Google Play Store SQL Analysis

Emir Dogan

21/03/2025

Project Objective

The goal of this project is to perform an in-depth analysis of the Google Play Store dataset using SQL within MySQL Workbench. We aim to derive insights about app performance, user engagement, and monetization patterns across various app categories.

Data Source

1. **Dataset Title:** Google Play Store Apps
2. **Source:** Kaggle - Google Play Store Dataset
3. **Publisher:** Lavanya Gupta
4. **License:** This work is licensed under the Creative Commons Attribution 3.0 Unported License. To view a copy of this license, visit http://creativecommons.org/licenses/by/3.0/.
5.

Tools & Technologies

1. **Database System:** MySQL (via MySQL Workbench)
2. **Query Language:** SQL (Standard MySQL)
3. **Documentation:** Google Docs
4. **Code Sharing:** GitHub

Research Questions

1. What are the top-rated apps across categories?
2. Which app categories are the most downloaded?
3. Do paid apps receive better ratings than free apps?
4. Which genres tend to have higher user engagement?
5. How does the price of an app relate to its popularity?

Creating Tables in MySQL Workbench

To begin the analysis, a new MySQL database schema was created and two relational tables were defined to represent the structure of the imported datasets.

The schema was implemented using **MySQL Workbench**, and follows a normalized design, separating application metadata from user reviews. Data types were initially defined with flexibility (e.g., `VARCHAR`, `TEXT`) to facilitate the raw import of potentially inconsistent data. Data transformation and cleaning operations will follow in subsequent steps.

The primary table, `playstore_apps`, stores high-level metadata for each application listed on the Google Play Store. The table was created using the following SQL script:

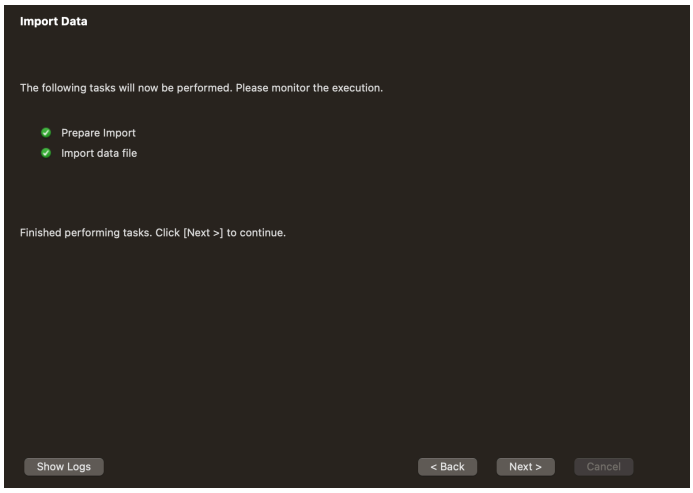
```sql
CREATE DATABASE IF NOT EXISTS playstore_db;
USE playstore_db;

CREATE TABLE playstore_apps (
    app_name VARCHAR(255),
    category VARCHAR(100),
    rating DECIMAL(3,2),
    reviews VARCHAR(50),
    size VARCHAR(50),
    installs VARCHAR(50),
    type VARCHAR(10),
    price VARCHAR(20),
    content_rating VARCHAR(50),
    genres VARCHAR(100),
    last_updated VARCHAR(50),
    current_version VARCHAR(50),
    android_version VARCHAR(50)
);
```

The secondary table, `app_reviews`, captures user feedback associated with the apps found in the primary table. It includes sentiment classification and polarity metrics derived from natural language processing. The table was created as follows:

```
CREATE TABLE app_reviews (
    app_name VARCHAR(255),
    translated_review TEXT,
    sentiment VARCHAR(20),
    sentiment_polarity DECIMAL(4,2),
    sentiment_subjectivity DECIMAL(4,2)
);
```

Data Importation Process

The CSV files were imported into MySQL using the **Table Data Import Wizard** in MySQL Workbench. This graphical tool allows for seamless loading of CSV files into existing table structures by mapping each column appropriately.



Data Cleaning and Preparation

Prior to conducting any analytical operations, the data required cleaning to resolve inconsistencies, ensure correct data types, and eliminate formatting anomalies. This step was performed entirely within the MySQL Workbench environment using SQL.

Cleaning the playstore_apps Table

Remove Special Characters from `installs` and Convert to Integer

The `installs` column initially included formatting characters such as commas and plus signs. These were removed and the column was converted to a numerical type.

```
UPDATE playstore_apps
SET installs = REPLACE(REPLACE(installs, '+', ''), ',', '');

ALTER TABLE playstore_apps
MODIFY COLUMN installs BIGINT;
```

Clean and Convert the `price` Column

The `price` column contained dollar signs, which were removed. The column was then converted to a decimal format.

```
UPDATE playstore_apps
SET price = REPLACE(price, '$', '')
WHERE price IS NOT NULL;

UPDATE playstore_apps
SET price = NULL
WHERE LOWER(TRIM(price)) IN ('nan', 'na', '');

ALTER TABLE playstore_apps
MODIFY COLUMN price DECIMAL(6,2);
```

Convert `reviews` to Integer

```
UPDATE playstore_apps
SET reviews = NULL
WHERE LOWER(TRIM(reviews)) IN ('nan', 'na', '');

ALTER TABLE playstore_apps
MODIFY COLUMN reviews BIGINT;
```

Normalize the `type` Field

One record contained an invalid value (`"0"`) in the `type` column. Additionally, `"nan"` string values were treated as null.

```
DELETE FROM playstore_apps
WHERE type = '0';

UPDATE playstore_apps
SET type = NULL
WHERE LOWER(TRIM(type)) IN ('nan', 'na', '');
```

Normalize the `rating` Field

Missing or invalid ratings were standardized. Both literal string values such as `"nan"` and actual SQL `NULL`s were handled.

```
UPDATE playstore_apps
SET rating = NULL
WHERE LOWER(TRIM(rating)) IN ('nan', 'na', '');
```

Convert `last_updated` to Date Format

The `last_updated` field was converted from string to proper `DATE` format after cleaning string `"NaN"` values.

```sql
ALTER TABLE playstore_apps
ADD COLUMN last_updated_temp DATE;

UPDATE playstore_apps
SET last_updated_temp = STR_TO_DATE(last_updated, '%M %d, %Y');

ALTER TABLE playstore_apps
DROP COLUMN last_updated;

ALTER TABLE playstore_apps
CHANGE COLUMN last_updated_temp last_updated DATE;

UPDATE playstore_apps
SET last_updated = NULL
WHERE TRIM(last_updated) = 'NaN';

ALTER TABLE playstore_apps
MODIFY COLUMN last_updated DATE;

UPDATE playstore_apps
SET last_updated = STR_TO_DATE(last_updated, '%M %d, %Y');
```

Cleaning the app_reviews Table

In the `app_reviews` dataset, invalid values appeared as lowercase `"nan"` strings. These were treated as missing data and removed entirely to maintain analytical accuracy.

```sql
USE playstore_db;
DELETE FROM app_reviews
WHERE LOWER(TRIM(sentiment)) = 'nan'
   OR LOWER(TRIM(translated_review)) = 'nan'
   OR LOWER(TRIM(sentiment_polarity)) = 'nan'
   OR LOWER(TRIM(sentiment_subjectivity)) = 'nan';
```

Additionally, rows with any `NULL` in core sentiment columns were excluded:

```
DELETE FROM app_reviews
WHERE sentiment IS NULL
   OR sentiment_polarity IS NULL
   OR sentiment_subjectivity IS NULL;
```

Exploratory Data Analysis

Objective 1:

To identify which applications have the highest number of installs across the Play Store.

```
USE playstore_db;

SELECT app_name, installs
FROM playstore_apps
WHERE installs IS NOT NULL
ORDER BY installs DESC
LIMIT 70;

SELECT DISTINCT app_name, installs
FROM playstore_apps
WHERE installs >= 1000000000;
```

These 20 applications have surpassed the **1,000,000,000+ install milestone**, making them the most downloaded and widely used apps on the Play Store:

- **Google Play Books**
- **Messenger – Text and Video Chat for Free**
- **WhatsApp Messenger**
- **Google Chrome: Fast & Secure**

- **Gmail**
- **Hangouts**
- **Skype – free IM & video calls**
- **Google Play Games**
- **Subway Surfers**
- **Facebook**
- **Instagram**
- **Google+**
- **Google Photos**
- **Maps – Navigate & Explore**
- **Google Street View**
- **Google**
- **Google Drive**
- **YouTube**
- **Google Play Movies & TV**
- **Google News**

Insights:

- The **Google ecosystem dominates** this list, with the majority of apps coming from Google itself.

- **Social media and messaging apps** like Facebook, Instagram, WhatsApp, and Messenger show the massive demand for connectivity.

- **YouTube, Play Movies & TV, and Play Books** demonstrate the global appetite for digital content consumption.

- **Subway Surfers** stands out as the only game to reach this level, showing exceptional popularity in the mobile gaming space.

Objective 2:

To examine which categories account for the largest user base.

```sql
SELECT category, SUM(installs) AS total_installs
FROM playstore_apps
WHERE installs IS NOT NULL
GROUP BY category
ORDER BY total_installs DESC;
```

An analysis of total installs by app category reveals that the **largest user bases** are concentrated in a few dominant areas, as detailed below:

1. **GAME** – *35.08 billion installs*
   The gaming category holds the largest user base, reflecting the immense popularity and diversity of mobile games. From casual puzzles to immersive action titles, mobile gaming continues to be the leading driver of user engagement.

2. **COMMUNICATION** – *32.65 billion installs*
   This category includes essential messaging and calling apps like WhatsApp, Messenger, and Skype. Its strong position emphasizes the critical role of communication tools in users' daily lives.

3. **PRODUCTIVITY** – *14.18 billion installs*
   Apps that enhance work efficiency, such as Google Drive, Microsoft Office, and task managers, attract a massive audience—especially as mobile devices become core tools for work and study.

4. **SOCIAL** – *14.07 billion installs*
   Platforms like Facebook, Instagram, and TikTok fall under this category. Their install volume underlines how central social networking has become in mobile user behavior.

5. **TOOLS** – *11.45 billion installs*
   Utilities such as file managers, battery savers, and cleaning tools make up this segment, showing steady demand due to their functional importance.

The categories with the highest number of installs are those that cater to **entertainment, communication, and daily utility**—indicating that users prioritize apps that either connect them, enhance productivity, or offer engaging experiences.

Objective 3:

To evaluate whether users rate paid apps more favorably than free ones.

```sql
SELECT type, ROUND(AVG(rating), 2) AS avg_rating, COUNT(*) AS app_count
FROM playstore_apps
WHERE rating IS NOT NULL
GROUP BY type;
```

An analysis of average user ratings across paid and free apps reveals a slight but noticeable difference in user sentiment:

- **Paid apps have a higher average rating (4.27)** compared to free apps (4.19).
- While the **difference in average rating is small (0.08 points)**, it suggests that users may perceive paid apps as higher quality or more valuable.
- The **sample size for paid apps is significantly smaller**, with only 647 apps compared to over 8,700 free apps. This could indicate that paid apps are more curated or specialized, potentially contributing to higher satisfaction.
- Free apps may suffer from **more variability in quality, user expectations, or ad-driven experiences**, possibly leading to slightly lower ratings.

Users tend to rate **paid apps slightly more favorably** than free ones, possibly due to perceived value, quality assurance, or fewer intrusive monetization methods. However, the dominance of free apps in volume shows that despite the marginal rating difference, free apps remain the overwhelming choice for users.

Objective 4:

To discover which app genres consistently receive high user satisfaction.

```sql
SELECT genres, ROUND(AVG(rating), 2) AS avg_rating, COUNT(*) AS app_count
FROM playstore_apps
WHERE rating IS NOT NULL
GROUP BY genres
HAVING app_count > 10
ORDER BY avg_rating DESC
LIMIT 10;
```

genres	avg_rating	app_count
Casual;Brain Games	4.47	13
Events	4.44	45
Simulation;Action & Adventure	4.42	11
Adventure;Action & Adventure	4.42	13
Word	4.41	28
Puzzle	4.39	121
Education;Pretend Play	4.38	23
Puzzle;Brain Games	4.37	19
Education;Education	4.37	50
Art & Design	4.36	56

Analyzing the average user ratings by genre provides insight into the types of applications that consistently deliver a satisfying user experience. Below are some of the top-performing genres in terms of user ratings:

- **Casual games with a brain-training focus** top the list, suggesting users enjoy simple yet mentally stimulating experiences.
- **Events** apps—typically used for scheduling and organization—also receive high satisfaction, likely due to their practical utility and reliability.
- **Genres combining simulation or adventure elements** tend to maintain strong engagement and receive positive feedback from users.
- **Word games**, which often blend entertainment with vocabulary challenges, also rank highly in terms of user ratings.

App genres that combine **mental engagement**, **organization**, and **interactive storytelling** consistently achieve higher user satisfaction. Developers targeting these genres may benefit from a strong baseline of user interest and appreciation.

Objective 5:

To explore the relationship between app price and user satisfaction.

```sql
SELECT
  CASE
    WHEN price = 0 THEN 'Free'
    WHEN price <= 1 THEN '$0.01–$1'
    WHEN price <= 5 THEN '$1.01–$5'
    WHEN price <= 10 THEN '$5.01–$10'
    ELSE 'Over $10'
  END AS price_range,
  ROUND(AVG(rating), 2) AS avg_rating,
  COUNT(*) AS app_count
FROM playstore_apps
WHERE rating IS NOT NULL
GROUP BY price_range
ORDER BY avg_rating DESC;
```

price_range	avg_rating	app_count
$0.01–$1	4.30	109
$1.01–$5	4.28	398
$5.01–$10	4.28	70
Free	4.19	8717
Over $10	4.15	70

User ratings were analyzed across different price ranges to assess whether there's a meaningful connection between how much users pay and how satisfied they are with the app.

- **Low-cost paid apps ($0.01–$1)** receive the **highest average user rating (4.30)**, suggesting a sweet spot where users perceive good value for money.

- Apps priced between **$1.01 and $10** maintain similarly high satisfaction, indicating that moderate pricing does not negatively impact user experience.

- **Free apps**, although overwhelmingly more common, receive slightly lower average ratings (4.19), likely due to a wider range of quality and ad-driven models.

- Interestingly, **apps priced over $10 have the lowest satisfaction (4.15)**, possibly due to higher expectations or niche use cases that may not justify their price point for all users.

There is a **mild positive correlation between paying a small amount and higher user satisfaction**, with **ultra-expensive apps showing diminishing returns in perceived value**. Developers may benefit from positioning their apps within the **low-to-mid price range** to maximize user appreciation while still generating revenue.

Objective 6:

To understand the general emotional tone of user feedback on Play Store apps.

```
3
4 •   SELECT sentiment, COUNT(*) AS total_reviews
5     FROM app_reviews
6     GROUP BY sentiment
7     ORDER BY total_reviews DESC;
8
```

- The **vast majority of user feedback (≈82%) is positive**, reflecting a generally satisfied and appreciative user base.

- **Neutral reviews** (≈12%) form a modest portion of feedback, often reflecting functional comments or suggestions without strong emotional lean.

- **Negative sentiment** accounts for only **5% of total reviews**, indicating that most users have positive experiences with the apps they use.

The emotional tone of user feedback on the Play Store is **overwhelmingly positive**, suggesting that most users are pleased with the app experiences provided. This trend is encouraging for developers and brands looking to build trust and engagement through mobile platforms.

Objective 7:

To highlight apps that receive a large number of positively classified reviews.

```sql
SELECT app_name, COUNT(*) AS positive_review_count
FROM app_reviews
WHERE sentiment = 'Positive'
GROUP BY app_name
ORDER BY positive_review_count DESC
LIMIT 10;
```

From the sentiment analysis of individual apps, the application **“10 Best Foods for You”** stands out with a total of **154 positively classified user reviews**.

- The **"10 Best Foods for You"** app appears to resonate very well with users, receiving a **notably high volume of positive feedback**.
- This volume of praise indicates a high level of user satisfaction, which could be attributed to the app’s usefulness, simplicity, or content quality.

Conclusion

This project successfully applied structured SQL-based analysis to the Google Play Store dataset to derive actionable insights on user behavior, app performance, and monetization trends. By creating a normalized schema in MySQL Workbench, performing extensive data cleaning, and executing targeted exploratory queries, the analysis highlighted key characteristics of high-performing applications and the factors influencing user engagement and satisfaction.

Findings revealed that:

- Applications developed by Google dominate the install rankings, reflecting the company's deep integration within the Android ecosystem.
- Gaming, communication, and productivity apps account for the largest user bases, confirming their central role in mobile user behavior.
- Paid apps tend to receive marginally higher ratings, though free apps remain overwhelmingly dominant in volume.
- Genres associated with cognitive engagement and organizational utility (such as word games and event planners) yield consistently high satisfaction scores.
- Low-cost paid apps ($0.01–$1.00) offer the best perceived value based on user ratings.
- Sentiment analysis of user reviews shows an overwhelmingly positive tone, with negative feedback forming a small minority.

By combining technical data processing with interpretative insight, this project demonstrates the power of SQL for large-scale application analytics. The methodology presented here can be extended to other app marketplaces, customer feedback datasets, or product ecosystems to support data-driven decision-making for developers, marketers, and platform strategists.