# Project 1 - What drives E-sports Prize Pools?

ArunK

10/18/2020

Videogames represent a large and growing proportion of the economy and consumer spending, largely due to the rise of E-sports; E-sports has drawn in competitive talent from across the world, as well as the attention of entrepreneurs, game developers, and viewers. As an avid E-sports fan, I've always been shocked by the huge cash prizes given out at tournaments, and wondered what contributes to a game's monetization in E-sports. I wanted to explore two other game performance metrics I felt would be intimately related with E-sports prize pools: consumer sales and game reviews. In order to get some insight to the relationship between a game's consumer success, audience reception, and E-sports monetization, I found three datasets containing relevant metrics for a number of game titles. The videogame sales dataset includes information on genre, publisher and number of copies sold by region (in millions) for over 15,000 games across over 100 platforms. The E-sports dataset has information on number of E-sports tournaments held, total cash prize pool and total online cash prize pool for 492 unique games. The review dataset has both Metacritic and User score for 10,451 unique titles; scores were pulled from the Metacritic site, and represent critical and popular reception respectively. All 3 datasets were found on Kaggle's data repository, posted by different individuals.

I initially expected to see a strong positive association between both consumer sales and reviews with E-sports distributions; as more people purchase and enjoy a game, I believe that the competitive pool would grow, enabling higher-skill competition and more viewership of potential events; both of these factors should incentivize tournament hosts and sponsors to offer larger prize-pools, thereby solidifying their own spot in the market and attracting new talent. Despite this belief, I feel that certain games might be more conducive than others to high E-sports prize pools; factors like genre or platform might inherently make games more competitive or appealing to viewers, which could result in increased prize distributions. I also believe that E-sports distributions and review score would increase for modern platforms; I believe that publishers for modern game platforms are more capable of incorporating user feedback when making games, and also would be able to quickly address issues through regular updates. Additionally, given the cash inflow to the E-sports industry as awareness and business interest grow, I'd expect prize-pools to increase to accommodate larger venues, higher competition, greater viewership, and attract new talent to a game. Furthermore, modern E-sports players are often able to compete in multiple games within a genre, and publishers would be hesitant to let competitors steal their player-base; the easiest way to combat this would be to offer larger prize pools than any prior game in the same genre. However, there are several cult-classic games that maintain their prevalence in E-sports and the videogame consumer community despite their age and relatively low prize pools, suggesting that E-sports prize pools and game sales may be unrelated in certain cases.

```
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------- tidyverse 1.3.0 --
## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.3     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0
## -- Conflicts ------------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
esports <- read.csv("GeneralEsportData.csv")
vg_sales <- read.csv("vg_sales.csv")
vg_reviews <- read.csv("result.csv")
inner_join(esports, vg_sales, by = c("Game" = "Name")) -> vg_dat

vg_dat %>% mutate(Name_Console = paste(Game, Platform, sep = "_")) -> vg_dat_cl
vg_reviews %>%mutate(Name_Console = paste(name, console, sep = "_")) -> vg_reviews_cl

inner_join(vg_dat_cl, vg_reviews_cl, by = "Name_Console") -> vg_dat_full
length(unique(esports$Game))
```

```
## [1] 492
```

```
glimpse(vg_reviews)
```

```
## Rows: 15,647
## Columns: 5
## $ metascore <int> 97, 97, 95, 94, 94, 93, 93, 92, 92, 91, 91, 91, 91, 90, 9...
## $ name      <chr> "Grand Theft Auto V", "Grand Theft Auto V", "The Last of ...
## $ console   <chr> "PS3", "X360", "PS3", "PS3", "PC", "X360", "WIIU", "3DS",...
## $ userscore <chr> "8.3", "8.3", "9.2", "8.5", "8.6", "8.5", "8.9", "9.1", "...
## $ date      <chr> "17-Sep-13", "17-Sep-13", "14-Jun-13", "26-Mar-13", "25-M...
```

```
vg_dat_full %>% select(-Year, -Genre.y, -Name_Console, -name, -date, -console) -> video_games

video_games %>% mutate(Genre = Genre.x) -> video_games
```

All three datasets were in wide format to begin which was kept for ease of joining; the sales and E-sports datasets were joined first on game to produce the `vg_dat` dataset. I then used the game and console variables in the `vg_dat` dataset and the review dataset to create the temporary variable `Name_Console`, which was then used to join the datasets; this allowed me to retain granularity in the console variable, since I felt this may be important for game performance. The original E-sports dataset includes information on 492 games, while the videogame sales dataset includes data on 16,598 games, and the review dataset contains information for 15,647 games. Since I'm primarily interested in the interaction of E-sports prize pools, game reception, and consumer sales, I chose to perform an inner join for each merge step; the resulting dataset includes games with E-sports prize pool, review and sales information. Including games that only appeared in two or fewer datasets would not be useful for my main topic of interest, and would unnecessarily increase the dataset size. The merge step produced a complete dataset that would be useful in examining the interaction of consumer success and E-sports prize pools; this dataset contained tournament and prize distribution data, as well as sales and review data by region and platform, for 111 unique games. The data source introduced a significant significant error to the dataset: the videogame sales data listed the `publisher` of `World of Tanks` for the XBox-360 as `"N/A"` instead of Wargaming.net, which was manually verified and fixed. The dataset was then made longer for relevant numeric variables prior to exploration.

```
## Conversion of Sales Data
video_games %>% mutate(NA_Sales = NA_Sales * 1000000, EU_Sales = EU_Sales * 1000000, JP_Sales = JP_Sale

video_games %>% mutate(userscore = as.numeric(userscore)*10, score_dev = (userscore - metascore)^2) -> v
video_games %>% mutate(ReleaseDate = as.character(ReleaseDate)) -> video_games
video_games %>% mutate(Dist_per_Tournament = TotalEarnings/TotalTournaments) -> video_games
video_games %>% pivot_longer(c(TotalEarnings:Rank, NA_Sales:Global_Sales, metascore, userscore, score_d


## Global Sales Breakdown
opts <- options(knitr.kable.NA = "")
```

```
video_games_long %>% select(Publisher, Metric, Value) %>% filter(Metric %in% c("Global_Sales","NA_Sales
```

Table 1: Sales Metric Distributions

| Metric | Total | StDev | Minimum | Maximum |
|---|---|---|---|---|
| Global_Sales | 422,530,000 | 3,338,738.5 | 10,000 | 14,980,000 |
| NA_Sales | 202,470,000 | 1,863,759.5 | 0 | 9,670,000 |
| EU_Sales | 144,900,000 | 1,181,347.6 | 0 | 6,060,000 |
| Other_Sales | 58,350,000 | 682,001.0 | 0 | 7,530,000 |
| JP_Sales | 16,790,000 | 216,739.7 | 0 | 1,870,000 |

```
video_games_long %>% select(Publisher, Metric, Value) %>% filter(Metric == "Global_Sales") %>% group_by
```

Table 2: Global Sales for Top 5 Publishers

| Publisher | Total | StDev | Minimum | Maximum |
|---|---|---|---|---|
| Activision | 155,600,000 | 4,946,252 | 20,000 | 14,760,000 |
| Electronic Arts | 92,250,000 | 2,299,689 | 30,000 | 8,490,000 |
| Microsoft Game Studios | 66,300,000 | 4,126,286 | 20,000 | 12,140,000 |
| Sony Computer Entertainment | 30,840,000 | 5,520,906 | 4,200,000 | 14,980,000 |
| Namco Bandai Games | 18,060,000 | 1,522,334 | 320,000 | 4,050,000 |

```
video_games_long %>% select(Genre, Metric, Value) %>% filter(Metric == "Global_Sales") %>% group_by(Gen
```

Table 3: Global Sales for Top 5 Genres

| Genre | Total | StDev | Minimum | Maximum |
|---|---|---|---|---|
| First-Person Shooter | 213,820,000 | 4,579,390 | 10,000 | 14,760,000 |
| Sports | 72,790,000 | 2,154,401 | 10,000 | 8,490,000 |
| Racing | 57,790,000 | 3,653,802 | 30,000 | 14,980,000 |
| Fighting Game | 53,700,000 | 1,487,525 | 20,000 | 7,070,000 |
| Third-Person Shooter | 19,130,000 | 3,187,396 | 20,000 | 6,760,000 |

```
## Genre Contribution for platforms
video_games_long %>% select(Genre, Platform, Metric, Value) %>% filter(Metric == "Global_Sales") %>%  g
```

Table 4: Global Sales for Top 10 Genre-Platform Pairs

| Genre | Platform | Total | StDev | Minimum | Maximum |
|---|---|---|---|---|---|
| First-Person Shooter | X360 | 110,370,000 | 5,528,375.5 | 340,000 | 14,760,000 |
| First-Person Shooter | PS3 | 66,340,000 | 4,973,378.5 | 350,000 | 13,460,000 |
| Racing | PS2 | 30,360,000 | 6,720,113.1 | 1,160,000 | 14,980,000 |
| Sports | PS4 | 25,580,000 | 3,035,551.6 | 200,000 | 8,490,000 |
| First-Person Shooter | PS4 | 21,540,000 | 2,319,853.4 | 950,000 | 7,600,000 |
| Racing | X360 | 20,660,000 | 1,732,016.4 | 1,290,000 | 5,510,000 |
| Third-Person Shooter | X360 | 19,110,000 | 343,947.7 | 6,110,000 | 6,760,000 |
| Fighting Game | PS2 | 16,310,000 | 1,584,653.5 | 40,000 | 4,050,000 |
| Sports | PS3 | 16,200,000 | 1,993,276.2 | 570,000 | 6,900,000 |
| Fighting Game | PS3 | 14,560,000 | 1,226,860.0 | 100,000 | 4,190,000 |

```
## E-sports Distribution Breakdown
video_games_long %>% select(Metric, Value) %>% filter(Metric %in% c("TotalEarnings","OnlineEarnings","T
```

Table 5: E-sports Earnings Distributions

| Metric | Average | StDev | Minimum | Maximum |
|---|---|---|---|---|
| TotalEarnings | 600,307.47 | 2,448,754.9 | 0 | 22,049,333 |
| OnlineEarnings | 103,173.55 | 635,456.7 | 0 | 5,019,639 |
| TotalPlayers | 128.45 | 410.9 | 0 | 3,494 |
| TotalTournaments | 66.19 | 233.0 | 0 | 2,607 |

```
video_games_long %>% select(Publisher, Metric, Value) %>% filter(Metric == "TotalEarnings") %>% group_by
```

Table 6: Total E-sports Earnings for Top 5 Publishers

| Publisher | Average | StDev | Minimum | Maximum |
|---|---|---|---|---|
| 505 Games | 4,598,705 | 5,202,202 | 4,901.26 | 9,103,505 |
| Wargaming.net | 3,656,497 | | 3,656,496.59 | 3,656,497 |
| Nintendo | 3,208,950 | | 3,208,949.74 | 3,208,950 |
| Activision | 1,696,628 | 4,912,199 | 50.00 | 22,049,333 |
| DreamCatcher Interactive | 1,015,000 | | 1,015,000.00 | 1,015,000 |

```
video_games_long %>% select(Platform, Metric, Value) %>% filter(Metric == "TotalEarnings") %>% group_by
```

Table 7: Total E-sports Earnings for Top 5 Platforms

| Platform | Average | StDev | Minimum | Maximum |
|---|---|---|---|---|
| PS4 | 1,822,807.6 | 4,816,584.62 | 0.0 | 22,049,333.3 |
| PC | 979,417.9 | 3,193,907.53 | 0.0 | 22,049,333.3 |
| GC | 543,405.8 | 1,305,912.30 | 50.0 | 3,208,949.7 |
| DS | 398,062.5 | 31,493.11 | 375,793.5 | 420,331.5 |
| X360 | 212,674.6 | 423,495.73 | 0.0 | 2,158,079.4 |

```
## Platform + Genre Distribution Breakdown
video_games_long %>% select(Genre, Platform, Metric, Value) %>% filter(Metric == "TotalEarnings") %>% g
```

Table 8: Total E-sports Earnings for Top 10 Genre-Platform Pairs

| Genre | Platform | Average | StDev | Minimum | Maximum |
|---|---|---|---|---|---|
| First-Person Shooter | PS4 | 4,445,973.0 | 8,683,319.5 | 46,235.07 | 22,049,333.3 |
| Fighting Game | GC | 1,604,499.9 | 2,269,034.8 | 50.00 | 3,208,949.7 |
| First-Person Shooter | PC | 1,440,032.1 | 4,372,846.0 | 4,901.26 | 22,049,333.3 |
| Sports | PC | 1,216,724.3 | 2,995,529.6 | 0.00 | 9,103,504.9 |
| Sports | PS4 | 1,151,631.1 | 2,830,778.9 | 0.00 | 9,103,504.9 |
| Fighting Game | PS4 | 709,903.3 | 904,616.2 | 1,994.00 | 2,202,726.4 |
| Fighting Game | PC | 704,571.9 | 1,024,294.5 | 7,938.22 | 2,202,726.4 |
| First-Person Shooter | X360 | 514,844.3 | 627,620.1 | 5,000.00 | 2,158,079.4 |
| Role-Playing Game | PC | 476,005.5 | | 476,005.50 | 476,005.5 |

| Genre | Platform | Average | StDev | Minimum | Maximum |
|---|---|---|---|---|---|
| First-Person Shooter | PS3 | 417,908.3 | 534,327.2 | 5,000.00 | 1,594,185.2 |

```
## Review Breakdown
video_games_long %>% select(Metric, Value) %>% filter(Metric %in% c("userscore", "metascore"))  %>% grou
```

Table 9: Review Distributions

| Metric | Average | StDev | Minimum | Maximum |
|---|---|---|---|---|
| metascore | 81.76 | 7.2 | 56 | 95 |
| userscore | 68.58 | 15.9 | 21 | 91 |

```
video_games_long %>% select(Publisher, Metric, Value) %>% filter(Metric %in% c("userscore")) %>% group_
```

Table 10: User score for Top 5 Publishers

| Publisher | Average | StDev | Minimum | Maximum |
|---|---|---|---|---|
| Nintendo | 91.00 | | 91 | 91 |
| Virgin Interactive | 86.00 | | 86 | 86 |
| Unknown | 85.50 | 2.12 | 84 | 87 |
| Atari | 85.33 | 4.62 | 80 | 88 |
| Sony Computer Entertainment | 84.33 | 1.53 | 83 | 86 |

```
video_games_long %>% select(Publisher, Metric, Value) %>% filter(Metric %in% c("metascore")) %>% group_
```

Table 11: Critic Score for Top 5 Publishers

| Publisher | Average | StDev | Minimum | Maximum |
|---|---|---|---|---|
| Sony Computer Entertainment | 92.33 | 3.06 | 89 | 95 |
| Nintendo | 92.00 | | 92 | 92 |
| NCSoft | 90.00 | | 90 | 90 |
| Sega | 89.00 | 5.66 | 85 | 93 |
| Microsoft Game Studios | 85.59 | 8.67 | 66 | 94 |

```
video_games_long %>% select(Platform, Metric, Value) %>% filter(Metric %in% c("score_dev")) %>% group_by
```

Table 12: Squared Deviation of Critic and User Reviews by Platform

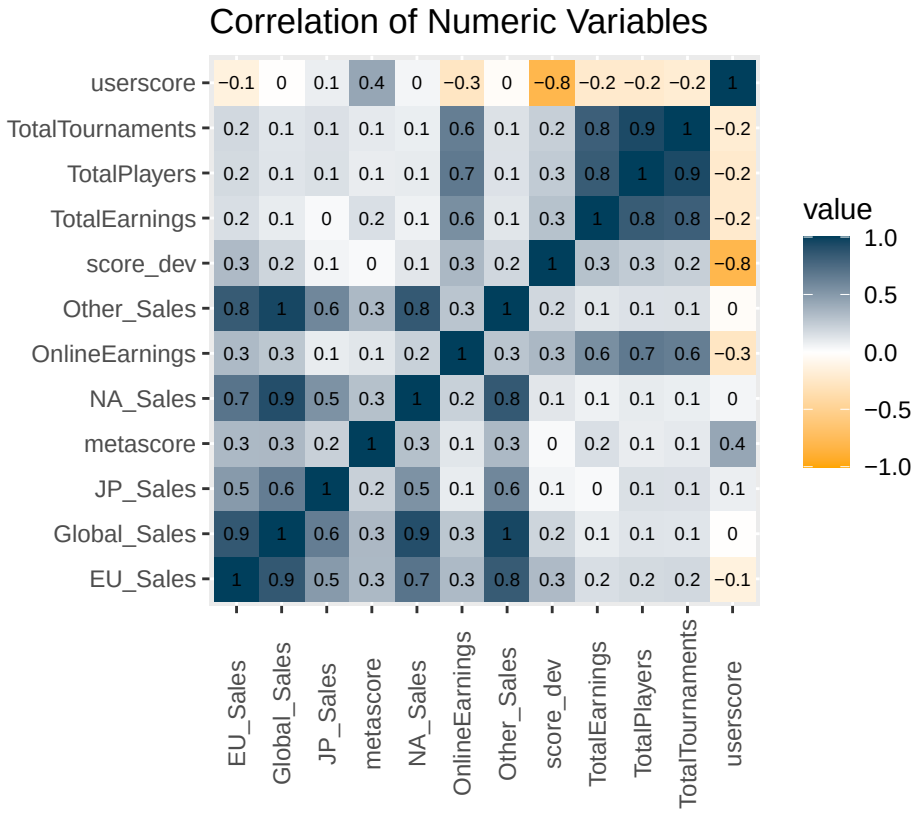
| Platform | Average | StDev |
|---|---|---|
| PC | 524.14 | 784.88 |
| PS4 | 518.00 | 549.06 |
| PS3 | 448.36 | 712.30 |
| X360 | 345.94 | 612.21 |
| GC | 257.67 | 281.44 |

```
## Correlation Matrix
video_games_cordat <- video_games %>% select_if(is.numeric) %>% select(-Dist_per_Tournament,-Rank)
video_games_cor <- video_games_cordat %>% cor(use = "pairwise.complete.obs", method = "spearman")
```

Taking a look at the distribution of sales metrics, we can see high variation evidenced by the global sales range of 10,000 to 14,980,000 copies; well-known publishers like Activision, EA, and Sony contribute to a majority of global sales, followed by cult-classic publishers like Bandai. Games in the first-person shooter genre dominated global sales with 213,820,000 total sales, followed by sports, racing, and fighting games; interestingly, the sales success of these genres seems to vary based on platform, which I'll explore below. These statistics suggest that overall, games successful in consumer markets tend to target widely-appealing categories like first-person shooters, racing, or fighting, and are usually published on consoles, such as the XBox-360, PS4, or PS3; the idea that publishers target certain mass-appeal platform-genre combinations is supported by the relatively low variation in global sales these top combination games experience.

Moving onto E-sports distributions, the massive variability in metrics is exacerbated by the highly skewed distribution of earnings, players, and number of tournaments. While 505 Games has the highest average E-sports prize pool of any publisher, the high standard deviation corresponds to the relative commercial failure of all games except their title Rocket League, which has become a staple in high-profile E-sports competitions. It seems like the highest average prize pools correspond to games made for the PS4 and PC; these high average prize pools might suggest an increase in spending by sponsors and hosts for tournaments played on modern game platforms, or could result from general increases in viewership. Regardless of platform, First-Person Shooter, Strategy, Fighting, and sports Games rank highly for average prize pools; this similarity with the genres that had high consumer success suggest that certain genres are more likely to appeal to game consumers, who then drive viewership and prize pools for the game up. Overall, user scores have a higher spread and lower average score than Metacritic scores; publishers like Nintendo, Sony, and Unknown rank highly for both scoring metrics. Games published on all platforms but Gamecube have high average disagreement between user and Metacritic scores, suggesting an area for further investigation.

```
library(ggpubr)
video_games_cor %>% as.data.frame %>%  rownames_to_column %>% pivot_longer(-1) %>%  na.omit %>%  ggplot
scale_fill_gradient2(low="#ffa600",high="#003f5c", limits = c(-1,1)) + theme(axis.text.x = element_text
```




The correlation heatmap shown indicates weak relationships between metrics for E-sports prize pools and consumer sales and reviews; this suggests that the three metrics of game performance may not be related in the way I initially thought, or that there are other factors which contribute more significantly to E-sports prize pool which are independent of consumer sales. The existence of strong relationships between variables from the same dataset, as well as the overall variability of all three performance attributes, seems to suggest that E-sports prize distributions, user reviews, and total consumer sales may be somewhat unrelated. Despite the weak association, there seems to be a negative correlation between user review score and metrics for E-sports prize distributions; this could indicate that games less popular among consumers have higher prize-distributions, potentially as publishers and sponsors attempt to incentivize player adoption. This idea is supported by the weak positive correlation between score deviation, calculated as the squared deviation between Metacritic and consumer scores, and metrics for consumer sales and E-sports distributions; it's feasible that games with split reviews generate loyal fanbases, which might drive up the number of sales and interest or competition in E-sports events. This correlation matrix was generated using `method = "spearman"`, to account for the unequal variance present in numeric variables, and the obvious violation of any normality assumptions.

```
library(ggExtra)
require(scales)
```

```
## Loading required package: scales
```

```
##
## Attaching package: 'scales'
```

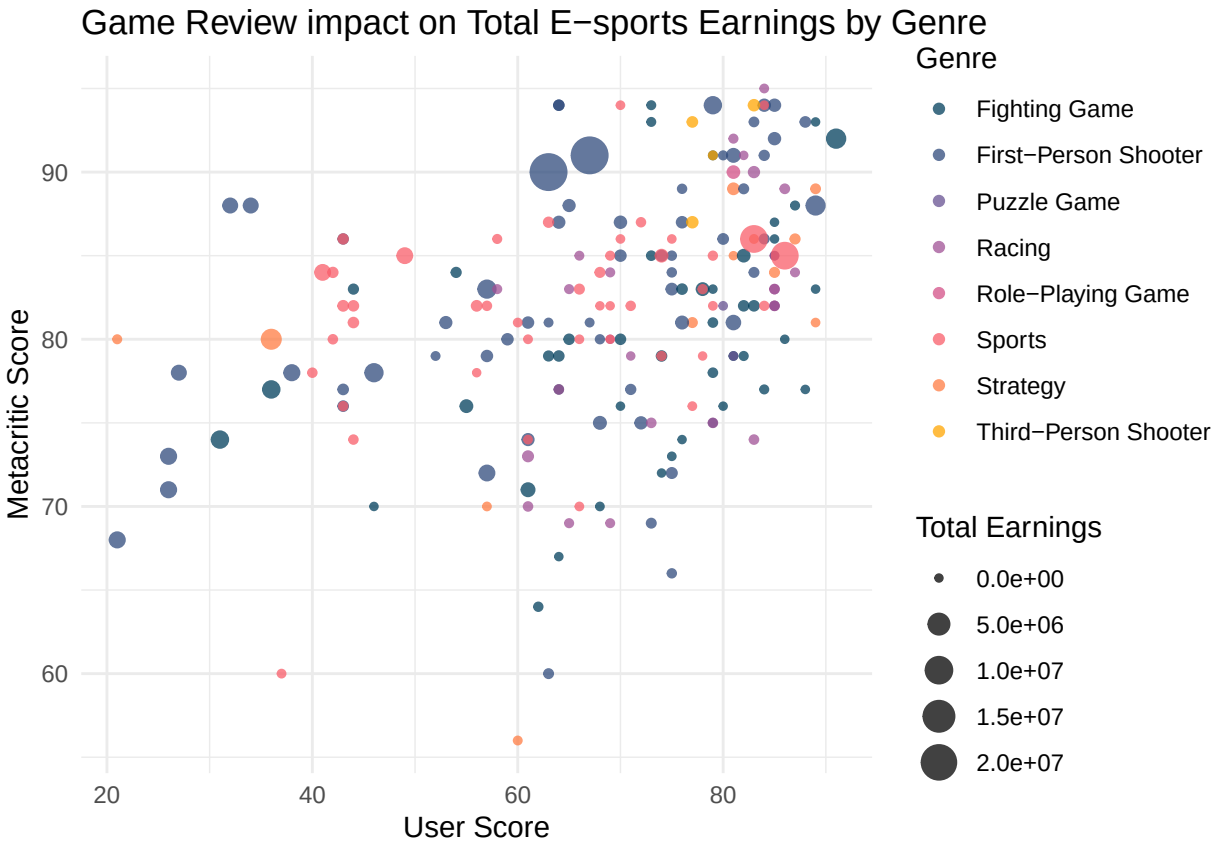
```
## The following object is masked from 'package:purrr':
##
##     discard
```

```
## The following object is masked from 'package:readr':
```

```
##
##     col_factor
```

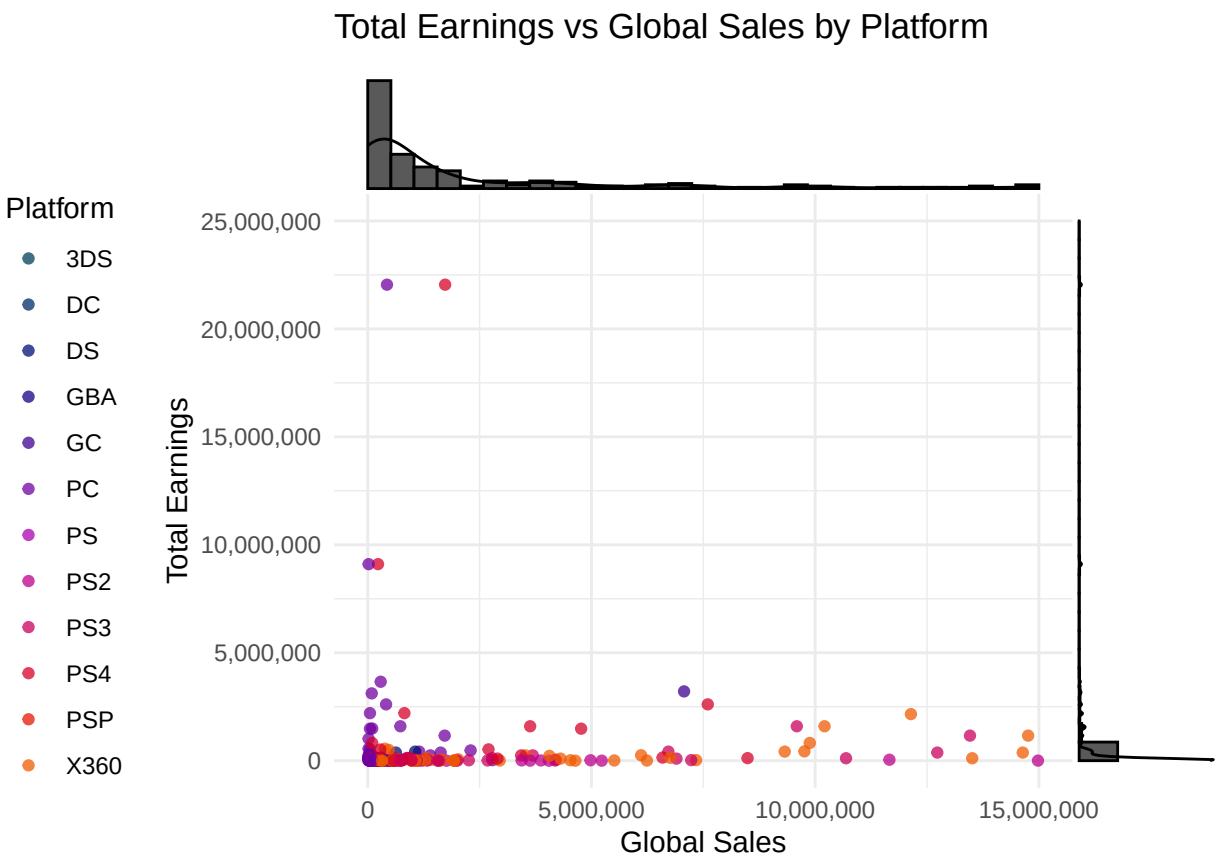
```
cust_pallete13 <- c("#003f5c","#002b6a","#000f77","#140085","#3f0092","#7100a0","#ab00ae","#bb008a","#c
cust_pallete8 <- c("#003f5c","#2f4b7c","#665191","#a05195","#d45087","#f95d6a","#ff7c43","#ffa600")
cust_pallete4 <- c("#003f5c","#7a5195","#ef5675","#ffa600")
cust_pallete3 <- c("#003f5c","#bc5090","#ffa600")
```

```
video_games %>% group_by(Genre) %>% ggplot(aes(x = userscore, y = metascore, color = Genre)) + geom_poi
```




Taking a look at how total E-sports earnings vary with game reviews, we can see a few key trends. First, there seems to be a positive association of Metacritic and user review scores, verified by their correlation of 0.4 shown in the correlation heatmap. Second, while top earning games seem to have higher critic and user reviews, it's important to recognize that this may be due to the incredibly large variation in E-sports distributions; as shown in the E-sports distributions table, Total Earnings can range from $0 to $22,049,333 with standard deviation of $2,448,755. The existence of games with more moderate distributions across review scores seems to support the idea that the largest distributions simply happen to be somewhat well reviewed games. Finally, it seems that certain game genres tend to generate higher distributions; there are a number of games with moderate distributions and low user scores for fighting games, first-person shooters, role-playing games, and strategy games. This suggests that poor consumer reception can be offset by the game's genre, supporting the idea that intrinsic game characteristics contribute to E-sports playability and prize pools.

```
video_games %>% group_by(Platform) %>% select(Platform, TotalEarnings, Global_Sales) %>% ggplot(aes(y =
scatter %>% ggExtra::ggMarginal(type = "densigram")
```

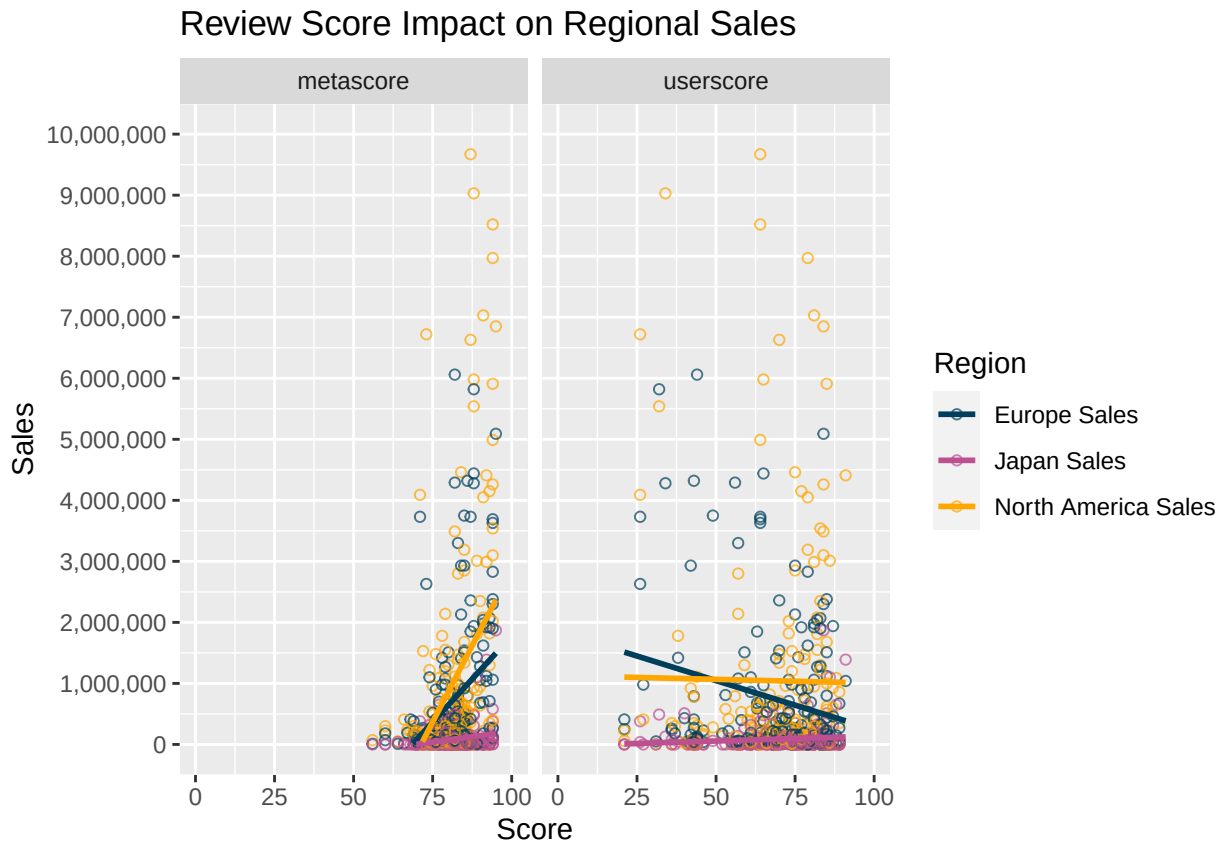



If we look at the relationship between total earnings and global sales, we see further evidence of their independence. Both global sales and total earnings are highly skewed right, with just 4 games generating over $9,000,000 in total distributions and 10 different games selling over 10 million copies; the majority of the 111 unique games in the dataset have total earnings less than $2,000,000 and copies sold less than 5 million. The lack of games that have large sales volume and high E-sports prize pools is directly contrary to my initial belief; it seems that games experience either E-sports or consumer-focused success. This idea might correspond to publishers selecting a primary market when developing a game, thereby enabling it to succeed among competitive players or in the casual marketplace. As shown in the global sales and E-sports earnings by genre-platform tables, First-Person Shooter Games on the XBox-360 or PS3 have the largest number of average global sales, while First Person Shooters on the PS4 have the largest average E-sports Earnings; this trend of certain platforms dominating E-sports distributions is further evidenced in the plot above. This behavior supports my initial idea that E-sports distributions increases as platforms become more modern; modern consoles offer higher quality of play and graphics, potentially attracting players and audiences. It's also possible that titles on modern platforms are seeing generally lower consumer sales: as prices for consoles and games continue to increase, consumers might be less willing to purchase new titles, and might instead watch others play these games through E-sports events. This interpretation suggests that games published on modern platforms are highly targeted at E-sports markets, while games on older platforms maintain consumer sales success due to accessibility. Similar to the interpretation presented for the prior plot, it's possible that this analysis is influenced by outlier games with high E-sports prize pools.

```
video_games %>% select(userscore, metascore, NA_Sales, EU_Sales, JP_Sales) %>% pivot_longer(userscore:m
```

```
## `geom_smooth()` using formula 'y ~ x'
```

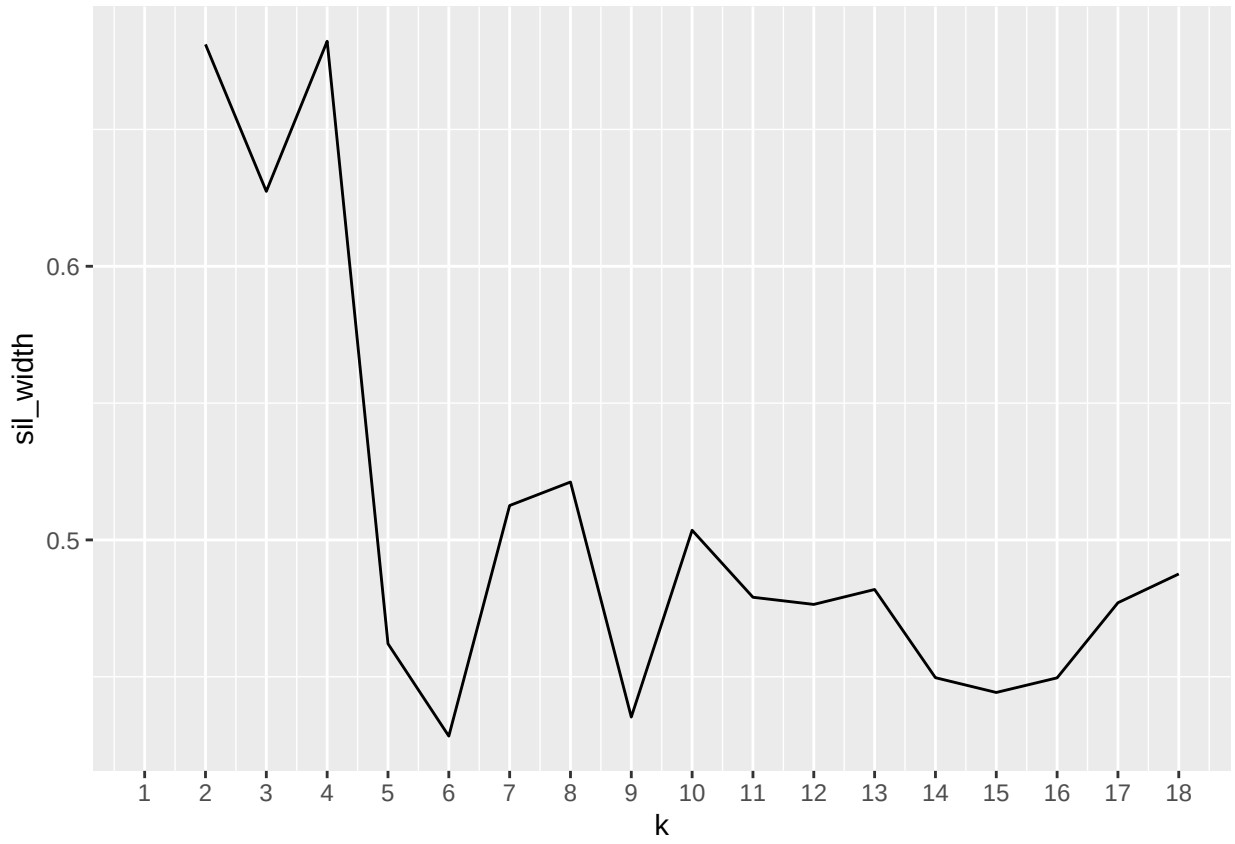
```
## Warning: Removed 86 rows containing missing values (geom_smooth).
```



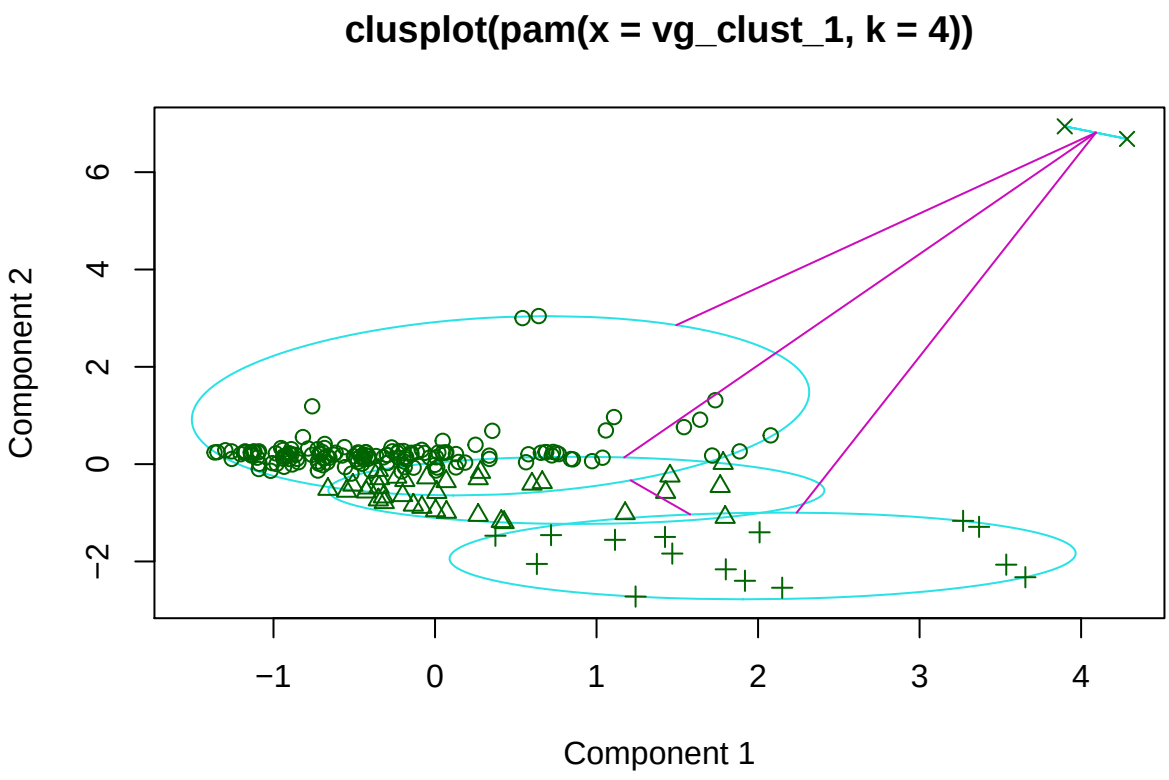

Now examining the effect of review scores on sales by region, we observe a few key trends. First, while user scores include a wide range of predictions, Metacritic scores are always between 50 and 100. This strong positive bias in Metacritic scores could be a result of the scoring methodology; Metacritic doesn't publicize the method they use to aggregate critic ratings, but even the worst performing games of all-time only have Metacritic ratings near 25. It's plausible that critics and users look for different things when rating a game and might differ on how important they find factors like ease of play, graphics, cross-play ability, or writing and directing. Furthermore, it's likely that any game that has historical E-sports activity and sales volume is at least moderately-well reviewed by critics; the disparity between consumer and Metacritic reviews was previously observed, and might be worth further investigation in subsequent studies. Second, Metacritic scores are positively associated with sales in Europe and North America, while user scores are negatively associated with sales in these areas. This suggests that consumers in European and North American markets are influenced by Metacritic reviews when choosing which games to purchase but seem to generate user reviews contrary to or independent of Metacritic score after playing. Finally, both metacritic score and user score seem to be positively associated with sales in Japan. It's worth noting that Metacritic only has an English-version site, and primarily reviews English-language games; this might help explain the relatively higher impact of Metacritic and user reviews on sales in North America and Europe.

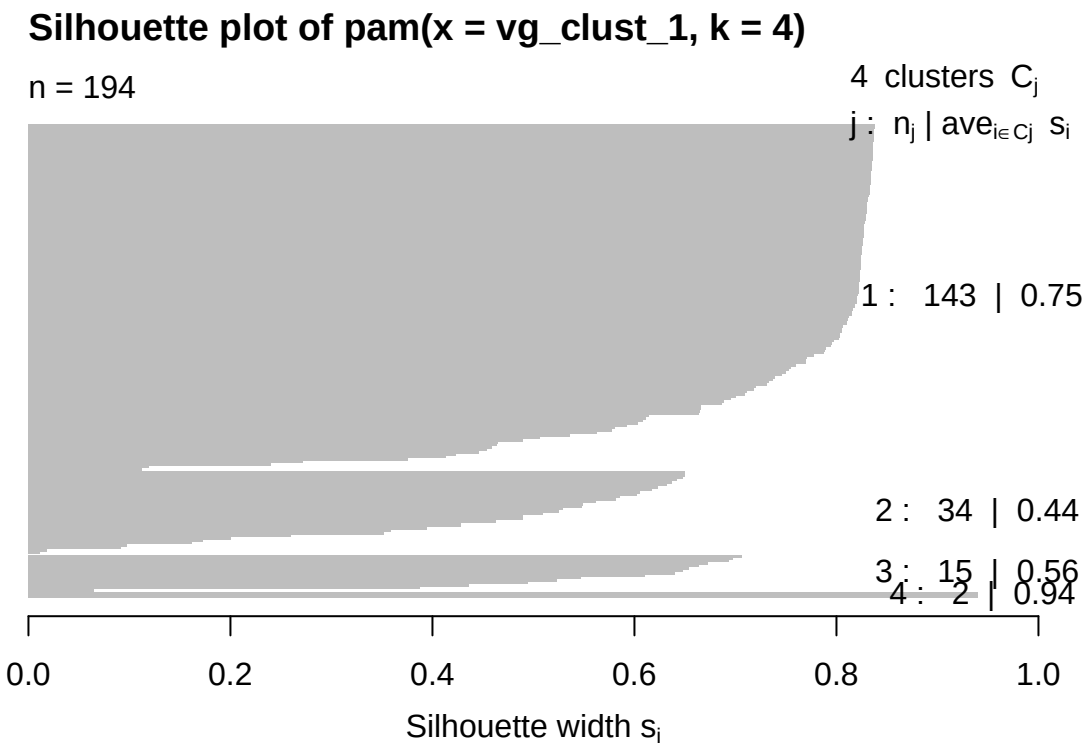
```
library(cluster)

vg_clust_1 <- video_games %>% select(TotalEarnings, Global_Sales, userscore)
sil_width <- c()
for (i in 1:18) {
  pam1 <- pam(vg_clust_1, k = i)
  sil_width[i] <- pam1$silinfo$avg.width
}

ggplot() + geom_line(aes(x=1:18, y = sil_width)) + scale_x_continuous(name = "k", breaks = 1:18)
```

```
## Warning: Removed 1 row(s) containing missing values (geom_path).
```



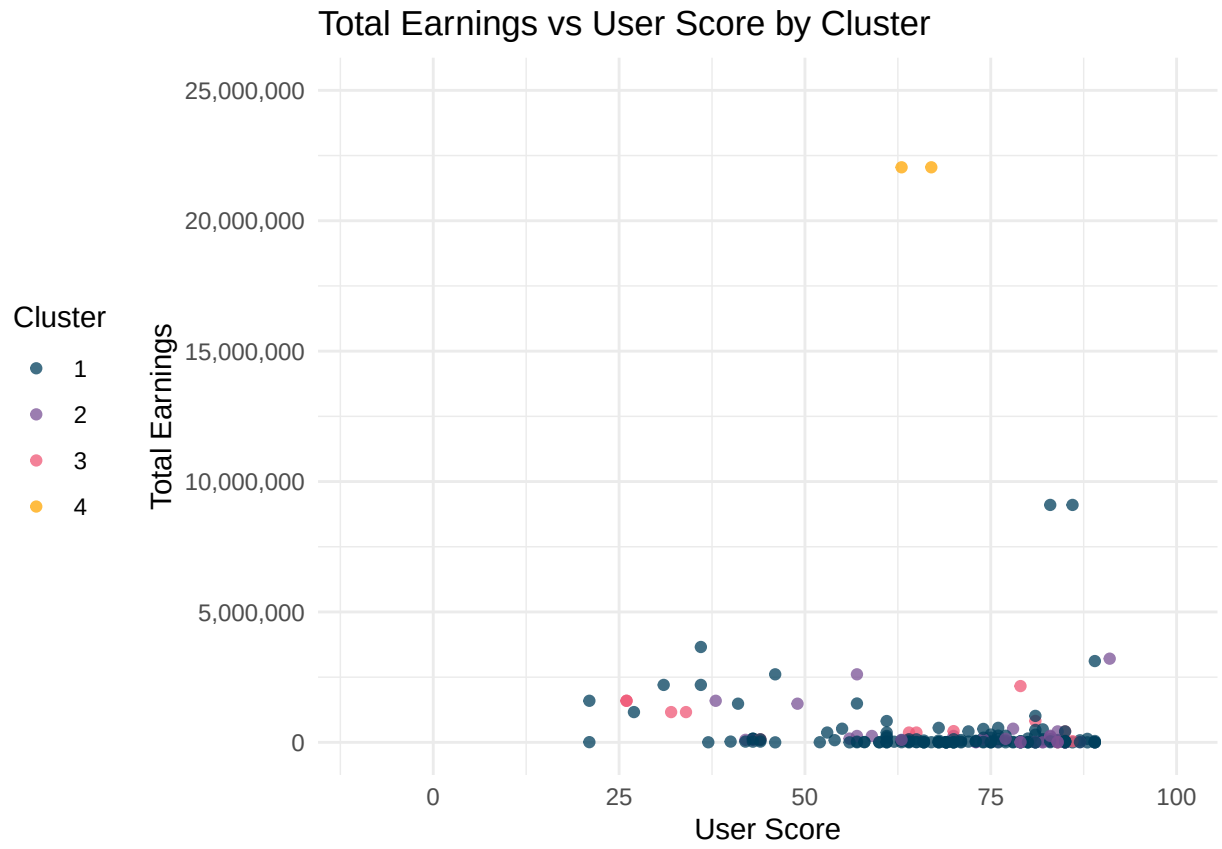

```
pam2 <- pam(vg_clust_1, k = 4)
plot(pam2)
```


clusplot(pam(x = vg_clust_1, k = 4))
Component 2
6
4
2
0
−2
−1
0
1
2
3
4
Component 1
These two components explain 70.06 % of the point variability.




```
vg_clust_1 %>% mutate(cluster = as.factor(pam2$clustering)) -> pamclust
pamclust %>% ggplot(aes(y = TotalEarnings, x = Global_Sales, color = cluster)) + geom_point(alpha = .75)
```

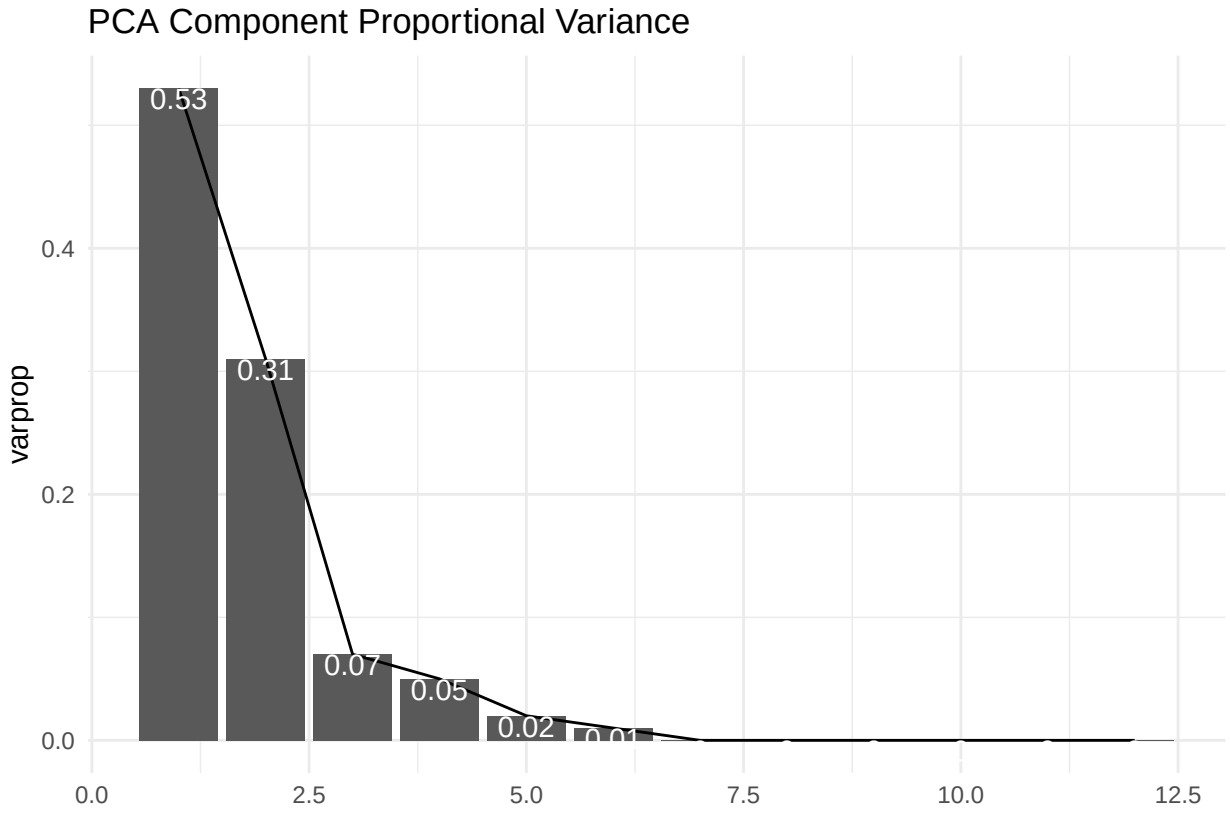

```
pamclust %>% ggplot(aes(y = TotalEarnings, x = userscore, color = cluster)) + geom_point(alpha = .75) +
```




```
video_games_cor %>%  scale() -> vg_pca
vg_pca1 <- princomp(vg_pca)
summary(vg_pca1)
```

```
## Importance of components:
##                           Comp.1    Comp.2     Comp.3    Comp.4     Comp.5
## Standard deviation     2.4050153 1.8443720 0.88670100 0.7632699 0.50092244
## Proportion of Variance 0.5258271 0.3092462 0.07147624 0.0529619 0.02281121
## Cumulative Proportion  0.5258271 0.8350733 0.90654959 0.9595115 0.98232270
##                           Comp.6      Comp.7      Comp.8       Comp.9
## Standard deviation     0.34836167 0.216555935 0.130959831 0.0805001718
## Proportion of Variance 0.01103235 0.004263316 0.001559134 0.0005891162
## Cumulative Proportion  0.99335505 0.997618364 0.999177499 0.9997666147
##                            Comp.10      Comp.11 Comp.12
## Standard deviation     0.0445760109 2.408770e-02       0
## Proportion of Variance 0.0001806382 5.274702e-05       0
## Cumulative Proportion  0.9999472530 1.000000e+00       1
```

```
eigval <- vg_pca1$sdev^2
varprop = round(eigval/sum(eigval), 2)

ggplot() + geom_bar(aes(y = varprop, x = 1:12), stat = "identity") + xlab("") + geom_path(aes(x = 1:12,y
```

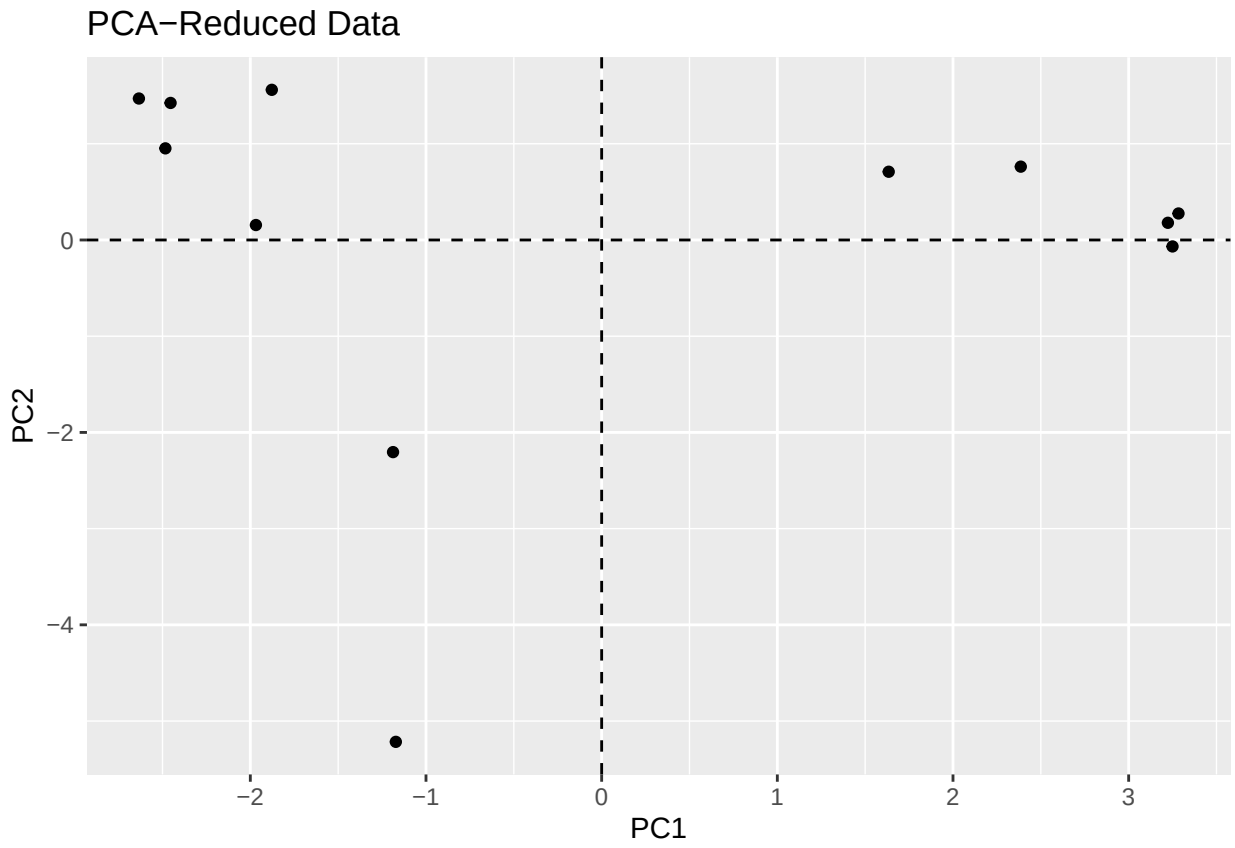



```
summary(vg_pca1, loadings = T)
```

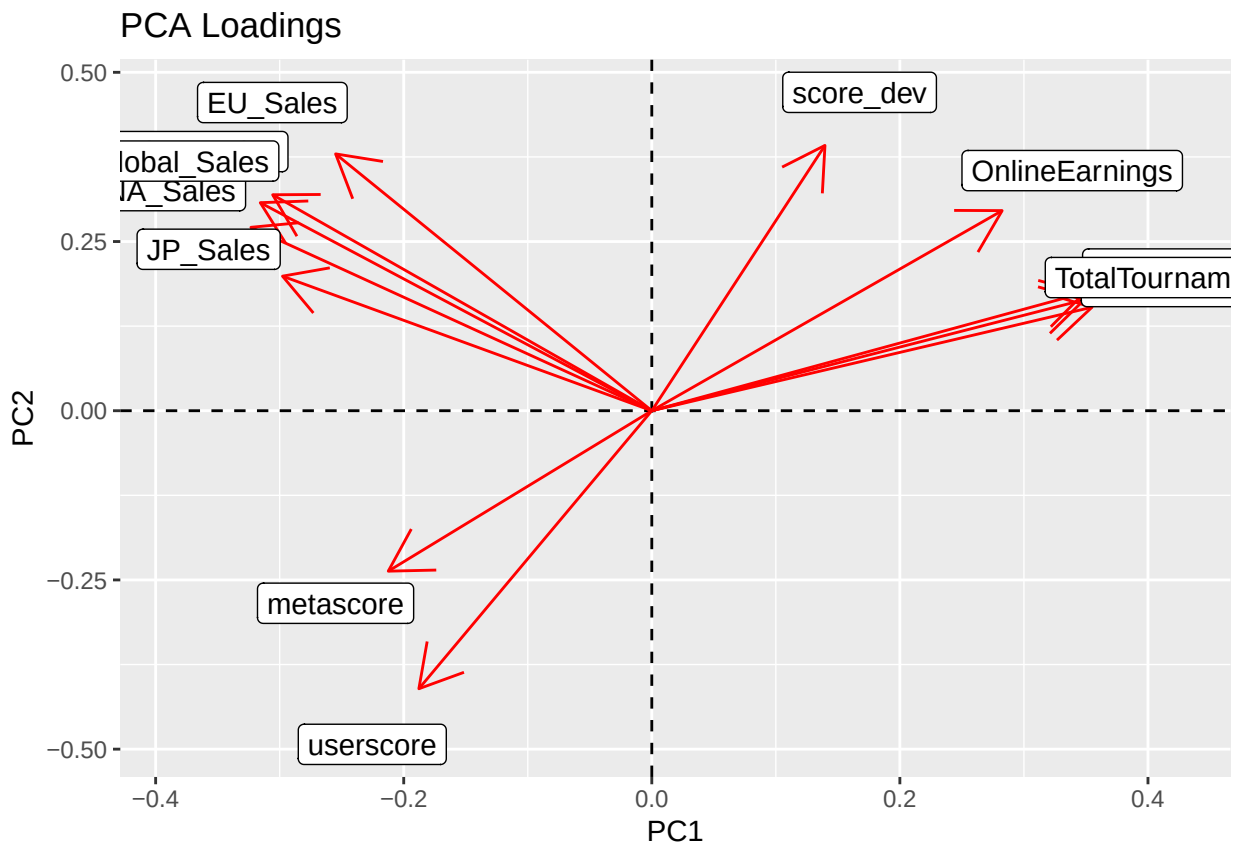
```
## Importance of components:
##                           Comp.1    Comp.2     Comp.3    Comp.4     Comp.5
## Standard deviation     2.4050153 1.8443720 0.88670100 0.7632699 0.50092244
## Proportion of Variance 0.5258271 0.3092462 0.07147624 0.0529619 0.02281121
## Cumulative Proportion  0.5258271 0.8350733 0.90654959 0.9595115 0.98232270
##                           Comp.6      Comp.7      Comp.8       Comp.9
## Standard deviation     0.34836167 0.216555935 0.130959831 0.0805001718
## Proportion of Variance 0.01103235 0.004263316 0.001559134 0.0005891162
## Cumulative Proportion  0.99335505 0.997618364 0.999177499 0.9997666147
##                           Comp.10      Comp.11 Comp.12
## Standard deviation     0.0445760109 2.408770e-02       0
## Proportion of Variance 0.0001806382 5.274702e-05       0
## Cumulative Proportion  0.9999472530 1.000000e+00       1
##
## Loadings:
##                  Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9
## TotalEarnings     0.355  0.153  0.284  0.117  0.137  0.440         0.721  0.110
## OnlineEarnings    0.282  0.296  0.214  0.156 -0.398 -0.742  0.106  0.199
## TotalPlayers      0.349  0.174  0.356         0.135                -0.334
## TotalTournaments  0.349  0.164  0.373         0.146  0.111        -0.507
## NA_Sales         -0.323  0.271  0.170        -0.241  0.228  0.609 -0.166  0.441
## EU_Sales         -0.255  0.379         0.183 -0.104        -0.745         0.404
## JP_Sales         -0.298  0.199  0.219 -0.400  0.696 -0.373         0.160
## Other_Sales      -0.306  0.319  0.153        -0.126  0.126               -0.772
```

```
## Global_Sales     -0.316  0.307  0.134        -0.108                       -0.105
## metascore        -0.212 -0.237  0.220  0.826  0.328 -0.119  0.104
## userscore        -0.188 -0.411  0.410        -0.146
## score_dev         0.140  0.392 -0.521  0.287  0.273         0.162
##                  Comp.10 Comp.11 Comp.12
## TotalEarnings
## OnlineEarnings
## TotalPlayers     -0.761
## TotalTournaments  0.633
## NA_Sales                -0.278
## EU_Sales                -0.133
## JP_Sales
## Other_Sales             -0.318   0.199
## Global_Sales             0.795  -0.339
## metascore                       -0.167
## userscore                0.315   0.703
## score_dev                0.245   0.554
```

```
data.frame(PC1 = vg_pca1$scores[,1], PC2 = vg_pca1$scores[,2] ) %>% ggplot(aes(PC1, PC2)) + geom_point()
```




```
vg_pca1$loadings[c(1:12), 1:2] %>% as.data.frame %>% rownames_to_column() %>% ggplot + geom_hline(aes(y
  geom_vline(aes(xintercept = 0), lty = 2) + ylab("PC2") + xlab("PC1") +
  geom_segment(aes(x=0,y=0, xend = Comp.1, yend = Comp.2), arrow = arrow(), col = "red") +
  geom_label(aes(x = Comp.1 *1.2, y = Comp.2 *1.2, label = rowname)) + ggtitle("PCA Loadings")
```




First performing PAM on the 18 numeric variables in the dataset, we can see that silhouette width is maximized with 4 clusters, after which consistency among clusters decreases; PAM with 4 clusters returns an average silhouette width of 0.68, suggesting the existence of a moderate structure in the clustering. The clustered scatterplot of total earnings and global sales shows support for my earlier interpretation; certain games have incredibly large E-sports earnings, while remaining games vary primarily in their global sales. The clustered plot of total earnings and user scores suggest that user scores are significantly less important in cluster formation, potentially owing to previously discussed variability. Next, plotting the relative contribution of PCA components in explaining observed variance, we find that 2 principle components should be included to describe the data. The first PC corresponds to high E-sports distribution metrics and low consumer sales and review metrics, while the second PC corresponds to high E-sports distribution metrics and consumer sales, but low score metrics. Visualizing the data along PC1 and PC2 indicates that most of the information present in the 18 numeric variables can be effectively reduced to the two shown. It's worth noting that given the loadings for PC1 and PC2, we see a single observation which is positive for PC1 and negative for PC2; this observation captures low total earnings and poor reviews, since global sales are nonnegative.

The plot of loadings provides further disproof to my initial thoughts on the relationship between game review scores, sales, and E-sports distributions. Total earnings and all sales metrics are near opposing, and review metrics are somewhat unrelated to both; the opposition of the global sales and total earnings loadings across the y-axis suggest that the two have antagonistic effects. The lack of points near `PC1 = 0` corroborates the idea that games are either successful in sales or in E-sports earnings, but not both. The lack of points in the second quadrant suggest scarcity of games that are unpopular with consumers and have high E-sports distributions. One calculated metric that is of interest is the score deviation; it seems that the deviation of metacritic and user scores contributes a novel impact compared to the initial review metrics. While the 3 game success datasets seem to be less related that I initially thought, I'm interested in exploring further factors that might be useful in explaining what drives E-sports prize pools, such as number of tournament sponsors, social media sentiment, or average player demographic; combined with outlier reduction, these additional factors may help inform more useful associations.