Capstone Project Milestone Report

By. Pratyush Pati

The initial problem I had set upon to explore in my project was the ability to find what variables of video games allowed them to win an Editor’s Choice award. The main purpose of the project’s finding would be beneficial to game developers as the success of developer companies are often measured by their ratings of their games and the awards they receive. By understanding what characteristics of a certain video game allow it possibly to win one of these prestigious award, IGN’s Editor’s Choice award, game developers will be able to know if their game will be successful as well as what type of project they should pursue to earn this achievement for their future games.

The dataset used for this project is from from Kaggle, specifically from the dataset provided in 20 Years of Games by Eric Grinstein (<https://www.kaggle.com/egrinstein/20-years-of-games>). The dataset includes about 18,000 video games reviews from IGN that range from games released from 1996-2016 in a csv file format. In more detailed view, the dataset contains, respectively: **Score\_phrase** (one word description of the game as a factor of either Amazing, Awful, Bad, Disaster, Good, Great, Masterpiece, Mediocre, Okay, Painful, Unbearable), **title** (video game’s title name), **url**(IGN’s url link to the review), **platform** (videogame platform from Nintendo, Sony, Microsoft, etc.), **score** (decimal value within the range of 0 – 10), **genre** (game’s category or style), **editor\_choice** ( have won *Y* or not won *N* the award), **release\_year, release\_month, release day**.

The dataset is pretty insightful to find relationships between these variables against the **editor\_choice**, but does have some limits. In my initial search to answer my project’s problem, I had wanted to obtain a dataset that included revenue and profits for each video game and/or each developer company since generally games that had higher commercial success tended to earn more prestigious awards. However, since most of these datasets are not readily available or hidden behind paywalls, I decided this Kaggle dataset would be appropriate. Additionally, since most of my variables are categorical variables I will be limited to fewer tests, specifically to Chi-Squared tests. For future development and a more accurate predictor of my problem statement, a better data set with more quantitative and categorical values would be helpful to explore numerous statistical tests between the variables.

In my initial exploratory data analysis (EDA) of the dataset, I have found some interesting findings. Since the dataset ranges of videogames from 1996-2016, there are numerous platforms that have been discontinued or appear solely for one specific video game. There is also some irregularities in the dataset with games that appear in the incorrect release year, mistaken a game’s release date with when IGN reviewed the game, generally with older games given recent release dates, and missing genre data. In preliminary data wrangling process, I decided it was best to remove these data and pick a recent sample year time period such as 2010 to avoid such mistakes. Sampling a certain year was also helpful to allow me to answer my problem and run future statistical tests as well as avoiding handling too much data at one time in preliminary plots and findings. Another interesting aspect of the dataset, which one of the Springboard advisors, *Miadad Rashid*, brought to my attention, was that the categorical variable **score\_phrase** is also ordinal as it has an inherent order going from Disaster to Masterpiece. It can allow the possibility to change them into a numerical rating and treat them as a quantitative variable in a correlation test. Also, since some games appear numerous times in the dataset due to releasing on multiple platform, I can possibly add additional variables to my data for indicating games that have such patterns.

Based on my initial findings, my best approach would be run chi-squared tests for my categorical variables against if the videogame won an Editor’s Choice award or not. By getting the observed total counts and running the statistical test on expected counts in different sampled years with an appropriate significance level, it would lead to conclusive result for my project’s problem statement. Also, since some of my categorical variables like **genre** and **platform** have sub-categories within them, if time-permitting, I would also like to find results how likely they are to win the award against each other. Another interesting factor I would also like to explore to see if a unique keyword or identifier in the videogame’s **title** lends to winning the award more.