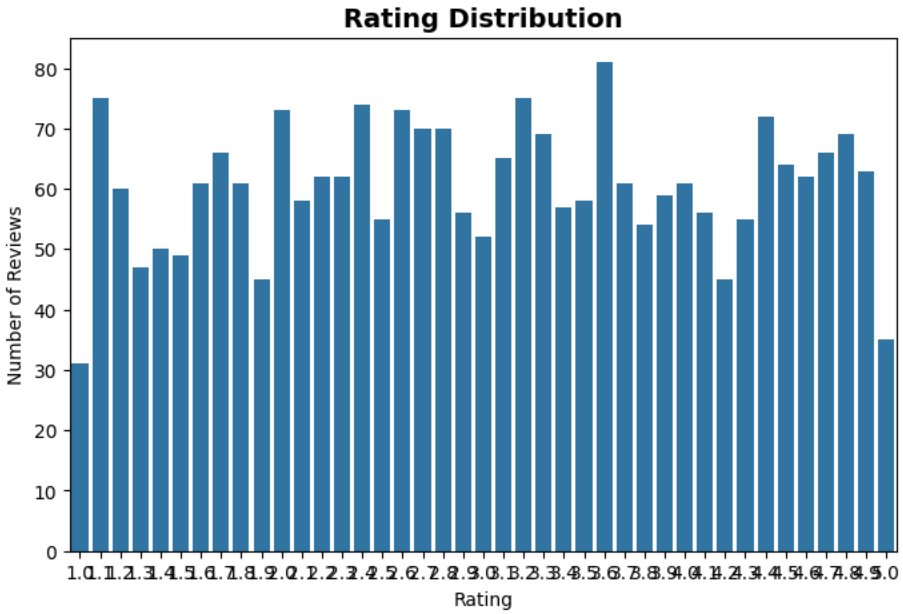
Benjamin Nguyen

Applied Data Science Project

Exploratory Data Analysis (EDA)

Overview

This exploratory data analysis (EDA) investigates user reviews from an application dataset. The analysis incorporates descriptive statistics and visualization to uncover underlying patterns, highlight trends, and provide insight into user behavior across different variables such as ratings, languages, purchase verification, app categories, device types, gender, helpful votes, and review volume over time.

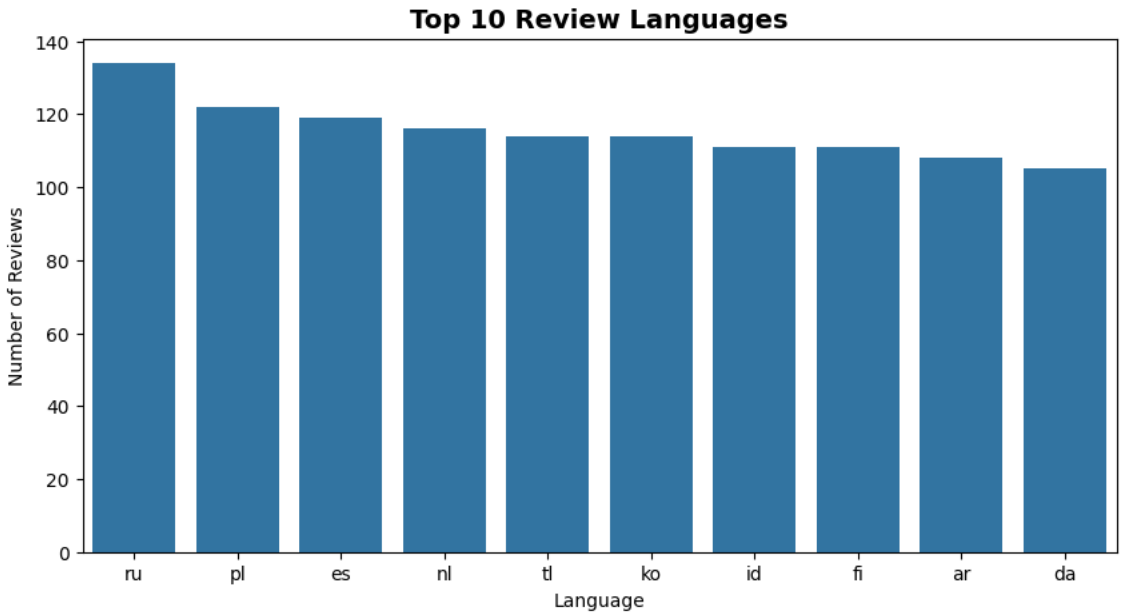


Rating Distribution

The distribution of ratings spans from 1 to 5 stars. The histogram indicates that reviews are relatively balanced across ratings, though there are slight variations. Peaks appear around higher ratings, suggesting that while users report dissatisfaction, the overall sentiment leans toward moderately positive experiences.

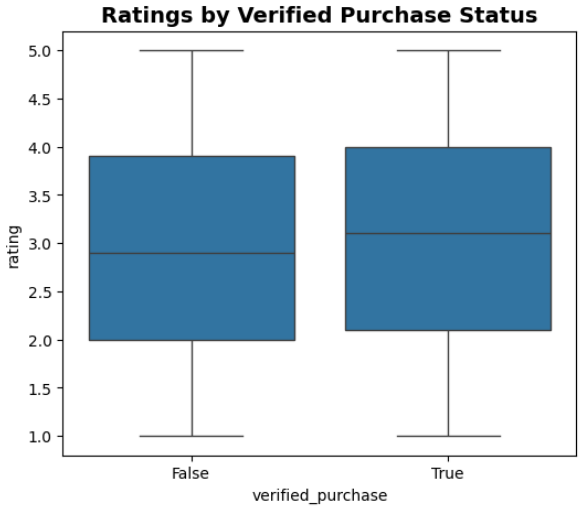
Top 10 Review Languages

Analysis of the top review languages reveals that Russian (ru), Polish (pl), and Spanish (es) dominate the dataset. This indicates that the application has a strong international user base, with significant adoption across multiple regions.



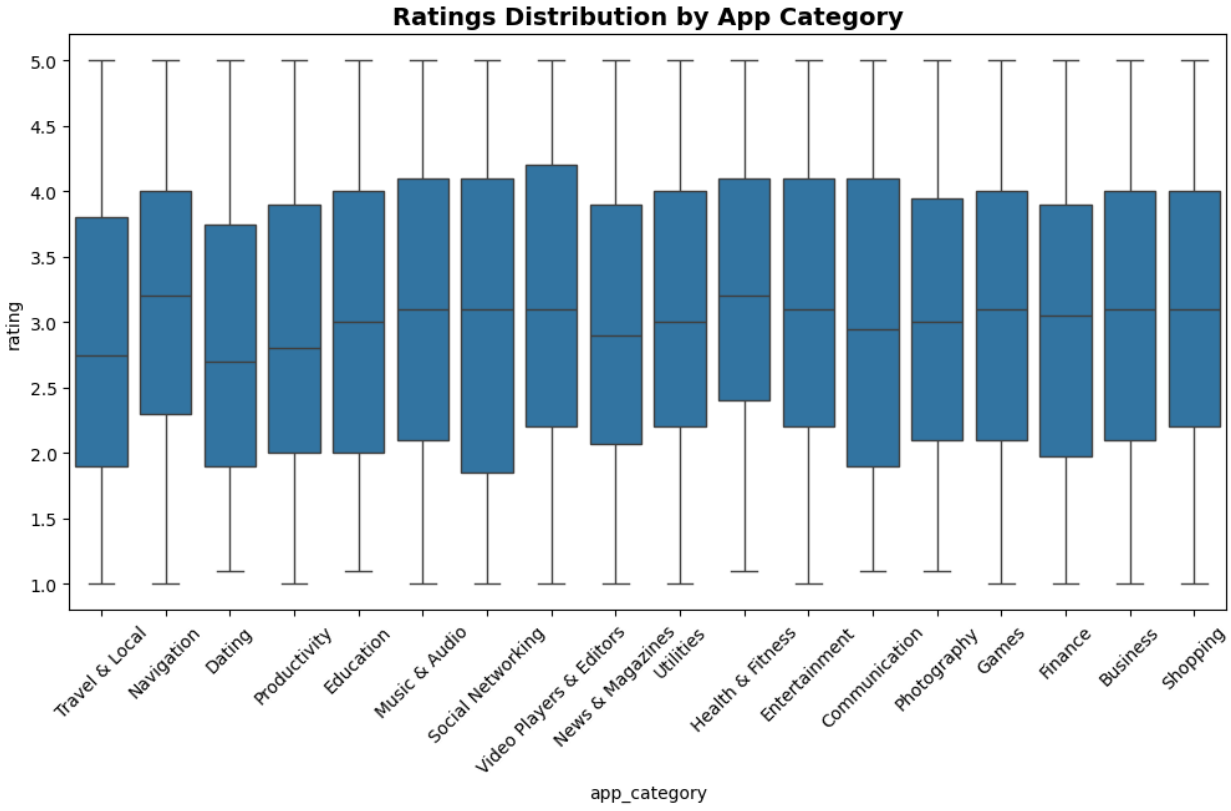
Ratings by Verified Purchase Status

The boxplot comparison between verified and non-verified purchases shows similar distributions, though verified purchases exhibit slightly higher median ratings. This may suggest that verified users are more satisfied with their experience, or that unverified reviews include more critical perspectives.



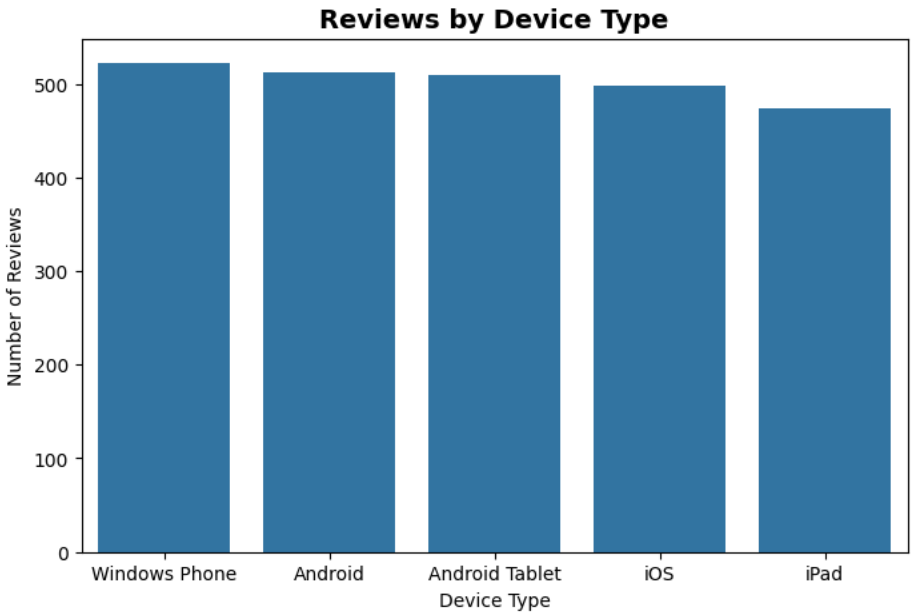
Ratings Distribution by App Category

When comparing ratings across app categories, the median remains consistent, hovering around 3 to 4 stars. Certain categories such as Health & Fitness and Navigation show slightly higher medians, whereas others like Dating and Productivity have wider variability, reflecting differing user satisfaction across app types.



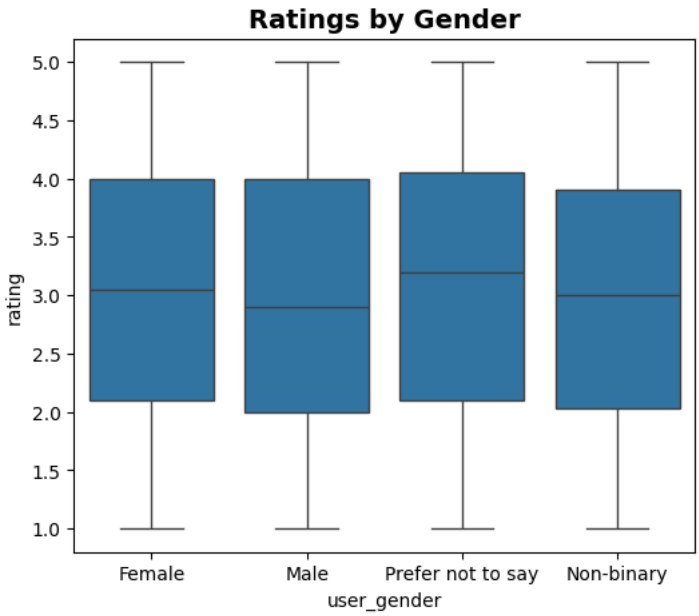
Reviews by Device Type

The bar chart shows reviews distributed across multiple device types, with Windows Phone, Android, and iOS devices receiving the majority of reviews. The relatively even distribution indicates that the application is widely used across platforms, supporting device compatibility.



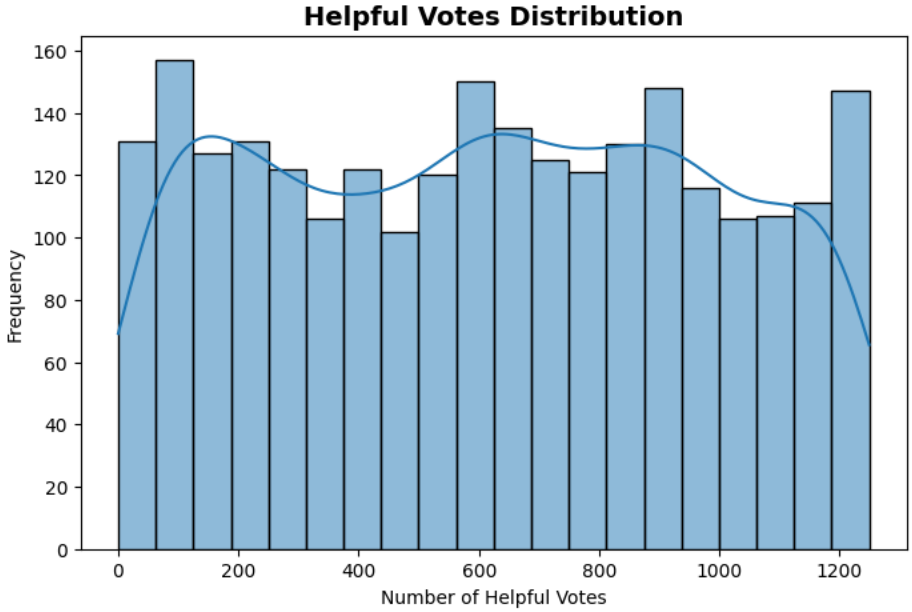
Ratings by Gender

Ratings grouped by gender (Female, Male, Non-binary, and Prefer not to say) show similar central tendencies. The medians range from 2.8 to 3.2 across groups, indicating that gender has limited impact on user perception. The consistency suggests that feedback is broadly representative across demographics.



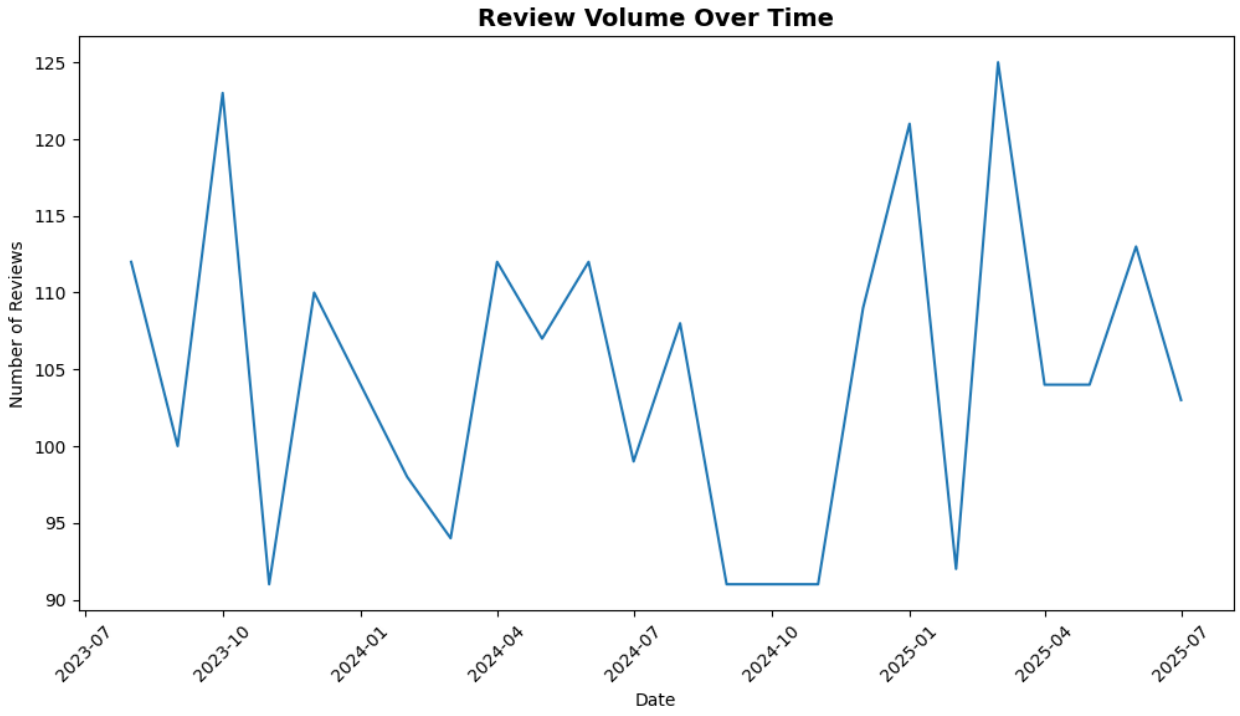
Helpful Votes Distribution

The histogram with density overlay demonstrates that helpful votes range widely, with peaks both at lower and higher values. This suggests a diverse engagement pattern: while many reviews receive only a few helpful votes, some garner significant attention from users, marking them as particularly valuable contributions.



Review Volume Over Time

The time series plot of review volume highlights fluctuations in user engagement over the months. Periodic peaks suggest cycles of user activity, potentially corresponding to app updates, marketing campaigns, or seasonal trends. This temporal analysis underscores the importance of context when evaluating user feedback.



Conclusion

The EDA highlights several key insights: ratings tend toward moderately positive values, usage spans multiple languages and device types, and review engagement varies over time. Verified purchases are generally associated with slightly higher ratings, while app category and gender show minor differences. These findings provide a strong foundation for further modeling and hypothesis testing in subsequent phases of the project.