**Mid-Semester Report**

Project Title: The prospect of the upcoming game named Black Myth: Wukong

Authors: Rui Sun, Xinyu Zhang

Team: IPhone XR

1. Executive Summary:
2. Decisions to be impacted:

The aim of our project is to give some feasible advice for the company who produced Black Myth to maximize their profits. Therefore, we need to focus on two aspects: revenue and cost, while the profit equals to the revenue minus the cost. As a result, there a re two main decisions that we want to have an impact on. First of all, we would like to offer some advisable analytics on how to properly price the game. Secondly, we want to make a list of regions where the game tends to be popular and would generate substantial profits. Since advertising and releasing the game would require fund as well, so, having such a list would help the company to reduce their cost.

1. Business value:

Since Black Myth: Wukong is the first Chinese game that reaches 3A standard, and it has already attracted huge attentions from players all over the world, the success of this game would help the entire single-player game industry in China to become prosperous again. Therefore, the decisions that we want to have an influence on would have a considerable business value.

1. Data assets:

We mainly use two data sets here, one is the sales of different games with variables regarding genres, regions, publishers and platforms. The other data set makes an extension based on the previous one, introducing reviews and ratings as another two important types of variables in terms of an analytics for games.

1. Data Processing:
2. Data Description:

One of the data set we used here is named as the sales and the ratings of video games, and here is a brief description of this data set. First of all, there are sixteen variables, including names, developers, sales, genres, ratings, the number of reviewers and the scores of those games have achieved, and mainly, we would focus on the sales, genres, scores, reviewer amount and the scores in this data set. After obtaining the data, we firstly clean out all unavailable data with the command of dropna() in python, and here is a brief description of the data set without missing values.

Moreover, here are some data visualizations for this data set. To further elaborate it, we would explain each figure one by one. First of all, the diagram1 is a heatmap for Pearson Correlations between different variables in the data set. The darker the square, the larger the correlations. For example, as shown in the diagram, the correlation between EU Sales and Global Sales is 0.94, which is quite large, and thus, the color of the square is the dark green. By this diagram, we could see the correlation between two quantified variables more directly. Secondly, the diagram2 and the diagram3 could be considered as the breakdown of details for the diagram1, showing the relationship between two specific variables. For instance, the diagram2 presents the relationship between User Scores and Critic Scores. As we could see from the diagram, there is a positive correlation between those two variables, which is in line with the logic. This is because the quality of a game is objective to some extent, meaning that a game that has gained a large amount of good comments among users would also win a favorable reception from professional reviewers.

Furthermore, the diagram4 group different games in accordance with their genres, and sum all the scores that a type of games has achieved as well as the number of reviewers that have made comments on this kind of games. By forming such a diagram, we could quantify the reputation and popularity for a genre of games to some extent. In addition, the bar chart1 involves three variables and shows the sales of games with various ratings (from for teenagers to adult-only) on three different platforms, which are windows, PS3 and X360. This chart would give us some direct hint on which platform is suitable for Black Myth to launch.

Apart from the data set mentioned above, there is another data set named as video game sales dataset. First, the data cleaning has been done. The unnecessary data has been omitted and a new attribute “Global Sales” has been calculated and added to the dataset. Therefore, after cleaning the data, the dataset contains 12 attributes and the type of each attribute has been shown in Figure2. Next, the relationship between attributes has been visualized. The correlation of each attribute has been found. According to the correlation matrix, the year and sales in different regions has negative correlation, and the sales in North America and PAL play an important role in total sales (Figure2).

For the genre analysis, the genre of Misc, action and sports have the greatest number of games in their genre. And the sports, action and shooter have the most global sales. Besides, the genre of action and sports has sold the most in a single year from 1970 to 2018 (Figure3). Since genre is a categorical variable, the statistical method of detecting outliers is suitable for genre. According to Figure 3(a), the sandbox and education games are extremely few in the dataset, which could affect the model. Therefore, those two genres could be considered as outliers based on the further analysis. Besides, according to the sales comparison of different genres in different regions, North America and PAL sell more games in all genres, and action, shooter and sports sell more in all the regions (Figure 4). The board game and education sell nearly zero in all the regions; therefore, it could be considered as outliers. According to the analysis of platforms, PS2, PS3 and X360 are the most popular game platforms (Figure 5). The IQue and CI28 could be considered as outliers since they do not contribute to the sales but will affect the machine learning model.

1. Data cleaning and outlier detection:

As mentioned above, we use dropna() to clean all Not Available data in the data set. Also, we have introduced some simple outlier detection methods at current stage, and we would try to explore more appropriate approaches to wipe out the outliers in the near future. First of all, we use boxplot to find the outliers for Critic Scores, Critic Counts, User Scores and User Count, one of the boxplots is presented as graph1. Secondly, we use scatter plots to locate the outliers (as shown in graph2). The reason why we use these two methods is that by graphs, we could see outliers more directly.

Apart from that, we also try some statistical methods to find the outliers. One method we apply here is using the z-score. After obtaining the z-score by numpy package in python, we could set a threshold to search out the outliers in our data set. Another method we use at present is IQR value, which is similar to z-score, as long as we get the value of IQR, we could set restrictions to the significance level we prefer, and then clean out the outliers.

To explain it further, for the outliers, since most of the attributes in our dataset are not continuous except sales, statistical approaches of outlier detection are not suitable. Sales do not have a specific confidence interval because the number of sales of a game can be very different from the other game. Our dataset is not 2D, and there are more than one factors that influence sales, so the depth-based approach is not applicable. In terms of the deviation-based approach, there is not a series of similar data in our dataset and each category of the dataset is independent, so it might be inefficient to use this outlier detection approach. Therefore, the distance-based approach and the density-based approach are selected as our outlier detection approaches.

Mahalanobis Distance is selected because the dataset is multi-dimensional, and there are several factors impacting sales. Multivariate outliers can be detected by Mahalanobis Distance as well. RKOF is another efficient approach to detecting outliers in our dataset. Since our data is clustering and has multidimensional domains, RKOF can provide accurate results by comparing density estimation to the density of neighboring observations. The reason why KNN is not used is that given the size of our dataset, the calculation required for KNN would be enormous, which makes it an inefficient approach for detecting outliers for our dataset. Due to the technical issue of finding the suitable outlier packages, the outlier detection part will be finished in the next week.

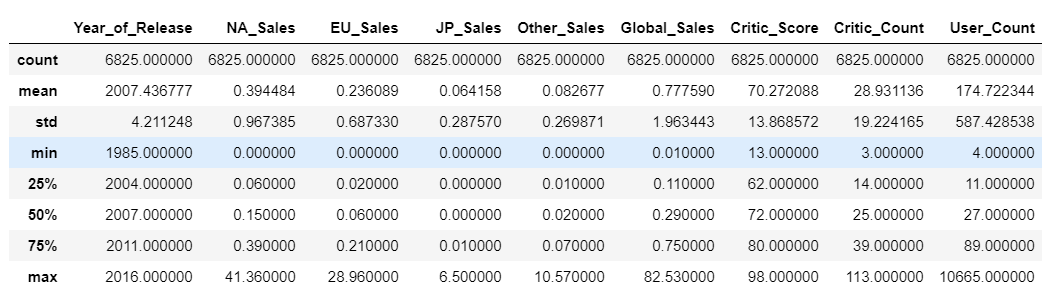


Figure1. A brief description for data set 1

|  |  |
| --- | --- |
| (a) | (b) |
| Figure 2. Data Summary.  (a) Information of Video Game Dataset after data cleaning.  (b) Correlation Matrix of all attributes | |

|  |  |
| --- | --- |
| (a) | (c) |
| (b) | |
| Figure 3. Genre Analysis. (a) Number of Games in each Genre. (b) The genre with the highest number of sales in each year. (c)Global Sales of each genre. | |

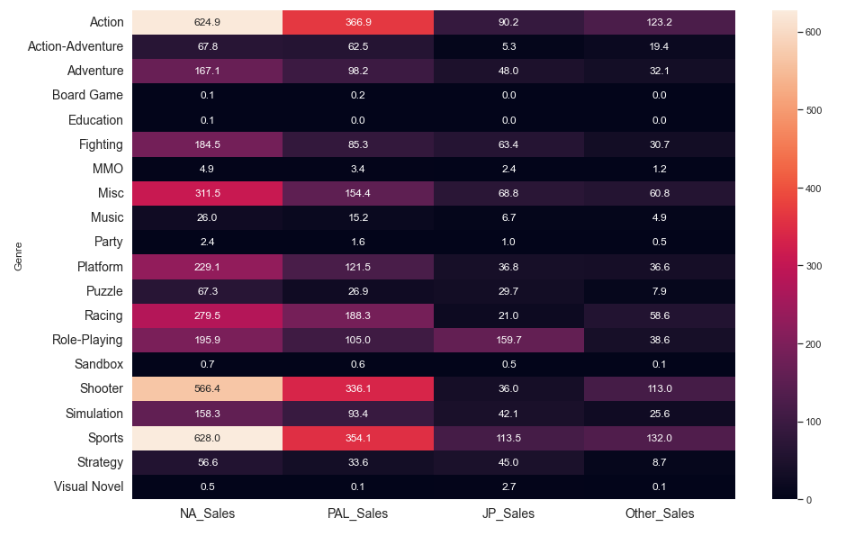


Figure 4. Sales Comparison by Genre.

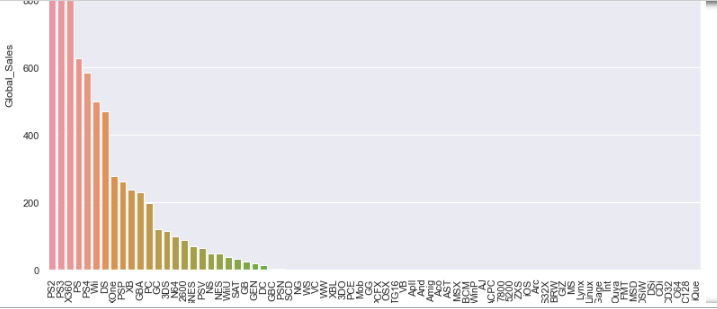


Figure 5. The Global Sales of different Platforms

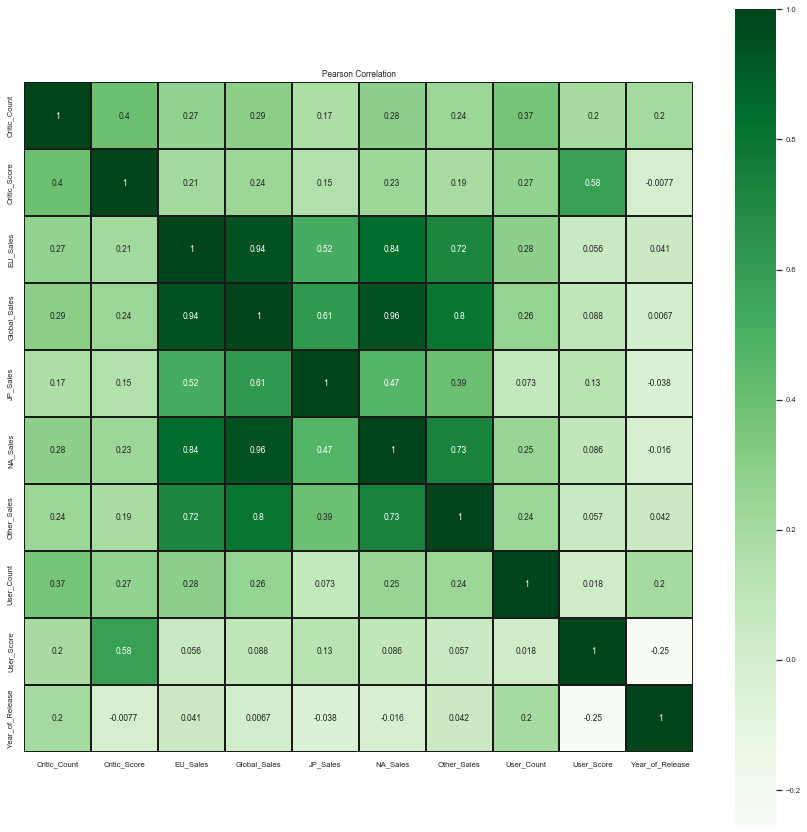


Diagram 1. Pearson Correlations for multiple variables

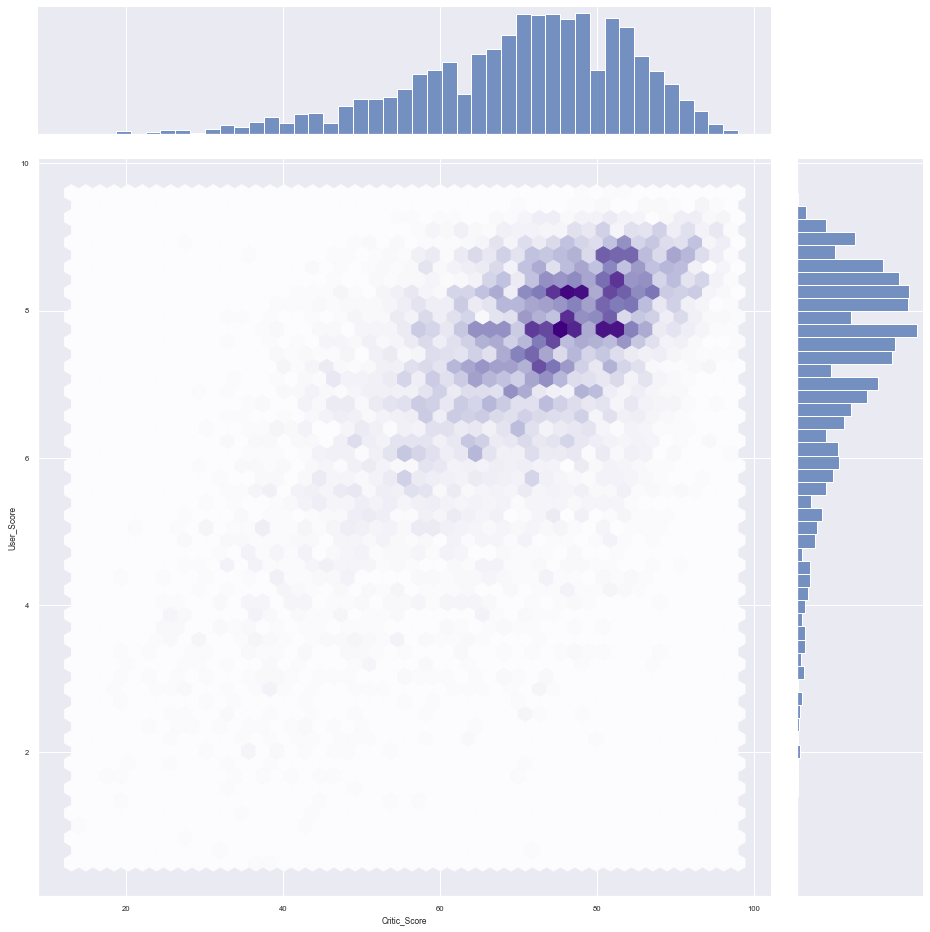


Diagram 2. Pearson Correlation between Critic Scores and User Scores

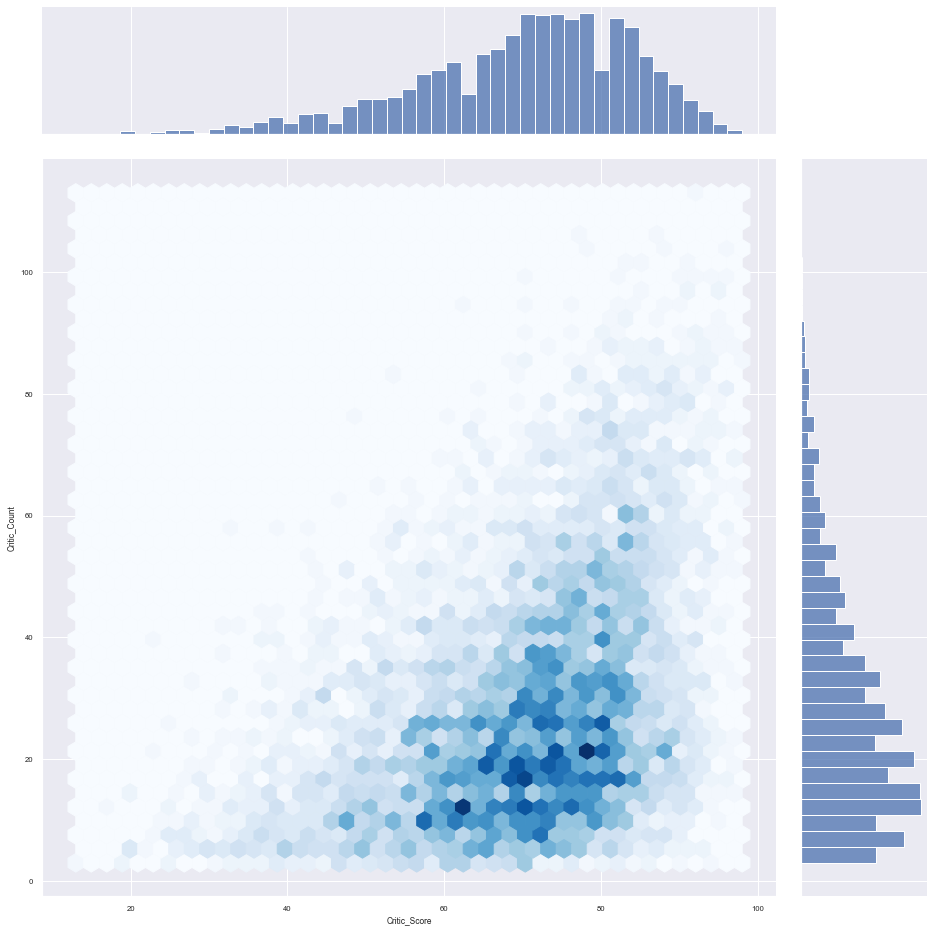


Diagram 3. Pearson Correlations between Critic Score and Critic Count

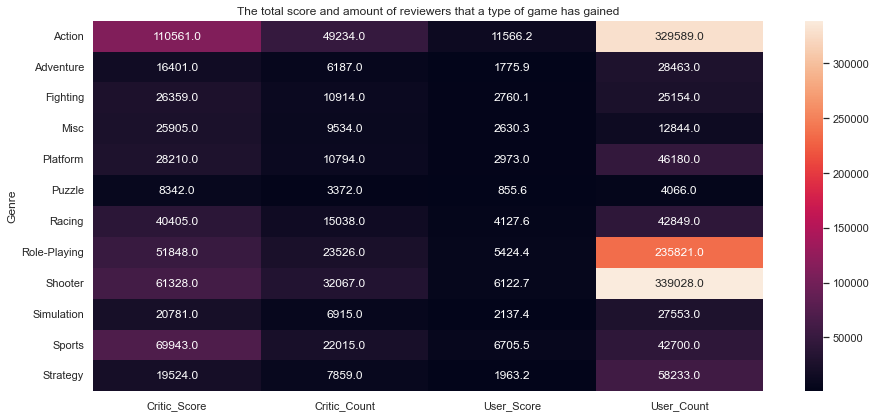


Diagram 4. Total score and number of reviewers for different genres

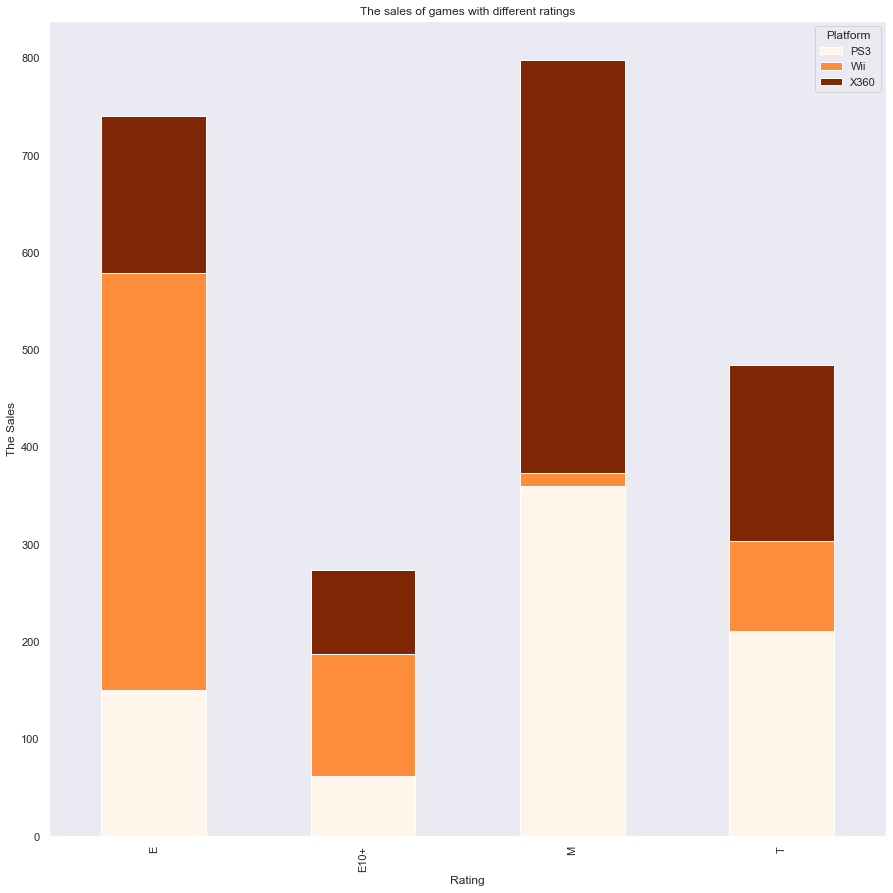
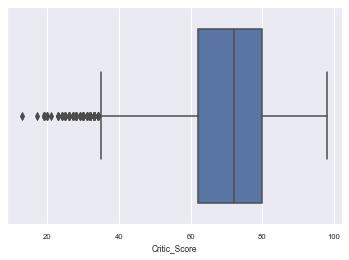
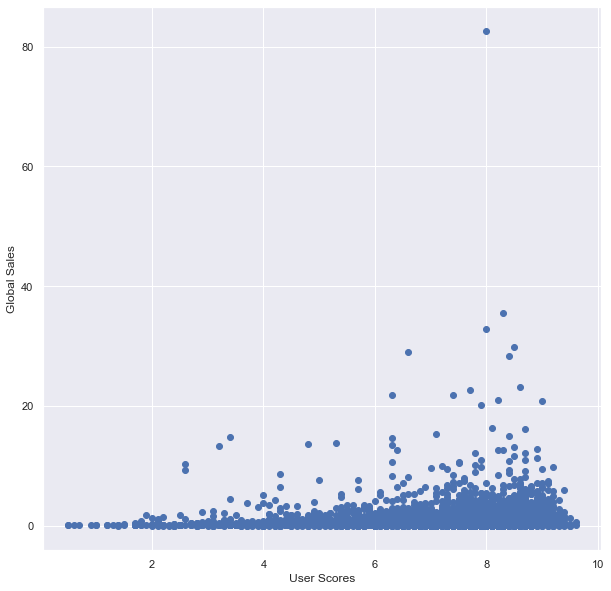


Chart 1. The sales of multi-rating games in different platforms



Graph 1. Boxplot for Critic Scores



Graph 2. Scatter Plot for User Scores and Global Sales

1. Model Updates:
2. Hypothesis Test

The hypothesis test is a suitable method to check the dependency between two variables in one attribute. Playstation 3 and XBox 360 are two of the consoles which are in the same generation. As per reports Playstation 3 has more number of units sold when compared with XBox 360 (87.4 million units vs 84 million units). However in 2015, both Microsoft and Sony stopped disclosing the sales figure. So by now,their sales figures might be almost equal. Since the number of units of each unit might be comparable now, let's hypothesize that, for a game, which is available in both platforms, PS3 will have the same game sales when compared with X360.

Null Hypothesis: A game which is available in both platforms, PS3 and X360 will have comparable sales.

Alternate Hypothesis: Sales would not be equal and will be independent.

The hypothesis test can be used for checking the dependency of sales in different regions. For example, 2K Games is an American video game publisher. Although video games published in Japan are mostly popular in North America, the vice versa may not be true.

Null Hypothesis: Sales of same games by 2K Games are dependent.

Alternate Hypothesis: Sales are independent

1. Regression Model and Random Forest Model

Most of the video game sales are in the North American Region and PAL region. Only the popular games perform well in other areas. In the correlation matrix below, we can see that there is a relation between sales in the PAL region and other regions (Figure 6). Therefore, using regression models, we might be able to predict sales of a game in these regions.

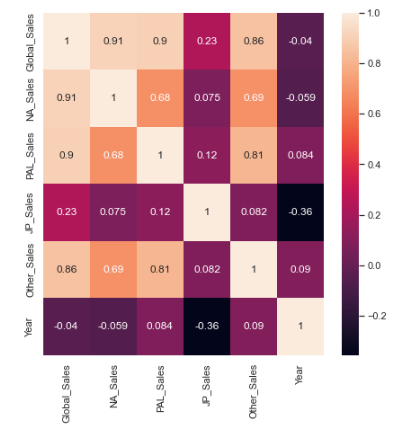


Figure 6. Correlation Matrix of Sales in Different Regions.

In the regression model, the cross validation linear regression model and OLS linear regression model are planned to be used for comparison of accuracy of linear regression model. And it is possible to choose a polynomial regression model with different degrees and come up with a better regression model.

Besides, we still want to be able to construct a model of how good a game will sell. One of the strongest continuous variables that affects the sales of a game is the year. We can first try to fit a Linear Regression model of video game sales over the years. For this, we will be using Sci-Kit Learn. In this part, we are not going to split the data because we want to take the entire dataset into account. Given the spread of the data, we do not need to worry about overfitting the linear regression model. If the prediction of the linear regression model does not have good performance. Random forest regressor would be our next plan.

A Random Forest Regressor is a forest of multiple decision trees that predict values based on decisions. A decision tree would make predictions by going through each node's decision, following the paths based on those decisions, until finally landing at a leaf node, which represents a prediction. The problem is that a singular tree can oftentimes be too brittle. If one were to make a decision tree model, one would have to figure out the decisions that have to be made first: the most important and significant decisions. For us to be able to figure out what the most significant decision is, we would have to maximize information gain. However, this will lead to being more prone to overfitting the model. Therefore, our solution to this is to use multiple trees.

To make sure the trees in our random forest are not all equally brittle, we can use a certain subset (with overlap) of our data for each tree we make. Sci-Kit Learn has a way of not only doing this, but also allowing us to specify the maximum depth of each tree, the number of trees in our forest, and the number of samples we will allow for the making of each tree. For our purposes, we will be using 5 trees and a maximum of 3000 out of the ~9000 samples for each tree, since the module's default of 100 trees would have been really expensive to compute and would not have made that much of a difference. We will not be specifying the maximum depth of each tree simply because we want the leaf nodes to be pure values.

1. Source Code:

We are so sorry that we fail to upload the code onto github, we include codes in our zip file.

1. Next Steps:
2. 10.15-10.17, we plan to prepare our presentation for the mid-term report.
3. 10.17-10.27, we intend to find more suitable outlier detection methods for our data sets here. Since during these days, we are going to finish our home assignment 2, which is a bunch of exercises for all kinds of outlier detection. Thus, we think it would be nice for us to conduct those two assignments in the same time.
4. 10.20-11.11, we would try to find a reasonable and feasible model to conduct our data analytics. Moreover, we would try to find the optimal result that we expect to get, such as a suitable pricing for Black Myth as well as the list of regions that are suitable to release the game, with the help of the machine learning applications.
5. After 11.12, we would pay efforts in finishing our final report and final presentation, trying to make our conducts more practical and persuasive.