FROM DOWNLOADS TO 5 STARS. THE SCIENCE OF APP VIRALITY

A Deep Dive into Consumer Behavior and Marketplace Dynamics

Presented by Team Carpe Diem

Vidhit T S | Padam Rathi | Md Faisal

Decoding the Digital Ecosystem

Context & Problem Statement

- Smartphones and apps drive modern communication and entertainment. With thousands of apps on the Play Store, understanding what makes an app successful is important.
- Our objective is to perform a data-driven analysis of the Play Store ecosystem and identify:
  - What factors influence app success,
  - How users rate and respond to apps, and
  - How metadata + user sentiment together reflect real usage trends.

Datasets Used

We used two datasets from the Google Play Store environment:

1.Googleplaystore Apps datasets

Metadata of ~10,000 apps (Category, Rating, Installs, Size, Reviews, Price, Type, Genres)

2.Googleplaystore User Reviews

~24,000 user reviews with Sentiment, Polarity, Subjectivity, and translated review text

These datasets allow us to combine app-level features with user opinions for deeper insights.

Exploratory Themes

1. **Market Dominance:** Which categories get the most installs? Does content rating (Everyone/Teen/Mature) affect popularity?
2. **Success Factors:** Do higher ratings lead to more installs? Do reviews influence visibility and trust?
3. **User Satisfaction:** Are Free apps rated better than Paid apps? How do sentiments reflect issues like ads, crashes, or privacy concerns?

Methodology used in the analysis

1. **Preprocessing**
   Cleaning raw datasets by fixing missing values, correcting formats, removing duplicates, and standardizing fields.
2. **Engineering**
   Creating useful features (numeric conversions, one-hot encodings, sentiment fields) to enhance analytical depth.
3. **Exploration**
   Analyzing patterns through visualizations (ratings, installs, categories) to understand app behaviour.
4. **Insights and Inferences**
   Extracting meaningful findings about category performance, user satisfaction, monetization, and sentiment trends.
5. **Prediction**
   Building a simple model (Random Forest) to estimate app success or ratings based on its attributes.

Preprocessing

- **Classified Missing App Type:**
  Filled missing Type values by checking the Price column and labeling apps as Free or Paid.
- **Removed Invalid Content Rating Rows:**
  Dropped the 2 rows where Content Rating was missing since they were too few to impact analysis.
- **Dropped Irrelevant Version Columns:**
  Removed Current Ver and Android Ver as they had inconsistent formats and were not useful for insights.
- **Imputed Missing Ratings:**
  Used a Random Forest Regressor to predict and fill missing Rating values based on other app features.
- **Standardized App Size:**
  Converted all Size values (M, k) into megabytes (MB) for uniform analysis.

Dataset Initially(Missing values)

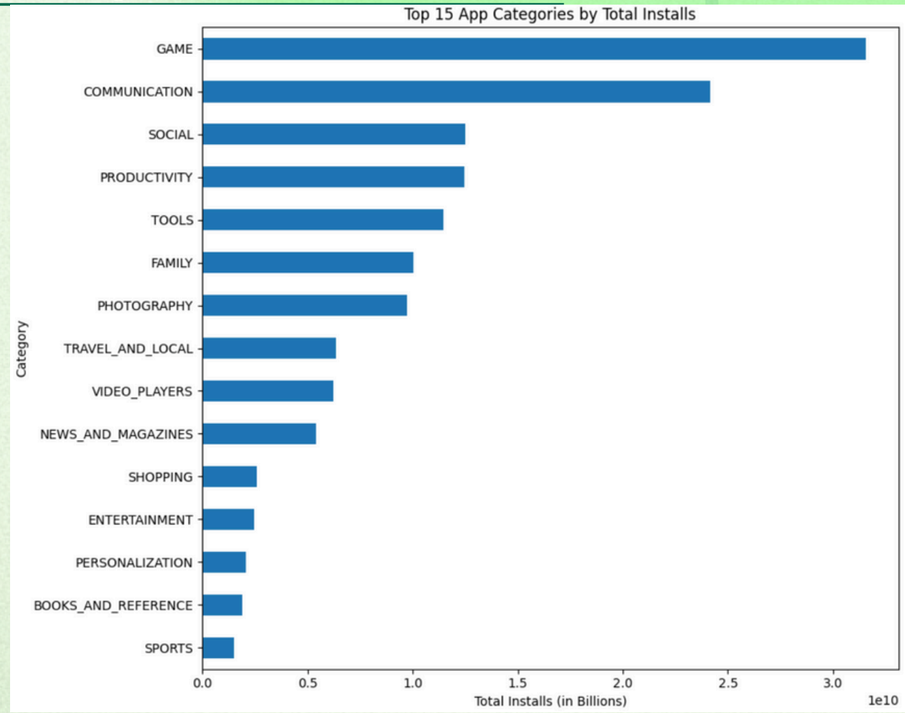
Rating	1474
Type	1
Content Rating	1
Current Ver	8
Android Ver	3

Feature Engineering

1. **Extracted Update Information**: Converted Last Updated into datetime and created new fields: Update Year, Update Month, and Days Since Update.
2. **Created Engagement Metric**: Added an Engagement Rate feature calculated as (Reviews ÷ Installs) to measure user interaction.
3. **Categorized Install Counts**: Binned Installs into Install Tiers (Low → Massive) for better comparison across apps.
4. **Separated Genres**:Split the Genres column into Primary Genre and Secondary Genre for more detailed analysis.

Google playstore Dataset analysis

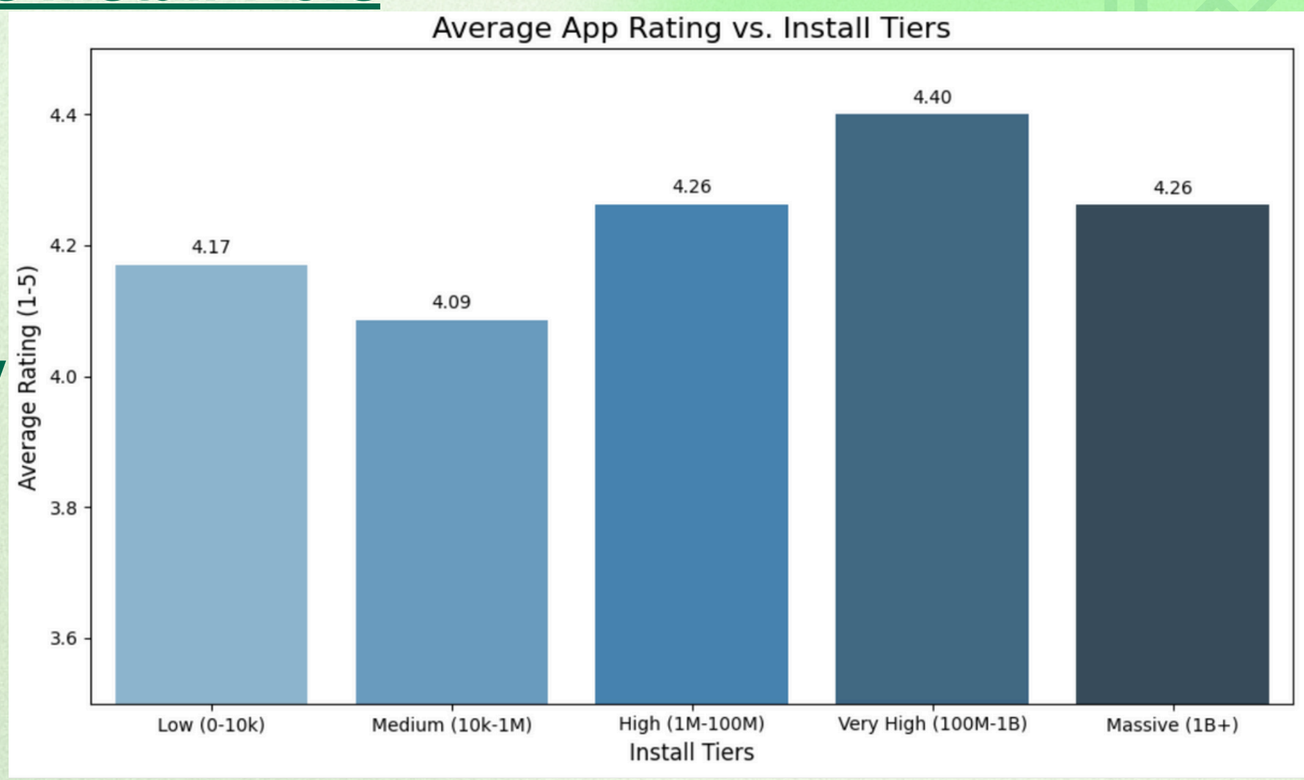
Which app categories lead the Play Store in total installs, and how does their popularity compare across the platform?



- Games dominate the Play Store ecosystem, accounting for the highest total installs by a large margin. This shows gaming apps drive the majority of user engagement on mobile devices.
- Communication and Social apps follow closely, highlighting that messaging, calling, and social networking remain core smartphone activities.
- Productivity, Tools, and Family categories also show very high install volumes, indicating that practical, everyday-use apps have broad, consistent demand across users.
- Categories like Sports, Books & Reference, and Personalization show significantly lower total installs, suggesting that these app types cater to more niche audiences and have limited mass-market demand compared to communication, gaming, or utility apps.

Rating Trends Across Install Tiers

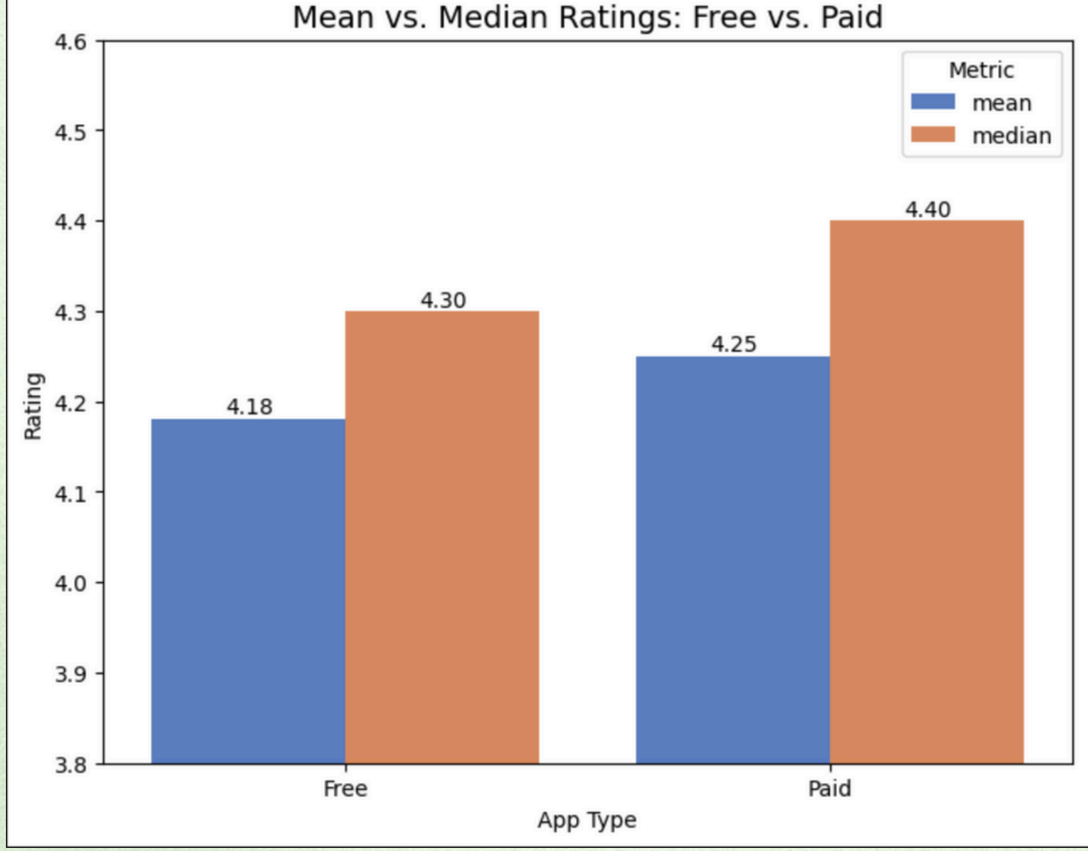
Is there a strong relationship between an app's rating and its install count, or are installs more heavily influenced by external factors like promotion and platform visibility?



- Ratings rise steadily from the Medium to Very High tiers, showing that only consistently high-quality apps manage to scale to tens or hundreds of millions of installs.
- Low-install apps often show high initial ratings because early reviews usually come from supportive or biased early users; ratings normalize once the app reaches a larger audience (Medium tier).
- At the 1B+ tier, ratings dip slightly because widely used apps receive harsher criticism — when an app becomes universal, even small issues generate large volumes of negative feedback.

Rating Distribution: Free vs. Paid Apps

Are paid apps rated better or worse than free apps, and what does this reveal about how monetization affects user experience?



- **Premium Quality Advantage:**Paid apps receive higher ratings (Mean 4.26, Median 4.40) than free apps (Mean 4.18, Median 4.30), showing users are more satisfied with paid apps—likely because they offer ad-free experiences, better features, and more polished quality to justify their price.
- Negative-Tail Skew in Both Categories:For both free and paid apps, the Median is higher than the Mean, revealing a left-skewed distribution where overall satisfaction is high, but a small number of extremely low (1-star) reviews disproportionately drags down the average rating.
- Expectations vs Reality Gap:The rating difference between free and paid apps is small (~0.08), suggesting that while paid apps do perform better, paying users expect more and rate more criticallyso even small bugs or shortcomings prevent paid apps from achieving much higher rating gaps.

Analysis of User Sentiment & App Reviews

- Sentiment Analysis & Opinion Mining of Google Play Store Reviews

Data Methodology (Preprocessing)

Data Source: googleplaystore_user_reviews.csv

Initial State: ~64,000 rows (Dirty, Noisy Data).

The Cleaning Pipeline:

Null Handling: Removed 42% of rows containing missing reviews/sentiment.

De-Duplication: Removed duplicate entries to prevent "spam bias" in sentiment counts.

Text Normalization: Converted text to lowercase; removed special characters & emojis.

Final Dataset: ~37,000 unique, high-quality reviews.

Presented by Juliana Silva

THANK YOU SO MUCH!

Presentations Templates are communication tools that can be used as lectures, reports, and more.

www.reallygreatsite.com