Steam Player Count Analysis

IST 687 Introduction to Data Science

December 2021

Team 4:

Taylor Sain

Cole Wood

Victoria Shu

Madeleine Gilad

Sahil Nanavaty

**Summary**

The data set we analyzed is Steam Charts which contains monthly average and peak player counts for around 1,000 games available on the video game digital distribution service known as Steam. The data set time frame contains 9 years' worth of data spanning from 2012 to November 2021. Some games have less than 9 years' worth of data due to their respective release dates.

When we explore the summary function of our data set, we can see that the monthly average player counts range from 0 to 1,584,886. Due to this large range, we will subset the data set a few different ways to gain some insight.

**Scope**

Our aim is to gain a better understanding of this market and to see which games are the most popular based on average/peak player counts and monthly gains/losses. In addition, we also hope to gain insight into the timeliness of player counts to see if there are times of the year in which a game release or update would be more favorable.

This sort of analysis could prove beneficial for an up-and-coming independent game studio hoping to feel out the highly competitive video game industry. Insight derived using our team’s methods could especially aid in determining consumer trends and ideal release schedules.

**Stakeholders**

* **Publishers**benefit from insight regarding optimal scheduling of releases and updates
* **Developers**benefit from insight regarding player preferences and trends

**Data Dictionary**

|  |  |
| --- | --- |
| Month | Month and Year |
| AvgPlayers | Average number of players playing the game |
| Gain | Exact change relative to previous month |
| PercentGain | Percent change relative to previous month |
| PeakPlayers | Peak player count of the month |
| AppID | Game’s unique identification number |
| Game | Game’s name/title |

**Data Questions**

**Which games are the most popular?**

There are a few ways to look at the games popularity so we will look at popularity by the most total players, average peak players, and average players in the last 3 months.

One way to measure success is by measuring the number of players per game. This dataset measures the average number of players per month as well as the peak players per month. Several of the top games are all competitive multiplayer games. After creating a subset, the data to show the sum of peak players per game we can see that there are 3 highly successful outliers.

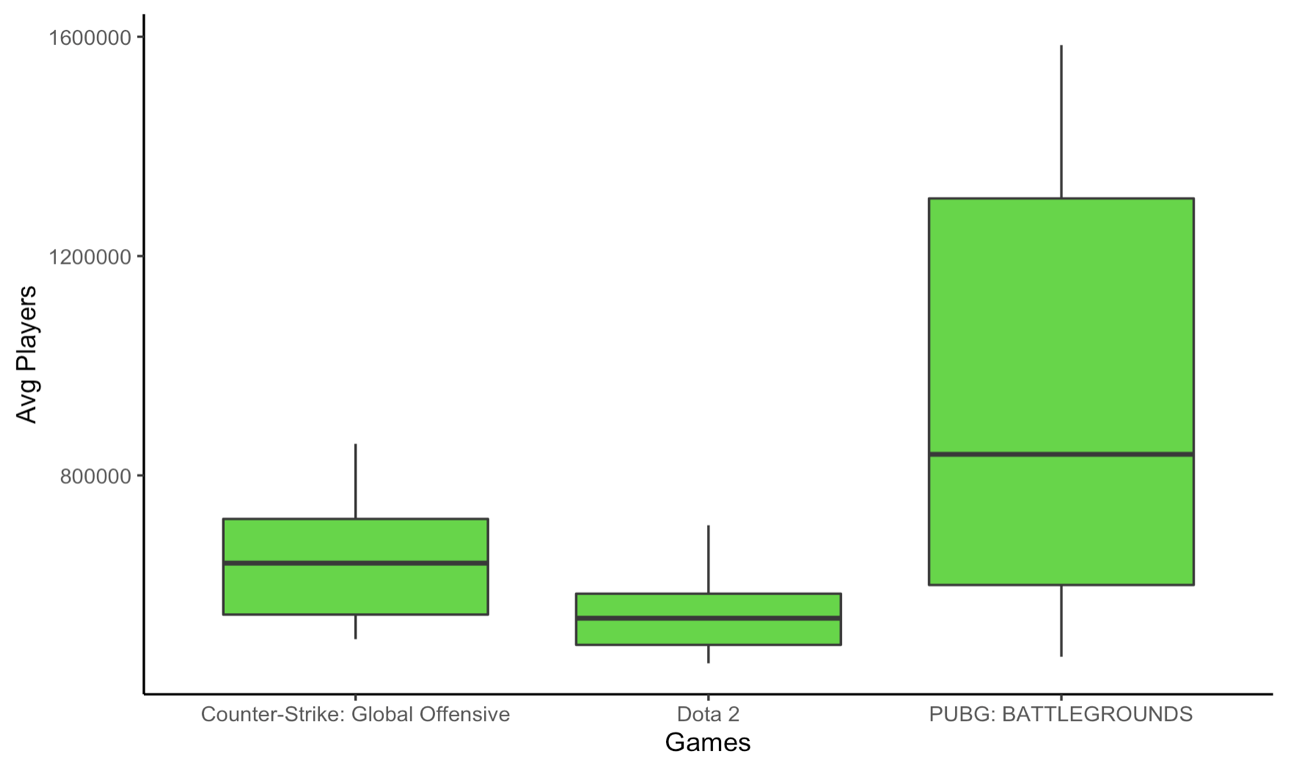
Dota 2    87.9 million

Counter-Strike: Global Offensive  65.3 million

PUBG: Battlegrounds  54.7 million

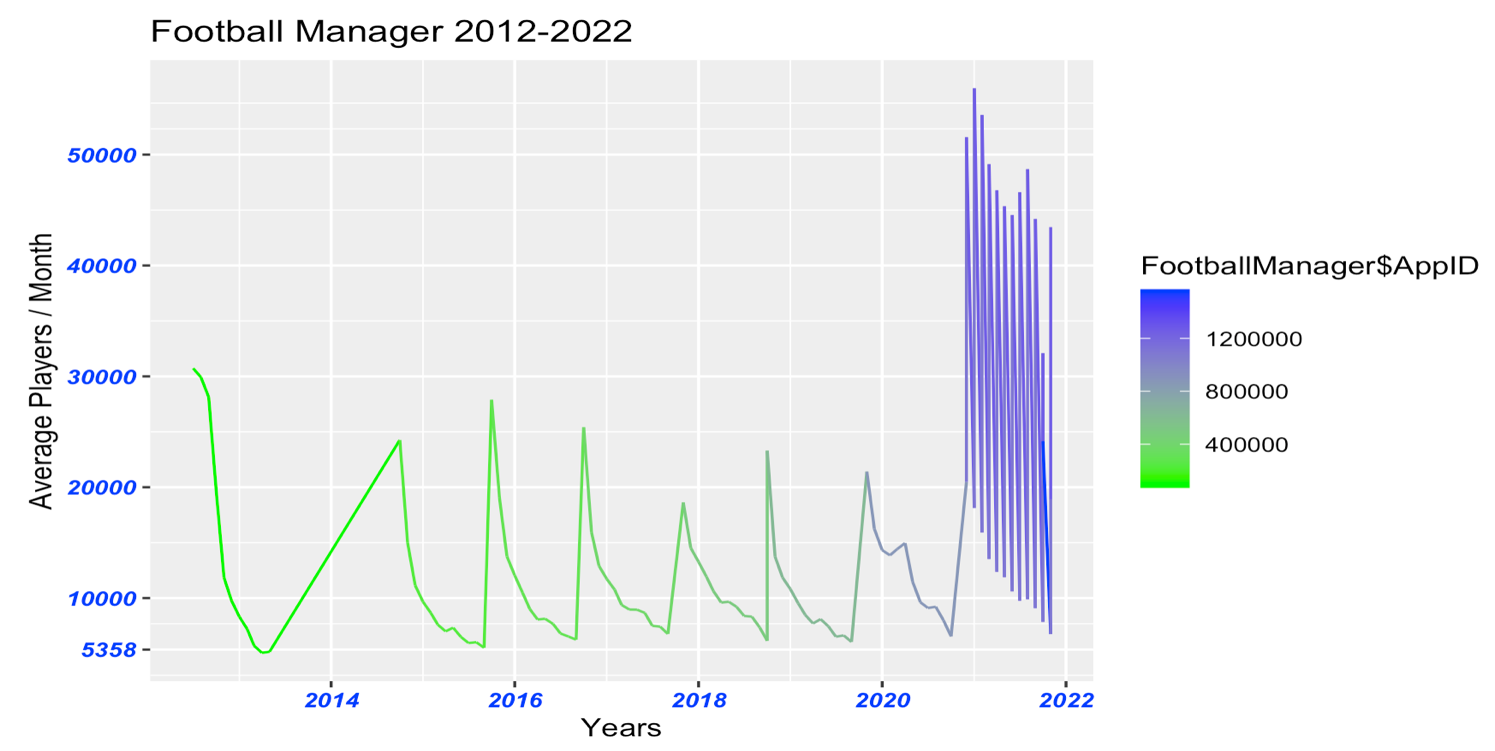
The next closest and still very successful game had 44 million less peak players,

Grand Theft Auto V 10.4 million



Another game that we should look at that has had consistently above average performance is in the sports genre. Football Manager is a very successful game, and the developers were not satisfied with just being above average. They looked for ways to innovate. And their work has paid off. For eight years the Average Players per month during the release of the new games would jump to 20,000 – 30,000 but the last 2 releases are above 50,000. We can see a few reasons for this. For one they made the game available on Xbox for the first time since 2008. So now it available on 7 gaming platforms. The second significant reason is that Football Manager mobile is getting more popular and has expanded to 3 more countries that all have football in their top sports. (Argentina, Canada, and Mexico). The developers have also made several improvements which reviewers are raving about.

What we can take away from this analysis is that we should continue to innovate. We should look for ways to improve or and expand our market. If we don’t do this, we could easily lose market share to one of the thousands of other games out there.

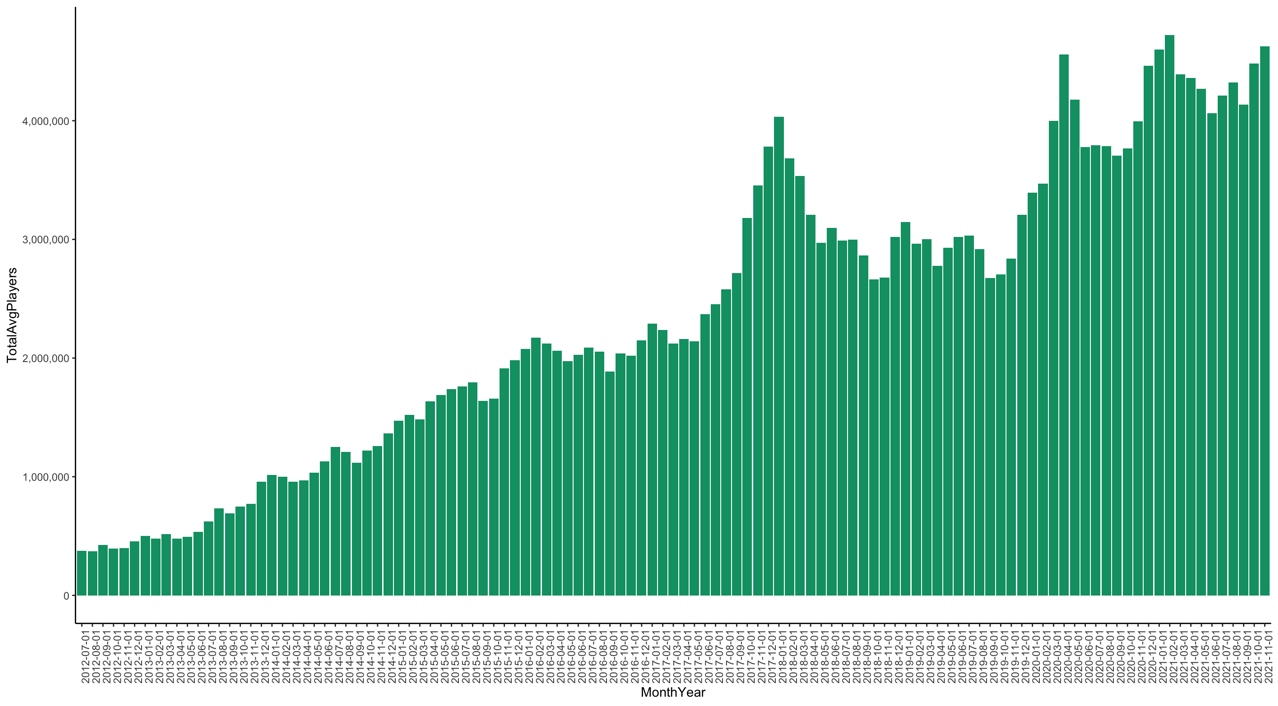


**Is there a specific seasons or month where the number of players increase the most on average?**

There has been a general increase in the number of players year after year. November and December each year have the most total players. Most games are released in the winter, especially sequel editions. For example, NBA 2k is released the winter before the following year.

As depicted in the figure below, it is prevalent that most people were playing steam games during the years 2018 – 2021. This may be linked to the stay-at-home mandate during the first outbreak of COVID-19. People around the world had nothing but time to get to know these games and have fun playing while staying safe at home.

As further investigation, it may be worth looking at the distribution of average players over the years 2018 – 2021 to create a predictive model. It would be interesting to see what predictions could be made about when people are going to play the most in 2022. However, this graph does tell that fall and winter are the times an analyst may expect a player surge.

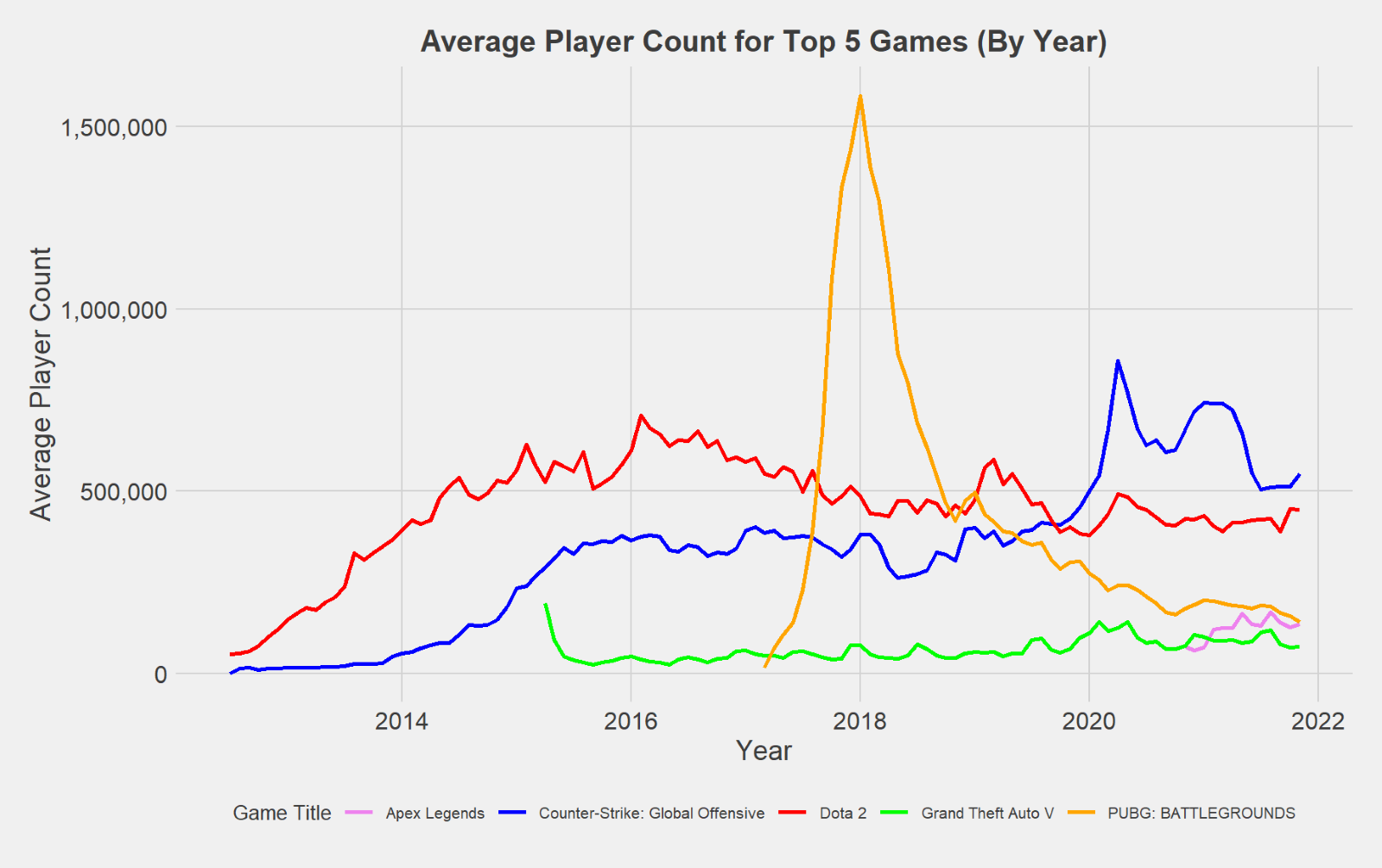


**What was the average number of players from the last 3 months?**

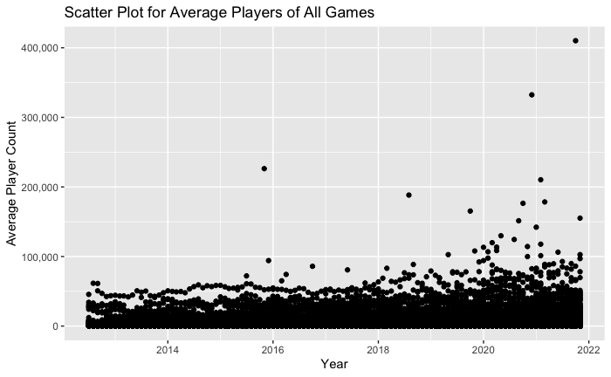
There were 4,421,500 average number of players each month in the last 3 months, with 1,320,785 average number of players for the top 5 games each month in the last 3. That means the top 5 games make up almost 30% of the industry, while the other 968 games make up the other 70%. This means that unless a game can break through the competition, it is most likely the game will get drowned in the sea of competition. We will also see this in a graph created to visualize the average number of players for each game over time.

We created a graph (1) with several lines representing the top 5 games along with points (2) to represent the other 968 games. We saw the top 5 games dominating the graph with a huge difference between those games and the cluster of games at the bottom. This could be due to the fact that there is such a high concentration of competitors in the gaming industry, so many games end up disappearing or getting lost in the sea of competitors with new games replacing those that left the industry.

We believe this could help our stakeholders at least understand how difficult and competitive the gaming industry is. However, we believe it would be more beneficial to also have data on user rating of the game, how many different operating systems the game supports, cost, and more to create different combinations and visualize it with a linear model showing the relationship between these different combinations and the average number of players.



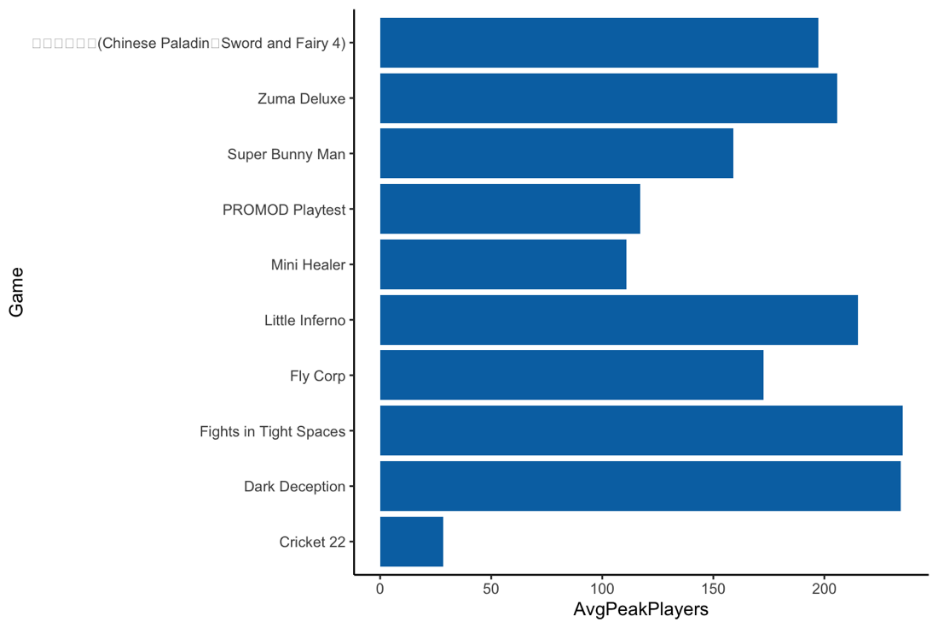
(1)

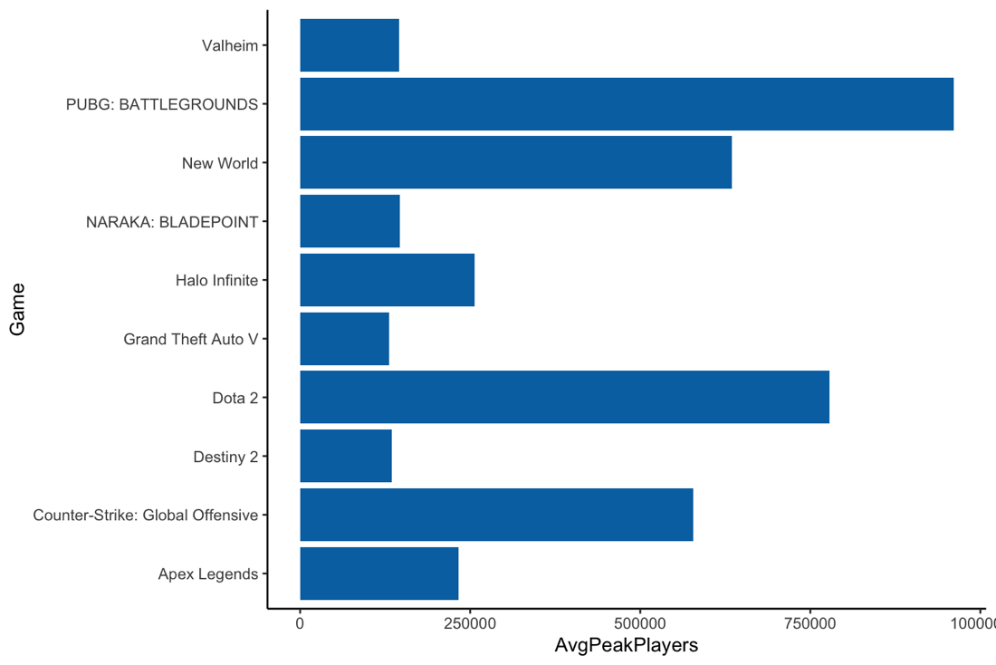
(2)

**What are the average peak players per game?**

As aforementioned, the data set used contained about 1,000 games over more than 50,000 rows. Thus, creating any bar plot of the full set would be difficult to interpret. To answer this question, a subset was created that only provided the average peak players per game. The graphs below represent the mean peak players for the bottom 10 and top 10 games.

The information that jumps out from the figures below are the games, Cricket22 and PubG: Battlegrounds. After doing further research, Cricket22 is an arcade style game on several gaming platforms. Cricket is a bat and ball game similar to baseball and was originated in Australia. Diving more deeply into the reviews, it is likely that Cricket22 is the best online Cricket game in 2021 but is performing against other games in its genre. This may explain why Cricket22 has such a low average of peak players from 2012-2021. On the other hand, PubG is the highest performing game in the data set. Its average peak players is just short of 1,000,000 players playing simultaneously. The aesthetics of this game are close to Call of Duty where players can play in squads against other teams. Being able to create an avatar and customize weapons may also play a part in why this game is so popular.





**What words are most prominent amongst the top performing and bottom performing games?**

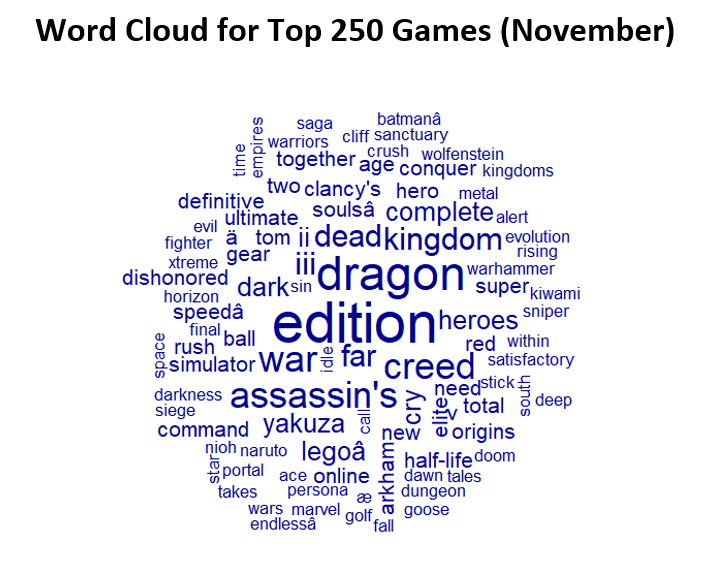
After reviewing the data available to us, the team decided to construct a visual display of text data. The logic was that the Game column contained only character data in the form of game titles, and that this data could be harnessed to construct meaningful visualizations(s) for our stakeholders.

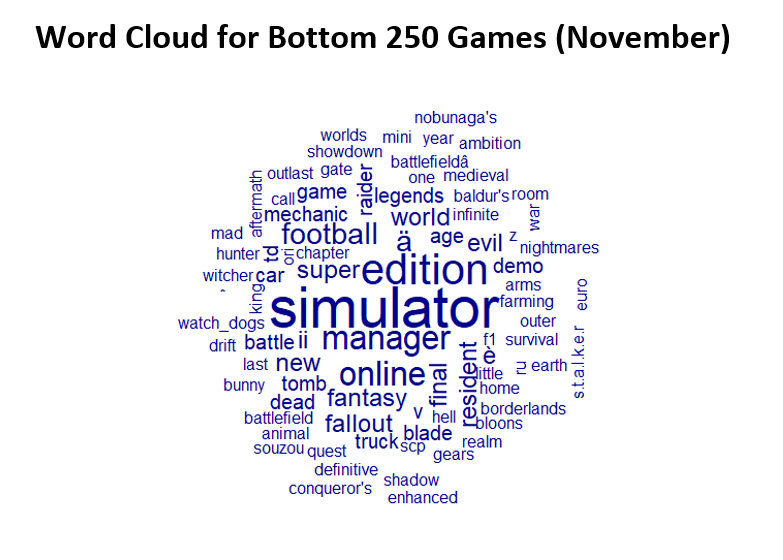
The aim was to construct a word cloud visual using two custom subsets, one to represent the top 250 games and the other to represent the bottom 250 games (both based on average player counts). After segregating the data into two subsets and constructing each one’s respective corpus, tokens, and document-feature matrix, the team moved forward with the creation of two distinct word clouds.

The first word cloud provided us with a high-level overview of the most prominent words present within the titles of the top 250 games. Some of the most prominent words in this visualization were words like “Edition,” “Dragon,” and “War.”

Conversely, the second word cloud gave us a focused look into the most prominent words present within the titles of the 250 lowest performing games. Prominent words in this low-performer visualization included “Edition,” “Simulator,” and “Manager.”

The insight derived from the word clouds could prove to be invaluable to an independent game studio. Developers could potentially detect trends in player preferences with regards to game naming conventions. Note that the word “Edition” appears in both word clouds. This indicates that the word “Edition” may not have any bearing on the success of a game, but rather appears commonly amongst all game titles.





**Data Acquisition**

The data set Steam Charts was acquired and downloaded from the data set repository Kaggle. The data set itself was in csv file format and was imported into RStudio using the Import Dataset feature.

**Data Cleaning**

The most challenging part in cleaning this dataset was converting the Month attribute into something useful in R.  We started by removing the rows with ‘Last 30 Days’ variable in the Month attribute. Then converted the Month from ‘chr’ to ‘Date’.

There was still one more column that needed to be converted to a number before we could start doing some analysis on the dataset. We created a function to convert the percentage data from ‘chr’ to ‘num’.

**Structure**

'data.frame': 50066 obs. of  7 variables:

 $ Month      : Date, format: "2021-11-01" "2021-10-01" "2021-09-01" ...

 $ AvgPlayers : num  548162 512436 512351 512082 506067 ...

 $ Gain       : num  35725.8 84.9 269 6014.6 -43279.7 ...

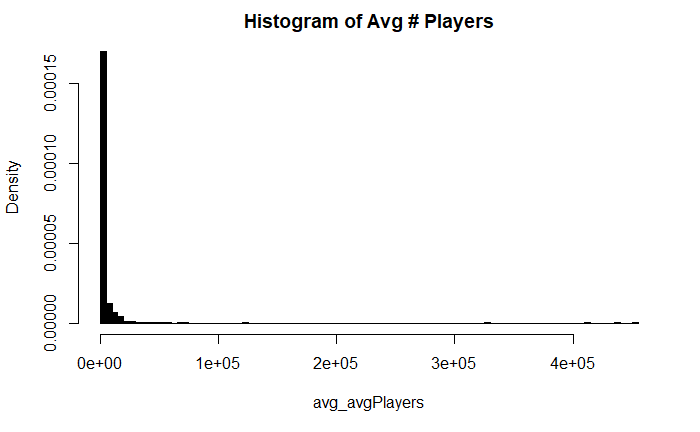
 $ PercentGain: chr  "+6.97%" "+0.02%" "+0.05%" "+1.19%" ...

 $ PeakPlayers: int  935593 864966 942519 802544 763523 929940 1087197 1148077 1198581 1123485 ...

 $ AppID      : int  730 730 730 730 730 730 730 730 730 730 ...

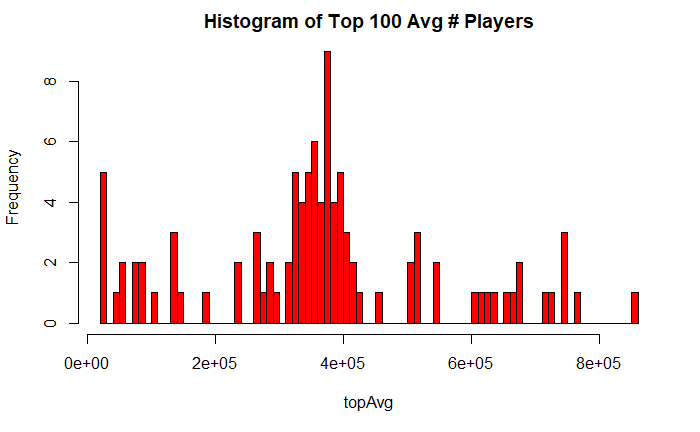
 $ Game       : chr  "Counter-Strike: Global Offensive" "Counter-Strike: Global Offensive" "Counter-Strike: Global Offensive" "Counter-Strike: Global Offensive" ...

**Descriptive Statistics**

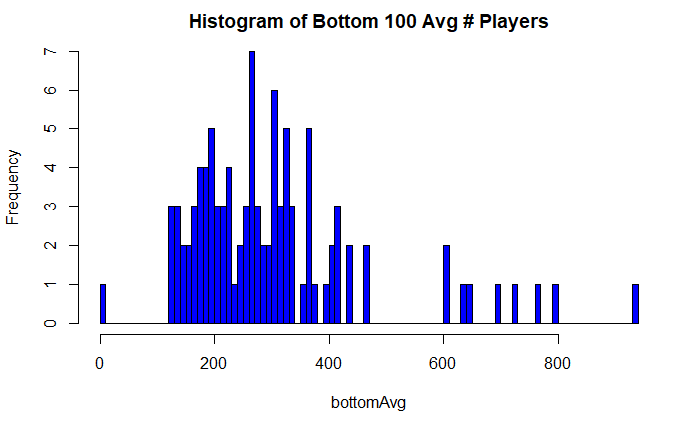


We initially wanted to create histogram for the total average number of players.  We found that the total average number of players has a mean of 5,339.163 players and a median of 716.33 players.  As you can see in our first histogram, it is extremely skewed to the right.  With the low mean and median player numbers of the total average number of players, one can conclude that most games in the dataset are relatively unpopular.

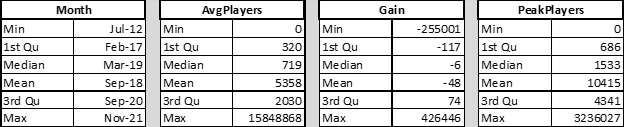
Due to the extreme degree of the positive skewedness in the original graph, we then decided it would be a promising idea to break up the average number of players and went with the top 100 and bottom 100.  We then recreated histograms for these which gave us a more readable graphs to analyze.



The histogram for the top 100 average player results appears to be normally distributed.  The top 100 have a mean of 366,418.1 players and a median of 361,846 players.  The mean and median for the top 100 average number of players are fairly close together given the large numbers given.



The histogram for the bottom 100 average number of players is slightly skewed to right, or positively skewed.  This makes sense given the bottom 100 would have the least number of average players per month in the dataset.  The bottom 100 had a mean of 304.0372 players and a median of 270.035 players in a month.  All of these factors conclude that the bottom 100 were reasonably unpopular games.



**Optimal Data**

We concluded that the data we had could have given us more information on the games.  While we understand that no dataset is perfect, we have a few suggestions that we felt could have been beneficial to work with.

One very important piece of information on the individual games would be the genre.  This can be one of the single biggest factors on determining what types of games people enjoy playing.  We know from experience that most players enjoy war related games, aka shooter games.  This data is lacking from the dataset we used.

Another area we felt they could’ve added was the sales information for each game.  This could help backup the data for game popularity and help someone model new games after the more popular and bestselling titles.

Geographic information about where players are both buying and playing from could be immensely helpful as well.  We believe this could help in the marketing campaigns for games, allowing companies to pinpoint their advertising and marketing strategies to allow for better use of money.  This info can also be helpful in determining the kind of games released.

User reviews and game ratings from publications can be helpful.  This can show and possibly even predict how a game will do when it is first launched.  By doing text mining one could probably determine with relative accuracy how well a game will do after being released.

Lastly, the dataset we used contained only games that are available on PC.  While some of the games are available on PC, mobile, and consoles, the information on what device is used is very important.  This can give companies ideas on what platforms to concentrate their efforts on to come up with the next popular game.  This allows them to see if only using maybe one or two platforms to launch on, or all platforms is cost effective and worthwhile to pursue.

**References**

John, J. (2020, January 19). *How to remove outliers in R: R-bloggers*. R. Retrieved December 17, 2021, from <https://www.r-bloggers.com/2020/01/how-to-remove-outliers-in-r/>

Mohr, Joseph. “Player Counts on Steam.” *Kaggle*, 4 Dec. 2021, https://www.kaggle.com/josephvm/player-counts-on-steam/metadata.

Prabhakaran, Selva. “If(Typeof \_\_ez\_fad\_position != 'Undefined'){\_\_ez\_fad\_position('Div-GPT-Ad-r\_statistics\_co-Box-3-0')};Outlier Treatment.” *Outlier Treatment With R | Multivariate Outliers*, <http://r-statistics.co/Outlier-Treatment-With-R.html>.

**Appendix**

library(tidyverse)

library(jmuOutlier)

library(modeest)

library(dbplyr)

library(ggplot2)

library(ggthemes)

library(scales)

library(zoo)

library(stringi)

library(lubridate)

library(purrr)

library(openair)

library(quanteda)

library(quanteda.textplots)

library(data.table)

# read in the csv file and declare the main dataframe

game <- read.csv("steam\_charts.csv")

cnames <- c("Month","AvgPlayers","Gain","PercentGain","PeakPlayers","AppID","Game")

colnames(game) <- cnames

#creating subset by removing Last 30 Days rows

game <- subset(game, Month!="Last 30 Days")

game1 <- sub(" ", " 01 ", game$Month)

game$Month <- game1

#converting Month column to date type

game$Month <- as.Date(game$Month, "%B %d %Y"); game$Month

class(game$Month)

#creating new Month and Year attribute

game$monthYear <- game$Month

game$monthYear <- format(game$monthYear, format="%B%Y")

#average players of all games from last 3 months

gamelast3months <- subset(game, (monthYear == "November2021" | monthYear == "October2021" | monthYear == "September2021"))

gameTotal3 <- sum(gamelast3months$AvgPlayers)

gameAvg3 <- gameTotal3 / 3

gameAvg3

#average players of top 5 games of last 3 months

game3monthsTop5 <- subset(gamelast3months, (Game == "Counter-Strike: Global Offensive" | Game == "Dota 2" | Game == "PUBG: BATTLEGROUNDS" | Game == "Apex Legends" | Game == "Grand Theft Auto V"))

gameTotal3Top5 <- sum(game3monthsTop5$AvgPlayers)

gameAvg3Top5 <- gameTotal3Top5 / 3

gameAvg3Top5

#percentage of industry for top 5 games in the last 3 months

gameAvg3Top5 / gameAvg3

#number of games in industry that are not the top 5

length(unique(game$Game))-5

#Linear model for the bottom games

gameBottom <- subset(game, Game!="Counter-Strike: Global Offensive" & Game!= "Dota 2" & Game != "PUBG: BATTLEGROUNDS" & Game != "Apex Legends" & Game != "Grand Theft Auto V")

ggplot(gameBottom, aes(x = Month, y = AvgPlayers)) + geom\_point() + scale\_y\_continuous(labels = scales::comma) + ggtitle("Scatter Plot for Average Players of All Games") + xlab("Year") + ylab("Average Player Count")

# Created a linear model to see if Months were responsible for AvgPlayers

bestMonth <- lm(formula = AvgPlayers ~ Month, data = game)

# Used summary to show residuals, coefficients, p-values, and Adjusted R2

summary(bestMonth)

# grouped Game by AvgPlayers and descending month. Took the first 12 lines of each

data\_new2 <- topGames %>%

  arrange (desc(Month, AvgPlayers)) %>%

  group\_by(Game) %>%

  slice(1:12)

# Total number of PeakPlayers

peak <- tapply(game$PeakPlayers, game$Game, sum)

gameName <- rownames(peak)

totalPeakPlayers <- data.frame(gameName, peak)

summarise(totalPeakPlayers)

# or use group\_by

totalPeakPlayersGroupBy =

  game %>% group\_by(Game) %>%

  summarise(totalPeakPlayers =sum(peak),)

# Dota 2 has the most players of all with 87,132,203

totalPeakPlayers[which.max(totalPeakPlayers$peak),]

# ALL Football Manager games

FootballManager <- data\_new2[903:1011,]

FootballManager

#  Line cart of all Football Manager releases

ggplot(FootballManager) +

  aes(x=FootballManager$Month, y=FootballManager$AvgPlayers, color= FootballManager$AppID,  group = .5) +

  geom\_line() +

  theme(axis.text.x = element\_text(face= "bold.italic", color = "blue")) +

  scale\_y\_continuous(breaks = c(5358,10000,20000,30000,40000,50000,60000)) +

  labs(title = "Football Manager 2012-2022", x= "Years", y="Average Players / Month") +

  theme(axis.text.y = element\_text(face= "bold.italic", color = "blue")) +

  scale\_color\_gradient(low= "green", high= "blue")

# Box plot of top games per avg player

ggplot(gameTop10) +

  aes(x=gameTop10$Game, y=gameTop10$AvgPlayers, group= gameTop10$Game) +

  geom\_boxplot(fill= "0099f8") +

  stat\_summary\_bin(fun.data = get\_alt\_text, geom="text")+

  labs(x = "Games", y= "Avg Players") +

  theme\_classic()

# Function to convert percent data (chr) to numeric

percToNum <- function(x) {

  removePerc <- sub("%", "", x)

  toNumeric <- as.numeric(removePerc)

}

# Creates "gameLast30" data frame - Subset of all "Last 30 Days" fields

gameLast30 <- subset(steam\_charts, steam\_charts$Month == "Last 30 Days")

gameLast30 <- na.omit(gameLast30)

gameLast30[["PercentGain"]] <- percToNum(gameLast30[["PercentGain"]])

# Creates "gameLast30Top" and "gameLast30Bottom" data frames - Subset of top 10 and bottom 10 performers from "gameLast30" based on Percent Gain

gameLast30Top <- gameLast30[with(gameLast30, order(-PercentGain)),]

gameLast30Top <- gameLast30Top[1:10,]

gameLast30Bottom <- gameLast30[with(gameLast30, order(PercentGain)),]

gameLast30Bottom <- gameLast30Bottom[1:10,]

# Creates "gamePUBG" data frame - Subset of PUBG's data

gamePUBG <- subset(game, game$Game == "PUBG: BATTLEGROUNDS")

gamePUBG <- na.omit(gamePUBG)

# Text Mining and Word Cloud

# Top 250 and Bottom 250 performing games of November (based on percent gain)

gameLast30.Top250 <- gameLast30[with(gameLast30, order(-PercentGain)),]

gameLast30.Top250 <- gameLast30.Top250[1:250,]

gameLast30.Bottom250 <- gameLast30[with(gameLast30, order(PercentGain)),]

gameLast30.Bottom250 <- gameLast30.Bottom250[1:250,]

# Corpus, Tokens, DFM, and Word Cloud for Top 250 (November)

gameCorpus.Top <- corpus(gameLast30.Top250$Game)

toks.Top <- tokens(gameCorpus.Top, remove\_punct = TRUE, remove\_numbers = TRUE, remove\_symbols = TRUE)

toksNoStops.Top <- tokens\_select(toks.Top, pattern = stopwords("english"), selection = "remove")

gameDFM.Top <- dfm(toksNoStops.Top, remove = c("?", "?", "?", "?", "?", "?", "?", "?", "?", "'", "?", ">", "S", "Y", "?", "T", "?"))

gameWC.Top <- textplot\_wordcloud(gameDFM.Top, min\_count = 2)

# Corpus, Tokens, DFM, and Word Cloud for Bottom 250 (November)

gameCorpus.Bottom <- corpus(gameLast30.Bottom250$Game)

toks.Bottom <- tokens(gameCorpus.Bottom, remove\_punct = TRUE, remove\_numbers = TRUE, remove\_symbols = TRUE)

toksNoStops.Bottom <- tokens\_select(toks.Bottom, pattern = stopwords("english"), selection = "remove")

gameDFM.Bottom <- dfm(toksNoStops.Bottom, remove = c("?", "?", "?", "?", "?", "?", "?", "?", "?", "'", "?", ">", "S", "Y", "?", "T", "?"))

gameWC.Bottom <- textplot\_wordcloud(gameDFM.Bottom, min\_count = 2)

#Histogram, mean, & median of Total Avg # Players

hist(avg\_avgPlayers, main = "Histogram of Avg # Players", breaks = 99, col = "black", freq = F )

mean(game$AvgPlayers)

median(game$AvgPlayers)

 #Naming Top 100 & bottom 100 Total Avg # Players

topAvg <- head(game$AvgPlayers,100)

bottomAvg <- tail(game$AvgPlayers,100)

#Histogram, mean, & median of Top 100 Avg # Players

hist(topAvg, col = "red",  breaks = 99, main = "Histogram of Top 100 Avg # Players")

mean(topAvg)

median(topAvg)

#Histogram, mean, & median of Bottom 100 Avg # Players

hist(bottomAvg, col = "blue",  breaks = 99, main = "Histogram of Bottom 100 Avg # Players")

mean(bottomAvg)

median(bottomAvg)

# For each game average # of peak players

#group the data by APPID and then PeakPlayers

PPGroupBy =

game %>% group\_by(Game,PeakPlayers)

#summarise the data into summary

summary <- summarise(PPGroupBy)

#find the average and store in variable

average\_peak\_pergame <- with(summary, by(PeakPlayers, Game, mean))

#put back into a matrix

average\_peak\_pergame\_2 <- t(sapply(average\_peak\_pergame, I))

#create data frame for avg peak per game

new\_avg\_peak.df <- as.data.frame(average\_peak\_pergame\_2)

#stack two create two columns

df <- stack(new\_avg\_peak.df)

df <- na.omit(df)

#update the column names

colnames(df) <- c('AvgPeakPlayers','Game')

#store into new dataframe and order in ascending order

newdf <- df[order(df$AvgPeakPlayers),]

#set top 10 and lowest 10 to veriables

low <- head(newdf, n = 10)

high <- tail(newdf, n = 10)

#visuals

#create bar graph for top 10 lowest 10 avg Peak Players

p1<-ggplot(data=low, aes(x=Game, y=AvgPeakPlayers)) +

geom\_bar(stat="identity", show.legend = FALSE,  fill = "#0072B2")+

theme\_classic()

p1+coord\_flip()

p2<-ggplot(data=high, aes(x=Game, y=AvgPeakPlayers)) +

geom\_bar(stat="identity", show.legend = FALSE,  fill = "#0072B2")+

theme\_classic()

p2+coord\_flip()