Characteristics Classification of Mobile Apps on Apple Store Using Clustering

Boxi Fu

University of California, Berkeley, USA
Email: 1907096379@qq.com

Abstract

This research is interested in the user ratings of Apps on Apple Stores. The purpose of this research is to have a better understanding of some characteristics of the good Apps on Apple Store so Apps makers can potentially focus on these traits to maximize their profit. The data for this research is collected from kaggle.com, and originally collected from iTunes Search API, according to the abstract of the data. Four different attributes contribute directly toward an App's user rating: rating_count_tot, rating_count_ver, user_rating and user_rating_ver. The relationship between Apps receiving higher ratings and Apps receiving lower ratings is analyzed using Exploratory Data Analysis and Data Science technique "clustering" on their numerical attributes. Apps, which are represented as a data point, with similar characteristics in rating are classified as belonging to the same cluster, while common characteristics of all Apps in the same clusters are the determining traits of Apps for that cluster. Both techniques are achieved using Google Colab and libraries including pandas, numpy, seaborn, and matplotlib. The data reveals direct correlation from number of devices supported and languages supported to user rating and inverse correlation from size and price of the App to user rating. In conclusion, free small Apps that many different types of users are able to use are generally well rated by most users, according to the data.

Keywords

Mobile Apps, Clustering, User Rating, Pairplot, Scatter Plot, Functionality

1. Introduction

As smart phones entering people's life, Apps for different operating systems for smart phones create brand new markets for Apps developers. Gradually, mobile Apps become profitable and grow faster than ever as new technologies and fea-

tures are added to mobile devices. Mobile Apps Stores like Apple Store allow users to rate their experience with Apps, and users usually use ratings from other users to determine whether to download an App or not. To maximize a mobile App company's profit, it is important to understand what the users think a good App is like. This research aims to have a better understanding on what traits are users for mobile Apps are looking for when using them.

Previously, the same data set has been used with a focus on popularity of different Genres, and the relationship between how willing users are to pay for Apps within different Genre and the Genres themselves [1]. For the purpose of this research, clustering is used to deal with numerical attributes of this data, thus Genres of Apps are not the focus of this research specifically.

For this research, the importance of Genre is reduced, while other numerical attributes are emphasized in attempting to understand what constitute a good App in general. To understand multiple numerical attributes and provide an easy visual representation of the result, clustering is chosen. This research hopes to find some meaningful correlations between one or some of the attributes and user rating.

2. Materials and Methods

2.1. Data

Mobile App Store (7200 apps).

https://www.kaggle.com/ramamet4/app-store-apple-data-set-10k-apps.

https://colab.research.google.com/drive/1T4r-xIgi7-3Svw5NqOWmHnErPiJArjpj.

The ever-changing mobile landscape is a challenging space to navigate. The percentage of mobile over desktop is only increasing. Android holds about 53.2% of the smartphone market, while iOS is 43%. To get more people to download your app, you need to make sure they can easily find your app. Mobile app analytics is a great way to understand the existing strategy to drive growth and retention of future users.

With millions of apps around nowadays, the following data set has become very key to getting top trending apps in iOS app store. This data set contains more than 7000 Apple iOS mobile application details. The data was extracted from the iTunes Search API at the Apple Inc website. R and linux web scraping tools were used for this study [2]. Given this rich dataset, the reasons for the study and the interests in data analysis are outlined in the next section.

2.2. Reason and Interest

Smartphones are one of the most commonly used technologies today, and Apple is among the best in its field. What is different from Apple to other smart device companies like Samsung and Google is that Apple iOS has built an insanely well-rounded app market. Many users have been driven to purchase Apple devices for the apps they support.

This dataset has 7195 unique apps, providing a large enough set for us to analyze and find a pattern while not too large to contain many outliers and null entries. In fact, the 16 columns it contains cover pretty much all information about an app. Data collected in July 2017 is not too out of date yet, ensuring that analyzing this dataset would yield useful results. Multiple hypotheses can be formulated based on the dataset (see from Table 1).

"id," "currency," and "vpp_lic" are less interesting for data analyzing purpose: "id" are assigned independently; "currency" are all the same: "US Dollar;" and "vpp_lic" has no direct influence on ratings. Some potential correlations are: what aspects determine an app's price on the AppStore? (size_byte, number of version, prime_genre, sup_devices, or lang). Does price have an effect on user's rating of the app? (rating_count_tot, rating_count_ver, user_rating, user_rating_ver, ver, cont_rating). Do positive ratings help the developer to carry out more versions? Does higher price help the developer to carry out more versions? (ver, price, rating). What are some characteristics of a good App based on rating, on profits made, or on long lasting effects (ver)? (Do users generally like games better than music apps? Do they criticize Finance apps harsher than social media apps?) [1]. However, despite how many questions this dataset can potentially answer, it has some defects that have to be acknowledged detailed in the next section.

Table 1. Attributes of the DataSet.

"id": App ID
"track_name": App Name
"size_bytes": Size (in Bytes)
"currency": Currency Type
"price": Price amount
"rating_count_tot": User Rating counts (for all version)
"rating_count_ver": User Rating counts (for current version)
"user_rating": Average User Rating value (for all version)
"user_rating_ver": Average User Rating value (for current version)
"ver": Latest version code
"cont_rating": Content Rating
"prime_genre": Primary Genre
"sup_devices.num": Number of supporting devices
"ipadSc_urls.num": Number of screenshots shown for display
"lang.num": Number of supported languages
"vpp_lic": Vpp Device Based Licensing Enabled

DataSet Shape: This dataset contains information about 7195 Apps from the Apple Store. It provides 14 attributes for each App, as listed above. There are no missing values in any column or row. Given this information mentioned above, the dataset is only cleaned slightly as described below.

2.3. Drawbacks

This is one of the highly reviewed datasets on kaggle, with a high score of 491 votes. There must have been similar analysis of these datasets previously. Also, being collected two years ago makes this dataset relatively early. It is not too early to produce any value, but later dataset would be preferred. Finally, product review spam cannot be addressed by the dataset alone. To improve the quality of this study, spam recognition and further cleaning of the dataset are required [3].

2.4. Clean Up

The dataset is loaded into Google Colab without any encoding due to the nature of this dataset containing characters of different languages.

This dataset contains no missing values, thus there is no need to fill in any dummy value. "Unnamed: 0," "currency," and "vpp_lic" columns are dropped since: "Unnamed: 0" is just an index number for all the Apps, "currency" column for this dataset contains only USD, and "vpp_lic" is a license number unrelated to the focus of this research [4].
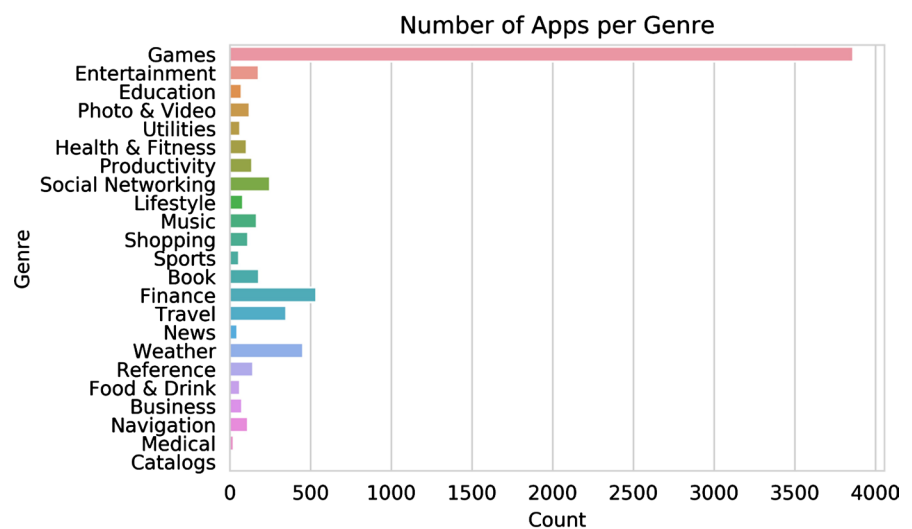
3. Results and Discussion
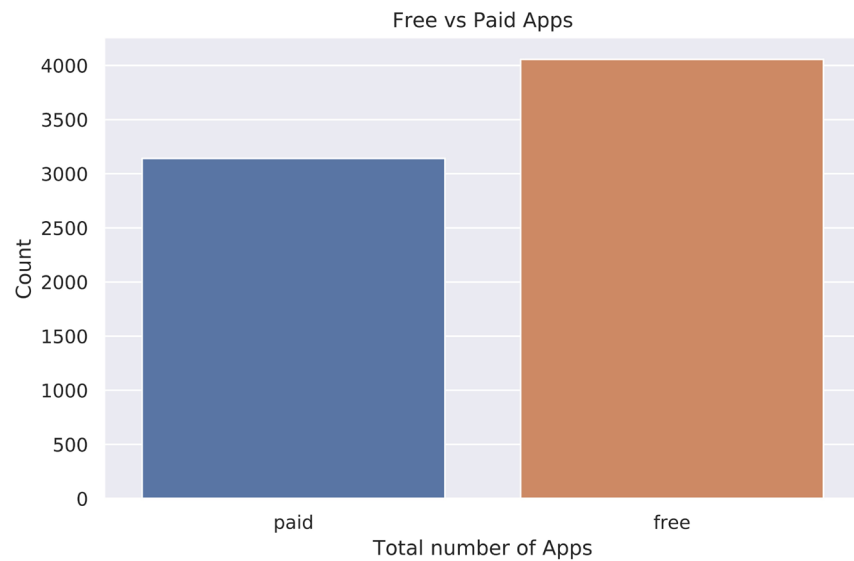
3.1. Exploratory Data Analysis

Large number of Apps having Games as their prime genre may lead to bias and error; however, this research did not end up discussing the relationship between prime genre and rating (see from **Figure 1**).

A number of Apps for Free and for a price have an almost even distribution. Outlier and bias effect may be less significant for this dataset (see from **Figure 2**) [1].

Most Apps with the highest rating_count_tot are free, which may be a trend for Apps receiving high ratings. There are Apps that charge users for downloading that receive high ratings as well, for example, Baby Connect (Activity Log) (see from **Table 2**).

Figure 1. Number of apps per genre.

Figure 2. Apps as free or paid.

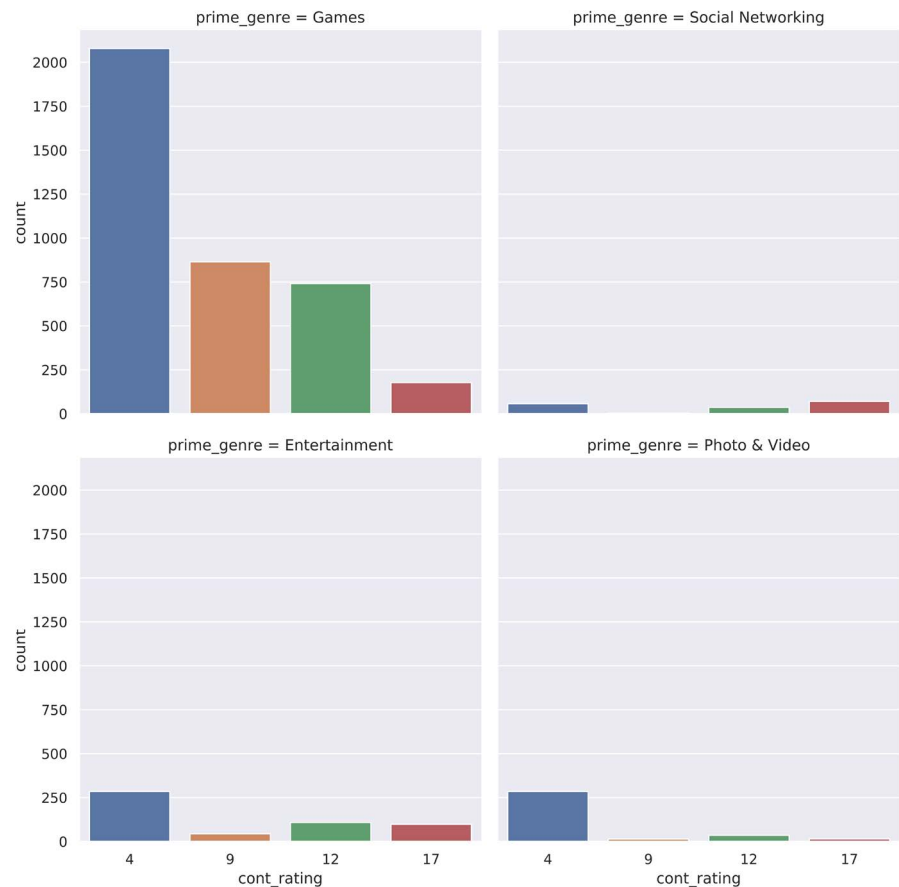**Table 2.** Apps with the most rating_count_tot within each genre.

prime_genre	track_name	price	rating_count_tot
Social Networking	Facebook	0.00	2,974,676
Photo & Video	Instagram	0.00	2,161,558
Games	Clash of Clans	0.00	2,130,805
Music	Pandora—Music & Radio	0.00	1,126,879
Reference	Bible	0.00	985,920
Health & Fitness	Calorie Counter & Diet Tracker by MyFitnessPal	0.00	507,706
Weather	The Weather Channel: Forecast, Radar & Alerts	0.00	495,626
Utilities	Google—Search made just for mobile	0.00	479,440
Travel	Google Earth	0.00	446,185
Shopping	Groupon—Deals, Coupons & Discount Shopping App	0.00	417,779
News	Twitter	0.00	354,058
Navigation	Waze—GPS Navigation, Maps & Real-time Traffic	0.00	345,046
Lifestyle	Zillow Real Estate—Homes for Sale & for Rent	0.00	342,969
Entertainment	Netflix	0.00	308,844
Food & Drink	Starbucks	0.00	303,856
Sports	ESPN: Get scores, news, alerts & watch live sports	0.00	290,996
Book	Kindle—Read eBooks, Magazines & Textbooks	0.00	252,076
Finance	Chase Mobile℠	0.00	233,270
Education	Duolingo—Learn Spanish, French and more	0.00	162,701
Productivity	Evernote—stay organized	0.00	161,065
Business	Indeed Job Search	0.00	38,681
Catalogs	CPlus for Craigslist app—mobile classifieds	0.00	13,345
Medical	Baby Connect (Activity Log)	4.99	3483

The graph above contains some of the attributes from the original cleaned dataset without "id," "ver," and "track_name" attributes given the fact that plotting the trend of these three attributes provides no useful information and/or the attribute contains non-numeric values. The above pairplot reveals some characteristics of Apps according to different genres. Large number of reddish pink dots match the previously observed concerns of overly large number of Apps with Games as their prime genre. Eventually, this concern is dodged since genre is not focused on by this research. The general trends of all Apps can also be observed by comparing two attributes at a time (see from Figure 3).
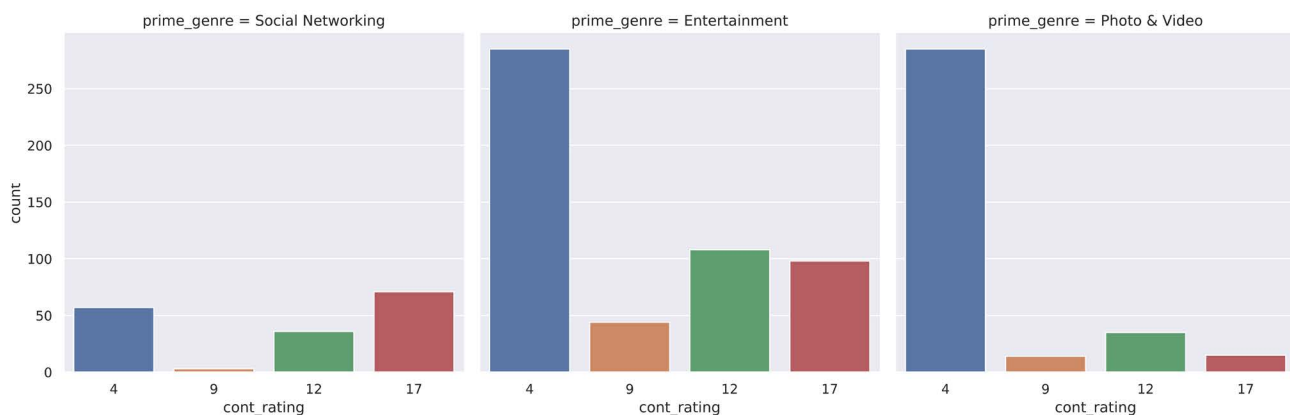
The content rating of Apps with different prime genre is further explored. In general, most Apps have a 4 + content rating, with only a few other levels of content rating. The concerns of overly large number of Apps with Games as prime genre is resolved by plotting a separate graph with Apps with Games as prime genre removed. Similar trend is observed (see from Figure 4 & Figure 5).

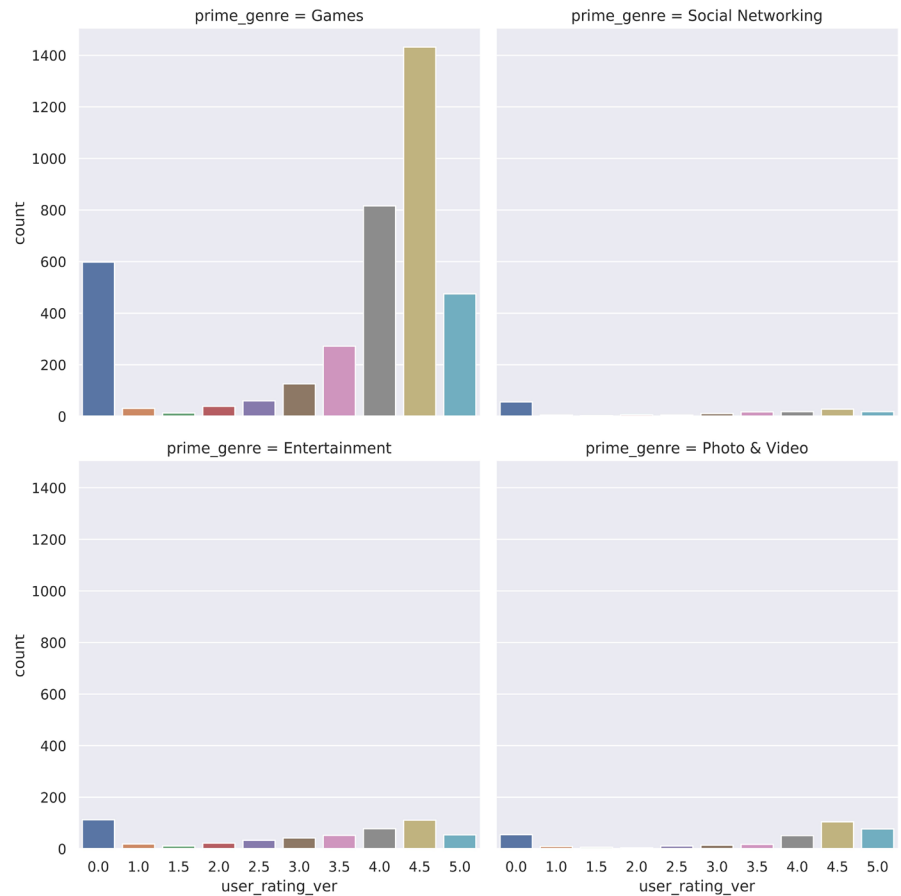**Figure 3.** Pairplot of apps colored by genre.

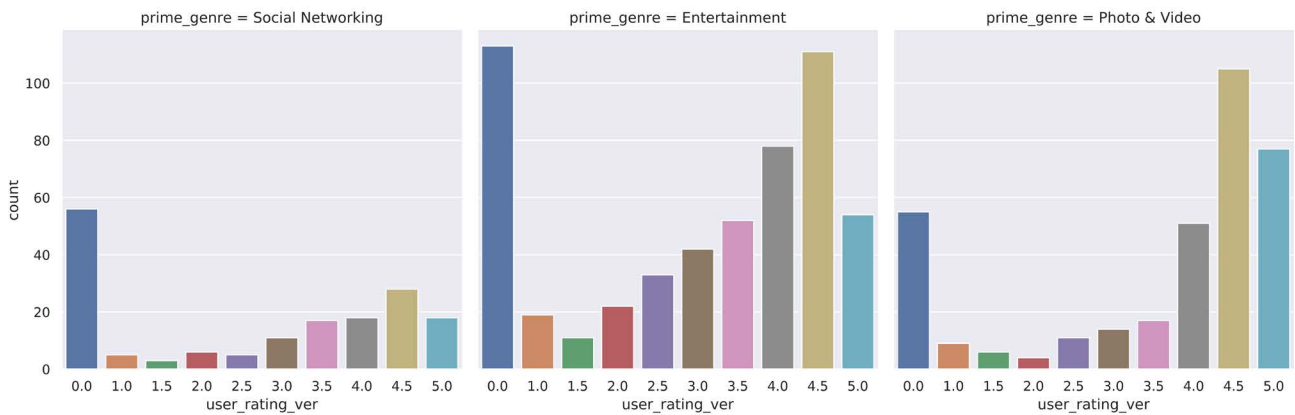**Figure 4.** Content rating of apps with different prime genre.

Figure 5. Content rating of apps with different primve genre with games removed.

Similarly, the user rating for current version of Apps with different prime genre is explored by two graphs above. In general, a polarized distribution can be observed: a large number of Apps with 0.0 user rating for their current version and a relatively larger number of Apps with rating within the range of rating 4.0 to 4.5 can be seen across all genres (see from **Figure 6** & **Figure 7**).

Most Apps have either no screenshots or 5 screenshots, according to the above picture. The previously observed spike in 0.0 rating and a general large amount

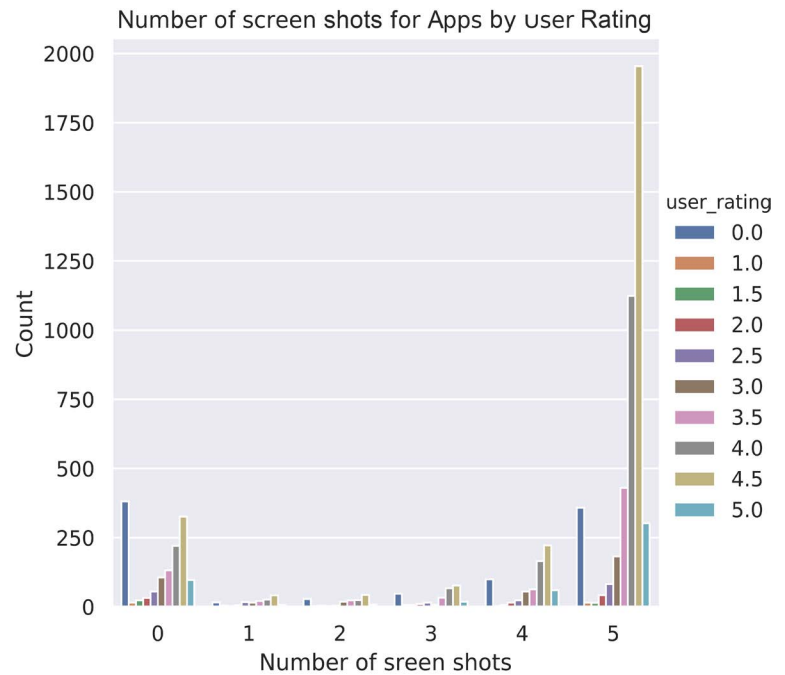**Figure 6.** User rating for current version with different prime genre.

Figure 7. User rating for current version with different prime genre with games removed.

of Apps with rating around 4.0 to 4.5 remains within each column with respect to the number of screenshots. Generally, it is more likely to have a higher rating with either 0 or 5 screenshots, according to this dataset (see from **Figure 8**).
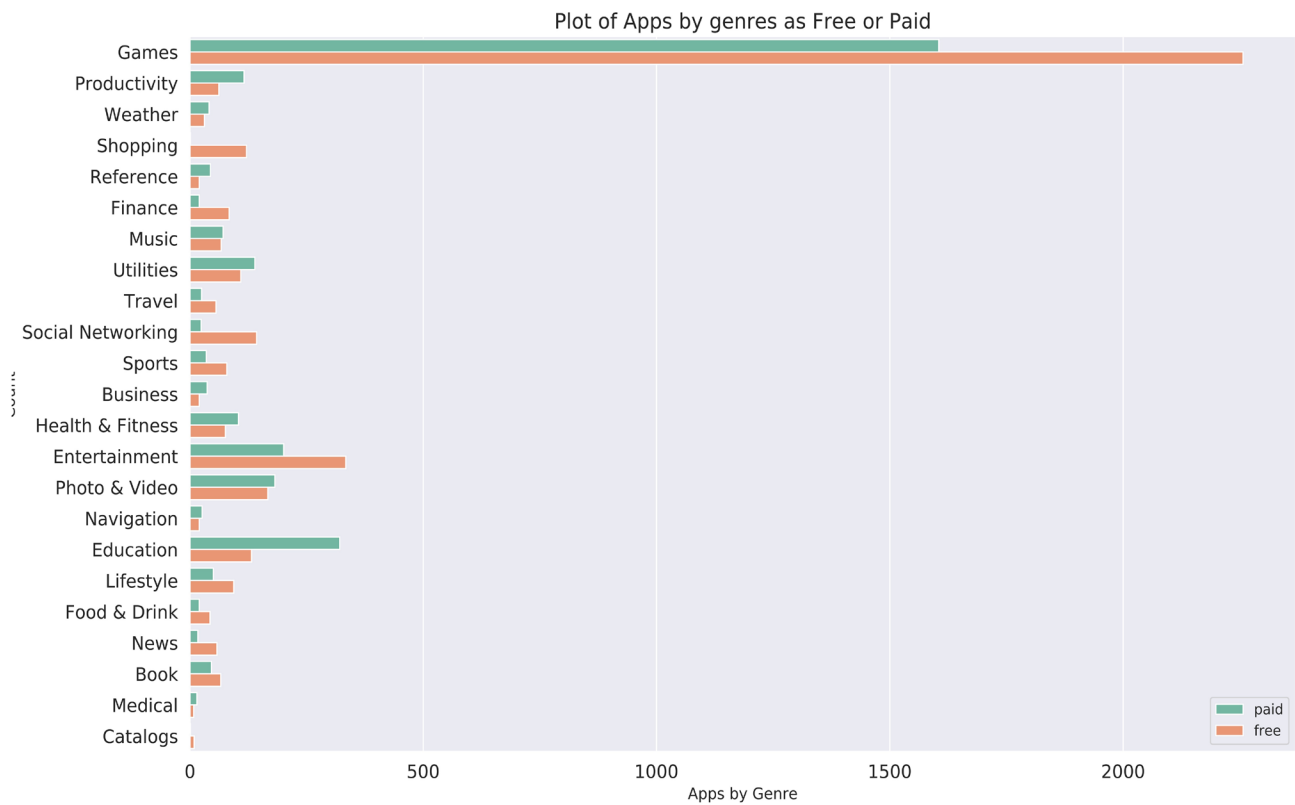
In general, the ratio of number of Free Apps and number of Paid Apps is about 3 to 2 or 4 to 3. Exceptions include Apps with prime genre Education, Productivity, Medical and Health & Fitness. This may be caused by a higher expectation from these Apps under these genres in creating more reward afterward

in the long run; users see these Apps as worthy investments, thus willing to pay for them. This is something Games, Social Networking, and Entertainment Apps do not provide for the general group of users (see from **Figure 9**).

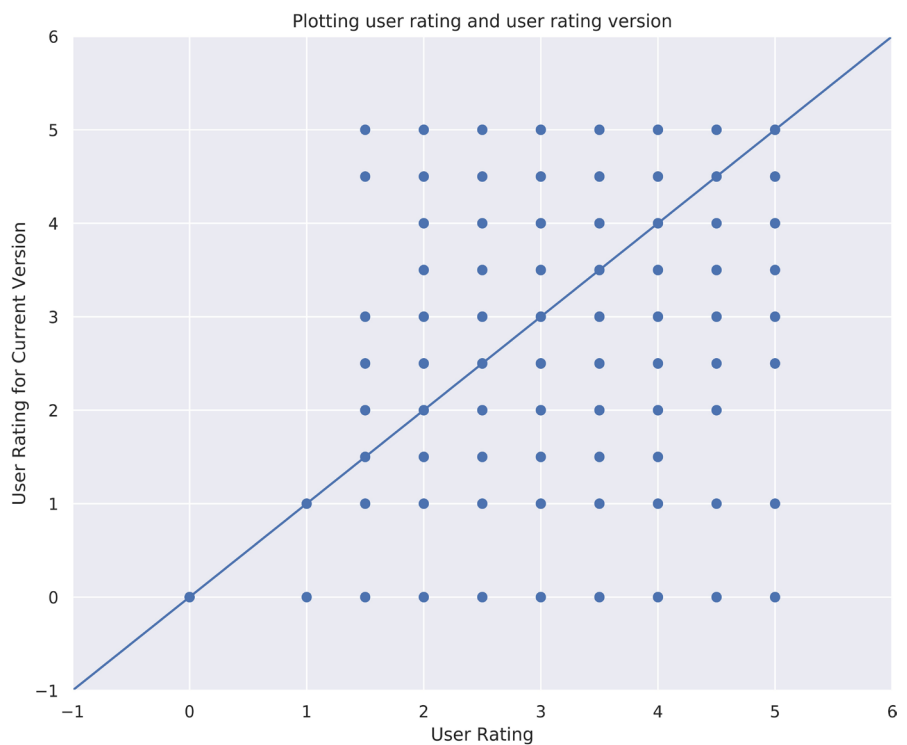**Figure 8.** Number of screenshots for apps by user rating.

Figure 9. Plot of apps by genres as free or paid.

3.2. Clustering Analysis

"Prime_genre," "ver," "track_name," "id," and "cont_rating" attributes are removed from the previously read dataset. These attributes either contain non-numeric values or does not provide any additional information through clustering. The Standard Scaler method from public library panda is used to standardize the dataset. This method divides the difference between each numeric value and the average value of the column by the standard deviation of the column. The standardized dataset is stored as df, for data frame.

User Rating for all versions and User Rating for Current Version have a relatively even correlation in which almost every possible combination of User Rating and User Rating for Current Version exists. This means that sometimes both attributes agree while other times they contradict one another significantly. The Apps with dot representation close to the line y = x on the above graph have a good representation of the two rating, *i.e.* User Rating for Current Version reflects the average User Rating for all versions; however, Apps with dot representation further away from the line y = x either have a significantly bad or a significantly good user rating for the current version compared to all its past versions. This shows that using either only user_rating or only user_rating_ver to define an App's success can lead to error (see from **Figure 10**).

A large distribution of price for user rating 4.0 is noticeable. This shows that most Apps fall in the rating between 3.5 and 4.5, which agrees with the previous observation. Notice that all of the Apps with user rating 5.0 have price under 50 dollars; in fact, most Apps have price under 50 dollars. Apps with price over 50
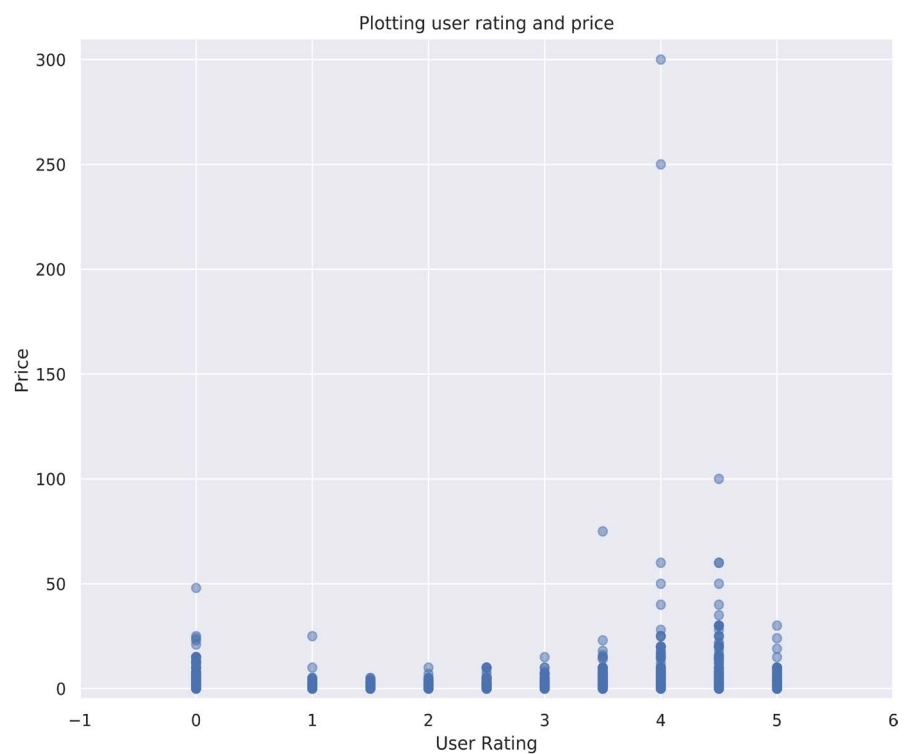

Figure 10. Scatter plotting of user rating of all version and user rating of current version.

dollars all end up having user ratings above 3.0. Thus giving an App a high price (above 50 dollars) may improve its rating to some degree, according to this dataset (see from Figure 11) [1].
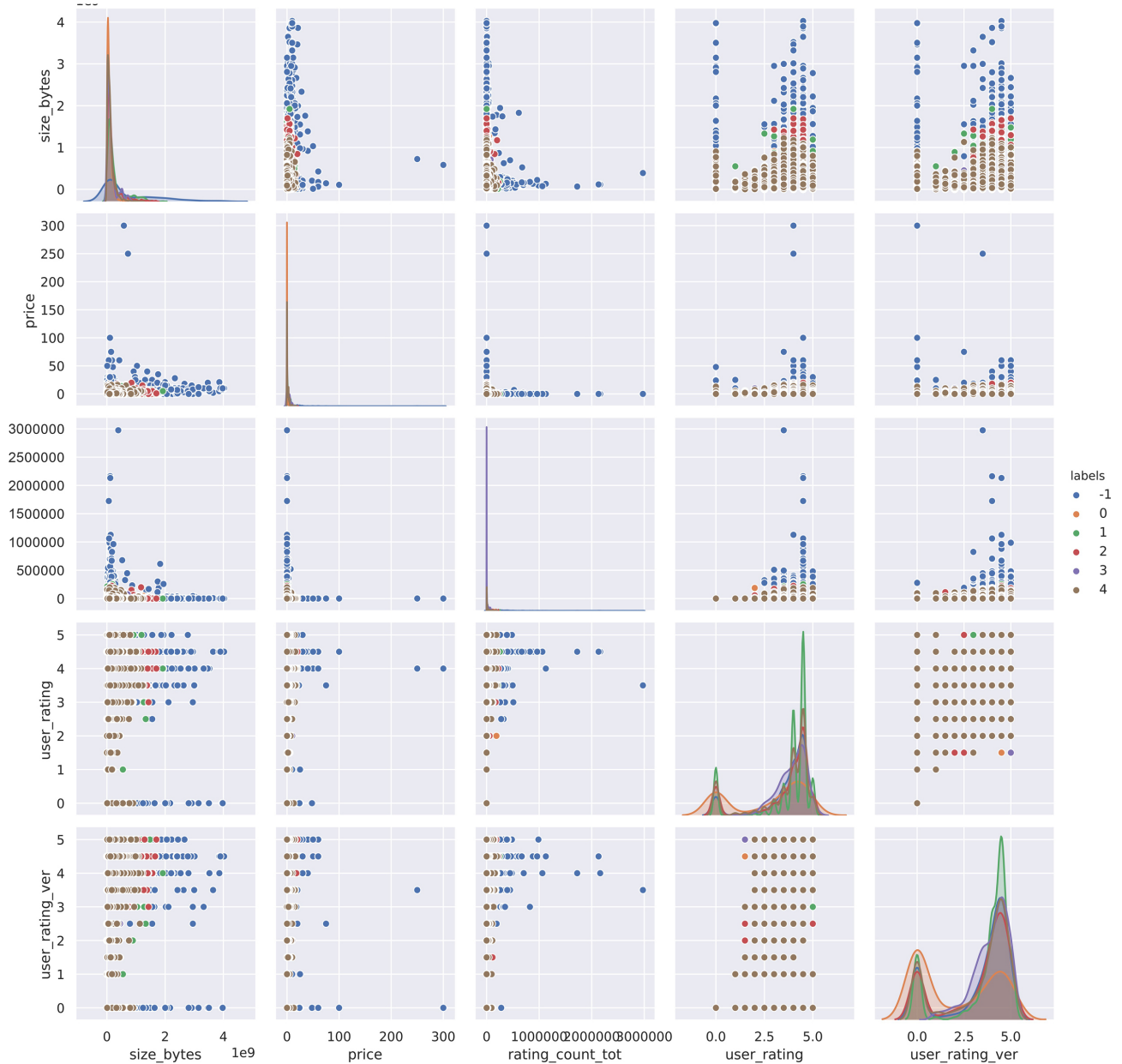
From the above pairplot of scatterplots, the behaviors of Apps in this dataset with respect to attributes "size_bytes," "price," "rating_count_tot," "user_rating," "user_rating_ver" can be observed. The clusters are assigned by the method fit provided by the public library hdbscan. A total number of 6 clusters are generated. From the graph, dark blue cluster represented by number −1 represents the relatively successful Apps within this dataset: they have a relatively large total rating and high user rating on average and for current version. By considering all three of the above ratings the success of an App can be determined more precisely and accurately. By tracking those characteristics of this cluster we can see what makes them successful. In contrast, the brown cluster represented by number 4 marks the relatively bad Apps within this dataset. Characteristics of Apps within this cluster 4 can also be found and used to identify the reason of a less successful App within this dataset (see from Figure 12).

From this pairplot, the characteristics of the −1 cluster can be observed (see from Figure 13).

Size in bytes generally has no correlation with other attributes for this cluster. Size distributes almost evenly across the full range of most other attributes except for number of language supported; in this case, a polarized distribution is observed: most Apps that support many languages are relatively small, while Apps that support only a few languages can either be large or small.
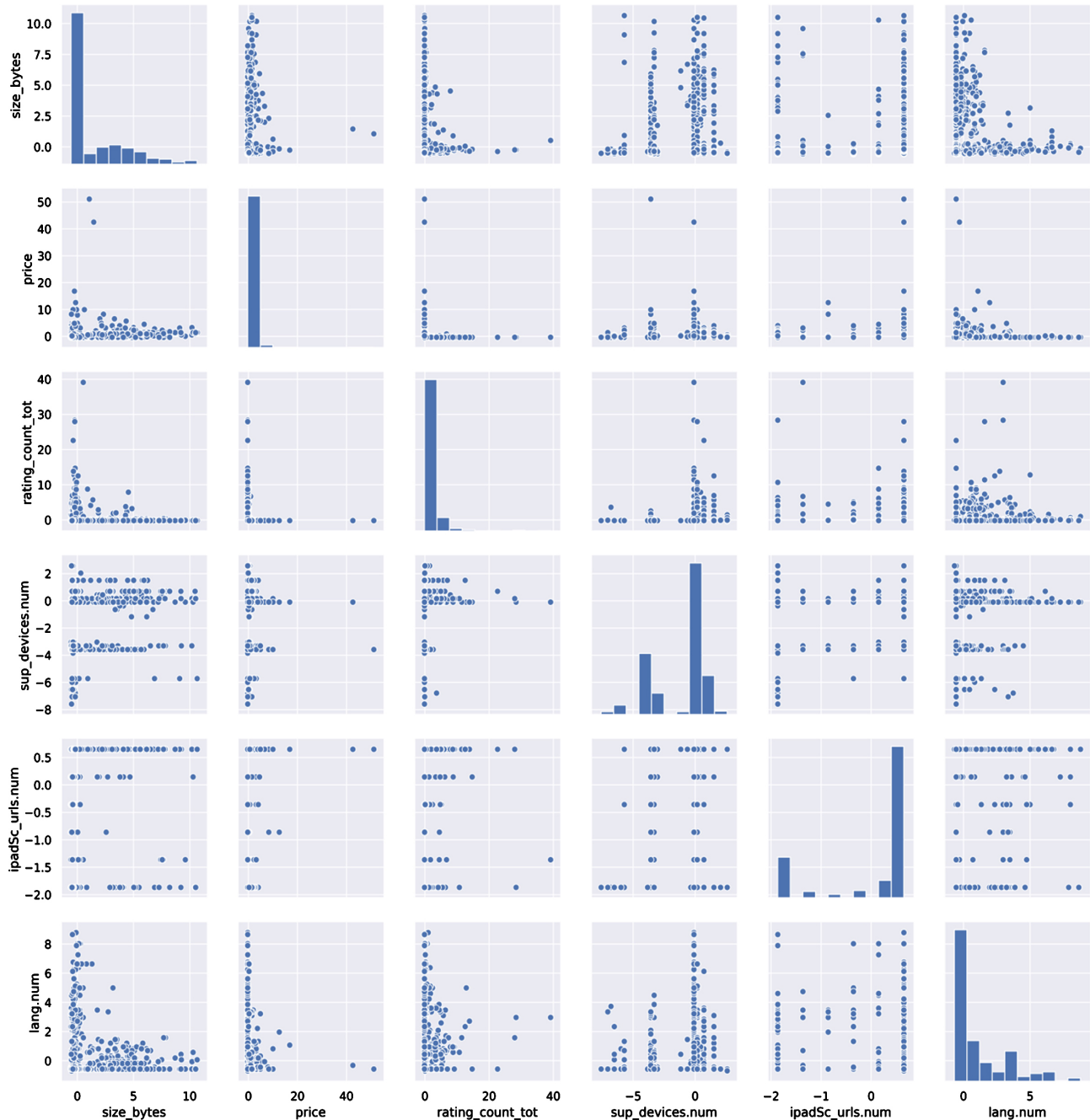

Figure 11. Scatter plotting of user rating and price.

Figure 12. Clustering of all apps with respect to size, price, and rating.

Price is an important attribute for this cluster. Price values are relatively low for Apps across all size in bytes. Similarly, price values are low for most rating count except for 0 rating count: this can be caused by the fact that users who do not want to spend money on these Apps do not rate them. The Apps with the highest few prices receive relatively high user rating, despite the fact that most of the free Apps also receive high user rating. In general, the more device an App supports, the higher the price, which makes sense. Similar correlation can be observed for number of screenshots, but weaker. This trend, however, is broken for a number of languages supported by the Apps; the more language is supported, the lower the price. In fact, most Apps that support many languages are free.

Figure 13. Behaviors of apps within cluster −1.

For the number of devices supported by an App, Apps that only support a very small number of devices are generally small in bytes, while those that support many devices can either be small or large. The number of devices supported generally has no correlation with respect to price, given the almost even distribution of price for Apps that support different number of devices. Apps that support more devices receive a higher total rating count since more users are able to download, use, and rate these Apps. No clear correlation can be observed between number of devices supported and user rating. No clear correlation can

be observed between number of devices supported and the number of screenshots either. Generally, Apps that only support a few devices support only a few languages, while Apps that support more devices have a larger range of number of languages supported.

Generally, the more languages supported, the higher an App can get in user rating. There are still a significant amount of Apps that support only a few amount of language receiving high user rating. No clear correlation can be made between language supported and other attributes yet discussed.

From this pairplot, the characteristics of the 4 cluster can be observed (see from Figure 14).

Size in bytes does have an inversely related relationship with price, rating_count_tot, and rating_count_ver: *i.e.* the larger the App, the cheaper and less total rating it receives. The Apps that cost a lot but are small in size, which may lead to limitation in its functionality, are subject to bad rating since users may feel like the money is not a worthy investment. Apps in this cluster have a similar size in bytes and user rating relationship compared to Apps in cluster −1. Compared to Apps in cluster −1, some Apps in this cluster are large in size but support way fewer devices, which limit their potential scope of users causing a lower total rating. Similarly, some Apps in this cluster that support the same number of language as Apps in cluster −1 are larger in size, which may cause users to complain.

Across the whole row (column) of price, one can notice that price is generally higher for Apps in this cluster than in cluster −1. Apps with similar or even worse functionality or performance that costs more would certainly get users disappointment.

A similar trend between the number of devices supported by an App and other attributes is observed, but with a relatively larger number of Apps supporting way fewer devices than those Apps in cluster −1 that have similar characteristics in other attributes. The most noticeable difference is probably the number of languages supported by an App and the number of devices it supports. Majority of Apps in cluster 4 have an inversely related relationship for these two attributes: the more it supports in one, the less it supports in the other. This, as described previously, would certainly cause a limitation in potential user population as well as dissatisfaction of users who will have trouble using these Apps.

The number of screenshots has little relationship with other attributes besides having a larger range than Apps in cluster −1: *i.e.* there are more Apps with drastically different characteristics in other attributes that share the same number of screenshots. Screenshot should not be a major factor in affecting an App being bad/unsuccessful.

The number of languages supported also behaves similarly to Apps in cluster −1, but with some slight differences. More Apps that support a larger number of language becomes larger, more costly, and support less number of devices than Apps in cluster −1. This would cause the same issues described previously, leading to a lower rating.
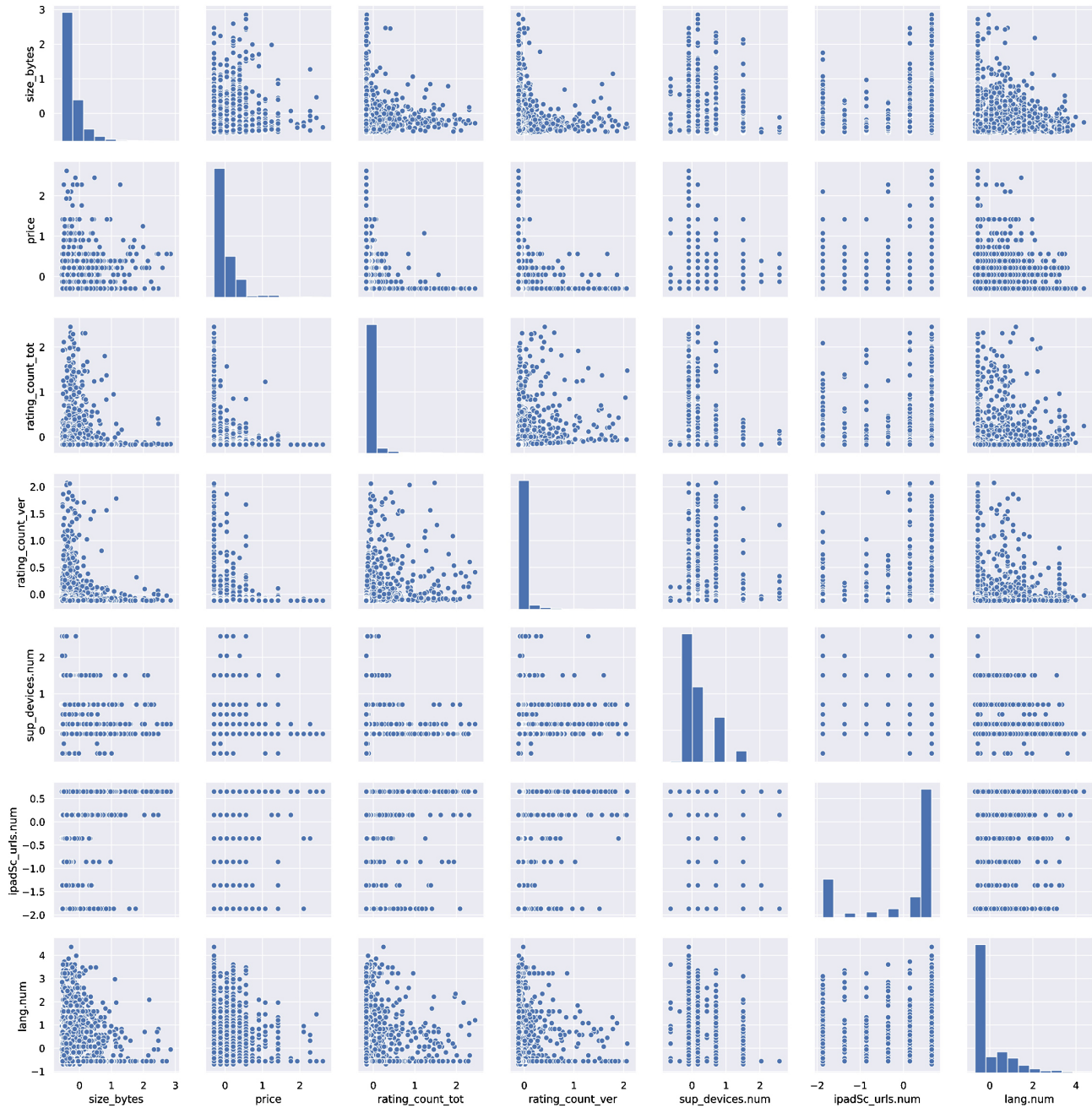
Figure 14. Behaviors of Apps within Cluster 4.

4. Conclusions

Smaller Apps that have many functionalities and support many devices received high user ratings. Apps that support many different languages, and larger Apps that do not cost as much while having at least as much functionality as what the users expect from its size typically are also successful.

Many free Apps receive high rating, so if cost is not a problem, making one's App free can be really helpful. On the other hand, Apps that provide exceptionally unique/useful/entertaining functionality with higher than average price can

also receive a good rating; except the definition of this functionality is not defined in this analysis. Thus in general, higher-priced apps received lower rating by their users, agreeing to other studies [5].

The numbers of devices and languages supported are loosely related to an App's rating. Generally, the more devices and languages supported, the larger the user population and the more likely the App would receive more rating and hopefully higher user rating. However, there are Apps that only support a few number of devices and languages that receive high user ranking. In this case, they match the characteristics of a successful App in other attributes.

The number of screenshots is even more loosely related to an App's rating. In general, a large number of screenshots is preferred, since many Apps that received a large number of total rating have at least 4 screenshots. There is a significantly large number of Apps with only a few screenshots receiving high user rating, so number of screenshots is really a minor factor.

5. Future Works

The relationship between prime genre and user rating is really interesting, as shown from the Exploratory Data Analysis part of this paper. This paper did not focus on genre mainly because the practice of clustering requires numeric value and the definition of distance. This relationship can be analyzed by splitting all Apps by genre and redo the clustering process above to find trending characteristics.

The practice of hierarchy clustering is considered, but eventually not used in this paper. The question of what information this hierarchical structure of Apps will reveal as well as how detailed this hierarchy should be are two major issues. A hierarchy clustering visualization was attempted and graphed with public library scipy.cluster.hierarchy, but, with over seven thousand leaves, such a clustering tree simply is illegible and impossible to be interpreted. Potentially, by sampling and providing a bound for how small a cluster can get, hierarchy clustering may produce some useful information. Finally, by splitting the whole dataset into smaller subset may yield even more surprising correlations not captured by the dataset, for example, the completeness of reviews other than the numerical score an App receives [6].

Acknowledgements

report. Professor Pradeep also instructed most data science knowledge to me with the help from CIS's slides.

Finally, thank you to my parents who provided the fund for this research project with CIS and all the supports I received along the way. I will not be able to finish this research without your supports.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

[1] Ruchigarg, D.S. (2018) Ruchigarg-DS/Mobile-Analytics. GitHub, Github.com/Ruchigarg-DS/Mobile-Analytics.

[2] Perumal, R. (2017) Mobile App Store (7200 Apps). Kaggle.
https://www.kaggle.com/ramamet4/app-store-apple-data-set-10k-apps

[3] Sheibani, A. (2012) Opinion Mining and Opinion Spam: A Literature Review Focusing on Product Reviews. 6th *International Symposium on Telecommunications* (*IST*), Tehran, 6-8 November 2012, 1109-1113.
https://doi.org/10.1109/ISTEL.2012.6483152

[4] Cathaypath Institute of Science (2018) Application of Data Analysis and Machine Learning Research Group.

[5] Finkelstein, A., Harman, M., Jia, Y., Martin, W., Sarro, F. and Zhang, Y. (2014) App Store Analysis: Mining App Stores for Relationships between Customer, Business and Technical Characteristics. UCL Department of Computer Science, RN/14/10, 1-12.

[6] Martin, W., Harman, M., Jia, Y., Sarro, F. and Zhang, Y. (2015) The App Sampling Problem for App Store Mining. *Proceedings of the* 12th *Working Conference on Mining Software Repositories*, Florence, 16-17 May 2015, 123-133.
https://doi.org/10.1109/MSR.2015.19