Code

Analytical Insights into iOS App Profitability

Krishnasurya Gopalakrishnan, Monica Muniraj, Shashank Shivakumar, Srinivas Saiteja Tenneti

Introduction

In the ever-evolving realm of mobile applications, a pivotal question echoes in the minds of developers and investors alike: where should one invest their resources, and what factors truly define an app's success? As the year 2021 ended, the digital landscape stood witness to a remarkable phenomenon: Apple's App Store, adorned with a staggering 1.6 million iOS applications. Despite the vast expanse of Google's Play Store, housing 3.5 million apps, the App Store emerged as a beacon of innovation and opportunity.

Welcome to our exploration within this vibrant ecosystem – the second-largest app marketplace globally – where innovation seamlessly intertwines with consumer preferences. Here, at the heart of Apple's App Store, our mission unfolds. We delve deep into the intricate web of variables that shape an app's destiny, exploring the realms of price, size, user ratings, and categories. These elements are not mere data points; they are the building blocks of informed decisions.

Our primary objective is clear: to identify the core variables that significantly influence an app's market triumph and profitability. For developers and investors, this knowledge is paramount, guiding their strategic choices in the ever-competitive app market. This isn't just an exploration; it's a journey where insights illuminate pathways to success, transforming raw data into boundless opportunities.

Join us in this odyssey, where data paints a vivid picture of potential and where every insight is a steppingstone towards informed, strategic decisions. Welcome to a realm where exploration isn't just a venture; it's the key to unlocking the infinite possibilities within the digital cosmos of mobile applications.

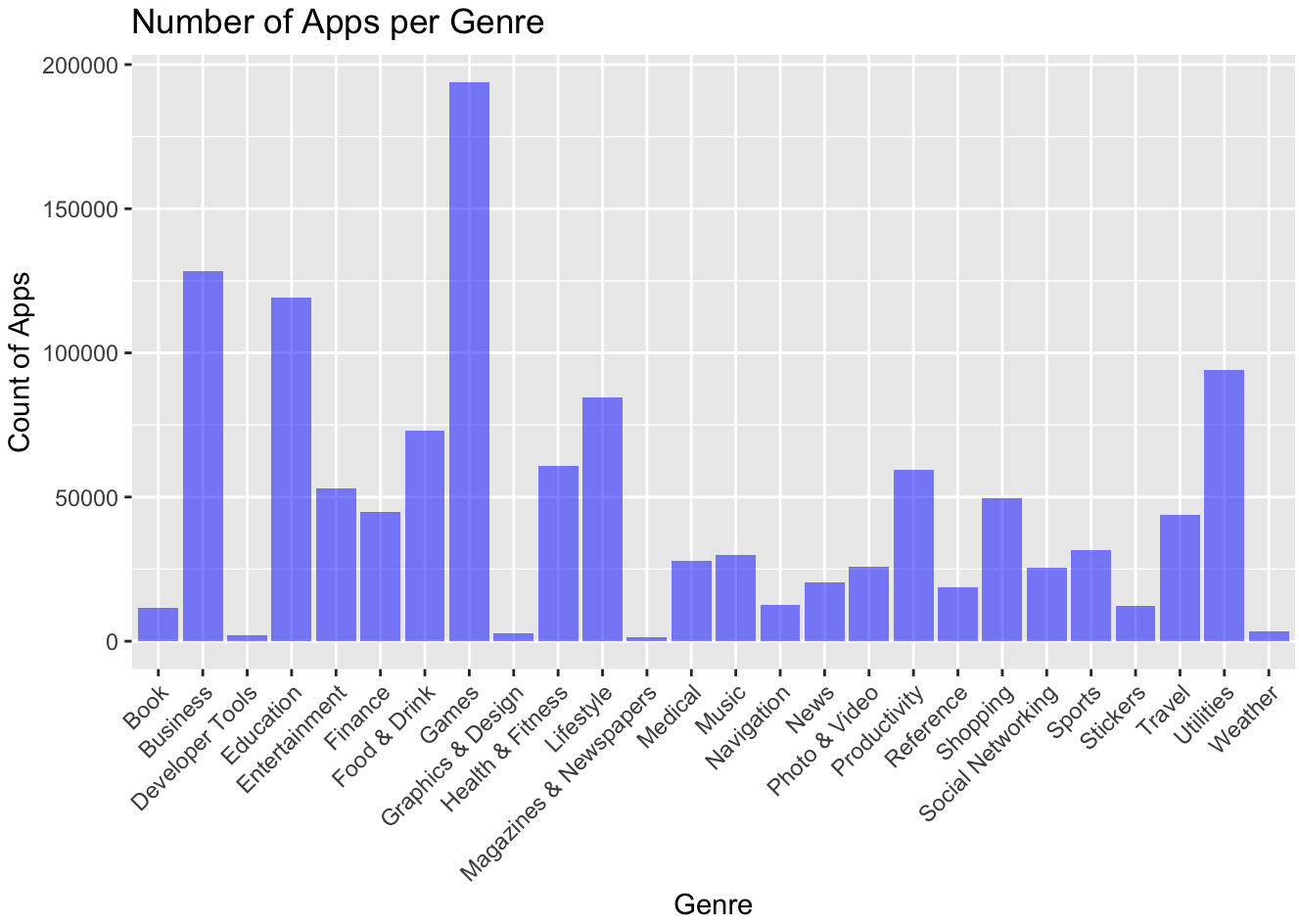
Descriptive statistics

Show

```
## 'data.frame':    1230372 obs. of  20 variables:
##  $ App_Id                : chr  "com.hkbu.arc.apaper" "com.dmitriev.abooks" "n
o.terp.abooks" "fr.antoinettefleur.Book1" ...
##  $ App_Name              : chr  "A+ Paper Guide" "A-Books" "A-books" "A-F Book
#1" ...
##  $ AppStore_Url          : chr  "https://apps.apple.com/us/app/a-paper-guide/id
1277517387?uo=4" "https://apps.apple.com/us/app/a-books/id1031572002?uo=4" "http
s://apps.apple.com/us/app/a-books/id1457024164?uo=4" "https://apps.apple.com/us/ap
p/a-f-book-1/id500454885?uo=4" ...
##  $ Primary_Genre         : chr  "Education" "Book" "Book" "Book" ...
##  $ Content_Rating        : chr  "4+" "4+" "4+" "4+" ...
##  $ Required_IOS_Version  : chr  "8.0" "10.0" "9.0" "8.0" ...
##  $ Released              : POSIXct, format: "2017-09-28 03:02:41" "2015-08-31 1
9:31:32" ...
##  $ Updated               : POSIXct, format: "2018-12-21 21:30:36" "2019-07-23 2
0:31:09" ...
##  $ Version               : chr  "1.1.2" "1.3" "1.3.1" "1.2" ...
##  $ Currency              : chr  "USD" "USD" "USD" "USD" ...
##  $ Free                  : chr  "True" "True" "True" "False" ...
##  $ DeveloperId           : int  1375410542 1031572001 1457024163 439568839 6567
31821 1146730227 1228484249 1535421441 1179894999 586867546 ...
##  $ Developer             : chr  "HKBU ARC" "Roman Dmitriev" "Terp AS" "i-editeu
r.com" ...
##  $ Developer_Url         : chr  "https://apps.apple.com/us/developer/hkbu-arc/i
d1375410542?uo=4" "https://apps.apple.com/us/developer/roman-dmitriev/id1031572001?
uo=4" "https://apps.apple.com/us/developer/terp-as/id1457024163?uo=4" "https://app
s.apple.com/us/developer/i-editeur-com/id439568839?uo=4" ...
##  $ Average_User_Rating   : num  0 5 0 0 0 0 0 0 0 0 ...
##  $ Reviews               : int  0 1 0 0 0 0 0 0 0 0 ...
##  $ Current_Version_Score : num  0 5 0 0 0 0 0 0 0 0 ...
##  $ Current_Version_Reviews: int  0 1 0 0 0 0 0 0 0 0 ...
##  $ Price                 : num  0 0 0 2.99 0 0 0 0 0 0 ...
##  $ Size_Bytes            : num  21993472 13135872 21943296 81851392 64692224
...
```

Number of Apps per Genre

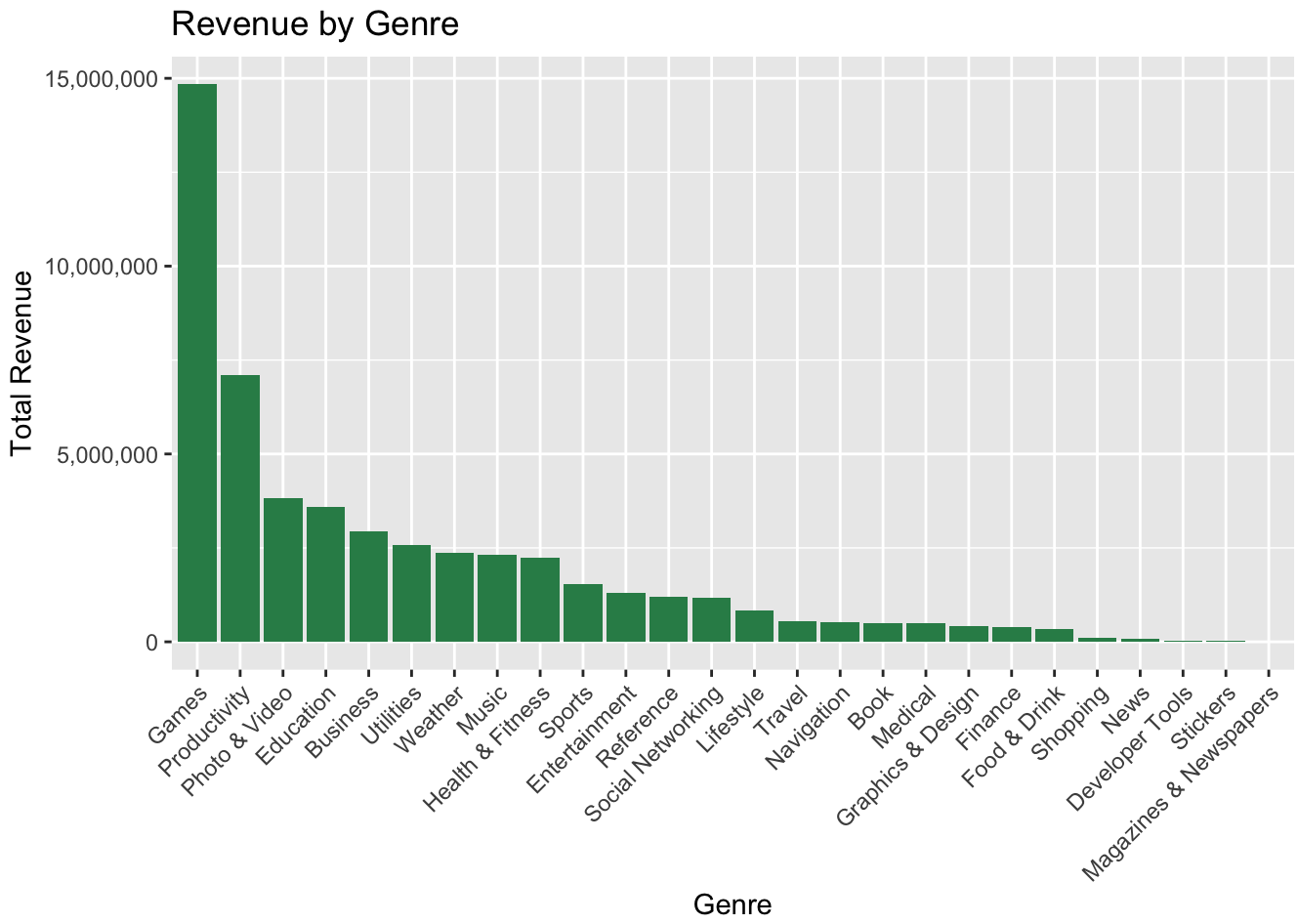
Show



The graph is illustrating the distribution of mobile apps across various categories. The most notable observation is the dominance of the “Games” genre, with its count significantly surpassing other categories. In contrast, “Stickers”, “Navigation”, and “News” are among the least populated genres. The distribution is varied: while a few genres like “Games” show extreme popularity, many genres exhibit moderate or low counts. Genres such as “Book”, “Business”, and “Weather” occupy a mid-range position in the distribution. The design adopts uniform shades of blue for each bar, emphasizing the differences in bar heights. In essence, the chart reveals a skewed distribution of apps, highlighting the prominence of entertainment-focused apps and suggesting potential market saturation or user preference trends in the app industry.

Total Revenue vs Genre of Apps

Show

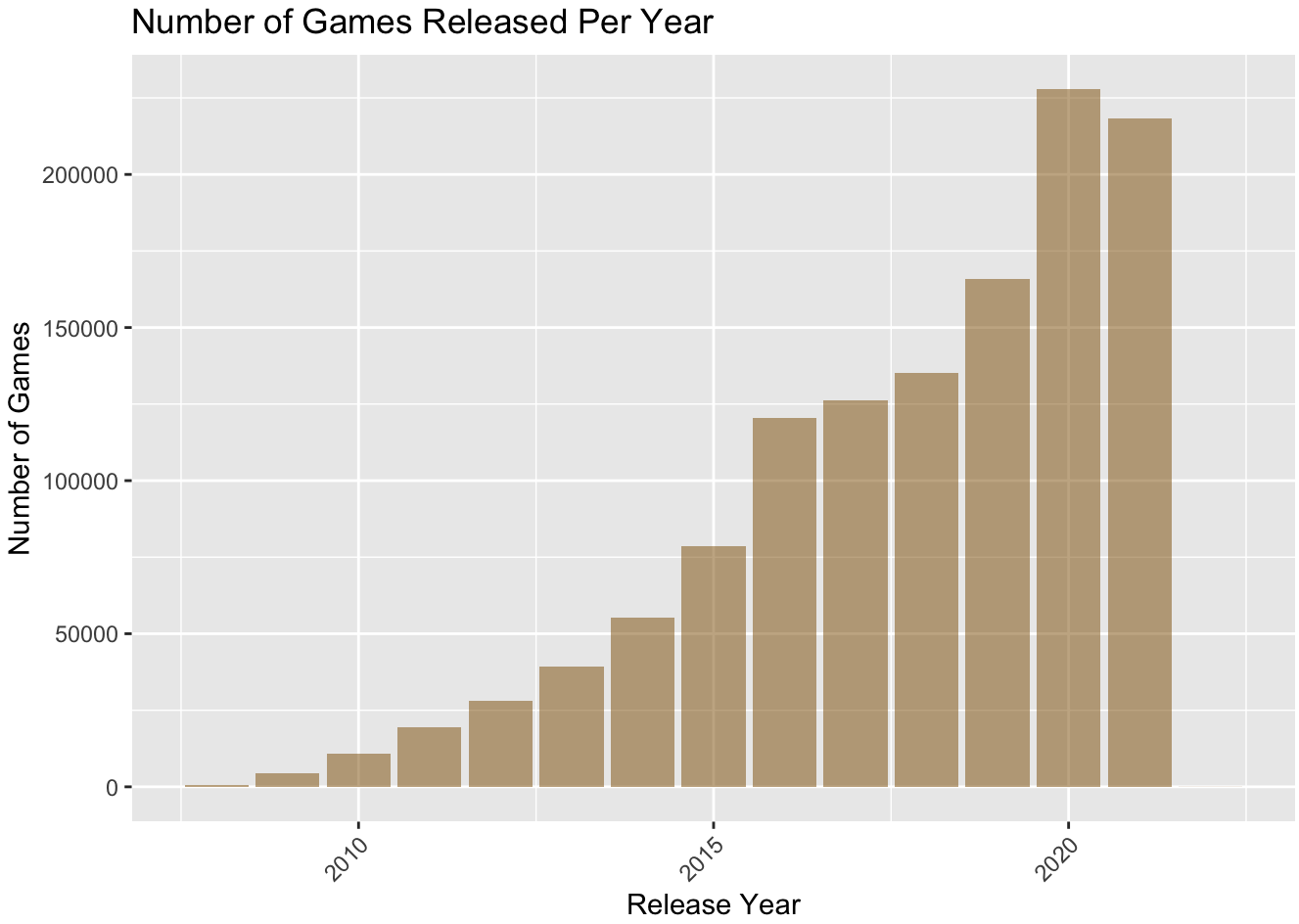


The chart presents the distribution of estimated revenue across various app genres. Visually, the distribution is notably right skewed. The “Games” genre stands out significantly, generating the highest revenue by a large margin. Following it, genres like “Photo & Video”, “Business”, and “Education” have a noticeable but considerably lower revenue. As we traverse to the right, many genres such as “Developer Tools”, “Magazines & Newspapers”, and “Stickers” show minimal revenue. This disparity suggests that while “Games” are dominant revenue earners, other genres struggle to achieve comparable financial success. The consistent green shade for bars accentuates the stark differences in revenue heights across genres. For EDA purposes, this chart vividly illustrates the prevailing revenue trends in the app market, emphasizing the disproportionate contribution of the gaming sector. It’s essential to note that the depicted revenue was estimated by multiplying the number of reviews with app prices, as there was no direct “Revenue” column.

Gaming apps

Frequency over years

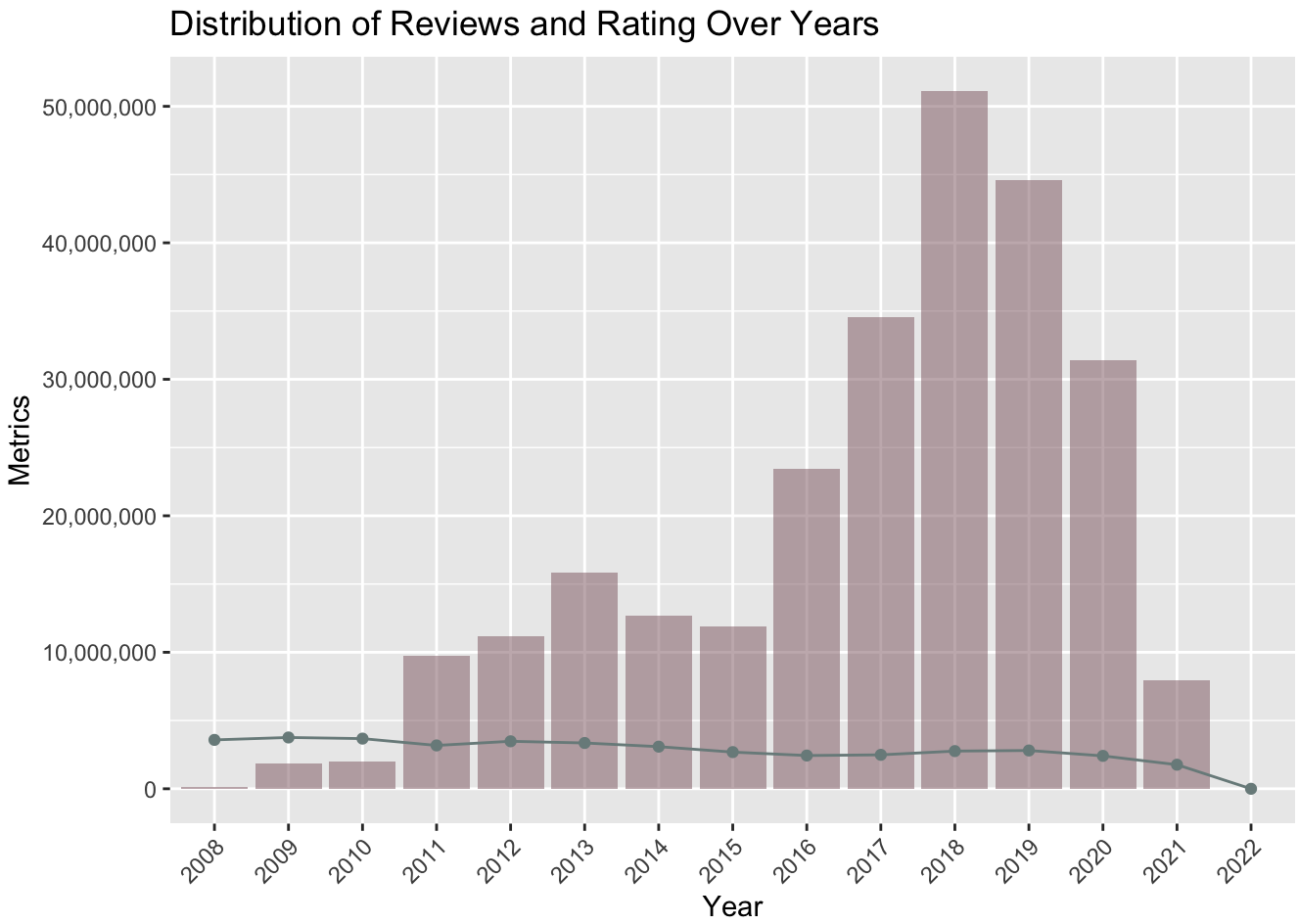
Show



The graph depicts the annual trend of game releases from 2010 to beyond 2020. The distribution showcases a pronounced upward trend(left skewed). Starting from 2010, where game releases are notably few, there's a consistent and almost exponential rise in the number of games released each year. Around 2015, there's a slight dip in releases, but the general trend remains ascendant. Post-2015, the growth appears more linear until around 2020, where a minor decline is observable. Notably, the highest number of games was released just before 2020, nearly reaching 200,000. The shade of the bars darkens progressively, mirroring the growth in numbers, reinforcing the visual impact. This EDA suggests that the gaming industry has seen remarkable expansion over the decade, with developers increasing their output every year, underscoring the booming nature of the gaming sector during this period.

Reviews & Rating vs Year

Show

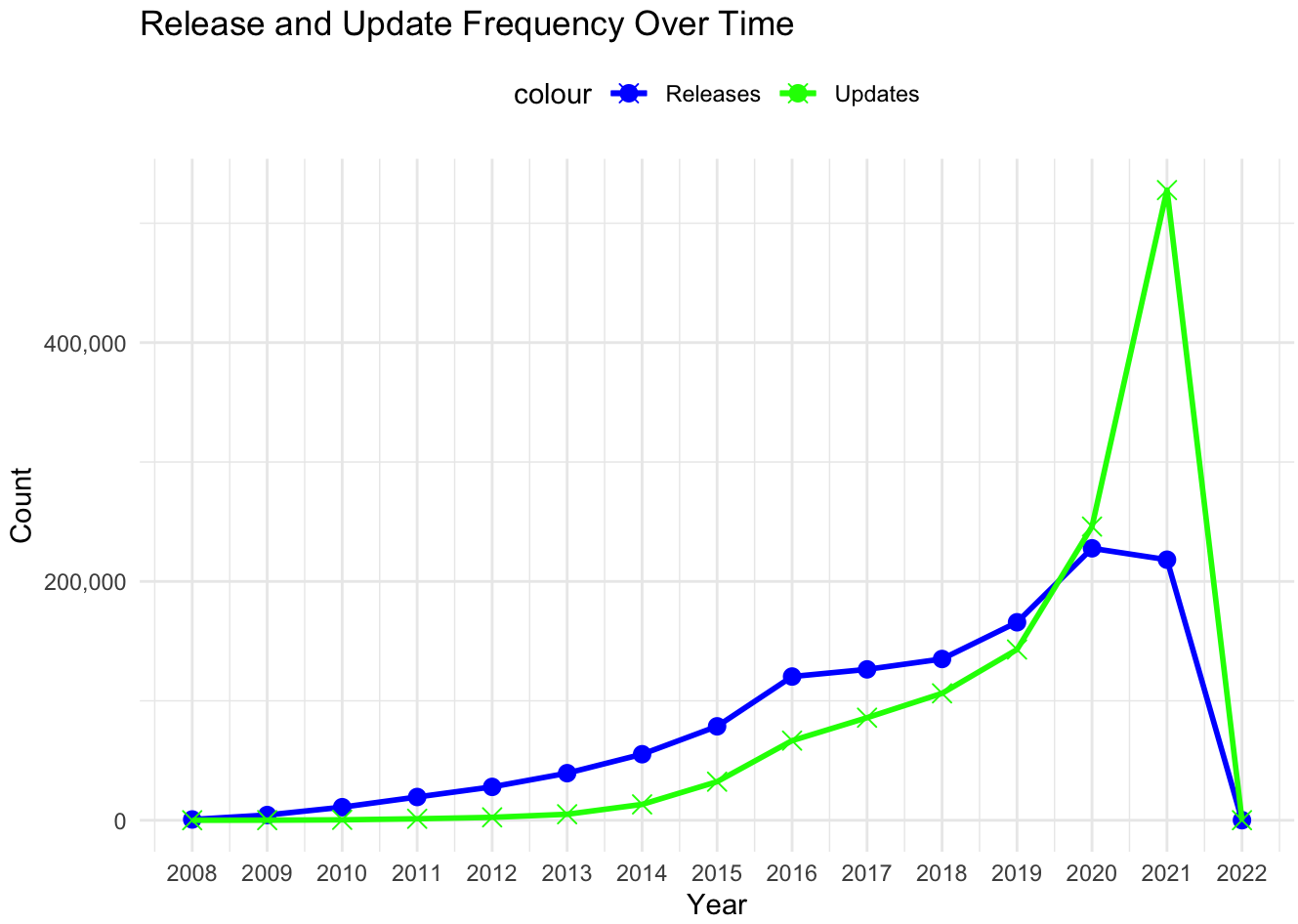


The graph displays the evolution of two key metrics from 2008 to 2022: total reviews (represented by bars) and average rating (illustrated by the line graph). From 2008 to 2014, the total reviews remained relatively low, with slight fluctuations but started showing a notable increase from 2015 onwards. The peak occurred around 2019 with nearly 50 million reviews, followed by a decline in 2020 and 2021. The average rating, although scaled for visualization purposes, remained fairly stable over the years, hovering around a constant value, indicating a consistency in user feedback despite the varying number of reviews.

The EDA unveils a period of explosive growth in the total reviews from 2015 to 2019, suggesting increased user engagement or product releases. However, the steady average rating underscores that while the quantity of feedback changed, the qualitative aspect of the products or services maintained a consistent standard over the years.

Release and Update Frequency over the years

Show



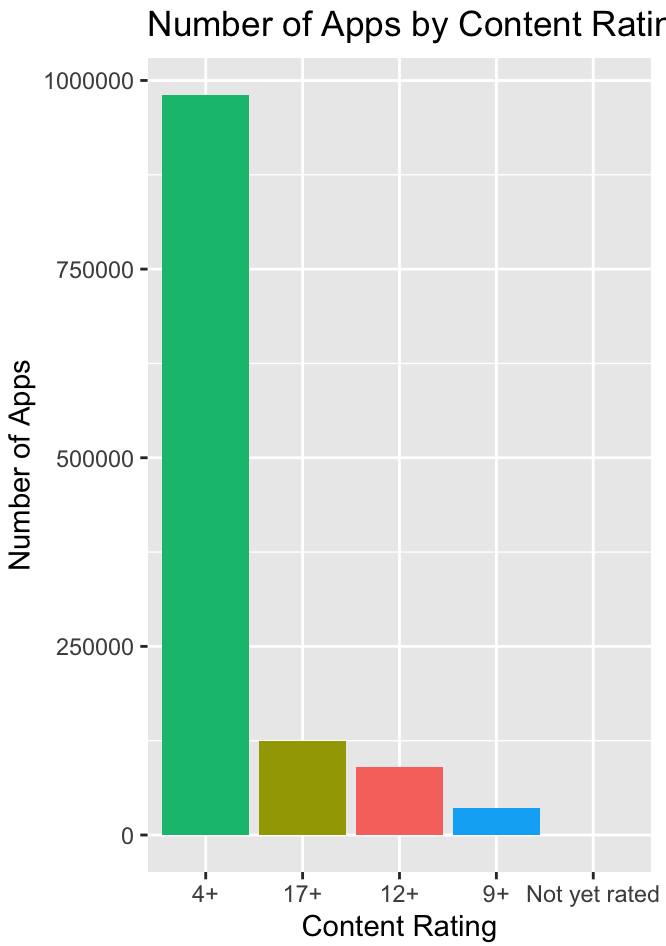
The graph illustrates the evolution of app releases and updates from 2008 to 2021. For much of the early period (2008-2016), the frequencies of both updates and releases were relatively low, with a slight edge for updates. A discernible growth in both metrics begins post-2016. The number of releases followed a moderate upward trajectory, while updates saw a steeper rise.

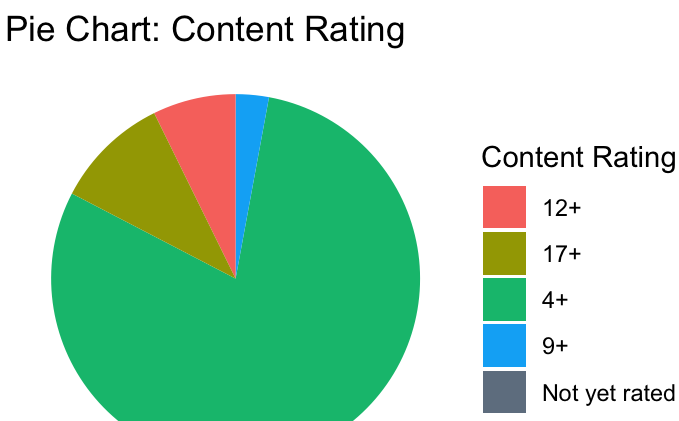
A significant point of interest is the pronounced spike in updates in 2020, reaching nearly 400,000. This surge aligns with the onset of the COVID-19 pandemic and the consequent lockdowns. With the world shifting to online platforms for myriad purposes, there was an evident heightened demand for app functionalities and refinements. Developers seemingly responded to this demand, resulting in the spike. Though updates receded post-2020, they stayed well above the numbers seen in the pre-pandemic years. The data for 2022 is excluded, as observations are available only up to 2021.

Content Rating

Frequency and Distribution

Show





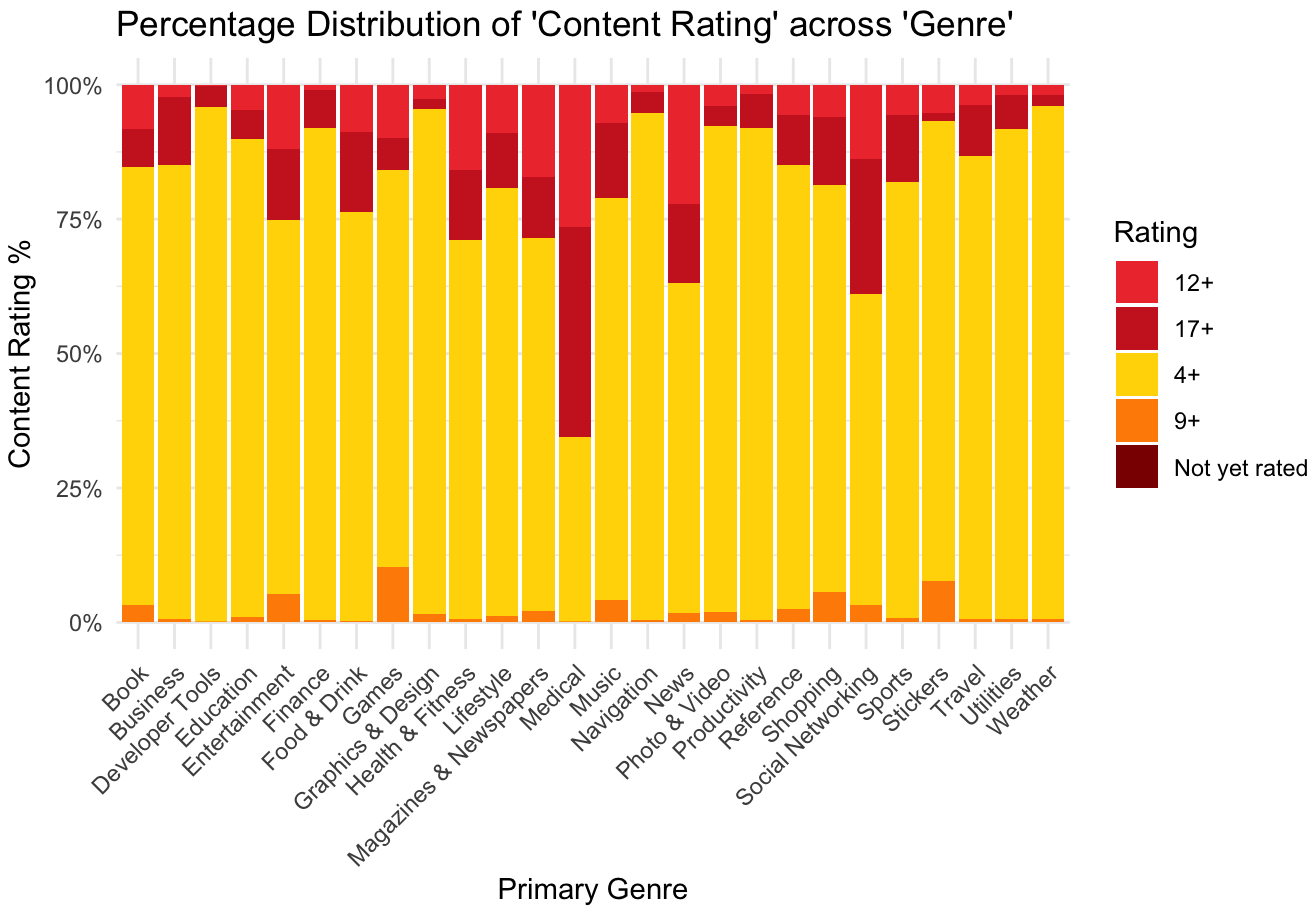
Show

```
##   Content_Rating  Count Percentage
## 1             4+ 980969      79.73
## 2             9+  35697       2.90
## 3            12+  89569       7.28
## 4            17+ 124127      10.09
## 5  Not yet rated     10       0.00
```

The bar chart shows the prevalence of apps whose minimum maturity rating is 4+ among apps with content rating of 9+, 12+ and 17+. While checking for the distribution of 4+ apps using a pie chart we can see that about a whopping 80% of the market is dominated by them. But this still does not give us the full story on whether all of these apps are from just a few genres like Games, Education etc.

Genre vs. Content Rating

Show



Stacked Percentage Bar Chart

This graph hows that apps belonging to the 4+ category take up almost over 75% of the total apps in each category and are almost evenly distributed in all the genres with a few exceptions in Medicine, Social Networking and News.

Pearson’s Chi-squared and ANOVA tests

Pearson’s Chi-squared test between Genre and Content_Rating

Show

```
##
##  Pearson's Chi-squared test
##
## data:  contingency_table
## X-squared = 2e+05, df = 100, p-value <2e-16
```

The Pearson’s Chi-squared test is a statistical test that determines if there’s a significant association between two categorical variables. In this instance, the test was conducted to examine the relationship between “Primary Genre” and “Content_Rating”.

Null hypothesis (H_0)- No dependence between the two variables
Alternate hypothesis (H_1) - Casual dependence exists between the two variables

The test output provides several key pieces of information:

1. **X-squared**: The chi-squared statistic is 167,518. This value represents the discrepancy between the observed frequencies in the contingency table and the expected frequencies under the assumption that “Genre” and “Content_Rating” are independent.
2. **df**: Degrees of freedom, which is 100 in this case. It’s an important parameter that determines the shape of the chi-squared distribution and is calculated based on the dimensions of the contingency table.
3. **p-value**: The p-value is less than 2.2e-16, which is virtually zero. A p-value this low suggests that the observed data is highly inconsistent with the assumption of independence between the two variables. Thus we reject the null hypothesis and adopt the alternate hypothesis.

In conclusion, even though the graph depicts an almost even distribution between “Genre” and “Content_Rating”, the proportions of Content Rating are not uniform across different genres. This disparity, possibly influenced by specific genres, results in a statistically significant association between the two variables. Hence, knowing an app’s genre might give insights into its content rating and vice versa.

ANOVA test for Primary Genre vs. Average UserRating

Show

```
##                    Df  Sum Sq Mean Sq F value Pr(>F)
## Primary_Genre      25  250995   10040    2263 <2e-16 ***
## Residuals     1230346 5458134       4
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The results of ANOVA (Analysis of Variance) test comparing the `Primary_Genre` to the `Average_User_Rating` depicts.

1. **Primary_Genre**: Represents the different levels or categories of the `Primary_Genre` variable. The `25` under this indicates there are 25 different genres being compared.
2. **Df (Degrees of Freedom)**:
   For Primary_Genre: 25, which means there are 26 genres being compared (df is one less than the number of levels).
   For Residuals: 1230346, representing the variability within each genre group.
3. **Sum Sq (Sum of Squares)**:
   For Primary_Genre: 250995, represents the between-group variability.
   For Residuals: 5458134, represents the within-group variability.
4. **Mean Sq (Mean Squares)**:
   For Primary_Genre: 10040, the average variance explained by genres.
   For Residuals: 4, the average variability within each genre.
5. **F value**: 2263 indicates the test statistic used to determine if the means of several groups are equal. It’s calculated by dividing the Mean Sq of Primary_Genre by the Mean Sq of Residuals.
6. **Pr(>F)**: Indicates the p-value, which is less than 2e-16 (extremely small), signifying strong evidence against the null hypothesis.
7. **Signif. codes**: A shorthand to describe the significance of results. The `<2e-16 *` indicates that the results are highly significant.

Null Hypothesis (H_0): The average user rating is the same across all primary genres.

Alternate Hypothesis (H_1): The average user rating is not the same for at least one primary genre when compared to others.

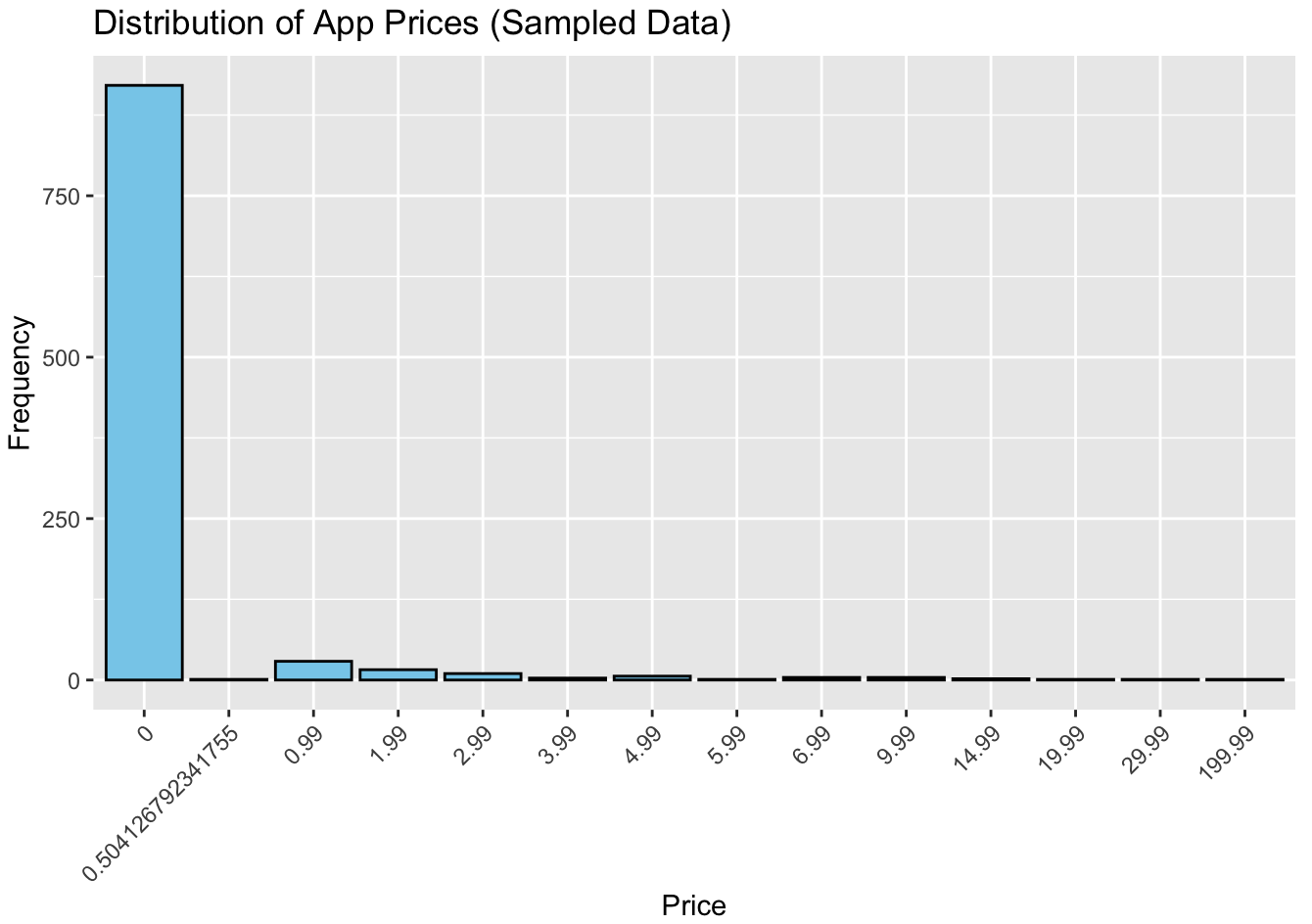
Given the extremely small p-value (much less than 0.05), we would reject the null hypothesis, suggesting that there's a statistically significant difference in the average user rating across different primary genres.

Analysis of Price

Frequency

Frequency of 'Price' for sampled data

Show

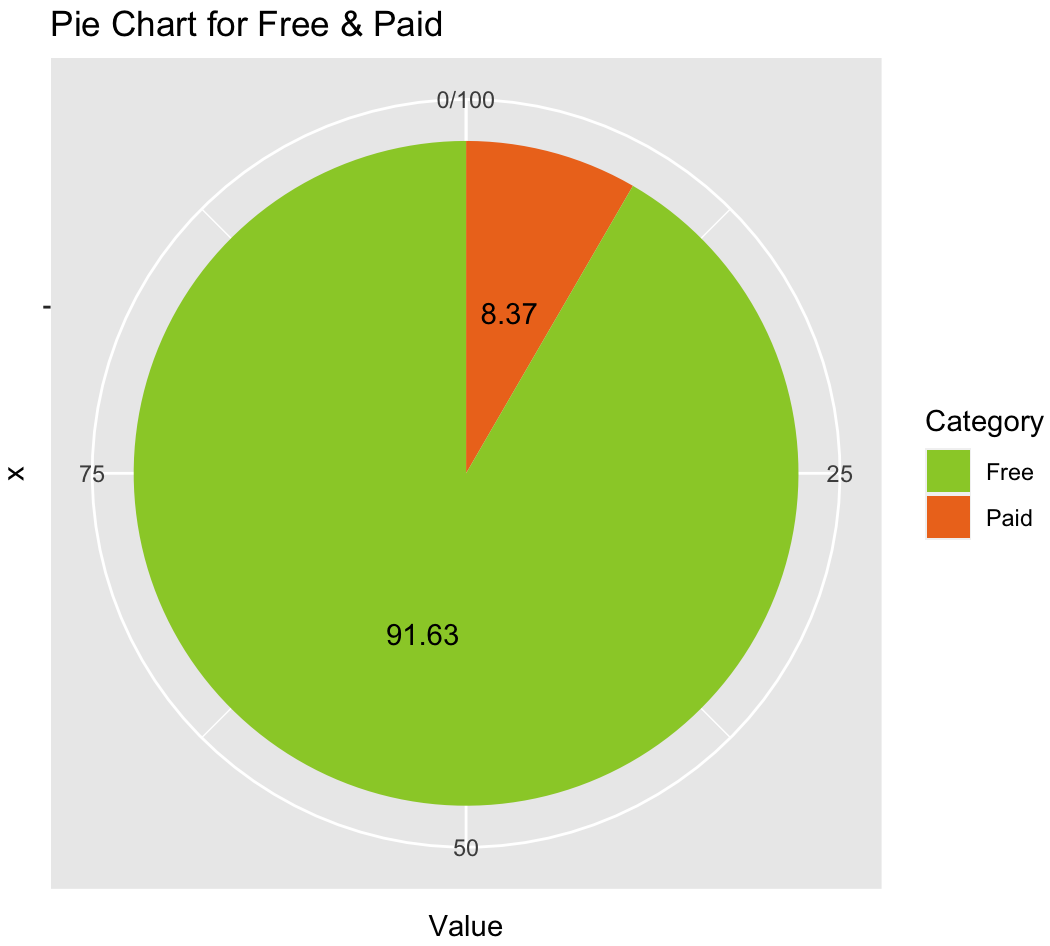


The above bar chart shows the distribution of Price (in USD) for a sampled dataset. A staggering 90% of the market consists of apps that are free (Price = 0). There is an instant drop from price 0 to about 1 USD. Followed by a steady decline as the price increases.

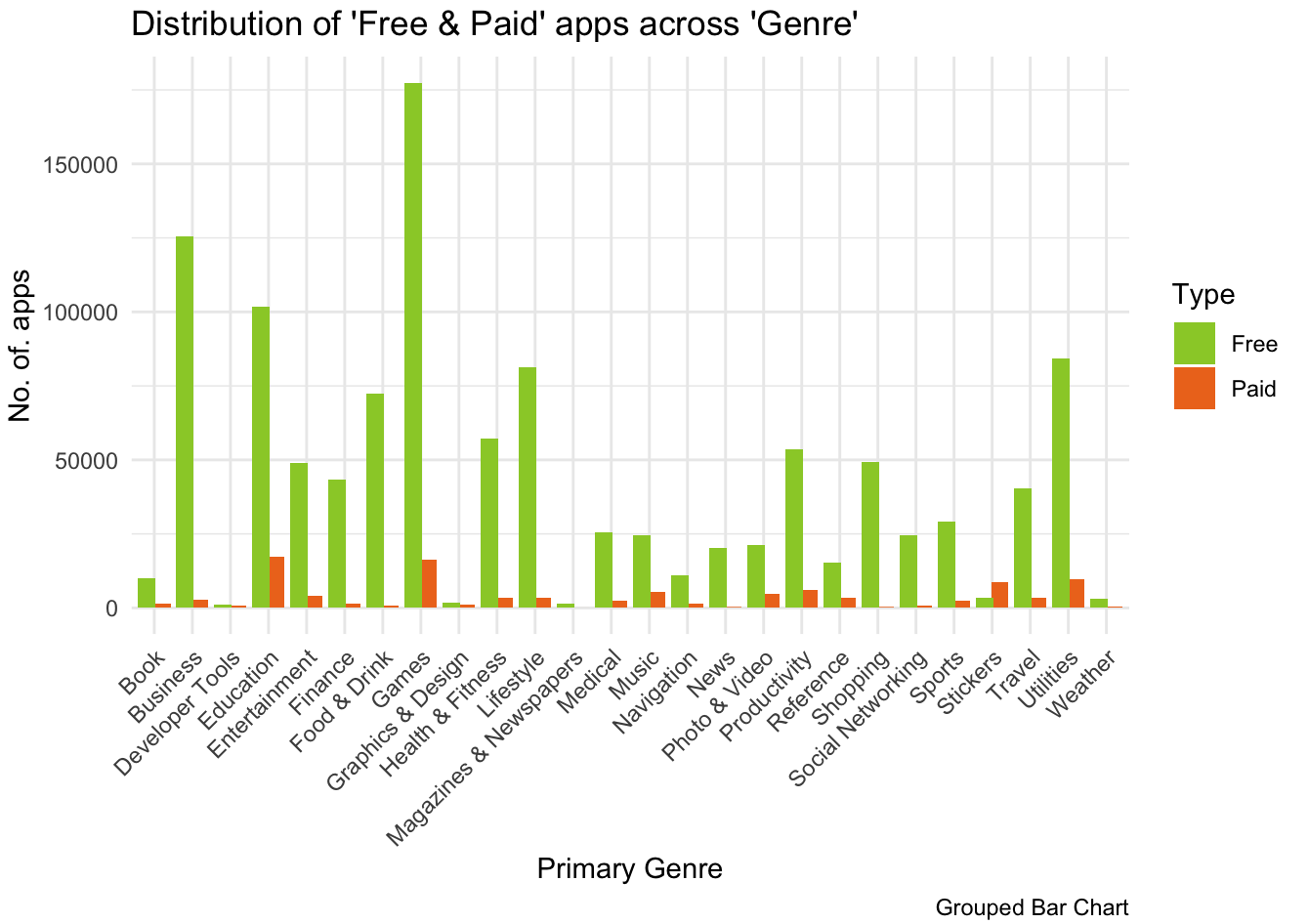
Distribution

Distribution of Free and Paid apps

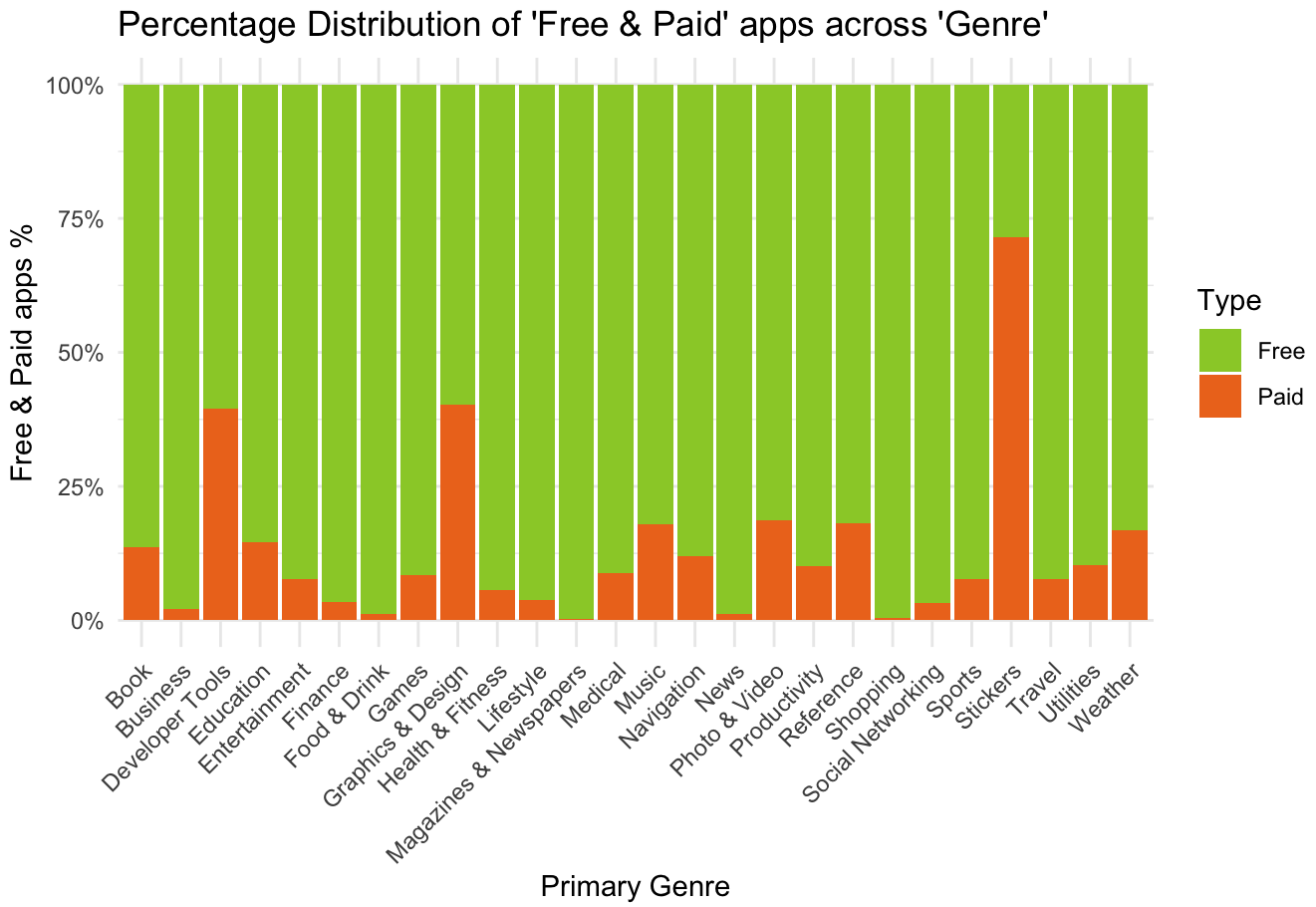
Show


Pie Chart for Free & Paid
0/100
25
50
75
8.37
91.63
Value
Category
Free
Paid

Show


Distribution of 'Free & Paid' apps across 'Genre'
No. of. apps
150000
100000
50000
0
Book
Business
Developer Tools
Education
Entertainment
Finance
Food & Drink
Games
Graphics & Design
Health & Fitness
Lifestyle
Magazines & Newspapers
Medical
Music
Navigation
News
Photo & Video
Productivity
Reference
Shopping
Social Networking
Sports
Stickers
Travel
Utilities
Weather
Primary Genre
Type
Free
Paid
Grouped Bar Chart

Show



Stacked Percentage Bar Chart

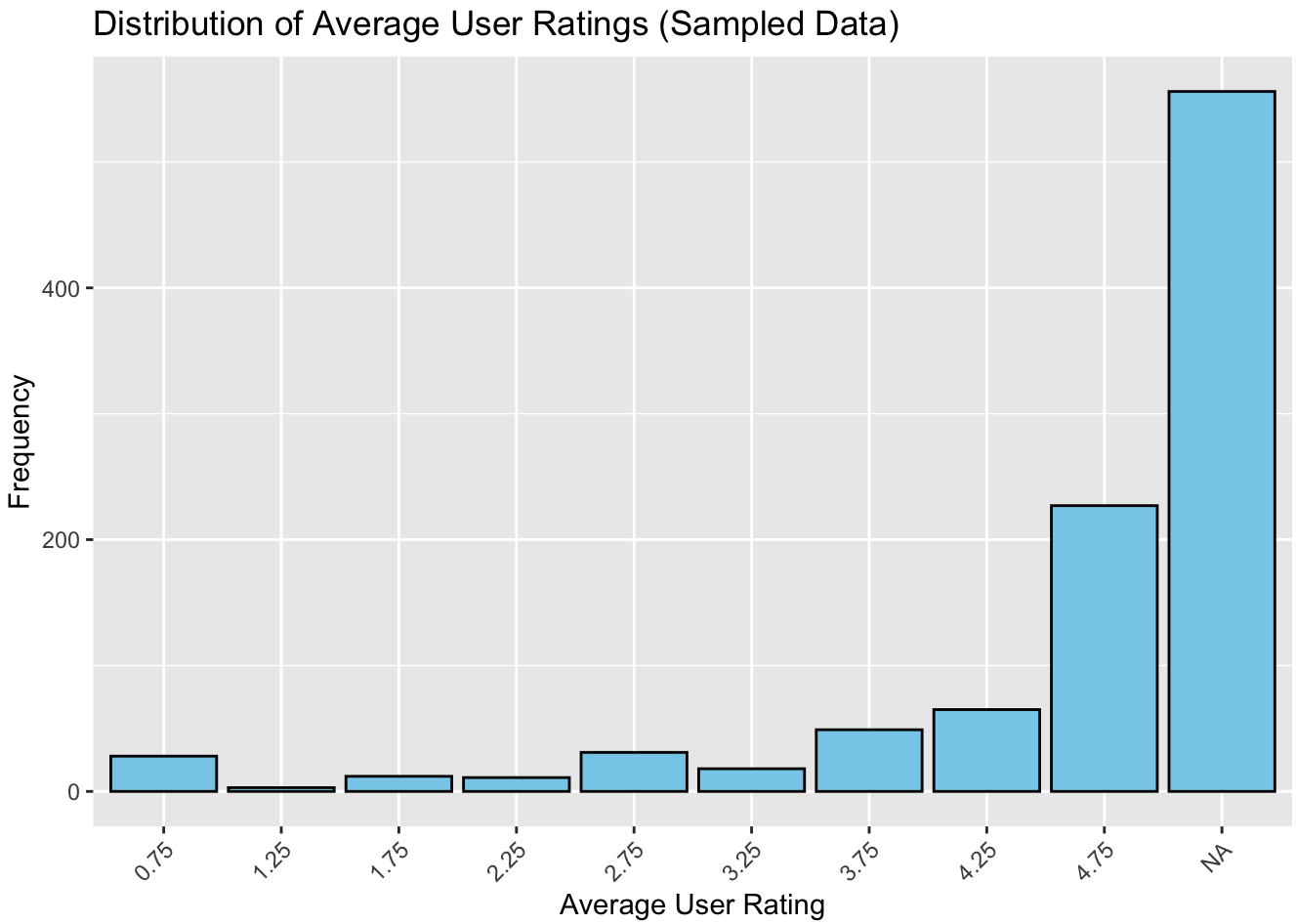
The pie chart shows that 8.37% of the content is free, while 91.63% of the content is paid.It's evident that the majority of apps on the App Store are free, with only a smaller portion being paid. The bar chart shows the distribution of free and paid apps across genres in the Apple App Store.The most popular genre is Games, with over 500,000 free and paid apps combined whereas Books, Developer tools, Magazines & Newspapers genres have the fewest free and paid apps.The number of free apps in each genre is typically several times higher than the number of paid apps. The stacked percentage bar chart shows the percentage distribution of free and paid apps across genres. We see that stickers app despite accounting for a very little percentage of the market, has the highest paid apps followed by developer tools and graphics & design.

Analysis of Ratings

Frequency

** Frequency of 'Average_User_Rating' for sampled data**

Show



From the bar plot obtained from the above, we could see from the distribution of the average user rating that there is a big disparity between ratings that are given and ratings that are not given (NA). Additionally, we can infer that the majority of users either provide ratings above 4.25 or do not provide any ratings.

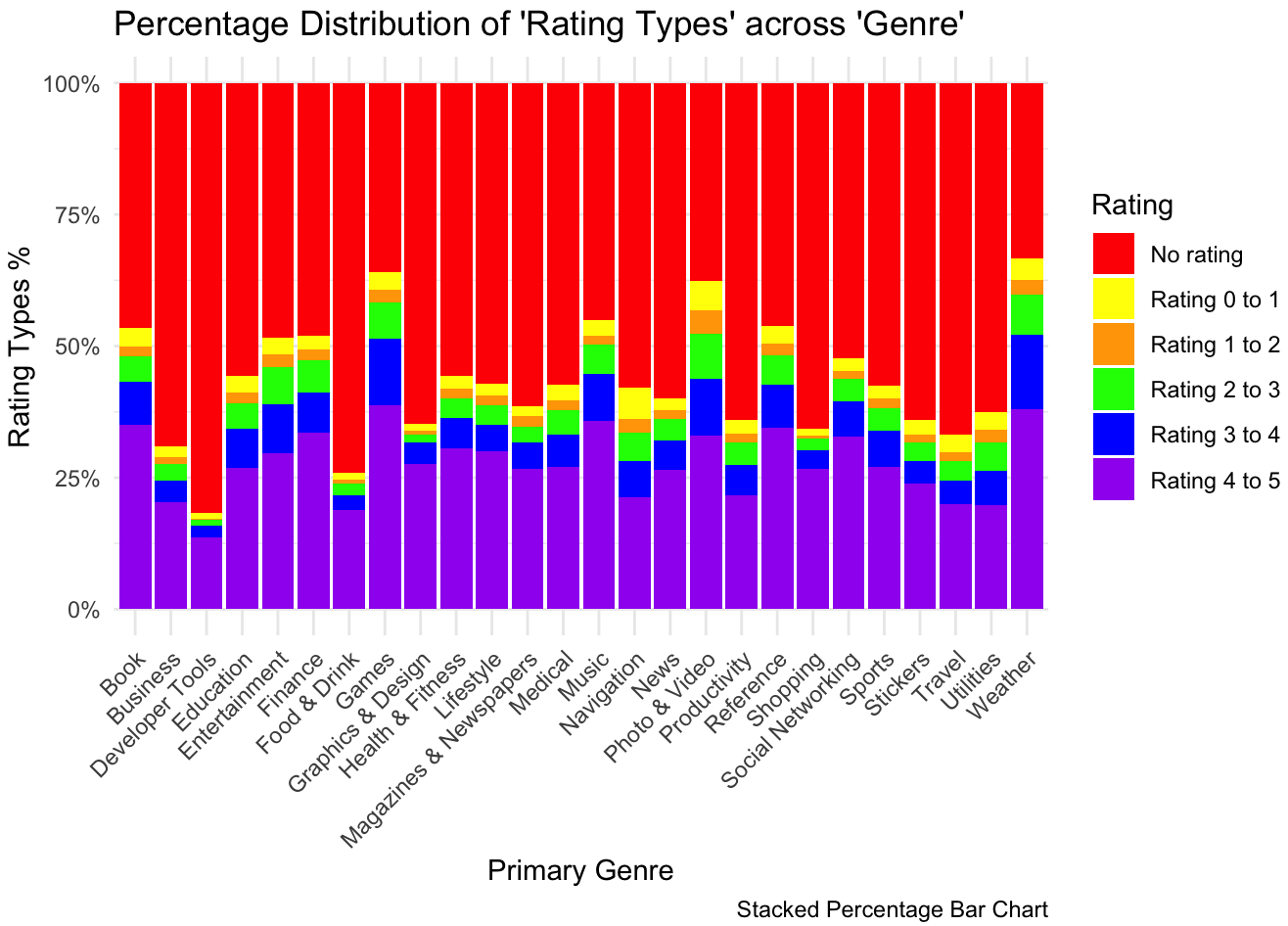
Creating User_Rating_Type column for various ranges of the ratings

Show

Distribution

Distribution of Rating types across Genre

Show

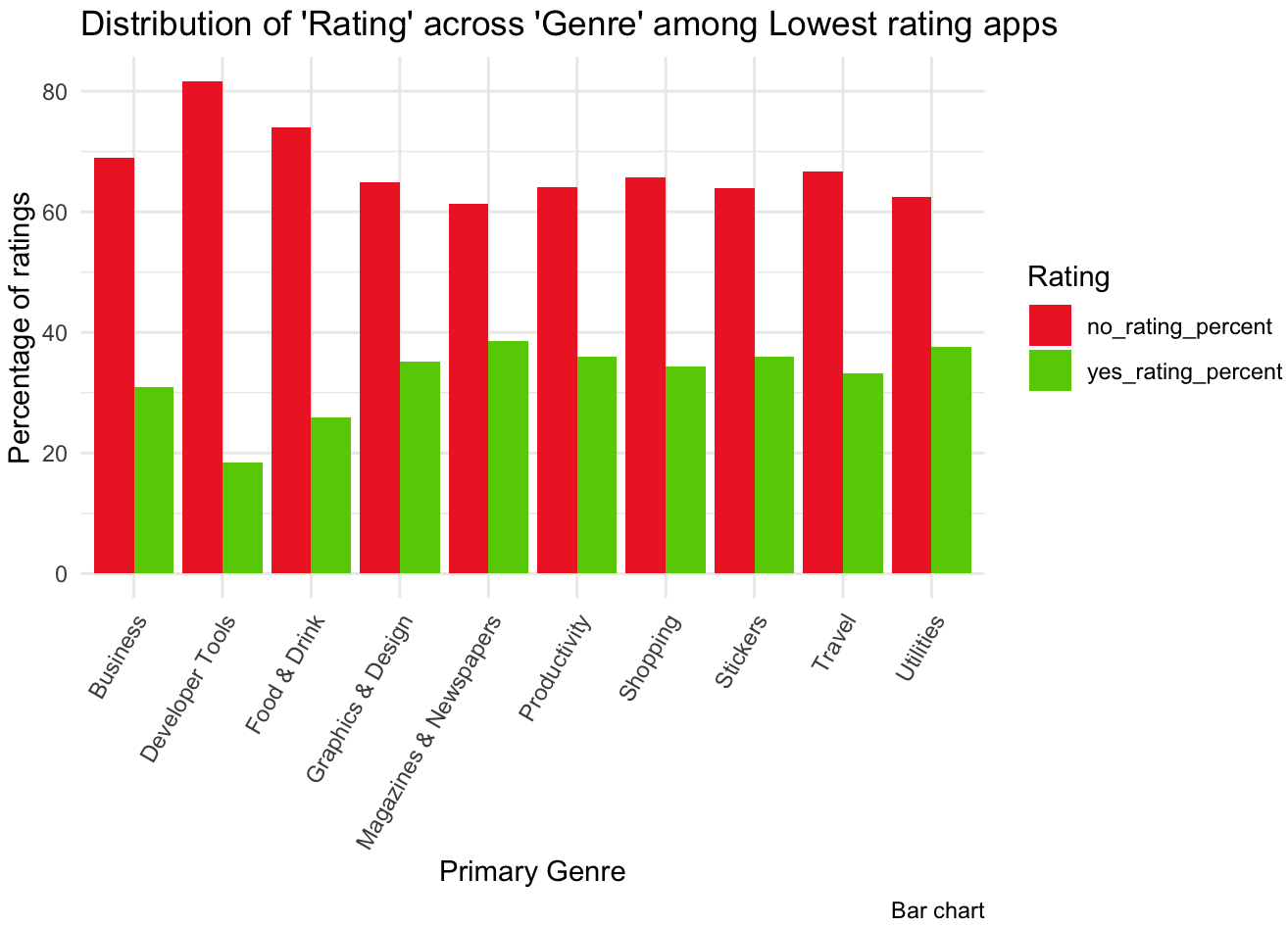


Stacked Percentage Bar Chart

From the above plot, according to category, the most no ratings (NA) are contributed by developer tools and food and drink applications, whereas weather apps and gaming apps contribute a higher number of ratings, indicating high user involvement.

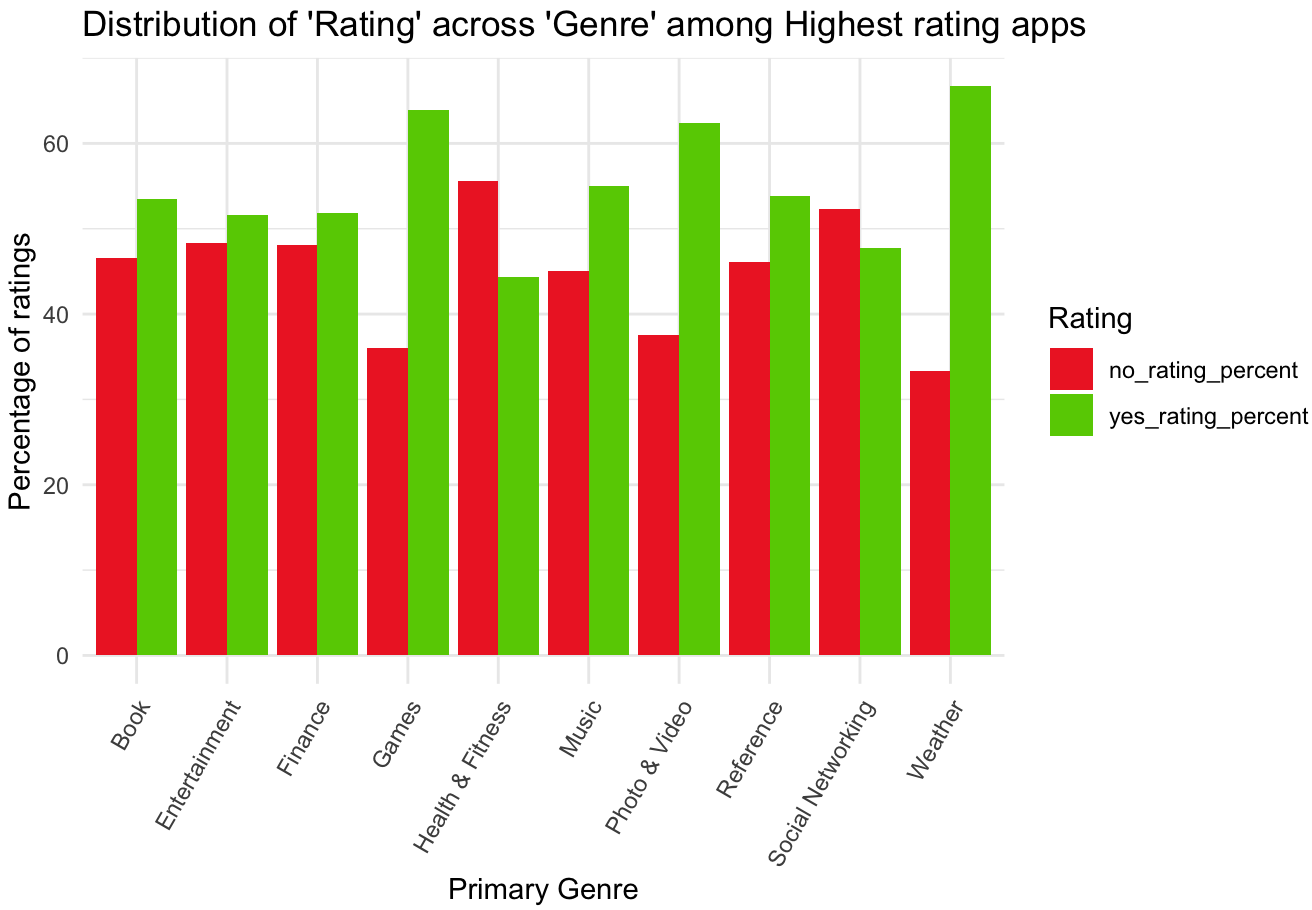
Most and least rated

Show



Bar chart

Show



Bar chart

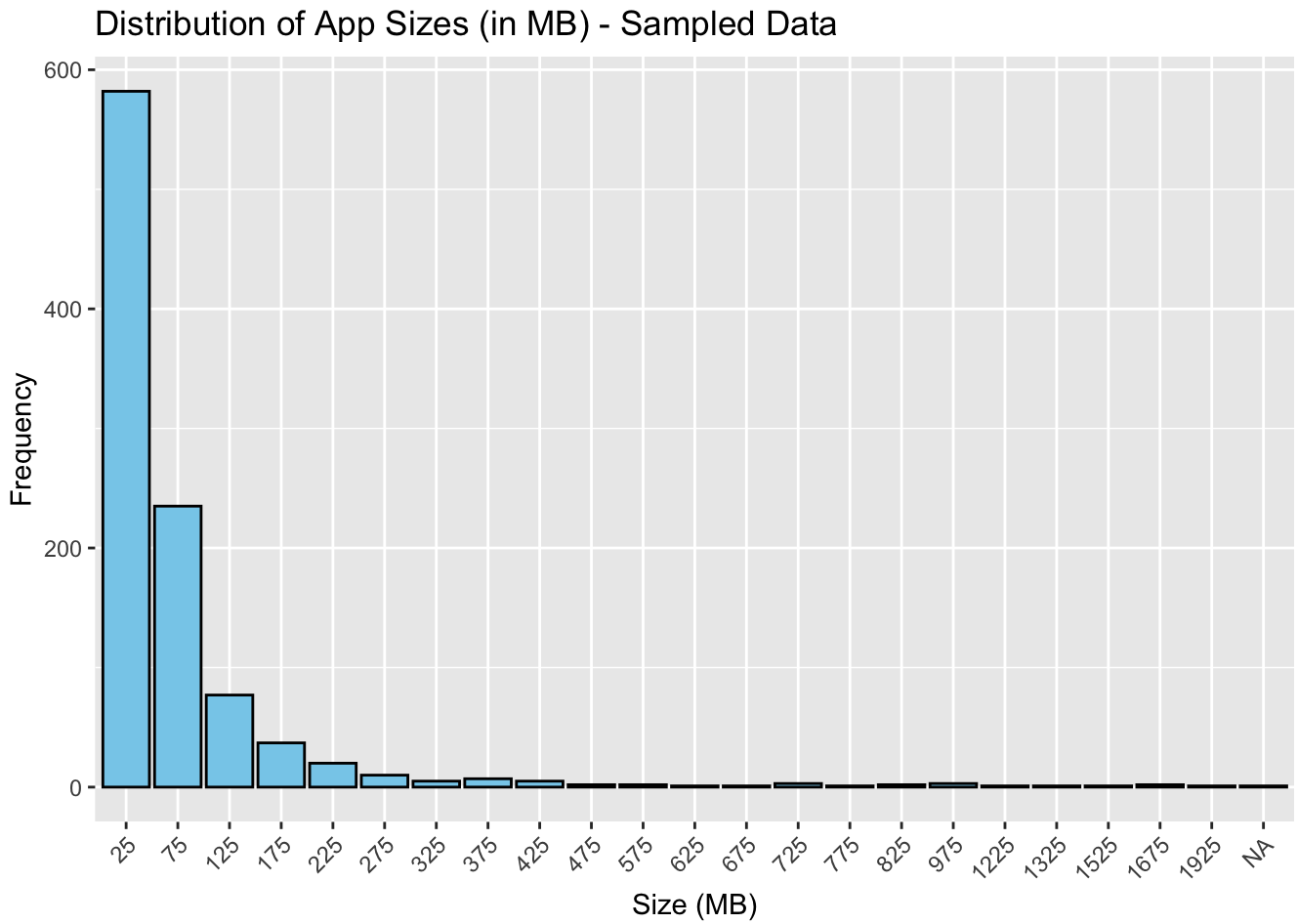
According to the aforementioned plots, the highest rated apps distribution of “Rating” across “Genre” shows that photo and video apps, games, and weather apps all contribute significantly to higher ratings. Developer tools and food & drink apps give the majority of the rating in the no ratings category and the least amount of rating in the ratings category among lowest rated apps, respectively, in the distribution of “Rating” across “Genre” among lowest rated apps.

Analysis of App size

Frequency

Frequency of ‘Size_Bytes’ for sampled data

Show



The Distribution of App Sizes (in MB) graph from the aforementioned plots shows that the bulk of apps fall into the 0–25 MB and 25–75 MB groups. This suggests that the majority of apps are modest, with usual sizes under 25MB, making the download and storage of apps easy on the device.

Key Takeaways from the Mobile App Market Analysis:

1. The gaming genre is ruling the app world, capturing a significant share of the market:
    - Within the mobile app ecosystem, gaming apps are the dominant force. They have a substantial market share compared to other types of apps. This could be due to the popularity of mobile gaming, the variety of game genres available, and the high user engagement that gaming apps tend to attract.
2. Apps with content rating of age 4+ are prevalent, constituting a whopping 80% of the market:
    - This conclusion highlights that a significant portion of the app market is occupied by applications that has a minimum maturity of ages 4 and above. Despite the genre category, these apps take up almost over 75% of the total apps in each category with exception for a few. This could suggest a high demand for kid-oriented entertainment and educational tools.
3. Free apps occupy a staggering 90% of the market:
    - A significant majority of apps are available to users at no cost. This may be driven by various monetization strategies such as in-app advertisements, in-app purchases, or subscription models, which enable developers to generate revenue without charging upfront.
4. Weather apps receive the most user ratings, while developer apps tend to have fewer ratings:
    - Weather apps receive a high number of user ratings, indicating that users are actively engaged with these apps. It's likely that weather apps provide real-time information that users find valuable. On the other hand, apps developed by developers have fewer ratings, possibly indicating that they are niche or used by a specific audience.

5. Most apps are compact, with sizes typically falling below 25MB, ensuring easy downloads and storage:
    - Majority of apps are designed to be lightweight having relatively small file sizes. This is a user-friendly approach as it minimizes the use of the device's internal storage space and ensures that apps can be quickly downloaded even with limited internet bandwidth.

These conclusions collectively provide insights into the current state of the mobile app market, including the dominance of gaming and children's apps, greater user engagement with weather apps, the prevalence of free apps, and the importance of keeping app sizes manageable for users.