Exploratory Data Analysis of Google Play Store App Data

Name: Jeevan EG

USN: 1RVU22CSE069

Institution: RV University

Abstract

This project focuses on an exploratory data analysis (EDA) of the Google Play Store dataset. The primary objective is to uncover patterns and insights within the data, such as popular app categories, app ratings, and relationships between app characteristics. Key findings reveal trends in app pricing, ratings, and install counts, which can aid developers and marketers in understanding the Google Play market landscape.

Table of Contents

Introduction

Background

The dataset contains detailed information about various Google Play Store apps, including their categories, user ratings, number of installs, and other related metadata. By examining these aspects, the analysis aims to reveal patterns that can be used to improve app performance, target appropriate categories, and increase user engagement.

Objective

The primary goal of this analysis is to:

- Analyze the characteristics of Google Play apps.
- Identify trends in app ratings, installs, and categories.
- Gain insights into free vs. paid app distribution. This foundation is crucial for any subsequent predictive modeling or advanced analysis.

Data Description

Dataset Overview

The dataset comprises approximately 10,000 rows and 13 columns. It contains various attributes of Google Play Store apps, covering both numerical (ratings, installs) and categorical (category, content rating) data.

Variable Descriptions

- **Category**: App category (e.g., Game, Family).
- **Rating**: User rating of the app, a float between 1 and 5.
- **Reviews**: Count of user reviews.
- **Size**: Size of the app in KB or MB.
- **Installs**: Number of app installs.
- **Price**: Cost of the app in USD.
- **Content Rating**: Age-appropriate rating (e.g., Everyone, Teen).
- **Genres**: App genre.
- **Last Updated**: Last updated date.
- **Current Version**: App's current version.
- **Android Version**: Minimum Android OS requirement.

Data Quality

The dataset contained missing values in key columns, particularly in ratings, inconsistent formats in size and installs columns, and duplicate entries. Addressing these quality issues was crucial for reliable analysis

Data Preprocessing

The preprocessing steps were necessary to address data quality issues and prepare the dataset for analysis. Key steps included:

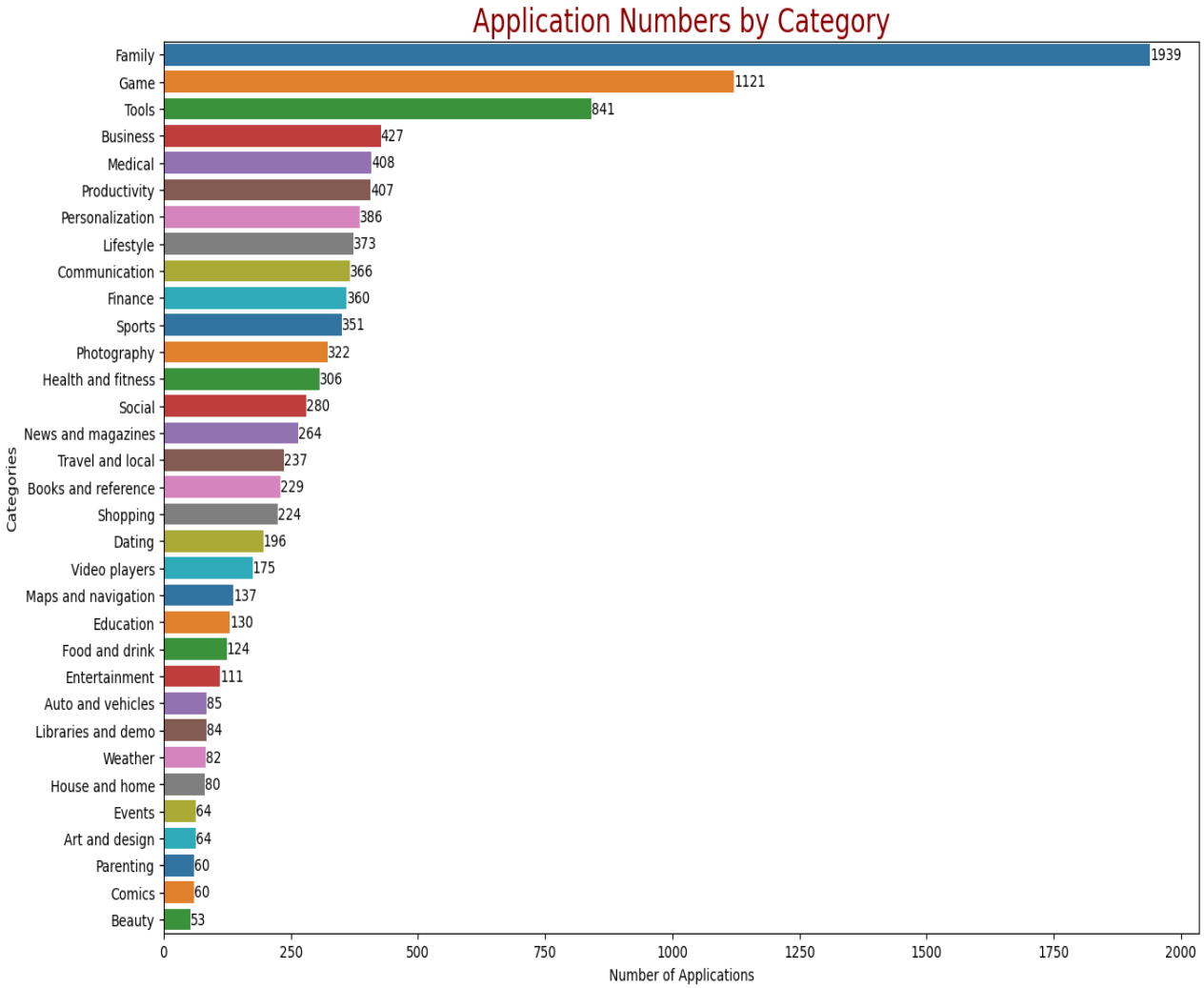
1. **Handling Missing Values**: Missing values in the Rating column were filled with the median value to maintain consistent statistical measures across the dataset.
2. **Removing Duplicates**: Duplicate rows were identified and removed to ensure data accuracy and prevent over-representation.
3. **Data Type Conversion**:
    - **Installs**: Converted from object to numerical format by removing symbols (like “+” and commas).
    - **Size**: Converted to a standard numerical format, with KB and MB standardized for comparison.
4. **Data Standardization**: Categories like Price and Content Rating were standardized to uniform formats, which facilitated better analysis across different data types.

Exploratory Data Analysis (EDA)

The EDA explored the distribution and relationships within the dataset's variables to reveal insights into Google Play Store app trends.

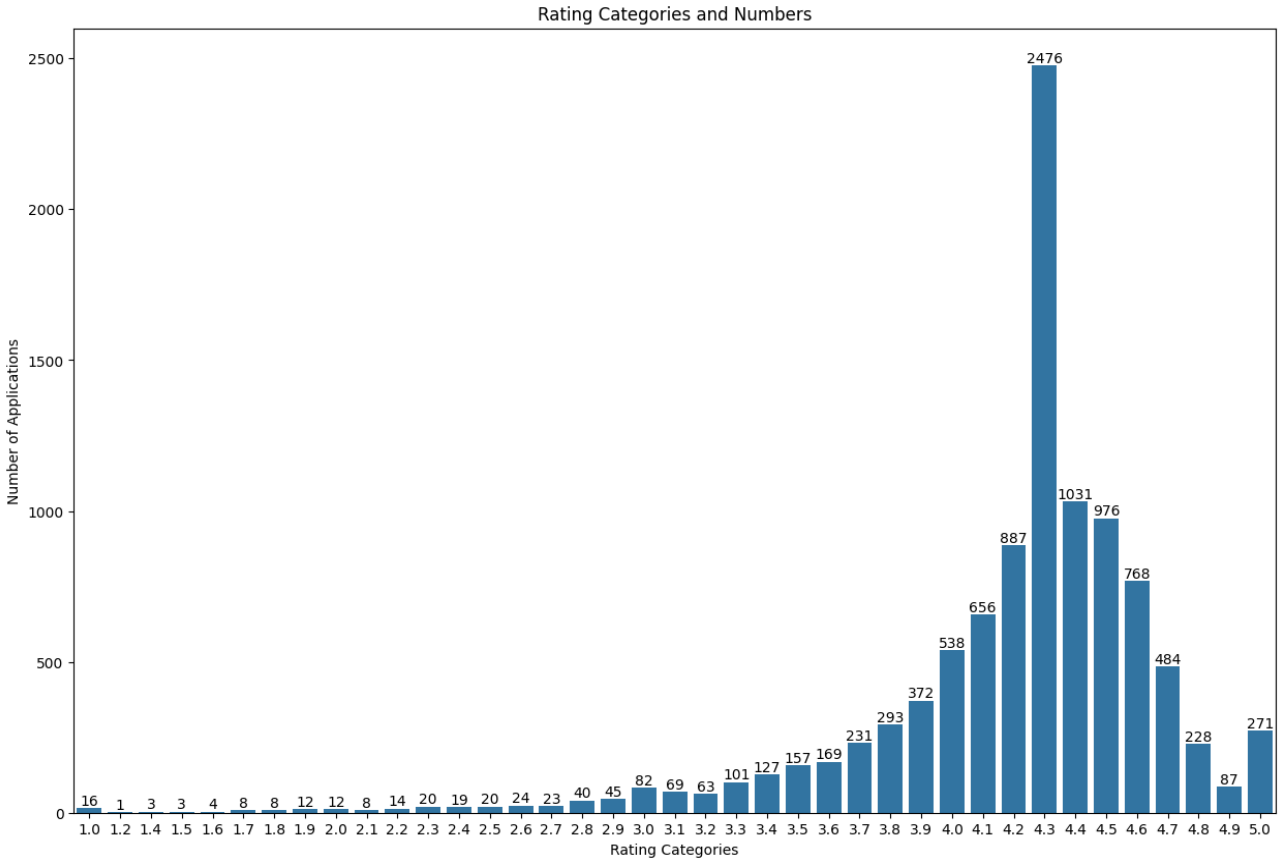
1. App Category Distribution

- The analysis showed that apps in the **Family**, **Game**, and **Tools** categories are the most prevalent. These categories make up a significant portion of the total apps available, suggesting high user demand and competition in these areas.



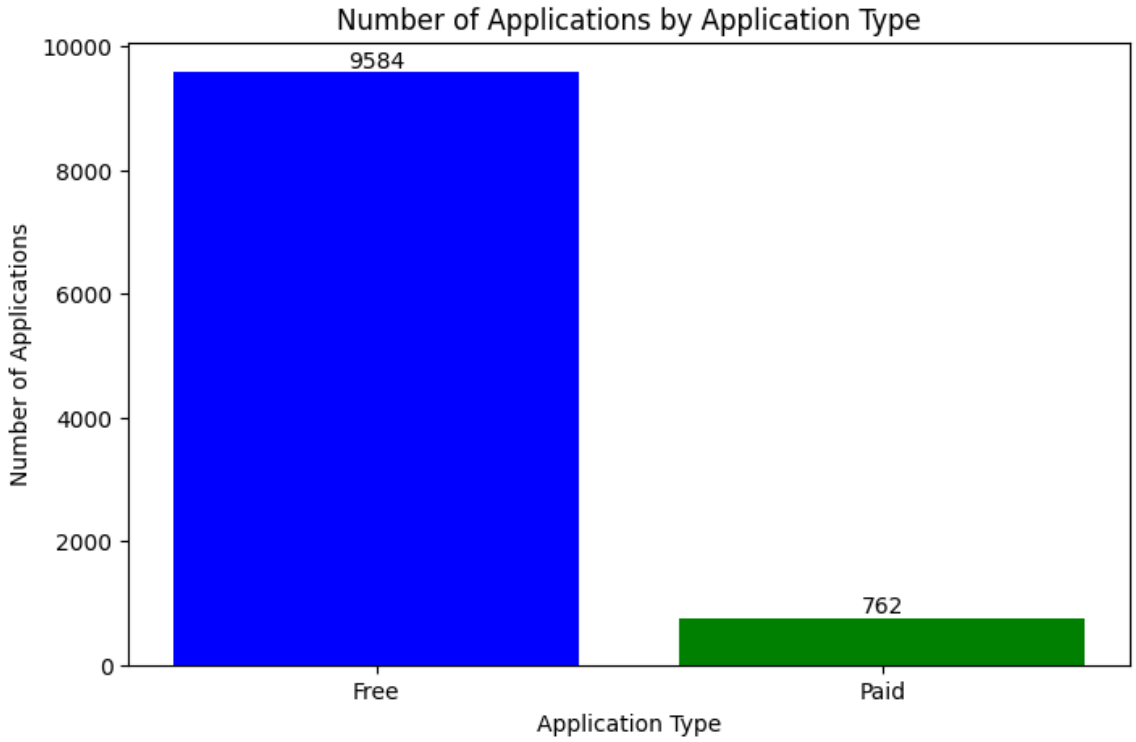
2. Rating Distribution

- Most apps have user ratings between **4.0 and 4.5**, indicating that the average app on Google Play is rated positively. However, a few categories show ratings below this range, indicating areas for potential improvement.



3. Free vs. Paid Apps

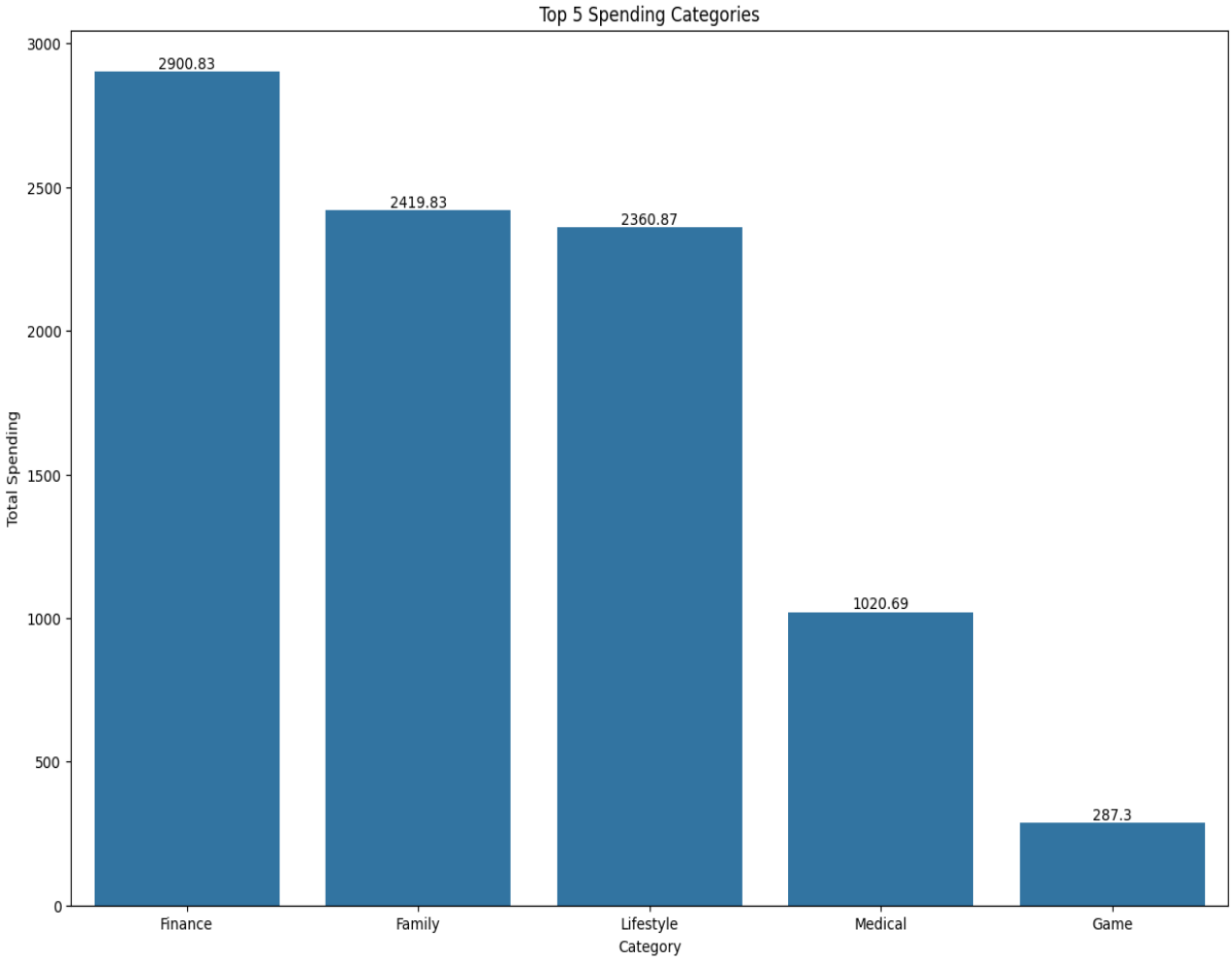
- The dataset revealed a strong preference for **free apps** over paid ones. The analysis found that free apps had significantly higher install counts than paid apps, which suggests that users are more likely to download free apps.
- **Paid Apps**: Tend to have lower install counts but may have higher-quality or niche audiences.



4. Calculate total spend by category

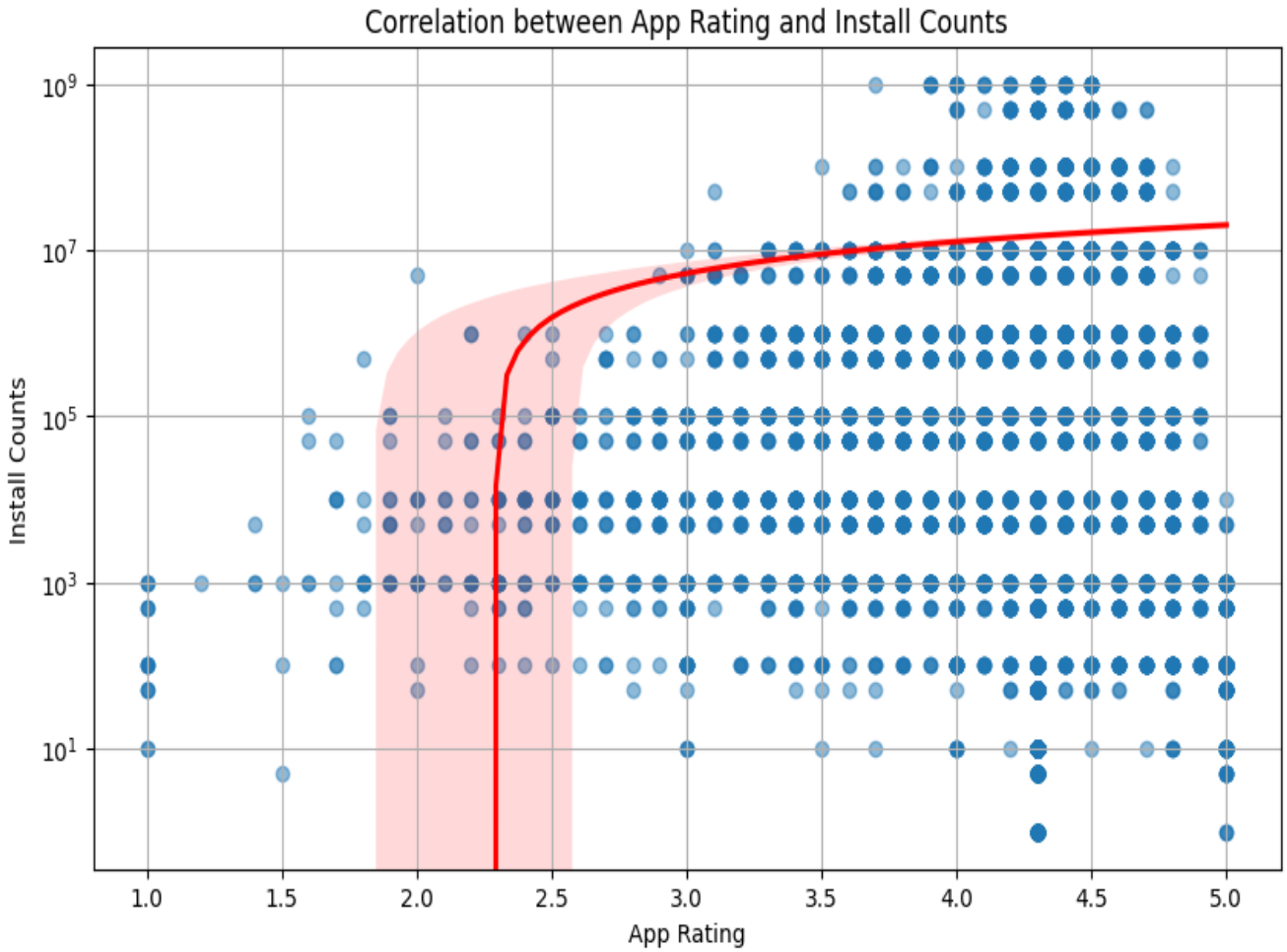
- **Finance dominates spending:** The "Finance" category significantly outspends other categories, with a total expenditure of approximately 2900.83. This suggests a strong financial focus in the person's spending habits.

- **Family and Lifestyle are close contenders:** The "Family" and "Lifestyle" categories follow closely behind "Finance,". This indicates a balance between personal well-being and family commitments.



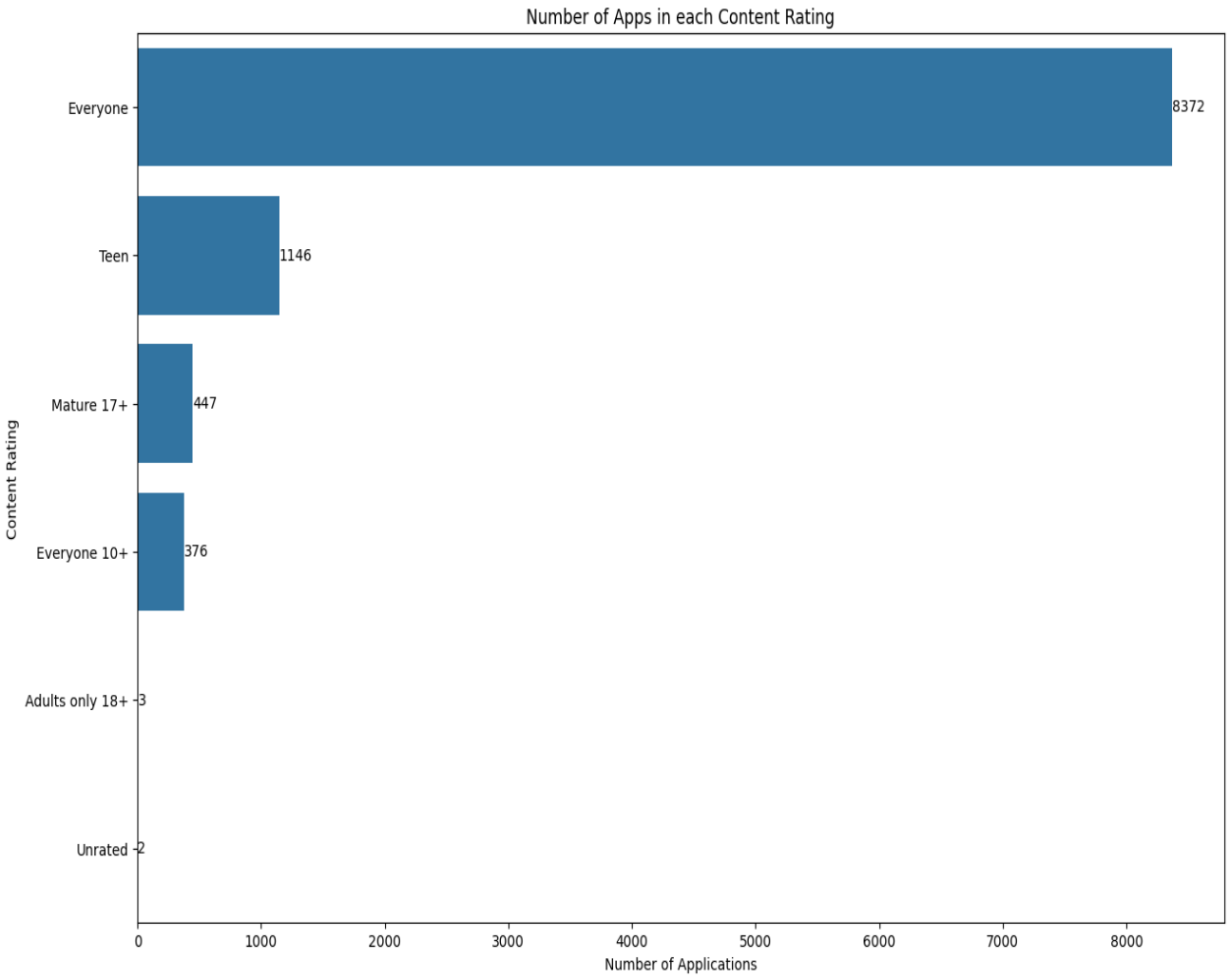
5. Install Count and Rating Relationship

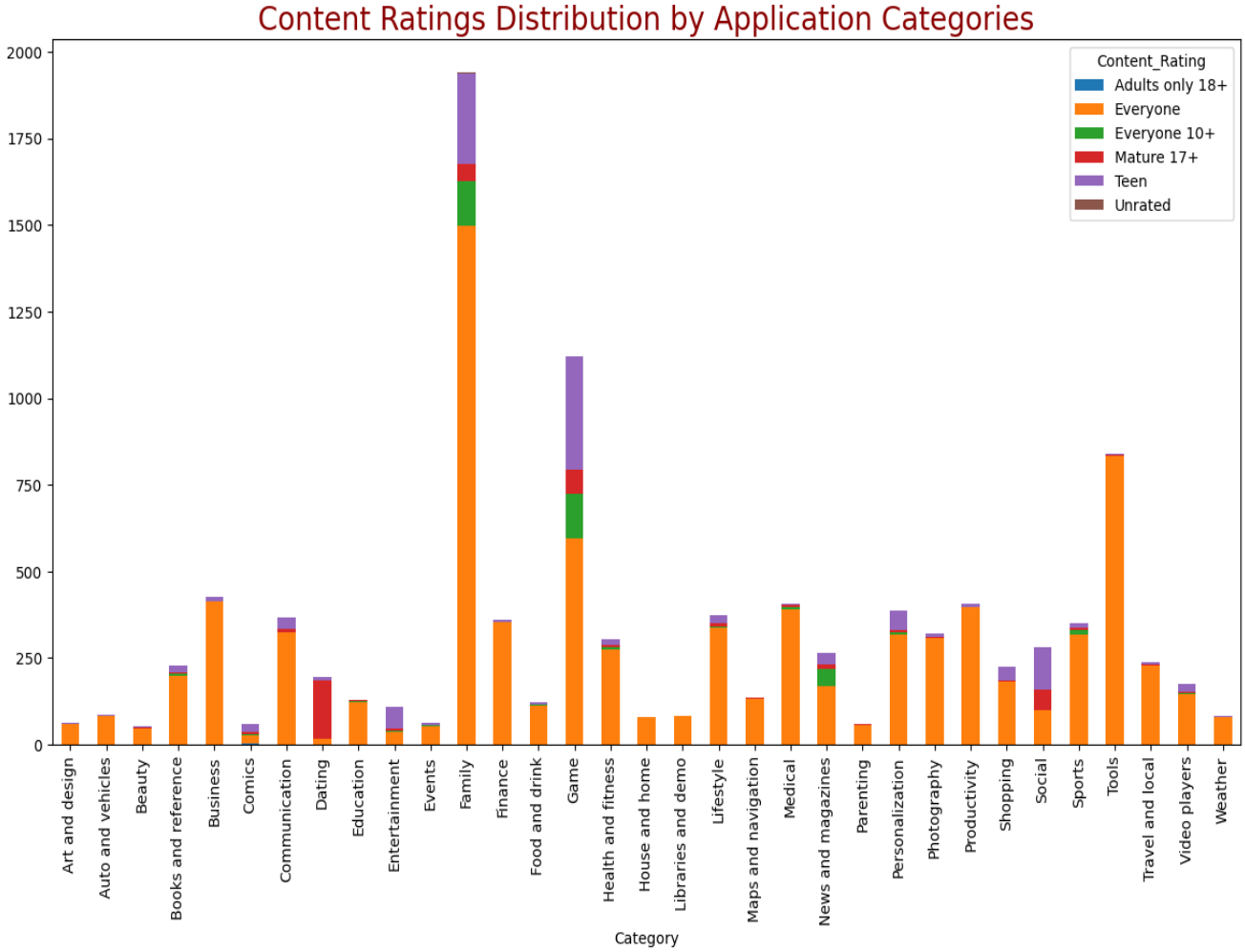
- There is a positive correlation between install counts and ratings, indicating that more highly-rated apps tend to have more downloads. However, certain high-install apps maintain moderate ratings, likely due to popularity rather than quality.



6. Content Rating Analysis

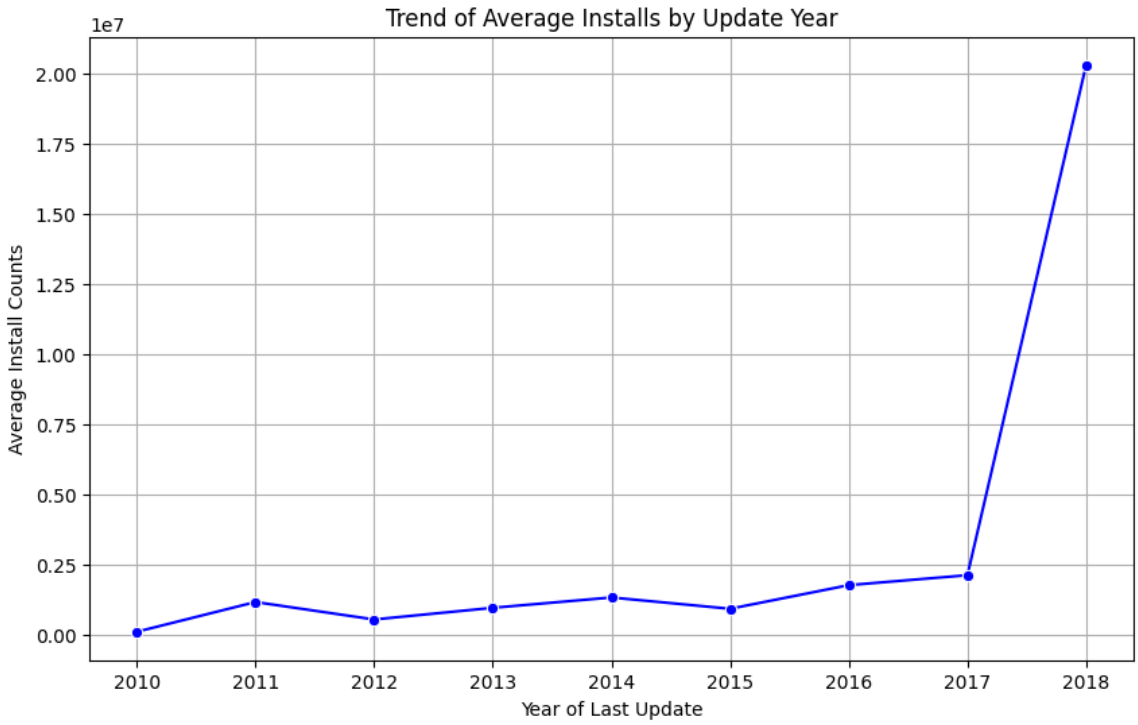
- Most apps are rated **“Everyone”**, making them accessible to all ages, followed by **Teen** and **Mature**. Apps aimed at all age groups have higher installs, reflecting a larger potential user base.




Content Ratings Distribution by Application Categories
Content_Rating
Adults only 18+
Everyone
Everyone 10+
Mature 17+
Teen
Unrated
2000
1750
1500
1250
1000
750
500
250
0
Art and design
Auto and vehicles
Beauty
Books and reference
Business
Comics
Communication
Dating
Education
Entertainment
Events
Family
Finance
Food and drink
Game
Health and fitness
House and home
Libraries and demo
Lifestyle
Maps and navigation
Medical
News and magazines
Parenting
Personalization
Photography
Productivity
Shopping
Social
Sports
Tools
Travel and local
Video players
Weather
Category

7. Update Frequency

- Frequent updates can indicate an app is actively maintained. Popular apps tend to have recent update dates, suggesting that maintenance and responsiveness to user feedback positively impact app popularity.



Visualizations Used

- **Histograms**: Used for rating distributions and install counts.
- **Bar Charts**: Displayed the distribution of categories and content ratings.
- **Scatter Plots**: Showed relationships between installs, ratings, and prices.

Key Findings

1. **Dominant Categories**: Family, Game, and Tools categories have the highest app counts, indicating competitive areas.
2. **User Preference for Free Apps**: Free apps have a substantially larger user base, suggesting that pricing strategies should consider user sensitivity to app costs.
3. **High Ratings and Installs Correlation**: Highly-rated apps generally have more installs, emphasizing the importance of quality for user acquisition.
4. **Price Variation**: Most paid apps are priced modestly, targeting affordability. High-priced apps are rare and typically serve specialized markets.
5. **Content Accessibility**: Apps targeting all age groups tend to perform better in terms of installs, showing the benefit of broader content appeal.

Conclusion

This analysis offers valuable insights into the Google Play Store landscape, where app category, pricing, and content rating significantly influence user engagement and install counts. Free apps with high ratings are more likely to attract a large audience, while specific app categories like Family, Game, and Tools show strong competition and user interest. These insights are essential for developers and marketers aiming to succeed in the competitive app market, enabling them to make data-informed decisions regarding app development, updates, and marketing strategies.