Unified Mentor, Gurugram, Haryana

Dissertation Project Report

On

Google Play Store App Rating Prediction

Presented by:

Swapnil Prabhakar Chavan

Introduction

- **Objective:** Analyze Google Play Store dataset to discover app trends, pricing effects, and factors affecting ratings & installs.
- **Why important:** Helps app developers & marketers improve app visibility and user satisfaction.
- Dataset contains both free & paid apps across multiple categories.

Dataset Overview

- **Source:** googleplaystore.csv
- **Total apps:** ~10,000+
- **Key columns:** Name, Category, Rating, Reviews, Size, Installs, Type, Price, Last Updated
- Missing values in ratings, size, and price handled

Columns: ['App', 'Category', 'Rating', 'Reviews', 'Size', 'Installs', 'Type', 'Price', 'Content Rating', 'Genres', 'Last Updated', 'Current Ver', 'Android Ver']

First rows:

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
0	Photo Editor & Candy Camera & Grid & ScrapBook	ART_AND_DESIGN	4.1	159	19M	10,000+	Free	0	Everyone	Art & Design	January 7, 2018	1.0.0	4.0.3 and up
1	Coloring book moana	ART_AND_DESIGN	3.9	967	14M	500,000+	Free	0	Everyone	Art & Design;Pretend Play	January 15, 2018	2.0.0	4.0.3 and up
2	U Launcher Lite – FREE Live Cool Themes, Hide ...	ART_AND_DESIGN	4.7	87510	8.7M	5,000,000+	Free	0	Everyone	Art & Design	August 1, 2018	1.2.4	4.0.3 and up
3	Sketch - Draw & Paint	ART_AND_DESIGN	4.5	215644	25M	50,000,000+	Free	0	Teen	Art & Design	June 8, 2018	Varies with device	4.2 and up
4	Pixel Draw - Number Art Coloring Book	ART_AND_DESIGN	4.3	967	2.8M	100,000+	Free	0	Everyone	Art & Design;Creativity	June 20, 2018	1.1	4.4 and up

Tools & Libraries

- **Tools:**
- **Python 3** with Jupyter Notebook
- **Libraries:**
- **Pandas** – Used for data cleaning & manipulation
- **Numpy** – Used for numerical calculations
- **Matplotlib** – Used for plotting and graphs
- **Seaborn** – Used for advanced data visualization

Data Cleaning Steps

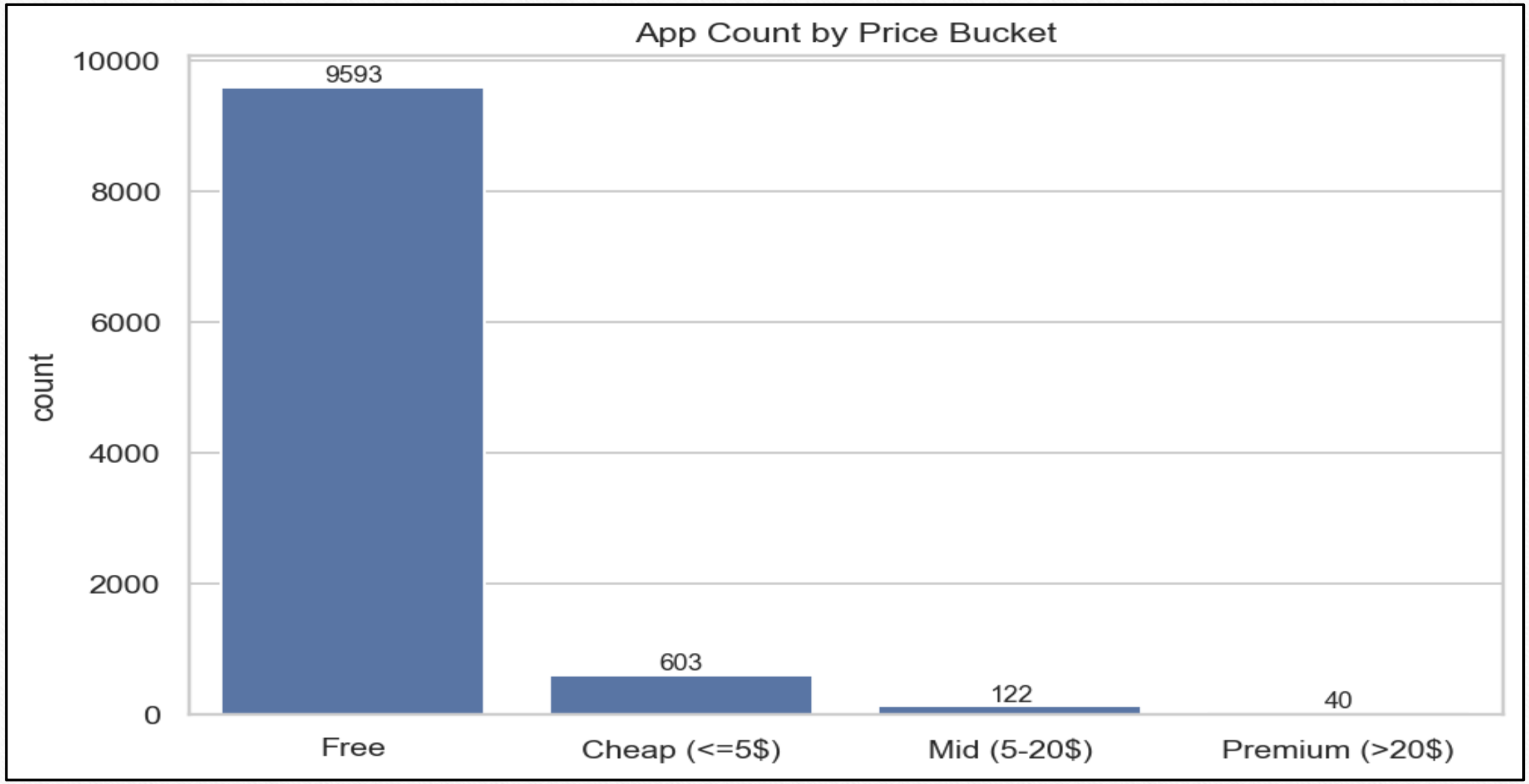
- Removed duplicates
- Converted numeric fields: Reviews, Installs, Price, Size
- Parsed Last Updated → Date & App Age
- Imputed missing ratings by category median
- Created Price Bucket, Size Bucket, Rating Group

Feature Engineering

- **Price Bucket**: Free, Cheap, Mid, Premium
- **Size Bucket**: Small, Medium, Large
- **Rating Group**: Low, Medium, High
- **App Age (days)**: Time since last update
- **Installs per MB**: Efficiency metric

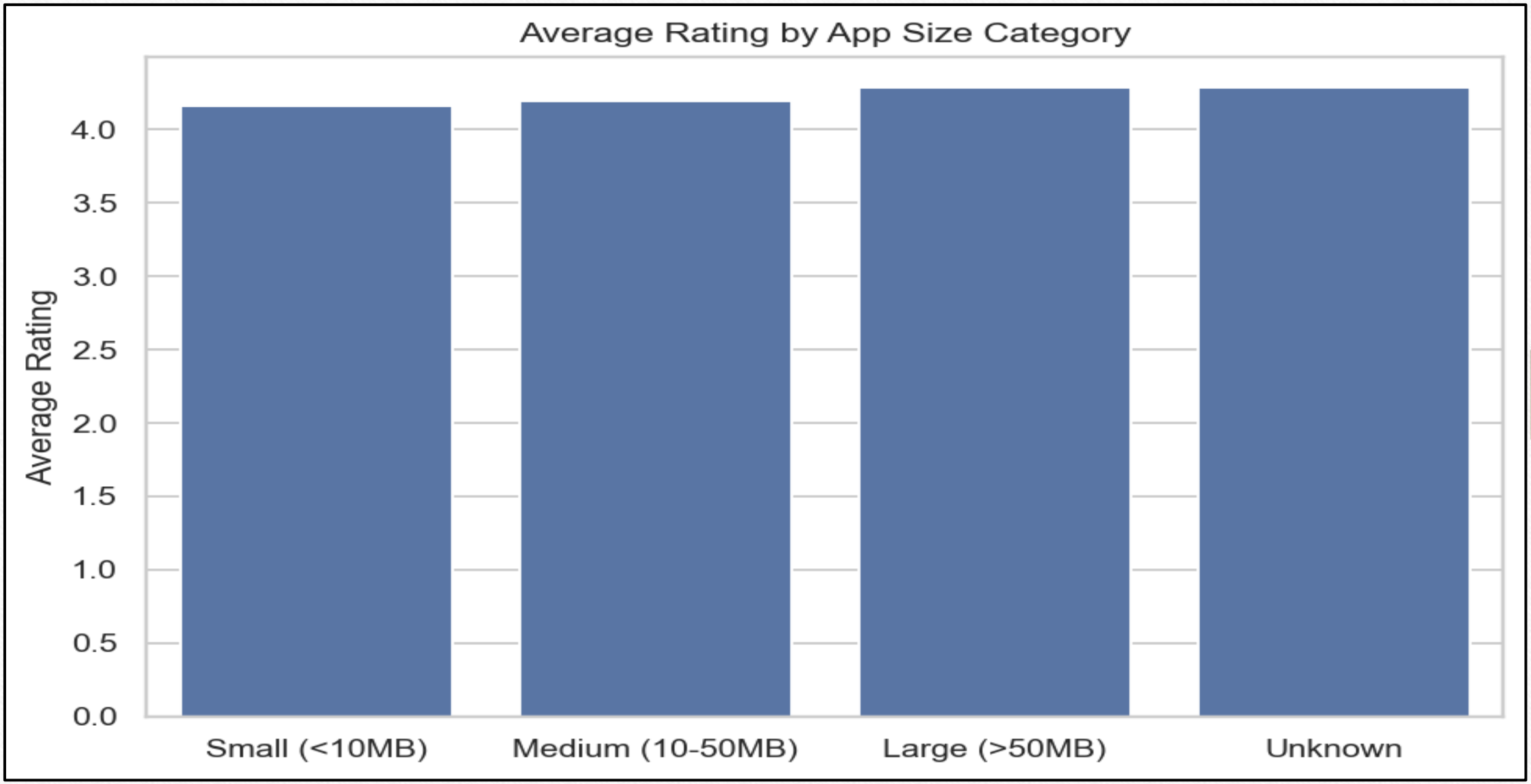
App Count by Price Bucket

- Majority of apps are free (~90%)
- Few premium apps (> $20)
- Cheap apps dominate paid segment


App Count by Price Bucket
10000
8000
6000
4000
2000
0
count
9593
603
122
40
Free
Cheap (<=5$)
Mid (5-20$)
Premium (>20$)

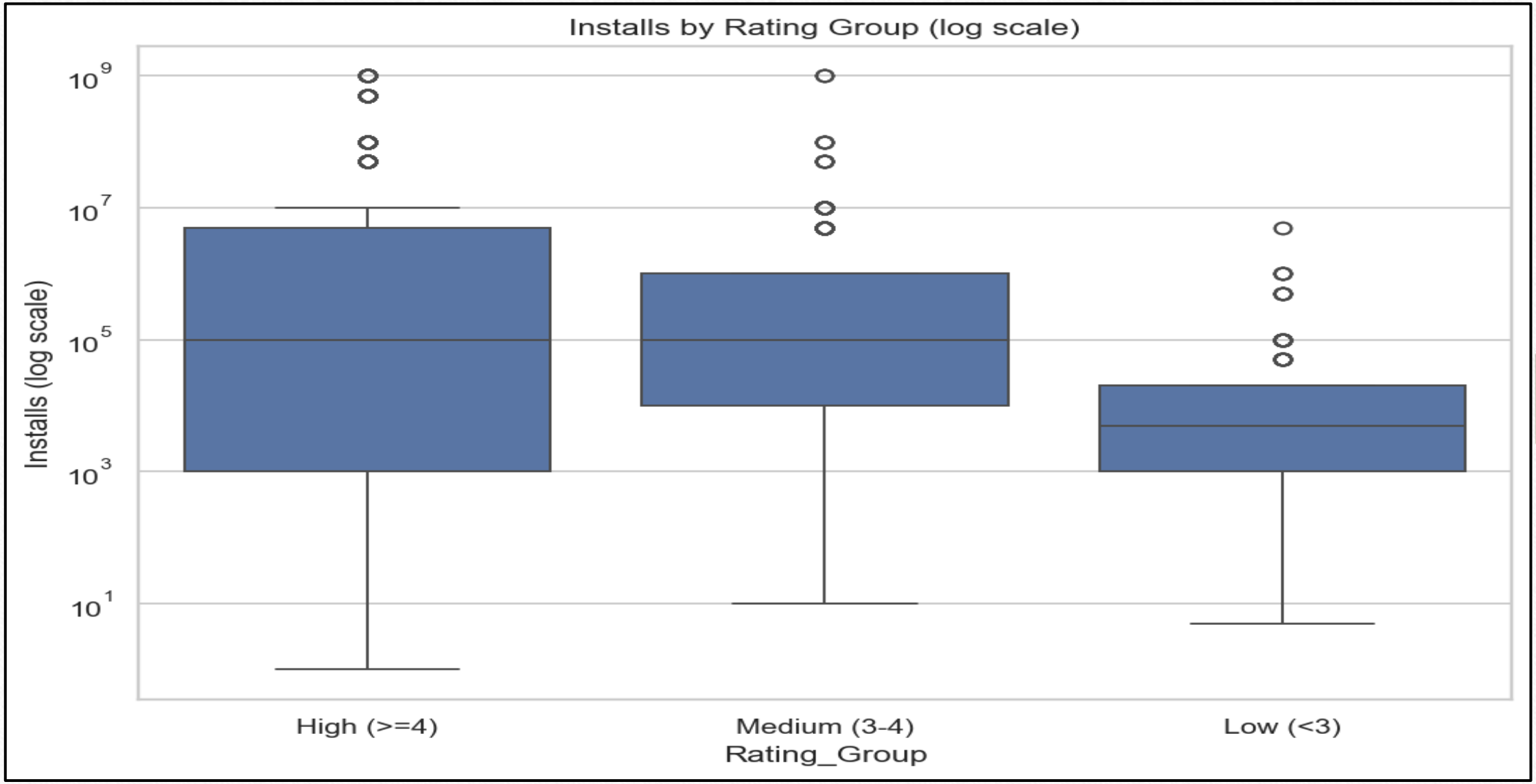
Average Rating by Size Category

- Small (<10MB) & Medium (10–50MB) apps have slightly higher ratings than large apps
- Possible reason: smaller apps run on more devices, better user reach


Average Rating by App Size Category
Average Rating
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Small (<10MB)
Medium (10-50MB)
Large (>50MB)
Unknown

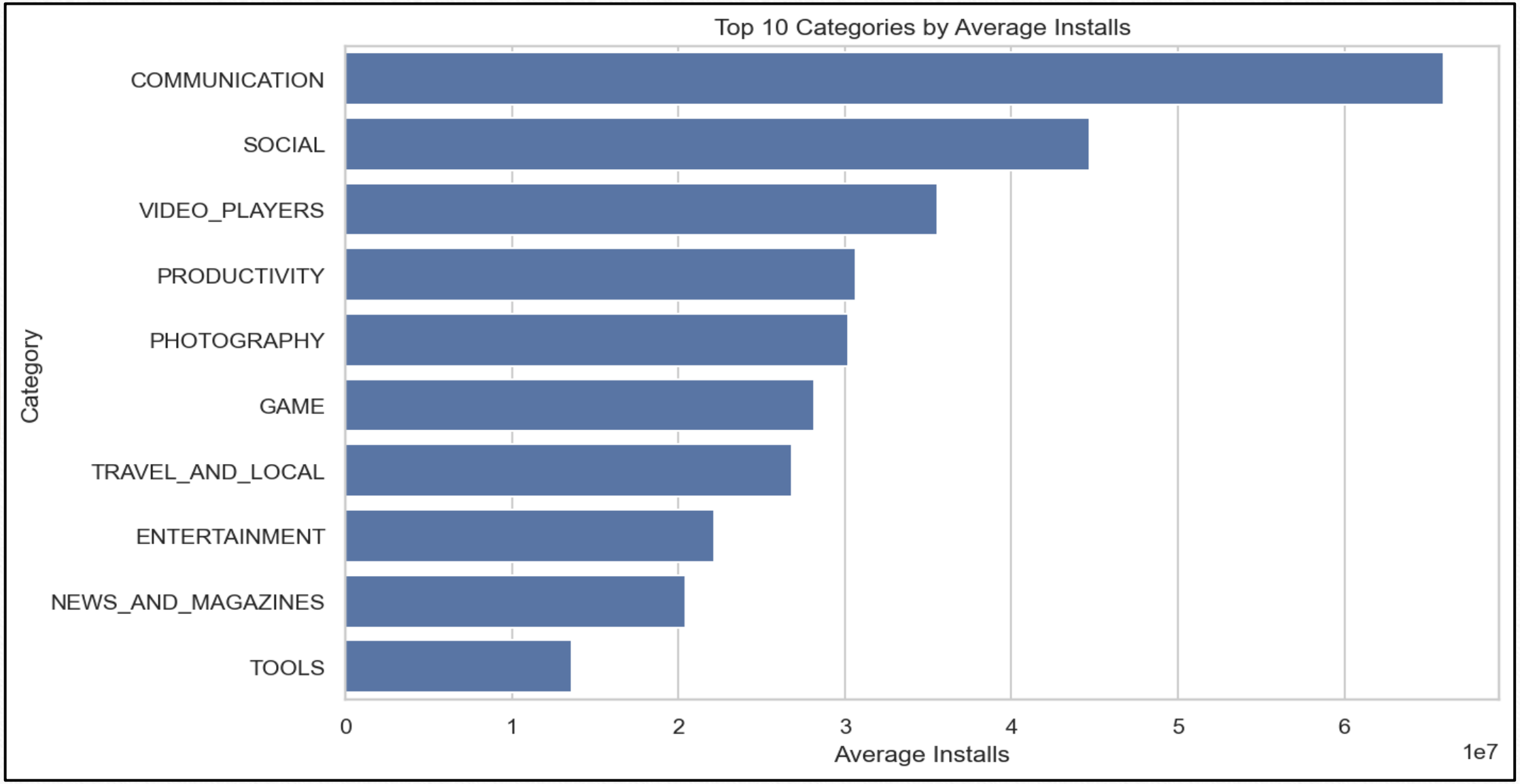
Installs by Rating Group

- High-rated apps (≥4) have significantly more installs
- Low-rated apps have limited reach


Installs by Rating Group (log scale)
Installs (log scale)
10^9
10^7
10^5
10^3
10^1
High (>=4)
Medium (3-4)
Low (<3)
Rating_Group

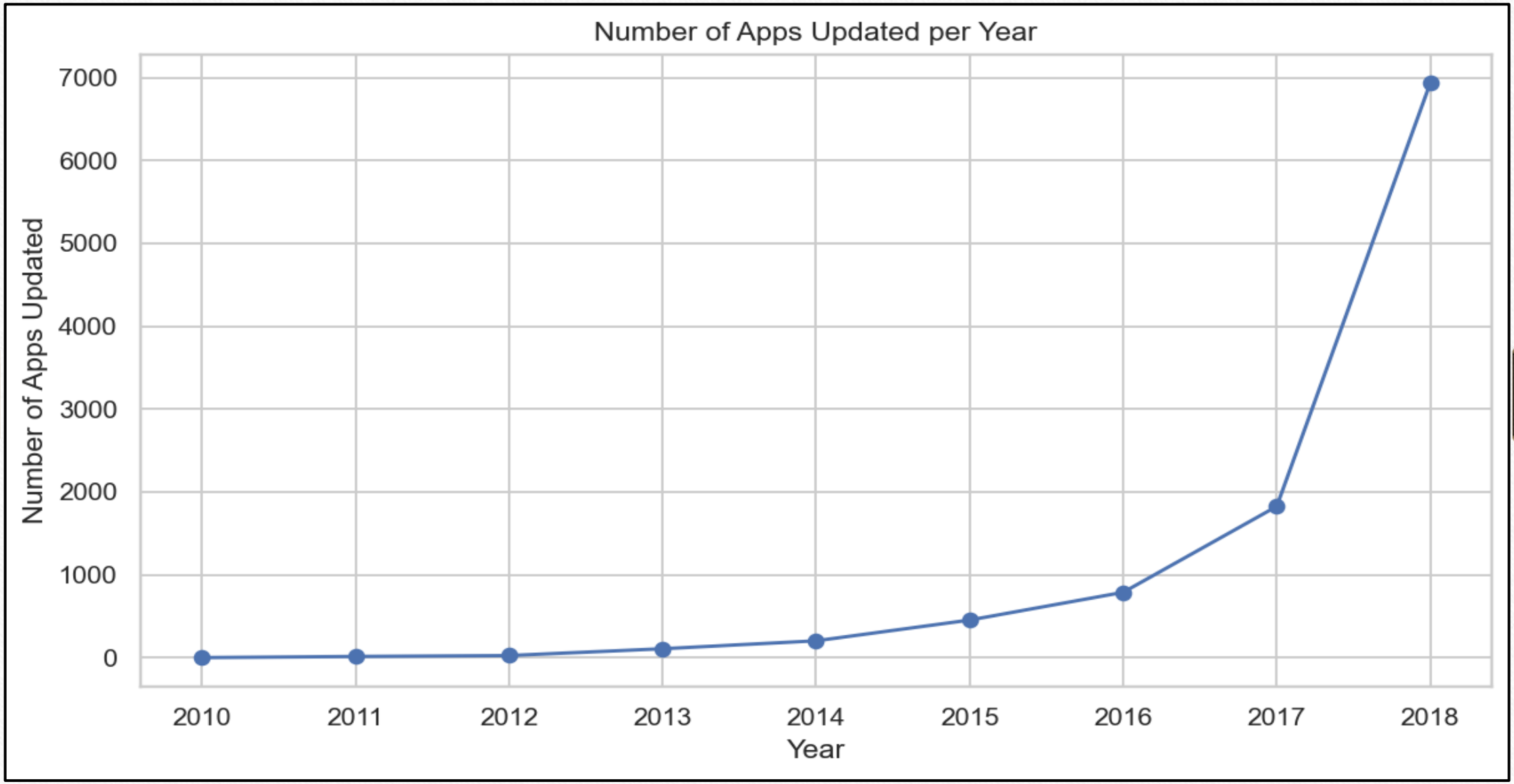
Top 10 Categories by Average Installs

- Game, Communication, and Social apps dominate install counts
- Utility apps have lower but consistent installs


Top 10 Categories by Average Installs
COMMUNICATION
SOCIAL
VIDEO_PLAYERS
PRODUCTIVITY
PHOTOGRAPHY
GAME
TRAVEL_AND_LOCAL
ENTERTAINMENT
NEWS_AND_MAGAZINES
TOOLS
Category
0
1
2
3
4
5
6
Average Installs
1e7

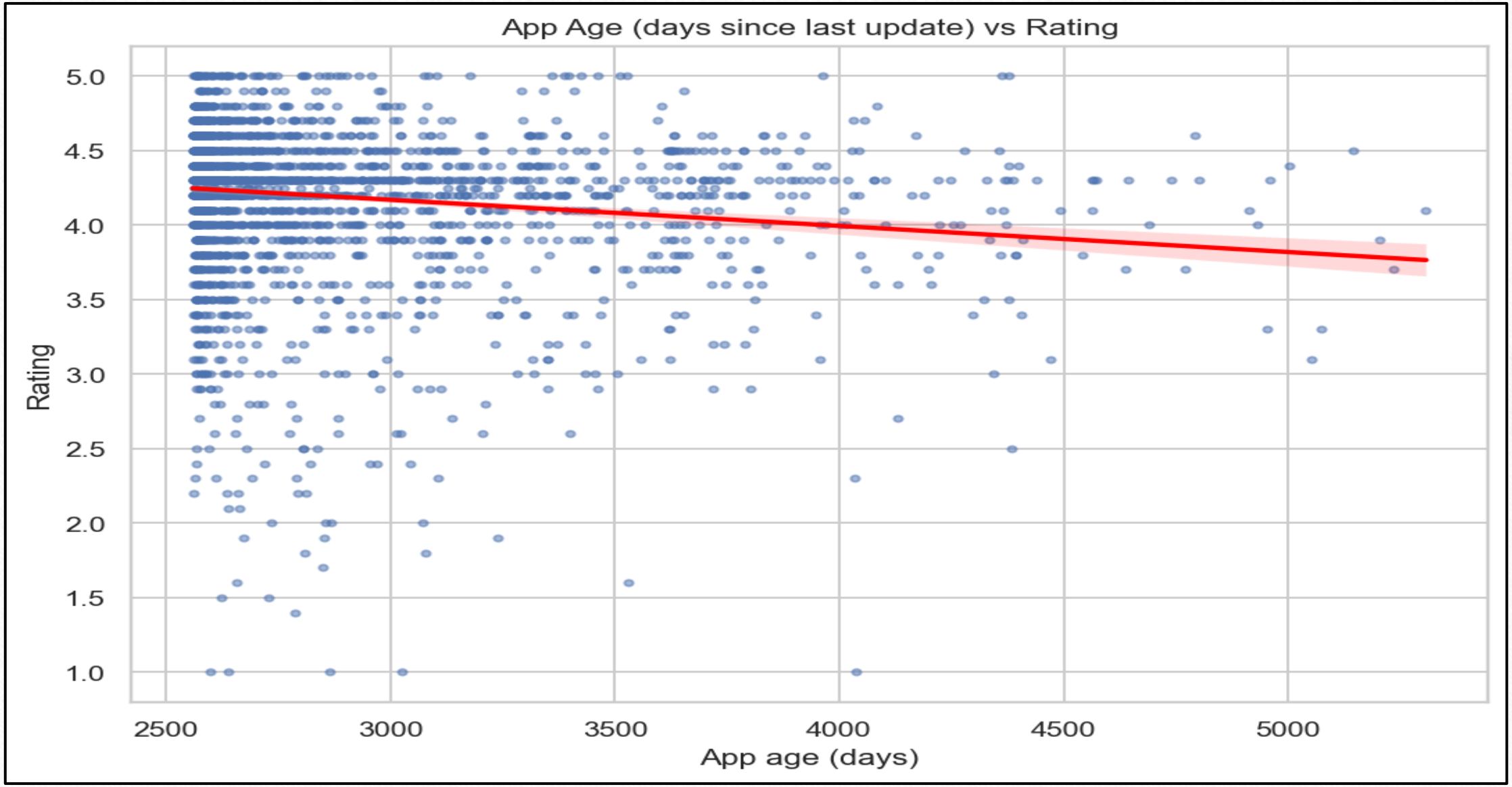
Apps Updated per Year

- Peaks in recent years — indicates active development
- Some old apps remain un-updated for long


Number of Apps Updated per Year
Number of Apps Updated
7000
6000
5000
4000
3000
2000
1000
0
2010
2011
2012
2013
2014
2015
2016
2017
2018
Year

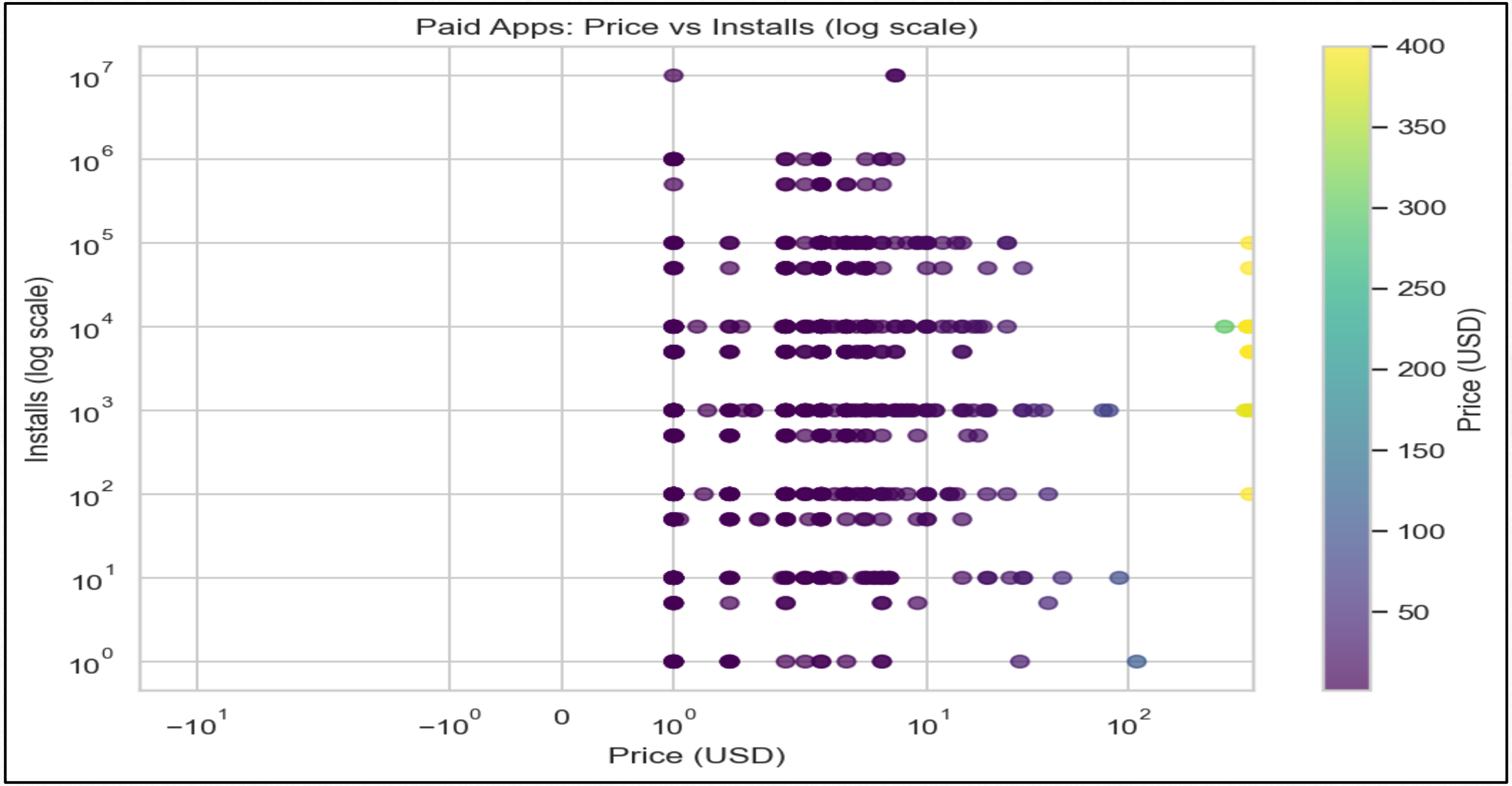
App Age vs Rating

- Slight negative correlation: older apps tend to have lower ratings
- Frequent updates may boost ratings


App Age (days since last update) vs Rating
Rating
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
2500
3000
3500
4000
4500
5000
App age (days)

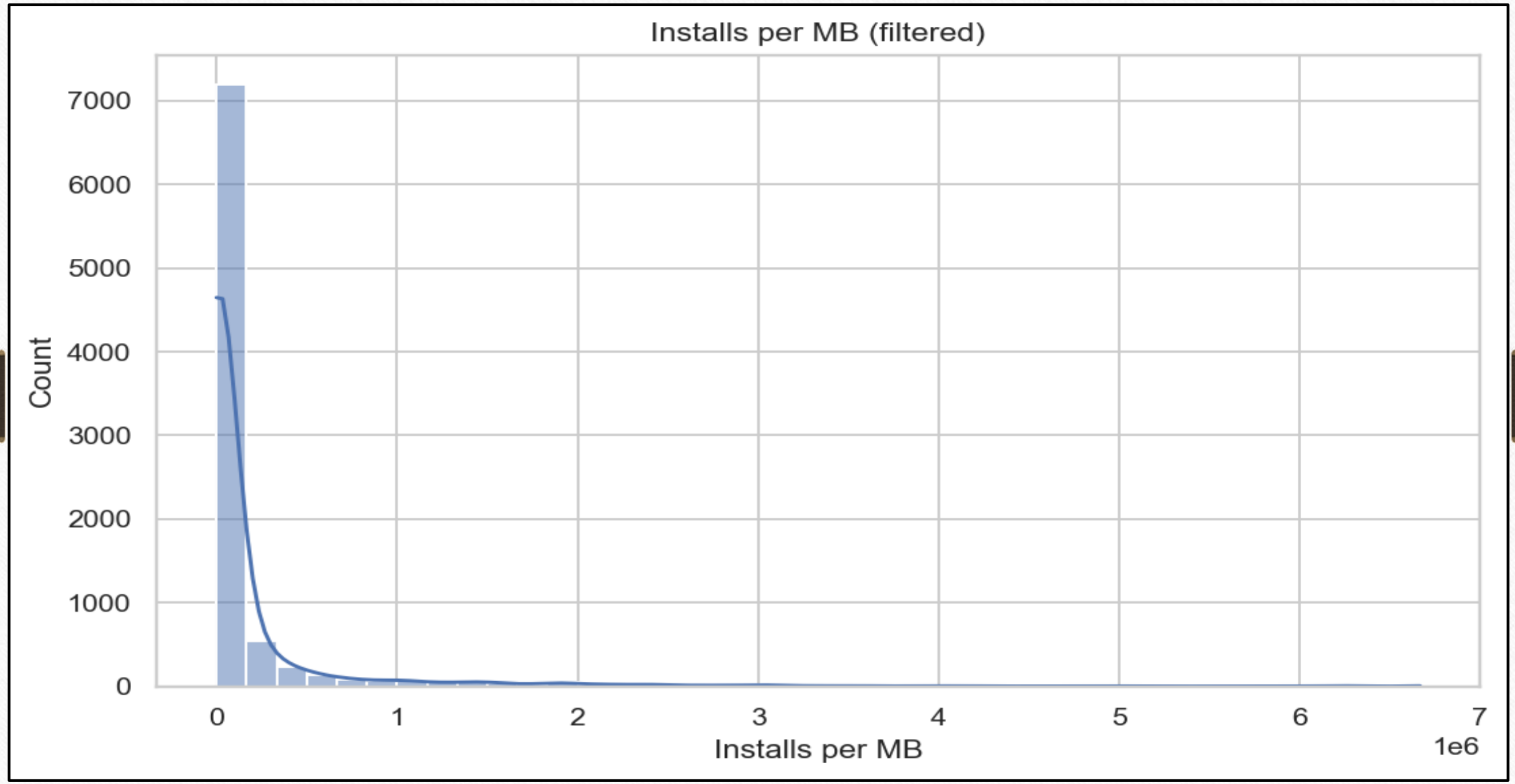
Paid Apps: Price vs Installs

- Higher-priced apps have very low installs
- Most paid apps cluster below $10


Paid Apps: Price vs Installs (log scale)
Installs (log scale)
10^7
10^6
10^5
10^4
10^3
10^2
10^1
10^0
-10^1
-10^0
0
10^0
10^1
10^2
Price (USD)
400
350
300
250
200
150
100
50
Price (USD)

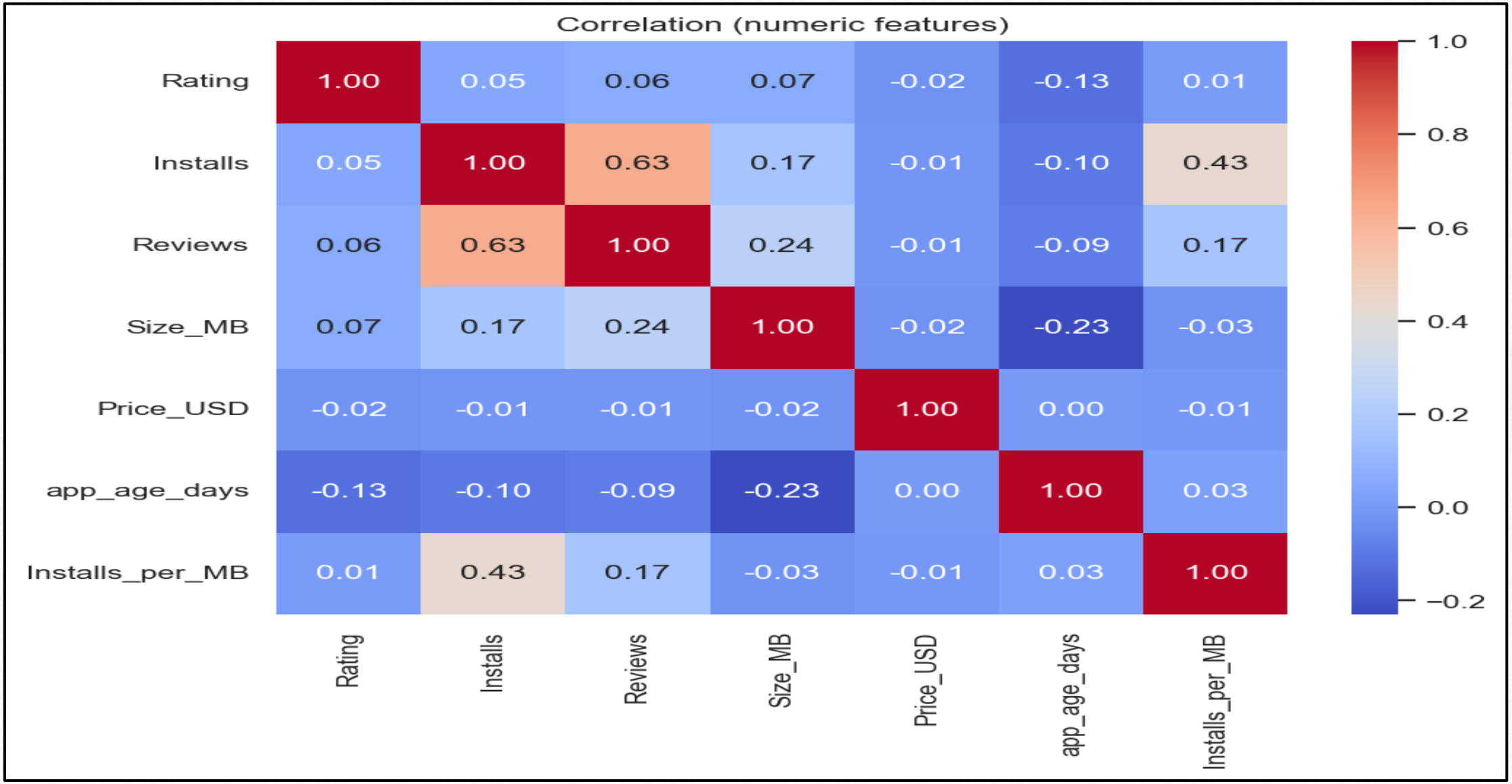
Installs per MB Distribution

- Some apps achieve very high installs per MB (efficient)
- Large inefficient apps may struggle with adoption


Installs per MB (filtered)
Count
7000
6000
5000
4000
3000
2000
1000
0
0
1
2
3
4
5
6
7
Installs per MB
1e6

Numeric Correlation Heatmap

- Reviews and Installs strongly correlated
- Price has negative correlation with installs


Correlation (numeric features)
Rating
Installs
Reviews
Size_MB
Price_USD
app_age_days
Installs_per_MB
1.00 0.05 0.06 0.07 -0.02 -0.13 0.01
0.05 1.00 0.63 0.17 -0.01 -0.10 0.43
0.06 0.63 1.00 0.24 -0.01 -0.09 0.17
0.07 0.17 0.24 1.00 -0.02 -0.23 -0.03
-0.02 -0.01 -0.01 -0.02 1.00 0.00 -0.01
-0.13 -0.10 -0.09 -0.23 0.00 1.00 0.03
0.01 0.43 0.17 -0.03 -0.01 0.03 1.00
1.0
0.8
0.6
0.4
0.2
0.0
−0.2

Conclusion & Recommendations

- Developers should focus on user ratings & frequent updates
- Consider app size optimization for wider reach
- Paid apps should balance price & value

Thank you!