Google Play Store App Analysis

San Jose State University
Project Partners: Tien Nguyen and Shayna Gaulden

Report prepared for MATH 250
Instructor Dr. Guangliang Chen

TABLE OF CONTENTS

I. INTRODUCTION

The Google Play Store is one of the most popular Android app stores. The data set used in this report consists of some of the apps that are found in this store. The data set is officially called Google Play Store Apps and can be found on Kaggle [3]. It was originally webscraped by a user Lavanya Gupta; a Computer Science graduate at Carnegie Mellon University. It was posted for the use of exploratory data analysis or any other desired tasks. The file which was used in this report is a csv file where each row is an app in the app store and each column is an attribute of the application.

Researchers in the past have used Google Play Store Apps dataset to do exploratory data analysis, machine learning tasks and more. For example, Lengkong and Maringka did a rating classification task of the apps [1]. Businge et used the Google Play Store Apps dataset to study Android App popularity [2]. However, before building any models, there are three crucial initial steps that should be performed. Those are cleaning data, visualizing data and doing dimension reduction on the data. In this project, we complete the three steps mentioned in order to have a better understanding of the data. We use Matlab with the statistics and machine learning package.

II. SUMMARY INFORMATION

Variables (total 13)

Numerical (total 5)

- Rating: The average user rating, (scale from 0-5).
- Reviews: The number of user reviews for the app.
- Size: The size of the app.
- Installs: The number of user downloads/installations for the app.
- Price: The price of the application.

Categorical (total 5)

- App: The application name.
- Category: Category that the app belongs to.
- Type: Specifies whether the app is a free or a paid application.
- Content Rating: Age group the app is targeted at.
- Genre: The genre for an app can belong to multiple genres apart from its category.

Unused (total 3)

- Last Updated: The data the application last received an update.
- Current Version: The current version of the app.
- Android Version: The android version the app is compatible with.

The data set has 10841 instances and 13 variables, but only 10 of the 13 are used in this report. First we needed to clean and preprocess the dataset. There were a total of 1476 missing variables, 1474 from the rating variable, presumably apps that had never received a rating, 1 from reviews variable, and 1 from the content rating variable. After removing all the rows with missing values there were still 9366 entries. In the Size column we discovered most sizes were numbers, as expected, but some sizes were entered as “Varies with device” so all of the rows with this invalid entry were also removed leaving 7729 entries still to work with.

The next step was to convert variables into a usable data type. The size and installs variable both had special characters that needed to be removed before they could be converted into the numeric data type double. Size variables had “$” characters in front of each entry. Some installs entries had a “+” character because apps with a very large number of installations were binned together and rounded down and denoted with “+”. The size variable had some sizes listed in KB and some sizes in MB. Each size was labeled with the measurement that was used so we removed the special characters and converted all the size data into MB. For the genre variable, we discovered some genres had subgenres which were separated by “;”. To account for this we removed all the sub genres only keeping the main genre type. Originally, there were over 120 unique genres in the data set but after removing sub-genres that number was reduced to 48.

III. VISUALIZATION OF 1 VARIABLE

A. Categorical variables

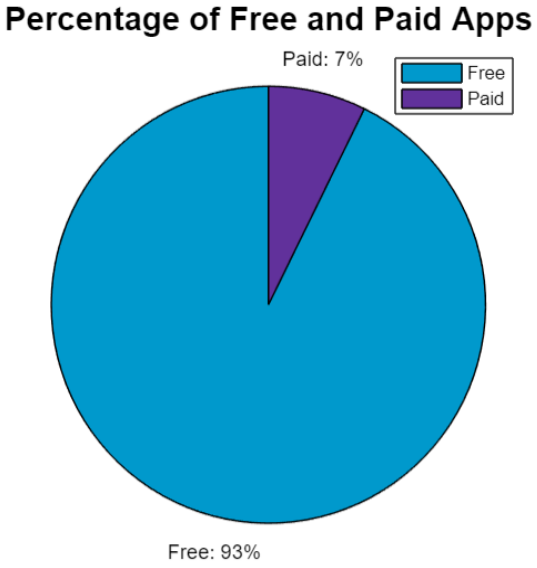

Fig. 1. Pie chart displaying the percentage of apps that are free versus those that cost money.

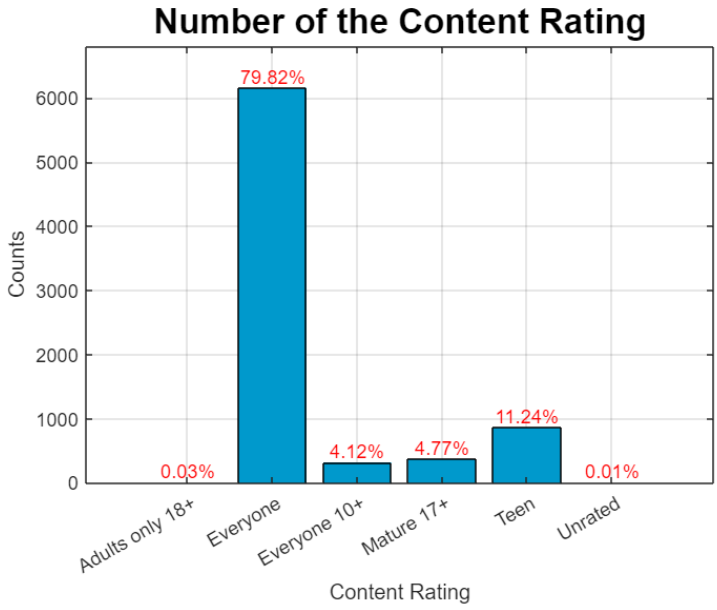

Fig. 2. Distribution of app content rating.

Most of the apps in the Google Play Store are free. Fig. 1 shows that around 93% of all Android applications are free while only 7% must be purchased. Fig. 2 displays the 6 unique content ratings given to apps in the Google App Store and number of apps in each content

rating class. The content rating that was most common was “Everyone”, which accounted for approximately 80% of all apps**.** The percentage of apps that are “Unrated” or rated for “Adults only 18+” is extremely small compared to the other content ratings. Respectively, there are only 1 and 2 apps in those content rating categories. This seemed too small a sample size to be considered in our data analysis so those 3 apps were removed effectively removing the two content rating groups with them.

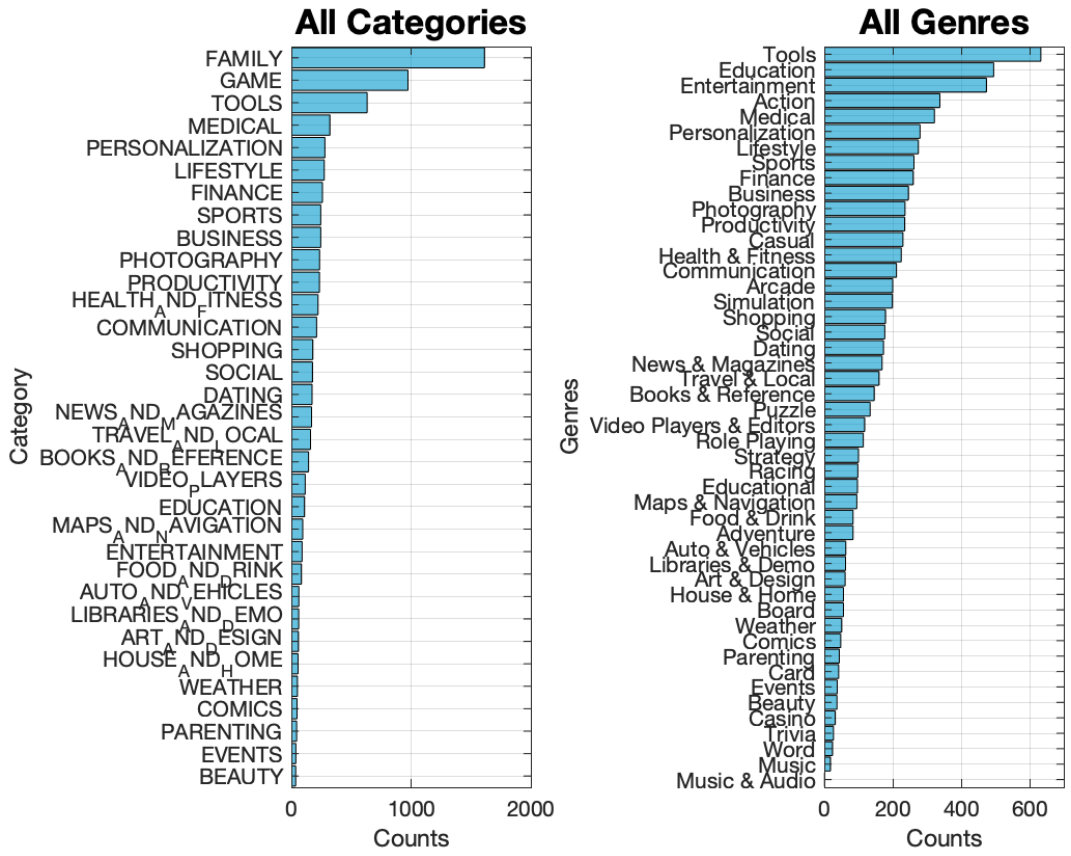

Fig. 3. Frequencies of all 33 unique app categories (left), and 48 genres (right).

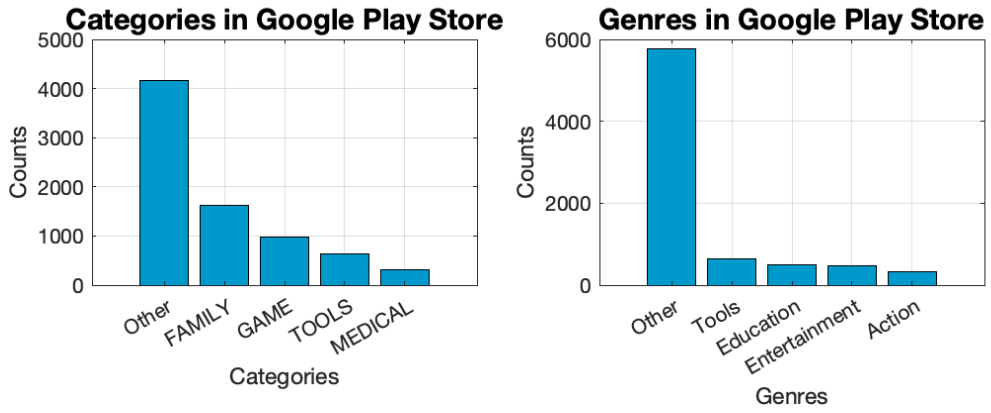

Fig 4. Histograms of the top 4 variables from app categories (left) and genres (left) with all other classes binned into "Other".

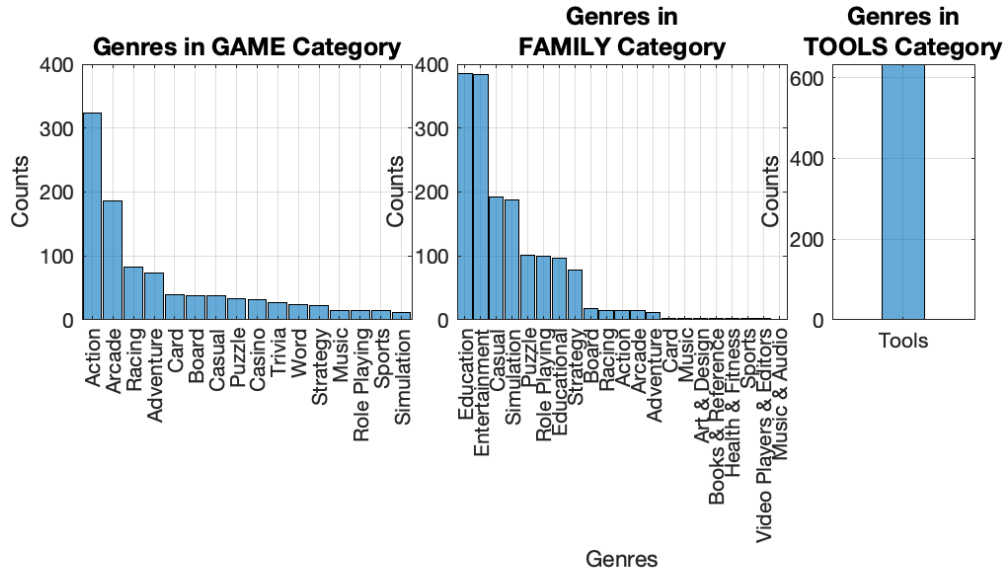

Fig. 5. Top 3 app categories and their unique genres.

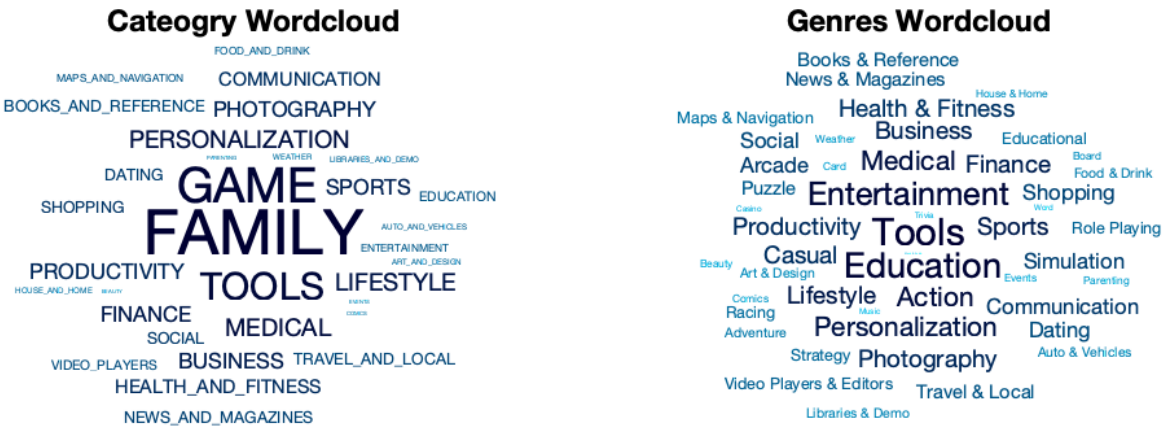

Fig. 6. Word clouds of category (left) and genre (right) variables.

The top two categories of apps fall under “Family” and “Game” while the “Beauty” and “Events” categories have the least number of apps; this can be seen in the graph on the left in Fig. 3 and the word cloud in on the left in Fig. 6. In Fig. 3 it appears that the number of apps in the “Family” category is almost two times higher than the second highest category, “Game”.

Originally we had interpreted the app genre to be just different types of categories, but this was not the case. In Fig. 3 we noticed that the names of the different genres seem to be a subclass of the app categories. To confirm our suspicion, we created histograms of genres for all apps in one category at a time. Three of these plots can be seen in Fig. 5. It turns out our observation was correct, the genre variable is unique to the applications category. For example the “Family” category of app has unique genres that are different and do not overlap with the “Game” category of app. Not every category has more than one genre, in fact only the “Game” and “Family” categories have multiple genres. Refer to Fig. 5 again to see the graph on the far right, the “Tools” category only has one genre also called “Tools”.

With all the genres together in Fig. 3 and 6, we can see that “Tools”, ”Education”, “Entertainment”, and “Action” are genres that have the highest number of apps. Notice that the genre with the highest frequency “Tools”, is also the name of the third highest category and is the only genre among the “Tools” category which is why it is the most frequent genre. The next most frequent genres are “Education” and “Entertainment” which are the top two genres in the largest category, “Family”. Finally, the fourth largest genre “Action” is the largest genre of the “Gaming” category. Understanding this breakdown and the relationship between category and genre explains the frequency distribution seen among the different genres.

B. Numeric Variables

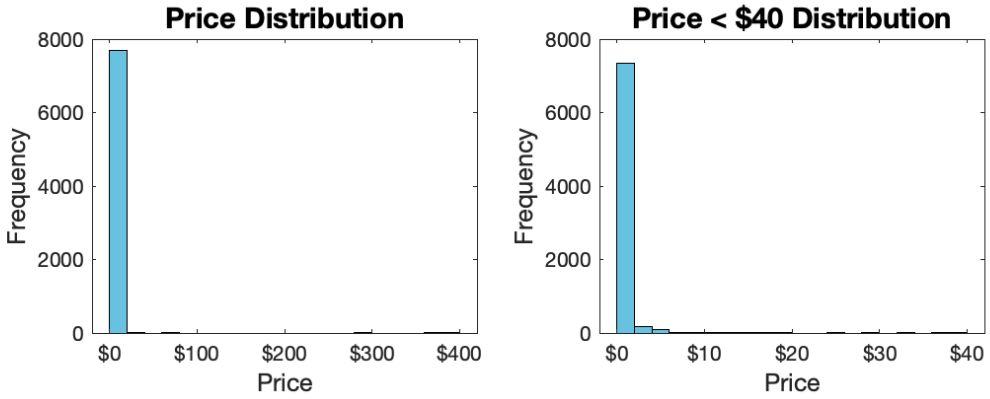

Fig. 7. Histogram of all prices compared to prices under $40.

A group of outliers was discovered by looking at the price variable and noticing how high the costs of some of the apps were, see Fig. 7. Most of the apps that are not free have a very low cost so it was surprising to see the maximum app price at $400. Upon further investigation, there were 15 apps between $379.99 and $400 all of these apps were named some variation of "I Am Rich". It appears that the original app "I Am Rich" was released on the IOS app store at $999.99 as an art installation, but was removed the day after it was released. We believe these apps to be knock off versions of that app. There were another 2 apps priced at $79.99 that were categorized as medical apps. These 17 high priced apps were removed from the dataset because their price was so much higher than every other app, refer to Appendix A. Fig. 45 for details on the removed outliers. In Fig. 7 we show the histogram of price after removing the outliers, notice the remaining apps are all under $40.

Most variables had a very large range of data with most of the apps falling within a small interval. This means a lot of our data was extremely skewed. The following plots compare the original variables to their log transformations. We believed a log transformation would help to reduce the range of our data and could potentially fix the skew shape.

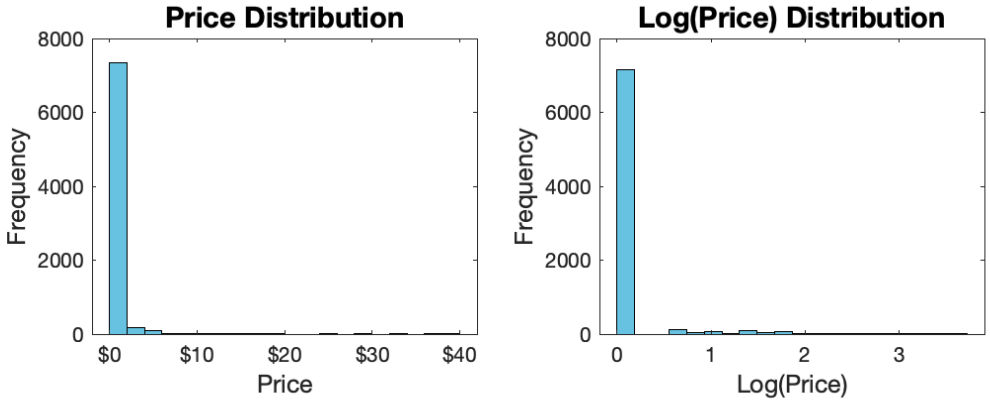

Fig. 8. Price distribution (left) compared to a log transformation of price distribution + $1 (right).

Price				
Mean	Median	Mode	Min	Max
0.3482	0	0	0	400

Fig. 9. Price summary statistics table.

Referring to Fig. 9 the median and mode are both 0, this is inline with the fact that free apps account for 93% of all apps on the app store. This caused a very skewed histogram for the price variable in Fig. 8. In order to try a log transformation on the price variable, we first needed to shift the data since the minimum price is 0 which is outside of the domain of a logarithmic function. We added $1 to every price before taking the log. In Fig. 8 The graph on the right shows the log transformation of price + $1 and while the transformation brought the range of numbers closer together, it did not help with the skewness of the data set so we decided at this point that price would not be considered for a log transformation. Once we get to the dimension reduction section of this report we discover the log transformation of price is indeed useful.

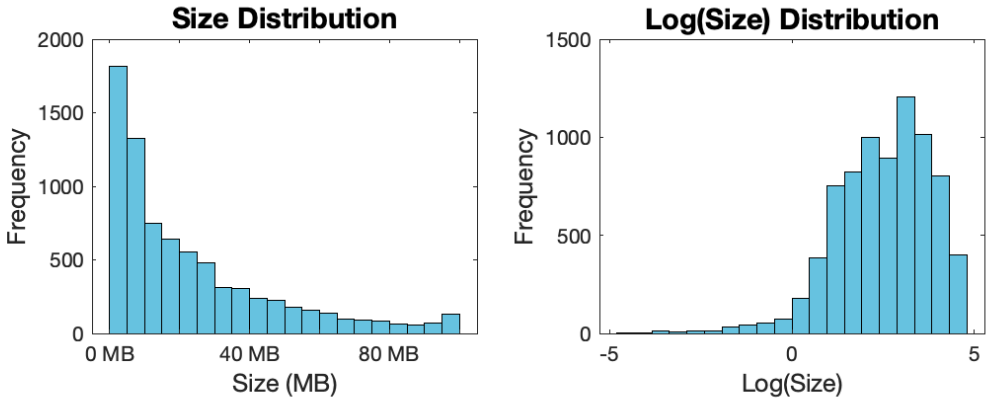

Fig. 10. Size distribution (left) compared to a log transformation of size distribution (right).

Size (MB)				
Mean	Median	Mode	Min	Max
22.9826	14	14	0.0085	100

Fig. 11. Size summary statistics table.

Referring to Fig. 10, it is clear that although the size variable was not as skewed as some of the other variables and had a less extreme range of data, the log transformation improves the histogram. It brings the values closer together and makes the shape of the data look better. Most apps have a smaller size, the median and mode being 14 MB which is the same as the mode, the mean is not much higher at approximately 23 MB see Fig. 11. Size will be considered for a log transformation going forward.

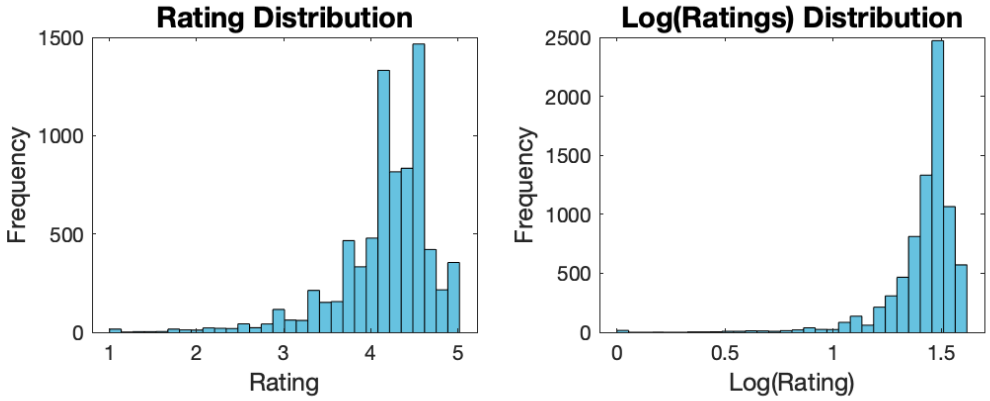

Fig. 12 Rating Distribution (left) compared to the log transformation of rating distribution (right).

Rating				
Mean	Median	Mode	Min	Max
4.1743	4.3	4.4	1	5

Fig. 13. Rating summary statistics table.

The rating distribution by itself did not appear to need a log transformation but one was attempted anyway for the sake of comparison, see Fig 12. Rating will not be considered for a log transformation. Most applications have a high rating hovering near 4.3 which we can see from Fig. 13.

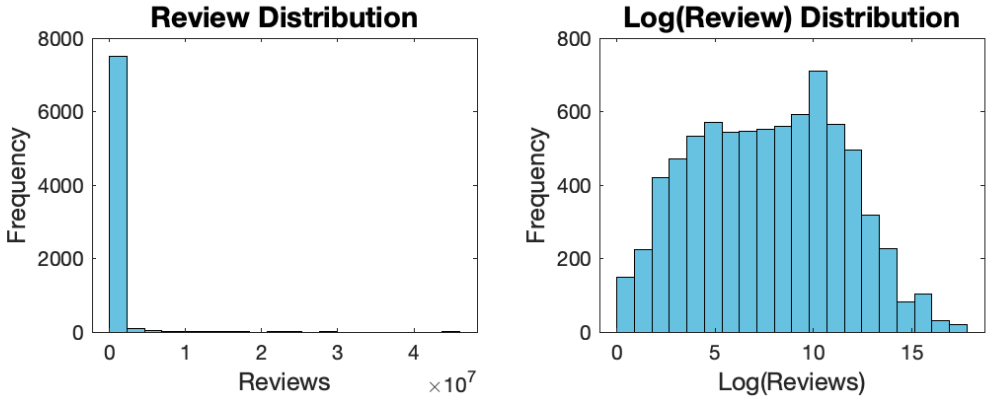

Fig. 14. Review Distribution (left) compared to the log transformation of review distribution (right).

Review				
Mean	Median	Mode	Min	Max
295,320	2367	2	1	44,893,888

Fig. 15. Review summary statistics table.

The number of reviews for each application on the Google Play App store is also extremely skewed see Fig. 14. The review variable had a very large range between only 1 review and almost 45 billion reviews on a single app, see Fig. 15. Because the range of this data is so large the mean is high but the mode number of reviews is only 2, see Fig. 15. In the histogram of Fig. 14 it is apparent visually how skewed the data is. The log transformation on the right in Fig. 14 has helped fix the extreme skew of this data and bring the range of data much closer together.

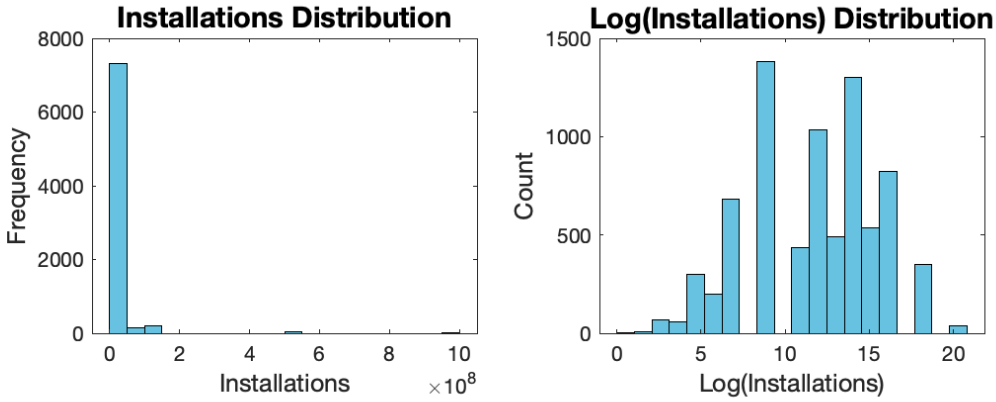

Fig. 16. Installation Distribution (left) compared to the log transformation of the installation distribution (right).

Installations				
Mean	Median	Mode	Min	Max
8,436,300	100,000	1,000,000	1	1,000,000,000

Fig. 17. Installations summary statistics table.

From the graph in Fig. 16, it appears that most applications do not have any installations, but that is not the case since the mode is 1 billion installations. It only appears that most of the data is close to 0 because of how big the range is along the x-axis. The range of installations goes from 1 to 1 trillion installations for a single app. This is causing the mean to be very high while the median and mode are much lower see Fig. 17. The log transformation in the graph on the left in Fig. 16 does appear to have less skewed data and will be considered going forward.

IV. VISUALIZATION OF 2 VARIABLES

A. Both Numeric

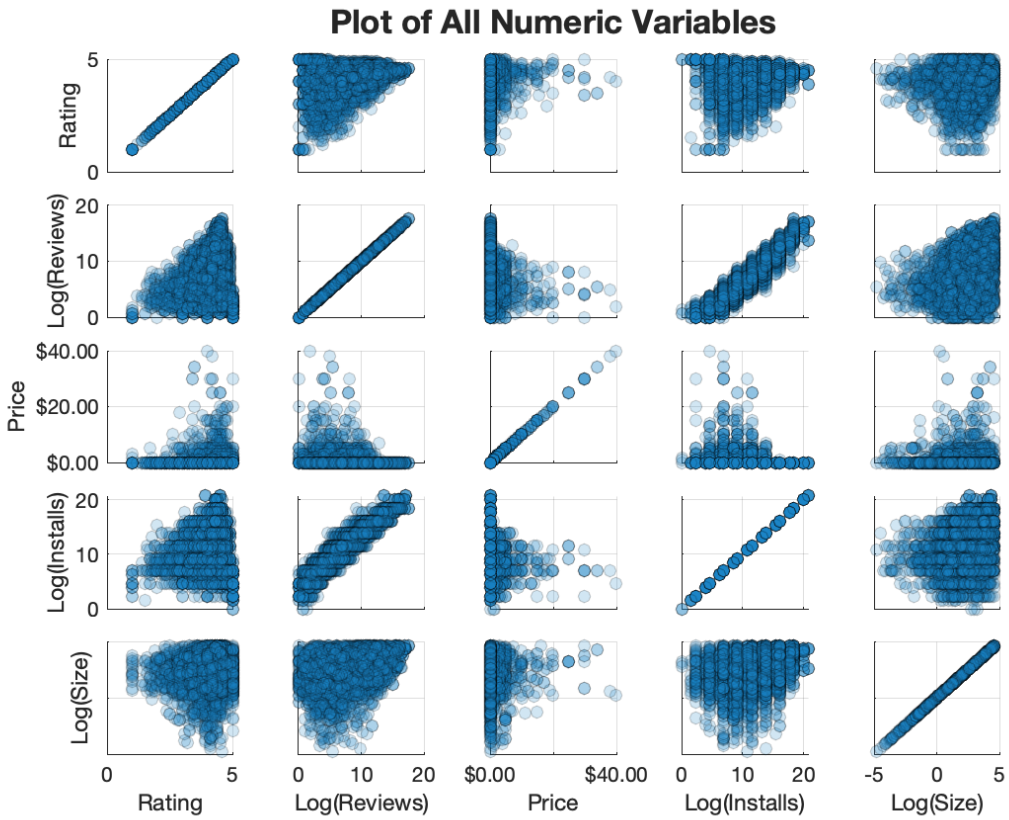

Fig. 18. All numeric variables plotted against each other.

We are interested in seeing if there is a relationship between any of the numerical variables. Refer to Fig. 18 to see all the variables plotted against each other. When comparing the number of installations directly to the number of reviews there is indeed a strong positive correlation and a linear relationship. This makes sense intuitively that apps that have more installations have also been reviewed more. For rating and number of reviews there is somewhat of a positive correlation similar to the graph comparing rating to the number of installations. While an app can have a high rating with a low number of reviews or a low number of installations it seems that most apps with many reviews or many installations also have a high rating. This could be due to the fact that apps with a higher rating could be attracting more people who may be more likely to install and then review them. The variables rating and log transformed size have a higher density of larger apps having high ratings. We can also see that apps with a higher price tend to be larger sized apps, with a lower number of installations, and a lower number of reviews, but higher ratings.

B. Numeric and Categorical

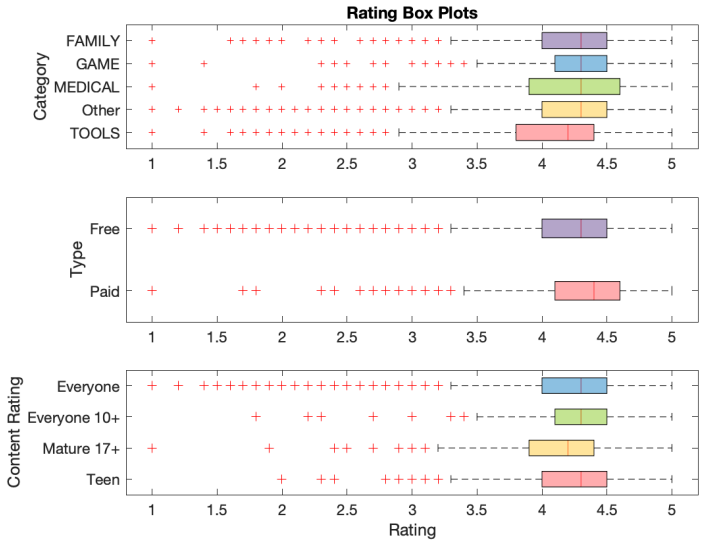

Fig. 19. App ratings grouped by category (top), type (middle) and content rating (bottom).

The rating for each application seemed to be pretty consistent across the different groupings for categories, types, and content rating, see Fig. 19. It was noticeable that paid apps seemed to have a slightly higher median rating and had fewer outliers among the lower ratings when compared to the free apps. We wondered if part of this was due to a common practice of paying for fake reviews on apps in the Google Play Store.

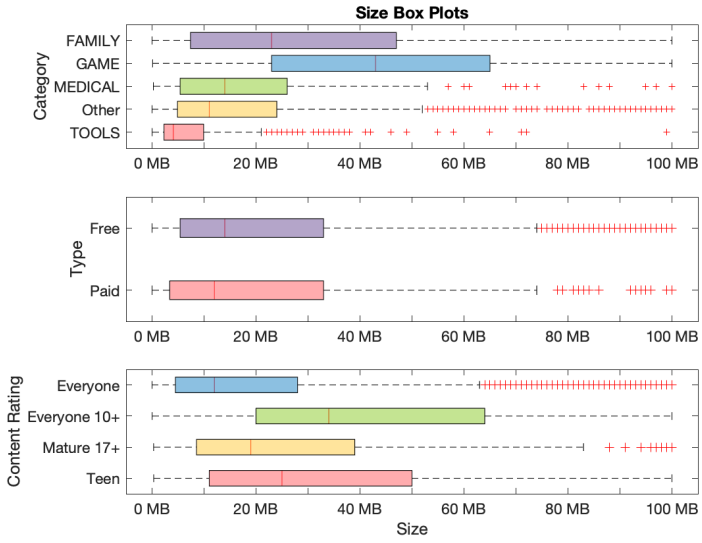

Fig. 20. App Size grouped by category (top), type (middle) and content rating (bottom).

Fig. 21. App Price for paid apps only grouped by category (top) and content rating (bottom).

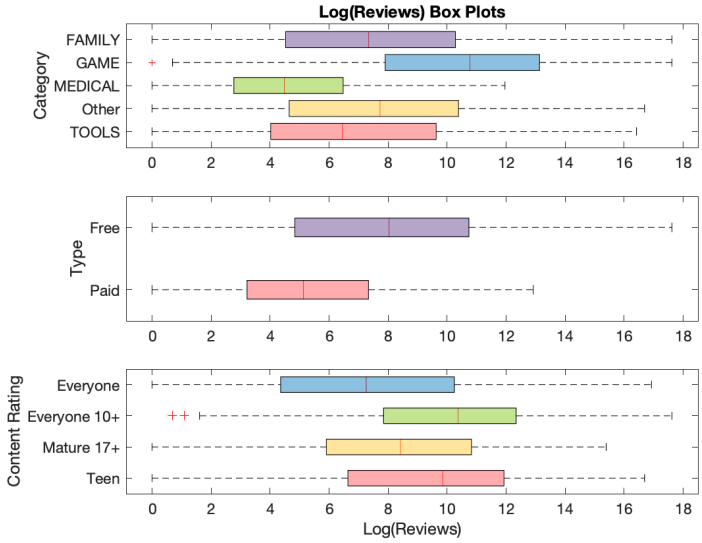

Fig. 22. App Reviews grouped by category (top), type (middle) and content rating (bottom).

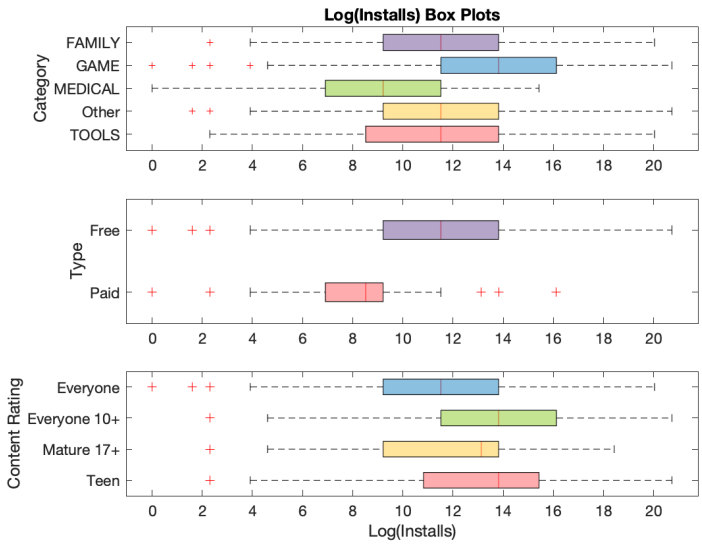

Fig. 23. App Installations grouped by category (top), type (middle) and content rating (bottom).

In Fig. 20, 22 and 23, the apps in the category “Game” have a larger median size, number of reviews and number of installations compared to all other categories. Apps in the “Medical” category have the lowest median size, number of reviews and number of installations compared to all other categories. This could be due to the fact that apps in the “Medical” category have a higher median price compared to all the other categories, refer to Fig. 21. The app category of “Tools” also had a much smaller size compared to all other categories.

Apps with a content rating of “Everyone” have the smallest median size, number of reviews and number of installations compared to apps with a different content rating, refer to Fig. 20, 22 and 23. This is interesting when referring back to Fig. 2 we see that apps with a content rating of “Everyone” make up 79.82% of all apps on the Google Play App Store. We also note that Fig. 22 and 23, confirm that free applications have a higher median number of installations and reviews.

V. VISUALIZATION OF 3 VARIABLES

A. Two Numeric and One Categorical

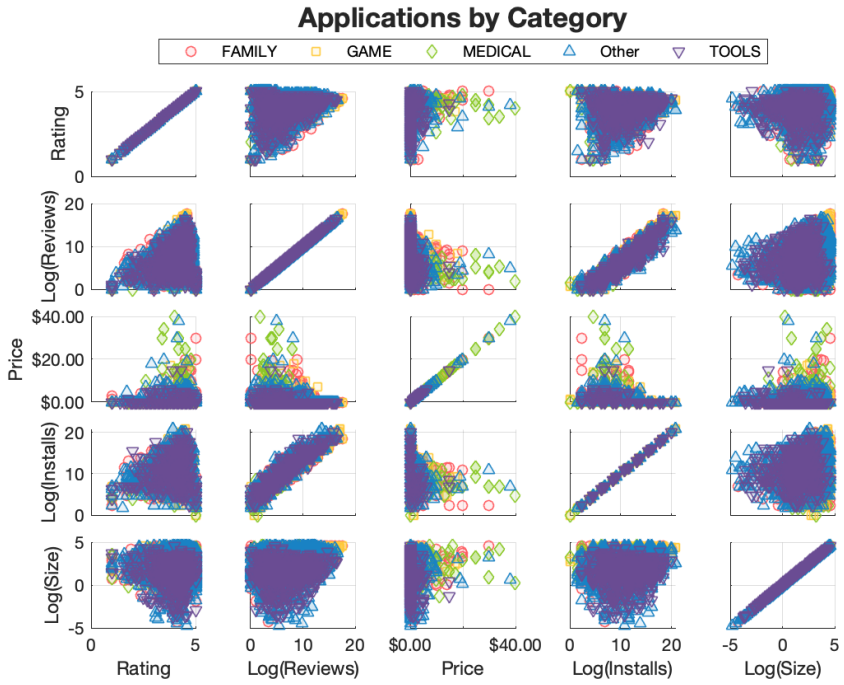

Fig. 24. All numerical variables plotted against each other and grouped by category.

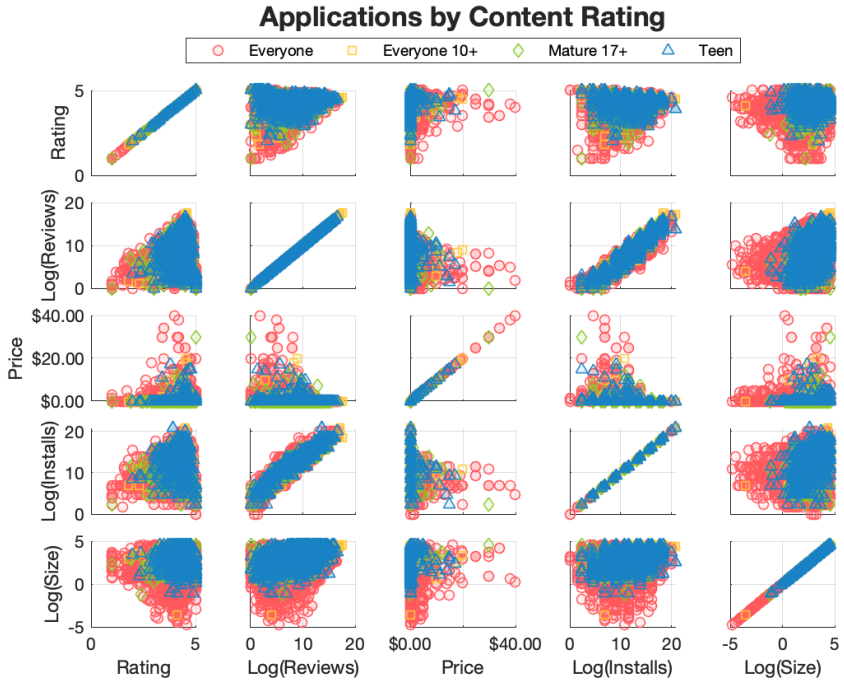

Fig. 25. All numerical variables plotted against each other and grouped by content rating.

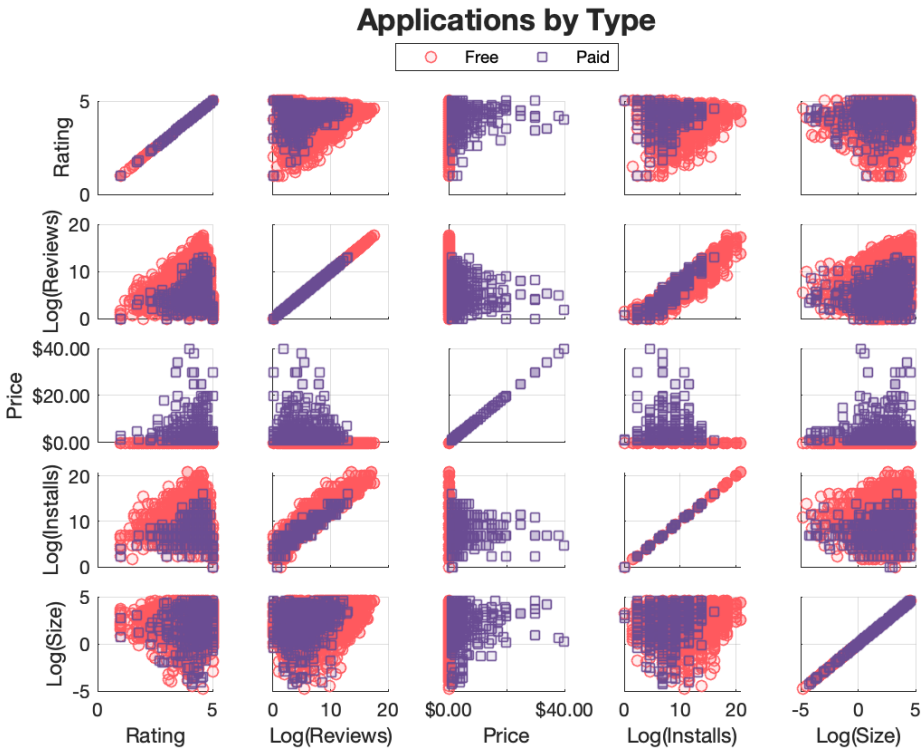

Fig. 26. All numerical variables plotted against each other and grouped by type.

Fig. 26 shows already what we have seen throughout the report between the relationship of free and paid apps. Although the two groups overlap in all of the plots we can notice where we see more free apps due to the free apps having a higher number of reviews, a higher number of installations, and lower ratings than the paid apps. We also see that because free apps have a price of 0 the plots comparing the price have all free apps along either the x = 0 or y = 0 axis.

A. Three Numeric

Originally we graphed our 3D numeric plots grouped by all the categories but we noticed that there may be 3 categories among those that have a larger difference between each other. It appeared that the categories “Game”, “Medical”, and “Tools” had data points with the least amount of overlap between these groups. This was confirmed by looking at the grouped box plots in Fig. 19-23. These categories all have at least one numeric variable with a significantly different median than all other categories. We wanted to see if there was a separation between these categories visually so we graphed 3D plots with apps that were only in the “Game”, “Medical”, or “Tools” category.

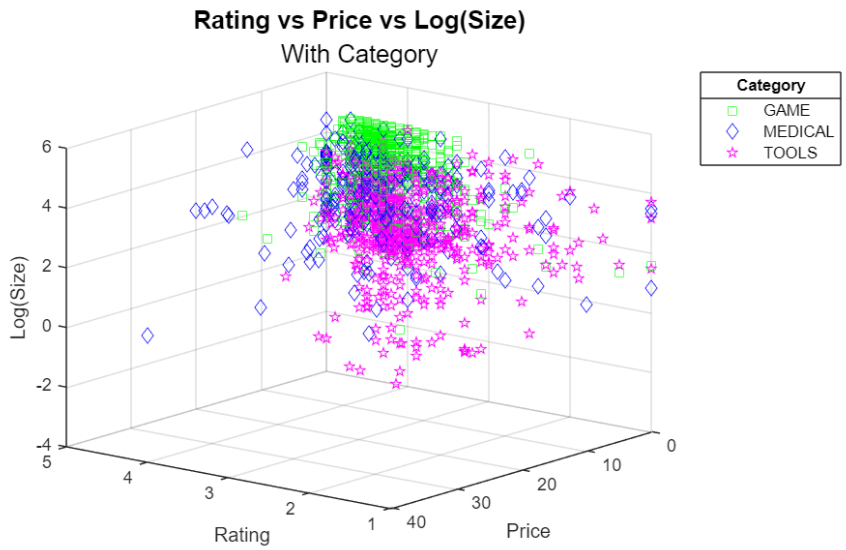

Fig. 27. Size vs rating vs price for apps in the "Game", "Medical" and "Tools" categories.

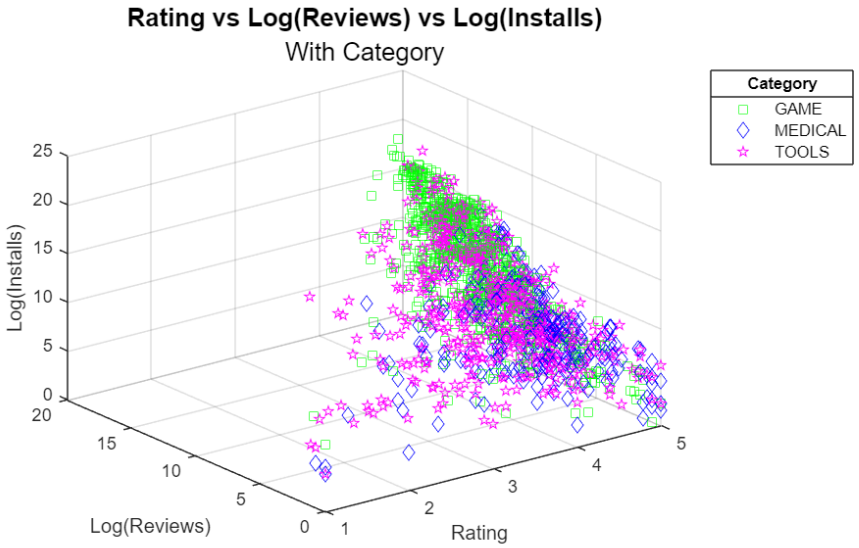

Fig. 28. Installations vs Reviews vs price for apps in the "Game", "Medical" and "Tools" categories.

Fig. 27 and 28 indicate that there is some separation between apps in the three categories chosen although there is still a lot of overlap between the data. We hoped that we could use a dimension reduction method to gain further insight into these 3 categories and the other categorical variables in our data set.

VI. DIMENSION REDUCTION METHODS

Before performing dimension reduction methods, we normalized the numerical variables to a common range [0,1] because there are big differences in the magnitudes of our data.

.

A. Principal Component Analysis (PCA)

The first set of PCA plots seen in Fig. 29 and 30 were done on app ratings, a log transformation of the number of reviews, a log transformation of the size, a log transformation of the number of installations and the price of the apps. These plots seemed to have a trail along the second principal component which we discovered to be due to the variation in the price of apps. We then did a PCA reduction again but this time with a log transformation of price + $1, refer to Fig. 31, 32 and 33 to see the difference. The PCA plots with a log transformation on the app price had a smaller variance along the second principal component. It also appeared to show a separation between two clusters of type “free” and “paid” apps. These clusters were even better defined when looking at a PCA dimension reduction plot using the first 3 principal components, refer to Fig. 33. In Fig. 33, not only is there a clear separation but there is very minimal overlap between the two groups.

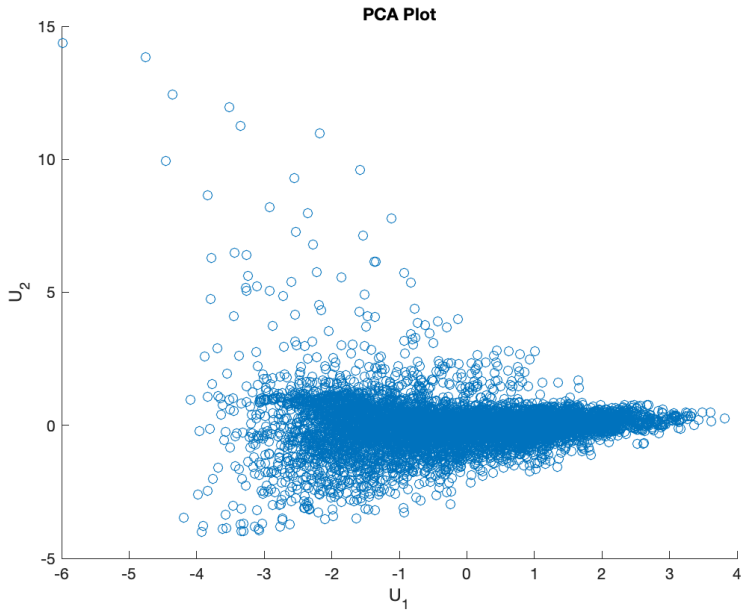

Fig. 29. PCA plot without a log transformation on price.

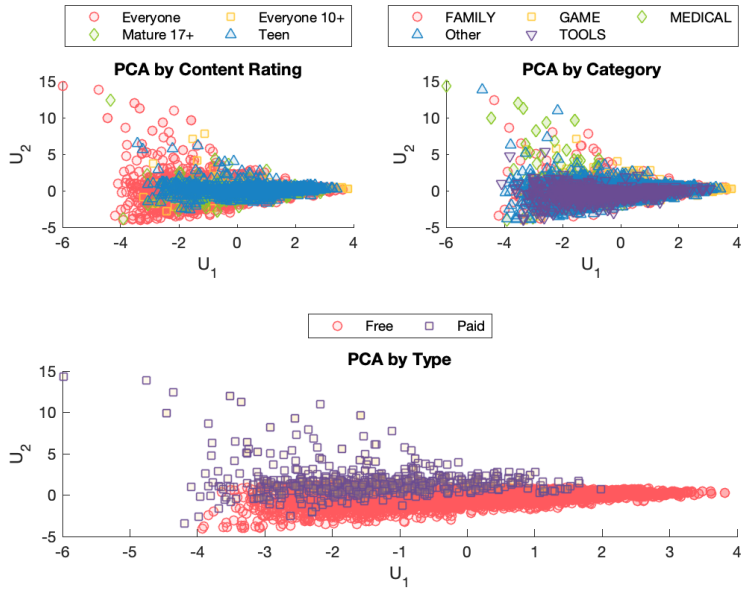

Fig. 30. PCA plot grouped by content rating, type and category without a log transformation on price.

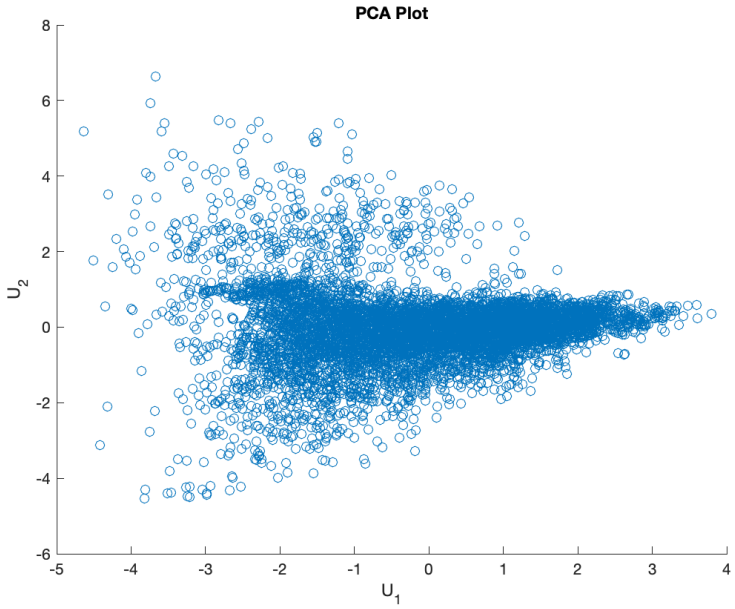

Fig. 31. PCA plot with a log transformation on price.

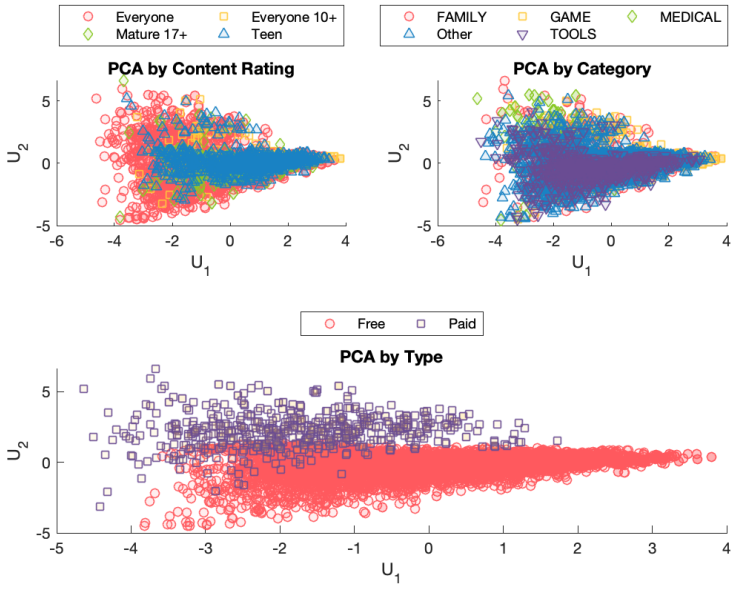

Fig. 32. PCA plot grouped by content rating, type and category with a log transformation on price.

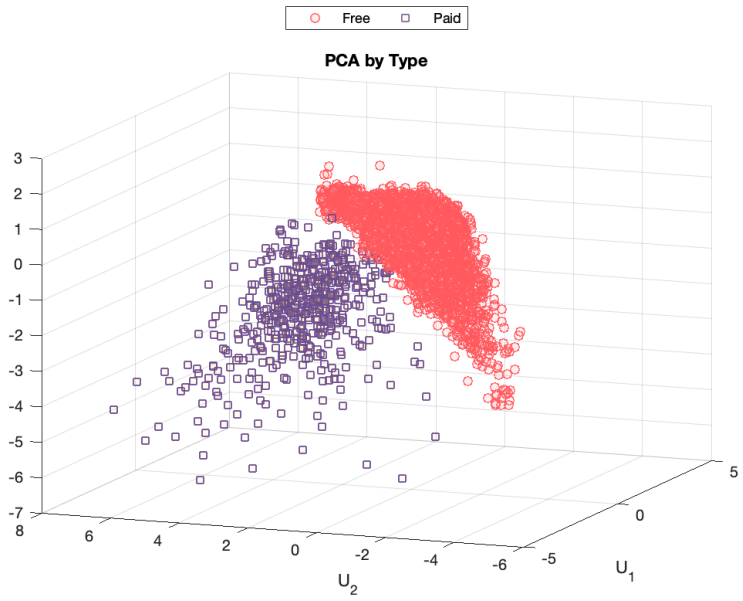

Fig. 33. 3D PCA plot grouped by type with a log transformation on price.

	Variable	PCA 1	PCA 2	PCA 3
1	"Rating"	0.1535	0.7314	0.6567
2	"Reviews_log"	0.6315	0.0374	-0.0839
3	"Size_MB_log"	0.3759	0.1514	-0.3580
4	"Price_log"	-0.2068	0.6577	-0.6538
5	"Installs_log"	0.6274	-0.0905	-0.0773

Fig. 34 Principal component coefficient table corresponding to PCA done in Fig. 31

	Percentage of total variance explained
1	45.2598
2	20.9743
3	17.6624
4	15.3609
5	0.7427

Fig. 35 Percentage variance explained by each principal component, corresponding to PCA done in Fig. 31

Fig. 34 shows the correlation between the principal components and the original variables. The first principal component is best correlated with two of the original variables which are log(Reviews) and log(Installs). This means that an application's number of reviews and installation will increase and decrease together. The second principal component is best correlated with the original variables rating and log(Price + $1). This means that an application's price and its rating will increase and decrease together as well. From Fig. 35, we see that 45% of the variance is explained by the first principal component. A total of approximately 84% of variance is explained by the first three principal components.

Next, we wanted to see the PCA plot for apps in the "Game", "Medical", or "Tools" category because of our early analysis. The best visible separation we saw for a PCA plot grouped by the three categories was when we used the first three principal components, see Fig. 36. There is still a lot of overlap and not a very clear separation between these app categories.

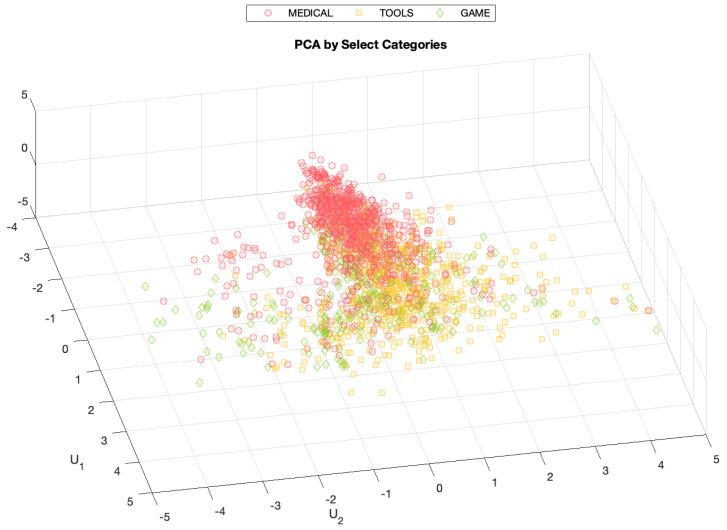

Fig. 36. 3D PCA plot grouped by apps in the "Game", "Medical", or "Tools" category.

B. Linear discriminant analysis (LDA)

Fig. 37. 1D LDA plot for free and paid apps.

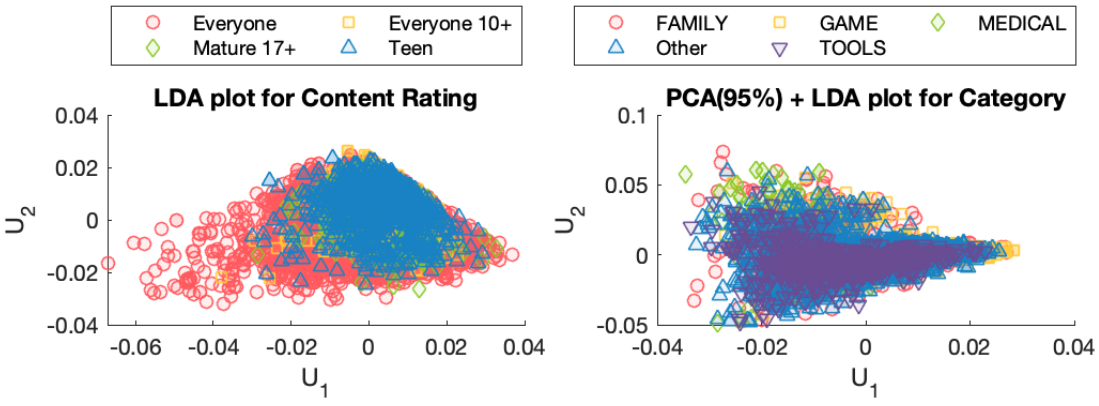

Fig. 38. 2D LDA plots for 4 different content ratings (left) and 5 different categories (right).

Using an LDA dimensions reduction method did not show much separation between categories, content rating or even type see Fig. 37 and 38. This is likely due to the heavy overlap

of data between the different groups and non linear separation. LDA seeks to project the data onto a line or plane that maximizes the difference between each group mean and the minimizes the variance within each group. This means that this method will not work well for data that has a lot of overlap across all dimensions.

C. Multidimensional Scaling (MDS)

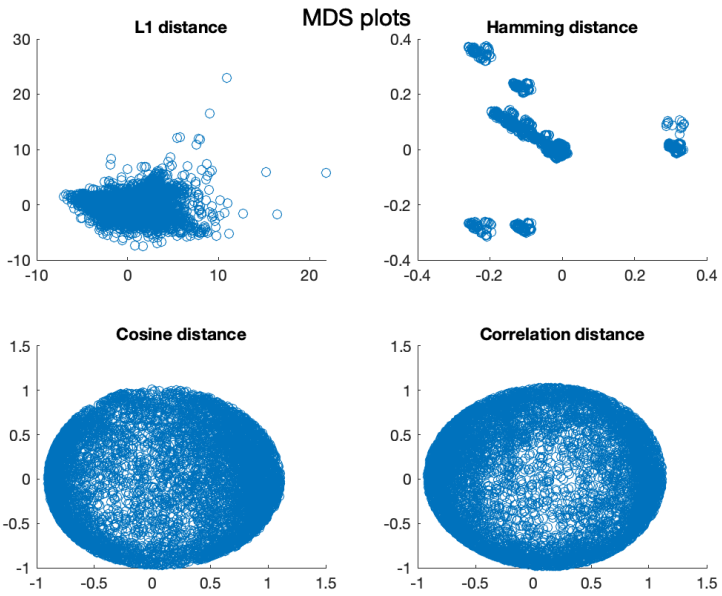

Fig. 39. MDS done on pairwise distances using L1, hamming, cosine and correlation distances.

$$\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}} \tag{1}$$

$$1-\frac{x\cdot y}{||x||_2||y||_2} \tag{2}$$

$$max(|x_1-y_1|,\ldots,|x_n-y_n|) \tag{3}$$

$$\sum_{i=1}^{n}|x_i-y_i| \tag{4}$$

Eq. 1-4 were used to compute a matrix of pairwise distances between the data points that would be used for classical multidimensional scaling. Eq. 1 is correlation distance, Eq. 2 is cosine distance, Eq. 3 is chebyshev distance and Eq. 4 is the L1 distance. To find the dissimilarity vectors in euclidean space the Matlab built in function of cmds() was used. Fig. 39 shows the resulting graphs. There does not appear to be any clear separation using MDS on cosine, correlation or L1 pairwise distances by themselves so we decided to color the points by groups.

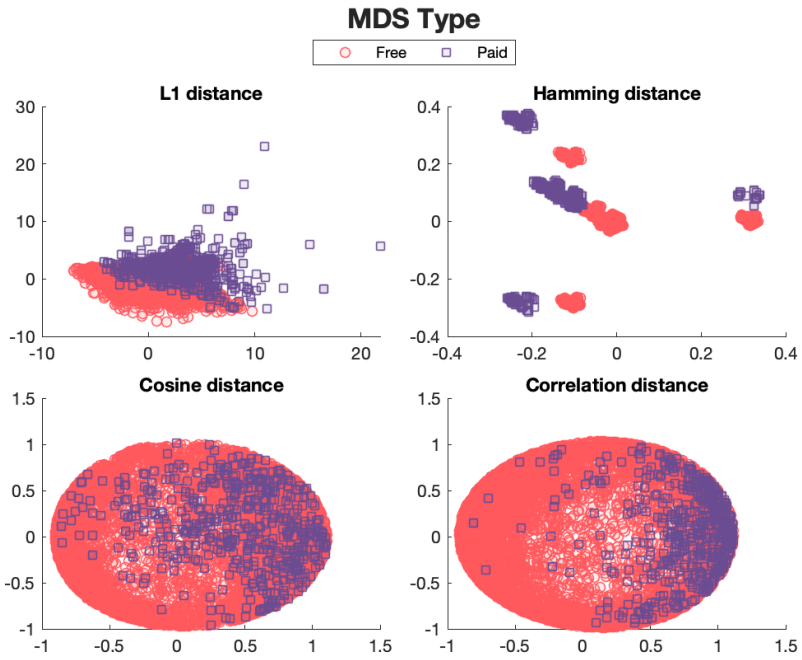

Fig. 40. MDS done on pairwise distances using L1, hamming, cosine and correlation distances grouped by type.

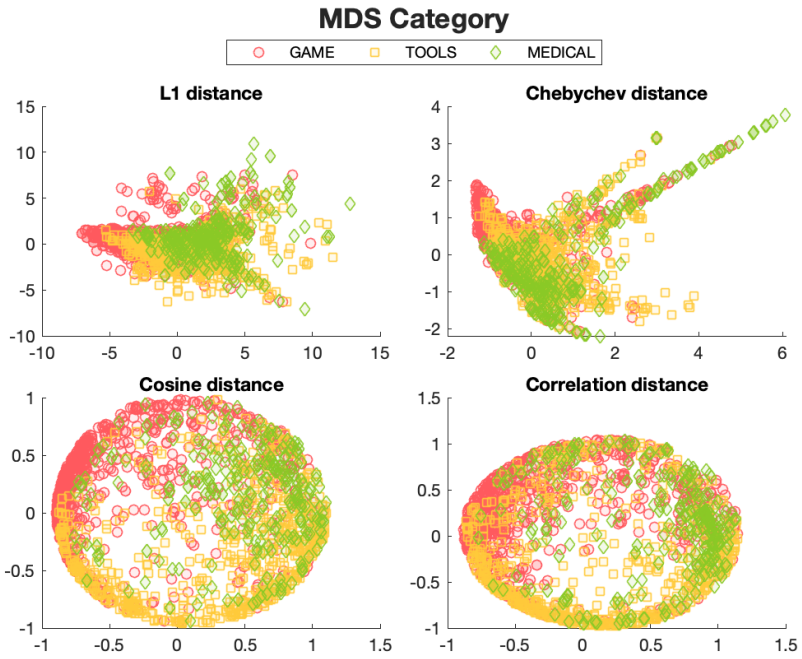

Fig. 41. MDS done on pairwise distances using L1, hamming, cosine and correlation distances grouped by category.

Fig. 40 shows that there is indeed some separation if the data is grouped by type. Hamming distance has the most well separated clusters with almost no overlap but we were unable to determine what grouping each individual cluster fell under. To see other attempted groupings, see Appendix B. Fig. 46 and 47.

In Fig. 41, we once again tried to see if there was a separation between the 3 most distinct app categories “Game”, “Medical”, and “Tools”. In the top two graphs with L1 and Hamming distances it appears each category has the same shape but shifted over. The bottom two graphs with cosine and correlation distances do have some separation. The top left edge of the circle is mostly apps in the “Game” category, the right section of the circle is mostly apps in the “Medical” category, and finally the bottom left is mostly apps in the “Tools” category. Because it appeared there might be some separation, we tried MDS again in the third dimension. Only the MDS graph with the distance of L1 showed what we were looking for which can be seen in Fig. 42. This graph did show that despite there still being some overlap we were seeing clear clusters forming for the 3 categories.

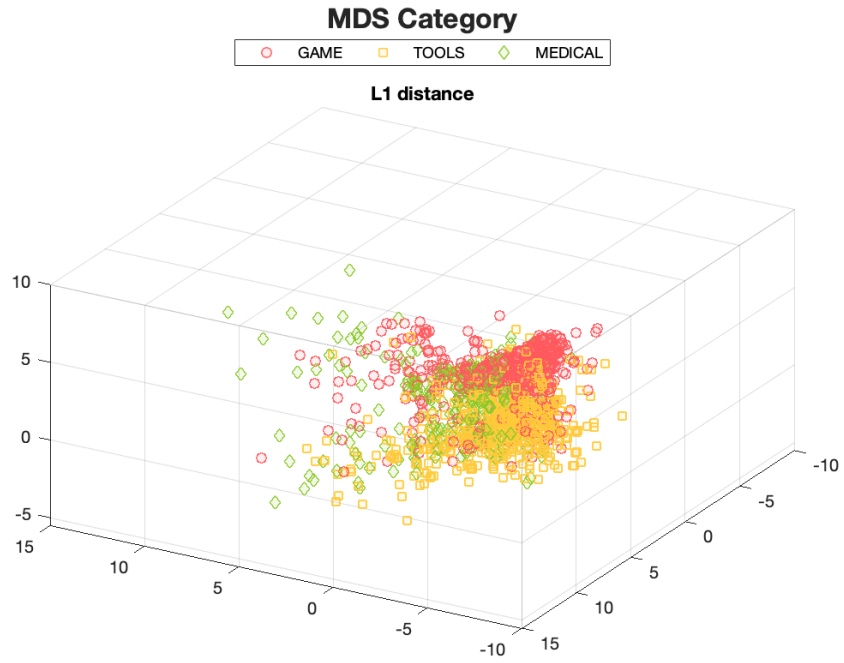

Fig. 42. MDS done on pairwise distances using L1 grouped by category.

D. Laplacian Eigenmaps

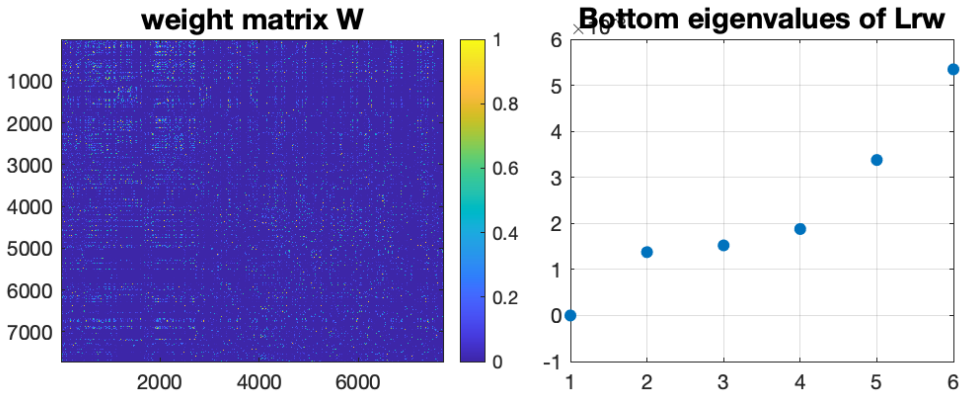

Fig. 43. Weight matrix with 10-nearest neighbors (left) and the bottom eigenvalues of Lrw (right).

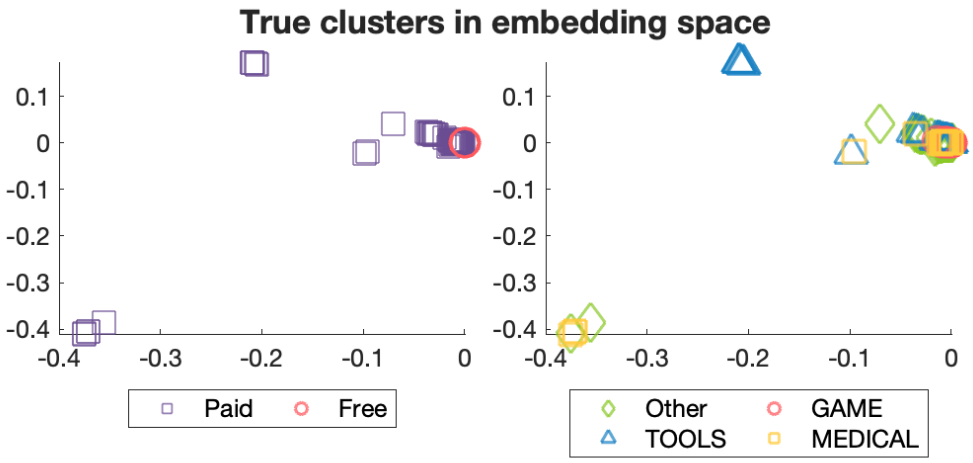

Fig. 44. True clusters in embedding space grouped by type (left) and category (right).

For the weight matrix in Fig. 43, we noticed more connections occurred in the upper right hand corner of the graph but no real patterns emerged from this graph. We believe this data set to be very noisy with a lot of overlapping data. In Fig. 44, there seems to be clusters within the apps that cost money with one of the clusters including and overlapping with the apps that are free. Within the apps that are paid there seems to be a tight cluster of paid apps in the "Tools" category, there is another tight cluster on the bottom right of both graphs that is paid but has a mix of categories. Because the categories overlap in most clusters, it is possible that what we are seeing is outlier and not a real group of clusters. It is also possible that our data does have several unique clusters but we have yet to determine what defining factors separate those apps from the rest.

VII. CONCLUSION

In this project, we have gone through preprocessing, cleaning, transforming, visualizing data, and performing data reduction methods. Some outliers were detected and removed and log transformations used on several of the variables. There are some interesting relationships between different features of apps on the Google play store. The apps that were the most distinct from each other with the best separations were the free and paid apps. Next, we were able to see some distinction and separation between three app categories "Game", "Medical" and "Tools". Within the three categories and within the app types, there was too much overlap to be able to fully separate these groups.

For further research, we would've liked to try keeping some of the missing entries we removed by using different approaches such as replacing missing data with median or mean. We also want to try to analyze the csv that is included in the Google Play store dataset which contains the most relevant 100 reviews for each app, possibly with natural language processing.

VIII. REFERENCES

[1] O. Lengkong and R. Maringka, "Apps Rating Classification on Play Store Using Gradient Boost Algorithm," 2020 2nd International Conference on Cybernetics and Intelligent System (ICORIS), 2020, pp. 1-5, doi: 10.1109/ICORIS50180.2020.9320756.

[2] J. Businge, M. Openja, D. Kavaler, E. Bainomugisha, F. Khomh and V. Filkov, "Studying Android App Popularity by Cross-Linking GitHub and Google Play Store," 2019 IEEE 26th International Conference on Software Analysis, Evolution and Reengineering (SANER), 2019, pp. 287-297, doi: 10.1109/SANER.2019.8667998.

[3] https://www.kaggle.com/datasets/lava18/google-play-store-apps

APPENDIX A

High App Prices

App Name	Price
Vargo Anesthesia Mega App	79.99
Vargo Anesthesia Mega App	79.99
most expensive app (H)	399.99
I'm rich	399.99
I'm Rich - Trump Edition	400
I am rich	399.99
I am Rich Plus	399.99
I am rich VIP	299.99
I Am Rich Premium	399.99
I am extremely Rich	379.99
I am Rich!	399.99
I am rich(premium)	399.99
I Am Rich Pro	399.99
I am rich (Most expensive app)	399.99
I Am Rich	389.99
I am Rich	399.99
I AM RICH PRO PLUS	399.99

Fig. 45. Tables of the 17 highest cost apps and their prices.

APPENDIX B

MDS Graphs

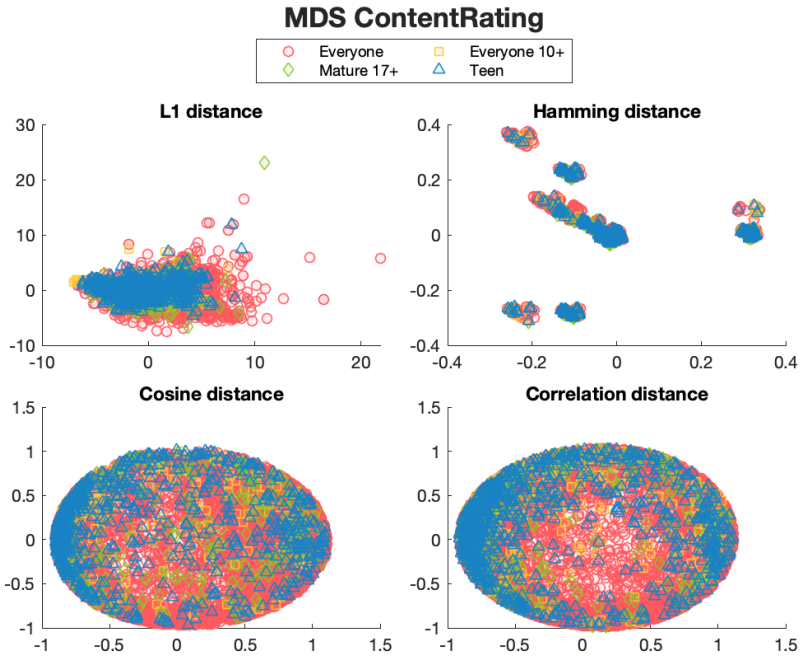

Fig. 46. MDS graphs for four different distance measures grouped by content rating.

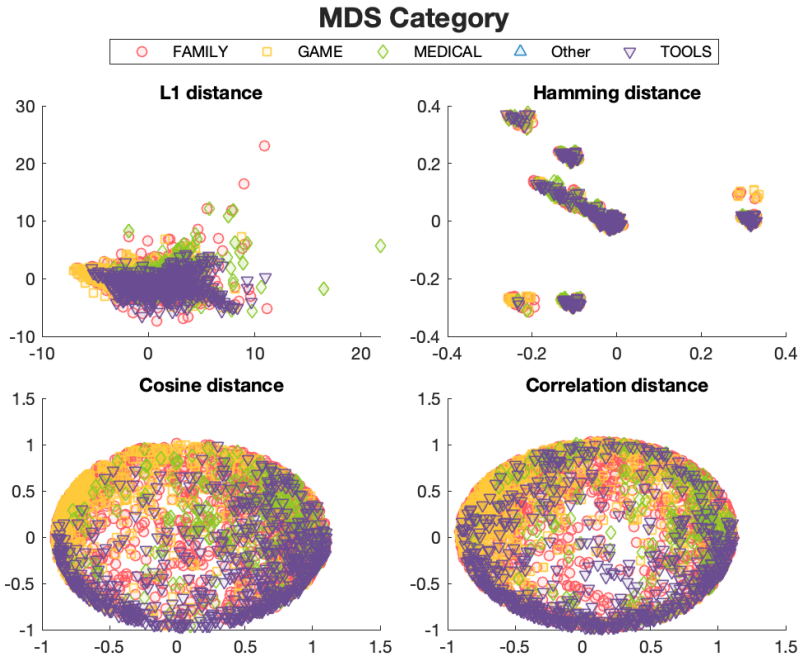

Fig. 47. MDS graphs for four different distance measures grouped by categories.