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 27088 rows and 18 columns (9 numerical, 9 categorical). Preliminary analysis revealed a relatively clean dataset with only 13 duplicate entries and no missing or constant columns. This suggests a good foundation for further, more in-depth analysis. The EDA included univariate and bivariate analyses of all features, encompassing descriptive statistics and visual inspection to identify potential patterns and relationships. While no immediately striking patterns emerged from this initial scan, two noteworthy bivariate observations were recorded and require further investigation. This initial automated data exploration provides a solid baseline understanding of the dataset's structure and quality. The identified duplicate entries will be addressed, and the two significant bivariate observations will be prioritized for deeper analysis in subsequent reports, which will inform more targeted modeling and 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	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 has 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, potentially requiring different preprocessing and modeling 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 dataset with generally high quality. The absence of missing values is a significant positive, indicating a complete dataset ready for analysis. The extremely low rate of duplicate rows (0.05%) is also negligible and unlikely to significantly impact subsequent analyses. The lack of constant columns suggests a reasonable level of variability within the features. However, the presence of two quasi-constant columns, 'english' and 'required_age', warrants attention. While not problematic in themselves, their high dominance in a single value (98.1% and 97.8% respectively) suggests these features might offer limited predictive power in modeling tasks or may be redundant. The presence of quasi-constant columns could impact the reliability of insights derived from analyses, particularly those involving machine learning models. These features might not contribute meaningfully to model performance and could even negatively affect model training by potentially introducing bias or increasing computational complexity without providing additional information. Insights drawn from analyses exclusively focused on these variables would be limited and potentially misleading, reflecting the dominant value rather than true underlying patterns. Furthermore, the small number of duplicate rows should be investigated to understand their origin and ensure they are not indicative of a larger data entry or collection problem. To address these issues, the quasi-constant columns ('english' and 'required_age') should be carefully evaluated for their relevance to the research question. If they are deemed redundant or unhelpful, they can be removed from the dataset. If they are deemed important, exploring the minority class (values other than 1 for 'english' and 0 for 'required_age') might provide valuable insights. Further investigation into the nature of the 13 duplicate rows is warranted to determine if they represent genuine duplicates or errors. This could involve examining the unique identifiers within those rows to understand the source of the duplication. Removing these duplicates is a simple solution if no further information is needed from them.

3. Univariate Analysis

3.1. Numerical Features

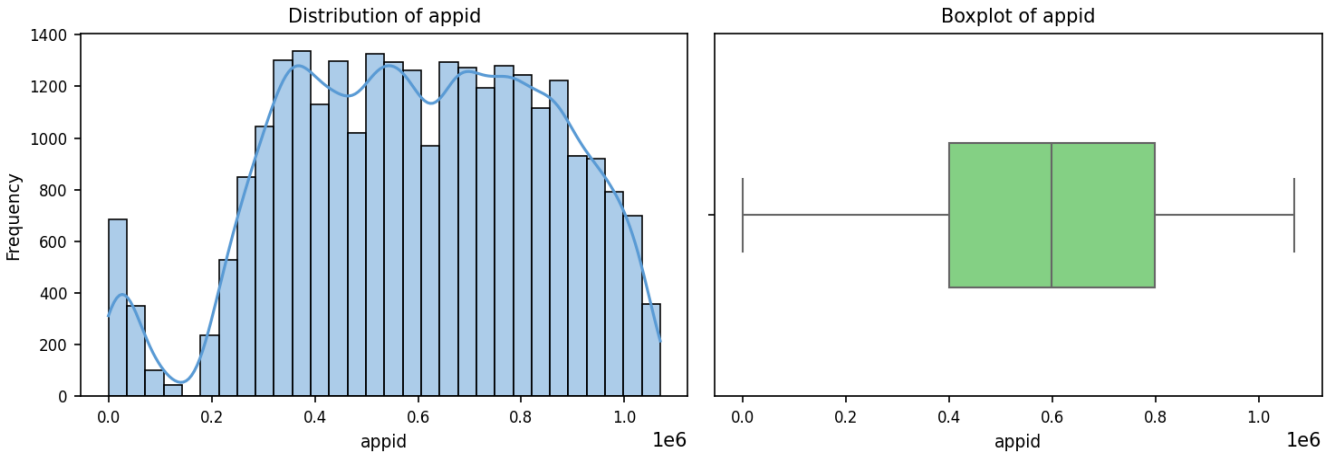

Figure 1: *Distribution (histogram and KDE) and boxplot for 'appid'. The histogram shows shape, central tendency, and spread. The boxplot highlights median, quartiles, and potential outliers.*

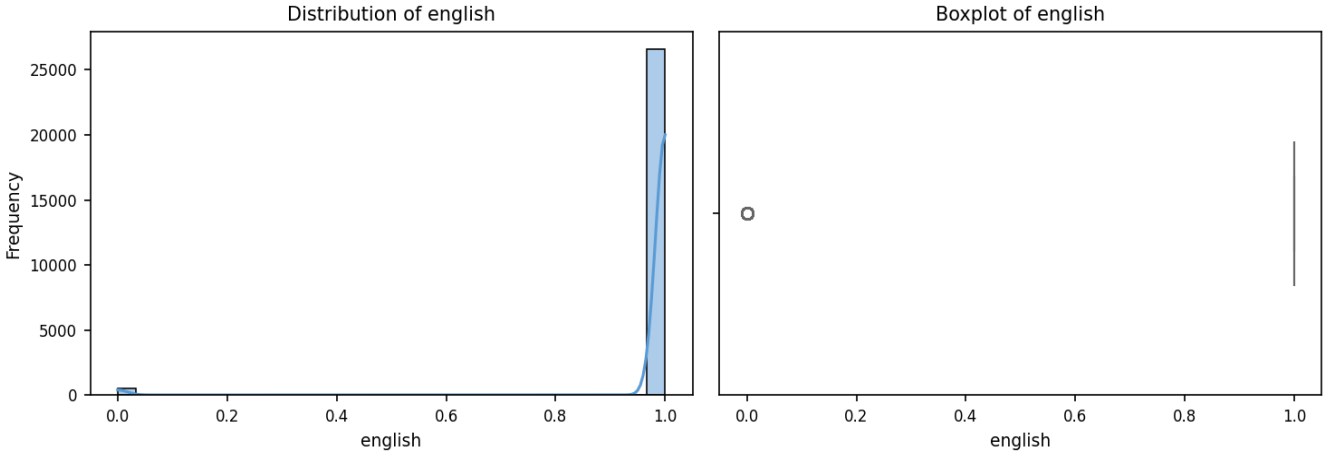

Figure 2: *Distribution (histogram and KDE) and boxplot for 'english'. The histogram shows shape, central tendency, and spread. The boxplot highlights median, quartiles, and potential outliers.*

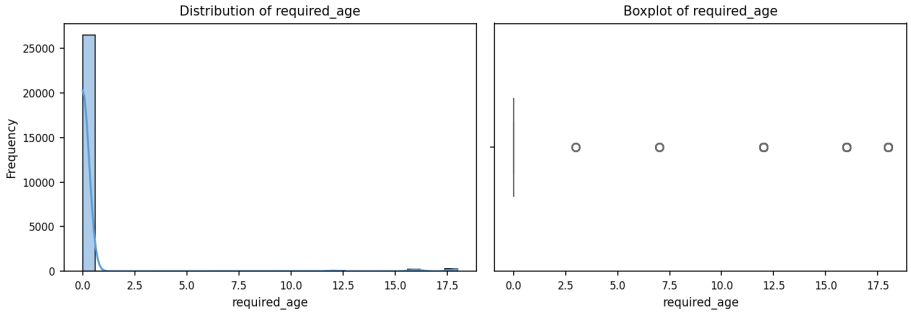

Figure 3: *Distribution (histogram and KDE) and boxplot for 'required_age'. The histogram shows shape, central tendency, and spread. The boxplot highlights median, quartiles, and potential outliers.*

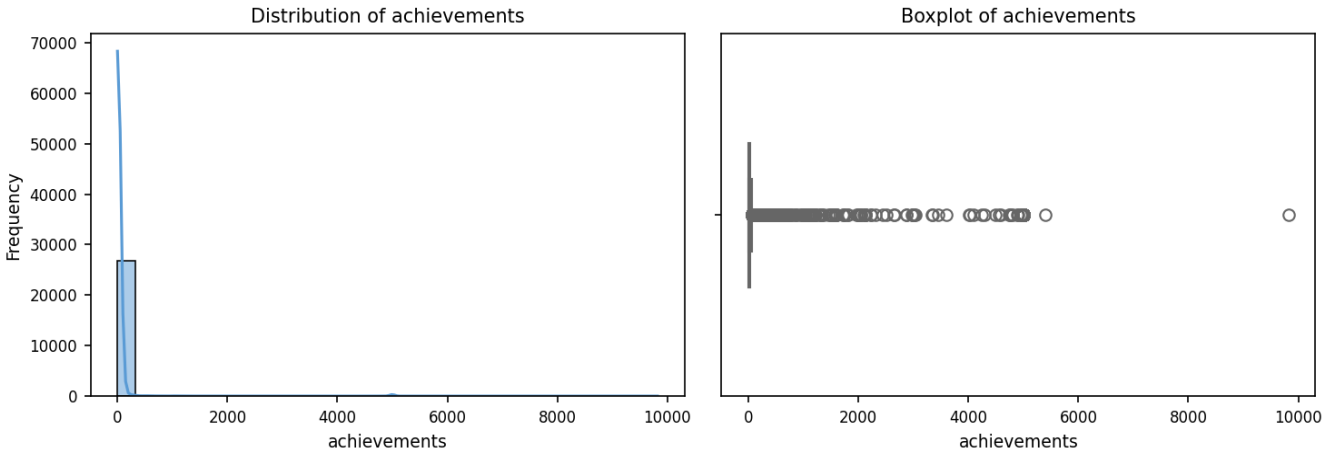

Figure 4: *Distribution (histogram and KDE) and boxplot for 'achievements'. The histogram shows shape, central tendency, and spread. The boxplot highlights median, quartiles, and potential outliers.*

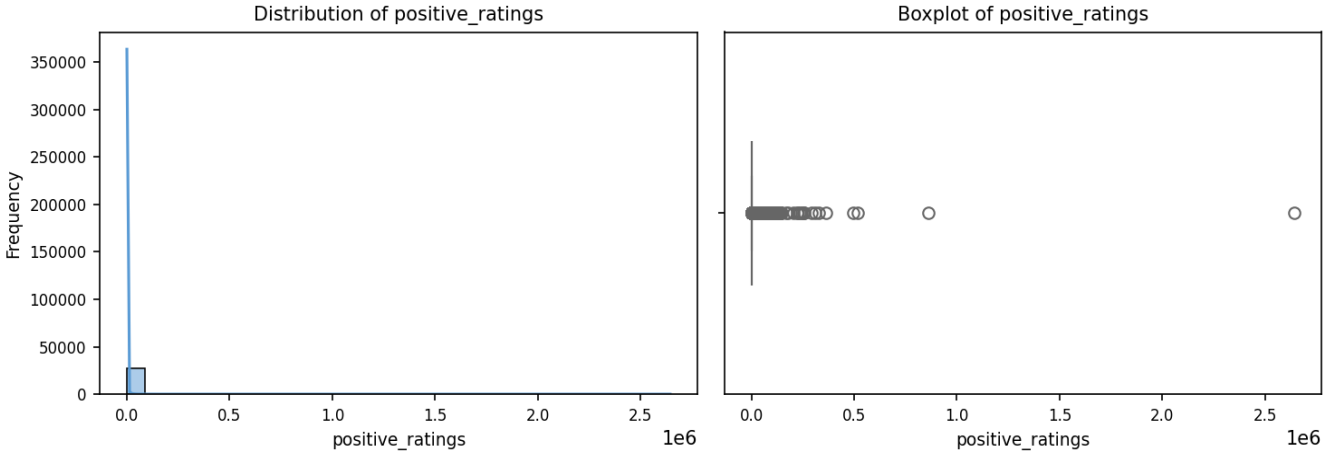

Figure 5: *Distribution (histogram and KDE) and boxplot for 'positive_ratings'. The histogram shows shape, central tendency, and spread. The boxplot highlights median, quartiles, and potential outliers.*

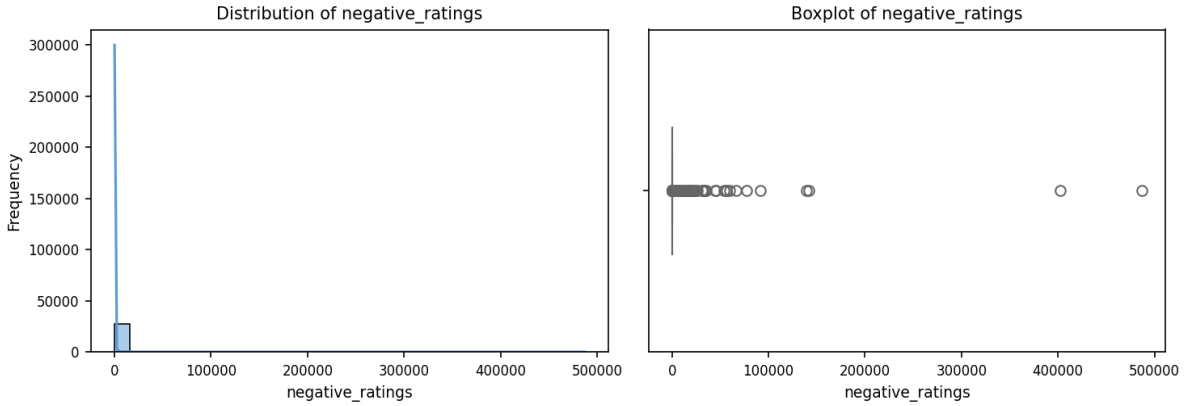

Figure 6: *Distribution (histogram and KDE) and boxplot for 'negative_ratings'. The histogram shows shape, central tendency, and spread. The boxplot highlights median, quartiles, and potential outliers.*

Observations on Numerical Feature Distributions:

The analysis of the numerical features reveals highly skewed distributions for several variables, indicating a substantial departure from normality. Specifically, 'english', 'required_age', 'achievements', 'positive_ratings', and 'negative_ratings' exhibit extreme right skewness, as evidenced by their significantly higher means than medians and exceptionally large skewness and kurtosis values. This suggests the presence of a long tail of high values, with a concentration of data points at the lower end of the range. In contrast, 'appid' shows a relatively symmetrical distribution, although the boxplot hints at the presence of outliers. The high standard deviations across most features, particularly for 'achievements', 'positive_ratings', and 'negative_ratings', highlight significant variability within these datasets. The presence of potential outliers is a recurring theme, consistently flagged by the boxplots and supported by the large discrepancies between means and medians. The extreme values for 'achievements', 'positive_ratings', and 'negative_ratings' (with maximum values orders of magnitude larger than their means) are particularly noteworthy. This suggests that a small number of data points exert a disproportionate influence on the mean, potentially masking underlying patterns in the data. The high kurtosis values further emphasize the presence of heavy tails and extreme values, indicating distributions that are far from normal. These outliers should be carefully investigated to determine their nature (e.g., errors, genuine extreme cases) and their impact on subsequent analyses. The striking characteristic across multiple features is the prevalence of right skewness and the presence of outliers, indicating the need for robust statistical methods that are less sensitive to extreme values. Transformations, such as logarithmic transformations, might be considered to address the skewness and improve the normality of these distributions before applying certain statistical models. Careful attention should be paid to outlier detection and handling to avoid misleading conclusions from analyses that might be unduly influenced by these extreme data points.

3.2. Categorical Features

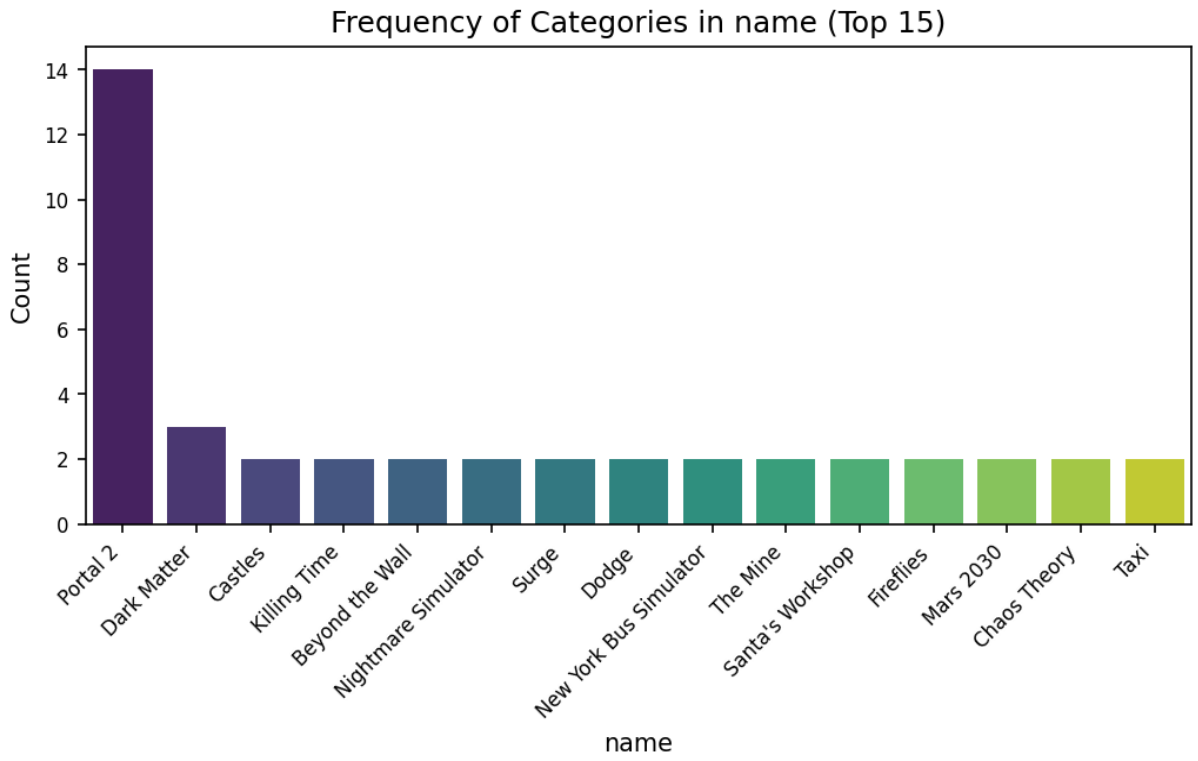

Figure 7: *Bar chart showing frequency of top categories in 'name'. Total unique values: 27033.*

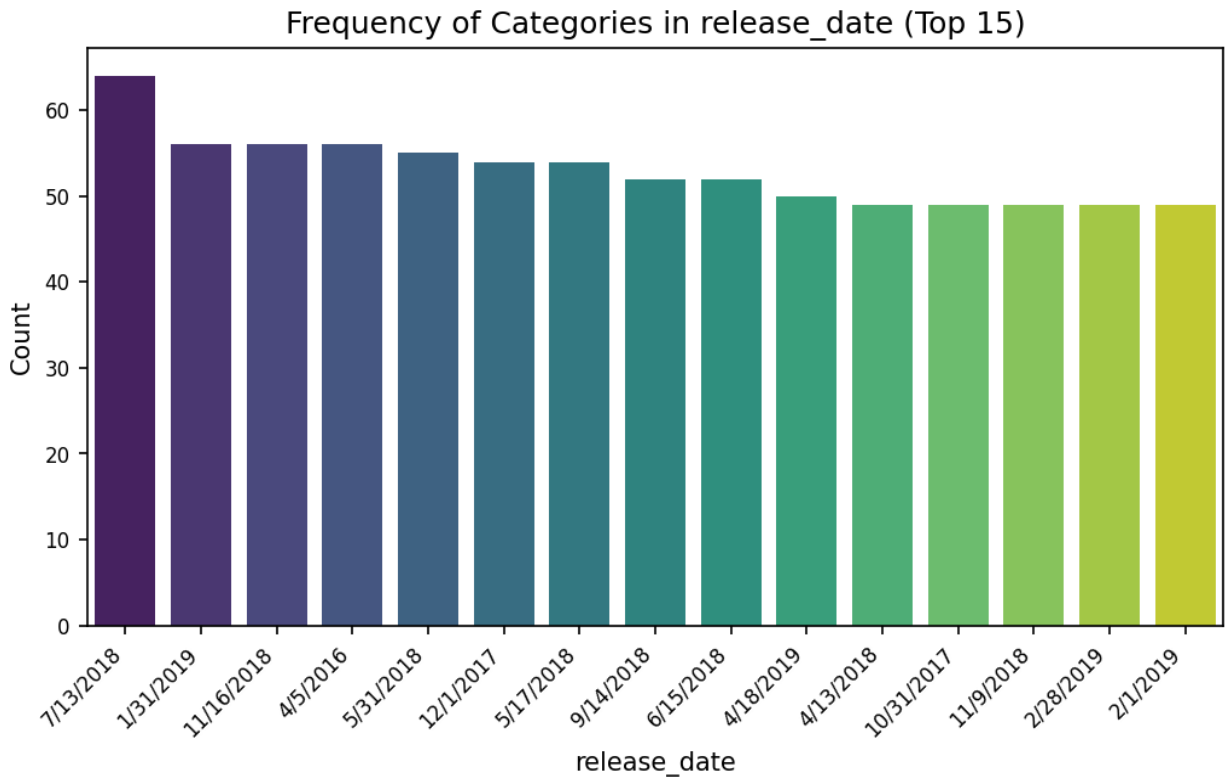

Figure 8: *Bar chart showing frequency of top categories in 'release_date'. Total unique values: 2619.*

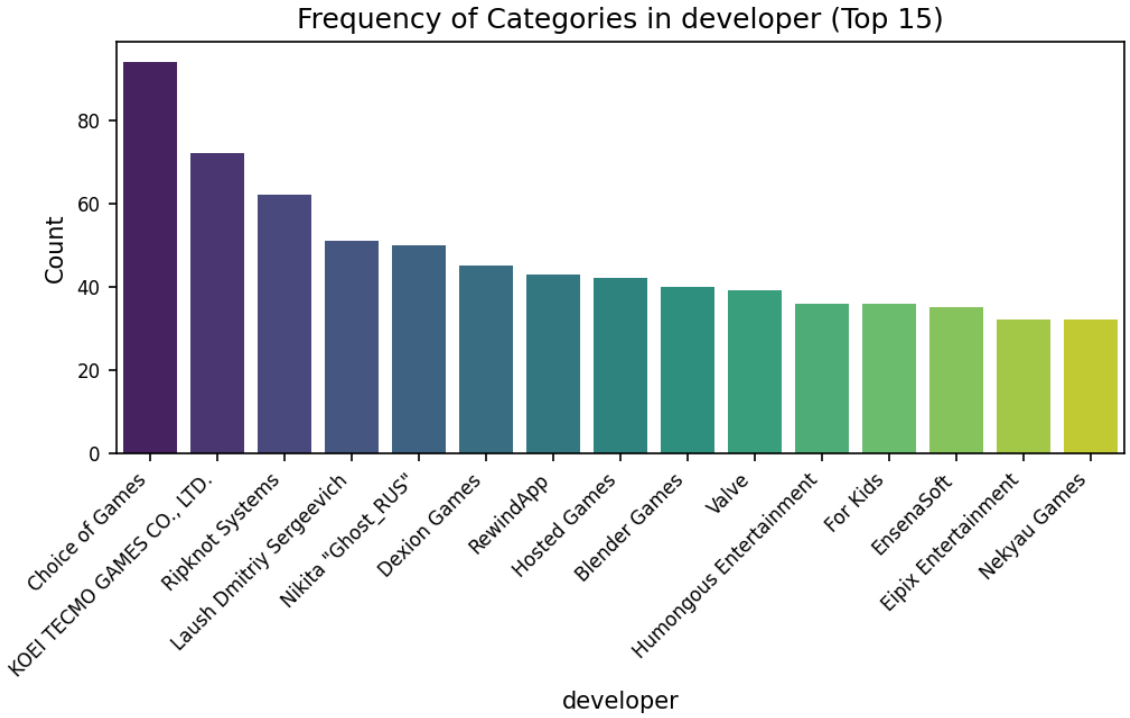

Figure 9: *Bar chart showing frequency of top categories in 'developer'. Total unique values: 17112.*

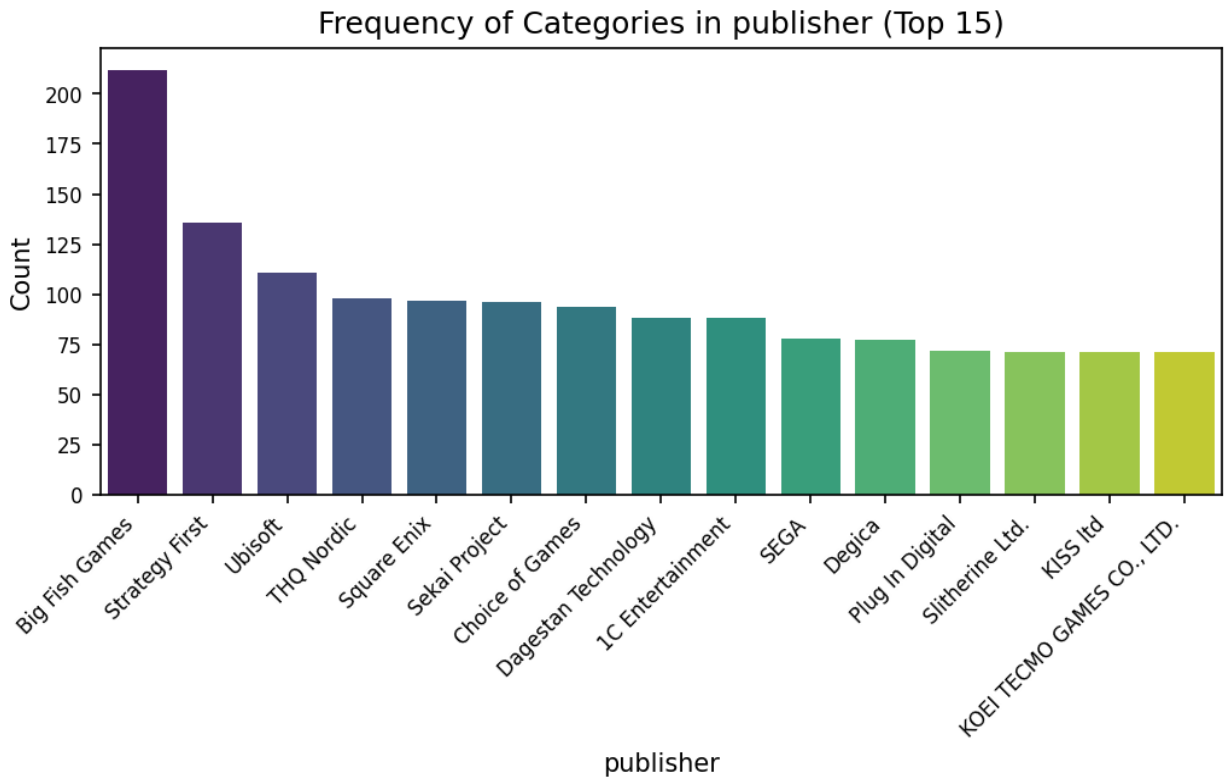

Figure 10: *Bar chart showing frequency of top categories in 'publisher'. Total unique values: 14353.*

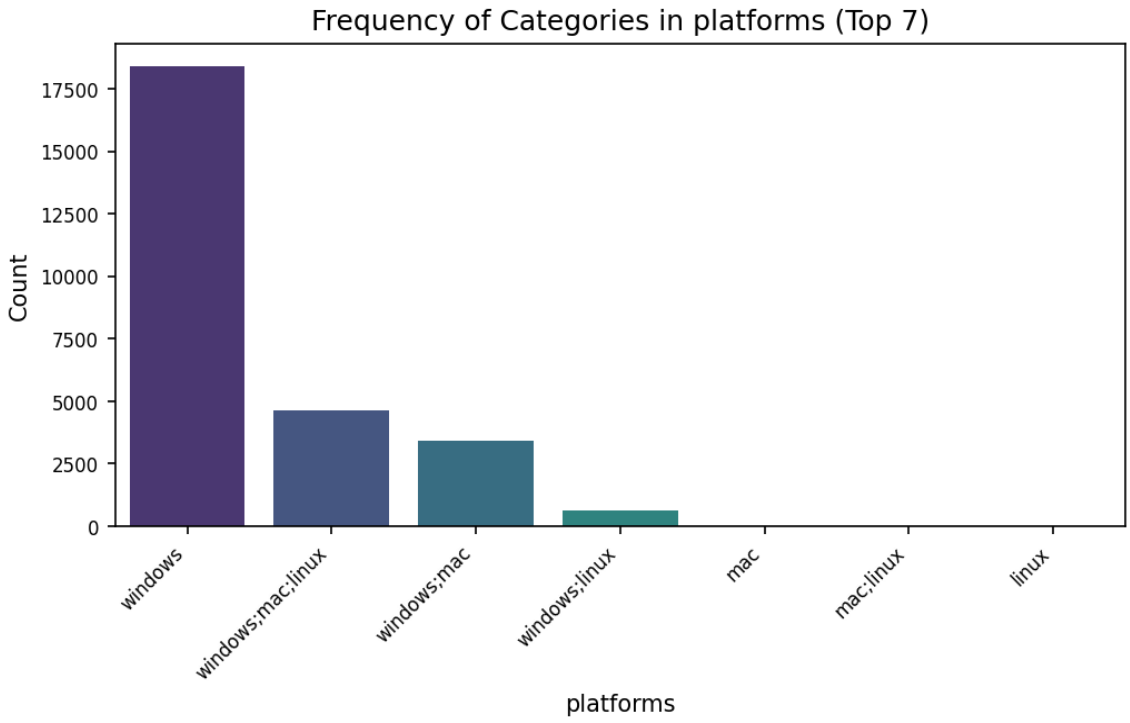

Figure 11: *Bar chart showing frequency of top categories in 'platforms'. Total unique values: 7.*

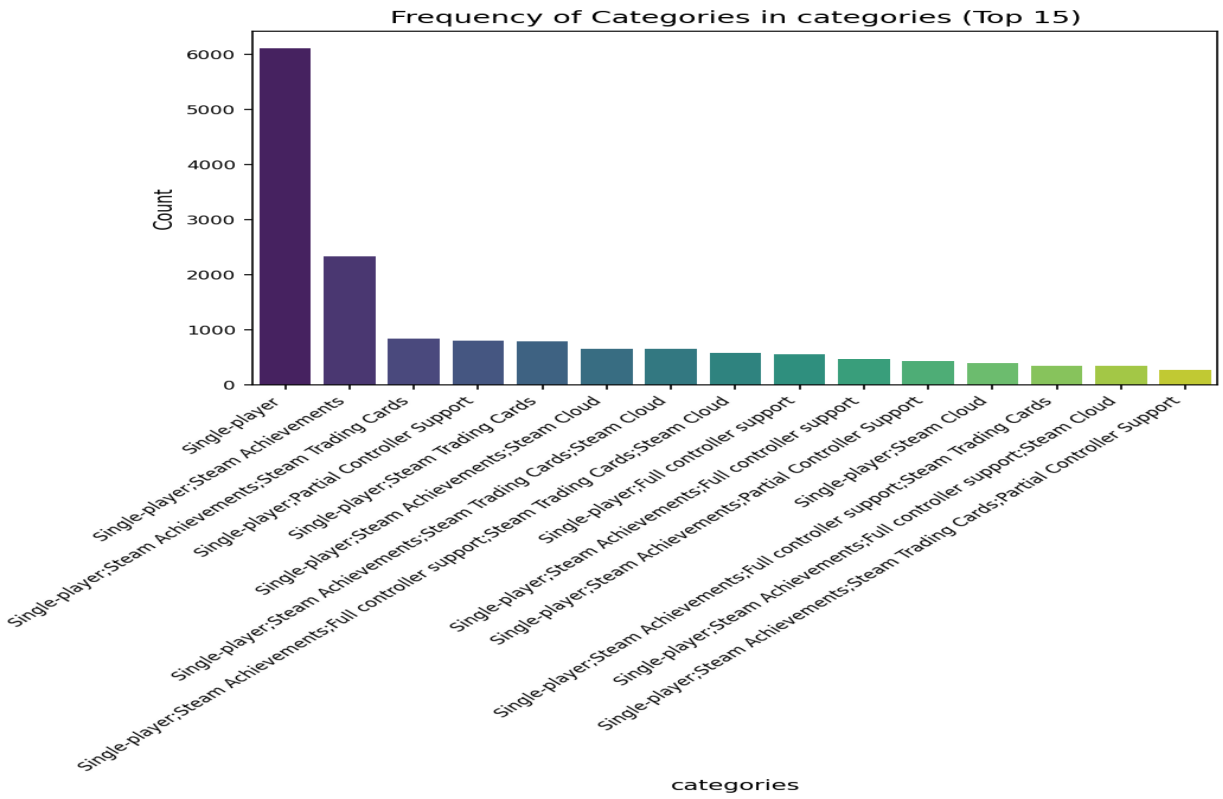

Figure 12: *Bar chart showing frequency of top categories in 'categories'. Total unique values: 3333.*

Observations on Categorical Feature Distributions:

The analysis of categorical features reveals a diverse dataset with significant variations in cardinality and distribution. Features like 'name', 'developer', and 'publisher' exhibit extremely high cardinality (27033, 17112, and 14353 unique values, respectively), indicating a large number of distinct game titles, developers, and publishers. This high cardinality poses challenges for model training, potentially leading to overfitting or the curse of dimensionality. Effective dimensionality reduction techniques, such as feature hashing or embedding methods, will likely be necessary for these features. In contrast, 'platforms' has a very low cardinality (7 unique values), suggesting a relatively small number of gaming platforms represented in the dataset. While most features are characterized by a highly skewed distribution with a dominant top category (e.g., 'windows' for platforms, 'Single-player' for categories), the degree of skewness varies. 'publisher' shows a relatively strong concentration with 'Big Fish Games' representing 0.8% of the data. Conversely, the top categories for 'name', 'release_date', and 'developer' represent a much smaller fraction (0.1%, 0.2%, and 0.3%, respectively), suggesting a more even spread across the remaining categories for these features. This uneven distribution needs to be considered when choosing appropriate encoding schemes; for example, one-hot encoding might be inefficient for high-cardinality features with skewed distributions. The insights gleaned from this analysis highlight the need for careful feature engineering. High-cardinality features require dimensionality reduction to prevent overfitting and improve model performance. The skewed distributions of many features suggest the need for encoding methods that handle class imbalance effectively, such as target encoding or frequency encoding, potentially in conjunction with techniques like binning or clustering to reduce the number of unique values. The relatively low cardinality of 'platforms' and 'categories' suggests that simpler encoding methods like one-hot encoding may be suitable.

4. Bivariate Analysis

4.1. Numerical vs. Numerical Features

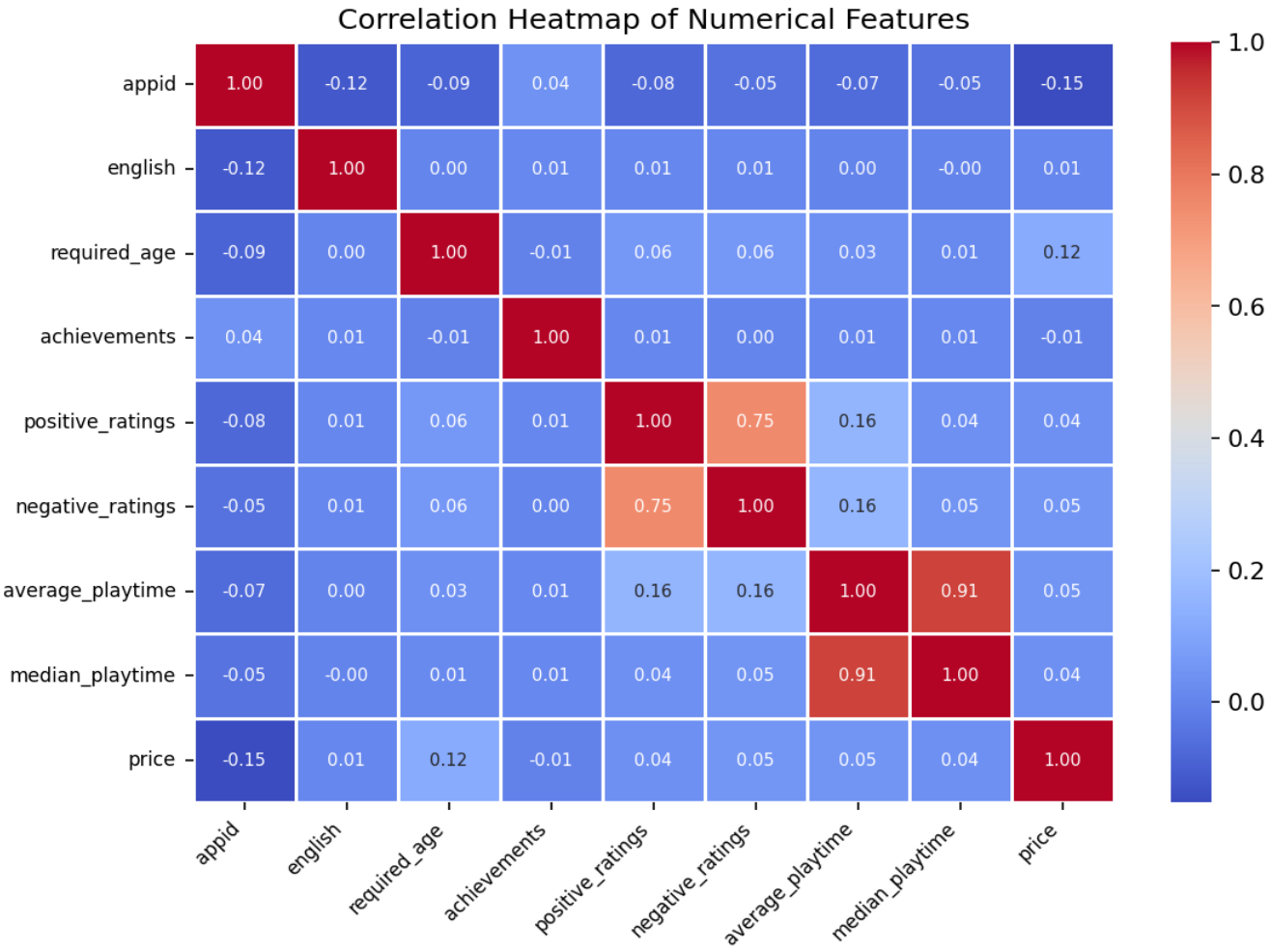

Figure 13: *Heatmap visualizing linear correlations (Pearson's r) between numerical features.*

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

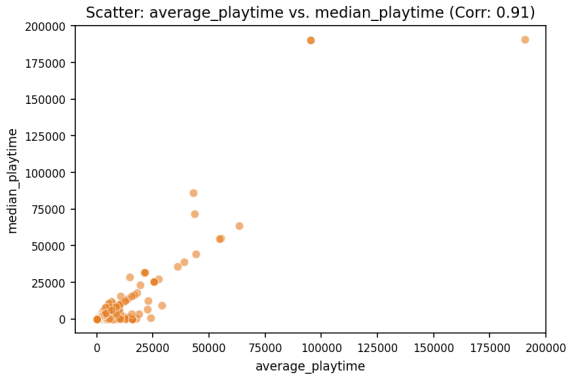

Figure 14: *Scatter plot for 'average_playtime' and 'median_playtime'. Correlation: 0.91.*

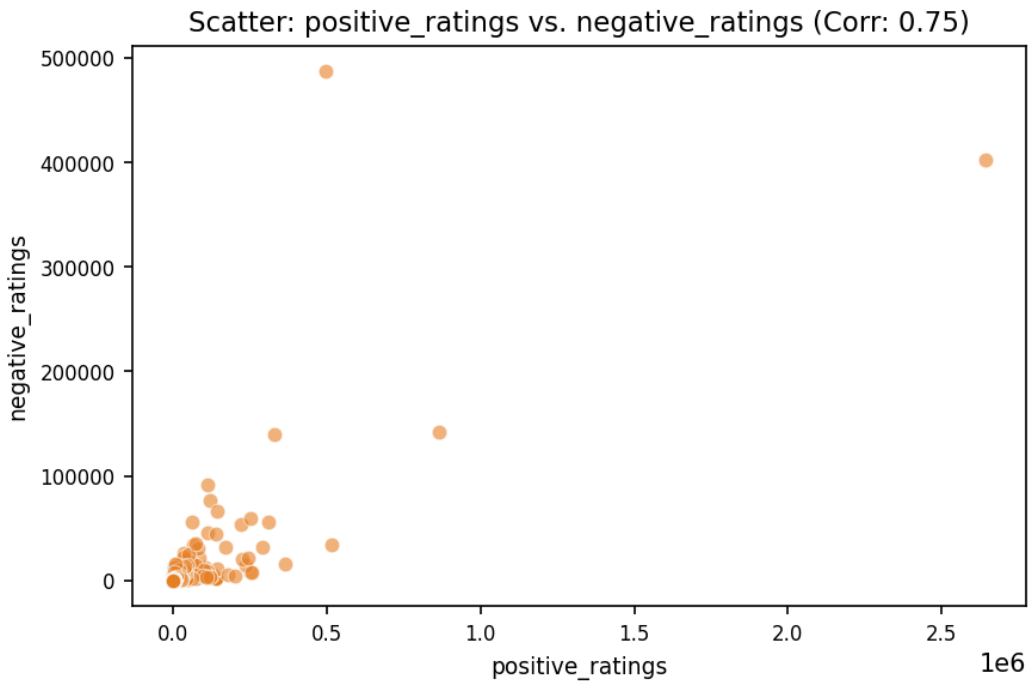

Figure 15: *Scatter plot for 'positive_ratings' and 'negative_ratings'. Correlation: 0.75.*

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 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 strongest positive correlations observed are between 'average_playtime' and 'median_playtime' (0.91) and 'positive_ratings' and 'negative_ratings' (0.75). The high correlation between average and median playtime strongly suggests that games with longer average playtime also tend to have longer median playtimes, which is expected as it reflects a consistent playtime distribution. The correlation between positive and negative ratings, while not as strong as the playtime correlation, still indicates a tendency for games with more positive ratings to also have a higher number of negative ratings. This might suggest that highly popular games attract more players and therefore receive more of both positive and negative feedback, rather than implying a direct causal relationship between positive and negative feedback. The scatter plots likely show a strong linear trend for the playtime variables and a more moderate positive trend for the ratings variables, confirming these correlations.

4.2. Numerical vs. Categorical Features

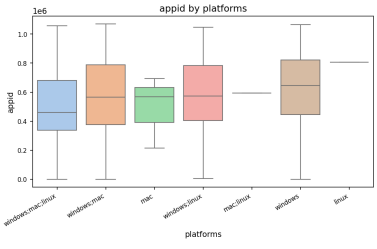

Figure 16: *Box plot of 'appid' across categories of 'platforms'.*

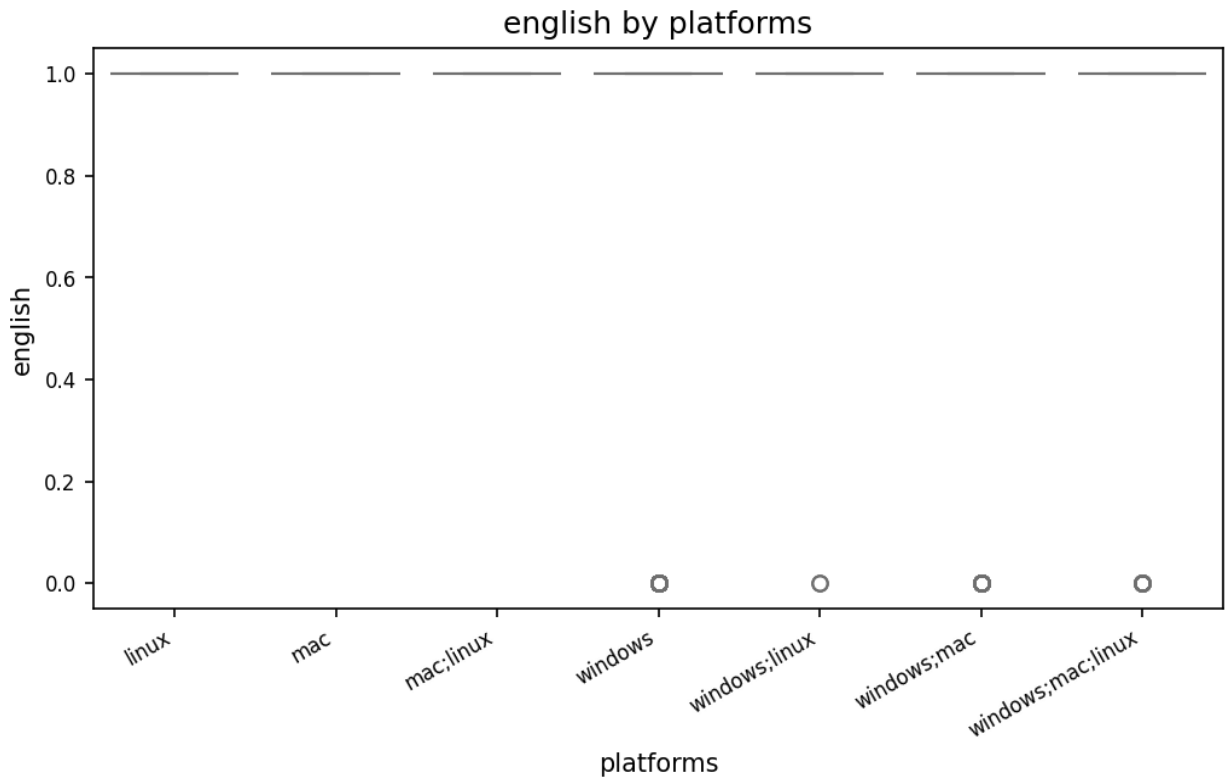

Figure 17: *Box plot of 'english' across categories of 'platforms'.*

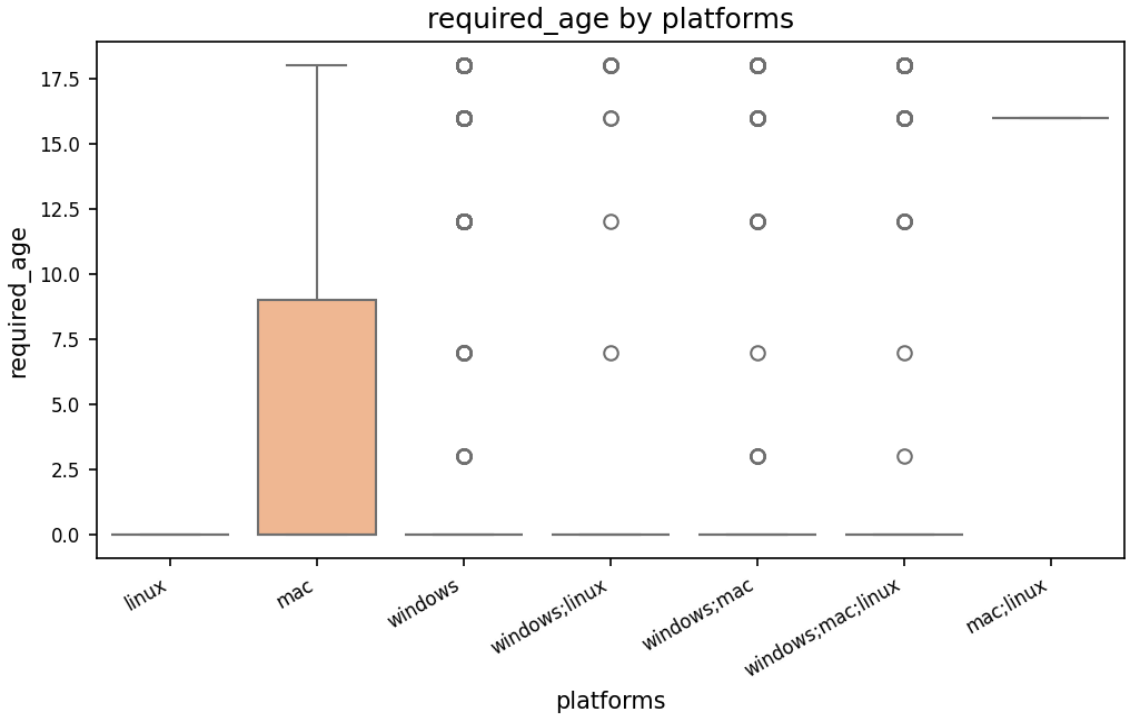

Figure 18: *Box plot of 'required_age' across categories of 'platforms'.*

Interpretation of Numerical vs. Categorical Interactions:

Box plots comparing numerical distributions across categories provide a powerful visual summary of data, revealing both central tendency and variability within each category. They show 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 (e.g., comparing the distribution of 'appid' across different 'platforms'), we can quickly assess whether the numerical variable behaves differently depending on the category. For instance, a longer box in one category indicates greater variability in the numerical data within that category compared to others with shorter boxes. The position of the median within the box and relative to the medians of other categories highlights differences in central tendency. Significant differences in medians across categories indicate that the average or typical value of the numerical variable is systematically different depending on the category. For example, if the median 'appid' value is significantly higher for the 'iOS' platform than for the 'Android' platform, it suggests that the 'iOS' platform tends to have apps with higher identifiers (which might correlate with app age or other factors). Significant differences in the spread (IQR or whisker lengths) suggest that the variability of the numerical variable is also affected by the category. A larger IQR in one category implies more heterogeneity or dispersion within that group. Combining observations about median differences and spread differences paints a more complete picture of how the numerical variable is distributed across the different categories, enabling more informed conclusions about the relationships between the categorical and numerical variables.

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. Specifically, no missing values were detected, indicating a high level of data completeness. However, the presence of 13 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 variation to the dataset. Univariate analysis covered all 18 features, examining the distribution of both numerical and categorical variables. While specific details on these distributions are absent from the provided log, the analysis itself suggests a foundation for further exploration into the characteristics and potential outliers within each feature. This will be crucial for understanding the underlying patterns and variability within the dataset. Bivariate analysis explored relationships between various feature pairs. The log indicates that observations were gathered, but lacks specifics on the nature or strength of identified correlations or associations. Further details regarding the specific feature pairs analyzed and the results are needed to fully interpret the findings of this section. The limited information suggests a need for more detailed reporting on the bivariate analysis to understand the interdependencies between variables. The overall findings suggest a dataset suitable for further analysis, given the minimal data quality concerns. However, the limited detail on the univariate and bivariate analyses necessitates a more comprehensive report to fully understand the underlying structure and relationships within the `temp_steam.csv` dataset. The absence of unexpected findings in the initial quality assessment is noteworthy but requires further investigation given the limited insight into the feature distributions and interrelationships.

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 found), and initial observations from univariate and bivariate analyses. This foundational understanding serves as a crucial first step in guiding further, more in-depth investigations. Given the report's findings of 13 duplicate rows, the next steps should prioritize addressing data quality issues and exploring potential relationships further: 1. **Duplicate Row Investigation:** Investigate the 13 duplicate rows to determine the cause of duplication (e.g., data entry errors, data merging issues). Decide on an appropriate strategy for handling duplicates (removal or consolidation) based on the nature of the duplicates. 2. **Bivariate Analysis Deep Dive:** The report mentions "various feature pairs" were analyzed with only two general observations noted. Expand on the bivariate analysis by generating correlation matrices for numerical features and exploring the relationships between numerical and categorical features (e.g., using box plots to visualize the distribution of numerical features across different categories). This will help identify any strong correlations or significant differences that warrant further investigation. 3. **Visualization and Exploration of Numerical Features:** The report indicates nine numerical features were analyzed. Create histograms and box plots for each numerical feature to identify the distribution (e.g., presence of skewness, outliers) and potential data transformations needed for subsequent modeling. 4. **Exploratory Data Analysis (EDA) of Categorical Features:** Similarly, create frequency tables and bar charts for the nine categorical features to understand their distributions and identify any categories with low frequencies that might need further attention. This visualization may also reveal insights for potential feature engineering.