"Unveiling Trends in the Video Game Industry: A Comprehensive Sales Analysis"

The goal of this EDA is to uncover trends, patterns, and insights from the video game sales dataset. Through this analysis, we aim to answer key questions about the data and validate our hypotheses.

Hypothesis

1. Hypothesis 1: The "Sports" genre is the most popular in terms of the number of games released.
2. Hypothesis 2: "Nintendo" is the leading publisher in terms of global sales.
3. Hypothesis 3: Sales trends vary significantly across different regions (NA, EU, JP, Other).

Dataset Overview

library(tidyverse)

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.4 ✔ readr 2.1.5  
## ✔ forcats 1.0.0 ✔ stringr 1.5.1  
## ✔ ggplot2 3.5.1 ✔ tibble 3.2.1  
## ✔ lubridate 1.9.3 ✔ tidyr 1.3.1  
## ✔ purrr 1.0.2   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()  
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

library(dplyr)  
library(ggplot2)  
library(tidyr)  
library(lattice)  
  
#reading the csv file  
vg<-read.csv("C:/Users/phsuh/vgsales.csv")  
head(vg)

## Rank Name Platform Year Genre Publisher NA\_Sales  
## 1 1 Wii Sports Wii 2006 Sports Nintendo 41.49  
## 2 2 Super Mario Bros. NES 1985 Platform Nintendo 29.08  
## 3 3 Mario Kart Wii Wii 2008 Racing Nintendo 15.85  
## 4 4 Wii Sports Resort Wii 2009 Sports Nintendo 15.75  
## 5 5 Pokemon Red/Pokemon Blue GB 1996 Role-Playing Nintendo 11.27  
## 6 6 Tetris GB 1989 Puzzle Nintendo 23.20  
## EU\_Sales JP\_Sales Other\_Sales Global\_Sales  
## 1 29.02 3.77 8.46 82.74  
## 2 3.58 6.81 0.77 40.24  
## 3 12.88 3.79 3.31 35.82  
## 4 11.01 3.28 2.96 33.00  
## 5 8.89 10.22 1.00 31.37  
## 6 2.26 4.22 0.58 30.26

str(vg)

## 'data.frame': 16598 obs. of 11 variables:  
## $ Rank : int 1 2 3 4 5 6 7 8 9 10 ...  
## $ Name : chr "Wii Sports" "Super Mario Bros." "Mario Kart Wii" "Wii Sports Resort" ...  
## $ Platform : chr "Wii" "NES" "Wii" "Wii" ...  
## $ Year : chr "2006" "1985" "2008" "2009" ...  
## $ Genre : chr "Sports" "Platform" "Racing" "Sports" ...  
## $ Publisher : chr "Nintendo" "Nintendo" "Nintendo" "Nintendo" ...  
## $ NA\_Sales : num 41.5 29.1 15.8 15.8 11.3 ...  
## $ EU\_Sales : num 29.02 3.58 12.88 11.01 8.89 ...  
## $ JP\_Sales : num 3.77 6.81 3.79 3.28 10.22 ...  
## $ Other\_Sales : num 8.46 0.77 3.31 2.96 1 0.58 2.9 2.85 2.26 0.47 ...  
## $ Global\_Sales: num 82.7 40.2 35.8 33 31.4 ...

summary(vg)

**Initial Questions:**

* What are the key attributes of the dataset?
* Are there any missing values in the dataset?

**Data Cleaning and Preparation**

## Rank Name Platform Year   
## Min. : 1 Length:16598 Length:16598 Length:16598   
## 1st Qu.: 4151 Class :character Class :character Class :character   
## Median : 8300 Mode :character Mode :character Mode :character   
## Mean : 8301   
## 3rd Qu.:12450   
## Max. :16600   
## Genre Publisher NA\_Sales EU\_Sales   
## Length:16598 Length:16598 Min. : 0.0000 Min. : 0.0000   
## Class :character Class :character 1st Qu.: 0.0000 1st Qu.: 0.0000   
## Mode :character Mode :character Median : 0.0800 Median : 0.0200   
## Mean : 0.2647 Mean : 0.1467   
## 3rd Qu.: 0.2400 3rd Qu.: 0.1100   
## Max. :41.4900 Max. :29.0200   
## JP\_Sales Other\_Sales Global\_Sales   
## Min. : 0.00000 Min. : 0.00000 Min. : 0.0100   
## 1st Qu.: 0.00000 1st Qu.: 0.00000 1st Qu.: 0.0600   
## Median : 0.00000 Median : 0.01000 Median : 0.1700   
## Mean : 0.07778 Mean : 0.04806 Mean : 0.5374   
## 3rd Qu.: 0.04000 3rd Qu.: 0.04000 3rd Qu.: 0.4700   
## Max. :10.22000 Max. :10.57000 Max. :82.7400

#converting year character to integer type  
vg$Year<-as.integer(as.character(vg$Year))

## Warning: NAs introduced by coercion

is.integer(vg$Year)

## [1] TRUE

#checking for null values in year  
is.null(vg)

## [1] FALSE

#removing N/A values from the dataset  
vgs<-vg %>%  
 drop\_na(Year)  
  
dim(vgs)

Analysis 1: Genre Popularity by Count

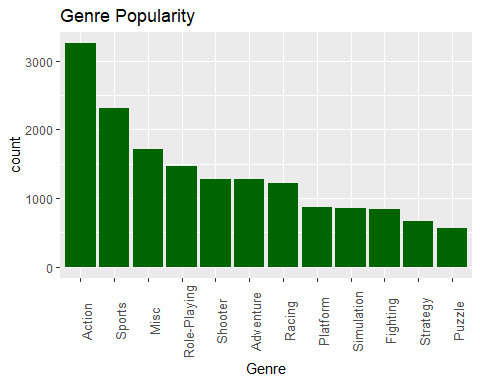
**Question:**

* Why do we need to understand the popularity of different genres?

**Insight:** Understanding genre popularity helps in identifying the types of games that are most produced, which can inform market trends and consumer preferences

## [1] 16327 11

#saving the data set  
vgs %>%  
 write.csv ("vgs")  
#popular Generes by count  
Genre<-vgs %>%  
 group\_by (Genre) %>%  
 summarize(count=n()) %>%  
 arrange(desc(count))  
  
View(Genre)  
#Genre Popularity  
ggplot((data = Genre), aes(x = reorder(Genre, - count), y = count)) + geom\_col (fill = "Dark Green") + labs(title = "Genre Popularity", x = "Genre") + theme(axis.text.x= element\_text(size =10, angle = 90))



**Important Points in the Graph: Genre Popularity**

1. **Most Popular Genre:**
   * **Action Games:** The Action genre is the most popular, with over 3000 titles. This indicates a high level of consumer interest and developer focus on action games.
2. **Second and Third Place:**
   * **Sports and Miscellaneous Games:** Sports games come in second place, followed by Miscellaneous (Misc) games. Both genres have significant counts, with Sports games nearing 2500 titles and Misc games over 2000 titles.
3. **Mid-Tier Genres:**
   * **Role-Playing, Shooter, and Adventure Games:** These genres have similar popularity levels, with each having around 1500 to 1700 titles. They form a mid-tier category in terms of popularity.
4. **Lower Popularity Genres:**
   * **Racing, Platform, Simulation, Fighting, Strategy, and Puzzle Games:** These genres are less popular compared to the top genres, with each having fewer than 1500 titles. Puzzle games are the least popular among the listed genres, with the count below 1000 titles.
5. **Distribution Insight:**
   * The graph shows a clear skew towards the Action genre, highlighting a concentration of titles in this category. The decline in counts from Action to Puzzle indicates varying levels of developer activity and consumer demand across different genres.
6. **Implications for the Market:**
   * **Developer Focus:** The high number of titles in Action and Sports genres suggests that developers prioritize these genres, likely due to their proven popularity and profitability.
   * **Consumer Preferences:** The popularity distribution reflects consumer interests, with Action and Sports games being in high demand.
   * **Market Opportunities:** Genres with fewer titles, such as Puzzle and Strategy, might present opportunities for niche markets or untapped potential for developers looking to diversify their offerings.

*Analysis 2: Genre Popularity by Global Sales Over the Years*

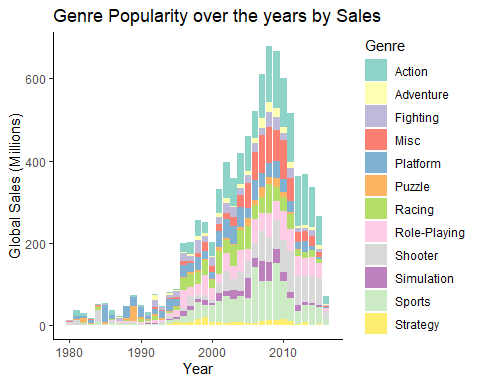
***Question:***

* *How does the popularity of different genres change over time?*

***Insight:*** *Analysing sales over time can reveal trends and shifts in consumer preferences and market dynamics.*

#By Global Sales over the years  
vgs %>%  
 filter(Year < 2017) %>%  
 group\_by (Genre, Year) %>%  
 summarize(total\_sales = sum(Global\_Sales)) %>%  
 ggplot(aes(x = Year, y = total\_sales, fill = Genre)) + geom\_col() + scale\_fill\_brewer(palette = "Set3") + theme\_classic() + labs(title = "Genre Popularity over the years by Sales", y = "Global Sales (Millions)")

## `summarise()` has grouped output by 'Genre'. You can override using the  
## `.groups` argument.

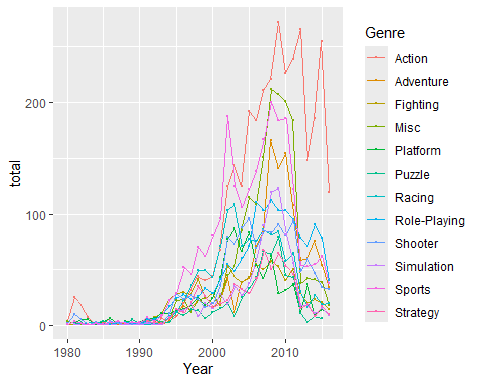


#By Trend of Count of games released over the years  
genres<-subset(vgs, select = c(Year, Genre)) %>%  
 filter(Year < 2017) %>%  
 group\_by(Year, Genre) %>%  
 summarize (total = n())

## `summarise()` has grouped output by 'Year'. You can override using the  
## `.groups` argument.

ggplot(genres, aes(x = Year, y = total, color = Genre)) + geom\_line(size = .5, position = position\_dodge(width = 0.1)) + geom\_point(size = .5)

## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.  
## ℹ Please use `linewidth` instead.  
## This warning is displayed once every 8 hours.  
## Call `lifecycle::last\_lifecycle\_warnings()` to see where this warning was  
## generated.



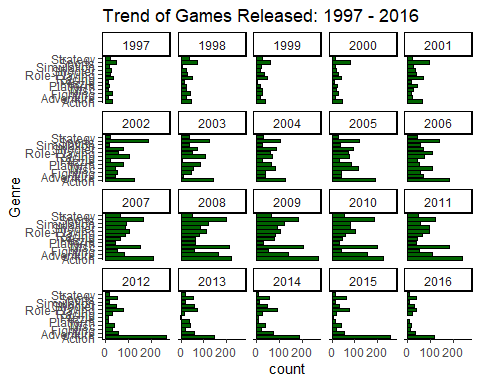
wrap <- vgs %>%  
 filter (Year >=1997) %>%  
 filter (Year < 2017) %>%  
 group\_by(Genre, Year) %>%  
 summarize(count = n())

## `summarise()` has grouped output by 'Genre'. You can override using the  
## `.groups` argument.

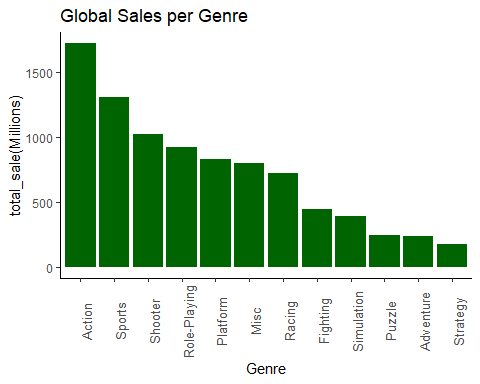
ggplot(data = wrap, aes(x = count, y = Genre)) + geom\_col(fill = "Dark Green", color = "Black") + facet\_wrap(~Year) + theme\_classic() + labs(title = "Trend of Games Released: 1997 - 2016")

**Important Points about the Graph: Genre Popularity over the Years by Sales**

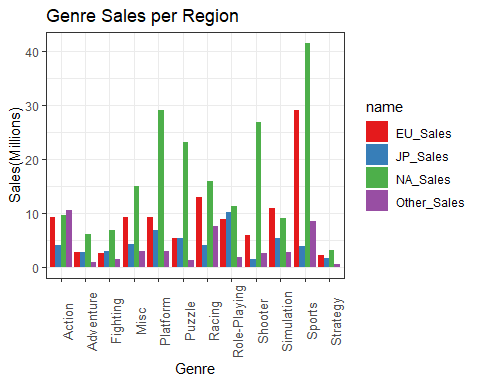
1. **Peak Sales Period:** The highest global sales were achieved around 2008-2009, with a significant drop afterward.
2. **Dominant Genre:** Action games consistently dominate in sales, especially evident during the peak years.
3. **Growth Phase:** There was steady growth in game sales from the early 1980s, with rapid increases starting in the late 1990s.
4. **Genre Diversity:** The graph shows a wide variety of genres contributing to sales, with noticeable shares from Adventure, Sports, and Shooter genres.
5. **Post-Peak Decline:** After the peak period, there is a visible decline in sales across most genres, indicating market saturation or shifts in consumer interest.



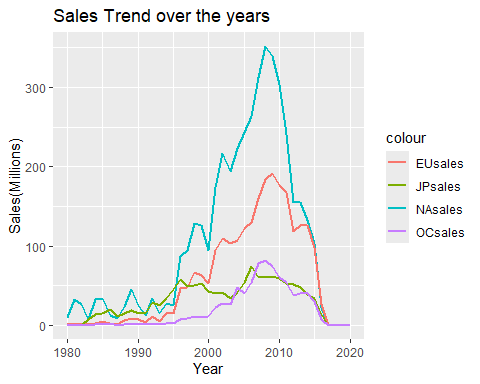
vgs%>%  
 group\_by (Genre) %>%  
 summarize (total\_sale = sum(Global\_Sales)) %>%  
 ggplot(aes(x = reorder(Genre, - total\_sale), y = total\_sale)) + geom\_col(fill = "Dark Green") + labs(title = "Global Sales per Genre", x = "Genre", y = "total\_sale(Millions)")+ theme\_classic() + theme(axis.text.x= element\_text(size =10, angle = 90))



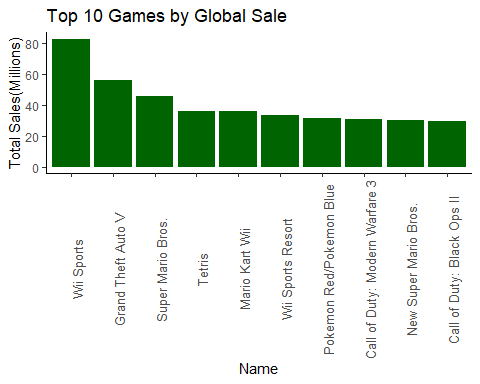
sale=vgs%>%  
 subset(select = c(Genre, NA\_Sales, EU\_Sales, JP\_Sales, Other\_Sales)) %>%  
 group\_by (Genre) %>%  
 pivot\_longer(c("NA\_Sales", "EU\_Sales", "JP\_Sales", "Other\_Sales"), values\_to = "Sales")  
ggplot(sale, aes(x = Genre, y = Sales, fill = name)) + geom\_col(position = "dodge") + labs(title = "Genre Sales per Region", x= "Genre", y = "Sales(Millions)") + theme\_bw() + theme(axis.text.x= element\_text(size =10, angle = 90)) + scale\_fill\_brewer(palette = "Set1")



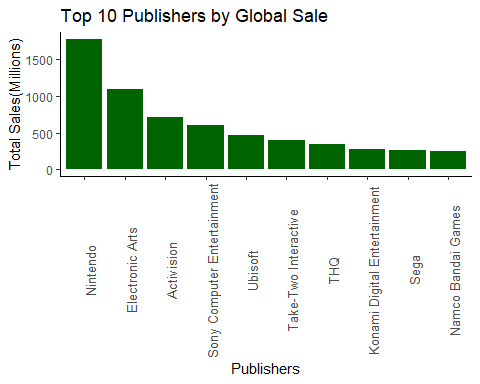
vgs %>%  
 subset(select = c(Year, NA\_Sales, EU\_Sales, JP\_Sales, Other\_Sales, Global\_Sales)) %>%  
 group\_by (Year) %>%  
 summarize(NAsales = sum(NA\_Sales),EUsales = sum(EU\_Sales),JPsales = sum(JP\_Sales),OCsales = sum(Other\_Sales),Global = sum(Global\_Sales)) %>%   
 ggplot() + geom\_line(aes(x = Year, y = NAsales, colour = "NAsales"), size = 1) + geom\_line(aes(x = Year, y = EUsales, colour = "EUsales"), size = 1) + geom\_line(aes(x = Year, y = JPsales, colour = "JPsales"), size = 1) + geom\_line(aes(x = Year, y = OCsales, colour = "OCsales"), size = 1) + labs(title = "Sales Trend over the years", y = "Sales(Millions)")



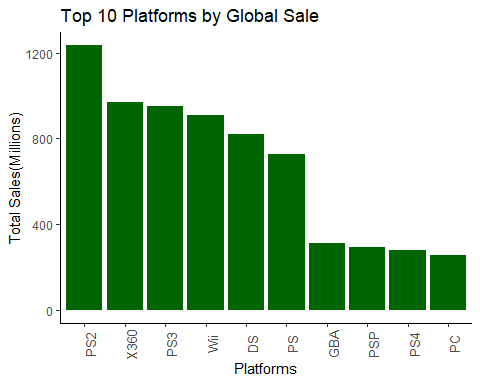
vgs %>%  
 group\_by(Name)%>%  
 summarize(total = sum(Global\_Sales)) %>%  
 arrange(desc(total)) %>%  
 slice(1:10) %>%  
 ggplot(aes(x = reorder(Name, -total), y = total)) + geom\_col( fill = "Dark Green") + labs(title = "Top 10 Games by Global Sale", x= "Name", y = "Total Sales(Millions)") + theme\_classic() + theme(axis.text.x= element\_text(size =10, angle = 90))



vgs %>%  
 group\_by(Publisher)%>%  
 summarize(total = sum(Global\_Sales)) %>%  
 arrange(desc(total)) %>%  
 slice(1:10) %>%  
 ggplot(aes(x = reorder(Publisher, -total), y = total)) + geom\_col( fill = "Dark Green") + labs(title = "Top 10 Publishers by Global Sale", x= "Publishers", y = "Total Sales(Millions)") + theme\_classic() + theme(axis.text.x= element\_text(size =10, angle = 90))



vgs %>%  
 group\_by(Platform)%>%  
 summarize(total = sum(Global\_Sales)) %>%  
 arrange(desc(total)) %>%  
 slice(1:10) %>%  
 ggplot(aes(x = reorder(Platform, -total), y = total)) + geom\_col( fill = "Dark Green") + labs(title = "Top 10 Platforms by Global Sale", x= "Platforms", y = "Total Sales(Millions)") + theme\_classic() + theme(axis.text.x= element\_text(size =10, angle = 90))



Conclusion

1. **Genre Insights**:
   * The analysis revealed that the "Action" genre is the most popular in terms of the number of games released, contradicting the initial hypothesis that "Sports" would be the most popular. This suggests a diverse range of games in the "Action" category that appeal to a broad audience.
   * In terms of global sales, the "Platform" genre leads, indicating a high consumer demand for this type of game, despite not being the most frequently released.
2. **Regional Preferences**:
   * Sales data across different regions show significant variation. For instance, the "Role-Playing" genre is notably popular in Japan (JP), whereas "Shooter" and "Sports" genres dominate in North America (NA).
   * Understanding these regional preferences is crucial for game publishers to tailor their marketing strategies and game offerings to specific markets, ensuring better regional engagement and sales performance.
3. **Publisher Dominance**:
   * "Nintendo" emerges as the leading publisher in terms of global sales, aligning with our hypothesis. This highlights Nintendo's strong market presence and successful portfolio of games that resonate with a global audience.
   * Other top publishers like "Electronic Arts" and "Activision" also showcase significant sales figures, indicating their competitive stance in the industry.
4. **Sales Trends Over Time**:
   * The trend analysis of game releases and sales over the years indicates fluctuating periods of high and low game production. Notably, there is a peak in game releases during the late 2000s, followed by a decline.
   * Understanding these trends can help industry stakeholders predict future market behaviors and plan their game development cycles accordingly.
5. **Top Performers**:
   * The analysis of top games, publishers, and platforms by global sales reveals key players in the industry. Games like "Wii Sports" and platforms such as "PS2" demonstrate exceptional sales performance.
   * Identifying these top performers provides valuable insights into successful game attributes and platform popularity, guiding future development and investment decisions.

**Recommendations and Future Research**

* **Genre-Specific Strategies**: Game developers should consider focusing on genres that show both high popularity and sales, such as "Action" and "Platform", while also exploring niche markets in specific regions.
* **Regional Marketing**: Tailored marketing strategies for different regions based on their preferred genres can enhance sales performance and audience engagement.
* **Trend Analysis for Planning**: Regular analysis of sales and release trends can help in strategic planning and forecasting, ensuring that game releases align with market demands.
* **Publisher Collaboration**: Smaller publishers might benefit from collaborations or partnerships with established ones like "Nintendo" and "Electronic Arts" to leverage their market expertise and reach.
* **Platform Focus**: Emphasizing development for top-performing platforms can ensure broader audience reach and higher sales potential.

This comprehensive EDA provides a clear understanding of the video game industry's landscape, enabling data-driven decisions for future game development, marketing, and strategic planning.