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). The dataset exhibits minimal data quality issues, with only 13 duplicate rows identified; no missing values or constant columns were detected. Preliminary univariate and bivariate analyses, including descriptive statistics and visual inspections, have been conducted. Key findings from this initial scan include the dataset's relatively large size and the absence of significant data quality problems. While initial bivariate analysis revealed two noteworthy observations (details provided in the full report), no immediately obvious, strong patterns emerged across features. Further investigation is warranted to fully understand the relationships within the data. This automated EDA provides a crucial foundation for subsequent, more in-depth analysis. The identified characteristics, combined with the lack of major data quality concerns, suggest the dataset is suitable for further modeling and predictive analysis, pending more thorough investigation of the two notable bivariate observations.

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 exhibits a roughly even split between numerical and categorical features, which is a relatively typical mix for many datasets. This suggests that both quantitative and qualitative analyses will be necessary to fully understand the data, requiring different techniques for each feature type.

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 generally high level of cleanliness in the dataset with 27,088 rows. The absence of missing values is a significant positive, indicating a comprehensive data collection process. The identification of only 13 duplicate rows (0.05% of the total) represents a negligible issue and can be easily addressed. The lack of constant columns confirms that the dataset is not trivially redundant, suggesting the presence of meaningful variation across features. However, the presence of two quasi-constant columns, `english` and `required_age`, warrants attention. While not entirely problematic, their high dominance by a single value (98.1% and 97.8% respectively) limits their potential predictive power in many machine learning models. These columns may be less informative than anticipated and could potentially lead to biased or less accurate models if included without careful consideration. The insights derived from analyses involving these features might be skewed, as they offer limited variability to explain any outcome. To address the identified issues, the 13 duplicate rows should be removed. For the quasi-constant columns, their usefulness needs careful evaluation. Depending on the analytical goals, they could be removed entirely, transformed (e.g., using techniques like binning or creating interaction terms with other features), or potentially used as stratification variables if their near-constant nature is relevant to the analysis (e.g., if the near-constant value represents a specific subgroup of interest). Further investigation into the reasons behind the high dominance in these columns is also recommended, as it might reveal underlying issues in data collection or representation.

3. Univariate Analysis

3.1. Numerical Features

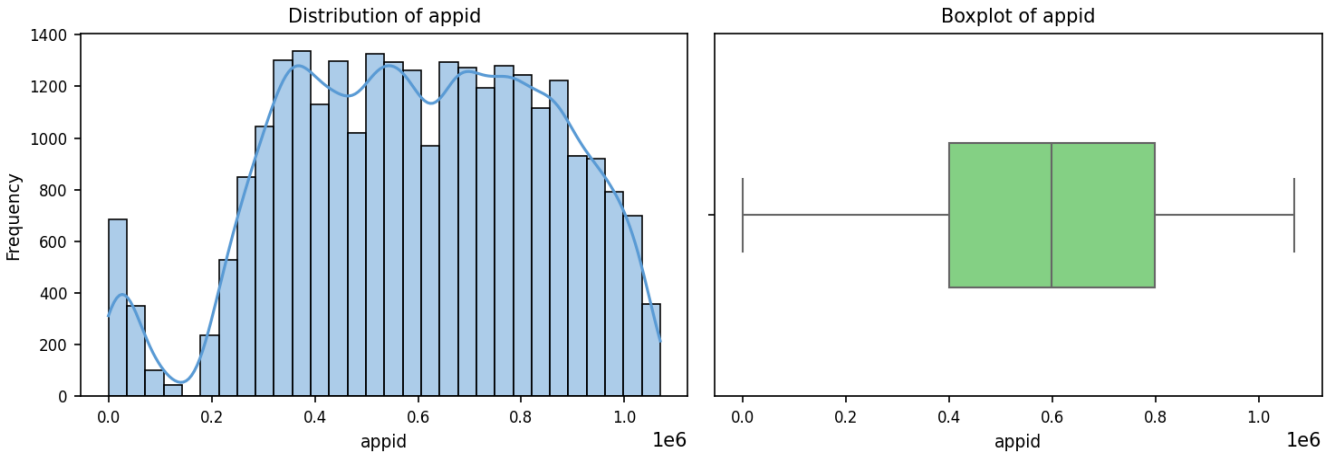

Figure 1: *Distribution (histogram and KDE) and boxplot for 'appid'. The histogram shows shape, central tendency, and spread. The boxplot highlights median, quartiles, and potential outliers.*

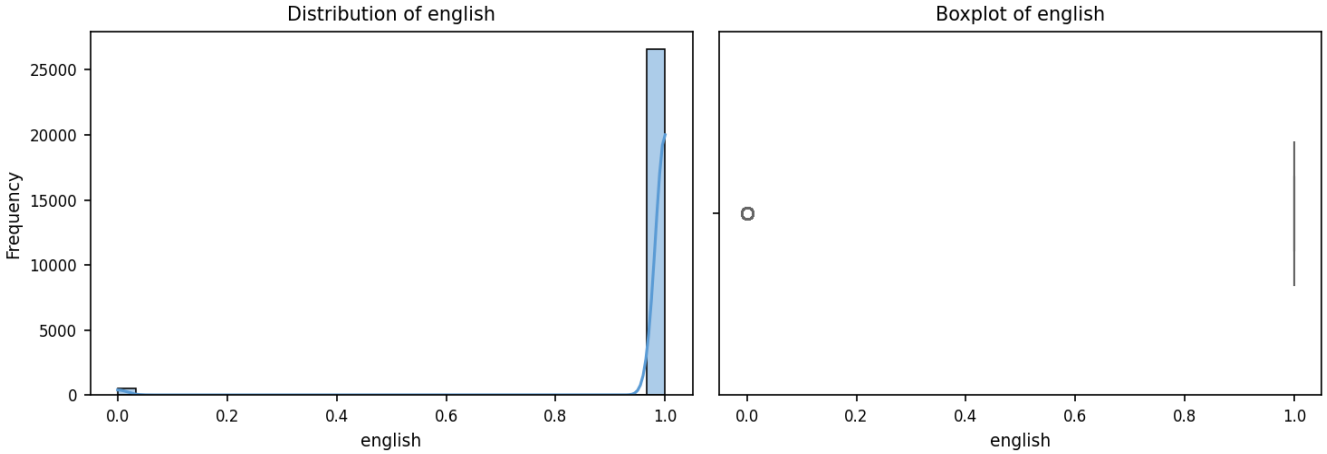

Figure 2: *Distribution (histogram and KDE) and boxplot for 'english'. The histogram shows shape, central tendency, and spread. The boxplot highlights median, quartiles, and potential outliers.*

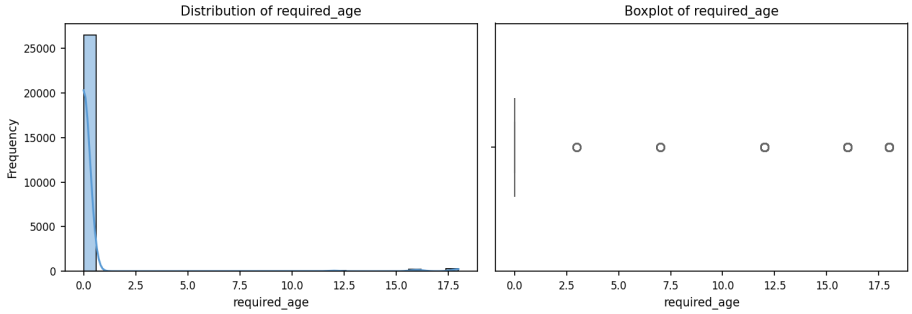

Figure 3: *Distribution (histogram and KDE) and boxplot for 'required_age'. The histogram shows shape, central tendency, and spread. The boxplot highlights median, quartiles, and potential outliers.*

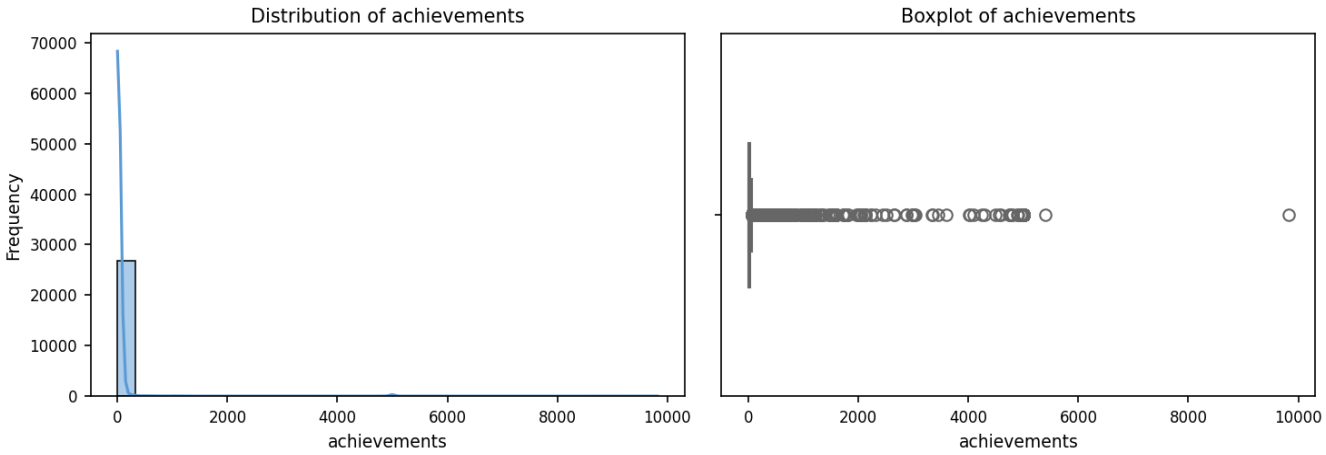

Figure 4: *Distribution (histogram and KDE) and boxplot for 'achievements'. The histogram shows shape, central tendency, and spread. The boxplot highlights median, quartiles, and potential outliers.*

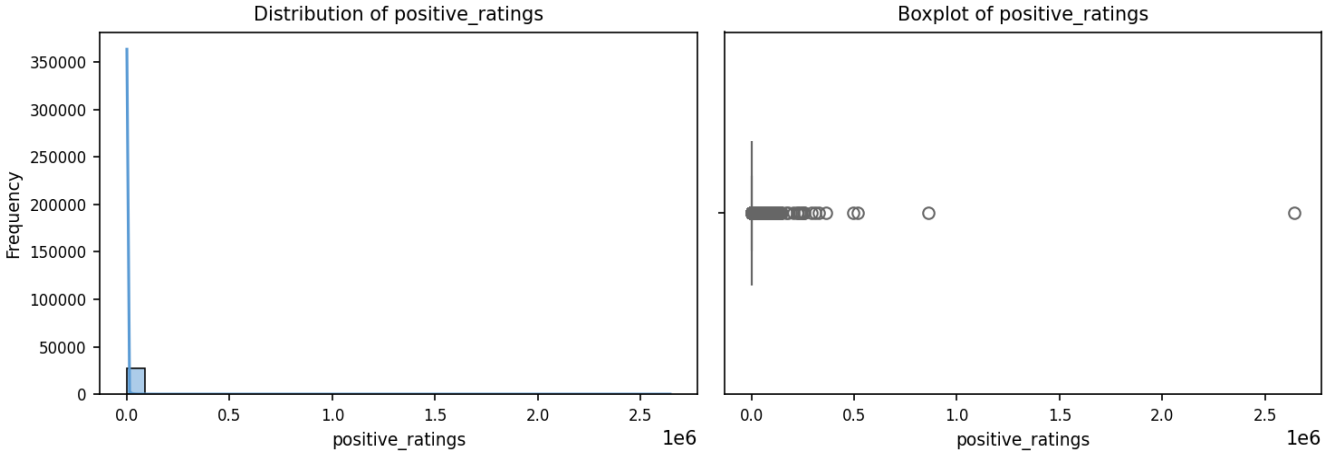

Figure 5: *Distribution (histogram and KDE) and boxplot for 'positive_ratings'. The histogram shows shape, central tendency, and spread. The boxplot highlights median, quartiles, and potential outliers.*

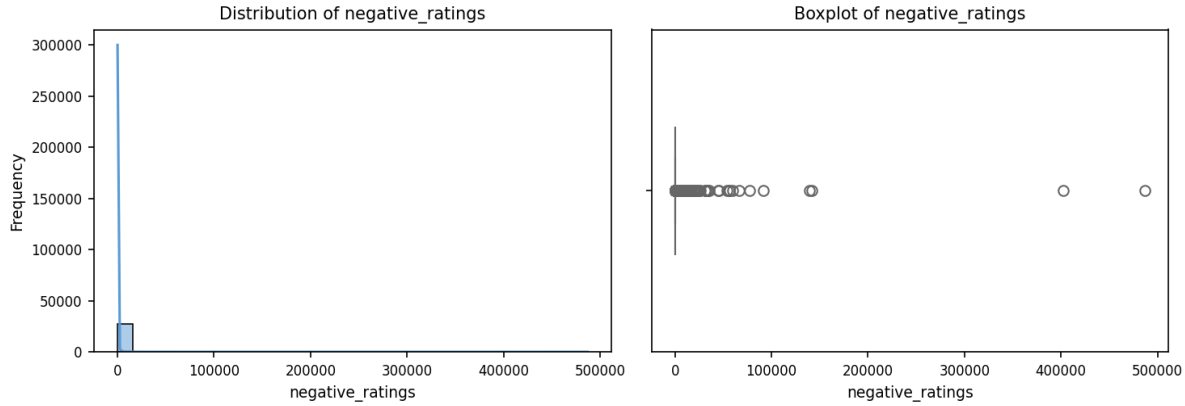

Figure 6: *Distribution (histogram and KDE) and boxplot for 'negative_ratings'. The histogram shows shape, central tendency, and spread. The boxplot highlights median, quartiles, and potential outliers.*

Observations on Numerical Feature Distributions:

The numerical features exhibit a striking lack of symmetry, with most displaying significant right skewness. This is particularly evident in 'achievements', 'positive_ratings', and 'negative_ratings', where the means are drastically higher than the medians, indicating the presence of very high values that pull the mean upwards. The extreme skewness is further confirmed by the high values of skewness and kurtosis for these features. The 'english' and 'required_age' features also show strong right skewness, although to a lesser extent than the others. In contrast, 'appid' shows a slight left skew, suggesting a longer tail towards lower values. The high standard deviations across most features highlight substantial variability and wide ranges of values, indicating a heterogeneous dataset. The boxplots consistently suggest the presence of numerous outliers across all features except potentially 'appid'. This is supported by the large discrepancies between the means and medians, and the extremely high maximum values relative to the other percentiles for 'achievements', 'positive_ratings', and 'negative_ratings'. For example, the maximum value for 'positive_ratings' is orders of magnitude larger than the mean, clearly indicating the influence of extreme data points. This prevalence of outliers necessitates careful consideration during further analysis, as they could significantly distort model training and interpretation. Robust statistical methods may be required to mitigate their impact. In summary, the data is characterized by highly skewed distributions, substantial variability, and a considerable number of potential outliers. This suggests the need for data preprocessing techniques such as outlier handling (e.g., winsorization, trimming, or transformation) and potentially feature scaling to improve the performance and interpretability of subsequent analyses. Understanding the underlying reasons for the skewness and outliers will be crucial for drawing meaningful conclusions from this dataset.

3.2. Categorical Features

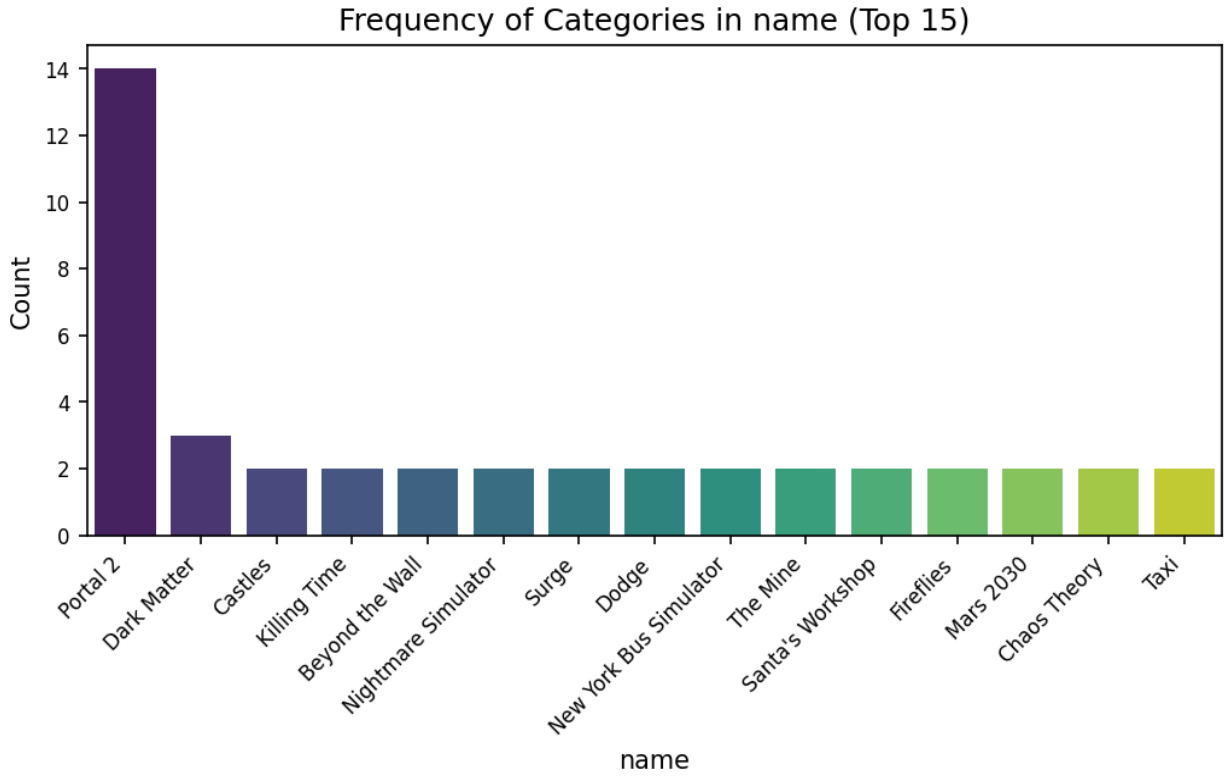

Figure 7: *Bar chart showing frequency of top categories in 'name'. Total unique values: 27033.*

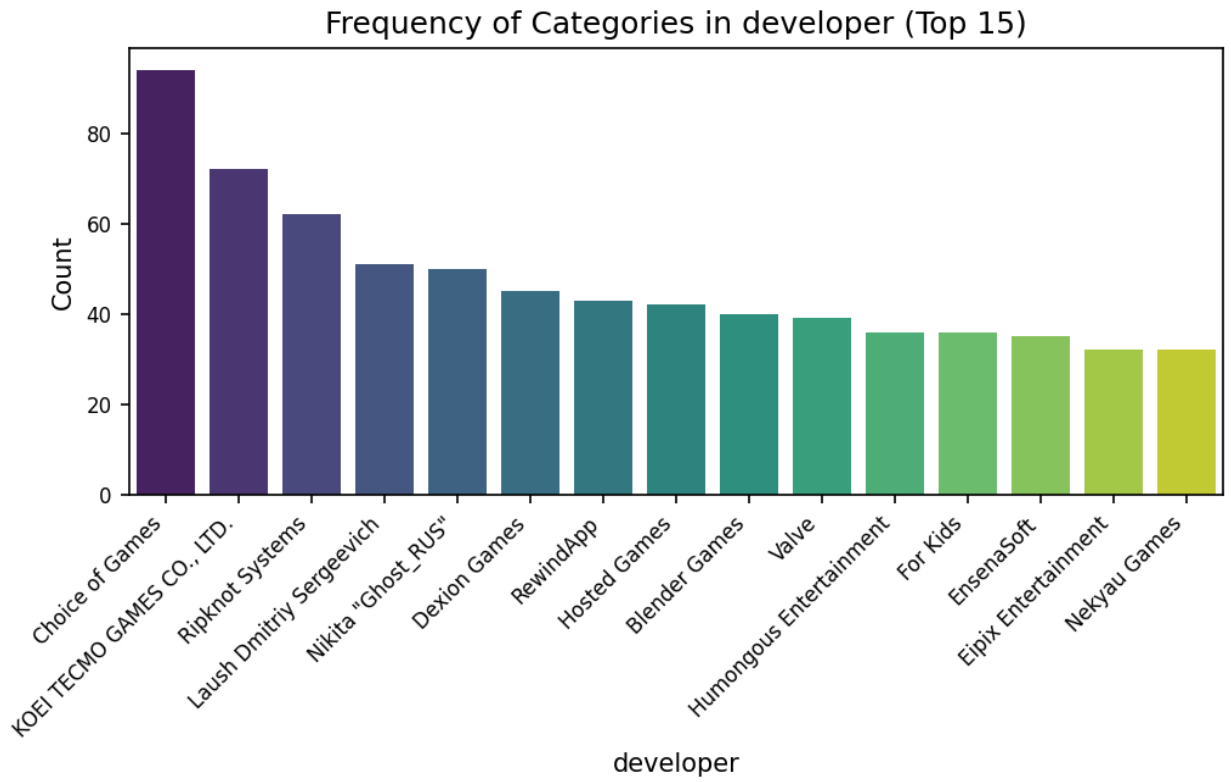

Figure 9: *Bar chart showing frequency of top categories in 'developer'. Total unique values: 17112.*

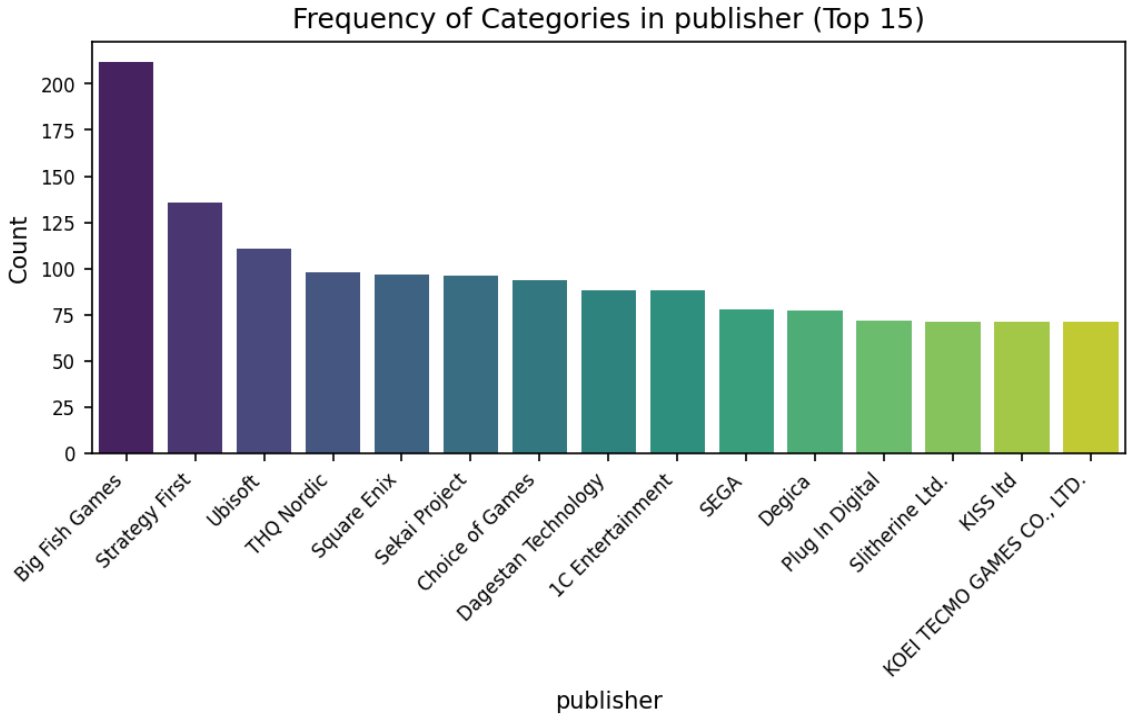

Figure 10: *Bar chart showing frequency of top categories in 'publisher'. Total unique values: 14353.*

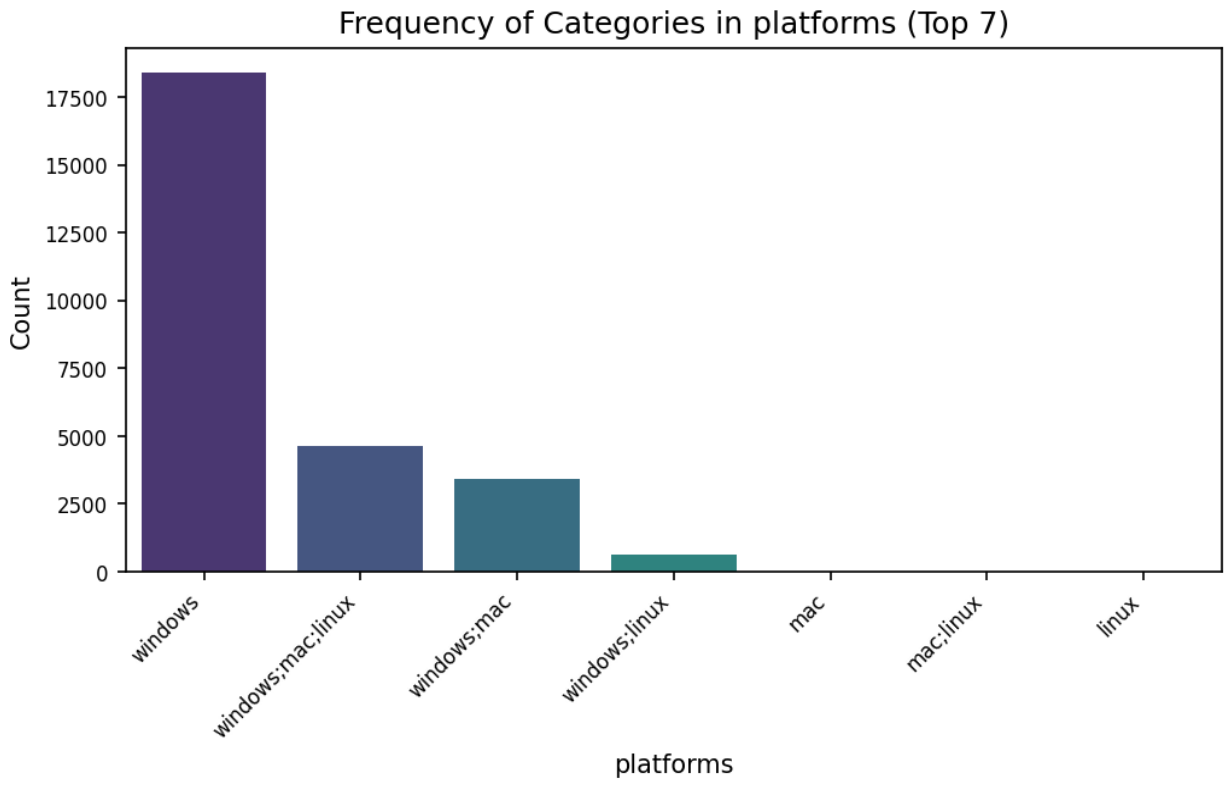

Figure 11: *Bar chart showing frequency of top categories in 'platforms'. Total unique values: 7.*

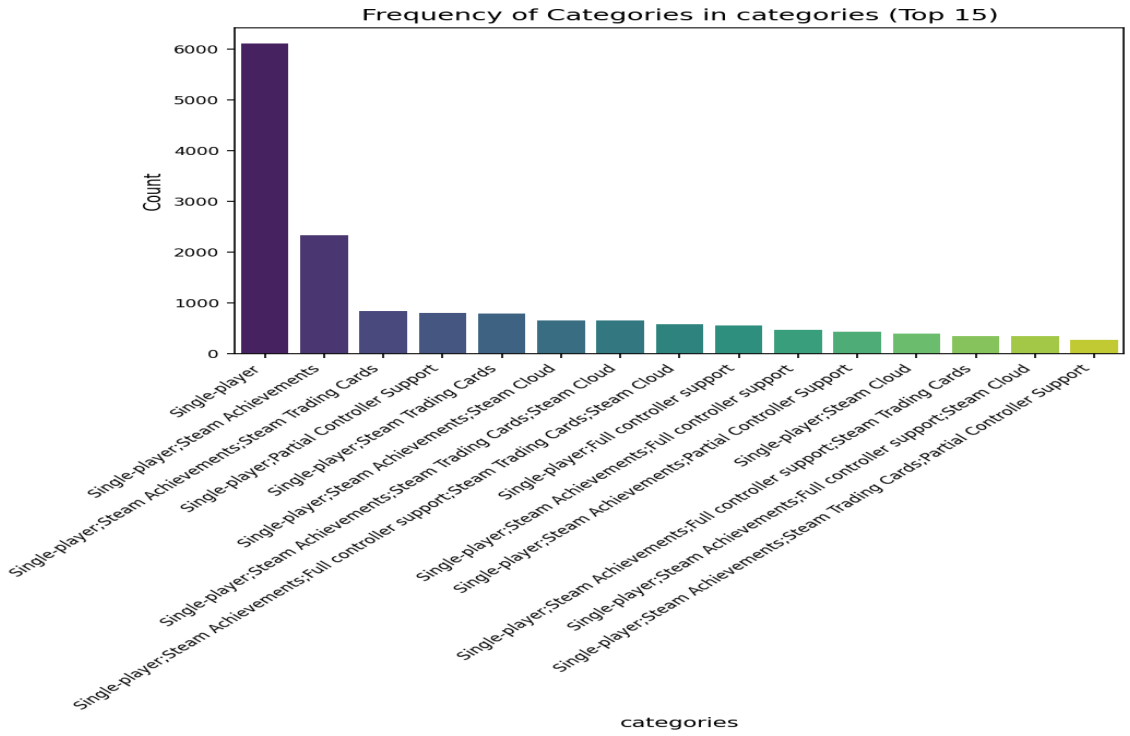

Figure 12: *Bar chart showing frequency of top categories in 'categories'. Total unique values: 3333.*

Observations on Categorical Feature Distributions:

The analysis of categorical features reveals a significant variation in cardinality and distribution. Features like 'name', 'developer', and 'publisher' exhibit extremely high cardinality, with tens of thousands of unique values. This high dimensionality presents a substantial challenge for model training, potentially leading to overfitting and increased computational cost. In contrast, 'platforms' has very low cardinality (only 7 unique values), suggesting it might be a relatively straightforward feature to incorporate into analysis. The distribution of values within these features is also uneven; while 'platforms' shows a strong dominance of 'windows' (67.9%), other features like 'name', 'release_date', 'developer', and 'publisher' have top categories representing only a tiny fraction (0.1% to 0.8%) of their respective total unique values. This indicates a long tail distribution, where a small number of categories dominate, while the vast majority are sparsely populated. The feature 'categories' occupies a middle ground, with a moderate cardinality (3333 unique values) and a dominant category ('Single-player' at 22.6%). This suggests that while encoding might still require careful consideration, the distribution is less skewed than the high-cardinality features. The high cardinality of several features necessitates careful consideration of feature encoding strategies. One-hot encoding would create an excessively large number of sparse columns, leading to computational inefficiency and potential overfitting. Techniques like target encoding, frequency encoding, or embedding layers (especially for neural networks) would be more appropriate for handling these features. For features with low cardinality like 'platforms', one-hot encoding might be perfectly suitable. In summary, the data reveals a clear need for dimensionality reduction and careful feature engineering. High-cardinality features require sophisticated encoding methods to avoid the curse of dimensionality, while features with skewed distributions might benefit from techniques that capture the importance of dominant categories. The analysis highlights the importance of understanding the unique characteristics of each categorical feature before proceeding with model building.

4. Bivariate Analysis

4.1. Numerical vs. Numerical Features

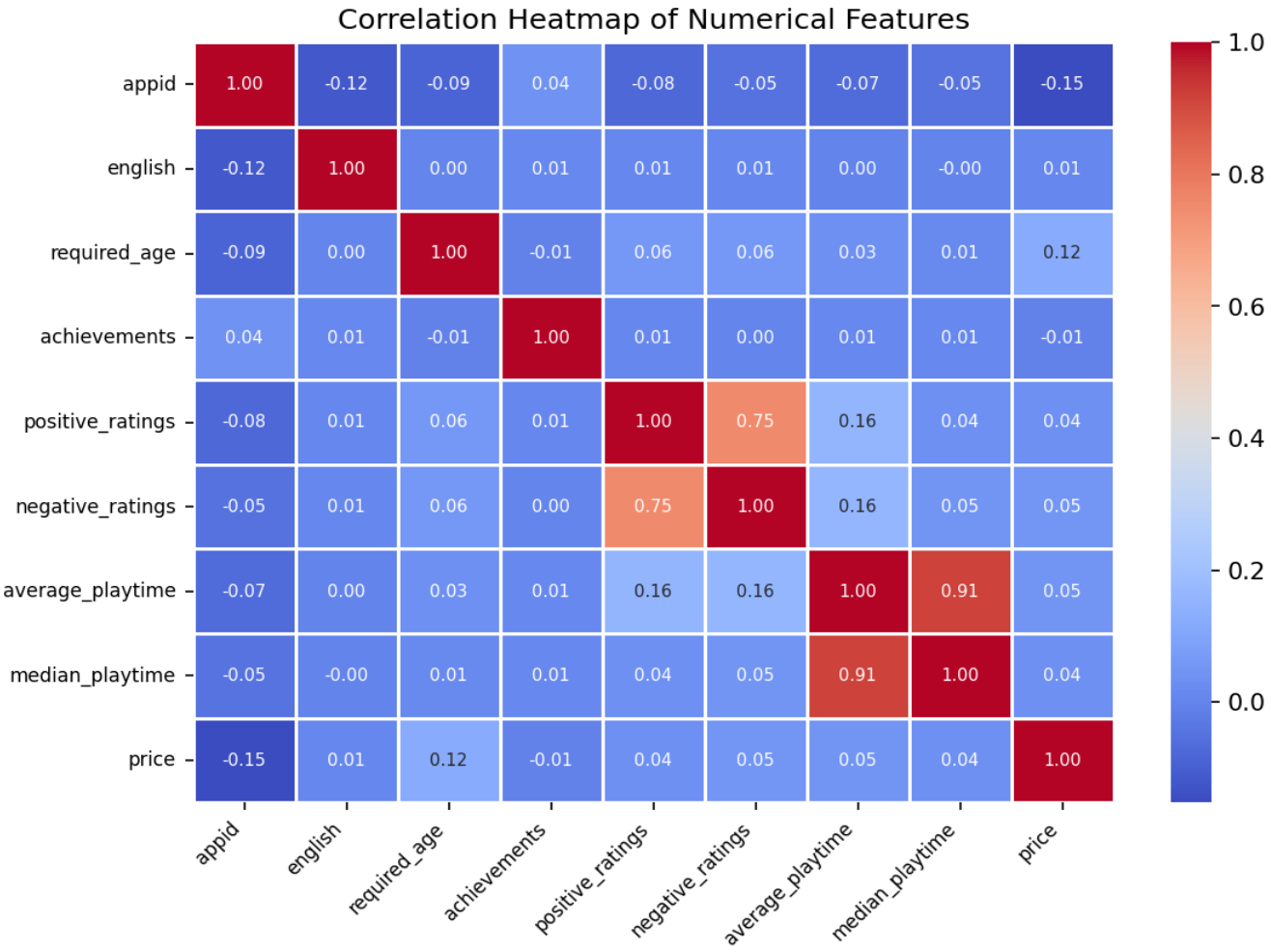

Figure 13: *Heatmap visualizing linear correlations (Pearson's r) between numerical features.*

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

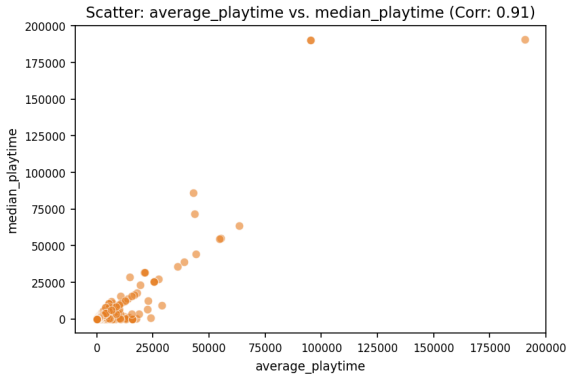

Figure 14: *Scatter plot for 'average_playtime' and 'median_playtime'. Correlation: 0.91.*

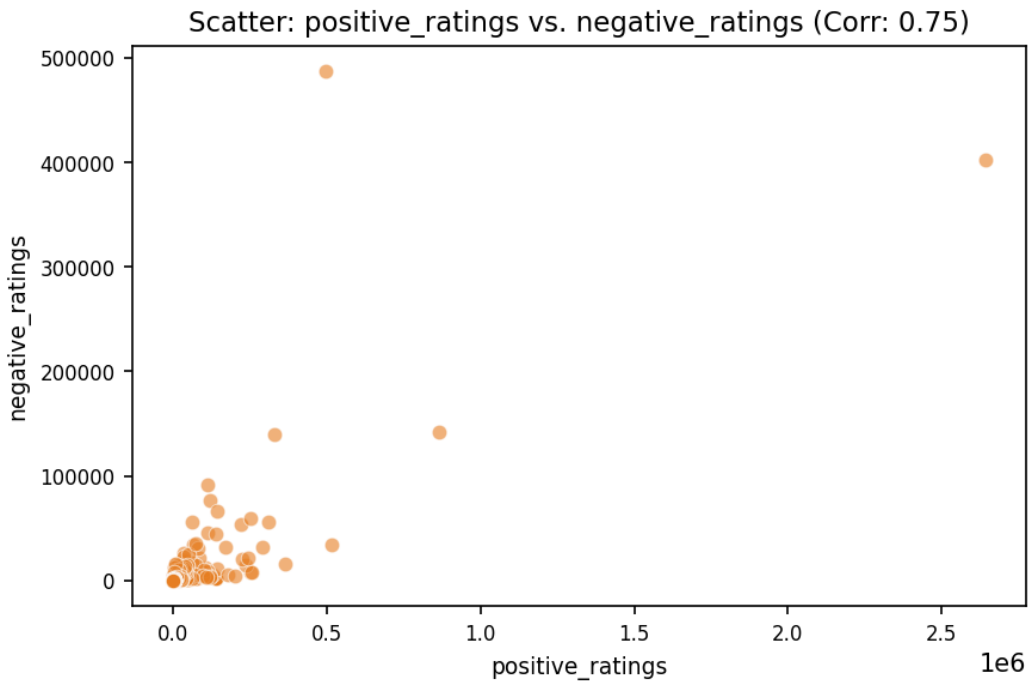

Figure 15: *Scatter plot for 'positive_ratings' and 'negative_ratings'. Correlation: 0.75.*

Interpretation of Numerical Correlations:

A correlation matrix displays the pairwise correlations between multiple variables. Each cell in the matrix shows the correlation coefficient (ranging from -1 to +1) between two variables. A value of +1 indicates a perfect positive correlation (as one variable increases, the other increases proportionally), -1 indicates a perfect negative correlation (as one increases, the other decreases proportionally), and 0 indicates no linear correlation. The provided analysis reveals several correlations, with two standing out as particularly strong. The strongest positive correlation is between `average_playtime` and `median_playtime` (0.91). This very high correlation suggests that games with longer average playtime also tend to have longer median playtime. This is intuitive; if a game has a long average playtime, it's likely that a significant portion of players are also spending a considerable amount of time playing, resulting in a similarly high median. Another strong positive correlation is observed between `positive_ratings` and `negative_ratings` (0.75). This suggests that games with a higher number of positive ratings also tend to have a higher number of negative ratings. This might indicate that popularity (more ratings in general) is the driving factor; games with a larger player base are more likely to receive both positive and negative feedback. The relatively weak correlation between `negative_ratings` and `average_playtime` (0.16) suggests little relationship between these two variables. The scatter plots likely visually confirm these relationships, showing a tight clustering of points for the first two correlations and a more dispersed pattern for the last one.

4.2. Numerical vs. Categorical Features

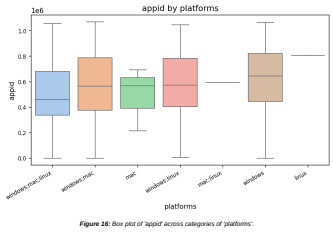

Figure 16: Box plot of 'appid' across categories of 'platforms'.

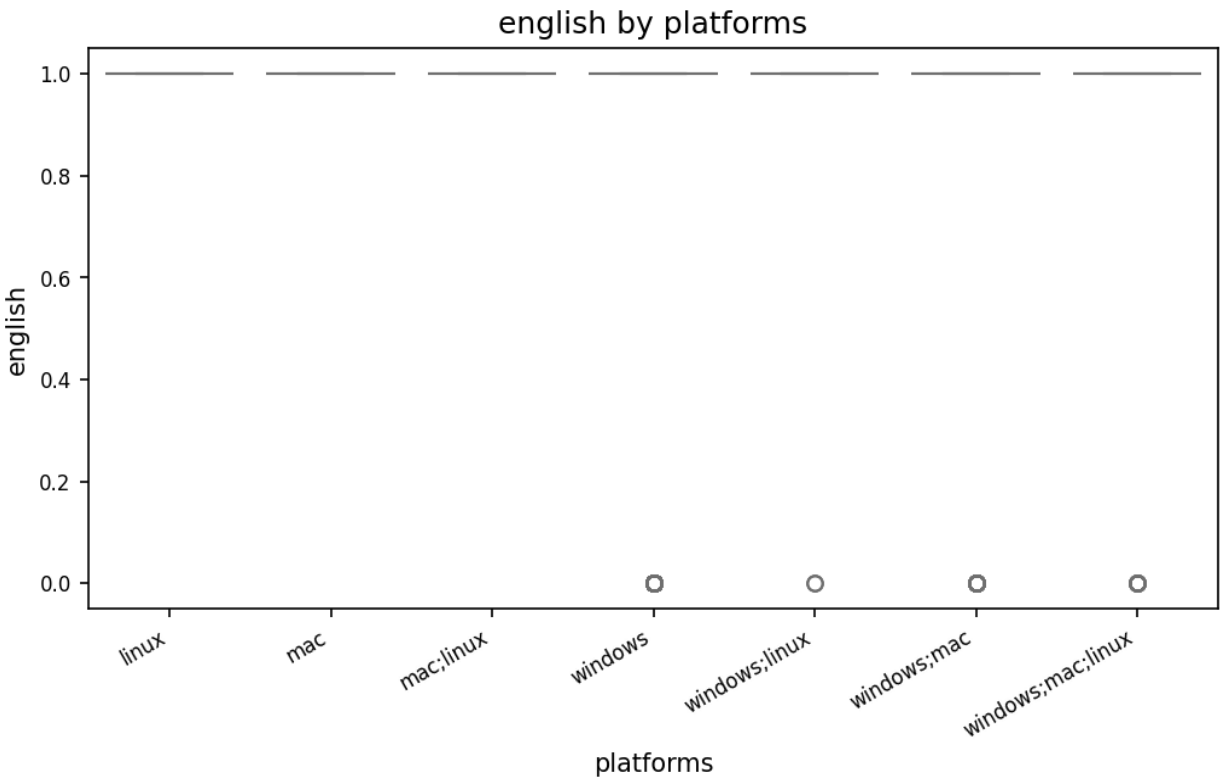

Figure 17: *Box plot of 'english' across categories of 'platforms'.*

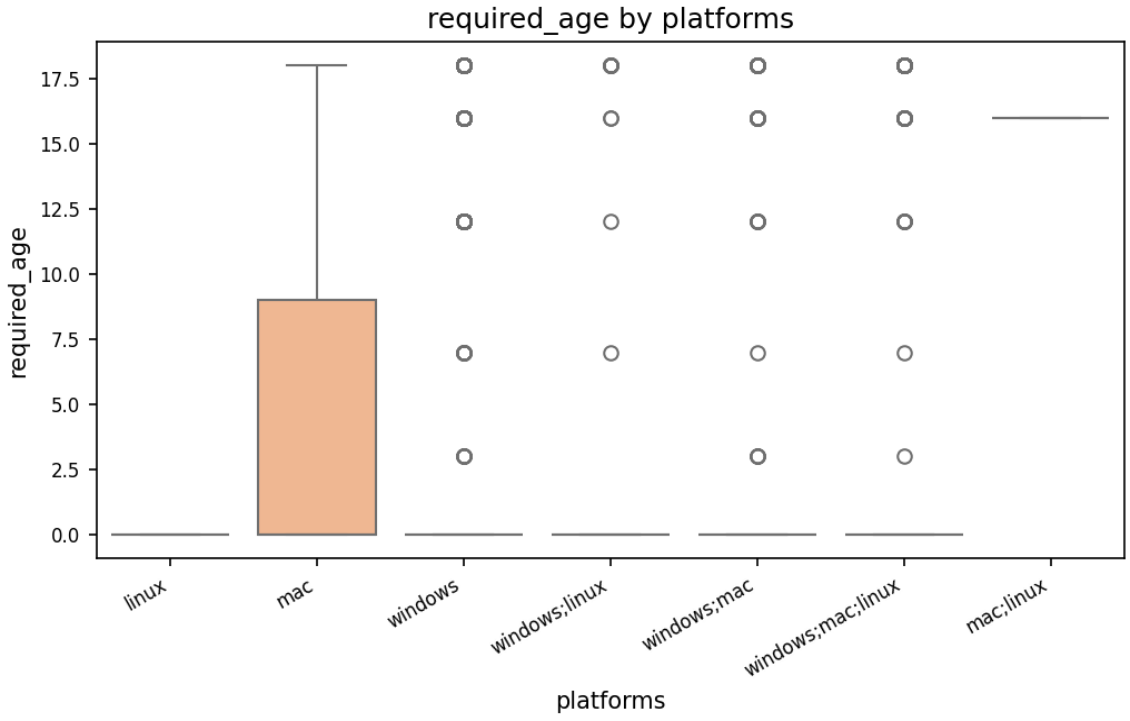

Figure 18: *Box plot of 'required_age' across categories of 'platforms'.*

Interpretation of Numerical vs. Categorical Interactions:

Box plots visualizing numerical distributions across categories offer a powerful way to compare the central tendency and variability of a numerical variable within different groups. They reveal not only the median (the middle value) of each category but also the interquartile range (IQR, the spread of the middle 50% of the data), outliers, and the minimum and maximum values. By comparing the boxes and whiskers across categories, we can quickly assess whether the distributions are similar or significantly different. For example, a box plot of 'appid' by 'platforms' might show whether certain platforms tend to have higher or lower application IDs, potentially indicating differences in the number of apps available on each platform. Similarly, a plot of 'english' (perhaps representing a score or rating) by 'platforms' could highlight whether user ratings or English language proficiency differs across platforms. Significant differences observed in the medians suggest a systematic difference in the average value of the numerical variable between categories. For instance, if the median 'english' score is considerably higher for platform A than for platform B, this suggests users on platform A tend to exhibit higher English proficiency. Differences in the spread (IQR or the presence of many outliers) indicate varying levels of consistency or variability within each category. A large IQR for platform C in the 'english' score box plot might suggest a wider range of English proficiency levels among users on that platform compared to platforms with smaller IQRs. In short, comparing medians reveals differences in central tendency, while comparing spreads reveals differences in data variability across categories.

4.3. Categorical vs. Categorical Features

5. Key Findings & Insights Summary

Key Findings & Insights The automated analysis of the `temp_steam.csv` dataset revealed a dataset comprising 270,88 rows and 18 columns, evenly split between numerical and categorical features. A notable data quality issue was the presence of 13 duplicate rows, which may indicate data entry errors or inconsistencies requiring further investigation. The absence of missing values is positive, simplifying subsequent analysis. No constant columns were identified, suggesting all features contribute some level of variance to the dataset. Univariate analysis examined the distributions of the nine numerical and nine categorical features. While the specific details of these distributions are not provided in the log, the analysis successfully covered all features. Further detailed reports on feature skewness, central tendency, and categorical frequency distributions would be needed to draw more specific conclusions about the characteristics of the data. Bivariate analysis explored relationships between feature pairs, revealing several observations though specifics remain undisclosed. The log only mentions that various feature pairs were analyzed and two observations were made, suggesting a need for a more comprehensive report detailing the nature and strength of these relationships (e.g., correlations, dependencies). The limited information prevents a deeper understanding of the interactions between features. The analysis revealed no explicitly mentioned surprising or unexpected findings. However, the limited information provided in the log about the univariate and bivariate analyses prevents a conclusive assessment of whether any surprising patterns were present. A more detailed report with specific findings from these analyses is necessary for a complete understanding of the dataset.

6. Conclusion & Potential Next Steps

This automated report provides a foundational, high-level overview of the `temp_steam.csv` dataset, highlighting its structure, data quality (with only 13 duplicates identified), and initial observations from univariate and bivariate analyses. This initial assessment establishes a solid basis for further, more in-depth investigation. Given the report's findings of 13 duplicates in a dataset of 27,088 rows, a first step should be to **investigate and resolve the 13 duplicate rows**. This involves determining the nature of the duplicates (exact duplicates or near duplicates with slight variations) and deciding whether to remove them, keep one representative row, or merge them based on relevant information. Next, the report mentions "various feature pairs" were analyzed in the bivariate analysis, but lacks specific details on the findings. Therefore, a crucial next step is to **generate a detailed report on bivariate analysis results**, including visualizations (scatter plots, correlation matrices, etc.) and statistical measures (correlation coefficients) for all feature pairs. This will highlight potential relationships between variables for further investigation. This is especially important given the limited information provided about the bivariate analysis' "observations gathered: 2". Finally, given the dataset's size and the presence of both numerical and categorical features, a significant next step would be to **perform more robust statistical analysis**. This could involve testing for significant differences between groups within categorical variables using ANOVA or t-tests (if relevant numerical features exist), or exploring more sophisticated modeling techniques (regression, classification) to uncover more complex relationships within the data, depending on the research goals.