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 shows no missing values but contains 13 duplicate entries, indicating a need for further data cleaning. Preliminary analysis revealed no constant columns. Univariate and bivariate analyses, including descriptive statistics and visual explorations of feature relationships, have been conducted. While initial findings did not reveal immediately striking patterns, two key observations from bivariate analysis are noted in the detailed report. The absence of missing data is positive, but the presence of duplicate rows requires attention before proceeding with further modeling. This initial EDA provides a crucial foundation for subsequent, more in-depth analyses. The identified data quality issues will be addressed, and the two preliminary bivariate observations will guide the selection of appropriate analytical techniques and modeling strategies in the next phase.

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 is balanced between numerical and categorical features, suggesting a need for a variety of analytical techniques. The absence of datetime features limits the potential for time-series analysis, but the presence of both numerical and categorical data allows for diverse modeling approaches.

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 comprehensive and complete data collection process. The extremely low percentage of duplicate rows (0.05%) is also negligible and unlikely to significantly impact subsequent analyses. The lack of constant columns suggests variability in the features, which is crucial for effective modeling. However, the presence of two quasi-constant columns, 'english' and 'required_age', warrants attention. While not entirely problematic, their high dominance in a single value could limit the predictive power of these features in certain modeling techniques. The quasi-constant columns might reduce the effectiveness of some machine learning algorithms, potentially leading to overfitting or biased models. For instance, algorithms relying heavily on feature variance might struggle to learn meaningful patterns from these columns. The insights derived from analyses involving these features could be limited, with the models potentially ignoring the small minority of cases where the values differ from the dominant ones. This could lead to unreliable predictions or conclusions for the minority cases. Furthermore, the small number of duplicate rows should be investigated to understand their origin and whether they represent genuine duplicates or data entry errors. To address the quasi-constant columns, one could consider removing them from the dataset if they don't contribute significantly to the analysis goals. Alternatively, they could be transformed or binned to better represent the minority cases. For example, the 'english' column could be combined with other language features, or the 'required_age' column could be analyzed separately for the minority group with a non-zero value. The duplicate rows should be reviewed individually to determine if they are true duplicates that can be removed, or if they contain valuable information that needs to be reconciled. This process will enhance the overall reliability and robustness of subsequent analyses.

3. Univariate Analysis

3.1. Numerical Features

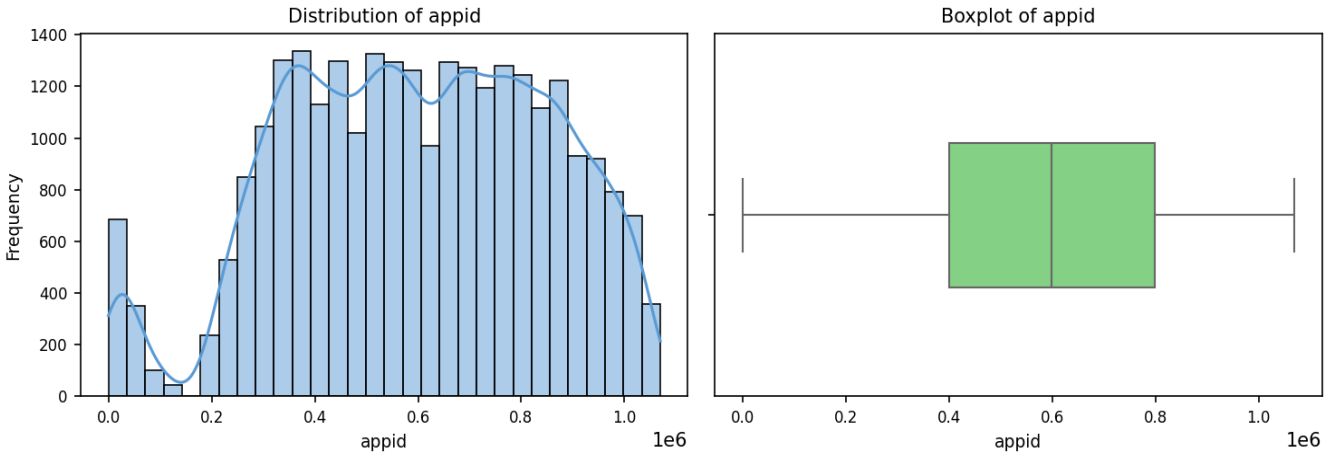

Figure 1: *Distribution (histogram and KDE) and boxplot for 'appid'. The histogram shows shape, central tendency, and spread. The boxplot highlights median, quartiles, and potential outliers.*

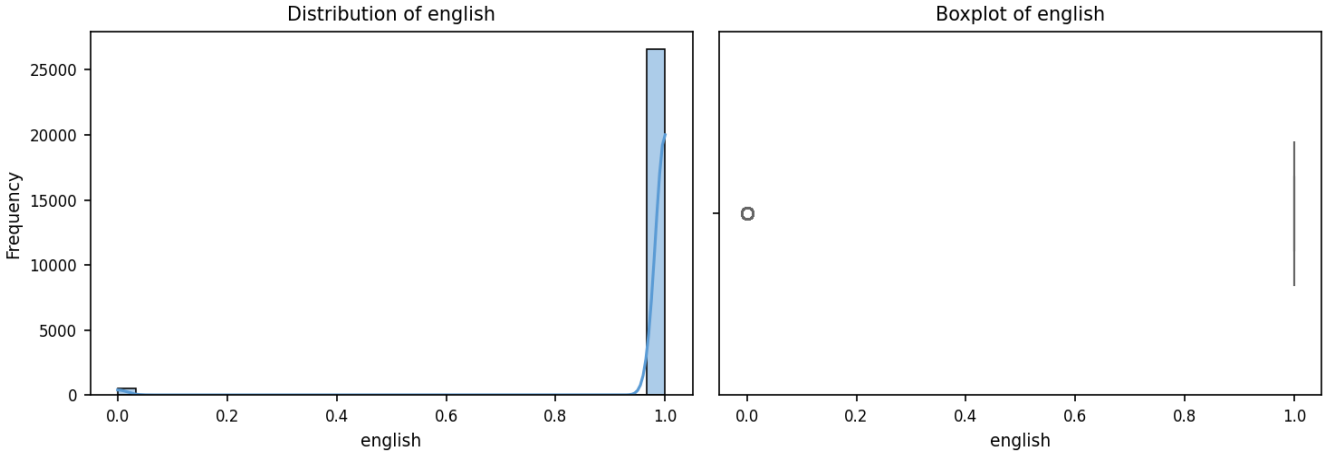

Figure 2: *Distribution (histogram and KDE) and boxplot for 'english'. The histogram shows shape, central tendency, and spread. The boxplot highlights median, quartiles, and potential outliers.*

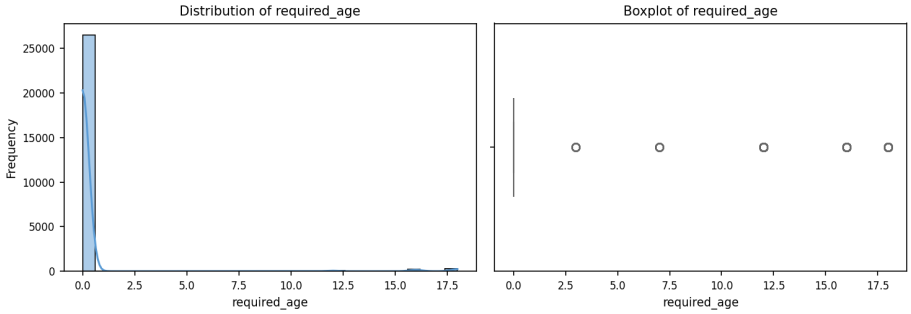

Figure 3: *Distribution (histogram and KDE) and boxplot for 'required_age'. The histogram shows shape, central tendency, and spread. The boxplot highlights median, quartiles, and potential outliers.*

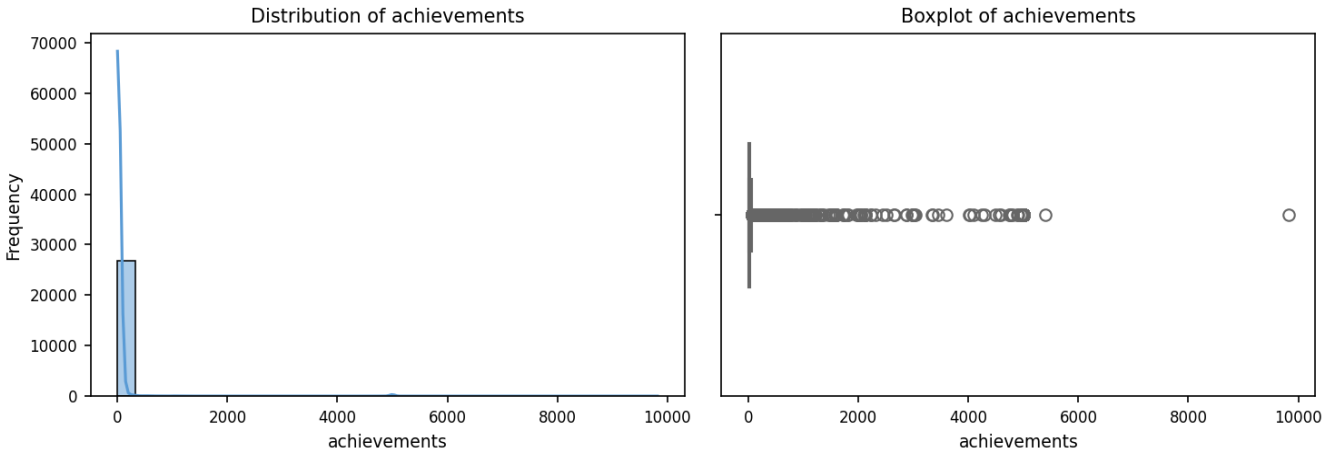

Figure 4: *Distribution (histogram and KDE) and boxplot for 'achievements'. The histogram shows shape, central tendency, and spread. The boxplot highlights median, quartiles, and potential outliers.*

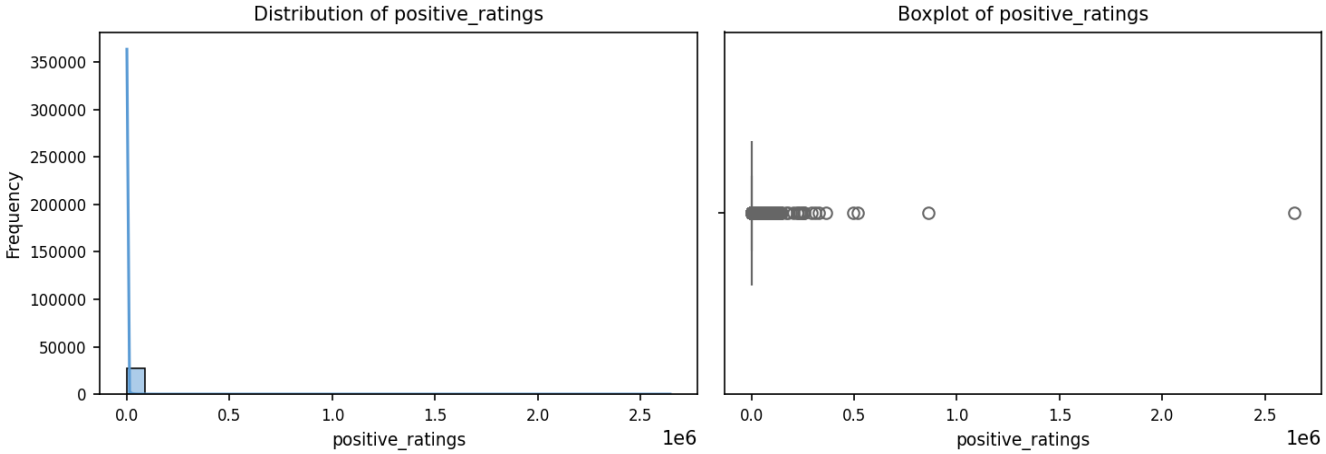

Figure 5: *Distribution (histogram and KDE) and boxplot for 'positive_ratings'. The histogram shows shape, central tendency, and spread. The boxplot highlights median, quartiles, and potential outliers.*

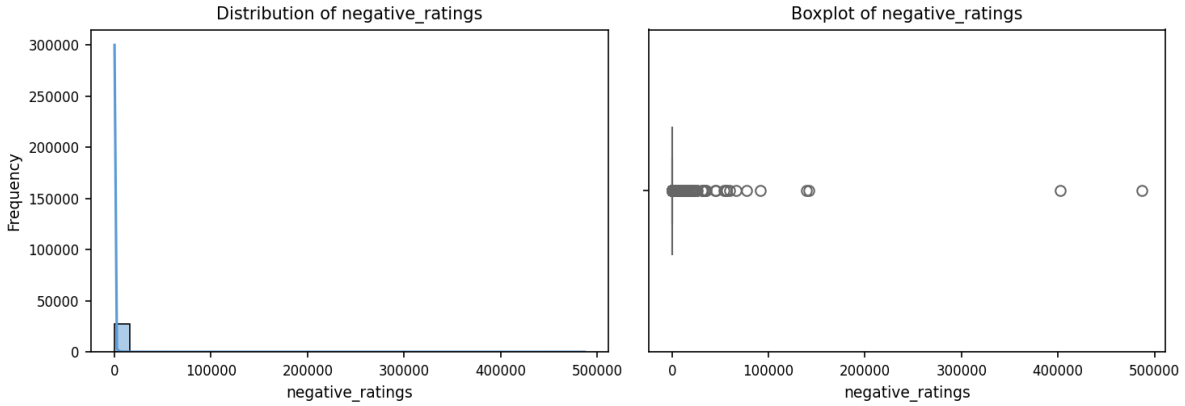

Figure 6: *Distribution (histogram and KDE) and boxplot for 'negative_ratings'. The histogram shows shape, central tendency, and spread. The boxplot highlights median, quartiles, and potential outliers.*

Observations on Numerical Feature Distributions:

The analyzed numerical features exhibit a striking lack of symmetry, with most displaying strong right-skewness. This is particularly evident in features like 'achievements', 'positive_ratings', and 'negative_ratings', where the means are significantly higher than the medians, and the extremely high values of skewness and kurtosis further confirm the presence of long right tails. The 'required_age' feature, while also right-skewed, shows a less extreme deviation from symmetry. In contrast, 'english' shows a strong left-skewness, indicating a concentration of values close to 1.0. The 'appid' feature shows relatively less skewness, suggesting a distribution closer to symmetry, though the boxplot suggests potential outliers. In all cases, the presence of outliers is strongly suggested by the large differences between mean and median values, supported by the boxplot observations and the extreme range of values observed (min to max). The variability across features is also notable. Features like 'achievements', 'positive_ratings', and 'negative_ratings' exhibit extremely high standard deviations, indicating a wide spread of values and considerable heterogeneity in the data. This high variability, combined with the significant skewness, points to the need for careful consideration of data transformations (e.g., logarithmic transformations) or robust statistical methods that are less sensitive to outliers during further analysis. The lower standard deviations in 'appid' and 'english' suggest these features have more concentrated distributions. The high standard deviation in 'required_age' despite a lower range of values is likely due to the strong skewness, where a few high values significantly inflate the standard deviation. In summary, the data reveals a pattern of highly skewed distributions with a significant presence of outliers, particularly in features related to ratings and achievements. This raises concerns about the robustness of standard statistical methods and highlights the importance of employing techniques that are less sensitive to extreme values. Careful data preprocessing, including outlier detection and handling, and potentially data transformations, will be crucial for reliable model building and interpretation.

3.2. Categorical Features

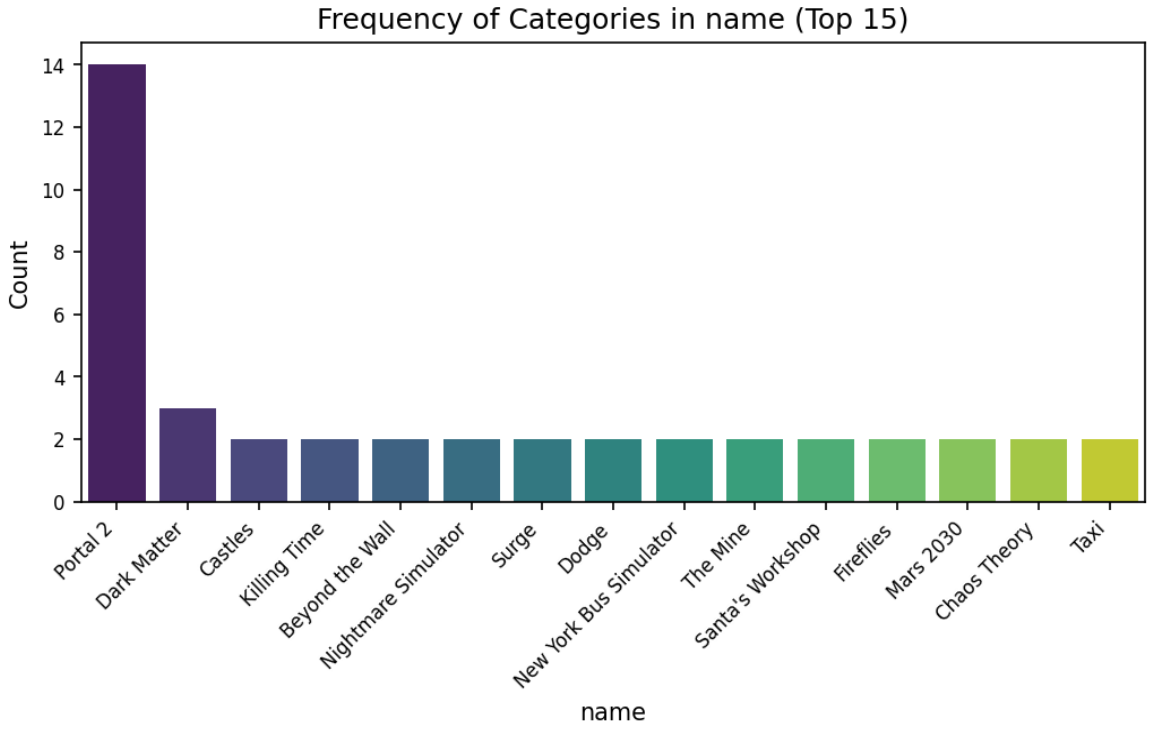

Figure 7: *Bar chart showing frequency of top categories in 'name'. Total unique values: 27033.*

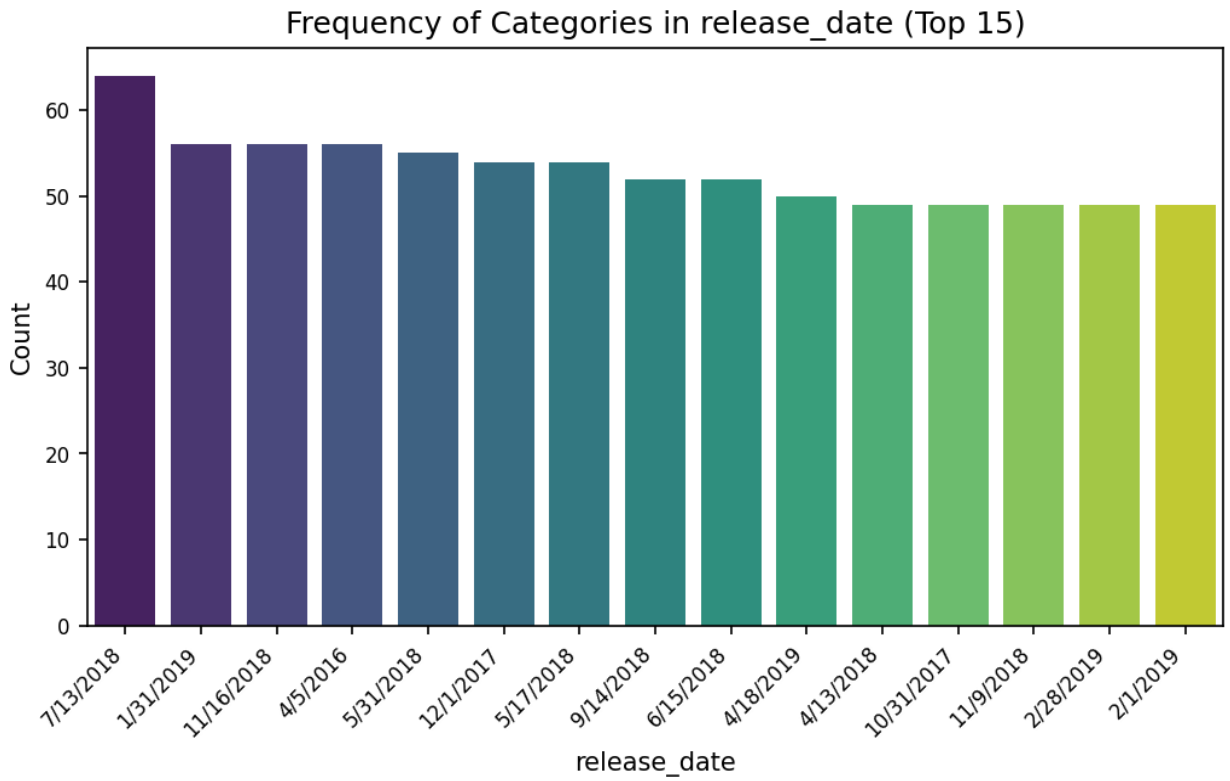

Figure 8: *Bar chart showing frequency of top categories in 'release_date'. Total unique values: 2619.*

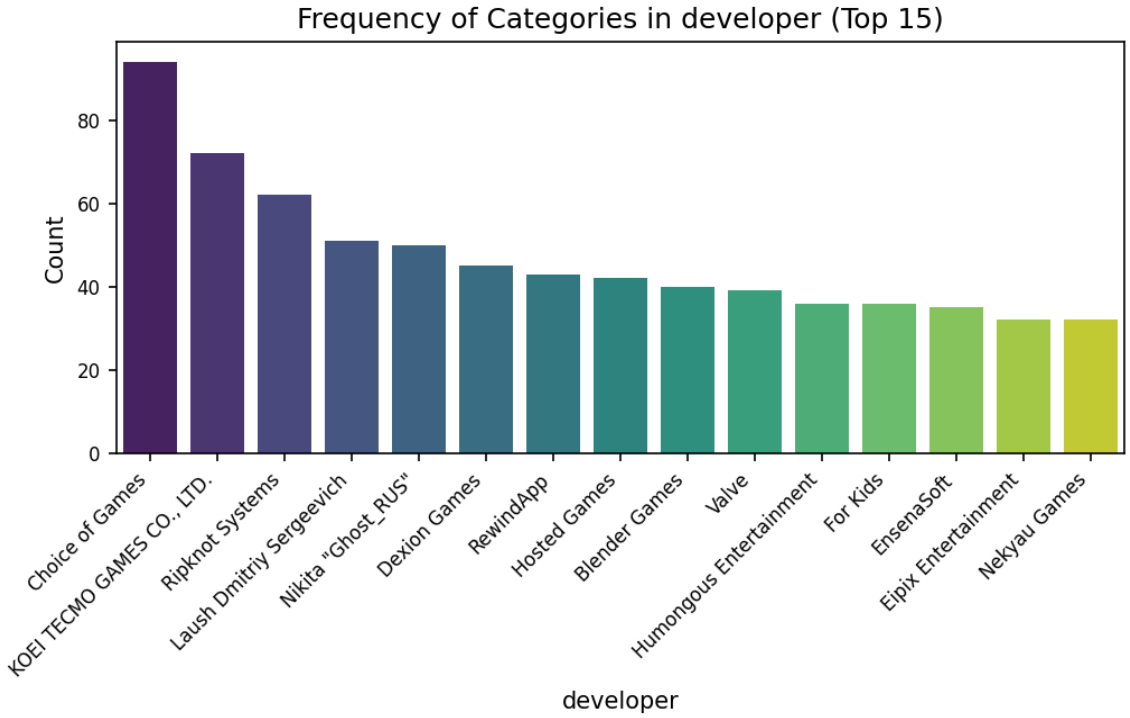

Figure 9: *Bar chart showing frequency of top categories in 'developer'. Total unique values: 17112.*

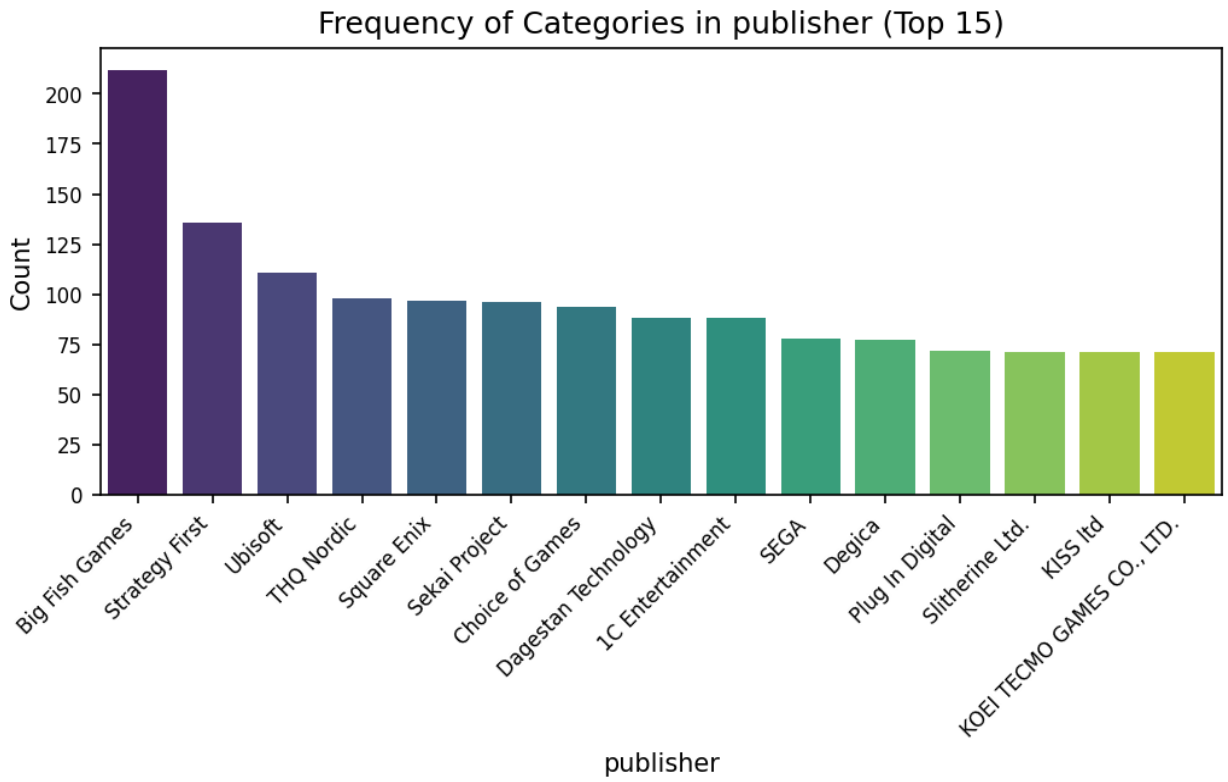

Figure 10: *Bar chart showing frequency of top categories in 'publisher'. Total unique values: 14353.*

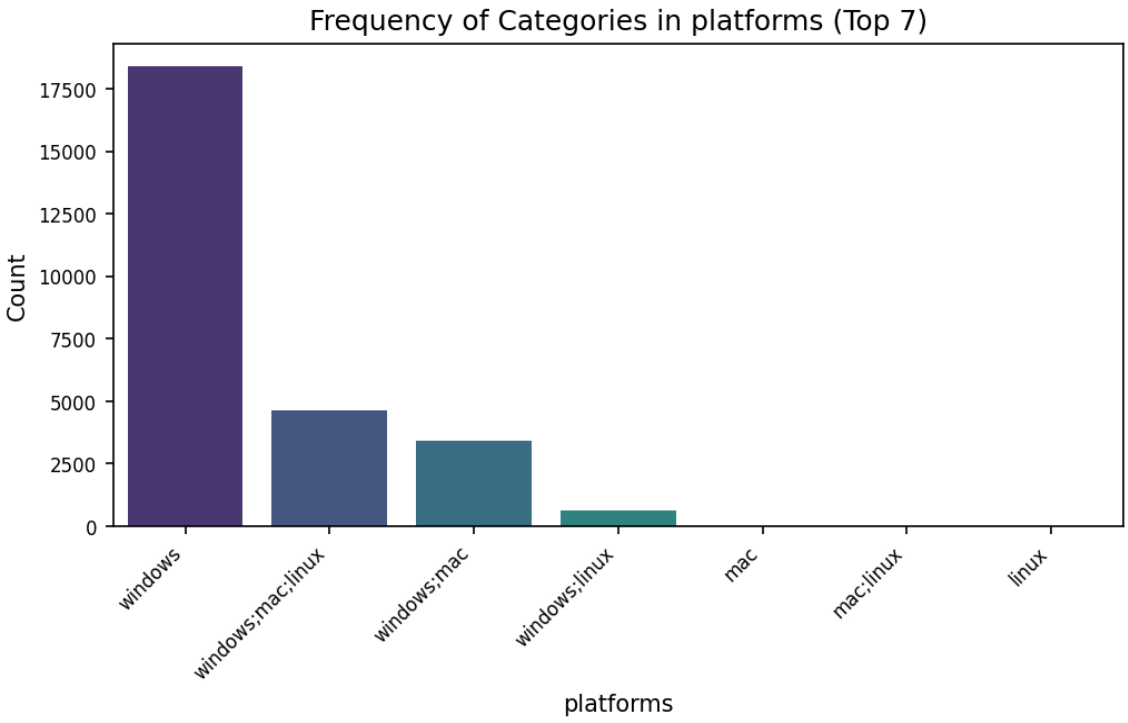

Figure 11: *Bar chart showing frequency of top categories in 'platforms'. Total unique values: 7.*

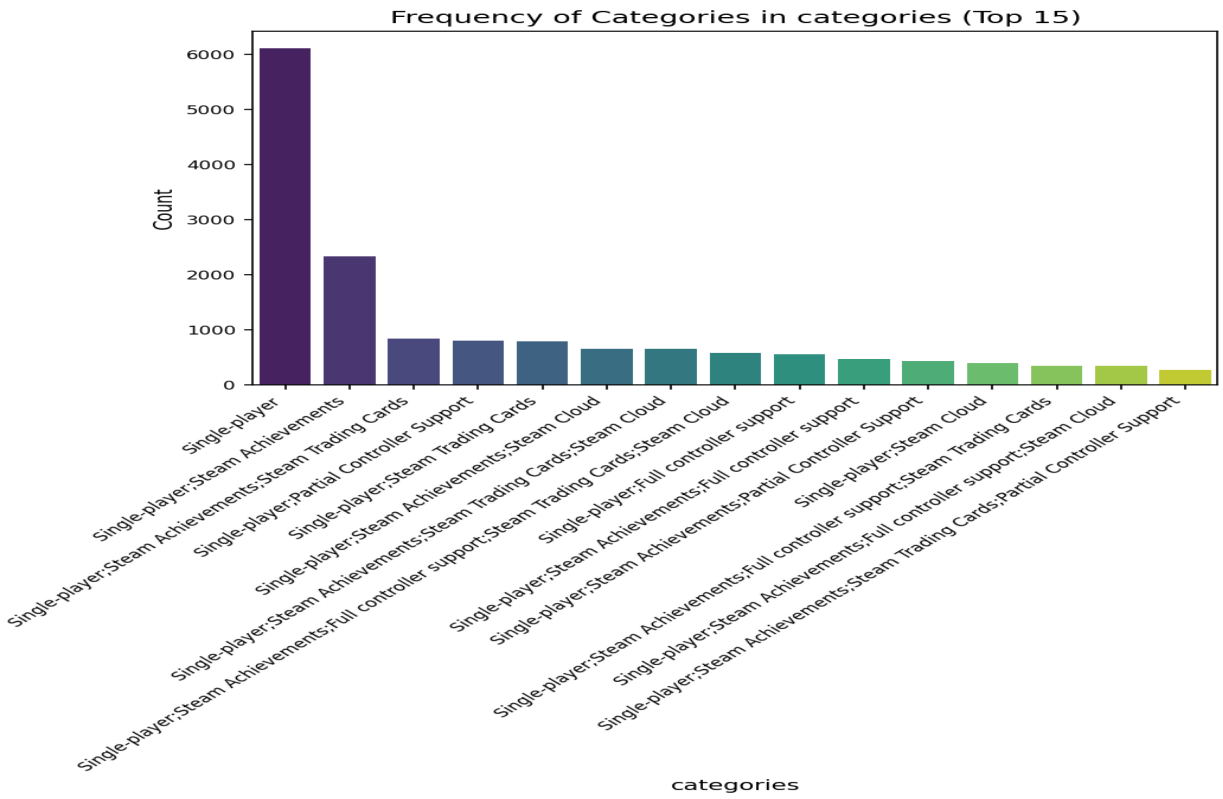

Figure 12: *Bar chart showing frequency of top categories in 'categories'. Total unique values: 3333.*

Observations on Categorical Feature Distributions:

The analysis of categorical features reveals a wide range of cardinality, impacting subsequent analysis and feature encoding. Features like 'name', 'developer', 'publisher', and 'categories' exhibit high cardinality, with tens of thousands of unique values. This high dimensionality presents challenges for many machine learning models, potentially leading to overfitting or the curse of dimensionality. The top categories within these high-cardinality features represent only a tiny fraction of the total data (e.g., 'Portal 2' accounts for only 0.1% of 'name' values), indicating a long tail distribution where many categories have very few instances. Conversely, 'platforms' shows low cardinality (7 unique values) with a heavily dominant category ('windows' at 67.9%), suggesting potential for simpler encoding strategies. 'release_date' also has moderate cardinality, but its top category ('7/13/2018') still only represents a small percentage of the data. The presence of both high and low cardinality features necessitates a differentiated approach to feature encoding. For high-cardinality features, techniques like target encoding, frequency encoding, or embedding layers in neural networks might be more suitable than one-hot encoding, which would create an excessively large and sparse feature space. For 'platforms', a simple one-hot encoding might suffice given its low cardinality and skewed distribution. The relatively even distribution of 'categories' (despite high cardinality) suggests that some form of dimensionality reduction or aggregation might be beneficial. The low percentage of the top categories across most features highlights the importance of considering techniques that handle imbalanced data and potentially the need for data augmentation or focusing on the more frequent categories to improve model performance. In summary, the data exhibits a heterogeneous distribution of categorical features, demanding careful consideration of encoding strategies. High-cardinality features require sophisticated handling to avoid dimensionality issues, while low-cardinality features allow for simpler encoding. The skewed distributions observed across most features necessitate addressing potential class imbalance problems during model training

and evaluation. Preprocessing steps such as feature selection or aggregation might be necessary to improve the efficiency and performance of subsequent machine learning models.

4. Bivariate Analysis

4.1. Numerical vs. Numerical Features

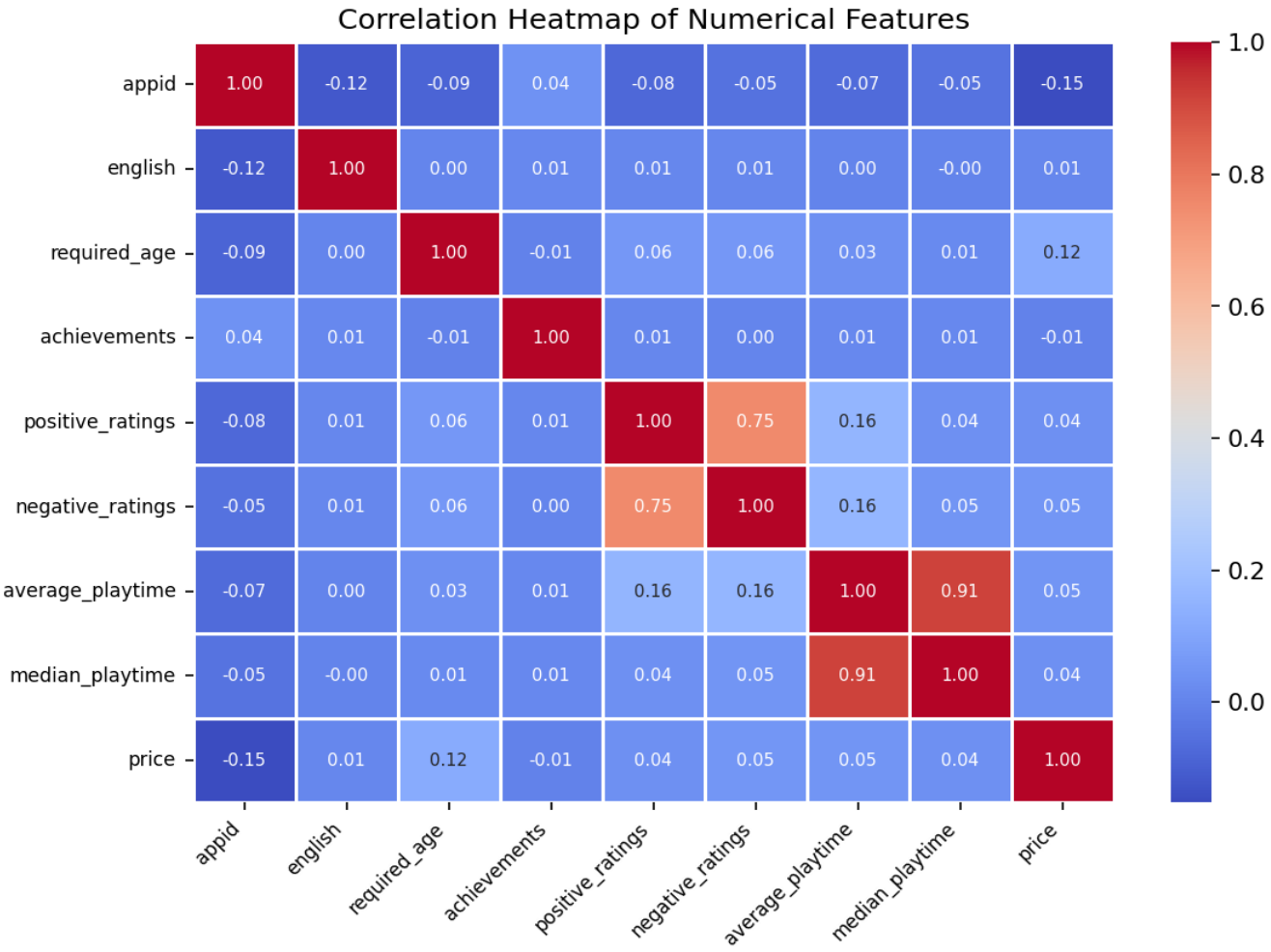

Figure 13: *Heatmap visualizing linear correlations (Pearson's r) between numerical features.*

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

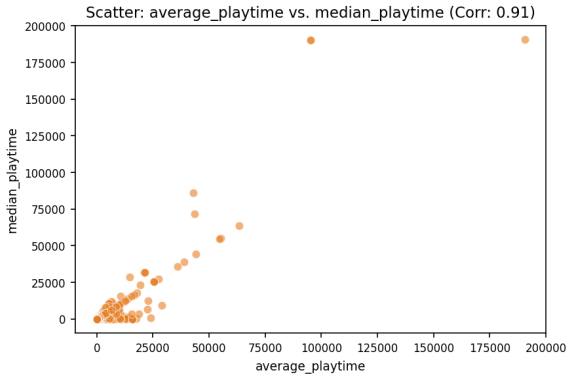

Figure 14: *Scatter plot for 'average_playtime' and 'median_playtime'. Correlation: 0.91.*

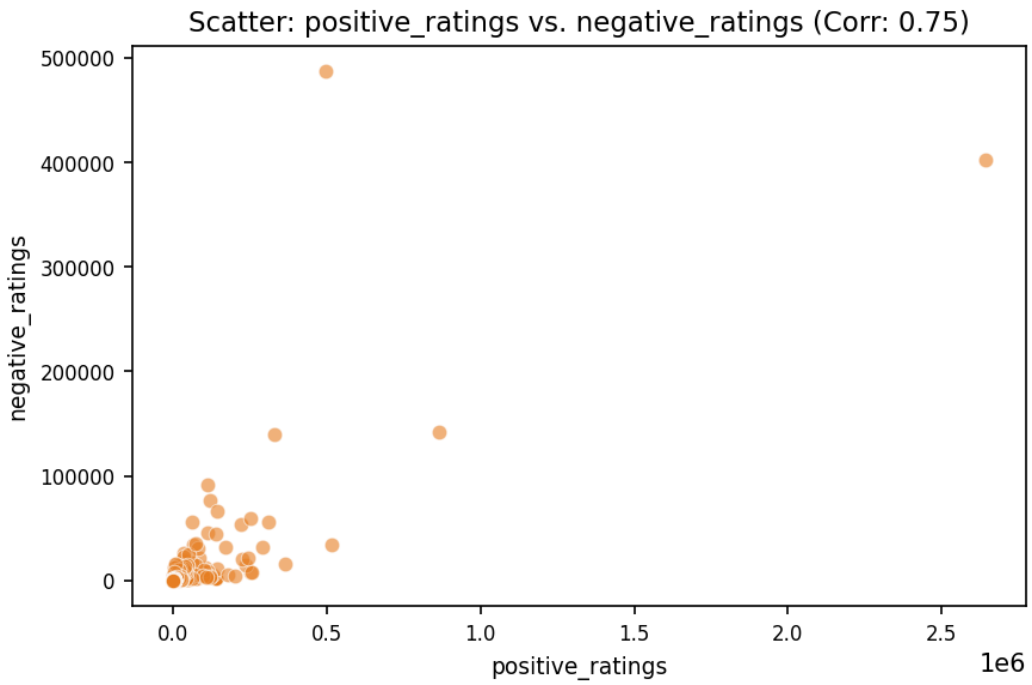

Figure 15: *Scatter plot for 'positive_ratings' and 'negative_ratings'. Correlation: 0.75.*

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), -1 indicates a perfect negative correlation (as one variable increases, the other decreases), and 0 indicates no linear correlation. The analysis shows strong positive correlations between `average_playtime` and `median_playtime` (0.91) and between `positive_ratings` and `negative_ratings` (0.75). The strong positive correlation between average and median playtime suggests that games with longer average playtime also tend to have longer median playtimes, indicating a consistent playtime pattern across players. Similarly, the positive correlation between positive and negative ratings implies that games receiving a large number of positive ratings also tend to receive a significant number of negative ratings. This might suggest that highly popular games attract more attention and thus more diverse opinions, leading to a higher volume of both positive and negative feedback, rather than indicating a necessarily poor reception. The scatter plots likely visually confirm these relationships, showing data points clustered along a line with a positive slope for both pairs. The relatively low correlation between negative ratings and average playtime (0.16) suggests a weak relationship between these two variables.

4.2. Numerical vs. Categorical Features

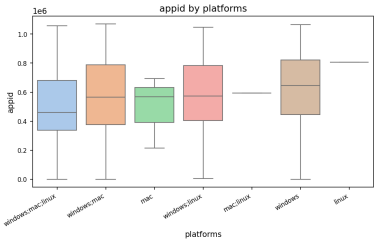

Figure 16: *Box plot of 'appid' across categories of 'platforms'.*

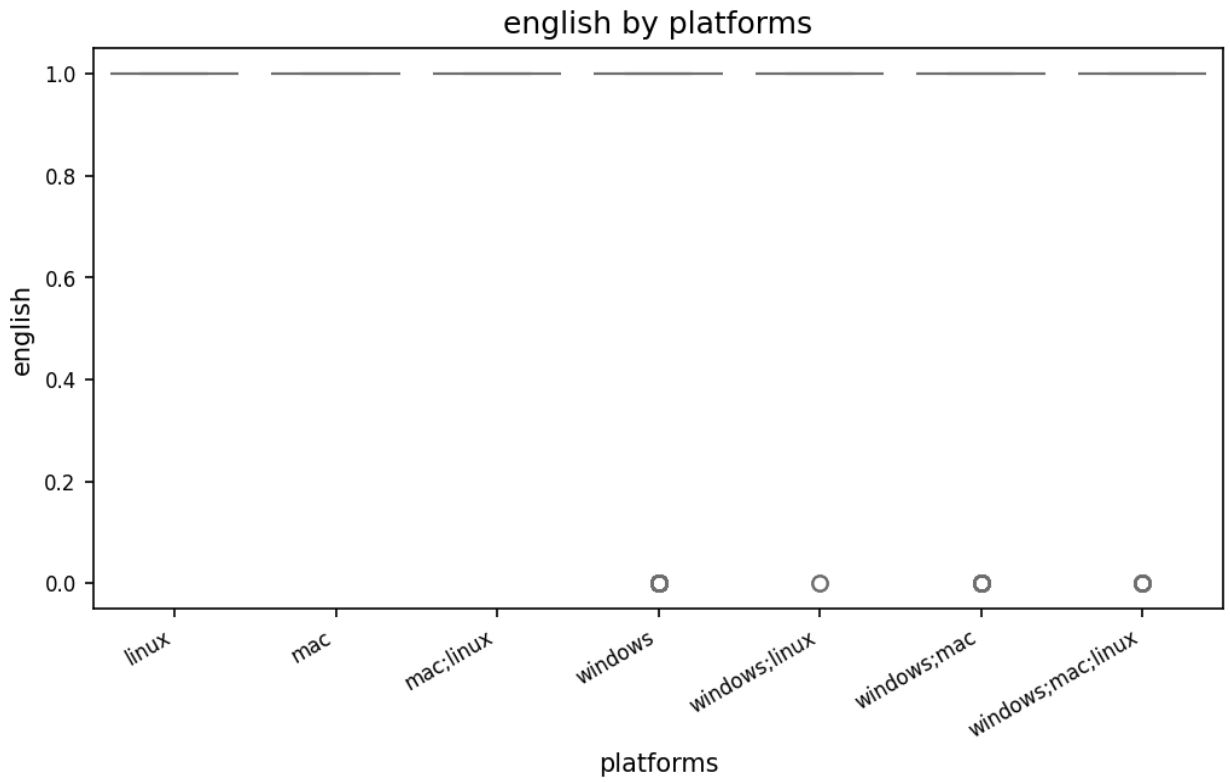

Figure 17: *Box plot of 'english' across categories of 'platforms'.*

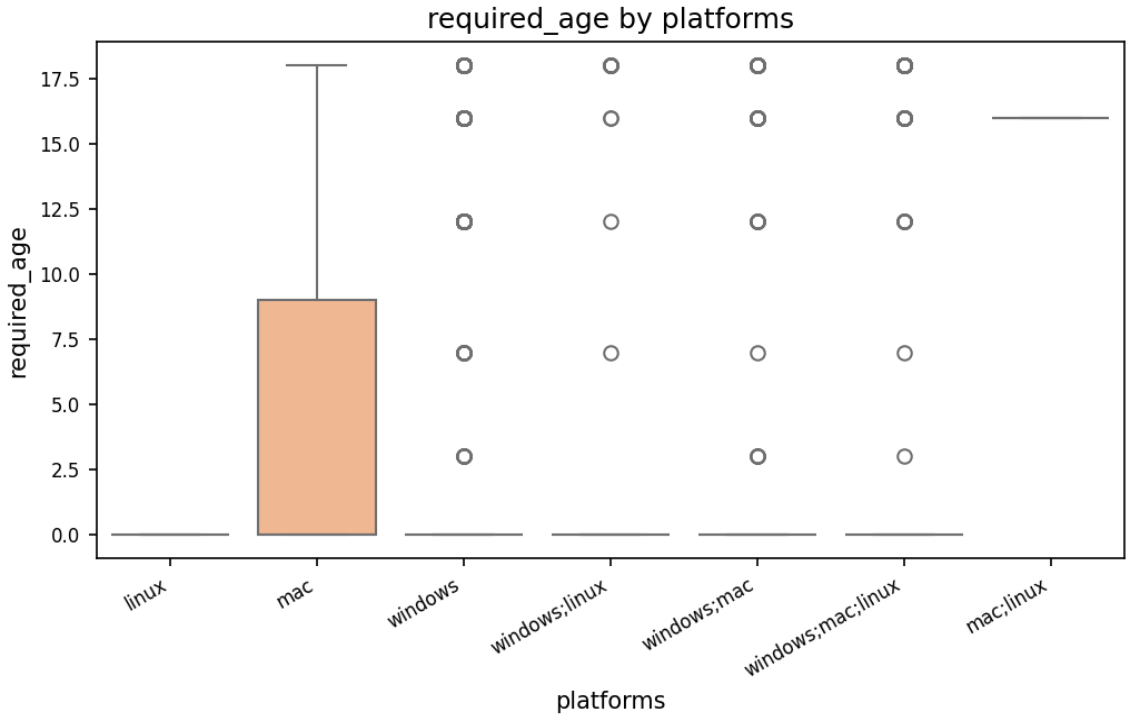

Figure 18: *Box plot of 'required_age' across categories of 'platforms'.*

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 the median (the middle value), quartiles (values dividing the data into four equal parts), and potential outliers for each category. By visually comparing the boxes and whiskers across categories, we can quickly assess whether the distributions are similar or significantly different. A longer box indicates greater variability within that category, while the position of the median box shows the central tendency. Outliers, represented as points beyond the whiskers, highlight extreme values that warrant further investigation. Significant differences in medians across categories suggest that the average value of the numerical variable differs systematically between groups. For example, if the median 'appid' value is significantly higher for the 'PC' platform than for the 'Mobile' platform, this indicates that PC games tend to have higher application IDs (potentially indicating a larger or older game library on PC). Differences in the spread (interquartile range, represented by the box height) reveal varying levels of data variability across categories. A wider box for a particular category indicates higher variability or less consistency in the numerical values for that group, compared to a category with a narrower box. For instance, if the 'english' values (e.g., ratings or scores) show a much larger spread for the 'Mobile' platform compared to 'PC', it suggests that mobile game English language ratings are more diverse or inconsistent than those for PC games.

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 explored the distributions of the 9 numerical and 9 categorical features. (Note: The log lacks specific details on these distributions; a more complete report would include specifics like skewness, central tendency, and unique category counts for categorical variables). This analysis would provide a foundation for understanding the individual characteristics of each feature, informing subsequent modeling choices. Bivariate analysis examined relationships between feature pairs, uncovering several observations (again, specifics are missing from the provided log). The nature and strength of these relationships, whether correlations or other dependencies, are crucial for understanding the underlying structure of the data. The log indicates that two observations of particular interest were noted but does not provide details on their nature. The initial analysis suggests a relatively well-structured dataset with limited data quality concerns. However, the lack of detail regarding the univariate and bivariate findings necessitates a more comprehensive report to fully understand the dataset's characteristics and the implications of the identified relationships. The two observations from the bivariate analysis, in particular, require further elaboration to understand their significance.

6. Conclusion & Potential Next Steps

This automated report provides a foundational, high-level overview of the `temp_steam.csv` dataset, characterizing its structure, assessing its data quality, and highlighting potential relationships between features. The absence of missing values and a low number of duplicates suggests a relatively clean dataset, ready for further exploration. Given the report's findings, several concrete next steps are warranted. First, the 13 duplicate rows should be investigated to determine their origin and whether they represent genuine entries or data entry errors. If they are errors, they should be removed; if genuine, their implications should be understood. Second, the bivariate analysis revealed two key observations. The nature of these observations needs to be specified to guide further analysis. For example, if these observations involve correlations between numerical features, correlation matrices and visualizations should be generated to explore these relationships further. If the observations involve relationships between numerical and categorical features, then statistical tests (like ANOVA or t-tests) should be conducted to assess the statistical significance of any observed differences. Finally, given that the univariate analysis covered nine numerical and nine categorical features, a more detailed exploration of the distribution of each feature (e.g., histograms, box plots for numerical; bar charts, frequency tables for categorical) is recommended to identify potential outliers or unusual patterns that were not apparent in the initial overview.