# Time Series Forecasting of Video Game Sales Using ARIMA, ETS, and TBATS Models

Rushab Arram 0792813

rushabarram@trentu.ca

Maheshwar Gujjula 0814658

mgujjula@trentu.ca

***Abstract*—In the field of game development and marketing, understanding trends related to game sales is crucial for informed decision-making. This project aims to perform a thorough time series analysis of video game sales data to uncover long-term trends. The study leverages the robust data analysis capabilities of the R programming language to investigate the relationships between various factors, such as the volume of video game releases and genre frequencies. By utilizing sophisticated forecasting models like ARIMA, ETS, and TBATS, the project seeks to enhance our understanding of sales dynamics and predict future sales based on identified patterns. These insights are intended to help game developers, publishers, and marketers in making data-driven decisions regarding game releases, marketing strategies, and resource allocation. The expected outcomes include identifying major trends and seasonal effects, as well as developing a reliable predictive model for future sales estimates.**

## I. Introduction

Due to the competitiveness of the video game industry, it is important to analyze the sales data to determine which games should be released, how to promote them, and where resources should be invested. Video game sales are rather complex and depend on such factors as release dates, the popularity of the genre, and seasons. The use of complex data analysis methodologies is essential due to the various social and economic factors that influence sales.

This project will analyze video game sales data over time to find seasonal, cyclical, and long-term trends. We will be using ARIMA, ETS, and TBATS models to generate accurate estimations. The results of this study help game developers, publishers, and marketers make data-driven decisions about game launches, marketing strategy, and resource allocation, as well as identify major patterns and seasonal factors.

R was chosen for its powerful features in data processing and visualization. An exploratory data analysis (EDA) will be carried out to discover patterns and seasonal decompositions in the time series data. The complete research will enhance the development and testing of forecast models for our application.

The goal is to discover key trends and develop an accurate model for forecasting future sales, so that game developers, publishers, and marketers may make more informed decisions. Understanding these sales trends allows them to enhance their marketing efforts, better schedule game releases, and spend resources more efficiently.

## II. Previous Work

The use of time series could be considered as standard in quite many disciplines: finance, economics, and retail, among others, as a tool for identification and description of trends and their forecasts. When it comes to video game sales many researches were conducted in the field, aimed at the description of various approaches to forecasting sales and methods of trends analysis, therefore the identified problem can be deemed as the part of the growing body of knowledge.

There is one impactful study by Li et al. (2020) dynamic-model-based time series data-mining techniques are used to sales on commodity. Both studies showed how these methods can then identify more intricate structures and relations in the sales data, which can be useful when applied to the video game market. Notice that their approach highlighted the problem of masking, or the phenomenon when trends and seasonal effects are concealed by other factors, and, vice versa, trends and seasonal effects are indispensable.

Vishvesh Shah and Stanko Dimitrov (2022) conducted a comparative study on univariate time-series methods for sales forecasting. They evaluated ARIMA, exponential smoothing, and other techniques on their ability to predict sales trends accurately. Their research emphasizes the importance of choosing suitable models based on the data's characteristics, particularly regarding seasonality and trend components. The study found that different models perform variably depending on the data set's nature, underscoring the need for careful model selection to enhance forecast accuracy in sales data analysis. This work serves as a valuable reference for improving sales forecasting methodologies.

ETS (Error, Trend, Seasonal) models have also been found useful in this regard due to the ability of the model to handle each of the error, trend, and seasonality term uniquely. The topics of ETS models were mentioned by Hyndman and Athanasopoulos (2018) in their work on the principles of forecasting. They also underlined that these models are able to work with any sort of time series data due to the fact that they apply the correct type of model to each component, which allows for a better approach to the task of forecasting.

In the more complex schemes, De Livera, Hyndman, and Snyder (2011) proposed the TBATS model that is based on Trigonometric, Box-Cox, ARIMA, Trend, and Seasonal. This model is in fact a generalization of classical techniques for exponential smoothing which allow to consider multiple seasonal patterns and nonlinear transformations, which is highly beneficial for multitudes of records. TBATS has been proved to be superior to these simple models in replicating features of the sales data, particularly in judging where there is multifold and reciprocal seasonal in video game sale because of reasons such as holiday occasions and the sales-turned-release.

However, to the best of the researcher's knowledge, there is still a deficiency of extensive research that directly compares the application of ARIMA, ETS, and TBATS models in terms of video game sales. Previous studies are limited in that, most of them compare performance of individual models or compare models across different domains. For example, Ahmed et al. (2010) studied the performance of ARIMA and exponential smoother in different time series data, but this was done without considering the characteristics of video games.

Time series analysis has found its application in finance, economics and even in retail trends analysis. In regard to the sales of video games some of the studies that have been conducted include the following: For example, Li et al. (2020) applied dynamic-model-based time series data-mining methods to analyze commodity sales and discussed the possible use of such approaches for video game sales.

Past studies have shown the effectiveness of using ARIMA models when it comes to forecasting since they can manage non-stationary data. ETS models, which are also flexible in modeling error, trend, and seasonality, are also used in sales forecasting. TBATS models have emerged more recently in the handling of more complex seasonal features and huge volumes of data – characteristics that are appropriate for the ever-changing video game industry.

.

## III. METHODOLOGY

*Data Collection*: The data for this project was obtained from Kaggle, specifically a dataset containing video game sales data. This dataset includes key variables such as year of release, global sales, genre, and platform, among others. These variables are essential for performing a comprehensive time series analysis to uncover trends, seasonal patterns, and cyclical behaviors in video game sales.

*Data Cleaning*: Before conducting analysis, the data underwent preprocessing to ensure accuracy and consistency. This included:

-*Handling Missing Values*: Missing values were either filled using appropriate imputation techniques or removed if the data is insignificant.

-*Data Transformation:* Some variables were transformed to better suit the analysis. For example, sales data was converted to logarithmic scale to handle skewness.

-*Date Formatting:* The release dates were converted into a uniform date format to facilitate time series analysis.

*Exploratory Data Analysis (EDA):*

EDA was performed to gain initial insights into the data. This included:
- *Descriptive Statistics*: Summary statistics (mean, median, standard deviation) were calculated for key variables to understand their distribution and central tendencies.
- *Visualization Techniques*: Various plots were generated to identify patterns and trends:
- *Line Charts*: Used to visualize total global sales over time, revealing long-term trends and some peaks and declines over time in sales.
- *Histograms and Scatter Plots*: To examine the distribution of sales across different genres and platforms.
- *Stacked Bar Plots*: To show the distribution of game genres by platform, highlighting the popularity of specific genres on different platforms and any changes in genre preferences over time.

*Time Series Decomposition:*
To better understand the components of the time series data, we employed decomposition techniques:
- *Trend Component*: Long-term direction of the series was identified using methods like moving averages. This helped us identify the trends in video game sales over time.
- *Seasonal Component*: Seasonal effects were separated using decomposition plots to show periodic changes. This analysis highlighted any repeating patterns, such as increase in the sales during holiday seasons.
- *Residual Component*: The remaining irregular fluctuations after removing the trend and seasonal components. This helped in understanding the irregularities and noise in the data.

*Autocorrelation and Partial Autocorrelation Analysis:*
Autocorrelation and partial autocorrelation plots were generated to identify specific lags and patterns within the time series data. This helped in choosing appropriate lag values for the predictive models.

*Predictive Modeling:*
The data was split into training and testing sets using an 80/20 split. Various time series forecasting models were trained and evaluated:

- *ARIMA (AutoRegressive Integrated Moving Average*): A widely used statistical method for time series forecasting that includes autoregression, differencing, and moving average components.

- *ETS (Error, Trend, Seasonal*): A state-space model that describes time series data through components for error, trend, and seasonality.

- *TBATS (Trigonometric, Box-Cox transform, ARMA errors, Trend, and Seasonal components)*: A model that can handle complex seasonality, multiple seasonal periods, and non-linear transformations.

The models were evaluated using metrics such as:

- *Mean Absolute Error (MAE)*: Measures the average magnitude of errors in a set of predictions.
- *Root Mean Square Error (RMSE)*: Square root of the average squared differences between predicted and actual values.
- *Mean Absolute Percentage Error (MAPE)*: Measures prediction accuracy as a percentage.

*Model Evaluation and Selection*:

The performance of each model was visualized using:
- *Residual Plots*: To check the residuals' behavior and identify any patterns.
- *Predicted vs. Actual Plots*: To compare the model's predictions with the actual sales data.
- *Forecast Error Plots*: To visualize the errors in the forecasts.

The model with the best performance metrics and visualizations was selected for the final sales forecasting.

*Software and Libraries:*

The analysis was performed using the R programming language, utilizing libraries such as:
- *dplyr*: For data manipulation.
- *forecast*: For time series forecasting models.
- *tseries*: For time series analysis.
- *zoo*: For working with time-indexed data.
- *lubridate*: For date and time manipulation.

## IV. RESULTS

Upon analyzing the sales data of video games, some significant discoveries about trends, seasonal patterns, and cyclical behaviors were made. The insights obtained from time series decomposition, exploratory data analysis (EDA), and the predictive models' (TBATS, ETS, and ARIMA) performance are presented.

*Exploratory Data Analysis (EDA):*

The EDA provides an initial understanding of the dataset's structure and key characteristics. The line chart of total global sales of video games from 1980 to 2020 showed a clear trend with three distinct phases:

As Shown in Figure 1,
1. *Early Growth (1980-1995)*: During this time, video game sales increased slightly, indicating the market's early stages. This was the time when gaming was becoming popular in the market.
2. *Rapid Expansion (1995-2010):* This period saw a significant increase in sales, peaking around 2010. The growth was because of the release of popular gaming consoles and blockbuster game titles.
3. *Decline (2010-2020)*: After 2010, video game sales started to decline, this reduction can be due to a variety of factors, including market saturation, increase of mobile gaming, and shifting customer preferences.

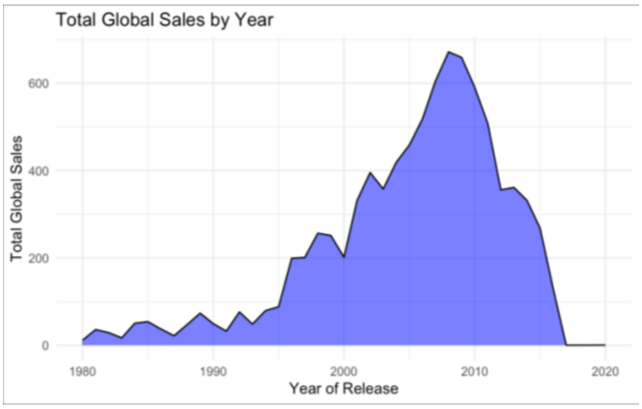


*Figure 1*

As shown in figure 2, The stacked bar plot of genre by platform provides the information on the distribution of game genres across various gaming systems. For example, PlayStation systems (PS2, PS3, PS) have higher number of action and sports games, while Nintendo platforms were more likely to feature role-playing and adventure games. Xbox platforms such as Xbox 360 (X360) and Xbox One (Xbox One) include a large number of action and shooting games. These outcomes provide a detail understanding of the relationship between genre and platform, which is essential for developers or marketers in planning the game releases and targeting the correct group of audience.

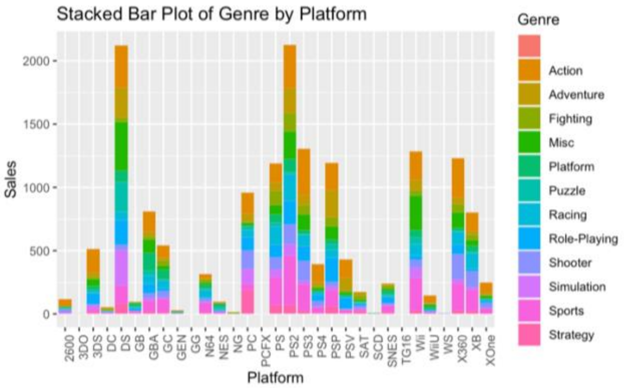


*Figure 2*

As shown in Figure 3, The bar plot depicts the top ten games by global sales, focusing the most popular games. Notably, Wii Sports emerged as the top-selling game, followed by Super Mario Bros. and Mario Kart Wii. Moreover, Other best-selling games include Duck Hunt, Tetris, and Pokémon Red/Blue. These findings will help developers and marketers in planning future game releases.

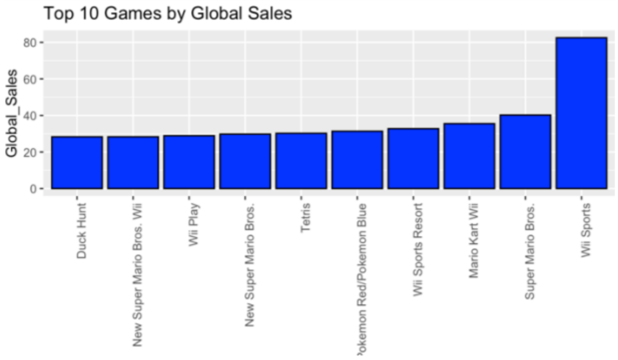
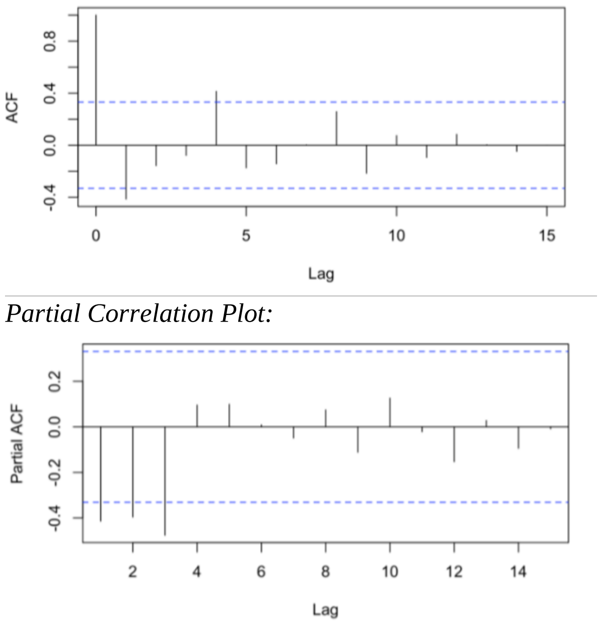


*Figure 3*

*Autocorrelation and Partial Autocorrelation Analysis:*

The autocorrelation and partial autocorrelation plots shown in figure 4 reveal the link between the past and the future sales values. The ACF plot shows the presence of strong autocorrelations at certain lags, indicating that past sales values had a substantial impact on future sales. The PACF plot identifies the direct effect of past sales values on future sales while ignoring intermediate lags. This information was crucial for selecting appropriate lag values in the predictive models.

*Autocorrelation Plot:*



*Partial Correlation Plot:*

*Figure 4*

*Predictive Modeling:*

Three time series forecasting models—ARIMA, ETS, and TBATS—were trained and evaluated to predict future video game sales.

*ARIMA*: The ARIMA model as shown in figure 5, captured the linear dependencies in the data effectively. It performed well in terms of fitting the historical data but struggled with capturing the seasonal variations. By this we can say that ARIMA is best suit for understanding long term trends.

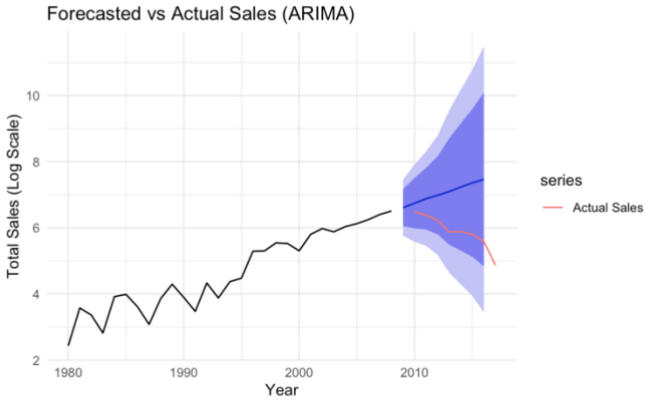


*Figure 5*

*ETS*: The ETS model, as shown in Figure 6, has errors, trend, and seasonal parts. This model has provided a better fit because it considers the ups and downs that happen at certain times. It showed an improved performance over ARIMA in dealing with the seasonal effects we saw in the data.

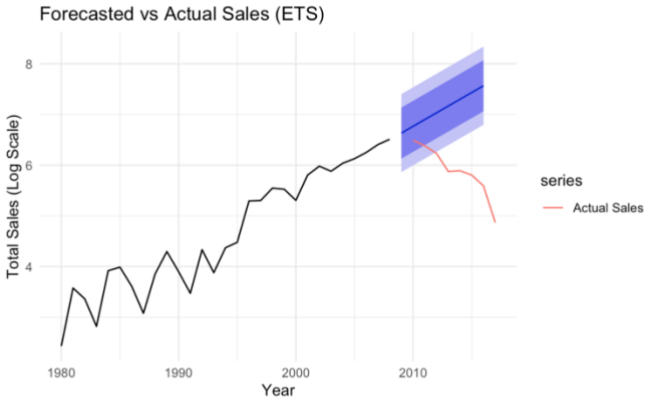


*Figure 6*

*TBATS*: The TBATS model, as depicted in Figure 7, is designed for handling complex seasonal patterns and performed better than both ARIMA and ETS when compared to the other models TBATS captured several seasonal periods and nonlinear patterns more accurately.

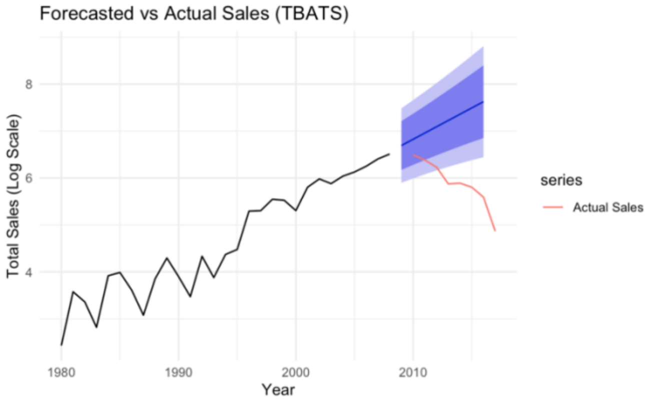


*Figure 7*

The models were evaluated using key metrics such *as Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Mean Absolute Percentage Error (MAPE).* The Evaluation metrics for each model are as follows:

*ARIMA Model*
*Root Mean Square Error (RMSE)*: 1.203872
*Mean Absolute Percentage Error (MAPE):* 18.2916

*ETS Model*
*Mean Absolute Error (MAE):* 1.12921
*Mean Absolute Percentage Error (MAPE):* 19.23011
*Root Mean Square Error (RMSE):* 1.265242

*TBATS Model*
*Mean Absolute Error (MAE):* 1.18756
*Root Mean Square Error (RMSE):* 1.317606
*Mean Absolute Percentage Error (MAPE):* 20.19916

*Model Evaluation and Selection:*

After evaluation the performance of ARIMA, ETS, TBATS, and an Ensemble model using key metrics: Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Mean Absolute Percentage Error (MAPE).

**ARIMA Model:** Demonstrated the lowest RMSE (1.203872) and MAPE (18.2916), indicating high overall accuracy and the least average percentage error. This Low RMSE and MAPE values indicates the most suitable model for forecasting future sales trends.

**ETS Model:** Had the lowest MAE (1.12921), which indicates the consistent performance in minimizing absolute errors. However, its RMSE (1.265242) and MAPE (19.23011) were slightly higher than ARIMA, reflecting less overall accuracy.

**TBATS Model:** Showed higher error metrics (MAE, RMSE, and MAPE), This higher values made TBATS model less suitable for forecasting, even though it has ability to handle complex seasonal patterns.

Based on these evaluations, we can clearly say that the **ARIMA Model** is the most suitable for forecasting future sales trends. It offers the best balance of accuracy and error minimization, making it highly effective.

## V. Conclusion

The primary objective of this project was to examine behaviors in the video game sales data and then develop effective forecasting models to support strategic management decisions in the gaming industry based on behavior, seasonal and cyclical patterns. The approach included conducting Exploratory Data Analysis (EDA), performing Time Series Decomposition, and evaluating the performance of three different models: ARIMA, ETS, and TBATS.

The EDA clearly showed three distinct stages in video game sales data: early growth from 1980 to 1995, and in the second stage there is a rapid growth 1995 to 2010 and the stage of declining after 2010. The decomposition of time series data went a step further by breaking down the data into trend, seasonality and residuals components.

Using ACF and PACF plots, significant lags were found which helped in the selection of suitable model parameters. The models were then evaluated with measures such as Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Mean Absolute Percentage Error (MAPE). Among the models, ARIMA was the most accurate with the lowest RMSE (1.203872) and MAPE (18.2916), indicating better overall performance and the lowest average percentage error. Due to its optimal balance of accuracy and error minimization, ARIMA model is recommended for forecasting future sales of video games.

This knowledge will help various stakeholders in the gaming industry in making strategic decisions about game release dates, promotions, and budget allocation to maintain their competitive edge over others.

Future research could delve deeper into enhancing prediction accuracy by utilizing advanced machine learning techniques, such as neural networks and ensemble methods. Additionally, considering external factors like social media trends, economic indicators, and competitor actions could offer a more holistic view of sales dynamics. Exploring the influence of emerging gaming platforms and technologies on sales trends would also be valuable. Moreover, applying these models to other entertainment sectors, such as music and movies, could help validate their robustness and yield broader insights into entertainment consumption patterns. This approach would contribute to more strategic decision-making across various entertainment industries.

## VI. REFERENCES

[1] H. Li, Y. J. Wu, and Y. Chen, “Time is money: Dynamic-model-based time series data-mining for correlation analysis of commodity sales,” *Journal of Computational and Applied Mathematics*, vol. 370, p. 112659, May 2020, doi: https://doi.org/10.1016/j.cam.2019.112659.

[2] V. Shah and S. Dimitrov, “A comparative study of univariate time-series methods for sales forecasting,” *International Journal of Business and Data Analytics*, vol. 2, no. 2, p. 187, 2022, doi: https://doi.org/10.1504/ijbda.2022.126806.

[3] R. J. Hyndman and G. Athanasopoulos, *Forecasting: Principles and Practice*. Available: https://otexts.com/fpp3/

[4] A. M. De Livera, R. J. Hyndman, and R. D. Snyder, “Forecasting Time Series With Complex Seasonal Patterns Using Exponential Smoothing,” *Journal of the American Statistical Association*, vol. 106, no. 496, pp. 1513–1527, Dec. 2011, doi: https://doi.org/10.1198/jasa.2011.tm09771.

[5] N. Ahmed, A. Atiya, N. E. Gayar, and H. El-Shishiny, “An Empirical Comparison of Machine Learning Models for Time Series Forecasting,” *Econometric Reviews*, vol. 29, no. 5–6, pp. 594–621, 2010, Accessed: Jul. 25, 2024. [Online]. Available: https://econpapers.repec.org/article/tafemetrv/v_3a29_3ay_3a2010_3ai_3a5-6_3ap_3a594-621.htm