Final Project: Video Game Sales Modeling

Group 6

Team Members:

Ayushi Agarwal

Caleb Paul

Jacinth Attada

Lakshmi Ramya Marineni

Xinbo Ye

OPIM 5604: Predictive Modeling

Professor Jose Cruz

MSBAPM, University of Connecticut

School of Business

Contents

Introduction

The video game industry is a big business with a lot of players, including private corporations and investors. The video game sector is larger than the movie and music industries combined with significant expansion potential. Top companies have benefited from the huge demand that was created due to COVID-19. A greater understanding of the industry's driving elements is in high demand among these investors. The video game industry has always prioritized innovation. The corporation can expect that technological advances, controls, and experiences would be implemented to generate more revenue to the stakeholders. In 2020, the gaming industry generated $155 billion in revenue, by 2025, analysts predict the industry will generate more than $260 billion in revenue.

Abstract

Video games are a billion-dollar business and have been for many years. Analysis and modeling are often performed in a cycle, enabling iterative refinement and data modeling to uncover interesting insights about video game sales. As a team we have worked on the video game sales dataset sourced from Kaggle. The dataset's most essential goals are to figure out which area, genre, region, and platform publisher to invest in. In this project, we analyzed the data set in JMP by applying data
preprocessing methods to clean data and remove any unnecessary variables, columns,
and rows. We performed data exploration. Clustering and correlation analysis were the two main approaches of data exploration that we employed. This aided in our ability to understand the data to uncover patterns and points of interest concerning the three objectives. Lastly, we performed various modelling techniques on the data and listed down the insights.

The related data dictionary:

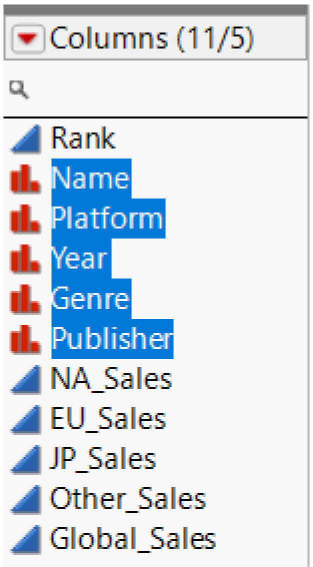
- Rank - Ranking of overall sales
- Name - The games name
- Platform - Platform of the games release (i.e. PC, PS4, etc.)
- Year - Year of the game's release
- Genre - Genre of the game
- Publisher - Publisher of the game
- NA_Sales - Sales in North America (in millions)
- EU_Sales - Sales in Europe (in millions)
- JP_Sales - Sales in Japan (in millions)
- Other_Sales - Sales in the rest of the world (in millions)
- Global_Sales - Total worldwide sales.

Data Cleaning

We initially preview the data and see that there are 11 variables in the dataset of which 5 are nominal and 6 are continuous variables of 16,598 records. We then analyzed the distribution of all variables in the dataset.

1. Previewing the Data

There are 11 variables in the dataset. 5 of them are nominal variables, and 6 are continuous variables.



We looked at the distribution to view the overall quality of the data. There are only 271 missing values and some outliers. We will proceed to remove or transform the outliers in the following steps. Additionally, some variables do not provide any business value, and we would like to remove those variables.

1.1. Exclude variable *Rank*

	Rank
1	1
2	2
3	3
4	4
5	5
6	6
7	7
8	8
9	9
10	10
11	11

The records in the original dataset are ordered by global sales in descending order. There are some ranks missing, but the overall rank does not provide much valuable information for our further analysis.

1.2. Exclude variable *Name*

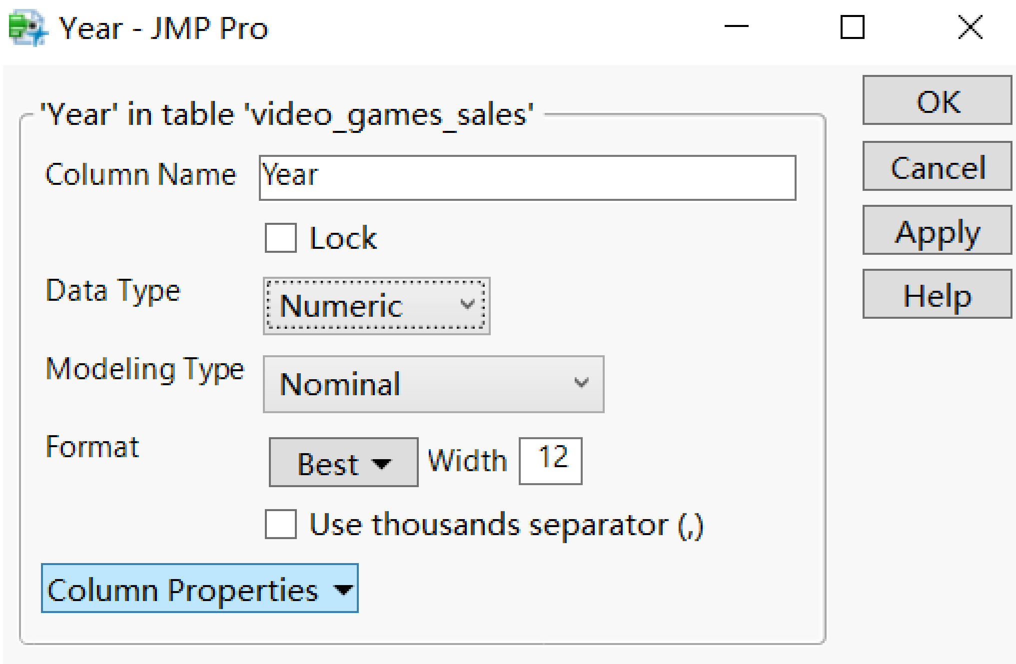
Frequencies

Level	Count	Prob
'98 Koshien	1	0.00006
.hack//G.U. Vol.1//Rebirth	1	0.00006
.hack//G.U. Vol.2//Reminisce	1	0.00006
.hack//G.U. Vol.2//Reminisce (jp sales)	1	0.00006
.hack//G.U. Vol.3//Redemption	1	0.00006
.hack//Infection Part 1	1	0.00006
.hack//Link	1	0.00006
.hack//Mutation Part 2	1	0.00006
.hack//Outbreak Part 3	1	0.00006
.hack//Quarantine Part 4: The Final Chapter	1	0.00006
.hack: Sekai no Mukou ni + Versus	1	0.00006
[Prototype 2]	3	0.00018
[Prototype]	2	0.00012
¡Shin Chan Flipa en colores!	1	0.00006
007: Quantum of Solace	6	0.00036
007: The World is not Enough	2	0.00012
007: Tomorrow Never Dies	1	0.00006
007 Racing	1	0.00006
1/2 Summer +	1	0.00006
1 vs. 100	1	0.00006
2 Games in 1: Disney Princess & The Lion King	1	0.00006
2 Games in 1: Disney's Brother Bear / The Lion King 1 1/2	1	0.00006
2 Games in 1: Sonic Advance & ChuChu Rocket!	1	0.00006
2 Games in 1: Sonic Battle & ChuChu Rocket!	1	0.00006
2 Games in 1: Sonic Pinball Party & Columns Crown	1	0.00006

N Missing 0
11493 Levels

There are in total 11493 unique names in the variable. Some of the values have multiple records, and others have only one. The reason for this case is that some games are published on multiple platforms, and therefore their sales are counted separately. We are not comparing sales on different platforms for each game in our analysis, so we decided to exclude this variable from the dataset.

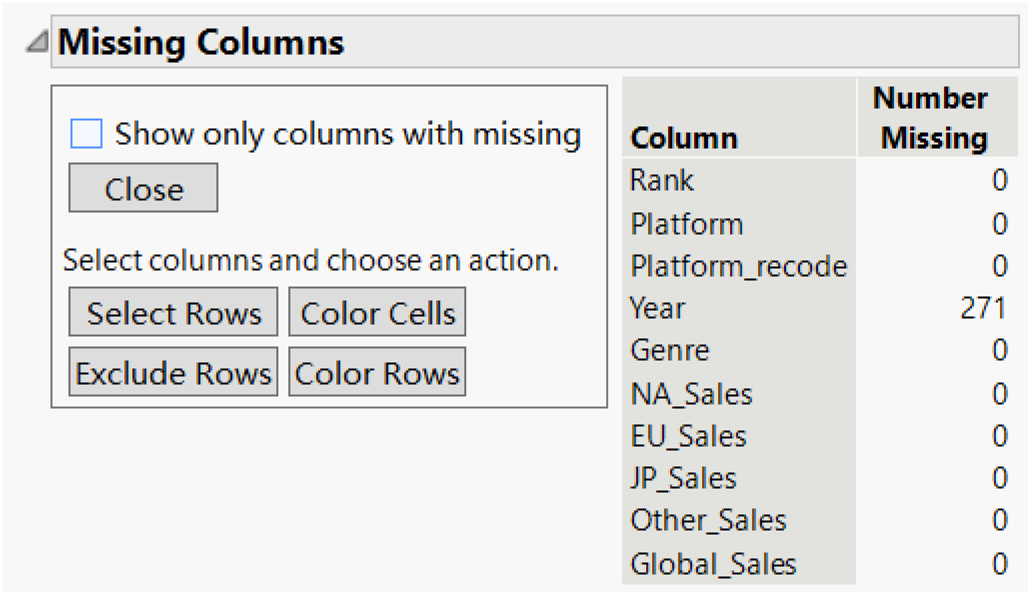
1.3. Change the type of variable *Year* into numeric



The variable *Year* was originally a character type variable. It has in total 40 levels. Adding a complex character variable into the predictive model would have noise and lower the performance of the model. Therefore, we decided to change its type to numeric.

1.4. Exclude Missing Values

By performing analyze——screening——exploring missing values in JMP, we found 271 missing values in the variable Year. Since the number of missing values is relatively small compared to the total number of records, we decided to exclude those records.



Missing Columns

Show only columns with missing

Close

Select columns and choose an action.

Select Rows | Color Cells

Exclude Rows | Color Rows

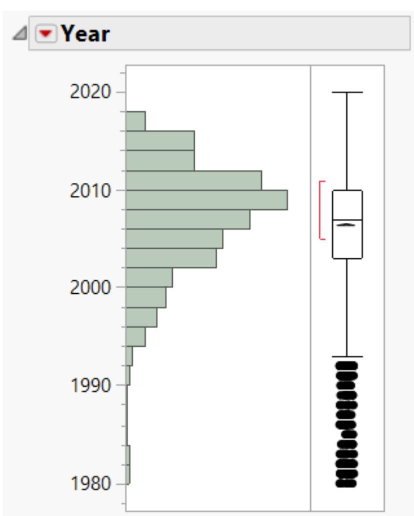
Column	Number Missing
Rank	0
Platform	0
Platform_recode	0
Year	271
Genre	0
NA_Sales	0
EU_Sales	0
JP_Sales	0
Other_Sales	0
Global_Sales	0

2. Handling Outliers

We now start cleaning the outliers in the dataset one variable at a time.

2.1. *Year*

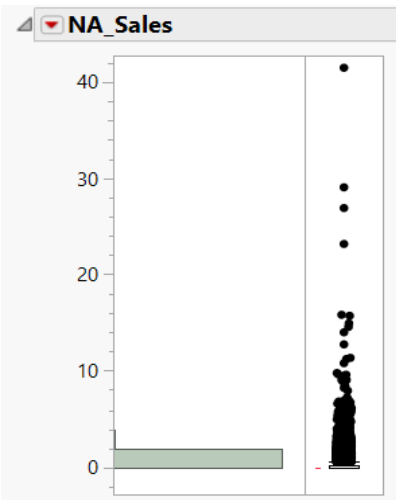
We do not transform or delete outliers. According to the box plot, there are a certain number of data characterized as outliers. We would like to keep the original time range as it is, so we don't change it.

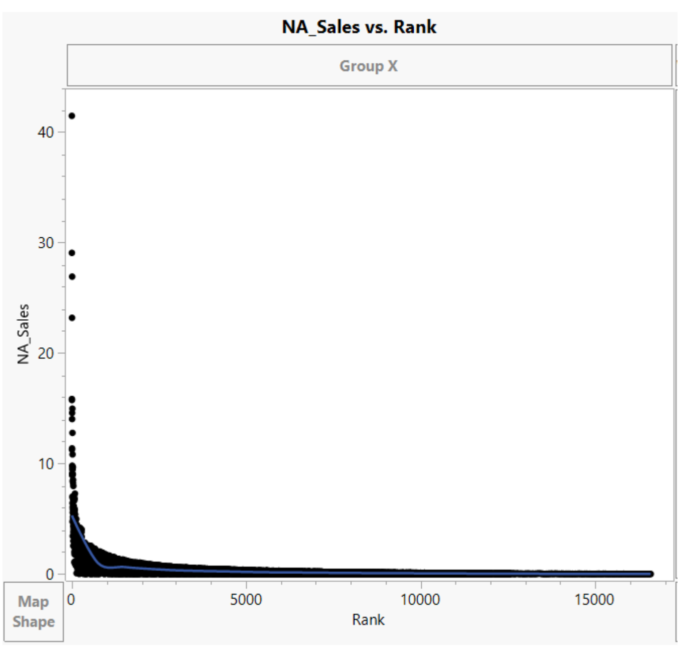


2.2. *NA_Sales*

We do not transform or delete outliers. According to the box plot, there are a certain number of data characterized as outliers. As we plot the distribution of NA_Sales with a scatter plot, we found that higher ranks come with higher sales. There are a great number of low sales records, therefore JMP characterized those high sales records as outliers. We cannot simply decide to remove the outliers.

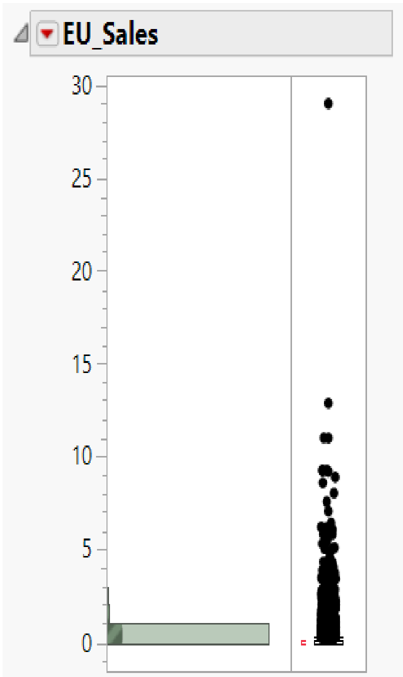
Additionally, we decided not to fit a SHASH distribution to the variable in order to keep the original information. When fitted with the SHASH distribution, there will be an information loss.

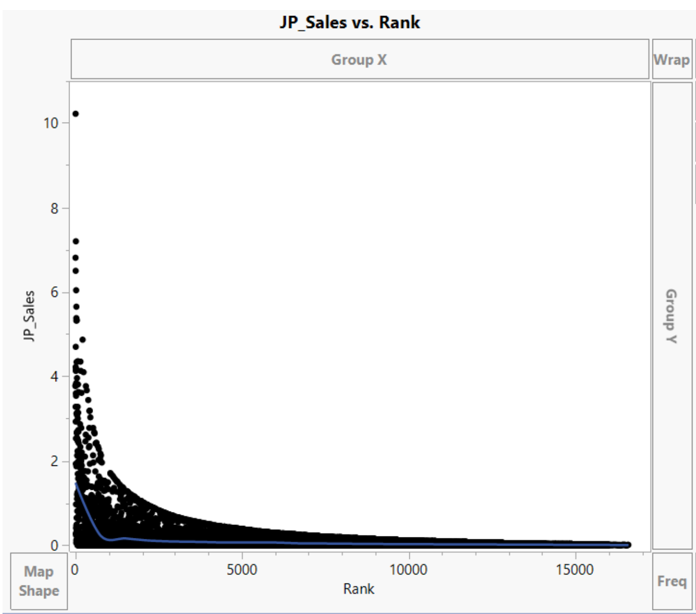




2.3. *EU_Sales*

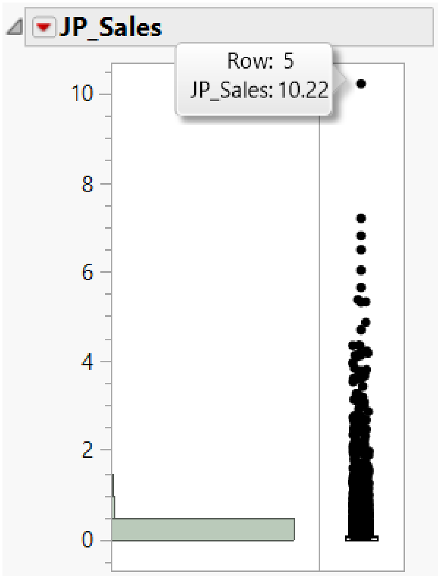
We do not transform or delete outliers. The reasoning behind this correlate to the last variable. The only difference between NA_Sales and EU_Sales is the region. The distribution of sales follows the ranking, so we don't need to exclude those high points in the scatter plot.

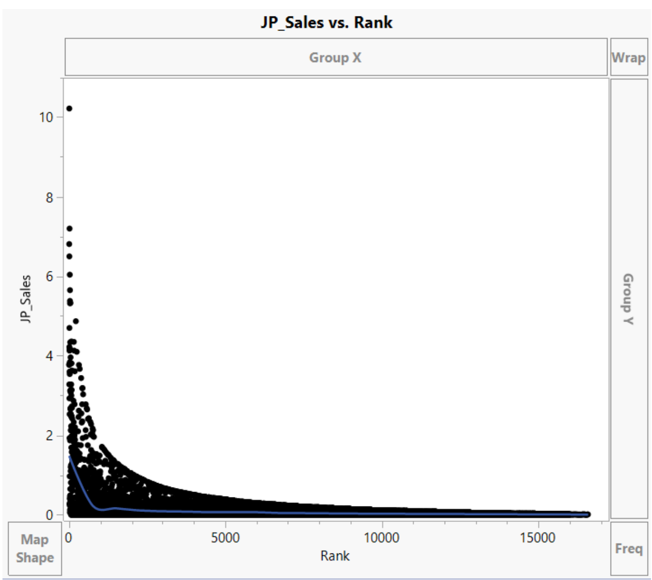




2.4. *JP_Sales*

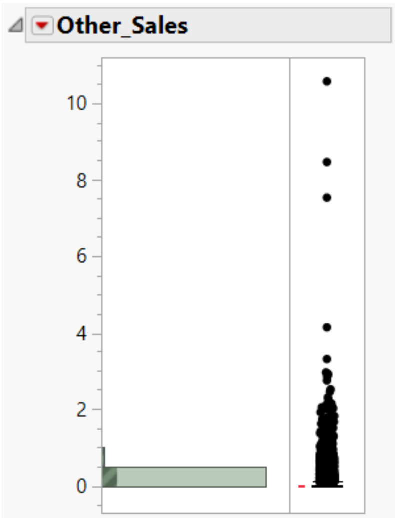
We do not transform or delete outliers. The reasoning is the same above. We also discovered that the top 1 record in JP_Sales is different from that in EU and NA. It shows that Pokemon Red & Blue has dominant popularity in Japan.

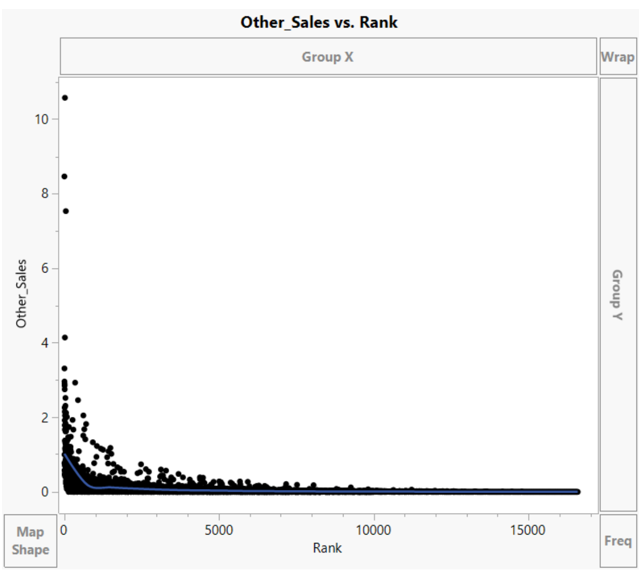




2.5. *Other_Sales*

We do not transform or delete outliers. The reasoning is the same above. From the box plot, we can see that there are three points far away from the other records. We found the dominant popularity of these two games in North America and other areas. We would explore more in further steps.

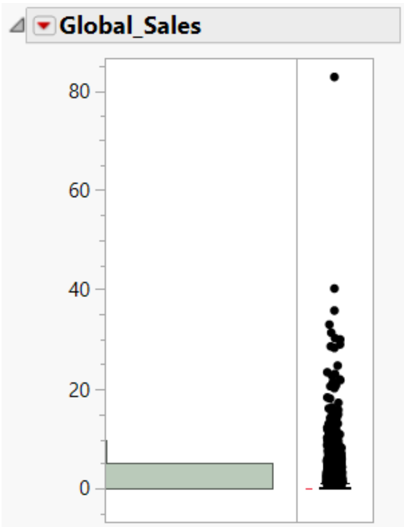


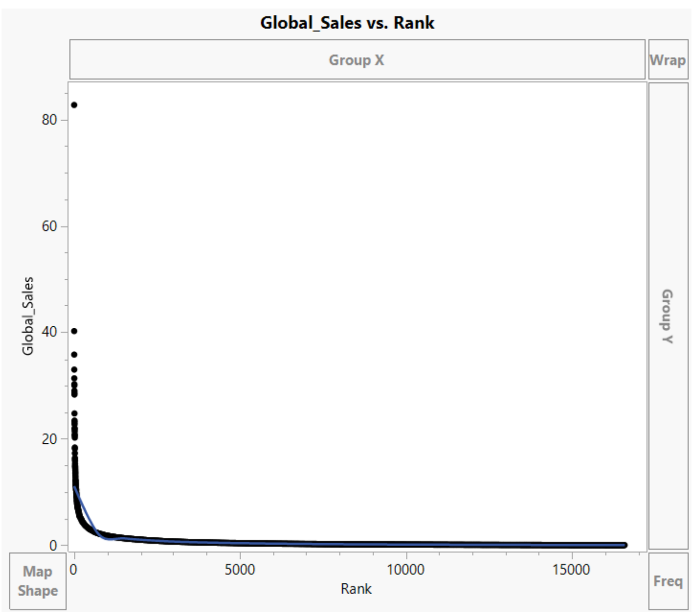


	Rank	Name	Platform	Year	Genre	Publisher	NA_Sales	EU_Sales	JP_Sales	Other_Sales
1	18	Grand Theft Auto...	PS2	2004	Action	Take-Two Interac...	9.43	0.4	0.41	10.57
2	48	Gran Turismo 4	PS2	2004	Racing	Sony Computer E...	3.01	0.01	1.1	7.53

2.6. *Global_Sales*

We do not transform or delete outliers. Global_Sales is calculated by summing all the regional sales. Since we did not transform or delete any outliers in each of the regional sales, we should not make changes to the Global_Sales to prevent information inconsistency. When compared to regional sales, global sales seem to fit better to a curve function.

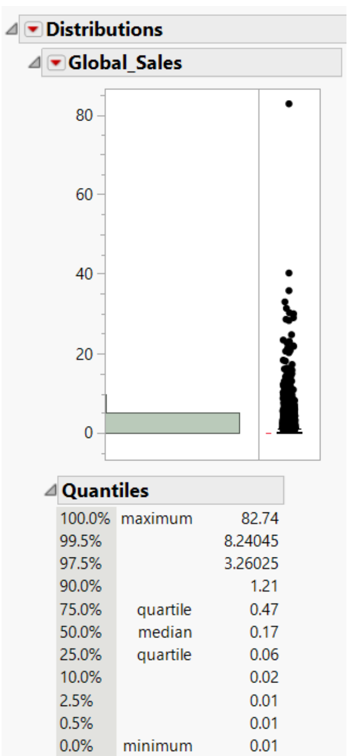




2.7. Add Data Binning Column

From the previous steps, we did not make any changes to the outliers. To prevent an overfitting problem, we decided to use data binning to transform the target variable into a categorical variable. It will help improve the model performance and extract important information from a wide range of numbers.

We divided the Global_Sales into 5 levels. We characterized the top 2.5% of data as Top Sales, 2.5% to 25% of data as High Sales, 25% to 50% of data as Medium Sales, 50% to 75% of data as Moderate Sales, and the last of data as Low Sales.



2.8. Add Binned Platform Column

After performing the distribution analysis, we found out that the levels of the platform are too many to analyze. Therefore, we consider transforming the value of platforms into the companies that developed the platforms. In this way we could narrow down the number of different platforms into 9 categories. The new column could be used for further analysis.

	Platform	Platform_rec...
	Wii	Nintendo
	NES	Nintendo
	Wii	Nintendo
	Wii	Nintendo
	GB	Nintendo
	GB	Nintendo
	DS	Nintendo
	Wii	Nintendo
	Wii	Nintendo
	NES	Nintendo

2.9. Add Global Sales_Sucess or not Column

For the logistic fit, a new column Global Sales_Sucess or not was created. This column was created based on the global sales column which was taken as 0 and 1 depending upon the threshold global sales value. This allowed us to categorize it into two groups, allowing us to determine the likelihood of the game being sold and successful over time.

Data Exploration

Correlation analysis

In the data exploration phase, we used a variety of methods to explore the data set. We used correlation analysis to find relationships within the data, linear regression to establish significant variables against a target variable, as well as clustering to assign groups and for identification. Lastly, we explored charts and graphs which allowed us to identify trends.

We performed correlation analysis on all variables to see how the relationship of each variable can affect another. From the multivariate analysis we can see that Global Sales and North American sales have the highest positive correlation which is at 0.9410. This means that these two variables have a strong relationship and are related to each other. Also notably, we can see that rank has a negative correlation with the sales variables. This means that as one variable increases, the other will decrease and vice versa.

Multivariate

Correlations

	Rank	Year	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales
Rank	1.0000	0.1788	-0.4014	-0.3791	-0.2678	-0.3330	-0.4274
Year	0.1788	1.0000	-0.0914	0.0060	-0.1693	0.0411	-0.0747
NA_Sales	-0.4014	-0.0914	1.0000	0.7677	0.4498	0.6347	0.9410
EU_Sales	-0.3791	0.0060	0.7677	1.0000	0.4356	0.7264	0.9028
JP_Sales	-0.2678	-0.1693	0.4498	0.4356	1.0000	0.2902	0.6118
Other_Sales	-0.3330	0.0411	0.6347	0.7264	0.2902	1.0000	0.7483
Global_Sales	-0.4274	-0.0747	0.9410	0.9028	0.6118	0.7483	1.0000

Clustering analysis

We conducted a cluster analysis with global sales to determine which regions are affecting each video game genre and which regions are selling more than others.

We conducted cluster analysis on the target variable global sales. This is done so we can view how regional sales affect our overall sales. It also allows us to identify and create groups within the data set. We first conducted a logistic regression to determine the most important variables to use for predicting sales. Here are the results of the logistic regression.

Response Global_Sales

Effect Summary

Source	LogWorth	PValue
NA_Sales	32647.40	0.00000
EU_Sales	28491.75	0.00000
JP_Sales	27829.76	0.00000
Other_Sales	22871.28	0.00000
Rank	15.191	0.00000
Platform	2.985	0.00104

Remove Add Edit Undo FDR

These are the variables that were used in the K Means clustering analysis. We can see that their P-Values are below 0.05 which means they are significant variables and will be used in our cluster analysis with our target variable Global Sales.

After performing the K-Means cluster we can conclude that the optimal cluster size is 8. Here are the results.

Iterative Clustering

Cluster Comparison

Method	NCluster	CCC	Best
K Means Cluster	3	-95.054	
K Means Cluster	4	-126.76	
K Means Cluster	5	-61.775	
K Means Cluster	6	-14.406	
K Means Cluster	7	-15.26	
K Means Cluster	8	1.92926	Optimal CCC
K Means Cluster	9	-41.25	
K Means Cluster	10	-17.495	

Columns Scaled Individually

However, for interpretability and for visual purposes, we will use a K-Means cluster size of 5 to analyze. Here are the results of the K-Means cluster with size 5.

K Means NCluster=5

Columns Scaled Individually

Cluster Summary

Cluster	Count
1	1
2	2
3	10893
4	88
5	5614

Step	Criterion
58	0

Cluster Means

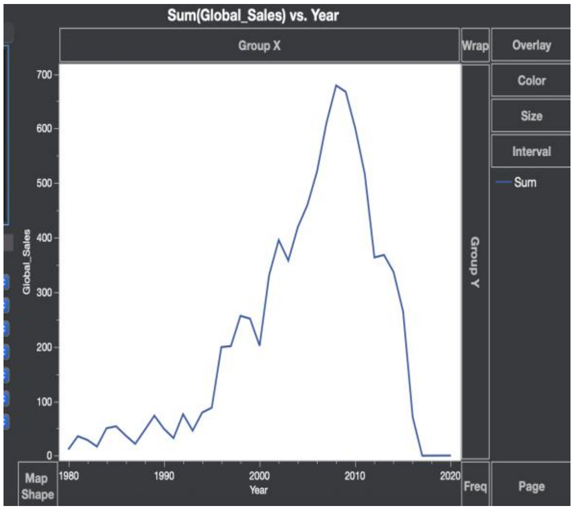
Cluster	Global_Sales	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Rank
1	82.74	41.49	29.02	3.77	8.46	1
2	16.235	6.22	0.205	0.755	9.05	33
3	0.10966951	0.05463968	0.02234646	0.02513265	0.00712568	11152.6704
4	14.5089773	6.84681818	4.30136364	2.18534091	1.17636364	47.3181818
5	1.12821696	0.55954934	0.31755611	0.14600285	0.10510331	2900.45885

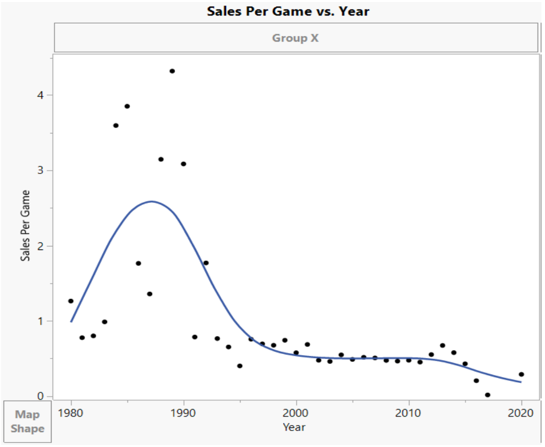
Cluster Standard Deviations

Here we can conclude that video games with a high number of global sales will also have a high number of sales in North America. Video Games with a medium number of sales will have a large amount of their sales from North America and Europe. Video Games with a low number of sales will have a larger number of sales coming from Japan and Other Regions. We can conclude that the majority of sales come from North America. When there are video games that do not sell well, we can conclude that Europe and Japan and Other Regional Sales will account for those sales. By determining that the North American region is the bestselling region will make it a large focus and a target for our team for our exploration.

Data Visualization

<u>Trend of Global Sales Vs. Year and Sales per game Vs. Year:</u>

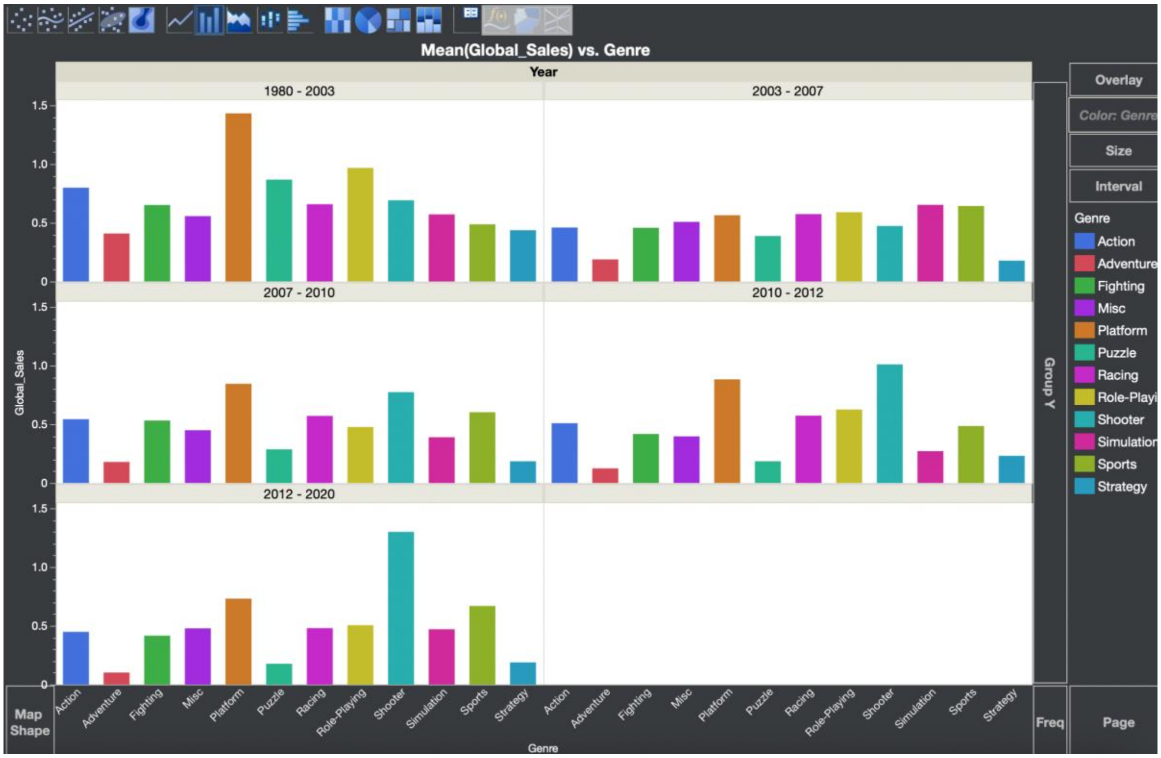




From our analysis, we found a significant drop on the global sales at the point of 2010. We wanted to check the trend of video games sales year-wise to predict the potential sales of video games in the upcoming years for the investors. Market in terms of sales for video games globally is not that good, so we need more specific strategies to improve the sales globally, investors need to be a little cautious before investing.

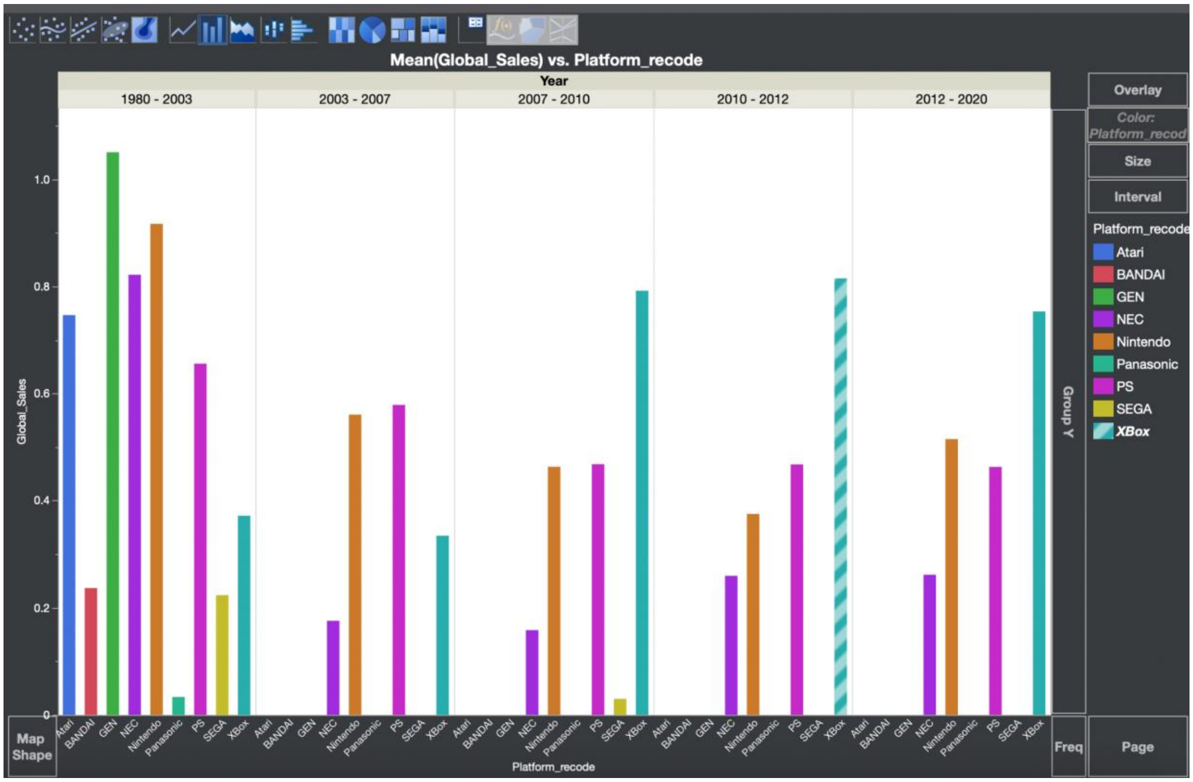
We would like to figure out whether there's a problem with the video games market. Therefore, we arrived at a second graph showing the sales per game do not have a significant drop. The market sales drop is due to a decrease in the number of games released. So we need more specific strategies.

<u>Genres Popularity with Highest Sales:</u>



From the plot, it can be seen that In the year 1980-2003, Platformers were the most popular games, then were Role-Playing and puzzle games. In the years 2003-2007, shooter and simulation became the most popular games in terms of sales. In the year 2007-2010, again Platformers were the most popular then in 2010-2012, shooters came ahead in popularity and till 2020 they are the most popular whereas platformers occupy the second place now.

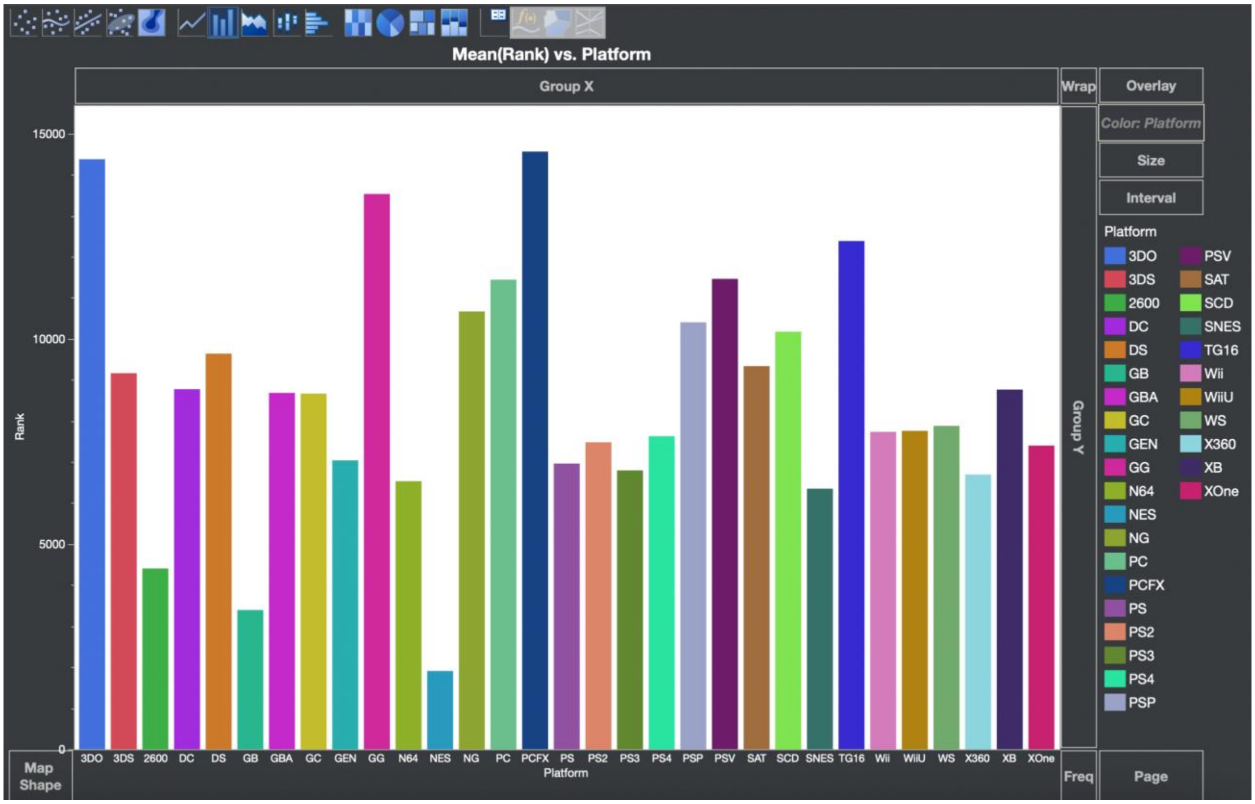
Platform Popularity with Highest Global Sales:



From the plot, it can be seen that In the year 1980-2003, GEN and Nintendo were the most popular platforms for highest revenue generation.

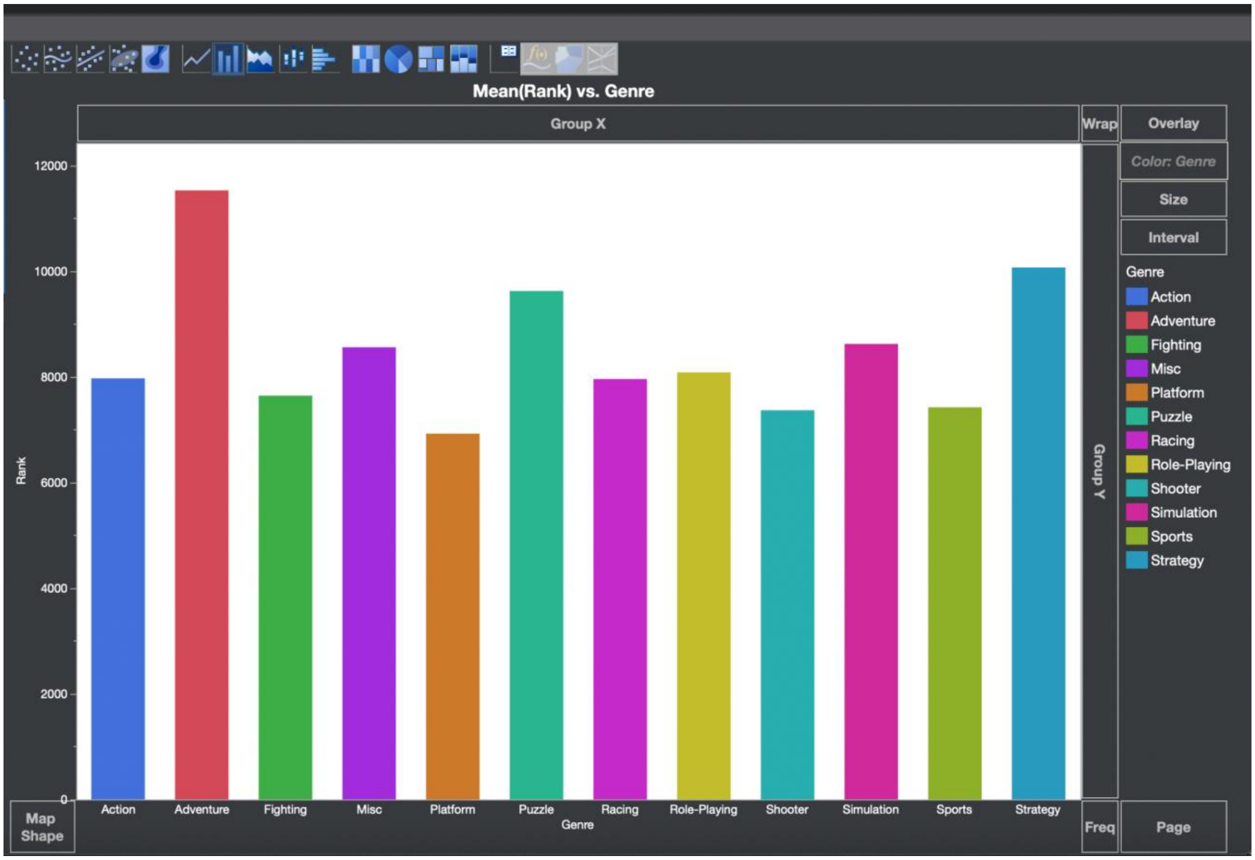
In the years 2003-2007, PS became the most popular platform in terms of revenue. In the year 2007 till 2020, XBOX sales were highest and were on peak which means Microsoft GAMES are making more money in the current generation.

Platforms with Highest Rank:



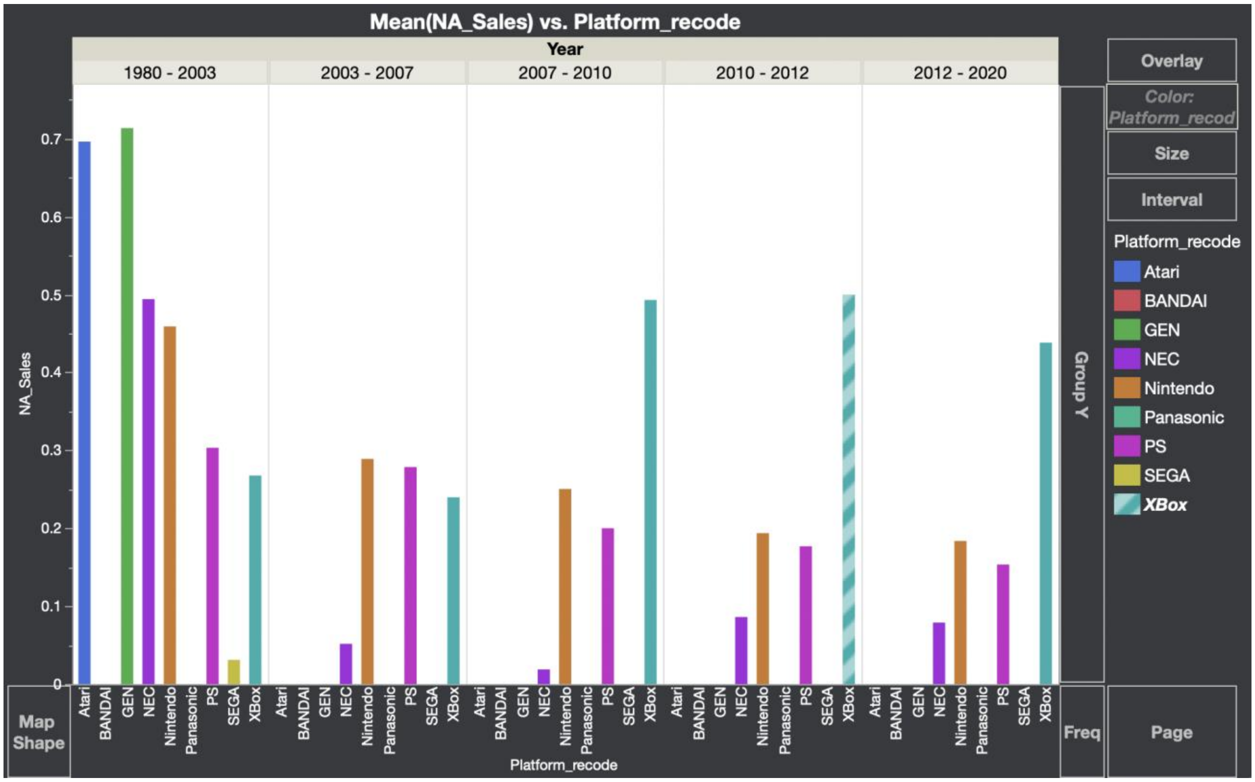
From the graph, it can be analyzed that PCFX (home console play stations) holds the highest-ranking platform which are developed and marketed by sony computer entertainment, then are 3DO.

Genres with Highest Rank:



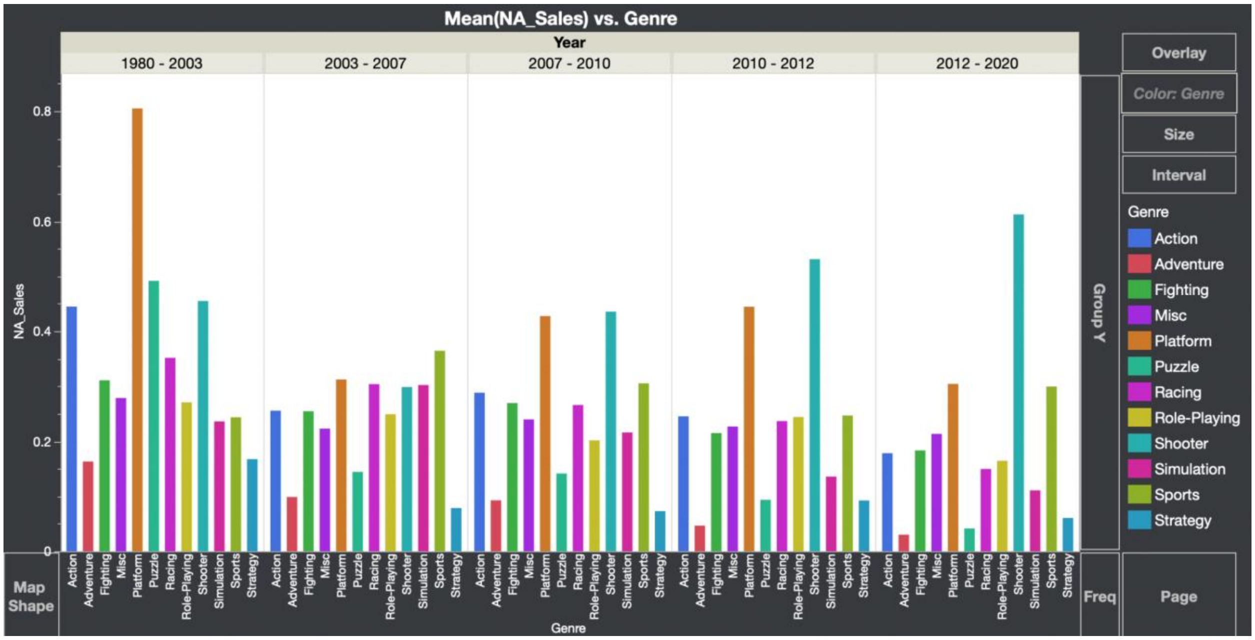
From the graph, it can be analyzed that adventure games hold the highest-ranking genres. Then there are strategy and puzzle games.

Highest Selling Platform in North America Year-Wise:



From the above graph, it can be seen that between the years 1980-2003, GEN was the highest selling platform for sales of video games. Then in the years 200-2007, Nintendo sales drastically increased. Then comes XBOX as the highest selling platform in the years from 2007 till 2020 in North America.

North-American Sales by Genre:



In North American sales analysis we wanted to see what genres have the highest sales over time. This will allow us to know the historical popularity of each genre from this specific region.

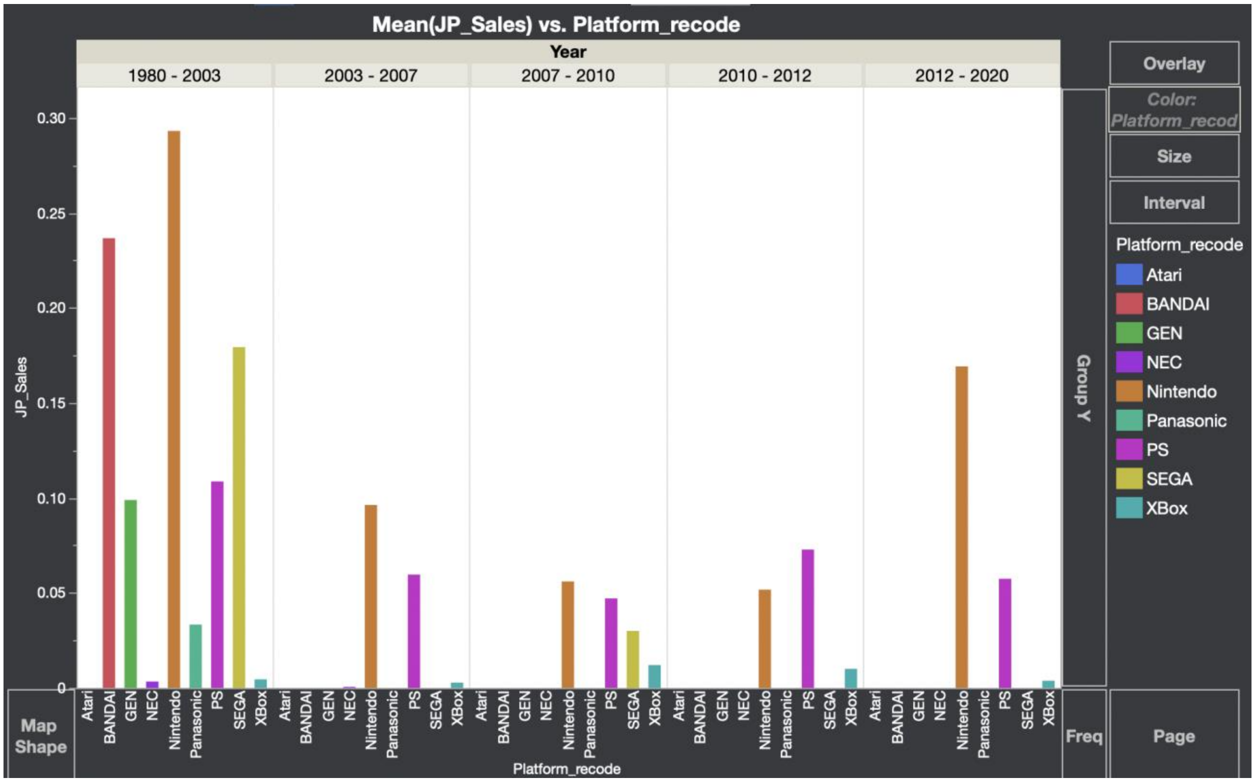
From our analysis we can see that from 1980 through 2003 the most popular genre has been the platform genre. From 2003 - 2007 the most popular genres are Sports and Platform.

From 2007 - 2020 we can see that the shooter genre has been more popular. In the earlier years the market was evenly distributed but over time Shooters have been dominating the market more and more.

Highest Selling Platform in Europe Year-Wise:

From the above graph, it can be seen that between the years 1980-2003, GEN was the highest selling platform for sales of video games. Then in the years 200-2007, PS sales drastically increased. Then comes XBOX as the highest selling platform since 2007 till 2020 in Europe as the highest selling platform in Europe.

<u>Highest Selling Platform in Japan Year-Wise:</u>

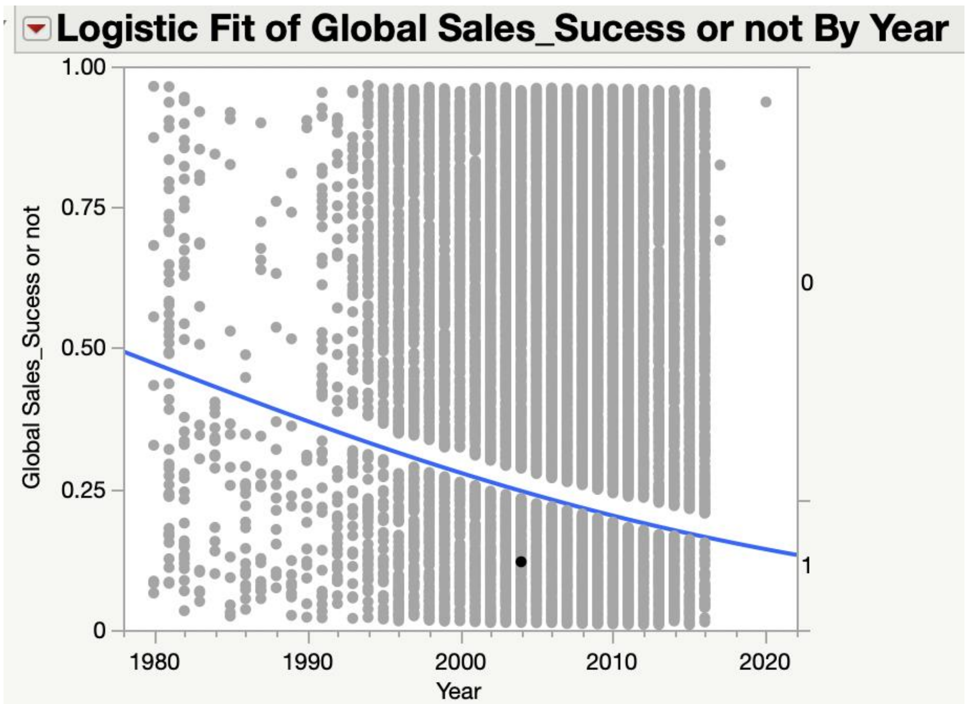


From the above graph, it can be seen that between the years 1980-2003, Nintendo was the highest selling platform for sales of video games and remained popular in the market till 2010. Then came PS as the highest selling platform for two years and after that again Nintendo became the highest selling platform in Japan.

Modeling

For the modeling, the data is partitioned into training (80%) and validation (20%) dataset. Based on these validation columns, the models are developed.

Logistic Fit:



The logistic model is used to model the probability of a certain class or event existing like successful or not. Over the past few years, the probability of a game being successful or not has been determined. It's less likely for the game to be successful based on the amount of sales.

Investors may reconsider investing in the gaming sector, but there are other criteria that lead to large sales, and games may be produced based on such features.

Decision Tree

As we are exploring the dataset, we found the decision tree might be a good model to fit different variables to the response variable. We put Platform_recode, Year, genre, NA_Sales, EU_Sales, JP_Sales, Other_Sales into the decision tree and run the model on the best r-square.

Fit Details

Measure	Training	Validation	Definition
Entropy RSquare	0.9498	0.9196	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.9911	0.9850	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.0737	0.1180	∑ -Log(ρ[j])/n
RASE	0.1419	0.1756	√ ∑(y[j]-ρ[j])²/n
Mean Abs Dev	0.0426	0.0534	∑ |y[j]-ρ[j]|/n
Misclassification Rate	0.0255	0.0373	∑ (ρ[j]≠ρMax)/n
N	13278	3320	n

Confusion Matrix

Training

Actual binning	Predicted Count				
	Top Sales	High Sales	Medium Sales	Moderate Sales	Low Sales
Top Sales	307	23	0	0	0
High Sales	14	2938	50	2	2
Medium Sales	0	42	3165	43	3
Moderate Sales	0	0	49	3089	80
Low Sales	0	0	0	30	3441

Actual binning	Predicted Rate				
	Top Sales	High Sales	Medium Sales	Moderate Sales	Low Sales
Top Sales	0.930	0.070	0.000	0.000	0.000
High Sales	0.005	0.977	0.017	0.001	0.001
Medium Sales	0.000	0.013	0.973	0.013	0.001
Moderate Sales	0.000	0.000	0.015	0.960	0.025
Low Sales	0.000	0.000	0.000	0.009	0.991

Validation

Actual binning	Predicted Count				
	Top Sales	High Sales	Medium Sales	Moderate Sales	Low Sales
Top Sales	74	10	0	0	0
High Sales	5	701	20	0	0
Medium Sales	0	14	773	18	2
Moderate Sales	0	0	9	780	31
Low Sales	0	0	0	15	868

Actual binning	Predicted Rate				
	Top Sales	High Sales	Medium Sales	Moderate Sales	Low Sales
Top Sales	0.881	0.119	0.000	0.000	0.000
High Sales	0.007	0.966	0.028	0.000	0.000
Medium Sales	0.000	0.017	0.958	0.022	0.002
Moderate Sales	0.000	0.000	0.011	0.951	0.038
Low Sales	0.000	0.000	0.000	0.017	0.983

From the fit detail results, we could see the model performance is relatively good. The misclassification rate is 0.03 on the validation dataset. The accuracy rate is high on the prediction of low sales, but relatively low on the high sales. It might be that the number of records that belong to low sales is greater than that belong to high sales.

Column Contributions

Term	Number of Splits	G^2	Portion
NA_Sales	31	19508.9654	0.5260
EU_Sales	35	9519.41471	0.2566
JP_Sales	24	7278.41314	0.1962
Other_Sales	18	695.092563	0.0187
Year	4	89.2185616	0.0024
Platform_recode	0	0	0.0000
Genre	0	0	0.0000

From the column contribution, we could see that Other_Sales, Year, Platform_recode, and Genre do not play a significant role in the model. The greatest contribution comes from the NA_Sales. It matches our analysis from the correlation matrix. In the decision tree model, we would be able to predict the sales of a new game if we have the sales of that game in the North American market.

After we removed the significant variables, we fitted a decision tree model again. The model performance dropped vastly, and we lost our view on the top sales. From the column contribution, we could conclude that all three variables are important in this model. We are not able to draw any conclusions from these three variables in the decision tree. In the visualization, we would explore further. The optimal number of splits for this partitioning will be 112.

The investors can now invest on the games with characteristics which have high probability of sales in regions based upon the leaf report.

Leaf Report for Decision Tree :

NA_Sales>=0.14&NA_Sales>=0.44&EU_Sales>=1.01&NA_Sales<1.87&EU_Sales<1.39&JP_Sales>=0.33 has a probability of 0.7944 of top sales being a top sales

NA_Sales>=0.14&NA_Sales>=0.44&EU_Sales>=1.01&NA_Sales<1.87&EU_Sales>=1.39&JP_Sales<0.07&EU_Sales<2.19 has a probability of 0.4180 being a top sales

NA_Sales>=0.14&NA_Sales>=0.44&EU_Sales>=1.01&NA_Sales<1.87&EU_Sales>=1.39&JP_Sales<0.07&EU_Sales>=2.19 has a probability of 0.9479 being a top sales

NA_Sales>=0.14&NA_Sales>=0.44&EU_Sales>=1.01&NA_Sales<1.87&EU_Sales>=1.39&JP_Sales>=0.07 has a probability of 0.9819 being a top sales

NA_Sales>=0.14&NA_Sales>=0.44&EU_Sales>=1.01&NA_Sales>=1.87 has a probability of 0.9954 as a top sales

Fit Details

Measure	Training	Validation	Definition
Entropy RSquare	0.0559	0.0542	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.1598	0.1554	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	1.3849	1.3873	∑ -Log(ρ[j])/n
RASE	0.7327	0.7327	√ ∑(y[j]-ρ[j])²/n
Mean Abs Dev	0.7213	0.7210	∑ |y[j]-ρ[j]|/n
Misclassification Rate	0.6472	0.6449	∑ (ρ[j]≠ρMax)/n
N	13278	3320	n

Confusion Matrix

Training

Actual binning	Predicted Count: Top Sales	High Sales	Medium Sales	Moderate Sales	Low Sales
Top Sales	0	175	44	76	35
High Sales	0	1477	350	752	427
Medium Sales	0	1206	403	1046	598
Moderate Sales	0	977	260	1212	769
Low Sales	0	670	144	1065	1592

Actual binning	Predicted Rate: Top Sales	High Sales	Medium Sales	Moderate Sales	Low Sales
Top Sales	0.000	0.530	0.133	0.230	0.106
High Sales	0.000	0.491	0.116	0.250	0.142
Medium Sales	0.000	0.371	0.124	0.322	0.184
Moderate Sales	0.000	0.304	0.081	0.377	0.239
Low Sales	0.000	0.193	0.041	0.307	0.459

Validation

Actual binning	Predicted Count: Top Sales	High Sales	Medium Sales	Moderate Sales	Low Sales
Top Sales	0	38	14	20	12
High Sales	0	357	79	191	99
Medium Sales	0	303	73	283	148
Moderate Sales	0	238	69	340	173
Low Sales	0	159	47	268	409

Actual binning	Predicted Rate: Top Sales	High Sales	Medium Sales	Moderate Sales	Low Sales
Top Sales	0.000	0.452	0.167	0.238	0.143
High Sales	0.000	0.492	0.109	0.263	0.136
Medium Sales	0.000	0.375	0.090	0.351	0.183
Moderate Sales	0.000	0.290	0.084	0.415	0.211
Low Sales	0.000	0.180	0.053	0.304	0.463

Column Contributions

Term	Number of Splits	G^2	Portion
Year	7	773.291744	0.3549
Platform_recode	7	753.271648	0.3457
Genre	2	652.416967	0.2994

Bootstrap Forest

We fitted the bootstrap forest using the Platform_recode, Year, Genre, NA_Sales, EU_Sales, JP_Sales, Other_Sales variables on default settings. From the overall statistics, we found that the model did not improve compared to the decision tree model. The column contribution shows the similar results as the decision tree

Column Contributions

Term	Number of Splits	G^2	Portion
NA_Sales	3996	11681.5044	0.4989
EU_Sales	4272	5074.16257	0.2167
JP_Sales	2498	4869.58551	0.2080
Other_Sales	2026	1576.16226	0.0673
Year	2518	122.369044	0.0052
Genre	688	51.7278547	0.0022
Platform_recode	366	38.4149527	0.0016

Confusion Matrix

Training

Actual binning	Predicted Count Top Sales	High Sales	Medium Sales	Moderate Sales	Low Sales
Top Sales	304	26	0	0	0
High Sales	4	2943	55	2	2
Medium Sales	0	17	3191	42	3
Moderate Sales	0	0	29	3116	73
Low Sales	0	0	0	31	3440

Validation

Actual binning	Predicted Count Top Sales	High Sales	Medium Sales	Moderate Sales	Low Sales
Top Sales	75	9	0	0	0
High Sales	4	702	20	0	0
Medium Sales	0	12	775	18	2
Moderate Sales	0	0	9	783	28
Low Sales	0	0	0	16	867

Training

Actual binning	Predicted Rate Top Sales	High Sales	Medium Sales	Moderate Sales	Low Sales
Top Sales	0.921	0.079	0.000	0.000	0.000
High Sales	0.001	0.979	0.018	0.001	0.001
Medium Sales	0.000	0.005	0.981	0.013	0.001
Moderate Sales	0.000	0.000	0.009	0.968	0.023
Low Sales	0.000	0.000	0.000	0.009	0.991

Validation

Actual binning	Predicted Rate Top Sales	High Sales	Medium Sales	Moderate Sales	Low Sales
Top Sales	0.893	0.107	0.000	0.000	0.000
High Sales	0.006	0.967	0.028	0.000	0.000
Medium Sales	0.000	0.015	0.960	0.022	0.002
Moderate Sales	0.000	0.000	0.011	0.955	0.034
Low Sales	0.000	0.000	0.000	0.018	0.982

Cumulative Validation

Per-Tree Summaries

Column Contributions

Term	Number of Splits	G^2	Portion
NA_Sales	4156	11150.7884	0.4767
JP_Sales	2601	4885.83301	0.2089
EU_Sales	4223	4593.70518	0.1964
Other_Sales	1946	2534.50163	0.1083
Year	2449	126.166327	0.0054
Genre	674	53.4730749	0.0023
Platform_recode	382	47.9217652	0.0020

We tried to remove the significant variables for the bootstrap forest and fit the model again using the default settings. The model performance also dropped vastly. Based on the column contribution, the North America has the most contribution and investing the games which are popular in this region will benefit the investors.

Overall Statistics

Measure	Training	Validation	Definition
Entropy RSquare	0.1194	0.0640	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.3121	0.1807	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	1.2918	1.3731	$\sum$ -Log(ρ[j])/n
RASE	0.7085	0.7246	$\sqrt{\sum(y[j]-\rho[j])^2/n}$
Mean Abs Dev	0.6944	0.7101	$\sum$ |y[j]-ρ[j]|/n
Misclassification Rate	0.5795	0.6395	$\sum(\rho[j] \neq \rho Max)/n$
N	13278	3320	n

Confusion Matrix

Training

Actual binning	Predicted Count Top Sales	High Sales	Medium Sales	Moderate Sales	Low Sales
Top Sales	0	168	69	44	49
High Sales	0	1361	696	424	525
Medium Sales	0	791	1224	595	643
Moderate Sales	0	634	694	1008	882
Low Sales	0	447	466	568	1990

Validation

Actual binning	Predicted Count Top Sales	High Sales	Medium Sales	Moderate Sales	Low Sales
Top Sales	0	38	22	7	17
High Sales	0	272	184	130	140
Medium Sales	0	223	212	177	195
Moderate Sales	0	181	163	230	246
Low Sales	0	104	140	156	483

Training

Actual binning	Predicted Rate Top Sales	High Sales	Medium Sales	Moderate Sales	Low Sales
Top Sales	0.000	0.509	0.209	0.133	0.148
High Sales	0.000	0.453	0.232	0.141	0.175
Medium Sales	0.000	0.243	0.376	0.183	0.198
Moderate Sales	0.000	0.197	0.216	0.313	0.274
Low Sales	0.000	0.129	0.134	0.164	0.573

Validation

Actual binning	Predicted Rate Top Sales	High Sales	Medium Sales	Moderate Sales	Low Sales
Top Sales	0.000	0.452	0.262	0.083	0.202
High Sales	0.000	0.375	0.253	0.179	0.193
Medium Sales	0.000	0.276	0.263	0.219	0.242
Moderate Sales	0.000	0.221	0.199	0.280	0.300
Low Sales	0.000	0.118	0.159	0.177	0.547

Neural Network

The neural network model is run to identify the underlying insights from the data. A neural network with various nodes was run, and the best performing model was found to be a neural network with a first layer and second layer with three nodes. This approach aids in the classification of data into categories, allowing for improved decision-making and insight into which game genres to invest in.

The misclassification rate for training dataset was 0.0108621 and validation the dataset is 0.0125999.

Neural

Validation Column: Validation

Model Launch

Model NTanH(3)NTanH2(3)

Training

binning

Measures	Value
Generalized RSquare	0.9963533
Entropy RSquare	0.9785373
RASE	0.0944945
Mean Abs Dev	0.0213686
Misclassification Rate	0.0108621
-LogLikelihood	411.84616
Sum Freq	13073

Confusion Matrix

Actual	Predicted Count				
binning	Top Sales	High Sales	Medium Sales	Moderate Sales	Low Sales
Top Sales	330	0	0	0	0
High Sales	0	2956	7	0	0
Medium Sales	0	8	3170	18	0
Moderate Sales	0	0	34	3075	54
Low Sales	0	0	0	21	3400

Confusion Rates

Actual	Predicted Rate				
binning	Top Sales	High Sales	Medium Sales	Moderate Sales	Low Sales
Top Sales	1.000	0.000	0.000	0.000	0.000
High Sales	0.000	0.998	0.002	0.000	0.000
Medium Sales	0.000	0.003	0.992	0.006	0.000
Moderate Sales	0.000	0.000	0.011	0.972	0.017
Low Sales	0.000	0.000	0.000	0.006	0.994

Validation

binning

Measures	Value
Generalized RSquare	0.9957108
Entropy RSquare	0.9749191
RASE	0.103625
Mean Abs Dev	0.022884
Misclassification Rate	0.0125999
-LogLikelihood	119.72914
Sum Freq	3254

Confusion Matrix

Actual	Predicted Count				
binning	Top Sales	High Sales	Medium Sales	Moderate Sales	Low Sales
Top Sales	82	0	0	0	0
High Sales	0	715	2	0	0
Medium Sales	0	0	786	3	0
Moderate Sales	0	0	4	782	24
Low Sales	0	0	0	8	848

Confusion Rates

Actual	Predicted Rate				
binning	Top Sales	High Sales	Medium Sales	Moderate Sales	Low Sales
Top Sales	1.000	0.000	0.000	0.000	0.000
High Sales	0.000	0.997	0.003	0.000	0.000
Medium Sales	0.000	0.000	0.996	0.004	0.000
Moderate Sales	0.000	0.000	0.005	0.965	0.030
Low Sales	0.000	0.000	0.000	0.009	0.991

Model Comparison:

Measures of Fit for binning

Creator	.2 .4 .6 .8	Entropy RSquare	Generalized RSquare	Mean -Log p	RASE	Mean Abs Dev	Misclassification Rate	N
Partition		0.9437	0.9899	0.0825	0.1492	0.0448	0.0278	16598
		.	.	.	.	.	0.0278	16598
Bootstrap Forest		0.9459	0.9903	0.0794	0.1457	0.0600	0.0242	16598
		.	.	.	.	.	0.0242	16598
Neural Model NTanH(3)NTanH2(3)		0.9757	0.9958	0.0357	0.0999	0.0249	0.0118	16327
		.	.	.	.	.	0.0280	16598

Based on the misclassification rate, we can conclude that the Bootstrap forest performs well when compared to other variables.

Conclusion

We used the dataset that contains data about video games sales. This data set contains video games sales for different regions over the years and includes information such as platform, year, genre NA sales, EU sales, JP sales, Global sales etc. We performed data cleaning, and data preprocessing by methods of missing value analysis, outlier analysis and previewing the distribution. This allowed us to exclude and hide columns which did not add value to our predictions. We also used histograms and boxplots to find different outliers in the distributions. The number of missing values is relatively small compared to the total number of records, so we decided to exclude those records. We transformed the value of platforms into the companies that developed the platforms.

We also performed exploratory data analysis by using correlation analysis and k-means cluster analysis. These methods of data exploration were used to extract important variables and allow us to view important information from the data to develop insights. K-Means clustering allowed us to find groups in the dataset which have not been explicitly labelled.

In our insights and analysis, we performed data visualization to gain a clear understanding of the information by giving it visual context through graphs and charts. Visualization was conducted using tabulate and graphs to analyze the data set. By methods of data pre-processing, data exploration and analysis, the data set has been fully evaluated.

After data cleaning and exploration, we performed logistic fit, decision tree classification, bootstrap model, Neural network and compared all the models. From model comparison, we have concluded that Bootstrap Forest is the best model for classification because of its lower misclassification rate compared to other models. Investors may use this methodology to design and invest in games that have a high likelihood of being top sellers.

We concluded that video games sales in the North American region have shown consistency over the years. The European region has gradually matched North America whereas Japan has fallen behind in terms of sales scale. The popularity for the Xbox platform is thriving in the North American region and providing many opportunities to reach new players and expand sales, thanks to popularity for the shooter genre that have strengthened its key franchises and can produce impressive sales and earnings growth. By knowing the trends in genre, regions preferences, popularity of publisher platforms, can help in predicting how successful the games could be.

References:

[1]. https://en.wikipedia.org/wiki/Video_game_industry

[2].https://www.statista.com/statistics/292056/video-game-market-value-worldwide/

[3]. www.kaggle.com/gregorut/videogamesales