

Capstone Project

Play Store App Review Analysis



Team member

Biswajeet sethi

PVN Malleswara Rao

Sanjay P Malviya

Abhilasha Rani Goel

Points for Discussion

- ▶ Introduction
- ▶ Objective
- ▶ Data Pipeline
- ▶ Data Summary
- ▶ Data Analysis Questions
- ▶ Data Cleaning
- ▶ Data Processing
- ▶ Exploratory Data Analysis (EDA)
- ▶ Bivariate Analysis
- ▶ Single Variate Analysis
- ▶ User sentiments Analysis
- ▶ Conclusion

Introduction To Google Play Store

- ▶ Google Play store is a digital distribution service for mobile applications on the Android operating system allowing users to browse and download applications developed with the Android software development kit.
- ▶ With computer science growing and open-source projects expanding, the Google Play store is increasing in popularity.
- ▶ While many public datasets provide Apple's App Store data, there are not many counterpart datasets available for Google Play store apps, yet the Google Play store data has enormous potential to drive application-making businesses.
- ▶ Unlike web development or desktop development, mobile development is unique in its convenience.
- ▶ With smartphones increasing in usage, mobile applications are growing in popularity as well.
- ▶ Actionable insights can be drawn for developers to work on and capture the Android market.

Objective

- ▶ 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.
- ▶ The objective of this project is to deliver insights to understand customer demands better and thus help developers to popularize the product.



Data Pipeline

- ▶ Understanding the Data: In this part we go through each columns, differentiate independent and dependent feature.
- ▶ Data Cleaning: In this process identify the errors and corruptions and either remove or manually replace with mean median and mode values and correct the data type.
- ▶ Data Processing: In this part process the each column to change the data type or to remove any symbol present in the particular column.
- ▶ EDA: At last we do some exploratory data analysis (EDA) to get some insight into data set and underlying structure of data set on selected features, visualize the data using different plots.

Data Summary

In first data set (**Play store data**) we have 10841 rows and 13 columns.

Following are the 13 columns available in the dataset:

- ▶ Apps: this column contains information about some different apps present in google play store app
- ▶ Category: it contains categories of different apps
- ▶ Rating: rating of apps by users
- ▶ Reviews: responses of app users
- ▶ Size: it contains the size of each app
- ▶ Installs: no. of installed each app
- ▶ Type: in the type we can able to know is our app is free or paid
- ▶ Price: in this column the price of each app is present
- ▶ Content rating: from this column we can able to know the app belongs to which age group people

Data Summary

- ▶ Genres: this consist of genres for each app
- ▶ Last updated: in the column we can see when app was last updated
- ▶ Current version: from this column we can find the app's current version
- ▶ Android version: from this we can find the android version used for the app

In the second data set (**user reviews**) we have 64295 rows and 5 columns.

We have Apps, translated Review, sentiment, sentiment polarity, sentiment subjectivity in this data set.

Data Analysis questions

- ▶ Is there any app which has the rating greater than 5?
- ▶ Is there any correlation between the columns?
- ▶ Are the reviews more than installs as only those who installed can review the app?
- ▶ Is there any Null values, Duplicate files and outliers?
- ▶ What is the distribution of Rating?
- ▶ What are the top Content Rating values? Are there any values with very few records?
- ▶ Which category have a greater number of apps and which category have least number of apps?
- ▶ Find the skew, mean and median of rating.
- ▶ What is the range of rating given to maximum apps?
- ▶ Does the installation number affect all other columns?
- ▶ Which type of apps are installed mostly, either free or paid?
- ▶ What is the maximum price of paid apps on play store and which app/s is/are expensive?

Data Analysis questions

- ▶ Which app under the paid category earn most money?
- ▶ The apps under which age group are installed mostly?
- ▶ Which user sentiment has the maximum frequency?
- ▶ Which apps have the most positive sentiment and which apps have the most negative sentiments?
- ▶ The apps under which category are installed most and apps under which category have least installation number?

Data Cleaning

- ▶ Check for the NULL values.
- ▶ Check for the apps which have rating more than 5.
- ▶ Check for the duplicate files.
- ▶ Check for the apps which have reviews greater than the number of installations.

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
2454	KBA-EZ Health Guide	MEDICAL	5.0	4	25000000.0	1	Free	0.00	Everyone	Medical	August 2, 2018	1.0.72	4.0.3 and up
5917	Ra Ga Ba	GAME	5.0	2	20000000.0	1	Paid	1.49	Everyone	Arcade	February 8, 2017	1.0.4	2.3 and up
6700	Brick Breaker BR	GAME	5.0	7	19000000.0	5	Free	0.00	Everyone	Arcade	July 23, 2018	1.0	4.1 and up
7402	Trovami se ci riesci	GAME	5.0	11	6100000.0	10	Free	0.00	Everyone	Arcade	March 11, 2017	0.1	2.3 and up
8591	DN Blog	SOCIAL	5.0	20	4200000.0	10	Free	0.00	Teen	Social	July 23, 2018	1.0	4.0 and up
10697	Mu.F.O.	GAME	5.0	2	16000000.0	1	Paid	0.99	Everyone	Arcade	March 3, 2017	1.0	2.3 and up

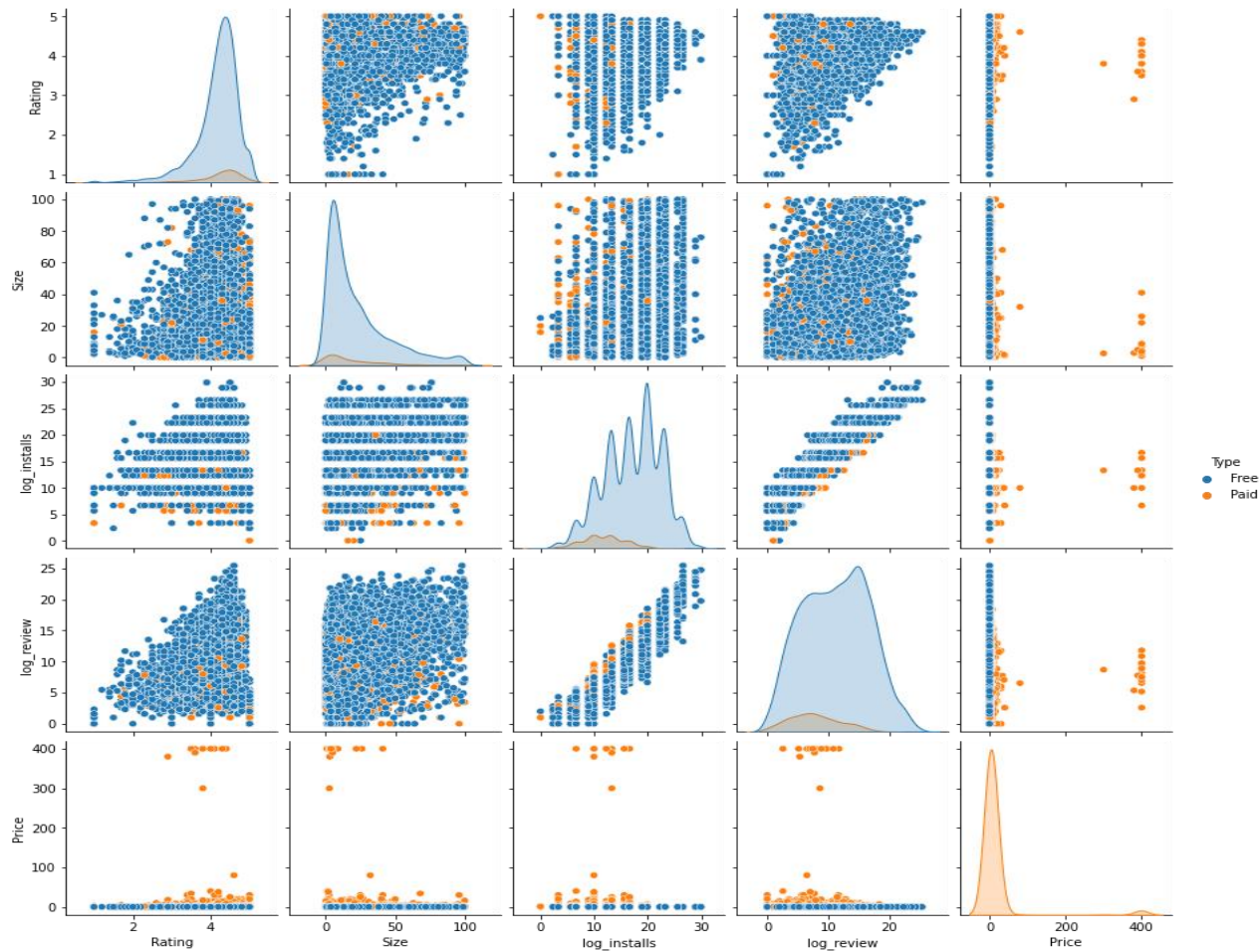
Data Processing

- ▶ Check for the columns which have numeric values but have object data type, like review, size, install and price.
- ▶ Remove the symbols from the numeric columns.
- ▶ Convert the data type from object to either int64 or float64.
- ▶ Convert all the values of a column to same unit.



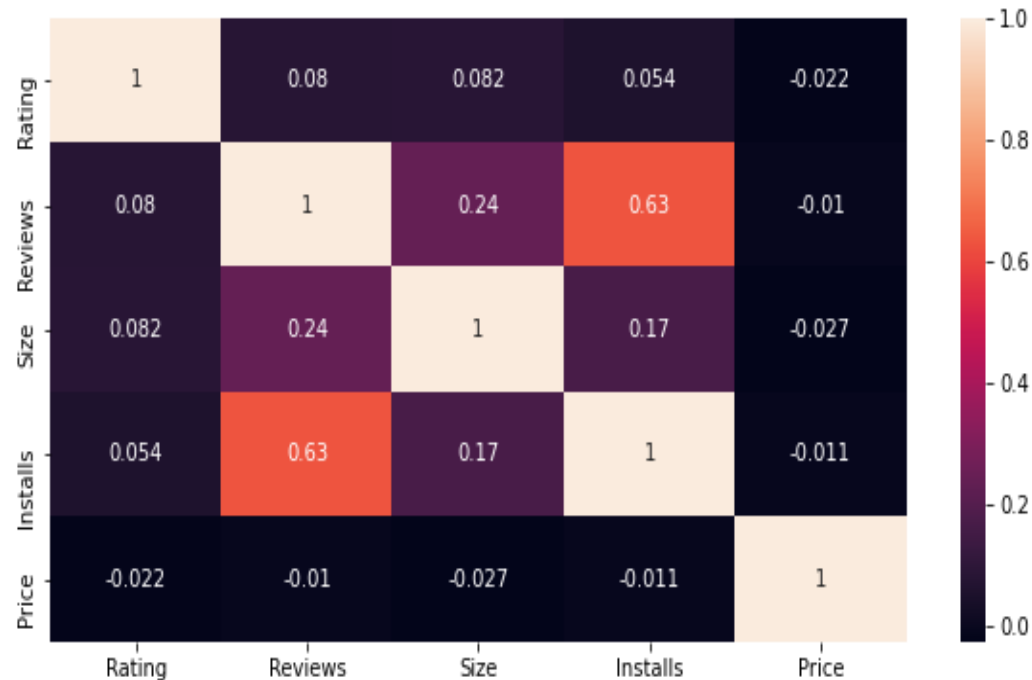
Exploratory Data Analysis (EDA)

Bivariate Analysis



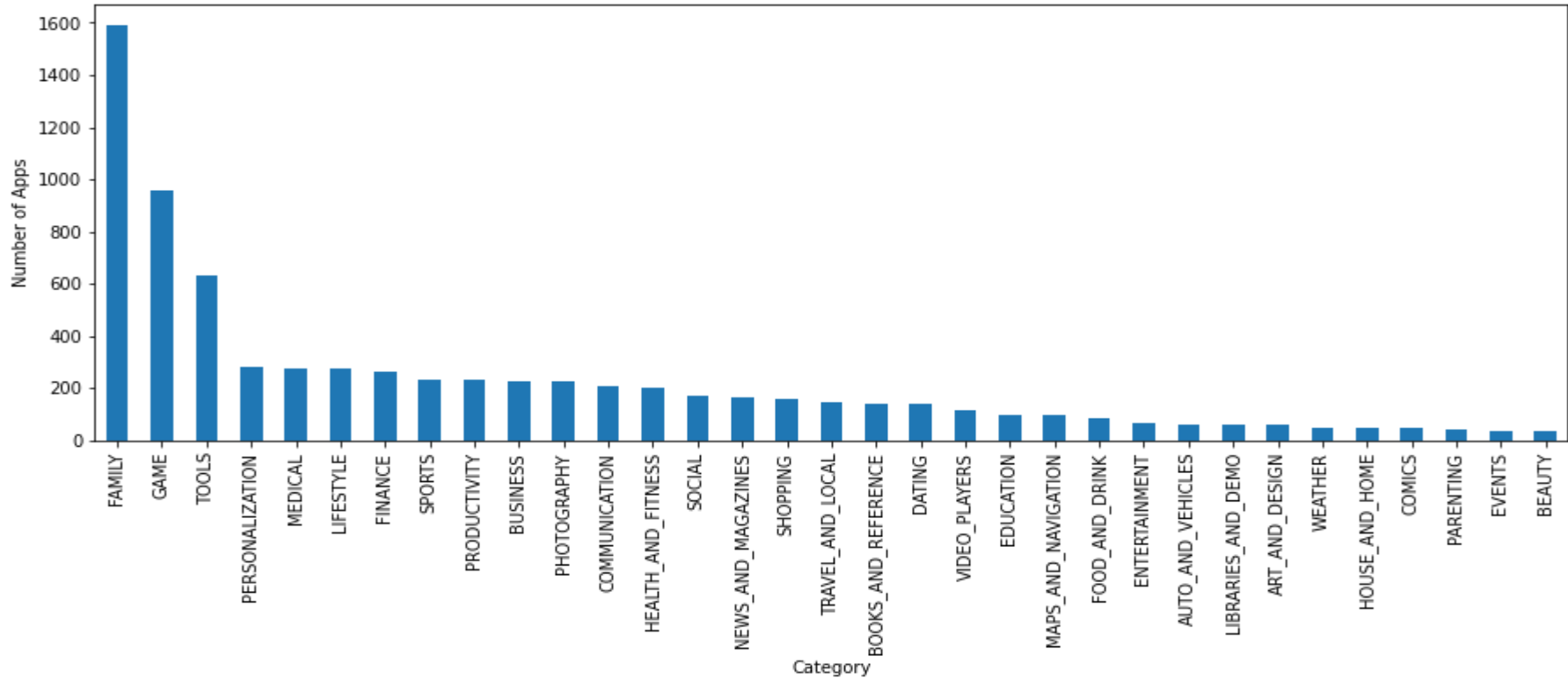
Heatmap for Different Quantities

- ▶ From the heat map we can say that there is maximum cross correlation between the Install and Review.
- ▶ Light color reflect the maximum correlation and dark color reflect the minimum correlation.

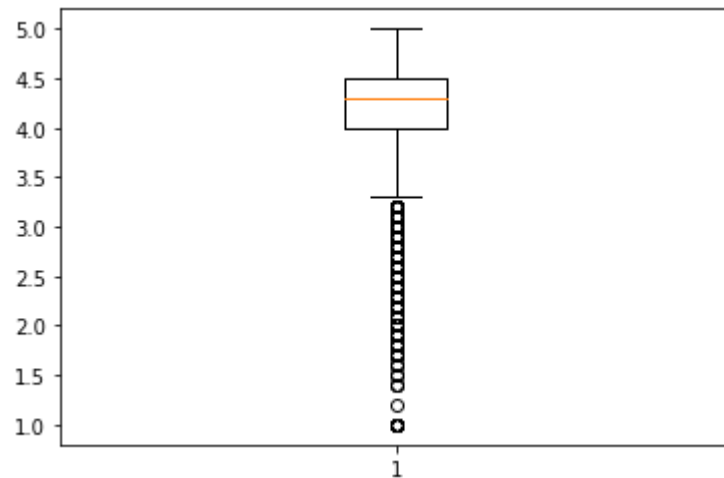
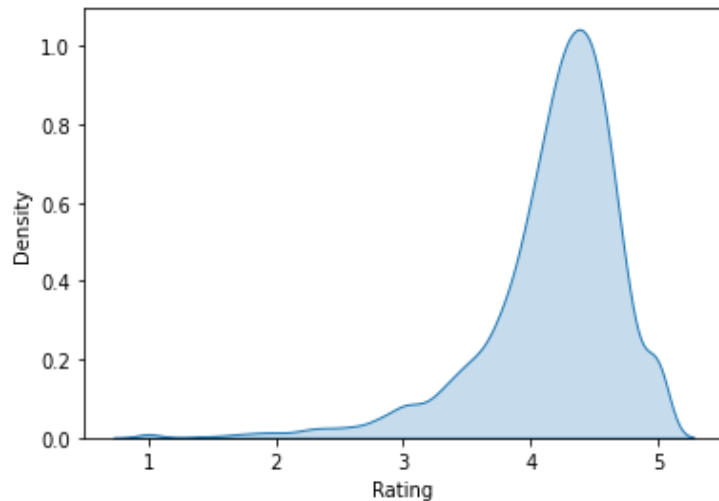


Category wise Number of Apps

Category wise number of apps in Play Store



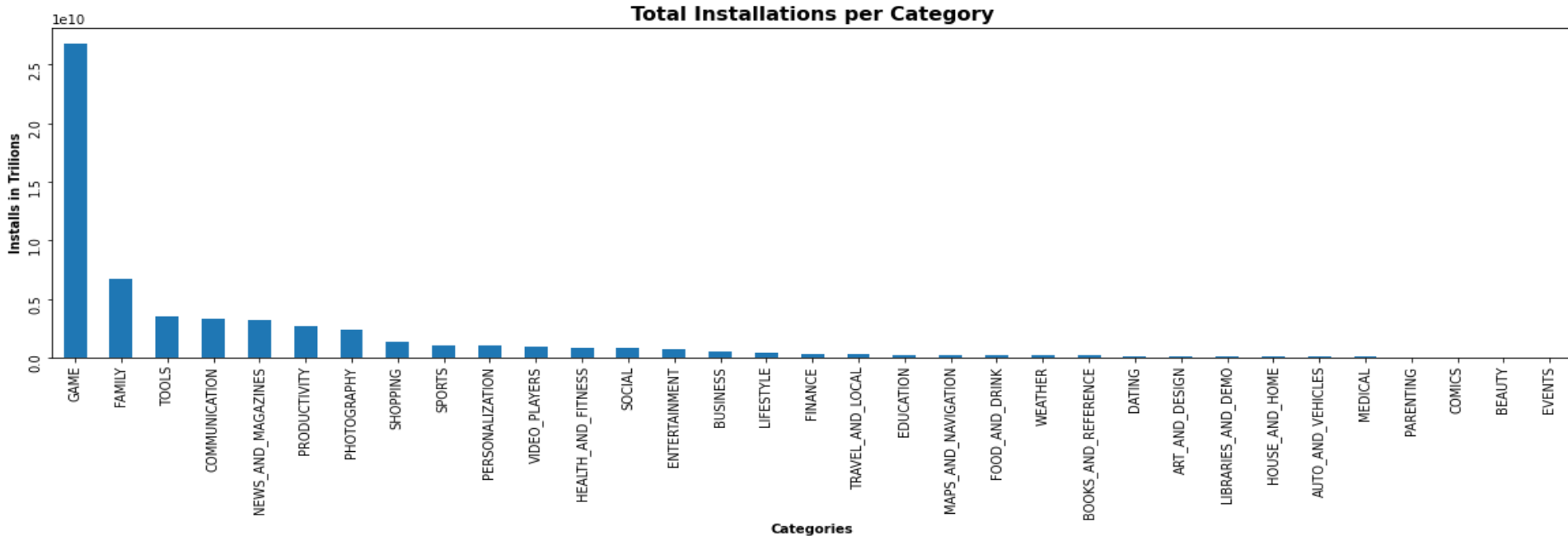
Rating Distribution



- ▶ the skew distribution of ratings -1.733457613883763
- ▶ the mean distribution of ratings 4.171309267241382
- ▶ the median of distribution of ratings 4.3
- ▶ Maximum number of apps have rating between 3.8 to 4.5

Categories in Demand

- ▶ The following bar graph shows the information of different category being installed.
- ▶ Gaming category have the maximum no of installation/downloads compared with other category.



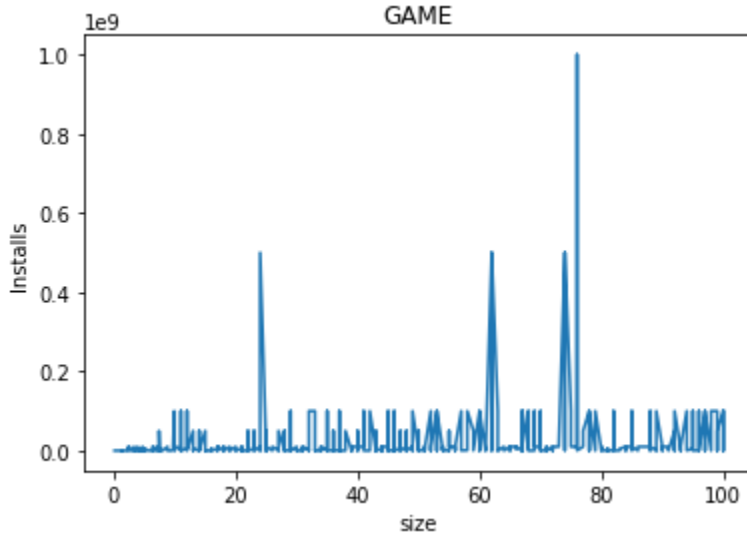
Installation Strategy

Let's consider three top categories in demand:

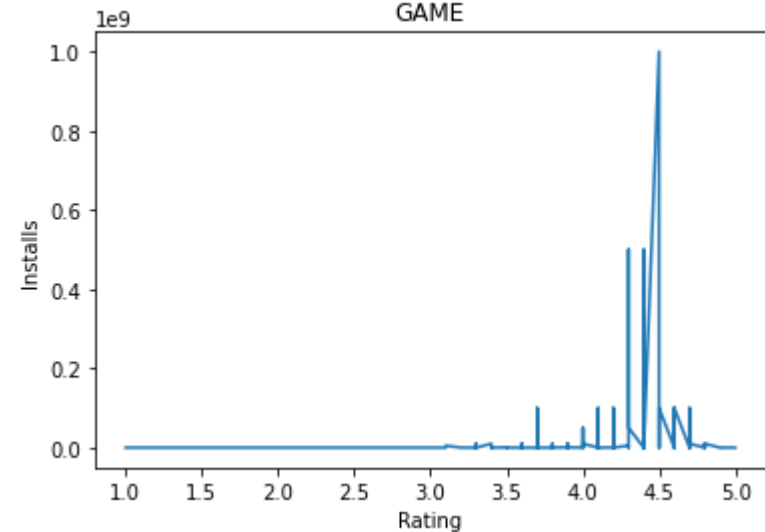
1. Game
2. Family
3. Communication

Game

- ▶ Game apps are available for every size
- ▶ And the most installed games have the file size approx. 60-80MB

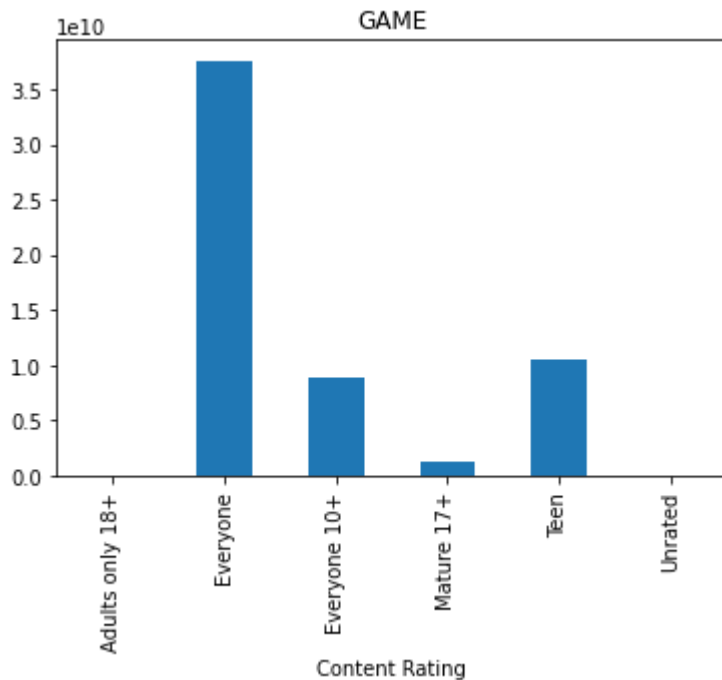


- ▶ Mostly gaming category have the rating between 4.2 - 4.5.
- ▶ And the app with 4.5 rating have the most installed files

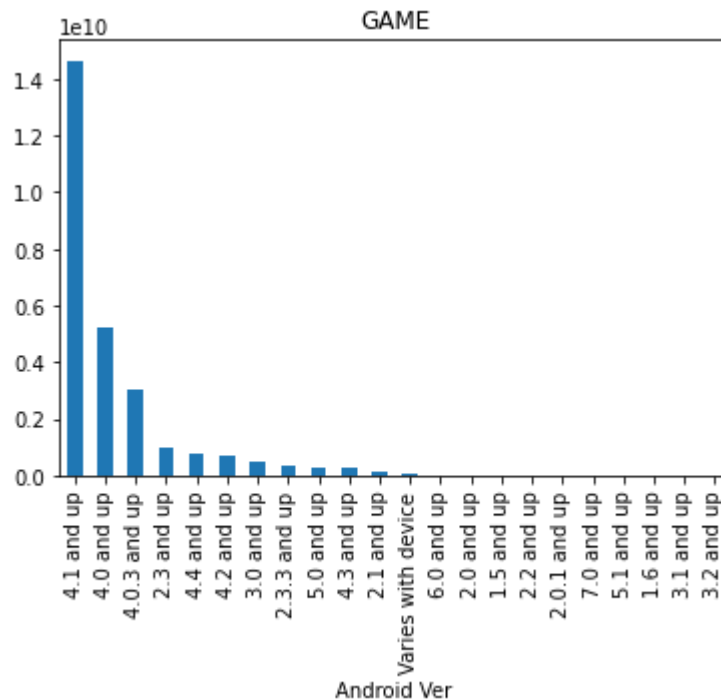


Game

- ▶ Game apps are used by almost every age group.

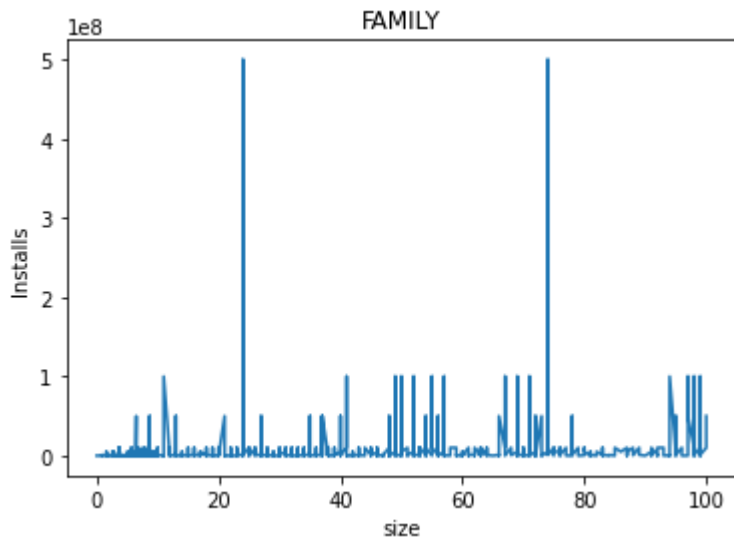


- ▶ In the game category the most installed android version is 4.1 and above.

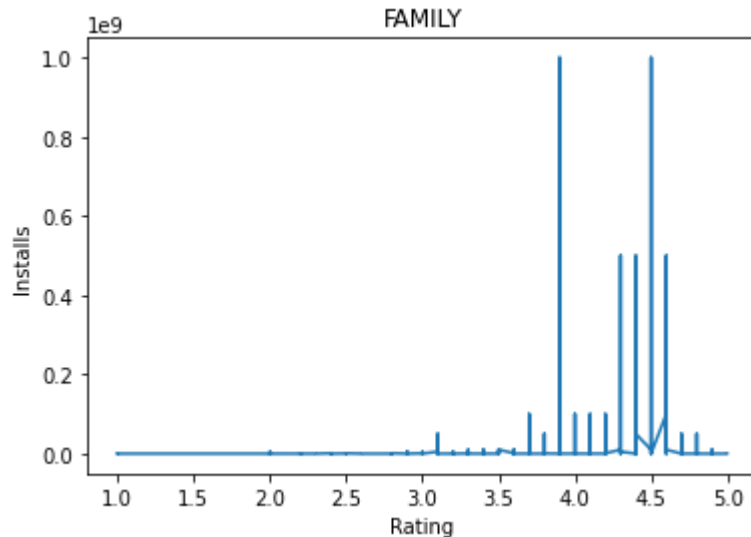


Family

- ▶ Similar to games category, family category is also mostly installed in all file sizes.
- ▶ The common app which are downloaded have the file size between 20-30MB & 70-80MB.

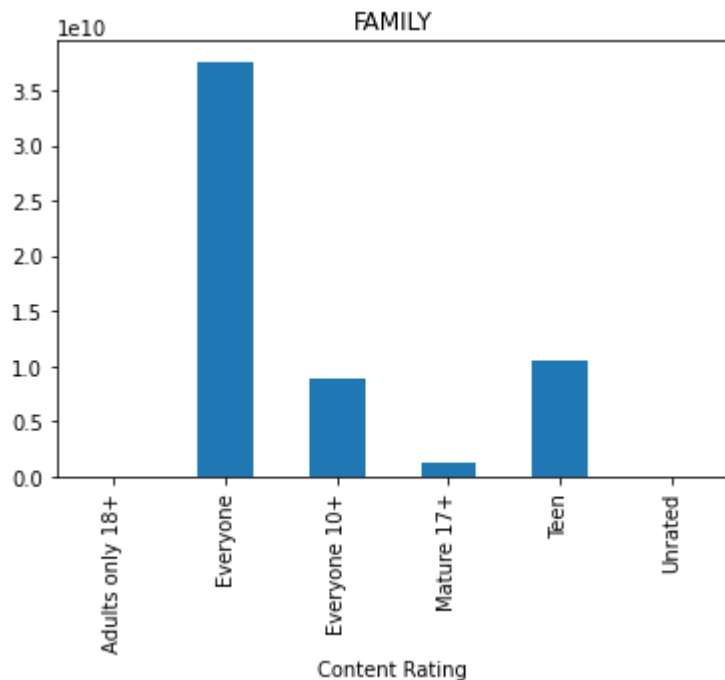


- ▶ Mostly family category have apps rating between 3.9 - 4.7.
- ▶ And the app with 4.5 rating have the most installed files.

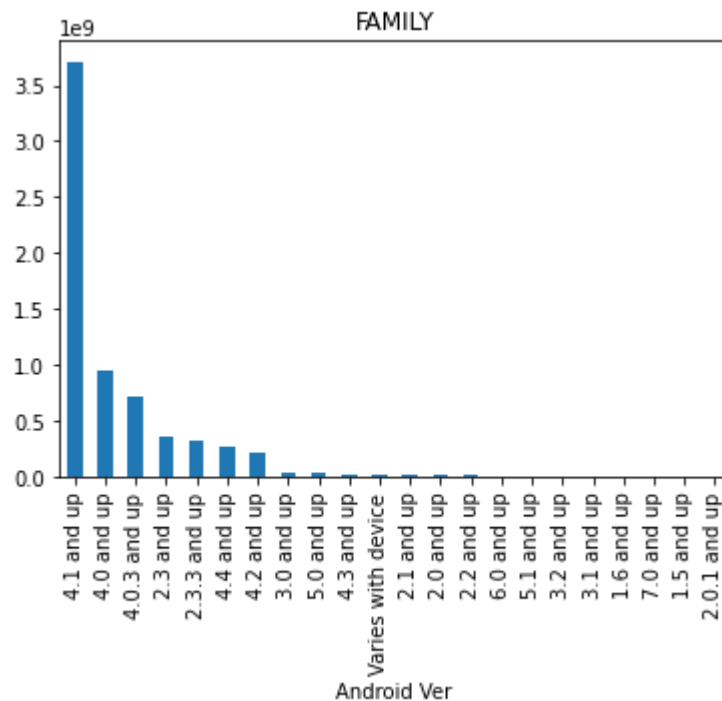


Family

- Family apps are used by almost every age group but the age group of 17+ is least downloaded.

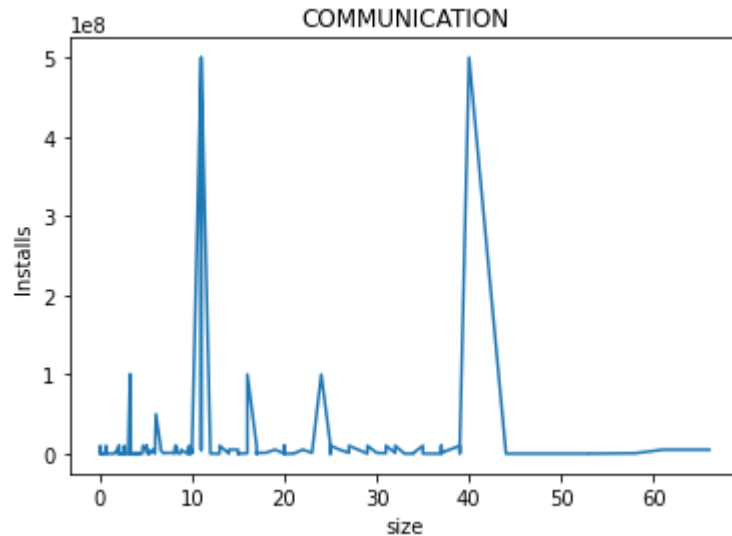


- In the Family category the most installed android version is 4.1 and above.

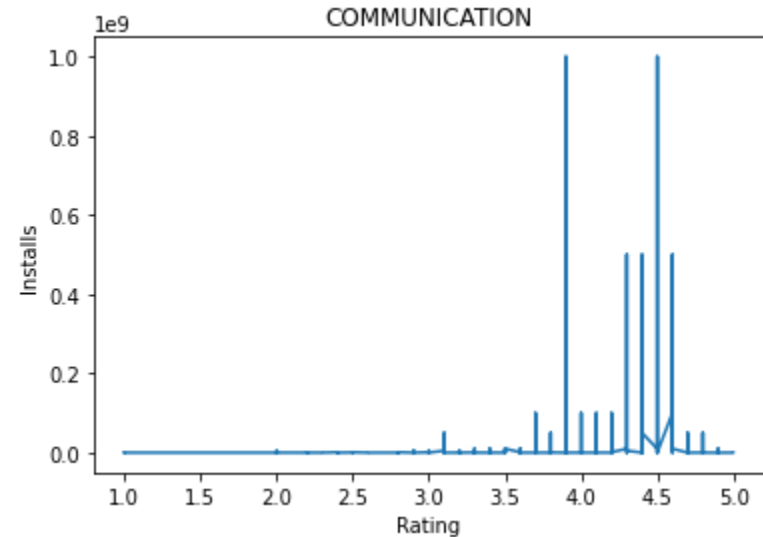


Communication

- ▶ Apps in the communication category which have the size of 10M & 40M are mostly downloaded

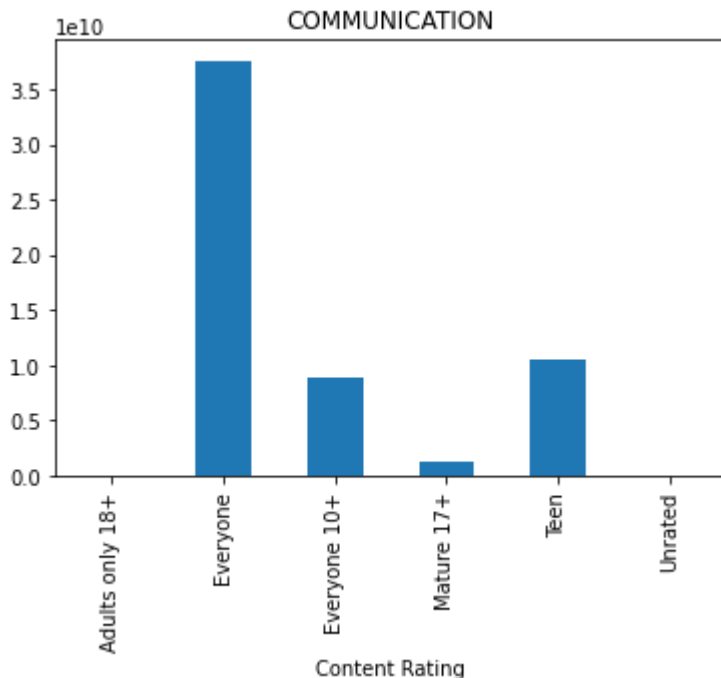


- ▶ Mostly communication category have apps rating between 4.0-4.6.
- ▶ And the app with rating 4.0 and 4.5 are the most installed apps.

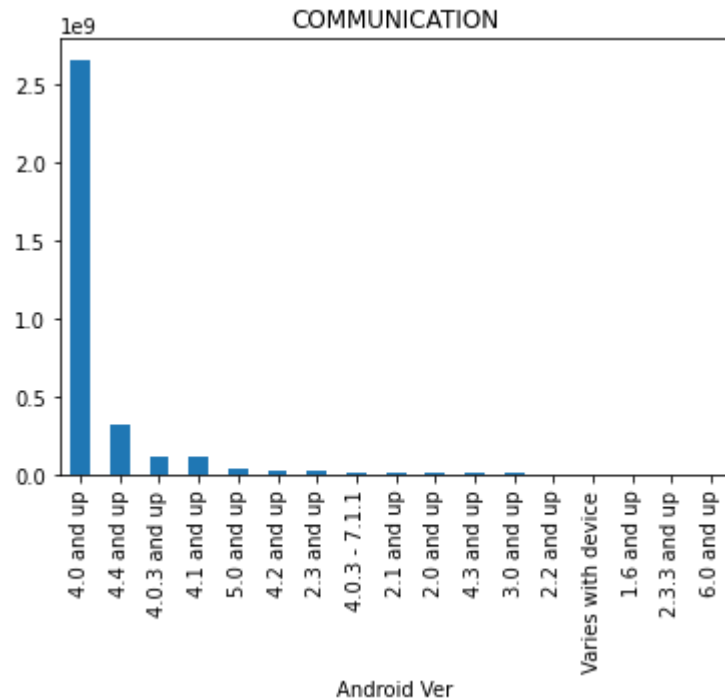


Communication

- Apps in communication category are used by almost every age group.

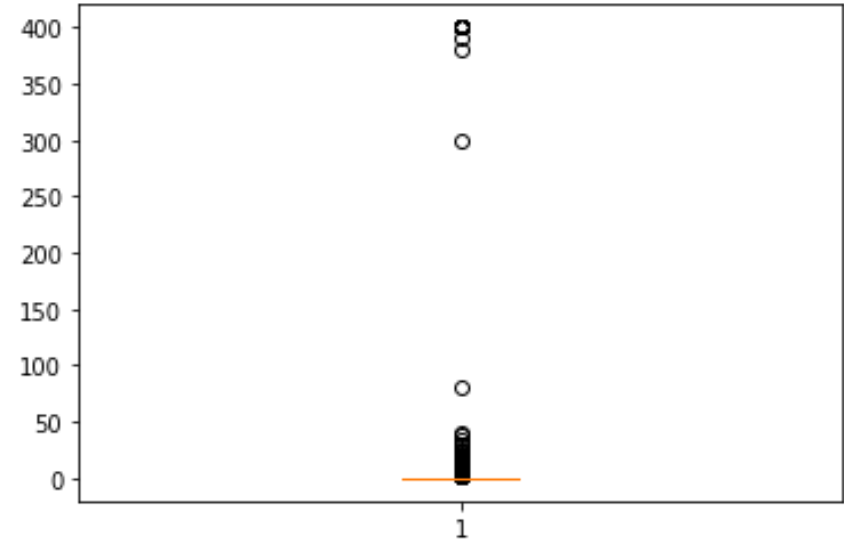


- Communication category apps are used on android version 4.0 and up.




Price

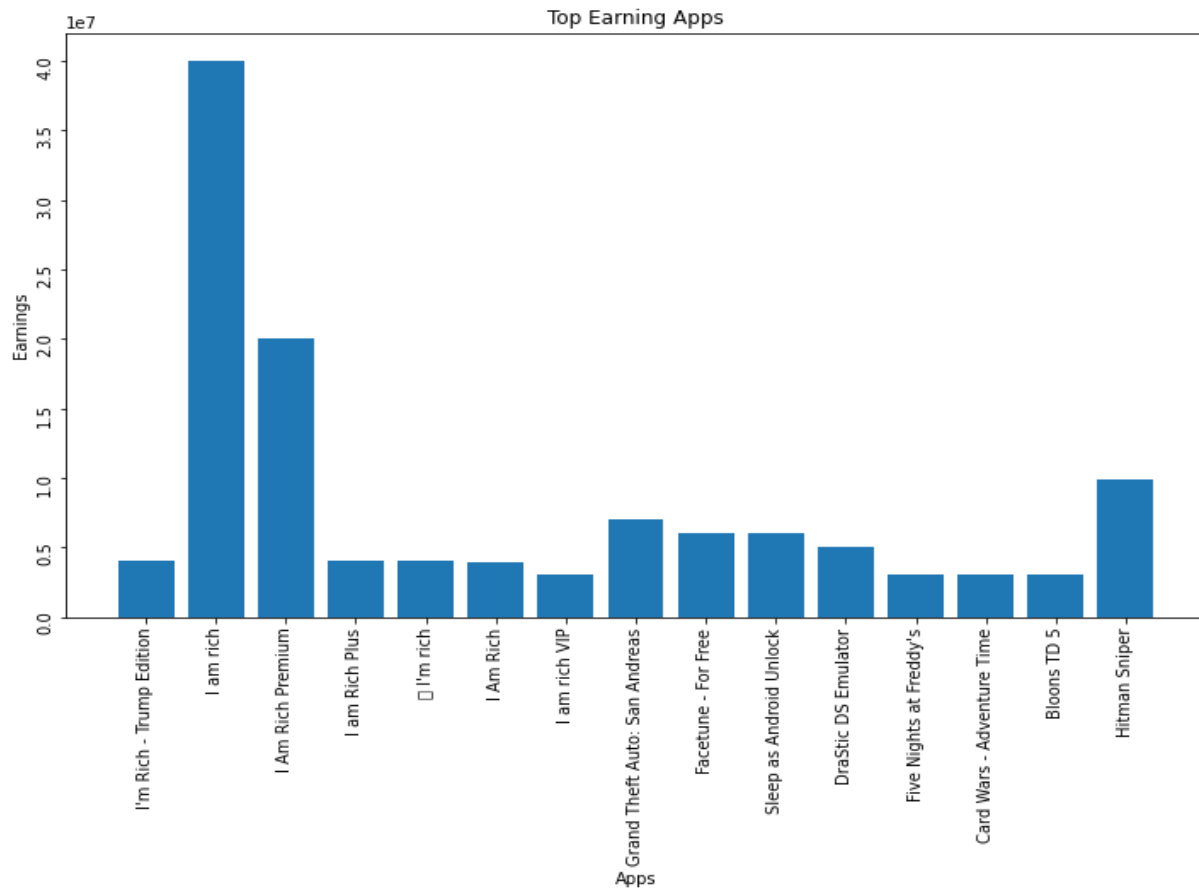
- ▶ From the boxplot we can check that most apps price range is in between 0 to 50\$.
- ▶ Some apps also have price more than 250\$.
- ▶ Lets check the apps which has price more than 250\$.



Most Expensive Apps

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
4197	most expensive app (H)	FAMILY	4.3	6	1.500	100	Paid	399.99	Everyone	Entertainment	July 16, 2018	1.0	7.0 and up
4362	 I'm rich	LIFESTYLE	3.8	718	26.000	10000	Paid	399.99	Everyone	Lifestyle	March 11, 2018	1.0.0	4.4 and up
4367	I'm Rich - Trump Edition	LIFESTYLE	3.6	275	7.300	10000	Paid	400.00	Everyone	Lifestyle	May 3, 2018	1.0.1	4.1 and up
5351	I am rich	LIFESTYLE	3.8	3547	1.800	100000	Paid	399.99	Everyone	Lifestyle	January 12, 2018	2.0	4.0.3 and up
5354	I am Rich Plus	FAMILY	4.0	856	8.700	10000	Paid	399.99	Everyone	Entertainment	May 19, 2018	3.0	4.4 and up
5355	I am rich VIP	LIFESTYLE	3.8	411	2.600	10000	Paid	299.99	Everyone	Lifestyle	July 21, 2018	1.1.1	4.3 and up
5356	I Am Rich Premium	FINANCE	4.1	1867	4.700	50000	Paid	399.99	Everyone	Finance	November 12, 2017	1.6	4.0 and up
5357	I am extremely Rich	LIFESTYLE	2.9	41	2.900	1000	Paid	379.99	Everyone	Lifestyle	July 1, 2018	1.0	4.0 and up
5358	I am Rich!	FINANCE	3.8	93	22.000	1000	Paid	399.99	Everyone	Finance	December 11, 2017	1.0	4.1 and up
5359	I am rich(premium)	FINANCE	3.5	472	0.965	5000	Paid	399.99	Everyone	Finance	May 1, 2017	3.4	4.4 and up
5362	I Am Rich Pro	FAMILY	4.4	201	2.700	5000	Paid	399.99	Everyone	Entertainment	May 30, 2017	1.54	1.6 and up
5364	I am rich (Most expensive app)	FINANCE	4.1	129	2.700	1000	Paid	399.99	Teen	Finance	December 6, 2017	2	4.0.3 and up
5366	I Am Rich	FAMILY	3.6	217	4.900	10000	Paid	389.99	Everyone	Entertainment	June 22, 2018	1.5	4.2 and up
5369	I am Rich	FINANCE	4.3	180	3.800	5000	Paid	399.99	Everyone	Finance	March 22, 2018	1.0	4.2 and up
5373	I AM RICH PRO PLUS	FINANCE	4.0	36	41.000	1000	Paid	399.99	Everyone	Finance	June 25, 2018	1.0.2	4.1 and up

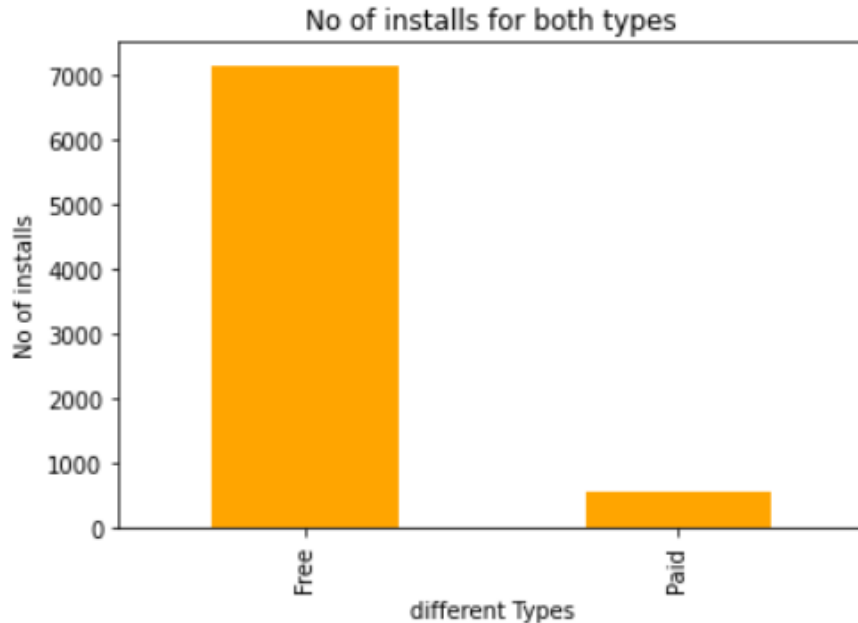
Highest Earning Apps



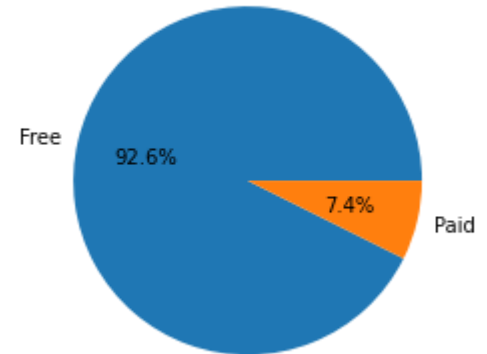
▶ “I am rich” is the highest earning app in the play store.

Different Types & Ratio

- ▶ Free apps have more installs than paid app.
- ▶ 92.6% apps are free in play store, only 7.4% apps are paid.

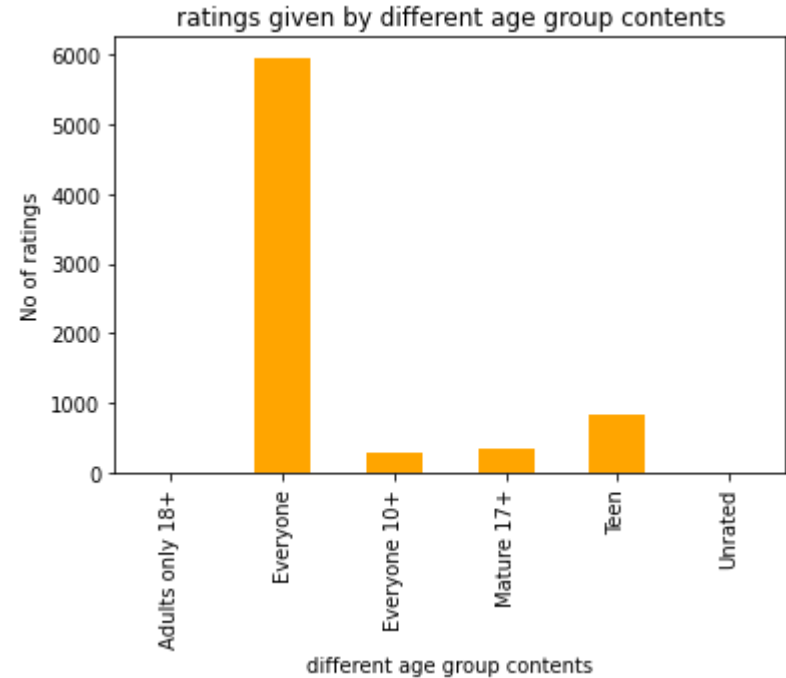


Percent of Free Vs Paid Apps in play store

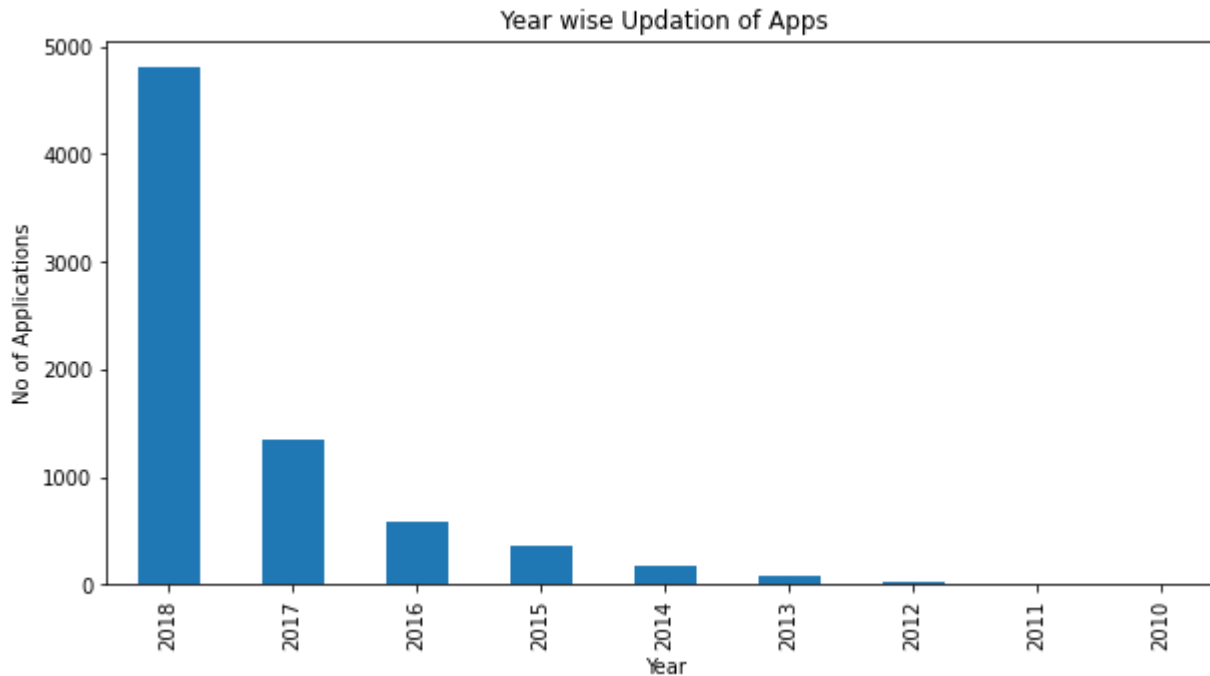


Rating by different age group

- ▶ Bar graph shows different age group contents vs the rating.
- ▶ From the different group contents apps which are available for everyone has maximum ratings and content with 18+ and unrated have least rating.



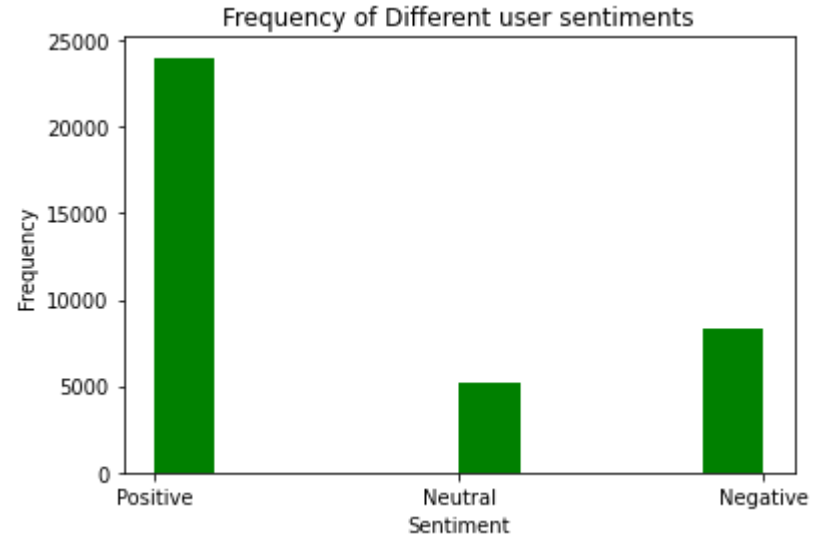
Last Update



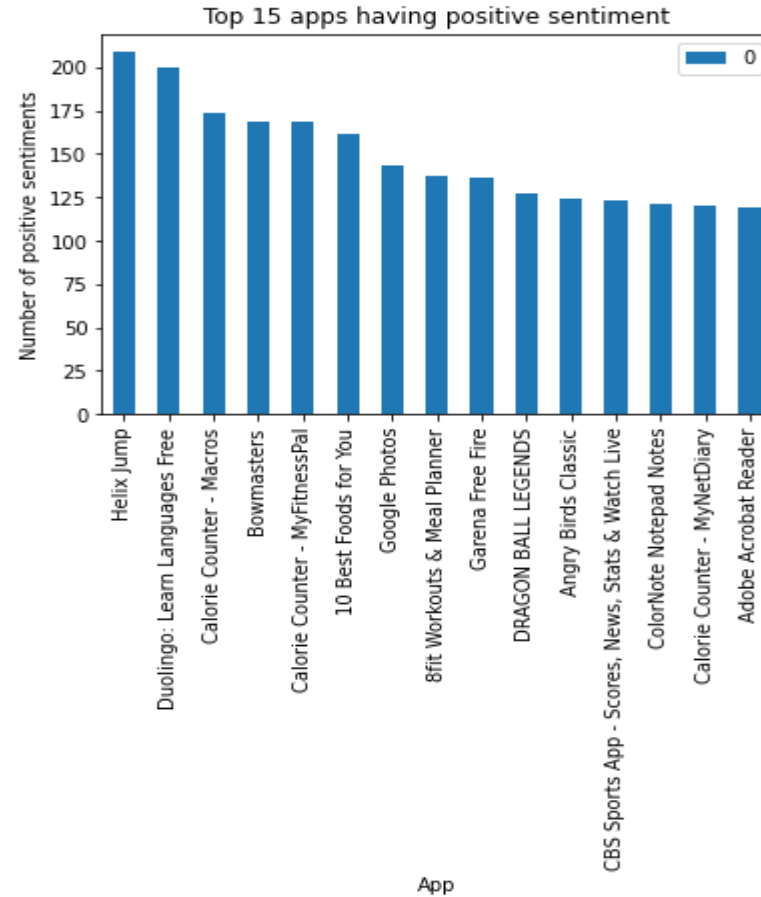
- ▶ Mostly apps are last updated in 2018.

Sentiment Frequency

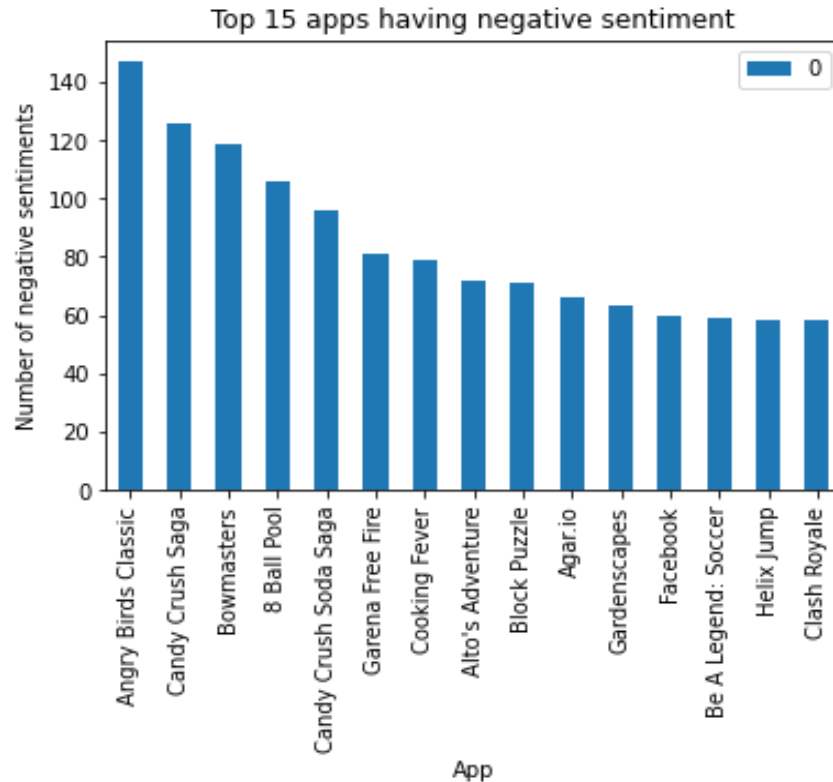
- ▶ From the second dataset we found that we have three types of sentiments.
- ▶ Among three we have most app with positive sentiments



Top 15 Apps having Positive Sentiment



Top 15 Apps having Negative Sentiment



Observation/ Conclusion

For the apps to be popular and mostly downloadable, a developer should focus on:

- ▶ The app should be free. For no ads application, the price of the app should be less than 10\$.
- ▶ Paid app should be designed in small size and to meet the user expectation.
- ▶ The size of the app should be as small as possible, preferably between 2MB – 50MB.
- ▶ There is a positive correlation between installs and review.
- ▶ The Game category have a good potential for developing an app, because this is the demanding category.
- ▶ Apps which are available for everyone are most installed apps.
- ▶ Users prefer the apps which are compatible with android version 4.1 and above.