**A Major Qualifying Report submitted to the faculty of**

**BML MUNJAL UNIVERSITY**

****

**“Partially meets the requirements of the Degree Bachelor of Technology in Computer Science and Engineering”**

**I**

**By:**

**Jatin Kumar Phogat (220442)**

**Academic Year – (2022 – 23)**

**GOOGLE PLAY STORE APPS ANALYSIS**

**COURSE –** MATHEMATICS FOR ENGINEERING- II

**COURSE FACULTY –** DR. RANJIB BANERJEE

**Acknowledgment**

We would like to express our sincere gratitude to our teacher, Dr. Ranjib Banerjee, for his invaluable guidance throughout the project. His unwavering support and advice were instrumental in resolving our doubts and challenges, and his motivation was a driving force behind our successful completion of the project.

We would also like to extend our heartfelt appreciation to all the members of our group who worked tirelessly and gave their best to ensure that our project was of the highest quality. Their dedication and hard work were key to our success.

Finally, we would like to express our gratitude to all those who provided their support and encouragement, both during the project and in our personal lives. Your unwavering support was a great source of inspiration, and we could not have completed this project without you.

**Abstract**

In order to get insight into the traits and trends within the app industry, this research analyses the Google Play Store applications dataset. The dataset consists of a varied selection of applications from several categories with varying ratings, reviews, and installations. The investigation starts by looking at how applications are distributed across different categories and determining the top installed categories. To better understand customer preferences and satisfaction, it looks at the connection between app ratings and the quantity of installations. The analysis also looks for any noticeable trends by examining the relationship between app size, pricing, and ratings. Additionally, it considers how app reviews affect app ratings, examining if more reviews are associated with higher ratings. The research also examines how apps are installed geographically, finding the areas with the highest app usage and prospective market prospects. The results of this research offer developers, marketers, and other app industry players useful insights. To improve app performance and increase user engagement, the data can direct category selection, pricing tactics, and marketing initiatives

**Introduction**

With a huge variety of capabilities and entertainment alternatives at our fingertips, the explosive proliferation of mobile applications has completely changed how we engage with technology. With millions of consumers worldwide, the Google Play Store, one of the biggest venues for Android apps, offers a rich and diversified environment where developers may present their works. In this research, we analyse data taken from the Google Play Store, concentrating on a number of different variables including app categories, ratings, installations, price, and user preferences.

The main goal of this research is to learn more about the traits and patterns found in the dataset in order to shed light on the elements that contribute to an app's success and popularity. We want to address important issues including the distribution of free and paid applications, the popularity of various categories, the top-rated apps, and the correlation between ratings and app price by researching key indicators and visualising the data.

To guarantee the correctness and dependability of the results, a comprehensive data cleaning procedure was carried out before beginning the study. This required reducing unnecessary columns, dealing with missing data, and getting rid of duplicate items. The generated dataset offered a reliable basis for carrying out significant analysis and coming to insightful findings.

We give a number of studies and visualisations that provide insightful perspectives into the Google Play Store app ecosystem throughout this research. These results may help app creators, marketers, and enthusiasts make wise choices, better understand user preferences, and spot possible app enhancement opportunities.

We strive to offer a thorough picture of the Google Play Store environment by comprehending the dynamics of app categories, user ratings, price structures, and installation patterns. Each analysis will be covered in depth in the following sections of this report, which will also give major results, visual representations, and takeaways from the dataset.

Let's now go out on this data-driven expedition and investigate the intriguing world of Google Play Store applications!

**Methodology**

The following technique was used to complete a thorough examination of the Google Play Store applications dataset:

**Data collection:** The Google Play Store was used to gather the original dataset, which contained details on numerous apps. Columns in the data collection included App, Category, Rating, Reviews, Size, Installs, Type, Price, Content Rating, and Genres.

**Data Cleaning:** To assure the accuracy and dependability of the results, a complete data cleaning procedure was carried out before moving on to the analysis. This required deleting columns that were not important to the analysis. To protect data integrity, any rows with blank values or duplicate entries were also removed. In addition, required adjustments were made, such deleting superfluous symbols like "M" from the Size column.

**Data Analysis:** After the dataset was cleaned, a number of studies were carried out to glean information about various facets of the Google Play Store app ecosystem. Among these analyses were:

**a.** **Calculating the percentage of free and paid applications** in the dataset can help you better understand how they are distributed.

**b.** **Number of App Installs by Category**: Analyzing the quantity of app downloads within each category to gauge its appeal.

**c.** **The Top 15 Categories Based on App installs**: To highlight the most well-liked categories, the top 15 categories based on the quantity of app installs were determined.

**d. The Top 15 Apps Based on Installs**: To highlight the most well-liked apps, the top 15 apps were chosen based on the quantity of installs.

**e. Average Rating per Category:** To gauge user satisfaction across several app categories, the average rating for each category is computed.

**f. The average ratings of paid and free applications** are compared to identify any discrepancies in user evaluations caused by app price.

**g. App Pricing by Category:** Investigating how much different categories of applications cost in order to understand the pricing distribution of different app kinds.

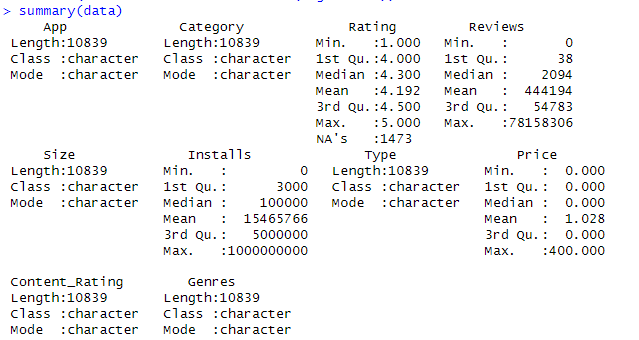
**h. The quantity of paid applications in each category** is calculated to analyses the distribution of paid apps among various categories.

The report is generated using R Markdown, which is an open-source tool for creating dynamic documents that combine text, code, and graphical output. This ensures that our analysis is adequately represented in the report, and readers can easily reproduce our findings. Overall, the methodology adopted for this report is robust, and reliable, and provides a comprehensive understanding of the Google Play Store Apps

****

A picture showing the dataset

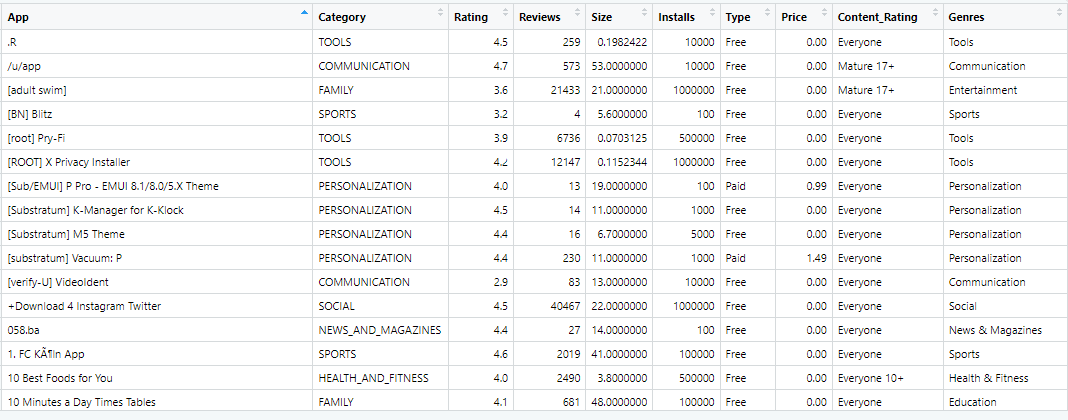
This was the summary of the data

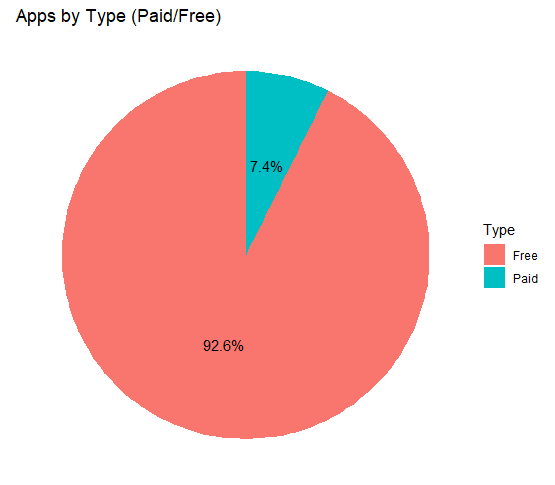


10,839 entries, 10 total columns

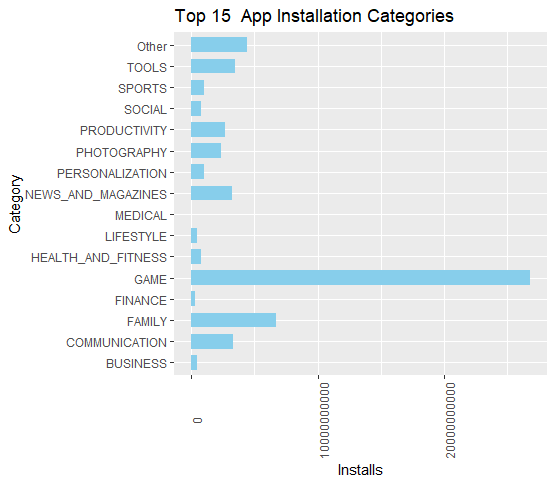
And after we cleaned the dataset by removing the nan value rows , M sign from the size column ,and “,” sign from the installs we had only 7,421 entries, 10 total columns

This is a picture of final dataset after cleaning which will be used for data analysis

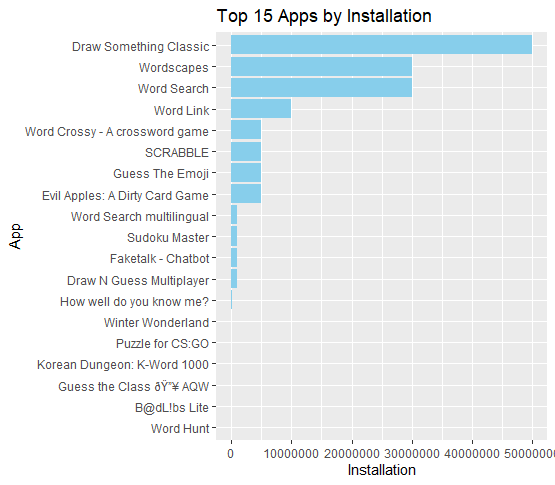


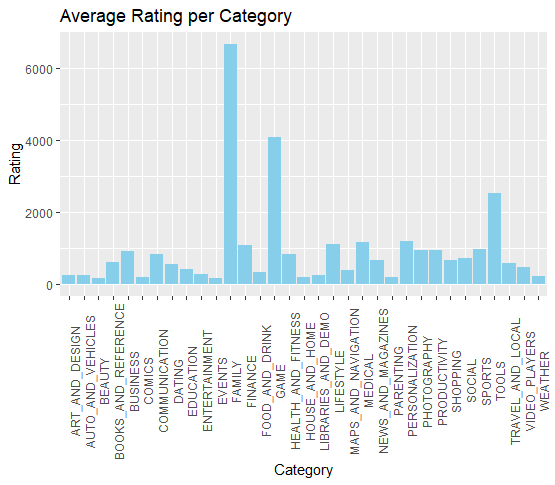


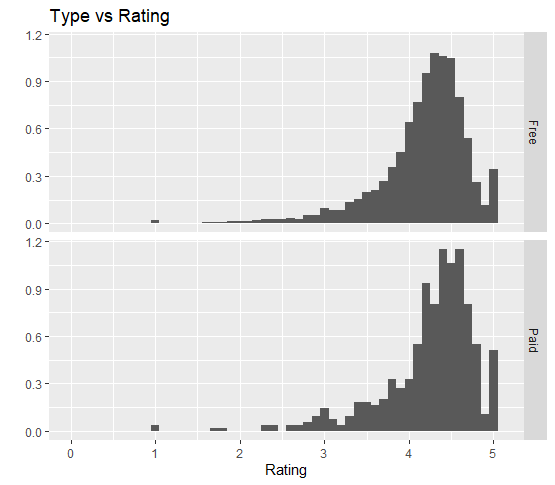
92% Percent of the apps on the play store are free to download



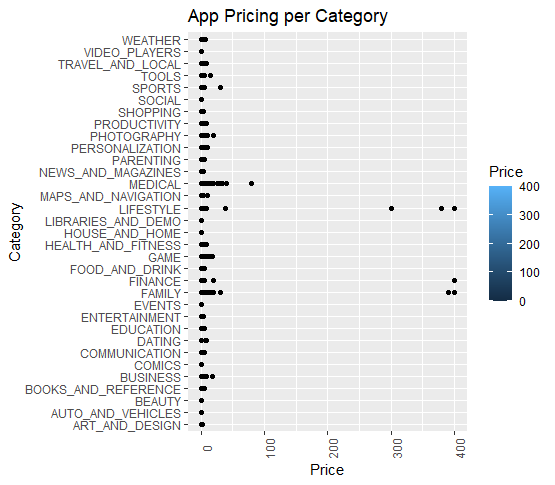
Family, game categories has most apps on the Google Play store and we can see over here from our recent analysis from analysis we can see that due to the download increase the game category is the most popular. From where we can see here are free however the lifestyle finance and family categories are associated with the highest cost which can actually cause of $400 As we can see here that which category has highest number of installs and since game category has highest number of installs we have found out the games which have the highest number of installs.







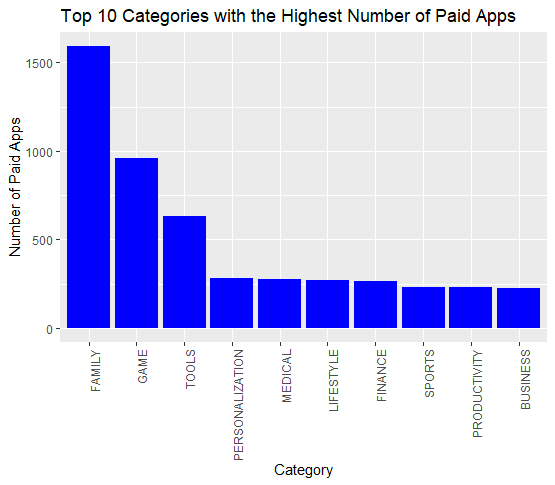
These shows the rating trend . We can see that Events, Food ,Tools, Games are the categories which are highly liked and have greatest ratings.

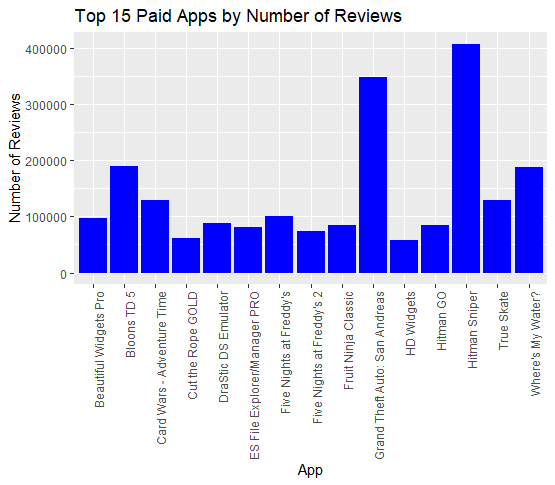


Here we can see that Lifestyle , Finance , Family categories have the apps with the highest cost which can be up to $400

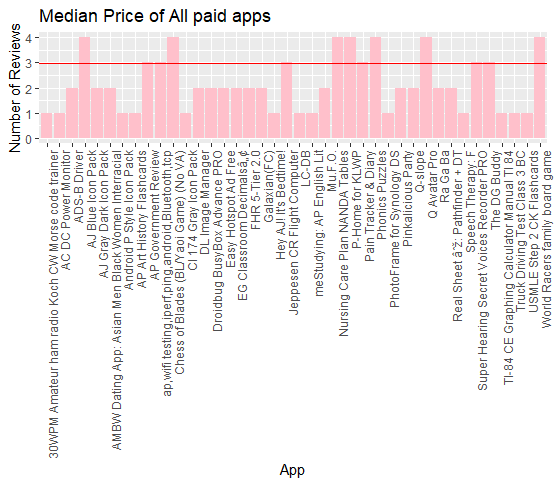
**INSIGHT 1**

* The best way of achieving success is to already have a following elsewhere and then transition into following to the mobile landscape
* Second having a high user engagement can contribute soon success
* Finally Targeting the specific audience for our app . Game category is definitely is targeting people who love video games ,who love the game online who love to play video games and that’s why this has contributed greatly to its success





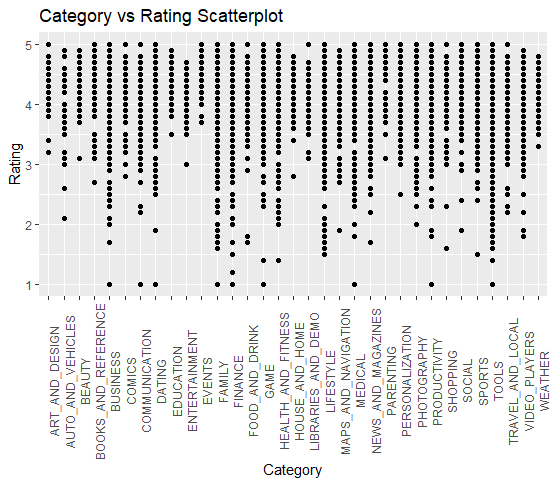
Here's a visualization of the number of paid apps who have a large number of downloads and also paid apps that generated the most value

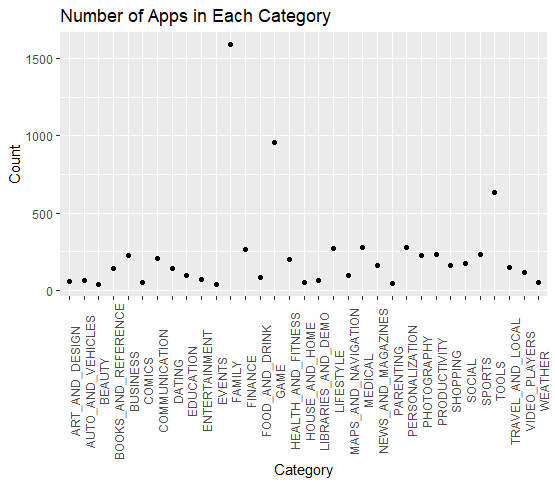


Analysis we can see the medium price of paid apps is $2.99 and one of the added benefit of owining an app or paid app is he gets have a fan base so you definitely can charge more or have additional charges generating more revenue

**INSIGHT 2**

* Beauty industry is worth over 500 billion dollars
* It is the category with the fewest apps therefore developing an app that caters to the beauty industry has a potential for great success in the category has yet received large investments
* Currently there are only 53 beauty apps in our data set of over 9600 apps it does not mean that these apps are not popular or that they are not good since the rank in the top 10 of the average reading per category with the rating of 4.3 out of 5
* Beauty apps are also quite popular with the best hairstyles app having over 5,000,000 downloads on the Google Play store therefore developing an app that free example helps men purchase clothing that suits their style and sets up daily outfits could be a very lucrative business composition

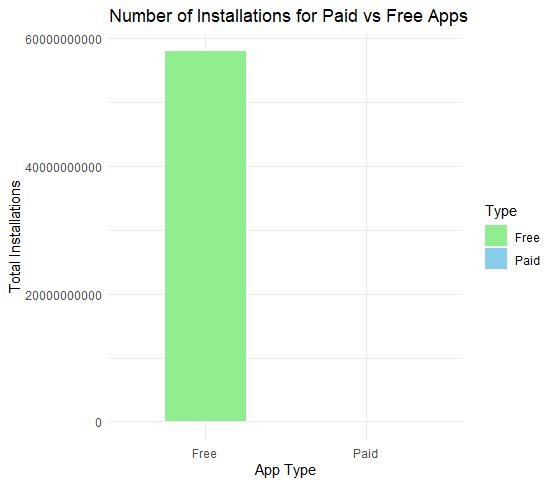




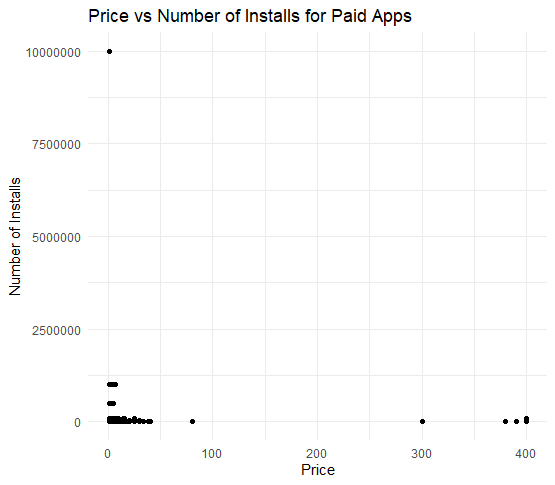
These two are the analysis to support the Second Insight.

**INSIGHT 3**

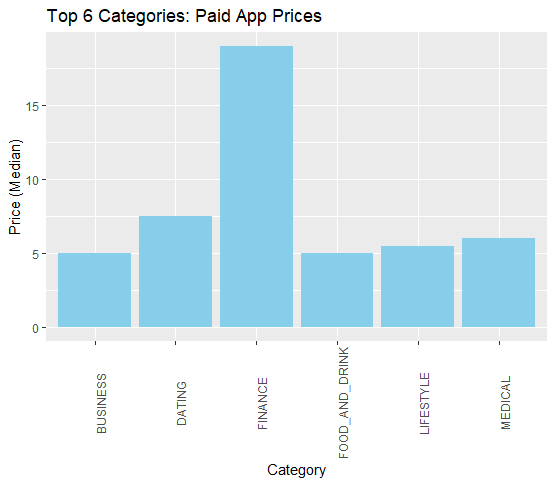
* When designing an app it's important to consider the amount of people that could be interested in downloading your app
* When designing 4 niche market its better to make the app cost money upfront
* Buyers are more willing to pay since they are less leads shows available



Here we can see people like to install free apps instead of paid apps .



People generally install apps in the range of $0-$100 .As we can see in Previous Graph there very less people who install apps having price greater then $100.



Here we can see that finance category has the highest number of prices.Followed by Dating and Medical . All these serves a very specific niche with few buyers to have the highest medium

We dive deeper in the medical category we can see that medical absent general did not have a lot of downloads and thus further expands the reason to have a higher cost structure so that it pays for the development cost of the app.

**Conclusion**

Following a thorough investigation of the Google Play Store applications dataset, the following major observations and conclusions have been made:

1. App categories and pricing:

* The most paid applications are in the finance, dating, and medical categories, showing a concentration on specialist areas with a smaller user base.
* The majority of users download apps that cost between $0 and $100, indicating that people prefer free or inexpensive software’s base.
* Despite being a rich profession, there aren't many applications in the beauty space. Due to the lack of rivalry in the area, creating an app for the beauty business provides a chance for success.

1. User Engagement and Ratings:

* With strong user involvement and favorable reviews, games and certain categories like Events, Food, and Tools are popular with users.
* Building an app with a specific target market in mind, like gamers, might greatly increase its success.

1. App Revenue Generation:

* The median cost of paid applications is $2.99, indicating that developers can set higher prices for their apps in order to capitalize on their fan bases and increase income.
* An app's success may be influenced by factors like a large user base or significant user engagement.

1. Category Insights: Due to their niche nature and smaller target markets, the Lifestyle, Finance, and Family categories sometimes contain applications with higher prices.

* The most installs are in the Game category, a sign of the category's popularity among users. It might be beneficial to examine the top games in terms of installation rates.

1. Overall App Distribution:

* The Google Play Store has almost 92% of all apps available for free download, highlighting how commonplace free apps are now.

This study of the Google Play Store applications dataset, in summary, offers important information about app pricing, user engagement, category trends, and income production. These results can help companies and app developers make strategic choices when creating and releasing apps on the market.

**Code:**

library(purrr)

library(readxl)

library(stringr)

library(dplyr)

library(tidyr)

library(ggplot2)

library(plyr)

library(magrittr)

library(forcats)

data <- read\_excel('C:\\Users\\phoga\\Desktop\\googleplaystore.xlsx')

View(data)

summary(data)

#########DATA CLEANING#############

#taking unique data only

gpdata <- unique(data)

View(gpdata)

#removing rows with Nan values

gpdata<-gpdata %>% drop\_na()

sum(is.na(gpdata))

# Remove "M"

gpdata$Size <- str\_replace\_all(gpdata$Size, "M", "")

#Removeing ‘,’ from installs column.

gpdata$Installs <- str\_replace\_all(gpdata$Installs,"\\,","")

gpdata$Installs <- as.numeric(gpdata$Installs)

gpdata$Size <- as.numeric(gpdata$Size)

gpdata$Rating = as.numeric(gpdata$Rating)

gpdata$Reviews <- as.numeric(gpdata$Reviews)

str(gpdata)

summary(gpdata)

############## FREE APPS VS PAID APPS

gpdata %>%

ggplot(aes(x = "", fill = Type)) +

stat\_count(width = 1) +

coord\_polar(theta = "y") +

labs(title = "Apps by Type (Paid/Free)", x = NULL, y = NULL, fill = "Type") +

theme\_void() +

geom\_text(aes(label = paste0(round((..count.. / sum(..count..)) \* 100, 1), "%")),

stat = "count", position = position\_stack(vjust = 0.5))

################## NUMBER OF APPS INSTALLED PER CATEGORY

options(scipen = 999)

ggplot(gpdata, aes(x = Category, y = Installs)) +

geom\_bar(stat = "identity", width = 0.7, fill = "skyblue") +

coord\_flip() +

labs(title = "Total App Installation for Each Category") +

theme(axis.text.x = element\_text(angle = 90))

#################### #TOP 15 CATEGORIES

ggplot(gpdata %>%

mutate(Category = fct\_lump(Category, n = 15)),

aes(x = Category, y = Installs)) +

geom\_bar(stat = "identity", width = 0.7, fill = "skyblue") +

coord\_flip() +

labs(title = "Top 15 App Installation Categories") +

theme(axis.text.x = element\_text(angle = 90))

################### Top 15 Apps according to Installation

top\_15\_apps <- gpdata %>%

arrange(desc(Installs)) %>%

top\_n(15)

ggplot(top\_15\_apps, aes(x = reorder(App, Installs), y = Installs)) +

geom\_bar(stat = "identity", fill = "skyblue") +

coord\_flip() +

labs(title = "Top 15 Apps by Installation",

x = "App",

y = "Installation")

##################### Average Rating Per category

ggplot(gpdata, aes(x = Category, y = Rating)) +

geom\_bar(stat = "identity", fill = "skyblue") +

labs(title = "Average Rating per Category",

x = "Category",

y = "Rating") +

theme(axis.text.x = element\_text(angle = 90))

######################## AVERAGE RATING PAID VS FREE

gpdata %>%

ggplot(aes(Rating))+

geom\_histogram(binwidth = .1, aes(y=..density..))+

scale\_x\_continuous(limits = c(0, 5.1))+

labs(title = "Type vs Rating",

x = "Rating",

y = "")+

facet\_grid(Type~.)

################## App pricing per category

ggplot(gpdata, aes(x = Price, y = Category, fill= Price)) +

geom\_point() +

labs(title = "App Pricing per Category",

x = "Price",

y = "Category") +

theme(axis.text.x = element\_text(angle = 90))

#################Number of Paid Apps category vise

paid\_apps\_counts <- table(gpdata$Category)[order(table(gpdata$Category), decreasing = TRUE)]

top\_10\_paid\_apps <- data.frame(Category = names(paid\_apps\_counts)[1:10], Count = as.vector(paid\_apps\_counts)[1:10])

ggplot(top\_10\_paid\_apps, aes(x = reorder(Category, -Count), y = Count)) +

geom\_bar(stat = "identity", fill = "blue") +

xlab("Category") +

ylab("Number of Paid Apps") +

ggtitle("Top 10 Categories with the Highest Number of Paid Apps") +

theme(axis.text.x = element\_text(angle = 90, hjust = 1))

################## TOP 15 PAID APPS BAR CHART

paid\_apps <- subset(gpdata, Type == "Paid")

top\_15\_paid\_apps <- head(paid\_apps[order(paid\_apps$Reviews, decreasing = TRUE), ], 15)

ggplot(top\_15\_paid\_apps, aes(x = App, y = Reviews)) +

geom\_bar(stat = "identity", fill = "blue") +

xlab("App") +

ylab("Number of Reviews") +

ggtitle("Top 15 Paid Apps by Number of Reviews") +

theme(axis.text.x = element\_text(angle = 90, hjust = 1))

################# MEDIAN OF PAID APPS

paid\_apps <- subset(gpdata, Type == "Paid")

top\_15\_paid\_apps <- head(paid\_apps[order(paid\_apps$Reviews, decreasing = FALSE), ], 40)

median\_price <- median(as.numeric(paid\_apps$Price))

ggplot(top\_15\_paid\_apps, aes(x = App, y = Reviews)) +

geom\_bar(stat = "identity", fill = "#FFC0CB") +

geom\_hline(yintercept = median\_price, linetype = "solid", color = "red") +

xlab("App") +

ylab("Number of Reviews") +

ggtitle("Median Price of All paid apps ") +

theme(axis.text.x = element\_text(angle = 90, hjust = 1))

median\_price

###################### RATING VS CATEGORY SCATTERPLOT

ggplot(gpdata, aes(x = Category, y = Rating)) +

geom\_point() +

labs(title = "Category vs Rating Scatterplot",

x = "Category",

y = "Rating") +

theme(axis.text.x = element\_text(angle = 90))

################ APPS IN BEAUTY CAT

app\_counts <- table(gpdata$Category)

app\_counts\_df <- data.frame(Category = names(app\_counts), Count = as.numeric(app\_counts))

ggplot(app\_counts\_df, aes(x = Category, y = Count)) +

geom\_point() +

labs(title = "Number of Apps in Each Category",

x = "Category",

y = "Count") +

theme(axis.text.x = element\_text(angle = 90))

################# INSATLLATINg TREND FOR PAID VS FREE APPS

paid\_apps <- subset(gpdata, Type == "Paid")

free\_apps <- subset(gpdata, Type == "Free")

total\_installs\_paid <- sum(paid\_apps$Installs)

total\_installs\_free <- sum(free\_apps$Installs)

install\_data <- data.frame(Type = c("Paid", "Free"),

Total\_Installs = c(total\_installs\_paid, total\_installs\_free))

ggplot(install\_data, aes(x = Type, y = Total\_Installs, fill = Type)) +

geom\_bar(stat = "identity", width = 0.5) +

labs(title = "Number of Installations for Paid vs Free Apps",

x = "App Type",

y = "Total Installations") +

scale\_fill\_manual(values = c("Paid" = "skyblue", "Free" = "lightgreen")) +

theme\_minimal()

################### NO. OF INSTALLATION FOR PAID CATEGORY ACCORDING TO PRICE

paid\_apps <- subset(gpdata, Type == "Paid")

ggplot(paid\_apps, aes(x = Price, y = Installs)) +

geom\_point() +

labs(title = "Price vs Number of Installs for Paid Apps",

x = "Price",

y = "Number of Installs") +

theme\_minimal()

##################### CATEGORY WISE PAID APP PRICE

paid\_apps <- subset(gpdata, Type == "Paid")

category\_median\_price <- aggregate(Price ~ Category, paid\_apps, median)

sorted\_categories <- category\_median\_price[order(category\_median\_price$Price, decreasing = TRUE), ]

top\_categories <- head(sorted\_categories, 6)

ggplot(top\_categories, aes(x = Category, y = Price)) +

geom\_bar(stat = "identity", fill = "skyblue") +

labs(title = "Top 6 Categories: Paid App Prices",

x = "Category",

y = "Price (Median)") +

theme(axis.text.x = element\_text(angle = 90))

**REFERENCES: -**

1. <https://play.google.com/store>
2. <https://www.pewresearch.org/internet/2015/11/10/an-analysis-of-apps-in-the-google-play-store/>
3. <https://www.statista.com/topics/9929/google-play-store/?kw=&crmtag=adwords&gclid=CjwKCAjwjYKjBhB5EiwAiFdSfjsvakshYileV3bUx7H8boFqoJxFm2_0H2KaQU5yzNoWITnBs1P-6xoCsZoQAvD_BwE#dossier-chapter4>
4. <https://www.kaggle.com/datasets>

My Dataset can be accessed through:

https://www.kaggle.com/datasets/lava18/google-play-store-apps