Capstone Project 1 Report - Video Game Sales

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 or even reinforcing certain business decisions that they have already made. 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. Other questions about video games may also provide insight into how successful a video game is. These others questions that can be explored are who made the game, when was the game released, what kind of a video game is it, and who reviewed the game?

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. The Metacritic ratings contains both a critic score and a user score. The critic score comes from professional video game critics and the user score comes from normal everyday gamers. 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.

This dataset contained 'tbd' values that were placeholders for missing values. These needed to be converted into a form that wouldn't be interpreted as a real value in python. Upon importing the dataset through pandas, all the 'tbd' values were converted to NaNs. This dataset was compiled in 2016 but it had many upcoming or to be released games, so any upcoming games beyond 2016 were removed. The main question of this capstone project was to determine whether or not a video game's score had an affect on its sales. Therefore, all video games that had no critic and user scores, and had zero global sales were filtered out. The Year_of_Release column is one that contains values for years, but it was of data type float64. It was converted to int64. Any potential outliers were observed using the pandas DataFrame describe method. Two video games called "Wii Sports" and "Grand Theft Auto V" looked to be very large outliers of global sales, but were kept in the dataset. They 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. In order to perform comparable analysis on the Metacritic scores of a video game, 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. Figure 1 below shows what the cleaned dataset looks like.

	Name	Genre	Year_of_Release	Publisher	Global_Sales	Critic_Score	User_Score
0	Wii Sports	Sports	2006	Nintendo	82.53	76.0	80.0
1	Grand Theft Auto V	Action	2013	Take-Two Interactive	56.57	97.0	83.0
2	Mario Kart Wii	Racing	2008	Nintendo	35.52	82.0	83.0
3	Wii Sports Resort	Sports	2009	Nintendo	32.77	80.0	80.0
4	Call of Duty: Modern Warfare 3	Shooter	2011	Activision	30.59	88.0	34.0

Figure 1: The cleaned dataset, ready for analysis.

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 EDA, potential storytelling, and visualizations of the dataset.

The first question asked about what a scatter plot of the Metacritic scores and global sales looked like. Inspecting the plots revealed that there seemed to be an obvious positive trend (Figure 2). 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.

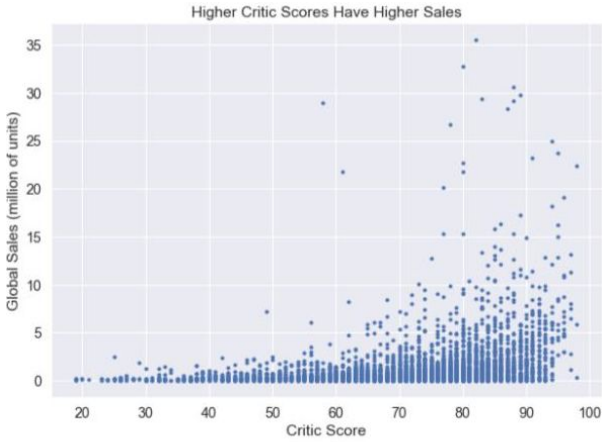

Figure 2: A scatter plot of critic scores and global sales in millions of units sold. Many of the data points are towards the bottom. Higher scores tend to have higher sales.

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 (Figure 3) showed that action games performed the best, by a significant margin. They 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.

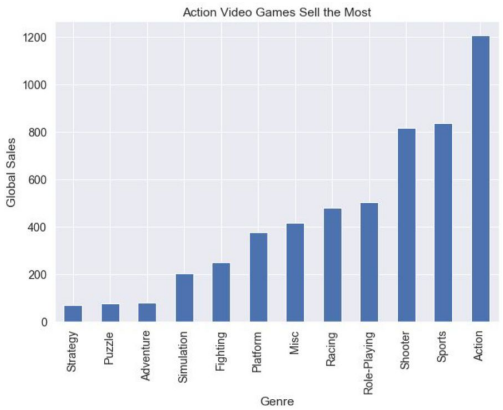

Figure 3: A bar graph of video game genres and their total global sales.

Capstone Project 1 Report - Video Game Sales

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 a bar chart (figure 4) was used to visualize the data. With the bar chart, it can be seen that Nintendo and Electronic Arts sold the most and 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.

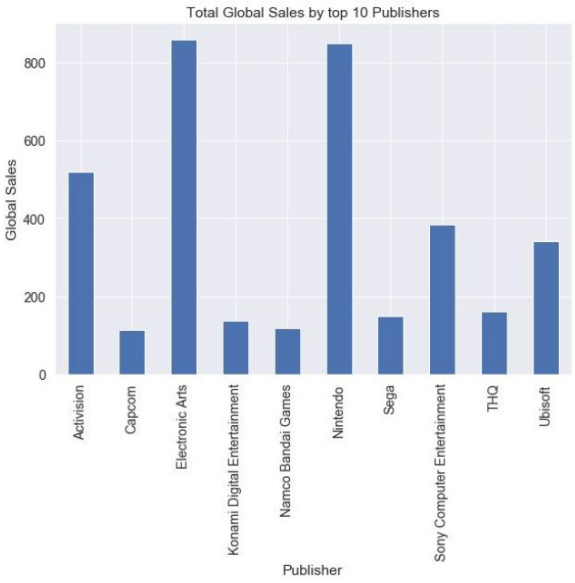

Figure 4: A bar plot of total global sales for the top 10 publishers.

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 (figure 5). 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.

Publisher	Activision	Electronic Arts	Nintendo	Sony Computer Entertainment	Ubisoft
Genre					
Action	60.0	44.0	37.0	52.0	66.0
Adventure	NaN	6.0	17.0	16.0	15.0
Fighting	3.0	12.0	6.0	9.0	12.0
Misc	21.0	9.0	40.0	26.0	27.0
Platform	12.0	4.0	41.0	30.0	17.0
Puzzle	1.0	5.0	24.0	3.0	1.0
Racing	13.0	43.0	15.0	33.0	17.0
Role-Playing	18.0	12.0	50.0	22.0	26.0
Shooter	38.0	52.0	11.0	33.0	43.0
Simulation	5.0	43.0	14.0	5.0	21.0
Sports	31.0	156.0	23.0	50.0	20.0
Strategy	8.0	7.0	12.0	3.0	12.0

Figure 5: A table of the top 5 publishers and their total sales for each video game genre.

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(figure 6). 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.

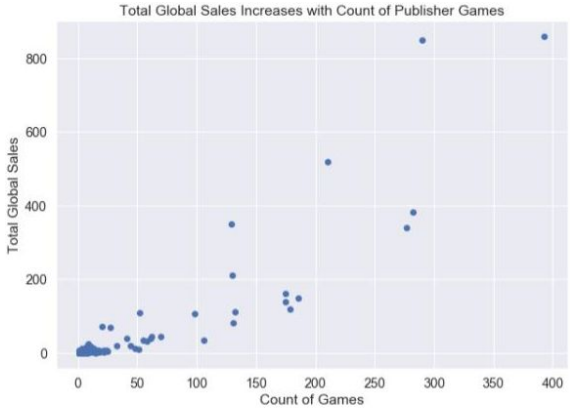

Figure 6: A scatter plot of count of games per publisher and their total global sales.

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 (figure 7). 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.

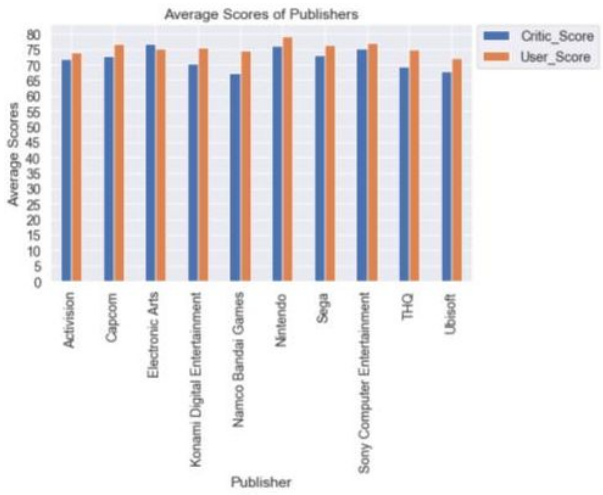

Figure 7: A bar plot of average critic and user scores for the top 10 publishers.

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.

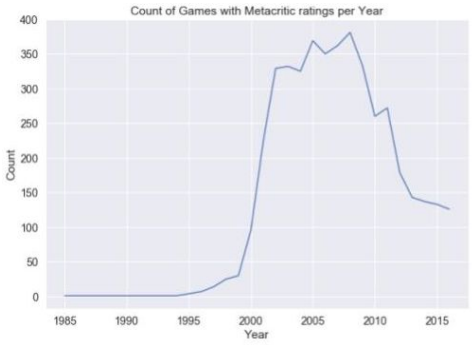

Figure 8: A line plot of the count of games released over the years. There is an interesting spike at about year 2001 and a drop at about year 2009.

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.

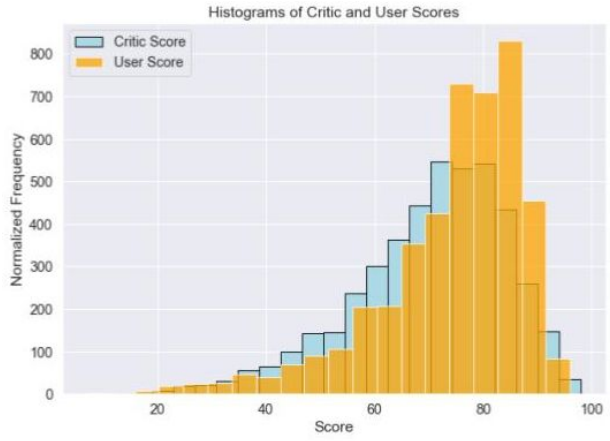

Figure 9: Histogram plots of both critic score(blue) and user score (orange).

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 (figure 9). 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 (Figure 10). 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.

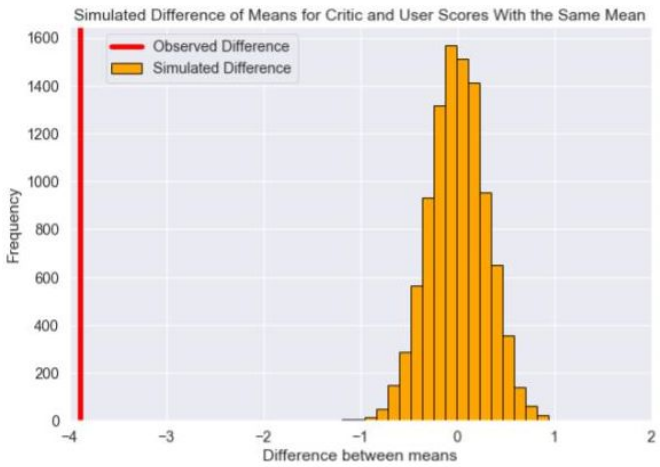

Figure 10: A sampling distribution of the differences of means for simulated critic and user score where their means are equal. The actual observed difference is the vertical red line.

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 visualization (figure 2) that many of the data point had low global sales, with a few stretching out towards the higher values. The log transformation 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(figure 11). 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.

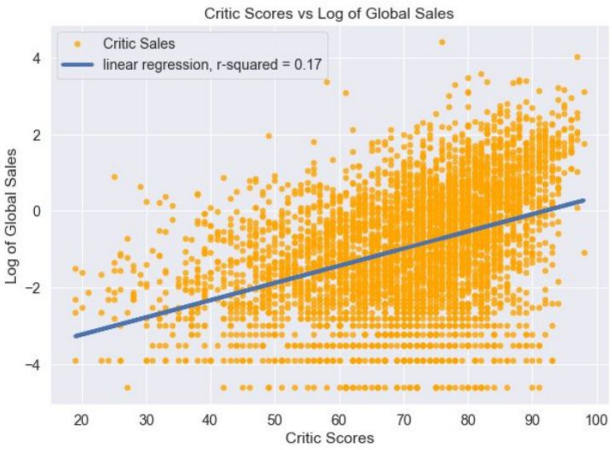

Figure 11: A scatter plot of critic scores and log of global sales. A linear regression line was fit to this plot and obtained an R-squared value of 17%.

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 (figure 6). 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 (figure 12). 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,

Figure 12: A scatter plot and linear regression fit on the log of count of games per publisher and the log of global sales. Fitting a line produced and R-squared value of 69%.

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.

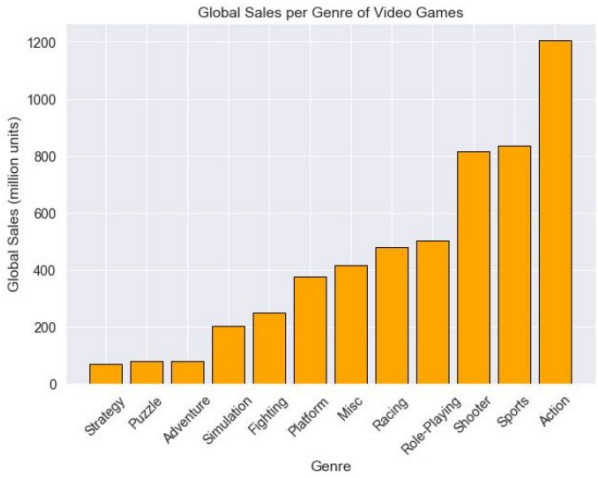

Figure 13: A bar graph showing the total global sales of each genre. Action games performed by almost 400 million units better than the next highest genre.

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 (figure 13). 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 2.5^{th} and the 97.5^{th} 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 (figure 14). 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.

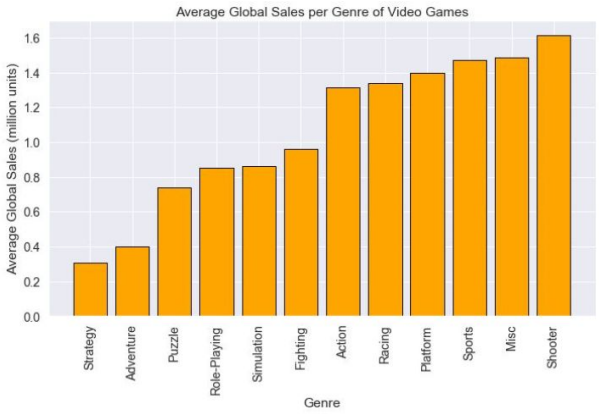

Figure 14: A bar graph of average global sales per video game genre.

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.

Machine Learning, In-depth Predictive Analysis:

Machine learning is a very powerful way to analyze data in order to perform predictions or to discover hidden insights. It allows for the production of models that fit to a dataset and generalizes to new unseen data. Machine learning can be used in regression analysis or classification problems. During the statistical inference portion of this capstone project a simple linear regression analysis was performed on a video game's global sales value based on its critic score (figure 11). By using machine learning, multiple additional features can be fed into a linear regression model in order to increase its predictive power as well as framing the problem in a classification context.

In order to feed additional features into the model the features first need to be transformed into a format that the scikit-learn library accepts. This means that the categorical features of this dataset, such as genre and publishers, will need to be encoded. Some of these categorical features will be turned into dummy variables through a method called one hot encoding. One hot encoding is a way to prevent bias that comes from ordinal values formed from the encoding of categorical features. This turns the category values into additional features that have binary values indicating whether or not the record contains that category. While performing feature engineering, care must be taken to avoid adding too many features to the model. Having too many features, or dimensions, may cause what is known as the curse of

dimensionality which can lead to various problems that occur in analyzing data in a high dimensional space.

For the linear regression model three new columns were added: log of global sales, count of publishers, and pub class. The log of global sales was added because it is generally advisable to perform log transformations when working with money. One of the features of a video game is the publisher. There is a larger amount of publishers in this dataset and performing one hot encoding on them could potentially lead to the curse of dimensionality. In order to circumvent that situation, a new column was created to categorize a video game based on its publisher's video game count. A video game was given a class of 0 if the publisher had a video game count less than 50, a class of 1 if the count was between 50-149, a class of 2 for counts between 150-199, and a class of 3 for 200+ video games. Now that the publisher class can be represented more concisely, this new feature can be one hot encoded in order to feed it into the linear regression model.

With the dataset transformed into a state that scikit-learn accepts, machine learning can be performed in order to predict the log of global sales for a video game. This was done by splitting the dataset into a training set and a hold-out set, cross-validating the model, fitting it on the training set, predicting on the hold-out set, and then calculating metrics for the model performance. The reason why the dataset is split into a training set and a hold-out set is because the model will need to be fitted and cross-validated on the training set and then tested on the hold-out set in order to properly measure how well the model performs to unseen data. Cross-validation was used to observe how the model would fit to different samples of the training data. The performance metrics that were calculated for the linear regression model were the adjusted R-squared and the root mean squared error values. The adjusted R-squared value is the R-squared value that takes into account the addition of more features. The root mean squared error is a value that is analogous to the standard deviation of the predicted values.

The cross-validation had produced 5 values for the adjusted R-squared that ranged from 0.28-0.32. This indicates that the model had a stable performance for different sections of the dataset. After training the entire model on the training set and testing it with the hold-out set the adjusted R-squared and the root mean squared error(RMSE) values were calculated to be 0.299 and 1.2769, respectively. With the simple linear regression model of the critic scores and the log of global sales, the R-squared value was calculated to be 17%. Introducing the additional features of a video game increased the R-squared value by about 13%. The range of values for log of global sales ranged from about -4 to 4. The calculated RMSE value shows that the errors of the prediction were on average about 1.28, which is a relatively large error given the range. This indicates that the linear regression model may not be able to predict the global sales of a video game very accurately. It seems that the linear regression model has room for improvement.

In order to turn this prediction problem into a classification one the dataset would need to be labeled. To keep it simple, each video game was given a binary label of whether or not it was a high selling video game. The range of values for the global sales was between 0.01-82 million units sold. It was decided that any video game that sold more than 1 million units would be high selling. The dataset for this classification problem will be very similar to the one used in the

linear regression model. However, the target variable will be the high selling labels instead of the log of global sales. Some additional features, such as user scores and count of games per publisher, were also added to the dataset.

Just like with the linear regression model the categorical features, as well as the target variable, were encoded and the dataset was split into a testing and a hold-out set. The testing set was used in some initial classifier modeling as well as hyperparameter tuning through grid search cross validation. Grid search cross validation is a method of obtaining the most optimal hyperparameters for a machine learning model by fitting a model with various sets of hyperparameters and then cross validating it to produce a score. The hyperparameters that produce the highest scores were chosen as the most optimal hyperparameters for the model, given the dataset.

The first classifier model that was looked at was logistic regression. Without tuning any of the hyperparameters and fitting and testing it with the training set, it was able to produce an accuracy score of about 79%, out-of-the-box. The accuracy of a classifier model shows the percentage of predictions the model correctly produced. A few other models such as random forest and SVM were also ran out-of-the-box and produced scores of 79% and 76%, respectively (figure 15). The logistic regression, SVM , and random forest classifiers had their hyperparameters optimized and scored 79%, 79%, and 80%, respectively. From these results it was seen that logistic regression and random forest did not change very much, but the SVM classifier was brought up to the same level as the other two in terms of accuracy score.

Classifier	Accuracy
Random Forest	0.790516
KNN	0.757751
SVM	0.756913
NaiveBayes	0.676465
Logistic Regression	0.790519

Figure 15: A table of out-of-the-box classifiers and their accuracy scores.

Logistic Regression

	precision	recall	f1-score	support
0	0.81	0.91	0.86	626
1	0.70	0.50	0.58	260
accuracy			0.79	886
macro avg	0.76	0.71	0.72	886
weighted avg	0.78	0.79	0.78	886

SVM

	precision	recall	f1-score	support
0	0.81	0.92	0.86	626
1	0.71	0.48	0.57	260
accuracy			0.79	886
macro avg	0.76	0.70	0.72	886
weighted avg	0.78	0.79	0.78	886

Random Forest

	precision	recall	f1-score	support
0	0.80	0.93	0.86	626
1	0.73	0.45	0.56	260
accuracy			0.79	886
macro avg	0.77	0.69	0.71	886
weighted avg	0.78	0.79	0.77	886

Figure 16: A classification report from the results of testing the hold-out set for logistic regression, SVM, and random forest.

One down side to accuracy is that it is not robust to unbalanced data. Take, for example, trying to predict whether or not a plane will crash. There are far more occurrences of planes not crashing than crashing. With an arbitrary ratio of 1 crash for every 100 flights, even if the classifier misclassified the 1 crash, the accuracy score would be 99%, which is not very representative of the data. Therefore some other metrics that takes this situation into account are precision, recall, and f-1. The precision of a model represents how many of the predicted positives are actually true. The recall represents the proportion of actual positives being predicted correctly. There will always be a trade-off between precision and recall. The f-1 score is the harmonic mean of precision and recall. Which metric to use for scoring the performance of a classifier depends on the dataset and the problem being solved.

For this problem of classifying video games as being high sales or not, it was determined that the precision would be the most suitable metric for scoring the model. Using the best hyperparameters obtained from the grid search cross validation as well as the hold-out set for scoring, it was seen that all three models performed very similarly to each other (figure 16). But it looks like the random forest model scored a 73% in predicting the positive class correctly and outscored both the logistic regression and SVM models by about 2%. One final way to judge how well a model performed is by looking at an ROC curve.

A Receiver Operator Characteristic (ROC) curve is a way to visualize the performance of a binary classifier model. It looks at the ratio of true positives and false negatives given varying threshold values. A threshold determines whether or not the classifier will classify a record as a certain class, given its probability of being that class. The model curve that is closer to the upper left portion of the plot indicates a good classifier. A curve that is towards the diagonal middle of the plot indicates that the classifier does no better than randomly guessing the class. The area under the curve (AUC) is a way to represent these curves numerically and allows for comparisons of performance between different models.

After plotting the curve for the three classifiers the AUCs were calculated to be 0.81, 0.80, and 0.84 for logistic regression, SVM, and random forest, respectively (figure 17). Given all the results that have been observed, it looks as though the best classifier for prediction given this dataset is the random forest model. The random forest classifier tied in best out-of-box performance, scored the highest after tuning the hyperparameters, had the highest precision in classifying the positive class, and had the highest AUC for the ROC curve. The logistic regression model had also performed well for this dataset and was a strong contender for the best classifier model.

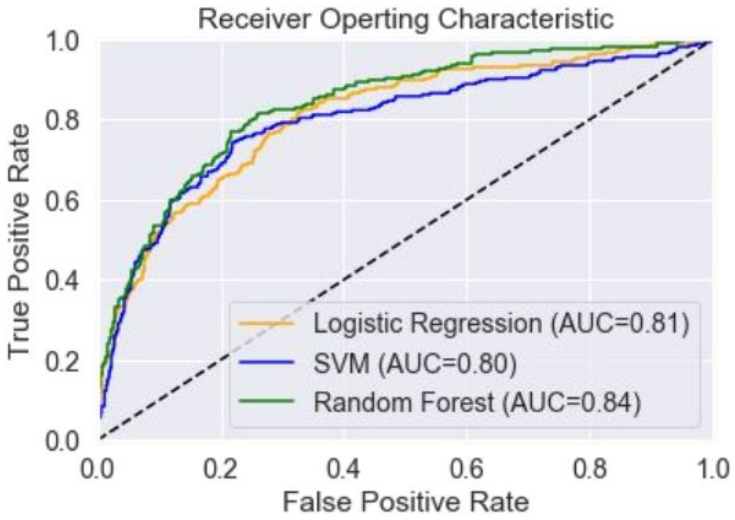

Figure 17: An ROC curve for the logistic regression, SVM, random forest classifiers.

Summary and Conclusion:

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 different features 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. It was seen that by applying machine learning and performing linear regression on many different features of a video game, the R-squared value for explaining the variance in the log of global sales was able to be increased. However, the relatively high value for the root mean squared shows that the linear regression model has room for improvement. It was also seen that by converting this problem into a classification one, a video game was able to be predicted as having a global sales value larger than 1 million units by using a random forest classifier.

One area that could have been improved on was the data cleaning and wrangling step. Many features and data points were discarded, which could have been used to increase predictive power and perhaps would have led to more analytical insights. The critic score of a video game was the chosen score for this capstone project. However, as it was seen, there was a statistically significant difference between the average values for critic score and user score. Would looking at user scores provide varying results? It was also seen that when plotting the count of games per year, there was a very big spike and a big drop. A deeper look could potentially lead to some other insights there. Another avenue that could be explored is to look at applying a neural network to this dataset. It is known that neural networks have very power predictive capabilities and may do a better job at classifying video games as being high selling.