Capstone Project - 1

Project Title – Play Store Review Analysis

Team Members

Indrashis Chakraborty

Shaurabh Pandey

Analysing the data to discover key factors responsible for app engagement and success.

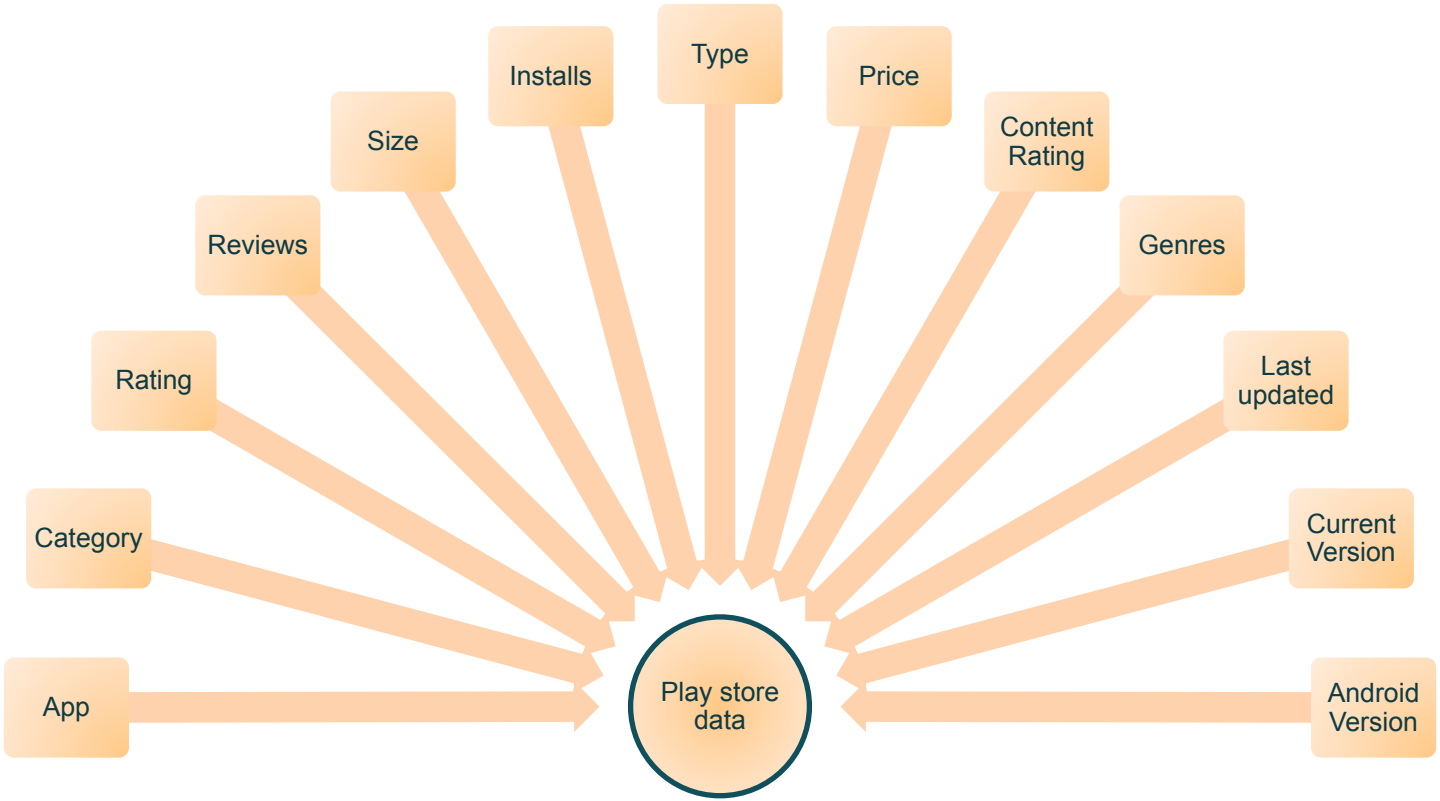
Content

- Defining the problem statement.
- Defining the Variables.
- Data cleaning and arranging.
- Eda on the two datasets

Google play store

- Google Play, also branded as the Google Play Store and formerly Android Market, is a digital distribution service operated and developed by Google.
- It serves as the official app store for certified devices running on the Android operating system and its derivatives as well as Chrome OS, allowing users to browse and download applications developed with the Android software development kit (SDK) and published through Google.
- 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.
- In this project we will be analysing the data to find the key factors that are responsible for app engagement and success

Defining the Variables(1st DataFrame)



Defining the Variables (continued)

- App – Application name
- Category – Category of the application .
- Rating – User rating of application between 1 to 5.
- Reviews – Number of reviews the app has received
- Size – Size of the application in mb
- Installs – Number of times the app has been installed
- Type – Type of application, whether it is free or paid.
- Price – Price of the app
- Content Rating – It states the age category to which the application belongs.
- Genres – Genre of the app
- Last Updated – The date, when the application last updated
- Current Version – The current version of the application on play store
- Android Version – The android version in which the application will install and run properly.

Data cleaning and arranging.

- First we dropped all the duplicate observations having the same app name
- While going through the process of data cleaning we observed that one of the app has a rating of above 5 which is not allowed on google play store , on a second look we found that value of all the column for this app is out of place and hence necessary corrections were performed

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
9300	Life Made WI-Fi Touchscreen Photo Frame	1.9	19.0	3.0M	1,000+	Free	0	Everyone	NaN	February 11, 2018	1.0.19	4.0 and up	NaN
9300	Life Made WI-Fi Touchscreen Photo Frame	NaN	1.9	19	3.0M	1,000+	Free	0	Everyone	NaN	February 11, 2018	1.0.19	4.0 and up

- Null values of Rating were replaced with mean values from the database.
- Data type of Rating and Reviews were changed from string to float and int respectively

- Converted the app size from mb to its equivalent value in bytes and so changing it from a string to a number
- In installs columns the data was in string format indicating the floor level of installation, so we removed the “+” and comma separator and converted the data to an integer for better analysis

- The format of last updated column was changed to date format
- The dollar sign was removed from price column and changed the data to int for better analysis
- In the 2nd Dataframe all the apps having null/na in their translated review column were dropped

Exploratory Data Analysis

For Play store Data

- Highest Rated and Most Installed Apps
- Relation between category and installs
- Relation between category and rating
- Relation between content rating and installs
- Relation between content rating and rating
- Comparison between different type of apps(free & paid)
- Analysis of paid apps Category by installation and price
- Installation Rate depending on date of last update

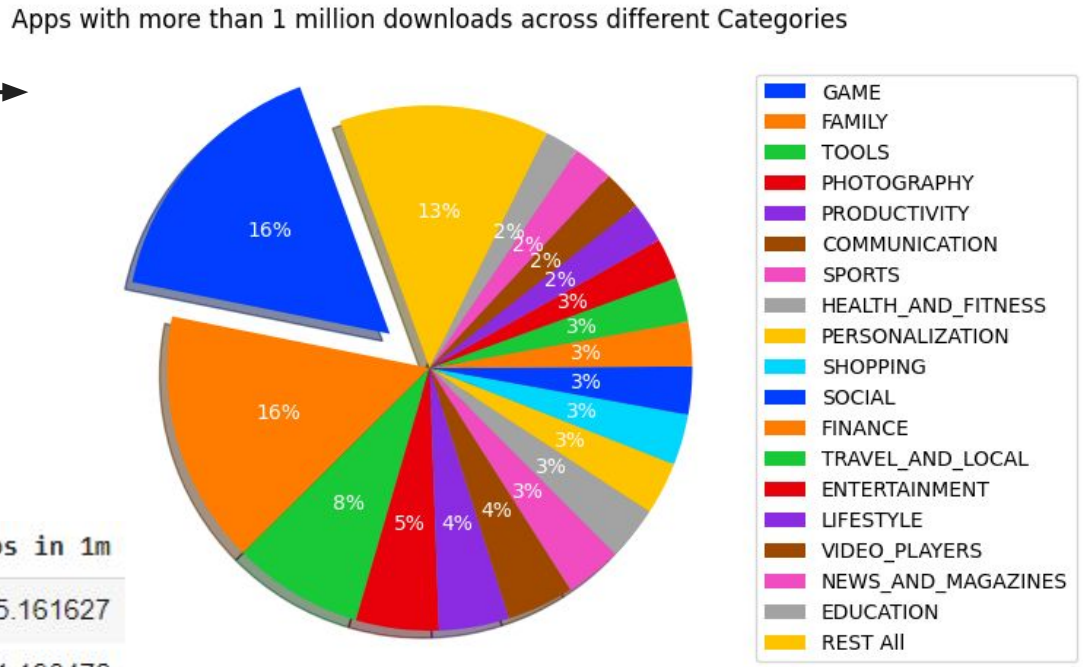
Highest Rated and Most Installed Apps

- Our dataset had a total of 9660 unique apps
- Of this 9660 apps , 3395 apps have more than 1 million downloads
- Of all the apps having more than 1 million downloads, 8 apps had the highest Rating of 4.9 with three of them being from Health and Fitness category
- Facebook , Whatsapp and Instagram are the top 3 apps when it comes to no. of Reviews and Installs ,
- Facebook is the most popular app on google play store with more than 1 Billion downloads and almost 70 million user reviews

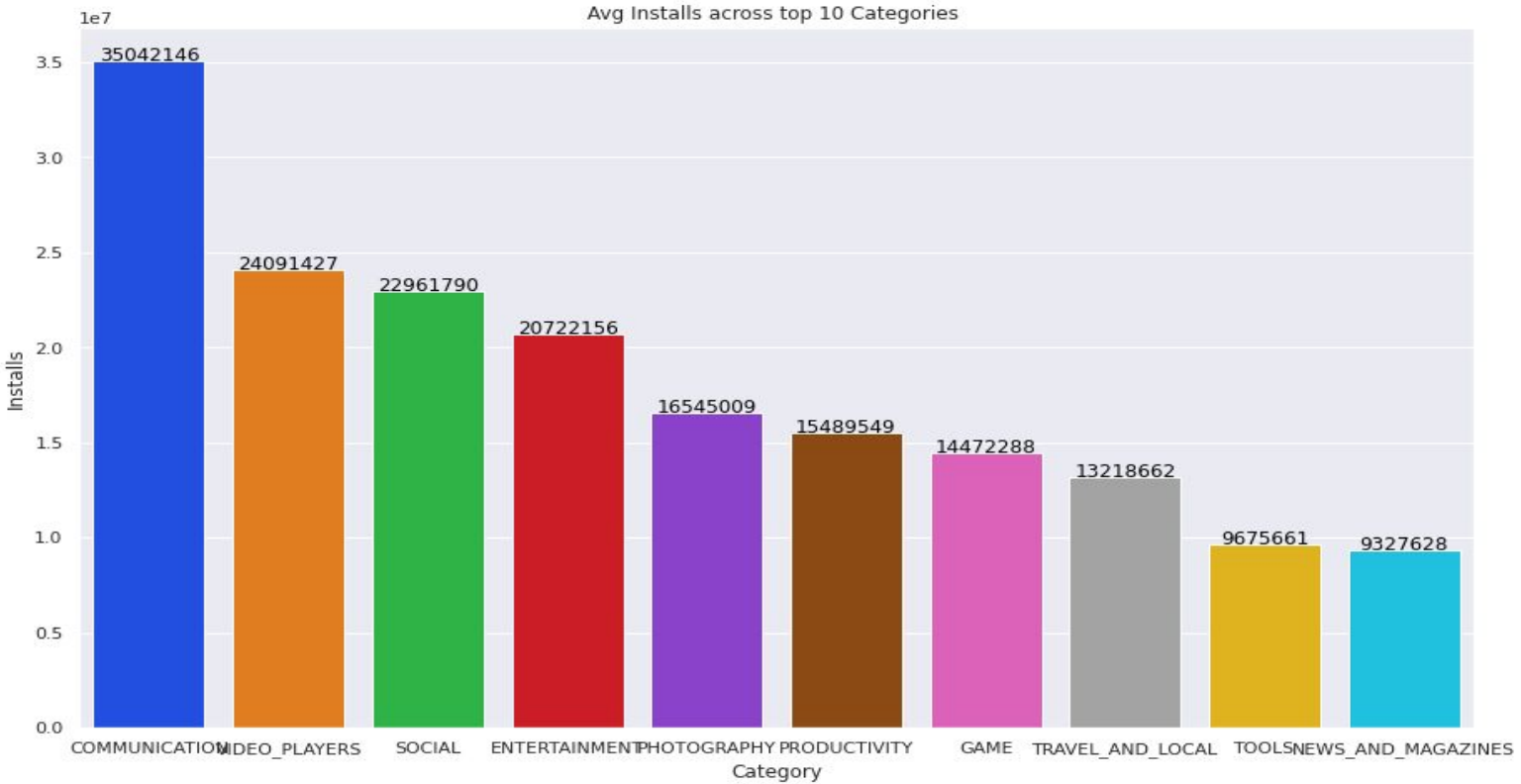
Relation Between Category and Installs

Of all the apps having more than 1 million downloads Game and Family were the two most popular categories

Family , Food_and_drink and Photography are the three most successful categories with more than 40% of their apps having 1 million+ downloads

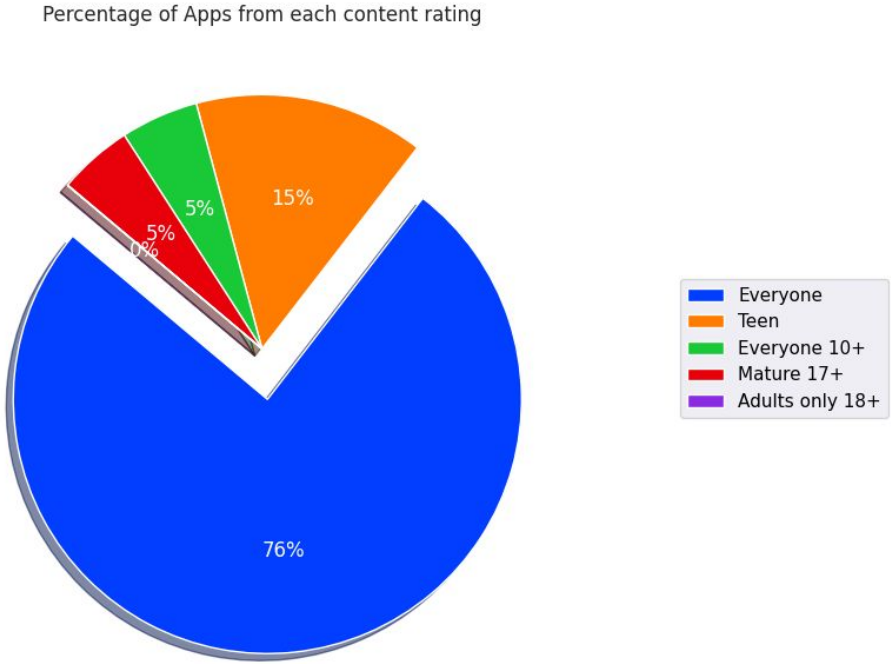


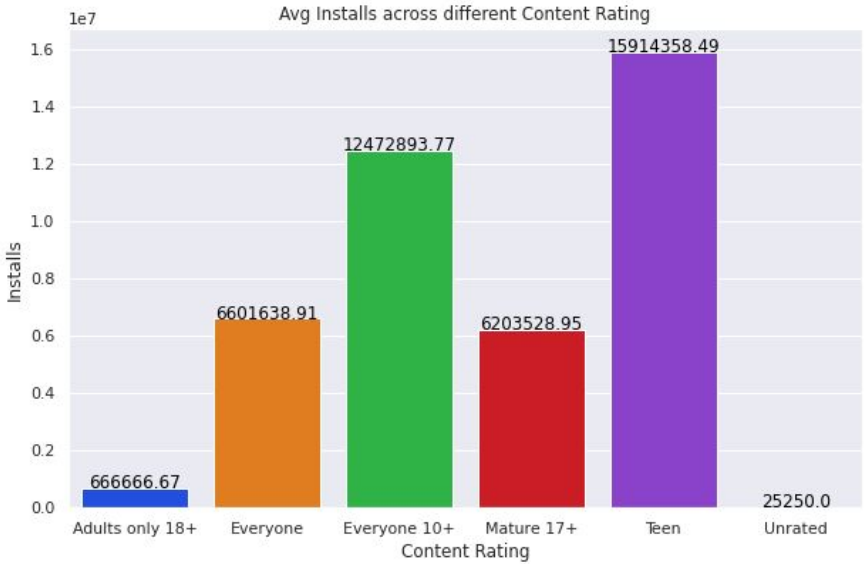
	Category	Percentage of apps in 1m
0	FAMILY	55.161627
1	PHOTOGRAPHY	41.190476
2	FOOD_AND_DRINK	41.176471



Communication Category has the highest number of Avg Installs among all the other Categories

Relation Between Content Rating and Installs

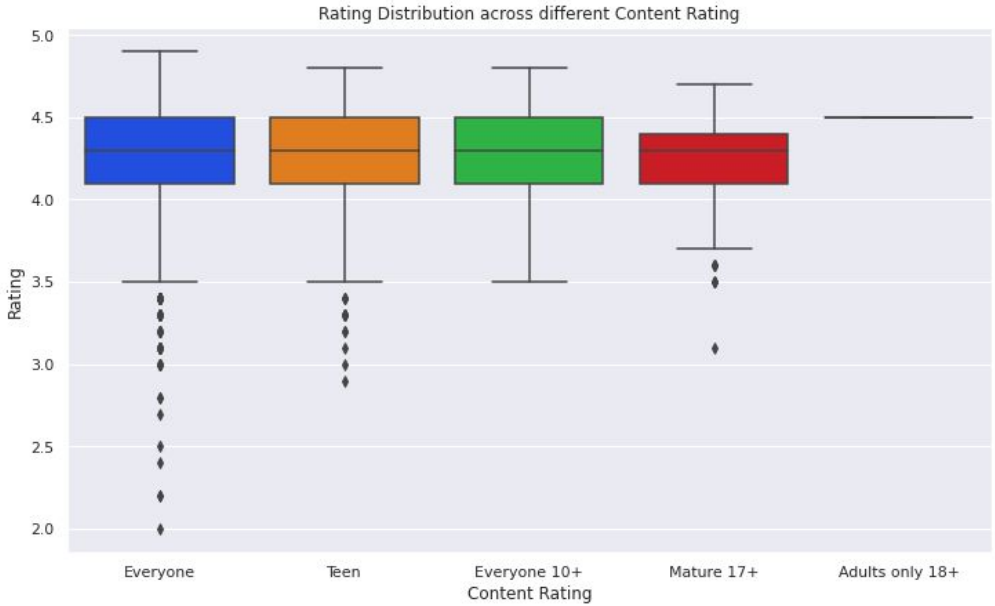




From apps having more than 1 million download almost 76% of them have a Content Rating of everyone , which says that most of the apps on google play store are suitable for all age groups

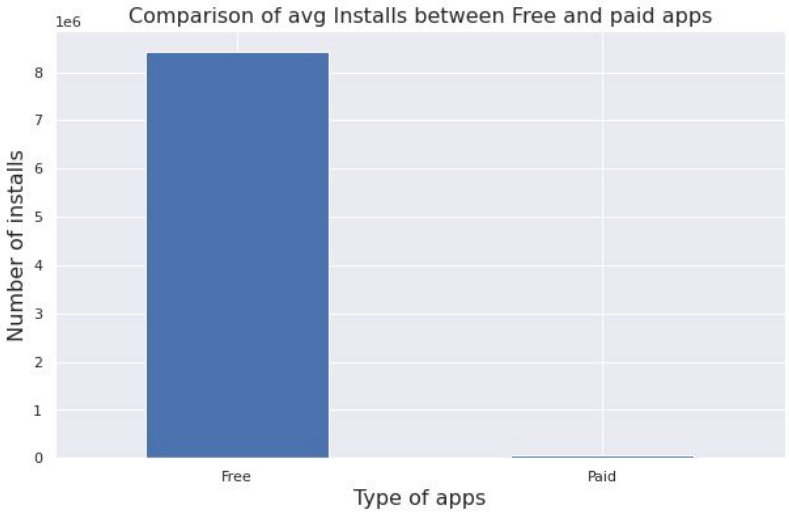
Apps made for teens have the highest average installation rates proving that majority of the audience on google play store belong to a young age group

Relation Between Content Rating and Rating

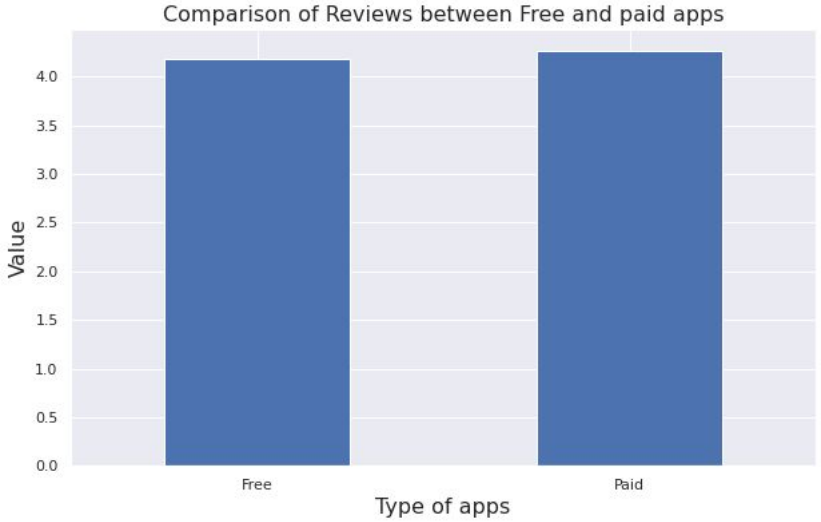


- Everyone , teen and Mature 17+ all three have some apps with abnormal low rating
- Median value for all the Content Ratings is almost same at around 4.3
- Overall Rating for apps carrying Mature 17+ content rating is slightly on the lower side

Comparison between different types of apps

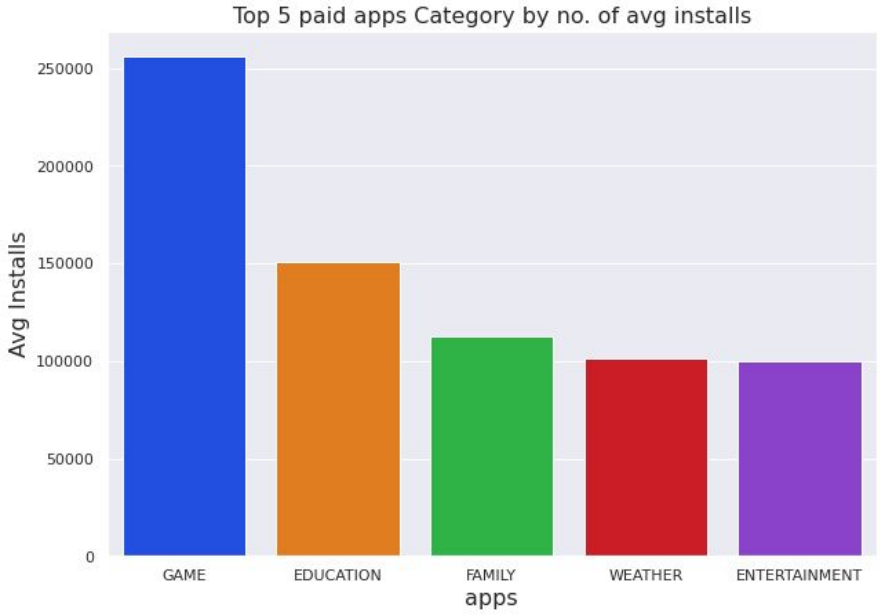


There can be seen a huge difference in avg installs between free and paid apps , which proves that the audience for paid apps is quite small as compared to free apps

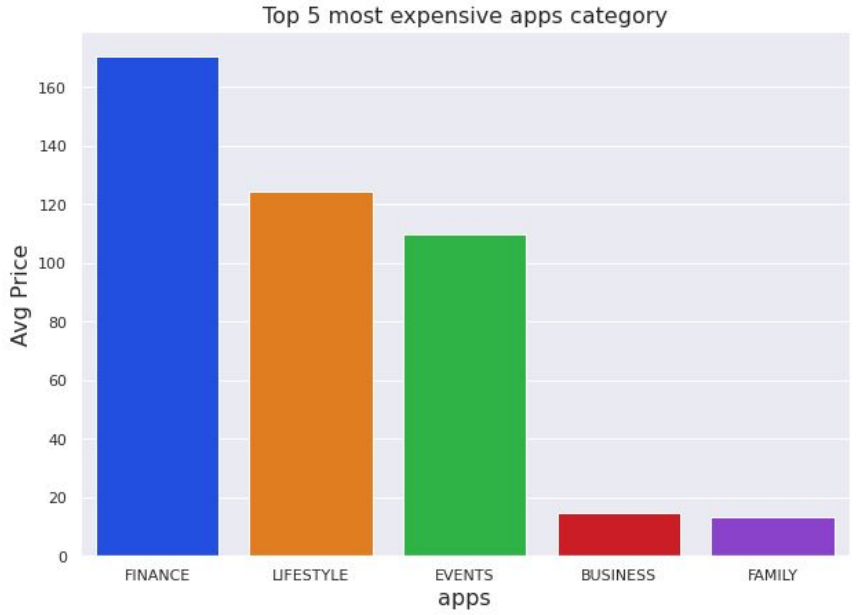


Paid apps hold a slight edge over Free apps in terms of avg rating proving that sometimes a price tag brings an inch better quality

Analysis of paid apps Category by installation and price

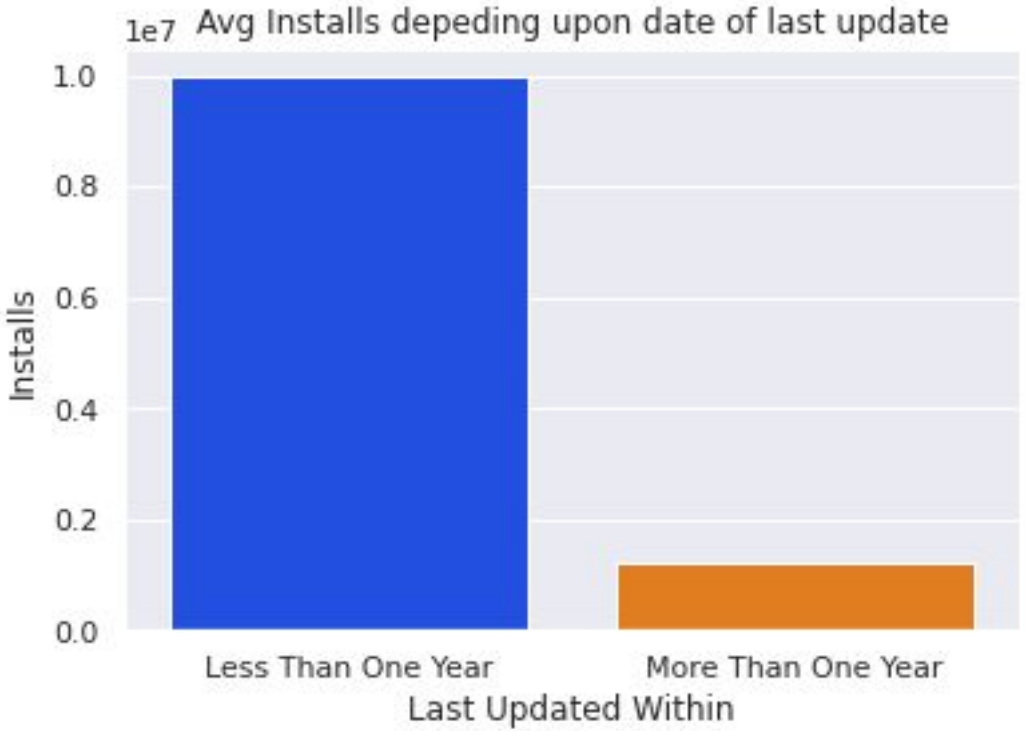


Gaming and education are the two categories for which people are more open to pay for the app . Premium gaming and premium education are the two things that people are ready to pay for



Paid apps that belong to finance category have the most highest avg price of almost around 163 dollars followed by Lifestyle , so premium finance software carry a high price tag with them

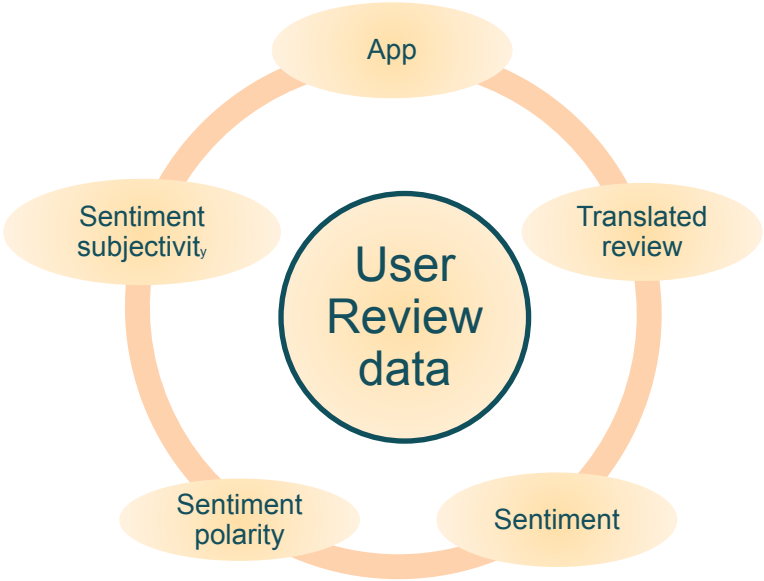
Installation Rate depending on date of last update



Those apps that are updated less than a year before have almost 10 times more installation as compared to the ones that not updated since one year which means that regular updates and keeping the app bug free is really important for any app's success

Defining the Variables (2nd DataFrame)

- Translated Review – Review of the user.
- Sentiment – It is the emotion expressed in the review , can be positive, negative or neutral
- Sentiment Polarity – It denotes the level of sentiment on a scale of -1 to 1 where -1 is for extreme negative and 1 is for extreme positive
- Sentiment Subjectivity – Subjectivity is when text is an explanatory article which must be analysed in context.



Exploratory Data Analysis

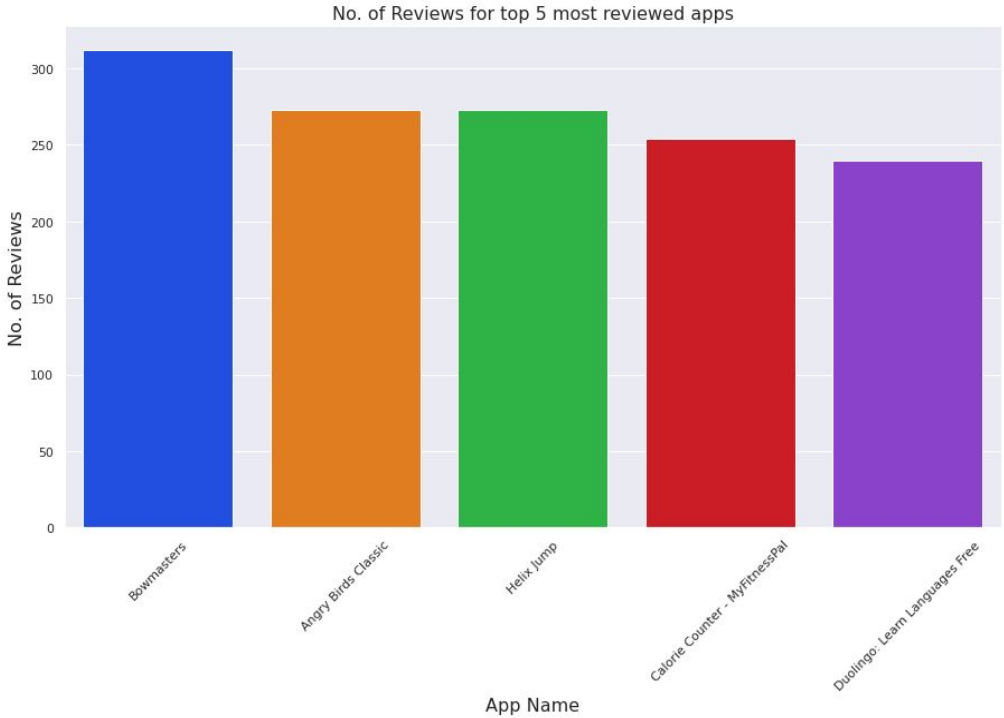
For user Review data

- Top 5 most reviewed Apps
- Word-cloud of the most reviewed App
- Sentiment Analysis of top 5 most reviewed app
- Top 5 apps with the most overall positive sentiment

For Merged Data

- Data Selection
- Correlation
- Top apps

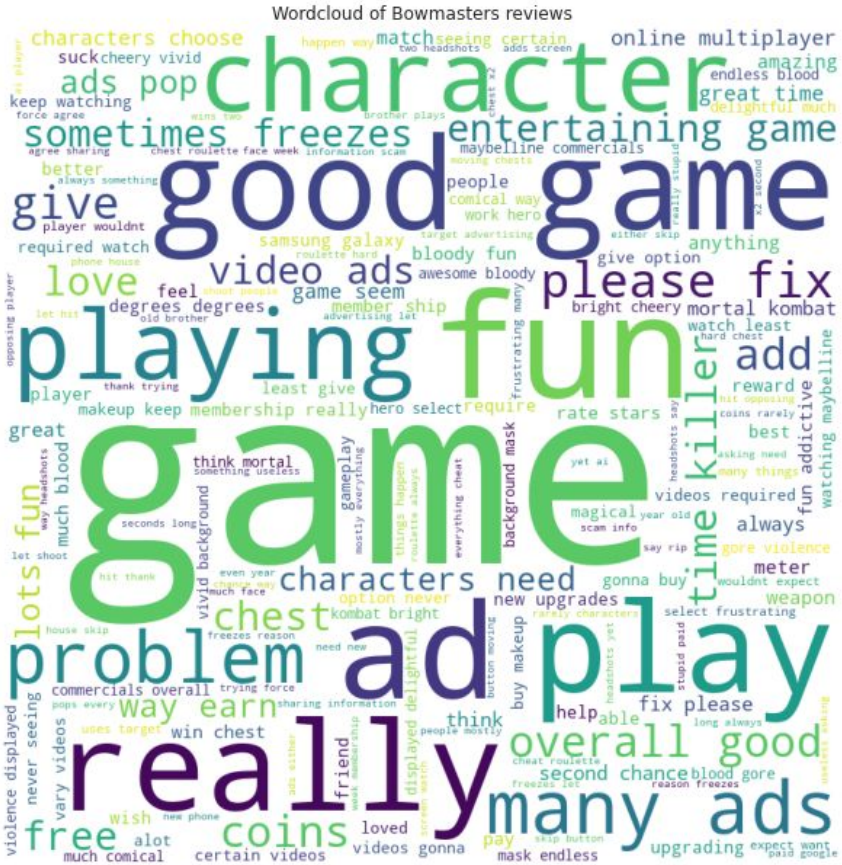
Top 5 most reviewed Apps



- Bowmasters, Angry Birds Classic , Helix Jump , Calorie Counter – MyFitnessPal , Duolingo: Learn Languages Free are the top 5 apps with most number of Translated Reviews
- Bowmasters is the most reviewed app with a total of 312 translated Reviews

Bowmasters	312
Angry Birds Classic	273
Helix Jump	273
Calorie Counter - MyFitnessPal	254
Duolingo: Learn Languages Free	240

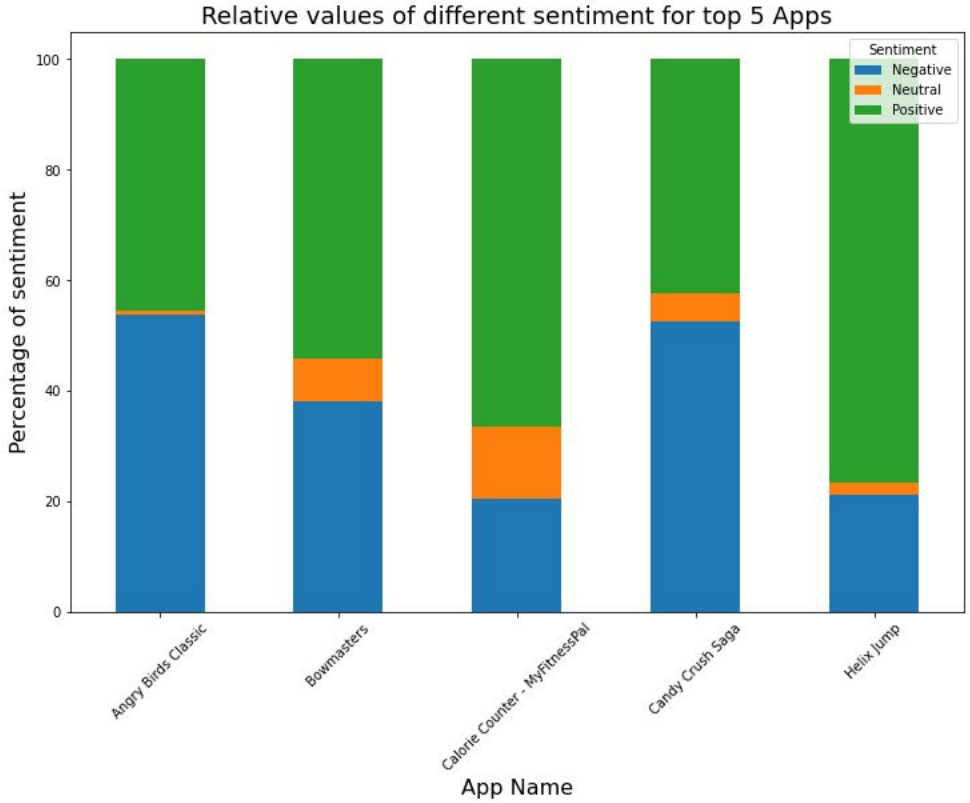
Word-Cloud of most reviewed App



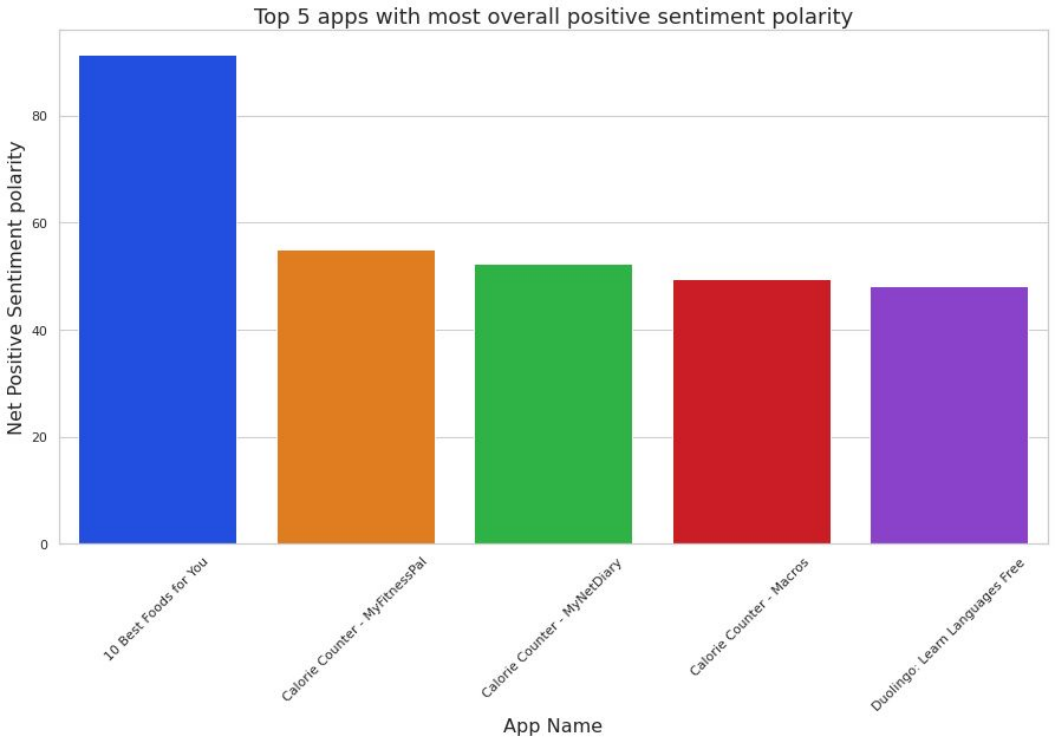
- Bowmasters is the app with most number of translated Reviews
- From the word-cloud we can see that it is a good fun game to play
- Word-cloud also makes it clear that a fair numbers of users were bothered by the number of in game ads

Sentiment Analysis of top 5 most reviewed app

- In case of angry birds classic and candy crush saga we can see that the amount of translated reviews with a positive and negative sentiment are almost equal
- Helix jump, Calorie Counter – MyFitnessPal and Bowmasters all three have majority of their reviews in the positive bracket



Top 5 apps with the most overall positive sentiment



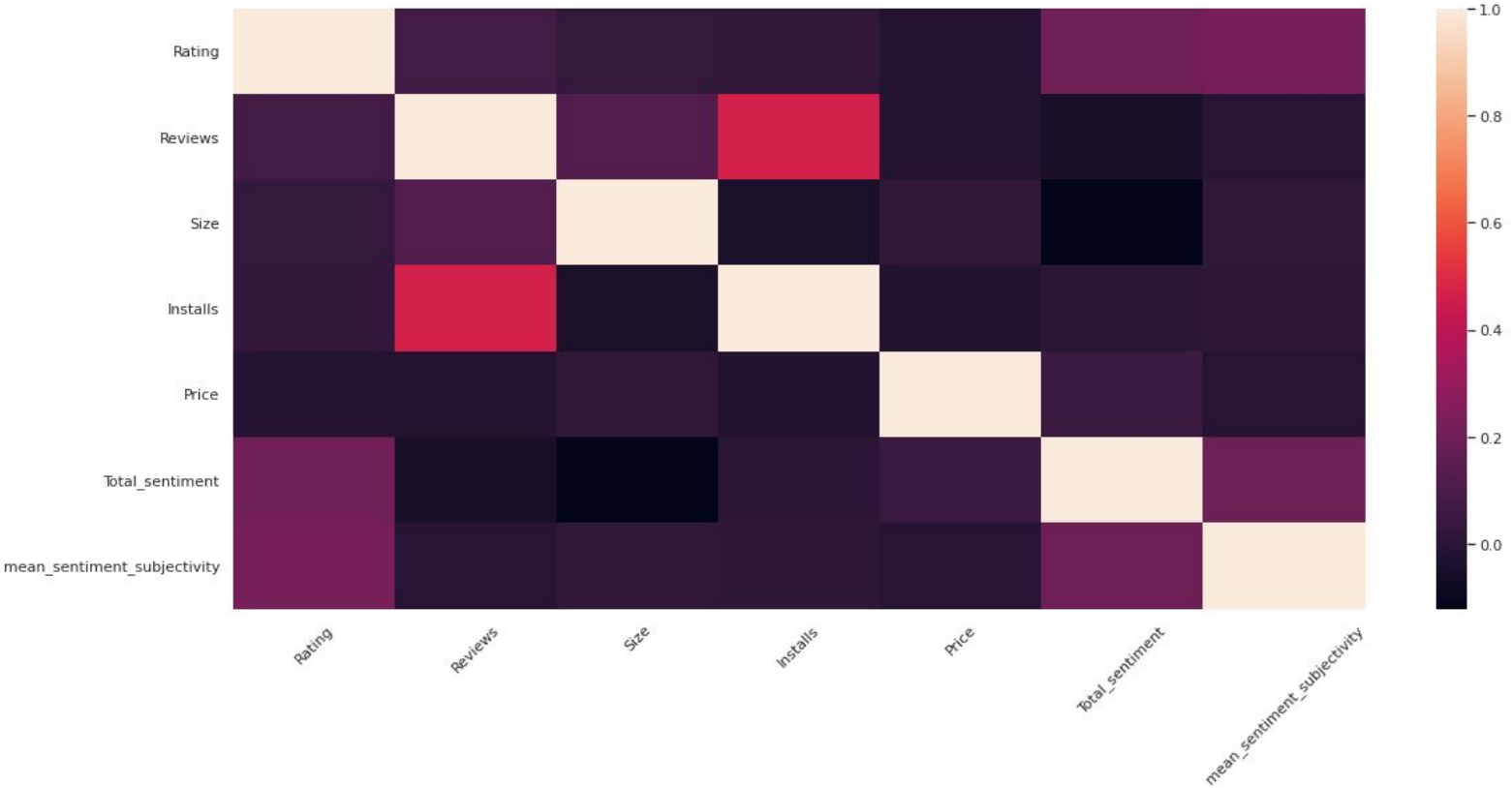
- 10 Best food for you , Calories Counter – MyFitnessPal , Calories Counter – MyNetDiary , Calorie Counter – Macros and Duolingo: Learn Languages Free are the 5 topmost in terms of net positive sentiment polarity
- We can notice that in general reviews for Calorie counter apps or in a more broader perspective fitness apps have a strong positive sentiment polarity to it

Data selection:

In merged data, we filtered apps by means values of Installs, reviews, rating and total sentiments , so that we can get those apps which are good according to all the possible parameters

	Rating	Reviews	Installs	Price	Total_sentiment	mean_sentiment_subjectivity
count	816.000000	8.160000e+02	8.160000e+02	816.000000	816.000000	816.000000
mean	4.282598	7.113919e+05	2.670946e+07	0.099767	7.965328	0.491992
std	0.313643	3.640515e+06	1.099350e+08	1.278179	8.764741	0.085183
min	2.600000	1.140000e+02	1.000000e+03	0.000000	-9.726559	0.000000
25%	4.100000	7.777750e+03	1.000000e+06	0.000000	2.009767	0.452758
50%	4.300000	4.065050e+04	3.000000e+06	0.000000	6.478120	0.497103
75%	4.500000	1.983450e+05	1.000000e+07	0.000000	11.262122	0.537958
max	4.900000	7.815831e+07	1.000000e+09	29.990000	91.322167	0.916667

Correlation Heatmap



Top Apps Overall

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver	Total_sentiment
0	Google Photos	PHOTOGRAPHY	4.5	10858556	Varies with device	1000000000	Free	0.0	Everyone	Photography	August 6, 2018	Varies with device	Varies with device	28.174331
1	Google	TOOLS	4.4	8033493	Varies with device	1000000000	Free	0.0	Everyone	Tools	August 3, 2018	Varies with device	Varies with device	9.893660
2	Dropbox	PRODUCTIVITY	4.4	1861310	61M	500000000	Free	0.0	Everyone	Productivity	August 1, 2018	Varies with device	Varies with device	17.961120
3	Clash Royale	GAME	4.6	23133508	97M	100000000	Free	0.0	Everyone 10+	Strategy	June 27, 2018	2.3.2	4.1 and up	11.410923
4	Duolingo: Learn Languages Free	EDUCATION	4.7	6289924	Varies with device	100000000	Free	0.0	Everyone	Education;Education	August 1, 2018	Varies with device	Varies with device	48.097857
5	Flipkart Online Shopping App	SHOPPING	4.4	6012719	Varies with device	100000000	Free	0.0	Teen	Shopping	August 6, 2018	Varies with device	Varies with device	8.162738

The value of all the variables of these apps are above the mean value of the dataset

Challenges:

- Huge size of data was to be handled keeping in mind not to miss anything which is of any relevance for analysis.
- A fair number of duplicate apps were present in the first database and so they needed to be dropped so to not have polluted results from our EDA
- Data types of all the variables in both the datasets needed to be handled and transformed carefully so we can perform meaningful analysis on them

Conclusion -

- Facebook is the most popular app on google play store in terms of no. of reviews and installs
- A total of 3395 apps out of 9660 apps had more than one million downloads
- From all the apps with more than 1 million downloads , 32% of them were from Game and Family Category and also it is the Family category which had 55% of its total apps with more than 1 million downloads the highest among all other categories
- Communication category had the highest average installs
- 76% of apps on play store have a content rating of everyone although it's the apps with a content rating of teen which is seen having the highest number of avg installs , so apps made for young audience have a higher probability of success
- A huge gap was seen in numbers of avg installs between free and paid app with the former having as much as 10 times more installs than the later , proving that the audience for paid apps is relatively very small
- Paid apps slightly edged free apps when it comes to rating giving the notion that a price also bring some extra quality
- Game and Education were the two most downloaded category of paid apps and hence they are the two field that somewhat can prefer higher quality than affordability

Conclusion (continued) -

- On an avg finance apps were the most expensive one out of all categories
- Apps that have been last updated more than a year before can be seen having almost 10 times less installs as compared to the ones that are more frequently updated ,so bug free and constant improvement is one of the major factor behind any app's success
- Bowmasters had the most number of translated reviews in our 2nd database
- We noticed that even though Bowmaster is a fun good game to time-pass but it also does have an annoying number of in-game ads
- Fitness apps and in particular apps coming under Calorie Counter banner had a strong sense of positive sentiment polarity attached with all its reviews
- 10 Best Foods For You had the highest overall positive sentiment polarity out of all the apps in our 2nd database