REPORT

EXPLORATORY ANALYSIS ON

Google Play Store Apps

DATA SET LINK :	**https://www.kaggle.com/lava18/google-play-store-apps**

ABSTRACT

The assignment is meant to do analysis on a dataset chosen from the kaggle website and provide valid deductions and insights based on the data description and data visualisations.The dataset chosen consists of data from the official Google Play Store App. It consists of several variables which can be further correlated which can help provide deep insights into how the Play Store manages and sorts data based on Ratings, Reviews, Popularity and many more. The data will be cleaned off the garbage values and will be replaced by the meaningful ones. The outliers will be properly scrutinised. The data will be made available as visual representations for easy analysis and conclusion will be reached.

Data Set

The dataset was chosen from the kaggle website **www.kaggle.com**.

The dataset contains all the details of the applications on Google Play. There are 13 features that describe a given app.

Purpose:

The Google Play Store apps dataset 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!

Variables and their meanings:

App	Application name
Category	Category the app belongs to

Rating	Overall user rating of the app (as when scraped)
Reviews	Number of user reviews for the app (as when scraped)
Size	Size of the app (as when scraped)
Installs	Number of user downloads/installs for the app (as when scraped)
Type	Paid or Free
Price	Price of the app (as when scraped)
Content Rating	Age group the app is targeted at - Children / Mature 21+ / Adult
Genres	An app can belong to multiple genres (apart from its main category). For eg, a musical family game will belong to Music, Game, Family genres.
Last Updated	Date when the app was last updated on Play Store (as when scraped)
Current Ver	Current version of the app available on Play Store (as when scraped)
Android Ver	Min required Android version (as when scraped)

Introduction

The dataset chosen contains data from the official Google Play Store which will help us analyse how the Play store manages and presents apps based on Ratings, Reviews , Popularity and many more.

Choosing and analysing the dataset:

In this decade, almost everybody owns an android phone and hence arrives the need of Apps in our day to day life. Therefore,there's a need to understand the different aspects of the one that fulfills your needs.
Google Play Store has about 3.8 Million Apps in comparison to Apple App Stores 2M. You can now easily understand the popularity of the Play Store. It has an App for almost everything you can imagine and hence it makes Play Store quite an amazing dataset to work on.
Though this dataset merely contains information about 0.28 percent(10841 out of 3.8 Million) of the Total Applications available on Play Store.
The public datasets (on Kaggle and the like) provide Apple App Store data, there are not many counterpart datasets available for Google Play Store apps anywhere on the web.
On digging deeper, we found out that iTunes App Store page deploys a nicely indexed appendix-like structure to allow for simple and easy web scraping.

On the other hand, Google Play Store uses sophisticated modern-day techniques (like dynamic page load) using JQuery making scraping more challenging.

Through this report, we will try to answer questions and draw inference.

- What % of Apps are available on Google Play for different age groups?
- Which are the most popular App categories?
- What is the general trend of people based on content rating?
- Which App Genres are most Popular?
- How does type(Free/Paid) varies based on app content rating?
- What has been the past trend in the number of Apps updated in each year?
- How does rating varies based on Type(Free/Paid) of App?
- How does the Type(Free/Paid) vary for top 3 categories ?
- Variation of Price of Apps and Rating ?
- Does size of App matter ?
- What role do ratings play in the market ?

Processing (Data Cleaning)

Data cleansing or data cleaning is the process of detecting and correcting corrupt or inaccurate records from a record set, table, or database and refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying, or deleting the dirty or coarse.

- The **Installs column** consists of values in the form of object as well as it also contains special characters like '+' and ',' which are not suitable for manipulation hence using *strip()* and *replace()* functions we got rid of those objects and convert the values to numeric which are easy to work with. There's a 'Free' value which was cleaned and replaced with nan(np.nan).The outliers are of relevance because there apps which get downloaded several times, even multiple times by the same user.

```
1,000,000+        1579
10,000,000+       1252
100,000+          1169
10,000+           1054
1,000+             907
5,000,000+         752
100+               719
500,000+           539
50,000+            479
5,000+             477
100,000,000+       409
10+                386
500+               330
50,000,000+        289
50+                205
5+                  82
500,000,000+        72
1+                  67
1,000,000,000+      58
0+                  14
Free                 1
0                    1
Name: Installs, dtype: int64
```

Converted_to

```
1000000       1579
10000000      1252
100000        1169
10000         1054
1000           907
5000000        752
100            719
500000         539
50000          479
5000           477
100000000      409
10             386
500            330
50000000       289
50             205
5               82
500000000       72
1               67
1000000000      58
0               15
Name: Installs, dtype: int64
```

- The **Review column** consists of values terminating with 'M' which were cleaned using *split() function* and the object(string) were transformed to float and stored back in the dataframe.

  df['Reviews'] = [float(i.split('M')[0]) if 'M' in i else float(i) for i in df['Reviews']]

- For the **Size column,** it can be seen that data has metric prefixes (Kilo and Mega) along with another string. Replacing 'k' and 'M' with their values to convert values to numeric. Though the data in 'Kbs' was changed to 'Mbs' for uniformity. [1Mb=1024Kb] .

  df["Size"] = [float(i.split('M')[0]) if 'M' in i else float(0) for i in df["Size"]]

```
Varies with device    1695
11M                    198
12M                    196
14M                    194
13M                    191
Name: Size, dtype: int64
```

- For the **Price column** data contained '$' sign which was removed and the values were converted to numeric for easy analysis.

```
array(['0', '$4.99', '$3.99', '$6.99', '$1.49', '$2.99', '$7.99', '$5.99',
       '$3.49', '$1.99', '$9.99', '$7.49', '$0.99', '$9.00', '$5.49',
       '$10.00', '$24.99', '$11.99', '$79.99', '$16.99', '$14.99',
       '$1.00', '$29.99', '$12.99', '$2.49', '$10.99', '$1.50', '$19.99',
       '$15.99', '$33.99', '$74.99', '$39.99', '$3.95', '$4.49', '$1.70',
       '$8.99', '$2.00', '$3.88', '$25.99', '$399.99', '$17.99',
       '$400.00', '$3.02', '$1.76', '$4.84', '$4.77', '$1.61', '$2.50',
       '$1.59', '$6.49', '$1.29', '$5.00', '$13.99', '$299.99', '$379.99',
       '$37.99', '$18.99', '$389.99', '$19.90', '$8.49', '$1.75',
       '$14.00', '$4.85', '$46.99', '$109.99', '$154.99', '$3.08',
       '$2.59', '$4.80', '$1.96', '$19.40', '$3.90', '$4.59', '$15.46',
       '$3.04', '$4.29', '$2.60', '$3.28', '$4.60', '$28.99', '$2.95',
       '$2.90', '$1.97', '$200.00', '$89.99', '$2.56', '$30.99', '$3.61',
       '$394.99', '$1.26', '$1.20', '$1.04'], dtype=object)
```

df['Price'] = [float(i.split('$')[1]) if '$' in i else float(0) for i in df['Price']]

- The **Type column** had no redundant data hence did not require any cleaning.

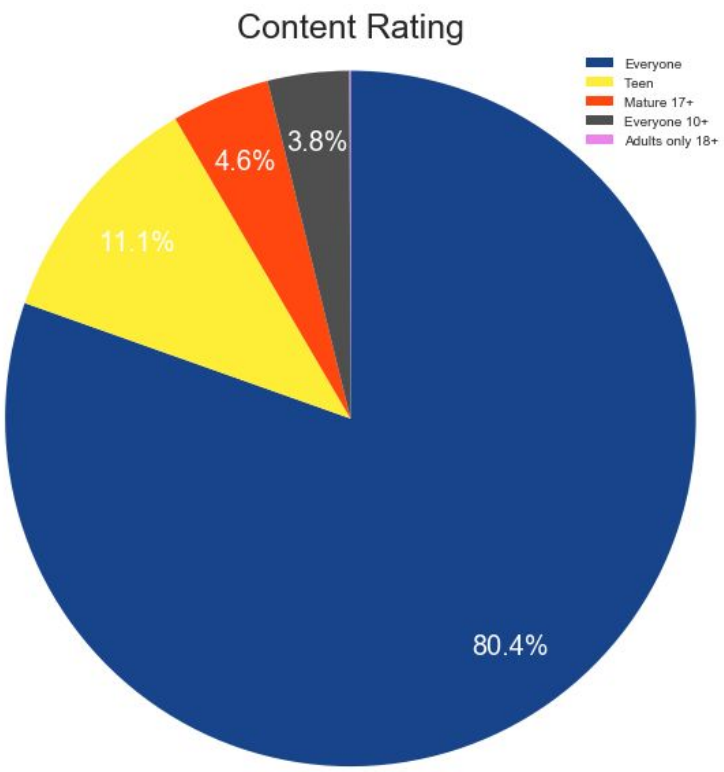
```
Free    10039
Paid      800
Name: Type, dtype: int64
```

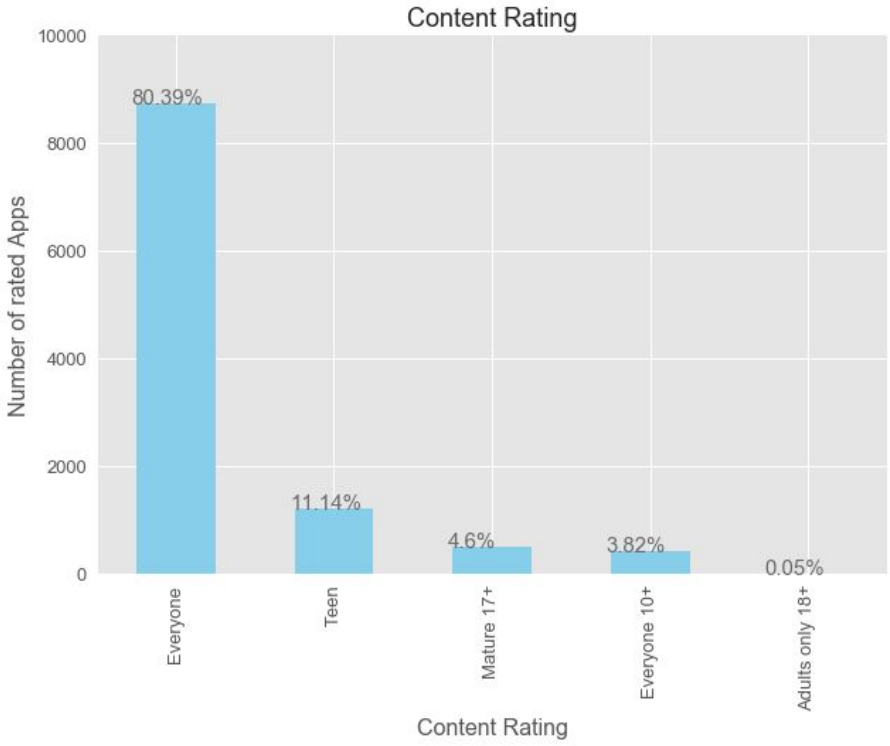
- The **Category column** had no redundant or duplicate data. No abnormalities found.
- The **Rating column** was analyses and was supposed to have values within the range 1-5 which was achieved . The values had no abnormalities and were of the type numeric(float). Yet the outliers were handled using a simple code : `df = df[df["Rating"]<=5].`

- The **Content Rating column** was analysed and we ended up deducing that the column had NAN values which were replaced with the neighbouring values.

EXPLORATORY ANALYSIS

The following scripts of the document contains **Descriptive Analysis** and **Data Visualisation** of the Google Play Store dataset along with *Answers to the Questions related to the dataset and Inferences drawn which have been dealt together rather than segregating them.*

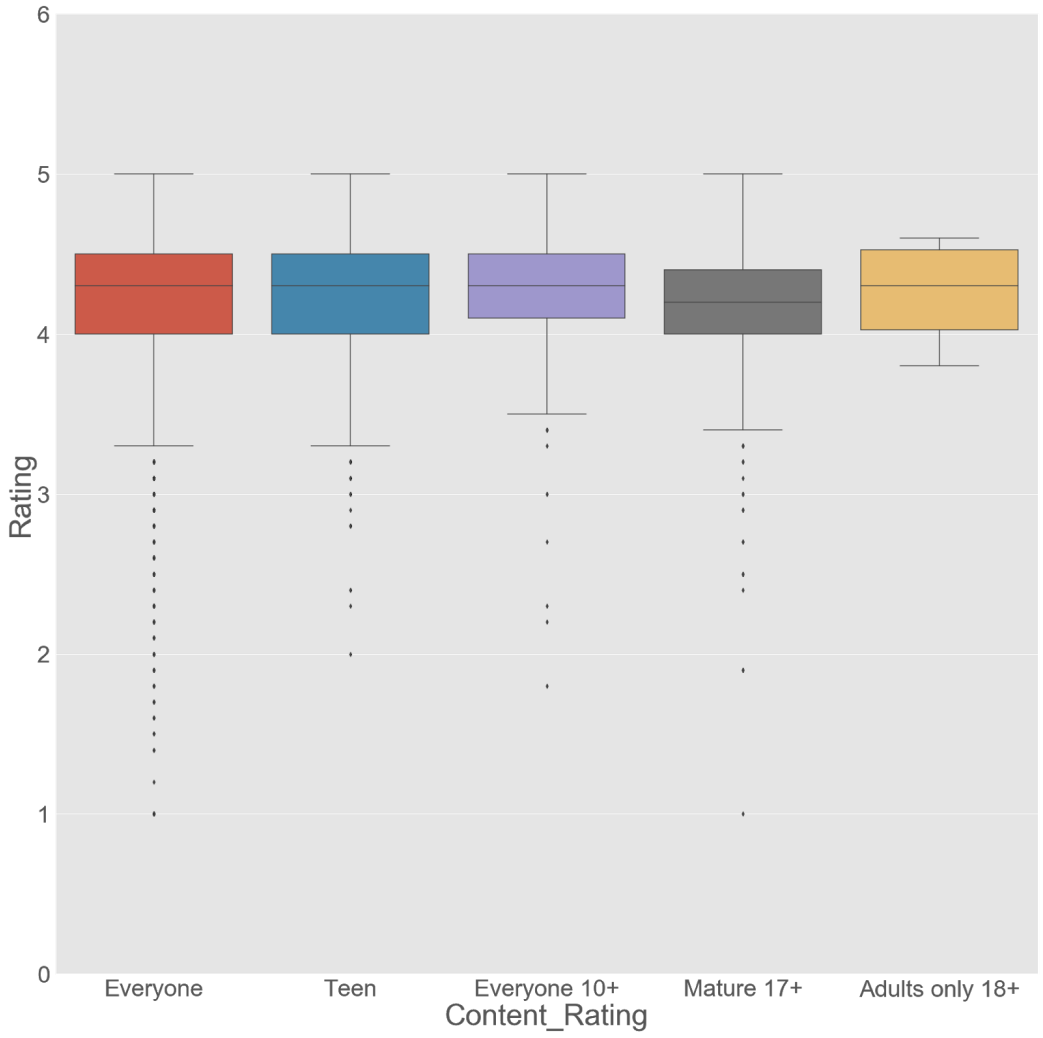




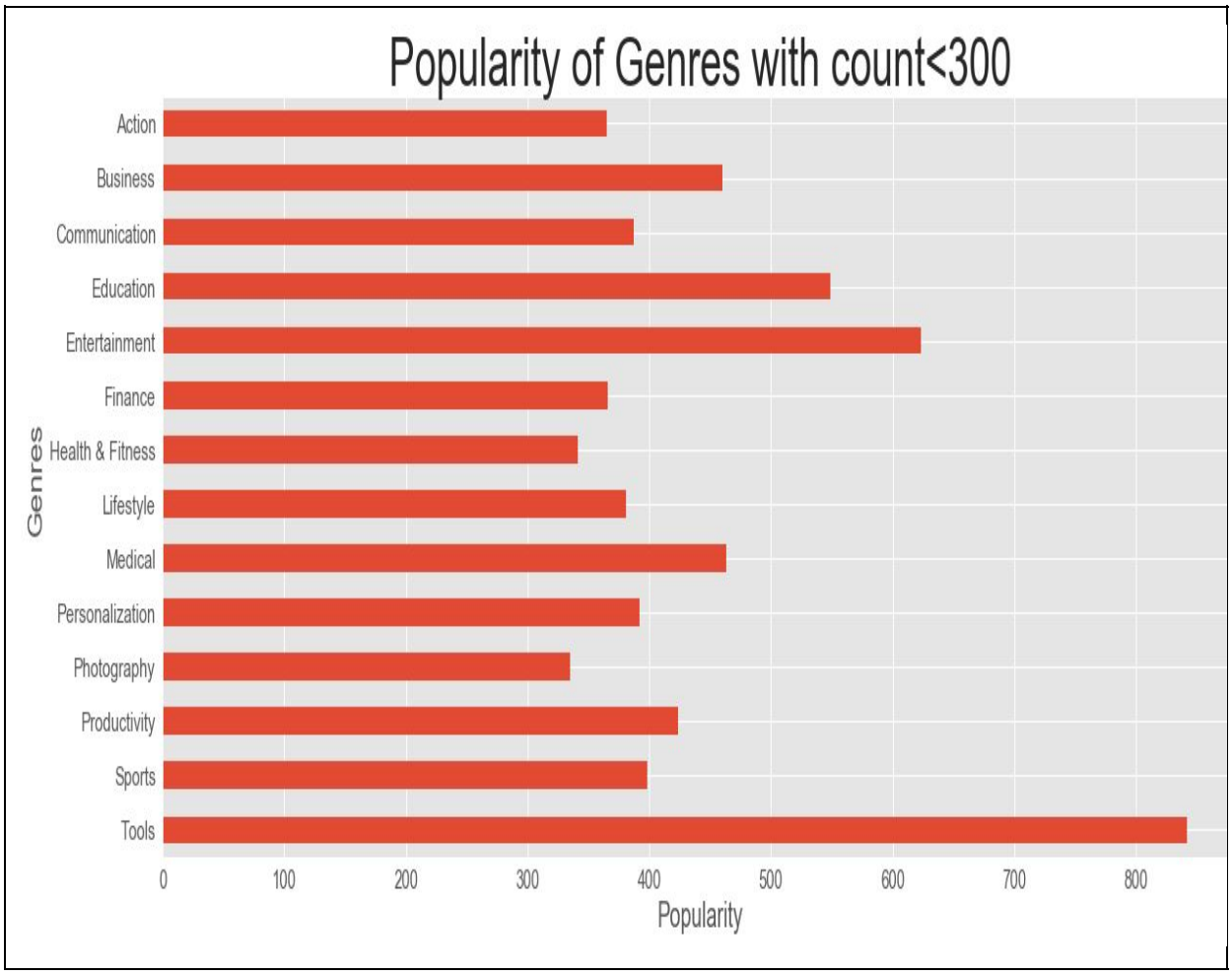
What % of Apps are available on Google Play for different age groups?

The rated Apps present in the dataset were analysed to see the number of apps on the store and

which AGE groups the apps are targeted at - Children / Mature 21+ / Adult. We come to deduce that ~80.4% of the apps are meant for all users irrespective of their age. Next ~11.14% apps are meant for teenagers. Other 4.6% of apps are meant for 17+ age group and ~0.03% are meant for 18+ age groups. Hence,we come to a conclusion that most of the apps are GENERAL PURPOSE and therefore, meant for Everyone.



As per the visualisation, it is safe to say that apps that are meant for all age groups have a higher range of ratings and therefore have been more reviewed. Trend of outliers tell us that different versions of the same apps must be present which are not so popular and hence have been rated low. The median ratings for apps of all content ratings fall in the same range i.e 4 - 4.5 which tells us that good apps of all content are present on Google Play.


Popularity of Genres with count<300
Action
Business
Communication
Education
Entertainment
Finance
Health & Fitness
Lifestyle
Medical
Personalization
Photography
Productivity
Sports
Tools
Genres
0
100
200
300
400
500
600
700
800
Popularity

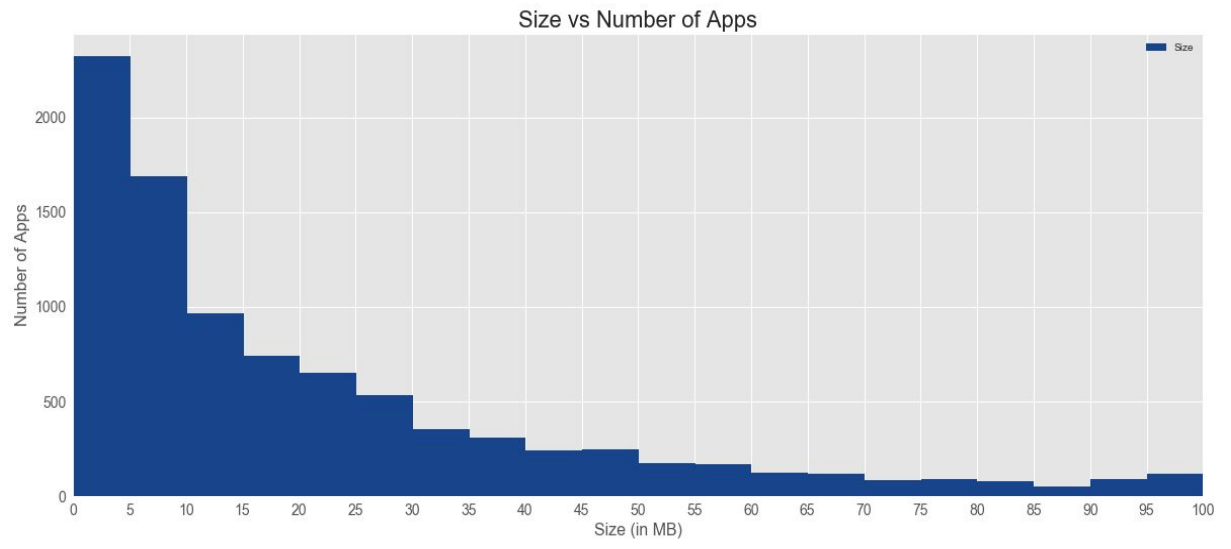

Popularity of Genres
Genres
Word
Weather
Video Players & Editors;Music & Video
Video Players & Editors;Creativity
Video Players & Editors
Trivia;Education
Trivia
Travel & Local;Action & Adventure
Travel & Local
Tools;Education
Tools
Strategy;Education
Strategy;Creativity
Strategy;Action & Adventure
Strategy
Sports;Action & Adventure
Sports
Social
Simulation;Pretend Play
Simulation;Education
Simulation;Action & Adventure
Simulation
Shopping
Role Playing;Pretend Play
Role Playing;Education
Role Playing;Brain Games
Role Playing;Action & Adventure
Role Playing
Racing;Pretend Play
Racing;Action & Adventure
Racing
Puzzle;Education
Puzzle;Creativity
Puzzle;Brain Games
Puzzle;Action & Adventure
Puzzle
Productivity
Photography
Personalization
Parenting;Music & Video
Parenting;Education
Parenting;Brain Games
Parenting
News & Magazines
Music;Music & Video
Music & Audio;Music & Video
Music
Medical
Maps & Navigation
Lifestyle;Pretend Play
Lifestyle;Education
Lifestyle
Libraries & Demo
House & Home
Health & Fitness;Education
Health & Fitness;Action & Adventure
Health & Fitness
Food & Drink
Finance
February 11, 2018
Events
Entertainment;Pretend Play
Entertainment;Music & Video
Entertainment;Education
Entertainment;Creativity
Entertainment;Brain Games
Entertainment;Action & Adventure
Entertainment
Educational;Pretend Play
Educational;Education
Educational;Creativity
Educational;Brain Games
Educational;Action & Adventure
Educational
Education;Pretend Play
Education;Music & Video
Education;Education
Education;Creativity
Education;Brain Games
Education;Action & Adventure
Education
Dating
Communication;Creativity
Communication
Comics;Creativity
Comics
Casual;Pretend Play
Casual;Music & Video
Casual;Education
Casual;Creativity
Casual;Brain Games
Casual;Action & Adventure
Casual
Casino
Card;Brain Games
Card;Action & Adventure
Card
Business
Books & Reference;Education
Books & Reference;Creativity
Books & Reference
Board;Pretend Play
Board;Brain Games
Board;Action & Adventure
Board
Beauty
Auto & Vehicles
Art & Design;Pretend Play
Art & Design;Creativity
Art & Design;Action & Adventure
Art & Design
Arcade;Pretend Play
Arcade;Action & Adventure
Arcade
Adventure;Education
Adventure;Brain Games
Adventure;Action & Adventure
Adventure
Action;Action & Adventure
Action
0
100
200
300
400
500
600
700
Popularity

Which App Genres are most Popular?
It is evident that the most popular Genres are **Tools** , **Entertainment** and **Education**.

- Many developers create tons of **Tools** and utility apps that work really well.The genre is quite diverse. Tools and utility apps come in handy in our day to day lives which make our lives easier and easy to manage. Tools like *Google Assistant, Google Search , Shareit* are the most popular ones.

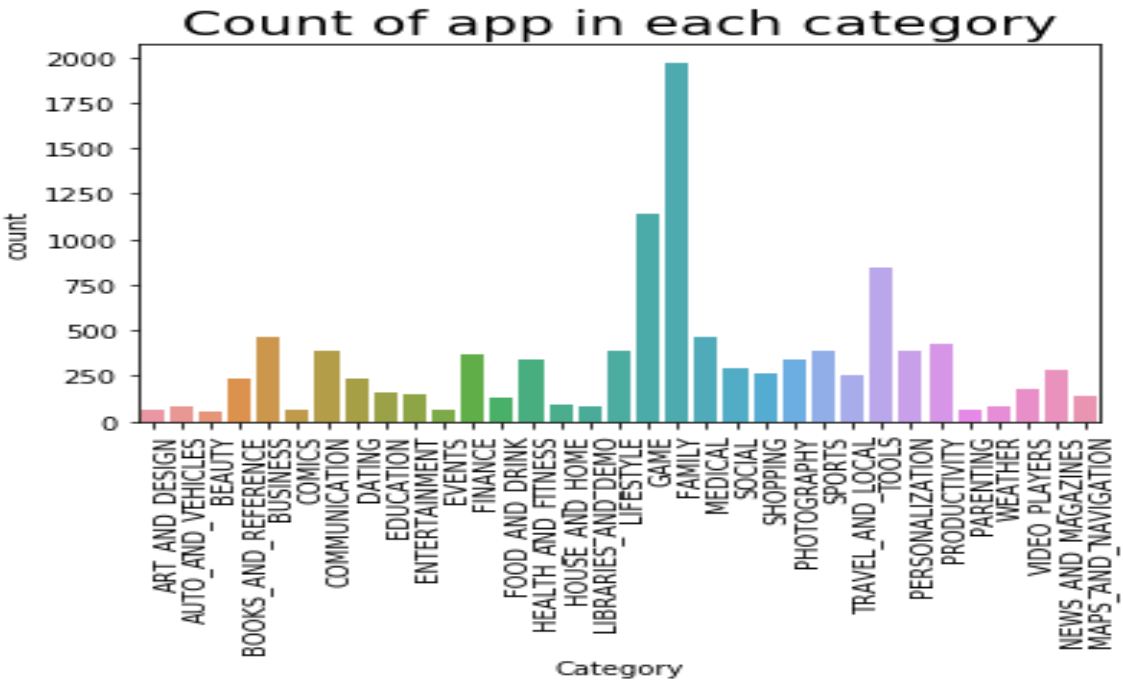
- From ebooks to music to live television - Google Play Store offers several such platforms for **Entertainmen**t Apps like *Netflix & YouTube* are the most popular ones.

- The quality of an **Educational** app is a pivotal attribute in the ever progressing world of teaching and education, and it can add a great deal of value to the learner's educational experience. The clichéd adage, 'Quality over quantity' has been overused, but quite deservedly so.



Does size of the App matter ?
The above visualisation is a **histogram** of Number of apps based on **Size**. It is evident that apps with size between 0-20 MB have a higher frequency. This is because the app size and conversion rate has a mutual connection which directly affects the number of apps download.Conversion rate is 87.94% of apps with the size between 0-20 MB while the conversion rate is 76.31% of 100+ MB sized apps. Well, the data doesn't indicate a gigantic

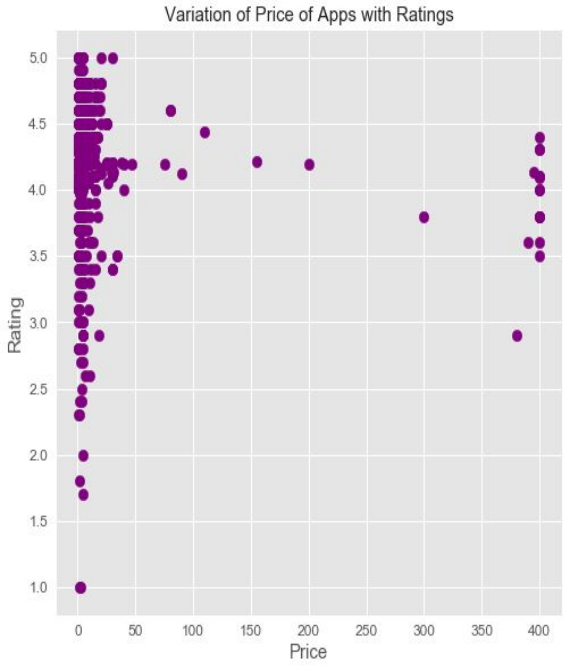
difference between the two factors, but it demonstrates that when the size goes up, the conversion rate comes down. On the basis of percentages revealed by data, the most recommendable size of an app for app developers is between 0-20 MB for better conversion rate.

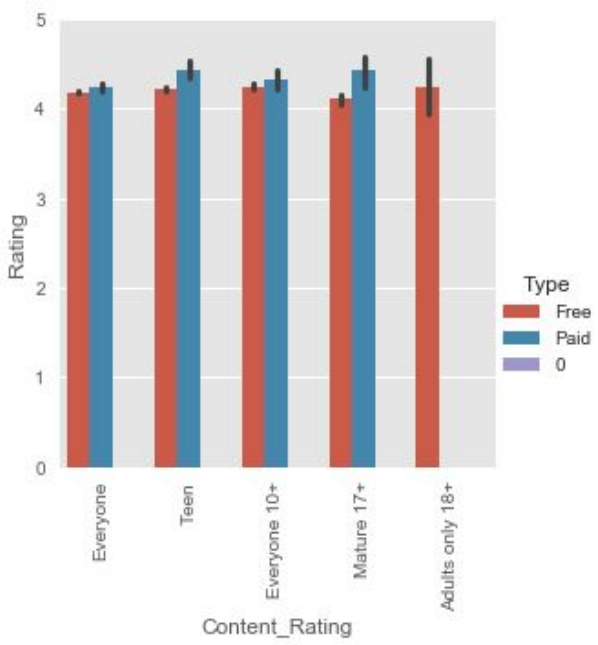


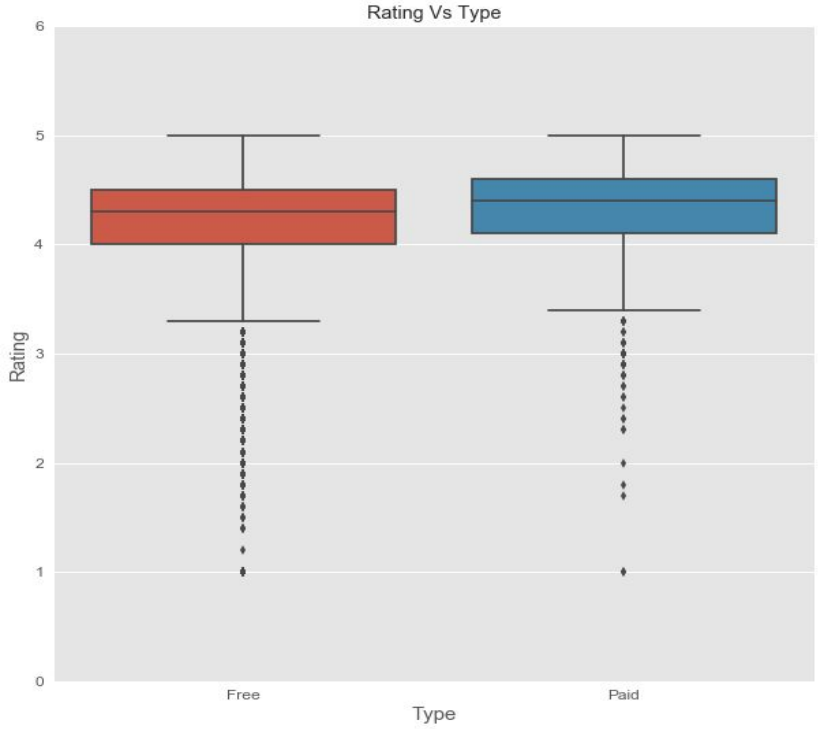
What sorts of apps are present on Google Play Store ?
The bar-plot depicts the several app Categories that are present on Google Play Store out of which:

More than 50%(i.e >1800) apps fall in Family category(if you are having a hard time managing your family, then it might be a good idea to get some help from family management apps.).
~800 fall in the Games category, ~750 fall in the Tools category and so on.

State why the frequency of some categories are relatively higher than the others.
This is because the app developers understand the need of the general population and apps are built to fulfill user demands. Thus higher numbers simply mean greater need.







How does rating varies based on Type(Free/Paid) of App?

It is obvious that general population uses the apps that are available for free. People also prefer to use the app which is already well rated. As well as the general population don't generally check apps that are paid and therefore ~97% of the apps that fall under the Paid section go unnoticed and unrated.

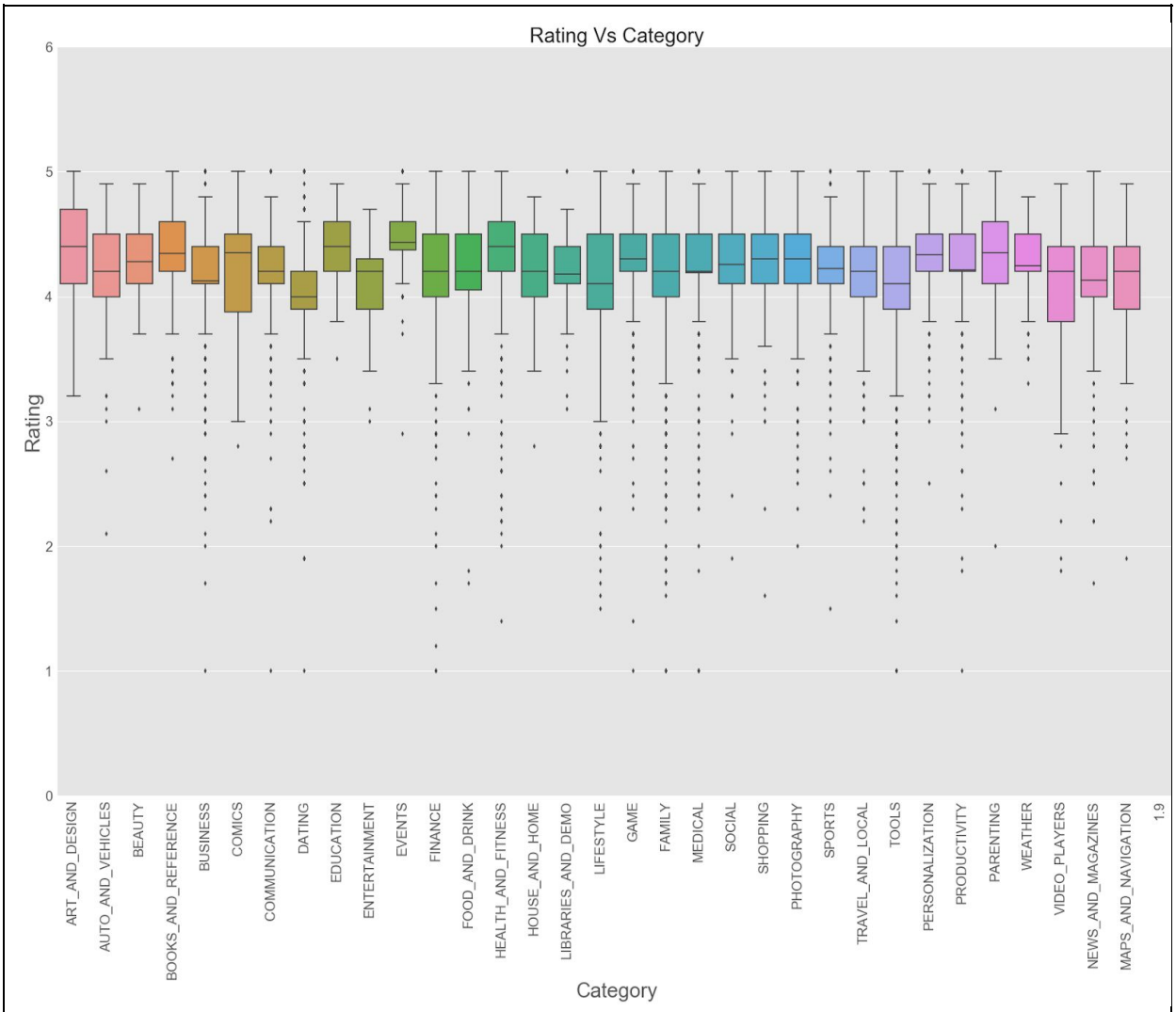
As visually evident, the points are concentrated along price value=0 and the sparse as the price

increases.It can be noticed, that there are some outliers present which are relevant and can't be removed as there are a few costly apps present on Play store like Zollinger's Atlas of Surgery($249) and CyberTuner($999.99) and Agro($999.99).

Variation of Price of Apps and Rating ? What role do ratings play in the market ?

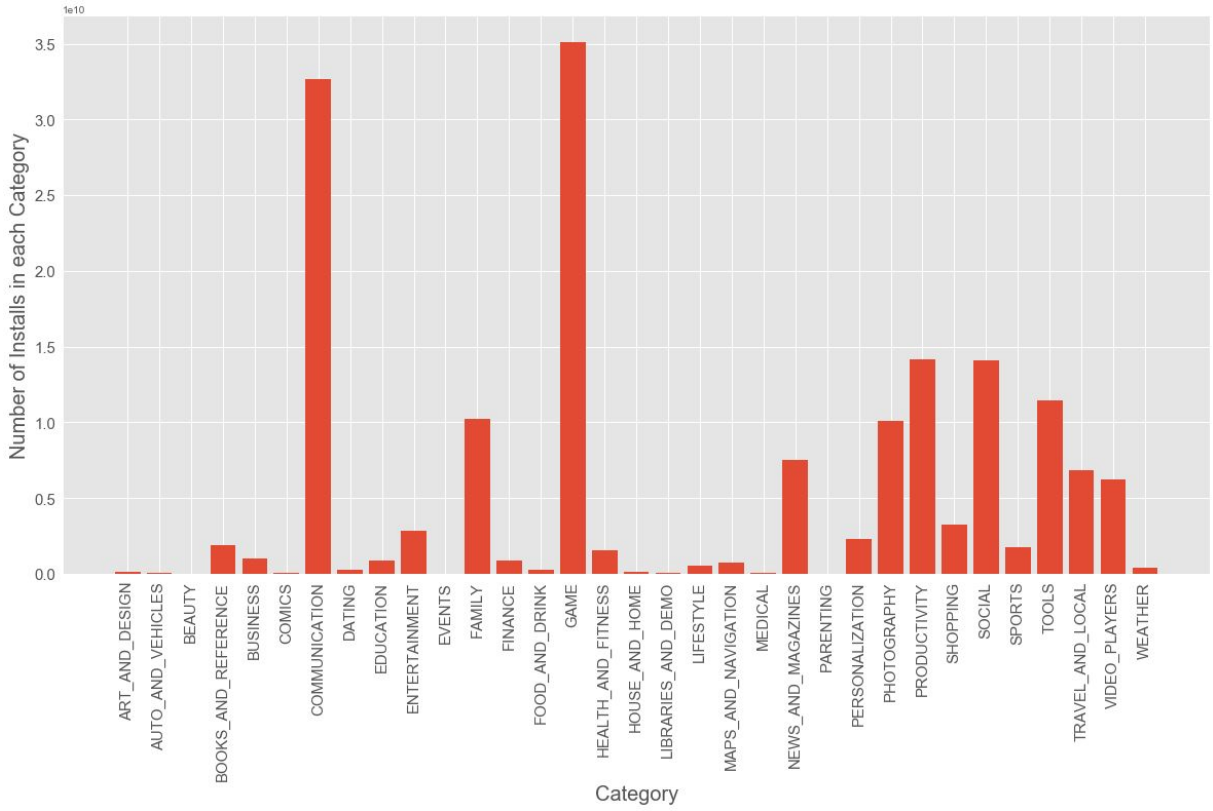
Jumping From...	To...	Is Expected to Increase App Store Conversion by...
★	★★	30%
★	★★★	340%
★	★★★★	730%
★	★★★★★	770%
★★	★★★	280%
★★	★★★★	540%
★★	★★★★★	570%
★★★	★★★★	89%
★★★	★★★★★	97%
★★★★	★★★★★	4%

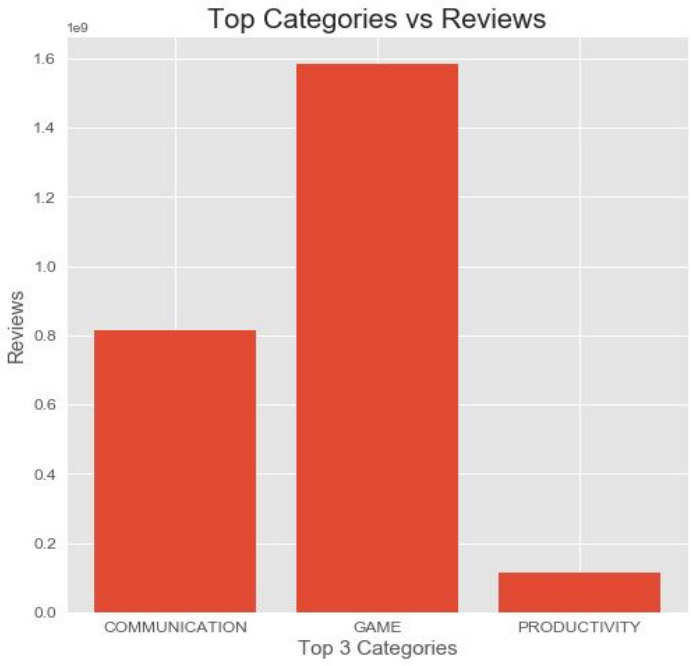
Ratings and reviews impact App's discoverability. Ratings play a big role in the customer adoption process, and data from our research on the value of a rating revealed that 59% of people usually or always check ratings before downloading an app—even if everything else checks out.In addition, the opportunity cost of a rating is huge. Take a look at the image above to see how a change in your average rating will impact your app store conversion.



The above comparative boxplots give a brief idea about the Rating of apps under each category. The ratings range from as low as 1 to as high as 5.

There are several outliers which can be overlooked in this case as there are some not so good apps present in every category on the store and therefore, get a really low rating. There are some extremely good and useful apps too.

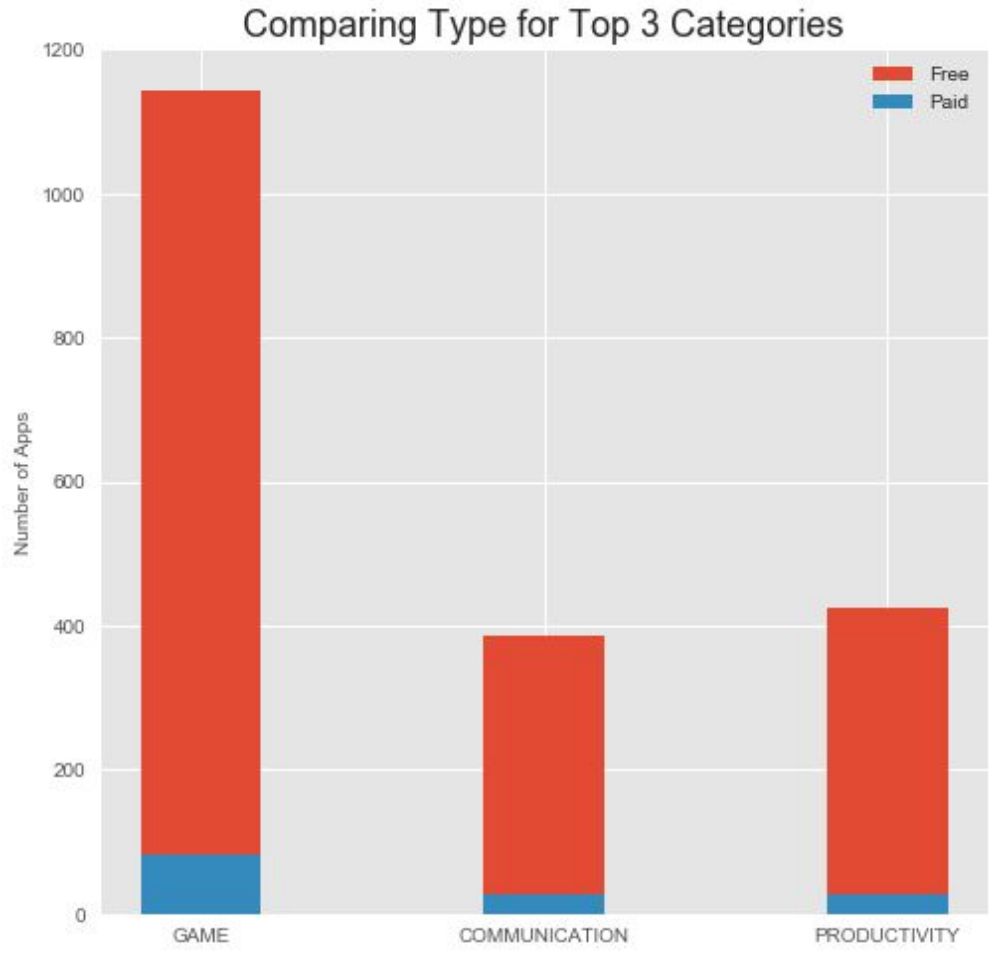




Which are the most popular App categories?
Is the hypothesis "*More the Number of Installs , more popular the App is***" *True ?***

It is evident that the most popular app categories are 1. Game 2. Communication 3. Productivity based on both the number of Installs and Reviews(*two Variables*). Thus, the hypothesis **"More the Number of Installs** , **more popular** the App is" True.

- In this decade, almost every teenager has an Android phone and games are a basic necessity for them.Not just them,there are games present for almost every age group and in different genres.
- Next basic need of people as said "*Humans Are By Nature Social Animals*" is communication as the graphs also signify. Communication platforms like Whatsapp messenger , Instagram, Google Duo are most popular these days.
- It's no surprise that productivity apps have become so popular. Most of us have too many priorities,commitments,work,and other things-and not enough time. But luckily, there are apps - most of them free and downloadable to make your life easier, more balanced,more productive, and most importantly, more in control of your time. Apps like Google Docs.

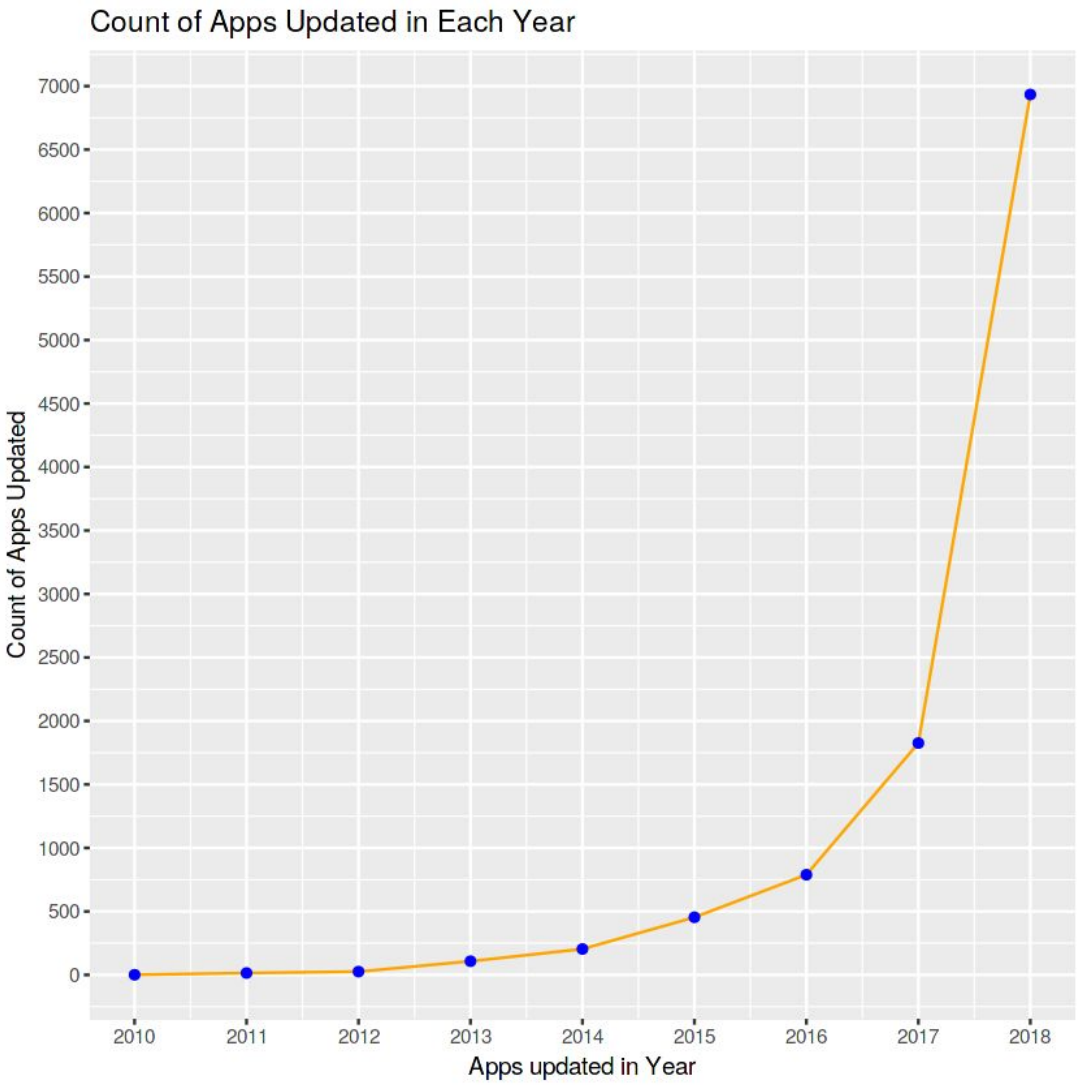


How does the Type(Free/Paid) vary for top 3 categories ?

Top 3 categories ranked as 1.Game 2.Communication & 3.Productivity.

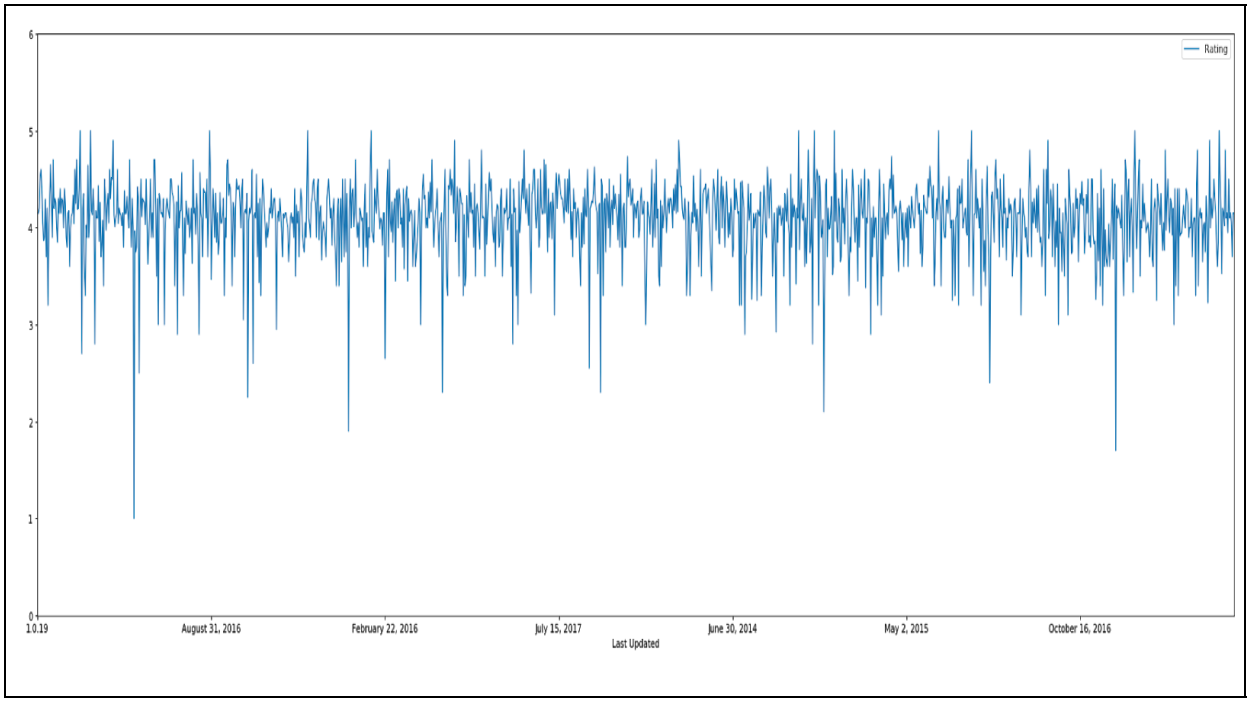
It is safe to say that even though Communication is a more popular category than Productivity, the

number of free apps provided by the former is lesser than the latter. Though the number of paid apps for the two is almost equal. The genre of Gaming takes a lead in both types as it is anyway the most popular genre and therefore, the apps falling in this genre can provide a very good conversion rate if they're paid.



What has been the past trend in the number of Apps updated in each year?
It can be deduced that there were only a few apps which were last updated in starting years 2014-2016. Most probable hypothesis would be that the developers have stopped support for these apps.Developers do provide good support and regular updates on Apps which has been shown by the Steep Reading for Year 2018. Almost 70 percent of Apps were updated recently.

Variation of Last update based in rating :



CONCLUSION

The following conclusions can be made after deeply analysing the dataset :-

- Most of the apps available on the store are meant for everyone ~80.4% without any restrictions based on age groups. The apps are targeted to the general public.

- The app size plays an important role. Apps with size between 0-20 Mb are most downloaded. Apps with size in that range are generally of relevance and of daily use. The app size and conversion rate has a mutual connection which directly affects the number of apps download.Conversion rate is 87.94% of apps with the size between 0-20 MB while the conversion rate is 76.31% of 100+ MB sized apps.

- It was noticed that the number of apps of a certain category were higher than the all other categories. This is probably because the app developers understand the need of the general population and apps are built to fulfill user demands. Thus higher numbers simply mean greater need. They understand the market really well.

- Rating of apps also depends on it's type i.e whether it's paid or open source. Through analysis, it is safe to conclude that the apps that are paid generally go unnoticed and unrated by the regular users. Users generally prefer apps that are open source without membership or license so that they don't have to go through the trouble of renewing their licenses and this also restricts the usage of the same app on multiple devices through sharing as licence keys are meant for a single device.

- Ratings and reviews impact app's discoverability. Ratings play a big role in the customer adoption process, and data from our research on the value of a rating revealed that **59% of people usually or always check ratings before downloading an app**—even if everything else checks out. *People act according to the opinion of the majority.* In addition, the opportunity cost of a rating is huge.

- Today,android technology is in the reach of every other person irrespective of their age group. Thus, it could be concluded that ***Games*** is the most popular genre. Hence, signifying that the majority of the population making use of the app store is **under 20 years old**.

- It was proved that more the number of installs an app gets , greater is the popularity of that app in the market.

- Approximately, *90%* of the apps present on the store have been made open source and available to the general population. A free app means that it has a near zero user acquisition barrier, since the user doesn't have to provide payment information up front, or that they are not presented with any perceivable cost. After they have experienced the app, it would be easier for them to decide whether or not to pay for extra content, if there are any.

- The number of apps that have received updates over a period of last 8 years has been significantly increased every year. It is evident from the analysis that the growth in the number is rather exponential. Thus, the growth in the number of updates is due to the advancement in android technology over the years.
