PLAY STORE APP REVIEW ANALYSIS

GAURAV BISHT

DATA SCIENCE TRAINEE

ALMA BETTER

CONCEPT

Mobile apps are everywhere. According to state counter global stats over 71.86% of mobile operating system market is shared by android worldwide - 'july 2022', so clearly andriod dominates in global market share.They are trouble free to create, profitable or booming & are very useful and because of all these factors, apps are being developed progressively. In this project, We will do a comprehensive analysis of the Android app market by comparing over ten thousand apps in Google Playstore all across different categories. We'll look for different insights in the data to conceive strategies to operate and grasp info. The data for this project was taken from the Google Play Store website.

1. ABOUT

Google play also branded as the google play store is a digital app which is run by google and is widely used by billions of people, downloading there favourite app according to their need, it's a app used for certified devices running over Andriod OS as well as Chrome OS, over 3+ million apps are there in play store, The total revenue of the platform is over $11.2 billion (82,000 crores) in 2019.

2. INSPECTING THE DATASET

A) Play Store Data.csv - This **data.csv** contains all the required information about the applications present in google play store app, as it contains 13 attributes or

bascially subsets that describes the given app. It has 10841 apps out of which 9659 are unique.

B) User Review.csv – This **user.csv** files contains 4 attributes i.e. App, Review, Sentiment Category & Sentiment Score.

1. ATTRIBUTES (googleplaystore.csv)

Above file holds all the details of the apps on Google Play.

Here i have taken 9 important features out of 13 that describe a given app.

App: Name of the app

Rating : Gives current average rating (out of 5) of the app on Google Play

Reviews : Number of user reviews given on the app

Category : Category of the app. for examples are: FAMILY, SOCIAL, TOOLS, COMMUNICATION etc.

Size : Size of the app in MB (megabytes), KB (kilobytes), GB (gigabytes)

Type : paid or free

Price : Price of the app in US$ on Google Play Store

Last Updated : Date on which the app was last updated on Google

PlayInstalls : NO. of times any given app is or was downloaded/installed from Google Play.

Android Version : It tells you on which version the app will run or will work.

Current Version : It will the current version of the app

Content Rating : Sutable age group for different Applications.

Genre : App can belong to any version.

2. ATTRIBUTES (user_reviews.csv)

Above file holds 100 reviews for each app, passed and accepted via sentiment investigator engine & it's sentiment score.

Review : Preliminary processed user review text

App : Name of the app on which the user review was provided. Matches the app column of the apps.csv file

Sentiment Category : User review - Positive, Negative or Neutral

Sentiment Score : Sentiment score of the user review.

3. COMPLICATION STATEMENT

Play store is a digital marketplace app which is used to download a variety of apps for andriod smartphones, as smartphones are increasing day by day and peoples need and requirement over apps is also increasing accordingly, also it's a very profitable & money making market for app developers.

As peoples demand changes with time and trend according to their need and comfort so developers need to know this thing too, like social, entertainment and games categories also are in hype.

So it's always important to find out what type of app is required in market and what's trending as what people are downloading and liking the most before developing the app,

Also it's important to keep in mind that how size, rating, type, price & review affect the sentiment of apps and user

4. ASSEMBLING THE DATASET (ANALYSIS METHODOLOGY)

Loading the Dataset – We are given two different dataset play store dataset & user review of apps, so here we'll import various python inbuild libraries as – numpy & pandas

Importing libraries – Using Matplotlib, seaborn & plotly to work on the dataset.

Data cleaning – Here we'll remove null values, find and remove identical or duplicate data remove duplicate rows also filling the missing values with mode & numerical values with median, fixing prize, size and other attributes.

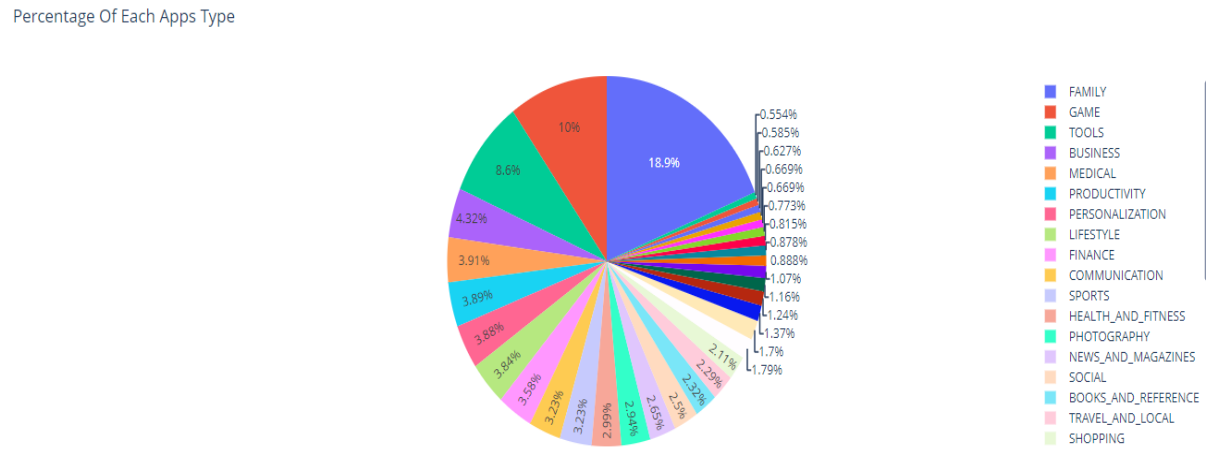
Exploratory Data Analysis – Analyzing the data and using different statical graphs and also data visualizations methods to make it more presentable

5. EXPLORATORY DATA ANALYSIS

After loading the dataset we performed this method by comparing our target variable. This process helped us figuring out various aspects and relationships among the diff.

Categories and their respective attributes or subtitles. It gave us a better idea of which category behaves in which manner compared to the each other gives a better corelationship.

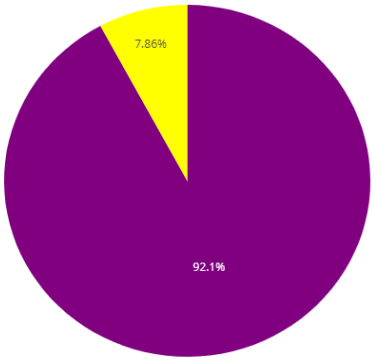
Figure 1. Number of apps available for download by category.




```
x = data_df.groupby(['Category'],as_index=False)['App'].count()
px.pie(x,values='App',names='Category',title='Percentage Of Each Apps Type').
show()

print('Top 10 Comman Apps Category in Google Play Store','\n'
      ,x.sort_values(by ='App',ascending =False))
x.head()
```

Figure 2. Percentage of apps in paid (yellow) & free (purple) category



```
x = data_df.groupby('Type',as_index=False)['App'].count()
print('\n'*2+'Most Of Apps In Google Play are Free'+'\n'*2)
px.pie(x,values='App',names='Type',color='Type',
       color_discrete_map={'Free':'purple','Paid':'yellow'}).show(
)
```

Figure 3. Heirarchical data spanning outward radically from the roots to leaves (just like a doughnut)


```
amount = data_df.groupby(['Type','Category'],as_index=False)['App'
].count()
px.sunburst(amount, values='App', path=['Type','Category'], title=
'Amount Of Apps in Paid and Free Category', color='Category')
```

6. Data Analysis & Visualization

In this data set there are various features that can be usedto analyse the data set. In this section we will be analysing different features to find which feature determines whether an app will be successful or not. The top five genres of the google play store include Tools, Entertainment, Education, Business and Medical. From figure 1 we can see that the most number of apps in Google Play store belongs to the categories of Family and game. This shows that apps that belongs to gaming and Family category are more common and apps in this category have high chances of being successful. Figure 2 shows the % of free & paid app, also figure 3 shows plotly sunburst plots to visualize

hierarchical data spanning outwards radically from the roots to leaves (just like a doughnut).

Here the sectors are determined by the categories and the inner ring tells about the % of free & paid apps.

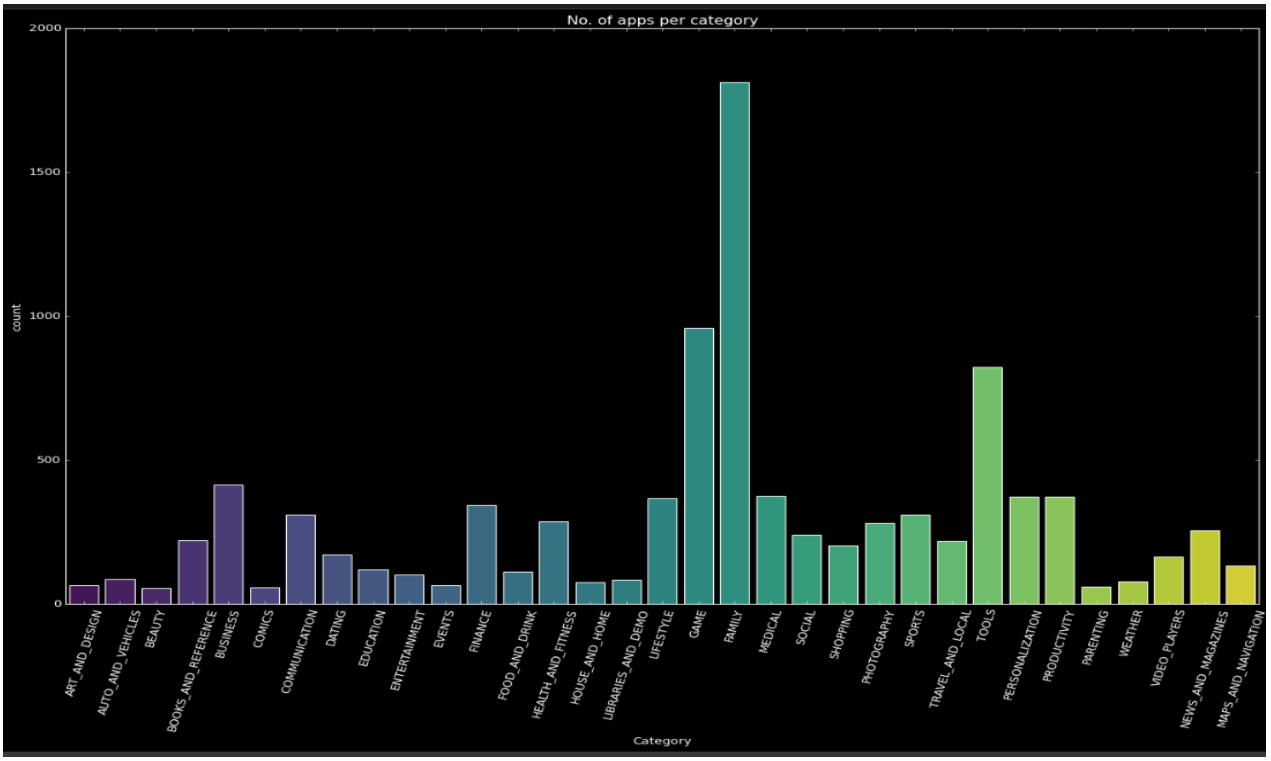
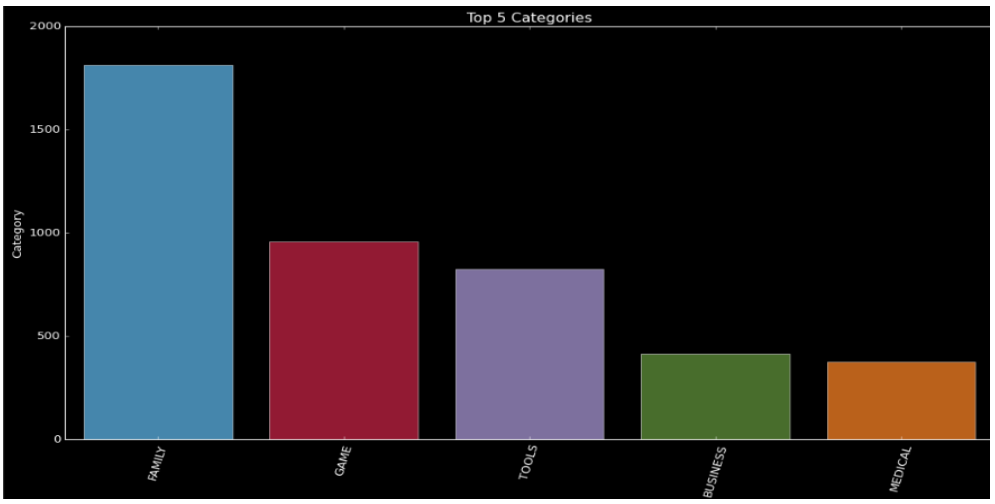


Figure 4. Number of apps category wise

```
plt.figure(figsize=(20,12))
z = sns.countplot(data_df['Category'], palette= "viridis" )
z.set_xticklabels(z.get_xticklabels(), rotation = 75)
plt.title('No. of apps per category')
plt.style.use('dark_background')
plt.show()
```

Here by looking at the figure 4 we can see that "Genre - Family" has the highest no. of applications hold & 'Beauty" has the lowest.

Top 5 categories holding max no. of applications are shown in figure down below which are as follows - Family, Games, Tool, Business & Medical.



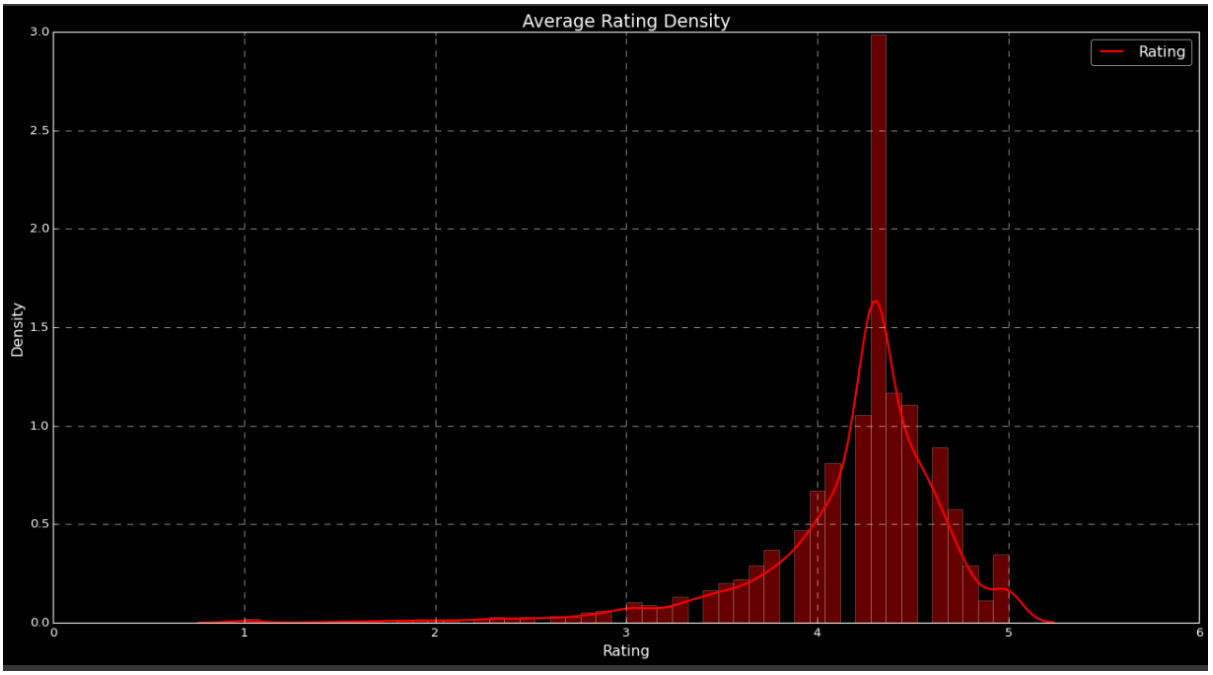


Figure 5. Average rating of apps in play Store

```
overall_rating = data_df.Rating.mean()
print('The average rating in playstore is {:.2f} '.format(rating_2))
```

```
plt.figure(figsize=(20,10))
sns.distplot(data_df['Rating'],color = 'red')

plt.style.use('tableau-colorblind10')
```

```
plt.legend(['Rating'])
plt.title('Average Rating Density')
plt.show()
```

While making the comprehensive analysis about the average rating of apps in play store app we get to know that the average mean of rating is 4.25 & also their are 265 apps whose ratingis = 5. The above figure 5 shows the representative graph of rating of apps in play store.

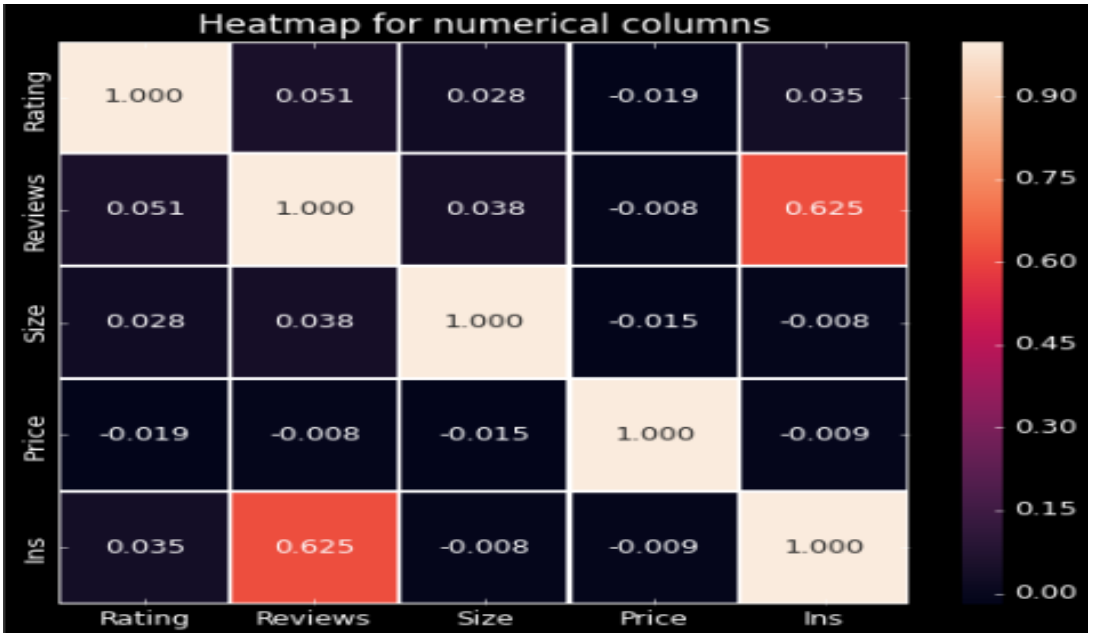


Figure 6. Heatmap for numerical columns

```
data_df['Android Ver'].value_counts()
sns.heatmap(data_df.corr(), annot = True, linewidths=1.0, fmt=".3f")
plt.title("Heatmap for numerical columns", size=15)
```

From figure 6 we can see that installs and reviews have the strongest inverse correlation 0.625. This is reasonable because popular apps tends to get more number of reviews.There is no correlation found between installs and other features like size,rating,number of installs and price. There is no correlation between rating and price also. Since installs parameter is independent and not correlated to any other parameters, we must only use installs to show the popularity of an app.

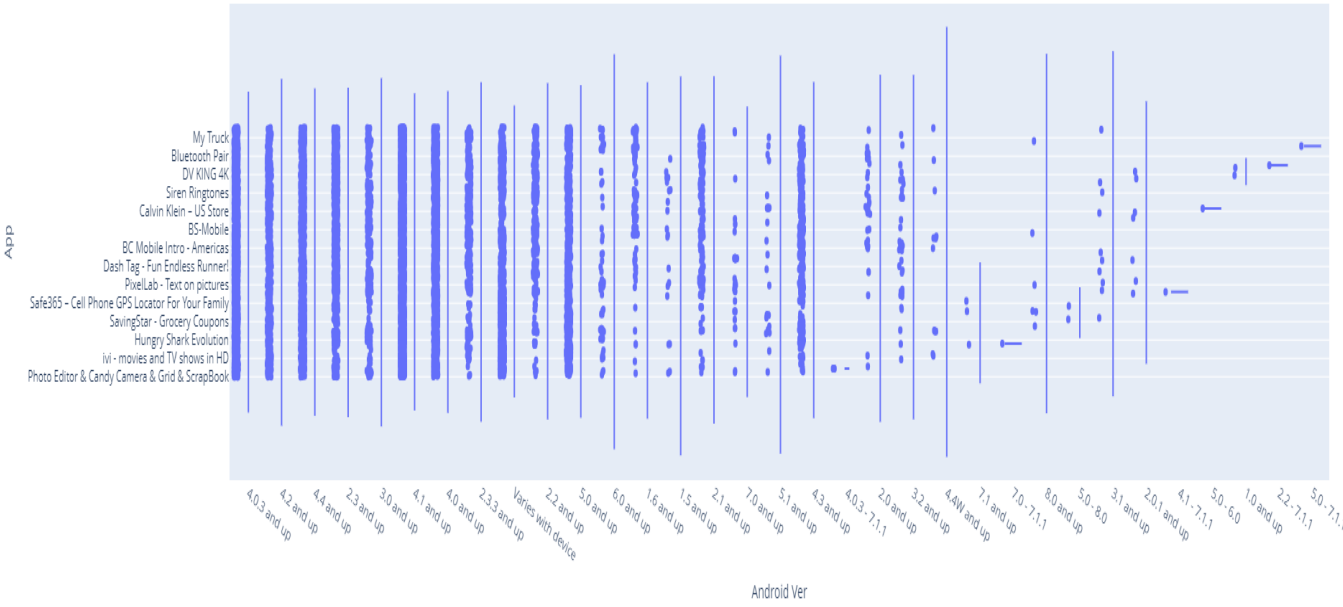


Figure 7. Apps running over different Android Version

```
a = px.violin(data_df, x='Android Ver', y = 'App', points='all')
a.show()
```

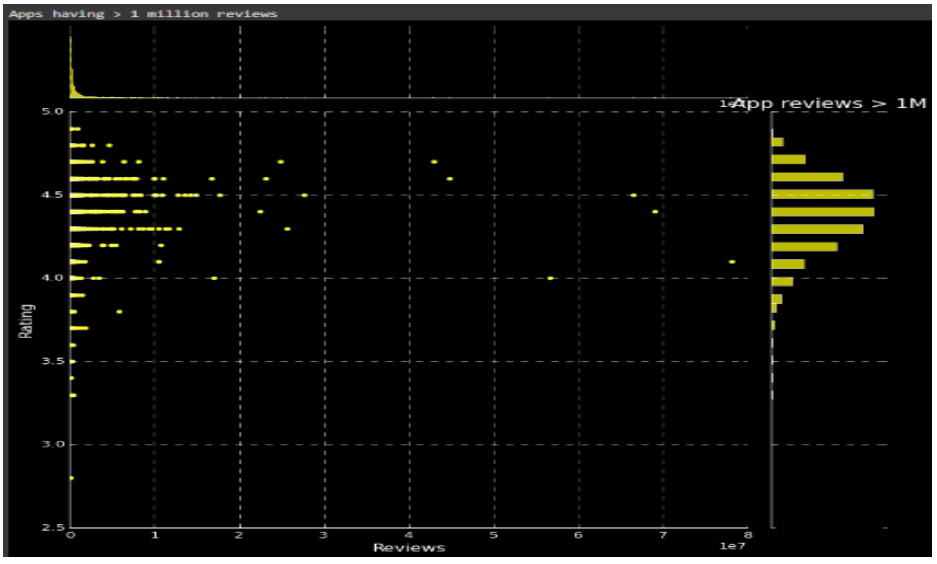


Figure 8.1 Review vs Rating

In Play Console, you can see an overview of your app's ratings, individual user reviews, and clustered data about your app's reviews. Users can rate your app on Google Play

with a star rating and review. Users can only rate an app once, but they can update their rating or review at any time.

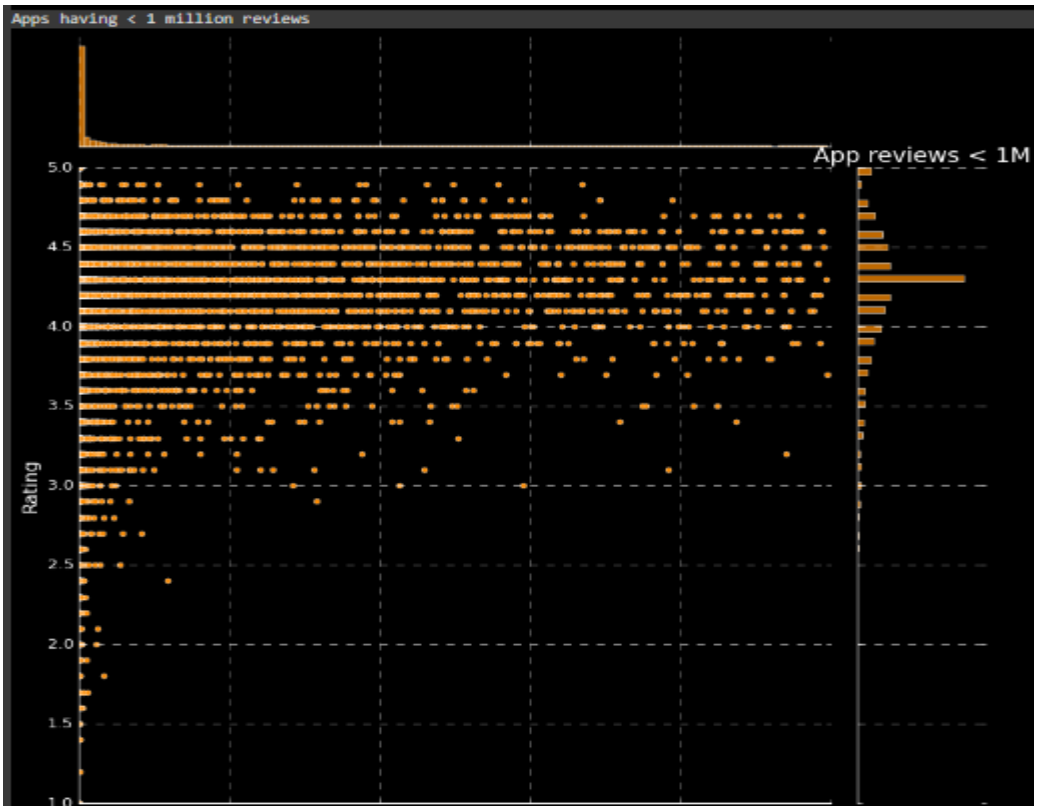


Figure 8.2 Review vs Rating

```
print('Apps having > 1 million reviews')
sns.jointplot(x='Reviews',y='Rating',data =data_df[data_df['Reviews'] > 100000],color='yell
ow', height=10)

plt.title('App reviews > 1M')
plt.style.use('dark_background')
plt.show()

print('Apps having < 1 million reviews')
sns.jointplot(x='Reviews',y='Rating',data =data_df[data_df['Reviews']<100000],color ='darko
range', height=10)

plt.title('App reviews < 1M')
plt.style.use('dark_background')
plt.show()
```

From the figure 8.1 & 8.2 the above observation shows that the the most reviewed apps are likely to have better rating,

Hence the conclusion by the above figures is that the apps which are having higher rating also have the highest no. of reviews or vice versa.

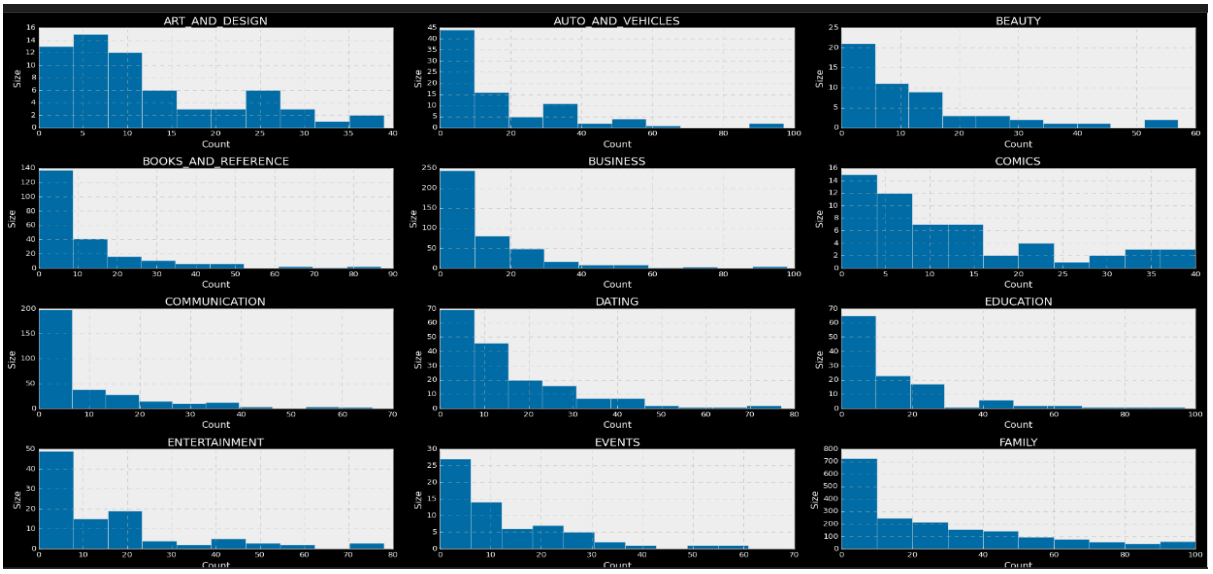


Figure 9.1 Different Categories & their sizes

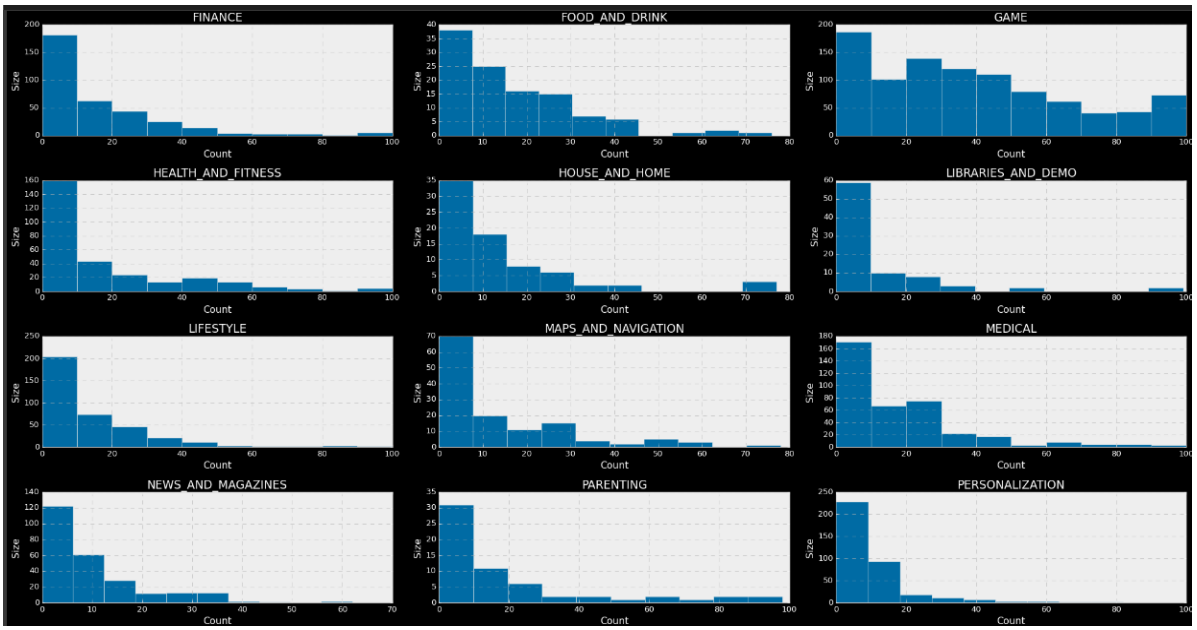


Figure 9.2

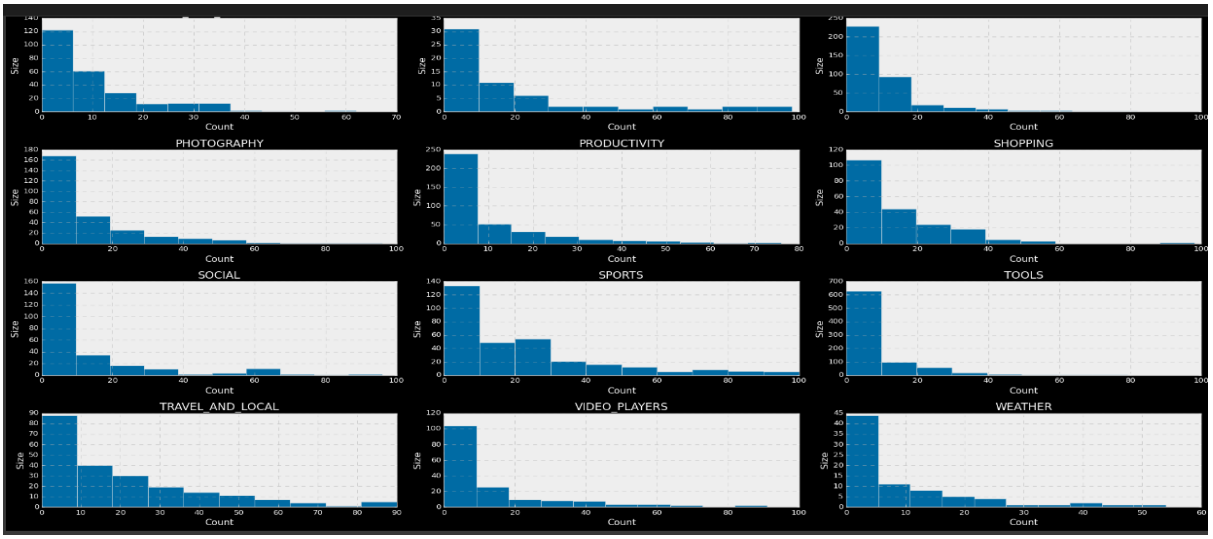


Figure 9.3

Clearly by the above figures 9.1, 9.2, 9.3 we can see that most of the apps are low to mid range in size as compared to games where apps are slightly higher in size.

```
ncount=1
plt.figure(figsize=(25,35))
for x in np.unique(data_df.Category.values):
    plt.subplot(11,3,ncount)
    plt.hist(data_df[data_df.Category == x].Size.values)
    plt.title(x)
    plt.ticklabel_format(useOffset=False, style='sci')
    plt.xlabel("Count")
    plt.ylabel("Size")
    ncount = ncount + 1
plt.tight_layout()
```

7. Technologies Used

PYTHON: Python is a high-level interpreted language that supports different platforms like Windows, Linux, Mac, Rasberry Pi, etc. Python can be used for creating web applications, database systems, handle big data, perform complex mathematical calculations. Python can be treated in an object-oriented, functional or procedural way.

GOOGLE COLAB: Colab notebooks allow you to combine executable code and rich text in a single document, along with images, HTML, LaTeX and more. When you create your own Colab ...Colab, or "Colaboratory", allows you to write and execute Python in your browser, with Zero configuration required, Access to GPUs free of charge & Easy sharing.

PYTHON PACKAGES: Following are some of the python packages used in this project.

Matplotlib: It is a 2D based plotting package that provides required modules and functions. A developer can customize font properties, styles, axes properties, etc.

Pandas: It is used for data analysis and manipulation. Pandas can convert data structures and dataset formats to data frames on which operations like loading data, rename attributes, mapping, crosstab, sub-data frames, plotting, etc. can be performed.

NumPy: It provides structures for multiple dimensional array objects and tools for related operations. NumPy is usually used for high performance scientific computational tasks.

8. INFERENCES & CONCLUSION

After analyzing and exploring the data we got to know that this dataset gives us so many intresting insights and useful info about the play store app & thus it will deliver useful info to the customers as well as will direct the developers to get new apps in market and to popularize the product. Also while using the visualization libraries i got to know about them alot as each library has it's own strength & weaknesses which are as follows. –

MATPLOTLIB - It's great for distribution analysis but low - level interface.

SEABORN - It has simple & short code but it dosen't have wide varities as matplotlib.

PLOTLY - It's very interactive, has versatile graphics & high - level interface. Bascially much better and revised version of matplotlib & seaborn.

Hence, this data set contains a large amount of data that can be used for various purposes. Currently, the this data set can be used for future developers and Google plays store team to glance at the google play store market and what categories of the apps should be made to keep google play store popular in the future. It can be used to improve business values and google play store in general.

Some points to talk about –

Family & games app have the highest share ratio in play store app.

from the above dataset we got to know that over 92.1% app in play store are free & 7.86% are paid.
Most of apps which are built as social media or communication apps have highest reviews follwed by 'game' & 'family'

top 5 categories in Google play store are - family, games, tools, business, medical.

The average rating in playstore is 4.25 & there are over 265 apps in the play store whose rating is 5.0

Most of the apps are low to mid range in size as compared to games where apps are slightly higher in size.

There is high positive corelationship between reviews and install which is 0.065.

The most reviewed apps are likely to be better rated in play store app.

9. REFERENCES

1. Python learning almabetter