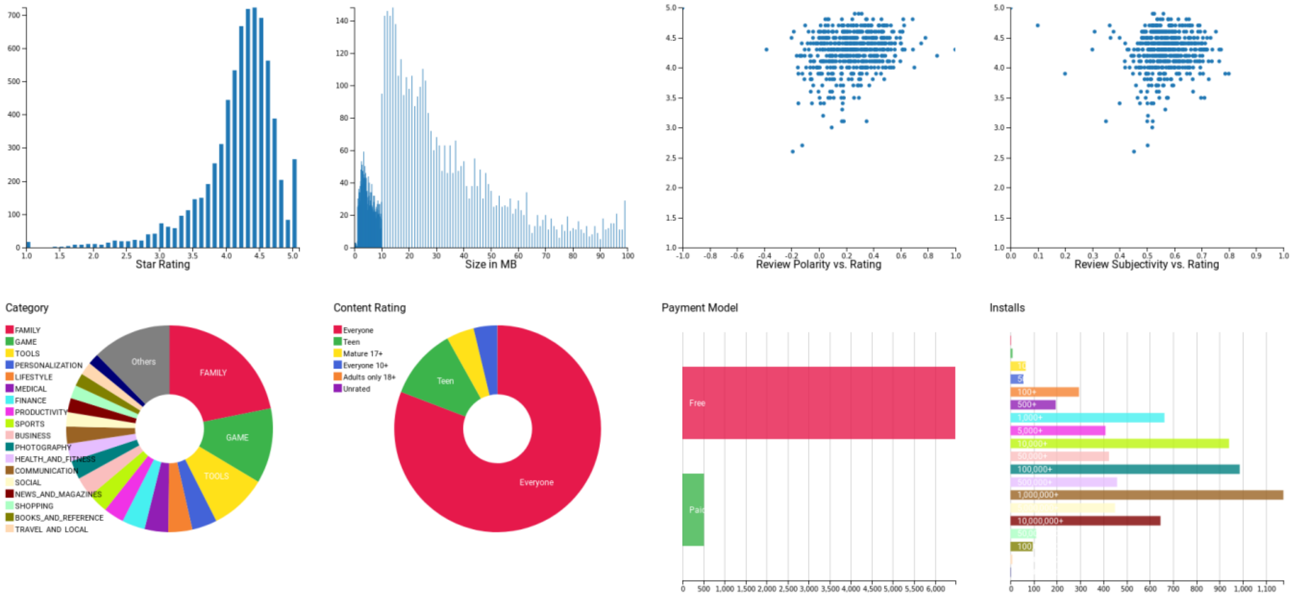
Exploratory Data Analysis of Google Play Store Apps

Tarun C
8th sem, , ECE
NITTE MEENAKSHI INSTITUTE OF TECHNOLOGY,
BANGALORE-560064



Abstract

The Google Play Store hosts millions of applications across various categories, catering to diverse user needs. This project aims to explore and analyze the Google Play Store dataset to uncover insights into app characteristics such as ratings, reviews, installs, pricing, and size.

The analysis began with extensive data cleaning, including the handling of missing values, duplicate removal, data type conversions, and transformation of categorical and numerical fields. Missing ratings were intelligently imputed based on install categories. Additionally, new features such as Size_MB and Installs_category were derived to facilitate deeper analysis.

Exploratory data analysis (EDA) and visualization techniques were then applied to reveal trends in app categories, rating distributions, relationships between reviews and installs, and genre performance. Correlation analysis helped identify strong relationships among numerical features, while category-wise insights highlighted the most installed, reviewed, and highest-rated genres.

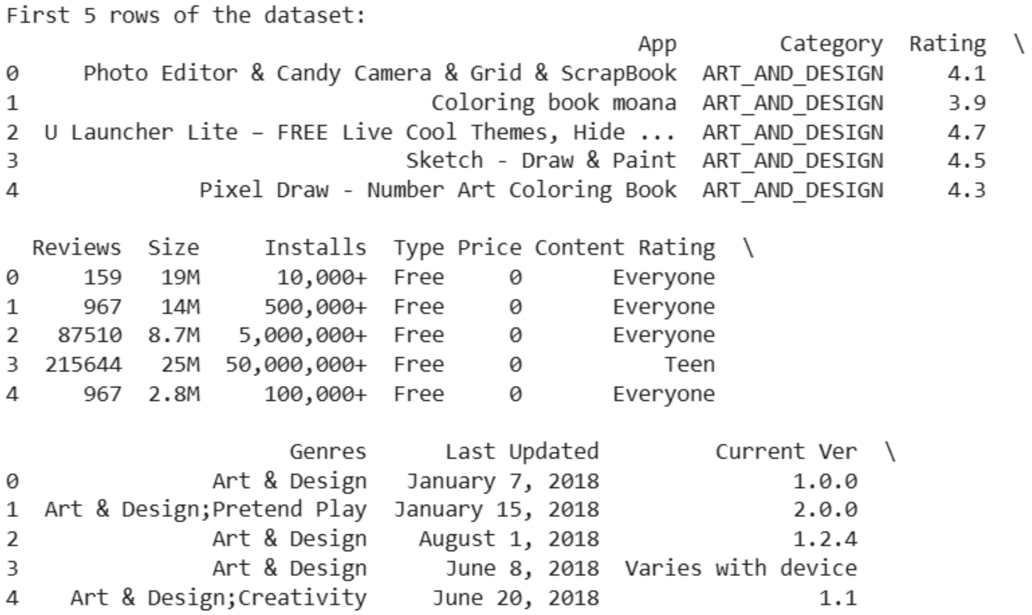
The findings provide valuable guidance for developers, marketers, and business strategists by identifying what characteristics contribute to app popularity and high user engagement on the platform.

KEYWORDS: Google Play Store, Exploratory Data Analysis, Data Cleaning, App Ratings, Installs, Reviews, App Categories, Python, Pandas, Seaborn, Correlation Analysis, Mobile Applications, Data Visualization, App Market Trends, Genre Performance

Data Collection

The dataset and the sources used for this process are listed below :

https://drive.google.com/drive/folders/13RYJ7YfjwlavX3Twg5KTR1DYg2_PVhca?usp=sharing

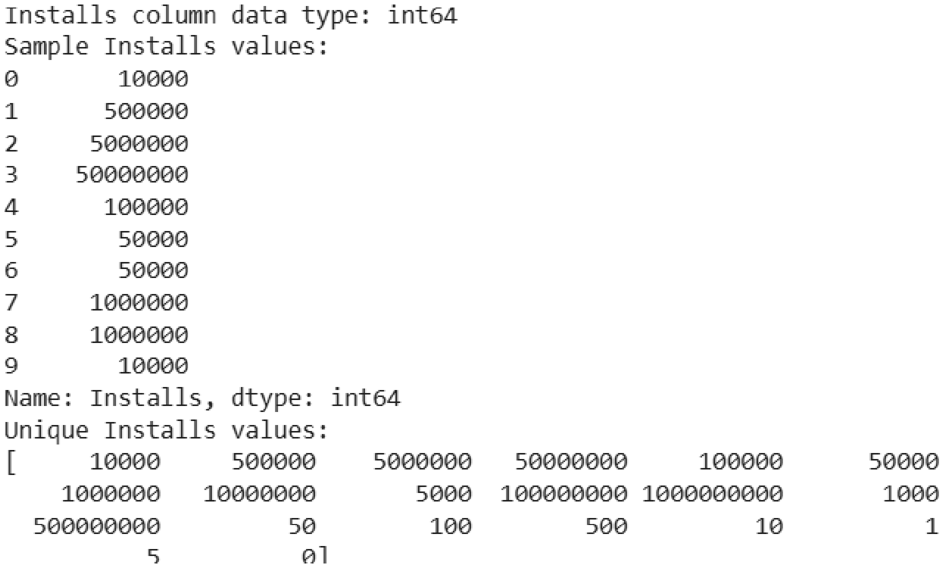


```
First 5 rows of the dataset:
                                                 App        Category  Rating  \
0     Photo Editor & Candy Camera & Grid & ScrapBook  ART_AND_DESIGN     4.1
1                                Coloring book moana  ART_AND_DESIGN     3.9
2  U Launcher Lite – FREE Live Cool Themes, Hide ...  ART_AND_DESIGN     4.7
3                              Sketch - Draw & Paint  ART_AND_DESIGN     4.5
4              Pixel Draw - Number Art Coloring Book  ART_AND_DESIGN     4.3

  Reviews  Size     Installs  Type Price Content Rating  \
0     159   19M      10,000+  Free     0       Everyone
1     967   14M     500,000+  Free     0       Everyone
2   87510  8.7M   5,000,000+  Free     0       Everyone
3  215644   25M  50,000,000+  Free     0           Teen
4     967  2.8M     100,000+  Free     0       Everyone

                     Genres      Last Updated         Current Ver  \
0              Art & Design   January 7, 2018               1.0.0
1  Art & Design;Pretend Play  January 15, 2018               2.0.0
2              Art & Design    August 1, 2018               1.2.4
3              Art & Design      June 8, 2018  Varies with device
4    Art & Design;Creativity     June 20, 2018                 1.1
```

Figure 1 : *output after loading dataset*

Data Cleaning

The raw dataset contained various inconsistencies, missing values, and formatting issues that needed to be resolved before analysis. A systematic cleaning process was applied to ensure data quality and integrity.

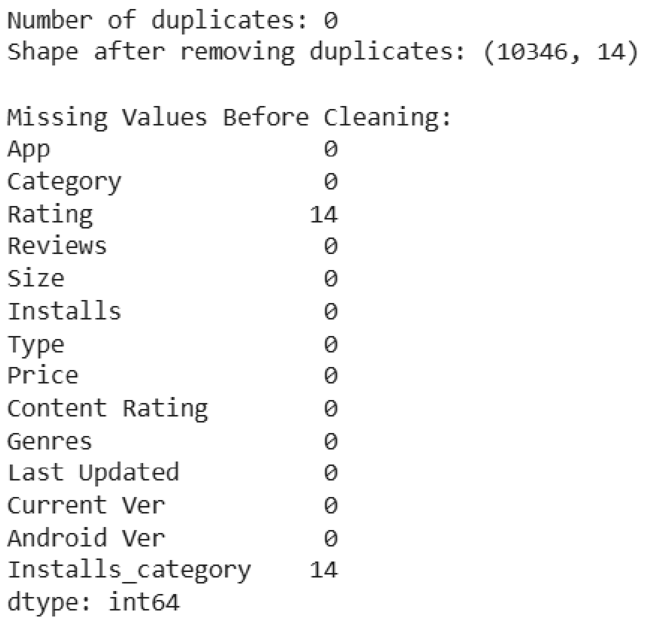


```
Installs column data type: int64
Sample Installs values:
0       10000
1      500000
2     5000000
3    50000000
4      100000
5       50000
6       50000
7     1000000
8     1000000
9       10000
Name: Installs, dtype: int64
Unique Installs values:
[      10000      500000     5000000    50000000      100000       50000
    1000000    10000000        5000   100000000  1000000000        1000
  500000000          50         100         500          10           1
          5           0]
```

Figure 2 : *output after data cleaning*

1. Duplicate and Invalid Rows

- **Duplicate entries** were identified and removed (483 duplicates dropped).
- A specific invalid row (index 10472) was also excluded from the dataset.

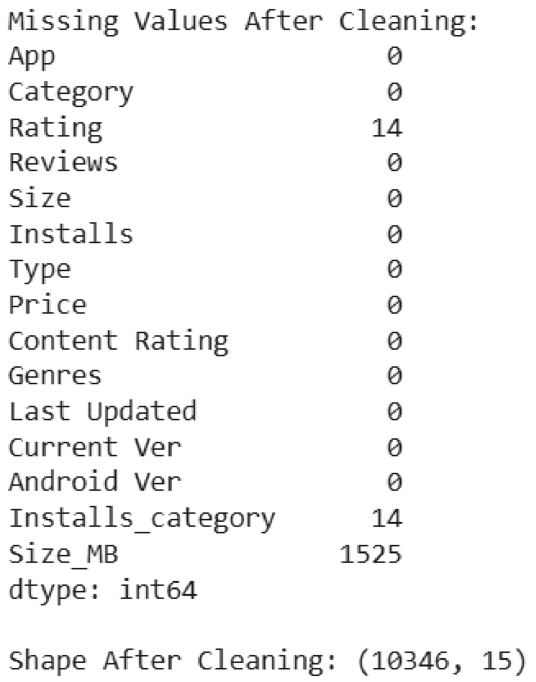


```
Number of duplicates: 0
Shape after removing duplicates: (10346, 14)

Missing Values Before Cleaning:
App                  0
Category             0
Rating              14
Reviews              0
Size                 0
Installs             0
Type                 0
Price                0
Content Rating       0
Genres               0
Last Updated         0
Current Ver          0
Android Ver          0
Installs_category   14
dtype: int64
```

Figure 3 : *output duplicates and invalid rows*

2. Handling Missing Values

- Columns with significant importance such as `Category`, `Type`, `Genres`, `Current Ver`, and `Android Ver` had missing entries dropped.
- **Missing `Rating` values** were imputed based on the app's `Installs_category`, using predefined averages from the dataset.



```
Missing Values After Cleaning:
App                     0
Category                0
Rating                 14
Reviews                 0
Size                    0
Installs                0
Type                    0
Price                   0
Content Rating          0
Genres                  0
Last Updated            0
Current Ver             0
Android Ver             0
Installs_category      14
Size_MB              1525
dtype: int64

Shape After Cleaning: (10346, 15)
```

Figure 4 : *output missing values after cleaning*

3. Data Type Conversion

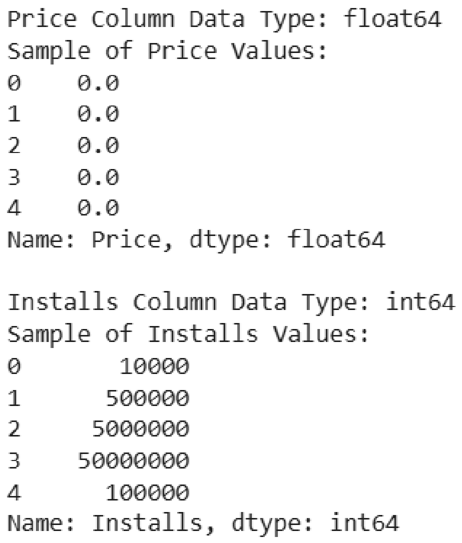
- **`Reviews`**: Converted to integer.
- **`Price`**: Converted from string (e.g., "$4.99") to float, with invalid entries replaced by `0.0`.
- **`Size`**: Converted to megabytes (`MB`) and stored in a new column `Size_MB`. Values in kilobytes were normalized and non-numeric entries (like "Varies with device") were handled as `NaN`.

4. Feature Engineering

- Created a new column `Installs_category` using `pd.cut()` to categorize apps based on install count ranges (e.g., Low, Moderate, Very High, Top Notch).
- This enabled more targeted imputation and categorical analysis.

5. Final Checks

- Verified data types and missing values post-cleaning.
- Final dataset contained **10,346 clean entries** (from an original 10,841).



```
Price Column Data Type: float64
Sample of Price Values:
0    0.0
1    0.0
2    0.0
3    0.0
4    0.0
Name: Price, dtype: float64

Installs Column Data Type: int64
Sample of Installs Values:
0       10000
1      500000
2     5000000
3    50000000
4      100000
Name: Installs, dtype: int64
```

Figure 5 : *data cleaning*

Exploratory Data Analysis (EDA)

After cleaning and preparing the dataset, various visual and statistical techniques were applied to uncover patterns, relationships, and insights within the Google Play Store app data.

1. Distribution of App Ratings

- Most app ratings were between **4.0 and 4.5**, suggesting generally positive user feedback.
- A smooth distribution curve with a slight skew was observed, visualized using a histogram and KDE plot.

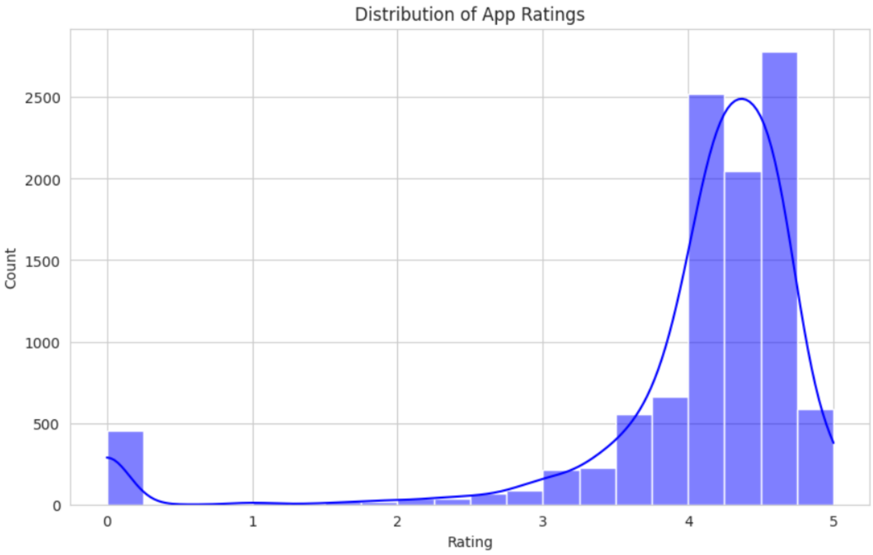


Figure 6 : *distribution of app ratings*

2. App Count by Category

- The **FAMILY** and **GAME** categories had the highest number of apps.
- Other dominant categories included **TOOLS**, **PRODUCTIVITY**, and **COMMUNICATION**.

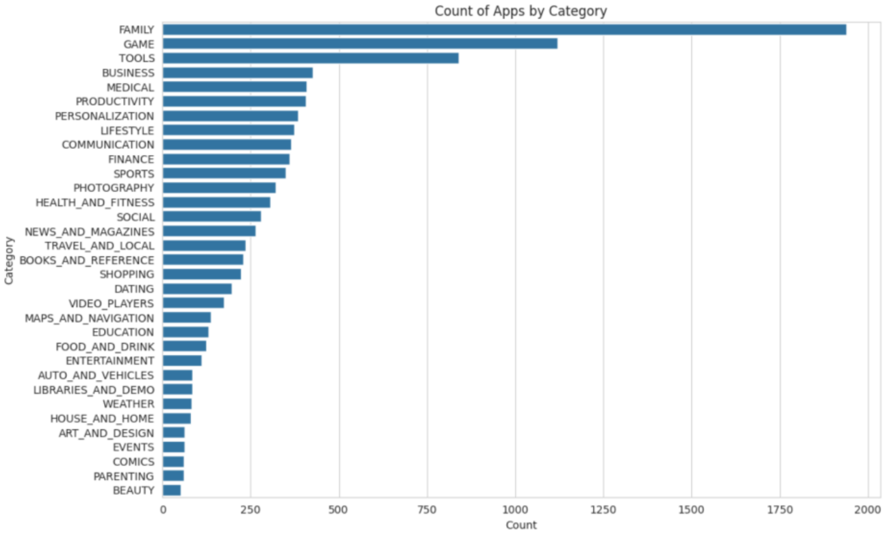


Figure 7 : *distribution of app ratings*

3. Relationship Between Installs and Ratings

- A scatter plot (with `log10(Installs)` for scale) showed a broad spread.
- No direct correlation, but clusters appeared for highly installed, moderately rated apps (often from large developers).

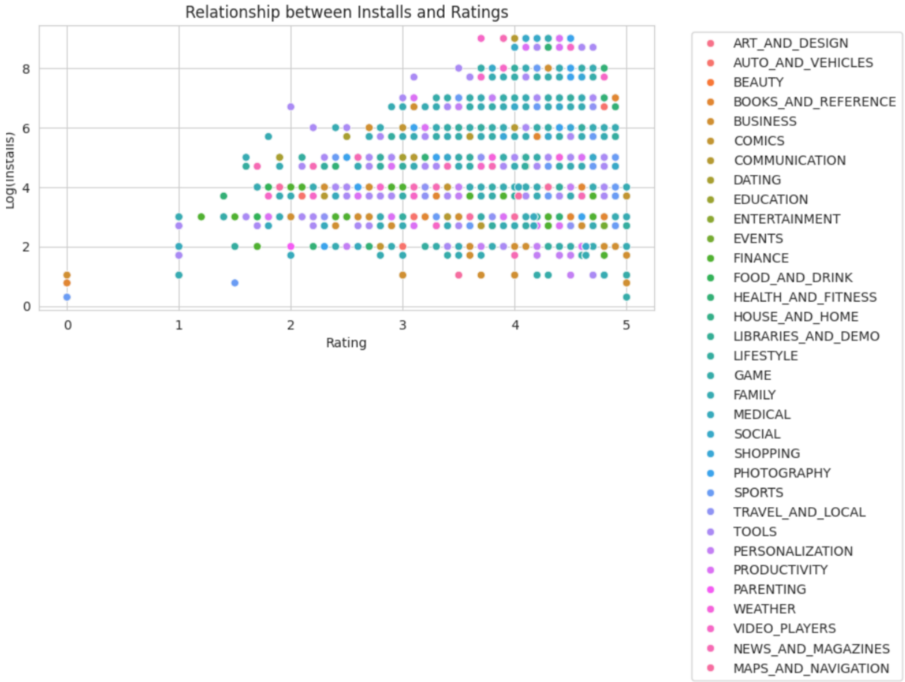


Figure 8 : *relationship between install and ratings*

4. Correlation Analysis

- A heatmap of numerical features revealed:
  - **Reviews and Installs** had the highest positive correlation (r ≈ 0.62), indicating popular apps attract more feedback.
  - **Price** had weak correlation with other variables.
  - **Rating** showed mild positive correlation with **Reviews** and **Installs**.

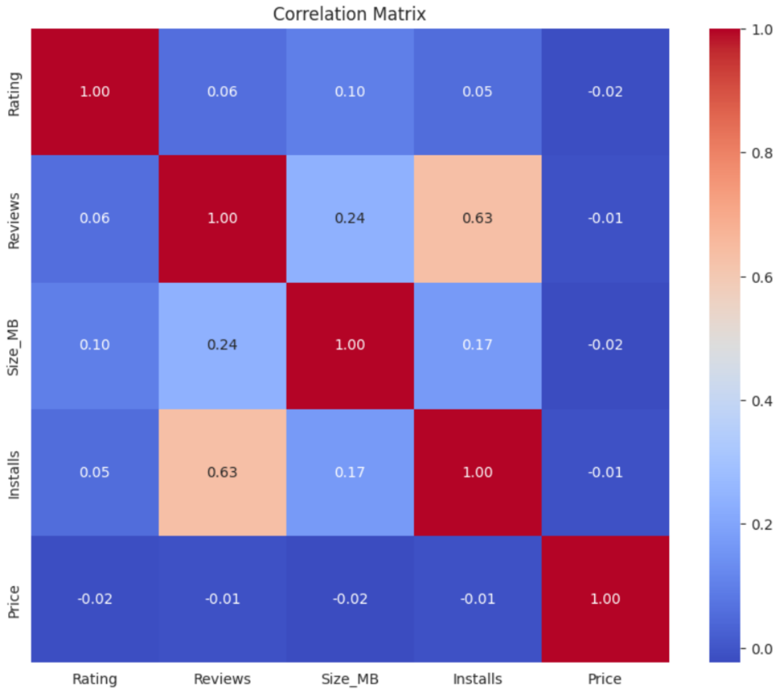


Figure 9 : *correlation matrix*

5. Ratings by Install Category

- Apps with higher installs (e.g., **Top Notch**, **Very High**) generally had more stable and higher ratings.
- Apps with **Very Low** or **Low** installs showed a wider rating distribution.

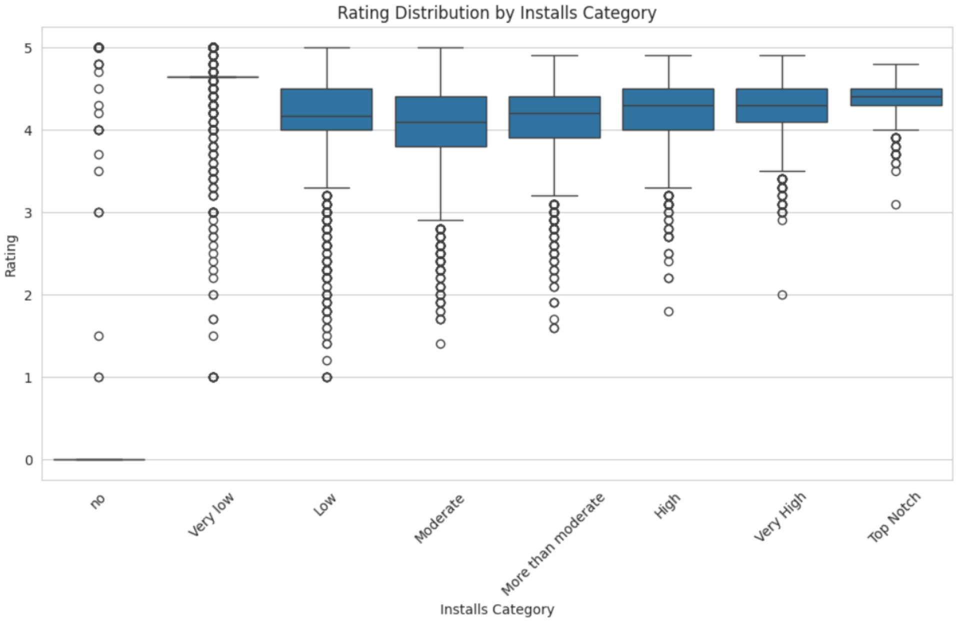


Figure 10: *correlation matrix*

CONCLUSION

This project involved a comprehensive exploratory data analysis of the Google Play Store dataset to uncover meaningful trends and patterns in app characteristics, user feedback, and market behavior.

Key findings include:

- The dataset initially contained 10,841 apps, reduced to 10,346 after cleaning.
- FAMILY and GAME categories dominate the Play Store in terms of number of apps.
- Apps in the EVENTS, EDUCATION, and BOOKS_AND_REFERENCE categories consistently achieved the highest average ratings.
- The GAME and COMMUNICATION categories recorded the highest number of installs, highlighting their popularity among users.
- A strong positive correlation (r ≈ 0.62) was observed between the number of reviews and installs, indicating that app popularity drives user engagement.
- Rating distributions showed that most apps cluster between 4.0 and 4.5, with very few rated below 3.0.
- Missing ratings were more prevalent in apps with fewer installs, supporting the assumption that less popular apps receive less feedback.

These insights can assist developers, marketers, and stakeholders in understanding the dynamics of app success on the Google Play Store, helping to optimize product development and user acquisition strategies.

Next Steps

Based on the insights derived from this exploratory analysis, several avenues for further investigation and enhancement are recommended:

1. **Advanced Feature Engineering**
   - Derive new features such as:
     - App age (based on Last Updated)
     - Sentiment analysis on app descriptions or reviews
     - Popularity index combining installs and reviews
2. **Predictive Modeling**
   - Build regression models to predict app ratings or review counts
   - Use classification models to predict app success categories (e.g., high install or high rating)
3. **Cluster Analysis**
   - Segment apps using clustering algorithms (e.g., KMeans) based on install count, reviews, and ratings
   - Identify groups of similar apps or market niches
4. **Interactive Dashboards**
   - Develop dashboards using Plotly, Dash, or Tableau for dynamic exploration of app performance and trends
5. **Category-Specific Studies**
   - Deep-dive into specific categories like **GAME**, **FAMILY**, or **TOOLS** to understand user expectations and monetization strategies
6. **Textual Data Analysis**
   - Apply NLP techniques to analyze app names, descriptions, or user reviews to extract keywords, sentiment, or thematic trends
7. **Comparative App Store Analysis**
   - Compare Google Play Store trends with those from the Apple App Store or other platforms to uncover cross-platform differences

References

1. Google Play Store Apps Dataset
   Kaggle. (n.d.). *Google Play Store Apps*. Retrieved from https://www.kaggle.com/datasets/lava18/google-play-store-apps
2. Cleaned Google Play Store Dataset
   Kaggle. (n.d.). *Cleaned Google Play Store Dataset*. Retrieved from https://www.kaggle.com/datasets/harshvir04/cleaned-google-play-store-dataset
3. Google Play Store Reviews Dataset
   Kaggle. (n.d.). *Google Play Store Reviews*. Retrieved from https://www.kaggle.com/datasets/prakharrathi25/google-play-store-reviews
4. Analyzing Google Play Store Datasets with Python
   Medium. (2023, September 17). *Analyzing Google Play Store Datasets with Python*. Retrieved from https://medium.com/@kabila2022/analyzing-google-play-store-datasets-with-python-fb41e07a2518

5. Data Cleaning Case Study: Google Play Store Dataset
   Tung M. Phung. (n.d.). *Data Cleaning Case Study: Google Play Store Dataset*. Retrieved from https://tungmphung.com/data-cleaning-case-study-google-play-store-dataset/