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 27,085 rows and 18 columns of data related to Steam games. The dataset comprises nine numerical and nine categorical features, with no datetime variables. Initial quality checks revealed ten duplicate entries, but no missing values or constant columns, suggesting a relatively clean dataset for further analysis. Preliminary univariate and bivariate analyses, including descriptive statistics and visual inspections (details in the full report), have been performed on all features. While two initial observations regarding relationships between features have been noted (detailed in the full report), further investigation is required to fully understand the underlying patterns and relationships within the data. This initial EDA provides a solid foundation for subsequent, more in-depth analyses. The absence of significant data quality issues and the identification of potential areas of interest suggest that the dataset holds value for further modeling and predictive tasks. The next steps will involve more detailed analysis of the identified patterns and a deeper dive into specific feature relationships.

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 exhibits a roughly even split between numerical and categorical features, which is a relatively typical mix for many datasets. The absence of datetime features might limit certain types of time-series or temporal analyses, but the presence of both numerical and categorical data allows for a wide range of analytical 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 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 generally high level of cleanliness in the dataset with 27,085 rows. The absence of missing values is a significant positive, indicating a complete dataset ready for analysis. The presence of only 10 duplicate rows (0.04%) is negligible and poses minimal risk to the integrity of the analysis. The absence of constant columns is also positive, suggesting a reasonable level of variability in the features. However, the identification of two quasi-constant columns, 'english' and 'requiredage', warrants attention. While not problematic in themselves, these high proportions of a single value (98.1% and 97.8% respectively) suggest limited variability and potential redundancy. This could impact modeling performance, potentially leading to overfitting or reduced predictive power, especially if these features are included without careful consideration. Insights derived from analyses involving these columns might be overly reliant on the dominant value and less representative of the overall dataset. To address the quasi-constant columns, several strategies can be employed. One approach is to carefully evaluate the importance of these features in the context of the overall analysis goals. If they provide minimal additional information given their high skew, they could be removed. Alternatively, binning or transforming these variables (e.g., creating a new binary variable indicating the presence or absence of the minority value) might be explored to better capture the subtle variations present within the small minority. The duplication issue is minor and could be resolved simply by removing the duplicate rows. The overall data quality is good, but attention to the quasi-constant columns is crucial for robust and reliable analysis.

3. Univariate Analysis

3.1. Numerical Features

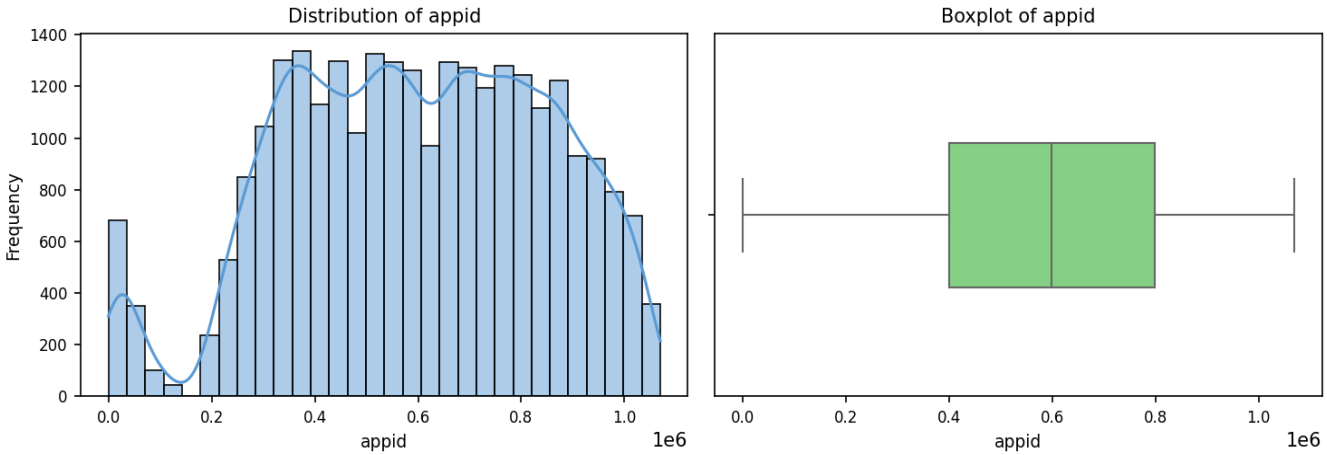


Figure 1: *Distribution (histogram and KDE) and boxplot for 'appid'. The histogram shows shape, central tendency, and spread. The boxplot highlights median, quartiles, and potential outliers.*

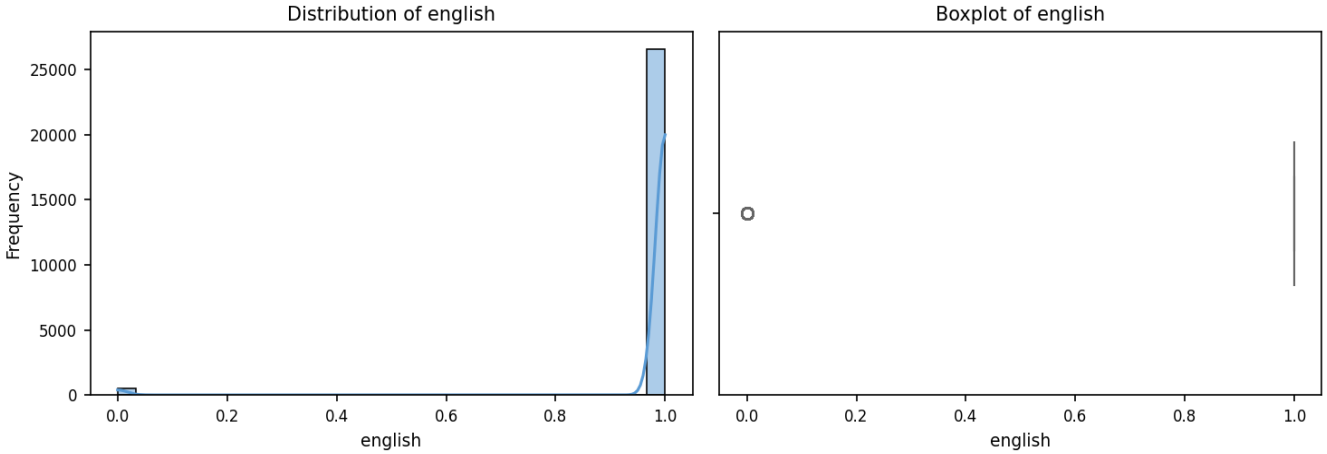


Figure 2: *Distribution (histogram and KDE) and boxplot for 'english'. The histogram shows shape, central tendency, and spread. The boxplot highlights median, quartiles, and potential outliers.*

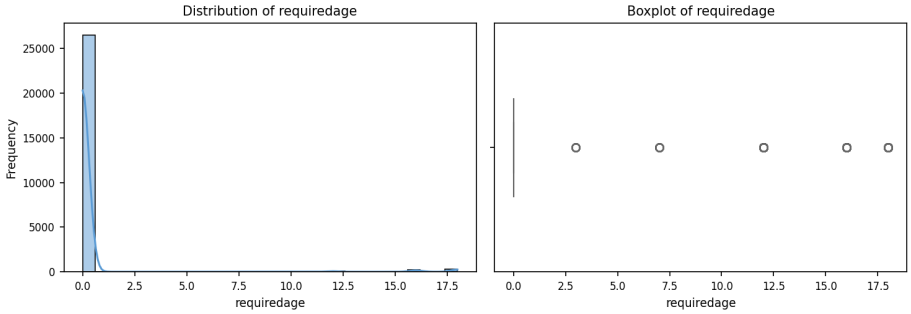


Figure 3: *Distribution (histogram and KDE) and boxplot for 'requiredage'. The histogram shows shape, central tendency, and spread. The boxplot highlights median, quartiles, and potential outliers.*

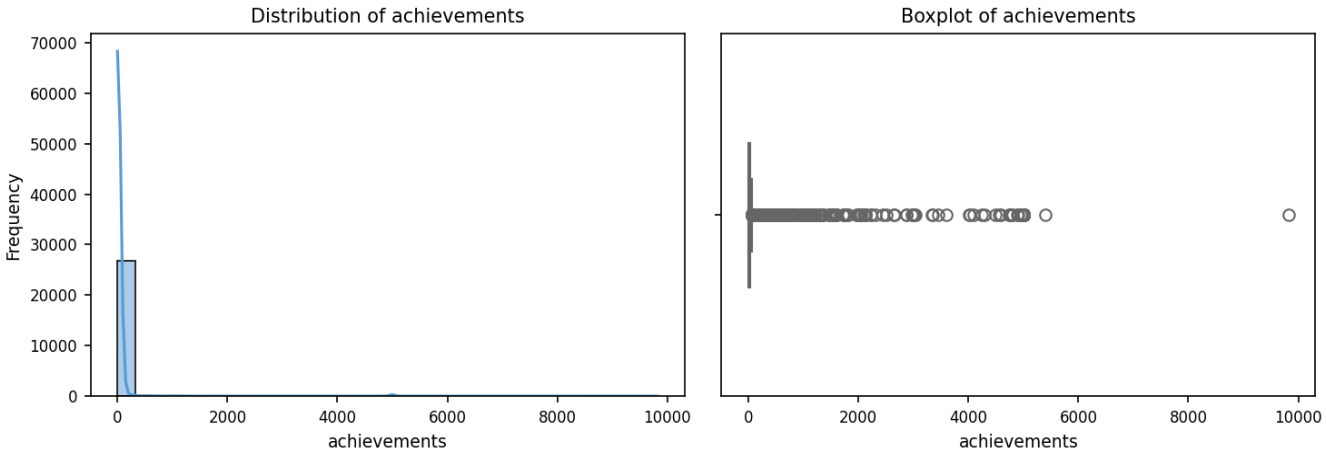


Figure 4: *Distribution (histogram and KDE) and boxplot for 'achievements'. The histogram shows shape, central tendency, and spread. The boxplot highlights median, quartiles, and potential outliers.*

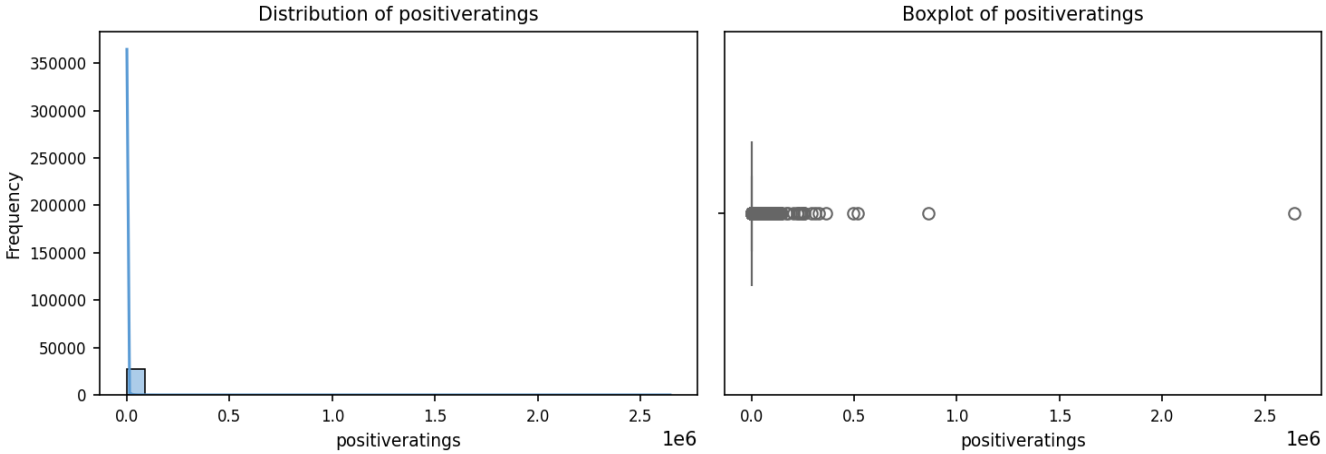


Figure 5: *Distribution (histogram and KDE) and boxplot for 'positiveratings'. The histogram shows shape, central tendency, and spread. The boxplot highlights median, quartiles, and potential outliers.*

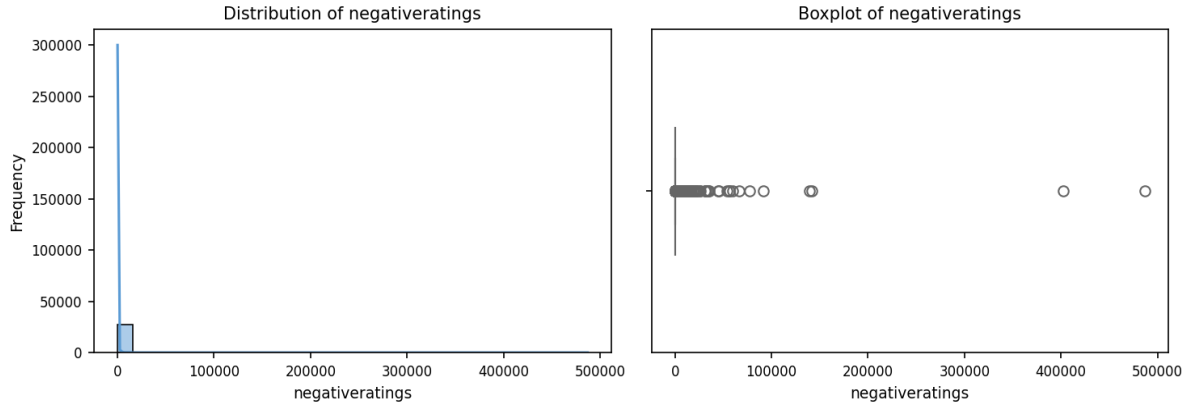


Figure 6: *Distribution (histogram and KDE) and boxplot for 'negativeratings'. The histogram shows shape, central tendency, and spread. The boxplot highlights median, quartiles, and potential outliers.*

Observations on Numerical Feature Distributions:

The analysis of the numerical features reveals a striking pattern of highly skewed distributions, predominantly right-skewed, with a significant presence of potential outliers. Features like 'achievements', 'positiveratings', and 'negativeratings' exhibit extremely high skewness and kurtosis values, indicating long right tails and a concentration of data points around the lower end of their range. The large discrepancies between means and medians further emphasize this right skew, where the mean is heavily influenced by a small number of extremely high values. The high standard deviations for these features also highlight their substantial variability. In contrast, 'appid' shows a relatively symmetric distribution, although the boxplot suggests the presence of outliers. 'english' shows a strong left skew, indicating a concentration of values close to 1.0. 'requiredage' also exhibits a right skew, with a mean significantly higher than the median. The presence of potential outliers, indicated by both the boxplots and the extreme differences between minimum/maximum values and the mean/median, is a pervasive issue across most features. This suggests that these features may be influenced by a small number of extreme data points that could disproportionately affect statistical analyses. Careful consideration should be given to handling these outliers, potentially through transformation techniques or robust statistical methods, to avoid biased results. The high variability, as evidenced by the large standard deviations, particularly in features like 'achievements', 'positiveratings', and 'negativeratings', further underscores the need for robust analysis techniques that are less sensitive to extreme values. In summary, the data exhibits a clear pattern of heavily skewed distributions with numerous potential outliers, especially in the features related to ratings and achievements. This necessitates a cautious approach to data analysis, requiring careful consideration of outlier treatment and the use of robust statistical methods to accurately reflect the underlying data patterns and avoid misinterpretations driven by extreme values. The relatively symmetric distribution of 'appid' stands in contrast to the others and might warrant separate investigation.

3.2. Categorical Features

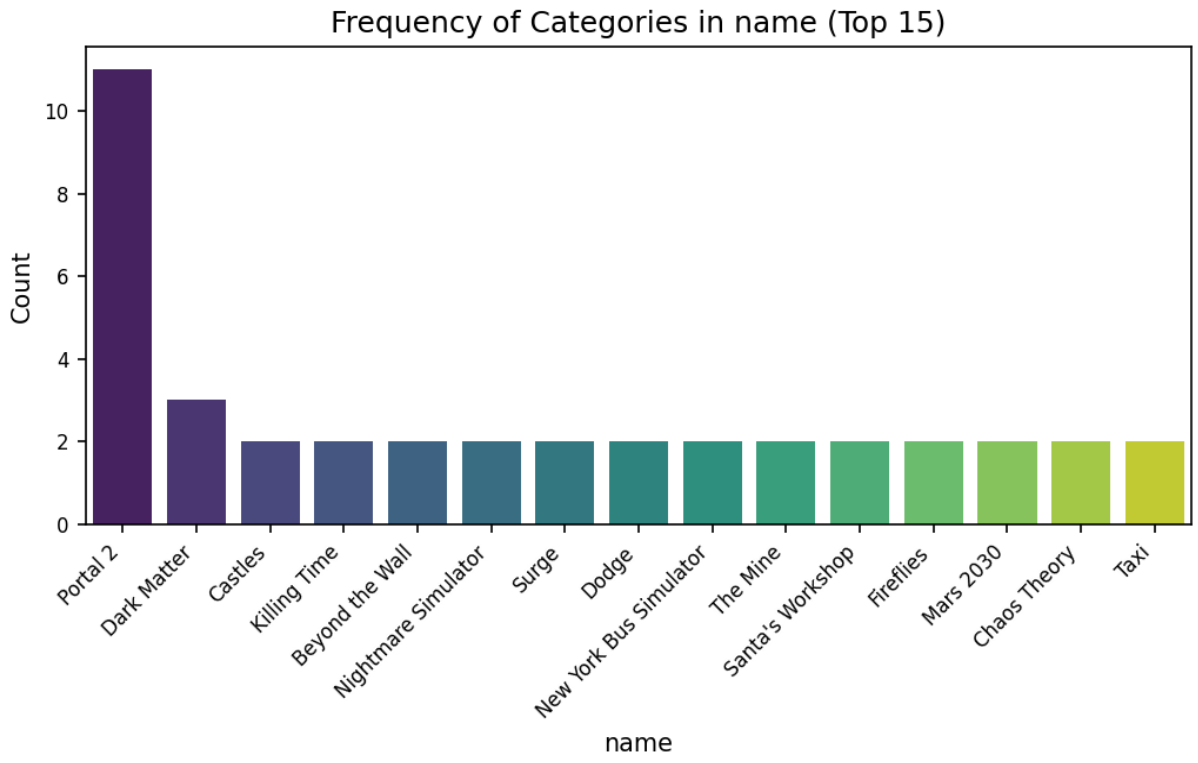


Figure 7: *Bar chart showing frequency of top categories in 'name'. Total unique values: 27033.*

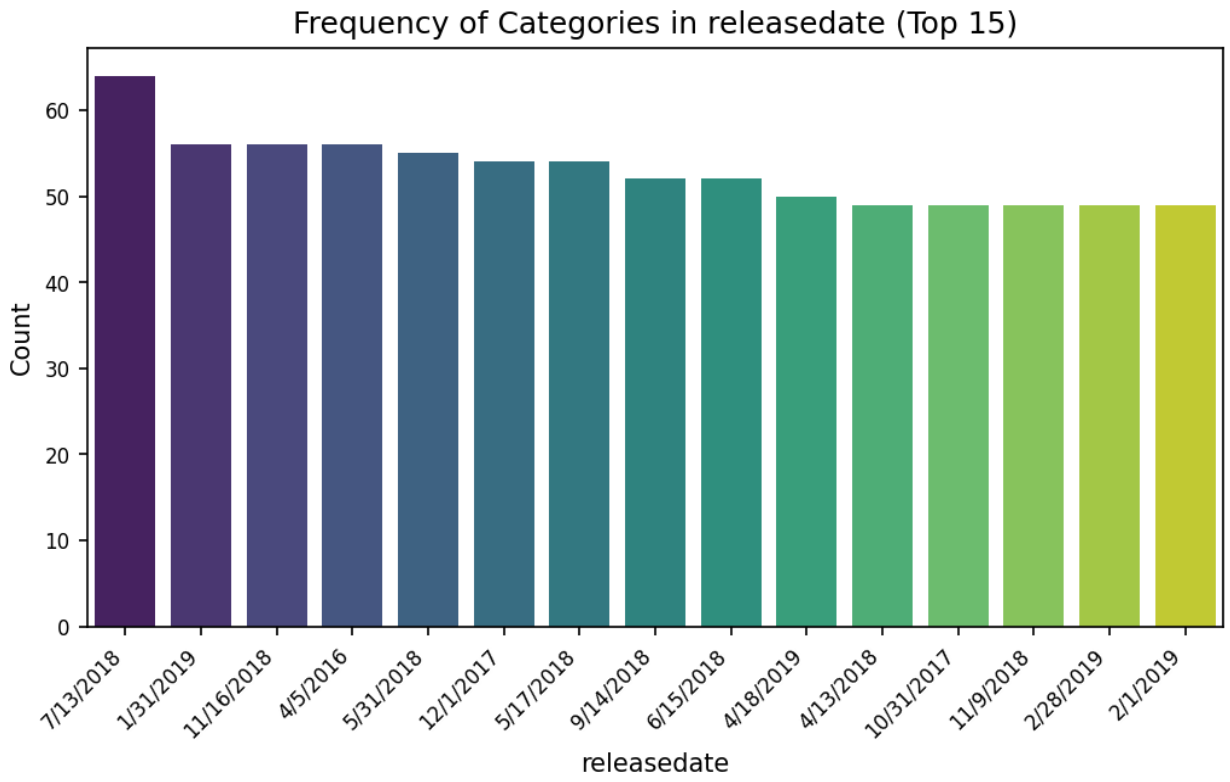


Figure 8: *Bar chart showing frequency of top categories in 'releasedate'. Total unique values: 2619.*

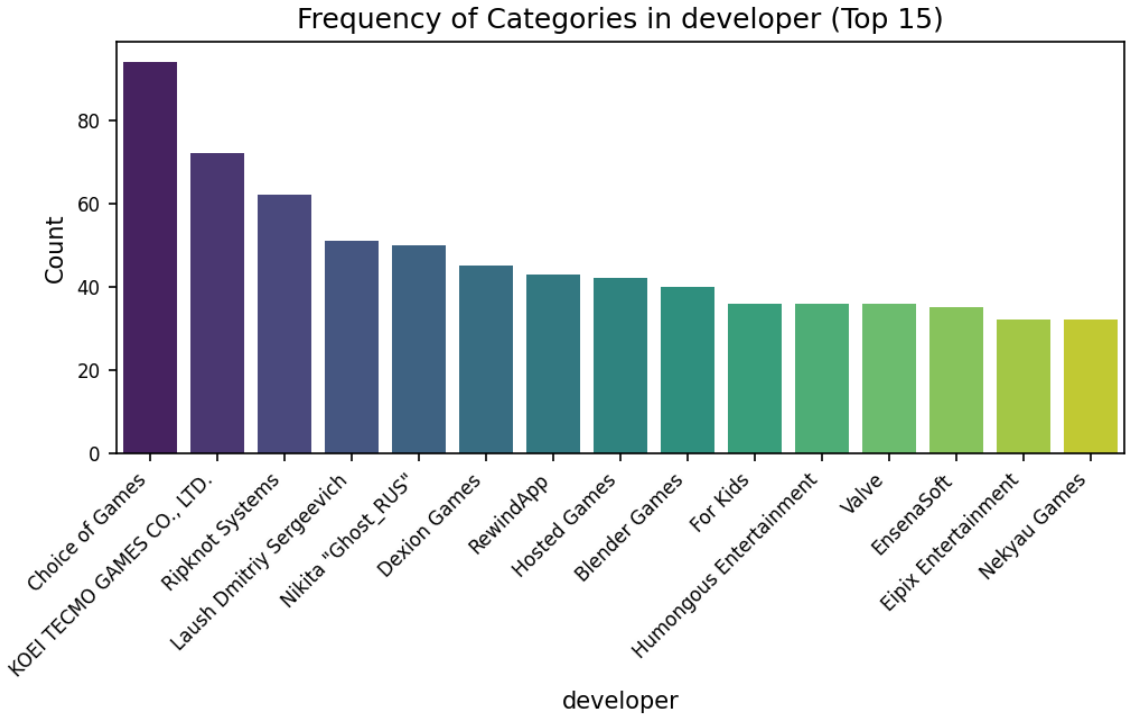


Figure 9: *Bar chart showing frequency of top categories in 'developer'. Total unique values: 17112.*

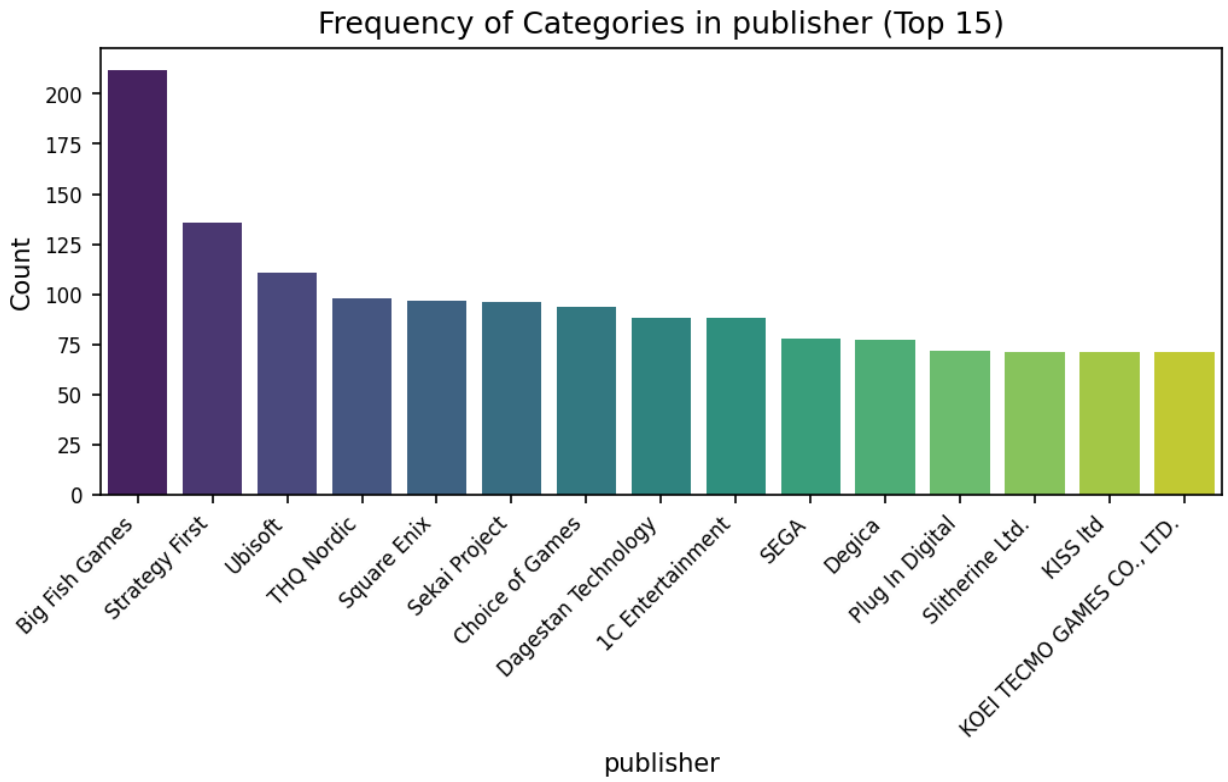


Figure 10: *Bar chart showing frequency of top categories in 'publisher'. Total unique values: 14353.*

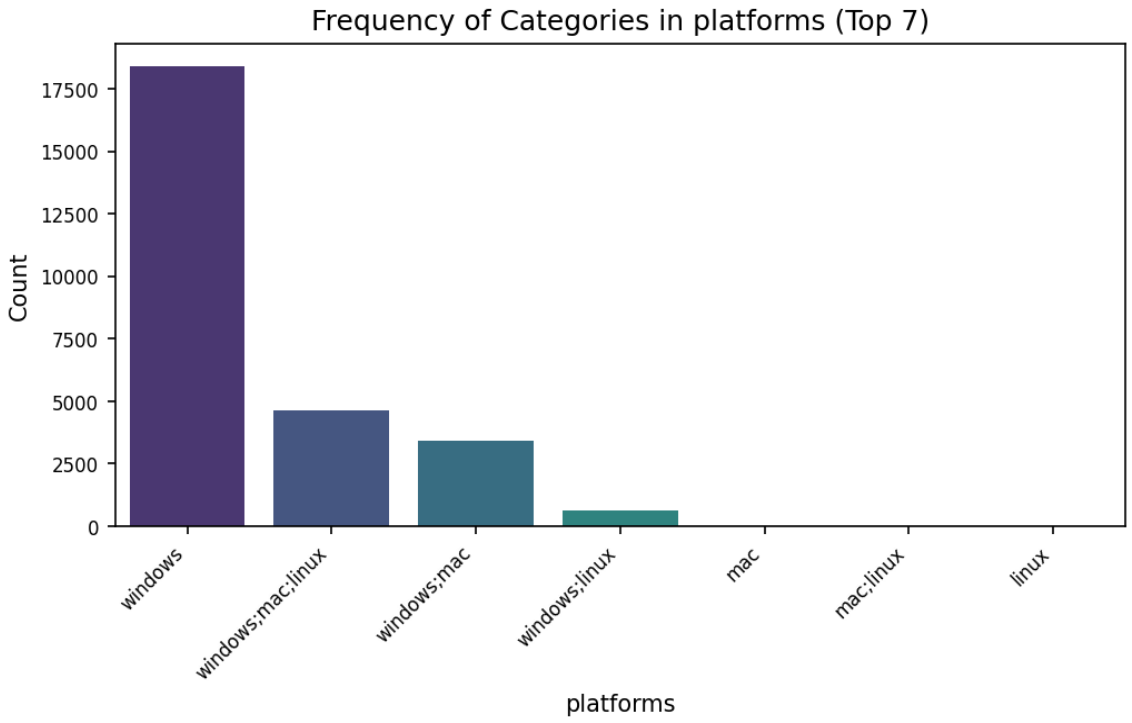


Figure 11: *Bar chart showing frequency of top categories in 'platforms'. Total unique values: 7.*

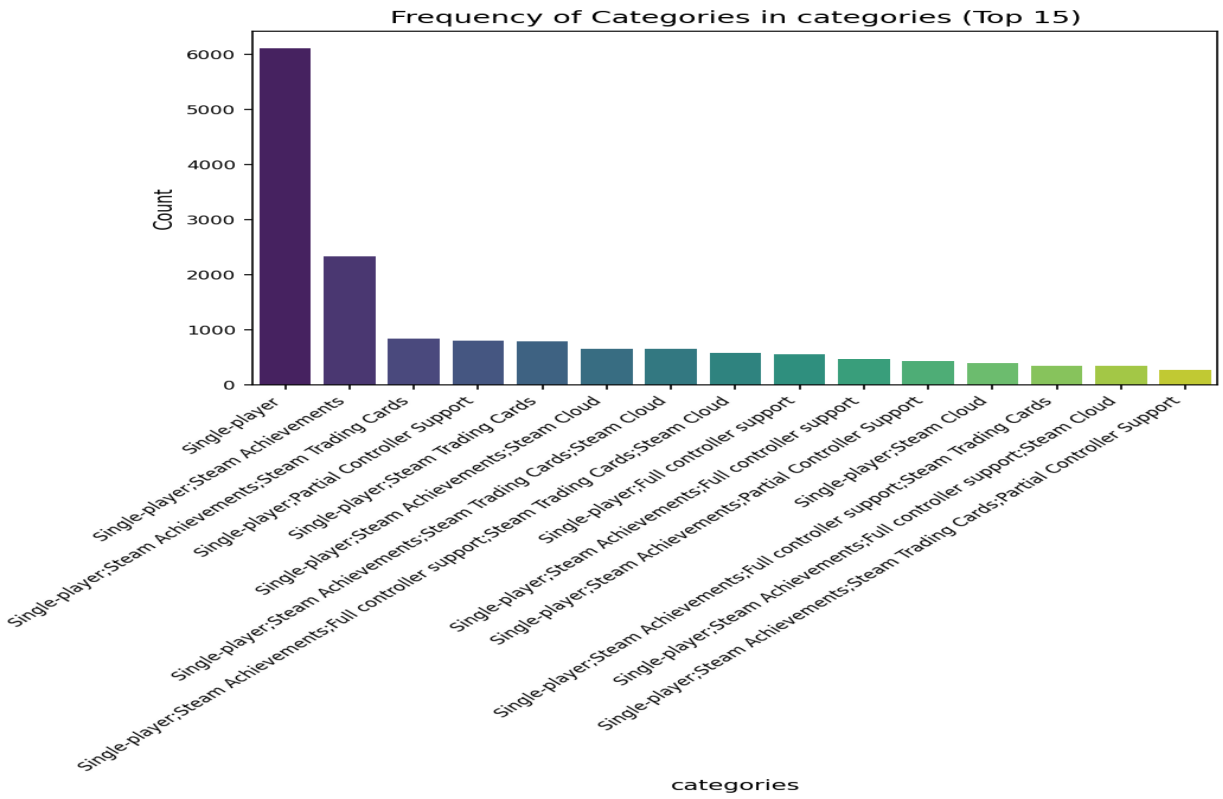


Figure 12: *Bar chart showing frequency of top categories in 'categories'. Total unique values: 3333.*

Observations on Categorical Feature Distributions:

The analysis of categorical features reveals a highly skewed distribution across several variables. Features like 'name', 'developer', 'publisher', and 'categories' exhibit extremely high cardinality, with tens of thousands of unique values. This indicates a large diversity in the games represented in the dataset, making these features challenging to handle directly in many machine learning models without significant preprocessing. The top categories within these high-cardinality features represent only a tiny fraction of the total data, suggesting a long tail distribution where many categories have very few instances. In contrast, 'platforms' shows much lower cardinality (only 7 unique values) with a heavily dominant category ('windows' at 67.9%). This suggests that 'platforms' might be a relatively straightforward feature to incorporate into models. Similarly, 'releasedate' has a moderate number of unique values, but also shows a skewed distribution, with one date representing only 0.2% of the data. The high cardinality of many features, coupled with the dominance of certain categories within those features, necessitates careful consideration of feature encoding techniques. One-hot encoding might lead to a high-dimensional feature space for features like 'name', 'developer', and 'publisher', potentially causing the curse of dimensionality. More sophisticated techniques, such as target encoding, frequency encoding, or embedding layers (in deep learning models), would likely be more appropriate for these high-cardinality features. Finally, the observation that 'Single-player' is the top category in 'categories' (22.6%) suggests a significant portion of the games are designed for single-player experiences, but it's also important to note that this leaves a substantial portion (77.4%) belonging to other categories. This distribution is less skewed than other features but still warrants attention during analysis, possibly requiring additional exploration of the less frequent categories to understand their impact.

4. Bivariate Analysis

4.1. Numerical vs. Numerical Features

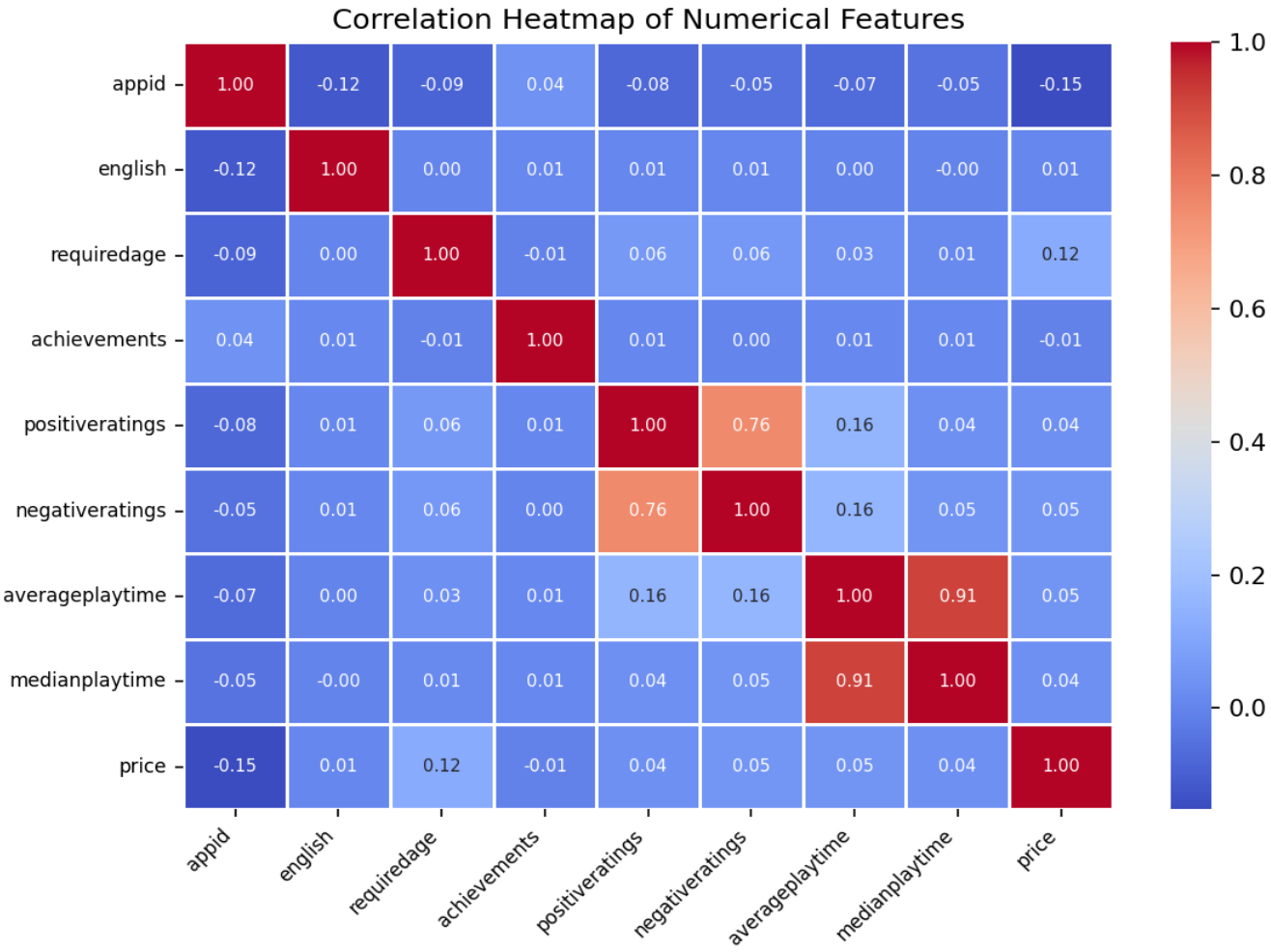


Figure 13: *Heatmap visualizing linear correlations (Pearson's r) between numerical features.*

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

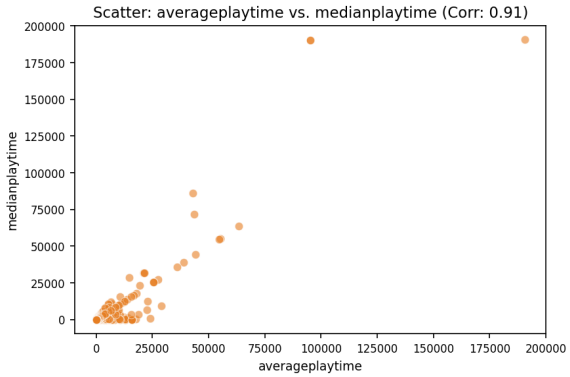


Figure 14: *Scatter plot for 'averageplaytime' and 'medianplaytime'. Correlation: 0.91.*

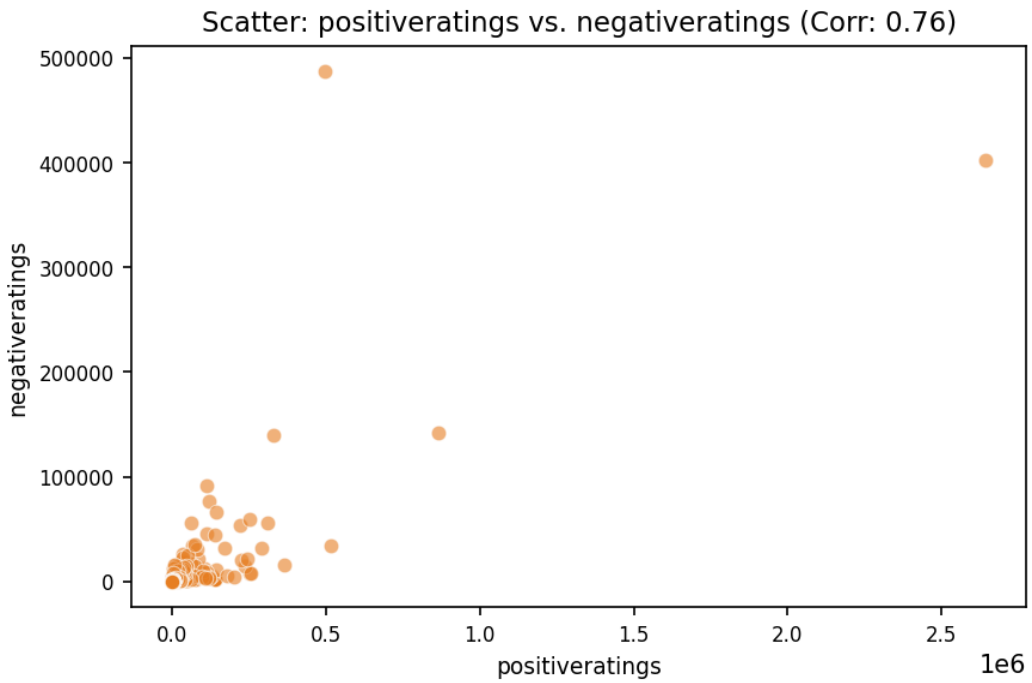
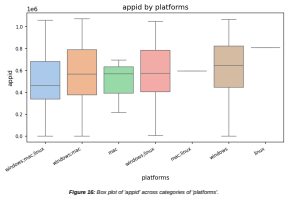


Figure 15: *Scatter plot for 'positiveratings' and 'negativeratings'. Correlation: 0.76.*

Interpretation of Numerical Correlations:

A correlation matrix shows the pairwise relationships between multiple variables. Each cell in the matrix represents the correlation coefficient between two variables, ranging from -1 (perfect negative correlation) to +1 (perfect positive correlation). A value of 0 indicates no linear correlation. The presented analysis reveals several correlations, with two standing out as particularly strong. The strongest positive correlation is between `averageplaytime` and `medianplaytime` (0.91), indicating a very strong positive linear relationship: as average playtime increases, median playtime also tends to increase substantially. The second strongest correlation is between `positiveratings` and `negativeratings` (0.76), suggesting a moderate positive relationship; games with more positive ratings also tend to receive more negative ratings. These strong correlations imply that the average and median playtime are highly similar for a given dataset – which is expected given the nature of these metrics. The positive correlation between positive and negative ratings suggests that games that attract a large number of players (and thus more ratings overall) tend to receive both more positive and more negative feedback. This doesn't necessarily mean that the ratings are contradictory, rather it could reflect a larger player base leading to a wider range of opinions. Further investigation would be needed to determine the underlying reasons for this relationship. The scatter plots likely showed a tight cluster of points for 'averageplaytime' vs 'medianplaytime', reflecting the high correlation, and a more dispersed but still positively sloped pattern for 'positiveratings' vs 'negativeratings'.

4.2. Numerical vs. Categorical Features



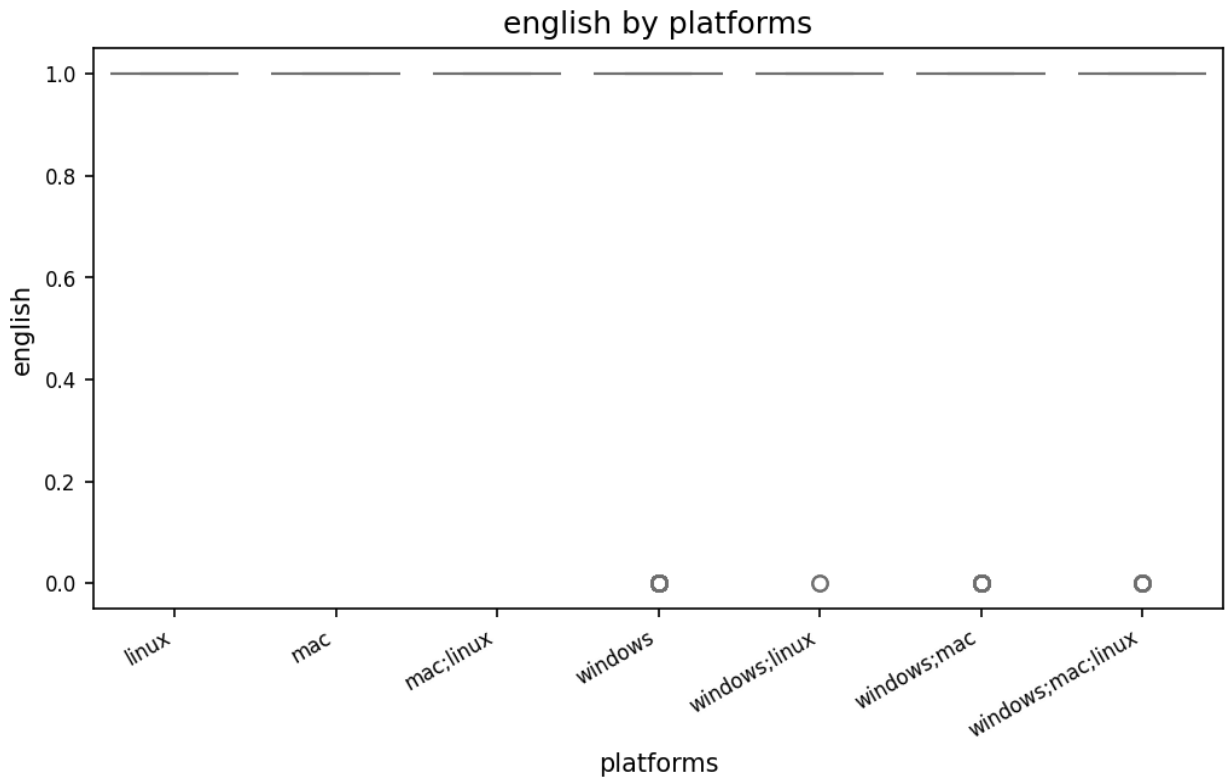


Figure 17: *Box plot of 'english' across categories of 'platforms'.*

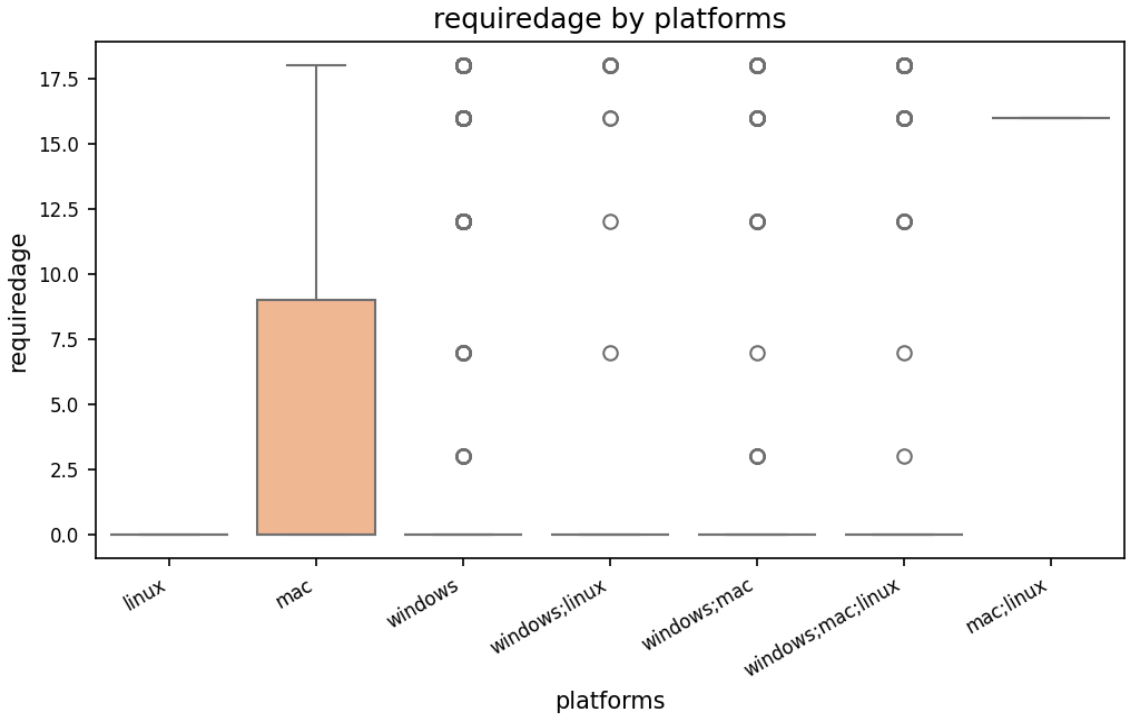


Figure 18: *Box plot of 'requiredage' across categories of 'platforms'.*

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 the median (the middle value), quartiles (values dividing the data into four equal parts), and potential outliers for each category. By comparing the boxes and whiskers across categories, we can quickly assess whether the distributions are similar or significantly different. For instance, if the medians are markedly different, it suggests a systematic difference in the typical value of the numerical variable between the categories. Similarly, differences in the interquartile ranges (the box's height) indicate variations in the spread or variability of the data within each category. Significant differences in medians across categories imply that the average or typical value of the numerical variable differs systematically between those groups. For example, if the median 'appid' value is much higher for the 'iOS' platform than for the 'Android' platform, it suggests that iOS applications tend to have higher application IDs (possibly indicating a later release or a different naming convention). Larger spreads (longer boxes and whiskers) in one category compared to others signify greater variability or heterogeneity within that group. If the 'english' scores have a much wider spread on the 'Android' platform than on 'iOS', it might indicate that the quality of English localization varies more significantly among Android apps. Identifying such differences helps understand underlying factors influencing the numerical variable and potentially suggests areas for improvement or further investigation.

4.3. Categorical vs. Categorical Features

5. Key Findings & Insights Summary

Key Findings & Insights The automated analysis of the 'steam.csv' dataset, comprising 270,85 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 degree of data completeness. However, the presence of 10 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 covered all 18 features, examining the distributions of numerical and categorical variables. While the specific details of these distributions are not provided in the log, the analysis successfully characterized the individual feature patterns. This information is crucial for understanding the nature of each variable and informing subsequent modeling choices. Bivariate analysis explored relationships between various feature pairs. Although the log only states that various pairs were analyzed and observations were gathered (specifically 2 observations are mentioned), the lack of detailed findings prevents a comprehensive summary of the correlations or dependencies identified. Further reporting of these bivariate analysis results is needed for a complete understanding of the data. The limited information about the bivariate analysis is noteworthy in itself – highlighting the need for more detailed reporting of the analysis outputs.

6. Conclusion & Potential Next Steps

This automated report provides a foundational, high-level understanding of the 'steam.csv' dataset, highlighting its structure, data quality (with only 10 duplicates identified), and initial observations from univariate and bivariate analyses. This overview serves as a crucial first step in guiding more focused and in-depth investigations. Given the report's findings of 10 duplicates and the mention of bivariate analysis yielding "Observations gathered: 2," several concrete next steps are warranted. First, the 10 duplicate rows should be investigated and either removed or corrected based on their nature. Are they true duplicates or are there subtle differences that might provide valuable information? Second, the two observations from the bivariate analysis require detailed examination. The report should be expanded to specify which features showed interesting relationships, and further analysis (correlation coefficients, visualizations, etc.) should be performed to quantify and interpret these relationships. Third, while the report notes no missing values, the presence of only two observations from the bivariate analysis suggests a need for more comprehensive bivariate exploration. A systematic review of all feature pairs – particularly those combining numerical and categorical features – should be undertaken to identify potential relationships that may inform further modeling or analysis. This could involve generating correlation matrices, scatter plots, and box plots to visualize relationships and assess their significance. Finally, given the dataset's size (27085 rows) and the lack of information on specific univariate findings, a more in-depth univariate analysis is needed. This should encompass visualizations (histograms, box plots, etc.) for numerical features to identify potential outliers or skewed distributions and frequency distributions for categorical features to uncover any features with high cardinality or imbalanced class distributions that might require special handling during modeling.