GROUP 4 Report

Analysis of Apps of Google Play Store

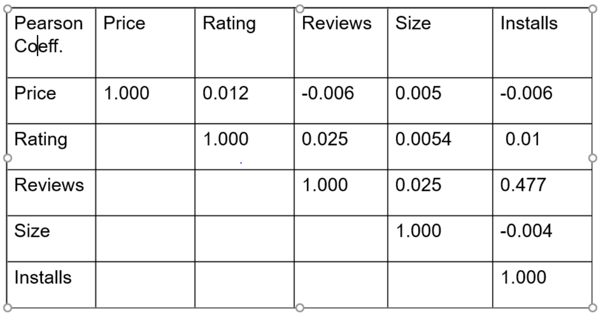
1 Details of Dataset

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 dataset is chosen from Kaggle. It is the web scraped data of 32k Play Store apps for analyzing the Android market. It consists of in total of 31997 rows and 9 columns, which are as follows:
1) App (Name)
2) Category (App)
3) Rating (App)
4) Reviews (User)
5) Size (App)
6) Installs (App)
7) Broad category (App)
8) Price (App)
9) Content Rating (Everyone/Teenager/Adult)

2 Data Cleaning

The raw data can have random sorting. To solve this, we sort the data according to category It should be noted that each and every piece of raw data may lead to a more accurate result. The current dataset holds values that are in the string format. We should convert the strings to a numerical format. Now to take care of null values, instead of dropping the rows that contain null values, we have replaced them with zero. Converting our data into appropriate forms Size: The size of the app is in "string" format. We needed to convert it into a numeric value. If the size is "10M", then 'M' was removed to get the numeric value of '10'. If the size is "512k", which depicts app size in kilobytes, the first 'k' was be removed and the size was converted to an equivalent of 'megabytes'. Installs: The value of installs is in "string" format. It contains numeric values with commas. It should be removed. And also, the '+' sign should be removed from the end of each string. Category and Content Rating: The Category and Content Rating consists of categorical values that should be converted to numeric values if we need to perform regression. So, these were converted to numeric values. Price: The price is in "string" format. We should remove the dollar sign from the string to convert it into numeric form.

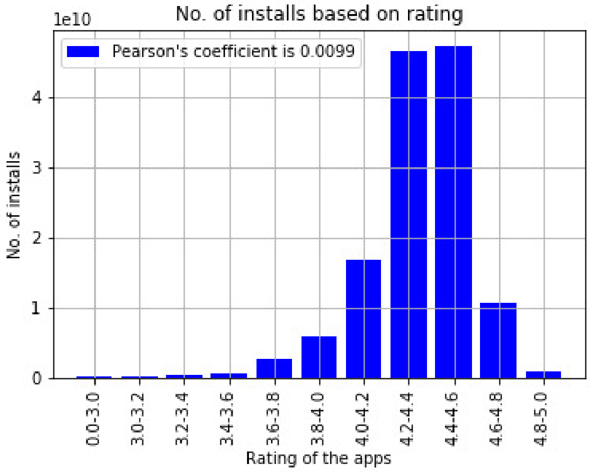
3 Correlation/relation between different attributes



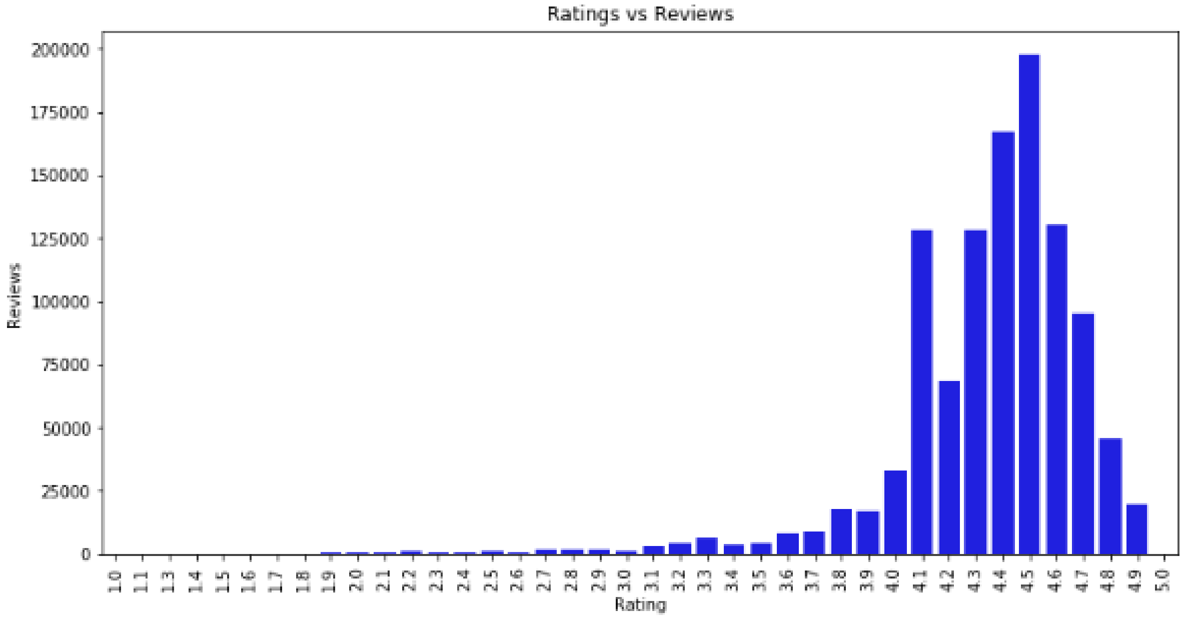
Pearson Coeff.	Price	Rating	Reviews	Size	Installs
Price	1.000	0.012	-0.006	0.005	-0.006
Rating		1.000	0.025	0.0054	0.01
Reviews			1.000	0.025	0.477
Size				1.000	-0.004
Installs					1.000

Inference from Pearson coefficient matrix-
1) Generally we think Install and price of an app will have a fair negative correlation but according to our analysis they are totally uncorrelated as paid apps have more and better features which make them popular so people prefer to install them and hence correlation approaches -0.006.
2) Installs and Size are uncorrelated (slightly negative) because large sized generally apps have more and better features.
3) Installs and Rating of an app are uncorrelated as generally high rated apps are downloaded by most but when no. of installs increases expectation from the app increases and hence people start rating it negatively.

4 Rating Vs Installs

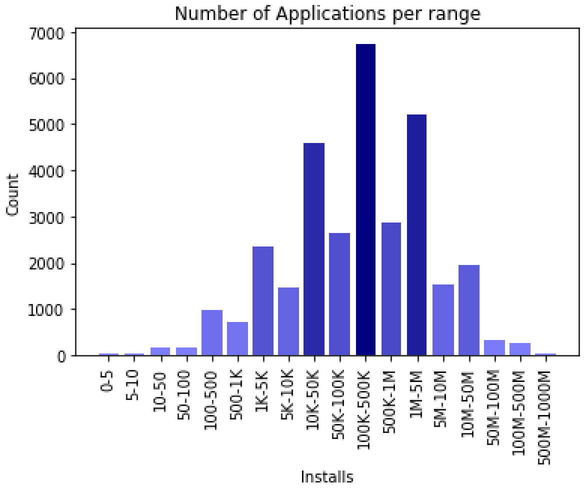


We can see the count of ratings is increasing as rating increases as it reaches a peak value at 4.4 and thereafter it starts decreasing. The reason behind this decrease might be that the new apps, which are having relatively less no. of downloads are positively rated by a small no. of users.
It would be better to plot rating vs review because people writing review will give more meaningful rating



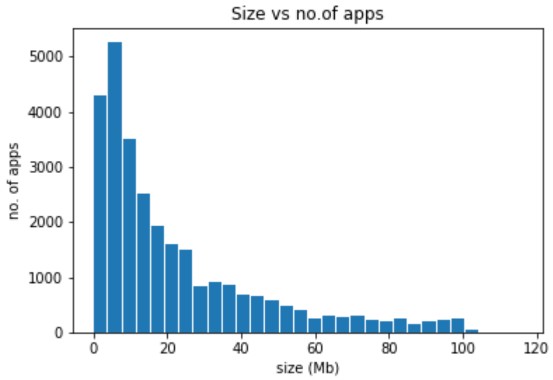
The plot is similar to the previous one, as expected.

5 No of Apps per Range



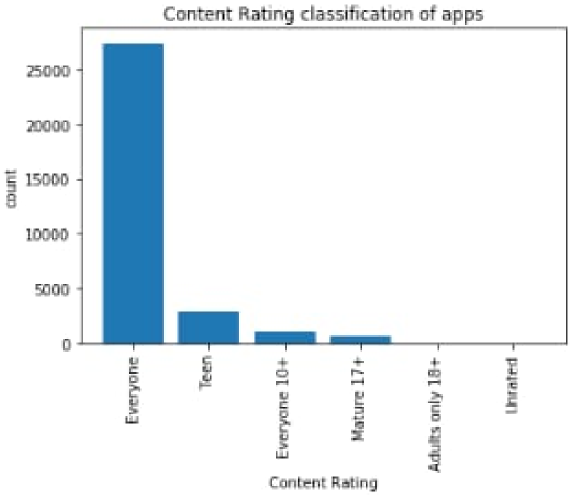
As it can be seen the graph is a sort of bell curve i.e no. of very popular apps (large no. of installs) as well as unpopular apps (very few installs) is very low. Most of the apps have avg. installs i.e in the range of 100K-500K.

6 App Size vs No of Apps



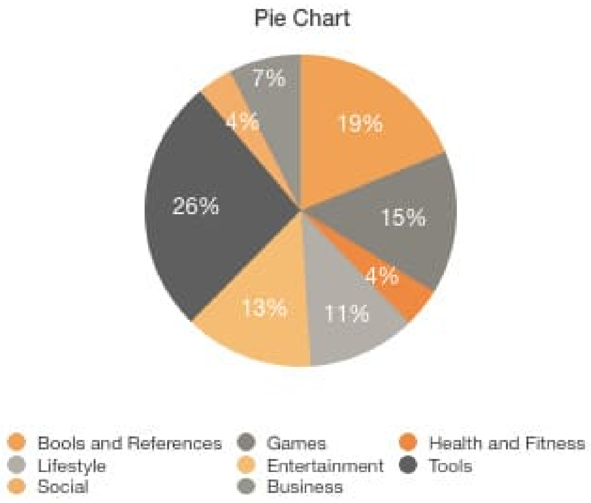
In this graph we can see that no. of apps is inversely proportional to the size of apps is inversely proportional to the size of apps. The simple reason behind this might be android development has become easy over the years and many developers are developing small sized apps which are easy to develop for beginners.

7 Content Rating Classification of Apps



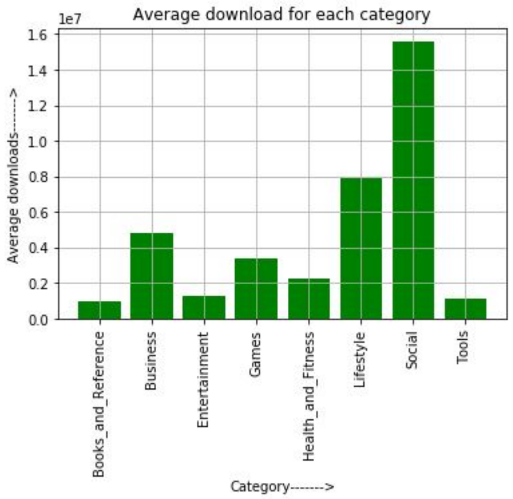
Almost all apps have content rating “Everyone” because most of the app developers want to target the whole population.

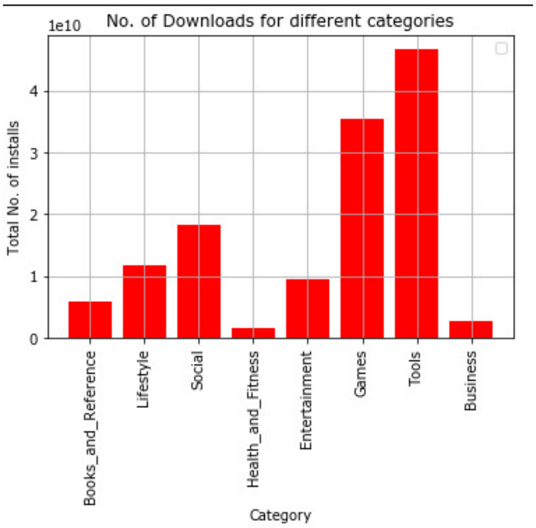
8 Category wise distribution of Apps



Max. apps (26 percent) belong to 'tool' category and minimum belong to 'social' (4 percent) and 'Health and fitness' (4 percent) category.
The percentage of apps depends on foreign factors such as-
1. Variety of apps in category ex. tools
2.Monopoly of popular apps eg. Social
3.Requirement

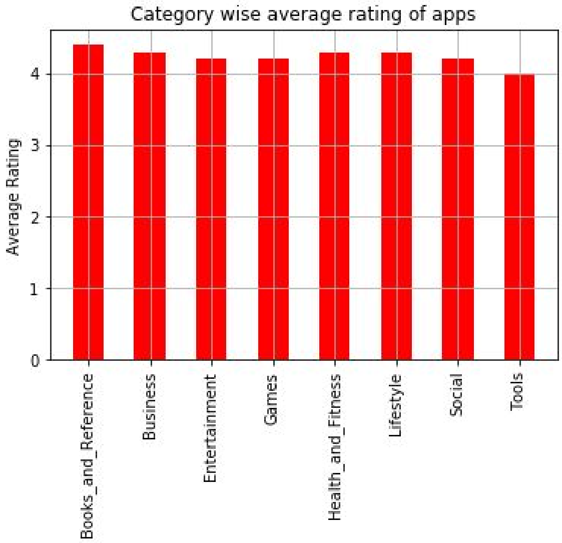
9 Average Download and Total Downloads for each categories





It can be seen that average no. of downloads is maximum for the “social category” because the no. of such apps is low (4 percent) but they are very popular. Average downloads for “Books and Reference” category is the lowest. On the other hand the total download for “social ” apps is relatively low where as that of “tools” apps is high. This increase in average no. of downloads can be attributed to the popularity and no. of apps.

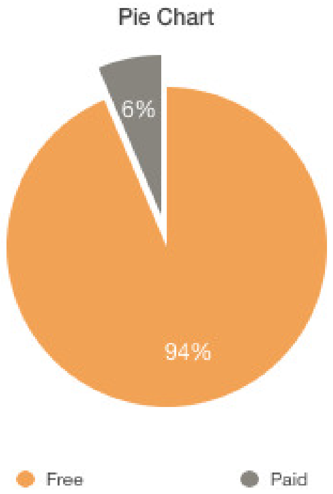
10 Category wise Average Rating



1.The average rating of apps is uniform irrespective of the category.
2.It should be noted that average rating is also around 4.4.
3.It can be concluded that there are no outliers in category w.r.t rating.

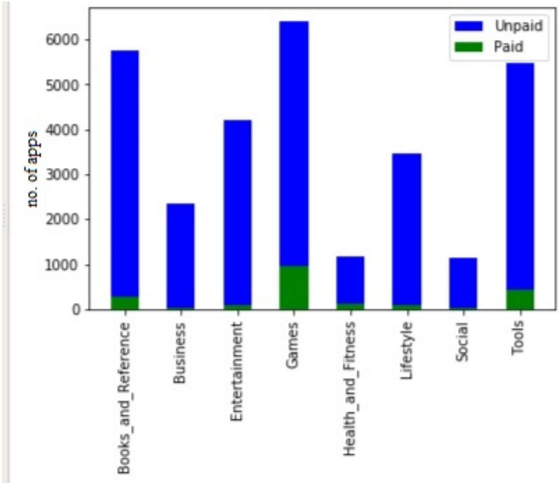
11 Paid and Free Apps

11.1 Distribution



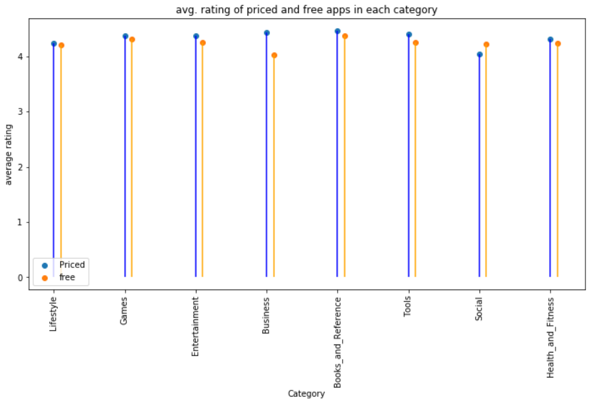
Majority of the apps (94percent) are free. Then, what are sources of income of the developers of free apps?
1.In-app advertising- The developer weaves in Google AdMobs into his apps which display the advertisement when a user is using that app and the developer gets an incentive for each click on the Ad.
2.Freemium-Some apps (mainly gaming) have goodies cost money to get unlocked (ex. PUB-G: where there is UC money which you can buy by paying and get exclusive items)]

11.2 No of Apps in each category



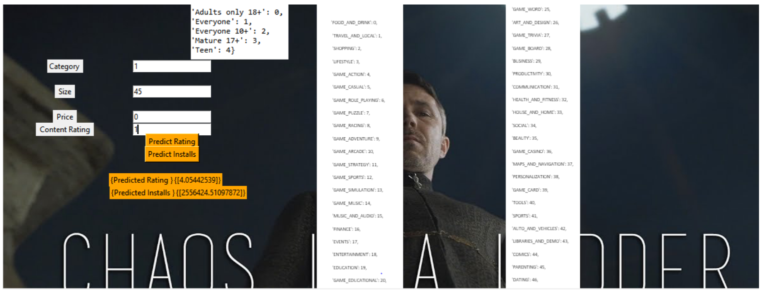
First basic observation is that the no. of free apps is always greater than no. of paid apps. It can be seen that there are more paid gaming apps than any other category.

11.3 Average Rating in paid and free



Generally, the average rating of paid apps is greater than average rating of free apps in each category. This is obvious because paid apps promise better experience. But in the case of social apps the average rating of free apps is greater than avg. rating of paid apps because-
1.As it can be seen in the graph that the no. of paid social apps is negligible when compared to free social apps.
2.Even free social apps like Facebook, Instagram provide quite satisfactory experience.

12 Regression Implementation



```
'Adults only 18+': 0,
'Everyone': 1,
'Everyone 10+': 2,
'Mature 17+': 3,
'Teen': 4}
```

Category: 1

Size: 45

Price: 0

Content Rating: 1

```
{Predicted Rating }{[4.05442539]}
{Predicted Installs }{[2556424.51097872]}
```

```
'FOOD_AND_DRINK': 0,
'TRAVEL_AND_LOCAL': 1,
'SHOPPING': 2,
'LIFESTYLE': 3,
'GAME_ACTION': 4,
'GAME_CASUAL': 5,
'GAME_ROLE_PLAYING': 6,
'GAME_PUZZLE': 7,
'GAME_RACING': 8,
'GAME_ADVENTURE': 9,
'GAME_ARCADE': 10,
'GAME_STRATEGY': 11,
'GAME_SPORTS': 12,
'GAME_SIMULATION': 13,
'GAME_MUSIC': 14,
'MUSIC_AND_AUDIO': 15,
'FINANCE': 16,
'EVENTS': 17,
'ENTERTAINMENT': 18,
'EDUCATION': 19,
'GAME_EDUCATIONAL': 20,
```

```
'GAME_WORD': 25,
'ART_AND_DESIGN': 26,
'GAME_TRIVIA': 27,
'GAME_BOARD': 28,
'BUSINESS': 29,
'PRODUCTIVITY': 30,
'COMMUNICATION': 31,
'HEALTH_AND_FITNESS': 32,
'HOUSE_AND_HOME': 33,
'SOCIAL': 34,
'BEAUTY': 35,
'GAME_CASINO': 36,
'MAPS_AND_NAVIGATION': 37,
'PERSONALIZATION': 38,
'GAME_CARD': 39,
'TOOLS': 40,
'SPORTS': 41,
'AUTO_AND_VEHICLES': 42,
'LIBRARIES_AND_DEMO': 43,
'COMICS': 44,
'PARENTING': 45,
'DATING': 46,
```

For regression we have used Random Forest Regression. In this regression, we use multiple decision trees to reach at the final predictions. Each decision tree is trained using a method call bagging. Bagging (Bootstrap Aggregation) is used when our goal is to reduce the variance of a decision tree. Here idea is to create several subsets of data from training sample chosen randomly with replacement. Now, each collection of subset data is used to train their decision trees. As a result, we end up with an ensemble of different models. Average of all the predictions from different trees are used which is more robust than a single decision tree. A collection of such random decision trees is called a random forest.In our model, the inputs are Category, Size, Price and content rating, and the predicted values are Installs and Ratings, i.e. we have used two random trees.

13 Ending note

We took this dataset because we wanted to conclude for entrepreneurs that where they should invest there time and money. Our focus was to decide whether 'social' apps or 'gaming' apps are more profitable. We see that though average number of social apps is quite high, still the total number of social apps is low and competition is high and market is dominated by apps like Facebook and Instagram. On the other hand, gaming apps are popular despite competition and scope is quite high, especially for 'action games'. The dataset contains immense possibilities to improve business values and have a positive impact. It is not limited to the problem taken into consideration for this project. Many other interesting possibilities can be explored using this dataset.

Ashutosh Sharma and ***Team***