Final Project Report

Christopher Hainzl, Christopher Barbieri, Christopher Hakkenberg,

Ryan Podzielny, Peter Bitanga

Introduction to Data Science DATA101

Dr. Tweneboah

December 16th, 2022

Table of Contents

Executive Summary

The data regarding the names of the video games, their sales, release platforms, and publisher companies were collected from Kaggle. All three of the video game datasets that we obtained from that website contained the names, sales, platforms, and companies associated with the video games, along with ratings and scores that were collected from Metacritic. It also includes the sales of the games outside of the United States, in countries such as Europe and Japan. However, for the sake of our analysis, we were only concerned with the global sales, user & critic scores that correspond with each of the games in our datasets. While conducting our data analysis, we wanted to see if there was a correlation between the user scores and critic scores of the games, as well as their global sales. We were able to determine that such a relationship does exist with the help of linear regression.

To put together our stock dataset, we looked at the high and low ends of each of the companies we analyzed, along with their total volume (in billions) and last trade. All this data was collected with the help of Market Watch and Investors Business Daily. With this information, we wanted to determine if a relationship exists between the global sales of video games and the stocks of the companies that published them. By combining the stock data with the video game sales data we collected, we were able to determine that this type of correlation exists as well. To get to this conclusion, we performed a linear regression analysis similar to our analysis of the video game sales dataset.

Overall, our analysis of the video game dataset in combination with the stock dataset can prove useful to people who are interested in investing in certain companies. It also provides insight into general information regarding video game sales and publications and can act as a stepping stone for people who are more interested in these topics.

Report

Initial Proposal: A Brief Look into our Data

1. Background of Data and Questions Raised

The Video Games Sales dataset contains names of video games, genres, platforms, sales numbers, ratings collected from Metacritic, and the names of the companies that developed the games. We collected our stock market data on Microsoft, Sony, Nintendo, and Electronic Arts from the New York Stock Exchange.

We analyzed the sales of games, user/critic scores, the relationship between user/critic scores, video game sales, and the relationship between video game sales and the value of the company's stock. Using the stock market with the video game dataset proved to be somewhat complicated, but with the proper procedure, we were able to combine it with the sales dataset. We also planned on determining whether or not video game sales influence the stock of the companies that sell them as well as whether or not critic/user scores influence the sales of video games.

2. Our Toolkit

The tools that were used in the data analysis were R-Markdown, Google Sheets, and the Investor.com website. We used R Markdown for tidying and organizing our data as well as to create several graphs and ultimately perform the mathematical analysis. This allowed for our analysis to be easily reproducible when provided with the same data. Investor.com and Google Sheets were utilized to obtain the stock market data from the NY Stock Exchange and record said data in a form that we could utilize, respectively. While analyzing the data, we applied our skills in exploratory data analysis, the creation of various types of graphs, and performing single and multiple linear regressions.

3. Brief EDA and Initial Findings

We began the analysis by importing the data as a data frame using the "read.csv" command. We then looked at the number of missing values in the various columns using the "is.na" and "colsums()" functions in combination. The "Critic_score", "Critic_Count Rating", and "user_score" columns had missing values, which we had to impute based on the rest of the values in the dataset. The "head()" command was used to view a small

portion of the data to get an idea of what was contained in the data frame. The "count()" function as well as a pipeline function (%>%) was used to determine the number of released games per listed year by counting the rows and organizing them by the "Year_of_Release" column. We then used ggplot to create several graphs including a line graph of the number of released titles per year, a histogram of the "Global_Sales" (to more easily visualize the distribution of the sales data), and Q-Q plots of Global Sales before and after the $\log_{10}$ transformation.

We used a line graph to visualize the distribution of the years of release, a histogram to visualize the global sales of the companies being analyzed, a multiple linear regression model to analyze the relationship between the companies' global sales and the scores that users and critics were giving their games, and a single linear regression model to analyze the relationship between the global sales of the games and the stocks of the companies that released them. [More information found in EDA section]

EDA and Data Wrangling - Video Game Dataset

Our first objective was to get a better understanding of our data, primarily what the data looked like, and what it meant for the models we wanted to use, mainly linear regression. Our first look was at the number of publications released per year, viewable with a line graph (*see Appendix Figure 1*). We essentially counted the number of times a year appeared in our dataset and graphed accordingly. What we mainly wanted to highlight was which years were most relevant in terms of activity from all the game publishers, to see what years would be most important to us in the stock dataset. It also proved useful in terms of global sales, and when publishers started getting into the video game industry. From here we also took a look at the total publications and total sales from each publisher and sorted them in descending order, to get a feel for which companies dominated in terms of these aspects. The companies we mainly saw were Nintendo, Electronic Arts (EA), Sony, and Microsoft, which were in the top six in both categories of the total number of publications and total global sales.

These companies were what we focused on for the majority of our questions. Mainly because of how big these companies were, and how most people knew them. Specifically, Sony and Microsoft, as they are both the companies that participated in the "Console Wars, " meaning

they had large competition, which was interesting to us to explore in a more in-depth look. Once we picked out the companies we wanted to focus on, then came further exploratory data analysis on them. Mainly filtering out the global sales and the user/critic scores for these companies specifically (*see the R Markdown file for code*).

After filtering, we needed to get a visual sense of the data and to see if it was fit for the models we wanted to use. After using a histogram to see the frequency of the global sales, as well as a Q-Q plot to see linearity, it was clear that we would have to perform some sort of transformation. We saw that the histogram of the sales was heavily skewed to the right and that the Q-Q plot's points were not in the realm of the line. To continue with performing a linear regression on the global sales, doing a log10 transformation was necessary. This drastically improved the skewness, with the new histogram being of roughly normal shape, and the Q-Q plot's points were essentially all on the line - meaning linearity was dramatically improved. With this transformation, we felt comfortable with using a multiple linear regression model (*see Appendix Figures 2.1 and 2.2*).

With the log transformation, we also wanted to get more visuals to try and get a better understanding of the companies we were looking at. We did a density ridge plot (*see Appendix Figure 4*) as well as a box plot of the sales and found some relatively interesting information on the companies as a whole. The main takeaway from the density ridge plot was that the shape and distribution of each company's sales were roughly the same. The box plot (*see Appendix Figure 3*) shows similar information (as to be expected), with the notable expectation of giving us a brief glimpse at the summary statistics, and some information on the outliers - of which there weren't many. We also used a correlation matrix to see if there was a relationship between these variables. Which there was, being relatively moderate, but enough for us to continue our exploration and continue to seek out multiple linear regression.

Once we were comfortable with our results from the EDA on global sales and with each company, we had to address the missing values found in critic and user scores. Since we discovered such a large proportion of missing data, around one-third from each column, we needed to impute them to continue with our analysis. From further EDA of looking at Q-Q plots and histograms of both the user and critic scores, we noticed several issues, specifically for user

scores. This column had a moderate left skew and the point of the Q-Q plot did not fall in line, so we decided to input with the median as it is a robust statistic. The critic scores by contrast had a roughly normal shape, so we felt comfortable imputing with the mean to continue with our analysis. From here we directly used multiple learning regression.

After completing our model for the critic and user scores against the global sales, our next focus was on the data wrangling for the stock market merged with video game sales. Our objective here was to get the total amount of global sales per year for each company. We did this by filtering for each company at a time, then filtering by year and using the sum() function on the global sales and inserting it into a new data frame. Afterwards, we performed a simple inner join by the company name to get a new data frame that was used for our simple linear regression - for both the stock market data and video game sales (*see R-Markdown file for code*). With our EDA and data wrangling completed, we felt very confident to use linear regression to try and determine a relationship.

EDA and Data Wrangling - Stocks Dataset

Our main hypothesis for the stock analysis was to figure out if the sales of video games had an inverse or a proportional relationship with the stock prices. Our initial results for the stocks started with a bunch of raw data within the exchange which was not limited to the following categories: high and low prices, volume, and date. Finally, we took a look at the last category, which is the last price recorded before the end of that day. After collecting all this information, we then performed an inner join between our stock and game sale datasets to ensure that it would be ready for our second linear regression. Within the inner join and the first and second linear regressions like mentioned before, we could perform a simple linear regression that could determine a relationship between video game sales and the stock price.

Models and Results

Our models consist of two explorations. One is the multiple linear regression with critic and user scores against global sales, and the other is a simple linear regression with global sales against stock market prices. Our original hypothesis was that both the user score and critic score would have a significant impact on Global Sales. A heightened user score and/or critic score will likely result in higher global sales. For our hypothesis testing, H_0: There is no relationship

between user scores and critic scores on global sales. H_a: There is some relationship between user scores and critic scores on global sales.

In regards to the critic and user scores, our approach was very simple. All our data was tidy and wrangled at this point, so we simply used R's lm() function to create the linear regressions. From our multiple linear regression, we were able to determine that global sales had a relationship with critic scores, but not user scores. Specifically, we found that the critic scores had a statistically significant correlation between global sales, while user scores did not - the p-value for critic scores was way below .05. The coefficient estimate was also positive, meaning if you have a higher critic score you'd see an increase in total global sales. This means we reject the null hypothesis H_0 and accept the alternative hypothesis H_a. Also for this linear regression, we did assumption checking to see if we could "trust" our model. With the variance inflation factor being within the range of 1-5, being 1.280917, it indicates a moderate correlation between the two variables. With the partial residual plots, we saw it follow along the line, so we know linearity was not violated. And with the Q-Q plot, almost all the points fell along the line of reference so we can assume normality (*see R-Markdown file for code)*.

For the simple linear regression with the stock market data and the video game sales, our initial hypothesis was that there was a statistically significant relationship between global sales per year and the yearly change in the company's stock market value. A heightened global sales would likely result in a large positive change in stock market value. For our hypothesis testing, H_0: There is no relationship between global sales and stock price. H_a: There is some relationship between global sales and stock price.

Like the critic and user scores against the global sales, we used R's lm() function to create our linear regression model. In our model we did find that there was a statistically significant correlation between these two variables, with a p-value much below 0.05 - so we reject the null hypothesis H_0 and accept the alternative hypothesis H_a. Though the coefficient of global sales ended up being negative, which was unexpected - we will discuss this in the External Factors section. (*see R-Markdown file for code)*

Potential Sources of Error and External Factors

Our data set contained some erroneous data inputs. For example, the "Imagine: Makeup Artist" game has an incorrect release date. The release date listed is 2020, however, the game was released in 2009. We discovered this because the data set was last updated in December 2016, thus having a release date of 2020 is impossible, any game with a release date past 2016 is an erroneous data point. In addition, "Imagine: Makeup Artist" is a Nintendo DS game, the Nintendo DS was obsolete by 2020, and games were no longer being released for that console by that year. In our stock price analysis, we only used video game sales to predict stock prices. In reality, many external factors affect the stock price. Most notably the 2008 recession that lasted several years probably had a negative effect on stock prices, independent of video game sales. In addition, companies we analyzed such as Sony are not solely dedicated to video games. Sony sells many products other than video games, such as televisions, audio/video equipment, and semiconductors. The same thing applies to Microsoft; not only do they make video games, but they also make computing devices such as laptops and tablets. The sale of these products may affect stock prices, stock by-backs and overall investor confidence may also have an effect.

There could also be some accuracy issues within the critic/user scores against the global sales. As a large portion of the data was missing, around one-third, we needed to impute with the median or mean. This could have slightly thrown off our results, but likely, this is not the case. We felt it was important to mention this as such a significant part of the data was missing.

Appendix

Code for our EDA, dating wrangling, models, etc. can be found within the R-Markdown file. This also includes the visuals we created to get a better understanding of what data we were working with. Below are some visuals we thought were most notable.

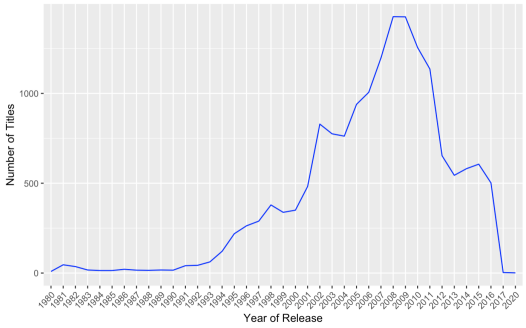

Figure 1 - Line graph of total publications per year

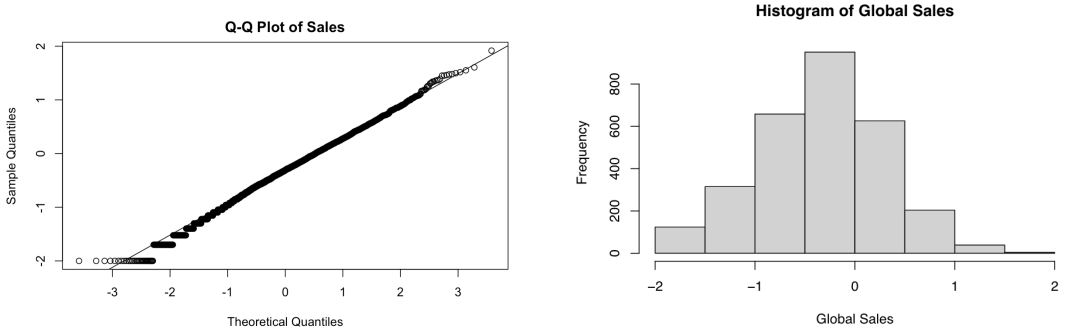

Figure 2.1 and 2.2 - Q-Q plot and Histogram of Log transformed sales

Figure 3 - Barplot of global sales by publisher

Figure 4 - Density Ridge plot of publishers against global sales