A Hot App is Born - Visualize Google Play Store Data

Group 6:
Yijun Chen - yc3166(In class)
Chaochen Wu - cw2661(In class)

Group 7
Qi Sun - qs479(Online）

Project Description

With the AppStore and Google Play now serving millions of people across the world, and new apps entering the market every day, individual app downloads start to reach millions in just a few days. People like to live at speed these days. Teenagers and students are being told more and more that they have to accomplish goals in their early years and having a productive day at work usually means how fast you completed your tasks. Based on that logic, apps are keeping people at the speed they need, and bringing out new and improved versions of programs every day.

Apps are popular because they're convenient, and constantly innovating. They changed the world when they first started being developed, and now we can't break away from the model. Worldwide number of mobile app downloads in 2017 reached 178.1 billion, and so far of 2018, it already reached 205.4 billion. In Google play store and Apple store, there are thousands of Apps, what makes an App successful? What's the standards of a successful App? Should we set our target users as everyone, or set up specific target users like teenages/students, children? What type of App do we want to develop: financial, education, entertaining? Which aspects will affect the popularity of Apps: size, cost or the name?

Therefore, the goal of our project is to analyze what makes an App successful? Questions been mentioned in previous paragraph are what we need to solve in order to create a successful App.

Datasets

In our project, we will use the Google Play Store Apps data set which contains App information in the Google Play Store. Here is attributes that we will use in our project.
App, type: ***Categorical***
The App name, short text.
Category, type: ***Categorical***
The App category, including business, beauty, art … etc.

Rating, type: ***Quantitative***
The App rating by users, from 0 to 5.
Reviews, type: ***Quantitative***
The App reviews number, from 0 to infinity.
Size, type: ***Quantitative***
The App size, express by MB or *Varies with device*.
Installs, type: ***Ordinal***
The App installs number, express by range, for example 100+, 1,000+, 10,000+, and etc.
Type, type: ***Categorical***
The App is free or paid, only two categories.
Content Rating, type: ***Ordinal***
The App's content rating, which restrict specific user to install, including *Everyone*, *Teen*, and *Mature 17+*.

Analytical Questions

Question 1:
Who are our target users?
Proxy Tasks:
Identify the correlation between App install number and Content rating.
Proxy Values:
Content rating -> Content rating, App install number-> installs or reviews.

Question 2:
What kind of APP is more attractive to our target users?
Proxy Tasks:
Compare install numbers for different categories of App.
Proxy Values:
categories of App -> category, install numbers ->installs

Question 3:
Can we have more detailed analysis of target users' preferred category?
Proxy Tasks:
Identify the correlation between top 3 target users and 5 genres.
Proxy Values:
Target users -> top 3 content rating
genres -> category.

Question 4:
Does name contribute to APP popularity?
Proxy Tasks:

Identify the correlation between popularity and names length of APP
Proxy Values:
Names length -> name, popularity -> installs/reviews

Question 5:
Is free downloading always better?
Proxy Tasks:
Compare the popularity of APP with different access type both free and non-free
Proxy Values:
Access type -> type, popularity->installs/reviews

Question 6:
Is high rating App always popular?
Proxy Tasks:
Identify the correlation between rating and popularity
Proxy Values:
Rating -> Rating, popularity->installs/reviews

Data Analysis

Question 1: We first aggregate records by Content Rating, and we find 6 types of content rating, including Adult only 18+, Everyone, Everyone 10+, Mature 17+, Teen, and Rate Pending. Because Adult only 18+ and Rate Pending had only few records, 3 and 2 respectively, so we won't not consider them in visualization. Then we calculate average reviews number for each rating. We find that Everyone 10+ and Teen have higher number of reviews compared with other Content Rating. Apps for Mature 17+ has relatively less number of reviews.

Question 2: Then we investigate the correlation between reviews number and App's category. By aggregate records by Category, there are 33 categories. The category with the highest number of records is FAMILY (1972), and the one with the lowest number of record is COMICS and PARENTING (60). By calculating the average number of Reviews we find that SOCIAL, GAME, and COMMUNICATION category have high reviews number, and BEAUTY, MEDICAL, and EVENT have relative low reviews number. This result is intuitive because almost everyone's smartphone have Facebook, Whatsapps, and some games.

Question 3: As we aggregate rating and category data in Question 1 and Question 2, we aggregate records by rating and category information together. Because some ratings have extremely high or low numbers of observations, we choose three ratings which have similar numbers of observations (Teen, Everyone 10+, Mature 17+). Some categories have a small number of records, lower than 10, so we chose the top 5 categories which have the highest number of records for each rating. Our result shows that Social Apps with Everyone 10+ rating have the highest average review numbers. Interestingly, Apps in the same categories have different average reviews number with different ratings.

Question 4: We also want to investigate name length effect on Apps reviews number. We collect words number of each App name. For App name which longer than 10 words we treat them as one group. Then we calculate average reviews number of each name length. The result shows 10 words name has the highest average reviews. Apps with one word names are the least popular group, which 1 / 7 of Apps with 10 words name.

Question 5: We first divide records to two groups by their types (Free or Paid), we find that most of Apps are free for download, only about 800 of Apps are paid to download. Then we calculate mean of reviews number, we find that free downloading Apps have higher average reviews compare with paid downloading Apps. We also study the correlation between Apps prices and their downloading times. Apps with price higher than 5$ have small reviews number. Some paid Apps with low cost is hot, but most of them have low number of reviews.

Question 6: It’s possible that an App has 4.7/5 rating but only 10 users, or over 20K downloads but only has 2.5/5 rating. Therefore, it’s important to study the correlation between rating and number of downloading times. We are going to projection all the Apps in dataset as dots in our visualization graph, each dot is a representative as an App, X-axis is a mark of App rating, and Y-axis as App downloading times. By study the distribution of dots, we are able to find the relationship of rating and number of downloading times.

Sketches

Question 1

Information our solution will contain:

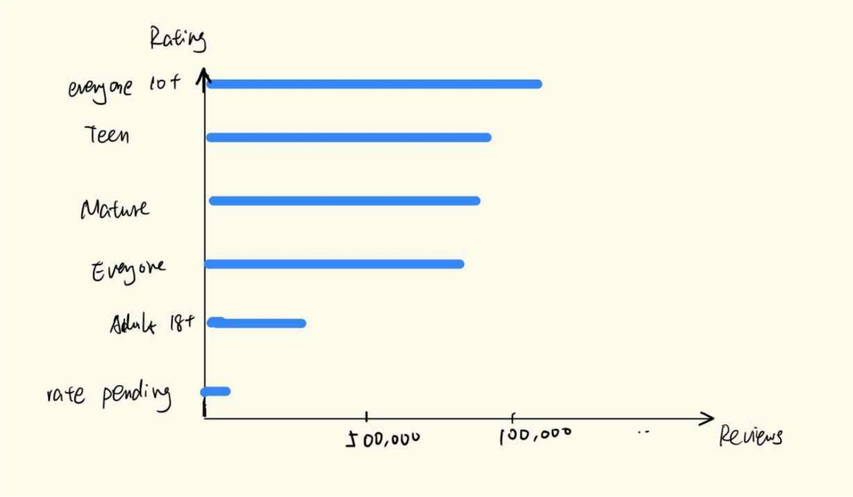
App install number

Content rating

Highlighted insights:

X axis represent different age range of target users.

It includes five groups: Adult only 18+, Everyone, Everyone 10+, Mature 17+, Teen, and Unrated. Y axis represent the number of App installation of each age group.



Question 2

Information our solution will contain:

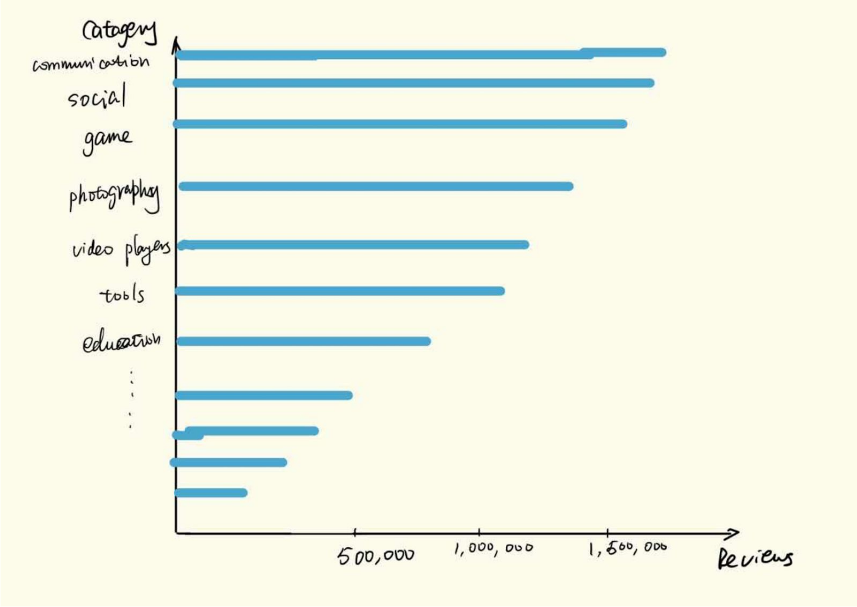
Categories of App

Number of installation

Highlighted insights:

We briefly group kinds of potentially app in the store into 32 categories: Art and design, Vehicles, Beauty, Communication, Education, Entertainment, Game, Weather, etc. Different bars represent different kinds of Apps.

The height of bars represents the number of installation of different kinds of App. The taller a bar is, the more popular this kind of App is.



Question 3

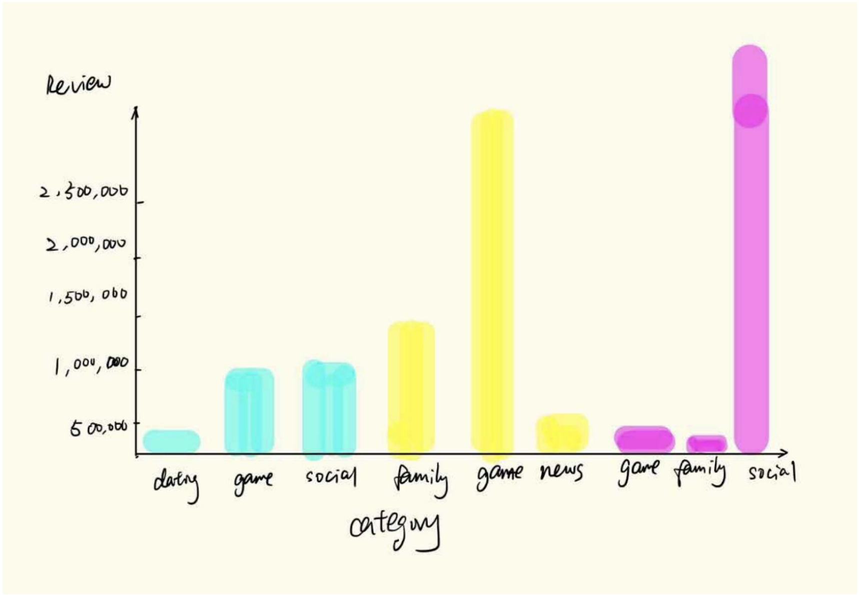
Information our solution will contain:

Number of installation

The size of App

Highlighted insights:

We divide the Apps by different number of installations. For each group, we use box plot to represent the distribution of Apps of different size. The more intensive of the slots, the more Apps belongs to that range of the size.



Question 4

Information our solution will contain:

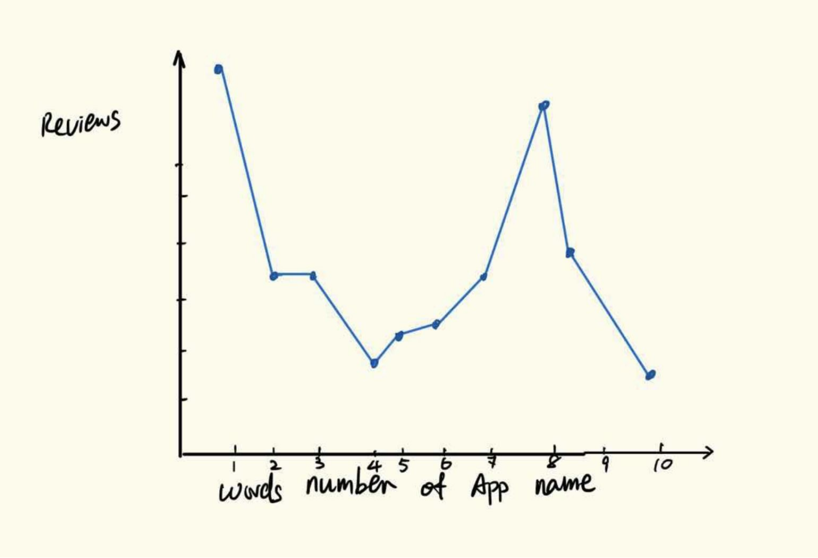
Length of App name

Number of installation

Number of reviews

Highlighted insights:

We use the number of installation and reviews to represent the popularity of an App. We group the App by different length of their names. X axis represents different groups of App of different length. Y axis represents the number of installation and reviews.



Question 5

Information our solution will contain:

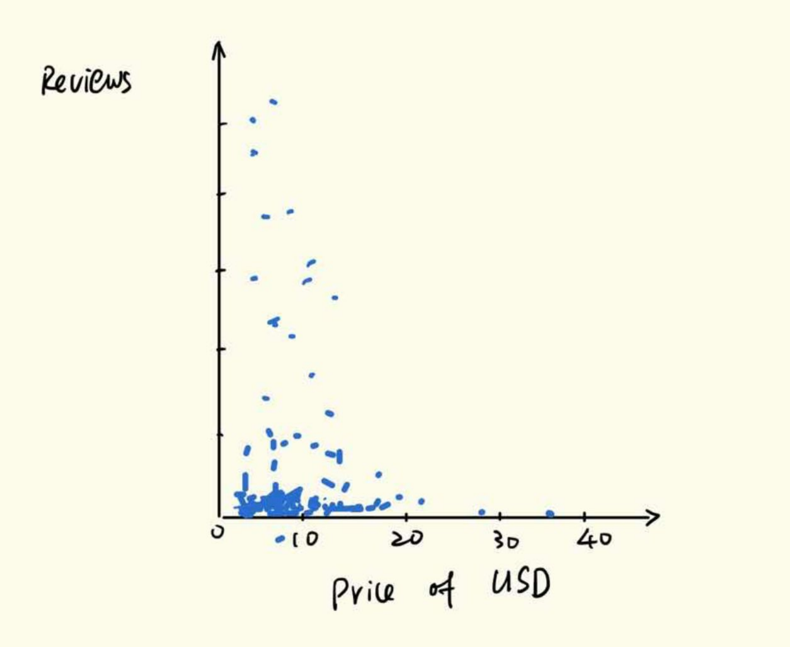
Access type

Number of installation

Number of reviews

Highlighted insights:

Y axis represents the number of installation X axis represents the price of App



Question 6

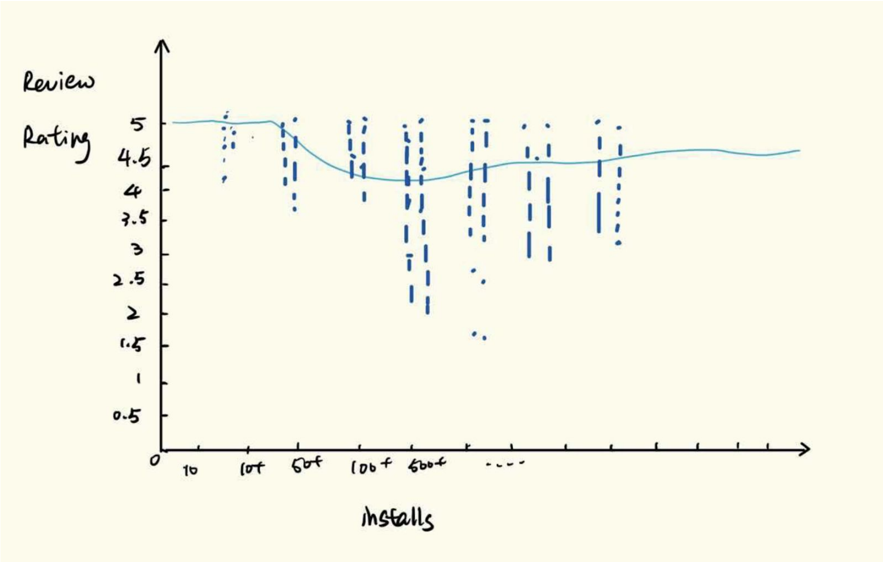
Information our solution will contain:

App rating App downloading times

App divided by genres

Highlighted insight:

The Y axis this Review Rating (from 0 - 5). The X axis is App installs number. Each App is represented by a dot, and the line shows median values.



Storyboard

AppStore and Google Play now serving millions of people across the world, and new apps entering the market every day, individual App downloads started reaching millions in just a few days. People can't live without their phones and apps. The first thing in the morning is to reach the cell phone and check messages and news. People like to live at speed these days. Teenagers and students are being told more and more that they have to accomplish goals in their early years and having a productive day at work, which usually depends on how fast you completed your tasks. Based on that logic, apps are keeping people at the speed they need, and bringing out new and improved versions of programs every day.

Apps are popular because they're convenient, and constantly innovating. They changed the world when they first started being developed, and now we can't break away from the model. Worldwide number of mobile app downloads in 2017 reached 178.1 billion, and so far of 2018, it already reached 205.4 billion. In Google play store and Apple store, there are thousands of Apps, what makes an App successful? What's the standards of a successful App? Should we set our target users as everyone, or set up specific target users like teenagers/students, children? What type of App do we want to develop: financial, education, entertaining? Which aspects will affect the popularity of Apps: size, cost or the name? Those are very interesting topics to discuss about.

In our project, we assume there's a company wants to develop a mobile App which will create great benefits. This company don't have any experience of App developing, but they want to find out what makes an App popular. They need us to do marketing research for them. We use information visualization techniques to analyze the data set for the current market. We propose the purpose of our research:

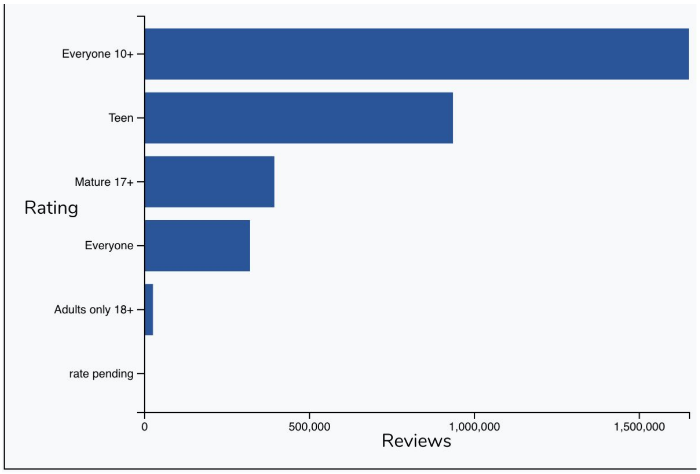
-- Decide which kind of App should be developed and target users.
-- Name choosing.
-- Effect of user rating.

We design 6 detailed questions to find answers for above questions:
1: Who are our target users?
2: What kind of APP is more attractive to our target users?
3: Can we have more detailed analysis of target users' preferred category?
4: Does name contribute to APP popularity?
5: Is free downloading always better?
6: Is high rating App always popular?
We created 6 graphs with D3.js to find answers to these questions.

We first aggregate records by Content Rating. The Entertainment Software Rating Board (ESRB) ratings provide concise and objective information about the content in video games and apps so consumers, especially parents, can make informed choices. We use the same rating system and divided users into 6 groups:

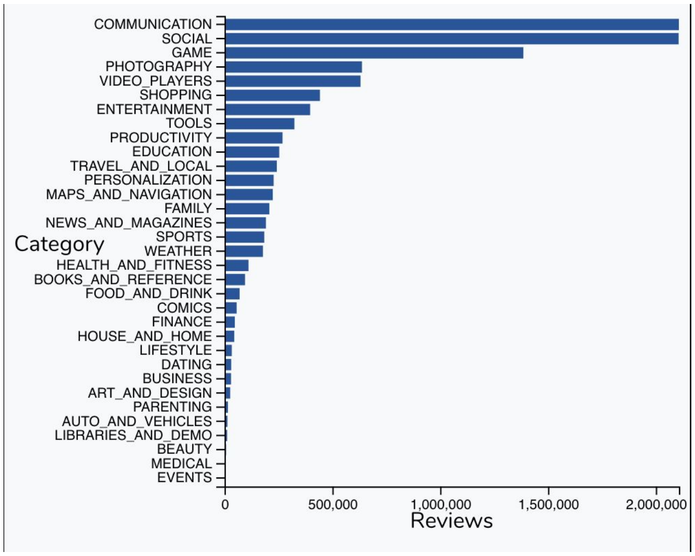
EVERYONE
Content is generally suitable for all ages. May contain minimal cartoon, fantasy or mild violence and/or infrequent use of mild language.
EVERYONE 10+
Content is generally suitable for ages 10 and up. May contain more cartoon, fantasy or mild violence, mild language and/or minimal suggestive themes.
TEEN
Content is generally suitable for ages 13 and up. May contain violence, suggestive themes, crude humor, minimal blood, simulated gambling and/or infrequent use of strong language.
MATURE
Content is generally suitable for ages 17 and up. May contain intense violence, blood and gore, sexual content and/or strong language.
ADULTS ONLY
Content suitable only for adults ages 18 and up. May include prolonged scenes of intense violence, graphic sexual content and/or gambling with real currency.
RATING PENDING
Not yet assigned a final ESRB rating.

For a company, it's important to set accurate target users, we design a graph with D3.js to represent the top 3 group of target users.
X axis represents different age range of target users. It includes six groups: Adult only 18+, Everyone, Everyone 10+, Mature 17+, Teen, and Rate Pending. Y axis represents the number of App installation of each age group.



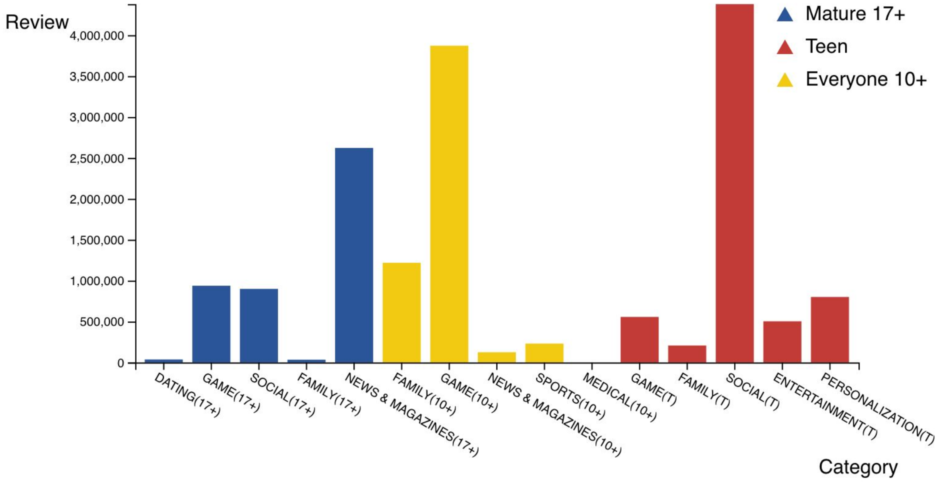
From the graph, we can easily conclude people of which age download the App mostly. Everyone 10+, Teen and Mature 17+ have highest number of reviews compared with other groups. We should consider them as our potential target users if we want to build a popular app.

After we set up our potential target users, the next step is setting the target genre: which type of app should we build? What kind of APP is more attractive to our target users? A video sharing App? A social communication App? Or Games? We provide following graph to find answer of our questions:



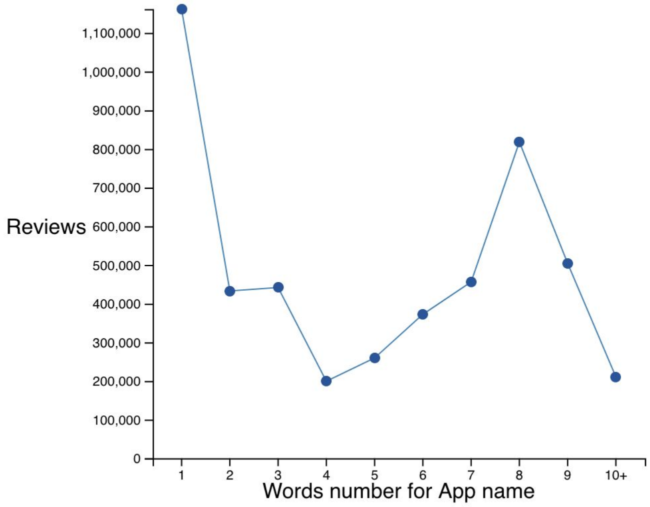
From this graph, we can find the top 5 popular genres are Communication, Social, Game, Photography, and Video Players. The top 5 least popular genres are Auto & Vehicles, Beauty, and Libraries & Demo, Medical, Parenting. Comics and Lifestyle have downloads, but not a lot. This provide different options for the company: do they want to create an App which has a lot of competitors or an App has potential market？Should we get into a market with a lot of competitors? Or develop a new market with less competitors, but also less users? These are questions we need to think about.

At this point, we need more detailed analysis of target users download behavior. We design a graph of further research from question 1 and 2: Identify top 5 downloading genres of our potential target users: mature 17+, teen, and everyone 10+.



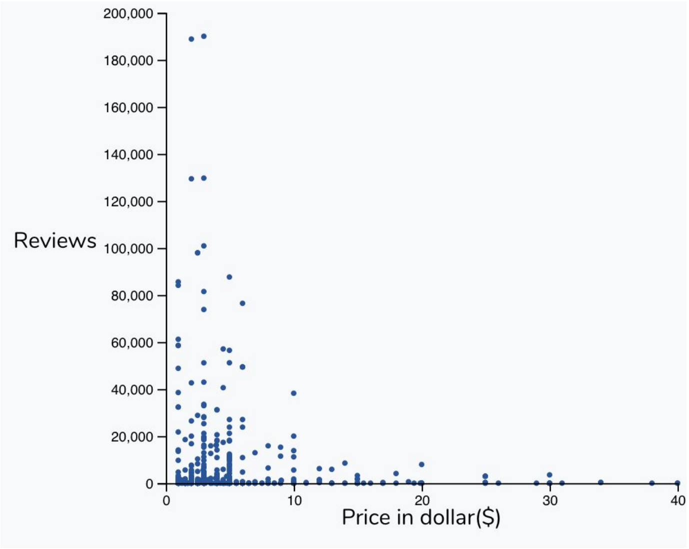
This is a surprise, the analysis can be very different when we see it from different perspectives. If the company want to develop an App which targets users are Mature 17+, they should consider News & Magazines; if they want their target users to be Everyone 10+, Game is the genre they want to consider; and if their target users are Teen, it's better to choose Social genre.

So far we are making good progress, now it's time to get into next part: does name contribute to APP popularity? Would we want to download an app just because it has an interesting name? Would "Candy Crush" be the same popular if we change the name into "Sugar Mash"? Would users still like "Instagram" if we change the name into "You Can Only Do Photo Post"? Is "facebook" gonna be more popular if it has a shorter name "FB"? We create the following graph to get answers of our question. X axis represents different groups of App of different length, and Y axis represents the number of installation and reviews.



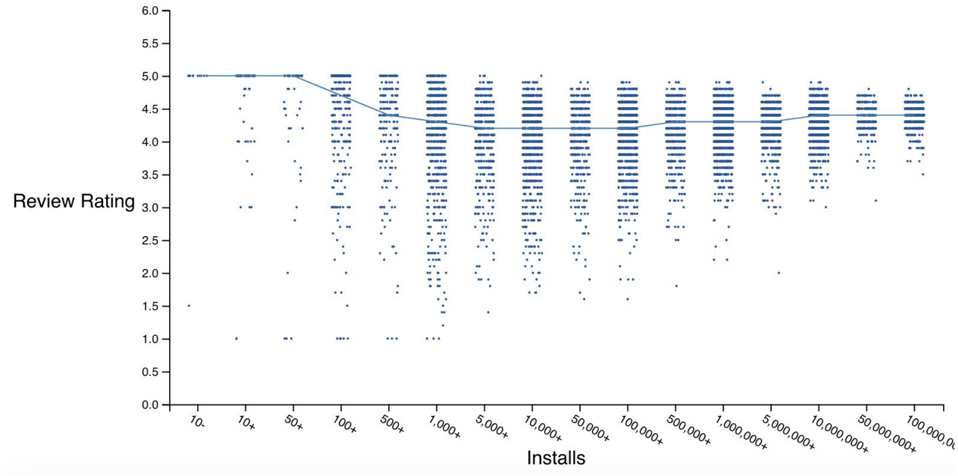
From this graph, we can see Apps with shorter names seems have more downloads, and Apps with about 8 characters are as popular as Apps with short names, and Apps with 4 letters have least downloads. So it’s better to give our App a name with short name with 1 or 2 letters, or 8 letters.

Free stuff is full of our life right now, sometimes we get free cookies from school; stores might offer free gifts to get more customers into their stores. Is free stuff always good? Can we apply this “free is good rule” on our APP development? We design this graph to solve this question. We use Y axis to represent the number of installation, and X axis to represent the price of App



Yes, we can! Free apps definitely have tremendously more downloads compare with charged Apps. And most users purchased 5$ or less cost Apps. Free App and low price App is more popular. Therefore, we can create an App with lower price, maybe a lite version with free downloading, and in-app purchase if uses like our App.

Finally we get to our last problem: is high rating App always popular? Do users check App rating when they try to download an app? If there are more than one App users can download of the same genre, do they prefer higher rating? Or more downloads? Let’s find out. Each dot is a representative as an App Review, Y-axis is a mark of App rating, and X-axis as App downloading times.



From the graph, we can see Apps with higher rating doesn't necessary mean they have more downloads, maybe because they are new released, maybe because the targets users are not big amount of people. But users tend to try an App with high rating even without a lot of downloads. Users prefer higher rating Apps when they have more than one choice. There's another interesting thing we can find from this graph：Apps with the most downloads doesn't have the highest rating. That's something we should pay attention on, and we can develop more interesting research for it: Do a lot of users uninstall the App？Or they keep the app? These are questions we want to do more effort if we are going further of research.

For conclusion, the company needs to be very clear about the target users: setting Everyone 10+, since they are the most active users. The company also need to set App genres: communication, social, and game are the top 3 genres, so choosing one of them is a good idea. App name should be either 1 or 2 letters, or 8 letters since they can make users interested in downloading. Most users prefer free or low price(5$ or less) Apps, therefore it's better to create a free or low cost App. Rating is very important for a beginning, it's better to create test app before officially release to App store.

Implementation

GitHub repository:
https://github.com/NYU-VIS-FALL2018/storytelling-group-6
Demo page:
https://nyu-vis-fall2018.github.io/storytelling-group-6/