Reflective Diary

Business Intelligence and Visualization

Google Play Store Data

Dhanushka Gunasinghe

23052728

This reflective diary document chronicles my learning journey of the Google Play Store App data analysis undertaken as part of the Business Intelligence and Data Visualization module. This reflects the key milestones from initial data interaction to the final presentation and describes entire learning opportunities alone with the thought process. Formative feedback and the learning experience shaped business intelligence and data visualization abilities, analytical skill and the outcome. This entire journey describes key milestones as below.

1st Milestone - Data Exploration, Data Cleaning and Understanding

Introduction

Google play store App data is publicly available dataset that provides valuable information and insights to different set of stakeholders and different format of scenarios. Analysis will reveal world trends on mobile and online consumer direction as well.

The initial step for any Data Analytic project is Data Exploration followed by Data Cleaning and Data Understanding.

Understanding the context of the data is very crucial before the analytical process. Successful data analysis will help us to achieve better results.

Pre-Stage Feeling

Data Exploration, Data Cleaning and Understanding in the journey of data exploration, data cleaning and understanding I feel with mix anticipation and curiosity. Interacting and dealing with the new data set is more existing and interesting but my previous experiences will extremely be helpful to accomplish this milestone successfully.
From my previous experiences I clearly understand the importance of this milestone and at this stage I must use my technical knowledge as well as critical thinking and the domain understanding for interpreting or the transformation the data effectively.

Approach

Data Exploration, Data Cleaning and Understanding is fundamentally crucial for marking the initial step of the any data related analytical process. So, I allocate significant time to explore the dataset, and this will help to make solid foundation to successful outcomes in the project. [1][2][3]

Google play store is one of the largest and most diversified mobile applications repositories and that hosted more than millions of android base mobile applications under various categories such as entertainment, education, health etc. Millions of mobile users use google play store to download mobile applications perform their day today tasks and google app store has good consumer feedback mechanism, rating and price strategy etc. With the other competitive platforms like Apple play store it's more challenging for access outside parties and difficult to find balance dataset using Apple play store data.

For the tasks of Data Exploration, Data Cleaning and Understanding I get my approach by interpreting using python plotting and numerical libraries such as Pandas, Matplotlib, Seaborn and NumPy which is highly effective in completing tasks efficiently and accurately. I strongly believe that seamlessly visualizing and interpreting data will help to avoid inconsistencies and enhance meaningful insights.[4]

Why do I need data explore, clean data and get better understanding of data? Can we analysis straightforward without above all?. [5] paper describes the importance of handling null values and missing data and clearly describes how statistical power reduced with the noisy data.

[6][7] research papers describe how to handle missing data according to the context. As example above two papers help to get broad knowledge about the methodologies of handling missing data including removing. Various imputation techniques and Interpolation

Solution execution

Initially I load the data set into python analytical environment and start from basic analysis. Here I listed my workarounds in detail.

```
In [13]: import pandas as pd
         import matplotlib.pyplot as plt

In [14]: data_path = 'D:/UWE/Data Visualization/Data Set/googlePlayStore.csv'

         data_googlePlay
```

Out[14]:

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Price ($)	Content Rating	Genres	Last Updated	Current Ver	Android Ver
0	Photo Editor & Candy Camera & Grid & ScrapBook	ART_AND_DESIGN	4.1	159	19M	10,000+	Free	0	0.0	Everyone	Art & Design	January 7, 2018	1.0.0	4.0.3 and up
1	Coloring book moana	ART_AND_DESIGN	3.9	967	14M	500,000+	Free	0	0.0	Everyone	Art & Design;Pretend Play	January 15, 2018	2.0.0	4.0.3 and up
	U Launcher Lite –											August		4.0.3

Image 1

- Image 1: Load the data from csv.

```
In [26]: playStore_null= pd.DataFrame({'Null Values' : data_googlePlay.isna().sum().sort_values(ascending=False),
                                      'Percentage Null Values' : (data_googlePlay.isna().sum().sort_values(ascending=False)) / (data_googlePlay.shape[0]) * (100)})
         playStore_null
```

Image 2

- **Image 2: Check for the Null Values**

	Null Values	Percentage Null Values
Rating	1474	13.596532
Current Ver	8	0.073794
Android Ver	3	0.027673
Type	1	0.009224
Price ($)	1	0.009224
Content Rating	1	0.009224
App	0	0.000000
Category	0	0.000000
Reviews	0	0.000000
Size	0	0.000000
Installs	0	0.000000
Price	0	0.000000
Genres	0	0.000000
Last Updated	0	0.000000

Image 3: [Columns with Null Values]

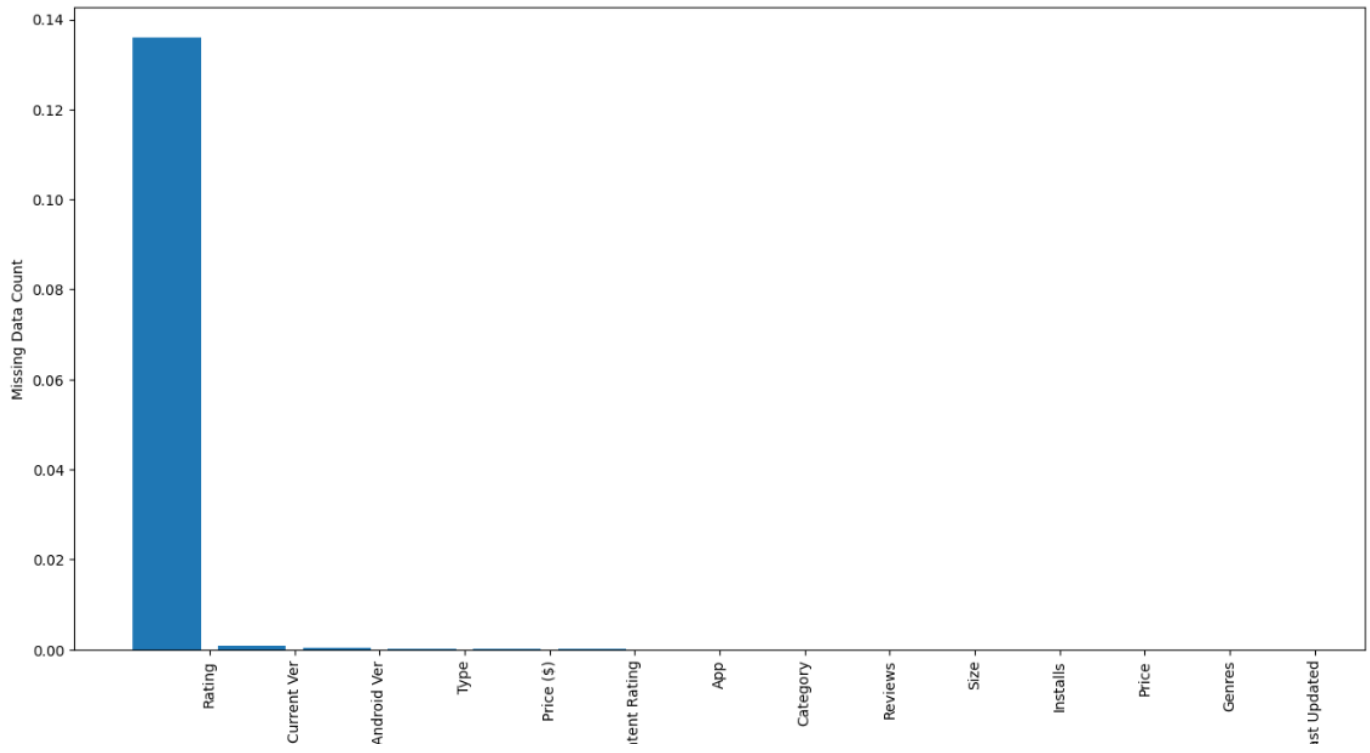
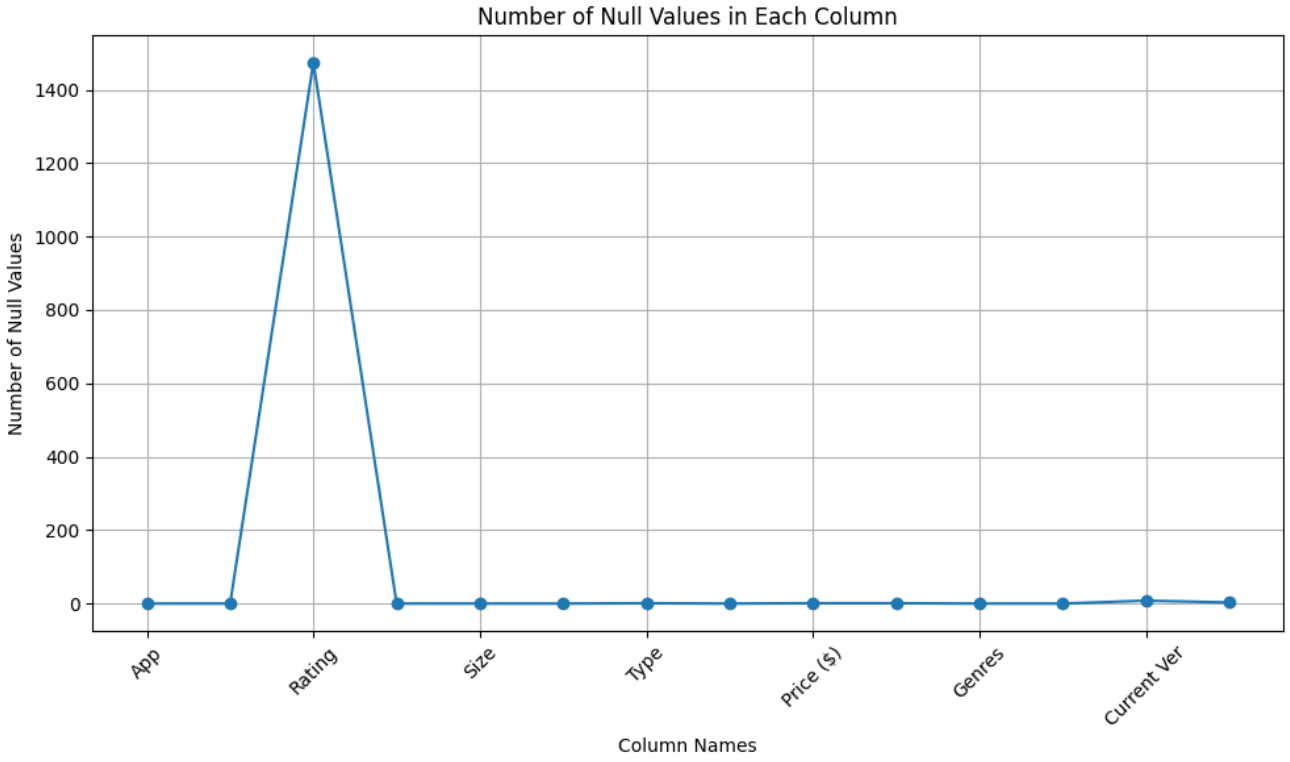


Image 4: [Null Value Counts with the feature columns]



According to null value analysis I found there 6 columns with null value Rating is with highest null values with 13.5% followed by the current android version contain only 0.07% of null values.

After full inspection of null values 'Rating' is a float datatype which contains a higher % of null values compared to the others. According to articles read related to the missing data handling [8] rating is significantly important aspects for our analysis specially for understanding application prospective. I decided to handle rating data with mean imputation which is less effect the original architecture of data. [Image5]

```
In [42]: #fill null value with rating mean
         mean_rating=data_googlePlay['Rating'].mean() # rating mean
         data_googlePlay['Rating'].fillna(mean_rating, inplace=True)

C:\Users\user\AppData\Local\Temp\ipykernel_1500\1576736188.py:3: FutureWarning: A value is trying to be set on a copy of
a DataFrame or Series through chained assignment using an inplace method.
The behavior will change in pandas 3.0. This inplace method will never work because the intermediate object on which we ar
e setting values always behaves as a copy.

For example, when doing 'df[col].method(value, inplace=True)', try using 'df.method({col: value}, inplace=True)' or df[col] = df
[col].method(value) instead, to perform the operation inplace on the original object.


 data_googlePlay['Rating'].fillna(mean_rating, inplace=True)
```

Image 5

. I decided to replace categorical variable ‘current version’, ‘android version’ and ‘type’ by mode [9] which is effective way to handle. [Image 6]

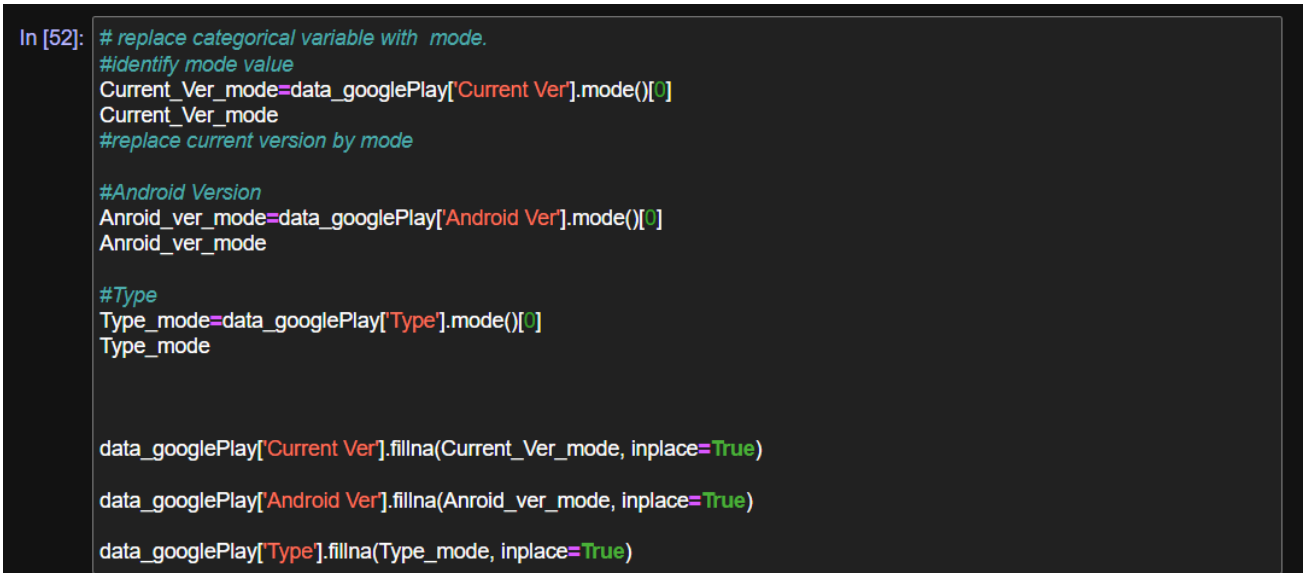


Image 6

Finally, I dropped ‘price’ and ‘content rating’ columns as below [Image 7]

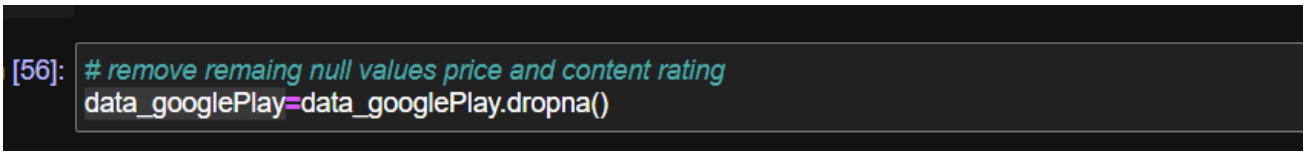


Image 7

After handling the missing values my next step in this milestone is explorer and pro processing the feature attributes, data types and data formats. After close inspection I expect to do below changers.

Feature Attribute	Existing Data Type	Expected Data Type	Existing Format	Expected Format	Comment	Reference
Reviews	Object	Int	Number Ex: 11	Number Ex: 11	Data type conversion using python	Image 8
Installs	Object	Int	Ex:10,100+	Ex: 10000	Data type	Image 9

					and format conversion on Tableau calculated fields by eliminate '+' and ',' for calculation purpose	
Price	Object	float	Ex: $420.32	EX: 420.32	Data type and format conversion using python by discard '$'	Image 10
Size	Object	Float	Ex. 120M	Ex 120	conversion on Tableau calculated fields by removing 'M' or 'K' and converted kilo bytes in Mega bytes.	Image 11

```
data_googlePlay['Reviews']=data_googlePlay['Reviews'].astype(int)
```

```
C:\Users\user\AppData\Local\Temp\ipykernel_1500\1703409144.py:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/index
ew-versus-a-copy
  data_googlePlay['Reviews']=data_googlePlay['Reviews'].astype(int)
```

```
data_googlePlay['Reviews']
```

```
0         159
1         967
2       87510
3      215644
4         967
        ...
10836      38
10837       4
10838       3
10839     114
10840  398307
Name: Reviews, Length: 10840, dtype: int32
```

Image 8

Install

```
INT(REPLACE([Installs],"+",""))
```

The calculation is valid. 6 Dependencies Apply

Image 9

```
data_googlePlay['Price'] = data_googlePlay['Price'].apply(lambda x: str(x).replace('$', '') if '$' in str(x) else str(x))
data_googlePlay['Price'] = data_googlePlay['Price'].apply(lambda x: float(x))
```

```
C:\Users\user\AppData\Local\Temp\ipykernel_1500\2137020522.py:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#retu
ew-versus-a-copy
  data_googlePlay['Price'] = data_googlePlay['Price'].apply(lambda x: str(x).replace('$', '') if '$' in str(x) else str(x))
C:\Users\user\AppData\Local\Temp\ipykernel_1500\2137020522.py:2: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
```

Image 10

Image 11

Google play store review's sentiment analysis is another important aspect to successful in-depth review. In this scenario sentiment analysis has main three categories positive, negative and neutral. For ease of the future analysis I convert above categorical features into as below.

Positive – 2

Negative- 1

Neutral – 0

Learning outcome

The initial stage of any data driven project, Data Exploration, Data Cleaning, and Understanding is in the top notch. I thought these steps were not a routine or traditional process of data analytics, but this helped to build basic fundamentals of the project.

To effectively achieve the Data Cleaning and Missing Data Handling process, I gone through the literature and found two important findings [10][11] on missing data handling and how those different techniques apply in different scenarios. Especially 'rating' has 13% of null occasions in total. So, I decided to use imputation with means other than drop, to ensure the power of statistics.

On another occasion, I employed mode imputation techniques, to fill the categorical variable 'type'. This method effectively addresses the gaps in the data set by replacing missing values with the most occurring instance and protects the statistical power of the dataset.

Meantime In this project I must options to handle missing data. The first option is handling outside the visualization environment. In three scenarios I used python code to handle missing data while in other occurrences I used calculated field to handle missing data inside the tableau work environment. [12][13] Tableau official guidance and documentation help me to succeed in this task.

I am extremely happy to learn the outcome and experience of this milestone for pre-better analysis.

Conclusion

Reflection of the Data Exploration, Data Cleaning, and Understanding milestone, I recognize the profound impact these initial stages have on the success of any data analytics project. I can recognize the importance of how crucial this in the success of data analytics. This is transformative learning experience for enchase effectiveness of future data- drive outcomes.

Post Stage Feeling and Future Thinking

Started with quite challengeable and later got good confidence. There is relief of accomplishments of the milestone with positive vibe and move forward with the process of planning and preparing for future stages or goals. It involves setting new objectives, stake holder understanding, scope definitions and strategizing on how to achieve them.

2nd Milestone – Identify Business Questions, Define Scope and Understand Stakeholders

Introduction

As we transition from the initial data exploration phase, our Business Intelligence and Data visualization on Google Play Store data analytics enters the critical second milestone. In this phase we are focused on identifying stakeholders first, and define business goals and scope accordingly. Hence this is a group assignment we focus on diversifying our tasks based on business questions in this stage.

Pre-Stage feeling

This stage is very crucial for successful implementation of the project with greater level of understanding among group members. So, l am working closely with the team and defining all the requirements is very important for this stage. Reflecting on these pre-stage feelings underscores the importance of a thoughtful and strategic approach, ensuring the effective team collaboration is directed towards delivering valuable and impactful insights.

Approach

Because of the novelty of the project and related domain I go through the literature behind it. According to this research [14] describes how to identify stake holders base on specific criteria such as

- Project involvement matrix
- Potential positive beneficiaries
- Potential negative beneficiaries
- Indirect or direct support measures
- Internal and external identities,

Defining the objectives, goals or the scope, is directly related to the stake holders' behaviors and requirements. [15] In this article describe how to ensure objectives set to expectations of stake holders using SMART technique. I also decided to use tangible and measurable set of objectives.

Solution Execution

Google play store is complex e-commence environment, which many stakeholders involve and beneficial in different ways. As a team we sit together and discuss and share our ideas and different views to obtain a final decision.

- App developers / UX UI Developers – existing app developers and potential developers are key stakeholders in the project. They want to get updates on the existing context of the environment.
- Product Managers – To understand the direction of market, user preferences and various performance matrixes will help product managers to enhancements of the products
- Sales and Marketing teams – Understand various sales and marketing related insights and use them to improve profitability of the products is very important. Such as understanding demographics diversification, age diversification, gender diversification and preference segmentation.

- Executives-Decision Markers- They look for high level of insights to make strategic decisions about resource allocation, investment opportunities and overall business strategies
- Regulatory Bodies – Legal and regulatory bodies interested in compliance handling, privacy protection and maintaining legal aspects of the apps.

After successful identification of stakeholders our next step was define business questions related to the each and individual stakeholders. We write down all the business-related questions to the purpose of later prioritization.[16] This research papers discussed how to understand business questions effectively with the root cause identification method.

Business Problems and Scope Definition

- App developers / UX UI Developers –
  - What is the most trending user engaging application segment over time?
  - How do rating and review reflect usability and application interaction?
  - How frequently do top rated and top reviewed applications change their implementation by updating?
  - How does the complexity of the app (e.g., size, number of features) impact user ratings and engagement?
  - What segments are users highly interested in, influence in going to paid versions?

Product Managers

- Identify potential areas of growth
- Understanding android version which are commonly used
- What are the future expectations based on reviews and ratings
- Which application version has higher rating distinguishes from other versions

- How does app size affect downloads and user engagement.
- How product rating navigates the future product road map.
- What are the spending patterns of users across different app categories?
- Analyze the successful version of the products by review as well as rating.

Learning outcome

At the end of this milestone, I am fortunate to learn complete 360-degree knowledge of google play store platform, their business models, strategies etc. Sametime, I get full knowledge of ecommerce business and monetization strategies.

This stage focuses on identifying stakeholders and their business problems and goals individually. To approach this milestone, I enchase my knowledge by relevant literature to understanding best methodologies and practice.

Throughout this process I clearly understand the importance of stakeholder identification and the importance of well define scope with well-marked objectives.

As a summary, Milestone 2 is significance of identifying and understanding stakeholders, define objectives and formulating business questions with the collaborative of the team with diversified ideas and feeling make room to sharpen my domain knowledge, business knowledge as well as greater learning outcomes for the next step.

Post Stage Felling

Upon reaching this milestone, I feel great and confident about the approaches we undertook. Even the pre stage feeling was uncertainty, and working as a team strengthened my abilities to acquire solid groundwork for our project. This process has underscored how these foundational steps were essential in navigating the complexities of stakeholder identification and scope definition.

3rd Milestone – Business Intelligence and Visualization on Google play store data set

Introduction

In this milestone I mainly focus on three steps. In data preparation, I will discuss managing data sources and the way data is handled inside the tableau. Next step is thoroughly analyzing all the data sources and identify the relationship between each other. Effective management of relationships between data sources enables us to move forward to Data Visualization with greater confidence.

Pre-Stage Felling

In this stage I have annoying about the use of new visualization tools, and I have doubt about learning curve of new application. I go through tableau help documentation which is extremely helpful to continue and make my pre-stage confidence in best level. Sametime, I decided go through my business intelligence lecture series and re do all the given exercises to boost my knowledge.

Approach

Base on the defined research problems identified necessary data requirements and import the relevant data files into tableau environment. The next step is analytics data file(s) (already cleaned and imputed) and Identify relationships of those data sources. After creation of necessary relationships, we can move toward to our next objective. Our next objective is the most important and more valuable for all of us stakeholders. We can start solving our business problems and challenges in tableau environments by implementing more insightful and attractive presence of information.[17][18]

Solution Execution

My first step was importing two sources of data into the tableau working environment.

- Google play store csv

- Google_user_review.csv

Two tables (csv Files) related by application name as I decide to explorer other field and datatype to ensure effectuate data understanding. For the further analytical purposes, I added some extra sheets and all the details,

listed below.

Sheet Name	Data Description	Purpose	Remark
Google play store	App name		Given [String]
Google play store	Category	App Segmentation	
Google play store	Rating	Customer Understanding	Given [String]
Google play store	Review	Customer Understanding	Given [Integer]
Google play store	Size	Size / Weight Determination	Given [Mb/Kb][String]
Google play store	Installs	Customer Understanding	Given Number [String]
Google play store	Price	Paid/free	Given [String]
Google play store	Last Update	App engagement	Give [Date]
Google play store	Current Ver	App engagement	Give [String]
Google play store	Android Ver	App Compatibility	Give [String]
Google play store	Rating_1		Created [Int]
Google play store	Price_1		Create[float]
Google play store	Review		Created [Int]
Google play store	Type_Subcat		Created
Google_user_review	Sentiment	Customer Understanding	Given[String]
Google_user_review	Sentiment_Polarity	Customer Understanding	Given [Float]
Google_user_review	Sentiment_Subjectivity	Customer Understanding	Given [Float]
Google_user_review	Count_neg	Customer Understanding	Created [float]
Google_user_review	Count_neutral	Customer Understanding	Created [float]
Google_user_review	Count_pos	Customer Understanding	Created [float]

In this level my all data is ready, before the moving directly into the problem-solving visualization.

First, I move forward with App developers / UX UI Developers prospective answer the business questions with visualization.

- **Business Problem: What is the most trending user engaging application segment over time?**

- **Visualization**

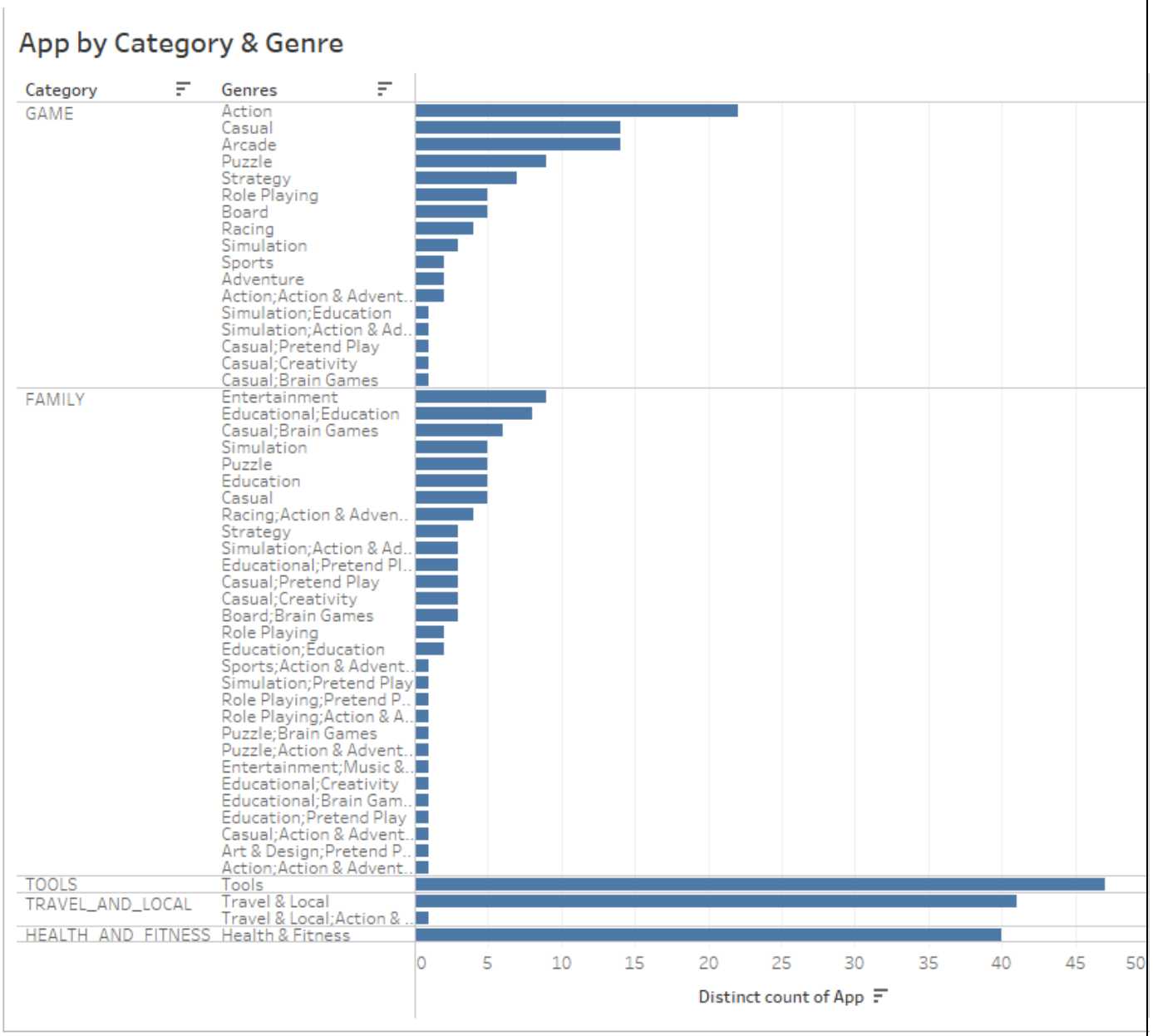


Chart 1

cha

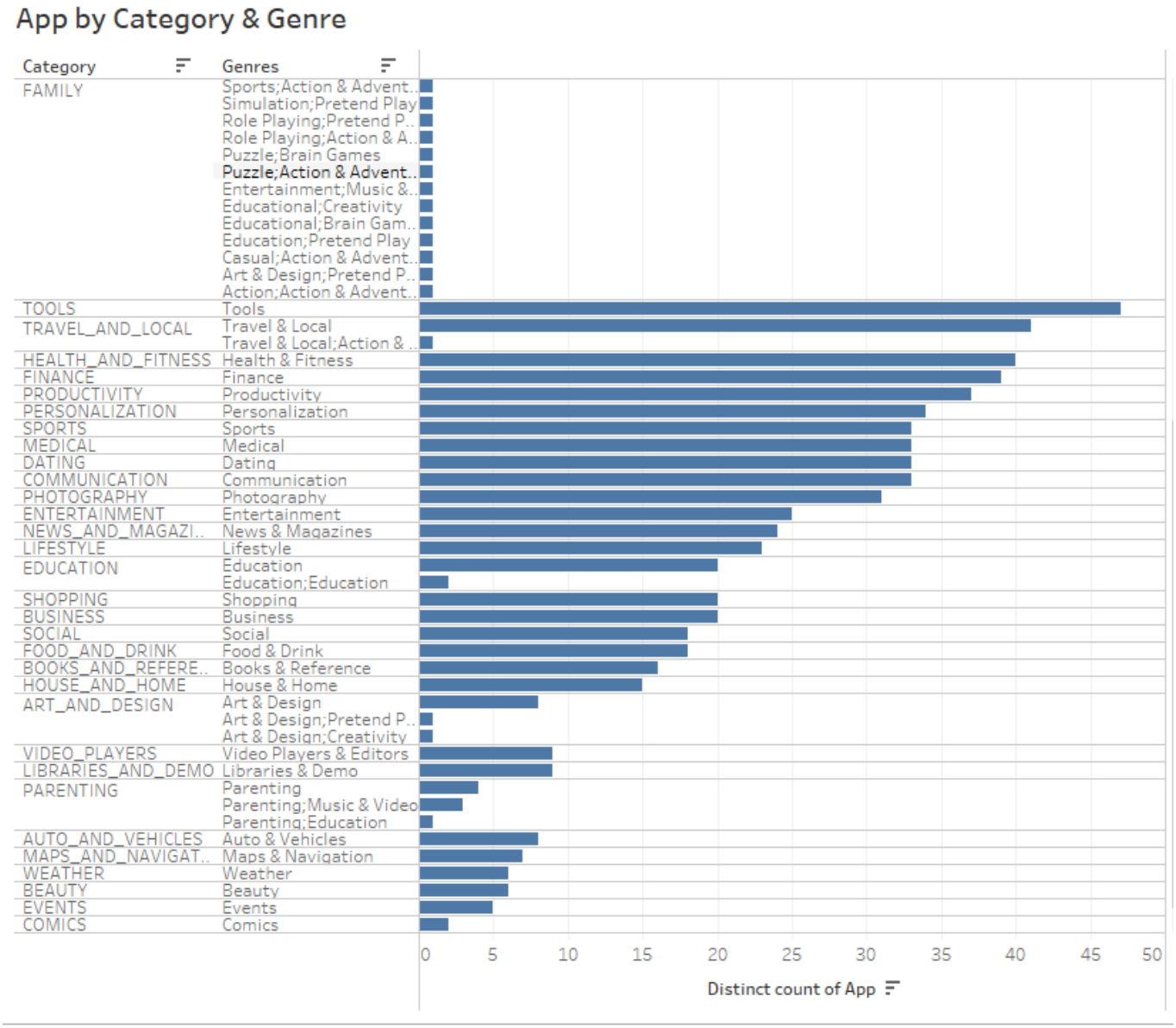


Chart 2

In Chart 1 and Chart 2, I visualize the number of apps across different categories and genres in google play store. The above charts provide insights for answering below three questions and it helped to obtain initial overview to me.

1. Identification of distinct segments within the Google Play Store.
2. Exploration of the genres associated with these segments.
3. Quantification of the number of apps within each category and genre.

Conclusion and Understanding

Some segments and genres such as TRAVEL_AND_LOCAL, HEALTH_AND_FITNESS, TOOLS and FINANCE etc. has highest number apps in play store. I have two explanations on above identification

Explanation 1

- Segments with the highest app count in the sense more user engagements, popularity and high profitability in particular segments and because of that the app developers are more favorable in those segments.

Explanation 2

- App development and deployment in google play is notably popular, because it's more accessible for any developer to easily publish their apps. The categories highlighted in the chart such as tools, travels, health and fitness apps may be functionally eased to develop and deploy, that would be another positive reason to high number of apps.

I decided to further analysis verify the above explanations. I add user rating factor over the app count charts (chart1 and chart2) to verify the actual user engagement on the apps.

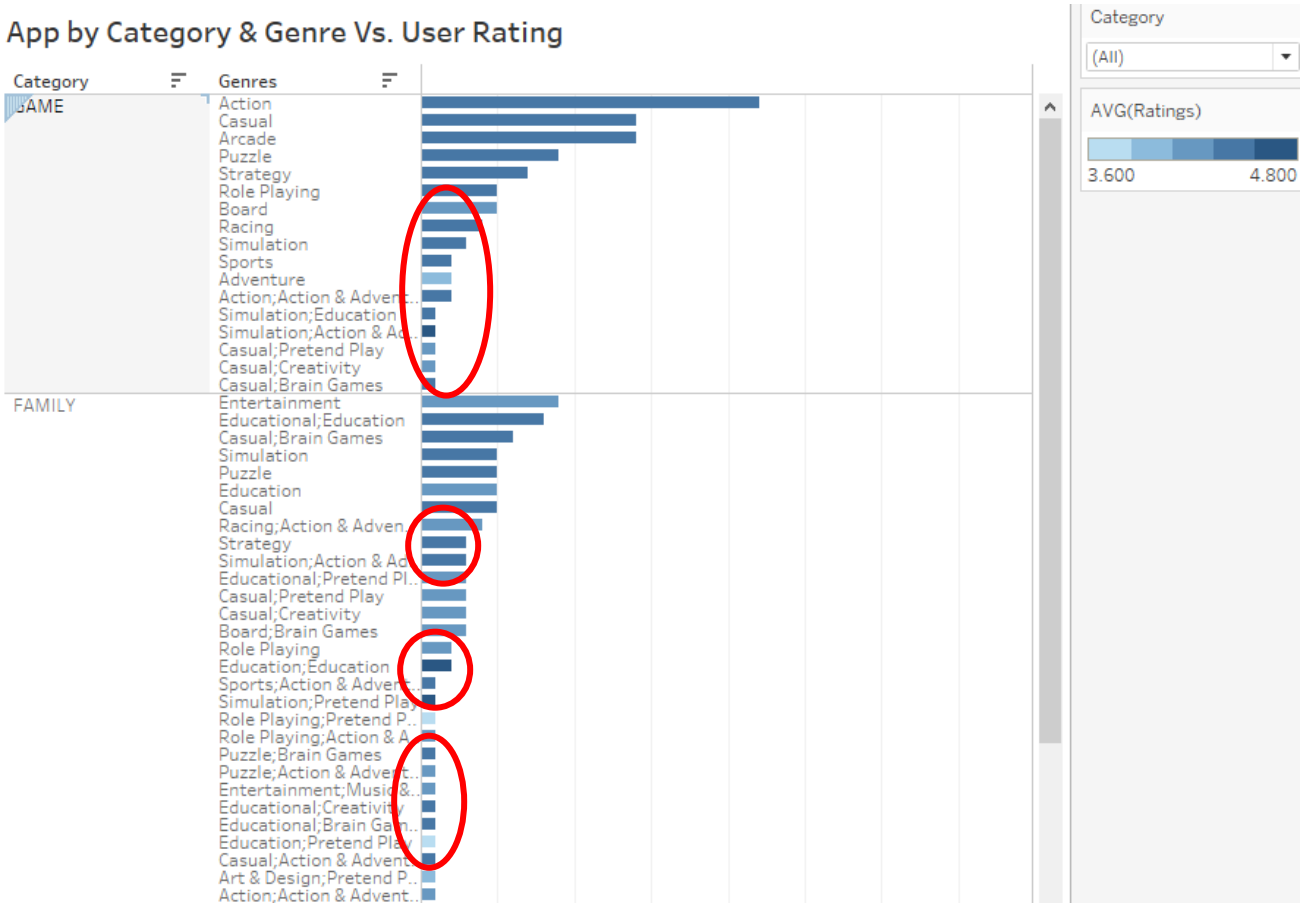


Chart 3

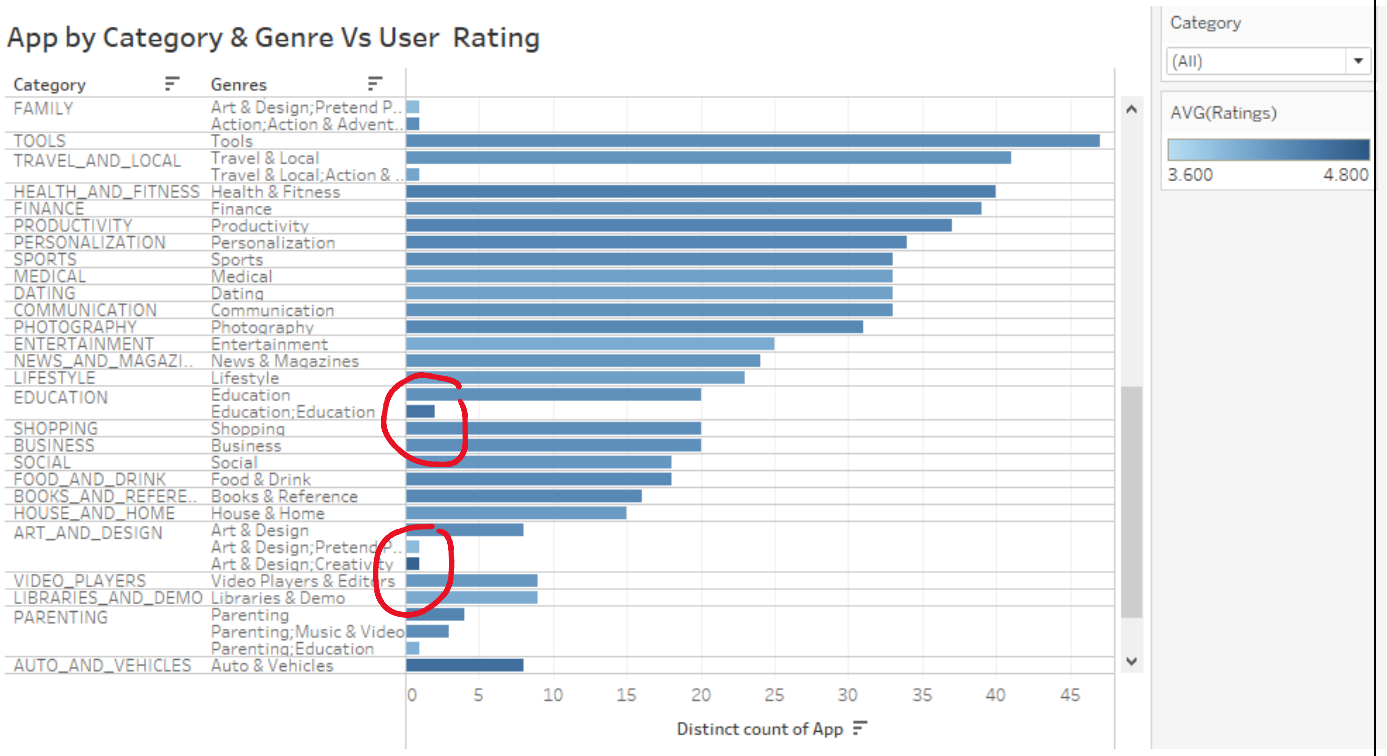


Chart 4

Explanation

Chart3 and 4 illustrate the distribution of user ratings across the application categories and genres. I used color gradient darker blue for higher ratings and when it's come to lighter with the fewer ratings. Upon close observation, I observed that categories and garners with the highest numbers of apps get highest number of reviews.

Moreover, I noticed little diversification (highlighted in red) which tells another good story for all stakeholders.

Conclusion and My reflection

Reflections of the insights chart3 and 4:

1. The first explanation is supported by the data visualization in above charts, which demonstrates higher the number of apps have more user interaction. This is a good co-relation to app developers to choose their categories more attractive and competitive landscape.
2. The second insight is interesting not only for app developers but also for opportunity hunters. The areas circled red have fewer applications, but those

apps get higher number of ratings suggest a significant untapped potential. These are the niches where competition is very less and apps with fully satisfied user requirements can gain receptive audience.

- **Business Problem: How do rating and review reflect usability and application interaction?**

- **Visualization**

1 Chart 5

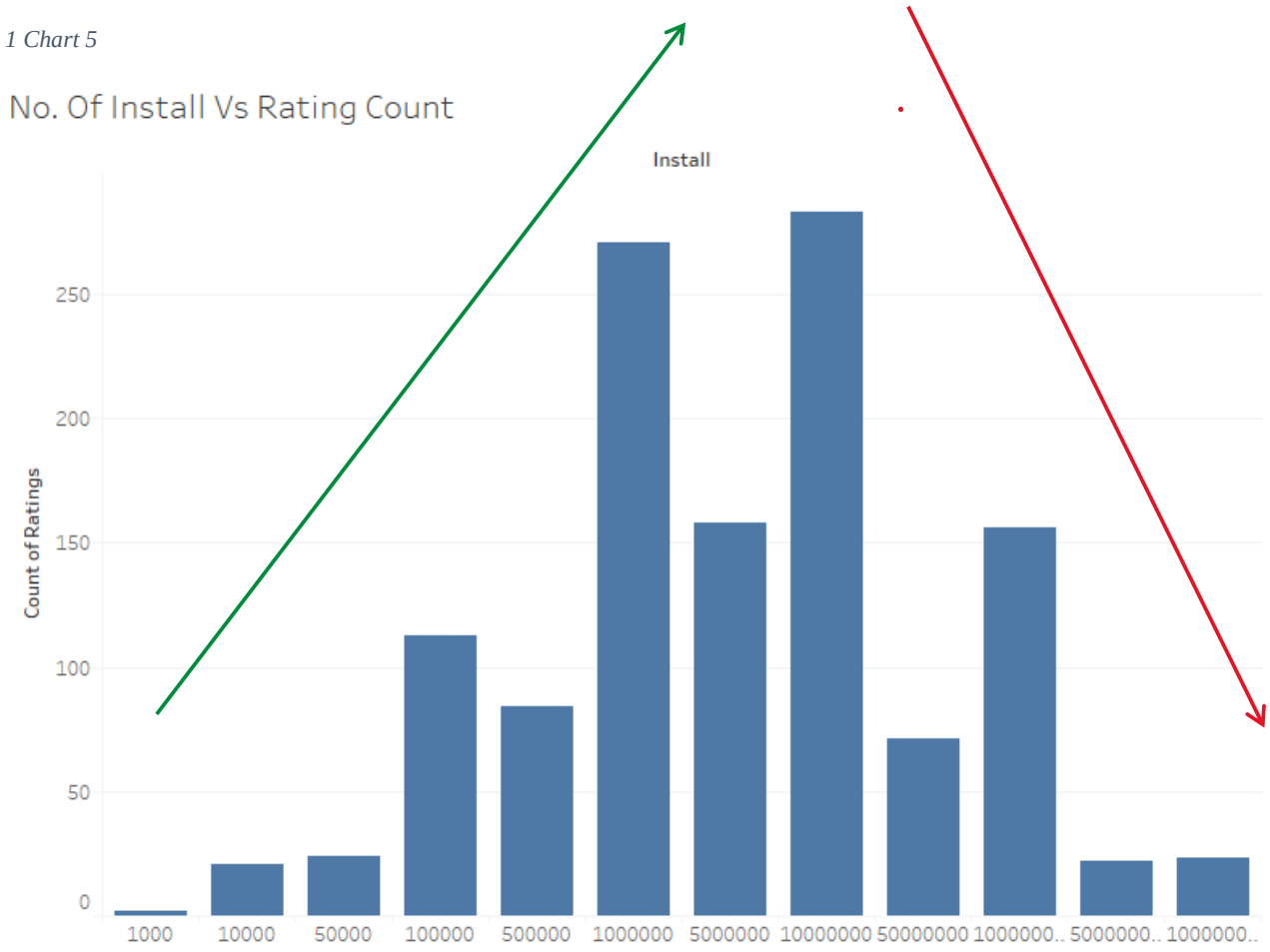
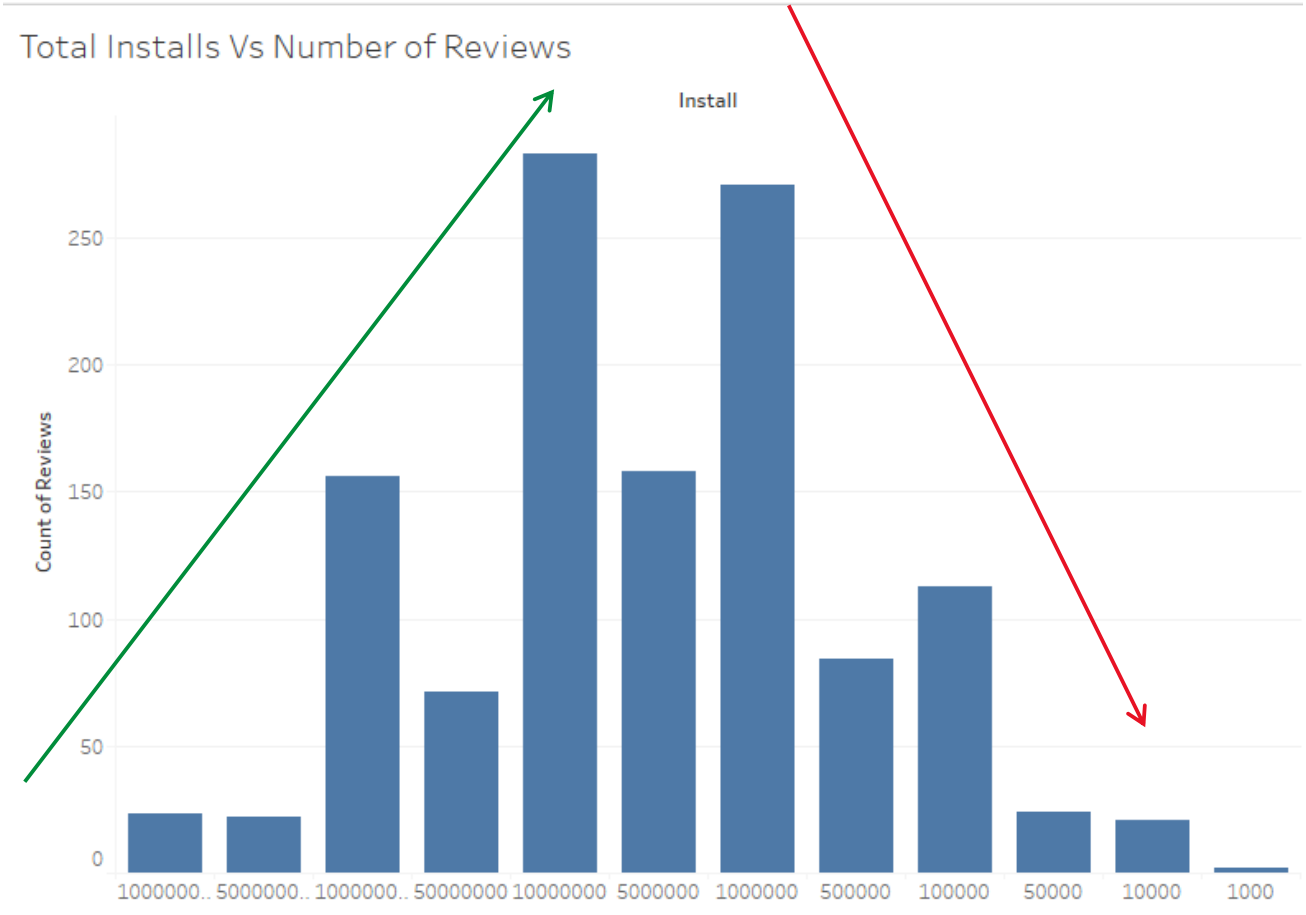


Figure 1 Chart 5



Chat 6

Explanation

1. Charts 5 and 6 visualize the correlation of reviews and ratings with the number of installs, they offer deeper insights of user engagement metrics.
2. In both visualizations have same pattern, mid-range installations(1M-10M) received highest ratings and reviews. This explains another interesting story. Apps with fewer or higher installs tend to garner fewer ratings and reviews.

Conclusion and my reflection

Explanation 1, I clearly understand that apps with installations between 1M and 10M are in sweet spot, they have a large enough user base to generate substantial feedback but are not so widespread that user feedback becomes diluted.

In further thinking I found that this user feedback dilution happens due to number of downloads between 1M and 10M higher and count the variable not fair for the analysis. To overcome the bias, I decided to move with the average.

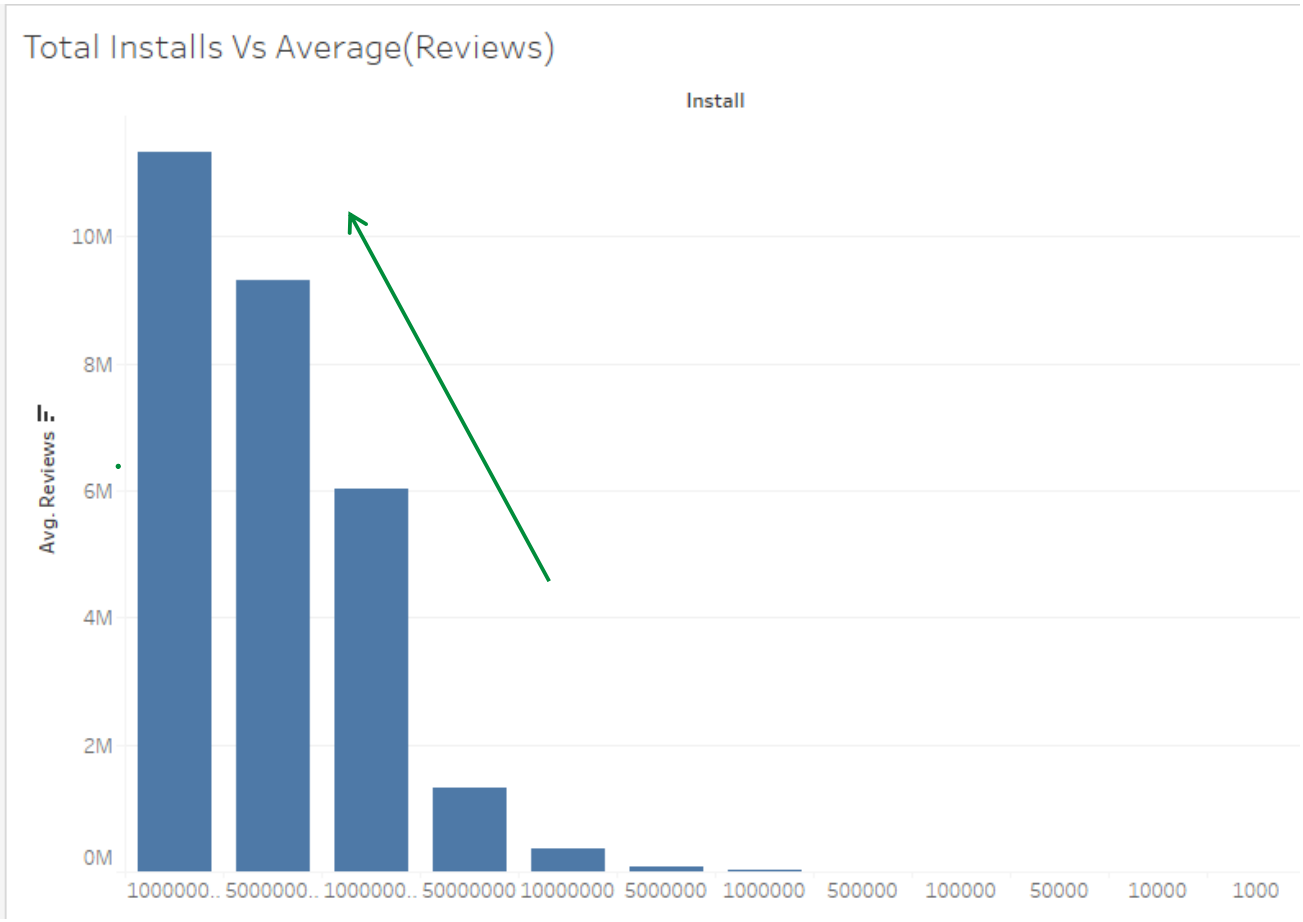


Chart 7

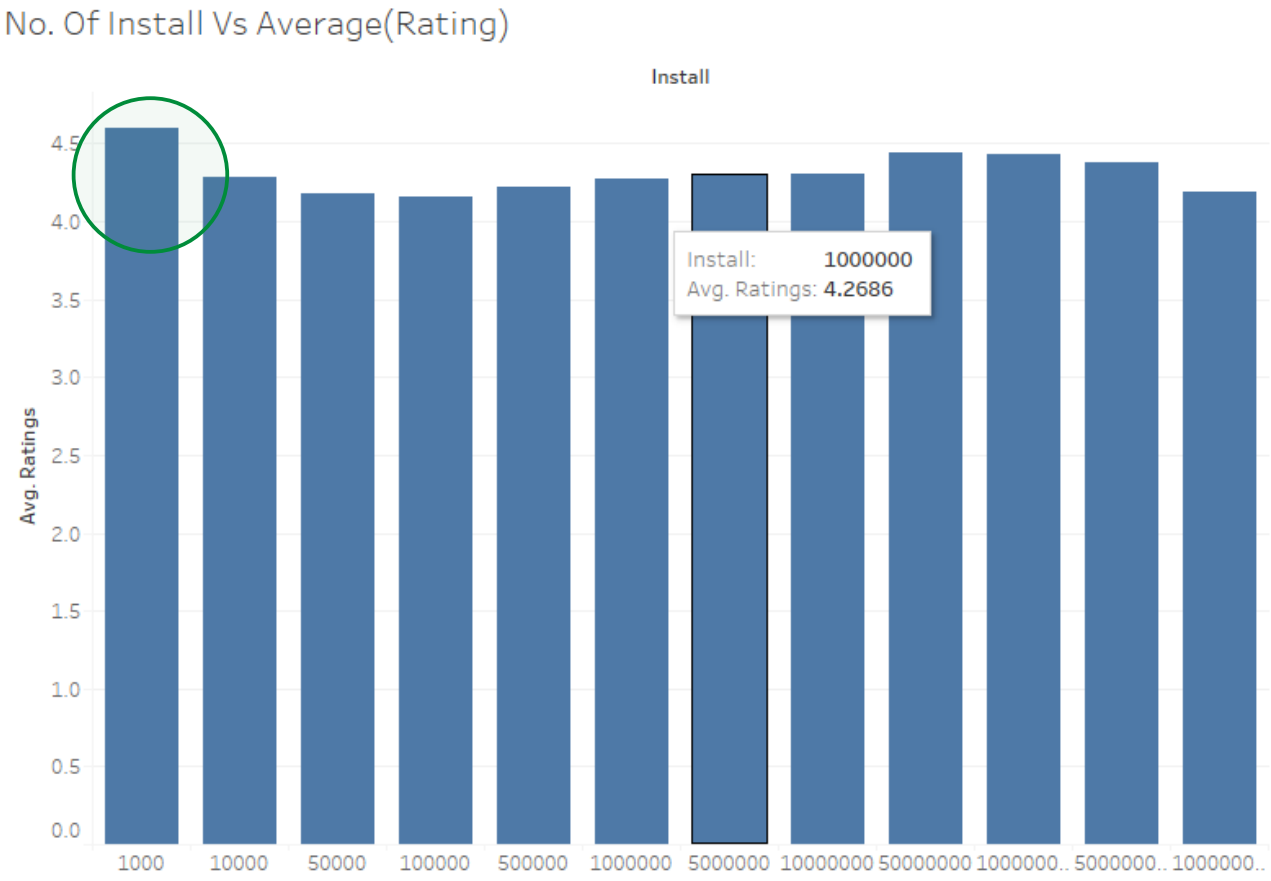


Chart 8

Explanation

Clear explanation can made by chart 7 Is higher the review average count for higher the installations. This reflects a large and engaged user base for highly installed apps, leading to significant feedback. In chart 8 apps installs between 10000 to 500M average rating remain stable around 4 and 4.3

Conclusion and my reflection

My reflection of chart 7 emphasizes the user engagement significantly higher the install apps and feedback. This level of engagement is crucial for app developers, as reviews provide valuable insights into user experience and areas for improvement. However, it also highlights the pressure on developers and product managers to continuously meet the needs of a vast audience, balancing new features with stability and performance.

I have two conclusions for chart 8.

1. Some apps with less than 1000 installs manage to get more than 4.5 reviews which are the most dominant and those have pure solid niche users.

App developers and product managers highly focused to build such a champion apps targeting dedicated segment with dedicate userbase is great.

2.My second reflection is number of installs increased, rating stabilize around 4 and 4.3. Because apps become larger and diversified with wider audience with different expectations, behaviors and enthusiastic.

- **How frequently do top rated and top reviewed applications change their implementation by updating?**

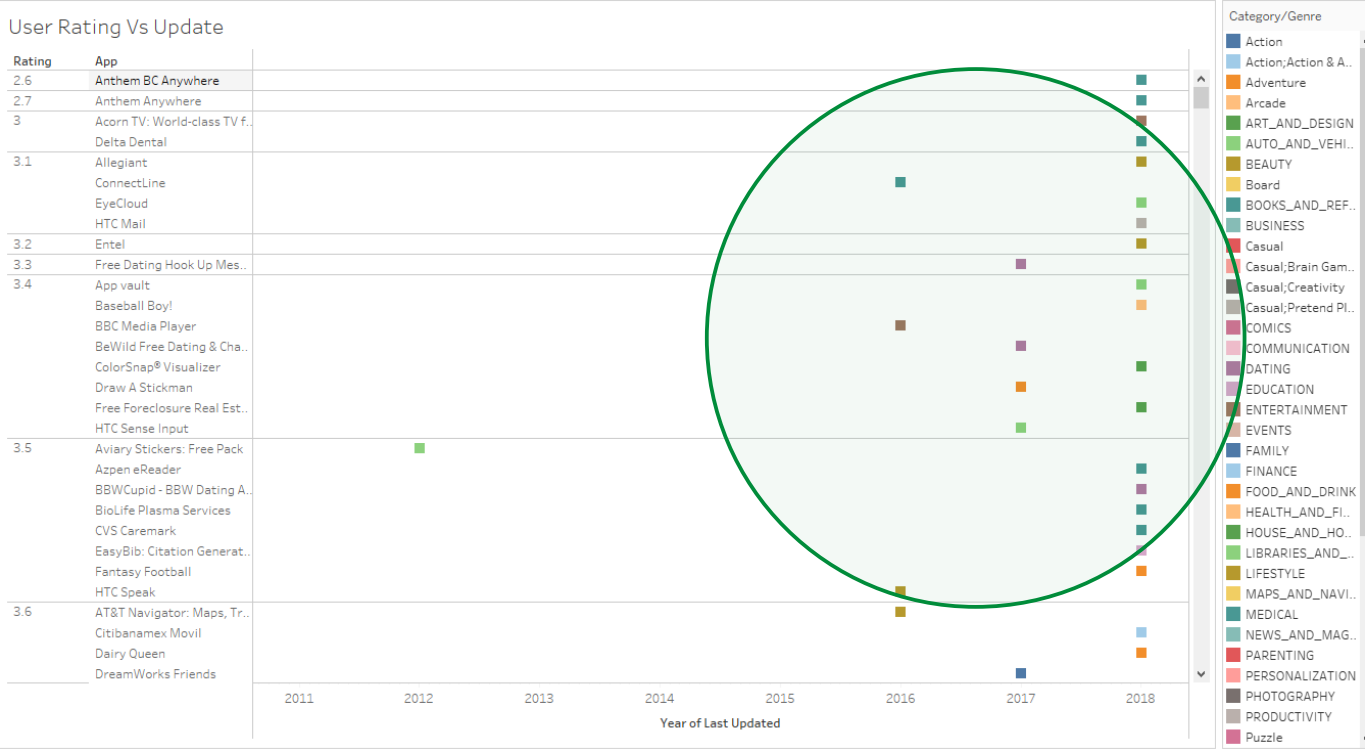
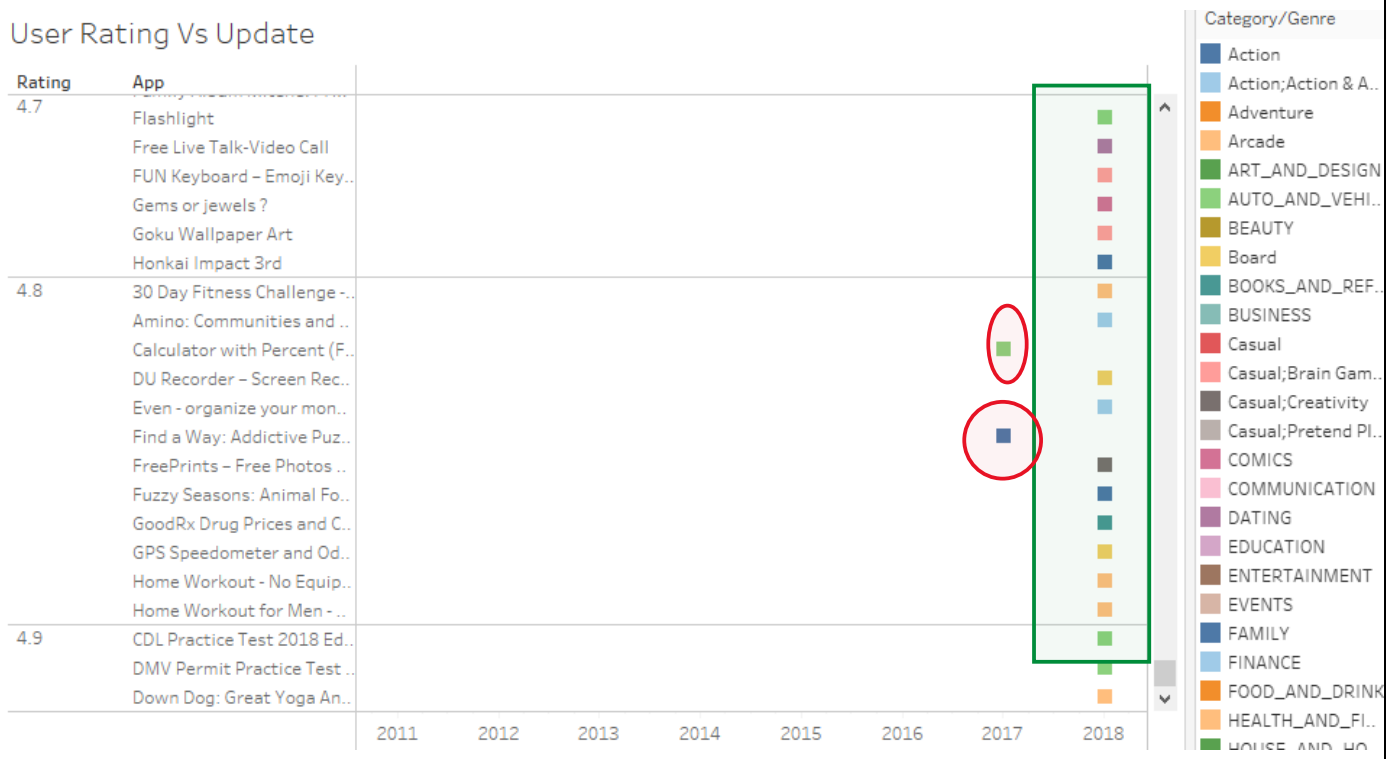


Chart 9



Chat 10

Chart 11

Explanation

Chart 9, apps with low rated have less frequency of update cycle and according to chart 10, higher rating apps they always maintain good levels of update cycle. Chart 11 proves the above evidence.

Conclusion and my reflection

The analysis indicates that there is a strong positive correlation between the frequency of updates and high user ratings. Applications that maintain a high rating tend to have recent updates, emphasizing the importance of continuous improvement and responsiveness to user feedback. Sametime high user satisfaction is observed across a different genre, indicate that maintaining quality and frequent updates can lead to success in any app category. Reflection of red circled outliers are the champion apps, but developers have opportunity to build new apps with new features can attract users more.

- **What segments are users highly interested in, influence in going to paid versions?**

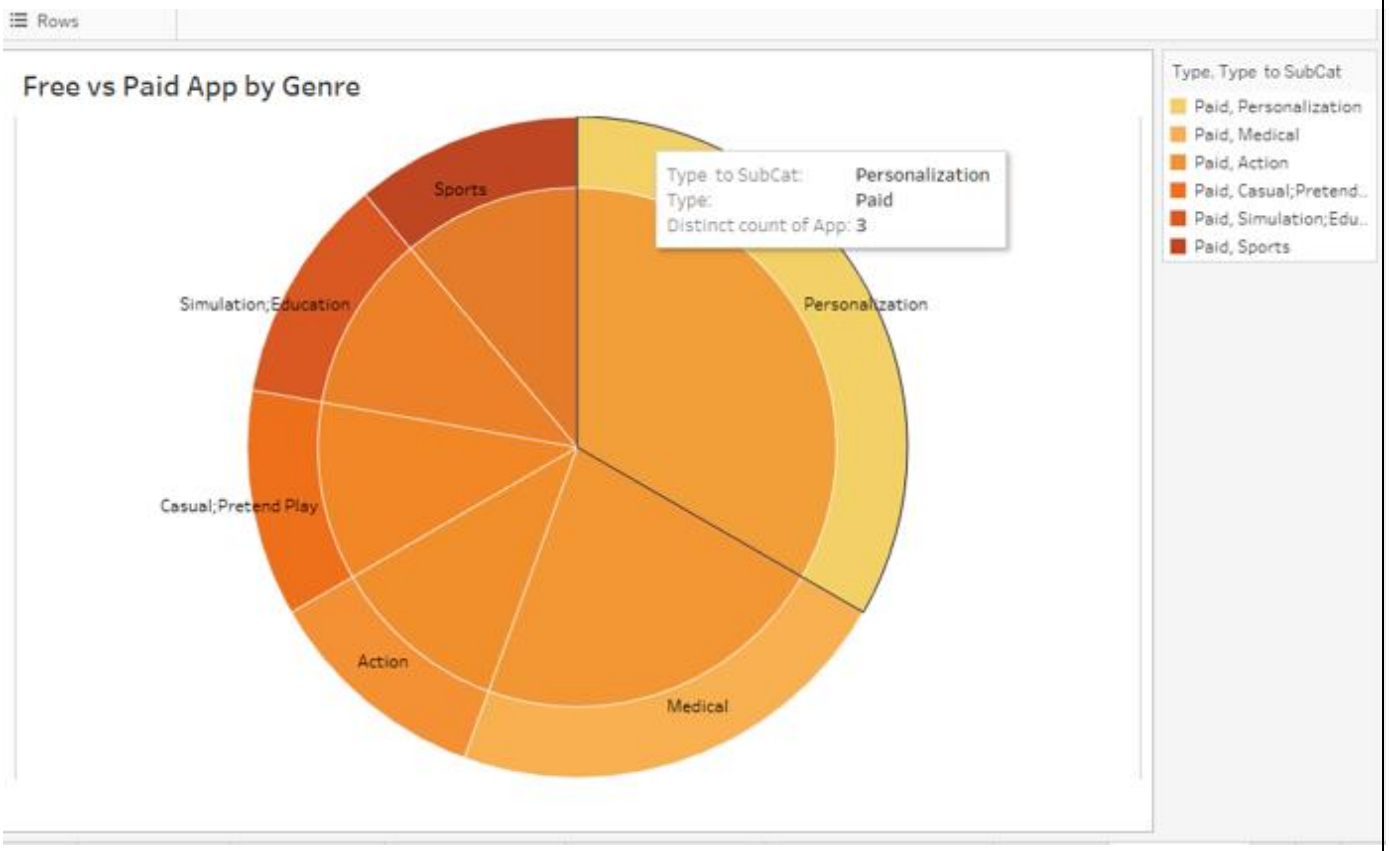


Chart 11

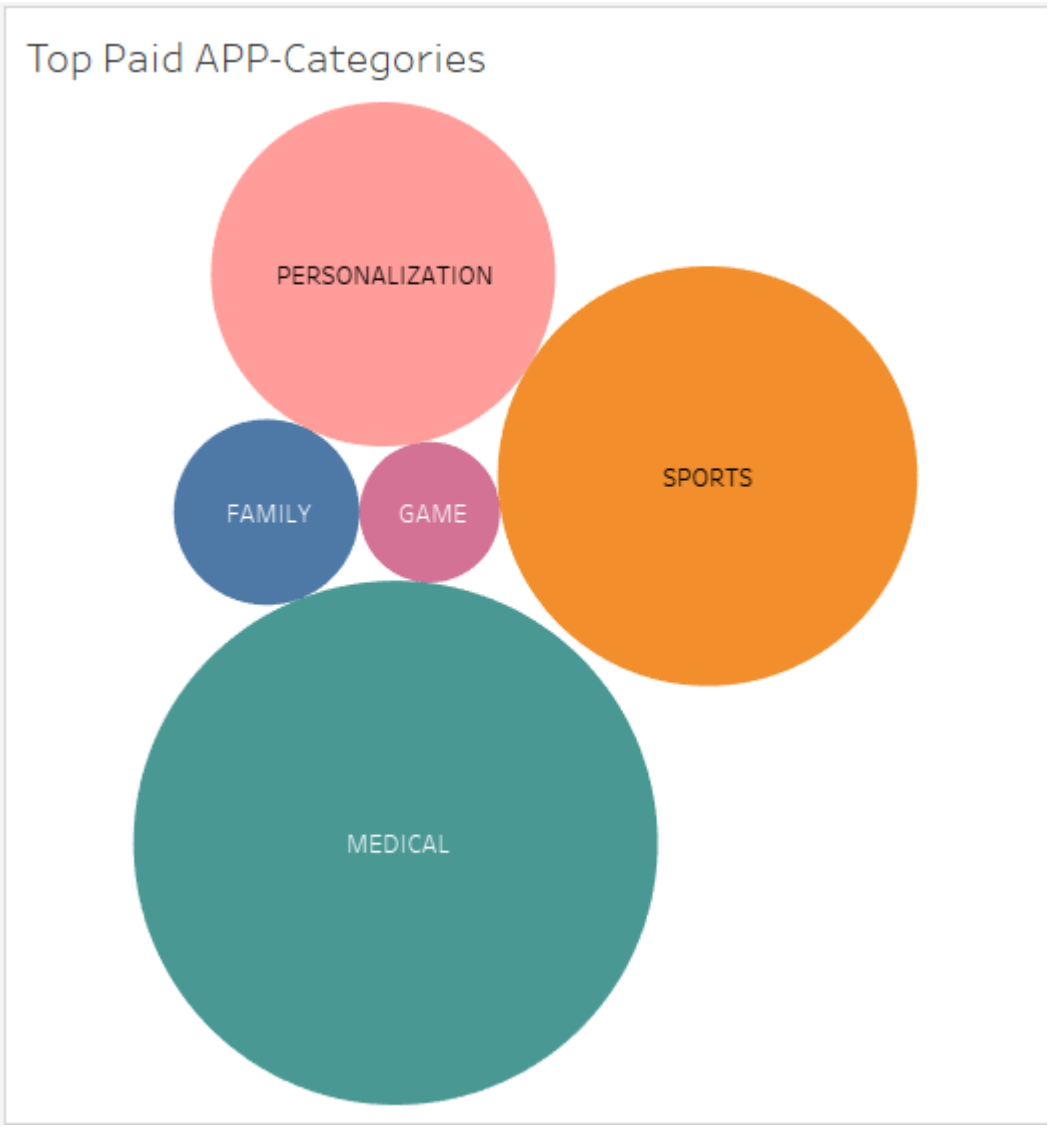

Top Paid APP-Categories
PERSONALIZATION
FAMILY
GAME
SPORTS
MEDICAL

Chart 12

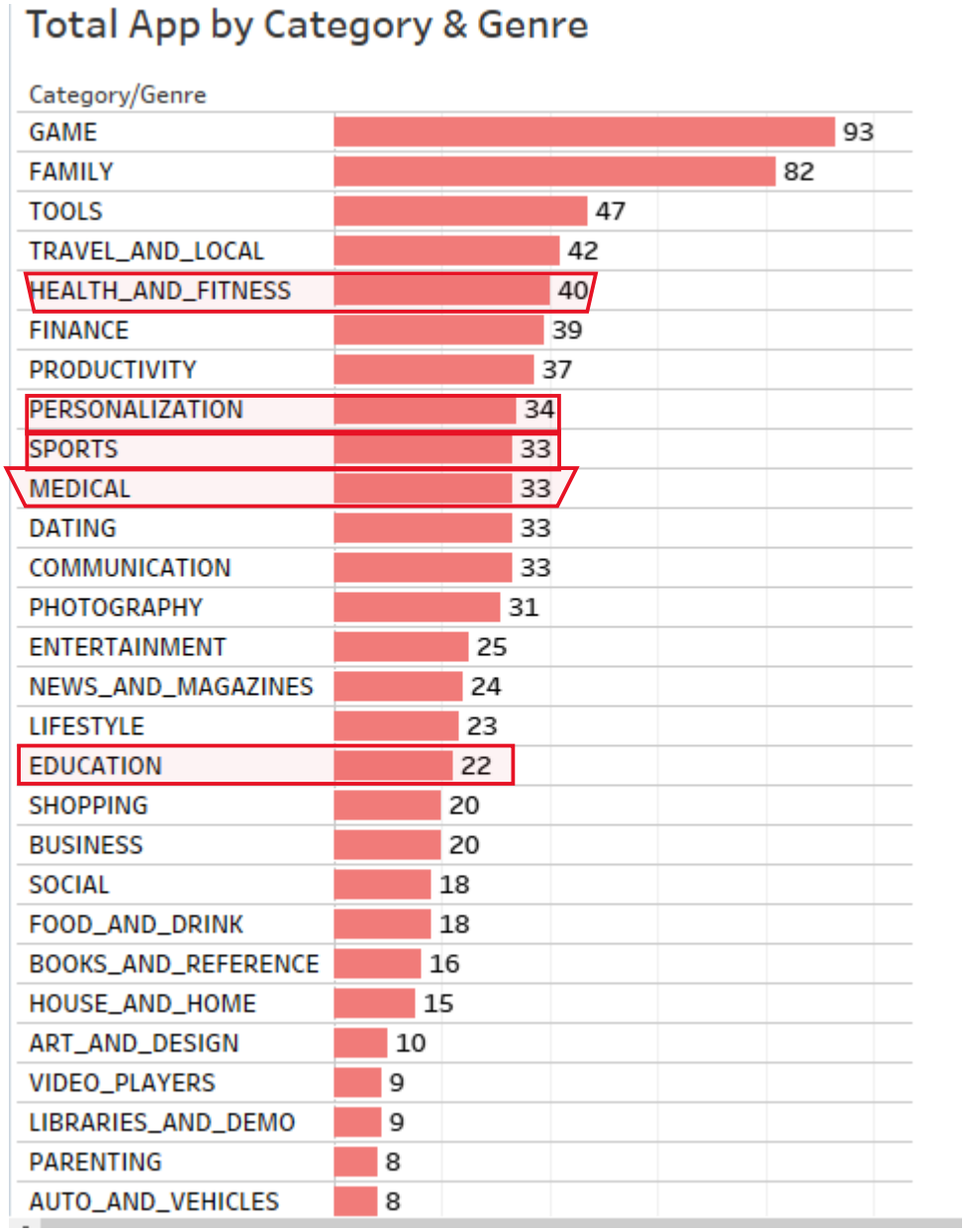

Total App by Category & Genre
Category/Genre
GAME 93
FAMILY 82
TOOLS 47
TRAVEL_AND_LOCAL 42
HEALTH_AND_FITNESS 40
FINANCE 39
PRODUCTIVITY 37
PERSONALIZATION 34
SPORTS 33
MEDICAL 33
DATING 33
COMMUNICATION 33
PHOTOGRAPHY 31
ENTERTAINMENT 25
NEWS_AND_MAGAZINES 24
LIFESTYLE 23
EDUCATION 22
SHOPPING 20
BUSINESS 20
SOCIAL 18
FOOD_AND_DRINK 18
BOOKS_AND_REFERENCE 16
HOUSE_AND_HOME 15
ART_AND_DESIGN 10
VIDEO_PLAYERS 9
LIBRARIES_AND_DEMO 9
PARENTING 8
AUTO_AND_VEHICLES 8

Chart 13

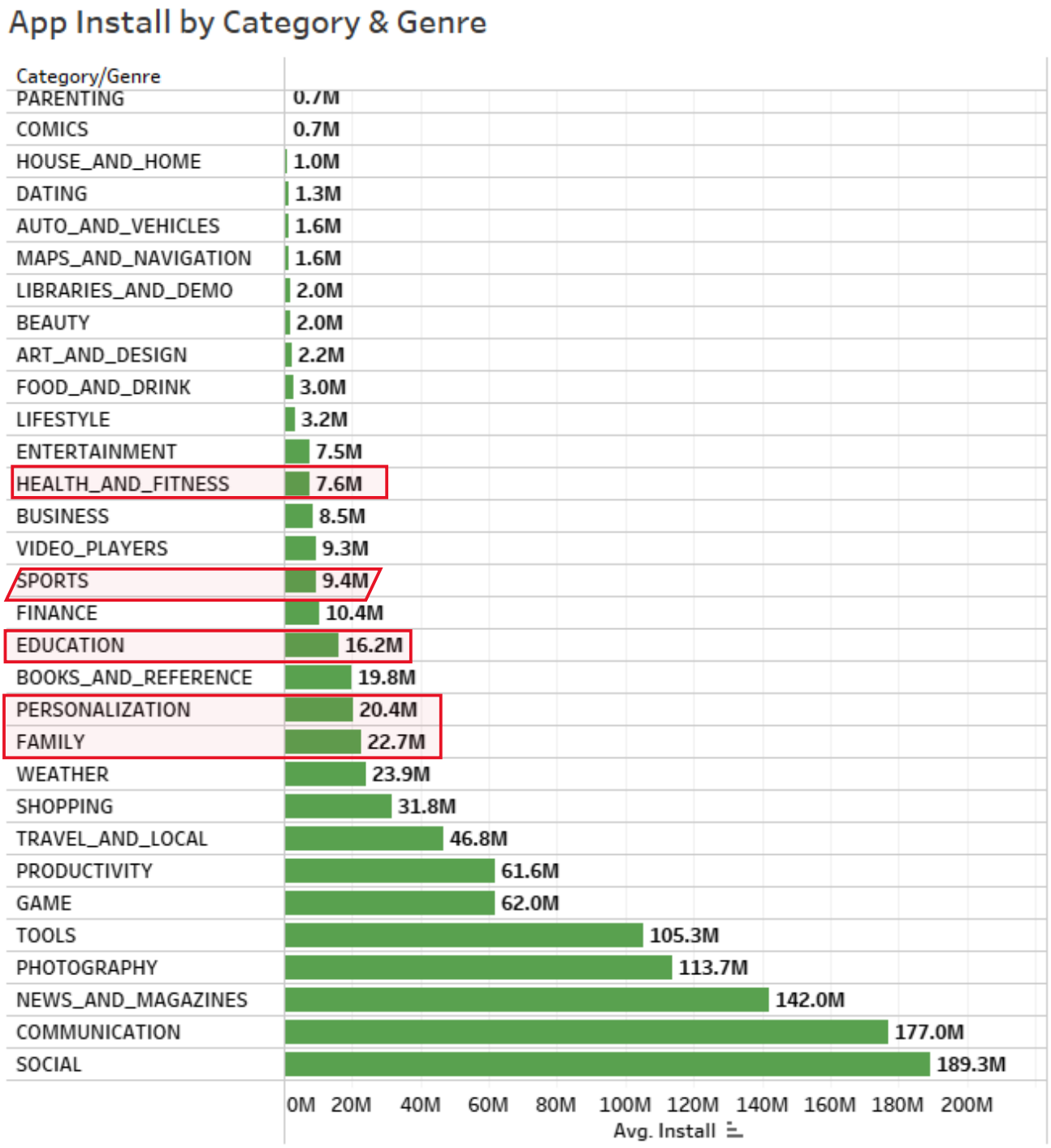

App Install by Category & Genre
Category/Genre
PARENTING 0.7M
COMICS 0.7M
HOUSE_AND_HOME 1.0M
DATING 1.3M
AUTO_AND_VEHICLES 1.6M
MAPS_AND_NAVIGATION 1.6M
LIBRARIES_AND_DEMO 2.0M
BEAUTY 2.0M
ART_AND_DESIGN 2.2M
FOOD_AND_DRINK 3.0M
LIFESTYLE 3.2M
ENTERTAINMENT 7.5M
HEALTH_AND_FITNESS 7.6M
BUSINESS 8.5M
VIDEO_PLAYERS 9.3M
SPORTS 9.4M
FINANCE 10.4M
EDUCATION 16.2M
BOOKS_AND_REFERENCE 19.8M
PERSONALIZATION 20.4M
FAMILY 22.7M
WEATHER 23.9M
SHOPPING 31.8M
TRAVEL_AND_LOCAL 46.8M
PRODUCTIVITY 61.6M
GAME 62.0M
TOOLS 105.3M
PHOTOGRAPHY 113.7M
NEWS_AND_MAGAZINES 142.0M
COMMUNICATION 177.0M
SOCIAL 189.3M
0M 20M 40M 60M 80M 100M 120M 140M 160M 180M 200M
Avg. Install

Chart 14

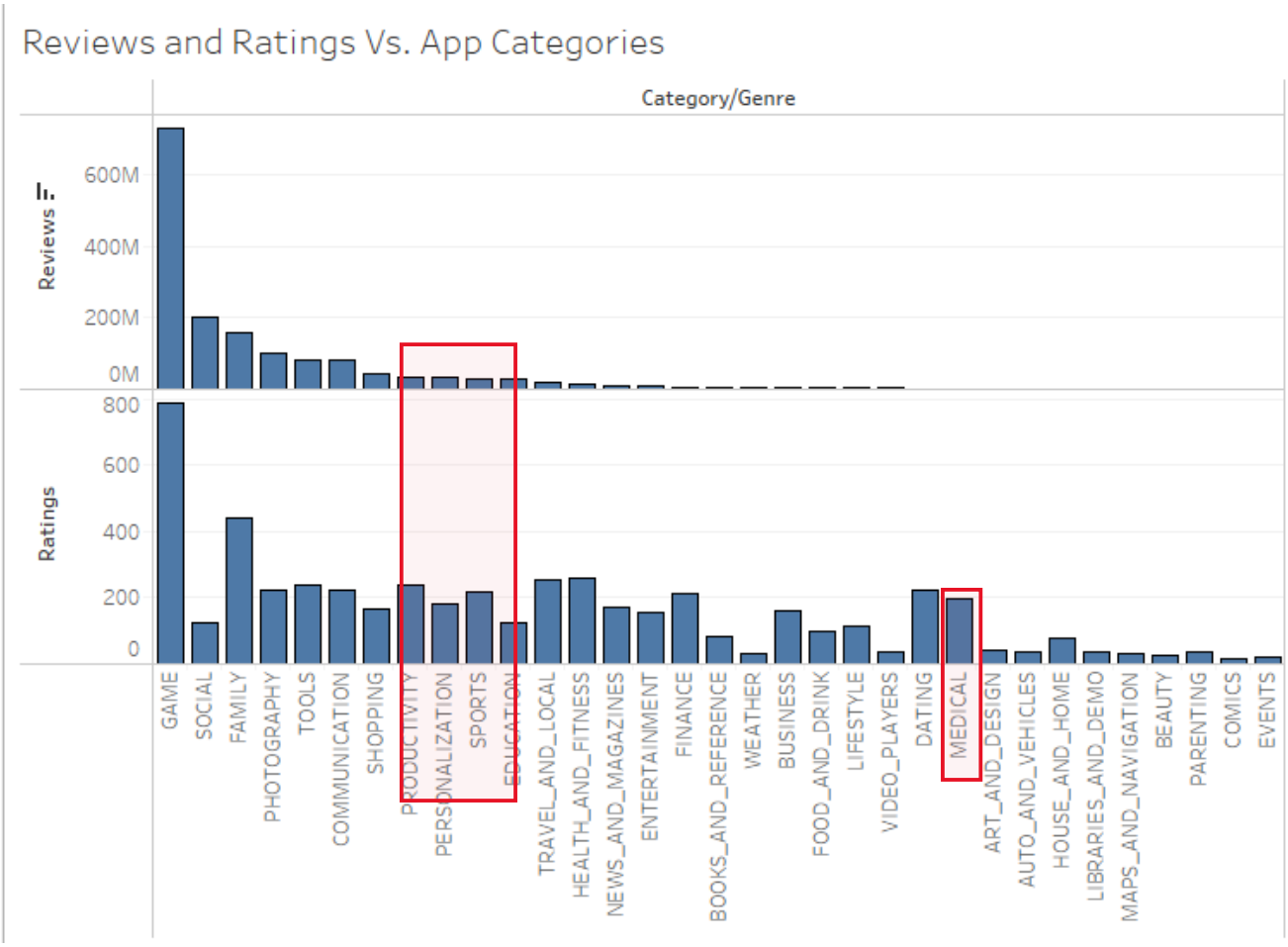


Chart 15

App segments with high potential to growth, highlighted in RED

Explanation

Charts 11 and 12, visualization describes the paid apps and belonging categories. Medical, Personalization, Games and Sports are at the top of the list. Chart 12, 13 illustrate the app integration and chart 14,15 illustrate the user engagement.

Conclusion and My Reflection

1. Medical, Sports, Family and sports categories have more potential for monetization. But according to chart 13 and 14 those categories have low user engagement. Specially categories like medical there have number of apps but with less user engagements. Developers and product managers have better

opportunities to understand market gap and full fill with better business advantage.

2. My next reflection is that categories with high potential to monetize have a smaller number of applications. For Example, customers are willing to pay categories like Education (chart 11) but very a smaller number of effective and positive user engagement on it.

Chart 13 and chart 14 evidence proves that segments like Education, Sport, Personalization, Medical has more opportunities to monetization and make profits.

4th Milestone – Stories beyond the Data and Future[19][20]

After analyzing all the visualization, I could build interesting stories for both app developer prospective and project manager prospective.

The first interesting thing I noticed, apps backed by tech giant with high impression among users dominant general and well popular segments. (chart 16).

When evaluate deeper details, I found that certain segments like Education, Spors, Art, Design and Auto. Etc. has high user ratings yet has surprisingly low app counts. That indicates beautiful story High satisfaction but low competition. (chart 18 and 19). Another interesting finding I observed from chart 11 and 12, was most of them are highly paid apps. So, this is a great opportunity for any developer or project manager to touch in.

While analyzing chart16, the chart showed that apps backed by big tech companies typically lead in the number of installs, thanks to their brand names, existing popularity and unlimited spends on promotions.

So, some developers entering the market are very scared because this giant fear

But in chart 17, I found that highly rated apps tend to focus on very specific niches or segments, excelling in areas where they directly address the needs and preferences of their targeted audience. Most of these apps come with less investment and lack of ground advantage. Unlike the giants that aim for broad appeal, these apps find strength in specialization, dominating their chosen fields through precision and dedication. It suggests that success may not necessarily require competing against the big names on their terms. Instead, there's considerable opportunity in identifying and serving a niche market exceptionally well.

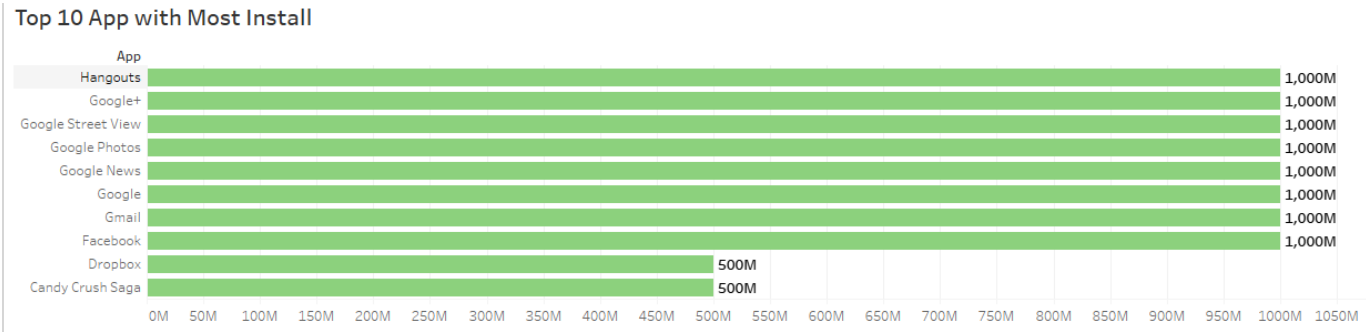


Chart 16

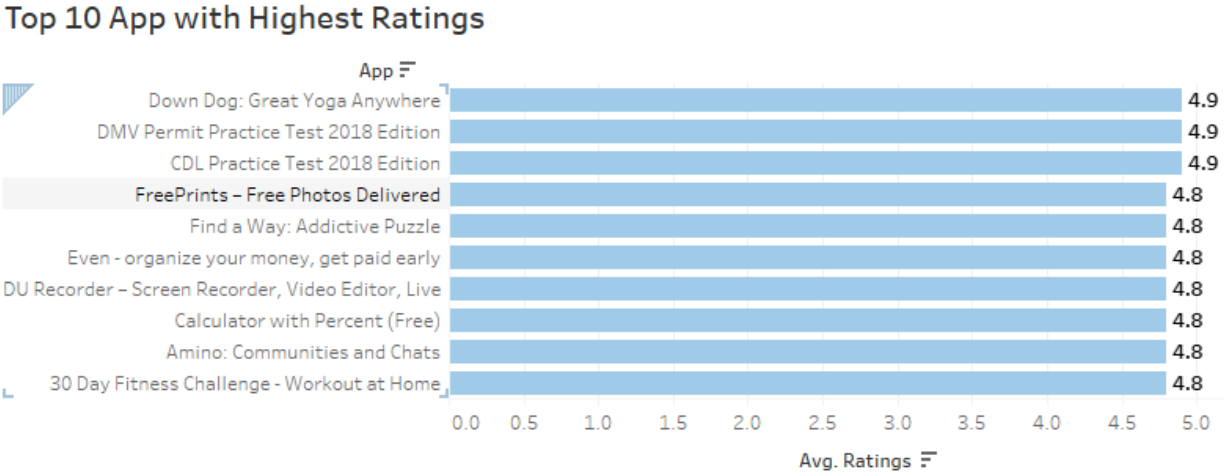


Chart 17

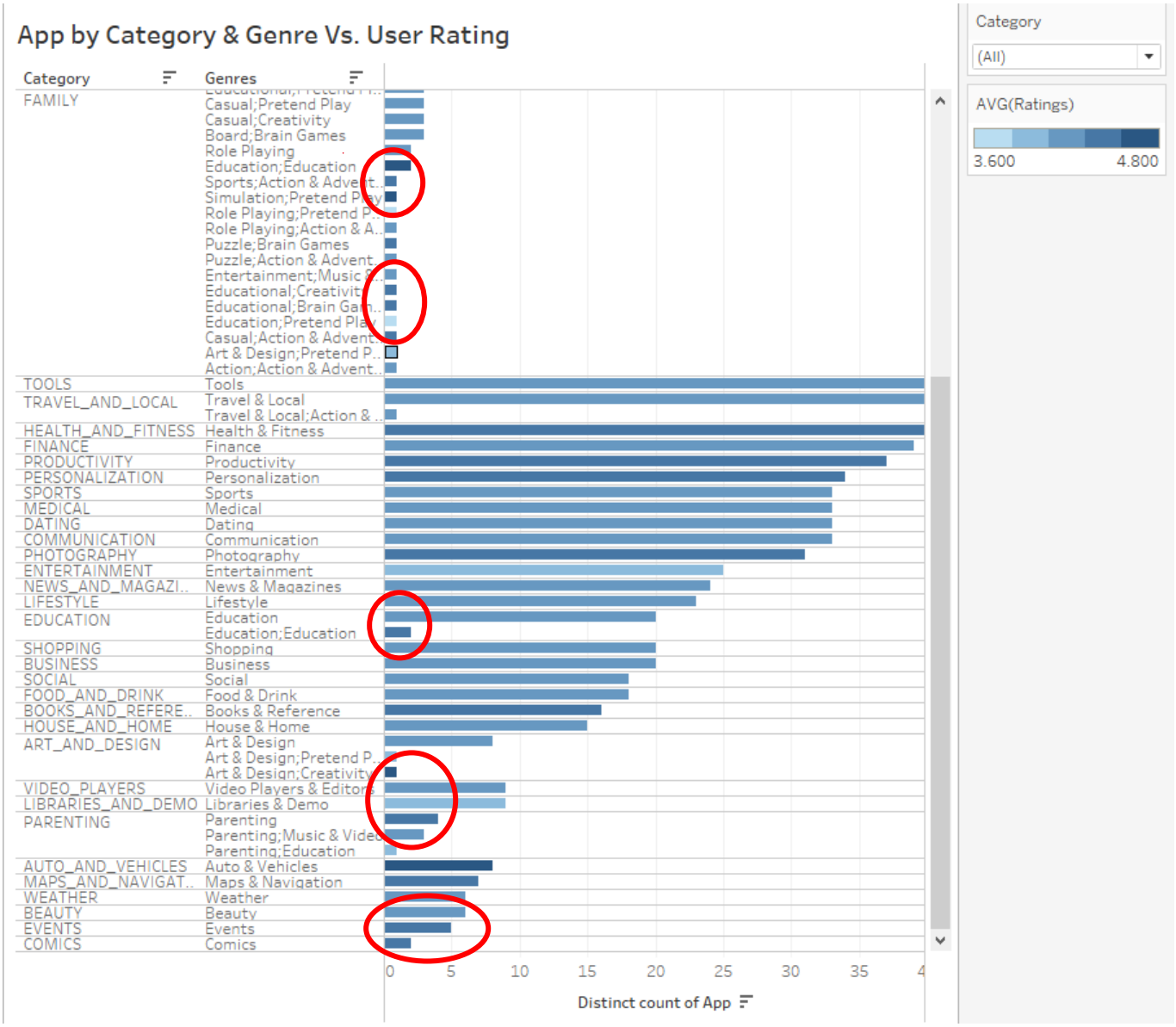

App by Category & Genre Vs. User Rating
Category
Genres
FAMILY
Casual;Pretend Play
Casual;Creativity
Board;Brain Games
Role Playing
Education;Education
Puzzle;Brain Games
TOOLS
Tools
TRAVEL_AND_LOCAL
Travel & Local
HEALTH_AND_FITNESS
Health & Fitness
FINANCE
Finance
PRODUCTIVITY
Productivity
PERSONALIZATION
Personalization
SPORTS
Sports
MEDICAL
Medical
DATING
Dating
COMMUNICATION
Communication
PHOTOGRAPHY
Photography
ENTERTAINMENT
Entertainment
NEWS_AND_MAGAZI..
News & Magazines
LIFESTYLE
Lifestyle
EDUCATION
Education
Education;Education
SHOPPING
Shopping
BUSINESS
Business
SOCIAL
Social
FOOD_AND_DRINK
Food & Drink
BOOKS_AND_REFERE..
Books & Reference
HOUSE_AND_HOME
House & Home
ART_AND_DESIGN
Art & Design
VIDEO_PLAYERS
LIBRARIES_AND_DEMO
Libraries & Demo
PARENTING
Parenting
AUTO_AND_VEHICLES
Auto & Vehicles
MAPS_AND_NAVIGAT..
Maps & Navigation
WEATHER
Weather
BEAUTY
Beauty
EVENTS
Events
COMICS
Comics
0
5
10
15
20
25
30
35
Distinct count of App
Category
(All)
AVG(Ratings)
3.600
4.800

Chart 18

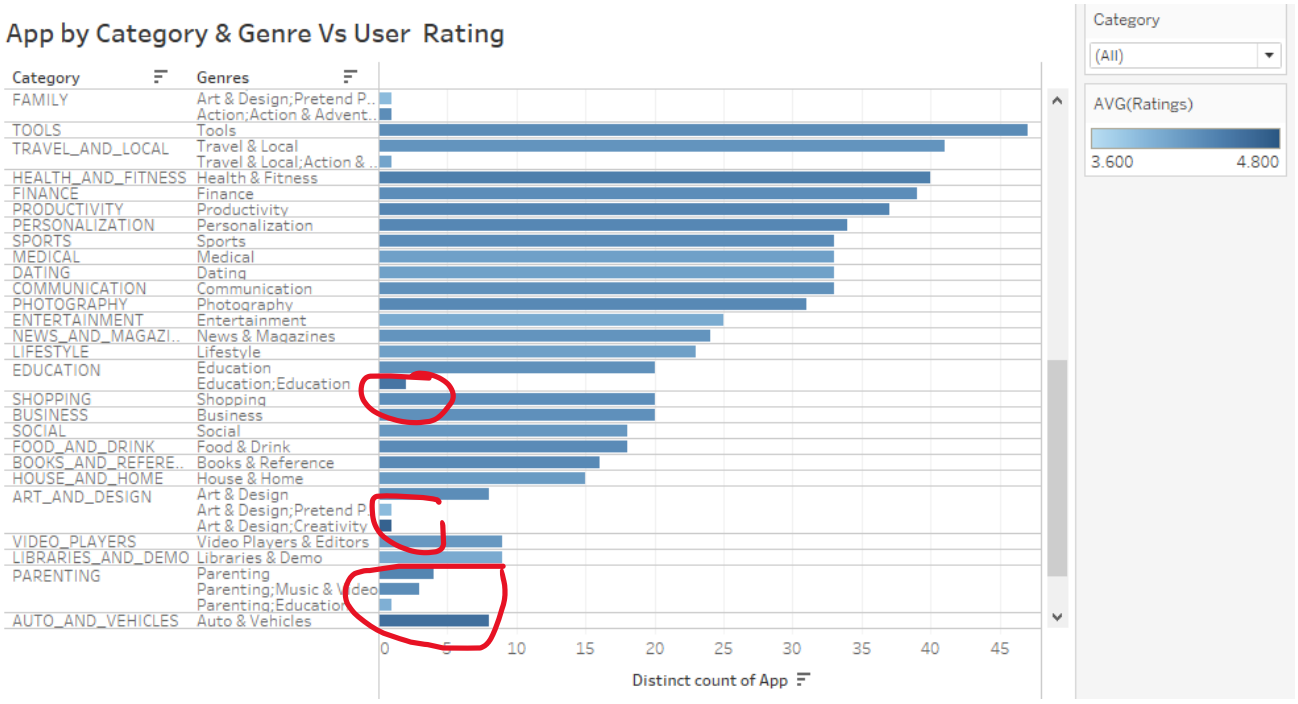

App by Category & Genre Vs User Rating
Category
Genres
FAMILY
Art & Design;Pretend P..
Action;Action & Advent..
TOOLS
Tools
TRAVEL_AND_LOCAL
Travel & Local
Travel & Local;Action & ..
HEALTH_AND_FITNESS
Health & Fitness
FINANCE
Finance
PRODUCTIVITY
Productivity
PERSONALIZATION
Personalization
SPORTS
Sports
MEDICAL
Medical
DATING
Dating
COMMUNICATION
Communication
PHOTOGRAPHY
Photography
ENTERTAINMENT
Entertainment
NEWS_AND_MAGAZI..
News & Magazines
LIFESTYLE
Lifestyle
EDUCATION
Education
Education;Education
SHOPPING
Shopping
BUSINESS
Business
SOCIAL
Social
FOOD_AND_DRINK
Food & Drink
BOOKS_AND_REFERE..
Books & Reference
HOUSE_AND_HOME
House & Home
ART_AND_DESIGN
Art & Design
Art & Design;Pretend P..
Art & Design;Creativity
VIDEO_PLAYERS
Video Players & Editors
LIBRARIES_AND_DEMO
Libraries & Demo
PARENTING
Parenting
Parenting;Music & Video
Parenting;Educatio..
AUTO_AND_VEHICLES
Auto & Vehicles
0
5
10
15
20
25
30
35
40
45
Distinct count of App
Category
(All)
AVG(Ratings)
3.600
4.800

Chart 19

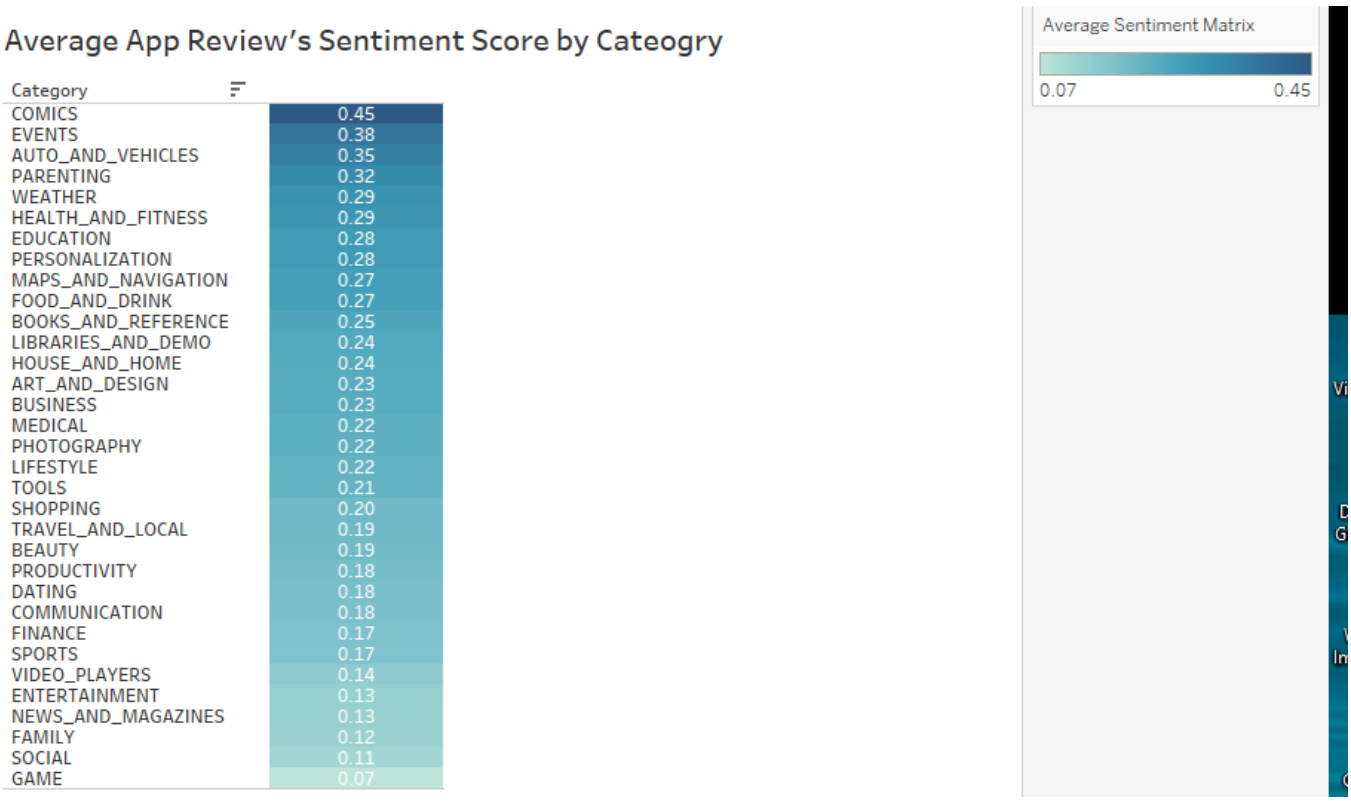

Average App Review's Sentiment Score by Cateogry
Category
COMICS 0.45
EVENTS 0.38
AUTO_AND_VEHICLES 0.35
PARENTING 0.32
WEATHER 0.29
HEALTH_AND_FITNESS 0.29
EDUCATION 0.28
PERSONALIZATION 0.28
MAPS_AND_NAVIGATION 0.27
FOOD_AND_DRINK 0.27
BOOKS_AND_REFERENCE 0.25
LIBRARIES_AND_DEMO 0.24
HOUSE_AND_HOME 0.24
ART_AND_DESIGN 0.23
BUSINESS 0.23
MEDICAL 0.22
PHOTOGRAPHY 0.22
LIFESTYLE 0.22
TOOLS 0.21
SHOPPING 0.20
TRAVEL_AND_LOCAL 0.19
BEAUTY 0.19
PRODUCTIVITY 0.18
DATING 0.18
COMMUNICATION 0.18
FINANCE 0.17
SPORTS 0.17
VIDEO_PLAYERS 0.14
ENTERTAINMENT 0.13
NEWS_AND_MAGAZINES 0.13
FAMILY 0.12
SOCIAL 0.11
GAME 0.07
Average Sentiment Matrix
0.07
0.45

Chart20

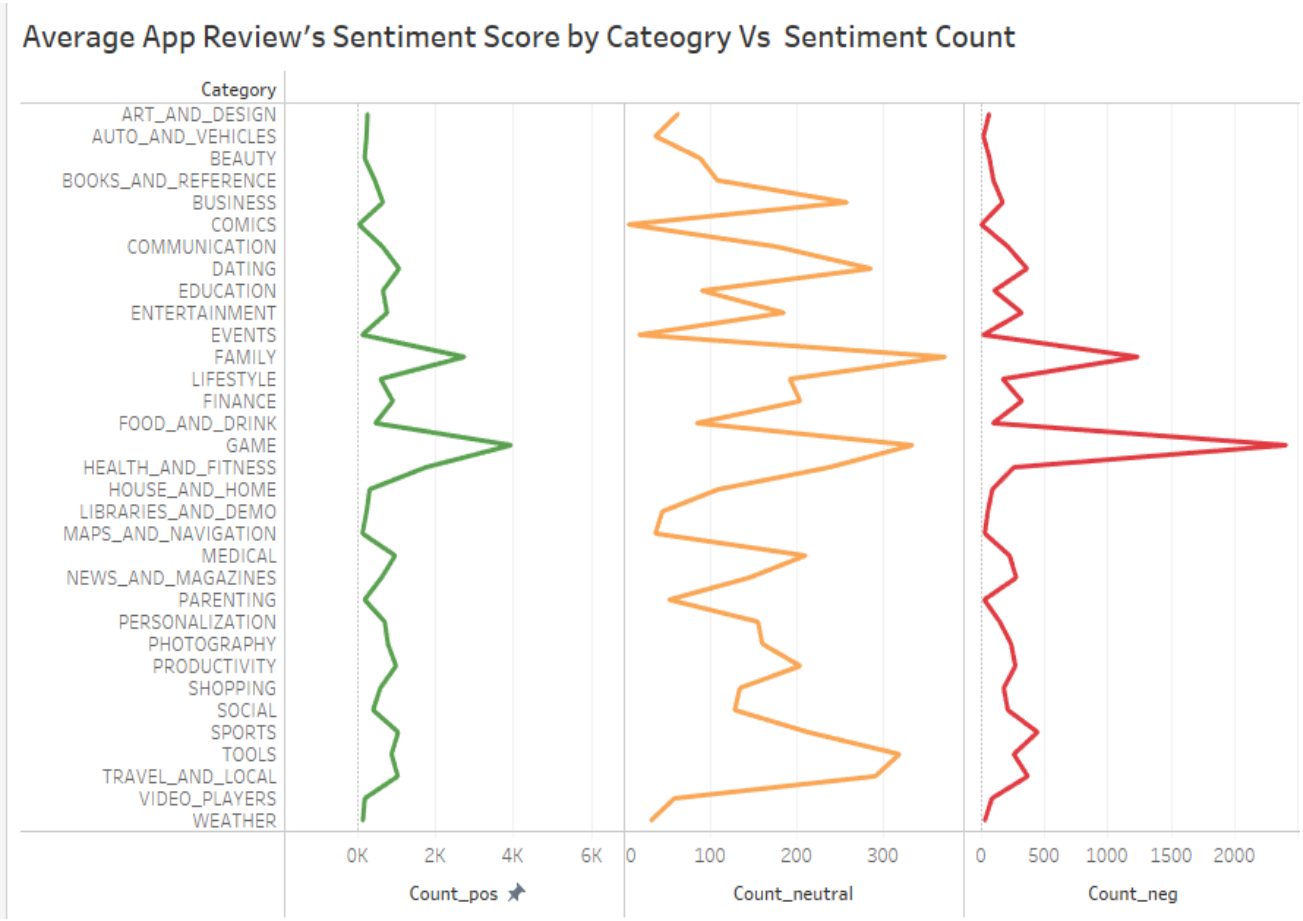


Chart 21

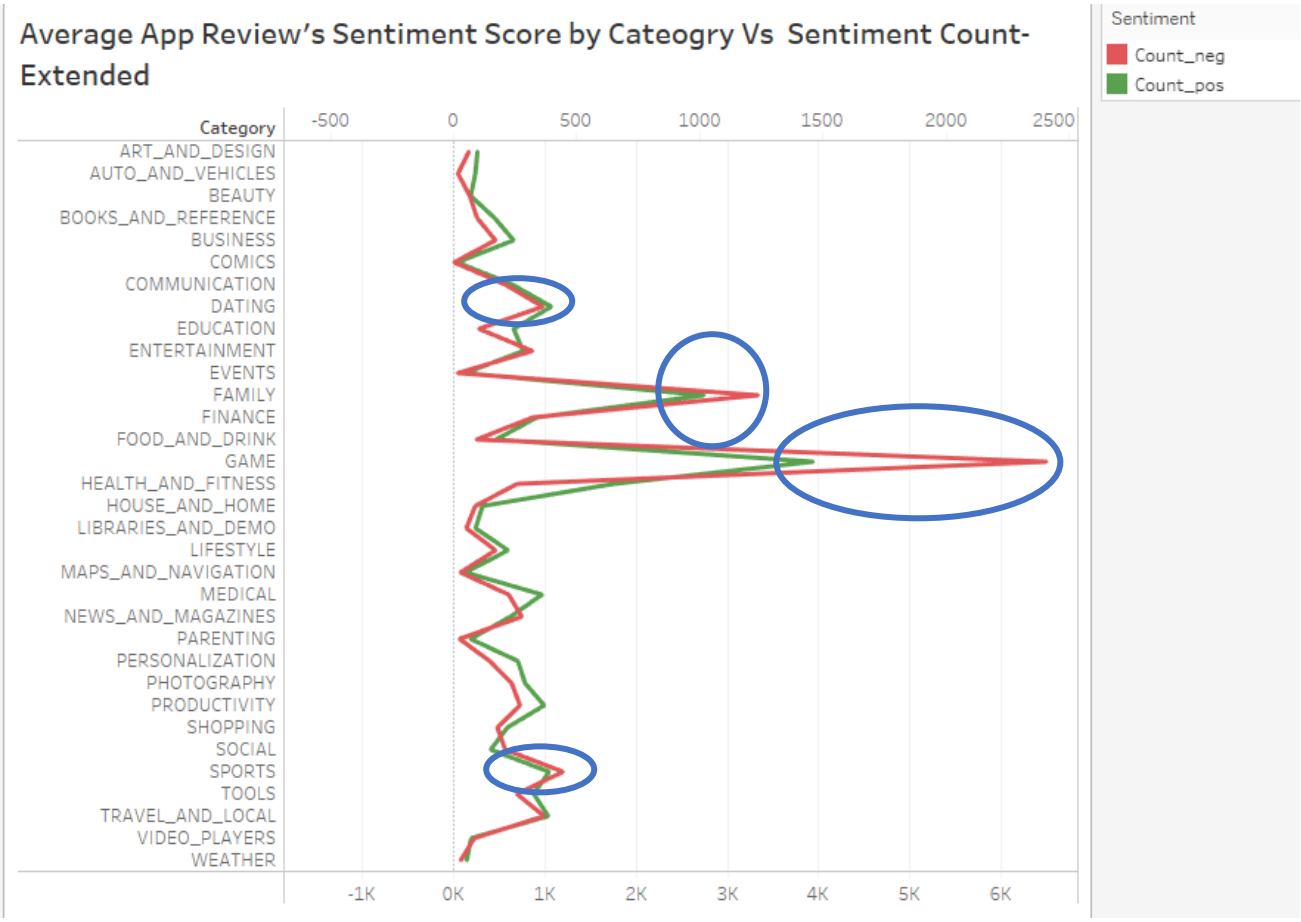

Average App Review's Sentiment Score by Cateogry Vs Sentiment Count- Extended
Sentiment
Count_neg
Count_pos
Category
ART_AND_DESIGN
AUTO_AND_VEHICLES
BEAUTY
BOOKS_AND_REFERENCE
BUSINESS
COMICS
COMMUNICATION
DATING
EDUCATION
ENTERTAINMENT
EVENTS
FAMILY
FINANCE
FOOD_AND_DRINK
GAME
HEALTH_AND_FITNESS
HOUSE_AND_HOME
LIBRARIES_AND_DEMO
LIFESTYLE
MAPS_AND_NAVIGATION
MEDICAL
NEWS_AND_MAGAZINES
PARENTING
PERSONALIZATION
PHOTOGRAPHY
PRODUCTIVITY
SHOPPING
SOCIAL
SPORTS
TOOLS
TRAVEL_AND_LOCAL
VIDEO_PLAYERS
WEATHER
-500
0
500
1000
1500
2000
2500
-1K
0K
1K
2K
3K
4K
5K
6K

Chart 22

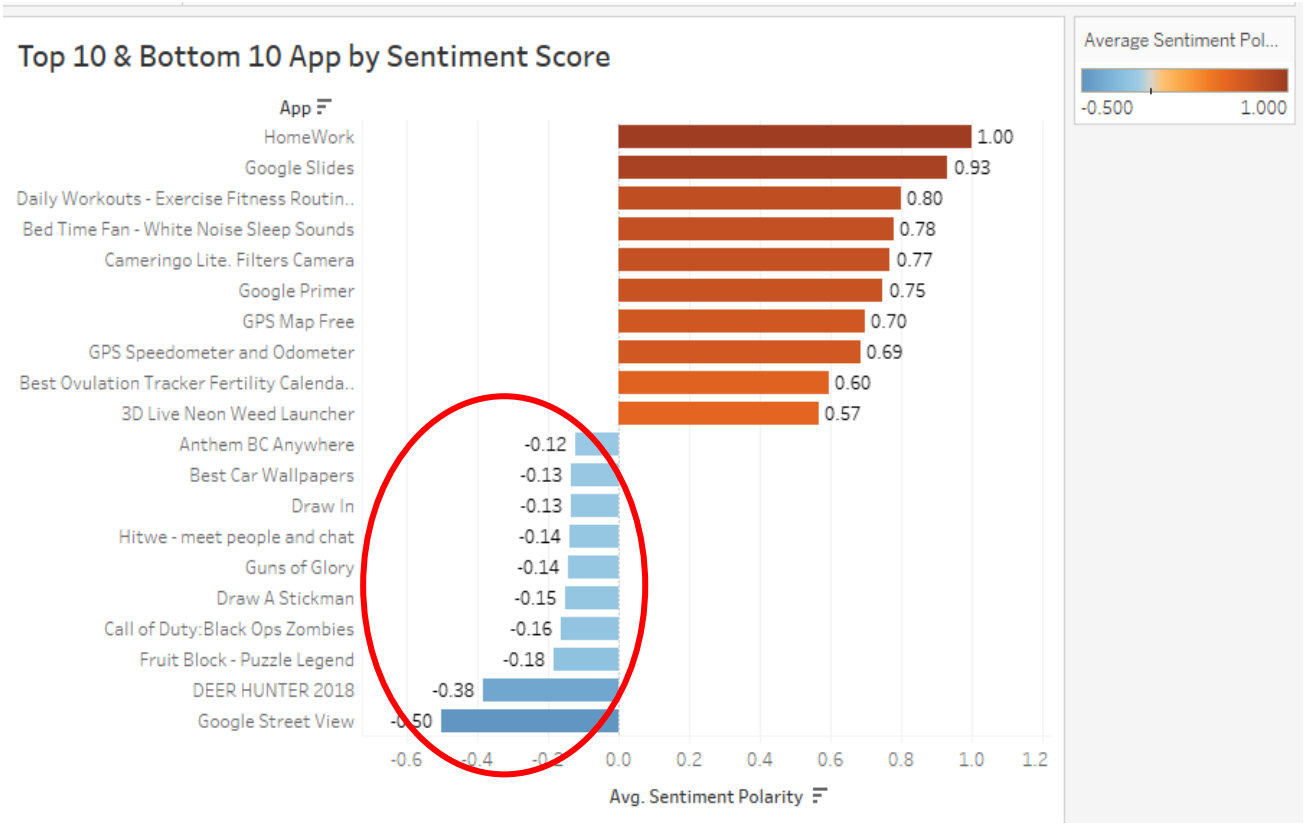


Chart 23

I would like to continue my further analysis based on user review sentiments which gives more interesting and valuable insights on user behavior patterns and future market capitalization opportunities.

Because of the high level of interest, I create few visualizations to identify sentiment behavior that guide to further analysis.

Chart 20 – Categories with highest positive sentiment and lower positive sentiment describe the story about user satisfaction. This is a good indication to identify categories or the segments with effective business opportunities and create advantages over the sentiment.

Chart 21 and Chart 22– These charts analyze the positive and negative sentiment counts over the categories. App developers, Product managers and the opportunity seeker are easy to determine which sector has low user sentiment and satisfaction. Further studies I like to go deeply on low positive sentiment segments [circled in blue].

Chart 23 – Indicates the high positive sentiment and low positive sentiment applications. Most of the low positive apps are gaming related. From use of this product managers and developers moving forward with different strategic changers to uplift the sentiment.

This is a very interesting analysis for me. I learn BI, analysis as well as good domain knowledge of google app store. In future I would like to move forward this analysis with different stake holder related business questions. Especially for android versions and sentiment analysis on reviews etc.

1.Rimsha, M. (2020) Https://www.researchgate.net/publication/343769728_analysis_of_google_play_store_data_set_and_predict_the_popularity_of_an_app_on_google_play_store. Analysis of Google Play Store Data Set and Predict the Popularity of an App on Google Play Store [online]. 11, pp. 1-15. [Accessed 01 June 2024].

2. Laura, C.E.C.I. (2024) Google Play Store - Statistics & Facts. Statista [online]. [Accessed 01 June 2024].

3.Bhimasen, M., Bhramara Bar, B. and Snehasis, D. (2023) Play Store App Analysis & Rating Prediction Using Classical Ml Models & Artificial Neural Network. Research Gate [online]. [Accessed 02 June 2024].

4. Parul,P.and , (2019) A Guide to Handling Missing Values in Python. *Kaggle* [online]. [Accessed 05 June 2024].

5. Nirajan, K. (2017) 6 Most Popular Techniques for Handling Missing Values in Machine Learning with Python. Dataaspirant [online]. [Accessed 05 June 2024].

6. Jiaxu, P., Jungpil, H. and Ke-wei, H. (2022) Handling Missing Values in Information Systems Research: A Review of Methods and Assumptions. Information System Research [online]. [Accessed 05 June 2024].

7. Amalia, K., Laura, B. and John, B. (2012) A Review of the Reporting and Handling of Missing Data in Cohort Studies with Repeated Assessment of Exposure Measures. Bmc Medical Research Methodology [online]. [Accessed 06 June 2024].

8.Nasima, T. (2020) Effective Strategies For Handling Missing Values in Data Analysis. Analytics Vidya [online]. [Accessed 06 June 2024].

9.Zhongheng, Z. (2016) Missing Data Imputation: Focusing on Single Imputation. National Library Medicine [online]. [Accessed 07 June 2024].

10.Rima, H. and Ahcène, B. (2014) Handling Missing Data Problems with Sampling Methods. Ieee [online]. [Accessed 07 June 2024].

11.Jason, W. (2013) Dealing with Missing Or Incomplete Data. Best Practices in Data Cleaning [online]., pp. 105-135. [Accessed 08 June 2024].

12.(2020) Get Started with Calculations in Tableau. Tableau Desktop and Web Authoring Help [online]. [Accessed 08 June 2024].

13. (2020) Create a Simple Calculated Field. Tableau Desktop and Web Authoring Help [online]. [Accessed 08 June 2024].

14. Benn, S., Jason, A. and O'Leary, B. (2016) Defining and Identifying Stakeholders: Views From Management and Stakeholders. Research Gate [online]., pp. 105-135. [Accessed 10 June 2024].

15. Annette, B., Tephen, H. and Robert, B. (2019) How to Engage Stakeholders in Research: Design Principles to Support Improvement. Health Research Policy and Systems [online]., pp. 105-135. [Accessed 10 June 2024].

16.(2023) Root Cause Analysis: The Art of Solving Problems in Business. Motion [online]. [Accessed 11 June 2024

17. Pavleenk, K. (2024) Tableau For Beginners – Data Visualisation Made Easy. Analytics Vidya [online]. [Accessed 11 June 2024].

18. Parul, P. (2018) Data Visualization with Tableau. Data Camp [online]. [Accessed 11 June 2024].

19. Ryan, M. (2017) Data Storytelling: How to Tell a Great Story with Data. Thoughtspot [online]. [Accessed 15 June 2024]
20. Pratibha Kumari, J. (2022) Data Storytelling & Data Visualization. Linkeding Article [online]. [Accessed 15 June 2024].