527 - Final Report

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

Authors: Rui Sun, Xinyu Zhang

Team: IPhone XR

1. Executive Summary:
   a. 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: sales volume and retail price. As a result, there are two main decisions that we want to have an impact on. First of all, we intend to estimate the amount of sales this game could achieve. Secondly, we would like to offer some advisable analytics on how to properly price the game.

   b. 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.

   c. Data assets:

      We mainly use two datasets here, one is the sales of different games with variables regarding genres, regions, publishers and platforms. The other dataset 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.

2. Data Processing:
   a. Data Description:

      One of the datasets we used here is named as the sales and the ratings of video games, and here is a brief description of this dataset. 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 dataset. 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 dataset without missing values.

Moreover, here are some data visualizations for this dataset. 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 dataset. 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.

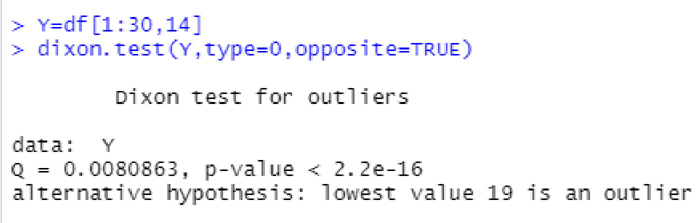
Apart from the dataset mentioned above, there is another dataset named as video game sales dataset. First, the data cleaning has been done. The unnecessary data has been omitted and a new attribute "Global Sales" has been calculated and added to the dataset. Therefore, after cleaning the data, the dataset 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.

b. Data cleaning and outlier detection：

As mentioned above, we use dropna() to clean all Not Available data in the dataset. Moreover, In order to filter out outliers, we choose user_count (the number of user comments of a specific game) as our beacon to conduct multiple kinds of outlier detections. First of all, we try to use statistical approaches to filter out the user_count with extreme small values. This is because we would use user_socre (the average score of a game obtained from user evaluation) as one of the indicators in the future prediction of game sales. Therefore, if the user_count is extremely small, which means only a few people have left comments and rated the game, then that corresponding user_score tends to be meaningless. To elaborate it further, we use user_score as an indicator for the quality of the game, but if there is only one person who rated the game, then the score of this game would be extremely subjective, which in other words, this score would not be qualified as a reference of the quality of that game under that circumstance. Consequently, we conduct the statistical outlier detection for user_count of our dataset, and below are the results.

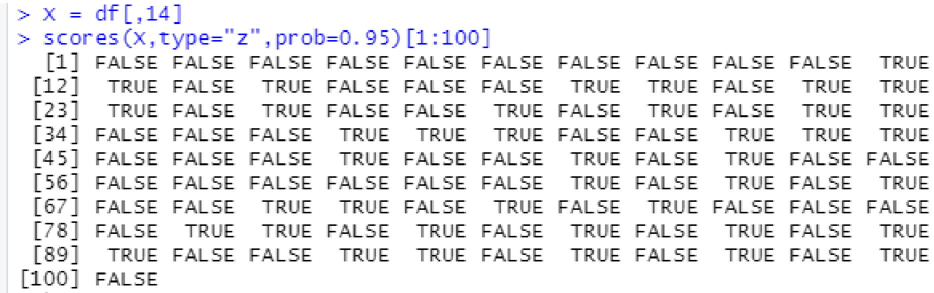
1. Dixon test (small sample size)



```
> Y=df[1:30,14]
> dixon.test(Y,type=0,opposite=TRUE)

        Dixon test for outliers

data:  Y
Q = 0.0080863, p-value < 2.2e-16
alternative hypothesis: lowest value 19 is an outlier
```

We do not choose this method because it can only be used by small sample size (e.g., 30), and we have thousands of data to process.

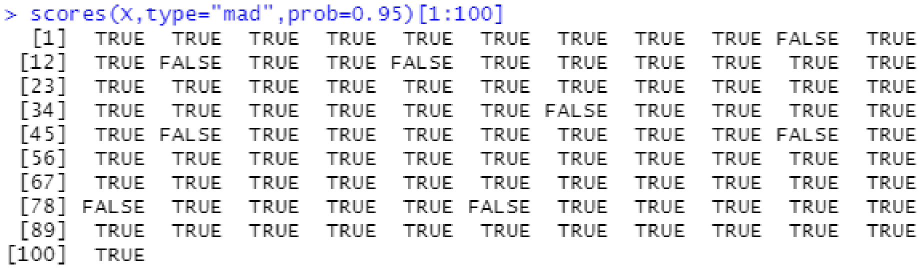
2. Normal score (Deviation with respect to the mean)



```
> X = df[,14]
> scores(X,type="z",prob=0.95)[1:100]
  [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE
 [12]  TRUE FALSE  TRUE FALSE FALSE FALSE  TRUE  TRUE FALSE  TRUE  TRUE
 [23]  TRUE FALSE  TRUE FALSE FALSE  TRUE FALSE  TRUE FALSE  TRUE  TRUE
 [34] FALSE FALSE FALSE  TRUE  TRUE  TRUE FALSE FALSE  TRUE  TRUE  TRUE
 [45] FALSE FALSE FALSE  TRUE FALSE FALSE  TRUE FALSE  TRUE FALSE FALSE
 [56] FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE  TRUE FALSE  TRUE
 [67] FALSE FALSE  TRUE  TRUE FALSE  TRUE FALSE  TRUE FALSE FALSE FALSE
 [78] FALSE  TRUE  TRUE FALSE  TRUE FALSE  TRUE FALSE  TRUE FALSE  TRUE
 [89]  TRUE FALSE FALSE  TRUE  TRUE FALSE  TRUE FALSE  TRUE FALSE  TRUE
[100] FALSE
```

We do not choose this method because normal score approach would introduce a threshold like 3, for any z-score larger than 3 would be considered as an outlier. However, this kind of deviation is double-tailed, which means we not only filter out the user_count with extreme small values, but also those with extreme large values, and normally, those game would be considered as the most popular games that have great importance to our analytics here.

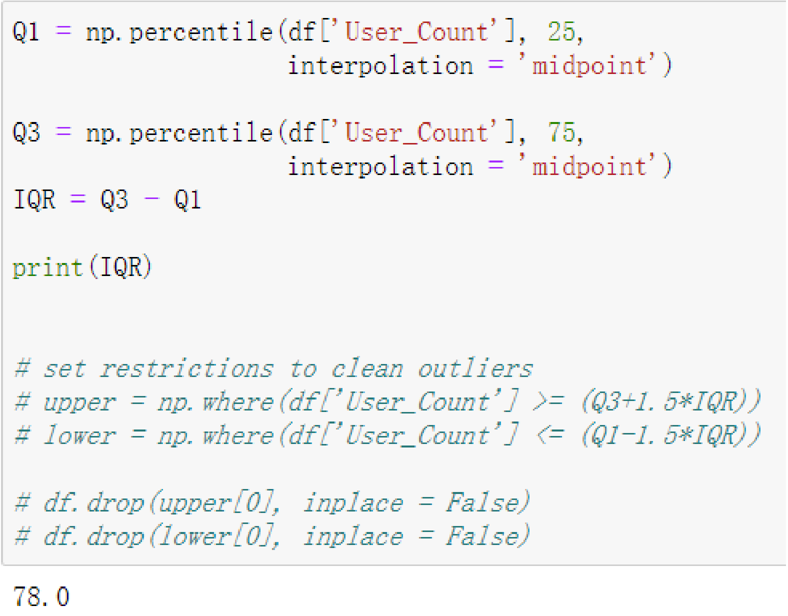
3. Median Absolute Deviation (Deviation with respect to the median)



```
> scores(x,type="mad",prob=0.95)[1:100]
  [1]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE
 [12]  TRUE FALSE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
 [23]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
 [34]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE
 [45]  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE
 [56]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
 [67]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
 [78] FALSE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE
 [89]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
[100]  TRUE
```

We do not choose this method for the same reason that mentioned above in the Normal score method.

4. Interquantile range score



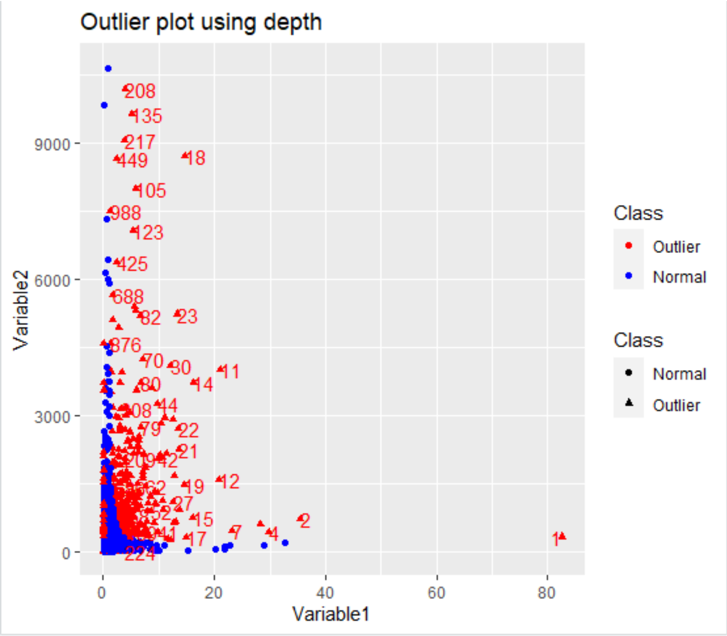
```
Q1 = np.percentile(df['User_Count'], 25,
                   interpolation = 'midpoint')

Q3 = np.percentile(df['User_Count'], 75,
                   interpolation = 'midpoint')
IQR = Q3 - Q1

print(IQR)


# set restrictions to clean outliers
# upper = np.where(df['User_Count'] >= (Q3+1.5*IQR))
# lower = np.where(df['User_Count'] <= (Q1-1.5*IQR))

# df.drop(upper[0], inplace = False)
# df.drop(lower[0], inplace = False)
```

```
78.0
```

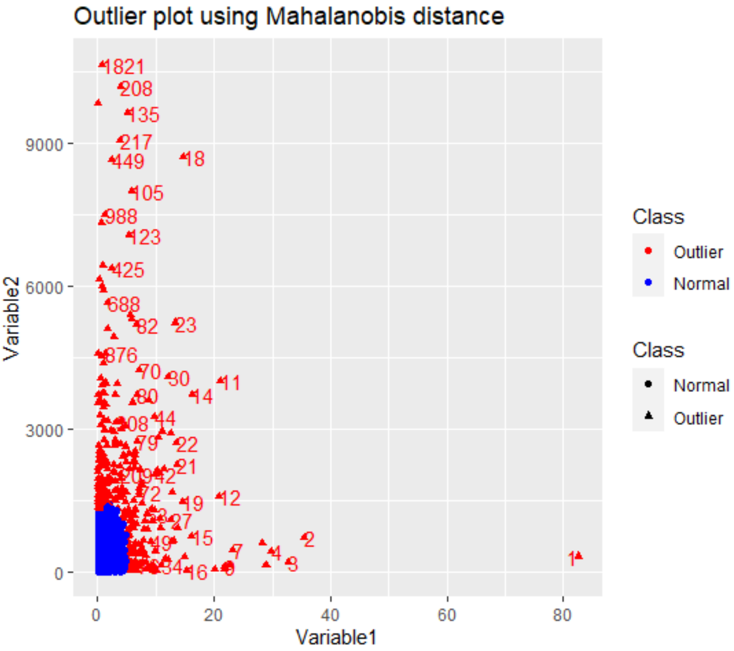
We would choose this method since we can just drop the lower tail.

Also, we conduct some other outlier detections. In this part, we introduce another variable: global_sales, which is the global sale volume of specific games accordingly. We select these two variables based on the following logic. A popular game that sells well tends to attract a lot of users to comment on that game and give the game a score, therefore, if a game with significant sales volume only has a few reviews from users, we tend to regard this as unnormal, and vice versa. Consequently, we choose global_sales and user_counts as two variables of our further outlier detection.

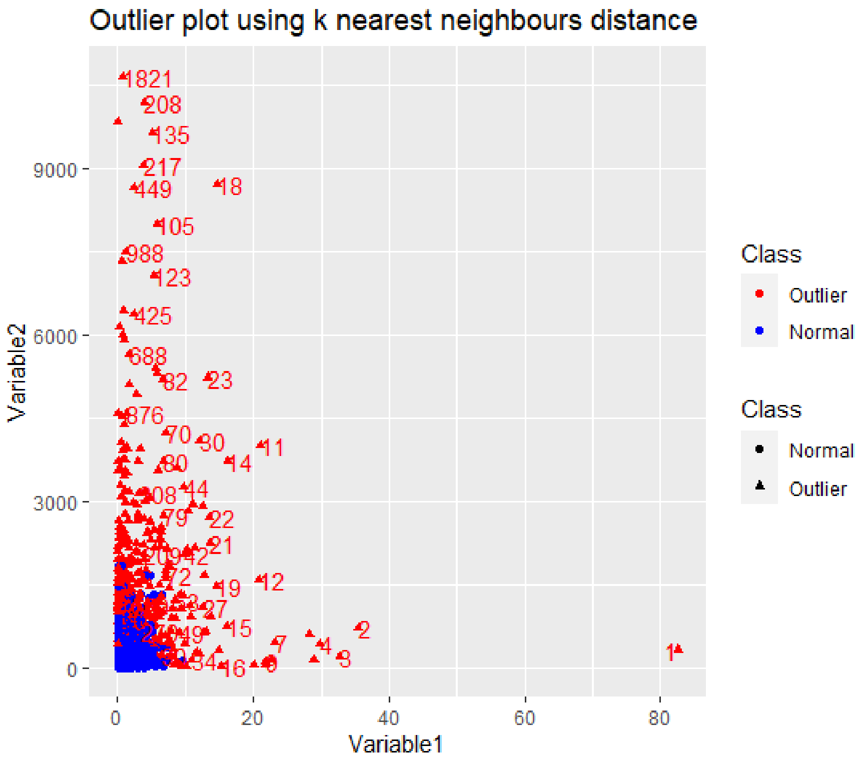
5. Depth-based Approach



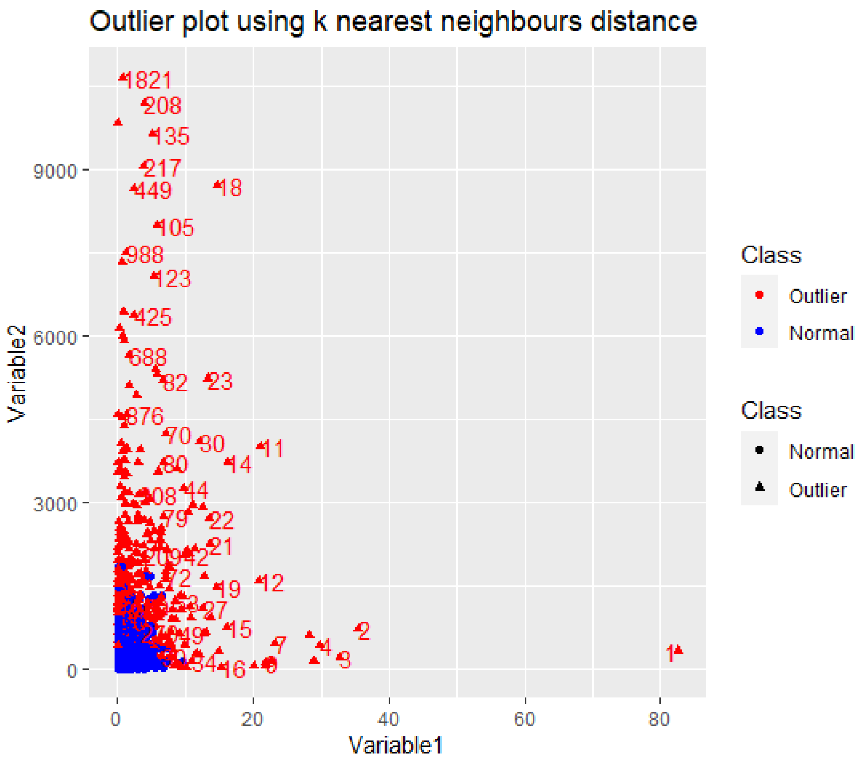
6. Outlier detection using Mahalanobis Distance



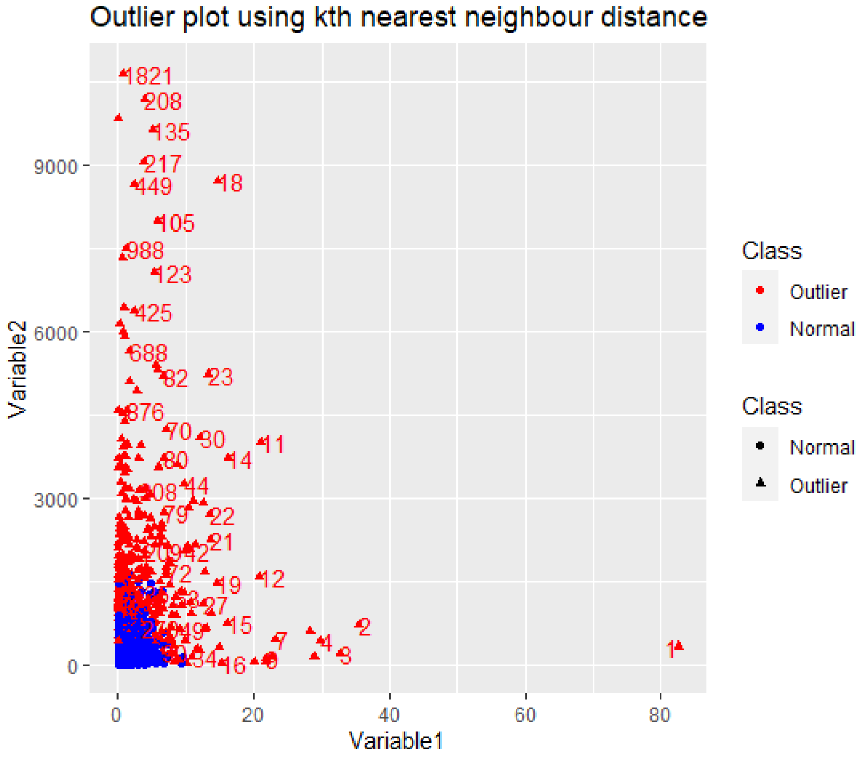
7. Outlier detection using k Nearest Neighbours Distance method



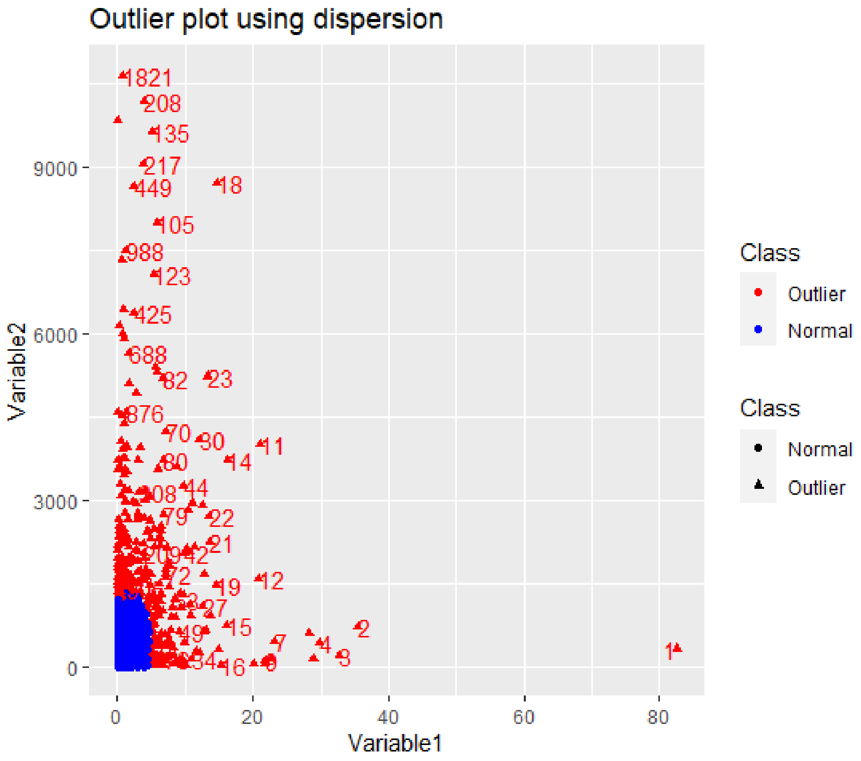
8. Outlier detection using kth Nearest Neighbour Distance method

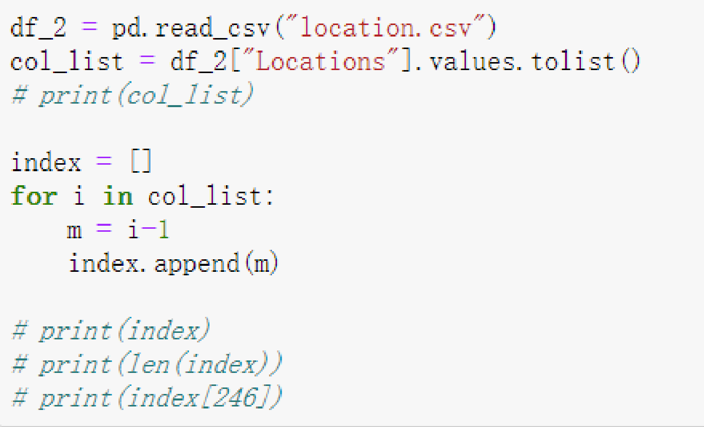


9. Outlier detection using genralised dispersion

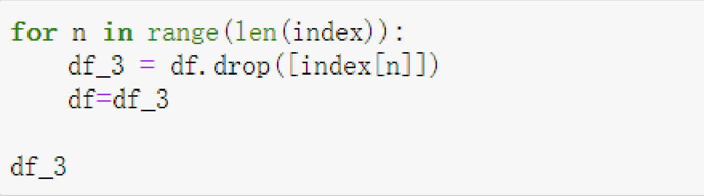


10. Joint assessment of outlier detection



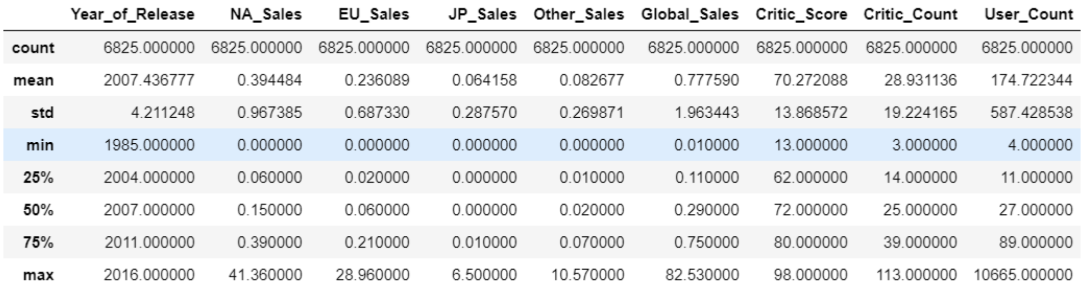


```
df_2 = pd.read_csv("location.csv")
col_list = df_2["Locations"].values.tolist()
# print(col_list)

index = []
for i in col_list:
    m = i-1
    index.append(m)

# print(index)
# print(len(index))
# print(index[246])
```



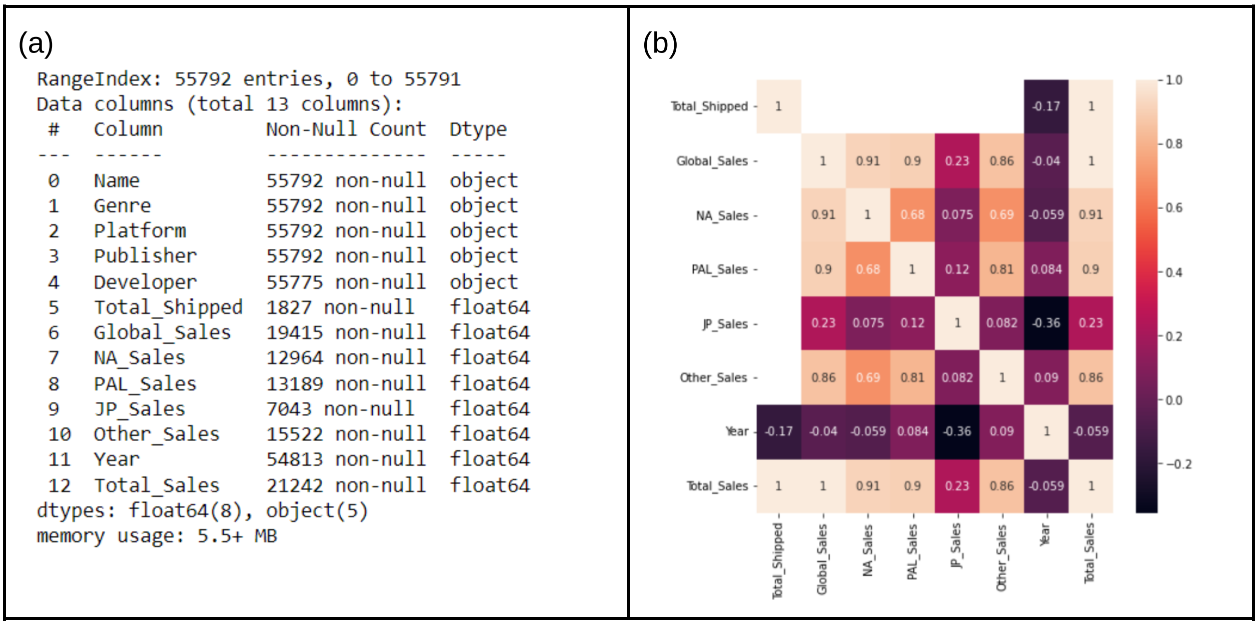
```
for n in range(len(index)):
    df_3 = df.drop([index[n]])
    df=df_3

df_3
```

We choose this joint assessment of outlier detection since this method could be considered as a combination of 5-9 above, which is more straightforward and powerful. After obtaining the locations of those outliers, we drop the rows that contain them.



	Year_of_Release	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales	Critic_Score	Critic_Count	User_Count
count	6825.000000	6825.000000	6825.000000	6825.000000	6825.000000	6825.000000	6825.000000	6825.000000	6825.000000
mean	2007.436777	0.394484	0.236089	0.064158	0.082677	0.777590	70.272088	28.931136	174.722344
std	4.211248	0.967385	0.687330	0.287570	0.269871	1.963443	13.868572	19.224165	587.428538
min	1985.000000	0.000000	0.000000	0.000000	0.000000	0.010000	13.000000	3.000000	4.000000
25%	2004.000000	0.060000	0.020000	0.000000	0.010000	0.110000	62.000000	14.000000	11.000000
50%	2007.000000	0.150000	0.060000	0.000000	0.020000	0.290000	72.000000	25.000000	27.000000
75%	2011.000000	0.390000	0.210000	0.010000	0.070000	0.750000	80.000000	39.000000	89.000000
max	2016.000000	41.360000	28.960000	6.500000	10.570000	82.530000	98.000000	113.000000	10665.000000

Figure1. A brief description for dataset 1



(a)

```
RangeIndex: 55792 entries, 0 to 55791
Data columns (total 13 columns):
 #   Column         Non-Null Count  Dtype  
---  ------         --------------  -----  
 0   Name           55792 non-null  object 
 1   Genre          55792 non-null  object 
 2   Platform       55792 non-null  object 
 3   Publisher      55792 non-null  object 
 4   Developer      55775 non-null  object 
 5   Total_Shipped  1827 non-null   float64
 6   Global_Sales   19415 non-null  float64
 7   NA_Sales       12964 non-null  float64
 8   PAL_Sales      13189 non-null  float64
 9   JP_Sales       7043 non-null   float64
 10  Other_Sales    15522 non-null  float64
 11  Year           54813 non-null  float64
 12  Total_Sales    21242 non-null  float64
dtypes: float64(8), object(5)
memory usage: 5.5+ MB
```

(b)

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

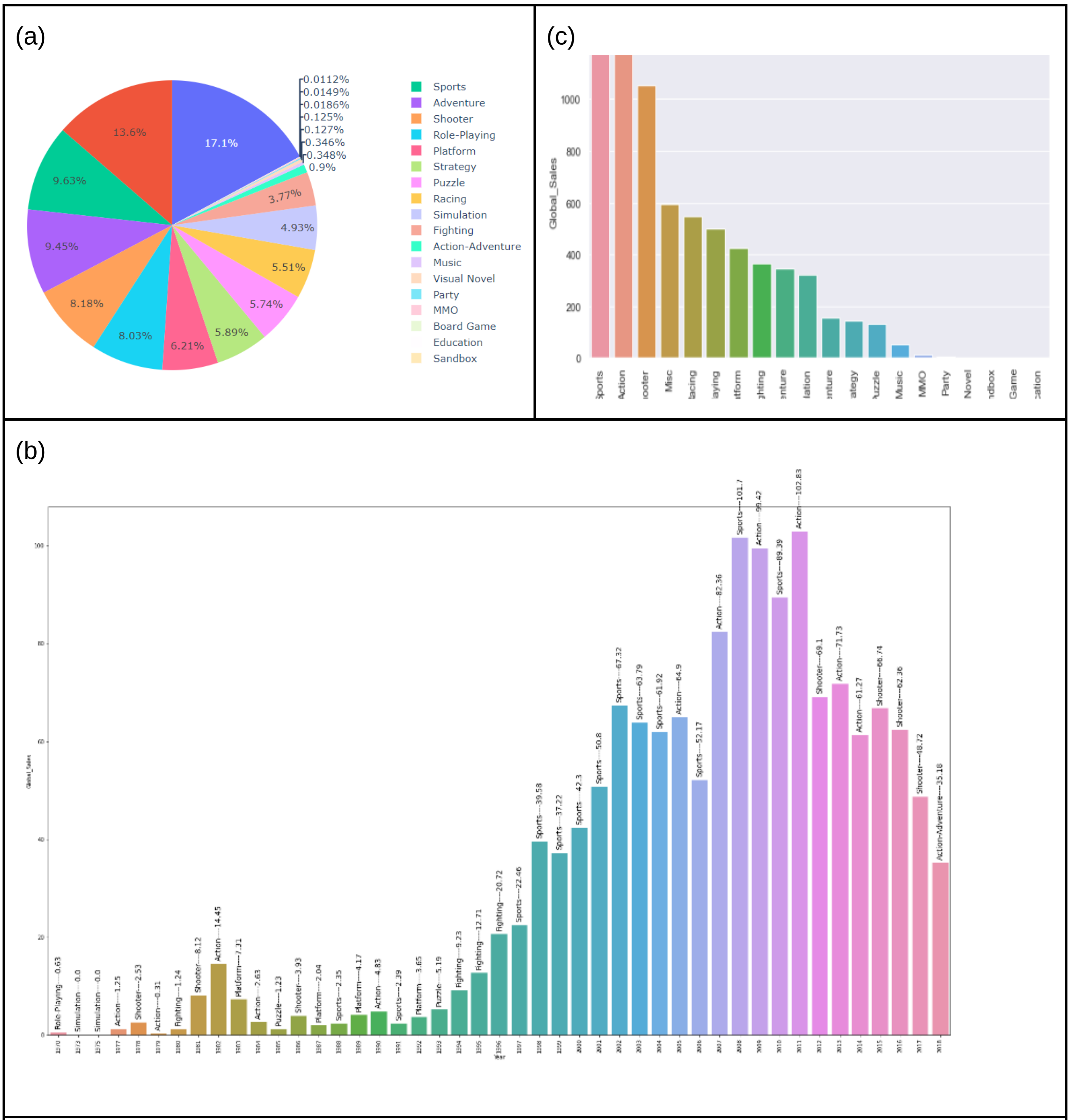


Figure 3. Genre Analysis. (a) Number of Games in each Genre. (b) The genre with the highest number of sales in each year. (c)Global Sales of each genre.

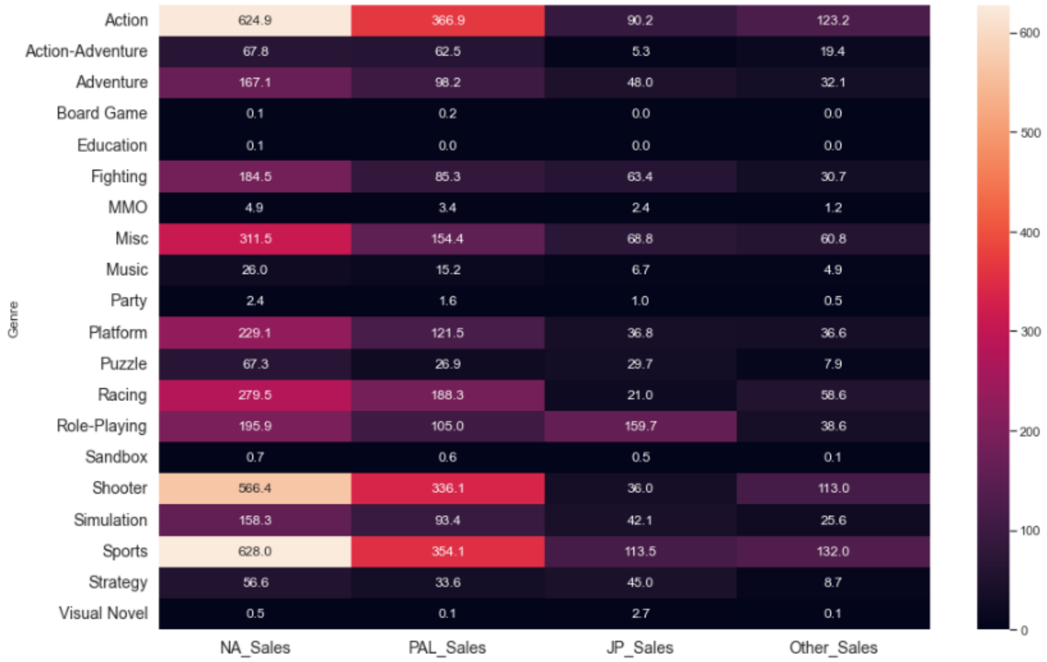


Figure 4. Sales Comparison by Genre.

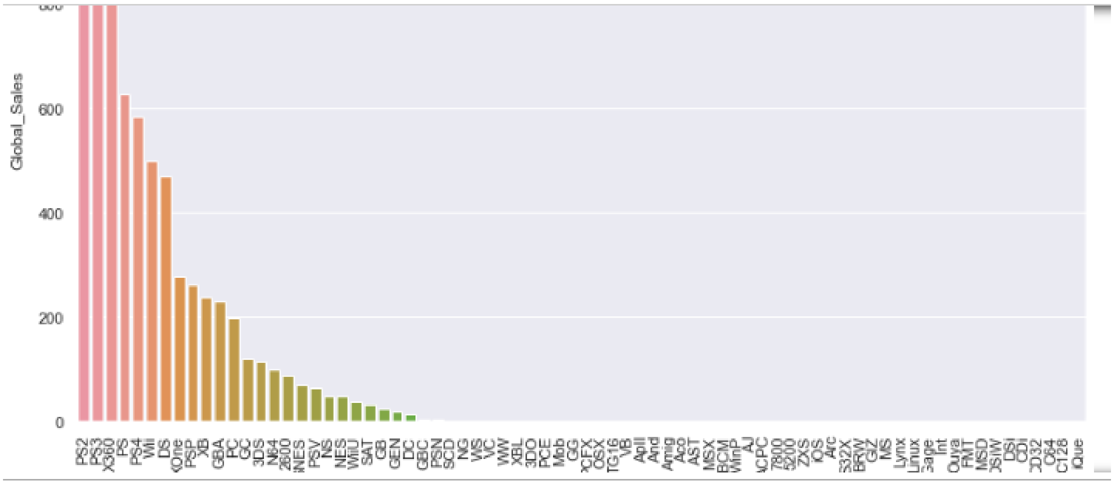
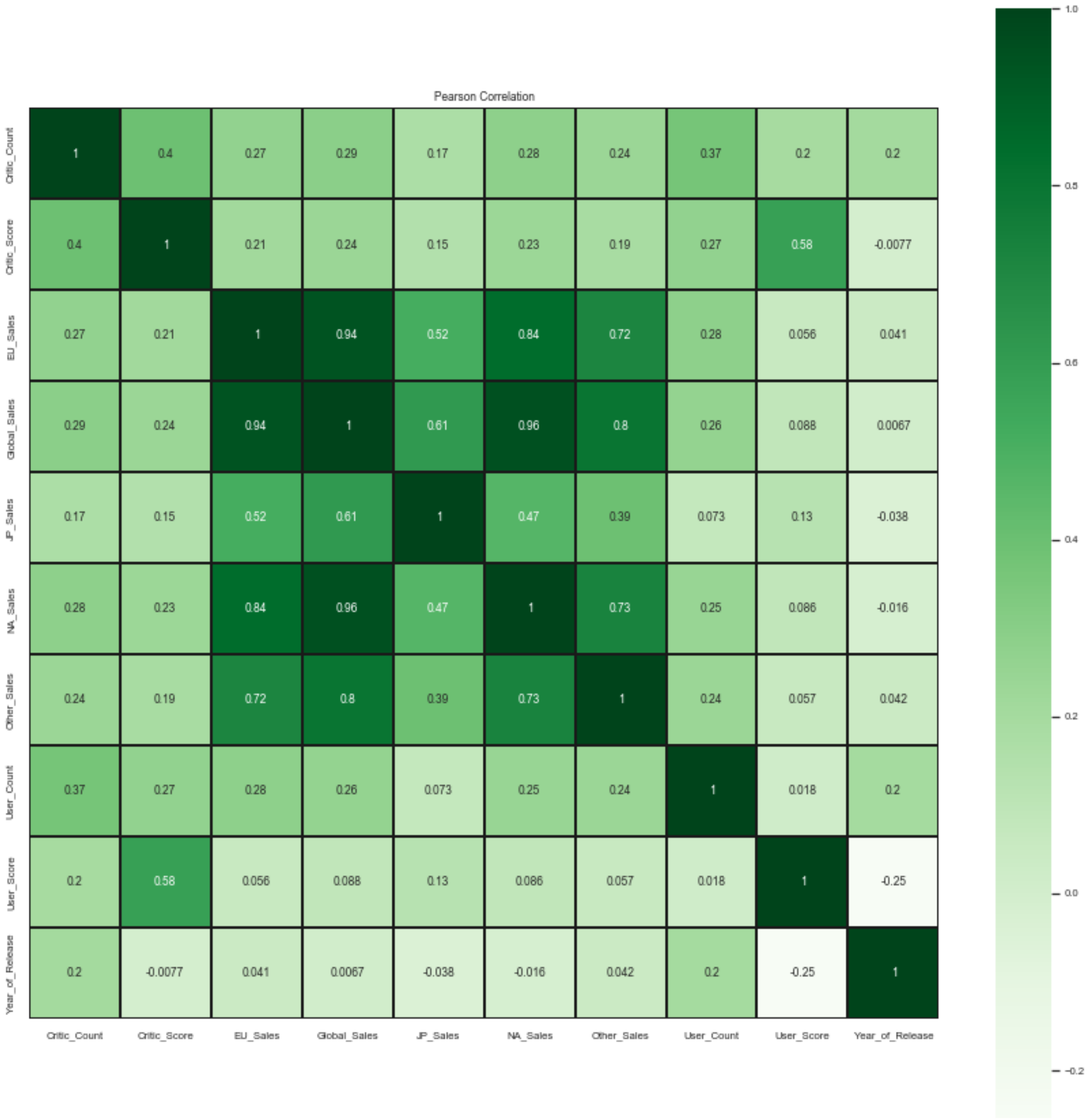


Figure 5. The Global Sales of different Platforms



	Critic_Count	Critic_Score	EU_Sales	Global_Sales	JP_Sales	NA_Sales	Other_Sales	User_Count	User_Score	Year_of_Release
Critic_Count	1	0.4	0.27	0.29	0.17	0.28	0.24	0.37	0.2	0.2
Critic_Score	0.4	1	0.21	0.24	0.15	0.23	0.19	0.27	0.58	-0.0077
EU_Sales	0.27	0.21	1	0.94	0.52	0.84	0.72	0.28	0.056	0.041
Global_Sales	0.29	0.24	0.94	1	0.61	0.96	0.8	0.26	0.088	0.0067
JP_Sales	0.17	0.15	0.52	0.61	1	0.47	0.39	0.073	0.13	-0.038
NA_Sales	0.28	0.23	0.84	0.96	0.47	1	0.73	0.25	0.086	-0.016
Other_Sales	0.24	0.19	0.72	0.8	0.39	0.73	1	0.24	0.057	0.042
User_Count	0.37	0.27	0.28	0.26	0.073	0.25	0.24	1	0.018	0.2
User_Score	0.2	0.58	0.056	0.088	0.13	0.086	0.057	0.018	1	-0.25
Year_of_Release	0.2	-0.0077	0.041	0.0067	-0.038	-0.016	0.042	0.2	-0.25	1

Diagram 1. Pearson Correlations for multiple variables

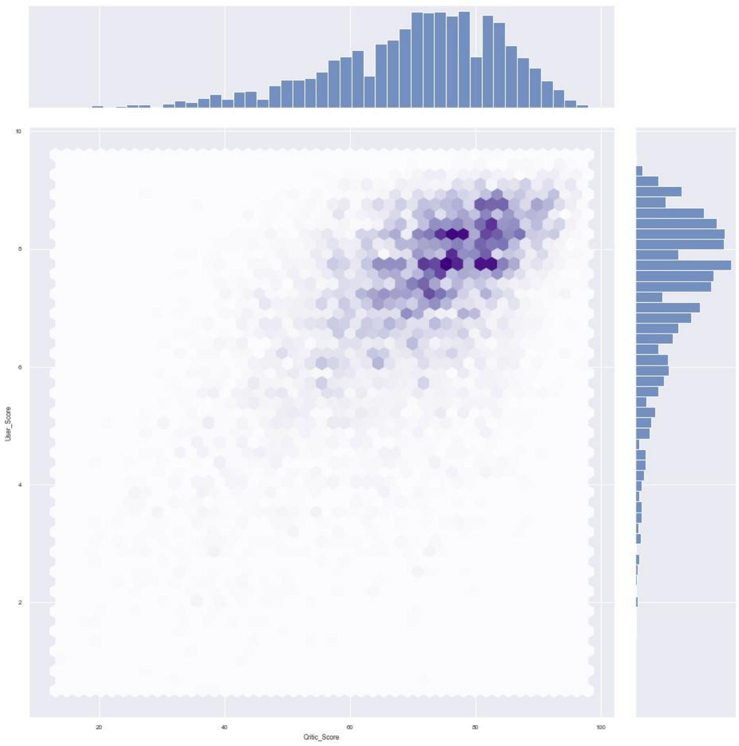


Diagram 2. Pearson Correlation between Critic Scores and User Scores

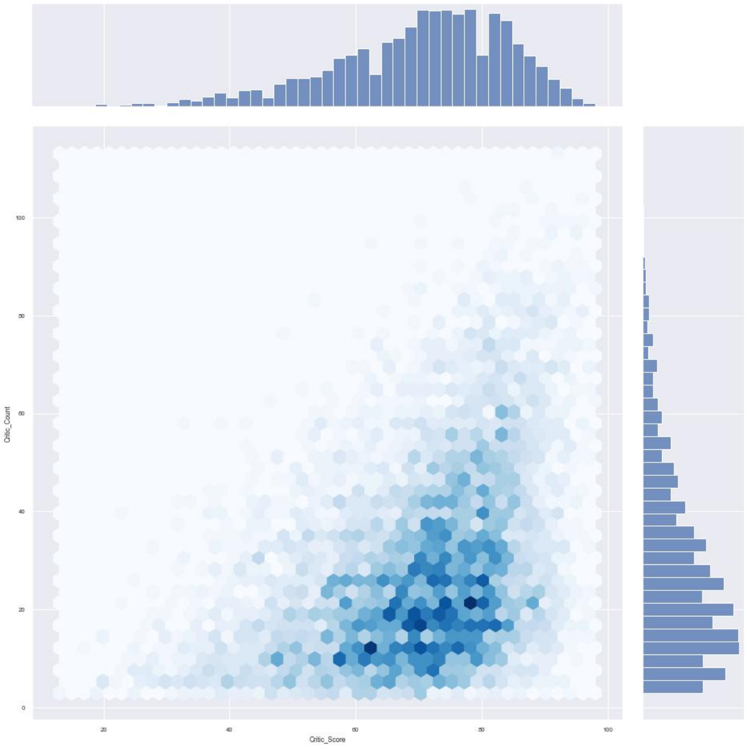


Diagram 3. Pearson Correlations between Critic Score and Critic Count

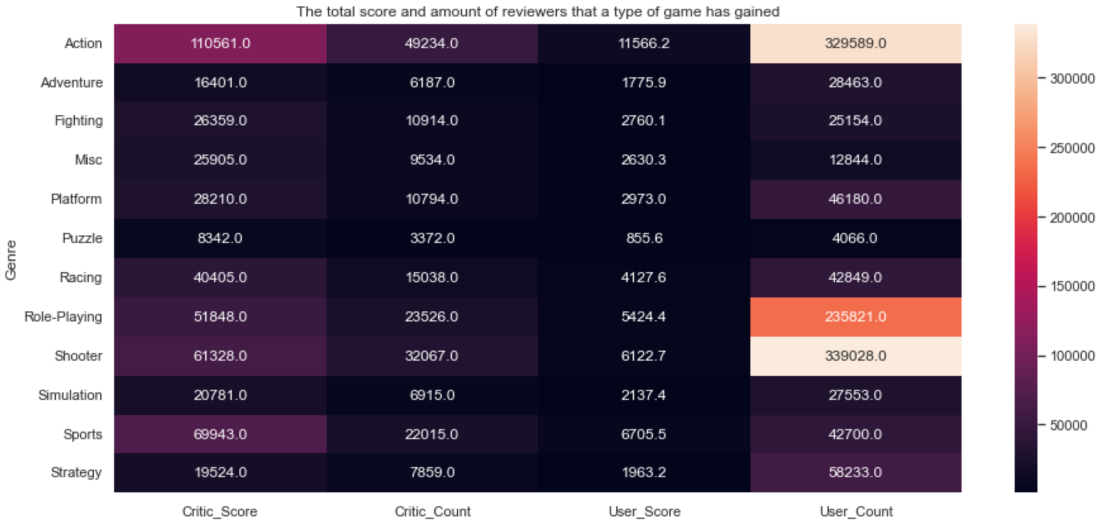


Diagram 4. Total score and number of reviewers for different genres

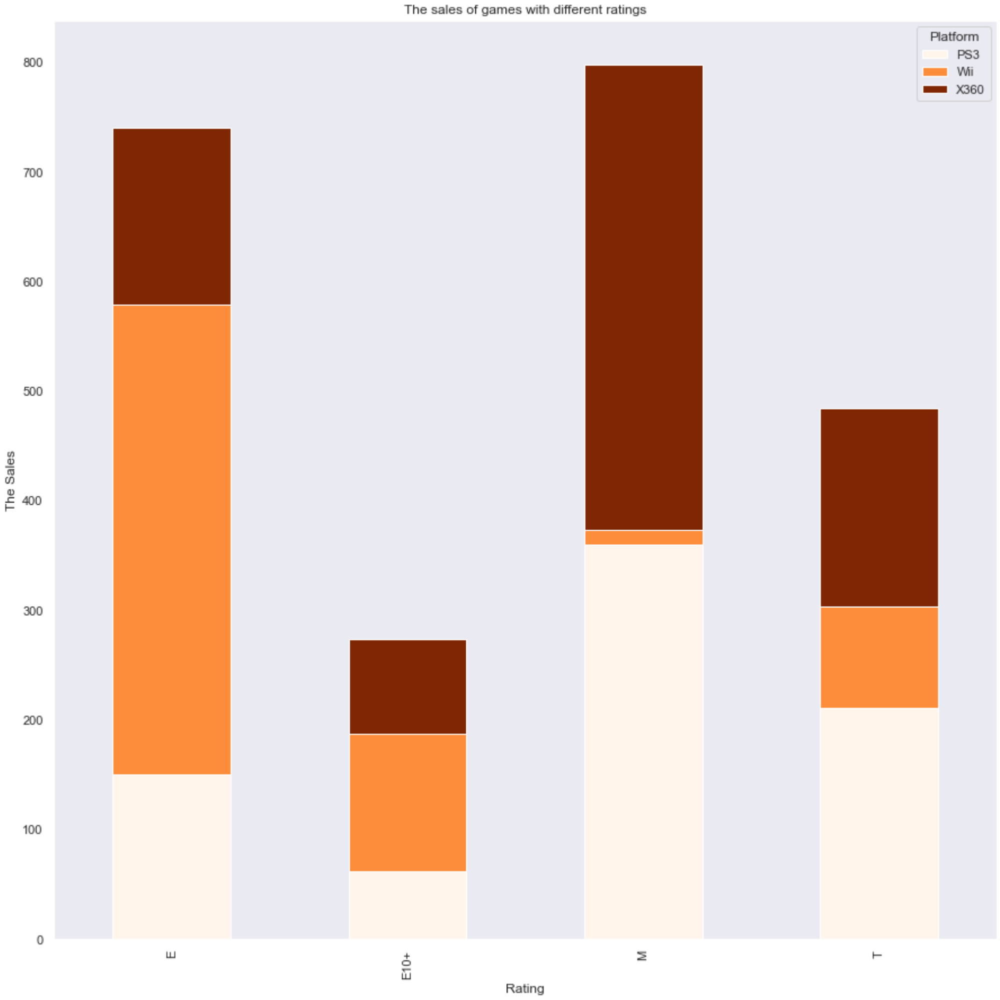
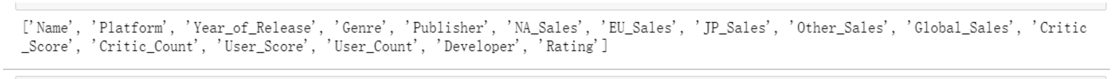


Chart 1. The sales of multi-rating games in different platforms

3. Model & Insight:

A. Predict the sales volume:

First of all, we want to predict the amount of sales that Black Myth could achieve by the application of data we obtained. In order to find a proper model to process the data, we need to check the information of our dataset. Below are columns we got from the dataset. In order to estimate sales volume, we choose Global_Salses as the target value (i.e., Y), and for independent variables (i.e., X1, X2, ..., Xn), we could divide them into two types: numerical variables and categorical variables. In terms of numerical variables, we should check the Pearson correlation values at Diagram1, which would help us to get rid of the variables that are almost no correlation with our target value (Y) here, for example, the correlation between Y and Year_of_Release is 0.0067, which is even less than 0.01. After removing Year_of_Release, we take four numerical variables into consideration, which are Critic_Score, Critic_Count, User_Score and User_Count.



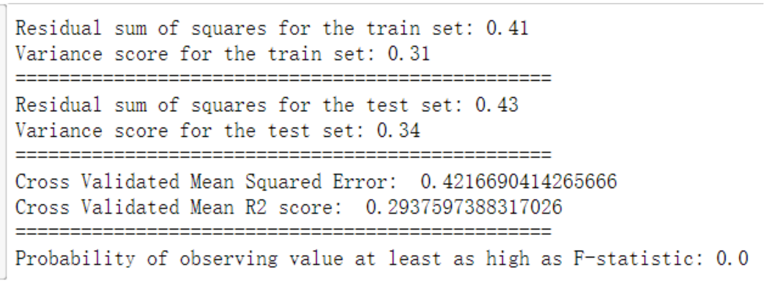
```
['Name', 'Platform', 'Year_of_Release', 'Genre', 'Publisher', 'NA_Sales', 'EU_Sales', 'JP_Sales', 'Other_Sales', 'Global_Sales', 'Critic
_Score', 'Critic_Count', 'User_Score', 'User_Count', 'Developer', 'Rating']
```

For categorical variables, we choose Platfrom, Genre and Rating to analyze with. This is because the producer of Black Myth: Wukong is an individual start-up in China named as Game Science, where according to the encoding method we applied, variables like Publisher and Developer that includes most famous game companies such as Nintendo and Microsoft will not have influence on the prediction of sales volume for Black Myth.

After selecting all independent variables we plan to use in the analytics, three different encoding methods are applied to search for a more proper way to deal with those categorical variables involved. Since those categorical variables are nominal, we use one-hot, target and frequency encoding methods relatively. In addition, under each encoding method, we conduct two different models to find a better estimation, which are multi-linear regression model and random forest model.

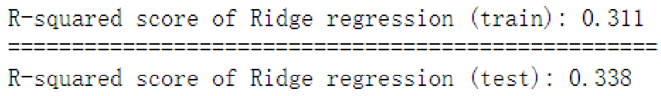
1. One-hot Encoding Method:

The performance of multi-linear regression:



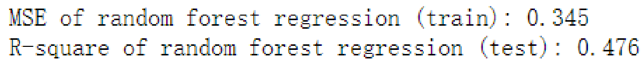
```
Residual sum of squares for the train set: 0.41
Variance score for the train set: 0.31
=================================================
Residual sum of squares for the test set: 0.43
Variance score for the test set: 0.34
=================================================
Cross Validated Mean Squared Error:  0.4216690414265666
Cross Validated Mean R2 score:  0.2937597388317026
=================================================
Probability of observing value at least as high as F-statistic: 0.0
```

After normalizing the data and using Ridge regression to deal with multicollinearity issues:

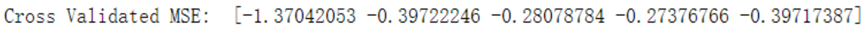


```
R-squared score of Ridge regression (train): 0.311
====================================================
R-squared score of Ridge regression (test): 0.338
```

The performance of random forest regression:



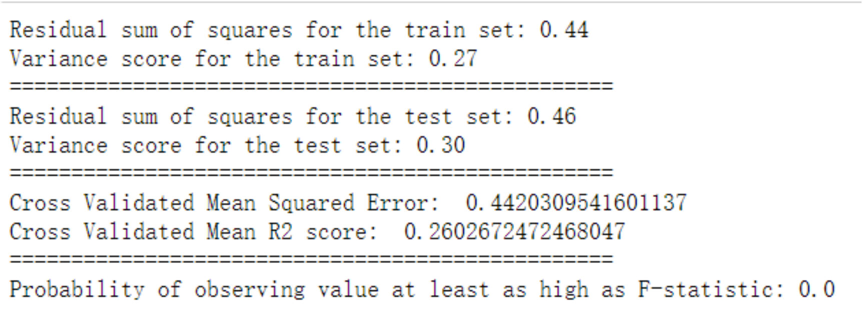
```
MSE of random forest regression (train): 0.345
R-square of random forest regression (test): 0.476
```



```
Cross Validated MSE:  [-1.37042053 -0.39722246 -0.28078784 -0.27376766 -0.39717387]
```

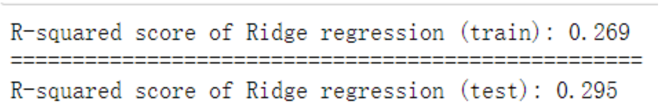
2. Target Encoding Method:

The performance of multi-linear regression:



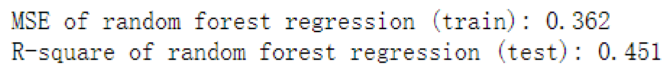
```
Residual sum of squares for the train set: 0.44
Variance score for the train set: 0.27
=================================================
Residual sum of squares for the test set: 0.46
Variance score for the test set: 0.30
=================================================
Cross Validated Mean Squared Error:  0.4420309541601137
Cross Validated Mean R2 score:  0.2602672472468047
=================================================
Probability of observing value at least as high as F-statistic: 0.0
```

After normalizing the data and using Ridge regression to deal with multicollinearity issues:

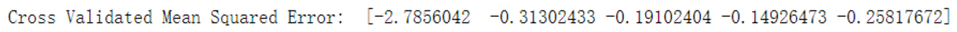


```
R-squared score of Ridge regression (train): 0.269
====================================================
R-squared score of Ridge regression (test): 0.295
```

The performance of random forest regression:



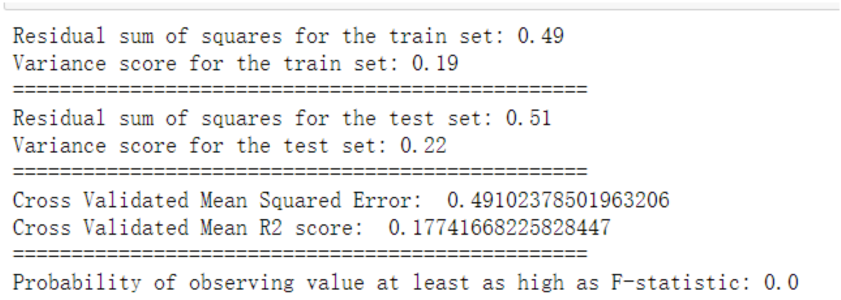
```
MSE of random forest regression (train): 0.362
R-square of random forest regression (test): 0.451
```



```
Cross Validated Mean Squared Error:  [-2.7856042  -0.31302433 -0.19102404 -0.14926473 -0.25817672]
```

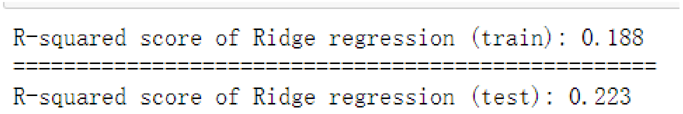
3. Frequency Encoding Method:

The performance of multi-linear regression:



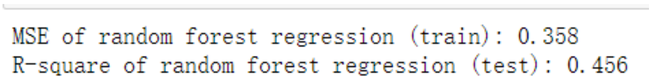
```
Residual sum of squares for the train set: 0.49
Variance score for the train set: 0.19
==================================================
Residual sum of squares for the test set: 0.51
Variance score for the test set: 0.22
==================================================
Cross Validated Mean Squared Error:  0.49102378501963206
Cross Validated Mean R2 score:  0.17741668225828447
==================================================
Probability of observing value at least as high as F-statistic: 0.0
```

After normalizing the data and using Ridge regression to deal with multicollinearity issues:

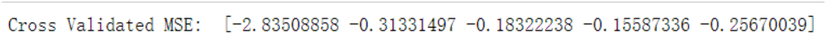


```
R-squared score of Ridge regression (train): 0.188
=====================================================
R-squared score of Ridge regression (test): 0.223
```

The performance of random forest regression:



```
MSE of random forest regression (train): 0.358
R-square of random forest regression (test): 0.456
```



```
Cross Validated MSE:  [-2.83508858 -0.31331497 -0.18322238 -0.15587336 -0.25670039]
```

Among all metrics, we select R2 score as our standard of reference and consequently choose random forest model under one-hot encoding, where the R2 score is 47%, meaning the model would perform better than other models presented above. By the application of this model, we predict the global sales volume of Black Myth on PS platforms (both PS3 and PS4), which is 4.91 million.

B. Price the game properly:

To estimate the price of the game when it is published, the profit formula was applied as shown in equation 1.

$$Profit = Revenue - Costs \tag{1}$$

For the game company, since the platform such as Play Station, will take a commission fee for each sale of the game on its platform, revenue can be represented by sales volume times the difference between game price and platform commission fee. Costs include R&D costs, advertising costs, and operation costs. Equations 2 and 3 show the detailed formula for revenue and costs.

$$Revenue = Sales\ Volume * (Price - Commission\ Fee) \tag{2}$$

$$Costs = R\&D\ Cost + AD\ Cost + OP\ Cost \tag{3}$$

Combining equations 1, 2, and 3, the detailed formula for the profit and the relationship between profit and the price is shown in equation 4.

$$Profit = Sales\ Volume * (Price - Commission\ Fee) - (R\&D + AD + OP\ Cost\ ) \tag{4}$$

With this equation, the price can be inferred when the rest terms are available. First, the profit was set to zero to see what price can make the revenue cover the total cost. The sales volume was set to 4.91 million, which is estimated by the sales model mentioned above. As the analysis is conducted based on the assumption that the game will be published on Play Station, Nintendo will take the commission fee for every sale of the game. According to Nintendo’s policy, for games whose sales are above 100 thousand, middleman’s interests are 10.3%, and retailer interests are 29.3% for every sale of the game (CESA).

Therefore, 60.7% of the price times sales volume will be the revenue of the game. Based on the news released by the developer, the estimated cost of R&D is $72 million (Doolan). The developer team has approximately 80 people with an average monthly salary of 25 thousand RMB. The total labor cost for three years is about 72 million RMB. The base of the company is in the new first-tier city, therefore, based on the average rental cost of building and equipment in that city is about 30 million RMB. Based on the scale of the game, 120 people should be needed, but the team only has 80 people, therefore, outsourcing is necessary, which is about 45 million RMB. The advertising is estimated to be 30 million RMB according to the average advertising cost for similar games. Adding the operation cost and advertising cost together, 177 million RMB is required, which is $25 million in US dollars. Therefore, plus the R&D cost, the total cost is about $97 million. Using equation 4, the price that leads to zero profit can be calculated in equation 5.

$$0 = 4.91\,m * Price * (1 - 39.6\%) - 97\,m \tag{5}$$

The price is calculated to be $32.6, which will lead to zero profit. There are correlations between sales volume and pricing. How much customers would like to pay for a game depends on the platform and the game's genre. In Play Station, after all those years, game companies found that $60 is a threshold, games exceeding $60 usually experience an unexpected decrease in sales volume. Therefore, most games set a price of less than $60 no matter how much investment developers put into the game.

To obtain a more specific price for the game, a similar game, God of War 3, was analyzed to compare with Black Myth: Wukong. God of War 3's sales volume was 5.2 million, and its R&D cost was 44 million, as shown in Table 1 below.

	Black Myth: Wukong	God of War 3
Sales Volume (million)	4.91	5.2
R&D Cost (million $)	72	44

Table 1: Black Myth: Wukong and God of War 3 R&D cost and sales volume (*God of war III*).

From the table, it can be found that the expected sales volume of Black Myth: Wukong is close to the sales volume of God of War 3. Besides, the sales volume of Black Myth: Wukong is estimated using datasets from 2019 and before. With more advertising methods and good reviews from customers under the introduction video on YouTube, it is expected that the actual sales volume would be higher than the sales volume estimated by the model. The table above also reveals that Black Myth: Wukong's R&D cost is higher than God of

War 3. God of War 3's initial price is $59.9 (*God of war III*). Therefore, it is expected that the initial price of Black Myth: Wukong will be higher than the initial price of God of War 3 or as same as God of War 3. With the threshold of price on Play Station mentioned above, Black Myth: Wukong's initial price should be $59.9.

4. Conclusion

In conclusion, our project mainly focuses on the prospect of an upcoming video game named Black Myth: Wukong. By random forest model with one-hot encoding method, the potential sales volume of this game on PS platform (both PS3 and PS4) tends to be 4.91 million, which is quite considerable. Accordingly, the suggested retail price is $59.9.

Moreover, there are some deficiencies in our model and could be improved in the future. On the one hand, for the prediction of sales volume, since it requires a multi-linear regression, the good-of-fit value (i.e., R2 score) is 47.6%, which is relatively low. From my perspective, there are two main reasons. Firstly, the independent variables we choose here do not meaningfully account for the sales volume of a video game. For instance, according to the Diagram1, the Pearson correlation between User_Count and Global_Sales is only 0.088, which is quite insignificant. Secondly, most variables that we take into consideration are categorical variables. Therefore, a more advanced encoding method might be helpful to improve the performance of the prediction as well.

On the other hand, based on the profit formula with the sales volume predicted by the previous model, costs, and statistics of similar games, the launch price of Black Myth: Wukong on Play Station will be $59.9. There are several advantages of the price model discussed in the previous section. First of all, it is a simple and understandable model that uses the profit formula to obtain a suitable launch price for the game. Also, the model can be applied to different games with available data on costs, sales volume, and commission fees taken by the platform. Moreover, users can adjust the price based on different levels of profit expected.

However, there are also flaws in the price model. First, the price model can only estimate the launch price of games. The price of games has a lifecycle, which means that in different periods, the price will be different in order to stimulate the purchase. For example, on holidays and black Friday, lots of games will offer a lower price to customers. The model cannot estimate the price of games on special days.

Moreover, this model does not include any impacts on customers on the price. The model only considers the costs and sales. When estimating the sales, the players' review is part of the data, but Black Myth has not been published, so the player's review is missing from the model. It is certain that players' reviews and other customer's behaviors will influence the launch price of the game. Therefore, including customers' actions in the model is necessary for future improvement.

To further improve the accuracy of the model, one major step is to gather customers' interests in the game before launch to increase the accuracy of the sales model. Also, including time series to anticipate the change of price after the game launches based on similar games' pricing strategies will lead to a more comprehensive price model instead of only estimating the launch price. An improved version of the price model should be able to anticipate the price of the game throughout different periods of a game's lifecycle, including intro, growth, maturity, decline, and niche, as shown in Figure 6 below.

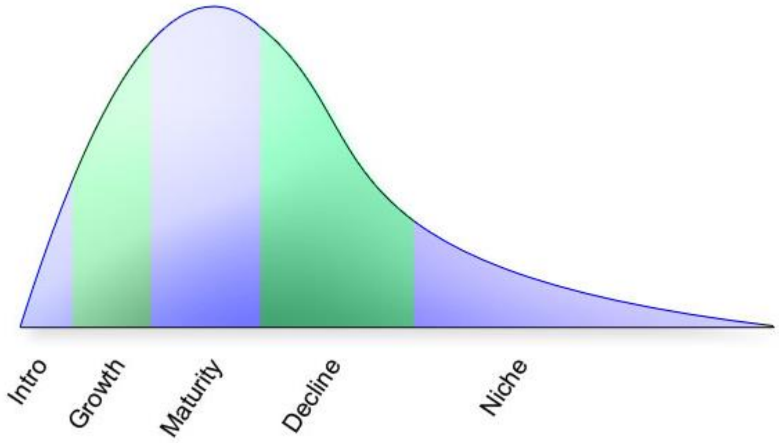


Figure 6: Games' general lifecycle. (Cook et al.)

5. Code Source:

   https://github.com/xinyuyvonne/ESE527_IPhoneXR.git

Dataset Link:

https://www.kaggle.com/datasets/rush4ratio/video-game-sales-with-ratings/code?datasetId=576&sortBy=voteCount

https://www.kaggle.com/datasets/ashaheedq/video-games-sales-2019

Reference:

“Business Models and Pricing Strategies in Videogames Industry.” https://dspace.lib.uom.gr/bitstream/2159/16047/8/ArampatzisPaschalisMsc2014.pdf.

Cook, Daniel, et al. “Game Genre Lifecycle: Part I.” *LOSTGARDEN*, 21 Sept. 2019, https://lostgarden.home.blog/2005/05/06/game-genre-lifecycle-part-i/.

Doolan, Liam. “Nintendo Switch Worldwide Software Sales Update (CESA 2022) - Super Mario 3D All-Stars, Pokémon, Zelda & More.” *Nintendo Life*, Nintendo Life, 30 Aug. 2022, https://www.nintendolife.com/news/2022/08/nintendo-switch-worldwide-software-sales-update-cesa-2022-super-mario-3d-all-stars-pokemon-zelda-and-more.

“God of War III.” *Wikipedia*, Wikimedia Foundation, 13 Dec. 2022, https://en.wikipedia.org/wiki/God_of_War_III.