What Makes a Video Game Popular?

Kevin Liu

November 20, 2017

Abstract

The goal of this analysis is to find some of the biggest factors that determine a game's popularity, as well as what helps them continue to have a significant player count. Specifically, which factors have the biggest impact on a game's current and lasting popularity.

Krista Lofgren (2017) of Big Fish Games states that "eSports," also known as competitive gaming may have an impact on the popularity of certain games, two of which are included in our data set (Dota 2 and Counter-Strike: Global Offensive), with both having a total of 118 and 851 tournaments, and 772 and 4349 player participants, respectively. Futhermore, Ben Casselmann (2015) of Five Thirty Eight also states that "eSports" is becoming popular at an alarming rate, with 58 million people watching competitive gaming in 2012, 74 million in 2013, and 89 million in 2014. A popular video game streaming website called "Twitch" garners 55 million active users, 1 million peak concurrent viewers per month, and 100 million unique viewers per month.

An article from The Economist (2014) explain that rising video game budgets, although allows developers to create better-looking games, also cause publishers to take less risks in the field of creativity, as there is more money at stake. Likewise, another article from Rock, Paper, Shotgun (2013) suggests that one reason for a vastly popular game *Dota 2* to have such a large player count is due to its competitive nature and direction.

Methods

The data was obtained from http://steamcharts.com/top. Note that the "peak" players are in regards to a 30-day timeframe. Additional information such as the genre, competitiveness, release date, and cost were found through https://en.wikipedia.org and http://store.steampowered.com. As we drew data directly from the Steam website, no game listed is a console-exclusive, as Steam is a PC-based gaming client. The main reason for using Steam as a means for data collection is because it is objectively the biggest provider and distributor of video games worldwide for the PC market. We have a dataset containing 40 games of varying genre, price, release date, whether they are competitive games, and their respective current player amount, peak monthly player amount, and total hours played by all players.

Analysis and Results

We first read the dataset into R. In order to simplify and have a more comprehensible plot, when plotting the data, we will represent hours and players in terms of thousands (ie. 87243 will be represented as 87.2).

```
library(tidyverse)
```

```
## Loading tidyverse: ggplot2
## Loading tidyverse: tibble
## Loading tidyverse: tidyr
## Loading tidyverse: readr
## Loading tidyverse: purrr
## Loading tidyverse: dplyr
```

```
## Conflicts with tidy packages
----------------------------------------------
```

```
## filter(): dplyr, stats
## lag():    dplyr, stats
```

```
library(smmr)
library(ggplot2)
#import steam charts dataset
steamcharts=read_csv("steamcharts.txt")
```

```
## Parsed with column specification:
## cols(
##   Title = col_character(),
##   Current = col_integer(),
##   Peak = col_integer(),
##   Hours = col_integer(),
##   Genre = col_character(),
##   Competitive = col_character(),
##   `Release Date` = col_character(),
##   Cost = col_double()
## )
```

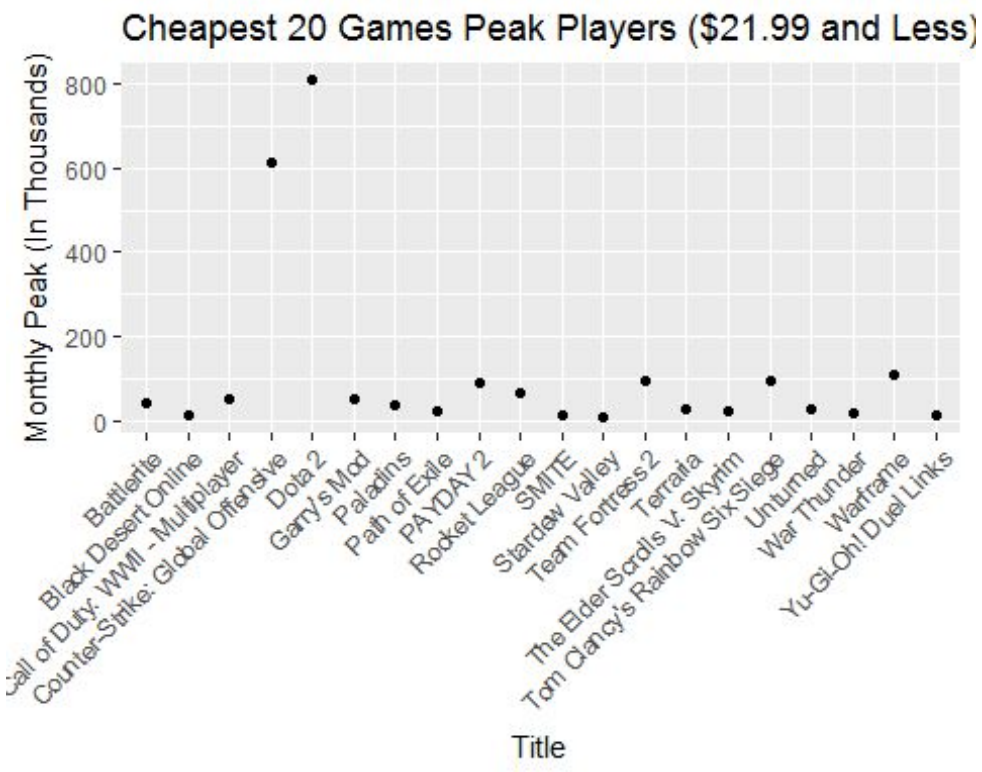
```
#display steam charts dataset
steamcharts
```

```
## # A tibble: 40 x 8
##                               Title Current    Peak     Hours    Genre
##                               <chr>   <int>   <int>     <int>    <chr>
##  1    PLAYERUNKNOWN'S BATTLEGROUNDS  435904 2866566 908580926  Shooter
##  2                          Dota 2  310166  807834 345842688     MOBA
##  3 Counter-Strike: Global Offensive  203780  610301 237387797  Shooter
##  4                  Team Fortress 2   49278   97248  40693708  Shooter
##  5                         Warframe   45149  111079  46019398      RPG
##  6   Tom Clancy's Rainbow Six Siege   38613   96326  24069565  Shooter
```

```
##  7                         PAYDAY 2   36158   92426  31037313  Shooter
##  8                     Rocket League   29062   70004  23131336   Sports
##  9             ARK: Survival Evolved   28575   63378  25588310  Sandbox
## 10          Sid Meier's Civilization V   26172   47074  18807141 Strategy
## # ... with 30 more rows, and 3 more variables: Competitive <chr>, `Release
## #   Date` <chr>, Cost <dbl>
```

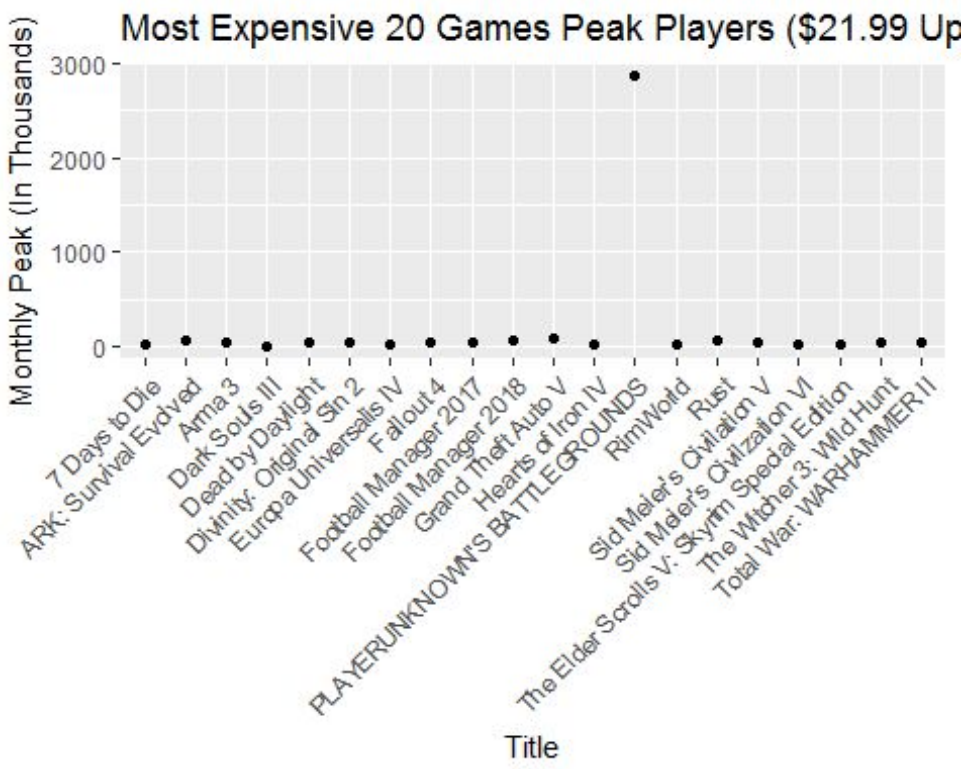
The data is ordered by decreasing amount of current players (on November 20, 2017). Before we proceed, we should note that intuitively, there are many possible factors that could affect a game's popularity.

```
#graph scatterplots to compare cheapest games peak players vs most expensive
games peak players
bycost=steamcharts %>% arrange(Cost)
cheap20=head(bycost,n=20)
exp20=tail(bycost,n=20)
#convert to thousands
n_sc=steamcharts %>% mutate(Current=(Current/1000), Peak=(Peak/1000),
Hours=(Hours/1000)) %>% arrange(Cost)
all_cheapest20r=head(n_sc,n=20)
all_me20r=tail(n_sc,n=20)
plot1=ggplot(all_cheapest20r,aes(x=all_cheapest20r$Title,y=all_cheapest20r$Pe
ak))+geom_point()
plot1=plot1+labs(title="Cheapest 20 Games Peak Players ($21.99 and
Less)",x="Title",y="Monthly Peak (In Thousands)")
#make x-axis easier to read
plot1=plot1+theme(axis.text.x=element_text(angle=45, vjust=1, hjust=1))
plot1
```



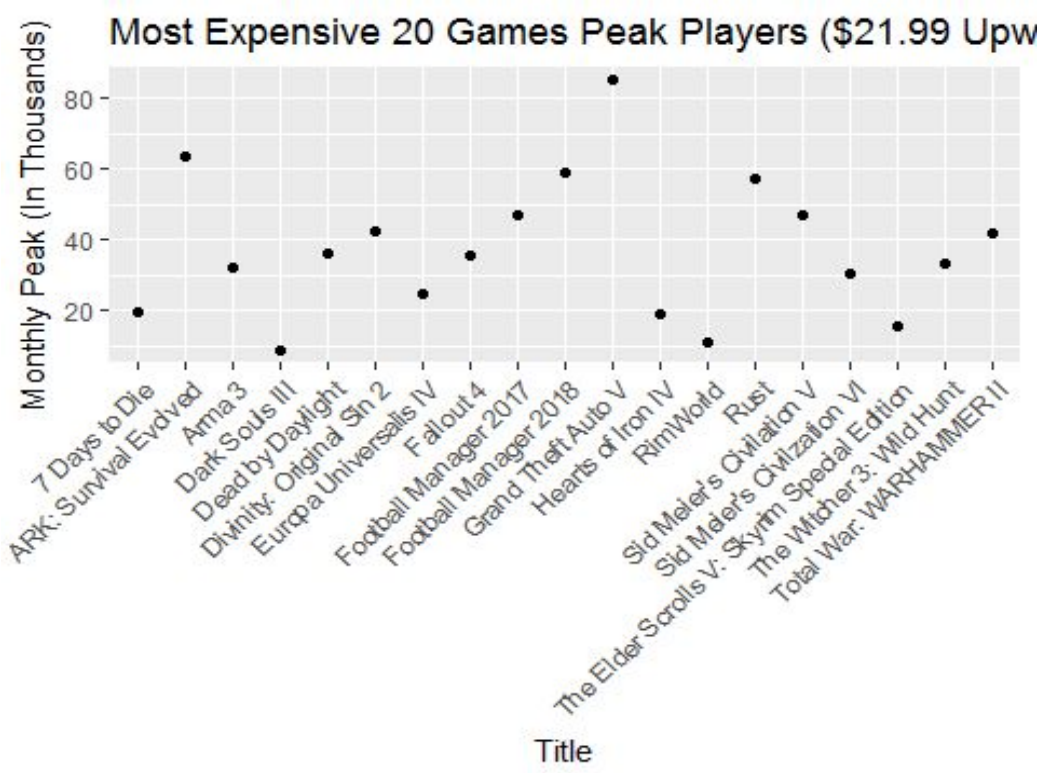
It seems that the monthly peak of players for the cheaper games, excluding the two outliers (Counter-Strike: Global Offensive and Dota 2) hover around 10 to 100 thousand, with most being around 10 to 50 thousand. We will compare these results to the games that are more expensive ($21.99 and over).

```
plot2=ggplot(all_me20r,aes(x=all_me20r$Title,y=all_me20r$Peak))+geom_point()
plot2=plot2+labs(title="Most Expensive 20 Games Peak Players ($21.99
Upwards)",x="Title",y="Monthly Peak (In Thousands)")
plot2=plot2+theme(axis.text.x=element_text(angle=45, vjust=1, hjust=1))
plot2
```



At first, it may seem like the distribution of peak players is uniform with the exception of PLAYERUNKNOWN'S BATTLEGROUNDS, but it is because of this outlier that causes the distribution to *seem* uniform, because the scaling of the y-axis must be extremely large in order to display the one datapoint. The outlier causes it to be quite difficult to actually see the range of peak players for the rest of the 19 games, so we'll remove it from the plot for now.

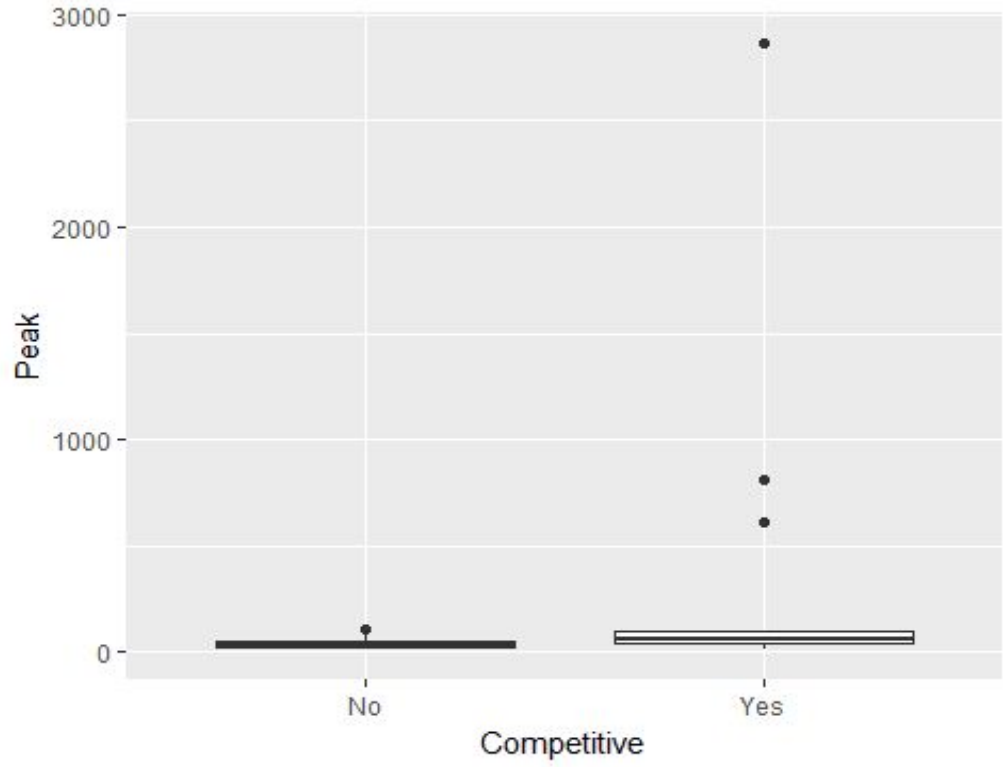
```
#remove the outlier from the plot
new_all_me20r=all_me20r[-6,]
#replot the data
new_plot2=ggplot(new_all_me20r,aes(x=new_all_me20r$Title,y=new_all_me20r$Peak
))+geom_point()
new_plot2=new_plot2+labs(title="Most Expensive 20 Games Peak Players ($21.99
Upwards)",x="Title",y="Monthly Peak (In Thousands)")
new_plot2=new_plot2+theme(axis.text.x=element_text(angle=45, vjust=1,
hjust=1))
new_plot2
```



Most of the more expensive games have peak players hovering around 25 to 60 thousand players monthly. Comparing the two scatterplots, we can see that the more expensive games have a higher monthly peak player count compared to cheaper games, presumably due to the fact that more expensive games usually have higher production value, and tend to be newer (as older games tend to drop in price after a while). A majority of the games in the cheaper price range were released between 2011 and 2015, whereas a majority of the games in the expensive price range were released between 2015 and 2017.

Now we will compare the plots of the competitive games against the non-competitive games. We will include all the data, especially since Dota 2 and Counter-Strike: Global Offensive are foundational to the competitive gaming scene, and therefore are crucial points of data.

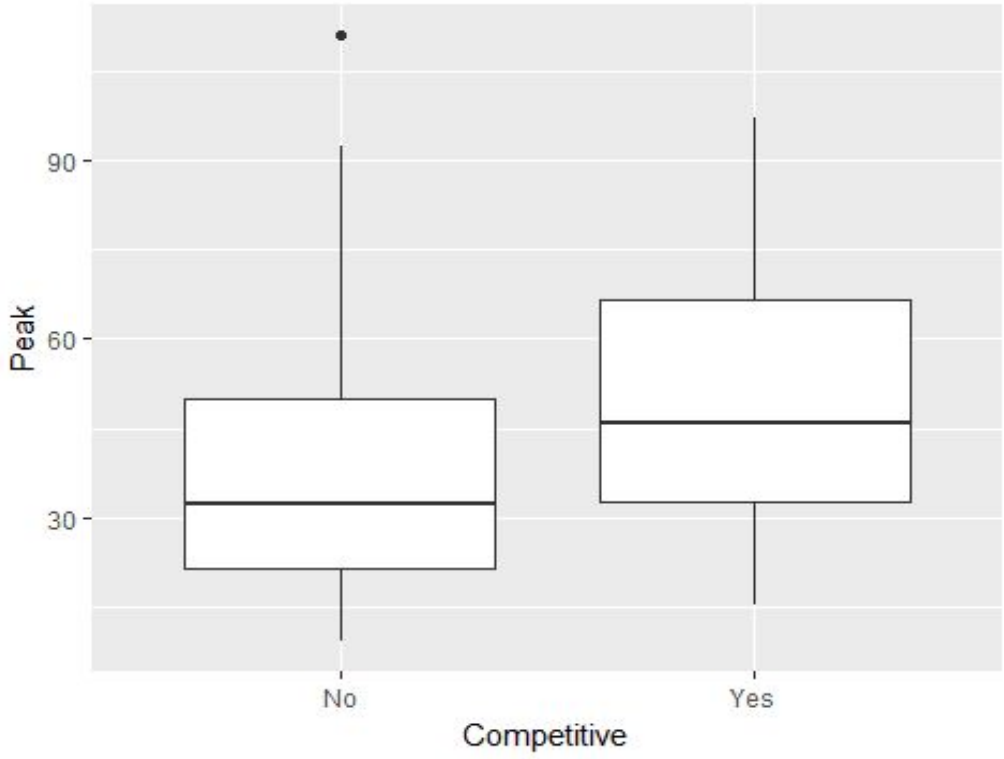
```
#boxplot by competitive
ggplot(n_sc,aes(x=Competitive,y=Peak))+geom_boxplot()
```



It is difficult to see the distribution of the two boxplots; however, it is important to note the three outliers on the side of the competitive games (Dota 2, Counter-Strike: Global Offensive, and PLAYERUNKNOWN'S BATTLEGROUNDS) have a *much* higher peak player count compared to the rest, as these are specific games that are extremely popular in the competitive gaming industry.

For the sake of clarity, we will display the plot with the removal of the three outliers from the competitive side.

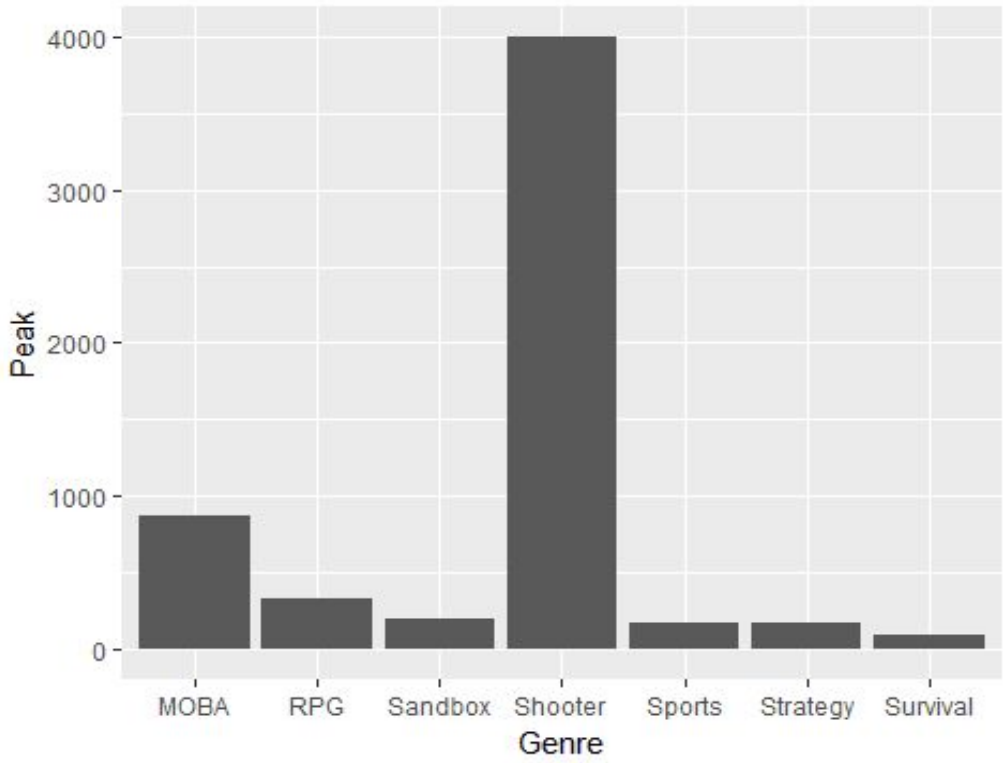
```
#remove outliers
n_sc0=n_sc[-26,]
n_sc0=n_sc0[-16,]
n_sc0=n_sc0[-1,]
ggplot(n_sc0,aes(x=Competitive,y=Peak))+geom_boxplot()
```



Here, we see a substantial difference between peak players of non-competitive games and competitive games. Competitive games have a higher minimum and maximum peak player count, as well as a higher median. This could be due to the rise of "eSports" as stated by the literature reviews, and an ever-growing playerbase attached to the competitive gaming industry.

Finally, we will compare the peak players of all genres, to see if genre has an impact on the popularity of a game. We include the outliers into this plot, as all of the data is important to see the cumulative peak players for each genre.

```
ggplot(data=n_sc,aes(x=Genre, y=Peak))+geom_bar(stat="identity")
```



It seems that the shooter genre has a substantial amount of the total monthly peak players out of all the other genres, which makes sense, as most of the games in the top 40 Steam charts games are in fact, shooting games. This could also tie-in with competitive games, since most competitive games in the data set are in fact, shooters.

We will also compare the median peak players of the two groups (cheap and expensive) instead of the mean, as the few outliers in both groups impact the calculate of the mean substantially, due to how large they are in comparison with the rest of the data. The same applies for the median of peak players between competitive and non-competitive games.

```
#show last 20 games (most expensive) and first 20 games (most cheap)
all_cheapest20=head(steamcharts,n=20)
all_me20=tail(steamcharts,n=20)
#calculate the median of both groups
median_headplayers=median(all_cheapest20$Peak)
median_tailplayers=median(all_me20$Peak)
allmedian=median(n_sc$Peak)
cheap=all_cheapest20r$Peak
expensive=all_me20r$Peak
comp_sc=n_sc %>% arrange(Competitive)
comp_sc=comp_sc[28:40,]
nocomp_sc=n_sc %>% arrange(Competitive)
nocomp_sc=nocomp_sc[1:27,]
median(cheap)
```

```
## [1] 42.121
```

```
median(expensive)
```

```
## [1] 36.1145
```

```
median(comp_sc$Peak)
```

```
## [1] 56.174
```

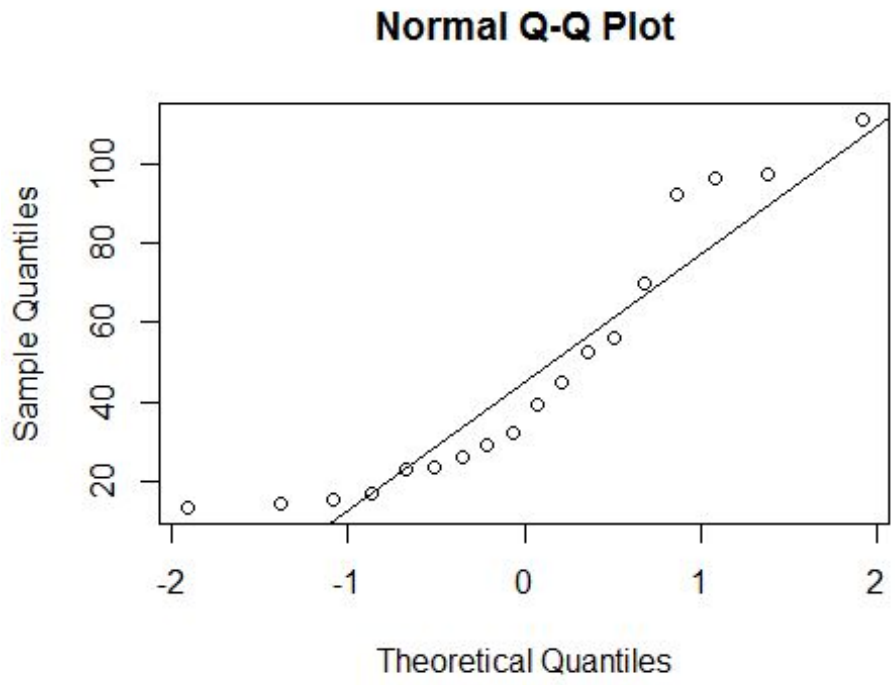
```
median(nocomp_sc$Peak)
```

```
## [1] 32.111
```

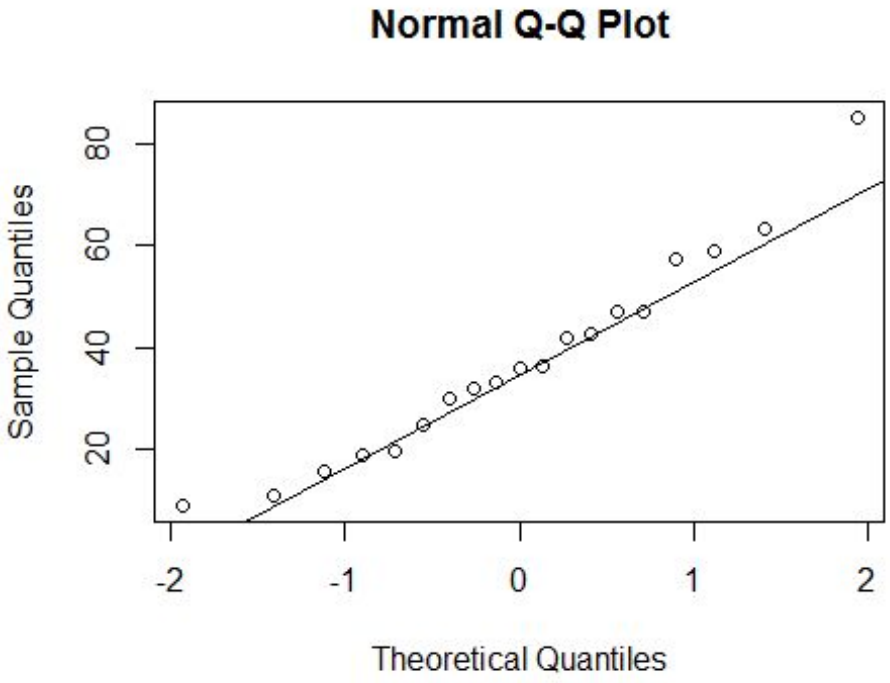
Using the calculated medians, we actually see that just barely, cheaper games have more peak median monthly players than expensive games (42 thousand and 36 thousand, respectively). Also, we can see that competitive games have a much higher peak monthly player count than non-competitive games (56 thousand to 32 thousand).

Before conducting any tests, we need to check to see if the data is approximately normally distributed across both groups, as this will affect our choice of testing the means or medians. This does not seem likely due to the number of outliers, but we will draw a normal quantile plot to be sure, except with the three outliers removed (Dota 2, Counter-Strike: Global Offensive, and PLAYERUNKNOWN'S BATTLEGROUNDS).

```
#draw normal quantile plots to assess normality
n_sc2=all_cheapest20r[-1,]
n_sc2=n_sc2[-15,]
qqnorm(n_sc2$Peak)
qqline(n_sc2$Peak)
```



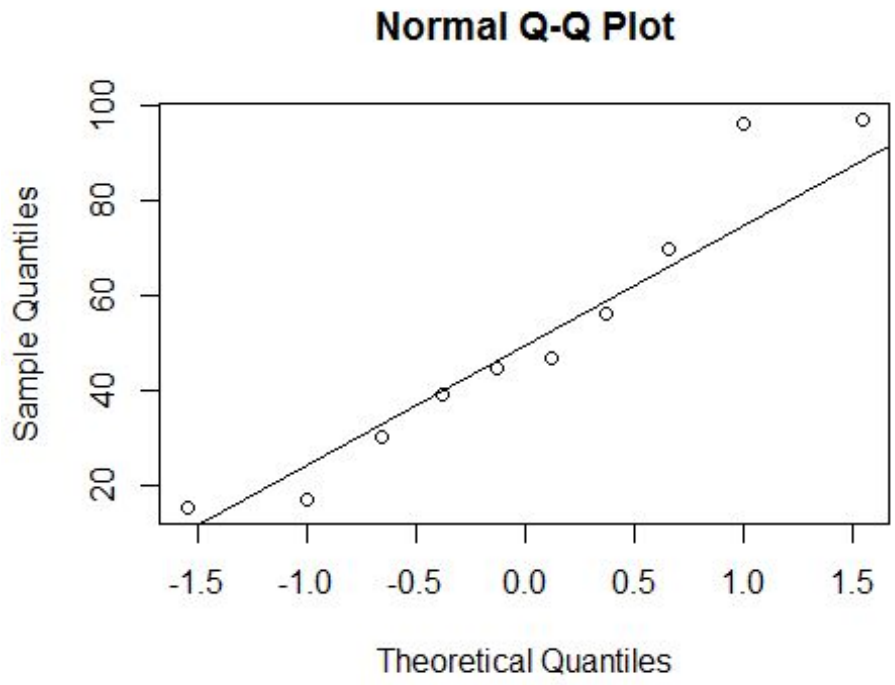

```
n_sc1=all_me20r[-6,]
qqnorm(n_sc1$Peak)
qqline(n_sc1$Peak)
```



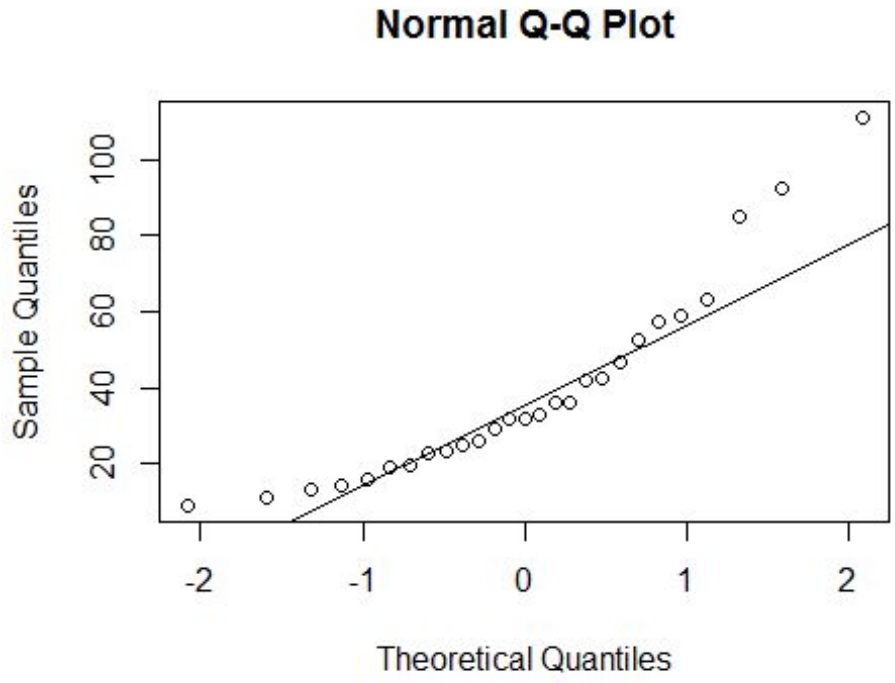
Both plots look surprisingly approximately normal, though not perfect by any means. Therefore, it is relatively okay to conduct a test on the means, given that we remove the three outstanding outliers first.

Now we look at the normal quantile plots for competitive and non-competitive games, with removal of the three outliers. We can remove these outliers as they aren't crucial to testing the average player count of the group.

```
comp_sc=comp_sc[-12,]
comp_sc=comp_sc[-8,]
comp_sc=comp_sc[-1,]
qqnorm(comp_sc$Peak)
qqline(comp_sc$Peak)
```




```r
qqnorm(nocomp_sc$Peak)
qqline(nocomp_sc$Peak)
```



Both plots start off having approximate normality, but strays off at the tail of the plot. As a result, it would be safer to conduct a test on the medians rather than mean, especially since with the removal of the three outliers in the competitive group, we have a much smaller sample size.

We first conduct a two-sample t-test of peak monthly players on the games' price, as we have two sample groups of cheap games and expensive games.

```
#conduct test on means
cheapmean=mean(n_sc2$Peak)
expmean=mean(n_sc1$Peak)
#difference in mean players is 10 thousand, with cheaper games having more
players
cheap1=c(n_sc2$Peak)
exp1=c(n_sc1$Peak)
newdf=data.frame(Peak=c(cheap1,exp1),
Cost=factor(rep(c("Cheap","Expensive"),times=c(length(cheap1),length(exp1))))
)
t.test(Peak~Cost,data=newdf,alternative="two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  Peak by Cost
## t = 1.1311, df = 27.569, p-value = 0.2677
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -8.172269 28.296204
## sample estimates:
##     mean in group Cheap mean in group Expensive
##                47.43639                37.37442
```

Our p-vaue is 0.2677, and so there is no statistically significant difference between the means. However, we do see that the mean of the cheap sample group is higher than the expensive sample group by approximately 10 thousand players.

Now we will conduct Mood's median test on the competitive games against non-competitive games' median peak players, in order to see the ratio of games whose peak players are higher than the median peak of all games.

```
comp_sc_orig=n_sc %>% arrange(Competitive)
comp_sc_orig=comp_sc_orig[28:40,]
newdf2=data.frame(Peak=c(comp_sc_orig$Peak,nocomp_sc$Peak),Competitive=factor
(rep(c("Yes","No"),
times=c(length(comp_sc_orig$Peak),length(nocomp_sc$Peak)))))
#conduct median test using smmr
#for reference purposes, population median is 37.8675 thousand
median_test(newdf2,Peak,Competitive)
```

```
## $table
##      above
## group above below
##   No     10    17
##   Yes    10     3
##
## $test
##        what      value
## 1 statistic 5.58404558
## 2        df 1.00000000
## 3   P-value 0.01812481
```

From this, we see that 77% of competitive games have a median peak monthly player count above the population median of 37.87 thousand, whereas only 37% of non-competitive games have a median peak monthly player count above the population median. We also have a p-value of 0.0181, which is less than our alpha of 0.05. This implies that we reject the null hypothesis that the medians of both groups are equal. Our alternative hypothesis is that at least one median is different. With a p-value of 0.0181, we can also say that this difference is statistically significant.

Conclusions

It appears that some of the biggest factors in a game's popularity are price, and competitive nature of the game itself. Through the plotted data and tests, cheaper competitive games continue to garner more players than expensive, non-competitive games. Looking at the data frame, a large amount of competitive games are on the cheaper side of price, with many of them being absolutely free. On the other hand, although there are a number of free non-competitive games, a majority of them are quite expensive, costing well over $29.99. The medians and means of cheaper and competitive games are also higher than that of the expensive and non-competitive games.

Notably, the three constant outliers (Dota 2, PLAYERUNKNOWN'S BATTLEGROUND, and Counter-Strike: Global Offensive) are all competitive games, two of which are shooters, and have a *significantly* larger player count than any other game on the list.

These results make sense, especially in the context of today's increasing popularity of "eSports" and competitive gaming. It is unsurprising that with an extremely large following towards the competitive aspect of video gaming, that many players would gravitate towards that particular aspect. In terms of cheaper games being more popular than expensive games, this makes sense, as most video game players tend to be young, and are not financially capable of purchasing overly expensive games. On a more general scale, people tend to enjoy cheaper items in general (a game would have to be *very* good to justify somebody spending $79.99 on it). Expensive games usually are made by large-scale production teams, and in the video game development industry, large corporations spend much more money on the development of games, and tend to "play it safe," focusing on the graphical and visual aspects of a game, and prefer a more derivative approach to gameplay, as opposed to taking risks and being more creative. As a result, cheaper games have a much

smaller budget, and are able to be more creative, and focus on gameplay rather than graphics, which are more appealing to gamers.

These results are important for the gaming industry in many aspects including streaming culture, competitive gaming, and game development. Game developers especially should see this as a cue to incorporate a competitive aspect to their games, specifically one that many people can gain access to, focusing *not* on how good the game looks, but rather on how competitive it is. A video game that produces a clear winner and loser, rather than a story-driven game would provide the game with a lasting player base, as it is more likely for a player to continue playing a game where he/she is able to continuously beat different people around the world, than a player to play through a game only to experience its story once and never play it again.

Some glaring limitations of this study is the fact that we only used the top 40 games, when there is an *extremely* large amount of games worldwide, so our sample size is significantly small and is by no means generalizable to a large extent. We also only drew from Steam Charts, which only accounts for games that are on PC, and so there is no extension to console exclusives.

The take-away from this analysis is one that has been continuously stated by other articles in the literature review, which is that competitive gaming is growing at a rapid pace, and it is the optimal time for developers to cater to this demographic, in a way that is cost-efficient for the consumer.

Further research is unquestionably needed, as there are countless more factors that can affect a game's popularity, such as timing of release, promotional investments and advertising, company reputation and brand, and more. Eventually, with enough research and analysis (an ample amount), we could be able to formulate an algorithm for the development of a video game that is guaranteed to garner worldwide success, or at least, with 99% confidence.

Appendix

eSports: Competitive gaming has been on the growth since the rise of notable video games like *Starcraft*, *Dota*, and *Counter-Strike*, with many games being streamed to audiences all over the world. These games have their own tournaments that are held live, with large money prize-pools (a notable one being *Dota 2*, with their *The International* prize pool being 2 million US dollars).

Twitch: A streaming website mainly used for people to stream and watch other people playing various video games, typically on a competitive level. Some popular streamers include "loltyler1," "AdrenalineCyber," "imaqtpie," and "summit1g," garner over 1 million total views, and tens of thousands of viewers at any given time.

Steam: A video-game client for PC, created by Gabe Newell. Used for the distribution, purchase, and playing of a majority of video games developed for PC. The indisputably most popular platform for PC gaming.

RPG: Abbreviation for "Role-Playing Game," players essentially play as a character in a built world, and act as the character in the video game. Often is associated with MMORPG, short for "Massively Multiplayer Online Role-Playing Game," where players create their own character and play in a world inhabited by other players playing worldwide. Some examples include *World of Warcraft*, *Ragnarok*, and *Destiny*.

Large game developers: Includes *Nintendo*, *Capcom*, *Ubisoft*, and *Squarenix*.

Small game developers: Includes *PUBG Corporation*, *Team Silent*, and *Team Meat*.

References

Casselman, F. B. (2015, May 22). Resistance is futile: eSports is massive ... and growing. Retrieved November 24, 2017, from http://www.espn.com/espn/story/_/id/13059210/esports-massive-industry-growing

Lofgren, K. (2017, April 5). 2017 Video Game Trends and Statistics - Who's Playing What and Why? | Big Fish Blog. Retrieved November 24, 2017, from https://www.bigfishgames.com/blog/2017-video-game-trends-and-statistics-whos-playing-what-and-why/

Michael Cohen (2013, September 11). Why Is Dota 2 The Biggest Game On Steam? Retrieved November 25, 2017, from https://www.rockpapershotgun.com/2013/09/11/why-is-dota-2-the-biggest-game-on-steam/

Top Games By Current Players. (n.d.). Retrieved November 24, 2017, from http://steamcharts.com/top

Why video games are so expensive to develop. (2014, September 24). Retrieved November 25, 2017, from https://www.economist.com/blogs/economist-explains/2014/09/economist-explains-15