Introduction

There are a lot of factors that may influence the sales of a video game. Sure, personal interest plays a huge role, but it is entirely possible that there are more general factors. Some such factors include the games publisher, platform or genre. The goal of this study is to determine if the sales of a video game can be predicted based on some of these factors. Along with this, it aims to find out which factors (if any) drive sales the most in different regions.

To answer this question, a dataset of video game sales and relevant information will be used. Once the data has been cleaned up and analyzed, a linear and lasso regression will be performed to see how much some of these factors influence different regions' sales. Once the models have been fit, we will analyze the weights of each feature to determine its significance in overall sales. Lastly, a neural network will be used to verify the results of the regression.

Upon concluding the analysis, I found that the factors in the dataset only account for ~18% of the variance in the data, meaning there are likely other factors which play a large role in how well a game sells. These results were supported by the results of the neural network. In terms of significant factors by region, most countries were dominated by Nintendo sales, followed by varying consoles. However, Nintendo was second to SquareSoft in Japan, which seemed to be more influenced by different publishers than release platforms.

Methods

The dataset contained the following information:

- Name | Platform
- Release Year | Genre
- Publisher | Sales (NA, EU, JP, Other, Global)
- Critic/User Score | Critic/User Count
- Developer | Age Rating.

Of the above columns, only critic and user count were dropped, as much of the data for those columns was missing and I felt they wouldn't play a very significant role in the sales. Once the non-numeric features were encoded, I performed a linear/lasso regression. I elected to use a regression model for this study as I felt it would give insight into the most important features by analyzing its weights.

To ensure accurate results, I first performed a linear regression, and then a lasso regression. Once the linear regression was completed, I immediately performed a lasso regression, in hopes of reducing the number of features since there were so many. Using the results of the lasso regression, I was able to compare the weights of the different features to reach a conclusion about sales in different regions.

The results of both regression models were compared. To *further* verify the results, I fit a neural network with two hidden layers and five possible outputs, signifying the five regions. The results will be discussed in the following section.

Results

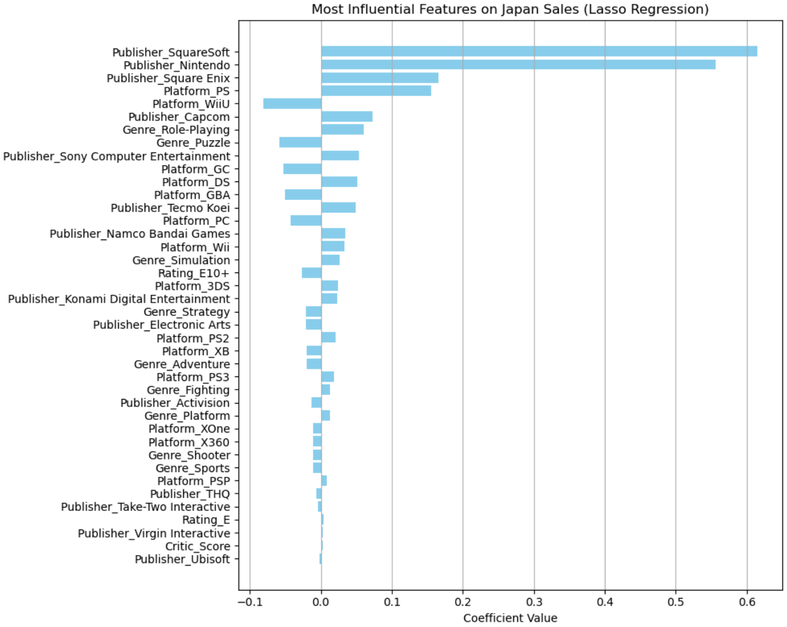
The following table shows the results of both the linear and lasso regression on each of the regions sales data:

	Linear R^2	Linear MSE	Lasso R^2	Lasso MSE
NA Sales	0.18406	0.58642	0.17072	0.59928
EU Sales	0.15987	0.25129	0.15315	0.25329
JP Sales	0.18531	0.03984	0.21536	0.03837
Other Sales	0.18463	0.03077	0.17766	0.03103
Global Sales	0.18725	1.97587	0.17687	2.0011

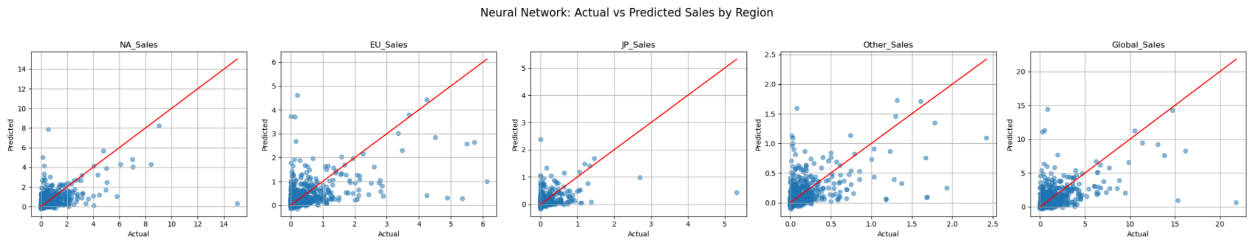
Note: The alpha values for the lasso regression was almost negligible.

As shown by the above table, the factors provided by the dataset don't account for much of the variance in sales data. That being said, some regions were predicted relatively accurately, such as Japan and Other Regions, while Global sales and NA sales had relatively high MSE. We also find that the lasso and linear regressions agree with one another for the most part, with little difference among their values.

From here, I plotted visualizations of the various features impact on each regions' sales. All of the graphs are included at the end of this report, but here is the
Japanese one as an example:



Finally, a neural network was used to verify the results of the regression. The results of the neural network are displayed in the following figure:



The results of these models give us a lot of insight into the questions originally asked.

Discussion

Now that we have our data, what does it show us? First of all, it shows that the information given to us was not enough to completely predict video game sales. This conclusion is supported both by the regressions and the neural network. This answers the first question, that there may be some outside factors that influence sales – maybe advertisement or proximity to holidays.

In my opinion, the more interesting question was also answered – which factors influence sales in different regions. The overall conclusion is that Nintendo produced games *tend* to do well, regardless of the regions, as the weight of those games was the highest in four out of the five. In each of these regions, specific platforms came in second and third. Usually, this console was the Wii, which aligns with Nintendo being on top. Variations came in what console followed it, which was the XBox 360 in NA and the PS4 in Europe.

The only region that was *not* dominated by Nintendo was surprisingly Japan, whose most influential factors were all publishers. In order, the strongest ones were SquareSoft, Nintendo, and Square Enix, in that order. While Nintendo is a frontrunner for the most popular, it's interesting that another publisher topped it in its home country.

Overall, this leads us to conclude that Nintendo games tend to sell well, regardless of the country, and that Japan sales are very influenced by the developer of the game, whereas platforms tend to be more influential around the world.

In the future, it may be interesting to explore datasets with more information, or a more recent dataset, as the one used for this experiment was from 2016, thus predating the Nintendo Switch (one of Nintendo's best selling consoles).

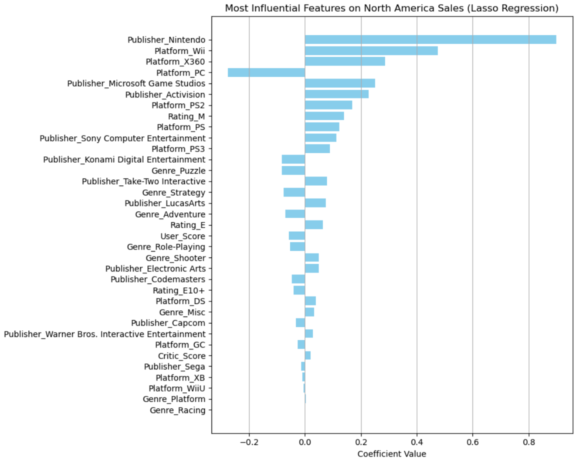
Citations

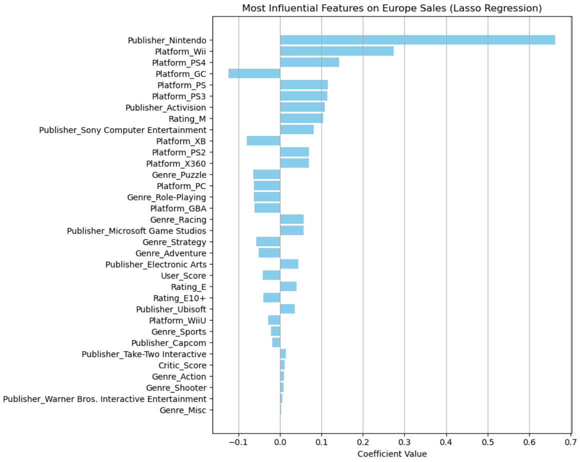
URL: https://www.kaggle.com/datasets/rush4ratio/video-game-sales-with-ratings

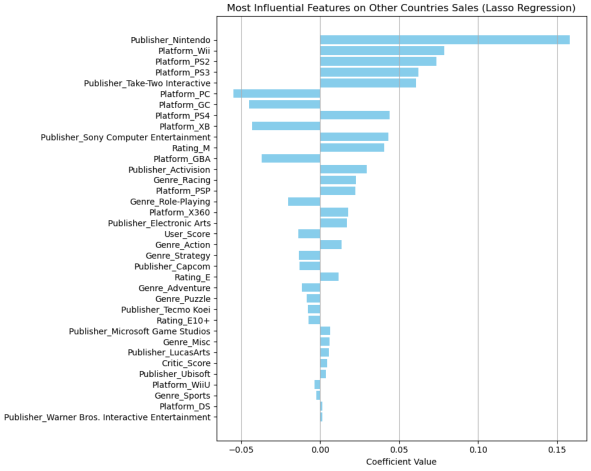
- Citation: Rush Kirubi. (2017). *Video Game Sales with Ratings*. Kaggle.

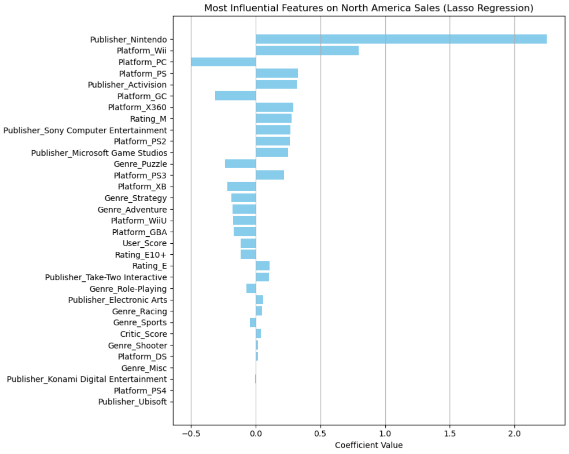
AI Usage: AI was used to help brainstorm a topic and find the dataset.

Additional Figures


Most Influential Features on North America Sales (Lasso Regression)
Publisher_Nintendo
Platform_Wii
Platform_X360
Platform_PC
Publisher_Microsoft Game Studios
Publisher_Activision
Platform_PS2
Rating_M
Platform_PS
Publisher_Sony Computer Entertainment
Platform_PS3
Publisher_Konami Digital Entertainment
Genre_Puzzle
Publisher_Take-Two Interactive
Genre_Strategy
Publisher_LucasArts
Genre_Adventure
Rating_E
User_Score
Genre_Role-Playing
Genre_Shooter
Publisher_Electronic Arts
Publisher_Codemasters
Rating_E10+
Platform_DS
Genre_Misc
Publisher_Capcom
Publisher_Warner Bros. Interactive Entertainment
Platform_GC
Critic_Score
Publisher_Sega
Platform_XB
Platform_WiiU
Genre_Platform
Genre_Racing
−0.2
0.0
0.2
0.4
0.6
0.8
Coefficient Value


Most Influential Features on Europe Sales (Lasso Regression)
Publisher_Nintendo
Platform_Wii
Platform_PS4
Platform_GC
Platform_PS
Platform_PS3
Publisher_Activision
Rating_M
Publisher_Sony Computer Entertainment
Platform_XB
Platform_PS2
Platform_X360
Genre_Puzzle
Platform_PC
Genre_Role-Playing
Platform_GBA
Genre_Racing
Publisher_Microsoft Game Studios
Genre_Strategy
Genre_Adventure
Publisher_Electronic Arts
User_Score
Rating_E
Rating_E10+
Publisher_Ubisoft
Platform_WiiU
Genre_Sports
Publisher_Capcom
Publisher_Take-Two Interactive
Critic_Score
Genre_Action
Genre_Shooter
Publisher_Warner Bros. Interactive Entertainment
Genre_Misc
−0.1
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Coefficient Value


Most Influential Features on Other Countries Sales (Lasso Regression)
Publisher_Nintendo
Platform_Wii
Platform_PS2
Platform_PS3
Publisher_Take-Two Interactive
Platform_PC
Platform_GC
Platform_PS4
Platform_XB
Publisher_Sony Computer Entertainment
Rating_M
Platform_GBA
Publisher_Activision
Genre_Racing
Platform_PSP
Genre_Role-Playing
Platform_X360
Publisher_Electronic Arts
User_Score
Genre_Action
Genre_Strategy
Publisher_Capcom
Rating_E
Genre_Adventure
Genre_Puzzle
Publisher_Tecmo Koei
Rating_E10+
Publisher_Microsoft Game Studios
Genre_Misc
Publisher_LucasArts
Critic_Score
Publisher_Ubisoft
Platform_WiiU
Genre_Sports
Platform_DS
Publisher_Warner Bros. Interactive Entertainment
−0.05
0.00
0.05
0.10
0.15
Coefficient Value


Most Influential Features on North America Sales (Lasso Regression)
Publisher_Nintendo
Platform_Wii
Platform_PC
Platform_PS
Publisher_Activision
Platform_GC
Platform_X360
Rating_M
Publisher_Sony Computer Entertainment
Platform_PS2
Publisher_Microsoft Game Studios
Genre_Puzzle
Platform_PS3
Platform_XB
Genre_Strategy
Genre_Adventure
Platform_WiiU
Platform_GBA
User_Score
Rating_E10+
Rating_E
Publisher_Take-Two Interactive
Genre_Role-Playing
Publisher_Electronic Arts
Genre_Racing
Genre_Sports
Critic_Score
Genre_Shooter
Platform_DS
Genre_Misc
Publisher_Konami Digital Entertainment
Platform_PS4
Publisher_Ubisoft
−0.5
0.0
0.5
1.0
1.5
2.0
Coefficient Value