

Capstone Project - 1

Play Store App Review Analysis

Presented By
Gaurav Kumar

Content

- Introduction
- Objective & Problem Statement
- Description of Data
- Import Library and data
- Data Cleaning
- Data analysis & visualizations
- Conclusion

Objective

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

Problem Statements

1. Top 10 installed apps in any category
2. Distribution of paid apps and free apps
3. Top Categories Apps in Google Play store
4. Which apps have the highest number of review?
5. Which app's have maximum number of installs in any category?
6. Top 10 expensive apps in the play store
7. Percentage of Review Sentiments
8. Positive review
9. Negative review

Data Set Description

There are two Data set. 1). Play Store Data 2). User Review Data

Play Store Data:

1. App – Application name
2. Category – Apps Category
3. Rating – Rating given to the apps
4. Review – Number of reviews given to the apps
5. Size – size of the app
6. Installs – No of users installed of the apps
7. Type – Free or Paid
8. Price – Price of the apps
9. Content Rating – An App belong which age group
10. Genres – Type of Genres the application belong to
11. Last Updated – When last time app updated
12. Current Ver – Current version of the app
13. Android Ver – Android version of the app

User review data set

App - App name

Translated Review – Reviews given by the consumer

Sentiment - Review sentiment i.e. Positive, Negative, Neutral

Sentiment Polarity – sentiment in numerical form range from -1 to 1

Sentiment Subjectivity – a measure of expression, of opinion, evaluations, feelings, and speculations

Import library & data



```
[177] # import libraries
import numpy as np
import pandas as pd
import matplotlib
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
from datetime import datetime
# import warnings
```

Import Playstore Data

```
[178] # mounting drive
from google.colab import drive
drive.mount('/content/drive')

Drive already mounted at /content/drive; to attempt to forcibly remount, call drive.mount("/content/drive", force_remount=True).

[179] #load data
file_path = '/content/drive/MyDrive/Colab Notebooks/AlmaBetter/Modules/Python for data science/EDA Project/Play Store Data.csv'
psd_df= pd.read_csv(file_path)
```

```
[180] #First 5 rows of the play store dataframe
psd_df.head()
```

	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	19M	10,000+	Free	0	Everyone	Art & Design	January 7, 2018	1.0.0	4.0.3 and up
1	Coloring book moana	ART_AND_DESIGN	3.9	967	14M	500,000+	Free	0	Everyone	Art & Design;Pretend Play	January 15, 2018	2.0.0	4.0.3 and up
2	U Launcher Lite – FREE Live Cool Themes, Hide ...	ART_AND_DESIGN	4.7	87510	8.7M	5,000,000+	Free	0	Everyone	Art & Design	August 1, 2018	1.2.4	4.0.3 and up
3	Sketch - Draw & Paint	ART_AND_DESIGN	4.5	215644	25M	50,000,000+	Free	0	Teen	Art & Design	June 8, 2018	Varies with device	4.2 and up
4	Pixel Draw - Number Art Coloring Book	ART_AND_DESIGN	4.3	967	2.8M	100,000+	Free	0	Everyone	Art & Design;Creativity	June 20, 2018	1.1	4.4 and up

```
[181] # getting the last five rows
psd_df.tail()
```

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
10836	Sya9a Maroc - FR	FAMILY	4.5	38	53M	5,000+	Free	0	Everyone	Education	July 25, 2017	1.48	4.1 and up
10837	Fr. Mike Schmitz Audio Teachings	FAMILY	5.0	4	3.6M	100+	Free	0	Everyone	Education	July 6, 2018	1.0	4.1 and up
10838	Parkinson Exercises FR	MEDICAL	NaN	3	9.5M	1,000+	Free	0	Everyone	Medical	January 20, 2017	1.0	2.2 and up
10839	The SCP Foundation DB fr nn5n	BOOKS_AND_REFERENCE	4.5	114	Varies with device	1,000+	Free	0	Mature 17+	Books & Reference	January 19, 2015	Varies with device	Varies with device
10840	iHoroscope - 2018 Daily Horoscope & Astrology	LIFESTYLE	4.5	398307	19M	10,000,000+	Free	0	Everyone	Lifestyle	July 25, 2018	Varies with device	Varies with device

```
[183] # Print the Shape of the Two DataFrames
psd_df.shape

(10841, 13)

[184] # column names
psd_df.columns

Index(['App', 'Category', 'Rating', 'Reviews', 'Size', 'Installs', 'Type', 'Price', 'Content Rating', 'Genres', 'Last Updated', 'Current Ver', 'Android Ver'],
      dtype='object')
```

Data Cleaning cont....

Handling null value

```
# Removing the null value and Duplicate present in the data set
```

```
# this function give type, non null value, null value, unique value and non null value %  
def playstoreinfo():  
    temp = pd.DataFrame(index = psd_df.columns)  
    temp['Data Type'] = psd_df.dtypes  
    temp['non null value'] = psd_df.count()  
    temp['null value'] = psd_df.isnull().sum()  
    temp['unique values'] = psd_df.nunique()  
    temp['non null value percentage'] = psd_df.isnull().mean()  
    return temp  
playstoreinfo()
```

	Data Type	non null value	null value	unique values	non null value percentage
App	object	10841	0	9660	0.000000
Category	object	10841	0	34	0.000000
Rating	float64	9367	1474	40	0.135965
Reviews	object	10841	0	6002	0.000000
Size	object	10841	0	462	0.000000
Installs	object	10841	0	22	0.000000
Type	object	10840	1	3	0.000092
Price	object	10841	0	93	0.000000
Content Rating	object	10840	1	6	0.000092
Genres	object	10841	0	120	0.000000
Last Updated	object	10841	0	1378	0.000000
Current Ver	object	10833	8	2832	0.000738
Android Ver	object	10838	3	33	0.000277

Android Ver - There are 3 null Value

```
# Null value in Android Ver column  
psd_df[psd_df['Android Ver'].isnull()]
```

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
4453	[substratum] Vacuum: P	PERSONALIZATION	4.4	230	11M	1,000+	Paid	\$1.49	Everyone	Personalization	July 20, 2018	4.4	NaN
4490	PI Dark [substratum]	PERSONALIZATION	4.5	189	2.1M	10,000+	Free	0	Everyone	Personalization	March 27, 2018	1.1	NaN
10472	Life Made WI-Fi Touchscreen Photo Frame		1.9	19.0	3.0M	1,000+	Free	0	Everyone	NaN	February 11, 2018	1.0.19	4.0 and up

```
# delete Null value in Android Ver column  
psd_df = psd_df[psd_df['Android Ver'].notna()]  
  
# New shape of dataframe after deleting null value in android ver columns  
psd_df.shape  
  
(10838, 13)
```

Current Ver - 8 Null value in this columns

```
# Null value in Current Ver  
psd_df[psd_df['Current Ver'].isnull()]
```

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
15	Learn To Draw Kawaii Characters	ART_AND_DESIGN	3.2	55	2.7M	5,000+	Free	0	Everyone	Art & Design	June 6, 2018	NaN	4.2 and up
1553	Market Update Helper	LIBRARIES_AND_DEMO	4.1	20145	11k	1,000,000+	Free	0	Everyone	Libraries & Demo	February 12, 2013	NaN	1.5 and up
6322	Virtual DJ Sound Mixer	TOOLS	4.2	4010	8.7M	500,000+	Free	0	Everyone	Tools	May 10, 2017	NaN	4.0 and up
6803	BT Master	FAMILY	NaN	0	222k	100+	Free	0	Everyone	Education	November 6, 2016	NaN	1.6 and up
7333	Dots puzzle	FAMILY	4.0	179	14M	50,000+	Paid	\$0.99	Everyone	Puzzle	April 18, 2018	NaN	4.0 and up
7407	Calculate My IQ	FAMILY	NaN	44	7.2M	10,000+	Free	0	Everyone	Entertainment	April 3, 2017	NaN	2.3 and up
7730	UFO-CQ	TOOLS	NaN	1	237k	10+	Paid	\$0.99	Everyone	Tools	July 4, 2016	NaN	2.0 and up
10342	La Fe de Jesus	BOOKS_AND_REFERENCE	NaN	8	658k	1,000+	Free	0	Everyone	Books & Reference	January 31, 2017	NaN	3.0 and up

```
# delete NaN value in Current Ver column  
psd_df = psd_df[psd_df['Current Ver'].notna()]  
  
# new shape of dataframe  
psd_df.shape  
  
(10830, 13)
```



Rating - There are 1470 NaN values

```
psd_df[psd_df["Rating"].isnull()]
```

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content	Rating	Genres	Last Updated	Current Ver	Android Ver
23	Mcqueen Coloring pages	ART_AND_DESIGN	NaN	61	7.0M	100,000+	Free	0	Everyone	Art & Design;Action & Adventure		March 7, 2018	1.0.0	4.1 and up
113	Wrinkles and rejuvenation	BEAUTY	NaN	182	5.7M	100,000+	Free	0	Everyone 10+	Beauty		September 20, 2017	8.0	3.0 and up
123	Manicure - nail design	BEAUTY	NaN	119	3.7M	50,000+	Free	0	Everyone	Beauty		July 23, 2018	1.3	4.1 and up
126	Skin Care and Natural Beauty	BEAUTY	NaN	654	7.4M	100,000+	Free	0	Teen	Beauty		July 17, 2018	1.15	4.1 and up
129	Secrets of beauty, youth and health	BEAUTY	NaN	77	2.9M	10,000+	Free	0	Mature 17+	Beauty		August 8, 2017	2.0	2.3 and up
...
10824	Cardio-FR	MEDICAL	NaN	67	82M	10,000+	Free	0	Everyone	Medical		July 31, 2018	2.2.2	4.4 and up
10825	Naruto & Boruto FR	SOCIAL	NaN	7	7.7M	100+	Free	0	Teen	Social		February 2, 2018	1.0	4.0 and up
10831	payermonstationnement.fr	MAPS_AND_NAVIGATION	NaN	38	9.8M	5,000+	Free	0	Everyone	Maps & Navigation		June 13, 2018	2.0.148.0	4.0 and up
10835	FR Forms	BUSINESS	NaN	0	9.6M	10+	Free	0	Everyone	Business		September 29, 2016	1.1.5	4.0 and up
10838	Parkinson Exercises FR	MEDICAL	NaN	3	9.5M	1,000+	Free	0	Everyone	Medical		January 20, 2017	1.0	2.2 and up
1470 rows × 13 columns														

1470 rows × 13 columns

- The `Rating` column contains 1470 NaN values which approximately 13.5% of the rows in the entire dataset. Deleting all these rows is not good as we will lose large amount of data. It impacts the final quality of the analysis.
- The NaN values will be replaced by the mean or median of the rest values in the `Rating` column.

```
112] # Finding Median of all non NaN values of rating column
median_rating = psd_df[~psd_df['Rating'].isnull()]['Rating'].median()
median_rating

4.3

113] # Replacing the NaN values in the 'Rating' column with its median value
psd_df['Rating'].fillna(value=median_rating, inplace = True)
```

Type - There is 1 NaN value in this columns

```
[ ] psd_df[psd_df['Type'].isnull()]
```

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content	Rating	Genres	Last Updated	Current Ver	Android Ver
9148	Command & Conquer: Rivals	FAMILY	NaN	0	Varies with device	0	NaN	0	Everyone 10+	Strategy	June 28, 2018	Varies with device	Varies with device	

The `Type` column has two types of entities, i.e. `Free` and `Paid`. If the type is `Paid` then the price will be printed in the `Price` column; else, it will show as '0'. In this case, the price of the app is printed as '0', which means the app is of type-free. Hence we can replace this NaN value with `Free`.

```
[ ] # Replacing the NaN value in 'Type' column corresponding to row index 9148 with 'Free'
psd_df.loc[9148, 'Type'] = 'Free'
```


Data Cleaning cont....

Handling duplicates values

Handling the duplicates in the App column

```
[115] # Duplicate values in App column
psd_df['App'].value_counts()
```

```
ROBLOX          9
CBS Sports App - Scores, News, Stats & Watch Live      8
Candy Crush Saga  7
8 Ball Pool      7
ESPN             7
..
Meet U - Get Friends for Snapchat, Kik & Instagram    1
U-Report        1
U of I Community Credit Union                       1
Waiting For U Launcher Theme                         1
iHoroscope - 2018 Daily Horoscope & Astrology         1
Name: App, Length: 9649, dtype: int64
```

```
[116] #deleting the duplicate values from the 'App' column
psd_df.drop_duplicates(subset = 'App',inplace = True)
```

```
# shape
psd_df.shape
```

```
(9649, 13)
```

```
[117] # Checking whether the duplicates in the 'App' column are removed or not
psd_df[psd_df['App'] == 'ROBLOX']
```

App	Category	Rating	Reviews	Size	Installs	Type	Price	Content	Rating	Genres	Last Updated	Current Ver	Android Ver
1653 ROBLOX	GAME	4.5	4447388	67M	100,000,000+	Free	0	Everyone	10+	Adventure>Action & Adventure	July 31, 2018	2.347.225742	4.1 and up

```
[118] # We have successfully handled all the duplicate values in the App column.
```

Changing the data type of last updated column from string to date time

```
[121] psd_df['Last Updated'] = pd.to_datetime(psd_df['Last Updated'])
```

```
psd_df.head()
```



	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	19M	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	14M	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	8.7M	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	25M	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	2.8M	100,000+	Free	0	Everyone		Art & Design;Creativity	2018-06-20	1.1	4.4 and up

Data Cleaning

Changing data type



```
[208] #Changing the datatype of the Price column from string to float.
[209] psd_df['Price'].value_counts()

0      8896
$0.99   143
$2.99   124
$1.99    73
$4.99    70
...
$18.99    1
$389.99    1
$19.90    1
$1.75     1
$1.04     1
Name: Price, Length: 92, dtype: int64
```

To convert this column from string to float, we must first drop the \$ symbol from the all the values. Then we can assign float datatype to those values.

```
[210] # Creating a function remove-dollar which dropps the $ symbol if it is present and returns the output which is of float datatype.
def remove1(val):
    if '$' in val:
        return float(val[1:])
    else:
        return float(val)

[211] # The drop_dollar funtion applied to the price column
psd_df['Price'] = psd_df['Price'].apply(lambda x: remove1(x))
```

[126] psd_df.head()

	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	19M	10,000+	Free	0.0	Everyone	Art & Design	2018-01-07	1.0.0	4.0.3 and up
1	Coloring book moana	ART_AND_DESIGN	3.9	967	14M	500,000+	Free	0.0	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	8.7M	5,000,000+	Free	0.0	Everyone	Art & Design	2018-08-01	1.2.4	4.0.3 and up
3	Sketch - Draw & Paint	ART_AND_DESIGN	4.5	215644	25M	50,000,000+	Free	0.0	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	2.8M	100,000+	Free	0.0	Everyone	Art & Design;Creativity	2018-06-20	1.1	4.4 and up

[127] psd_df[psd_df['Price'] != 0].head()

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
234	TurboScan: scan documents and receipts in PDF	BUSINESS	4.7	11442	6.8M	100,000+	Paid	4.99	Everyone	Business	2018-03-25	1.5.2	4.0 and up
235	Tiny Scanner Pro: PDF Doc Scan	BUSINESS	4.8	10295	39M	100,000+	Paid	4.99	Everyone	Business	2017-04-11	3.4.6	3.0 and up
427	Puffin Browser Pro	COMMUNICATION	4.0	18247	Varies with device	100,000+	Paid	3.99	Everyone	Communication	2018-07-05	7.5.3.20547	4.1 and up
476	Moco+ - Chat, Meet People	DATING	4.2	1545	Varies with device	10,000+	Paid	3.99	Mature 17+	Dating	2018-06-19	2.6.139	4.1 and up
477	Calculator	DATING	2.6	57	6.2M	1,000+	Paid	6.99	Everyone	Dating	2017-10-25	1.1.6	4.0 and up

Data Cleaning

Changing data type

changing the data type of installs from string to integer

```
[214] psd_df['Installs'].value_counts()
```

```
1,000,000+    1416
100,000+      1112
10,000+       1029
10,000,000+   937
1,000+        886
100+          709
5,000,000+   607
500,000+     504
50,000+      468
5,000+       467
10+          384
500+         328
50+          204
50,000,000+  202
100,000,000+ 188
5+           82
1+           67
500,000,000+ 24
1,000,000,000+ 20
0+           14
0            1
Name: Installs, dtype: int64
```

We need to remove '+', ',' symbol from all the entities, to convert Installs column from string datatype to integer datatype

```
[215] # Creating a function convert_plus which drops the '+' symbol if it is present and returns the output which is of integer datatype.
def remove_plus(val):
    if '+' and ',' in val:
        new = int(val[:-1].replace(',',''))
        return new
    elif '+' in val:
        new1 = int(val[:-1])
        return new1
    else:
        return int(val)
```

```
[216] # the remove_plus function applied to the dataframe
psd_df['Installs'] = psd_df['Installs'].apply(lambda x: remove_plus(x))
```

```
[217] psd_df.head()
```

	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	19M	10000	Free	0.0	Everyone	Art & Design	2018-01-07	1.0.0	4.0.3 and up
1	Coloring book moana	ART_AND_DESIGN	3.9	967	14M	500000	Free	0.0	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	8.7M	5000000	Free	0.0	Everyone	Art & Design	2018-08-01	1.2.4	4.0.3 and up
3	Sketch - Draw & Paint	ART_AND_DESIGN	4.5	215644	25M	50000000	Free	0.0	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	2.8M	100000	Free	0.0	Everyone	Art & Design;Creativity	2018-06-20	1.1	4.4 and up

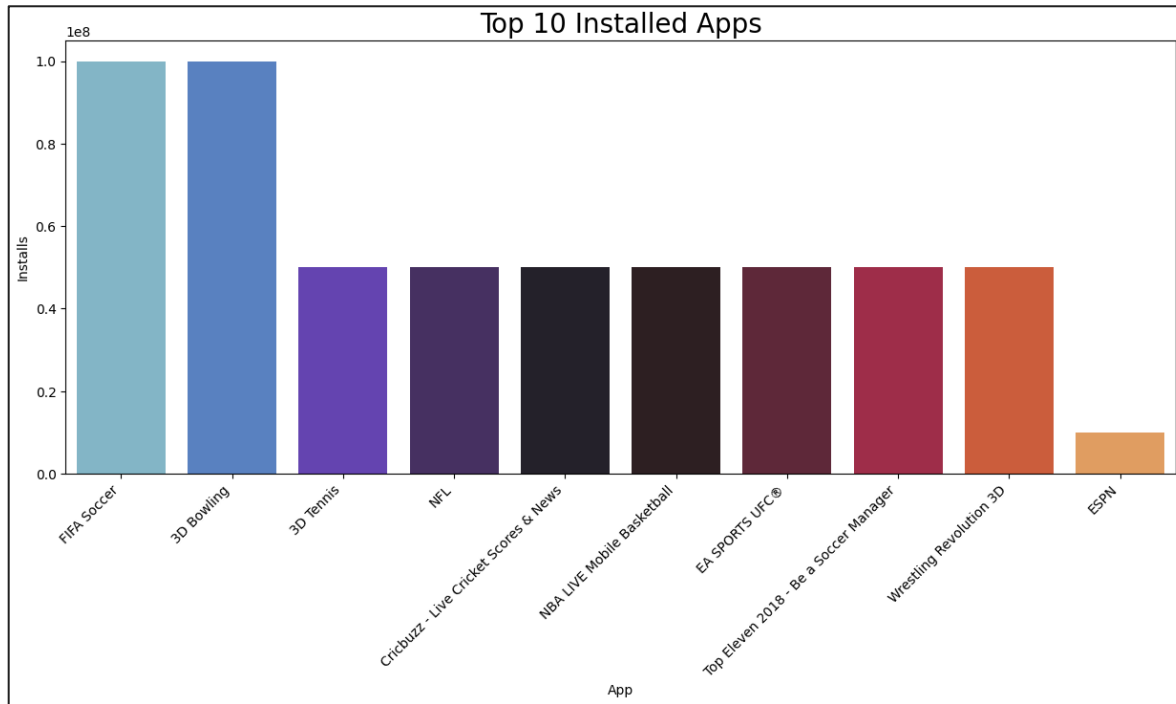
```
[139] #changing data type of reviews column from string to integer
```

```
[140] psd_df['Reviews'] = psd_df['Reviews'].astype(int)
psd_df.head()
```

	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	19.0	10000	Free	0.0	Everyone	Art & Design	2018-01-07	1.0.0	4.0.3 and up
1	Coloring book moana	ART_AND_DESIGN	3.9	967	14.0	500000	Free	0.0	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	5000000	Free	0.0	Everyone	Art & Design	2018-08-01	1.2.4	4.0.3 and up
3	Sketch - Draw & Paint	ART_AND_DESIGN	4.5	215644	25.0	50000000	Free	0.0	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	100000	Free	0.0	Everyone	Art & Design;Creativity	2018-06-20	1.1	4.4 and up

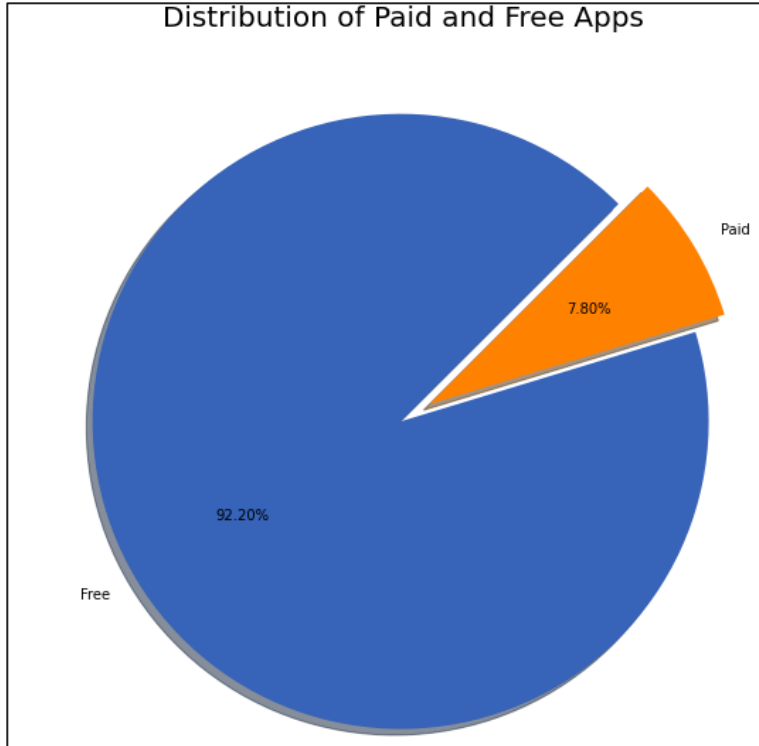
Data Visualization

Top 10 installed apps in any category



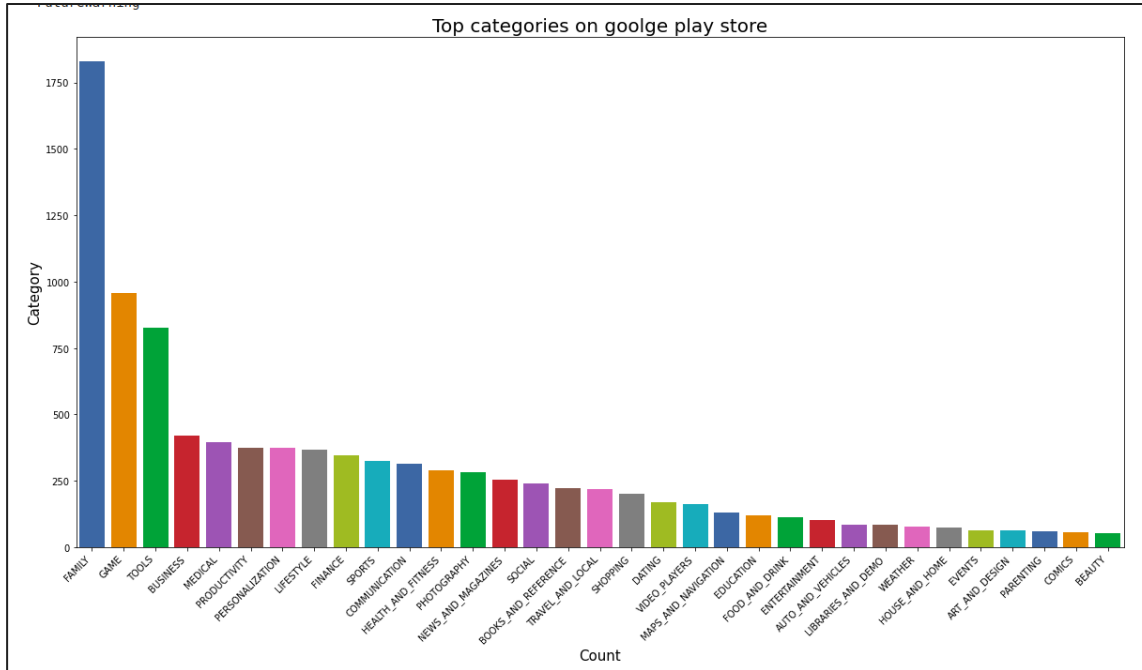
* From the above graph we can see that in the Sports category FIFA Soccer and 3D Bowling has the highest installs.

Distribution of paid apps and free apps



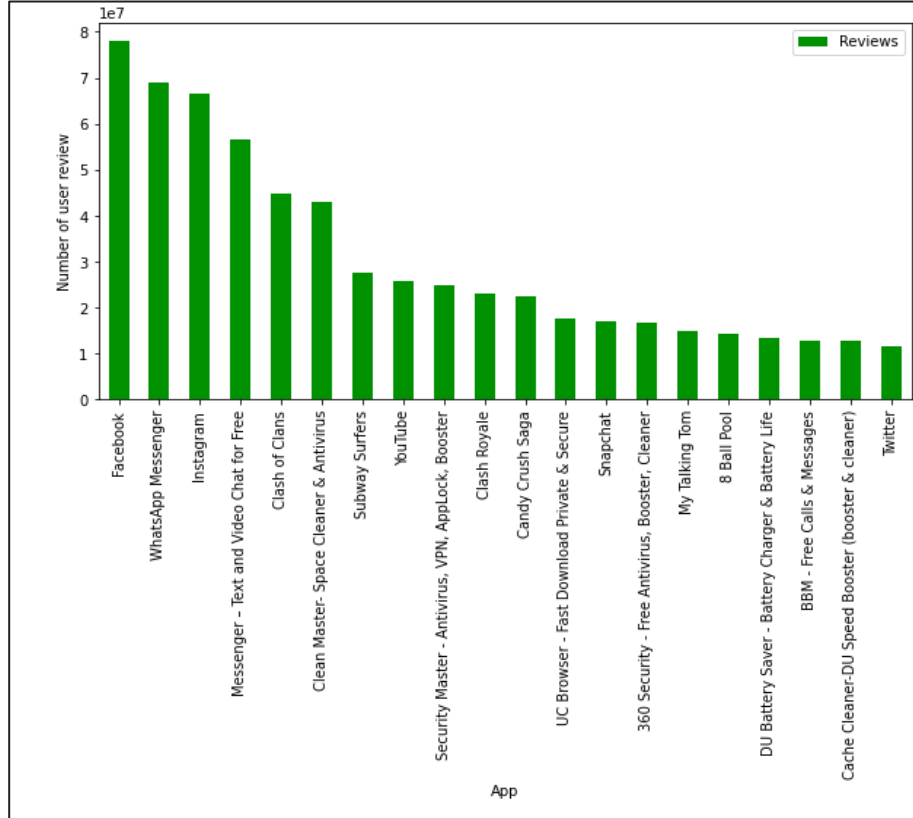
* Free apps in play store are **92.20%** and
Paid apps in play store are **7.80%**

Top Categories Apps in Google Play store



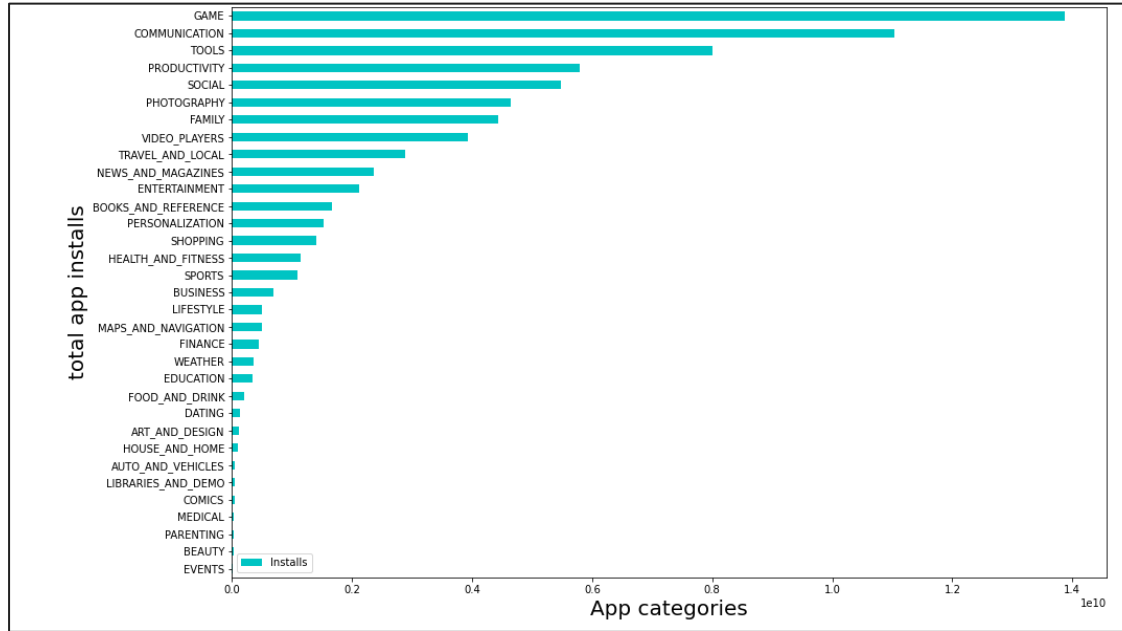
* So there are all total 33 categories in the dataset From the above output we can come to a conclusion that in play store most of the apps are under` FAMILY & GAME` category and least are of `EVENTS & BEAUTY` Category.

Which apps have the highest number of review?



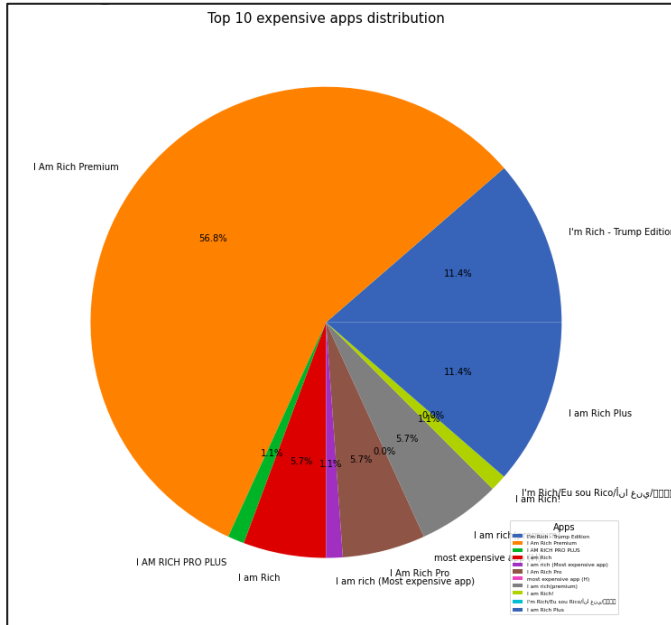
- * Top 20 apps with the highest number of review
- * Facebook has the highest number of user review

Which app's have maximum number of installs in any category?



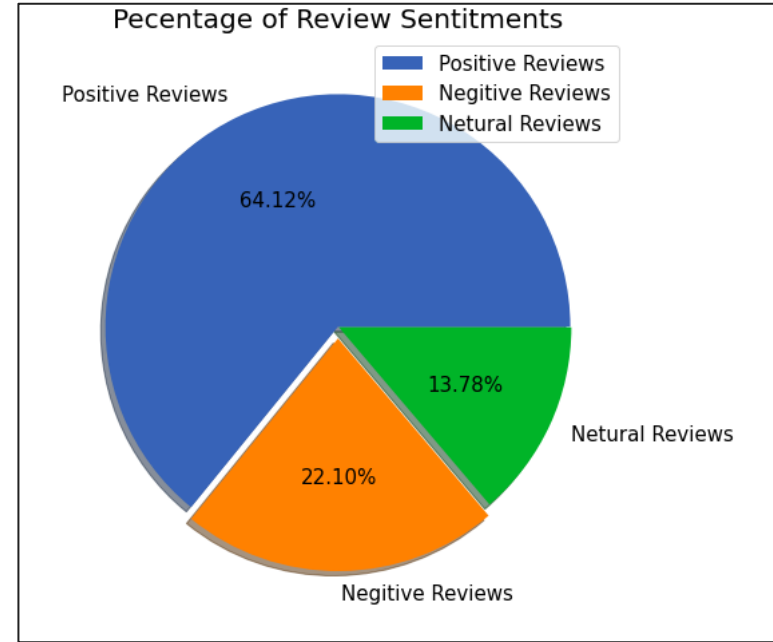
This tells us the category of apps that has the maximum number of installs. The `Game`, `Communication` and `Tools` categories have the highest number of installs compared to other categories of apps.

Top 10 expensive apps in the play store



* From the above graph we can interpret that the **I am Rich Premium** app is the **most expensive app** in the play store.

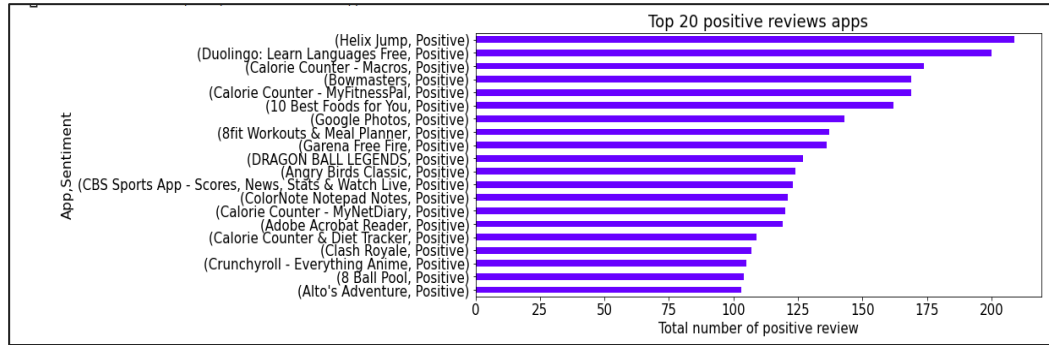
Percentage of Review Sentiments



Findings:

1. Positive reviews are **64.12%**
2. Negative reviews are **22.10%**
3. Neutral reviews are **13.78%**

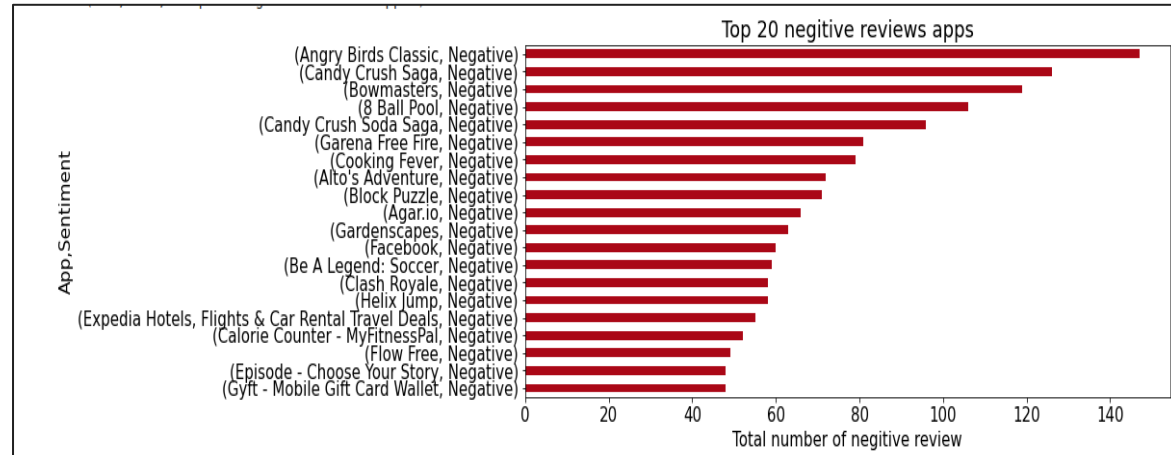
Positive review



* It is clear that Helix jump as max no. of positive review

Negative review

* It is clear that Angry Birds Classic has got highest negative reviews



Conclusion

In the initial phase, we focused more on the problem statements and data cleaning, in order to ensure that we give them the best results out of our analysis.

- * In the Sports category FIFA Soccer and 3D Bowling has the highest number of installs.
- * Percentage of free apps = 92.20%
- * Maximum apps in the play store are from Family category
- * Category with the highest number of installs: Game
- * Most popular app in the Play Store based on the number of reviews: Facebook
- * I am Rich Premium app is the most expensive app in the play store.
- * Overall percentage of review sentiment in which Positive sentiment count is 64%, Negative 22% and Neutral 14%.
- * Helix Jump has the highest number of positive reviews
- * Angry Birds Classic has the highest number of negative reviews.

Q&A