Google Play Store Dataset Summary

Dataset Overview

Dataset Dimensions: 10,841 rows × 13 columns
Total Applications: 10,841 apps
Unique Applications: 9,660 apps
Duplicate Rows: 483

Column Information

The dataset contains 13 columns with the following attributes:

Column	Data Type	Non-Null Count	Description
App	Object	10,841	Application name
Category	Object	10,841	App category classification
Rating	Float64	9,367	User rating (out of 5)
Reviews	Object	10,841	Number of user reviews
Size	Object	10,841	Application size
Installs	Object	10,841	Number of installations
Type	Object	10,840	Free or Paid
Price	Object	10,841	Application price
Content Rating	Object	10,840	Age-appropriate content rating
Genres	Object	10,841	Application genre(s)
Last Updated	Object	10,841	Last update date
Current Ver	Object	10,833	Current version
Android Ver	Object	10,838	Minimum Android version required

Missing Values Analysis

Column	Missing Values
Rating	1,474 (13.6%)
Type	1 (0.01%)
Content Rating	1 (0.01%)
Current Ver	8 (0.07%)
Android Ver	3 (0.03%)

Note: The Rating column has the highest number of missing values, accounting for approximately 14% of the dataset.

Statistical Summary

Rating Distribution

Metric	Value
Count	9,367
Mean	4.19
Standard Deviation	0.54
Minimum	1.00
25th Percentile	4.00
Median (50th)	4.30
75th Percentile	4.50
Maximum	19.00*

*The maximum value of 19.00 appears to be an outlier or data entry error, as ratings should be capped at 5.0.

Categorical Variables Summary

Key Findings

- **Total Categories:** 34 unique app categories
- **Most Common Category:** FAMILY (1,972 apps)
- **Total Genres:** 120 unique genres
- **Most Common Genre:** Tools (842 apps)

Content Rating Distribution

- **Total Content Ratings:** 6 unique ratings
- **Most Common:** Everyone (8,714 apps, 80.4%)

App Type Distribution

- **Free Apps:** 10,039 (92.6%)
- **Paid Apps:** Remaining percentage
- **Total Price Points:** 93 unique prices

Installation Statistics

- **Total Install Ranges:** 22 unique ranges
- **Most Common:** 1,000,000+ installations (1,579 apps)

Size Information

- **Unique Size Values:** 462
- **Most Common:** "Varies with device" (1,695 apps)

Update Information

- **Total Unique Dates:** 1,378
- **Most Frequent Update Date:** August 3, 2018 (326 apps)

Version Information

- **Current Versions:** 2,832 unique versions
- **Most Common:** "Varies with device" (1,459 apps)
- **Android Versions:** 33 unique minimum requirements
- **Most Common:** 4.1 and up (2,451 apps)

Top Applications

Most Frequently Listed App: ROBLOX (9 entries)

Data Quality Issues & Recommendations

Identified Issues

1. **Missing Values:** Significant missing data in the Rating column (13.6%)
2. **Duplicate Entries:** 483 duplicate rows requiring investigation
3. **Data Inconsistencies:**
    - Rating outlier (19.00) needs verification
    - Price column contains special characters (e.g., "$")
    - Installs column contains special characters (e.g., "+")
4. **Data Type Mismatches:** Several numerical columns stored as objects (Reviews, Size, Installs, Price)

Preprocessing Requirements

- Remove or handle duplicate entries
- Clean Price column (remove "$" symbols, convert to float)
- Clean Installs column (remove "+" and "," symbols, convert to integer)
- Convert Reviews to numeric format
- Investigate and handle the Rating outlier (19.00)
- Address missing values through imputation or removal
- Standardize date formats in Last Updated column
- Parse and standardize Size column

Potential Analysis Opportunities

1. **Rating Analysis:** Factors influencing app ratings
2. **Price Strategy:** Relationship between pricing and installations/ratings
3. **Category Performance:** Popular categories and their characteristics
4. **Size Optimization:** Impact of app size on installations
5. **Content Rating Trends:** Distribution across different age groups
6. **Update Frequency:** Correlation between updates and user engagement
7. **Predictive Modeling:** Forecasting app success based on features
8. **Market Segmentation:** Identifying app clusters and market niches

---

Conclusion

This dataset provides comprehensive information about Google Play Store applications suitable for exploratory data analysis and predictive modeling. After appropriate data cleaning and preprocessing, the dataset can yield valuable insights into app performance factors, market trends, and user preferences.