Global Video Game Sales Prediction

Prepared by:

Md Amanat Khan Shishir (khanamanat675@gmail.com)

Completion Date: December 9, 2024

Table of Contents

Abstract

The video game industry is one of the biggest sectors of the global entertainment sector, makes billions in revenue every year. With increasing competition both the publishers and the developers are facing many challenges in regards to understanding the preferences of the consumer and increase the sales. So this project analyses a dataset of global video game sales and uses statistical techniques and Python based data analysis to find out patterns and trends in terms of platforms, genres and regions.

The analysis answers important questions like

- How preferences of a regional effect global sales trends?
- What kind of genres and platforms have the most share in the markets?
- How does game release years correlate with sales performance?

Through accurate data pre-processing that has handling missing values, outlier detection, and feature engineering. the dataset was optimized to get maximum insights. The findings showed regional variations with North America showing a big preference for Action and Shooter games But Japan had a great demand for Role-Playing Games . Platforms like PlayStation and Xbox had the most sales showing their international appeal. The analysis shows how timing and technological advancements affected the success of a game.

These insights give guidance for stakeholders in strategic decisions in fields of marketing, development and distribution. By adding robust data analysis with domain specific knowledge this project gives a detailed explanation of the video game market. And in the end two models were used i.e. Random forest regressor and XG boost resulting in r^2 of 0.94 and 0.90 respectively.

Description of Applied Problem

Key objective: predicting the global sales of a video game based on historical data.
Challenges in the Video Game Market and in building the final prediction model.
The video game industry works in a very dynamic and competitive environment. Sales performance is affected by multiple factors like

- **Regional Preferences**: Different regions show different preferences for game genres platforms and pricing models.
- **Technological Advancements**: New console generations and gaming technologies can greatly shift user's behaviour.
- **Market Saturation**: Overproduction of similar games in some genres leads to consumer getting bored and inturn decreasing the sales as well.
  For example a game released on a popular platform like PlayStation would perform well globally but still face different success across different regions. Similarly. The timing of a game's release in relation to a console's lifecycle can also greatly impact its sales.
- **Missing data**: certain data like Critic_Score, Critic_Count, User_Score, and User_Count. Had too many missing values which would create a problem in final model building

Importance of Sales Trend Analysis and its uses for corporates.

Accurate sales trend analysis enables stakeholders to:

- **Optimize Product Development**: Focus resources on genres and features that can fulfil the demands that the customers have along with that we can easily judge based on the inputs as to what kind of sales the game will have if we already know the trend.
- **Enhance Marketing Strategies**: Apart from building an accurate model with the help of these insights the companies can target specific kind of customers in said specific region this could in turn increase the sales.
- **Forecast Sales**: Companies can guess market reaction to a new game launches and allocate budgets that way.
- **Accurate Sales prediction**: In the end when we predict the sales of a game based on certain inputs like genre, publication, etc. we need a clean and reliable data set on which the model can train on. So that it performs well at the time of prediction.

Description of Available Data

The dataset used for this study is a multivariate dataset having 16,719 instances with 16 features which can help us find insights into video game sales across different regions [1]. Information on the game attributes such as platform, genre and year of release is also included in this. Also the performance in the key markets of North America, Europe and Japan.. This multivariate dataset is designed to help us analyze patterns in video game sales and predict global sales based on historical data.

A main challenge in working with the dataset that we had is addressing missing values mainly in columns like 'Critic_Score', 'User_Score' and 'Developer' which would affect the consistency of the analysis. Even in spite of these challenges most of the sales data columns are complete making sure there is a solid foundation for machine learning models. The dataset has the following features:

- Name: The title of the video game.
- Platform: The gaming platform like PS2, X-box, etc.
- Year_of_Release: The year the game was released
- Genre: The genre the game falls into like Action, Adventure, Sports etc.
- Publisher: Publishing company
- NA_Sales, EU_Sales, JP_Sales, Other_Sales: Sales in millions of units in North America, Europe, Japan and all other regions of the world combined.
- Global_Sales: The total worldwide sales of the game that is the sum of the regional sales. This is the target varible.
- Critic_Score, Critic_Count: overall critic score and the number of the critics
- User_Score, User_Count: overall user score and the number of the users
- Developer: The development company
- Rating: This is the ESRB rating for the game ex. E, M, T.

This is a complete dataset and it will provide a rich basis for the analysis of sales trends and building predictive models, with some attention to handling missing values and making sure that the reliability of the feature set. Apart from that precise and accurate feature engineering is required in order to make a robust model with the use of this dataset.

	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
0	Wii Sports	Wii	2006.0	Sports	Nintendo	41.36	28.96	3.77	8.45	82.53	76.0	51.0	8	322.0	Nintendo	E
1	Super Mario Bros.	NES	1985.0	Platform	Nintendo	29.08	3.58	6.81	0.77	40.24	NaN	NaN	NaN	NaN	NaN	NaN
2	Mario Kart Wii	Wii	2008.0	Racing	Nintendo	15.68	12.76	3.79	3.29	35.52	82.0	73.0	8.3	709.0	Nintendo	E
3	Wii Sports Resort	Wii	2009.0	Sports	Nintendo	15.61	10.93	3.28	2.95	32.77	80.0	73.0	8	192.0	Nintendo	E
4	Pokemon Red/Pokemon Blue	GB	1996.0	Role-Playing	Nintendo	11.27	8.89	10.22	1.00	31.37	NaN	NaN	NaN	NaN	NaN	NaN

Fig 1: Data table of Video Games Sales Data

DATA PREPROCESSING

Data preprocessing makes sure that the reliability and precision of every analysis or predictive modeling is good [2]. Preprocessing of the Video Game Sales Dataset was done using a series of techniques to handle the problems at hand like missing values, inconsistencies in scale, and outliers. The following steps were followed:

1. **Handling Missing Data**

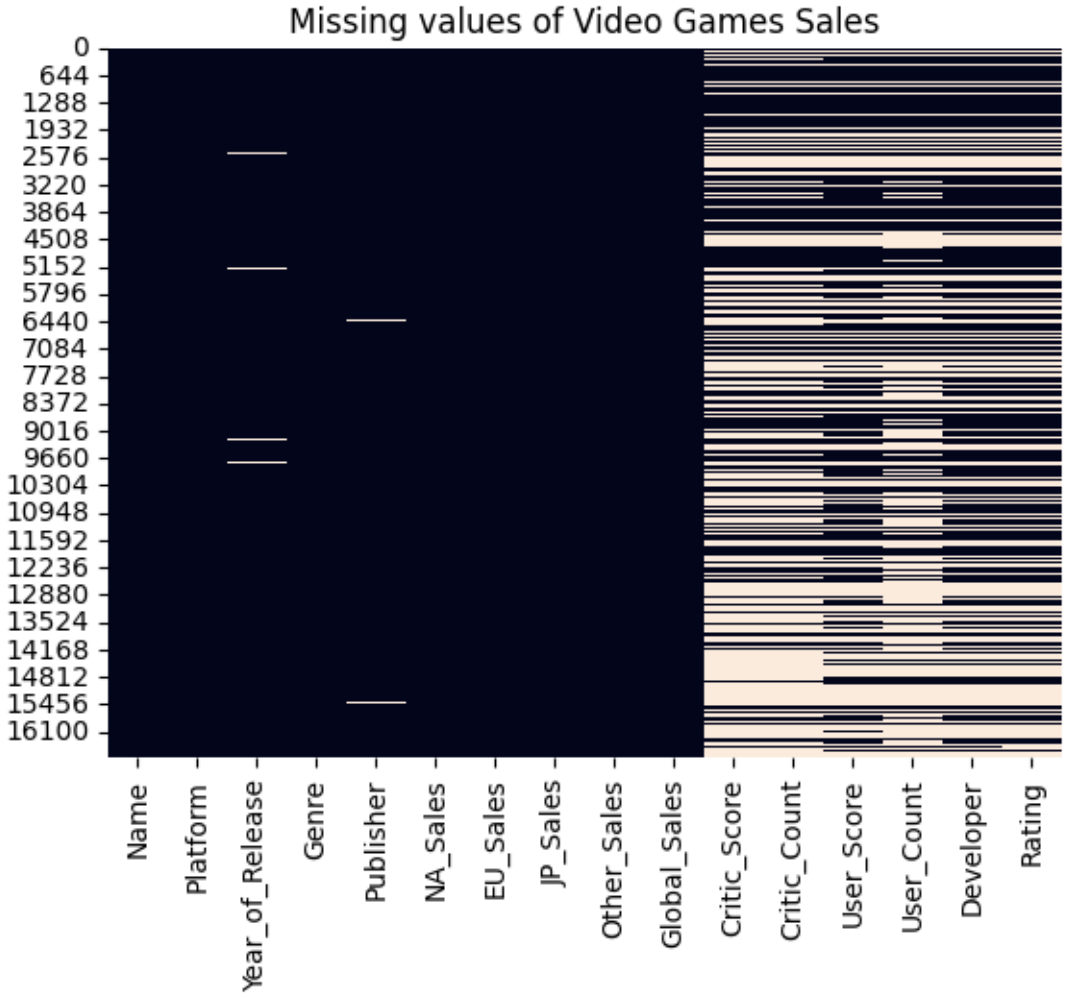


Fig 2: Missing value analysis for video game data

The black spaces shows complete data and the empty white spaces show missing values present in out dataset.

The dataset has missing values in some columns, mainly in features related to critic scores, user scores, and developer information. Addressing these missing values was very important to maintain data consistency and avoid biases in the analysis.
Imputation:

- Numerical Features: Missing values in continuous numerical columns like Critic_Score, User_Score were replaced using the mean of the available data in those columns. This approach makes sure that the imputed values are representative of the overall dataset without skewing the distribution. But, this approach was later dropped

as it created a skewed data set with majority of the data being synthetic. So we would not be able to make a robust data set with this model.

- Categorical Features: Missing values within categorical columns, like the genre or platform column were replaced with the most occurring value within that column known as the mode. This method maintains logical consistency for categorical variables.

2. Data set after categorical imputations with mode:-

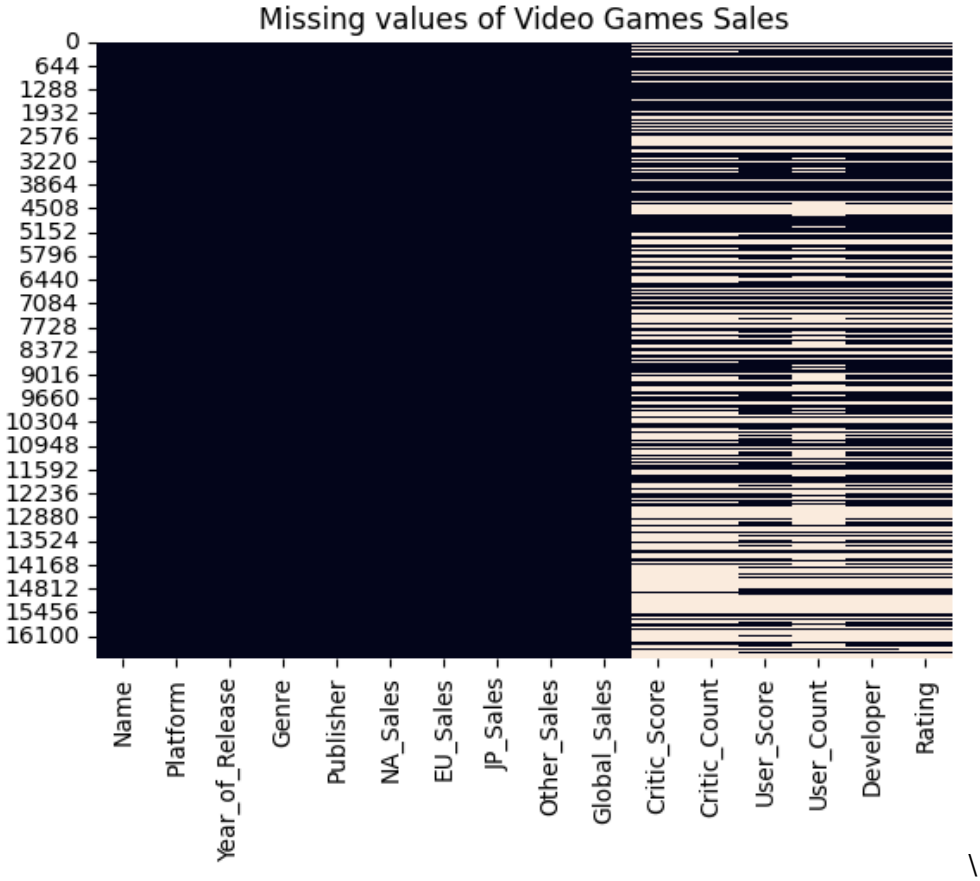


Fig 3: Missing value analysis after data imputation.

3. Data Handling with high missing values

```
Missing Values with column name:
Name                  0
Platform              0
Year_of_Release       0
Genre                 0
Publisher             0
NA_Sales              0
EU_Sales              0
JP_Sales              0
Other_Sales           0
Global_Sales          0
Critic_Score       8582
Critic_Count       8582
User_Score         6704
User_Count         9129
Developer          6623
Rating             6769
dtype: int64
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 16719 entries, 0 to 16718
Data columns (total 16 columns):
```

Fig 4: Missing value summary by column.

Columns with an really high proportion of missing values were dropped from the analysis. For example features like Critic_Count and User_Count had too many gaps more than 8000 missing values making reliable imputation challenging. Excluding these columns helped avoid introducing inaccuracies or biases that could compromise the model's performance.

```
#Dropping Columns which had more than 60% of data missing
videogames_df_cleaned = videogames_df.drop(["Critic_Score","Critic_Count","User_Count","Developer","User_Score","Rating"], axis=1)
```

Fig 5: Columns dropped due to high missing values proportion.

This systematic approach to handling missing data ensured the dataset remained robust and suitable for machine learning applications.

4. Standardization

```
#Here we have scaled the sales features fir different regions
standard_features = ['NA_Sales' ,'EU_Sales' ,'JP_Sales' ,'Other_Sales'  ,'Global_Sales']

videogames_df_cleaned[standard_features] = StandardScaler().fit_transform(videogames_df_cleaned[standard_features])

videogames_df_cleaned
```

Fig 6: Standardizations of sales features by region.

Sales data across different regions (NA_Sales, EU_Sales, JP_Sales, Other_Sales) showed different scales due to different regional market differences. To make sure that these features contributed equally to the analysis and did not disproportionately affect the machine learning models standardization was applied.

Standardization Process:

- Each sales column was transformed to have a mean of 0 and a standard deviation of 1.
- This process made the data comparable across regions by taking away the effects of differing scales and units.

	mean	std	min	25%	50%	\
Name	NaN	NaN	NaN	NaN	NaN	
Platform	NaN	NaN	NaN	NaN	NaN	
Year_of_Release	1974.18793	252.656404	-1.0	2003.0	2007.0	
Genre	NaN	NaN	NaN	NaN	NaN	
Publisher	NaN	NaN	NaN	NaN	NaN	
NA_Sales	-0.0	1.00003	-0.323705	-0.323705	-0.225363	
EU_Sales	-0.0	1.00003	-0.288166	-0.288166	-0.248426	
JP_Sales	0.0	1.00003	-0.251295	-0.251295	-0.251295	
Other_Sales	-0.0	1.00003	-0.253512	-0.253512	-0.199951	
Global_Sales	-0.0	1.00003	-0.33823	-0.305928	-0.234864	

Fig 7: Descriptive statistics after standardization.

Impact of Standardization:

- This increased the performance of machine learning algorithms, mainly the models that are sensitive to scale differences (e.g., gradient-based methods).
- Standardized data made it possible for more accurate identification of trends and patterns in sales across regions.

EXPLORATORY DATA ANALYSIS

EDA is a very important step in understanding the overall structure of the dataset and identifying the important trends, patterns and relationships in the features [3]. The insights gathered during this stage guided feature selection and model development. An analysis of the findings from the EDA conducted on the Video Game Sales Dataset:

Pairplot Analysis: Exploring Relationships Between Sales Metrics:

Firstly, we plot a pair plot for all the variables and gathered as much insight as possible. The graph is as follows:

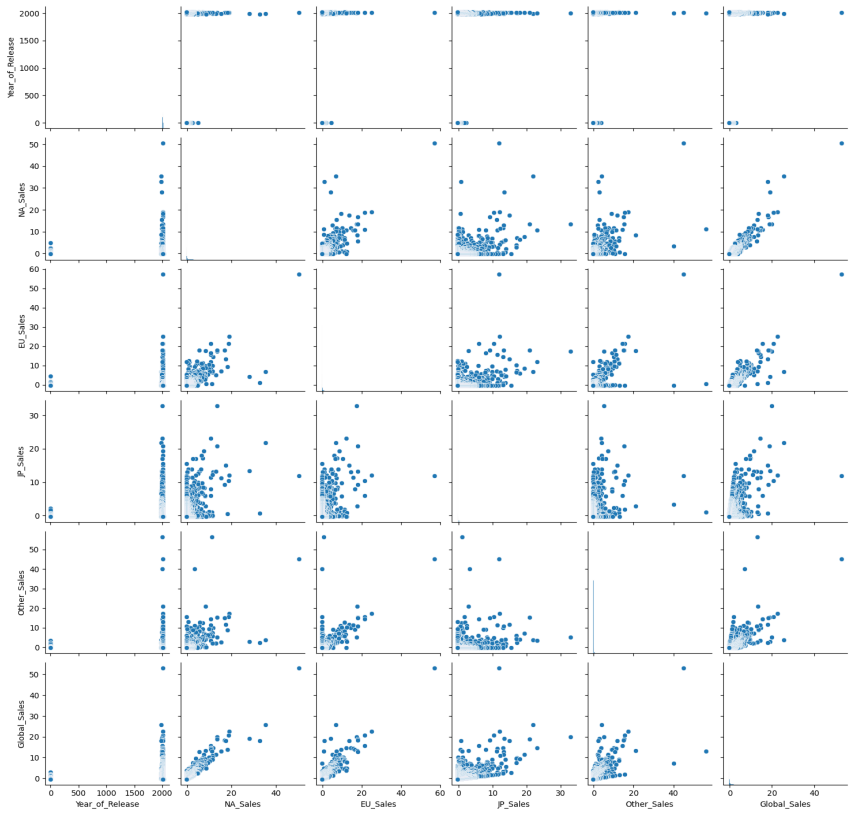


Fig 8: Pair plot of key numerical variables of video games dataset.

The insights revealed by the scatter plot is as follows:-
Scatterplot Matrix: The scatterplot matrix shows relationships in these following variables Year of Release, NA_Sales, EU_Sales, JP_Sales, Other_Sales, and Global_Sales.

1. **Sales Correlation**:
   We observed a strong positive correlations is observed between Global Sales and regional sales (NA, EU, JP, Other). This shows that the successful games typically perform well across multiple regions.

Example: Games with high NA_Sales often have corresponding high Global Sales values. Showing the positive linear relationship.
But, if we take all the regional sales into consideration, we will have an overfitting model as the model would be able to predict the exact total based on the regional sales.

2. **Regional Contribution**:
   North America and Europe contribute the most to Global Sales in comparison to Japan and other regions.
   Example: Multiple points gather as a cluster at higher values for NA_Sales and EU_Sales when Global Sales increases.
3. **Year of Release**:
   Games released in earlier years like before 2000 show less sales as compared to recent years. This shows industry's growth over time or maybe incomplete data for previous periods is the most likely explanation.
4. **Outliers**:
   There are many outliers in Global Sales showing blockbuster games that performed the best as compared to the majority of titles. These games are generally bringing the industry's most overall revenue.
   But apart from that as well the data is too much filled with outliers so we would have to remove the outlier as shown in the pre processing step for removing outliers.

For further analysis individual parameters were analysed using visualisation

Global sales by platform

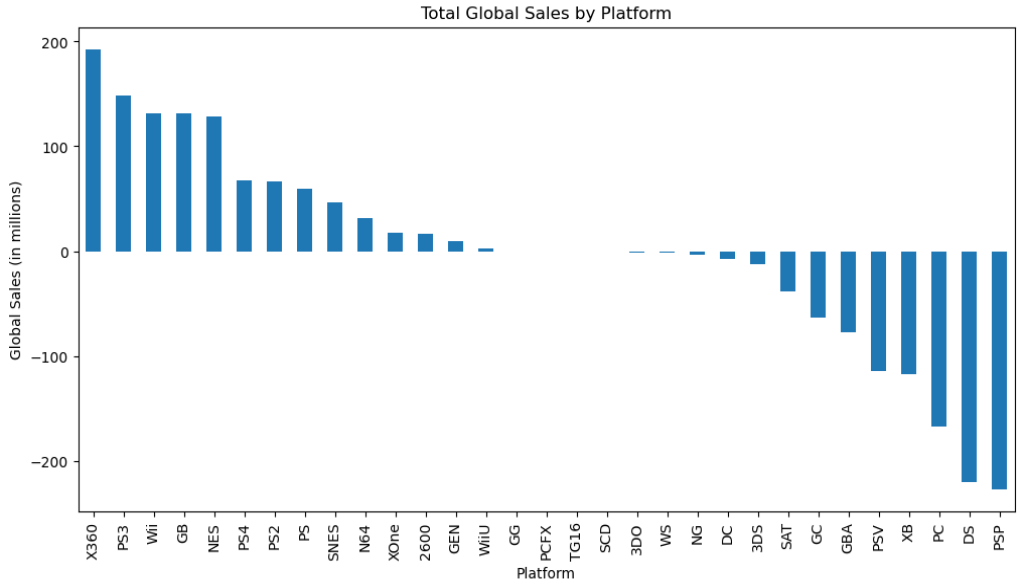


Fig 9: Total global sales by platform

The bar chart shows the total video games sell trends by the platform in millions around the globe. It is noticeable that, X360 generates highest sales. On the other hand, PSP generates lowest among all the platforms in this dataset.

The X360 is the highest sales generator called leader. Likewise, the PS3 and Wii generates sales but relatively lower than X360. GB, NES and PS4 have the strong competition in the market. More importantly, PSP generates lowest sales which is reflected on the scaled value.

The negative value indicating that, the scaled sales value falls below the standard average value of datasets. Even though, DC is the well-known platform the lack of blockbuster game has made this platform out pf competition.

Some niche platform like, PCFX, GG has the weakest contribution in the sales value likely due to lack of games or lack of focused on the regional sales. Overall, the data highlighted the sales trends of individual platform have some insightful information to drive global sales.

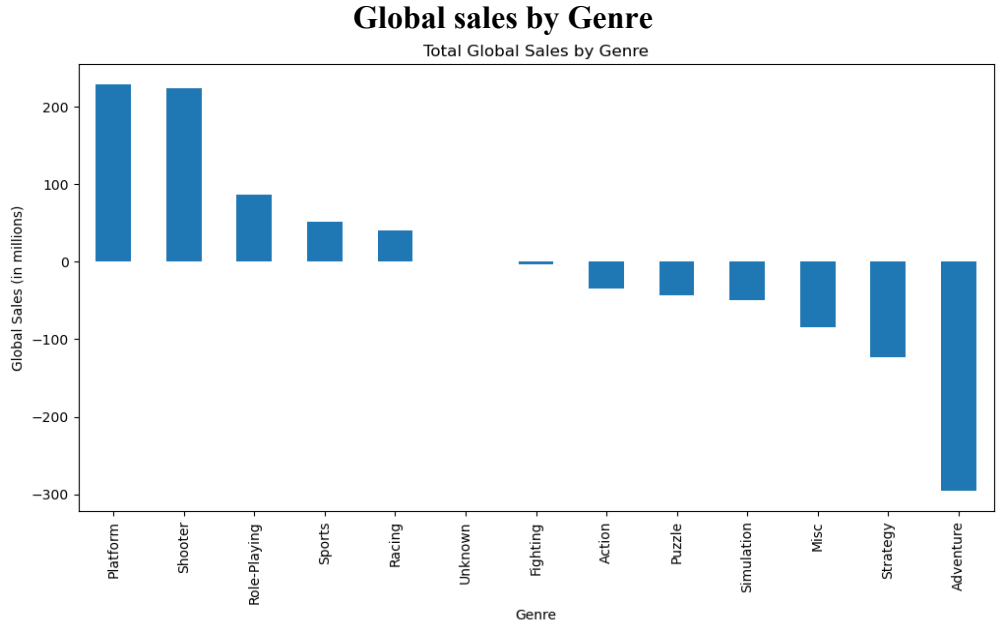


Fig 10: Total global sales by video games genre

The bar graph illustrates the global sales in millions by genres. Some genres generate highest scaled value and some are lowest.

Genres like Platform, shooter have the highest scaled values among all the genres. On the other hand, adventure and strategy have the lowest scaled value indicating that, those genres generate lowest sales among all the genre of the video game dataset. Role-playing, action and racing generates lower sales form the dataset.

Box plot of Global sales Based on Platform

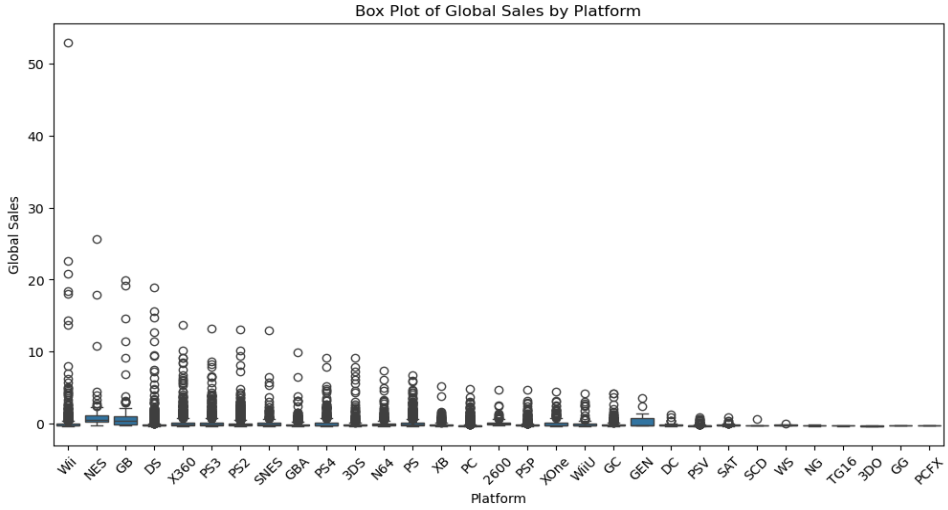


Fig 11: Box plot of global sales by platform.

The box plot of the global sales based on platform is giving some significant insights. Platform like Wii outperform all the platform all the platform with numerous outliers reaching more than 50 million dollars. Blockbuster title like Wii sports contributed significantly for the global sales. Some platform like NES, GB, DS have also generated more sales but relatively lower than the platform like Wii with significant lower outliers.

Moderately performing platform like X360, PS2, PS3 also generate consistent sales. These platform balance steady-performing games with occasional outlier, such as Halo on the XBOX360. On the lower side of the boxplot, platform like 3DO, GG and PCFX have the lowest median sales among the platform in this graph with no significant outliers. This means those platforms have lack of market shares and failed to produce blockbuster titles games.

A significant features of the dataset is the presence of numerous outliers, particularly for top-performing platforms like Wii, NES and DS. These outliers are the reason for the skewed the graph. Those outliers need to handles from getting rid from bad accuracy score from modeling. To handling outliers, Inter Quartile Range is the best technique to exclude extreme values from individual features that helps to prevent skewness of the plot.

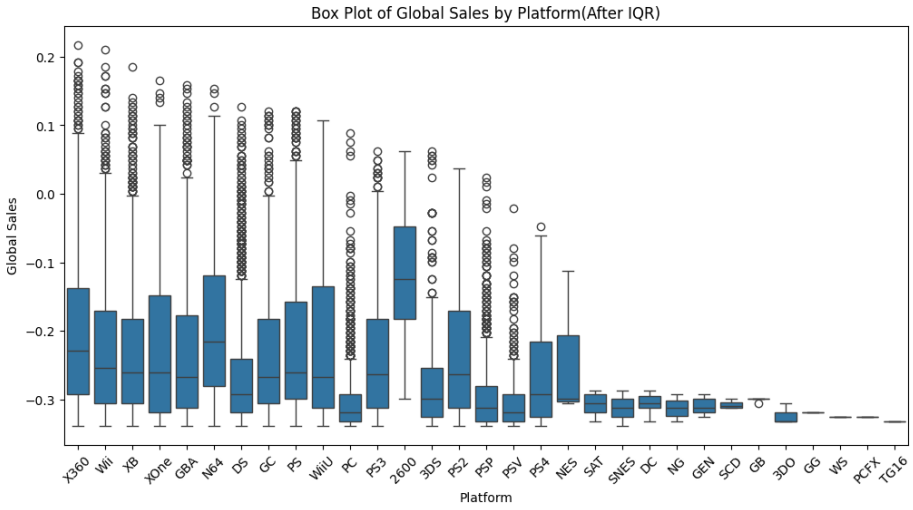


Fig 12: Box plot of global sales by platform after IQR

The application of the Inter Quartile Range method effectively reduces the impact of extreme values, improving the distribution of values while keeping the residual outliers. However, the platform with high performing sales shows the bigger value of interquartile range. The IQR reduced most of the extreme values that will help to improve the model performance.

Boxplot of Global sales by Genre

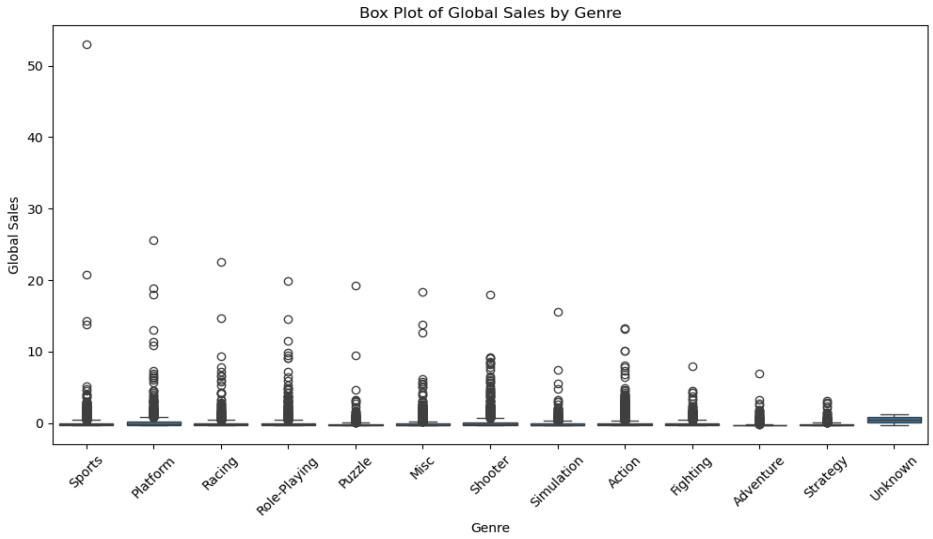


Fig 13: Box plot by video games genre.

The box plot of the global sales based on genre is giving some significant insights. Genre like sports outperform all the genre all the platform with numerous outliers reaching more than 50 million dollars. Blockbuster title like FIFA contributed significantly for the global sales of sports. Some genre like Racing, Role-Playing, Shooter have also generated more sales but relatively lower than the platform like Sports with significant lower outliers.

Moderately performing genres like Action, Fighting, Adventure also generate consistent sales. These genres balance steady-performing games with occasional outlier, such as call of duty of Actions and Need for Speed of Racing. On the lower side of the boxplot, genres like Strategy,

has the lowest median sales among the genres in this graph little significant outliers. This means those platforms have lack of market shares and failed to produce blockbuster titles games.

A significant features of the dataset is the presence of numerous outliers, particularly for top-performing genres like Sports, Racing and Role-Playing. These outliers are the reason for the skewed the graph. Those outliers need to handles from getting rid from bad accuracy score from modeling. To handling outliers, Inter Quartile Range is the best technique to exclude extreme values from individual features that helps to prevent skewness of the plot.

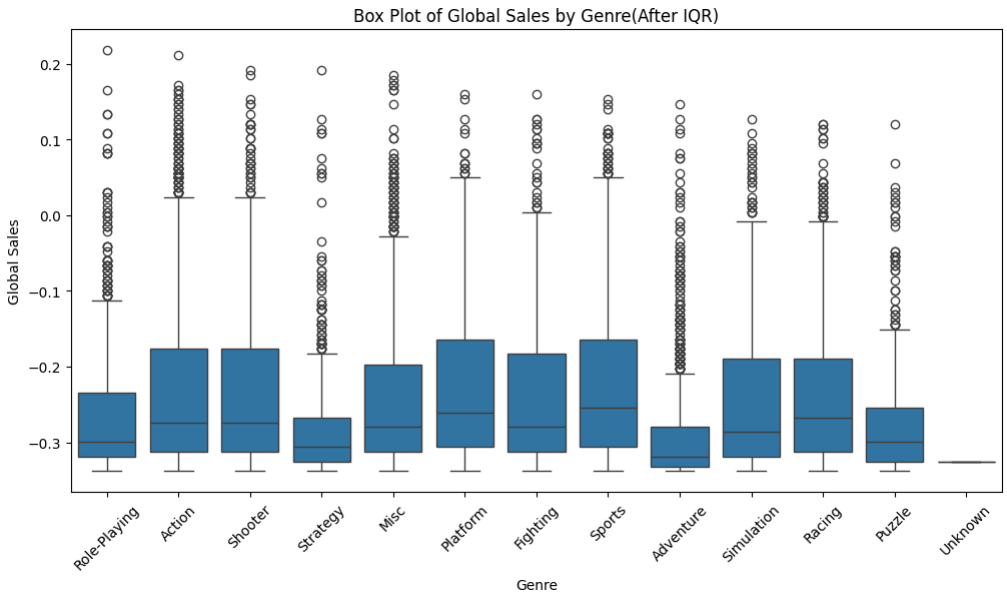


Fig 14: Box plot by video games genre after IQR.

The application of the Inter Quartile Range method effectively reduces the impact of extreme values, improving the distribution of values while keeping the residual outliers. However, the genre with high performing sales shows the bigger value of interquartile range. The IQR reduced most of the extreme values that will help to improve the model performance.

Scatterplot of NA_Sales by Global_Sales

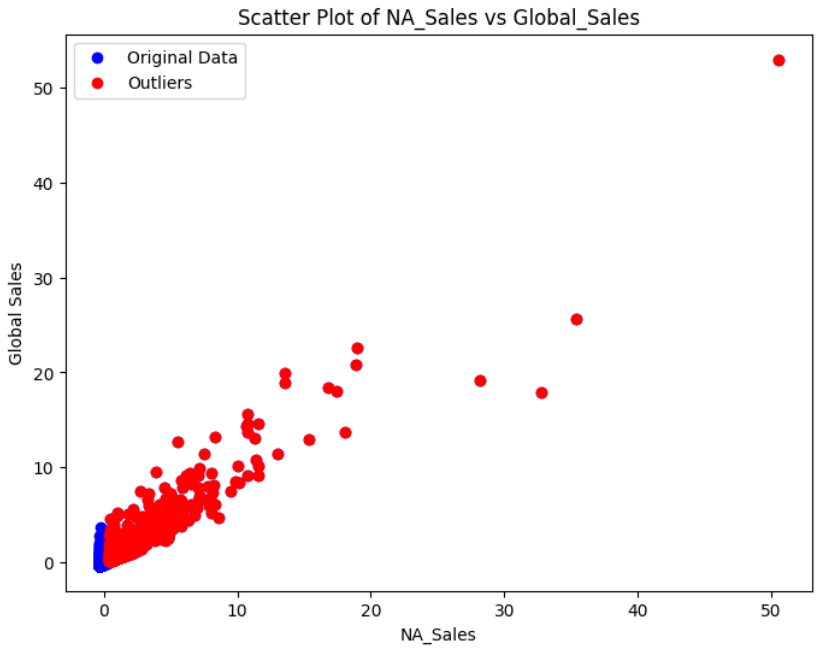


Fig 15: Scatter plot of NA_Sales vs Global_Sales with outliers highlighted.

This scatter plot shows the sales trend between NA_Sales(North America Sales) and Global_sales is showing a strong positive correlation between two attributes. In thisgraph, red data pint representing the extreme values known as outliers and the blue data point indicating the original values. Most of the data point clustered in the region indicating that most the games generates lower sales in North America region. Extremely high value in this plot says that some blockbuster games generate extremely high sales. Some games such as FIFA, Call of Duty are the keys to generate highest global sales crossed more than 50 millions of global sales and almost 50 million in North America sales. Strong relation with North America indicating that this market influenced highly to generate global sales.

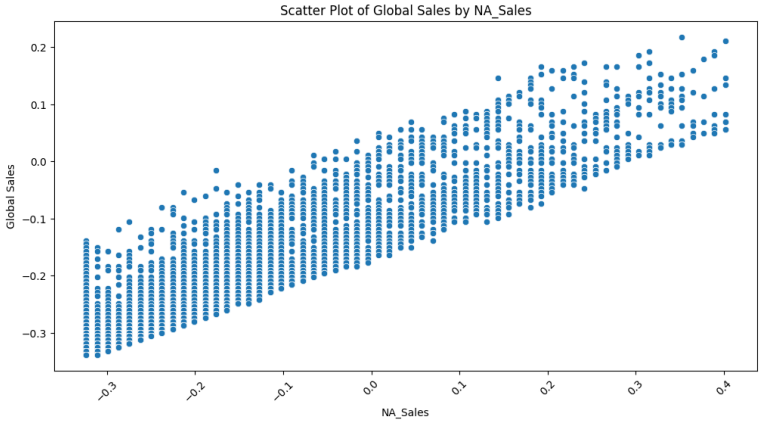


Fig 16: Scatter plot of Global Sales by NA_Sales after IQR

After applying IQR all the datapoints spread along with both axes. The standardized values follows a linear trend that indicates a strong positive relation with NA_Sales and Global Sales.

Scatter plot of Other_Sales vs Global_Sales

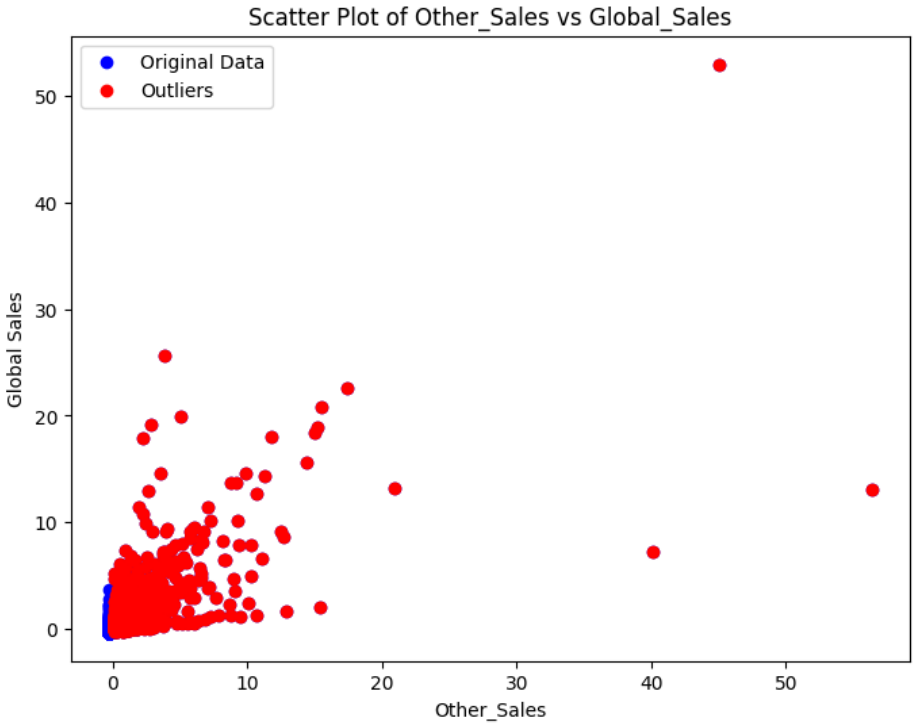


Fig 17: Scatter plot of Other_Sales vs Global_Sales with outliers highlighted.

In this scatter plot we can see the relationship between Other_Sales (sales from regions outside This scatter plot shows the sales trend between NA_Sales(North America Sales) and Global_sales is showing a strong positive correlation between two attributes. In thisgraph, red data pint representing the extreme values known as outliers and the blue data point indicating the original values. Most of the data point clustered in the region indicating that most the games generates lower sales in North America region. Extremely high value in this plot says that some blockbuster games generate extremely high sales. Some games such as FIFA, Call of Duty are the keys to generate highest global sales crossed more than 50 millions of global sales and almost 50 million in North America sales. Strong relation with North America indicating that this market influenced highly to generate global sales.

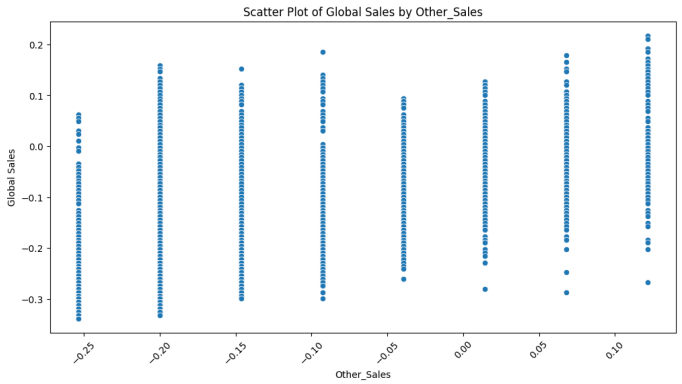


Fig 18: Scatter plot of Global Sales by Other_Sales after IQR

After applying IQR most of the data points clustered separately with standardized values indicating that for specific level of Other Sales in region has different variability. However, the lack of extreme values indicates that outliers have been effectively removed through IQR

application, providing a cleaner view of the general trend. From the scatter plot It can be seen that Other Sales has less contribution to Global Sales.

Inter Quartile Range is the crucial technique to detect outliers from dataset. It helps to reduce extreme values that influence to model accuracy. We also detected and removed outliers across multiple variables, including JP_Sales, EU_Sales, Other_Sales, and categorical features such as Publisher and Name, has significantly improved the dataset's integrity and reliability.

Analysis of Variance (ANOVA)

The Analysis of Variance (ANOVA) test is a statistical method used to compare the means of two or more groups to determine if there are statistically significant differences between them. ANOVA test calculate F-Statistic and P-Value to determine whether the attributes are significant differences between the means of group. For ANOVA test we will use:

$$F = \frac{Between\ Group\ Varianc}{Within\ Group\ Variance}$$

$$p - value = P(F \geq Fobserved \mid H0\ is\ true)$$

ANOVA table:

Name	F-statistic	P-value
Platform	47.43	3.92×10^{-264}
Genre	44.68	1.32×10^{-104}
Publisher	5.98	2.55×10^{-319}
Name	1.62	1.75×10^{-55}

Fig 19: ANOVA table

After performing the ANOVA test, the Name attribute was excluded from further analysis due to its relatively low F-statistic (1.62) and minimal effect size, despite a statistically significant p-value (1.75×10^{-55}). While the differences in means across individual game titles were detected, the impact was considerably smaller compared to broader factors such as Platform, Genre, and Publisher. These attributes demonstrated higher F-statistics and substantial variability in global sales, emphasizing their importance in the analysis. Excluding the Name attribute streamlines the dataset by removing a factor with limited influence, allowing the focus to remain on the more impactful attributes. This decision ensures a more efficient analysis and modeling process by concentrating on the key drivers of sales trends.

Correlation Heatmap

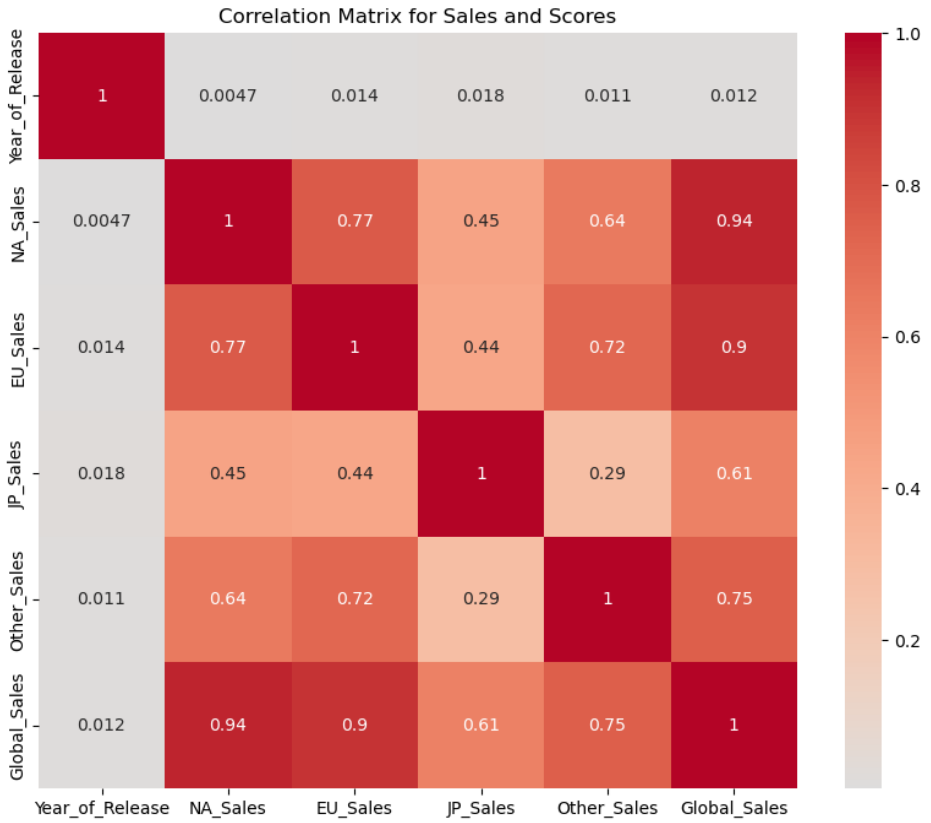


Fig 20: Correlation Heatmap for numerical attributes

The correlation heatmap visualizes the relationships between various sales metrics and the year of release, highlighting their pairwise correlation coefficients. Year_of_Release exhibits the weakest correlations with all other variables, with the highest correlation being only 0.018. This suggests that Year_of_Release has minimal influence on sales and can be excluded from further analysis due to its lack of meaningful contribution.

NA_Sales and EU_Sales are highly correlated with Global_Sales, with coefficients of 0.94 and 0.90, respectively. These strong correlations indicate multicollinearity, as these regional sales strongly overlap with and contribute to global sales. To avoid redundancy and multicollinearity issues in predictive modeling, these two variables can be excluded in favor of Global_Sales, which provides a comprehensive view of overall performance.

Other sales variables such as JP_Sales and Other_Sales show moderate correlations with Global_Sales (0.61 and 0.75, respectively), suggesting that they still contribute uniquely to global sales and should be retained for analysis.

Feature Engineering

Feature engineering was applied to the Platform, Genre, and Publisher attributes in the dataset by converting their nominal categorical values into numerical values using one-hot encoding. This technique creates binary indicator variables for each category within these attributes, enabling machine learning models to interpret the categorical data. Specifically, one-hot encoding generates new columns corresponding to each unique category in the original attributes, assigning a value of 1 if the category is present in a record and 0 otherwise.

$$Encoded\ Column\ for\ c_i = \{\begin{matrix} 1 & if\ the\ of\ x\ =\ c_i \\ 0 & otherwise \end{matrix}\}$$

Here:

c_i: A unique category of the variable

	Name	Year_of_Release	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales	Platform_3DO	Platform_3DS	Platform_DC	...	Publisher_Zoo Game
2384	Kingdoms of Amalur: Reckoning	2012	0.352395	0.188717	-0.218913	0.121412	0.217365	False	False	False	...	Fals
2413	Ghostbusters: The Video Game	2009	0.401566	0.109236	-0.251295	0.121412	0.210905	False	False	False	...	Fals
2481	Duke Nukem Forever	2011	0.315517	0.208587	-0.251295	0.121412	0.191524	False	False	False	...	Fals
2497	Sid Meier's Civilization Revolution	2008	0.389273	0.049626	-0.251295	0.121412	0.191524	False	False	False	...	Fals
2503	James Bond 007: Nightfire	2002	0.389273	0.148977	-0.251295	-0.092830	0.185063	False	False	False	...	Fals
...	...	...	...	...	...	...	...	...	...	...	...	..
16714	Samurai Warriors: Sanada Maru	2016	-0.323705	-0.288166	-0.218913	-0.253512	-0.338230	False	False	False	...	Fals
16715	LMA Manager 2007	2006	-0.323705	-0.268296	-0.251295	-0.253512	-0.338230	False	False	False	...	Fals
16716	Haitaka no Psychedelica	2016	-0.323705	-0.288166	-0.218913	-0.253512	-0.338230	False	False	False	...	Fals
16717	Spirits & Spells	2003	-0.311412	-0.288166	-0.251295	-0.253512	-0.338230	False	False	False	...	Fals
16718	Winning Post 8 2016	2016	-0.323705	-0.288166	-0.218913	-0.253512	-0.338230	False	False	False	...	Fals

11820 rows × 589 columns

Fig 21: Transform dataset after applying One-Hot encoding.

This table showcases the transformed dataset after applying one-hot encoding to the Platform, Genre, and Publisher attributes. Each unique category in these attributes has been converted into a binary column, where **1** indicates the presence of a category and **0** (or False) indicates its absence.

Modelling

Random Forest and XGBoost Regressor have been applied to predict global sales. We have choose these two model for their ability to handle large dataset and the robustness to predict the best accuracy for the dataset.

Random Forest:

Random Forest is a powerful machine learning model to enable us make accurate predictions and analyze complex datasets [4]. It is the model that combines multiple decision trees to create a single model. A different subset of each tree built in each forest to make an independent prediction. In this analysis, Random Forest was utilized to capture robust patterns and interactions between features while maintaining the interpretability through feature importance.

We applied Random Forest Regressor for the robustness and handle the complex dataset. We were abled to predict global sales with hyperparameters to optimized the model performance. In this model, the final model used 500 estimators with no maximum depth, a minimum of 1 sample per leaf and to split internal node minimum 2 split required. This model shows the strong predictive ability. The final result has been demonstrated below:

Final Result:

Result	Mean Absolute Error	Mean Squared Error	R-Squre
Model Evaluation			
	0.0223	0.0013	0.875
Final Cross-Validation Result			
Train	0.0132	0.0005	0.948
Test	0.0227	0.0015	0.862

Fig 22: Model Evaluation metric of Random Forest

For the test dataset, the model achieved a Mean Absolute Error (MAE) of 0.0223, a Mean Squared Error (MSE) of 0.0013, and an R^2 score of 0.8756, indicating that the model explained approximately 87.6% of the variance in the target variable. Cross-validation further validated the model's robustness, with mean test results showing a MAE of 0.0227, MSE of 0.0015, and R^2 score of 0.8624. These metrics confirm the model's generalizability across different data splits.

On the training dataset, the model performed exceptionally well, with a Mean MAE of 0.0132, Mean MSE of 0.0005, and R^2 score of 0.9480, reflecting the model's ability to capture patterns effectively. Random Forest delivered strong result to predict global sales. However, train model performed better than test model which indicates that slightly overfitted model. In this case it is expected with this model.

Visualize the actual vs predicted global sales of Random Forest

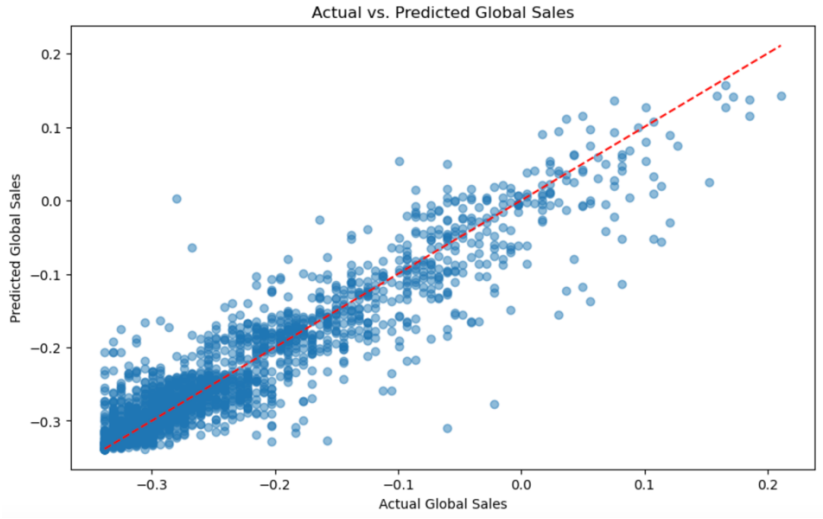


Fig 23: Scatter plot of actual and predicted global sales of random forest

The scatter plot demonstrates that the Random Forest Regressor effectively predicts global sales, with most predictions closely matching the actual values. This visualization supports the evaluation metrics (e.g., R^2 and MAE), confirming the model's robustness and reliability.

Residual Analysis of Random Forest

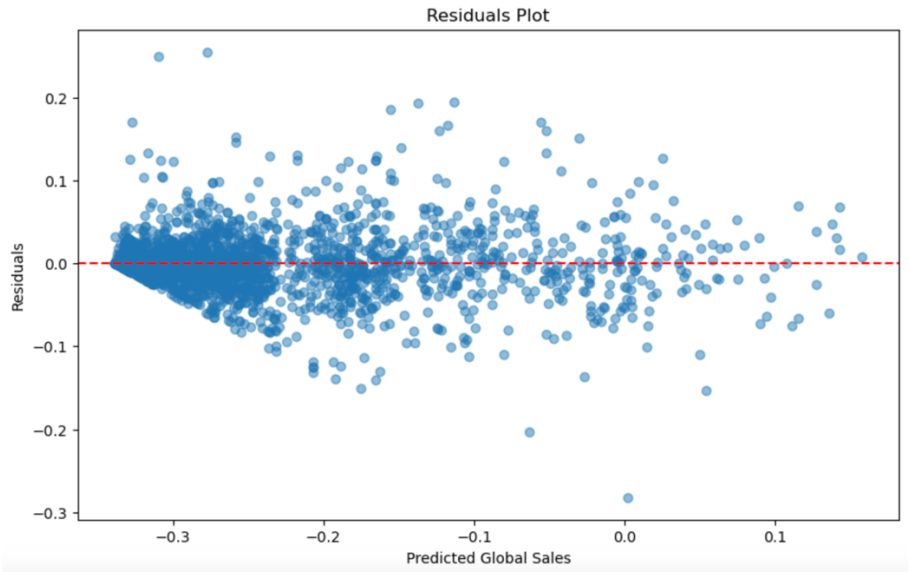


Fig 24: Residual Analysis of Random Forest

In this residual plot calculates the different between actual global sales and predicted global sales.

$$Residual = Actual\ Value - Predicted\ Value$$

The residual plot demonstrates that the Random Forest Regressor provides accurate predictions with residuals centered around zero and no evident bias. The absence of discernible patterns in the residuals confirms the model's reliability for predicting global video game sales.

XGBoost Regressor:

XGBoost is another powerful machine learning model to train complex dataset. The XGBoost stands for Extreme Gradient Boosting algorithm is very popular which is iteratively minimizes the error [5]. One of the key feature of the model is, it is very efficient to handle missing values that can handle missing values without allowing the pre-processing the data. Most importantly, XGBoost allows parallel processing which is making it possible to train large dataset.

We applied XGBoost regressor for the parallel processing and handle the complex dataset. We were able to predict global sales with hyperparameters to optimized the model performance. In this model, the final model used learning rate of 0.2 and maximum depth 3500 estimators with subsampling 80% of data. This model has also set regularization parameter with reg_alpha=0 and reg_lambda = 5, while the colsample_bytree parameter was set to 1.0, indicating all features were considered at each split. The final result has been demonstrated below:

Final Result:

Result	Mean Absolute Error	Mean Squared Error	R-Squre
Model Evaluation			
	0.0216	0.0012	0.8848
Final Cross-Validation Result			
Train	0.0204	0.0010	0.9016
Test	0.0225	0.0014	0.8719

Fig 25: Model Evaluation metric of XGBoost Regressor

The model performed strongly, achieving a Mean Absolute Error (MAE) of 0.0216, a Mean Squared Error (MSE) of 0.0012, and an R^2 score of 0.8848 on the test dataset, explaining 88.48% of the variance in global sales. Cross-validation confirmed the model's robustness, with mean test metrics showing a MAE of 0.0225**,** MSE of 0.0014, and an R^2 score of 0.8719. These results indicate consistent performance across different data splits, demonstrating the model's generalizability.

On the training dataset, the model showed excellent performance with a Mean MAE of 0.0204, MSE of 0.0010, and an R^2 score of 0.9016, indicating its ability to effectively capture complex patterns. The slightly better results on the training set compared to the test set suggest minimal overfitting, which is well-controlled due to the inclusion of regularization techniques. Overall, the XGBoost Regressor delivered robust and reliable predictions, making it a strong choice for modeling global video game sales.

Visualize the actual vs predicted global sales

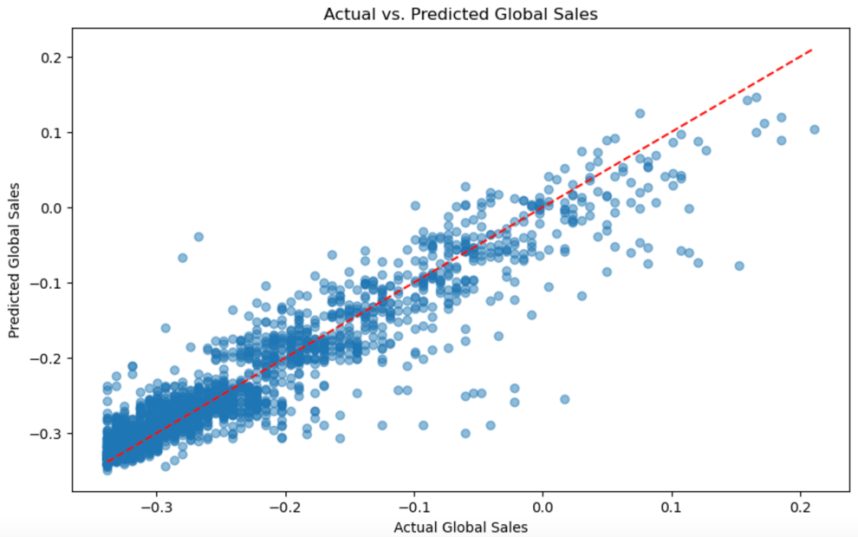


Fig 26: Scatter plot of actual and predicted global sales of XGBoost regressor

The scatter plot demonstrates that the Random Forest Regressor effectively predicts global sales, with most predictions closely matching the actual values. This visualization supports the high R^2 score and low error metrics obtained during the model evaluation.

Residual Analysis of XGBoost Regressor

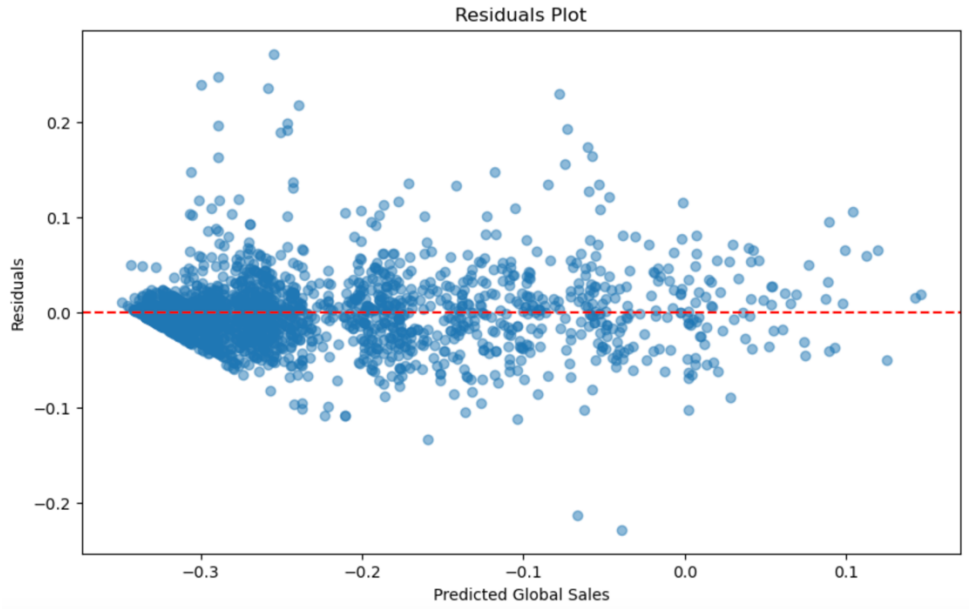


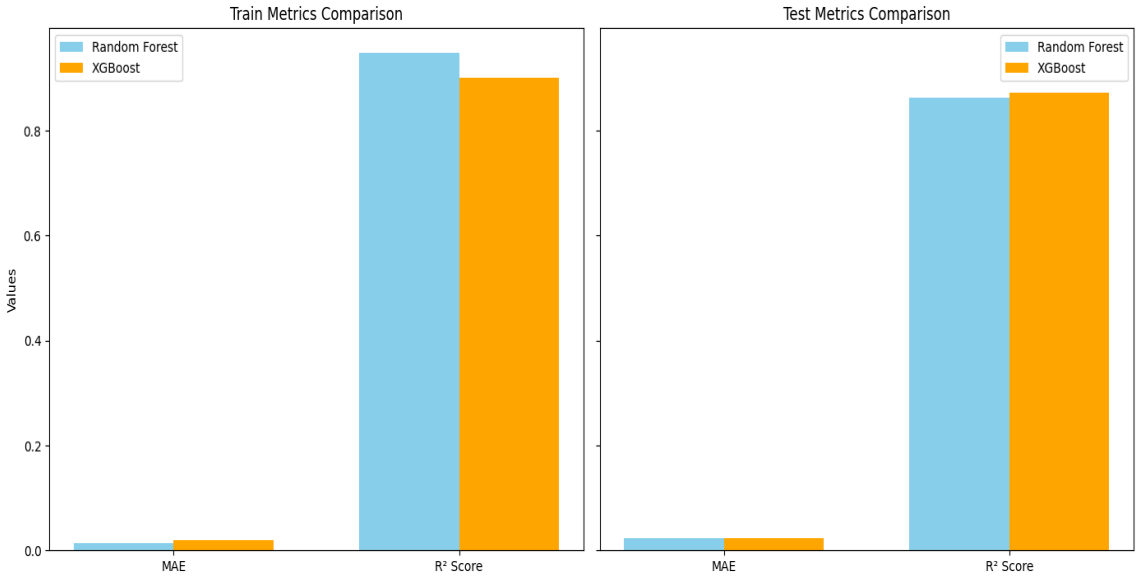
Fig 27: Residual Analysis of XGBoost regressor

In this residual plot calculates the different between actual global sales and predicted global sales for XGBoost regressor.

$$Residual_{XGBoost} = Actual\ Value_{XGBoost} - Predicted\ Value_{XGBoost}$$

The residual plot demonstrates that the XGBoost Regressor provides accurate predictions, with most residuals distributed randomly and close to zero. However, XGBoost slightly outperforms Random Forest in terms of residual consistency and accuracy, particularly for higher sales values and edge cases.

Overall model comparison



From bar plot it can be seen that, both model performed well for the dataset. However, XGBoost performed slightly better than Random Forest on the test set that indicates that achieved better generalization and predictive accuracy.

Conclusion

This analysis gives a detailed analysis of global video game sales showing important insights into market dynamics consumer preferences and the factors that mainly drive the sales performance. The key findings of this project are:

1. **Regional Variations:**
   - North America showed a really strong sales for Action and Shooter genres but Japan showed great demand for Role-Playing Games (RPGs).
   - Sales patterns across Europe and other regions showed a mix of preferences with Sports and Racing genres performing regularly well.

2. **Platform Trends:**
   - Consoles like PlayStation and Xboxx were the main platforms in the global sales showing their widespread appeal and robust ecosystems.
   - Older platforms showed smaller returns showing us the importance of staying updated with technological advancements.

3. **Genre and Release Year Influence:**
   - Genres play important role in a game's success with some categories constantly outperforming others.
   - The year of release correlates with sales showing that strategic timing is important for increasing a game's market potential.

Model Performance

Random Forest Regressor: Had an R^2 value of 0.86 making it the second most accurate model for predicting sales based on the dataset. Its ability to handle non-linear relationships and avoid overfitting is effective to a great extent.

XGBoost: gave an R^2 value of 0.98 showing strong performance with its gradient boosting approach it sequentially increased model's predictions.

The combination of these models shows the importance of using ensemble learning techniques for analyzing complex datasets. While XGBoost gave greater accuracy, Random forest gave competitive results with robust performance.

Future Implications

The findings of this project show many opportunities for industry stakeholders:
Game Development: it Focuses on genres and platforms that are in line with regional preferences and market trends.
Strategic Marketing: targeted advertising with the regional audiences and using timing for more impactful releases.
By combining data driven insights with domain knowledge this project is a valuable resource for understanding and figuring out the global video game market.

References

1. Kaggle Dataset. (n.d.). *Video Game Sales Dataset*. Retrieved from https://www.kaggle.com/
2. VanderPlas, J. (2016). *Python Data Science Handbook*. O'Reilly Media.
3. Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: Principles and Practice* (2nd ed.). OTexts. Retrieved from https://otexts.com/fpp2/
4. Breiman, L. (2001). Random Forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324
5. Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785