Google Play Store Apps Rating Prediction – Project Report

Name: Monu Ramkesh Gupta
Course / Internship: Data Analyst Internship
Tools: Python, Pandas, NumPy, Scikit-learn, Jupyter Notebook, Power BI

GitHub Repository Link:

---

2. Abstract This project analyzes the Google Play Store dataset to predict app ratings using Exploratory Data Analysis (EDA), data cleaning, visualization, and machine learning techniques. A Power BI dashboard is created to showcase actionable insights for developers.

---

3. Introduction Google Play Store hosts millions of apps, and understanding app ratings helps developers create better apps and capture user attention. This project focuses on analyzing app ratings, comparing Free vs Paid apps, understanding category and content rating distribution, and predicting ratings using machine learning.

---

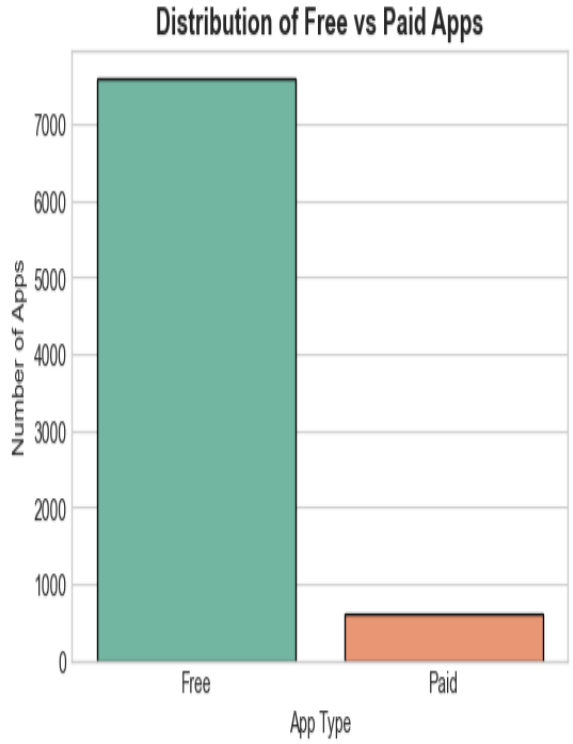
4. Dataset Description - Source: Google Play Store (scraped) - Size: 10,841 rows, 13 columns - Columns: - App, Category, Rating, Reviews, Size, Installs, Type, Price, Content Rating, Genres, Last Updated, Current Version, Android Version

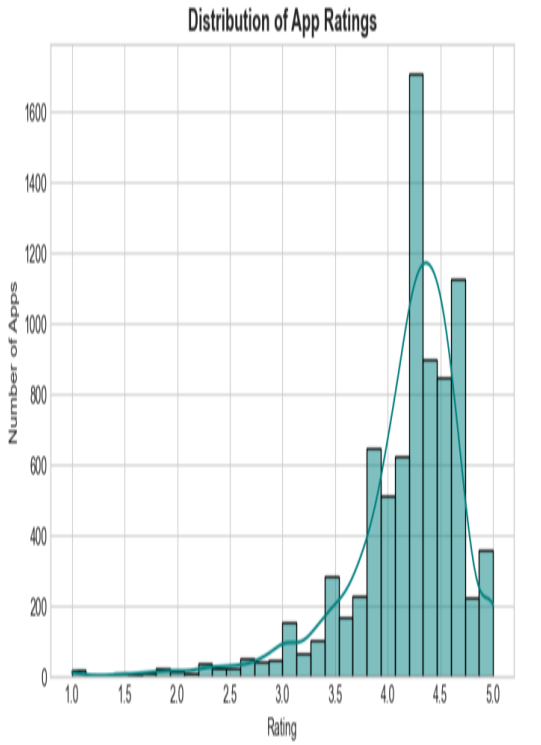
---

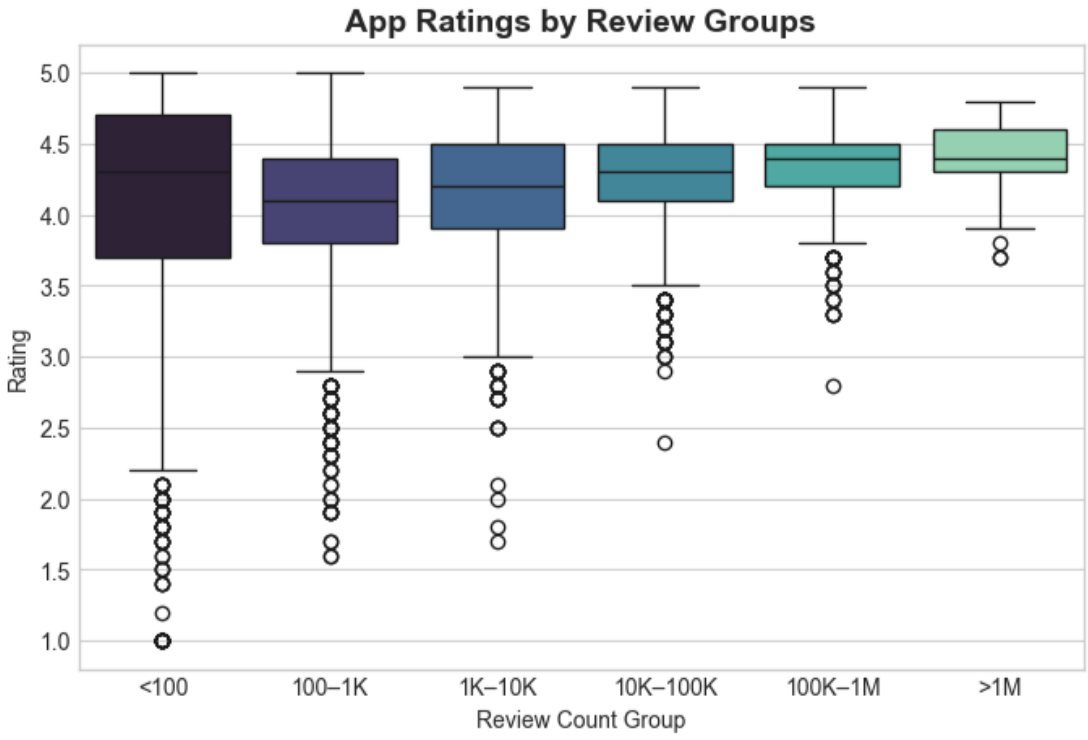
5. Data Cleaning & Preprocessing - Removed duplicate apps and invalid ratings (>5) - Standardized text columns: App, Category, etc. - Converted Size to MB, handling 'Varies with device' - Encoded categorical columns for ML - Result: **8,190** cleaned rows (~24% removed)

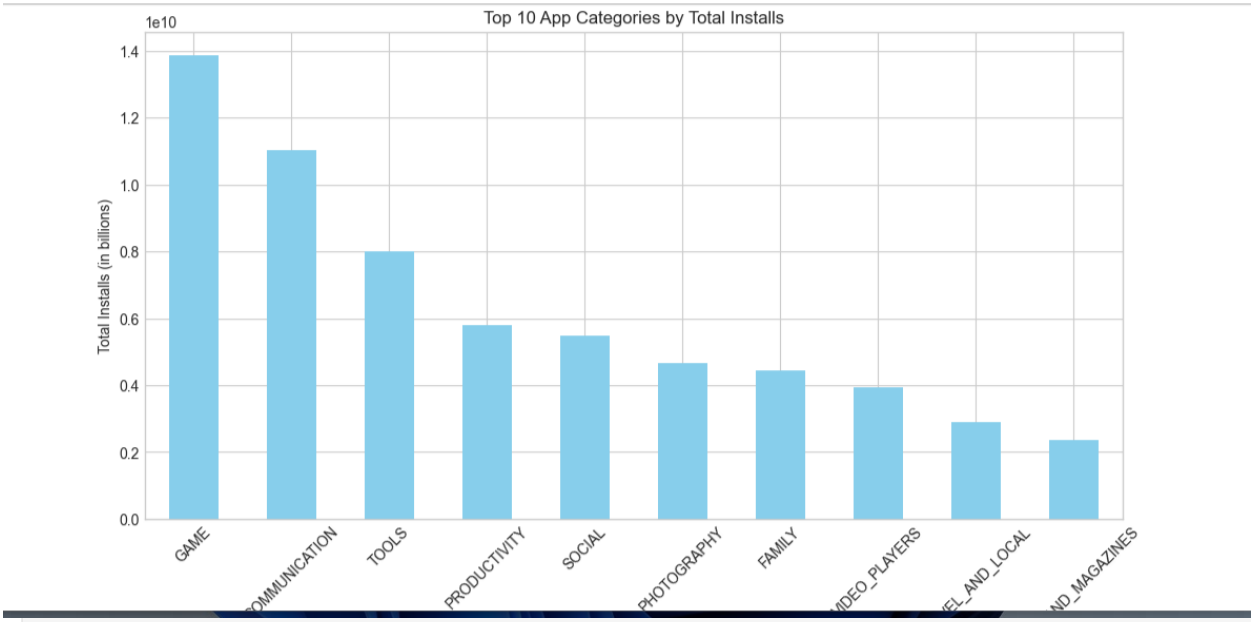
---

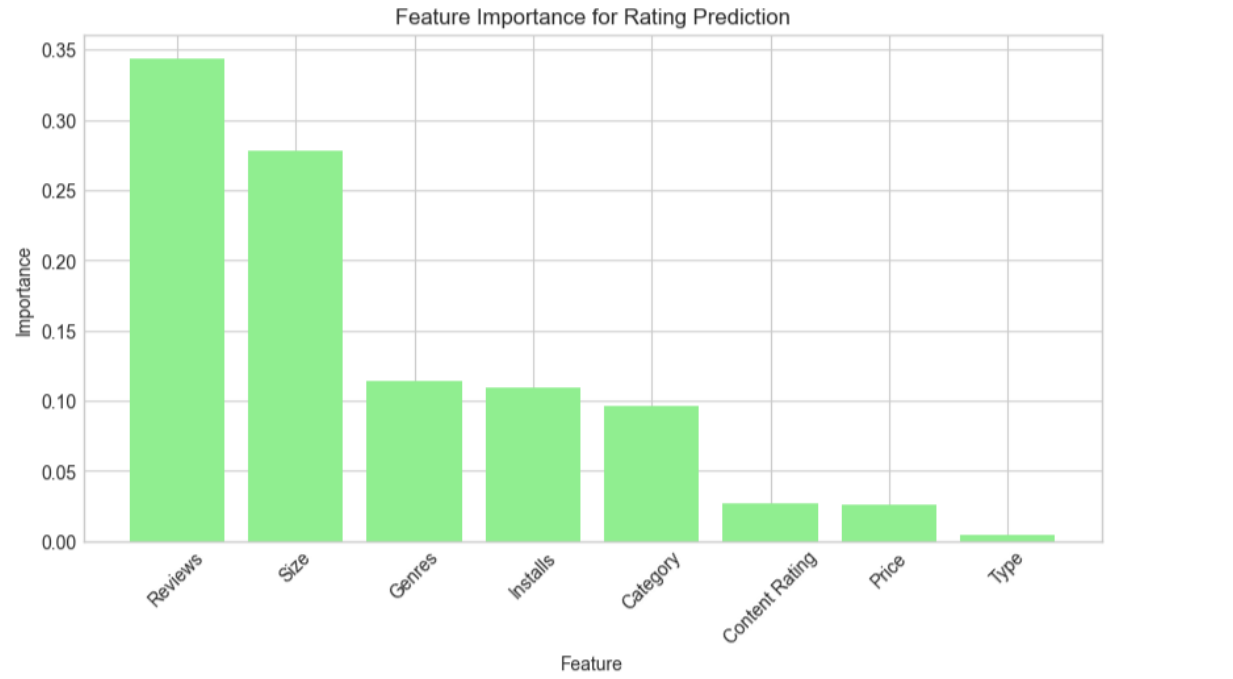
6. Exploratory Data Analysis (EDA) - **Free vs Paid apps:** Free apps dominate (~7k), Paid apps (~1k) - **Content Rating Distribution:** Majority apps for Everyone, then Teen - **Top Categories:** Game, Communication, Tools, Productivity, Social dominate - **Rating Analysis:** Most apps have ratings between 4–5; Free vs Paid comparison via Donut chart - **Reviews vs Ratings:** Higher reviews generally maintain high ratings - **Size vs Ratings:** Small apps (0–10 MB) dominate; larger apps have wider rating spread


Distribution of Free vs Paid Apps
Number of Apps
0
1000
2000
3000
4000
5000
6000
7000
Free
Paid
App Type


Distribution of App Ratings
Number of Apps
0
200
400
600
800
1000
1200
1400
1600
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
Rating


App Ratings by Review Groups
Rating
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
<100
100–1K
1K–10K
10K–100K
100K–1M
>1M
Review Count Group





---

7. Machine Learning Model - **Goal:** Predict Rating based on Reviews, Size, Installs, Genre - **Algorithm:** Random Forest Regressor - **Train-Test Split:** 80% train, 20% test - **Performance:** - R^2 Score: 0.12 - MAE: 0.35 - RMSE: 0.51 - Low R^2 indicates rating influenced by factors beyond dataset features

---

8. Power BI Dashboard - **Background:** Dark gradient for professional look - **Visuals:** 1. Average Rating: Free vs Paid → Donut chart 2. Content Rating Distribution → Donut chart 3. Top Categories (Top 8 + Others) → Donut chart 4. KPI Cards: Total Apps, Average Rating, Total Free, Total Paid, Total Reviews - **Slicers:** Type, Content Rating, Top Categories

---

9. Key Insights - Most apps are Free and target Everyone - Top categories dominate the Play Store - Paid apps slightly outperform Free apps in average rating - Reviews and installs moderately correlate with ratings

---

10. Conclusion - Cleaned and analyzed Google Play Store dataset - Built a Random Forest model for rating prediction - Created an interactive Power BI dashboard for insights - Slicers enable filtering by app type, content rating, and top categories for detailed analysis

---

11. References - Google Play Store Dataset (Scraped) - Python: Pandas, NumPy, Scikit-learn - Power BI - Matplotlib / Seaborn (for charts)

End of Report.