DS 7331 - Lab 1 (Video Game Sales with Ratings Dataset Analysis)

Eric Graham

Introduction

The purpose of this analysis is to explore the factors that contribute to critical ratings of video games. Through exploratory data analysis, I examined patterns in game characteristics and their relationship to critical quality. After cleaning and preprocessing the data, I then fit PCA and LDA models to identify the most impactful features driving overall variance in the dataset and to determine whether feature combinations could effectively predict critical response to a game.

Dataset Information

This project uses the Video Games Sales with Ratings dataset from Kaggle. It includes 16719 observations and 18 features.

Variable Glossary

Variable	Type	Description
Name	object	Name of the game
Platform	object	Console on which the game is running
Year_of_Release	float64	Year of the game released
Genre	object	Game's category
Publisher	object	Publisher
NA_Sales	float64	Game sales in North America (in millions of units)
EU_Sales	float64	Game sales in the European Union (in millions of units)

Variable	Type	Description
JP_Sales	float64	Game sales in Japan (in millions of units)
Other_Sales	float64	Game sales in the rest of the world, i.e. Africa, Asia excluding Japan, Australia, Europe excluding the E.U. and South America (in millions of units)
Global_Sales	float64	Total sales in the world (in millions of units)
Critic_Score	float64	Aggregate score compiled by Metacritic staff
Critic_Count	float64	The number of critics used in coming up with the Critic_score
User_Score	object	Score by Metacritic's subscribers
User_Count	float64	Number of users who gave the user_score
Developer	object	Party responsible for creating the game
Rating	object	The ESRB ratings (E.g. Everyone, Teen, Adults Only..etc)

Exploratory Data Analysis Highlights

Bins for Target Variable

Helpfully, Metacritic supplies their own descriptive bins for scores:

Score Range	Video Games Classification
90-100	Universal acclaim
75-89	Generally favorable
50-74	Mixed or average
20-49	Generally unfavorable
0-19	Overwhelming dislike

I opted to start with these bins because of their business relevance and interpretability: the official Metacritic classifications have an immediately-recognizable meaning to stakeholders. However, for LDA there is a risk that severely unbalanced classes would affect model performance by causing larger classes to have an outsized influence on feature extraction, so I will keep an eye out for extremely small bins.

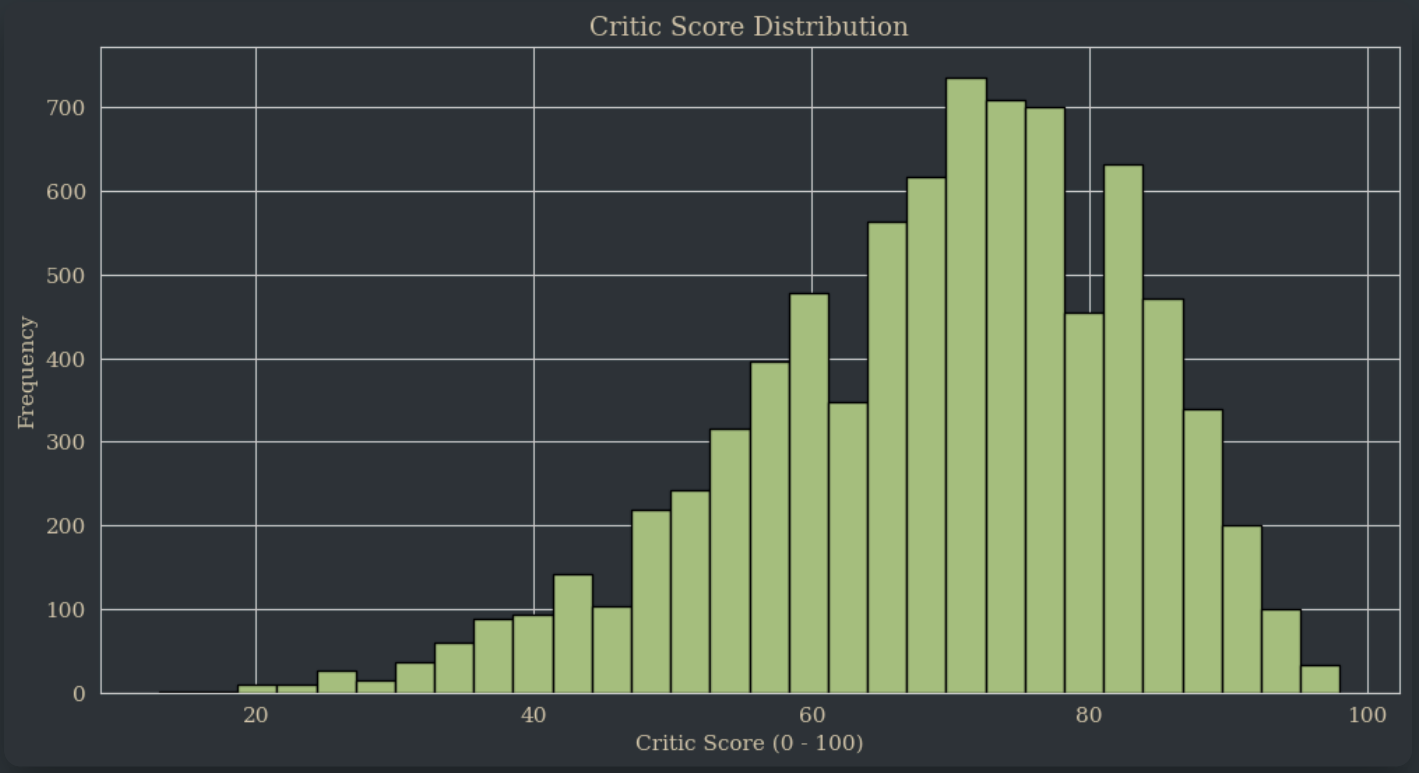
```
critic_category
Mixed or Average         4326
Generally Favorable      2673
Generally Unfavorable     870
Universal Acclaim         257
Overwhelming Dislike       11
Name: count, dtype: int64
```

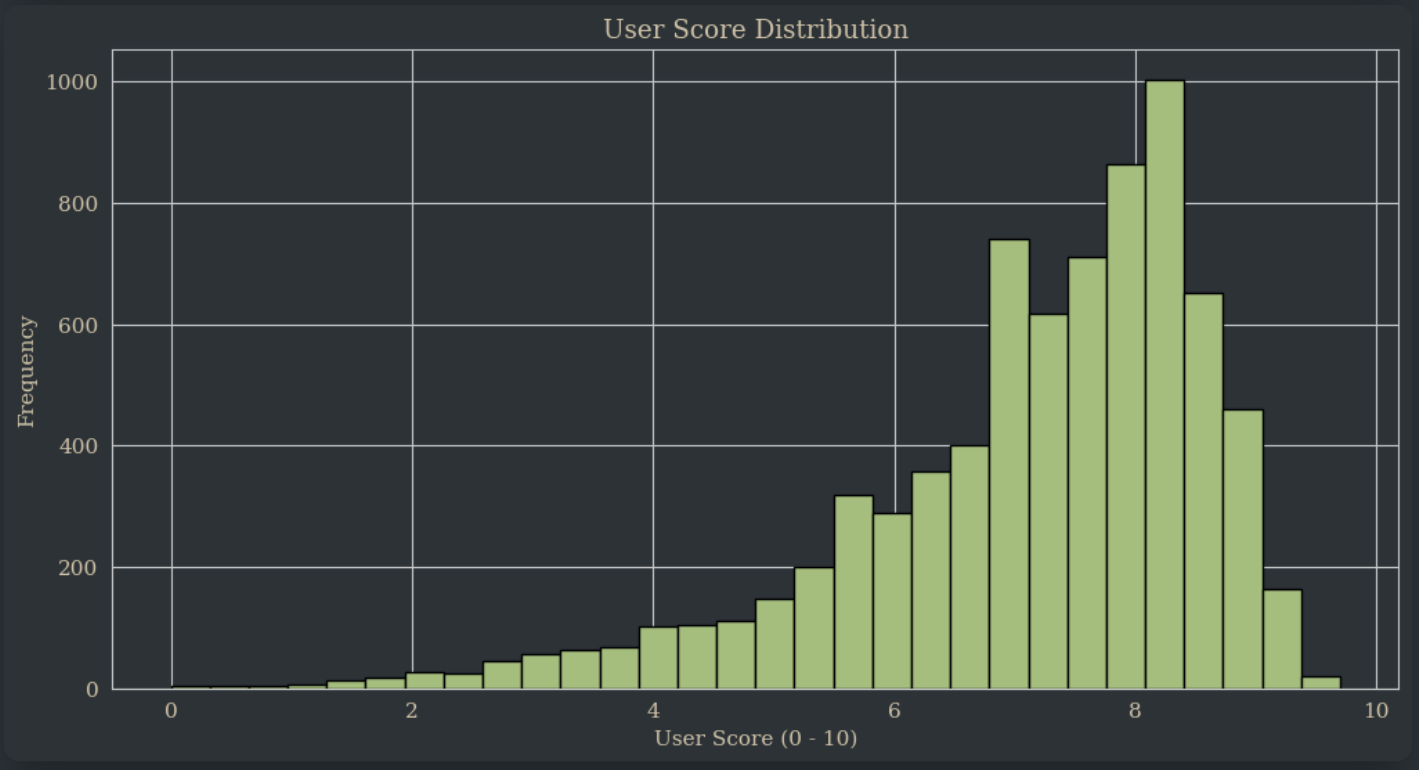
The above counts show that there are only 11 observations in the "Overwhelming Dislike" bin, which could have a negative impact on LDA, so I grouped it with "Generally Unfavorable" to create the below distribution:

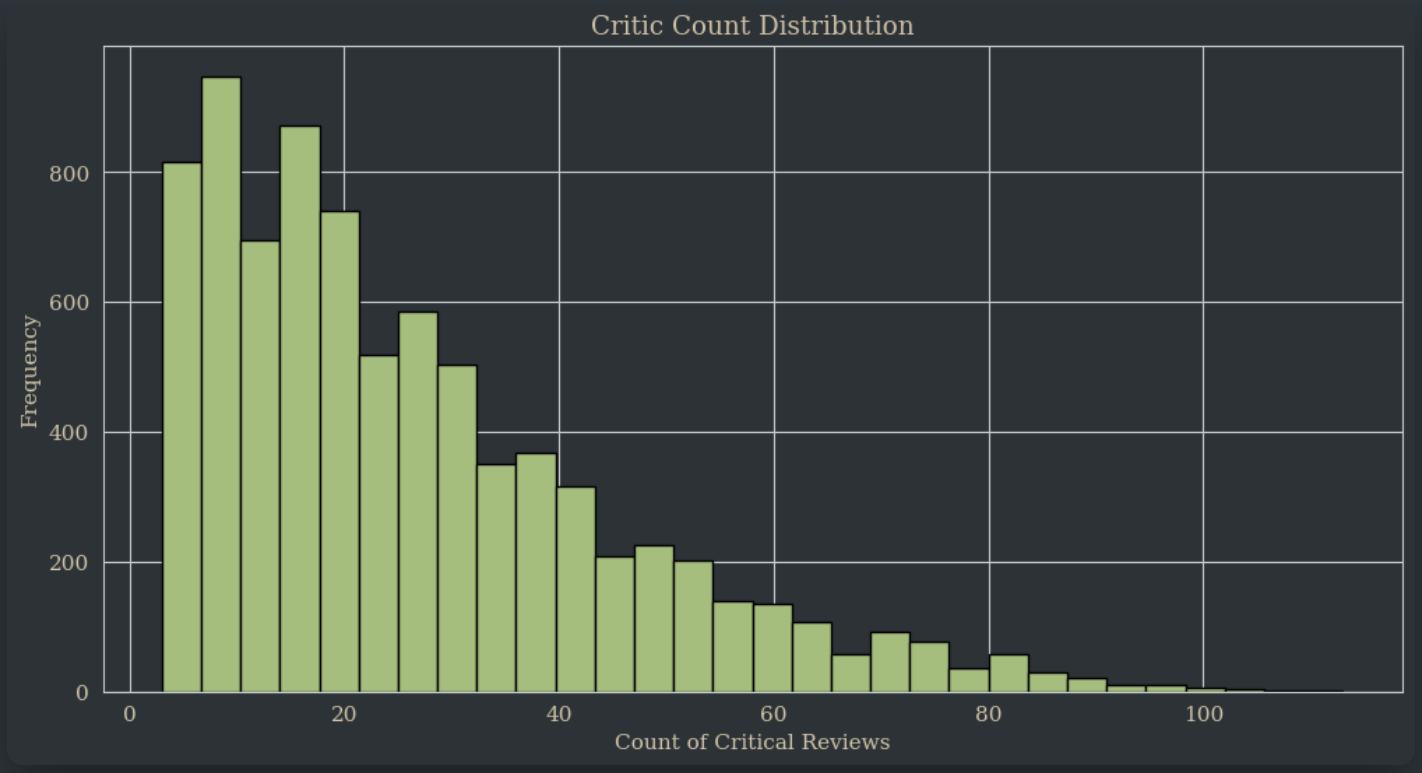
```
critic_category
Mixed or Average                              4326
Generally Favorable                           2673
Generally Unfavorable / Overwhelming Dislike   881
Universal Acclaim                              257
Name: count, dtype: int64
```

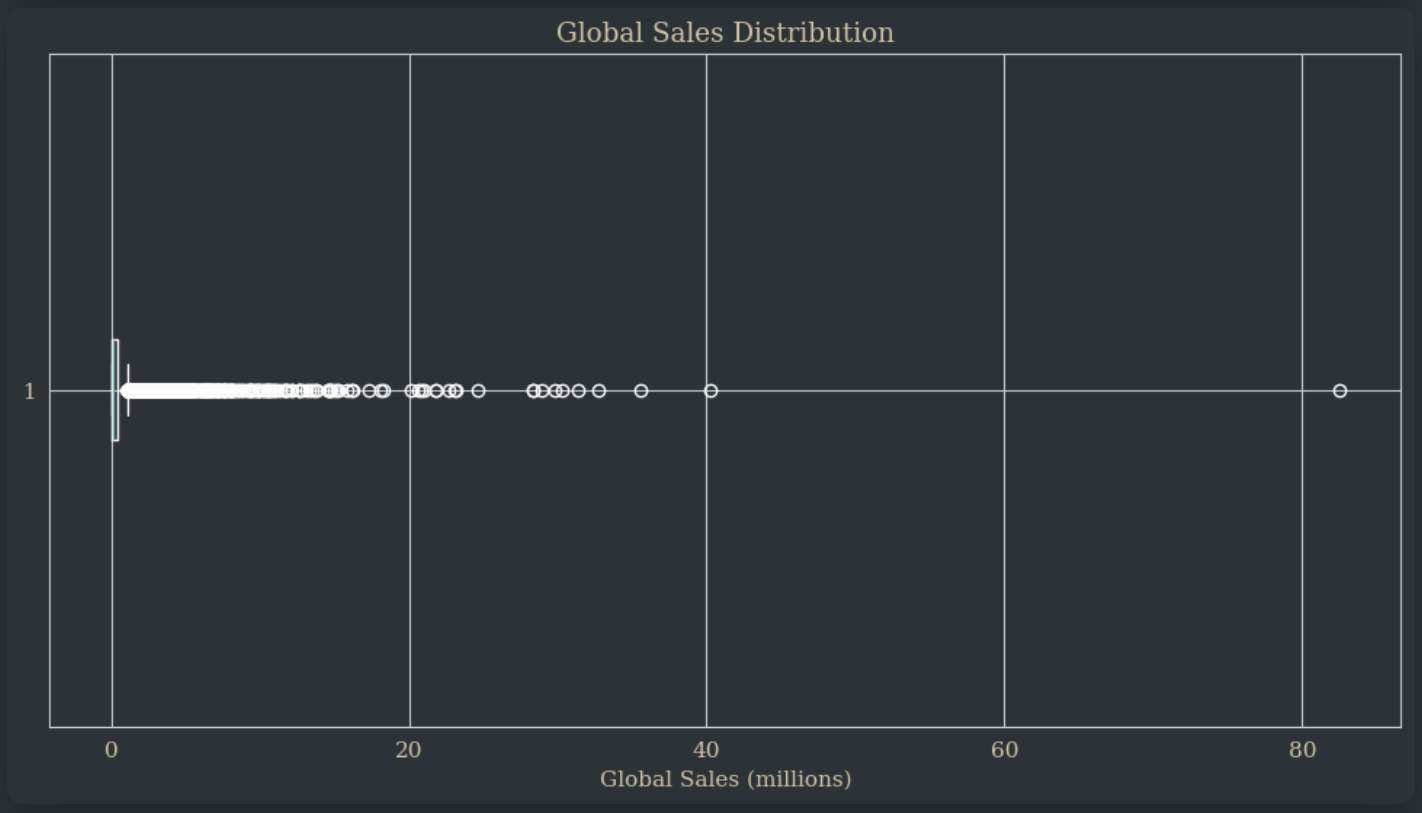
Univariate Analysis

Numeric Variable Plots

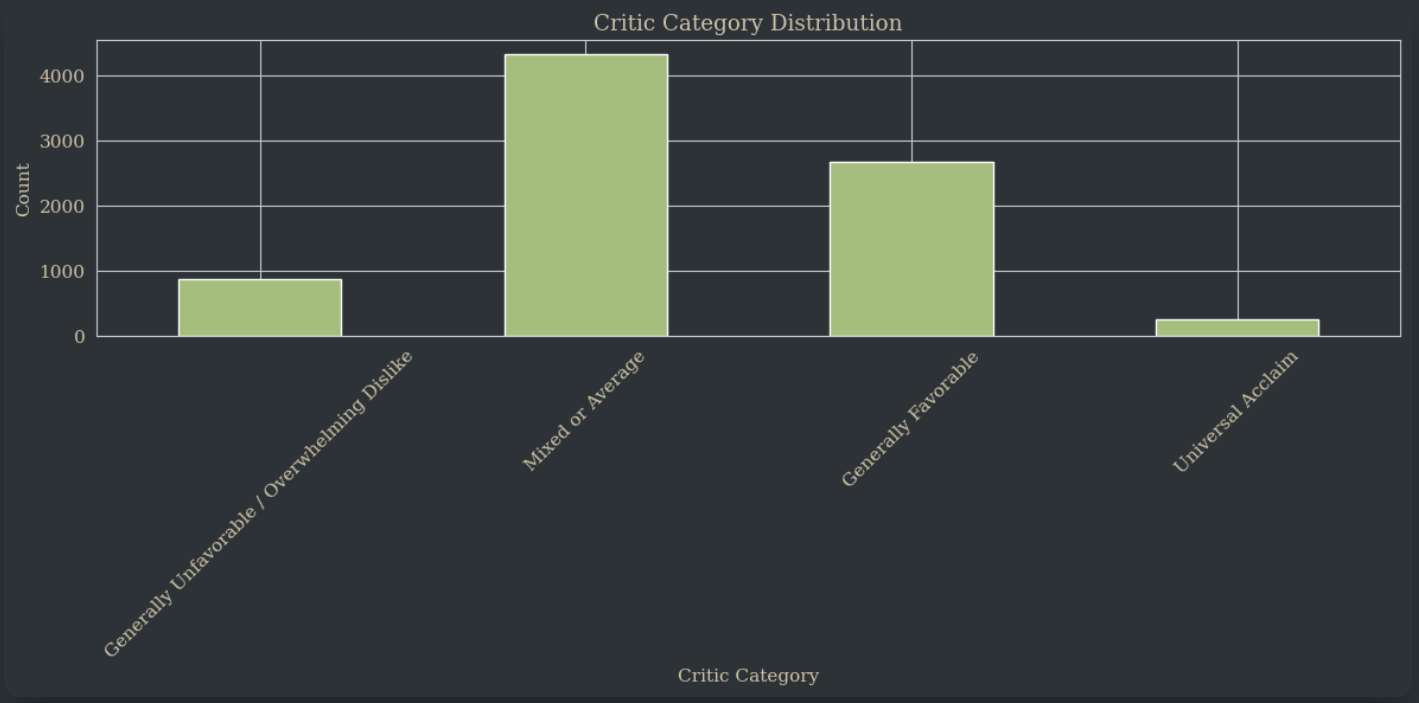


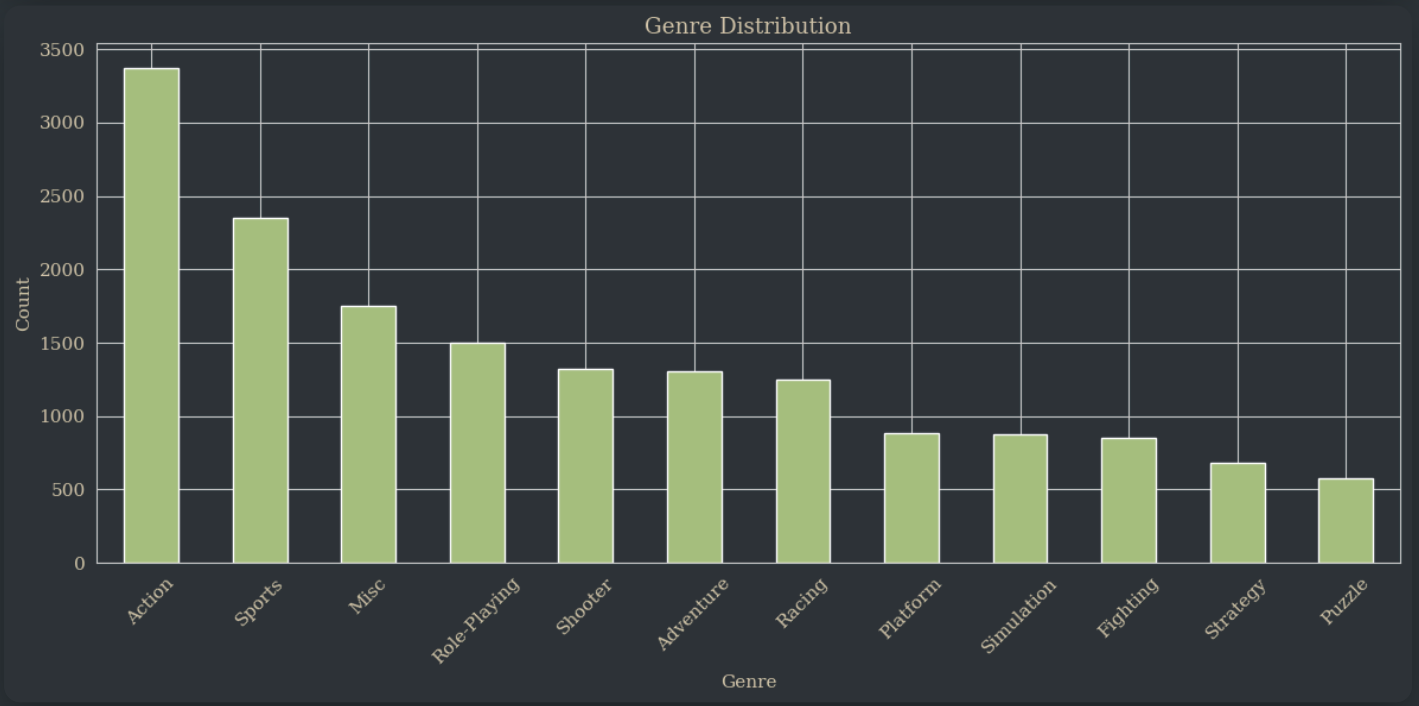



Critic Count Distribution
Frequency
0
200
400
600
800
0
20
40
60
80
100
Count of Critical Reviews


Global Sales Distribution
1
0
20
40
60
80
Global Sales (millions)

Categorical Variable Plots





Findings

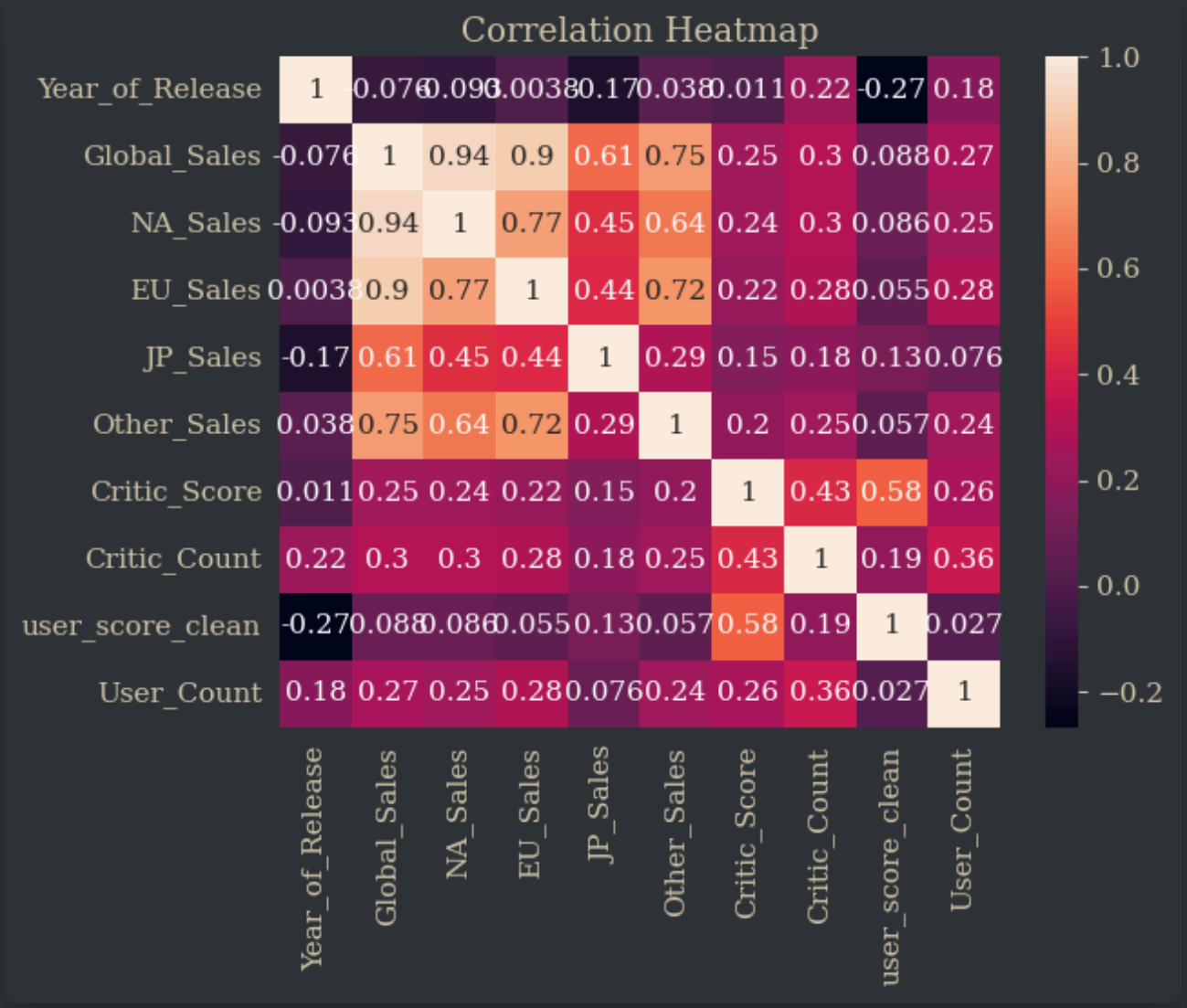
- Critic_Score has a moderate left skew, indicating that very few titles receive very high or low scores. While it would be inaccurate to say that the bins are normally distributed, it is true that the largest bins is "Mixed or Average," followed by "Generally Favorable," which reflects the distribution of the numeric ratings.
- Interestingly, User_Score is also moderately left-skewed.
- Sales figures generally have a right skew, indicating that most games don't sell many

copies.

- Likewise, Critic_Count has a heavy right skew, which indicates that most games aren't widely reviewed.
- Genre is moderately imbalanced in favor of Action and Sports

Bivariate Analysis

Correlation Heatmap and Matrix

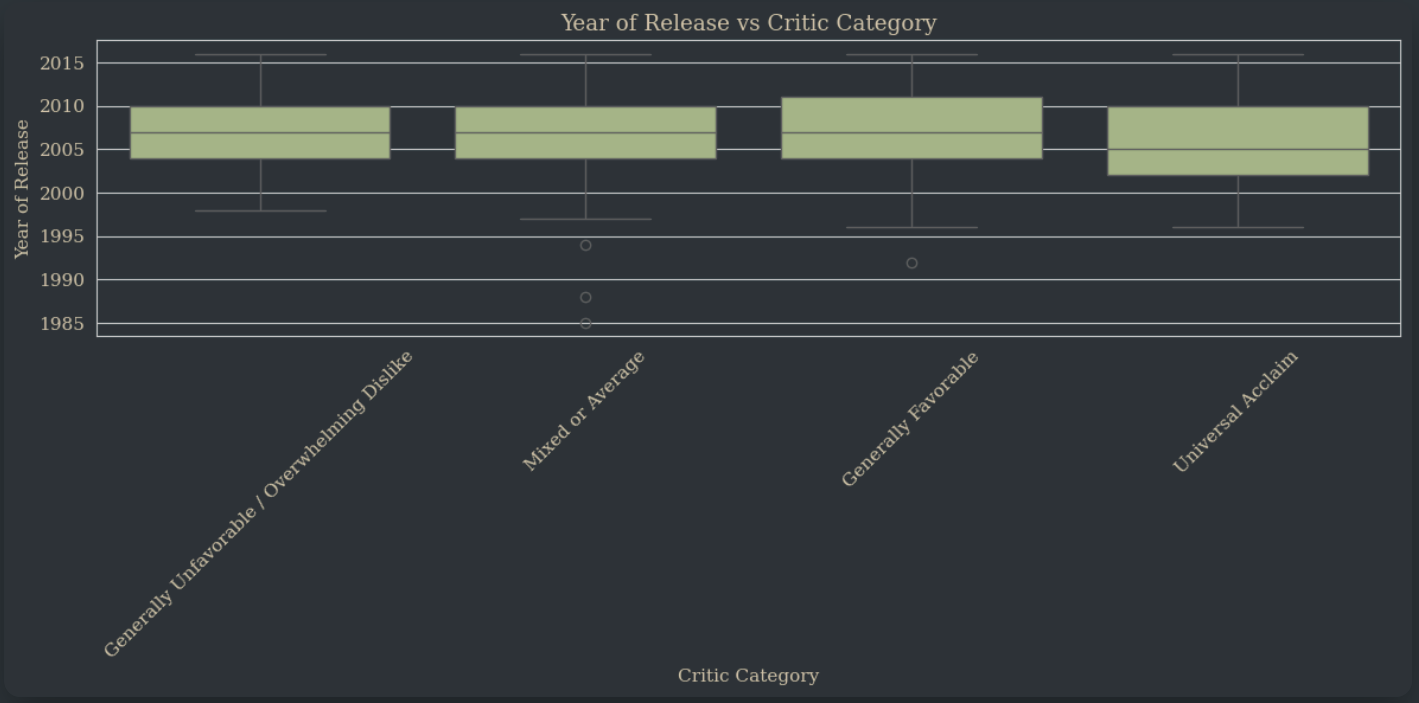


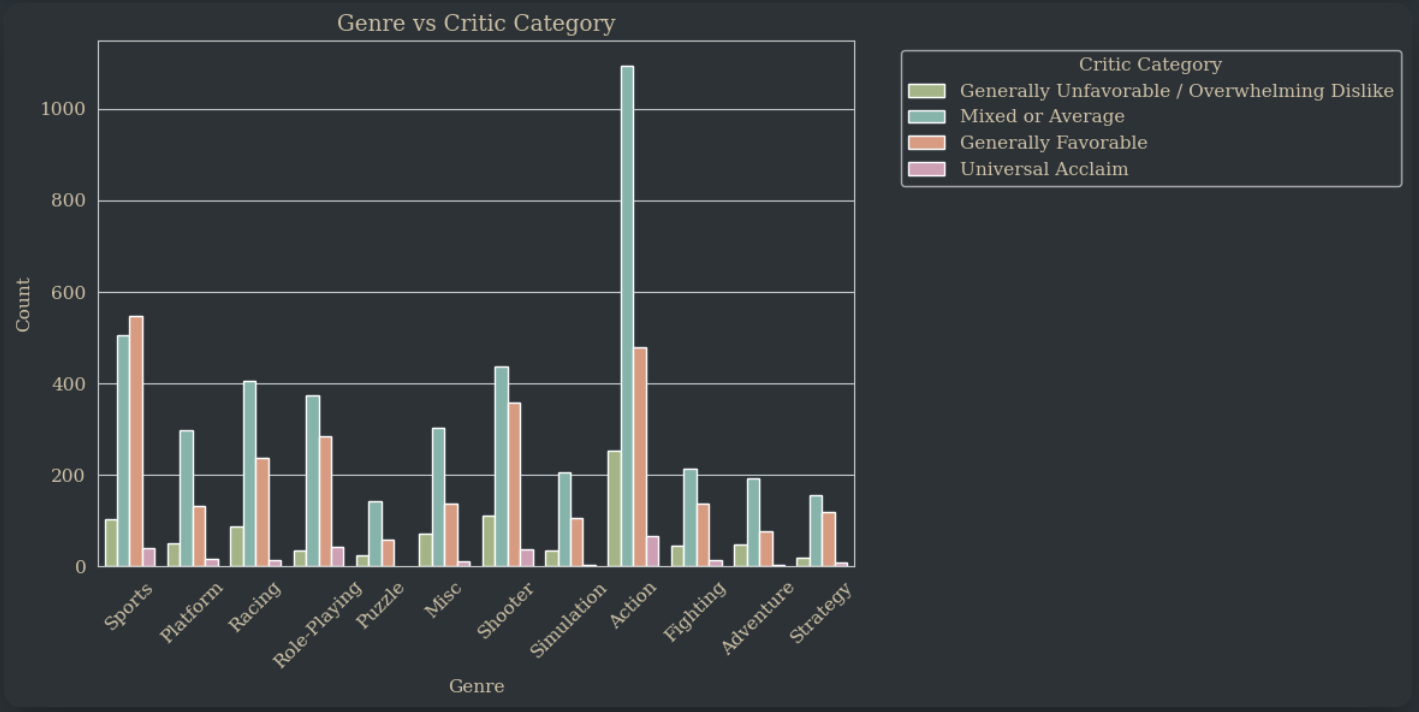
```
Critic_Score        1.000000
user_score_clean    0.580878
Critic_Count        0.425504
User_Count          0.264376
Global_Sales        0.245471
NA_Sales            0.240755
EU_Sales            0.220752
Other_Sales         0.198554
JP_Sales            0.152593
Year_of_Release     0.011411
Name: Critic_Score, dtype: float64
```

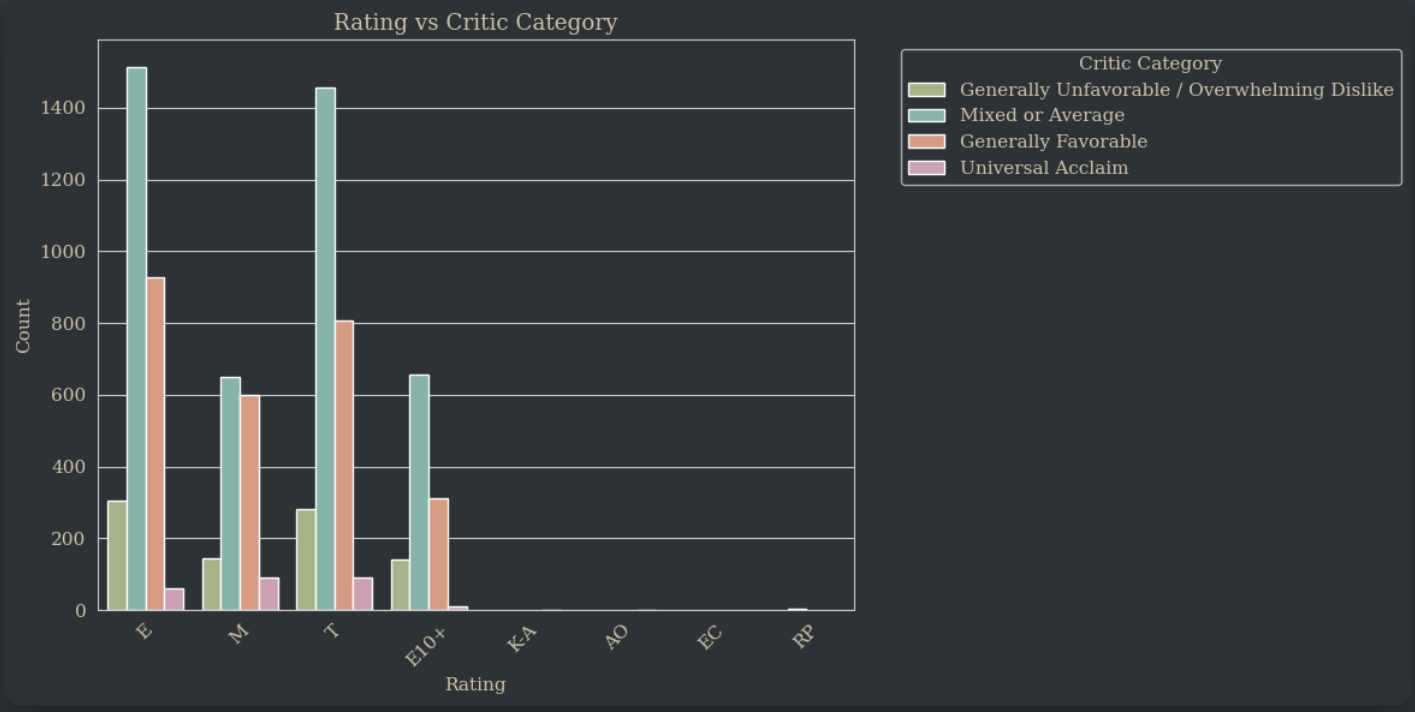
Findings

- Based on the above correlation table, I am immediately suspicious of the relationship between Critic_Count and Critic_Score, as games that receive critical acclaim likely get reviewed more, potentially creating a "leaky variable" effect. From a domain perspective, I don't think this has business value: advising a game publisher to seek out more reviews doesn't seem like intuitively good advice to help them get better reviews, and could even backfire.
- I'm slightly concerned about the relationship with user_score_clean, based on the idea that a consumer's perception of a game being "critically acclaimed" might lead them to give it a higher review. Aside from that, one might intuit that a game which pleases a critic will also please players because critics have to play the games to review them. This is less worrisome to me than the Critic_Count relationship because I think user_score_clean has a less direct path to "leakiness." From a domain perspective, telling publishers that making a game that players like will improve their reviews seems intuitively sensible.

Target-Feature Plots







Findings

- The mean of Year seems to be a bit lower for games receiving "Universal Acclaim," which indicates to me that older games were more likely to receive high scores.
- The relationship between critical_score_category and Genre looks ripe for analysis; Sports games skew towards higher critical ratings, while Platform games seen to trend more towards "mixed or average" reviews.
- As with Genre, the M Rating category seems to track with a higher proportion of Generally Favorable reviews.

EDA Summary

As noted in greater detail in the notebook and above:

- Most games receive "Mixed or Average" or "Generally Favorable" critic scores
- Sales data is heavily right-skewed, indicating that most games don't sell many copies
- Critic_Count is right-skewed, showing that most games don't receive many reviews
- Critic_Count is likely a leaky variable
- User and critic scores moderately correlated (0.58)
- Older games more likely to receive "Universal Acclaim" score
- Genre and Rating are potentially useful categorical predictors

Data Cleaning and Preprocessing

Missing Values

As seen below, a significant number of observations lack values for key variables, including our target variable!

	Missing_Count	Missing_Percentage
Name	2	0.011962
Platform	0	0.000000
Year_of_Release	269	1.608948
Genre	2	0.011962
Publisher	54	0.322986
NA_Sales	0	0.000000
EU_Sales	0	0.000000
JP_Sales	0	0.000000
Other_Sales	0	0.000000
Global_Sales	0	0.000000
Critic_Score	8582	51.330821
Critic_Count	8582	51.330821
User_Score	6704	40.098092
User_Count	9129	54.602548
Developer	6623	39.613613
Rating	6769	40.486871

Missing Values to Drop

1. **critic_score_category** is our target variable, so we can't properly conduct analysis without it. I will drop all records for which this value is null. This will also drops null values for the **Critic_Score** feature (from which our binned target was derived) and **critic_count**, which I also regard as being redundant with the target variable.
2. Because of overlap and redudancy among these features, I will drop the original Critic_Score and Critic_Count columns from the model entirely.
3. **User_Score** and **user_score_clean** are potentially-valuable indicators of game quality, but null or TBD values in these features make up more than 50% of our dataset, which would be too much synthetic data to impute, so these features will be dropped as well.
4. There are a small number of missing records for **Name** and **Genre** so little is lost by dropping the null observations for these features.
5. There are a moderate number of missing records for **Year_of_Release**. I can't put my finger on it but something feels wrong about imputing year values, so I'm dropping these.

Missing Values to Impute

1. **Rating** is an interesting case because ESRB ratings weren't implemented until 1994, so there is likely a time-based component to this missingness.
2. **Publisher** and **Developer** are strictly categorical features with a multitude of possible labels.

For these three features, I chose to implement an "Unknown" category to preserve potential meaningfulness within

Feature Engineering

Categorical Features

High Cardinality Features

I note that Publisher and Developer have far too many unique values to make one-hot encoding practical: even though I believe in the power of my PC and its abundance of RAM, from a statistical perspective we encounter the "curse of dimensionality." For features with such high cardinality, I chose to only use the top 15 values and lump the remaining into an "Other"

category.

One-Hot Encoding

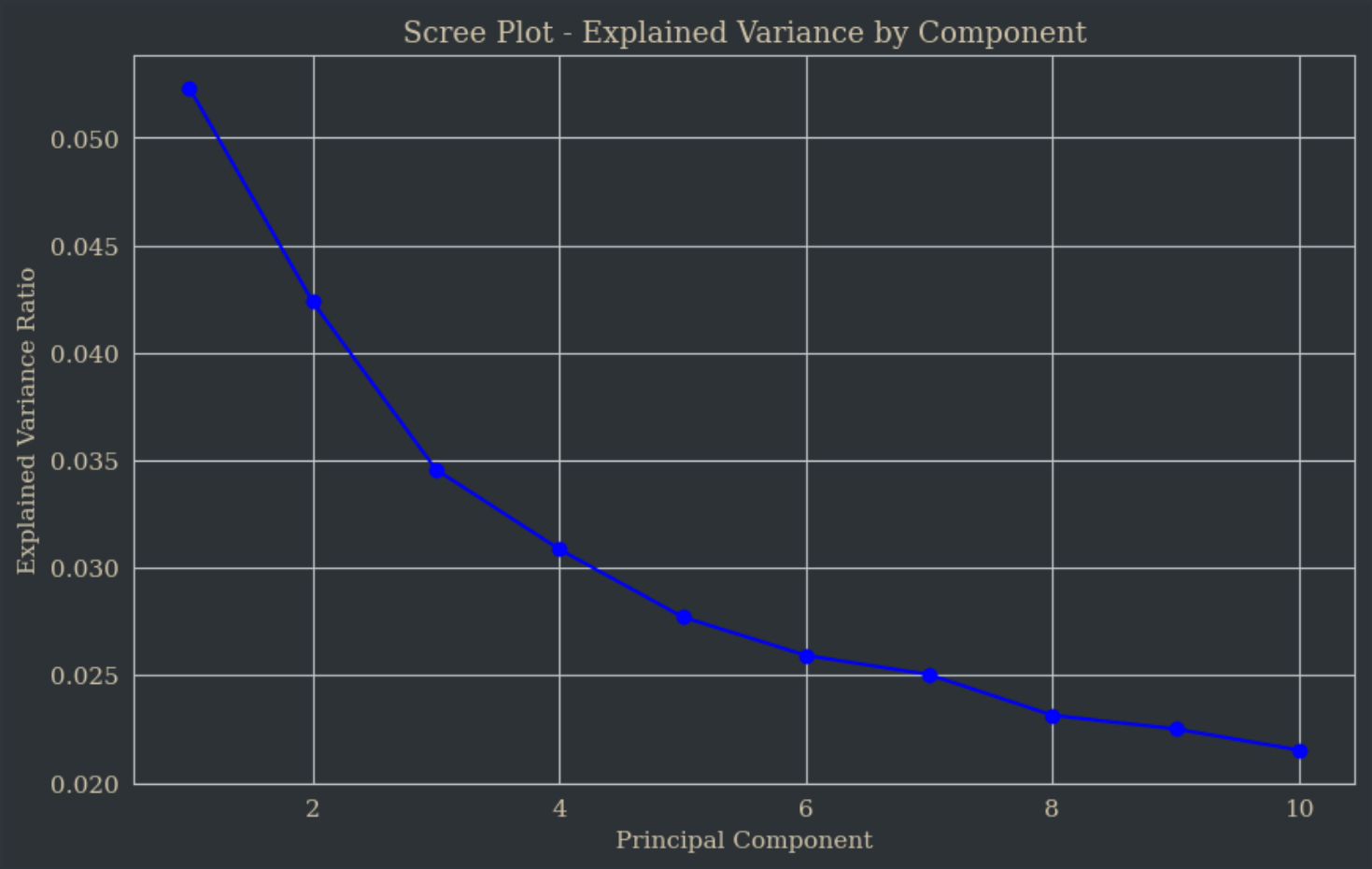
With the high-cardinality features have been transformed, I feel confident one-hot encoding the categorical variables.

Feature Scaling

Because the sales variables and Year_of_Release are on much larger scales than the one-hot encoded categorical features, I standardized all features using StandardScaler to prevent one of these larger feature types from dominating the PCA and LDA analysis. Note that the Name variable isn't needed for PCA or LDA, so it is dropped from our X.

Principal Component Analysis

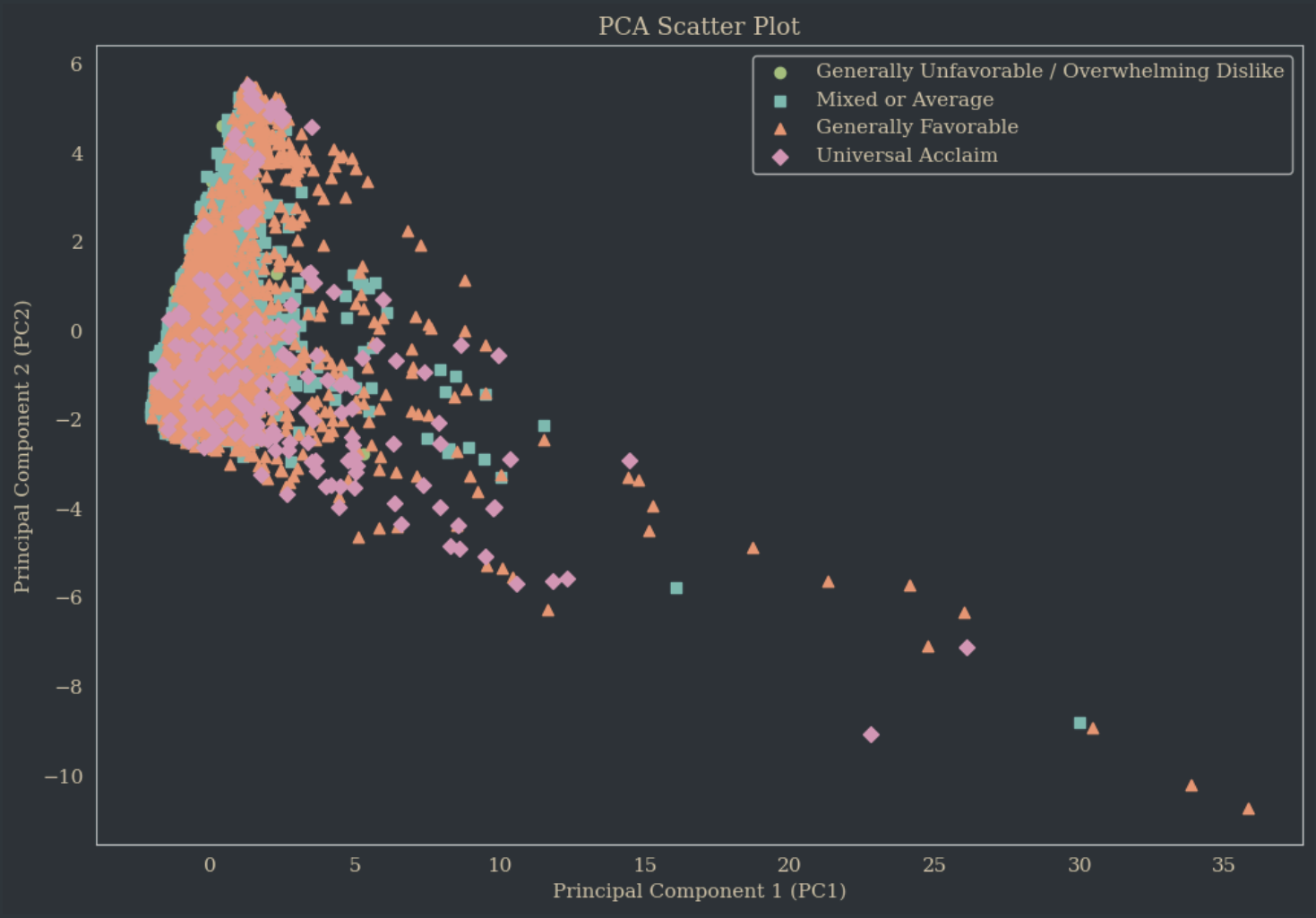
Scree Plot



Analysis

The above Scree plot shows that PC1 and PC2 explain the most variance at 5.2% and 4.2%, followed by a steep dropoff and gradual decline of variance explained in the later principal components. This confirms that the data is complex and highly dimensional, with very distributed variance.

PCA Scatter Plot



Analysis

The above scatterplot shows where individual games lie on a 2D plot of PC1 and PC2, with color/shape coding to indicate our critical score categories. The significant overlap of the critical score categories shows that the PC1 and PC2 data patterns (which I will discuss below, in section 5.5) don't effectively distinguish between critical quality levels. From a business perspective, this tells me that critical response to a game is a very complex and nuanced outcome.

Feature Contributions

PC1

```
Global_Sales          0.404746
JP_Sales              0.365230
NA_Sales              0.364443
EU_Sales              0.359019
Other_Sales           0.337074
Developer_Nintendo    0.307009
Publisher_Nintendo    0.246304
Developer_Other       0.204038
Publisher_Other       0.165987
Rating_E              0.136028
dtype: float64
```

PC2

```
Rating_E                    0.386190
Genre_Sports                0.366101
Developer_Other             0.289871
Publisher_Electronic Arts   0.281128
Year_of_Release             0.235201
Rating_M                    0.221180
Developer_EA Sports         0.210954
Developer_EA Canada         0.201888
Developer_EA Tiburon        0.152468
Rating_T                    0.151824
dtype: float64
```

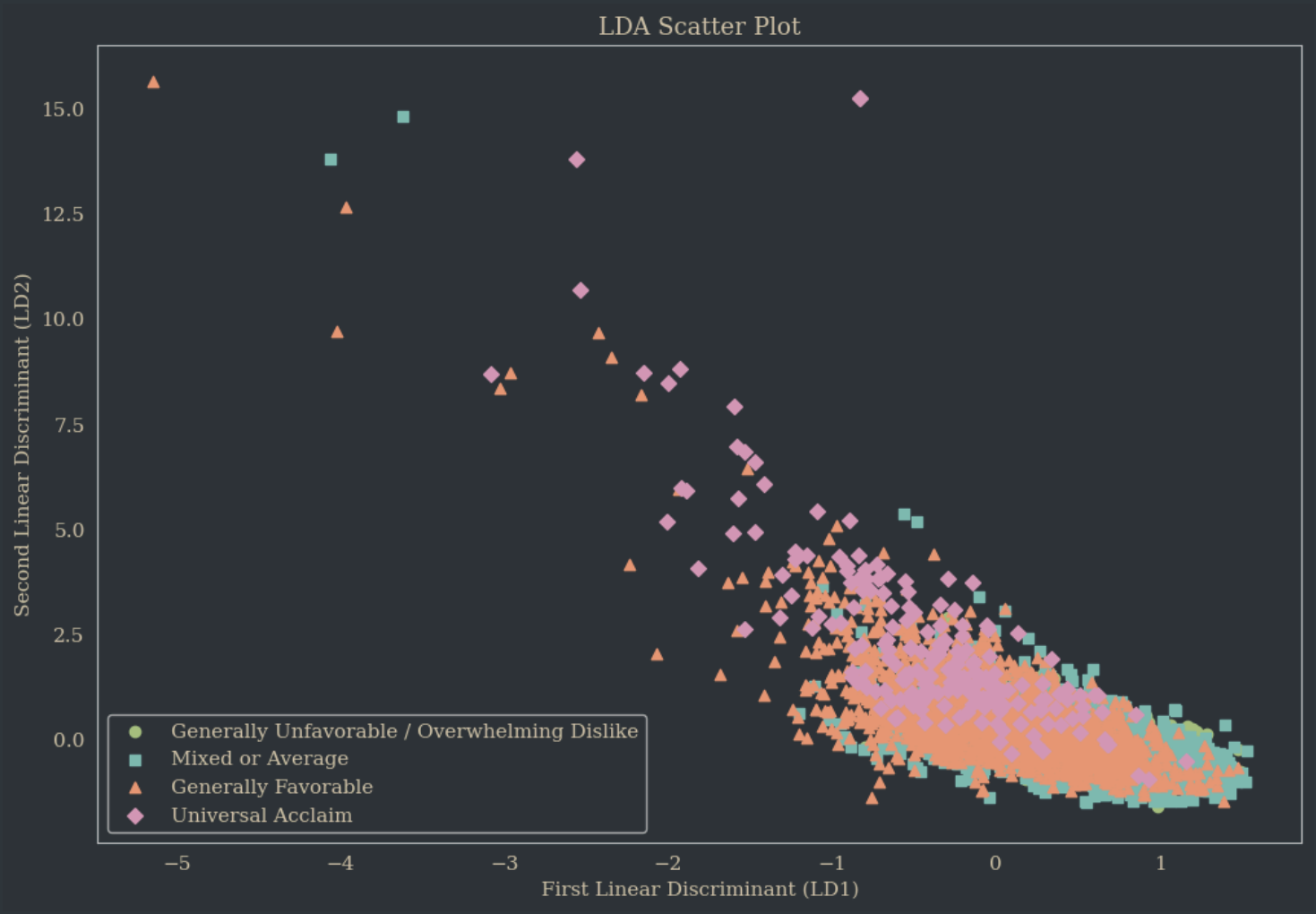
Analysis

The above lists of top contributing features for PC1 and PC2 suggest two game profiles that explain the most variance in the dataset:

1. PC1 is our "commercially successful" bucket, dominated by sales figures and the developer/publisher Nintendo
2. PC2 is more complex: it is dominated by the "E for Everyone" rating, followed by the "sports" Genre and EA Sports publishers/developers. A significant amount of variance is also explained by "Developer_Other", and "M for Mature" rating, which is a fruitful area for further research.

Linear Discriminant Analysis

LDA Scatter Plot



Analysis

In our LDA scatterplot, we see the clear value of supervised learning in predictive analysis: there is distinct clustering of critical score categories (with some overlap), which contrasts sharply with PCA's mixed results. LDA was able to identify linear combinations of features that effectively separate critical quality levels.

Confusion Matrix and Classification Report

```
[[ 23   1 468  33]
 [  1   1 170   0]
 [  9   4 831   6]
 [  3   0  28  19]]
```

	precision	recall	f1-score	support
Generally Favorable	0.64	0.04	0.08	525

Generally Unfavorable / Overwhelming Dislike	0.17	0.01	0.01	172
Mixed or Average	0.56	0.98	0.71	850
Universal Acclaim	0.33	0.38	0.35	50
accuracy			0.55	1597
macro avg	0.42	0.35	0.29	1597
weighted avg	0.53	0.55	0.42	1597

Analysis

Unfortunately, the LDA shows generally poor accuracy (55%) and exhibits bias in that it is more effective at predicting the "Mixed or Average" class (which, as seen in EDA, is the largest class). From a business standpoint, a model that is only effective at identifying average critical response is not particularly useful.

PCA vs LDA Comparison

The ability of LDA to distinguish feature combinations that predict critical score category is a function of it being a supervised learning model. However, while LDA identified some patterns it struggled to translate them into reliable predictions (especially for the minority classes).

In contrast, PCA didn't distinguish between the critical score categories because that's not what it was designed to do: it was able to show us the the combination of features that explain the most variance in the dataset as a whole. It revealed the general patterns in our data, not patterns that predict a target.

Interpretation and Limitations

Post-Modeling EDA Summary

Critical vs. Commercial Success

In light of our PCA analysis, the distinction between critical and commercial success that was seen in EDA makes sense: PC1 was dominated by the commercial success features, but that didn't effectively discriminate between critical quality score categories.

Complexity of Critical Quality

The PCA/LDA analysis confirms the observation from EDA that critical success is a complex and nuanced phenomenon which can't be easily explained by casual examination. Critical quality explains little overall variance in the PCA, and while the LDA had some success in predicting critical score it was heavily biased.

This provides a quantitative explanation for something that critics themselves would likely assert based on intuition: the quality of a work of art is a combination of many subtle, interconnected features.

Preprocessing Effects on PCA/LDA

One-hot encoding was essential for bringing categorical variables into the models, I can tell we're going to be doing that a lot!

This made feature scaling even more critical, in order to prevent the large scale of the sales figures from exercising outsized influence compared to our binary categorical features.

Limitations and Next Steps

A notable limitation of LDA is that it requires the assumptions of linearity be met. This wasn't confirmed prior to modeling, and may explain some of the issues with model performance. Another possible issue is class imbalance: I took a light approach in feature engineering, but may have left too much imbalance in the dataset.

Ultimately, the LDA model's bias limits its usefulness. Different binning or sampling approaches may be more fruitful than the ones used here, but it seems likely that a different approach might be needed to predict critical scores from this data.

Colophon

This notebook was written entirely in Jupyter running in a Miniconda environment on Ubuntu 22.04 running under WSL.

The PDF report was generated from Typora using the Everforest theme.

Claude.ai was used for generating markdown tables, repetitive code repurposing, plot formatting (to match the Everforest theme), and moral support.