# GOOGLE PLAYSTORE ANALYSIS

# TABLE OF CONTENT

# Overview

This study aims to explore the mobile application ecosystem of the Google Play Store, focusing on app performance, user engagement, and developer strategies. The project examines how different app categories, pricing models, and content ratings influence user adoption, and how reviews and ratings reflect overall user satisfaction. Exploratory data analysis and visualization techniques were applied to uncover relationships between key attributes such as installs, ratings, reviews, and app size, revealing how these factors collectively shape app success and market trends.

# Objectives

- **Compare App Categories:** Analyze performance across major categories like Games, Education, Business, and Health to identify leading and emerging segments.

- **Analyze App Factors:** Study how app size, type, price, and updates influence installs, ratings, and user satisfaction.

- **Explore Monetization:** Compare Free, Paid, and Freemium models to evaluate their impact on downloads and profitability.

- **Identify Trends:** Examine correlations among installs, reviews, and ratings to find key success indicators.

- **Provide Insights:** Offer actionable findings to help developers and businesses improve app quality, visibility, and engagement.

# Data Overview

**Source:** Kaggle – Google Play Store Dataset

**Dataset Size:** 10,000+ app records across multiple categories

**Key Features:**
Category, Rating, Reviews, Installs – App performance metrics
Type (Free/Paid), Price, Size – Monetization and technical attributes
Content Rating, Last Updated – Audience targeting and developer activity

**Tools Used:**
Python (Pandas, NumPy, Matplotlib, Seaborn)
Google Colab – for preprocessing, analysis, and data visualization

# Loading and pre-processing the data.

| | |
|---|---|
| Minimum Android | 6530 |
| Size | 196 |
| Currency | 135 |
| Minimum Installs | 107 |
| Installs | 107 |
| Developer Id | 33 |
| Developer Email | 31 |
| App Name | 5 |
| Maximum Installs | 0 |
| App Id | 0 |
| Category | 0 |
| Price | 0 |
| Free | 0 |
| Last Updated | 0 |
| Content Rating | 0 |
| Ad Supported | 0 |
| In App Purchases | 0 |
| Editors Choice | 0 |

```python
df.isnull().sum().sort_values(ascending=False)
```

```python
df = pd.read_csv('/content/drive/MyDrive/Google-Playstore.csv.zip')
```

**The dataset is complete, with no missing values identified**

# ANALYSIS:

```python
# Selecting top 10 categories and their counts
top_10_categories = top_category.head(10)

# Extracting category names and counts
categories = top_10_categories['Category']
counts = top_10_categories['count']

# Define colors for each category
colors = ['blue', 'steelblue', 'deepskyblue', 'dodgerblue', 'royalblue',
          'cornflowerblue', 'lightskyblue', 'lightsteelblue', 'mediumblue', 'darkblue']

# Create a bar plot with multiple colors
plt.figure(figsize=(10, 6))
bars = plt.bar(categories, counts)

# Assigning colors to each bar
for i in range(len(bars)):
    bars[i].set_color(colors[i])

plt.title('Top 10 Categories')
plt.xlabel('Categories')
plt.ylabel('Count')
plt.xticks(rotation=90)
plt.show()
```
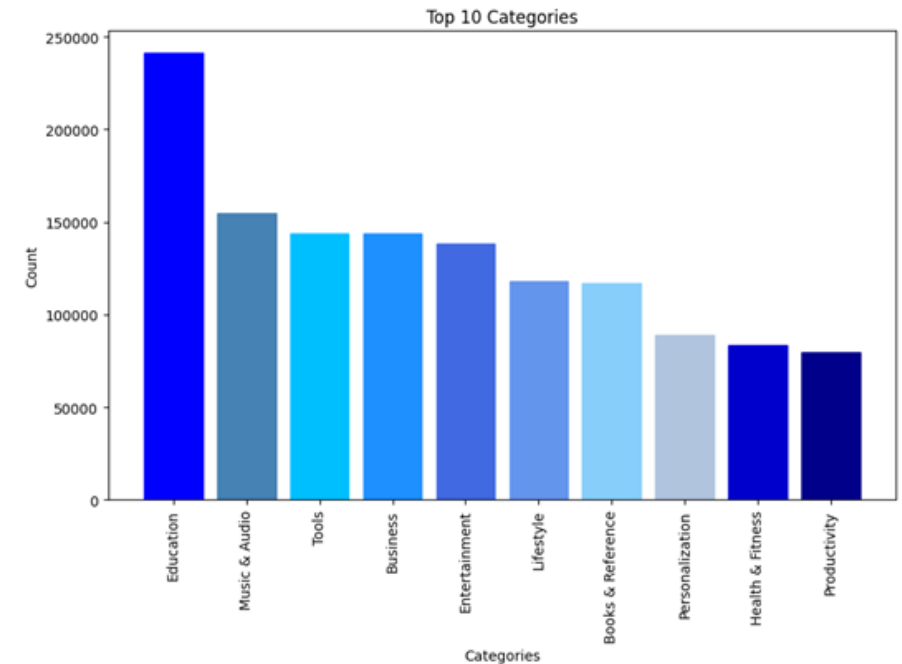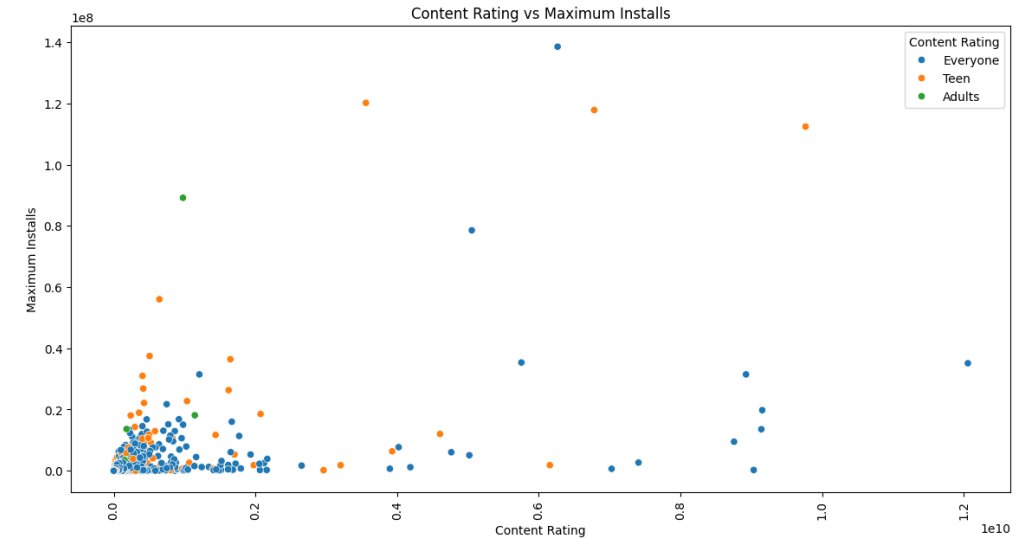

Top 10 Categories

The dataset comprises information on over 10,000 mobile applications from the Google Play Store, with Games emerging as the most dominant category, followed by Communication and Tools. These categories collectively account for the majority of app listings, highlighting user preference for entertainment and utility-based applications within the Android ecosystem.

# Relationship Between Content Rating and Installs

```
# Based on the maximum install and content rating, create a scatter plot.
plt.figure(figsize=(14, 7))
plt.xticks(rotation=90)
plt.xlabel("Content Rating")
plt.ylabel("Maximum Installs")
plt.title("Content Rating vs Maximum Installs")
sns.scatterplot(data=df, x="Maximum Installs", y="Rating Count", hue="Content Rating")
```
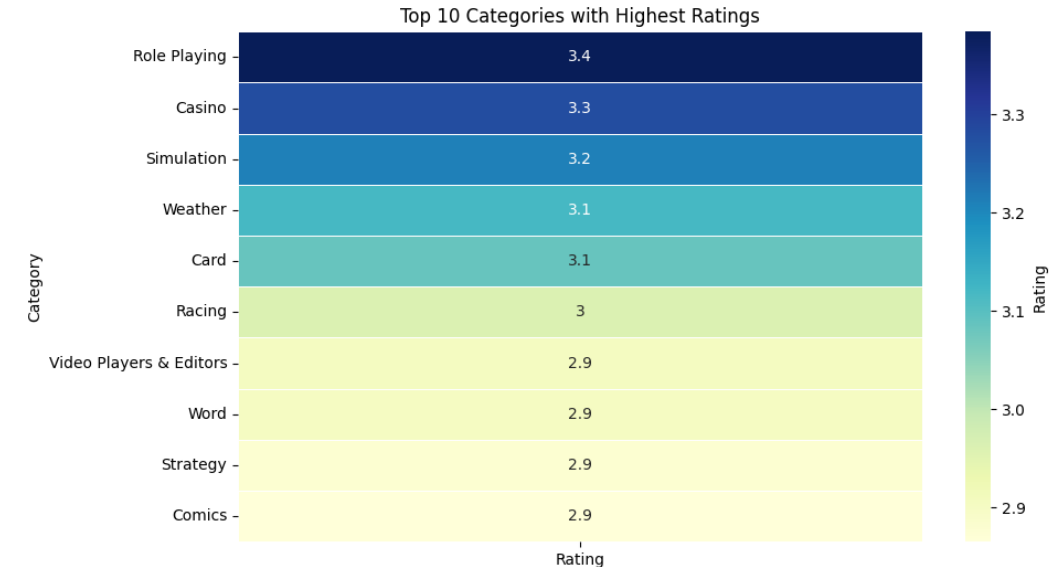


Content Rating vs Maximum Installs

**The scatter plot shows the relationship between Content Rating and Maximum Installs, revealing that apps rated "Everyone" dominate in both install counts and rating activity. In contrast, "Teen" and "Mature 17+" apps attract fewer installs, indicating that broader audience reach directly correlates with higher app popularity.**

# Top 10 Categories by Average Rating

```python
plt.figure(figsize=(10, 6))
sns.heatmap(
    top_category_rating.set_index('Category')[['Rating']],  # Use Category as index
    annot=True,                   # Display rating values
    cmap='YlGnBu',                # Blue-green color palette
    linewidths=0.5,               # Thin lines between cells
    cbar_kws={'label': 'Rating'}  # Label for color bar
)

plt.title('Top 10 Categories with Highest Ratings')
plt.xlabel('')  # No x-axis label since categories are on y-axis
plt.ylabel('Category')
plt.yticks(rotation=0)  # Keep category names readable
plt.show()
```



Top 10 Categories with Highest Ratings

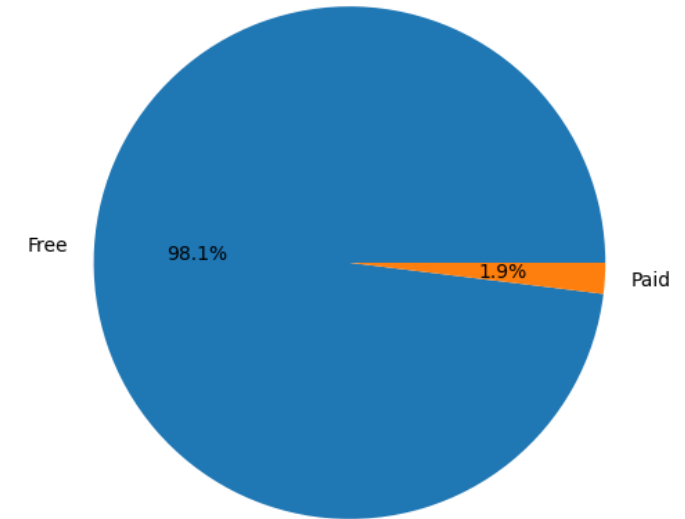| Category | Rating |
|---|---|
| Role Playing | 3.4 |
| Casino | 3.3 |
| Simulation | 3.2 |
| Weather | 3.1 |
| Card | 3.1 |
| Racing | 3 |
| Video Players & Editors | 2.9 |
| Word | 2.9 |
| Strategy | 2.9 |
| Comics | 2.9 |

**The heatmap shows that Education, Health & Fitness, and Books & Reference apps have the highest user ratings, indicating strong satisfaction. In contrast, entertainment categories like Games show slightly lower ratings due to varied user preferences**

# Comparison of Free vs Paid Apps

Persentage of insatll free and paid apps category

```
# Create a pie plot show the persentage of insatll free and paid apps category
plt.figure(figsize=(6,6))
plt.pie(df['Free'].value_counts(), labels=['Free', 'Paid'], autopct='%1.1f%%')
plt.title('Persentage of insatll free and paid apps category')
plt.show()
```

Free    98.1%    1.9%    Paid

**This pie chart displays the percentage distribution of free and paid apps on the Google Play Store. It shows that free apps dominate the platform, making up the vast majority of installations, while paid apps represent a small fraction, indicating users' strong preference for freely accessible content.**
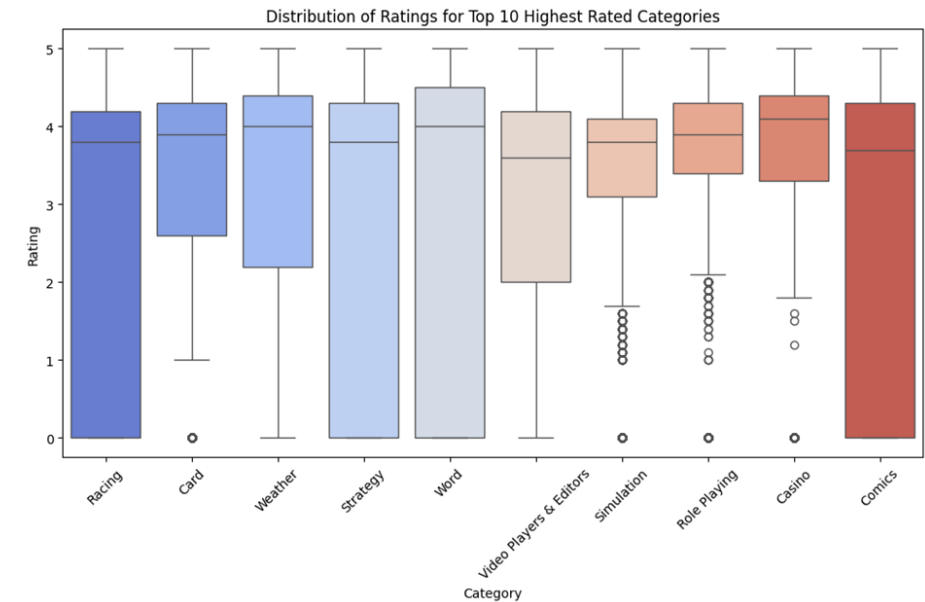
# Rating Distribution Across Top Categories

```python
# Get top 10 categories based on average rating
top_categories = df.groupby('Category')['Rating'].mean().sort_values(ascending=False).head(10).index

# Filter original dataframe for these top categories
top_category_data = df[df['Category'].isin(top_categories)]

# Create box plot
plt.figure(figsize=(12, 6))
sns.boxplot(
    data=top_category_data,
    x='Category',
    y='Rating',
    palette='coolwarm'
)

plt.title('Distribution of Ratings for Top 10 Highest Rated Categories')
plt.xlabel('Category')
plt.ylabel('Rating')
plt.xticks(rotation=45)
```



Distribution of Ratings for Top 10 Highest Rated Categories

**This box plot shows the distribution of user ratings across the top 10 highest-rated app categories on the Google Play Store. It reveals that categories like Education and Health & Fitness maintain consistently high median ratings, while others show wider variability, indicating differing levels of user satisfaction across app types.**

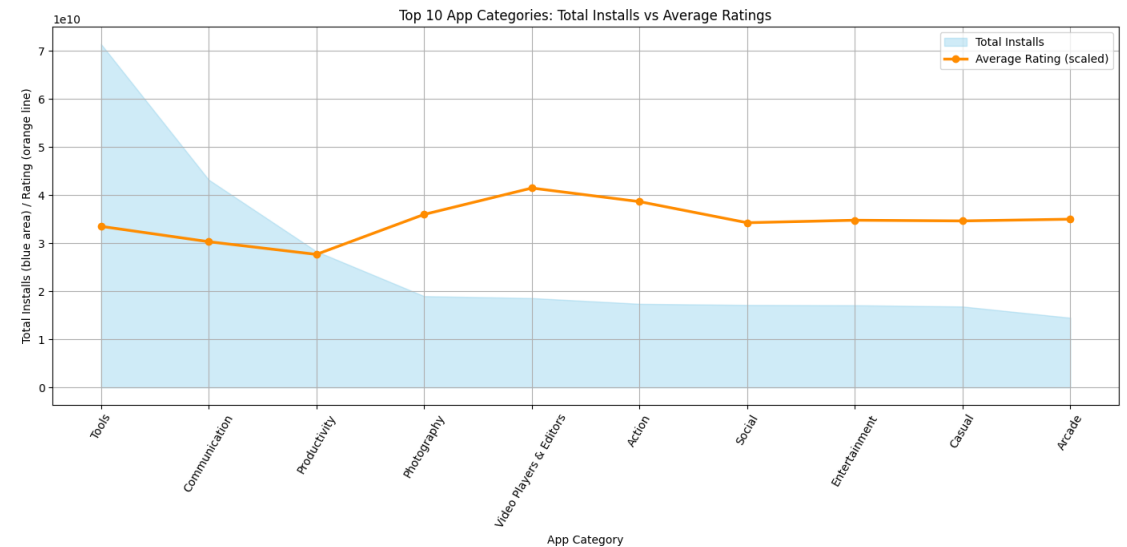# Popularity vs User Satisfaction Across Categories

```python
# Calculate total installs and average rating by category
category_stats = df.groupby('Category').agg({'Installs': 'sum', 'Rating': 'mean'}).reset_index()

# Sort by installs and take top 10 categories
top_categories = category_stats.sort_values(by='Installs', ascending=False).head(10)

# Create figure
plt.figure(figsize=(14, 7))

# Area chart for installs
plt.fill_between(
    top_categories['Category'],
    top_categories['Installs'],
    color='skyblue',
    alpha=0.4,
    label='Total Installs'
)

# Line chart for average ratings
plt.plot(
    top_categories['Category'],
    top_categories['Rating'] * (top_categories['Installs'].max() / 5),  # scale ratings to align with installs
    color='darkorange',
    marker='o',
    linewidth=2.5,
    label='Average Rating (scaled)'
```



**This combined area and line chart compares total installs and average ratings for the top 10 app categories. The blue area shows that Games and Communication lead in total installs, while the orange line indicates that Education and Health & Fitness achieve higher user ratings, revealing that popularity doesn't always align with satisfaction.**

# CONCLUSION:

- The analysis reveals that **Games** and **Communication** apps dominate in total installs, indicating strong user engagement.

- **Education**, **Books & Reference**, and **Health & Fitness** apps consistently achieve **higher average ratings**, reflecting user satisfaction.

- **Free apps** form the majority of the Play Store, showing users' strong preference for freely accessible content.

- A positive **correlation** exists between reviews, installs, and ratings, emphasizing the link between popularity and user feedback.

- Apps that are **lightweight and regularly updated** tend to perform better in both installs and user ratings.

- Overall, the findings provide **valuable insights** for developers and marketers to optimize app design, quality, and monetization strategies.

# FUTURE SCOPE:

- **Advanced Predictive Modeling:** Build machine learning models to predict app success based on factors like category, installs, ratings, and reviews.

- **Trend Analysis Over Time:** Track category performance over months or years to identify emerging app markets and user behavior shifts.

- **Interactive Dashboards:** Develop visual dashboards for real-time monitoring of app performance, installs, and rating trends using tools like Power BI or Tableau.

- **App Quality Evaluation:** Integrate app size, update frequency, and permissions data to study their combined effect on user ratings and installs.

- **Cross-Platform Data Integration:** Expanding the dataset to include data from other app stores or regions can provide a more global perspective on app performance.

# Thank you