Google Play Store Data Analysis & Insights Using Python & Machine Learning

Objective: To analyze Google Play Store apps using EDA, data cleaning, visualization, and basic machine learning techniques to uncover important patterns, trends, and app behavior on the platform.

Key Focus Areas:

- Data cleaning & preprocessing
- Exploratory Data Analysis (EDA)
- Visualization of app behaviors
- Metric comparison across categories
- Rating vs Installs relationship
- Category-wise performance insights

Data Understanding & Preprocessing

Dataset Includes: ✔ App name ✔ Category ✔ Rating ✔ Reviews ✔ Installs ✔ Size ✔ Content Rating ✔ Price ✔ Last Updated

Preprocessing Steps:

- Removed missing / inconsistent values
- Converted "1,000+", "10M+", "0" to numeric formats
- Parsed "Size" field to MB units
- Cleaned "Price" column → all apps were identified as FREE
- Encoded categorical features (Category, Content Rating)
- Scaled numeric features where required

Why Preprocessing Matters: Ensures reliable visualizations, valid statistical insights, and stable ML model interpretation.

Key EDA Insights

Top Installed Categories:

- **Game**, **Communication**, **Social**, **Productivity**, **Tools**

Rating Patterns:

- Most ratings fall between **4.0 to 4.5**
- Apps with **4.3–4.5 rating** show the *highest* install frequency
- Apps rated **above 4.3 get ~2.5× more installs** than apps below 4.1

Category Rating Comparison:

- Finance: **4.2**
- Game: **4.3**
- Social: **4.3**

Installs Comparison:

- Finance apps: **~50K installs**
- Game apps: **~5M installs**
- Social apps: ~1M installs

Business Findings & Patterns

Hidden Patterns Uncovered:

1. High Installs ≠ High Ratings

Tools, Dating, and Family apps get *high downloads* but show *lower* average ratings → indicates user dissatisfaction despite popularity.

2. App Quality Drives Growth

Categories with high median ratings (4.3+) consistently dominate install counts.

3. Platform Is Dominated by Free Apps

Entire dataset contains **free apps**, meaning market competition relies on quality and user trust, not pricing.

4. Rating Sweet Spot

The **4.3–4.5 rating segment** appears to be a *growth trigger zone*.

5. Category Influence on Visibility

Entertainment-oriented categories (Games, Social, Communication) naturally attract the largest user base.

Conclusion & Future Scope

Conclusion:

- User ratings strongly influence app visibility and install growth.
- App categories differ widely in user expectations, performance, and growth.
- Positive user feedback (≥4.3) significantly boosts download numbers.
- Some high-install categories struggle with quality issues (Tools, Dating, Family).

Future Scope:

- Build a **predictive model** to estimate installs based on app features.
- Identify ideal feature combinations for a “successful app profile”.
- Perform sentiment analysis on user reviews to understand negative feedback.
- Compare paid vs free markets using a richer dataset.
- Deploy an interactive **Streamlit dashboard** for dynamic exploration.

Acknowledgment

I would like to express my sincere gratitude to Unified Mentor Private Limited for providing me with the opportunity, guidance, and resources to work on this project. Their structured mentorship, learning environment, and practical exposure played a key role in helping me build real-world data analysis skills.

This project became a valuable learning experience thanks to their continuous support and encouragement.