رواد مصر الرقمية

وزارة الاتصـــــالات
وتكنولوجيا المعلومات

What Makes an App Successful on Google Play Store?

Name
Mohamed Ahmed Hassan
Roaa Sabry Aly
Menna Salah Mokhtar
Raneem Tamer Ahmed

1. Objective:

This project analyzes more than **9,600 real apps and 35,930 user reviews** from the Google Play Store to discover the **key factors behind app success**.

We will identify which categories and genres attract the most downloads and highest ratings, compare free versus paid apps in terms of quality and user satisfaction, examine the impact of pricing strategy, app size, required Android version, and user sentiment (positive/negative reviews) on performance, and reveal why some apps remain unrated.

The results will be presented through an interactive dashboard that allows exploration of success patterns by category, price, sentiment, and technical requirements.

This analysis will provide clear, actionable insights for developers, marketers, students, and anyone interested in understanding what truly drives app success on one of the world's largest digital platforms.

---

2. Dataset information:

2.1 Dataset Overview

The project uses the famous **Google Play Store Apps dataset** , containing real data scraped from the Google Play Store between 2010–2018.

- **Main file (googleplaystore.csv)**

  → **Original size:** 10,841 rows

  → **After professional cleaning:** 9,638 unique, up-to-date apps (one row per app – latest version only)

- **Reviews file (googleplaystore_user_reviews.csv)**

  → **Original size:** 64,295 rows

  → **After professional cleaning:** 35,930 high-quality, real user reviews from approximately 800 apps

Both files were cleaned extensively to remove duplicates, corrupted rows, missing values, and formatting errors to ensure 100% accuracy for analysis.

2.2 columns in reviews dataset

The file consists of five columns:

- **App:** The exact name of the application the review belongs to.
- **Translated_Review:** The actual text written by the user, already translated into English. This is the real voice of the user.
- **Sentiment:** A simple classification of the overall emotion of the review, automatically labeled as **Positive, Negative, or Neutral** by a machine learning model.
- **Sentiment_Polarity:** A numerical score from **-1.0 to +1.0** that measures how strongly positive or negative the review is.

  – A value close to **+1.0** means extremely positive (e.g., "Best app ever!!!")

  – A value close to **-1.0** means extremely negative (e.g., "Scam! Don't install!")

  – A value near **0.0** means neutral or factual " no emotion " (e.g., "The app opens normally").
- **Sentiment_Subjectivity:** A score from 0.0 to 1.0 that shows how much the review is based on personal opinion or emotion rather than objective facts:

  – **0.0** = 100% factual and objective (e.g., "The app size is 25 MB")

  – **1.0** = 100% subjective and emotional (e.g., "This is the most amazing game I've ever played!!!")

  – Most real reviews fall between **0.4 and 0.9** because users naturally express feelings, opinions, or experiences rather than pure facts.

---

3. Data cleaning :

3.1 Cleaning the Main Dataset (googleplaystore.csv)

The raw file contained 10,841 rows, but after full professional cleaning we obtained **9,638 unique, up-to-date, and perfectly clean apps**.

1. **Removed one completely corrupted row** A single row (index 10472) had all columns shifted due to a formatting error, causing impossible values such as Category = “1.9” and Rating = 19. This row would have destroyed any statistical analysis, so it was deleted immediately.
2. **Eliminated duplicate apps** Many apps appeared multiple times with different update dates. We normalized the app names by converting them to lowercase and stripping extra spaces to ensure accurate matching. Then We converted “Last Updated” to a real date format and kept only the **most recent version** of each app. This reduced the dataset from ~10,840 to **9,638 truly unique and current apps**.
3. **Converted Installs into a real integer** The original column contained values like “1,000,000+”. We removed the “+” and commas so that 5,000,000,000 becomes a real number. This allows correct mathematical operations and accurate sorting (e.g., finding the most downloaded apps).
4. **Transformed Size into clean Megabytes (MB)** Values were originally “19M”, “415k”, or “Varies with device”. We converted everything to Megabytes (e.g., 19M → 19.00 MB, 415k → 0.40 MB) and filled “Varies with device” with the **median size of the same Category**. Now we can compare app sizes fairly and create beautiful, gap-free charts.
5. **Cleaned Price column** Prices appeared as “$4.99”. We removed the dollar sign, converted to a real number, and rounded to two decimals. This enables accurate revenue calculations (Price × Installs) and clean currency display.
6. **Fixed the Type column (Free/Paid)** Some rows had missing Type. Using the Price column logic (Price = 0 → Free, Price > 0 → Paid), we corrected every single entry so the Free/Paid classification is now **100% accurate**.
7. **Created Android_Min_Version column** From the messy “Android Ver” field (e.g., “4.0.3 and up”), we extracted only the minimum required version as a clean number (e.g., 4.0). Missing values were filled with the **median per Category**. This powerful column lets us analyze how modern or old-targeted each app is.

8. **Handled missing Ratings honestly** Approximately 15% of apps had no rating. Instead of inventing fake ratings (which would be unethical), we kept the original Rating as a real number (with NULLs for calculation) and created a new column **Rating_Display** that shows “Unrated” for missing values and “number of rate” for rated ones. This keeps all calculations accurate while making tables beautiful and transparent.

9. **Beautified the Genres column** Many entries used semicolons (e.g., “Art & Design;Creativity”). We replaced “;” with “ / ” and fixed “&” spacing, turning ugly text into clean, readable genres like “Art & Design / Creativity”.

10. **Applied final rounding** All floating-point columns (Size_MB, Price, Rating, Android_Min_Version) were rounded to 1 or 2 decimals to eliminate ugly artifacts like 9.649999999 and present perfectly clean numbers in reports and dashboards.

3.2 Cleaning the User Reviews Dataset (googleplaystore_user_reviews.csv)

The raw file had 64,295 rows, but more than a third were empty or junk.

1. **Removed over 22,000 completely empty or “nan” rows** These contained no real review text — keeping them would only add noise.

2. **Cleaned the App column** Removed leading/trailing spaces so that every app name matches exactly with the main file. This guarantees a **perfect SQL join on the App column**.

3. **Kept only real English reviews** Dropped rows where Translated_Review was missing or literally “nan”.

4. **Standardized Sentiment** Converted all sentiment labels to consistent title case (Positive, Negative, Neutral) for reliable filtering and visualization.

5. **Converted Sentiment_Polarity and Sentiment_Subjectivity to real numbers** Originally stored as text. Conversion allows us to calculate average sentiment per app or category and create powerful correlation charts.

6. **Rounded polarity and subjectivity to 4 decimals** Produces clean, professional-looking values.

7. **Kept only reviews belonging to apps present in the cleaned main file** Result: ~35,930 high-quality reviews from exactly the same 9,638 apps.

4. Business Questions

This project **answers the following 11 key questions** to fully understand **what makes an app successful on Google Play Store:**

1. What are the 10 most downloaded apps of all time?

2. Which category has the highest average rating?

3. Which category has the most total downloads?

4. Do paid apps have better ratings than free apps?

5. Do paid apps or free apps get more downloads?

6. What is the most common price for paid apps?

7. Do smaller apps (in size) get more downloads or better ratings?

8. Do apps that support older Android versions get more downloads?

9. Which categories have the happiest users (based on review sentiment)?

10. Which genres are the real hidden winners (high rating + high installs)?

11. Which apps make the most money (Price × Installs)?

---

5. Database Design and Implementation in Microsoft SQL Server

A relational database named **GooglePlayStoreDB** was created in Microsoft SQL Server to store the cleaned data professionally.

- The cleaned main dataset (9,638 unique, up-to-date apps) was imported as the table **MainApps**.
- The cleaned reviews dataset (35,930 high-quality reviews belonging only to these apps) was imported as the table **UserReviews**.

Appropriate data types were applied (e.g., **BIGINT** for Installs and Reviews, **DECIMAL** for Rating, Price, Size, and sentiment scores, **NVARCHAR(MAX)** for review text, and **DATE** for Last Updated) to ensure accuracy and performance.

A **one-to-many relationship** was established:

- **App** in **MainApps** → **Primary Key** (guarantees one row per app)
- **App** in **UserReviews** → **Foreign Key** referencing MainApps(App)

This design ensures **referential integrity**, eliminates duplicates and orphan records, enables fast and accurate joins, and follows industry-standard relational database principles. The final schema is clean, fully normalized, and ready for analytical SQL queries and Power BI visualization.

6. SQL Queries and Analytical Functions

11 SQL queries were developed in Microsoft SQL Server to fully answer the 11 business questions.

- **5 queries use analytical window functions** as required:
- Large numbers **(Installs, revenue)** are formatted using **FORMAT(..., 'N0')** for professional presentation.
- Decimal values **(ratings, prices)** are rounded and cast using **CAST(... AS DECIMAL(10,2))** for clean display.
- **Results clearly prove:**

  - Free apps dominate downloads ·

  - The Events category has the highest average rating .

  - Paid apps have slightly higher ratings ·

  - Smaller apps and older-Android-compatible apps reach more users ·

  - Winner in User Happiness: Comics category .

  - Minecraft is the top-earning paid app.

  - The communication genre is the real hidden winner.