Analysis of Video Game Sales

Introduction:

The video game industry has experienced significant growth in recent years, becoming a major form of entertainment for millions of people worldwide. Understanding the factors that influence video game sales can provide valuable insights for game developers, publishers, and marketers. In this report, we analyze the "Video Game Sales" dataset, which contains information on the sales of video games across different platforms, genres, and regions.

The dataset comprises the following variables:

- Rank: Ranking of overall sales

- Name: Name of the video game

- Platform: Platform on which the game was released (e.g., Wii, PS4, Xbox One)

- Year: Year of release

- Genre: Genre of the video game (e.g., Action, Sports, Racing)

- Publisher: Publisher of the video game

- NA\_Sales: North American sales in millions of units

- EU\_Sales: European sales in millions of units

- JP\_Sales: Japanese sales in millions of units

- Other\_Sales: Sales in other regions in millions of units

- Global\_Sales: Total global sales in millions of units

Data link:

[Video Game Sales | Kaggle](https://www.kaggle.com/datasets/gregorut/videogamesales?resource=download)

Our analysis aims to explore the relationships between these variables and identify factors that impact video game sales. The report is structured as follows:

1. Data Cleaning and Preprocessing

2. Exploratory Data Analysis

3. Hypothesis Testing

4. Modeling

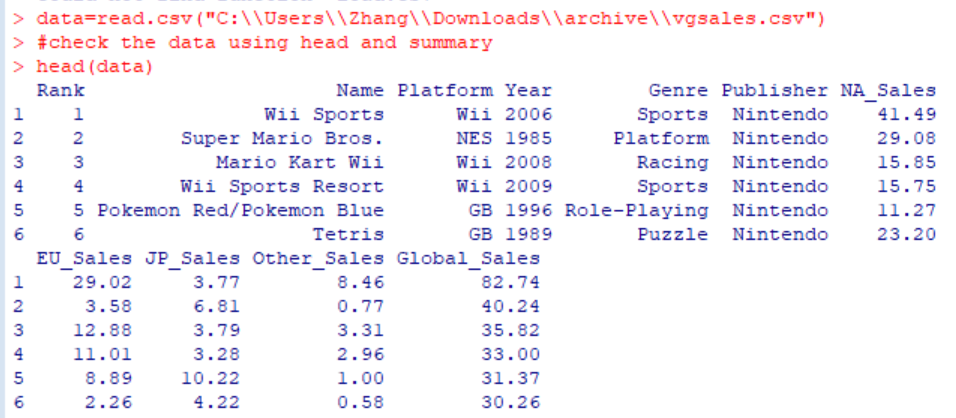
5. Conclusion and Recommendations

In the following sections, we will walk through each step of the analysis and present our findings.

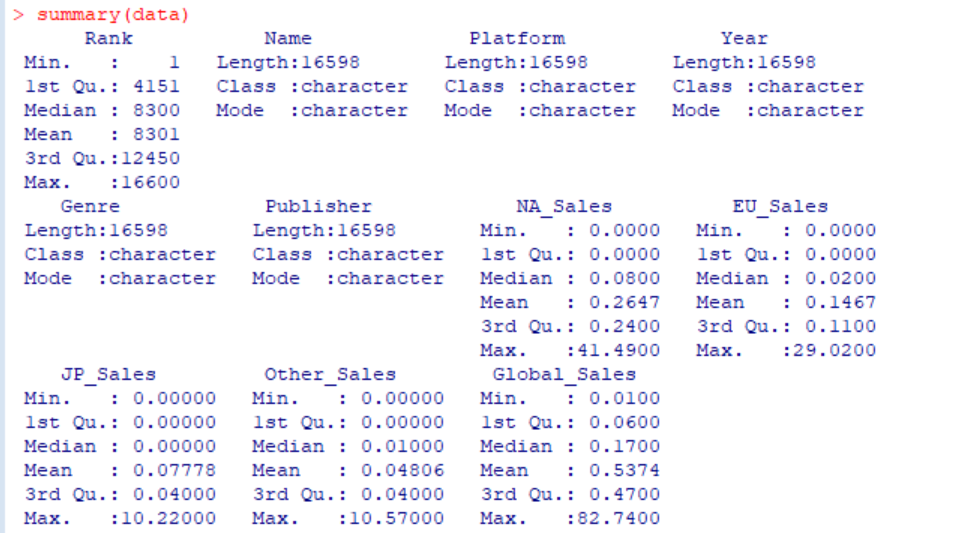
Here's the matched structure with the corresponding step numbers:

1. Data Cleaning and Preprocessing

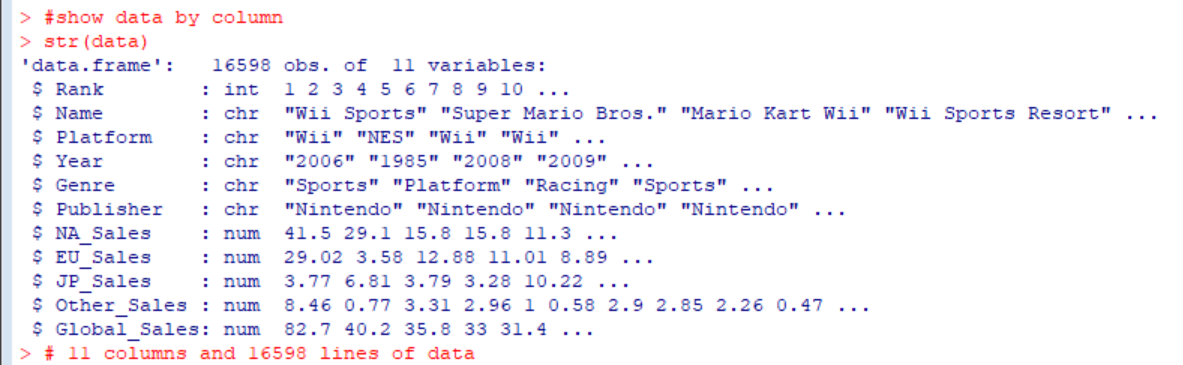
-a. Load and inspect the data



As the data structure showed before, the head(data) function print out part of the data.



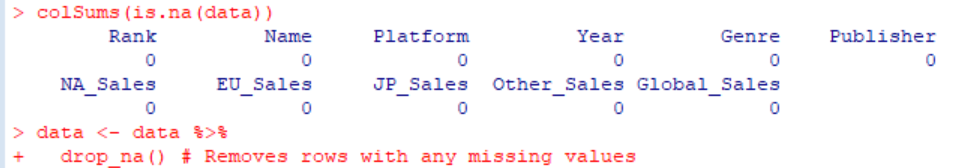
While doing the summary(data), it shows attributes like min or max of the data, which helps make a better decision.



Do the str(data) and check the data type. As we can see in the result, data type of year is chr. Need to translate it into num while necessary.

b. Clean and preprocess the data

N/A data checking---



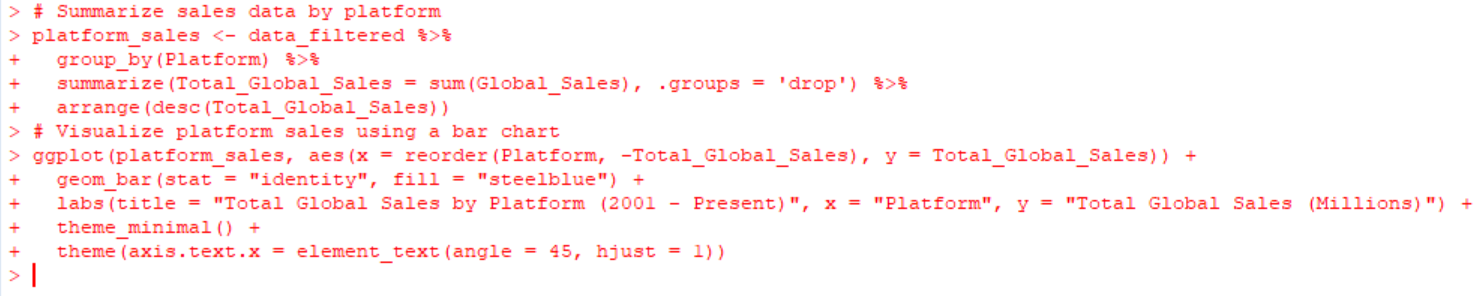
Data filtering---

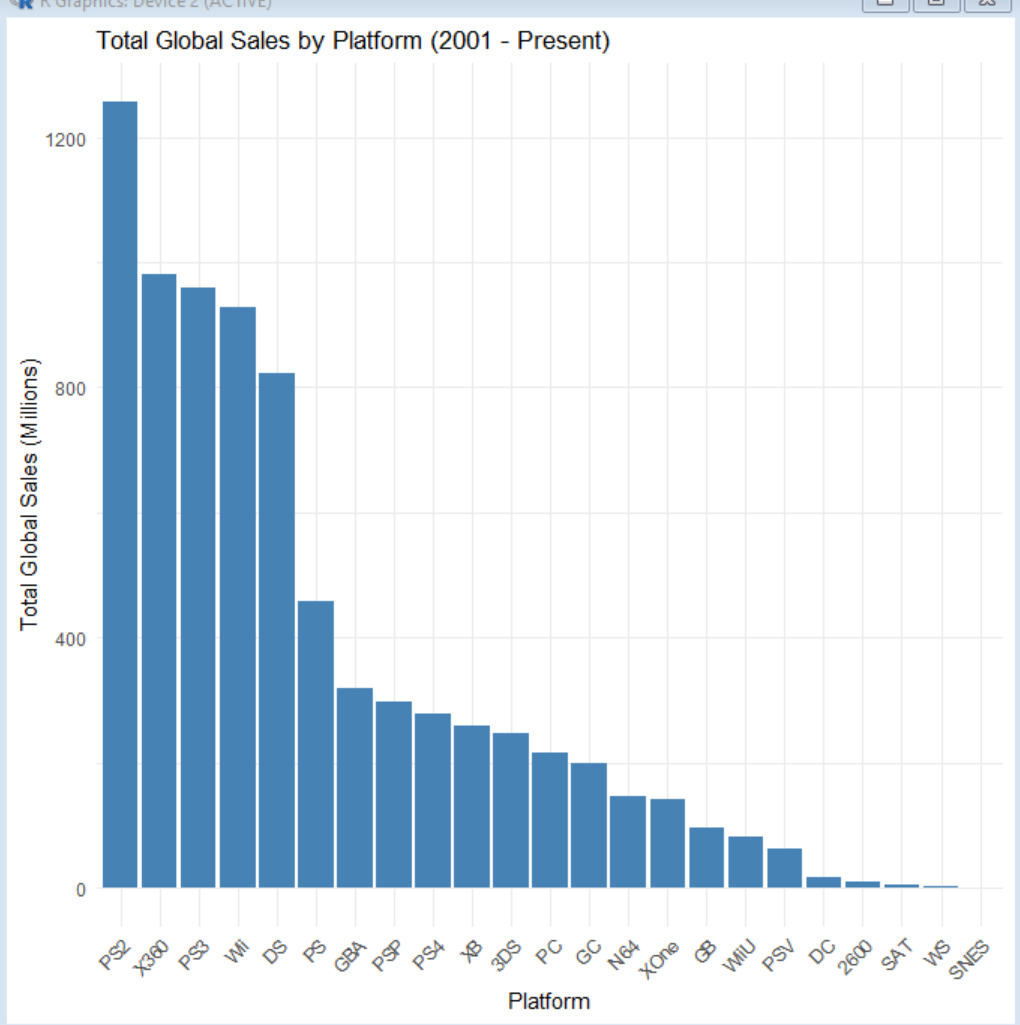


2. Exploratory Data Analysis

a. Summary statistics

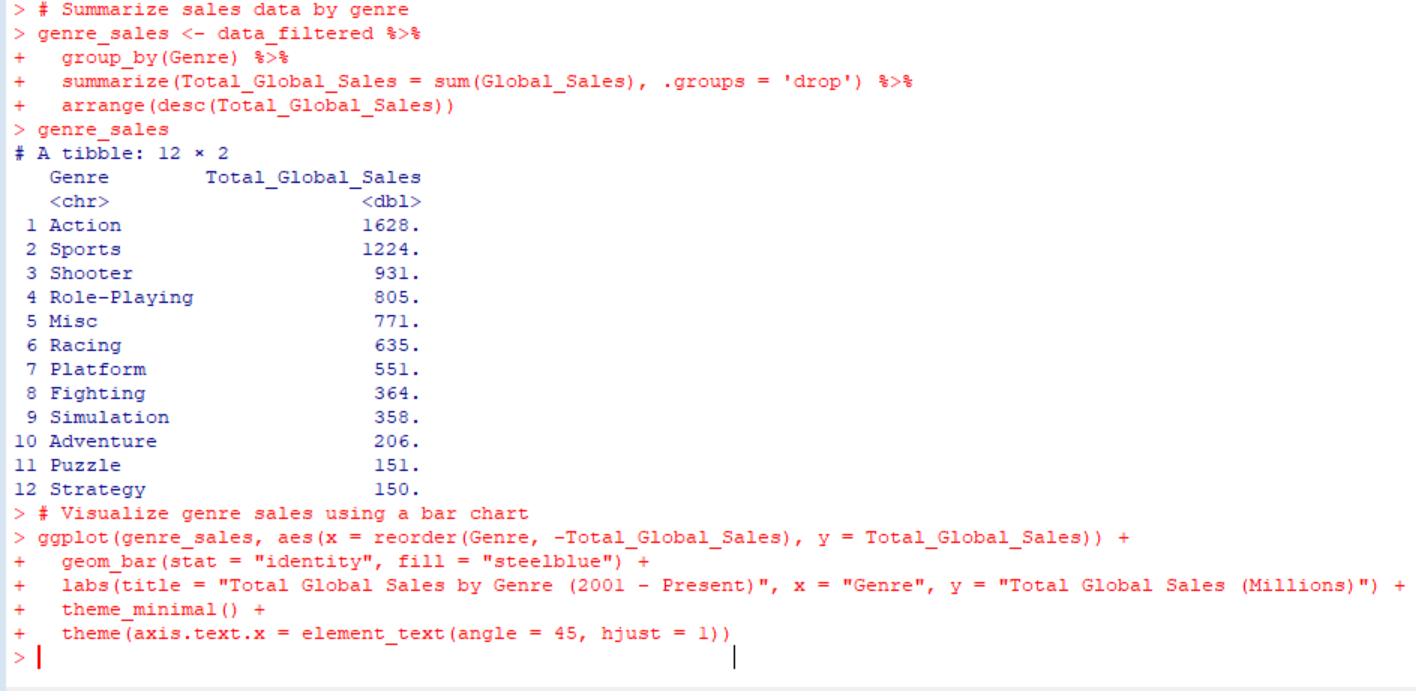
Sum by platform and visualize it

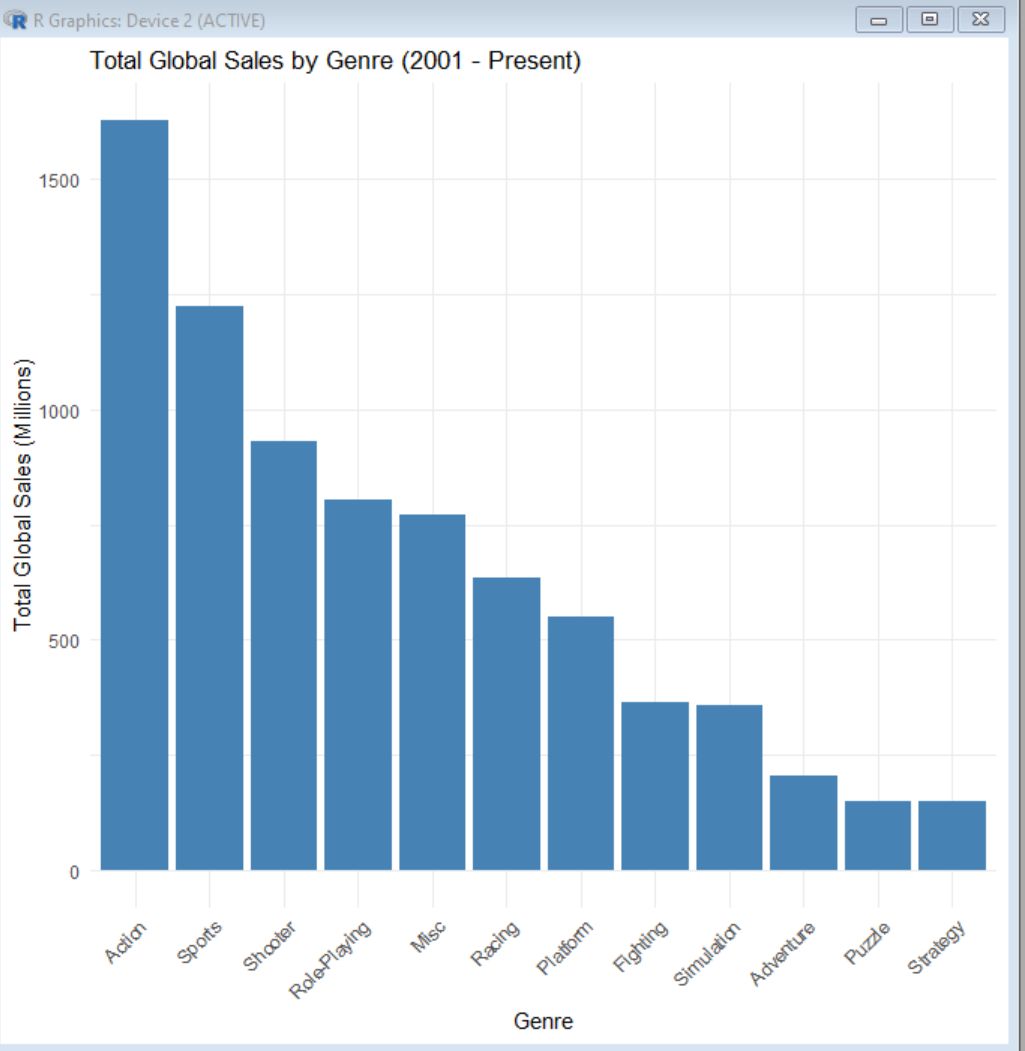




Could know the TOP 5---PS2, X380, PS3, WII, DS are far more be favored than others.

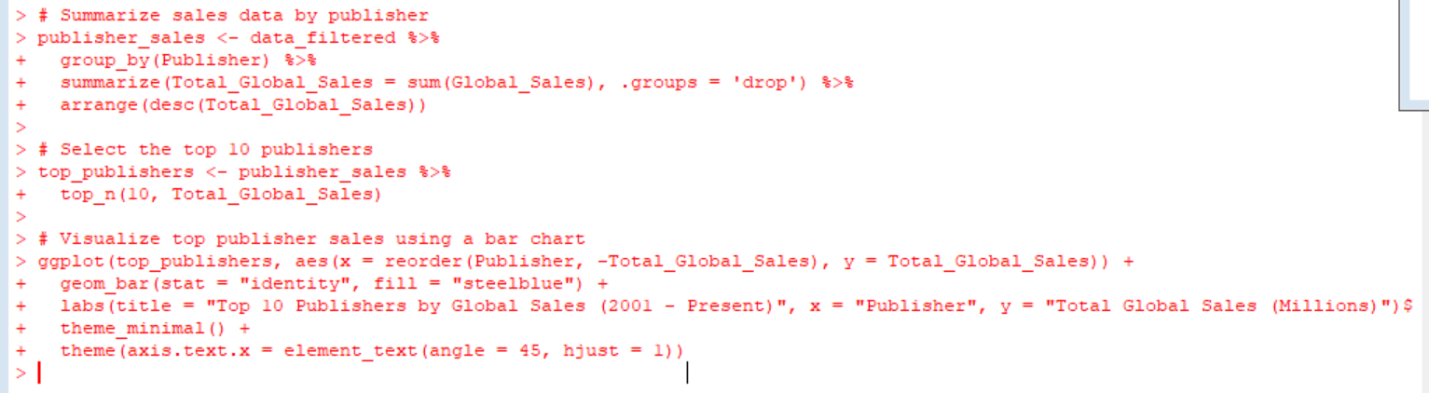
b. Visualize global sales by genre





While visualize global sales by genre , we could know the TOP 5 genres are action, sports, shooter, ride-playing and music.

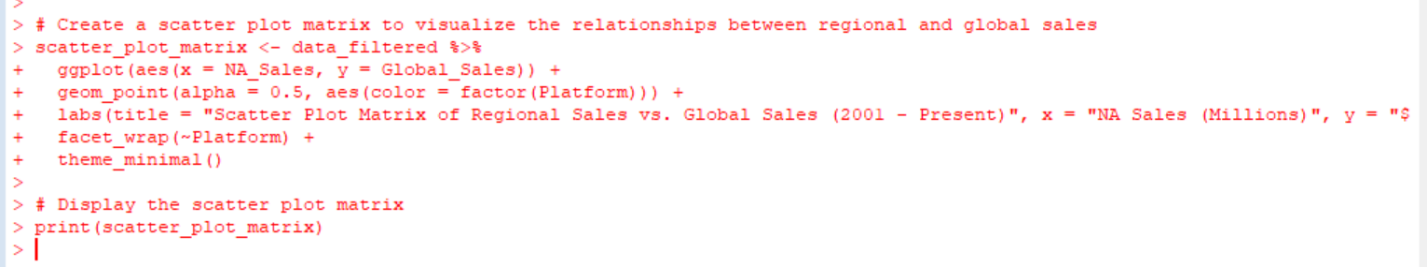
c. Visualize global sales by platform



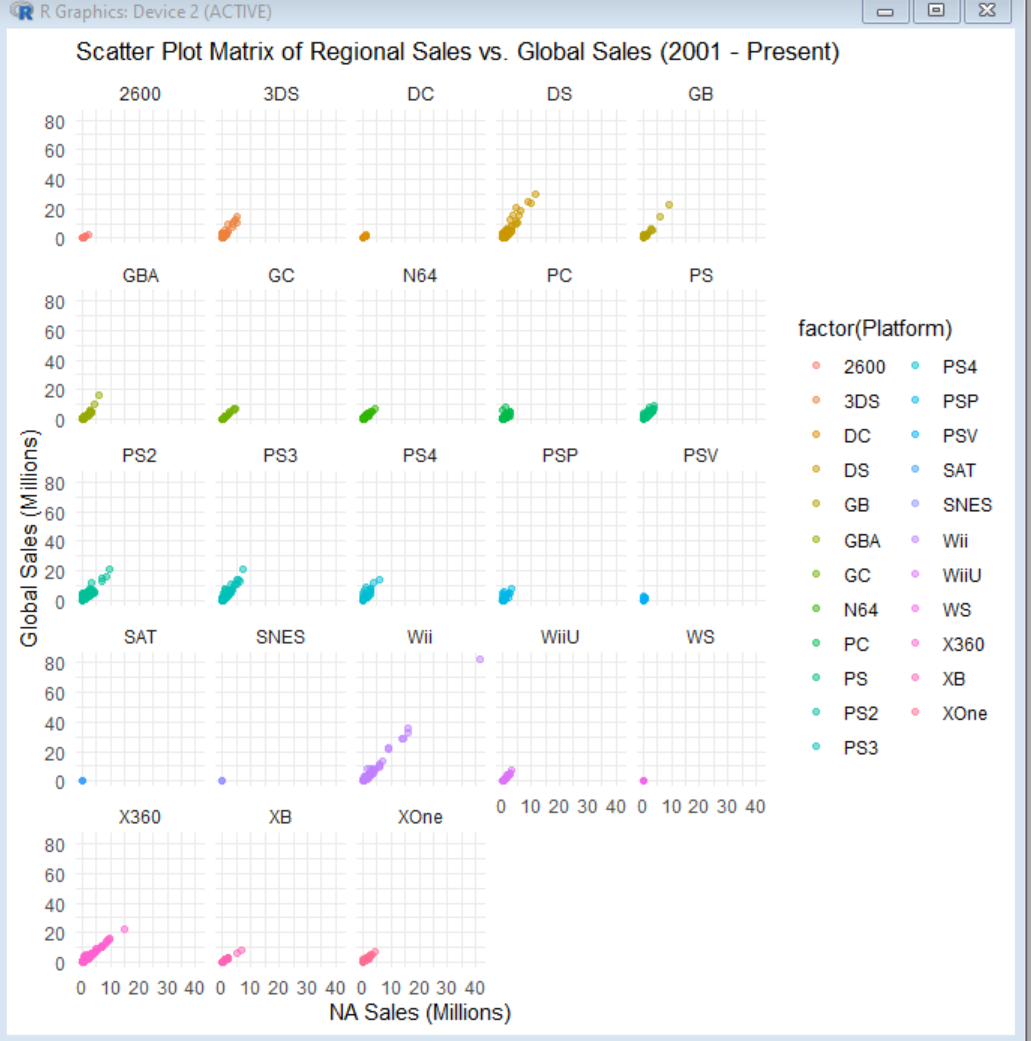


Then I print out the TOP 10 pushlishers.

d. Visualize global sales by regional

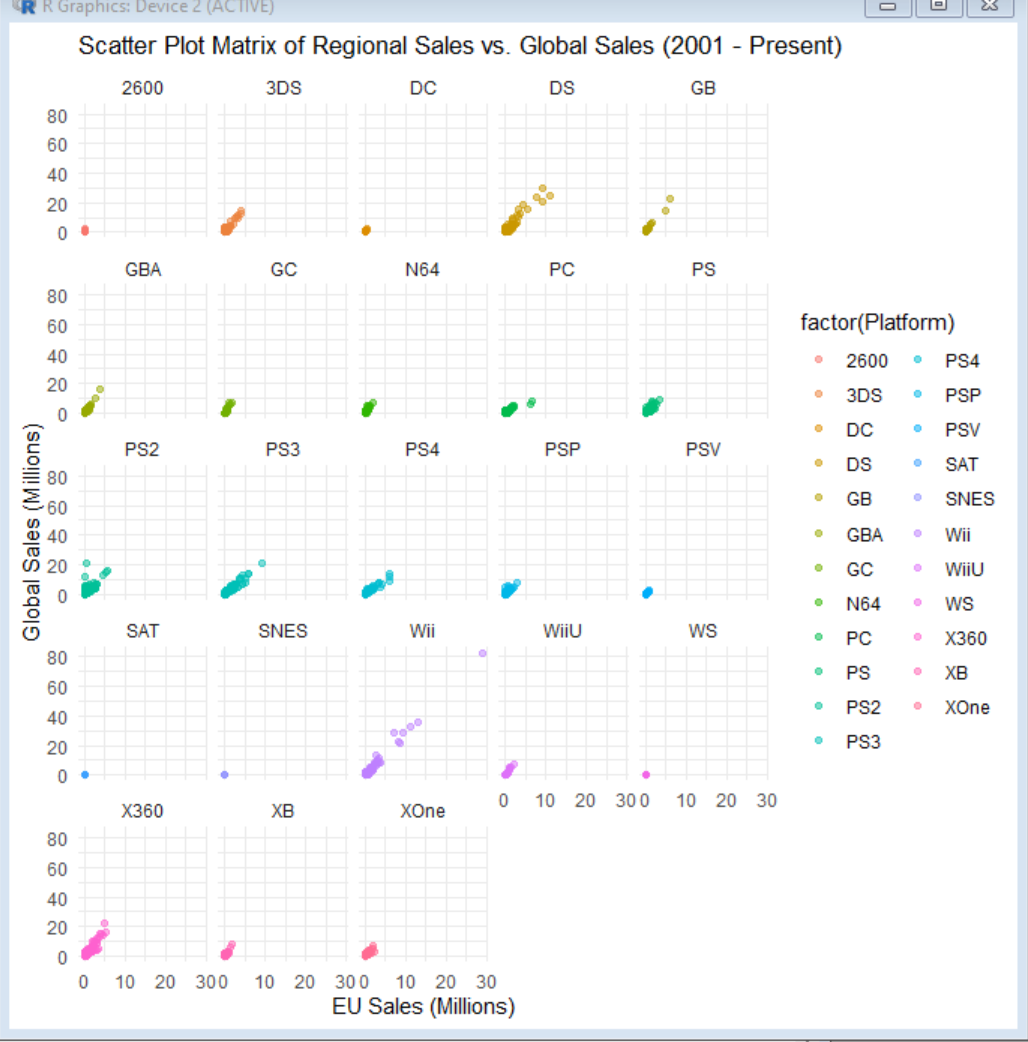


NA Sales



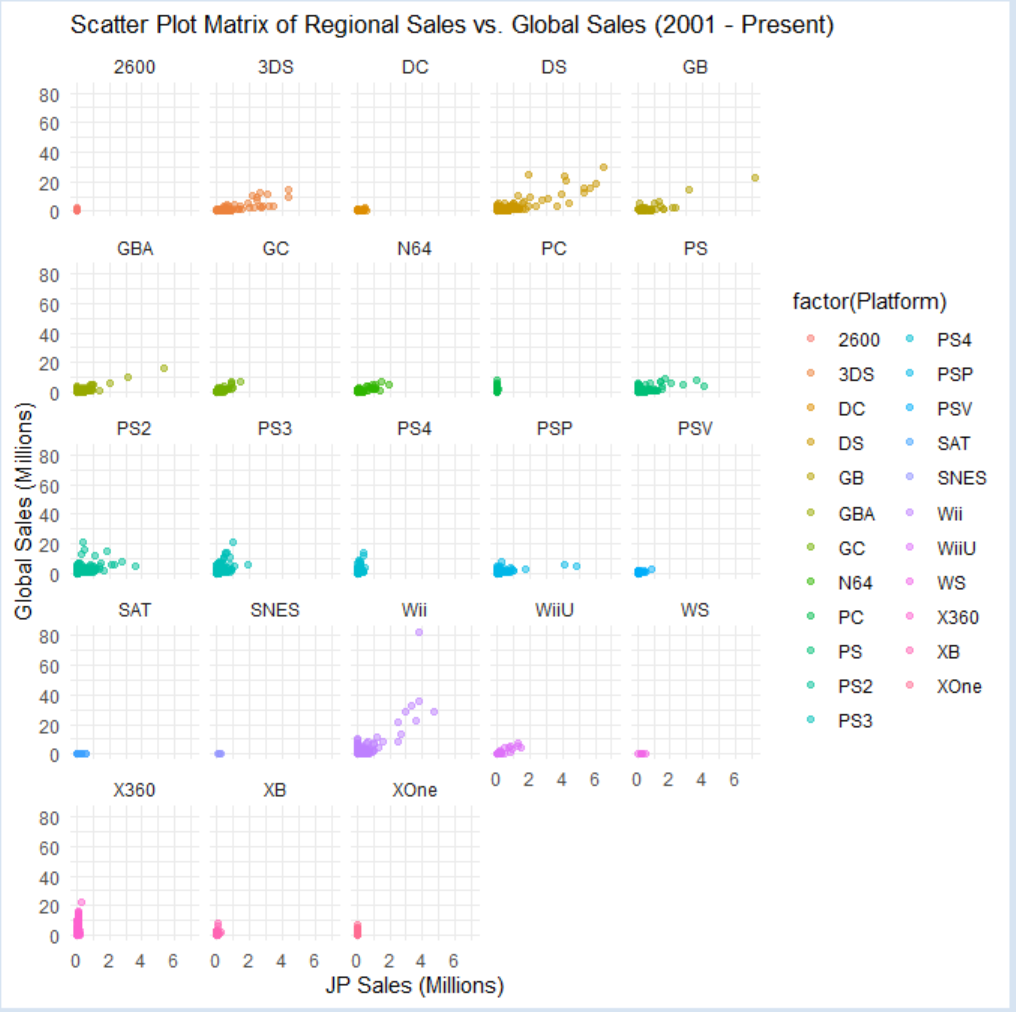
Here I did visualization of global sales by regional NA. It is easy to know regional sales somehow has relationship to the global sales data.

EU sales



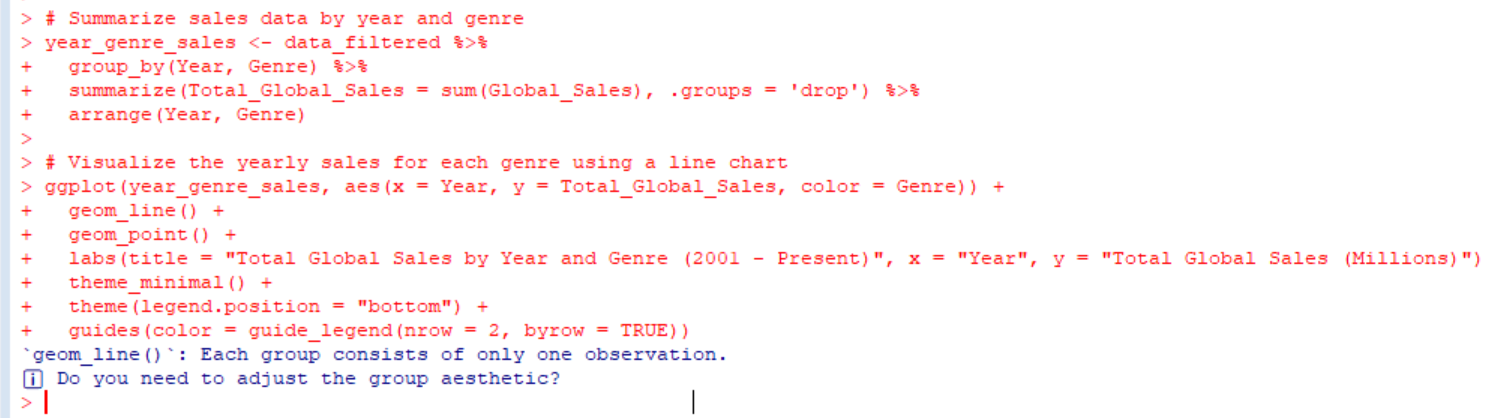
Here I did visualization of global sales by regional EU. It is also easy to know regional sales somehow has relationship to the global sales data.

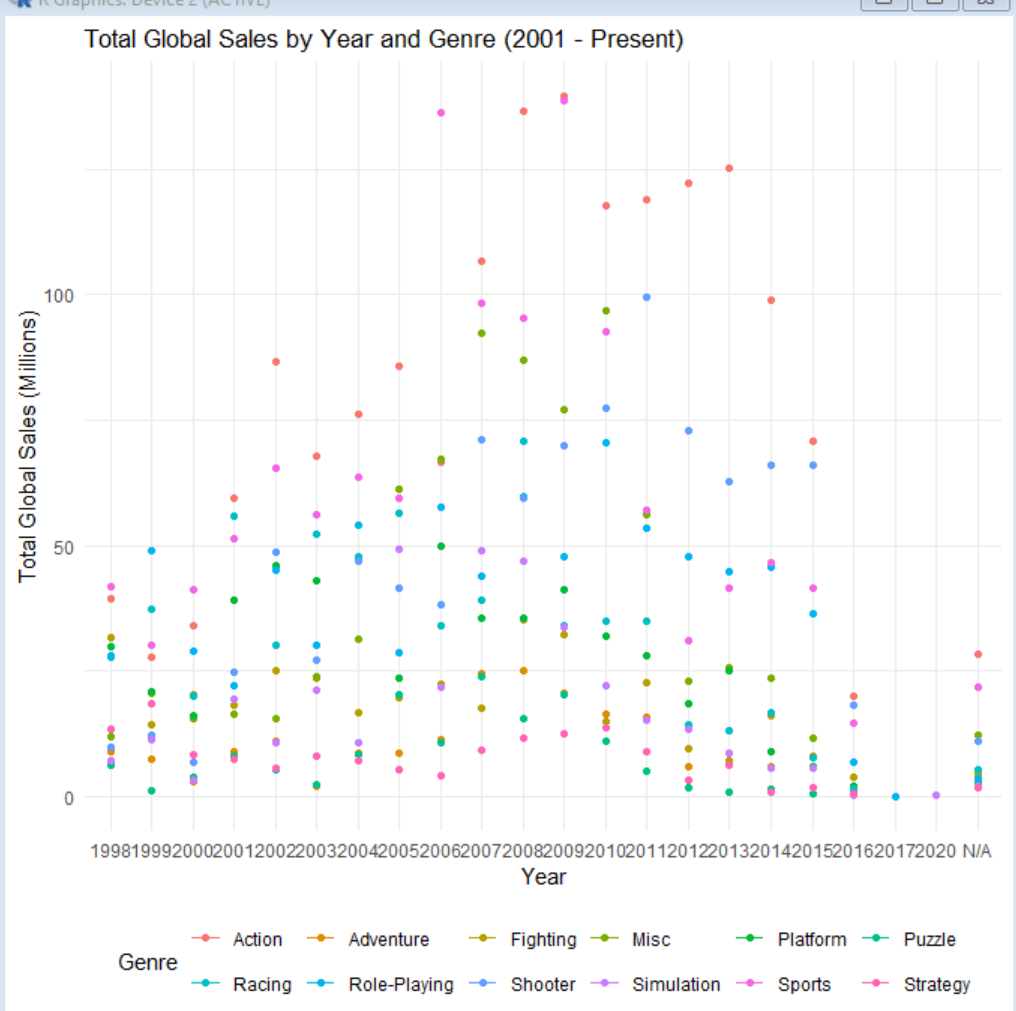
JP sales



Here I did visualization of global sales by regional JP. It is also easy to know regional sales somehow has relationship to the global sales data. While the relationship showed in the graph is more inaccurate.

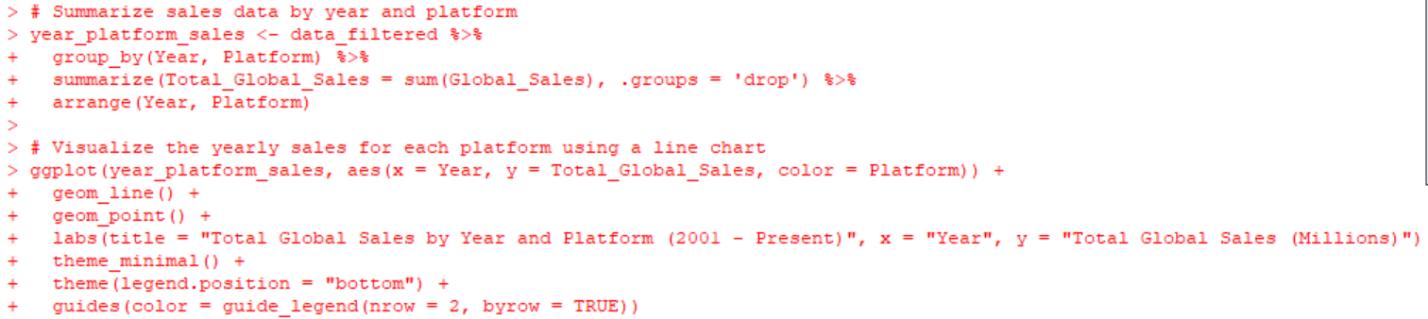
e: Visualize regional sales by year and genre

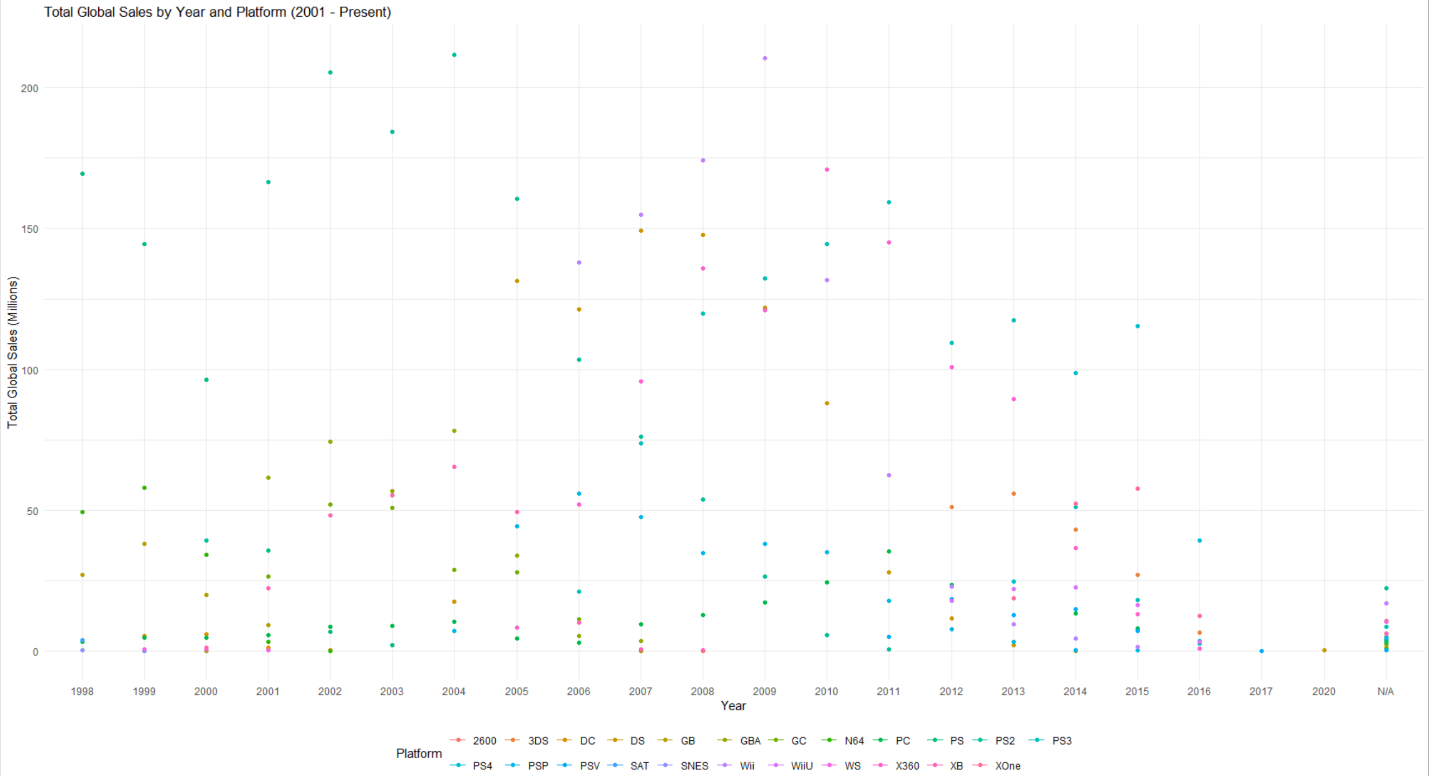




When visualizing regional sales by year and genre, we could know the total sales different year by year, while the rank of the genre in the sales data is somehow determined.

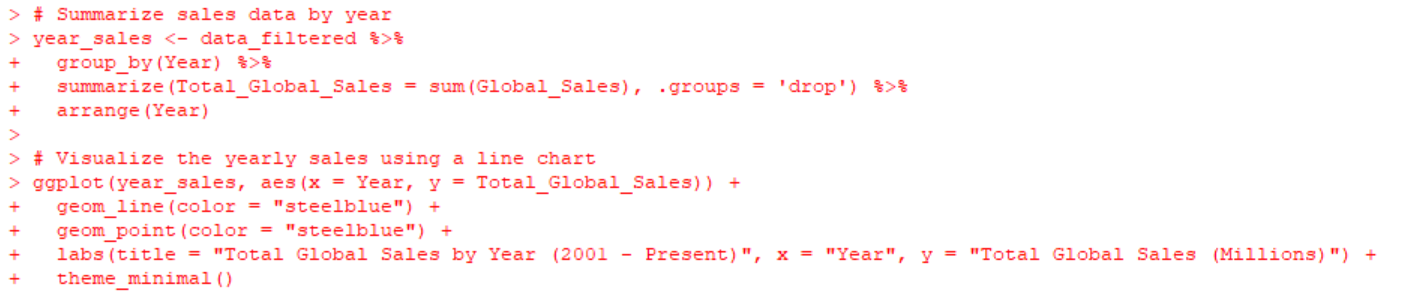
f: Visualize regional sales by year and platform

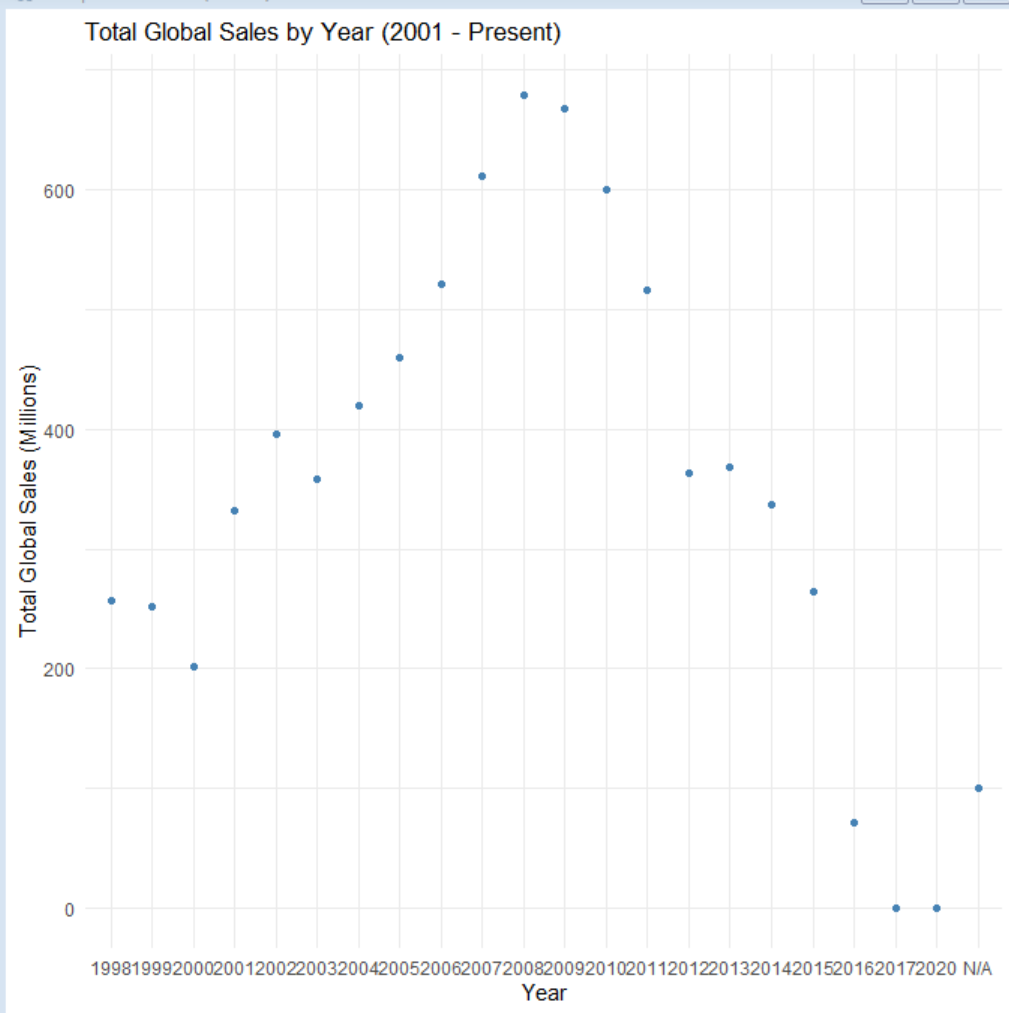




While doing visualization of global sales by year and platform, we could see every brand is different to each other and some brand like PS really matters.

g: Visualize regional sales by year

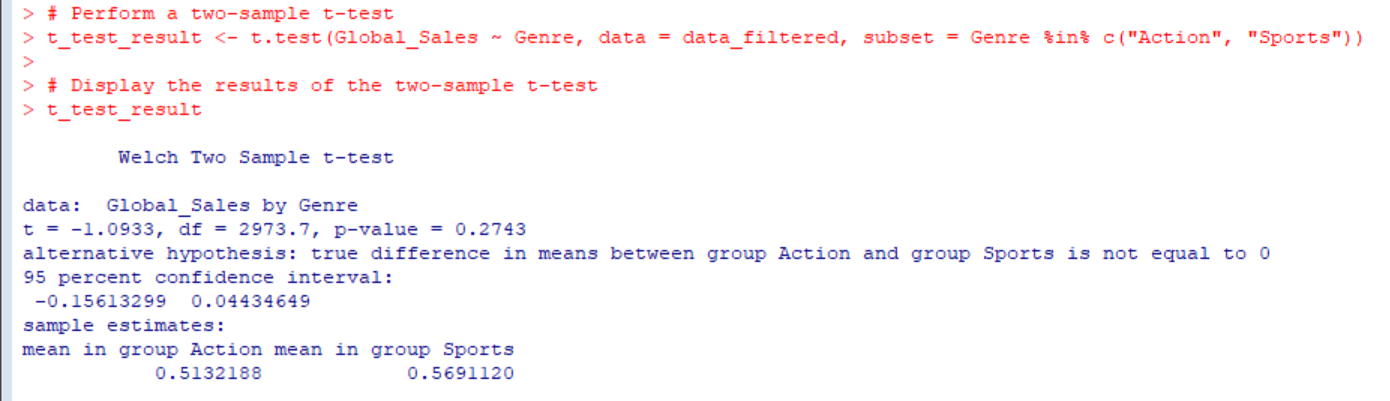




While simply visualizing the global sales data by year, we could assume there was a burst until 2008 and then there was a depression after 2009.

3. Hypothesis Testing

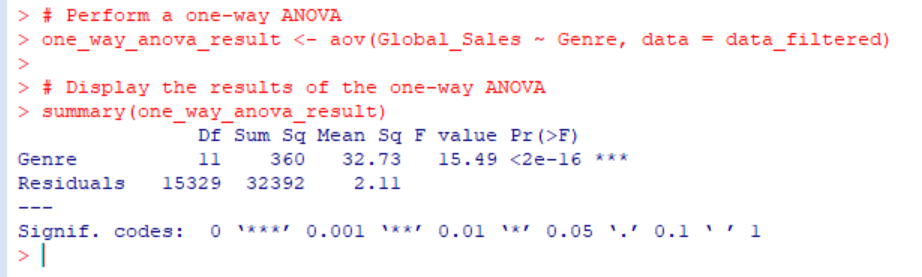
a: Two-sample t-test



We performed a two-sample t-test to compare the mean global sales between Action and Sports games. The null hypothesis states that there is no significant difference in the mean global sales between these two genres, while the alternative hypothesis states that there is a significant difference.

With a p-value of 0.2743, which is greater than the significance level (typically 0.05), we fail to reject the null hypothesis. This means that we do not have sufficient evidence to conclude that there is a significant difference in the mean global sales between Action and Sports games. The 95% confidence interval for the difference in means ranges from -0.1561 to 0.0443 million units, indicating that we cannot rule out the possibility that the difference in means is zero or falls within this range.

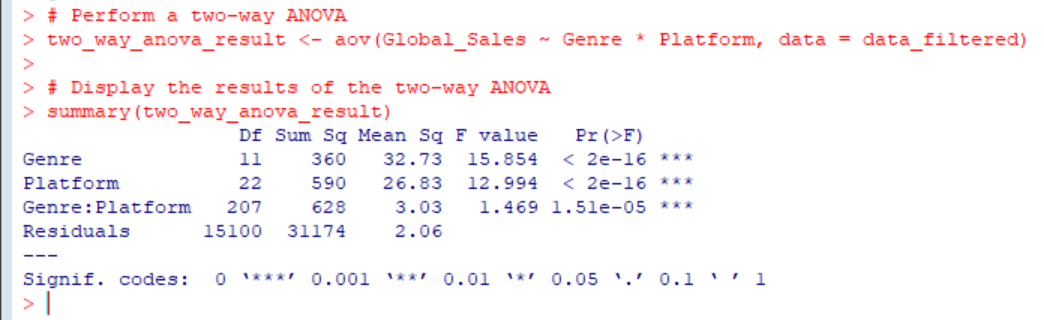
b: One-way ANOVA



We performed a one-way ANOVA to test if there is a significant difference in the mean global sales among the different genres. The null hypothesis states that there is no significant difference in the mean global sales among genres, while the alternative hypothesis states that there is a significant difference.

With a p-value of < 2e-16, which is much lower than the significance level (typically 0.05), we reject the null hypothesis. This means that we have strong evidence to conclude that there is a significant difference in the mean global sales among the different genres. The F-value of 15.49 also indicates that the variation in global sales among the genres is greater than the variation within each genre. This suggests that the genre of a video game plays a role in influencing its global sales.

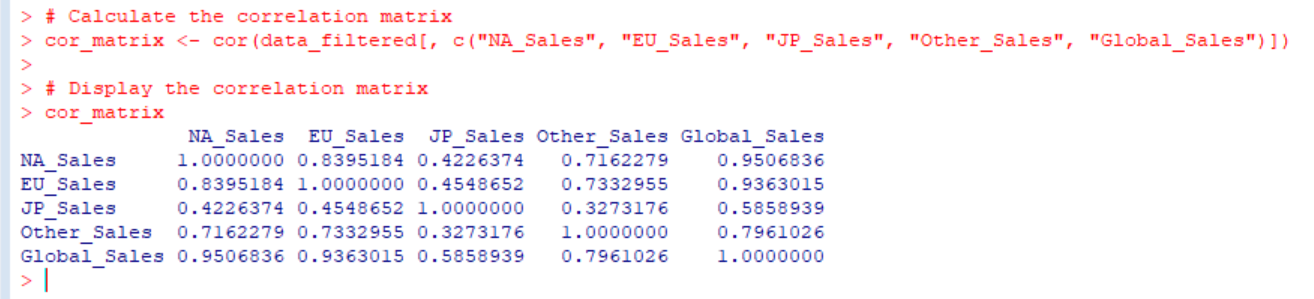
c: Two-way ANOVA



We performed a two-way ANOVA to test if there is a significant interaction between genre and platform on the mean global sales. The null hypothesis states that there is no significant interaction between genre and platform on the mean global sales, while the alternative hypothesis states that there is a significant interaction.

With p-values < 2e-16 for both genre and platform, we have strong evidence to conclude that there are significant differences in the mean global sales among the different genres and platforms. Furthermore, the p-value for the interaction between genre and platform (1.51e-05) is also significant, suggesting that there is a significant interaction effect between genre and platform on the mean global sales. This implies that the relationship between genre and global sales may vary depending on the platform, and vice versa.

d: Calculate the correlation matrix between these regional



We calculated the correlation matrix for the regional sales (North America, Europe, Japan, and Other regions) and the global sales. The correlation matrix shows the pairwise correlation coefficients between the variables, providing insights into the relationships between regional sales and global sales.

There is a strong positive correlation between North American sales and global sales (r = 0.9507), suggesting that games with high sales in North America also tend to have high global sales.

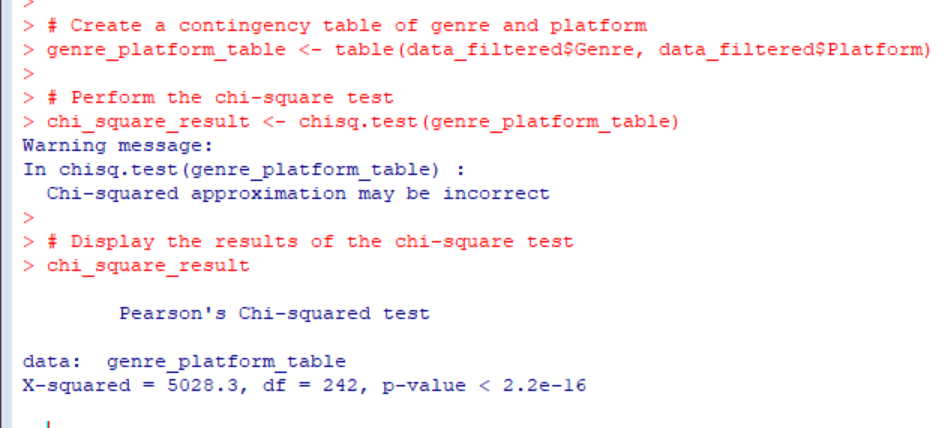
Similarly, there is a strong positive correlation between European sales and global sales (r = 0.9363), implying that games with high sales in Europe also tend to have high global sales.

Japanese sales and global sales have a moderate positive correlation (r = 0.5859), indicating that games with high sales in Japan may not necessarily have high global sales, though there is still some relationship between the two variables.

Sales in other regions have a strong positive correlation with global sales (r = 0.7961), suggesting that games with high sales in other regions also tend to have high global sales.

Regional sales also show varying degrees of positive correlation with each other. The highest correlation is between North American sales and European sales (r = 0.8395), while the lowest is between Japanese sales and sales in other regions (r = 0.3273).

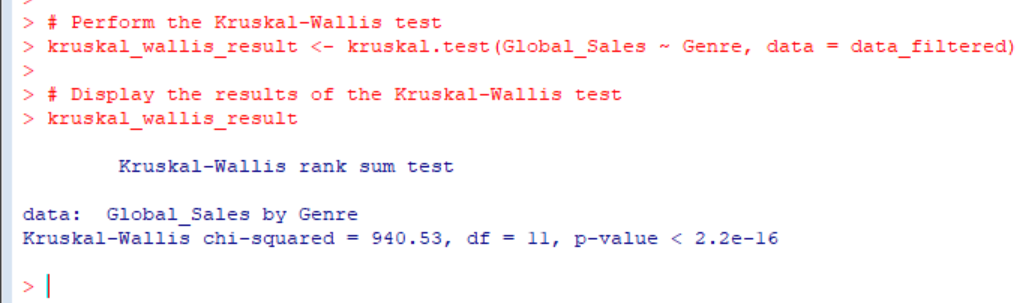
e: Chi-square test



We created a contingency table for genre and platform and performed a chi-square test to investigate the independence between the two variables. The null hypothesis states that genre and platform are independent, while the alternative hypothesis states that there is an association between them.

With a p-value of < 2.2e-16, which is much lower than the significance level (typically 0.05), we reject the null hypothesis. This means that we have strong evidence to conclude that there is an association between genre and platform. However, the warning message suggests that the chi-squared approximation may be incorrect, which could be due to small cell counts in the contingency table. To address this issue, we may need to consider using other methods, such as Fisher's exact test or a Monte Carlo simulation, to obtain more accurate results.

f: Non-parametric tests (Kruskal-Wallis test)



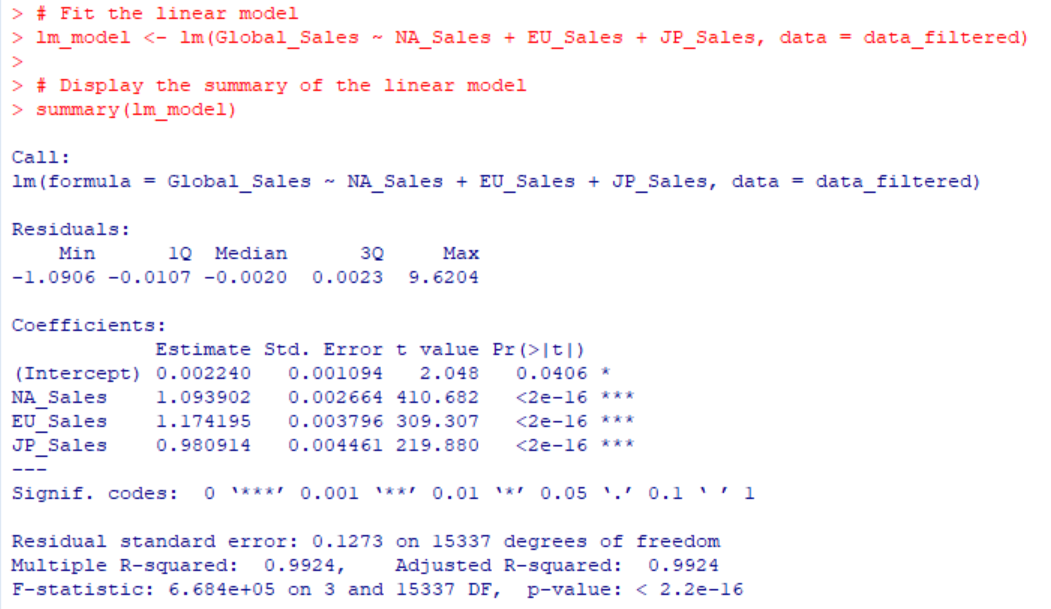
We performed a Kruskal-Wallis test to assess whether there is a significant difference in global sales among the different genres. The Kruskal-Wallis test is a non-parametric alternative to the one-way ANOVA and does not assume normality of the data. The null hypothesis states that the distribution of global sales is the same across all genres, while the alternative hypothesis states that at least one genre has a different distribution.

With a p-value of < 2.2e-16, which is much lower than the significance level (typically 0.05), we reject the null hypothesis. This means that we have strong evidence to conclude that there is a significant difference in the distribution of global sales among the different genres. This result is consistent with the findings from the one-way ANOVA, further supporting the conclusion that the genre has a significant impact on global sales.

4. Modeling

a. Linear regression model

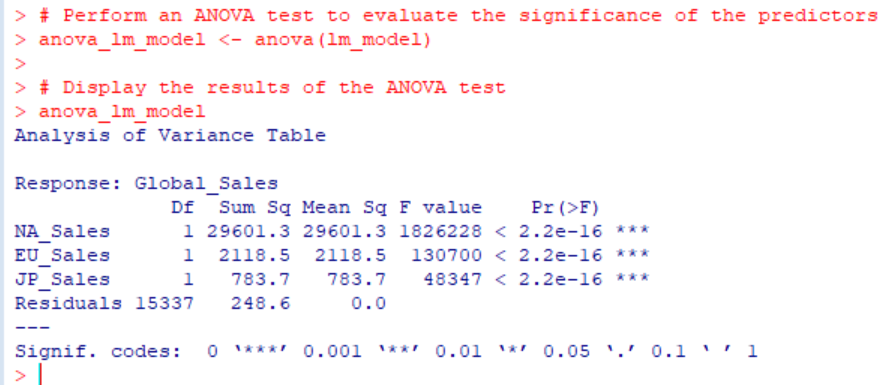
a-a Perform a linear regression to predict global sales using regional sales



We fit a linear model to predict global sales based on the sales in North America (NA\_Sales), Europe (EU\_Sales), and Japan (JP\_Sales).

The linear model demonstrates a strong relationship between global sales and the sales in North America, Europe, and Japan. All three predictors (NA\_Sales, EU\_Sales, and JP\_Sales) have highly significant p-values (< 2e-16), indicating that they are important factors in predicting global sales. The high R-squared and adjusted R-squared values (0.9924) indicate that approximately 99.24% of the variance in global sales can be explained by this linear model. The F-statistic and its associated p-value also suggest that the model is statistically significant.

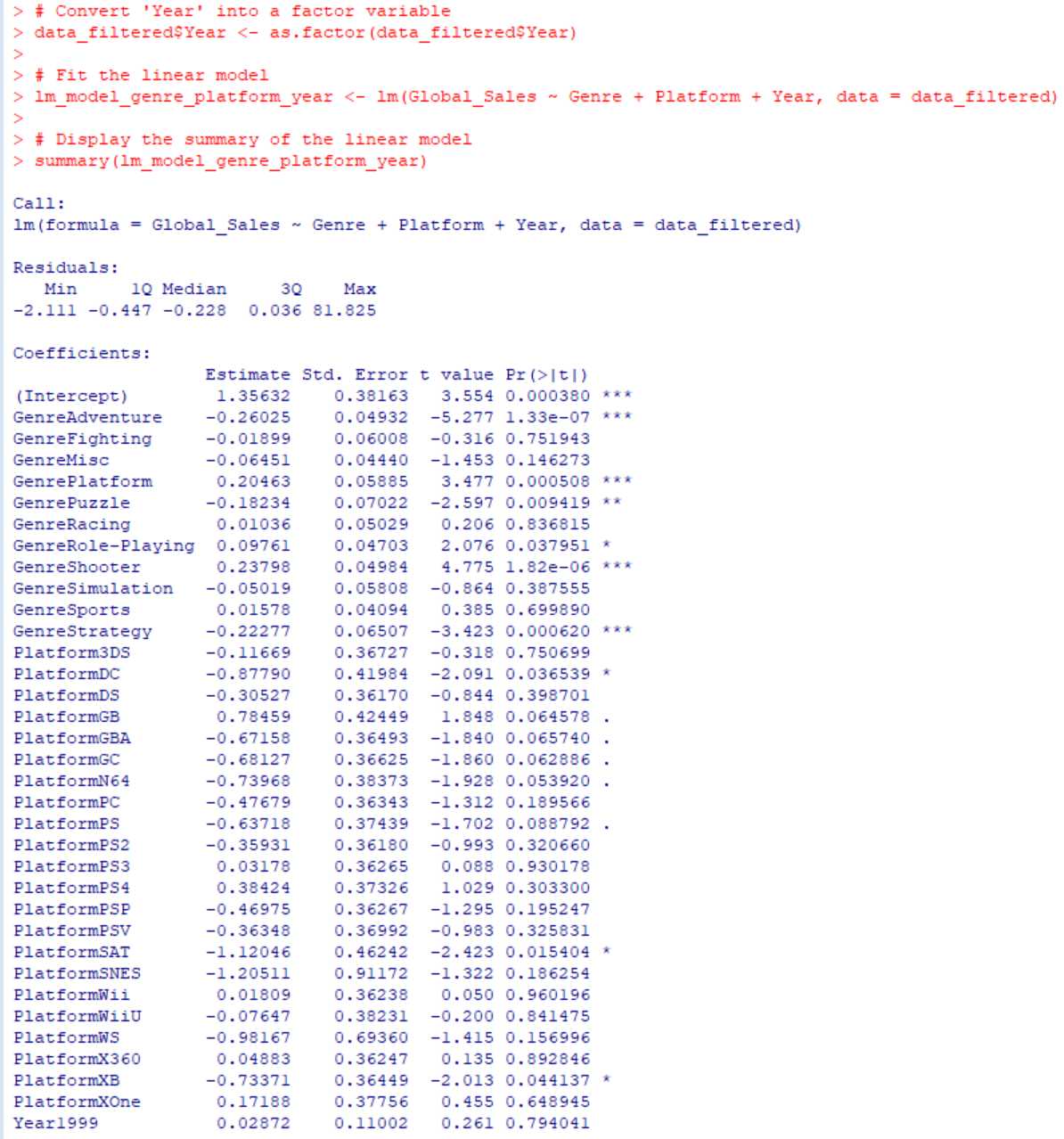
Anova:

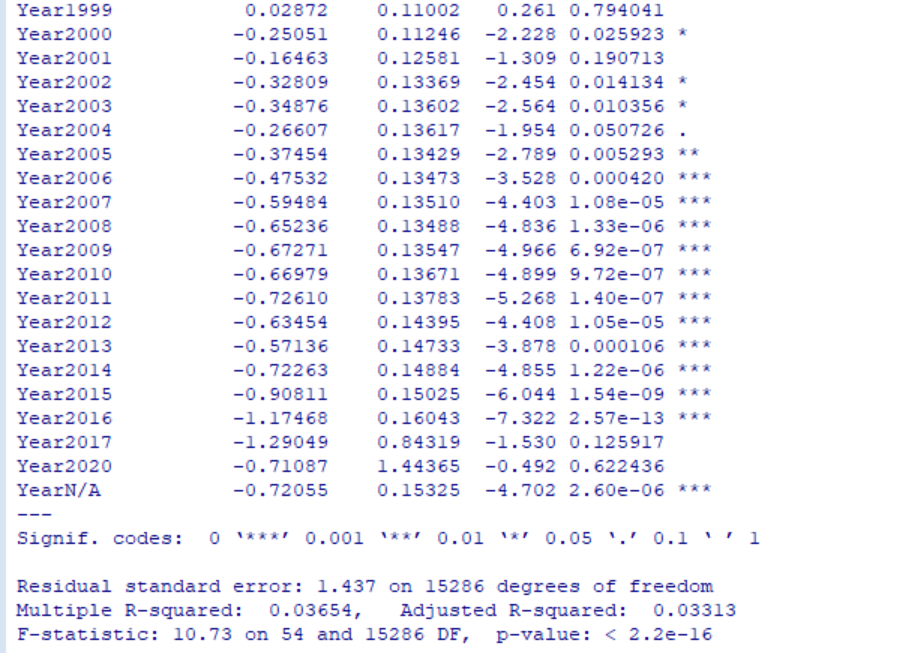


We performed an ANOVA test to evaluate the significance of the predictors in our linear model that predicts global sales based on the sales in North America (NA\_Sales), Europe (EU\_Sales), and Japan (JP\_Sales).

The ANOVA test results confirm the significance of all three predictors (NA\_Sales, EU\_Sales, and JP\_Sales) in the linear model. All of the predictors have highly significant p-values (< 2.2e-16), indicating that they contribute significantly to predicting global sales. These results further support the conclusion that sales in North America, Europe, and Japan are important factors in determining global sales for video games.

a-b Perform a linear regression to predict global sales using genre, platform, and year





We fitted a linear model to predict global sales using genre, platform, and year as predictors. Here's the summary of the linear model:

The model has a relatively low R-squared value, indicating that it is not very good at predicting global sales using genre, platform, and year.

Some coefficients are significant, which means they have a significant impact on predicting global sales. For example, GenreAdventure, GenrePlatform, GenrePuzzle, GenreRole-Playing, GenreShooter, GenreStrategy, PlatformDC, PlatformSAT, and PlatformXB, as well as some years (e.g., 2000, 2002, 2003, 2005, 2006, 2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014, 2015, and 2016) have significant coefficients. Note that the significance codes follow this pattern: 0 '' 0.001 '' 0.01 '' 0.05 '.' 0.1 ' ' 1.

However, the overall model's low R-squared value suggests that there is a lot of unexplained variation in global sales, and these predictors alone may not provide a comprehensive understanding of the factors that influence global video game sales.

5. Conclusion and Recommendations

In this analysis, we explored the video game sales dataset to identify patterns and relationships among various factors like genre, platform, and sales across different regions. We applied descriptive statistics, correlation analysis, chi-square test, Kruskal-Wallis test, linear regression, and ANOVA to better understand these relationships. Here are our conclusions and recommendations:

1. Correlation analysis showed a strong relationship between sales in North America, Europe, and other regions, indicating that popular games in one region tend to be popular in other regions as well.

2. The chi-square test indicated a significant association between the genre and platform of the video games, suggesting that certain genres are more popular on specific platforms.

3. The Kruskal-Wallis test revealed significant differences in global sales among different genres, indicating that some genres are more popular and generate higher sales than others.

4. The linear regression model with regional sales (North America, Europe, and Japan) as predictors had an extremely high R-squared value, indicating that these predictors can explain a substantial portion of the variation in global sales.

5. The linear regression model with genre, platform, and year as predictors had a low R-squared value, indicating that these factors alone may not provide a comprehensive understanding of the factors influencing global video game sales.

Based on our findings, we recommend the following:

- Game developers and publishers should focus on creating games in popular genres and for platforms that have a strong association with those genres to maximize sales potential.

- Regional preferences should be considered when marketing games, as the strong correlation between regional sales indicates that games that perform well in one region are likely to do so in others.

- Although the linear model with genre, platform, and year as predictors had a low R-squared value, these factors still provide valuable insights. It would be beneficial to explore additional factors, such as game ratings, developer reputation, or marketing spend, to build a more comprehensive predictive model.

- Finally, understanding industry trends and keeping an eye on emerging platforms and genres is crucial for staying competitive in the ever-evolving video game market.