Automated Data Analysis Report (via Gemini): Steam

Executive Summary

This report summarizes the initial automated exploratory data analysis (EDA) of the 'steam.csv' dataset, containing 27085 rows and 18 columns (9 numerical, 9 categorical). Preliminary analysis revealed a relatively clean dataset with only 10 duplicate entries and no missing values or constant columns. Initial univariate and bivariate analyses, including descriptive statistics and visualizations, have been conducted on all features. Two key observations from bivariate analysis are noted, though further investigation is needed to fully understand their implications. The dataset's size and apparent cleanliness suggest ample data for robust modeling. The balance of numerical and categorical features indicates the potential for diverse analytical approaches. The initial EDA has established a solid foundation for subsequent, more in-depth analyses. No major data quality issues were uncovered, facilitating efficient progression to more advanced modeling techniques. This initial scan provides a strong starting point for further investigation. The identified observations warrant further exploration, and the lack of significant data quality issues allows us to proceed directly to more sophisticated analyses to derive actionable insights from the Steam dataset.

1. Data Overview

This report provides an initial automated analysis of the dataset from 'steam.csv'.

1.1. Basic Information

Table 1: Dataset Dimensions

Metric	Value
Number of Rows	27085
Number of Columns	18
Total Data Points	487530

1.2. Data Types

Table 2: Summary of Feature Data Types

Data Type	Count
object	9
int64	8
float64	1

Data Types Distribution Interpretation:

> The dataset contains a roughly even mix of numerical and categorical features, lacking any datetime features. This suggests a moderately diverse dataset suitable for a range of analytical techniques, though the absence of temporal data limits time-series analysis possibilities.

2. Data Quality Assessment

2.1. Missing Values

No missing values were found in the dataset. This is excellent for data completeness.

2.2. Duplicate Records

The dataset contains 10 duplicate rows (representing 0.04% of the data). These may need to be investigated or removed depending on the analysis context, as they can skew results.

2.3. Feature Variance

No constant columns (columns with only one unique value) were identified.

The following columns are quasi-constant (one value is highly dominant), potentially offering limited information: english (dominant value: 1 at 98.1%); requiredage (dominant value: 0 at 97.8%). Their utility should be reviewed.

Data Quality Summary & Implications:

The data quality assessment reveals a relatively high level of cleanliness in the dataset of 27,085 rows. The absence of missing values is a significant positive, indicating a complete dataset ready for analysis. The low rate of duplicate rows (0.04%) is also negligible and easily addressed. The presence of quasi-constant columns, 'english' and 'requiredage', however, warrants further investigation. While not strictly problematic, their high dominance in a single value (98.1% and 97.8% respectively) suggests these variables may have limited predictive power in many modeling scenarios and could potentially be redundant. The implications for further analysis are primarily related to the quasi-constant columns. Including these variables in models might not significantly improve predictive accuracy and could even introduce noise or bias. Furthermore, analyses focusing on these variables would yield limited insights due to the lack of variability. The small number of duplicate rows is unlikely to significantly impact analysis, but their removal is a simple and recommended step to maintain data integrity. To address the identified issues, the duplicate rows should be removed. For the quasi-constant columns, a thorough investigation into the reasons for their high dominance is necessary. This might involve exploring data collection methods or the underlying population characteristics. Depending on the research question, these columns might be safely removed, or if they represent a meaningful categorical variable with a small minority class, techniques like oversampling or data augmentation could be considered to improve balance. If their inclusion is deemed necessary for context, careful consideration of their limited predictive value should be factored into any modeling interpretations.

3. Univariate Analysis

3.1. Numerical Features

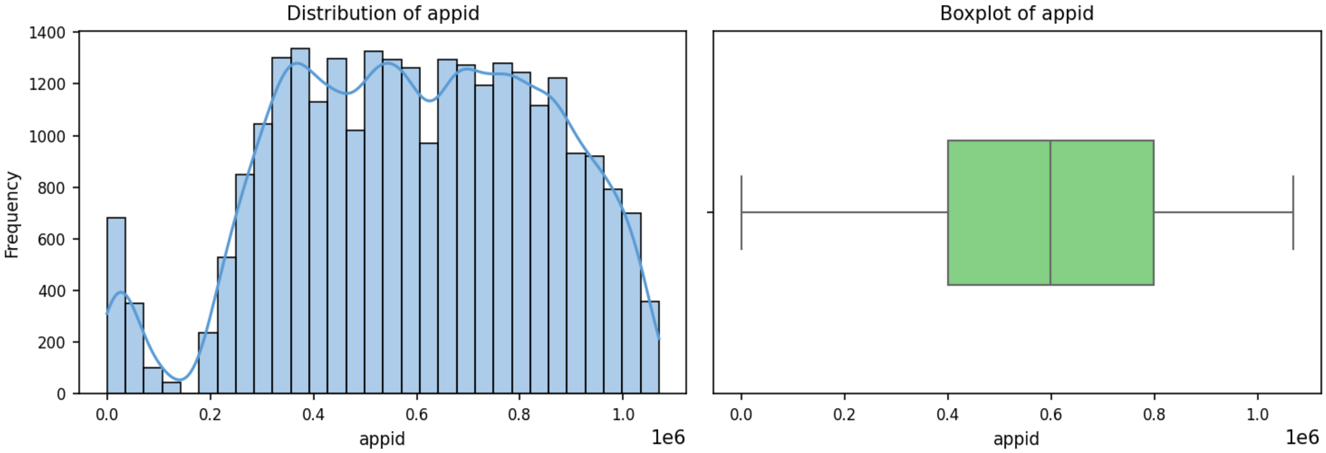

Figure 1: *Distribution (histogram and KDE) and boxplot for 'appid'. The histogram shows shape, central tendency, and spread. The boxplot highlights median, quartiles, and potential outliers.*

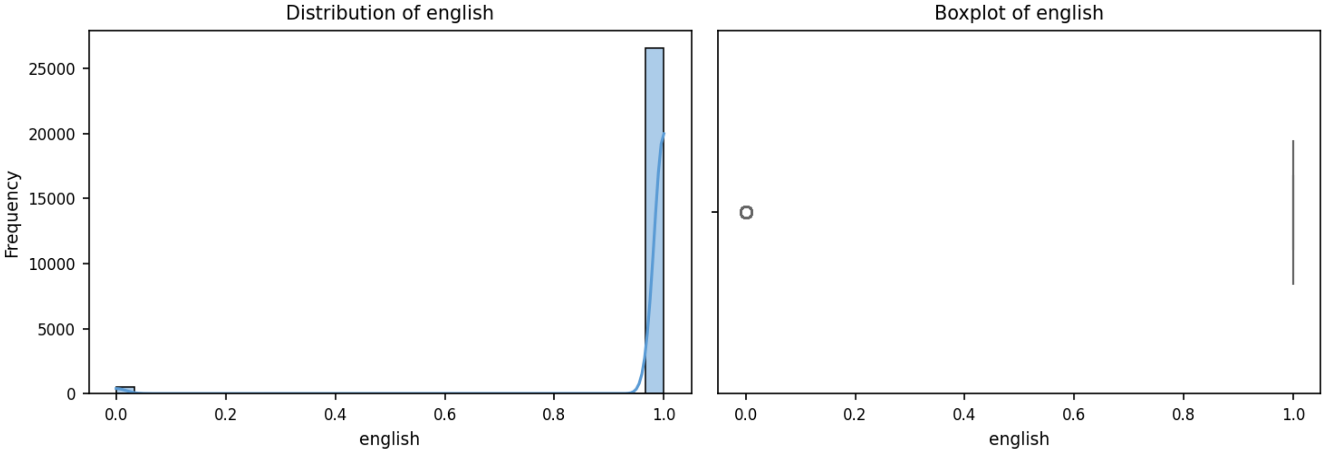

Figure 2: *Distribution (histogram and KDE) and boxplot for 'english'. The histogram shows shape, central tendency, and spread. The boxplot highlights median, quartiles, and potential outliers.*

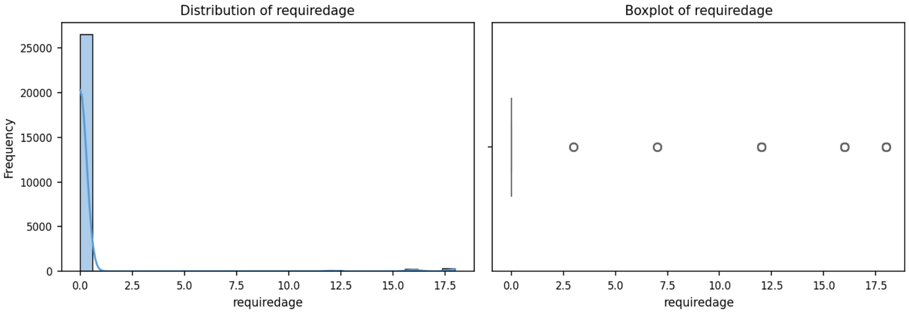

Figure 3: *Distribution (histogram and KDE) and boxplot for 'requiredage'. The histogram shows shape, central tendency, and spread. The boxplot highlights median, quartiles, and potential outliers.*

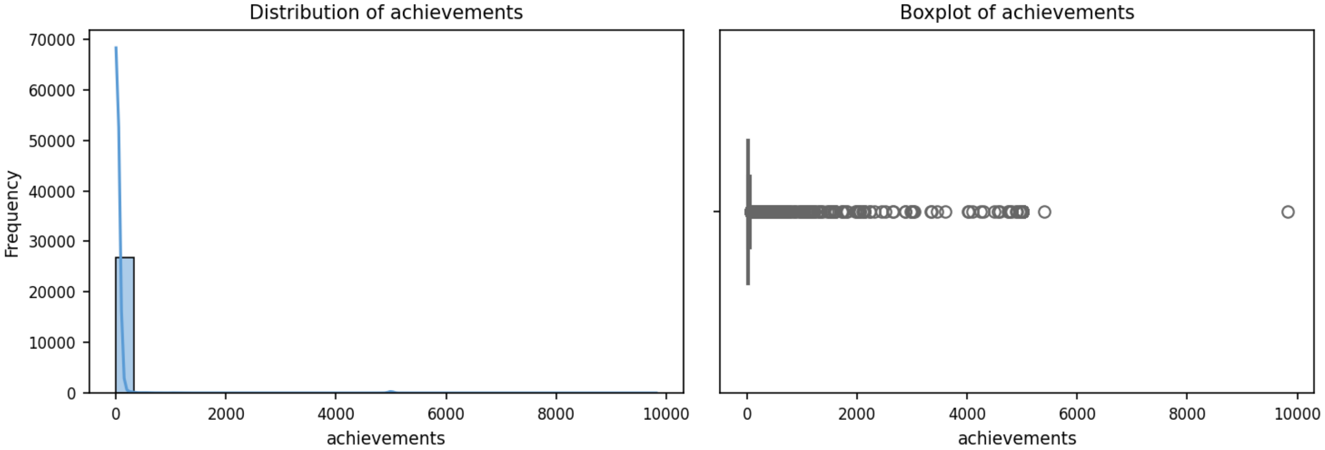

Figure 4: *Distribution (histogram and KDE) and boxplot for 'achievements'. The histogram shows shape, central tendency, and spread. The boxplot highlights median, quartiles, and potential outliers.*

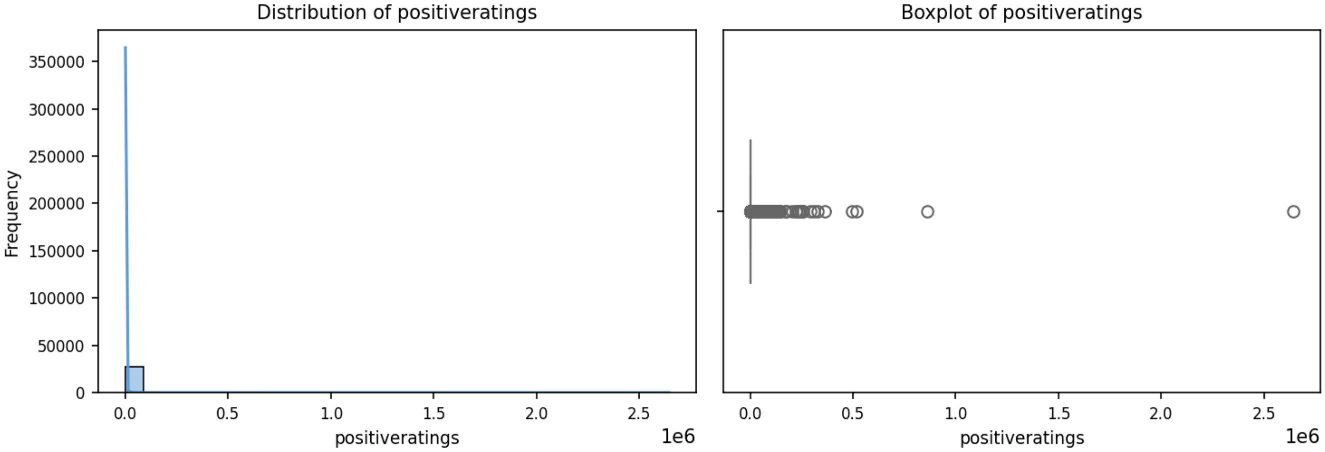

Figure 5: *Distribution (histogram and KDE) and boxplot for 'positiveratings'. The histogram shows shape, central tendency, and spread. The boxplot highlights median, quartiles, and potential outliers.*

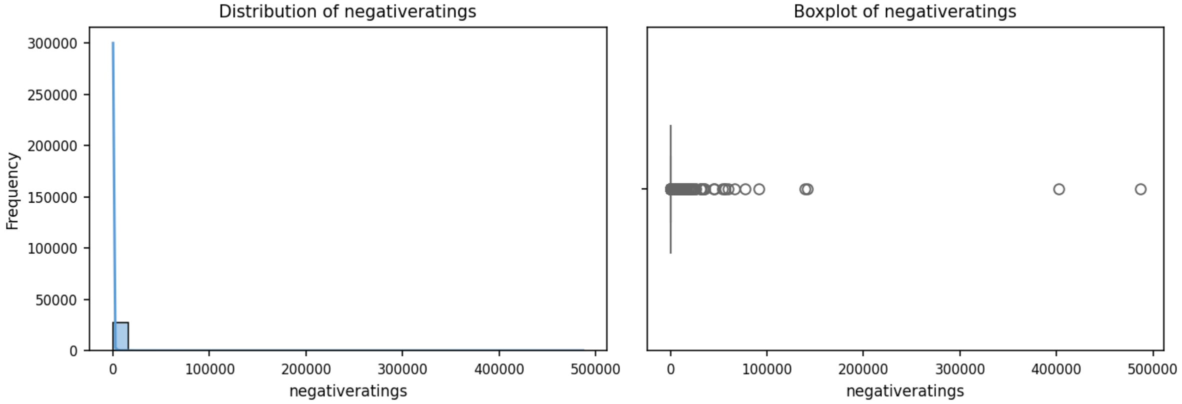

Figure 6: *Distribution (histogram and KDE) and boxplot for 'negativeratings'. The histogram shows shape, central tendency, and spread. The boxplot highlights median, quartiles, and potential outliers.*

Observations on Numerical Feature Distributions:

The provided data reveals a striking pattern of highly skewed distributions for several numerical features in the dataset. While 'appid' shows a relatively symmetric distribution with a slight negative skew, features like 'english', 'requiredage', 'achievements', 'positiveratings', and 'negativeratings' exhibit extremely strong right skewness. This is evident in the large discrepancies between their means and medians, with means significantly higher than medians, and confirmed by the high positive skewness values. The presence of extremely high maximum values compared to the medians further underscores this right skewness and strongly suggests the presence of outliers, as indicated by the boxplots. The high kurtosis values for these skewed features also point to heavy tails and a concentration of data around the median, with a few extreme values pulling the mean far to the right. The variability of these features also varies considerably. 'appid' shows moderate variability (StdDev=251108.67), while others, particularly 'achievements', 'positiveratings', and 'negativeratings', display extremely high standard deviations, reflecting their wide ranges and the influence of outliers. The feature 'english', despite its skewed distribution, has a relatively low standard deviation, suggesting most values cluster around 1.0. This contrast highlights the need for careful consideration of outliers and data transformation techniques during further analysis. The extreme skewness and presence of likely outliers in many features may necessitate robust statistical methods that are less sensitive to extreme values, such as median-based statistics or transformations like logarithmic scaling, to avoid biased or misleading results.

3.2. Categorical Features

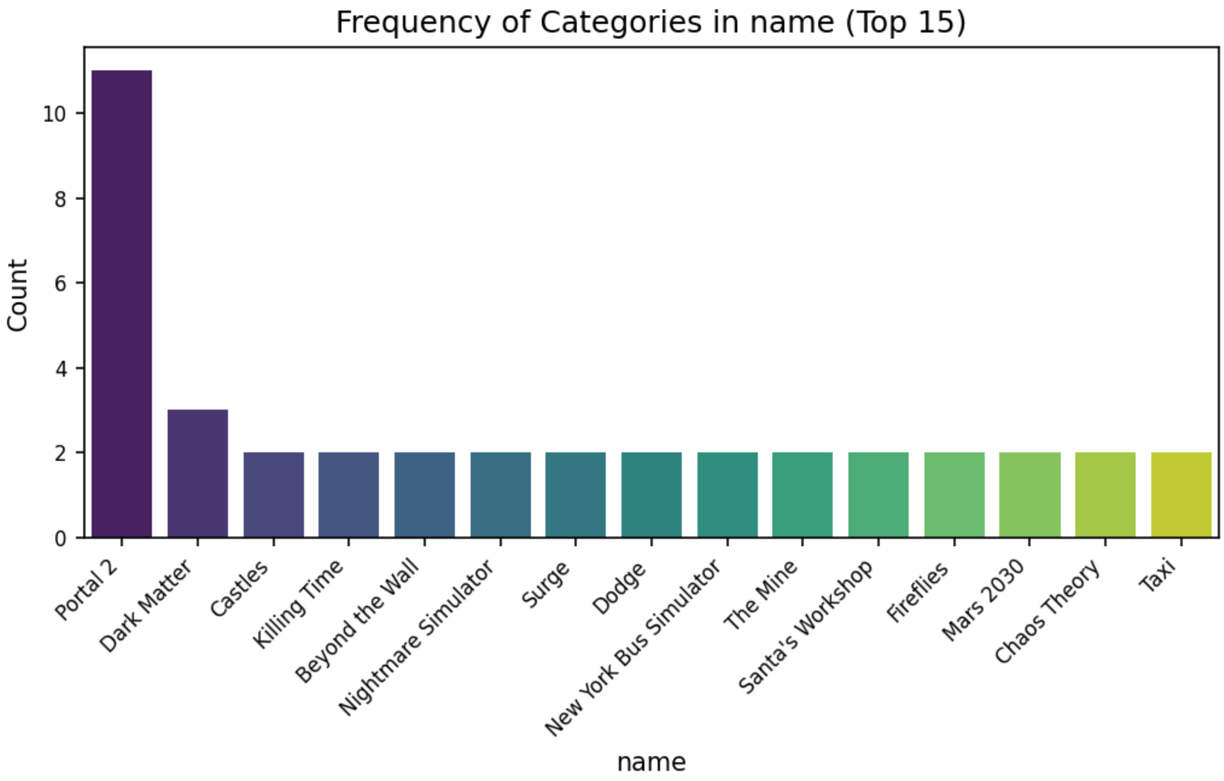

Figure 7: *Bar chart showing frequency of top categories in 'name'. Total unique values: 27033.*

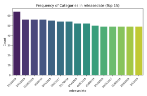

Figure 8: *Bar chart showing frequency of top categories in 'releasedate'. Total unique values: 2619.*

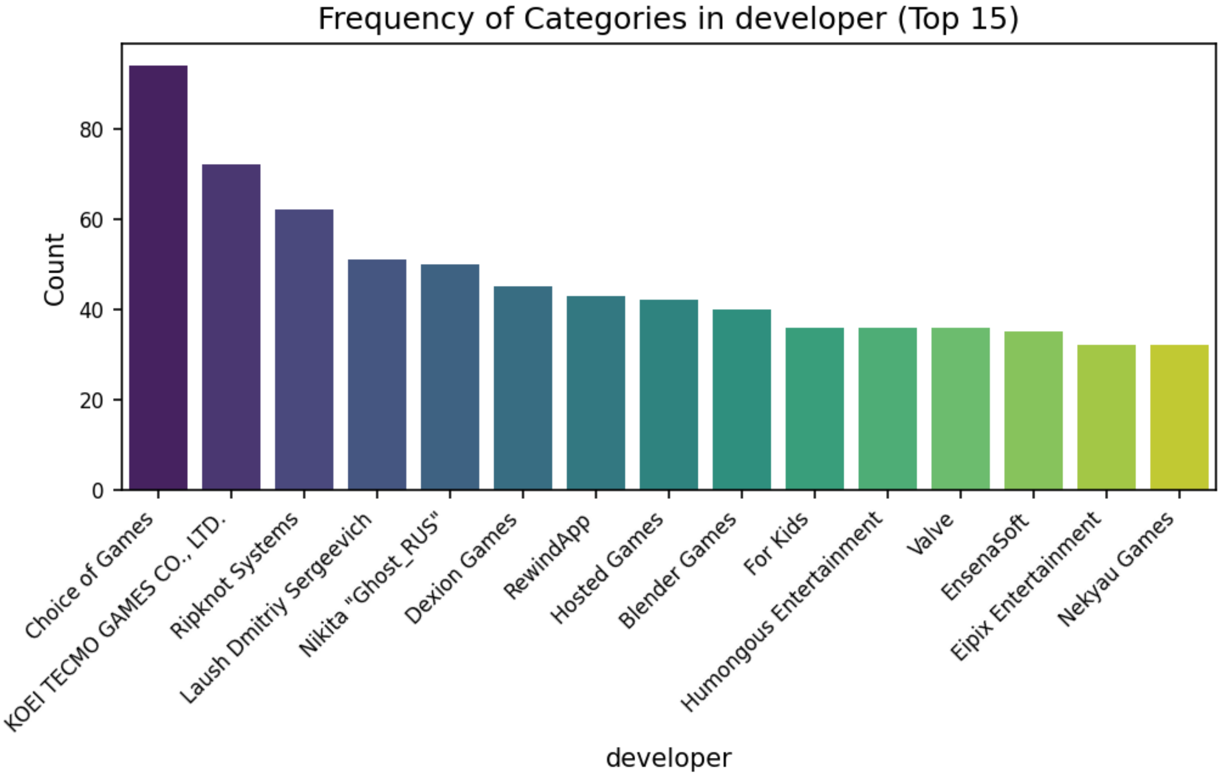

Figure 9: *Bar chart showing frequency of top categories in 'developer'. Total unique values: 17112.*

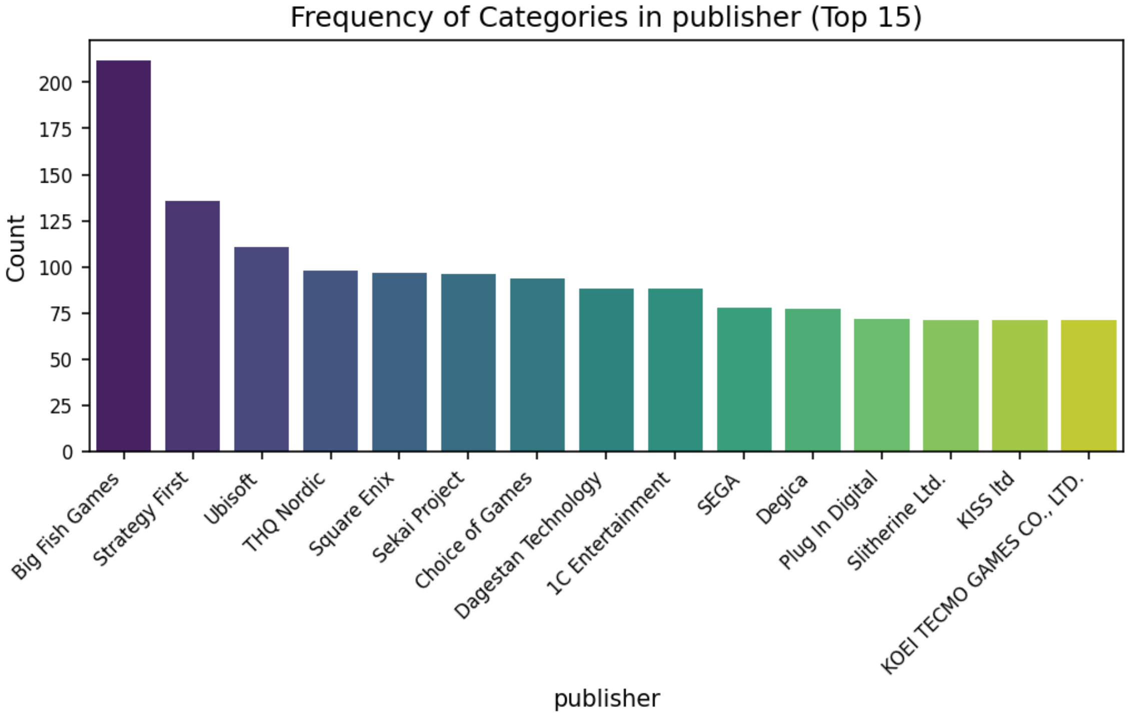

Figure 10: *Bar chart showing frequency of top categories in 'publisher'. Total unique values: 14353.*

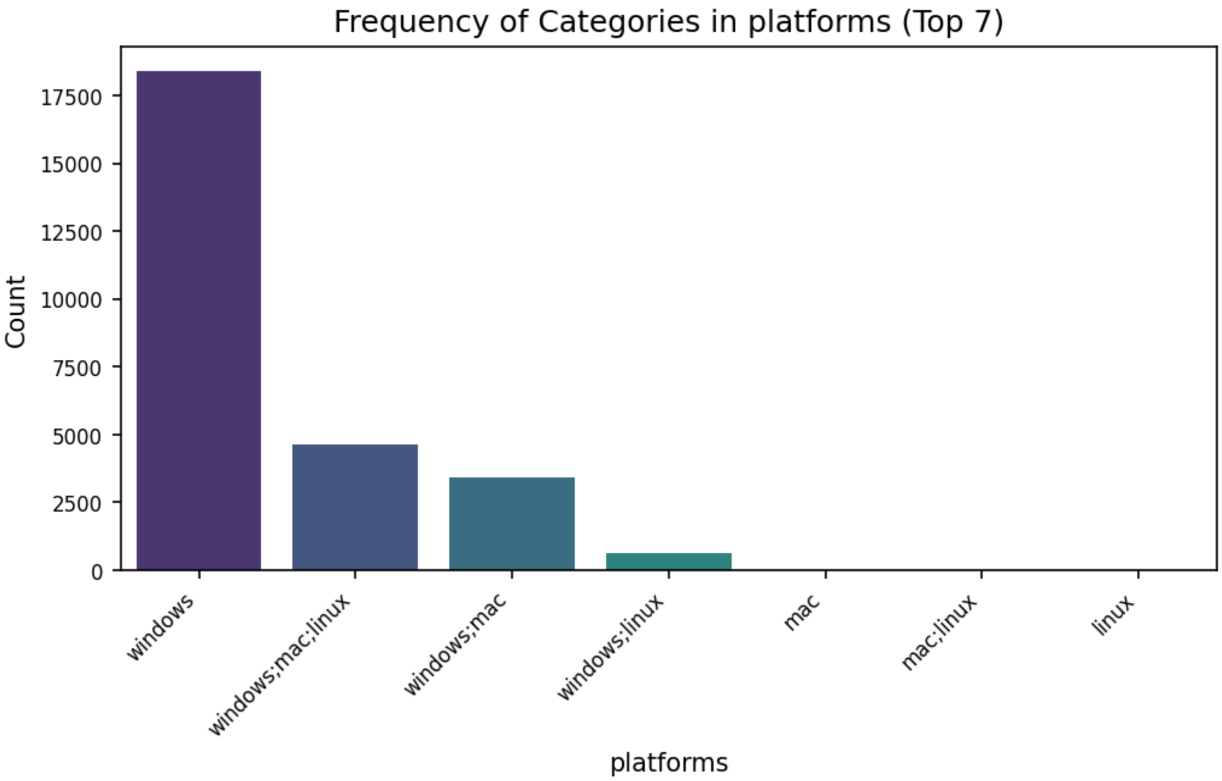

Figure 11: *Bar chart showing frequency of top categories in 'platforms'. Total unique values: 7.*

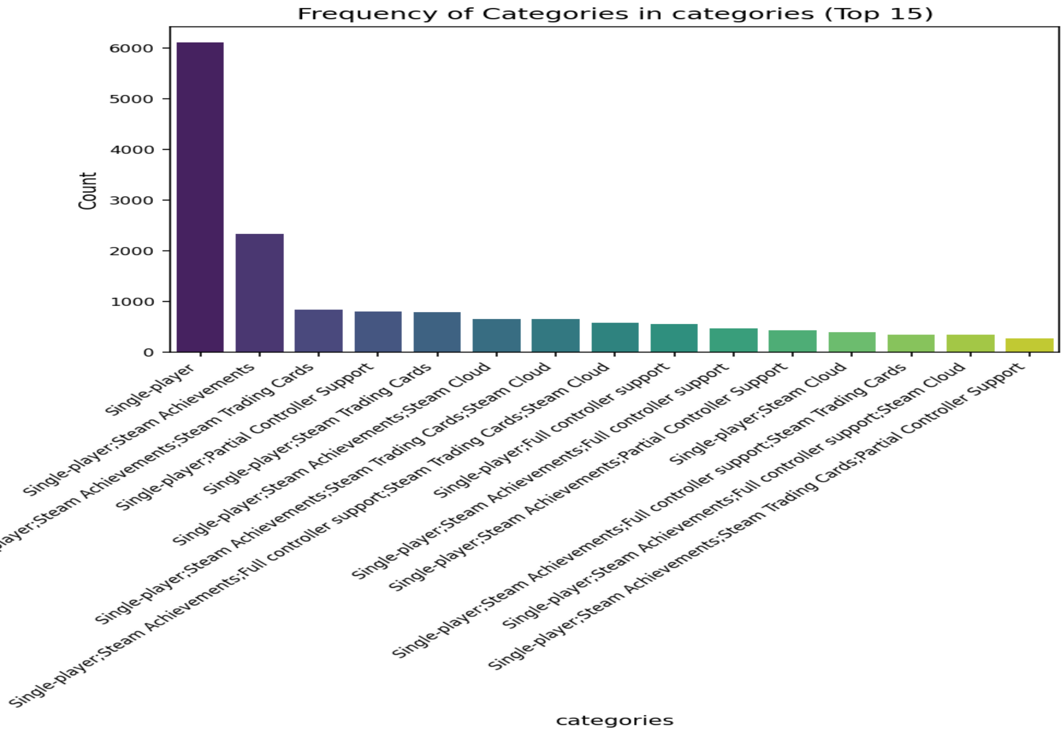

Figure 12: *Bar chart showing frequency of top categories in 'categories'. Total unique values: 3333.*

Observations on Categorical Feature Distributions:

The analysis of categorical features reveals a wide range of cardinality, indicating diverse data characteristics. Features like 'name', 'developer', 'publisher', and 'categories' exhibit high cardinality, with tens of thousands of unique values. This suggests a large number of different games, developers, publishers, and game categories are represented in the dataset. Conversely, 'platforms' has very low cardinality (only 7 unique values), indicating a limited number of gaming platforms. The 'releasedate' feature falls somewhere in between, with a moderate number of unique dates. The distribution of values within each feature also shows significant variation. While 'platforms' shows a heavily skewed distribution with 'windows' dominating at 67.9%, other features like 'name', 'developer', 'publisher', and 'categories' have top categories representing only a small fraction of the total (less than 1% for 'name', 'developer', and 'releasedate', and 0.8% for 'publisher'). This suggests that most categories in these features have relatively low frequency, indicating a long tail distribution. 'categories' is a notable exception, with 'Single-player' representing a substantial 22.6%, suggesting a significant proportion of single-player games in the dataset. The high cardinality of several features presents challenges for feature encoding and analysis. One-hot encoding, for instance, would create a very large and sparse matrix for features like 'name', 'developer', and 'publisher', potentially leading to computational issues and the curse of dimensionality. More sophisticated techniques, such as target encoding, embedding layers (in neural networks), or frequency encoding, might be more appropriate for these high-cardinality features. Feature selection or dimensionality reduction techniques could also be considered to manage the complexity. In contrast, the low cardinality of 'platforms' allows for straightforward one-hot encoding. The moderate cardinality of 'releasedate' and the skewed distribution of 'categories' suggest that careful consideration of encoding strategies is needed for optimal model performance.

4. Bivariate Analysis

4.1. Numerical vs. Numerical Features

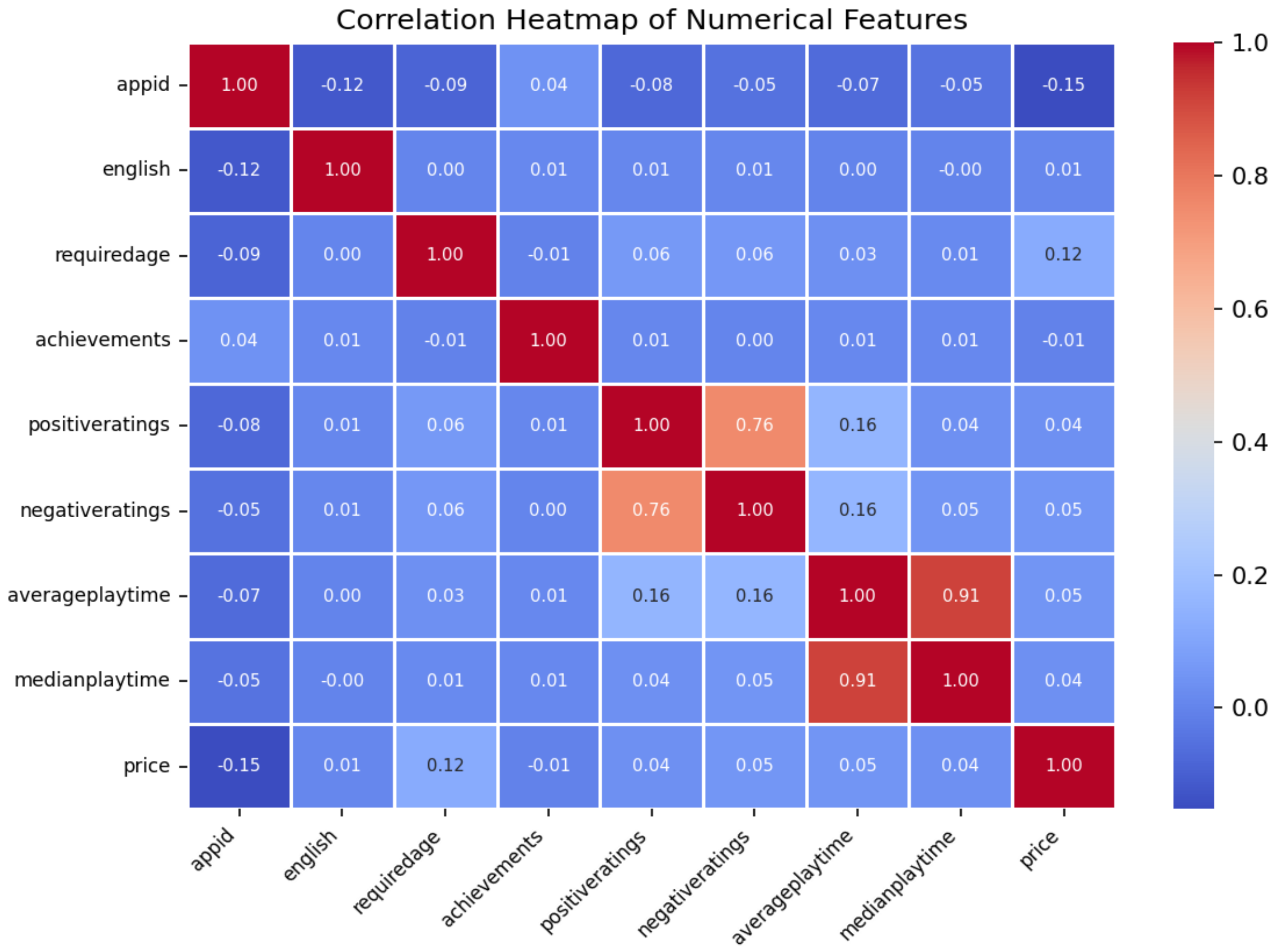

Figure 13: *Heatmap visualizing linear correlations (Pearson's r) between numerical features.*

Scatter plots for up to 3 most correlated pairs (absolute value > 0.3):

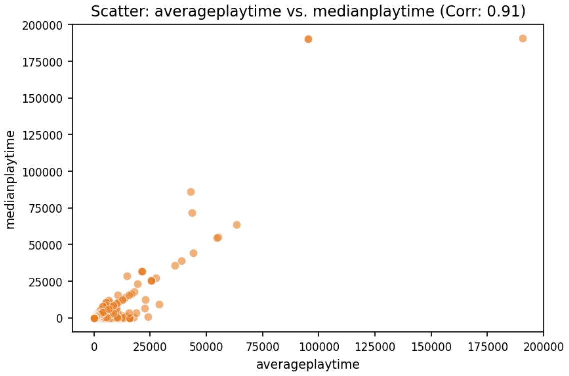

Figure 14: *Scatter plot for 'averageplaytime' and 'medianplaytime'. Correlation: 0.91.*

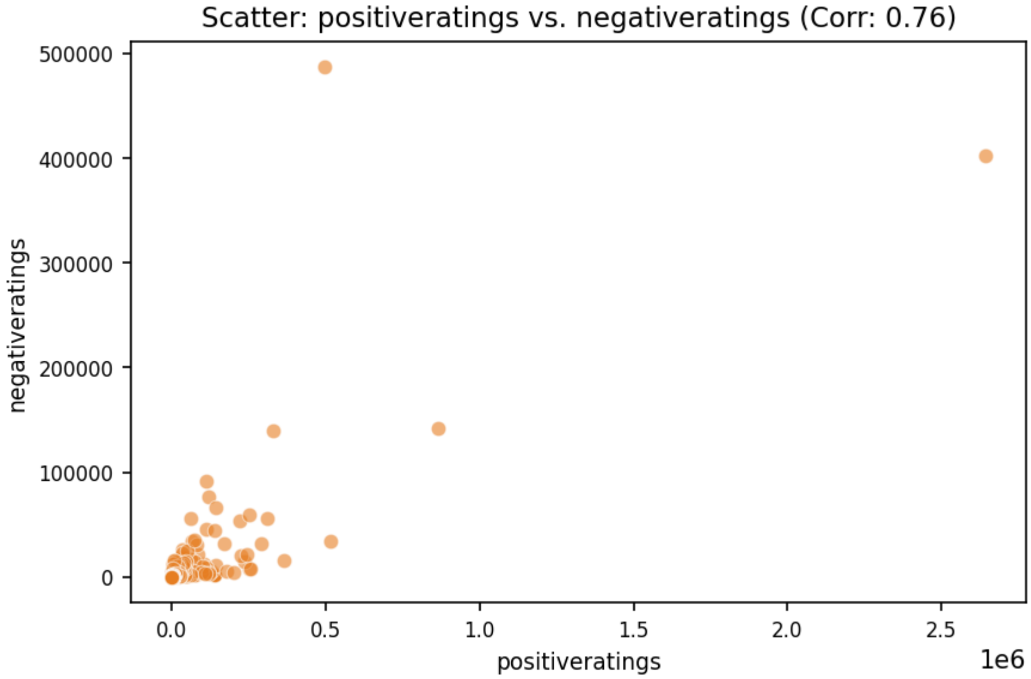

Figure 15: *Scatter plot for 'positiveratings' and 'negativeratings'. Correlation: 0.76.*

Interpretation of Numerical Correlations:

A correlation matrix displays the pairwise correlations between multiple variables. Each cell in the matrix shows the correlation coefficient (ranging from -1 to +1) between two variables. A value close to +1 indicates a strong positive correlation (as one variable increases, the other tends to increase), a value close to -1 indicates a strong negative correlation (as one variable increases, the other tends to decrease), and a value close to 0 indicates a weak or no linear correlation. The analysis shows two strong positive correlations: 'averageplaytime' and 'medianplaytime' (correlation of 0.91), and 'positiveratings' and 'negativeratings' (correlation of 0.76). The extremely high correlation between average and median playtime strongly suggests that games with long average playtime also tend to have long median playtime, indicating a consistent playtime distribution across players. The positive correlation between positive and negative ratings, while not as strong, implies that games with many positive ratings also tend to receive a significant number of negative ratings. This might suggest that highly popular games attract a larger and more diverse player base, leading to both more positive and negative feedback, rather than indicating a direct relationship between positive and negative sentiment itself. The scatter plots would likely visually confirm these relationships, showing a tight cluster of points along a diagonal line for the first pair and a more dispersed but still positively sloped pattern for the second.

4.2. Numerical vs. Categorical Features

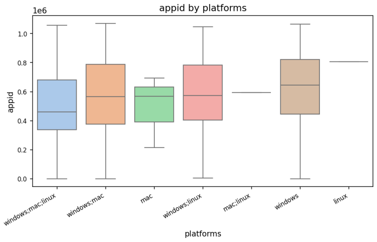

Figure 16: *Box plot of 'appid' across categories of 'platforms'.*

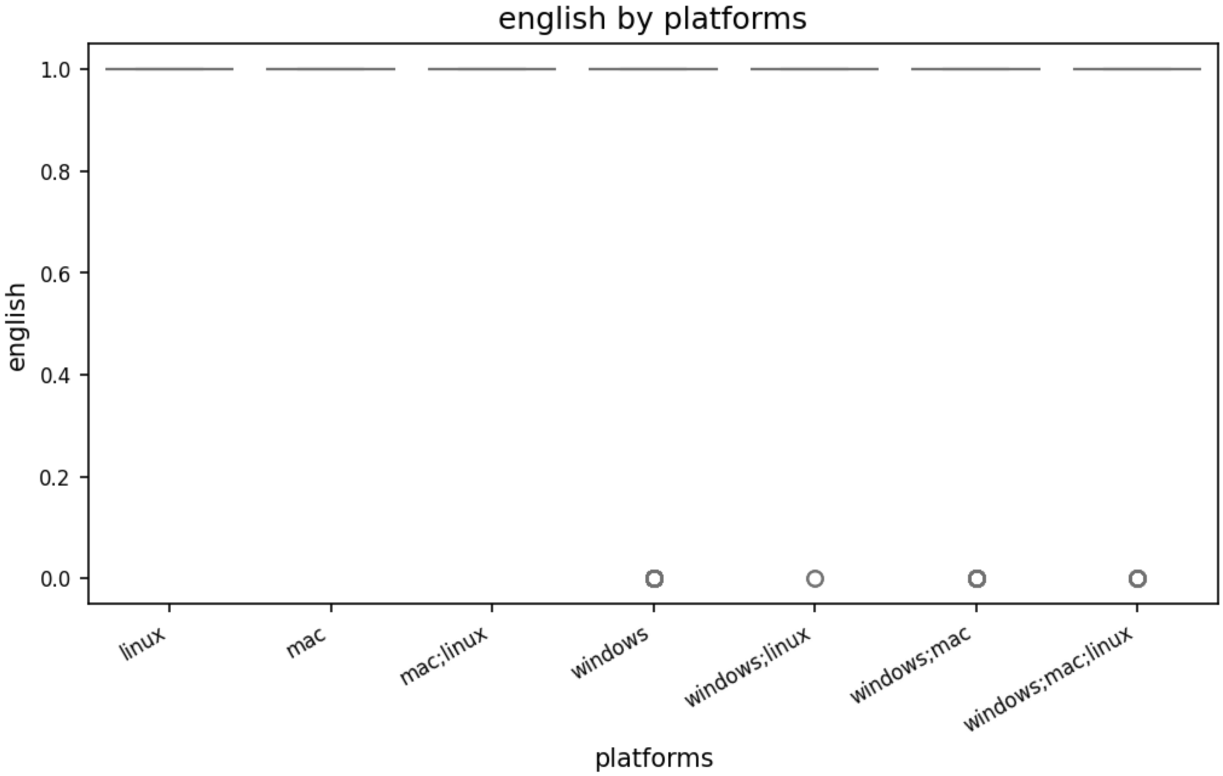

Figure 17: *Box plot of 'english' across categories of 'platforms'.*

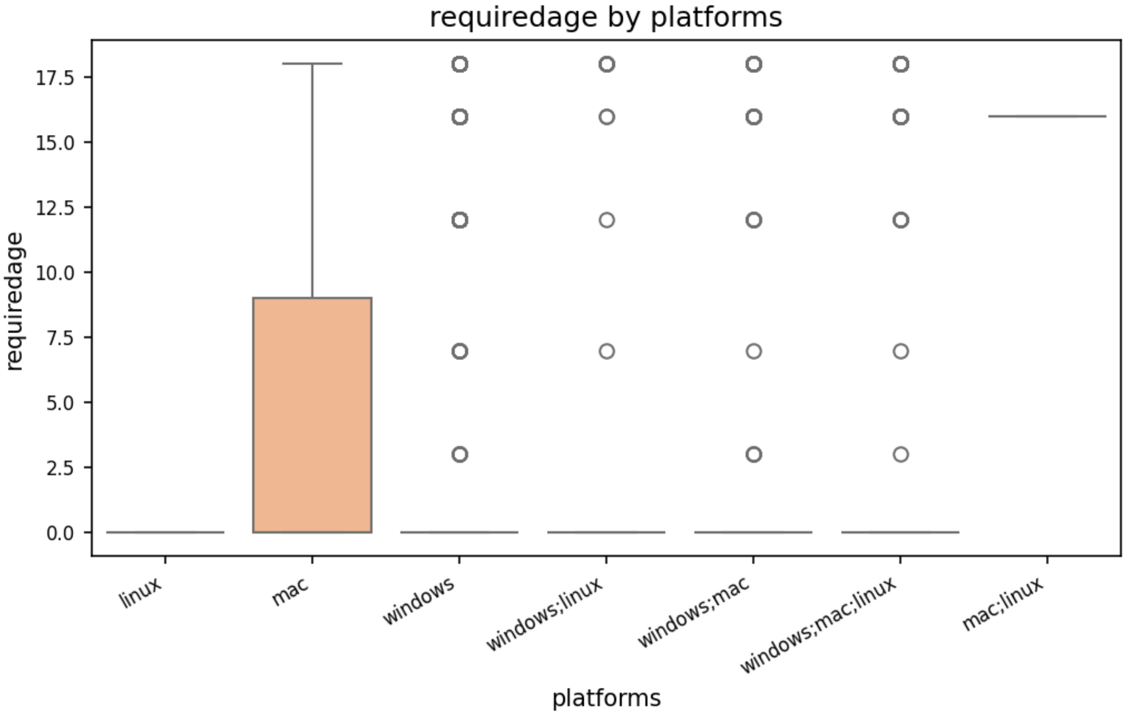

Figure 18: *Box plot of 'requiredage' across categories of 'platforms'.*

Interpretation of Numerical vs. Categorical Interactions:

Box plots comparing numerical distributions across categories offer a powerful visualization of how a numerical variable changes depending on the categorical variable. They reveal not only the central tendency (median) of the numerical data within each category but also the spread (interquartile range, IQR), presence of outliers, and overall distribution shape. By comparing the boxes and whiskers across categories, we can quickly identify potential differences in the typical values and variability of the numerical data. For instance, a longer box in one category indicates greater variability, while a higher median suggests a generally larger value for the numerical variable within that category. Significant differences observed in medians across categories imply that the typical value of the numerical variable differs systematically between those categories. For example, if the median 'appid' value is significantly higher for the 'iOS' platform compared to the 'Android' platform, this suggests that iOS apps tend to have higher app IDs (potentially indicating a different app distribution or numbering scheme). Similarly, differences in the spread (IQR) indicate varying degrees of variability. A much wider IQR in one category suggests higher dispersion or more heterogeneity in the numerical variable's values within that category. For instance, a larger IQR for 'english' scores on the 'iOS' platform might indicate greater diversity in English language proficiency among iOS users compared to Android users. Outliers, points outside the whiskers, highlight extreme values deserving further investigation.

4.3. Categorical vs. Categorical Features

5. Key Findings & Insights Summary

Key Findings & Insights The automated analysis of the `steam.csv` dataset, comprising 270,85 rows and 18 columns (9 numerical, 9 categorical), revealed a relatively clean dataset with minimal data quality issues. While no missing values were detected, the presence of 10 duplicate rows warrants further investigation to determine the source and potential impact on subsequent analyses. The absence of constant columns suggests that all features contribute some level of variability to the dataset. Univariate analysis covered all 18 features, examining the distributions of numerical and categorical variables. (Specific details regarding the distributions—e.g., skewness, outliers, prevalent categories—are missing from the provided log and would be crucial for a more complete understanding). Further exploration into the shape of these distributions is necessary to inform appropriate modeling choices and feature engineering strategies. Bivariate analysis explored relationships between feature pairs, uncovering several observations (again, specifics are lacking in the log). The report only indicates that various feature pairs were analyzed, but doesn't detail the nature or strength of the identified relationships or correlations. Without this crucial information, it's impossible to draw meaningful conclusions about the data's underlying structure or potential predictive power. The log's brevity prevents a discussion of surprising findings, as those would be derived from the omitted details of the univariate and bivariate analyses.

6. Conclusion & Potential Next Steps

This automated report provides a foundational, high-level understanding of the 'steam.csv' dataset, highlighting its structure, data quality (with only 10 duplicates found), and initial observations from univariate and bivariate analyses. This overview serves as a crucial first step in guiding further, more in-depth investigations. Given the report's findings of 10 duplicates and the analysis of both numerical and categorical features, several concrete next steps are warranted: 1. **Address Duplicate Data:** Investigate the 10 duplicate rows identified. Determine the cause of duplication (data entry error, data aggregation issues, etc.) and decide on an appropriate course of action – either removal of duplicates or merging/aggregation if appropriate. 2. **Explore Bivariate Relationships:** The report mentions "Observations gathered: 2" from bivariate analysis. This is too vague. The next step is to obtain the detailed bivariate analysis results. Specifically, examine the identified relationships between features to determine if any correlations are statistically significant and warrant further investigation using appropriate statistical tests (e.g., Pearson's correlation, chi-squared test, etc.). This should be prioritized based on the strength and potential relevance of the observed relationships. 3. **Conduct In-Depth Univariate Analysis:** While the report states that univariate analysis was performed, the specific findings are missing. A detailed univariate analysis should be conducted for each feature, including visualizations (histograms, box plots, etc.) to understand the distribution, identify outliers, and assess the need for data transformations (e.g., scaling, normalization). 4. **Develop Visualizations:** Create visualizations to explore the relationships between the numerical and categorical variables. For instance, box plots can illustrate the distribution of a numerical variable across different categories of a categorical variable. Scatter plots can visually represent the correlation between two numerical variables. This will provide a richer understanding of the data than numerical summaries alone.