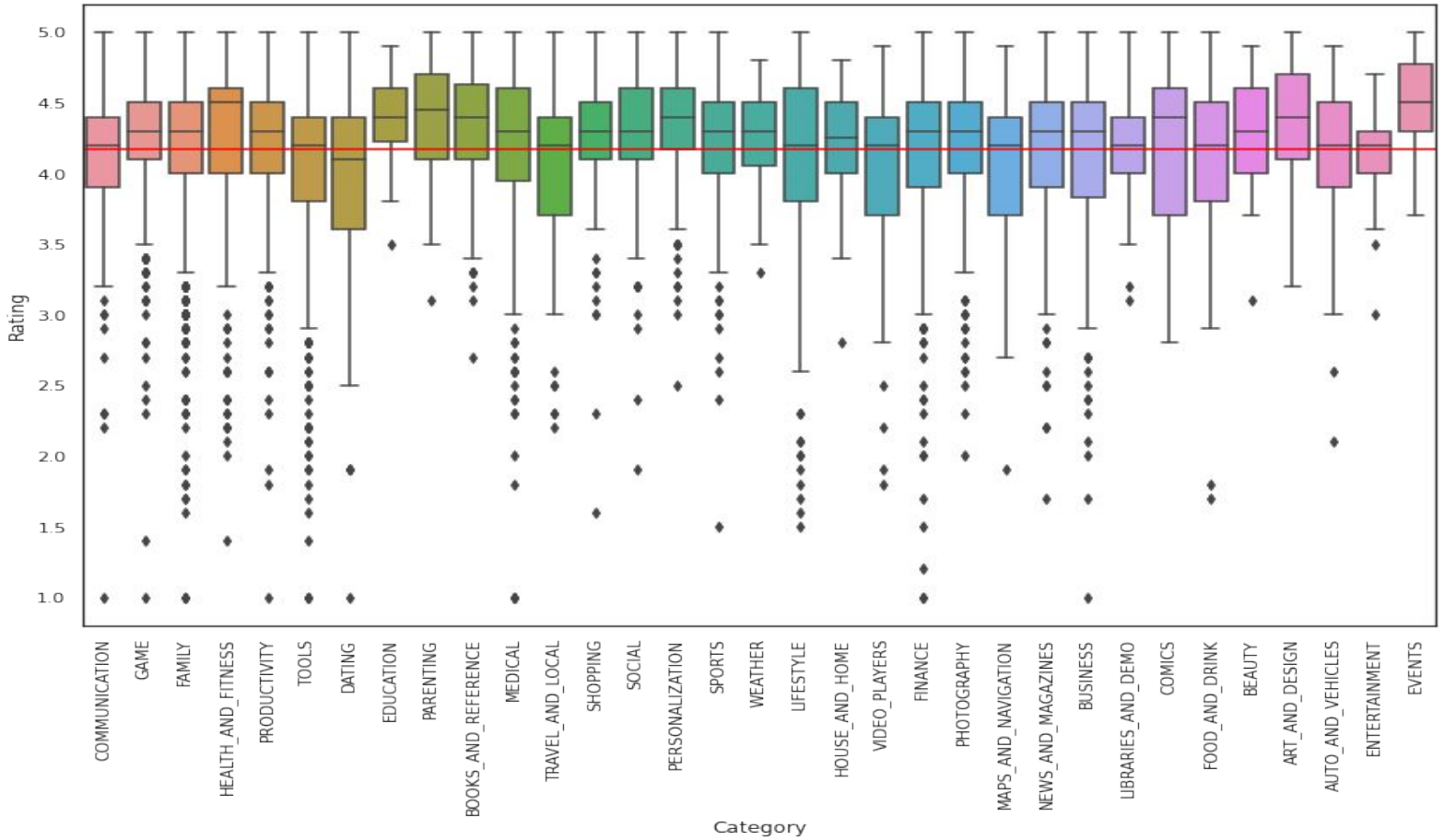
Analysis of app category w.r.t no. of Installs



Analysis on Category w.r.t Ratings



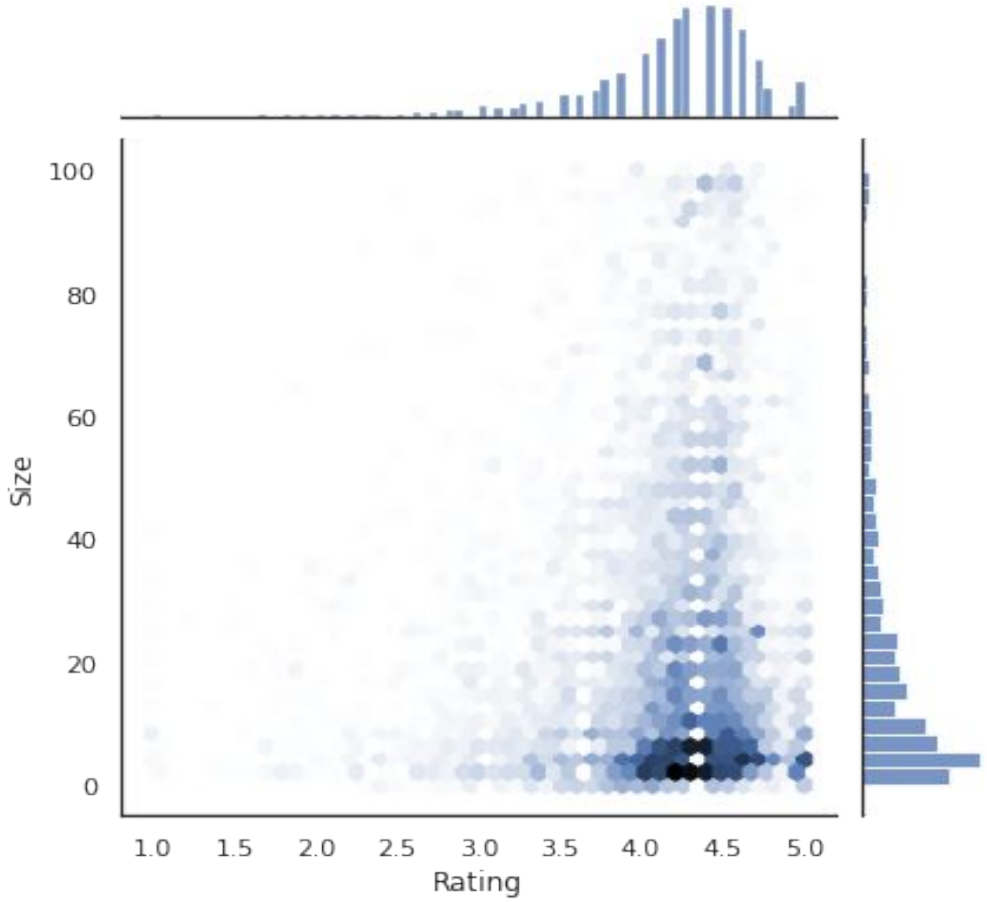
Box-plot for all Categories



- Performance of all the Category are mostly descent.
- **Health & Fitness** and **Book & Reference** category app have more than 50% of apps with a rating higher than 4.5 .
- The **Communication** and **Social** category app have most number of installs.
- Apps in **Dating** category having lower rating than the average ratings is 50%

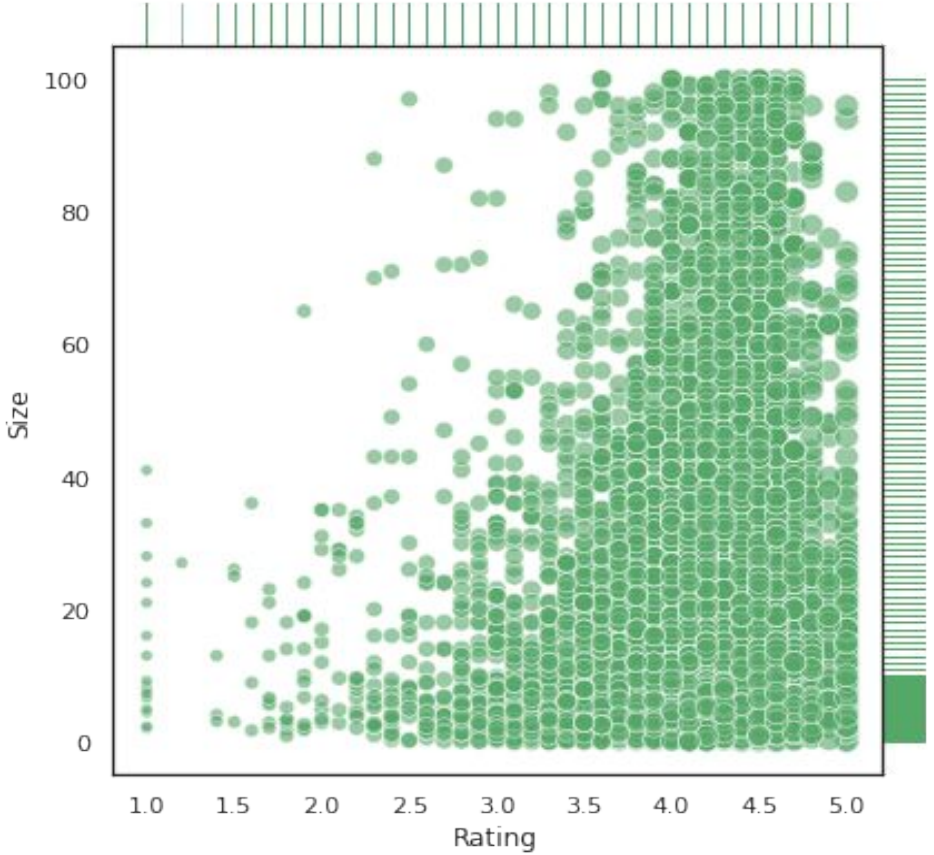
Light Vs Bulky?

How do app sizes impact the app rating?



Analysis on Size

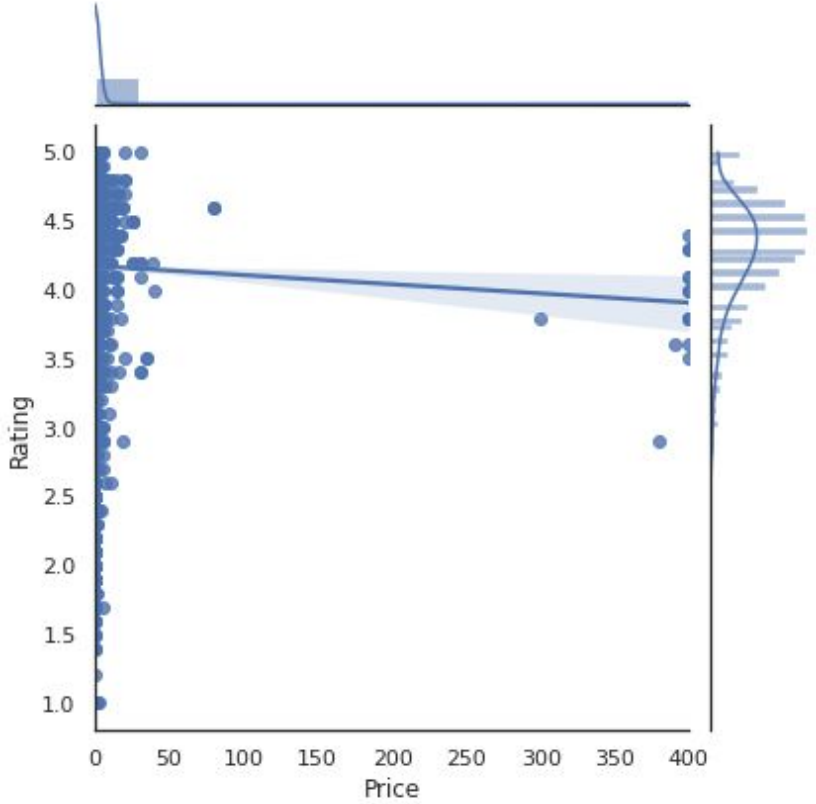
- Generally on increasing Rating, Size of App also increases.
- Most top rated apps are optimally sized between 2MB to 40MB neither too light nor too heavy.



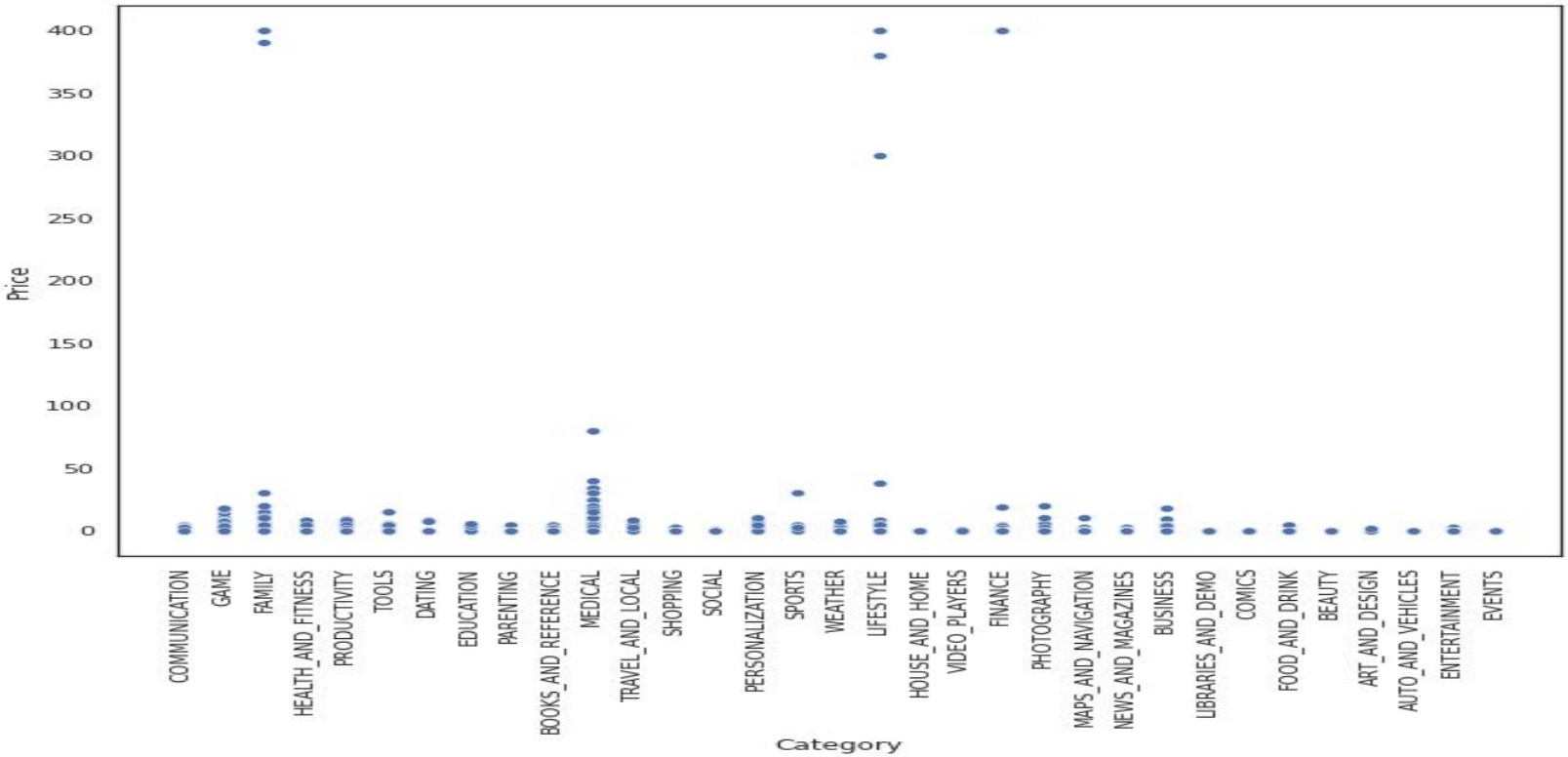
Free Vs Paid ?

How do app prices impact app rating?

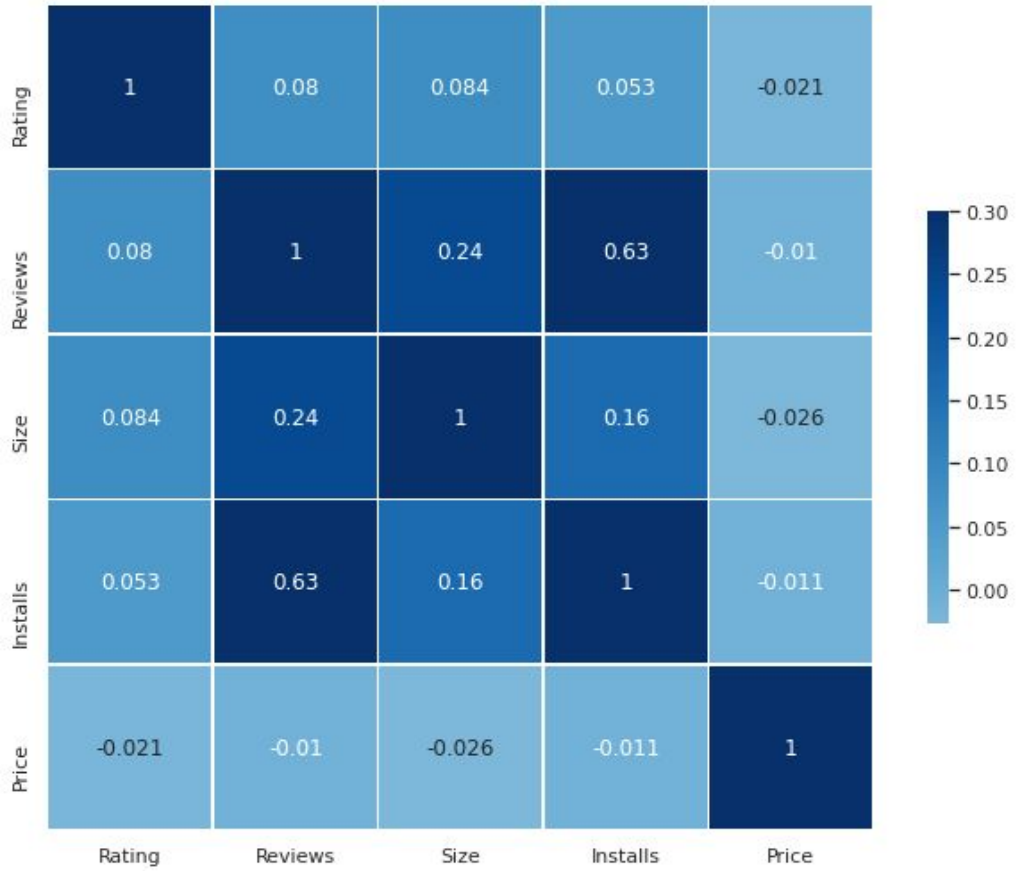
- Most top rated apps are optimally priced between ~1$ to ~30$.
- There are only a very few apps priced above 250$



Analysis on Category w.r.t Price



Exploring Correlations



Data Preparation:

- The features we have in both dataset are not sufficient to predict.
- We chose the features of both the dataset which is required for prediction.

Data preparation (Contd..)

- We sorted the **Install** column and we get the invalid entry just shown below in **Rating, Category & Installs** column we dropped that row and dropped the null values in the **Rating** column.

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating
10472	Life Made WI-Fi Touchscreen Photo Frame	1.9	19.0	3.0M	1,000+	Free	0	Everyone	NaN
420	UC Browser - Fast Download Private & Secure	COMMUNICATION	4.5	17714850	40M	500,000,000+	Free	0	Teen
474	LINE: Free Calls & Messages	COMMUNICATION	4.2	10790289	Varies with device	500,000,000+	Free	0	Everyone
3767	Flipboard: News For Our Time	NEWS_AND_MAGAZINES	4.4	1284017	Varies with device	500,000,000+	Free	0	Everyone 10+
3574	Cloud Print	PRODUCTIVITY	4.1	282460	Varies with device	500,000,000+	Free	0	Everyone

Contd...

- In **Current version and Android version** we replace the Nan value with common values using ‘mode()’ in **both columns.**

No. rows before operation	No. of rows after operation
10,841	9,363

- While doing the grouping by Category with the mean of Ratings & Installs, we’ll be left with these

	Category	Rating	Installs
0	ART_AND_DESIGN	4.358065	2.003760e+06
1	AUTO_AND_VEHICLES	4.190411	7.278055e+05
2	BEAUTY	4.278571	6.408619e+05
3	BOOKS_AND_REFERENCE	4.346067	1.079377e+07
4	BUSINESS	4.121452	3.306165e+06

No. rows before operation	No. of rows after operation
9,363	33

Contd…

- In the User Review dataset we have dropped all Nan values and we get the data that we need for the sentiment analysis.

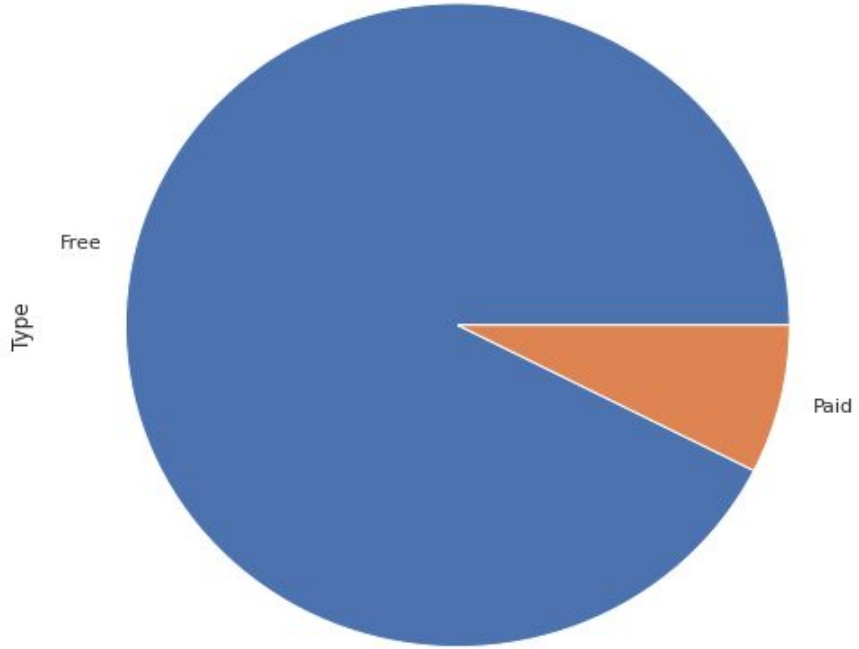
No. rows before operation	No. of rows needed for sentiment analysis
6,4295	3,7427

Characteristic Evaluation

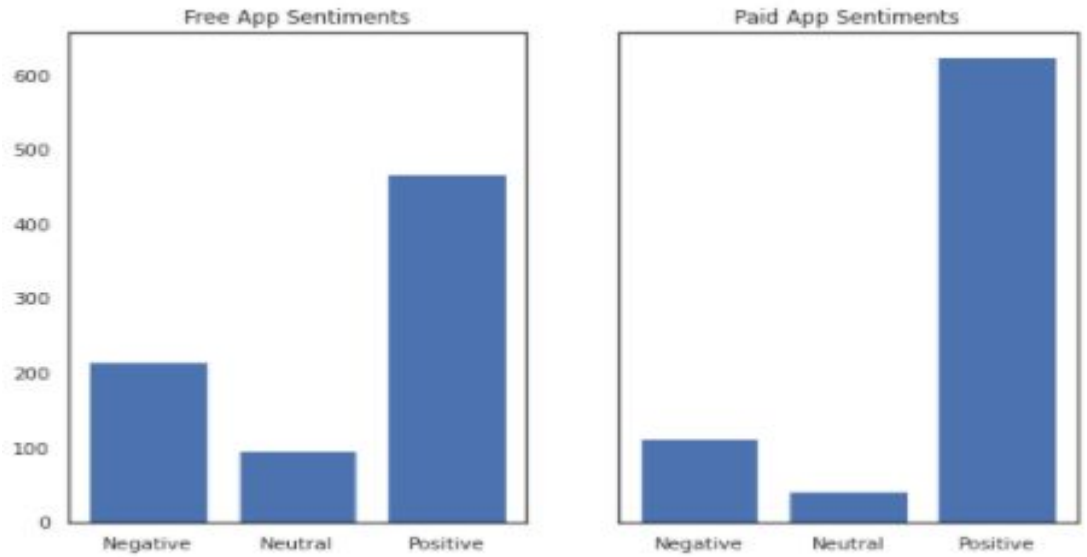
Type of Apps

- Paid apps have a relatively lower number of downloads than free apps

- **Free** −7149
- **Paid** - 577

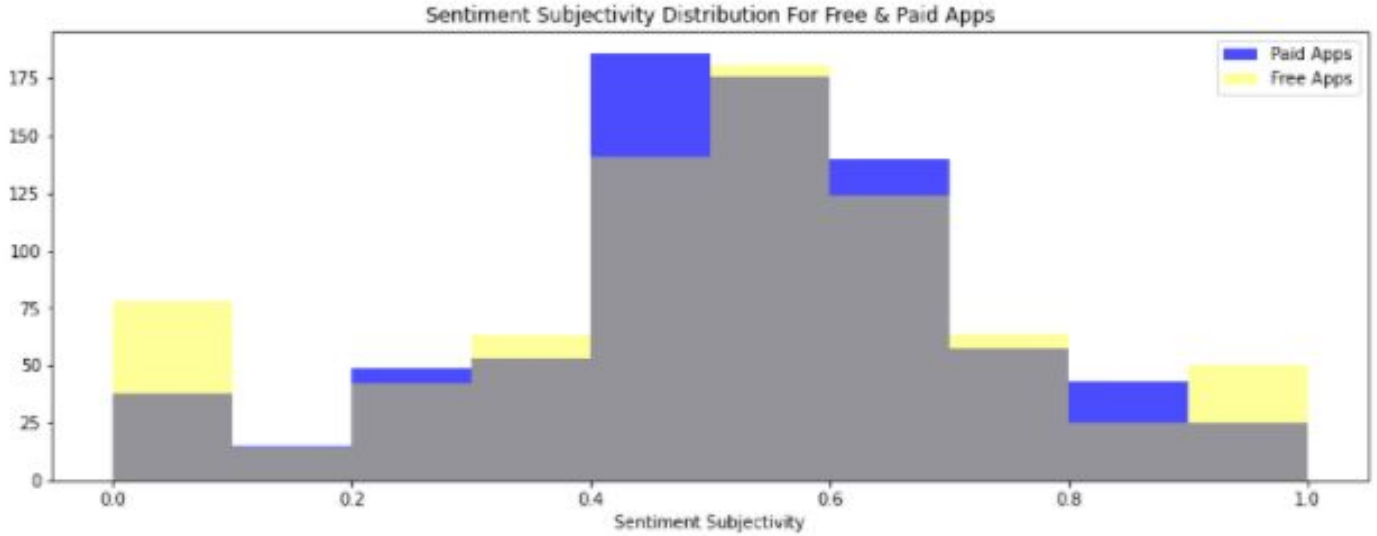


Analysis of Sentiment on free apps and paid apps



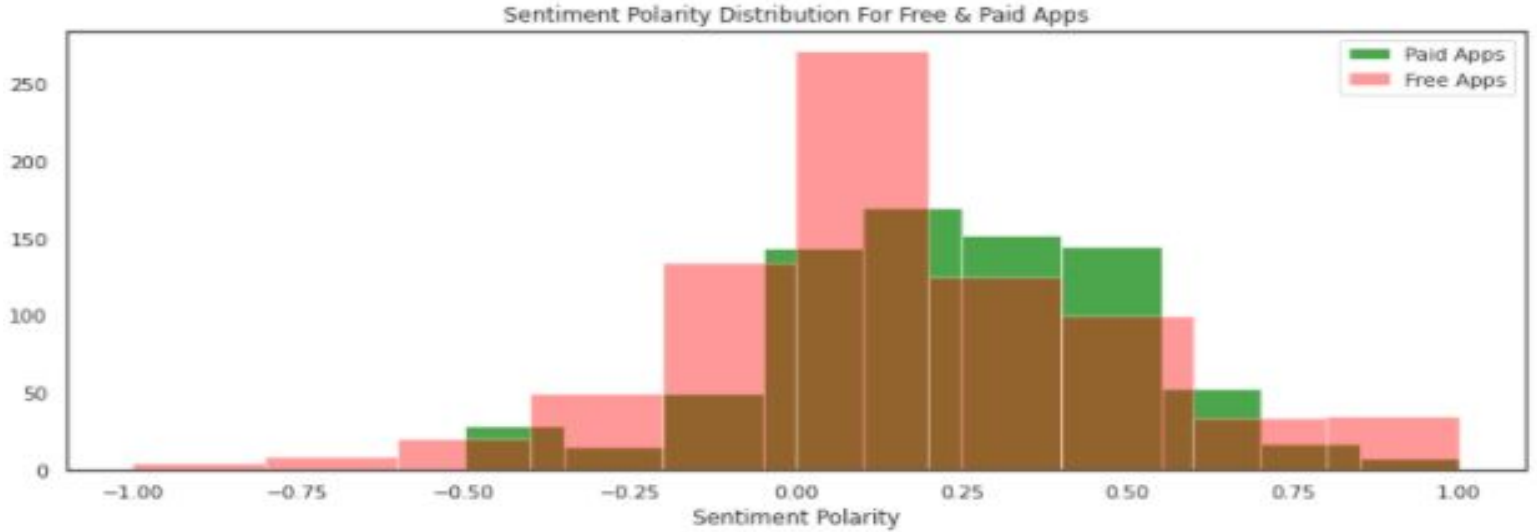
- As we can see Free apps have more negative and neutral reviews, indicating higher variance of sentiments for Free apps.

Analysis of sentiment subjectivity on Free and Paid apps



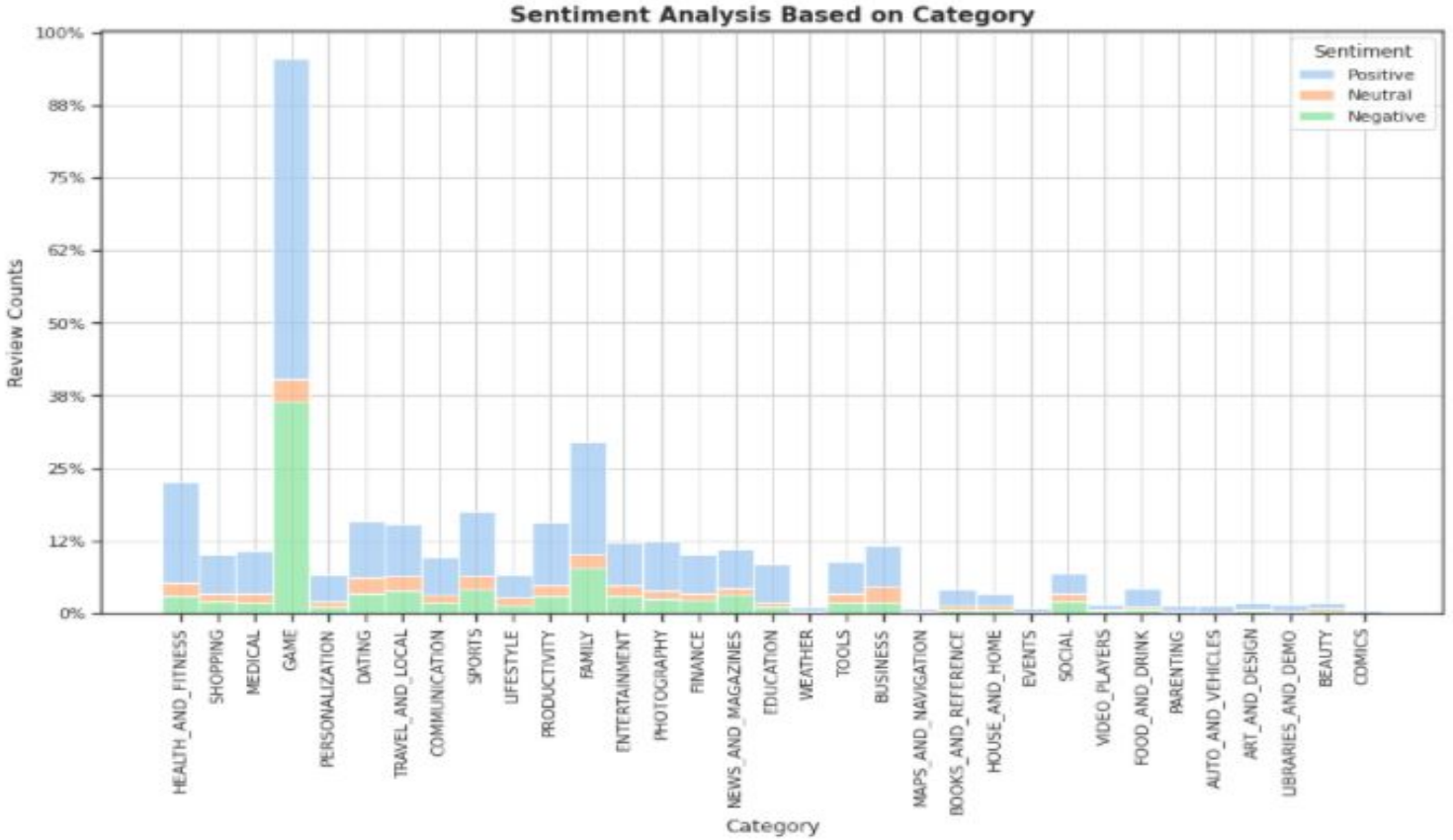
- Both paid and free apps seem to have a distribution of sentiment subjectivity that is very close to a normal distribution, although paid apps have more data around the mean.
- The overall subjectivity of reviews of paid apps seem to be slightly lower than that of free apps.

Sentiment Polarity distribution of Paid app & Free app



- The polarity of a sentiment measures how negative or positive the context is.
- When we look at this graph, we find that there is more neutral polarity in free apps, as shown by the larger red region just above the 0.00 on the x axis.
- However there is a more extreme positive polarity for free apps, as seen on the x axis in the range of 0.5 to 1.00.
- When it comes to Paid apps, the majority of opinions fall somewhere between 0 and 0.5.

Sentiment analysis based on Category



Sentiment analysis based on Category (Contd..)

- Family, Sports and Health & Fitness apps perform the best, Having more than 50% positive reviews.
- On the contrary, many Game and Social apps perform decent leading to 50% positive and 50% negative.

A quick look on reviews of Free & Paid Apps

Free Apps

Paid Apps

Challenges:

- There are lots of null values in both dataset.
- Making the data more accurate.
- The Play Store dataset contains lots of information in gibberish form.

Conclusion:

- Average rating of (active) apps on Google Play Store is 4.17.
- Users prefer to pay for apps that are light-weighted. Thus, a paid app that is bulky may not perform well in the market.
- Most of the top rated apps are optimally sized between ~2MB to ~40MB - neither too light nor too heavy.
- Most of the top rated apps are optimally priced between ~1to 30 - neither too cheap nor too expensive.
- **Medical and Family** apps are the most expensive and even extend up to 80$.
- Users tend to download a given app more if it has been reviewed by a large number of people.
- Paid apps have a slightly higher number of favourable reviews than free apps.
- Free apps get more negative and neutral feedback, suggesting a wider range of opinions.
- When it comes to free apps, users are more pessimistic and harsh than when it comes to paid apps.
- More than half users rate **Family, Sports and Health & Fitness** apps positively. Apps for games

Q&A