Google Play Store Dataset Analysis

By Anonymous Python: Indoo Park, Bryan Kim, Yunsueb Kim, Edward Lara

Abstract:

In this project, we aim to investigate what drives smartphone application success in the Google Play Store. We define "success" as an app that has many installs, but is also rated highly by users. To do so, we analyzed web scraped data of approximately 10,000 applications on the Google Play Store to explore the relationships between demographics that are provided on the Google Play Store. With this dataset we examined which variables were important in determining how "popular" and how "good" an app is by investigating how our independent variables influenced the number of 'Installs' and the app 'Rating'. To determine what variables influenced these dependent variables, we fit a multiple linear regression model and examined the significance levels of each variable. With this analysis, we confirmed some of our hypotheses in our EDA, where we anticipated that 'Last Updated' as well as certain factors in 'Category' were significant in influencing both 'Installs' and 'Ratings', and that 'Size' was an important feature in predicting 'Installs'. Our multiple linear regression models show that at the threshold of α=0.05, 'Last.Updated' and some factors in 'Category' were indeed significant in influencing both 'Installs' and 'Rating'. Also, our models show that that 'Size' of an app is in fact significant in determining the 'Installs'. However, some areas of our model output that surprised us was the fact that 'Content Rating' was insignificant in determining 'Rating' or 'Installs', and the fact that being paid or free had no influence in the number of 'Installs'. Overall, some of the limitations of our study include the fact that our dataset is a year and a half old, and that 'Installs' is measured in bins such as 1000+, 5000+ which would hinder the accuracy of our results. Our report concludes that a free application with a low file size under the category of 'communication' would be ideal for a client looking to create an app with the highest rating and highest amount of installation.

Problems to Be Solved:

As technology improves, the possible uses of smartphone applications increase alongside the possible profits from applications. The goal of this report is to provide insight into the current popular and highly rated applications that Google Play Store consumers have gravitated towards. As a possible application developer, this report will be useful in understanding the Google Play Store's current market and where there is an absence of a popular application that can be capitalized on.

Variables:

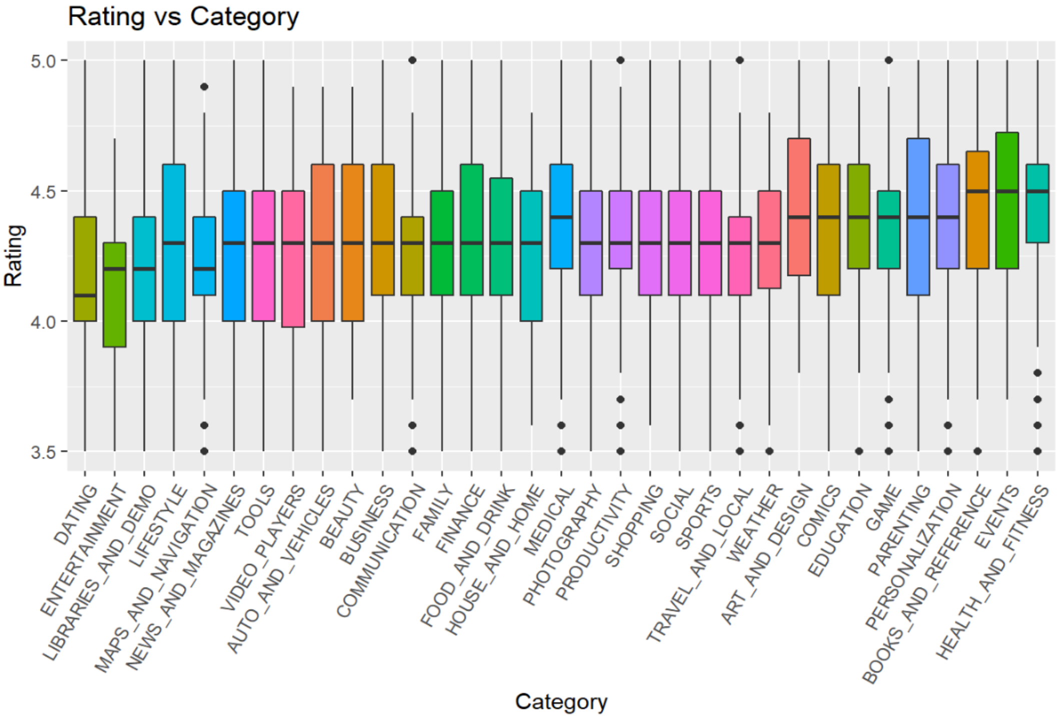
This data set was acquired from Kaggle, and the publisher web scraped the data directly from the Google Play Store. Thus, each variable measured was taken straight from the Google Play Store and is formatted as displayed when you enter an app's page on the store. Before data analysis could begin, data cleansing was necessary on the raw web-scraped data. The data cleansing process only required minor but necessary changes to the dataset. First, observations with incomplete columns (NA or NaN values) were removed from the dataset. From there, units in size of applications were standardized to be in megabytes (MB) instead of both megabytes and kilobytes. The data was scraped on September 4, 2018 so we created a new variable by finding the difference (in days) from the day each app was last updated and (09/04/2018). This shows how many days it had been since the last update upon data collection. The last step in our data cleansing process was the removal of unnecessary characters in our data. For example, the plus signs in the 'Installs' column, commas, and other unnecessary characters were removed. More information on the details of the variables and whether or not they are numerical or categorical is provided in the appendix.

Exploratory Data Analysis:

Before carrying out statistical analyses, we wanted to check the relationships between our variables against 'Ratings' and 'Installs' to guide us towards some hypotheses as to what variables were important.

Rating vs. App Category:

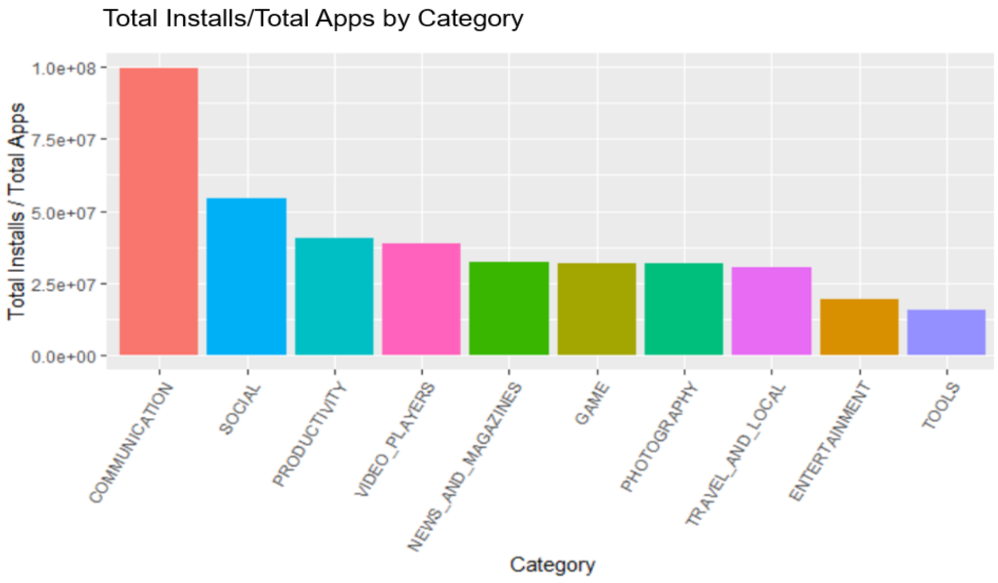
The following shows box plots of each application category against the average rating of that category. What is interesting to see for developers is the median rating but also the variance in each category. As a developer, categories with high median ratings and short variance whiskers would be the ideal category to join. Categories identified as books and reference, events, and health and fitness, respectively have the three highest median ratings. Health and fitness has short variance whiskers and highest rating which could signal a possible marketplace for developers to get involved in. Thus, with this graph we hypothesized that 'Category' would be significant in explaining 'Rating'.

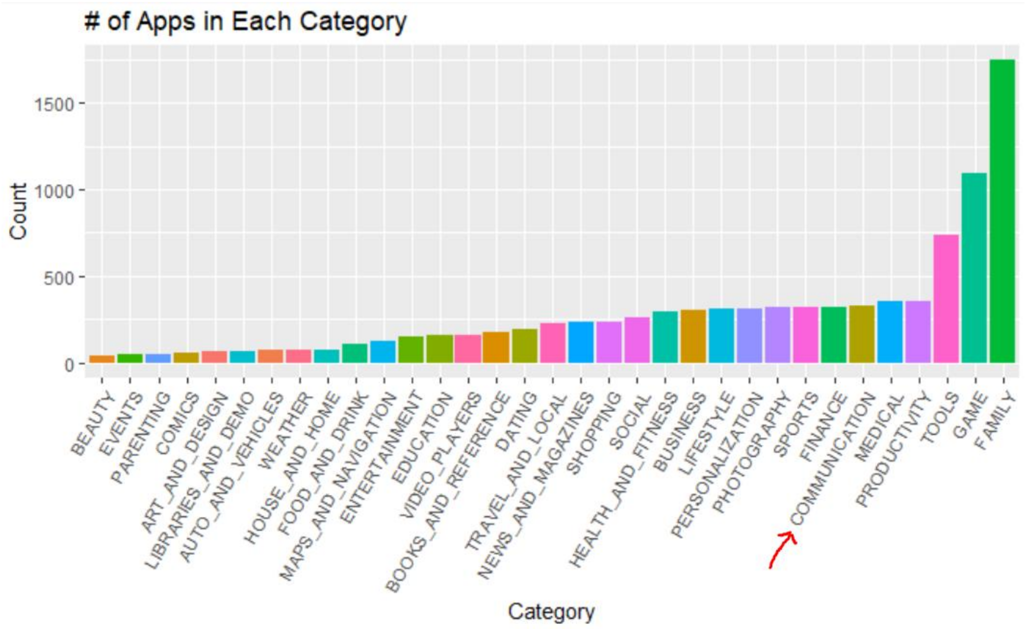


(Total Installs / Total Apps) vs App Category:

The plot below shows the average number of installs per app by category from highest to lowest. What is interesting in the plot below is how communication apps dominate as the most popular category. This includes applications such as Skype, Discord, Telegram, and WeChat. Thus this graph convinced us that the category of app matters in determining the number of 'Installs' and that communication apps would also be significant. It is important to note that the graph shown is the ratio of installs by total applications in the respective categories.

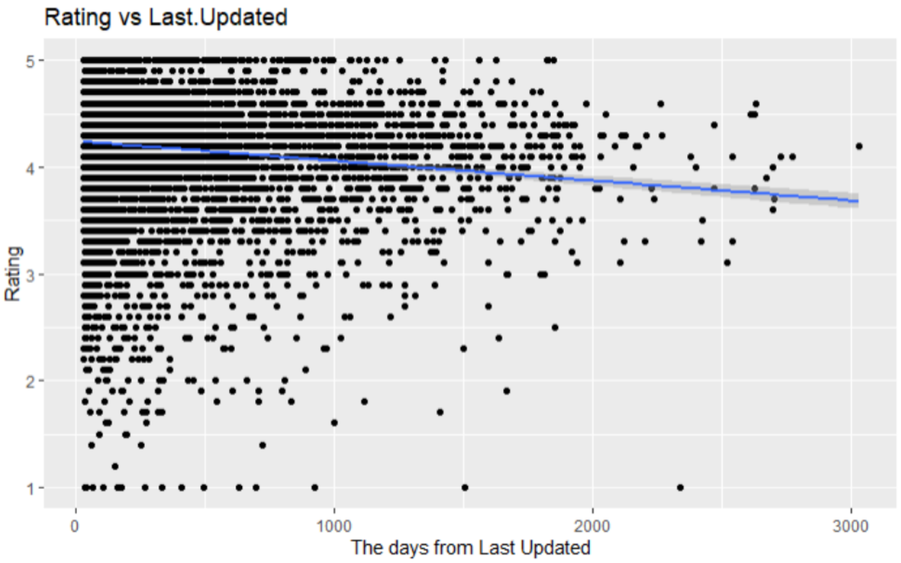
Moreover, the second plot shows that not too many apps are under the ‘communications’ category; thus with a low number of competing apps and the highest number of installs per app, this may be a great app category to create.

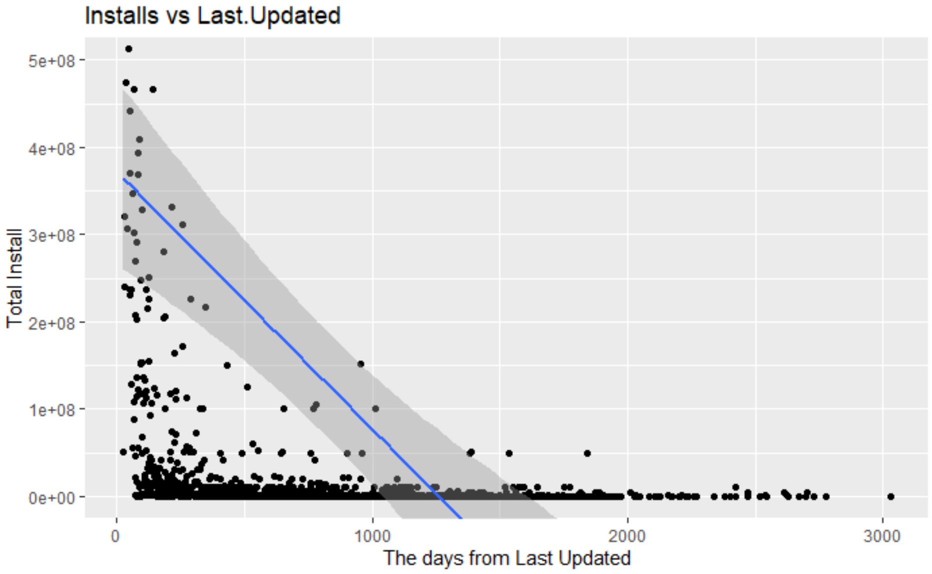




Rating vs Last Updated & Installs vs Last Updated:

Plotted below is the number of days since an application was last updated against the rating of that application and against the number of installations respectively. A line of best fit is plotted to show the average rating. As time passes without an update, the rating and installs tend to decrease. This is easily explainable as updates tend to fix bugs and other complaints consumers would have of the application and lead to low installations and ratings. As such, we hypothesised that ‘Last Updated’ would be significant in determining both ‘Rating’ and ‘Installs’.





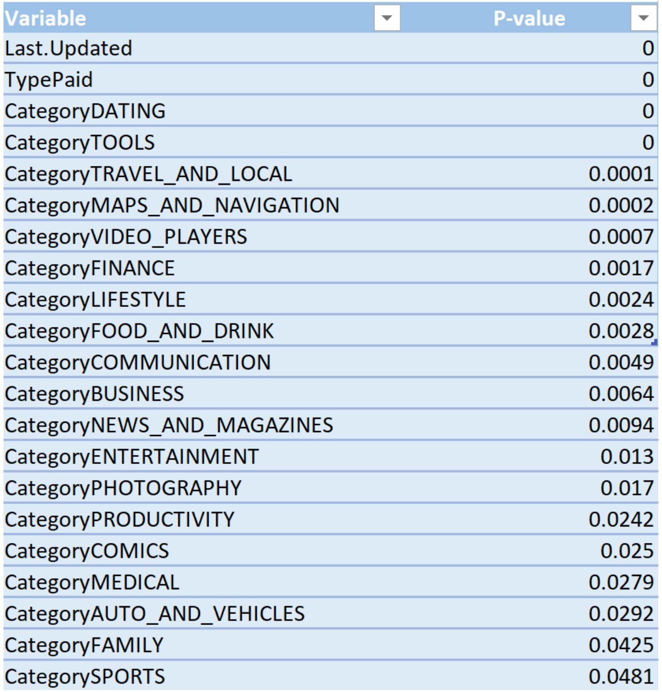
Statistical Analysis:

To verify and support our hypotheses, we used multiple linear regression to model the variables 'Installs' and 'Rating'. We attempted other models such as polr (an adaptation of ordinal regression) as well as other algorithms such as XGBoost. However, these methods gave us insignificant results or were difficult to interpret. Multiple linear regression was used since the interpretability of the significance of each independent variable was simple but important in the understanding of the data. Moreover, the multiple linear regression function in R automatically hot-encodes the categorical columns as factors and thus as variables in our model, which makes the data analysis easier to conduct. With the summary results of the fitted model, we were able to confirm our hypotheses generated in the EDA portion.

Summary and Interpretation of Results:

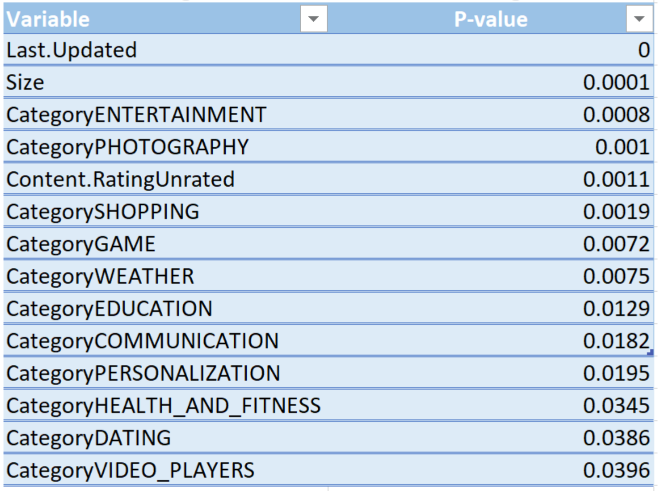
While we used multiple linear regression to model 'Rating' and 'Installs' against 'Category', 'Size', 'Type', 'Content.Rating', 'Last.Updated', our model was not used for the purpose of predicting rating or installs of a new app. Rather, we investigated the significance of each variable in the models and how important they are in influencing the number of installs and rating.

Table 1-1 . Significant Variables in Determining Rating



Variable	P-value
Last.Updated	0
TypePaid	0
CategoryDATING	0
CategoryTOOLS	0
CategoryTRAVEL_AND_LOCAL	0.0001
CategoryMAPS_AND_NAVIGATION	0.0002
CategoryVIDEO_PLAYERS	0.0007
CategoryFINANCE	0.0017
CategoryLIFESTYLE	0.0024
CategoryFOOD_AND_DRINK	0.0028
CategoryCOMMUNICATION	0.0049
CategoryBUSINESS	0.0064
CategoryNEWS_AND_MAGAZINES	0.0094
CategoryENTERTAINMENT	0.013
CategoryPHOTOGRAPHY	0.017
CategoryPRODUCTIVITY	0.0242
CategoryCOMICS	0.025
CategoryMEDICAL	0.0279
CategoryAUTO_AND_VEHICLES	0.0292
CategoryFAMILY	0.0425
CategorySPORTS	0.0481

Table 1-2 . Significant Variables in Determining Installs



Variable	P-value
Last.Updated	0
Size	0.0001
CategoryENTERTAINMENT	0.0008
CategoryPHOTOGRAPHY	0.001
Content.RatingUnrated	0.0011
CategorySHOPPING	0.0019
CategoryGAME	0.0072
CategoryWEATHER	0.0075
CategoryEDUCATION	0.0129
CategoryCOMMUNICATION	0.0182
CategoryPERSONALIZATION	0.0195
CategoryHEALTH_AND_FITNESS	0.0345
CategoryDATING	0.0386
CategoryVIDEO_PLAYERS	0.0396

Our statistical modelling resulted in the p-value table above, which shows the variables that were significant in determining the number of ‘Installs’ as well as ‘Rating’. With α=0.05, we found that the ‘Last.Updated’ variable was the most significant in predicting both the number of ‘Installs’ as well as the ‘Rating’ of the app. Updating an app more frequently would mean developers of an app are constantly fixing bugs, and updating new features which would attract more users while retaining them. For ‘Rating’, whether the app is paid or unpaid was significant with paid apps having higher ratings. We believe this is since free apps have ads and less revenue earned to go back into app development. For the number of ‘Installs’, the variable ‘Size’ had a big influence since the larger the app size, the less installations there would be. Also, we can confirm our hypothesis about ‘communication’ being important in explaining ‘Installs’ since Table 1-2 shows that ‘CategoryCOMMUNICATION’ is significant.

Overall Conclusion:

Based on the exploratory data analysis and statistical analysis, we would recommend an application developer to create an app under the ‘communication’ category. This is because there are few competitors (number of applications of that category), average rating is 4.3 which is moderately high, and has the highest installations when accounting for ratio of communication installations by total applications. The size of the application is insignificant in the rating, but a small application size may lead to higher number of installations. We recommend the application be free as those applications tend to have a much larger number of installations but lower rating. Application content should be everyone 10+ or teen as those tend to have the highest number of installations. We recommend constant updates to bugs and consumer complaints as application installations and ratings decrease over time that an application is not updated.

Limitations of the Study:

Some challenges of this study are that many assumptions with the data. We assumed the date the data was scraped which was over a year and a half ago. Because of this, the Play Store market could have changed significantly, and new apps may have emerged as popular or highly rated. In a number of installations, the data was generally given in varying bins rather than exact numbers. For example, 10,000+, or 50,000+, and so forth.

Recommendations for the Future:

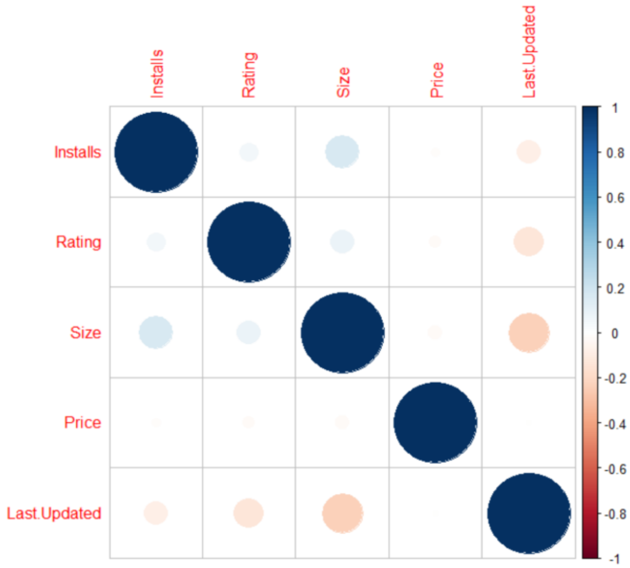
If we had to repeat our study, we would aim to collect data that is more specific in number of installations as that was a limitation in our data analysis. While 10,000 applications in the raw data was originally satisfactory, we would like the data to include more than 50,000 applications considering there are over 2.9 million applications currently on the Google Play Store. In the 10,000 original applications, only 647 were paid apps which gave us the information we needed on paid or free but we would have liked a larger number of applications to generalize to 2.9 million applications.

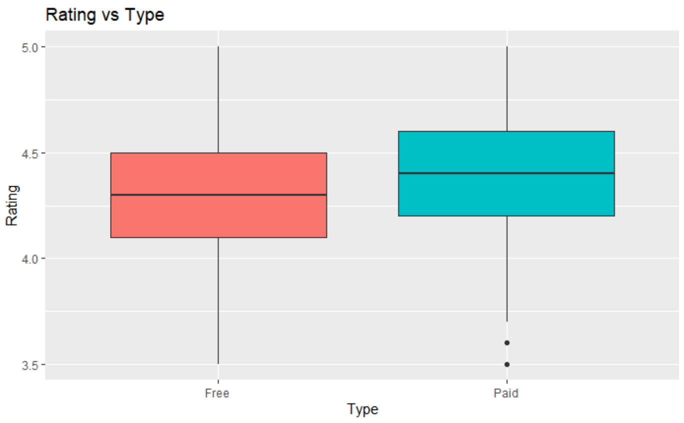
Appendix:

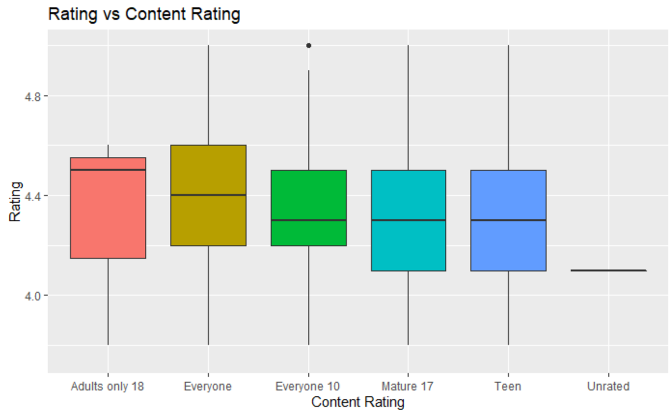
More Details on Variables:

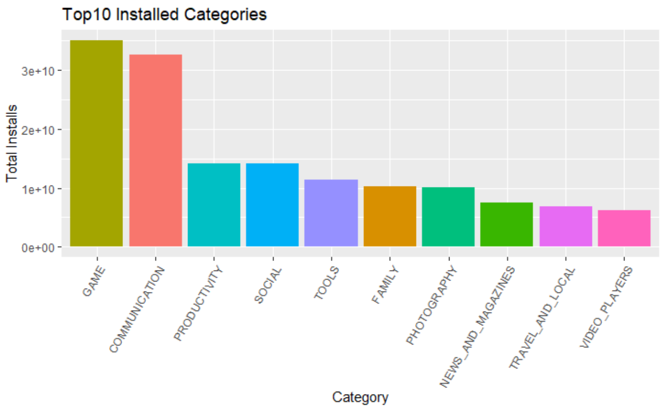
The first column entry is the application, which is straightforward in each row being the name of the application that is being evaluated. The second column entry (categorical) is the category of the application, which lists the first category the application falls into in the Play Store. The third column (numerical) is the average rating by all users who reviewed the application. This column will be used to determine whether an application is rated significantly higher (with an alpha of .05) than others based on specific column inputs. The fourth column entry (numerical) is the number of reviews an application has. This is important to consider because it could lead to skewed results if we use applications with too few reviews. Few results could have the application rated too highly or too low giving us biased information. The fifth column (numerical) is the size of the application which refers to the memory required to download the application. This could be useful but would require a deeper look at certain ranges of memory. The sixth column entry (numerical) tells us the number of device installations of the app. The seventh column entry (categorical) is whether an application is free to download or is paid. This, however, does not give information about in-app purchases which are sometimes necessary to use all functions of an app. The eighth column entry (numerical) specifically measures the price of each app. The options are the cost of the application or 0 if the application is free. The ninth column entry (categorical) is the content rating which is the audience the application is targeted to. The ratings go: everyone, everyone 10+, teen, mature (17+), adult (18+), and unrated. The tenth column entry (categorical) is titled “Genres” which is the category that the application is labelled under in the app store. If there is more than one category for the application, then two are shown separated by a semicolon. The eleventh column entry (numerical) is the day the application was last updated, however, as aforementioned, for our analysis we created a new column showing the difference in days between last update day and September 4, 2018. The twelfth column entry (categorical) is the current version of the application. The limitation is that it may be difficult to understand the update numbers. For example, version 1.0.15 compared to version 2.0.1 of a different application may signal completely different levels of updates. The thirteenth column entry (categorical) is titled Android Version which refers to the minimum Android software required to use the app.

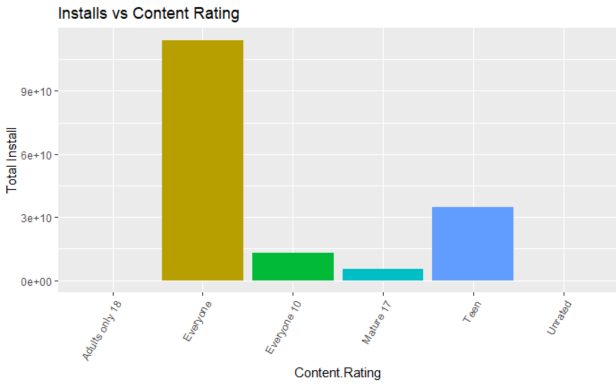
The following includes graphs and charts we created during our analysis that are interesting and insightful but were excluded from the report to keep the report concise.

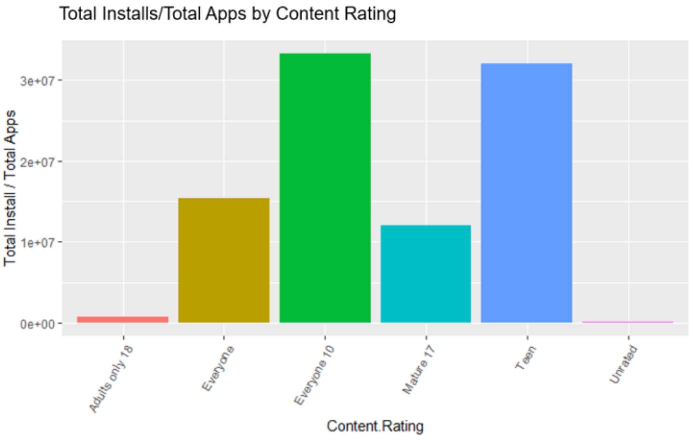



Rating vs Type


Rating vs Content Rating


Top10 Installed Categories


Installs vs Content Rating


Total Installs/Total Apps by Content Rating

Works Cited:

Gupta, Lavanya. “Google Play Store Apps.” Kaggle, 4 Sept. 2018.