

Predicting Video Game Sales in North America: Data and EDA

Date of Presentation: 11-05-2024

Student Name: Nirja Shah

Student's Pace Email Address: ns53074n@pace.edu

Class Name: Practical Data Science

Program Name: MS in Data Science

Seidenberg School of Computer Science and Information Systems

Pace University

Agenda

- Executive summary
- Project plan recap
- Data
- Exploratory data analysis
- Modeling methods
- Findings
- Recommendations and technical next steps

Executive Summary

- Problem:

How can a predictive model based on past game sales by genre help identify which genres to prioritize for marketing in the North American region?

- Solution:

The goal is to identify a prediction model to estimate sales in the North American region based on genre using previous data on game sales. With this knowledge, the business can focus its marketing efforts on the genres that are most likely to succeed.

Project Plan

Deliverable	Due Date	Status
Data and EDA	11-05-2024	Completed
Methods, Findings and Recommendations	11-12-2024	Completed
Final Presentation	12-03-2024	Not Started

Data

Data

- Data Source: Video Game Sales 2024 from [Kaggle](#)
- Sample Size: 64016 Rows
- There are total 14 attributes, some of them are listed below:
 - **Publisher:** Name of the publisher.
 - **Developer:** Name of the Developer.
 - **Console:** Console on which this game was released on.
 - **Genre:** Genre of the game.
 - **Critic score:** Metacritic score (out of 10).
 - **North America sales:** North American sales of copies in millions.
 - **Release Date:** Date of the game released.

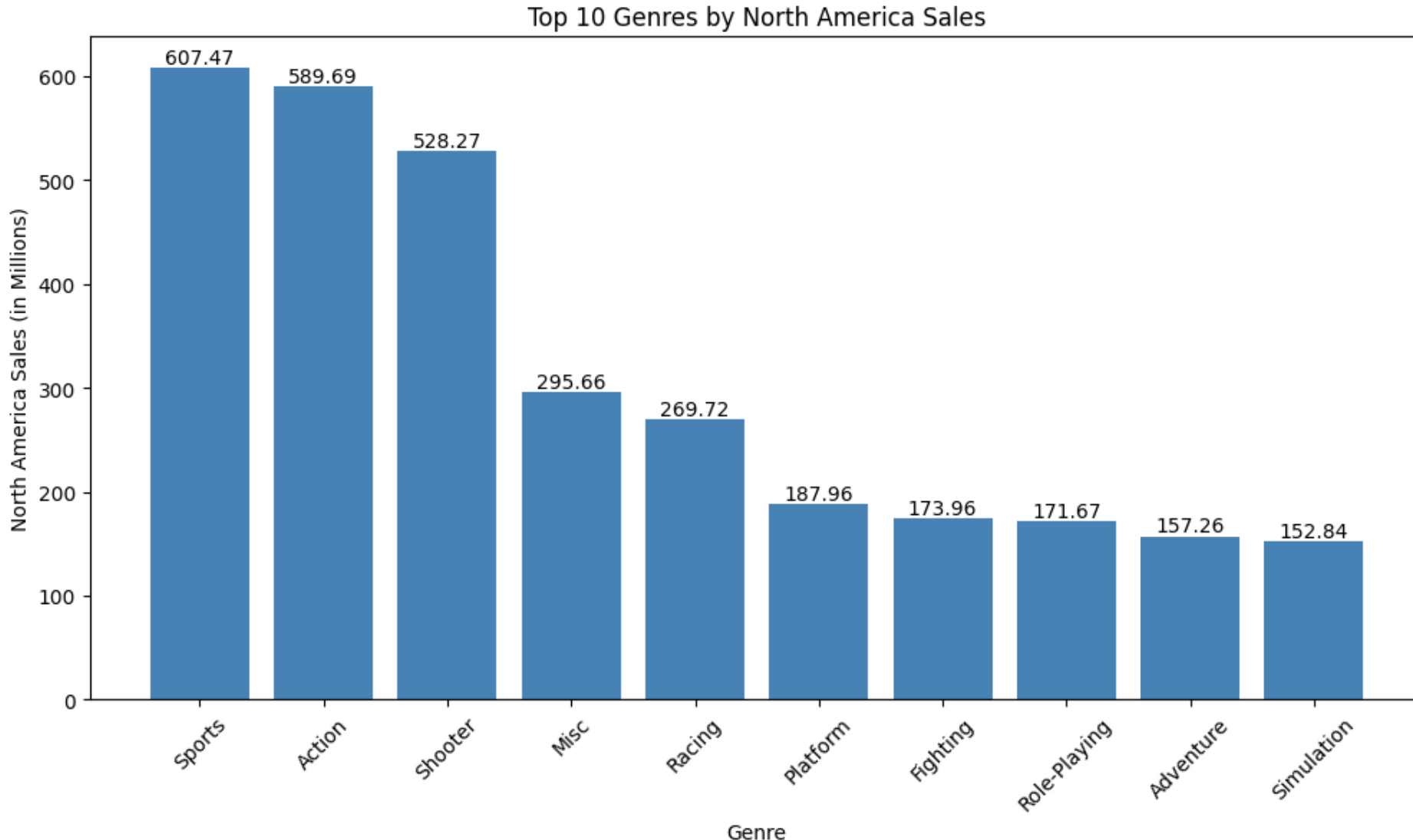
If you want to check out all the features in the dataset, please refer to [appendix slide](#)

Data (continued)

- Time Period: The data covers video game sales from as back as January 1, 1977, to as recent as January 1, 2024.
- Assumption:
 - Since the majority of game shops and platforms submit sales data for North America, we assume that the dataset represents a representative sample of video game sales in this region. This means that our study accurately reflects North American sales patterns.

Exploratory Data Analysis

Top 10 Genres



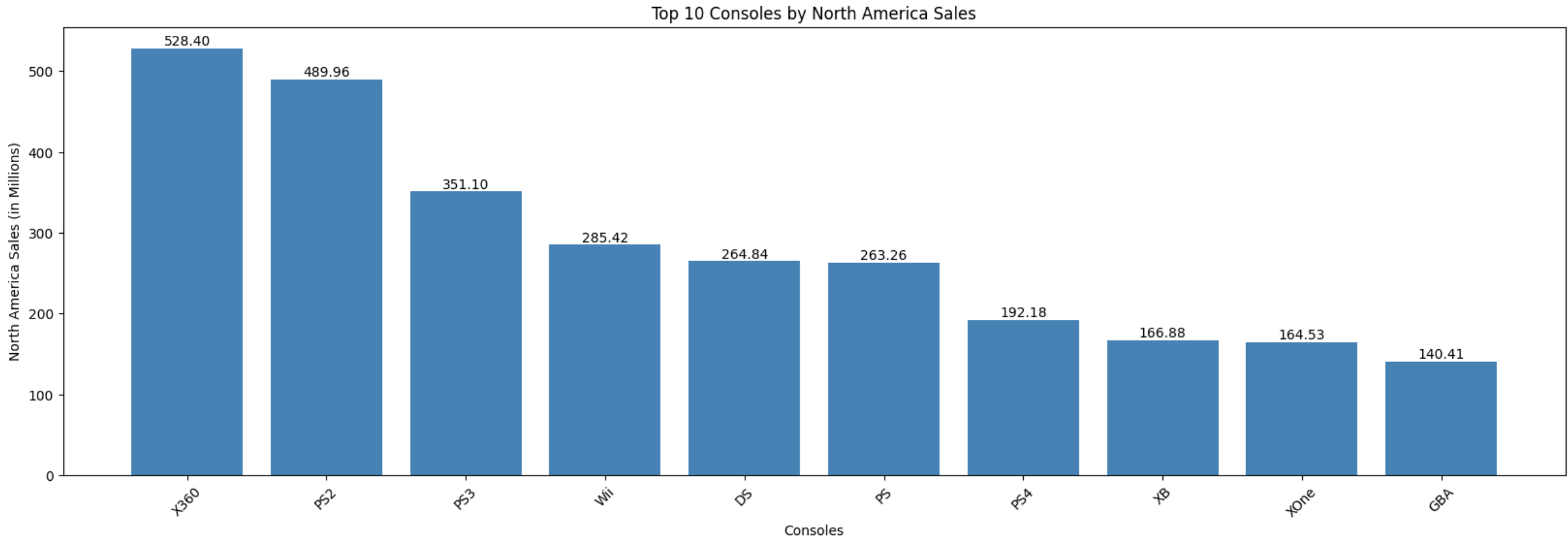
Key Points:

Sports, Action, and Shooter genres dominate North American game sales, while Racing, Platform, and other genres contribute moderately.

It's no surprise that sports-themed video games are a hit in North America. In 2023, the North American sports market made up 35%, or \$169.6 billion, of the total global sports market (source).

Note: For sales data across all genres, please refer to [slide 27](#) appendix

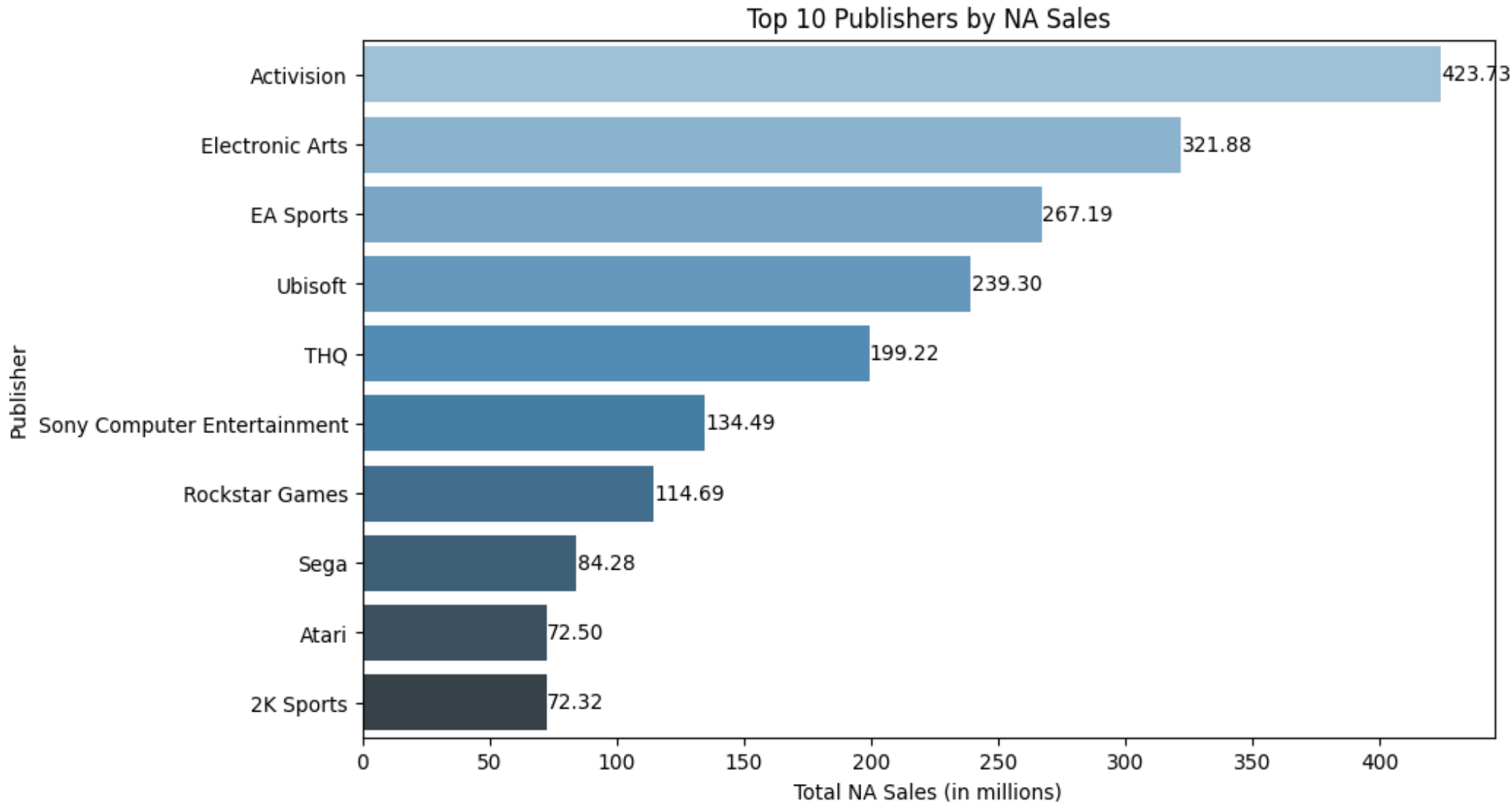
Top 10 Consoles



Key Points: The Xbox 360 and PS2 lead North American console sales, with other popular consoles like the PS3, Wii, and DS following behind.

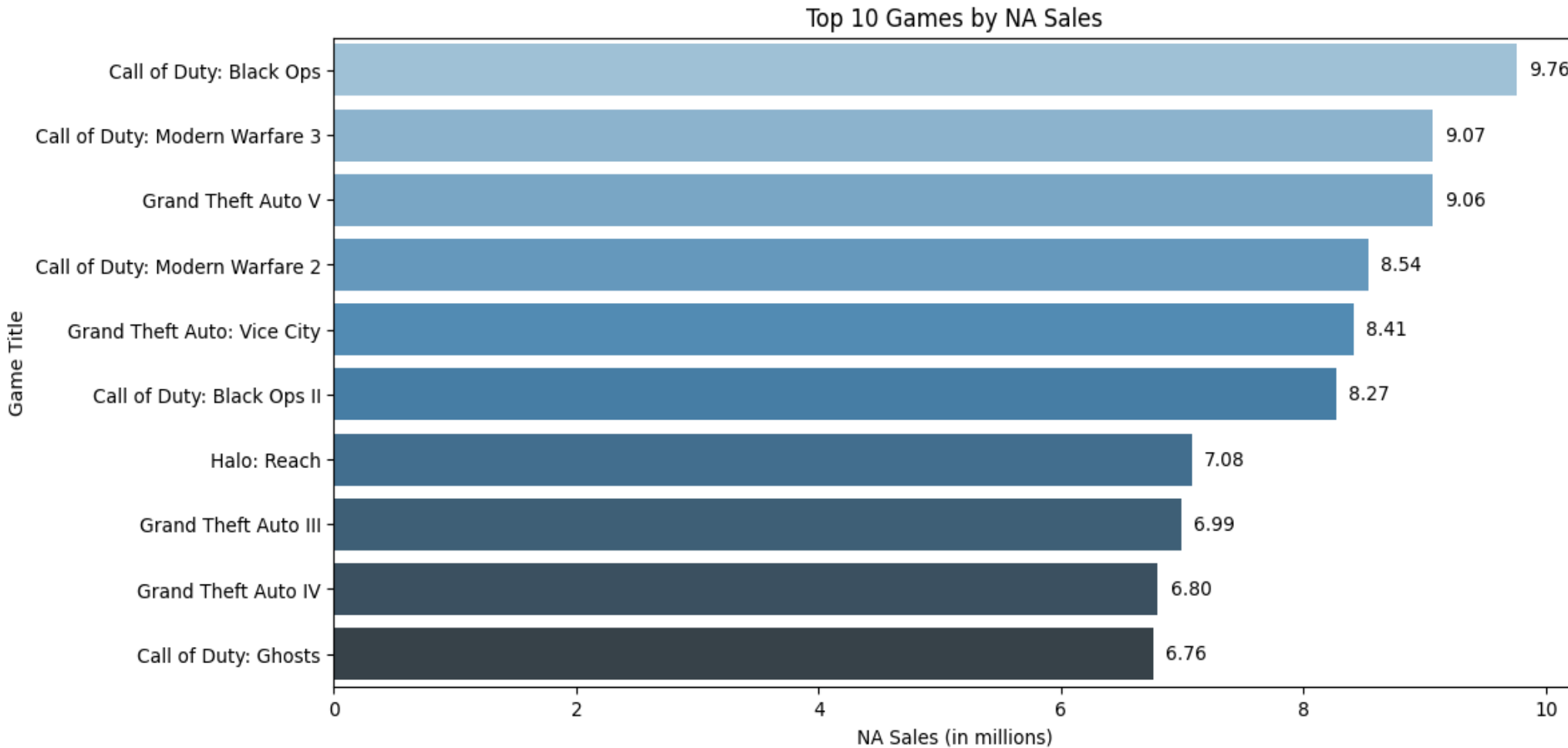
Note: For sales data across all consoles, please refer to [slide 28](#) appendix

Top 10 Publisher



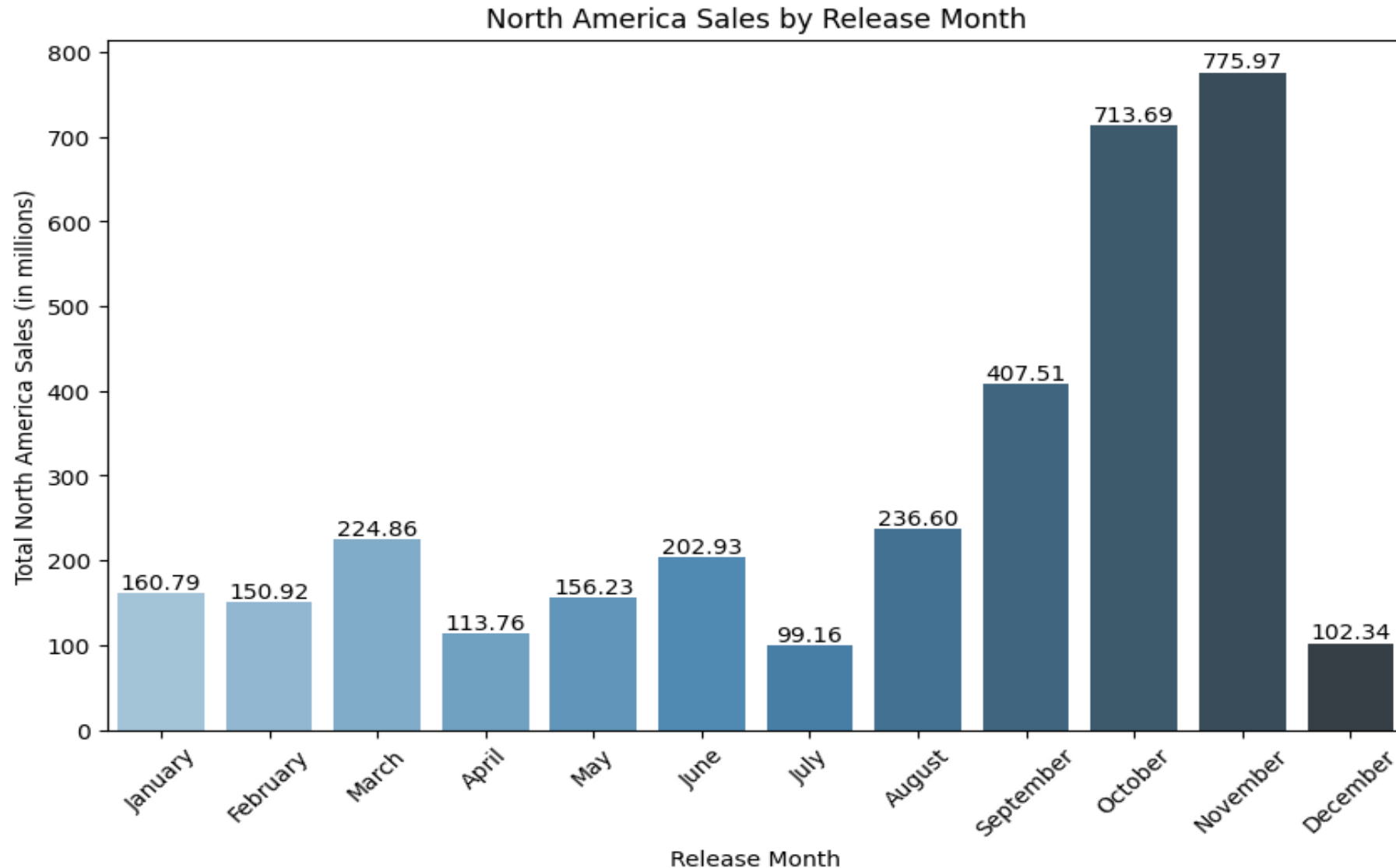
Key points:
The chart shows that Activision and Electronic Arts (EA) dominate North American video game sales, with Activision at the top (423.73 million) and EA, including EA Sports, following closely.

Top 10 Games



Key Points:
The *Call of Duty* and *Grand Theft Auto* franchises dominate the top 10 NA sales chart, with *Call of Duty: Black Ops* leading at 9.76 million units.

Sales by Month



Key Point:

This chart shows strong seasonality in NA game sales, peaking in October (713.69M) and November (775.97M) due to holiday shopping.

Sales dip in the summer, with July being the lowest (99.16M).

March and September also see moderate increases, likely due to release cycles.

Modeling methods

Modeling Methods

- Outcome Variable:
 - **North America sales:** Our main goal is to predict the sales..
- Features used:
 - **Console:** Popular consoles tend to drive higher sales.
 - Hypothesis: Games released on widely used consoles (e.g., PlayStation, Xbox) will perform better in North America.
 - **Genre:** Genre influences consumer preferences.
 - Hypothesis: Popular genres (like sports and shooters) should correlate with higher sales.
 - **Release Date:** Timing impacts demand.
 - Hypothesis: Games released during holiday seasons or major events will see a sales boost.
- These features capture key factors affecting North American game sales and provide insights for targeted marketing and release strategies.

Model Type: ARIMA

- ARIMA model was selected because it's specifically designed for forecasting, and we are trying to predict sales for North America region.
- **ARIMA** stands for:
 - AutoRegressive (AR): Looks at how previous data points influence the next one.
 - Integrated (I): Finds patterns in data changes rather than the data itself.
 - Moving Average (MA): Considers the past prediction errors to make the next prediction better.

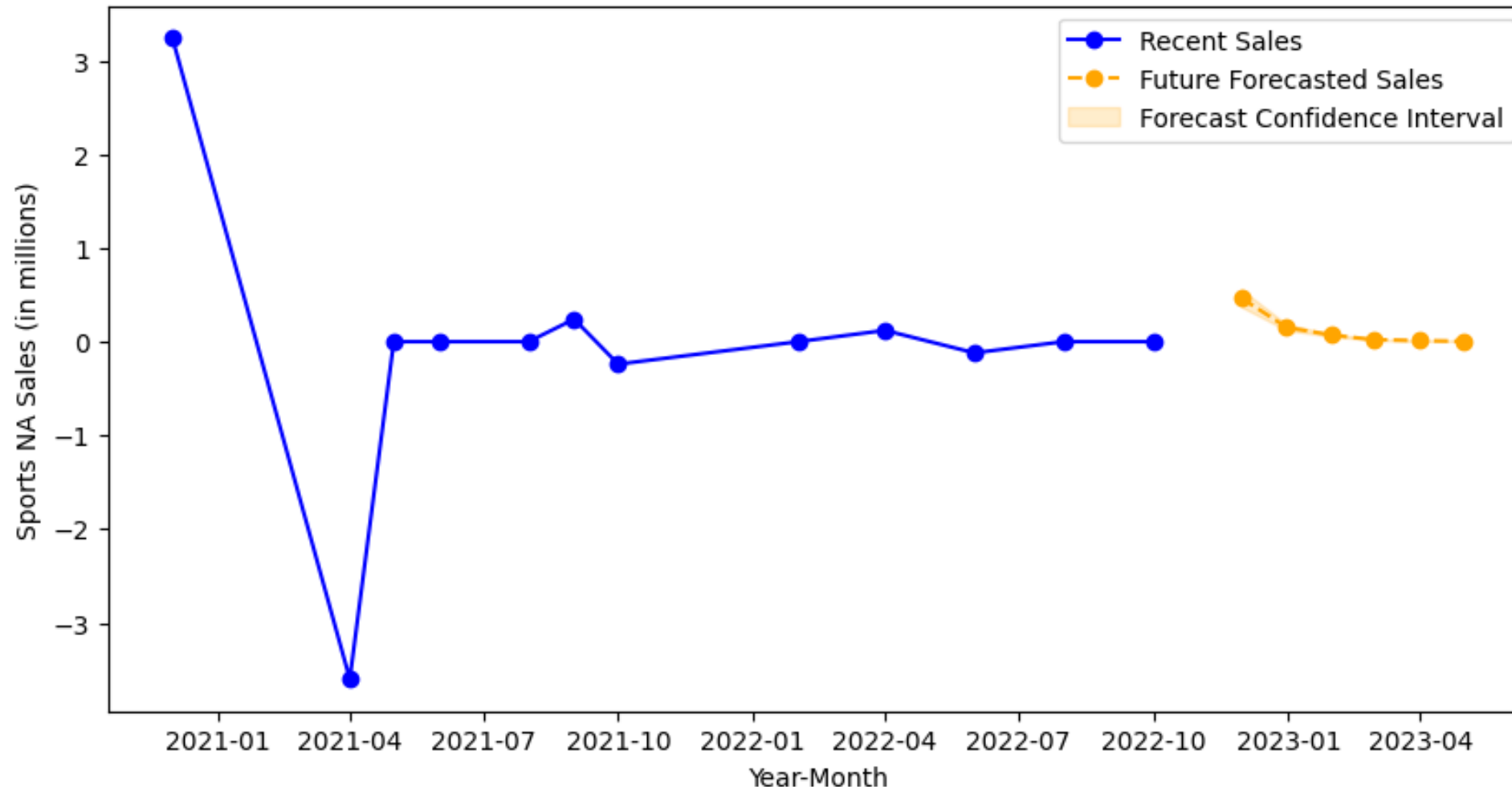
Why ARIMA?

- **Time series forecasting:** ARIMA is designed for forecasting based on time series data, like monthly sales.
- **Captures trends and seasonality:** It models both trends (e.g., increasing or decreasing sales) and seasonal patterns over time.
- **Predicts future values:** ARIMA uses historical data to forecast future values, adjusting for past behaviors and trends.
- **Handles noise and fluctuations:** The model is robust against random noise in the data, improving forecast reliability.
- **Proven effectiveness:** ARIMA is widely used and well-established for accurate forecasting in many industries.

Findings

Forecast based on specific Genre

Zoomed-In Future Forecast with Recent Sales (Sports)



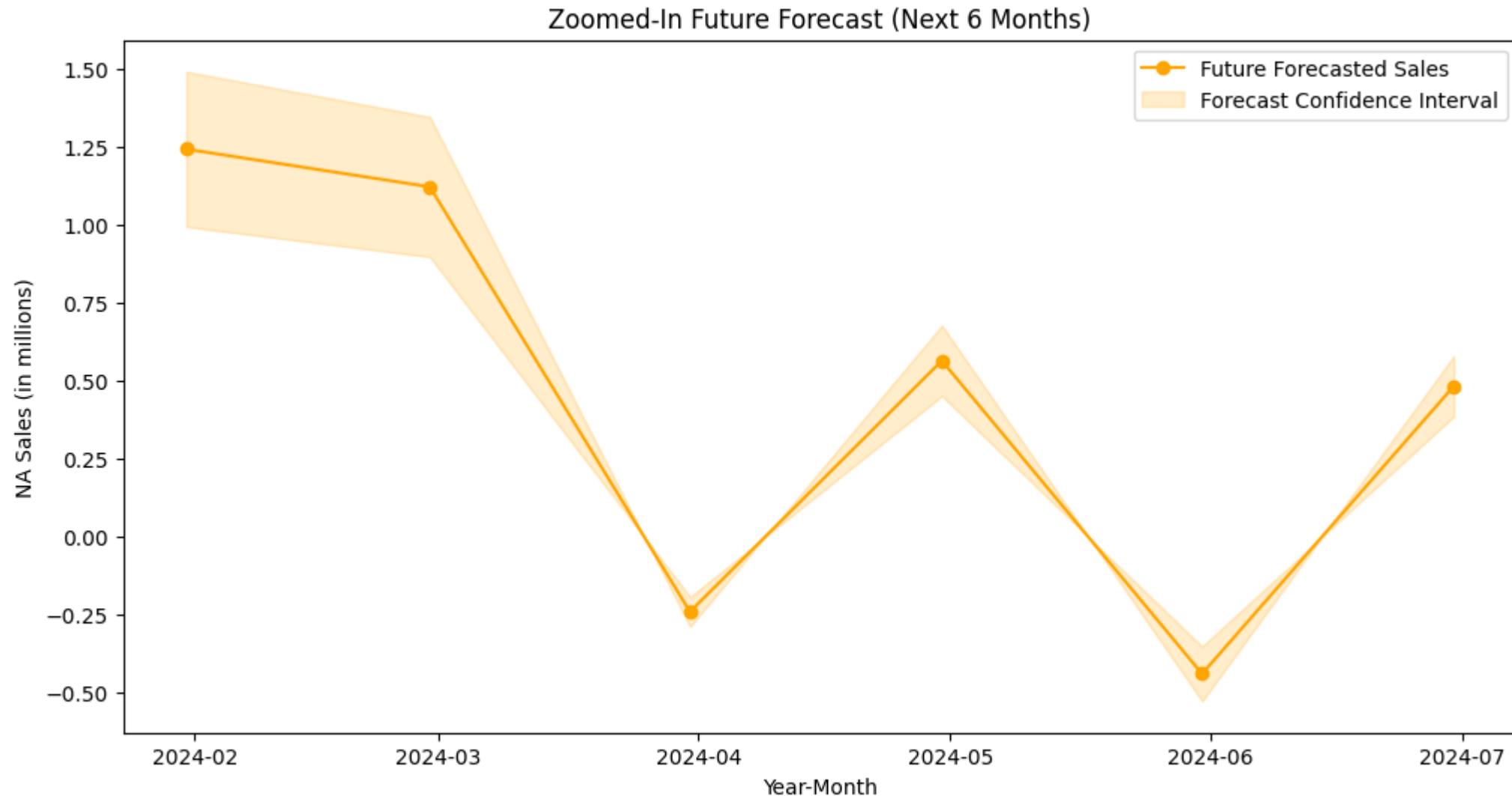
The drop in early 2021 suggests a significant decline or disruption, after which the sales stabilize but remain relatively flat.

The confidence interval around the forecast is relatively narrow, meaning the model has low uncertainty but doesn't expect significant growth in sales. This reflects limited optimism for a resurgence in demand, at least in the short term.

Forecast based on Genre (continued..)

- **Sharp Decline in Early 2021:** There is a noticeable, steep drop in sports game sales at the start of 2021. This may indicate a disruption or major change in the market, possibly due to external factors like shifts in consumer interest or industry changes.
- **Stabilization Post-Decline:** After the initial drop, sales seem to stabilize, fluctuating slightly but staying close to zero for most of 2021 and 2022. This suggests a period of recovery or steady, but low, demand for sports games in North America.
- **Predicted Gradual Decline in 2023:** The forecasted sales from early 2023 onward show a slight downward trend, suggesting that demand for sports games may continue to decrease in the near future.
- **Low Confidence in Sales Growth:** The confidence interval around the forecast is relatively narrow, meaning the model has low uncertainty but doesn't expect significant growth in sales. This reflects limited optimism for a resurgence in demand, at least in the short term.
- **Implications for the Market:** These findings could imply that the North American market for sports games may be saturated or facing competition from other genres. Publishers might need to innovate, target new demographics, or consider alternative strategies to reignite growth in this segment.

North America sale forecast



North America sales forecast (continued..)

- **Initial Sales Decline:** The forecast shows a drop in sales from February 2024 through April 2024, with sales dipping below zero by April. This suggests a possible decrease in demand for sports games in early 2024.
- **Seasonal Fluctuation:** After hitting a low in April, sales are projected to rise through July, indicating a seasonal or cyclical demand pattern where interest may pick up following a period of decline.
- **Forecast Uncertainty:** The shaded confidence interval around forecasted values shows the expected range of variation, especially around turning points (March and April). This reflects some uncertainty and potential volatility, possibly due to external factors like major sports events or new releases.
- **Need for Market Interventions:** With a projected sales dip, publishers might consider promotional campaigns or new releases around April and May to help counter the decline.
- **Recovery Signs:** The upward trend from April onward suggests a temporary low point, with potential for publishers to capitalize on renewed interest by aligning with popular sports events, like playoffs or summer sports seasons.

Recommendations & Data Science Next Steps

Seasonal Fluctuations in Sports Game Sales

- **Insight:** The forecast shows a distinct pattern where sales drop significantly around early spring (April) and then gradually increase through the summer (June and July). This could suggest a seasonal trend in sports game demand.
- **Actionable Recommendation:** Plan marketing campaigns and promotional activities strategically. For example, introduce discounts, bundles, or in-game events in March and April to counteract the expected sales decline. This could help maintain customer engagement and boost sales during traditionally low months.

High Forecast Uncertainty and Sales Volatility

- **Insight:** The forecast includes a wide confidence interval, indicating high uncertainty in future sales estimates. This suggests that sports game sales may be sensitive to external factors, like game releases or sports seasons, leading to fluctuations.
- **Actionable Recommendation:** Invest in more frequent game releases or partnerships around key sports events (such as championships or new sports seasons) to capitalize on potential sales spikes. Additionally, consider conducting further market research to understand external factors affecting demand and integrate those insights into future sales forecasting.

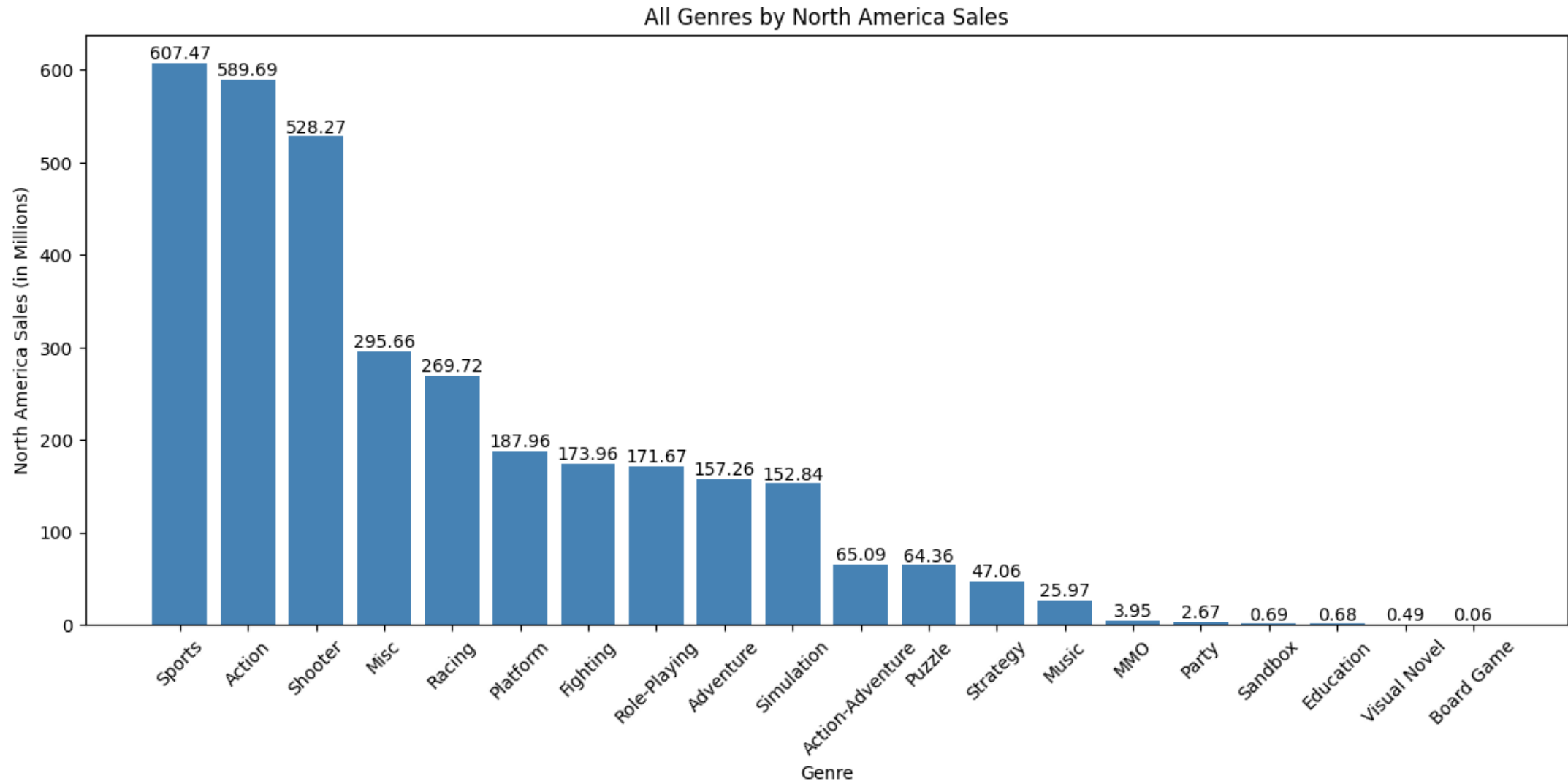
Appendix

Data Features

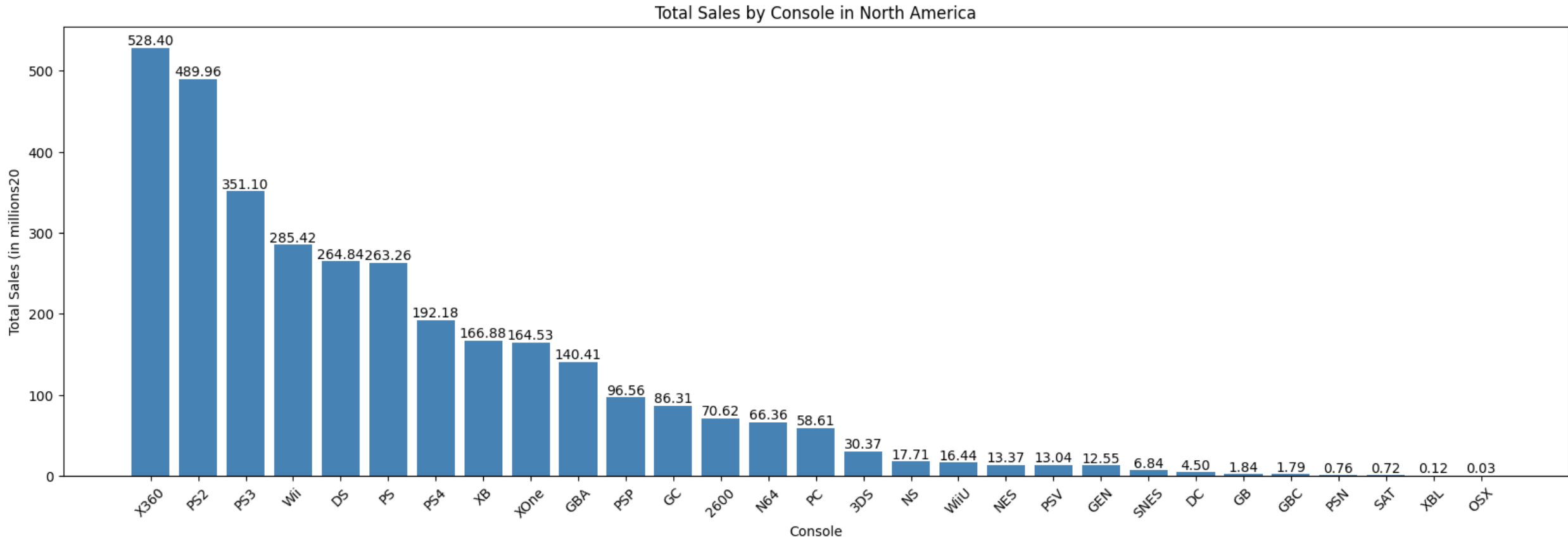
- All features 14 features in the dataset are as follows:

1. **Image:** Cover image of the game.
2. **Title:** Name of the game.
3. **Publisher:** Name of the publisher.
4. **Developer:** Name of the Developer.
5. **Console:** Console on which this game was released on.
6. **Genre:** Genre of the game.
7. **Critic score:** Metacritic score (out of 10).
8. **Total Sales:** Global sales of copies in millions.
9. **North America sales:** North American sales of copies in millions.
10. **Japan Sales:** Japan sales of copies in millions.
11. **PAL Sales:** Europe/Africa sales of copies in millions.
12. **Other Sales:** Other sales of copies in millions.
13. **Release Date:** Date of the game released.
14. **Last Updated:** Date the data was last updated.

All Genres



All Console



Note: There are total 81 unique consoles, here about 29 are only visible because all the other genres had less than 0.1 sales.