Final Project Part 3
Final Data Analysis Report
Predicting if a Video Game Will Experience High Sales

Contributions:

Names of Group Members	Contribution to Proposal
Mojan Majid	Steps for checking assumptions, fixing violations to get a new model, and methods section of the final report
Luis Rodrigo Nieto	Data cleaning, limiting the categories, fitting the preliminary model, fix violations to get new model, and methods section of the report
David Runesha	Discussion and ethics sections of the report
Jacob Rebelo	Introduction and results sections of the report

Introduction

The amount of sales a business is capable of generating is key to its financial success, understanding how to reach the desired sales target is imperative. In this report we aim to look at the video game industry to gain knowledge of key factors which result in increased sales. In our model we will use Global_Sales as our numerical response variable, which is the total number of sales world-wide of a particular video game. To study if any potential significant relationship with sales exists we will use 5 predictor variables that we believe should showcase a desired outcome of a positive linear relationship. Our first predictor variable will be the platform the game was developed and released on (Xbox, Playstation, etc), this variable is categorical. Second, User_score, this numerical variable is an overall score given to the game by users who play it. Research indicated that these variables were among 3 key factors that were associated with a game being a blockbuster game (Cox. J, 2014). Further research yielded similar sentiment, the platform has a large impact on total sales of any particular game (Babb, J., Terry, N., & Dana, K. 2013). The third variable is Critic_Score, this numerical variable indicates a score which professional reviews have attached to any particular game. More research suggested that critic reviews predicted sales more than any other factor, hence illustrating these variables' importance in our model (Sherrick, B., Schmierbach, M. 2016). We were unable to find any research which combined more than 3 variables of interest; therefore, 2 additional variables were selected to enhance context and build a more robust model, Year of Release and Genre. About Genre, different styles of video games have different target audiences which could lead to changes in sales. Different years could also impact the sales of a video based on external social trends; for example games released between the years of 2020-2021 may experience high sales as lockdowns from the covid pandemic forced many to remain indoors and find new ways to entertain themselves. The style and year of release provide additional context and may help eliminate certain games which may be seen as outliers due to external context that the articles researched do not provide. After completing our model we look to examine the linear equation which is produced, specifically the beta coefficients for each predictor variable. With the intercept, they will allow us to understand which factors are potentially associated with producing increased sales and which characteristics of a video game may hinder its performance in the global market.

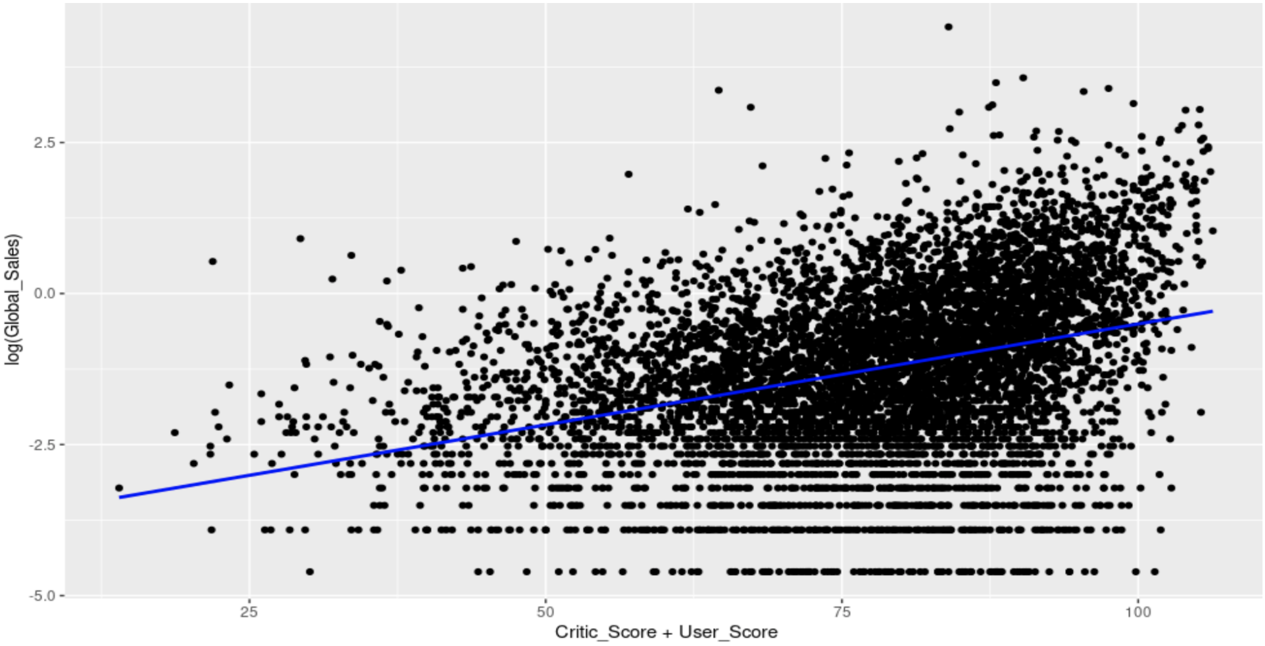
Methods

First of all —after determining our research question and cleaning the data of the dataset— we fitted the preliminary model with the selected variables. We have five variables, three of them are numerical and two of them are categorical, and our response. After looking at the amount of data we were going to work with, we decided to limit certain categories from the categorical variables in order to get a clearer view of the plots that would determine if our model violates assumptions or not. So, the categorical variables were limited, and only the most popular ones were kept. We assessed and considered that 5 categories per variable would be enough to have clear plots, but still not lose enough data that would change the objective of our research. Also, the year of release was limited to be higher than 2000 because the data obtained before this year was really scarce and did not add any value to the overall analysis.

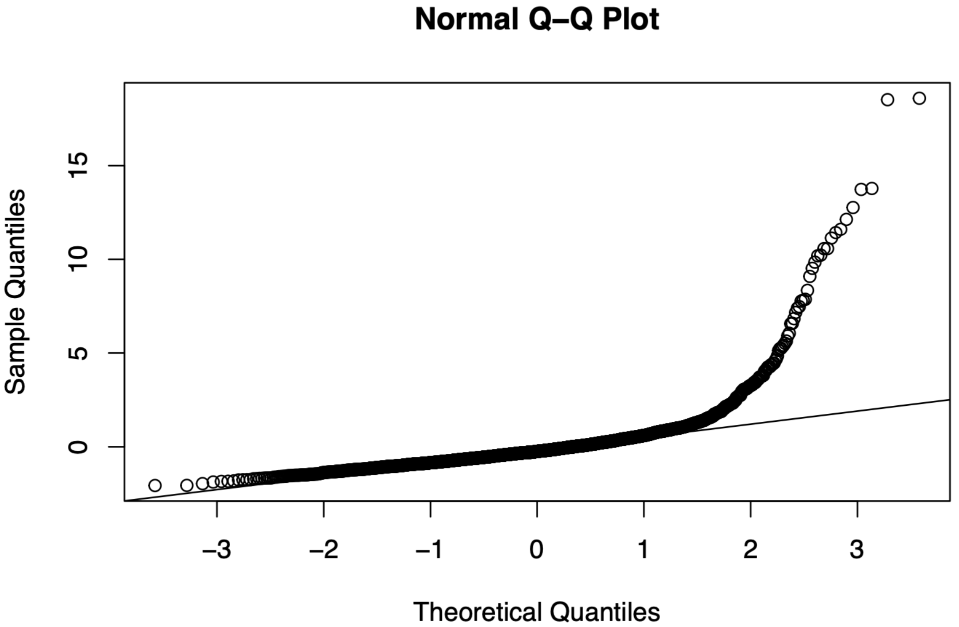
After fitting the preliminary model with the arranged dataset, checking the assumptions for violation was next. First we needed to check the additional conditions for multiple linear models. For the conditional mean response condition, we had to create a scatterplot of Response vs Fitted values and search for random diagonal scatter or an easily identifiable non-linear trend, which meant that the condition was satisfied. For the conditional mean predictor condition, all pairwise scatterplots of predictors were created and a lack of curves or other non-linear patterns were searched. After checking these conditions now the assumptions could be checked. For our numerical variables, a residual versus each predictor and a residual versus fitted values scatter plots were made to check for linearity, uncorrelated errors, and constant variance assumption violations. Regarding our categorical variables, a box plot for residuals versus each categorical predictor was made, but violations were clearly seen on the scatter plots. Finally, a normal quantile-quantile (QQ) plot was made to check for violations in the linearity assumption.

After analyzing the scatter and QQ plots for any violations, for every one, transformations were needed to be applied in order to correct it. For example: a Box-Cox transformation to response or predictors for linearity and normality violations, or a variance stabilizing transformation to response for constant variance violations. After applying the transformations, assumptions and conditions needed to be checked again until no more violations appeared. When everything checked out, an ANOVA test needed to be made to make sure a statistical linear relationship existed for at least one predictor. Our null hypothesis, which is that all slopes are zero, was tested by checking if the p-value given by the code was smaller than the alpha we chose. If the null hypothesis was rejected, a statistically significant linear relationship was established for at least one predictor, and a hypothesis test for individual coefficients in our model could be performed. When comparing values given by the R code to our alpha, we could claim if a significant linear relationship existed or not. If it did exist, now a partial F test needed to be made to compare our model to a more reduced model. A reduced model needs to be built without the coefficients that didn’t have a significant linear relationship. As all coefficients were significant in our model, we did not proceed with a partial F test, but if that had not been the case, we could compare the reduced model to our original full model and see which one is better. If the null hypothesis is rejected we conclude there exists a significant linear relationship between Y and at least one of its predictors; therefore, proving that the original model is better than the reduced one. If the null wasn’t rejected, we would need to start again from the very beginning but now work with the reduced model and see if this one does satisfy all of the conditions previously mentioned.

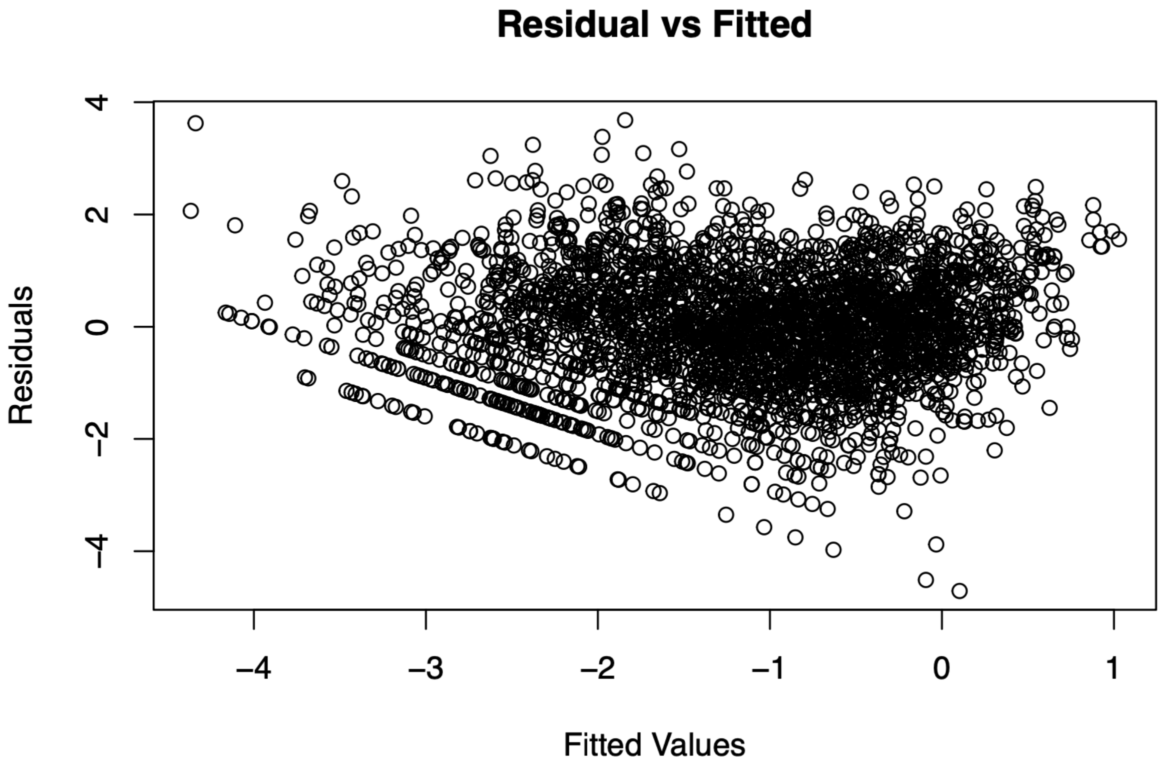
Results

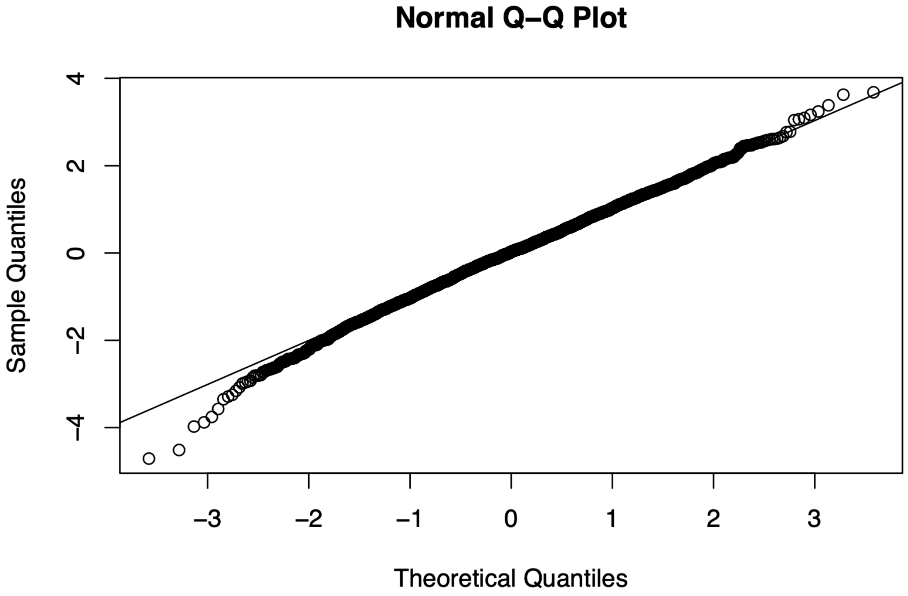


Above is a plot of our sample data with a preliminary simple model. To get an understanding as to whether our data would be useful in answering our research question about total sales, we compared the variables User score and Critic score to the percent change (log) of total global sales. To help illustrate a potential relationship we ran a simple linear regression model and highlighted it in blue. As we can see from our preliminary model we expected to see a positive linear relationship. We added the variables platform, year of release and genre to our model based on our external research and desire to understand what type of games tend to sell better. We began by running a multiple linear regression on our new model and found that each variable was statistically significant. There was potential to add other predictor variables in our model however they were broken down by region across the world, or included the developer or title of the game. These additional variables could hurt how easily the model can be interpreted; some games were more popular in certain regions, since we wished to view global sales as a whole, further breaking down sales by region would not have helped us in our analysis. Furthermore it would be very complicated to run a model having a variable with so many different categorical responses such as developer name or title and would have made the model too complex. It is for these reasons we choose our 5 predictor variables. To ensure our model does not violate any assumptions and cannot be improved we carried out the steps described in our methods section to arrive at a final model. One major violation we found our model to have was a violation of the normality assumption, this can be seen in the qq-plot below as many values vary from the derived diagonal line.



In order to correct this violation we performed a box-cox transformation to our model; after performing the transformation we found our 95% confidence interval on the maximum likelihood estimate was very close to 0, meaning our box-cox adjustment was reasonable. We proceeded by checking the residual and qq plots again for the model, this time we did not observe any clear violations of assumptions. The qq-plot, and residual vs fitted values scatterplot below show that the transformation has significantly improved previous assumption violations:





Without further violations we could begin testing for significance of our predictors, an ANOVA analysis was used to determine if our model was significant and at what level. In the table below we can see the output that was generated, we observe significant statistical significance for all of our predictors except for one being Genre2shooter. At this point we could have made further changes to the model and removed the predictor without significant significance, however after careful consideration we decided against removing the variable for the following reasons. The aim of our model is to provide potential game developers with information about what type of games will sell best relative to their competitors, removing a genre of game could affect the significance of other predictors in the genre category, their interpretation and the information provided by the model. In conclusion, the predictor Genre2shooter does not appear to be statistically significant; therefore we will claim that our model has a limitation in correctly predicting the projected sales of video games in the shooter genre. However for other styles of games we can be confident about its accuracy, as a whole the increased numbers of predictors makes for a more accurate model although interpretations will be more complex. As such our final model with predictor estimates along with other relevant information such as standard error, t-values and p - values are included in the table below.

	Estimate	**Std. Error**	**t-value**	**Pr(>|t|)**
Intercept	-49.639739	17.484269	-2.839	0.004555
Critic_Score	0.059145	0.001787	33.099	2e-16
User_Score	-0.132174	0.018119	-7.295	3.84e-13
Platform2PS2	2.259776	0.084532	26.733	2e-16
Platform2PS3	2.174448	0.067621	32.157	2e-16
Platform2X360	2.161901	0.067404	32.074	2e-16
Platform2XB	1.404005	0.091028	15.424	2e-16
Year_of_Release	0.021705	0.00869	2.498	0.012555
Genre2Racing	-0.300751	0.064419	-4.669	3.17e-06

Genre2Role	-0.233194	0.066504	-3.506	0.000461
Genre2Shooter	0.064563	0.054163	1.192	0.233357
Genre2Sports	-0.283759	0.056905	-4.987	6.51e-07

Discussion

The final model presented in this study aimed to predict video game sales using a combination of five predictor variables: platform, User_score, Critic_Score, Year of Release, and Genre. The model underwent a rigorous methodology, including data cleaning, preliminary model fitting, and assumption checks. The decision to limit certain categories to categorical variables was made to enhance clarity in data visualization without compromising the research objectives. Upon examining the results, the model demonstrated statistical significance for most predictor variables, except Genre2shooter. While this limitation is acknowledged, the final model, despite its limitation in predicting sales for shooter genre games, offers valuable insights for developers aiming to understand the dynamics of the video game market. The interpretability of the model may be complex due to the increased number of predictors, but it contributes to a more accurate representation of the factors influencing global sales. One notable improvement made to the model involved addressing a violation of the normality assumption through a Box-Cox transformation. This adjustment enhanced the model's robustness and ensured that the assumptions were met. In conclusion, the final model provides a valuable tool for game developers seeking information on the factors influencing video game sales. The model's accuracy is enhanced by considering a range of predictors, although it necessitates careful interpretation due to its complexity. Despite the limitations, the model significantly contributes to understanding the dynamics of video game sales in a global context.

Ethics

In this project, opting for manual variable selection was a reasonable and sound choice based on several factors inherent to the nature of the research. The decision to manually select variables stems from an understanding of the video game industry and the desire to incorporate domain-specific knowledge into the model-building process by using external research findings that identified key predictors associated with blockbuster games, such as platform and review scores. The complexity of the gaming market might not be fully encapsulated by automated selection methods alone. Ethically, the decision to choose manual selection over automated methods is justifiable in this context. The goal of the research is to offer valuable insights to potential game developers, and the inclusion of variables like Genre2shooter, despite its non-significance, aligns with the ethical imperative of providing comprehensive information. Automated methods, while efficient, might overlook certain specific nuances and potentially yield a model that lacks the depth required for practical application in the gaming industry. The ethical equivalence of the two methods largely depends on the goals and context of the research. In this case, manual selection was not chosen merely for practical reasons but was driven by a commitment to providing meaningful and relevant insights. The decision to forego certain automated methods is not reflective of negligence; rather, it reflects a conscientious choice to prioritize the richness and accuracy of the model over the expediency of automated processes.

References

Cox, J. (2014). "What Makes a Blockbuster Video Game? An Empirical Analysis of US Sales Data," *Managerial and Decision Economics, John Wiley & Sons, Ltd., vol. 35*(3), pages 189-198, April. https://doi.org/10.1002/mde.2608

Babb, J., Terry, N., & Dana, K. (2013). The Impact Of Platform On Global Video Game Sales. International Business & Economics Research Journal (IBER), 12(10), 1273–1288. https://doi.org/10.19030/iber.v12i10.8136

Sherrick, B., Schmierbach, M. (2016) The Effects of Evaluative Reviews on Market Success in the Video Game Industry. The Computer Games Journal 5, 185–194. https://doi.org/10.1007/s40869-016-0027-y