Automated Data Analysis Report (via Gemini): Steam

Executive Summary

This report summarizes the initial automated exploratory data analysis (EDA) of the 'steam.csv' dataset, containing 27085 rows and 18 columns (9 numerical, 9 categorical). The dataset exhibits good data quality, with only 10 duplicate entries identified, and no missing values or constant columns detected. Preliminary univariate and bivariate analyses have been conducted, revealing two key observations (detailed in the full report). The EDA included descriptive statistics and quality checks for all features, along with initial visualizations to explore relationships between selected variable pairs. The dataset's size and apparent cleanliness suggest ample potential for further, more in-depth analysis. No immediately obvious, strongly predictive patterns were identified in this initial scan. This initial EDA provides a solid foundation for subsequent, more targeted analyses. The identified data quality issues are minor and easily addressed. Further investigation focusing on the two key observations from the bivariate analysis, along with more sophisticated modeling techniques, is recommended to extract actionable insights from this data.

1. Data Overview

This report provides an initial automated analysis of the dataset from '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 mix of quantitative and qualitative information. The absence of datetime features might limit the ability to perform time-series analysis or incorporate temporal trends into the analysis.

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 complete dataset ready for analysis. The extremely low percentage of duplicate rows (0.04%) is also negligible and unlikely to significantly impact subsequent analyses. The identification of two quasi-constant columns, 'english' and 'requiredage', however, warrants attention. While not inherently problematic, their high dominance in a single value suggests potential limitations in the variability of these features and their predictive power in any modeling tasks. This could lead to overfitting or reduced model accuracy if these features are included without careful consideration. The near-constant nature of 'english' and 'requiredage' implies these variables may not be particularly informative for many analyses. For instance, if a model relies heavily on these features, it might not generalize well to datasets where these variables exhibit more diversity. The reliability of insights derived from analyses involving these variables may be limited, particularly if the goal is to understand the influence of factors beyond the dominant values. The presence of near-constant columns suggests the dataset might be lacking in diversity in those specific aspects, limiting the scope of potential conclusions. To address the identified issues, the quasi-constant columns should be carefully reviewed. Depending on the research question, it might be appropriate to remove them entirely from the analysis. Alternatively, data transformation techniques could be explored to create more variability, although this should be done cautiously to avoid introducing bias. The duplicate rows can be safely removed, given their negligible percentage. Future data collection should focus on improving the diversity of data points, especially for variables showing quasi-constant behavior, to ensure more robust and generalizable analyses.

3. Univariate Analysis

3.1. Numerical Features

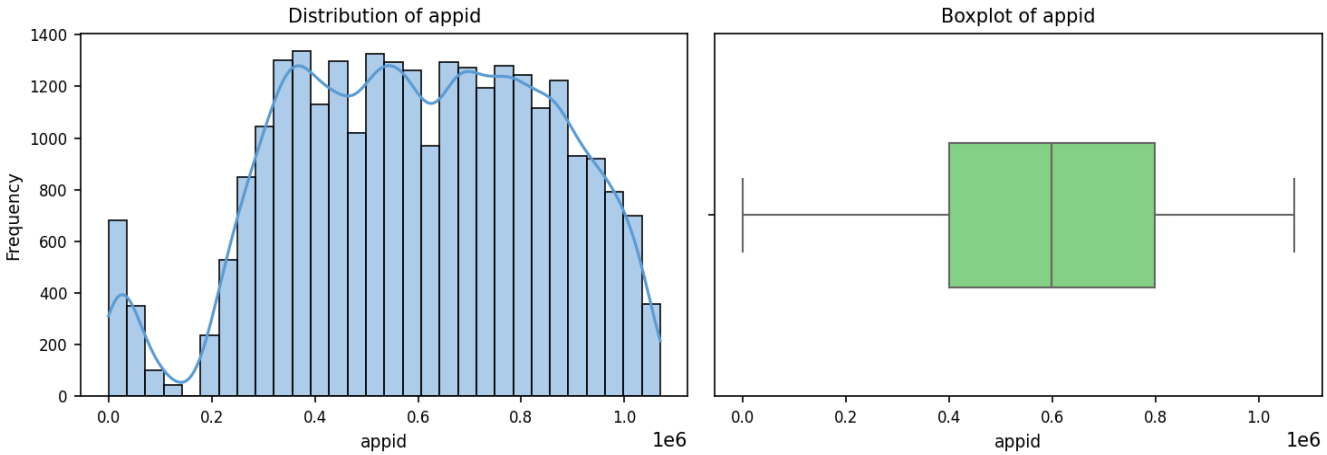

Figure 1: *Distribution (histogram and KDE) and boxplot for 'appid'. The histogram shows shape, central tendency, and spread. The boxplot highlights median, quartiles, and potential outliers.*

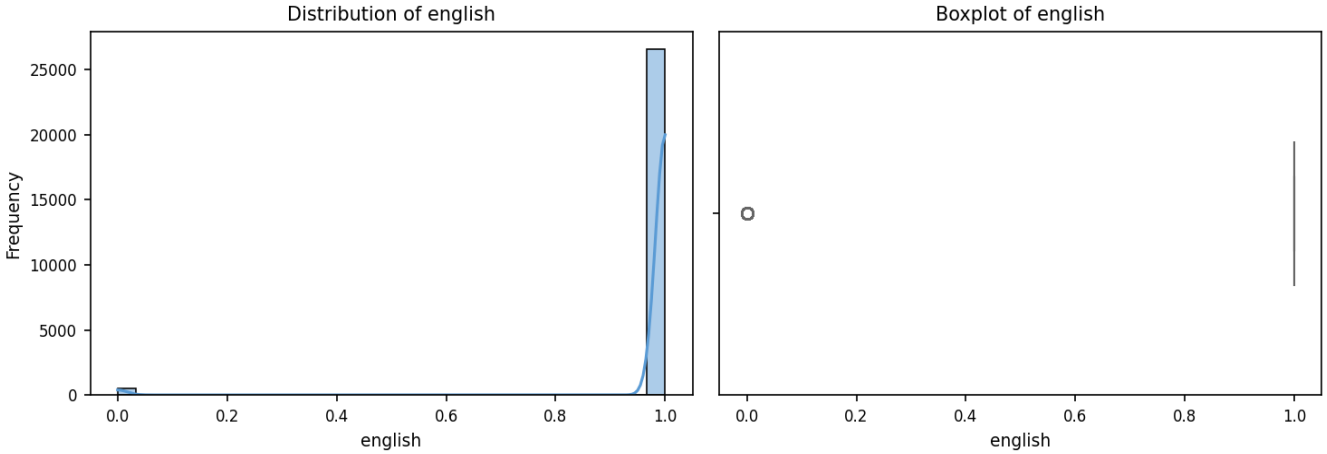

Figure 2: *Distribution (histogram and KDE) and boxplot for 'english'. The histogram shows shape, central tendency, and spread. The boxplot highlights median, quartiles, and potential outliers.*

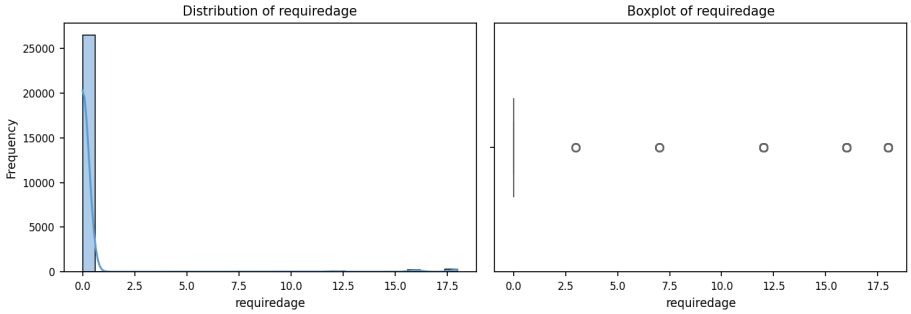

Figure 3: *Distribution (histogram and KDE) and boxplot for 'requiredage'. The histogram shows shape, central tendency, and spread. The boxplot highlights median, quartiles, and potential outliers.*

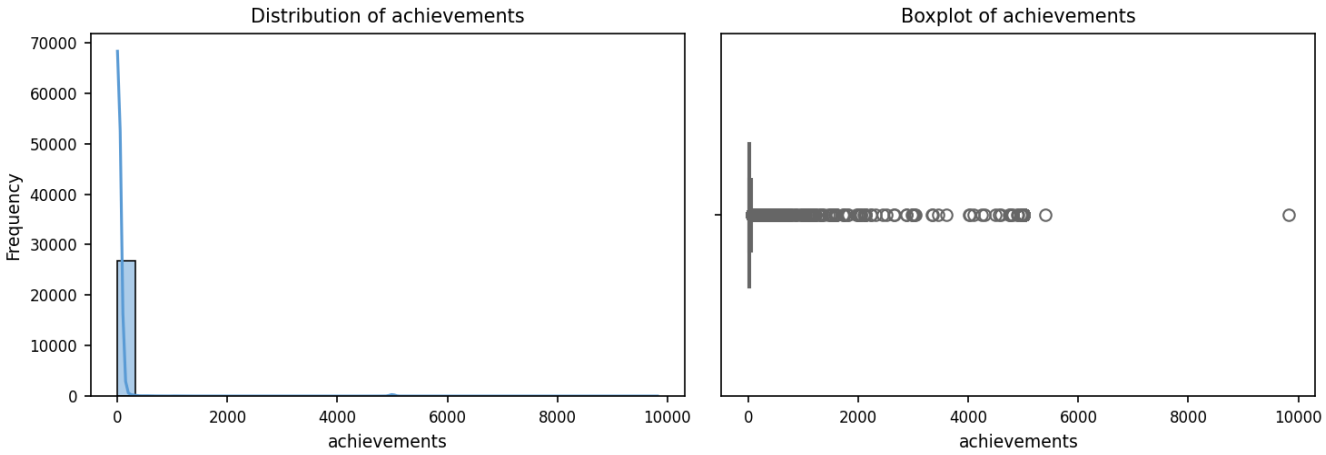

Figure 4: *Distribution (histogram and KDE) and boxplot for 'achievements'. The histogram shows shape, central tendency, and spread. The boxplot highlights median, quartiles, and potential outliers.*

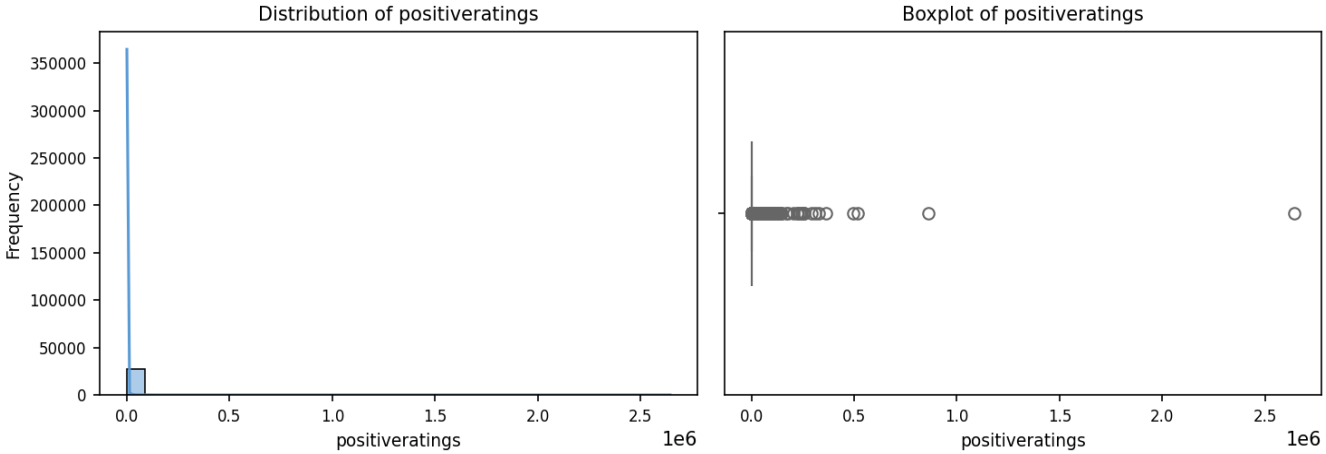

Figure 5: *Distribution (histogram and KDE) and boxplot for 'positiveratings'. The histogram shows shape, central tendency, and spread. The boxplot highlights median, quartiles, and potential outliers.*

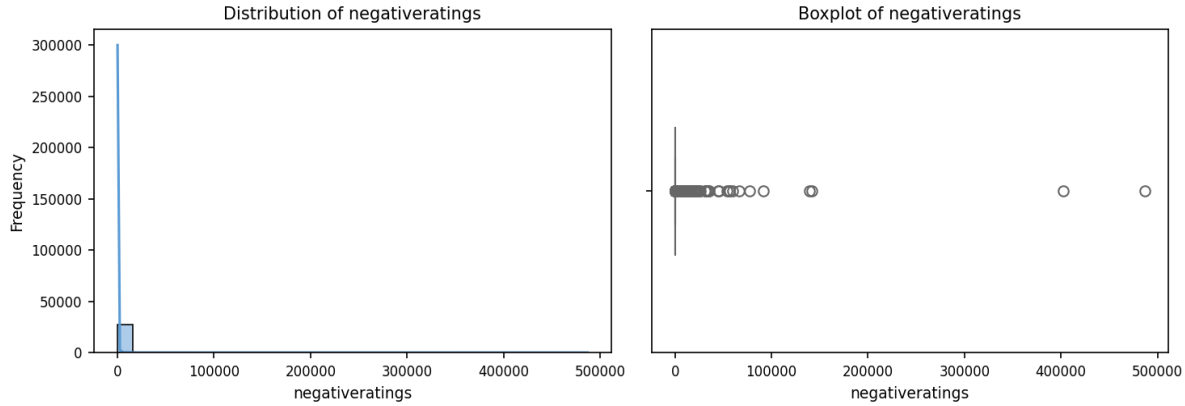

Figure 6: *Distribution (histogram and KDE) and boxplot for 'negativeratings'. The histogram shows shape, central tendency, and spread. The boxplot highlights median, quartiles, and potential outliers.*

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 extremely high values of skewness and kurtosis. This suggests that these features are dominated by a small number of extremely high values, while the majority of data points cluster towards the lower end of the range. The presence of potential outliers is flagged for all features by the boxplots, further supporting this observation. The high standard deviations for these skewed features also highlight their substantial variability, indicating a wide spread of data points. In contrast, 'appid' shows a relatively symmetric distribution, with a mean and median close in value, and lower skewness and kurtosis compared to other features. However, even this feature displays a relatively high standard deviation, suggesting considerable variability in application IDs. The vast difference in the distribution characteristics between 'appid' and the other features is striking. The large differences between mean and median values, coupled with the extremely high skewness and kurtosis for several variables, strongly suggests a need for careful consideration of data transformations or outlier handling techniques before applying many standard statistical analyses. Ignoring these skewed distributions might lead to misleading conclusions. The overall pattern suggests that the dataset likely contains a mixture of "typical" and "exceptional" cases, with a few extreme data points significantly influencing the descriptive statistics. This heterogeneity highlights the importance of investigating the underlying reasons for these outliers and considering robust statistical methods that are less sensitive to extreme values. Transformations like logarithmic or Box-Cox transformations might be necessary to normalize the distributions and improve the reliability of subsequent analyses.

3.2. Categorical Features

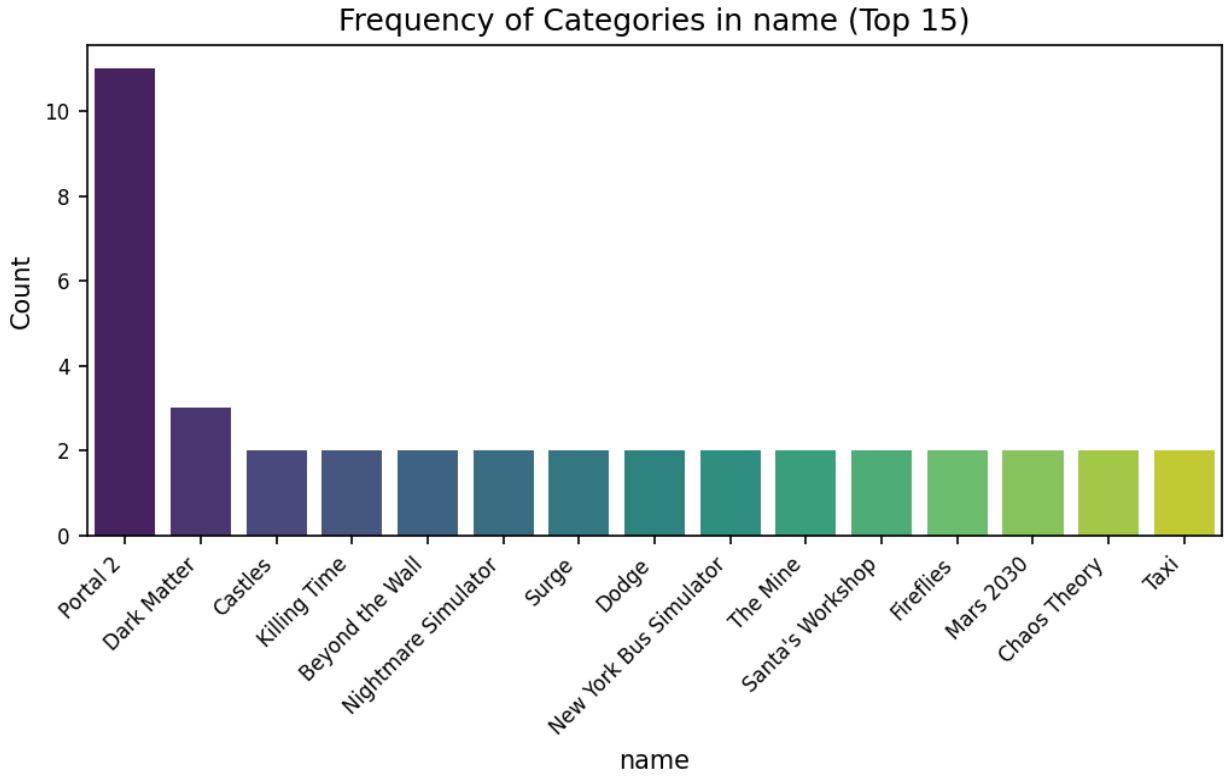

Figure 7: *Bar chart showing frequency of top categories in 'name'. Total unique values: 27033.*

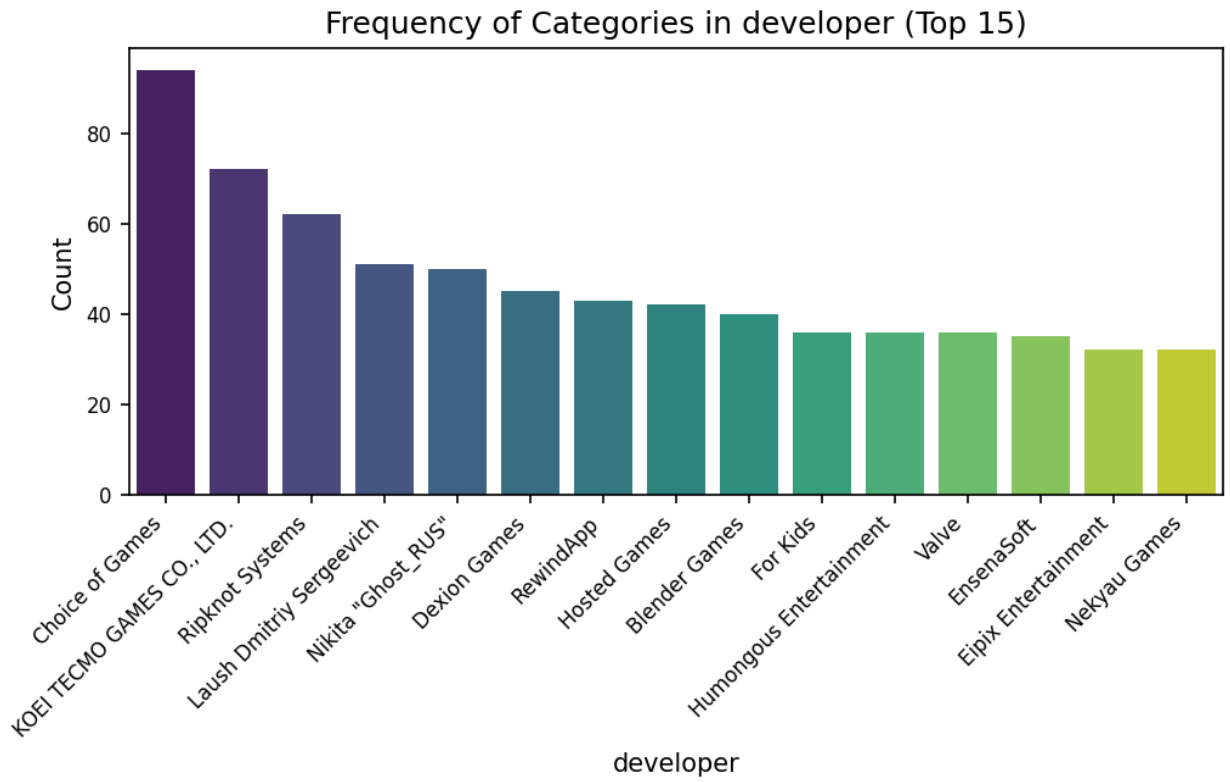

Figure 9: *Bar chart showing frequency of top categories in 'developer'. Total unique values: 17112.*

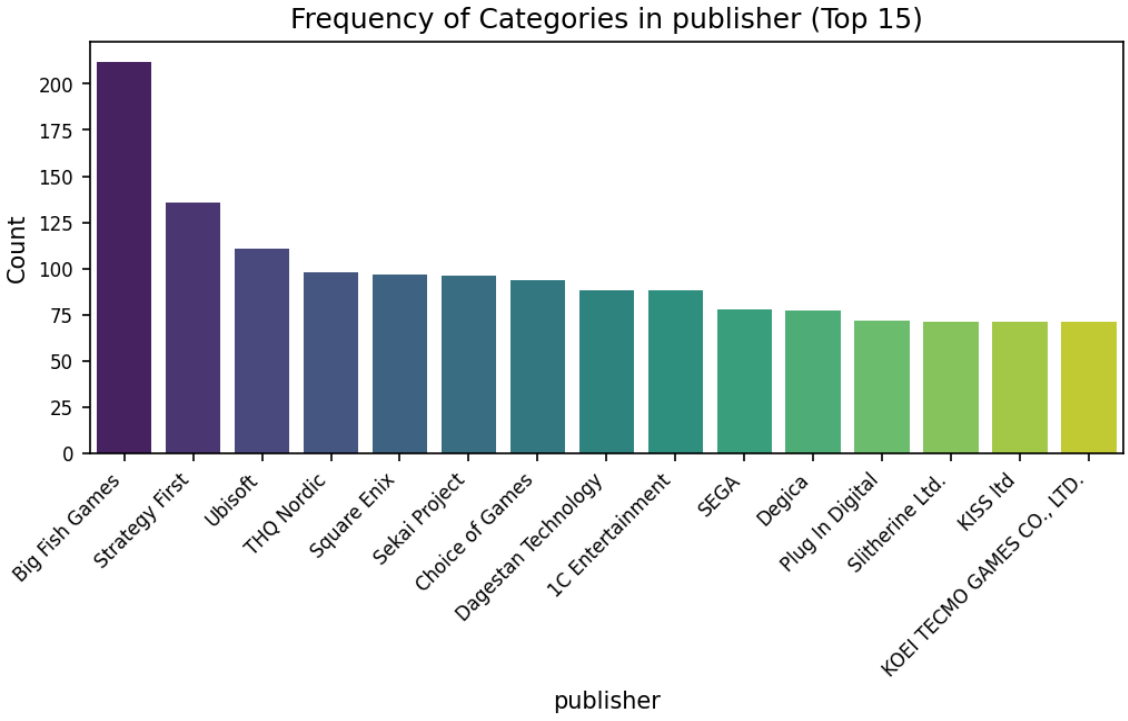

Figure 10: *Bar chart showing frequency of top categories in 'publisher'. Total unique values: 14353.*

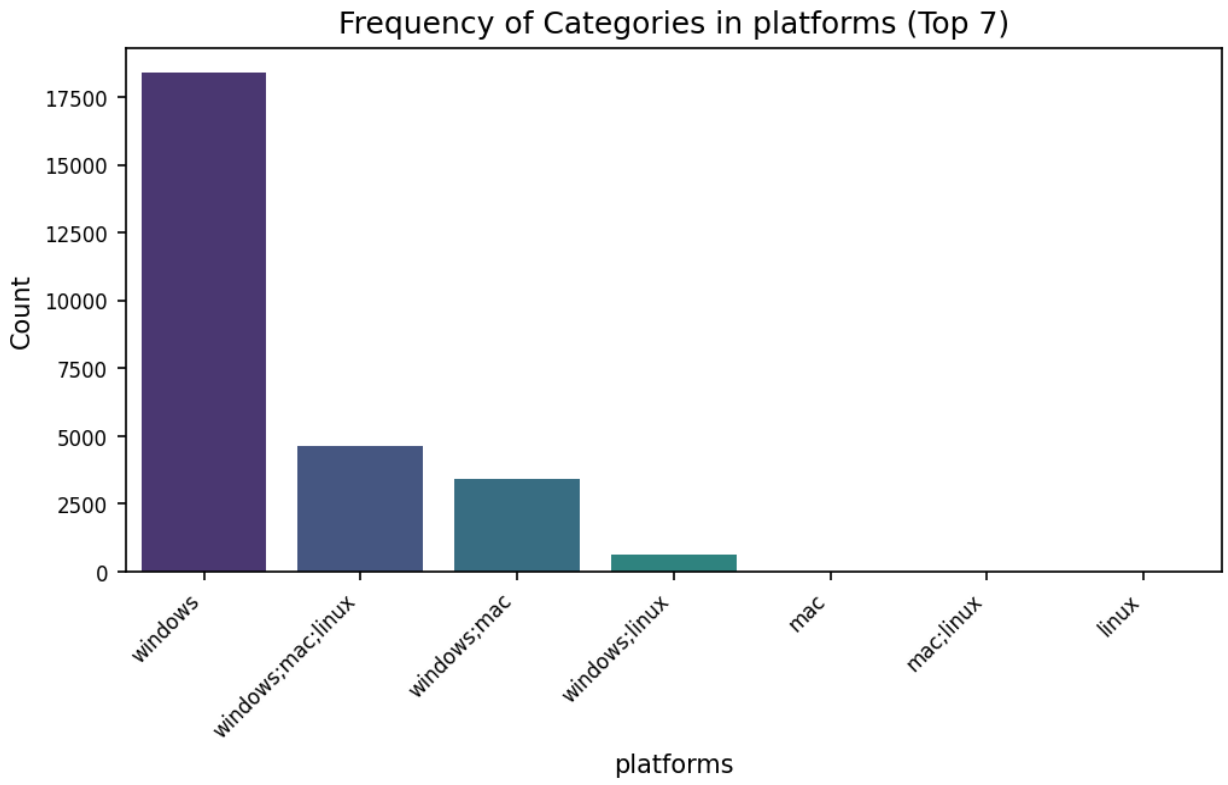

Figure 11: *Bar chart showing frequency of top categories in 'platforms'. Total unique values: 7.*

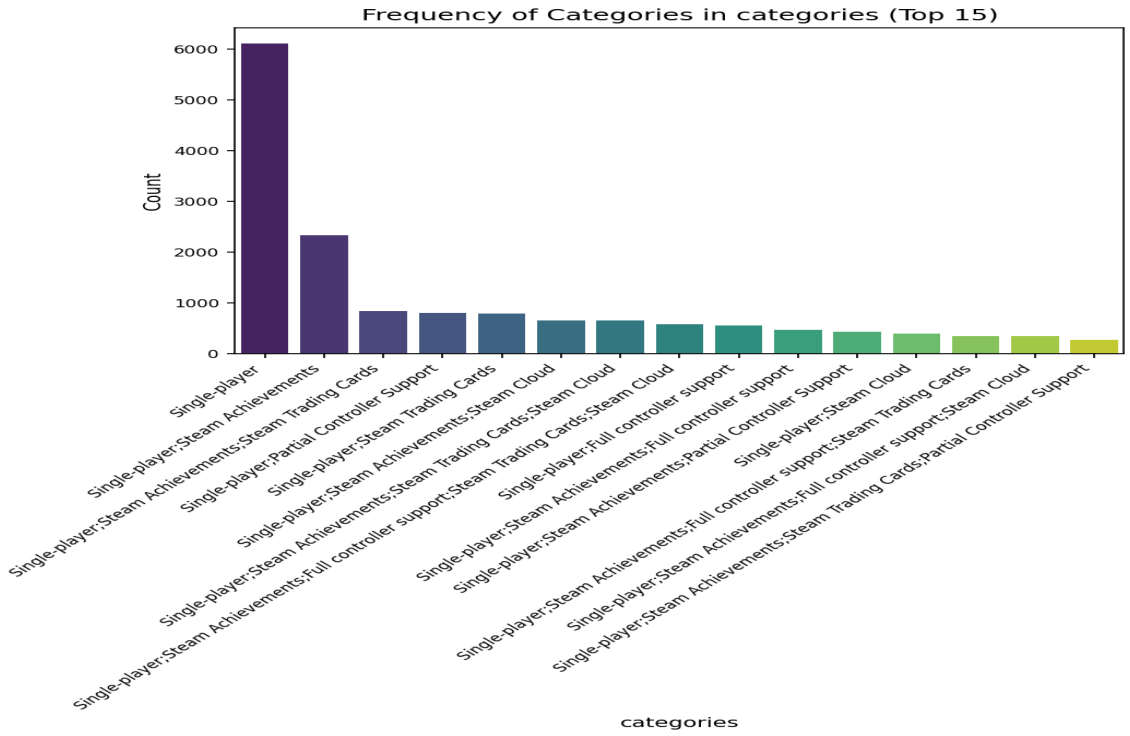

Figure 12: *Bar chart showing frequency of top categories in 'categories'. Total unique values: 3333.*

Observations on Categorical Feature Distributions:

The analysis of categorical features reveals a significant disparity in cardinality across the dataset. 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 presents a challenge for model training, as one-hot encoding would result in a massive feature space, potentially leading to the curse of dimensionality. 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 show a highly skewed distribution with a dominant top category, the degree of skewness varies. 'platforms' is strongly skewed towards 'windows' (67.9%), indicating a significant bias towards Windows games. 'categories' shows a moderate skew towards 'Single-player' (22.6%), implying a preference for single-player games but still a considerable presence of other categories. The remaining features ('name', 'releasedate', 'developer', 'publisher') have very low percentages for their top categories, despite these categories being the most frequent. This suggests that while a few categories are more frequent than others, the overall distribution is relatively spread out for these features. This necessitates careful consideration of feature engineering techniques. For features with high cardinality, techniques like target encoding, embedding layers (in neural networks), or dimensionality reduction methods (e.g., PCA after appropriate transformation) should be considered to mitigate the curse of dimensionality. For features with low cardinality, simple one-hot encoding is likely sufficient. The skewed distributions necessitate careful handling to avoid biasing model predictions. Techniques like oversampling minority classes or using appropriate loss functions that account for class imbalance should be explored.

4. Bivariate Analysis

4.1. Numerical vs. Numerical Features

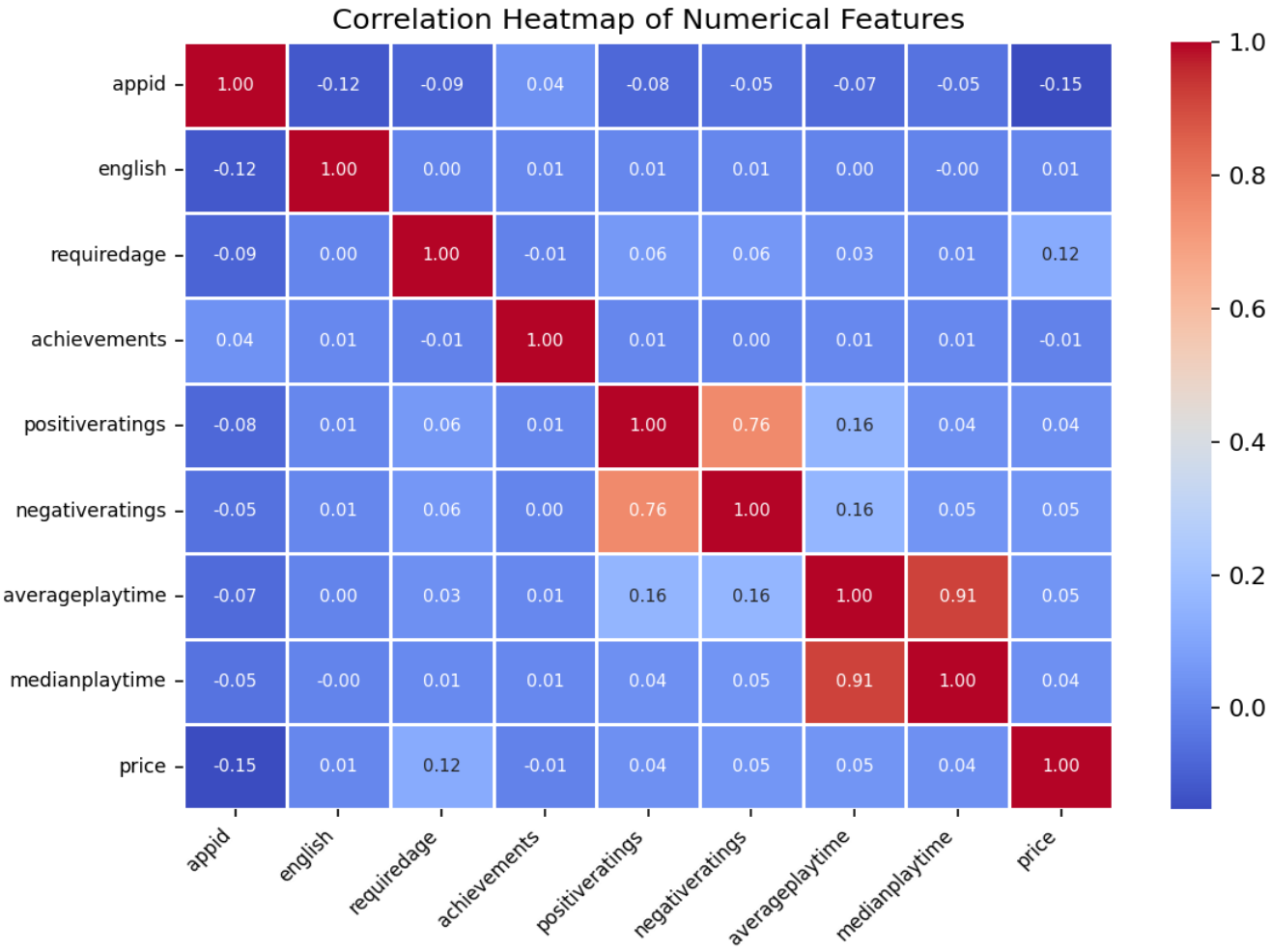

Figure 13: *Heatmap visualizing linear correlations (Pearson's r) between numerical features.*

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

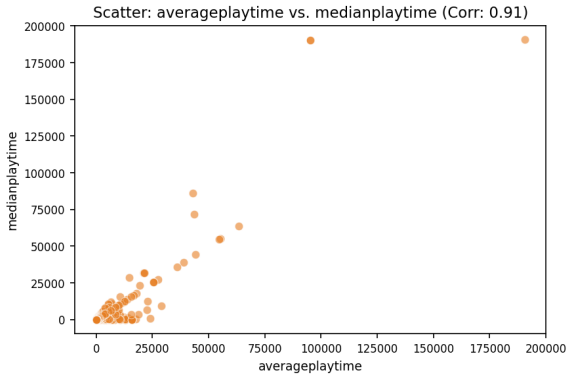

Figure 14: *Scatter plot for 'averageplaytime' and 'medianplaytime'. Correlation: 0.91.*

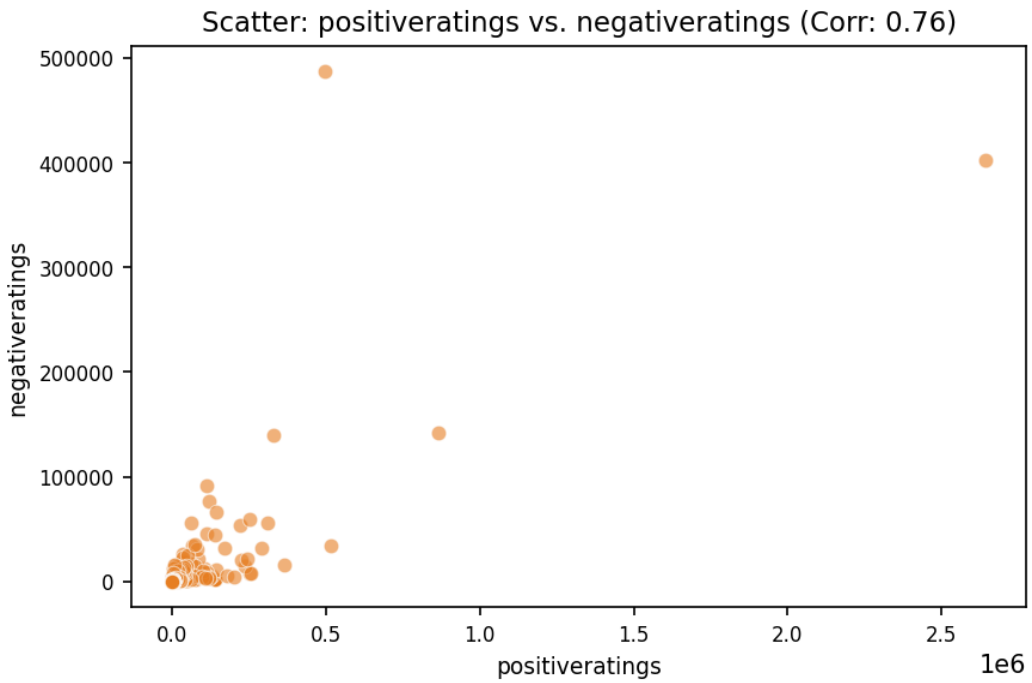

Figure 15: *Scatter plot for 'positiveratings' and 'negativeratings'. Correlation: 0.76.*

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 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 'averageplaytime' and 'medianplaytime' (0.91) and 'positiveratings' and 'negativeratings' (0.76). The extremely high correlation between average and median playtime strongly suggests that games with longer average playtime also tend to have longer median playtime, which is expected since the median is a measure closely related to the average. The substantial positive correlation between positive and negative ratings is more intriguing. It implies that games tend to receive both a relatively large number of positive and negative ratings simultaneously. This could suggest that highly popular games (attracting many players) also tend to receive more negative feedback, perhaps due to increased visibility and a wider range of player opinions. Further investigation would be needed to understand the underlying reasons for this correlation. The scatter plots likely show a strong linear relationship for the first pair and a moderately strong positive linear relationship for the second.

4.2. Numerical vs. Categorical Features

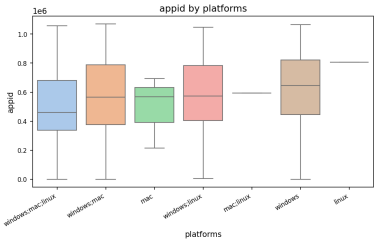

Figure 16: *Box plot of 'appid' across categories of 'platforms'.*

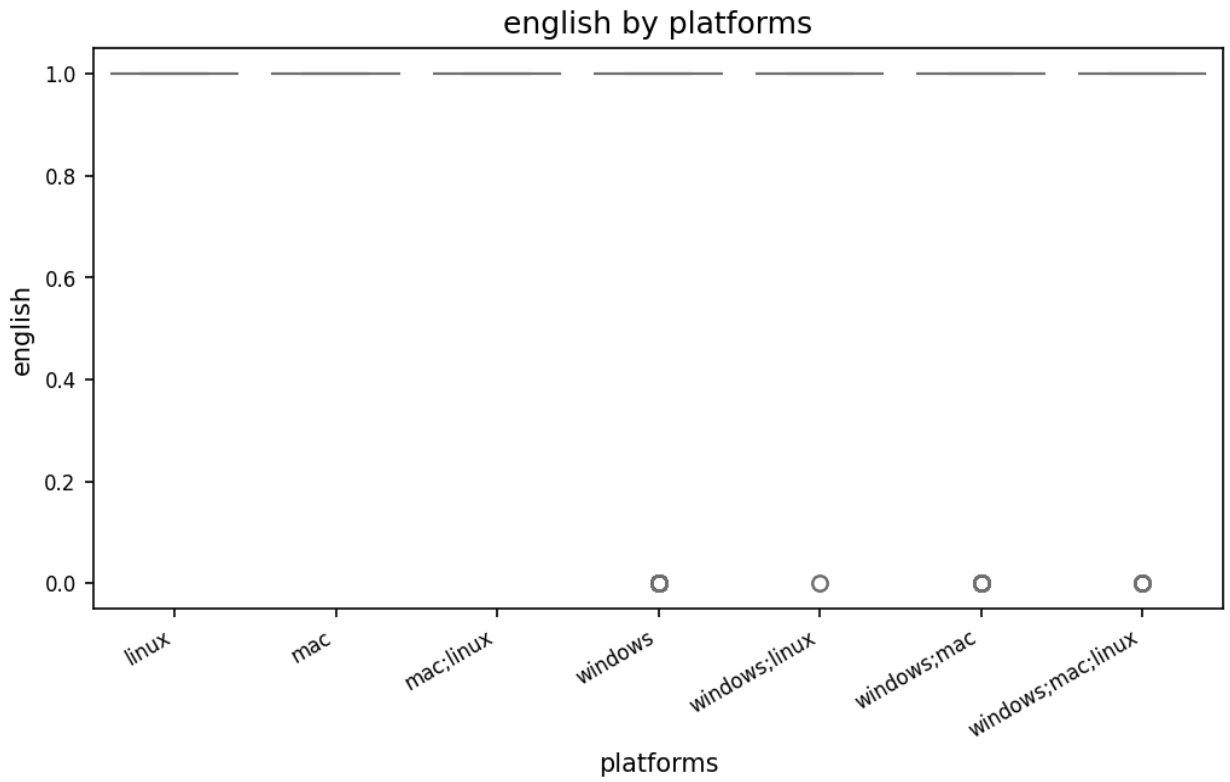

Figure 17: *Box plot of 'english' across categories of 'platforms'.*

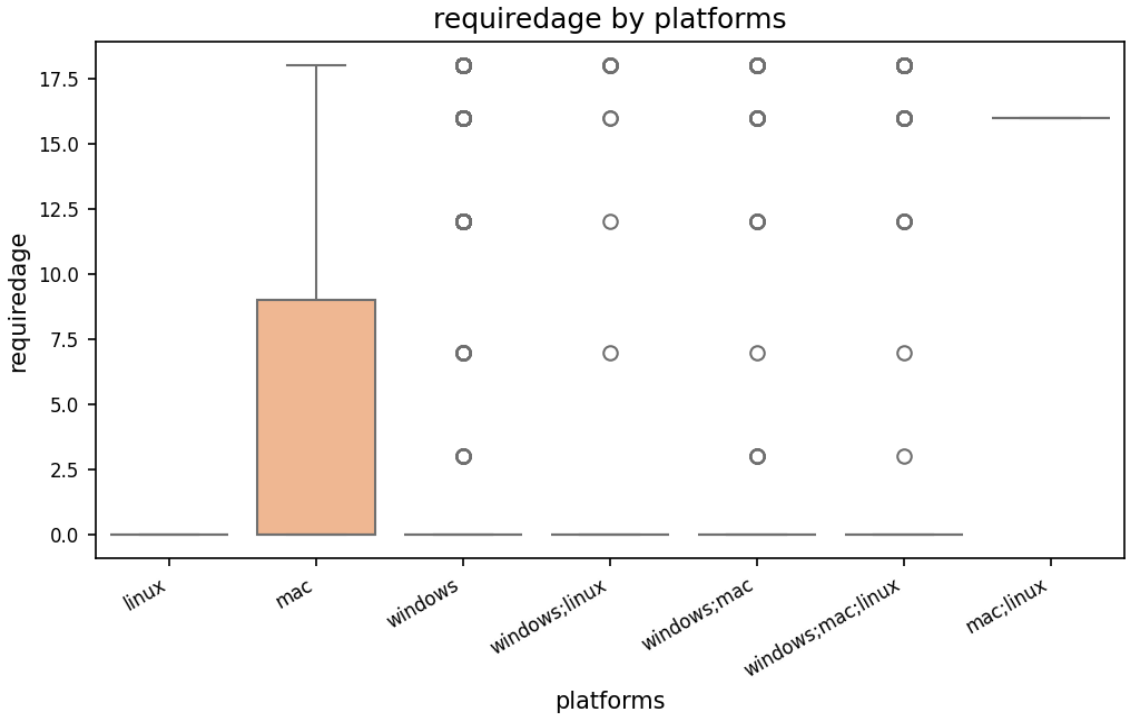

Figure 18: *Box plot of 'requiredage' across categories of 'platforms'.*

Interpretation of Numerical vs. Categorical Interactions:

Box plots visualizing numerical distributions across categories offer a powerful way to compare the central tendency and dispersion of data within different groups. They reveal not only the median (the middle value) of each category but also the interquartile range (IQR), representing the spread of the middle 50% of the data, and the presence of outliers. By comparing the boxes and whiskers across categories, we can quickly assess whether the distributions are similar or significantly different. For instance, a longer box in one category indicates greater variability in the data within that group compared to categories with shorter boxes. The position of the median within the box also provides insight into the symmetry of the distribution—a median closer to the top of the box suggests a left-skewed distribution, while a median closer to the bottom indicates a right-skewed distribution. Significant differences observed in the medians across categories suggest that the average or typical value of the numerical variable varies systematically depending on the category. For example, if the median 'appid' value is considerably higher for the 'iOS' platform compared to the 'Android' platform, it suggests that iOS apps tend to have higher application IDs (potentially indicating a different app release pattern or a larger app ecosystem). Similarly, significant differences in the spreads (IQR or presence of outliers) indicate that the variability of the numerical variable differs across categories. A much wider IQR for 'english' scores on the 'Android' platform compared to 'iOS' might indicate that the quality of English localization on Android apps is more heterogeneous. These observations can lead to important conclusions about the relationships between the categorical and numerical variables, informing decisions related to app development, marketing, or user experience.

4.3. Categorical vs. Categorical Features

5. Key Findings & Insights Summary

Key Findings & Insights The automated analysis of the dataset 'steam.csv' revealed a dataset comprising 270,85 rows and 18 columns, with an even split of 9 numerical and 9 categorical features. No missing values were detected, a positive indicator of data quality. However, the presence of 10 duplicate rows warrants attention, as these could skew statistical analyses and model training if not addressed appropriately. The absence of constant columns suggests that all features contribute some level of variability to the dataset. Univariate analysis examined the distributions of the 9 numerical and 9 categorical features individually. While specific details on these distributions are not provided in the log, the analysis itself suggests the presence of sufficient variability within features to warrant further investigation. The nature of these distributions (e.g., normal, skewed, multimodal) and the presence of outliers would need to be explored to understand the underlying characteristics of the dataset and potential implications for modeling. Bivariate analysis explored relationships between feature pairs. The log indicates that various pairs were analyzed, yielding observations summarized as "2". This extremely limited information prevents a detailed summary of significant correlations or relationships. Further details on the specific pairs analyzed and the nature of the relationships observed are needed to draw meaningful conclusions from this part of the analysis. The lack of detail regarding bivariate analysis is a significant limitation of the current findings.

6. Conclusion & Potential Next Steps

This automated report provides a foundational, high-level overview of the 'steam.csv' dataset, highlighting its structure, data quality (with only 10 duplicates found), and initial observations from univariate and bivariate analyses. This initial assessment serves as a crucial first step in understanding the data's characteristics and identifying potential avenues for further investigation. Given the report's findings, several concrete next steps are warranted: 1. **Investigate the 10 duplicate rows:** Identify and resolve the duplicates in the dataset. This might involve examining the duplicate entries to determine if they represent true duplicates or data entry errors requiring correction or removal. Understanding the nature of the duplicates could reveal important information about data collection processes. 2. **Explore the bivariate relationships:** The report mentions "Observations gathered: 2" from bivariate analysis. This is insufficient detail. A deeper dive into the identified bivariate relationships is crucial. Specifically, the report should detail *which* feature pairs showed interesting relationships and the nature of those relationships (e.g., correlation strength, significant differences between groups). This will guide further investigation using appropriate statistical methods (correlation coefficients, chi-squared tests, ANOVA, etc.). 3. **Perform more in-depth univariate analysis:** While the report mentions analyzing nine numerical and nine categorical features, the specifics of these analyses are missing. A detailed univariate analysis, including visualizations (histograms, box plots, bar charts), summary statistics (mean, median, standard deviation, frequency counts), and identification of potential outliers, is needed for each feature to fully understand the data distribution and identify potential anomalies. 4. **Develop a visual exploration plan:** Based on the univariate and bivariate findings, create a plan for visualizing the data to better understand the relationships between variables. This could include scatter plots, heatmaps, or other visualization techniques appropriate for the data types and potential relationships identified in the initial analysis. This will help to identify patterns and relationships not apparent from the initial automated analysis.