

Deliverable 1: Descriptive Statistics

What Makes a Video Game Successful?

Ioannidis Efstratios (tem2505)

Dataset: Video Game Sales with Ratings 2016

1 Purpose

The goal of this project is to investigate which factors are most strongly associated with the commercial success of video games. Success is measured using global sales figures. Using a Video Game dataset Shams that includes information such as genre, platform, publisher, ESRB rating, release year, and review metrics, this deliverable focuses on descriptive statistics and exploratory data analysis to identify patterns and variables that may influence global sales.

2 Variables

2.1 Dependent Variable (Y)

- Global Sales

2.2 Independent Variables (X)

- Platform
- Year of Release
- Genre
- Publisher
- NA Sales, EU Sales, JP Sales, Other Sales (if needed for advanced analysis)
- Critic Score
- Critic Count
- User Score
- User Count
- Developer
- Rating (ESRB content rating)

Comment: In my variables I have also the variable "Name". "Name" is just the title of the game ,not useful for analysis.

3 Descriptive Statistics

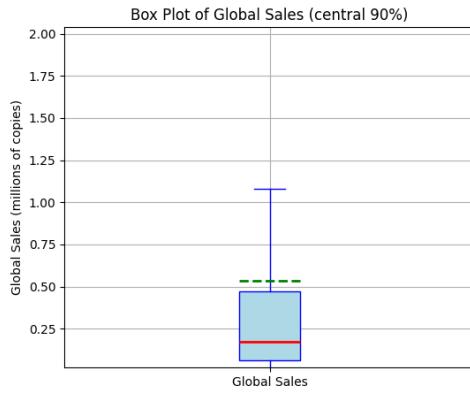
The dataset includes both numerical variables (e.g., Global Sales, Critic Score, User Count) and categorical variables (e.g., Genre, Platform, Publisher). For this deliverable, descriptive statistics such as mean, standard deviation, minimum, and maximum will be calculated only for the numerical variables. Although User Score represents a rating, it is stored as text because of non-numeric entries like "tbd"(to be determined), which prevent it from being recognized as purely numerical.

Variable	Mean	Std Dev	Min	Max
Year_of_Release	2006.487356	5.878995	1980.00	2020.00
NA_Sales	0.263330	0.813514	0.00	41.36
EU_Sales	0.145025	0.503283	0.00	28.96
JP_Sales	0.077602	0.308818	0.00	10.22
Other_Sales	0.047332	0.186710	0.00	10.57
Global_Sales	0.533543	1.547935	0.01	82.53
Critic_Score	68.967679	13.938165	13.00	98.00
Critic_Count	26.360821	18.980495	3.00	113.00
User_Count	162.229908	561.282326	4.00	10665.00

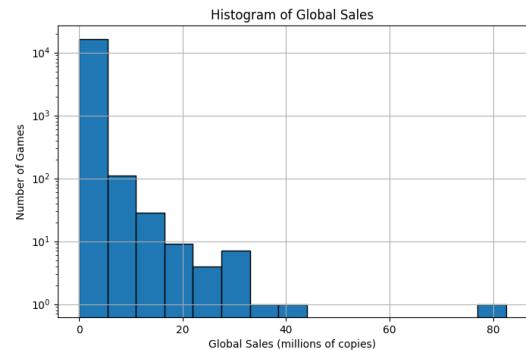
4 Plots

4.1 Plot for each variable

Global Sales



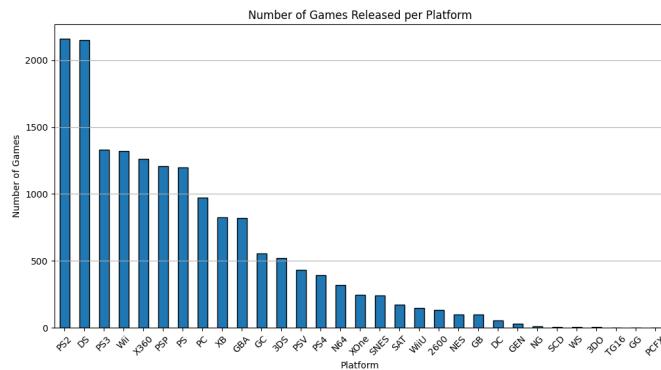
(a) Box plot of "Global Sales"



(b) Histogram of Global Sales

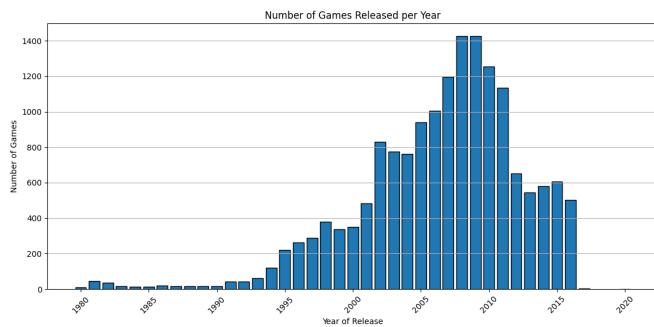
The box plot in panel (a) shows that most games sell relatively few copies (the middle 50 percent lie between about 0.05 M and 0.5 M, median = 0.17 M), with a small number of outliers reaching tens of millions. Likewise, the histogram in panel (b) (with a log-scaled y-axis) reveals a heavily right-skewed distribution: over ten thousand titles sell under 1 M copies, while only a handful break into the multi-million-copy range..

Platform



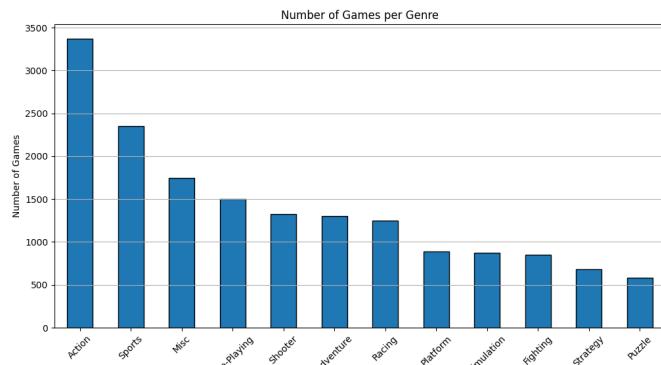
The bar plot shows that certain platforms, such as the DS, PS2, and PS3, had the highest number of game releases, while others had significantly fewer titles.

Year of Release



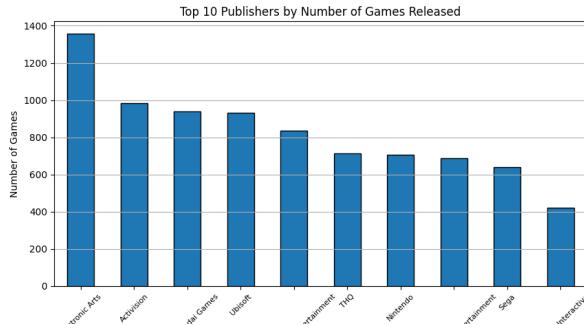
The bar plot shows that video game releases increased steadily through the 1990s and peaked between 2007 and 2009, followed by a gradual decline in subsequent years.

Genre



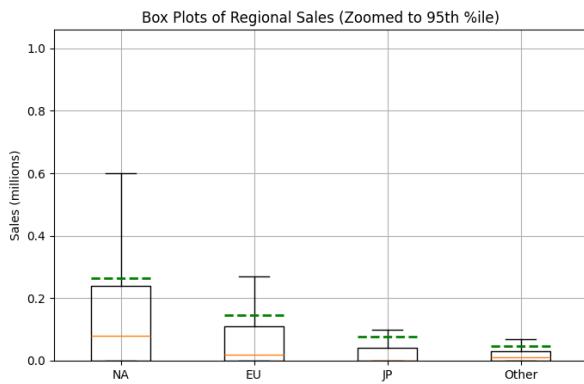
The bar plot shows that Action and Sports are the most common genres among video games, while Puzzle and Strategy genres have fewer releases.

Publisher



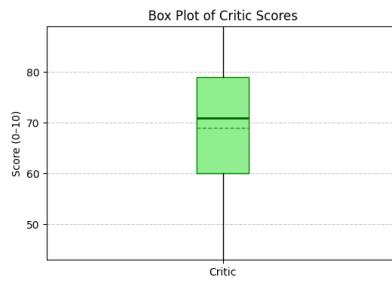
The bar plot shows that major companies like Electronic Arts, Activision, and Namco Bandai published the highest number of games in the dataset.

NA Sales, EU Sales, JP Sales, Other Sales

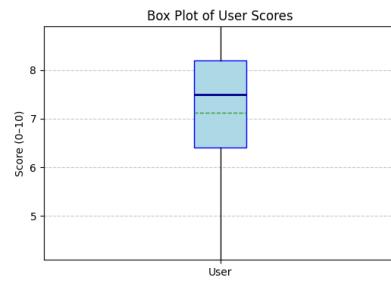


The box plots show that North America and Europe have the highest median game sales, while Japan and Other regions generally have lower sales with a few notable outliers.

Critic and User scores



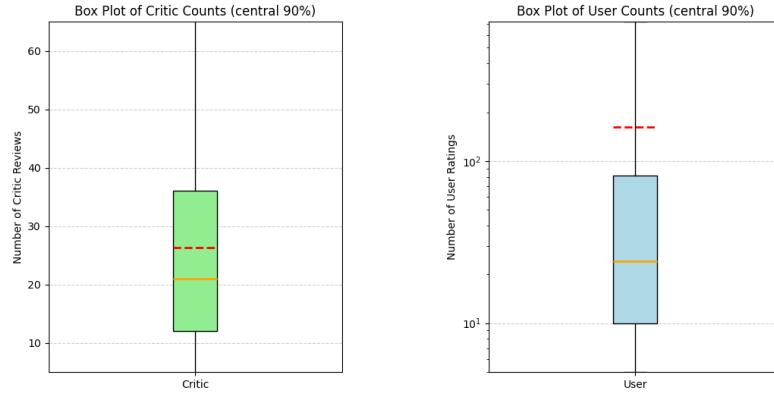
(c) Box plot of "Critic Scores"



(d) Box plot of "User Scores"

Critic scores panel (c) cluster around a median of 71 (IQR 60–79) with a few below 50 or above 90, whereas user scores panel (d) center near 7.5 (IQR 6.3–8.2) with far fewer extremes—showing critics vary more widely while users' ratings are tighter at the top.

Critic and User count

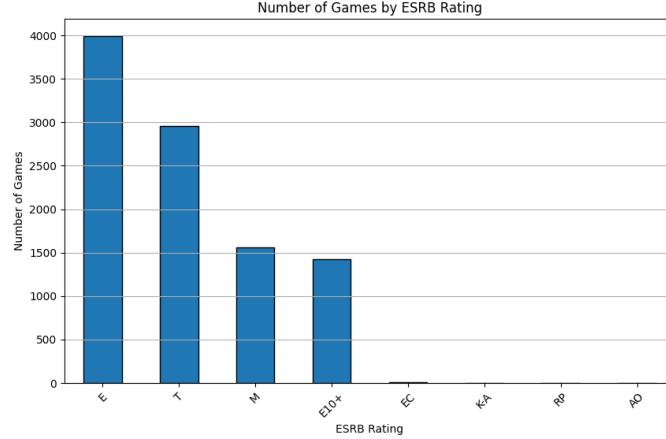


(e) Box plot of "Critic Counts"

(f) Box plot of "User Counts"

Critic count (number of critic reviews per game) in panel (e) is tightly clustered—most games receive about 12–36 reviews (median 26) with very few extreme values—whereas user count (number of user ratings per game) in panel (f) (log-scaled) spans two orders of magnitude: most titles fall between 10 and 90 ratings (median 50), but a small number rack up hundreds, producing a pronounced right tail.

Rating (ESRB content rating)



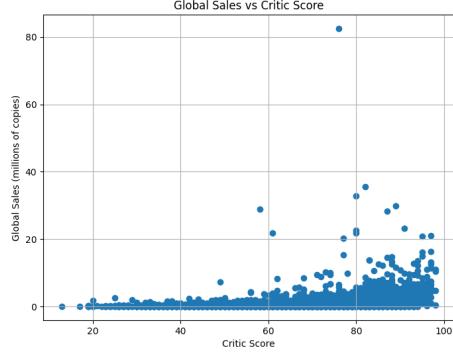
The bar plot shows that most video games are rated E (Everyone) or T (Teen), while fewer titles are rated M (Mature) or E10+. Very few games are rated EC (Early Childhood), KA (old rating system), RP (Rating Pending), or AO (Adults Only).

Developer

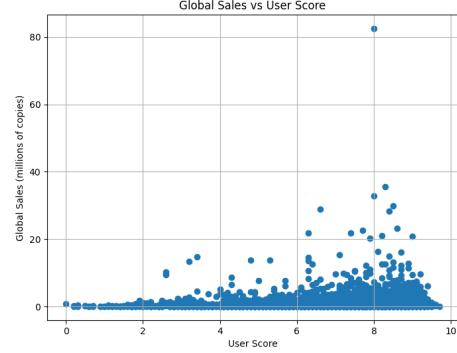
A plot for Developer was not included because there are too many unique developers in the dataset, which would result in a crowded and unreadable graph.

4.2 Plots between independent and dependent variables

Global sales, Critic scores and User scores



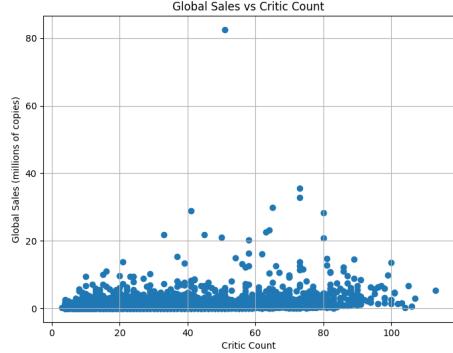
(g) Scatter plot of "Critic Scores"



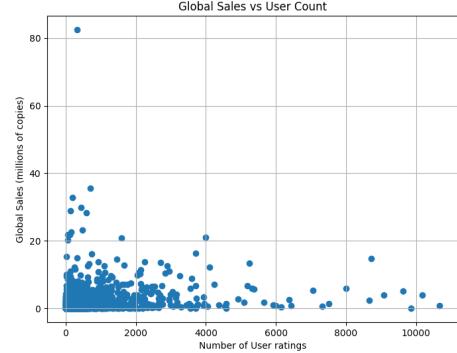
(h) Scatter plot of "User Scores"

Panels (g) and (h) reveal very similar patterns: most games with low-to-moderate scores sell only a few million copies, while a handful of high-scoring titles (especially those above 75 from critics or 8 from users) break into the tens of millions. Critic scores (g) show a slightly tighter cluster in the 60–90 range, whereas user scores (h) spread more evenly from 4–9, but in both cases only the top-rated games achieve blockbuster sales.

Global sales, Critic counts and User counts



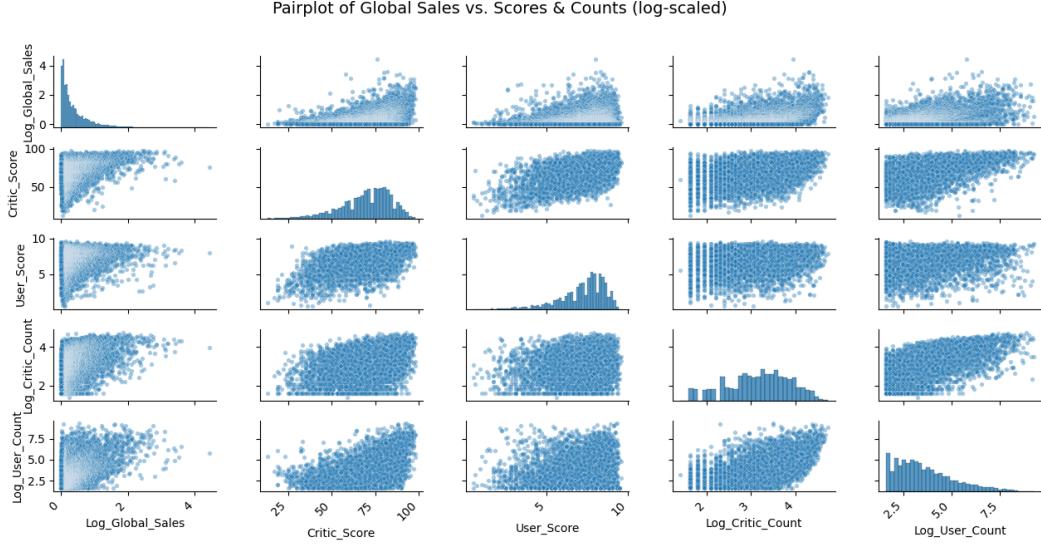
(i) Scatter plot of "Critic Scores"



(j) Scatter plot of "User Scores"

Panel (i) shows that critic review counts are fairly evenly spread from a few reviews up to around 100, with games receiving 10–30 reviews covering the full range of sales from near zero to 80 million copies. In contrast, panel (j) reveals that most user ratings cluster at low counts—fewer than 2 000 votes—even though those few titles that do rack up thousands of ratings can still achieve blockbuster sales. Both plots suggest more reviews tend to accompany higher sales, but critic coverage is much more uniformly distributed than player-vote counts.

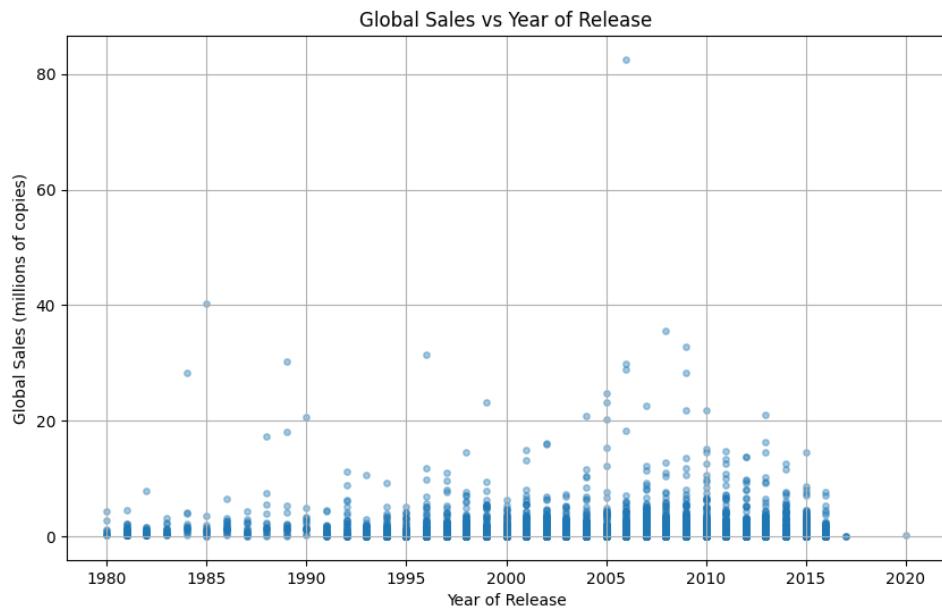
4.2.1 Global Sales , Scores and Counts



The pairplot provides a unified view of how global sales, review scores, and review counts relate to one another, and how each variable is distributed individually. Along the diagonal, we see that *Log Global Sales* follows an exponential-like decay—indicating a small number of blockbuster titles and many low-selling games. *Critic Score* is roughly normally distributed around 75–80, while *User Score* is left-skewed, with most games rated between 6 and 9. While *Log Critic Count* shows a uniform like distribution and *Log User Count* show heavy right tails, suggesting that only a few games receive extensive attention from users while critics are more uniformly distributed.

The off-diagonal scatterplots reveal that higher global sales are generally associated with better critic scores and more reviews, and that critic and user scores themselves are moderately correlated. In summary, blockbuster games are rare but consistently earn higher ratings and attract more reviews, while mid- and low-tier titles tend to cluster in the lower left of each subplot.

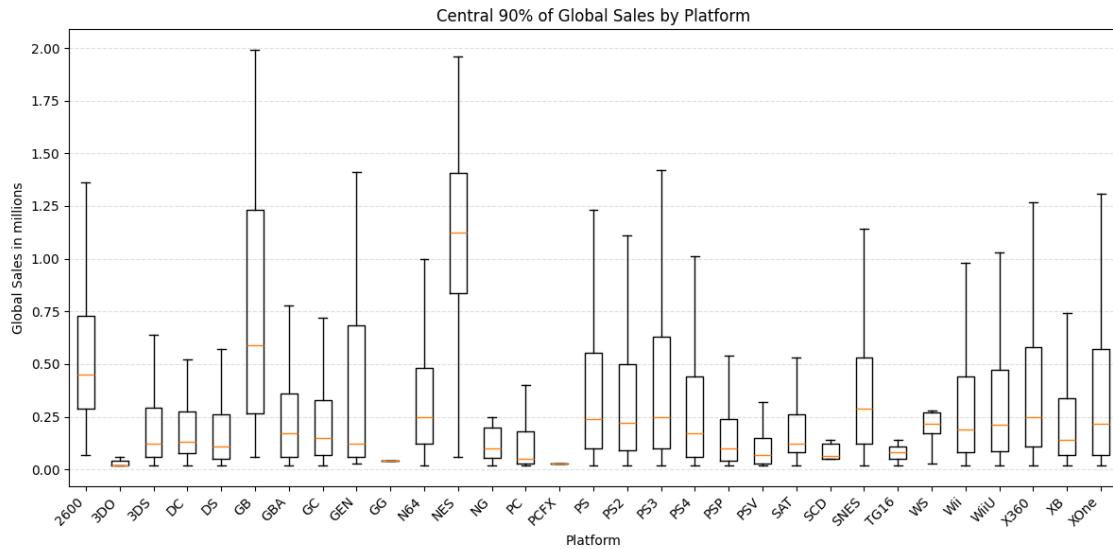
Global sales and Year of release



The scatter plot shows the distribution of global video game sales by year of release:

- **Most games each year** sell fewer than 5 million copies.
- **Top-selling games** in the 1980s and '90s remained under 5 million copies.
- **From the mid-2000s onward**, several titles began to exceed 10 million in sales.
- **An exceptional outlier** around 2006 reached nearly 80 million copies.

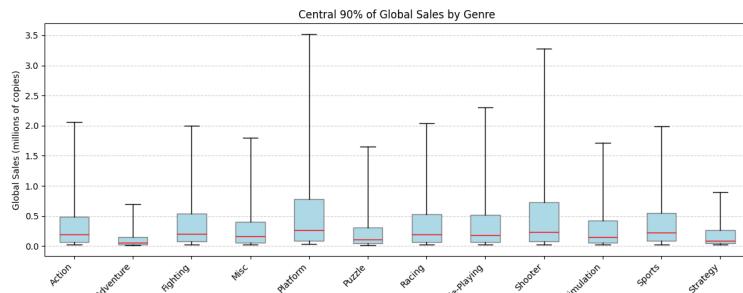
Global sales and Platform



The box-and-whisker chart of raw Global Sales by platform (central 90%) reveals a clear hierarchy:

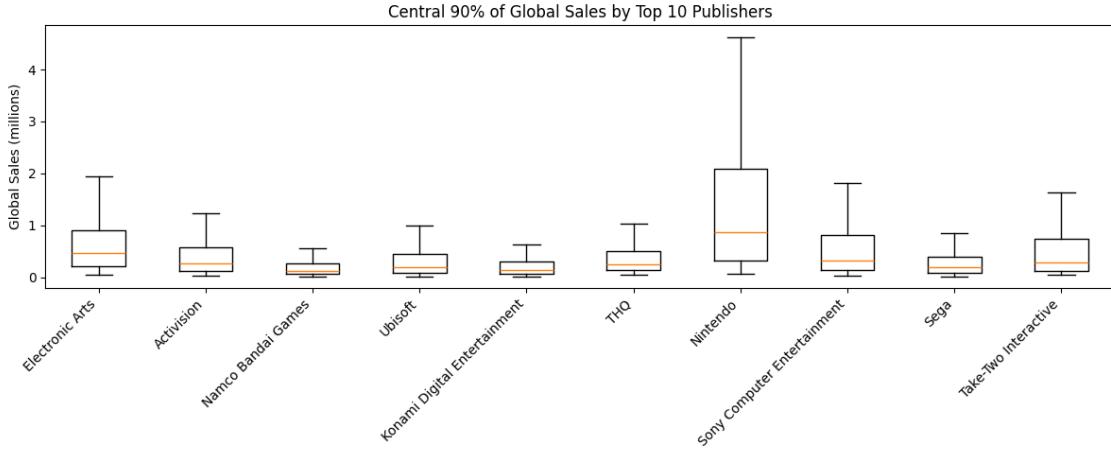
- **Classic home consoles** like NES, Game Boy (GB), and Genesis (GEN) sit at the top, with medians around 0.3–0.6 million copies and very wide IQRs, reflecting both consistent sellers and big hits.
- **Modern consoles** (e.g., PS4, Xbox 360, Xbox One) cluster in the 0.2–0.5 million range, showing moderate spread.
- **Handheld platforms** (DS, GBA, PSP, 3DS) generally lie lower—medians around 0.1–0.3 million.
- **PC and niche systems** (PC-FX, 3DO) have medians nearly at zero, indicating most titles sold under 100K copies.
- **Whiskers across all platforms** reveal that outliers exist everywhere—"blockbuster" titles reaching several million copies.

Global sales and Genre



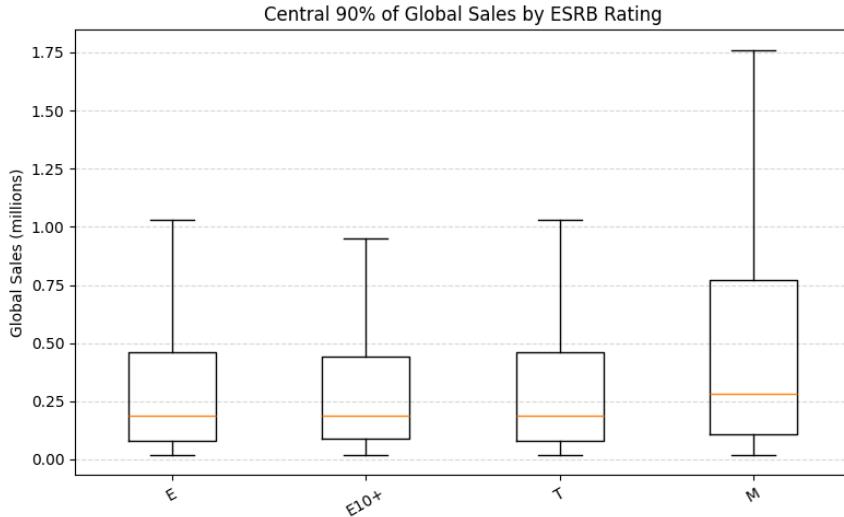
The box plot presents the central 90% of global sales across genres. Shooter and Platform titles lead in sales, with median values around 0.3 million copies and the widest spreads. Adventure, Puzzle, and Strategy genres lag behind, with medians below 0.1 million. The remaining genres—such as Action, Racing, Sports, and Role-Playing—occupy the middle range, typically showing median sales around 0.1–0.2 million copies.

Global sales and Publisher



Nintendo outperforms all others (median 0.9 million; central 90% up to 2 million), with Sony ranking next (median 0.3 million). EA, Activision, Ubisoft, Sega, Konami, THQ, and Take-Two cluster lower, showing medians between approximately 0.1 and 0.4 million and generally tighter interquartile ranges. Overall, platform holders (Nintendo and Sony) dominate in higher sales, while third-party publishers tend toward more modest but consistent performance.

Global sales and ESRB Rating



M-rated titles clearly outperform the rest, with a median around 0.30 million and the top 90% reaching up to approximately 1.75 million units. In contrast, E, E10+, and T games all cluster more tightly at lower sales levels, with medians near 0.18 million and 90th percentiles below or around 1.05 million copies.

5 Observations and Next Steps

Overall, our exploratory analysis revealed several clear patterns: higher critic and user scores tend to coincide with greater global sales, albeit with many low-selling outliers. Games that attract more reviews—whether

from critics or players—also generally perform better in the market. Platforms such as the NES and PlayStation stand out for their stronger sales performance, while genres like Shooter and Platform show higher median global sales. We also observed a gradual rise in top-selling games over time, with a peak around the mid-2000s.

Next, we shift from description to inference—testing whether these observed patterns hold up statistically. We will conduct hypothesis tests and compute confidence intervals for score–sales relationships, run ANOVA to compare mean sales across genres and platforms, and develop a multiple linear regression model. If appropriate, we may also explore logistic regression to identify which factors best predict whether a game becomes “successful” (e.g., selling over one million units).

References

- Shams, Rushdi. *Video Game Sales with Ratings*. <https://www.kaggle.com/datasets/rush4ratio/video-game-sales-with-ratings>. Accessed: 2025-05-02, 2018.
- Triantafillou, Sofia. *MEM 264 – Applied Statistics*. <https://elearn.uoc.gr/course/view.php?id=5842section-0>, 2025.