



Video Game Analytics & Prediction

Insights From Steam, RAWG, and VGSales (2000-2020)

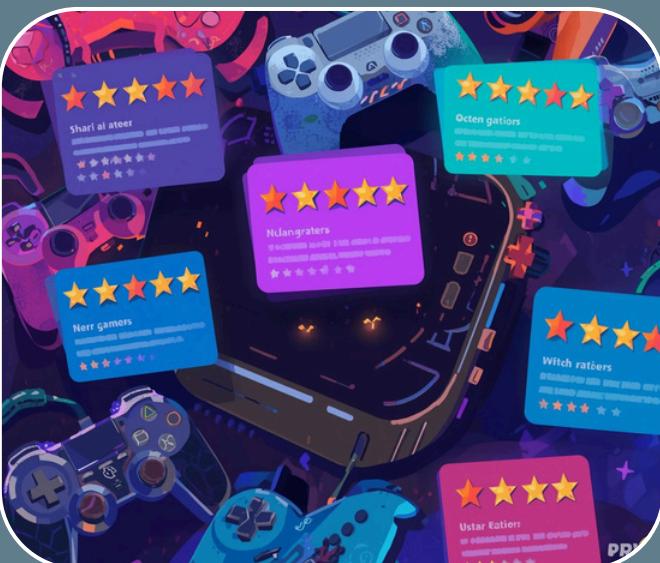
Sinan Disi | DSCI510 Final Project

Introduction:

Goal: Analyze two decades of video game data (2000-2020) to identify the features that influence positive player feedback and apply a Ridge regression model to quantify their impact



Objective #1: Investigate ratings by platforms and genre

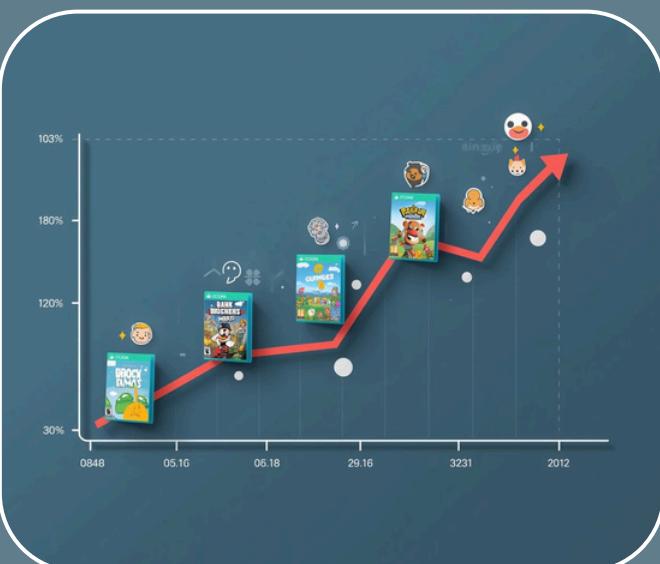


Objective #2: Examine whether game price or feedback volume effect positivity

Objective #3: Identify years with high positivity (rating)

Objective #4: Link feedback patterns to global sales

Objective #5: Build a Ridge regularized linear regression model to predict positive ratings

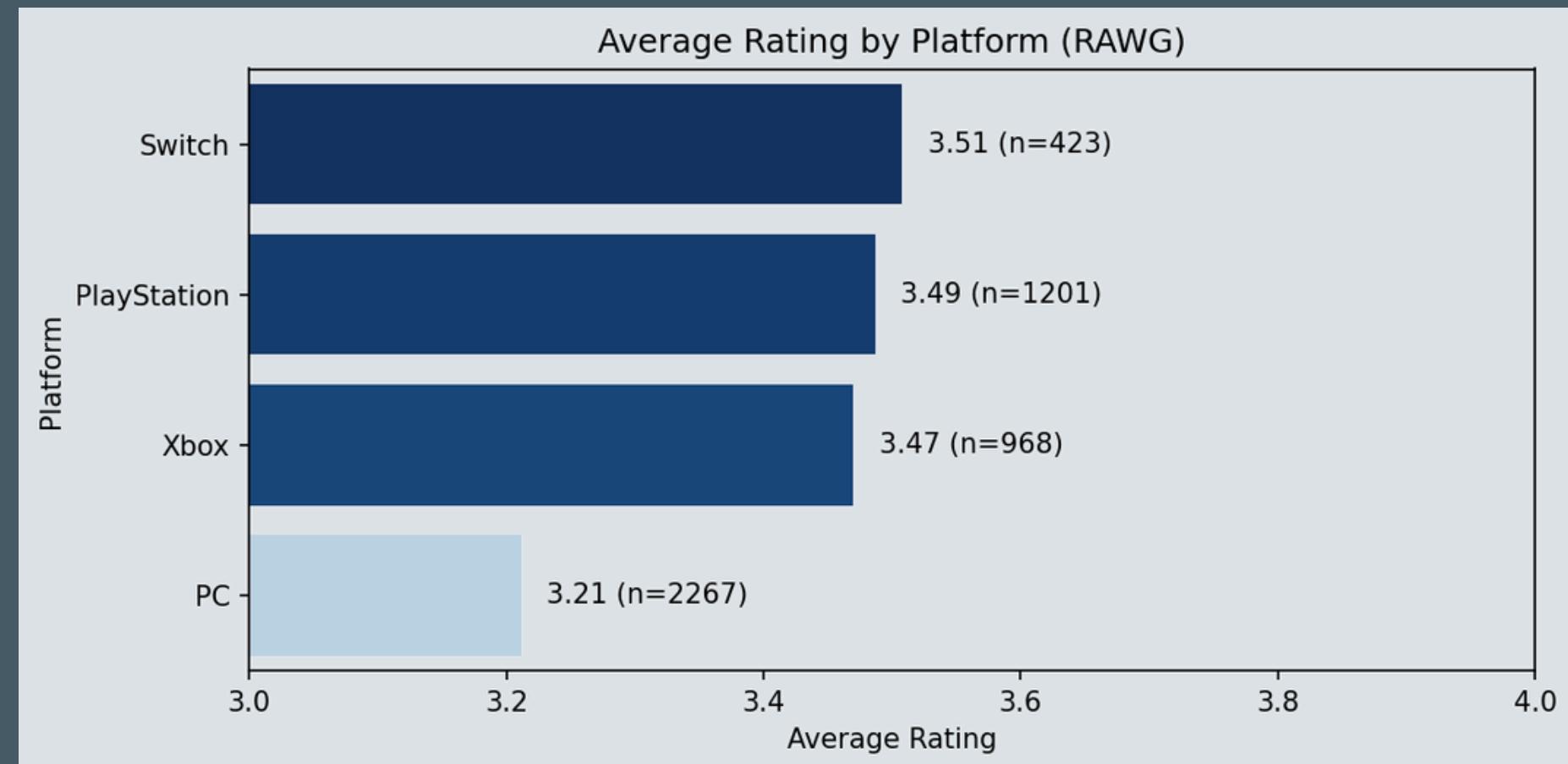
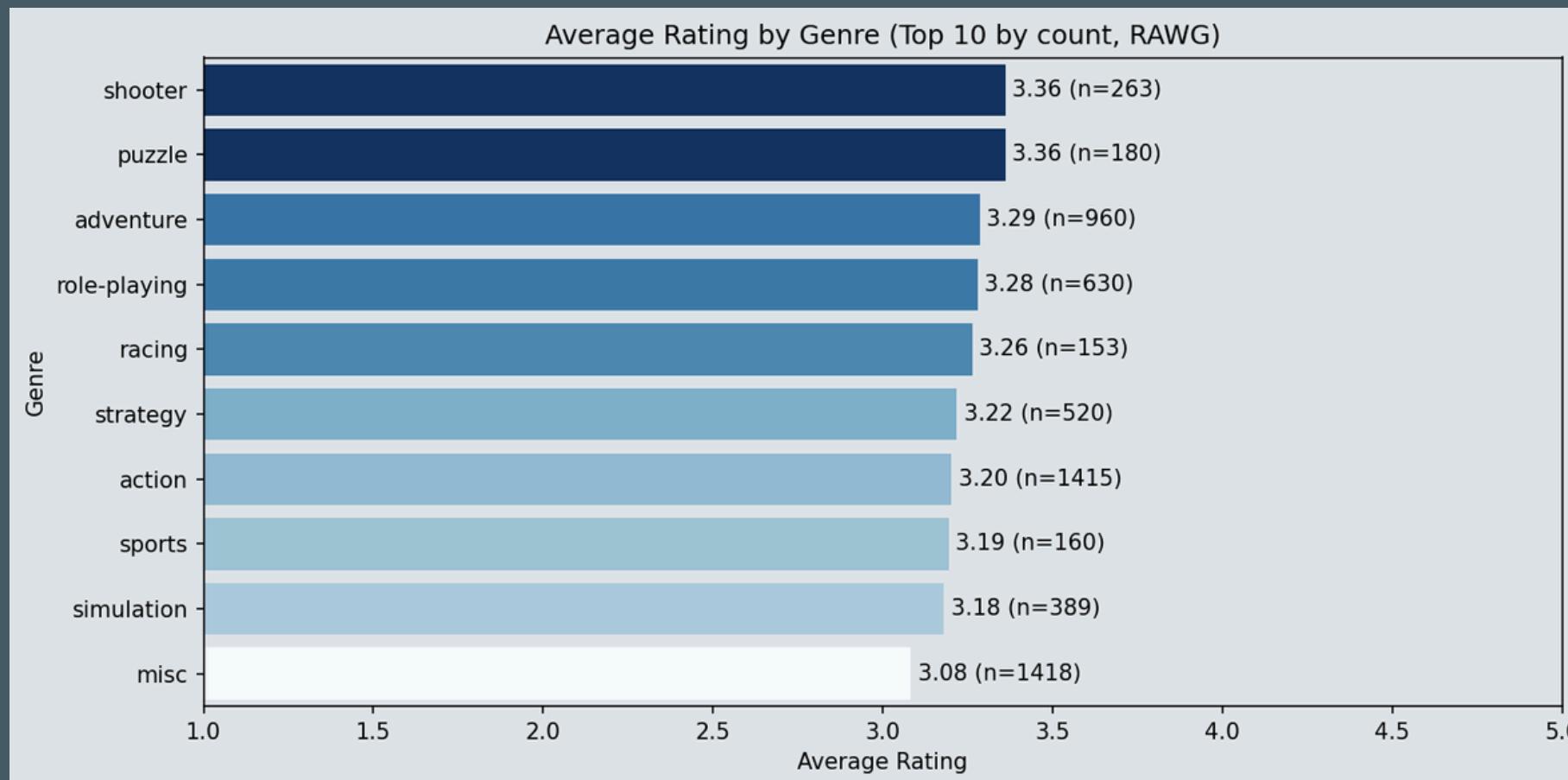


Data Sources

#	Name / Short Description	List of fields	Type	Format	Raw Data Size
1	Steam Games Dataset (Kaggle)	name, release_date, price, positive_ratings, negative_ratings, genres, median_play_time, owners	API Call	CSV	27075 (≈ 6000+ used after filtering)
2	RAWG Video Games Database	name, released date, rating, genres, platforms	API Call	JSON --> CSV	10000 (≈ 3000 used after filtering)
3	Video Game Sales Dataset (Kaggle)	Global_Sales, Genre, Name, Platform, Year_of_Release, Platform	API Call	CSV	16587 (≈ 9000 used after filtering)

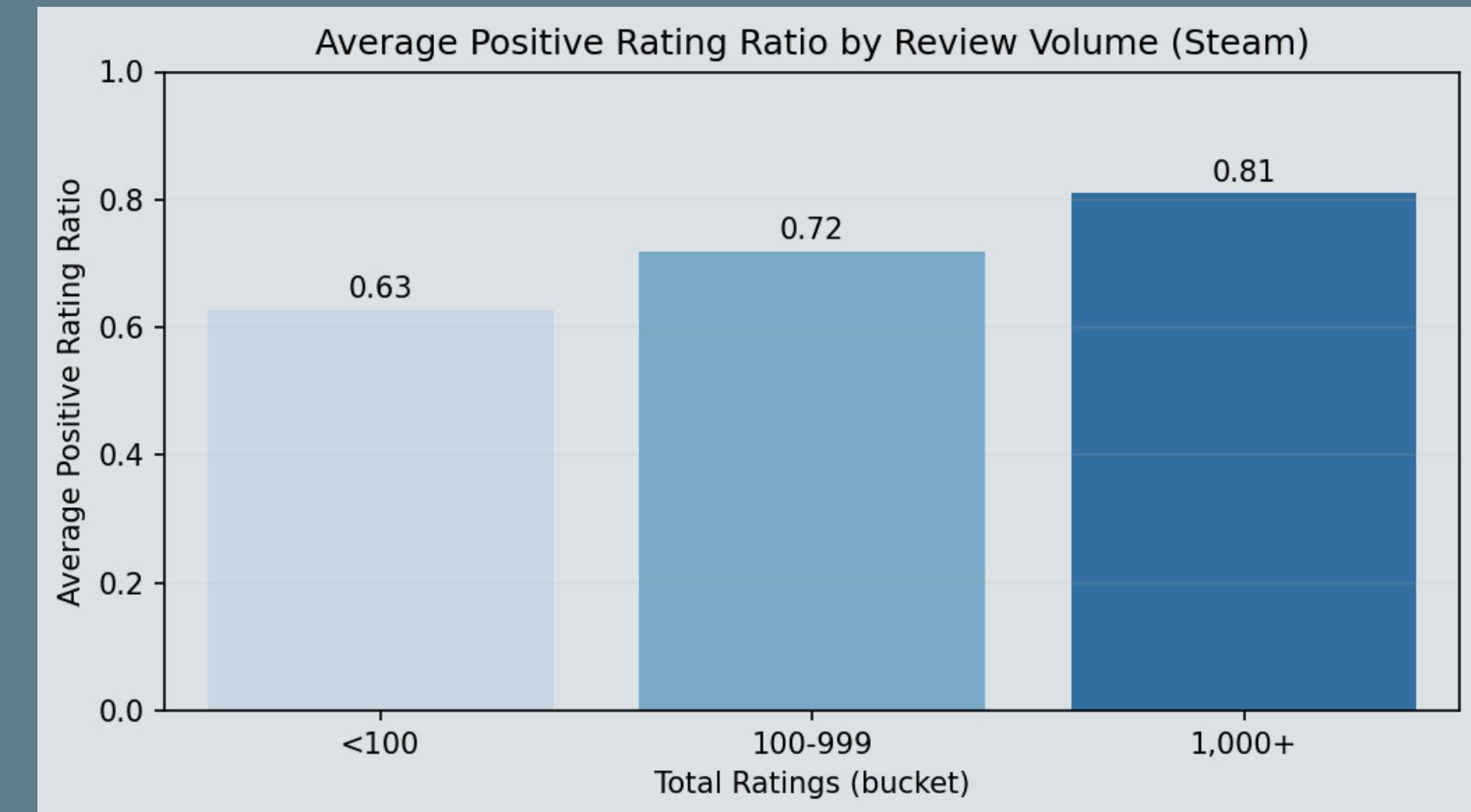
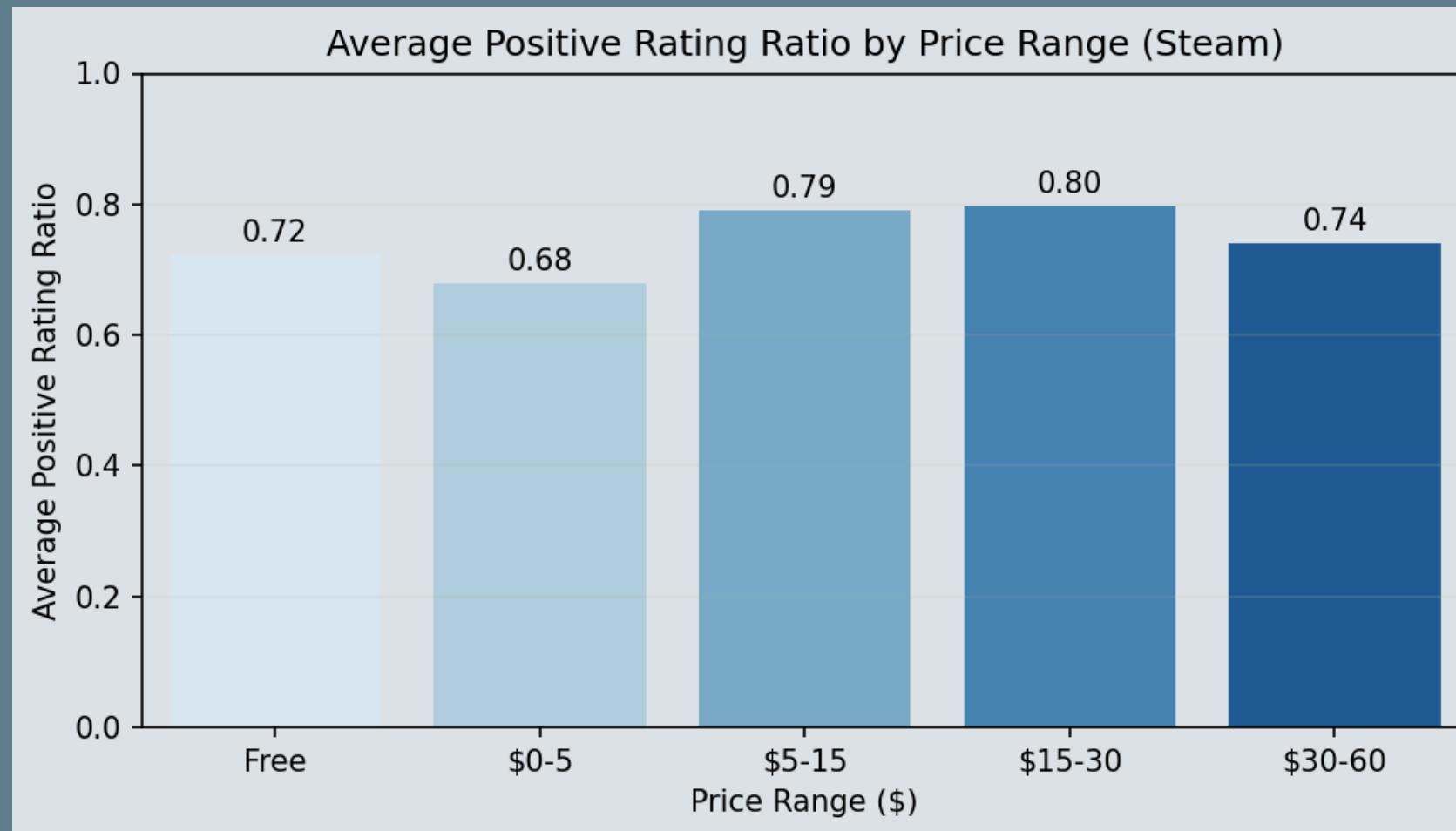
Approach: Collected data via APIs into CSV files, then cleaned and filtered all datasets (2000-2020) by removing non-gaming titles and entries with missing key fields. Consolidated and standardized platforms and genres, and performed binning and bucketing to analyze game feedback volume and price.

How Platform and Genre Influence Game Rating



- Shooter & Puzzle genres rate the highest.
- Simulation and Misc have the lowest genre ratings.
- PC has the lowest platform ratings.
- Switch, PlayStation, and Xbox have similar ratings

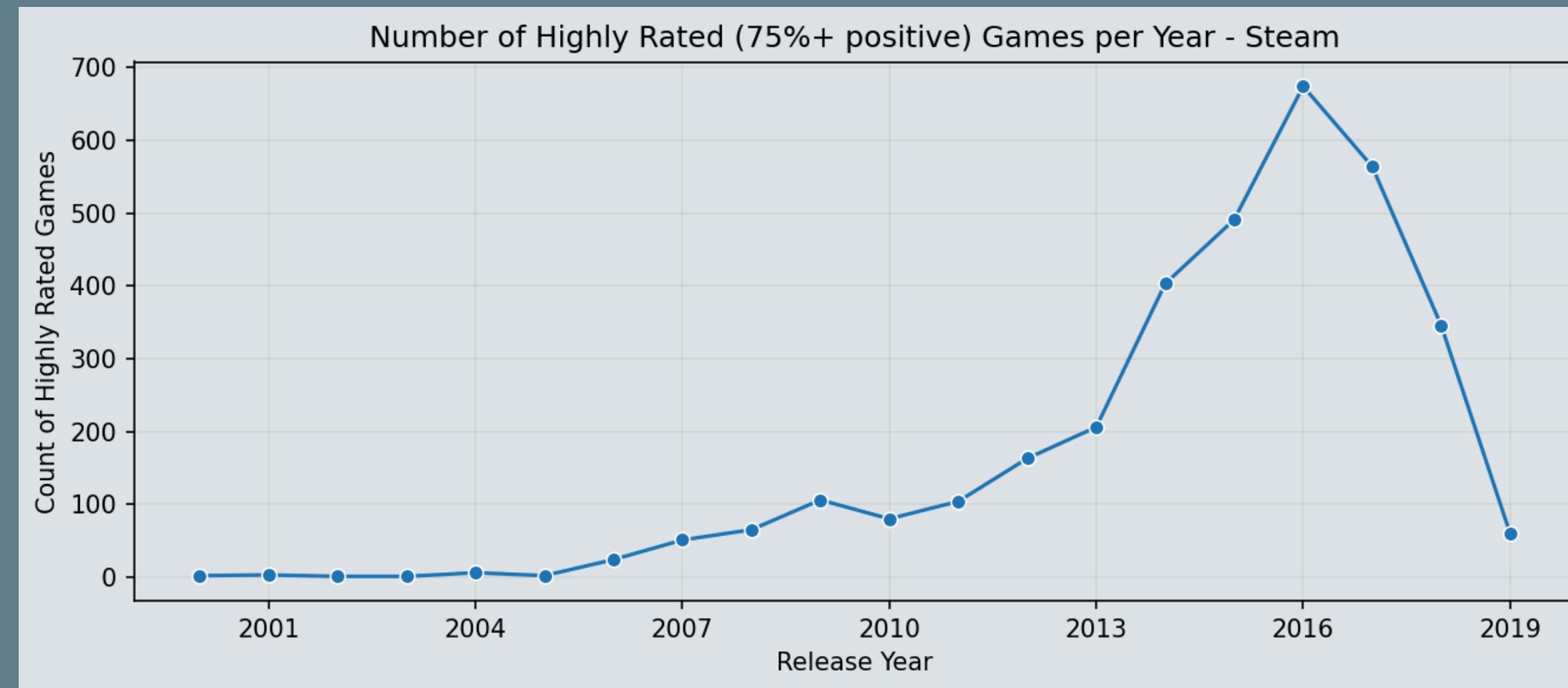
How Rating Volume and Price Affect Feedback Positivity



- Free games are not rated the highest
- Mid-priced games (\$5-30) have the best positivity
- Very cheap games (\$0-5) have the lowest positivity
- Note: Correlation between price and average positive rating ratio: 0.189 (n=6,073)

- Games with more reviews tend to have higher positivity
- Titles with 1,000+ reviews show the strongest positive feedback

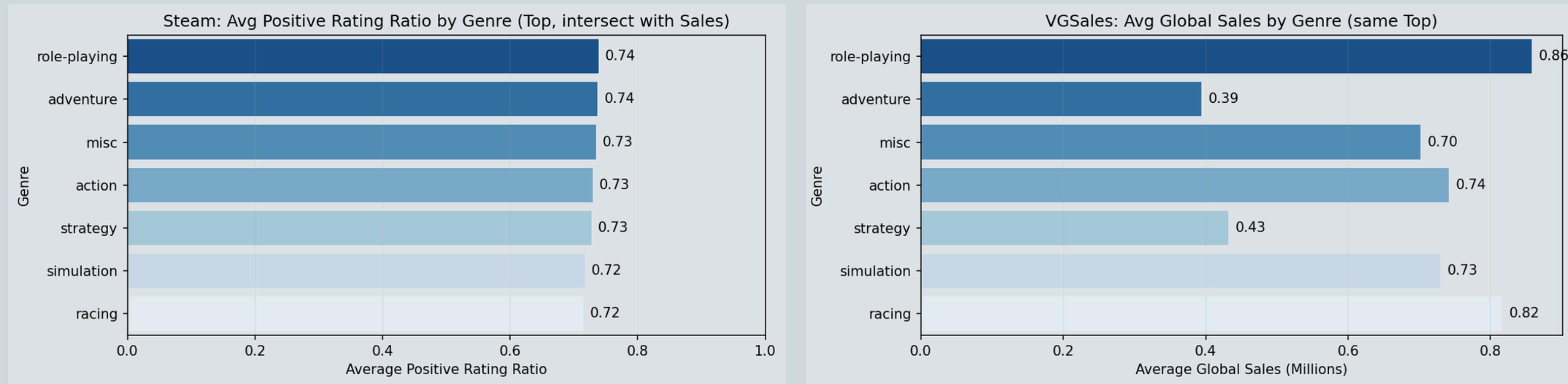
How Release Timing Influences Positive Ratings



- Only games with 75%+ positive ratings
- The volume of highly rated games stays low between 2000-2007
- It rises gradually between 2008-2012
- The peak is in 2016

Note: Steam's massive expansion between 2012-2016 aligns with the volume of highly rated games. The decline after 2016 may indicate a market plateau.

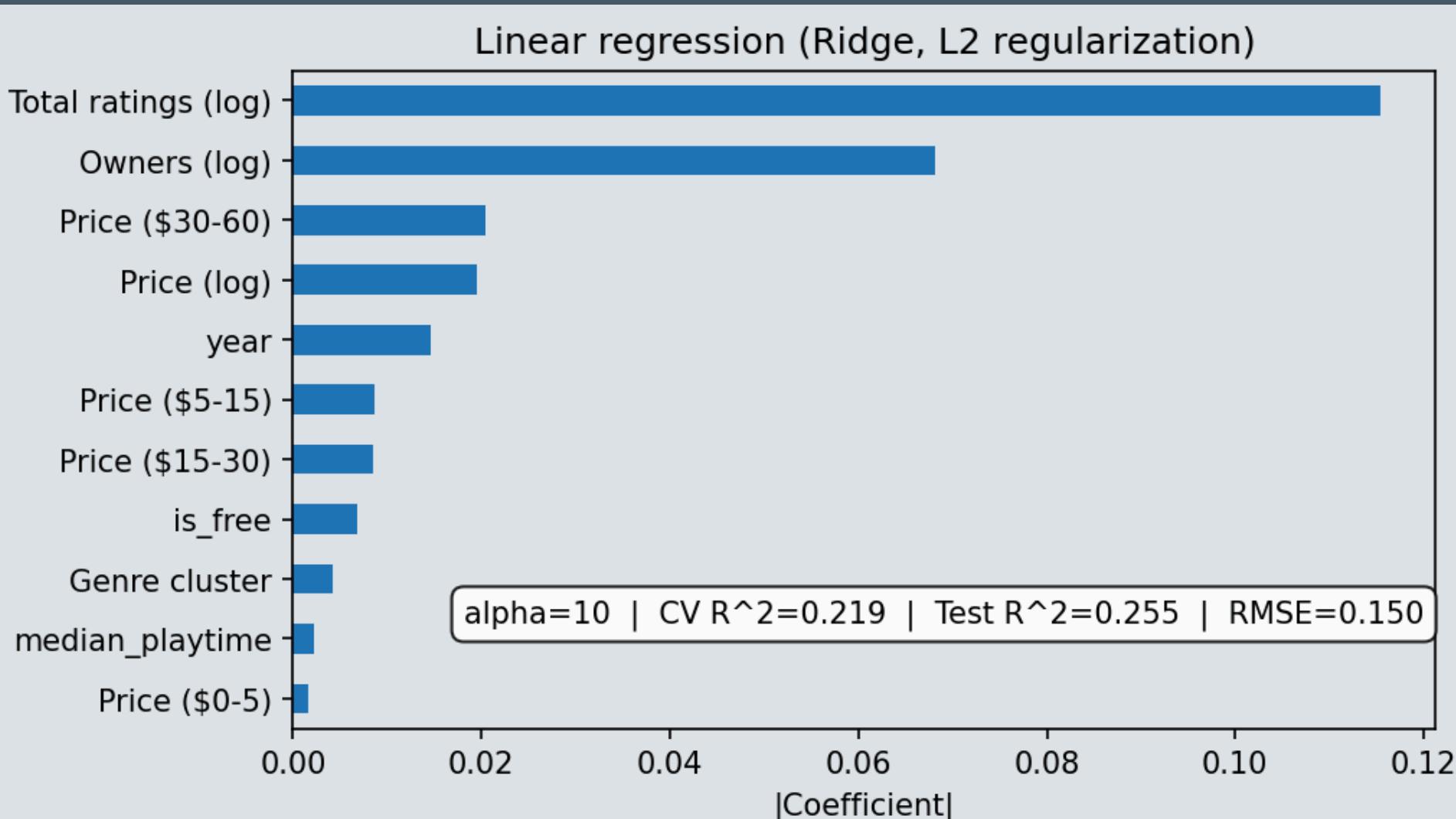
How Global Sales Relate to Positive Ratings



- Genres with similar positive rating ratios ($\approx 0.72\text{--}0.74$) show very different global sales outcomes
- RPG titles achieve high average global sales, while Adventure titles remain much lower despite similar ratings
- Overall relationship is weak and slightly negative ($\text{corr} \approx -0.32$)
- Conclusion: High player ratings do not necessarily translate into higher global sales

Ridge regularized linear regression model

- Model tuning: A Ridge penalty of $\alpha = 10$ was selected via 5-fold cross-validation, achieving a balanced level of L2 regularization that keeps coefficients in check and limits overfitting
- Cross-validation fit: CV $R^2 \approx 0.22$ indicates the model explains about 22% of the variance in positive feedback % on unseen folds, indicating a genuine yet modest signal
- Test performance: Test $R^2 \approx 0.26$ is close to the CV value, suggesting stable and limited overfitting to the training data
- Error scale: RMSE ≈ 0.15 on a target in $[0, 1]$ implies that the model is usually off by about 0.15, which adds some noise
- Total ratings and owners are the most influential features; playtime or genre had minimal influence



- Note #1: Ridge handles connected features and keeps the model stable.
- Note #2: L2 keeps all inputs, just softens their influence.
- Note #3: Standardization puts features on the same scale for evaluation.

Challenges

- Cleaning and filtering large datasets
- Managing missing or low-quality Steam data (many games with 0 playtime or no feedback)
- Modeling a noisy target variable (positive rating ratio)
- Applying log scaling and building game-price bins and feedback-volume buckets to stabilize skewness
- Handling inconsistent genre labels: standardized genres (canonical labels → 12 unified labels → consolidated to 5 clusters for the prediction model).
Note: Steam and VGSales only had 8 overlapping genres, so cross-dataset analysis was limited to that



Thank You!

