Google Play Store Content Analysis

Presented by Sana Khan
Data Science Intern, Unified Mentor

ABOUT GOOGLE PLAYSTORE

Google Play

- Launched by Google in 2012 as the official Android app store
- Hosts over 3.5 million apps across diverse categories
- Available in 190+ countries with 2.5+ billion Android users
- Offers apps, games, movies, books, and more
- Supports Free, Freemium, Paid apps and in-app purchases
- Uses Google Play Protect for app security and safety
- Provides user ratings and reviews for quality feedback
- Enables developers to publish via Google Play Console
- Major source of global digital revenue for app developers

PROJECT OBJECTIVE

- Analyze app ratings, installs, and categories
- Understand pricing trends and app types
- Explore user preferences by category and type
- Identify the most popular and best-rated apps

PROBLEM STATEMENTS

- Which are the most common app categories?
- Which categories receive the most installs?
- What is the rating distribution of apps?

- Are paid apps rated better than free ones?
- What content rating is most common?
- How does app size affect rating?

Google Play

TOOLS, LIBRARIES & PROJECT WORKFLOW 📊

1

🔧 Tools & Libraries
Python – Core programming language
pandas – Data manipulation & preprocessing
matplotlib – Data visualization
seaborn – Statistical data visualization
Jupyter Notebook – Interactive coding environment

2

📈 Project Workflow

- Load the Dataset – Using pandas.read_csv()
- & Data Cleaning
  - Removed duplicates and null values
  - Converted data types (e.g., Installs, Price)
  - Cleaned invalid entries (e.g., Rating > 5)
-

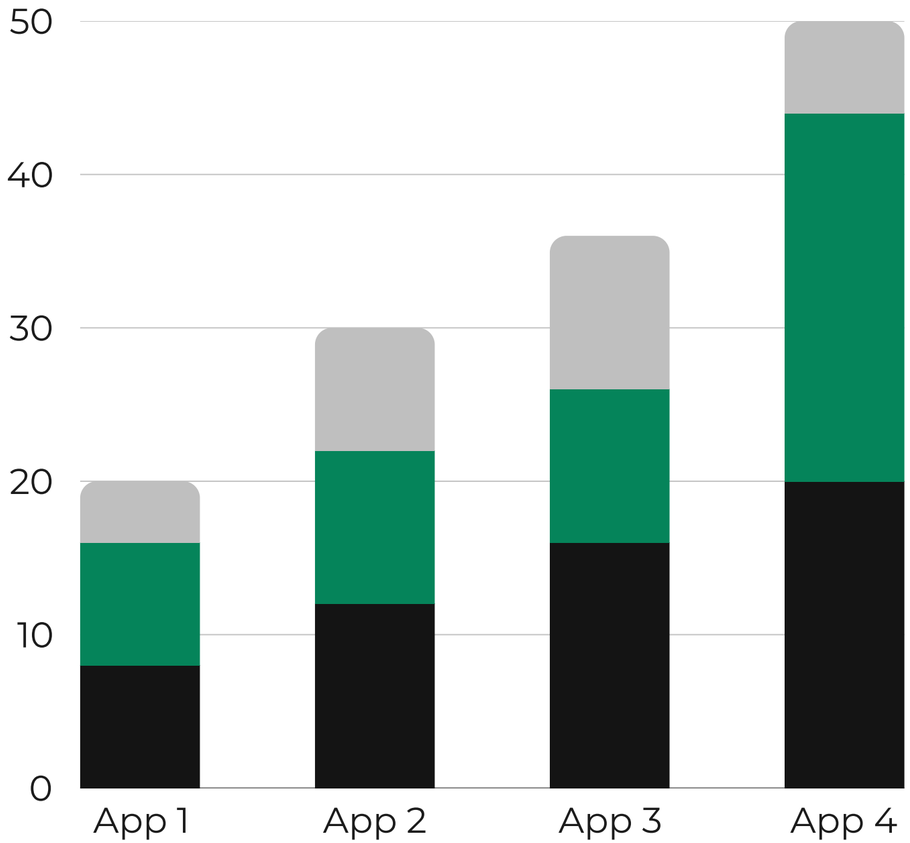
3

- Exploratory Data Analysis (EDA)
- Visualized distribution of Ratings, Installs, Categories
- Analyzed Free vs Paid apps, Content Ratings, App Sizes

4

- Insight Extraction
- Identified most installed & highest rated app categories
- Observed user preferences by content rating and app type
- 5. Visualization
- Plots: Bar charts, histograms, boxplots, heatmaps

DATA SET OVERVIEW



- **Source: Internship dataset provided by Unified Mentor**
- **Records: ~10,000 apps**
- **Fields: App name, Category, Rating, Reviews, Size, Installs, Price, Type, Content Rating, etc.**

IMPORT LIBRARIES AND LOAD THE DATASET

```
]: # Importing necessary libraries
import pandas as pd  # For data manipulation
import numpy as np  # For numerical operations
import matplotlib.pyplot as plt  # For visualization
import seaborn as sns  # For better-looking plots

# Setting styles for plots
sns.set(style="darkgrid")
```

```
: # Load the dataset
df = pd.read_csv("googleplaystore (1).csv")

# Display the first few rows to understand the data
df.head()
```

Out[18]:

	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

DATA CLEANING

```
]: # Get basic information about the dataset
   df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10841 entries, 0 to 10840
Data columns (total 13 columns):
 #   Column          Non-Null Count  Dtype  
---  ------          --------------  -----  
 0   App             10841 non-null  object 
 1   Category        10841 non-null  object 
 2   Rating          9367 non-null   float64
 3   Reviews         10841 non-null  object 
 4   Size            10841 non-null  object 
 5   Installs        10841 non-null  object 
 6   Type            10840 non-null  object 
 7   Price           10841 non-null  object 
 8   Content Rating  10840 non-null  object 
 9   Genres          10841 non-null  object 
 10  Last Updated    10841 non-null  object 
 11  Current Ver     10833 non-null  object 
 12  Android Ver     10838 non-null  object 
dtypes: float64(1), object(12)
memory usage: 1.1+ MB
```

```
]: # Check for missing values in each column
   df.isnull().sum()
```

```
]: App                0
   Category           0
   Rating          1474
   Reviews            0
   Size               0
   Installs           0
   Type               1
   Price              0
   Content Rating     1
   Genres             0
   Last Updated       0
   Current Ver        8
   Android Ver        3
   dtype: int64
```

```
]: # Check for duplicate rows
   df.duplicated().sum()
```

```
]: 483
```

```
# Summary of numerical columns
df.describe()
```

	Rating
count	9367.000000
mean	4.193338
std	0.537431
min	1.000000
25%	4.000000
50%	4.300000
75%	4.500000
max	19.000000

```
df.dropna(inplace=True)
```

```
df.fillna(value="Unknown", inplace=True)  # Example for categorical columns
```

```
df.dtypes
```

```
App                object
Category           object
Rating            float64
Reviews            object
Size               object
Installs           object
Type               object
Price              object
Content Rating     object
Genres             object
Last Updated       object
Current Ver        object
Android Ver        object
dtype: object
```

```
[6]: df['Reviews'] = df['Reviews'].astype(int)  # Convert to integer
     df['Installs'] = df['Installs'].str.replace(',', '').str.replace('+', '').astype(int)  # Clean and convert

[1]: df.drop_duplicates(inplace=True)

[2]: df.info()
     df.head()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 8886 entries, 0 to 10840
Data columns (total 13 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   App             8886 non-null   object
 1   Category        8886 non-null   object
 2   Rating          8886 non-null   float64
 3   Reviews         8886 non-null   int32
 4   Size            8886 non-null   object
 5   Installs        8886 non-null   int32
 6   Type            8886 non-null   object
 7   Price           8886 non-null   object
 8   Content Rating  8886 non-null   object
 9   Genres          8886 non-null   object
 10  Last Updated    8886 non-null   object
 11  Current Ver     8886 non-null   object
 12  Android Ver     8886 non-null   object
dtypes: float64(1), int32(2), object(10)
memory usage: 902.5+ KB
```

]:

	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	10000	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	500000	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	5000000	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	50000000	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	100000	Free	0	Everyone	Art & Design;Creativity	June 20, 2018	1.1	4.4 and up

]:
```
df.describe()
```

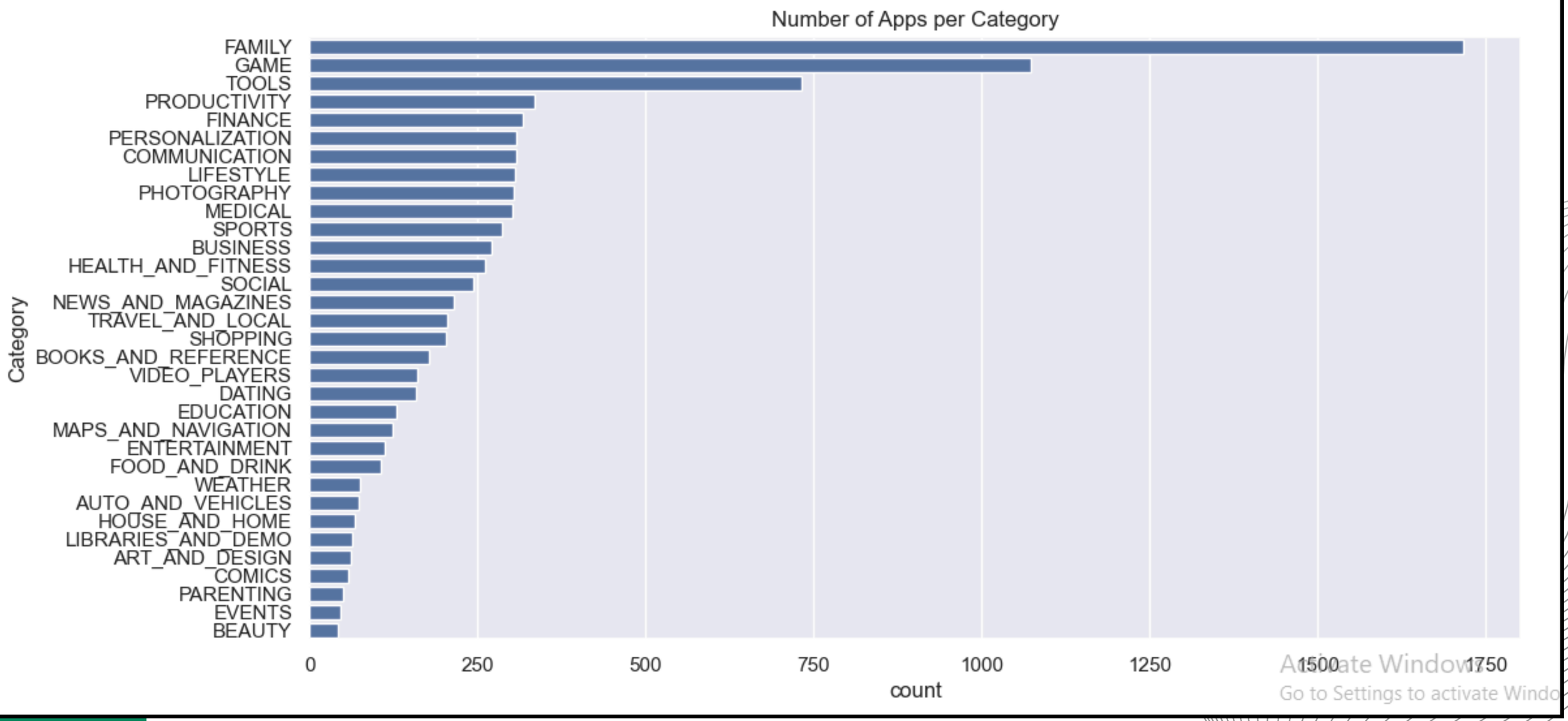
]:

	Rating	Reviews	Installs
count	8886.000000	8.886000e+03	8.886000e+03
mean	4.187959	4.730928e+05	1.650061e+07
std	0.522428	2.906007e+06	8.640413e+07
min	1.000000	1.000000e+00	1.000000e+00
25%	4.000000	1.640000e+02	1.000000e+04
50%	4.300000	4.723000e+03	5.000000e+05
75%	4.500000	7.131325e+04	5.000000e+06
max	5.000000	7.815831e+07	1.000000e+09

```
df['Category'].value_counts()
```

```
Category
FAMILY                 1717
GAME                   1074
TOOLS                   733
PRODUCTIVITY            334
FINANCE                 317
PERSONALIZATION         308
COMMUNICATION           307
LIFESTYLE               305
PHOTOGRAPHY             304
MEDICAL                 302
SPORTS                  286
BUSINESS                270
HEALTH_AND_FITNESS      262
SOCIAL                  244
NEWS_AND_MAGAZINES      214
TRAVEL_AND_LOCAL        205
SHOPPING                202
BOOKS_AND_REFERENCE     177
VIDEO_PLAYERS           160
DATING                  159
EDUCATION               129
MAPS_AND_NAVIGATION     124
ENTERTAINMENT           111
FOOD_AND_DRINK          106
WEATHER                  75
AUTO_AND_VEHICLES        73
HOUSE_AND_HOME           68
LIBRARIES_AND_DEMO       64
ART_AND_DESIGN           61
COMICS                   58
PARENTING                50
EVENTS                   45
BEAUTY                   42
Name: count, dtype: int64
```

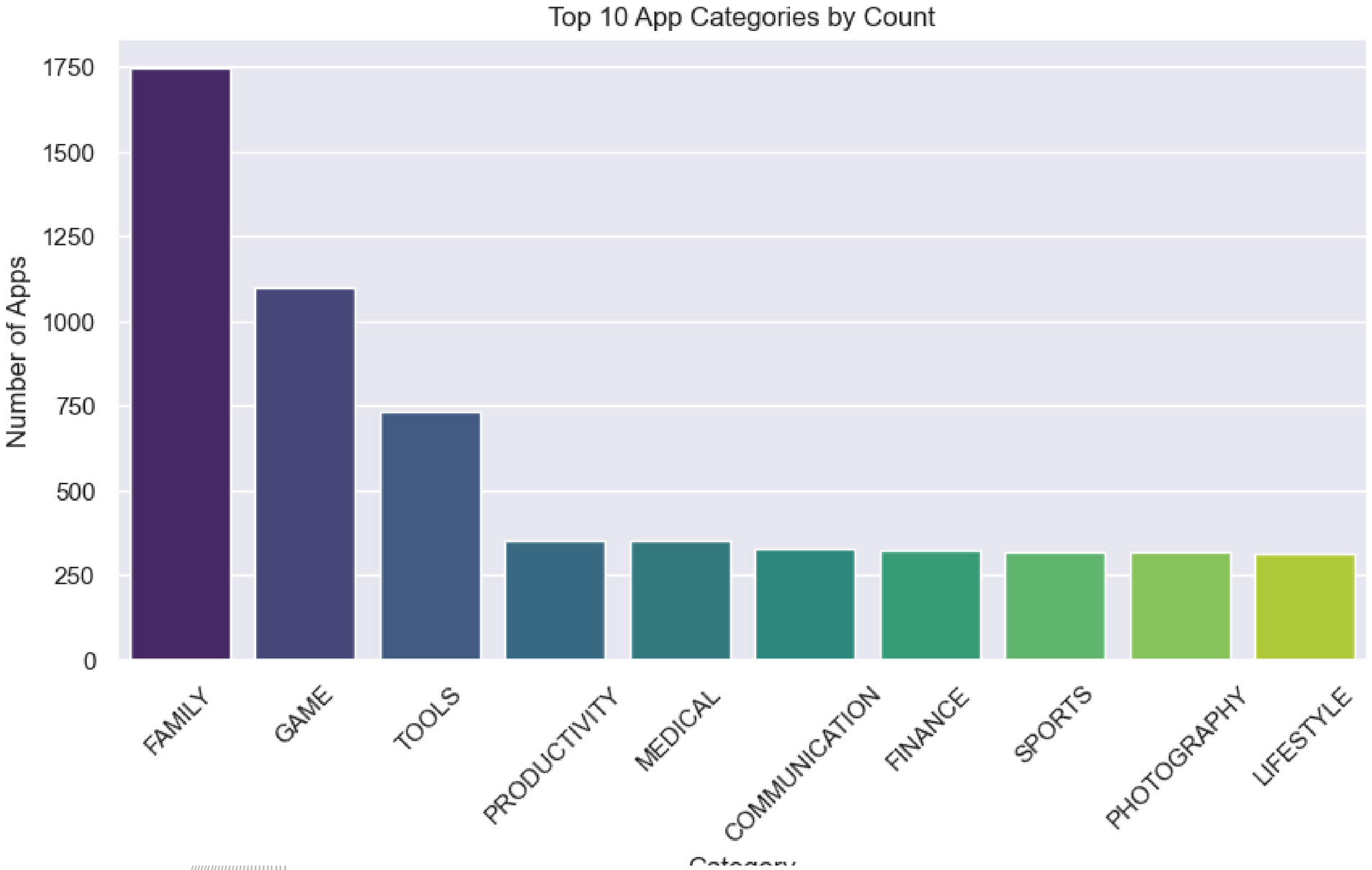
```
import matplotlib.pyplot as plt
import seaborn as sns

plt.figure(figsize=(12,6))
sns.countplot(y=df['Category'], order=df['Category'].value_counts().index)
plt.title('Number of Apps per Category')
plt.show()
```



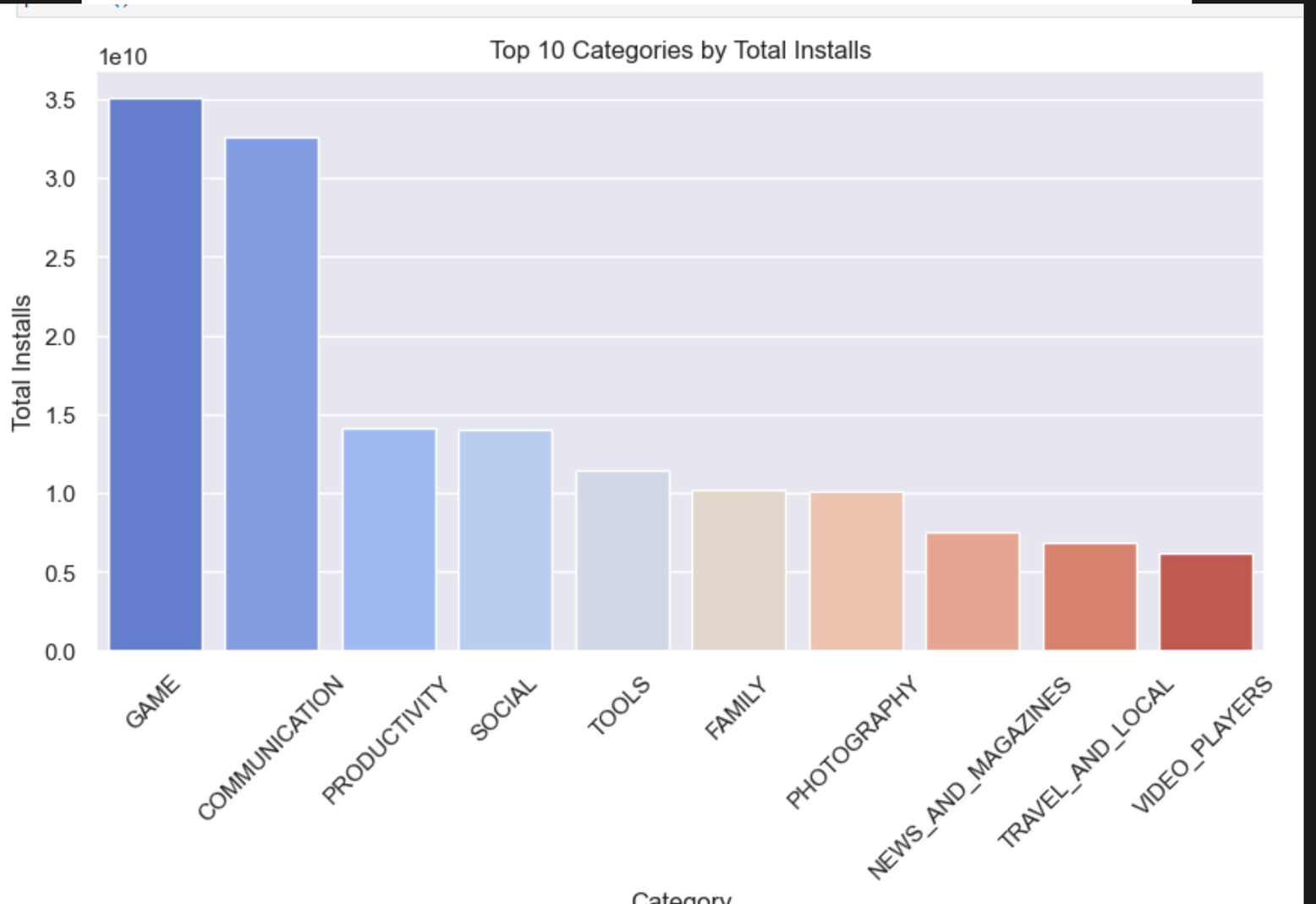
In [ ]:
```
sns.heatmap(df.corr(), annot=True, cmap="coolwarm")
plt.title("Correlation Matrix")
plt.show()
```

In [27]:
```
sns.boxplot(x=df['Reviews'])
plt.title("Boxplot of Reviews")
plt.show()
```

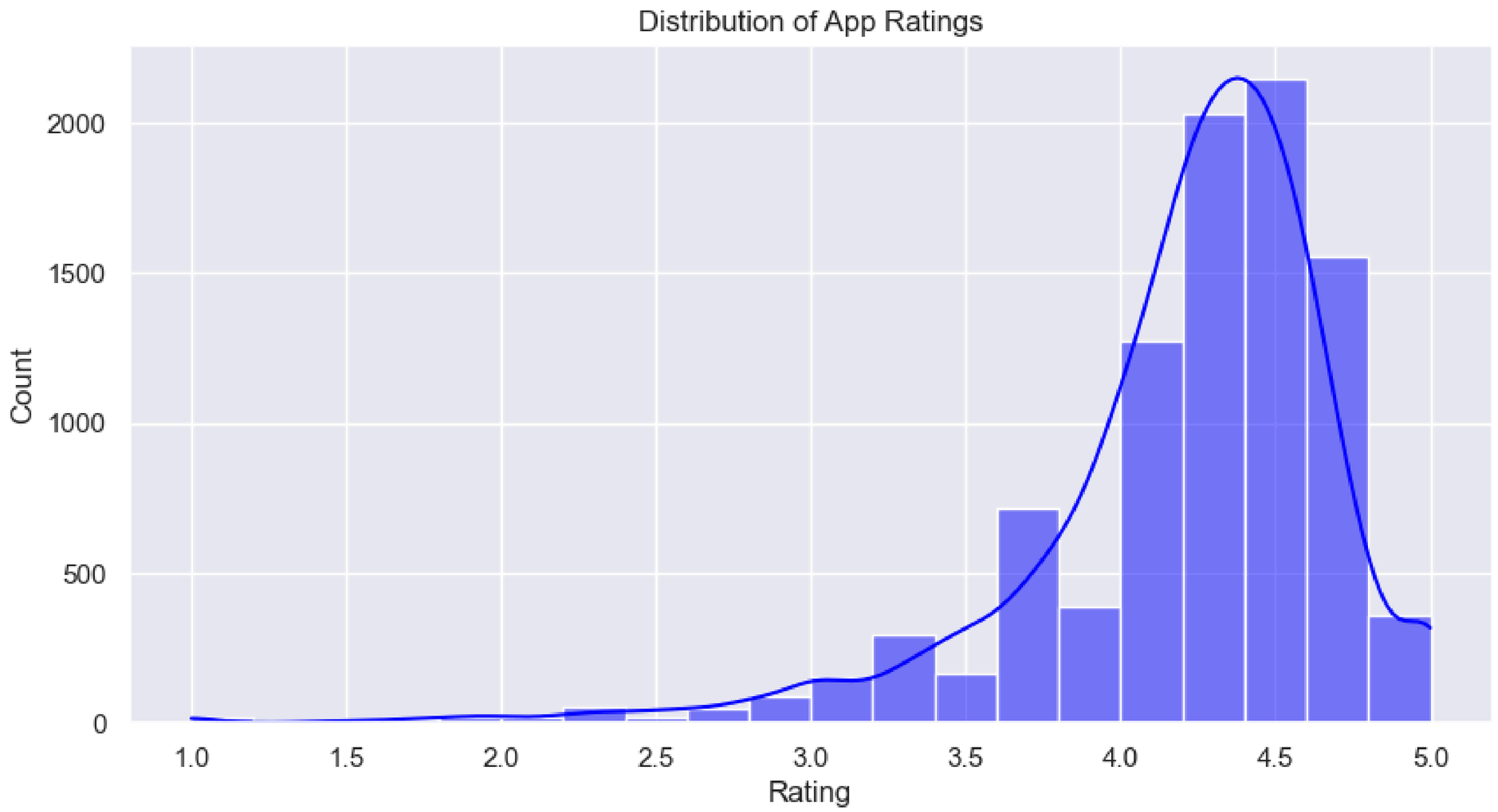
]:
```
top_categories = df['Category'].value_counts().head(10)
plt.figure(figsize=(10, 5))
sns.barplot(x=top_categories.index, y=top_categories.values, palette="viridis")
plt.xticks(rotation=45)
plt.title("Top 10 App Categories by Count")
plt.xlabel("Category")
plt.ylabel("Number of Apps")
plt.show()
```


Top 10 App Categories by Count
Number of Apps
1750
1500
1250
1000
750
500
250
0
FAMILY
GAME
TOOLS
PRODUCTIVITY
MEDICAL
COMMUNICATION
FINANCE
SPORTS
PHOTOGRAPHY
LIFESTYLE
Category

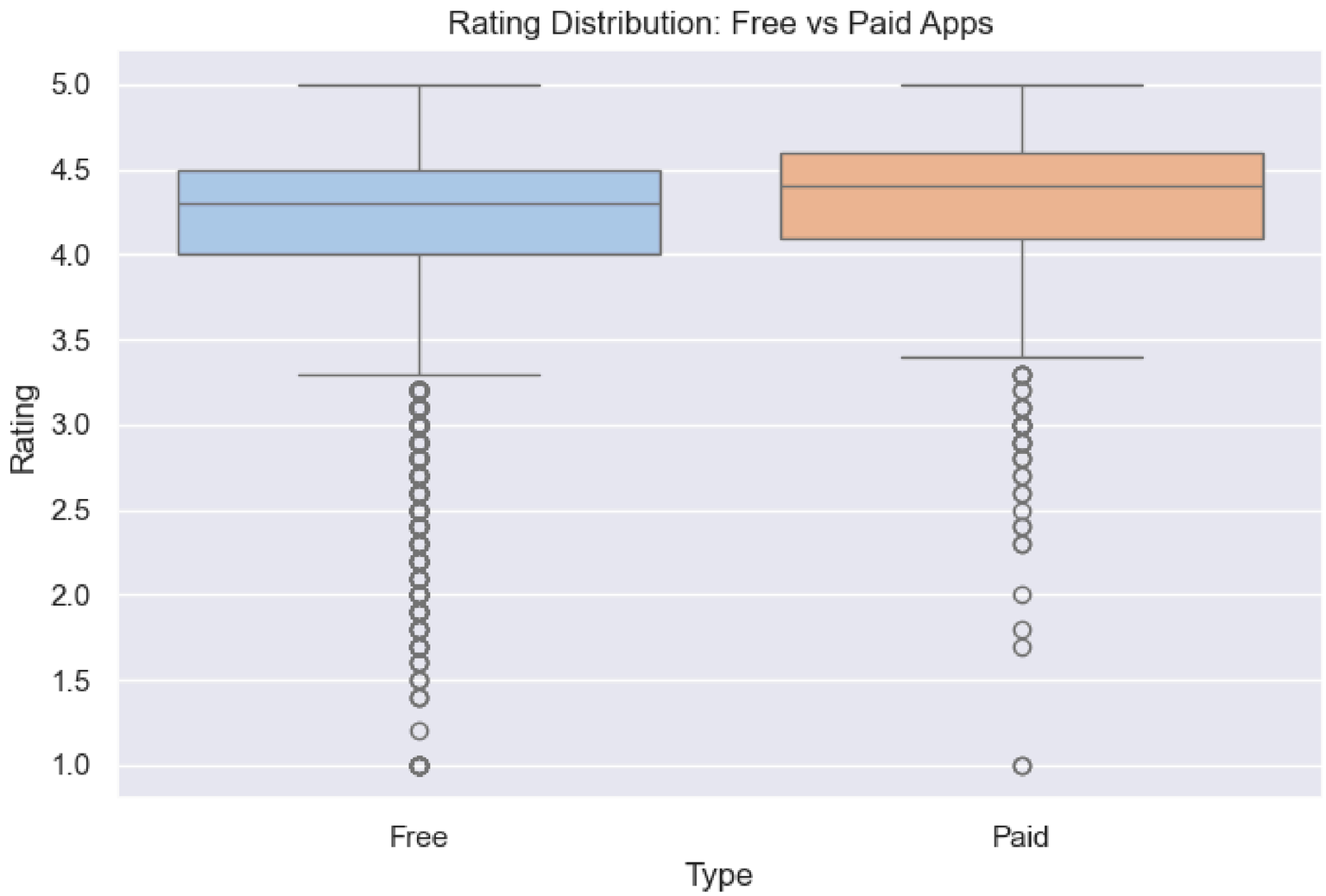
```
2]: top_installs = df.groupby('Category')['Installs'].sum().sort_values(ascending=False).head(10)
    plt.figure(figsize=(10, 5))
    sns.barplot(x=top_installs.index, y=top_installs.values, palette="coolwarm")
    plt.xticks(rotation=45)
    plt.title("Top 10 Categories by Total Installs")
    plt.xlabel("Category")
    plt.ylabel("Total Installs")
    plt.show()
```



```
]: plt.figure(figsize=(10, 5))
   sns.histplot(df['Rating'], bins=20, kde=True, color='blue')
   plt.title("Distribution of App Ratings")
   plt.xlabel("Rating")
   plt.ylabel("Count")
   plt.show()
```



In [34]:
```
plt.figure(figsize=(8, 5))
sns.boxplot(x=df['Type'], y=df['Rating'], palette="pastel")
plt.title("Rating Distribution: Free vs Paid Apps")
plt.show()
```

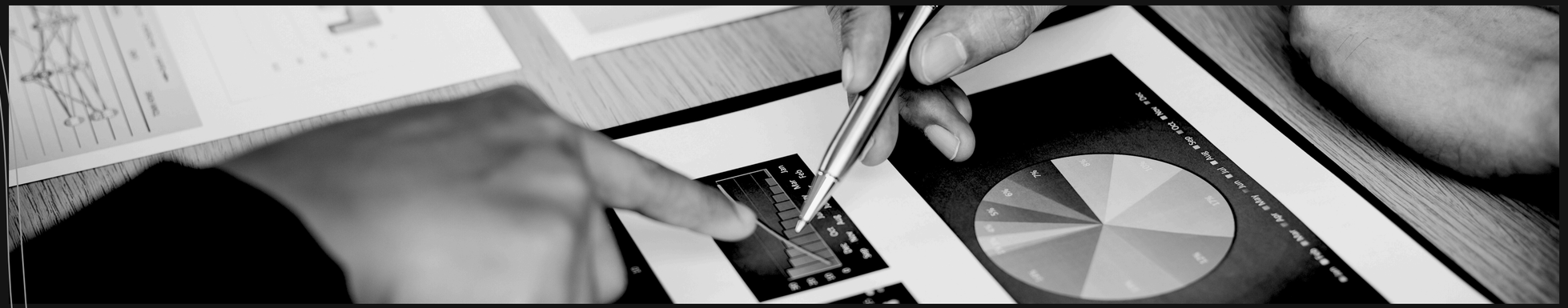


KEY INSIGHTS

- Free apps are 85%+ of the total
- Categories like Communication and Social lead in installs
- Paid apps rated slightly higher on average
- TV-MA rating most common for mature apps

RECOMMENDATIONS

- Target high-install categories
- Explore quality paid apps as a monetization strategy
- Prioritize app quality & ratings to build user trust
- Design apps for all age groups (Everyone rating)



GET IT ON Google Play

Skills Gained

- Data wrangling with pandas
- Data visualization with matplotlib & seaborn
- Statistical interpretation
- Presenting data-driven business insights

Challenges Faced

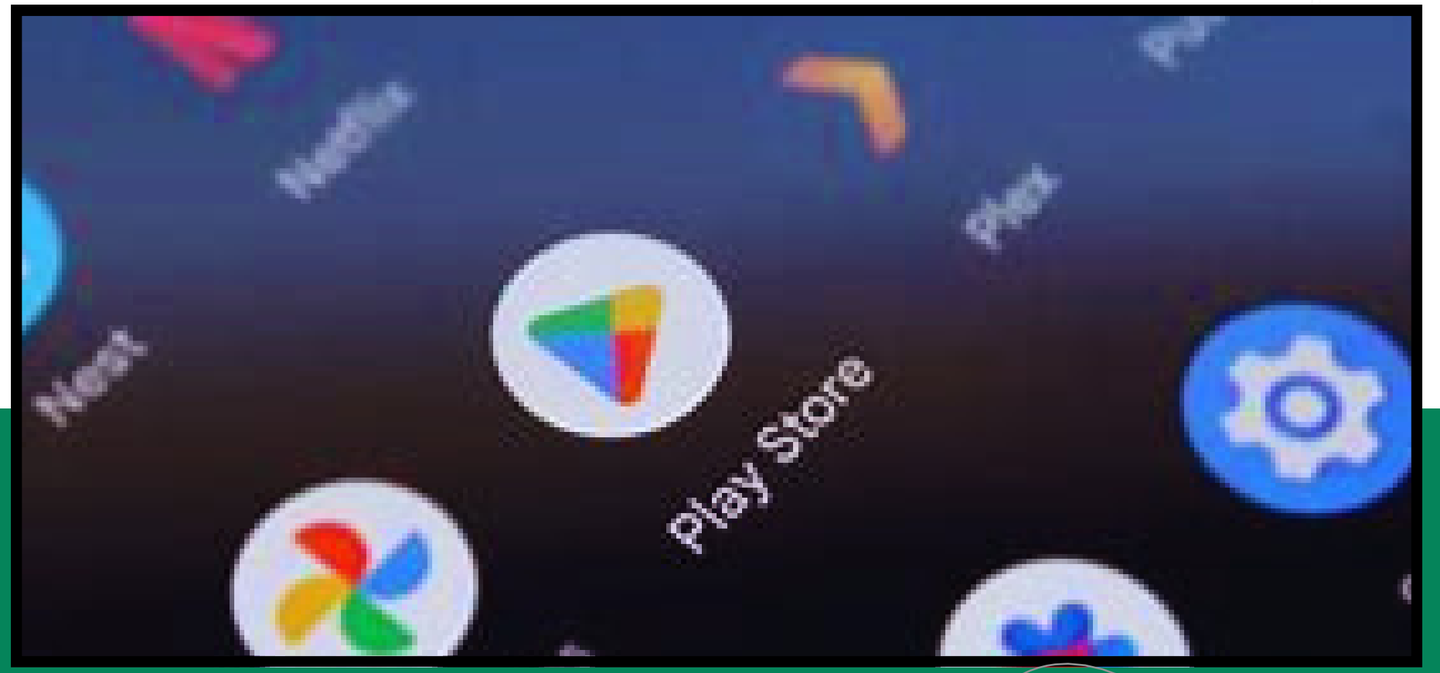
- Cleaning inconsistent install values
- Managing missing/blank fields
- Formatting issues in pricing and ratings
- Dealing with outliers and duplicates

CONCLUSIONS

- Valuable experience working with real-world data
- Gained confidence in handling messy datasets
- Learned to extract insights from app market data
- Understood how to apply EDA in practical scenarios

Google Play

THANK YOU



sanaekhan25@gmail.com