Prof. Tweneboah
CMPS-240
12 Dec 2020

Project Report: Group 1

The Executive Summary:

Millions of mobile apps are downloaded and uploaded on Google App Store platform every day. Various features of an app play a big role in determining whether the app will be successful or not. Using the dataset containing data extracted from the Google Play Store, this project analyses various features of apps and investigates correlations and patterns established from the analysis.

The dataset contains about 600k applications available in the Google Play Store. This dataset was published in December 2020 and was imported from Kaggle. The dataset provides information of apps with features such as ratings, number of installs, price, category, release date, last updated date, Ad supported, in app purchases (any additional fees excluding the download fee, may be the app prompts the user to subscribe for their pro version, etc) and so on.

Our role in this project has been to study the relations between various features of apps and deduct factors that renders applications popular or successful. We answer questions like, what is happening behind the scenes with various categories of applications, how do app companies make revenue, what are the relations of ads and in app purchases in making a website more popular?

Data Cleaning and Data Preparation:

The dataset consisted of more than 6k Null values in Ratings column and around 150 in Installs column. The rows containing null values were dropped from both Installs and Ratings columns. After that, the dataset's rating category was checked to see if it had any invalid ratings, such as less than or equal to zero and greater than 5. The datasets did have a large number of zero ratings. So, those data with rating zero were removed from the dataframe. Similarly, the dataset was analyzed to see if there was any data with 0 installs having ratings higher than 0. There were some apps which had 0 installs, but they had ratings higher than 0 and even 5. Thus, those data with 0 numbers of installs were removed for finding the correlation between ratings and installs.

For finding the correlation between content ratings and ratings, ratings with null values and 0s were removed. Similarly, the numbers of "Unrated" and "Adults only 18+" Content Ratings were negligible. Thus, those content ratings were first replaced with NaN and then dropped. Lastly, the dataset's Category portion was checked to see if it contained any faulty values. However, the Category section was completely fine as it did not have any such improper values.

Analysis: Paid Apps

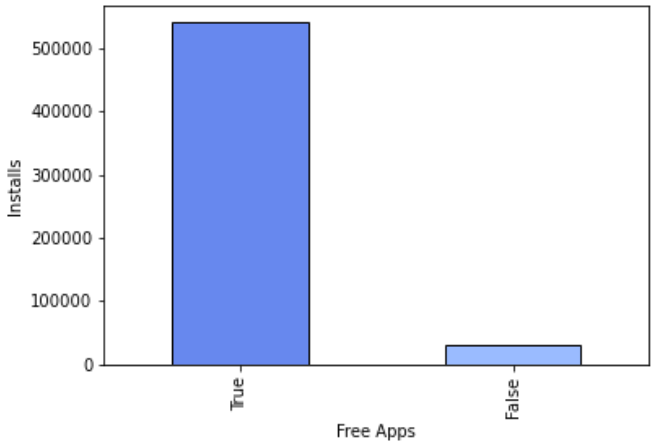


Figure 1.

The term “Google Play Store” seems a misnomer because most of the Apps available in the “Store” are free. Paid Apps only account for a tiny fraction of all apps in Google Play App Store, as seen from fig. 1. This is for a reason; Apps with high price do not correlate with high installs or ratings.

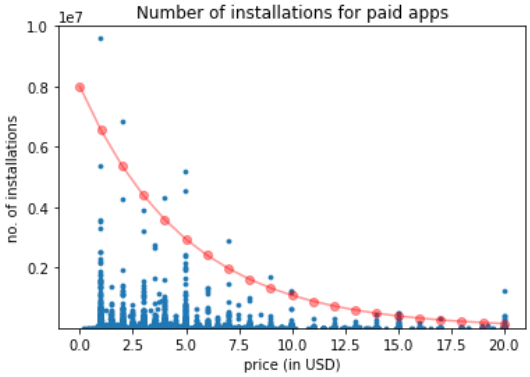


Figure 2.

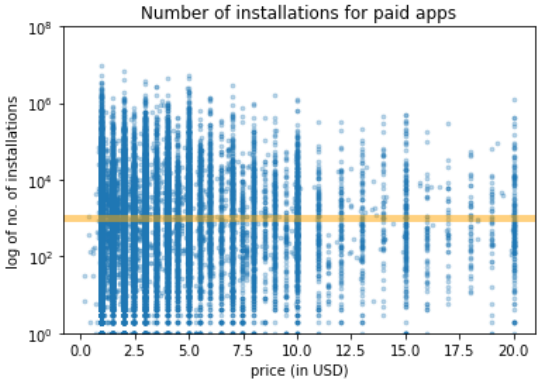


Figure 3.

In the figure 2, we observe that the number of installations of paid apps decreases dramatically with the price of the apps. The red curve is an exponential curve: $y(x) = A.\exp(-\alpha x)$. It approximately fits the data for apps with price less than 20 USD. Figure 3. is the same as the previous figure, but with a log scale. The uniformity of the figure in log scale supports the claim that the number of installations for paid apps has exponential loss with price.

Analysis: In app purchases and Ads

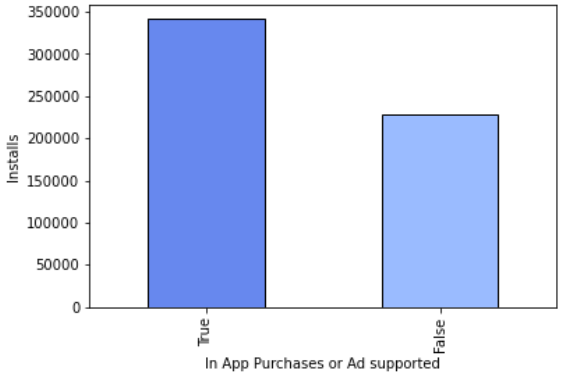


Figure 4.

If almost all Apps are free then how do Apps make money? From fig. 4, we see that In app purchases and Ad supported accounts for the majority of all Apps in Play Store. The heatmap provides more detail on this statistics; we find that the majority of the Apps are Ad Supported, without In App Purchases. This suggests that major contribution of revenues of App Companies come from Ads

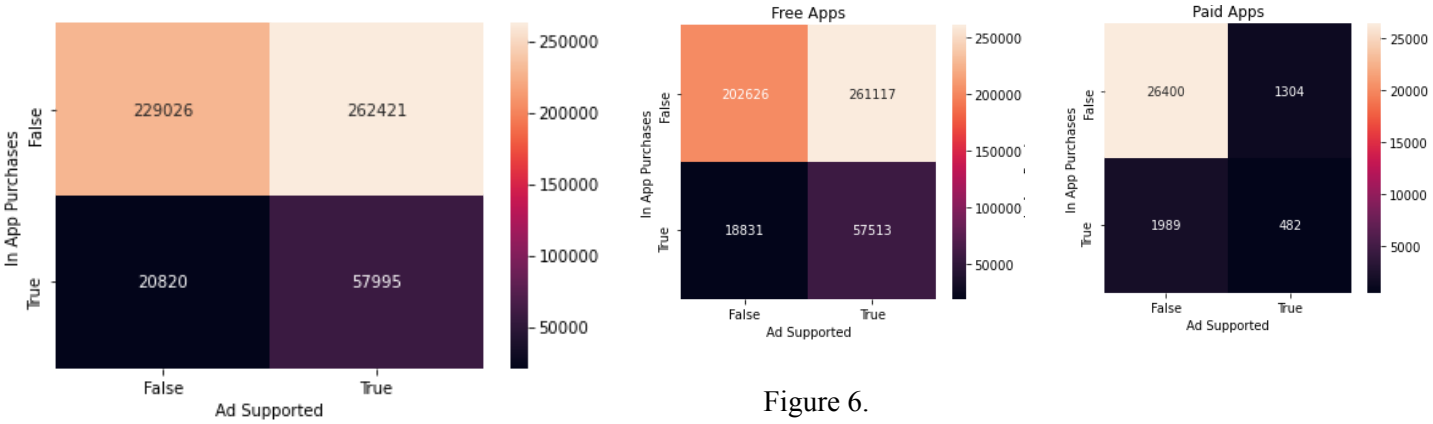
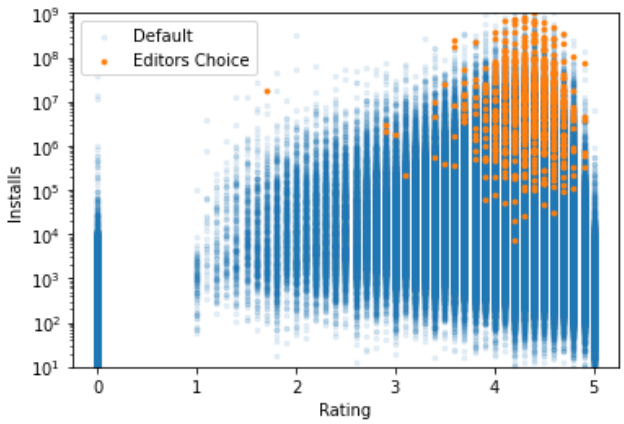


Figure 6.

Dissecting fig. 5 into paid and free apps, we find that even though Apps are labeled "Free" in the App Store, they contain Ads and In App Purchases. Majority of the Free Apps are Ad Supported; this means that App Companies can generate revenue even though their Apps are free of cost. Among the paid apps, only a fraction contains In App Purchases and supports Ad. This means that Paid Apps mainly generate their revenue through client's purchase

Analysis: Editors Choice



Editors Choice are Apps that are recommended by the Google Play Store editors. It represents only a tiny fraction of all Apps. The Editors Choice are supposedly the apps of best quality. The figure shows that Editors Choice Apps usually have high ratings and high installs

Figure 7.

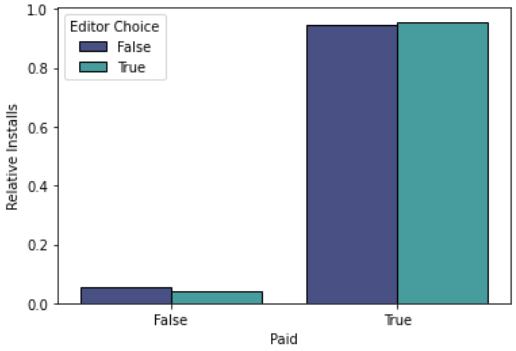


Figure 8.

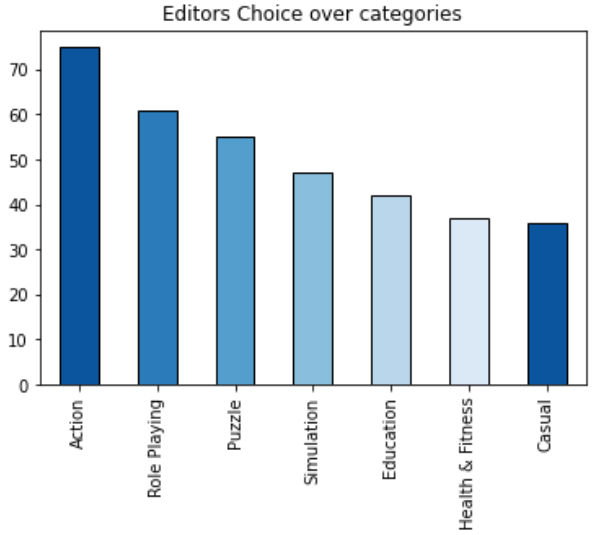


Figure 9.

The fig. 7 shows that editors choice directly corresponds to the usual demographic of the paid and free apps. However, on a close inspection, we find that Editors prefer free apps over paid apps. The preference of Editors on categories of Apps are shown in the figure on the right. It can be observed that Editors frequently choose games, then education apps and health & fitness apps. This provides some information on the demographics that the Editors in Google are trying to attract. It also suggests that mostly gamers follow the editors preferences

Analysis: Last Updated and Released date

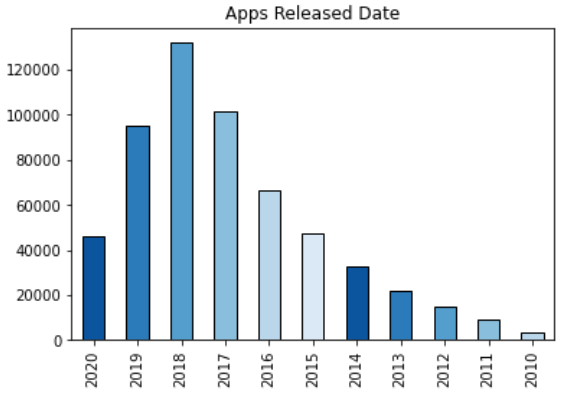


Figure 10.

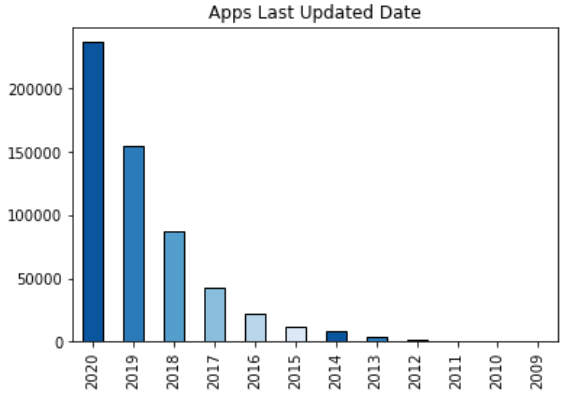
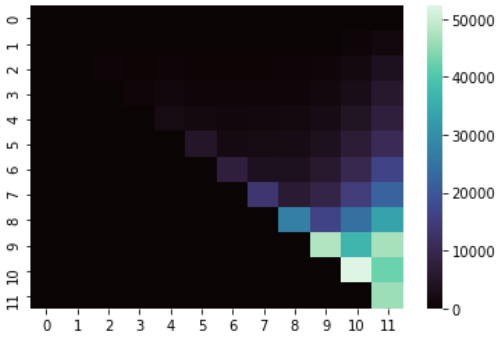


Figure 11.

Release of Apps in Google Apps Store peaks at year 2018, then decreases gradually. We observe that there are not a lot of apps released in 2020. This is because this data was scrapped in the middle of year 2020. The Last Updated figure shows that most apps were updated in 2020 which is obvious since Apps get updated continuously.



The figure on the left shows the correlation between Released and Last Updated dates. The numbers on the labels correspond to years after 2009. The dates gets more correlated as we get closer to 2020

Figure 12.

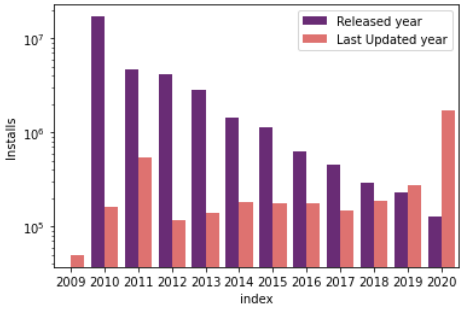


Figure 13.

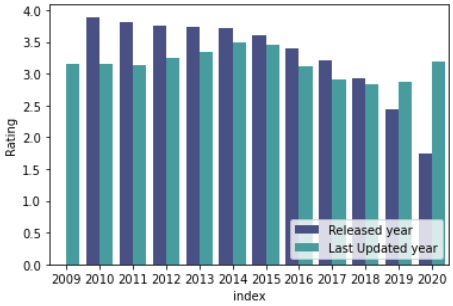
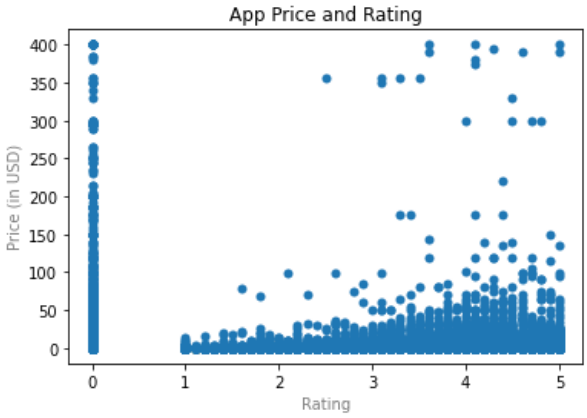


FIgure 14.

The figures above compares Released dates and Last Updated Dates with Installs and Ratings. Old Apps have high Installs; They have spent more time on the App Store market. And, Apps with latest updates have high installs; This is because regularly updated apps are relevant in the market We see a downward trajectory of Ratings with release time. This tells us that users are less likely to rate 5 stars today than in the past

Analysis: App Price and Rating



Now, we attempt to see if there is a correlation between price and rating for apps. The data set was grouped by currency, but due to the vast majority of the currency being in USD (96%), all apps requiring another form of currency were filtered out.

Figure 15.

Figure 15 shows the rating and price of apps. For comparison, the $50 mark will be used because it is closer to the peak of the main cluster. The majority of highly rated apps tend to cost less than $50 USD. The majority of apps that cost more than $50 USD have a rating of 0 (70%). It can be concluded that apps that cost greater than $50 USD tend to have a lower rating than apps that cost less.

Analysis: Content Rating and App Size

For this part, we tried to find if the content rating of apps show a difference in average app size. The data set was grouped by content rating. The size of some apps are dependent on the device that they are on. These apps (3%) were filtered out to provide these figures. The total number of apps that were considered is 1,080,882.

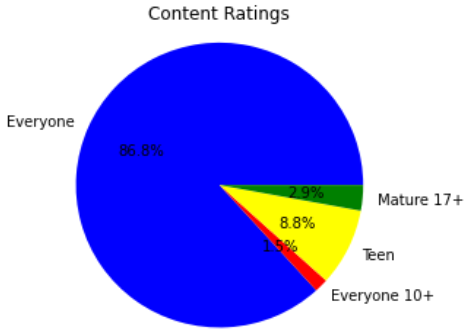


Figure 16.

Figure 16 shows the proportion of apps with different content ratings. The “Adults only 18+” and “Unrated” ratings were disregarded in this chart because their sections will not be visible. The chart reveals that the vast majority of apps are rated “Everyone” with 937,670 (86.75%) apps under this rating. The ratings with the fewest apps are “Adults only 18+”, with 52 apps and “Unrated” with 66 apps (0.01% combined).

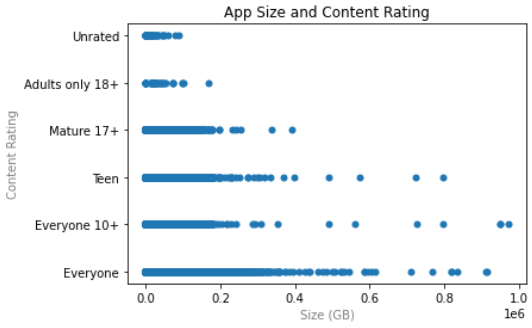


Figure 17.

Figure 17 shows the size of apps (labels in gigabytes) by content rating. The content rating with the most apps is “Everyone” with an average app size of 14,583.84 bytes. Despite being the content rating with the most apps, “Everyone” does not have the highest app size average. The content rating with the highest average app size is “Everyone 10+” with an average of 29,218.49 bytes. The content rating with the lowest app size average is “Unrated” with 11,992.25 bytes on average.

All other content ratings have an average app size between 11,000 and 28,000 bytes. There seems to be no definite correlation between content rating and app size, but the “Everyone 10+” and “Adult only 18+” ratings may have the highest averages due to being part of the top 3 smallest categories. “Unrated” may have the smallest average app size because they are apps that have not been properly categorized yet.

Analysis: In app purchases and Price

For this part, we will try to find out if the availability of in app purchases affect the price of apps. The data was broken into two groups: apps with in app purchases, and apps without in app purchases. For price, only prices in USD were considered, leaving the total number of apps considered at 1,075,136.

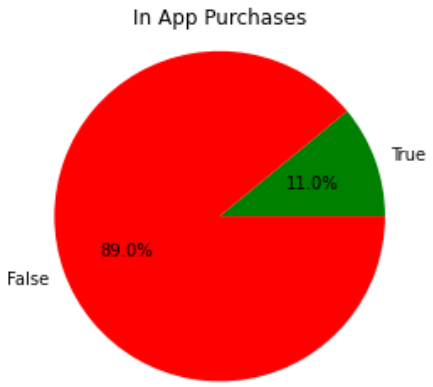


Figure 18.

Figure 18 shows the proportion of apps that have in-app purchases and apps that do not have them. The graph shows that the majority of apps do not have in-app purchases. There are 118,056 (10.98%) apps with in-app purchases, and 957,080 (89.02%) apps without them.

Figure 19.

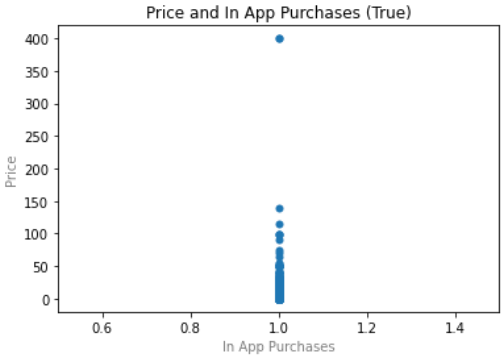


Figure 20.

Figure 19 shows the prices of apps without in app purchases. The majority of apps without in app purchases are $150 USD or less (956,969 apps). The average price for all apps without in app purchases is $0.22 USD. Figure 20 shows the prices of apps with in-app purchases. It is revealed that the majority of apps with in-app purchases cost less than $50 USD (118,043 apps), which is a clear distinction from apps without in-app purchases. The average price for all apps with in-app purchases is $0.11 USD, a lower average than apps without in app purchases. It can be concluded that apps with in-app purchases cost less than apps without them.

Analysis: Rating and Category

In this part, the way that the figures were produced is that first the dataset was grouped by “Category”. Then the average rating for each category was computed and then sorted. The goal here is to see which categories had higher average ratings and which ones had low ones.

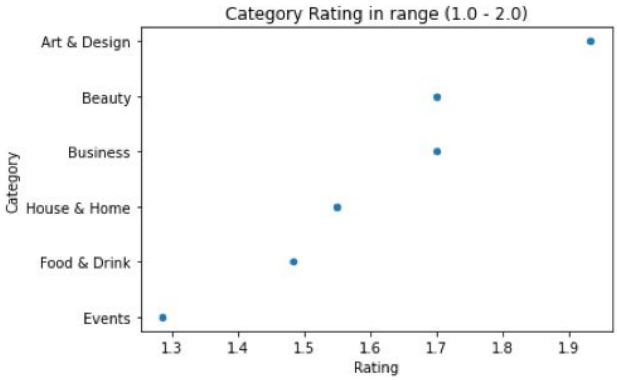


Figure 21: The category rating in range 1.0 – 2.0

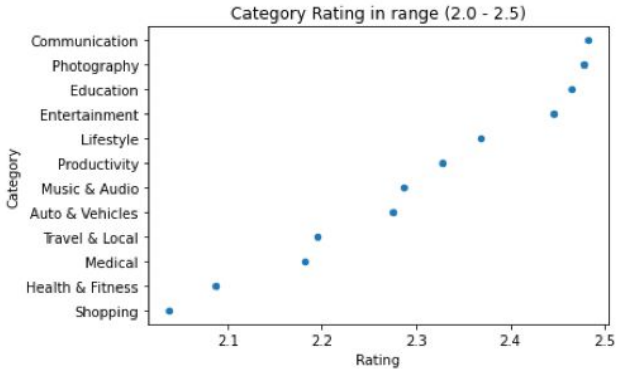


Figure 22: The category rating in range 2.0 – 2.5

Figure 21 for “Rating and Category” demonstrates the average rating that the apps have been given in each category. Note that only apps with an average rating of 1.0 – 2.0 are included in this specific figure. As demonstrated, the “Events” category has the least average rating among the other categories that are present in the dataset. Figure 22 for “Rating and Category” demonstrates the same type of information as figure 21 but applies the information to apps with an average rating of 2.0 – 2.5. Communication, Photography, and Education seem to be the top categories within this range.

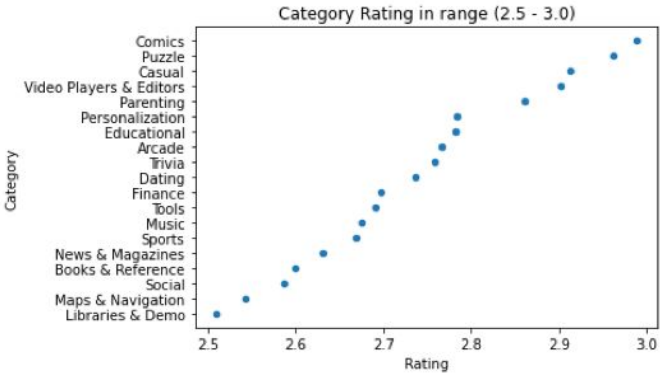


Figure 23: The category rating in range 2.5 – 3.0

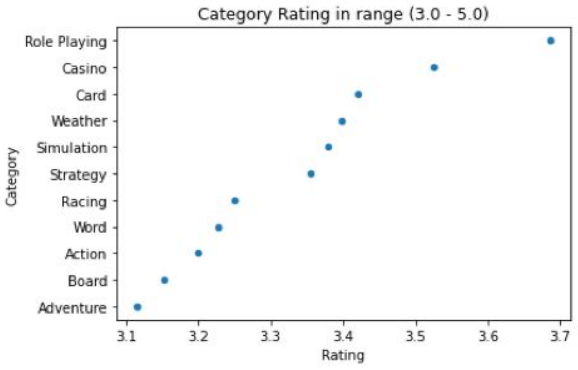


Figure 24: The category rating in range 3.0 – 5.0

Figure 23 applies the same information to apps with an average rating of 2.5 – 3.0. What is significant to note here is the number of the categories in this figure. Based on the figure, we

can conclude that the average rating for most of the categories falls within this range. Figure 24 applies the same information to apps with an average rating of 3.0 or higher. From this figure, we can conclude that Role Playing apps have had the highest rating among other apps on average. Casino and Card apps follow after this category.

Analysis: Ad Supported and Free

In this part, we attempted to see if there is any type of connection between the number of applications that support apps and whether or not those apps are free. We grouped the dataset by the "Ad supported" column, we then applied the .count() function to evaluate the number of free apps in the applications that support ads and the ones that do not.

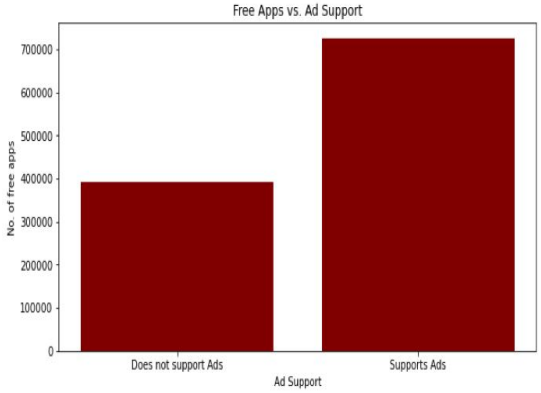


Figure 25

As demonstrated in this figure, apps that support ads tend to have a higher number of free applications. In other words, when an app is free, it is more likely to support ads. And this makes sense, because since the app developers are not making any money directly from customers through application installations, they could be receiving money from the advertisement that they publish in the application.

Analysis: In App Purchases and Category

In this part we attempted to see which categories support the "in App Purchases" feature more than the others. The dataset was initially grouped by "Category". Then for each category, the number of apps that support "in App Purchases" was computed.

Figure 26 demonstrates the number of apps in each category that support the "In App Purchases" feature in the range of 0 – 5,000. This figure demonstrates that the applications that are part of the "Parenting" category support the "In App Purchases" feature the least. And if we think about it, this makes sense because usually parenting apps simply provide a single functionality to the parents to control their children and there are not many associated features to purchase. Figure 27 demonstrates the same information as figure 26 but within the range of 5,000 – 10,000. As demonstrated, not many categories are included within this range.

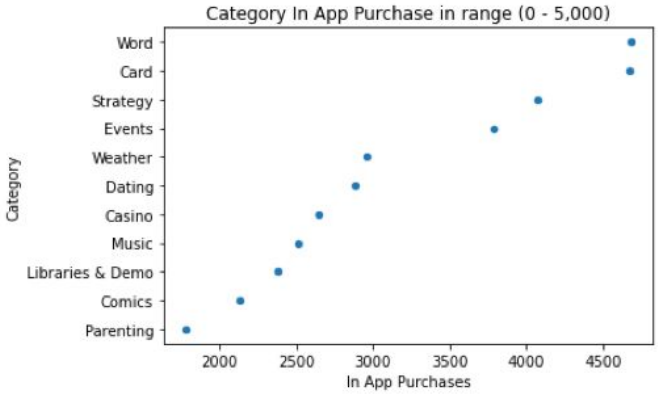


Figure 26.

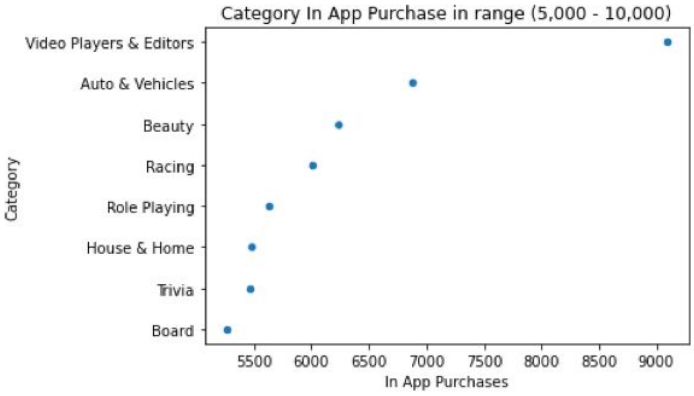


Figure 27.

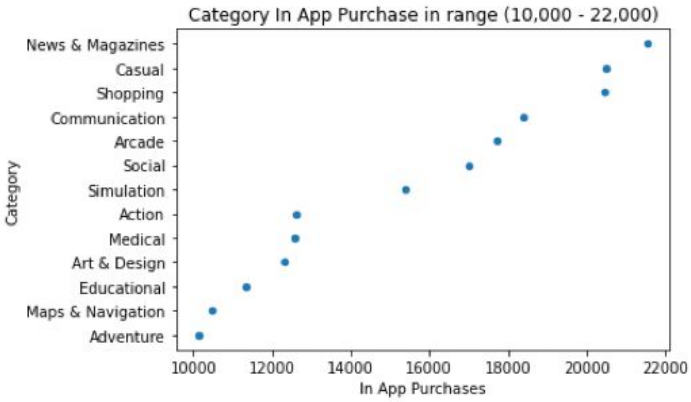


Figure 28.

Figure 29.

Figure 28 demonstrates the same information as the previous figures but within the range of 10,000 – 22,000. Interestingly, the “Action” and “Medical” category have about the same number of apps that support “In App Purchases”. Figure 29 demonstrates the information within the range of 22,000 and above. As shown, most of the apps that support “In App Purchases” are part of the Education, Music & Audio, and Entertainment category. This makes sense as Education applications would sell books and other tools inside their applications. Music & Audio applications would be selling songs and recordings, and the Entertainment applications would be selling different types of tokens (for example for games).

Analysis: Category vs Number of Installs

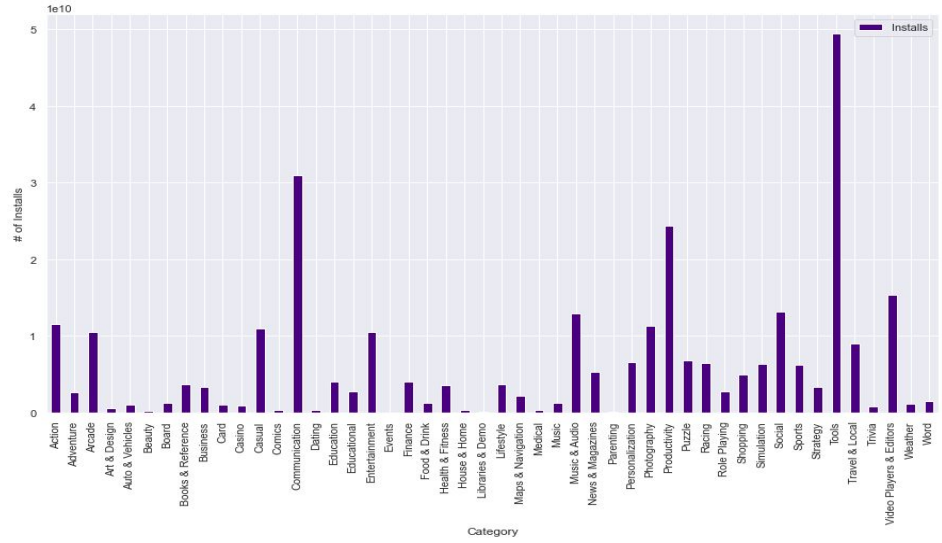


Figure 30.

Figure 30 demonstrates the relation between the application categories and the number of installs per category. So from the bar graph above we can see that the category Tools have the highest number of installs. And we can see that it leads other app categories on the number of installations with a large difference. As we saw in Figure 30 Education category has the highest number of apps but, the number of installs for educational applications is not as high as for other categories like Communication, Tools, Video Players & Editors, entertainment and so on.

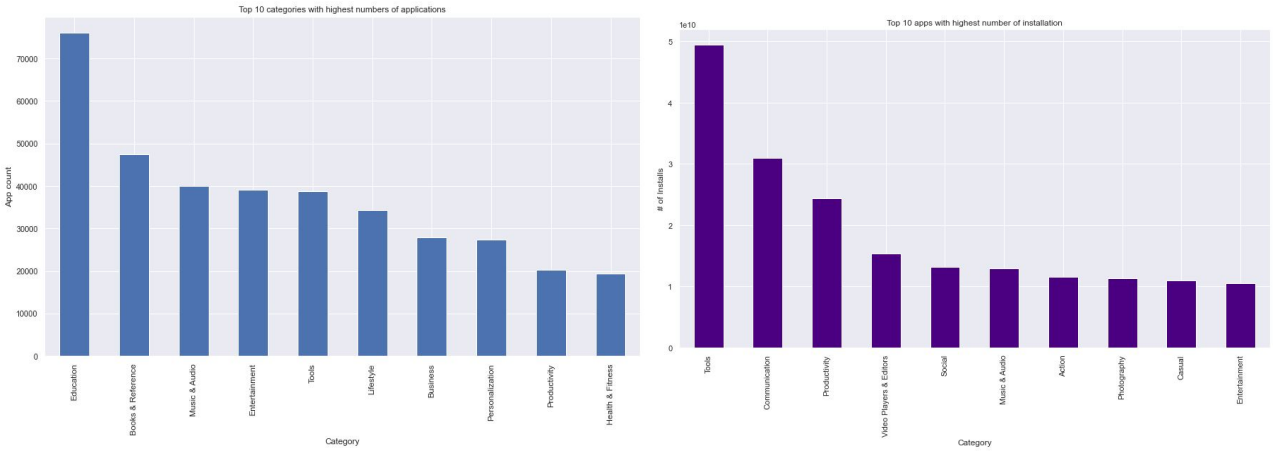


Figure 31. Comparison between Application Categories with highest app count, and those with highest installation.

The bar graph on the left shows the top 10 categories with highest app count. So, we can tell that the category Education has the highest number of applications however, it does not even come under the top 10 categories of applications to have highest installations. Whereas, the Category Tools, that has the highest number of installations only ranks fifth on highest number

of apps. Therefore, we can conclude that individual apps in category tools have a higher installation rate than apps in the education category and the same conclusion can be drawn for other categories by looking at the bar chart.

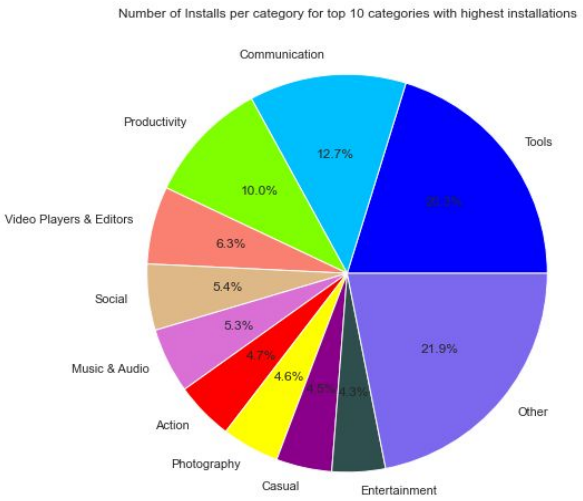


Figure 32: Pie Chart representing top 10 application categories with highest number of installations.

Analysis: Price Vs Category

We tried to find the correlation between the categories of application and their prices. Generally, which category of applications are priced and which of them are free.

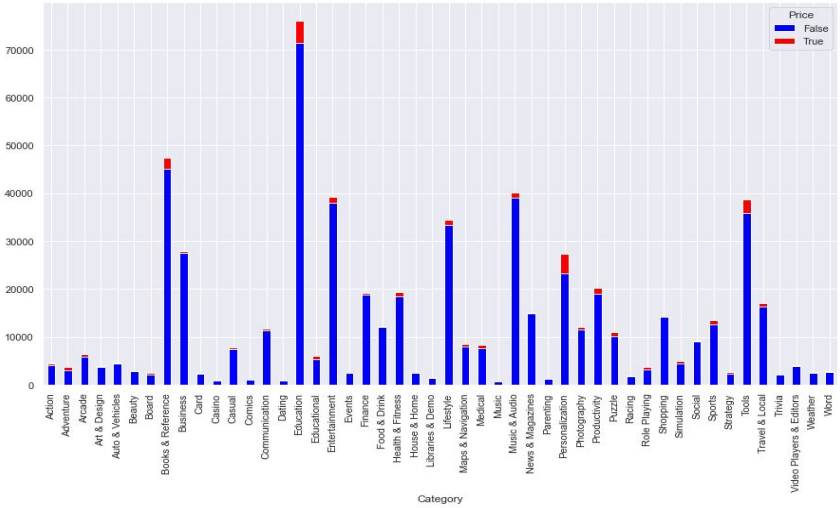


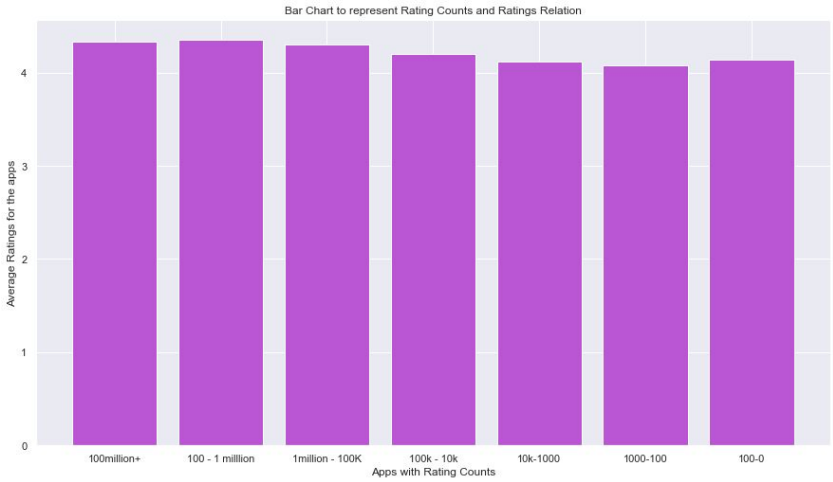
Figure 33.

The above stacked bar graph is representing the applications that are free and priced in the various categories. The color blue represents the population of the apps that are free, and red represents the apps that are priced. From the bar graph above we can tell that the apps in the educational and personalization sectors are priced more than the apps in other sections.

Then comes tools and books that are leading in the number of priced apps which also makes sense, because we usually pay for things like books, videos, video editors, etc. But overall, in all categories the applications are mostly free rather than being priced.

Analysis: Rating vs rating count

How would ratings relate to rating counts? There were apps that had 100 million + rating counts and apps with just 5 rating count, so would there be any significant differences in the average rating among these apps with high and low range of rating counts?



Bar graph representation for apps with rating counts in a particular range vs their ratings. All the apps were grouped into 7 ranges (as shown on the bar graph) for the value of their rating counts. And these groups of apps were graphed to see the ratings relation between them.

Figure 34.

From the above chart we can see that there is no significant relationship between the number of rating counts and ratings. As we see in the above bar chart that the apps that have different ranges for rating counts have similar ratings. For instance, apps where just less than 100 people have given ratings and apps where 100 million+ people have given ratings have almost the same average rating. Therefore, we can conclude that the number of rating counts does not affect the rating for an application in this dataset.

Analysis: Installs vs Ratings

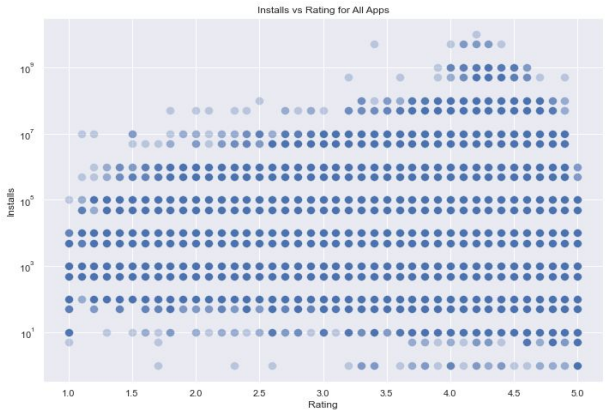


Figure 35: Installs vs Rating for all apps in general

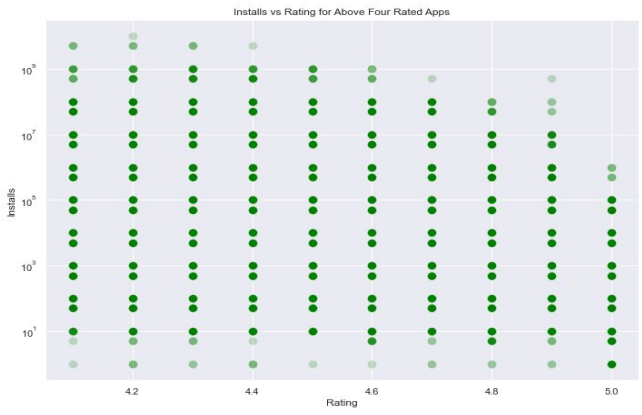


Figure 40: Installs vs Ratings for 4+ rated Apps

Since the Installs number had a drastic difference between the lower and upper value, y-axis was set to represent the log value of Installs. Similarly, apps with ratings zero were

dropped to establish the relationship between Installs and Ratings. In the above figure, we can see that rating alone does not have a conclusive impact on the number of installs although there are mostly the applications with ratings between 4 and 5 that have won the race of installation. While saying that, we cannot conclude the proportional relationship between ratings and install as the number of apps in the install range of 1K to 100M have evenly distributed ratings from 1 to 5. On the other hand, there are hardly any applications with ratings less than 4 that have been installed for more than 100 M. Therefore, we can conclude that the ratings can impact the tendency of crossing 100 M installation. However, it does not have a direct impact upon the general number of installs for applications since the highest rated applications do have less installs than the median rated applications.

Overall, the applications with median ratings between 4 and 5 had the highest numbers of installations. Also, as the number of installs increases, the possibility of retaining the rating of 5 decreases as more installs means more people rating the application, and one or more ratings lower than 5 certainly lowers the average rating for that application.

Analysis: Rating vs Content Rating

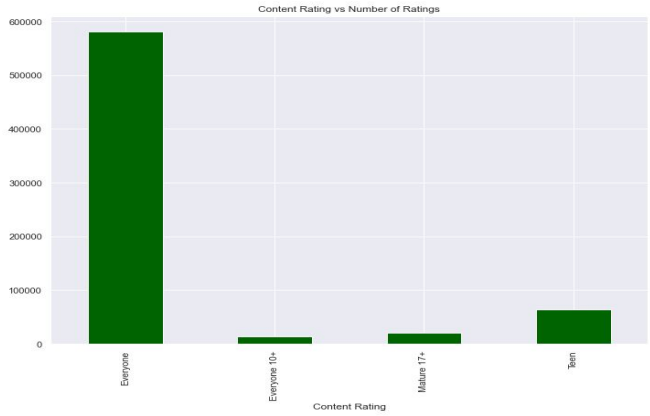


Figure 41: Content Ratings vs Number of Total Ratings

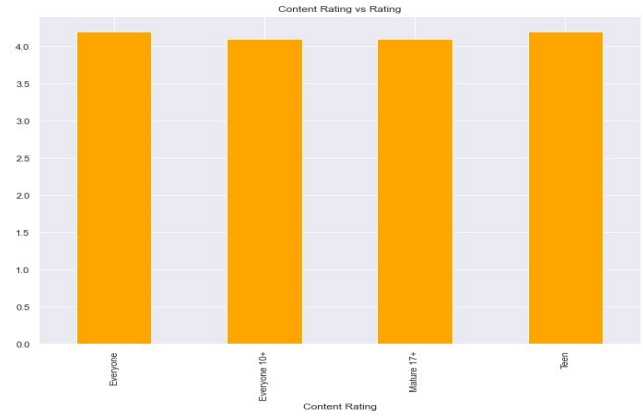


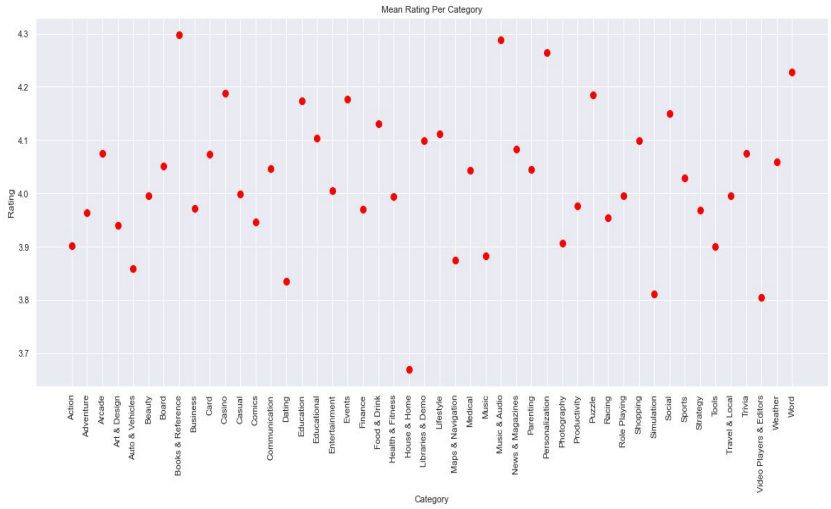
Figure 42: Content Rating Vs Rating

For the content rating, we dropped two categories which were “Unrated” and “Adults only 18+” as they had negligible numbers of ratings. There were very few applications under these two content ratings, so they were dropped. It is obvious for the application with content rating “Everyone” to have the maximum number of ratings as those are the applications installed by most of the users. Similarly, when the more people install applications, more people are likely to leave a rating for that application. In this sense, we can conclude that applications with content ratings “Everyone” are mostly have the highest numbers of ratings.

In figure 41, we can see that the ratings have a huge number of outliers when compared with the content rating. Thus, we used the median to find the correlation between content rating and rating as Median serves the best for our purpose of dealing with the outliers. We used groupby for content rating and median of rating as median helps to measure the central tendency

of ratings. After doing that, we can see that the content rating does not affect the rating as they don't have a distinct correlation. Overall, the rating is independent of the content rating.

Analysis: Category Vs Mean Rating Distribution



For finding the correlation between category and Mean Rating Distribution, we used groupby for category and mean rating. Overall, the dataset did not have strange categories for any application. Thus, there was no real need for data cleaning for the categories part. We took the mean values of rating for each category to find which categories had the highest and lowest numbers of ratings.

Figure 43. Category Vs Mean Rating Distribution

After that, we used seaborn's scatter plot to plot categories and mean rating distribution. We can see in the above figure the "Books and Reference" category had the highest mean rating followed by the "Music and Audio" and "Personalization". On the other hand, "Houses and Home", "Video Players and Editor", "Simulation" and "Dating" had lowest mean ratings.

Conclusion and Future Work:

Online mobile applications represent a large and still-growing market. There may be many factors that go into determining the popularity of an application. These factors may be of different types. Focusing mainly on those factors of an application that are external to the application itself, such as the app rating, price, and so on, this project studies the factors that classify the overall popularity and success of an app. It is found that some features of an application are more consequential in accurately determining the popularity of an app than others. However, almost every feature has some contribution in aiding in the investigation of the dynamics of the online market platform.

In the future, to improve the classification capabilities, additional features would be helpful. These features may include factors that are internal to the app (i.e., the software features of an app). Furthermore, the use of machine learning algorithms, such as Neural Networks, could result in classifications of apps according to their potentiality, so that developers of apps can optimize their apps for more success.