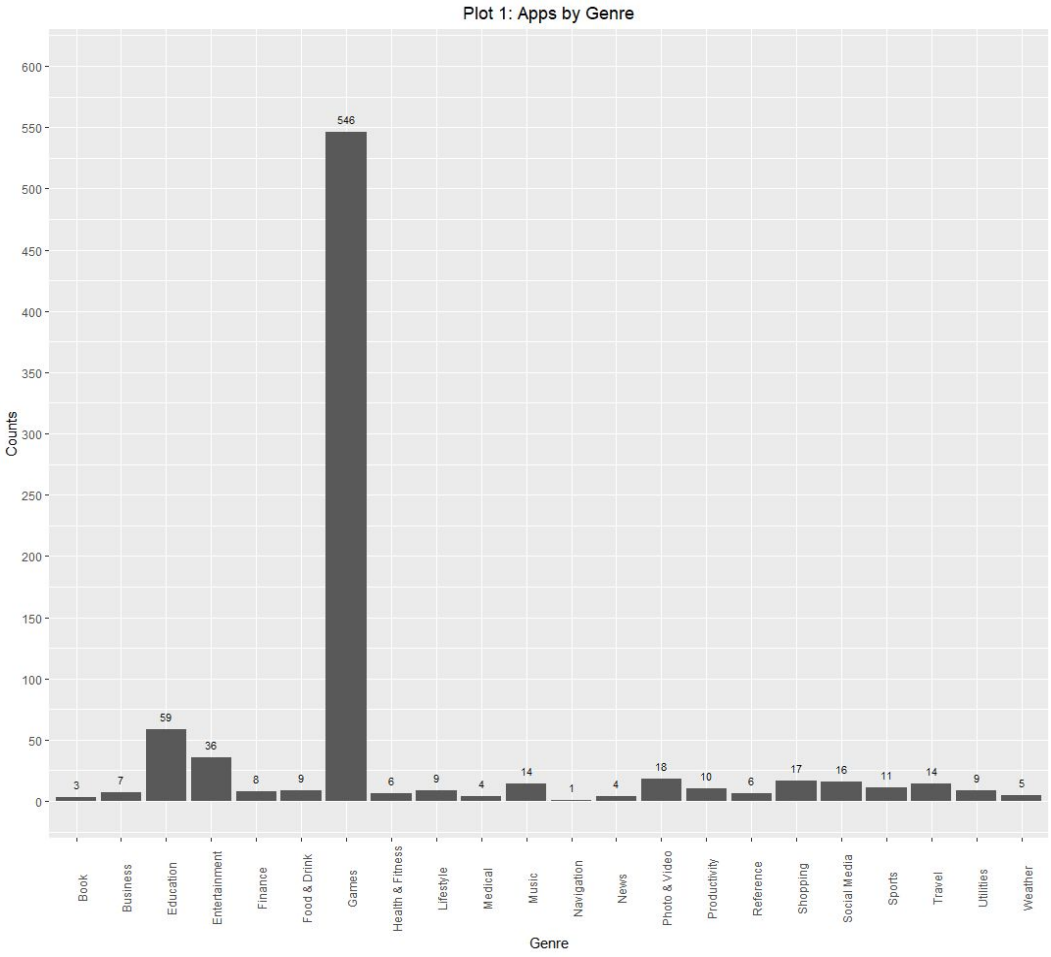
Module 6 Final Project
Brad Viles
ALY 6000
October 25, 2020

Abstract

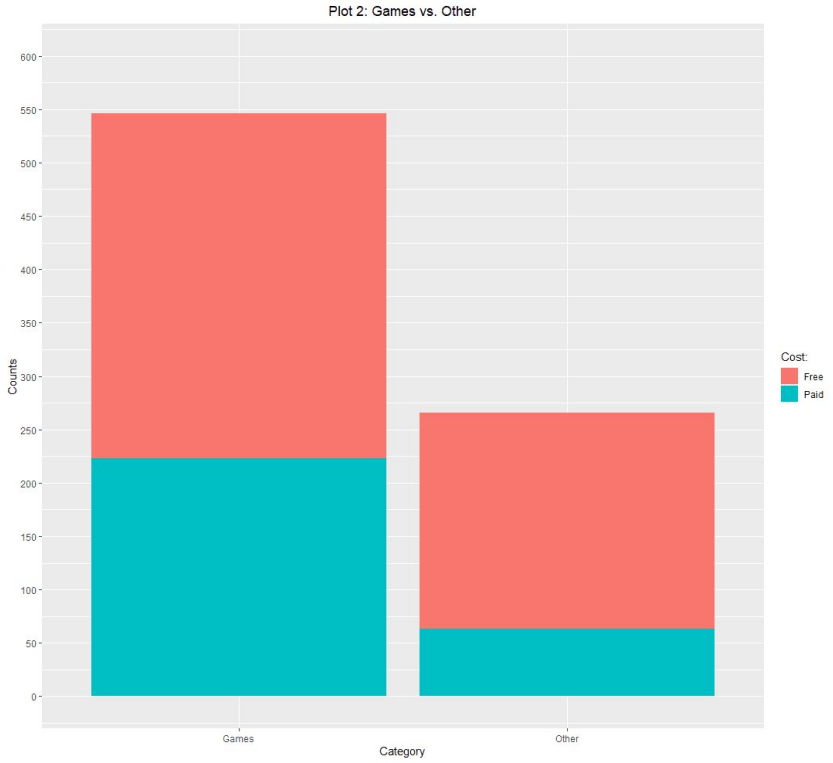
I chose a collection of three data sets to work with in this project that summarized characteristics of mobile applications available on the Apple store and the Google Play store. The main goal of this analysis is to understand the characteristics of apps that are present in both mobile application stores in order to increase potential audience for future app development. I will explore characteristics such as genre, the relationship between ratings and reviews, and content rating of the applications. The datasets that I worked with are as follows: Apple - 7197 objects of 17 variables, Google Play - 10841 objects of 13 variables, and Google Play 32K - 32000 objects of 12 variables (Prakash and Koshy, 2019) (Ramanathan 2017). In order to best understand the intersection of the two application stores, a cleaned data set combining the Apple data and the Google Play 32K data was created with 812 objects of 16 variables. It can be assumed that these do not encompass all apps available on each respective store, especially as the Apple data was collected in July of 2017 and the Google data was collected in April of 2019. Nonetheless, we will be able to make some starting generalizations to create a roadmap to create our own successful multi-platform app.

Initial Analysis

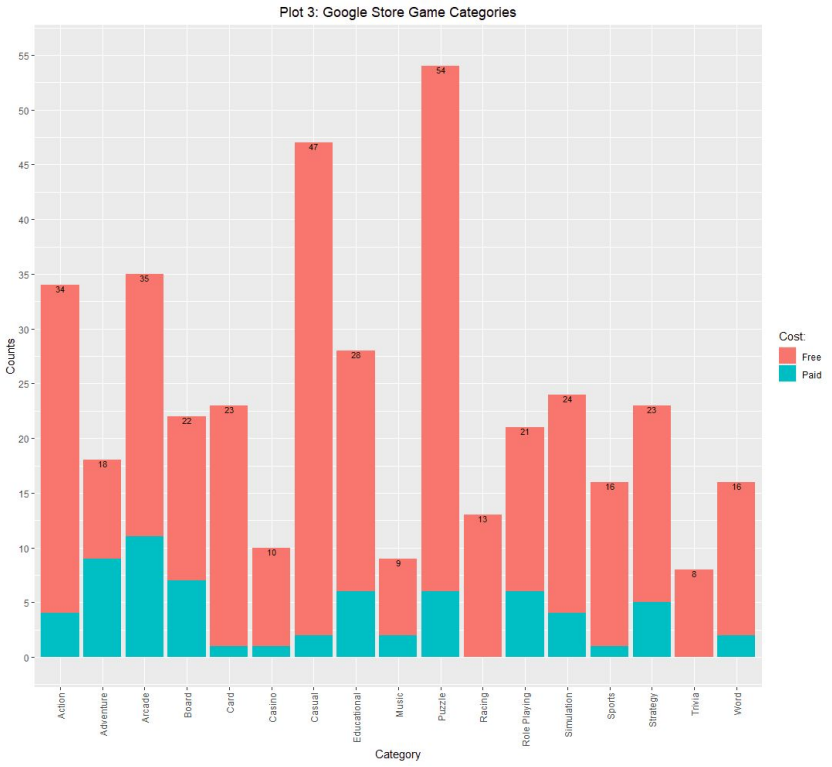


I began my analysis by breaking down the combined dataset into genres in order to see which of the 22 notated genres was most prevalent within the data. Unsurprisingly the dataset was dominated by games as shown in Plot 1. Upon seeing this I decided it would be more beneficial to separate games

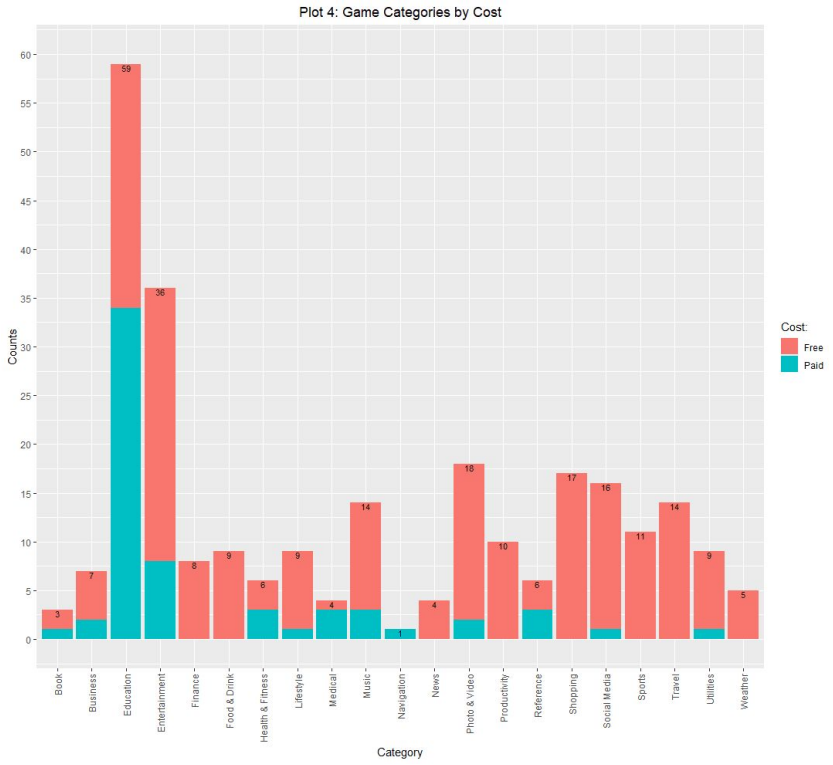
into its own category compared against the rest of the population. Additionally, Plot 2 is segmented by free versus paid apps. This shows that even when compared against the rest of the population games are far more prevalent, and there are almost as many paid games as there are other apps within the population. In total the combined data includes 266 non-games to 546 games. On initial inspection of the combined dataset it appears that games are the most successful application type on both stores, and notably the data does not notate "freemium" apps that include additional in-app purchases. Plot 3 breaks down the games data further by category.



Notably, this plot uses only data from the Google Play store, as the Apple data did not subset their games data by any specific subgenre or category. Puzzle and casual games top the charts with 54 and 47 apps respectively, but also have some of the largest counts of free apps. I will again reiterate that this data does not include apps with in-app purchases which can still generate large amounts of revenue. For strictly paid apps, arcade, and adventure games edge out the field, and notably have a high ratio of their sample as paid apps. On the other end of the spectrum are



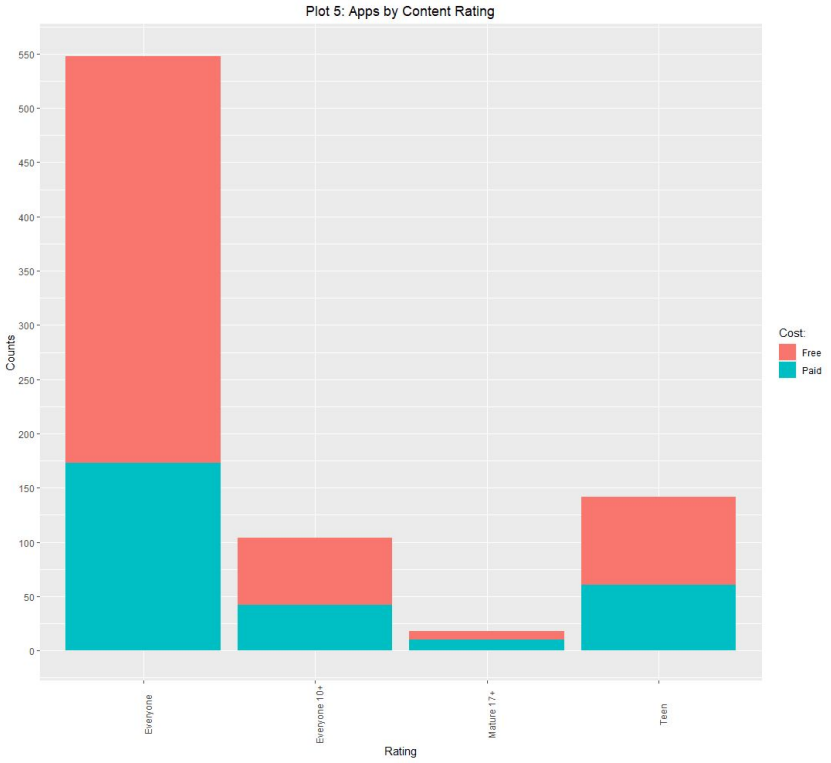
racing and trivia games which logged no instances of paid apps on the Google store.



Plot 4 switches gears to explore the other side of Plot 2 within the combined dataset. Logging an enormous count of paid applications are educational apps. 34 of the 59 educational apps in the population are paid, which given its own category would still be second to entertainment apps by only 2 counts. There exists a large population of apps that are entirely or primarily free, with only paid entertainment apps getting close to double digits with 8 counts. Other notable features from this plot are that 75% of medical apps, 50% of health & fitness apps, and 50% of reference apps are paid. The sole navigation app in the combined sample is paid as well, but due to the small sample size we may guess that each platform utilizes its own proprietary navigation application, and so there is no overlap within the respective application stores of navigation apps.

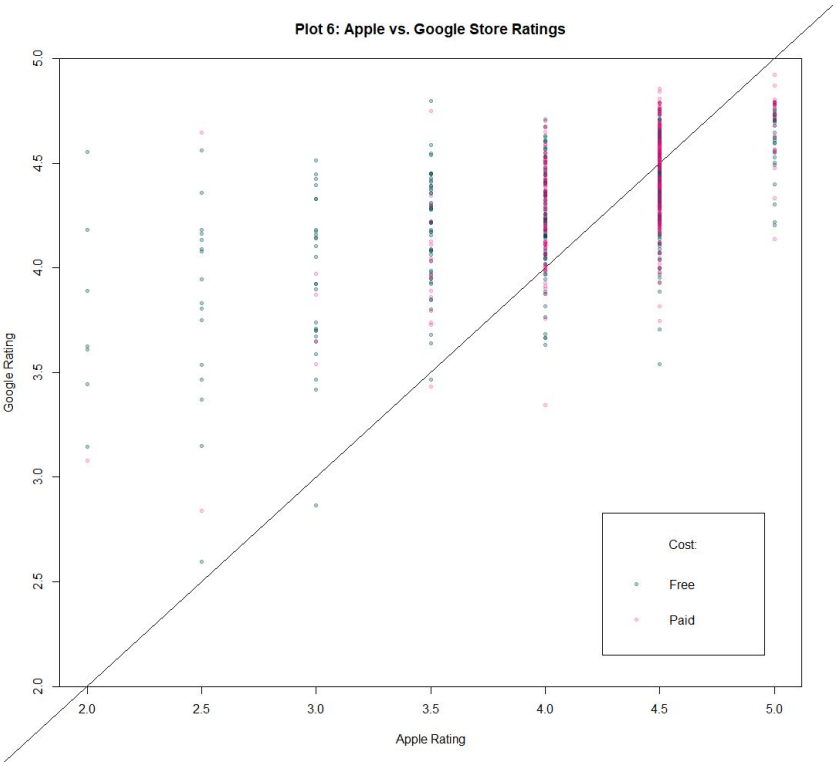
By breaking down the different categories of apps in this manner, we appear to be faced by the primary decision of whether we should aim to develop a game or a different type of app. From there the choice is how revenue within the app is generated. There is an option to simply charge a flat price for download, or create a free app with other revenue models like sponsored ads, data collection for resale, or freemium in-app purchasing. There are clearly certain genres both within and outside the games subset that lend themselves towards one revenue model over another.

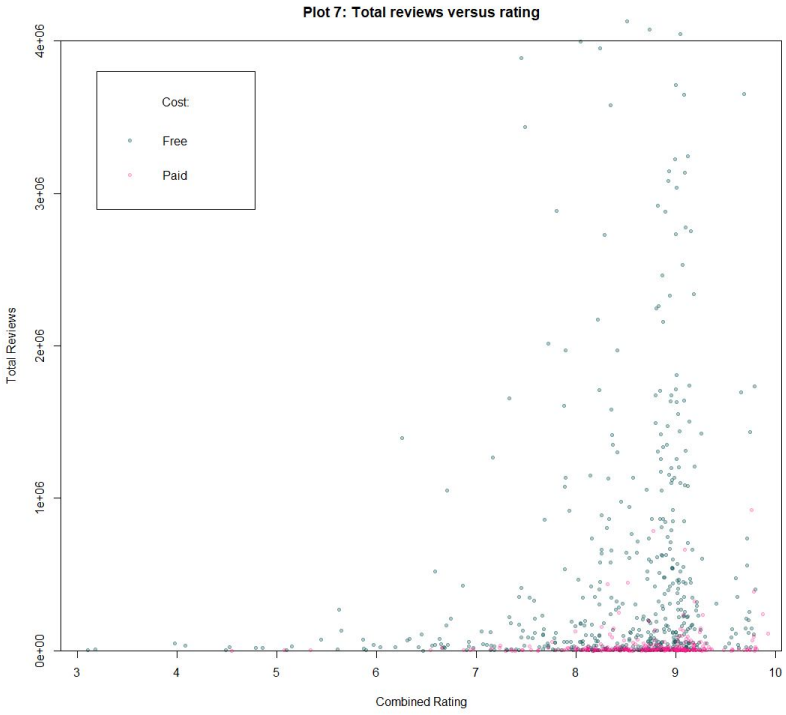
The final aspect of free versus paid apps that I was interested in exploring was a breakdown by content rating. Are you more likely to find a paid game in a rating bracket for teens or adults as they are a more likely target audience? Plot 5 shows that most apps in the combined dataset are rated for everyone, with about 30% of those being paid apps. The other age ranges are split fairly evenly down the middle of free versus paid, meaning that if we do decide to develop an app targeted at a higher age range or with a higher content rating, it is more viable to use a pay per download model than with an app



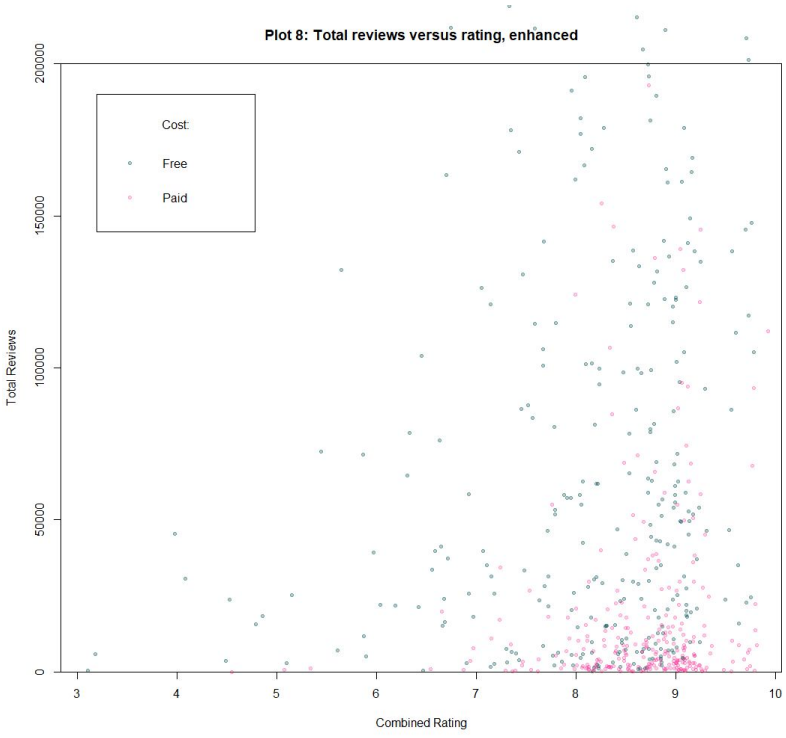
rated for all ages. The caveat to that is that there are far more apps designed for all ages, meaning there is a greater potential target audience with a lower rating. This could actually mean increased revenue if we were to implement a different revenue model into a free app rated for everyone.

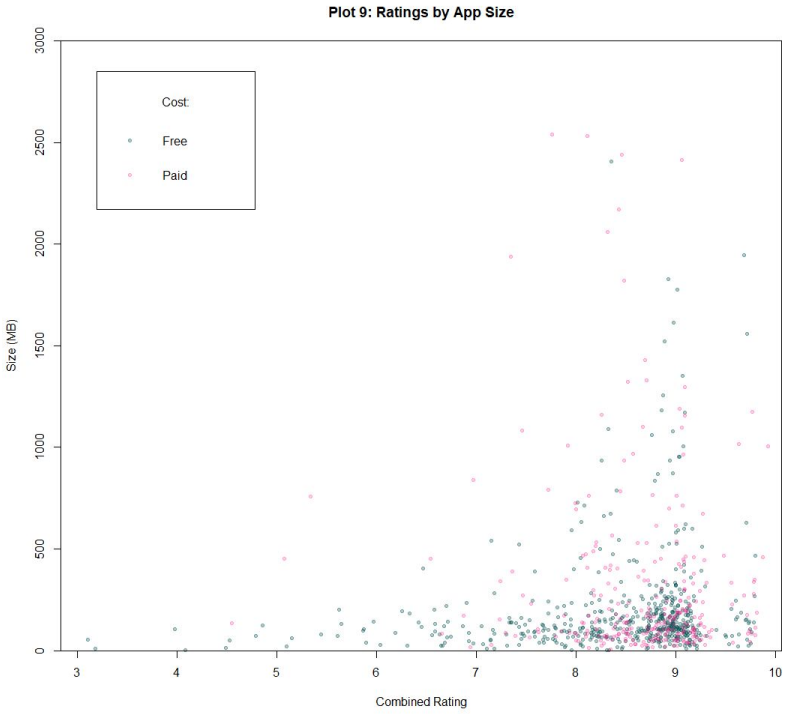
The other major indicators of success that I wanted to explore in my combined dataset were ratings and reviews of apps. The Apple rating data was rounded to the nearest .5 as you can see in Plot 6, while the Google rating data was a raw decimal. Included on the plot is a y=x reference line denoting what an equal rating between the two mobile stores would look like. Plot 6 shows that on average apps are higher rated in the Google Play store than in the Apple store. This seems to primarily be true for free apps. What we see from the paid apps, is a cluster in the top right corner of the graph, fairly equally bisected by the reference line. What this tells us is that while there is some variance between ratings between the two stores, paid apps are consistently highly rated on both platforms. Looking closer we can see that paid apps that are lower rated on the Apple store maintain a relatively higher performance in the Google





store, consistent with the trend shown by the free apps. This made me curious about the sample size of ratings and reviews for these apps, which is shown in Plot 7. This plot uses a combined rating of the sum of apple and google ratings, and a sum of the review count between the two platforms. This makes it vastly apparent that free apps in general have a much higher review count than paid apps. It also shows that apps with a high number of reviews are generally highly rated, with very few falling below 7.5/10 in a combined rating. My assumption from this is not that high numbers of reviews create a good app, but rather good apps generate a high number of reviews. In order to dive deeper into the paid portion of this plot, Plot 8 scales the y limit down from 4 million to 200,000. We see more of the same trend, but with far fewer total reviews for paid apps, with the bulk garnering fewer than 50,000 reviews. My hypothesis is that this is due to a greatly decreased number of downloads leading to a greatly decreased number of reviews. In comparison to the free apps shown in Plot 7, the bulk of the combined ratings for paid apps fall between 8/10 to 9.5/10, again indicating a very high quality in paid applications. Once again I find myself wishing there was more data on the





other revenue models or total revenue numbers for free download apps in order to better examine potential business models for our application development. A final characteristic between free and paid apps that I wanted to examine was application size. Plot 9 shows a breakdown of the size of each app compared against its combined rating from both platform stores. Most apps from the sample are under 500MB, but for those larger than 500MB there seems to be a range of ratings from 8/10 to 9.5/10. From this plot it is hard to differentiate much between free versus paid apps, however the conclusion that can be drawn is that there is correlation between larger apps and a generally high rating.

Conclusion

In conclusion, this data suggests that there are several directions to take with our potential app. I would caution that this data does not paint a full picture however, and there are attributes that were not appropriately present within the data sets that I would want to examine further to strongly recommend a direction. Specifically, more data regarding the revenue generation of free download apps would be very important to further analysis. There are many apps that are not classified as “paid” within this dataset that will still generate revenue. Additionally, a key metric regarding application success and audience that I would want to explore further is downloads. One of the datasets did include data on downloads, but in a largely segmented factor that was not easily analyzed or generalized across my sample. This alongside the ratings and reviews data that I explored here would give a better understanding of potential success of different categories of application.

If the goal is simply to create the best paid app, my recommendation based purely on this dataset is to proceed with the development of an education application. There is

strong precedent within that subset of the industry for a paid download, and educational apps hold a strong market share outside of games. If the goal is to maximize revenue with our application, I will wait on further data regarding revenue models in order to make a recommendation. I hypothesize that the best revenue model is a freemium model that makes use of in-app purchases based on the prevalence of games within the market, and the relatively small share of them that are paid to download.

Bibliography:

Prakash, G. and Koshy, J. (April 2019). Google PlayStore App analytics (Version 1) [Data set]. https://www.kaggle.com/gauthamp10/google-playstore-apps?select=Google-Playstore-Full.csv

Ramanathan. (July 2017). Mobile App Store (Version 7) [Data set]. https://www.kaggle.com/ramamet4/app-store-apple-data-set-10k-apps

Appendix A: R code for cleaning data

```
setwd("C:/NECPS/ALY6000/M6")
library(FSA)
library(FSAdata)
library(magrittr)
library(dplyr)
library(plotrix)
library(ggplot2)
library(moments)

###### CLEANING DATA

### Load 2 data sets from google play store and apple store
apple <- read.csv("AppleStore.csv")
gplay <- read.csv("googleplaystore32k.csv")
gplay2 <- read.csv("googleplaystore.csv")

### Convert google values to numeric
gplay$Rating <- as.character(gplay$Rating)
gplay$Rating <- as.numeric(gplay$Rating)

gplay$Price <- as.character(gplay$Price)
gplay$Price = as.numeric(gsub("\\$", "", gplay$Price))

### Content rating cleaning
gplay$Content.Rating <- as.character(gplay$Content.Rating)
gplay$Content.Rating[gplay$Content.Rating %in% c("$0.99","0","100,000+")] <- NA
gplay$Content.Rating[gplay$Content.Rating %in% "Adults only 18+"] <- "Mature 17+"
gplay$Content.Rating <- as.factor(gplay$Content.Rating)

apple$cont_rating <- as.character(apple$cont_rating)
apple$cont_rating[apple$cont_rating %in% "4+"] <- "Everyone"
apple$cont_rating[apple$cont_rating %in% "9+"] <- "Everyone 10+"
apple$cont_rating[apple$cont_rating %in% "12+"] <- "Teen"
apple$cont_rating[apple$cont_rating %in% "17+"] <- "Mature 17+"
apple$cont_rating <- as.factor(apple$cont_rating)

### Reviews
apple$rating_count_tot <- as.numeric(apple$rating_count_tot)

### Slice columns to be merged
Capple <- apple[c(3,4,6,7,9,12,13)]
Cgplay <- gplay[c(1,3:5,7,8)]

### Merge Capple and Cgplay
names(Capple)[1] <- "App Name"
names(Cgplay)[1] <- "App Name"
apps <- merge(Capple,Cgplay, by="App Name")
apps$`App Name` <- droplevels(apps$`App Name`)

### Drop duplicate records
apps = apps[order(apps[ ,"App Name"],-apps[,"Reviews"]),]
apps = apps[!duplicated(apps$`App Name`),]

### Additional cleaning
apps$prime_genre <- as.character(apps$prime_genre)
apps$prime_genre[apps$prime_genre %in% "Social Networking"] <- "Social Media"
apps$prime_genre <- as.factor(apps$prime_genre)

### Create Free vs. Paid factor
apps$freepaid <- NA
apps$freepaid <- as.character(apps$Price)
apps$freepaid[apps$freepaid %in% "0"] <- "Free"
apps$freepaid[!apps$freepaid %in% "Free"] <- "Paid"
apps$freepaid <- as.factor(apps$freepaid)

gplay$freepaid <- NA
gplay$freepaid <- as.character(gplay$Price)
gplay$freepaid[gplay$freepaid %in% "0"] <- "Free"
gplay$freepaid[!gplay$freepaid %in% "Free"] <- "Paid"
gplay$freepaid <- as.factor(gplay$freepaid)

### Clean name to fit on plots with LAS=2
apps$prime_genre <- as.character(apps$prime_genre)
apps$prime_genre[apps$prime_genre %in% "Social Networking"] <- "Social Media"
apps$prime_genre <- as.factor(apps$prime_genre)
```

Appendix B: Coding plots

```
###### PLOTS ######

### PLOT 1 ### CLEANING ###
genre <- apps$prime_genre
GENREcounts <- data.frame(genre)

### Plot 1 ### genre counts - COMBINED DATA
p1 <- ggplot(data=GENREcounts, aes(x=genre)) +
  geom_bar()+
  geom_text(stat='count', aes(label=..count..), vjust=-1, cex=3) +
  theme(plot.title = element_text(hjust = 0.5)) +
  theme(axis.text.x = element_text(angle = 90)) +
  scale_y_continuous(name="Counts",limits=c(0,600), breaks=seq(0,600,50)) +
  labs(title="Plot 1: Apps by Genre",x="Genre",y="Counts")
print(p1)



### PLOT 2 ### CLEANING ###
gamesdata <- as.character(genre)
gamesdata[!gamesdata %in% "Games"] <- "Other"
apps$gameother <- gamesdata

### Plot 2 ### genre - games vs non games and free vs paid - COMBINED DATA
p2 <- ggplot(apps) +
  geom_bar(aes(x=gameother, fill=freepaid)) +
  labs(title="Plot 2: Games vs. Other",x="Category",y="Counts") +
  theme(plot.title = element_text(hjust = 0.5)) +
  scale_y_continuous(name="Counts",limits=c(0,600), breaks=seq(0,600,50)) +
  scale_fill_discrete(name="Cost:")
print(p2)



### PLOT 3 ### CLEANING ###
games <- gplay
gamecats <- c("GAME_ACTION","GAME_ADVENTURE","GAME_ARCADE","GAME_BOARD","GAME_CARD",
              "GAME_CASINO","GAME_CASUAL","GAME_EDUCATIONAL","GAME_MUSIC","GAME_PUZZLE",
              "GAME_RACING","GAME_ROLE_PLAYING","GAME_SIMULATION","GAME_SPORTS",
              "GAME_STRATEGY","GAME_TRIVIA","GAME_WORD")
games <- games %<>% filterD(Category == gamecats)
gamecatsclean <- c("Action","Adventure","Arcade","Board","Card",
                   "Casino","Casual","Educational","Music","Puzzle",
                   "Racing","Role Playing","Simulation","Sports",
                   "Strategy","Trivia","Word")
levels(games$Category) <- gamecatsclean
GPfree <- games
GPfree <- GPfree %<>% filterD(freepaid == "Free")
GPpaid <- games
GPpaid <- GPpaid %<>% filterD(freepaid == "Paid")

### Plot 3 ### ggplot with free/paid of game categories
p3 <- ggplot(data=games, aes(x=Category)) +
  geom_bar(aes(x=Category, fill=freepaid)) +
  labs(title="Plot 3: Google Store Game Categories",x="Category",y="Counts") +
  theme(plot.title = element_text(hjust = 0.5)) +
  scale_fill_discrete(name="Cost:") +
  scale_y_continuous(name="Counts",limits=c(0,55), breaks=seq(0,60,5)) +
  geom_text(stat='count', aes(label=..count..), vjust=1, cex=3) +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
print(p3)



### PLOT 4 ### CLEANING ###
nongames <- apps
nongames <- nongames %<>% filterD(prime_genre != "Games")

### Plot 4 ### genre counts without games, free vs paid - COMBINED DATA
p4 <- ggplot(data=nongames, aes(x=prime_genre)) +
  geom_bar(aes(x=prime_genre, fill=freepaid)) +
  labs(title="Plot 4: Game Categories by Cost",x="Category",y="Counts") +
  theme(plot.title = element_text(hjust = 0.5)) +
  scale_fill_discrete(name="Cost:") +
  scale_y_continuous(name="Counts",limits=c(0,60), breaks=seq(0,60,5)) +
  geom_text(stat='count', aes(label=..count..), vjust=1, cex=3) +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
print(p4)



### PLOT 5 ### CLEANING ###
rated <- apps$Content.Rating
ratedcounts <- data.frame(rated)

### Plot 5 ### genre counts - COMBINED DATA
p5 <- ggplot(data=apps, aes(x=Content.Rating, fill=freepaid)) +
  geom_bar()+
  theme(plot.title = element_text(hjust = 0.5)) +
  theme(axis.text.x = element_text(angle = 90)) +
  scale_fill_discrete(name="Cost:") +
  scale_y_continuous(name="Counts",limits=c(0,550), breaks=seq(0,600,50)) +
  labs(title="Plot 5: Apps by Content Rating",x="Rating",y="Counts")
print(p5)

```

```
### PLOT 6 ### CLEANING ###
free <- apps
free <- free %<>% filterD(freepaid == "Free")
paid <- apps
paid <- paid %<>% filterD(freepaid == "Paid")

### Plot 6 ### Ratings for free apps versus paid apps - COMBINED DATA
plot(Rating~user_rating, data=free, pch=20, col=rgb(0,0.3,0.3,0.3),
     main="Plot 6: Apple vs. Google Store Ratings",
     ylab="Google Rating", xlab="Apple Rating",
     xlim=c(2,5), ylim=c(2,5))
abline(coef=c(0,1))
points(Rating~user_rating, data=paid, pch=20, col=rgb(1,0,0.5,0.2))
legend("bottomright", title="Cost:", inset=.05, c("Free", "Paid"),
       pch=20, col=c(rgb(0,0.3,0.3,0.3),rgb(1,0,0.5,0.2)))



### PLOT 7 ### CLEANING ###
free$totalreviews <- free$rating_count_tot + free$Reviews
paid$totalreviews <- paid$rating_count_tot + paid$Reviews
free$tenrating <- (free$user_rating + free$Rating)
paid$tenrating <- (paid$user_rating + paid$Rating)

### Plot 7 ### Number of reviews for free apps vs paid apps
plot(totalreviews~tenrating,
     data=free,
     pch=20,
     col=rgb(0,0.3,0.3,0.3),
     main="Plot 7: Total reviews versus rating",
     xlab="Combined Rating",
     ylab="Total Reviews",
     ylim=c(0,4000000),
     las=0)
points(totalreviews~tenrating, data=paid, pch=20, col=rgb(1,0,0.5,0.2))
legend("topleft", title="Cost:", inset=.05, c("Free", "Paid"),
       pch=20, col=c(rgb(0,0.3,0.3,0.3),rgb(1,0,0.5,0.2)))



### Plot 8 ### Zoom in on plot 7
plot(totalreviews~tenrating,
     data=free,
     pch=20,
     col=rgb(0,0.3,0.3,0.3),
     main="Plot 8: Total reviews versus rating, enhanced",
     xlab="Combined Rating",
     ylab="Total Reviews",
     ylim=c(0,200000),
     las=0)
points(totalreviews~tenrating, data=paid, pch=20, col=rgb(1,0,0.5,0.2))
legend("topleft", title="Cost:", inset=.05, c("Free", "Paid"),
       pch=20, col=c(rgb(0,0.3,0.3,0.3),rgb(1,0,0.5,0.2)))



### PLOT 9 ### CLEANING ###
apps$tenrating <- (apps$user_rating + apps$Rating)
apps$sizeMB <- (apps$size_bytes/1000000)

### Plot 9 ### size x rating
plot(sizeMB~tenrating,
     data=free,
     xlab="Combined Rating",
     ylab="Size (MB)",
     ylim=c(0,3000),
     main="Plot 9: Ratings by App Size",
     pch=20,
     col=rgb(0,0.3,0.3,0.3)
     )
points(sizeMB~tenrating, data=paid, pch=20, col=rgb(1,0,0.5,0.2))
legend("topleft", title="Cost:", inset=.05, c("Free", "Paid"),
       pch=20, col=c(rgb(0,0.3,0.3,0.3),rgb(1,0,0.5,0.2)))

```