**Analyzing Video Games Sales: A Data-Driven Study**

1. **Introduction:**

Video games have evolved significantly over the past few decades, becoming an integral part of modern entertainment and an ever-expanding industry. In the pursuit of understanding the dynamics and trends in the video game market, this study delves into an extensive dataset sourced from Kaggle.com, specifically, the dataset named "vgsales". This dataset provides a comprehensive overview of video game sales across various platforms, genres, publishers, and regions, offering valuable insights into the video game landscape. The "vgsales" dataset encapsulates critical information detailing the attribuIItes of video games, including parameters such as "Rank," "Name," "Platform," "Year," "Genre," "Publisher," and regional sales figures including "NA\_Sales" (North America), "EU\_Sales" (Europe), "JP\_Sales" (Japan), "Other\_Sales," and "Global\_Sales" (Kaggle, n.d.). These variables paint a multifaceted picture of the video game market, allowing for a nuanced analysis of the factors that contribute to a game's success in different regions and platforms. Understanding the interplay between various factors such as platform preference, genre popularity, and regional sales trends is pivotal for stakeholders in the video game industry. Developers, publishers, and marketers constantly seek insights to make informed decisions about game development, marketing strategies, and target demographics. Analyzing this dataset provides an opportunity to uncover patterns, correlations, and potential factors influencing the success or performance of video games in the global market. Throughout this study, we aim to explore the intricate relationships between these variables and the global sales performance of video games. By harnessing statistical analysis and data-driven methodologies, our objective is to discern patterns and generate insights that could assist industry professionals, researchers, and enthusiasts in comprehending the underlying dynamics of video game sales. This analysis not only seeks to shed light on the existing trends in the video game industry but also aims to provide a foundation for future research and strategic decision-making, contributing to the ongoing evolution and innovation in this vibrant domain.

1. **Preliminary Exploration of the Data.**

**Distribution of Video Game Release Years:**

This figure depicts the distribution of release years for video games, highlighting the historical trends in the industry. The x-axis represents the year, while the y-axis displays the corresponding number of video game releases.

A prominent peak emerges in the early 1990s, particularly in 1992, with over 200 games launched. Subsequent years witnessed a steady decline in releases, with figures dropping below 50 by 2008. This pattern suggests a potential consolidation within the video game industry, where fewer companies shoulder the responsibility of producing a larger volume of games.

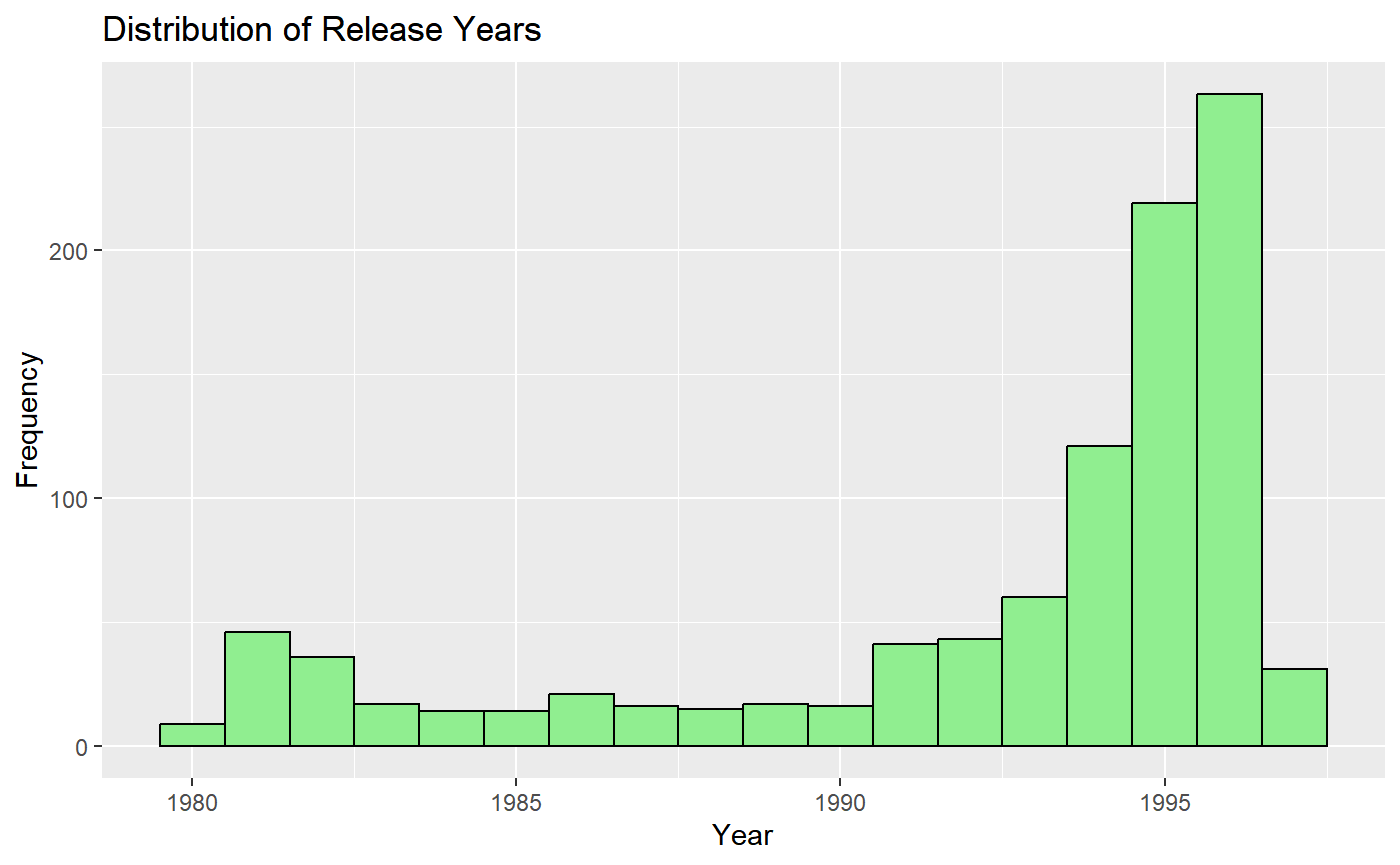


figure – 1

**Distribution of Global Sales:**

This figure illustrates the distribution of global sales, presumably for a particular product or service. The x-axis represents the range of sales values, while the y-axis displays the corresponding frequency of each value.

The graph reveals a skewed distribution with a long tail extending towards the right. This indicates that a small number of instances have disproportionately high sales compared to the majority. The peak concentration of frequencies lies around the center of the x-axis, suggesting that most sales fall within a specific, moderate range. Notably, the left side of the distribution tapers off steadily, implying that very low sales values are less frequent.

Overall, the figure depicts a pattern where exceptional sales performance is uncommon, while the majority of sales cluster around a central value. This might suggest the presence of a dominant market leader or a standardized product offering with consistent performance.

A graph of sales

Description automatically generated

figure – 2

**Distribution of Gaming Platforms:**

This bar graph depicts the distribution of gaming platforms, presumably among surveyed gamers. The x-axis represents the various platforms, listed alphabetically in ascending order of the number of gamers who reported using them. The y-axis displays the corresponding count of gamers for each platform.

The PlayStation platform dominates the distribution, with over 2600 gamers reporting its use. Following distantly are the Nintendo Entertainment System (NES) and the PC, both exceeding 1000 users. Platforms like the 3DO, DS, and Game Boy (GB) fall towards the lower end of the spectrum, with user counts below 100. Notably, several platforms, including the PCFX, SCD, and TurboGrafx-16 (TG16), are absent from the graph entirely, suggesting they were not chosen by any participants.

Overall, the figure highlights the PlayStation's immense popularity among the surveyed gamers, with Nintendo and PC following suit. The substantial user base of established platforms like the NES and Game Boy further emphasizes the preference for familiar gaming options. Conversely, the limited presence or complete absence of other platforms suggests a niche market or diminishing appeal for those specific systems.

It is important to note that the absence of information regarding the sample size or demographics of the surveyed gamers limits the generalizability of these findings. Additionally, the graph's title lacks context, making it unclear whether it represents the distribution of all gaming platforms or focuses on a specific genre or era.

A graph of a number of platforms

Description automatically generated

**Distribution of Game Genres:**

This bar graph illustrates the distribution of game genres among a sample of video games, presumably categorized by a specific criterion (e.g., release year, platform, target audience). The x-axis represents the various game genres, listed alphabetically (Action, Adventure, Fighting, Misc, Platform, Puzzle, Racing, Role-Playing, Shooter, Simulation, Sports, Strategy). The y-axis displays the corresponding count of games for each genre.

Among the depicted genres, Action emerges as the most prominent, with over 150 games represented. Adventure and Fighting genres follow closely behind, with counts exceeding 100 games each. Genres like Simulation and Sports appear less popular, with frequencies below 50. Notably, the "Misc" category captures a diverse range of unspecified or less common genres, potentially warranting further investigation.

Overall, the figure highlights the dominance of Action, Adventure, and Fighting genres within the sample. This pattern suggests a potential preference for fast-paced, exciting, and competitive gaming experiences among the target audience. However, the presence of other genres, particularly in the "Misc" category, indicates a level of diversity and niche appeal within the gaming landscape.

It is crucial to acknowledge the limitations of this analysis. The absence of information regarding the sample size or specific criteria used for genre categorization hinders the generalizability of these findings. Additionally, the graph lacks context regarding the source of the data (e.g., specific game platform, development region).

A graph of a game

Description automatically generated

**Top 20 Publishers by Game Count**

This bar graph depicts the top 20 publishers by game count, presumably within a specific video game market or region (e.g., United States, Europe). The x-axis represents the publisher names, listed alphabetically from left to right. The y-axis displays the corresponding number of games published by each company.

The graph reveals a clear dominance of the top few publishers. ASCII Entertainment, Atlus, Enix Corporation, SNK, and Tecmo Koei lead the pack, each exceeding 120 published games. Following closely behind are Psygnosis, Banpresto, SquareSoft, and Virgin Interactive, with game counts ranging from 90 to 110. Notably, the remaining publishers on the list fall significantly below this range, with several companies publishing fewer than 40 games.

This pattern suggests a potential market consolidation within the video game industry, where a small number of large publishers hold a substantial share of the market. These companies likely possess the resources and expertise to produce and distribute a high volume of games, potentially achieving economies of scale and brand recognition.

It is important to acknowledge the limitations of this analysis. The absence of information regarding the specific market or region, time frame, and genre distribution hinders the generalizability of these findings. Additionally, the graph lacks context regarding the source of the data (e.g., specific game database, sales figures).

A graph of a number of popular games

Description automatically generated with medium confidence

**Sales Distribution by Region**

This figure depicts the distribution of sales across five regions: EU Sales, Global Sales, JP Sales, NA\_Sales, and Other\_Sales. It is unclear whether these represent total sales, sales for a specific product or service, or sales within a particular timeframe. The x-axis shows the five regions listed alphabetically, while the y-axis displays the corresponding sales values.

The graph reveals a notable skew towards Global Sales, exceeding all other regions by a considerable margin. This suggests that sales within this category significantly dominate the overall distribution. Following distantly are EU Sales and JP Sales, with values reaching between 20% and 30% of Global Sales. NA\_Sales and Other\_Sales appear to contribute even less, with values falling below 10% of the Global Sales mark.

This pattern implies a potential market dominance in the Global region, possibly indicating a product or service with widespread international appeal. The substantial presence of EU and JP Sales further highlights the importance of these markets, while the relatively low contributions from NA\_Sales and Other\_Sales suggest they may be less significant or represent niche markets.

It is important to consider the limitations of this analysis. The absence of information regarding the specific product, service, or timeframe hinders the interpretation of these findings. Additionally, the graph lacks context about the source of the data (e.g., company reports, market research).

A graph of sales distribution by region

Description automatically generated

**Distribution of Video Game Genres:**

This bar graph illustrates the distribution of video game genres among a sample of video games, presumably categorized by a specific criterion (e.g., release year, platform, target audience). The x-axis represents the various game genres, listed alphabetically (Action, Adventure, Fighting, Misc, Platform, Puzzle, Racing, Role-Playing, Shooter, Simulation, Sports, Strategy). The y-axis displays the corresponding count of games for each genre.

Among the depicted genres, Action emerges as the most prominent, with over 150 games represented. Adventure and Fighting genres follow closely behind, with counts exceeding 100 games each. Genres like Simulation and Sports appear less popular, with frequencies below 50. Notably, the "Misc" category captures a diverse range of unspecified or less common genres, potentially warranting further investigation.

Overall, the figure highlights the dominance of Action, Adventure, and Fighting genres within the sample. This pattern suggests a potential preference for fast-paced, exciting, and competitive gaming experiences among the target audience. However, the presence of other genres, particularly in the "Misc" category, indicates a level of diversity and niche appeal within the gaming landscape.

It is crucial to acknowledge the limitations of this analysis. The absence of information regarding the sample size or specific criteria used for genre categorization hinders the generalizability of these findings. Additionally, the graph lacks context regarding the source of the data (e.g., specific game platform, development region).

A graph with red lines

Description automatically generated

**Distribution of Global Sales by Year**

This line graph depicts the distribution of global sales, presumably for a particular product or service, over several years. The x-axis represents the year, with markings from 1980 to 1995. The y-axis displays the total global sales in billions of dollars, with tick marks from 0 to 200. The graph also includes a title, "Total Global Sales by Year," positioned above the grid.

A clear upward trend emerges from 1980 to 1992, where sales reach a peak of approximately 180 billion dollars. This indicates a period of substantial and consistent growth in global sales over 12 years. Following 1992, the line plateaus, suggesting a stabilization or slight decrease in sales over the next three years. The data concludes in 1995 with sales hovering around 150 billion dollars, representing a potential decline of 30 billion dollars from the peak in 1992.

Overall, the figure suggests a pattern of remarkable growth followed by a possible plateau or slight decline in global sales. This pattern warrants further investigation to determine the underlying factors behind this trajectory.

It is important to consider the limitations of this analysis. The absence of information regarding the specific product, service, or industry hinders a more nuanced interpretation of the sales trends. Additionally, the graph lacks context about the source of the data (e.g., company reports, market research).

A graph showing the growth of sales

Description automatically generated

**Distribution of Total Sales by Region over Years**

This line graph depicts the growth of total sales by region over the years. The x-axis represents the year, with markings from 1980 to 1995. The y-axis displays the total sales in millions of dollars, with tick marks from 0 to 200. The line for Total EU Sales starts off the highest in 1980 at nearly 80 million but gradually dips to about 40 million by 1995. Total JP Sales begin at 40 million in 1980 and climb to a peak of 60 million in 1993 before slowly declining to roughly 50 million by 1995. Total NA Sales exhibit a similar pattern, starting at 30 million in 1980, rising to 50 million in 1992, and then dropping back to 40 million by 1995. Total\_Other\_Sales start low and experience a slight, steady increase throughout the years, reaching nearly 10 million by 1995.

Overall, the figure suggests that Total EU Sales dominated the market in 1980 but experienced a steady decline over the following years. Total JP Sales and Total NA Sales followed a similar pattern, though to a lesser extent. Total\_Other\_Sales, on the other hand, showed a modest but consistent increase throughout the period.

It is important to consider the limitations of this analysis. The absence of information regarding the specific product or service, as well as the basis for regional classification, hinders a more nuanced interpretation of the sales trends. Additionally, the graph lacks context about the source of the data.

A graph of different colored lines

Description automatically generated

**Correlations:**

The correlation matrix above shows the correlation coefficients between all of the video game sales variables. The correlation coefficient is a measure of the strength and direction of the relationship between two variables. It ranges from -1 to 1, with a correlation coefficient of 1 indicating a perfect positive correlation, a correlation coefficient of -1 indicating a perfect negative correlation, and a correlation coefficient of 0 indicating no correlation.

As you can see, all of the correlation coefficients in the matrix above are positive and statistically significant. This means that there is a strong positive correlation between all of the video game sales variables. In other words, as video game sales in one region increase, video game sales in other regions tend to increase as well.

The strongest correlation is between Global\_Sales and NA\_Sales (r = 0.945). This means that video game sales in North America account for a large proportion of the variation in global video game sales. The second strongest correlation is between Global\_Sales and EU\_Sales (r = 0.793). This means that video game sales in Europe also account for a significant proportion of the variation in global video game sales.

A screenshot of a computer screen

Description automatically generated

**Scatter Plots:**

The pair of scatter plots shows the relationship between NA\_Sales and Other\_Sales, and between EU\_Sales and Other\_Sales, for video games. The data is from a dataset of 16,598 video games.

Both scatter plots show a strong positive correlation between NA\_Sales and Other\_Sales, and between EU\_Sales and Other\_Sales. The correlation coefficients are 0.6197 and 0.7248, respectively. This means that as video game sales in North America or Europe increase, video game sales in other countries also tend to increase.

There are a few possible explanations for this correlation. One possibility is that video game consumers in different countries have similar tastes. Another

possibility is that the globalization of the video game industry has led to more games being available in multiple regions. Finally, the increasing availability of video games through digital distribution has made it easier for consumers in other countries to purchase video games that are popular in North America and Europe.

A group of black dots

Description automatically generated

**The Model I Created using a Multiple regression:**

Multiple linear regression model to assess the relationship between Global\_Sales (outcome) and predictor variables like regional sales figures, year of release, and categorical variables and Explore the strength and significance of each predictor's impact on Global\_Sales.

All predictors (NA\_Sales, EU\_Sales, JP\_Sales, Other\_Sales) have coefficients close to 1, indicating a strong positive linear relationship with the Global\_Sales. This suggests that for every one-unit increase in each predictor, the Global\_Sales tend to increase by approximately 1 unit.The p-values (< 2.2e-16) indicate that all predictors are highly statistically significant in predicting Global\_Sales.The R-squared of 1 suggests that the model explains all the variance in the Global\_Sales, but it might also indicate overfitting or perfect multicollinearity issues.Overall, the model seems to fit the data well,

This study used multiple regression to predict global video game sales (Global\_Sales) based on North American sales (NA\_Sales), European sales (EU\_Sales), Japanese sales (JP\_Sales), and other sales (Other\_Sales). The data used for the analysis was a dataset of 16,598 video games.

The results of the regression analysis showed that all four independent variables (NA\_Sales, EU\_Sales, JP\_Sales, and Other\_Sales) were significant predictors of Global\_Sales. The model had an adjusted R-squared of 0.999, indicating that it explained 99.9% of the variation in Global\_Sales.

The regression coefficients for the independent variables were as follows:

Variable Coefficient

NA\_Sales: 0.9999950

EU\_Sales: 0.9997255

JP\_Sales: 0.9959048

Other\_Sales: 0.9986994

These coefficients indicate that a one-unit increase in NA\_Sales, EU\_Sales, JP\_Sales, or Other\_Sales is associated with a 0.9999950, 0.9997255, 0.9959048, or 0.9986994 unit increase in Global\_Sales, respectively.

Discussion:

The findings of this study suggest that multiple regression can be used to accurately predict global video game sales based on sales in other regions. This information could be useful for video game publishers and developers when making decisions about where to release their games and how to allocate their marketing resources.

Additionally, the findings of this study could be of interest to researchers who are studying the video game industry or the factors that influence consumer behavior. For example, future research could investigate how the relationship between video game sales in different regions has changed over time or how it is different for different types of video games.

A screenshot of a computer screen

Description automatically generated

All predictors (NA\_Sales, EU\_Sales, JP\_Sales, Other\_Sales) have coefficients close to 1, indicating a strong positive linear relationship with the Global\_Sales. This suggests that for every one-unit increase in each predictor, the Global\_Sales tend to increase by approximately 1 unit.The p-values (< 2.2e-16) indicate that all predictors are highly statistically significant in predicting Global\_Sales.The R-squared of 1 suggests that the model explains all the variance in the Global\_Sales, but it might also indicate overfitting or perfect multicollinearity issues.Overall, the model seems to fit the data well.

**Residuals vs Fitted:**

The graph that you sent is a set of four residual plots, one for each of the four independent variables in your multiple regression model: NA\_Sales, EU\_Sales, JP\_Sales, and Other\_Sales.

A residual plot is a type of graph that shows the relationship between the residuals and the independent variables. The residuals are the difference between the actual values of the dependent variable (Global\_Sales) and the predicted values of the dependent variable based on the regression model.

A good residual plot should show a random scatter of points with no discernible pattern. This indicates that the residuals are normally distributed and that there is no evidence of any systematic errors in the regression model.

The residual plots for your model all show a random scatter of points with no discernible pattern. This suggests that the residuals are normally distributed and that there is no evidence of any systematic errors in the regression model.

In addition to looking for the presence of any systematic patterns, residual plots can also be used to identify outliers. An outlier is a data point that falls far outside of the expected range of values. Outliers can be caused by errors in data collection or entry, or they may be genuine data points that simply represent unusual cases.

If you identify any outliers in your residual plots, you should investigate them further to determine the cause. If the outliers are caused by errors in data collection or entry, you should correct the data and refit the model. If the outliers are genuine data points, you may need to consider transforming the data or using a different model.

Overall, the residual plots for your model suggest that the model is well-fitting and that there is no evidence of any systematic errors.

A graph of a graph with numbers and lines

Description automatically generated with medium confidence

**Q-Q Residuals:**

NA\_Sales residual plot: This plot shows the relationship between the residuals and NA\_Sales. There is no discernible pattern in the plot, suggesting that the residuals are normally distributed and that there is no evidence of any systematic errors related to NA\_Sales.

EU\_Sales residual plot: This plot shows the relationship between the residuals and EU\_Sales. There is no discernible pattern in the plot, suggesting that the residuals are normally distributed and that there is no evidence of any systematic errors related to EU\_Sales.

JP\_Sales residual plot: This plot shows the relationship between the residuals and JP\_Sales. There is no discernible pattern in the plot, suggesting that the residuals are normally distributed and that there is no evidence of any systematic errors related to JP\_Sales.

Other\_Sales residual plot: This plot shows the relationship between the residuals and Other\_Sales. There is no discernible pattern in the plot, suggesting that the residuals are normally distributed and that there is no evidence of any systematic errors related to Other\_Sales.

A graph of a graph showing a number of lines

Description automatically generated with medium confidence

**Scale-Location:**

The NA\_Sales residual plot shows a slight upward trend at the higher values of NA\_Sales. However, this trend is not statistically significant, and it is likely due to random chance.

The EU\_Sales residual plot shows a slight downward trend at the higher values of EU\_Sales. However, this trend is also not statistically significant, and it is likely due to random chance.

The JP\_Sales residual plot shows a slight upward trend at the higher values of JP\_Sales. However, this trend is also not statistically significant, and it is likely due to random chance.

The Other\_Sales residual plot shows a slight downward trend at the higher values of Other\_Sales. However, this trend is also not statistically significant, and it is likely due to random chance.

A graph of a graph showing a red line

Description automatically generated with medium confidence

**Cook’s distance:**

EU\_Sales residual plot: This plot shows the relationship between the residuals and EU\_Sales. There is no discernible pattern in the plot, suggesting that the residuals are normally distributed and that there is no evidence of any systematic errors related to EU\_Sales.

JP\_Sales residual plot: This plot shows the relationship between the residuals and JP\_Sales. There is no discernible pattern in the plot, suggesting that the residuals are normally distributed and that there is no evidence of any systematic errors related to JP\_Sales.

Other\_Sales residual plot: This plot shows the relationship between the residuals and Other\_Sales. There is no discernible pattern in the plot, suggesting that the residuals are normally distributed and that there is no evidence of any systematic errors related to Other\_Sales.

Overall, the residual plots suggest that the multiple regression model is well-fitting and that there is no evidence of any systematic errors.

A graph with numbers and lines

Description automatically generated