The Success of Games

Prepared by:
Derek Tao
Jan Danel
Jay Ralyea

May 6, 2021

Contents

List of tables

List of figures

Executive summary

As the world becomes more and more digital, understanding what makes games successful in gaining users is important for businesses to grow. Entrepreneurs and start-ups need to be creative and clever in order to make their idea stand out against millions of games that are present in online-stores such as the Google Play store. We decided that analyzing successful, and unsuccessful, games already found in the Google Play store could help a game developers learn the details of what successful games have in common in order to make their product thrive. The data utilized came from Mr. Prakash's Google Play Store Apps dataset, retrieved from his Kaggle account, gauthamp10, on March 3rd, 2021. For this analysis, a successful game has more installs than the typical game on the Google Play store. The results received from these three methods show that the category of the game, its size, content rating, whether it supports advertisement, and if it provides in-app purchases are the most significant variables that entrepreneurs and start-ups should consider when developing new games.

1 Introduction

Some of the most successful video games in the world can be found on the Google Play store on an Android phone. In theory, all one needs to create a successful game is some stellar code and a great idea.

Game developers need to be creative and clever in order to make their idea stand out against the millions of games on the Google Play store. Analyzing successful, and unsuccessful, games already found in Google Play could help a developer learn the details of what would make their product thrive. Although games can be very profitable, they also require time and money to be put in up front. A relatively basic game could cost up to $120,000 to create and develop (Lastovetska 2021). Users may not play the game if there are considerable paywalls, but they also may not use it if the game is lacking in quality. What is the middle ground that will push a game to success? We aim to answer these questions through our analysis and research.

1.1 Business problem

One major obstacle that a game developer faces right away is how to create a successful game, which is attractive to new customers and retains them, all the while being profitable. Thus, the specifications of a game and how it monetizes users are a major issue for any new developer. When considering what type of game to create, one should consider multiple specifications such as the game category, package size, and minimum Android version support, to name a few.

Similarly, when considering the package size and the minimum Android version support, one should consider how accessible they want their game to be while balancing other opportunities. If the developer's app only supports the latest Android version, then they might miss out on the market of customers who are currently running the previous version. However, it might also be unnecessary to create a game that supports all Android versions, as the majority of the market is most likely using some of the most recent versions. In the same manner, the developer needs to find a balance between how large they want their game to be and how practical it's actually going to be. The game could end up having great details, but if it starts to lag or takes up a lot of storage, then a user might be unhappy with it and find other games that take up less space and still provide similar functionality. These are just a few of the many relevant game characteristics/specifications that could

impact game popularity and success.

Once the developer has a game that is or is likely to be successful, he or she must consider how to monetize it. Should the developer consider charging for installing the game or allow consumers to download it for free and then charge them after a trial? Similarly, should the developer create a game that is completely free and only make money off of advertising? Lastly, would a combination of in-app purchases and advertising be a better way to maximize profits? These are all strategies that a developer should consider, and each of these strategies may be affected by the game's category. It is possible that a certain type of consumer is willing to pay to download the game but not for anything else. There could also be a game category whose customers are willing to sit through advertisements if it means that they do not have to pay. In other words, it is important to find patterns in profit strategies that lead certain games to be more popular than others in their respective type of category. Once again, determining the most relevant specifications for an app and how to monetize its users are both very important pieces of information needed in order to create a successful and profitable game.

1.2 Intended audience

Our intended audience is game developers looking to increase their game's total install count. By using data scraped from the Google Play store, developers can utilize subsequent analysis to increase their app's total number of installs, a proxy for game success.

Developers can take advantage of a game's success by including ads in the game itself. Successful games with enough downloads are inherently more likely to attract advertisers than unsuccessful games, thereby generating more revenue for the developer. Additionally, game developers can include in-app purchases, which offer perks to the user. Games with more total installs, and therefore a larger user base, will, in general, sell more in-app purchases which means more revenue for the developer. These two mechanisms contributed to the nearly 40 billion dollars in gross app revenue generated in 2020 and are a hallmark of many games found on the Google Play store (Sharma 2020). If game developers hope to maximize their revenue, they should seek to maximize the game's installs.

1.3 Data

To answer the proposed business problem, we are using a dataset that contains a wide variety of game applications from the Google Play store, as well as information on the app characteristics that may be correlated with the popularity and demand of those apps in the store. The data that we will conduct our analysis on includes information from 128,622 different game applications and contains a total of 17 different variables. The variables that are helpful in identifying specific apps and their development details are the game's name, identification number, last date it was updated, and date it was released on the Play store. The variables in the dataset that will allow us to measure and compare the performance of the apps are the game's install count, rating, content rating, and number of ratings. Lastly, the variables we will use to analyse any presence of correlation with popularity are whether or not the game supports ads or in-app purchases, as well as the game's category, price, install size, and minimum Android version, which is the most version of the Android operating system required to play the game.

	App.Name	App.Id	Category	Rating	Rating.Count	Installs	Minimum.Installs	Maximum.Installs	Price	Currency	Size	Minimum.Android	Released	Last.Updated	Content.Rating	Ad.Supported	In.App.Purchases
1	World War 2: Offline Strategy	com.skizze.wwii	Strategy	4.3	17297	1,000,000+	1e+06	2161778	0.00	USD	86M	5.1 and up	19-Jul-18	26-Nov-20	Everyone 10+	TRUE	TRUE
2	Little Pandaâ€™s Dream Town	com.sinyee.babybus.village	Educational	4.0	44700	10,000,000+	1e+07	29969311	0.00	USD	90M	4.2 and up	16-Aug-18	29-Sep-20	Everyone	TRUE	TRUE
3	Baby Panda: Dental Care	com.sinyee.babybus.dentistll	Educational	4.1	10990	10,000,000+	1e+07	12520805	0.00	USD	87M	4.2 and up	27-Apr-20	29-Sep-20	Everyone	TRUE	TRUE
4	Memory Match-7	com.style7.memorymatch7	Board	4.0	7	1,000+	1e+03	2530	0.00	USD	8.5M	2.3 and up	21-Nov-16	21-Nov-16	Everyone	TRUE	FALSE
5	Chess Clock	com.chess.clock	Puzzle	4.4	13534	1,000,000+	1e+06	2465256	0.00	USD	1.0M	2.2 and up	7-May-14	21-Aug-19	Everyone	FALSE	FALSE
6	Moving Cube	com.apperztech.movingcube	Casual	5.0	11	10+	1e+01	44	0.00	USD	7.0M	5.1 and up	7-Feb-20	17-May-20	Everyone	TRUE	FALSE
7	High Climb	inc.apperz.highclimb	Casual	4.9	15	100+	1e+02	422	0.00	USD	14M	4.4W and up	7-Mar-20	15-May-20	Everyone	TRUE	FALSE
8	Chess	cc.chess.full	Puzzle	4.4	1187	10,000+	1e+04	12457	1.99	USD	6.9M	4.1 and up	29-Feb-12	12-Aug-20	Everyone	FALSE	FALSE
9	Slydris	com.radiangames.slydris	Puzzle	4.6	522	5,000+	5e+03	6453	1.99	USD	37M	2.3 and up	7-Jun-13	29-Jul-14	Everyone	TRUE	FALSE
10	Chess Free	cc.chess	Puzzle	4.3	39326	1,000,000+	1e+06	3569670	0.00	USD	8.3M	4.1 and up	28-Feb-11	6-Sep-20	Everyone	TRUE	FALSE
11	Mate in 3-4 (Chess Puzzles)	com.chessking.android.learn.attack2	Board	4.6	5877	100,000+	1e+05	384063	0.00	USD	10M	4.1 and up	3-Dec-15	1-Apr-20	Everyone	TRUE	TRUE
12	unWorded	com.bentostudio.unworded	Puzzle	4.2	255	1,000+	1e+03	3991	0.99	USD	74M	2.3 and up	15-Feb-17	22-Feb-17	Everyone	FALSE	FALSE
13	Untangle	com.ctgames.untangle	Puzzle	4.4	2953	100,000+	1e+05	282567	0.00	USD	5.8M	3.0 and up	21-May-14	26-Apr-17	Everyone	TRUE	FALSE
14	Learn Chess: From Beginner to Club Player	com.chessking.android.learn.beginnerstoclub	Educational	4.2	10431	1,000,000+	1e+06	1305209	0.00	USD	11M	4.1 and up	24-Apr-15	30-Jan-20	Everyone	TRUE	TRUE
15	InBlock	io.github.aleksdev.inblock	Puzzle	4.8	309	10,000+	1e+04	10653	0.00	USD	7.3M	4.1 and up	19-Jun-17	17-Oct-20	Everyone	TRUE	TRUE
16	Zircon - crystal puzzle	com.remaginarium.quartz	Puzzle	4.5	1175	50,000+	5e+04	82280	0.00	USD	11M	4.0.3 and up	18-Apr-16	7-Dec-19	Everyone	TRUE	TRUE
17	2248 Hexa	com.vector.game.puzzle.twodots.numberlink2248hexa	Puzzle	4.5	6346	500,000+	5e+05	622774	0.00	USD	11M	4.1 and up	4-Jan-18	25-Jun-20	Everyone	TRUE	TRUE
18	Unblock Red Wood - slide puzzle	com.coolandroidappzfree.unblockme	Puzzle	4.5	206	10,000+	1e+04	12401	0.00	USD	6.4M	2.3 and up	20-Apr-17	28-Jun-17	Everyone	TRUE	FALSE
19	Ocean Hunter : Match 3 Puzzle	com.superday.aos.oceanhunter	Puzzle	4.6	157	5,000+	5e+03	8354	0.00	USD	44M	4.4 and up	26-Aug-20	19-Nov-20	Everyone	TRUE	TRUE
20	Jewel Witch - Best Funny Three Match Puzzle Game	com.smile.jewel.google	Puzzle	4.4	9874	1,000,000+	1e+06	1047260	0.00	USD	49M	4.1 and up	29-Dec-17	5-Sep-20	Everyone	TRUE	TRUE
21	Block Puzzle: Star Finder	com.bitmango.go.blockpuzzlestarfinder	Puzzle	4.2	20610	10,000,000+	1e+07	10867285	0.00	USD	36M	4.4 and up	7-May-18	3-Nov-20	Everyone	TRUE	TRUE
22	Line Connect Puzzle - Connect Color Dots free	com.yang.flowlinedotconnect	Board	4.8	299	5,000+	5e+03	6211	0.00	USD	15M	4.1 and up	15-Nov-19	5-Jul-20	Everyone	TRUE	FALSE
23	Block Puzzle - Fun Brain Puzzle Games	blockpuzzle.game.block.puzzle	Puzzle	4.7	678	50,000+	5e+04	77615	0.00	USD	44M	5.0 and up	19-Jun-20	27-Nov-20	Everyone	TRUE	TRUE
24	Block Sudoku Puzzle	com.bigc.block.soduku	Puzzle	4.6	8406	1,000,000+	1e+06	2250227	0.00	USD	47M	6.0 and up	19-Aug-20	25-Nov-20	Everyone	TRUE	FALSE
25	Blockpuz	com.sg.block.puzzle.quest.story	Puzzle	4.3	1094	500,000+	5e+05	814723	0.00	USD	15M	4.2 and up	10-Dec-17	4-Aug-20	Everyone	TRUE	FALSE
26	Sudoku Cafe	com.bitmango.sudokucafe1	Puzzle	4.5	2918	100,000+	1e+05	212781	0.00	USD	36M	4.2 and up	9-Oct-12	20-Oct-20	Everyone	TRUE	TRUE
27	UnpuzzleX	com.kekgames.unpuzzlex	Puzzle	4.7	2932	50,000+	5e+04	83360	0.00	USD	9.1M	5.1 and up	10-Jun-20	12-Nov-20	Everyone	TRUE	TRUE

Figure 1: Cleaned Data

1.4 Data collection

The data we will use in our analysis was gathered by Gautham Prakash. This framework allows for a myriad of information to be collected from the millions of apps on the Google Play Store. We downloaded Mr. Prakash's Google Play Store Apps dataset from his Kaggle account (gauthamp10) on March 3rd, 2021.

From initial viewing of the data set, the data appear to be very appropriate for answering the business problem. It contains over 140,000 games across all 17 game genres

featured on the Google Play Store. This ensures that analysis of the business problem will be of interest and benefit to game developers within all game genres. Fortunately, there are only a few games with missing data, and these missing values are not widespread across the rows. The data values are relatively clean and do not appear to contain input errors. The one quality issue that may affect our analysis is the imprecision of the install counts of the games. The "Installs" and "Minimum Installs" variables only provide a range of installs. In order to adequately perform analysis, a numeric variable is needed to determine whether or not a game is more successful than others. An exact number of installs is provided as a variable called "Maximum Installs." This variable will be used as the target throughout our analysis and will be referred to as the "install count." This also allowed us to create our target variable. Games in the data set were assigned a 1 if they were in the 75th percentile of total installs, and a 0 if they were not. Additionally, to balance the number of successes and failures in the analysis, a random sample of 500 successful games and 500 failed games was collected from the original data.

Variables	Description
App Name	The name of the application
App Id	Unique application id
Category	Under what category is the application categorized (what type of game)
Rating	The average rating of the application by Google Play Store reviews
Rating Count	How many individuals have made a review on the Google Play Store of the specific app
Installs	The approximate number of current app installations in categories (0+, 1+, 5+, 10+, 100+, 500+, 1000+, 5000+, etc.)
Minimum Installs	The minimum number of app installations in the current Installs category (if Installs in 500+, then Minimum Installs is 500)
Maximum Installs	The maximum number of installations the app has had since its launch at a given time
Price	The US dollar amount needed to download the application ($0 if free)
Currency	The currency used to make purchases of and in the application (all in US dollars)
Size	The size of the application in Kb or Mb
Minimum Android	The minimum version of Android that is supported in the application
Released	The date on which the app was released

Variables	Description
Last Updated	The date of the last update
Content Rating	Age suitability rating based on the content of the application (Teen, Everyone, etc.)
Ad Supported	Whether the app supports advertisement (TRUE) or not (FALSE)
In App Purchases	Whether the app has in app purchases (TRUE) or not (FALSE)

1.5 Data preparation

Our first step was to remove column variables in the data set that are certain to be irrelevant to our business problem. We removed the columns: free, developer ID, developer email, developer website, privacy policy, and editor's choice. This left us with the following columns: ad supported, app id, app name, category, content rating, currency, in-app purchases, last update, maximum installs, minimum Android version, price, installs, minimum installs, rating, rating count, released, and size. We chose not to include the variable free because of the redundancy with price. If the price of a game is 0 dollars, we will be able to count the app as free. Additionally, information such as the developer's ID or email is unlikely to be important in determining an app's success. We also decided to remove all apps whose general currency was not US dollars. Similarly, we also dropped all the apps who did not specify their download size nor the minimum Android version required for the game to run. We also decided to remove any of the remaining apps which did not have information on all the column variables that we are interested in using for our analysis.

In our preparation process, we emphasized the importance of having complete data because we believe that games with complete information on the Google Play store are a better representation of the store as a whole. Consistency in the data was also important to us, especially for quantitative data because it must be scaled properly to avoid misleading analysis in the future. This is why keeping currency constant (USD) and avoiding ambiguous data, such as "varies with device" in the size and minimum Android version columns, were necessary decisions. Lastly, we created the response variable. For games considered particularly successful, with a high install count, they received a "1" and those with lower install counts received a "0."

1.5.1 Descriptive analytics

Offering in-app purchases is one of the ways game developers seek to generate revenue through their app. Game developers can take advantage of players' desires to finish tasks early and often by selling in-game currency for US dollars which the players can then use on various perks in the game. This is commonly referred to as a "pay to win" strategy (**???**). These kinds of in-app purchases can be considered obtrusive and can distract from the quality of the game. The plots below compare games with in-app purchases to those without in-app purchases.

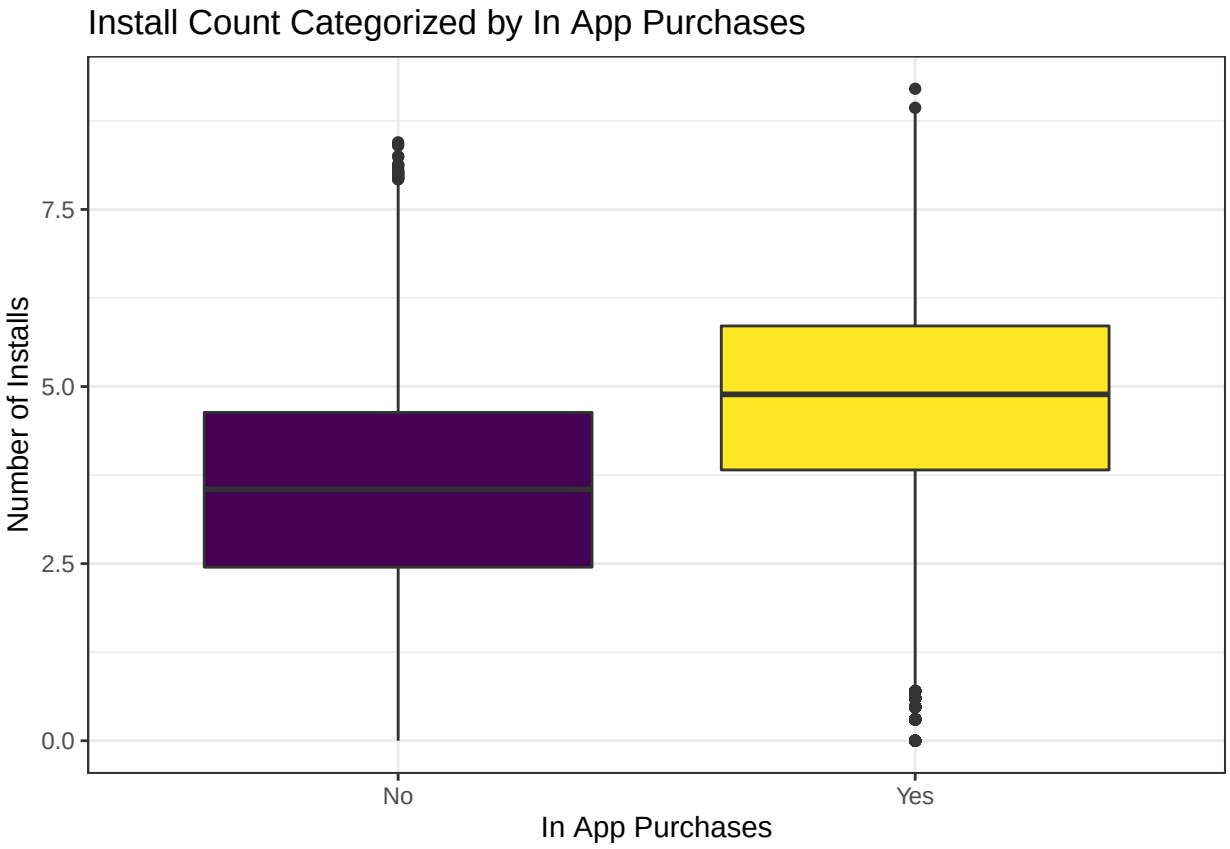


Figure 2: Boxplots for the True and False values of in-app purchases with respect to number of installs (log base ten)

From the plot we can see that the median of the install count for games with in-app purchases is higher than the median for those without in-app purchases. This may imply, at the very least, in-app purchases within a game do not deter gamers from downloading said game. Therefore, including in-app purchases may be a lucrative method of generating revenue.

Displaying ads within a game is another way game developers can monetize their app. Similar to in-app purchases, players may consider ads obnoxious as they are an added component completely irrelevant to the game. The plots below compare games with ad support to those without ad support.

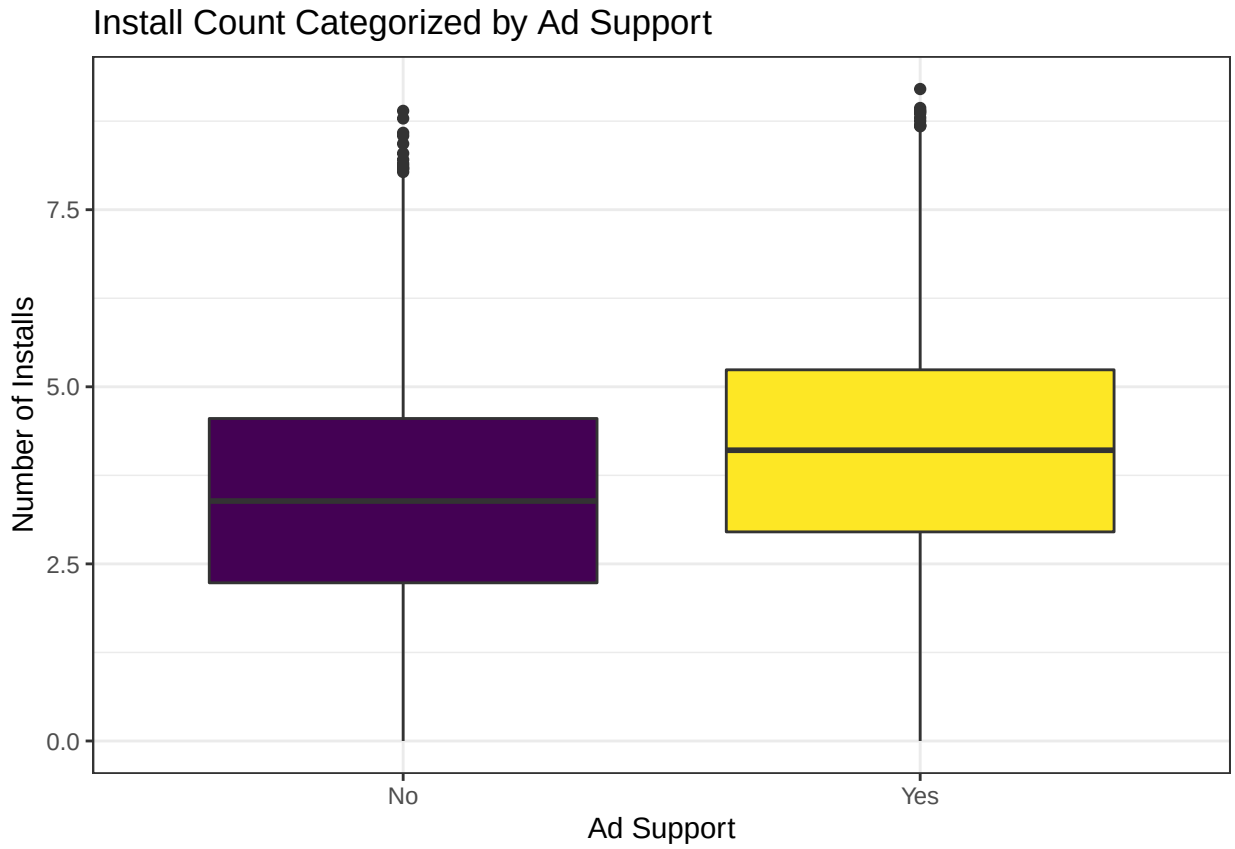


Figure 3: Boxplots for the True and False values of ad support with respect to number of installs (log base ten)

In the above plot it is apparent that apps with ad support have a higher total number of installs than apps without ad support. While certainly not as large a difference as with in-app purchases, it may be the case that ads do not detract from gamers' experiences enough to render the game unplayable. Therefore, the inclusion of ads in a developer's game may be a reliable source of revenue.

The graph below shows the medians of the install counts for games in all 16 game categories in the Google Play Store.

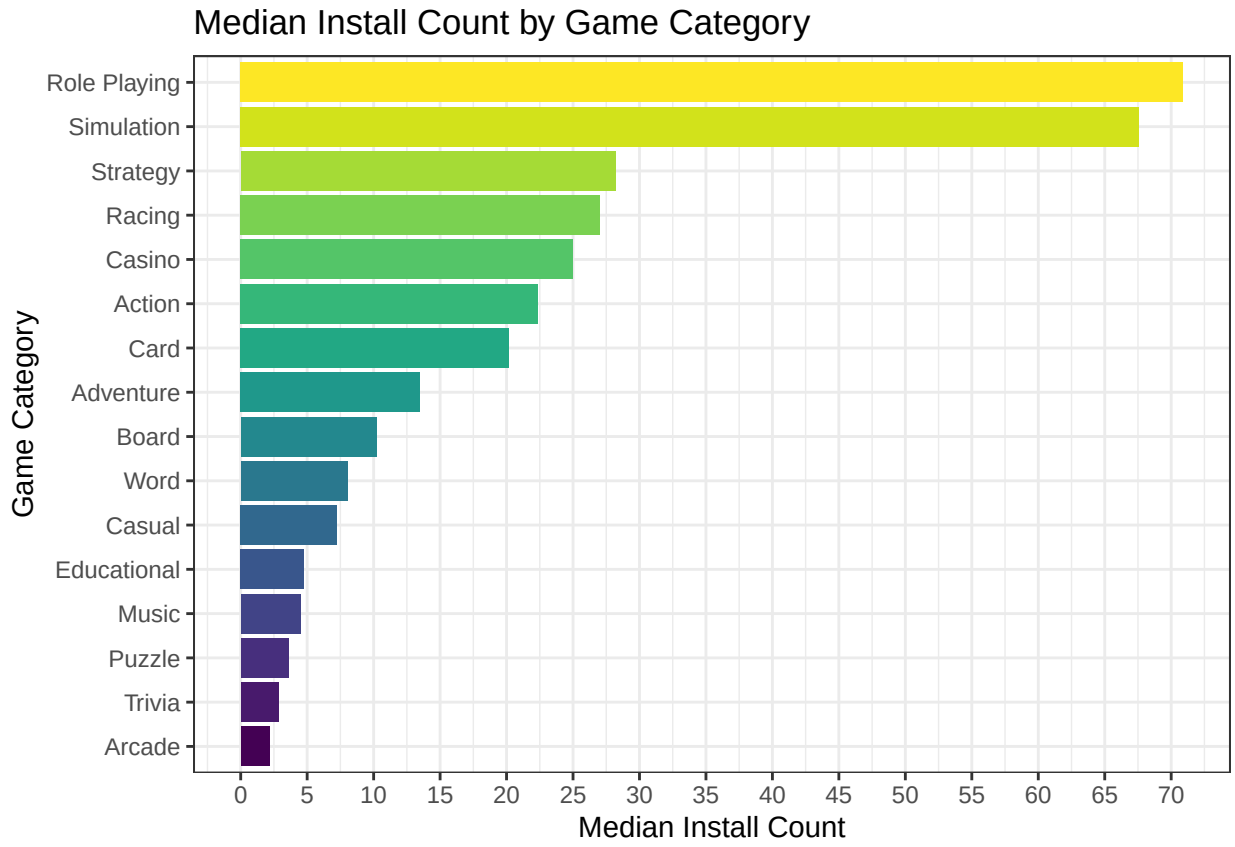


Figure 4: Shown are the median install counts, in thousands of installs, for each of the 16 game categories in the Google Play Store

As shown in the figure, the role playing and simulation genres are by far the most popular categories in the store, based on the median of the install counts. The medians for these two categories are over twice as high as those of the next most popular categories. Action, card, casino, racing, and strategy are also relatively popular categories, while trivia, puzzle, music, education, and arcade are the least popular.

The plot below shows the relationship between game size in MB and install count for all games in the data set. A Log10 transformation was applied to the install counts to narrow the scale of the install counts, since the range of install counts in the data is very wide.

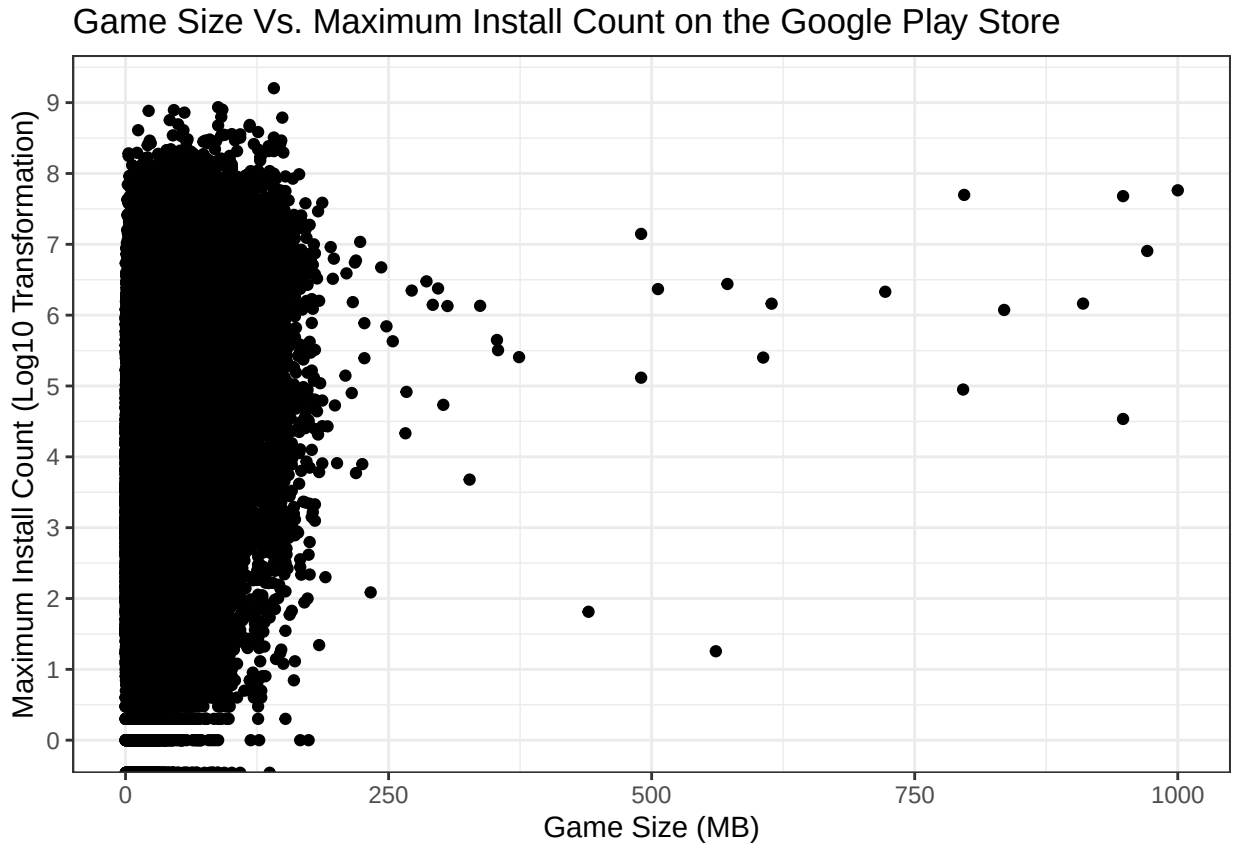


Figure 5: Shown are the Log10 transformations of the maximum install counts for all games in the dataset, plotted against the size of each game in megabytes

Since game size and install counts are both continuous quantitative variables, the above plot is appropriate for examining the presence of a correlation between the two variables. Based on the plot, it is pretty clear that there is no noticeable relationship between game size and install counts.

Table 2: Revenue Strategy by Category

Category	Ad Supported	In App Purchases
Action	90.06%	89.07%
Adventure	81.26%	77.33%
Arcade	97.38%	80.86%
Board	96.54%	69.51%
Card	89.09%	62.44%
Casino	82.09%	91.34%
Casual	96.78%	72.93%
Educational	90.02%	74.14%
Music	94.75%	79.81%
Puzzle	96.32%	74.06%
Racing	96.83%	84.15%
Role Playing	79.64%	89.12%
Simulation	95.42%	71.08%
Strategy	55.15%	93.49%
Trivia	97.43%	74.50%
Word	98.16%	84.52%

The table above shows that 11 out of the 16 game categories had a value greater than 90.00% for the total number of installations regarding the ad supported variable. This leads us to believe that at least 90% of users in these categories are willing to watch ads to play, which is a very promising statistic when considering advertisement revenue. Similarly, we are able to determine that only 3 categories have more installations that support in-app purchases than support advertisements; casino, role playing, and strategy. Including these three categories, only 7 categories had a value greater than 80.00% for in-app purchases. This statistic is not as high as for ad supported and the number of categories is less than half of all categories being analyzed, so we do not expect in-app purchases to be more effective at generating revenue than advertisements. However, it is also very important to notice that all values are greater than 55.00% in both columns. The only category which had similar percentages for these two columns was action, with ad supported at 90.06% and in-app purchases at 89.07%. Meanwhile, for the rest of the categories, the difference between the values of ad supported and in-app purchases was mostly greater than 10 basis points.

1.5.2 Insight summary

The first visualization demonstrates that the inclusion of in-app purchases, a hallmark of pay to win gaming, likely does not decrease the chances a person installs the game. Similarly, from the second visualization we can see it is unlikely that ad support in a game negatively impacts the chances a gamer downloads the app. It may be the case that ads do not detract from gamers' experiences enough to render the game unplayable. These initial results are surprising given individuals frequently consider both in-app purchases and ads to be nuisances.

Before beginning game development, developers should consider the game categories that are the most popular and lucrative. The third visualization gives us an idea of which game categories produce the highest median of install counts, and thus should potentially be targeted by game developers. It also shows which game categories produce the lowest median of install counts, and thus should potentially be avoided. Role-playing and simulation are by far the most popular categories by this metric, while trivia, puzzle, music, educational, and arcade appear to be the categories that developers should consider avoiding if their goal is to maximize install counts. The presence of the arcade category among the least popular categories is particularly surprising, considering the longevity of arcade games in the gaming industry.

The business problem originally proposed the possibility that the size of a game, or the amount of space it takes up on a device, can be related to the game's popularity. The fourth visualization allows us to examine the presence or absence of a relationship between the game size variable and the games' install counts. The vast majority of the games are concentrated within a size range of 0-125 MB, and the points within this range are very uniformly distributed. as opposed to showing a clear positive or negative direction. Based on this observation, it is safe to conclude that a game's size is not correlated with its install count, and that the game size variable is not significant to the business problem as previously believed.

Lastly, from the table Revenue Strategy by Category, it is interesting to see that 13 of the 16 game categories had a percentage of ad-supported games higher than the percentage of games containing in-app purchases, and that all but two categories had an ad-supported percentage larger than 80.00%. Meanwhile, in-app purchases only had seven categories with a value larger than 80.00%. This leads us to believe that for games

in general, using advertisements is a more effective revenue strategy than having in-app purchases.

2 Predictive Analytics

In order to be a successful game on the Google Play store, the game should be installed on a large number of devices. For our purposes, successful games are those that fall within the top 25 percent of maximum installs. We hope to discover what variables may lead to a mobile game's success in regard to total installs and how influential said variables are in determining success. To accomplish our goal we built three separate models, namely logistic regression, decision tree modelling, and random forests.

3 Process

The first predictive analytics process used to gain insight of successful games was logistic regression. The variables included in the model are category, size, minimum android version, content rating, ad support, and in app purchases. To balance the number of successes and failures in the analysis, a random sample of 500 successful games and 500 failed games was collected from the original data. The decision tree and random forest models provided additional information about relevant predictor variables. In the decision tree we included the variables category, minimum android version, content rating, ad supported, and in app purchases, in order to gain insight into their significance in classifying if a game is successful or not. This returned a complex tree which we shortened in order to facilitate interpretation. In a similar manner, the variables in the decisions tree were also used in the random forest. Here, the sample of the data set used in logistic regression was also used for the decision trees and random forests. For all three models the data are separated into three sections – two for exploration and comparison and one for testing performance.

4 Assessments

For the first iteration of logistic regression, we used the following variables: category, size, minimum android version, content rating, ad support, and in-app purchases. After testing the model with and without the minimum android version variable included, it was determined that the variable was insignificant, so it was removed. Further testing also showed the

Table 3: VIF Values for Each Explanatory Variable

	GVIF	Df	GVIF^(1/(2*Df))
Size	1.036176	1	1.017927
Price	1.015316	1	1.007629
Content.Rating	1.034374	3	1.005649
Ad.Supported	1.037526	1	1.018590
In.App.Purchases	1.034189	1	1.016951

category variable is insignificant, leading to its removal. Thus, the remaining variables in the model were size, content rating, ad support, and in-app purchases. Some content ratings appeared to be insignificant; however, one final test showed that the model improved with content rating's inclusion. Below is the summary information for the best logistic regression model:

```
Coefficients:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)               -2.364e+00  4.022e-01  -5.877 4.18e-09 ***
Size                       1.255e-05  3.573e-06   3.512 0.000445 ***
Price                     -1.857e+00  1.086e+00  -1.710 0.087254 .
Content.RatingEveryone 10+ 8.828e-01  4.652e-01   1.898 0.057747 .
Content.RatingMature 17+   1.099e+00  7.069e-01   1.555 0.119921
Content.RatingTeen         6.836e-01  2.278e-01   3.002 0.002685 **
Ad.SupportedTRUE           1.196e+00  3.480e-01   3.437 0.000587 ***
In.App.PurchasesTRUE       1.174e+00  1.862e-01   6.304 2.89e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 6: Logistic Regression Summary

Additionally, we checked for multicollinearity among the explanatory variables. Multicollinearity means that the explanatory variables exhibit linear relationships with each other, which is undesirable for logistic regression. Since the VIF for each variable is close to 1, we concluded that there is negligible multicollinearity.

To measure the performance of the logistic regression model, we generated a confusion matrix and an ROC curve for the model using the testing data. The confusion matrix shows us the number of true positives, true negatives, false positives, and false negatives that the model yields for predicted successes and failures in the testing data:

The confusion matrix has an accuracy of 63.18%, which indicates moderate strength for the model. Next, plotting the ROC curve for the model allows us to see the performance of the model at various classification thresholds for success. The closer the ROC curve is to the upper left corner of the plot, the more useful the model.

Table 4: Logistic Regression Confusion Matrix

Predicted	Truth	
	No	Yes
No	52	41
Yes	33	75

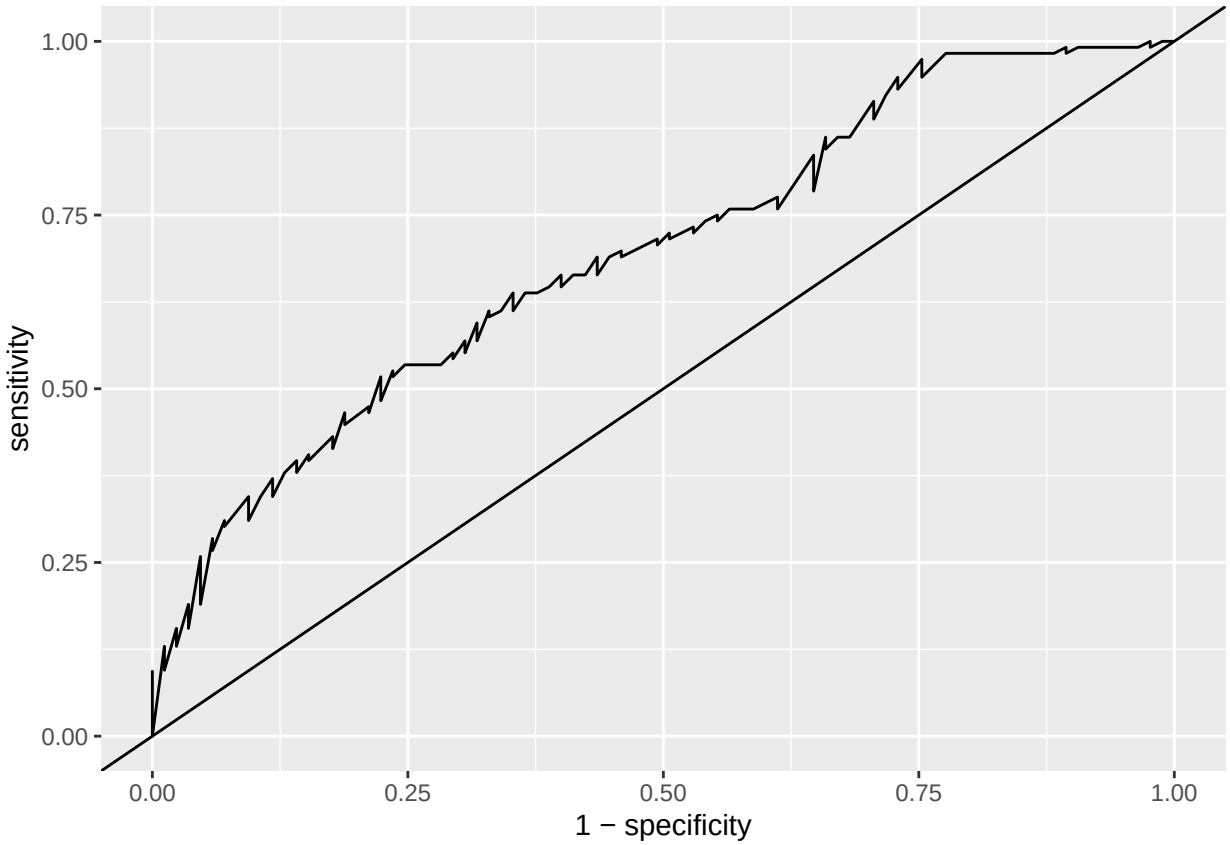


Figure 7: ROC Curve

The area under the ROC curve is 0.694. An area of 1 indicates maximum usefulness for the model, so we conclude that the model has moderately strong usefulness.

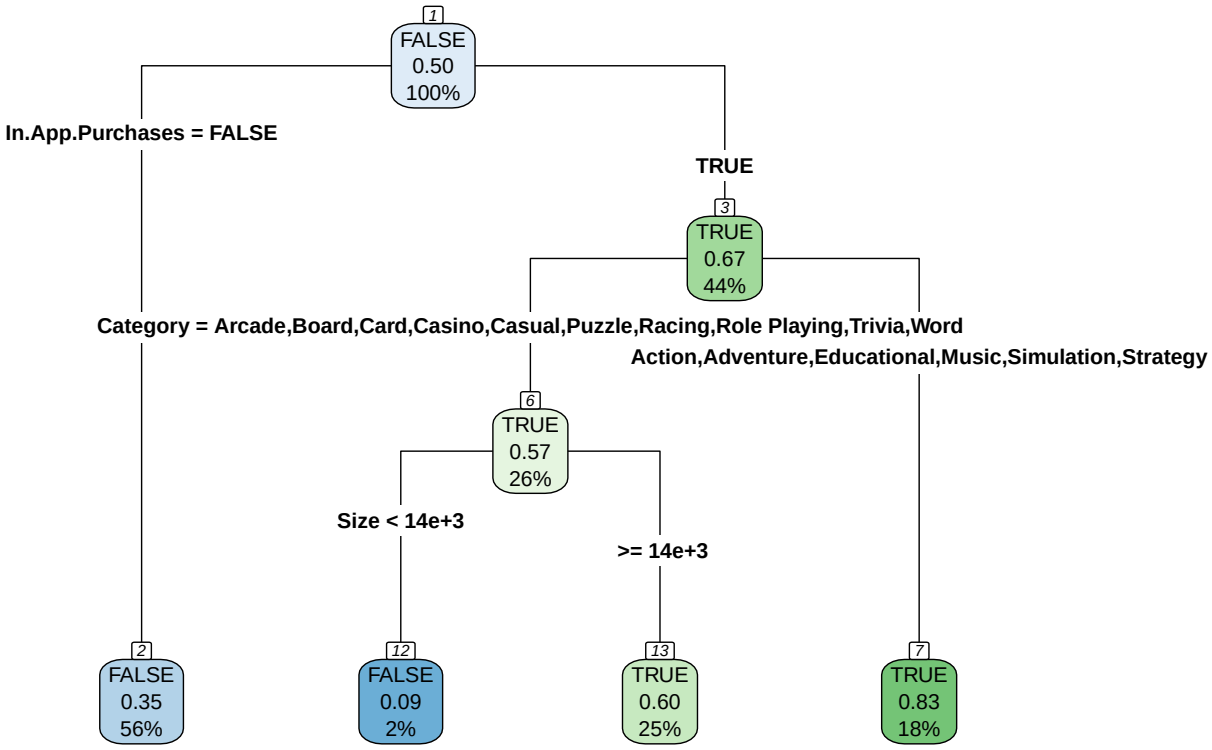


Figure 8: Decision Tree

The decision tree above is a pruned tree, with a complexity parameter of 0.015, that is more accurate and easier to interpret than the original tree. This pruned tree shows that in-app purchases was the most important predictor, followed by the category that the game is in, and then the size of the game. Surprisingly, the variable ad-supported is not included which is in contradiction to our previous expectations. Additionally, we expected that in-app purchases would negatively impact a game's chance of success. However, in the very first split, which is the most important, we can see that the inclusion of in-app purchases leads to a game's success.

From the decision tree visualization we can see that a successful game provides in-app purchases and falls in the categories of Action, Adventure, Educational, Music, Simulation, or Strategy. Similarly, a successful game can also provide in-app purchases, be in any game category, but the size of the game has to be at least 14,000.0 megabytes large. Meanwhile, a non-successful app does not provide in-app purchases, or it is smaller than 14,000.0 megabytes and falls in the following categories: Arcade, Board, Card, Casino, Casual, Puzzle, Racing, Role Playing, Trivia, or Word. By using these variables to split the

Table 5: Decision Tree Confusion Matrix

Predicted	Truth	
	FALSE	TRUE
FALSE	72	38
TRUE	28	62

data, we get an accuracy of 0.67 as can be seen in the confusion matrix below.

Table 6: Summary of Important Variables From the Decision Tree

	FALSE	TRUE	MeanDecreaseAccuracy	MeanDecreaseGini
Category	11.3282614	6.9462256	12.3729798	59.55598
Content.Rating	0.2688097	-0.3051276	-0.1273118	14.11734
Size	3.5020067	9.3675517	9.5065373	67.31864
Ad.Supported	7.0578120	7.9526063	10.3728666	10.78433
In.App.Purchases	27.6893509	26.9315600	34.8294165	27.92081

Table 7: Random Forest Confusion Matrix

Predicted	Truth	
	FALSE	TRUE
FALSE	65	20
TRUE	35	80

After creating a random forest on the most significant variables in classifying success or not we get the previous importance table. We are able to see that the three most significant variables in predicting true success are in-app purchases, size and ad-supported. Similarly, for classifying false failures, the three most important variables are in-app purchases, category and ad-supported.

The random forest returned very similar results as the decision tree. The table of importance shows that size is the most important variable, followed by category, in-app purchases, content rating, and lastly ad-supported, when looking at the mean decrease in the gini coefficient. Although our random forest model is similar to our decision tree, we are able to see from the confusion matrix below that the random forest has an accuracy rate of 0.725 which is 0.055 percentage points higher than the decision tree.

5 Results

From the logistic regression model, we found that size, content rating, ad support, and in-app purchases are the significant variables for predicting the probability of a game achieving the 75th percentile of maximum install count in the Google Play Store. Surprisingly, game category was not found to be a significant predictor, despite our EDA suggesting that there may be a relationship between game category and install count. When taking the exponential of the coefficients of the predictor variables, the result is greater than 1 for all of the predictor variables. This means that the predictor variables are positively correlated

with probability of success, according to the model. Specifically, increasing the size of the game will increase the probability of a game's success. In regards to content rating, the rating "Everyone 10+" has the most positive impact on a game's success. Finally, the presence of ad support and in-app purchases both increase the probability of a game's success.

The results from the decision tree and random forest were both very similar and showed that three of the most important variables that would lead to a game being successful (being in the top 25th percentile of maximum downloads) are in-app purchases, category, and size. Additionally, ad-supported was a significant variable in classifying true success in the random forest model even though it is not present in our pruned decision tree. As for performance, the accuracy of the decision tree was 0.67 while the accuracy for the random forest was slightly higher at 0.725. Even though the accuracy for the random forest was higher, the main reason why we believe that the random forest is the better model is because it more accurately classified successful games, which is what we are most interested in. The random forest correctly classified 80 games as successful while the decision tree only correctly classified 62 as successful.

6 Insight Summary

From our three predictive analytics methods, we conclude that the most important factors game developers should consider when developing new games are category, size, content rating, ad support, and in-app purchases. Each category attracts a different type of mobile gamer, so some categories may attract more people than others, which is why "category" appears to be an important variable. Size is likely important because if a game is too large, it seems unlikely average gamers would free up space on their phone for that specific game. In terms of content rating, games like "Call of Duty," while popular, do not appeal to everyone with its mature content. On the other hand, games such as "Words with Friends," while lighthearted with content for all ages, do not have the thrill associated with more intense games. "Subway Surfers," the most downloaded game on the Play Store, and "Asphalt 9," a popular racing game, strike this balance optimally with an "Everyone 10+" rating. Lastly, it seems the inclusion of ad support and in-app purchases in a game as a positive impact are flukes. It seems unlikely a gamer would install a game because the game has in-app purchases or has ads, both of which could be considered nuisances in a game.

7 Conclusions

The insight that we gained from our analysis shows that the categories classified by our models with the most successful games are Action, Adventure, Educational, Music, Simulation, or Strategy. In the same way, we found that the successful games in our data tend to be larger games, starting at 14,000 megabytes. There was also a clear positive relationship between the size of the game and the probability of it being a successful game. As for content rating, we found that being classified as everyone 10 and up was most common among successful apps. Each of these variables likely contributes to the success of a game, whereas our results for in-app purchases and ad support were unexpected. The models suggested the inclusion of in-app purchases and ads within a game will increase a game's install count. Given gamer's supposed aversion to these features, it is possible these features are included once a game is successful, rather than including them at the beginning.

8 Recommendations

For these reasons stated in our conclusion, we believe that a game developer who wants to develop a successful game should consider creating a game that is at least 14,000 megabytes, can be classified as everyone 10 and up, provide in-app purchases, support advertisements, and be in the following categories: Action, Adventure, Educational, Music, Simulation, or Strategy. A developer does not have to fulfill all of these criterias in order for the game to be successful, it is still possible that creating a game that fits into one of the six categories, without fulfilling the other criteria, will help the game become successful. In a similar manner, it is statistically probable that one could create a game based on these suggestions and not have it become a successful game. The 5 criterias that we specify are variables that successful games share and would increase the probability of being classified as a successful game based on our models, it does not guarantee a game to become successful.

References

Lastovetska, Anastasiia. 2021. "App Development Cost: Understand Your Budget to Build Powerful Apps." MLSDev; Accessed March 3, 2021. https://mlsdev.com/blog/app-development-cost.

Prakash, Gautham. 2020. "Google Play Store Apps." Accessed March 3, 2021. https://www.kaggle.com/gauthamp10/google-playstore-apps.

Sharma, Avinash. 2020. "Top Google Play Store Statistics 2019-2020 You Must Know." AppInventiv; Accessed March 4, 2021. https://appinventiv.com/blog/google-play-store-statistics/.

Appendix A: Data

App Name	App Id	Category	Rating
World War 2: Offline Strategy	com.skizze.wwii	Strategy	4.3
Little Pandaâ€™s Dream Town	com.sinyee.babybus.village	Educational	4.0
Baby Panda: Dental Care	com.sinyee.babybus.dentistII	Educational	4.1
Memory Match-7	com.style7.memorymatch7	Board	4.0
Chess Clock	com.chess.clock	Puzzle	4.4
Moving Cube	com.apperztech.movingcube	Casual	5.0

Rating Count	Installs	Minimum Installs	Maximum Installs
17297	1,000,000+	1e+06	2161778
44700	10,000,000+	1e+07	29969311
10990	10,000,000+	1e+07	12520805
7	1,000+	1e+03	2530
13534	1,000,000+	1e+06	2465256
11	10+	1e+01	44

Price	Currency	Size	Minimum Android	Released	Last Updated
0	USD	86M	5.1 and up	19-Jul-18	26-Nov-20
0	USD	90M	4.2 and up	16-Aug-18	29-Sep-20
0	USD	87M	4.2 and up	27-Apr-20	29-Sep-20
0	USD	8.5M	2.3 and up	21-Nov-16	21-Nov-16
0	USD	1.0M	2.2 and up	7-May-14	21-Aug-19
0	USD	7.0M	5.1 and up	7-Feb-20	17-May-20

Content Rating	Ad Supported	In App Purchases
Everyone 10+	TRUE	TRUE
Everyone	TRUE	TRUE
Everyone	TRUE	TRUE
Everyone	TRUE	FALSE
Everyone	FALSE	FALSE
Everyone	TRUE	FALSE

Appendix B: Data preparation details

R

First, we loaded in the necessary package and imported the data.

```
apps <- read.csv("Google-Playstore.csv") # reading in the data
```

Next, we read in the data and created a list of columns to drop and categories to keep in the data.

```
dropcol<-c('Free','Developer.Id','Developer.Website','Developer.Email',
           'Privacy.Policy','Editors.Choice') # vectorizing the columns to drop

keepcat<-c('Action','Adventure','Arcade','Board','Card','Casino','Casual',
           'Educational','Music','Puzzle','Racing','Role Playing','Simulation',
           'Strategy','Trivia','Word') # vectorizing the categories to keep
```

We then used the functions found in the dplyr package to select and filter said columns and categories.

```
apps <- apps %>%
  select(!all_of(dropcol))

apps <- apps %>%
  filter(Category %in% keepcat)
```

The drop_na function in the tidyr package then removed all entries that had an NA value in any column.

```
apps <- drop_na(apps)
```

Afterward, we filtered to only apps whose currency is the US dollar.

```
apps <- apps %>%
  filter(Currency == "USD")
```

Similarly, for Size and Minimum.Android we removed all observations which were classified as "Varies with device" in order to be able to apply the same analysis on all apps.

```
apps <- apps %>%
  filter(Size != "Varies with device", Minimum.Android != "Varies with device")
```

In our final preparation step, we found empty inputs in the variables Minimum.Android and Released so we removed those observations from our dataset as they were incomplete data. This left us with 128622 apps (observations) which is the same amount, and the same data, that we reached while cleaning the data in Excel.

```
apps <- apps %>%
  filter(Released != "", Minimum.Android != "")
```

Finally, in order to create our response variable, we determined the 75th percentile of install count to be 140,167. Any game with more downloads is a success, and any game with fewer is a failure.

```
apps <- apps %>%
  mutate(good_value=as.numeric(Maximum.Installs>140167))
```

To show the first few rows and all the column of the data, we piped the head() of the data, which shows the first 5 rows, into the select() function to pick a few columns of the data, and then piped that into the kable() function to finally print the specified columns in select() in a clear and organized manner. We printed a total of 5 tables so that the values inside the table were clear enough to read and understand, making sure that we printed all the different columns present in our finalized and clean data.

```
head(apps) %>% select(App.Name : Rating) %>%
  kable(col.names = c("App Name", "App ID", "Category", "Rating"))
```

App Name	App ID	Category	Rating
World War 2: Offline Strategy	com.skizze.wwii	Strategy	4.3
Little Pandaâ€™s Dream Town	com.sinyee.babybus.village	Educational	4.0
Baby Panda: Dental Care	com.sinyee.babybus.dentistII	Educational	4.1
Memory Match-7	com.style7.memorymatch7	Board	4.0
Chess Clock	com.chess.clock	Puzzle	4.4
Moving Cube	com.apperztech.movingcube	Casual	5.0

```
head(apps) %>% select(App.Name, Rating.Count : Maximum.Installs) %>%
  kable(col.names = c("App Name", "Rating Count", "Installs",
                      "Min. Installs", "Max. Installs"))
```

App Name	Rating Count	Installs	Min. Installs	Max. Installs
World War 2: Offline Strategy	17297	1,000,000+	1e+06	2161778
Little Pandaâ€™s Dream Town	44700	10,000,000+	1e+07	29969311
Baby Panda: Dental Care	10990	10,000,000+	1e+07	12520805
Memory Match-7	7	1,000+	1e+03	2530
Chess Clock	13534	1,000,000+	1e+06	2465256
Moving Cube	11	10+	1e+01	44

```
head(apps) %>% select(App.Name, Price : Minimum.Android) %>%
  kable(col.names = c("App Name", "Price", "Currency", "Size",
                      "Min. Android Version"))
```

App Name	Price	Currency	Size	Min. Android Version
World War 2: Offline Strategy	0	USD	86M	5.1 and up
Little Pandaâ€™s Dream Town	0	USD	90M	4.2 and up
Baby Panda: Dental Care	0	USD	87M	4.2 and up
Memory Match-7	0	USD	8.5M	2.3 and up
Chess Clock	0	USD	1.0M	2.2 and up
Moving Cube	0	USD	7.0M	5.1 and up

```
head(apps) %>% select(App.Name, Released : Content.Rating) %>%
  kable(col.names = c("App Name", "Released", "Last Updated",
                      "Content Rating"))
```

App Name	Released	Last Updated	Content Rating
World War 2: Offline Strategy	19-Jul-18	26-Nov-20	Everyone 10+
Little Pandaâ€™s Dream Town	16-Aug-18	29-Sep-20	Everyone
Baby Panda: Dental Care	27-Apr-20	29-Sep-20	Everyone
Memory Match-7	21-Nov-16	21-Nov-16	Everyone
Chess Clock	7-May-14	21-Aug-19	Everyone
Moving Cube	7-Feb-20	17-May-20	Everyone

```
head(apps) %>% select(App.Name, Ad.Supported : good_value) %>%
  kable(col.names = c("App Name", "Ad Supported", "In App Purchases",
                      "Good Value"))
```

App Name	Ad Supported	In App Purchases	Good Value
World War 2: Offline Strategy	TRUE	TRUE	1
Little Pandaâ€™s Dream Town	TRUE	TRUE	1
Baby Panda: Dental Care	TRUE	TRUE	1
Memory Match-7	TRUE	FALSE	0
Chess Clock	FALSE	FALSE	1
Moving Cube	TRUE	FALSE	0

Excel

The first step was to remove column variables in the data set that are certain to be irrelevant to our business problem. The exact columns removed were Free, Developer ID, Developer Email, Developer Website, Privacy Policy, and Editor's Choice. Here is the data set after this step:

App Name	App Id	Category	Rating	Rating Count	Installs	Minimum Installs	Maximum Installs	Price	Currency	Size	Minimum Android	Released	Last Updated	Content Rating	Ad Supported	In App Purchases
HTTrack Website Copier	com.httrack.android	Communication	3.6	2848	100,000+	100000	351560	0	USD	2.7M	2.3 and up	12-Aug-13	20-May-17	Everyone	FALSE	FALSE
World War 2: Offline Strategy	com.skizze.wwii	Strategy	4.3	17297	1,000,000+	1000000	2161778	0	USD	86M	5.1 and up	19-Jul-18	26-Nov-20	Everyone 10+	TRUE	TRUE
WPSApp	com.themausoft.wpsapp	Tools	4.2	488639	50,000,000+	50000000	79304739	0	USD	5.8M	4.1 and up	7-Mar-16	21-Oct-20	Everyone	TRUE	FALSE
OfficeSuite - Office, PDF, Word, Ex	com.mobisystems.office	Business	4.2	1224420	100,000,000+	100000000	163660067	0	USD	59M	4.4 and up	22-Dec-11	23-Nov-20	Everyone	TRUE	TRUE
Loud Player Free	com.arthelion.loudplayer	Music & Audio	4.2	665	50,000+	50000	73463	0	USD	29M	5.0 and up	24-Sep-16	22-Nov-20	Everyone	FALSE	FALSE
eCharge+	com.innogy.echarge	Maps & Navigation	3.5	377	10,000+	10000	38029	0	USD	12M	6.0 and up	4-Feb-19	6-Nov-20	Everyone	FALSE	FALSE
Jobonji: hire local pros handyman	com.pack.jobonji	Lifestyle	4.4	3346	100,000+	100000	224897	0	USD	9.4M	5.0 and up	10-Jan-18	30-Sep-19	Everyone	FALSE	TRUE
Little Pandaâ€™s Dream Town	com.sinyee.babybus.villa	Educational	4	44700	10,000,000+	10000000	29969311	0	USD	90M	4.2 and up	16-Aug-18	29-Sep-20	Everyone	TRUE	TRUE
Baby Panda: Dental Care	com.sinyee.babybus.den	Educational	4.1	10990	10,000,000+	10000000	12520805	0	USD	87M	4.2 and up	27-Apr-20	29-Sep-20	Everyone	TRUE	TRUE
Excel Reader	com.xcelfilereaderMS	Tools	3.3	1141	100,000+	100000	193274	0	USD	3.3M	5.0 and up	13-Jan-20	25-Jun-20	Everyone	TRUE	FALSE
AndroXLS editor for XLS sheets	com.entertain.androxls	Productivity	3	8831	5,000,000+	5000000	6286237	0	USD	8.4M	4.1 and up	19-Dec-16	21-Oct-20	Everyone	TRUE	FALSE
Office HD: Presentations BASIC	softmaker.applications.p	Business	3.1	768	100,000+	100000	180332	0	USD	21M	4.0 and up	5-Oct-15	27-Jun-17	Everyone	FALSE	TRUE
ã,¡ãƒ¼ãƒ3ãƒ°ã,¯ã,¬ã,µã,¤ãƒ¯ï¼šæ„›ã	jp.co.excite.woman.topic	News & Magazines	3.2	200	50,000+	50000	94585	0	USD	16M	4.4 and up	20-May-12	24-Nov-20	Everyone	TRUE	TRUE
Google Slides	com.google.android.apps	Productivity	4.2	448679	500,000,000+	500000000	800199122	0	USD	Varies with device	6.0 and up	25-Jun-14	24-Nov-20	Everyone	FALSE	FALSE
Grab - Transport, Food Delivery, P	com.grabtaxi.passenger	Maps & Navigation	4.4	6088890	100,000,000+	100000000	165573138	0	USD	153M	5.0 and up	30-May-13	23-Nov-20	Everyone	FALSE	FALSE
Gofood - Food delivery solution b	com.customer.gofood	Food & Drink	2.8	192	50,000+	50000	50401	0	USD	26M	5.0 and up	18-Jun-20	16-Oct-20	Everyone	FALSE	FALSE
Truecaller: Caller ID, block fraud &	com.truecaller	Communication	4.4	14074045	500,000,000+	500000000	681961526	0	USD	87M	5.1 and up	31-May-12	23-Nov-20	Everyone	TRUE	TRUE
Snapseed	com.niksoftware.snapsee	Photography	4.5	1281438	100,000,000+	100000000	181354482	0	USD	Varies with device	Varies with device	6-Dec-12	14-Apr-20	Everyone	FALSE	FALSE
3D Digital Clock Live Wallpaper-7	style_7.a3ddigitalclock_7	Tools	4.3	574	100,000+	100000	148146	0	USD	1.5M	4.0 and up	16-Jan-19	2-Jan-20	Everyone	TRUE	FALSE
Internet speed test Meter- Speed	com.internet.speedtest.c	Tools	4.9	329074	10,000,000+	10000000	13161972	0	USD	8.1M	5.0 and up	16-Jan-19	16-Nov-20	Teen	TRUE	TRUE
Loud Player - Audio player	com.arthelion.loudplayer	Music & Audio	4.1	64	500+	500	688	1.49	USD	26M	Varies with device	24-Sep-16	22-Nov-20	Everyone	FALSE	FALSE
SoundSeeder -Play music simultar	com.kattwinkel.android.s	Music & Audio	3.7	7056	1,000,000+	1000000	1801574	0	USD	20M	4.1 and up	8-Jun-13	6-Jul-20	Everyone	TRUE	TRUE
Fuel Music Player	music.musica.musicplaye	Music & Audio	4.4	589	100,000+	100000	227988	0	USD	12M	5.0 and up	22-Jan-19	16-Nov-20	Everyone	TRUE	TRUE
TempWorks	com.employeecompanion	Business	3.9	40	10,000+	10000	20514	0	USD	46M	4.1 and up	26-Jan-18	19-Jun-20	Everyone	FALSE	FALSE
Qwick Professionals	com.qwickmobile	Lifestyle	4.1	254	10,000+	10000	19909	0	USD	52M	5.0 and up	26-Apr-19	17-Sep-20	Everyone	FALSE	FALSE
LibreOffice & OpenOffice docume	at.tomtasche.reader	Productivity	3.9	31881	5,000,000+	5000000	6709911	0	USD	Varies with device	Varies with device	5-Oct-10	14-Nov-20	Everyone	TRUE	TRUE
All PDF - PDF Reader, PDF Viewer	tz.co.wadau.allpdf	Books & Reference	4.2	9281	500,000+	500000	904096	0	USD	Varies with device	5.0 and up	24-Feb-18	16-Oct-20	Everyone	TRUE	TRUE
Docs To Goâ„¢ Free Office Suite	com.dataviz.docstogo	Business	4.2	226610	50,000,000+	50000000	70125112	0	USD	Varies with device	Varies with device		14-Nov-20	Everyone	TRUE	TRUE
Zoho Sign - Upload, Scan and Sign	com.zoho.sign.zohosign	Business	4	63	10,000+	10000	26306	0	USD	31M	5.0 and up	12-Jul-17	23-Nov-20	Everyone	FALSE	TRUE
Office Delve - for Office 365	com.microsoft.delvemob	Productivity	4	1510	500,000+	500000	741523	0	USD	13M	5.0 and up	13-Apr-15	1-Sep-20	Everyone	FALSE	FALSE
Xodo PDF Reader & Editor	com.xodo.pdf.reader	Productivity	4.7	319893	10,000,000+	10000000	16603209	0	USD	Varies with device	4.1 and up	20-Jan-14	9-Nov-20	Everyone	FALSE	FALSE
MEGA	mega.privacy.android.ap	Productivity	3.7	755515	100,000,000+	100000000	104740945	0	USD	Varies with device	Varies with device	8-Feb-16	25-Nov-20	Everyone	FALSE	TRUE
Adobe Scan: PDF Scanner with OCI	com.adobe.scan.android	Business	4.7	1128455	50,000,000+	50000000	56777956	0	USD	Varies with device	Varies with device	31-May-17	7-Nov-20	Everyone	FALSE	TRUE
PDF Reader for Android 2020	com.ivilamobie.pdfreade	Tools	4.4	70313	5,000,000+	5000000	9307179	0	USD	18M	4.4 and up	23-Dec-19	24-Nov-20	Everyone	TRUE	TRUE
Cute	com.contextlogic.cute	Shopping	4.5	267438	10,000,000+	10000000	13614116	0	USD	27M	5.0 and up	11-May-15	28-Oct-20	Everyone	FALSE	FALSE
Wish - Shopping Made Fun	com.contextlogic.wish	Shopping	4.4	10810840	500,000,000+	500000000	524216973	0	USD	23M	5.0 and up	26-Jul-12	20-Nov-20	Teen	FALSE	FALSE
QR & Barcode Reader	com.teacapps.barcodesca	Productivity	4.5	335907	50,000,000+	50000000	76427891	0	USD	5.1M	6.0 and up	17-Jun-16	22-Nov-20	Everyone	TRUE	TRUE
Memory Match-7	com.style7.memorymatch	Board	4	7	1,000+	1000	2530	0	USD	8.5M	2.3 and up	21-Nov-16	21-Nov-16	Everyone	TRUE	FALSE
Image from Video Grabber-7	style_7.imagefromvideog	Tools	0	0	100+	100	222	0	USD	1.2M	4.0 and up	12-Feb-20	12-Feb-20	Everyone	TRUE	FALSE
Google Text-to-Speech	com.google.android.tts	Tools	4.3	2292232	5,000,000,000+	5000000000	7769956479	0	USD	Varies with device	Varies with device	10-Oct-13	15-Oct-20	Everyone	FALSE	FALSE
Android TV Home	com.google.android.tvlau	Tools	1.1	226	10,000,000+	10000000	39240688	0	USD	5.6M	8.0 and up	8-Jan-18	19-Nov-20	Everyone	TRUE	FALSE
Qwick Business	com.qwickmobilebusines	Lifestyle	0	0	100+	100	168	0	USD	52M	5.0 and up	19-Dec-19	17-Sep-20	Everyone	FALSE	FALSE
Streamline3 Admin	com.ubtsupport.streamlir	Business	4.7	31	1,000+	1000	3629	0	USD	23M	6.0 and up	18-Dec-18	23-Nov-20	Everyone	FALSE	FALSE
N Docs - Office. PDF. Text. Markup	net.siava.docs	Productivity	4.3	1265	100,000+	100000	113425	0	USD	64M	7.0 and up	20-Jul-17	25-Nov-20	Everyone	FALSE	TRUE

Figure 9: Data after column removal

With only relevant variables remaining, filters were added to the columns. First and foremost, all rows containing blanks were filtered out. Specifically, there were blanks in the Category,

Rating, Minimum Android Version, and Release Date columns. Here is an example of the Excel filter menu for the Rating column:

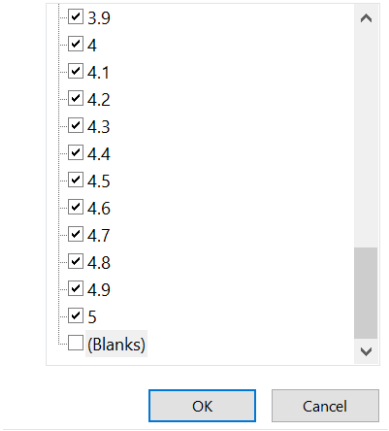


Figure 10: Select relevant results

Next, the Category column was filtered to include only apps that are games. The games are categorized by their genre, which include Action, Adventure, Arcade, Board, Card, Casino, Casual, Educational, Music, Puzzle, Racing, Role Playing, Simulation, Strategy, Trivia, and Word.

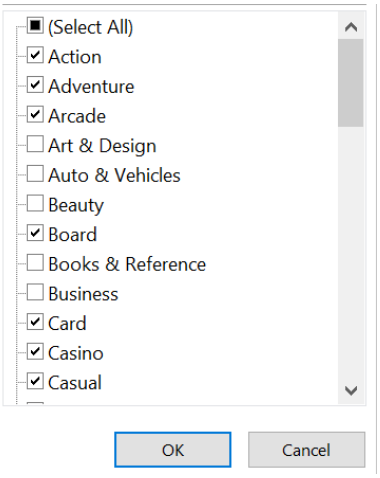


Figure 11: Filter to chosen categories

Another column that needed to be filtered was the Currency column, which contained the specific international currency that can be used to pay for the game in the store. Any games that had a currency other than USD were filtered out.

Additionally, the Size and Minimum Android Version columns contained a "varies with device" option, indicating that some games have different values for these variables depending on the device the game is played on. Rows containing this option were removed.

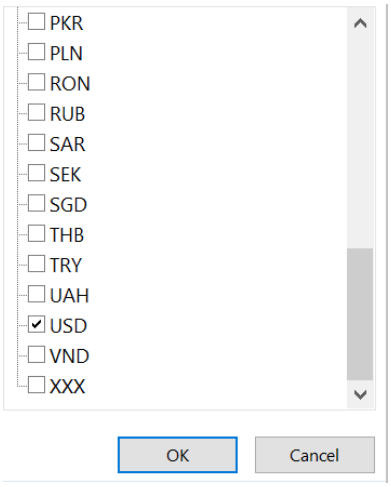


Figure 12: Remove all non-USD currencies

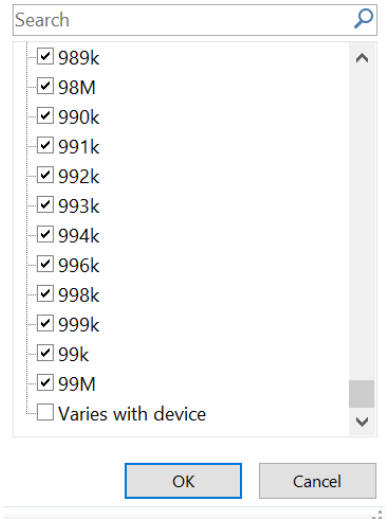
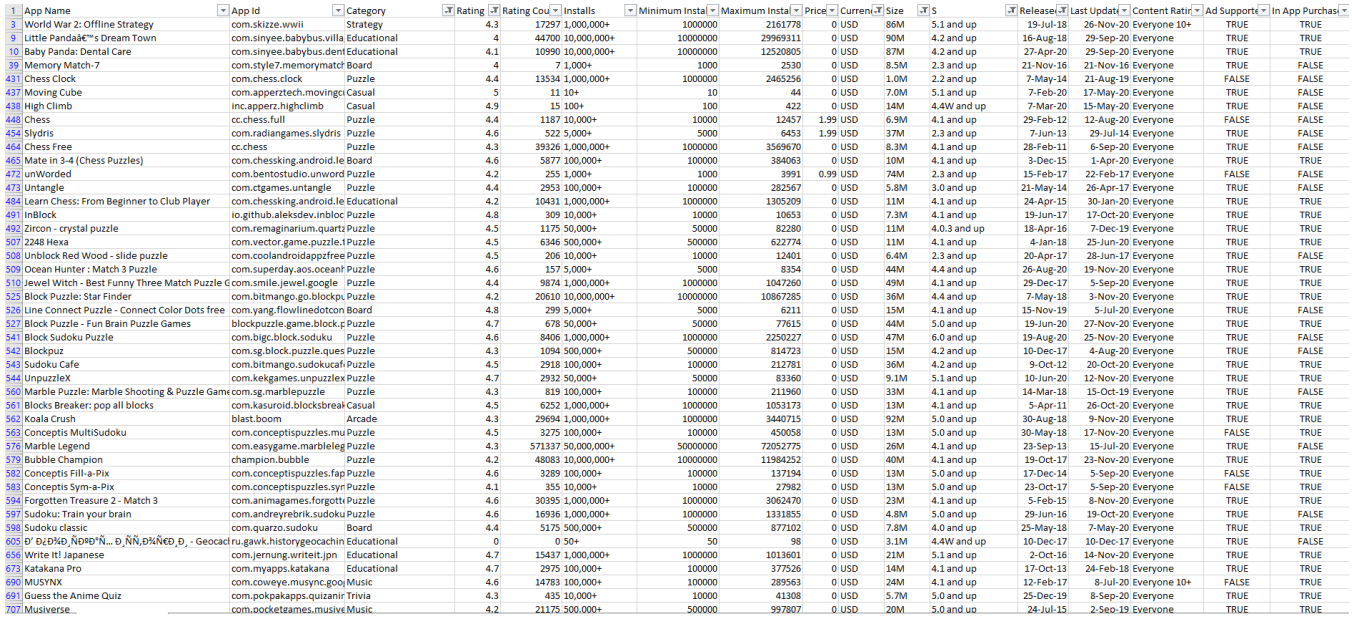


Figure 13: Remove rows with uninformative data

After all these filters were applied, 128622 rows remained in the data.



1	App Name	App Id	Category	Rating	Rating Cou	Installs	Minimum Insta	Maximum Insta	Price	Curren	Size	S	Release	Last Updat	Content Ratir	Ad Supporte	In App Purchas
3	World War 2: Offline Strategy	com.skizze.wwii	Strategy	4.3	17297	1,000,000+	1000000	2161778	0	USD	86M	5.1 and up	19-Jul-18	26-Nov-20	Everyone 10+	TRUE	TRUE
9	Little Pandaâ€™s Dream Town	com.sinyee.babybus.villa	Educational	4	44700	10,000,000+	10000000	29969311	0	USD	90M	4.2 and up	16-Aug-18	29-Sep-20	Everyone	TRUE	TRUE
10	Baby Panda: Dental Care	com.sinyee.babybus.dent	Educational	4.1	10990	10,000,000+	10000000	12520805	0	USD	87M	4.2 and up	27-Apr-20	29-Sep-20	Everyone	TRUE	TRUE
39	Memory Match-7	com.style7.memorymatch	Board	4	7	1,000+	1000	2530	0	USD	8.5M	2.3 and up	21-Nov-16	21-Nov-16	Everyone	TRUE	FALSE
431	Chess Clock	com.chess.clock	Puzzle	4.4	13534	1,000,000+	1000000	2465256	0	USD	1.0M	2.2 and up	7-May-14	21-Aug-19	Everyone	FALSE	FALSE
437	Moving Cube	com.apperztech.movingcu	Casual	5	11	10+	10	44	0	USD	7.0M	5.1 and up	7-Feb-20	17-May-20	Everyone	TRUE	FALSE
438	High Climb	inc.apperz.highclimb	Casual	4.9	15	100+	100	422	0	USD	14M	4.4W and up	7-Mar-20	15-May-20	Everyone	TRUE	FALSE
448	Chess	cc.chess.full	Puzzle	4.4	1187	10,000+	10000	12457	1.99	USD	6.9M	4.1 and up	29-Feb-12	12-Aug-20	Everyone	FALSE	FALSE
454	Slydris	com.radiangames.slydris	Puzzle	4.6	522	5,000+	5000	6453	1.99	USD	37M	2.3 and up	7-Jun-13	29-Jul-14	Everyone	TRUE	FALSE
464	Chess Free	cc.chess	Puzzle	4.3	39326	1,000,000+	1000000	3569670	0	USD	8.3M	4.1 and up	28-Feb-11	6-Sep-20	Everyone	TRUE	FALSE
465	Mate in 3-4 (Chess Puzzles)	com.chessking.android.le	Board	4.6	5877	100,000+	100000	384063	0	USD	10M	4.1 and up	3-Dec-15	1-Apr-20	Everyone	TRUE	TRUE
472	unWorded	com.bentostudio.unword	Puzzle	4.2	255	1,000+	1000	3991	0.99	USD	74M	2.3 and up	15-Feb-17	22-Feb-17	Everyone	FALSE	FALSE
473	Untangle	com.ctgames.untangle	Puzzle	4.4	2953	100,000+	100000	282567	0	USD	5.8M	3.0 and up	21-May-14	26-Apr-17	Everyone	TRUE	FALSE
484	Learn Chess: From Beginner to Club Player	com.chessking.android.le	Educational	4.2	10431	1,000,000+	1000000	1305209	0	USD	11M	4.1 and up	24-Apr-15	30-Jan-20	Everyone	TRUE	TRUE
491	InBlock	io.github.aleksdev.inbloc	Puzzle	4.8	309	10,000+	10000	10653	0	USD	7.3M	4.1 and up	19-Jun-17	17-Oct-20	Everyone	TRUE	TRUE
492	Zircon - crystal puzzle	com.remaginarium.quartz	Puzzle	4.5	1175	50,000+	50000	82280	0	USD	11M	4.0.3 and up	18-Apr-16	7-Dec-19	Everyone	TRUE	TRUE
507	2248 Hexa	com.vector.game.puzzle.t	Puzzle	4.5	6346	500,000+	500000	622774	0	USD	11M	4.1 and up	4-Jan-18	25-Jun-20	Everyone	TRUE	TRUE
508	Unblock Red Wood - slide puzzle	com.coolandroidappzfree	Puzzle	4.5	206	10,000+	10000	12401	0	USD	6.4M	2.3 and up	20-Apr-17	28-Jun-17	Everyone	TRUE	FALSE
509	Ocean Hunter : Match 3 Puzzle	com.superday.aos.oceanh	Puzzle	4.6	157	5,000+	5000	8354	0	USD	44M	4.4 and up	26-Aug-20	19-Nov-20	Everyone	TRUE	TRUE
510	Jewel Witch - Best Funny Three Match Puzzle G	com.smile.jewel.google	Puzzle	4.4	9874	1,000,000+	1000000	1047260	0	USD	49M	4.1 and up	29-Dec-17	5-Sep-20	Everyone	TRUE	TRUE
525	Block Puzzle: Star Finder	com.bitmango.go.blockpu	Puzzle	4.2	20610	10,000,000+	10000000	10867285	0	USD	36M	4.4 and up	7-May-18	3-Nov-20	Everyone	TRUE	TRUE
526	Line Connect Puzzle - Connect Color Dots free	com.yang.flowlinedotcon	Board	4.8	299	5,000+	5000	6211	0	USD	15M	4.1 and up	15-Nov-19	5-Jul-20	Everyone	TRUE	FALSE
527	Block Puzzle - Fun Brain Puzzle Games	blockpuzzle.game.block.p	Puzzle	4.7	678	50,000+	50000	77615	0	USD	44M	5.0 and up	19-Jun-20	27-Nov-20	Everyone	TRUE	TRUE
541	Block Sudoku Puzzle	com.bigc.block.soduku	Puzzle	4.6	8406	1,000,000+	1000000	2250227	0	USD	47M	6.0 and up	19-Aug-20	25-Nov-20	Everyone	TRUE	FALSE
542	Blockpuz	com.sg.block.puzzle.ques	Puzzle	4.3	1094	500,000+	500000	814723	0	USD	15M	4.2 and up	10-Dec-17	4-Aug-20	Everyone	TRUE	FALSE
543	Sudoku Cafe	com.bitmango.sudokucafe	Puzzle	4.5	2918	100,000+	100000	212781	0	USD	36M	4.2 and up	9-Oct-12	20-Oct-20	Everyone	TRUE	TRUE
544	UnpuzzleX	com.kekgames.unpuzzlex	Puzzle	4.7	2932	50,000+	50000	83360	0	USD	9.1M	5.1 and up	10-Jun-20	12-Nov-20	Everyone	TRUE	TRUE
560	Marble Puzzle: Marble Shooting & Puzzle Game	com.sg.marblepuzzle	Puzzle	4.3	819	100,000+	100000	211960	0	USD	33M	4.1 and up	14-Mar-18	15-Oct-19	Everyone	TRUE	FALSE
561	Blocks Breaker: pop all blocks	com.kasuroid.blocksbreak	Casual	4.5	6252	1,000,000+	1000000	1053173	0	USD	13M	4.1 and up	5-Apr-11	26-Oct-20	Everyone	TRUE	TRUE
562	Koala Crush	blast.boom	Arcade	4.3	29694	1,000,000+	1000000	3440715	0	USD	92M	5.0 and up	30-Aug-18	9-Nov-20	Everyone	TRUE	TRUE
563	Conceptis MultiSudoku	com.conceptispuzzles.mu	Puzzle	4.5	3275	100,000+	100000	450058	0	USD	13M	5.0 and up	30-May-18	17-Nov-20	Everyone	FALSE	TRUE
576	Marble Legend	com.easygame.marbleleg	Puzzle	4.3	571337	50,000,000+	50000000	72052775	0	USD	26M	4.1 and up	23-Sep-13	15-Jul-20	Everyone	TRUE	FALSE
579	Bubble Champion	champion.bubble	Puzzle	4.2	48083	10,000,000+	10000000	11984252	0	USD	40M	4.1 and up	19-Oct-17	23-Nov-20	Everyone	TRUE	TRUE
582	Conceptis Fill-a-Pix	com.conceptispuzzles.fap	Puzzle	4.6	3289	100,000+	100000	137194	0	USD	13M	5.0 and up	17-Dec-14	5-Sep-20	Everyone	FALSE	TRUE
583	Conceptis Sym-a-Pix	com.conceptispuzzles.syn	Puzzle	4.1	355	10,000+	10000	27982	0	USD	13M	5.0 and up	23-Oct-17	5-Sep-20	Everyone	FALSE	TRUE
594	Forgotten Treasure 2 - Match 3	com.animagames.forgotte	Puzzle	4.6	30395	1,000,000+	1000000	3062470	0	USD	23M	4.1 and up	5-Feb-15	8-Nov-20	Everyone	TRUE	TRUE
597	Sudoku: Train your brain	com.andreyrebrik.sudoku	Puzzle	4.6	16936	1,000,000+	1000000	1331855	0	USD	4.8M	5.0 and up	29-Jun-16	19-Oct-20	Everyone	TRUE	FALSE
598	Sudoku classic	com.quarzo.sudoku	Board	4.4	5175	500,000+	500000	877102	0	USD	7.8M	4.0 and up	25-May-18	7-May-20	Everyone	TRUE	TRUE
605	Ð' Ð¿Ð¾Ð¸ÑˆÐ°Ñ… Ð¸ÑÑ‚Ð¾Ñ€Ð¸Ð¸ - Geocach	ru.gawk.historygeocachin	Educational	0	0	50+	50	98	0	USD	3.1M	4.4W and up	10-Dec-17	10-Dec-17	Everyone	TRUE	FALSE
656	Write it! Japanese	com.jernung.writeit.jpn	Educational	4.7	15437	1,000,000+	1000000	1013601	0	USD	21M	5.1 and up	2-Oct-16	14-Nov-20	Everyone	TRUE	TRUE
673	Katakana Pro	com.myapps.katakana	Educational	4.7	2975	100,000+	100000	377526	0	USD	14M	4.1 and up	17-Oct-13	24-Feb-18	Everyone	TRUE	TRUE
690	MUSYNX	com.coweye.musync.goo	Music	4.6	14783	100,000+	100000	289563	0	USD	24M	4.1 and up	12-Feb-17	8-Jul-20	Everyone 10+	FALSE	TRUE
691	Guess the Anime Quiz	com.pokpakapps.quizanir	Trivia	4.3	435	10,000+	10000	41308	0	USD	5.7M	5.0 and up	25-Dec-19	8-Sep-20	Everyone	TRUE	TRUE
707	Musiverse	com.pocketgames.musive	Music	4.2	21175	500,000+	500000	997807	0	USD	20M	5.0 and up	24-Jul-15	2-Sep-19	Everyone	TRUE	TRUE

Figure 14: Remaining data

Appendix C: Analytics details

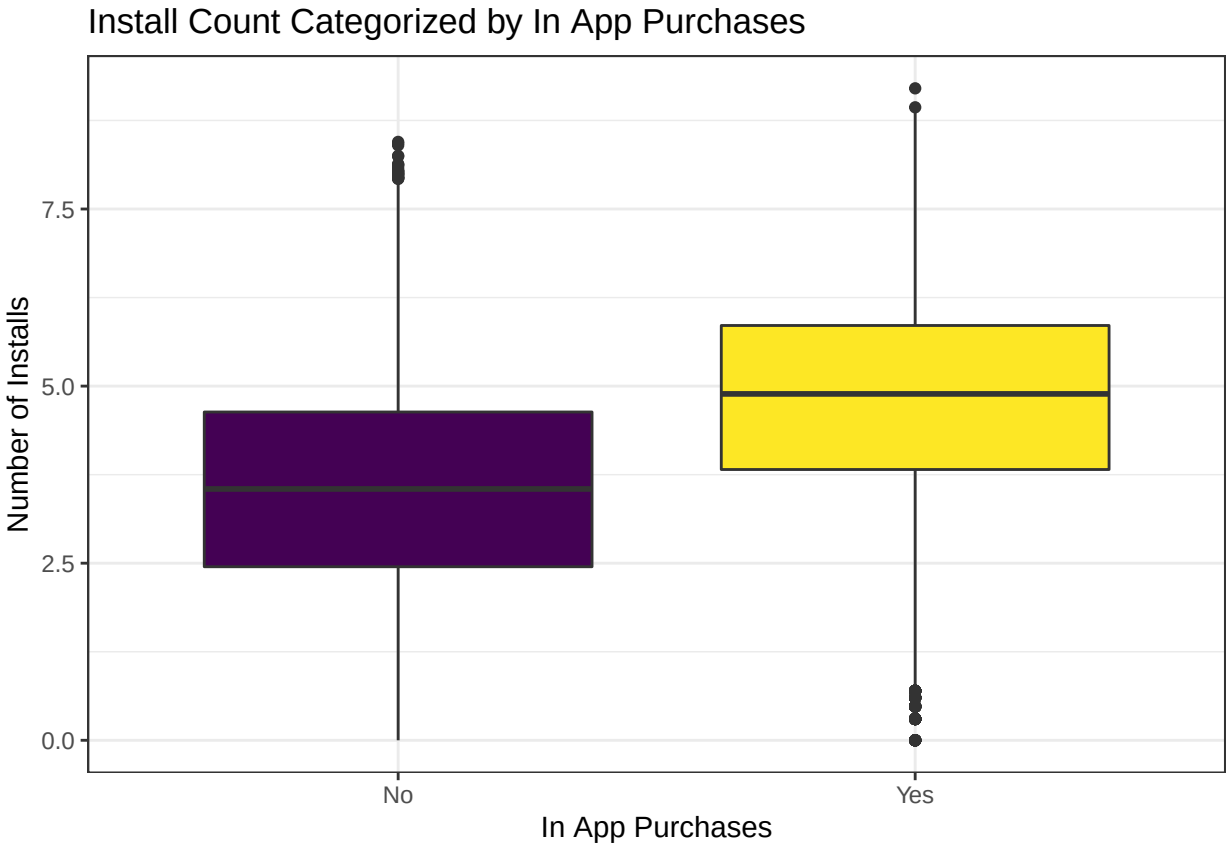
Descriptive analytics

First, the data must be mutate the install count by the log10 transformation. This will scale the boxplot so that it is readable as there is an observation greater than one billion and many observations close to or equal to zero.

```
apps.log <- apps %>%
  mutate(Maximum.Installs = log10(apps$Maximum.Installs))
```

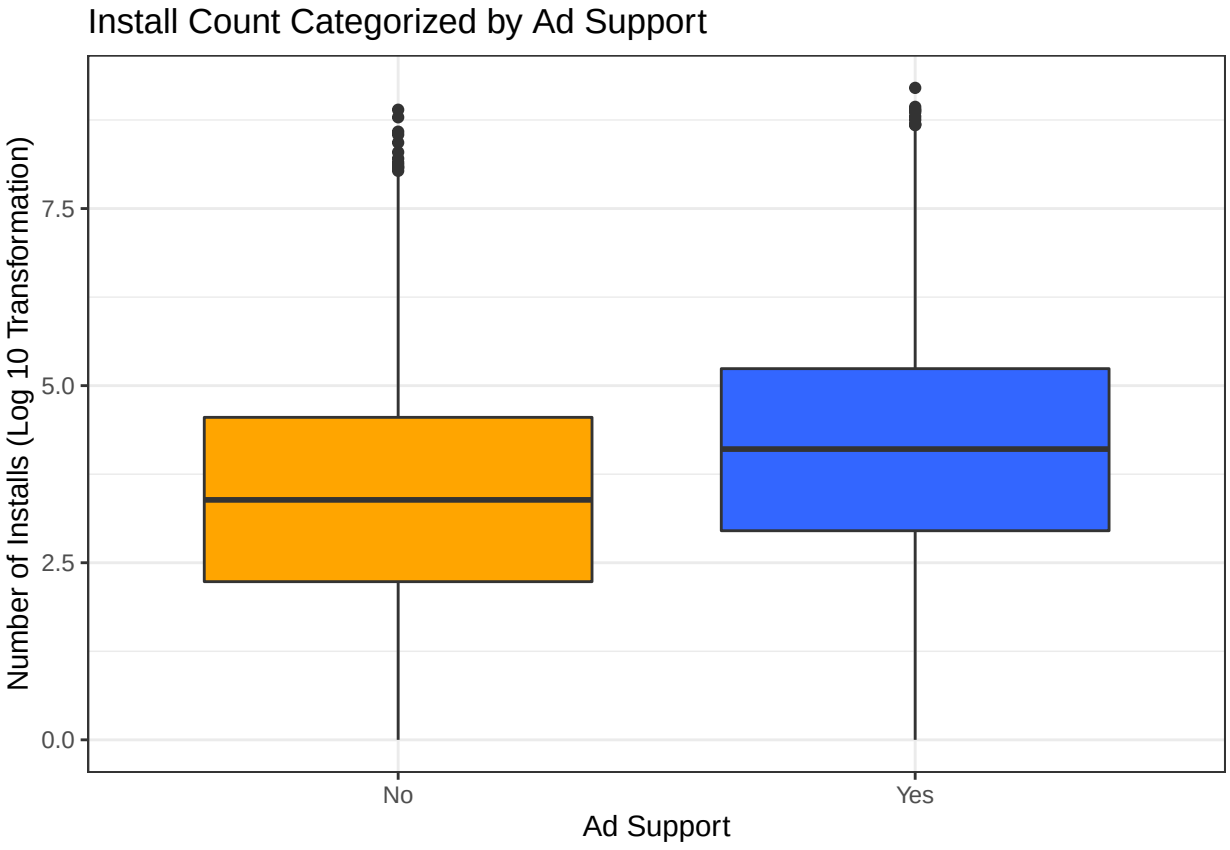
Here we create the boxplot. First, place the boolean in app purchases column on the x-axis and the install count (scaled by the log10 transformation) on the y-axis. Next, set the fill of the boxplot equal to the in-app purchases value, and alter aesthetic appearances like the title, the x and y-axis labels, and the theme. Then, change "FALSE" and "TRUE" x-axis ticks to "No" and "Yes" respectively.

```
apps.log %>%
  ggplot(aes(x = In.App.Purchases, y = Maximum.Installs,
             fill = In.App.Purchases)) + geom_boxplot() +
  labs(title = "Install Count Categorized by In App Purchases") +
  xlab("In App Purchases") +
  ylab("Number of Installs") +
  theme_bw() + theme(legend.position = "none") +
  scale_x_discrete(labels = c("FALSE" = "No", "TRUE" = "Yes")) +
  scale_fill_viridis_d()
```


Install Count Categorized by In App Purchases
Number of Installs
7.5
5.0
2.5
0.0
No
Yes
In App Purchases

For the ad support boxplot, place the boolean ad supported column on the x-axis and the install count (scaled by the log10 transformation) on the y-axis. Then, set the fill of the boxplot equal to the in app purchases value, and alter aesthetic appearances like the title, the x and y-axis labels, and the theme. Finally, change "FALSE" and "TRUE" x-axis ticks to "No" and "Yes" respectively.

```
apps.log %>%
  ggplot(aes(x = Ad.Supported, y = Maximum.Installs,
             fill = Ad.Supported)) +
  geom_boxplot(fill = c("orange", "#3366FF")) +
  labs(title = "Install Count Categorized by Ad Support") +
  xlab("Ad Support") + ylab("Number of Installs (Log 10 Transformation)") +
  theme_bw() + theme(legend.position = "none") +
  scale_x_discrete(labels = c("FALSE" = "No", "TRUE" = "Yes"))
```



In order to create a meaningful plot of the install count by category, we must first find the median of the install count for each of the categories.

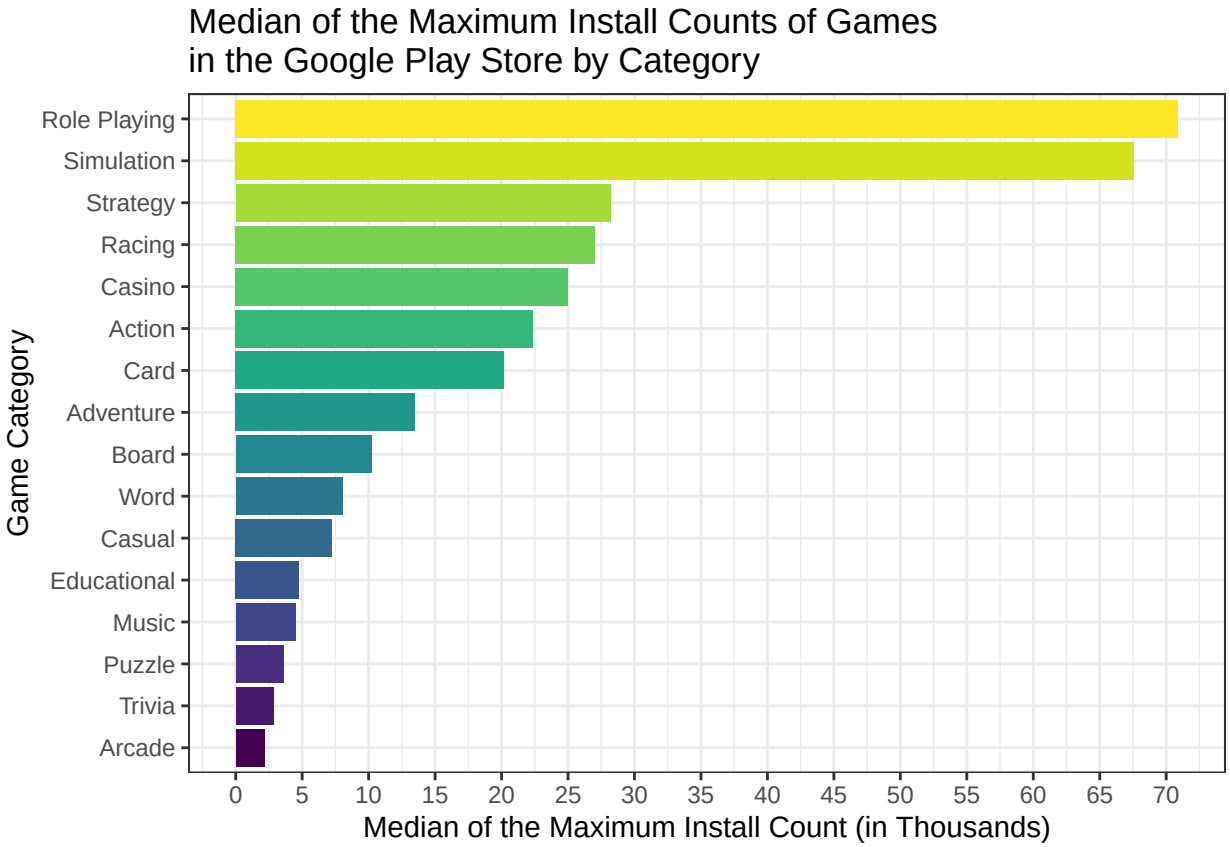
```
apps.summary <- apps %>%
  select(Maximum.Installs) %>%
  aggregate(by = list(apps$Category), median)
```

The columns are then renamed and the median for each category is then scaled by dividing by 1,000.

```
apps.summary <- apps.summary %>%
  mutate(Category = Group.1,
         median = Maximum.Installs / 1000) %>%
  select(Category, median)
```

To create the plot, change aesthetic information such as the theme, legend position, x-axis labels, y-axis labels, and fill color.

```
ggplot(apps.summary, aes(x = reorder(Category, median), y = median)) +
  geom_bar(stat = "identity", aes(fill = reorder(Category, median))) +
  theme_bw() + theme(legend.position='none') +
  labs(x="Game Category",
       y="Median of the Maximum Install Count (in Thousands)",
       title="Median of the Maximum Install Counts of Games \nin the Google Play Store b
  scale_y_continuous(breaks=seq(0,90,5)) +
  coord_flip() + scale_fill_viridis_d()
```


Median of the Maximum Install Counts of Games
in the Google Play Store by Category
Game Category
Role Playing
Simulation
Strategy
Racing
Casino
Action
Card
Adventure
Board
Word
Casual
Educational
Music
Puzzle
Trivia
Arcade
0
5
10
15
20
25
30
35
40
45
50
55
60
65
70
Median of the Maximum Install Count (in Thousands)

To display a scatterplot of all game download sizes, the values need to be type "character."

```
apps.size <- apps
apps.size$Size <- as.character(apps$Size)
```

Each of the values in the size column is measured in kilobytes, megabytes, or gigabytes. The unit abbreviation must be removed, and the units must be standardized.

```
for (i in 1:nrow(apps.size)){
  if (str_contains(apps.size[i,11],"k")){
    apps.size[i, 11] <- str_sub(apps.size[i, 11], end = -2)
    apps.size[i, 11] <- gsub(",", "", apps.size[i, 11])
    apps.size[i, 11] <- as.numeric(apps.size[i, 11]) / 1000
  }
  else if (str_contains(apps.size[i,11],"G")){
    apps.size[i, 11] <- str_sub(apps.size[i, 11], end = -2)
    apps.size[i, 11] <- gsub(",", "", apps.size[i, 11])
    apps.size[i, 11] <- as.numeric(apps.size[i, 11]) * 1000
  }
  else{
    apps.size[i, 11] <- str_sub(apps.size[i, 11], end = -2)
    apps.size[i, 11] <- gsub(",", "", apps.size[i, 11])
    apps.size[i, 11] <- as.numeric(apps.size[i, 11])
  }
}
```

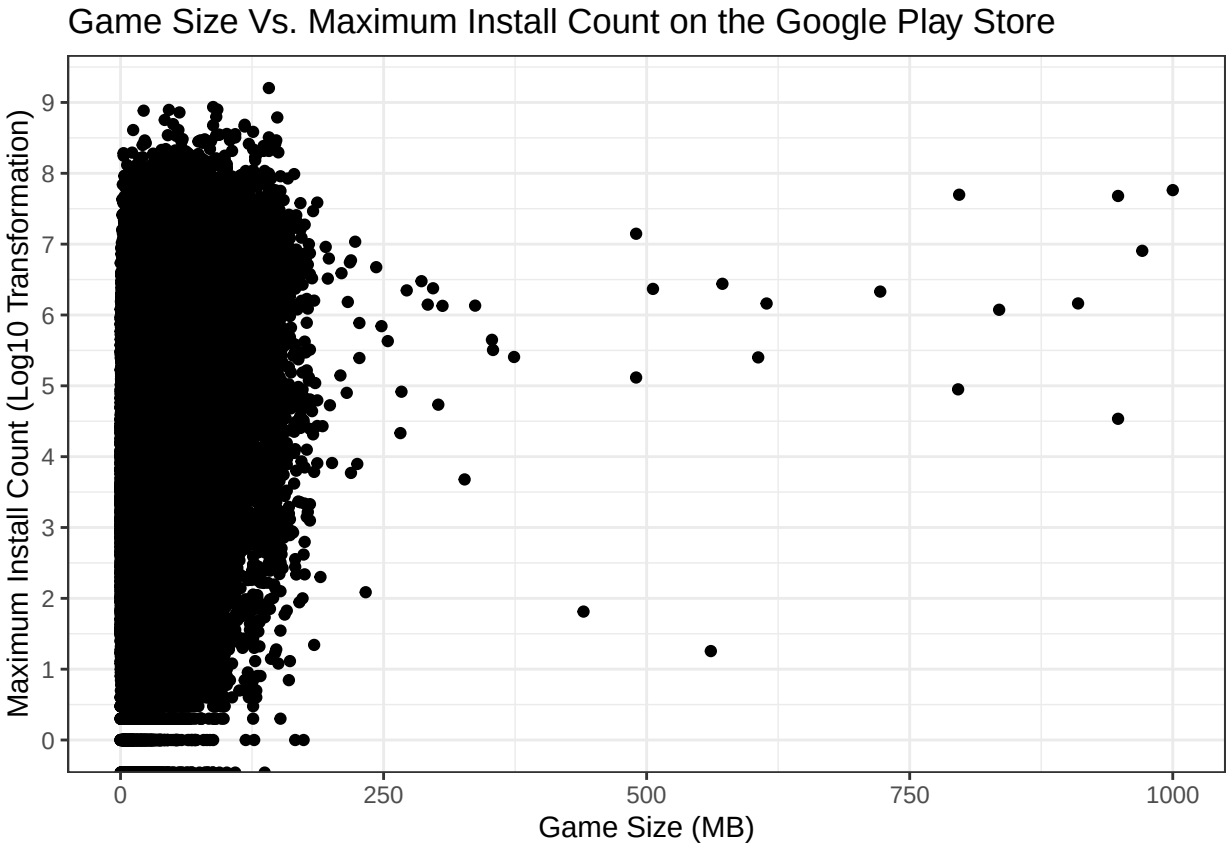
The app size values then need to be converted from "character" type to "numeric" type.

```
apps.size$Size<-as.numeric(apps.size$Size)
```

The scatterplot of size values is now able to be constructed.

```
ggplot(apps.size,aes(x = Size, y = log10(Maximum.Installs))) +
  geom_point() + theme(axis.title.y = element_text(vjust = 3),
        plot.title = element_text(hjust = 0.5))+
  labs(x = "Game Size (MB)",
```

```
      y = "Maximum Install Count (Log10 Transformation)",
      title="Game Size Vs. Maximum Install Count on the Google Play Store") +
scale_y_continuous(breaks = seq(0, 10, 1)) + theme_bw()
```



For the table of game categories with percentages representing the number of games with in-app purchases or ad support, first the data frame is copied.

```
apps.table <- apps
```

Next, a column is created with a numeric binary representation of the ad support variable.

```
apps.table$Ads<-ifelse(apps.table$Ad.Supported == 'TRUE', 1, 0)
```

Next, the number of installations per category that have ads is calculated.

```
apps.table$Ad.Games <- apps.table$Ads * apps.table$Maximum.Installs
```

The process above is then repeated for the in-app purchases variable.

```
apps.table$IAP <- ifelse(apps.table$In.App.Purchases == 'TRUE', 1, 0)
```

```
apps.table$IAP.Games <- apps.table$IAP * apps.table$Maximum.Installs
```

The table is then created by grouping games by their category and summarizing the relevant data.

```
tbl<-apps.table %>% group_by(Category) %>%
  summarise(`Ad Supported` = percent(sum(Ad.Games) / sum(Maximum.Installs),
                                     .01),
            `In App Purchases`=percent(sum(IAP.Games)/sum(Maximum.Installs),
                                       .01))
```

Finally, the table is wrapped in a kable() function.

```
kable(tbl, caption = 'Revenue Strategy by Category')
```

Table 8: Revenue Strategy by Category

Category	Ad Supported	In App Purchases
Action	90.06%	89.07%
Adventure	81.26%	77.33%
Arcade	97.38%	80.86%
Board	96.54%	69.51%
Card	89.09%	62.44%
Casino	82.09%	91.34%
Casual	96.78%	72.93%
Educational	90.02%	74.14%
Music	94.75%	79.81%
Puzzle	96.32%	74.06%
Racing	96.83%	84.15%
Role Playing	79.64%	89.12%
Simulation	95.42%	71.08%
Strategy	55.15%	93.49%
Trivia	97.43%	74.50%
Word	98.16%	84.52%

Predictive analytics

9 Predictive Analytics Details

Load necessary packages.

```
library(tidyverse)
library(viridis)
library(knitr)
library(stringr)
library(sjmisc)
library(scales)
library(tidymodels)
library(knitr)
library(modelr)
library(car)
library(rlang)
library(rpart)
library(rpart.plot)
```

```
library(randomForest)
library(kableExtra)
```

9.1 Logistic Regression

Reading data, setting binary success variable, and removing irrelevant columns.

```
apps <- read_csv("Google-Playstore (1).csv")

dropcol<-c('Free','Developer.Id','Developer.Website','Developer.Email',
           'Privacy.Policy','Editors.Choice', 'Installs', 'Minimum.Installs',
           'App.Name', 'App.Id', 'Rating', 'Rating.Count')

keepcat<-c('Action','Adventure','Arcade','Board','Card','Casino','Casual',
           'Educational','Music','Puzzle','Racing','Role Playing','Simulation',
           'Strategy','Trivia','Word') # vectorizing the categories to keep

apps <- apps %>%
  select(!all_of(dropcol)) %>%
  filter(Category %in% keepcat) %>%
  drop_na() %>%
  filter(Currency == "USD") %>%
  filter(Size != "Varies with device",
         Minimum.Android != "Varies with device") %>%
  filter(Released != "", Minimum.Android != "")

apps <- apps %>%
  mutate(success = as.numeric(Maximum.Installs>140167)) %>%
  drop_na()
```

Balancing the data so there's an equal number of successful games and unsuccessful games by under-sampling the unsuccessful games.

```
set.seed(1)

play1 <- apps[apps$success == 1,]
```

```
play0 <- apps[apps$success == 0,]
play <- rbind(play1[sample(1:nrow(play1),500,replace=F),],
              play0[sample(1:nrow(play0),500,replace=F),])
```

Adjusting size variable to be numeric. Some size measurements were in megabytes (M) and others were in kilobytes (k). Depending on the unit, the appropriate number of zeroes were added to put the number in bytes.

```
size_format <- function(string){
  string <- gsub(",","", string)
  if (str_detect(string, "M")){
    pd <- sub("\\.", "", string)
    sub("M", "000", pd)
  }else{
    sub("k", "", string)
  }
}

play$Size <- play$Size %>%
  sapply(function(x) size_format(x))

play$Size <- as.numeric(play$Size)
```

Convert the minimum Android version to a number floored to the tenths place.

```
version_format <- function(string){
  num <- substr(string, 1, 3)
  as.numeric(num)
}

play$Minimum.Android <- play$Minimum.Android %>%
  sapply(function(x) version_format(x))
```

Convert variables to factors.

```
play$Category <- as.factor(play$Category)
```

```
play$Content.Rating <- as.factor(play$Content.Rating)
play$success <- as.logical(play$success)

str(play)

## tibble [1,000 x 12] (S3: tbl_df/tbl/data.frame)
##  $ Category        : Factor w/ 16 levels "Action","Adventure",..: 10 13 4 10 14 1 2 1
##  $ Maximum.Installs: num [1:1000] 338703 3459869 270723 8876516 2201402 ...
##  $ Price           : num [1:1000] 0 0 0 0 0 0 0 0 0 0 ...
##  $ Currency        : chr [1:1000] "USD" "USD" "USD" "USD" ...
##  $ Size            : num [1:1000] 23000 35000 13000 12000 103000 49000 68000 55000 35
##  $ Minimum.Android : Named num [1:1000] 4.1 4.1 4 4.4 4.4 2.3 4 5 4.4 6 ...
##   ..- attr(*, "names")= chr [1:1000] "4.1 and up" "4.1 and up" "4.0 and up" "4.4 and
##  $ Released        : chr [1:1000] "8-May-17" "30-Aug-17" "27-Mar-15" "8-Sep-13" ...
##  $ Last.Updated    : chr [1:1000] "2-Apr-20" "21-Jun-20" "18-Nov-19" "30-Oct-20" ...
##  $ Content.Rating  : Factor w/ 5 levels "Everyone","Everyone 10+",..: 1 1 1 1 2 4 4 1
##  $ Ad.Supported    : logi [1:1000] TRUE TRUE TRUE TRUE TRUE TRUE ...
##  $ In.App.Purchases: logi [1:1000] TRUE FALSE FALSE FALSE TRUE FALSE ...
##  $ success         : logi [1:1000] TRUE TRUE TRUE TRUE TRUE TRUE ...
```

Generating training, validation, and testing data.

```
set.seed(1)

play.div <- play %>%
  initial_split(prop = 0.6,strata=Category)

play.div2 <- play.div %>%
  testing() %>%
  initial_split(prop=0.5,strata=Category)

play.train <- training(play.div)
play.validate <- training(play.div2)
play.test <- testing(play.div2)
```

Comparing logistic regression models: Create the initial model and view the summary.

```
LogReg.mod1<-glm(success ~ Category + Size + Price + Minimum.Android +
                   Content.Rating + Ad.Supported + In.App.Purchases,
                 data=play.train, family="binomial")
summary(LogReg.mod1)
```

```
##
## Call:
## glm(formula = success ~ Category + Size + Price + Minimum.Android +
##     Content.Rating + Ad.Supported + In.App.Purchases, family = "binomial",
##     data = play.train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.2294  -0.9143  -0.0301   0.9546   2.0380
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                -1.114e+00  7.341e-01  -1.518 0.129018
## CategoryAdventure          -6.883e-01  5.398e-01  -1.275 0.202246
## CategoryArcade             -1.211e+00  4.804e-01  -2.520 0.011744 *
## CategoryBoard              -9.585e-01  6.401e-01  -1.497 0.134281
## CategoryCard               -9.689e-01  5.744e-01  -1.687 0.091605 .
## CategoryCasino             -1.946e+00  9.350e-01  -2.082 0.037364 *
## CategoryCasual             -1.110e+00  4.733e-01  -2.345 0.019030 *
## CategoryEducational        -7.050e-01  5.713e-01  -1.234 0.217161
## CategoryMusic              -7.478e-01  6.780e-01  -1.103 0.270090
## CategoryPuzzle             -1.327e+00  4.693e-01  -2.827 0.004696 **
## CategoryRacing             -2.875e-01  5.771e-01  -0.498 0.618363
## CategoryRole Playing       -8.897e-01  5.653e-01  -1.574 0.115517
## CategorySimulation          1.108e-01  4.856e-01   0.228 0.819571
## CategoryStrategy           -3.076e-01  6.592e-01  -0.467 0.640752
## CategoryTrivia             -1.855e+00  6.819e-01  -2.721 0.006506 **
## CategoryWord               -8.762e-01  6.837e-01  -1.282 0.199988
## Size                        1.042e-05  3.685e-06   2.828 0.004683 **
## Price                      -2.002e+00  1.193e+00  -1.678 0.093256 .
## Minimum.Android            -9.674e-02  1.281e-01  -0.755 0.450300
```

```
## Content.RatingEveryone 10+  5.619e-01  4.739e-01   1.185 0.235832
## Content.RatingMature 17+    1.003e+00  7.349e-01   1.365 0.172147
## Content.RatingTeen          6.777e-01  2.574e-01   2.633 0.008474 **
## Ad.SupportedTRUE            1.213e+00  3.637e-01   3.336 0.000851 ***
## In.App.PurchasesTRUE        1.288e+00  2.027e-01   6.357 2.05e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 831.35  on 599  degrees of freedom
## Residual deviance: 669.14  on 576  degrees of freedom
## AIC: 717.14
##
## Number of Fisher Scoring iterations: 8
```

Create second model and evaluate it against the first model.

```
LogReg.mod2<-glm(success ~ Category + Size + Price +
                   Content.Rating + Ad.Supported + In.App.Purchases,
                 data=play.train, family="binomial")
anova(LogReg.mod2, LogReg.mod1, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: success ~ Category + Size + Price + Content.Rating + Ad.Supported +
##     In.App.Purchases
## Model 2: success ~ Category + Size + Price + Minimum.Android + Content.Rating +
##     Ad.Supported + In.App.Purchases
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       577     669.71
## 2       576     669.14  1  0.56917   0.4506
```

```
summary(LogReg.mod2)
```

```
##
```

```
## Call:
## glm(formula = success ~ Category + Size + Price + Content.Rating +
##     Ad.Supported + In.App.Purchases, family = "binomial", data = play.train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.1700  -0.9244  -0.0318   0.9613   2.0151
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 -1.459e+00  5.754e-01  -2.535 0.011231 *
## CategoryAdventure           -6.844e-01  5.387e-01  -1.270 0.203944
## CategoryArcade              -1.202e+00  4.791e-01  -2.508 0.012136 *
## CategoryBoard               -9.590e-01  6.390e-01  -1.501 0.133445
## CategoryCard                -9.734e-01  5.728e-01  -1.699 0.089250 .
## CategoryCasino              -1.929e+00  9.357e-01  -2.062 0.039241 *
## CategoryCasual              -1.097e+00  4.718e-01  -2.326 0.020030 *
## CategoryEducational         -6.805e-01  5.688e-01  -1.197 0.231499
## CategoryMusic               -7.386e-01  6.767e-01  -1.091 0.275108
## CategoryPuzzle              -1.313e+00  4.677e-01  -2.808 0.004980 **
## CategoryRacing              -2.680e-01  5.756e-01  -0.466 0.641550
## CategoryRole Playing        -8.881e-01  5.638e-01  -1.575 0.115199
## CategorySimulation           1.263e-01  4.838e-01   0.261 0.794046
## CategoryStrategy            -2.814e-01  6.581e-01  -0.428 0.668873
## CategoryTrivia              -1.859e+00  6.805e-01  -2.731 0.006306 **
## CategoryWord                -8.795e-01  6.832e-01  -1.287 0.197995
## Size                         9.867e-06  3.602e-06   2.740 0.006151 **
## Price                       -1.997e+00  1.185e+00  -1.686 0.091883 .
## Content.RatingEveryone 10+   5.733e-01  4.745e-01   1.208 0.226961
## Content.RatingMature 17+     9.877e-01  7.328e-01   1.348 0.177725
## Content.RatingTeen           6.692e-01  2.572e-01   2.602 0.009266 **
## Ad.SupportedTRUE             1.201e+00  3.632e-01   3.307 0.000943 ***
## In.App.PurchasesTRUE         1.256e+00  1.976e-01   6.356 2.07e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 831.35  on 599  degrees of freedom
## Residual deviance: 669.71  on 577  degrees of freedom
## AIC: 715.71
##
## Number of Fisher Scoring iterations: 8
```

Create third model and evaluate it against the second model.

```
LogReg.mod3<-glm(success ~ Size + Price + Content.Rating +
                   Ad.Supported + In.App.Purchases,
                 data = play.train, family="binomial")
anova(LogReg.mod3,LogReg.mod2,test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: success ~ Size + Price + Content.Rating + Ad.Supported + In.App.Purchases
## Model 2: success ~ Category + Size + Price + Content.Rating + Ad.Supported +
##     In.App.Purchases
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       592     702.48
## 2       577     669.71 15   32.772 0.005047 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(LogReg.mod3)
```

```
##
## Call:
## glm(formula = success ~ Size + Price + Content.Rating + Ad.Supported +
##     In.App.Purchases, family = "binomial", data = play.train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.6698  -0.9328  -0.0419   0.9851   1.9704
```

```
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 -2.364e+00  4.022e-01  -5.877 4.18e-09 ***
## Size                         1.255e-05  3.573e-06   3.512 0.000445 ***
## Price                       -1.857e+00  1.086e+00  -1.710 0.087254 .
## Content.RatingEveryone 10+   8.828e-01  4.652e-01   1.898 0.057747 .
## Content.RatingMature 17+     1.099e+00  7.069e-01   1.555 0.119921
## Content.RatingTeen           6.836e-01  2.278e-01   3.002 0.002685 **
## Ad.SupportedTRUE             1.196e+00  3.480e-01   3.437 0.000587 ***
## In.App.PurchasesTRUE         1.174e+00  1.862e-01   6.304 2.89e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 831.35  on 599  degrees of freedom
## Residual deviance: 702.48  on 592  degrees of freedom
## AIC: 718.48
##
## Number of Fisher Scoring iterations: 8
```

Create final model and evaluate it against the third model.

```
LogReg.mod4<-glm(success ~ Size + Content.Rating +
                   Ad.Supported + In.App.Purchases,
                 data=play.train,family="binomial")
anova(LogReg.mod4,LogReg.mod3,test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: success ~ Size + Content.Rating + Ad.Supported + In.App.Purchases
## Model 2: success ~ Size + Price + Content.Rating + Ad.Supported + In.App.Purchases
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1       593     714.14
## 2       592     702.48  1   11.659 0.0006391 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(LogReg.mod4)
```

```
##
## Call:
## glm(formula = success ~ Size + Content.Rating + Ad.Supported +
##     In.App.Purchases, family = "binomial", data = play.train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.5668  -0.9286  -0.4048   0.9860   2.1170
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 -2.705e+00  3.799e-01  -7.121 1.07e-12 ***
## Size                         1.254e-05  3.487e-06   3.597 0.000322 ***
## Content.RatingEveryone 10+   8.224e-01  4.580e-01   1.796 0.072553 .
## Content.RatingMature 17+     9.247e-01  6.512e-01   1.420 0.155618
## Content.RatingTeen           6.883e-01  2.264e-01   3.040 0.002365 **
## Ad.SupportedTRUE             1.512e+00  3.256e-01   4.643 3.44e-06 ***
## In.App.PurchasesTRUE         1.206e+00  1.839e-01   6.558 5.47e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 831.35  on 599  degrees of freedom
## Residual deviance: 714.14  on 593  degrees of freedom
## AIC: 728.14
##
## Number of Fisher Scoring iterations: 4
```

Check the third model, which is the best model, for multicollinearity.

```
vif(LogReg.mod3)
```

```
##                       GVIF Df GVIF^(1/(2*Df))
## Size              1.036176  1        1.017927
## Price             1.015316  1        1.007629
## Content.Rating    1.034374  3        1.005649
## Ad.Supported      1.037526  1        1.018590
## In.App.Purchases 1.034189  1        1.016951
```

Use validation data to generate matrix for logistic regression.

```
LogReg.add <- play.validate %>%
  add_predictions(LogReg.mod3, type = "response")  %>%
  rename(prob_success = pred) %>%
  mutate(pred_success = prob_success >= 0.5,
         success = as.factor( if_else(success == 1, "Yes", "No" ) ),
         pred_success = as.factor( if_else( pred_success,
                                            "Yes", "No" )))
log_table<-LogReg.add %>%
  conf_mat(truth = success, estimate = pred_success)
log_table<-as.matrix(log_table)
log_table<-as.table(log_table[[1]])
log_table %>% kable() %>%
  add_header_above(c("Predicted" = 1,"Truth" = 2))
```

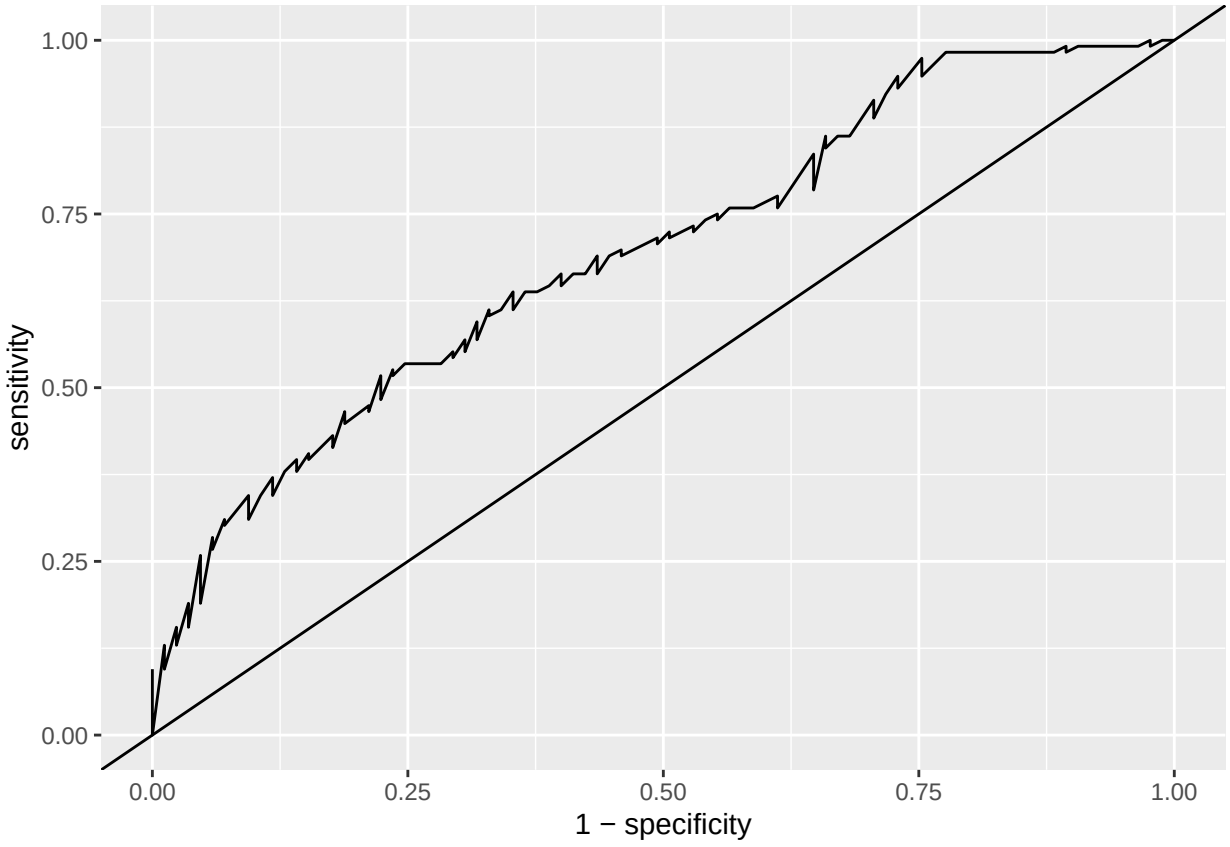
Predicted	Truth	
	No	Yes
No	52	41
Yes	33	75

Use validation data to generate ROC curve.

```
LogReg.add %>%
  metrics(truth = success, estimate = pred_success) %>%
  filter(.metric == "accuracy")
```

```
## # A tibble: 1 x 3
##   .metric  .estimator .estimate
##   <chr>    <chr>          <dbl>
## 1 accuracy binary         0.632
```

```
LogReg.add %>%
  roc_curve(event_level = "second", truth = success, prob_success) %>%
  ggplot(aes(x = 1 - specificity, y = sensitivity)) +
  geom_line() +
  geom_abline(slope = 1, intercept = 0)
```




```
LogReg.add %>%
  roc_auc(event_level = "second", truth = success, prob_success)
```

```
## # A tibble: 1 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 roc_auc binary         0.694
```

9.2 Decision Tree

Decision Tree model

```
df4<-apps
df4$success<-ifelse(df4$Maximum.Installs>140167,TRUE,FALSE)
df4$Category<-as.factor(df4$Category)
df4$Content.Rating<-as.factor(df4$Content.Rating)
df4$In.App.Purchases<-as.factor(df4$In.App.Purchases)
df4$Ad.Supported<-as.factor(df4$Ad.Supported)
df4$success<-as.factor(df4$success)
df4$Minimum.Android<-as.factor(trimws(gsub('[A-Z,a-z]*','',df4$Minimum.Android),which='b

size_format <- function(string){
  string <- gsub(",","", string)
  if (str_detect(string, "M")){
    pd <- sub("\\.", "", string)
    sub("M", "000", pd)
  }else{
    sub("k", "", string)
  }
}

df4$Size <- df4$Size %>%
  sapply(function(x) size_format(x))

df4$Size <- as.numeric(df4$Size)
```

Dividing the data into training, validation, and testing

```
set.seed(1)

df4st<-df4[df4$success==TRUE,]
df4sf<-df4[df4$success==FALSE,]

df4<-rbind(df4st[sample(1:nrow(df4st),500,replace=F),],df4sf[sample(1:nrow(df4sf),500,re

df4.div <- df4 %>%
  initial_split(prop = 0.6)
df4.train <- training(df4.div)
```

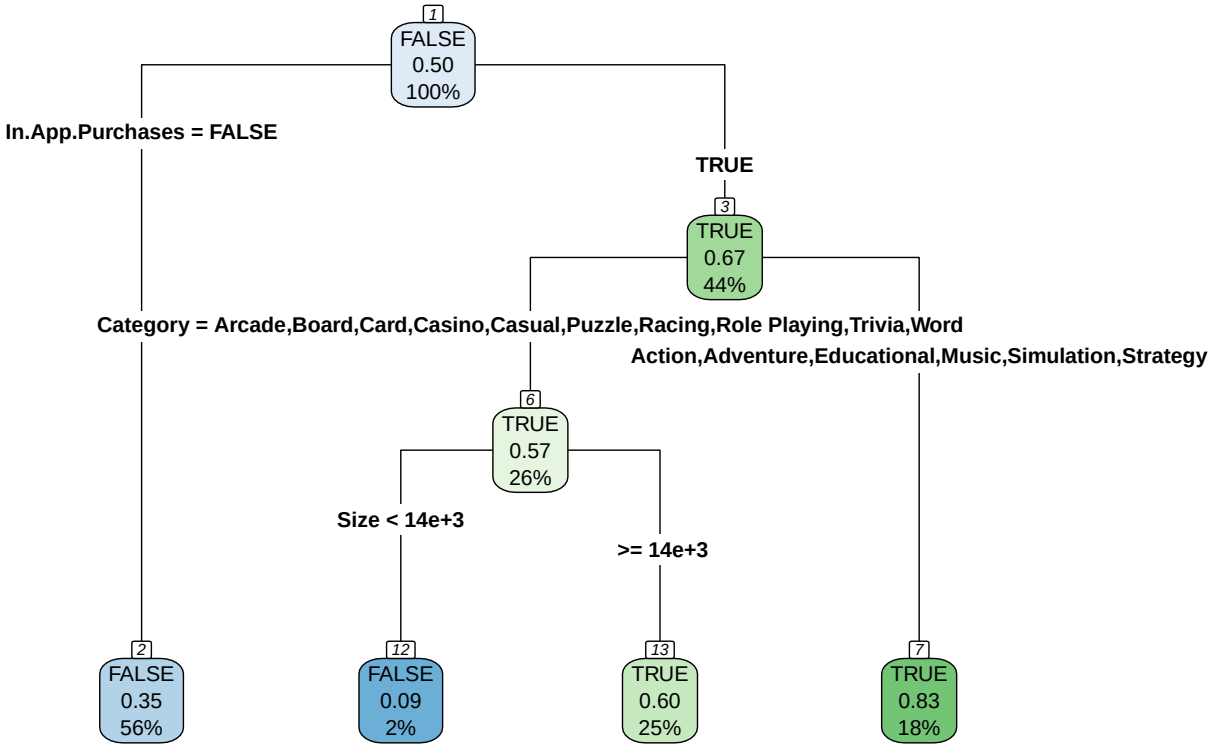
```
df4.div2 <- df4.div %>%
  testing() %>%
  initial_split(prop = 0.5)
df4.validate <- training(df4.div2)
df4.test <- testing(df4.div2)
```

Building the decision tree and prunning it

```
DT.mod <- rpart(success ~ Category + Content.Rating + Size +
                  Ad.Supported + In.App.Purchases,
                data = df4.train)
DT.mod.prune <- DT.mod %>% prune(cp = 0.015)
```

Decision Tree graph

```
DT.mod.prune %>%  rpart.plot(type = 4, nn = TRUE)
```



Decision Tree confusion matrix

```
DT.add.prune <- df4.validate %>%
  add_predictions(DT.mod.prune, type = "class") %>%
  rename(pred_value = pred) %>%
  mutate(method = "DT.mod.prune")

DT_table<-DT.add.prune %>%
  conf_mat(truth = success, estimate = pred_value)
DT_table<-as.matrix(DT_table)
DT_table<-as.table(DT_table[[1]])
DT_table %>% kable() %>%
  add_header_above(c("Predicted" = 1,"Truth" = 2))
```

Predicted	Truth	
	FALSE	TRUE
FALSE	72	38
TRUE	28	62

Decision Tree accuracy

```
DT.add.prune %>%
  metrics(truth = success, estimate = pred_value) %>%
  filter(.metric == "accuracy")

## # A tibble: 1 x 3
##   .metric  .estimator .estimate
##   <chr>    <chr>          <dbl>
## 1 accuracy binary          0.67
```

9.3 Random Forest

Random Forest model

```
RF.mod <- randomForest(success ~ Category + Content.Rating + Size +
                         Ad.Supported + In.App.Purchases,
                       data = df4.train,
                       mtry = 2, importance = TRUE)
```

Random Forest importance output

```
RF.mod %>% importance()
```

```
##                         FALSE       TRUE MeanDecreaseAccuracy MeanDecreaseGini
## Category          11.3282614  6.9462256           12.3729798         59.55598
## Content.Rating     0.2688097 -0.3051276           -0.1273118         14.11734
## Size               3.5020067  9.3675517            9.5065373         67.31864
## Ad.Supported       7.0578120  7.9526063           10.3728666         10.78433
## In.App.Purchases  27.6893509 26.9315600           34.8294165         27.92081
```

Random Forest confusion matrix

```
RF.add <- df4.validate %>%
  add_predictions(RF.mod) %>%
  rename(pred_value = pred) %>%
  mutate(method = "RF.mod")


RF_table<-RF.add %>%
  conf_mat(truth = success, estimate = pred_value)
RF_table<-as.matrix(RF_table)
RF_table<-as.table(RF_table[[1]])
RF_table %>% kable() %>%
  add_header_above(c("Predicted" = 1,"Truth" = 2))
```

Predicted	Truth	
	FALSE	TRUE
FALSE	65	20
TRUE	35	80

Random Forest accuracy

```
RF.add %>%
  metrics(truth = success, estimate = pred_value) %>%
  filter(.metric == "accuracy")
```

```
## # A tibble: 1 x 3
##   .metric  .estimator .estimate
##   <chr>    <chr>          <dbl>
## 1 accuracy binary         0.725
```

Appendix D: Comment incorporation

9.4 Deliverable 1

9.4.1 Introduction

Introduction 1: "Motive/context is clear. Introduction could be more succinctly written (fewer words so it is more concise). I think the questions at the end are an appropriate way to end and get more specific about the business problem in the next section."

Introduction 2: "I really enjoyed reading this deliverable and believe that the introduction text provides sufficient and succinct context for the chosen business problem, which is also very interesting and relevant, especially with the rise of use in technology due to the pandemic we are currently facing. Personally, I feel that there are some parts in the introduction that only serve to add some fluff and I think that it is not necessary."

Both of the comments above point out that our introduction could be written more succinctly. For this reason we went back and made sure to get rid of the excessive verbiage.

Introduction 3: "The context provided was very clear. I think the explanation could be improved by providing cited empirical evidence to bolster some of the claims made in the first two paragraphs."

This comment stated that adding empirical evidence could bolster our claim. For this reason, we decided to include the cost of development to make the reader understand that it is an important problem. However, we did not include empirical evidence from a study or research project as we were worried about the length of our introduction. The first two comments had already told us that we had an introduction that was too long, and we did not want to compromise it.

Introduction 4: "The context to the business problem addressed was concise and engaging for the reader. I really enjoyed how the writing brought the topic to life. Overall, I though that the explanation for the problem was clear in terms of the motivation for the analysis. There was enough background information for me to understand the topic without being too lengthy. The only thing I might still have questions about is what"the next level" for an app typically is. I personally don't know much about the field, so I would be interested know to how app success is measured (i.e. number of users, number of downloads, revenue?)."

Comment 4 in the introduction told us that we did a good job explaining our topic and showing how it is a problem worth analyzing. However, it also stated that the measuring of success was not explained and so someone without knowledge would not be able to make an assumption.

9.4.2 Business Problem

Business problem 1: "I see a few business problems– how to come up with relevant specifications, how to create profits, general strategy approach to achieve success. It is hard to determine what"the" problem you will try to solve is. Explanation could be improved with defining or mentioning what app category is, or what the package size is. I don't know what a package size is, so as a reader I would have liked a small explanation."

The first business problem feedback is extremely important as it tells us that our business problem is not clear, more specifically that they find multiple problems. This is a serious problem as it is important that our readers understand what we are going to try and find insight into. We made sure to go back and fix this by clearly stating what our business problem is.

Business problem 2: "I believe that the business problem is very easy to understand and relevant with the use of delivery apps and streaming apps becoming increasing more popular. I was impressed with the depth of the business problem presented and do not think there's much to change here."

This previous comment states that our business problem is clear and that they do not have any further suggestions. For this reason, there was nothing that we could do to incorporate this feedback; it was really just a comment.

Business problem 3: "This was a very in-depth exploration of the chosen business problem. This section of the report was well written and easy to logically understand. I really have no suggestions other than to lead with your problem statement (the last sentence in the passage) and provide the evidence afterwards. I think your team is on the right track."

The third feedback from the business problem tells us that we did a good job explaining our business model but suggested that we should lead off with our question of interest. The way we incorporated this was not through literally putting our business question first but stating the first sentence in the context of what we are trying to answer.

Business problem 4: "This section discusses many interesting and important issues that come up while determining what characteristics will make a successful app, but I feel that the central business problem isn't very clear. While going through the various strategies an app could employ to grow within the market provides a lot of background to a relevant issue, the section is very long which makes it difficult to determine the specific business problem you've chosen to answer. I don't have any questions about the business problem, my only suggestion would be to shorten the length and highlight which part of successful app development you're choosing to analyze."

This last comment provides us with two problems. We see that the length of the text is very long and that the business question is not entirely clear. In order to fix both problems we decided to get rid of excessive verbiage that explains interesting parts related to our business problem but are not directly related to it. We then made sure that our business problem was clearly stated and alluded to.

9.4.3 Intended Audience

Intended audience 1: "Earlier in the document, App store was mentioned, not Google Play. No big deal but thought I would note that. Intended audience mentioned include the developers to increase chance of success and profitability, good work there. There's a lot of extra info, like"developers should look to apps within their own app-category". Also, I think the info in the last paragraph about how an app will become successful is a little too specific for the scope of this section. How does your business problem affect if the user has more positive experience? The link there is not clear."

The first comment from the intended audience tells us that we include too much information that is not relevant and that we also get into too much detail that is unnecessary. Once again, we see that our texts include too much verbiage and we went back through the section and reduced the length of it. We made sure to not cut out any sentences that were directly related to our business problem and also made our descriptions a little less specific to reduce their length.

Intended audience 2: "This section, too, was understandable and succinct. However, personally, I felt that this section was more of a preview to the Required Data section rather than a section about the intended audience. If this section could be a bit more simple, I believe that this section will be easier to understand. In regards of the explanation of the benefits, I believe that it is very easy to understand."

This second feedback was similar to the first one as it states that the explanation could have been made simpler. Just as before, we decided to go back through the section and generalize the text a little bit more in the parts where we went deeply into specifics.

Intended audience 3: "Overall, a great explanation of the state of the Google Play store; however, the exact audience isn't stated directly necessarily. My one suggestion would be to clearly indicate who stands to benefit from this information (likely the google play store and developers). Other than that, great job delineating the different types of developers and their respective different customers/competitors. It was also very easy to understand the benefits accrued to the audience."

The third feedback from the intended audience section says that the intended audience is not clear. To fix this, we specifically stated who our audience is in one short sentence and made sure that all the other sentences do not mislead the reader on who the intended audience is.

Intended audience 4: "I feel like the intended audience for this analysis was implied, but not explictly stated. From what I gathered, the intended audience is app developers looking to make their app stand out. If that's correct it might help to frame this section diffrently. The benefits were well-explained and easy to understand."

Feedback 4 was similar to the third one because in feedback 4 they state that the intended audience is implied but not specifically stated. To fix this we did the same as for feedback 3; we went back and specifically stated who our intended audience is, and checked that all sentences were written to support this.

9.4.4 Required Data

Required data 1: "Good work on required data. Scope, different data channels, anticipated challenges are all well articulated."

Required data 2: "The data required was described in detail and seemed appropriate for the business problem selected. Obtaining the data also seems reasonable and potential issues were also addressed well. I don't really have any questions or suggestions. Good job!"

Required data 3: "It was very easy to understand what data is needed. My one suggestion, for your sake, might be to limit the scope of your analysis to only a few classes of apps (games & education, or some other combination), otherwise, great job!"

The first three required data feedbacks state that we did a good job describing the data necessary for our analysis and how it could help us answer our business problem. The only thing that was suggested was that we limit the analysis of categories. This suggestion was extremely helpful and the rest of our analysis was shaped by it. We made sure to narrow the scope to only games that are in the Google Play store.

Required data 4: "The required data section is very thorough and easy to understand in regards to what data is needed in order to research this business problem. One thing that could help with this section is compacting the information given into a smaller section. It's good to give a very thorough explanation for the data but if the explanation becomes too long, it gets harder to read it and may have an opposite effect."

The last feedback from the required data section informs us that, once again, our explanation is too long. We decided to go back and remove all the details that had already been stated or alluded to. Similarly, we made sure to only go into some detail where we thought there could be some confusion. These changes reduced the length of the text and made it easier to read.

9.5 Deliverable 2

9.5.1 Data

Data 1: "Description is succinct and easy to understand. I think Figure 1 is a helpful but perhaps would be more useful if it were larger. The data in the kable is helpful and is clear in demonstrating what the data looks like. I am not sure if you need to include the code that you used to create the kables."

This comment tells us that in general our data section looks fine. There are two suggestions which are making figure 1 larger, and that we do not need to show the code for the kable. We decided to go back and make the image larger, and hid the code that creates the kabled table as well.

Data 2: "The description is very succinct and easy to follow, especially with the graphics provided. In addition, the variable names are very self explanatory. However, I was and am confused about the difference between Installs, Minimum Installs, and Maximum Installs. Also I would suggest that you try to give a brief description of the variables for clarity sake and how they can be used. Lastly, I would suggest to make sure that none of the code goes over the page as it can limit the understanding of the code."

This comment also tells us that in general, our section looks good. One problem that they state is that we do not state the difference between Installs, Minimum Installs, and Maximum Installs. We fixed this by removing all but Maximum Installs as it was the only accurate and numeric variable that we could use as a measurement of install counts. They also suggested that we describe each of the variables a little bit more and that we make sure that none of the code runs off the page. The way that we fixed this was through expanding the descriptions in the data dictionary and by putting code on new lines so that when it knits, the entire code is showing.

Data 3: "The file did not knit correctly, as there are warnings at the top. Great description of contents, loved the table describing the variables. You only need to include 3 - 4 rows for the data table. This will allow for more compressed stacking. Also, you do not need to include the R Code for the construction of the tables. Also, the third table runs off the page."

The third feedback for the data section tells us that the file didn't knit correctly, that we should not include the r code, and that the table runs off the page. The way we fixed the knitting problem was by specifying that warning must be suppressed. Similarly, we also specified the echo to be false so that the code is not shown when the file knits. Lastly, we fixed the table by changing the number of variables, thus columns, in each table.

Data 4: "This team did a great job in terms of describing the dataset, what variables were included, and what the variables represented. I clearly understood the description provided. There were small issues with reading the data shown in the table as the words in the table sometimes overlapped with each other, making it difficult to read. The data in the table was still able to be understood for the most part. The only suggestions I have for this section are in regards to formatting and style. Figure 1 appears a bit small on the page, and I needed to zoom in quite a bit to read it. I also noticed that some of the R code was running off the page and I couldn't see all of the code."

This feedback's main concerns were regarding the format and style. It states that figure 1 is too small and that some of the r code runs off the page, which is similar feedback to what we received in the third feedback. What we did was resize figure 1 and made it larger, and then put code on a new line so that way all of it is showing when the file knits.

Data 5: "The data description provides a clear description of the contents of the clean data. A suggestion is that the data table presented could be larger for text reading."

This comment is very straightforward, the suggestion is to just make the table larger. We fixed this by making figure 1 larger as stated in the two previous feedbacks for this data section.

9.5.2 Data Collection

Data collection 1: "The data collection section is clear and useful in understanding why the dataset will help you answer your business problem. One recommendation I would make is that you should put an in-text citation in the first paragraph after you use a quote. Also, I think you could elaborate a little on"The one quality issue that may affect our analysis is the imprecision of the install counts of the games". Overall very well done!" This first data collection feedback suggested that making an in-text citation after the quote would be a good idea and that we should elaborate on the quality issues. We addressed these two suggestions by making sure to cite the quote properly and by expanding the quality issue without using too much technical jargon and verbiage.

Data collection 2: "The description provided in this section is very well put together and displays a lot of understanding about the task at hand. The information is not too drawn out nor is there much of a lack thereof. One thing that would be improved is expanding upon the example provided about the Android version and release month. This seems like an interesting example but just saying that there could be new insights from those variables doesn't exactly convince me that they could be important."

This comment says that we should explore a little more in depth the android version and the release month. We decided to explain them a little bit but not too much as we thought that it would expand the example too much and we tried to keep the section succinct.

Data collection 3: "Great job, no feedback."

Data collection 4: "I clearly understood how the data was collected, and I really liked how they included the date the dataset was accessed on as well. The sources included seem thoroughly addressed and the data collection description is succinct. I understand the considerations discussed, and they brought up some really good points about the scope and depth of the dataset. The only question I have is about the distribution of number of games between the 17 genres. Are there some genres with significantly more or less games in them? Having genres in the dataset with relatively few games could affect the usefulness of the analysis to certain developers."

The two above feedbacks inform us that we did a good job in the section. The only question,

or suggestion, is that we expand on the description of games, especially their count. This is a problem that we encountered as there is no fair representation of games but a way that we fixed this later on is by taking random samples of our dataset so that there is more of a balance between the games.

Data collection 5: "The first quote needs an in-text citation. Other than that, I fully understand the data collection summary, as well as, the source of the data, appropriateness, quality, and usefulness. The additional drawback about the install counts being imprecise, but not being a result of the data necessarily also added another layer to the data portion."

This comment once again suggested that we fix the in text citation. We made sure to go back and fix it before submitting the revised version as well as the final project. The rest of the comment seems like it is stating that we did a good and clear job.

9.5.3 Data Preparation

Data preparation 1: "Data prep section is clear and concise! I would suggest to maybe briefly add to why you didn't use size and minimum android version supported? When reading I wondered why these, esp the size variable, would not be useful. I see you have addressed that at the very end of the section, but I think it would have been more helpful to me as a reader earlier. Other than that, I think explaining the reason for why you have chosen your specific methodology is very well done. In terms of details, a lot of the information is repetitive. My understanding of the purpose of the details section is to note your procedures specific to R and Excel. For example, I read the reason for taking out Sports category 3 separate times. A lot of the information in the R and Excel sections is the same as the general Data preparation section. In the excel part, if Figure 2 and Figure 3 could be larger, that would be more easily understood by the reader."

Data preparation 2: "The data preparation is very straight to the point and is easy to follow. However, one thing I would like to see expanded upon is why you removed some columns like Free, Developer ID, etc. since some columns could be seen as something that could contribute to the business problem to others. Simply stating that you removed those columns can confuse the people reading. Regarding the considerations, I believe that this section is very well explained and discusses the considerations very well."

Data preparation 3: "I would like justification as to why the column"Free" was irrelevant. Other than that, the logic was very easy to follow and I have no other suggestions."

Data preparation 4: “The description provided was very through and easy to understand. The reasoning behind each of the decisions was well explained for the most part. I had one question about the methodology in terms of the exclusion of certain categories of games. It is my understanding that the business problem only wants to address what makes a successful gaming app, but the dataset appears to only include gaming apps. I am curious as to which categories in the dataset are not gaming apps. The additional considerations were well addressed and I understood their emphasis on complete data for their analysis.”

All four of these feedbacks for the data preparation section clearly say that we did not do a good job explaining why we thought some variables were not important, and why we decided to drop them. We made sure that we went back through this section and gave an explanation into why each variable that was removed did not matter for our analysis.

Data preparation 5: “I understand the data preparation process described. I believe a place of improvement could be clearly detailing where your group talks about appropriateness, quality, and usefulness. These topics could be further discussed after the data cleaning that was described.”

This last feedback was similar to the others but not exactly. Where the other four stated that we did not clearly say why variables were not important, this comment states that we should more clearly address why kept the variables that we did and that we could do this in a further section.

9.5.4 Data Preparation Details in R

Data preparation details in R 1: “All writing is clear and concise. My understanding of the purpose of the details section is to note your procedures specific to R and Excel and any issues you might have run into for one and not the other. For example, I read the reason for taking out Sports category 3 separate times.”

Data preparation details in R 2: “One suggestion I have for this section is to explain very shortly how you would go about the steps described using function names and such; even putting them in parentheses would be sufficient. Also, a lot of the information in this section is already provided in the Data preparation section, which makes this section sound redundant. Finally, with the comments in the code, all of them go off the page, making it hard to understand exactly what is going on in the code.”

Data preparation details in R 3: “No need to detail not including”Sports" again, it is just redundant. The code and explanations for many of the lines run off the page. Combat this

by entering your code on multiple lines. No need to include the tables again. Please include the str() output for the final cleaned dataset."

The first three feedback comments for the data preparation details in R section show us that we were too repetitive in our description of the cleaning section. They also inform us that we should not go into the detail of the cleaning such as the reasons why we remove or modify certain variables, but rather we should only describe the code itself. These two were easy fixes as we removed essentially everything that we had written and only explained what our code did. So, we essentially replaced the text that we had with pseudo code.

Data preparation details in R 4: "The data preparation in R was easy to understand, but i would suggesting breaking up the big paragraph into more digestible chunks. I would also once again work on making all of the R script readable on the page. Overall, I ultimately understood the process and the results, but the formatting made it a bit more difficult."

This comment let us know that it would be easier to digest the code if the description was separated into smaller chunks. We made sure to fix this as we replaced the text, as stated previously, to make sure that we explained each section clearly.

Data preparation details in R 5: "Yes, I understand the process and results shown in the data preparation R details. A place of improvement is making sure the first table is readable because the App ID slightly overlaps with the category variables."

The last feedback on the data preparation details in R section told us that the table we included was not easy to read as variable names overlapped a little. We were able to go back and fix this by specifying the number of variables, thus columns, in each table so that there would be no overlap.

9.5.5 Data Preparation Details in Excel

Data preparation details in Excel 1: "A lot of the information in the R and Excel sections is the same as the general Data preparation section. If Figure 2 and Figure 3 could be larger, that would be more easily understood by the reader."

Data preparation details in Excel 2: "This section has the same problems as the section for R. The description of the steps and in general have already been provided in the Data preparation section and now the Data preparation details in R section. In addition, the figures provided are too small and illegible, providing not much information. Like i suggested

in the preparation details in R section, try to include some functions or technical terms regarding the data preparation. Another thing to consider is providing screenshots of the steps taken to prepare the data in Excel."

Data preparation details in Excel 3: "No need to reexplain the motivation behind removing certain elements of the data set for a third time. Please still detail the process, but the justification only needs to be provided in the original Data Prep section. I would rec. making the screenshots larger and adding a few more screenshots that detail the process you went through as you cleaned."

The first three feedbacks for the data preparation details in Excel were very similar to the previous section. These also stated that we should not have repeated the same information that we did in the data cleaning section. For this reason, we made sure to go back and only specify the steps taken in Excel. We removed all text that described anything but the process on how to get the same results we got when cleaning in Excel.

Data preparation details in Excel 4: "The process of data preparation in Excel was explained clearly and I understood the results provided. My only suggestion would be to consider breaking up the paragraph after Figure 2 to make the process easier to follow and to make Figures 2 and 3 larger as I had difficulty reading them."

This feedback was very similar to the fourth feedback in the data preparation details in R, as it also states that we should break up the text into smaller chunks so that the reader can follow along more easily. We did follow this suggestion and made sure that we described each step clearly and in small chunks.

Data preparation details in Excel 5: "The data preparation details in Excel are clear and succinct. An area of improvement is describing more of the technical process of how you omitted the Sports category using Excel or how you filtered out certain values such as currency."

This last feedback is similar to the previous one. It states that we should not describe the reasoning behind the data cleaning but rather just describe the steps that we took to clean the data in Excel. So, we went back through this section and removed all data cleaning explanations, and replaced it with just the description of the steps in Excel so that other people can replicate our process.

9.6 Deliverable 3

9.6.1 Descriptive Analytics

Descriptive Analytics 1: “It might be helpful for a brief explanation of what ‘median number of installs’ means while considering the log transform. Someone who might be unfamiliar with a logarithmic transform might benefit from this. Why did you decide to use a log transform? In the intro,”an example of this would be upgrading a building may take 24 hours to complete" is a confusing sentence. The thought process and logic linking what is shown in the visualizations to what this could mean for developers is clear and concise. For the third visual, I think there clarity could be improved with what is meant by ‘median maximum’. Maybe “median of the maximum” would be less confusing. Also, in the caption, what is meant by ‘during the games’ launch time? This is not mentioned anywhere. The table is well-done and I found it very informative. The write-up for table 1 could be more concise with verbiage”.

We explained our reasoning for the log transform in more detail. We also clarified the syntax regarding the median install count. We removed the part about launch time in the caption as well.

Descriptive Analytics 2: “The purpose of this section is to briefly introduce and explain the visualizations. I believe that your team provided a wealth of context and insights that could be trimmed out of this section. For the revision, please move some of the general conversation surrounding the context/insights to the insight summary section. Very well written nonetheless.”

We moved many of the details about insights to the next section

Descriptive Analytics 3: “I understand the explanation behind Figure 1, but I feel like the background information could be more succinct. You could consider cutting down on the background information for in-app purchases and play to win strategies. The visualization was also easy to interpret. My only suggestion would be to consider relabeling true and false to yes and no, as that might be more intuitive to understand. For Figure 2, the explanation was easily understood and succinct. In terms of the visualization, my suggestion would be to possibly change the colors of the bars for Figure 2, in order to better differentiate it from Figure 1 at a glance and to also consider changing True and False to Yes and No. The Figure 3 explanation was clear and very insightful. In terms of the visualization, I liked the color gradation and it would be interesting to see if you could change the gradation to correlate to the median maximum install counts. I would also suggest sorting the game categories by largest to smallest median maximum install counts

to make the graph easier to interpret at a glance. The reasoning behind the explanation for Figure 4 was clear and easily understood. I have no outstanding questions, but I do wonder if the game genre has any effect on the size of game. For example, people looking to download a racing game understand that it will take up a lot space on their phone anyways so they don't care about the size of the game, while others looking for a word game would be more contentious of the space. In terms of the visualization, my only suggestion would be to possibly introduce some color in the form of a gradation. For Table 1, the description was clear and easy to understand, but due to the length my suggestion would be to split up the paragraphs even further so the information is more easily digestible. I would also possibly consider the same things for the graphs. The number of games column may not be totally necessary, and splitting the median maximum installs, ad supported, and in app purchases into separate tables then sorting them highest to lowest may be easier to understand."

We did not feel that renaming the boxplot x-axis was particularly necessary. We also sorted the barplot accordingly. We kept the color gradation of the scatterplot as is because we did not see a reason to add color to such a dense plot. We also fixed the table to remove columns that we did not feel were necessary. We still felt that game size was a relevant variable because there is still size variation within categories that could impact install count.

9.6.2 Insight Summary

Insight Summary 1: "Insight summary is understandable and concise. It also does a good job of summing up without being too repetitive. I don't have any suggestions for this section."

No suggestion provided

Insight Summary 2: "Great succinct description of the results. Once you bring over some of the context (albeit abbreviated) and list the actual values from the visualizations from the previous section, this will be a very strong section."

We initially included actual values in the insight summary, but after Professor Martinet's comments advised us to trim down the section significantly, we decided to remove them for brevity's sake.

Insight Summary 3: "I understood the descriptions provided and I really liked how they tied all of their insights back to addressing their business problem. My only suggestion would be to maybe include in the names of the genres that have the highest and lowest

install counts in the second paragraph."

We included the names of the highest and lowest installed genres

9.6.3 Descriptive Analytics Details

Descriptive Analytics Details 1: "Code is very well-commented and detailed about the process. In the bar chart, I do wonder about the color scheme? Does the color signify anything? Other than that, I do not have any suggestions."

No suggestion provided

Descriptive Analytics Details 2: No comment.

No comment provided

Descriptive Analytics Details 3: "The code was well-commented and I clearly understood how the visualizations were created. The colors helped a lot with the understanding of the code as well. My only suggestion would be to fix the instances where the code runs off the page, such as page 12 of the pdf."

We fixed the code run-off in the knitting

9.7 Deliverable 4

9.7.1 Predictive Analytics

Predictive Analytics 1: "This was written very clearly and I think defining what was considered a success was very important. I would maybe be cautious about the wording of"what causes a game to be successful". Cause is a strong word so maybe phrasing it more like insights into factors that contribute to a game being successful or explain how you would determine a cause? I am a little confused because you said you used logistic regression, k-means, decision trees, and random forest, but you built models using linear regression, decision trees, and random forests. Were those meant to be the same because I don't really understand why they are different. Overall, very clear about the goals of the modelling."

We addressed the issue of including extra models. Additionally, we changed the text from "We hope to discover insight into what causes a game to be successful in regard to total installs, and which variables are most important in determining success." to "may lead to."

Predictive Analytics 2: "The goal is nicely stated and the models are listed out clearly. The process part said that the team has used k-means clustering, but didn't include it in the final result, so maybe briefly mention why you choose not to include k-means clustering (maybe the accuracy rate is too low?)"

We removed the mention of k-means clustering as we did not ultimately build the model.

Predictive Analytics 3: "Good general introduction; however, I am a bit confused. You began by mentioning 5 separate methods you intend to use, but then mention a different set of 3 models you built. Please make this more transparent in the revised deliverables."

We removed the two extra methods we did not use and corrected by changing "linear" to "logistic"

Predictive Analytics 4: "The descriptions are clear and understandable. No suggestions :)"

Thank you. No suggestions.

Predictive Analytics 5: "This is an interesting proposal and the methodology appears sound! I think there is some repetition and general editing needed for the paragraph though."

We went back and addressed the repetition, mistaken inclusions, and made all around edits to ensure high quality work.

9.7.2 Process

Process 1: "The process was laid out very clearly and I understood exactly how all the techniques worked. You did a great job explaining in non-technical terms. Really good work here. If you do end up using linear regression to create a model like you mentioned in the previous section, remember to add an explanation for that process too! Also, it may be valuable to include how many observations were in the original data and why you chose to take 500 random samples as opposed to another number?"

The computational constraints of including 500 observations were already fairly high, so including any more observations in the model may have caused R or the computer to crash. As mentioned previously, we ultimately did not create a linear regression model.

Process 2: “I really like how the team put in consideration the balance of successful the failed games and selected 500 each. The description is great but I am not entirely sure that words like ‘75th percentile’,”splitting" are very non-technical (maybe check with the TAs and Professor Martinet). For example, maybe changing the ‘75th percentile’ to ‘top 25%’? Also did you split the data in 60%, 20%, 20%?”

These words do happen to be too technical, so the verbiage was changed. We did split the data into those proportions; however, the mention of that also happens to fall into the “too technical” range.

Process 3: Great general description that was very straight forward.

Thank you. No suggestions.

Process 4: “This section was very descriptive and provided good explanations for each of the methodologies, but I felt that the decision tree explanation was still a little too technical for a non-technical explanation. For example, mentioning the purity of a tree and the homogeneity/heterogeneity of the groups may be a little too technical for this section. Also, I have a question on the success/failure method – was it games at or above the 75th percentile, or only games in the 75th percentile?”

We removed the technical explanations. To be specific, the 75th percentile includes the 25 percent of all games that happen to have the highest install counts.

Process 5: “The process is clear and makes sense. There is good review of the methodologies, but some parts are overexplained relative to others. Structuring the paragraph more in an understandable way could help improve this, but overall this paragraph helps define the process!”

We slightly restructured the paragraphs and improved the flow through general edits.

9.7.3 Assessments

Assessments 1: “Your group did a fantastic job laying out the visualizations so that they were easily to follow. I really like that it wasn’t R output, and it was actually a screenshot! The processes were really concise and easy to understand. Only thing I would consider changing is potentially moving the figure captions to be on the top rather than the bottom of the screenshot so it is a tad easier to follow and identify.”

We ultimately decided against moving the captions to the top of the figures because, in our experience, captions typically come at the bottom of figures and plots. We were also required to change the screenshots as they are screenshots of R output.

Assessments 2: "The assessment is really nice and I understand everything clearly. Just a small mistake about the title of the figures: all 4 figures are in the pattern of"Figure x: Figure x"."

We adjusted the figure captions so they did not label the figure twice.

Assessments 3: "Double listed Figure X titling. I rec. another read through for grammatical errors. I would convert the variable importance table from a table to a graph (it is just easier to interpret this way)."

We were unable to convert the variable importance table to a graph, so it remains a table. We also altered the captions so that they did not double list "Figure X."

Assessments 4: "The description was very thorough and I liked the further explanation for keeping one of the not-as-significant variables in the final logistic regression model. I also liked the explanation of multicollinearity so readers know what it actually means and how it is relevant to the model. The visualizations were clear and easy to follow."

Thank you. No suggestions.

Assessments 5: "I think there is a bit too much background provided: I think providing a final model with a brief explanation of how you got there would suffice. In addition the graphics could be made more consistent in style to make it more professional."

We aimed to make the graphics more consistent as we agree with this commenter. We also removed some of the background information and included more evaluation informations so as to justify the model's inclusion.

9.7.4 Results

Results 1: "You guys did a really good job explaining all the performance metrics and it is displayed in a very organized manner. The only thing I would consider changing is moving the figure captions to be on top of the output rather than the bottom for readability improvement! Great job!"

As stated previously, we decided to leave the figure captions at the bottom of the figure instead of the top.

Results 2: "I like the ROC curve and the confusion matrix. Same thing with the figure titles"Figure x: Figure x". Also I notice that the matrix numbers are a little bit weird. Before you said that you select 1000 samples, 500 success and 500 failure. If you are using 60%, 20%, 20%, the testing dataset would have 200 games (like figure 5, the total is 201). However, figure 7 and 8 are something like 20,000. So maybe check that part as well."

There was erroneous data included in that model which we did not initially notice. We made sure to only include the relevant data so that the data matched the other models. As stated previously, we removed the issue with "Figure x: Figure x."

Results 3: "Great discussion of the model results."

Thank you. No suggestions.

Results 4: "The description is clear and understandable. I would suggest decreasing the size of the confusion table because it's fairly large (to be the same size as fig 7 and 8) and decrease the size of the ROC curve as well so it's easier to connect the explanation of ROC to the ROC visualization. In the last part of this section I would also suggest explaining what the gini coefficient means because you explained the assessment measures for the other types of models but not random forest."

We attempted to decrease the size of the confusion matrix with moderate success. The same goes for the ROC curve.

Results 5: "The information here is on point and useful! I think organization would help: the graphics seem a bit haphazardly placed, but the text is overall understandable and helps clarify the main results well!"

We attempted to place the graphics with more care on the final deliverable.

9.7.5 Insight Summary

Insight Summary 1: "The insights were valuable and succinct. Very good job on this section. If the random forest model and decision tree results were very similar is there a reason you included both? Maybe just using one of the other would be better because they provide similar insights. Overall, very well organized and great job!"

Even though the results from the decision tree and random forest were similar, we believe that the inclusion of both supports our ultimate conclusion.

Insight Summary 2: "The finding is really interesting: I didn't expect that in-app purchases would be such an important variable! Also the game category being a non-important variable is surprising as well. Overall a great deliverable, even though I didn't get to read your team's deliverable 1-3 I could understand everything clearly and I really enjoy it. It feels like your team is doing a much better job than mine hhh (we crammed until 11:37pm on Sunday and the whole thing is really messy). Great job!"

Thank you. No suggestions.

Insight Summary 3: "Good discussion of the insights here, but I push you guys to go a bit deeper. Instead of just listing the important variables in each model, I suggest you discuss reasons WHY they could be important or discuss the context that these results were generated in. In doing so, try to remain succinct and cohesive."

We moved most of the initial insight summary section up to the Results section and included reasoning for why certain variables were more important than others.

Insight Summary 4: "This section was understandable and very clear, and I liked the end of the description where the difference in accuracy scores was addressed and further reasoning for choosing the model with the lower accuracy score was given. No suggestions."

Thank you. No suggestions.

Insight Summary 5: "This is a very textbook summary and I think there has been a good deal of effort put into making sure the insights are understandable, so I think this section works very well!"

Thank you. No suggestions.

9.7.6 Predictive Analytics Details

Predictive Analytics Details 1: "the code chunks were very easy to follow and broken up in a very methodical way. The rubric says the results should be output so make sure eval is set to true and the results are printed out, not just the code!"

We made sure to include the output in this section as well.

Predictive Analytics Details 2: "The code is really neat and nicely divided. However, the references part is kind of off so just fix that and it would be great! (our team didn't cite correctly last deliverable and we got 7/10 in format...)"

We fixed the references section so that it was located on the very last page of the deliverable.

Predictive Analytics Details 3: "Please include the outputs in this section as well. Also, the Reference header is missing."

As stated previously, the output was added and the reference header was included as well.

Predictive Analytics Details 4: "It was easy to understand the process of this section – both the explanations and code were easy to follow. I would suggest to state the size/split for the training, validation, and testing data to state that you used a 60/20/20 split. I was wondering why the summary data from the models was not included – although it may be somewhat long it would be easier to see exactly what the models were telling you and seeing the difference instead of just describing the code, especially when the description of the code mentions a graph/visual. Also, references should be on their own page."

The data split is referenced in the final deliverable and summary information could not be presented in a format acceptable for a deliverable, hence it was not included. We ensured the final deliverable included output in this section as well as moved the references to their own page.

Predictive Analytics Details 5: "The code being colored is very useful, and I think the overall style is very appealing. This section has been broken up nicely, and the code is very readable! I think making sure code doesn't run over different pages would help"

Unfortunately, the code highlighting needed to be removed. In order to make the code slightly more understandable, we made each comment into a sentence explaining the code chunk.