Capstone Project 1: Milestone Report

Introduction:

The video game entertainment industry is a multi-billion dollar industry that is continuously growing every year. There are some video games out there that are selling millions of copies and some that are barely making a scratch in the industry. What makes one game sell more than another? Are there trends or relationships that help in explaining a video game's global sales? These questions, as well as many others, are important ones to ask if you are interested in the video game entertainment industry. The results of this capstone project may be helpful in identifying the factors that affect how well a video game performs in sales, potentially providing direction for companies. Given the dataset being used, the main question for this capstone project is if the Metacritic score of a video game has an effect on how well a video game sells.

The Dataset, Cleaning and Wrangling:

The dataset that was used for this capstone project came from a kaggle competition on video game sales with Metacritic ratings. This dataset contained numerous columns related to video games, such as: name, platform, year of release, genre, publisher, NA/EU/JP/other/global sales, and critic/user score. Coming from a kaggle competition, this dataset was well organized and in a relatively clean state. However, there were still aspects of the dataset that needed to be dealt with and cleaned. There were duplicate and missing values; columns and rows needed to be filtered as well.

Upon importing the dataset through pandas, all the 'tbd' values were converted to NaNs. This dataset was compiled in 2016, so any upcoming games beyond 2016 were removed. All video games that had no critic and user scores, and had zero global sales were filtered out. The Year_of_Release column was of data type float64. It was converted to int64. Any potential outliers were observed using the pandas DataFrame describe method. A video game called "Wii Sports" was seen to be a very large outlier of global sales, but it was kept in the dataset. It will be removed and dealt with as necessary during the analysis phase. Duplicate video games were observed in the dataset. This was due to video games being remade/re-released on different years. They were also released on different platforms. To deal with the duplicates the entire dataset was sorted by descending global sales and grouped by video game name. Aggregations were calculated on year of release by first value, publisher by mode, global sales by sum, critic and user scores by max, and genre by mode. The user scores were multiplied by 10 in order to bring them to the same scale as critic scores. The cleaned dataset was then sorted by descending global sales and exported to a csv for the next step.

Exploratory Data Analysis, Storytelling and Visualizations:

Now that the video game dataset had been cleaned, and the relevant rows and columns were obtained, exploratory data analysis and visualizations were explored and created. After importing the cleaned dataset, the pandas info and describe methods were called to see what can be explored. Through this exploration, 7 questions were produced. These questions will provide the basis for the visualizations and storytelling of the dataset.

The first question asked about what a scatter plot of the Metacritic scores and global sales looked like. Inspecting the plots reveals that there seems to be an obvious positive trend.

However, most of the data points seemed to be on the lower end of the global sales with a few stretching out to higher global sales values. Perhaps a deeper look at the trend will require a log transformation of the global sales. Taking the log should provide a better visualization as well as normalize any potential outliers in the global sales.

The second question explored the relationship between video game genres and global sales. Performing the pandas series unique method on genre column revealed that there are 12 different genres that video games could be categorized under. Plotting a bar chart of genres and video games showed that action games performed the best, by a significant margin. It sold about 400 million units over the next two highest selling genres, sports and shooters. In order to explain why action games had such a high total global sales value, a deeper analysis will need to be performed.

Question three took a look at how well popular publishers performed. It was seen that the dataset contained 315 unique publishers. For the initial analysis, only the top 10 publishers who released the most games were looked at. By retrieving the top 10 publisher names from the value counts method, the dataset was filtered accordingly. In order to create visualizations for the publishers and global sales, the data needed to be grouped by publishers with an aggregation on the sum of global sales. For this analysis, both a pie plot and a bar chart were used to visualize the data. The main reason for this was to show the differences between analyzing a pie plot and a bar chart. By looking at the pie plot, the only thing that could have been inferred was that both Electronic Arts and Nintendo performed better than the rest of the top 10 publishers. However, with the bar chart, it can be seen that the amount by which they out performed the other publishers was over 200 millions units. The relationship between count of games a publisher produces and global sales could be explored further.

The previous questions looked at the relationships between genres and global sales as well as publisher game counts and global sales; question four asked about the genres that some publishers tend to produce. The top 5 publishers with the most games produced were looked at for this exploration. The dataset was grouped by both publisher and genre with an aggregation on the count of each genre, producing a multilevel index. In order to visualize the data, the DataFrame needed to be unstacked. This multilevel index was unstacked using the pandas unstack method. The visualization revealed that Electronic Arts produced a very large amount of sports games, a majority of the video games made were action games, and Nintendo produced more role-playing games than any other genres they produced. These results indicated that there may be variances in the dispersion of video game genres.

The fifth question explored the relationship between the total amount of video games produced per publisher and their total global sales. It was clearly seen that as the count of games increased, the total amount of global sales also increased. For a deeper exploration, perhaps the log of both global sales and video games counts could be performed, because a majority of the data points were towards the bottom left of the visualization.

Looking at which publishers had the highest average ratings was what question six entailed. Once again, only the top 10 highest producing publishers were looked at for this initial analysis. After grouping the dataset by publishers and aggregating both critic and user scores by their means, a bar chart was created. The visualization showed that all the average scores for the top 10 publishers ranged between 65-80. It was also seen that Nintendo had the highest average score for users. Electronic arts and Nintendo had nearly the same values for the highest critic scores. Also, it was seen that average user scores were higher than critic scores, with Electronic Arts being an exception.

The final question asked about the amount of video games released each year. The dataset was grouped by year of release and, using the pandas size method, aggregated by the count of games. Analyzing the visualization shows that there was a huge spike in count of games in the dataset at about year 2001. The count of games also started to drop at around the year 2009. This is also another potential observation that could be explored further.

Statistical Inference:

From the previous section where visualizations were created for the video game variables, some potential avenues to explore were discovered. For this part of the analysis, the focus was put into looking at the differences between average critic and user scores, the correlation between critic scores and global sales, the correlation between count of video games per publisher and global sales, and the differences in analyzing the trends in video game genres versus sum total global sales and average global sales.

Metacritic is a multi-media rating website that assigns scores to video games or other forms of media entertainment. A Metacritic score can either be a score given by professional critics or a score given by the users of the website. Before even looking at the main question of this capstone project of the relationship between Metacritic score and global sales, it was first important to understand the differences between critic and user scores.

The first step was to plot a histogram of the critic and user scores in the dataset. Simply by observing the visualization, it was seen that there was a higher frequency of critic scores below 75 and a higher frequency of user scores above 75. Calculating the average of the critic scores and user scores gave them a value of 70.45 and 74.34, respectively, with a difference of 3.88 in favor of user scores. In order to ensure that this observed difference of means was just due to chance, a hypothesis test was performed. The null hypothesis for this test was that the means of critic and user scores were the same, their difference would be equal to 0. The alternative hypothesis was that there is a difference between the means of the critic and user scores. The significance level for this hypothesis test was 0.05.

In order to perform a hypothesis test, the null was assumed to be true. To see what the probability, the p-value, of getting the observed difference of means assuming the null hypothesis was true, a simulation of the null hypothesis had to be ran. The simulation started with creating a new mean value to shift the means of the datasets of critic score global sales and user score global sales to make them equal. This was done by concatenating the two datasets and taking the mean of that dataset. Then the critic score global sales dataset was subtracted by its original mean and the new mean from the concatenation was added to it. The same was done for the user scores. Bootstrap resampling was then used to resample the data and to create a replicate difference of the means. The bootstrapping was performed 10,000 times to create a distribution of the possible differences of means for critic and user scores, where the means were the same. A histogram of the simulated differences was plotted and a vertical line of the actual observed difference was also plotted. Just from the visualization, it was very clear that that the observed difference was nowhere near the distribution of possible differences of the simulation. The p-value was calculated by adding up each time a datapoint in dataset was at least or more extreme than the observed value and then dividing that by the length of the simulated dataset. The p-value was 0, which was lower than the significance level of 0.05. This meant that the null hypothesis of the means of the critic and user scores being the same could be rejected and the alternative hypothesis of the means being different could be accepted. The concluding inference that was drawn from this was that there was a statistically significant difference between the critic scores and the user scores.

After determining the difference between the Metacritic scores, their effects on global sales could now be observed. For the analysis only the critic scores were used. It was known from the previous visualizations that many of the data point had low global sales, with a few stretching out towards the higher values. The log was performed on the global sales values, allowing the data points to spread out a little more. With the log of global sales and the critic scores, a linear regression analysis as well as a hypothesis test were performed. The null hypothesis of a linear regression test is that there is no correlation between the two variables, the slope of the regression line is 0. The alternative hypothesis is that there is a correlation between the variables and the slope is not 0. The significance level of this hypothesis was 0.05.

The scipy stats linregress function performs a linear regression on two datasets and returns the slope, the intercept, the r value, and the p-value. Using the outputs of this function a line was fitted to the plot of critic scores and log of global sales. Taking the square of the r value returns the R^2 value, the coefficient of determination or a goodness of fit measure. The R^2 of the fitted line was 0.17. This meant that 17% of the variation in the log of global sales could be explained by the critic scores in the linear model. A quick look at the user score and log of global sales R^2 value showed that it was an even lower value of 0.05, or 5%. Also, the p-value, which was practically 0, was below the significance level of 0.05, allowing for rejection of the null and acceptance of the alternative stating that there was a statistically significant correlation between critic scores and the log of global sales. The results of performing this analysis revealed that there may be other variables that could also contribute to the global sales of a video game.

It was seen that there seemed to be a strong correlation between the count of video games per publisher and their global sales from the visualizations made previously. This visualization contained many of the data points towards the bottom left of the plot. In order to achieve a more representative and appropriate visualization of the variables, the log of both the count of games per publisher and their total global sales was taken and linear regression was performed. A hypothesis test for the correlation between the log of game counts and the log of global sales was ran. Once again, the null hypothesis was that there is no correlation between the variables. The alternative hypothesis is that there is a correlation. The significance level was 0.05. Running the linear regression function returned the slope, the intercept, the r-value, and the p-value. A line of the was plotted using the slope and intercept. The R^2 value for these two variables was 0.69. This meant 69% of the variability in the log of global sales could be explained by the log of count of video games with the linear regression model. The p-value was practically zero for this test and was below the significance level, allowing for rejection of the null hypothesis. There looked to be a decent correlation between the log of count of games per publisher and the log of global sales.

Another variable of video games that was seen to have an affect on the global sales was the genre. For this dataset, all of the video games could be put into 1 of 12 genres. It was seen that action games had a total global sales value of about 1200 million units. That was about 400 million units over the next two highest selling genres. After looking a bit deeper, it was discovered that the proportion of action games for this dataset was about 0.198. That meant that nearly 20% of all the video games in this dataset were action games. So it made sense as to why the global sales values was so high. In order to determine what the true population proportion of action games could be, bootstrap resampling was used in order to calculate a 95% confidence interval.

The genres data was bootstrap resampled and then the proportion of action games was calculated for that resampled data. This was repeated 10,000 times to obtain a distribution of possible values that the action game proportions could take. Taking the 2.5th and the 97.5th percentile of this distribution returned the upper and lower bound of the 95% confidence interval. There was 95% confidence that the true proportion of action video games was between the values of 18.6% and 21%.

In order to confirm that the observed video game proportion was not due to chance, a hypothesis test with a significance level of 0.05 was needed. The null hypothesis was that the proportions of all the video game genres were equal, action games had a proportion of 1/12, or 8%. The alternative hypothesis was that there was a difference in the proportion of the video game genres, action video game proportion was not equal to 1/12, or 8%. The p-value was practically 0 and the null hypothesis was rejected. The alternative hypothesis of there being a statistically significant difference in the proportion of video game genres was accepted. The results of this analysis revealed that the large global sales value of action video games was likely due to it have a very large proportion of 20%. This lead to the thought of looking at the average global sales instead of sum total for genres, forming the basis of the last analysis.

After removing the outliers of “Wii Sports” and “Grand Theft Auto V” the plot of each genre with their average global sales was made. It was seen that the shooter genre performed the best with an average of 1.6 million copies sold, followed closely by sports and misc. In order to confirm that the variation of sales for the genres were not significantly different, a hypothesis test was performed using the genres of shooter and sports. The null hypothesis stated that there was no difference between the standard deviations of sales for shooters and sports. The alternative hypothesis stated that there was a difference between the standard deviations. The significance level was 0.05.

Once the dataset was separated into shooter and sports the standard deviations were calculated to be 3.52 and 2.70, respectively, with a difference of 0.82. To simulate a dataset of equal variation, permutation resampling was performed on the concatenated shooter and sports datasets. The resampled data was then split back into the original ratio of the separate datasets and the standard deviations were calculated. A difference of standard deviations of the resampled datasets was then taken. This was repeated 10,000 times to obtain a distribution of standard deviation differences under an assumed true null hypothesis. The p-value was calculated to be 0.11. This value was above the significance level, allowing for the acceptance of the null hypothesis stating that there was no statistically significant difference between the standard deviations of global sales for shooters and sports. The shooter genre had the highest average global sales and it was not due to an outlier that the other top selling genre, sports, did not observe.

Summary:

This project started off with a question about how the Metacritic score of a video game was related to its global sales. In an attempt to answer this question, other inferences and insights were discovered about the other variables of a video game. It was seen that the average user scores was different from the average critic scores. This meant that they may interact with other video game variables a little differently. A correlation between the critic scores and global sales was observed, but it was not a very strong one. This led to the thought that there may be other variables that contribute to the global sales. There was a relatively strong correlation seen between the count of games a publisher produces and the total global sales of that publisher. A line was fit to this data, allowing for predictions of how much global sales a publisher would expect to see if they sold a certain amount of video games. It was also seen that genres have a significant correlation to the global sales. Initially, it was observed that action games performed abnormally well in total global sales compared to the other genres. After taking a deeper look, it was found that the proportion of action games could explain this observation. Of the 12 video game genres, 20% of our dataset were action games. Looking at the mean, instead provided a better representation of which genres performed best in global sales. After confirming that the standard deviations of the global sales values for the top 2 games were not different, it was inferred that the highest selling genre was shooter.