OPIM 5604: FINAL PROJECT

GROUP #9

Group Lead : Naveen Abboju

Group member #1: Hany Jagwani

Group member #2: Karol Kuczynski

TITLE: VIDEO GAME SALES

UNIVERSITY OF CONNECTICUT
MSBAPM

Table of Contents	Page Number
Executive Summary	2
Introduction	3
About the Data Set	4
Data Source	5
SEMMA	5
Analyzing the Missing Values	7
Multivariate Analysis	7
Distribution Analysis	8
Exploring Outliers in the Data Set	11
Transforming the Variables	12
After the transformation of variables	13
Principal Component Analysis	15
Clustering	16
Data Visualization	18
Make Validation Column	26
Modeling	**27-34**
Regression Tree	27
Bootstrap Forest	28
Boosted Tree	29
Model Screening	30
Conclusion	34

EXECUTIVE SUMMARY:

The video game sales dataset contains sales figures for video games released between 1980 and 2020 for more than 17,000 games from 58 different platforms. It contains information such as the title of the game, the platform the game was released on, the year it was released, the genre the game belongs to, the publisher, the number of players, the sales in different regions, and the total global sales. This dataset is useful to gain valuable insights into the video game industry, such as which genres are the most popular, which platforms are the most successful, and which publishers have the highest sales. Additionally, the dataset can be used to identify trends in the industry, such as the increasing popularity of mobile gaming and the increasing focus on digital sales. This dataset is ideal for those looking to gain a better understanding of the video game industry and its development over time.

It can be used to identify trends in the video game industry, analyze the performance of different platforms, and compare the success of different genres. The dataset is an invaluable resource for those interested in video game history, sales, and analytics.

The video game sales dataset is to provide an analysis of the sales of video games across all platforms worldwide. The dataset includes titles, platforms, year of release, genre, publisher, NA_sales, EU_sales, JP_sales, and Other_sales. This allows for the analysis of the different types of games, the sales of those games across different regions, and the trends in video game sales over time.

Based on our findings, developers should utilize the Decision Tree model derived from our dataset to calculate which type of video game will sell the best in any given area, as well as tell the gamer which game will be worthwhile to pick up and play.

INTRODUCTION:

Video game sales data and market analysis

This dataset offers data on the sales figures and performance of video games. It covers all regions and provides detailed information on the platforms, release dates, genres, publishers, developers, and other relevant data.

This data set is perfect for anyone looking to analyze the video game market and identify trends. It can be used to track the sales of individual games, discover which genres are most popular, and forecast future sales.

Data on video game sales is essential for anyone involved in the video game industry, from developers and publishers to retailers and investors. This dataset provides the most comprehensive and up-to-date data available, making it an invaluable resource for anyone looking to understand the video game market.

It is important to analyze the most successful video games in terms of their genre and region. The success of a video game also depends on the type of platform it is created for. This can help in understanding the target audience of the video game and can be a useful guide for game developers.

From this analysis, we predict the most successful video games which will help us understand the reasons behind the success or failure of video games. The analysis can be used to modify the production of video games by the most lucrative strategy since the development of video games requires a substantial investment.

ABOUT THE DATA SET:

The dataset includes information such as the name of the game, the platform on which it was released, the region in which it was released, the release date, the genre, the publisher, and the number of copies sold. This dataset contains 16,598 items and includes a list of video games that have sold more than 100,000 copies. It was created from data that was scraped from Charts.

Fields included in the dataset:

- Rank - Ranking of overall sales
- Name - The games name
- Platform - Platform of the game's 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.

Predictors of the dataset:

- Overall sales: the total number of units sold
- Average price: the average price of a game
- Platform: the platform the game was released on

- Genre: the genre of the game
- Year: the year the game was released

Reviews

- Overall reviews: the total number of reviews for the game
- Positive reviews: the number of positive reviews for the game
- Negative reviews: the number of negative reviews for the game

Forecasts

- Sales forecast: the projected sales for the game in the future
- Price forecast: the projected price for the game in the future

DATA SOURCE:

There are 16,598 records in the data collection regarding video game sales that were obtained from Kaggle. The variables in this data set are the rank of total sales, name of the game, gaming platforms, year of release, genre, publisher, sales in North America, Europe, Japan, the rest of the globe, and worldwide sales.

SEMMA (SAMPLE, EXPLORE, MODIFY, MODEL, & ASSES):

SAMPLE: This stage involves choosing a portion of the suitable volume dataset from the larger dataset to create the model. To identify variables influencing the process at this early stage, the information will be separated into training, validation, and testing sets. We'll use JMP's Make Validation Column tool to build these samples.

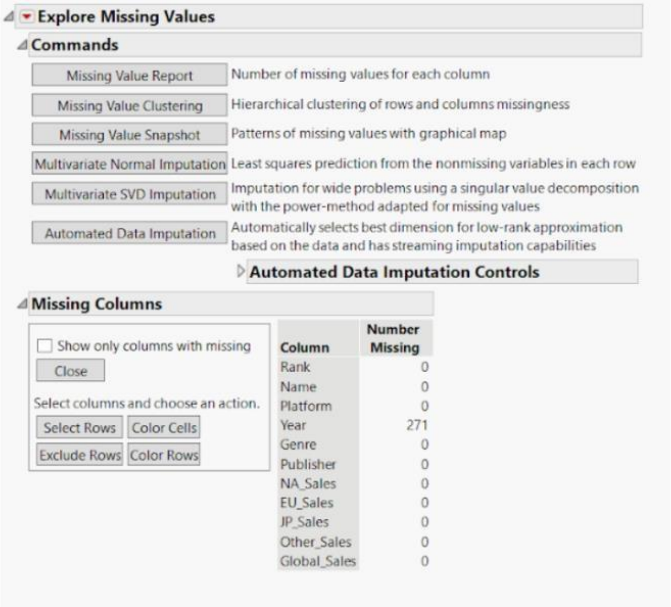
EXPLORE: The analysis of the correlations among the various data elements as well as the identification of missing information is done in this step using both univariate and multivariate techniques. While univariate analysis focuses on each factor separately, multivariate analysis examines the relationships between variables. All these elements could have an impact on the analysis's conclusion, which will rely primarily on data visualization. JMP's distribution, multivariate, and correlation tools will be used for this.

MODIFY: Using application logic, we will extract what we discover through research. Before being passed on to the modeling stage, the data is analyzed, cleaned, improved, and transformed. From this, we derive JMP's formula columns and tools like the missing data pattern, outlier analysis, principal component analysis, recode, binning, transformations, and standardization.

MODEL: The variables are modified, and the data is cleaned. Data mining methods can be applied to develop a predicted model of how the data produces the desired outcome. The JMP modeling tools fit model, partition, bootstrap forest, boosted tree, and model screening will all be used.

ASSES: We assess the effectiveness and dependability of the model. The data can now be used to determine performance efficacy after being tested.

ANALYZING THE MISSING VALUES:



Missing Value Report	Number of missing values for each column
Missing Value Clustering	Hierarchical clustering of rows and columns missingness
Missing Value Snapshot	Patterns of missing values with graphical map
Multivariate Normal Imputation	Least squares prediction from the nonmissing variables in each row
Multivariate SVD Imputation	Imputation for wide problems using a singular value decomposition with the power-method adapted for missing values
Automated Data Imputation	Automatically selects best dimension for low-rank approximation based on the data and has streaming imputation capabilities

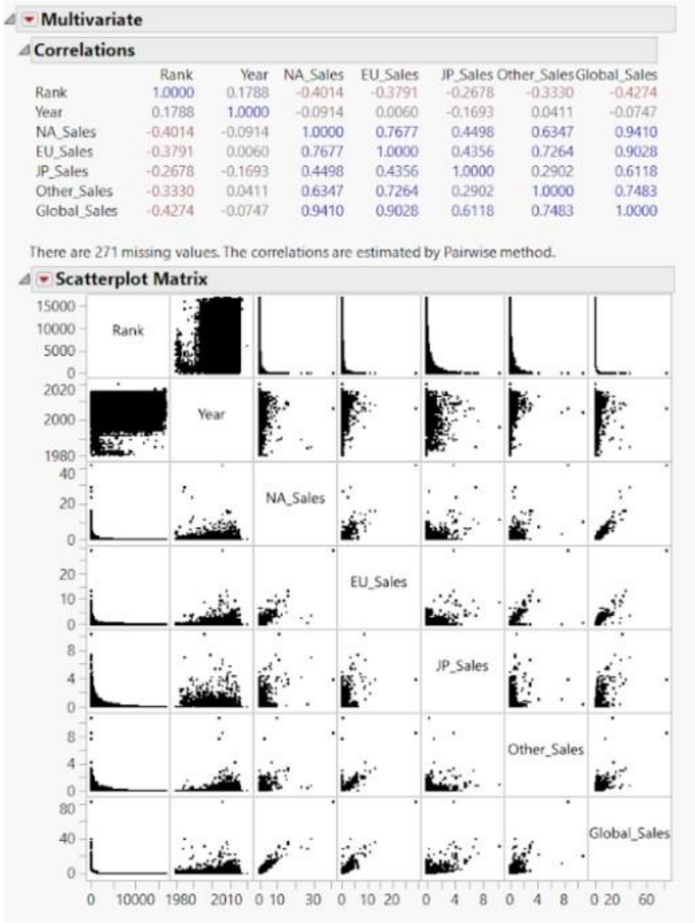
Select columns and choose an action.

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

By analyzing the missing values in the dataset, there are 271 missing values present in the year. The number is low compared to the total number of records present in the original dataset. And no missing values are present in the remaining fields of the dataset. So, we can exclude the missing values from the year field.

MULTIVARIATE ANALYSIS:

The multivariate analysis is used to find the correlation among the variables. By performing multivariate analysis on the video game sales data set we found that all the variables are strongly correlated to the target variable. Hence, we are not excluding the fields as they are strongly correlated.



Multivariate

Correlations

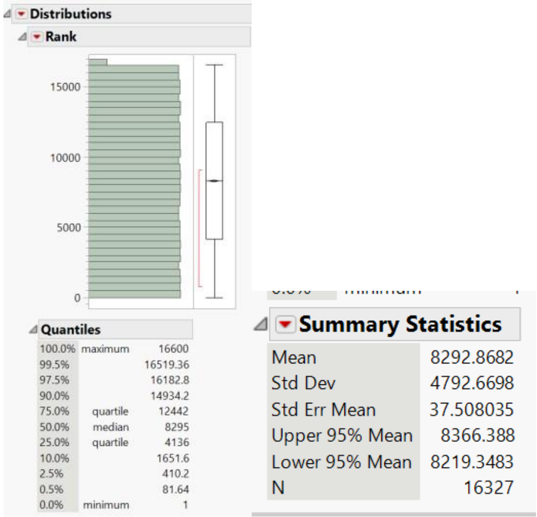
	Rank	Year	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales
Rank	1.0000	0.1788	-0.4014	-0.3791	-0.2678	-0.3330	-0.4274
Year	0.1788	1.0000	-0.0914	0.0060	-0.1693	0.0411	-0.0747
NA_Sales	-0.4014	-0.0914	1.0000	0.7677	0.4498	0.6347	0.9410
EU_Sales	-0.3791	0.0060	0.7677	1.0000	0.4356	0.7264	0.9028
JP_Sales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

There are 271 missing values. The correlations are estimated by Pairwise method.

Scatterplot Matrix

DISTRIBUTION ANALYSIS:

Rank: The distribution analysis for the rank variable depicts that it is uniformly distributed.



Distributions

Rank

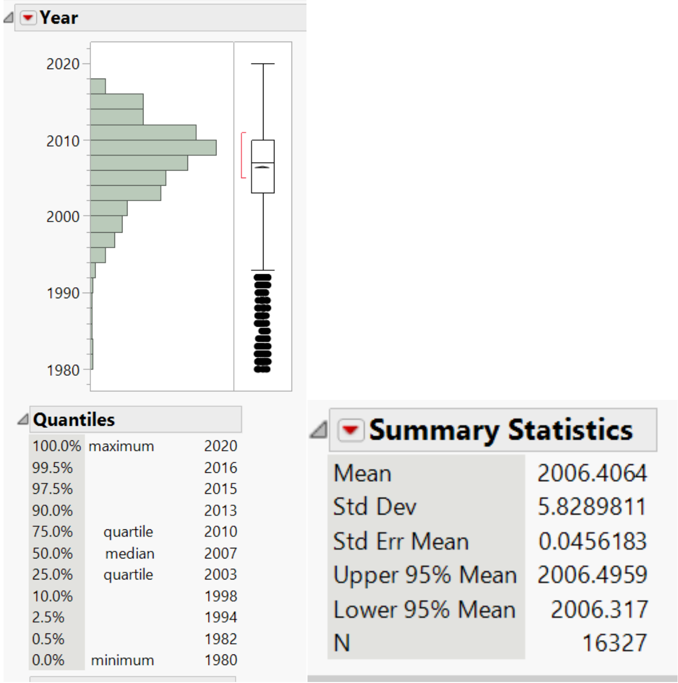
Quantiles

100.0%	maximum	16600
99.5%		16519.36
97.5%		16182.8
90.0%		14934.2
75.0%	quartile	12442
50.0%	median	8295
25.0%	quartile	4136
10.0%		1651.6
2.5%		410.2
0.5%		81.64
0.0%	minimum	1

Summary Statistics

Mean	8292.8682
Std Dev	4792.6698
Std Err Mean	37.508035
Upper 95% Mean	8366.388
Lower 95% Mean	8219.3483
N	16327

Year: The year variable is normally distributed with few outliers present in the analysis.



Quantiles

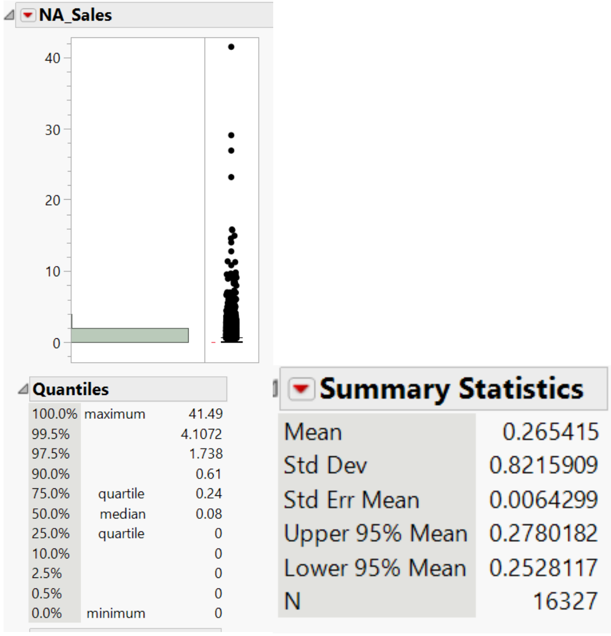
100.0%	maximum	2020
99.5%		2016
97.5%		2015
90.0%		2013
75.0%	quartile	2010
50.0%	median	2007
25.0%	quartile	2003
10.0%		1998
2.5%		1994
0.5%		1982
0.0%	minimum	1980

Summary Statistics

Mean	2006.4064
Std Dev	5.8289811
Std Err Mean	0.0456183
Upper 95% Mean	2006.4959
Lower 95% Mean	2006.317
N	16327

The sales distribution in North America, Europe, Japan, other, and worldwide

is not evenly distributed and it has outliers in the data.

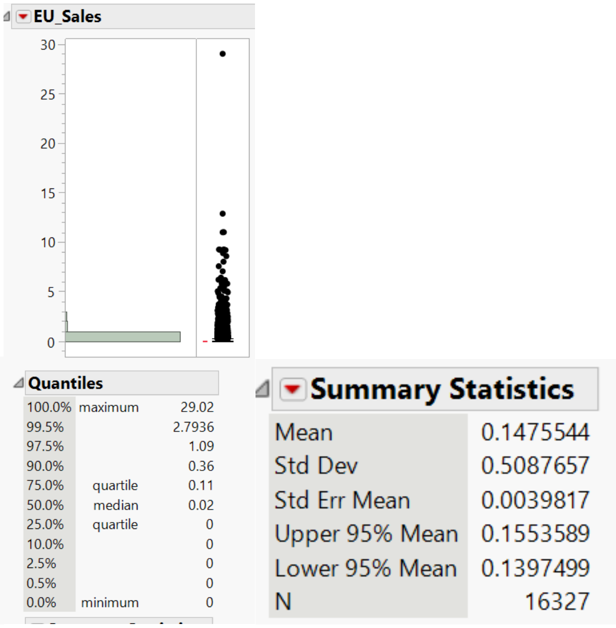


Quantiles

100.0%	maximum	41.49
99.5%		4.1072
97.5%		1.738
90.0%		0.61
75.0%	quartile	0.24
50.0%	median	0.08
25.0%	quartile	0
10.0%		0
2.5%		0
0.5%		0
0.0%	minimum	0

Summary Statistics

Mean	0.265415
Std Dev	0.8215909
Std Err Mean	0.0064299
Upper 95% Mean	0.2780182
Lower 95% Mean	0.2528117
N	16327

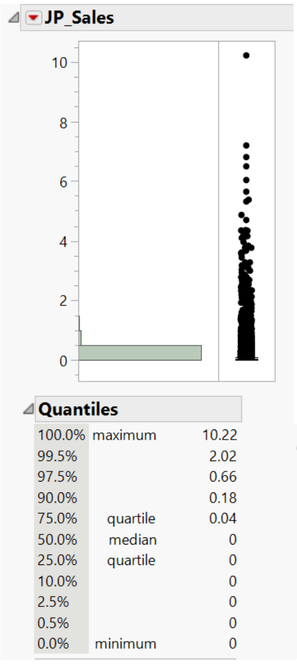


Quantiles

100.0%	maximum	29.02
99.5%		2.7936
97.5%		1.09
90.0%		0.36
75.0%	quartile	0.11
50.0%	median	0.02
25.0%	quartile	0
10.0%		0
2.5%		0
0.5%		0
0.0%	minimum	0

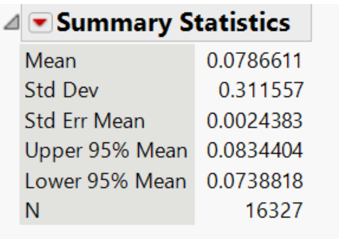
Summary Statistics

Mean	0.1475544
Std Dev	0.5087657
Std Err Mean	0.0039817
Upper 95% Mean	0.1553589
Lower 95% Mean	0.1397499
N	16327



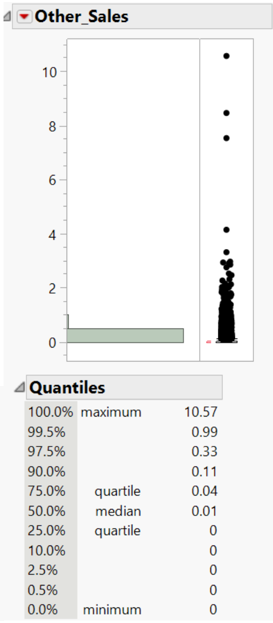
Quantiles

100.0%	maximum	10.22
99.5%		2.02
97.5%		0.66
90.0%		0.18
75.0%	quartile	0.04
50.0%	median	0
25.0%	quartile	0
10.0%		0
2.5%		0
0.5%		0
0.0%	minimum	0



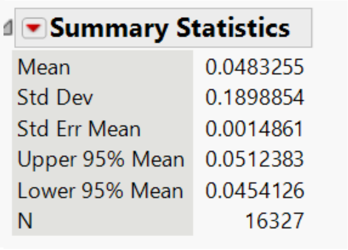
Summary Statistics

Mean	0.0786611
Std Dev	0.311557
Std Err Mean	0.0024383
Upper 95% Mean	0.0834404
Lower 95% Mean	0.0738818
N	16327



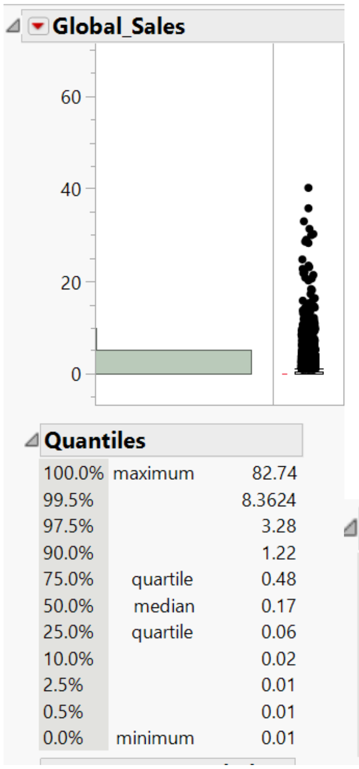
Quantiles

100.0%	maximum	10.57
99.5%		0.99
97.5%		0.33
90.0%		0.11
75.0%	quartile	0.04
50.0%	median	0.01
25.0%	quartile	0
10.0%		0
2.5%		0
0.5%		0
0.0%	minimum	0



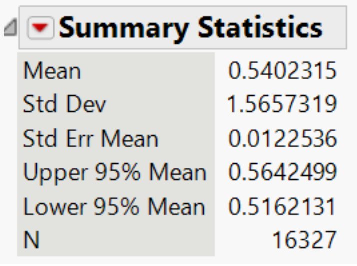
Summary Statistics

Mean	0.0483255
Std Dev	0.1898854
Std Err Mean	0.0014861
Upper 95% Mean	0.0512383
Lower 95% Mean	0.0454126
N	16327



Quantiles

100.0%	maximum	82.74
99.5%		8.3624
97.5%		3.28
90.0%		1.22
75.0%	quartile	0.48
50.0%	median	0.17
25.0%	quartile	0.06
10.0%		0.02
2.5%		0.01
0.5%		0.01
0.0%	minimum	0.01

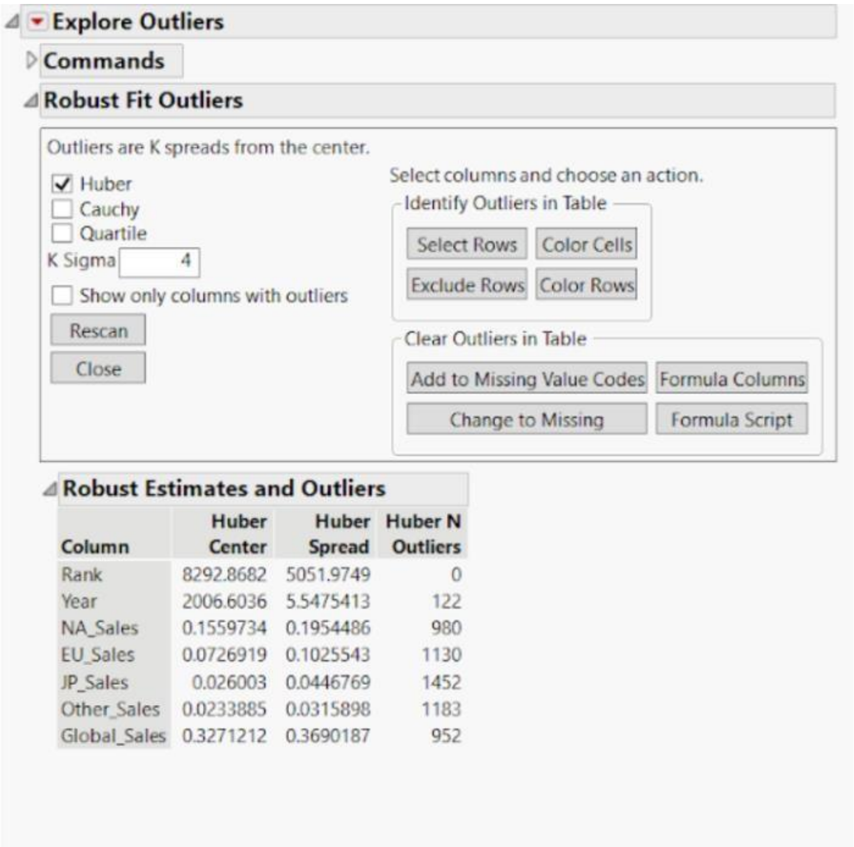


Summary Statistics

Mean	0.5402315
Std Dev	1.5657319
Std Err Mean	0.0122536
Upper 95% Mean	0.5642499
Lower 95% Mean	0.5162131
N	16327

EXPLORING OUTLIERS IN THE DATASET:

After exploring the outliers in the dataset in each variable of the dataset we found the outliers present in the columns: year, NA_Sales, EU_Sales, JP_Sales, Other_sales, & Global_sales. The below picture shows the number of outliers present in the variables as well as the outliers present outside the whiskers by performing the distribution analysis.



Explore Outliers

Commands

Robust Fit Outliers

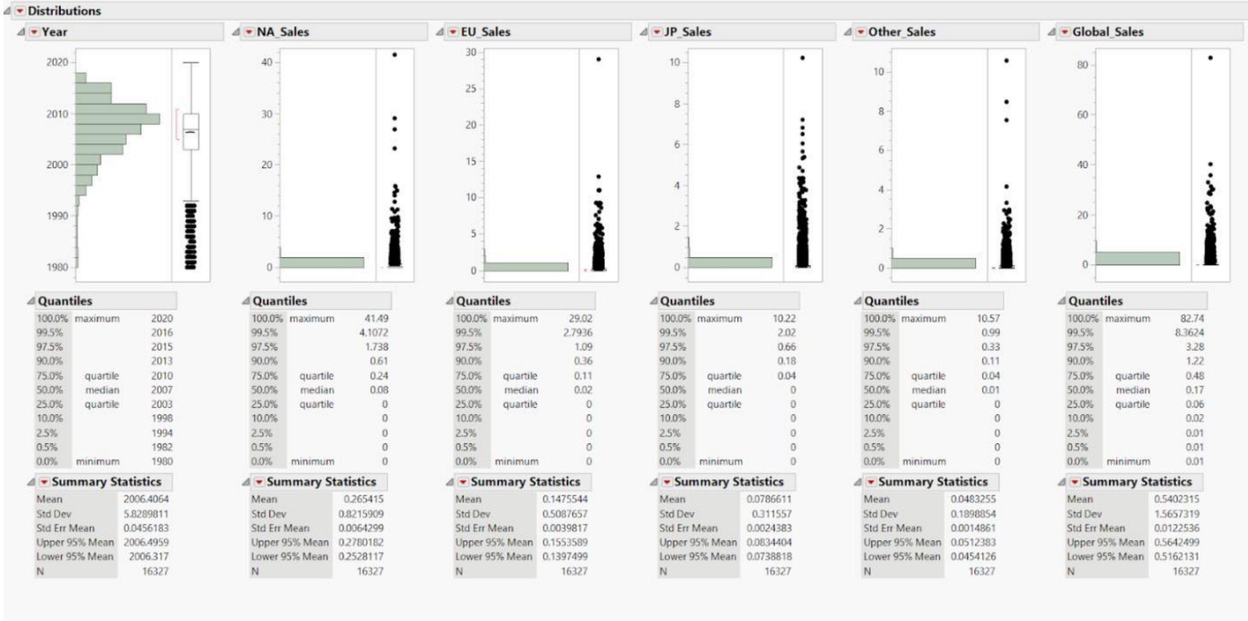
Outliers are K spreads from the center.

☑ Huber
□ Cauchy
□ Quartile
K Sigma 4
□ Show only columns with outliers

Select columns and choose an action.

Robust Estimates and Outliers

Column	Huber Center	Huber Spread	Huber N Outliers
Rank	8292.8682	5051.9749	0
Year	2006.6036	5.5475413	122
NA_Sales	0.1559734	0.1954486	980
EU_Sales	0.0726919	0.1025543	1130
JP_Sales	0.026003	0.0446769	1452
Other_Sales	0.0233885	0.0315898	1183
Global_Sales	0.3271212	0.3690187	952



Quantiles

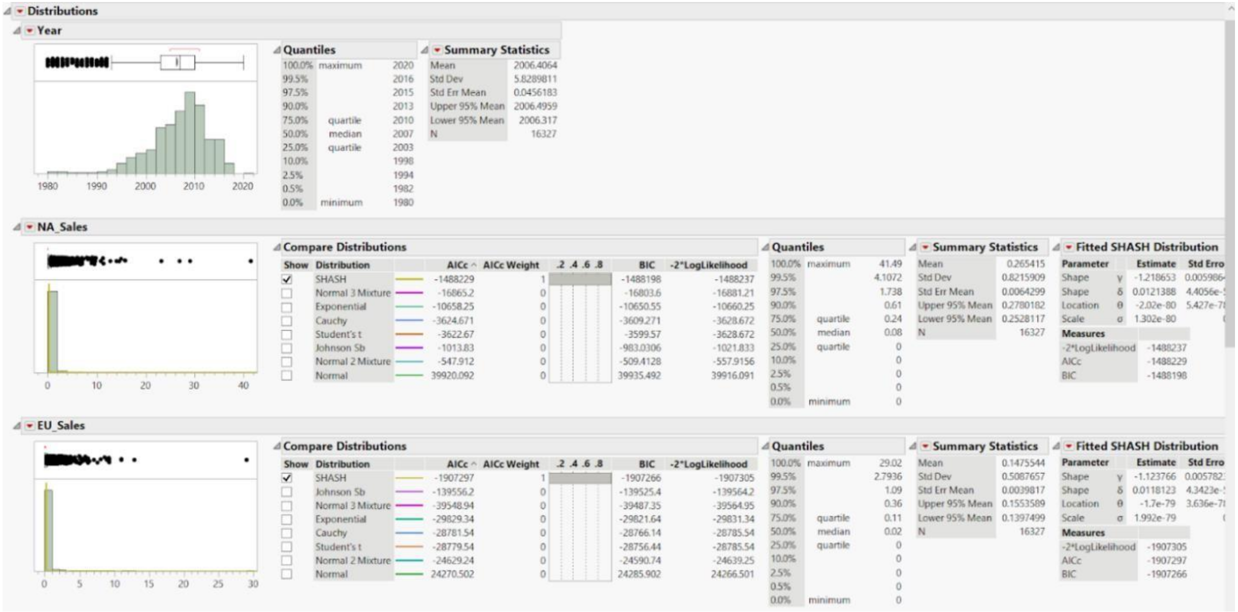
		Year	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales
100.0%	maximum	2020	41.49	29.02	10.22	10.57	82.74
99.5%		2016	4.1072	2.7936	2.02	0.99	8.3624
97.5%		2015	1.738	1.09	0.66	0.33	3.28
90.0%		2013	0.61	0.36	0.18	0.11	1.22
75.0%	quartile	2010	0.24	0.11	0.04	0.04	0.48
50.0%	median	2007	0.08	0.02	0	0.01	0.17
25.0%	quartile	2003	0	0	0	0	0.06
10.0%		1998	0	0	0	0	0.02
2.5%		1994	0	0	0	0	0.01
0.5%		1982	0	0	0	0	0.01
0.0%	minimum	1980	0	0	0	0	0.01

Summary Statistics

	Year	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales
Mean	2006.4064	0.265415	0.1475544	0.0786611	0.0483255	0.5402315
Std Dev	5.8289811	0.8215909	0.5087657	0.311557	0.1898854	1.5657319
Std Err Mean	0.0456183	0.0064299	0.0039817	0.0024383	0.0014861	0.0122536
Upper 95% Mean	2006.4959	0.2780182	0.1553589	0.0834404	0.0512383	0.5642499
Lower 95% Mean	2006.317	0.2528117	0.1397499	0.0738818	0.0454126	0.5162131
N	16327	16327	16327	16327	16327	16327

TRANSFORMING THE VARIABLES:

Transforming the variables with their best fit will minimize the outliers present in the variables. So that makes data clean to perform further operations. Here the best fit is SHASH for all variables except for the Global_Sales variable. The Global_Sales variable has lognormal as its best fit.



Year

Quantiles

100.0%	maximum	2020
99.5%		2016
97.5%		2015
90.0%		2013
75.0%	quartile	2010
50.0%	median	2007
25.0%	quartile	2003
10.0%		1998
2.5%		1994
0.5%		1982
0.0%	minimum	1980

Summary Statistics

Mean	2006.4064
Std Dev	5.8289811
Std Err Mean	0.0456183
Upper 95% Mean	2006.4959
Lower 95% Mean	2006.317
N	16327

NA_Sales

Compare Distributions

Show	Distribution	AICc ^	AICc Weight	BIC	-2*LogLikelihood
☑	SHASH	-1488229	1	-1488198	-1488237
☐	Normal 3 Mixture	-16865.2	0	-16803.6	-16881.21
☐	Exponential	-10658.25	0	-10650.55	-10660.25
☐	Cauchy	-3624.671	0	-3609.271	-3628.672
☐	Student's t	-3622.67	0	-3599.57	-3628.672
☐	Johnson Sb	-1013.83	0	-983.0306	-1021.833
☐	Normal 2 Mixture	-547.912	0	-509.4128	-557.9156
☐	Normal	39920.092	0	39935.492	39916.091

Quantiles

100.0%	maximum	41.49
99.5%		4.1072
97.5%		1.738
90.0%		0.61
75.0%	quartile	0.24
50.0%	median	0.08
25.0%	quartile	0
10.0%		0
2.5%		0
0.5%		0
0.0%	minimum	0

Summary Statistics

Mean	0.265415
Std Dev	0.8215909
Std Err Mean	0.0064299
Upper 95% Mean	0.2780182
Lower 95% Mean	0.2528117
N	16327

Fitted SHASH Distribution

Parameter		Estimate	Std Erro
Shape	γ	-1.218653	0.005986
Shape	δ	0.0121388	4.4056e-
Location	θ	-2.02e-80	5.427e-7
Scale	σ	1.302e-80	(

Measures	
-2*LogLikelihood	-1488237
AICc	-1488229
BIC	-1488198

EU_Sales

Compare Distributions

Show	Distribution	AICc ^	AICc Weight	BIC	-2*LogLikelihood
☑	SHASH	-1907297	1	-1907266	-1907305
☐	Johnson Sb	-139556.2	0	-139525.4	-139564.2
☐	Normal 3 Mixture	-39548.94	0	-39487.35	-39564.95
☐	Exponential	-29829.34	0	-29821.64	-29831.34
☐	Cauchy	-28781.54	0	-28766.14	-28785.54
☐	Student's t	-28779.54	0	-28756.44	-28785.54
☐	Normal 2 Mixture	-24629.24	0	-24590.74	-24639.25
☐	Normal	24270.502	0	24285.902	24266.501

Quantiles

100.0%	maximum	29.02
99.5%		2.7936
97.5%		1.09
90.0%		0.36
75.0%	quartile	0.11
50.0%	median	0.02
25.0%	quartile	0
10.0%		0
2.5%		0
0.5%		0
0.0%	minimum	0

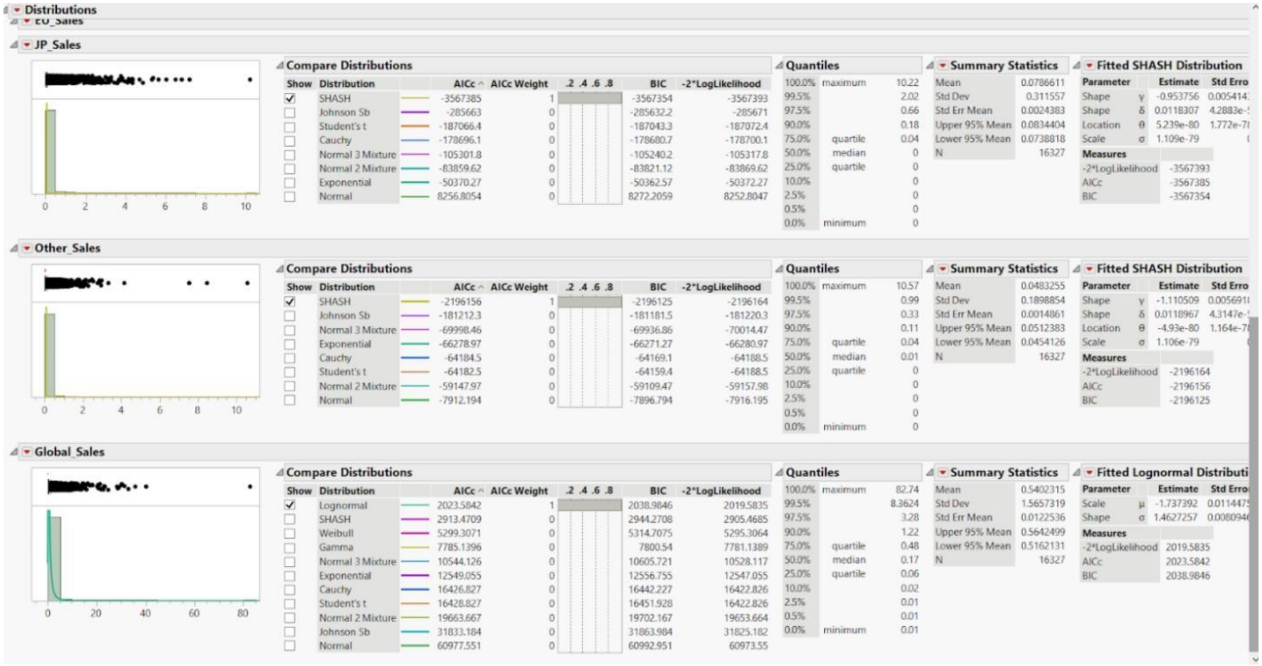
Summary Statistics

Mean	0.1475544
Std Dev	0.5087657
Std Err Mean	0.0039817
Upper 95% Mean	0.1553589
Lower 95% Mean	0.1397499
N	16327

Fitted SHASH Distribution

Parameter		Estimate	Std Erro
Shape	γ	-1.123766	0.005782
Shape	δ	0.0118123	4.3423e-
Location	θ	-1.7e-79	3.636e-7
Scale	σ	1.992e-79	(

Measures	
-2*LogLikelihood	-1907305
AICc	-1907297
BIC	-1907266



Distributions

EU_Sales

JP_Sales

Compare Distributions

Show	Distribution	AICc	AICc Weight	BIC	-2*LogLikelihood
☑	SHASH	-3567385	1	-3567354	-3567393
□	Johnson Sb	-285663	0	-285632.2	-285671
□	Student's t	-187066.4	0	-187043.3	-187072.4
□	Cauchy	-178696.1	0	-178680.7	-178700.1
□	Normal 3 Mixture	-105301.8	0	-105240.2	-105317.8
□	Normal 2 Mixture	-83859.62	0	-83821.12	-83869.62
□	Exponential	-50370.27	0	-50362.57	-50372.27
□	Normal	8256.8054	0	8272.2059	8252.8047

Quantiles

100.0%	maximum	10.22
99.5%		2.02
97.5%		0.66
90.0%		0.18
75.0%	quartile	0.04
50.0%	median	0
25.0%	quartile	0
10.0%		0
2.5%		0
0.5%		0
0.0%	minimum	0

Summary Statistics

Mean	0.0786611
Std Dev	0.311557
Std Err Mean	0.0024383
Upper 95% Mean	0.0834404
Lower 95% Mean	0.0738818
N	16327

Fitted SHASH Distribution

Parameter		Estimate	Std Erro
Shape	γ	-0.953756	0.005414
Shape	δ	0.0118307	4.2883e-
Location	θ	5.239e-80	1.772e-7
Scale	σ	1.109e-79	

Measures	
-2*LogLikelihood	-3567393
AICc	-3567385
BIC	-3567354

Other_Sales

Compare Distributions

Show	Distribution	AICc	AICc Weight	BIC	-2*LogLikelihood
☑	SHASH	-2196156	1	-2196125	-2196164
□	Johnson Sb	-181212.3	0	-181181.5	-181220.3
□	Normal 3 Mixture	-69998.46	0	-69936.86	-70014.47
□	Exponential	-66278.97	0	-66271.27	-66280.97
□	Cauchy	-64184.5	0	-64169.1	-64188.5
□	Student's t	-64182.5	0	-64159.4	-64188.5
□	Normal 2 Mixture	-59147.97	0	-59109.47	-59157.98
□	Normal	-7912.194	0	-7896.794	-7916.195

Quantiles

100.0%	maximum	10.57
99.5%		0.99
97.5%		0.33
90.0%		0.11
75.0%	quartile	0.04
50.0%	median	0.01
25.0%	quartile	0
10.0%		0
2.5%		0
0.5%		0
0.0%	minimum	0

Summary Statistics

Mean	0.0483255
Std Dev	0.1898854
Std Err Mean	0.0014861
Upper 95% Mean	0.0512383
Lower 95% Mean	0.0454126
N	16327

Fitted SHASH Distribution

Parameter		Estimate	Std Erro
Shape	γ	-1.110509	0.005691
Shape	δ	0.0118967	4.3147e-
Location	θ	-4.93e-80	1.164e-7
Scale	σ	1.106e-79	

Measures	
-2*LogLikelihood	-2196164
AICc	-2196156
BIC	-2196125

Global_Sales

Compare Distributions

Show	Distribution	AICc	AICc Weight	BIC	-2*LogLikelihood
☑	Lognormal	2023.5842	1	2038.9846	2019.5835
□	SHASH	2913.4709	0	2944.2708	2905.4685
□	Weibull	5299.3071	0	5314.7075	5295.3064
□	Gamma	7785.1396	0	7800.54	7781.1389
□	Normal 3 Mixture	10544.126	0	10605.721	10528.117
□	Exponential	12549.055	0	12556.755	12547.055
□	Cauchy	16426.827	0	16442.227	16422.826
□	Student's t	16428.827	0	16451.928	16422.826
□	Normal 2 Mixture	19663.667	0	19702.167	19653.664
□	Johnson Sb	31833.184	0	31863.984	31825.182
□	Normal	60977.551	0	60992.951	60973.55

Quantiles

100.0%	maximum	82.74
99.5%		8.3624
97.5%		3.28
90.0%		1.22
75.0%	quartile	0.48
50.0%	median	0.17
25.0%	quartile	0.06
10.0%		0.02
2.5%		0.01
0.5%		0.01
0.0%	minimum	0.01

Summary Statistics

Mean	0.5402315
Std Dev	1.5657319
Std Err Mean	0.0122536
Upper 95% Mean	0.5642499
Lower 95% Mean	0.5162131
N	16327

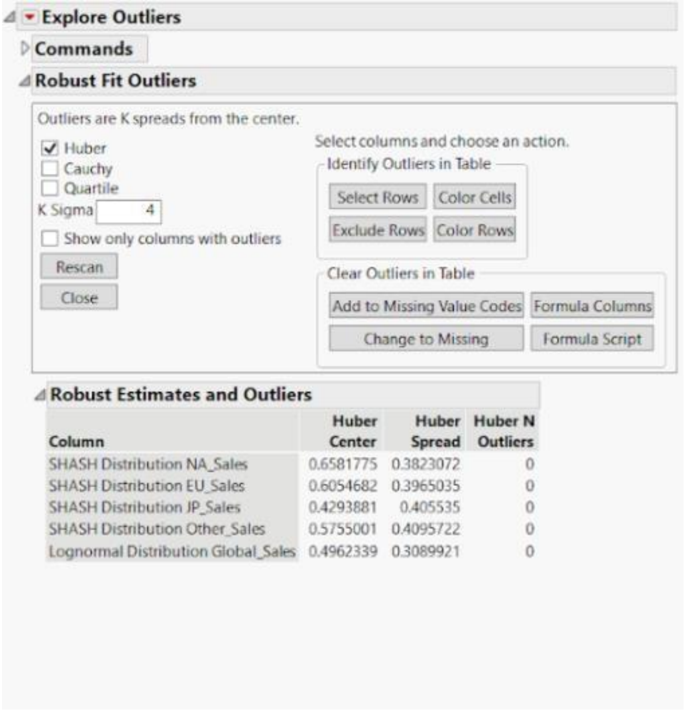
Fitted Lognormal Distributi

Parameter		Estimate	Std Erro
Scale	μ	-1.737392	0.011447
Shape	σ	1.4627257	0.008094

Measures	
-2*LogLikelihood	2019.5835
AICc	2023.5842
BIC	2038.9846

AFTER THE TRANSFORMATION OF VARIABLES:

By performing the outlier analysis on the transformed variables, we found zero outliers in all transformed variables. Hence, we can say that the data is now cleaned and ready to perform further analysis on it. The below pictures show the zero outliers present in each transformed variable and by performing the distribution analysis we can see there are no outliers present outside the whiskers.



Outliers are K spreads from the center.

☑ Huber
□ Cauchy
□ Quartile
K Sigma 4
□ Show only columns with outliers

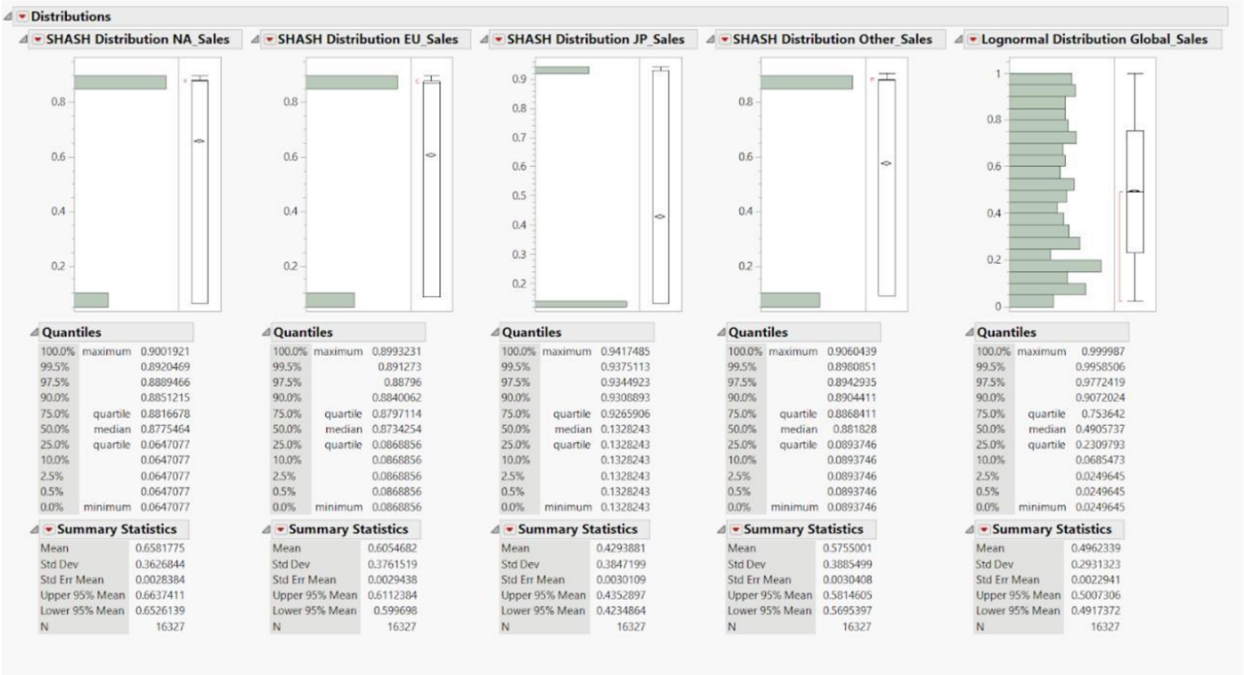
Select columns and choose an action.

Identify Outliers in Table

Clear Outliers in Table

Robust Estimates and Outliers

Column	Huber Center	Huber Spread	Huber N Outliers
SHASH Distribution NA_Sales	0.6581775	0.3823072	0
SHASH Distribution EU_Sales	0.6054682	0.3965035	0
SHASH Distribution JP_Sales	0.4293881	0.405535	0
SHASH Distribution Other_Sales	0.5755001	0.4095722	0
Lognormal Distribution Global_Sales	0.4962339	0.3089921	0


Distributions
SHASH Distribution NA_Sales
0.8
0.6
0.4
0.2
SHASH Distribution EU_Sales
0.8
0.6
0.4
0.2
SHASH Distribution JP_Sales
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
SHASH Distribution Other_Sales
0.8
0.6
0.4
0.2
Lognormal Distribution Global_Sales
1
0.8
0.6
0.4
0.2
0

SHASH Distribution NA_Sales

Quantiles

100.0%	maximum	0.9001921
99.5%		0.8920469
97.5%		0.8889466
90.0%		0.8851215
75.0%	quartile	0.8816678
50.0%	median	0.8775464
25.0%	quartile	0.0647077
10.0%		0.0647077
2.5%		0.0647077
0.5%		0.0647077
0.0%	minimum	0.0647077

Summary Statistics

Mean	0.6581775
Std Dev	0.3626844
Std Err Mean	0.0028384
Upper 95% Mean	0.6637411
Lower 95% Mean	0.6526139
N	16327

SHASH Distribution EU_Sales

Quantiles

100.0%	maximum	0.8993231
99.5%		0.891273
97.5%		0.88796
90.0%		0.8840062
75.0%	quartile	0.8797114
50.0%	median	0.8734254
25.0%	quartile	0.0868856
10.0%		0.0868856
2.5%		0.0868856
0.5%		0.0868856
0.0%	minimum	0.0868856

Summary Statistics

Mean	0.6054682
Std Dev	0.3761519
Std Err Mean	0.0029438
Upper 95% Mean	0.6112384
Lower 95% Mean	0.599698
N	16327

SHASH Distribution JP_Sales

Quantiles

100.0%	maximum	0.9417485
99.5%		0.9375113
97.5%		0.9344923
90.0%		0.9308893
75.0%	quartile	0.9265906
50.0%	median	0.1328243
25.0%	quartile	0.1328243
10.0%		0.1328243
2.5%		0.1328243
0.5%		0.1328243
0.0%	minimum	0.1328243

Summary Statistics

Mean	0.4293881
Std Dev	0.3847199
Std Err Mean	0.0030109
Upper 95% Mean	0.4352897
Lower 95% Mean	0.4234864
N	16327

SHASH Distribution Other_Sales

Quantiles

100.0%	maximum	0.9060439
99.5%		0.8980851
97.5%		0.8942935
90.0%		0.8904411
75.0%	quartile	0.8868411
50.0%	median	0.881828
25.0%	quartile	0.0893746
10.0%		0.0893746
2.5%		0.0893746
0.5%		0.0893746
0.0%	minimum	0.0893746

Summary Statistics

Mean	0.5755001
Std Dev	0.3885499
Std Err Mean	0.0030408
Upper 95% Mean	0.5814605
Lower 95% Mean	0.5695397
N	16327

Lognormal Distribution Global_Sales

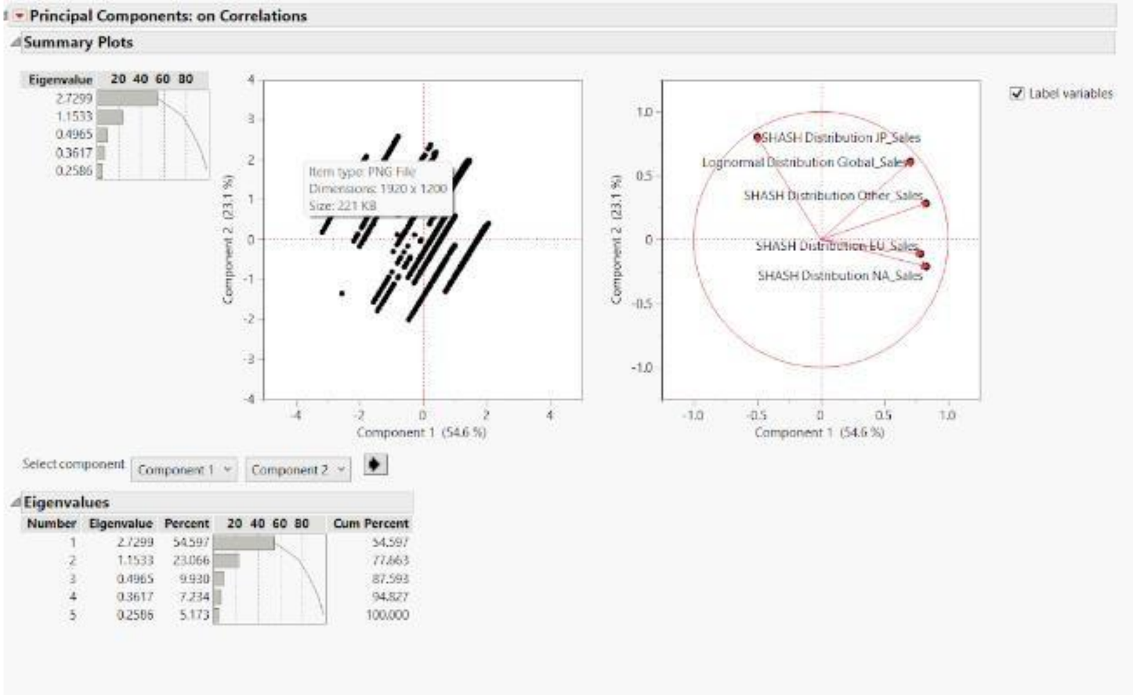
Quantiles

100.0%	maximum	0.999987
99.5%		0.9958506
97.5%		0.9772419
90.0%		0.9072024
75.0%	quartile	0.753642
50.0%	median	0.4905737
25.0%	quartile	0.2309793
10.0%		0.0685473
2.5%		0.0249645
0.5%		0.0249645
0.0%	minimum	0.0249645

Summary Statistics

Mean	0.4962339
Std Dev	0.2931323
Std Err Mean	0.0022941
Upper 95% Mean	0.5007306
Lower 95% Mean	0.4917372
N	16327

PRINCIPAL COMPONENT ANALYSIS:



Principal Components: on Correlations

Summary Plots

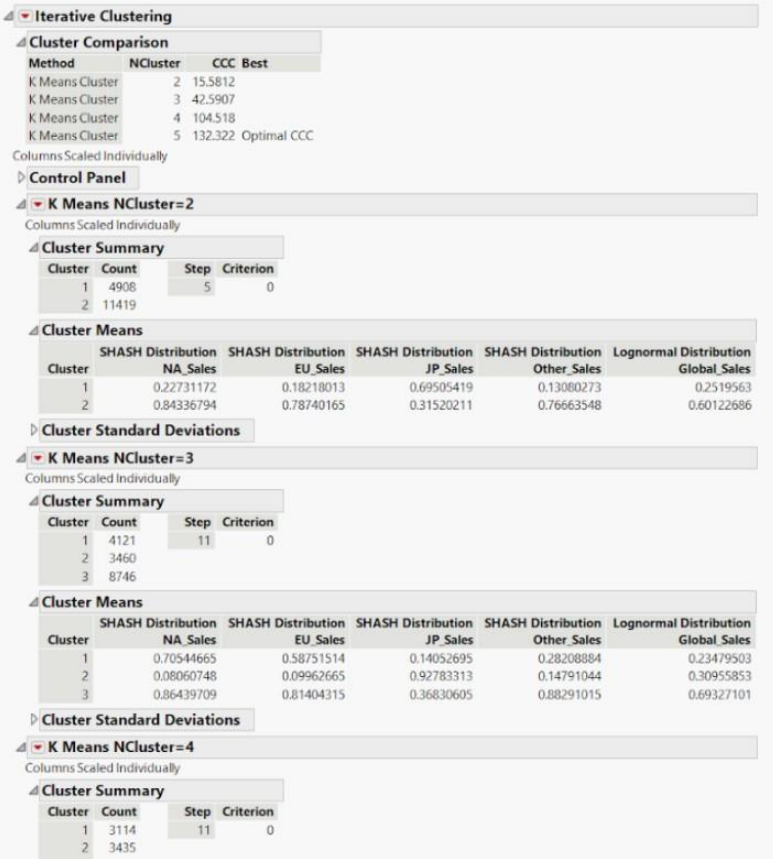
Eigenvalues

Number	Eigenvalue	Percent	20 40 60 80	Cum Percent
1	2.7299	54.597		54.597
2	1.1533	23.066		77.663
3	0.4965	9.930		87.593
4	0.3617	7.234		94.827
5	0.2586	5.173		100.000

The three principal components that should be saved are PC1, PC2, and PC3 as they cover more than 80% of variance. These components explain a large portion of the variance in the dataset, which can provide valuable insights into the underlying factors that influence the data. Additionally, these components can be used to reduce the dimensions of the data and make it more manageable.

CLUSTERING:

Clustering is a technique that is used to know the alike characteristics of the variables present in the dataset. Here we performed K-means on transformed variables. From the analysis, the ideal number of clustering is 5 with an optimal value of 132.32, and the number of clusters is 2473.



Iterative Clustering

Cluster Comparison

Method	NCluster	CCC	Best
K Means Cluster	2	15.5812	
K Means Cluster	3	42.5907	
K Means Cluster	4	104.518	
K Means Cluster	5	132.322	Optimal CCC

Columns Scaled Individually

Control Panel

K Means NCluster=2

Columns Scaled Individually

Cluster Summary

Cluster	Count
1	4908
2	11419

Step	Criterion
5	0

Cluster Means

Cluster	SHASH Distribution NA_Sales	SHASH Distribution EU_Sales	SHASH Distribution JP_Sales	SHASH Distribution Other_Sales	Lognormal Distribution Global_Sales
1	0.22731172	0.18218013	0.69505419	0.13080273	0.2519563
2	0.84336794	0.78740165	0.31520211	0.76663548	0.60122686

Cluster Standard Deviations

K Means NCluster=3

Columns Scaled Individually

Cluster Summary

Cluster	Count
1	4121
2	3460
3	8746

Step	Criterion
11	0

Cluster Means

Cluster	SHASH Distribution NA_Sales	SHASH Distribution EU_Sales	SHASH Distribution JP_Sales	SHASH Distribution Other_Sales	Lognormal Distribution Global_Sales
1	0.70544665	0.58751514	0.14052695	0.28208884	0.23479503
2	0.08060748	0.09962665	0.92783313	0.14791044	0.30955853
3	0.86439709	0.81404315	0.36830605	0.88291015	0.69327101

Cluster Standard Deviations

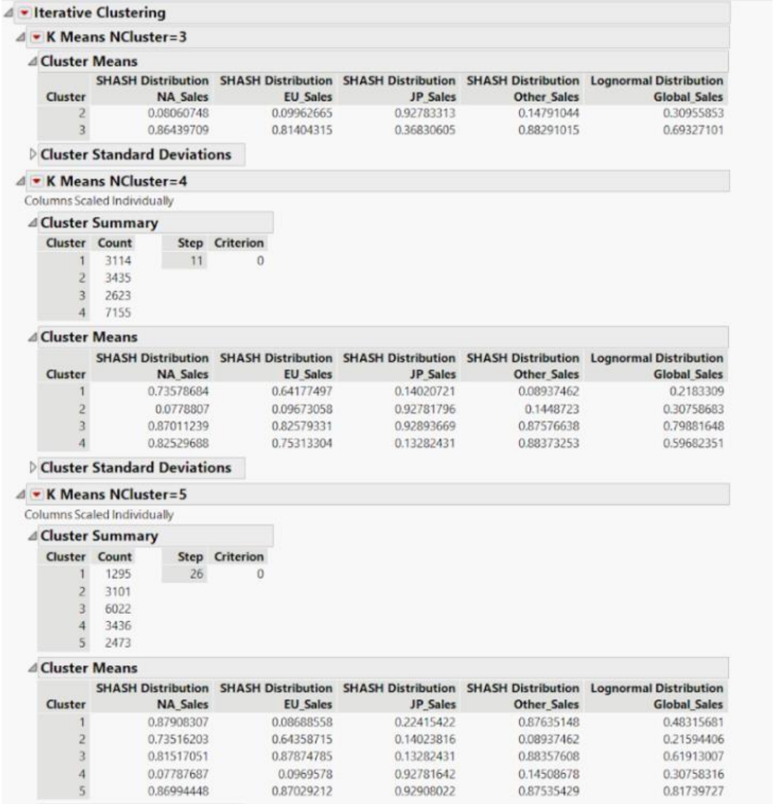
K Means NCluster=4

Columns Scaled Individually

Cluster Summary

Cluster	Count
1	3114
2	3435

Step	Criterion
11	0



Iterative Clustering

K Means NCluster=3

Cluster Means

Cluster	SHASH Distribution NA_Sales	SHASH Distribution EU_Sales	SHASH Distribution JP_Sales	SHASH Distribution Other_Sales	Lognormal Distribution Global_Sales
2	0.08060748	0.09962665	0.92783313	0.14791044	0.30955853
3	0.86439709	0.81404315	0.36830605	0.88291015	0.69327101

Cluster Standard Deviations

K Means NCluster=4

Columns Scaled Individually

Cluster Summary

Cluster	Count
1	3114
2	3435
3	2623
4	7155

Step	Criterion
11	0

Cluster Means

Cluster	SHASH Distribution NA_Sales	SHASH Distribution EU_Sales	SHASH Distribution JP_Sales	SHASH Distribution Other_Sales	Lognormal Distribution Global_Sales
1	0.73578684	0.64177497	0.14020721	0.08937462	0.2183309
2	0.0778807	0.09673058	0.92781796	0.1448723	0.30758683
3	0.87011239	0.82579331	0.92893669	0.87576638	0.79881648
4	0.82529688	0.75313304	0.13282431	0.88373253	0.59682351

Cluster Standard Deviations

K Means NCluster=5

Columns Scaled Individually

Cluster Summary

Cluster	Count
1	1295
2	3101
3	6022
4	3436
5	2473

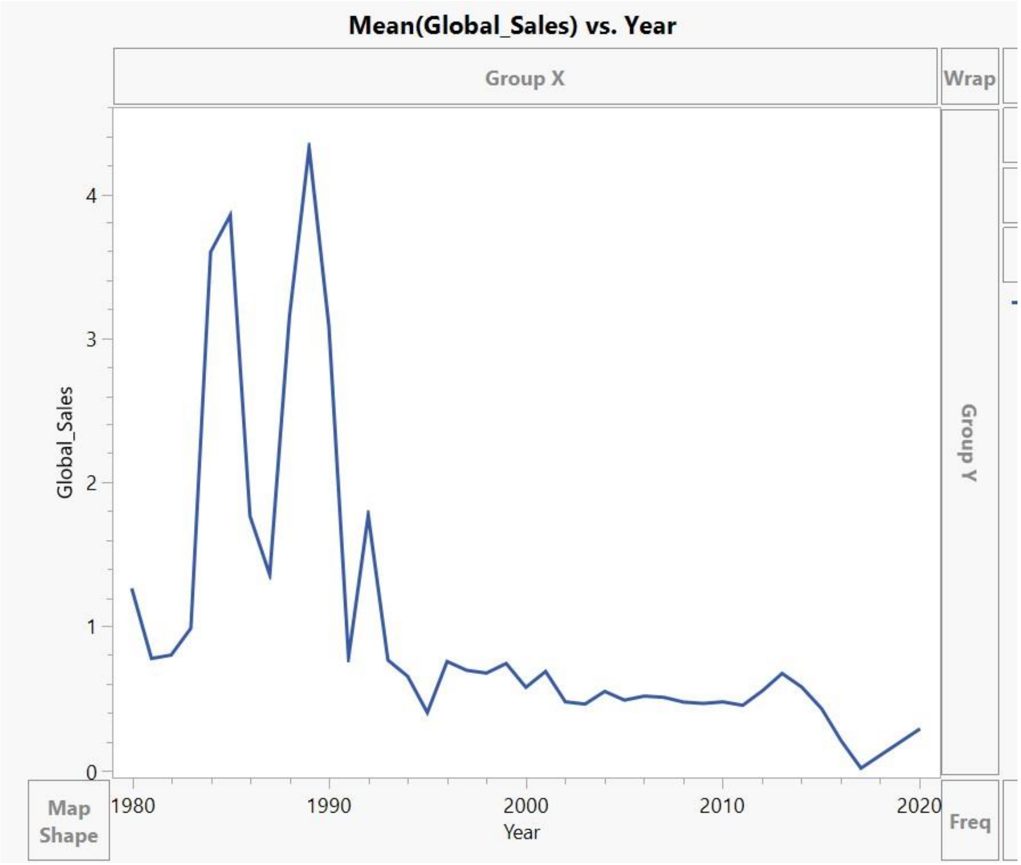
Step	Criterion
26	0

Cluster Means

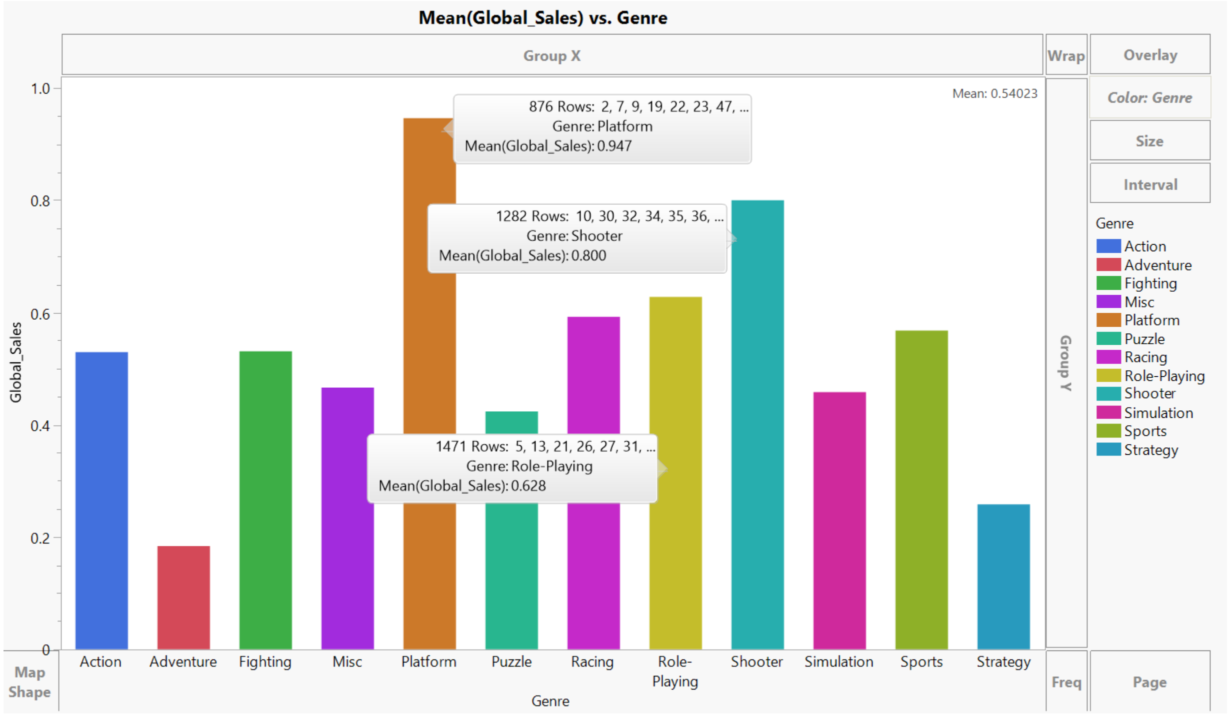
Cluster	SHASH Distribution NA_Sales	SHASH Distribution EU_Sales	SHASH Distribution JP_Sales	SHASH Distribution Other_Sales	Lognormal Distribution Global_Sales
1	0.87908307	0.08688558	0.22415422	0.87635148	0.48315681
2	0.73516203	0.64358715	0.14023816	0.08937462	0.21594406
3	0.81517051	0.87874785	0.13282431	0.88357608	0.61913007
4	0.07787687	0.0969578	0.92781642	0.14508678	0.30758316
5	0.86994448	0.87029212	0.92908022	0.87535429	0.81739727

DATA VISUALIZATION:

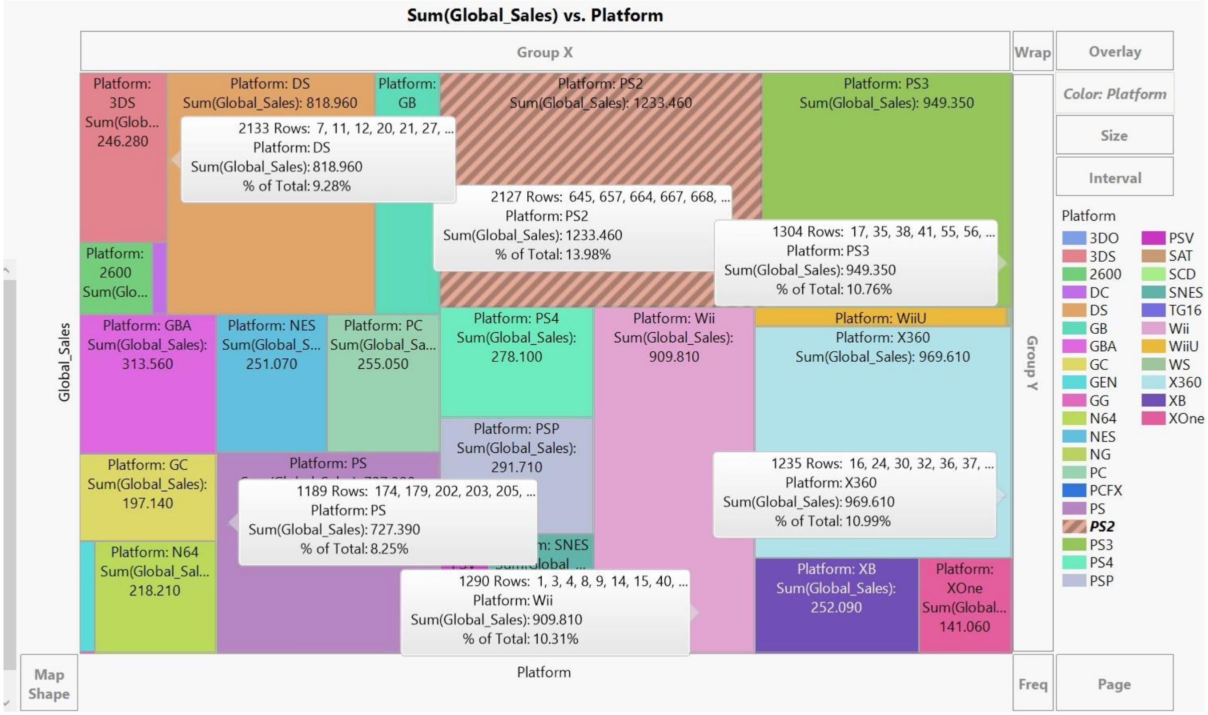
Global_Sales vs. Year: In the early 80s video game sales followed the upward trend and peaked in the mid-80s then declined gradually. And again, in the early start of the 90s, it gradually increased and reached its maximum sales. Then after it has seen some fluctuations as each year passes. From the visualization, we can say that at present video game sales started from their minimum point and it is raising their bar.



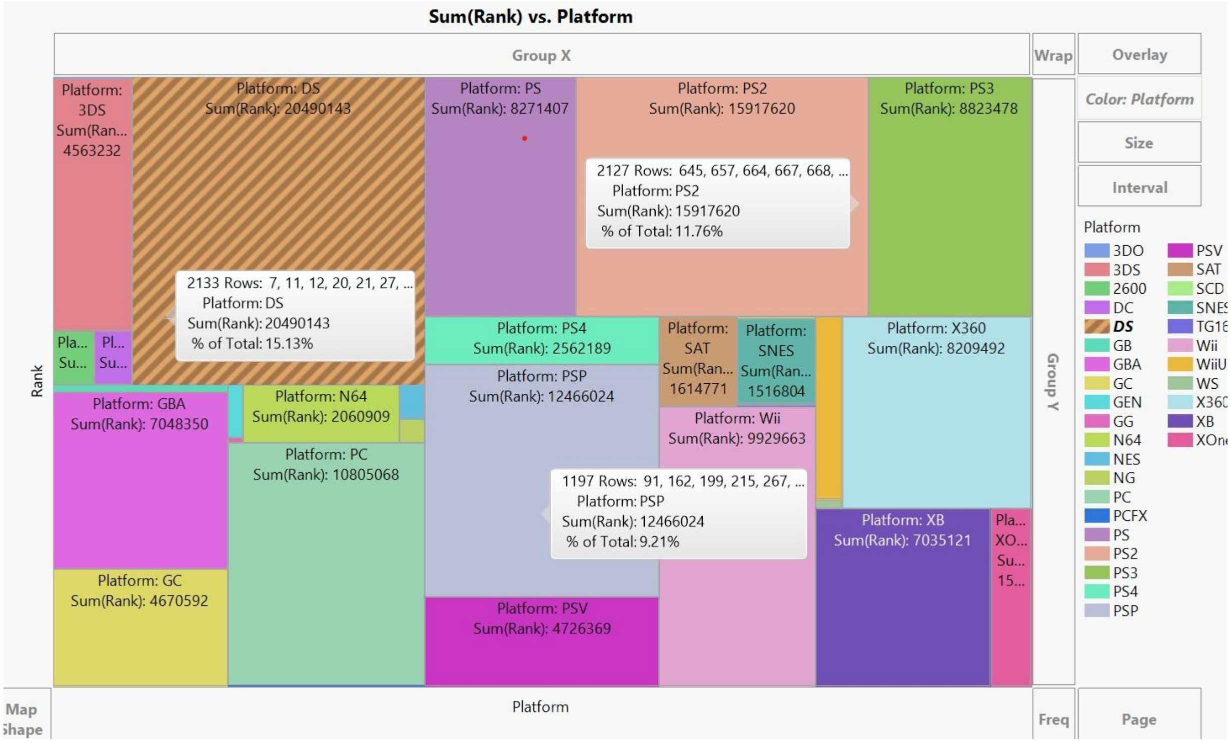
Global_Sales vs. Genre: The Visualization below depicts that the Platform genre has the highest sale which is followed by Shooter & Role-playing genres respectively.



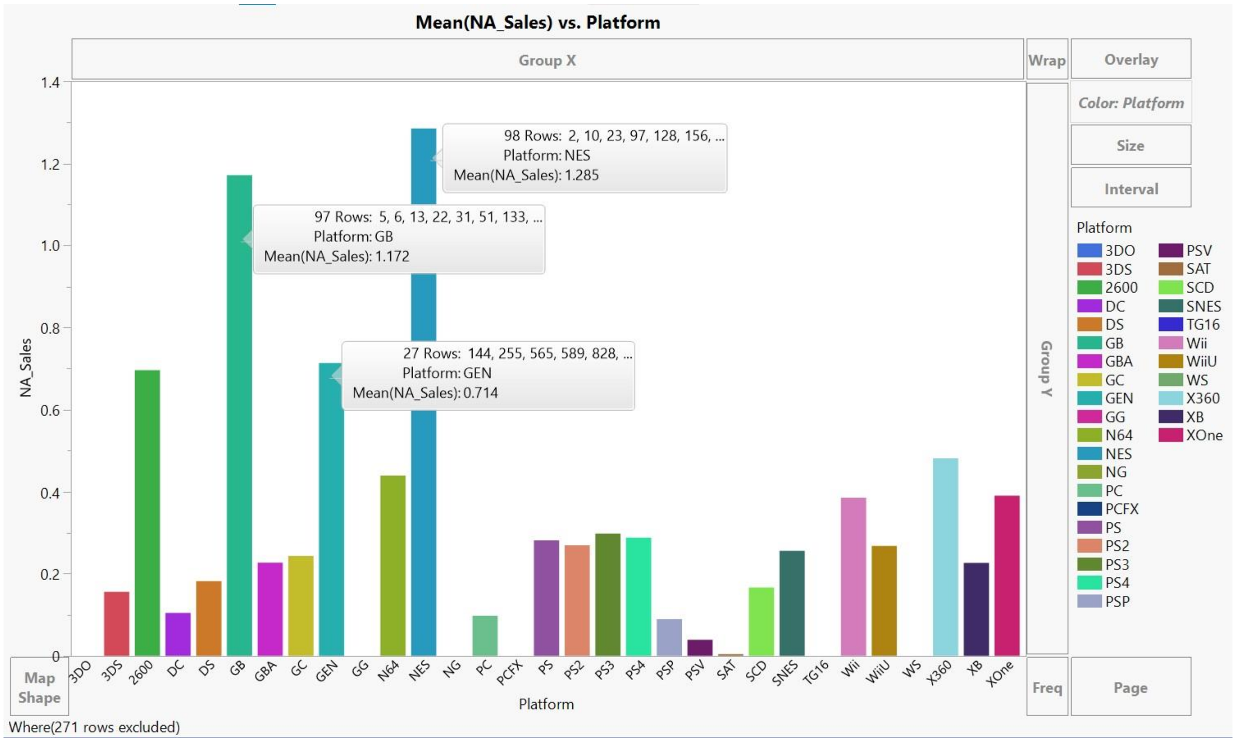
Global_Sales vs. Platform (gaming console): The picture below shows that Sony PS2 has the highest sales. And in the second position is Microsoft X360. And then it is followed by PS3 in third and Wii in the fourth position.



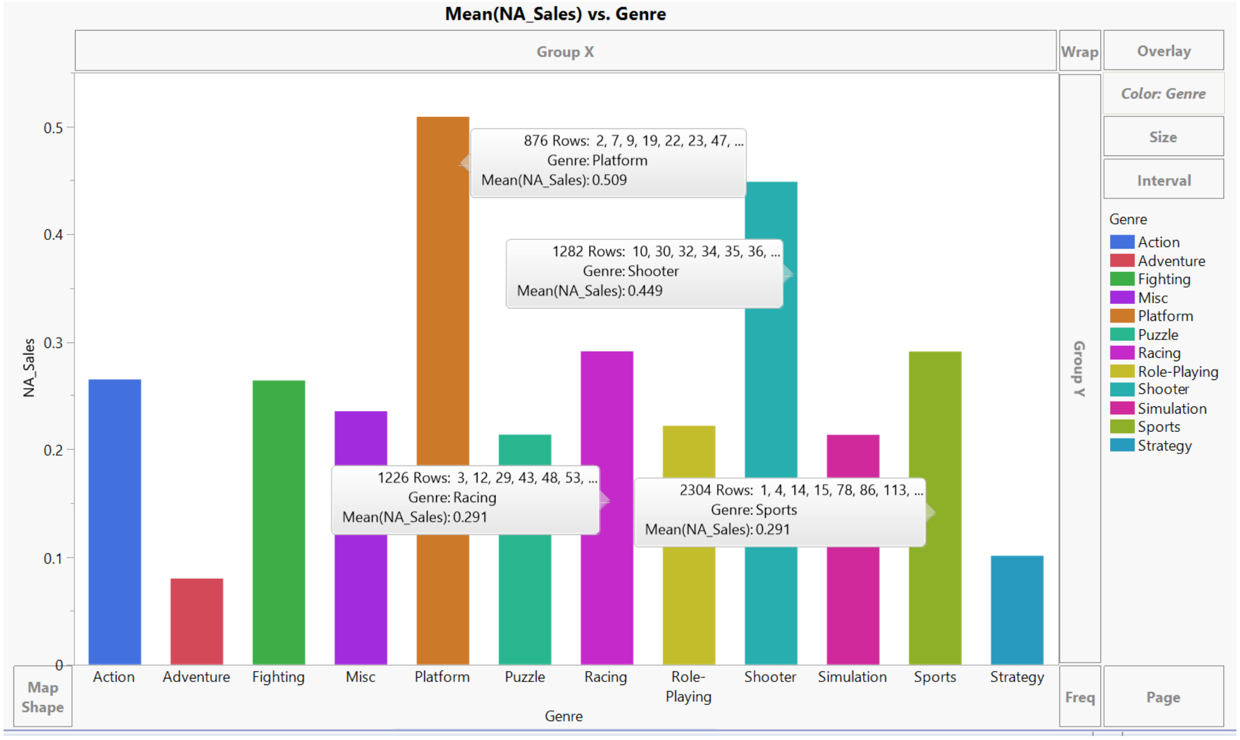
Rank vs. Platform: As per rankings the platform DS secured the highest rank and then PS2 & PSP are in the second and third position respectively.



NA_Sales vs. Platform: The bar plots below indicate which gaming platform has the highest sales in North America. As per the plot, the NES platform has the highest sales. GB gaming platform is in the second position which is then followed by the GEN platform.



NA_Sales vs Genre: The below box plots show the highest-selling gaming genre in North America. The platform genre (2D games) has the highest sales in North America. And shooter gaming genre is in the second position. Racing and Sports game genres share the third spot.



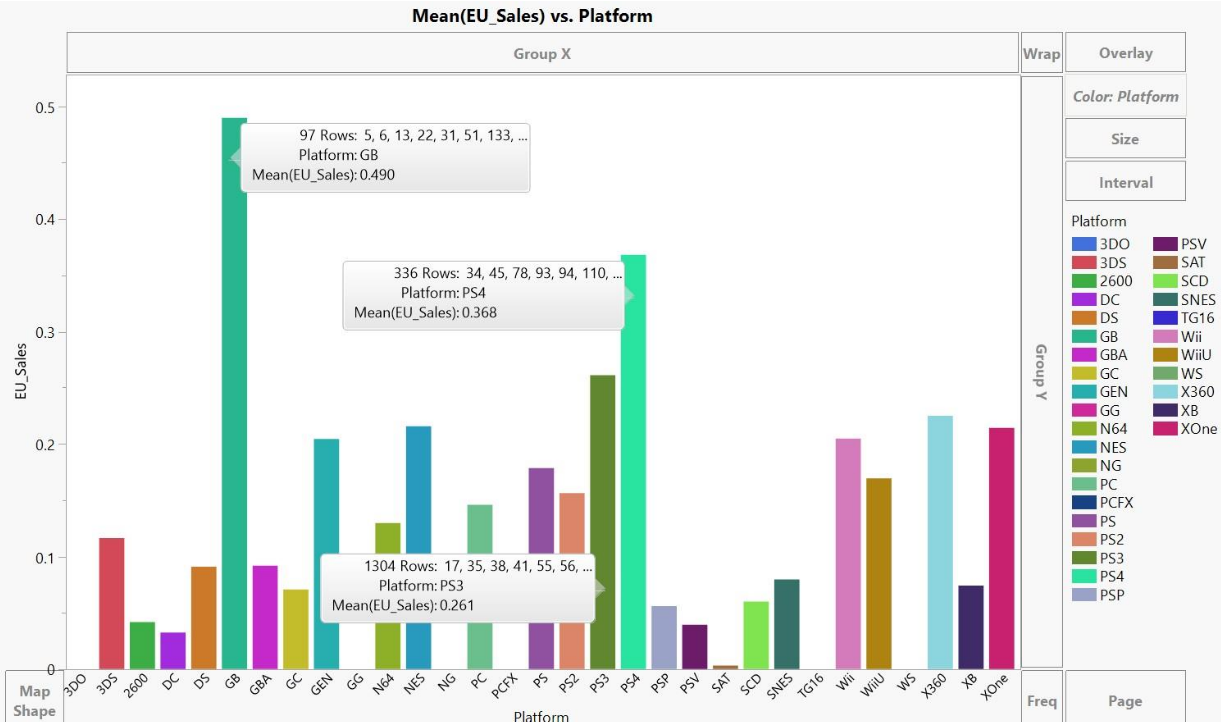
876 Rows: 2, 7, 9, 19, 22, 23, 47, ...
Genre: Platform
Mean(NA_Sales): 0.509

1282 Rows: 10, 30, 32, 34, 35, 36, ...
Genre: Shooter
Mean(NA_Sales): 0.449

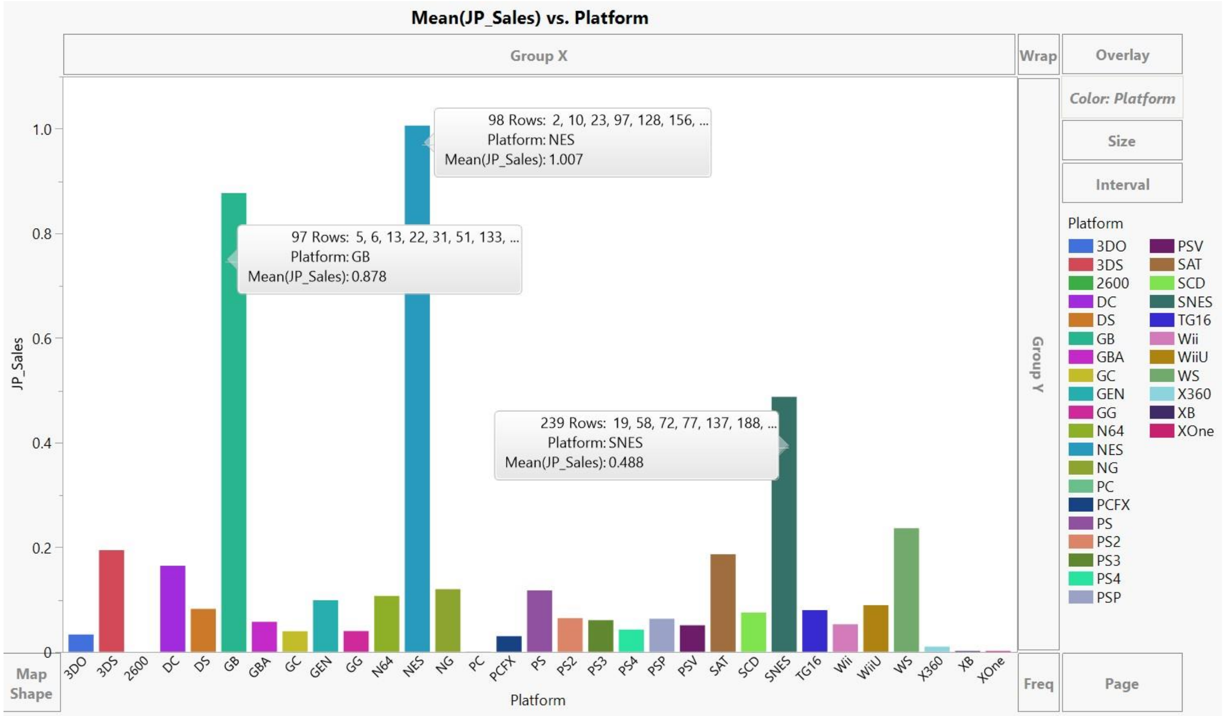
1226 Rows: 3, 12, 29, 43, 48, 53, ...
Genre: Racing
Mean(NA_Sales): 0.291

2304 Rows: 1, 4, 14, 15, 78, 86, 113, ...
Genre: Sports
Mean(NA_Sales): 0.291

EU_Sales vs Platform: The bar plot below indicates which gaming platform has the highest sales in Europe. As per the plot, the GB platform has the highest sales. The PS4 gaming platform is in the second position which is then followed by the PS3 platform.



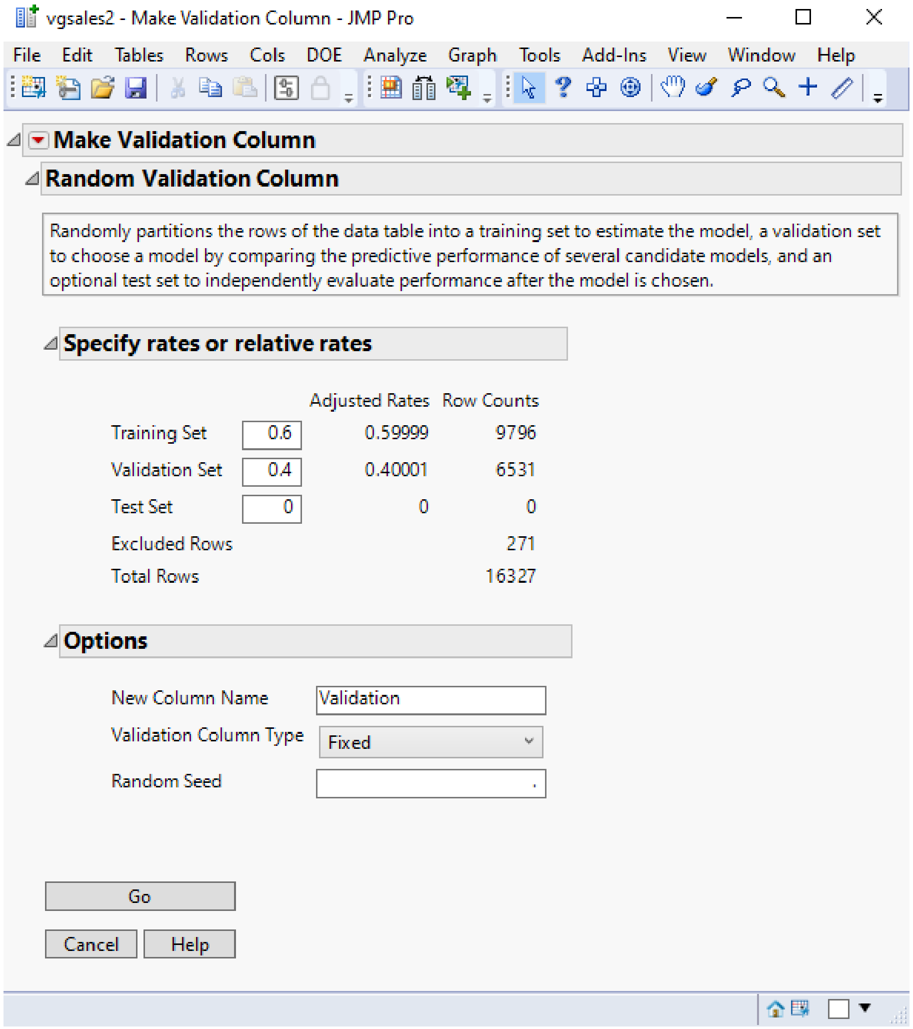
JP_Sales vs Platform: The bar plots below indicate which gaming platform has the highest sales in Japan. As per the plot, the NES platform has the highest sales. GB gaming platform is in the second position which is then followed by the SNES platform.



MAKING VALIDATION COLUMN:

We are considering the 60/40 split between the Training & Validation set to build the models.

This makes a validation column in the dataset of training with 60% and validation with 40%.



Make Validation Column

Random Validation Column

Randomly partitions the rows of the data table into a training set to estimate the model, a validation set to choose a model by comparing the predictive performance of several candidate models, and an optional test set to independently evaluate performance after the model is chosen.

Specify rates or relative rates

		Adjusted Rates	Row Counts
Training Set	0.6	0.59999	9796
Validation Set	0.4	0.40001	6531
Test Set	0	0	0
Excluded Rows			271
Total Rows			16327

Options

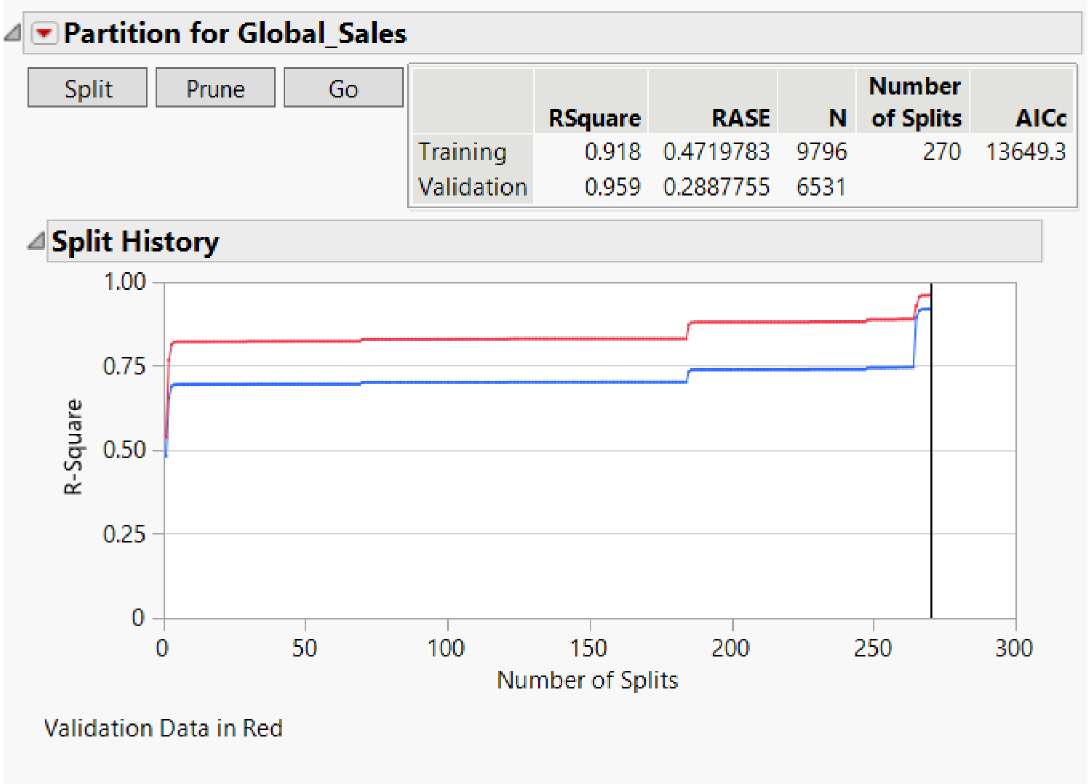
New Column Name: Validation

Validation Column Type: Fixed

Random Seed: .

MODELING

REGRESSION TREE:



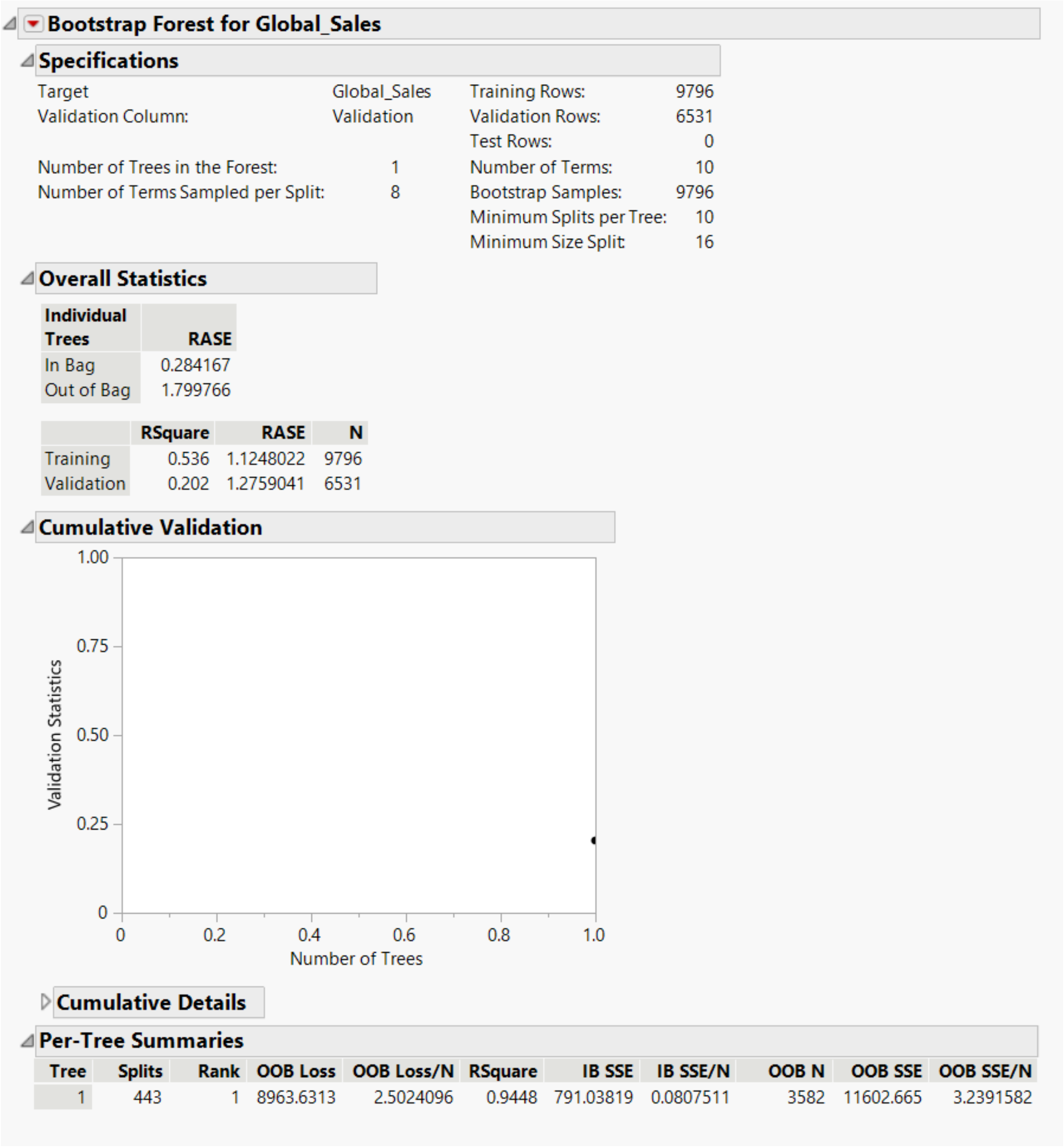
Partition for Global_Sales

	RSquare	RASE	N	Number of Splits	AICc
Training	0.918	0.4719783	9796	270	13649.3
Validation	0.959	0.2887755	6531		

Split History

Validation Data in Red

BOOTSTRAP FOREST:



Bootstrap Forest for Global_Sales

Specifications

Target	Global_Sales	Training Rows:	9796
Validation Column:	Validation	Validation Rows:	6531
		Test Rows:	0
Number of Trees in the Forest:	1	Number of Terms:	10
Number of Terms Sampled per Split:	8	Bootstrap Samples:	9796
		Minimum Splits per Tree:	10
		Minimum Size Split:	16

Overall Statistics

Individual Trees	RASE
In Bag	0.284167
Out of Bag	1.799766

	RSquare	RASE	N
Training	0.536	1.1248022	9796
Validation	0.202	1.2759041	6531

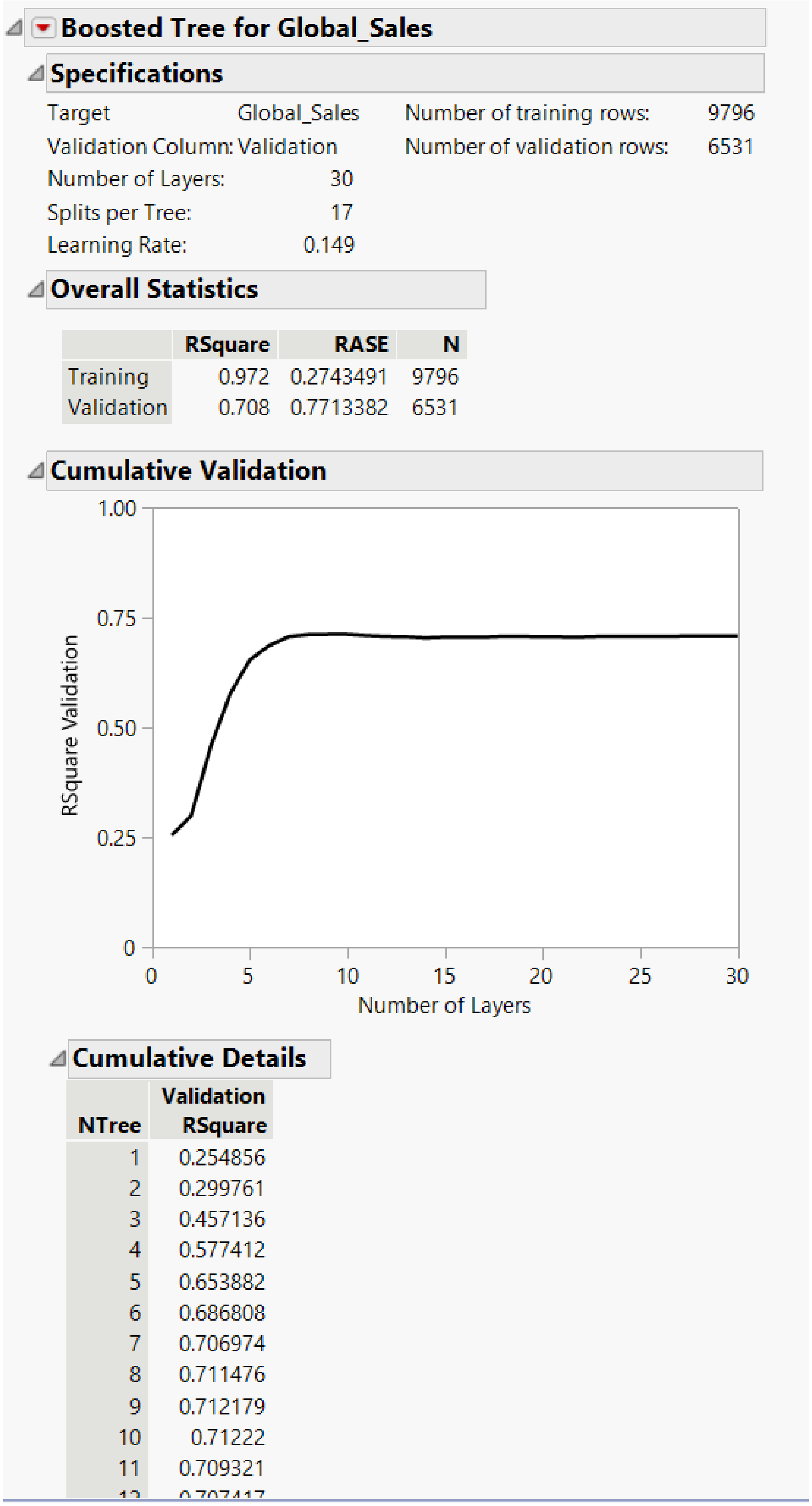
Cumulative Validation

Cumulative Details

Per-Tree Summaries

Tree	Splits	Rank	OOB Loss	OOB Loss/N	RSquare	IB SSE	IB SSE/N	OOB N	OOB SSE	OOB SSE/N
1	443	1	8963.6313	2.5024096	0.9448	791.03819	0.0807511	3582	11602.665	3.2391582

BOOSTED TREE:



Boosted Tree for Global_Sales

Specifications

Target	Global_Sales	Number of training rows:	9796
Validation Column:	Validation	Number of validation rows:	6531
Number of Layers:	30		
Splits per Tree:	17		
Learning Rate:	0.149		

Overall Statistics

	RSquare	RASE	N
Training	0.972	0.2743491	9796
Validation	0.708	0.7713382	6531

Cumulative Validation

Cumulative Details

NTree	Validation RSquare
1	0.254856
2	0.299761
3	0.457136
4	0.577412
5	0.653882
6	0.686808
7	0.706974
8	0.711476
9	0.712179
10	0.71222
11	0.709321
12	0.707417

MODEL SCREENING:

By making the validation column, we proceed to compare the five different sets of models. Among the five, by comparing them we will find the best model which will provide enough information to build a successful model. The five models we are selecting for our dataset are Decision Tree, Bootstrap Forest, Boosted Tree, Neural networks, and Logistic regression.

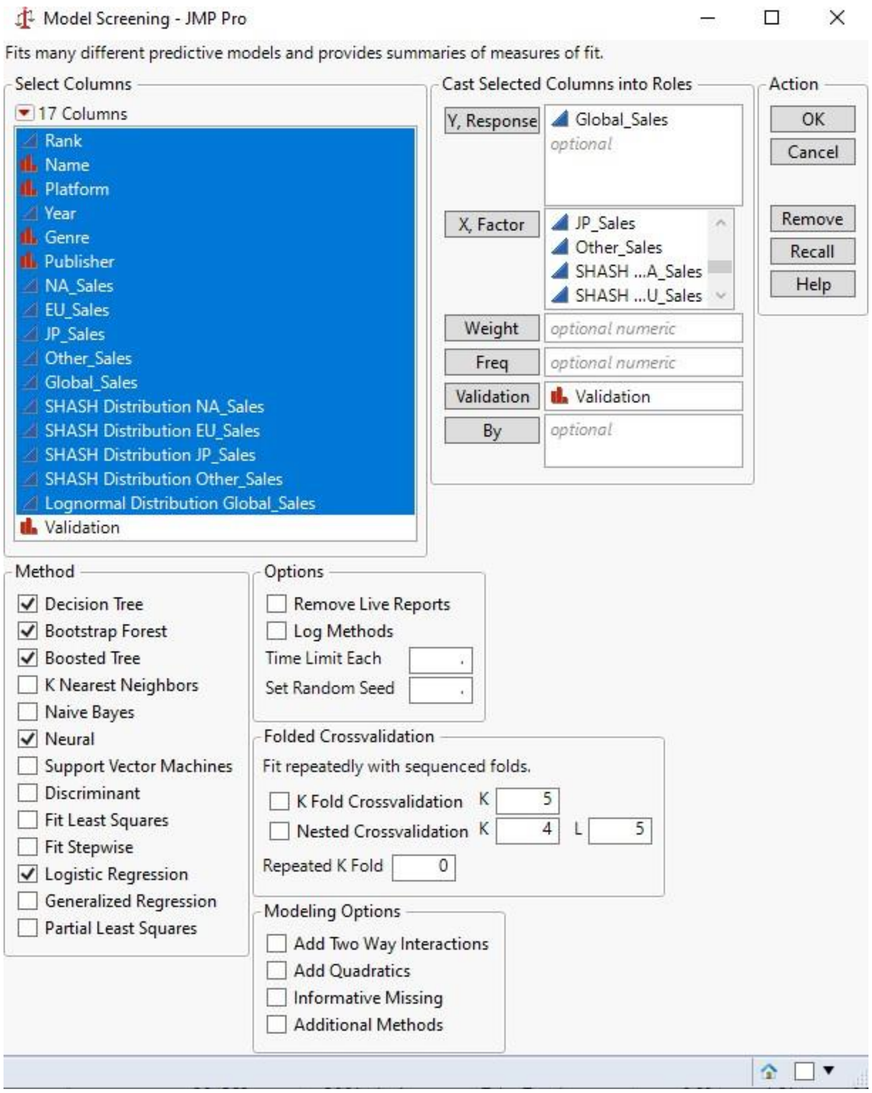
For the Decision Tree, JMP will create a model by partitioning the data into different subsets based on the values of the predictor variables. It will then use a cost function or error rate to evaluate the different splits and determine the best split point. The model will then be pruned to reduce the complexity of the tree and improve its accuracy.

For Bootstrap Forest, JMP will create a forest of decision trees, each one based on a different sample of the data. It will then use an optimization algorithm to find the best combination of trees that minimizes the prediction error.

For Boosted Tree, JMP will build a sequence of models where each model is based on the previous model's predictions. It will then combine the predictions from all the models and use an optimization algorithm to find the best combination of predictors that minimizes the prediction error.

For Neural networks, JMP will create a model based on a set of neurons connected in a network. It will use an optimization algorithm to find the best combination of weights and biases that minimizes the prediction error.

Finally, for Logistic Regression, JMP will create a model based on a set of predictor variables and their associated coefficients. It will use an optimization algorithm



Model Screening - JMP Pro

Fits many different predictive models and provides summaries of measures of fit.

Select Columns

17 Columns

- Rank
- Name
- Platform
- Year
- Genre
- Publisher
- NA_Sales
- EU_Sales
- JP_Sales
- Other_Sales
- Global_Sales
- SHASH Distribution NA_Sales
- SHASH Distribution EU_Sales
- SHASH Distribution JP_Sales
- SHASH Distribution Other_Sales
- Lognormal Distribution Global_Sales
- Validation

Cast Selected Columns into Roles

Y, Response	Global_Sales optional
X, Factor	JP_Sales Other_Sales SHASH ...A_Sales SHASH ...U_Sales
Weight	optional numeric
Freq	optional numeric
Validation	Validation
By	optional

Action

Method

- ☑ Decision Tree
- ☑ Bootstrap Forest
- ☑ Boosted Tree
- □ K Nearest Neighbors
- □ Naive Bayes
- ☑ Neural
- □ Support Vector Machines
- □ Discriminant
- □ Fit Least Squares
- □ Fit Stepwise
- ☑ Logistic Regression
- □ Generalized Regression
- □ Partial Least Squares

Options

- □ Remove Live Reports
- □ Log Methods
- Time Limit Each .
- Set Random Seed .

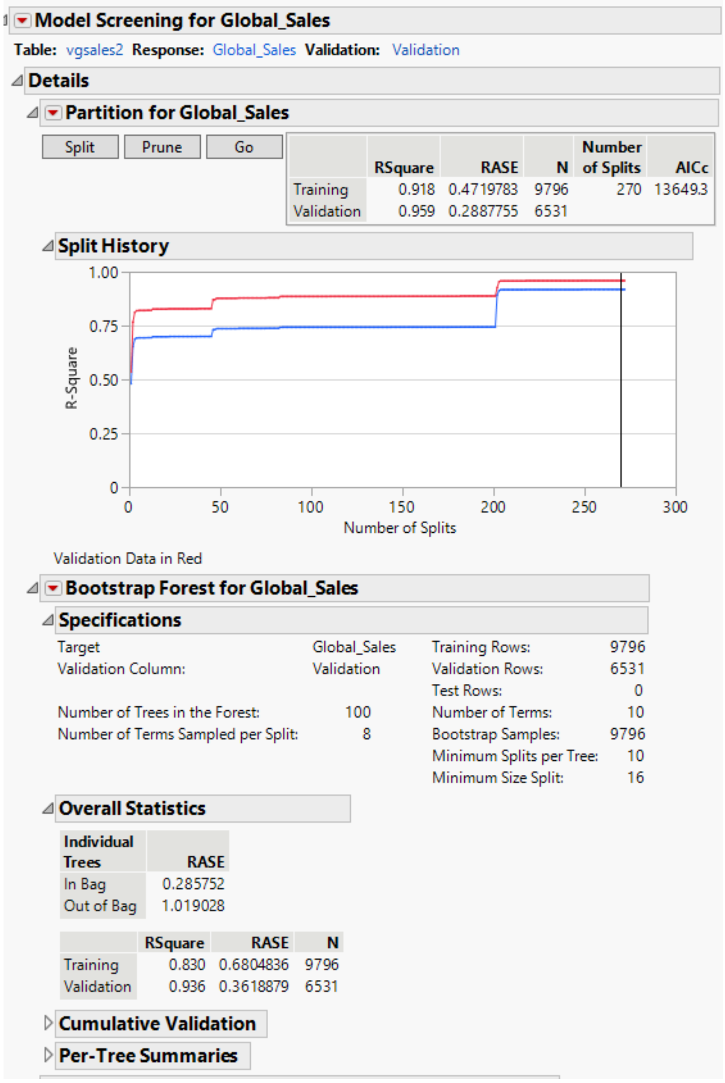
Folded Crossvalidation

Fit repeatedly with sequenced folds.

- □ K Fold Crossvalidation K 5
- □ Nested Crossvalidation K 4 L 5
- Repeated K Fold 0

Modeling Options

- □ Add Two Way Interactions
- □ Add Quadratics
- □ Informative Missing
- □ Additional Methods



Model Screening for Global_Sales

Table: vgsales2 **Response:** Global_Sales **Validation:** Validation

Details

Partition for Global_Sales

	RSquare	RASE	N	Number of Splits	AICc
Training	0.918	0.4719783	9796	270	13649.3
Validation	0.959	0.2887755	6531		

Split History

Validation Data in Red

Bootstrap Forest for Global_Sales

Specifications

Target	Global_Sales	Training Rows:	9796
Validation Column:	Validation	Validation Rows:	6531
		Test Rows:	0
Number of Trees in the Forest:	100	Number of Terms:	10
Number of Terms Sampled per Split:	8	Bootstrap Samples:	9796
		Minimum Splits per Tree:	10
		Minimum Size Split:	16

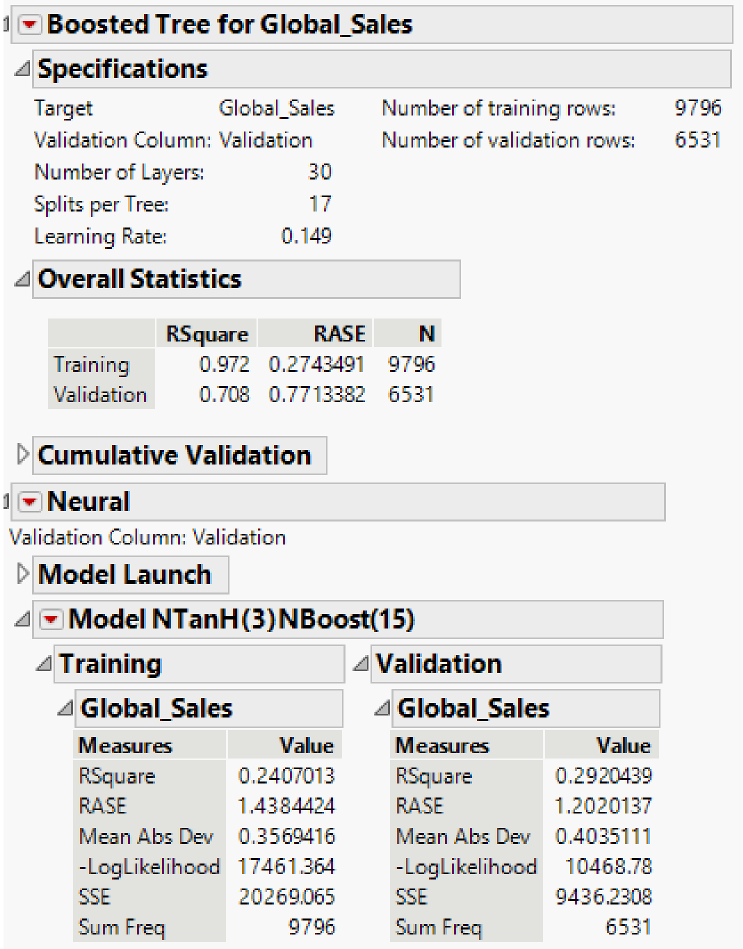
Overall Statistics

Individual Trees	RASE
In Bag	0.285752
Out of Bag	1.019028

	RSquare	RASE	N
Training	0.830	0.6804836	9796
Validation	0.936	0.3618879	6531

Cumulative Validation

Per-Tree Summaries



Boosted Tree for Global_Sales

Specifications

Target	Global_Sales	Number of training rows:	9796
Validation Column:	Validation	Number of validation rows:	6531
Number of Layers:	30		
Splits per Tree:	17		
Learning Rate:	0.149		

Overall Statistics

	RSquare	RASE	N
Training	0.972	0.2743491	9796
Validation	0.708	0.7713382	6531

Cumulative Validation

Neural

Validation Column: Validation

Model Launch

Model NTanH(3)NBoost(15)

Training

Global_Sales

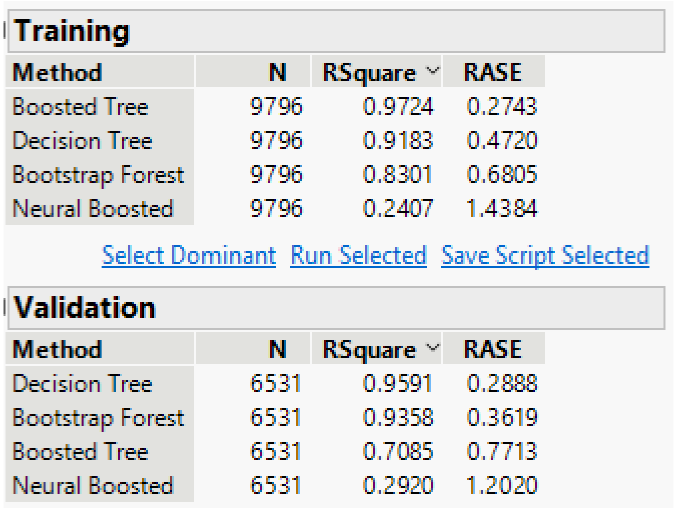
Measures	Value
RSquare	0.2407013
RASE	1.4384424
Mean Abs Dev	0.3569416
-LogLikelihood	17461.364
SSE	20269.065
Sum Freq	9796

Validation

Global_Sales

Measures	Value
RSquare	0.2920439
RASE	1.2020137
Mean Abs Dev	0.4035111
-LogLikelihood	10468.78
SSE	9436.2308
Sum Freq	6531

As we can see, the Decision Tree model is the most accurate when predicting the success of a video game. The model can take in many data points, as well as a variety of different data types, to accurately predict the success of a video game. This model can be used by game developers to make better decisions when developing a game, as well as by gamers to make better decisions when deciding which game to purchase.



Training

Method	N	RSquare	RASE
Boosted Tree	9796	0.9724	0.2743
Decision Tree	9796	0.9183	0.4720
Bootstrap Forest	9796	0.8301	0.6805
Neural Boosted	9796	0.2407	1.4384

Validation

Method	N	RSquare	RASE
Decision Tree	6531	0.9591	0.2888
Bootstrap Forest	6531	0.9358	0.3619
Boosted Tree	6531	0.7085	0.7713
Neural Boosted	6531	0.2920	1.2020

CONCLUSION:

Overall, the Decision Tree model was the best predictor of game success, with a correlation of 0.9591 and a RASE of 0.2888. This indicates that the model was able to accurately predict the success of games with a high degree of accuracy.