EMPIRICAL RESEARCH ON MARKET PREFERENCES APPS IN GOOGLE PLAY

A PROJECT PROPOSAL SUBMITTED FOR THE PARTIAL FULFILMENT OF THE REQUIREMENTS OF THE ADVANCED DIPLOMA IN DATA SCIENCE PROGRAM

BY

D. JAYENDRA PRABHASHI MAHAWATTA

COADDS 20.1F – 016

INDEPENDENT RESEARCH PROJECT

ADVANCED DIPLOMA IN DATA SCIENCE

NATIONAL INSTITUTE OF BUSINESS MANAGEMENT

COLOMBO, SRI LANKA

05TH NOV 2022

DECLARATION

I hereby declare that the work presented in this project report was carried out independently by myself and have cited the work of others and given due reference diligently.

Date: 11/01/2022

D.J. Prabhashi Mahawatta Signature of the Candidate

I certify that the above student carried out his/her project under my supervision and guidance.

Date: …… …….…….

W.M.S.G.D.C.Wanigasekara Signature of the Supervisor

ACKNOWLEDGEMENT

The completion of this paper would not have been possible without the presence of a group of people who allocated and extended their valuable time and assistance throughout the course of this research.

I would first like to express my sincere gratitude to my parents who gave me loving care and unstoppable encouragement throughout this research study as well as throughout my life to make me a productive scholar.

I would like to express my heartfelt gratitude to the lecturers of all the modules that have been completed so far, for providing me with all the necessary and imperative knowledge and skills without which this report could not have been materialized.

I would also like to express my sincere gratitude to my supervisor Associate Professor, Miss. W.M.S.G.D.C.Wanigasekara of the National Innovation Center at National Institute of Business Management. I would also like to thank Mr. Thurairasa Balakumar, Lecturer for Statistics and the Course Director of my degree program for his complete and unrelenting support with the completion of this paper.

Finally, I would also like to take this opportunity to sincerely thank all my colleagues and everyone else who provided insight and expertise that assisted the research in any manner possible, even though I have not mentioned them explicitly here.

EXECUTIVE SUMMARY

Mobile app distribution platforms such as Google play store gets flooded with several thousands of new apps everyday with many more thousands of developers working independently or in a team to make them successful. With immense competition from all over the globe, it is imperative for a developer to know whether he is proceeding in the right direction. Unlike making a movie where the presence of popular celebrities raises the probability of success even before the movie is released, it is not the case with developing apps. Since most Play Store apps are free, the revenue model is quite unknown and unavailable as to how the in-app purchases, in-app adverts and subscriptions contribute to the success of an app. Thus, an app's success is usually determined by the number of installs and the user ratings that it has received over its lifetime rather than the revenue it generated. In this thesis, on a smaller scale, I have tried to perform exploratory data analysis to dive deeper into the Google Play Store data, discovering relationships with specific features such as how the number of words in an app name for instance, affect installs, in order to use them to find out which apps are more likely to succeed and to identify the market preferences apps.

TABLE OF CONTENTS

LIST OF FIGURES

List of Tables

CHAPTER 01

INTRODUCTION

1.1 Background

The Google Play Store started life as the “Android Market” in 2008. It launched alongside the very first Android devices, and its purpose was to distribute apps and games. The android Market was extremely basic at the beginning. It didn’t support paid apps and games until 2009. However, as the Android platform grew, Android Market distributed apps and games. By 2012, it featured over 450,000 Android apps and games. By this time, Google’s ecosystem had expanded greatly compared to the humble beginnings of the Android Market. In fact, the Android Market was just one of the company’s online markets. 6 The creation of the Google Play Store in 2012 was the culmination of three separate online markets that Google was running at the time. It combined the Android Market, the Google Music Store, and the Google eBookstore. The Google eBookstore launched in 2010 with over three million eBooks. Despite the large library, it was mostly filled with public domain titles and scans. Google Music launched in 2011 in beta, and while fans loved the local uploading feature, its library of music to purchase wasn’t big.

Unlike Android apps and games, the Google Music and eBook stores aren’t exclusive to Android phones and tablets. Google was taking the same approach as Apple, which keeps the App Store, Apple Books, and iTunes as separate entities. However, Google’s stores weren’t nearly as popular, despite having wider availability. To more accurately reflect the

scope of what Google had to offer, all three stores were combined under the “Google Play” brand. The eBookstore became “Google Play Books” and Google Music became “Google Play Music,” all found in the Play Store. The Play Store is a digital marketplace. However, it used to sell physical devices, too. For a brief time, Google sold Nexus devices, Chromecasts, and Chromebooks through a “Devices” tab on the Play Store. At the time, this was the only place that Google had to sell goods. As the company’s hardware efforts grew, it was time for a new store. The Google Store was created in 2015 for the company’s hardware, and the “Devices” tab on the Play Store now directs to that instead.

1.2 Research Problem

The expansion of smart phones is driving the fast development of mobile app stores. Currently, the two largest global platforms for app distribution are Apple’s App Store (for iOS users) and Google play store (official app store for the Android OS). For my research, I have picked Google play store and did a thorough analysis of its features that were available to me to identify the market preferences apps. But the question arises, is it even necessary to do so? Well, the answer is yes, that is because an average of 6,140 mobile apps are released through the Google Play Store every day, according to the statistics shown by Statista in their recent report where all the app in together compete for user attention [2]. And the number of available apps in the Google play store is estimated to be around 2.6 million applications in March 2018 subsequent to outperforming 1 million apps in July 2013 [1]. The huge number of apps in the play store and the numbers of the apps released every day make it quite

competitive for the app developers, the companies who are building an app to come up with a unique idea that will definitely be bought by the end users. Because at the end of the day if the app does not perform well in the android market, then all the hard work behind building the app will go in vain. As the mobile industry is growing rapidly it is increasing the level of competition however, increased competition also leads to increased chances of failure. So, the developers need to do enough research as an enormous amount of time, effort and the money are invested into the process, so business cannot afford an app failure. If we look into the history of the revenue growth of play store, then between 2016 and 2017, play store has seen a revenue growth of 34.2%, with the percentage rise in app downloads was 16.7% [16]. Well, the developer does not get all of the money an app makes, he gets 70% of the total money and play store keeps 30 % when the app is launched in the play store, and there is no minimum threshold when it comes to the amount that an application is to earn, any amount above 1$ gets credited to the account. Building an app is therefore not an easy task to deal with, as there are a bunch of developers creating apps every now and then. But to reach the level of success that makes an app stand out in the crowd. According to the State of Mobile App Developers study, around one-third of the app developers have less than 10k downloads across all regions, globally only 15% have crossed >1000k download mark [18]. The lower the number of downloads the less it has a chance that it will do great business ahead in future in the android market. This is one big problem that I tried to solve in my research. Success is a thing that does not come to one so easily and for that, I analyzed the features of Google play store and came to a conclusion that will help developers to understand the market and the audience preferences.

1.3 Objectives of The Project

- The main objective of this project is to deliver insights to understand customer demands better and thus help developers to popularize the product.

- The study's goal is to uncover market preferences by examining Google Play Store categories, ratings, reviews, price, size, and so on.

- Determine what types of apps have the most downloads, ratings, and reviews, including paid and non-paid effects on the app's ratings and reviews.

- Determining the impact of numerous app factors on Google Play app installation.

1.4 Research Questions

- How does app rating and reviews affect app installation?

- What types of apps have the most downloads, ratings, and reviews?

- How does paid and non-paid effects affect the app's ratings and reviews?

- What is the relationship between price, size, rating and reviews?

- Does the number of words in an app name for instance, affect installs?

1.5 Scope of The Research

The study strictly focused on identifying the app market and the audience preferences. The explanations for all the analysis performed are described in detail in Chapter 4. From the EDA performed, I observed the importance of each feature and the correlations between each of them. Hence, from the analysis, it can be seen that identifying the market preferences will play a very important role for developers and can bring certain changes that might affect the lifetime of their app.

1.6 Justification of the Research

Due to the expansion and competition among the market I chose to do research relevant to this sector so I did a thorough analysis of my data of Google Play Store which I think will be a big contribution for the developers as through this they will get to know market preferences and will be able to decide what feature needs to be maintained or which one needs to be modified according to the current state of their app. My main motive for this research was to find something meaningful throughout the analysis that might matter to the developing community or to the end users.

1.7 Expected Limitations

Firstly, the sample is not representative for the whole Google Play Store, so insights should be derived with caution.

Secondly, Lack of strong explanatory variables such as marketing expenditure, in-app-purchases and cross platform effects (Facebook is a network not an app).

Lastly, one would assume, app developers want to maximize profit and "sell" as many installs as possible, or earn money with in app purchases. This is not always the case. Some developers are sub-contractors and program apps for third parties. They do not want to sell apps, but use them in a product bundle.

1.8 Research Outline

Chapter 1: Contains a brief overview of my research, my estimated goal and how my work can help the developers.

Chapter 2: Discussion of the literature review and background study of my thesis, including detailed description of the algorithms I used in our research.

Chapter 3: Description of the process of collecting data, data processing and description of the corresponding features.

Chapter 4: Methodology of the research

Chapter 5: Discoveries about the correlation between the features that I found by performing EDA with visual representations.

Chapter 6: Conclusion and proposed future work for the system.

CHAPTER 02

LITERATURE REVIEW

2.1 Introduction to the Research Theme

It has been observed that the significant growth of the mobile application market has a great impact on digital technology. Having said that, with the ever-growing mobile app market there is also a notable rise of mobile app developers; eventually resulting in sky-high revenue by the global mobile app industry. With immense competition from all over the globe, it is imperative for a developer to know that he is proceeding in the correct direction. To retain this revenue and their place in the market the app developers might have to find a way to stick into their current position. And also, the sentiment of Google Play Store users gives us an idea of how the applications market reacts, what are the needs and how to succeed in the android market. The opinion of the audience is an essential component for every business for further development and improvement of the applications.

2.2 Theoretical Explanation about the Key Words in the Topic

Android: Is a mobile operating system based on a modified version of Linux kernel and other open-source software, designed primarily for touchscreen mobile devices such as smartphones and tablets.

Google Play: Also branded as the Google Play Store and formerly Android Market, is a digital distribution service operated and developed by Google.

Sentiment analysis: Also known as opinion mining or emotion AI is the systematic identification, extraction, quantification, and study of affective states and subjective information using natural language processing, text analysis, computational linguistics, and biometrics.

Variable importance: Variable importance is determined by calculating the relative influence of each variable

PMF: The Probability Mass Function (PMF) is also called a probability function or frequency function which characterizes the distribution of a discrete random variable. Let X be a discrete random variable of a function, then the probability mass function of a random variable X is given by. $\mathbf{P_x\ (x) = P(\ X=x)}$

Correlation: Correlation is a statistical measure that expresses the extent to which two variables are linearly related (meaning they change together at a constant rate). It's a common tool for describing simple relationships without making a statement about cause and effect.

Linear Regression: Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable.

Logistic Regression: Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist. In regression analysis, logistic regression (or logit regression) is estimating the parameters of a logistic model (a form of binary regression).

2.3 Finding by Other Researchers

According to Statista.com published by L.Ceci, May 18, 2022 Mobile apps are projected to generate more than 613 billion U.S. dollars in revenues in 2025, with mobile games making up the biggest revenue share among all app categories. In 2020, gaming and video made up the largest shares of the mobile content market for the year. The ePublishing and education sectors still saw a limited market for their mobile content, despite the increase in apps usage brought by the COVID-19 pandemic disrupting regular school system settings.

As an indispensable part of the smartphone experience, the largest number of apps in the major app stores are free to download. However, in recent years, the growth of global consumer spending on apps has shown users' healthy appetite for premium services or paid app content. In the second quarter of 2021, Android consumers spent an average of 5.31 U.S.

dollars per handset, after peaking in the last quarter of 2020 reaching an average of 10.6 U.S. dollars per mobile device. As of September 2021, the number of paid apps has shrunk to make up only six percent and four percent of the total numbers in the Apple App Store and the Google Play Store, respectively. In comparison, apps offering subscription plans are becoming increasingly popular in the monetization landscape. In 2020, the leading subscription apps in the Apple App Store generated more than 10 million U.S. dollars in global revenues.

Mobile internet users are spoilt for choice when it comes to the sheer variety and availability of mobile apps. According to the same publisher, in the first quarter of 2022, gaming apps were the most popular app category in the Google Play Store, accounting for 13.63 percent of available apps worldwide. Mobile education apps ranked second with a 10.41 percent share.

Google Play is the biggest digital distribution platform for Android apps worldwide, offering users close to 3.3 million mobile apps to choose from. However, Google Play offers more than just apps – the platform also provides users with music, video, e-book downloads and rental services. Most mobile apps on Google Play are available for free. As of March 2022, over 96 percent of Android apps could be downloaded without having to pay for them upfront, although this does not preclude other mobile app monetization strategies such as in-app advertising and in-app purchases. As of the beginning of 2022, the ranking of leading Android apps in the Google Play Store worldwide based on revenue consists mainly of gaming apps, including the mobile version of the Roblox platform and Clash of Clans.

As reported by Statista Research Department, Sep 15, 2021, In 2020, global revenue from mobile apps increased to over 318 billion U.S. dollars. The number of mobile app downloads also increased this year, with games representing the biggest share. Almost 100 billion downloads were made in the mobile games segment. With around 19.6 billion downloads, mobile photo and video apps were the second most download apps that year.

2.4 The Research Gaps

Even though the data set is quite interesting and hard to obtain, it is neither representative for the whole Google Play Store, nor can we find any significant relation between installs – except rating count, which can be a proxy for an App developer's success -- and other observed features/variables.

Better data quality is necessary to find robust evidence.

2.5 Table for Variables

Sources: Kaggle (Google Play Store Apps)

Variable	Definition
App Name	Name of the app
App Id	Package name
Category	App Category
Rating	Average rating
Rating Count	Number of ratings
Installs	Approximate install count
Minimum Installs	Approximate minimum app install count
Maximum Installs	Approximate maximum app install count
Free	Whether app is Free or Paid
Price	App price
Currency	App currency

Size	Size of application package
Minimum Android	Minimum android version supported
Developer Id	Developer Id in Google Playstore
Developer Website	Website of the developer
Developer Email	Email-id of developer
Released	App launch date on Google Playstore
Last Updated	Last app update date
Content Rating	Maturity level of app
Privacy Policy	Privacy policy from developer
Ad Supported	Ad support in app
In App Purchases	In-App purchases in app
Editors Choice	Whether rated as Editor Choice.
Scraped Time	Scraped date-time in GMT

CHAPTER 03

DATA PREPARATION PROCESS - DATA PREPROCESSING AND DATA WRANGLING

3.1 Data Cleaning

Google-Playstore.csv

First, I drop the entries that are missing the app name, since it is difficult to impute data without this information. Then, I'll drop unnecessary columns from the dataframe. These columns are 'unnecessary' in the sense that we won't be able to analyzes meaningful insights from the information within these columns, since they either contain duplicate information that is present in other columns, or is just administrative data (IDs)

Dropped Columns: App Id, Developer Id, Developer Website, Developer Email, Privacy Policy, Scraped Time, Minimum Installs, Maximum Installs, Installs

Missing Data

Google-Playstore.csv

	Count	Percentage
Released	71053	3.071972
Rating	22883	0.989345
Rating Count	22883	0.989345
Minimum Android	6530	0.282324
Size	196	0.008474
Currency	135	0.005837
Average_Installs	107	0.004626
App Name	2	0.000086
Category	0	0.000000
Free	0	0.000000
Price	0	0.000000
Last Updated	0	0.000000
Content Rating	0	0.000000
Ad Supported	0	0.000000
In App Purchases	0	0.000000
Editors Choice	0	0.000000

Figure 3.1: Checking Null Values

There isn't much downside to dropping the entirety of the rows with missing data, since less than 5% of the data would be dropped. This wouldn't cause a meaningful effect on the insights that we are able to generate from the data.

So, I drop rows with these missing values. (Released, Rating, Minimum Android, Rating Count, Size, Average Installs)

Googleplaystore_user_reviews.csv

```
playstore_user.isnull().sum()
```

```
App                        0
Translated_Review      26868
Sentiment              26863
Sentiment_Polarity     26863
Sentiment_Subjectivity 26863
dtype: int64
```

Figure 3.2: Checking Null Values

I drop all the rows with the missing values.

3.2 Data Transformation

Google-Playstore.csv

Some columns with the object and float data type make more sense to be represented as an int. For example, the rating count column, which represents the number of people giving ratings makes more sense to be represented as an integer. Other columns also need to be rounded up, such as the Price and Rating columns.

I decided to remove the Installs column because it displays similar information to the Minimum Installs and Maximum Installs columns. The only difference is that the Installs column is formatted as a threshold, while the other 2 columns give out exact values.

Also, I will simplify some columns to remove unnecessary clutter, with one example being the Minimum Android column which has the 'and up' phrase in every single entry. This doesn't really provide any additional information, so it is best to remove it to declutter our dataset.

Some categories of interest like Music and Education are given with different labels: there are both 'Music & Audio' and 'Music' labels as well as 'Education' and 'Educational' for education. They should be merged together to represent a single category.

And I created two new columns named IsGame and WordCount.

Changings:

- Change the Rating Count dtype to int and Size dtype to float.
- Reformat release and update dates.
- Round up the Rating, Price, and Average Installs columns to 2 decimal points.
- Remove the 'and up' in the Minimum Android column for legibility.
- Collapse multiple categories into one.
- Create new columns named IsGame and WordCount

CHAPTER 04

METHODOLOGY

4.1 Introduction

This analysis approach is divided into three phases: data cleansing, data visualization, and liner regression. In the first step, I collected the raw data from Kaggle. Then I did basic data cleaning and data, I removed some unnecessary features and made it ready for data visualization.

4.2 Dataset

Kaggle dataset: https://www.kaggle.com/datasets/gauthamp10/google-playstore-apps

- Name: Google Play Store Apps
- Context: Google Play Store Android App Data. (2.3 million+ App Data)
- Content: The data was collected in the month of June 2021. The Dataset contains Application data of more than 2.3 million applications with the twenty-four attributes.
- Author: Gautham Prakash

4.3 Methods, Techniques and Tools

4.3.1 Methods

- Basic Statistics
- EDA
    1. Univariate Exploration
    2. Bivariate Exploration
    3. Multivariate Exploration
- Sentiment Analysis (Using googleplaystore_user_reviews.csv)
- Linear Regression

4.3.2 Techniques and Tools

Using Python data visualization tools and technologies, analyzes massive amounts of information and makes data-driven decisions about the Market Preferences App in Google Play Store. Use techniques like Pie Chart, Bar Chart, Histogram, Heat Map, Scatter Plot, Line Plot, Box Plot and WordCloud for visualization purposes. Help of pandas and scipy analyses, basic statistics, correlation and linear regression. Identify the features of app installation using simple linear regression and multiple linear regression models in Python. Use the cross table to identify the Relationship between categorical variables. And use sentiment analysis to identify the emotional tone behind reviews.

CHAPTER 05

DATA ANALYSIS, VISUALIZATION AND INTERPRETATION

5.1 Types of Variables

Categorical Data

Variable	Data Type
App Name	Object
Category	Object
Currency	Object
Minimum Android	Object
Content Rating	Object
Ad Supported	Boolean
In App Purchases	Boolean
Editors Choice	Boolean
Free	Boolean

IsGame	Boolean
Released	Date Time
Last Updated	Date Time

Numerical Data

Variable	**Data Type**
Rating	Float
Rating Count	Float
Price	Float
Average Installs	Float
Size	Float
WordCount	Integer

5.2 Data Analysis

Basic Statistics

Summary statistics of categorical variables

	Rating	Rating Count	Price	Average_Installs	WordCount
count	2.235309e+06	2.235309e+06	2.235309e+06	2.235309e+06	2.160285e+06
mean	2.206445e+00	2.759874e+03	1.047680e-01	2.420717e+05	3.585620e+00
std	2.108361e+00	1.987582e+05	2.660905e+00	1.791941e+07	2.125141e+00
min	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	1.000000e+00
25%	0.000000e+00	0.000000e+00	0.000000e+00	6.800000e+01	2.000000e+00
50%	3.000000e+00	6.000000e+00	0.000000e+00	6.030000e+02	3.000000e+00
75%	4.300000e+00	4.200000e+01	0.000000e+00	6.213500e+03	5.000000e+00
max	5.000000e+00	1.385576e+08	4.000000e+02	1.102881e+10	1.600000e+01

Figure 5.2.1: Summary statistics of categorical variables

```
print(f"The average rating of an app is {round(df['Rating'].mean(),2)}")
```

```
The average rating of an app is 2.21
```

```
most_installed = df.iloc[df['Average Installs'].idxmax()]
print(f"The most installed app in the Google Play Store is {most_installed['App Name']},
      with {int(most_installed['Average Installs']):,.0f} installations.")
```

```
The most installed app in the Google Play Store is Google Play services, with 11,028,813,508 installations.
```

That's over 11 billion installations! However, this isn't surprising, since most Android phones have this app pre-installed.

```
most_expensive = df.iloc[df['Price'].idxmax()]
print(f"The most expensive app in the Google Play Store is {most_expensive['App Name']},
      costing {most_expensive['Price']} dollars.")
```

```
The most expensive app in the Google Play Store is MESH Connect, costing 400.0 dollars.
```

5.3 Visualization and Interpretation

In this chapter I analyze the dataset to summarize their main features with visual representations to see what the data can tell us beyond the formal modeling.

5.3.1 Univariate Exploration

CATEGORY

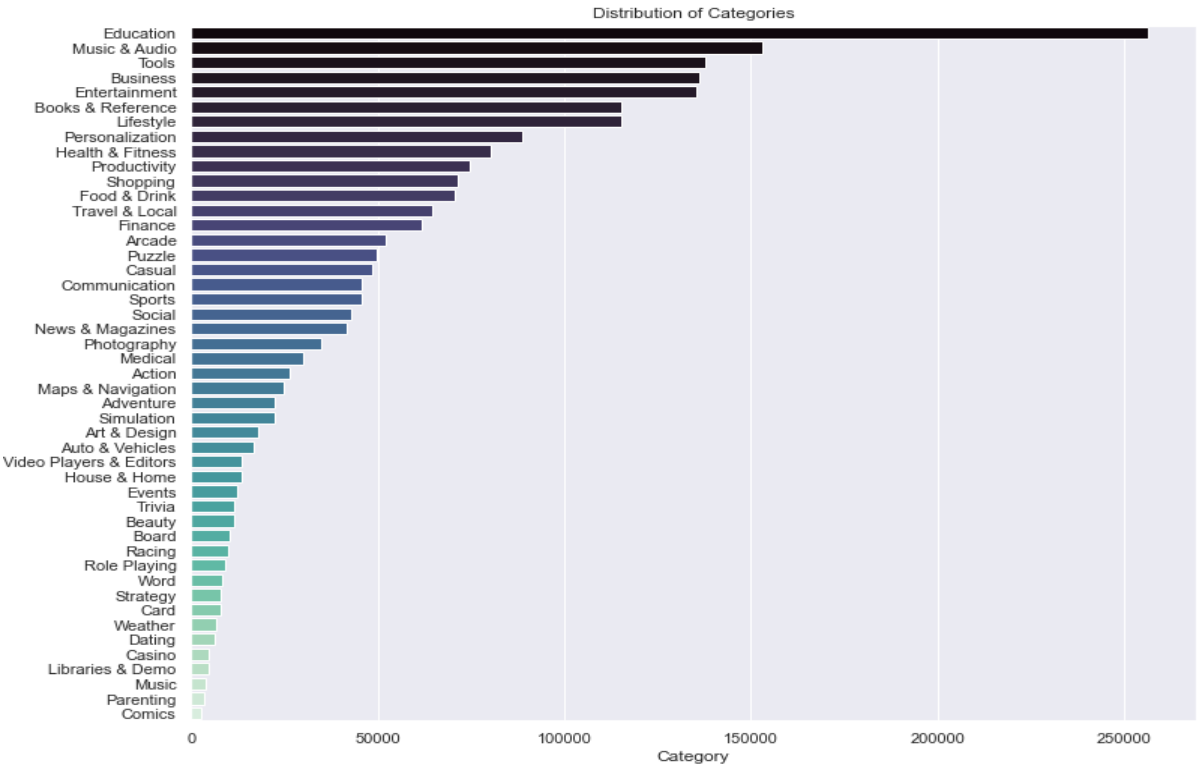

Figure 5.3.1.1: Distribution of Categories

As we can see, Maximum number of apps present in google play store comes under Education, Music & Audio, Tools, Business and Entertainment. and minorities are from Comics and Parenting.

MINIMUM ANDROID

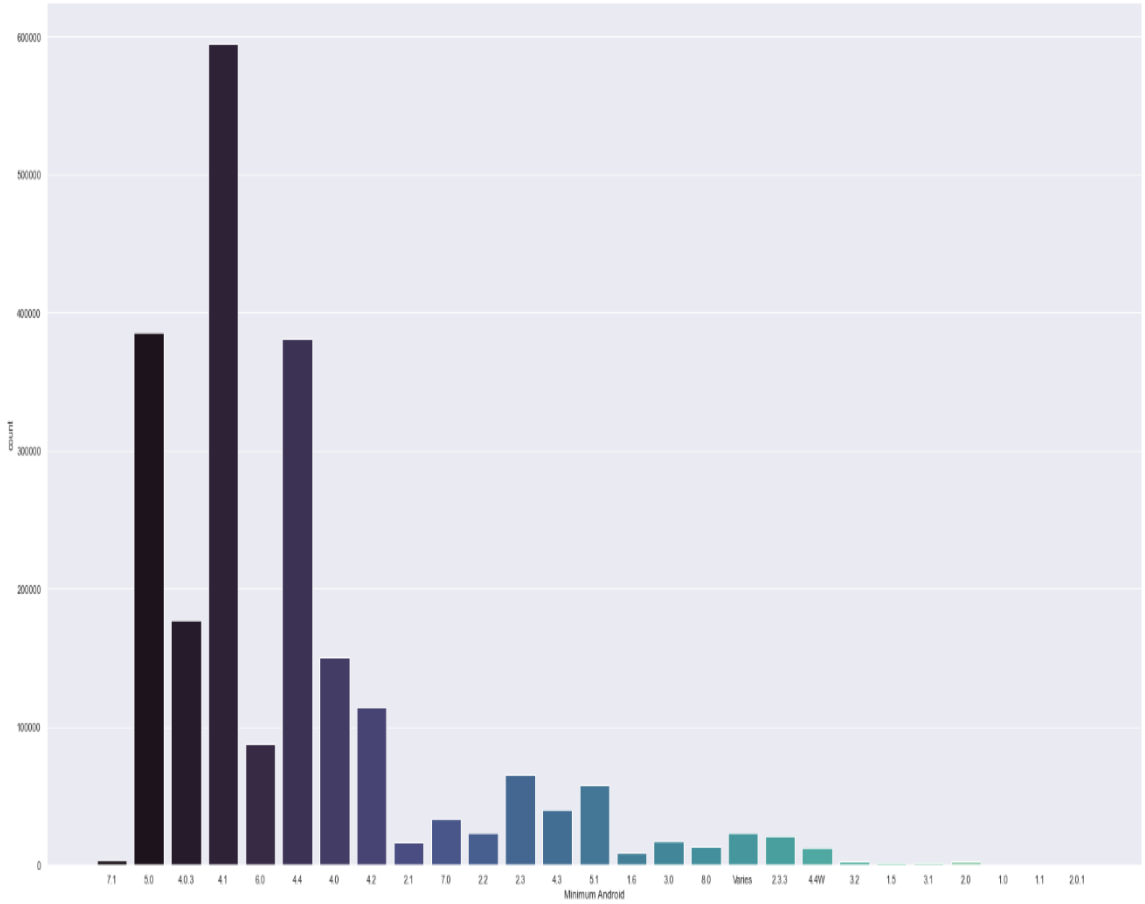

Figure 5.3.1.2: Minimum Android Version

The majority of apps are compatible with Android 4.1 version, according to the aforementioned graph. Android versions of 5 and 4.4 are also quite popular in google play store.

CURRENCY

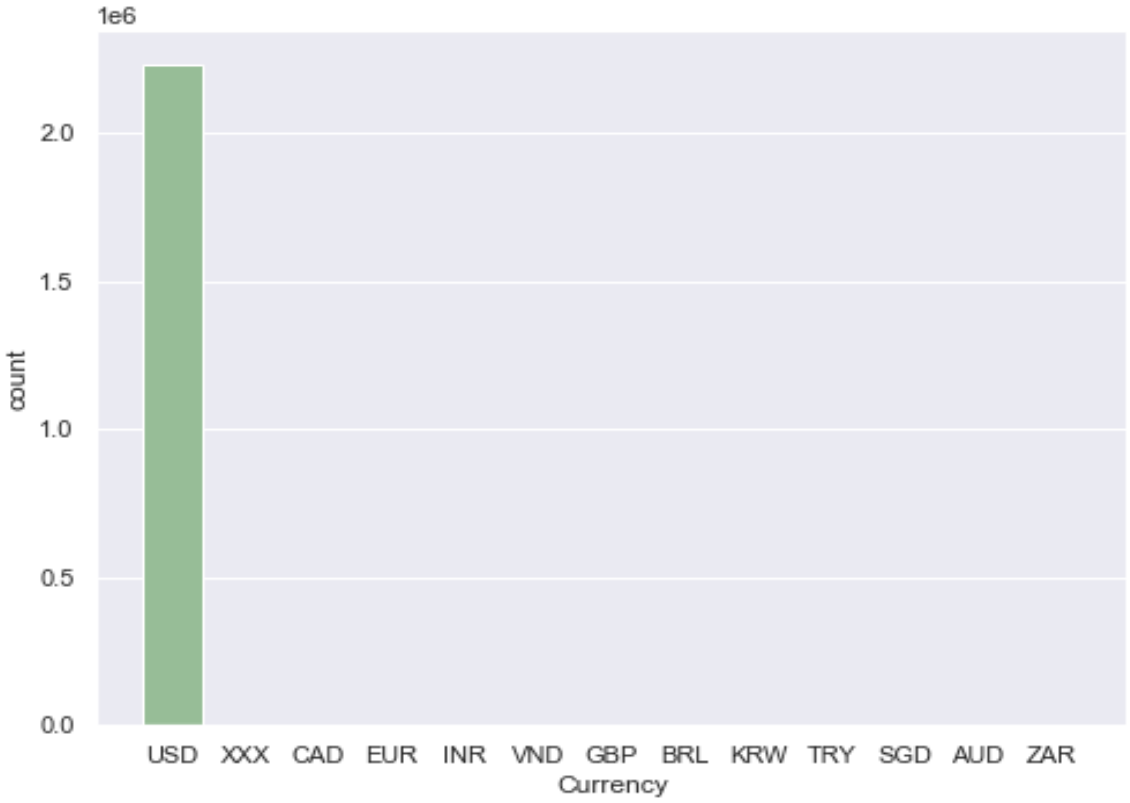

Figure 5.3.1.3: Distribution of Currency

Almost all the apps in Google Play Store are perches in USD. No significant quantity is coming from any other currencies.

CONTENT RATING

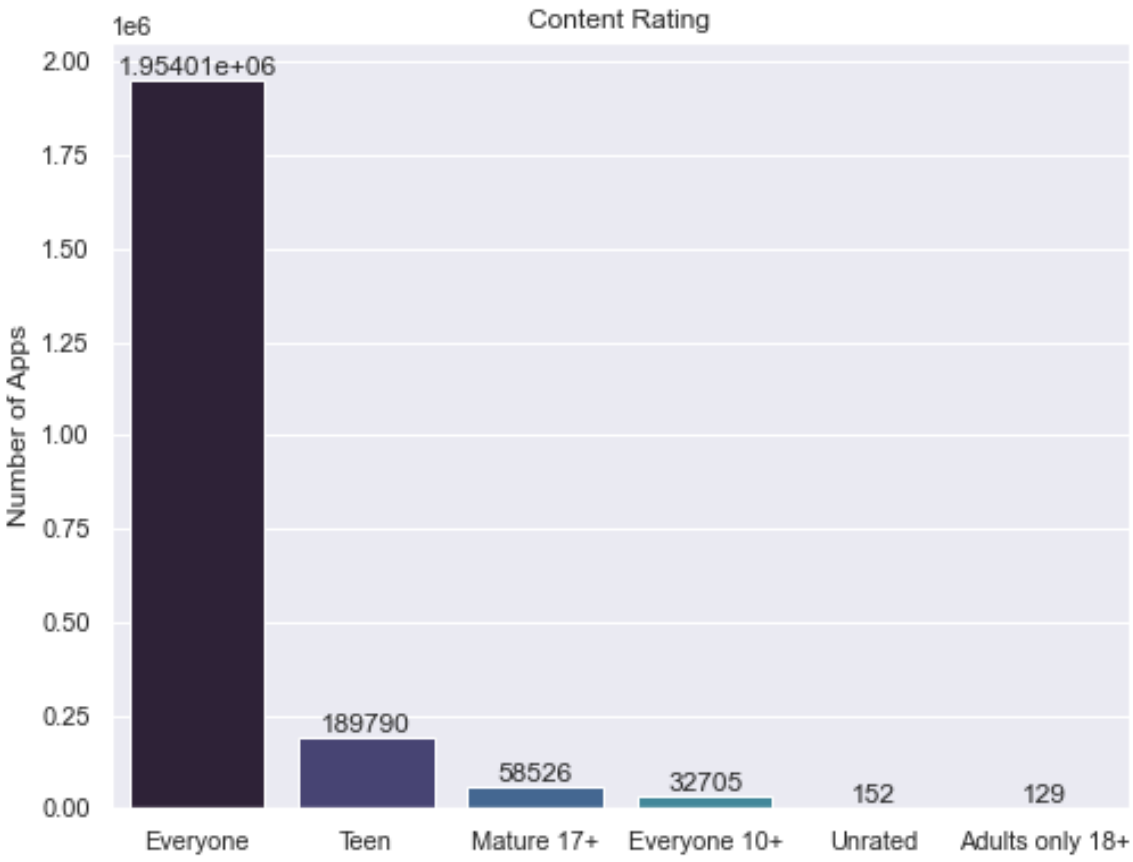

Figure 5.3.1.4: Distribution of Content Rating

Majority of the apps are built suitable for all age groups. Only 129 of the two million+ apps fall into the 18+ category. Regarding the above graph, there are 152 unrated apps. Despite the fact that teens are more susceptible to influence, there aren't many apps specifically made for them.

RELEASED

Figure 5.3.1.5: Number of App Released per Year

The number of apps released every year has been rising consistently, seeing an almost exponential increase in the last 6 years. The largest increase came between the years 2018 and 2019, where around 145k apps were released. That must've been a busy year for developers!

*Since the data was scraped in June 2021, the entries from 2021 aren't complete yet at the time.

LAST UPDATED

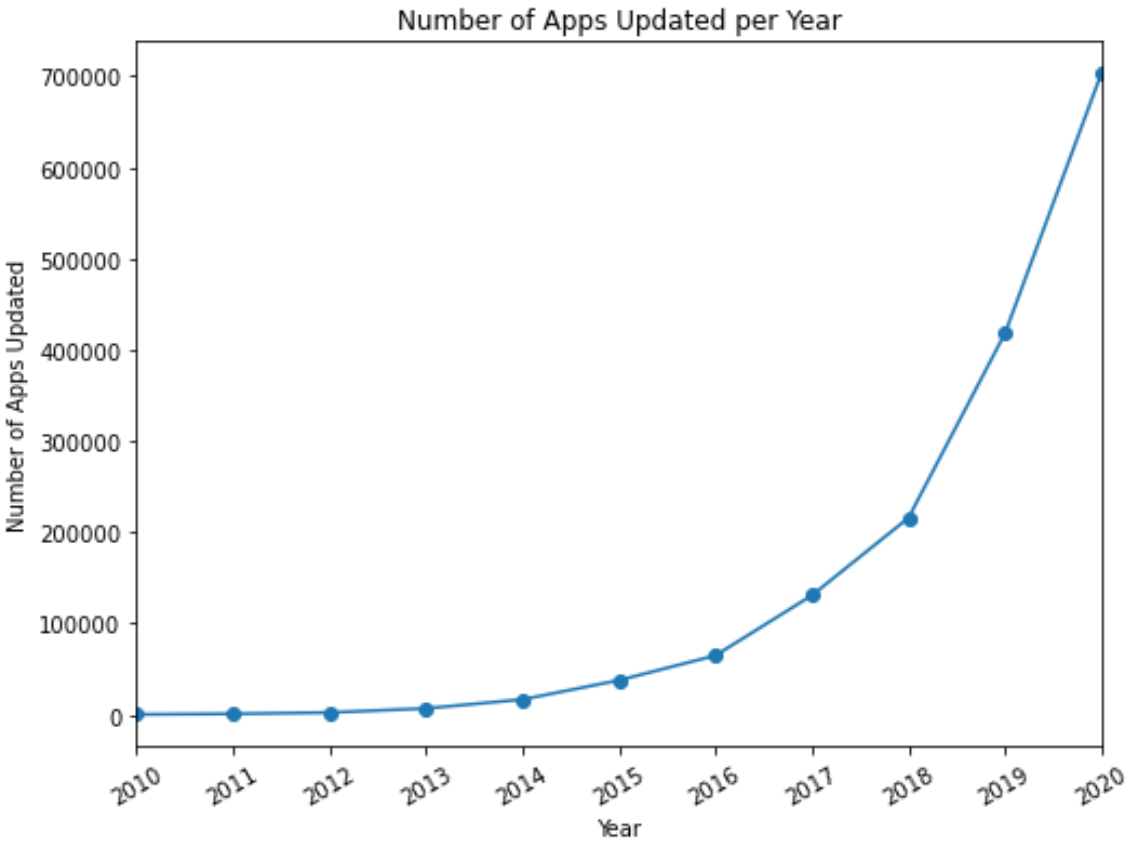

Figure 5.3.1.6: Number of Apps Updated

Figure 5.3.1.6 shows that the number of updates increases each year, with more updates occurring for apps starting in 2016 and later. However, only 700000 apps out of more than 2 million were upgraded in 2020.

FREE

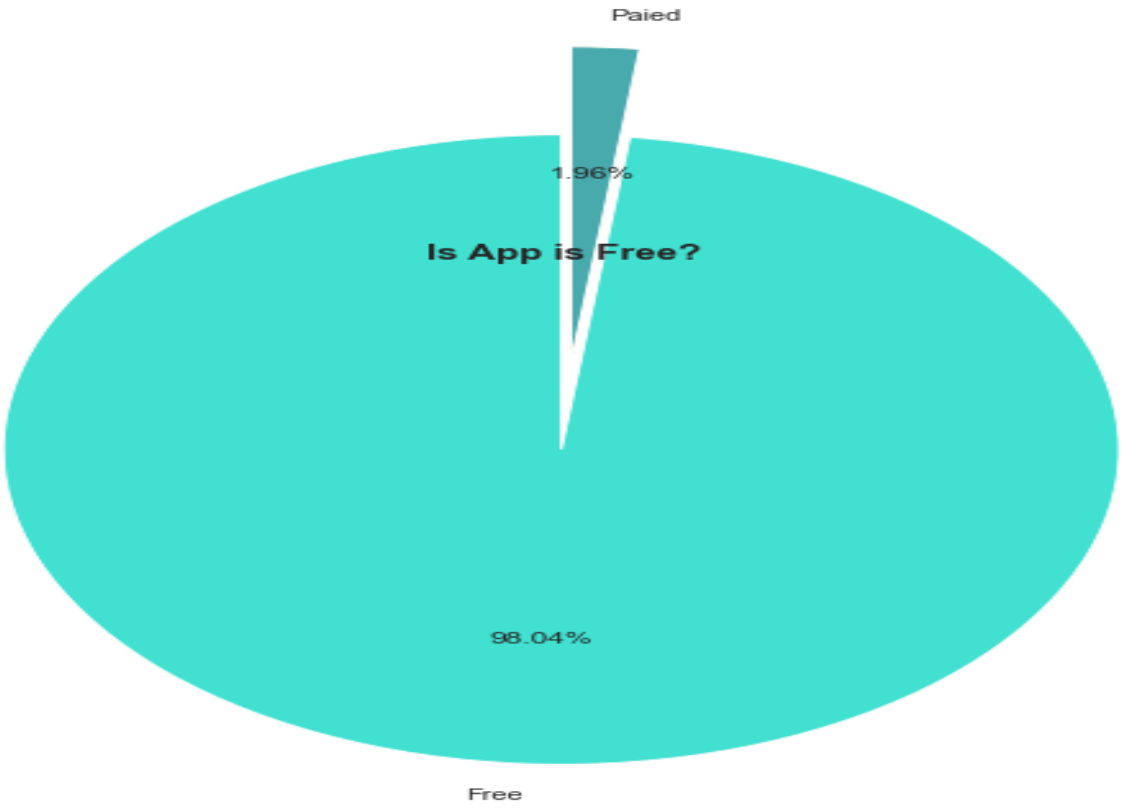

Figure 5.3.1.7: Distribution of Free and Paid Apps

Free apps dominate the Google Play Store, as is clear. Only 1.96% of the apps in the Google Play Store are paid, while 98% of them are free.

IN APP PURCHASES

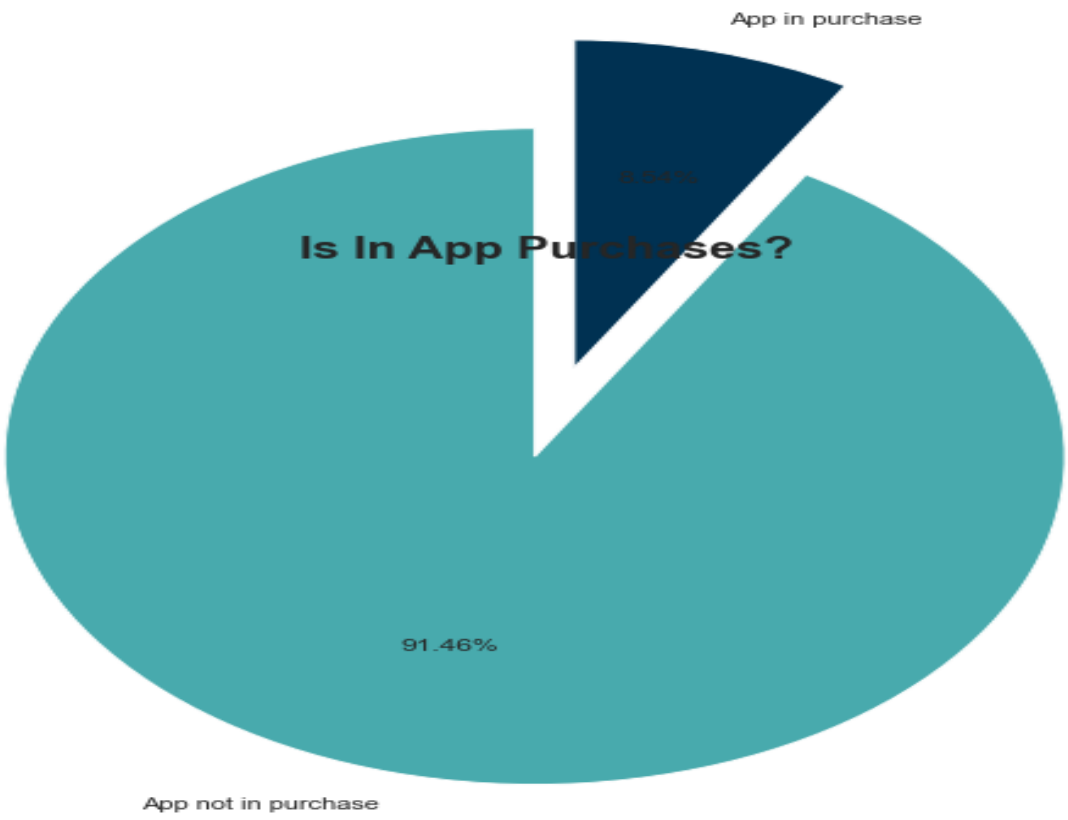

Figure 5.3.1.8: Distribution of In App Purchase

In-app purchases are extra content or subscriptions that you buy inside an app. Just 8.54% of app downloads involve in-app purchases in the android market.

EDITORS CHOICE

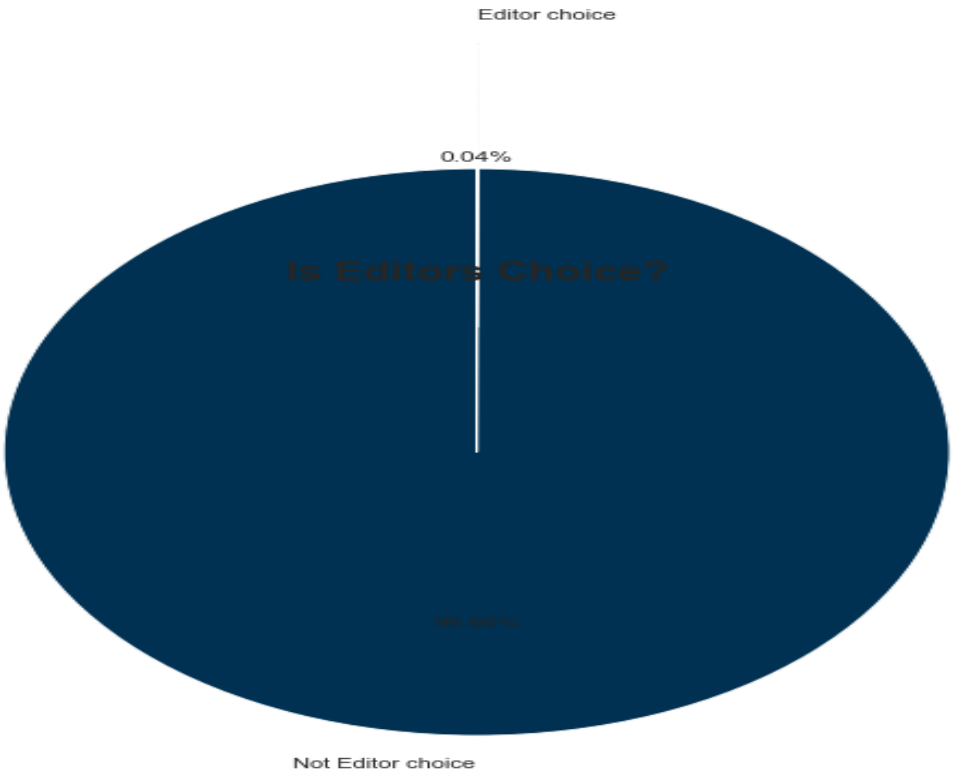

Figure 5.3.1.9: EDITORS CHOICE

Editors' choice refers to the apps and games that introduce users to the best innovative, creative, and designer apps on android. Google also released an improved editors' choice section that features app and game reviews accumulated and organized by the editorial team of play store. If your app gets featured on editor's choice google play, there is a high chance for your app to gain million downloads. However, only 0.04 percent apps out of more than 2 million are featured on editor's choice google play. It is evident that creating an app that can be listed on editor's choice in google play is not simple.

IS GAME

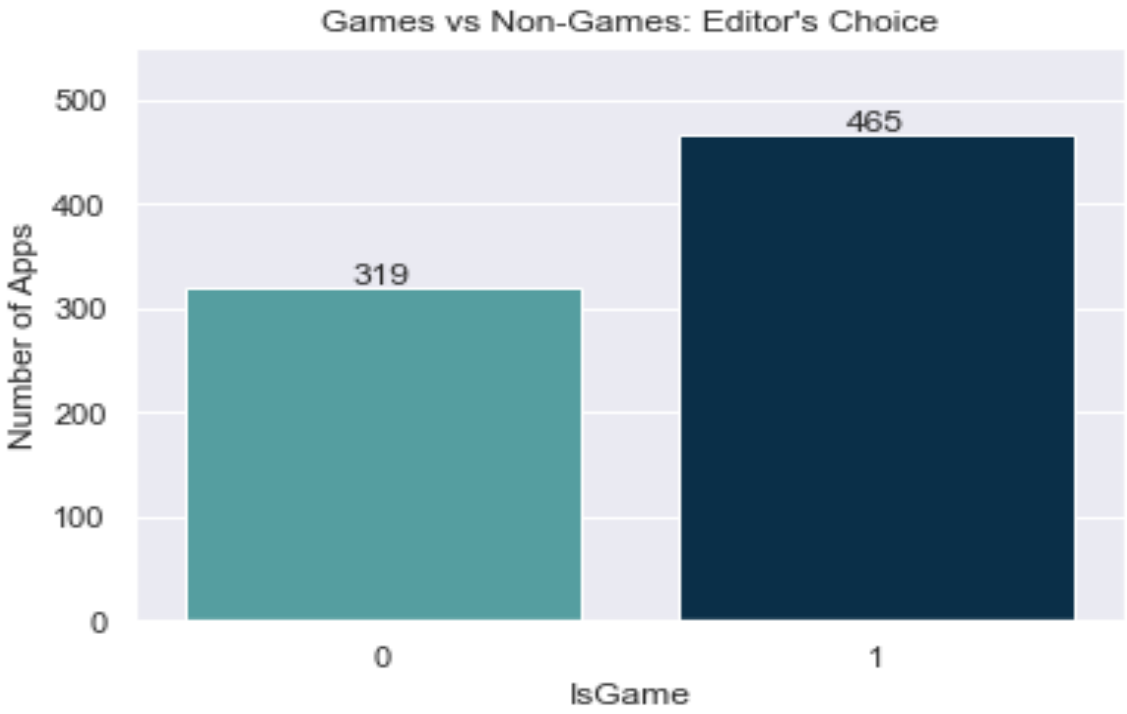

Figure 5.3.1.10: Games vs Non-Games: Editor's

Despite comprising less than 20% of the Play Store's population, over 60% of Editor's Choice apps were games. Per Google, Editor's Choice apps were "*the apps and games that introduce users to the best innovative, creative, and designer apps on Android*". So, it's clear to say that innovation (at least what's defined by the Play Store team) seems to have been more apparent from games.

AD SUPPORTED

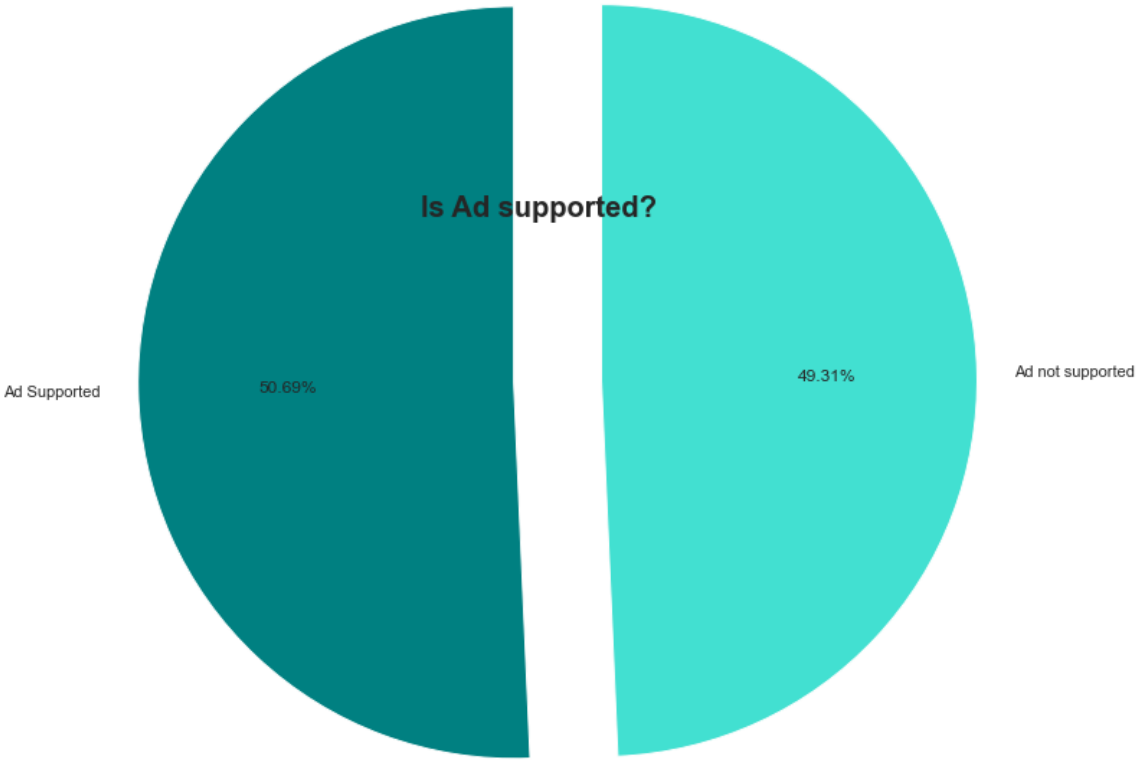

Figure 5.3.1.11: Ad Supported

The following graph illustrates that there wasn’t any significant difference between the apps with advertising and those that do not fit into this group.

The term "ad-supported" refers to whether or not the program or application features advertisements. or the software or app might include PUPs (potentially unwanted programs) Sometimes it's as simple as a company advertising their other products and it can be as bad as potentially unwanted programs (PUP's) you need to watch for during installation.

RATING

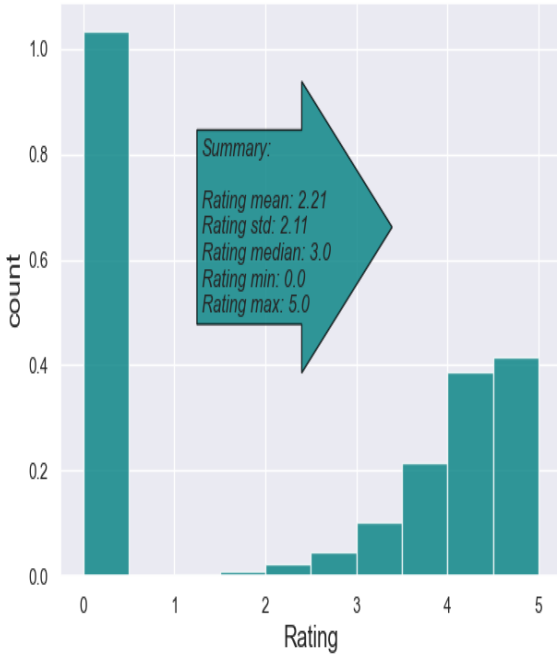

Figure 5.3.1.12: Rating

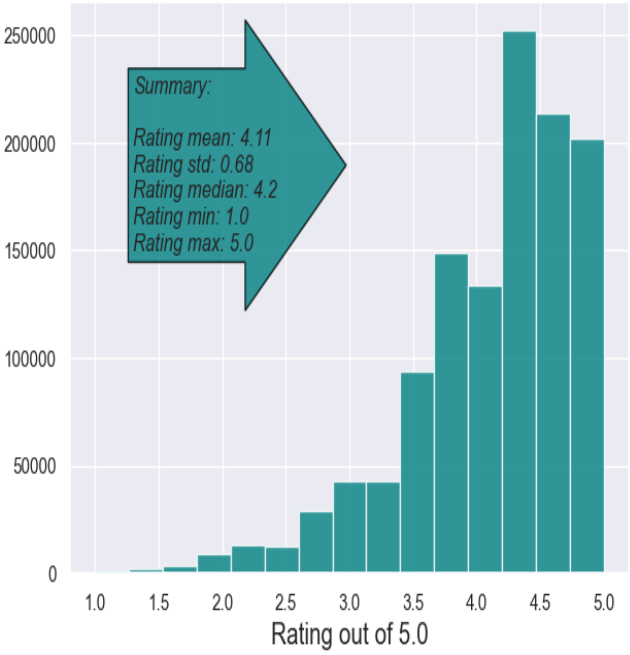

Figure 5.3.1.13: Rating Without Zero

According to figure 5.3.1.12, there are many more apps with no rating. So, I omit them for better visuals, that's shown in figure number 5.3.1.13. Histogram (figure number 5.3.1.13) shows that the majority of the apps are rated between ~4.2 and 4.5. It is also surprising to see so many 5-star ratings.

```
print('Apps with rating equal or lower than 1 star: ',len(data[data['Rating'] <= 1.0]))
```

```
Apps with rating equal or lower than 1 star:  1035085
```

```
print('Apps with rating equal or lower than 2 star: ',len(data[data['Rating'] <= 2.0]))
```

```
Apps with rating equal or lower than 2 star:  1050543
```

RATING COUNT

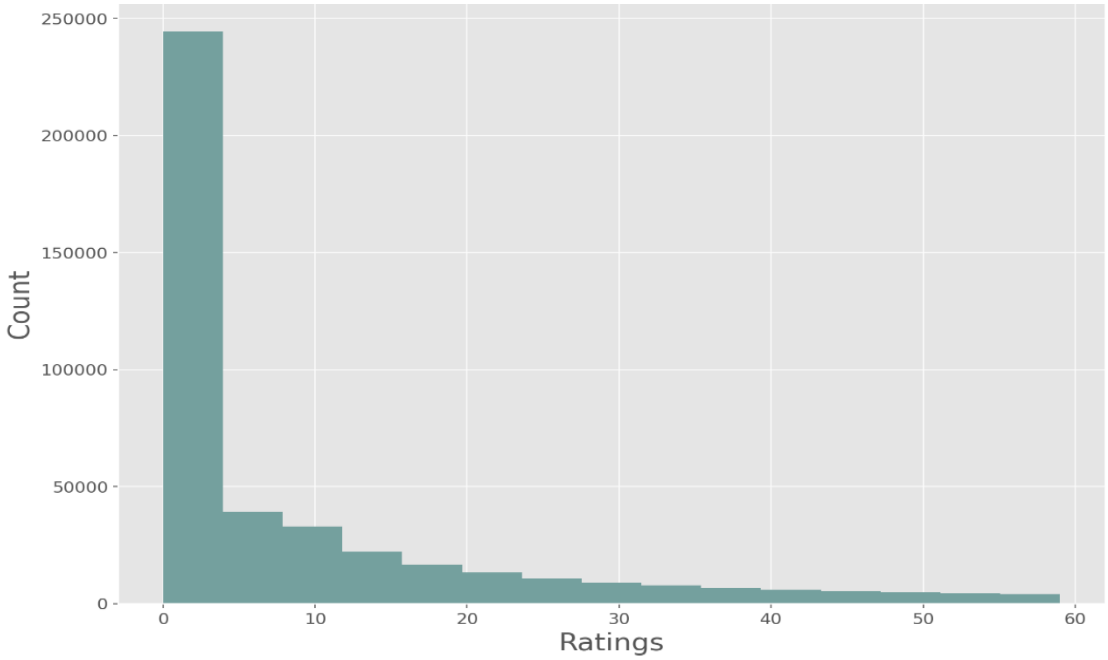

Figure 5.3.1.14: Rating Count

This histogram tells us that about a quarter of the apps have no more than 5 ratings. This shows how competitive the mobile market is. Only a small proportion of apps can go as popular as the ones with thousands of ratings. let's explore the apps that have more than 1 million ratings:

```
over_mln = data[data['Rating Count'] > 1e6]
over_mln.shape
```

```
(786, 16)
```

Figure 5.3.1.15: Rating Count of More than One Million

Out of the initial 2 million apps, only 786 have over 1 million ratings.

AVERAGE INSTALLS

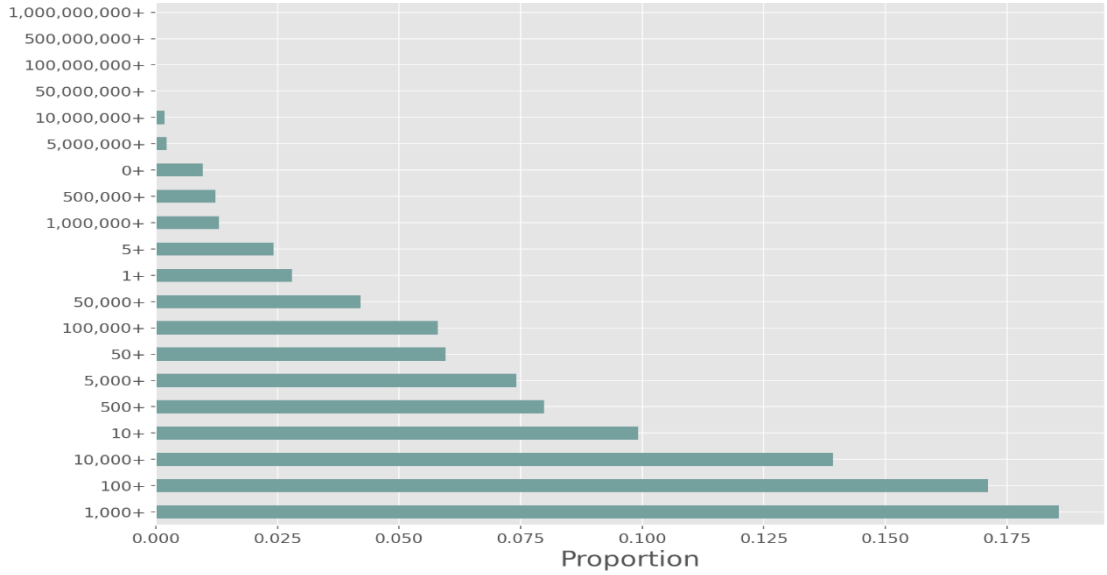

Figure 5.3.1.16: Proportion of Install Categories

The plot shows that the vast majority of installs are between 10 and 10k installs. Most of the app installs are relatively small compared to the maximum, which is around 1e9 (a billion). However, it's interesting to see that there are quite a lot of zero installs.

```
upper, lower = iqr_fence(data['Average_Installs(+)'])
print('Upper Fence:', upper)
print('Lower Fence:', lower)
```

```
Upper Fence: 1000000.0
Lower Fence: 0.0
```

We can see that the lower fence is 0 installs, while the higher fence is 1 million installs.

```
print('Total apps', len(data))
no_installs = [1e9, 1e8, 1e7, 1e6, 1e5, 1e4, 1e3, 1e2, 1e1]
for n in no_installs:
    print('Number of apps with less than ' + str(n) + ' installs:', len(data.loc[data['Average_Installs(+)']<n]))
```

```
Total apps 9620
Number of apps with less than 1000000000.0 installs: 9600
Number of apps with less than 100000000.0 installs: 9388
Number of apps with less than 10000000.0 installs: 8249
Number of apps with less than 1000000.0 installs: 6225
Number of apps with less than 100000.0 installs: 4610
Number of apps with less than 10000.0 installs: 3118
Number of apps with less than 1000.0 installs: 1775
Number of apps with less than 100.0 installs: 740
Number of apps with less than 10.0 installs: 156
```

Figure 5.3.1.17: Installs (Less than)

We can see that there are quite few a apps with just 10 or 100 Installs.

PRICE

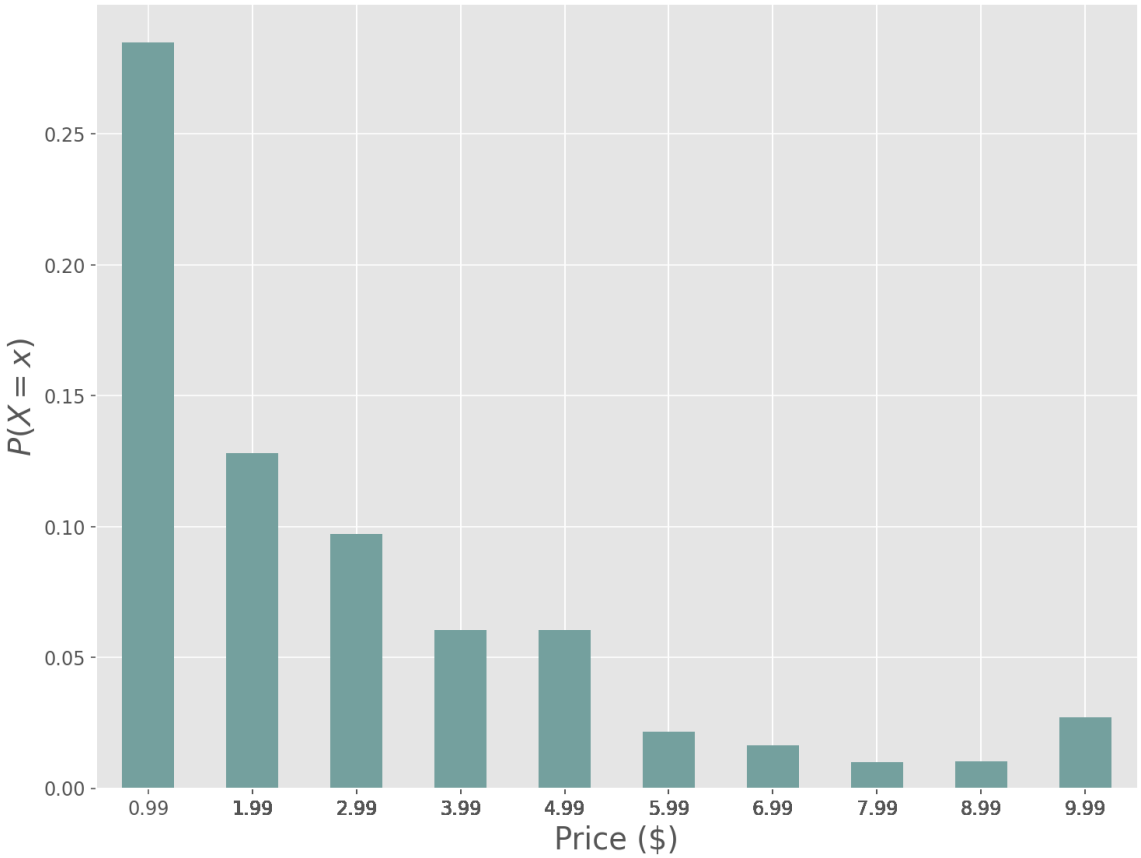

Figure 5.3.1.18: PMF of Price of Paid

It is clear that most apps cost about a dollar and almost all the apps are quite cheap.

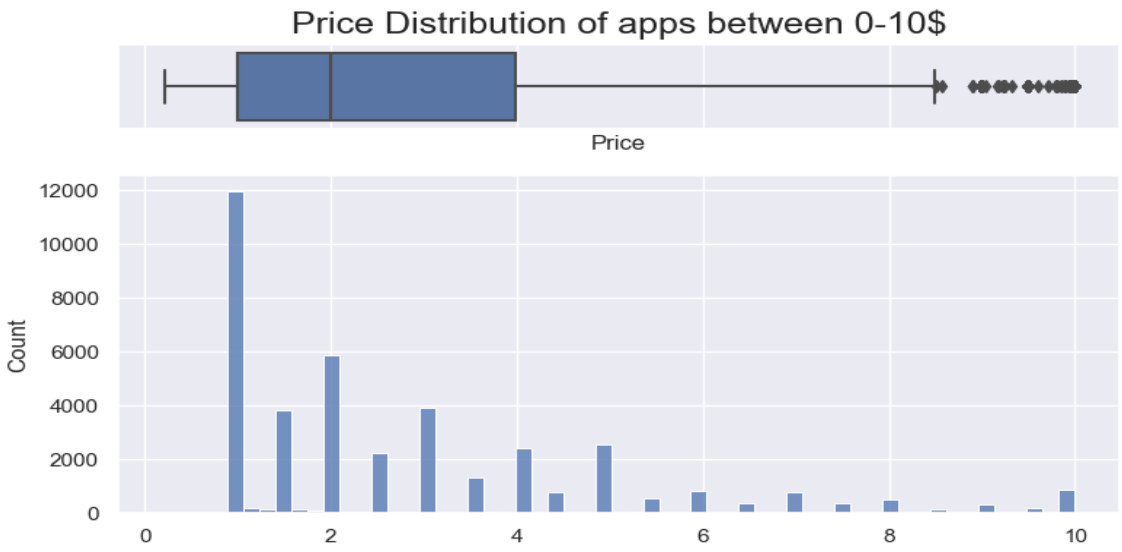

Figure 5.3.1.19: Price Distribution of apps between 0-10$Apps

Most of the apps are between 1 to 2$ and it is a right skewed distribution.

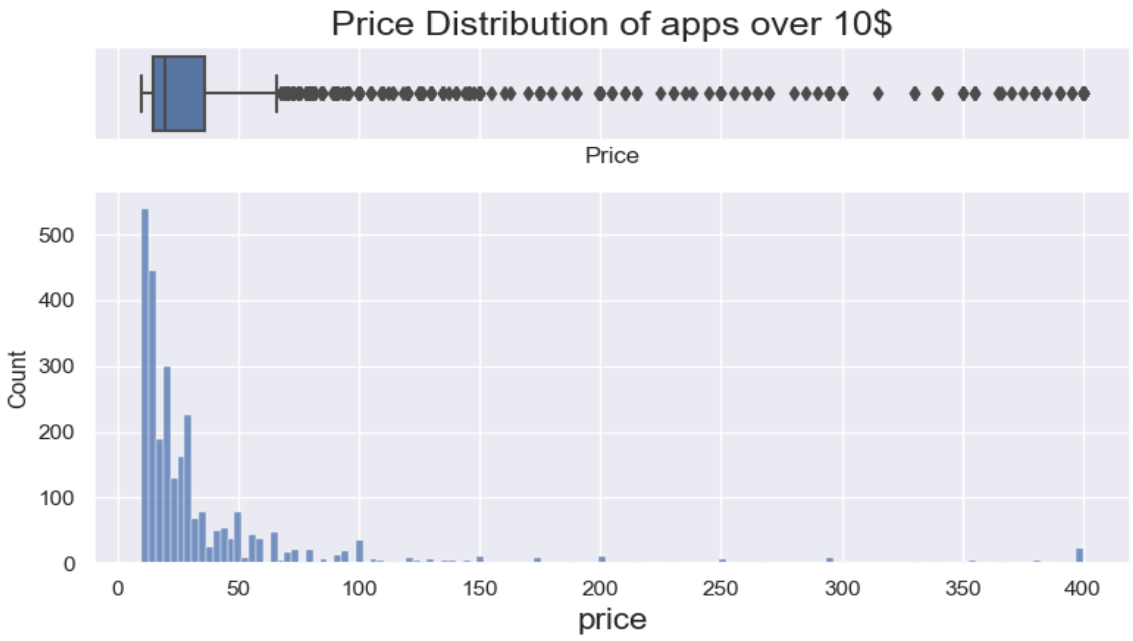

Figure 5.3.1.20: Price Distribution of Apps Over 10$

There are some apps with a price close to 400$.

SIZE

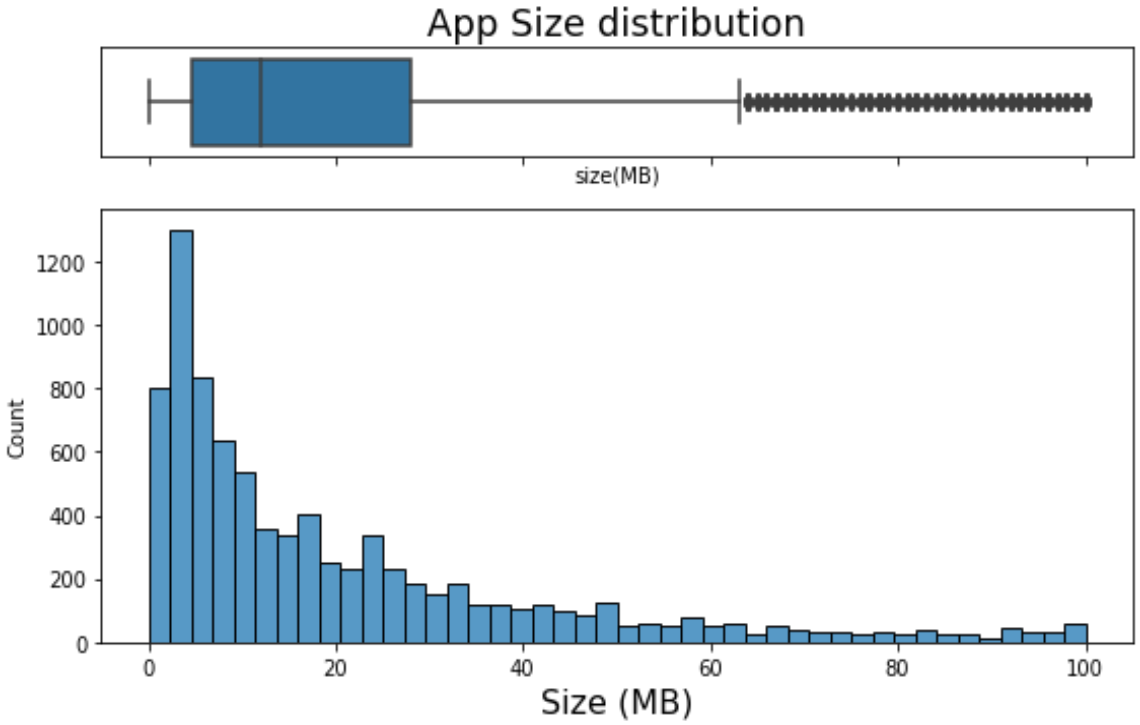

Figure 5.3.1.21: App Size Distribution

The distribution of app size is a right skewed long tail. Most of the apps have sizes lower than 20, and some apps have sizes around 100MB.

5.3.2 Bivariate Exploration

APP NAME VS INSTALLS

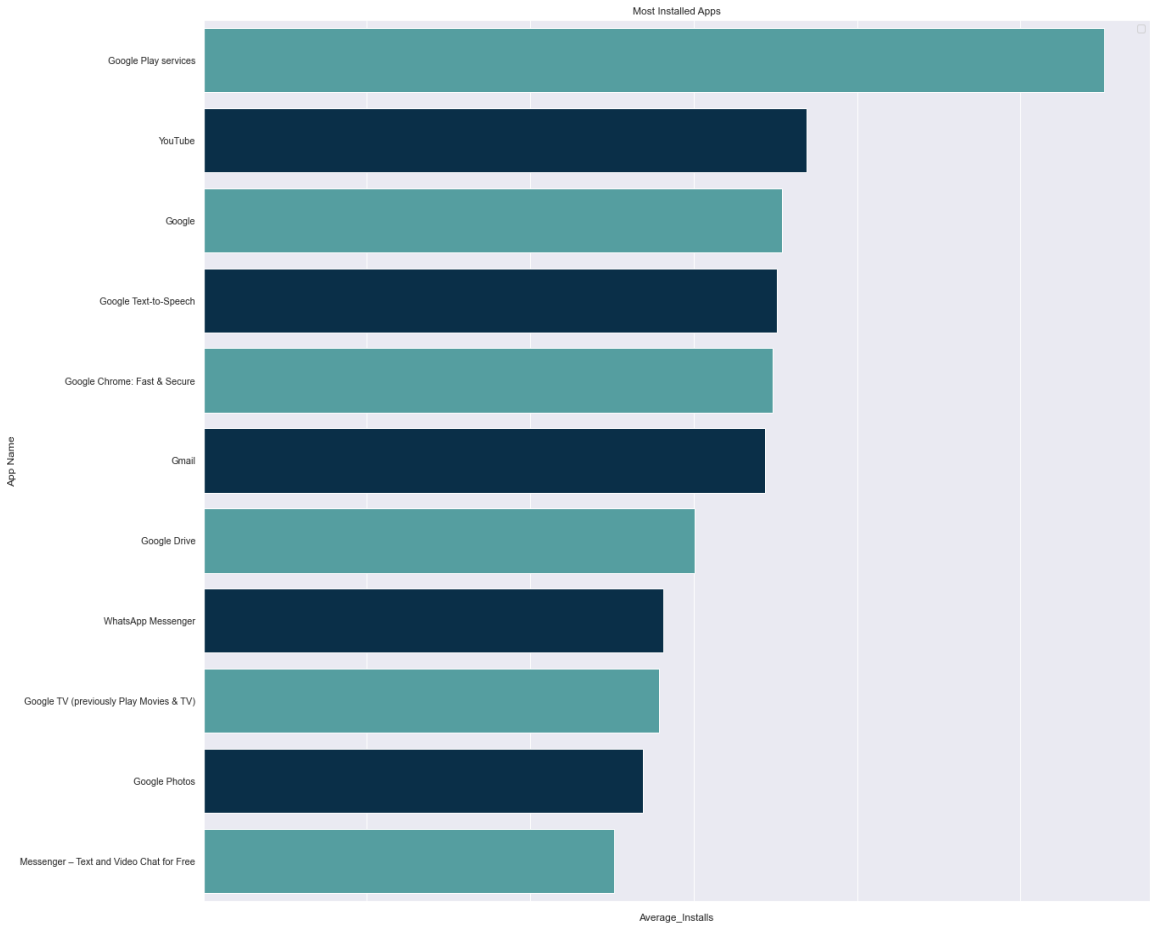

Figure 5.3.2.1: Most Installed Apps

The most installed app in the Google Play Store is Google Play services. This isn't surpris ing, since most Android phones have this app pre-installed. Almost all the other apps in t he list are also pre-installed apps.

CATEGORY VS. RATING

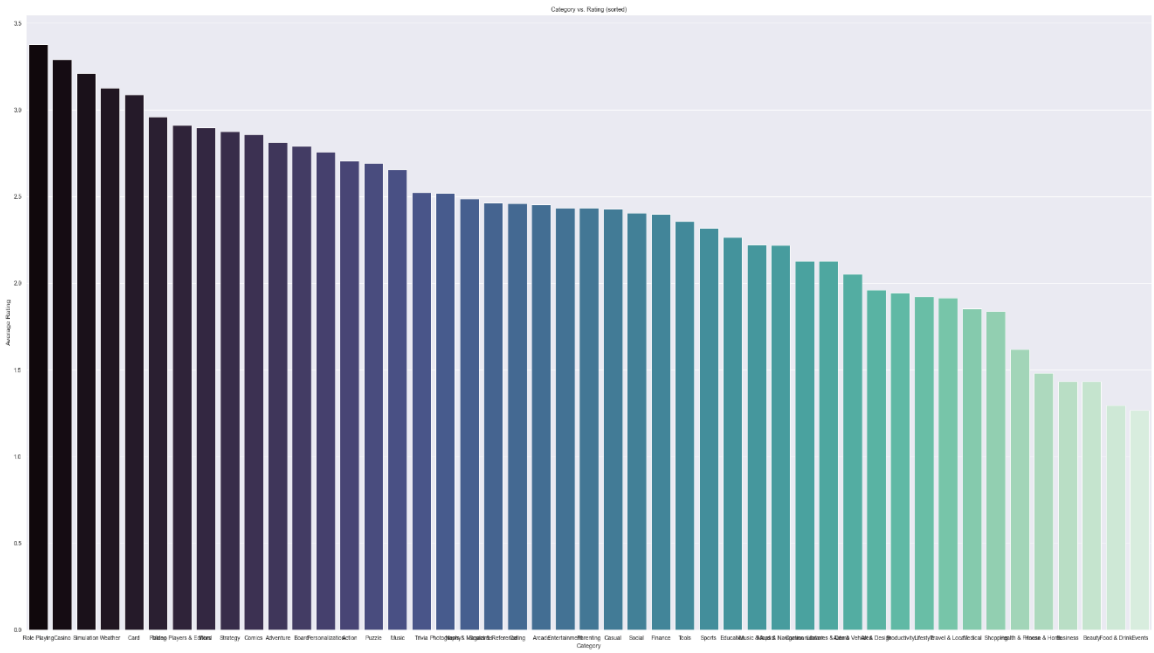

Figure 5.3.2.2: Category vs. Rating

The data are distributed quite uniformly. This, given some thought, should make sense, as each game within a category is rated based on games of that same category. Taking the average of all the scores for a particular category results in approximately the rating. However, the top 3 categories are based on games, which are Role Playing, Casino and Simulation. Least number of ratings are from Event and Food & Drink categories.

Highest Mean Rating Categories.

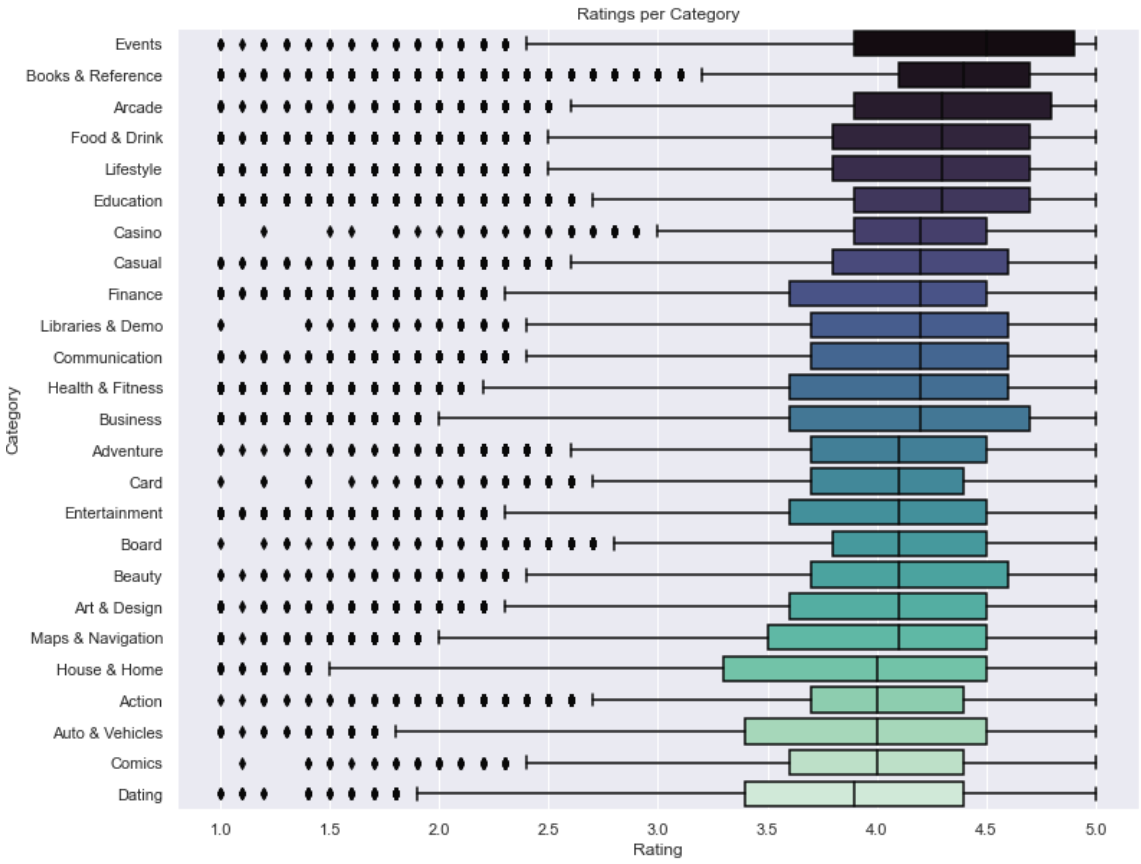

Figure 5.3.2.3: Top 25 Highest Mean Rating Categories

One thing to point out is the fact that the Event category holds the highest mean rating, while having the least number of installs on the Play Store (check the Figure 5.3.2.4). Actually, if you were to think about it, that would make perfect sense, since there are a smaller number of apps that result in a regression of the rating to the mean, causing slightly 'inflated' ratings for categories with lesser apps (although this does not apply to most categories)

CATEGORY VS INSTALLS

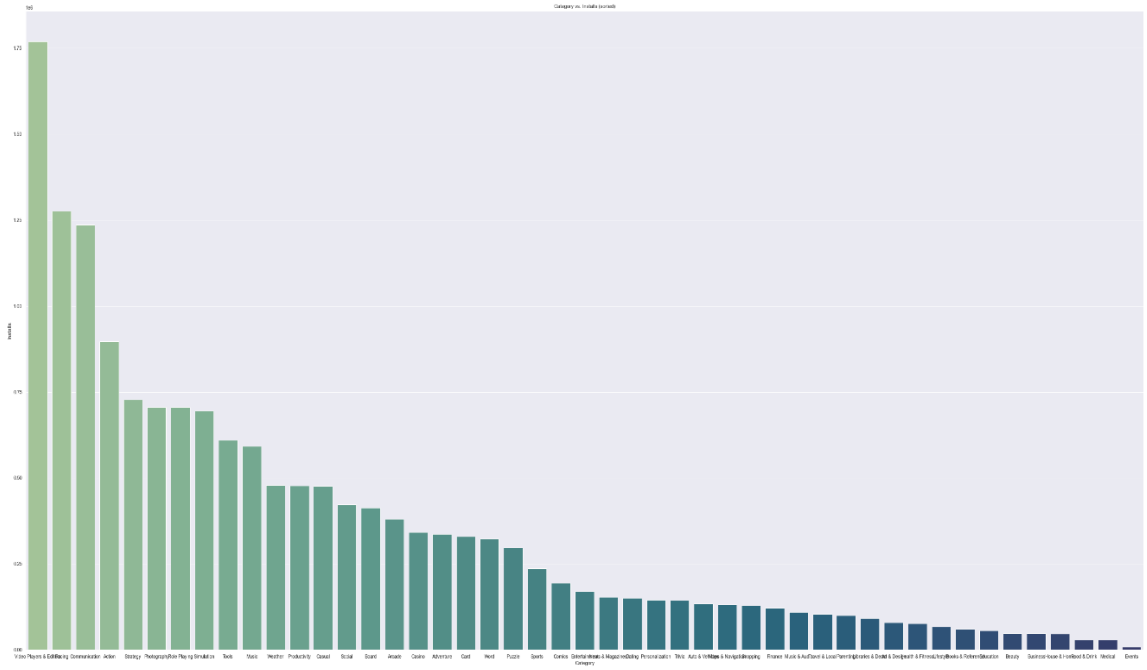

Figure 5.3.2.4: Category vs Installs

The data are visibly heavily skewed. The top categories are unsurprising; Video Players & Editors, Racing, Communication and Action apps are widely downloaded and used quite often. Education, which had the most apps but was dethroned by other categories and now ranks in the bottom 10 for installation. Tools category which previously occupied the 3rd spot in app number has now fallen by six places to the 9th spot. Hence, looking at categories by apps might be misleading for a developer. A developer wanting to attract a large user base should pick a category based on the number of installs and not by the number of apps in the Play Store.

Turning our eyes on the other side of the spectrum, we see high ranking apps in terms of rating ranked rather high in terms of installs. Given some thought, this seemingly strange occurrence feels much more understandable. Theoretically, the installs for a particular category should have little correlation with the ratings of that particular category. To conceive this counterintuitive thought, consider Figure 5.2.3.5

RATING VS INSTALLS

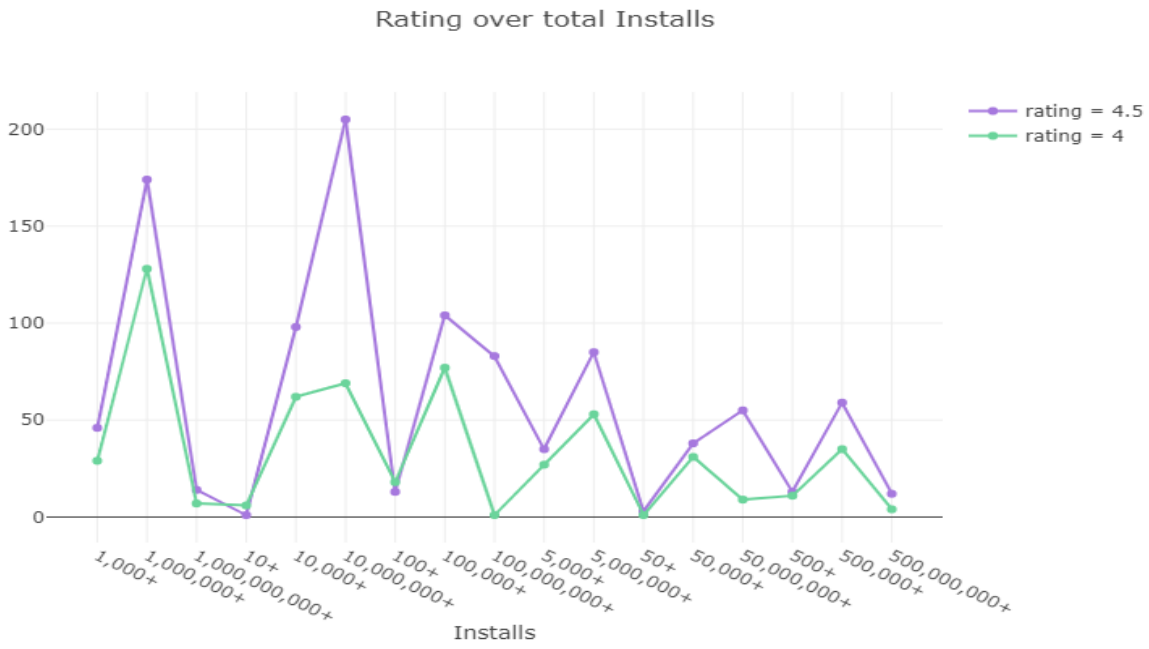

Figure 5.3.2.5: Install vs Rating

Most important things about apps are their ratings. We didn't even see the app's whose ratings are less than 3.5. Before downloading the app first, we see the rating of the app if the app rating is more than 4 then we say that yeah this is a good category of app then after we see other attribute of apps. Very few people see the reviews of apps. My personal observation is

that some of the good people who are passionate about the technology and have enough time give the review of the app. That's why review of the app directly affect to the number of installs of an app.

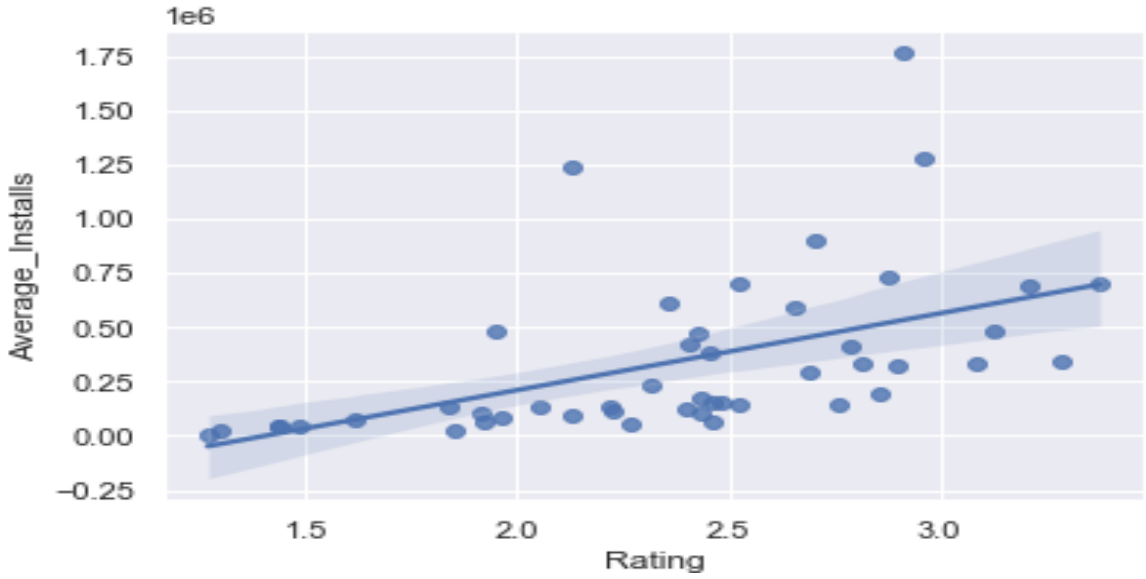

Figure 5.3.2.6: Relation Between Rating and Installs

The magnitude of the slope of this line is close to 0, but the slope of the line is negative. This section highlights the counterintuitive and inversely proportional relationship between rating and installs of apps from the Google Play Store. More specifically, a higher rating generally corresponds with a higher number of installs.

INSTALL VS APP NAME (WrordCount)

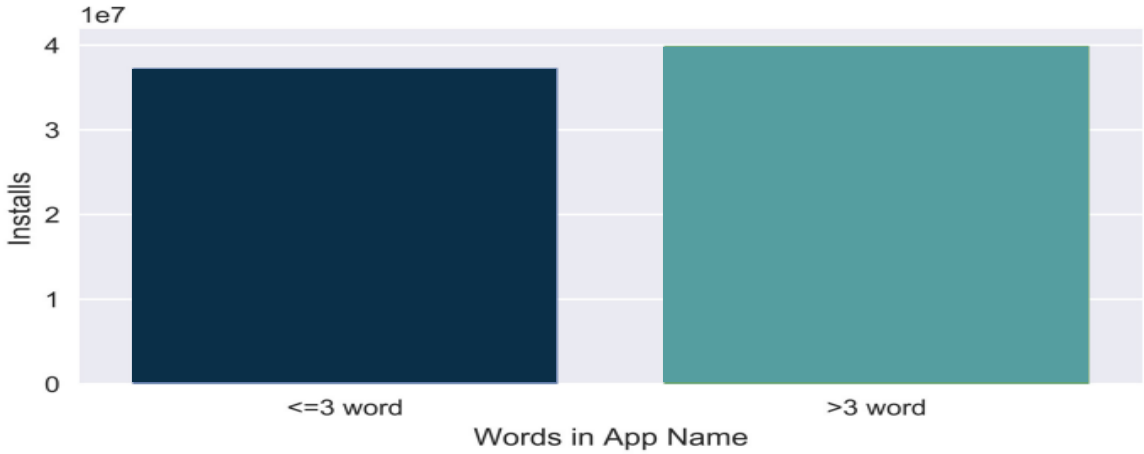

Figure 5.3.2.7: Less Than or Equal to Three Words

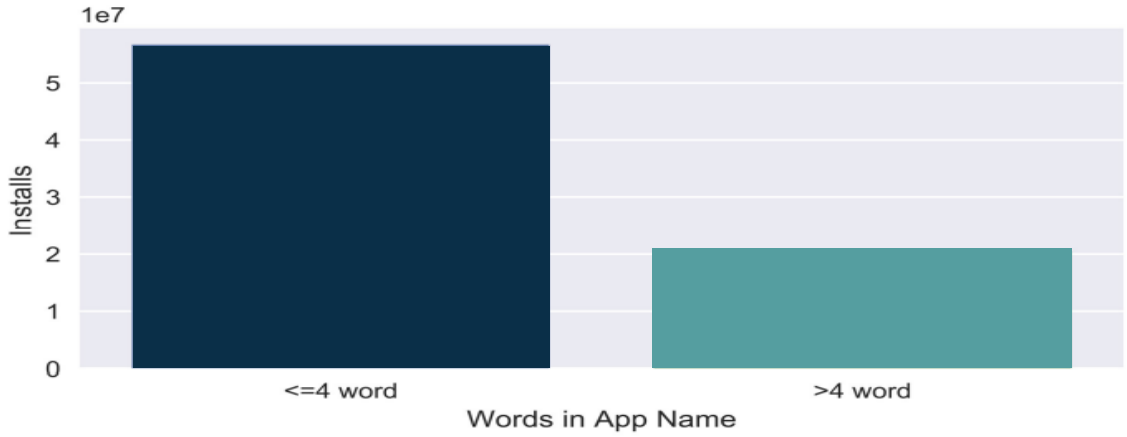

Figure 5.3.2.8: Less Than or Equal to Four Words

I decided to see how apps with a threshold on the number of words used in their names fared against those without any. Figure 5.3.2.7 shows that apps with less than or equal to three words in their names account for almost half the number of installs. When the threshold is increased to 4 words, we see the margin increasing significantly in Figure 5.3.2.8 It seems that most of the installs in the Play Store are contributed by apps having less than or equal to 4 words in their names. Hence, it is better that developers use a concise word or words to name their app.

TYPE(FREE) VS. INSTALL

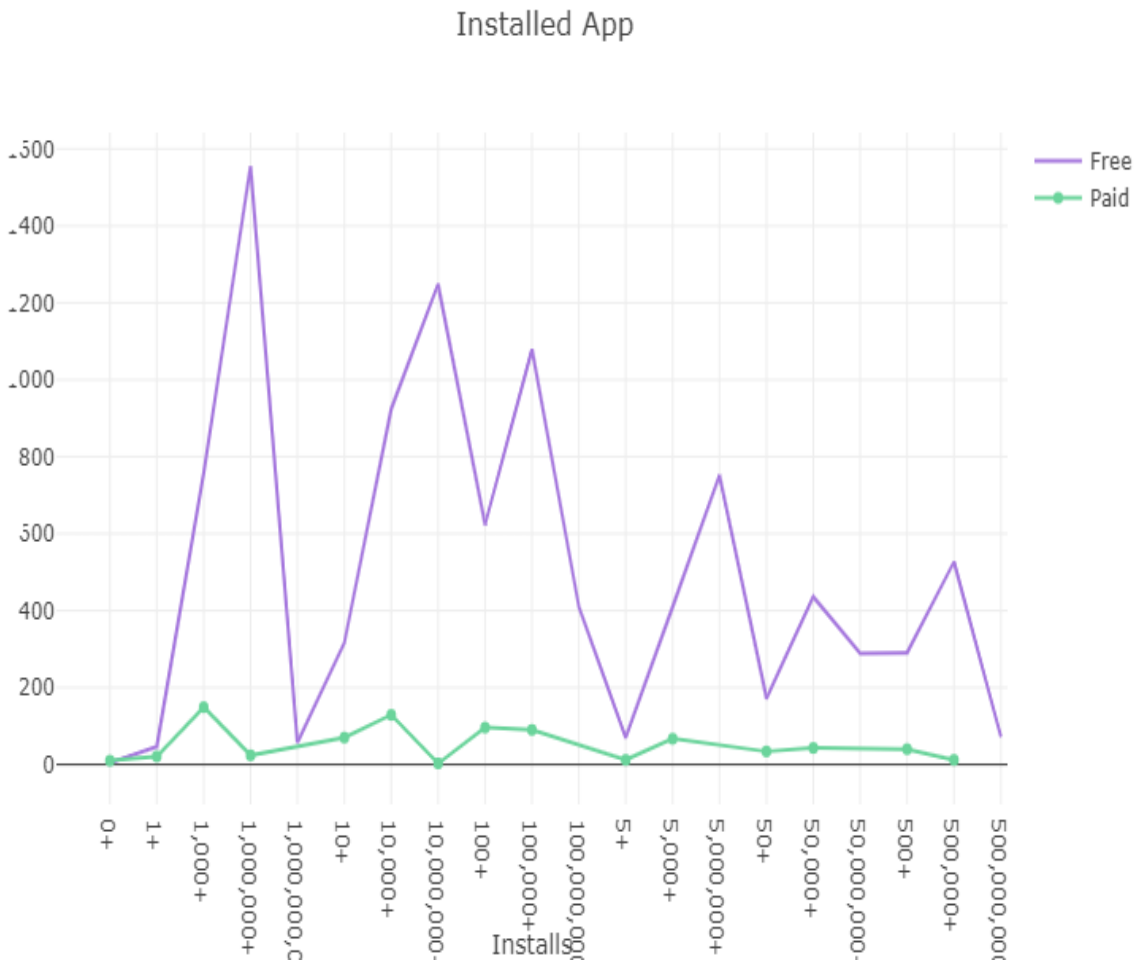

Figure 5.3.2.9: Type vs Installs

As expected, free apps are much more likely to be installed than paid apps. The difference between the two categories is incredibly high. I expected a patent difference, but not one that was pronounced.

*False = Not Free (Paid)

*True = Free

Distribution of Install

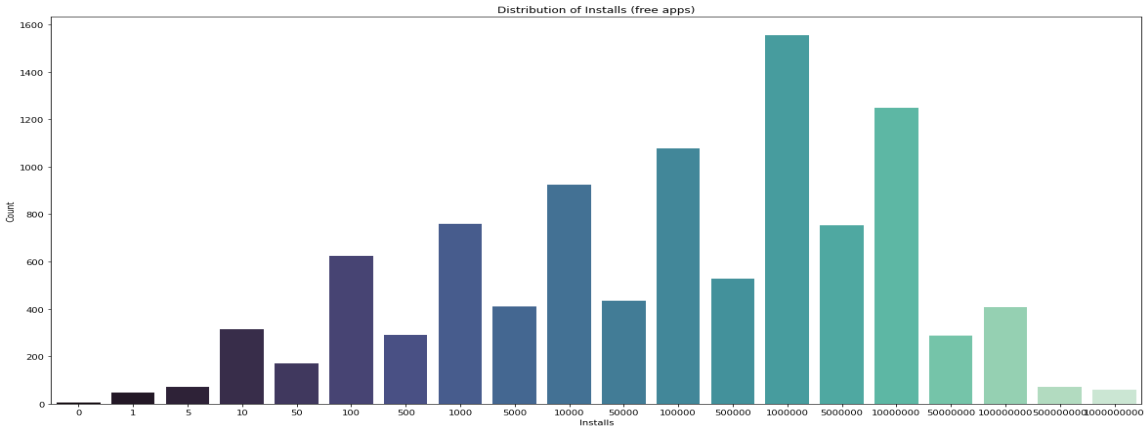

Figure 5.3.2.10: Distribution of Installs (free apps)

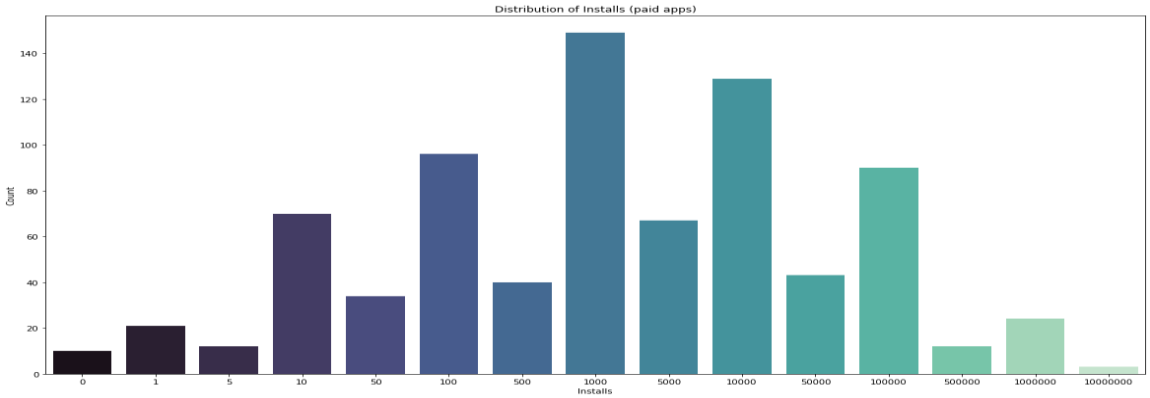

Figure 5.3.2.11: Distribution of Installs (paid apps)

Although the figure 5.3.2.10 gets slightly unintelligible toward the end, identifying the general distribution is still feasible. We can see that a large number of free apps receive a great deal of installs. By comparing the distributions of the two figures above, we can see that free apps receive a wider array and generally greater number of installs than paid apps.

In this section, we found that the number of installs of paid apps overall paled in comparison to the number of installs of free apps overall.

FREE VS. RATING

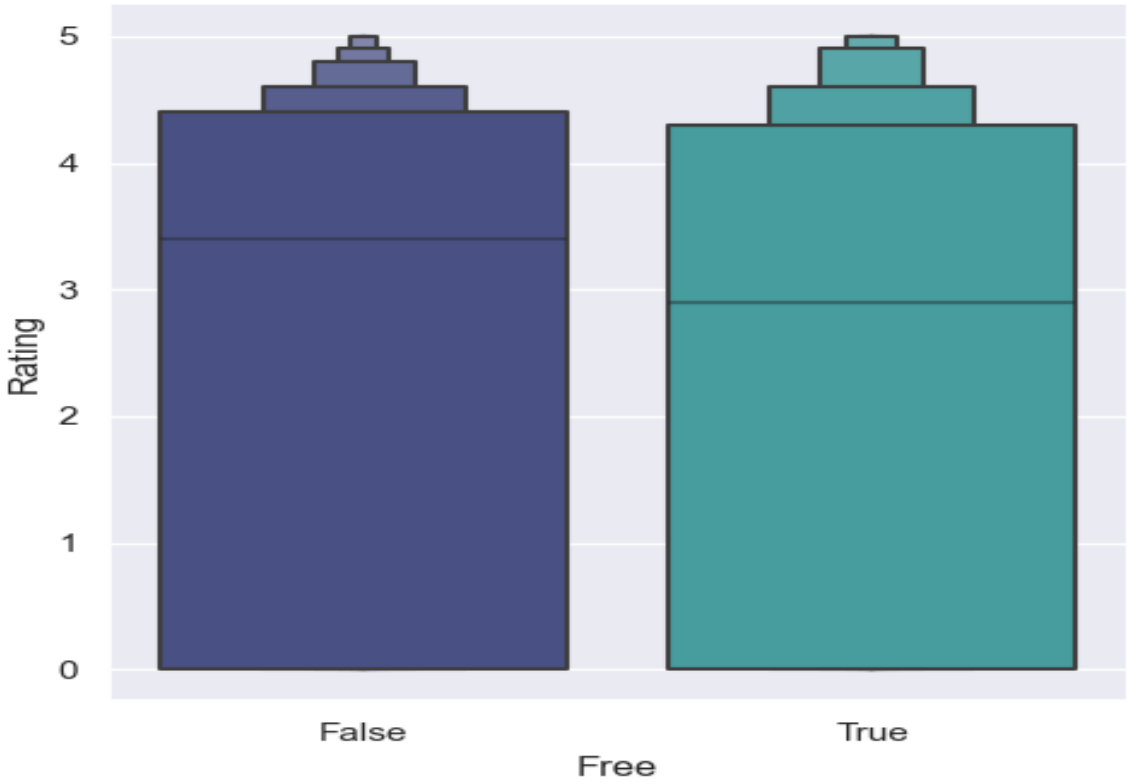

Figure 5.3.2.12: Type vs Installs

As visible in the graph above, the average rating for each type is approximately the same. While this result may seem counterintuitive to some, a reasonable process is unknowingly employed to judge most apps accurately. The average user generally rates based on sheer user experience, not on the price or the number of installations.

However, two anomalous cases conveniently counteract each other. Both cases result from experiences with paid apps yet entail different emotions. The first of these cases results from pungent dissatisfaction. If a user felt an app was not delivering the appropriate amount of quality for its price, the user, not overall disappointed with the quality of the app, would write a negative review and leave a rating toward the lower end of the spectrum. In contrast,

the other case is the product of abounding appreciation for a particular application. In this scenario, the user feels the app is too good for its price, thus giving it a higher rating. These two contrasting cases effectively cancel each other out, resulting in a moderately accurate average rating for the paid apps.

Distribution of Rating

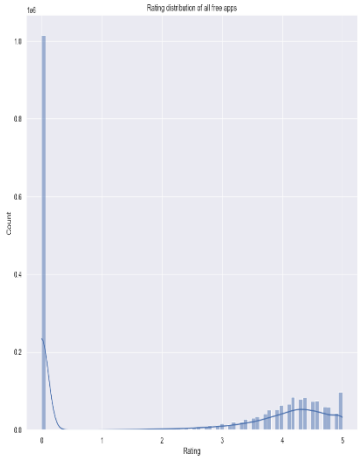

Figure 5.3.2.13: All Free

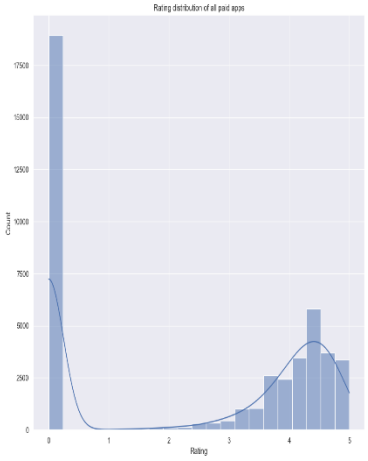

Figure 5.3.2.14: All Paid

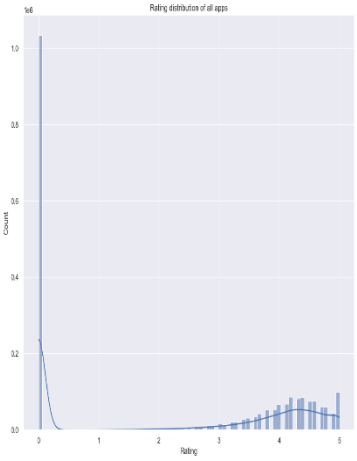

Figure 5.3.2.15: All

The Figures above have no distinctive features. Although there are some slight discrepancies in the height of the KDE curve, the graphs are nonetheless similar.

On a bit of a side note, these two graphs underscore why Install and Rating were picked as the two metrics for this exploratory data analysis. While they may initially seem tightly correlated and futile in conjunction, the preceding graphs differentiate them, each providing benefits to the exploration. The two previous graphs provided a high-level overview of the relationship between Type (Free/Paid) and one of the scoring metrics

CONTENT RATING VS. RATING

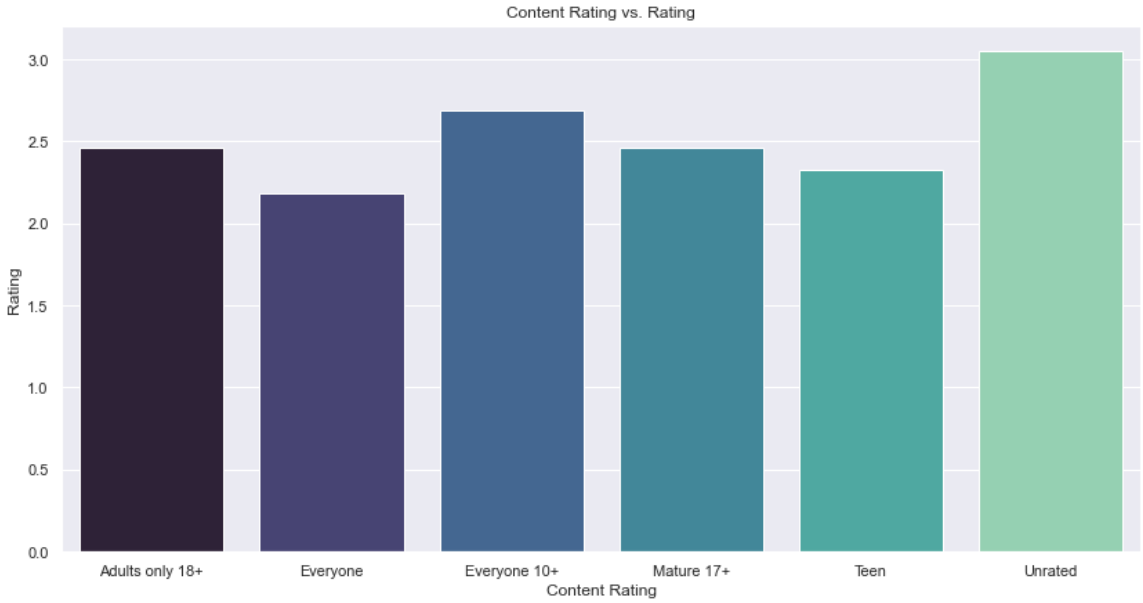

Figure 5.3.2.16: Content Rating vs. Rating

As visible in the graphic above, the ratings of the various content ratings are roughly the s ame. As said in earlier explanations of the lack of nuance and variation in average ratings, each show is evaluated based on the quality of said app with respect to the content rating. It is unreasonable to judge an app from the teen content rating based on apps from the eve ryone content rating.

Let’s take a look at the distribution of ratings and installs within each content rating.

Distribution of Ratings

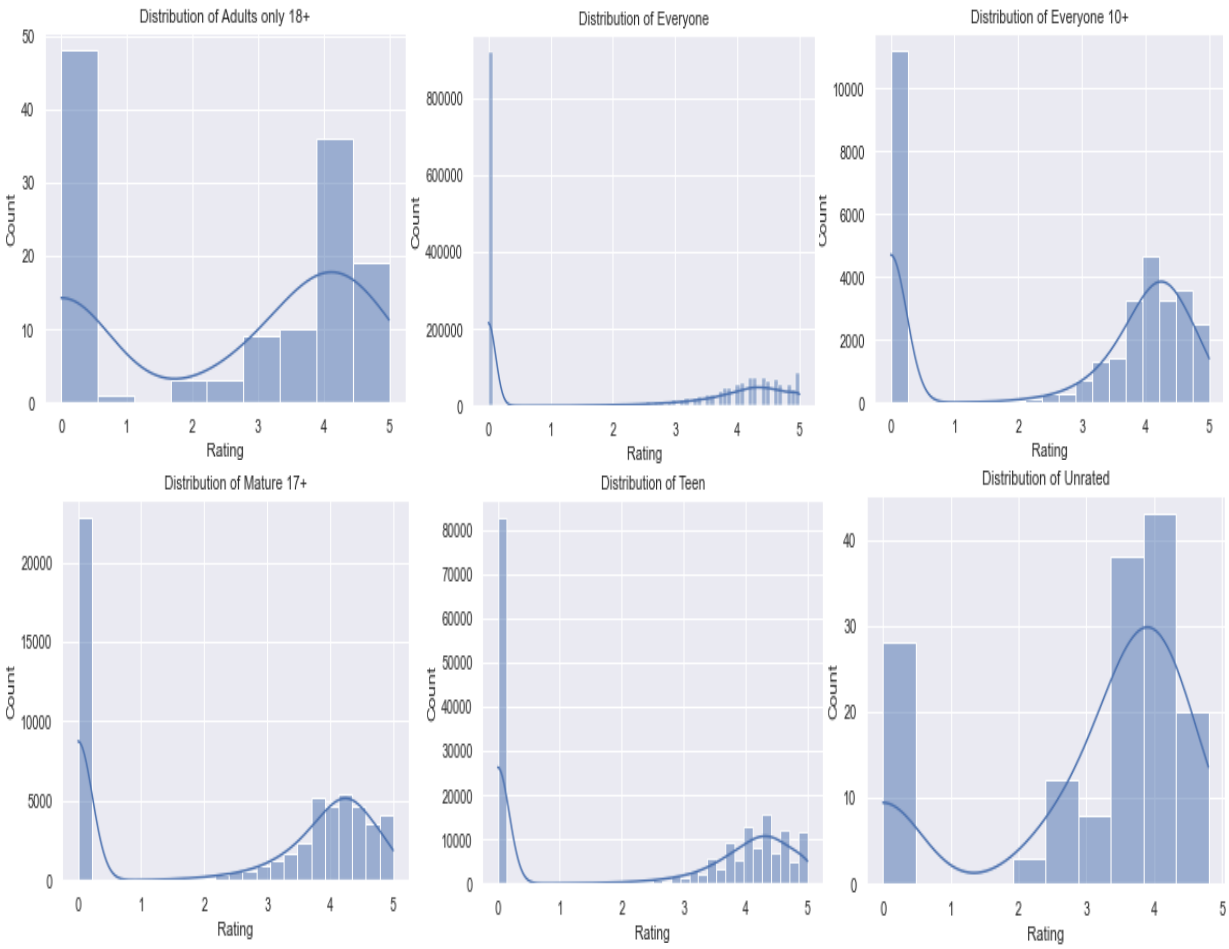

Figure 5.3.2.17: Distribution of Ratings

Nevertheless, we can see that the distribution is slightly skewed all the while being mostly normal. The content rating of an app does not necessarily affect the rating it receives.

CONTENT RATING VS INSTALLS

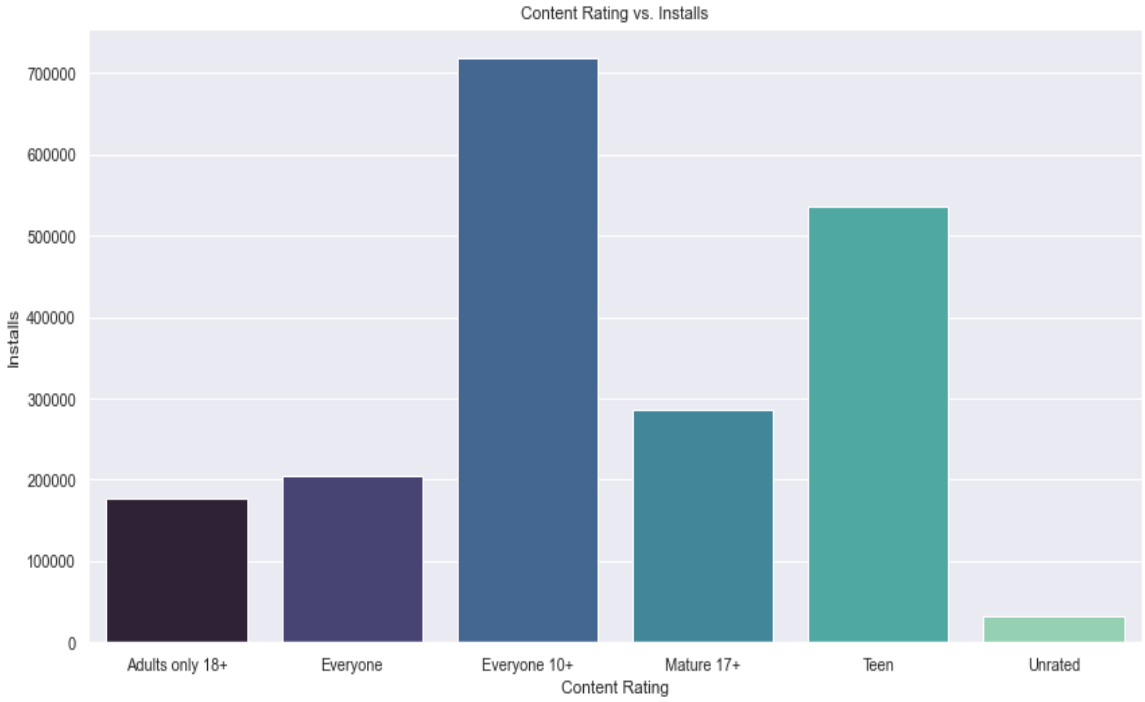

Figure 5.3.2.18: Content Rating vs. Installs

Contrary to the distribution of the ratings, the distribution of the installs has more discrepancies and variation. This is simply due to the popularity of each content rating.

Adult and Unrated apps are virtually invisible in Figure 5.3.1.4 and very low in Figure 5.2.3.18 Apps that fall under the Content Rating, 'Mature 17+', 'Teens' and 'Everyone 10+' has the highest chance to be downloaded. For prospective app developers and publishers, it will be worthwhile to invest in apps falling under these Content Ratings.

Distribution of Installs

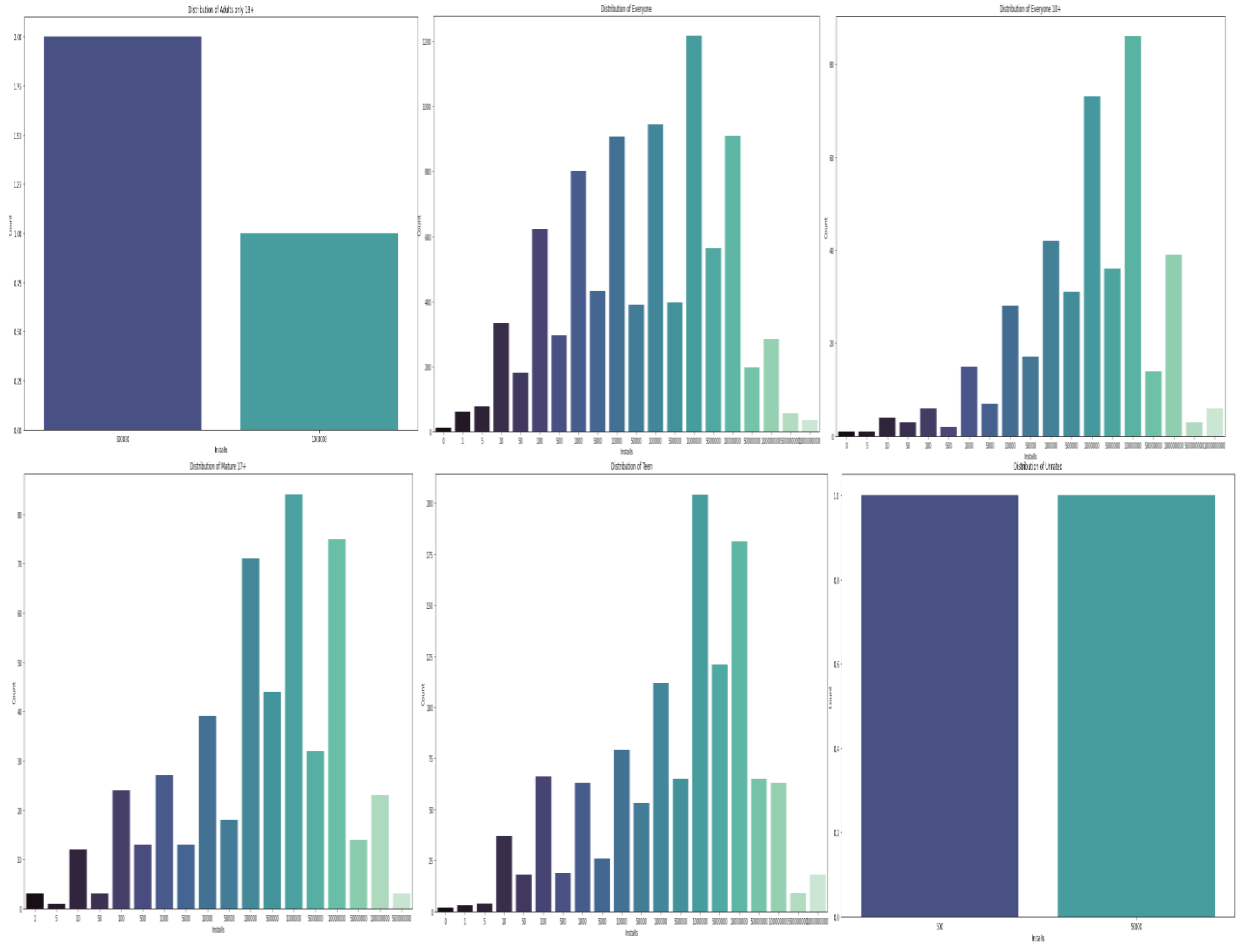

Figure 5.3.2.19: Distribution of Installs

Ignoring the awkward graphs in the Unrated and Adults only 18+ content ratings, we can deduce that the distribution of installs is fairly normal. The content rating can drastically influence the number of installs a particular app receives.

CONTENT RATING VS TYPE(FREE)

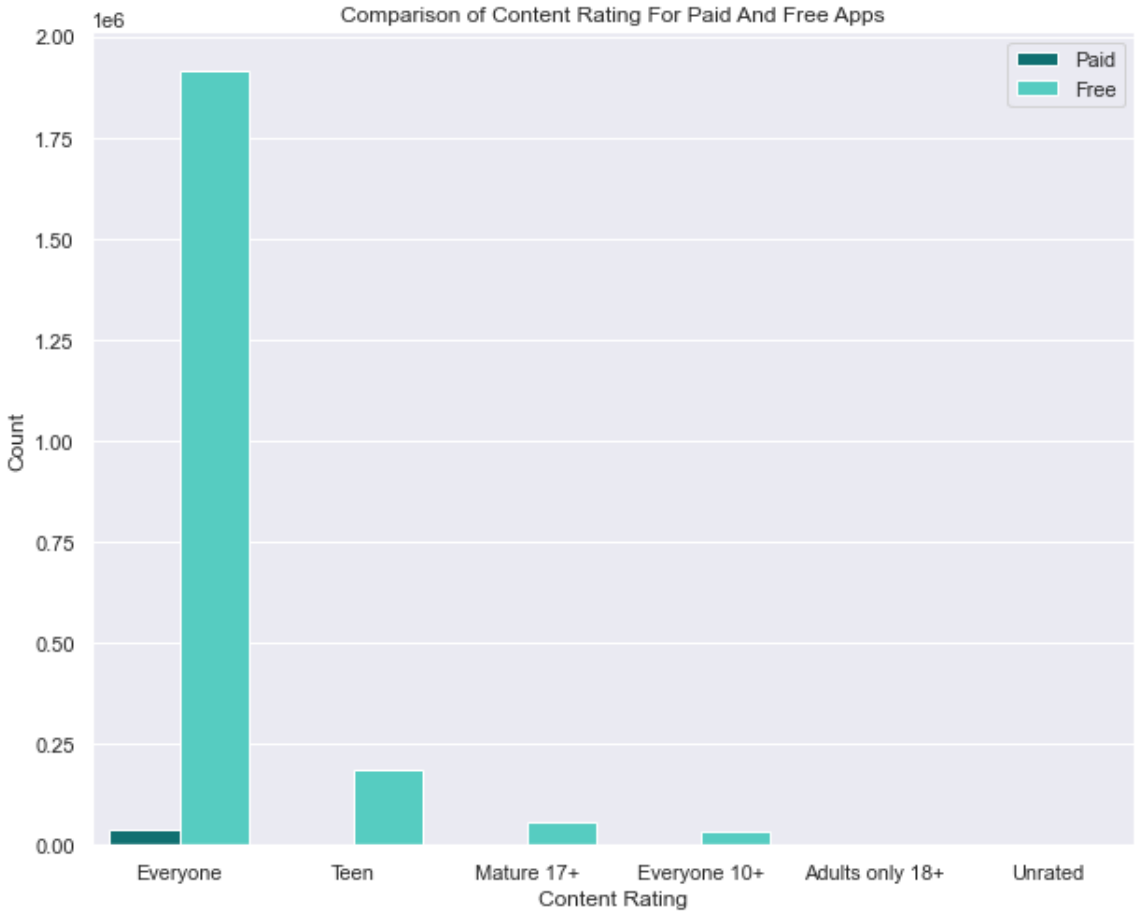

Figure 5.3.2.20: Content Rating for Paid and Free apps

We can clearly see that there was no paid app in other categories except EVERYONE.

MOST INSTALLED GAME PER CATEGORY

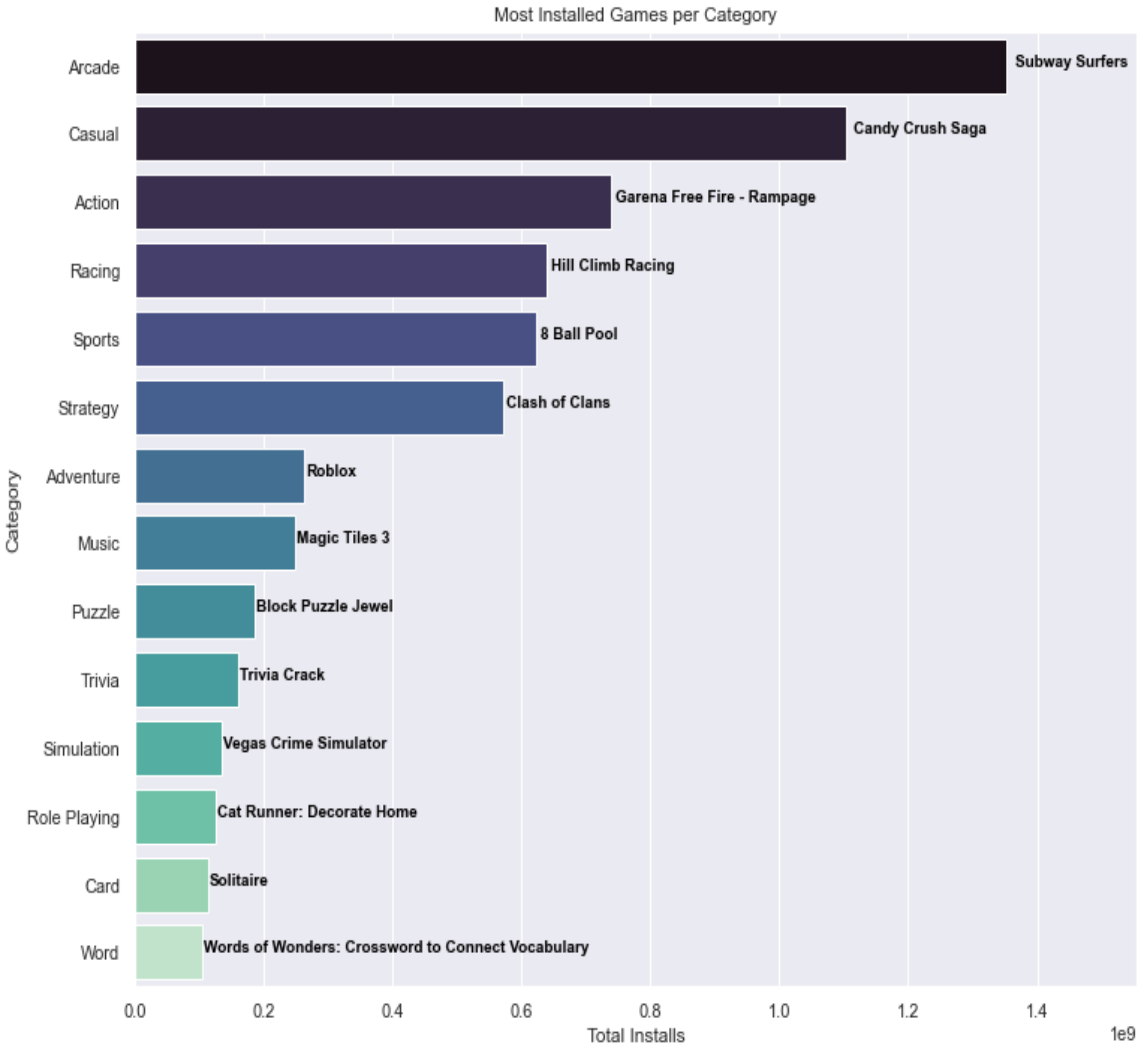

Figure 5.3.2.21: Most Installed Game per Category

Most installed Game is Subway Surfers and it's from Arcade category. Second highest installed game from Casual category which is Candy Crush Saga.

EDITORS CHOICE EFFECT ON APP INSTALLS AND RATINGS.

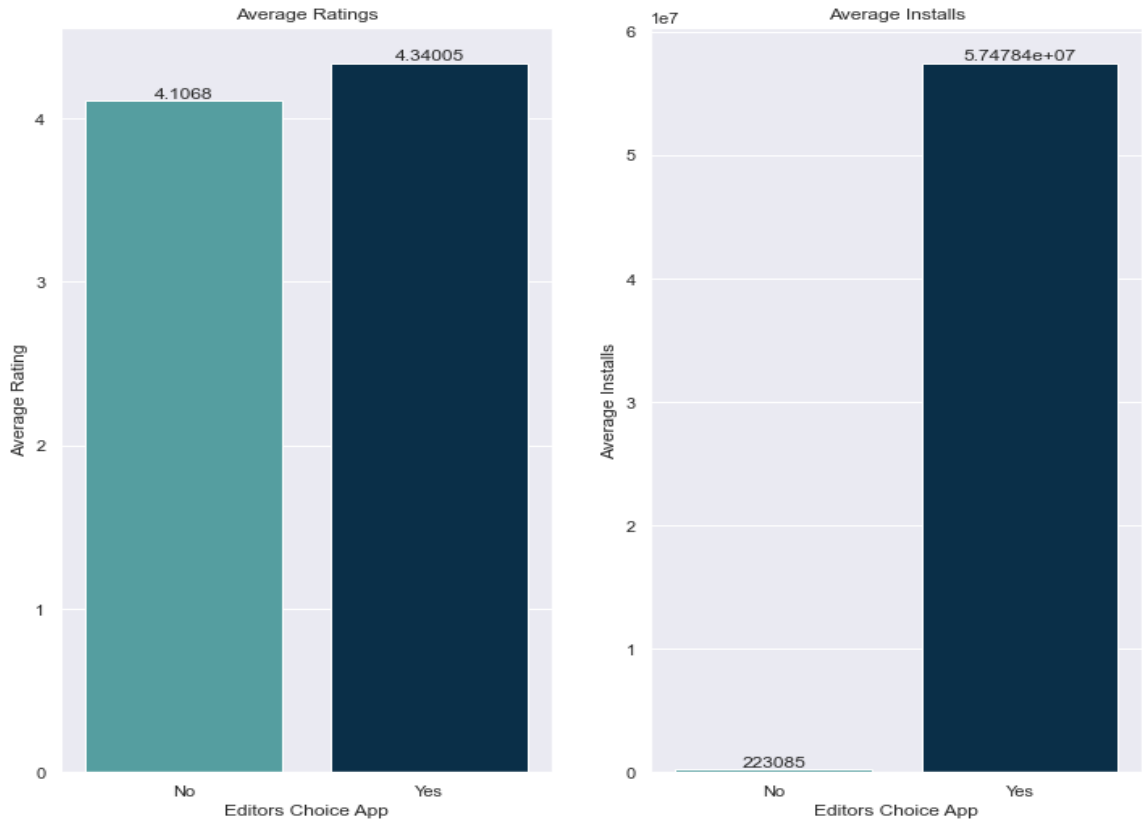

Figure 5.3.2.22: Editors Choice effect on app Installs and Ratings

We can infer that those installations play a larger role for an app to be labeled as an Editor's Choice app rather than mean ratings. This makes sense, since an app with two 5-star ratings (mean rating of 5) and an app with five thousand ratings that average to around 4.5 doesn't make the first app better simply because it has a higher rating!

If we take a different perspective, the 'Editors Choice' label doesn't seem to affect the mean ratings for apps (both categories hover around the 4.1-4.3 rating mark), while it has **very profound effects** towards average installations (55 million installations vs 223 thousand installations).

CATEGORY VS PRICE

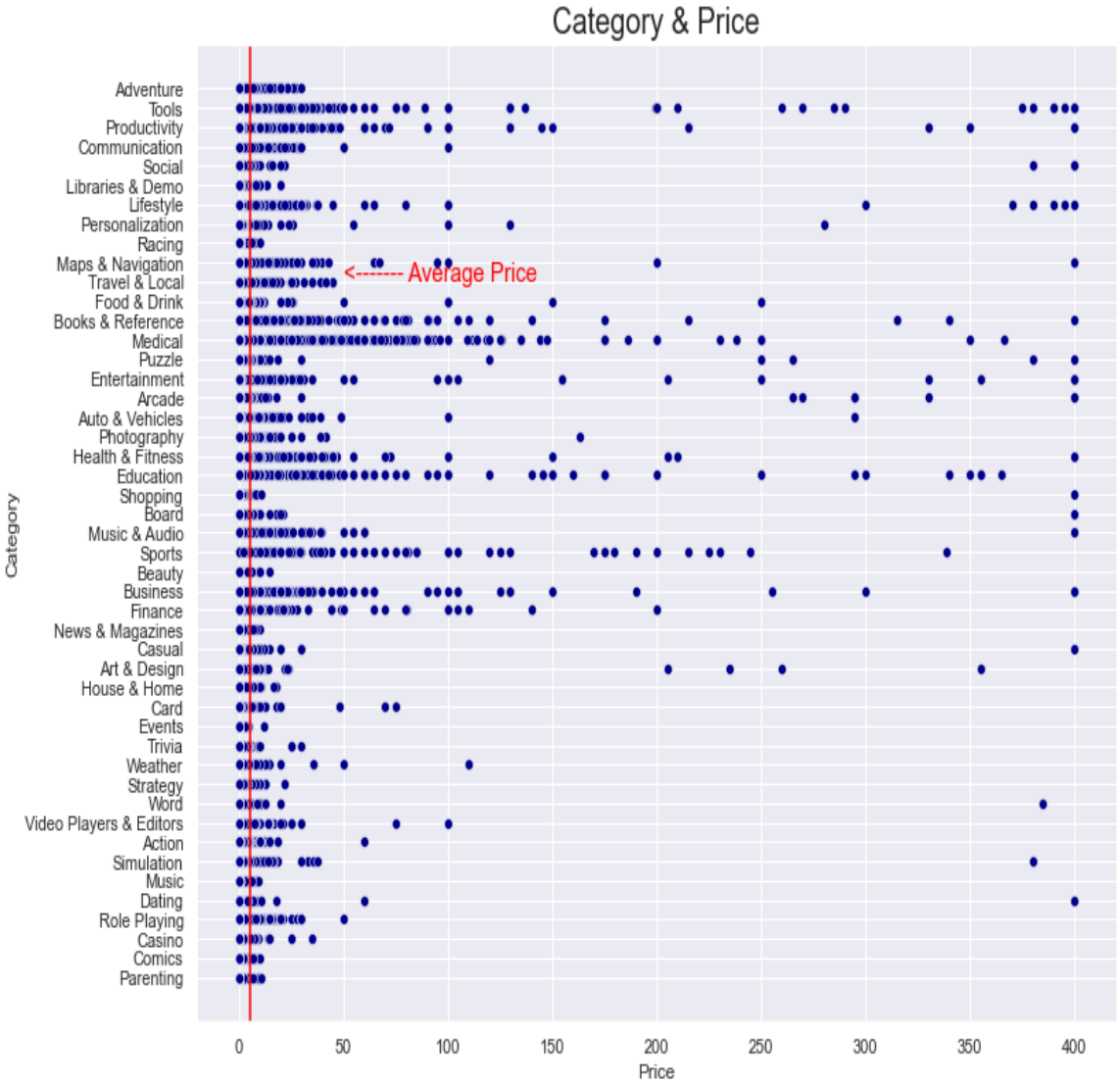

Figure 5.3.2.23: Price over Category

Very few categories have very few expensive apps and almost every app in each category is under 50 dollars. And the average price is around 1 dollar. Conclusion is we would prefer making a free app for Play Market rather than a paid app.

CATEGORY VS TYPE(FREE)

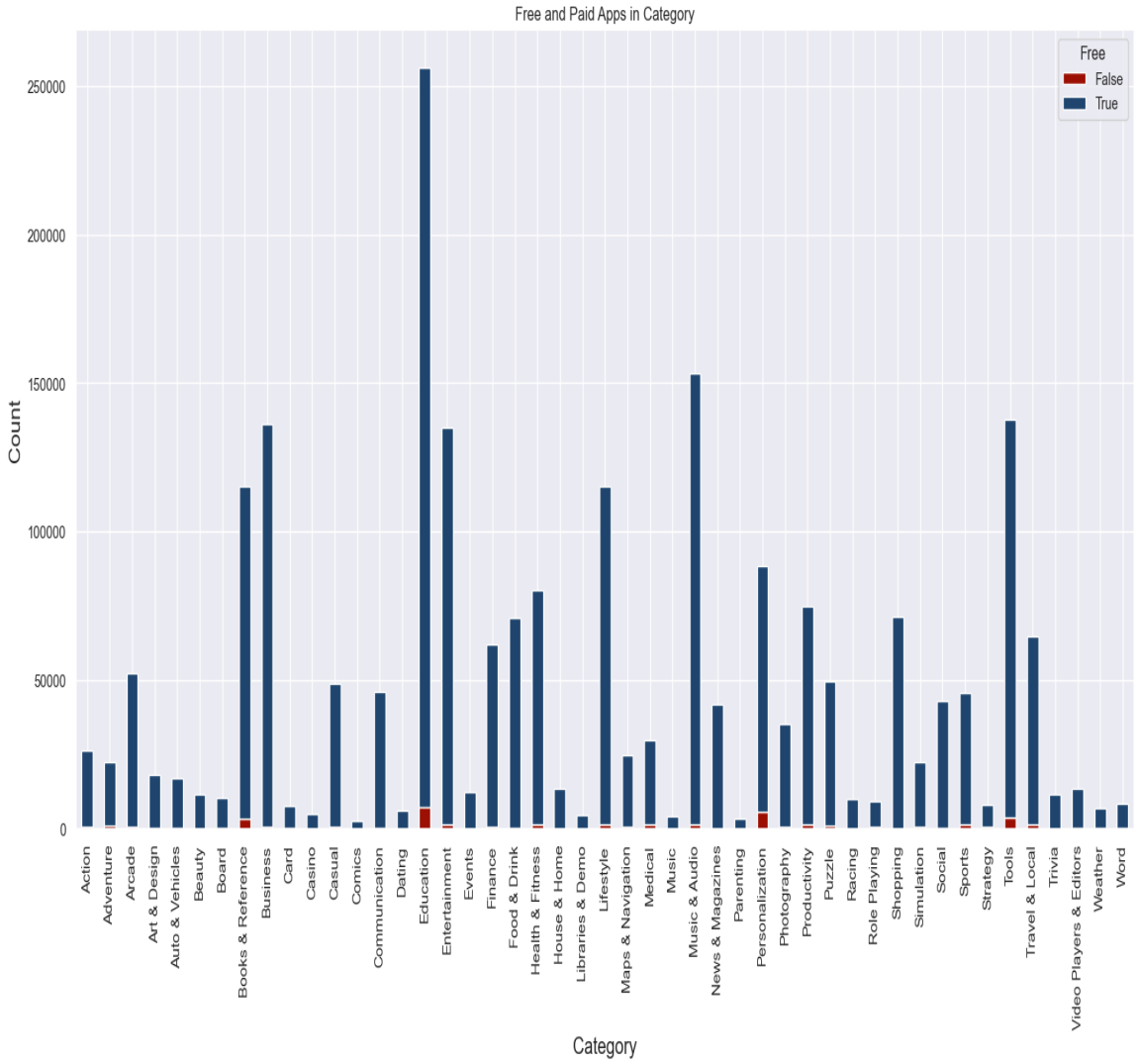

Figure 5.3.2.24: Type over Category

While Figure 5.3.1.7 and Figure 5.3.2.20 confirmed the dominance of ‘free’ apps, we can see that Tools, Personalization and Education – all have a small fraction of paid apps, but Entertainment is a different case with barely any visible slice in the bar chart in Figure 5.3.2.24 This could likely be due to the fact that entertainment apps are free to download, but require a subscription to use, such as Netflix, Amazon Prime Video, Crunchyroll, etc.

SIZE VS RATING

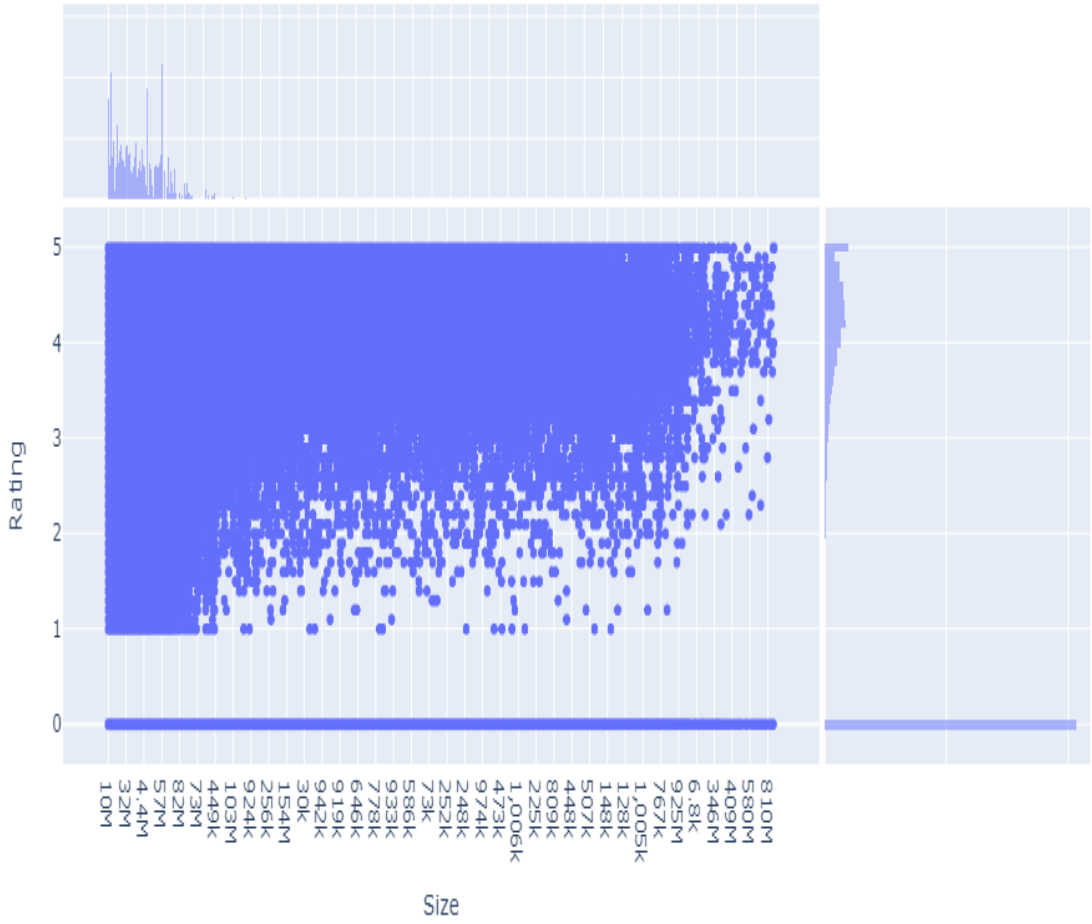

Figure 5.3.2.25: Size over Rating

This graph suggests that apps with a lower size have more varied ratings. Meaning that as we increase in size, the ratings tend to fall less and less

SIZE VS INSTALLS

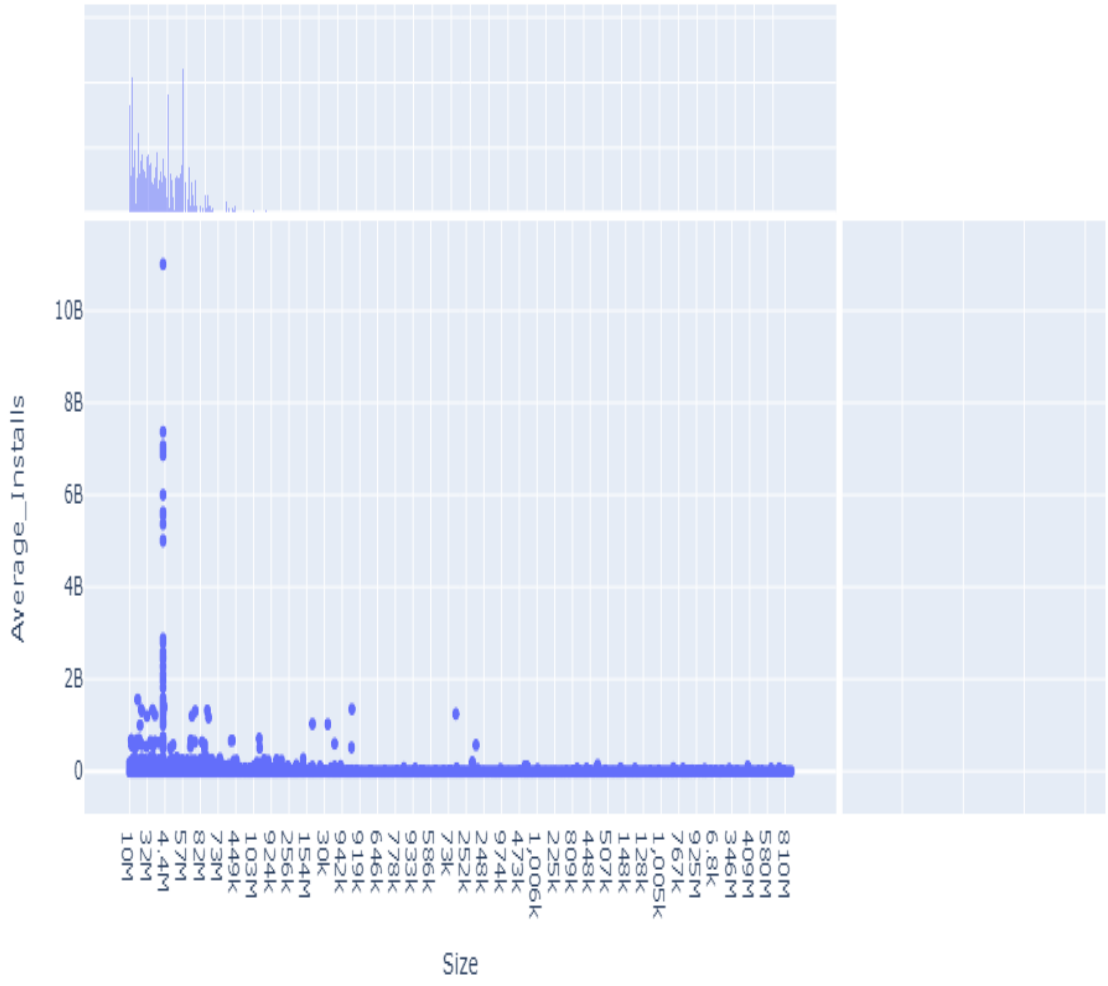

Figure 5.3.2.26: Size over Installs

From the above plot we can see that size impacts the number of installations. Applications with large size are less installed by the user. The apps with smaller sizes have more chance to be downloaded.

5.3.3 Multivariate Exploration

THE TRENDS OF THE TOP 3 MOST INSTALLED GAME GENRES

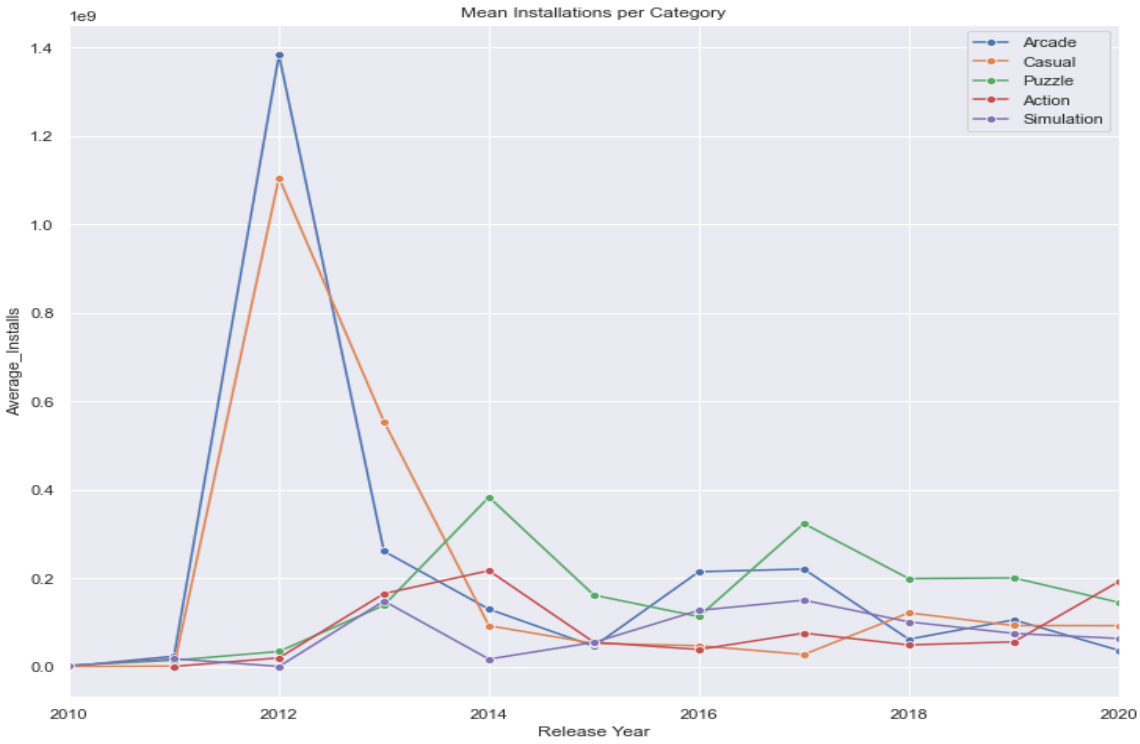

Figure 5.3.3.1: Mean Installations of Games per Category

For the past few years, the 'Puzzle' category has consistently had the highest average installations among the highly rated games, with 2020 being the lone exception, as action games saw a massive boost in average installation numbers this year. Games like Player Unknown's Battlegrounds (PUBG Mobile) and Garena's Free Fire battle royale games were a huge driving force, also depicted from the visualization above of the most installed games per category.

So, if you're a developer looking to dive in the mobile gaming market, these categories are a good bet, but puzzle games seem to consistently outperform other categories.

COMPARISON OF RATING FOR BOTH FREE AND PAID APPS ACORDING TO THE CATEGORIES

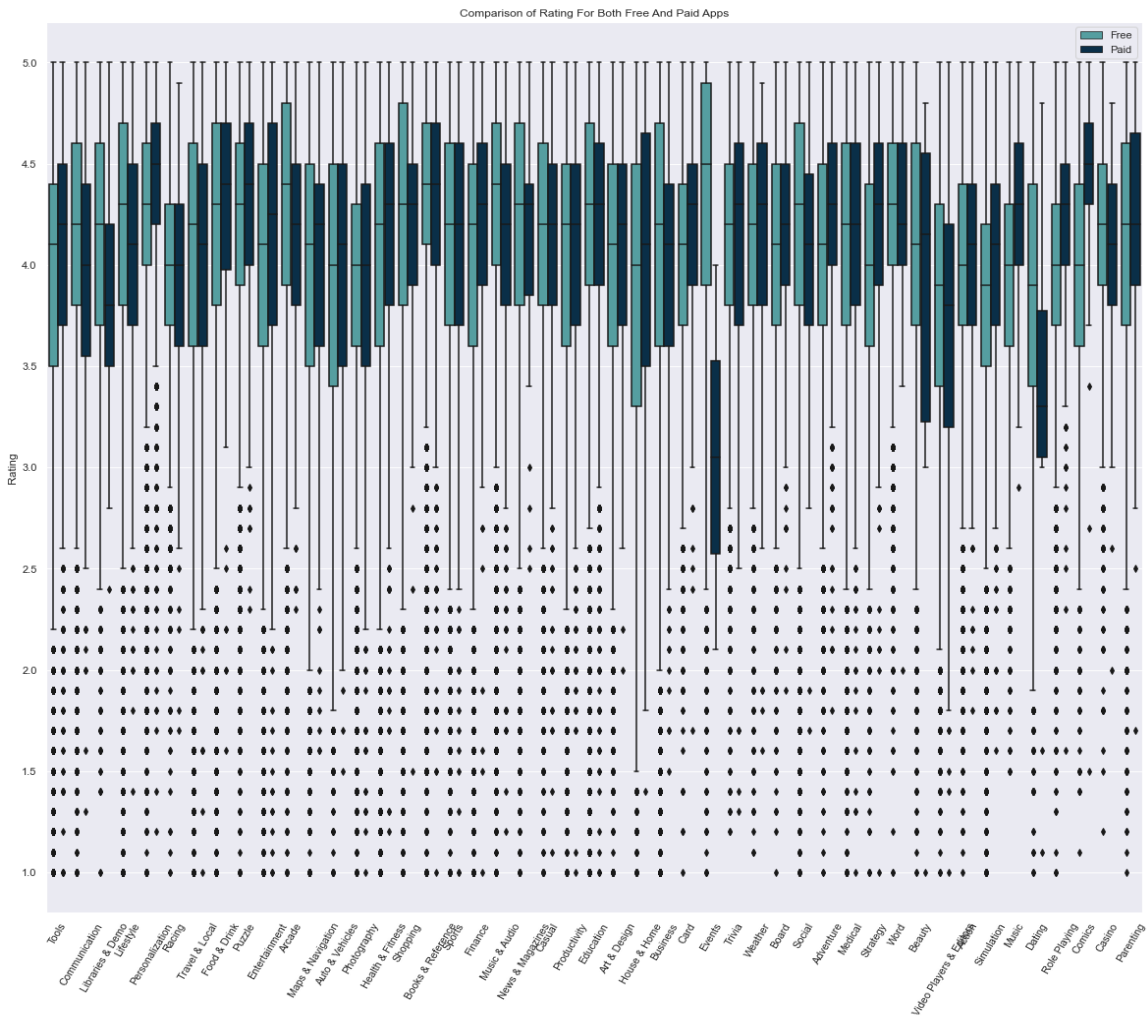

Figure 5.3.3.2: Comparison of Rating for Both Free and Paid Apps According to Category

All categories are skewed to the left. There are obvious outliers in every category. From this plot we can see that in most categories, paid apps have higher ratings than free apps. In particular it is also interesting to notice that free apps have lots of outlier values compared to paid apps.

MAINTENANCE OF APP VS. RATING GROUPING BY FREE OR PAID

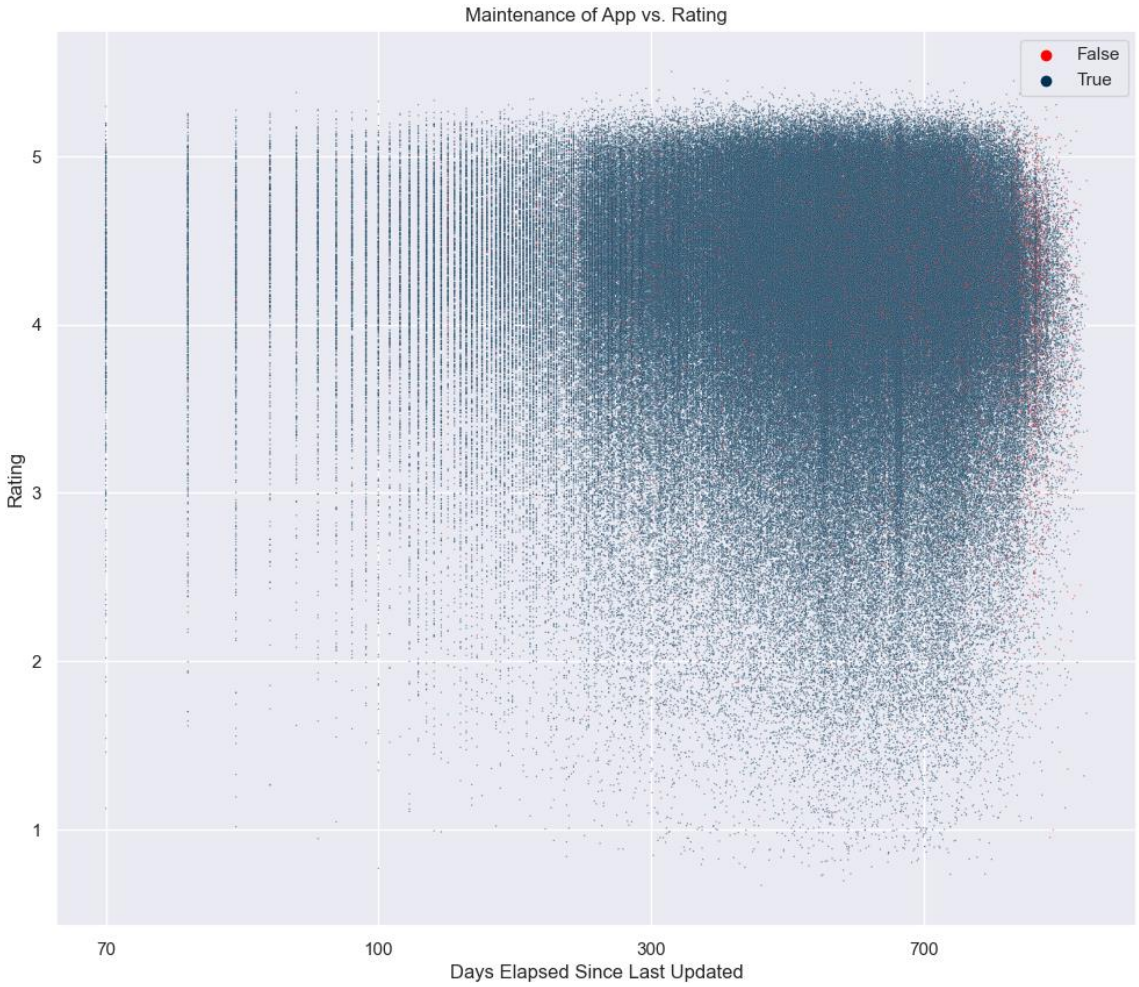

Figure 5.3.3.3: Maintenance of App vs Rating Grouping by Free or Paid

We can see very few apps were updated last year. (365 days have elapsed since the last update.) But we should also consider the possibility that not all apps were updated with relevant information by the dataset owner. And here low ratings could also indicate that a new update has not been well received by an app's user base. Surprisingly, paid apps are much less maintained than free apps.

NUMBER OF INSTALLED APPS CONTENT RATING WISE ACCORDING TO CATEGORY

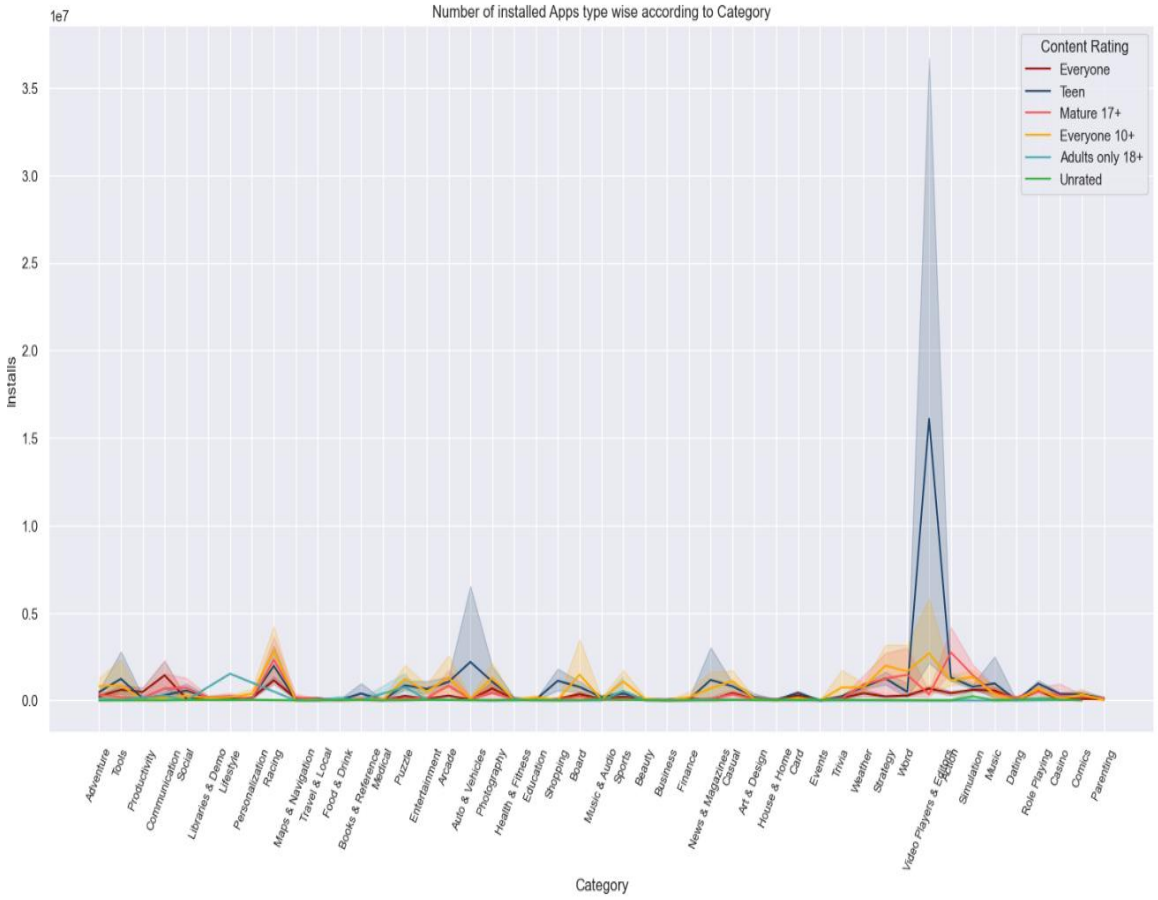

Figure 5.3.3.4: Number of Installed Apps Content Rating Wise According to Category

Most of the installations are done by the teens and the most are Video Players and Editors. Most of the Adults showing interest in downloading the Lifestyle, Spots, Puzzles and Communication Category. Unrated installations are barely seen in every category.

5.3.4 Sentiment Analysis

SENTIMENT REVIEWS

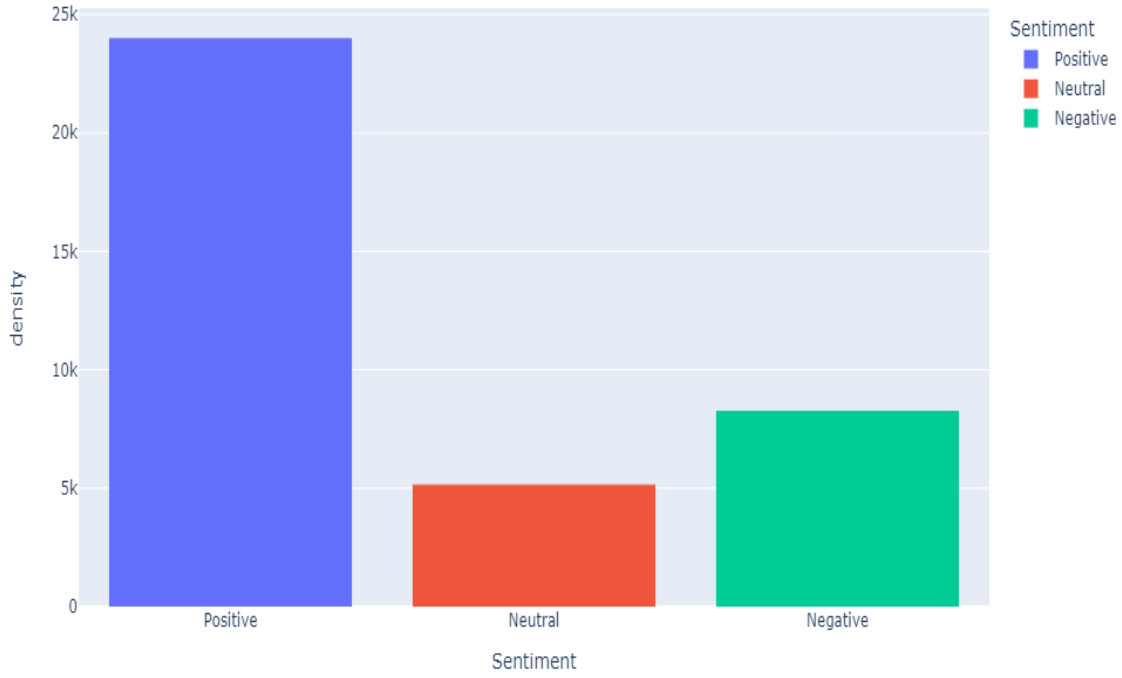

Figure 5.3.4.1: Sentiment Reviews

According to sentiment reviews most of the reviews are positive and negative reviews are less than half of positive.

SENTIMENT POLARITY DISTRIBUTION

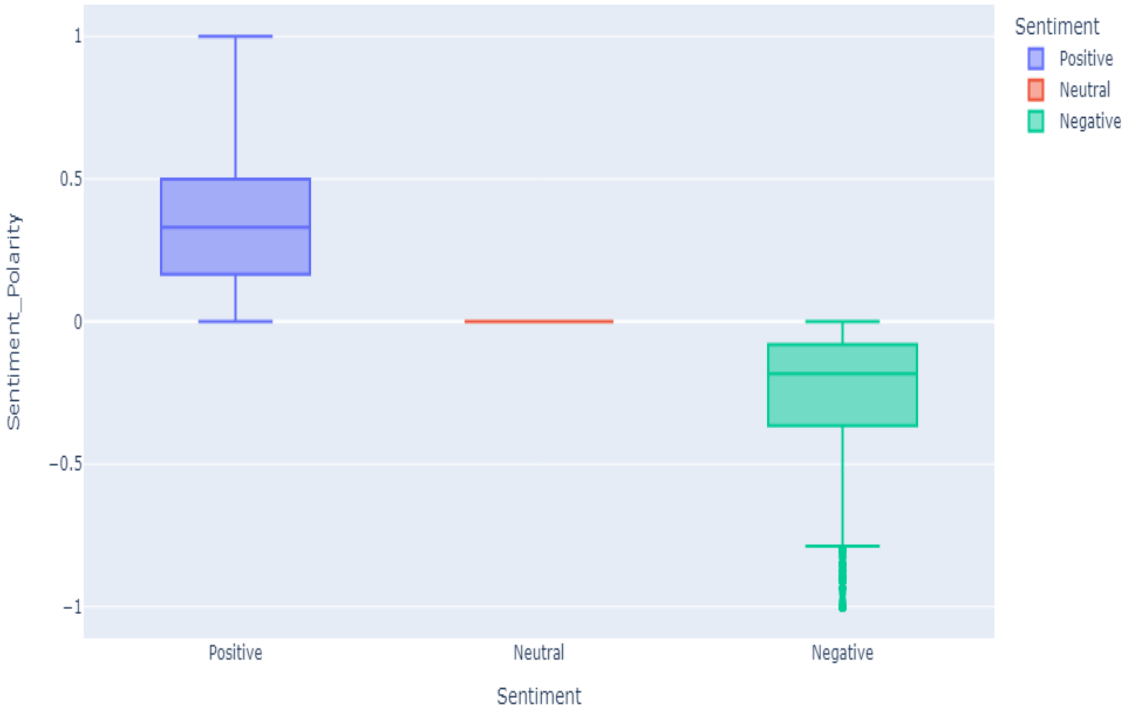

Figure 5.3.4.2: Sentiment Polarity Distribution

The boxplot of a sample of negative words from a population which is skewed to the left. The bottom whisker is much longer than the top whisker and the line is rising to the top of the box. And the other hand positive words have a positively skewed distribution while neutral words distribution is symmetrical.

WORDCLOUD

All

Figure 5.3.4.3: All Words

Positive

Figure 5.3.4.4: Positive Words

Negative

Figure 5.3.4.5: Negative Words

From above figures (figure number 5.3.4.3, 5.3.4.4, and 5.3.4.5), We can see that most feelings refer to the 'Games' category, as it appears in a larger size. Words like great, good, love, best, nice appear for positive words. As for the negative words, we have propaganda, hate, bad, malware, update, useless.

5.4 Analytics Models and Algorithms

Analyzing the dataset to derive insights for developers can be a major objective. One could define app installs as their success. So, we want to analyze: What factors do influence installs? Let's see, if the data convinces us of any significant relationship between the Installs and all other meaningful variables.

5.4.1 Correlation

Correlation is a statistical term describing the degree to which two variables move in coordination with one-another. Correlation coefficients are used to measure the strength of the linear relationship between two variables. There are different types of correlation:

- If the two variables move in the same direction, then those variables are said to have a positive correlation. In this case, the correlation coefficient is greater than zero.
- If they move in opposite directions, then they have a negative correlation. In this case, the correlation coefficient is less than zero.
- In certain cases, the correlation coefficient is exactly zero. This means that there is no relation between the variables.

The value of correlation coefficient varies anywhere between 1 and -1. The closer the value is to 0, the weaker is the relation and the closer it is to 1 or -1, the stronger is the relation.

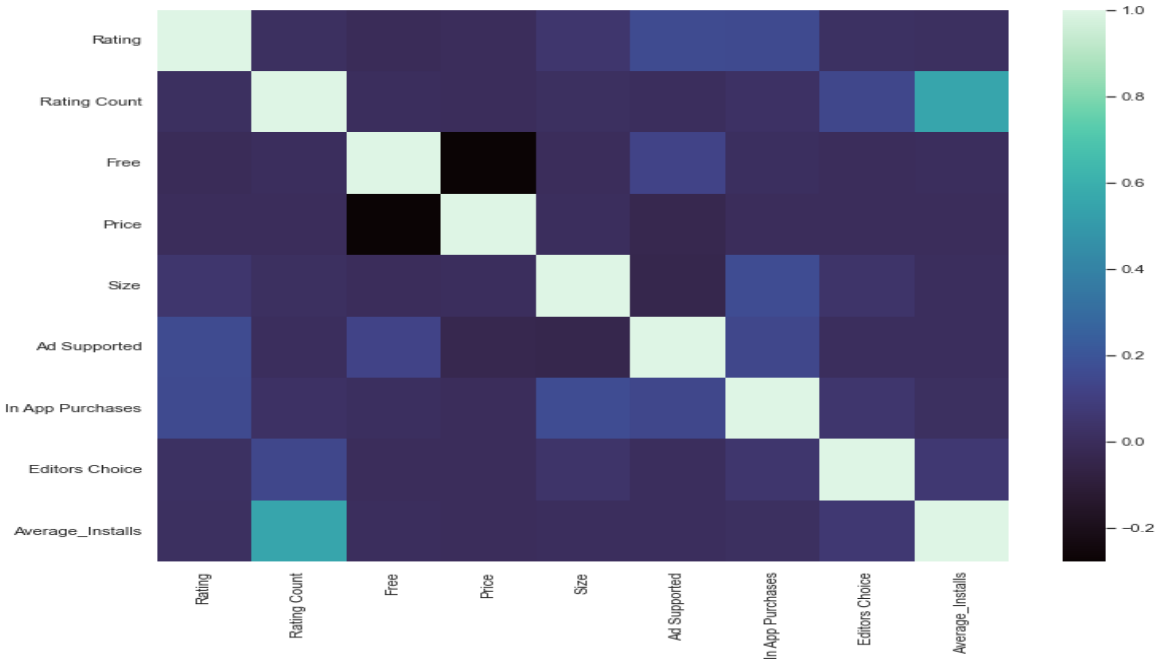

Figure 5.4.1.1: Heatmap

	Rating	Rating Count	Free	Price	Size	Ad Supported	In App Purchases	Editors Choice	Average_Installs
Rating	1.000000	0.022544	-0.008037	-0.004302	0.023156	0.160146	0.150603	0.017634	0.029322
Rating Count	0.022544	1.000000	0.002226	-0.000669	0.011376	0.010997	0.047024	0.180521	0.537291
Free	-0.008037	0.002226	1.000000	-0.276199	-0.062689	0.125345	0.008446	-0.003809	0.004069
Price	-0.004302	-0.000669	-0.276199	1.000000	0.012539	-0.034244	-0.002802	0.000583	-0.001156
Size	0.023156	0.011376	-0.062689	0.012539	1.000000	-0.081719	0.036460	0.012786	0.011644
Ad Supported	0.160146	0.010997	0.125345	-0.034244	-0.081719	1.000000	0.130729	0.006089	0.015636
In App Purchases	0.150603	0.047024	0.008446	-0.002802	0.036460	0.130729	1.000000	0.044063	0.050844
Editors Choice	0.017634	0.180521	-0.003809	0.000583	0.012786	0.006089	0.044063	1.000000	0.110929
Average_Installs	0.029322	0.537291	0.004069	-0.001156	0.011644	0.015636	0.050844	0.110929	1.000000

Figure 5.4.1.2: Correlation Values

According to figures 5.4.1 and 5.4.2, we can see that Installs and Rating Count has the strongest inverse correlation. This is reasonable because more ratings are conducted on apps that are the most popular. Since Installs was not correlated to Type (Free) this disproves our intuition that free apps lead to more installs. Since the Installs parameter is independent and not correlated to any other parameters, we must only use Installs and Rating Count for linear regression.

5.4.2 Linear Regression

Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable.

This form of analysis estimates the coefficients of the linear equation, involving one or more independent variables that best predict the value of the dependent variable. Linear regression fits a straight line or surface that minimizes the discrepancies between predicted and actual output values.

Types of Linear Regression

Simple Linear Regression - 1 dependent variable (interval or ratio), 1 independent variable (interval or ratio or dichotomous)

Multiple Linear Regression - 1 dependent variable (interval or ratio), 2+ independent variables (interval or ratio or dichotomous)

Logistic Regression - 1 dependent variable (dichotomous), 2+ independent variable(s) (interval or ratio or dichotomous)

Ordinal Regression - 1 dependent variable (ordinal), 1+ independent variable(s) (nominal or dichotomous)

Multinomial Regression - 1 dependent variable (nominal), 1+ independent variable(s) (interval or ratio or dichotomous)

Simple Linear Regression

For the model, I use Average Installs as dependent variable and Rating count as independent variable.

```
np.corrcoef(x, y)
```

```
array([[1.        , 0.95840999],
       [0.95840999, 1.        ]])
```

```
x = x.reshape(-1, 1)
```

```
from sklearn.linear_model import LinearRegression
regressor = LinearRegression()
regressor.fit(x, y)
```

```
LinearRegression()
```

```
predict = regressor.predict(x)
predict
```

```
array([ 9.18309358, 10.89347019, 15.16190254, ...,  5.69410581,
        8.86788063, 16.59769305])
```

Figure 5.4.2.1: Simple Linear Regression Model

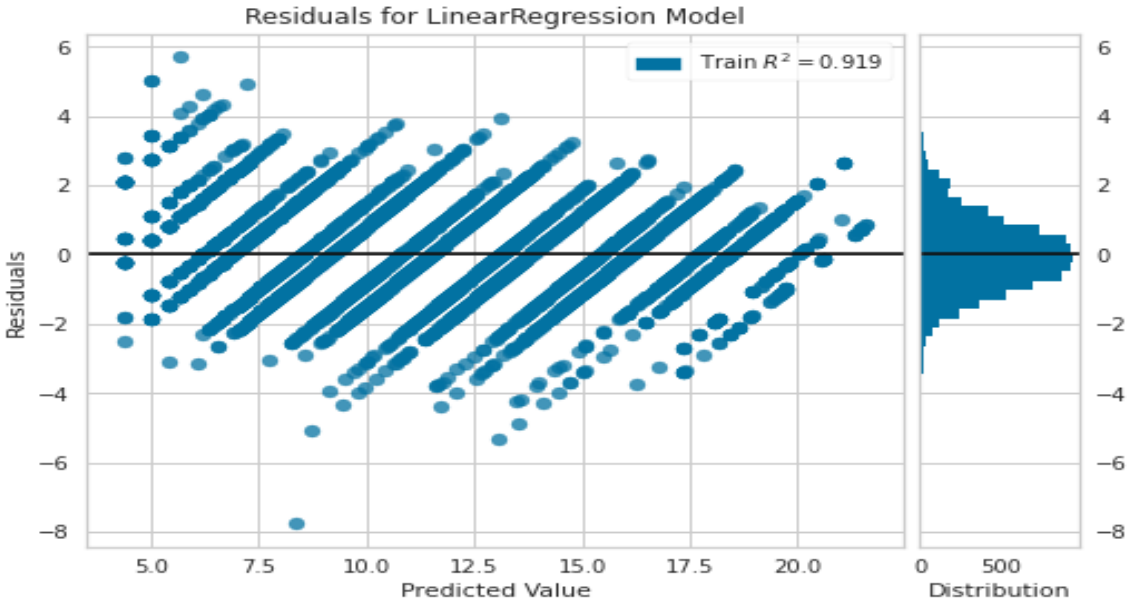

Figure 5.4.2.1: Residuals for Linear Regression Model

Confirmed! The number of downloads can be explained and predicted by the number of ratings. And 92% of the variability observed in the target variable is explained by the regression model.

CHAPTER 06

DISCUSSION AND RECOMMENDATIONS

6.1 Discussion

People are more interested to install the gaming Apps, the top Rating is given to the gaming apps. In App Purchases are correlated to App rating. So, we can say that if the app provides customer support and has subscription plans it will help to engage customers. Most people do not give rating, But the people who are given rating tend to give 4+ rating the most. Most of the Adults installed the Sports, lifestyle and communication Apps. Most of the installations are done by the teens and the most are Video Players and Editors. Video Players and Editors have more demand. People are mostly downloading the free apps. The installation of the free apps is high and the availability of the free apps also is very high. There are 20 applications with 1+ billion installations and Top 2 apps are Facebook and Whatsapp. Over 60% of Editor's Choice apps were games, despite only comprising less than 20% of all apps and for an app to be labeled as "Editor's Choice", average installations play a larger role than app ratings. The number of downloads can be explained and predicted by the number of ratings.

6.2 Recommendations and Future Work

My main motive for this research was to find something meaningful throughout the analysis that might matter to the developing community or to the end users. But just by looking at the EDA and the linear regression its difficult figure out the success of the app, so it will be better to create a success metric which I think will be a big contribution for the developers as through this they will get to know their success rate and will be able to decide what feature needs to be maintained or which one needs to be modified according to the current state of their app.

6.3 Conclusions

The Google Play Store is the largest app market in the world. It generates more than double the downloads of the Apple App Store, but makes only half the money as the App Store. So, I found data from Kaggle to conduct research on it. To deliver insights to understand customer demands better and thus help developers to popularize the product. I've only cracked the surface with this exploratory data analysis, and there are so many more insights to uncover. The average rating of apps on the Google Play Store is 4.19, where the number of downloads is related to the amount of user ratings, mostly for free apps. Users prefer smaller sized apps whose main categories are Video Players & Editors, Communication and Games. The number of downloads can be explained and predicted by the number of ratings with 0.919 r- squared. But the dataset is neither representative for the whole Google Play Store, nor can *we* find any significant relation between installs – except rating count, which can be a proxy for an App developer's success -- and other observed features/variables.

List of References

[1] (2018). Google play store: number of apps 2018. [online] https://www.statista.com/statistics/266210/number-of-available-applications-in-the-google-play-store/.

[2] (2018). Number of daily android app releases worldwide 2018 | statistic. [online] https://www.statista.com/statistics/276703/android-app-releases-worldwide/.

[3] Aralikatte, R., Sridhara, G., Gantayat, N., and Mani, S. (2018). Fault in your stars: an analysis of android app reviews. In Proceedings of the ACM India Joint International Conference on Data Science and Management of Data, pages 57–66. ACM.

[4] Breiman, L. (2001). Random forests. Mach. Learn., 45(1):5–32.

[5] Fu, B., Lin, J., Li, L., Faloutsos, C., Hong, J., and Sadeh, N. (2013). Why people hate your app: Making sense of user feedback in a mobile app store. In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 1276–1284. ACM.

[6] Grover, S. (2015). 3 apps that failed (and what they teach us about app marketing). [online] https://blog.placeit.net/apps-fail-teach-us-app-marketing/ .

[7] Zhong N, Michahelles F (2013) Google Play is Not a Long Tail Market: An Empirical Analysis of App Adoption on the Google Play App Market, in: Proceedings of the 28th Annual ACM Symposium on Applied Computing, SAC ‘13. ACM, New York, pp. 499–504.10.1145/2480362.2480460 .

[8] Martin W, Sarro F, Harman M (2016) Causal Impact Analysis for App Releases in Google Play, in: Proceedings of the 2016 24th ACM SIGSOFT International Symposium on Foundations of Software Engineering, FSE 2016. ACM, New York, pp. 435–446. 10.1145/ 2950290.2950320.

[9] Abu Farha, I., & Magdy, W. (2021). A comparative study of effective approaches for Arabic sentiment analysis. Information Processing & Management, 58(2), 102438. doi:10.1016/j.ipm.2020.102438

[10] Finkelstein A, Harman M, Jia Y, Martin W, Sarro F, Zhang Y (2017) Investigating the relationship between price, rating, and popularity in the Blackberry World App Store. Inf Softw Technol 87:119–139. [online] Investigating the relationship between price, rating, and popularity in the Blackberry World App Store – ScienceDirect

[11] Luiz, W., Viegas, F., Alencar, R., Mourão, F., Salles, T., Carvalho, D., Gonçalves, M. A., and Rocha, L. (2018). A feature-oriented sentiment rating for mobile app reviews. In Proceedings of the 2018 World Wide Web Conference on World Wide Web, pages 1909–1918. International World Wide Web Conferences Steering Committee.

[12] Data science a comprehensive analysis on "google play store apps" dataset from kaggle. [online] https://towardsdatascience.com/data-science-a-deep-analysis-on-google-play-store-apps-from-kaggle-8283bbc508b0

[13] Mucherino, A., Papajorgji, P. J., and Pardalos, P. M. (2009). k-Nearest Neighbor Classification, pages 83–106. Springer New York, New York, NY.

[14] Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs up?: sentiment classification using machine learning techniques. In Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10, pages 79–86. Association for Computational Linguistics.

[15] Ruiz, I. J. M., Nagappan, M., Adams, B., Berger, T., Dienst, S., and Hassan, A. E. (2016). Examining the rating system used in mobile-app stores. IEEE Software, 33(6):86–92.

[16] Saxena, P. (2018). How much money can you make with your app. [online] https://appinventiv.com/blog/how-much-money-can-you-earn-through-your-mobile-app.

[17] Tuckerman, C. (2014). Predicting mobile application success.

[18] Valentine, A. (2017). 4 mobile app developer success stories. [online] https://blog.proto.io/4-mobile-app-developer-success-stories/