Group 3

STAT 515- Applied Statistics and Visualization for Analytics

VISUALIZATION OF GOOGLE PLAY STORE DATASET

Sandeep Kumar Yedla

Bhavani Maddala

Shanmukh Alluri

Introduction:

Google Play is an advanced digital administration service or mobile applications on the Android working framework permitting clients to browse and download applications created with the Android programming development kit. With software engineering developing and open-source projects extending, the Google Play store is expanding in prevalence. While numerous public datasets give Apple's App Store information, there are relatively few partner datasets accessible for Google Play store applications, yet the Google Play store information can possibly drive application-production organizations. Dissimilar to web improvement or work area advancement, mobile application is extraordinary in its comfort. Android is the dominant mobile operating system today with about 85% of all mobile devices running Google's OS. The Google Play Store is the largest and most popular Android app store.

Analysis Questions:

1. what are the different categories of applications in the google Play Store and which has the highest number of applications?

2. What is the number comparison of paid vs free applications in various categories?

3. How is the application rating distributed across the Google play store dataset?

4. How is the rating distributed across various categories of the Play store dataset?

5. How many applications have been updated since the date of last update?

6. What is the correlation between Price, Rating, Size, Reviews and Installs?

Purpose, Data set:

The purpose of our project was to gather and analyse detailed information on apps in the Google Play Store in order to provide insights on app features and the current state of the Android app market. There are more than 3.04 million apps found on Google Play Store. With this project we will take you through a journey of analysing various apps found on the play store with the help of different R libraries. The dataset has been taken from Kaggle and it consists of 13 columns – App, Category, Rating, Reviews, Size, Installs, Type, Price, Content Rating, Genres, Last Updates, Current Ver and Android 10841 Rows.

Different Categories of Applications in the Google Play Store Using Pie chart:

The google app store data consists of applications under different category, it is obvious to explore the total number of applications in each category and understand which constitute to most of the data.

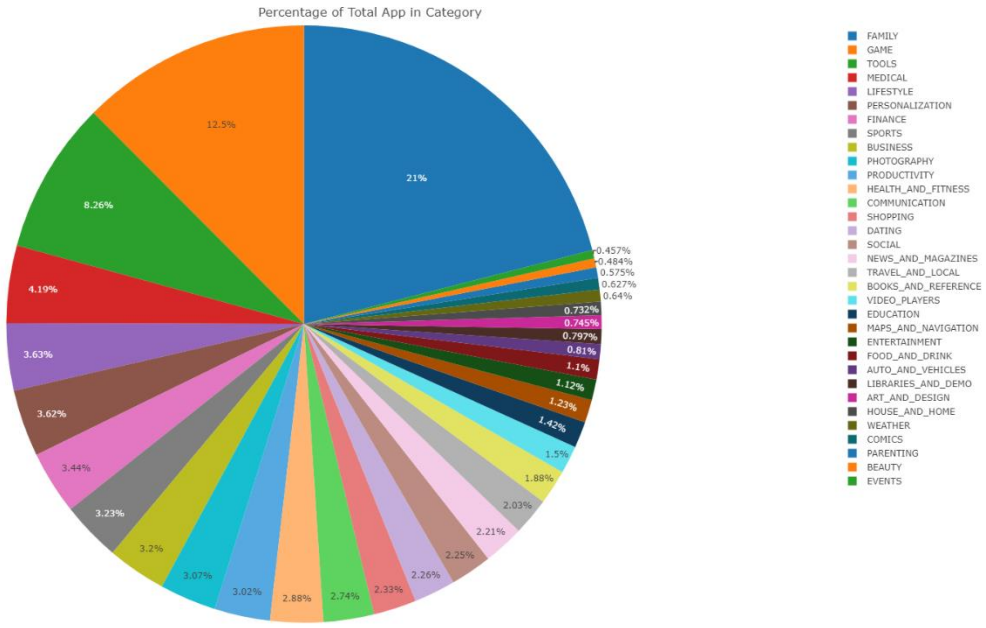


Figure 1

From the above pie chart, we can depict that most of the apps are under the category of family & Game and least apps are under Beauty & Comics Category. For example, 21% of the applications in the Google Play store belong under the category Family and 12.5% belong to the Category game. Going further, we can also depict the various applications under each category. For example, we can find the top 10 installed applications under game category and represent them.

Paid Vs Free Applications in various Categories Using Circular Bar Plot:

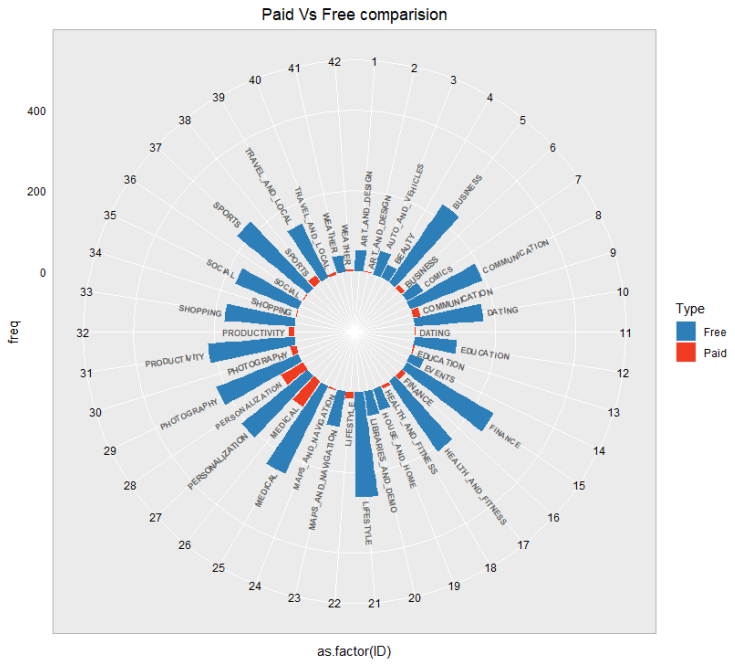


Figure 2

The Google play store consists of applications which are both paid and free. This would further help us compare the number of free and paid applications in each category. With the help of a circular bar graph, we can make the comparison between free and paid applications and come to a conclusion that most of the applications in the play store are for free and paid applications in most of the categories are very few (almost negligible). We can find good number of paid applications in Medical and personalization categories.

Distribution of ratings of the Applications in the Google Play store Data set using Density Curve:

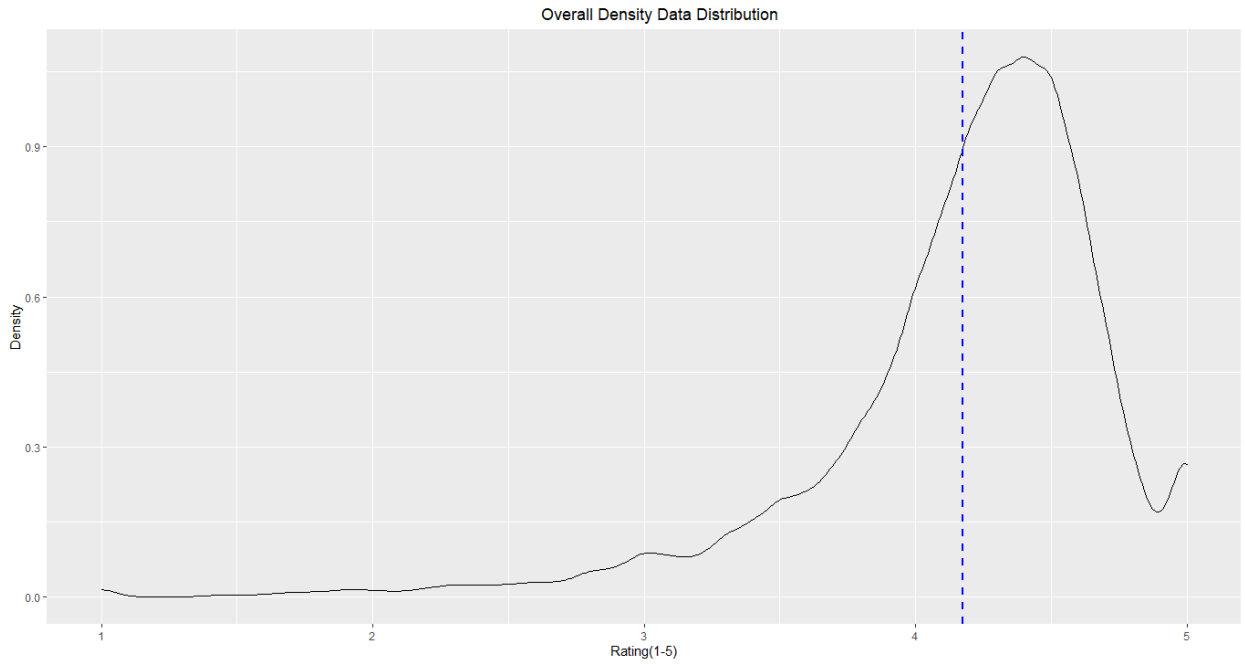


Figure 3

The analysis of the application ratings in the google play store plays a major role while depicting the current state of the App Market. Using a density curve, we can find the distribution of ratings in the Play store. From the above graph, we can conclude that most of the apps in the google play store are rated between 3.5 to 4.8 and the mean value of the ratings is 4.2.

Distribution of Ratings across various Categories of Applications using Side -by- Side Box Plot:

The analysis of application rating is a crucial aspect to visualize and analyze any category of applications in the google play store. understanding the range of its rating distribution give us insights of how the rating is distributed. The side-by-side box plot showcases the range of ratings of applications which are under a particular category.

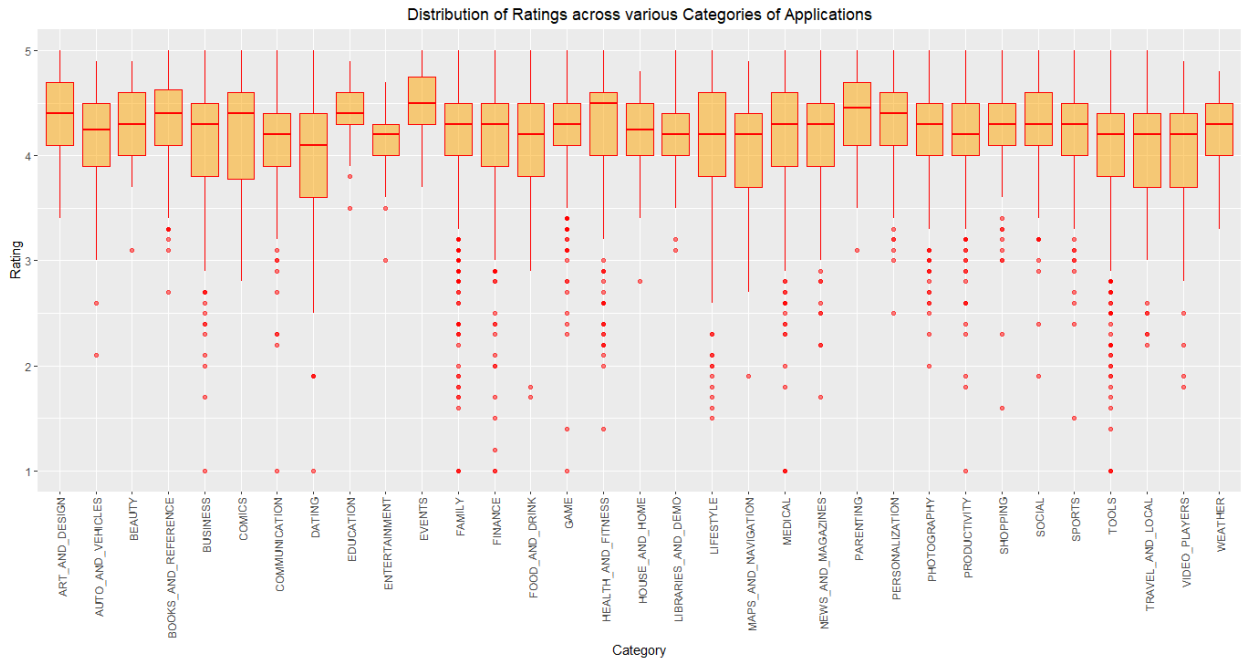


Figure 4

From the plot above, we can conclude that most ratings are continuous, and values are around 3.9 to 4.6. The outliers are for Family and Tools, because there are a greater number of installs for the family category and the ratings are distributed right from 1 to 4.5. However, the rating is mostly consistent for ART_AND_Design which lies in the range from 4.2 to 4.6. Going further, we can find which applications have highest and lowest rating within each category.

The frequency spread of Applications by date of last updates:

The play store dataset has the data of the applications which are updated to the new updates occurring. From this, we can get the insights of the applications which are updated over a particular period of time.

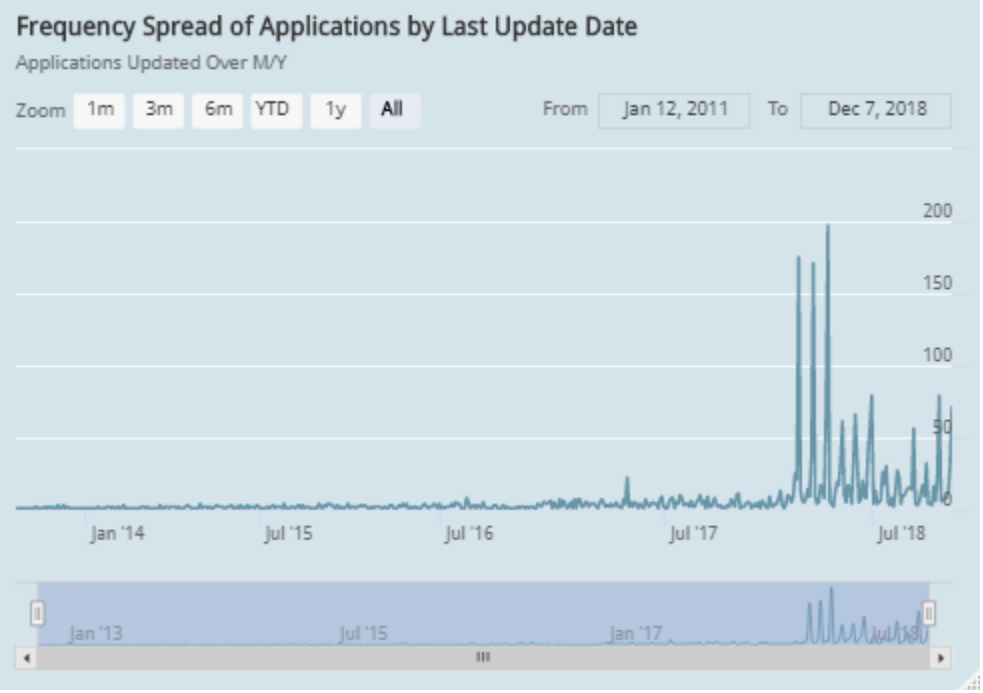


Figure 5

From the above high chart, we can see the number of applications updated over the years.

Between the year 2013 to 2017 there are not many applications which have been updated and from 2017 to 2018, we can observed increase in the number of applications updated.

Exploring the relationship between the numerical data using Correlation Plot:

It is always important to understand the dependency of the data to make conclusions or decisions. For that, understanding which factor affects the others by checking the proportionality whether they are positively or negatively corelated to each other. From the data set, the numerical data like Price, Rating, Size, Reviews, and Installs are major factors. Finding the relationship between them can help us to get insights on how to update specific applications and improve the business perspective and provide better services. By using correlation plot, we can get the insights of data dependency. The correlation plots give a scale ranging from -1 to +1 which means the more inclined towards +1 is a positive relation and more inclined towards -1 is a negatively corelated.

Figure 6

From the above Visualization, when we compare the number of installs with price, we can conclude that price and number of installs are negatively corelated, it almost constitutes to around -1 as there is minute dot representation. For example, if the price is high, the number of installs is decreased because people generally prefer free apps over paid apps unless it is necessary for them. Apart from that, when we compare number of installs with Rating, we can conclude that rating and number of installs are negatively corelated, as installs increase the ratings are decreasing, this might be because of the poor app quality and low ratings that they are negatively corelated. The number of installs is not affected by the size, as the correlation plot describes the zero scale in the plot for installs and size, so for the application developers whose applications are in the dataset can slightly be lenient towards applications of larger size with quality which might not affect the service/ their installs.

The number of installs is positively corelated with the Reviews, which is around 0.5, and might be affecting the number of installs. This could be because the users would download the application by the reviews. The developer has to consider the reviews, understand them and develop the next updated version of the application.

Conclusion:

Based on our Analyses, we can conclude that Google play store has 33 categories of applications and “Family” category has the highest number of applications and Play store has more free applications than the paid applications and the paid applications are mostly in medical and personalization category. Also, the distribution of ratings over the Google Play store is between 3.5 to 4.8 and the mean value of the ratings is 4.2.

The distribution of ratings across individual categories, Family, Tools have a greater number of outliers and the ratings are consistent for category Art_And_Design. The correlation between the applications is not strong enough and is mostly negative or negligible and also the number of applications updated between a specific time period can be calculated. Here, there are 200 applications which have been updated between 29th January and 26th March. There is no strong correlation between the application features and it is mostly negatively correlated or negligible.

REFERENCES:

https://www.kaggle.com/lava18/google-play-store-apps?select=googleplaystore.csv

https://www.rdocumentation.org/packages/janitor/versions/2.1.0

https://www.r-graph-gallery.com/piechart-ggplot2.html

https://www.r-graph-gallery.com/circular-barplot.html

https://www.statmethods.net/stats/regression.html

https://www.highcharts.com/blog/tutorials/highcharts-for-r-users/

APPENDIX:

1.) In the column Installs we had a '+' sign which was removed by using Excel and it was made numeric.

2.) The Size column contains the megabyte data and the kilobyte data, so for comparing converting them to one extension is required to mega byte data is converted into kilo byte using lambda function by dividing with 1000, so thus we have 19kb… for analysis.

```
import os
import csv
import pandas as pd
google_apps = pd.read_csv("cleaned_MyData.csv",encoding='cp1252');
google_apps.shape
```

```
In [2]: google_apps["Size"] = google_apps["Size"].apply(lambda x: str(x).replace(",", "") if "," in str(x) else x)
        google_apps["Size"] = google_apps["Size"].apply(lambda x: str(x).replace('M', '') if 'M' in str(x) else x)
        google_apps["Size"] = google_apps["Size"].apply(lambda x: str(x).replace("Varies with device", "NAN") if "Varies with device" in
        google_apps["Size"] = google_apps["Size"].apply(lambda x: float(str(x).replace('k', '')) / 1000 if 'k' in str(x) else x)
```

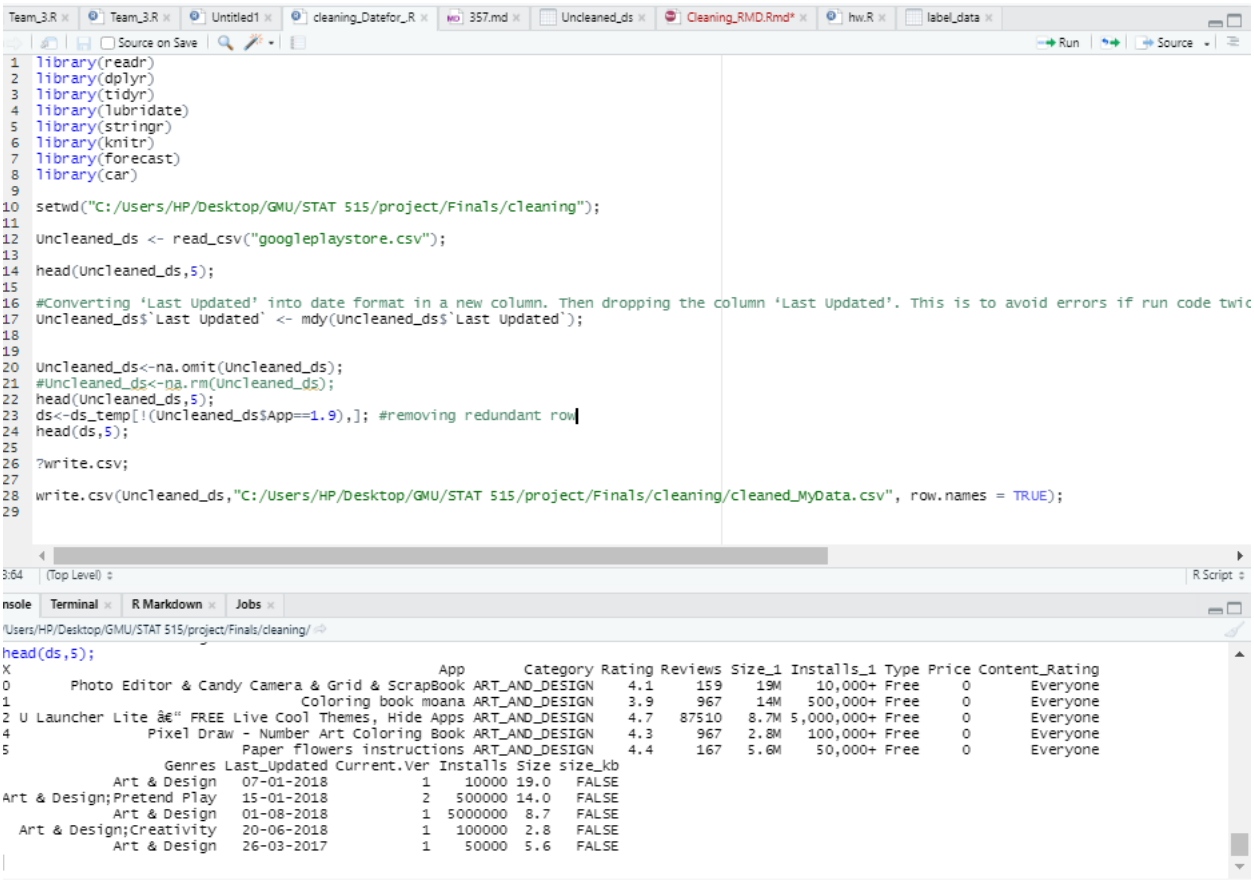
```
In [6]: df = pd.DataFrame(google_apps)
        df.head()
```

Out[6]:

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
0	Photo Editor & Candy Camera & Grid & ScrapBook	ART_AND_DESIGN	4.1	159	19	10,000+	Free	0	Everyone	Art & Design	2018-01-07	1.0.0	4.0.3 and up
1	Coloring book moana	ART_AND_DESIGN	3.9	967	14	500,000+	Free	0	Everyone	Art & Design;Pretend Play	2018-01-15	2.0.0	4.0.3 and up
2	U Launcher Lite – FREE Live Cool Themes, Hide ...	ART_AND_DESIGN	4.7	87510	8.7	5,000,000+	Free	0	Everyone	Art & Design	2018-08-01	1.2.4	4.0.3 and up
3	Sketch - Draw & Paint	ART_AND_DESIGN	4.5	215644	25	50,000,000+	Free	0	Teen	Art & Design	2018-06-08	Varies with device	4.2 and up
4	Pixel Draw - Number Art Coloring Book	ART_AND_DESIGN	4.3	967	2.8	100,000+	Free	0	Everyone	Art & Design;Creativity	2018-06-20	1.1	4.4 and up

```
In [4]: df.to_csv (r'C:\Users\HP\Desktop\GMU\playstore_sizecleaned.csv', index = True, header=True);
```

3.) Omitting the NA values which were created where created during the previous size conversion and the date format is changed with the lubridate library.



```
library(readr)
library(dplyr)
library(tidyr)
library(lubridate)
library(stringr)
library(knitr)
library(forecast)
library(car)

setwd("C:/Users/HP/Desktop/GMU/STAT 515/project/Finals/cleaning");

Uncleaned_ds <- read_csv("googleplaystore.csv");

head(Uncleaned_ds,5);

#Converting 'Last Updated' into date format in a new column. Then dropping the column 'Last Updated'. This is to avoid errors if run code twic
Uncleaned_ds$`Last Updated` <- mdy(Uncleaned_ds$`Last Updated`);


Uncleaned_ds<-na.omit(Uncleaned_ds);
#Uncleaned_ds<-na.rm(Uncleaned_ds);
head(Uncleaned_ds,5);
ds<-ds_temp[!(Uncleaned_ds$App==1.9),]; #removing redundant row
head(ds,5);

?write.csv;

write.csv(Uncleaned_ds,"C:/Users/HP/Desktop/GMU/STAT 515/project/Finals/cleaning/cleaned_MyData.csv", row.names = TRUE);
```

```
head(ds,5);
X                                                 App       Category Rating Reviews Size_1 Installs_1 Type Price Content_Rating
0       Photo Editor & Candy Camera & Grid & ScrapBook ART_AND_DESIGN    4.1     159    19M    10,000+ Free     0       Everyone
1                                 Coloring book moana ART_AND_DESIGN    3.9     967    14M   500,000+ Free     0       Everyone
2 U Launcher Lite â€“ FREE Live Cool Themes, Hide Apps ART_AND_DESIGN    4.7   87510   8.7M 5,000,000+ Free     0       Everyone
4                Pixel Draw - Number Art Coloring Book ART_AND_DESIGN    4.3     967   2.8M   100,000+ Free     0       Everyone
5                           Paper flowers instructions ART_AND_DESIGN    4.4     167   5.6M    50,000+ Free     0       Everyone
                  Genres Last_Updated Current.Ver Installs Size size_kb
            Art & Design   07-01-2018           1    10000 19.0   FALSE
Art & Design;Pretend Play   15-01-2018           2   500000 14.0   FALSE
            Art & Design   01-08-2018           1  5000000  8.7   FALSE
  Art & Design;Creativity   20-06-2018           1   100000  2.8   FALSE
            Art & Design   26-03-2017           1    50000  5.6   FALSE
```