Analyzing Video Game Sales

Data 101 Final: Peter Bitanga, Ryan Podzielny, Christopher Hainzl, Christopher Hakkenberg, Christopher Barbieri

Background



Purpose of the Data Analysis

- We will be attempting to see a relationship in the Stock Market using Linear Regression Models and Sales Data from Video Game Companies
- This type of analysis is fairly practical in modeling for a company's success or failure



Data

- Using "Video Games Sales Dataset" from kaggle data base
- Data current as of December 2016
- Contains: Name of Game, Platform, Year Released,
 Global Sales (in millions), Developer, Rating
- About 6,900 complete cases
- Stock data, 2004-2016 of Microsoft, EA, Nintendo,
 Sony

	Name <dr></dr>	Platform <dr>></dr>	Year_of_Release <chr></chr>	Genre <chr></chr>	Publisher <chr></chr>	NA_Sales <dbl></dbl>
1	Wii Sports	Wii	2006	Sports	Nintendo	41.36
2	Super Mario Bros.	NES	1985	Platform	Nintendo	29.08
3	Mario Kart Wii	Wii	2008	Racing	Nintendo	15.68
4	Wii Sports Resort	Wii	2009	Sports	Nintendo	15.61
5	Pokemon Red/Pokemon Blue	GB	1996	Role-Playing	Nintendo	11.27
6	Tetris	GB	1989	Puzzle	Nintendo	23.20

Our Data Analysis Approach

- Explore data
 - o "glimpse" to view basic information
 - "head" to get a sense of what dataset looks like
 - "is.na" to add up missing values by column

```
glimpse(Video_Games)
                                                                                                                                                                                                                                                                               .
  Rows: 16.719
 Columns: 16
   $ Name
                                                         <chr> "Wii Sports", "Super Mario Bros.", "Mario Kart Wii", "Wii Sports .
                                                         <chr> "Wii", "NES", "Wii", "Wii", "GB", "GB", "DS", "Wii", "Wii", "NES"...
  $ Platform
  $ Year_of_Release <chr> "2006", "1985", "2008", "2009", "1996", "1989", "2006", "2006",
                                                         <chr> "Sports", "Platform", "Racing", "Sports", "Role-Playing", "Puzzle...
  $ Genre
                                                         "Nintendo", "Nintend
  $ Publisher
                                                         <db7> 41.36, 29.08, 15.68, 15.61, 11.27, 23.20, 11.28, 13.96, 14.44, 26...
  $ NA Sales
                                                        <db1> 28.96, 3.58, 12.76, 10.93, 8.89, 2.26, 9.14, 9.18, 6.94, 0.63, 10.
 $ EU Sales
                                                         <db?> 3.77, 6.81, 3.79, 3.28, 10.22, 4.22, 6.50, 2.93, 4.70, 0.28, 1.93...
  $ JP_Sales
                                                         <db1> 8.45, 0.77, 3.29, 2.95, 1.00, 0.58, 2.88, 2.84, 2.24, 0.47, 2.74,...
 $ Other_Sales
  [r]
 colSums(is.na(Video_Games))
                                                                                   Platform Year of Release
                                                                                                                                                                                                                                                      Publisher
                                          Name
                                                                                                                                                                                                            Genre
              Global_Sales
                                                                                                                            Critic Count
                                                                      Critic_Score
                                                                                                                                                                                           User_Score
                                                                                                                                                                                                                                                   User_Count
```

8582

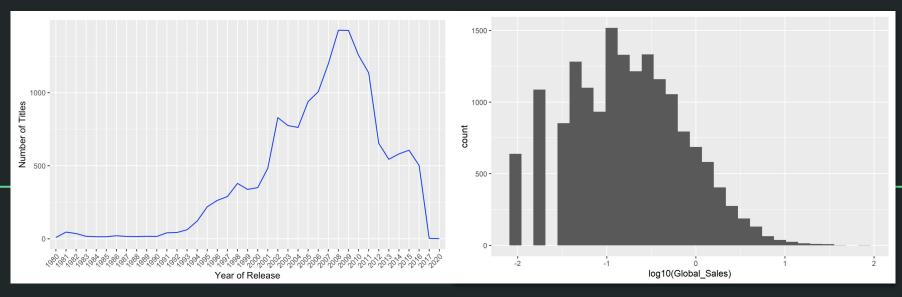
9129

9129

8582

Exploratory Data Analysis





Name <chr></chr>	Platform <chr></chr>	Year_of_Release <chr></chr>	Genre <chr></chr>	Publisher <chr></chr>	NA_Sales <dbl></dbl>	EU_Sales <dbl></dbl>
Imagine: Makeup Artist	DS	2020	Simulation	Ubisoft	0.27	0
1 row 1-7 of 16 columns						

```
# Basically get the highest sales
pubs_count <- table(sales$Publisher)
pubs_count <- as.data.frame(pubs_count)

colnames(pubs_count) <- c("pubs", "total_games")

pubs <- c(pubs_count$pubs)

sale_total <- c()

for (x in pubs) {
    currPub <- sales %>% filter(Publisher == x)
    total <- sum(currPub$Global_Sales)
    sale_total <- append(sale_total, total)
}

sale_numbers <- data.frame(pubs, sale_total)</pre>
```

```
# Total sales and games published by companies we are looking at
sale_numbers <- sale_numbers %>% arrange(desc(sale_total))
pubs_count <- pubs_count %>% arrange(desc(total_games))
```

head(pubs_count)

```
head(sale numbers)
##
                             pubs sale total
                         Nintendo
                                      1788.81
                  Electronic Arts
                                      1116.96
                       Activision
                                       731.16
                                       606.48
     Sony Computer Entertainment
## 5
                          Ubisoft
                                       471.61
## 6
            Take-Two Interactive
                                       403.82
```

```
##
                               pubs total games
                   Electronic Arts
                                            1356
## 1
                                            985
##
                        Activision
                Namco Bandai Games
                                            939
##
                                             933
##
                           Ubisoft
                                            834
     Konami Digital Entertainment
## 6
                                THO
                                             715
```

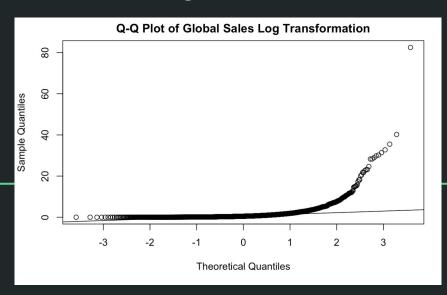
Seeing who the most popular publishers are through total sales and total games published

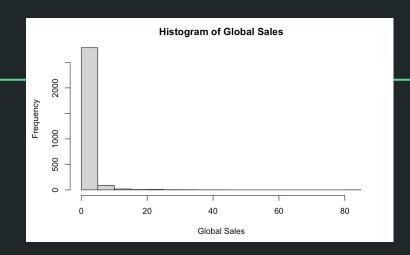
```
pubs_wanted = c("Nintendo", "Electronic Arts", "Microsoft Game Studios", "Sony Computer Entertainment")
looking <- sale_numbers %>% left_join(pubs_count, by="pubs")
looking <- looking %>% filter(pubs %in% pubs_wanted)
looking
```

##			pubs	sale_total	total_games
##	1		Nintendo	1788.81	706
##	2		Electronic Arts	1116.96	1356
##	3	Sony	Computer Entertainment	606.48	687
##	4		Microsoft Game Studios	248.32	191

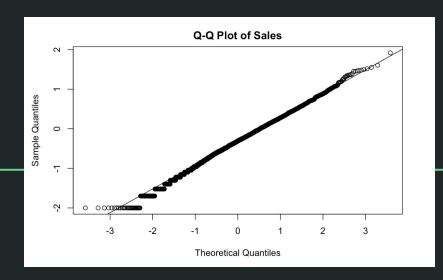
Most noteworthy publishers are Nintendo, EA, Sony and Microsoft

Without Log Transform

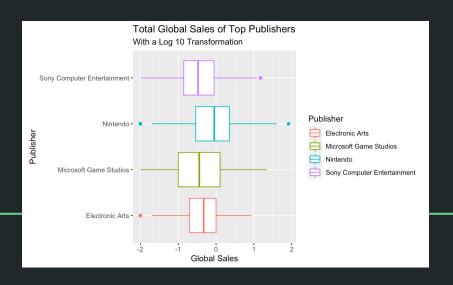


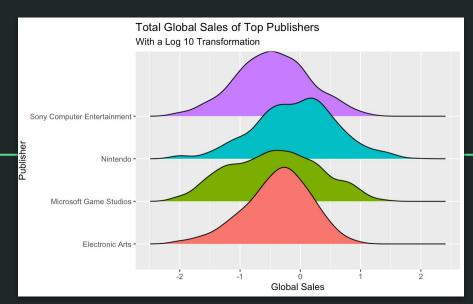


With Log Transform

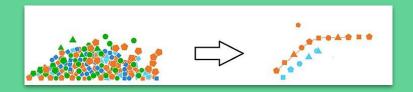








Data Wrangling



Data Wrangling - Sales Dataset

```
# Get the yearly sales for each game
new pubs <- sales %>% filter(Publisher %in% pubs wanted)
# Dropping the N/A (were normal NA's but rather strings)
new pubs <- new pubs %>% filter(Year of Release != "N/A")
# new pubs$Global Sales <- log10(new pubs$Global Sales)
years <- unique(new_pubs$Year_of_Release)
years <- sort(years, decreasing = FALSE)
size <- length(years) + 1
sales per year <- data.frame(matrix(ncol = size, nrow = 0))</pre>
colnames(sales per year) <- c("pubs", years)
# Basically for each company we want, we add the total sales for each year
for (x in pubs wanted) {
 # Get only one publisher
 currPub <- new pubs %>% filter(Publisher == x)
 # Pub needs to be the in the first column
 year tots <- c(x)
 for (year in years) {
   # Get the year we are on
   pub_year <- currPub %>% filter(Year_of_Release == year)
   # Add them all up and append to vector
   tot sales year <- sum(pub year$Global Sales)
   year tots <- append(year tots, tot sales year)
 # Add as row
 sales per year[nrow(sales per year) + 1.] = year tots
sales per year[,2:ncol(sales per year)] <- sapply(sales per year[,2:ncol(sales per year)], as.numeric)
sales_per_year
```

Getting total sales per year for each publisher, to align with stock market data

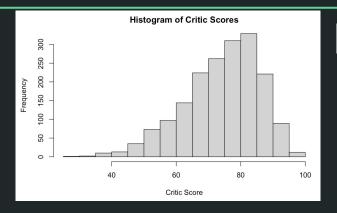
```
pubs 1983 1984 1985 1986 1987 1988
                      Nintendo 10.96 45.56 49.95 16.18 11.95 36.44 63.88 35.49
## 2
               Electronic Arts 0.00 0.00 0.00 0.00 0.00
         Microsoft Game Studios 0.00 0.00 0.00 0.00
                                                     0.00
## 4 Sony Computer Entertainment 0.00 0.00 0.00 0.00 0.00 0.00
                1993 1994 1995 1996 1997 1998 1999
## 1 15.97 38.11 20.04 24.99 16.72 73.70 25.80 48.41 65.33 34.05 45.37 48.31 38.14
          0.06 0.00 0.65 3.33 7.90 22.94 29.95 20.14 25.13 45.12 73.01 69.82
                0.00 0.00 0.00 5.12 0.00 0.00 0.09 0.99 13.49
          0.00 0.00 3.43 18.45 35.07 43.90 34.64 36.24 21.69 43.29 25.85 24.32
                         2007 2008
                                     2009 2010 2011 2012 2013 2014 2015
          126.69 204.42 103.09 90.01 127.39 59.74 51.69 56.11 53.48 48.88 27.61
          66.83 58.20 70.66 83.41 85.28 80.26 71.51 48.98 52.18 45.19 45.74
                  9.95 25.99 15.20 15.55 48.99 21.07 15.12 16.41 10.71 10.56
## 4 35.15 26.33 33.72 37.35 26.37 36.15 34.50 26.73 12.18 11.97 18.64 11.23
     2016
## 1 11.74
## 2 26.48
## 3 3.84
## 4 9.07
```

Data Wrangling - Sales Dataset

new_pubs\$User_Score[is.na(new_pubs\$User_Score)] <- median(new_pubs\$User_Score, na.rm=TRUE)

User score had a left skew, imputed with median





new_pubs\$Critic_Score[is.na(new_pubs\$Critic_Score)] <- mean(new_pubs\$Critic_Score, na.rm=TRUE)

Critic score had a more normal shape, imputed with mean

Data Wrangling (and EDA) - Sales Dataset

	new_pubs.Global_Sales	new_pubs.User_Score	new_pubs.Critic_Score
new_pubs.Global_Sales	1.00000000	0.07984132	0.2788627
new_pubs.User_Score	0.07984132	1.00000000	0.4227799
new_pubs.Critic_Score	0.27886275	0.42277990	1.0000000

Correlation matrix of the user and critic scores so see if there was a relationship.

From what we see, there is a small to moderate correlation between critic scores and global sales. Worthy for further exploration.

Data Gathering - Stock Market Dataset

End of year stock data, by Publisher

^	company	datestocks.	highstocks.	lowstocks.	total_volume [‡]	laststocks.
1	Microsoft(MSFT)	12/31/2004	27.440	26.6800	1830000000	26.720
2	Electronic Arts(EA)	12/31/2004	63.710	48.0500	140000000	61.680
3	Sony(SONY)	12/31/2004	39.200	35.7500	16500000000	38.960
4	Nintendo(NTDOY)	12/31/2004	3.170	2.8500	17300000000	3.170
5	Microsoft(MSFT)	12/30/2005	28.100	26.1000	1270000000	26.150
6	Electronic Arts(EA)	12/30/2005	58.230	52.1200	75524100	52.310
7	Sony(SONY)	12/30/2005	41.300	36.2000	10000000	40.800
8	Nintendo(NTDOY)	12/30/2005	3.060	2.7200	20888500	3.050
9	Microsoft(MSFT)	12/29/2006	30.260	28.8000	1140000000	29.860
10	Electronic Arts(EA)	12/29/2006	56.680	50.2100	77786700	50.360
11	Sony(SONY)	12/29/2006	43.780	39.2700	19630900	42.830
12	Nintendo(NTDOY)	12/29/2006	6.550	5.8200	12284500	6.500
13	Microsoft(MSFT)	12/31/2007	36.720	32.6300	1060000000	35.600
14	Electronic Arts(EA)	12/31/2007	60.350	53.9600	59614200	58.410
15	Sony(SONY)	12/31/2007	56.150	52.6100	16138500	54.300
16	Nintendo(NTDOY)	12/31/2007	15.680	13.6500	17405500	14.810
17	Microsoft(MSFT)	12/31/2008	21.250	18.4700	1550000000	19.440
18	Electronic Arts(EA)	12/31/2008	22.050	14.7900	204826200	16.040
19	Sony(SONY)	12/31/2008	21.890	18.0900	32184500	21.870

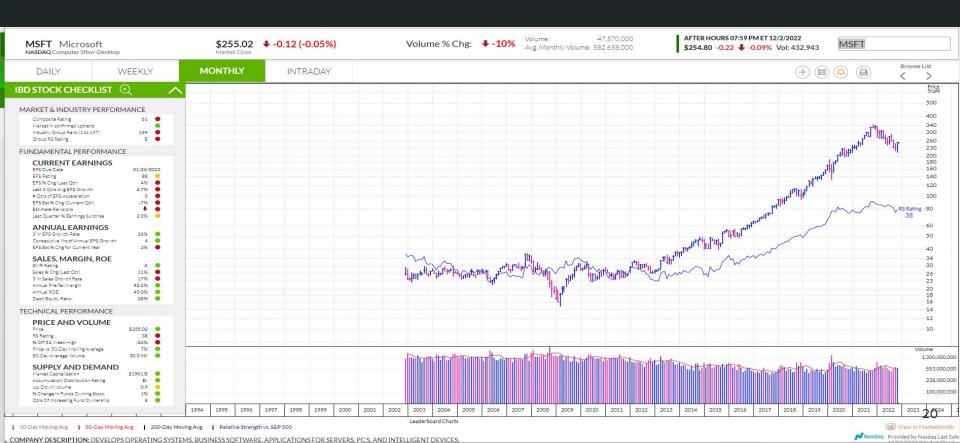
Data Wrangling - Stock Market Data (SONY)



Data Wrangling - Stock Market Data (NTDOY)



Data Wrangling - Stock Market Data (MSFT)



Data Wrangling - Stock Market Data (EA)



COMBANY DESCRIPTION: DEVELOPE VIDEO CAME SOCTIMADE AND CONTENT EAR CAME CONSOLES DOS HANDHELD DI ATECDMS AND MOBILE PHONES

Data Wrangling - Stocks and Sales

```
stocks <- read.csv("Stocks - DATA101 - Sheet1.csv")
stats_combined <- looking %>% left_join(sales, by = c("pubs" = "Publisher"))

{r}
stats_arranged <- stats_combined %>% arrange(Year_of_Release) %>% filter(Year_of_Release >= 2004)
stocks_with_stats <- stocks %>% inner_join(stats_arranged, by = c("company" = "pubs"))

{r}
stock_lm <- lm(change_.last. ~ sale_total, data = stocks_with_stats)
summary(stock_lm)|</pre>
```

Combining sales and stock data to prepare for linear regression

Our Models



MLR - Sales Hypothesis

We initially believed that both User Score and Critic Score will have a significant impact on the Global Sales. A heightened User Score and/or Critic Score will likely result in higher Global Sales.

H_o: 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

Multiple Linear Regression - Sales

```
2.5 % 97.5 % (Intercept) -1.79888331 -1.428911882 
User_Score -0.03350004 0.007913426 
Critic_Score 0.01658804 0.021795582
```

```
lin reg uc scores <- lm(Global Sales ~ User Score + Critic Score, data=new pubs)
summary(lin_reg_uc_scores)
##
## Call:
## lm(formula = Global Sales ~ User Score + Critic Score, data = new pubs)
## Residuals:
       Min
                 10 Median
## -2.07499 -0.30084 0.03558 0.35701 2.17428
## Coefficients:
                Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.613898 0.094318 -17.111 <2e-16 ***
## User Score -0.012793 0.010558 -1.212
                                           0.226
## Critic Score 0.019192 0.001328 14.456 <2e-16 ***
## Signif. codes: 0 '***' 0.001 '**' 0.05 '.' 0.1 ' ' 1
## Residual standard error: 0.5504 on 1777 degrees of freedom
   (1141 observations deleted due to missingness)
## Multiple R-squared: 0.1227, Adjusted R-squared: 0.1217
## F-statistic: 124.3 on 2 and 1777 DF, p-value: < 2.2e-16
```

MLR - Sales Hypothesis Test

There is a statistically significant relationship between Critic Scores and Global Sales, with its p-value way under .05.

This means we reject the null hypothesis, H₀, and accept the alternative hypothesis, H_a, as there is some relationship between the predictor and response.

SLR - Stocks & Sales Hypothesis

We initially believed that the Global Sales Total for the year will have a significant impact on the yearly change in Stock Market Value. A heightened Global Sales Total will result in a larger positive change in Stock Market Value will likely be seen.

H_o: There is no relationship between Global Sales and Stocks

H_a: There is some relationship between Global Sales and Stocks

Linear Regression - Stocks and Sales

```
stock lm <- lm(change .last. - sale total, data = stocks with stats)
summary(stock 1m)
##
## Call:
## lm(formula = change .last. - sale total, data = stocks with stats)
##
## Residuals:
      Min
               10 Median
                                      Max
## -43.780 -3.570
                    0.489
                            6.640
                                   22,665
##
## Coefficients:
                Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.4479796 0.3023355
                                     11.40 < 2e-16 ***
## sale total -0.0018242 0.0002321
                                     -7.86 4.03e-15 ***
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ! ' 1
## Residual standard error: 13.35 on 18862 degrees of freedom
     (1572 observations deleted due to missingness)
## Multiple R-squared: 0.003265, Adjusted R-squared: 0.003212
## F-statistic: 61.79 on 1 and 18862 DF, p-value: 4.032e-15
```

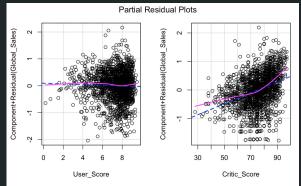
SLR - Sales Hypothesis Test

There is a statistically significant relationship between Global Sales and Stock market volume, with its p-value way under .05.

This means we reject the null hypothesis, H₀, and accept the alternative hypothesis, H_a, as there is some relationship between the Global Sales and Stocks.

Meaning behind the Results

- Extremely low p-value for Critic Score in predicting sales and for Total Sales in predicting change in stock price
- High p-value for User Score in predicting sales
- Low estimate values compared to intercept
- Negative estimate value for Total Sales in predicting change in stock price

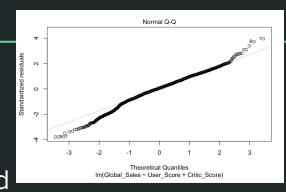


Linearity assumption is not violated

Assumption Checking - Sales

User_Score Critic_Score 1.280917 1.280917

Variance inflation factor indicates moderate correlation between user and critic scores in the model



Almost all points fall along line of reference, so we can assume normality

Potential Sources of Error

Name <chr></chr>	Platform <chr></chr>	Year_of_Release <chr></chr>	Genre <chr></chr>	Publisher <chr></chr>	NA_Sales <dbl></dbl>	EU_Sales <dbl></dbl>
Imagine: Makeup Artist	DS	2020	Simulation	Ubisoft	0.27	0
1 row 1-7 of 16 columns						

- Incorrect data entry
 - Actually released in 2009
- Missing values
- Imputing data can make things less accurate
- Not all companies sell only video games, stock prices can fluctuate from other factors

Citations

- SID_TWR. "Video Games Sales Dataset." Kaggle, 10 May 2019, <u>www.kaggle.com/datasets/sidtwr/videogames-sales-dataset?select=Videogames Sales as at 22 Dec 2016.csv</u>
- "Home Page." *MarketWatch*, www.marketwatch.com/.
- Investor's Business Daily. "Stock News & Stock Market Analysis IBD."
 Investor's Business Daily, www.investors.com/

Any Questions?