Springboard Data Science Career Track

**Metacritic Score Prediction Model**

horizontal line

# *Introduction*

### Background:

The video game industry continues to grow in popularity across the globe as our world becomes more technologically advanced. Year after year, video games become more accessible to the average person than at any previous time in history. With the continued saturation of video games by software companies, it has become increasingly difficult for a single game to gain notoriety. Metacritic is website millions of consumers visit per day to view the best-reviewed games. Metacritic review score offers a way to gain notoriety in the mind of video game consumers.

### *Problem:*

The Metacritic score is only received after the game has already been released. By that time, it is too late to make changes to a video game that would impact its commercial success. In my report, I plan to give a detailed analysis of Metacritic review scores and to develop a predictive model for Metacritic review scores. This predictive model will allow company executive to anticipate how well received a game would be by industry critics before the launch of their title.

### 

### 

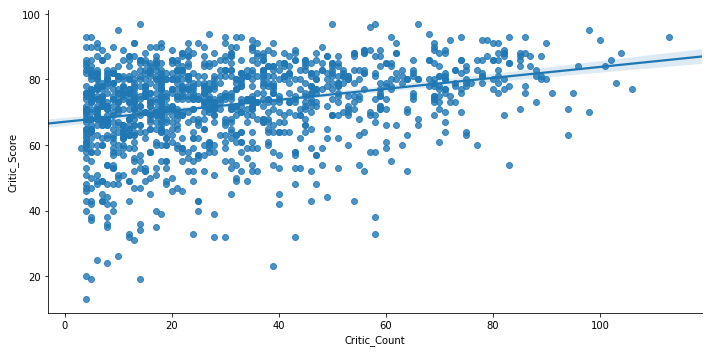
### 

### **The Dataset**:

I began my analysis with a dataset available from Kaggle found at the following link: “<https://www.kaggle.com/rush4ratio/video-game-sales-with-ratings/data>”. To build the best predictive model, my data set will only include video games that have been released after 2012. video games review scores consider more aspect of a game now than in recent years. With that in mind, I decided to keep the data used to construct the model as current as possible. After importing the data set from Kaggle, I noticed several columns contained sales data which is irrelevant in terms of the project goal. I proceeded to drop all columns containing sales data from the data frame. Next, I discovered that there were several missing values in the user count and the user rating columns. I felt the best way to quickly deal with this problem would be to calculate the mean of each column and replace those missing values with the mean of all know values. The 'ratings' column contains some missing values as well. The 'ratings' column contains data about each games ESRB rating. After some brainstorming, I couldn’t come up with a method to fill the missing values that would not jeopardize the accuracy of the model. The best thing to do in this situation is to drop the video games that have missing values in the rating column.

# *Exploratory Data Analysis:*

#### Critic Score and Critic Count



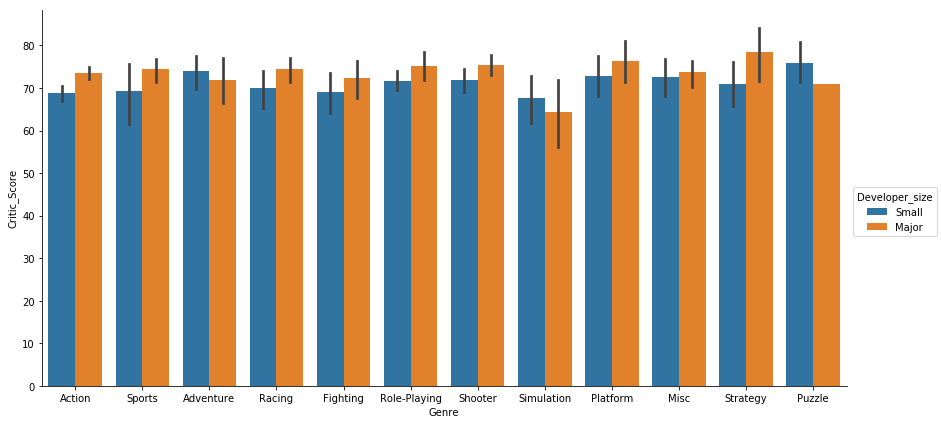
First, I wanted to explore the relationship between critic score (the target variable) and critic count. I came up with a scatter plot that displays a slightly positive relationship between critic count and critic score. Most games have between 5 - 60 games critic reviews. Next, I wanted to get a sense of which genre was on average rated the highest.

#### Violin plot of Critic Score and Genre

### 

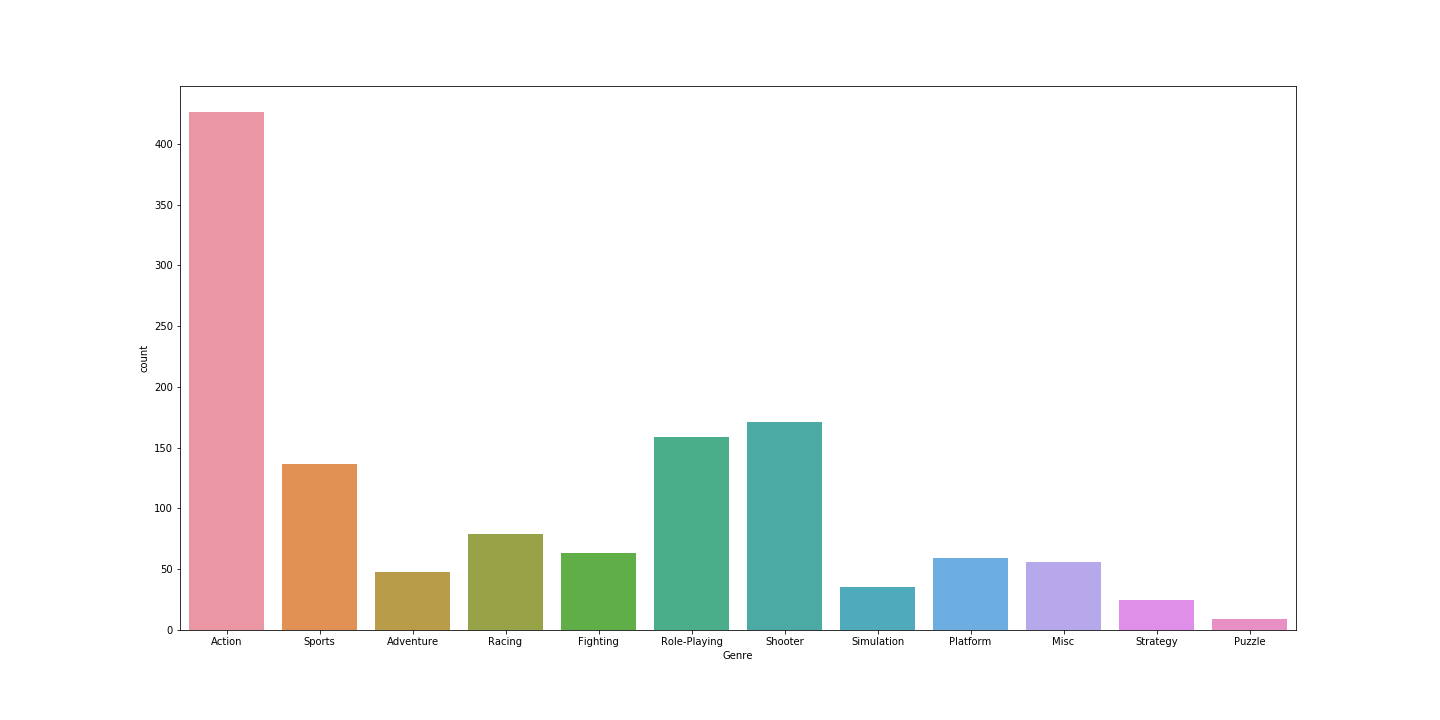
Looking at games by genre, there isn't any drastic difference in the average review score by genre. Racing games have the greatest variance in review score, while puzzle games have the lowest. The plot does show that games the fall into the Platform and Strategy genres are rated on average slightly higher than other categories.

#### Critic Score and Genre grouped by Developer



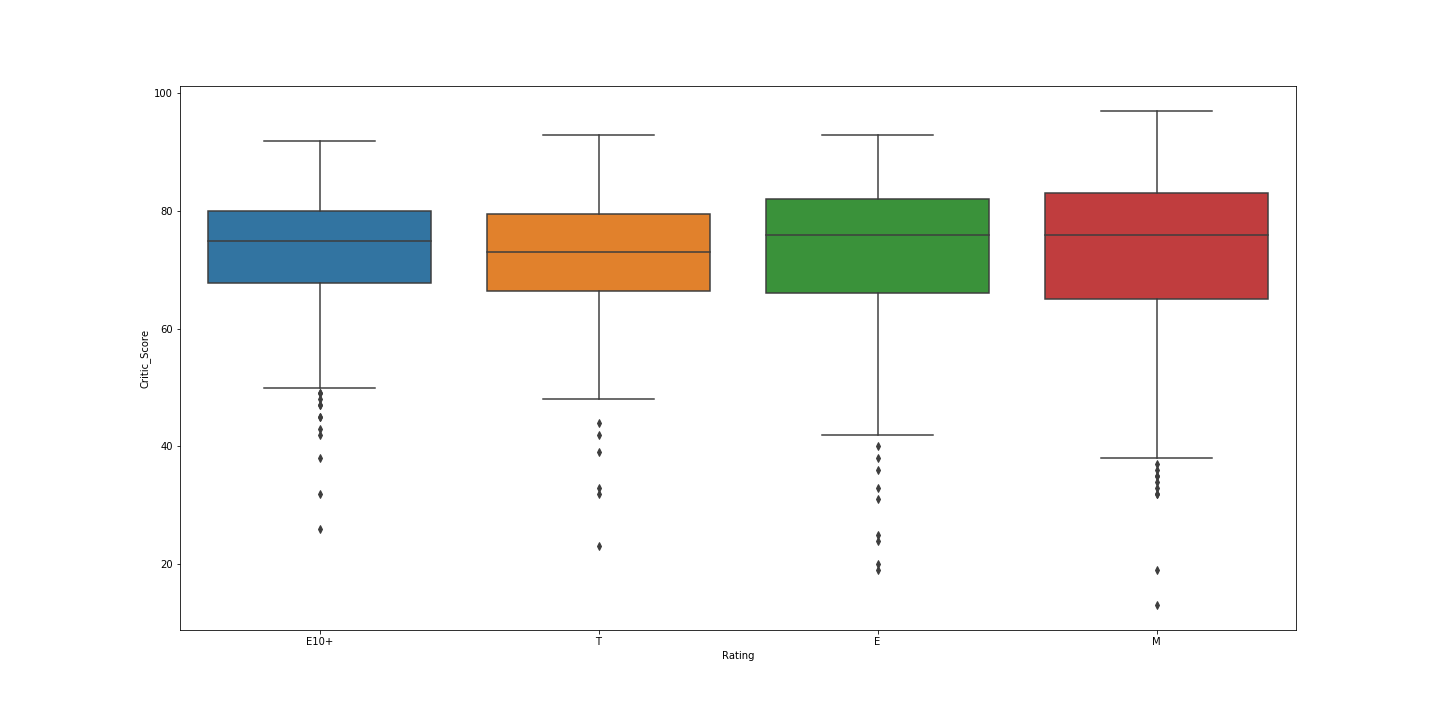
Taking a deeper look at critic scores by genre, I tried to split the games made by the major and small developer to see if there is a trend in the scores. The plot reveals that games developed by major developers rated higher on average than games made by small developers excluding games in Adventure, Simulation, and Puzzle genres. I suspect that major developers have a bigger budget to work on games which would lead to higher quality games. If we were able to get access to the number of developers working on the game or the budget for development, we would be able to get a better understanding of what exactly causes major developer's games to be rated higher.

#### Count of Games released by Genre



### The most popular type of game released between 2012 and 2016 were action games. The number of action games released during that period is significantly higher than any other category. Following the action games genre, Shooter, Role-playing, then Sports are the next highest genres released.

#### Critic Score and Ratings Chart



The box plot of Critic Scores and Ratings shows ‘E’ and ‘Mature’ rated games are rated the highest and have similar mean critic scores. Although, the variance of ‘Mature’ rated games critic score is larger than E rated games.

# ***Inferential Statistical Analysis for Prediction Model****:*

To develop the best model for predicting Metacritic review scores on video games, I need to verify which variables have a statistically significant relationship with Critic Scores variable. A chi-squared goodness of fit test allows us to test the statistical significance of a relationship between categorical variables. I will run the test on several variables within the data frame to see which variables should be included when creating the final model. To be clear, for each variable that is tested our null and alternate hypothesis is as follows:

H0: There is no relationship between the variable and Game Quality (Categorized Critic Scores).

Ha: There is a relationship between the variable and Game Quality (Categorized Critic Scores).

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | Chi Statistic | P-Value | Expected Value |
| Genre | 61.9599558931 | 0.00166336164951 | 33 |
| Platform | 26.4503350982 | 0.493729754709 | 27 |
| Rating | 33.5884026466 | 0.00010543964857 | 9 |
| Developer size | 11.1890509966 | 0.0107463201464 | 3 |
| Critic Count | 436.95798475 | 3.28801958791e-08 | 288 |
| Publisher size | 14.2250362929 | 0.00261430484993 | 3 |

After running a couple of chi-squared goodness of fit tests, I found that the variables 'Genre', 'Rating', and 'Critic Count' have a statistically significant relationship with 'Game Quality'(Categorized Critic Scores). The 'Platform' variable doesn't have a statistically significant relationship with 'Game Quality' based on the p-value of 'Platform' surpassing .05 alpha threshold. The test of 'Developer size' (Whether a game was made by a major developer or not) and 'Publisher size'(Whether a game is published by a major publisher or not) relationship with 'Game Quality' isn't valid since the expected value for both variables is 3 and need to be greater than 5 to be considered valid.

Next, I wanted to explore the relationship between critic score and YSR (years since platform release). I predict that there should be a positive correlation between YSR (the difference in platform release date and game release date). As console get older, scores should increase as developers become more familiar with each platforms capability. Surprisingly, I found almost no correlation between Critic Score and YSR. The calculated correlation from the data set is -0.04164. This leads me to believe that YSR should be left out of the variables used to construct my predictive model.

Lastly, I wanted to determine if the difference in proportion of high quality games developed by major developers and small developers is statistically significant. I decided to run a two-sample proportion test. I predict that there will be a significant difference in the number of high quality games created by Major developer and small developer.

H0: The null hypothesis is that there is no difference between the two proportions.

HA: The alternative is that there is a statistical difference between the two proportions.

The test resulted in a z-statistic of 9.2869213766 and a p-value of 1.58815467642e-20. The P-value from out proportion test is extremely low. We can conclude that there is a statistically significant difference between the number of high quality games generated by Major and Small developers. There is sufficient evidence to include this variable in our model development