Data Analytics Programming Report

Amey Athaley
Javeria Rangoonwala
Sahana Subramanian
Victor Nguyen

1. Description of the project goals

In this project, we investigate the different variables of Apps on Google Play Store that affect the application and the top 100 most relevant user reviews. We attempt to use our analysis to answer the following questions:

1. What makes an app popular?
2. What are some interesting trends that we can observe about user behaviour & sentiment on app usage?
3. Can we predict app popularity if given a set of features about the app?

Mobile app market has grown ~20% in the last 4 years. Android apps market comprises 90% of its market share. Our analysis has the potential to be scaled and applied to identify and solve the following problems for the much wider mobile app industry:

1. Prediction of app popularity to gauge revenue generated & optimize investment strategy for app development
2. Identifying untapped segments in the app market
3. Identifying fake/junk apps that spam the play store

2. Data Cleaning

We started the analysis by creating a data dictionary to understand the structure of the dataset and what each feature represents. Post that, we handled the missing values in some of the columns by either dropping the rows or imputing them with the mode values, depending on the percentage of nulls in each feature. We transformed and some of the columns like Installs, Size and Price to numeric type for ease of analysis.

3. Exploratory Data Analysis

After establishing a good sense of each feature, we proceeded with plotting a pairwise plot between all the quantitative variables to look for any evident patterns or relationships between the features.

3.1. Android Market Breakdown

We broke down the apps by category and found that the Family and Game categories have the highest market prevalence. The Business, Tools and

Medical apps are also catching up. We also checked how each category performed in terms of number of reviews, size and installation count. We found that the gaming category has the maximum number of reviews and install count but the family apps consume more space in the play store.

3.2. Sizing Strategy

We analysed the sizing distribution of the top-rated apps (rating greater than 4.5) and observed that most top-rated apps are optimally sized between ~2MB to ~40MB i.e., neither too light nor too bulky. We found that the Game and Family categories have the highest number of bulky apps. We also observed that despite this, these bulky apps are fairly highly rated indicating that they are bulky for a purpose. Majority of the paid apps that are highly rated have small sizes which implies that most paid apps are designed and developed to cater to specific functionalities and hence are not bulky. Users prefer to pay for apps that are light-weighted. A paid app that is bulky may not perform well in the market but it also depends on the category of the app.

3.3. Pricing Strategy

The initial intuition for pricing strategy for apps to be popular would be to make them free. That is confirmed by the data as free apps dominate the store at 92.6%, making it over 12.5 times their paid counterparts. It follows that cheaper apps are also more popular which is supported by the fact that 90% of paid apps are less than $10 and 80% of apps are less than $5. Categorically, the most expensive apps are Family and Medical apps which would suggest that people are willing to pay more for medical apps, possibly due to a belief in a reliability or highly specialized applications.

When cleaning the data, one of the major outliers we found were useless apps priced over $250 which we considered ‘junk’ apps. The names of these apps were typically something to the effect of “I am rich” which would do nothing other than prove the user paid an exorbitant amount it. An interesting insight for app developers for games is that all of the games are within $20.

3.4. Install Count and Rating

To answer the question of whether highly installed apps are rated better, we plotted the install count and rating. We saw that there was no pattern in apps that had installs less than 500 Million. However, for apps with installs greater than this threshold, the minimum rating rises to 3.7 out of 5. To confirm this, we bucketed the apps into bins based on number of installs and plotted the minimum rating for each bucket. We observed that the lowest rating for an app increases with the number of installs. We can say that popular apps are less likely to be rated badly.

3.5. Rating Distribution

The rating distribution revealed that most apps perform reasonably well with an average rating of 4.17. We broke down the average rating by category to check if any category performs exceedingly good or bad. We conducted a One-way Anova Test and confirmed that the average ratings across categories is statistically different. The Health and Fitness and Books and Reference produce the best apps with 50% apps having a rating greater than 4.5. Interestingly, half of the Dating apps have a rating lower than the average.

3.6. Basic Sentiment Analysis – User Reviews

We plotted the fraction of positive, negative and neutral reviews for each category and observed that the Health and Fitness apps perform the best with more than 85% positive reviews. On the other hand, Game and Social apps have a higher fraction of negative reviews. We compared the reviews between free and paid apps and found that people are harsher towards free apps whereas users are more tolerant when they are paying for it.

3.7. Frequency of Words in Reviews

We created a word cloud of commonly occurring words in positive and negative reviews and found that the words – "love", "great" and "good" were the most commonly occurring words in the positive reviews. On the other hand, the negative words that were prevalent were "bad", "hate" and "ads".

Our aim was to analyse the reviews and get a better idea of the common issues that people face with apps or the attributes that make an app popular. We extracted phrases from the reviews and observed that positive reviews had phrases like "user friendly", "free version", "works great" and "highly recommend". The negative reviews contained phrases like "waste time", "many ads", "spend money" and "takes forever". We can see that loading time and ads were one of the main concerns amongst users. On the other hand, usability is one of the reasons that users give positive reviews.

4. Predictive Modelling

We proceeded to predict the popularity of app based on its features and chose the install count as the measure of popularity. We split the number of installs into four buckets – Not so popular, Intermediate Popular, Popular and Extremely Popular, so that we have a classification problem at hand.

4.1. Feature Selection

We plotted a Pearson Correlation Matrix which showed a moderate positive correlation exists between the number of reviews and the number of installs. Our initial analysis also revealed that the size of the app affects the rating. Most of the top-rated apps were sized between 2 Mb and 40 Mb. Intuitively, the number of reviews and size of the app would explain the majority of the popularity of the app. The rating distribution showed that categories had statistically different mean ratings which in turn contributes to the popularity. We decided to use the features - ‘Installs’, 'Reviews', 'Category', 'Content Rating', 'Type', 'Genres' and 'Size'. We dropped the features – Android Ver, Current Ver and Last Updated since these did not have any significant relationship with the install count.

4.2. Implementation

We implemented three classification models to predict the popularity of the app, amongst which Random Forest with tuning gave us the best accuracy with 81.08 % accuracy.

Logistic Regression – 58.67% accuracy
Decision Tree Classifier – 76.10% accuracy
Random Forest Classifier – 78.16% accuracy
Random Forest with Hyperparameter tuning - 81.08% accuracy

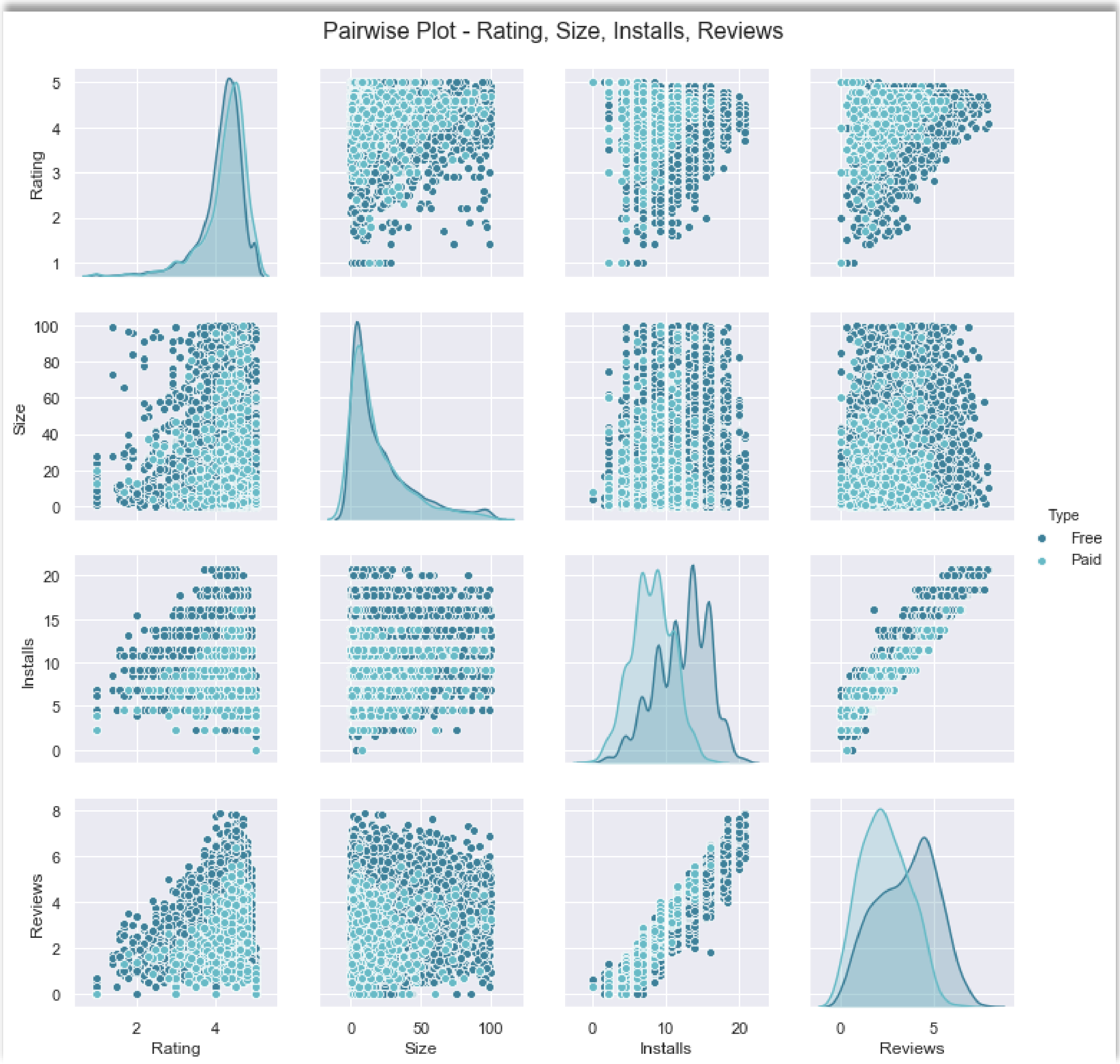
We used Randomized Search CV to loop over a set of random values for the parameters of the model. We ran the grid search and chose the optimal number of tress, number of features at each split, maximum number of levels and the method of selecting samples. Using these parameters, we predicted the popularity of the app on our test set and obtained an accuracy of 88.05%.

5. Findings and Conclusion

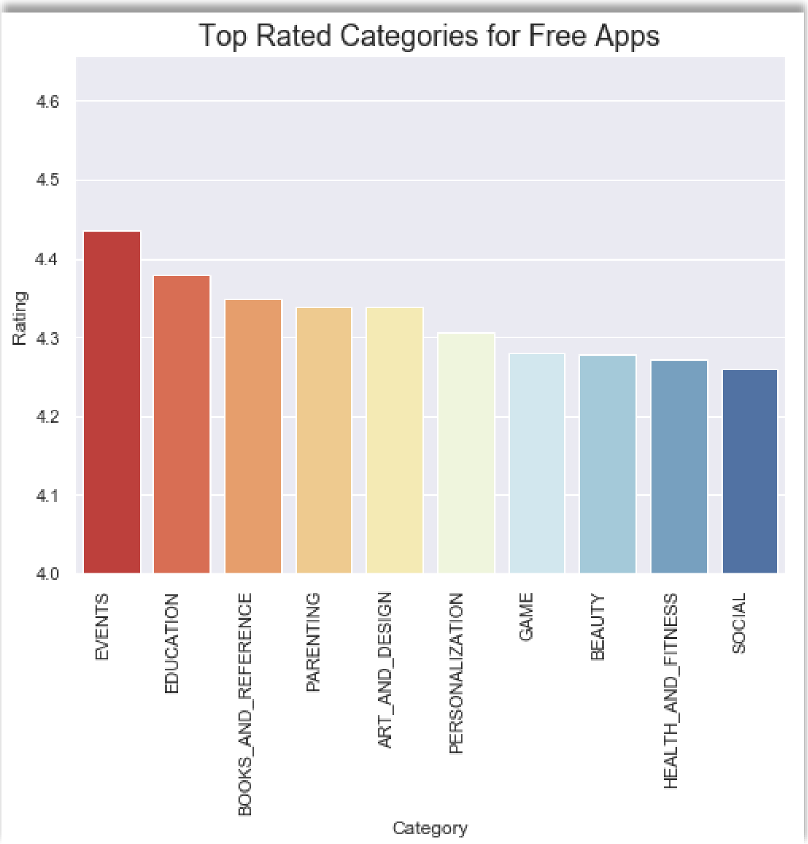
1. Developers should aim to keep the apps optimally sized (between 2 MB to 40 MB) to increase the likelihood of it being popular.
2. Paid apps designed to fulfil specific functions tend to be small sized and are more likely to meet user expectations.
3. Majority of the apps should be free. For apps with no ads, apps can be priced under $10.
4. Free apps tend to outperform paid apps both in volume, install counts and ratings.
5. There is a positive correlation between installs and ratings
6. The category (Game) is a potential unsaturated space for app developers. Apps in this category tend to have high installs.
7. Highly installed apps such as Communication, Productivity & Photography are not as highly rated indicating dissatisfaction among customers and potential untapped market segments.
8. For paid apps, users review harshly.
9. Basic sentiment analysis revealed some common issues like loading time and inconvenience with apps may contribute to a negative rating. Usability was one of the main reasons apps were rating positive.

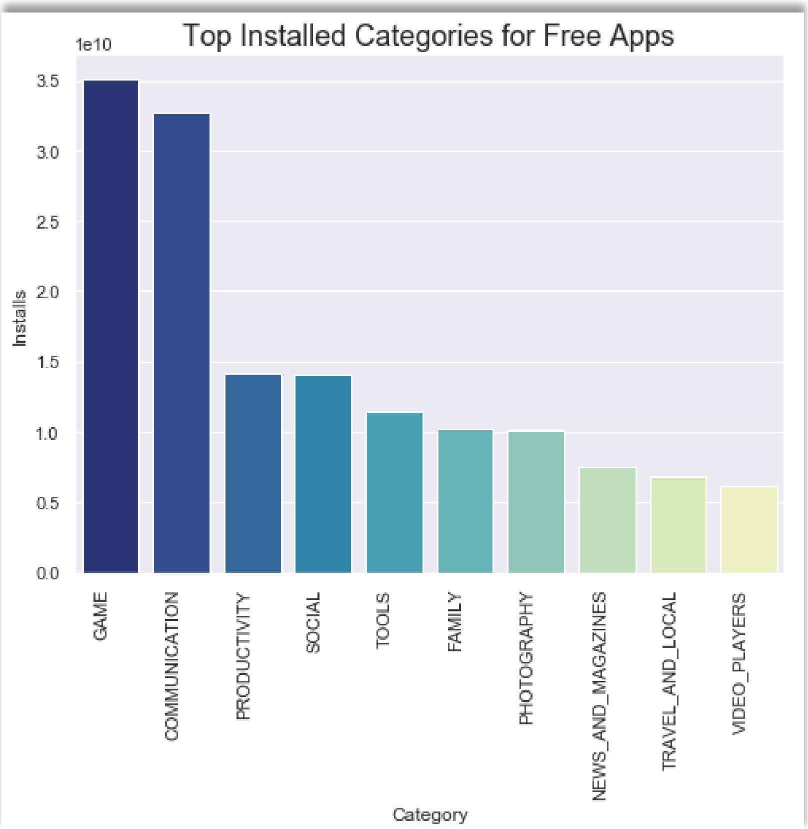
Appendix

3.0. Pairwise Bivariate plots

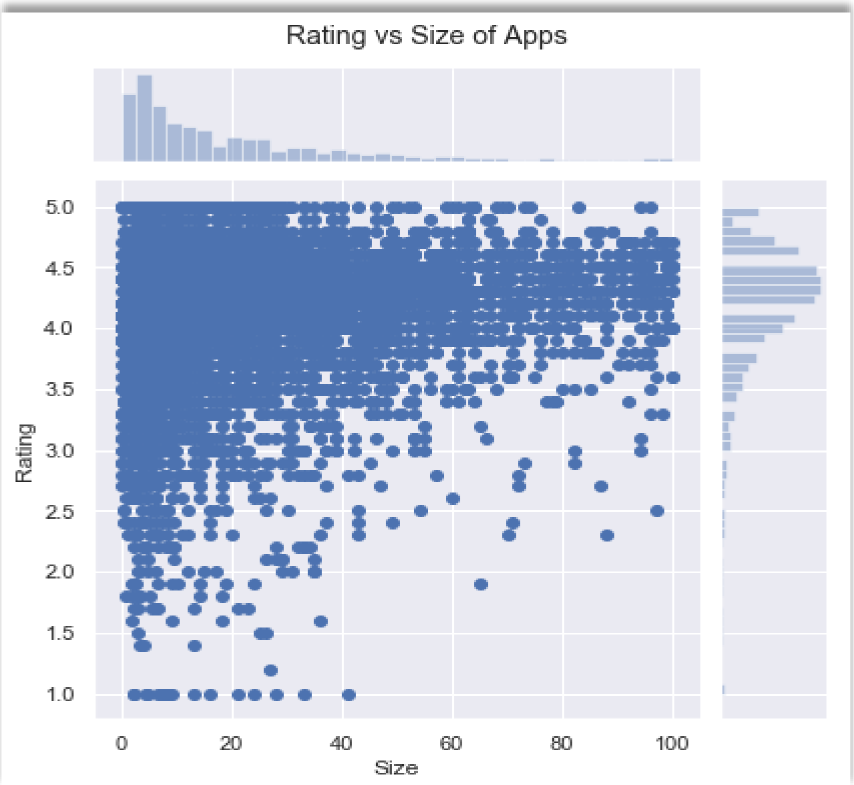


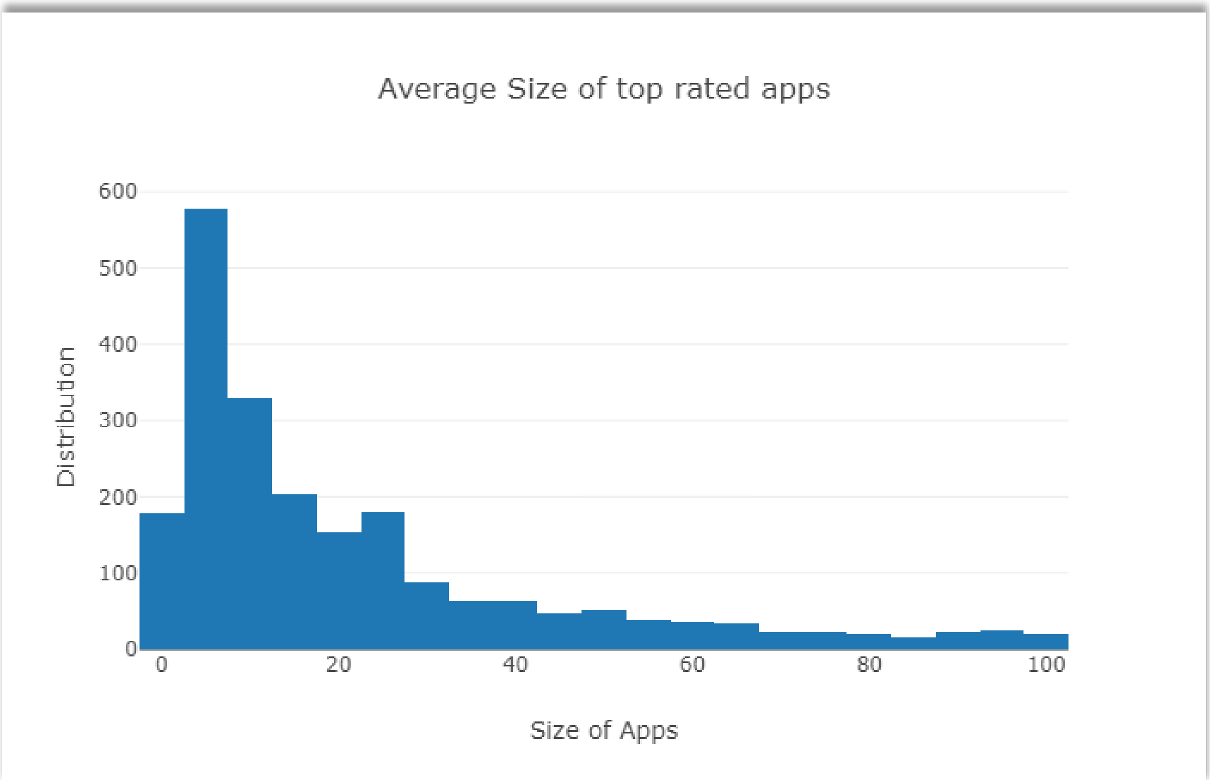
3.1. Android Market Breakdown

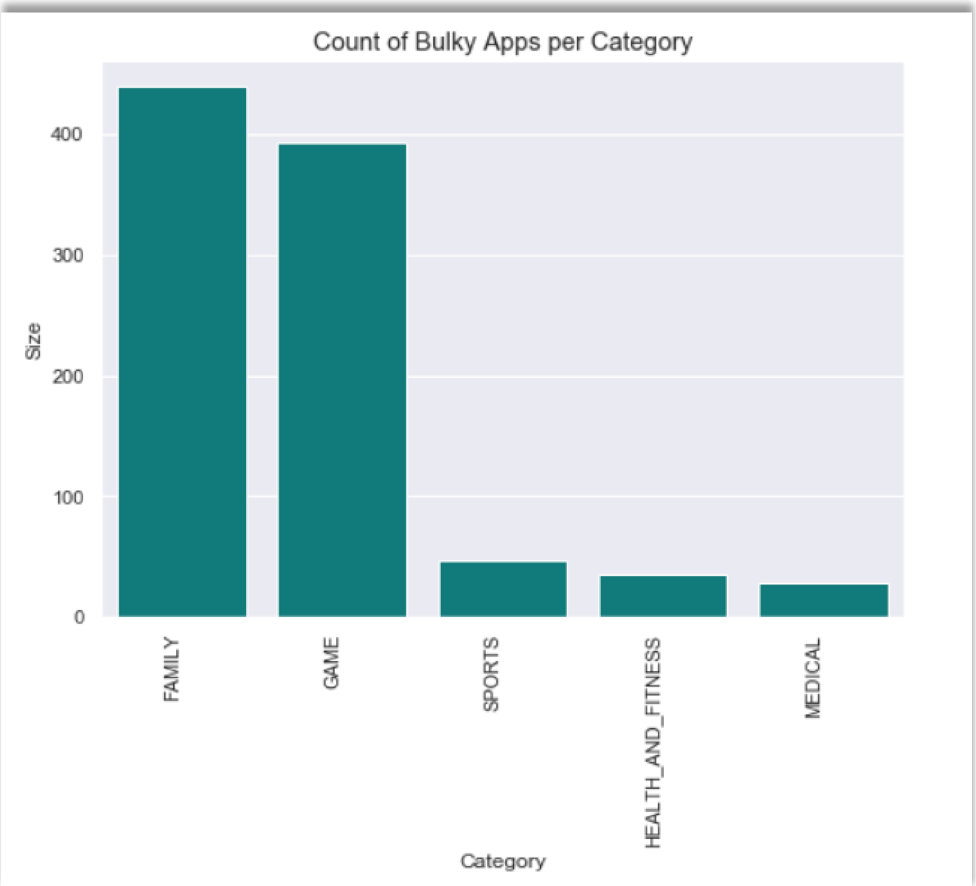


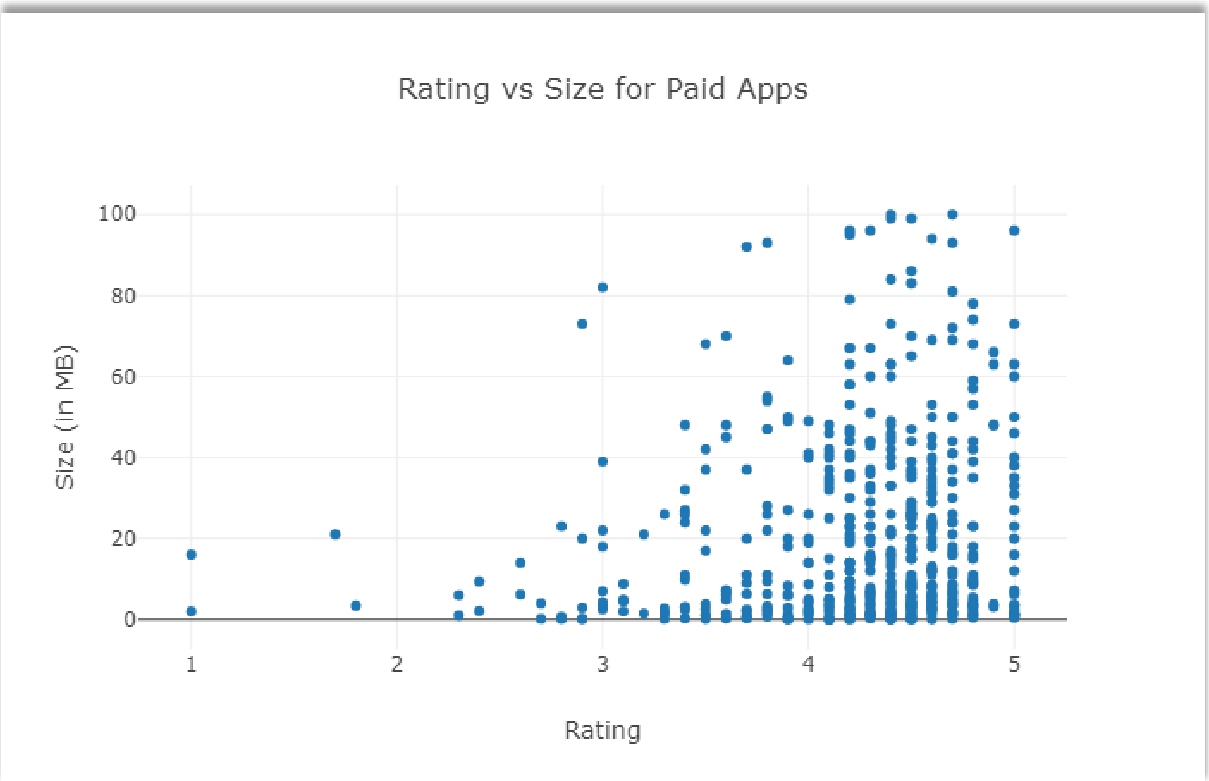


3.2. Sizing Strategy

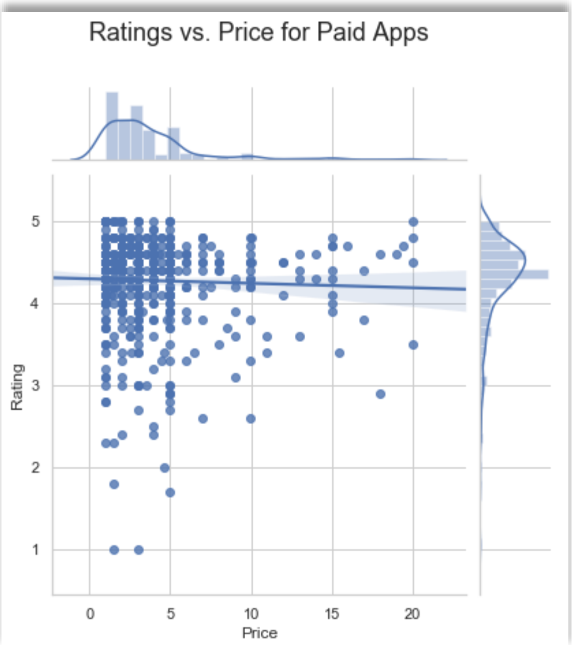


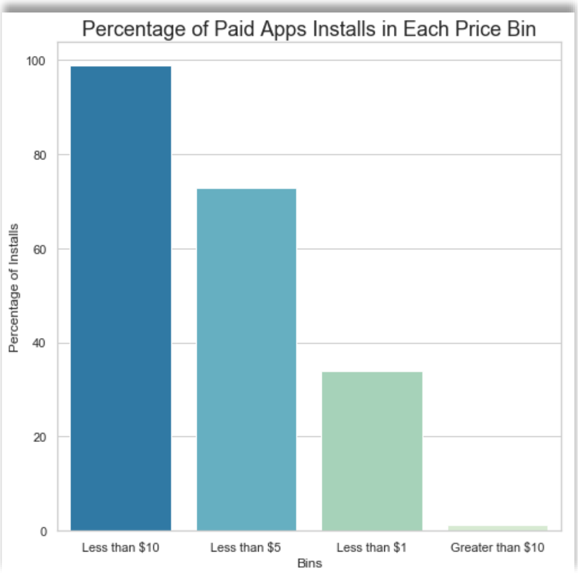


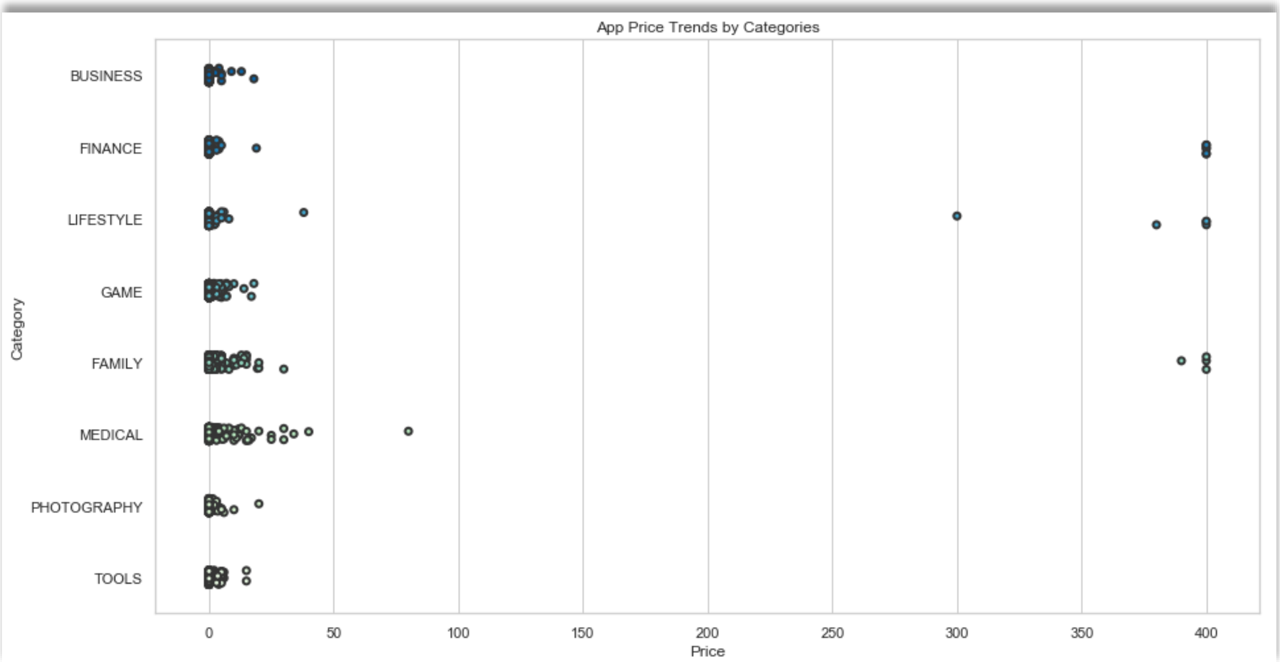

Count of Bulky Apps per Category
Size
0
100
200
300
400
FAMILY
GAME
SPORTS
HEALTH_AND_FITNESS
MEDICAL
Category

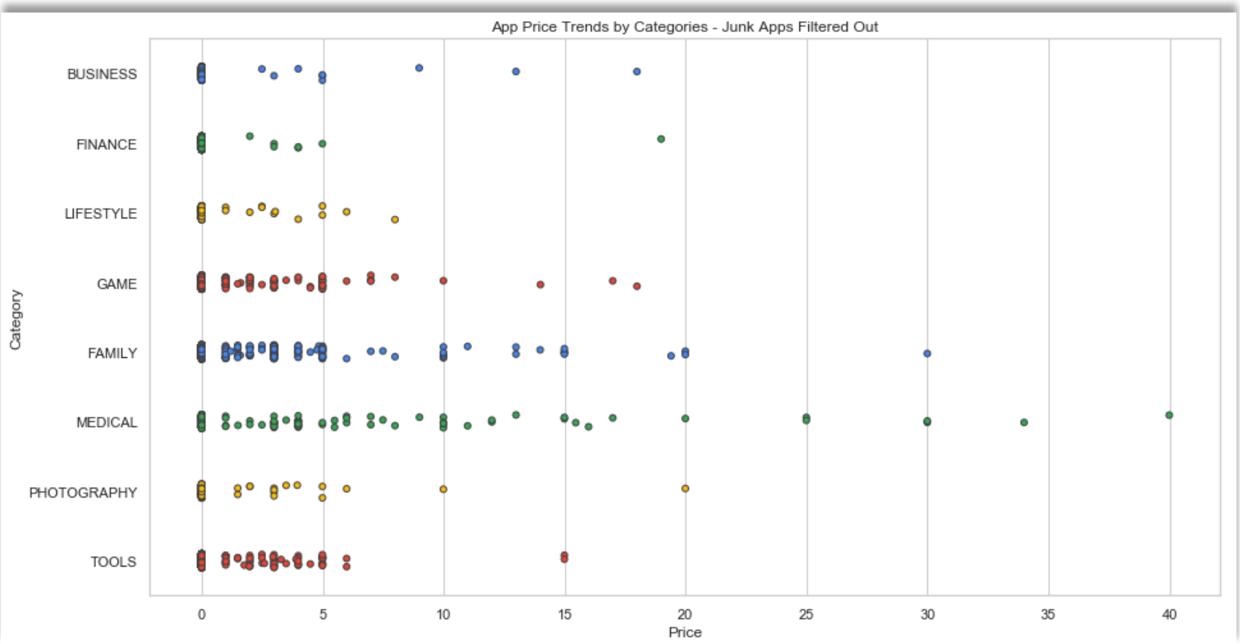

Rating vs Size for Paid Apps
Size (in MB)
0
20
40
60
80
100
1
2
3
4
5
Rating

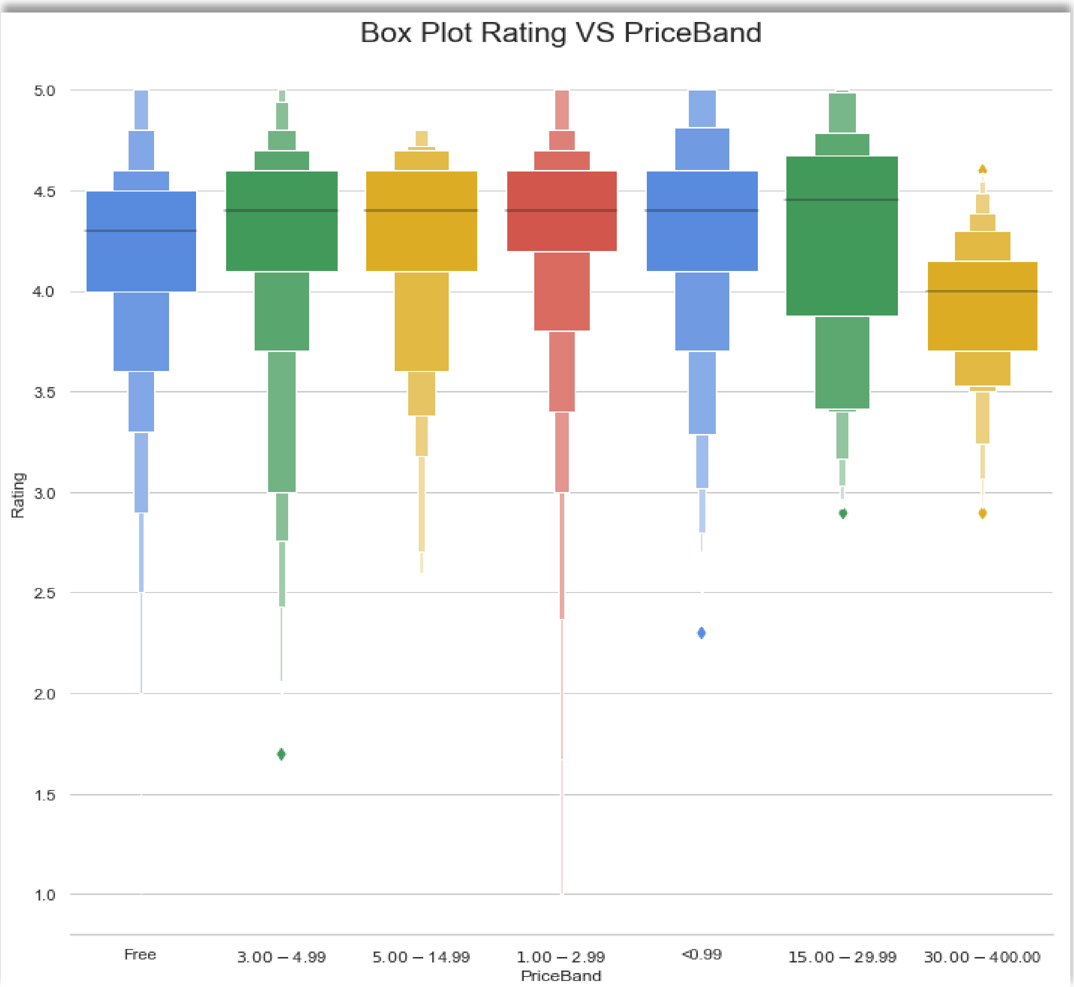
3.3. Pricing Strategy

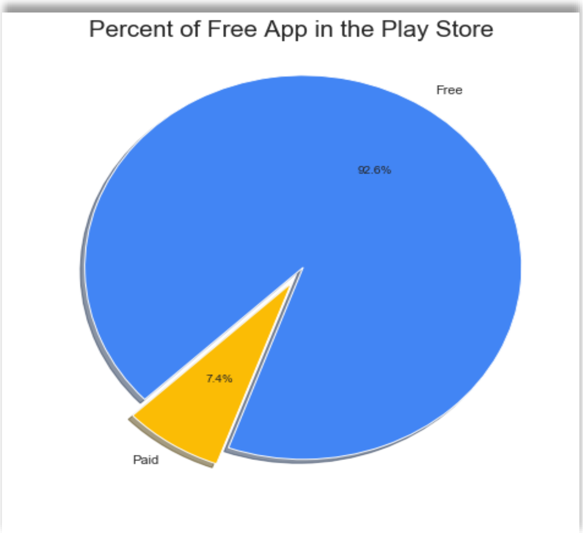




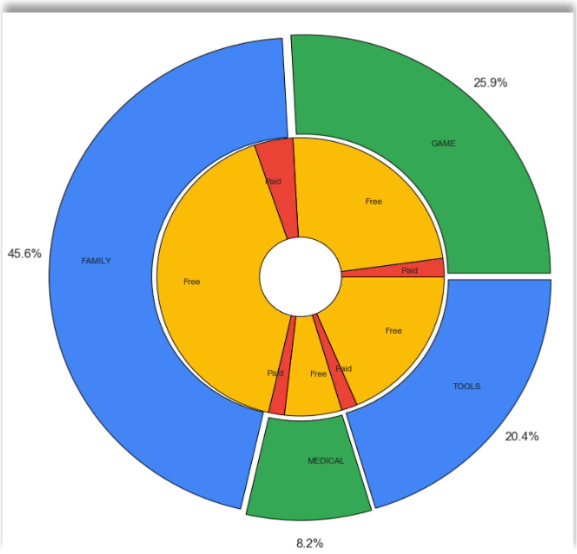


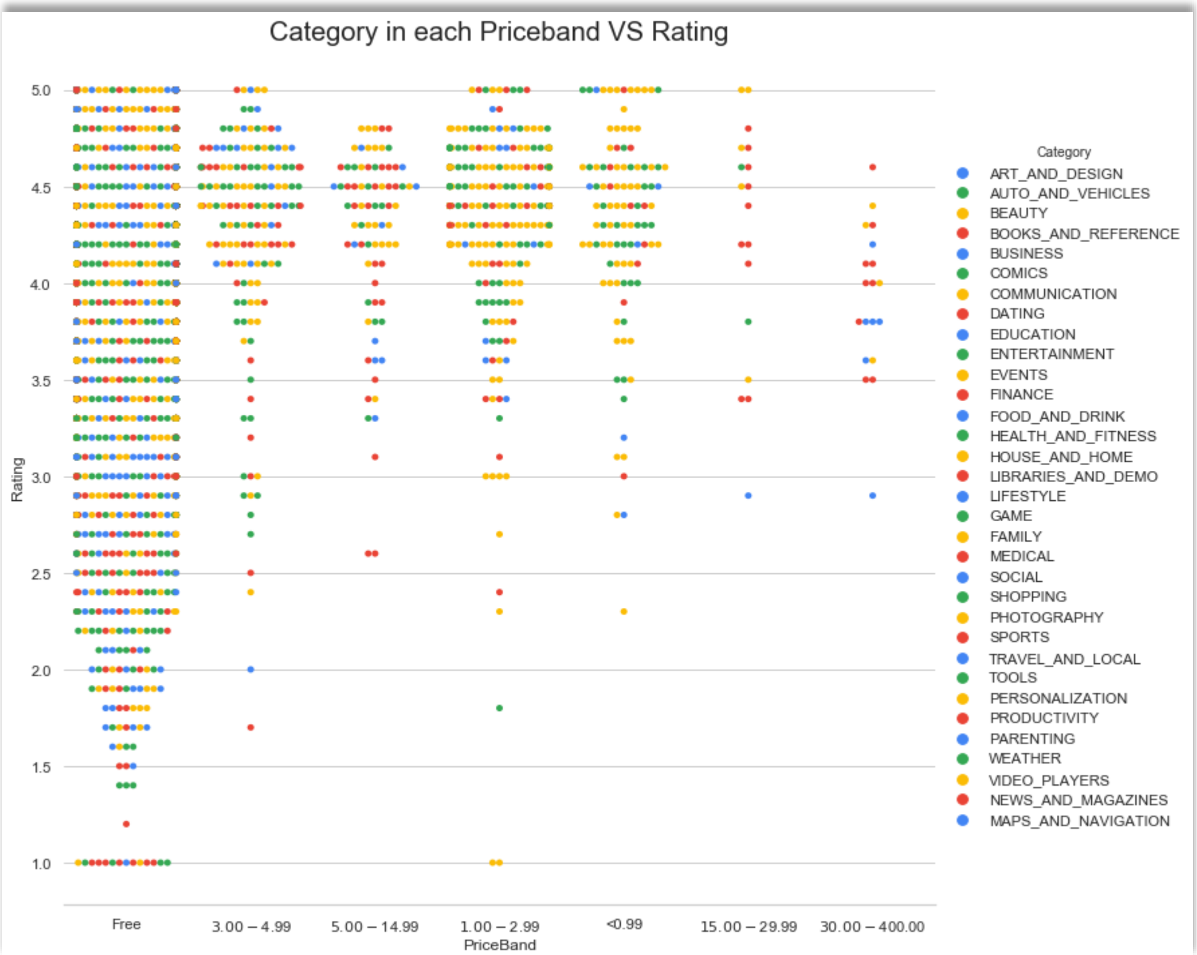

App Price Trends by Categories - Junk Apps Filtered Out
Category
BUSINESS
FINANCE
LIFESTYLE
GAME
FAMILY
MEDICAL
PHOTOGRAPHY
TOOLS
0
5
10
15
20
25
30
35
40
Price


Box Plot Rating VS PriceBand
Rating
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
Free
3.00 − 4.99
5.00 − 14.99
1.00 − 2.99
<0.99
15.00 − 29.99
30.00 − 400.00
PriceBand



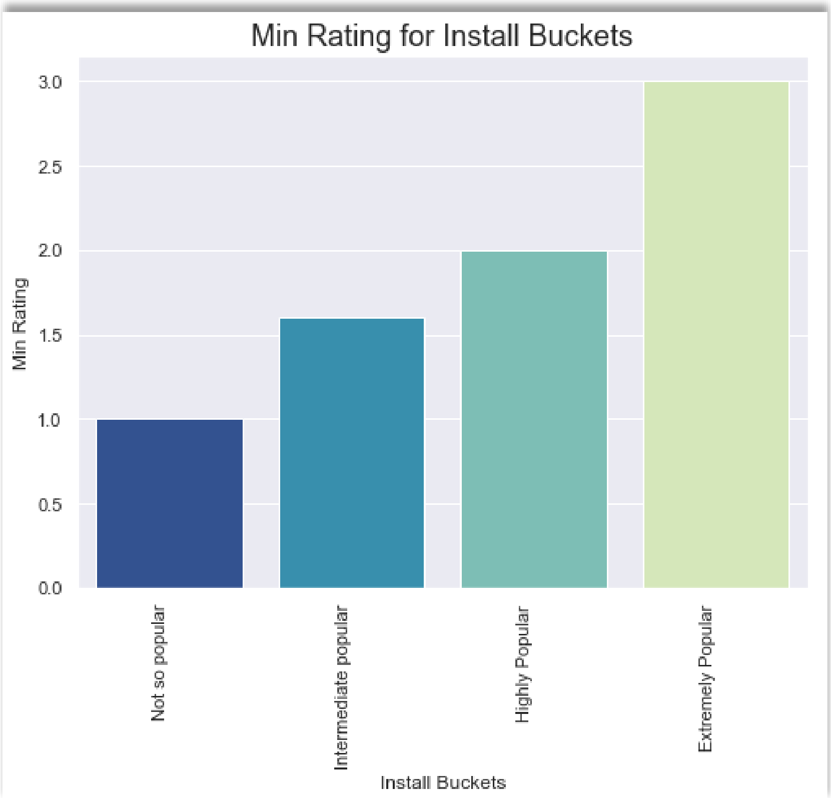
Percent of Free App in the Play Store



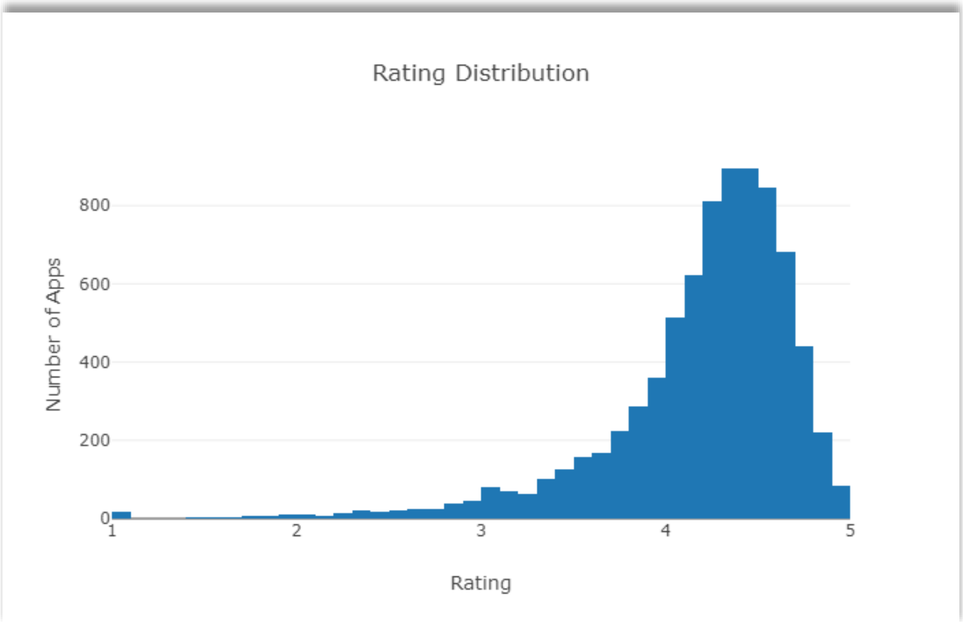


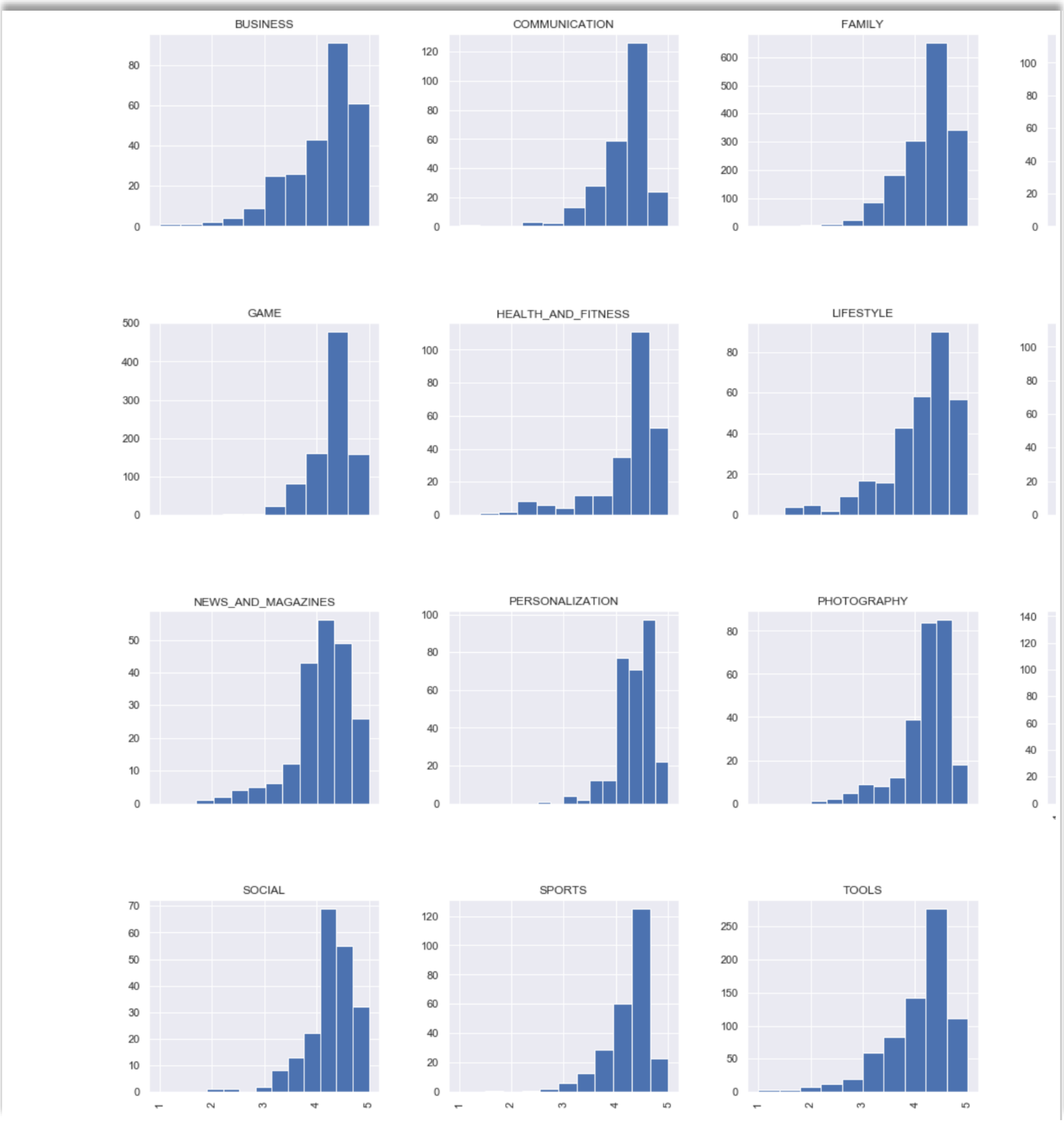
Category in each Priceband VS Rating

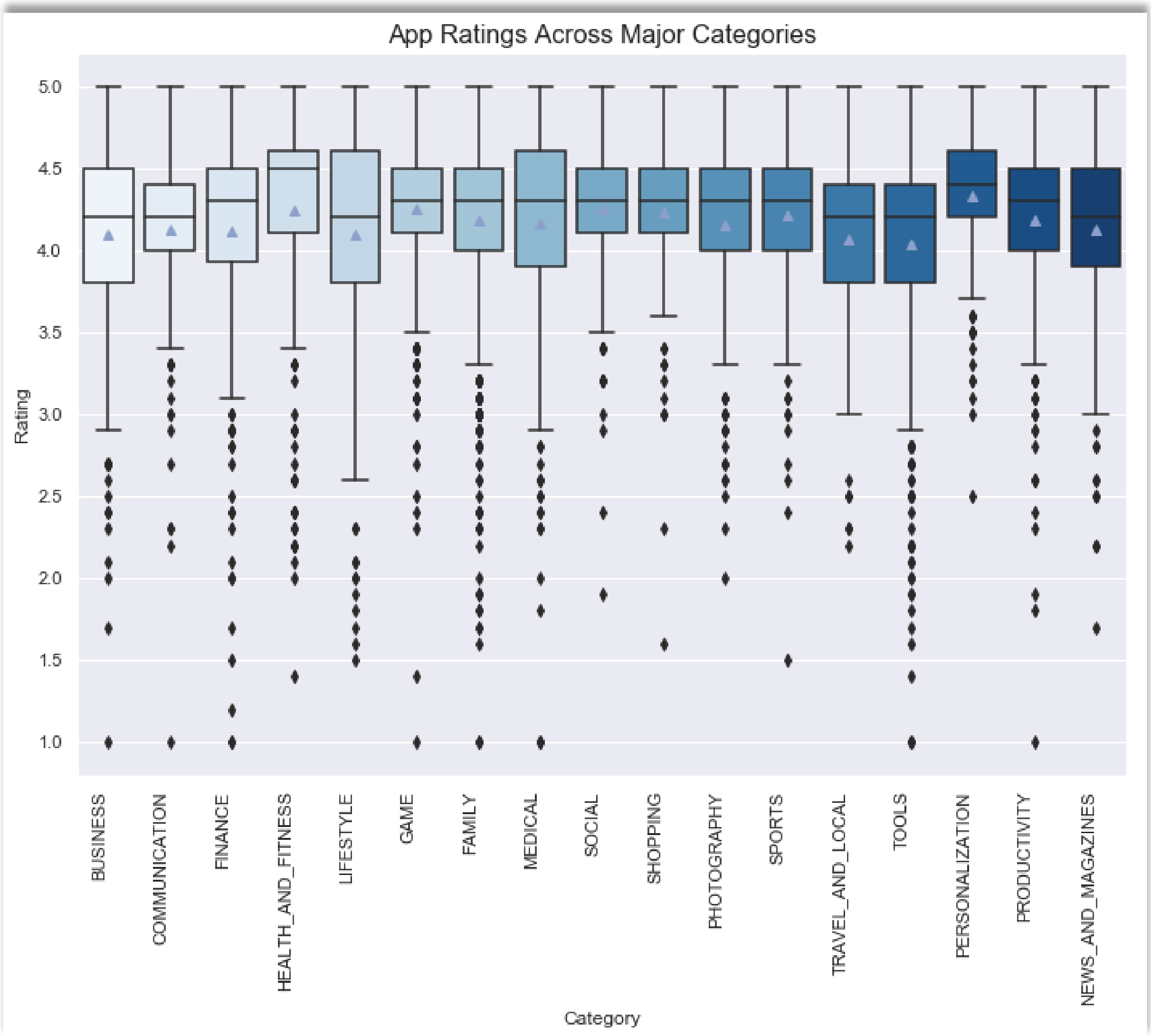
3.4. Install Count and Rating



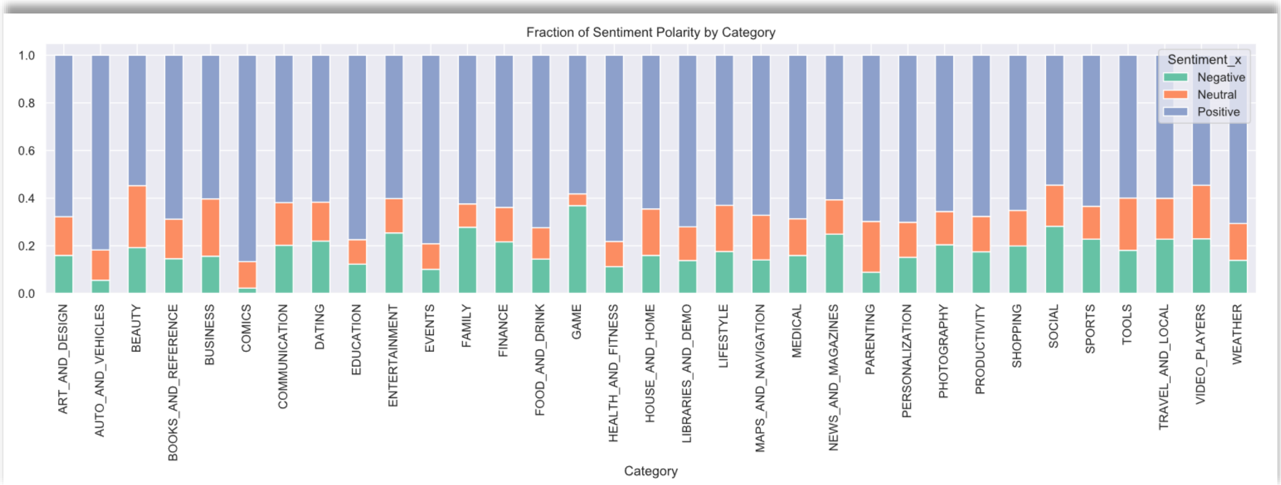
3.5. Rating Distribution

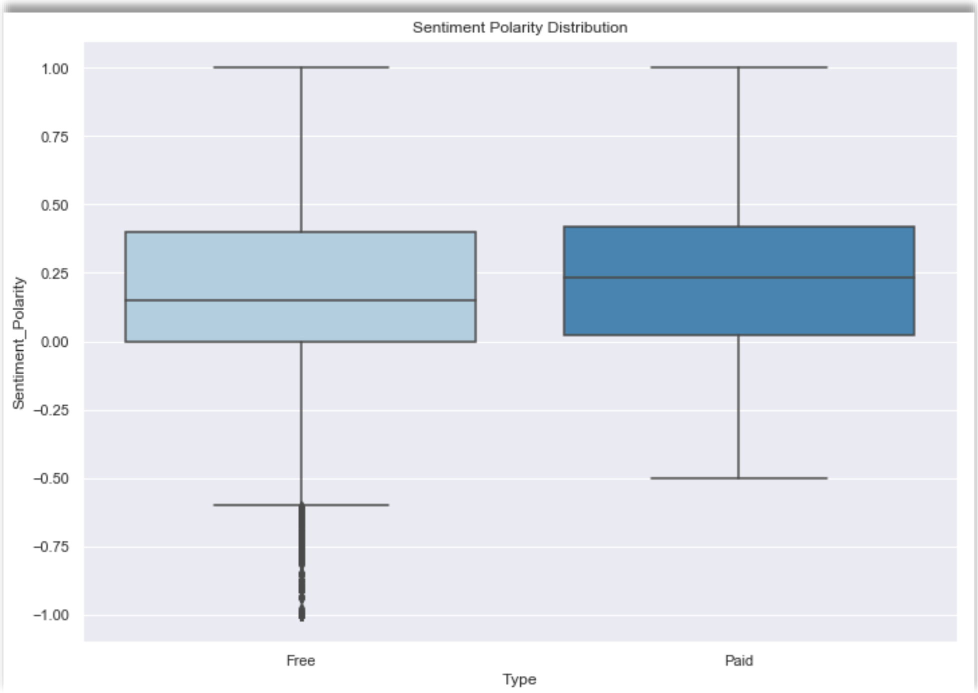



BUSINESS
0
20
40
60
80
COMMUNICATION
0
20
40
60
80
100
120
FAMILY
0
100
200
300
400
500
600
0
20
40
60
80
100
GAME
0
100
200
300
400
500
HEALTH_AND_FITNESS
0
20
40
60
80
100
LIFESTYLE
0
20
40
60
80
0
20
40
60
80
100
NEWS_AND_MAGAZINES
0
10
20
30
40
50
PERSONALIZATION
0
20
40
60
80
100
PHOTOGRAPHY
0
20
40
60
80
0
20
40
60
80
100
120
140
SOCIAL
0
10
20
30
40
50
60
70
1
2
3
4
5
SPORTS
0
20
40
60
80
100
120
1
2
3
4
5
TOOLS
0
50
100
150
200
250
1
2
3
4
5



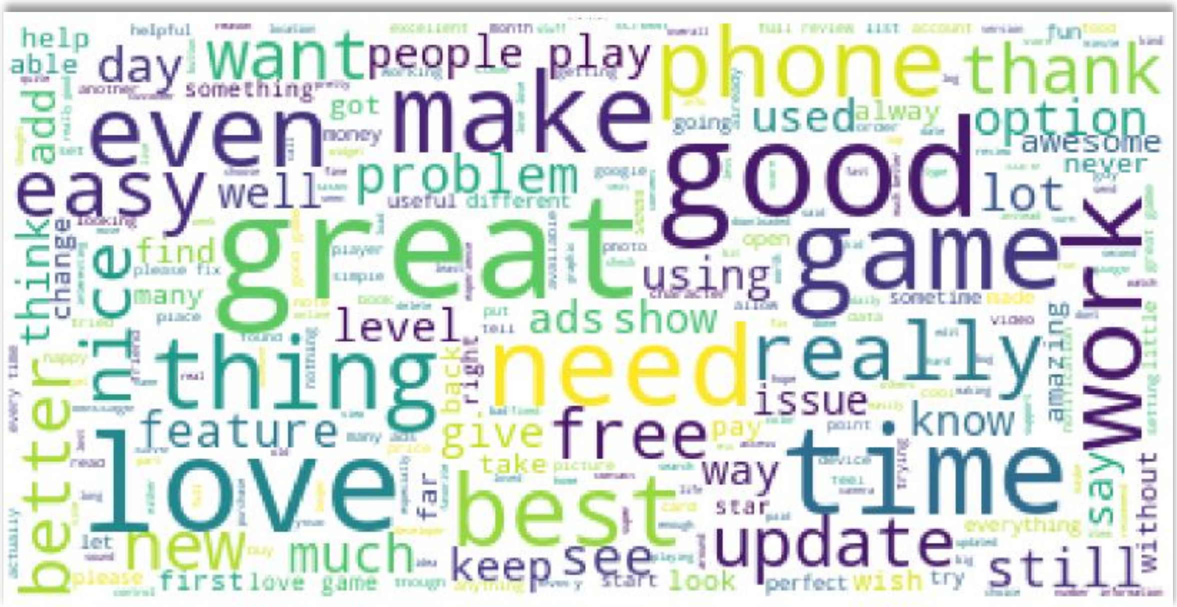
3.6. Basic Sentiment Analysis – User Reviews



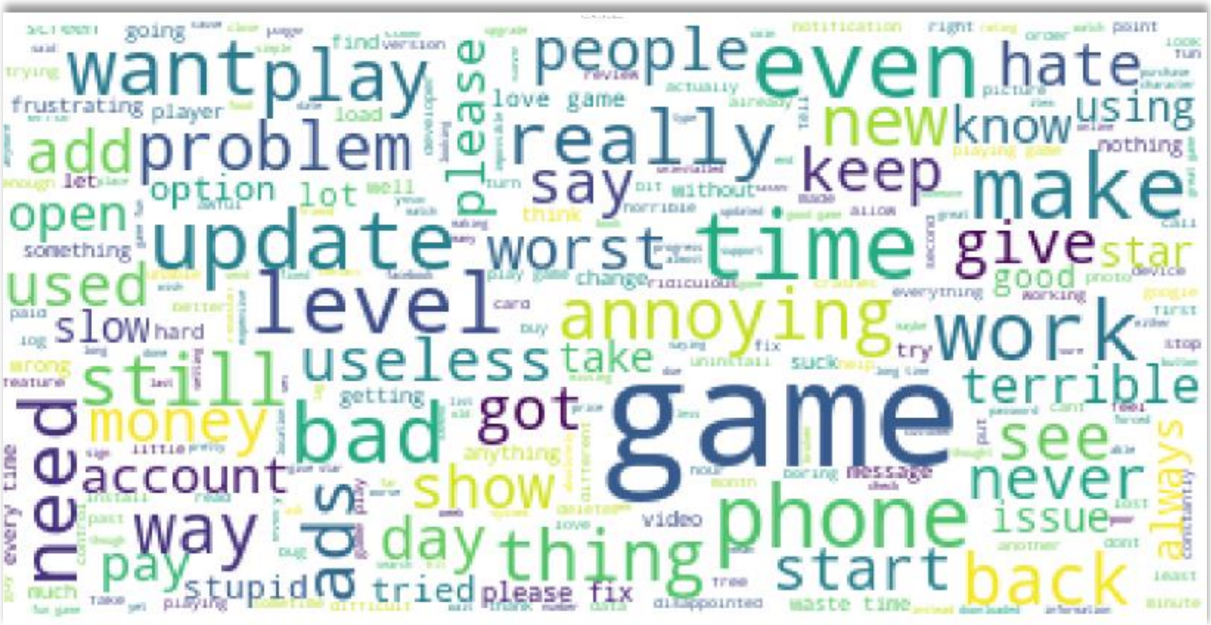


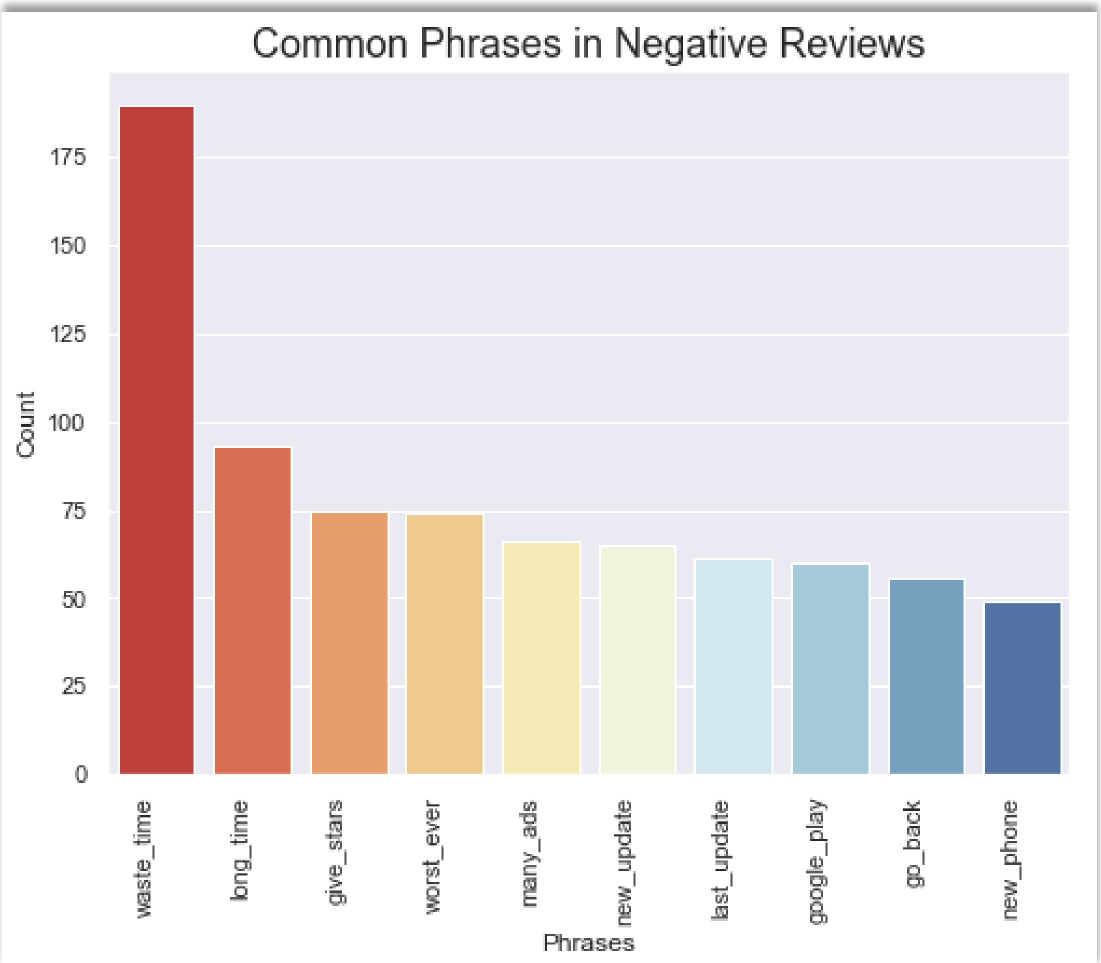
3.7. Frequency of words in reviews

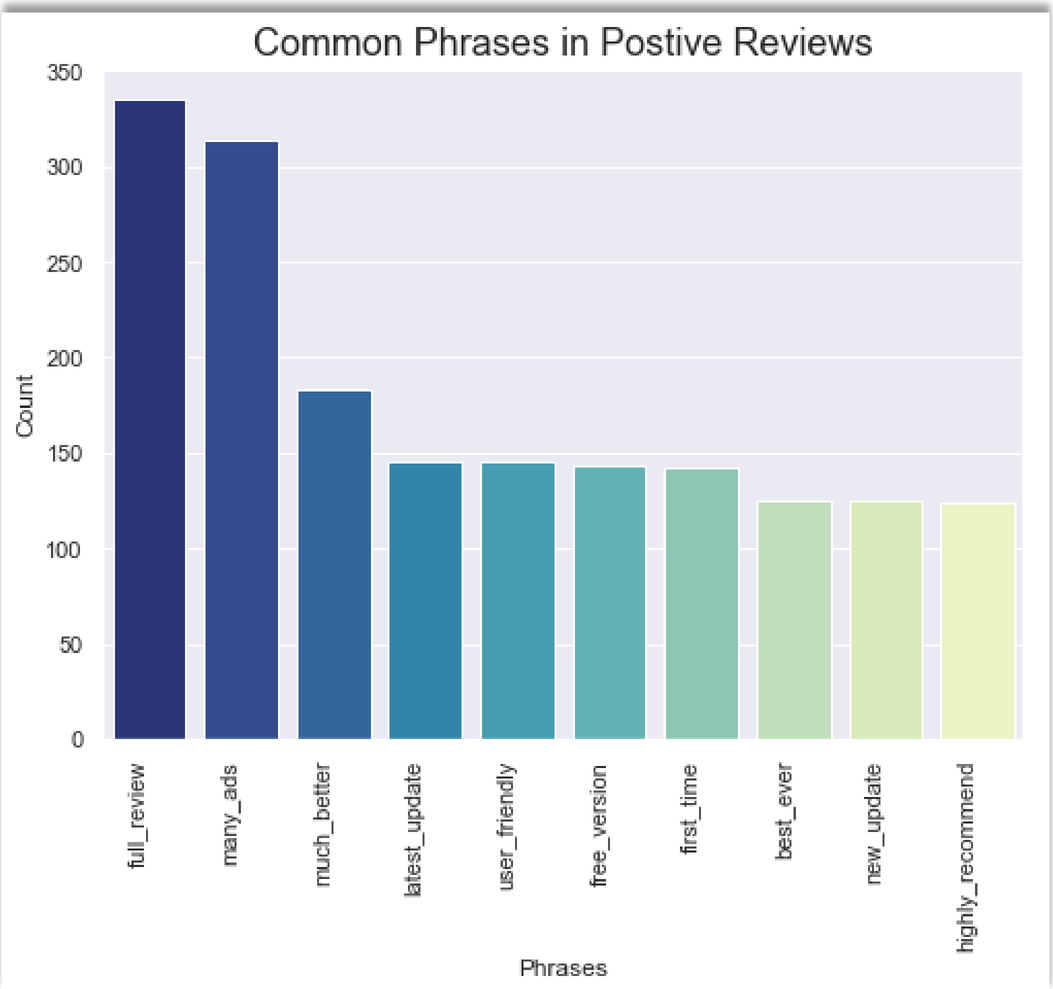
Positive Word Cloud



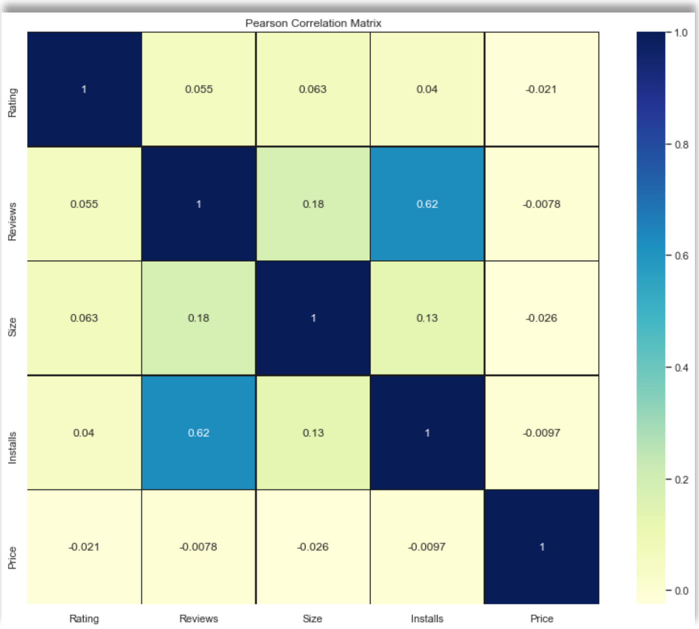
Negative Word Cloud





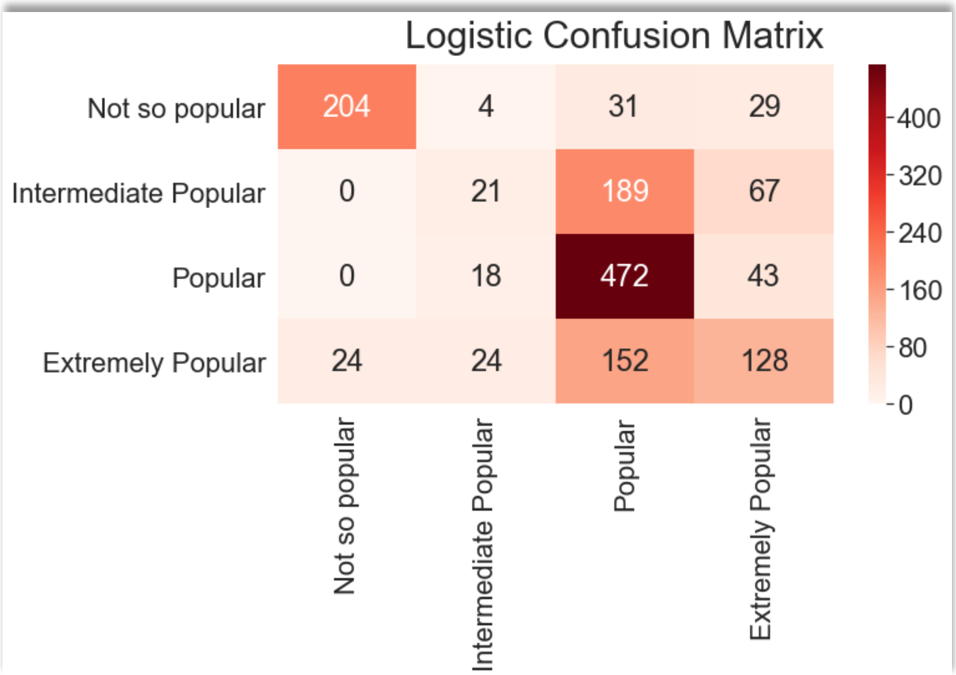


4.1. Feature Selection using Correlation Matrix

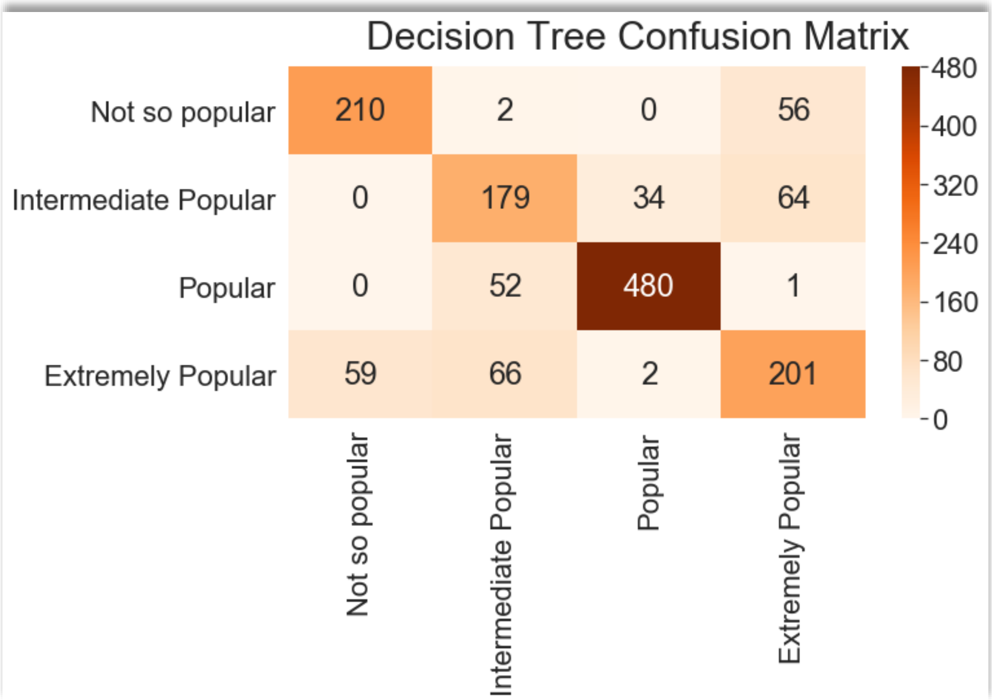


4.2. Implementation

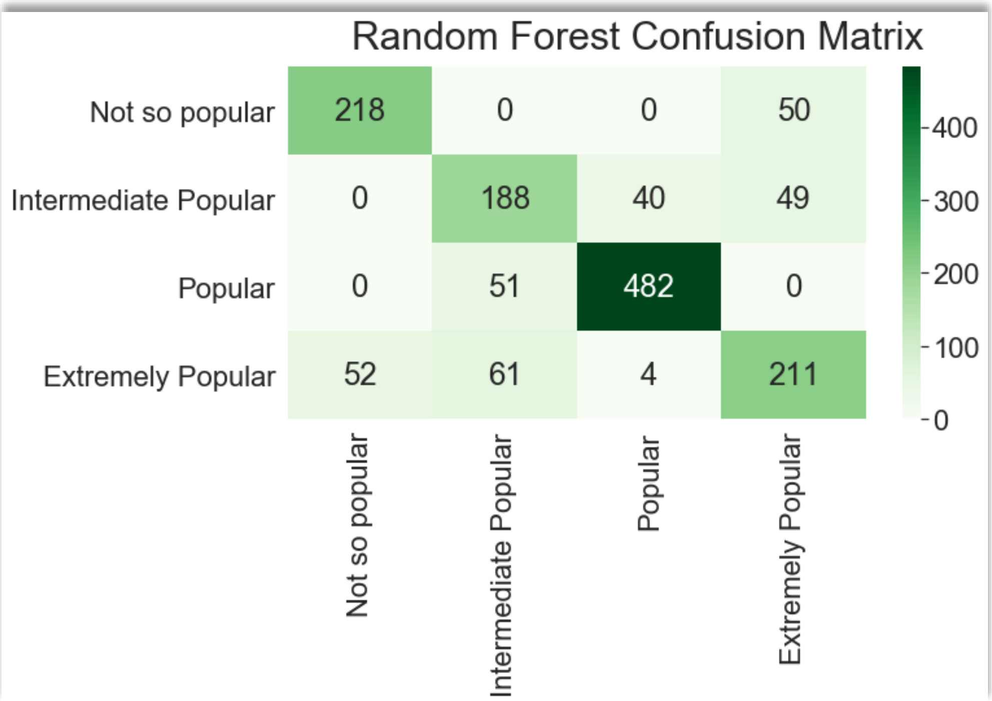
Logistic Regression Confusion Matrix



Decision Tree Confusion Matrix



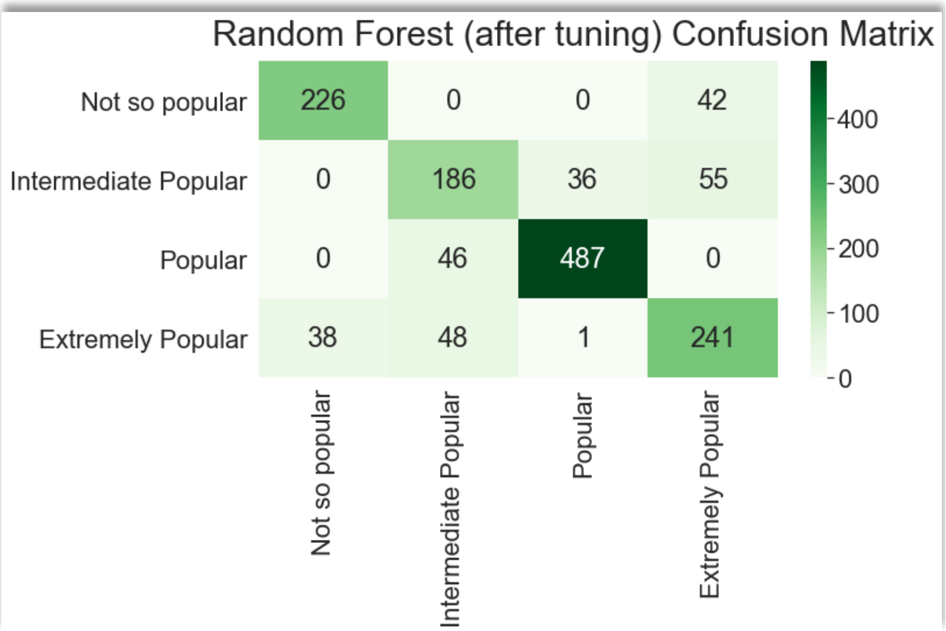
Random Forest Confusion Matrix



Random Forest Confusion Matrix

	Not so popular	Intermediate Popular	Popular	Extremely Popular
Not so popular	218	0	0	50
Intermediate Popular	0	188	40	49
Popular	0	51	482	0
Extremely Popular	52	61	4	211

Random Forest Confusion Matrix after Hypertuning



Random Forest (after tuning) Confusion Matrix

	Not so popular	Intermediate Popular	Popular	Extremely Popular
Not so popular	226	0	0	42
Intermediate Popular	0	186	36	55
Popular	0	46	487	0
Extremely Popular	38	48	1	241