Capstone Project

Play Store App Review Analysis

Team Members

Swapnil Patil
Harish Gawade
Tushar Wagh

Introduction

The Play Store apps data 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 is one of the largest and most popular android app stores with over a million apps.

Our main objective is to analyze the dataset and find out which features contribute to app success and how these features affect the user engagement with the app.

Steps Involved

- **Importing Libraries and data:-** First, we imported all the python libraries required for this, which include NumPy, Pandas, Matplotlib and Seaborn. We read the CSV into a dataframe and pandas dataframe did the work for us.
- **Discover and understand data:-** In this step, we observed the data by exploring few rows, checking shape, columns, data types etc.
- **Data Preparation:-** Here we carried out data cleaning and data transform to make data efficient for analysis and visualisation.
- **Data Visualisation:-** Visualized data with the help of graphs and plots to learn trends, patterns and get answers to the questions related to the data. This process helped us figuring out various aspects and relationships among features of the app.

Data Overview

10841 Apps

13 Columns

Different Data types

- String
- Integer
- Float
- Object

Features:-

- **App** - name of the app
- **Category** - category of the app
- **Rating** - app's rating by the users out of 5
- **Reviews** - number of the app's reviews
- **Size** - size of the app
- **Installs** - number of installs of the app
- **Type** - whether the app is free or paid
- **Price** - price of the app in $
- **Content Rating** - target audience of the app
- **Genres** - genre of the app
- **Last Updated** - date the app updated last time
- **Current Ver** - current version of the app
- **Android Ver** - android version required to run the app

Data Preparation

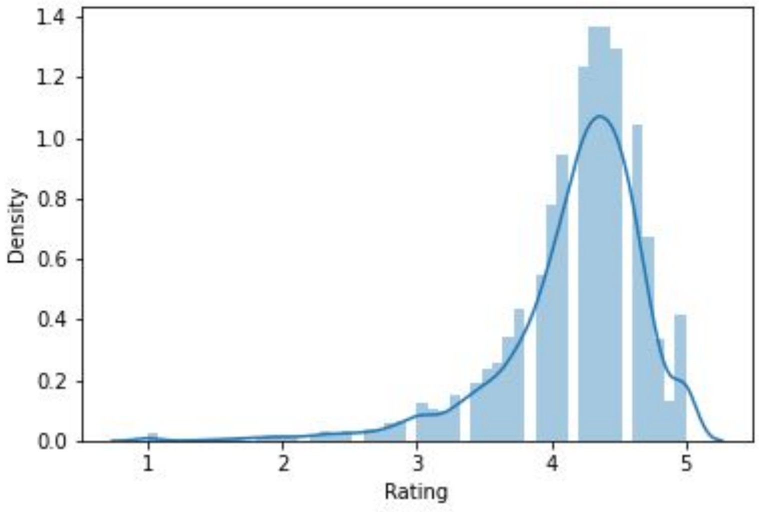
Missing Data

- Overall 1476 null values
- Rating has 99% of total missing values

Treating missing data

- Analyzed features one at a time
- Replaced rating missing values with median value
- Treated other columns missing data in best way possible

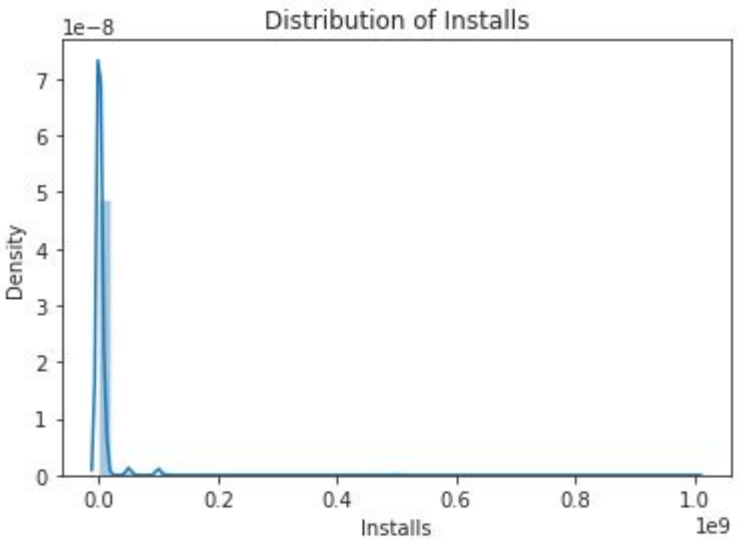
Distribution of Rating

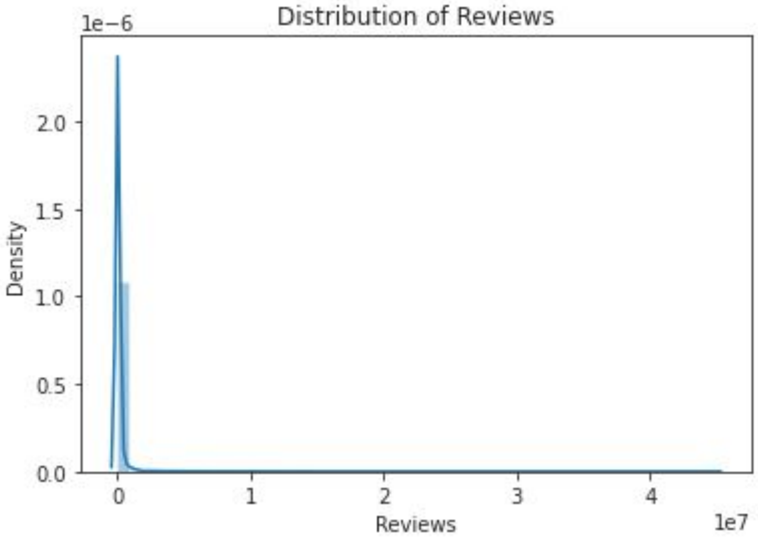


Data Preparation contd.

Data Transformation

- Changed the data types to their original one
- Performed Log Transformation on Installs and Reviews column

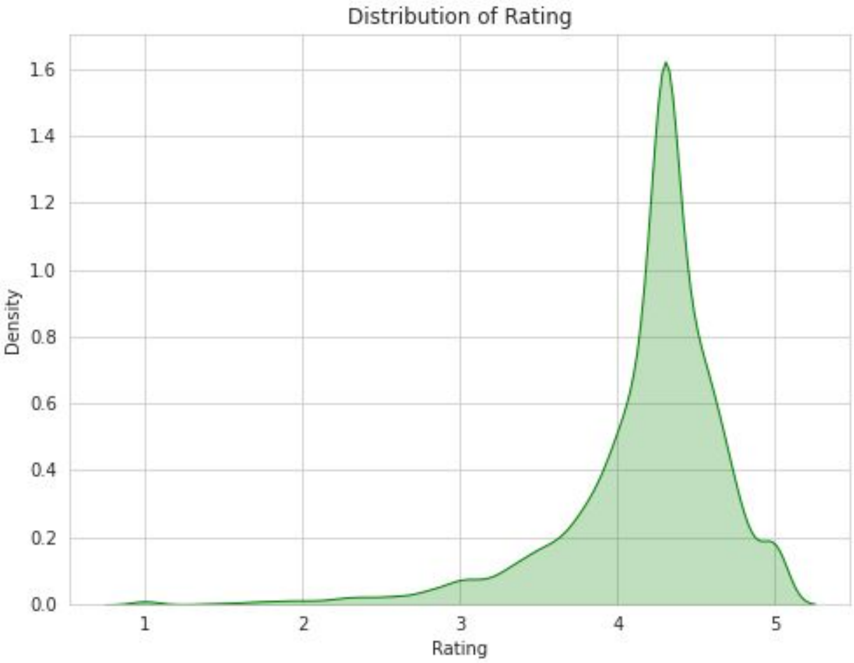




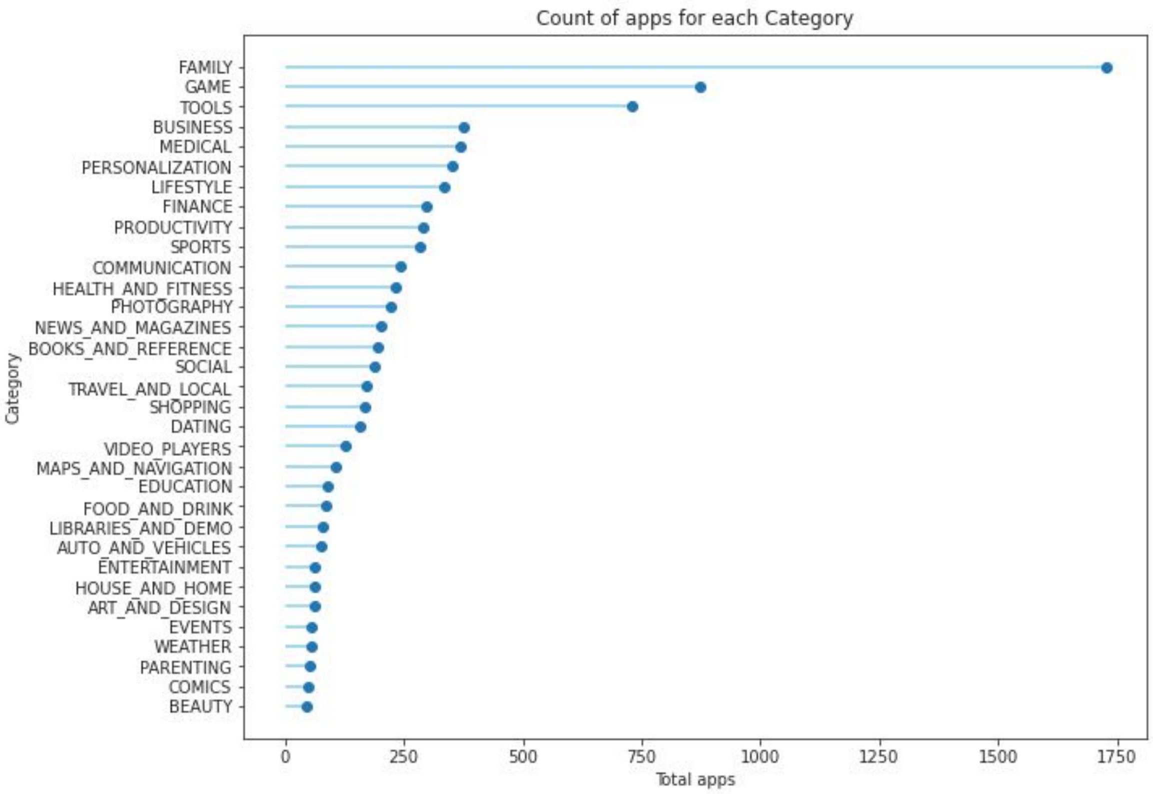
Data Insights

Distribution of Rating

- Distribution of rating is negatively skewed
- Average rating is around 4.18
- Most of the rating is above 3.5
- **Rating can be a variable to identify app success.**

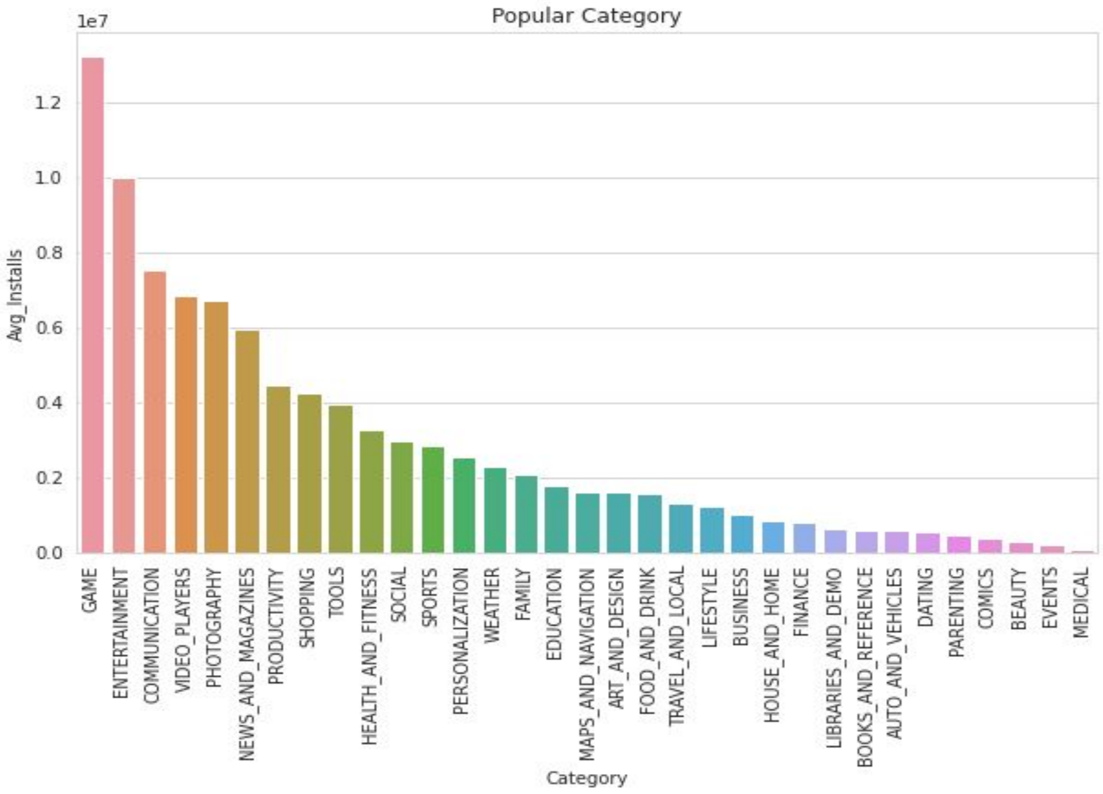


How many apps are there in each category?



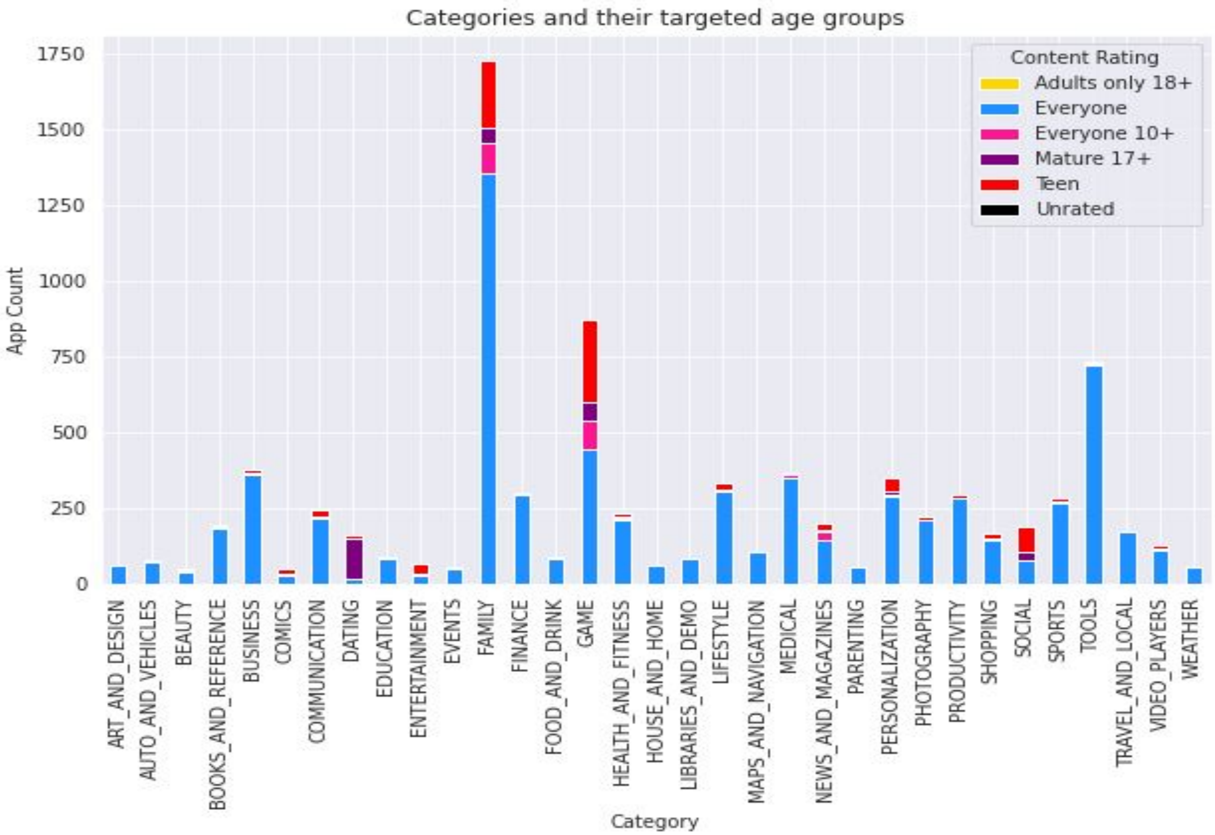
- Play store has 33 Categories in total.
- Family category has most 1726 apps.
- Game and Tools category have 873 and 731 apps respectively.

Which category apps are installed most?



- Game is the most popular category.
- Entertainment, Communication, Video players, Photography categories are also popular among smartphone users.
- Medical category has least number of installs.

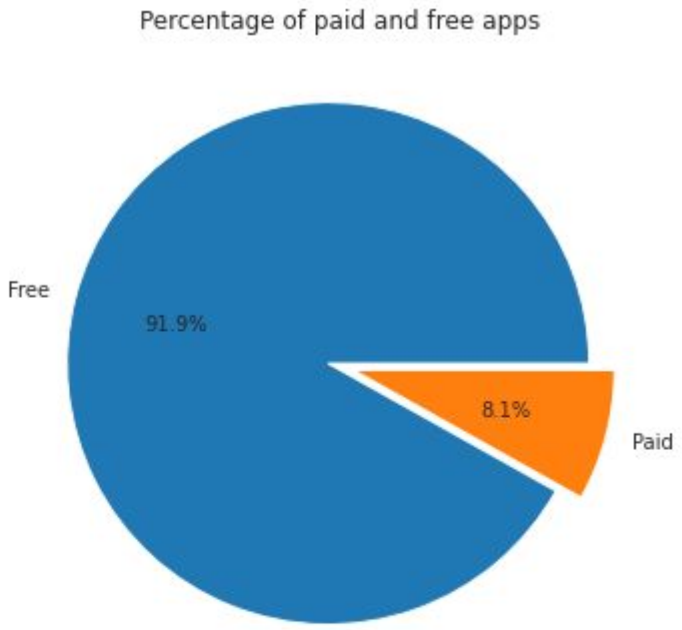
Which age groups do different categories target?

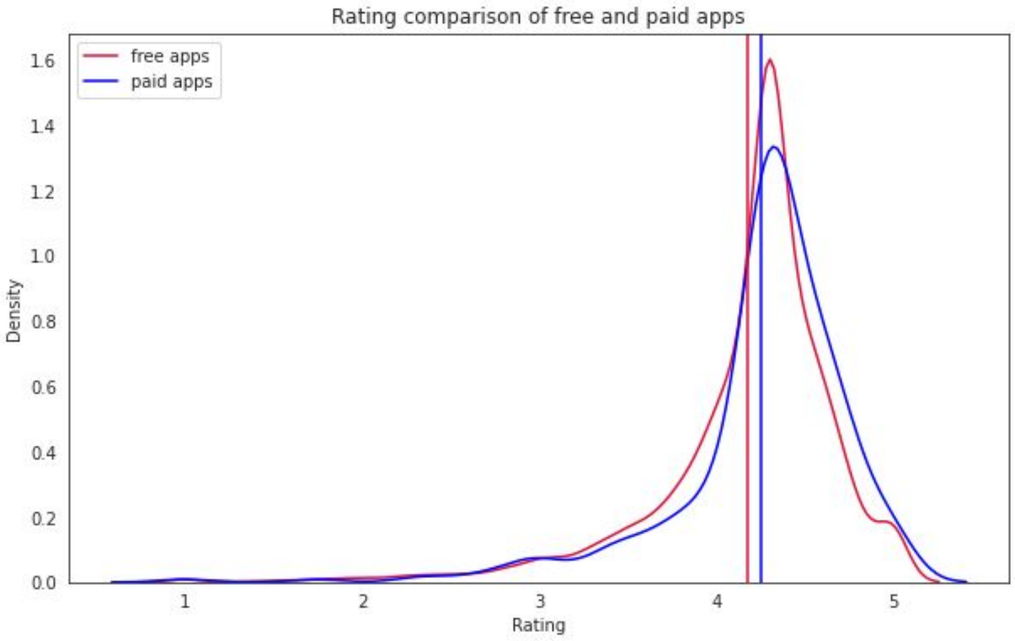


- Every category targets almost all age group audiences.
- Dating category has almost all apps for the mature audience.

What percentage of apps are paid?

Do paid apps get better rating?

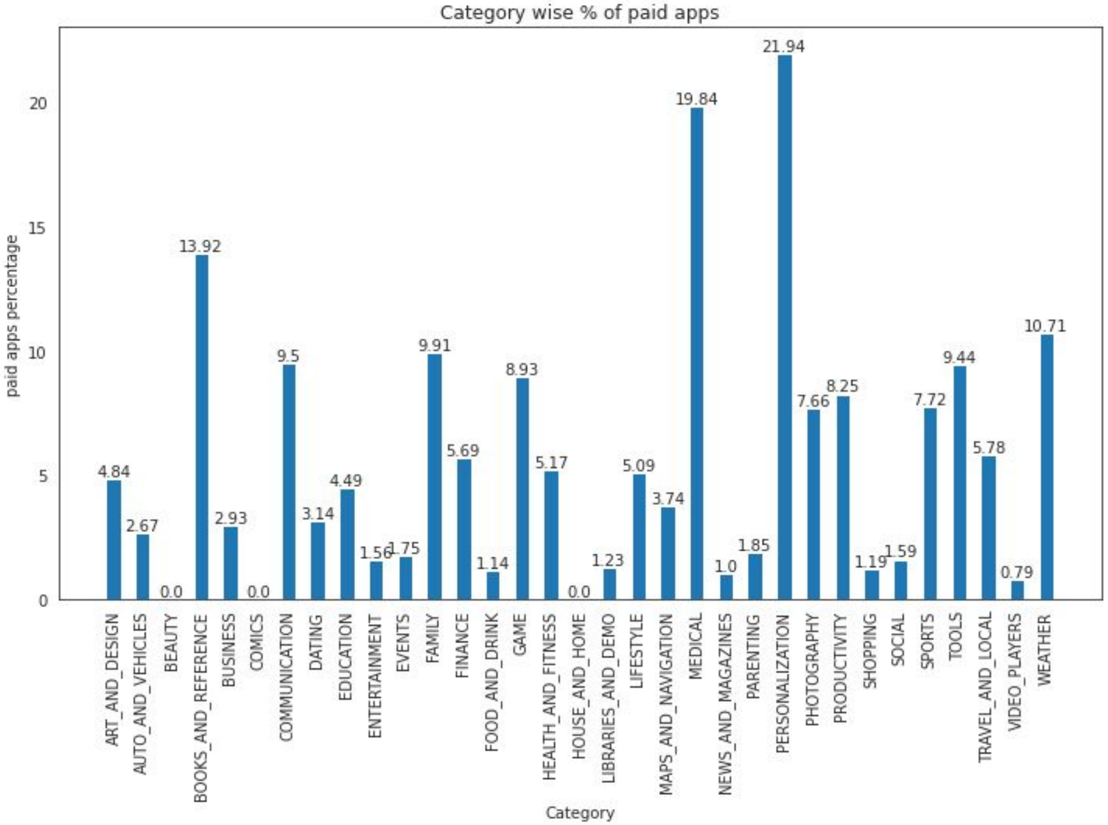




Free apps - 7747 **Paid apps - 685**

Average Ratings:- **Free apps - 4.18**
Paid apps - 4.26

Which type of apps are users willing to pay for?

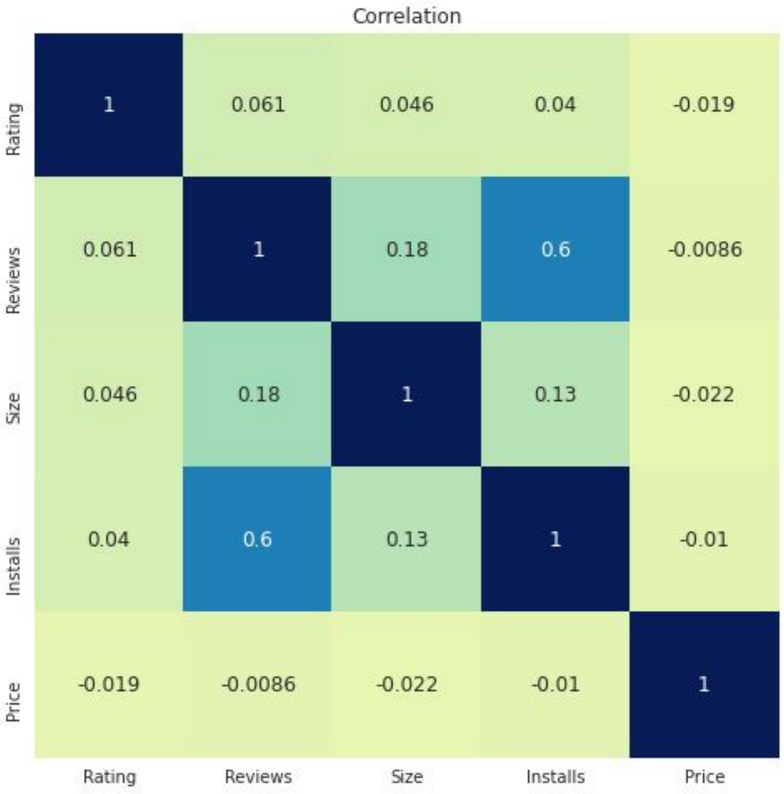


- Personalization and Medical category have high rate of paid apps.
- This type of apps generally do well as paid apps, since value is in the apps functionality.

Bivariate Analysis

Correlation Heatmap

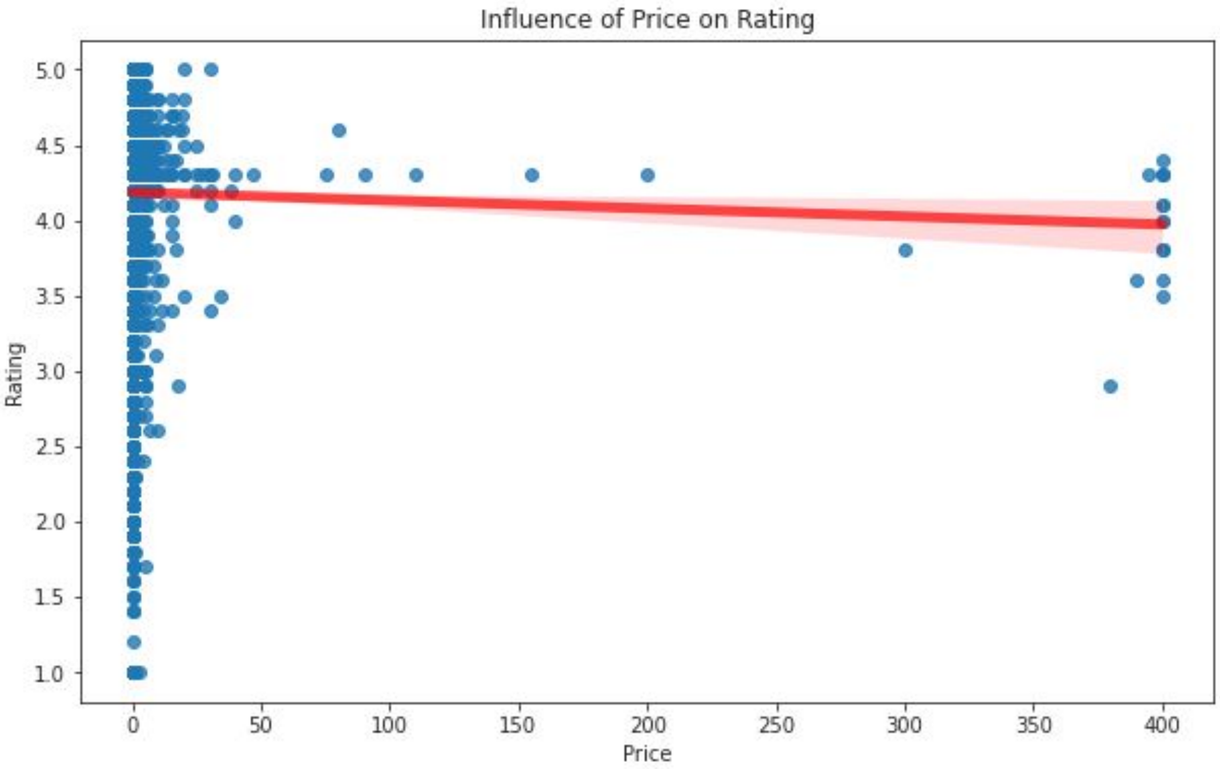
- Installs and Reviews have the strongest correlation
- Rating has negative relation with Price
- Installs has positive relation with Size but it is very weak



Correlation

	Rating	Reviews	Size	Installs	Price
Rating	1	0.061	0.046	0.04	-0.019
Reviews	0.061	1	0.18	0.6	-0.0086
Size	0.046	0.18	1	0.13	-0.022
Installs	0.04	0.6	0.13	1	-0.01
Price	-0.019	-0.0086	-0.022	-0.01	1

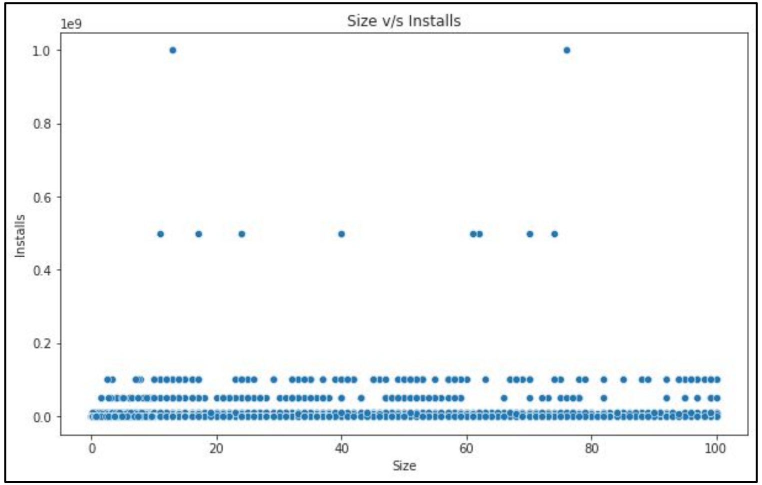
Does rating change with increasing price?

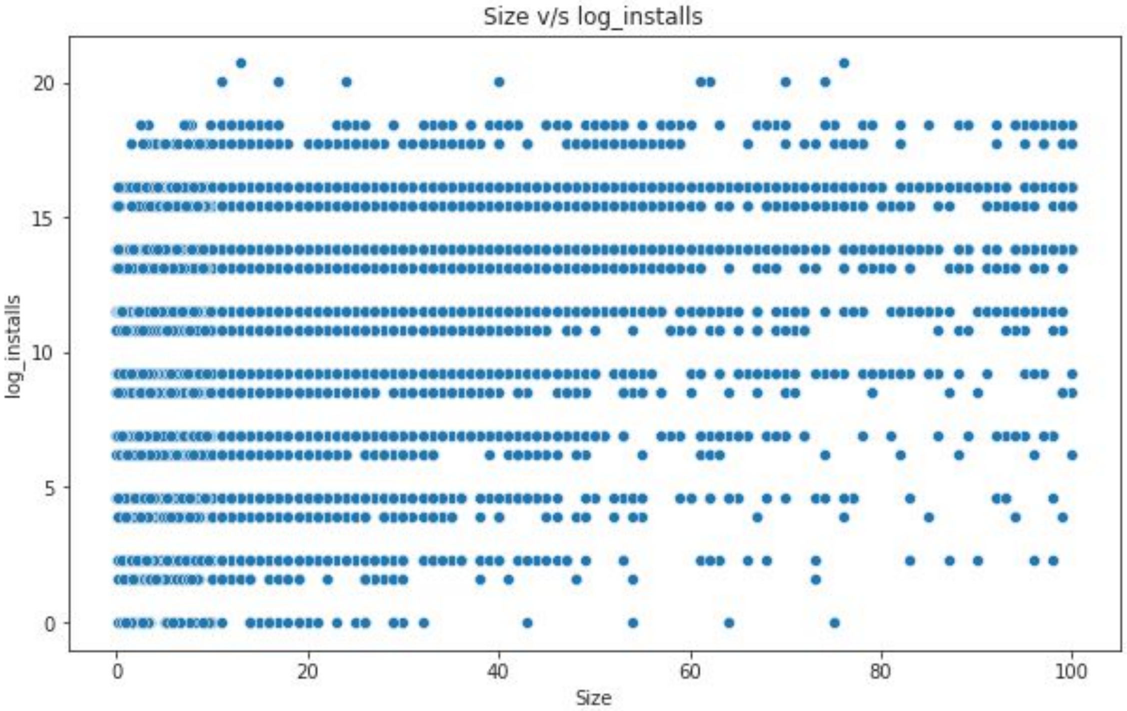


- Majority of apps cost below $100
- There is negative relation between price and rating.
- Rating decreases with increasing price.

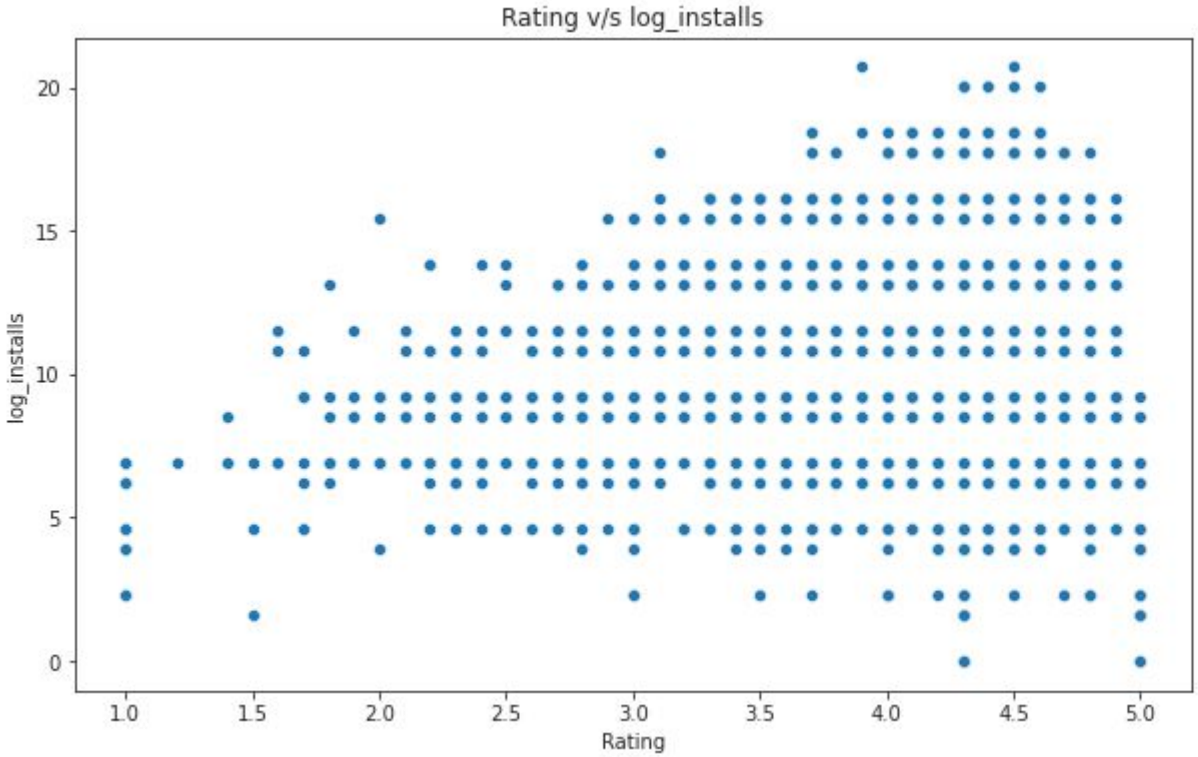
Does the size of an app influence the installs?

Before Log Transformation





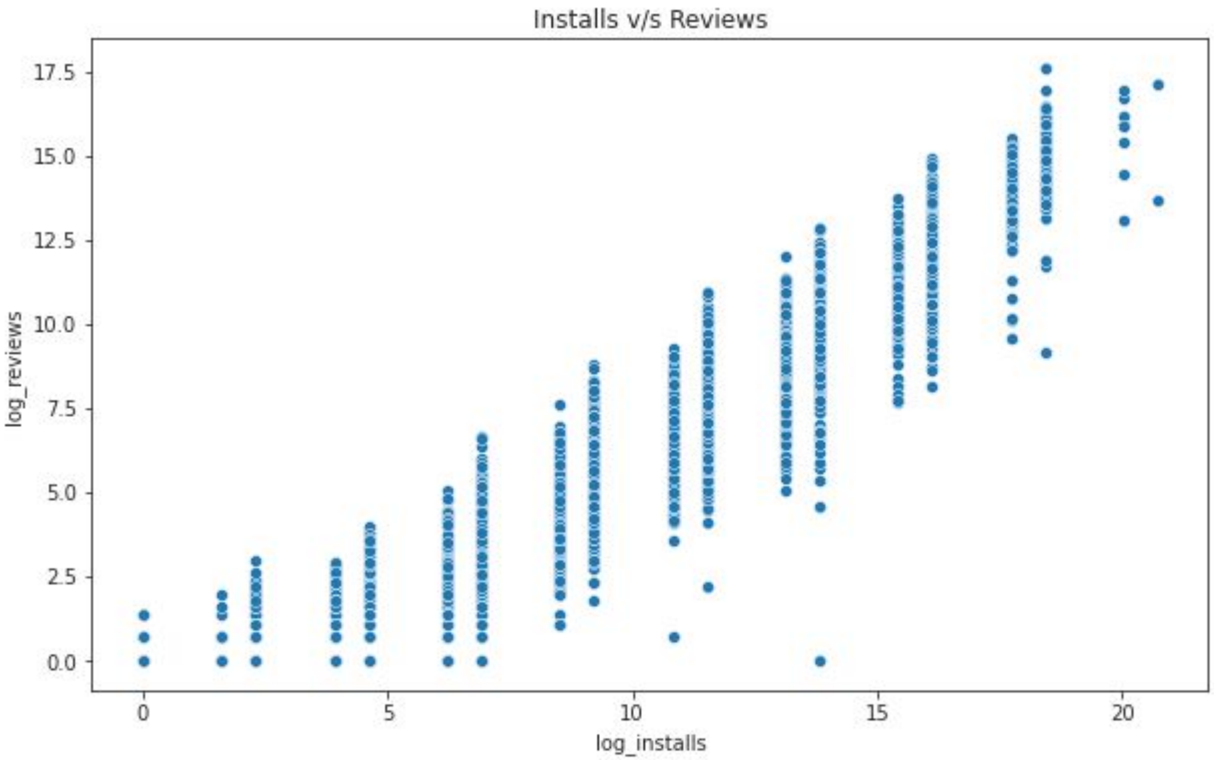
Do higher rated apps attract more users?



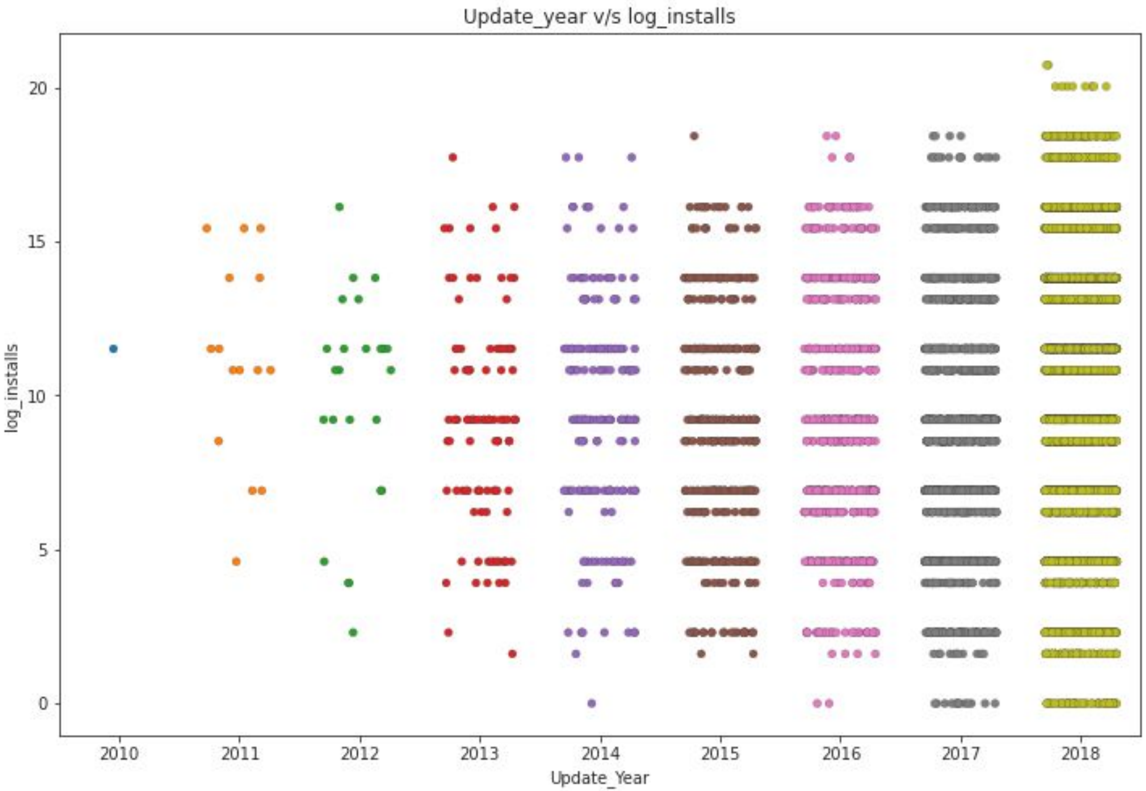
- Both are the most important features to look after.
- Apps with rating around 4.5 have more installs.
- Users prefer highly rated apps to download.

How reviews affect users decision to download apps?

- Popular apps tend to get more reviews.
- Users read the reviews which attract them to install the app.

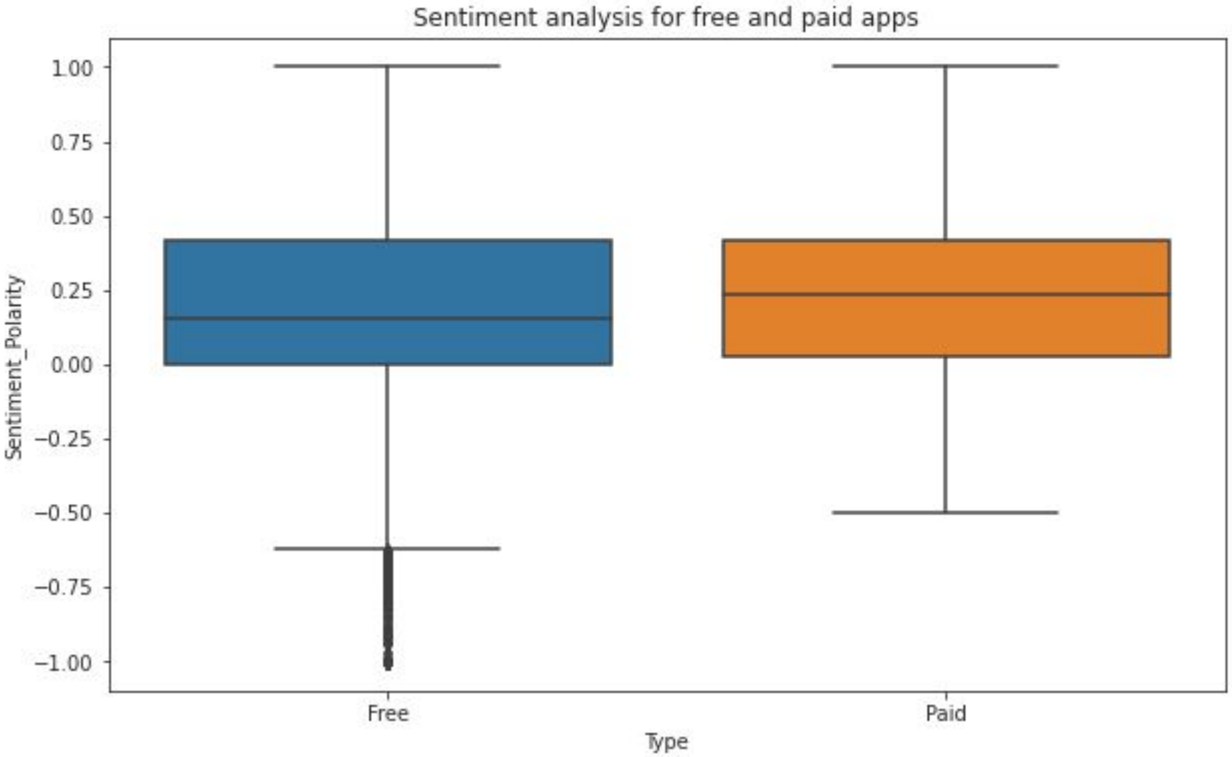


Are app updates important?



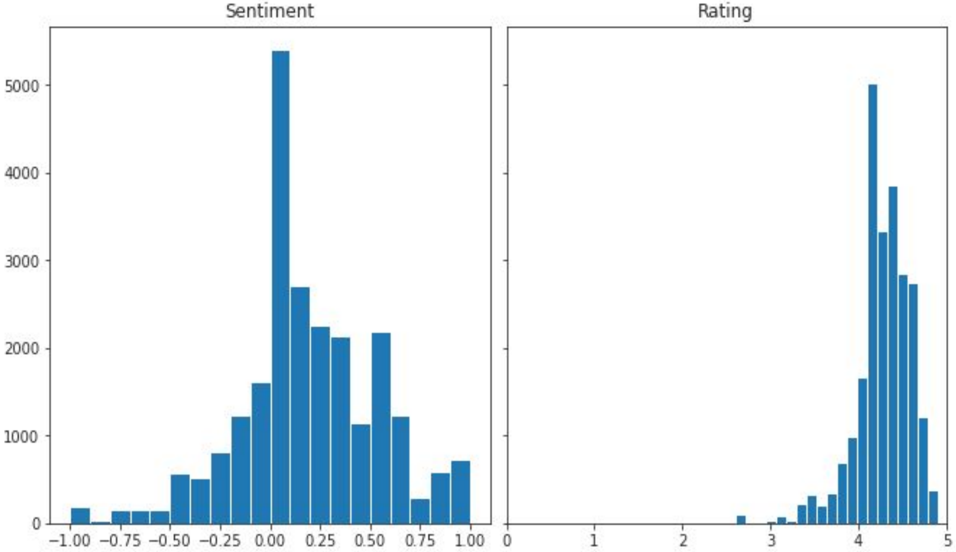
- Most of the apps get frequent updates.
- Updating apps can improve the user experience.
- leads to more convenience and increased engagement in the use of the app.

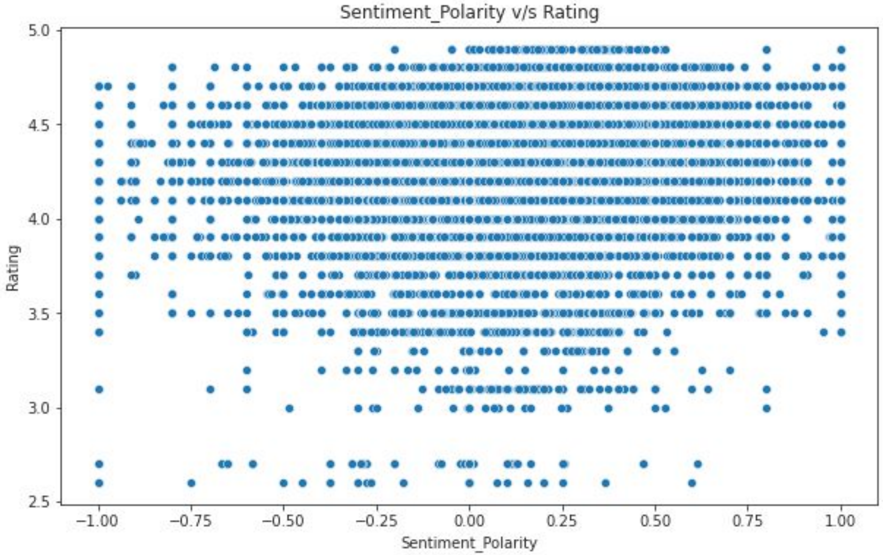
Sentiment analysis for free and paid apps



- Free apps get more negative reviews.
- Median polarity is higher for paid apps.

Are sentiment influences the final rating of the app?





- Positive trend for both rating and sentiment

Conclusion

- Family category has more apps on the play store but Game category is the most popular category.
- Approx. 91% apps on play store are free apps and Medical and Personalisation apps generally do well as paid apps.
- Users prefer apps that require less space. Bulky apps are downloaded less.
- App ratings and reviews have a significant impact on a user's decision to download or not download an app.
- Updating the app can improve user experience and happy users attract more new users.
- Sentiments in reviews also matters in attracting new users as other user's positive reviews about the app strengthen the decision to download.

Challenges

- Data cleaning was the most time consuming phase and whether to drop the null values or fill them was the real challenge.
- Installs and reviews were highly skewed and getting undesirable results lead to lots of search reaching to a solution of log transformation.
- Reviews dataset has so much data and it was little difficult to handle.

Q & A