STEAM GAMES DATA ANALYSIS II

MSBA 503: ANALYTICS PROGRAMMING II

CARLOS VELAZQUEZ | WILLIAM KILDAY

KV

ANALYTICS

Problem Description

Steam, a digital platform for video games, has provided data on **user-generated reviews** & seeks to explore how this data can be utilized to gain deeper insights.

Is it possible to analyze our customers' emotions toward the games developed by our top 5 developers?

Are there alternative methods for collecting data on genres and categories?

How feasible is it to predict whether a customer is likely to purchase a game, and what factors drive those decisions?

What methods can be explored to group reviews based on common themes to better understand user feedback and preferences?

Scope of Work

Objective

The objective is to extract actionable insights from Steam's data by leveraging advanced techniques such as predictive analytics, natural language processing, and clustering. To achieve this, Steam has engaged Kilday & Velazquez Analytics to analyze user-generated reviews, predict customer behavior, identify key purchase drivers, and group reviews based on common themes to inform marketing, development, and data collection strategies.

Project Scope

1. Descriptive Analysis on Sentiment Scores of Developers:
Kilday & Velazquez Analytics will conduct a detailed sentiment analysis of user-generated reviews to evaluate how players feel about various developers. This will include calculating sentiment scores, identifying trends, and analyzing patterns to provide a clear understanding of developer reputation across the platform.

2. Web Scraping Method for Data Collection:
A robust web scraping methodology will be implemented to gather game metadata. This process will ensure the availability of high-quality data to support downstream analyses and expand Steam's data collection capabilities.

3. Neural Network and KNN Classification for Purchase Prediction:
Predictive models using neural networks and K-Nearest Neighbors (KNN) will be developed to determine the likelihood of a customer purchasing a game. These models will identify patterns in user behavior, game features, and review content, helping Steam predict purchases with greater accuracy.

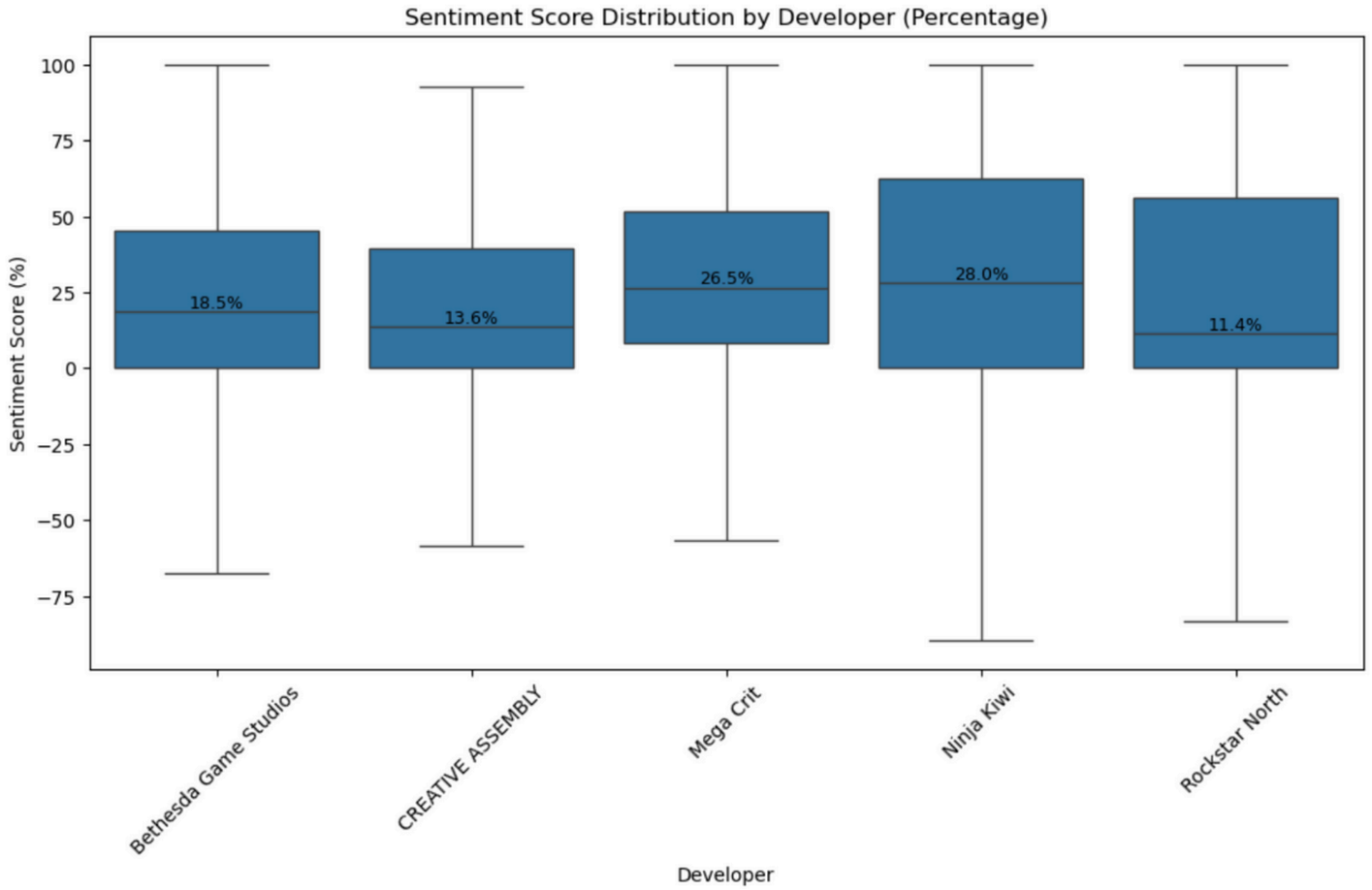
4. Decision Tree & Feature Importance Analysis to Understand Variables Influencing Purchases:
A decision tree model will be constructed to analyze and interpret the variables that most significantly influence purchase behavior. This will provide clear and actionable insights into the factors driving customer decisions, enabling Steam to refine its strategies.

5. K-Means Clustering for Grouping Reviews:
K-Means clustering will be applied to group reviews based on textual content. This analysis will uncover common themes and topics within reviews, such as gameplay highlights, user complaints, or unmet needs. These insights will inform development priorities and marketing campaigns.

AGENDA

- SENTIMENT SCORE
- WEB SCRAPING
- NEURAL NETWORK & KNN CLASSIFICATION
- DECISION TREE & RANDOM FOREST
- CLUSTER MODEL

Sentiment Score Across Top 5 Developers



Insights

Highest Median Score:
Ninja Kiwi (28%)

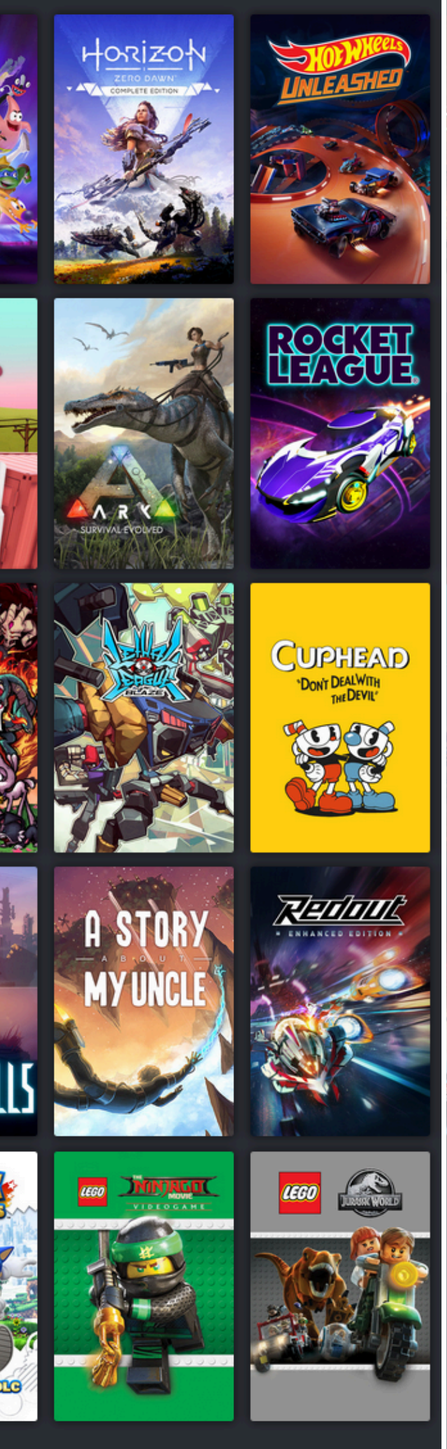
Consistent:
Mega Crit

Lowest Median Score:
Rockstar North (11.4%)

Greater Variability:
Rockstar North

Action Items

- Developer Performance Evaluations
- Target Improvement: Benchmark
- Resource Allocation



Understanding Game Popularity and Longevity

Why are some games so successful?

→ Explore New Data to Find the Most Successful Games Today and of All Time

→ Scrape the Data From Various Sources

→ Combine New Data Against the Entire Data Set

→ Compare Common Features of Successful Games to the Rest

Web Scraping Data

- Scraping the Top Steam Games Today and the Best Selling Steam Games of All Time (minimum 1 million copies sold)
- Creating a group that matches either list and comparing it to the rest of the games

Top games by current player count

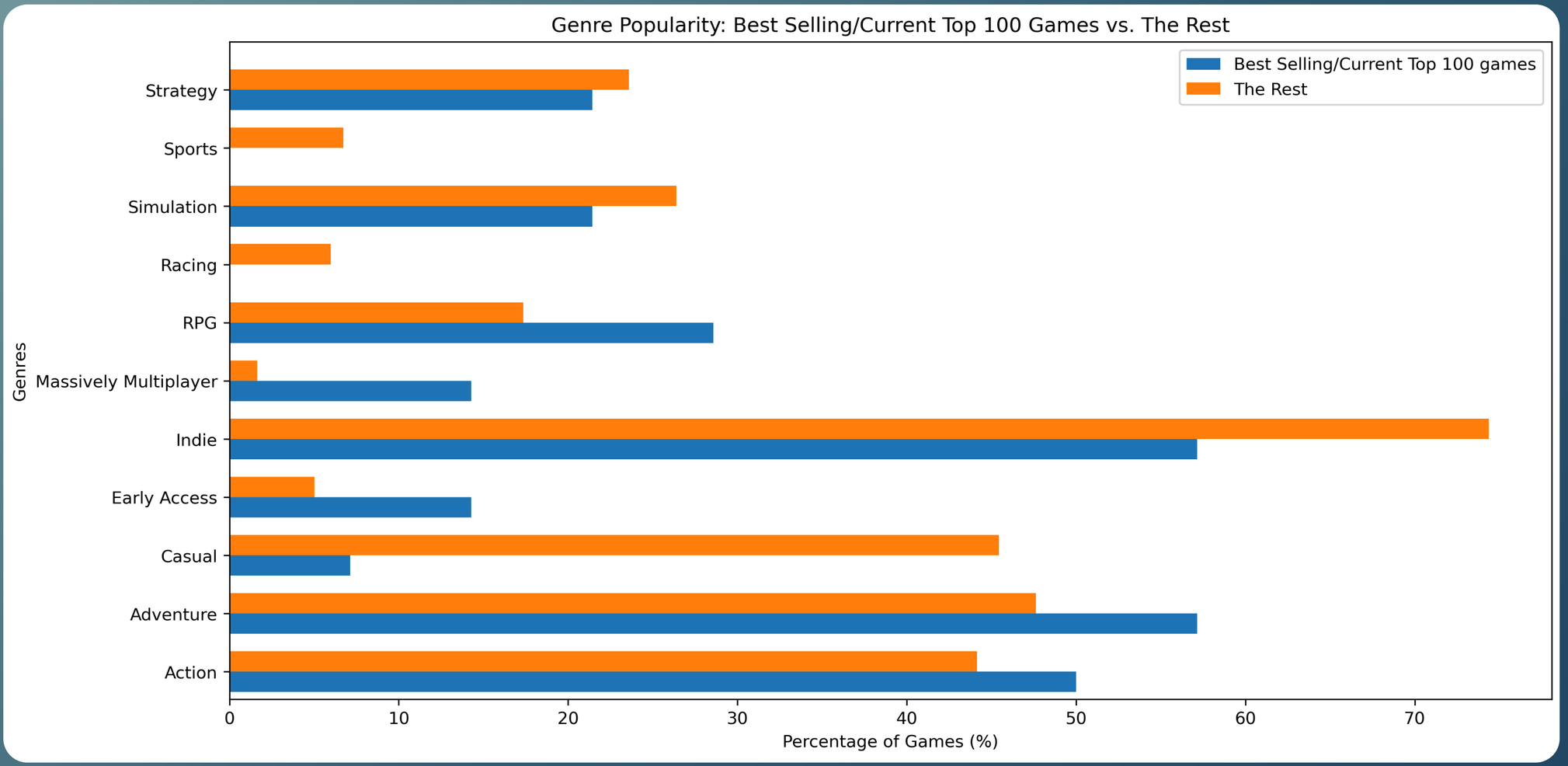
CURRENT PLAYERS	PEAK TODAY	GAME
661,362	1,460,230	Counter-Strike 2
401,323	460,883	Marvel Rivals
291,065	618,358	Dota 2
266,741	498,966	Path of Exile 2
212,667	750,493	PUBG: BATTLEGROUNDS
115,342	123,268	Banana
112,698	298,926	NARAKA: BLADEPOINT
99,245	122,668	Call of Duty®
95,226	192,997	Source SDK Base 2007
77,794	161,127	Grand Theft Auto V
71,083	118,350	Wallpaper Engine
70,370	144,786	Rust
57,719	96,300	Stardew Valley
57,018	144,354	Apex Legends

https://store.steampowered.com/stats/stats/

Game	Total copies sold	Series
PUBG: Battlegrounds	42 million[1]	*PUBG Universe*
Minecraft	33 million[2]	*Minecraft*
Terraria	32 million[5]	—
Diablo III	20 million[6][b]	*Diablo*

https://en.wikipedia.org/wiki/List_of_best-selling_PC_games

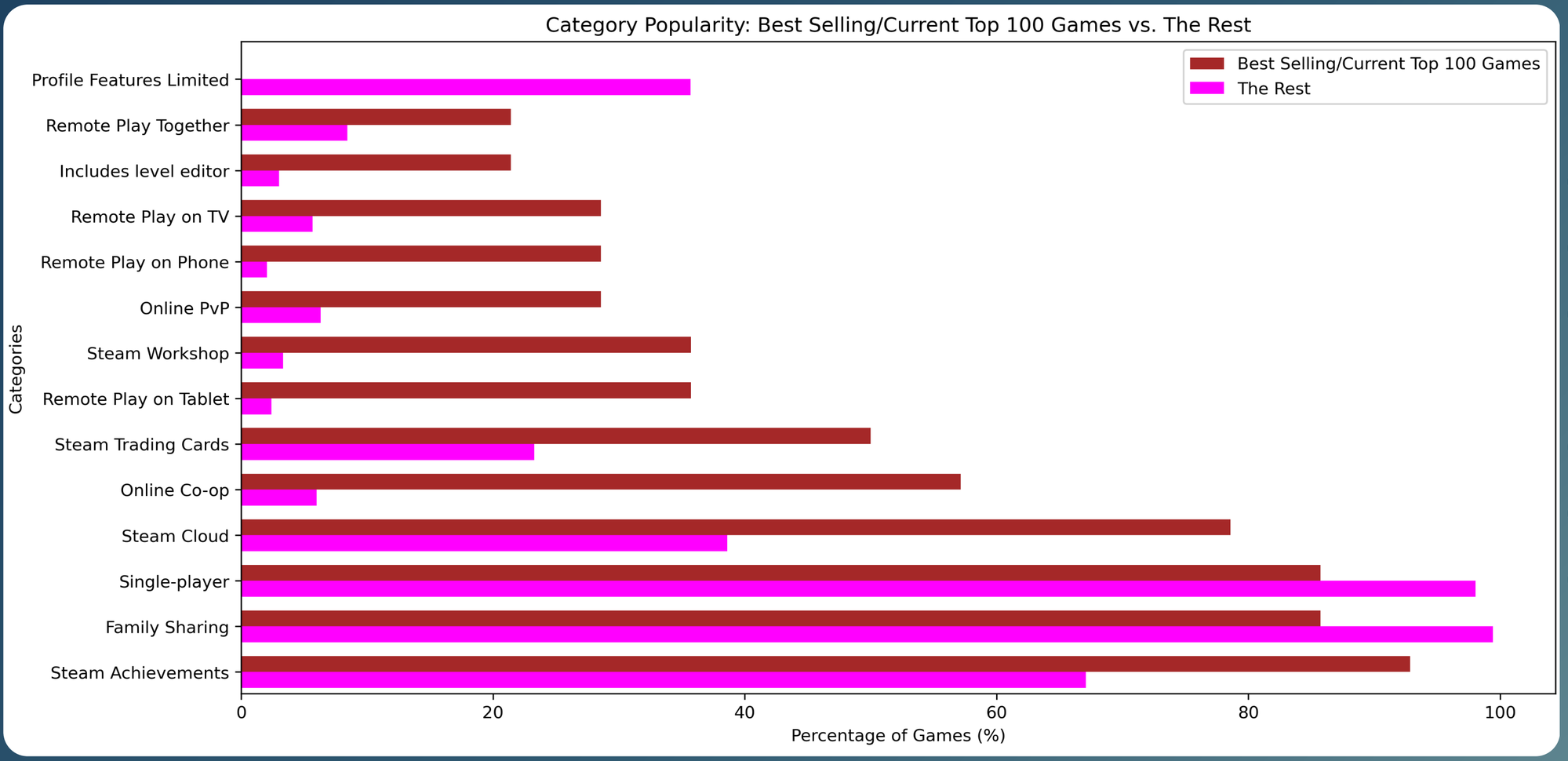
Comparing Genres: Best Selling/Current Top 100 Games vs. The Rest



Genres linked to success

- RPG
- Massive Multiplayer
- Adventure
- Action
- Early Access

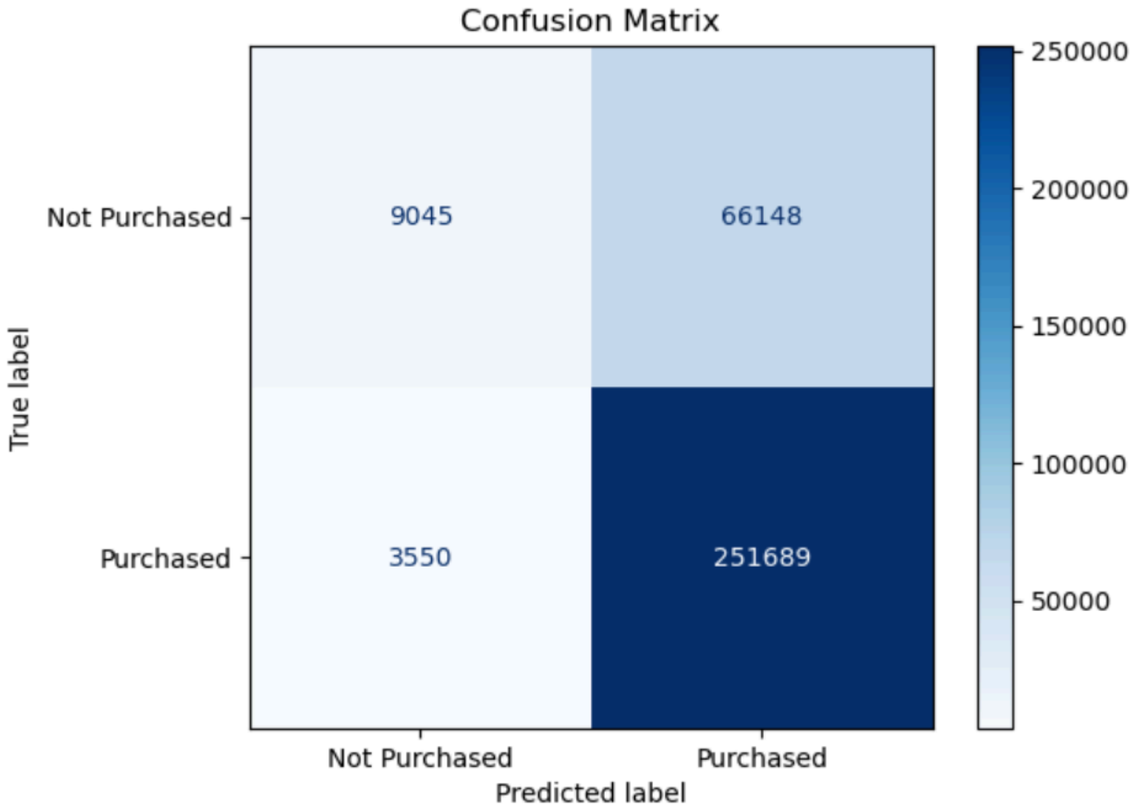
Comparing Categories: Best Selling/Current Top 100 Games vs. The Rest



Categories linked to success

- Online
- Steam Workshop
- Remote Play on Tablet
- Online Co-op
- Steam Cloud
- Steam Acheivements

Neural Network Classification



Test Accuracy: 78.9%

How feasible is it to predict whether a customer is likely to purchase a game, & what factors drive those decisions?

Matrix helps evaluate whether the model effectively uses the available data to predict purchases

Test Accuracy:
78.9%

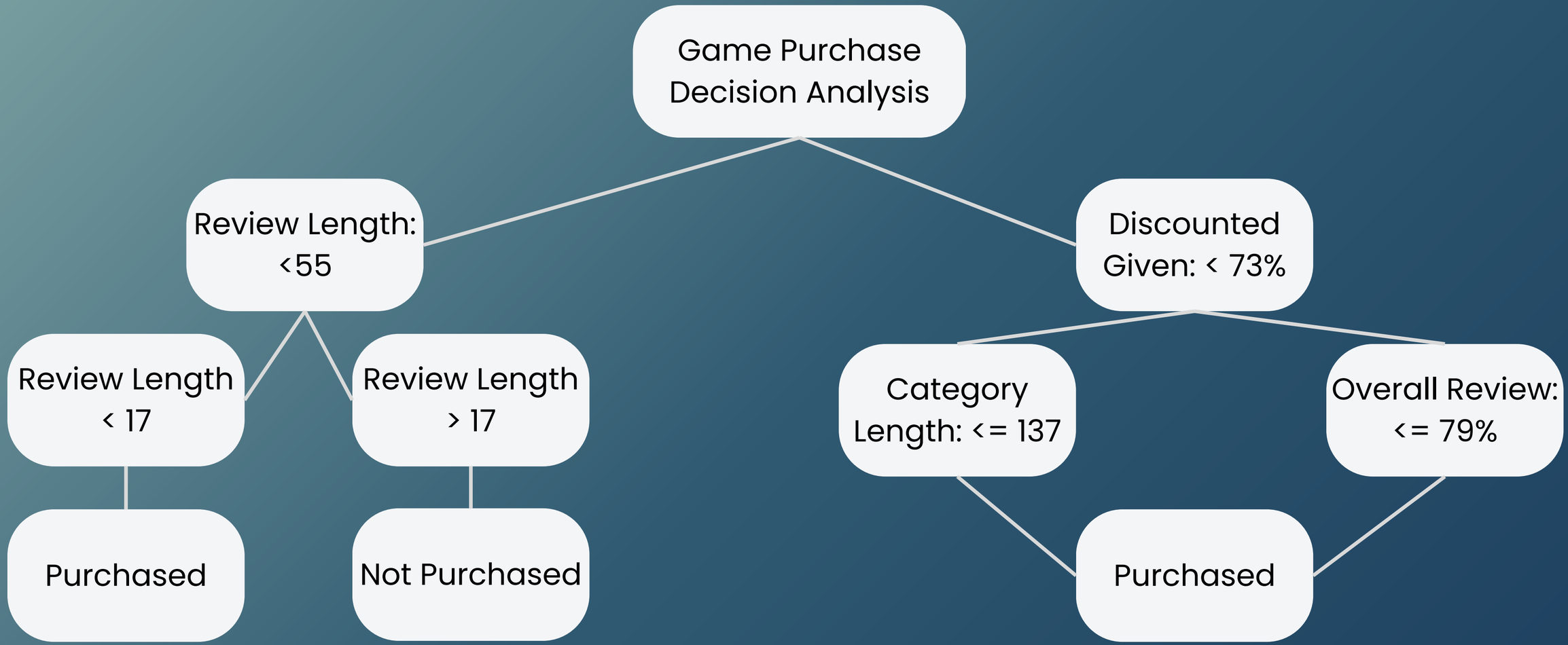
Correctly Predict Purchases:
251,689

Classification Report:

	precision	recall	f1-score
0	0.42	0.19	0.26
1	0.79	0.92	0.85
accuracy			0.76
macro avg	0.61	0.56	0.56
weighted avg	0.71	0.76	0.72

Score	Description
76%	Model Correctly predicted 76% of the total cases
92%	Model correctly identified 92% of all actual purchases
26%	Model struggles to identify non-purchases

Decision Tree



Feature Importance

Review Length (52%):

- This feature has the highest importance by far, suggesting that the length of user reviews plays a crucial role in the model's predictions.

Genre Length (9.1%):

- The diversity or number of genres associated with a game is a significant factor.

Received for Free (<1%):

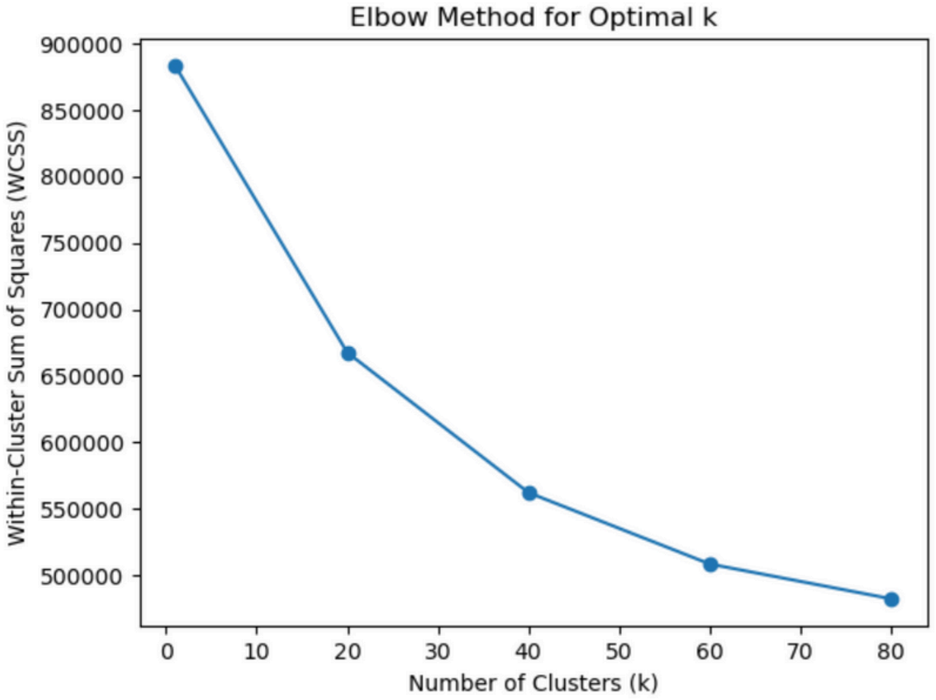
- Whether the game was received for free impacts the prediction moderately.

Discount Percentage (<1%):

- The percentage of discount offered is another key factor influencing the likelihood of purchase.

K Mean: Cluster Model

What methods can be explored to group reviews based on common themes to better understand user feedback and preferences?




```
Cluster 0:
simulator, truck, best, game, driving

Cluster 1:
hours, game, fun, played, play

Cluster 2:
good game, good, game, really, play

Cluster 3:
portal, f🦆k, games, s💩t, im

Cluster 4:
great game, great, game, play, fun

Cluster 5:
gta, online, game, rockstar, best
```

Clustering Reviews

Grouped By:

- TF-IDF (Term Frequency-Inverse Document Frequency)

Action Items:

- Improve game quality by addressing key concerns.
- Enhance marketing and recommendations by highlighting strengths.
- Better understand player preferences and emerging trends.

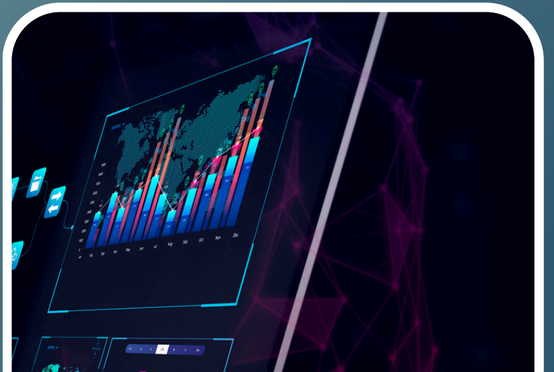
Recommendations

01 The sentiment score analysis provides a clear roadmap for enhancing developer performance and player satisfaction on Steam. By implementing these recommendations, Steam can maximize the impact of **high-performing developers**, address weaknesses, and foster a stronger, more engaged player community.

02 Focus future game development on genres blending action, adventure, RPG. Incorporating strong multiplayer features with steam achievements and co-op have been proven success factors of both current favorites and historical best-sellers.

03 Utilize the neural network and KNN models to **predict** users who are most likely to purchase games. Focus marketing efforts on these users through personalized promotions, discounts, and targeted game recommendations to maximize conversion rates. These findings are confirmed through feature importance analysis, providing a robust understanding of what drives customer decisions.

04 Identify player priorities and concerns. By acting on these insights, Steam can optimize marketing strategies, **improve** game recommendations, and support developers in creating better experiences.

Thank You!

Q&A