Lien Ly – Final Project
Data Analysis and Inference
Video Games Rating Vs. Platforms

I. Data source and definitions explained

The dataset was retrieved from IGN (Imagine Games Network – www.ign.com), where it provides articles, news, reports, rating, previews trailers for video games and other types of media. IGN users used scraping tool to retrieve data, then posted it on IGN forum.

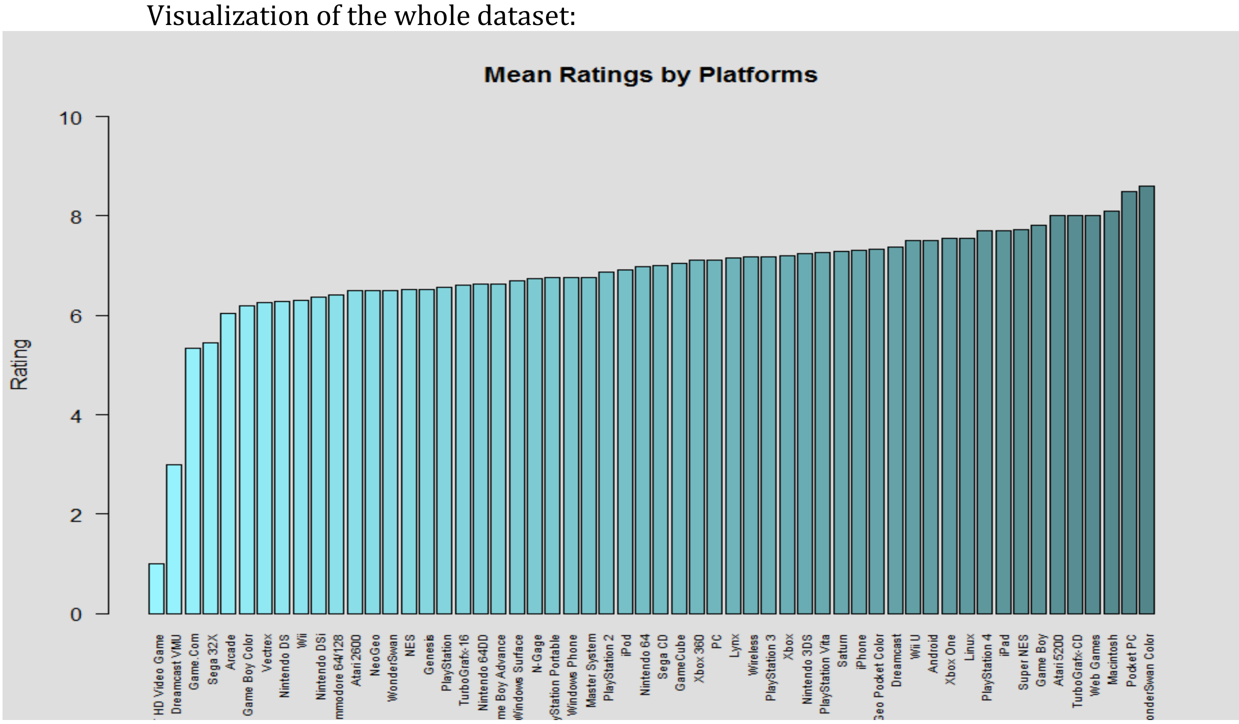
By using R, we got the summary of the data as follow:

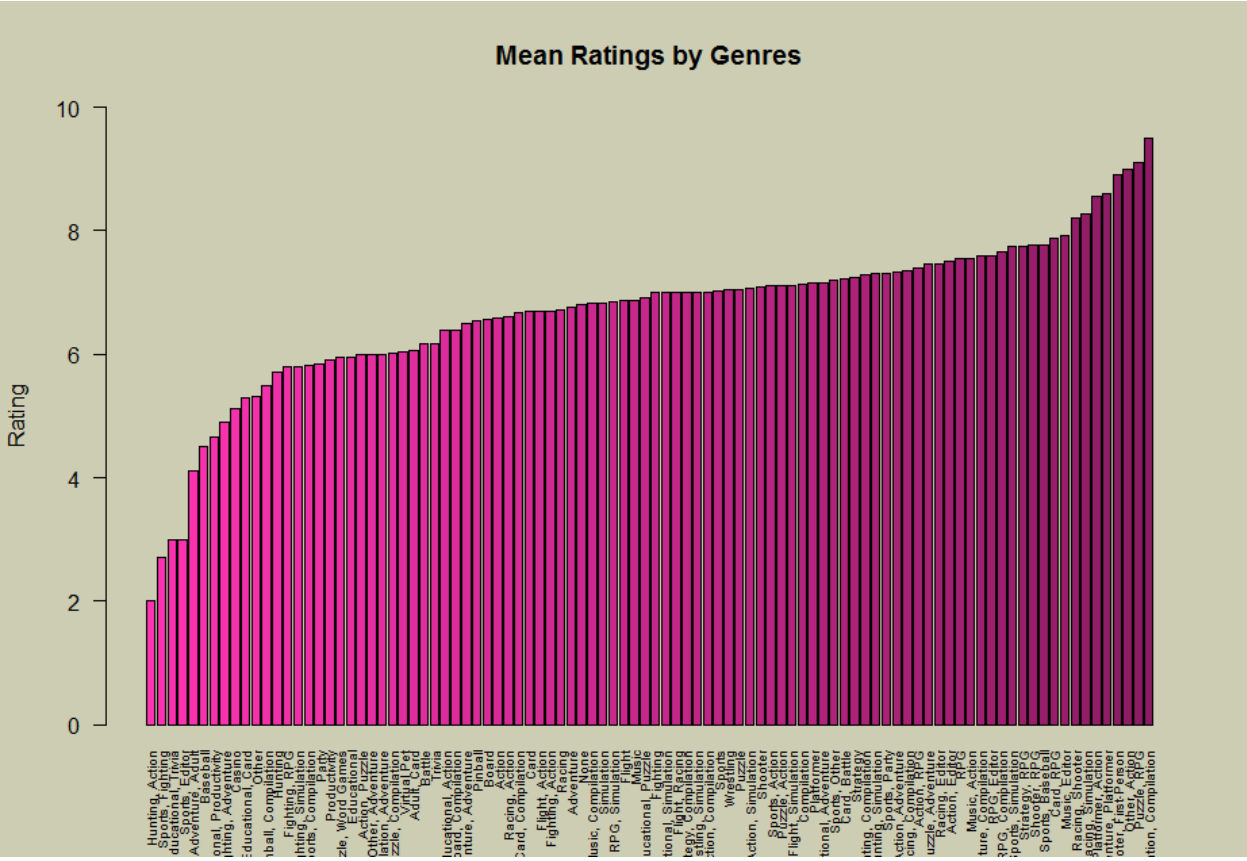
```
> summary(dat)
                           Game                Platform         Score               Genre     
 Cars                        :   10   PC            :3026   Min.   : 0.500   Action    :3628  
 Madden NFL 07               :   10   PlayStation 2:1683   1st Qu.: 6.000   Sports    :1854  
 Brain Challenge             :    9   Xbox 360      :1582   Median : 7.200   Shooter   :1472  
 Madden NFL 08               :    9   Wii           :1347   Mean   : 6.915   Racing    :1189  
 Need for Speed Undercover:    9   PlayStation 3:1295   3rd Qu.: 8.100   Strategy  :1013  
 Ratatouille                 :    9   Nintendo DS   :1040   Max.   :10.000   Adventure:1012  
 (Other)                     :17478   (Other)       :7561                    (Other)   :7366  
```

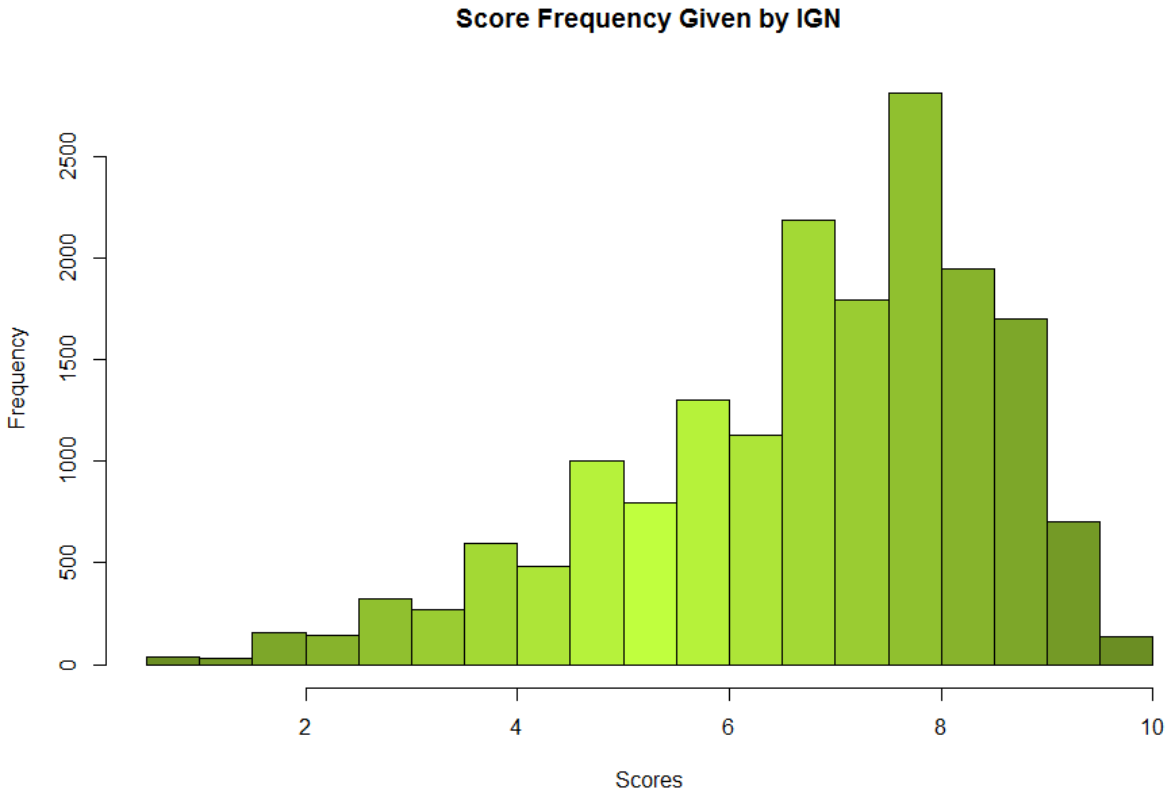
The dataset contains 4 fields, with a total of 17,534 observations

- Game (string): Game's title.
- Platform (string): System which the game was published, and intended to be played on.
- Score (float number): review rating or score determined by experts at IGN where "all reviews go through a stringent editing process for fairness, transparency and accuracy by the time they finally appear on IN and stand as the IGN Review." Possible minimum score can range from 0.0 ~ 0.9 (disaster) to 9.0 ~ 9.9 (amazing) or 10 (masterpiece). In this paper, I will use "score" and "rating" interchangeably.
- Genre: categorization of video games based on gameplay interaction

Visualization of the whole dataset:







II. Main features of data set presented with data visualization

Two fields we will be using is “Score” and “Platform”.

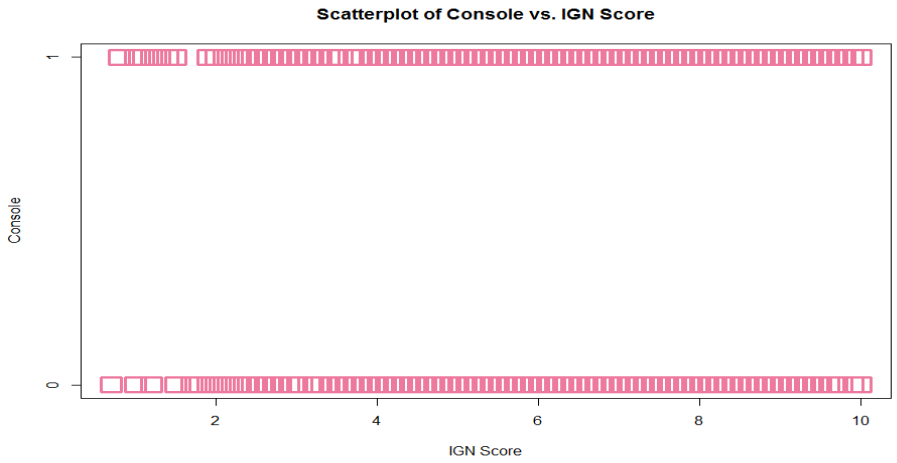
There are a total of 56 different platforms according to the dataset, our goal is to divide those platforms into two categories: “console” and “non-console”, where “console” is represented by value 1, and “non-console” is represented by value 0. Game platforms can be categorized as: Arcade, audio, console (home video game console), handheld (mobile, handheld console), online and PC. We will consider “home console” as “console” platform, the rest is “non-console”. Within the dataset, “DVD/HD Video Game”, “Wireless” can be either “console” or “non-console”, therefore I exclude those out.
I categorized all platforms into types in PlatformCategory.csv.

```
> summary(platformDat)
             Platform       Platform.type   MeanScore
 Android          : 1   Arcade       : 1   Min.   :3.000
 Arcade           : 1   Handheld     :16   1st Qu.:6.521
 Atari 2600       : 1   Home console:26   Median :6.987
 Atari 5200       : 1   Mobile       : 6   Mean   :6.942
 Commodore 64/128: 1   PC           : 5   3rd Qu.:7.465
 Dreamcast        : 1                      Max.   :8.600
 (Other)          :48
>
```

As you may suspect, there are a lot of games that were published in multiple platforms, many of them can be both console and non-console games. Hence, to maintain exclusivity for the console rating, I extract out all cross platform games. Here’s the main data that I will be working with, please see finalproject.R for more details on the procedure of getting the main data. Main data before extracting out cross platform games, total observations: 16,628. Main data after extracting out cross platform games, total observations: 8,646

```
> summary(mainDat)
     Score           Console
 Min.   : 0.700   Min.   :0.000
 1st Qu.: 5.800   1st Qu.:0.000
 Median : 7.000   Median :0.000
 Mean   : 6.799   Mean   :0.436
 3rd Qu.: 8.000   3rd Qu.:1.000
 Max.   :10.000   Max.   :1.000
```

This data includes “Score” and “Console”, where “Score” is the IGN review rating for a game, “Console” consisted of 1 and 0. 1 means the game that has the corresponding “Score” is a console game. 0 means the game that has the corresponding “Score” is a non-console game. Please ignore the “Mean” in the summary(mainDat) above, Console has only binary values.



III. Research questions

1. Develop a model that will provide the probability and odds of a game being a console game for any given IGN rating?

2. Approximately what rating is associated with a probability of 50% (the odds are even) for a game to be a console game?
3. Input rating of games whose platforms are unknown into the model to determine the probability and odds of the game being a console game. The game was also not included in the original dataset.
   Chosen game: Superhot (Rating: 7.5)
4. Determine how decreasing the rating from 7.5 to 6.5 would affect the probability and odds of the game being console game.

IV. Method for addressing research question and explanation (20)

By looking at the scatterplot of the data above, it doesn't make sense to have a best fitted line. Hence, I'm going to use logistic regression to address the research questions because logistic regression allows us to:

- Model the probability of an event occurring depending on the values of the independent variables (in this case IGN game ratings), which can be categorical or numerical (in our case it is numerical).
- Estimate the probability that an event occurs for a randomly selected observation (that we want to predict) versus the probability that the event does not occur.
- Predict the effect of a series of variables on a binary response variable. In our case, if we only have one independent variable IGN score, or more than one independent variables, and the one dependent variable that is binary (0 or 1).
- Classify observations by estimating the probability that an observation is in a particular category (such as a game being console game, or non-console game).

Reasons for not choosing other regression methods:

- Simple linear regression is one quantitative variable predicting another (our Console variable is binary, not quantitative)
- Multiple regression is simple linear regression with more independent variables
- Non-linear regression is still two quantitative variables, but the data is curvilinear.

Problems if we run linear regression for these research questions:

- Binary data does not have a normal distribution, which is a condition needed for most other types of regression (we can observe this through scatterplot above)
- Predicted values of the dependent variables can be beyond 0 and 1 which violates the definition of probability
- Probabilities are often not linear such as "U" shapes where probability is very low or very high at the extremes of x-values
- Violates assumption that error terms should be normally distributed

V. Data satisfaction of requirements of method demonstrated (10)

- Binary logistic regression requires the dependent variable to be binary. In our main dataset, the "console" data is binary. The game is either a console game (1), or non-console game (0).
- Logistic regression assumes that P(Y=I) is the probability of the event occurring, our dependent variables are coded accordingly.
- Logistic regression requires each observation to be independent. Our "Score" variables are scores of individual games, they are not related to one another, completely independent from one another.
- Logistic regression assumes linearity of independent variables and log odds.
- Logistic regression requires large sample size, our number of observations is 8,646.

VI. Method applied and method interpretation (20)

I used glm() function to apply the method on the data in R. The summary of the result is saved in mainDat.glm variable.

Based on the result, we got the p-value for predicting whether the given IGN score is for a console game or non-console game, is very small (p-value = 0.00035). Hence we can say that Score value does have an influence, or does help predict Console.
Based on the coefficients value, we can conclude that for every one unit change in Score, the log odds of game being Console (vs non-console) decreases by 0.04372.

Estimated Regression Equation: $\hat{p} = \frac{e^{\beta_0+\beta_1 x_1}}{1+e^{\beta_0+\beta_1 x_1}} = \frac{e^{0.03967-0.04372x_1}}{1+e^{0.03967-0.04372x_1}}$

Where $\hat{p}$ is the estimated probability of the game being a console game and x_1 is an IGN score. The odds of a given score being of a console game is $\frac{\hat{p}}{1-\hat{p}}$

```
> exp(mainDat.glm$coefficients)
(Intercept)       Score
  1.0404626   0.9572226
> exp(confint(mainDat.glm))
Waiting for profiling to be done...
                2.5 %    97.5 %
(Intercept) 0.8794354 1.2308336
Score       0.9345492 0.9804404
```

By using those functions above, and by looking at the result.
Value 0.9572226 tells us that for every unit increase of Score, we have 0.957 lesser likelihood, or 4.278% decrease of the game being a console game. This actually surprises me because I made an assumption before this study that console games tend to have higher scores.
Come back to our "Superhot" game with IGN rating of 7.5, based on the regression equation above we got the probability of Superhot being a console game is $\hat{p} = \frac{e^{0.03967-0.04372(7.5)}}{1+e^{0.03967-0.04372(7.5)}} = 0.4284$ *or 42.84%* chance of being a console game.
I created a table that has IGN Score and its probability of being a console game, probability of being a non-console game, and odds of being a console game accordingly. The data is saved in "oddstbl". By looking at this table, we can see that when the scores are at around *0.9 to 1.0, the odds of a score being a console game or non-console game is even, in other words, the probability of a score being a console game or non-console game is roughly 50%.*
Also, based on the same table, the odds of getting 7.5 is 0.7496, odds of getting 6.5 is 0.7831. *Hence, decreasing the rating from 7.5 to 6.5 will increase the odds of being a console game by (0.7831 – 0.7496)*100 = 3.35%.*
I made another table "intervalScore" where there're the scores are increased by 0.1 interval, as well as the odd ratio accordingly. The plot of this table is the exact same plot of function $y = e^{\beta_1 \times \delta} = e^{-0.04372 \times \delta}$ where δ is the score interval. Hence, we got an exponential regression line for score intervals and odds ratio.