Analysis of Trends and Sales of Video Game Industry

Group 4 • Vishal Orsu • Rishik Reddy Ragi • Sai Manish Reddy Pannala
STAT515 Final Project

I ABSTRACT

In this report, we conducted an exploratory data analysis on a video game sales dataset that includes information on game titles, platforms, genres, publishers, release years, and sales data. The dataset contains information on more than 16,000 video games from different regions of the world over the past few decades. Our analysis focused on understanding the distribution of video game sales across different platforms, genres, and regions. We utilized various data visualization techniques, such as bar charts, heatmaps, and scatter plots, to gain insights into the trends and patterns in the dataset. Our analysis revealed that the most popular video game platforms were the PlayStation 2, Xbox 360, and PlayStation 3. The most popular video game genres were Action, Sports, and Shooter. We also found that video game sales varied significantly by region, with the North American region contributing the highest share of global video game sales. Our findings suggest that there is a strong relationship between the popularity of a video game platform and the number of games released on that platform. Similarly, the popularity of a video game genre appears to be closely related to the number of games released within that genre. Our analysis provides insights that can be used by video game developers and publishers to better understand the video game market and consumer preferences and inform their decision-making regarding platform and genre selection.

Keywords • video games • sales • genre • platform • analysis • visualization • console • trends

II INTRODUCTION

The video game industry has grown significantly in recent years, with a wide variety of games available on various platforms. This growth has led to an increase in the amount of data available on video game sales, which can be analyzed to gain insights into the industry. In this report, we analyze a dataset of video game sales from different regions around the world, covering a wide range of platforms and genres.

The dataset [1] used in this analysis contains information on over 16,000 video games released between 1980 and 2020, including their sales in different regions, such as North America, Europe, Japan, and others. It also includes details on the platforms on which the games were released, such as Nintendo, Xbox, PlayStation, and others. Additionally, the dataset provides information on the genre of each game, such as action, sports, puzzle, and others.

The aim of this report is to provide insights into the video game industry by analyzing the sales data of games across different regions, platforms, and genres. We will explore the top-selling games, platforms, and genres, as well as identify any trends in the industry over time. Furthermore, we will create visualizations to help illustrate our findings and provide a better understanding of the data.

Our report will provide a comprehensive analysis of the video game industry which we gained by analyzing the dataset. It will highlight the potential insights that can be gained from analyzing such data and highlight the importance of data analysis in making informed decisions in the video game industry.

Three primary research questions that are focused on in this research are:

1. *Are certain genres of games more popular in specific regions?*
2. *How has the video game industry evolved over time?*
3. *Is there any relationship between critic and user scores and sales?*

III METHODOLOGY

This analysis was conducted using a dataset containing information on video game sales from various regions across the world, as well as the year of release, platform, genre, and publisher of each game. The dataset was obtained from Vgchartz [2], a website that hosts datasets for public use.

To begin the analysis, the dataset was first imported into RStudio, a popular software for data analysis. The dataset was then cleaned and preprocessed [3] to ensure that the data was in a suitable format for analysis. This included removing any missing or erroneous data and converting the data types to the appropriate format. After preprocessing, the data was analyzed using a variety of statistical and visualization techniques in order to gain insights into the video game industry. These techniques included descriptive statistics, data visualization using ggplot2, and hypothesis testing using statistical tests such as the t-test.One of the main goals of the analysis was to investigate the relationship between a game's genre, platform, and its sales performance. This was achieved by generating various plots and visualizations to explore the patterns and trends in the data, as well as performing statistical tests to determine if there were any significant differences between different groups.

This analysis provided valuable insights into the video game industry, including which genres and platforms are most popular, how sales performance varies across different regions, and which publishers have the most successful games. These insights could be used by industry professionals to inform decision-making and strategy development and could also be used by researchers and academics as a basis for further research.

IV DATA VISUALIZATIONS AND ANALYSIS

A. EXPLORATORY ANALYSIS:

The data visualizations and analysis in this study provide an in-depth understanding of the video game industry. The bar charts reveal the global sales trends across different platforms and genres. The heatmap [4] provides insight into the number of games released across platforms and genres. These visualizations give us a sense of which platforms and genres are most popular in the industry and how they compare to one another. The analysis of the data suggests that the industry has seen a shift towards more global sales on newer platforms such as PS4, XOne, and WiiU. Additionally, the analysis reveals that the most popular genres are action, sports, and shooter games. Overall, the data visualizations and analysis provide valuable insights into the video game industry that can be used to inform business decisions and future research.

- **Listed below are the few analyses we have conducted on the data provided:**

1. Univariate Analysis:

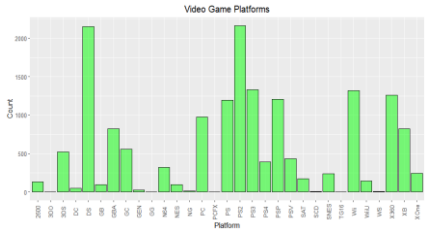


Figure 1

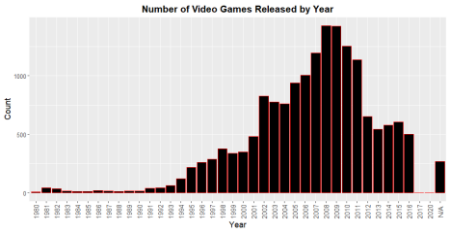


Figure 2

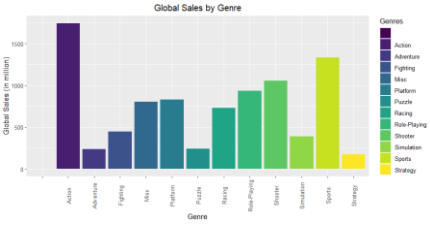


Figure 3

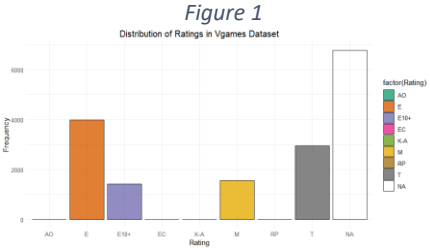


Figure 4

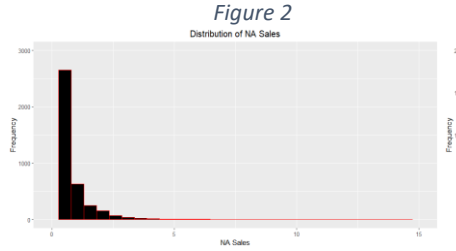


Figure 5

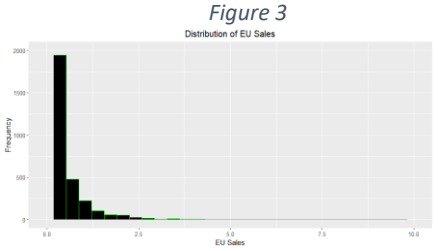


Figure 6

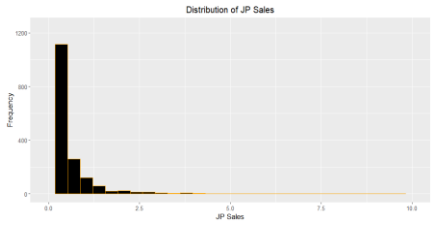


Figure 7

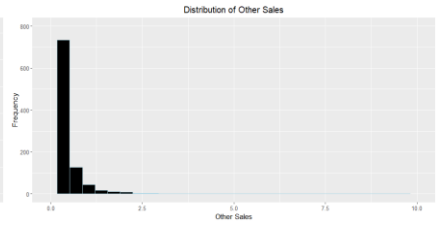


Figure 8

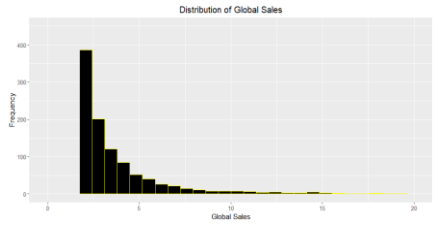


Figure 9

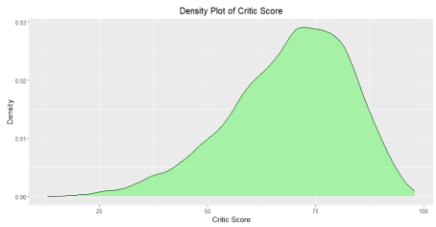


Figure 10

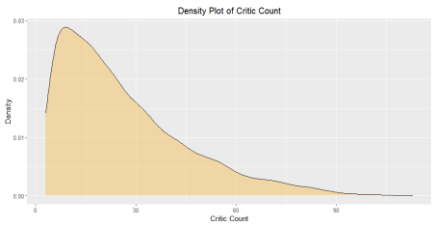


Figure 11

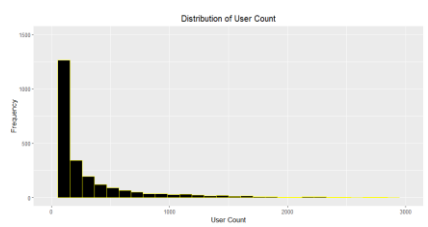


Figure 12

The graphs above show an exploratory analysis of the dataset for video game sales. Figure 1 shows a bar chart of the distribution(count) of video games across different platforms. Figure 2 represents a bar chart of the count of video games released in each year. Figure 3 displays a bar graph of the total global sales for each video game, separated by genre, in millions. Figure 4 depicts a bar chart of the frequency distribution of video game ratings in the dataset. Figure 5, 6 and 7 shows a histogram of the distribution of video game sales in North America, Europe, and Japan respectively in millions. A histogram of the distribution of video game sales in the rest of the world is shown in Figure 8. A histogram showing the breakdown of total global sales is shown in Figure 9. Figures 10 and 11 depict the density plot that evaluates the distribution of critic score and critic count in the dataset. Figure 12 depicts the histogram distribution of user counts in this dataset, allowing us to see the frequency of various user count values.

2. Research Question Analysis:

Figure 13 is a bar plot showing the total global sales of video games around the world, sorted by genre answering our first research question. The bar graph illustrates that action, sports, and shooter are the top three genres in terms of global sales, with action being the highest selling genre. In terms of global sales, the least popular genres are adventure, strategy, and puzzle games.

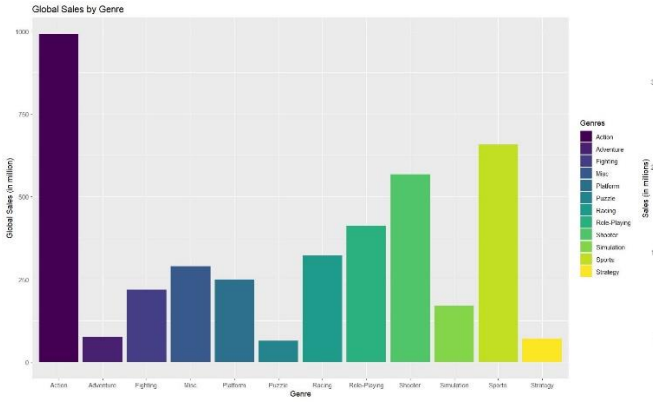


Figure 13

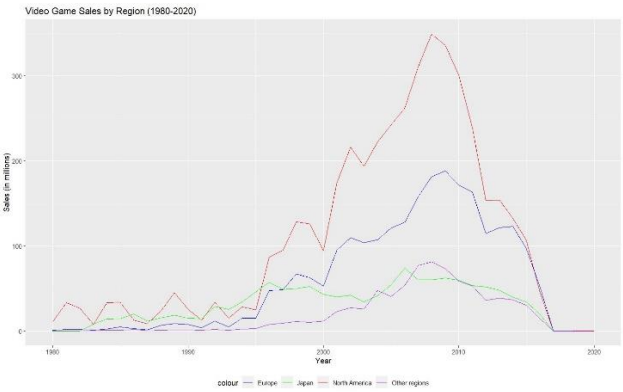


Figure 14

In order to address our second research question, Figure 14 shows a line plot of video game sales over time and estimates the total sales for each region (North America, Europe, Japan, and Other) as well as for the global market. The plot clearly shows that the graph has been increasing linearly for all the regions, despite a few ups and downs until 2008.

There has been an overall decline in video game sales till then, and there has been no major gain in sales since then that can make a significant change to the plot. A correlation diagram of total sales by region and genre is shown in Figure 15. The correlation plot reveals that the overall sales in all regions are highest for the action, sports, and shooter genres, while they are lowest for the adventure and puzzle genres. The region with the highest total sales for all genres is North America, whereas the region with the lowest is Japan. The role-playing genre has a very strong performance Japan, with total sales exceeding those of any other genre in that region. The Other sales category has the lowest total sales across all genres and regions, indicating that it is most likely an insignificant market for video games. There is a positive correlation between total sales in each region and genre, implying that video games with higher sales in one region are likely to have greater sales in other regions as well.

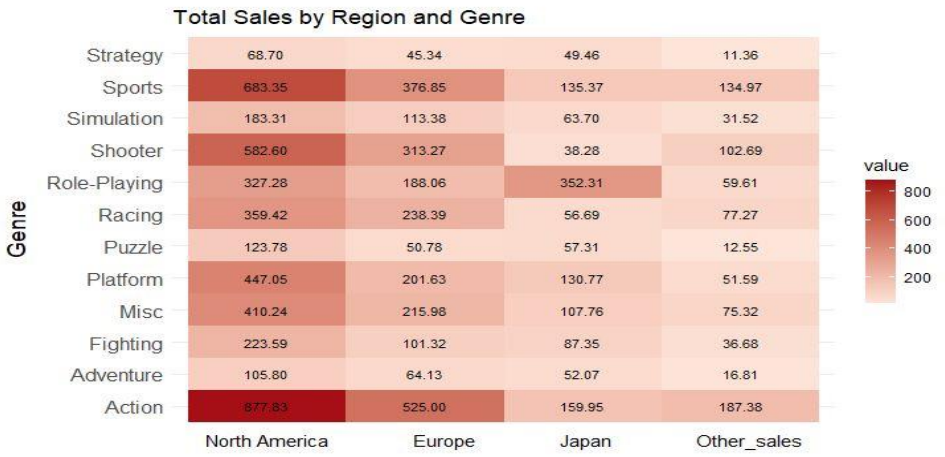


Figure 15

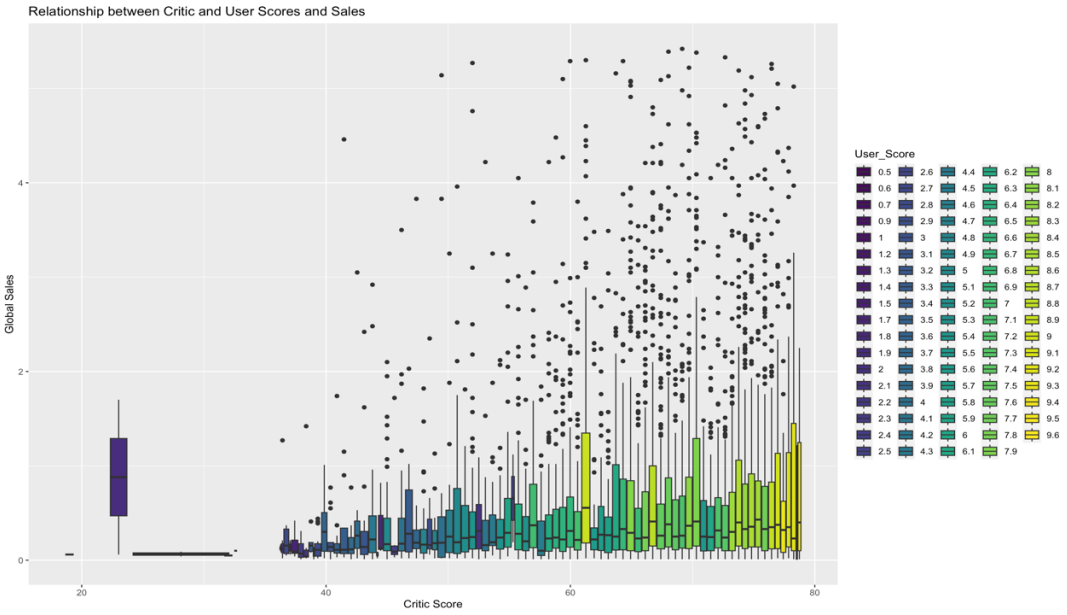


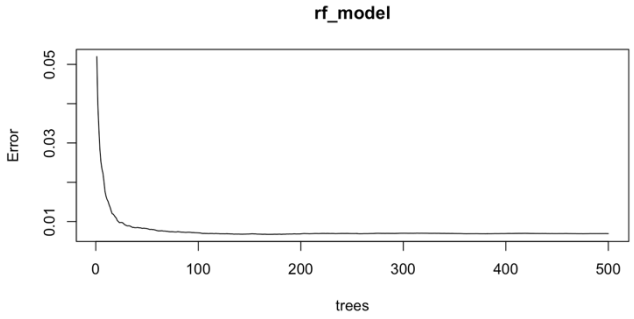
Figure 16

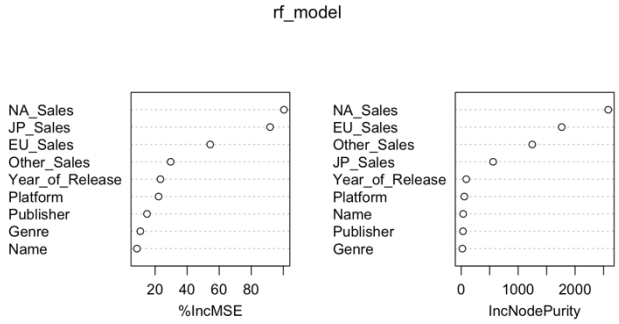
Figure 16 is a boxplot that visualizes the relationship between global sales, critic scores, and user scores of video games in response to our third research question. The plot helps us in understanding the relationship between the three variables and identifying any potential patterns or trends that exist.

B.-MODELS:

<u>RANDOM-FOREST:</u>

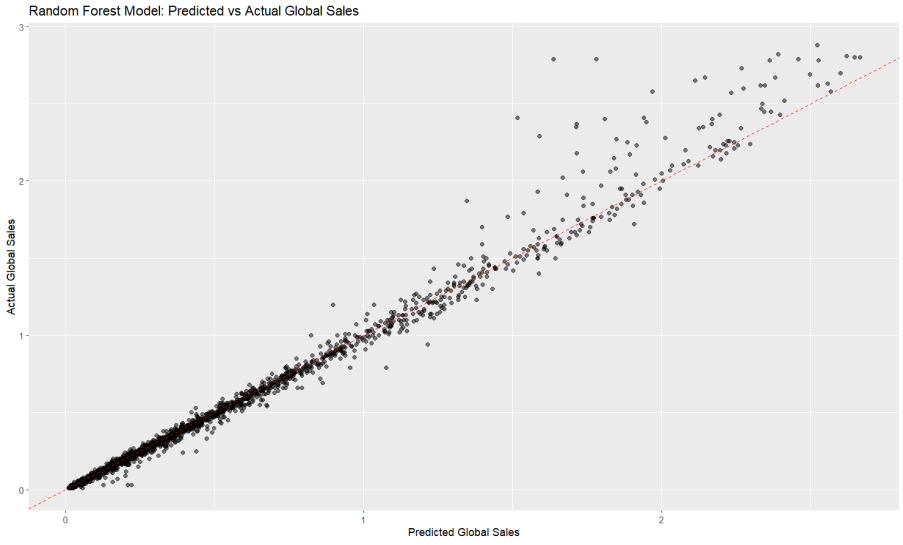
We have applied the random forest model [5] to predict the global sales of video games based on the factors like the game's name, platform, year of release, genre, and publisher. in order to increase the accuracy, we eliminated the outliers. we split the dataset into training and testing sets and then trained the random forest model using 500 trees. As we are working on a continuous target variable, The root mean squared error (RMSE) metric was used by us to assess the model's performance, it was 0.09848887 which is basically a low score indicating that the model has good predictive power.






```
RMSE: 0.09848887
```

To dig deeper and examine the model we have we plotted the variable, which revealed the most important features for predicting global sales. In order to see how effectively the model is, we have produced a scatter plot comparing the actual and predicted global sales. The plot displays a striking positive correlation between the two, with the majority of the points falling along the line of ideal prediction.



The random forest model was woking perfectly for predicting global video game sales based on a range of game features. The model proved to be effective in predicting the global sales of video games.

MULTIPLE LINEAR REGRESSION MODEL:

We also Multiple linear regression which is a statistical method that models the linear relationship between a dependent variable and multiple independent variables. In other words, it enables us to predict a dependent variable's value from two or more independent variables.

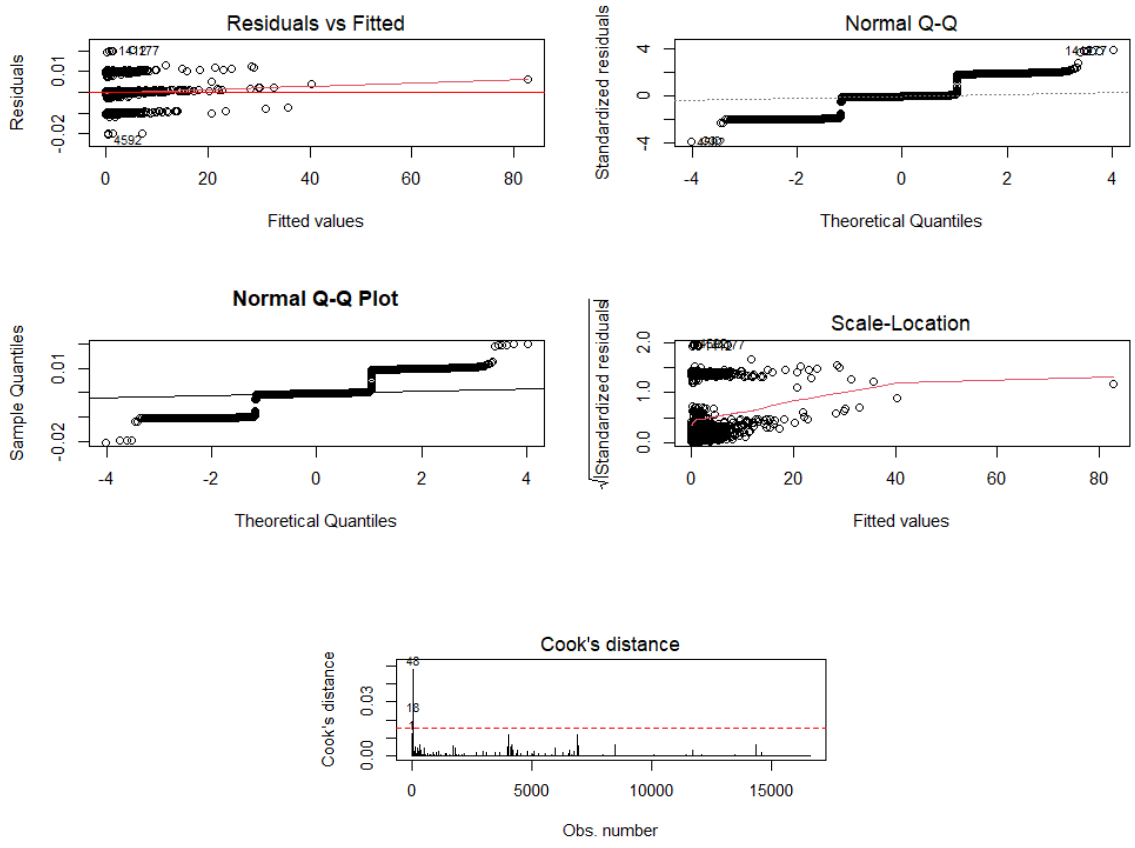


Figure 17

Figure 17 shows the plots of a multiple linear regression model performed on this video game sales dataset.

The Residuals versus Fitted plot above shows a highly random distribution of residuals around the horizontal line at y=0, showing that the linear regression assumption of constant variance is met. However, there is a little curvature in the residuals that can point to nonlinearity. The Normal Q-Q plot displays a reasonably linear pattern of residuals closely following the straight diagonal line, showing that the linear regression assumption of normally distributed residuals is met. However, there are a few deviations from the diagonal line towards the distribution's tails that could point to a very little non-normality. The Scale-Location plot demonstrates that the linear regression assumption of constant variance is satisfied since residuals are distributed fairly randomly and equally over the range of predicted values. Cook's distance is a measure of each observation's influence on the regression coefficients, and it can be utilized to identify influential observations that may have a big impact on the model.The overall appearance of this graph indicates that our model is a reasonable fit to the data, but there may be a few minor assumptions that need to be strengthened and there may be scope for improvement.

V CONCLUSION

Finally, this analysis provides an outline of the video game market from 1980 to 2020. The dataset contains data on the sales of video games across various geographies, subgenres, and platforms from which We found the most popular genres, platforms, and geographical locations through visualizations and analysis. Also, the relationship between video game sales and other elements like user and critic reviews has been looked at. It demonstrates the major shifts this industry went through over time.

Finally, our research shows that the random forest model is effective at forecasting video game sales across the globe. The model was successful in achieving a low root mean squared error, demonstrating its good predictive power, and it was able to pinpoint the key characteristics for forecasting global sales. A significant positive correlation was seen in the scatter plot comparison of the global sales, both actual and predicted. The report also highlights the shortcomings of linear regression in predicting video game sales, as well as the requirement to strengthen some of its underlying assumptions. An analysis like this can help game publishers and developers make decisions based on data and improve their marketing plans.

References

[1] R. KIRUBI, "Video Game Sales with Ratings," Kaggle, [Online]. Available: https://www.kaggle.com/datasets/rush4ratio/video-game-sales-with-ratings. [Accessed 01 May 2023].

[2] W. D'Angelo, "Video Game Charts, Game Sales, Top Sellers, Game Data," VGChartz, [Online]. Available: https://www.vgchartz.com/. [Accessed 05 May 2023].

[3] Arthurtok, "THE CONSOLE WARS: PS vs Xbox vs Wii," Kaggle, [Online]. Available: https://www.kaggle.com/code/arthurtok/the-console-wars-ps-vs-xbox-vs-wii. [Accessed 05 May 2023].

[4] H. Wickham, "Create Elegant Data Visualisations Using the Grammar of Graphics," ggplot2, 2016. [Online]. Available: https://ggplot2.tidyverse.org/. [Accessed 05 May 2023].

[5] R. p. b. A. L. a. M. W. Fortran original by Leo Breiman and Adele Cutler, "randomForest: Breiman and Cutler's Random Forests for Classification and Regression," 23 May 2022. [Online]. Available: https://cran.r-project.org/web/packages/randomForest/index.html. [Accessed 05 May 2023].