EDA on Play Store App Reviews

Abstract :

The Google Play Store serves as the official marketplace for Android devices, providing users with access to a wide range of applications, games, and digital media, including books, music, movies, and TV shows. It plays a crucial role in connecting app developers with millions of users worldwide. User ratings and reviews are critical indicators of app performance, influencing user engagement and download rates. Through this exploratory data analysis, we aim to uncover the key factors driving user satisfaction, positive feedback, and higher ratings. By understanding these parameters, app developers and businesses can optimize their strategies to enhance app visibility, engagement, and overall success on the platform.

Problem Statement:

The success of an app on the Google Play Store is heavily influenced by user engagement, ratings, and reviews. However, identifying the key factors that drive positive user feedback and high ratings remains a challenge for app developers and businesses. This project aims to analyze the Google Play Store dataset to uncover insights into the factors affecting app success, such as app category, pricing, size, installs, and user reviews. By understanding these relationships, we can provide actionable recommendations for improving app performance, user satisfaction, and overall engagement on the platform.

Data Summary:

The Google Play Store dataset comprises two main parts: app-related data and user review data. These datasets provide a comprehensive view of app features, user feedback, and performance metrics.

The contents of data are:

Play Store Data:

- **App:** Name of the app.
- **Category:** The category to which the app belongs (e.g., Education, Lifestyle, Games).
- **Rating:** Average user rating for the app, ranging from 1 to 5.
- **Reviews:** Total number of user reviews for the app.
- **Size:** The app's file size (in MB or specified as 'Varies with device').
- **Installs:** The number of times the app has been installed (e.g., 10k, 1M+).
- **Type:** Indicates whether the app is Free or Paid.
- **Price:** Cost of the app in dollars (0 for Free apps).
- **Content Rating:** Specifies the suitable audience age group (e.g., Everyone, Teen).
- **Genres:** The app's genre(s), providing additional classification details.
- **Last Updated:** Date of the app's most recent update.
- **Current Ver:** The current version of the app.
- **Android Ver:** Minimum Android version required to run the app.

User Review Data:

- **App:** Name of the app (to link with the Play Store dataset).
- **Translated Review:** User reviews translated into English.
- **Sentiment:** The sentiment of the review (Positive, Negative, Neutral).
- **Sentiment Polarity:** A numerical value indicating the sentiment's polarity,
- ranging from -1 (negative) to +1 (positive).
- **Sentiment Subjectivity:** A numerical value between 0 and 1 that indicates whether
- a review is factual (closer to 0) or opinionated (closer to 1).

Data Cleaning Process:

The dataset underwent several cleaning steps to ensure its quality and consistency for effective analysis. Below are the key actions performed:

1. **Removal of Duplicates:**

- Duplicate entries were identified and removed to ensure that each app was listed only once, reducing redundancy.

2. **Handling Missing Values:**

Missing values in critical fields like Rating and Current Ver were addressed:

- **Rating:** Missing values were replaced with the average rating of the app's respective category.
- **Current Ver and Android Ver:** Missing entries were either replaced with the most common versions or labeled as "Unknown" for consistency.

3. **Standardization of Formats:**

- Created a new Revenue column by combining data from Installs and Price to calculate app revenue.
- Standardized the Size column by converting all values to MB. Entries like "Varies with device" were either removed or replaced with a default value.
- Non-numerical characters in columns like Price and Installs were removed, and the data was converted to appropriate numerical formats for analysis.

4. **Dropped Irrelevant Columns:**

- **Last Updated:** Removed due to excessive missing values and lack of relevance

to the analysis objectives.

- **Translated Review:** Removed from the user reviews dataset, as the sentiment column already captured the necessary insights from the reviews.

Tools and Techniques Used:

Power BI: Utilized for advanced visualizations, dashboard creation, and integration of calculated metrics to provide actionable insights.

Techniques Applied:

1. **Data Cleaning:**

- Identified and removed duplicate records to avoid redundancy.
- Addressed missing values in key columns such as `Rating` and `Current Ver` using imputation techniques (e.g., meadian imputation for ratings).
- Standardized formats for numeric, text, and date fields to maintain consistency across the dataset.

2. **Data Transformation:**

- Added new columns such as `Revenue` by combining `Installs` and `Price` to evaluate app performance in monetary terms.
- Standardized app size (`Size` column) to MB and converted non-standard entries (e.g., "Varies with device").
- Categorized apps by install ranges for better segmentation and analysis.

3. **Data Analysis:**

- Merged sentiment data with app data to analyze how user reviews and sentiment impact ratings and installs.
- Computed metrics like average sentiment polarity and subjectivity for each app to quantify user opinions.
- Performed correlation analysis to identify relationships between variables like `Rating`, `Installs`, and `Price`.

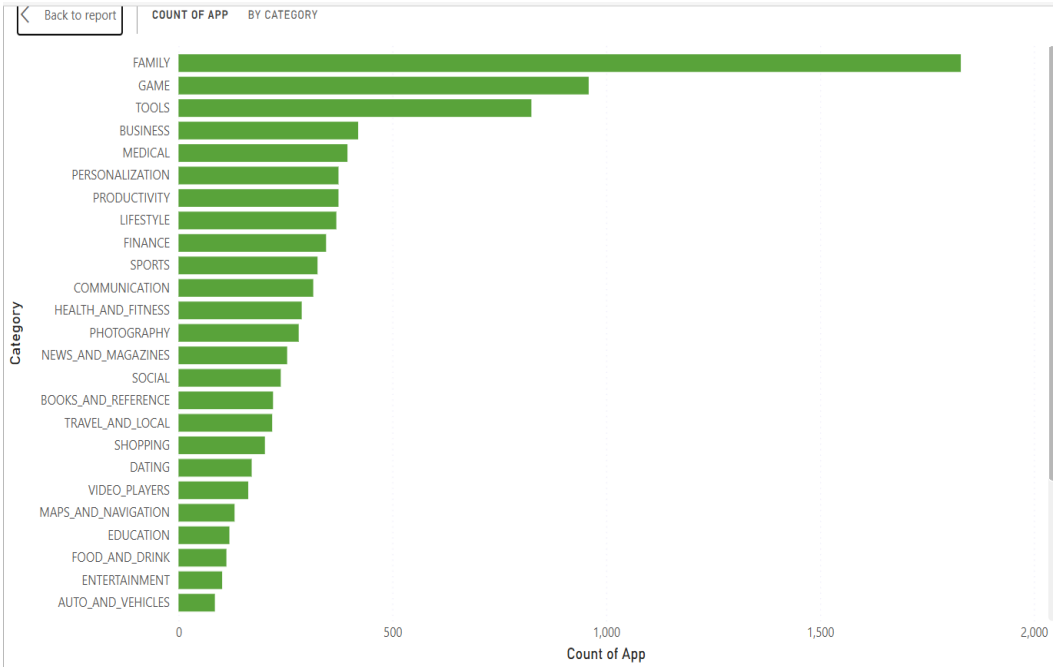
4. **Data Visualization:**

- Structured the dataset for advanced visualizations, including heatmaps, scatter plots, bar charts, and pie charts.
- Created visualizations to display category-wise performance, install trends, and sentiment distributions.

- Designed dashboards in Power BI to highlight actionable insights, such as the top-performing categories and price impact on installs.

Visualization:

Count of App By category:



This bar chart shows the count of apps categorized by their respective categories in the Google Play Store dataset.

Key Observations:

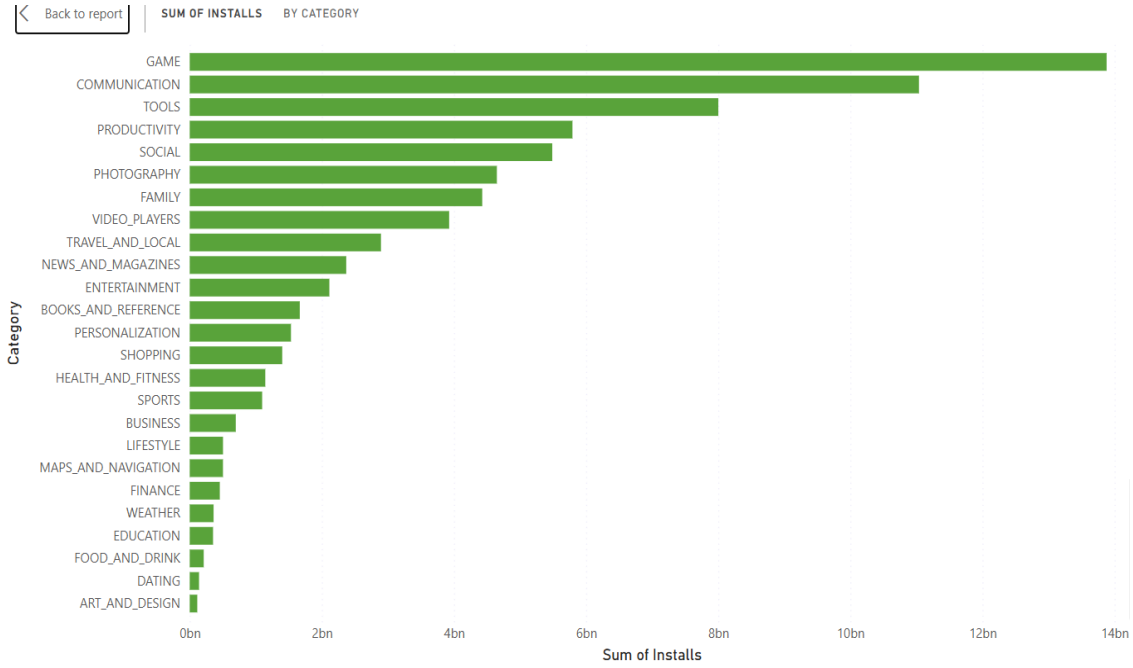
- The **Family** category has the highest number of apps, followed by **Games** and **Tools**.
- Categories like **Business**, **Medical**, and **Personalization** also have a significant number of apps.
- Niche categories such as **Auto and Vehicles**, **Entertainment**, and **Food and Drink** have fewer apps compared to other categories.

Insights:

- The dominance of the **Family** and **Games** categories suggests a high focus on entertainment and household-oriented applications in the Play Store.

- Businesses developing apps can explore less populated categories like **Auto and Vehicles** or **Food and Drink** to target less competitive markets.

Sum of Installs By Category:



This bar chart illustrates the total number of app installs across different categories in the Google Play Store dataset.

Key Observations:

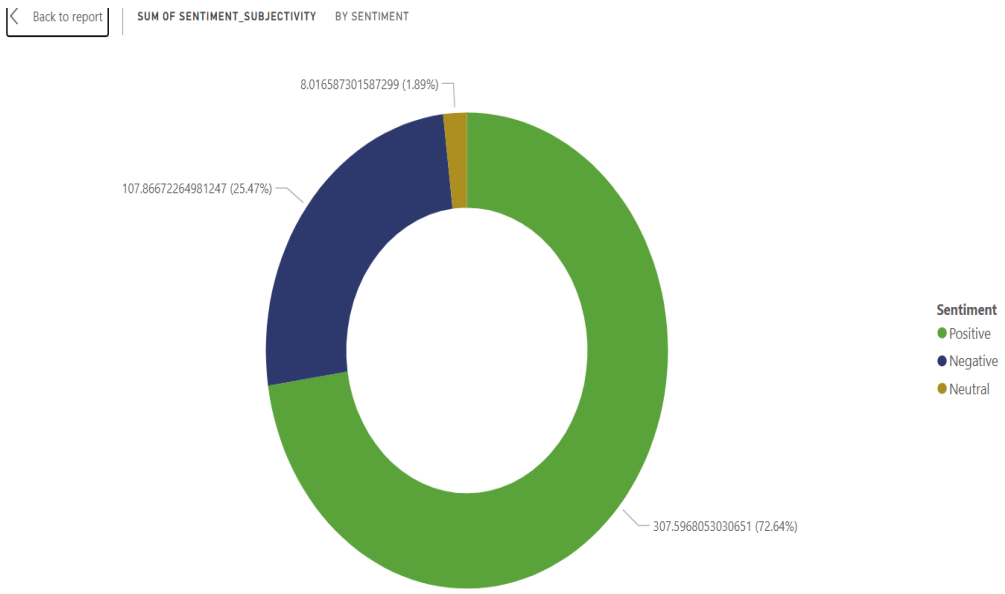
- The **Game** category has the highest total number of installs, indicating its immense popularity among users.
- **Communication** and **Tools** categories rank second and third in terms of total installs, showcasing their wide usage.
- Categories such as **Productivity**, **Social**, and **Photography** also have significant install counts, reflecting their importance to users.
- Lower install counts are observed in categories like **Art and Design**, **Dating**, and **Food and Drink**, suggesting niche or limited audiences.

Insights:

- High install counts in the **Game** and **Communication** categories highlight their dominant role in user engagement.

- Developers may focus on categories with moderate installs, such as **Travel and Local** or **News and Magazines**, for growth opportunities.

Sum of Sentiment_Subjectivity By sentiment:



This donut chart represents the distribution of sentiment subjectivity (positive,

Negative and neutral) for app reviews in the dataset.

Key Observations:

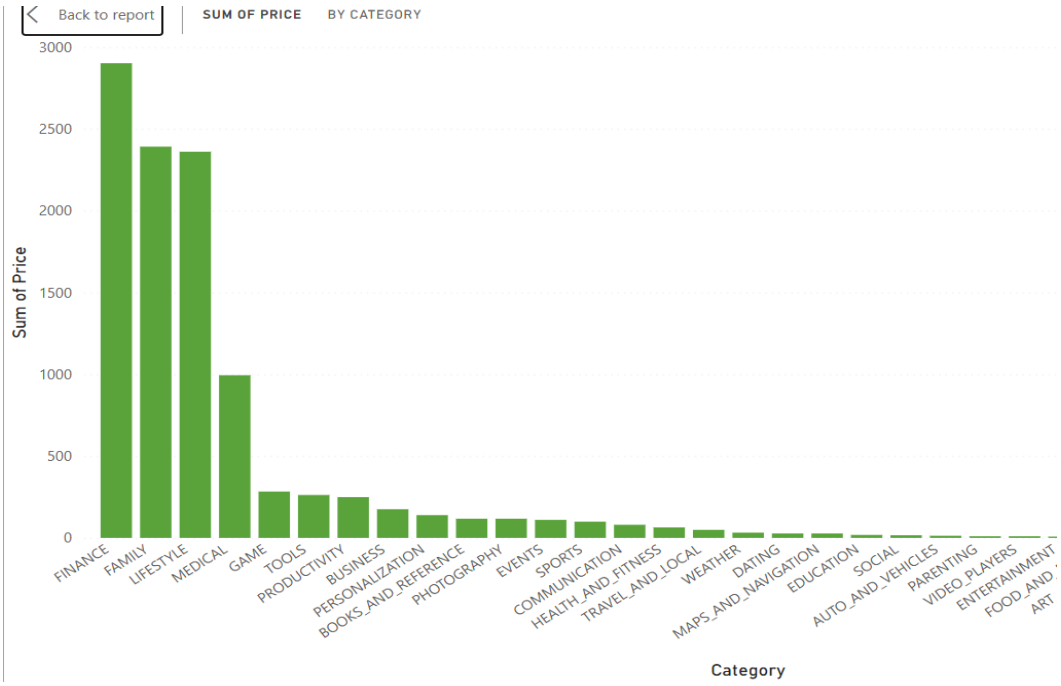
1. **Positive Sentiment** (Green) dominates the chart, accounting for **72.64%** of the total sentiment. This indicates that most user reviews are favorable.
2. **Negative Sentiment** (Blue) contributes **25.47%**, showing a smaller but still significant share of dissatisfaction among users.
3. **Neutral Sentiment** (Yellow) has the smallest proportion, representing only **1.89%** of the total sentiments.

Insights:

- The majority of app reviews are positive, suggesting that users generally have a good experience with the apps.
- The presence of a notable percentage of negative reviews highlights areas where app developers could focus on improvements.

- Neutral sentiments are minimal, implying that most user feedback is either explicitly positive or negative.

Sum of Price By Category:



This bar chart represents the **sum of app prices** grouped by category, highlighting the total revenue potential for paid apps in each category.

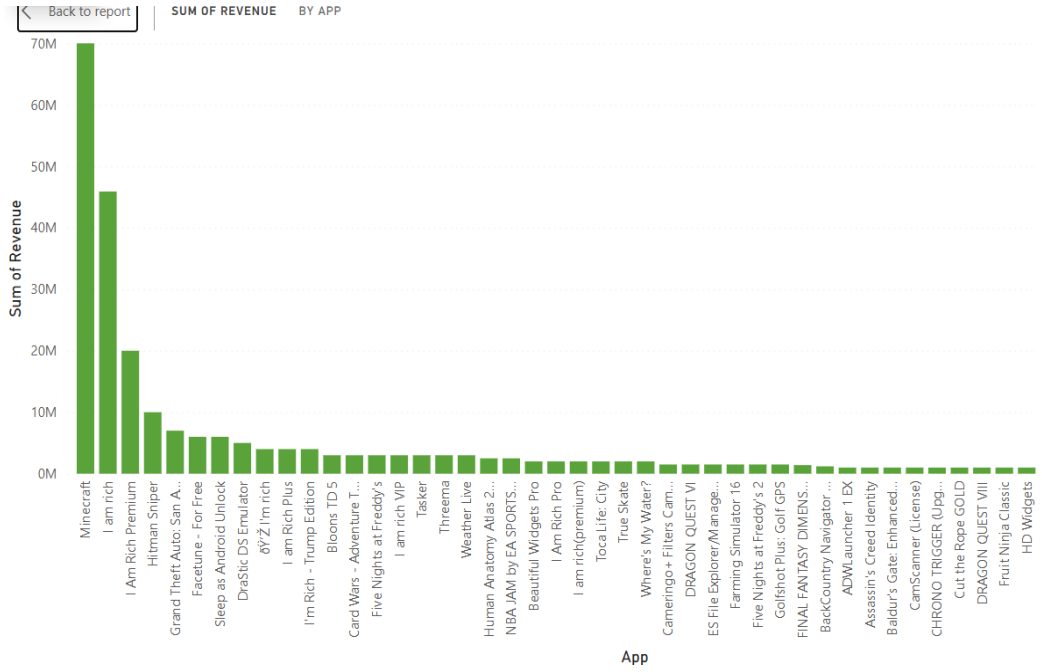
Key Observations:

- **Finance**, **Family**, and **Lifestyle** categories have the highest total app prices, suggesting that these categories are dominated by more expensive or numerous paid apps.
- **Medical** follows as the next category with significant app pricing sums, indicating a substantial number of paid apps in this field.
- Most other categories, including **Games**, **Tools**, **Productivity**, and **Business**, show much lower total prices compared to the top categories.
- Categories like **Weather**, **Dating**, and **Maps and Navigation** have minimal contributions to total app prices, indicating a smaller market for paid apps in these areas.

Insights:

- The **Finance** category's dominance in pricing may reflect the niche, value-driven nature of apps in this space.
- **Family** and **Lifestyle** categories likely cater to specialized audiences willing to pay for tailored apps.
- Developers could explore potential in categories with lower total prices by introducing premium offerings.

Sum of Revenue By App:



Bar Chart: Sum of Revenue by App

Description: The bar chart titled **"SUM OF REVENUE BY APP"** visually represents the revenue generated by various apps.

Key Points:

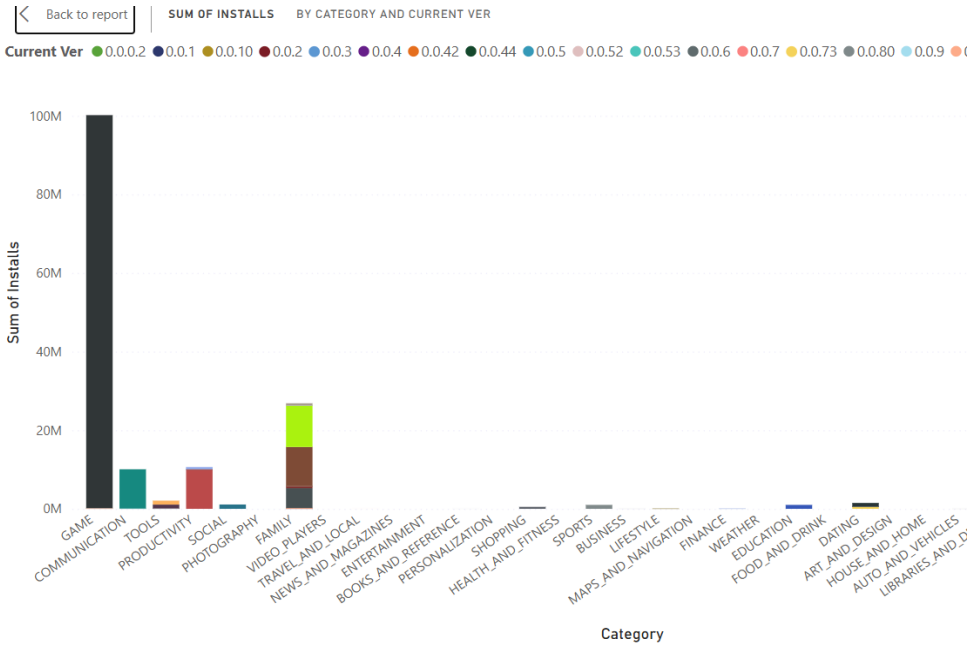
- **Minecraft**: Has the highest revenue, almost reaching 70M.
- **I Am Rich Premium** and **Grand Theft Auto: San Andreas**: Follow with significant revenues, around 50M each.
- **Other Apps**: The revenue decreases progressively for other apps, with many generating much lower revenue.

Insights:

- **Top Performers**: Minecraft, I Am Rich Premium, and Grand Theft Auto: San Andreas are the top revenue-generating apps.
- **Revenue Distribution**: The chart shows a clear decline in revenue among the other apps, highlighting the dominance of the top three apps.

  This chart helps in identifying the most financially successful apps and understanding the revenue distribution among different apps.

Sum of Installs By Category And Current ver:



Bar Chart: Sum of Installs by Category and Current Version

Description: This bar chart shows the sum of installs for various application categories, with different colors representing different current versions of the apps.

Key Points:

- **Game**: Has the highest number of installs, nearly reaching 100 million.
- **Communication, Social, and Family**: Also have significant install numbers, though much lower than the Game category.

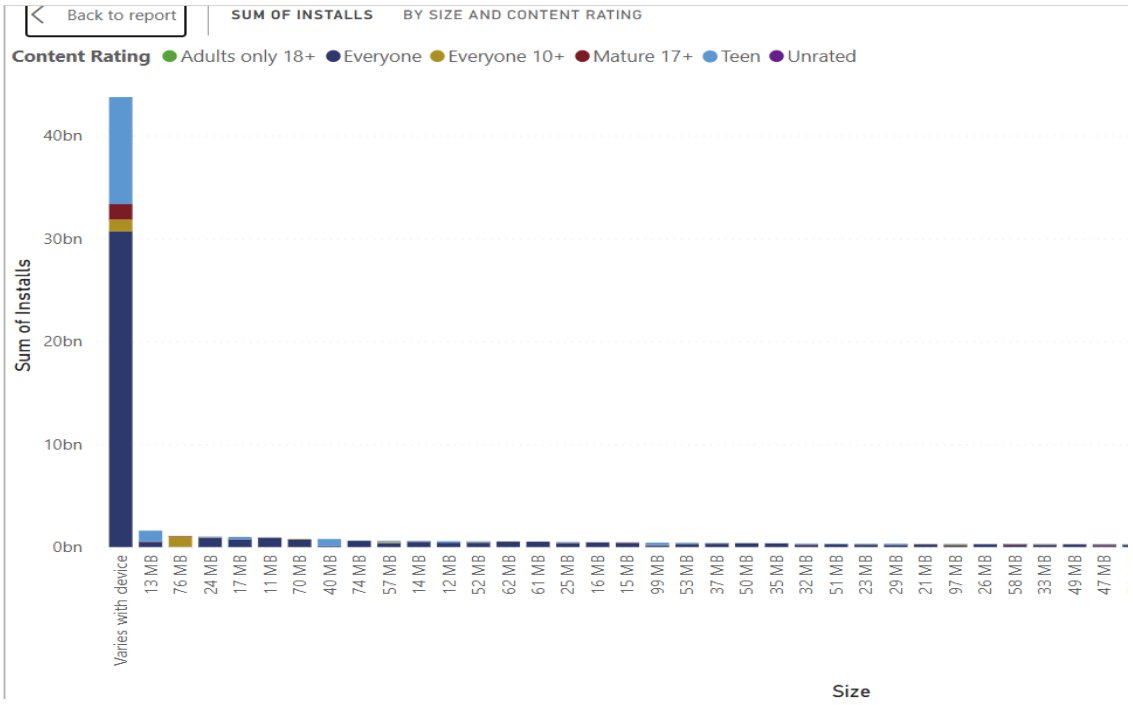
- **Other Categories**: Most other categories have lower install counts, indicating less popularity.

Insights:

- **Game Category**: Dominates in terms of the number of installs, indicating high popularity and usage.
- **Communication and Social Categories**: Also have substantial installs, suggesting their importance and frequent use.
- **Diverse Versions**: Different versions of apps within each category show varied usage and adoption rates.

This chart provides a visual representation of the popularity and usage of different categories of applications based on their install counts. It highlights which categories are most widely used and which ones have less traction.

Sum of Installs By Size And Content Rating:



Bar Chart: Sum of Installs by Size and Content Rating

Description: This bar chart shows the total sum of installs for applications of different sizes, categorized by content rating.

Key Points:

- **Varies with Device**: Applications with sizes that vary with the device have the highest sum of installs, reaching up to 40 billion. These installs are mainly for "Teen" rated apps, followed by "Everyone" and "Everyone 10+" rated apps.
- **Specific Sizes**: Other application sizes (e.g., 13 MB, 76 MB) have significantly lower sums of installs, with varying distributions of content ratings.

Content Ratings:

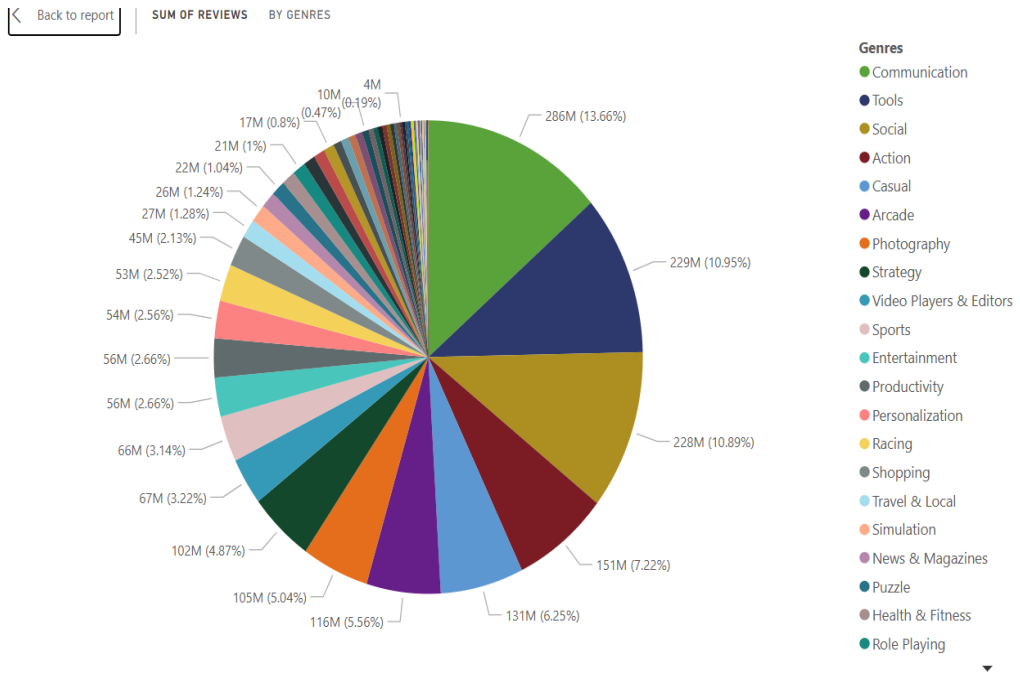
- **Teen (Blue)**: The majority of installs for "Varies with Device" sizes.
- **Everyone (Yellow)**: The next highest installs for "Varies with Device" sizes.
- **Everyone 10+ (Brown)**: Significant installs for "Varies with Device" sizes.
- **Mature 17+ (Red)**: Lower installs compared to Teen and Everyone ratings.
- **Unrated (Purple)**: The least installs among the content ratings.

Insights:

- **High Popularity**: Apps with sizes that vary with devices are highly popular, especially those rated for Teens and Everyone.
- **Content Rating Influence**: Content ratings like Teen and Everyone significantly influence the number of installs.

This chart provides a clear overview of how app size and content rating affect the total number of installs, highlighting the dominance of "Varies with Device" sized apps and the influence of content ratings on installs.

Sum of Reviews By Genre:



Pie Chart: Sum of Reviews by Genre

Description: This pie chart illustrates the total sum of reviews for various app genres.

Key Points:

- **Communication**: Has the highest number of reviews, with 286 million (13.66% of total reviews).
- **Tools and Social**: Follow closely with 229 million (10.95%) and 228 million (10.89%) reviews, respectively.
- **Action and Casual**: Also significant, with 151 million (7.22%) and 131 million (6.25%) reviews, respectively.
- **Other Genres**: Include Arcade, Photography, Strategy, Video Players & Editors, Sports, Entertainment, Productivity, Personalization, Racing, Shopping, Travel & Local, Simulation, News & Magazines, Puzzle, Health & Fitness, Role Playing, and Other, with varying percentages and millions of reviews.

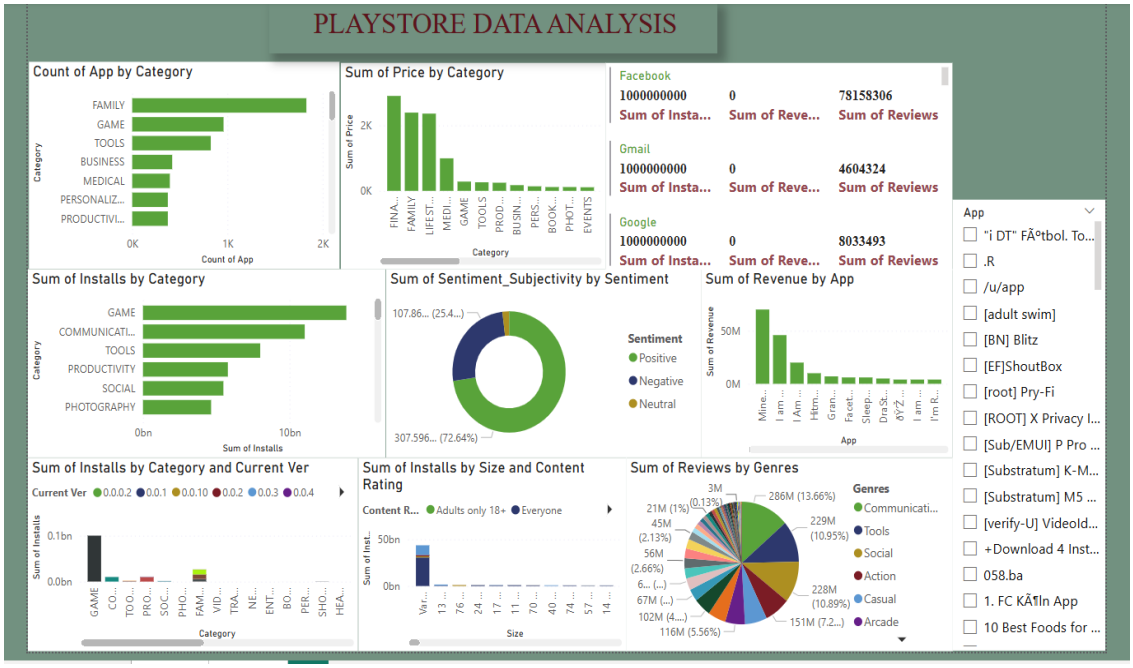
Insights:

- **Top Genres**: Communication, Tools, and Social dominate the review counts, highlighting their popularity and user engagement.

- **Genre Distribution**: The chart shows the distribution of reviews across various genres, indicating which genres have the most user activity.

This pie chart provides a clear overview of the review distribution among different app genres, helping identify the most popular and engaging categories.

Final Dashboard:



Summary of Play Store Data Analysis Dashboard:

1. **Most Apps Belong to the "Family" Category** – Family apps have the highest count, followed by Games and Tools.
2. **Games Have the Most Installs** – The "Game" category dominates in total installs, followed by Communication and Tools.
3. **Finance, Family & Lifestyle Apps Have the Highest Prices** – These categories generate the most revenue from paid apps.
4. **Positive Reviews are the Majority** – Over 72% of user sentiments are positive, while negative reviews are around 25%.
5. **Major Apps Like Facebook, Gmail & Google Have Billions of Installs** – But they have no direct revenue contribution in this dataset.
6. **Revenue is Mostly Concentrated in a Few Apps** – A small number of apps generate most of the revenue.

7. **Reviews Are Highest in Communication & Tools Categories** – These categories receive the most user feedback.
8. **Most Installs are from "Everyone"-Rated Apps** – Apps rated for all audiences are installed the most.

Insights:

- **Games and Communication apps dominate in popularity.**
- **Finance apps generate high revenue but are fewer in number.**
- **Users mostly leave positive reviews, indicating overall satisfaction.**
- **Targeting the right category can maximize installs or revenue depending on the strategy.**

Conclusion of Play Store Data Analysis Dashboard:

1. **Games and Communication Apps are the Most Popular**
    - These categories have the highest installs and reviews, showing strong user engagement.
2. **Finance and Lifestyle Apps Generate the Most Revenue**
    - Even though they have fewer installs, these apps tend to be paid or have higher monetization strategies.
3. **User Sentiment is Mostly Positive**
    - The majority of reviews are positive, indicating overall user satisfaction with Play Store apps.
4. **Free Apps Drive High Install Numbers**
    - Most installs come from apps rated for "Everyone," suggesting that free and widely accessible apps perform the best.
5. **A Few Apps Dominate the Market**
    - Revenue and installs are highly concentrated among top-performing apps like Facebook and Gmail, while smaller apps struggle for visibility.

Final Thought:

- **For maximum success, developers should focus on engaging and monetizing popular categories like Games and Communication, while also exploring high-value niches like Revenue and Lifestyle.**