Action games and games with higher critic scores tend to have more sales*

https://github.com/Liuxuan-Wang/STA304-PS5

Liuxuan Wang

23 December 2020

Abstract

Video game, as an entertainment full of excitement and positive emotion, has attracted nearly 40% of the world population to be video game players. Obtained video game sales data from Kaggle with over 16,000 video games, this paper conducts multiple linear regression models to see how factors such as number of years since publish, critics and players' evaluation and type of game affect the global video game sales, and get the result that action games and games with higher critic scores tend to have more global sales. From this study, video game producers can have a better sense on what kind of games tend to have a better sale and can predict a certain video game's sale in the future.

Keywords: video games, global sales, multiple linear regression model

Introduction

Playing game is always an attractive activity for young and old. Because of the enjoyment and positive emotion brought by the games, the love for the games includes both inner psychological motivation and the outer positive rewards Pang (2019). With the development of technology, video games can provide players with more vivid visual effects, better auditory effects and various of game machanics. As the types of games booming, the population of players is also surging. The recent survey has revealed that nearly 3.1 billion people around the world are video game players, which takes up about 40% of the world population Price and writer (2020). In this large market, producers are pursuing to write popular games and players with various of demand are seeking more interesting games. For the producer side, it is important to have an idea on which types of games are more welcomed by the consumers and the expect sales of certain games that you have designed. Therefore, in this paper, I am going to use R as programming tool R Core Team (2020) and statistical method to analyze the factors that might affect the video games sales and build up models that can predict certain types of games' sale. The packages used are tidyverse Wickham et al. (2019), janitor Firke (2020) and pROC Robin et al. (2011)

The data set used in this paper is from Kaggle Kirubi (2016). It includes data about over 16,000 games published from 1980 to 2020. For each observation, the information of platform, year of publish, genre, publisher, sales in different area in the world, worldwide sales, and user scores are shown in the data. In this data, I will mainly focus on the worldwide sales to see how these factors will affect the total sales of the video game around the world. Overall, from the multiple linear regression we build, action games and games with higher critic scores tend to have more global sales.

The paper will be followed a full analysis of data. In the Data section, the data will be carefully described and evaluated. The model built up based on the data will be shown in the Model section. Some discussions results of the model, as well as the limitation and next steps will be shown in the following.

*https://github.com/Liuxuan-Wang/STA304-PS5

Data

This original data was obtained from vrchartz, with 16719 games released from 1980 to 2020. It includes the data about the games name, platform, year of release, publisher, sales in different areas, rating and user's and critic's score. The sales are in millions of unit. Critic's score are in 100 while users' score are in 10. The critic's score are from Metactrc's staff and user's scores are from Metacritic's subscriber

Data Features and Strengths

- Data Completeless:

With 16719 games of 12 different genres, the data set has collected almost all the games from 1980 to 2016. The variables included also cover most of the features that can be used to evaluate a game.

- Broad Time span:

It has recorded the video games across over thirty years, which can give data users' a view of video games' history using data.

Data Weakness

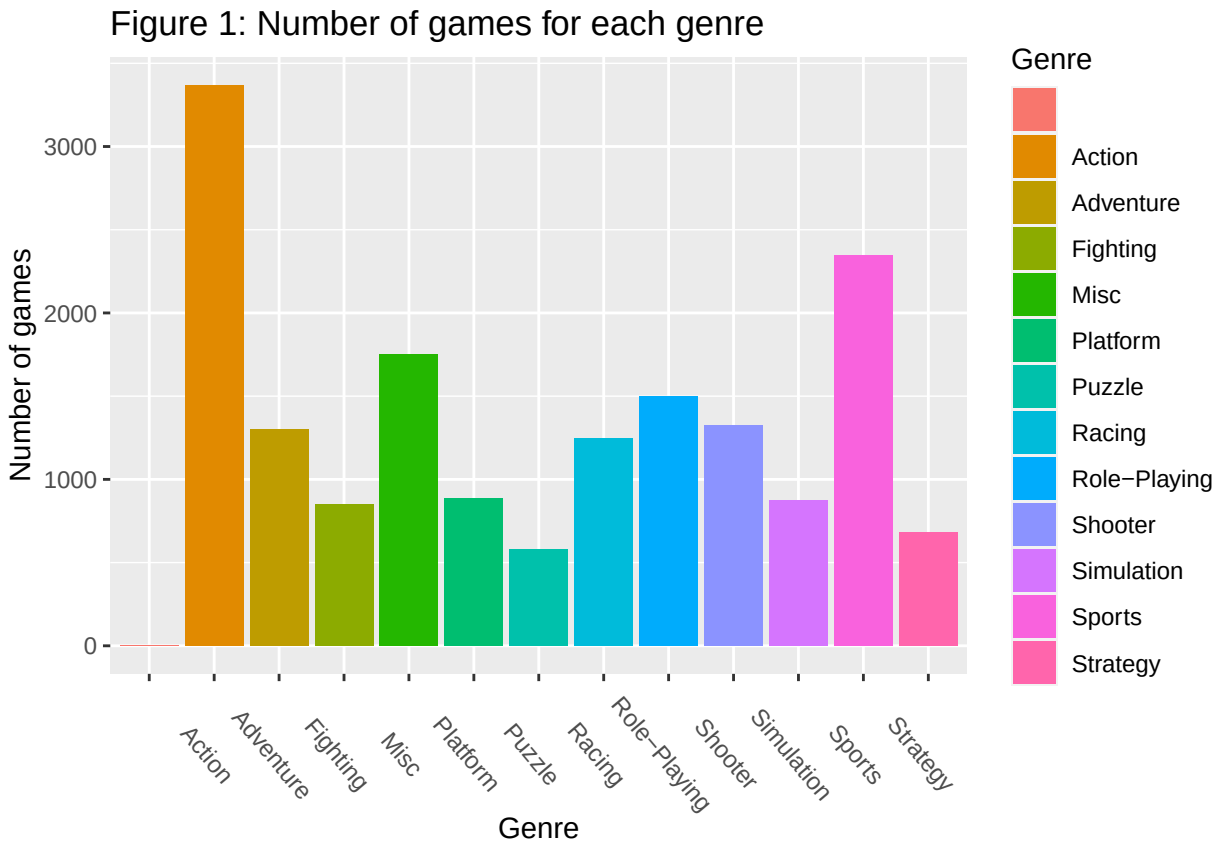
- Too many NA values:

Since there are a lot of old games and some of details of games are unavailable, therefore there are many NA values in this data set.

- The scores are from only one platform.

Though Metacritic is one of the largest platform for evaluating games and films, however, the scores collected here are only face to subscribers and staffs on this platform while cannot reflect what the other people think. There can be sampling error happening here.

Data Description

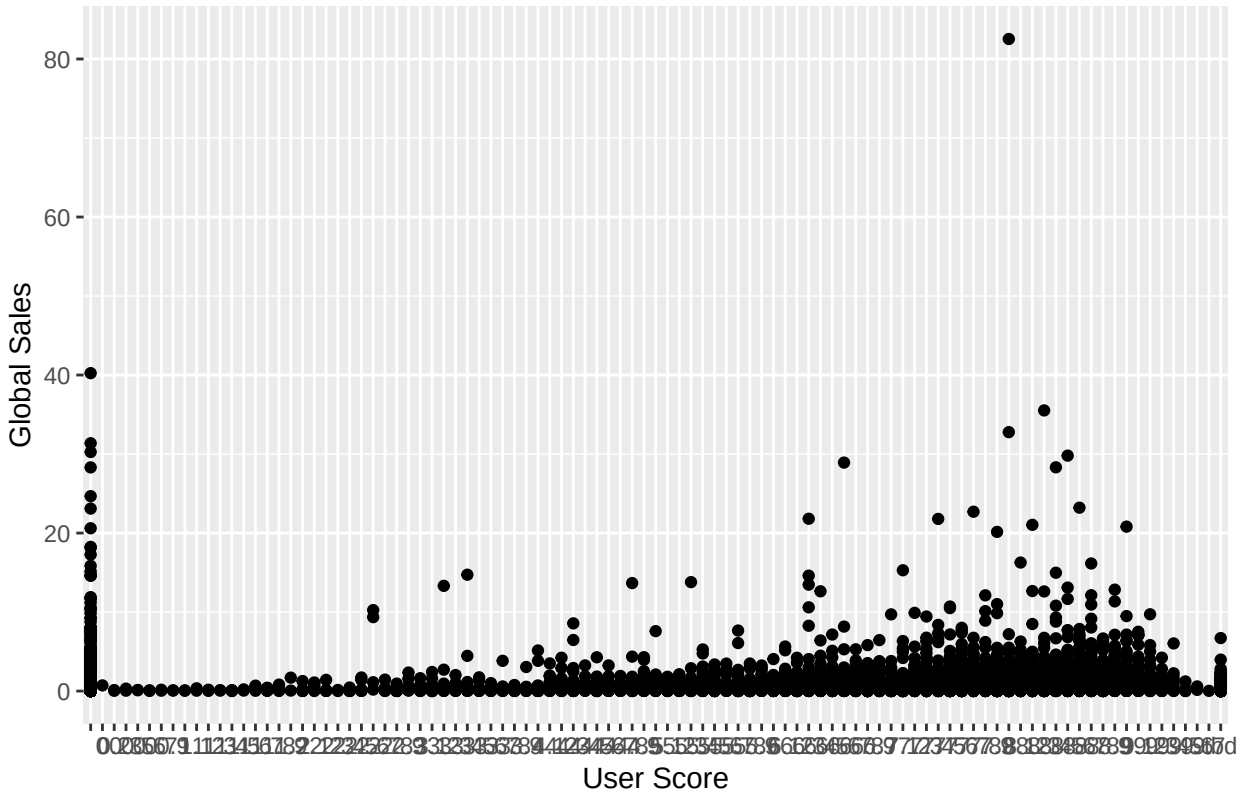


From 1980 to 2020, in this data set, action genre has the largest number of games released, with over 3500. Puzzle genre has only around 500 games released in this time period, which is the least.

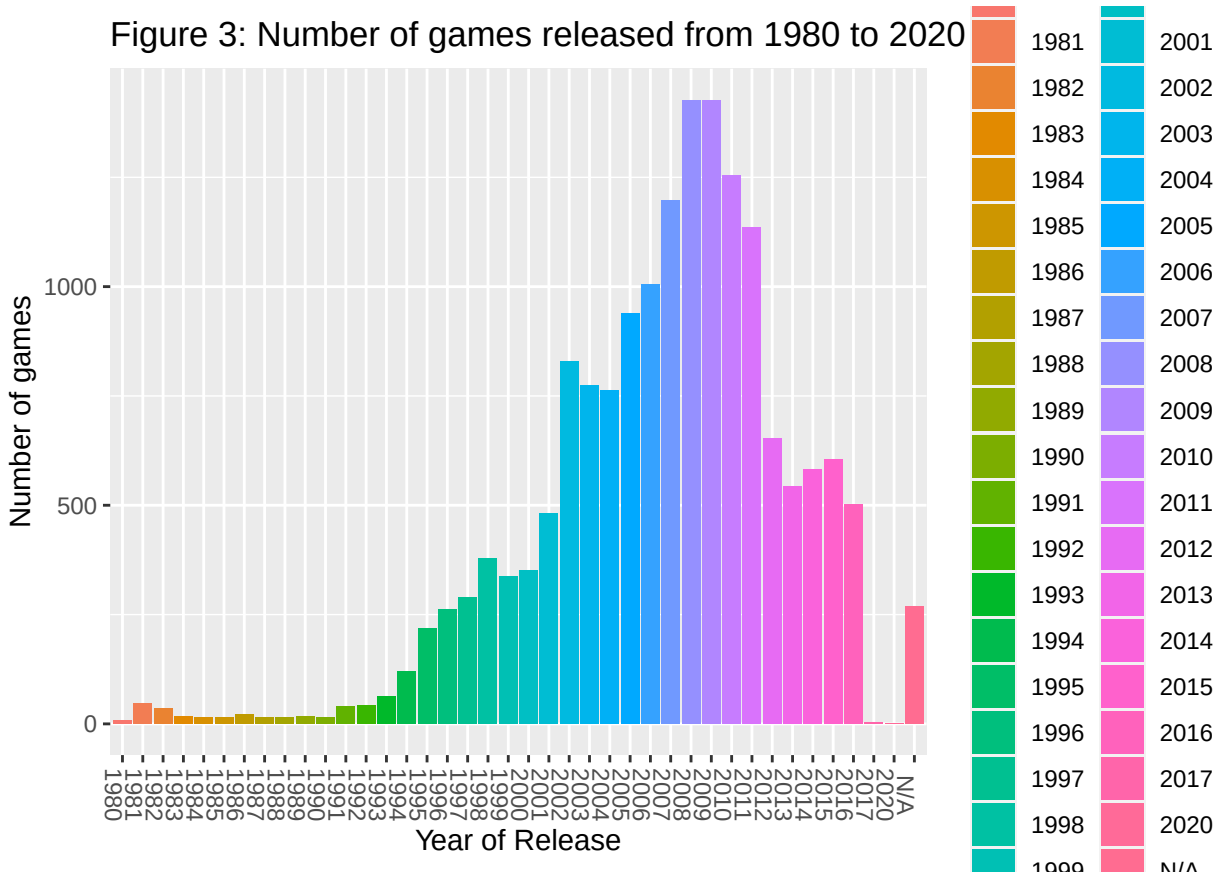
```
##                         Name Sales Publisher Year Player_score Critic_score
## 1                 Wii Sports 82.53  Nintendo 2006            8           76
## 2           Super Mario Bros. 40.24  Nintendo 1985                       NA
## 3              Mario Kart Wii 35.52  Nintendo 2008          8.3           82
## 4           Wii Sports Resort 32.77  Nintendo 2009            8           80
## 5   Pokemon Red/Pokemon Blue 31.37  Nintendo 1996                       NA
## 6                     Tetris 30.26  Nintendo 1989                       NA
## 7       New Super Mario Bros. 29.80  Nintendo 2006          8.5           89
## 8                    Wii Play 28.92  Nintendo 2006          6.6           58
## 9  New Super Mario Bros. Wii 28.32  Nintendo 2009          8.4           87
## 10                 Duck Hunt 28.31  Nintendo 1984                       NA
```

The top 10 games with the most sales are mostly released before 2010. All of them are from Nintendo and released before 2010. From here, we can see that these games have been popular for many years. Except for Wii Play and some old games without critic score records, the other games have good evaluation from both players and critics.

Figure 2: Scatter plot for global sales to user scores



The scatter plot above has no obvious linear pattern between these two variables. It shows that the global sales of video games are not simply affected by the users' evaluation. There are also a lot of other factors.




```
##    Year Frequency
## 29 2008      1427
## 30 2009      1426
## 31 2010      1255
## 28 2007      1197
## 32 2011      1136
## 27 2006      1006
## 26 2005       939
## 23 2002       829
## 24 2003       775
## 25 2004       762
## 33 2012       653
## 36 2015       606
## 35 2014       581
## 34 2013       544
## 37 2016       502
## 22 2001       482
## 19 1998       379
## 21 2000       350
## 20 1999       338
## 18 1997       289
## 40  N/A       269
## 17 1996       263
## 16 1995       219
## 15 1994       121
## 14 1993        62
```

```
## 2  1981        46
## 13 1992        43
## 12 1991        41
## 3  1982        36
## 7  1986        21
## 4  1983        17
## 10 1989        17
## 8  1987        16
## 11 1990        16
## 9  1988        15
## 5  1984        14
## 6  1985        14
## 1  1980         9
## 38 2017         3
## 39 2020         1
```

From the graph and frequency table above, in this data set, 2002-2011 have the most number of games released being recorded. There are 269 games which have unavailable year of release. In 1980s, the video game industry has just started, which leads to less games released. For games after 2016, this data set has its limitation on recording.

Model

I firstly clean the raw data. In this paper, I will mainly focus on building model predicting the global sales using year of release, genre, critic score and user score as explanatory variables. After selecting these variables and excluding NA values, using simple random selection to randomly select 1000 samples from the cleaned data. And build up a multiple regression model to see how these factors affect the global sales. The model's summary is shown below.

```
## Warning: NAs introduced by coercion
```

```
##
## Call:
## lm(formula = Global_Sales ~ ., data = SRSdata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.9264 -0.5924 -0.2241  0.2139 11.1542
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           2.462953   0.920162   2.677 0.007562 **
## Year_of_Release1997   2.572770   1.127489   2.282 0.022714 *
## Year_of_Release1998 -3.699695   1.004719  -3.682 0.000244 ***
## Year_of_Release1999 -3.804261   1.068935  -3.559 0.000390 ***
## Year_of_Release2000 -3.941566   0.934991  -4.216 2.72e-05 ***
## Year_of_Release2001 -3.351865   0.898082  -3.732 0.000201 ***
## Year_of_Release2002 -3.855151   0.886375  -4.349 1.51e-05 ***
## Year_of_Release2003 -3.749320   0.886618  -4.229 2.57e-05 ***
## Year_of_Release2004 -3.656148   0.887438  -4.120 4.12e-05 ***
## Year_of_Release2005 -3.910140   0.887176  -4.407 1.16e-05 ***
## Year_of_Release2006 -3.894499   0.886035  -4.395 1.23e-05 ***
## Year_of_Release2007 -3.739342   0.884490  -4.228 2.59e-05 ***
## Year_of_Release2008 -3.593241   0.882693  -4.071 5.07e-05 ***
```

```
## Year_of_Release2009 -3.819187   0.885167  -4.315 1.76e-05 ***
## Year_of_Release2010 -3.557041   0.887200  -4.009 6.56e-05 ***
## Year_of_Release2011 -3.429815   0.889478  -3.856 0.000123 ***
## Year_of_Release2012 -4.003999   0.894907  -4.474 8.58e-06 ***
## Year_of_Release2013 -4.039699   0.897865  -4.499 7.65e-06 ***
## Year_of_Release2014 -3.910451   0.894237  -4.373 1.36e-05 ***
## Year_of_Release2015 -3.540132   0.891354  -3.972 7.67e-05 ***
## Year_of_Release2016 -4.355701   0.903789  -4.819 1.67e-06 ***
## GenreAdventure      -0.330803   0.233033  -1.420 0.156060
## GenreFighting       -0.026169   0.190593  -0.137 0.890820
## GenreMisc            0.169198   0.181481   0.932 0.351406
## GenrePlatform        0.092306   0.186892   0.494 0.621488
## GenrePuzzle         -0.466466   0.311525  -1.497 0.134626
## GenreRacing          0.083241   0.154900   0.537 0.591126
## GenreRole-Playing   -0.167019   0.145772  -1.146 0.252179
## GenreShooter         0.061925   0.136397   0.454 0.649929
## GenreSimulation     -0.158294   0.205477  -0.770 0.441265
## GenreSports         -0.239212   0.134708  -1.776 0.076083 .
## GenreStrategy       -0.641281   0.228396  -2.808 0.005089 **
## Critic_Score         0.041340   0.003726  11.094  < 2e-16 ***
## Player_Score        -0.112621   0.037682  -2.989 0.002872 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.227 on 966 degrees of freedom
## Multiple R-squared:  0.2465, Adjusted R-squared:  0.2208
## F-statistic: 9.577 on 33 and 966 DF,  p-value: < 2.2e-16
```

Results

From the summary of multiple linear regression above, most of genre variables' coefficient are not statistically significant. For year of release, the year of 1996 is the base year, and for genre, action is the base variable. The coefficients for year of release show that the global sales will be more when the time of release becomes longer. The coefficients for genre just align with the situation we observe from the data part, that action games tend to have more sales. Critic score and player's evaluation both have siginificant affect on the global sales. Critic score has statistically significant positive relationship with globals sales. Surprisingly, the relationship between player's score and global sales is negative. Commonly, we think that the game with higher player evaluation should have more sales. This can be the effect of sampling.

Discussion

1) For some video games such as Mario and Pokemon, though they have been released for over 20 or 30 years, they are still populate and favorited by players from all over the world. Part of the reasons that why the coefficients for year of release are extremely close can be generated from these extremely popular old games which have larger sales growth than the other later games.

2) The circumstance shown in the coefficients of player score and critic score can be reasonable. Since players can be attracted by the high scores and pay for the game. However, player's evaluations mostly come out after they buy it. Therefore, the relationship between player's evaluation and the games' sales can be negative.

Limitation

- Sampling method

The sampling method used here is simple random selection, which can have some biases on sampling.

- Other factors

Except for the factors we used in this report, there are still other factors that can affect the sales. For example, for games on platforms that can expose to more players, the sales tend to be higher. Nintendo as large factory of video games starting in early years, tend to own more high sales games, as well.

Next steps

- Use more advanced sampling method:

Using stratified sampling method according to the platform or the genre of games can be a better choice with less biases than simple linear regression.

- Find more complete data from different platforms

Using data from different evaluation platforms can generate less bias on sampling error.

Reference

Firke, Sam. 2020. *Janitor: Simple Tools for Examining and Cleaning Dirty Data.* https://CRAN.R-project.org/package=janitor.

Kirubi, Rush. 2016. "Video Game Sales with Ratings." https://www.kaggle.com/rush4ratio/video-game-sales-with-ratings.

Pang, Chirlien. 2019. "Understanding Gamer Psychology: Why Do People Play Games?" *Sekg.* https://www.sekg.net/gamer-psychology-people-play-games/.

Price, Ben, and Ben Price(130 Articles Published)Ben is a writer. 2020. "New Report Shows Percentage of Global Population That Plays Video Games." *Game Rant.* https://gamerant.com/3-billion-gamers-report/.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Robin, Xavier, Natacha Turck, Alexandre Hainard, Natalia Tiberti, Frédérique Lisacek, Jean-Charles Sanchez, and Markus Müller. 2011. "PROC: An Open-Source Package for R and S+ to Analyze and Compare Roc Curves." *BMC Bioinformatics* 12: 77.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.