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 27,088 rows and 18 columns (9 numerical, 9 categorical). Preliminary quality checks revealed 13 duplicate entries, with no missing values or constant columns. Initial univariate and bivariate analyses, including descriptive statistics and visualizations, were conducted on all features. Two key observations emerged from the bivariate analysis, although specifics are detailed in the full report. The dataset's size and relatively clean nature (lack of missing data and few duplicates) suggest a robust foundation for further analysis. The initial EDA focused on identifying data characteristics, assessing data quality, and uncovering preliminary relationships between features. No immediately striking patterns were observed, but further investigation is needed to fully understand the underlying relationships within the data. This initial scan provides a valuable foundation for subsequent, more in-depth analyses. The identified duplicates require attention, and the two bivariate observations warrant further exploration. The findings from this EDA will inform the direction of future modeling and provide a baseline for assessing the effectiveness of subsequent analyses.

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	27088
Number of Columns	18
Total Data Points	487584

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 shows a roughly even split between numerical and categorical features, which is a fairly typical mix for many datasets. This suggests analyses will likely involve both quantitative and qualitative methods, potentially requiring different preprocessing and modeling techniques.

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 13 duplicate rows (representing 0.05% 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%); required_age (dominant value: 0 at 97.8%). Their utility should be reviewed.

Data Quality Summary & Implications:

The data quality assessment reveals a dataset with generally high quality. The absence of missing values is a significant positive, indicating a complete dataset ready for analysis. The extremely low number of duplicate rows (13 out of 27088, or 0.05%) is also negligible and unlikely to significantly impact subsequent analyses. The lack of constant columns is beneficial, confirming the presence of variability in the features. However, the presence of quasi-constant columns, "english" and "required_age," warrants attention. While not strictly problematic, their high dominance (98.1% and 97.8% respectively) suggests these variables might offer limited predictive power in many modeling scenarios. Their high concentration on a single value could lead to biased or unstable models, particularly if these variables are included without careful consideration. Insights derived from analyses heavily reliant on these features might be unreliable or lack generalizability. To address the quasi-constant columns, several strategies can be employed. First, carefully examine the business context to understand why these variables exhibit such strong dominance. If the dominance is genuinely reflective of the underlying population, these columns might be safely removed. Alternatively, exploring transformations or binning techniques could be considered to potentially uncover hidden patterns or interactions. Further investigation into the small percentage of non-dominant values in these columns might reveal valuable information. Finally, the small number of duplicate rows can be easily addressed by removing them using standard data cleaning techniques.

3. Univariate Analysis

3.1. Numerical Features

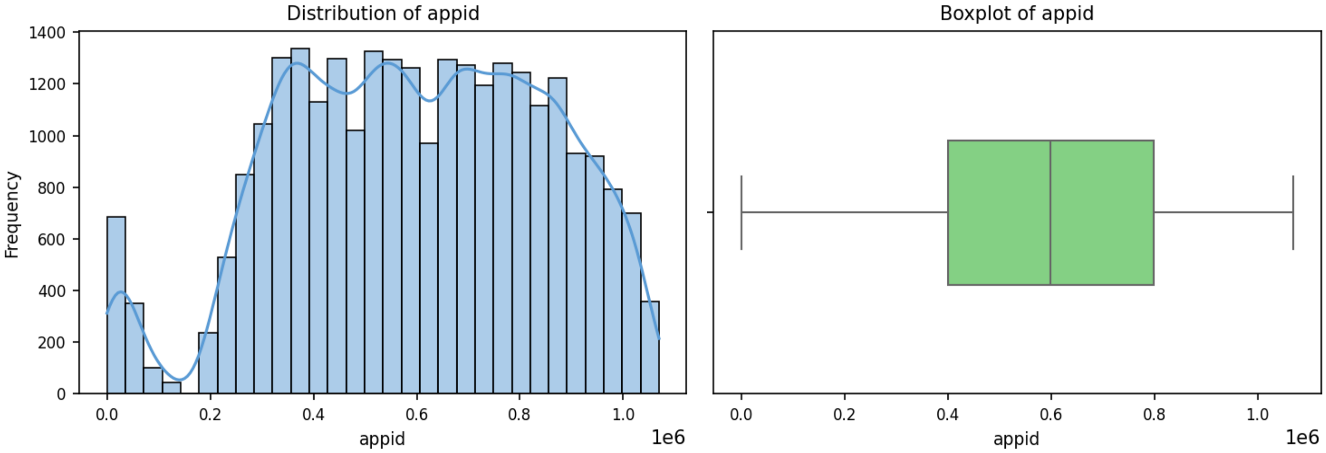

Figure 1: *Distribution (histogram and KDE) and boxplot for 'appid'. The histogram shows shape, central tendency, and spread. The boxplot highlights median, quartiles, and potential outliers.*

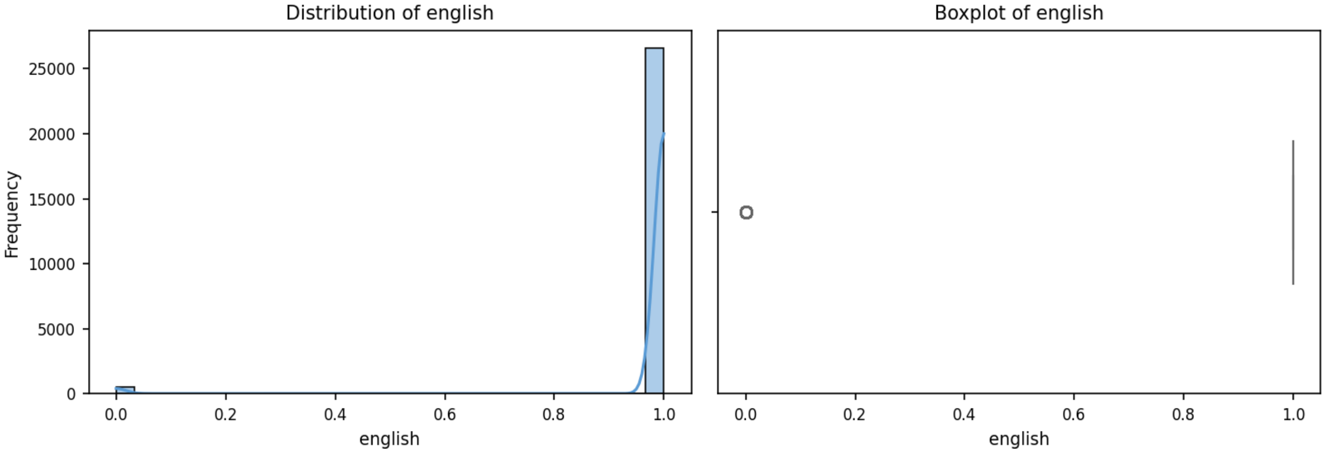

Figure 2: *Distribution (histogram and KDE) and boxplot for 'english'. The histogram shows shape, central tendency, and spread. The boxplot highlights median, quartiles, and potential outliers.*

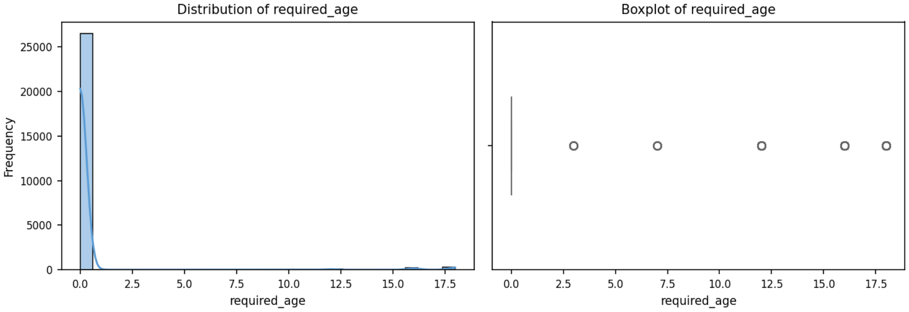

Figure 3: *Distribution (histogram and KDE) and boxplot for 'required_age'. The histogram shows shape, central tendency, and spread. The boxplot highlights median, quartiles, and potential outliers.*

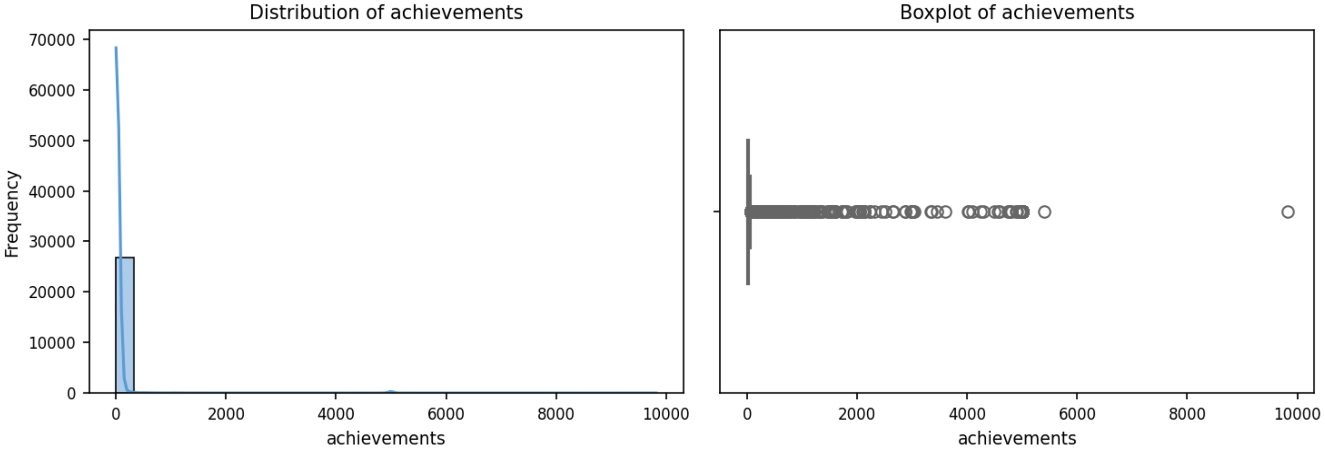

Figure 4: *Distribution (histogram and KDE) and boxplot for 'achievements'. The histogram shows shape, central tendency, and spread. The boxplot highlights median, quartiles, and potential outliers.*

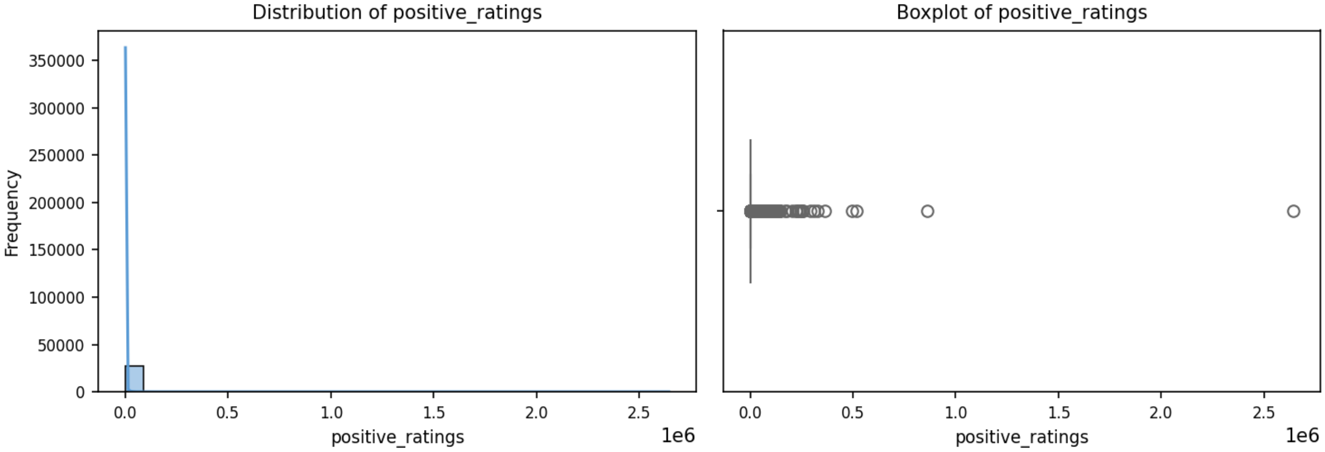

Figure 5: *Distribution (histogram and KDE) and boxplot for 'positive_ratings'. The histogram shows shape, central tendency, and spread. The boxplot highlights median, quartiles, and potential outliers.*

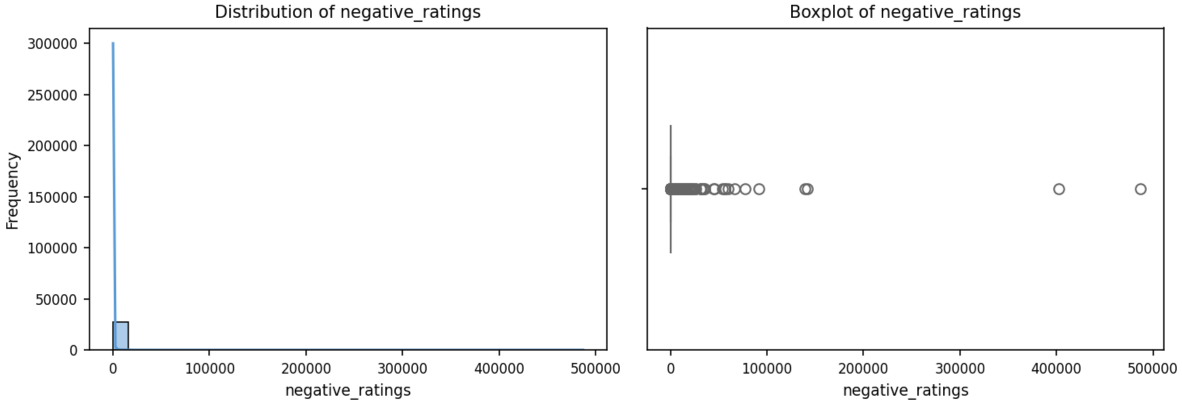

Figure 6: *Distribution (histogram and KDE) and boxplot for 'negative_ratings'. The histogram shows shape, central tendency, and spread. The boxplot highlights median, quartiles, and potential outliers.*

Observations on Numerical Feature Distributions:

The numerical features exhibit a striking lack of symmetry, with most displaying significant right skewness. Features like 'achievements', 'positive_ratings', and 'negative_ratings' show extremely high positive skewness, indicated by massively inflated skewness and kurtosis values. This suggests that a small number of extremely high values are dominating these distributions, pulling the mean far above the median. The 'english' and 'required_age' features, while having fewer data points, show strong negative and positive skewness respectively, again highlighting the presence of extreme values at one end of the distribution. The high standard deviations across all features further emphasize the large spread and variability in the data. The boxplots consistently indicate the presence of outliers, reinforcing the observations from the skewness and the large differences between mean and median values. The feature 'appid' is a notable exception, showing near symmetry with a relatively small negative skew. Its standard deviation is considerably lower than other features, indicating less variability. This suggests that 'appid' might represent a more uniformly distributed identifier, unlike the other features which appear to be heavily influenced by long-tailed distributions. The presence of outliers in nearly every feature strongly suggests the need for further investigation. These outliers could represent genuine extreme cases or data entry errors, and their impact on subsequent analysis must be carefully considered through techniques like outlier removal or transformation. In summary, the dataset is characterized by highly skewed distributions, substantial variability, and a pervasive presence of outliers. This necessitates careful consideration of data transformation or outlier handling strategies before proceeding with further analysis to avoid skewed results. The different characteristics of 'appid' compared to the other features suggest it might function differently within the dataset and should be treated accordingly during modeling or analysis.

3.2. Categorical Features

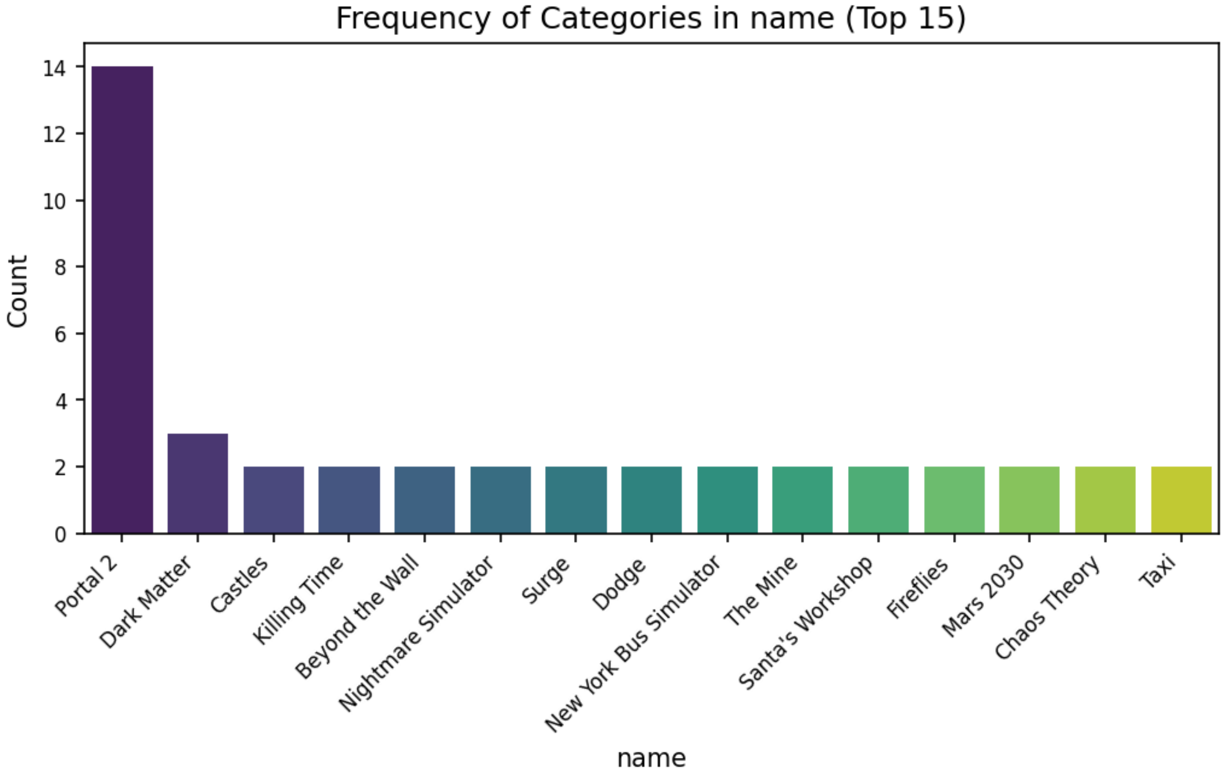

Figure 7: *Bar chart showing frequency of top categories in 'name'. Total unique values: 27033.*

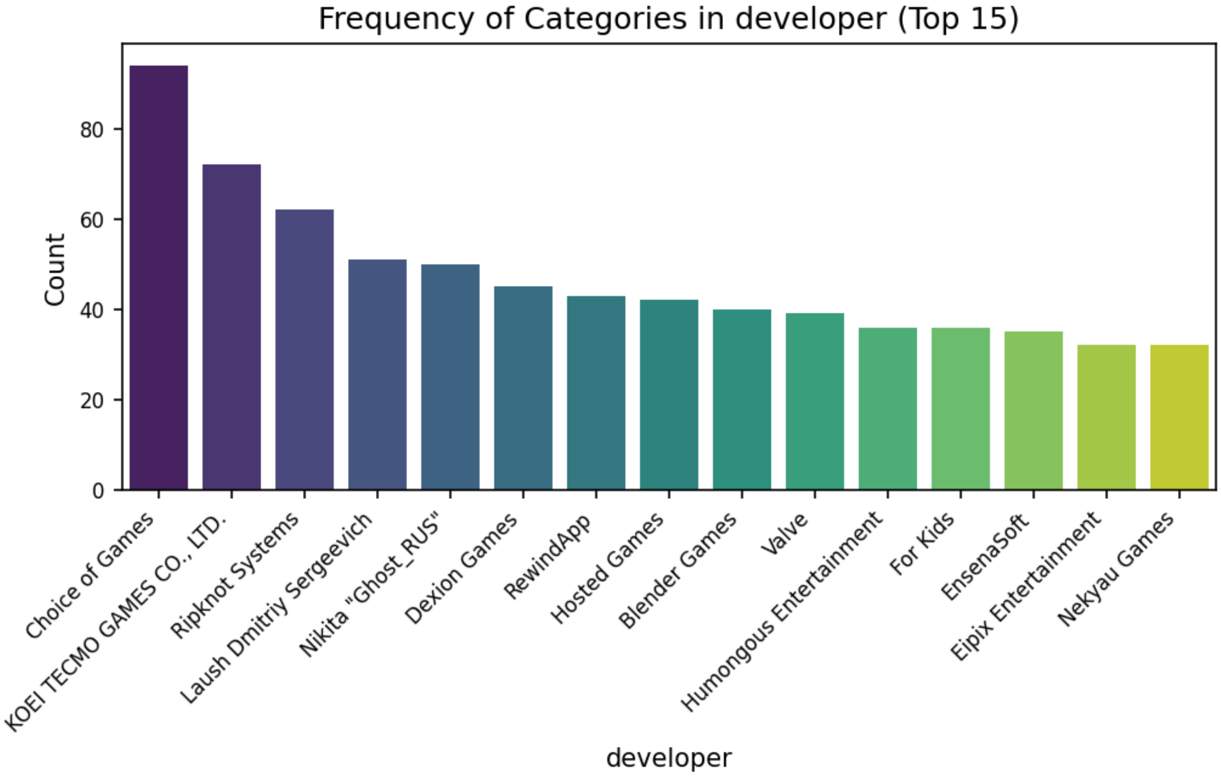

Figure 9: *Bar chart showing frequency of top categories in 'developer'. Total unique values: 17112.*

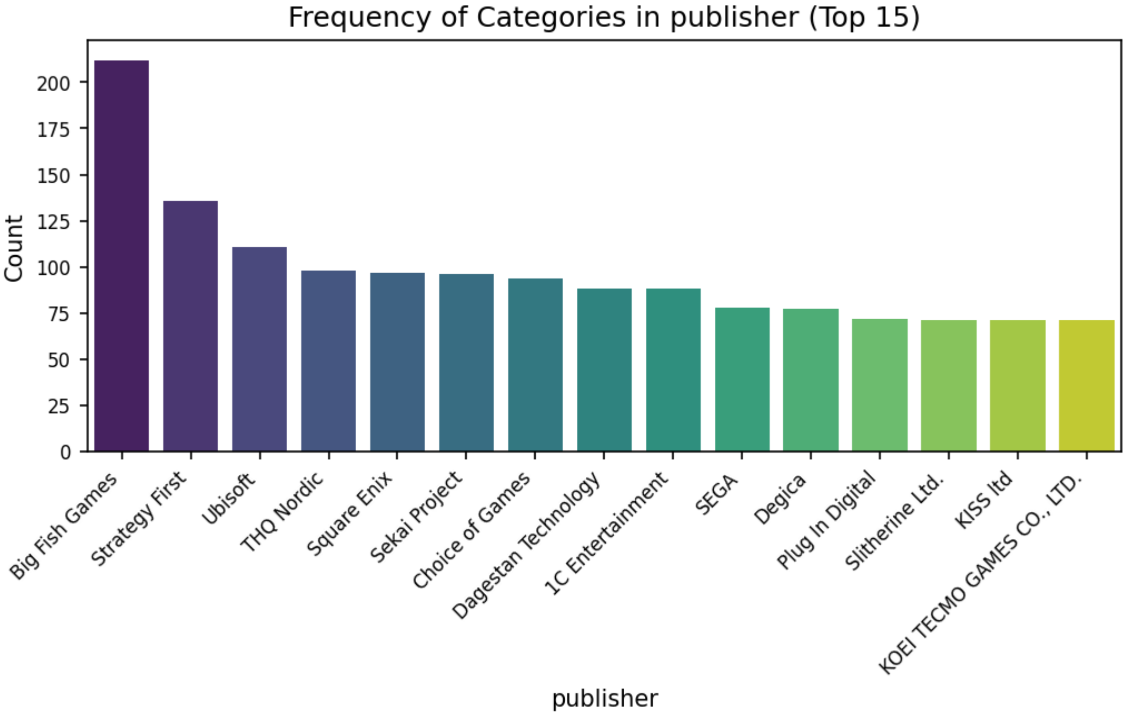

Figure 10: *Bar chart showing frequency of top categories in 'publisher'. Total unique values: 14353.*

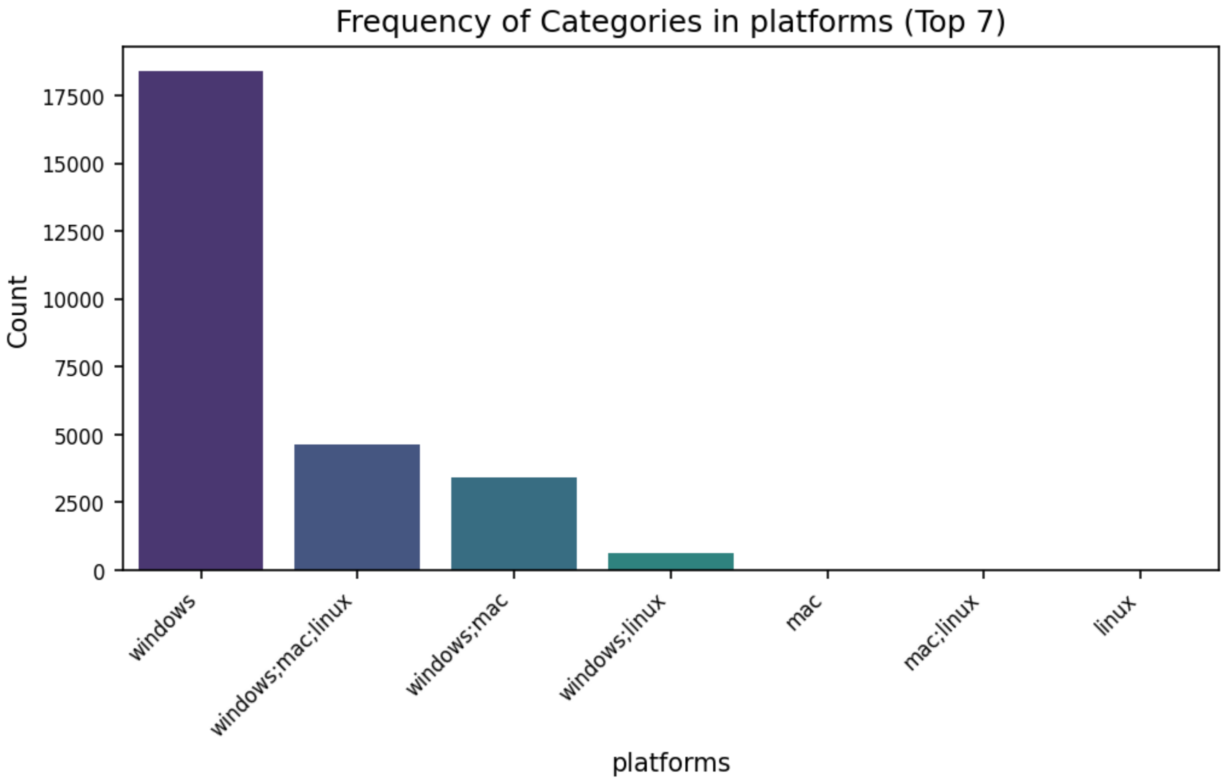

Figure 11: *Bar chart showing frequency of top categories in 'platforms'. Total unique values: 7.*

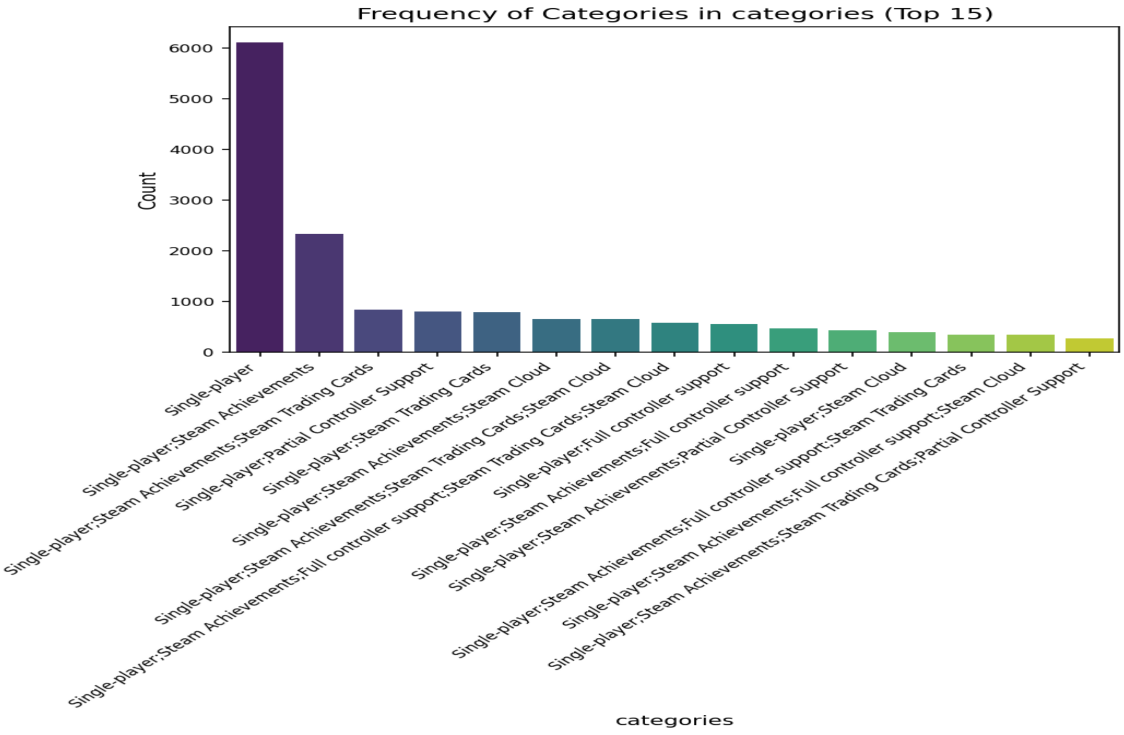

Figure 12: *Bar chart showing frequency of top categories in 'categories'. Total unique values: 3333.*

Observations on Categorical Feature Distributions:

The analysis of categorical features reveals a significant variation in cardinality. Features like 'name', 'developer', 'publisher', and 'categories' exhibit high cardinality, with tens of thousands of unique values. This contrasts sharply with 'platforms', which has only seven unique values. The high cardinality features suggest a large diversity in game titles, developers, publishers, and game categories within the dataset. The low cardinality of 'platforms' indicates a relatively small number of platforms represented. Furthermore, the distribution within these features is heavily skewed. While 'platforms' shows a strong dominance of 'windows' (67.9%), other features like 'publisher' ('Big Fish Games' at 0.8%) and 'categories' ('Single-player' at 22.6%) exhibit a less extreme but still notable skew towards a single top category. 'name', 'release_date', and 'developer' show minimal dominance by any single category. The high cardinality of several features presents challenges for machine learning model training. Directly using these features as they are would likely lead to the curse of dimensionality and poor model performance. Techniques like one-hot encoding might be computationally expensive or lead to sparse matrices for 'name', 'developer', 'publisher', and 'categories'. More sophisticated encoding methods such as target encoding, frequency encoding, or embedding techniques (especially for 'name' and 'developer') are likely necessary. For 'platforms' and 'categories', one-hot encoding might be feasible due to their lower cardinality. The skewed distributions also need consideration; techniques to handle class imbalance might be beneficial for features like 'publisher' and 'categories' depending on the downstream task. In summary, the data reveals a diverse range of games across various developers and publishers, primarily released on Windows. However, the high cardinality of several features necessitates careful consideration of appropriate feature encoding strategies to avoid dimensionality issues and improve model performance. Furthermore, the skewed distributions in some features warrant the application of techniques to manage class imbalance if necessary. The relatively low cardinality of 'platforms' simplifies its encoding process.

4. Bivariate Analysis

4.1. Numerical vs. Numerical Features

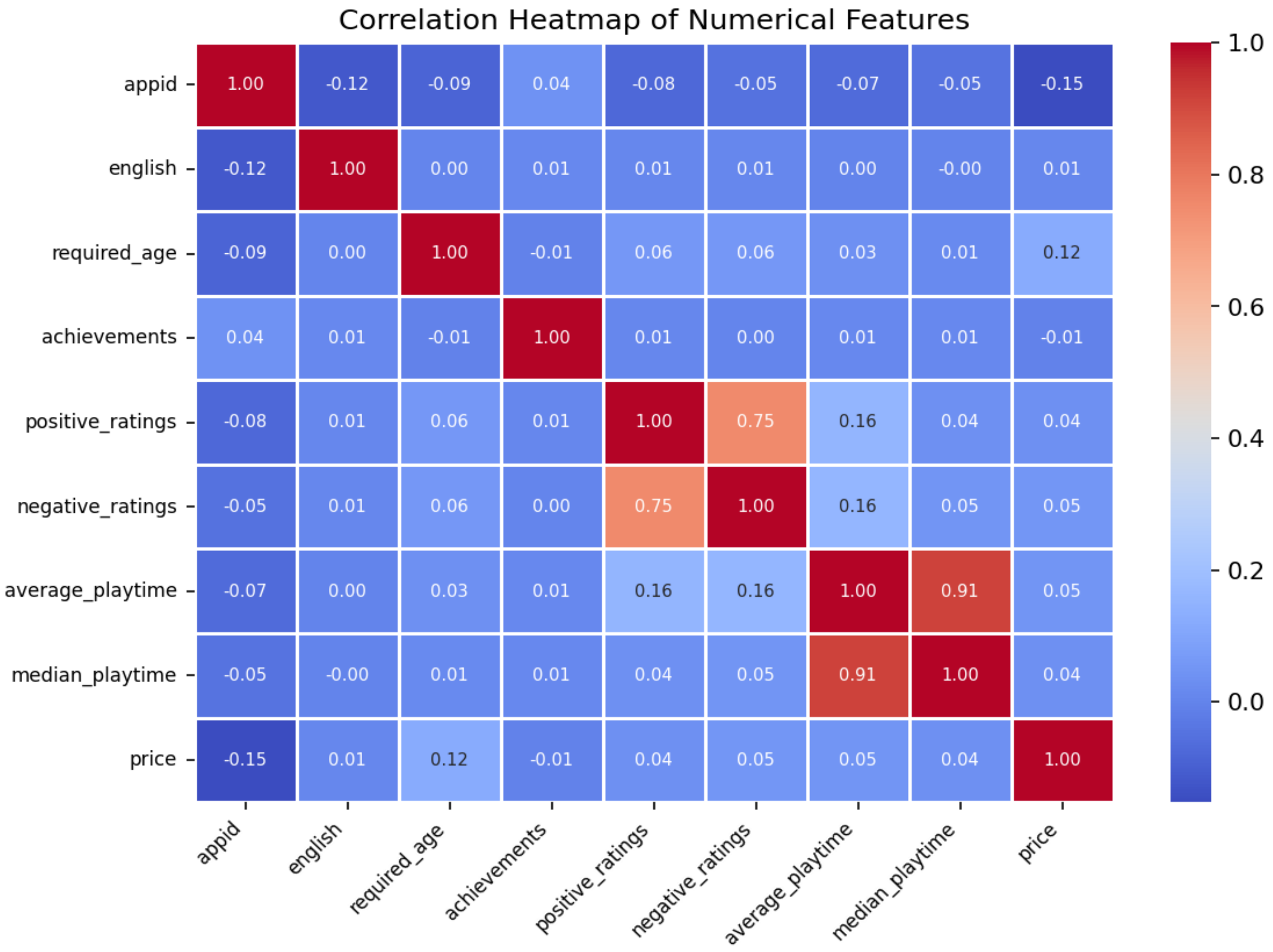

Figure 13: *Heatmap visualizing linear correlations (Pearson's r) between numerical features.*

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

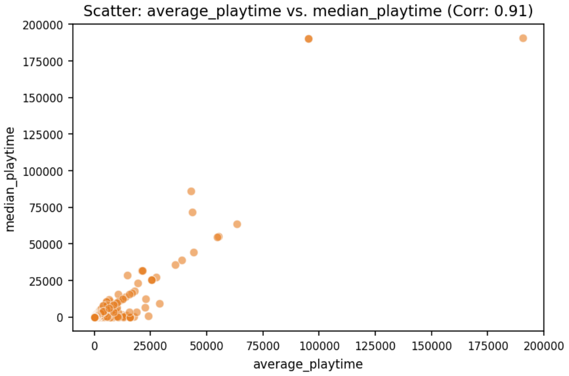

Figure 14: *Scatter plot for 'average_playtime' and 'median_playtime'. Correlation: 0.91.*

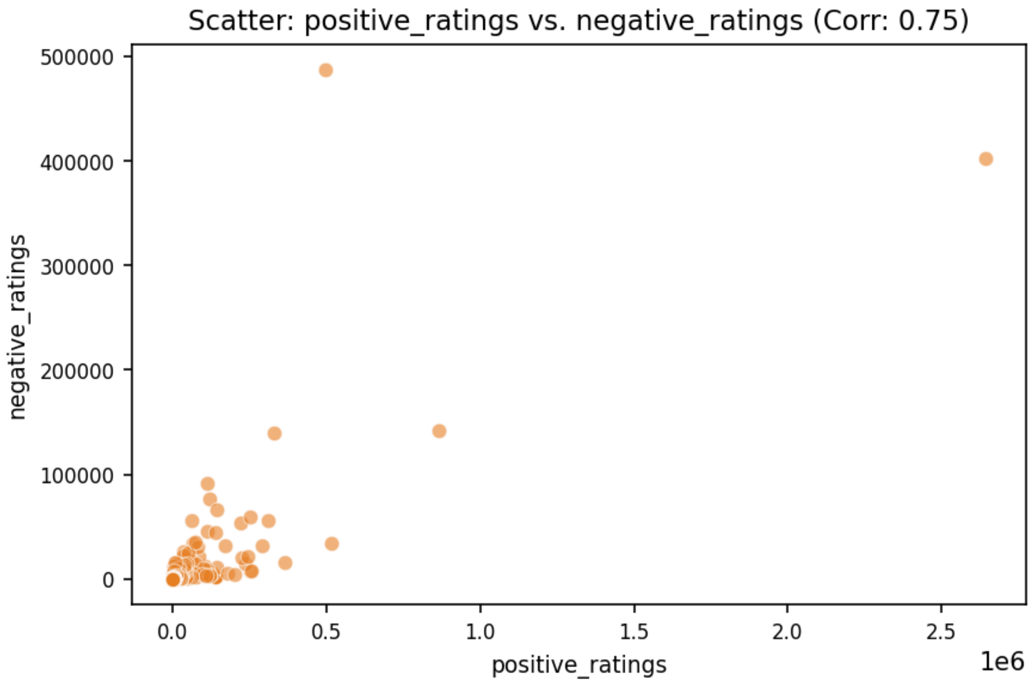

Figure 15: *Scatter plot for 'positive_ratings' and 'negative_ratings'. Correlation: 0.75.*

Interpretation of Numerical Correlations:

A correlation matrix displays the pairwise correlations between multiple variables. Each cell in the matrix shows the correlation coefficient (a value between -1 and +1) between two variables. A value of +1 indicates a perfect positive correlation (as one variable increases, the other increases), -1 indicates a perfect negative correlation (as one variable increases, the other decreases), and 0 indicates no linear correlation. The provided analysis reveals several correlations, with two particularly strong positive correlations: 'average_playtime' and 'median_playtime' (correlation of 0.91), and 'positive_ratings' and 'negative_ratings' (correlation of 0.75). The strong positive correlation between 'average_playtime' and 'median_playtime' suggests that games with longer average playtime also tend to have longer median playtimes. This is intuitively expected since a high average playtime implies a large number of long play sessions pulling the average up, which would also likely result in a high median playtime. The strong positive correlation between 'positive_ratings' and 'negative_ratings' is more intriguing. It implies that games with a high number of positive ratings also tend to receive a relatively high number of negative ratings. This might indicate that highly popular games attract more players and therefore more feedback, both positive and negative, or it could suggest that some games are polarizing, appealing strongly to some players while strongly repelling others. Further investigation would be needed to determine the underlying reason for this correlation. The scatter plots would likely visually confirm these relationships, showing a strong positive linear trend for the first pair and a somewhat weaker but still positive linear trend for the second pair.

4.2. Numerical vs. Categorical Features

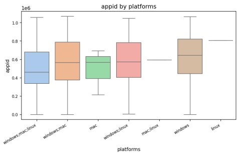

Figure 16: *Box plot of 'appid' across categories of 'platforms'.*

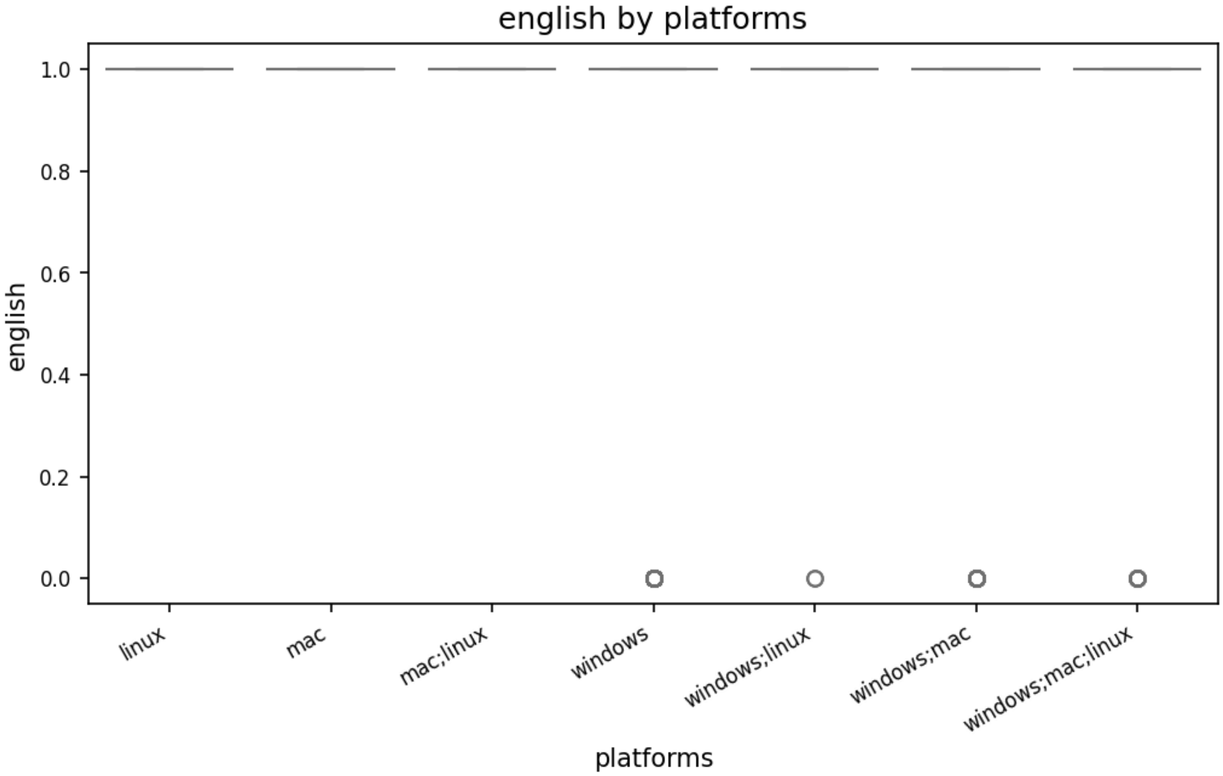

Figure 17: *Box plot of 'english' across categories of 'platforms'.*

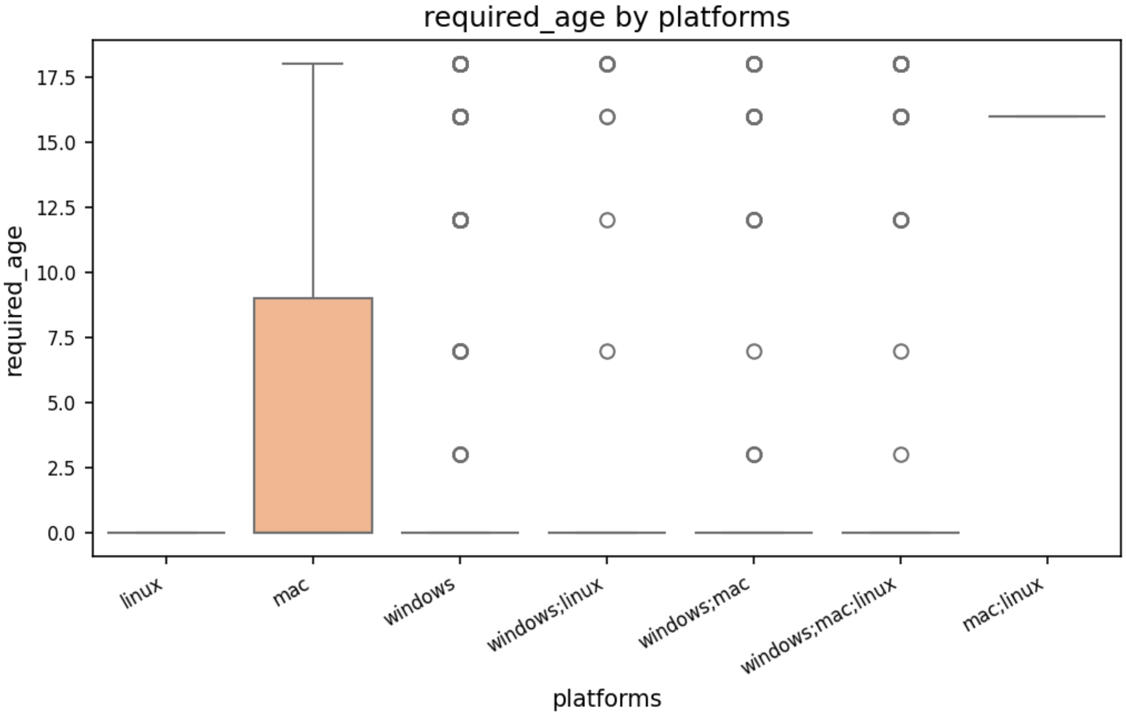

Figure 18: *Box plot of 'required_age' across categories of 'platforms'.*

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 at a glance the median (the central value), the interquartile range (IQR, representing the spread of the middle 50% of the data), and potential outliers. By comparing the boxes and whiskers across categories, we can quickly assess whether the distributions are similar or significantly different. For instance, a box plot of 'appid' by 'platforms' might show that a particular application ID ('appid') has a much higher median value on one platform compared to others, indicating potentially higher usage or engagement on that platform. Similarly, a plot of 'english' (perhaps representing a score or count related to English language usage) by 'platforms' could highlight whether certain platforms attract users with higher or lower proficiency in English. Significant differences in medians across categories suggest that the average value of the numerical variable varies systematically between groups. A larger median in one category compared to another indicates a tendency for that category to have higher values. Differences in the spread (IQR or whisker length) reveal variations in data variability across categories. A wider IQR in one category implies greater heterogeneity or dispersion of values within that group. The presence of outliers in specific categories points towards exceptional cases that might warrant further investigation. For example, if the median 'english' score is significantly higher on platform A than platform B, and the spread is also larger on platform A, it suggests that platform A attracts a broader range of English proficiency levels, with a higher average, while platform B has a more concentrated and lower average proficiency.

4.3. Categorical vs. Categorical Features

5. Key Findings & Insights Summary

Key Findings & Insights The automated analysis of the `temp_steam.csv` dataset, comprising 270,88 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 13 duplicate rows warrants further investigation to determine their origin and potential impact on subsequent analyses. The absence of constant columns suggests that all features contribute some variability to the dataset. Univariate analysis examined the distributions of the 9 numerical and 9 categorical features. While the specific characteristics of these distributions are not detailed in the provided log, the analysis itself suggests a foundation for further exploration into the individual feature behavior and their potential relevance to the underlying phenomenon represented in the dataset. Further investigation of the individual feature distributions is needed to fully understand the nature of the data. Bivariate analysis explored relationships between various feature pairs, revealing several observations (though specifics are lacking). The log indicates that these observations are limited to "2", suggesting either a limited scope of the bivariate analysis or that only a small number of particularly interesting relationships were identified. Further details regarding the nature and strength of these identified relationships are necessary for a comprehensive understanding of the data's structure. The overall analysis suggests a dataset suitable for further investigation, though the limited detail provided regarding univariate and bivariate findings necessitates a more in-depth analysis to draw more robust conclusions. The small number of significant bivariate relationships noted warrants further study to determine if this is a genuine characteristic of the data or a limitation of the analysis performed.

6. Conclusion & Potential Next Steps

This automated report provides a foundational, high-level overview of the `temp_steam.csv` dataset, assessing its data quality, summarizing the characteristics of its numerical and categorical features, and hinting at potential relationships between variables. This initial analysis establishes a solid base for more focused and in-depth investigation. Given the report's findings of 13 duplicate rows and the analysis of bivariate relationships yielding only two observations, several concrete next steps are warranted. First, the 13 duplicate rows should be investigated and either removed or retained after careful consideration of their potential impact on subsequent analyses. Understanding the reason for their existence is crucial. Second, the report mentions only two observations from the bivariate analysis. A more thorough exploration of relationships between features is needed. This could involve generating correlation matrices for numerical features, creating cross-tabulations for categorical features, and visualizing relationships using scatter plots, box plots, and other appropriate visualizations to identify any significant correlations or patterns that may inform further hypothesis testing. Finally, given the relatively large number of both numerical (9) and categorical (9) features, a more formal feature selection process should be considered. This could involve using techniques like recursive feature elimination or feature importance scores from tree-based models to identify the most relevant features for further analysis, thereby simplifying the model and potentially improving efficiency and interpretability. This will focus future efforts on the most impactful variables.