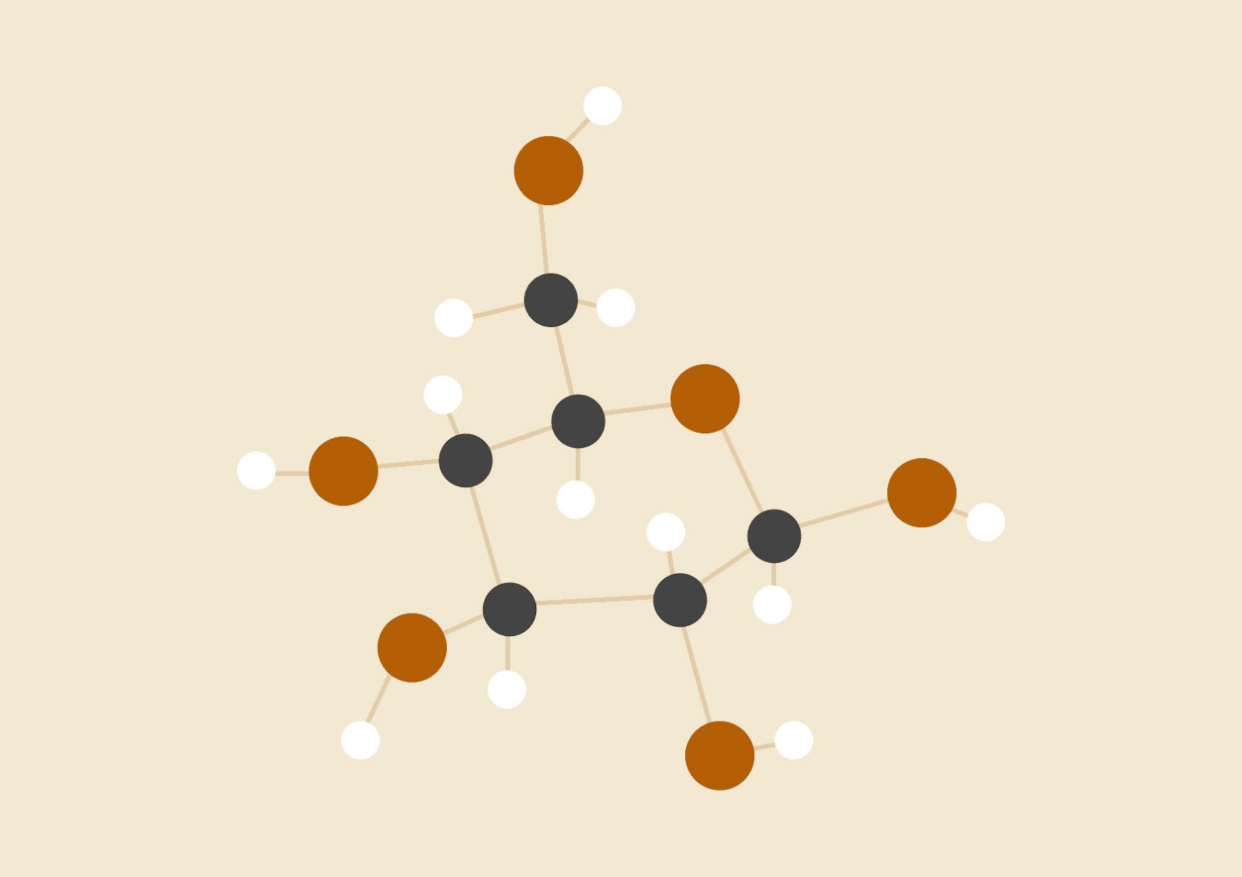
ANDROID MARKET APP STATISTICAL STUDY-2019

Category, Rating, Installs amount, Price and Reviews



Mingxing DING

05/01/2020

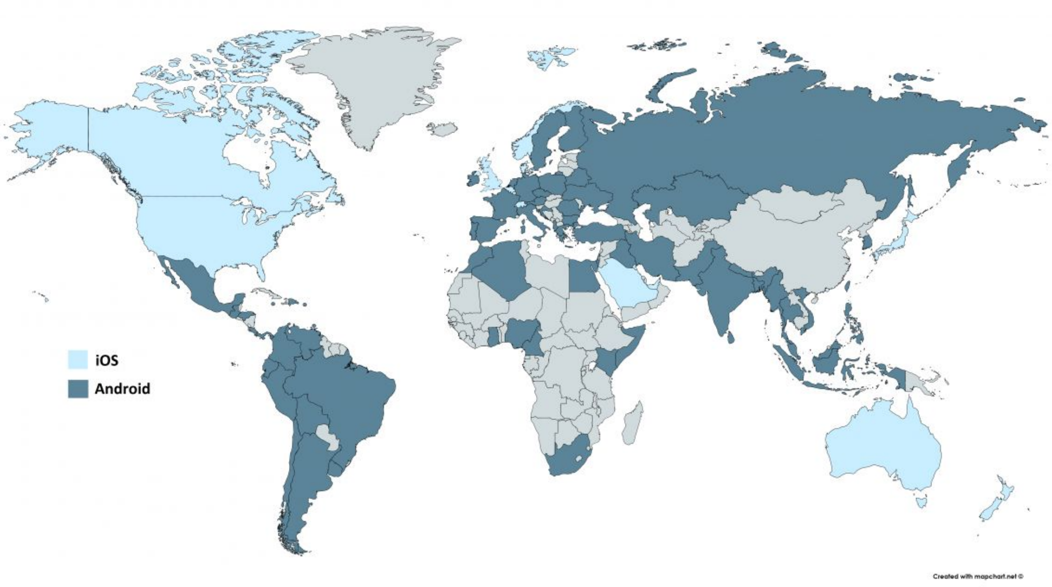
IBM Data Science

INTRODUCTION

1. Background

The two principal camps in the smartphone market, Android and IOS, in fact, have a sharp difference in the market shares. From a report of international data corporation IDC[1], in 2019, based on the global market, because of the release of 5G smartphone and the clearance of manufacturers, the market shares of Android has raised from 85.1% in 2018 to 87%. Correspondingly, there is a fall in the iSO part. But for Apple, if they release their 5G mobile phone in 2020, il will be a positive impact in terms of their market shares.

Otherwise, website Device Atlas[2] published a map of where Android (dark blue) and iOS (light blue) are the most popular mobile operating systems. (For grey countries, they don't have enough data to determine OS popularity).



These reports show that nowadays, Android still holds a large portion of the smartphone in the worldwide. The study of APP based on Android system will be strongly valuable for App development and marketing plan.

2. Interest

In view of the importance of Android market, it will be interesting for App making companies to learn about the actual situation of it before developing a new product, for example, the distribution of different category App, their average niveau of charge, what ranges are the rating and installs amounts and what are the characteristics of a good App. These aspects can help to better position the App in the market, find the potential opportunity and take advantage of the successful example.

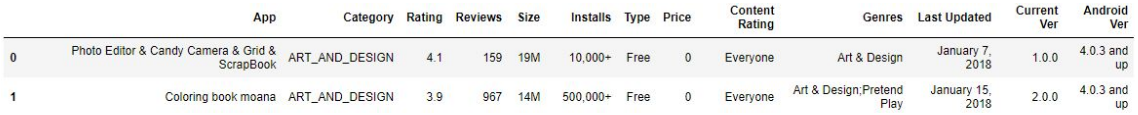
DATA

1. Data sources

Two datasets have been used in this study: 1- 'googleplaystore.csv' and 2-'googleplaystore_user_reviews.csv'. They are all download from Kaggle. The latest update of these datasets is on 2019-02-03.

2. Data cleaning

For the 1st dataset, there are 10,841 observations and 13 features in total. The attributes are: App name, Category, Rating, Reviews (reviews number), Size, Installs (installs amount), Type (free or paid), Price, Contents Rating (for everyone or for adults over 18 years old…), Genres (one App can belong to different category), Lasted Updated (date), Current version and Android version.



	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
0	Photo Editor & Candy Camera & Grid & ScrapBook	ART_AND_DESIGN	4.1	159	19M	10,000+	Free	0	Everyone	Art & Design	January 7, 2018	1.0.0	4.0.3 and up
1	Coloring book moana	ART_AND_DESIGN	3.9	967	14M	500,000+	Free	0	Everyone	Art & Design;Pretend Play	January 15, 2018	2.0.0	4.0.3 and up

At first, several columns have been removed because they are not in the scope of the this study. They are Reviews, Genres, Last Updated, Current version and Android version.

A missing values check shows that there are respectively 1474, 1, 1 vacants in Rating, Type and Content Rating columns. Recheck will be done after the following treatment.

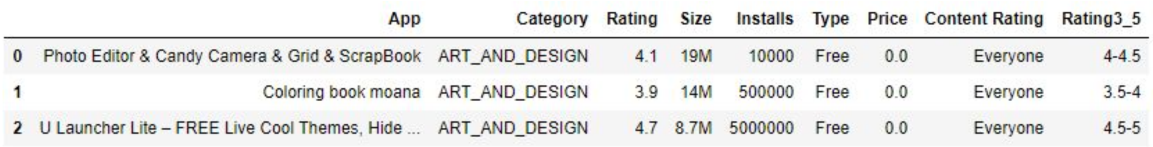
For the Content Rating feature, there are several distinct values: everyone, Teen, Mature 17+… and the category 'everyone' is predominant (8714/10,841). In order to normalize this feature, only the 'everyone' category is kept.

After this step, there isn't any missing value in Type columns and the row with one missing value in Content Rating value is removed. The missing data in Rating columns are not touched in this stage because we'll see later that they don't bother this study.

So now it rests 8714 rows in all but there are only 7903 distinct App names. This means that there are rows with the same App name. After inspection, it shows that sometimes there are several rows with one App name but without difference for the rest features. So only one row will kept for these Apps.

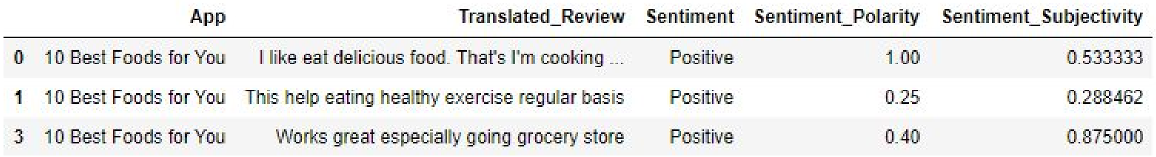
Another new column 'Rating 3-5' is added in order to classify the Apps into 'NoRatingInfo', '<=3', '3-3.5', '3.5-4', '4-4.5' and '4.5-5' items.

After these treatment, we finally get a data set of 7903 rows and 8 features.



	App	Category	Rating	Size	Installs	Type	Price	Content Rating	Rating3_5
0	Photo Editor & Candy Camera & Grid & ScrapBook	ART_AND_DESIGN	4.1	19M	10000	Free	0.0	Everyone	4-4.5
1	Coloring book moana	ART_AND_DESIGN	3.9	14M	500000	Free	0.0	Everyone	3.5-4
2	U Launcher Lite – FREE Live Cool Themes, Hide ...	ART_AND_DESIGN	4.7	8.7M	5000000	Free	0.0	Everyone	4.5-5

For the 2nd dataset 'googleplaystore_user_reviews.csv', it has 64,295 observations and 5 features initially. A missing value check shows that there are 26,868 rows having no reviews information which is the main study target by this dataset and most of these rows don't have any values in other columns, either. So all this row are deleted. At last, we got a dataset in size (37427, 5) , having 865 distinct Apps and no missing values.

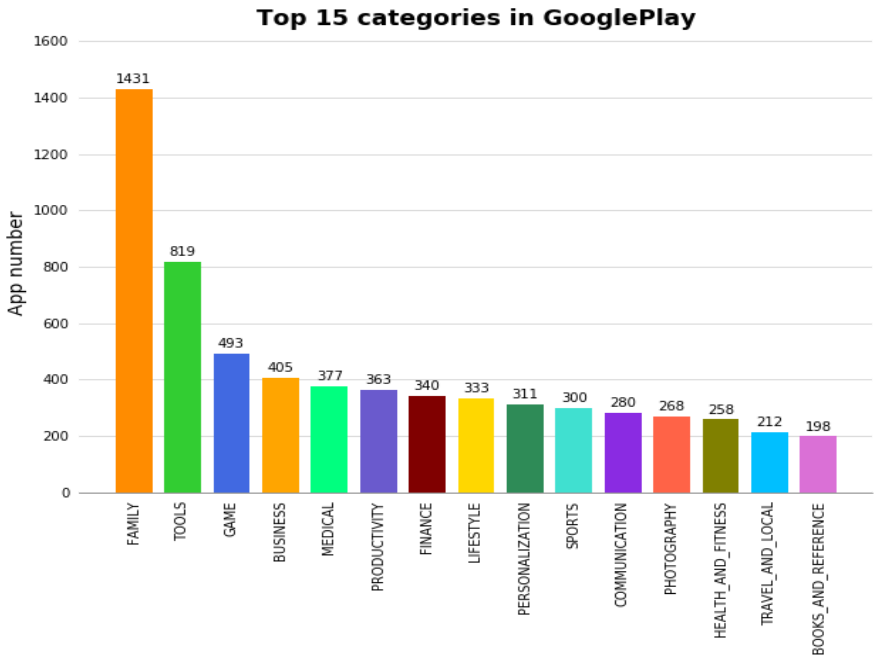


	App	Translated_Review	Sentiment	Sentiment_Polarity	Sentiment_Subjectivity
0	10 Best Foods for You	I like eat delicious food. That's I'm cooking ...	Positive	1.00	0.533333
1	10 Best Foods for You	This help eating healthy exercise regular basis	Positive	0.25	0.288462
3	10 Best Foods for You	Works great especially going grocery store	Positive	0.40	0.875000

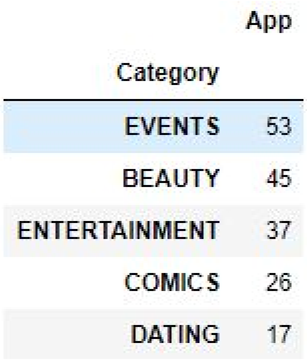
METHODOLOGY

1. Category distribution

There are in total 33 categories. The Top 15 numerous categories are as follows:



Compared to the numerous App categories, there are also some familiar categories having just few available Apps. The last 5 are:



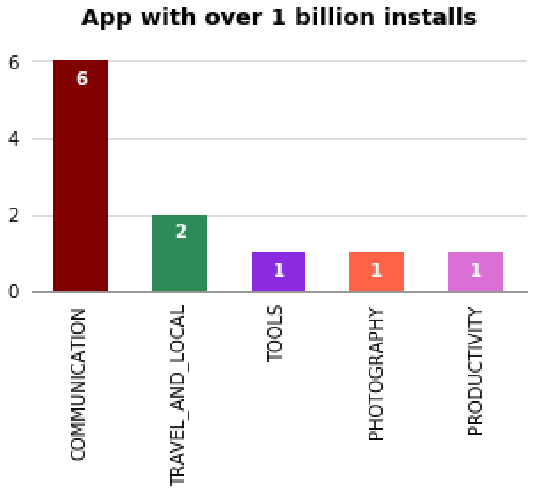
Category	App
EVENTS	53
BEAUTY	45
ENTERTAINMENT	37
COMICS	26
DATING	17

2. **Installs amount statistics and relationship with Category**

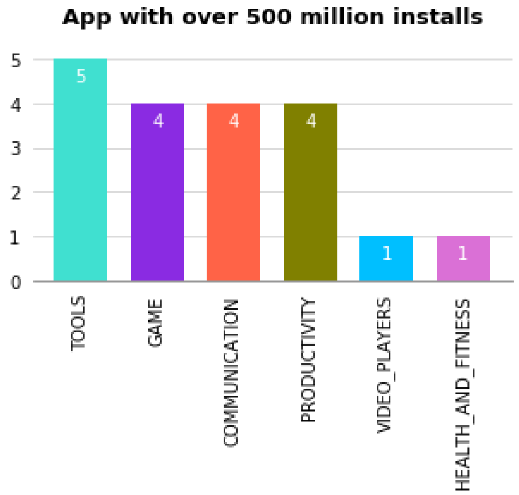
In this dataset, the installs amount is represented in the form of xxxxxx +, which is in object type. For facilitating the comparison, the datatype of this feature is changed to integer. There are 11 App having over 1 billion install and 19 App having over 500 million installs.

So what kind of App has so many installs?

For those 11 Apps who have over 1 billion installs, the COMMUNICATION category has occupied more than a half (6). And for those Apps who have over 500 million installs, TOOLS, GAME, COMMUNICATION and PRODUCTIVITY categories are evenly predominant.



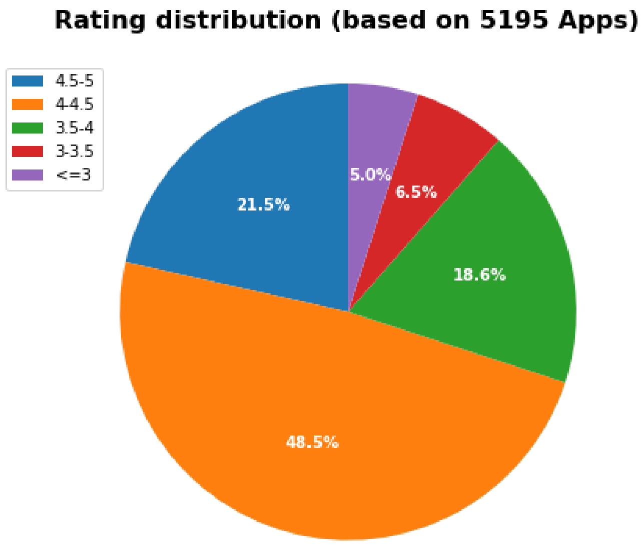
App with over 1 billion installs



App with over 500 million installs

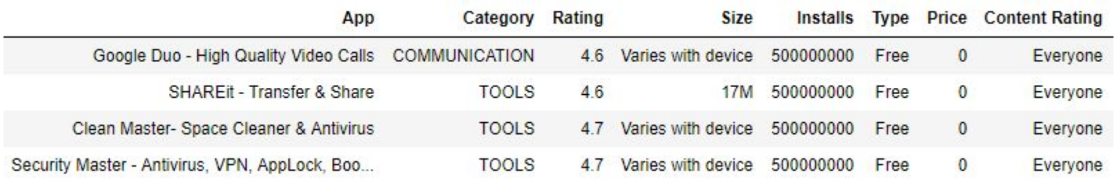
3. App Rating analysis and combination study with large installs amount

Like described in the Data section, in the cleaned dataset, there are still 1285 missing rating value. Apart from this part, there are 5195 Apps for this Rating analysis. Moreover, most of these Apps have a Rating over 3, so the analysis detailed the range of 3 - 5. Below is the distribution of Apps in different range of Rating:



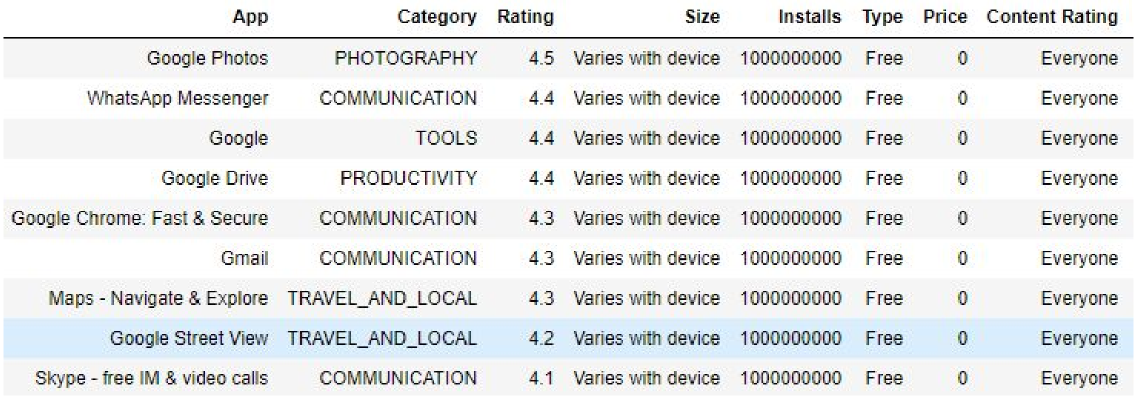
Rating distribution (based on 5195 Apps)

The Apps with large installs amount and high rating will be a good example as a successful product and are worth a special attention and deepen analysis. There are 4 Apps who have over 500 million installs and have a rating in range 4.5 - 5:



App	Category	Rating	Size	Installs	Type	Price	Content Rating
Google Duo - High Quality Video Calls	COMMUNICATION	4.6	Varies with device	500000000	Free	0	Everyone
SHAREit - Transfer & Share	TOOLS	4.6	17M	500000000	Free	0	Everyone
Clean Master- Space Cleaner & Antivirus	TOOLS	4.7	Varies with device	500000000	Free	0	Everyone
Security Master - Antivirus, VPN, AppLock, Boo...	TOOLS	4.7	Varies with device	500000000	Free	0	Everyone

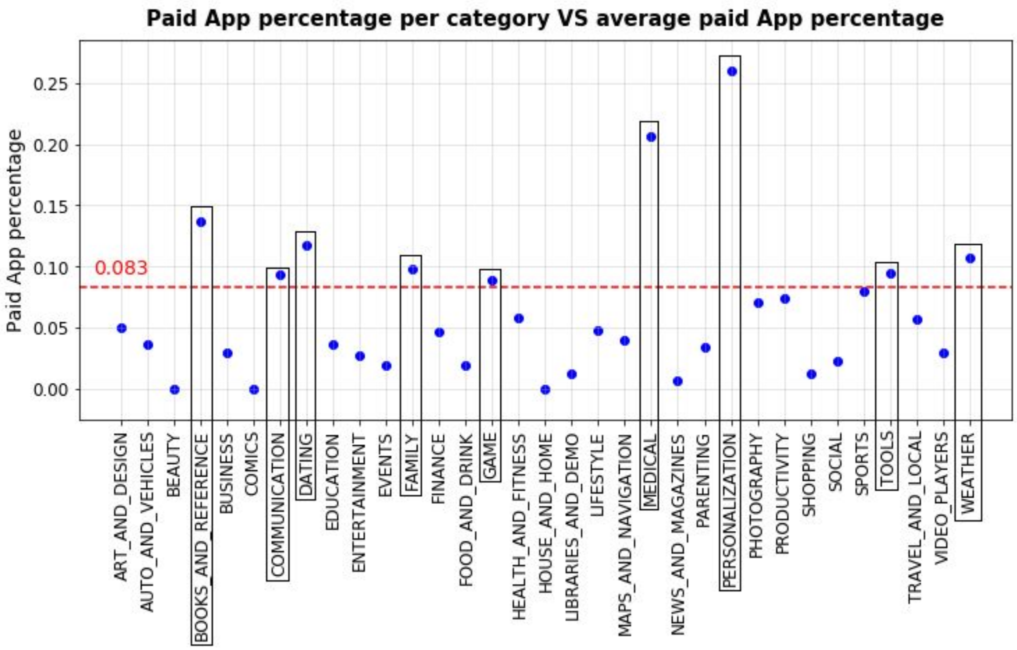
And there are 9 Apps who have over 1 billion installs and have a rating in range 4 - 4.5:



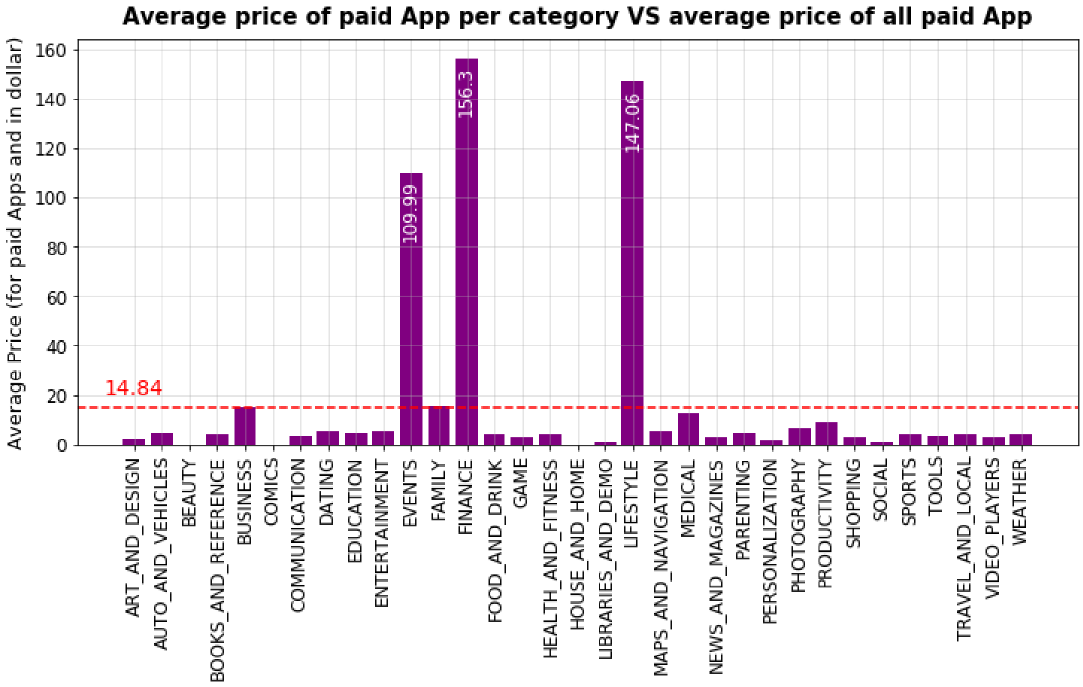
App	Category	Rating	Size	Installs	Type	Price	Content Rating
Google Photos	PHOTOGRAPHY	4.5	Varies with device	1000000000	Free	0	Everyone
WhatsApp Messenger	COMMUNICATION	4.4	Varies with device	1000000000	Free	0	Everyone
Google	TOOLS	4.4	Varies with device	1000000000	Free	0	Everyone
Google Drive	PRODUCTIVITY	4.4	Varies with device	1000000000	Free	0	Everyone
Google Chrome: Fast & Secure	COMMUNICATION	4.3	Varies with device	1000000000	Free	0	Everyone
Gmail	COMMUNICATION	4.3	Varies with device	1000000000	Free	0	Everyone
Maps - Navigate & Explore	TRAVEL_AND_LOCAL	4.3	Varies with device	1000000000	Free	0	Everyone
Google Street View	TRAVEL_AND_LOCAL	4.2	Varies with device	1000000000	Free	0	Everyone
Skype - free IM & video calls	COMMUNICATION	4.1	Varies with device	1000000000	Free	0	Everyone

4. Type and Price analysis

The dataset is grouped by the category. The paid App percentage and the average price for those paid Apps are also calculated, along with the total paid Apps percentage and total average price (of all those paid Apps).



In the schema above, the red line presents the global paid App percentage. Among the categories who have a paid App percentage larger than the average one, MEDICAL and PERSONALIZATION categories are especially remarkable.



Among the categories who have a much higher price than the global average price, it should pay attention to EVENTS category, because the result is based just on one paid App in this category, meanwhile, for FINANCE and LIFESTYLE category, they are both an average of 16 Apps in the category.

5. **Analysis of customs' reviews for popular Apps**

From the first 3 numerous categories: FAMILY, TOOLS and GAME, one high rating and large installs amount App from each is chosen as an representable example for analysis. For each App, the distribution of positive, negative and neutral reviews is plotted. In this dataset, there are two features which may also be interesting: sentiment polarity and sentiment subjectivity.

Sentiment polarity, also known as orientation, is the emotion expressed in a sentence. It can be positive, negative and neutral. And sentiment subjectivity express some personal feelings, views and beliefs. With TextBox, the polarity score is a float within the range [-1.0, 1.0]. The subjectivity is a float within the range [0.0, 1.0] where 0.0 is very objective and 1.0 is very subjective. Even though we can't confirm if this labelling standard is used in the dataset studied, we can still have a rough impression of the polarity and subjectivity of the reviews. So for each app studied, the average sentiment polarity and subjectivity are also calculated.

Besides, the positive reviews and negative reviews are examined by using Word Cloud in tentative to find some remarkable characteristics of these Apps.

For the Family category, it's 'Build a Bridge' who has been chosen:

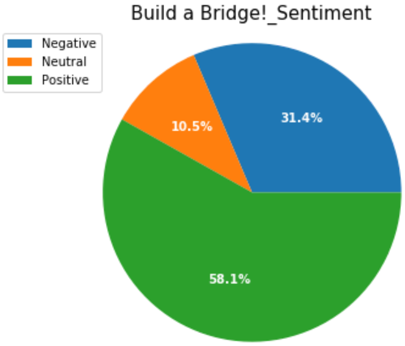
Rating: 4.6

Installs amount: 10 million +

Quantity of reviews: 86

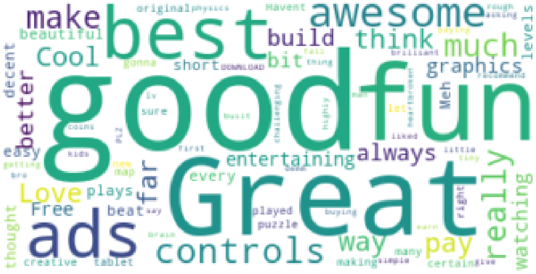
Average sentiment polarity: 0.09

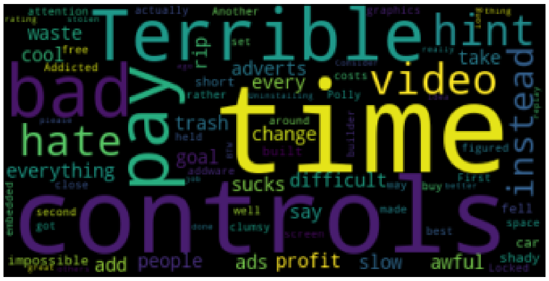
Average sentiment subjectivity: 0.51



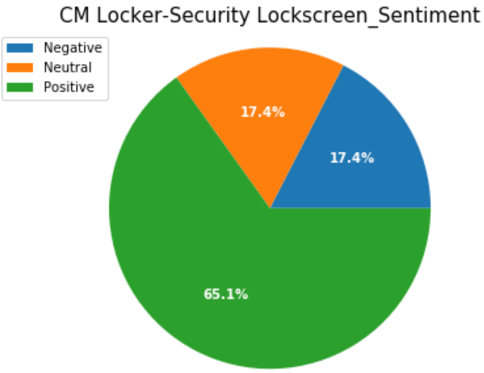
Word Cloud of positive reviews (left) [stopwords: "game", "play", "bridge",'level']

Word Cloud of negative reviews (right) [stopwords: "game", "play", "bridge", 'level','much','levels','watch']





For the Tools category, it's 'CM Locker - Security Lockscreen' who has been chosen:



Rating: 4.6

Installs amount: 100 million +

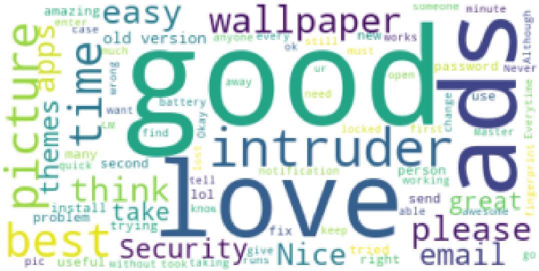
Quantity of reviews: 86

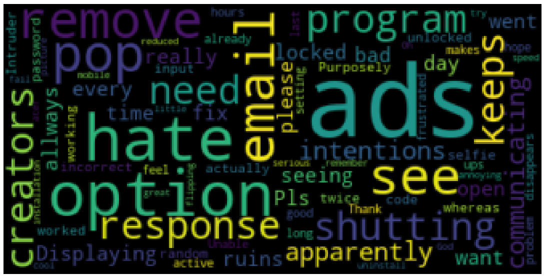
Average sentiment polarity: 0.24

Average sentiment subjectivity: 0.48

Word Cloud of positive reviews (left) [stopwords: phone", "screen", "even", "really", "locker", 'lock','unlock','app','lockscreen']

Word Cloud of negative reviews (right) [stopwords: "lock", 'screen', 'lockscreen', 'unlock', 'phone']





For Game category, it's Hill Climb Racing who has been chosen:

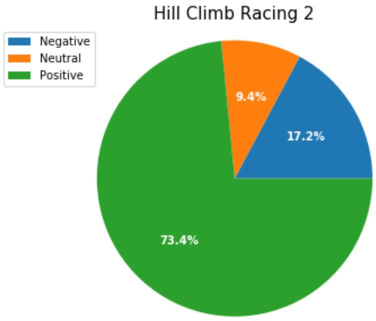
Rating: 4.6

Installs amount: 100 million +

Quantity of reviews: 64

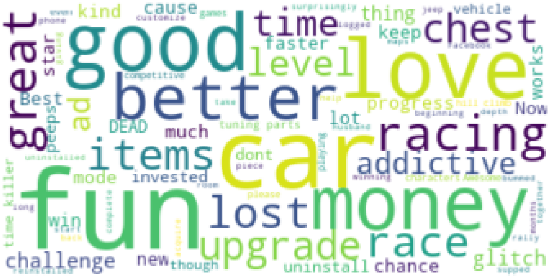
Average sentiment polarity: 0.16

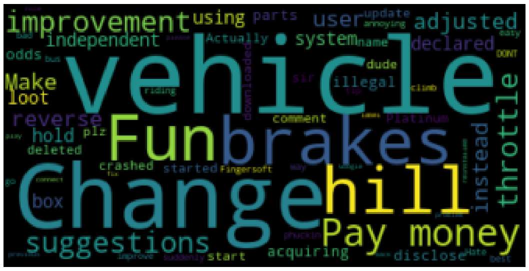
Average sentiment subjectivity: 0.48



Word Cloud of positive reviews (left) [stopwords: 'game', 'make', 'play', 'really', 'need', 'many', 'way']

Word Cloud of negative reviews (right) [stopwords: 'game', 'fully']





RESULTS AND DISCUSSION

1. In the category distribution part, we get that the N°1 numerous category is FAMILY and the second is TOOLS. The two categories have much more Apps than any other categories. Then from the 3rd to the 15th category, the difference is more gradual. In this study, we can also find that in google play store, there are still few Apps in categories like Dating, Comic and Entertainment. However, these categories have actually a realistic demand in our daily life. So it may be interesting to focus in this field in order to produce high quality and practical Apps.
2. In the installs amount analysis, among those large installs Apps, interestingly, the FAMILY category, which has most available Apps in the market, doesn't have impressive installs amounts. Instead, it's the category like COMMUNICATION, GAME, TOOLS, etc. who have Apps that achieve extremely high installation. This result corresponds well to our daily use habit of smartphone.

3. In the Rating analysis part, where the study is based on 5195 rating data available Apps, we found that about a half Apps (48.5%) have a rating between 4-4.5. 70% Apps can achieve a rating above 4. This means that in the general market, most of the products are very competitive. We find four Apps : Google Duo, SHAREit, Security Master and Clean Master, they all have a rating between 4.5 - 5 and over 500 million installs. And if we look for the Apps having a rating in range 4 - 4.5 and over 1 billion installs, we get: Google photos, Whatsapp, Google, Google Drive, Google Chrome, Gmail, Maps, Google Street View and Skype. It's obvious that the google series apps occupy a large part of these high rating and large installs Apps. Apart from that, we can find the App like Whatsapp, Skype and apps for mobile security and cleaning.
4. The Type and Price analysis shows that over all this dataset, 8.3% Apps are paid and their average price is 14.84 dollars. The paid Apps percentage higher than the average level are categories: BOOKS_AND_REFERENCE, COMMUNICATION, DATING, GAME, MEDICAL, PERSONNALISATION, TOOLS and WEATHERS. Among them, MEDICAL and PERSONALISATION category have remarkable higher paid percentage: 20.69% and 26.05% . In terms of average price, there are EVENTS, FINANCE and LIFESTYLE categories who have much higher price that the global average level: 109.99, 156.3 and 147.06 dollar respectively. But it should note that the results of EVENTS category is based only on one App. But for the other two categories, they have both 16 paid Apps for calculating this average. So the high price is therefore very impressive.

   The result of this part can give the developer a good reference when they decide if their App is going to be paying and if yes, how to position their price in a way to maximize the profit and not go too far away from the category average level.

5. In the review analysis part, 3 Apps chosen are the successful one in their category. They can get a positive reviews percentage up to 73.4% (Hill Climb Racing 2). It's Build a Bridge! who has a lowest positive reviews percentage among these 3, which is 58.1%. This result corresponds to its low average sentiment polarity compared to others two studied Apps, which is 0.09. The idea of Word Cloud study is to discover the characteristics of these Apps from their positive and negative reviews. Unfortunately, the result is not like expected. Because all the high frequency words that we can get are relatively 'neutral', like good, great, fun, best, etc. or terrible, hate, pay, etc. for negative reviews. It's difficult to get specifics directive points in technical level and product level.

CONCLUSION

From this statistical study based Apps in Android market, we can have a good view of the category distribution, the rating and installs amounts insight, the type and price market situation, etc. These can give App developing companies a vision of the market and help them choose a target app category and better position their product in terms of price. However, the reviews analysis doesn’t get an expected result. This is logical because the positive, negative and neutral is classified just by the emotional word like like, fun, hate, horrible, etc. So when we try to present the reviews in the word cloud form, we can get this kind of result. So if we want to discover the product-self level insight of the App, maybe it will need to remove the emotional type words.

Otherwise, it should be note that even we use Google Play dataset for the study of Android market, it’s not complete. For example, in a huge Android market in the world - China, the Google Play is generally not available. So in order to have a fully study of the Android market in a global scope, we need to combine other Android Apps download platform.

REFERENCE

[1] https://www.expreview.com/70381.html

[2] https://deviceatlas.com/blog/android-v-ios-market-share