Automated Data Analysis Report (via Gemini): Temp Steam

Executive Summary

This report summarizes the initial automated exploratory data analysis (EDA) of the `temp_steam.csv` dataset, containing 27,085 rows and 18 columns (9 numerical, 9 categorical). The dataset exhibits good data quality, with no missing values but 10 duplicate rows identified. Preliminary univariate and bivariate analyses, including descriptive statistics and visual inspections, were conducted on all features. Key findings include the absence of constant columns and the identification of a relatively small number of duplicate entries. Further investigation is needed to understand the nature of these duplicates and their potential impact on subsequent analyses. No immediately striking patterns emerged from the bivariate analysis, warranting further exploration of potential relationships between features. This initial EDA scan provides a foundational understanding of the dataset's structure and quality. The findings will inform subsequent, more in-depth analyses focusing on specific hypotheses and potentially requiring data cleaning and feature engineering to extract actionable insights.

1. Data Overview

This report provides an initial automated analysis of the dataset from 'temp_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 is balanced between numerical and categorical features, which is a fairly typical mix. The absence of datetime features might limit the ability to perform time-series analysis or investigate trends over time, but allows for other types of analysis such as clustering or classification.

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 dataset of 27,085 rows with generally high quality. The absence of missing values is a significant positive, indicating a complete dataset ready for analysis. The presence of only 10 duplicate rows (0.04%) represents a negligible issue, easily rectified with minimal impact on the overall dataset. The lack of constant columns ensures that no variables provide redundant or unhelpful information. However, the identification of two quasi-constant columns, 'english' and 'requiredage', warrants attention. While not entirely problematic, their high dominance in a single value suggests potential limitations in their analytical utility. The quasi-constant columns 'english' and 'requiredage' could limit the effectiveness of certain modeling techniques. For example, algorithms relying on variance across features might perform poorly if heavily weighted by these columns. Insights derived from analyses heavily reliant on these variables might be skewed or lack sufficient variability to draw robust conclusions. The low percentage of duplicates, while insignificant in the grand scheme, should still be investigated to understand their origin and ensure they are not indicative of a larger data entry or collection problem. To address the identified issues, the duplicate rows should be removed. For the quasi-constant columns, several strategies are possible depending on the research question. One approach is to remove them entirely if they are deemed irrelevant or not contributing meaningfully to the analysis. Alternatively, if they represent meaningful information despite their limited variance, they could be recoded or transformed to better capture the subtle variations present. Further investigation into the meaning and potential value of these columns is crucial before deciding on a course of action. Finally, a review of the data collection process could help prevent similar issues in future data gathering.

3. Univariate Analysis

3.1. Numerical Features

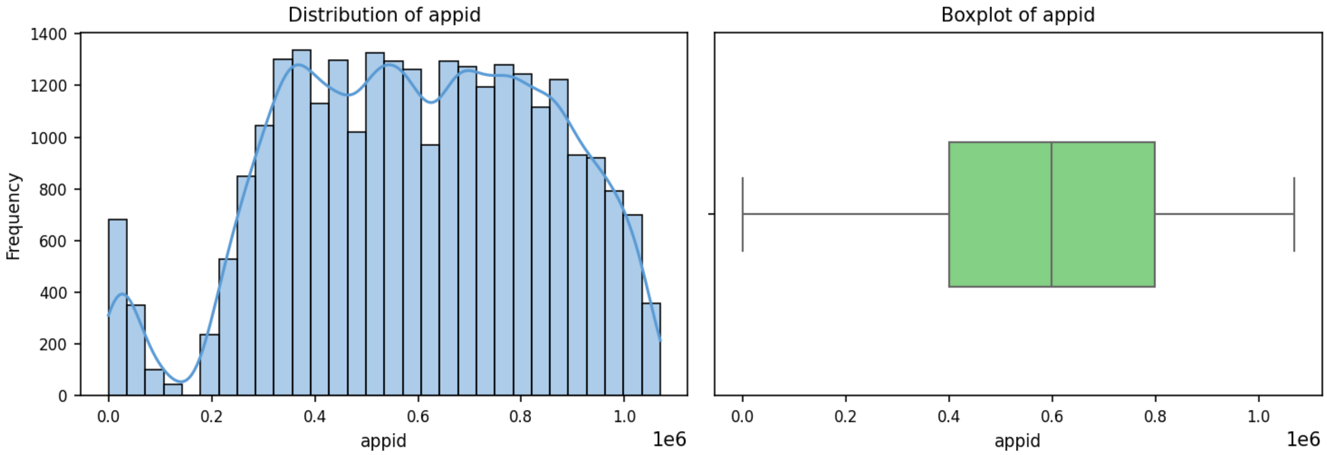

Figure 1: *Distribution (histogram and KDE) and boxplot for 'appid'. The histogram shows shape, central tendency, and spread. The boxplot highlights median, quartiles, and potential outliers.*

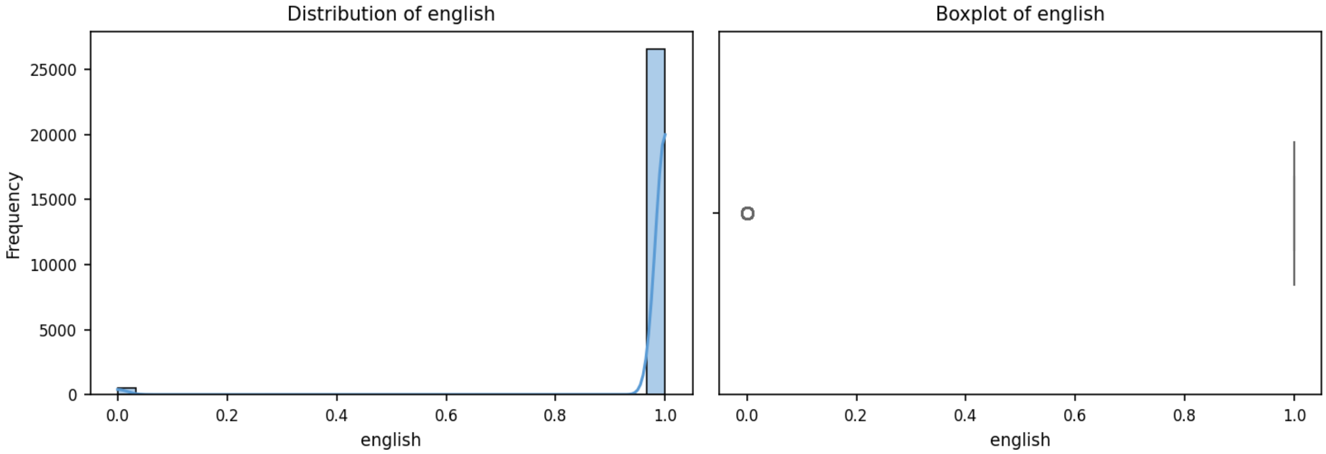

Figure 2: *Distribution (histogram and KDE) and boxplot for 'english'. The histogram shows shape, central tendency, and spread. The boxplot highlights median, quartiles, and potential outliers.*

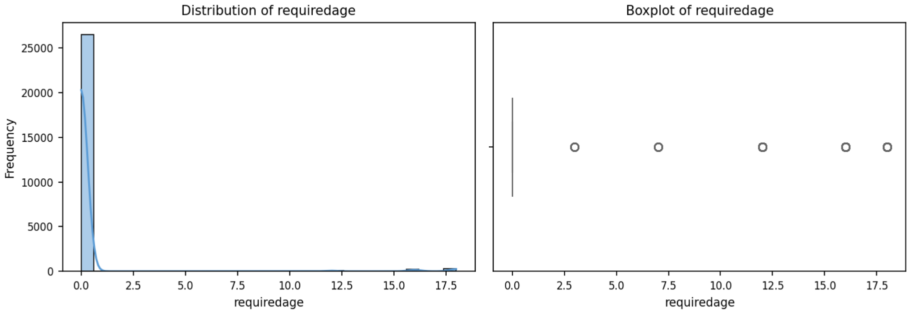

Figure 3: *Distribution (histogram and KDE) and boxplot for 'requiredage'. The histogram shows shape, central tendency, and spread. The boxplot highlights median, quartiles, and potential outliers.*

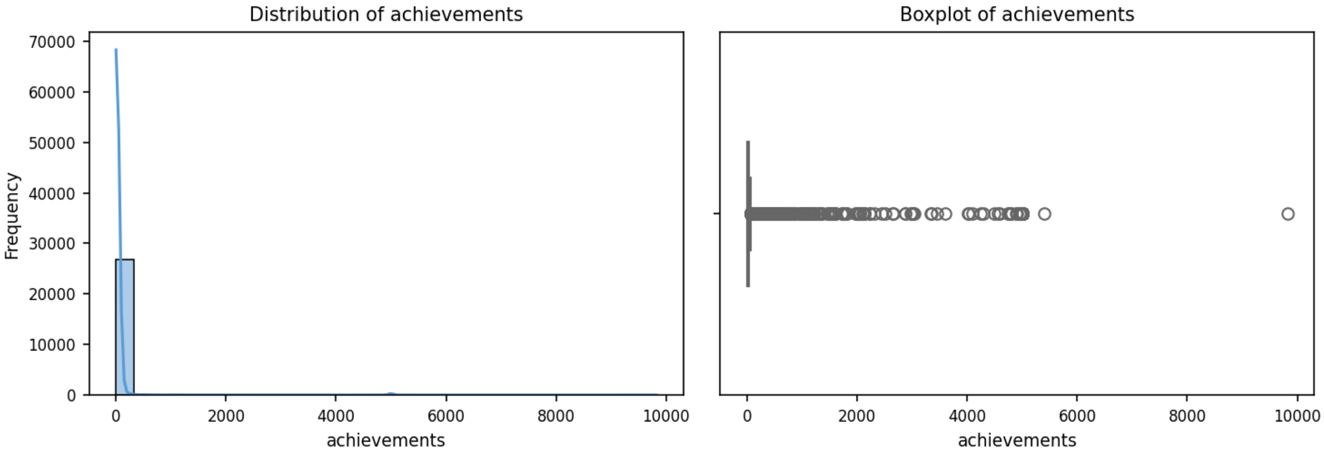

Figure 4: *Distribution (histogram and KDE) and boxplot for 'achievements'. The histogram shows shape, central tendency, and spread. The boxplot highlights median, quartiles, and potential outliers.*

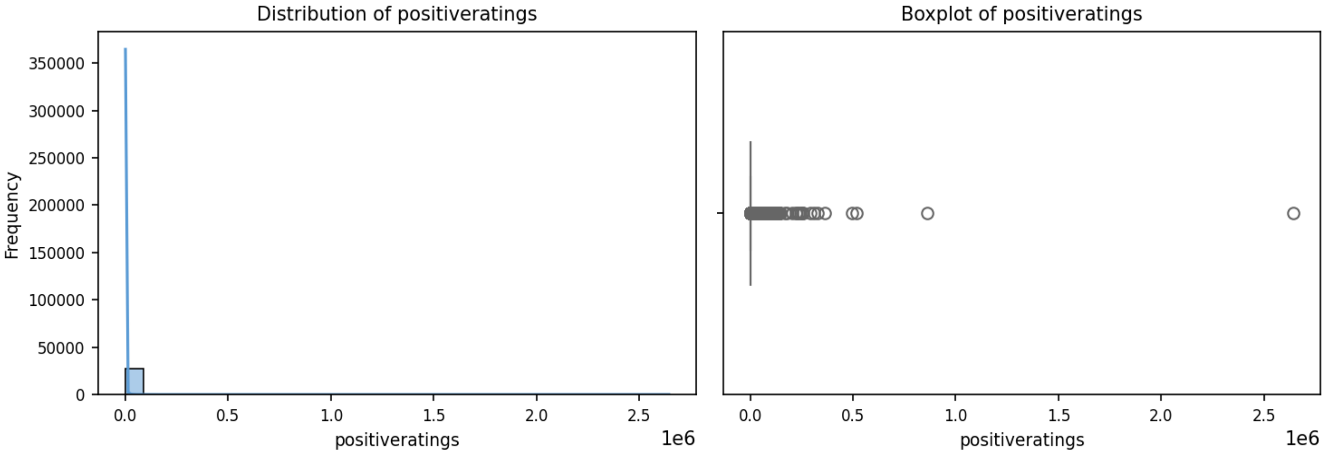

Figure 5: *Distribution (histogram and KDE) and boxplot for 'positiveratings'. The histogram shows shape, central tendency, and spread. The boxplot highlights median, quartiles, and potential outliers.*

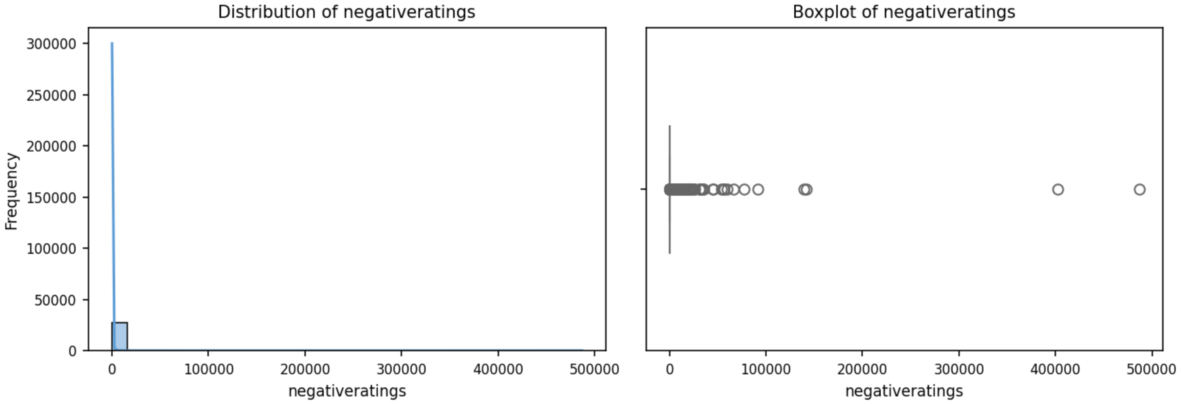

Figure 6: *Distribution (histogram and KDE) and boxplot for 'negativeratings'. The histogram shows shape, central tendency, and spread. The boxplot highlights median, quartiles, and potential outliers.*

Observations on Numerical Feature Distributions:

The analysis reveals highly skewed distributions for several numerical features in the dataset, indicating a significant departure from normality. Features like 'achievements', 'positiveratings', and 'negativeratings' exhibit extreme right skewness, with massive differences between their means and medians. This suggests the presence of a long tail of high values, far exceeding the typical values. The extreme skewness and kurtosis values (far exceeding 3 and 0 respectively, indicating heavy tails and peakedness) for these features, combined with the boxplot indications of outliers, strongly suggest the need for careful consideration of data transformation techniques (e.g., logarithmic transformation) or robust statistical methods to mitigate the influence of these extreme values in subsequent analyses. 'requiredage' also shows significant right skewness, though less extreme than the aforementioned features. In contrast, 'appid' shows a relatively symmetric distribution, although the boxplot hints at potential outliers. The standard deviation of 'appid' is substantial compared to its mean, indicating considerable variability in application IDs. The feature 'english' displays a strong left skew, with a mean close to 1 but a median of 1, suggesting a concentration of values near 1 and a few values significantly lower. This feature also shows evidence of outliers. The high standard deviation in features like 'achievements', 'positiveratings', and 'negativeratings' reflects the large spread and range of values observed, further emphasizing the dominance of a few extremely high values. Overall, the data suggests a need for careful preprocessing before applying many standard statistical methods. The extreme skewness and presence of many likely outliers in several features necessitate robust techniques to avoid biased results. Understanding the nature of these extreme values (e.g., are they errors or genuine, extreme observations?) is crucial for choosing appropriate data handling strategies.

3.2. Categorical Features

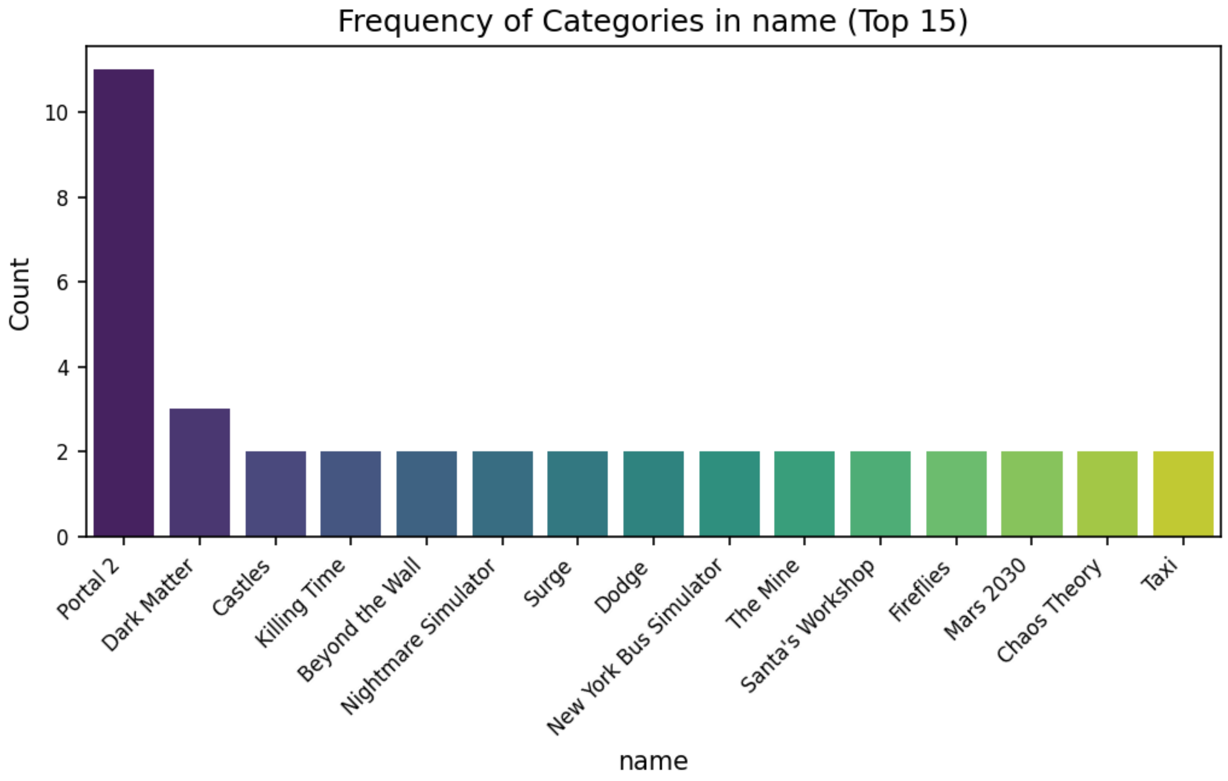

Figure 7: *Bar chart showing frequency of top categories in 'name'. Total unique values: 27033.*

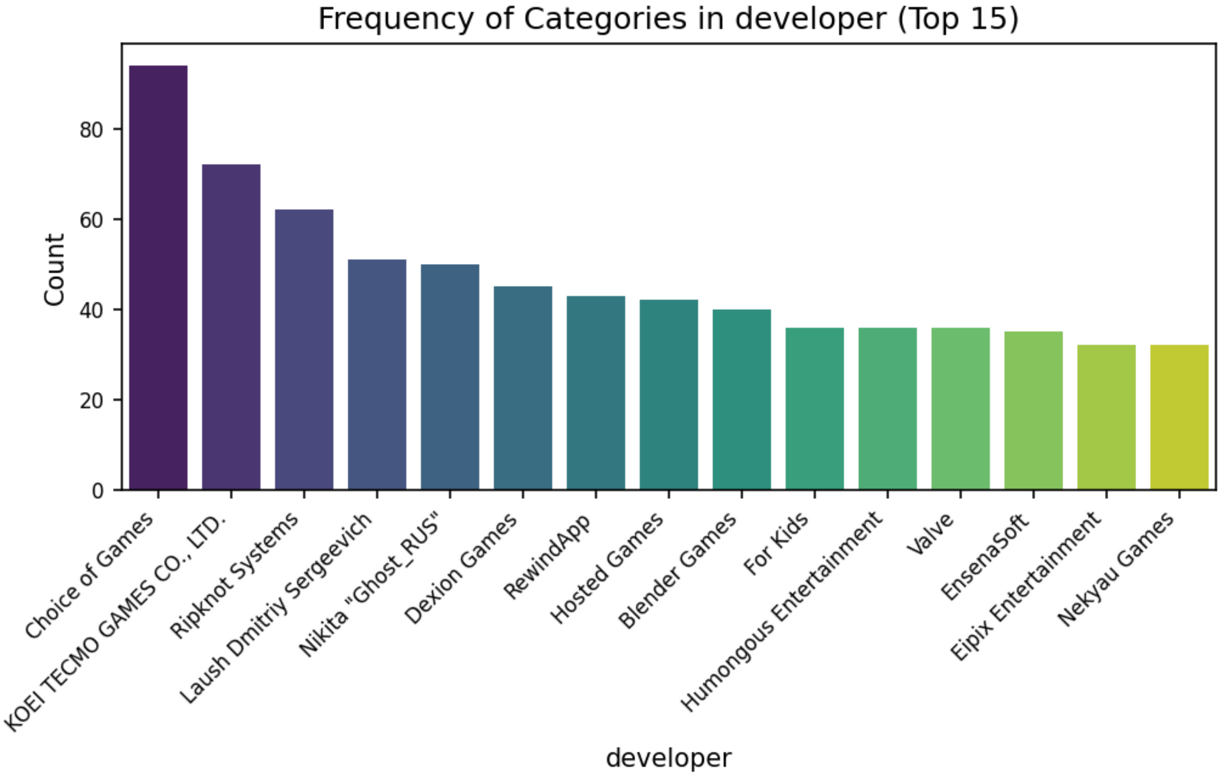

Figure 9: *Bar chart showing frequency of top categories in 'developer'. Total unique values: 17112.*

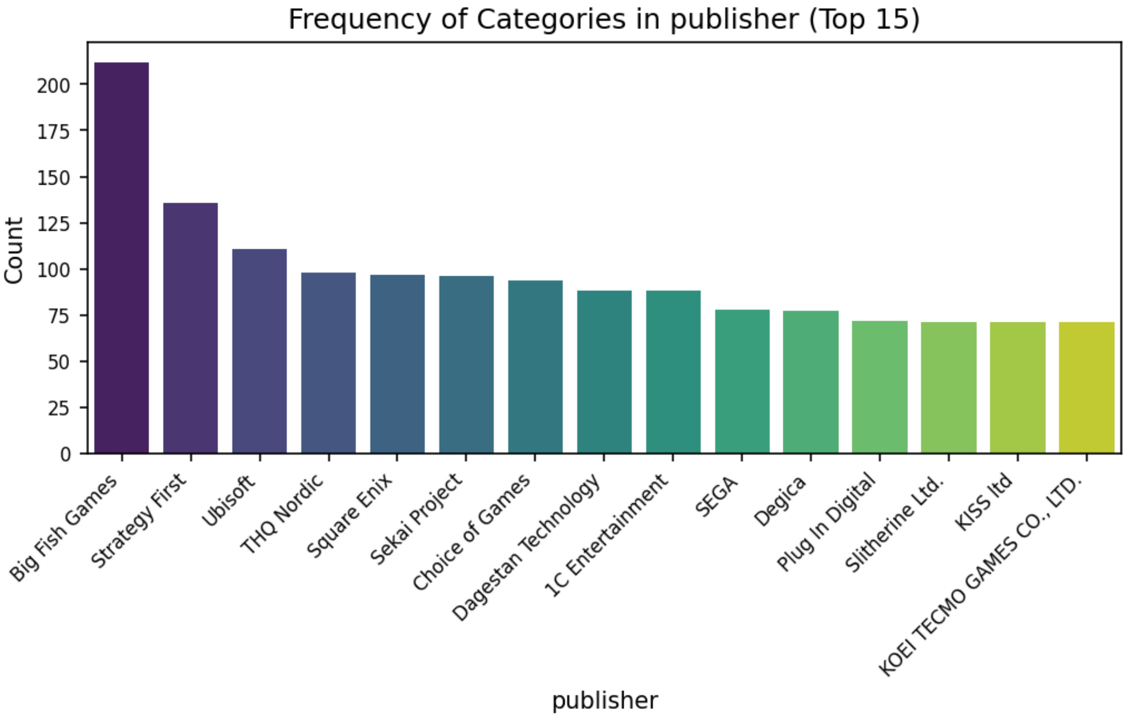

Figure 10: *Bar chart showing frequency of top categories in 'publisher'. Total unique values: 14353.*

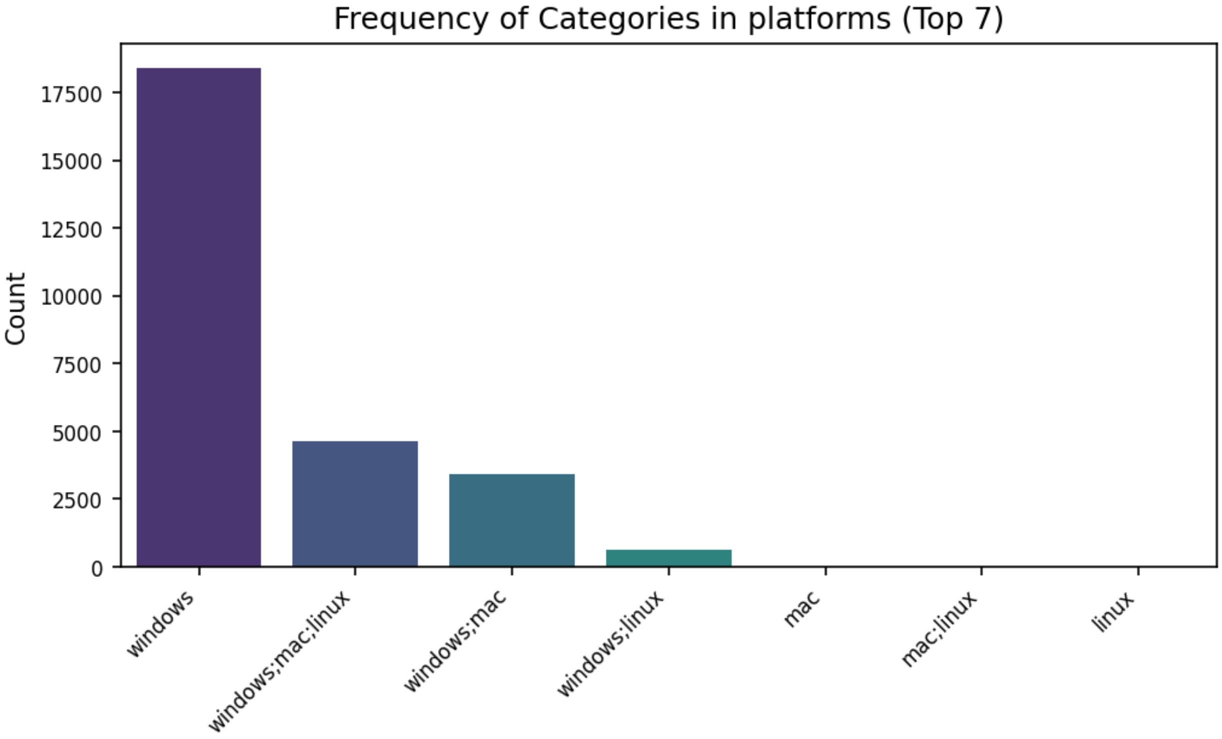

Figure 11: *Bar chart showing frequency of top categories in 'platforms'. Total unique values: 7.*

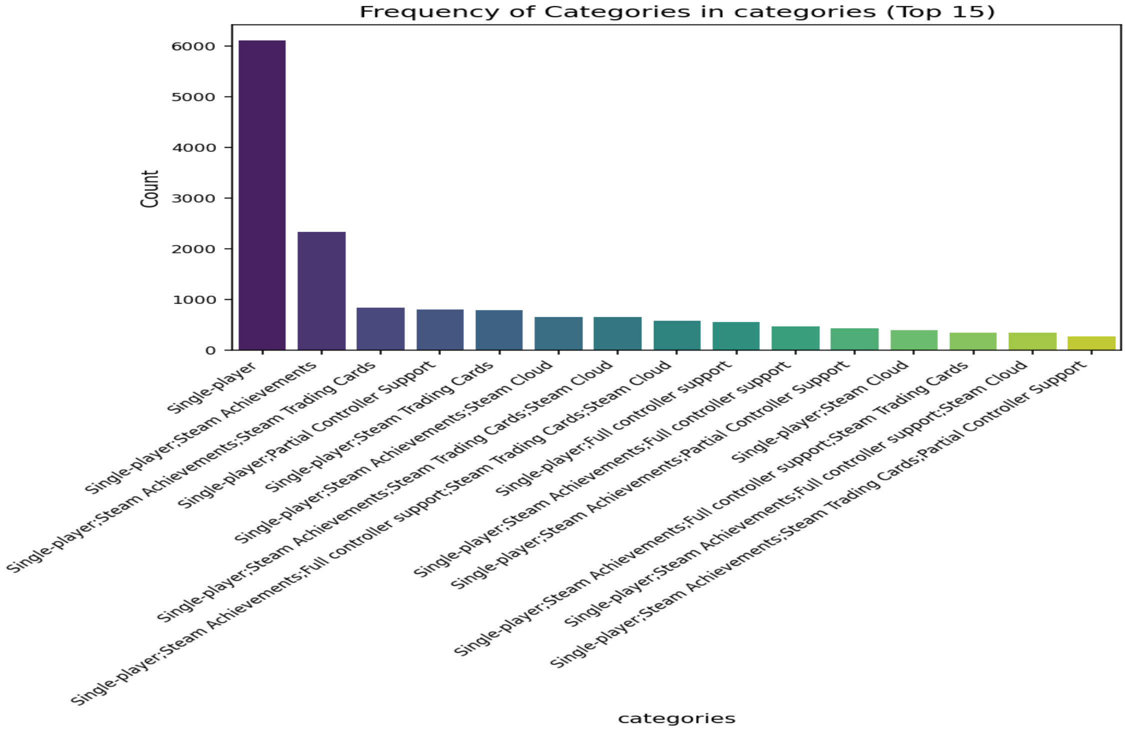

Figure 12: *Bar chart showing frequency of top categories in 'categories'. Total unique values: 3333.*

Observations on Categorical Feature Distributions:

The analysis of categorical features reveals a wide range of cardinality, indicating diverse data characteristics. Features like 'name', 'releasedate', 'developer', and 'publisher' exhibit very high cardinality (thousands of unique values), suggesting a large number of distinct games, developers, and publishers in the dataset. In contrast, 'platforms' has a relatively low cardinality (7 unique values), indicating a limited set of gaming platforms. The high cardinality features present a challenge for model training due to potential sparsity and the curse of dimensionality; techniques like embedding or feature hashing might be necessary for effective feature encoding. While most high-cardinality features lack a single dominant category (e.g., 'Portal 2' represents 0.0% of 'name'), 'platforms' shows a highly skewed distribution with 'windows' comprising 67.9% of the data. Similarly, 'categories' shows a notable skew, with 'Single-player' accounting for 22.6% of the entries. This uneven distribution in 'platforms' and 'categories' might indicate a bias towards Windows-based games and single-player titles within the dataset. This should be considered during analysis to avoid skewed results. For the high-cardinality features, strategies like frequency encoding or target encoding could be explored to handle the imbalanced distribution of categories. In summary, the dataset presents both high- and low-cardinality features, requiring careful consideration of feature encoding strategies. The high-cardinality features necessitate dimensionality reduction techniques to prevent overfitting and improve model performance. The skewed distributions in features like 'platforms' and 'categories' require attention during analysis to avoid biased interpretations. Further investigation into the underlying reasons for these skewed distributions could provide valuable insights into the dataset's composition and potential biases.

4. Bivariate Analysis

4.1. Numerical vs. Numerical Features

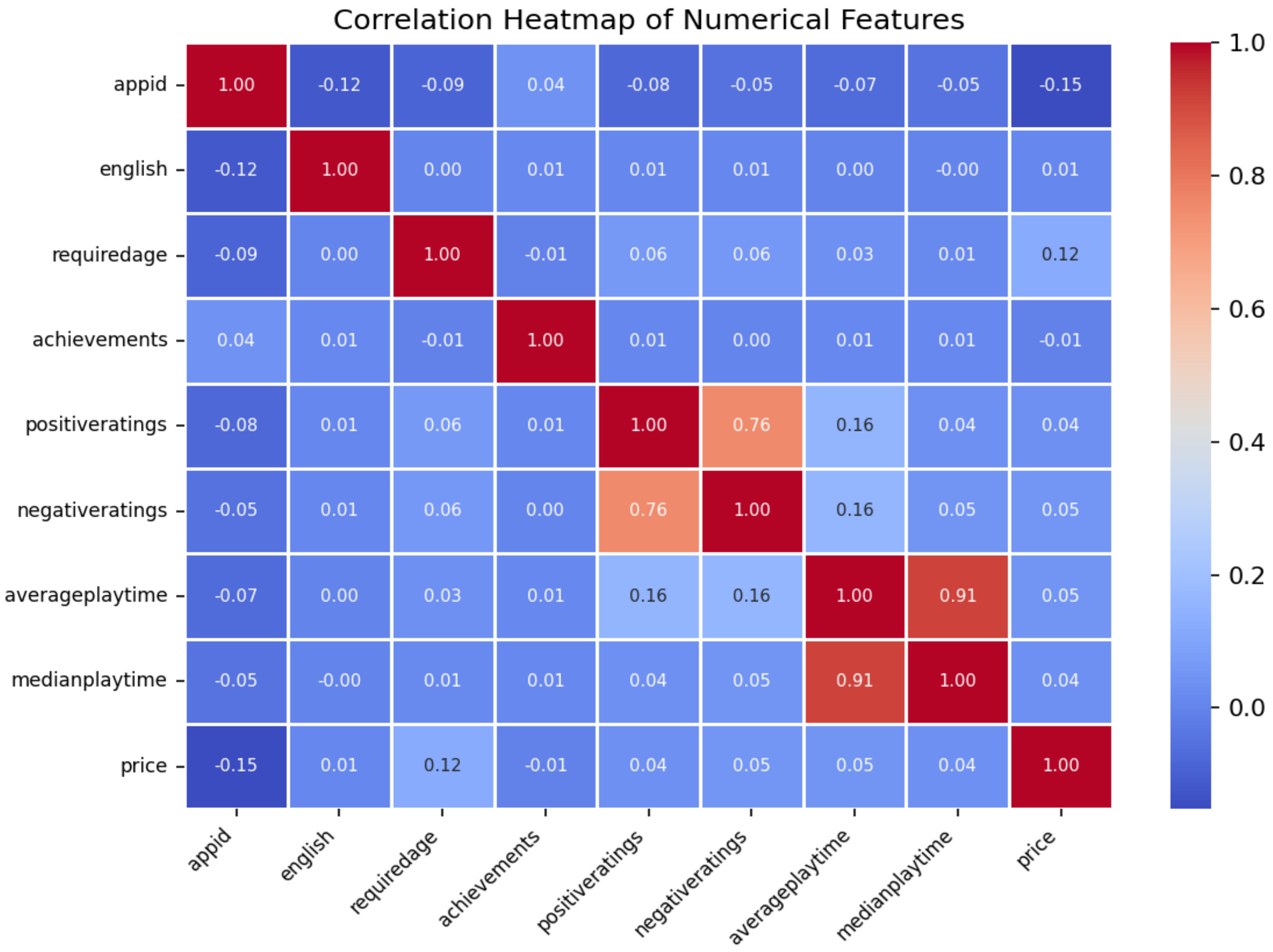

Figure 13: *Heatmap visualizing linear correlations (Pearson's r) between numerical features.*

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

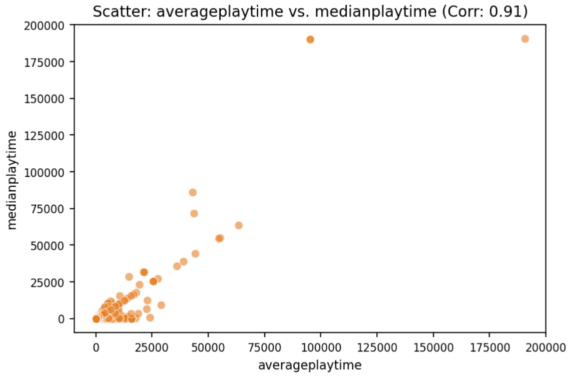

Figure 14: *Scatter plot for 'averageplaytime' and 'medianplaytime'. Correlation: 0.91.*

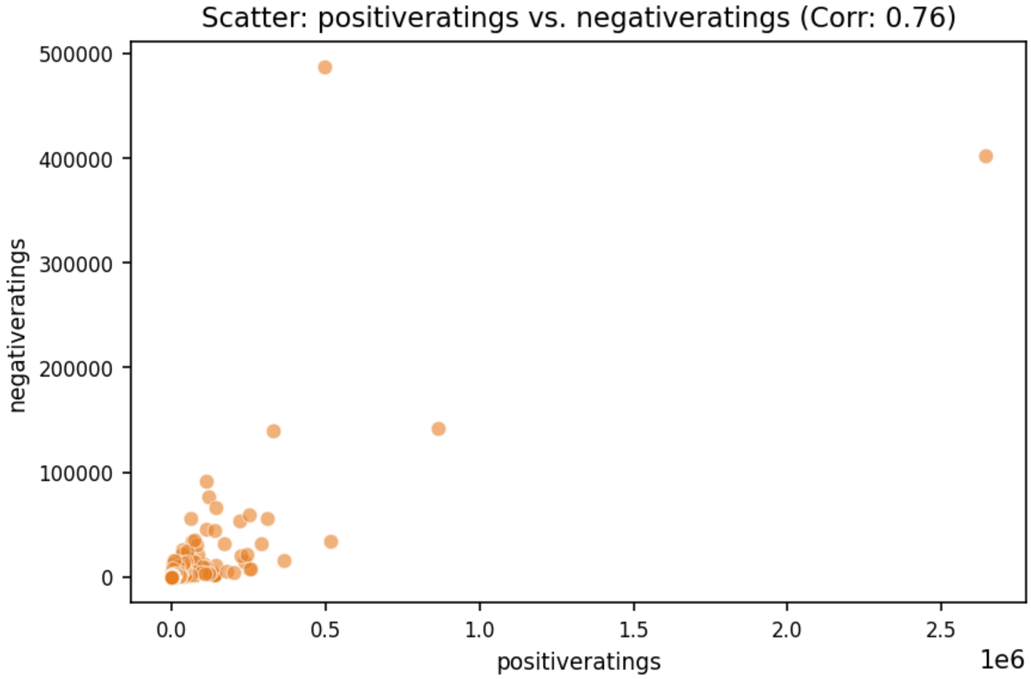

Figure 15: *Scatter plot for 'positiveratings' and 'negativeratings'. Correlation: 0.76.*

Interpretation of Numerical Correlations:

A correlation matrix displays the correlation coefficients between all pairs of variables in a dataset. Each cell in the matrix shows the correlation between two variables; a value of +1 indicates a perfect positive correlation, -1 a perfect negative correlation, and 0 indicates no linear correlation. The closer the absolute value is to 1, the stronger the relationship. In this analysis, two strong positive correlations stand out: the correlation between `averageplaytime` and `medianplaytime` (0.91), and the correlation between `positiveratings` and `negativeratings` (0.76). The strong positive correlation between average and median playtime suggests that games with longer average playtime also tend to have longer median playtimes. This is expected, as a similar distribution of playtime values leads to similar average and median values. The strong positive correlation between positive and negative ratings is more intriguing. It might imply that games which receive a large number of positive ratings also tend to receive a substantial number of negative ratings. This could indicate that highly popular games attract a larger overall player base, leading to both more positive and negative feedback. Further investigation would be needed to determine the underlying reasons, but it might suggest that highly engaging games (leading to more playtime) also evoke stronger reactions, both positive and negative, from players. The relatively weak correlation between negative ratings and average playtime (0.16) suggests that the amount of negative feedback doesn't strongly influence playtime. The scatter plots likely visually confirm these relationships, showing tightly clustered points for the `averageplaytime` vs `medianplaytime` and a more dispersed but still positively sloped cloud of points for `positiveratings` vs `negativeratings`.

4.2. Numerical vs. Categorical Features

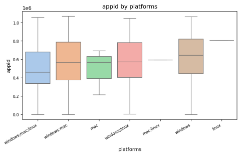

Figure 16: *Box plot of 'appid' across categories of 'platforms'.*

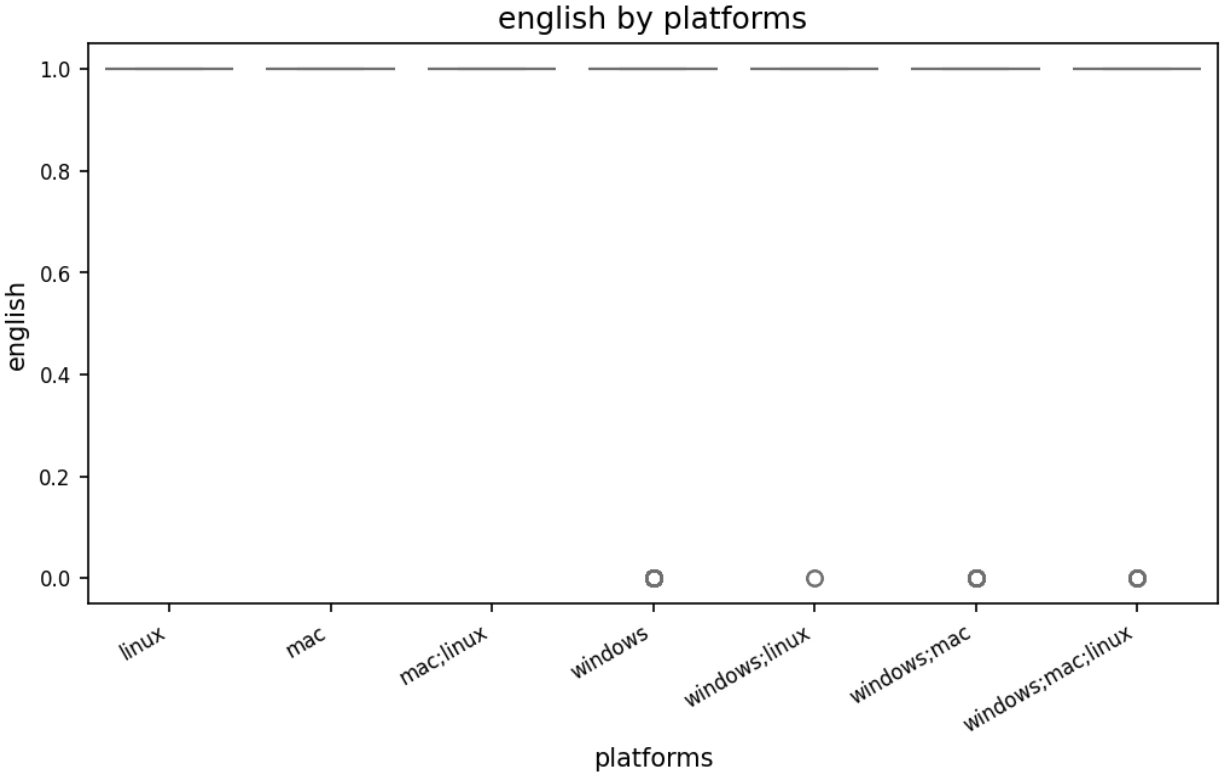

Figure 17: *Box plot of 'english' across categories of 'platforms'.*

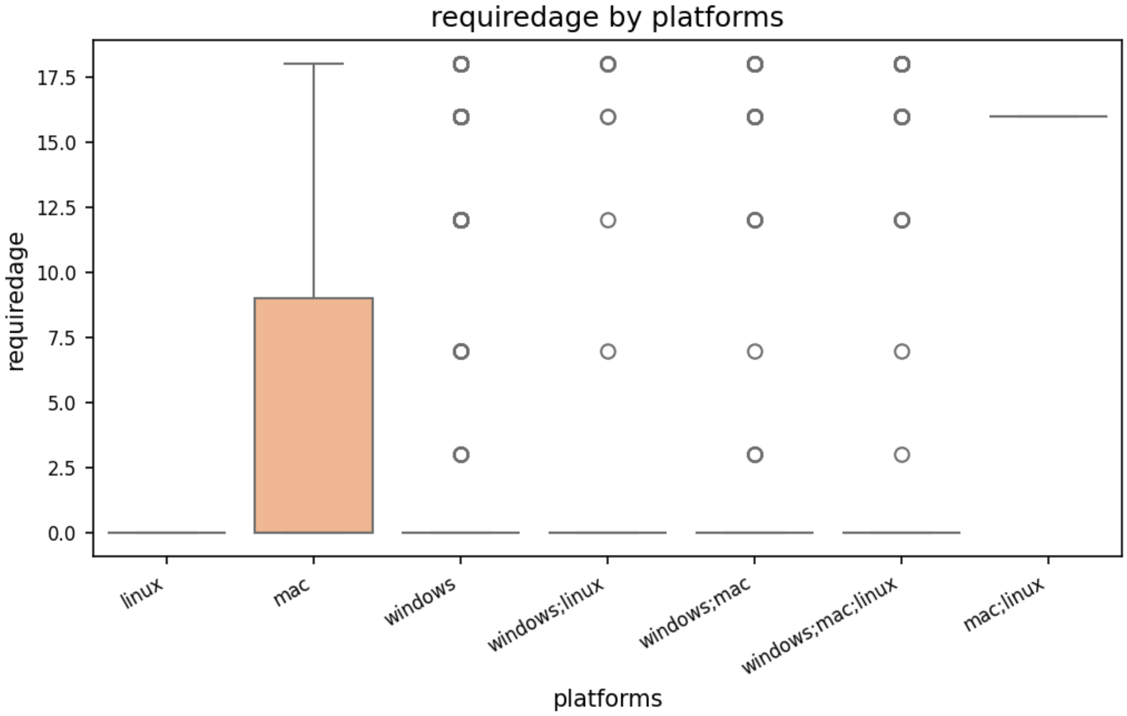

Figure 18: *Box plot of 'requiredage' across categories of 'platforms'.*

Interpretation of Numerical vs. Categorical Interactions:

Box plots visualizing numerical distributions across categories offer a powerful way to compare the central tendency and dispersion of data within different groups. They reveal not only the median (the middle value) of each category but also the interquartile range (IQR), representing the spread of the middle 50% of the data, as well as potential outliers. By visually comparing the boxes and whiskers across categories, we can quickly assess whether there are significant differences in the typical values and variability of the numerical variable. For example, a box plot showing 'appid' by 'platforms' might reveal whether certain platforms tend to have higher or lower application IDs, suggesting potential differences in the number of applications submitted or approved on each platform. Similarly, a 'english' (presumably representing a score or count related to English language proficiency) by 'platforms' plot could highlight whether certain platforms attract users with higher or lower English language skills. Significant differences observed in the medians indicate that the typical value of the numerical variable differs systematically between categories. For instance, if the median 'appid' is significantly higher for Platform A than for Platform B, it suggests Platform A has a larger number of applications. Differences in the spread (IQR or presence of outliers) reveal variations in the data's variability. A larger IQR for a particular category means that the data within that category is more spread out and less consistent, while the presence of outliers indicates extreme values that warrant further investigation. For example, if the 'english' scores on Platform C show a much larger IQR than on Platform D, it suggests greater heterogeneity in English language proficiency among users on Platform C. These insights can be crucial for understanding underlying patterns and making informed decisions based on the data.

4.3. Categorical vs. Categorical Features

5. Key Findings & Insights Summary

Key Findings & Insights The dataset `temp_steam.csv` comprises 270,85 rows and 18 columns, consisting of 9 numerical, 9 categorical, and no datetime features. Initial data quality assessment revealed the presence of 10 duplicate rows, while no missing values or constant columns were detected. The existence of duplicate rows suggests potential data redundancy and requires further investigation to determine the source and appropriate handling (e.g., removal or merging). This redundancy could skew statistical analyses if not addressed. Univariate analysis examined the distributions of the 9 numerical and 9 categorical features. While specific details of these distributions are not provided in the log, the analysis itself suggests a foundation for understanding the individual characteristics of each feature. Further investigation into the specific shapes of these distributions (e.g., normality, skewness) and the identification of outliers within numerical features will be crucial for subsequent modeling and analysis. Similarly, the frequency distributions of the categorical features will help in understanding the class balance and potential imbalances that may require further preprocessing. The bivariate analysis explored relationships between various feature pairs, revealing several observations (though the specific findings are not detailed). The lack of specifics prevents a detailed summary of the correlations or associations discovered. However, the fact that bivariate analysis was performed indicates that the relationships between different features were investigated. This analysis will be key to identifying potential predictors and understanding the underlying structure of the data. No explicitly surprising or unexpected findings are mentioned in the provided log. However, the discrepancy between the apparent cleanliness (no missing values or constant columns) and the existence of duplicate rows could be considered noteworthy. A deeper dive into the nature of these duplicates is warranted to understand their origin and potential impact on the reliability of the analysis.

6. Conclusion & Potential Next Steps

This automated report provides a foundational, high-level overview of the `temp_steam.csv` dataset, characterizing its structure, data quality, and highlighting initial observations from univariate and bivariate analyses. The absence of missing values and a relatively small number of duplicates suggests good initial data quality, paving the way for more in-depth exploration. Given the report's findings of 10 duplicate rows, the first next step is to **investigate and remove the 10 duplicate rows from the dataset**. This will ensure the subsequent analyses are based on clean and accurate data. The report mentions two observations from the bivariate analysis; therefore, the second step is to **fully document these two bivariate observations, including the specific feature pairs involved and the nature of the observed relationships**. This documentation should include visualizations (scatter plots, correlation matrices, etc.) to support the findings. The nature of these observations (e.g., correlations, interactions) will then guide the next steps. For example, if strong correlations are identified, a third step would be to **perform further analysis to determine the strength and significance of these correlations using appropriate statistical tests (e.g., Pearson's correlation, Spearman's rank correlation)**, and assess potential confounding variables. Finally, given that the report only states that "various feature pairs" were analyzed, a fourth step would be to **systematically explore all relevant bivariate relationships between numerical and categorical features**. This could involve creating visualizations (box plots, violin plots) to compare the distributions of numerical features across different categories of categorical features. This comprehensive bivariate analysis will help reveal potential interactions and relationships missed in the initial automated analysis.