EDA on Play Store App Reviews

Abstract:

Google play store is the official app store for all devices operating on the Android OS. It allows the users to browse and download the apps that are developed with the android software development kit (SDK). Apart from offering android applications and games, it also serves as a digital media store offering music, books, movies, and television programs. User ratings and reviews can significantly increase the number of app downloads; hence it is important to analyse the parameters which lead to users giving positive feedback and higher rating. Though this exploratory data analysis, we can understand and discover the key factors responsible for app engagement and success.

Problem Statement:

Two datasets are provided, one with basic information and the other with user reviews for the respective app. We must examine and evaluate the data in both datasets in order to identify the important characteristics that influence app engagement and success.

Data Summary:

We are provided with two datasets:

- **Play_store_data:** It contains the basic details of the app like number of user reviews, ratings, etc.
- **User reviews:** It contains the user reviews and its sentiment score for the respective app.

We need to explore and analyse the data to discover key factors responsible for app engagement and success.

The contents of play_store_data are:

- **App:** It contains the name of the app with a short description (optional).
- **Category:** This section gives the category to which an app belongs. In this dataset, the apps are divided among 33 categories.
- **Size:** The disk space required to install the respective app.
- **Rating:** The average rating given by the users for the respective app. It can be in between 1 and 5.
- **Reviews:** The number of users that have dropped a review for the respective app.
- **Installs:** The approximate number of times the respective app was installed.

- **Type:** It states whether an app is free to use or paid.
- **Price:** It gives the price payable to install the app. For free type apps, the price is zero.
- **Content rating:** It states which age group is suitable to consume the content of the respective app.
- **Genres:** It gives the genre(s) to which the respective app belongs.
- **Last updated:** It gives the day in which the latest update for the respective app was released.
- **Current Ver:** It gives the current version of the respective app.
- **Android Ver:** It gives the android version of the respective app.

The contents of User Reviews are:

- **App:** It contains the name of the app with a short description (optional).
- **Translated Review:** It contains the English translation of the review dropped by the user of the app.
- **Sentiment:** It gives the attitude/emotion of the writer. It can be 'Positive', 'Negative', or 'Neutral'.
- **Sentiment Polarity:** It gives the polarity of the review. Its range is [-1,1], where 1 means 'Positive statement' and -1 means a 'Negative statement'.
- **Sentiment Subjectivity:** This value gives how close a reviewer's opinion is to the opinion of the general public. Its range is [0,1]. Higher the subjectivity, closer is the reviewer's opinion to the opinion of the general public, and lower subjectivity indicates the review is more of a factual information.

Data Cleaning Process

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

1. **Removal of Duplicates**

- Duplicate entries in the dataset were identified and removed to ensure each app is listed only once.

2. **Handling Missing Values**

- Missing values were identified in fields such as "Rating" and "Current Ver."
- **Strategies used:**
    a. For "Rating," missing values were replaced with the mean rating for that app's category.

b. Missing values in "Current Ver" and "Android Ver" were either replaced with the most common version or labelled as "Unknown."

3. **Standardization of Formats**

- The "Revenue" column was added by using the app installation and price columns
- The "Size" column was standardized to MB (e.g., converting values like "Varies with device" to a default value or removing the entry).
- "Last Updated" column was deleted in the data set because most of the values are missing, we can't get insights from this.
- Translated Review column was deleted in the user reviews data because in the data set we have sentiment. The sentiment was taken from the translated review.

Tools and Techniques Used

- **Power BI**: Advanced cleaning, transformation, and integration of sentiment analysis data. The Power BI environment allowed for handling larger datasets and performing complex transformations.

- **Techniques Applied**

1. **Data Cleaning:**
    - Identified and removed duplicate records.
    - Addressed missing values using appropriate imputation strategies.
    - Ensured consistency in formats for numeric, textual, and date-based fields.
2. **Data Transformation:**
    - Created new fields for further analysis, such as grouping apps by install ranges or aggregating category-wise ratings.
    - Standardized key metrics like app size and installs for comparability.
3. **Data Analysis:**
    - Integrated sentiment analysis scores to provide insights into user feedback.
    - Calculated sentiment polarity and subjectivity for each review to quantify user opinions.
4. **Data Visualization:**
    - Ensured the dataset was structured to support visualizations like bar charts, scatter plots, and pie charts.
    - Generated derived metrics for trend and correlation analyses, such as category-wise average ratings or install growth rates.

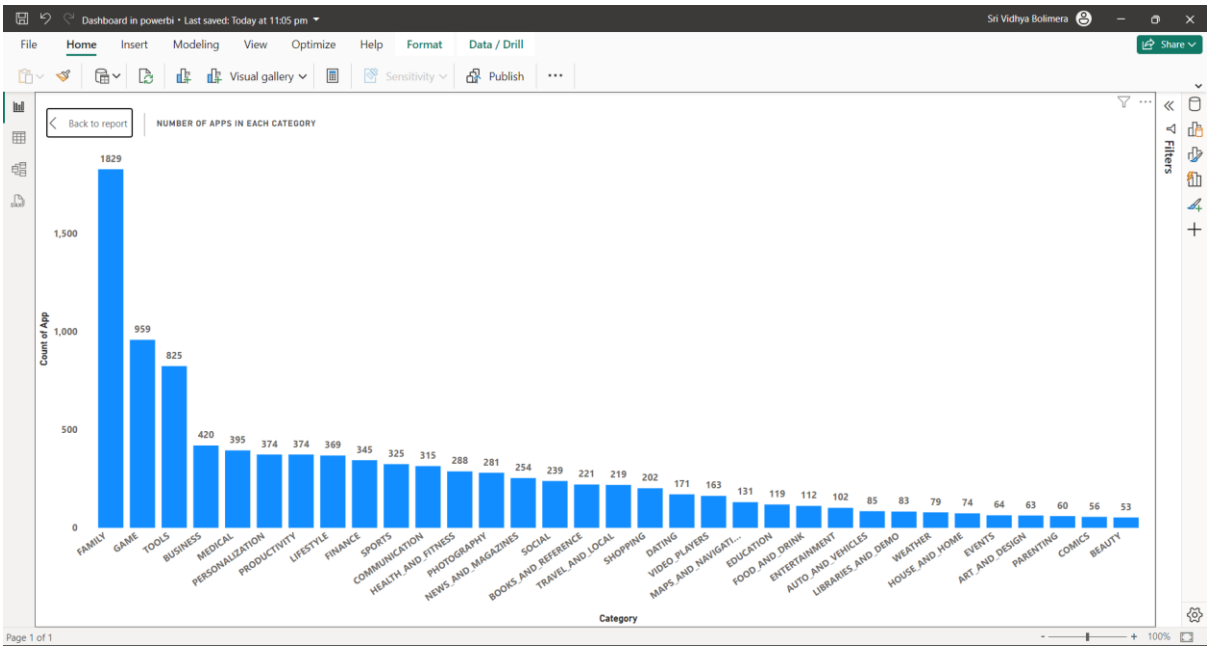
Cleaned Dataset for Analysis

The cleaned dataset was structured for efficient analysis and visualization. Below are the details of the final dataset:

- **Play store Columns:** The dataset included fields such as App, Category, Rating, Reviews, Size, Installs, Type, Content Rating, Genres, Last Updated, Current Ver and Revenue columns are present for the visualisation
- **Row Count:** A total of 9,649 rows remained after cleaning and preprocess
- **User Reviews:** Android Ver, Translated Sentiment, Sentiment Polarity, Sentiment Subjectivity, etc.
- **Row Count:** A total of 37,432 rows remained after cleaning.

Key Insights from Visualizations:

Number of Apps in Each Category:



1. **App Category Distribution:**

- The chart illustrates the distribution of apps across various categories.
- The length of each bar represents the number of apps within that category.

2. **Most Popular Categories:**

- "Games" is the most popular category with the highest number of apps.

- The top 5 categories with the highest number of apps are:
    - Games
    - Tools
    - Business
    - Medical
    - Personalization

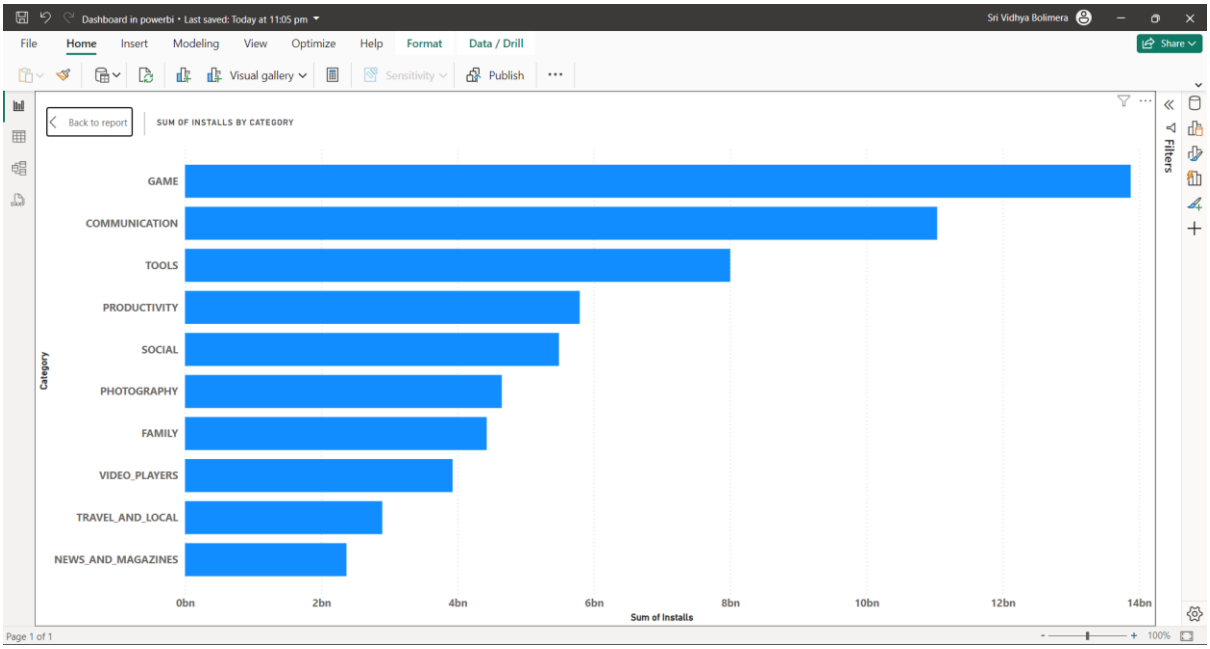
3. **Least Popular Categories:**

- The least populated categories include:
    - Beauty
    - Comics
    - Parenting

4. **Insights:**

- App developers tend to focus more on certain categories like "Games," "Tools," and "Business."
- The distribution of apps across categories is uneven, with some categories having significantly more apps than others.

Sum Of Installs by Category



1. **App Installation Distribution:**

- The chart illustrates the total number of installations for apps across various categories.
- The length of each bar represents the number of times apps in that category have been installed.

2. **Most Installed Categories:**

- "Games" is the category with the highest number of installations.
- The top 5 categories with the highest number of installations are:
    a. Games
    b. Communication
    c. Tools
    d. Productivity
    e. Social

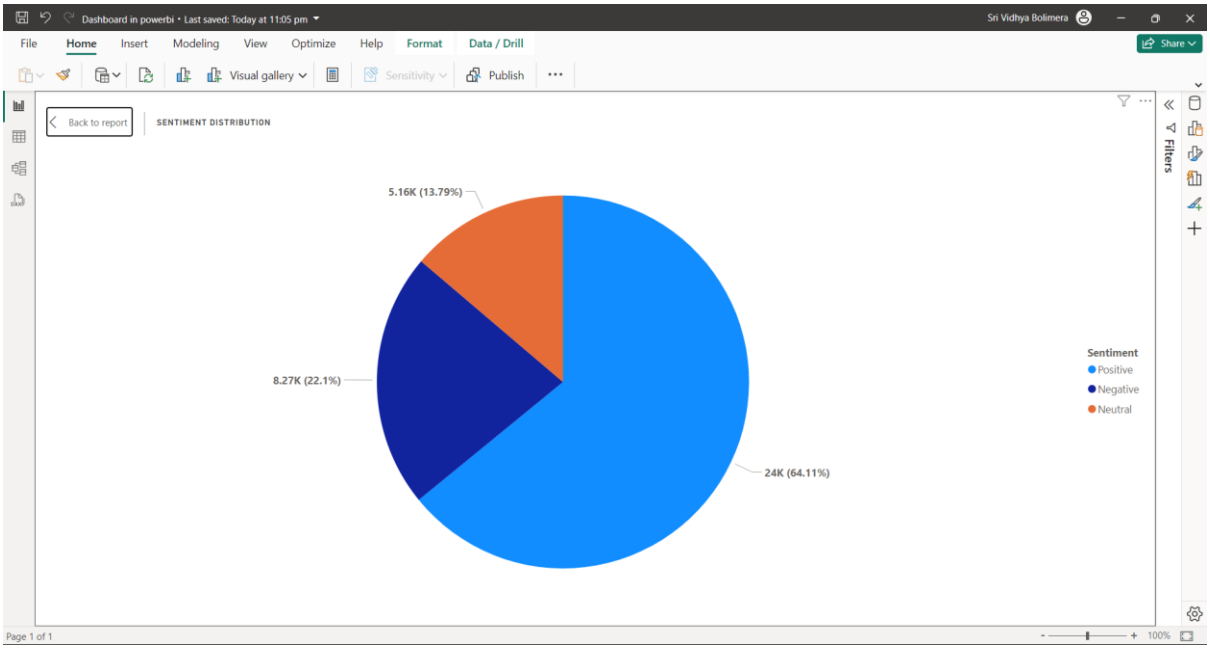
3. Least Installed Categories:

- The categories with the fewest installations include:
    a. News and Magazines
    b. Travel and Local
    c. Video Players

4.Insights:

- App categories like "Games," "Communication," and "Tools" are extremely popular and have been installed billions of times.
- The distribution of installations is uneven, with some categories having significantly more installations than others.

Sentiment Distribution:



1.**Sentiment Distribution:**

- The pie chart visualizes the distribution of sentiments, likely related to customer feedback or social media analysis.

2. **Sentiment Categories:**

- The chart shows three sentiment categories:
    a. Positive
    b. Negative
    c. Neutral

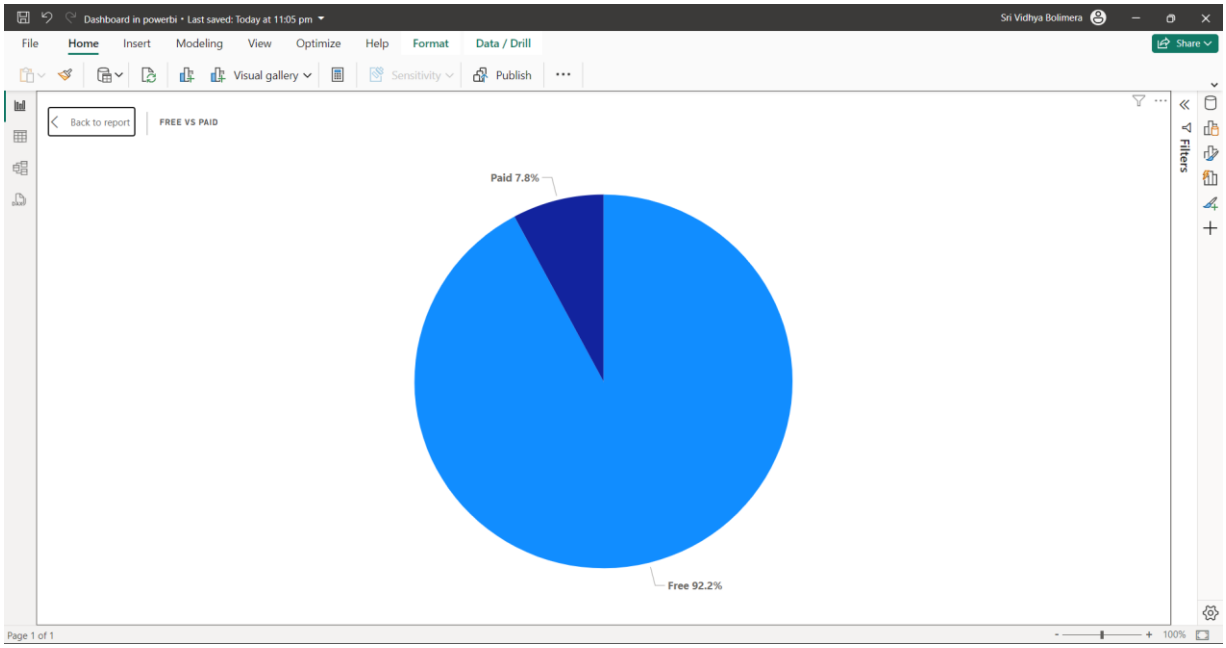
3.**Sentiment Proportions:**

- Neutral sentiment constitutes the largest proportion, accounting for approximately 64% of the total.
- Positive sentiment is the second largest, representing around 22% of the total.
- Negative sentiment makes up the smallest proportion, at about 14% of the total.

4.Insights:

- The data suggests that overall sentiment is predominantly neutral.
- Positive sentiment is higher than negative sentiment, indicating a generally positive sentiment.
- However, the significant presence of neutral sentiment suggests that there might be areas for improvement to increase positive sentiment.

Free Vs Paid



Free vs. Paid Apps:

- The pie chart illustrates the proportion of free apps compared to paid apps.

2. Distribution:

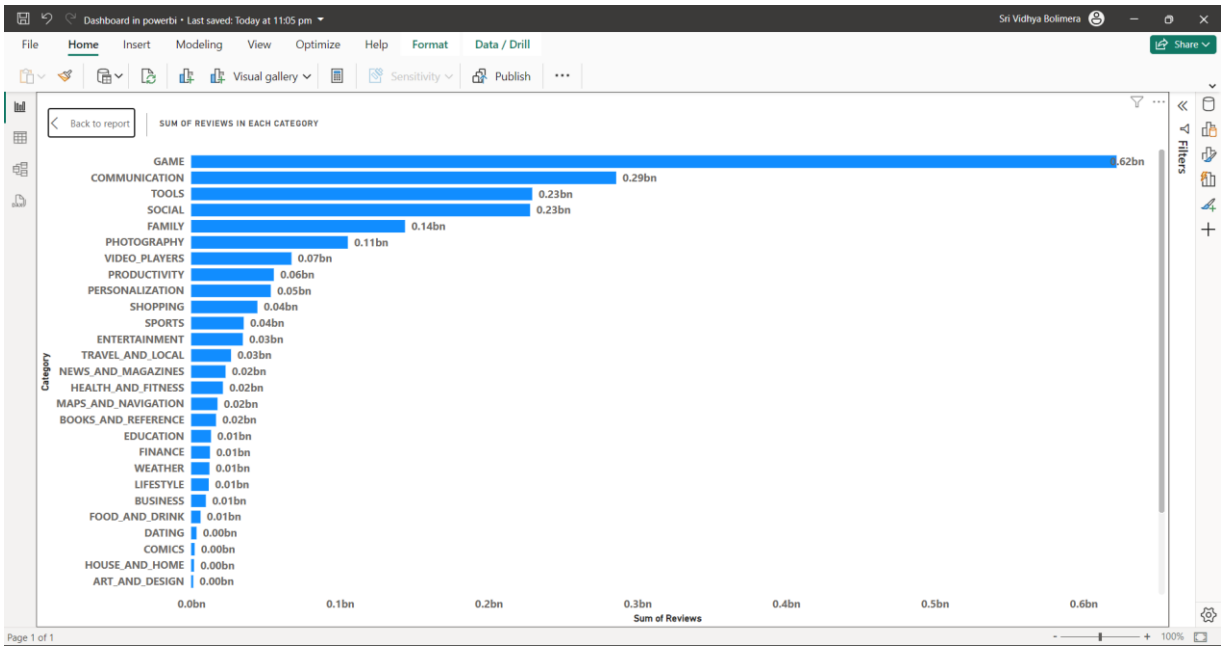
- The chart shows that the vast majority of apps are free, accounting for approximately 92.2% of the total.

- Paid apps make up a much smaller portion, representing about 7.8% of the total.

3.Insights:

- This distribution suggests that the app market is heavily dominated by free apps.
- Developers may rely on in-app purchases or advertisements within free apps to generate revenue.

Sum Of Revies in Each Category:



1.App Review Distribution:

- The chart illustrates the total number of reviews received for apps across various categories.
- The length of each bar represents the number of reviews for apps in that category.

2.Categories with the Most Reviews:

- "Games" has the highest number of reviews, exceeding 0.6 billion.
- The top 5 categories with the highest number of reviews are:
    a. Games
    b. Communication
    c. Tools
    d. Social

3.Categories with the Least Reviews:

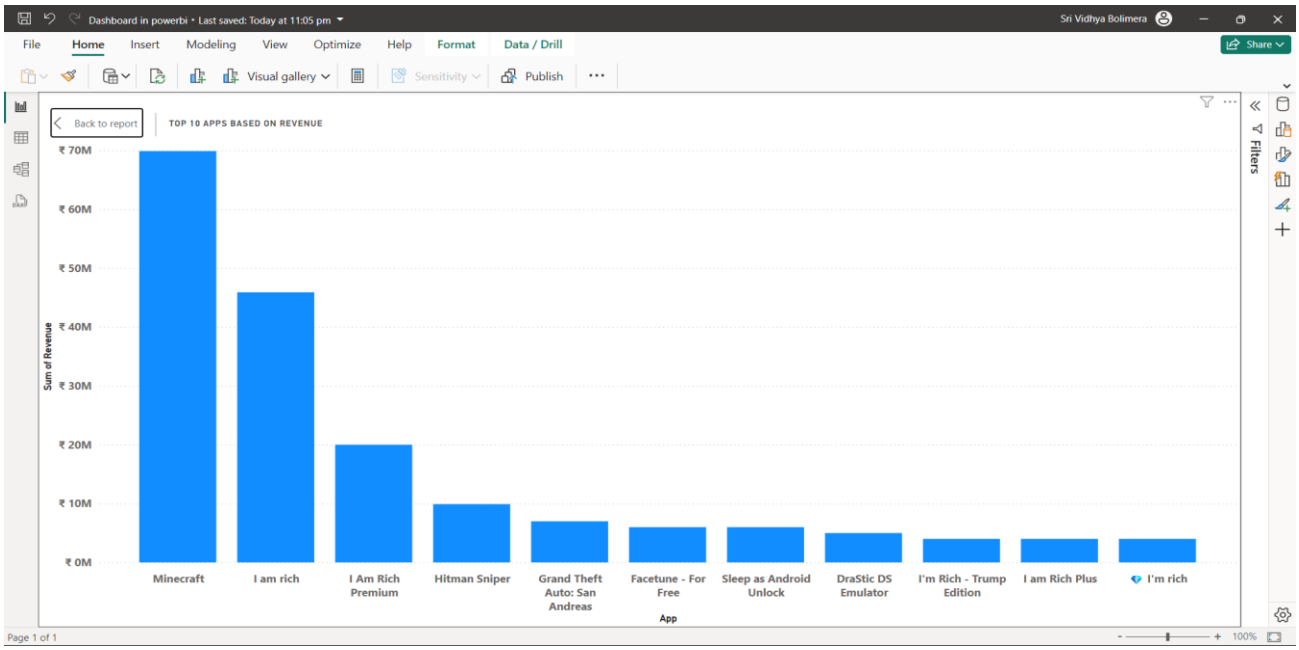
- The categories with the fewest reviews include:
    a. House and Home

b. Art and Design
c. Dating
d. d. Comics

4.Insights:

- App categories like "Games," "Communication," and "Tools" tend to receive the most reviews.
- The distribution of reviews is uneven, with some categories receiving significantly more reviews than others.

Top 10 Apps Based on Revenue:



1.Top 10 Apps by Revenue

- The chart visualizes the revenue generated by the top 10 highest-grossing apps.
- The length of each bar represents the revenue generated by that app.

2. Top-Grossing App:

- "Minecraft" is the top-grossing app, generating significantly more revenue than the others.

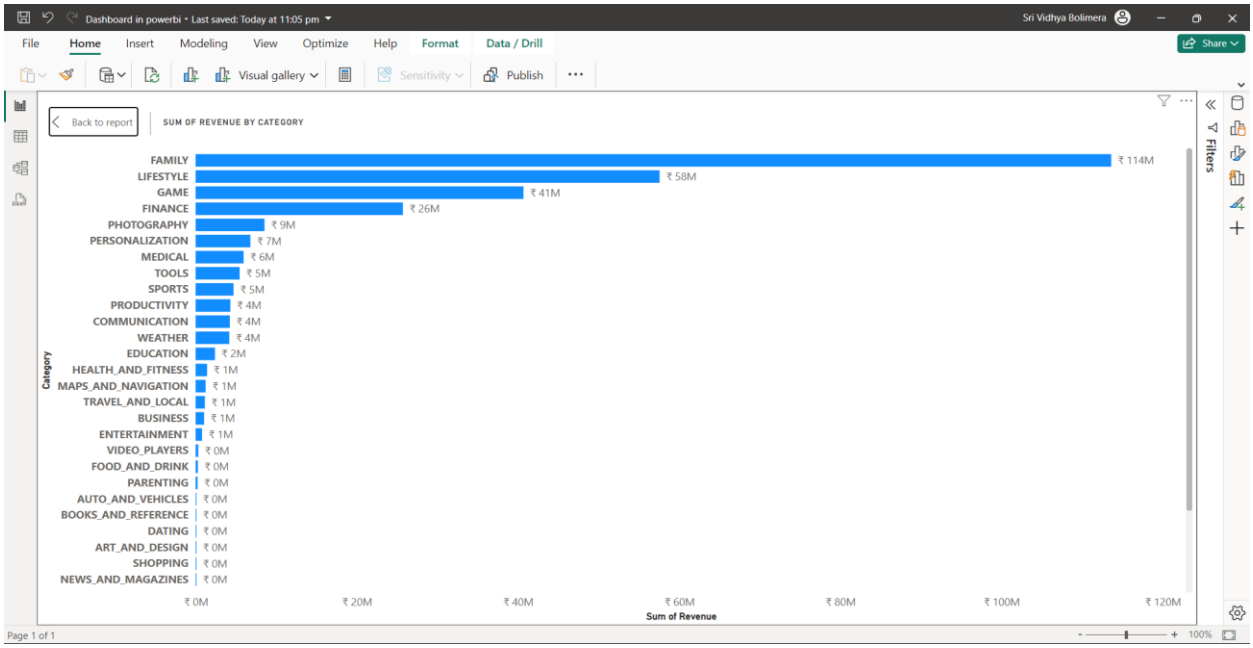
3.Other High-Revenue Apps:

- "I am Rich" and "I Am Rich Premium" rank second and third in terms of revenue.
- The remaining apps in the top 10 generated progressively lower revenue.

4.Insights:

- The chart highlights the significant revenue disparity between the top-grossing app and the rest.
- It suggests that a small number of apps generate a substantial portion of the total revenue.

Sum Of Revenue by Category:



1.App Revenue Distribution:

- The chart illustrates the revenue generated by apps across various categories.
- The length of each bar represents the total revenue earned by apps within that category.

2.Categories with Highest Revenue:

- "Family" is the category generating the highest revenue, exceeding 1.14 billion.
- The top 5 categories with the highest revenue are:
    a. Family
    b. Lifestyle
    c. Games
    d. Photography
    e. Personalization

3.Categories with Lowest Revenue:

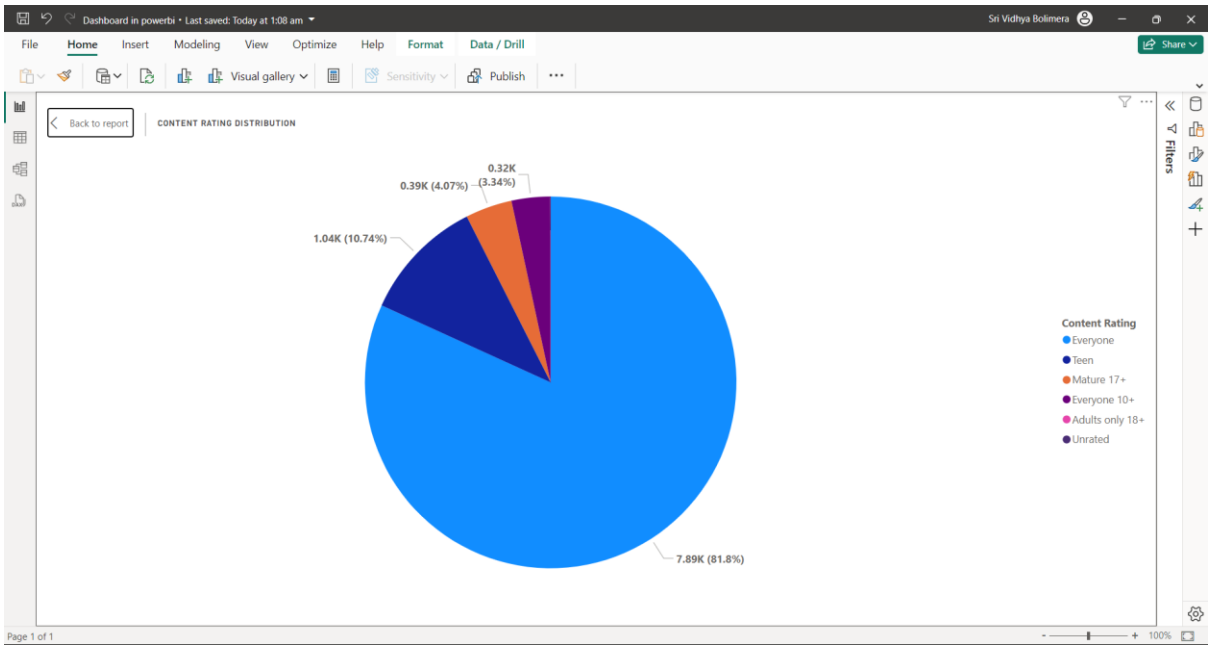
- The categories generating the least revenue include:
    a. News and Magazines
    b. Art and Design
    c. Food and Drink

d. Dating
e. Books and Reference

4. Insights:

- App categories like "Family," "Lifestyle," and "Games" contribute significantly to overall revenue generation.
- The distribution of revenue across categories is uneven, with some categories generating much more revenue than others.

Content Rating Distribution:



1. **Content Rating Distribution:**

- The pie chart visualizes the distribution of content ratings for a set of data, likely related to apps, movies, or similar content.

2. Content Rating Categories:

- The chart shows different content rating categories:
  - Everyone
  - Teen
  - Mature 17+
  - Adults only 18+
  - Unrated

3.Content Rating Proportions:

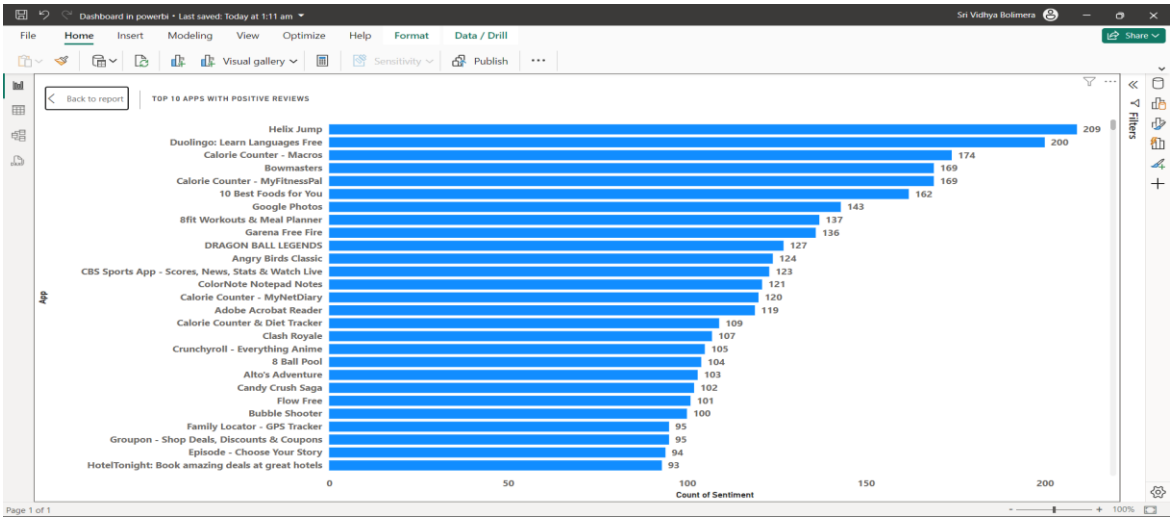
- Everyone content is the most prevalent, making up approximately 81.8% of the total.
- Teen content is the second largest, representing around 10.74% of the total.
- The remaining categories have much smaller proportions:

- Mature 17+: 4.07%
- Adults only 18+: 3.34%
- Unrated: 0.05%

4.Insights:

- The data suggests that the majority of the content is rated as suitable for everyone.
- There are relatively small proportions of content with higher age restrictions.

Top 10 Apps with Positive Reviews:



1. Top 10 Apps with Positive Reviews:

- The chart visualizes the number of positive reviews received by the top 10 apps with the highest number of positive reviews.
- The length of each bar represents the count of positive reviews for that app.

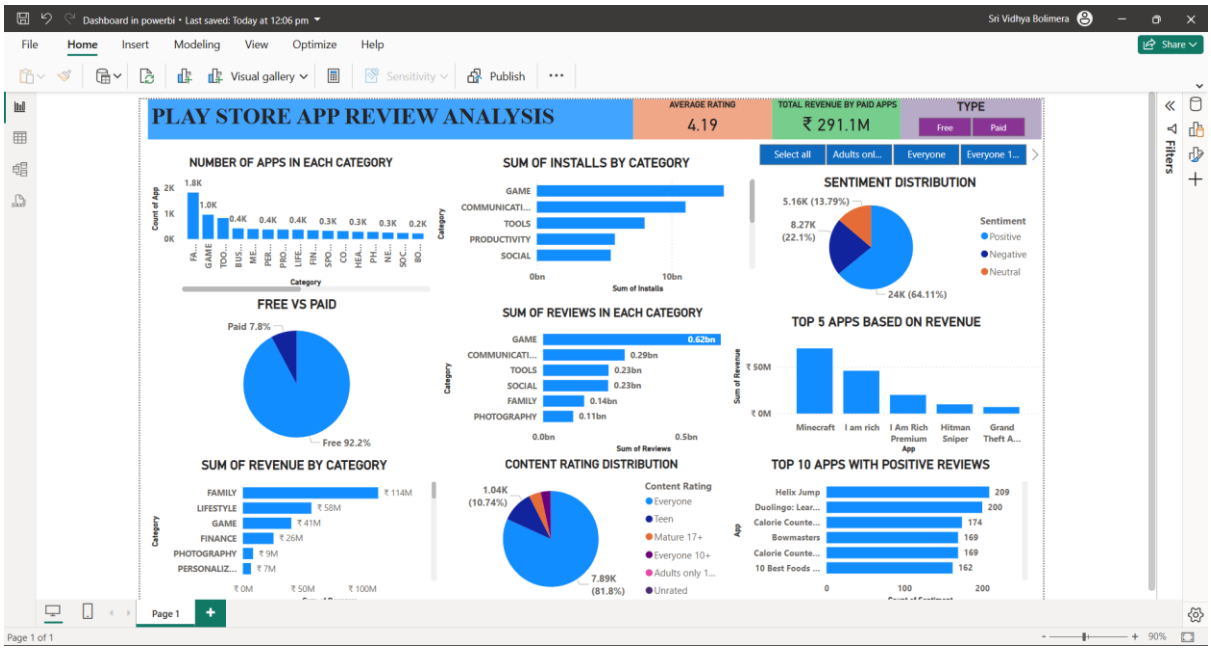
2.Apps with the Most Positive Reviews:

- "Hello Jump" has the highest number of positive reviews, exceeding 200.
- The top 10 apps with the highest number of positive reviews are:
  - Hello Jump
  - Duolingo: Learn Languages Free
  - Calorie Counter - Macros
  - Calorie Counter - MyFitnessPal
  - 10 Best Foods for You
  - Google Photos
  - Sfit Workouts & Meal Planner
  - Garena Free Fire
  - DRAGON BALL LEGENDS
  - Angry Birds Classic

3. Insights:

- The chart highlights the apps that have received the most positive feedback from users.
- This information can be valuable for app developers, marketers, and users alike.

Final Dashboard



1.App Market Overview:

- **Category Diversity:** The market showcases a wide range of app categories, with "Games" leading in terms of the number of apps available.
- **Free vs. Paid:** The majority of apps are free, which is a typical trend in app stores. This suggests that developers often rely on in-app purchases or advertisements for revenue.
- **Installation Popularity:** Categories like "Games," "Communication," and "Tools" are the most installed, highlighting their popularity among users.

2. App Performance & Revenue:

- **Revenue Leaders:** "Family" and "Lifestyle" apps generate the highest revenue, indicating their profitability.
- **Top-Grossing Apps:** A small number of apps generate a significant portion of the total revenue, emphasizing the importance of identifying and capitalizing on successful app concepts.
- **Reviews:** Categories like "Games" and "Communication" receive the most reviews, suggesting high user engagement and interaction.

3.User Sentiment and Content:

- **Neutral Sentiment:** A large portion of user sentiment is neutral, indicating a need for developers to focus on improving user experience and addressing neutral feedback.
- **Positive Reviews:** The "Top 10 Apps with Positive Reviews" chart highlights the apps that have resonated well with users, providing valuable insights for other developers.
- **Content Rating:** Most apps are rated for "Everyone," suggesting a focus on family-friendly content and a wide audience reach.

Overall, the dashboard provides a comprehensive view of the app market. It highlights key trends, identifies high-performing categories and apps, and offers valuable insights into user behavior and sentiment. This information can be used by developers, marketers, and businesses to make informed decisions regarding app development, marketing strategies, and user experience.

Conclusion:

- **App Market is Dynamic:** The market is diverse with numerous categories, and "Games" dominate in terms of both app numbers and installations.
- **Free Apps Dominate:** The majority of apps are free, with developers relying on in-app purchases or ads for revenue.
- **Revenue is Concentrated:** A small number of apps generate a significant portion of the total revenue, highlighting the importance of successful app concepts.
- **User Engagement Varies:** Categories like "Games" and "Communication" receive more reviews and installations, indicating higher user engagement.
- **Sentiment is Predominantly Neutral**: While positive sentiment exists, a significant portion of user feedback is neutral, suggesting areas for improvement in app development.
- **Content Rating is Primarily "Everyone":** This suggests a focus on family-friendly content and a wide audience reach for many apps.

Overall, the dashboard provides a snapshot of the app market, highlighting key trends, identifying successful categories and apps, and offering insights into user behavior. This information can be valuable for developers, marketers, and businesses in making informed decisions about app development and marketing strategies.