Exploratory Data Analysis on the Google Play Store Dataset

Emre Demirbaş

Contents

1 Introduction

This report presents an Exploratory Data Analysis (EDA) of the Google Play Store dataset. The aim is to understand how app characteristics—such as category, price, content rating, size, and popularity—relate to user satisfaction and market presence.

Dataset summary. The dataset contains app-level information (e.g., name, category, size, rating, reviews, installs, price, content rating). Prior to analysis, data cleaning steps were applied: removing duplicates, handling missing values, converting string-formatted numerics (e.g., "1,000+" installs, "19M" size) to numeric types, and standardizing categories.

2 Numerical Distributions

2.1 Ratings

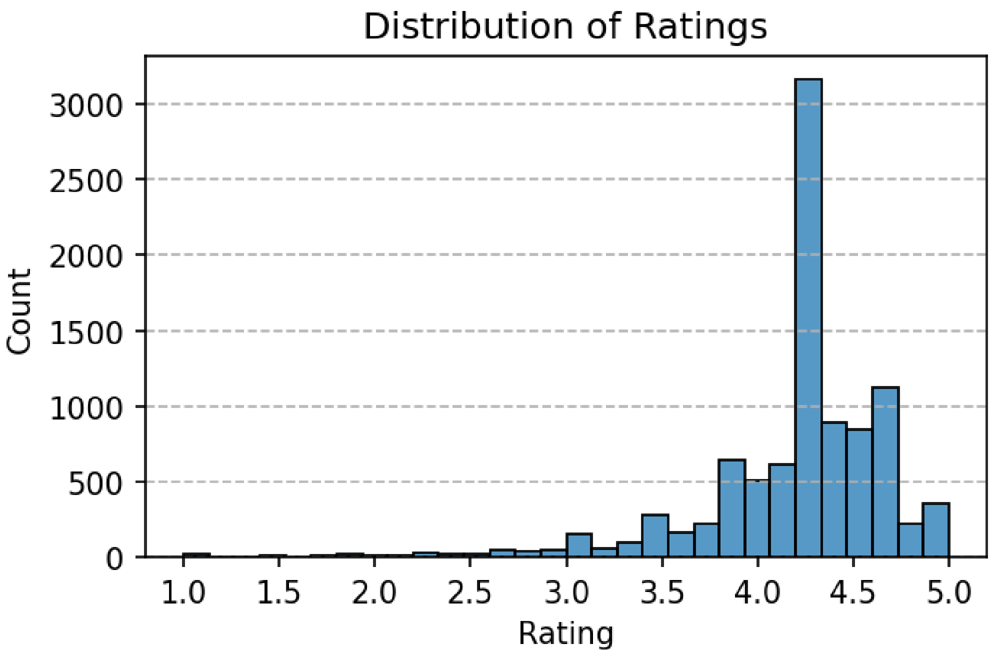

Figure 1: Distribution of App Ratings. Most apps are rated between 4.0 and 4.5, showing generally positive user feedback.

Insight: Ratings are right-skewed with a strong peak around 4.3–4.5; few apps fall below 3.0.

2.2 Reviews (filtered to < 1M)

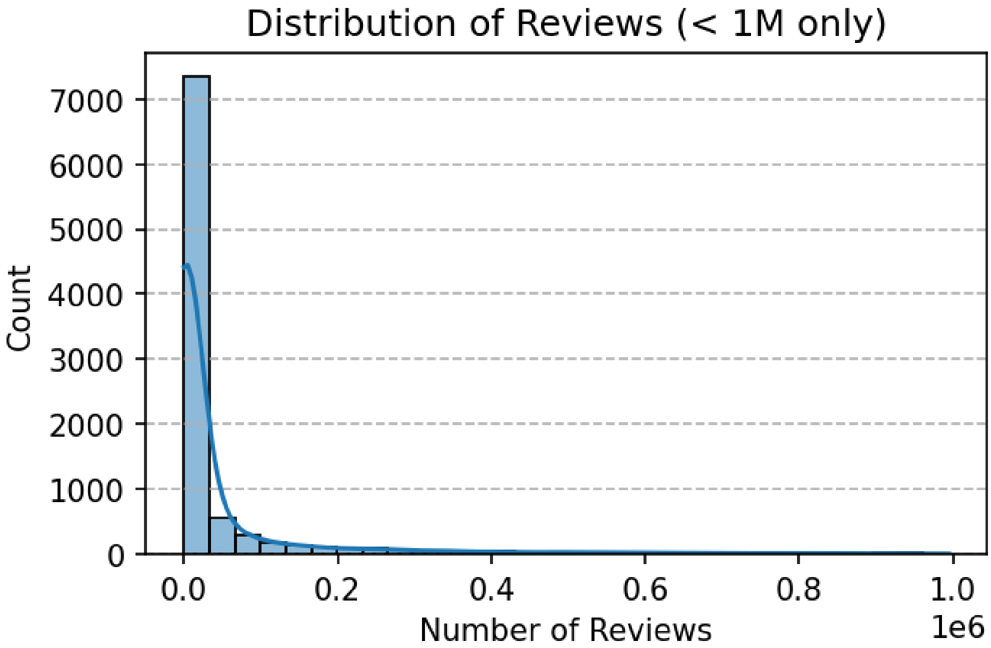

Figure 2: Distribution of Reviews (apps with fewer than one million reviews). The long tail indicates that most apps receive relatively few reviews, while a small group attracts very high engagement.

Insight: Engagement is concentrated in a small number of popular apps.

2.3 App Size (MB)

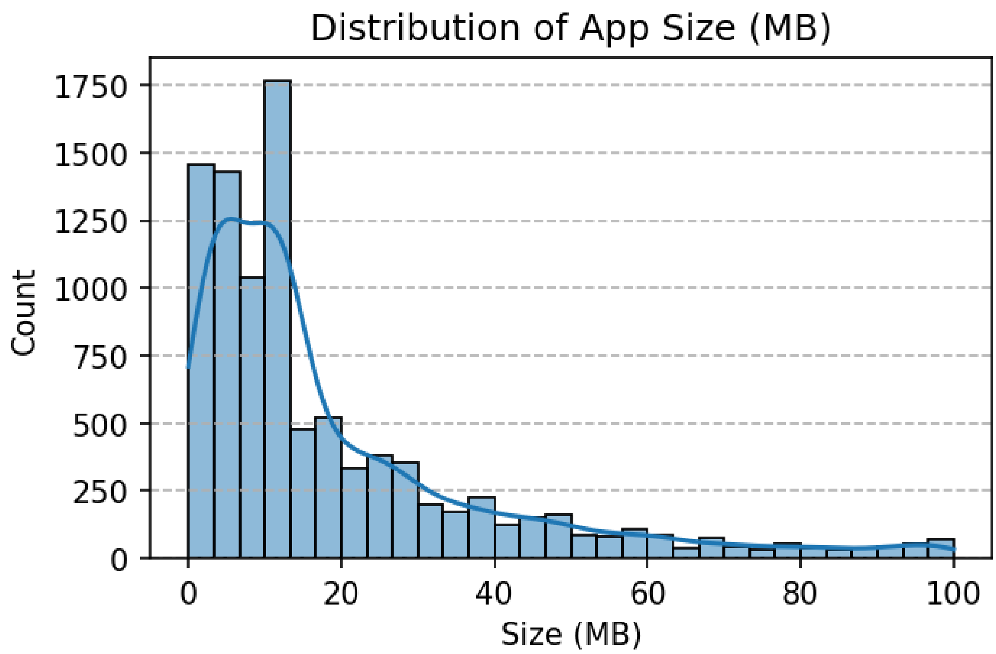

Figure 3: Distribution of App Size in MB. Most apps are below 20 MB, with a long tail toward larger sizes.

Insight: Developers tend to keep apps compact for better downloads and storage use.

2.4 Price (including free apps)

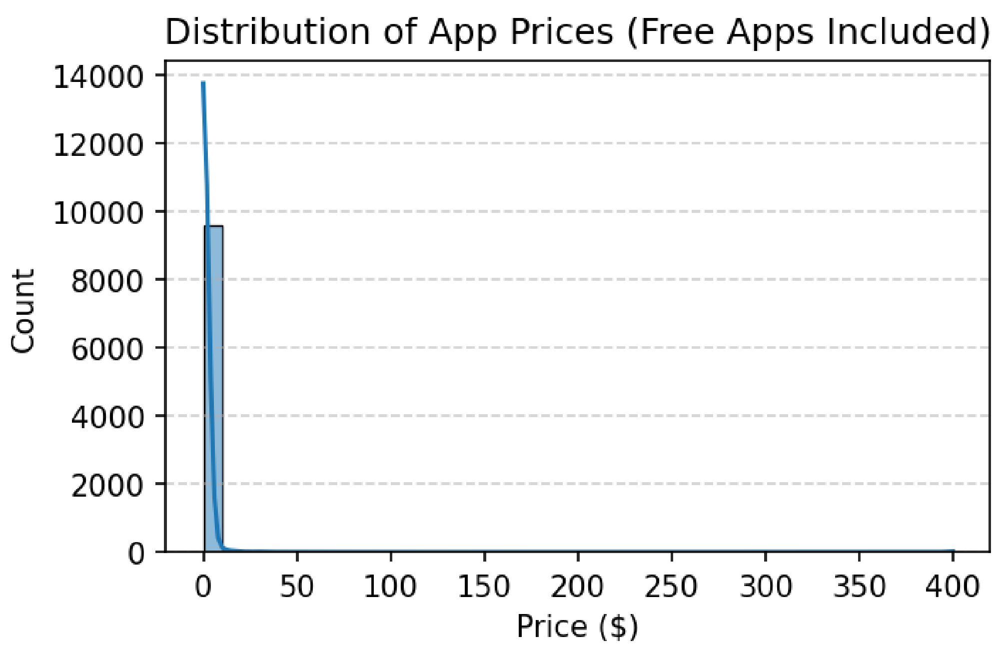

Figure 4: Distribution of App Prices (free apps included). The spike at zero reflects that most apps are free.

Insight: The ecosystem is dominated by free apps.

2.5 Price (paid apps only)

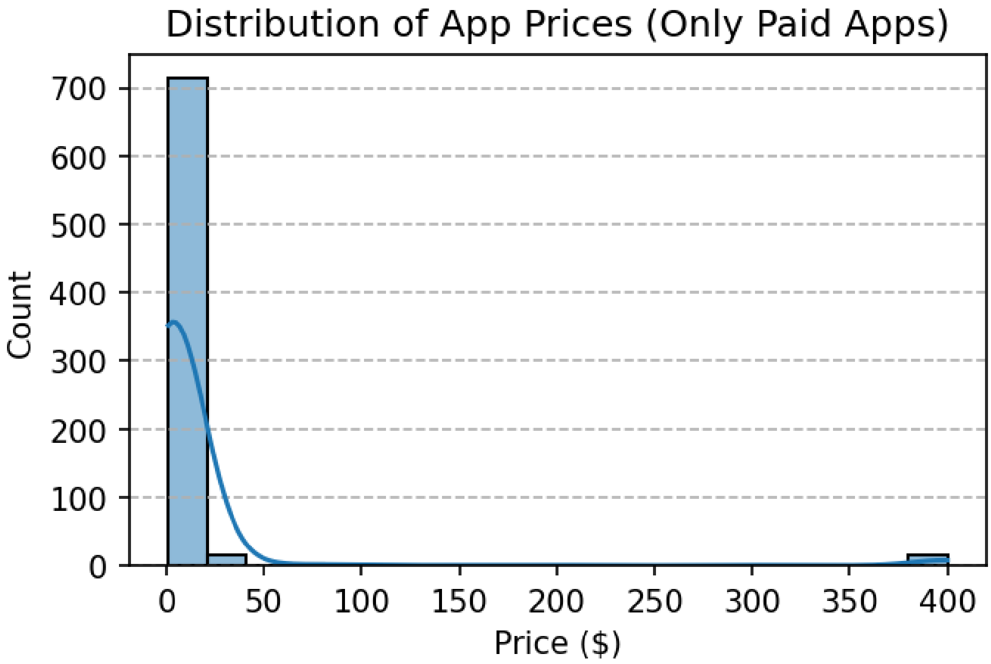

Figure 5: Distribution of App Prices (paid apps only). Most paid apps are under $20, with a small set at premium prices.

Insight: Paid pricing clusters at low values; very high prices are rare and likely niche.

3 Category-Level Analyses

3.1 Top 10 categories by number of apps

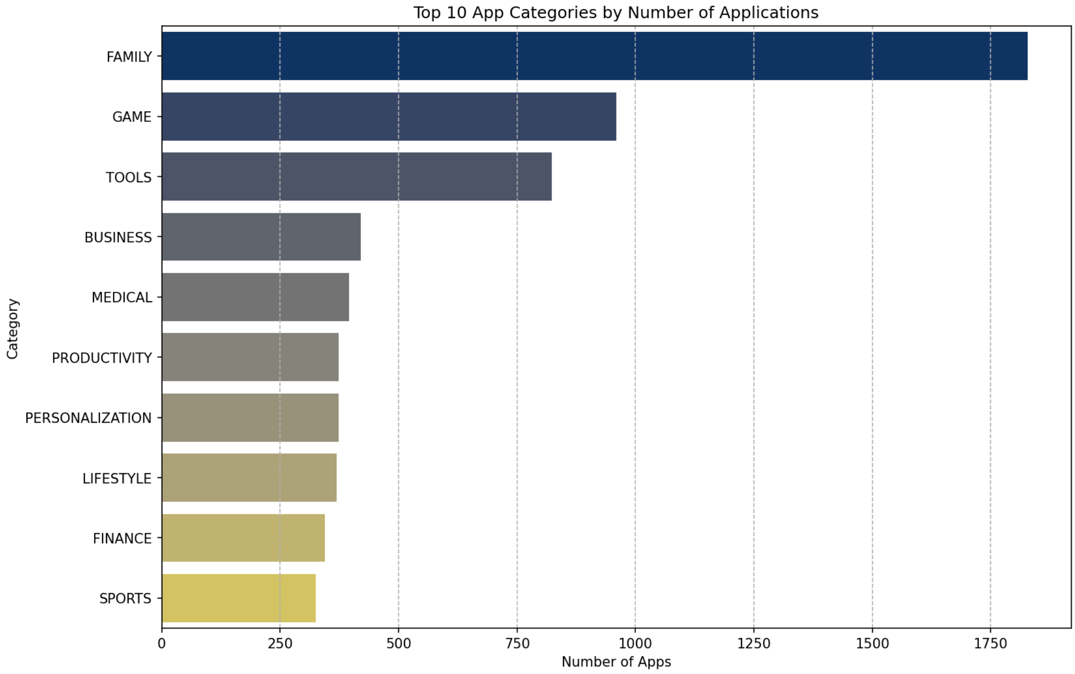

Figure 6: Top 10 App Categories by Number of Applications. Family, Game, and Tools lead the store.

Insight: The top ten categories account for about 64.4% of all apps, showing concentration in entertainment and general-utility segments.

3.2 All categories by app count (full list)

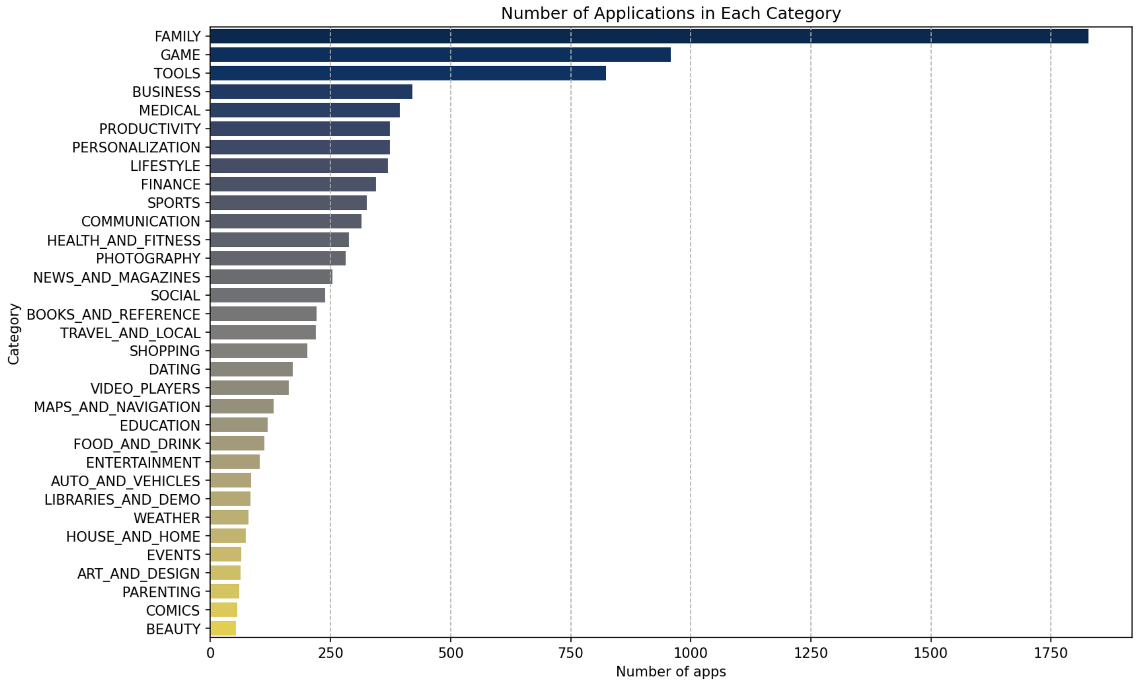

Figure 7: Number of Applications in Each Category (full view).

Insight: Beyond the leaders, counts decrease steadily; many categories are comparatively small.

3.3 Highest total spending by category

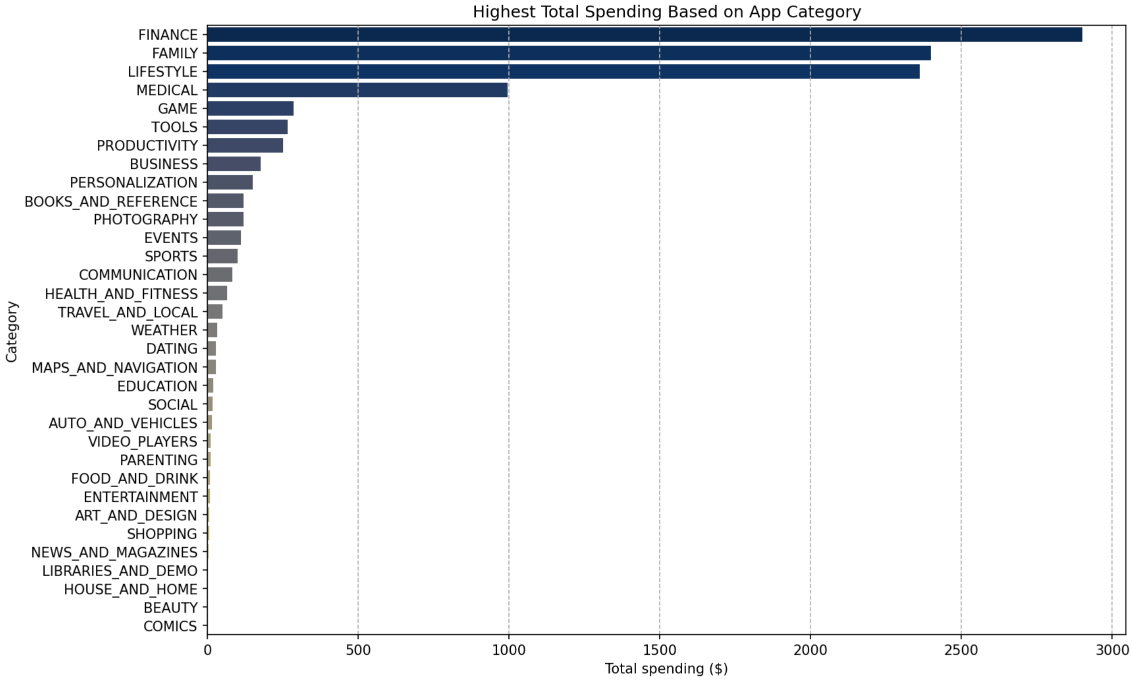

Figure 8: Highest Total Spending Based on App Category. Finance, Family, Lifestyle, and Medical categories account for the largest total spend.

Insight: Spending is not strictly tied to app counts; certain categories monetize better than others.

3.4 Most common words in app names

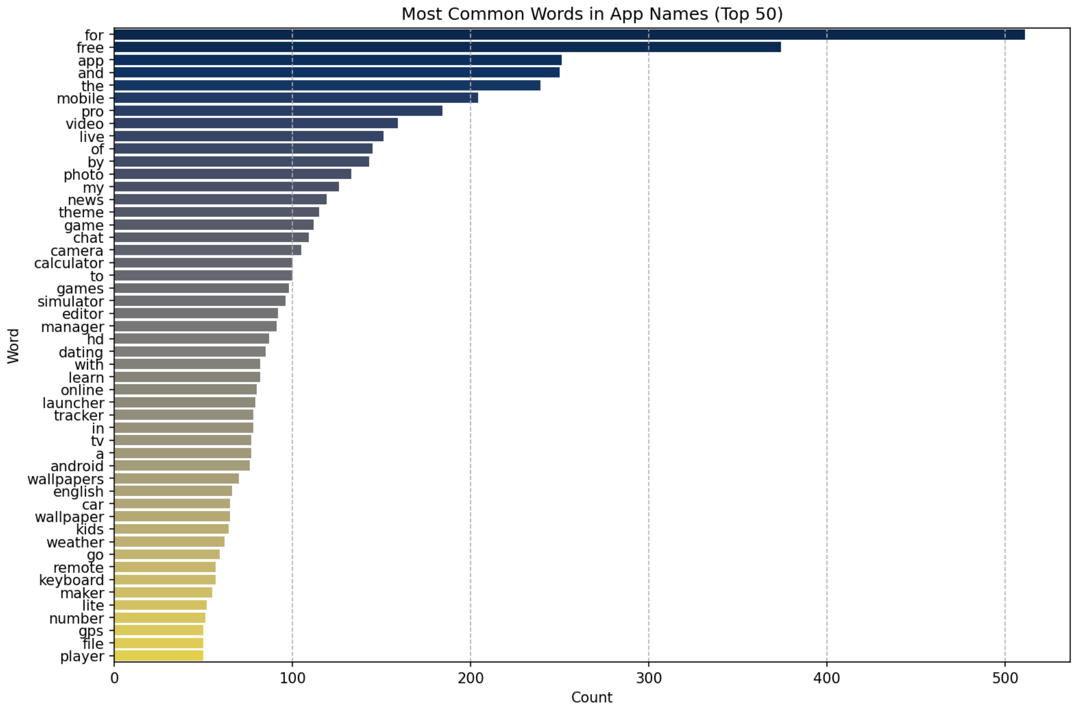

Figure 9: Most Common Words in App Names (Top 50).

Insight: Frequent tokens (e.g., *for*, *free*, *app*, *mobile*) reflect generic branding and feature-oriented naming.

4 Ratings by Type and Content

4.1 Ratings by app type (Free vs. Paid)

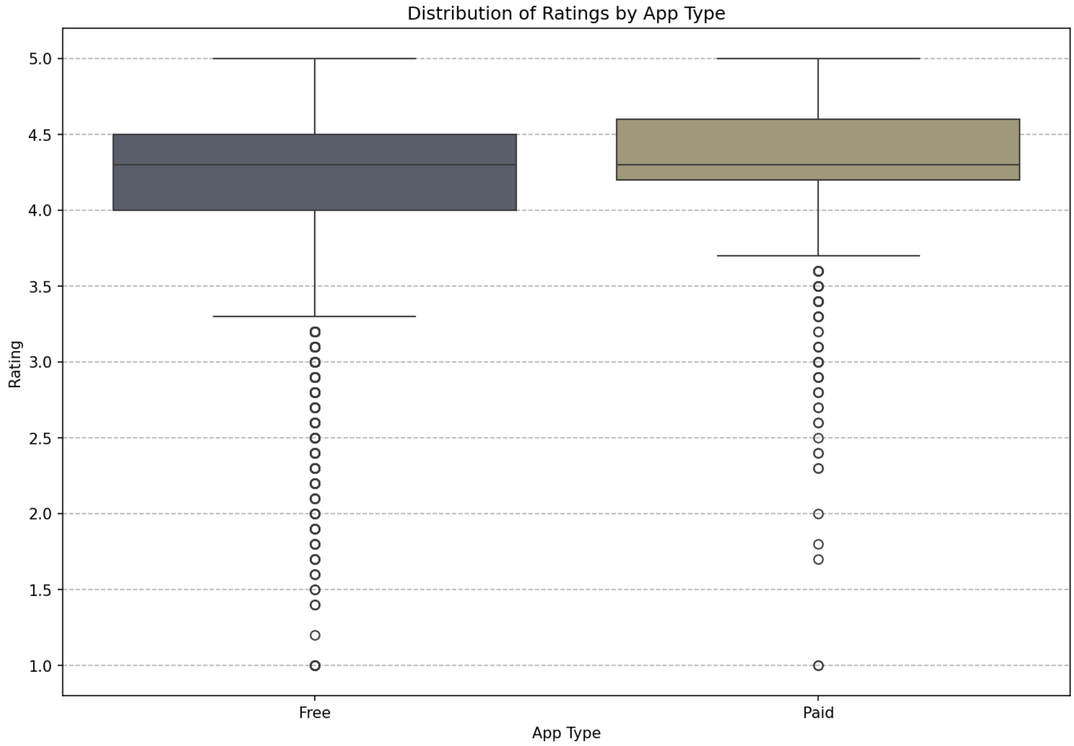

Figure 10: Distribution of Ratings by App Type. Paid apps show slightly higher and more consistent ratings than free apps.

Insight: Free apps have wider variance; paid apps tend to deliver a more polished experience.

4.2 Counts by rating category

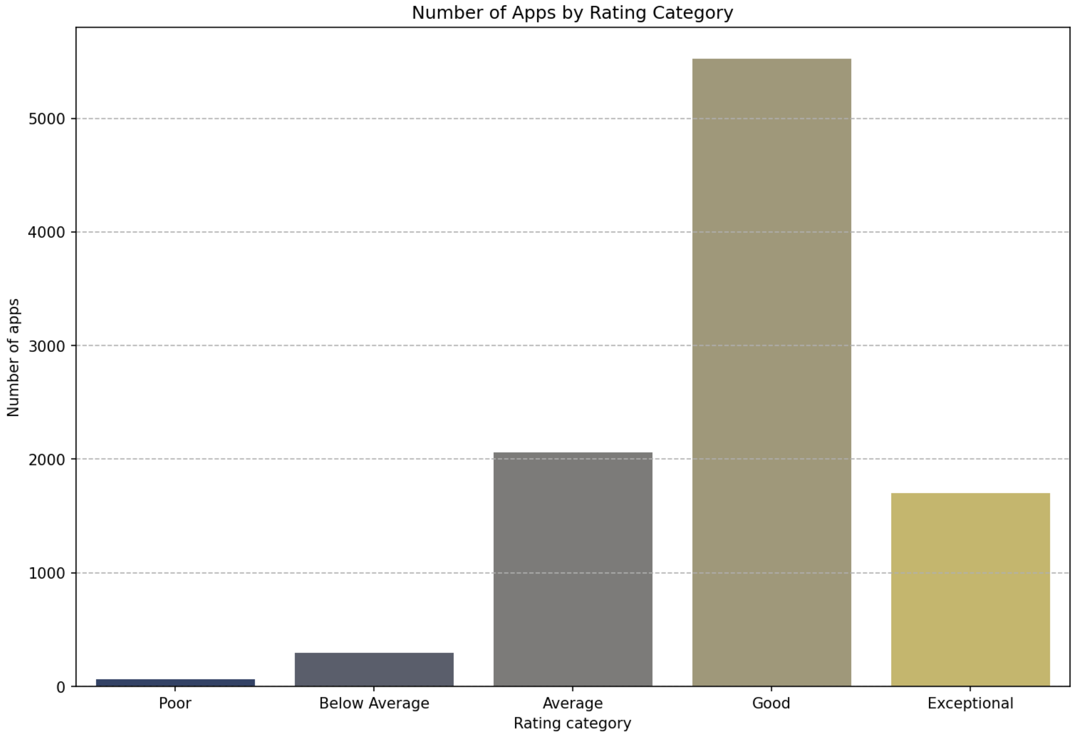

Figure 11: Number of Apps by Rating Category. Most apps are “Good” or “Exceptional”.

Insight: Overall user satisfaction is high across the store.

4.3 Average rating by content rating

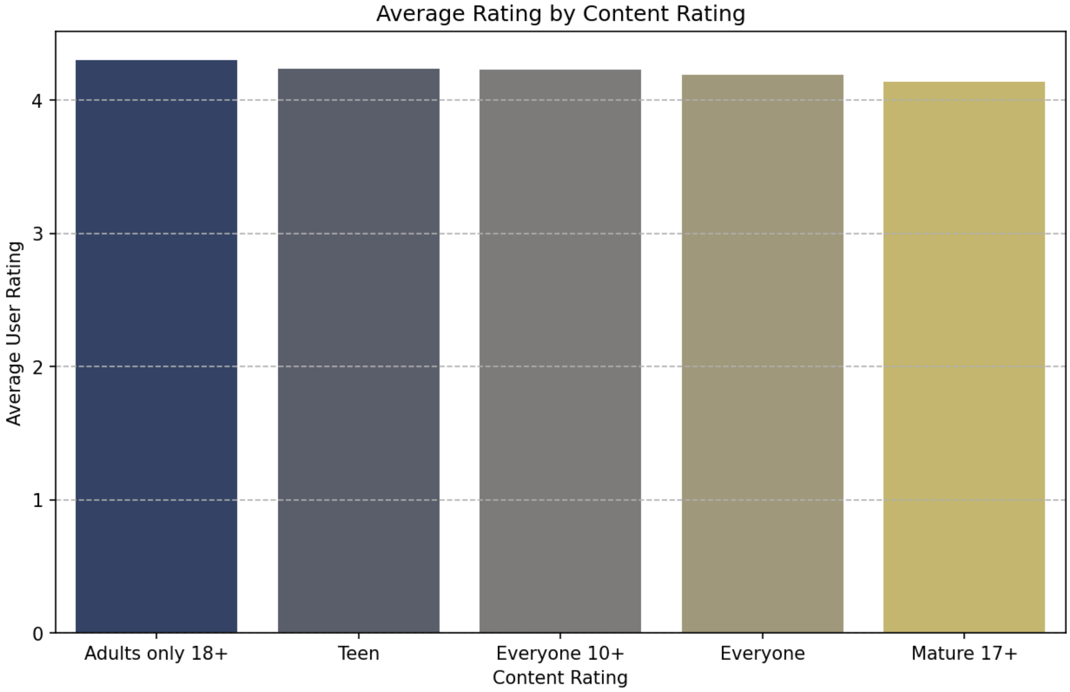

Figure 12: Average Rating by Content Rating. Adults-only apps show the highest mean, while general-audience apps are slightly lower.

Insight: Niche/mature audiences may rate more positively.

4.4 Rating distribution by content rating

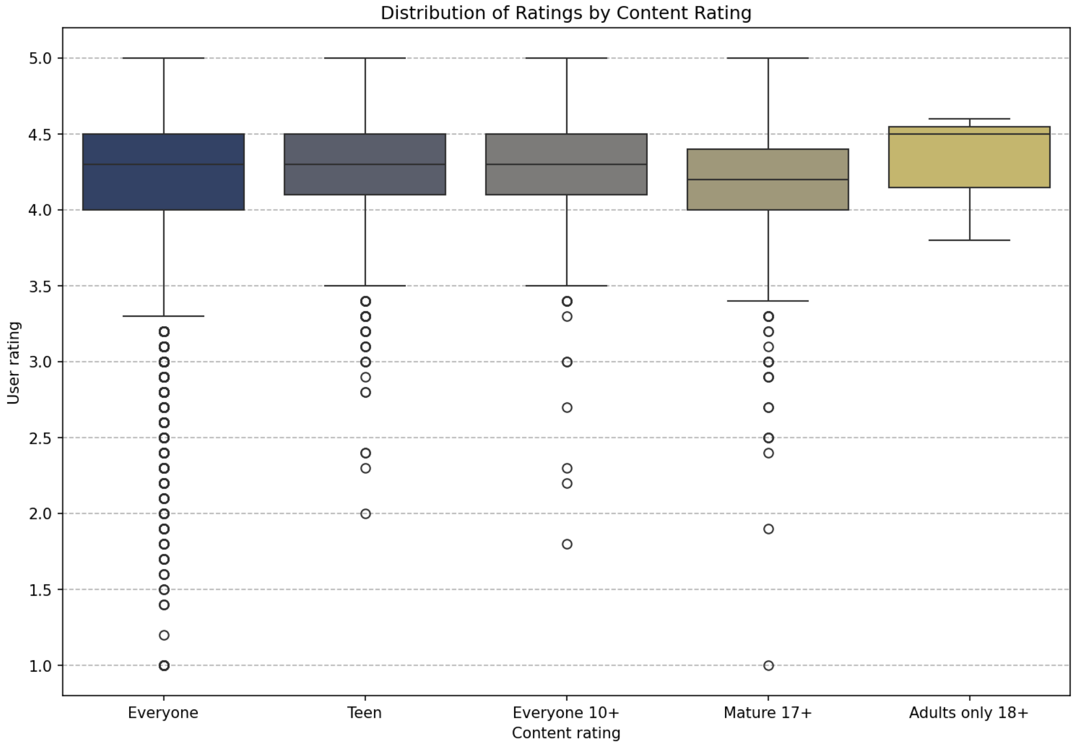

Figure 13: Distribution of Ratings by Content Rating. All content classes generally rate above 4.0 with modest differences in spread.

Insight: Apps for "Everyone" display more variability, likely due to broader appeal and diverse app quality.

5 Relationships and Correlations

5.1 Popularity vs. satisfaction by category

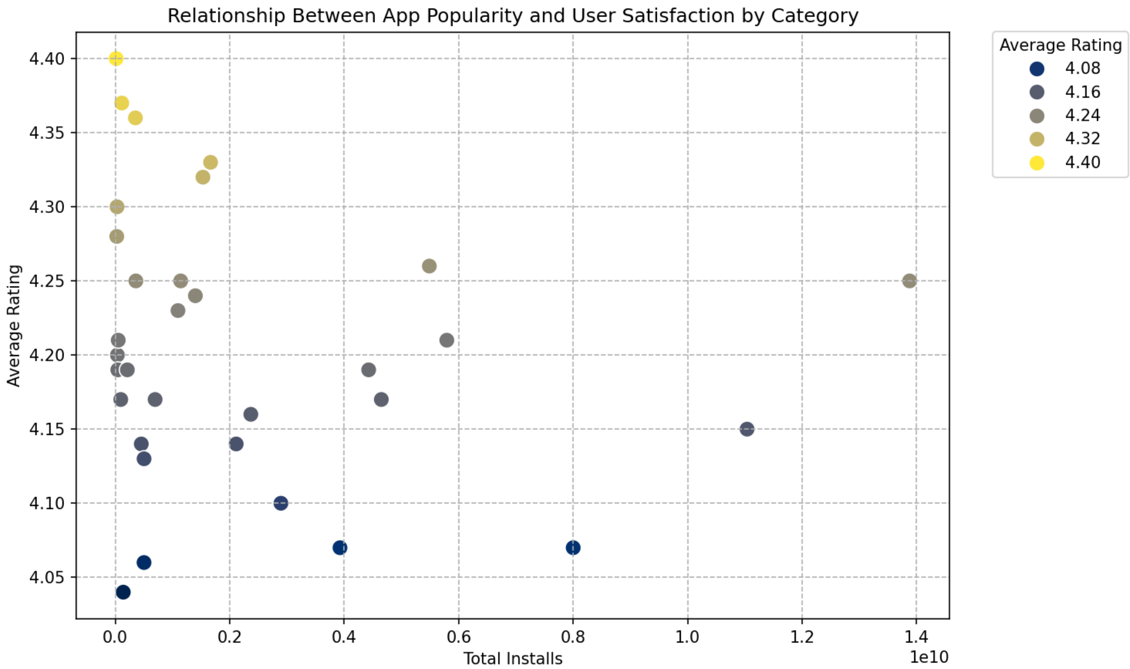

Figure 14: Relationship Between App Popularity and User Satisfaction by Category.

Insight: Categories with low or moderate installs can still achieve high ratings. Popularity does not guarantee better satisfaction.

5.2 Ratings vs. number of reviews

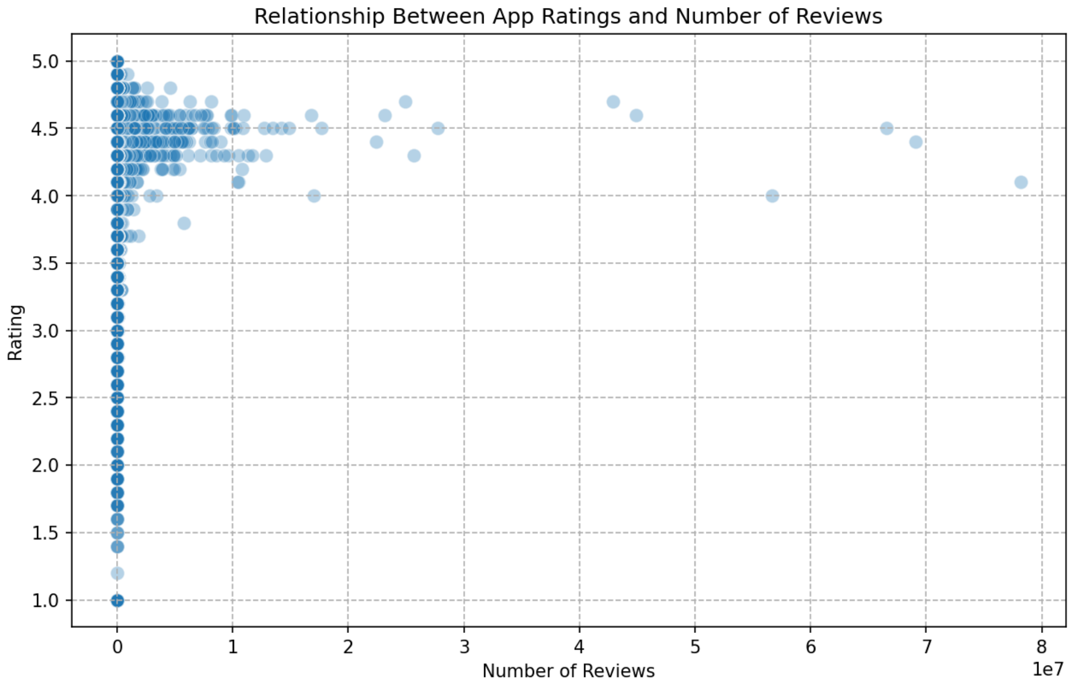

Figure 15: Relationship Between App Ratings and Number of Reviews.

Insight: The correlation is very weak (about 0.05). Apps with many reviews do not necessarily have higher ratings.

5.3 Content ratings vs. categories (heatmap)

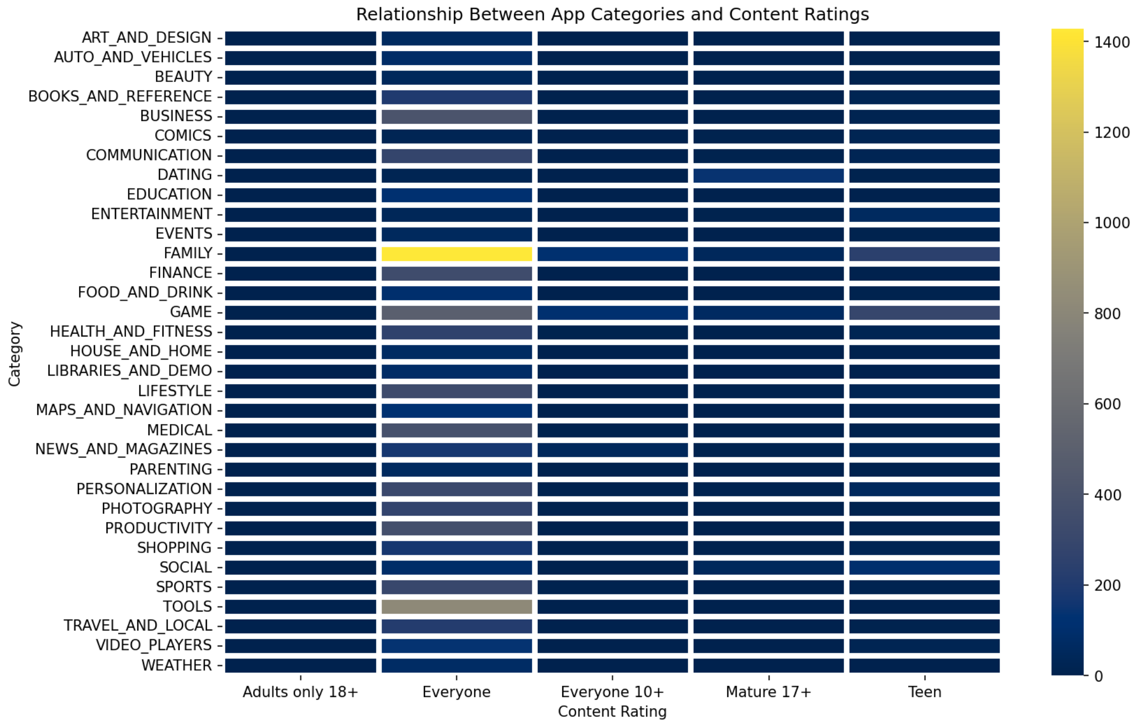

Figure 16: Relationship Between App Categories and Content Ratings. Most categories are rated “Everyone”; Dating and Games include more Teen/Mature apps.

Insight: The store mainly targets general audiences, with a few categories skewing mature.

5.4 Installs by Android version

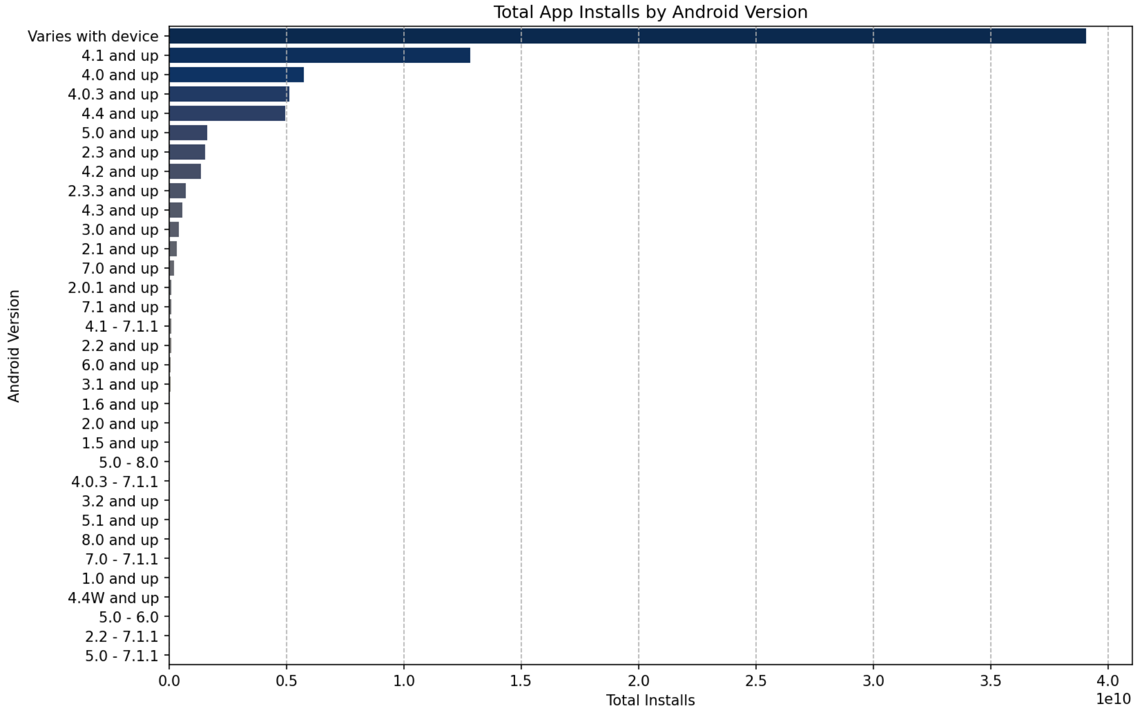

Figure 17: Total App Installs by Android Version. “Varies with device” dominates, followed by older minimum-version targets (e.g., 4.x).

Insight: Many apps support a wide range of versions to maximize reach; few specify only the latest Android versions.

6 Conclusion

The Google Play Store is dominated by free, well-rated apps concentrated in a handful of categories (Family, Game, Tools). Popularity and satisfaction are largely independent: categories or apps with fewer installs can still deliver high quality. Content ratings show that most apps target everyone, with a small share aimed at teens or mature users. Finally, broad Android version support reflects developers’ goal to reach the widest possible audience.