Video Game Sales (CS544_Final_Project)

Tommy Lee

10/3/2021

Background: Dataset is based on Video game sales from the early 1980s to Dec of 2016 from Kaggle. Almost all of my analysis required removing NA values for each markdown chunk depending on the dataframe column I am analyzing. I broke down my analysis into smaller questions and a summary/ conclusion at the end.

Big Question: How have video game sales been doing for the past decades?

Data

```
library(sampling)

data <- read.csv("C:/Users/Tommy Lee/Desktop/CS 544/CS544_Final_Project_Lee/Video_Games_Sales_as_at_22_

# Data consist of 2017 & 2020, but this dataset should only be up to Dec 22,2016
data <- subset(data,!(Year_of_Release %in% c('2017','2020')))

# Change year from string to numeric for analysis
data['Year_of_Release'] <- as.numeric(data[['Year_of_Release']])
```

```
## Warning: NAs introduced by coercion
```

I had to clean the data and remove any anticipated games that were going to be released further than 2016 such as 2017 or 2020. Also, I changed the "Year of Release" column from character to numeric type for my analysis.

Check data type for each data frame column

```
sapply(data,class)
```

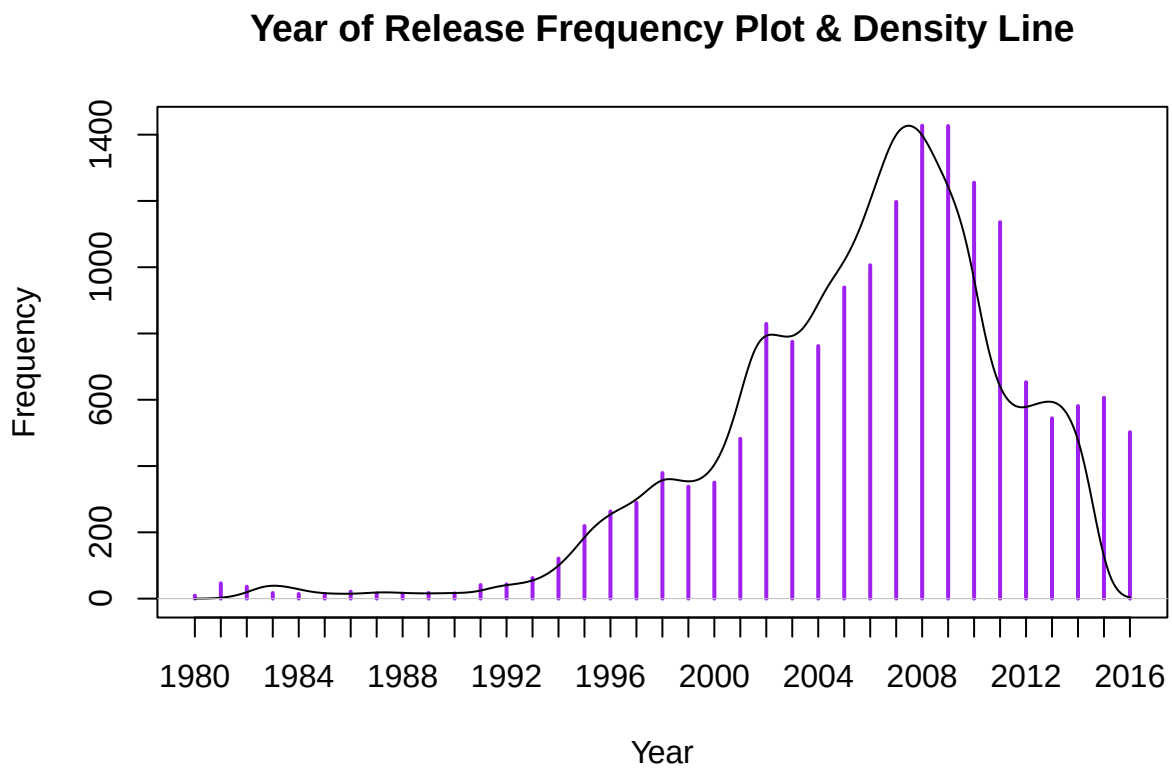
```
##           Name       Platform Year_of_Release          Genre      Publisher
##    "character"    "character"      "numeric"    "character"    "character"
##       NA_Sales       EU_Sales       JP_Sales    Other_Sales   Global_Sales
##      "numeric"      "numeric"      "numeric"      "numeric"      "numeric"
```

```
##    Critic_Score    Critic_Count      User_Score      User_Count       Developer
##       "integer"       "integer"       "numeric"       "integer"     "character"
##          Rating
##     "character"
```

Q: How is the distribution of video games over the years?

Analysis based on year (distribution)

```
# Remove N/A values for year
year_data <- subset(data,Year_of_Release != "N/A")


# Left skewed normal distribution
plot(table(year_data$Year_of_Release),xlab = 'Year', ylab = 'Frequency',
     main = 'Year of Release Frequency Plot & Density Line',col = 'purple')

# Show density plot on same plot for a better idea of distribution
par(new=TRUE)
plot(density(year_data$Year_of_Release),axes = FALSE,xlab = "", ylab = "", main = "")
```



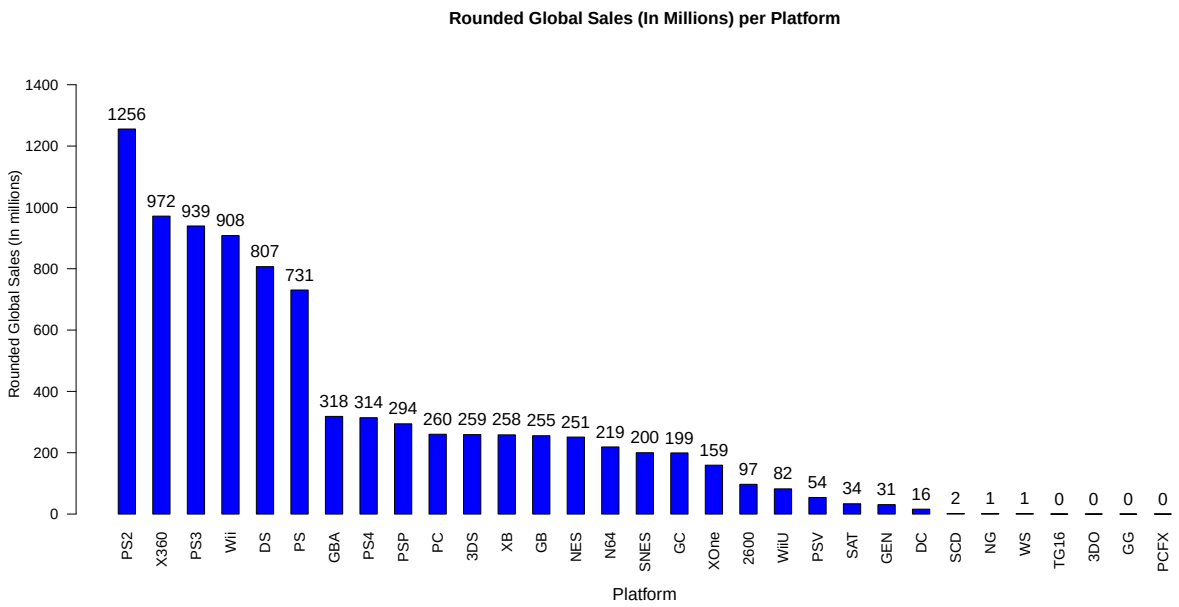
The distribution for year of release variable is a left-skewed normal distribution. This shows that in the 1980s and 1990s, video games were scarce and the 2000s is when video games start to stockpile. A simple explanation is that technology (specifically the technology to create 3D gaming) to create better games

was introduced and also improving in the 2000s. The year of release frequency starts going down in the mid-2000s probably because there was not a strong breakthrough in the advancement of technology for gaming compared to the creation of 3D gaming.

Q: Which gaming console had the most sales globally? Which gaming console had the least sales globally and why?

Analysis based on platform (categorical variable)

```
platform_data <- subset(data,Platform != "N/A")
agg_platform_data<- aggregate(platform_data[,c("NA_Sales","EU_Sales","JP_Sales",
                                               "Other_Sales","Global_Sales")],
                              by=  list(platform_data$Platform),FUN = sum)



platform_mx <- t(as.matrix(sort(agg_platform_data[,c("Global_Sales")],decreasing = TRUE)))
colnames(platform_mx)<- agg_platform_data[order(agg_platform_data$Global_Sales
                                                ,decreasing = TRUE),c('Group.1')]



x <- barplot(platform_mx  ,beside = TRUE,ylim = c(0,1500),
        ylab = "Rounded Global Sales (In millions)",col = "blue",
        main = "Rounded Global Sales (In Millions) per Platform",las = 2)
text(x,y = sort(agg_platform_data$Global_Sales,decreasing = TRUE) + 50 ,
     labels = round(sort(agg_platform_data$Global_Sales,decreasing = TRUE)), cex = 1.2)
title(xlab="Platform", line=4, cex.lab=1.2)
```



```r
# A few PS2 games
head(platform_data[(platform_data$Platform == 'PS2'),c('Name','Global_Sales')],10)
```

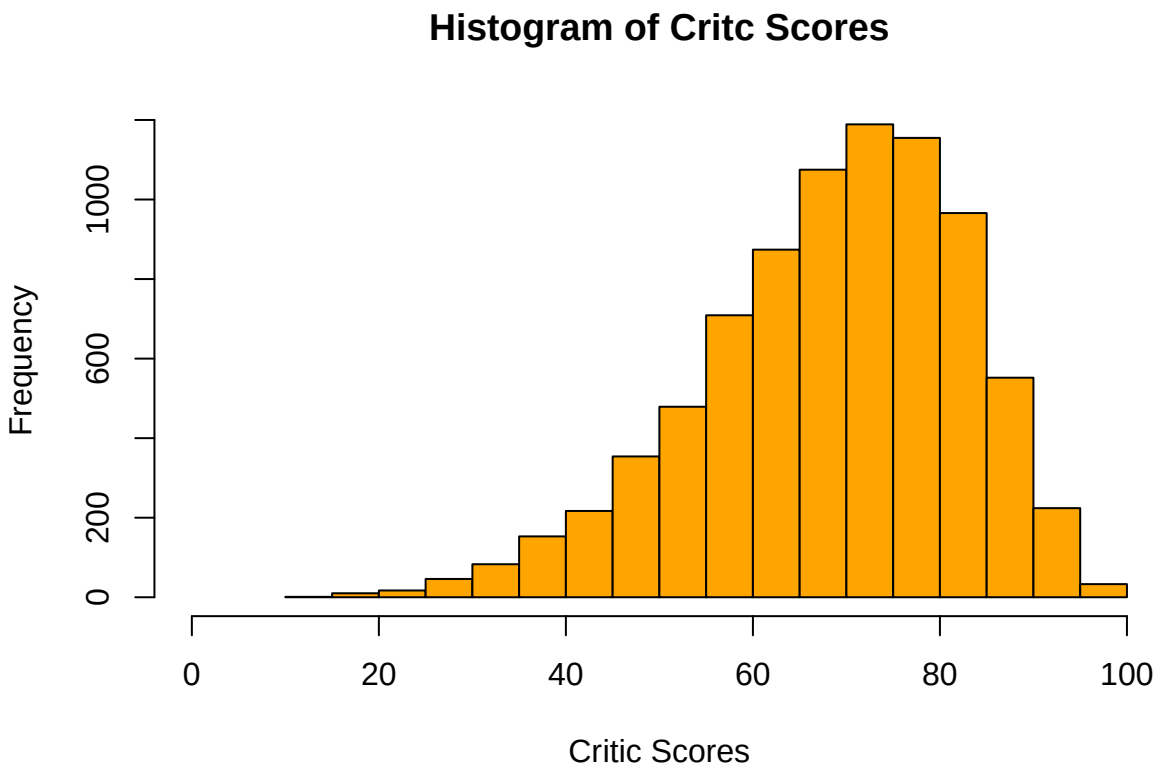
```
##                              Name Global_Sales
## 18  Grand Theft Auto: San Andreas        20.81
## 25    Grand Theft Auto: Vice City        16.15
## 29         Gran Turismo 3: A-Spec        14.98
## 39           Grand Theft Auto III        13.10
## 49                Gran Turismo 4        11.66
## 85               Final Fantasy X         8.05
## 103    Need for Speed Underground         7.20
## 114  Need for Speed Underground 2         6.90
## 115     Medal of Honor: Frontline         6.83
## 133               Kingdom Hearts         6.40
```

In the barplot above, we can see that the PS2 had the highest global sales for a total of about 1256 million sales. This makes sense because PS2 had the largest gaming library out of all the gaming consoles and had a lot of the most classic games we know today such as Final Fantasy, Kingdom Hearts and Grand Theft Auto. Lesser-known gaming consoles such as PCFX, GG and 3DO had the lowest global sales and that is most likely because these consoles were released before the 2000s when video game production was low.

Q: Are these games classified as highly-rated games based on critics?

Analysis based on Critc Score (numerical variable)

```r
critic_data <- subset(data,Critic_Score!= "N/A")


hist(critic_data$Critic_Score,
        col = "orange",breaks= 20,
        xlim = c(0,100), xlab = "Critic Scores", ylab= 'Frequency',
        main = 'Histogram of Critc Scores')
```

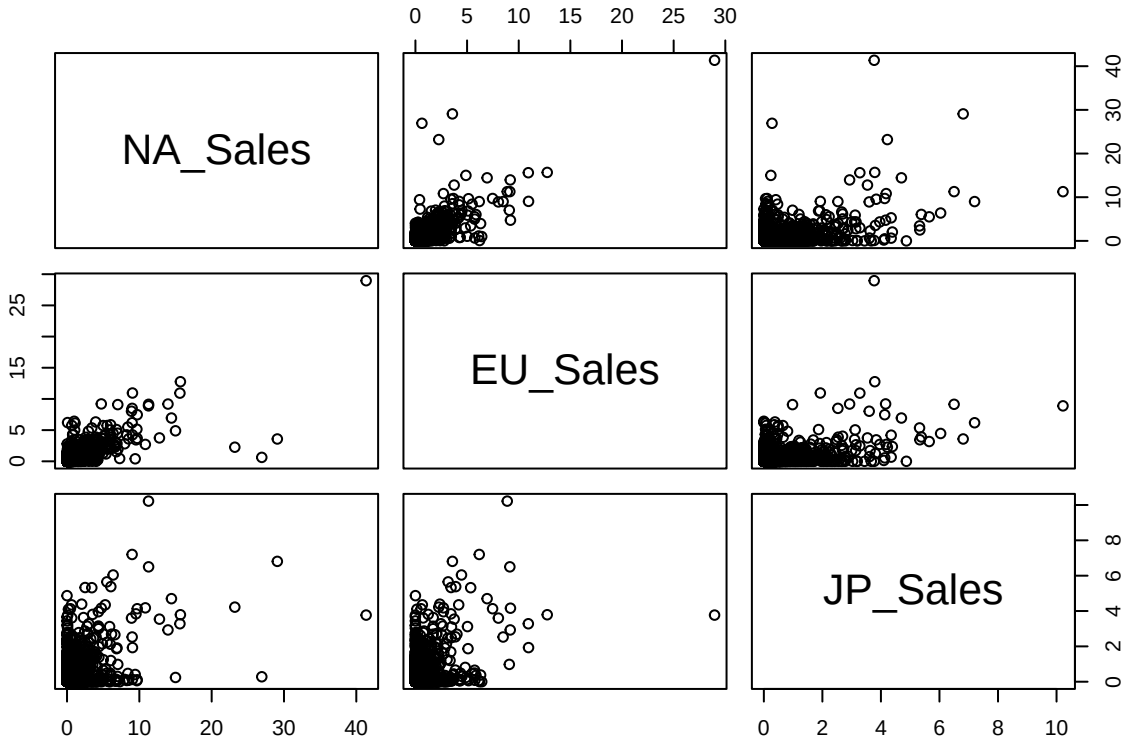


Based on the histogram above, the scores are very close to being normally distributed and most critics give a score of 75 to 85 meaning these video games are fairly rated.

Q: Is there any correlation between the different kind of sales? If so, what can we infer from the correlation?

Possible correlation between Sales

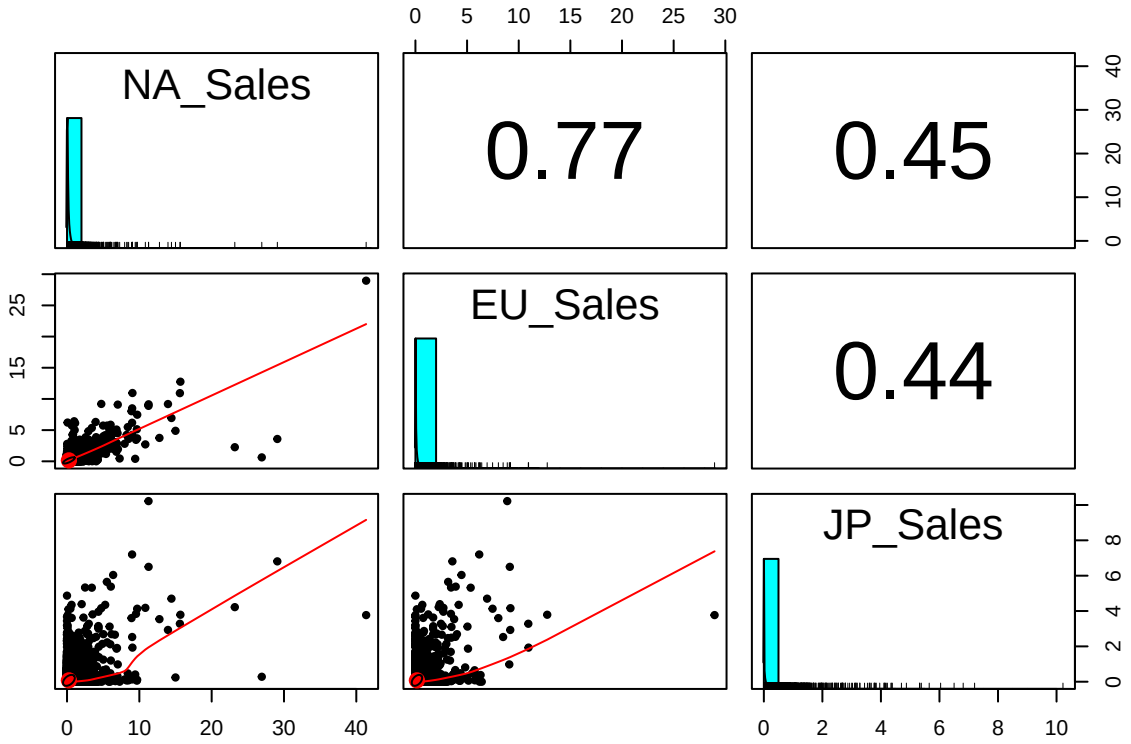
```
library(psych)
# Sales data only (excluding Global Sales)
sales_only <-data[,c("NA_Sales","EU_Sales","JP_Sales")]


# Pairwise plot
pairs(sales_only)
```




```
# Correlation Matrix
cor(sales_only)
```

```
##             NA_Sales  EU_Sales  JP_Sales
## NA_Sales  1.0000000 0.7653347 0.4495940
## EU_Sales  0.7653347 1.0000000 0.4350608
## JP_Sales  0.4495940 0.4350608 1.0000000
```

```
# Combo plot of pairwise plot and correlation matrix
#For fun, show correlation matrix and pairwise plots w/ linear regression line in one plot
pairs.panels(sales_only)
```



Using the pairs.panel function, I created a combo plot of the pairwise plot and correlation matrix, we can see that there is a moderately strong correlation between NA and EU Sales (0.77) while NA and EU Sales compared to JP Sales have relatively low correlation (less than 0.5). My inference on why there is such low correlation with JP sales is due to regional preference. To elaborate, as an example, US and JP culture are significantly different and that means what US considers entertainment can be very different in JP. A popular video game in Japan can be not so popular in the US due to audience preference.

Q: What is the average user score using the Central Limit Theorem?

Apply Central Limit Theorem to User Score

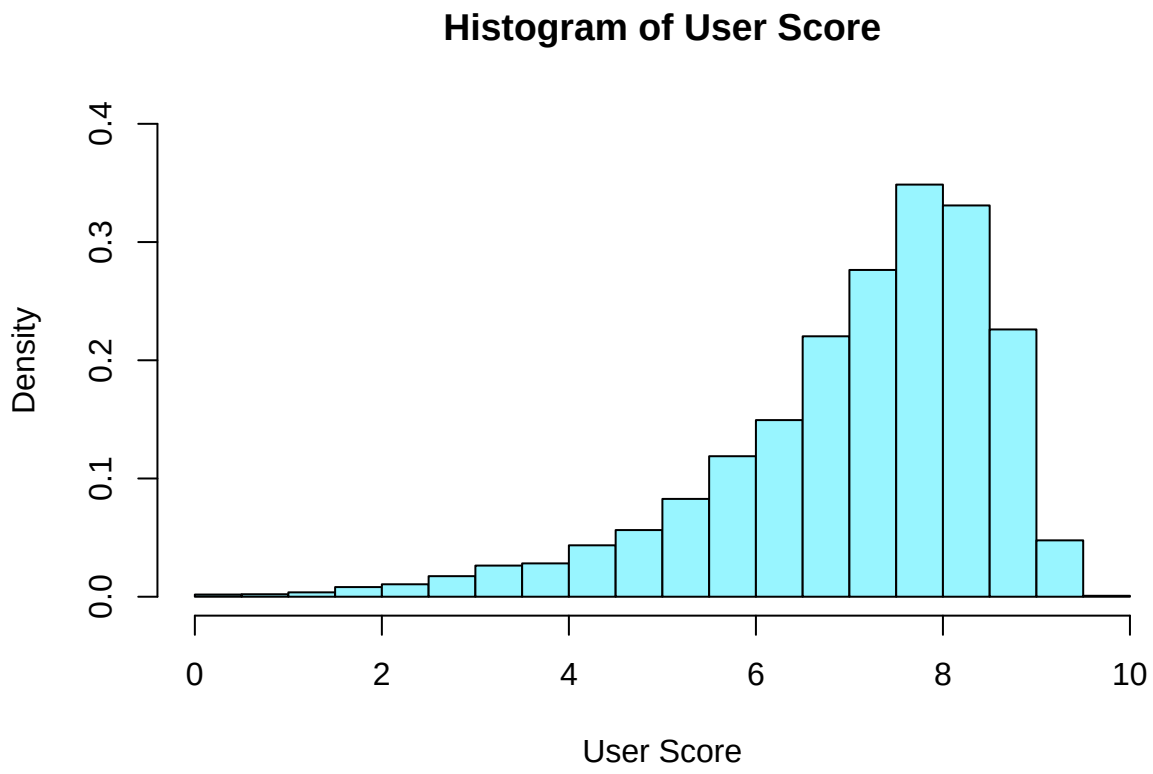
```
set.seed(7046)

# Remove all NA values in User Score
user_not_na <- data[!(is.na(data$User_Score)),]

# Check User Score Range
cat("The range of user scores is",range(user_not_na$User_Score)[1],"to",
    range(user_not_na$User_Score)[2])
```

```
## The range of user scores is 0 to 9.7
```

```r
# Original dataset stats and plot
cat("User Score Mean = ", mean(user_not_na$User_Score),
      "and SD = ", sd(user_not_na$User_Score), "\n")
```

```
## User Score Mean =  7.125046 and SD =  1.500006
```

```r
hist(user_not_na$User_Score, main = "Histogram of User Score"
     , xlab = "User Score", col = "cadetblue1",prob = TRUE,
     ylim = c(0,0.4))
```



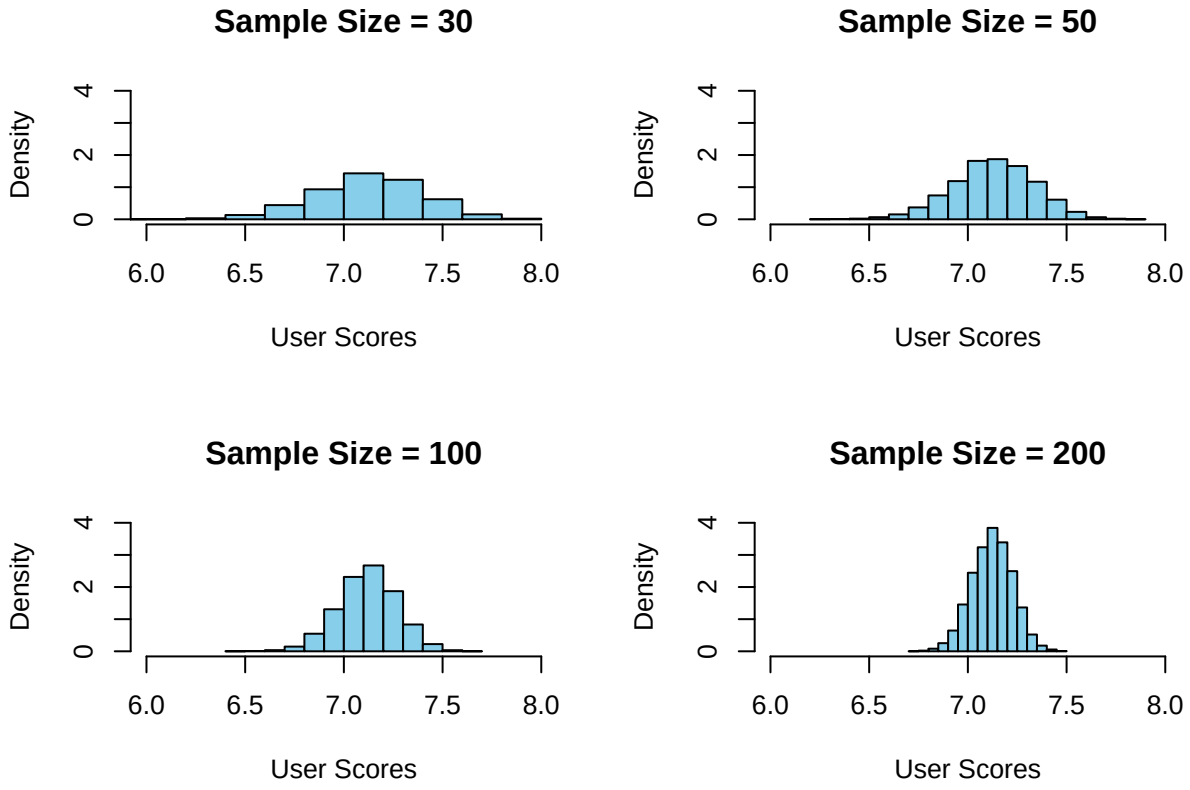

```r
# Create empty vector to store sample means
xbar <- numeric(nrow(user_not_na))


# Simple random sampling w/o replacement to prove CLT
par(mfrow = c(2,2))

for (size in c(30, 50, 100, 200)) {
  for (i in 1:nrow(user_not_na)) {
    s <- srswor(size,nrow(user_not_na))
    xbar[i] <- mean(user_not_na[s != 0,"User_Score"])
  }
  hist(xbar,xlim = c(6,8), xlab = "User Scores",
   main = paste("Sample Size =", size), col = "sky blue", prob = TRUE,
```

```
   ylim = c(0,4))

  cat("Sample Size = ", size, " Mean = ", mean(xbar),
      " SD = ", sd(xbar), "\n")
}
```

```
## Sample Size =  30  Mean =  7.128644  SD =  0.2754325
```

```
## Sample Size =  50  Mean =  7.130078  SD =  0.2098499
```

```
## Sample Size =  100  Mean =  7.119938  SD =  0.1489328
```




```
## Sample Size =  200  Mean =  7.122324  SD =  0.10373
```

Using the "User_Score" column from the dataset that ranges from 0 to 10 (estimated from 9.7) and is also very left-skewed, we want to apply the Central Limit Theorem to compare the true mean with the sample means. Using sample sizes 30,50,100 and 200 and simple random sampling without replacement, we can see that the standard deviation of the distribution of sample means significantly reduces as the sample size increases. The sample means are about 7.12 which is almost identical to the population mean of 7.125046. This shows that users highly rated the dataset's video games.

Q: Using various sampling methods, can we find the average amount of critics counts for the top 5 genres in this dataset?

sampling methods based on Top 5 genres

```
# Top 5 genres based on number of video games in that genre (pre-filtered no NA genre games)
genre_not_na <- data[(!is.na(data$Genre))& (data$Genre !='Misc'),]
top5 <- names(head(sort(table(genre_not_na$Genre),decreasing = TRUE),5))

# Subsetting only top 5 genres
subset_top5 <- subset(genre_not_na, Genre %in% top5)

# Sample size
samp_size <- 643
```

Simple sample randoming w/ replacement

```
# Simple Random sampling w/o replacement
s <- srswr(samp_size, nrow(subset_top5))

# Rows of original dataset that are sampled
row <- rep((1:nrow(subset_top5))[s!=0], s[s!=0])

# Simple sampling subset of original dataset
sample_top5 <- subset_top5[row,]

# Frequency
table(sample_top5$Genre)
```

```
##
##       Action    Adventure Role-Playing      Shooter       Sports
##          221           77           92           84          169
```

```
# Proportions
prop.table(table(sample_top5$Genre))
```

```
##
##       Action    Adventure Role-Playing      Shooter       Sports
##    0.3437014    0.1197512    0.1430793    0.1306376    0.2628305
```

Systematic sampling w/ equal sizes

```
set.seed(7046)
```

```r
# Systematic sampling equal sizes
k <- ceiling(nrow(subset_top5)/samp_size)


# Systematic sampling parameters
r <- sample(k,1)
system_rows <- seq(r, by = k, length = samp_size)


 # Subset of original dataset using systematic sampling equal sizes
system_genre <- subset_top5[system_rows,]

#Frequency
table(system_genre$Genre)
```

```
##
##       Action    Adventure Role-Playing      Shooter       Sports
##          227           85           89           71          143
```

```r
# Proportions table
prop.table(table(system_genre$Genre))
```

```
##
##       Action    Adventure Role-Playing      Shooter       Sports
##    0.3691057    0.1382114    0.1447154    0.1154472    0.2325203
```

Systematic sampling w/ unequal probabilities based on Critic Counts

```r
set.seed(7047)
# Inclusion probabilities
non_na_sales <- subset_top5[!is.na(subset_top5$Critic_Count),]


pik <- inclusionprobabilities(non_na_sales$Critic_Count,samp_size)


# Systematic sampling w/ unequal probabilities
s <- UPsystematic(pik)

# Using s variable as index, get subset of systematic sampling
system_unequal_samp <- subset_top5[s != 0,]

# Frequency
table(system_unequal_samp$Genre)
```

```
##
##       Action    Adventure Role-Playing      Shooter       Sports
##          436          156          191          170          307
```

```r
# Proportions table
round(prop.table(table(system_unequal_samp$Genre)),2)
```

```
##
##       Action    Adventure Role-Playing      Shooter       Sports
##         0.35         0.12         0.15         0.13         0.24
```

Stratified sampling by Genre

```r
set.seed(7046)
# Order data
order_genre <- subset_top5[order(subset_top5$Genre),]

# Remove blank value genre
order_genre <- order_genre[order_genre$Genre != '',]



samp <- floor(samp_size*table(order_genre$Genre)/sum(table(order_genre$Genre)))
sum(samp)
```

```
## [1] 641
```

```r
# Need three more for 643
add.one<-sample(1:length(samp),samp_size-sum(samp))
samp[add.one]<-samp[add.one]+1
sum(samp)
```

```
## [1] 643
```

```r
# Subset of original data using systematic sample unequal
st.1 <- strata(order_genre, stratanames = c('Genre'),
       size = samp ,method = "srswor",
       description = TRUE)
```

```
## Stratum 1
##
## Population total and number of selected units: 3369 220
## Stratum 2
##
## Population total and number of selected units: 1303 86
## Stratum 3
##
## Population total and number of selected units: 1498 98
## Stratum 4
##
## Population total and number of selected units: 1323 86
## Stratum 5
```

```
##
## Population total and number of selected units: 2348 153
## Number of strata  5
## Total number of selected units 643
```

```
#head(getdata(subset_top5,st.1))

# Frequency
table(st.1$Genre)
```

```
##
##       Action    Adventure Role-Playing      Shooter       Sports
##          220           86           98           86          153
```

```
# Percentages
prop.table(table(st.1$Genre))
```

```
##
##       Action    Adventure Role-Playing      Shooter       Sports
##    0.3421462    0.1337481    0.1524106    0.1337481    0.2379471
```

```
# Original mean
mean(non_na_sales$Critic_Count)
```

```
## [1] 27.9717
```

```
# Sample means
# Simple random sampling
mean(sample_top5[!is.na(sample_top5$Critic_Count),"Critic_Count"])
```

```
## [1] 28.20231
```

```
# Systematic sampling w/ equal sizes
mean(system_genre[!is.na(system_genre$Critic_Count),"Critic_Count"])
```

```
## [1] 26.96835
```

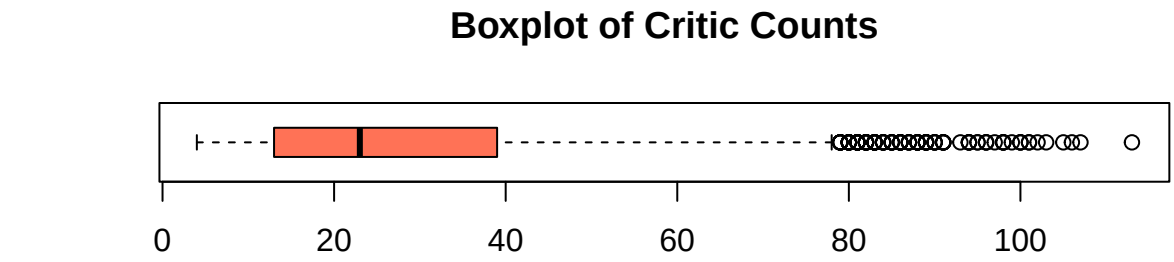
```
# Systematic sampling w/ unequal probabilities
mean(system_unequal_samp[!is.na(system_unequal_samp$Critic_Count),"Critic_Count"])
```

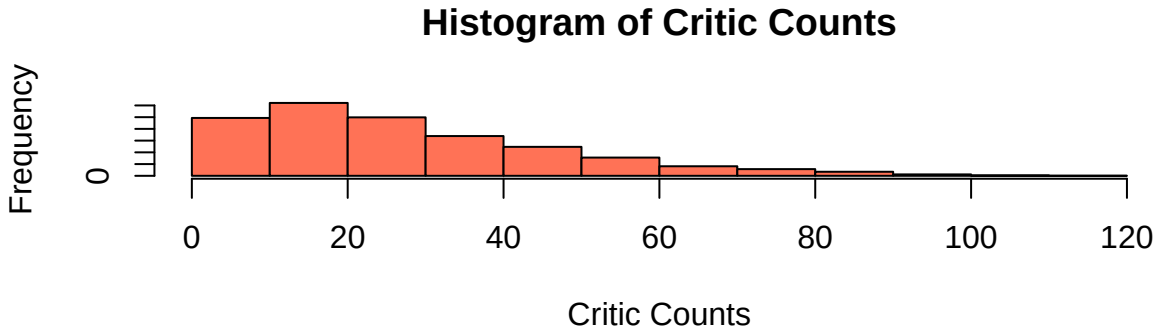
```
## [1] 29.69152
```

```
# Stratified sampling
test <- getdata(subset_top5,st.1)
mean(test[!is.na(test$Critic_Count),"Critic_Count"])
```

```
## [1] 27.18827
```

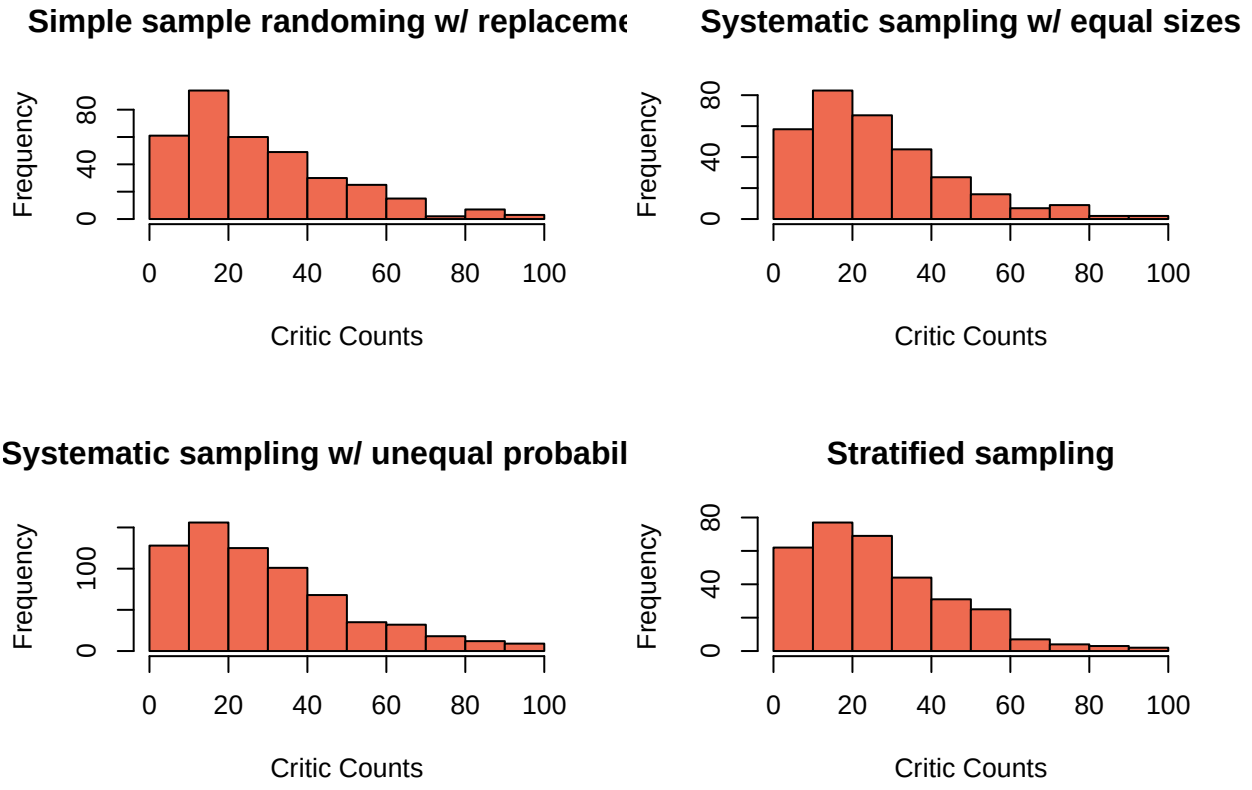
```r
par(mfrow = c(2,1))
# Boxplot of Critic Counts
boxplot(non_na_sales$Critic_Count,horizontal  = TRUE,col = c('coral1'),
        main = 'Boxplot of Critic Counts')
#Histogram of Critic Counts
hist(non_na_sales$Critic_Count,col = c('coral1'),
        main = 'Histogram of Critic Counts',xlab = 'Critic Counts')
```






```r
# How many outliers?
crit_IQR = fivenum(non_na_sales$Critic_Count)[4] - fivenum(non_na_sales$Critic_Count)[2]


crit_outlier_top = subset(non_na_sales,
      Critic_Count > fivenum(non_na_sales$Critic_Count)[4] + 1.5 *crit_IQR)

crit_outlier_bottom = subset(non_na_sales,
      Critic_Count < fivenum(non_na_sales$Critic_Count)[2] - 1.5 *crit_IQR)




par(mfrow = c(2,2))
hist(sample_top5$Critic_Count, main  = 'Simple sample randoming w/ replacement',
    xlab = 'Critic Counts', col= 'coral2')
```

```
hist(system_genre$Critic_Count, main  = 'Systematic sampling w/ equal sizes',
    xlab = 'Critic Counts', col= 'coral2')
hist(system_unequal_samp$Critic_Count,
    main  = 'Systematic sampling w/ unequal probabilities',
    xlab = 'Critic Counts', col= 'coral2')
hist(getdata(subset_top5,st.1)$Critic_Count,
     main  = 'Stratified sampling',
     xlab = 'Critic Counts', col= 'coral2')
```



We are interested in only the top 5 genres in this dataset (excluding the Misc. genre which is not a useful genre type).

As shown above in the histogram and boxplot of critic counts, the data is very right-skewed meaning we should see majority of critic counts in the range of 0 to 40. The average amount of critics in this subset is about 28. Using different sampling methods such as simple random sampling w/ replacement, systematic sampling w/ equal sizes and w/unequal probabilities and stratified sampling with a sample size of 643, we approximate about the same average as the population mean of 28-29 critics per video game. Also, using these sampling methods,the data converges closer to 28-29 in the plots better than the original histogram that looks very spread out visually. This average also shows that the scores are not biased as 28-29 critics is a fair amount of people to score per video game.

Find the most frequent words that appear in video game titles and their corresponding ratings.

```
library(stringr)
library(tm)
```

```
## Loading required package: NLP
```

```
library(sjmisc)
```

```
## Learn more about sjmisc with 'browseVignettes("sjmisc")'.
```

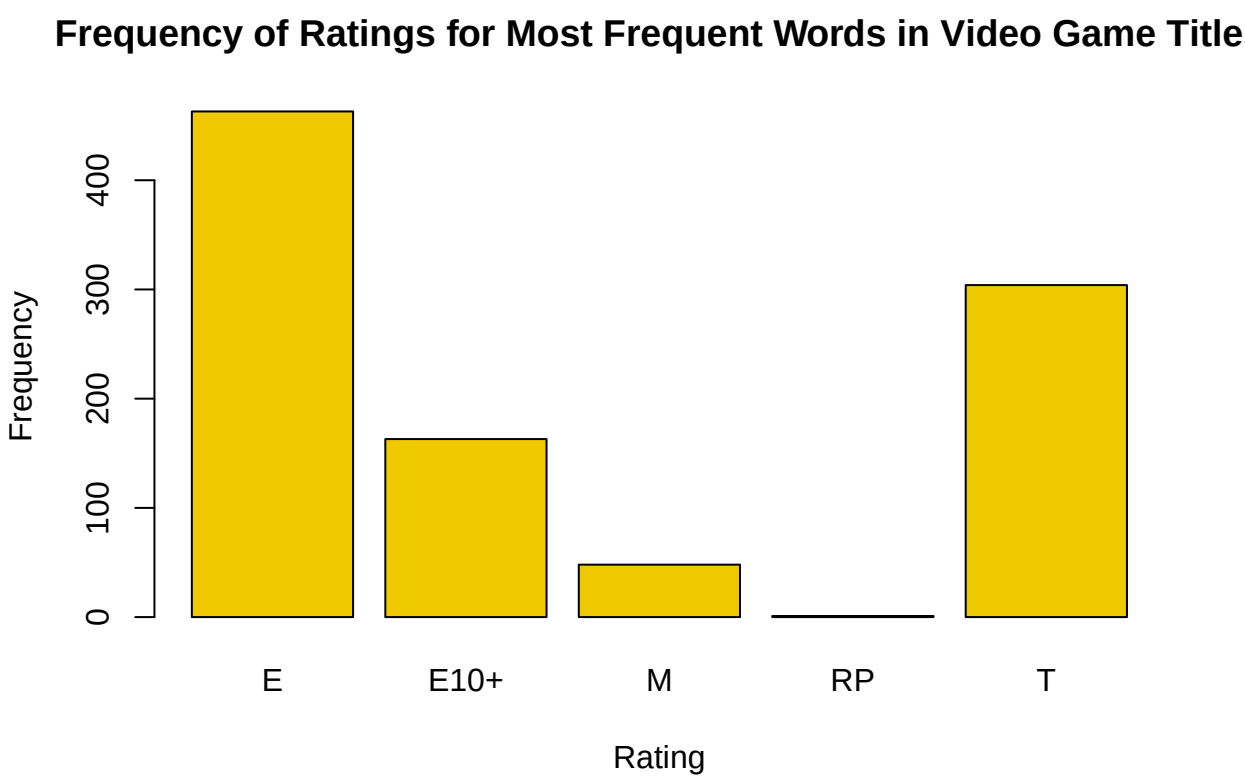
```
vid_titles <- data[data$Name != "",]

# Get all video game titles into one string separated by empty space
corps <- paste(vid_titles$Name, collapse = ' ')
# Remove punctuation
corps_fixed <- str_replace_all(corps,"[[:punct:]]"," ")
# Remove numbers
corps_fixed2 <- str_replace_all(corps_fixed,'[0-9]+',"")



# Split strings into a list of strings by empty space
spliter1<- strsplit(corps_fixed2, " ")
# Lower case all strings
unlist_vector <- tolower(unlist(spliter1))
# Remove all empty strings in vector, stopwords, and the word "game"
word_vec<- unlist_vector[unlist_vector !="" & !(unlist_vector %in% c(stopwords("en"),'game'))]


# Get Frequency of words as dataframe
word_freq_df <- as.data.frame(sort(table(word_vec),decreasing = TRUE))
# Get length of words per string
word_freq_df['word_length'] <- str_count(word_freq_df$word_vec)
# Grab words that have a frequency over 200 and length greater than 3
words_we_care <- word_freq_df[word_freq_df['word_length'] > 3 & word_freq_df['Freq'] > 200,]
# Words we care about as a vector
words_we_care_vec <- as.vector(words_we_care$word_vec)

word_freq_df[word_freq_df$word_vec %in% words_we_care_vec,c('word_vec','Freq')]
```

```
##    word_vec Freq
## 2     world  416
## 5     super  292
## 6      star  263
## 7    soccer  233
## 9    dragon  228
## 10     wars  218
```

```r
# Use for loop to grab all words that contain the words we care about
game_titles = c()
for (i in data$Name){
  if (any(str_contains(tolower(i),words_we_care_vec))){
    game_titles<- c(game_titles,i)
  }
}


# Skim the game titles/ names
head(data[data$Name %in% game_titles,'Name'],20)
```

```
##  [1] "Super Mario Bros."                   "New Super Mario Bros."
##  [3] "New Super Mario Bros. Wii"           "Super Mario World"
##  [5] "Super Mario Land"                    "Super Mario Bros. 3"
##  [7] "Super Smash Bros. Brawl"             "Animal Crossing: Wild World"
##  [9] "Super Mario 64"                      "Super Mario Galaxy"
## [11] "Super Mario Land 2: 6 Golden Coins"  "Super Mario 3D Land"
## [13] "Super Mario All-Stars"               "Super Mario 64"
## [15] "New Super Mario Bros. 2"             "Super Mario Kart"
## [17] "FIFA Soccer 13"                      "Star Wars Battlefront (2015)"
## [19] "Super Smash Bros. for Wii U and 3DS" "Super Mario Galaxy 2"
```

```r
# Barplot of ratings
barplot(table(data[data$Name %in% game_titles & data$Rating != "","Rating"]),
        xlab = 'Rating', ylab = 'Frequency', col = 'gold2',
        main = 'Frequency of Ratings for Most Frequent Words in Video Game Titles')
```



After splitting the name column (video game titles) into individual strings and filtering out stop-words, empty strings, digits and the word "game". I wanted to see if we can find the most commonly used words in video game titles. With some text mining techniques, I found that the words "world","super","star","soccer","dragon" and "wars" have the highest frequency (above 200 and longer than 3 characters). Skimming the top of the filtered dataset for video game titles with the words in the previous sentence, I see that most games are from the Mario series. The barplot indicates that majority of video games were tailored to everyone aka all ages. I can infer that these game titles were meant to attract the general audience and the Mario game series is a perfect example of a game that everyone can enjoy.

Big question: How are video game sales doing?

Conclusion: Based on the year of release variable, video games started rising rapidly in sales in the 2000s. A huge portion of that was due to the PS2 gaming console that had the largest total sales globally. Breaking the sales into different regions, we noticed that NA and EU Sales were strongly correlated positively while NA and EU Sales' correlation with Japan was weak and that can be inferred as a difference in culture of entertainment.

Exploring into the games themselves, I was interested in whether the sales were affected by users(gamers) and critics. Looking at the histogram of critic scores, the scores are close to being normally distributed and have a high frequency in the range of 75-85 on a 100 score scale which are consider fair scores. Similarly, using a simple random sampling with sample size of 30,50,100 and 200, I was able to prove the Central Limit Theorem and calculated the sample means for User Scores which are about 7.1 which matches the population mean with an insignificant margin of error. Also, using various sampling methods such as simple random sampling, systematic sampling and stratified sampling with a sample size of 643 on critic counts on the subset of top 5 genres, we can say the sample mean of critic counts is about 28-29 per video game concluding that the scores from critics are not biased.

Using text mining, I was able to identify the top most used words in a video game title. Skimming just the top of the titles that contain those words, the Mario franchise appeared multiple times and the barplot of age ratings show most games are tailored to the general audience and the Mario game series is a classic example of that.

From all the information above, video game sales are thriving. We can expect that video game sales will stay high as long as there are more production of rated E games, advancement in gaming technology in the near future and a gaming console that be on par or beat PS2 global sales. With higher sales, we should also expect video games to have higher scores from gamers and critics so people will know which games are worthy to purchase.

Resources:

https://www.kaggle.com/gregorut/videogamesales (dataset)

https://journals.sagepub.com/doi/full/10.1177/1046878120945735 (study on Poland (EU) vs US gaming preferences)

https://www.europeanbusinessreview.com/how-has-technology-changed-the-gaming-industry/ (3D gaming aspect)

https://www.ladbible.com/entertainment/gaming-the-ps2-was-the-best-video-games-console-of-all-time-fact-20170303 (PS2)