Yi Tang

Dr. Fatemeh Hendijani Fard

13 December 2019

Report of Mobile Applications Project

Dataset Description

The dataset is collected for 87 top of unique mobile applications from eight categories, which are education, entertainment, family, finance, game_action, health_and_fitness, lifestyle, and music_and_audio, with a total of 2715303 reviews from March 3rd, 2019 to May 9th, 2019. After removing duplicate reviews, there are 1478938 reviews left.

Initial Statistics & Analysis

The dataset has eight categories: education, entertainment, family, finance, game_action, health_and_fitness, lifestyle, and music_and_audio. Categories entertainment has the highest number of applications, which is fourteen, along with ten applications for categories education, family, finance, health_and_fitness. A strange situation happened: the sum of all applications for eight categories is 88 instead of 87. For further examine, the application names Duolingo: Learn Languages Free appears in two categories: education and family.

category	
EDUCATION	10
ENTERTAINMENT	14
FAMILY	10
FINANCE	10
GAME_ACTION	11
HEALTH_AND_FITNESS	10
LIFESTYLE	12
MUSIC_AND_AUDIO	11

category	
EDUCATION	137227
ENTERTAINMENT	226723
FAMILY	167172
FINANCE	185512
GAME_ACTION	252986
HEALTH_AND_FITNESS	154330
LIFESTYLE	137751
MUSIC_AND_AUDIO	217237

We count reviews for each category for the next step. Category game_action has 252986 reviews to become the first place in all categories in this period, though the category education is the lowest with only around half of the number of reviews in game_action. The reasons for high amounts of texts in category game_action can be gamers are like to give expression to their thoughts, or they can receive some rewards in games if they write reviews. There exist a significant difference in review numbers for eight categories. Recreation applications always have more reviewers because of more users.

category	contentRating	
EDUCATION	Everyone	10
ENTERTAINMENT	Everyone	4
	Mature 17+	1
	Teen	9
FAMILY	Everyone	7
	Everyone 10+	3
FINANCE	Everyone	10
GAME_ACTION	Everyone	5
	Mature 17+	2
	Teen	4
HEALTH_AND_FITNESS	Everyone	10
LIFESTYLE	Everyone	11
	Mature 17+	1
	Teen	1
MUSIC_AND_AUDIO	Everyone	2
	Teen	9

For the content rating part, all applications in categories education, finance, and health_and_fitness can work for everyone. Nine apps in categories music_and_audio and entertainment separately, four applications in category game_action and one application in lifestyle are generally used by teens. Besides, there are exist some apps that only allow users with above ten years old or seventeen years old. The apps with ages constraint may include some contents that should keep children away from them.

Text Process and Analysis

After dropping non-English reviews, removing non-ASCII characters, punctuations, and adjacent duplicate characters, I decide to remove the reviews that contain two or less number of words. From my previous experience, it is difficult to write a constructive suggestion in two words. The possible two words reviews could be "Greatest app.", "Worst app." Thus, I even thought it is useless to keep one-word and two-word reviews in some specific score-sub-groups. Our aim is receiving some useful feedback from reviews, which can help developers to improve applications. Due to the large size dataset we have, remove the words equal, or less than two terms, can help us save time and human resources to get higher productivity.

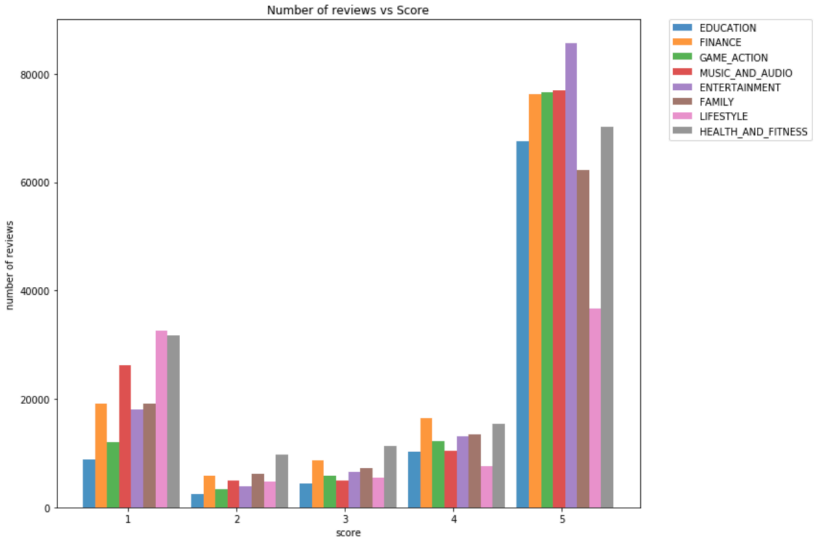
In this case, we can compare the number of reviews by each category before and after removing one-word and two-word reviews. It clearly shows that category music_and_audio takes game_action's place to become the category owned the highest number of reviews, and the number of reviews in category lifestyle falls to the bottom of the list. Category game_action only left among 60% of reviews after removing. This result may confirm what I wrote above: they chose to write one or two words as an application review to receive some rewards in games. Reviews in categories education, family, and health_and_fitness reviews still have 76% left, and the users may write reviews only when they have something want to express.

Category	before	after	percentage
EDUCATION	122543	93530	0.76
ENTERTAINMENT	198812	126159	0.63
FAMILY	145043	109821	0.76
FINANCE	169751	123519	0.73
GAME_ACTION	214488	127306	0.59
HEALTH_AND_FITNESS	142388	108325	0.76
LIFESTYLE	124402	87075	0.70
MUSIC_AND_AUDIO	193746	138319	0.71
Total	1311173	914054	0.70

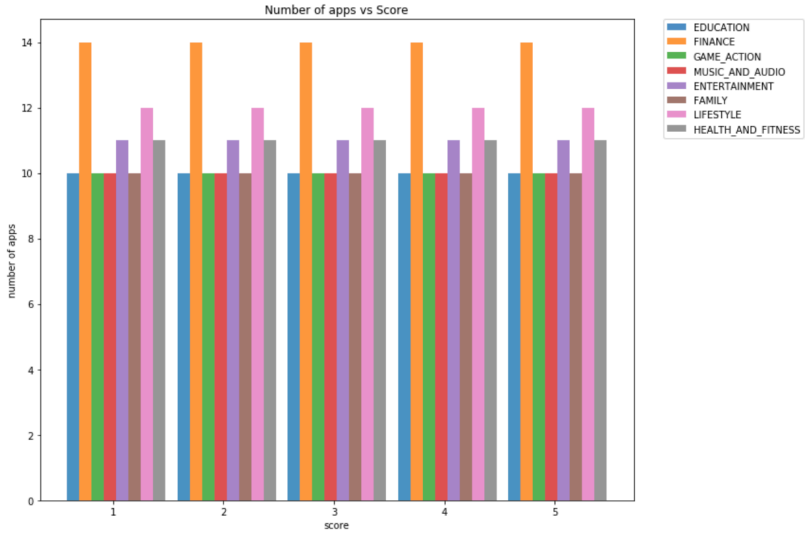
Categories Analysis

Introducing and exploring a new factor score here. First of all, focusing on the relationship between categories and scores given by reviewers. Above 60% of users of applications in

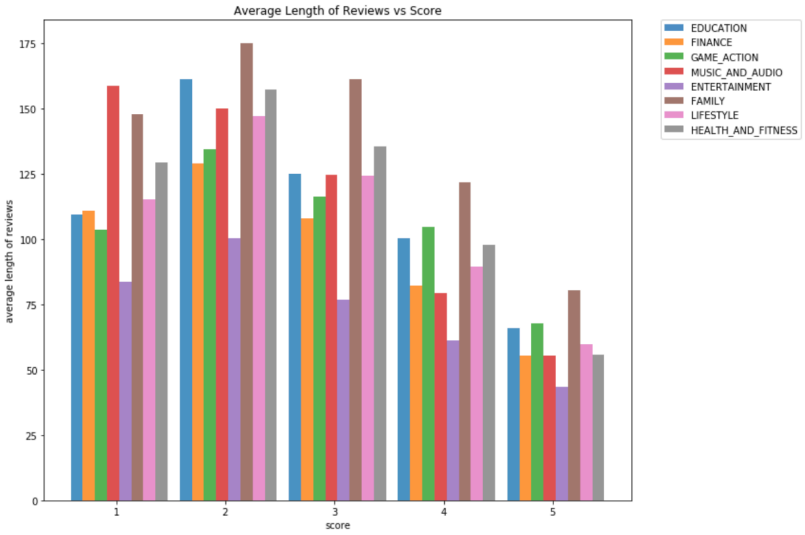
categories education, entertainment, family, finance, game_action, and music_and_audio gave rating 5 for good experiences, but only 40% of users in category lifestyle gave rating 5. There is a low percentage of users of applications that gave a rating 1 to 4 for all categories except lifestyle. There exists that 38% of users thought bad experience and gave rating 1 for lifestyle applications. The plot on the left shows that category entertainment has more than 80000 reviews for the score with five that gets the first prize, though category lifestyle has only 40000 reviews for the score with five with the almost same number in rating 1.



To count apps in each score-sub-group is useless for this dataset since all applications received score 1 to 5. Next, looking for the relationship between the average length of reviews in each category and the score-sub-groups. For all categories except music_and_audio, they follow the same tendency:



when the score goes up, the mean length of reviews reaches the peak in rating 2, followed by decreasing until hit bottom at score 5. However, the average length of reviews in category music_and_audio decreases along with the score increases. The category family has the highest average range of reviews for rating 2 to 5, with average equals to 175 in score 2, and category entertainment has the lowest average length of reviews all the time. It is an almost inverse relationship between the average length of reviews and the score-sub-groups in each category.

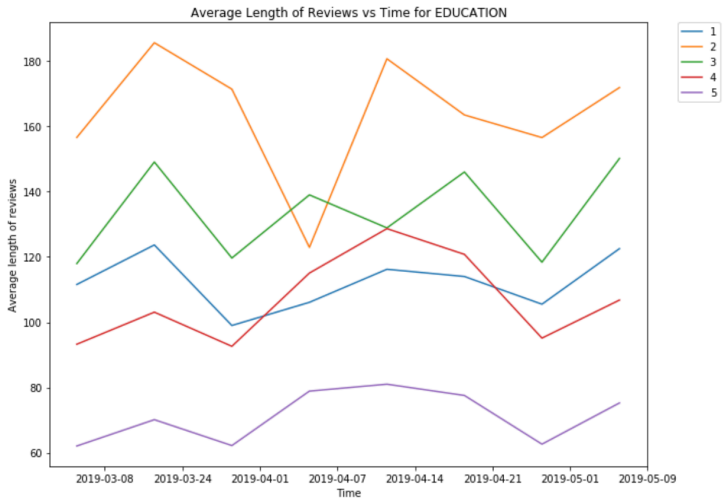


There exists some correlation between the length of the reviews and the score in each app-category. The highest one is -0.43 appeared in category finance, and others are around -0.3, which are all have low correlation coefficients. However, it gives the idea that the length of the reviews and the score are in a negative relationship in each app-category. The analysis of the connection between the average length of reviews and the score-sub-groups in each category above also gives a similar conclusion.

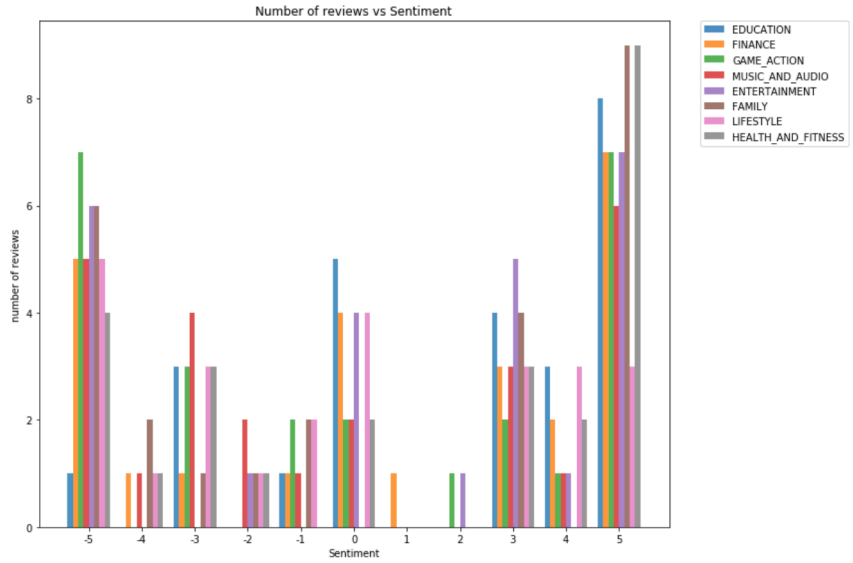
category	
EDUCATION	-0.216824
ENTERTAINMENT	-0.285170
FAMILY	-0.176173
FINANCE	-0.432277
GAME_ACTION	-0.237993
HEALTH_AND_FITNESS	-0.274127
LIFESTYLE	-0.262165
MUSIC_AND_AUDIO	-0.347001

The next step is comparing rating and the average length of reviews during the time for each app category along with date time. Since the date records in the dataset are chaotic, I even can find

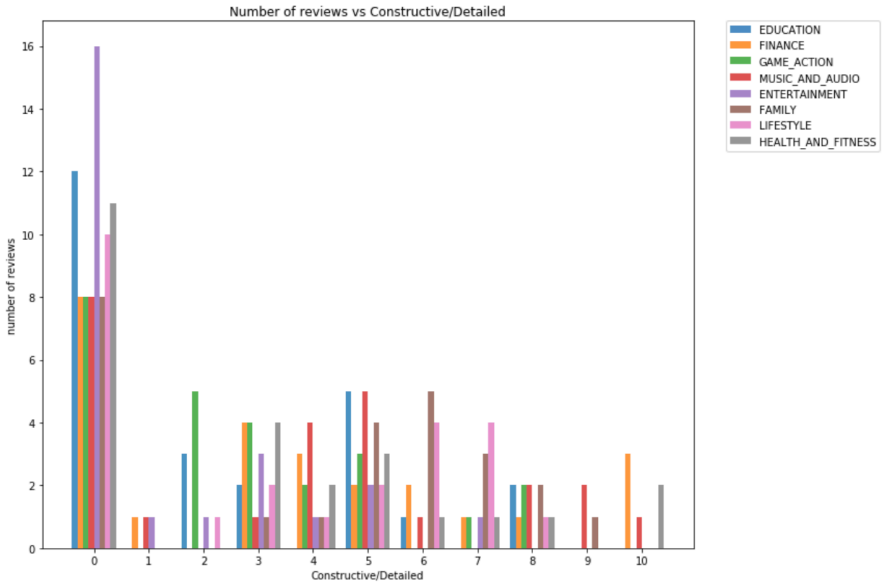
some 2014 in the date column. Therefore I decided to deal with them as mistakes made when putting reviews into csv file since if I choose to deal with them with mistakes made when grabbing from apps and remove all texts with the wrong dates. There will have some apps removed. Then I used the date when the data collected to analyze. I only put the line chart for education here since rating with 3 has the highest average length of reviews except on April 7th, 2019, a significant margin decline in rating 3 caused rating 2 reaches the first prize in that day. For further details, the average length of reviews is the smallest for all categories when the score is five, and the average length of reviews is largest for half of the categories when the score is 2.



Randomly selecting five reviews with each score-sub-group from each category, I used my thought to mark -5 to 5 in sentiment factor and 0 to 10 in the constructive factor. For the sentiment column, the result plots show on the left. Most of categories have the highest number of reviews in score five except lifestyle. The category lifestyle has the largest number of reviews in rating -5. Most of categories appear the



smallest number of reviews in score 1 or 2. For constructive/detailed column, all categories have the hugest number of reviews in rating 0 with the entertainment get the first place that means most of the feedbacks are useless. For non zero texts, most of them are moderately useful, which is in rate 5. Since these two columns were rated by myself, it has large randomness because of no criterion.



In conclusion, reviewers will write more if they think the apps need to be updated or fixed, for users who enjoy the apps will not write much feedbacks. Many reviews in category game_action are less than or equal to two words; the high score for apps may be cheating. Mobile applications in most categories have a high score, but apps in lifestyle have low scores with lots of complaints. The developers for lifestyle applications should improve their applications by listening to users' thoughts, fitting bugs, and adding more features.

Learnings & thoughts

This project is the first for me to clean texts and analyze such large dataset in python. Here are some libraries I used in this project, and I did not know previously: os, glob, langid, string, and

re. I used the os module to set the working directory that I can read data straightly. And then, I used the glob module to grab the csv file with parts of file names. Glob can catch the keywords both in folders and subfolders that did an excellent job for me since the data for March 24th, 2019, are put in subfolders. The package langid is surprising to me. Langid is a language identification tool that can distinguish sentences in different languages. I can simply enter a sentence, and it will show me what language the sentence used. The string is a package that I used to deal with punctuations, and re is a package that used to remove multiple characters is they occur more than twice continuously. I am still learning re, it is a functional package when I want to clean texts, but it is also difficult to understand.

Furthermore, I did not consider texts that need to be read and graded by analysts before. I think it has large randomness and bias because of no criterion. My advice is setting a standard for each score-sub-group and using it as a guideline to rate. Overall, I earned an achievement when I finished the project.