12/14/2023

AN ANALYSIS OF VIDEO GAME SALES AND RATINGS

CIS 5250: VISUAL ANALYTICS | R PROJECT

PRESENTED TO:

Dr. Shilpa Balan

PRESENTED BY:

Mitali Purohit
Rishabh Shah

MS in Information Systems
California State University, Los Angeles

Table of Contents

1.0 Introduction

1.1 Background

The rise of the video game industry represents a fascinating evolution in entertainment consumption. Over the years, video games have transitioned from niche hobbies to a cultural phenomenon, captivating audiences on a global scale. This transformation is not merely a shift in preferences but a testament to the industry's ability to innovate, adapt, and resonate with diverse audiences. As technology advanced, so did the gaming experience, from the early pixelated adventures to the immersive worlds of today. This project aims to delve into the intricacies of this transformative journey, exploring the factors that have propelled video games into a central position within the broader entertainment landscape.

1.2 Dataset Overview

The dataset comprises information on video game titles, platforms, genres, publishers, sales figures, and critic/user ratings. Spanning multiple years, this dataset offers a rich source of information to explore trends, patterns, and factors influencing the success of video games.

1.3 Project Objectives

Our primary objectives are to conduct Exploratory Data Analysis (EDA) to uncover trends, analyze global and regional sales, examine the impact of ratings, and identify key contributors to the industry's success. Through this analysis, we aim to provide valuable insights for game developers, publishers, and industry enthusiasts.

1.4 Methodology

We utilized the R programming language within the R Studio environment for data manipulation, statistical analysis, and visualization. The dataset was loaded and cleaned to ensure accuracy. Our approach involves a combination of descriptive statistics, visualizations, and advanced analytics to extract meaningful patterns from the data.

1.5 Expected Outcome

By the end of this analysis, we anticipate revealing trends in video game sales, understanding factors influencing ratings, and providing actionable insights for stakeholders in the video game industry.

2.0 Data Description

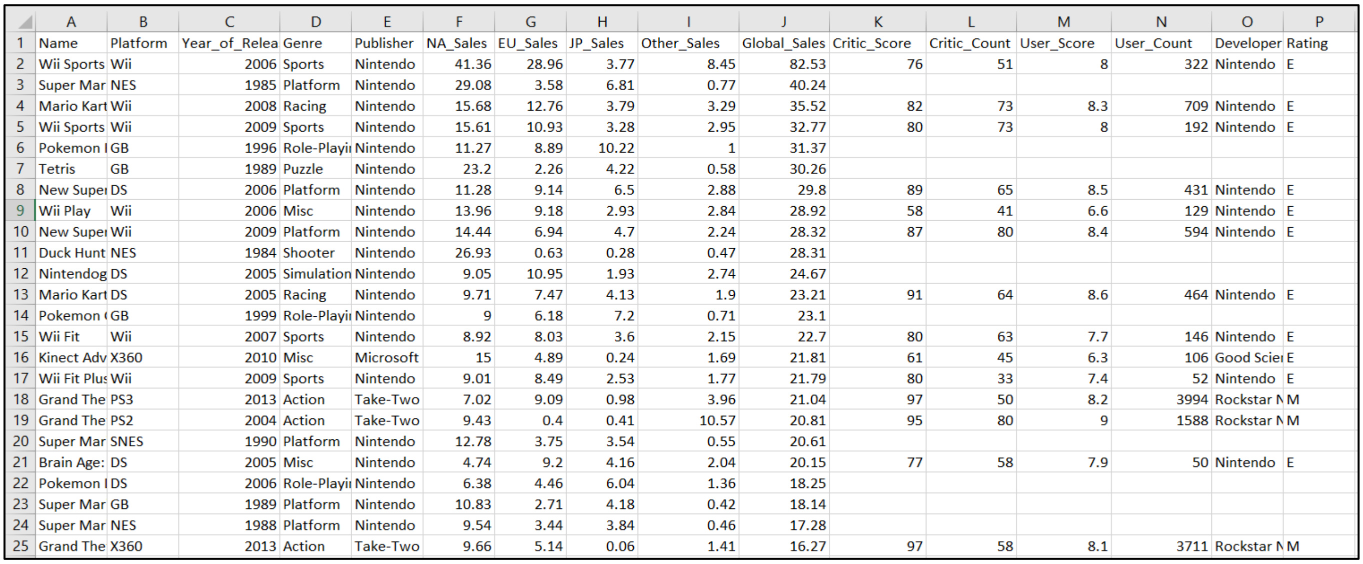
2.1 Dataset Structure

This data set is regarding Video game sales from 1980 to 2022. This data set is collected from Kaggle. The URL to the data set is https://www.kaggle.com/datasets/rush4ratio/video-game-saleswith-ratings. The Video Game Sales dataset contains information about the sales of Video games from 1980 to 2020 for different Genres, Publishers, developers, and Platforms the game was released, and Ratings based on Users and Critics in other parts of the world. This data set contains all the information on video game sales, like the game's name, the genre of the game, publisher, developer, sales in different locations like North America, Europe, Japan, and Global sales, and user and critic-scores based on the user and critic count. It has all the details of video game sales in different parts of the world. The dataset contains 16720 rows and 16 columns, shown below.

2.2 Analytical Opportunities

The dataset's depth and breadth open up a myriad of analytical opportunities. Researchers can employ statistical analyses to uncover correlations between critical acclaim, user reception, and sales performance. Furthermore, temporal trends and regional preferences can be scrutinized to identify emerging markets and evolving consumer behaviors. This dataset provides a fertile ground for machine learning models, enabling predictive analytics in forecasting sales trajectories or identifying potential blockbuster titles. As we embark on this analytical journey, the dataset's 16 columns become conduits for unlocking the intricacies of the video game industry, transforming raw data into actionable intelligence for stakeholders across the gaming spectrum.

Screenshot of Data (showing column/field names and some rows)



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	Name	Platform	Year_of_Relea	Genre	Publisher	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales	Critic_Score	Critic_Count	User_Score	User_Count	Developer	Rating
2	Wii Sports	Wii	2006	Sports	Nintendo	41.36	28.96	3.77	8.45	82.53	76	51	8	322	Nintendo	E
3	Super Mar	NES	1985	Platform	Nintendo	29.08	3.58	6.81	0.77	40.24						
4	Mario Kart	Wii	2008	Racing	Nintendo	15.68	12.76	3.79	3.29	35.52	82	73	8.3	709	Nintendo	E
5	Wii Sports	Wii	2009	Sports	Nintendo	15.61	10.93	3.28	2.95	32.77	80	73	8	192	Nintendo	E
6	Pokemon I	GB	1996	Role-Playi	Nintendo	11.27	8.89	10.22	1	31.37						
7	Tetris	GB	1989	Puzzle	Nintendo	23.2	2.26	4.22	0.58	30.26						
8	New Supe	DS	2006	Platform	Nintendo	11.28	9.14	6.5	2.88	29.8	89	65	8.5	431	Nintendo	E
9	Wii Play	Wii	2006	Misc	Nintendo	13.96	9.18	2.93	2.84	28.92	58	41	6.6	129	Nintendo	E
10	New Supe	Wii	2009	Platform	Nintendo	14.44	6.94	4.7	2.24	28.32	87	80	8.4	594	Nintendo	E
11	Duck Hunt	NES	1984	Shooter	Nintendo	26.93	0.63	0.28	0.47	28.31						
12	Nintendog	DS	2005	Simulation	Nintendo	9.05	10.95	1.93	2.74	24.67						
13	Mario Kart	DS	2005	Racing	Nintendo	9.71	7.47	4.13	1.9	23.21	91	64	8.6	464	Nintendo	E
14	Pokemon (	GB	1999	Role-Playi	Nintendo	9	6.18	7.2	0.71	23.1						
15	Wii Fit	Wii	2007	Sports	Nintendo	8.92	8.03	3.6	2.15	22.7	80	63	7.7	146	Nintendo	E
16	Kinect Adv	X360	2010	Misc	Microsoft	15	4.89	0.24	1.69	21.81	61	45	6.3	106	Good Scie	E
17	Wii Fit Plus	Wii	2009	Sports	Nintendo	9.01	8.49	2.53	1.77	21.79	80	33	7.4	52	Nintendo	E
18	Grand The	PS3	2013	Action	Take-Two	7.02	9.09	0.98	3.96	21.04	97	50	8.2	3994	Rockstar N	M
19	Grand The	PS2	2004	Action	Take-Two	9.43	0.4	0.41	10.57	20.81	95	80	9	1588	Rockstar N	M
20	Super Mar	SNES	1990	Platform	Nintendo	12.78	3.75	3.54	0.55	20.61						
21	Brain Age:	DS	2005	Misc	Nintendo	4.74	9.2	4.16	2.04	20.15	77	58	7.9	50	Nintendo	E
22	Pokemon I	DS	2006	Role-Playi	Nintendo	6.38	4.46	6.04	1.36	18.25						
23	Super Mar	GB	1989	Platform	Nintendo	10.83	2.71	4.18	0.42	18.14						
24	Super Mar	NES	1988	Platform	Nintendo	9.54	3.44	3.84	0.46	17.28						
25	Grand The	X360	2013	Action	Take-Two	9.66	5.14	0.06	1.41	16.27	97	58	8.1	3711	Rockstar N	M

The dataset comprises the following columns, each providing specific information about the video games:

Column Name	Description	Example
Name	Title or name of the video game. Date Type: Char	Wii Sports, Pokemon Red/Pokemon Blue, Duck Hunt
Platform	Gaming platform on which the video game is available. Data Type: Char	Wii, NES, GB, X360
Year_of_Release	Year when the video game was released. Data Type: Numeric	12/15/2006, 3/16/1985, 7/23/2008
Genre	Genre or category of the video game. Data Type: Char	Sports, Platform, Racing
Publisher	Company responsible for publishing the video game. Data Type: Char	Nintendo, Microsoft Game Studios, Take-Two Interactive
NA_Sales	Sales figures for the video game in North America (in millions of units). Data Type: Numeric Data Type: Numeric	41.36 million, 29.08 million, 15 million
EU_Sales	Sales figures for the video game in Europe (in millions of units). Data Type: Numeric	10.93 million, 2.26 million, 8.03 million
JP_Sales	Sales figures for the video game in Japan (in millions of units). Data Type: Numeric	0.58 million, 1.9 million, 2.15 million
Other_Sales	Sales figures for the video game in other regions (in millions of units). Data Type: Numeric	8.45 million, 1 million, 0.47 million
Global_Sales	Total global sales figure for the video game (sum of NA, EU, JP, and Other sales). Data Type: Numeric	40.24 million, 82.53 million, 28.31 million
Critic_Score	Score assigned to the video game by critics. Data	80, 82, 76

	Type: Numeric	
Critic_Count	Number of critics who provided a score for the video game. Data Type: Numeric	51, 80, 891
User_Score	Score assigned to the video game by users or players. Data Type: Numeric	8, 8.3, 7.6
User_Count	Number of users who provided a score for the video game.	322, 1588, 3994
Developer	Company or individual responsible for developing the video game. Data Type: Char	Game Arts, Treyarch, Polyphony Digital
Rating	Content rating assigned to the video game (e.g., E for Everyone, M for Mature). Data Type: Char	E, M, T

3.0 Data Cleaning

Data cleansing is the process of identifying and fixing problems in datasets. The purpose of data cleansing is to correct data that is inaccurate, incomplete, erroneous, redundant, or irrelevant to the purpose of the dataset. This is typically accomplished by replacing, modifying, or deleting data that falls into one of these categories. Combining multiple data sources can result in duplicated or mislabelled data. Inaccurate data can make results and algorithms look correct but unreliable. There is no absolute way to specify the exact steps of the data cleansing process, as the process is different for each data set. Our decisions are usually based on datasets. Therefore, if the data quality is low, the results will not be accurate. Therefore, data cleansing is essential to obtain quality data that 7 leads to better decisions. Not all data in a dataset is good data. There was a little junk data. The data set used for this analysis contained some null values. Some datasets had empty or missing values, so the data was removed and filtered while focusing on the required datasets. Unnecessary columns have been removed, and some queues have been split. Below are some steps that are performed to clean up the records.

1. Data cleaning category name:

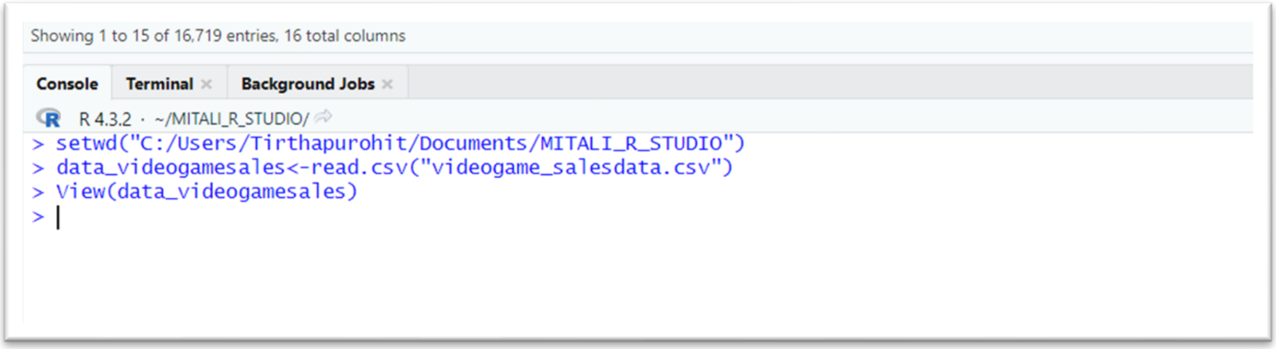
The "Publisher" and "Developer" columns often contain duplicate information, especially for smaller studios that handle both roles. Merging them eliminates redundancy, reduces data size, and improves storage efficiency.

Merging the columns clarifies the relationship between publisher and developer, eliminating confusion about who is responsible for each game. This ensures consistent interpretation and avoids misinterpreting data points.

Working with a single "Publisher/Developer" column simplifies various tasks:

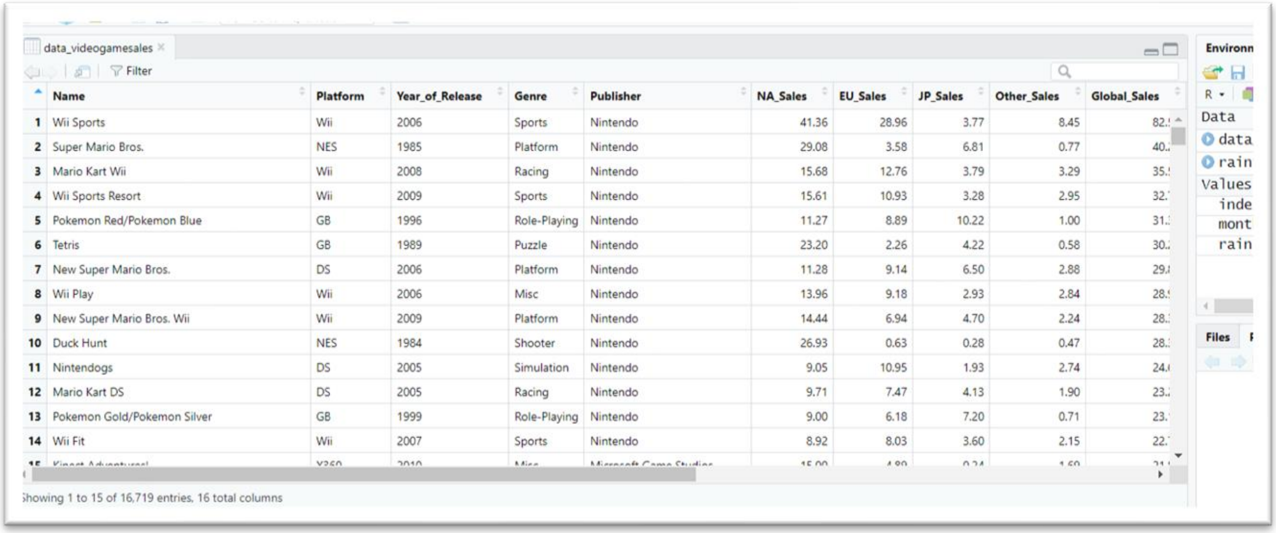
- Identifying dominant companies in the market based on combined game sales.
- Analyzing market share trends and their impact on sales performance.
- Building predictive models based on combined publisher/developer expertise.

```
> setwd("C:/Users/Tirthapurohit/Documents/MITALI_R_STUDIO")
>  data_videogamesales<-read.csv("videogame_salesdata.csv")
> View(data_videogamesales)
```



Showing 1 to 15 of 16,719 entries, 16 total columns

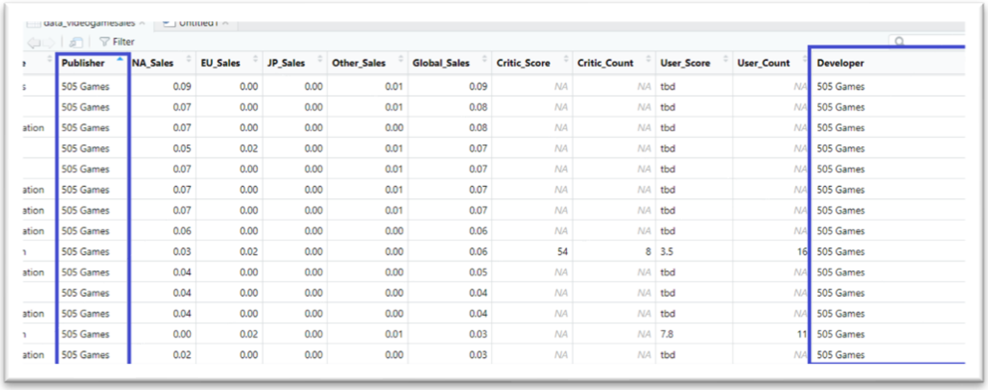
```
> setwd("C:/Users/Tirthapurohit/Documents/MITALI_R_STUDIO")
> data_videogamesales<-read.csv("videogame_salesdata.csv")
> View(data_videogamesales)
> |
```



	Name	Platform	Year_of_Release	Genre	Publisher	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales
1	Wii Sports	Wii	2006	Sports	Nintendo	41.36	28.96	3.77	8.45	82.!
2	Super Mario Bros.	NES	1985	Platform	Nintendo	29.08	3.58	6.81	0.77	40..
3	Mario Kart Wii	Wii	2008	Racing	Nintendo	15.68	12.76	3.79	3.29	35.!
4	Wii Sports Resort	Wii	2009	Sports	Nintendo	15.61	10.93	3.28	2.95	32.'
5	Pokemon Red/Pokemon Blue	GB	1996	Role-Playing	Nintendo	11.27	8.89	10.22	1.00	31.:
6	Tetris	GB	1989	Puzzle	Nintendo	23.20	2.26	4.22	0.58	30..
7	New Super Mario Bros.	DS	2006	Platform	Nintendo	11.28	9.14	6.50	2.88	29.!
8	Wii Play	Wii	2006	Misc	Nintendo	13.96	9.18	2.93	2.84	28.!
9	New Super Mario Bros. Wii	Wii	2009	Platform	Nintendo	14.44	6.94	4.70	2.24	28.:
10	Duck Hunt	NES	1984	Shooter	Nintendo	26.93	0.63	0.28	0.47	28.:
11	Nintendogs	DS	2005	Simulation	Nintendo	9.05	10.95	1.93	2.74	24.(
12	Mario Kart DS	DS	2005	Racing	Nintendo	9.71	7.47	4.13	1.90	23..
13	Pokemon Gold/Pokemon Silver	GB	1999	Role-Playing	Nintendo	9.00	6.18	7.20	0.71	23.'
14	Wii Fit	Wii	2007	Sports	Nintendo	8.92	8.03	3.60	2.15	22.'
15	Kinect Adventures!	X360	2010	Misc	Microsoft Game Studios	15.00	4.89	0.24	1.69	21.!

Showing 1 to 15 of 16,719 entries, 16 total columns

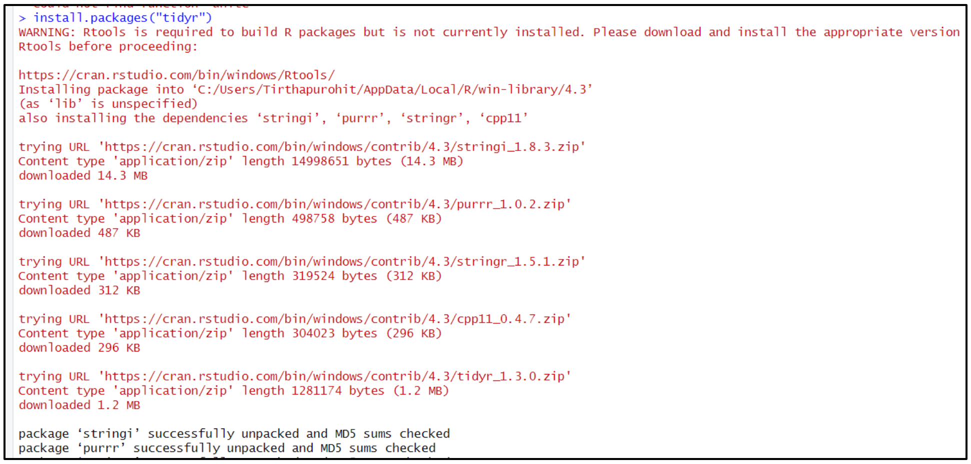
Screenshot of the data before cleaning:



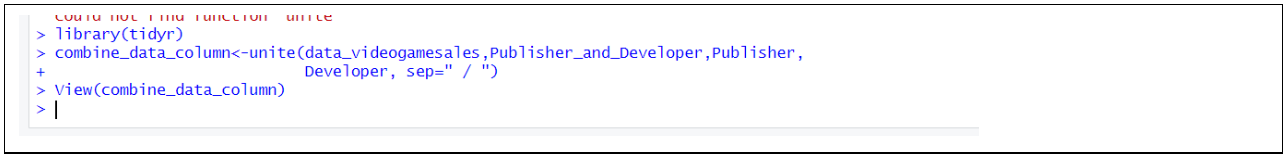
e	Publisher	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales	Critic_Score	Critic_Count	User_Score	User_Count	Developer
s	505 Games	0.09	0.00	0.00	0.01	0.09	NA	NA	tbd	NA	505 Games
	505 Games	0.07	0.00	0.00	0.01	0.08	NA	NA	tbd	NA	505 Games
ation	505 Games	0.07	0.00	0.00	0.00	0.08	NA	NA	tbd	NA	505 Games
	505 Games	0.05	0.02	0.00	0.01	0.07	NA	NA	tbd	NA	505 Games
	505 Games	0.07	0.00	0.00	0.01	0.07	NA	NA	tbd	NA	505 Games
ation	505 Games	0.07	0.00	0.00	0.01	0.07	NA	NA	tbd	NA	505 Games
ation	505 Games	0.07	0.00	0.00	0.01	0.07	NA	NA	tbd	NA	505 Games
ation	505 Games	0.06	0.00	0.00	0.00	0.06	NA	NA	tbd	NA	505 Games
n	505 Games	0.03	0.02	0.00	0.00	0.06	54	8	3.5	16	505 Games
ation	505 Games	0.04	0.00	0.00	0.00	0.05	NA	NA	tbd	NA	505 Games
	505 Games	0.04	0.00	0.00	0.00	0.04	NA	NA	tbd	NA	505 Games
ation	505 Games	0.04	0.00	0.00	0.00	0.04	NA	NA	tbd	NA	505 Games
n	505 Games	0.00	0.02	0.00	0.01	0.03	NA	NA	7.8	11	505 Games
ation	505 Games	0.02	0.00	0.00	0.00	0.03	NA	NA	tbd	NA	505 Games

```
> install.packages("tidyr")
> library(tidyr)
> combine_data_column<-unite(data_videogamesales,Publisher_and_Developer,Publisher,
+                       Developer, sep=" / ")
> View(combine_data_column)
```

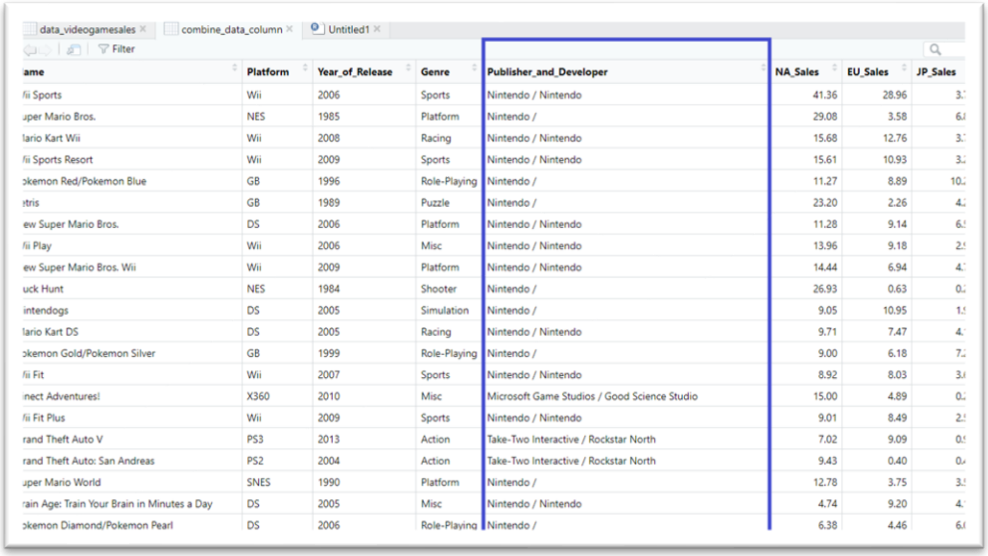
Screenshot of the results in R Studio



```
> install.packages("tidyr")
WARNING: Rtools is required to build R packages but is not currently installed. Please download and install the appropriate version
Rtools before proceeding:

https://cran.rstudio.com/bin/windows/Rtools/
Installing package into ‘C:/Users/Tirthapurohit/AppData/Local/R/win-library/4.3’
(as ‘lib’ is unspecified)
also installing the dependencies ‘stringi’, ‘purrr’, ‘stringr’, ‘cpp11’

trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.3/stringi_1.8.3.zip'
Content type 'application/zip' length 14998651 bytes (14.3 MB)
downloaded 14.3 MB

trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.3/purrr_1.0.2.zip'
Content type 'application/zip' length 498758 bytes (487 KB)
downloaded 487 KB

trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.3/stringr_1.5.1.zip'
Content type 'application/zip' length 319524 bytes (312 KB)
downloaded 312 KB

trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.3/cpp11_0.4.7.zip'
Content type 'application/zip' length 304023 bytes (296 KB)
downloaded 296 KB

trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.3/tidyr_1.3.0.zip'
Content type 'application/zip' length 1281174 bytes (1.2 MB)
downloaded 1.2 MB

package ‘stringi’ successfully unpacked and MD5 sums checked
package ‘purrr’ successfully unpacked and MD5 sums checked
```



```
  could not find function "unite"
> library(tidyr)
> combine_data_column<-unite(data_videogamesales,Publisher_and_Developer,Publisher,
+                       Developer, sep=" / ")
> View(combine_data_column)
> |
```



ame	Platform	Year_of_Release	Genre	Publisher_and_Developer	NA_Sales	EU_Sales	JP_Sales
/ii Sports	Wii	2006	Sports	Nintendo / Nintendo	41.36	28.96	3.
uper Mario Bros.	NES	1985	Platform	Nintendo /	29.08	3.58	6.
lario Kart Wii	Wii	2008	Racing	Nintendo / Nintendo	15.68	12.76	3.
/ii Sports Resort	Wii	2009	Sports	Nintendo / Nintendo	15.61	10.93	3.
okemon Red/Pokemon Blue	GB	1996	Role-Playing	Nintendo /	11.27	8.89	10.
etris	GB	1989	Puzzle	Nintendo /	23.20	2.26	4.
ew Super Mario Bros.	DS	2006	Platform	Nintendo / Nintendo	11.28	9.14	6.
/ii Play	Wii	2006	Misc	Nintendo / Nintendo	13.96	9.18	2.
ew Super Mario Bros. Wii	Wii	2009	Platform	Nintendo / Nintendo	14.44	6.94	4.
uck Hunt	NES	1984	Shooter	Nintendo /	26.93	0.63	0.
intendogs	DS	2005	Simulation	Nintendo /	9.05	10.95	1.
lario Kart DS	DS	2005	Racing	Nintendo / Nintendo	9.71	7.47	4.
okemon Gold/Pokemon Silver	GB	1999	Role-Playing	Nintendo /	9.00	6.18	7.
/ii Fit	Wii	2007	Sports	Nintendo / Nintendo	8.92	8.03	3.
inect Adventures!	X360	2010	Misc	Microsoft Game Studios / Good Science Studio	15.00	4.89	0.
/ii Fit Plus	Wii	2009	Sports	Nintendo / Nintendo	9.01	8.49	2.
rand Theft Auto V	PS3	2013	Action	Take-Two Interactive / Rockstar North	7.02	9.09	0.
rand Theft Auto: San Andreas	PS2	2004	Action	Take-Two Interactive / Rockstar North	9.43	0.40	0.
uper Mario World	SNES	1990	Platform	Nintendo /	12.78	3.75	3.
rain Age: Train Your Brain in Minutes a Day	DS	2005	Misc	Nintendo / Nintendo	4.74	9.20	4.
okemon Diamond/Pokemon Pearl	DS	2006	Role-Playing	Nintendo /	6.38	4.46	6.

4.0 Data Analysis and Visualization

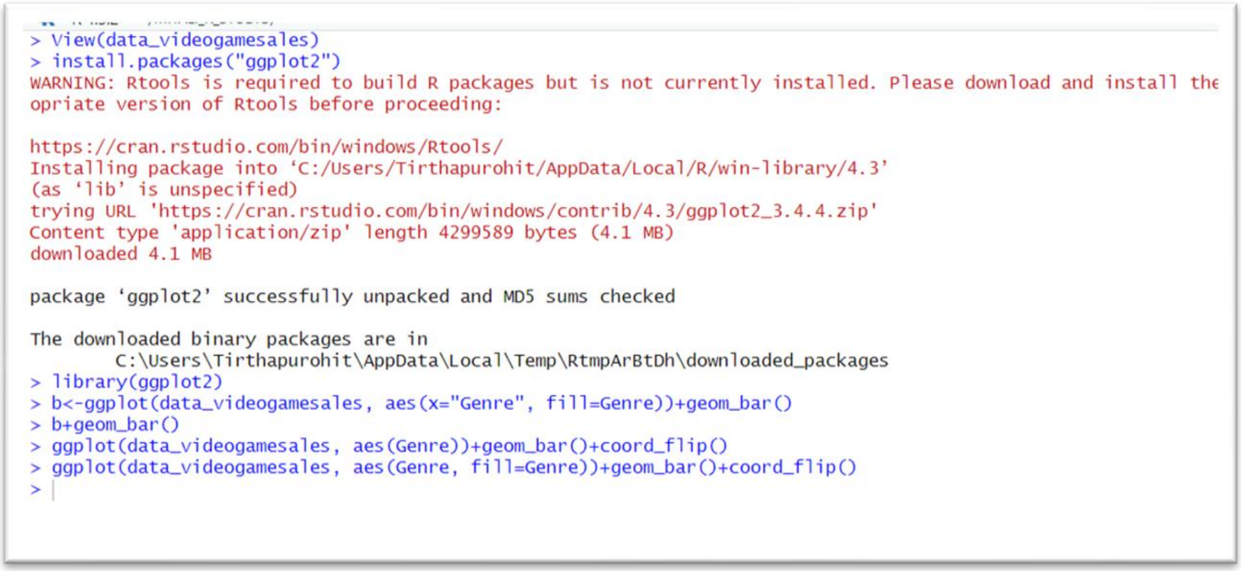
Data visualization is a graphical representation of information and data. Data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in your data using visual elements such as charts, graphs, and maps. Data visualization tools and technology are essential for analyzing large amounts of information and making informed decisions with big data.

1. Which game genres have gained the most popularity in the gaming community?

R Studio code:

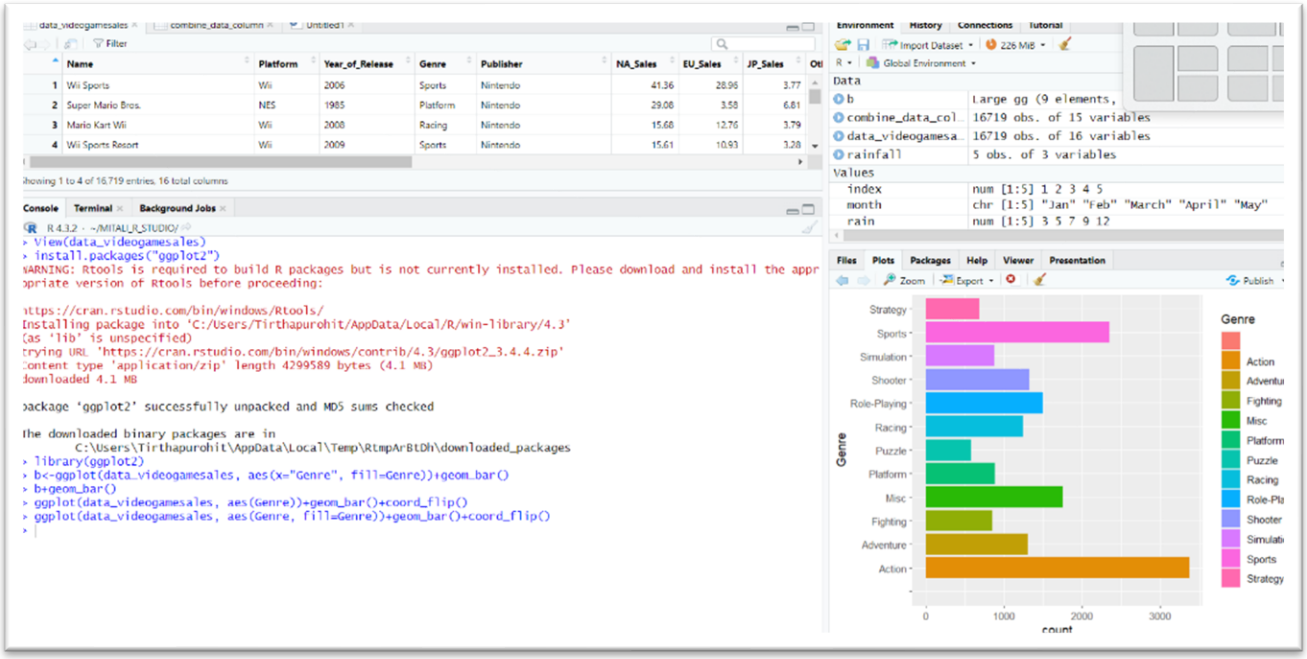
```
> View(data_videogamesales)
> install.packages("ggplot2")
> library(ggplot2)
> b<-ggplot(data_videogamesales, aes(x="Genre", fill=Genre))+geom_bar()
> b+geom_bar()
> ggplot(data_videogamesales, aes(Genre))+geom_bar()+coord_flip()
> ggplot(data_videogamesales, aes(Genre, fill=Genre))+geom_bar()+coord_flip()
```

Screenshot of the code in R Studio:



```
> View(data_videogamesales)
> install.packages("ggplot2")
WARNING: Rtools is required to build R packages but is not currently installed. Please download and install the
opriate version of Rtools before proceeding:

https://cran.rstudio.com/bin/windows/Rtools/
Installing package into ‘C:/Users/Tirthapurohit/AppData/Local/R/win-library/4.3’
(as ‘lib’ is unspecified)
trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.3/ggplot2_3.4.4.zip'
Content type 'application/zip' length 4299589 bytes (4.1 MB)
downloaded 4.1 MB

package ‘ggplot2’ successfully unpacked and MD5 sums checked

The downloaded binary packages are in
        C:\Users\Tirthapurohit\AppData\Local\Temp\RtmpArBtDh\downloaded_packages
> library(ggplot2)
> b<-ggplot(data_videogamesales, aes(x="Genre", fill=Genre))+geom_bar()
> b+geom_bar()
> ggplot(data_videogamesales, aes(Genre))+geom_bar()+coord_flip()
> ggplot(data_videogamesales, aes(Genre, fill=Genre))+geom_bar()+coord_flip()
> 
```

Screenshot of the code and results:



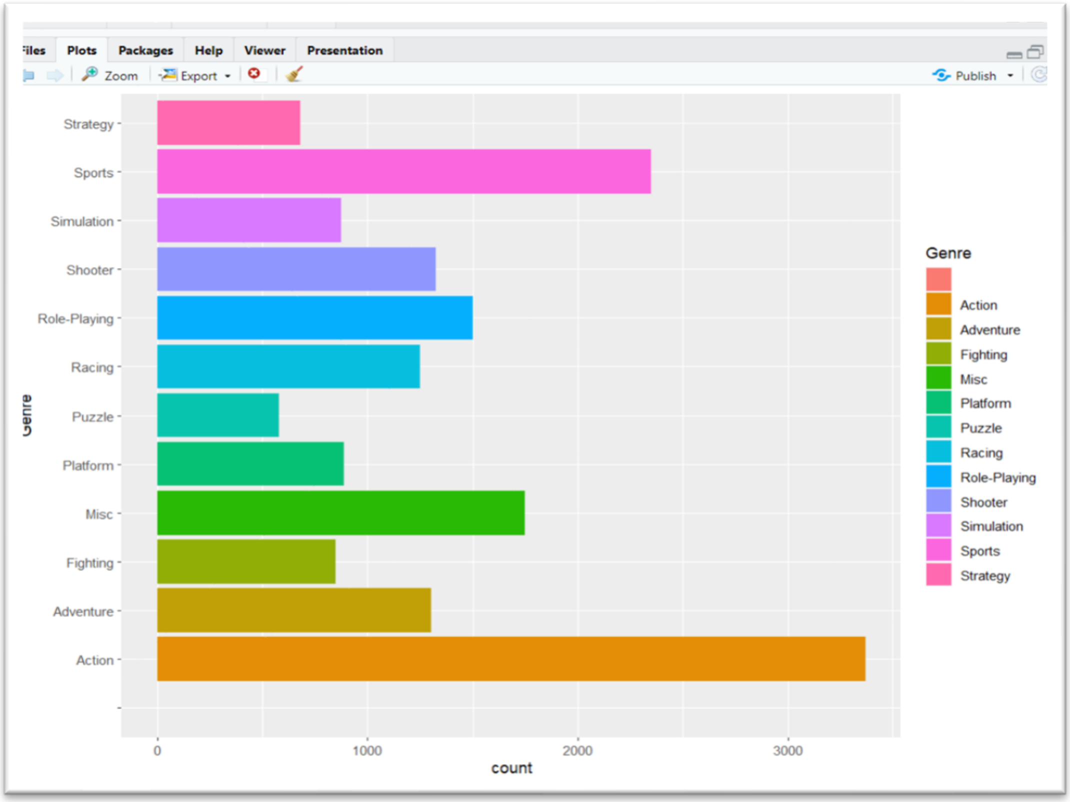
	Name	Platform	Year_of_Release	Genre	Publisher	NA_Sales	EU_Sales	JP_Sales	Otl
1	Wii Sports	Wii	2006	Sports	Nintendo	41.36	28.96	3.77	
2	Super Mario Bros.	NES	1985	Platform	Nintendo	29.08	3.58	6.81	
3	Mario Kart Wii	Wii	2008	Racing	Nintendo	15.68	12.76	3.79	
4	Wii Sports Resort	Wii	2009	Sports	Nintendo	15.61	10.93	3.28	

howing 1 to 4 of 16,719 entries, 16 total columns

```
> View(data_videogamesales)
> install.packages("ggplot2")
WARNING: Rtools is required to build R packages but is not currently installed. Please download and install the appr
opriate version of Rtools before proceeding:

https://cran.rstudio.com/bin/windows/Rtools/
Installing package into 'C:/Users/Tirthapurohit/AppData/Local/R/win-library/4.3'
(as 'lib' is unspecified)
trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.3/ggplot2_3.4.4.zip'
Content type 'application/zip' length 4299589 bytes (4.1 MB)
downloaded 4.1 MB

package 'ggplot2' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
        C:\Users\Tirthapurohit\AppData\Local\Temp\RtmpArBtDh\downloaded_packages
> library(ggplot2)
> b<-ggplot(data_videogamesales, aes(x="Genre", fill=Genre))+geom_bar()
> b+geom_bar()
> ggplot(data_videogamesales, aes(Genre))+geom_bar()+coord_flip()
> ggplot(data_videogamesales, aes(Genre, fill=Genre))+geom_bar()+coord_flip()
> 
```

```
Data
b                     Large gg (9 elements,
combine_data_col…     16719 obs. of 15 variables
data_videogamesa…     16719 obs. of 16 variables
rainfall              5 obs. of 3 variables
Values
  index               num [1:5] 1 2 3 4 5
  month               chr [1:5] "Jan" "Feb" "March" "April" "May"
  rain                num [1:5] 3 5 7 9 12
```

Screenshot of the Bar chart:



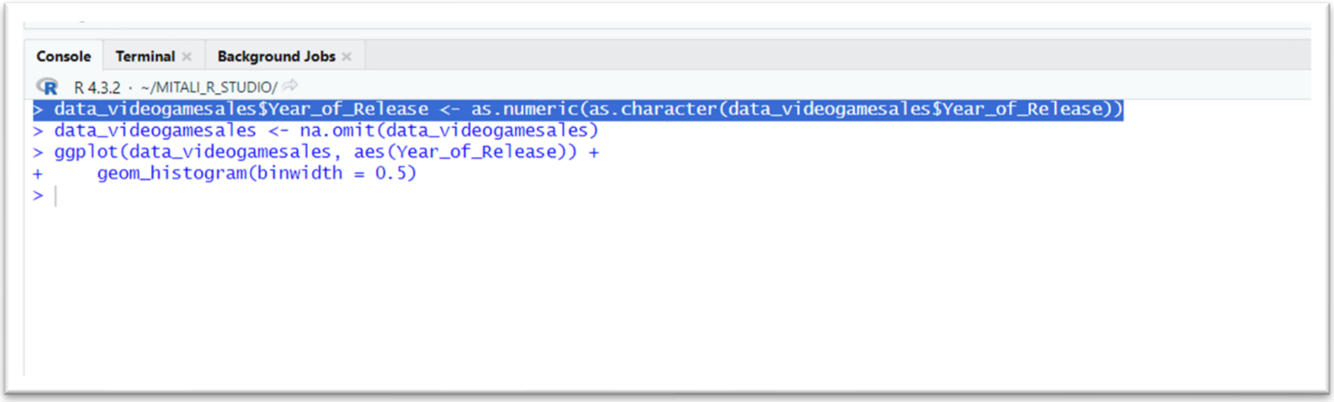
Insight: In the above visualization, we have used a bar chart and grouped the count of the games for the genre to show the genres that have gained the most popularity in the gaming community. Here, all the games developed till now are categorized into different genres. The game genre offe rs a variety of games to group together. Games with the same mechanics and playstyle are groupe d into one genre. Its primary purpose is to categorize games so that consumers know what they li ke and buy similar games. According to the data, the action genre has become the most popular c ompared to other genres. In a sense, action games have significantly impacted users. Almost all t he topdeveloped games are somewhere related to combats, fighting, adventure, or shooting, and t hey are mostly multiple people playing games. Sports, role-play, and misc games are also on the most popular list. Here the games are being developed on the current popularity of the specific ge nre. The sales of Action & Sports genre games are high compared to other genres. Puzzles and St rategies have a minor popularity compared to different genres. As these genres of games are the m ost popular among all, this visualization helps us know the genre's popularity.

2. How have "Video games" evolved since their inception in the 1980s?

R Studio code:

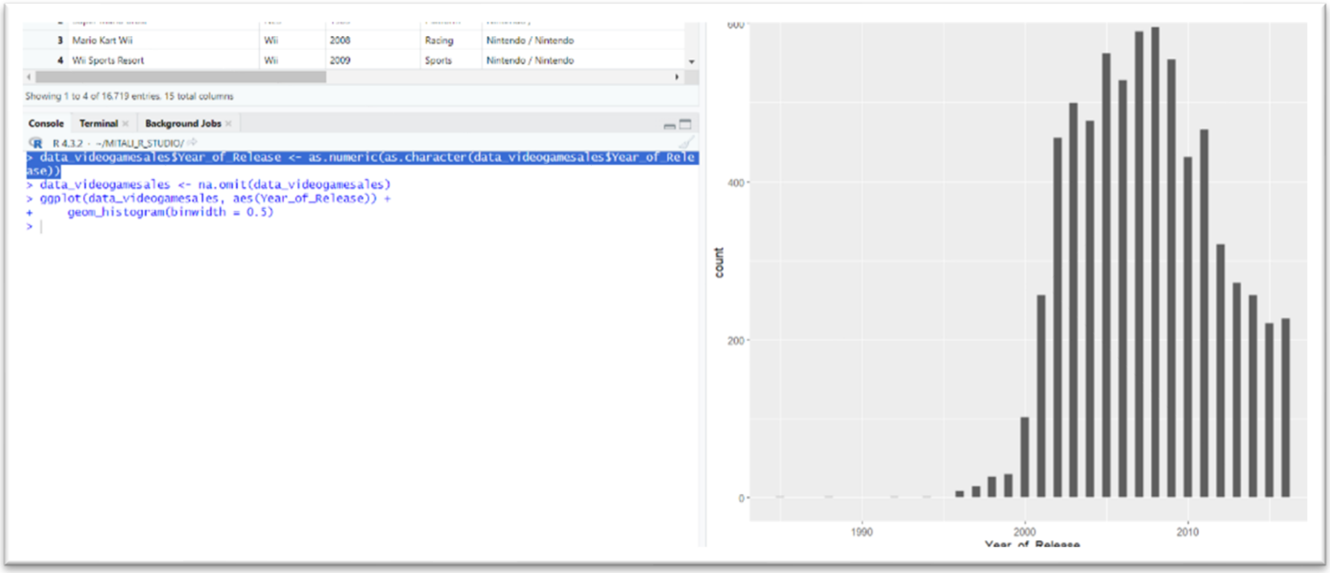
```
> data_videogamesales$Year_of_Release <- as.numeric(as.character(data_videogamesales$Year
_of_Release))
> data_videogamesales <- na.omit(data_videogamesales)
> ggplot(data_videogamesales, aes(Year_of_Release)) +
+    geom_histogram(binwidth = 0.5)
```

Screenshot of the code in R Studio:



```
> data_videogamesales$Year_of_Release <- as.numeric(as.character(data_videogamesales$Year_of_Release))
> data_videogamesales <- na.omit(data_videogamesales)
> ggplot(data_videogamesales, aes(Year_of_Release)) +
+     geom_histogram(binwidth = 0.5)
> |
```

Screenshot of the code in R Studio of Output:

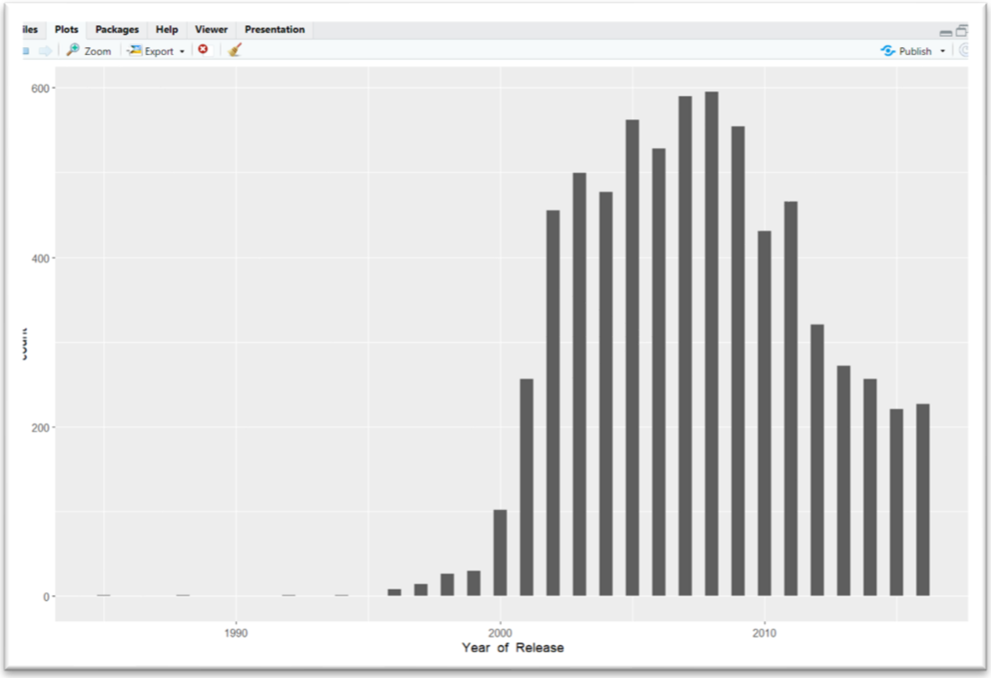


3	Mario Kart Wii	Wii	2008	Racing	Nintendo / Nintendo
4	Wii Sports Resort	Wii	2009	Sports	Nintendo / Nintendo

Showing 1 to 4 of 16,719 entries, 15 total columns

```
> data_videogamesales$Year_of_Release <- as.numeric(as.character(data_videogamesales$Year_of_Rele
ase))
> data_videogamesales <- na.omit(data_videogamesales)
> ggplot(data_videogamesales, aes(Year_of_Release)) +
+     geom_histogram(binwidth = 0.5)
> 
```

Screenshot the histogram chart:



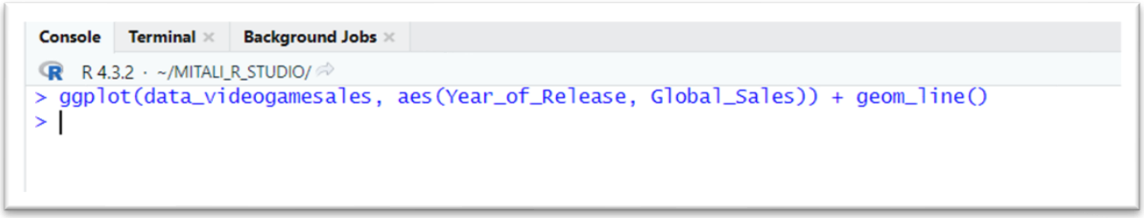
Insight: In the above visualization, we have used a histogram to represent the data for the respective years of the video game released. Since the inception of video games in the 1980s, in the very initial stages of the category, the number of games published per year has been less than 100 for the next ten years, i.e., the year 1990. It is the same phase where the purchase of personal computers has gradually increased. By seeing the number of games published per year, we can say that the publishing companies are also on the same levels where the competition is meager, keeping the gaming industry's demand in mind. The gaming industry has been seeing a golden phase since 1995 and is continuing; from 2005 to 2012, the industry has seen exponential growth. From this, we can say that seeing the development of the industry; there have been many publishers who entered the gaming field. Still, not many publishers could sustain themselves because of the competition. We can say that the games have been published by each publisher, who were the 20 companies that suffered because of developing good quality content. Also, by seeing that, we can tell the dominance of specific companies in the market. Later after seeing the drastic growth in games produced in the years 2005- 2012, the number of games produced took a hit but was steady as only the best in business were able to sustain and dominate the industry.

3. How did the games generate revenue globally during their evolution as an industry?

R Studio code:

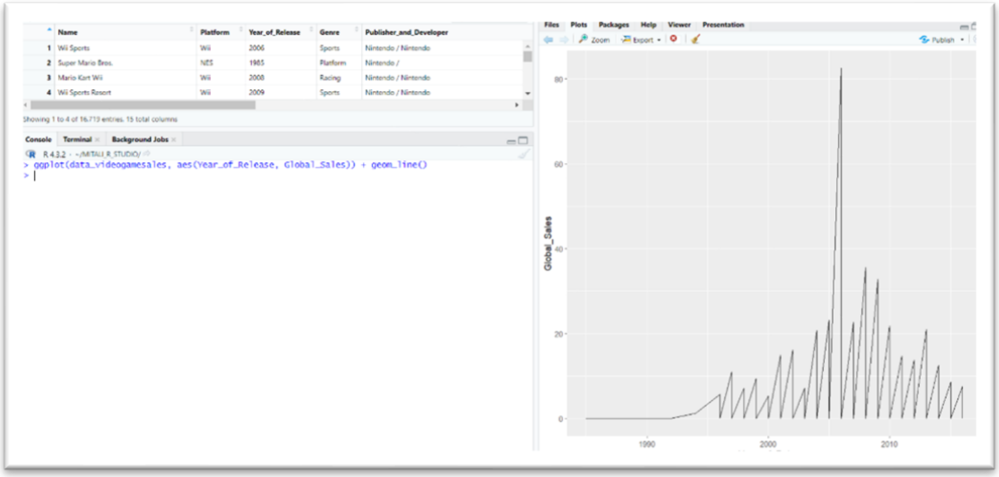
```
> ggplot(data_videogamesales, aes(Year_of_Release, Global_Sales)) + geom_line()
```

Screenshot of the code in R Studio:



```
R 4.3.2 · ~/MITALI_R_STUDIO/
> ggplot(data_videogamesales, aes(Year_of_Release, Global_Sales)) + geom_line()
> |
```

Screenshot of the code in R Studio Output:

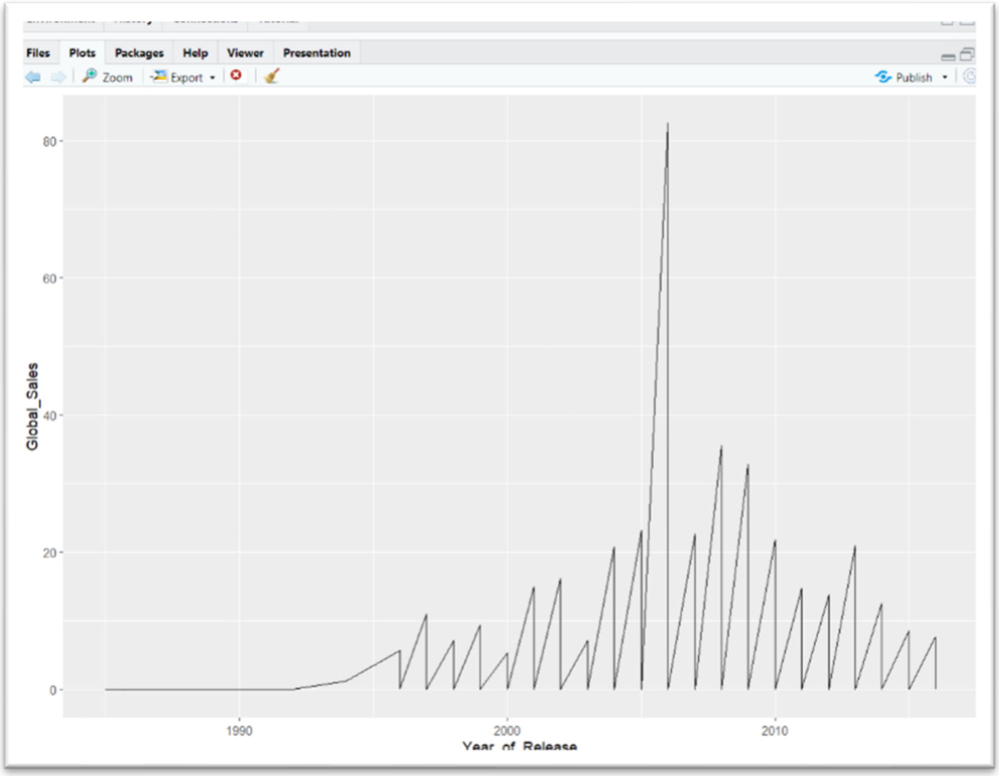


	Name	Platform	Year_of_Release	Genre	Publisher_and_Developer
1	Wii Sports	Wii	2006	Sports	Nintendo / Nintendo
2	Super Mario Bros.	NES	1985	Platform	Nintendo /
3	Mario Kart Wii	Wii	2008	Racing	Nintendo / Nintendo
4	Wii Sports Resort	Wii	2009	Sports	Nintendo / Nintendo

Showing 1 to 4 of 16,719 entries, 15 total columns

```
R 4.3.2 · ~/MITALI_R_STUDIO/
> ggplot(data_videogamesales, aes(Year_of_Release, Global_Sales)) + geom_line()
> |
```

Screenshot of the Line chart:



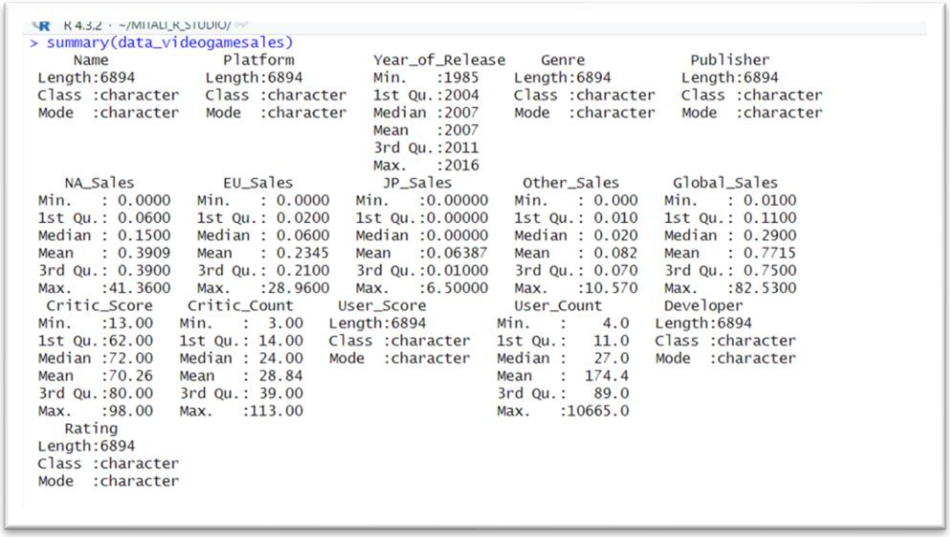
Insight: In the above visualization, we have used a ‘line chart’ and have grouped the sales for their release years to show the sales. We can see the sales of video games globally since their inception in the 1980s. We can see that the sales are fluctuating up and down. Few years sales will be very low, and next few years, there will be a sudden rise in sales. The year 2005 has the higher sales. Except for the year 2005, there is an up and down in video game sales. There could be many reasons for this variation in sales. In the initial years until 1983, the sales in all regions were steady as when the personal computer segment had not yet grown, it can be seen as one of the causes of fewer sales. After 1983, the personal computer segment increased drastically, which helped the gaming industry's growth. Personal computers were evolving in a few parts of the world. But throughout all these years, the ‘rest of the world’s countries didn’t show much growth though the personal computer market was on the rise, as the gaming culture in those countries didn’t evolve so much as compared to a few countries like North America, Japan, and Europe. Even though there weren’t 22 many developers in those parts of the world, it can also be said to be one of the reasons for the gaming industry not evolving as much as these countries in the rest of the parts of the world. In 2020, the gaming industry took a hit because of covid as the production of semiconductors has been stopped, and the complete supply chain has been disrupted. Apart from this, the games are available for free on torrent sites. The gaming industry has struggled for the past 6-8 years and is trying to overcome the issue.

5.0 Summary Statistics

1. Statistical analysis for all the variables

Code for summary statistics in R Studio:

```
>summary(data_videogamesales)
```



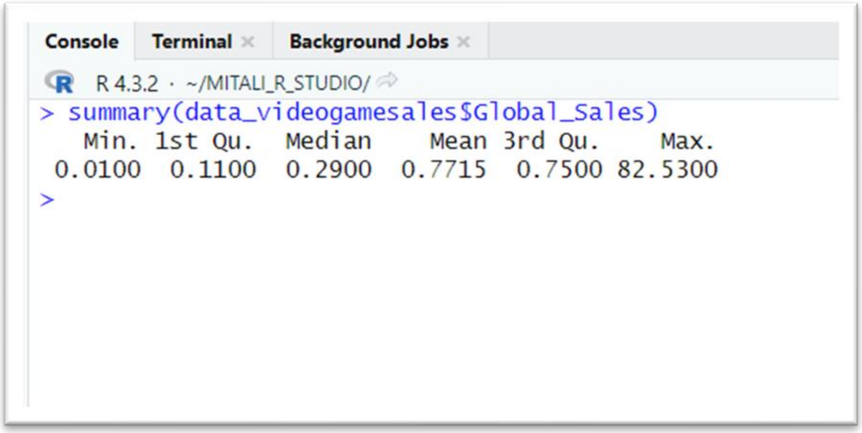
```
R 4.3.2 · ~/MITALI_R_STUDIO/
> summary(data_videogamesales)
     Name             Platform         Year_of_Release     Genre            Publisher        
 Length:6894        Length:6894        Min.   :1985    Length:6894        Length:6894       
 Class :character   Class :character   1st Qu.:2004    Class :character   Class :character  
 Mode  :character   Mode  :character   Median :2007    Mode  :character   Mode  :character  
                                       Mean   :2007                                         
                                       3rd Qu.:2011                                         
                                       Max.   :2016                                         
    NA_Sales          EU_Sales          JP_Sales         Other_Sales       Global_Sales    
 Min.   : 0.0000   Min.   : 0.0000   Min.   :0.00000   Min.   : 0.000   Min.   : 0.0100  
 1st Qu.: 0.0600   1st Qu.: 0.0200   1st Qu.:0.00000   1st Qu.: 0.010   1st Qu.: 0.1100  
 Median : 0.1500   Median : 0.0600   Median :0.00000   Median : 0.020   Median : 0.2900  
 Mean   : 0.3909   Mean   : 0.2345   Mean   :0.06387   Mean   : 0.082   Mean   : 0.7715  
 3rd Qu.: 0.3900   3rd Qu.: 0.2100   3rd Qu.:0.01000   3rd Qu.: 0.070   3rd Qu.: 0.7500  
 Max.   :41.3600   Max.   :28.9600   Max.   :6.50000   Max.   :10.570   Max.   :82.5300  
  Critic_Score    Critic_Count     User_Score          User_Count        Developer        
 Min.   :13.00   Min.   :  3.00   Length:6894        Min.   :    4.0   Length:6894       
 1st Qu.:62.00   1st Qu.: 14.00   Class :character   1st Qu.:   11.0   Class :character  
 Median :72.00   Median : 24.00   Mode  :character   Median :   27.0   Mode  :character  
 Mean   :70.26   Mean   : 28.84                      Mean   :  174.4                     
 3rd Qu.:80.00   3rd Qu.: 39.00                      3rd Qu.:   89.0                     
 Max.   :98.00   Max.   :113.00                      Max.   :10665.0                     
    Rating         
 Length:6894       
 Class :character  
 Mode  :character  
```

2. Statistical analysis of the variable Global_Sales

Code for Statistical analysis of the variable Global_Sales in R Studio

```
> summary(data_videogamesales$Global_Sales)
```

Screenshot of the code in R Studio:

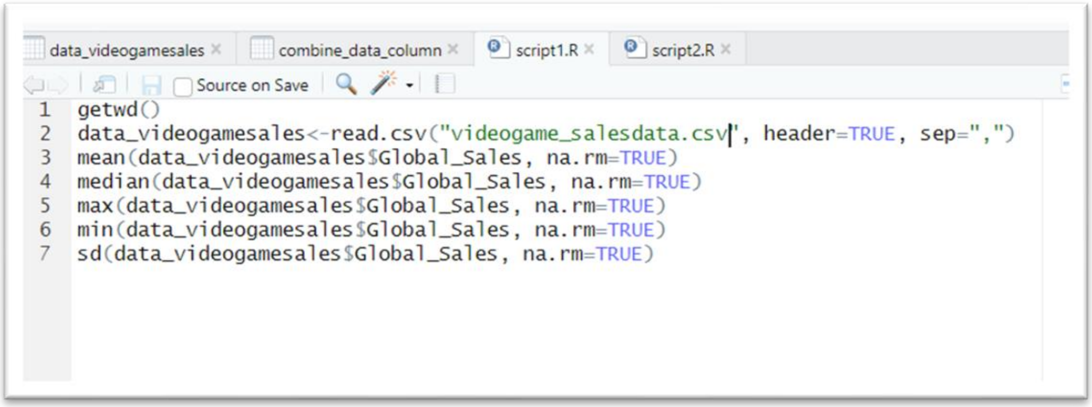


```
R 4.3.2 · ~/MITALI_R_STUDIO/
> summary(data_videogamesales$Global_Sales)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 0.0100  0.1100  0.2900  0.7715  0.7500 82.5300 
> 
```

Code for individual Statistical analysis of the variable Global_Sales in R Studio:

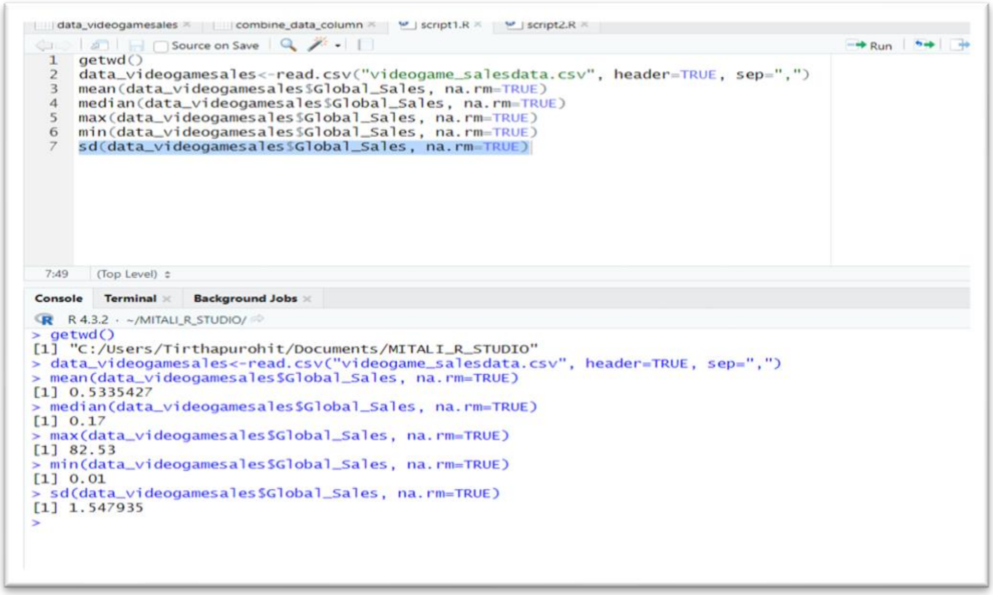
```
getwd()
data_videogamesales<-read.csv("videogame_salesdata.csv", header=TRUE, sep=",")
mean(data_videogamesales$Global_Sales, na.rm=TRUE)
median(data_videogamesales$Global_Sales, na.rm=TRUE)
max(data_videogamesales$Global_Sales, na.rm=TRUE)
min(data_videogamesales$Global_Sales, na.rm=TRUE)
sd(data_videogamesales$Global_Sales, na.rm=TRUE)
```

Screenshot of the code of the variable Global_Sales in R Studio:



```
getwd()
data_videogamesales<-read.csv("videogame_salesdata.csv", header=TRUE, sep=",")
mean(data_videogamesales$Global_Sales, na.rm=TRUE)
median(data_videogamesales$Global_Sales, na.rm=TRUE)
max(data_videogamesales$Global_Sales, na.rm=TRUE)
min(data_videogamesales$Global_Sales, na.rm=TRUE)
sd(data_videogamesales$Global_Sales, na.rm=TRUE)
```

Screenshot of the output in R Studio:



```
getwd()
data_videogamesales<-read.csv("videogame_salesdata.csv", header=TRUE, sep=",")
mean(data_videogamesales$Global_Sales, na.rm=TRUE)
median(data_videogamesales$Global_Sales, na.rm=TRUE)
max(data_videogamesales$Global_Sales, na.rm=TRUE)
min(data_videogamesales$Global_Sales, na.rm=TRUE)
sd(data_videogamesales$Global_Sales, na.rm=TRUE)
```

```
> getwd()
[1] "C:/Users/Tirthapurohit/Documents/MITALI_R_STUDIO"
> data_videogamesales<-read.csv("videogame_salesdata.csv", header=TRUE, sep=",")
> mean(data_videogamesales$Global_Sales, na.rm=TRUE)
[1] 0.5335427
> median(data_videogamesales$Global_Sales, na.rm=TRUE)
[1] 0.17
> max(data_videogamesales$Global_Sales, na.rm=TRUE)
[1] 82.53
> min(data_videogamesales$Global_Sales, na.rm=TRUE)
[1] 0.01
> sd(data_videogamesales$Global_Sales, na.rm=TRUE)
[1] 1.547935
>
```

Mean: The mean of the global sales is 0.5335427 million. The minimum sales are 0.01 million, and the maximum sales are 82.53 million. The data indicated that the average sales globally are around 0.53 UDS. This means value indicates the average sales of all the countries in the world. This mean value can be compared against an individual country to judge whether the sales are comparable or not. Similarly, their respective means can be used to compare the sales of a smaller group against a more comprehensive group. An example would be comparing the mean of the sales from a small country and a big country.

Standard Deviation: The standard deviation of the variable global sales is 1.547935 million. That means that each country's sales are at an average difference of 1.547935 million from the mean sales of all the countries. This value shows how the data is spread out from the mean. The standard deviation is more than the mean value, indicating that the data points were above the mean. Since the standard deviation is greater than the mean value, the data are more spread out.

Median: The median of the variable global sales is 0.17, and the mean value is 0.5362517. The median provides a helpful measure of the center of a dataset. We can see that the median value is not close to the mean value. Since the median value is not close to the mean, we can conclude that the dataset is not distributed symmetrically. We are comparing the median to the mean; we can say that the data set is not evenly distributed and the data is more spread out from the lowest to highest values

Minimum: We see the minimum global sales is 0.01 million. When we compare the sales of different countries for different publications and different years of release, we see that the minimum sales globally is 0.01 million

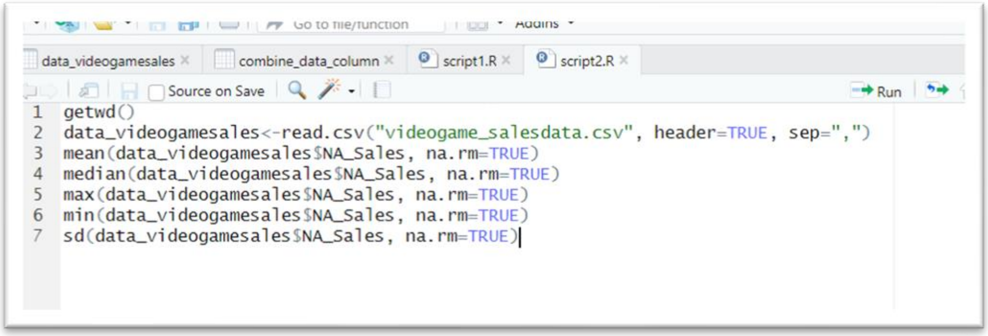
Maximum: We see the maximum global sales is 82.53 million. When we compare the sales of different countries for different publications and different years of release, we see that the maximum sales globally is 82.53 million

3. Statistical analysis of the variable NA_Sales

Code for individual Statistical analysis of the variable NA_Sales in R Studio

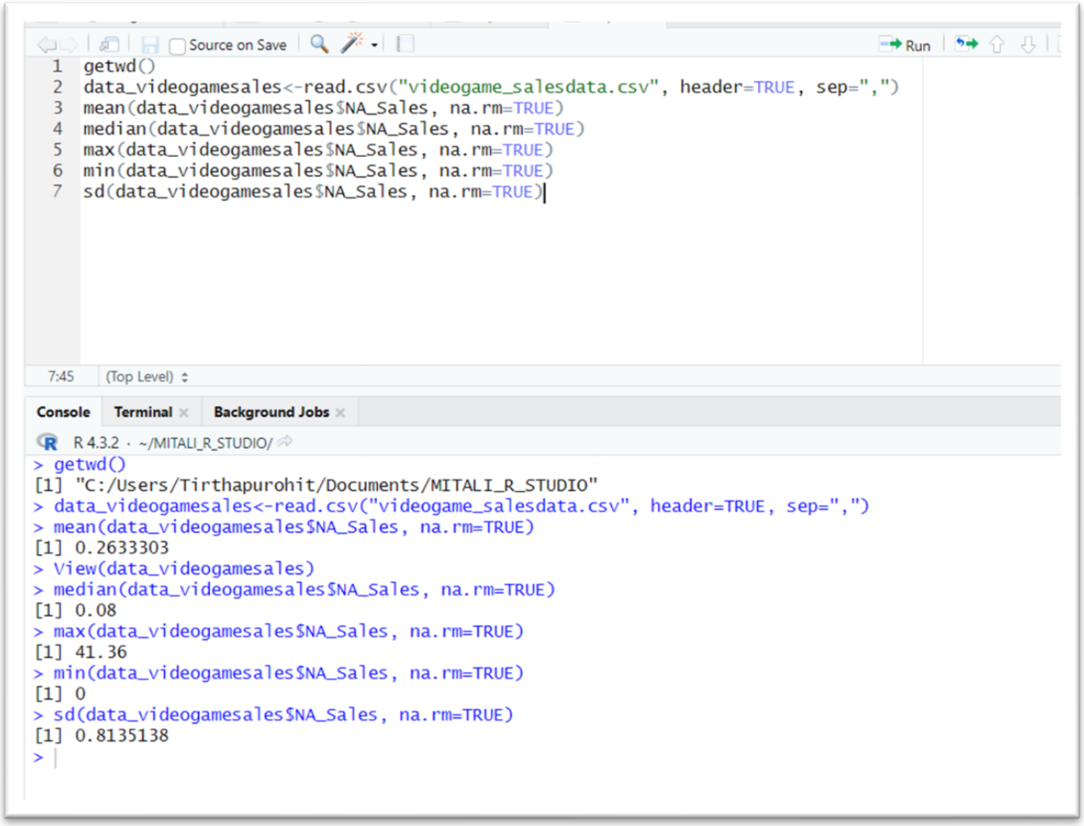
```
getwd()
data_videogamesales<-read.csv("videogame_salesdata.csv", header=TRUE, sep=",")
mean(data_videogamesales$NA_Sales, na.rm=TRUE)
median(data_videogamesales$NA_Sales, na.rm=TRUE)
max(data_videogamesales$NA_Sales, na.rm=TRUE)
min(data_videogamesales$NA_Sales, na.rm=TRUE)
sd(data_videogamesales$NA_Sales, na.rm=TRUE)
```

Screenshot of the code of the variable NA_Sales in R Studio



```
getwd()
data_videogamesales<-read.csv("videogame_salesdata.csv", header=TRUE, sep=",")
mean(data_videogamesales$NA_Sales, na.rm=TRUE)
median(data_videogamesales$NA_Sales, na.rm=TRUE)
max(data_videogamesales$NA_Sales, na.rm=TRUE)
min(data_videogamesales$NA_Sales, na.rm=TRUE)
sd(data_videogamesales$NA_Sales, na.rm=TRUE)
```

Screenshot of the output in R Studio



```
getwd()
data_videogamesales<-read.csv("videogame_salesdata.csv", header=TRUE, sep=",")
mean(data_videogamesales$NA_Sales, na.rm=TRUE)
median(data_videogamesales$NA_Sales, na.rm=TRUE)
max(data_videogamesales$NA_Sales, na.rm=TRUE)
min(data_videogamesales$NA_Sales, na.rm=TRUE)
sd(data_videogamesales$NA_Sales, na.rm=TRUE)
```

```
> getwd()
[1] "C:/Users/Tirthapurohit/Documents/MITALI_R_STUDIO"
> data_videogamesales<-read.csv("videogame_salesdata.csv", header=TRUE, sep=",")
> mean(data_videogamesales$NA_Sales, na.rm=TRUE)
[1] 0.2633303
> View(data_videogamesales)
> median(data_videogamesales$NA_Sales, na.rm=TRUE)
[1] 0.08
> max(data_videogamesales$NA_Sales, na.rm=TRUE)
[1] 41.36
> min(data_videogamesales$NA_Sales, na.rm=TRUE)
[1] 0
> sd(data_videogamesales$NA_Sales, na.rm=TRUE)
[1] 0.8135138
>
```

Mean The average annual sales in each region can be calculated from the output. The result shows that in the North American area, from the year 1985 to 2020, the average sales for all video games, irrespective of genre, developer, and publisher, are **0.2633303** million, which is higher than any individual region and which is almost 50 percent of the global average. By this, we can analyze that the craze for video games is always high in North America compared to the remaining parts of the world. Similarly, we can also see that Europe, Japan, Other regions, and global sales averages.

Median The median is the most central value of each column. If the median is lower than the mean, the mode will also be lower. As a result, we claim that our data is skewed to the right and thus positive. A longer or fatter tail on the right side of the distribution is referred to as a positive skew. When the median exceeds the mean, the data is skewed to the left, indicating that it is negative. Negative skew refers to the left side's longer or broader tail. The Median for North America is 0.08, which is lower than the mean value of 0.266. We can see that the median value is far from the mean value. We are comparing the median to the mean; we can say that the data set is not evenly distributed from the lowest to highest values

Standard Deviation Data are clustered around the mean when the standard deviation is low and spread out when the standard deviation is high. The standard deviation of the variable NA_Sales is 0.8135138. The standard deviation is greater than the mean value, which indicates that the data points were above the mean. Since the standard deviation is greater than the mean value, the data are more spread out from the mean. This value shows how the data is spread out from the mean.

Maximum Here, we see the maximum sales region-wise. In individual regions, North America always stands first in maximum sales with $ 41.36 million, followed by 28.96, 10.22, 10.57 & 82. 53 are sales in millions for Europe, Japan, and other regions and total global sales.

Minimum On the other side, the minimum sales for every region are zero. We see that the minimum sales made in North America for different years and for different publishers are zero.

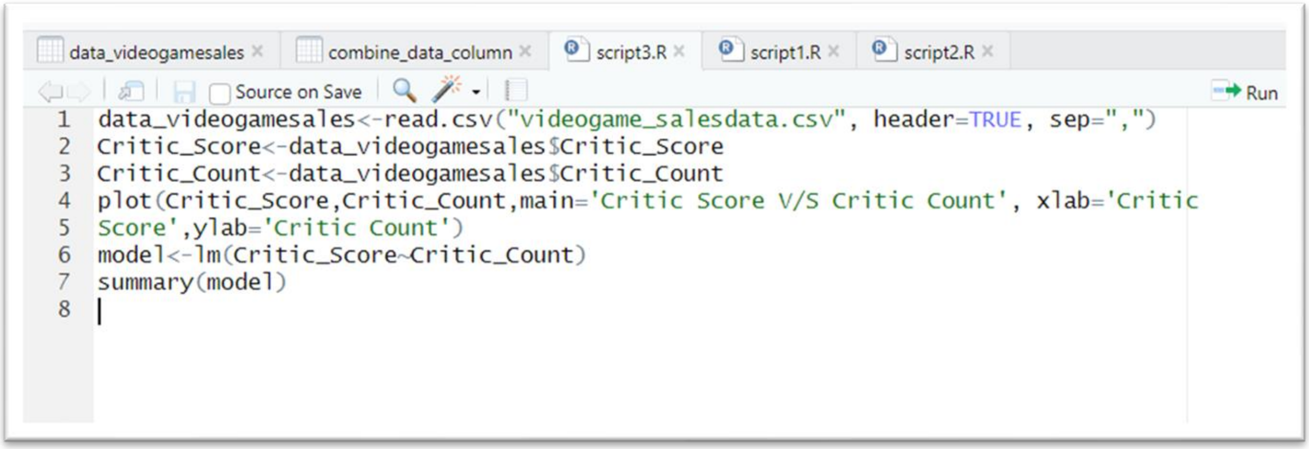
6.0 R Scripts

The correlation between the variables Critic_Score and Critic_Count with the Linear Regression in R Script.

Code for Linear Regression

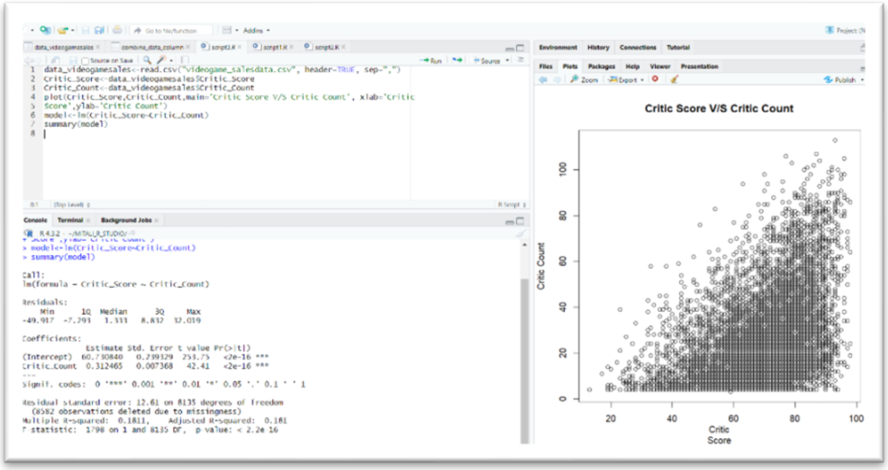
```
data_videogamesales<-read.csv("videogame_salesdata.csv", header=TRUE, sep=",")
Critic_Score<-data_videogamesales$Critic_Score
Critic_Count<-data_videogamesales$Critic_Count
plot(Critic_Score,Critic_Count,main='Critic Score V/S Critic Count', xlab='Critic
Score',ylab='Critic Count')
model<-lm(Critic_Score~Critic_Count)
summary(model)
```

Screenshot of the code in R Studio:



```
data_videogamesales<-read.csv("videogame_salesdata.csv", header=TRUE, sep=",")
Critic_Score<-data_videogamesales$Critic_Score
Critic_Count<-data_videogamesales$Critic_Count
plot(Critic_Score,Critic_Count,main='Critic Score V/S Critic Count', xlab='Critic
Score',ylab='Critic Count')
model<-lm(Critic_Score~Critic_Count)
summary(model)
```

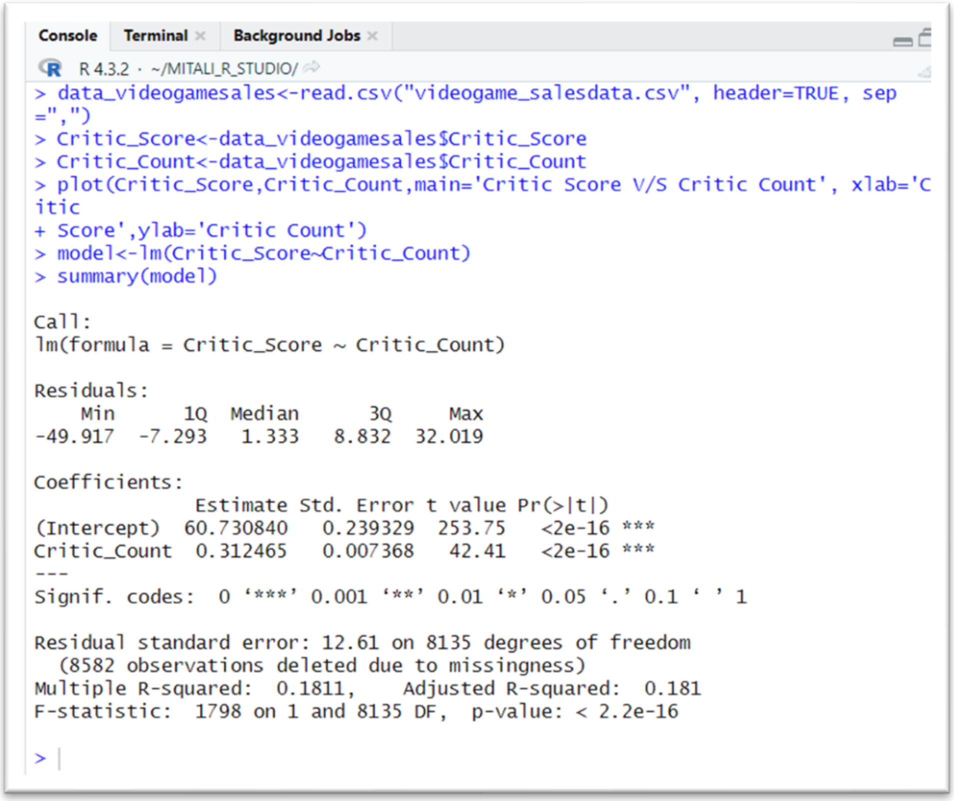
Screenshot of the output in R Studio:



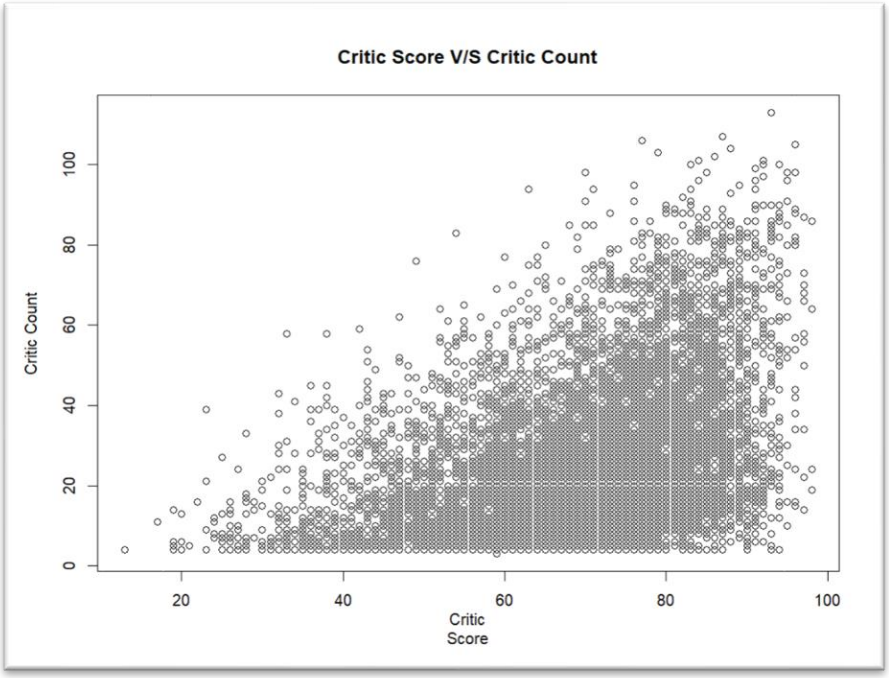
```
data_videogamesales<-read.csv("videogame_salesdata.csv", header=TRUE, sep=",")
Critic_Score<-data_videogamesales$Critic_Score
Critic_Count<-data_videogamesales$Critic_Count
plot(Critic_Score,Critic_Count,main='Critic Score V/S Critic Count', xlab='Critic
Score',ylab='Critic Count')
model<-lm(Critic_Score~Critic_Count)
summary(model)
```

```
+ Score',ylab='Critic Count')
> model<-lm(Critic_Score~Critic_Count)
> summary(model)

Call:
lm(formula = Critic_Score ~ Critic_Count)

Residuals:
    Min      1Q  Median      3Q     Max 
-49.917  -7.293   1.333   8.832  32.019 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  60.730840   0.239329  253.75   <2e-16 ***
Critic_Count  0.312465   0.007368   42.41   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.61 on 8135 degrees of freedom
  (8582 observations deleted due to missingness)
Multiple R-squared:  0.1811,	Adjusted R-squared:  0.181 
F-statistic:  1798 on 1 and 8135 DF,  p-value: < 2.2e-16
```

Screenshot of the Code in R Studio:



```
> data_videogamesales<-read.csv("videogame_salesdata.csv", header=TRUE, sep
=",")
> Critic_Score<-data_videogamesales$Critic_Score
> Critic_Count<-data_videogamesales$Critic_Count
> plot(Critic_Score,Critic_Count,main='Critic Score V/S Critic Count', xlab='C
itic
+ Score',ylab='Critic Count')
> model<-lm(Critic_Score~Critic_Count)
> summary(model)

Call:
lm(formula = Critic_Score ~ Critic_Count)

Residuals:
    Min      1Q  Median      3Q     Max 
-49.917  -7.293   1.333   8.832  32.019 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  60.730840   0.239329  253.75   <2e-16 ***
Critic_Count  0.312465   0.007368   42.41   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.61 on 8135 degrees of freedom
  (8582 observations deleted due to missingness)
Multiple R-squared:  0.1811,	Adjusted R-squared:  0.181 
F-statistic:  1798 on 1 and 8135 DF,  p-value: < 2.2e-16

> 
```



Linear Regression analysis allows you to understand the strength of relationships between variables. Using statistical measurements like R-squared / adjusted R-squared and regression analysis, we can tell you how much of the total variability in the data is explained by your model. Linear regression analysis predicts a variable's value based on another variable's value. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable. In our example, the dependent variable is the critic count, and the independent variable is the critic score

Linear Regression Equation

$\boldsymbol{Y = mx + c}$

In the above equation, m is the slope, c is the constant, and y is the y-intercept. From the linear regression table, we can write the liner regression equation as

$\boldsymbol{Y}$ **= 0.31 $*$ (Critic score) + 60.72**

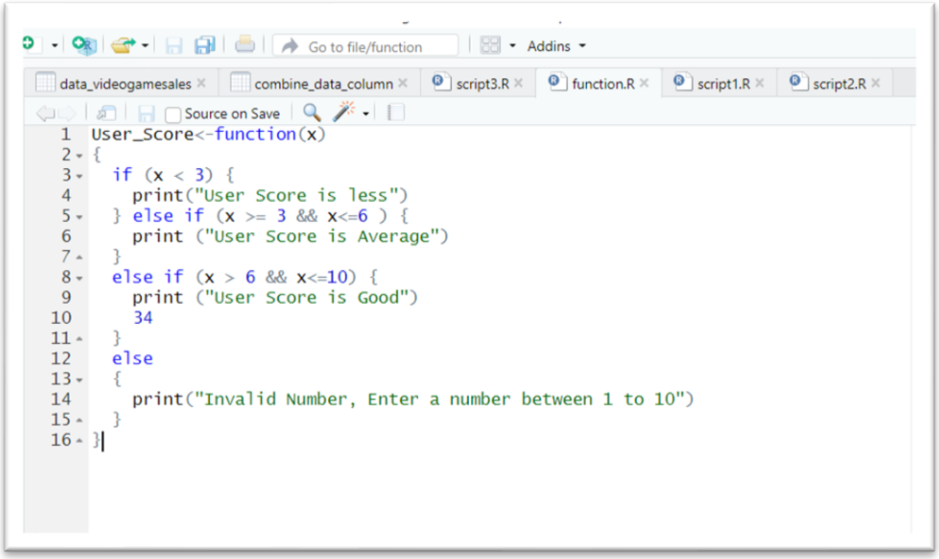
R square value is approximately 0.1823, which is 18.23%. This shows a correlation between critic count and critic score. This indicates very low correlations. There were 16448 observations. 8467 missing values are reported. The r-squared value is 18.23%, which is the correlation between the critic count and the critic score in this example. We can conclude that the critic score explains that the variation in base salary has significantly less impact on this variation.

7.0 R Function

Checking whether the User Score is good or not using R Function

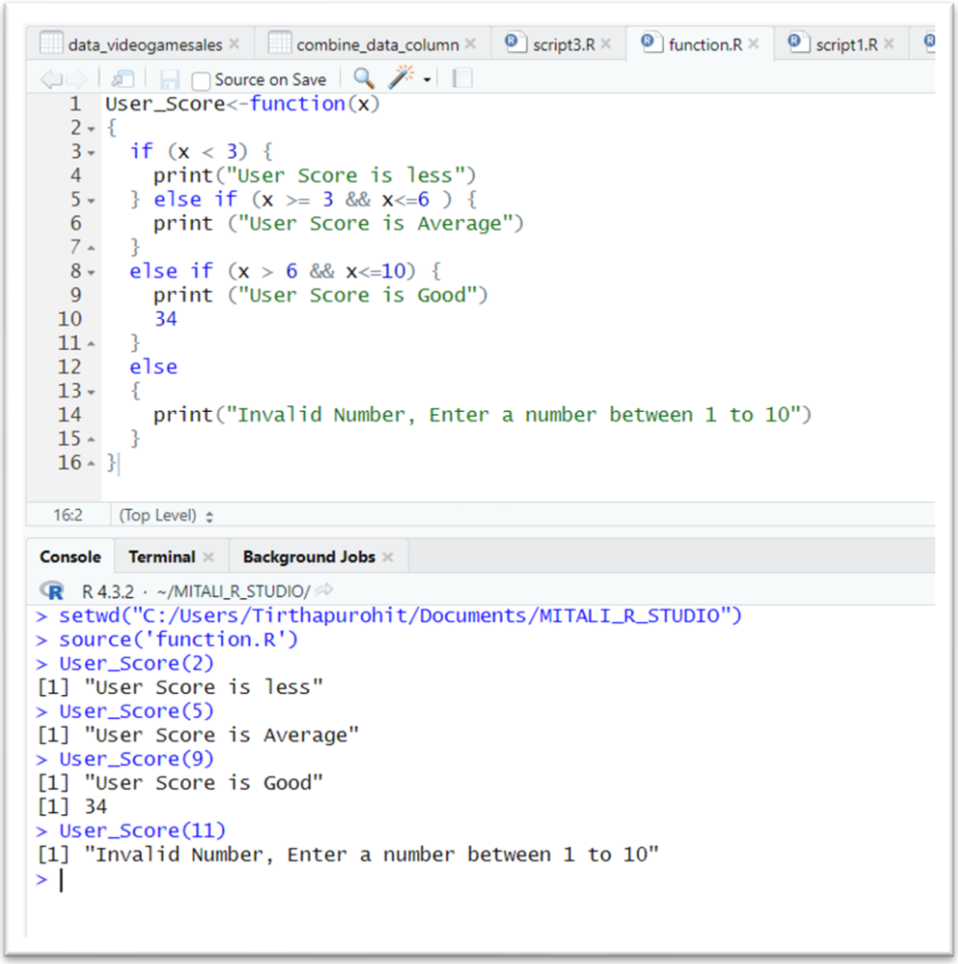
```
User_Score<-function(x)
{
 if (x < 3) {
   print("User Score is less")
 } else if (x >= 3 && x<=6 ) {
   print ("User Score is Average")
 }
 else if (x > 6 && x<=10) {
   print ("User Score is Good")
   34
 }
 else
 {
   print("Invalid Number, Enter a number between 1 to 10")
 }
}
```

Screenshot of Code in R studio:



```
User_Score<-function(x)
{
  if (x < 3) {
    print("User Score is less")
  } else if (x >= 3 && x<=6 ) {
    print ("User Score is Average")
  }
  else if (x > 6 && x<=10) {
    print ("User Score is Good")
    34
  }
  else
  {
    print("Invalid Number, Enter a number between 1 to 10")
  }
}
```

Screenshot of Output in R studio:



```
User_Score<-function(x)
{
  if (x < 3) {
    print("User Score is less")
  } else if (x >= 3 && x<=6 ) {
    print ("User Score is Average")
  }
  else if (x > 6 && x<=10) {
    print ("User Score is Good")
    34
  }
  else
  {
    print("Invalid Number, Enter a number between 1 to 10")
  }
}
```

```
> setwd("C:/Users/Tirthapurohit/Documents/MITALI_R_STUDIO")
> source('function.R')
> User_Score(2)
[1] "User Score is less"
> User_Score(5)
[1] "User Score is Average"
> User_Score(9)
[1] "User Score is Good"
[1] 34
> User_Score(11)
[1] "Invalid Number, Enter a number between 1 to 10"
> |
```

The above function can be used to check if the user scores a game receive is less, average, or good. We are using the function to check the user score. If the score is below 3, then the score is less. If the score is between 3 and 6, then the score is average; if the user score is above 6 and less than 10, then the user score is good. If the input value is anything other than 1 to 10, it shows “Invalid Number; enter a number between 1 to 10”. If and else if function is used to execute the program.

8.0 Conclusion

During our analysis of video game sales and ratings, several key findings have emerged, shedding light on critical aspects of the industry:

- **Sales Trends:** We observed notable trends in global sales, with certain platforms and genres standing out as dominant forces in the market.
- **Impact of Ratings:** The analysis suggests a correlation between high critic scores and user scores, indicating that well-received games tend to perform consistently well across both metrics.
- **Regional Variances:** Differences in sales across North America, Europe, and Japan highlight the importance of tailoring marketing and content strategies to regional preferences.
- **Time Trends:** Examining the year of release data revealed evolving preferences and technological shifts, influencing the success of video games over time.

9.0 Reference

1. Roettl, J., & Terlutter, R. (n.d.). The same video game in 2D, 3D or virtual reality – how does technology impact game evaluation and brand placements? PLOS ONE. Retrieved November 19, 2022, from https://journals.plos.org/plosone/article?id=10.1371%2Fjournal.pone.0200724
2. Comparing video game sales by Gaming Platform - swer.wtamu.edu. (n.d.). Retrieved November 20, 2022, from https://swer.wtamu.edu/sites/default/files/Data/303-1102-1-PB_0.pdf
3. Run R script file with RScript - Example. TutorialKart. (2021, June 21). Retrieved December 16, 2022, from https://www.tutorialkart.com/r-tutorial/r-script-file/
4. Grolemund, H. W. and G. (n.d.). R for data science. 19 Functions. Retrieved December 16, 2022, from https://r4ds.had.co.nz/functions.html
5. About linear regression. IBM. (n.d.). Retrieved November 2, 2022, from https://www.ibm.com/topics/linear-regression