**Capstone Project-1 (EDA)**

**Play Store App Review Analysis**



**By**

**Balkrishna Tiwari**

**Gmail:-** balkrishnatiwari942@gmail.com

**LinkedIn:-** [linkedin.com/in/balkrishna-tiwari-71a727b5](https://www.linkedin.com/in/balkrishna-tiwari-71a727b5)

**GitHub Link:-** https://github.com/BalKrishna-Tiwari/Playstore-App-Analysis.git

**Abstract:** *Mobile applications have emerged as best tools for various industries and business to connect with there end users, with deeper penetration of IT sector. There are also government organizations, NGOs and similar philanthropist group that promote or execute their work through there specialized apps. Beside being a business or communication tool, mobile applications are also used directly as a business product that charges fees for its users. There has been increase in number of applications being launched on app store year on year. Quite a large number of app designers put their effort either individually or in a group to come up with well-functioning apps. Most of the app available on app store is free and there is no positive revenue generated from most of these app which cannot be utilized to label certain app being successful over other. The success of app thus totally depends on the users review and rating provided to it. Since apps are meant to be one of the prominent faces of the business or organization and a final product of the creator, they are more interested in finding key parameters for a successful and trending apps. I have done exploratory data analysis of apps data and apps review data available through google to find solutions to key business and technical queries. The study aims to highlight the key parameters tried on a successful apps historically and hence predicting solutions for new app development with higher success probability. I have utilized Pandas, Matplotlib and Seaborn library from Python to carryout various data analytics and visualization on available data set.*

1. **Problem Statement**

Our aim is to get detailed insight of app industry and various parameters that affects success or acceptance of app within its users. Raw data of app and its reviews are available with us. In order to achieve above objective, we need to clean our source data and further standardize them. On next part we have to carry out exploratory data analysis on cleaned data which will reinforce our understanding for the versatility of app within audience. We will try to explore trends within apps and its feature that suggests better success rate for historically launched and managed apps, so that same principles can be adopted by app makers to launch new apps on google app store.

1. **Introduction**

With the advancement and spread of Information and digital technology, usability of mobile phones among common people have increased drastically. The resent push of governments for digital economy and governance has also boosted use and popularity of smart phones. We have seen growth in use of smart phones among masses which is now not limited to only youngsters but have found interest of users belonging to all ages. Apps are the categorized feature segments of mobile that is designed to function and accomplish specific purpose. These purpose very according to needs of the user, like payments app is designed to carryout seamless transactions of digital payments with quickness coupled with integration of banking transactions and record keeping features simultaneously. With wide variation of audience there is wide demand of categories. And since there is high demand for app that has attracted large number of developers and companies to try their hand in app building business. But as most of the apps are free to use, the best one is selected from among the pool of supplies in same category. So, it has become very important for developers to check historical trends about the performance of app, and to apply insights to make their own business decision. Data science is the field that helps to extract logical details of bulk data and funnels them to conclusive remarks or results. I have used method named exploratory data analysis that falls under branch of data science to explore apps data and apps review data. The main aim as discussed in problem statement will be to get detailed insights about the behavior of variables and its effect on rating and reviews, that ultimately drives apps acceptance within its users. Based on results I would also be suggesting certain key points that need to be considered while designing or launching a new app on app store.

**3 Available Data sets**

We have two data sets available with us from google app store, Play store data and User reviews data. Play store data contains list of apps that is available on google play store for users, this list appears to contain apps that was being used during or till 2018. With list of apps, play store data also contains certain features that describes apps as a summery. Another data set, Review data contains list of written reviews given to particular app by its users. For this data we do not have any time stamp mentioned which would have added advantage in analysis to visualize change in trend in reviews. Details of the columns from both data sets are discussed in below sections.

**3.1 App Store Data**

App store data contains list of apps and aggregated values of certain variables associated with app. The data structure contains 10841 rows and 13 columns. This implies it has a record of nearly 10K+ apps and various aggregated features in rest 12columns. The data contains list of app and values of variables associated to it like current rating, number of reviews, category of app, android version required, current version of app etc. It is evident that this data set forms a unique set to identify nature of app and its valuation in different parameters.

There are non-standard data within column and missing value within this dataset which we will rectify in data cleaning stages. Also, there is duplicity of certain rows which we will deal later. The details of the column available in this data set is listed below.

1. App: This column contains name of the app in string format.
2. Category: Category of the app is defined as the broad subject that the app aims to cover for. These values are again mentioned in string format.
3. Rating: It reflects value of aggregated rating received by an app from all reviews. Logically this rating should not be greater then 5 and should not be greater then 1, but there may be case were app is never been rated in such case only app may contain null value for rating.
4. Reviews: It contains number of reviews given to particular app. The value mentioned is not a simple number but number prefixed with M or K, which stands for million and thousands respectively. The values are mentioned as string.
5. Installs: It specifies apps category based on number of installs made by user for particular app. No exact number is given instead categorization is done as 10+, 100+, 1000+ and so on.
6. Type: This column specifies whether particular app is paid or free. Value mentioned as string.
7. Price: It gives the value or price applicable to user in order to install or buy particular app. The values mentioned is float number prefixed with currency symbol, which in this case is $. The final value is mentioned as string.
8. Content Rating: It categorizes the app based on target audience age suitable to use app.
9. Genres: It categorizes app based on genres associated with the app. There may be multiple genres category on single app.
10. Last Update: This column mentions last update date of the app.
11. Current Version: The column specifies current version of app available on app store.
12. Android version: It specifies the least android version required to run the app on mobile.

**3.2 User Review Data**

This data set contains list of comments made during each written review on app. It has 64295 rows and 5 columns. Mostly all column appears to be standardized as per value content, but there appears lots of missing value or nun value which requires closer look. There is also large number of duplicities in data which is mostly associated because of missing values. Columns of this data set is described briefly below.

1. App: This column contains the name of app in short description. The format is similar to App data set and is in string.
2. Translated Reviews: This column contains translated reviews of users translated in English and is mentioned in string format.
3. Sentiment: This column mentions categorization of translated reviews based on fact that comment is positive, negative or neutral.
4. Sentiment Polarity: Score of review based on its polarity towards positive or negative sentiment. Most negative sentiment have value of -1 while most positive comment will get value of +1, 0 polarity means neutral comment.
5. Sentiment subjectivity: Score of review based on its content registering subject matters.

**4 Python and its Libraries.**

Data science is an advanced and scientific approach towards solving various statistical and business problems using historically collected data. It involves use of various scientific and data logics that can be trained or scripted to do repeated work. Python is one such strong language that facilitates logical scripting environment or coding facility to perform such tasks. Python is simplest programming language as compared to other language and hence most data scientist prefer same for caring critical analysis. With more then 137,000 libraries it has establish itself as a great source of tools and programming community. I have used python in exploration of given app data.

I have used 3 libraries within my analysis, namely NumPy, Pandas, Matplotlib and Seaborn. Pandas is used to carryout excel like works on data set. Pandas is used when we have structured data as input, it is built on NumPy. Other python library Matplotlib is used to set axis and figures and used to plot various statistical charts. Seaborn is an advanced form of Matplotlib and is build on Matplotlib itself, we have flexibility to execute charts directly from dataset using Seaborn methods and functions.

**5 Data Cleaning and Standardization**

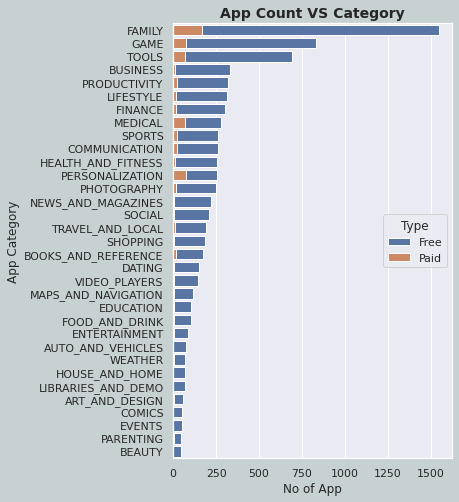
Both the dataset available with us have certain misleading values that need to be checked and sorted out. Mostly every data analysis task, starts with checking data for its value and application of necessary cleaning and exclusions of values, so that final data is cleaned one which will not land us on wrong conclusions. Also, in order to represent data on visual graphs the values need to be standard and should be of same type. Based on the primary and later inspection of data I have used below steps in order to convert data set into clean usable form.

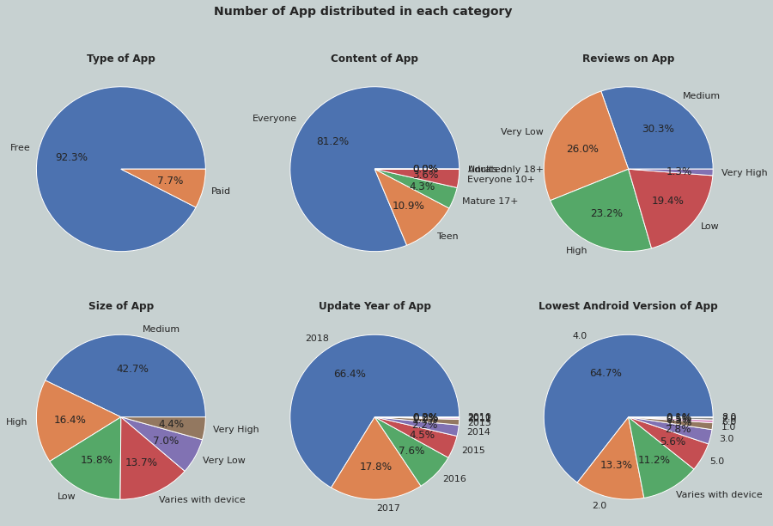
1. Inspecting App Data: I have used head() function to inspect kind of data available in 1st data set. It reflects first 5rows of structured data. By using info() method we got to know that data contains 10841 entries, different data type of each column, and number of null entries in each column.
2. Understanding Variables: With describe() method, I have tried to check out spread of numerical data along each column. By applying unique() method on each column we get the list of unique values mentioned in each column, and hence we can visualize the content of columns to take further necessary action.
3. Standardization of column variables: From above inspection we got to know that how columns contains unique data and how it need to be standardized. Size column contains details of size of app, but value is mentioned in string. I have defined function that will decode K to thousand and M to millions and convert actual size of all values into one-unit i.e. Millions, but the value will be in numbers. Similarly, number of reviews has been converted to numeric 1000 if suffix is K. Values in column installs contains value with + sign added in suffix, I have applied function to convert all values to numeric only. From price column I have removed dollar sign from values. Date available in Last update column is available as string, I have converted string into timestamp using strptime function.
4. Manipulated columns: Some columns have descriptive details that can not be converted into numbers or unique single data directly, I have manipulated them to convert their values either in single number or single value. Android version contains descriptive data of range of lowest android version to highest version which is required to run the app. I have captured only lower versions main version number and excluded all other details in new column naming it to Android version2. Tried to implement same thing on Current version but latter discovered that all app has different method of numbering version hence excluded this parameter from analysis. I have also manipulated genres to capture only one genre linked with app as primary genres.
5. Subcategorization based on values: I have also subcategorized apps based on value and its distribution within columns. With number of reviews being highly distributed throughout big span, app has been categorized into Very high reviews, High reviews, Medium Reviews, Low Reviews and Very low reviews. Similarly, I have categorized Rating and size of app too on similar principles.
6. Exclusion of Rows: There are certain rows that are not required for analysis that has been dropped out of data frame. There is one row that have size value mentioned as 1000+, which is dropped from data frame. There are 1173 apps with duplicate name, I have removed some of duplicate data by using drop\_duplicate() method and in next step shorted data for rating and excluded app name duplicates using drop\_duplicate() function considering app name as subset and keeping only first record. For App review data frame around 26868 rows do not have translated reviews, and hence has been dropped from final data frame.

**6. Exploratory Data Analysis**

The above cleaned App and Reviws data can now be utilized to get visual details of how the variable behave with each other. We have utilized Matplotlib and Seaborn liberaries to make EDA for this project. EDA is the process of generating summery statistics for numerical data in the dataset to discover outliers and patterns and chek for our hypothesis. In below sections we have created various graphical representation that will help in understanding app industry better.

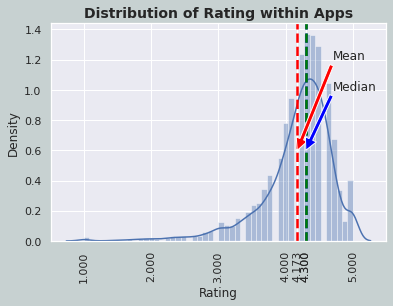
**6.1 Number of app in each category**



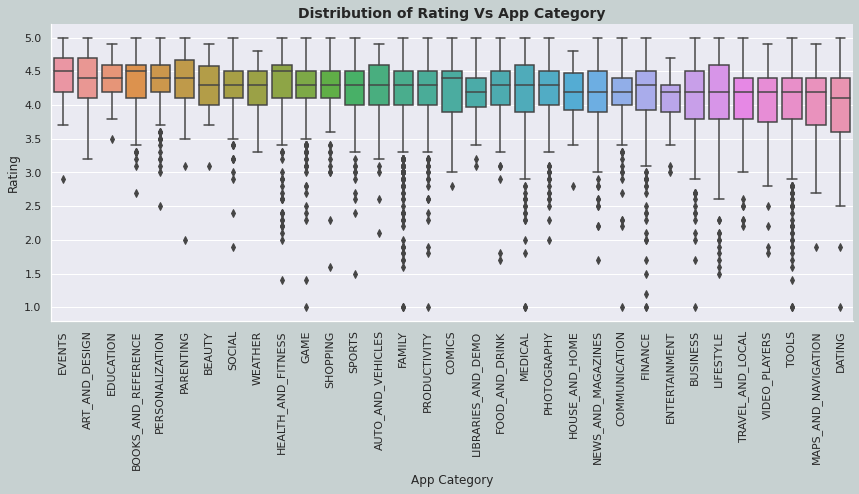


Category family have the highest number of apps dedicated to it followed by Game and Tools. Beauty have the lowest number of app. Also, contribution of paid app to free app is negligible. Ratio of paid app to free app appears higher in Personalization category. From pie chart we can also see the distribution of share of total apps among various app features. Nearly 92.3% app are free, 81.2% app belongs to audience of all age (Everyone), only 1.3% of app have got very high reviews ie above 4Million, 42.7% of the apps belong to medium size ie 5-30MB, most of the apps are updated in 2018 and most of the app requires minimum 4th version of android to run on mobile.

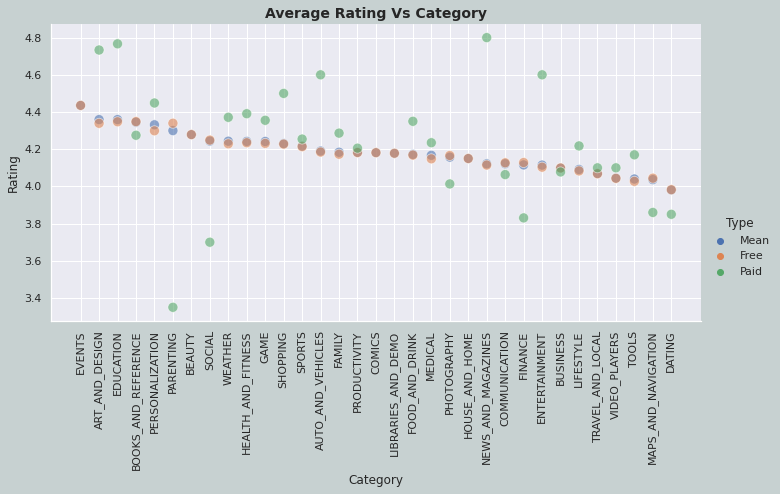
**6.2 Rating distribution**



Rating is normally distributed around 4.3 and lightly skewed towards left. Mean rating among all app is 4.17 while middle of app all rating is 4.3. From statistical analysis of rating distribution we can say that most of the apps have received higher ratings only. In order to check if above is valid for all category, we have used a plot of rating distribution against category.

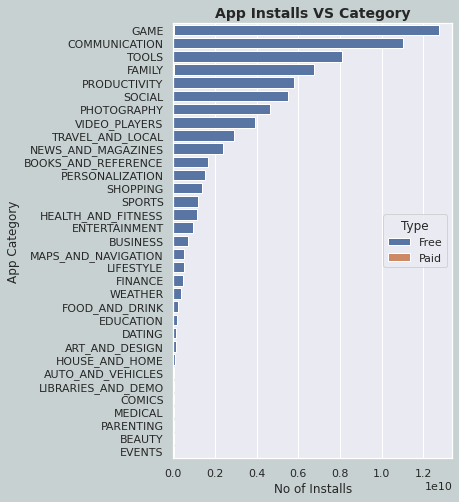


From this graph we conclude that statistical distribution of rating holds valid for all category if outliers are excluded.

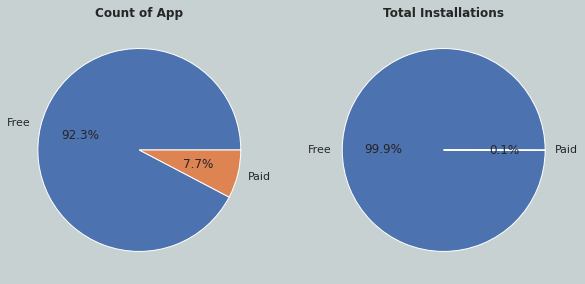


In above plot I have visualized comparison of average rating of app within category for only paid, free and combined. Since number of app that are free is comparatively high, it was obvious that mean rating for free app is equivalent to mean rating of free app which is visible in plot. While most of the paid app has rating higher then its free counter part but this is not valid for all category and there is an exception, like parenting, social and finance category.

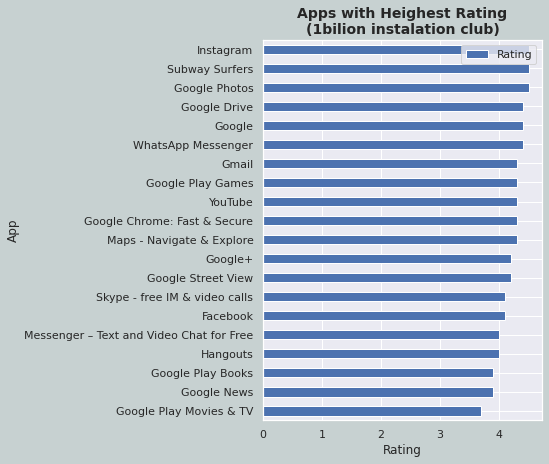
**6.3 Number of Installations**



Game has the highest number of installations, no doubt that’s why it has 2nd highest number of app on app store to offer. Communication and Tools follows there after for total number of installations.

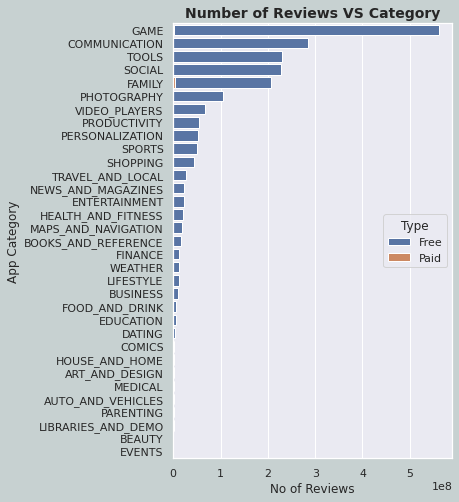


Although there are 7.7% of apps available as paid app but their contribution on total installation is only 0.1% of the total installation, and hence we can see most developers are offering free apps.

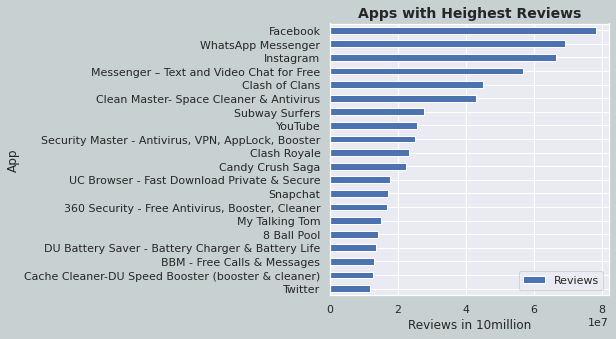


Since number of installations are categorized in multiples of 10 we have 20apps with highest number of installations or installation above 1Billion. Chart above shows these app with their rating.

**6.4 Apps with highest reviews**



Game have received highest number of reviews foolowed by Communication and Tools category. We can also note that contribution of reviews in paid is comperatively negligible and is not visible on graph except some trace in family category. Also if we observe closely the list is following similar order as that of category with number of installations.



Top 20 apps with highest number of reviews are mentioned in above chart. Facebook is the app which has received highest number of reviews followed by WatssApp and Instagram.

**6.5 Paid App Analysis**

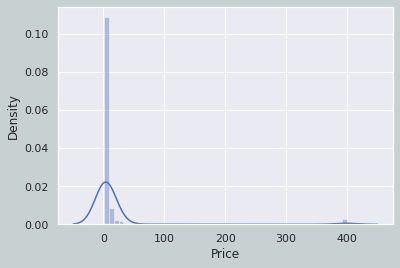
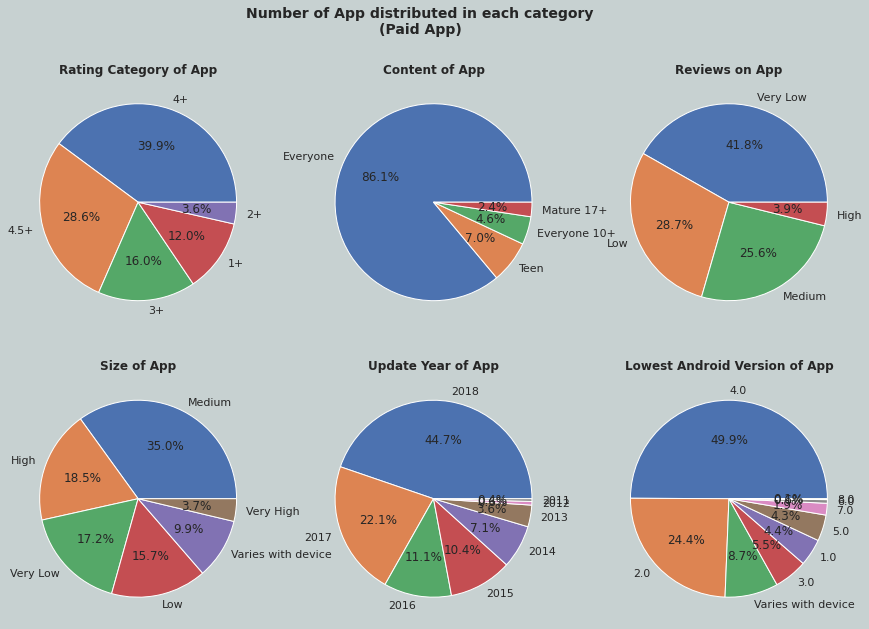
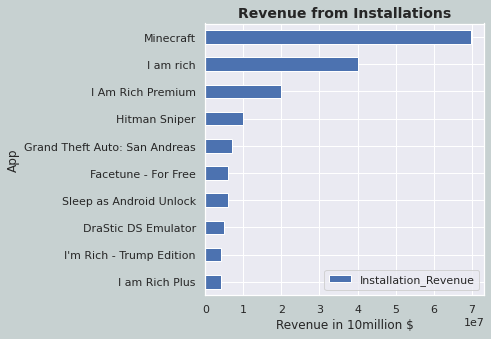


Chart above shows rate and density of app concentrated about rate. Rate of most of paid apps are concentrated between 0-1doller while some app have rate near 400dollers.

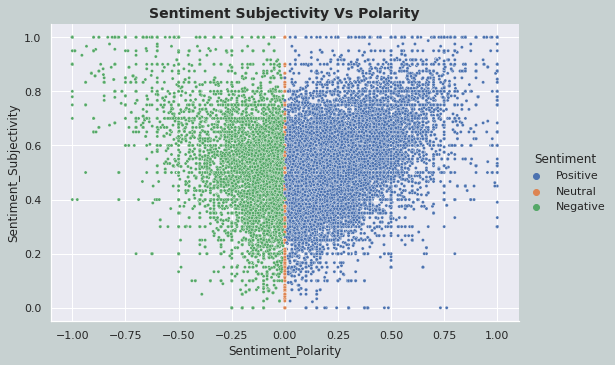


We can see distribution of share of number of apps in different category.

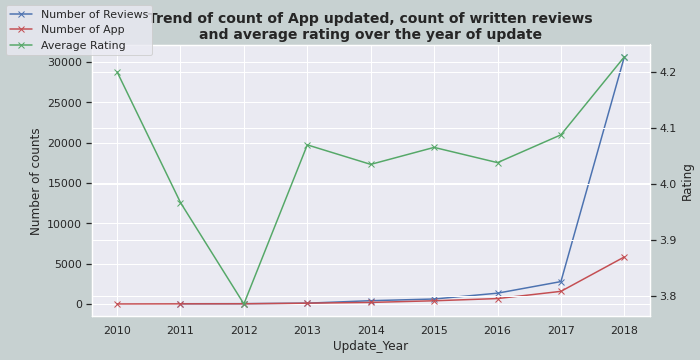


Minecraft is the app that has generated highest number of revenues from installation only, followed by I am Rich and I am Rich premium.

**6.6 App Review Analysis**

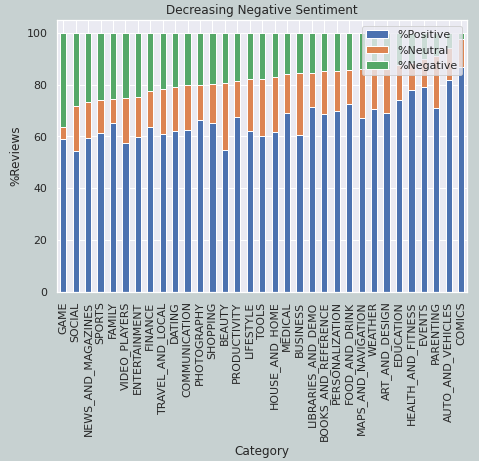


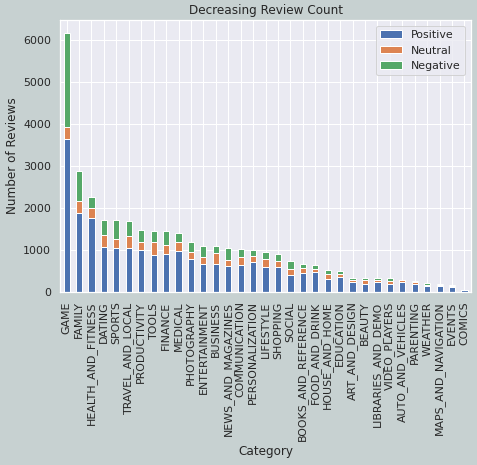
Sentiment\_Polarity have captured sentiments based on content being positive, negative or neutral. There is no clear relation between subjectivity and polarity.



In above graph I have plotted trend of number of app that has been updated, number of reviews on the app and the average rating of the app by year on year. We can see that Number of written reviews has increased drastically after 2017. Also average rating has shown a trend of improvement after 2016. 2012 was the update year of app that has recorded minimum average rating of 3.8.

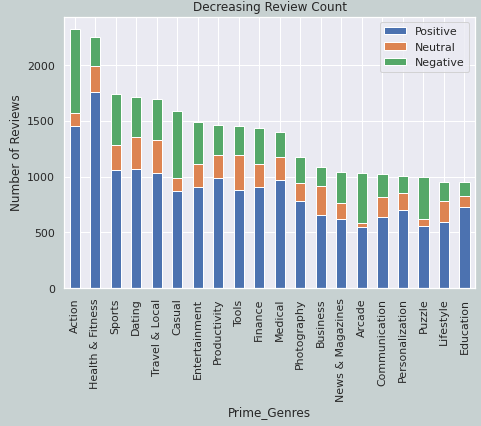
**6.6.1 Contribution of shares of positive, negative and neutral reviews on Category of app.**





Category game has received maximum number of written reviews followed by Family and Health&Fitness. Out of the total review received Game has received maximum share of negative reviews. Category Comics have maximum share of Positive+Neutral review, this is also the category which have received lowest number of reviews. By compering both chart, it can be concluded that users are more vocal and sttracted towards category Game but at the same time users have also recorded there pejorative for the apps available in this category.

**6.6.2 Genres with highest reviews.**



Action has received the highest number of reviews followed by Health&Fitness. Health&Fitness have also attracted highest percent of positive reviews. This is in line with our previous finding that category Game and Health&Fitness falls into top discussed category, of which above mentioned genres belong.

1. **Conclusion**
2. Reviews are highly correlated to number of installs.
3. Most of app have rating above 3.5
4. App suitable to everyone has heighset span of rating, while app belonging to adult only have smaller rating span.
5. 92% of total app on app store are free.
6. 81% of total app is suitable for everyone.
7. 42.7% of the app have size 5MB to 30MB.
8. Most of app are updated in the year 2018, or is lastly updated.
9. Most of the app requires android version 4.0 and above on any mobile to function.
10. Most of the paid apps are priced between 0-1$.
11. Minecraft have generated highest revenue from installations only.
12. Game, Social, News&Magzins are most negatively reviewed category as per % of share.
13. Game and Family have received most number of written reviews.
14. **Business applicable Conclusions**

* People prefer using free app most, but with advanced engagement features and usability, people are also ready to pay price if price is below 1$.
* To increase the scope of number of users it is better to launch app which is suitable for every age of audience.
* In order to maintain app relevant with user it is necessary that app is updated regularly. App that is lastly updated have more number of installs.
* Game is the most installed category of app at the same time it is the most negatively reviewed category by percentage, this signifies for existing market gap and hence this category shows potential of service improvement.
* Event is the least installed category that should be avoided by app maker because of its dimed demand.
* Dating is again most underrated category that have potential room for new apps.
* Game and Family are the top category which are discussed by user or have received written reviews. They hold good ground of interest from users point.

1. **References**

* pandas.pydata
* matplotlib.org
* seaborn.pydata
* Tradingeconomics
* Statista
* Wikipedia
* Stack Overflow
* W3Schools
* Investopidia
* Business of Apps
* Appypie
* Statology
* Educba
* Tutorialspoint.com
* Pythoncharts.com
* India.gov.in
* companiesmarketcap.com
* alltopeverything.com
* telecom.economictimes
* Sensortower
* Businessinsider
* TRAI
* Geeksforgeeks