Automated Data Analysis Report (via Gemini): Temp Steam

Executive Summary

This report summarizes the initial automated exploratory data analysis (EDA) of the `temp_steam.csv` dataset, containing 27,085 rows and 18 columns (9 numerical, 9 categorical). Preliminary analysis revealed a relatively clean dataset with only 10 duplicate entries and no missing values or constant columns. Univariate and bivariate analyses, including descriptive statistics and visual inspections (details in the full report), were conducted to identify initial patterns and potential relationships between features. Two key observations emerged from the bivariate analysis (specific details are provided in the subsequent sections). The dataset's size and apparent data quality suggest suitability for further, more in-depth analysis. However, the lack of immediately obvious, strong patterns necessitates a more comprehensive investigation to uncover potentially hidden insights. This initial EDA provides a solid foundation for subsequent modeling and analysis. The findings highlight the need for more focused investigation into specific feature interactions suggested by the bivariate analysis. Further exploration, including advanced statistical methods and potentially feature engineering, is recommended to extract maximum value from this dataset.

1. Data Overview

This report provides an initial automated analysis of the dataset from 'temp_steam.csv'.

1.1. Basic Information

Table 1: Dataset Dimensions

Metric	Value
Number of Rows	27085
Number of Columns	18
Total Data Points	487530

1.2. Data Types

Table 2: Summary of Feature Data Types

Data Type	Count
object	9
int64	8
float64	1

Data Types Distribution Interpretation:

The dataset is balanced between numerical and categorical features, suggesting a need for diverse analytical techniques. The absence of datetime features limits the ability to perform time-series analysis or investigate temporal trends.

2. Data Quality Assessment

2.1. Missing Values

No missing values were found in the dataset. This is excellent for data completeness.

2.2. Duplicate Records

The dataset contains 10 duplicate rows (representing 0.04% of the data). These may need to be investigated or removed depending on the analysis context, as they can skew results.

2.3. Feature Variance

No constant columns (columns with only one unique value) were identified.

The following columns are quasi-constant (one value is highly dominant), potentially offering limited information: english (dominant value: 1 at 98.1%); requiredage (dominant value: 0 at 97.8%). Their utility should be reviewed.

Data Quality Summary & Implications:

The data quality assessment reveals a dataset with generally high quality. The absence of missing values is a significant positive, indicating a robust data collection process. The extremely low rate of duplicate rows (0.04%) is also negligible and unlikely to significantly impact subsequent analysis. The identification of quasi-constant columns, 'english' and 'requiredage', however, warrants attention. While not problematic in themselves, these columns may indicate a potential imbalance in the data or limited variability in these specific features, which could affect the performance of certain analytical techniques, particularly in predictive modeling. The presence of quasi-constant columns ('english' and 'requiredage') could be a concern depending on the analysis goals. For example, if these variables were intended to be predictive features in a model, their high dominance of a single value might limit the model's ability to learn meaningful patterns. This could lead to overfitting or a model that doesn't generalize well to new data. Similarly, insights derived from analyses focusing on these variables might be skewed by the inherent imbalance. The negligible number of duplicates is unlikely to have a substantial impact, and their removal is straightforward. To address the quasi-constant column issue, a thorough investigation into the data collection and definition of these variables is needed. Understanding why these variables exhibit such limited variation is crucial. Options include: removing these columns if they are deemed irrelevant to the analysis, exploring data transformations (if appropriate) to better represent the limited variation, or focusing analytical efforts on other, more informative variables. The duplicate rows can be easily removed, ensuring a clean dataset for further analysis. A careful review of the data dictionary and data generation process could prevent similar issues in future data collections.

3. Univariate Analysis

3.1. Numerical Features

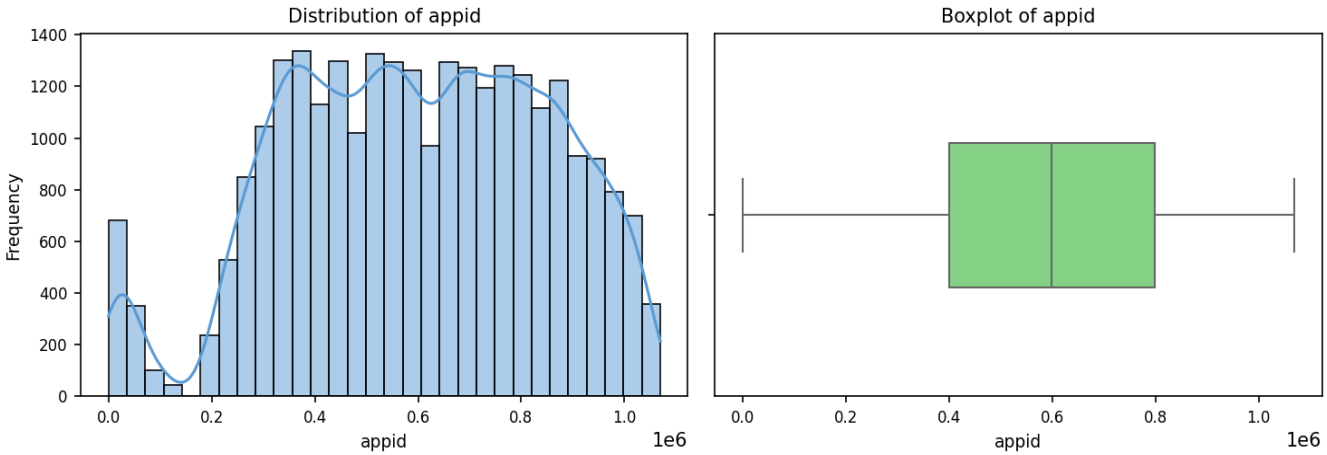

Figure 1: *Distribution (histogram and KDE) and boxplot for 'appid'. The histogram shows shape, central tendency, and spread. The boxplot highlights median, quartiles, and potential outliers.*

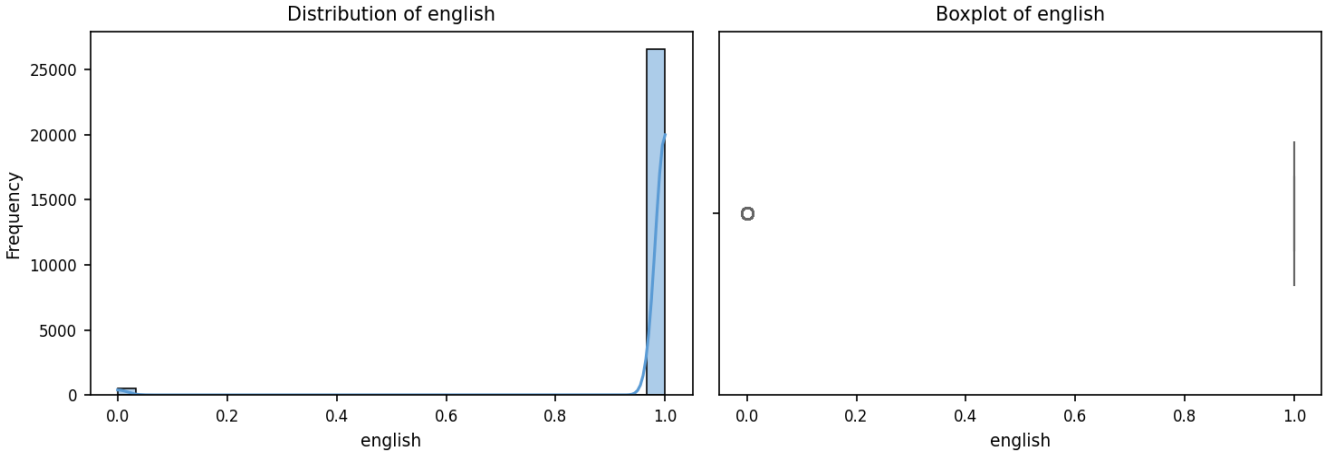

Figure 2: *Distribution (histogram and KDE) and boxplot for 'english'. The histogram shows shape, central tendency, and spread. The boxplot highlights median, quartiles, and potential outliers.*

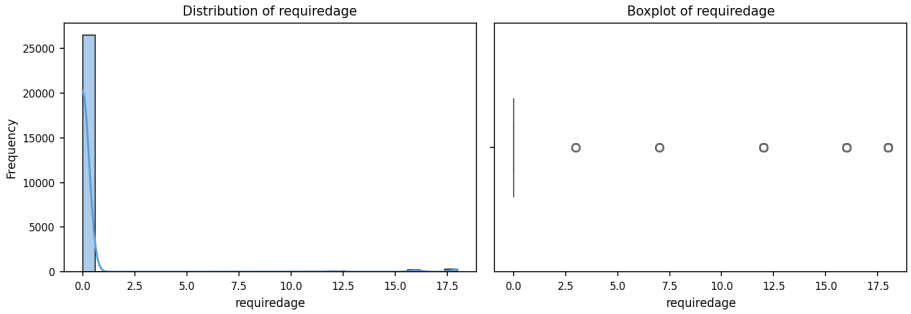

Figure 3: *Distribution (histogram and KDE) and boxplot for 'requiredage'. The histogram shows shape, central tendency, and spread. The boxplot highlights median, quartiles, and potential outliers.*

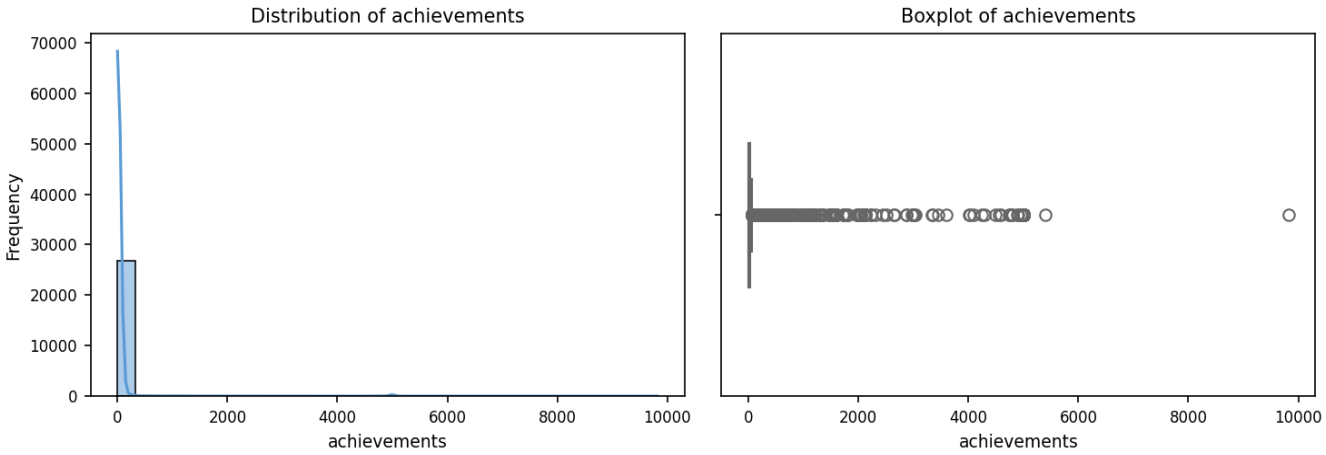

Figure 4: *Distribution (histogram and KDE) and boxplot for 'achievements'. The histogram shows shape, central tendency, and spread. The boxplot highlights median, quartiles, and potential outliers.*

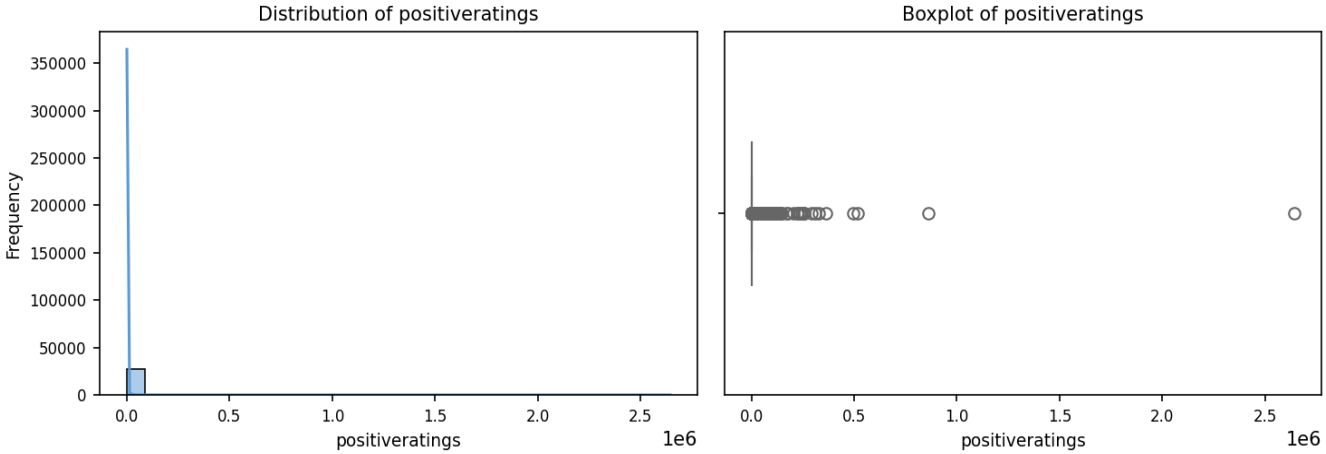

Figure 5: *Distribution (histogram and KDE) and boxplot for 'positiveratings'. The histogram shows shape, central tendency, and spread. The boxplot highlights median, quartiles, and potential outliers.*

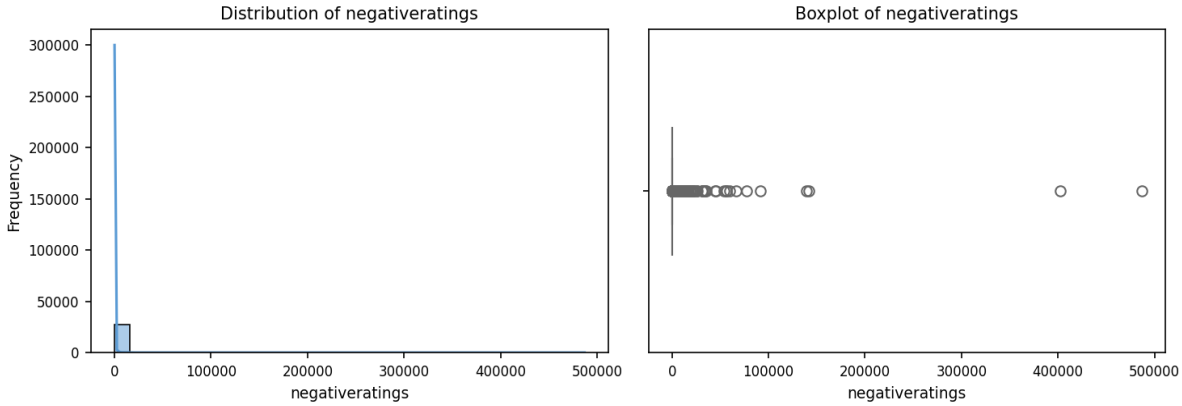

Figure 6: *Distribution (histogram and KDE) and boxplot for 'negativeratings'. The histogram shows shape, central tendency, and spread. The boxplot highlights median, quartiles, and potential outliers.*

Observations on Numerical Feature Distributions:

The analysis reveals highly skewed distributions for several numerical features in the dataset. Most notably, 'english', 'requiredage', 'achievements', 'positiveratings', and 'negativeratings' exhibit extreme right skewness, indicated by significantly larger means than medians and very high positive skewness values. This suggests a concentration of data points at the lower end of the range, with a long tail extending towards extremely high values. The kurtosis values for these features are also exceptionally high, confirming the presence of heavy tails and a sharp peak near the lower end. In contrast, 'appid' shows a relatively symmetric distribution with a slight negative skew. The presence of outliers is strongly suggested by the boxplots and the large discrepancies between mean and median for most features. The variability of the features is also striking. While 'appid' and 'english' have relatively low standard deviations compared to their means, features like 'achievements', 'positiveratings', and 'negativeratings' exhibit extremely high standard deviations, reflecting a wide spread of values and further emphasizing the influence of outliers. This high variability necessitates careful consideration during data analysis, as standard statistical methods might be highly sensitive to these extreme values. Transformations, such as logarithmic transformations, might be necessary to mitigate the impact of outliers and improve the normality of the distributions before applying certain statistical models. In summary, the data demonstrates a clear pattern of highly skewed distributions with numerous potential outliers for several features. This poses significant challenges for data analysis and model building, requiring careful attention to outlier handling and potentially data transformations to ensure reliable and robust results. The relatively normal distribution of 'appid' stands in stark contrast to the others, suggesting potentially different data generation mechanisms or characteristics for this feature compared to the rest.

3.2. Categorical Features

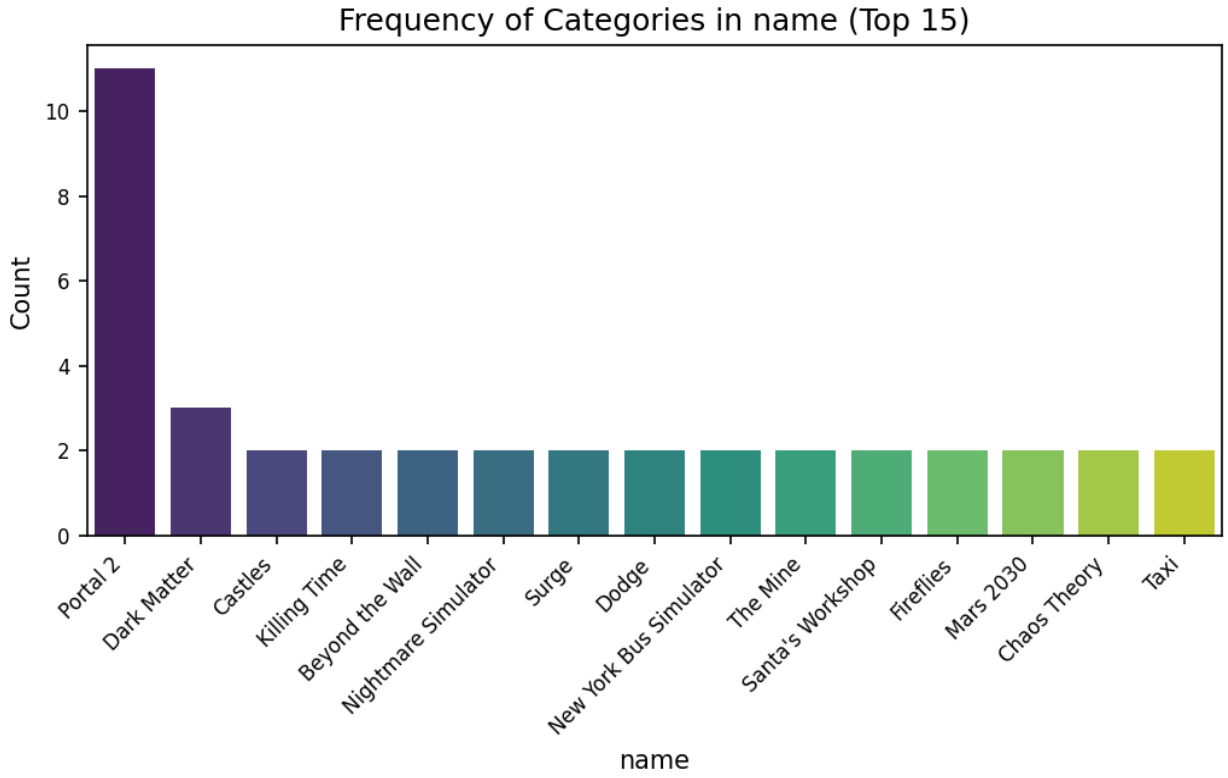

Figure 7: *Bar chart showing frequency of top categories in 'name'. Total unique values: 27033.*

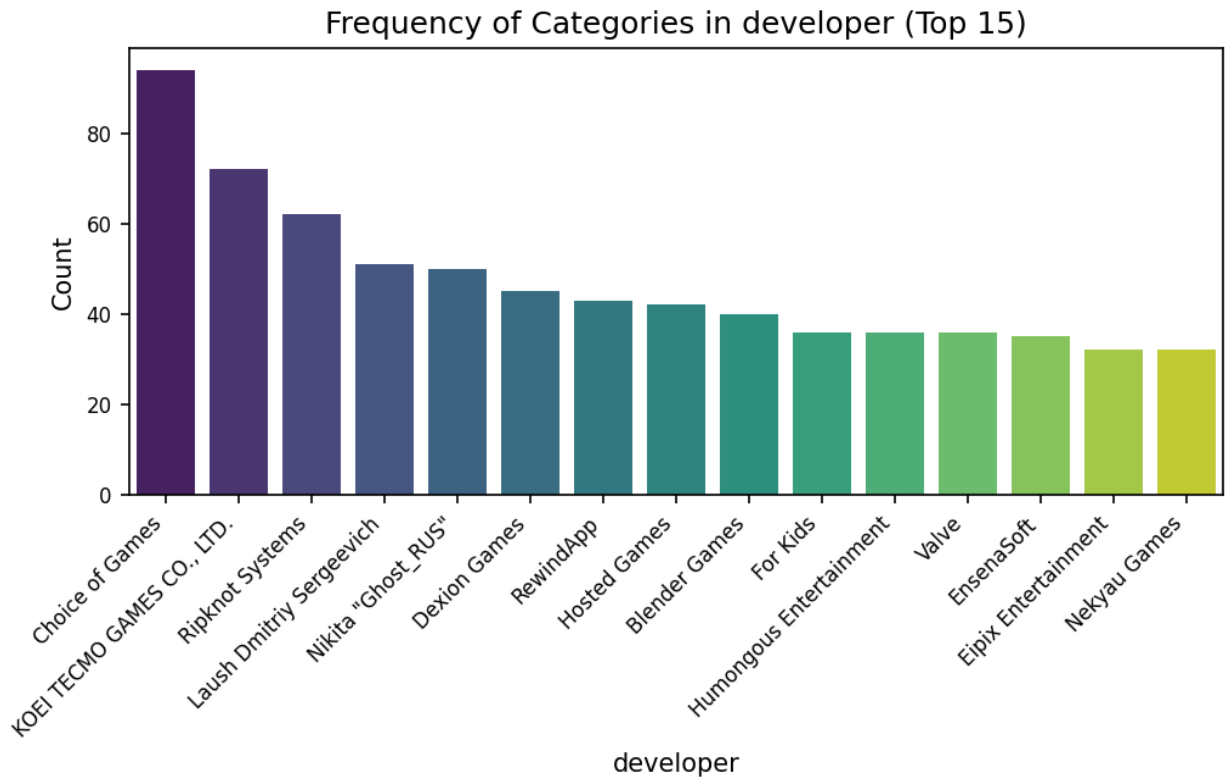

Figure 9: *Bar chart showing frequency of top categories in 'developer'. Total unique values: 17112.*

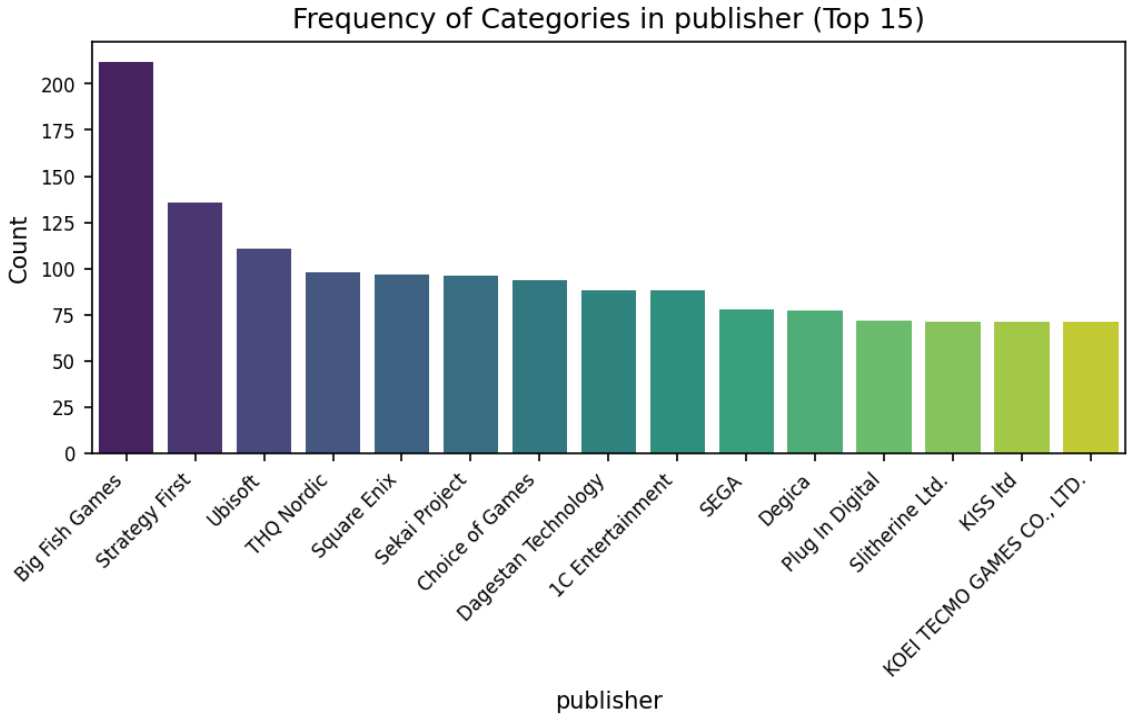

Figure 10: *Bar chart showing frequency of top categories in 'publisher'. Total unique values: 14353.*

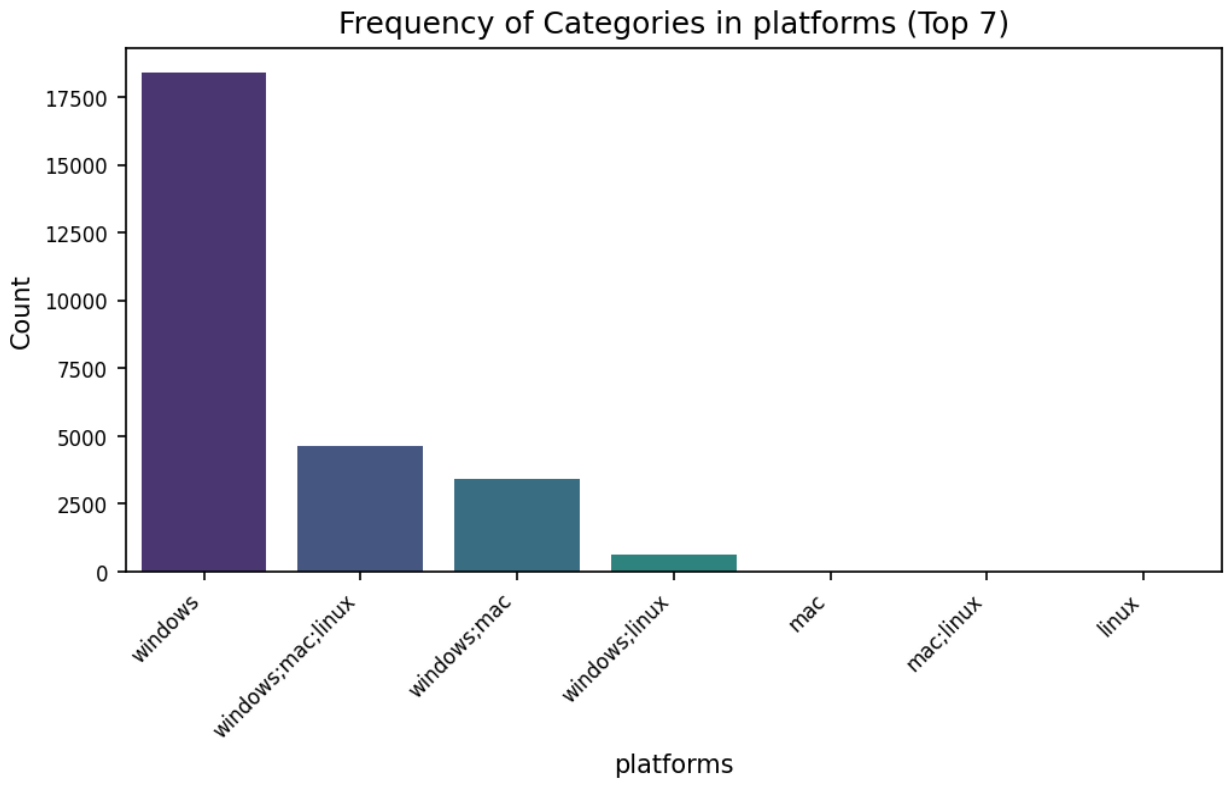

Figure 11: *Bar chart showing frequency of top categories in 'platforms'. Total unique values: 7.*

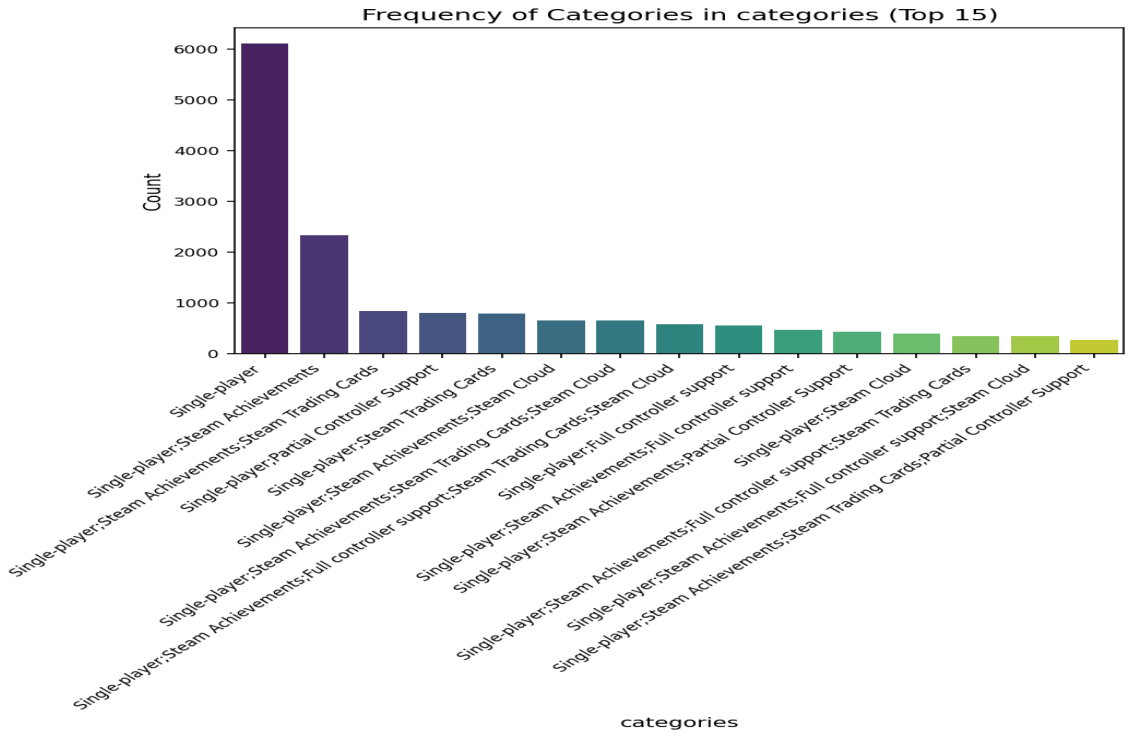

Figure 12: *Bar chart showing frequency of top categories in 'categories'. Total unique values: 3333.*

Observations on Categorical Feature Distributions:

The analysis of categorical features reveals a significant variation in cardinality. Features like 'name', 'developer', 'publisher', and 'categories' exhibit very high cardinality (thousands of unique values), indicating a large diversity in game titles, developers, publishers, and game categories within the dataset. In contrast, 'platforms' has a low cardinality (only 7 unique values), suggesting a relatively limited number of platforms represented. The 'releasedate' feature falls somewhere in between with a moderate number of unique values. The distribution of values within each feature also shows significant differences. While 'platforms' displays a heavily skewed distribution with 'windows' dominating at 67.9%, other features like 'name', 'developer', 'publisher', and 'categories' have top categories representing only a small percentage of the total (0.0% to 0.8%). This suggests that these features are relatively evenly distributed across their many unique values, with no single category overwhelmingly dominant. The 'categories' feature, however, shows a somewhat more concentrated distribution, with 'Single-player' representing a substantial 22.6%. The high cardinality of several features poses significant challenges for model training. One-hot encoding would create a massive number of features, potentially leading to the curse of dimensionality. For features like 'name', 'developer', and 'publisher', dimensionality reduction techniques such as target encoding or embedding methods might be necessary. For the lower cardinality features, like 'platforms' and potentially 'categories', one-hot encoding might be feasible. The highly skewed distribution of 'platforms' should also be considered during modeling.

4. Bivariate Analysis

4.1. Numerical vs. Numerical Features

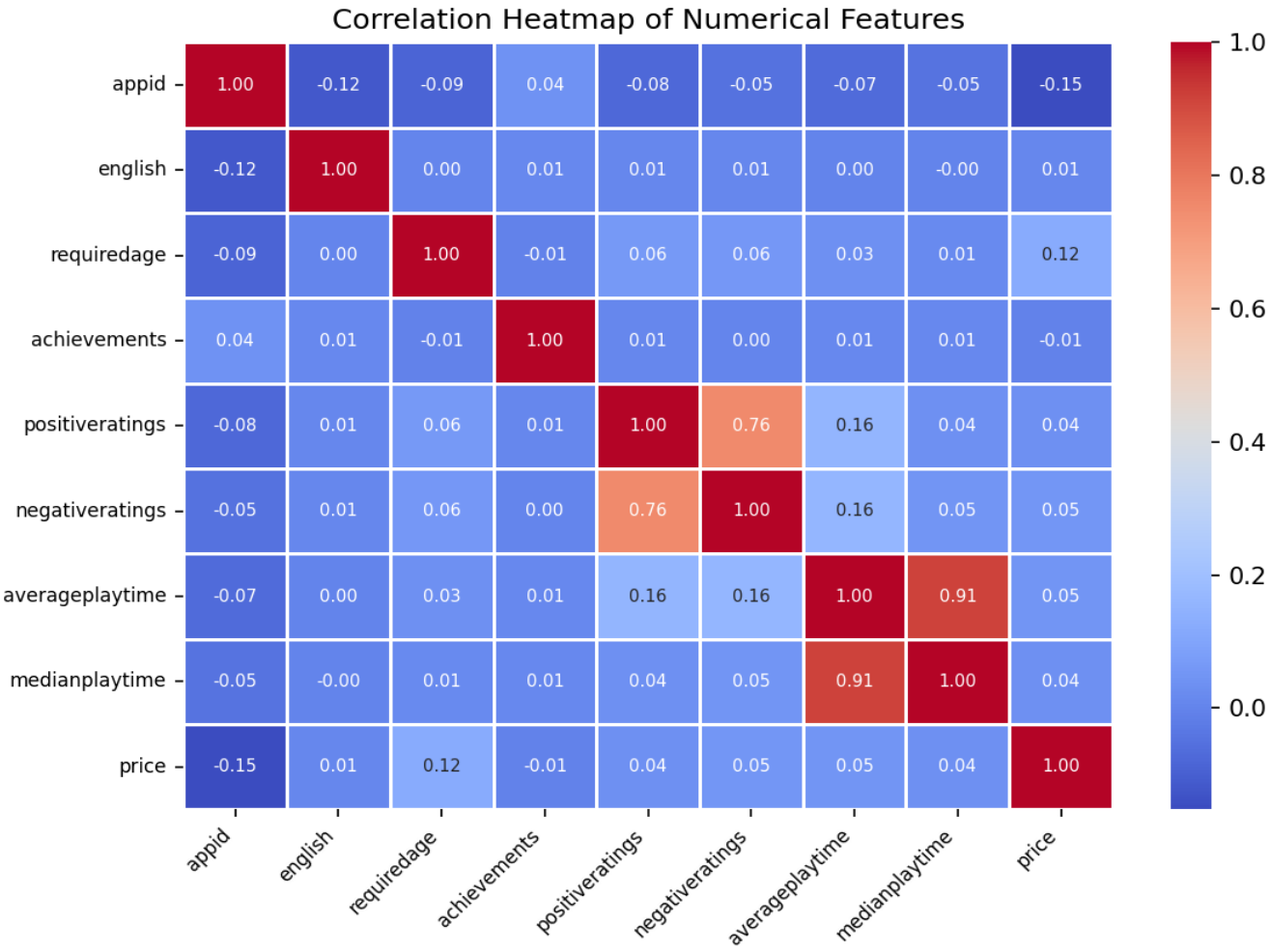

Figure 13: *Heatmap visualizing linear correlations (Pearson's r) between numerical features.*

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

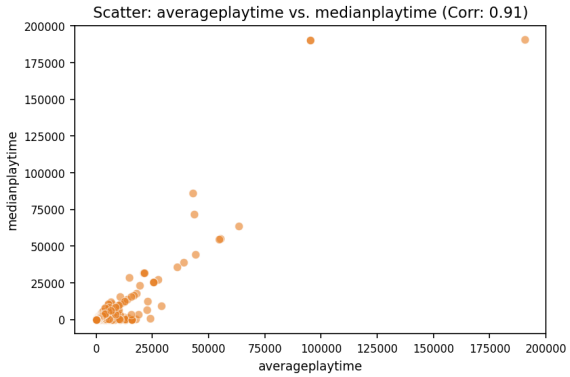

Figure 14: *Scatter plot for 'averageplaytime' and 'medianplaytime'. Correlation: 0.91.*

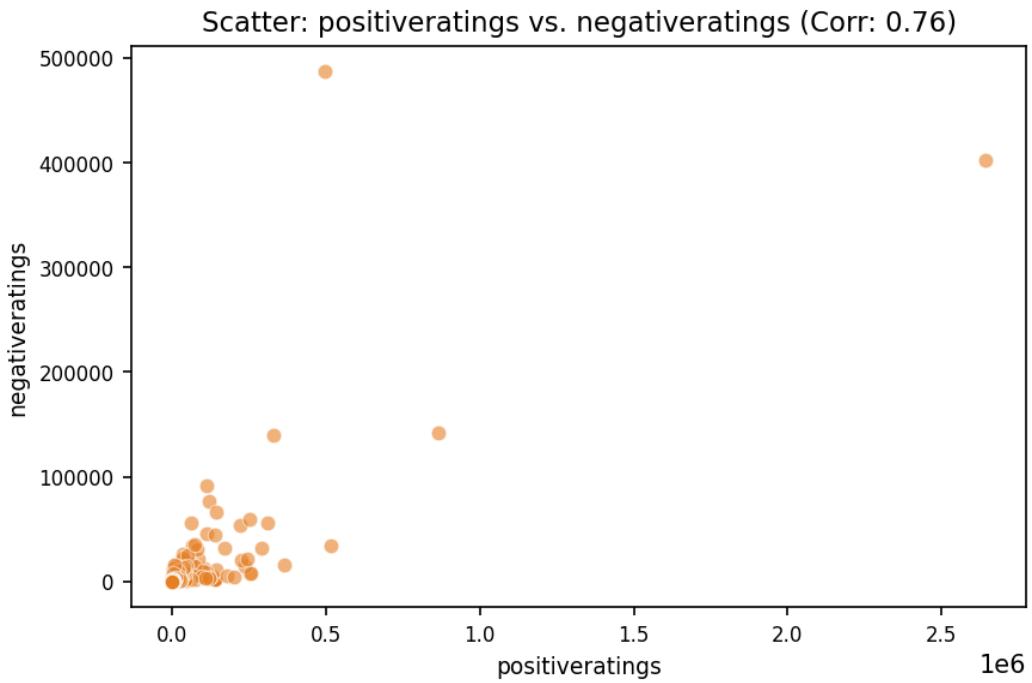

Figure 15: *Scatter plot for 'positiveratings' and 'negativeratings'. Correlation: 0.76.*

Interpretation of Numerical Correlations:

A correlation matrix displays the pairwise correlations between multiple variables. Each cell in the matrix shows the correlation coefficient (ranging from -1 to +1) between two variables; a value near +1 indicates a strong positive correlation (variables tend to increase together), a value near -1 indicates a strong negative correlation (one variable increases as the other decreases), and a value near 0 indicates a weak or no linear relationship. The provided analysis shows strong positive correlations between `averageplaytime` and `medianplaytime` (0.91) and between `positiveratings` and `negativeratings` (0.76). The strong positive correlation between average and median playtime suggests that games with longer average playtimes also tend to have longer median playtimes, indicating a consistent pattern in playtime duration. Similarly, the positive correlation between positive and negative ratings implies that games receiving many positive ratings also tend to receive a considerable number of negative ratings. This could suggest that highly popular games attract a larger overall audience, leading to more diverse feedback, including both positive and negative reviews, rather than indicating inherently poor quality. The scatter plots likely visually confirm these relationships, showing a strong clustering of points along a diagonal line for the playtime variables and a more dispersed but still positively sloped pattern for the ratings variables.

4.2. Numerical vs. Categorical Features

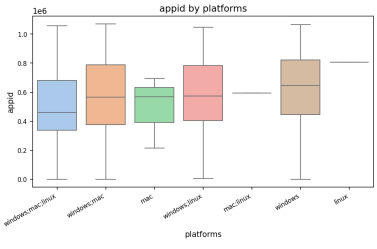

Figure 16: *Box plot of 'appid' across categories of 'platforms'.*

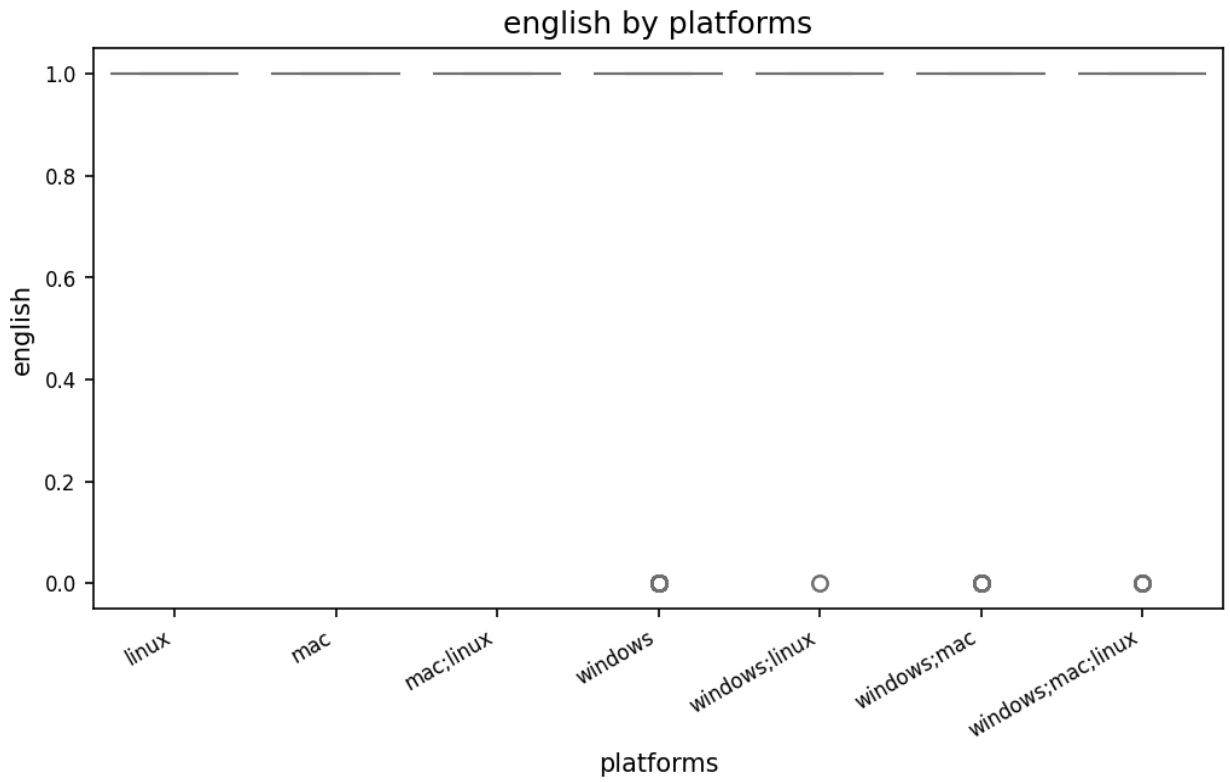

Figure 17: *Box plot of 'english' across categories of 'platforms'.*

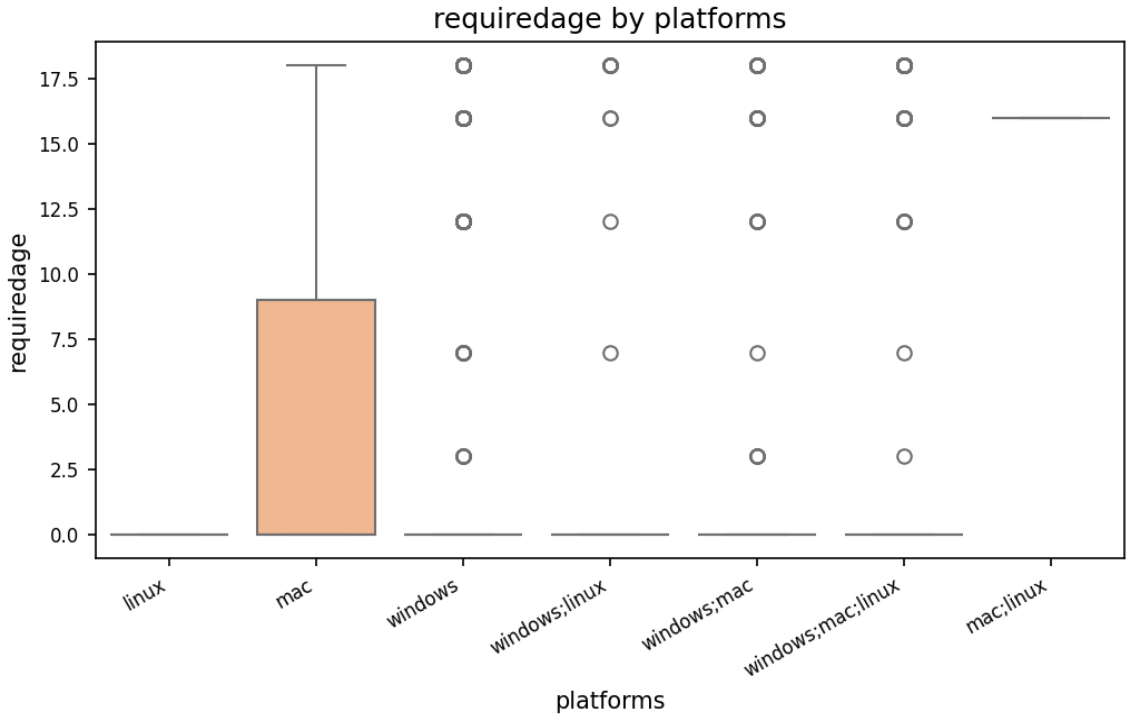

Figure 18: *Box plot of 'requiredage' across categories of 'platforms'.*

Interpretation of Numerical vs. Categorical Interactions:

Box plots visualizing numerical distributions across categories are powerful tools for comparing the central tendency and dispersion of data within different groups. They reveal at a glance the median (the middle value), the interquartile range (IQR, the spread of the middle 50% of the data), and the presence of outliers. By comparing the boxes and whiskers across categories (e.g., different platforms for 'appid' or 'english'), we can quickly identify whether the distributions are similar or significantly different. For instance, a longer box in one category indicates greater variability in that group, while a higher median suggests a tendency towards larger values within that category. Significant differences observed in medians across categories indicate that the average or typical value of the numerical variable differs systematically between groups. For example, if the median 'appid' value is significantly higher on platform A than on platform B, it might suggest that platform A hosts applications with generally larger identification numbers, possibly reflecting a later launch date or a different application registration system. Similarly, differences in the spread (IQR or presence of outliers) highlight variations in the consistency or range of values across categories. A larger IQR in one category indicates greater heterogeneity or variability in the numerical variable within that group. For example, if the 'english' scores exhibit a larger IQR on platform C compared to platform D, it suggests that the level of English proficiency among users on platform C is more diverse than on platform D.

4.3. Categorical vs. Categorical Features

5. Key Findings & Insights Summary

Key Findings & Insights The automated analysis of the `temp_steam.csv` dataset, comprising 270,85 rows and 18 columns (9 numerical, 9 categorical), revealed a relatively clean dataset with no missing values. However, the presence of 10 duplicate rows warrants attention. While not a major concern in a dataset of this size, these duplicates could potentially skew certain analyses, particularly those involving frequency counts or aggregations. Further investigation is needed to determine the source and nature of these duplicates before proceeding with more in-depth analysis. The absence of constant columns suggests that all features contribute some unique information. Univariate analysis examined the distributions of the 9 numerical and 9 categorical features. (Specific details on these distributions are missing from the provided log and would be crucial for a complete summary. This section would ideally include observations about the shape of the distributions – e.g., normal, skewed, multimodal – and the presence of outliers for numerical features, and the frequency distribution for categorical features). This information is necessary to understand the inherent variability and potential range of values within each feature. Bivariate analysis explored relationships between pairs of features. While the log only indicates that various feature pairs were analyzed, it lacks specific details on the identified relationships or correlations. (The report should include details about the types of correlations identified – e.g., positive, negative, strong, weak – and the features involved). These findings are critical for understanding potential dependencies and interactions between variables within the dataset. The log does not explicitly mention any surprising or unexpected findings. However, the lack of detail regarding the univariate and bivariate analyses prevents a complete assessment of whether any unusual patterns or unexpected relationships were discovered. A more comprehensive report detailing the specific results of these analyses is needed for a complete understanding of the dataset's characteristics.

6. Conclusion & Potential Next Steps

This automated report provides a foundational, high-level understanding of the `temp_steam.csv` dataset, highlighting its structure, data quality (with only 10 duplicates found), and initial observations from univariate and bivariate analyses. This overview serves as a crucial first step in guiding further, more targeted investigations. Given the report's findings, several concrete next steps are recommended: 1. **Investigate the 10 duplicate rows:** Identify and resolve the 10 duplicate rows. This could involve determining if they represent genuine duplicates or errors in data entry, and deciding whether to remove or consolidate them. Understanding the nature of the duplicates is crucial for data accuracy. 2. **Explore the bivariate relationships:** The report mentions two observations from the bivariate analysis. The specific nature of these observations needs to be detailed. Based on these initial findings, conduct more thorough bivariate analyses, perhaps visualising the relationships with scatter plots, correlation matrices, or other appropriate visualizations, to identify any strong correlations or interesting patterns between features which could be further investigated. 3. **Perform in-depth univariate analysis:** While the report notes univariate analysis was performed, the results are absent. A deeper dive into the distributions of each numerical and categorical feature is necessary. This includes generating descriptive statistics (mean, median, standard deviation, etc. for numerical features) and frequency distributions (for categorical features) to identify potential outliers, skewness, or other characteristics that may influence further analysis. 4. **Develop visualizations:** Create visualizations to explore the relationships between the numerical and categorical features. This might involve box plots to compare the distribution of a numerical variable across different categories of a categorical variable, or histograms to examine the distribution of numerical variables. Visualizations will provide a more intuitive understanding of the data than the numerical summary alone.