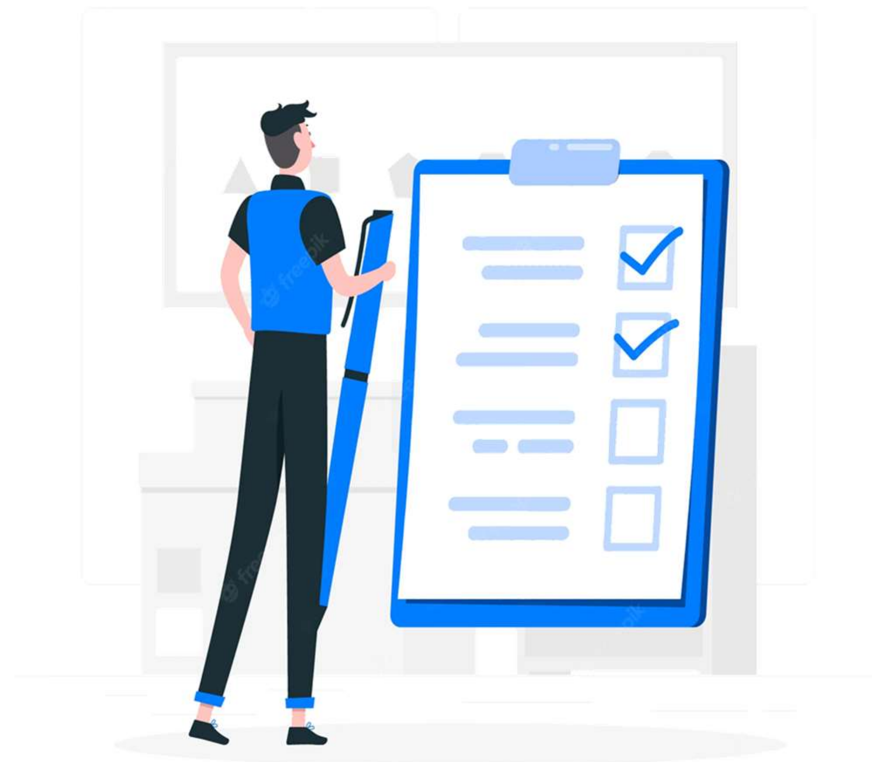
Capstone Project

Play Store App Review Analysis

Ajay Pandey
Manjiri Kulkarni
Prasad Wagh
Shahrukh Ahmad

Content:

- Problem Statement
- Why Google Play Store and it’s data?
- EDA and process involved
- Data Understanding
- Data Cleaning and Manipulations
- Correlation Heatmaps
- Exploratory Data Analysis(EDA):
  - Bivariate Analysis
  - App Categories
  - App Price
  - App Ratings
  - App Installs
  - App Revenue
  - Sentiments
- Challenges faced during data exploration
- Conclusions of our analysis
- Predictions



Problem Statement:

- The Play Store apps data has enormous potential to drive app-making businesses to success. Actionable insights can be drawn for developers to work on and capture the Android market.
- Each app (row) has values for category, rating, size, and more. Another dataset contains customer reviews of the android apps.
- Our main objective is to perform EDA on the given dataset to discover key factors responsible for app engagement and success.
- We need to analyze the data and come up with meaningful insights that would actually help business to strategize their moves.

Why Google Play Data is important to analyze??

In 2021, the Google Play Store had **70% of worldwide downloads** throughout the whole year.

As a large section of apps are free, people tend to download them from Google Play Store.

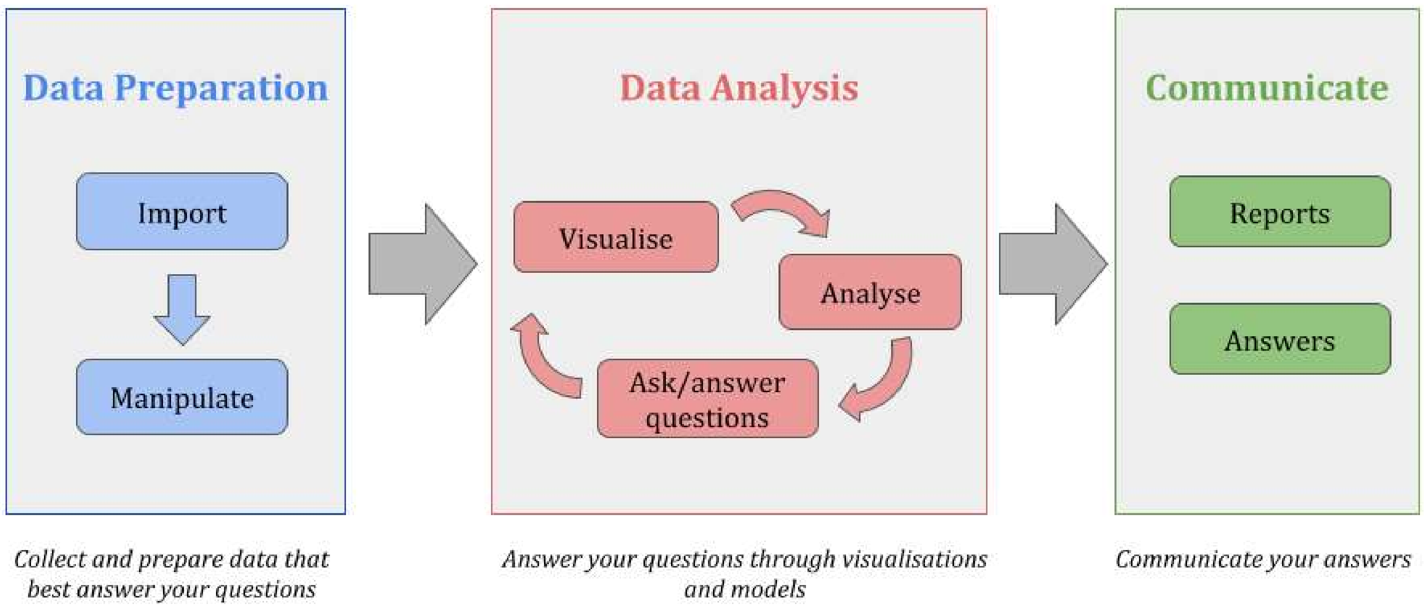
Mobile app market is set to grow **20% by 2023**.

It offers a ready-made market for apps and games. They have a capacity of about a little **over 2 billion users monthly**!!

Knowing the important insights from the play store data can help developers and business managers because they can predict the profit and manage their revenues accordingly.

So, let's get started!!

Flowchart And EDA Process:



Exploring our databases:

We have 2 databases: Play store data set + User reviews data set

Play store data set : It contains basic details of the app like number of users, reviews, ratings, etc. It has **10841 rows and 13 columns**. The features of play store data are:

- App: It contains name of the app with short description.
- Category: This column give the category to which an app belongs. This data set contains 33 categories.
- Rating: The average rating given by the users for the respective app.
- Reviews: The number of users that have dropped a review for this respective app.
- Size: The disk space required to install the respective app.

Exploring our databases continued..

- Installs: The approximate number of times the respective app was installed
- Type: It states whether an app is free to use or paid.
- Price: It gives the price payable to install the app. Price is 0 for free app.
- Content Rating: It states which age group is suitable to consume the content of the respective app.
- Genres: It gives the genres to which respective app belongs.
- Last Updated: It gives the date at which the latest update for the respective app was released.
- Current Ver: It gives the current version of the respective app.
- Android Ver: It gives the android version of the respective app.

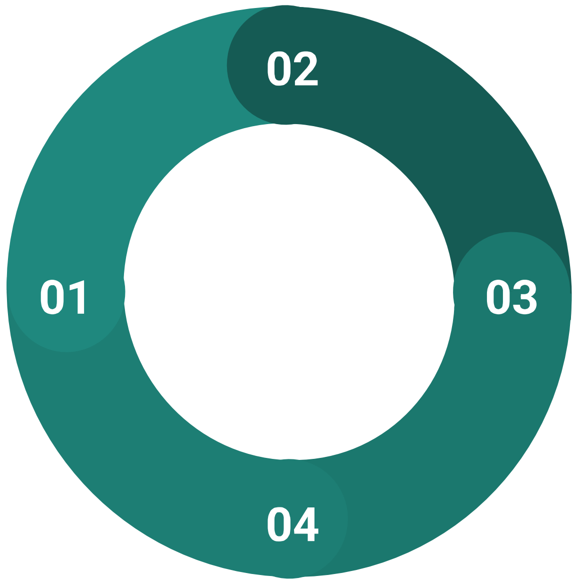
Exploring our databases continued..

User reviews data set : It contains the user reviews and its sentiment score for the respective app.

- It has **64295 rows and 5 columns**.
- App: It contains name of the app with short description.
- Translated Review: It contains the English translation of the review dropped by the user of the app.
- Sentiment: It gives the attitude/emotion of the writer. It can be ‘Positive’ , ‘Negative’ or ‘Neutral’.
- Sentiment Polarity: It gives the polarity of the review. Its range is [-1,1], where 1 means ‘Positive statement’ and -1 means a ‘Negative statement’.
- Sentiment Subjectivity: This value gives how reviewer’ s opinion is to the opinion of the general public. Its range is [0,1].

Data Cleaning and Manipulation:

In the size column, converted KBs→MBs



Dropped duplicate rows from both the datasets

From the Installs and Price columns, changed ‘$’, ‘,’, ‘+’ to an empty string.

In ‘Rating’ column, replaced the null values with the mode.

Dealing with Null Values of Play Store Data

There were 5 columns Rating, Current Ver, Android Ver, Type and Content Rating with missing values.

```
#Checking for null values
df_data.isna().sum().sort_values(ascending=False)
```

```
Rating            1474
Current Ver          8
Android Ver          3
Type                 1
Content Rating       1
App                  0
Category             0
Reviews              0
Size                 0
Installs             0
Price                0
Genres               0
Last Updated         0
dtype: int64
```

```
#filling null values from rating column with mode.
df_data= df_data.fillna(df_data["Rating"].mode()[0])
```

```
#Dropping null values from Type,Content Ratings,Current ver and android ver columns.
df_data.dropna(subset=["Type","Content Rating", "Current Ver", "Android Ver"], inplace= True)
```

```
#Checking for null values
df_data.isna().sum().sort_values(ascending=False)
```

```
App               0
Category          0
Rating            0
Reviews           0
Size              0
Installs          0
Type              0
Price             0
Content Rating    0
Genres            0
Last Updated      0
Current Ver       0
Android Ver       0
dtype: int64
```

Dealing with Duplicate Values in Play Store Data

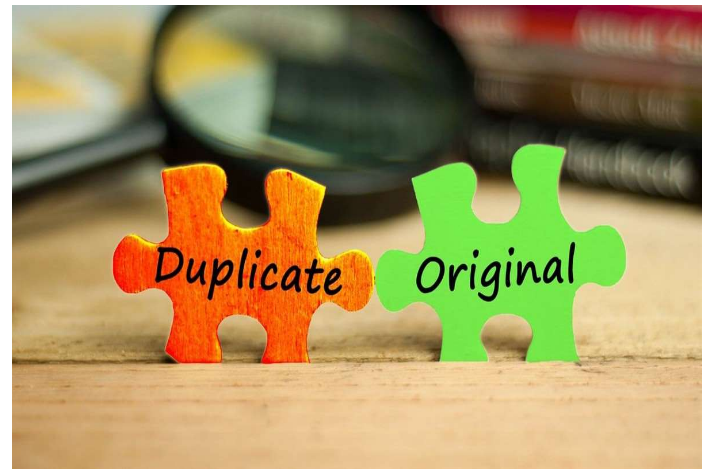
```
# Determining duplicate values in our play store dataset.
df_data.duplicated().sum()
```

```
483
```

```
# Dropping the duplicate values from Play store dataset.
df_data= df_data.drop_duplicates()
```

```
#Rechecking our play store dataset wheather they have any more duplicate values.
df_data.duplicated().sum()
```

```
0
```




```
# convert free values in Type column to 0
df_data[df_data['Type']!='Free'][df_data[df_data['Type']!='Free']['Price']=='0']
#Changing the 'Reviews' column values into valid numeric values
df_data['Reviews'] = pd.to_numeric(df_data['Reviews'])
```

Dealing with Symbolic Values from Price and Installs column

```python
# List of character needs to be remove
list_of_chars = ['+', ',' , '$' ]
# List of column names to clean
list_of_columns = ['Installs', 'Price']

# Loop for each column
for col in list_of_columns:
    # Replace each character with an empty string
    for char in list_of_chars:
        df_data[col] = df_data[col].astype(str).str.replace(char, '')
    # Convert col to numeric
    df_data[col] = pd.to_numeric(df_data[col])

# Typecasting the str type to timestamp in "Latest Updated" column.
df_data["Last Updated"]= pd.to_datetime(df_data["Last Updated"])
```

```python
#defining function to convert all unti in MB and removing unit symbol
def convert(i):
  if 'k' in i:
    return float(i[:-1])/1024
  elif 'M' in i:
    return float(i[:-1])
  else:
    return

df_data['Size']=df_data['Size'].apply(convert)
```

Dealing with Duplicate Values in User Review Data

AI

```
[15] #finding duplicate values
     df_reviews.duplicated().sum()

33616
```

```
[17] #Dropping duplicate values
     df_reviews= df_reviews.drop_duplicates()
```

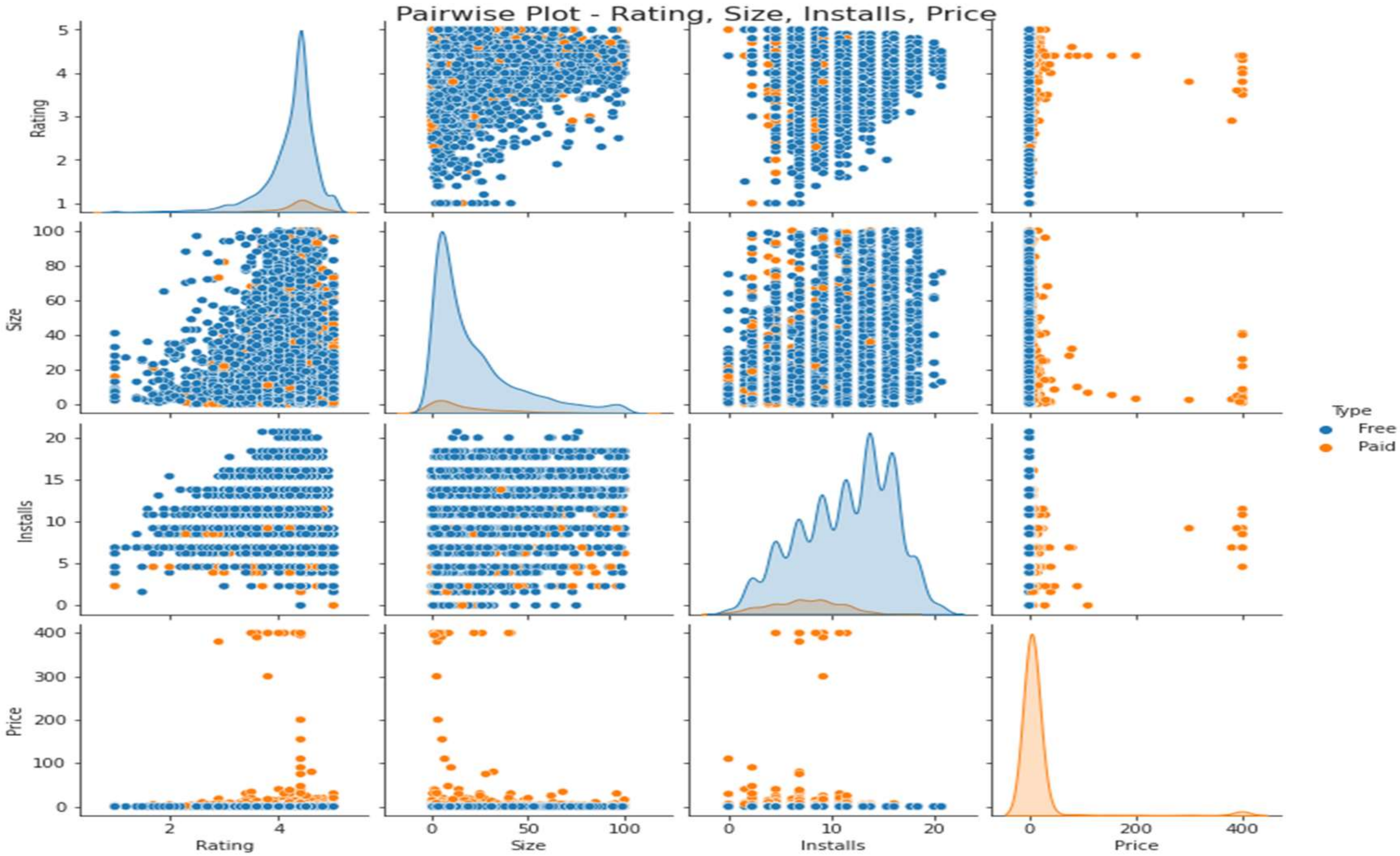
```
[18] #Rechecking to verify if duplicate values are removed
     df_reviews.duplicated().sum()

0
```

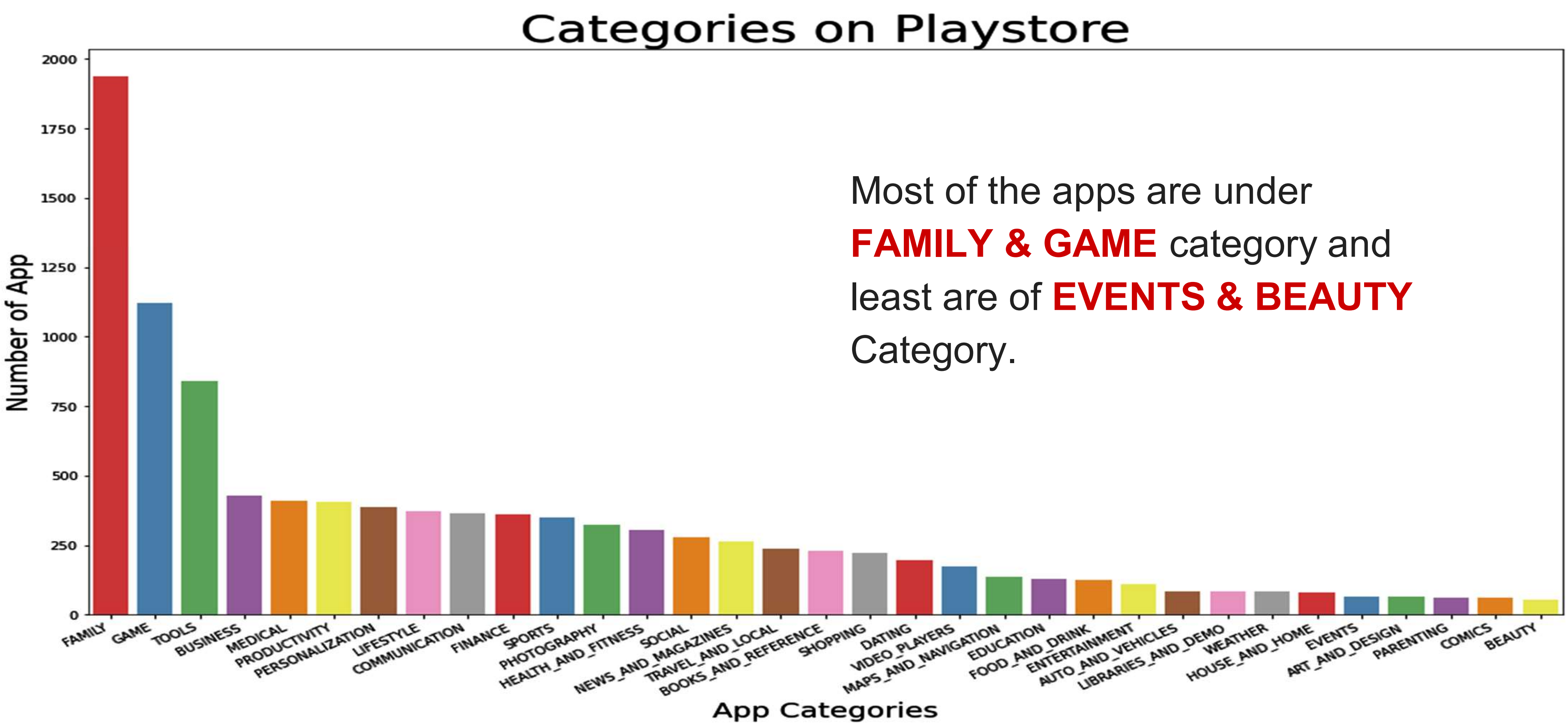
Now, our data is ready for the analysis!

Bivariate Analysis

Here, we are exploring 2 columns at the same time, for the purpose of determining the empirical relationship between them.

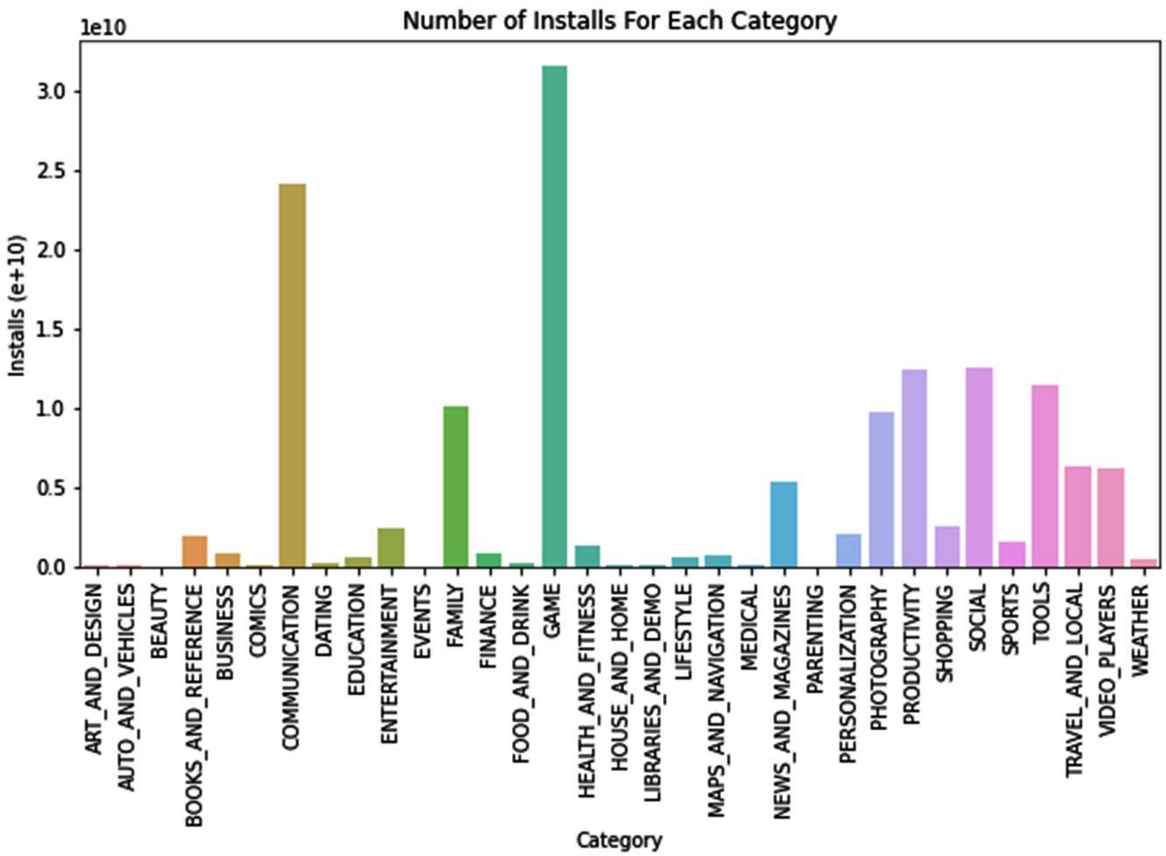


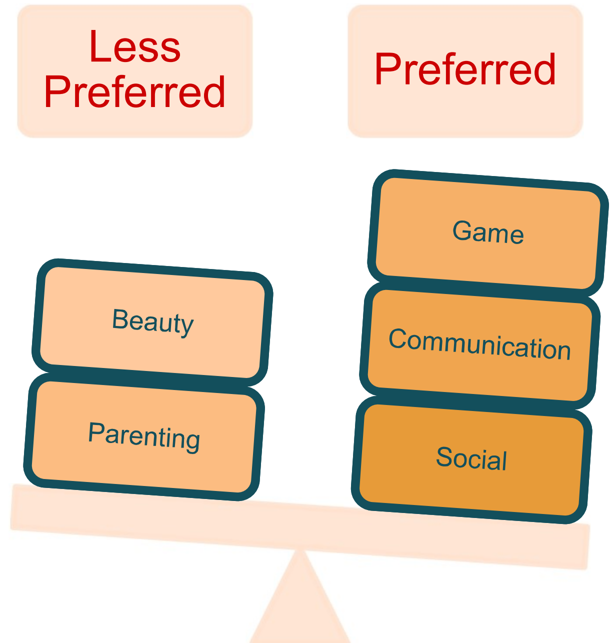
Let’s explore app categories!



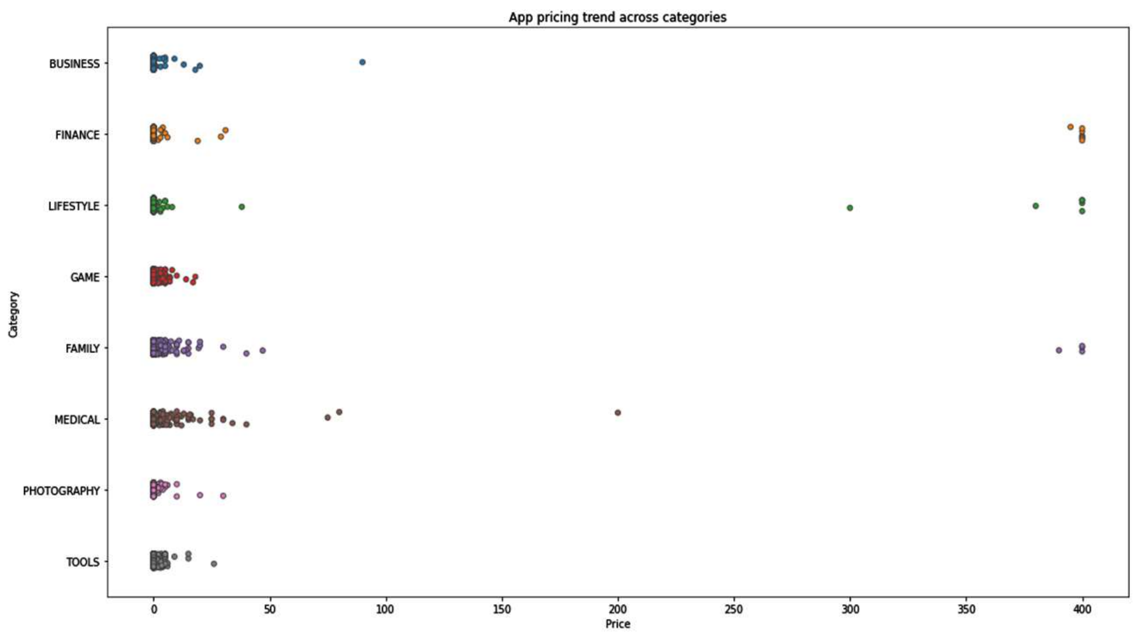
Most of the apps are under **FAMILY & GAME** category and least are of **EVENTS & BEAUTY** Category.

What are the installs across various categories?





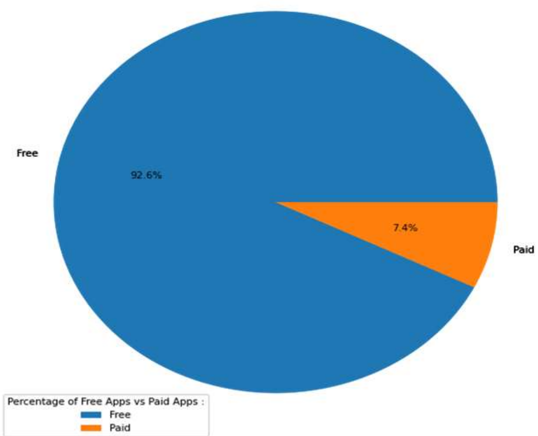
How does the price vary across major categories?

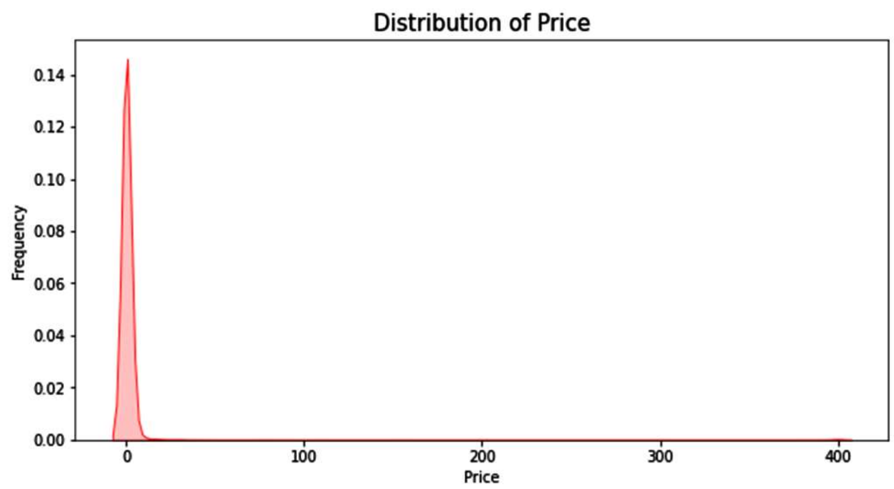


- Many factors should be considered while selecting the right pricing for your mobile app. It is important to re-evaluate the app price before entering the market.
- Here we can see that different apps categories demand different price ranges. Some apps that are simple and easy are free like tools and games.

All Game apps are comparatively low in price, may be that's the reason game apps have more number of downloads, as we have seen earlier. Hence, **Lower the app price → More are the installs!!!**

How many apps are paid?

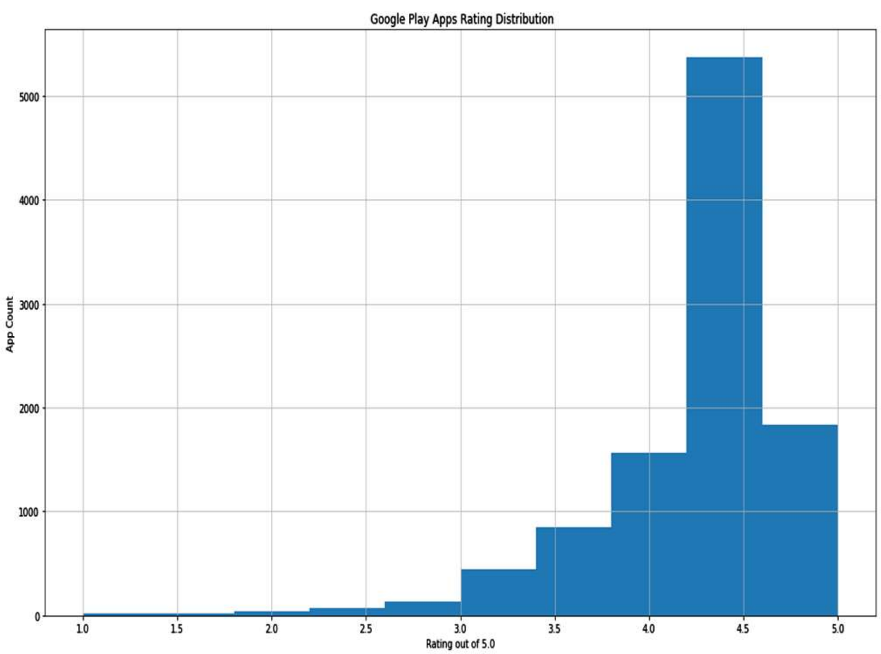


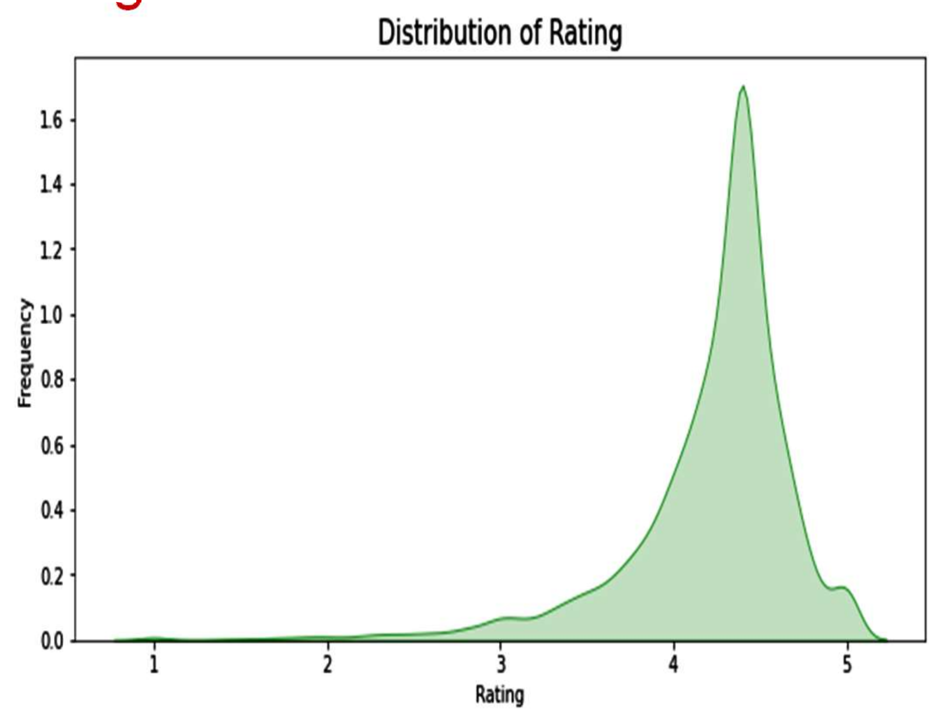


- From the above pie chart, we can conclude **92.6% apps** on google play store are free.
- Even from the **7.4%** of paid apps, most of the apps has price range under 100 dollars.

So less the price → More are the installs!!

Now let’s see how ratings are performing!



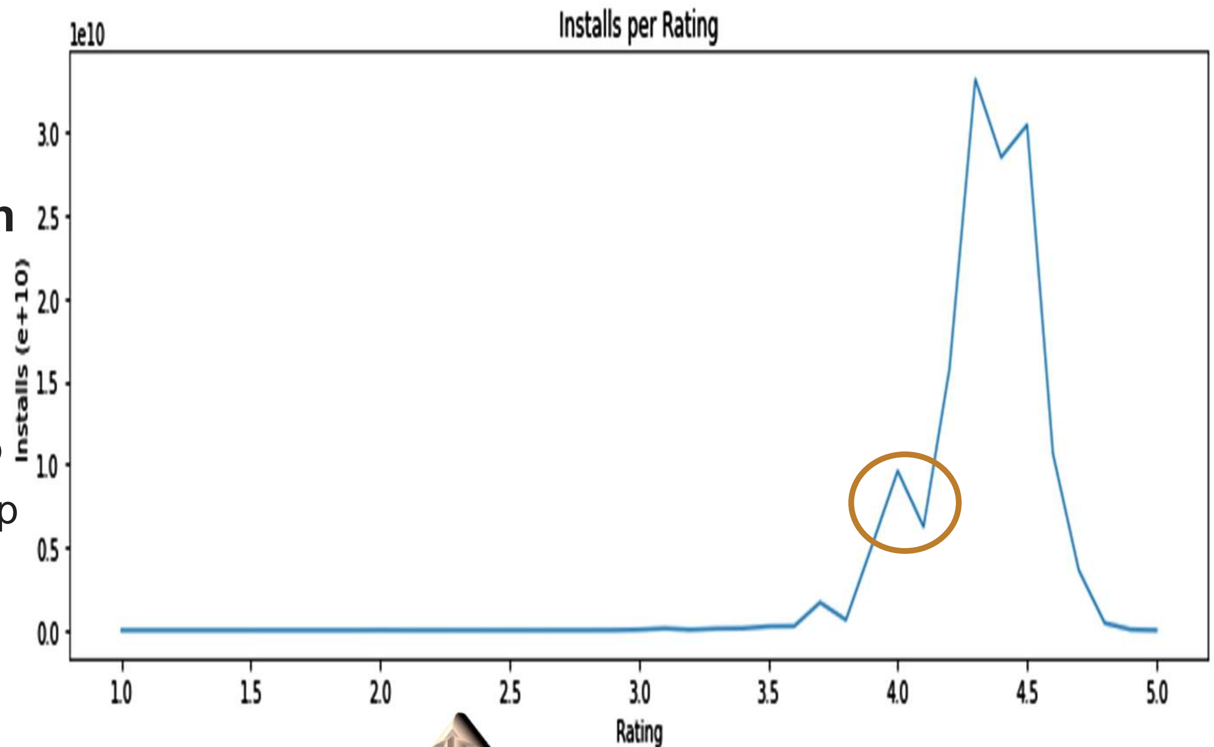


- We can observe that average rating across all app categories is **4.217**.
- A small spike at rating 1 emphasizes that a few apps has been rated poorly and hence the subsequent installs are affected.

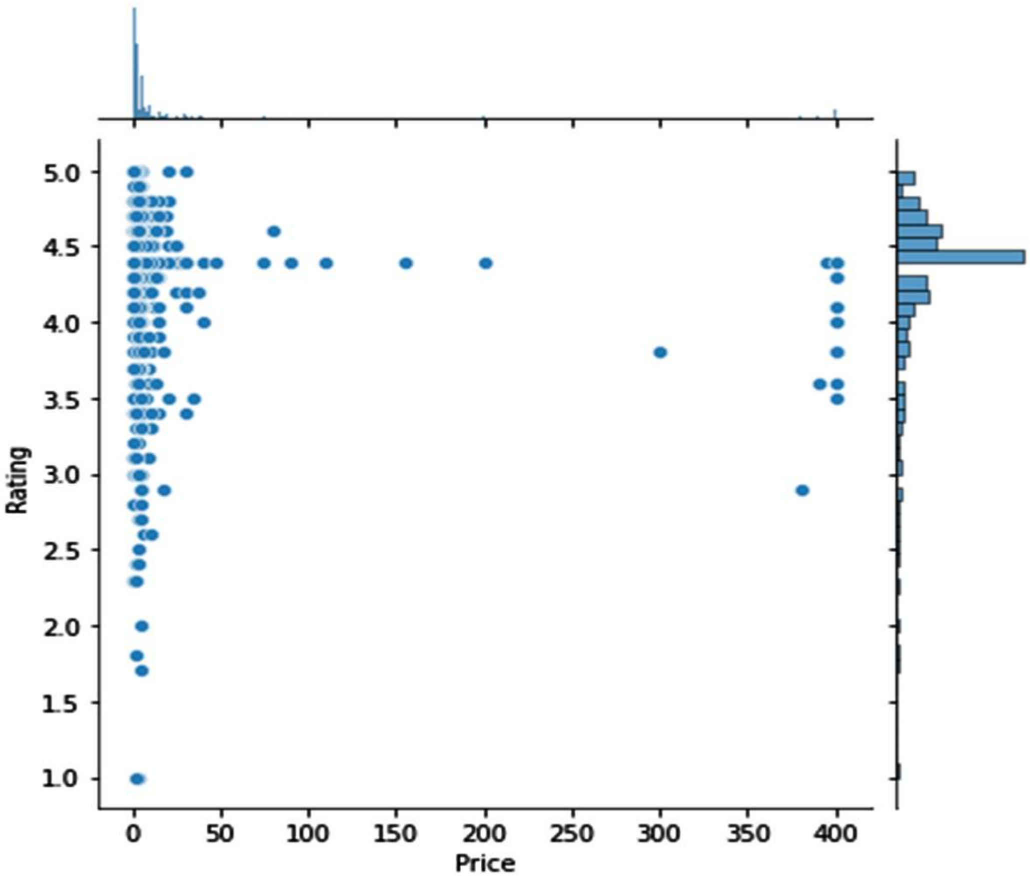
How does installs perform against rating?

- Here, we can observe that higher the rating, more the no of installs. **But this correlation slightly changes after 4.5 ratings**. The probable reason could be
    - Increase in the size of an app
    - Increase in the price of an app

- One interesting insight is after 4.0, the ratings are slightly dipped. The reason is increase in price! But as soon as the prices are decreased, the ratings are again up.



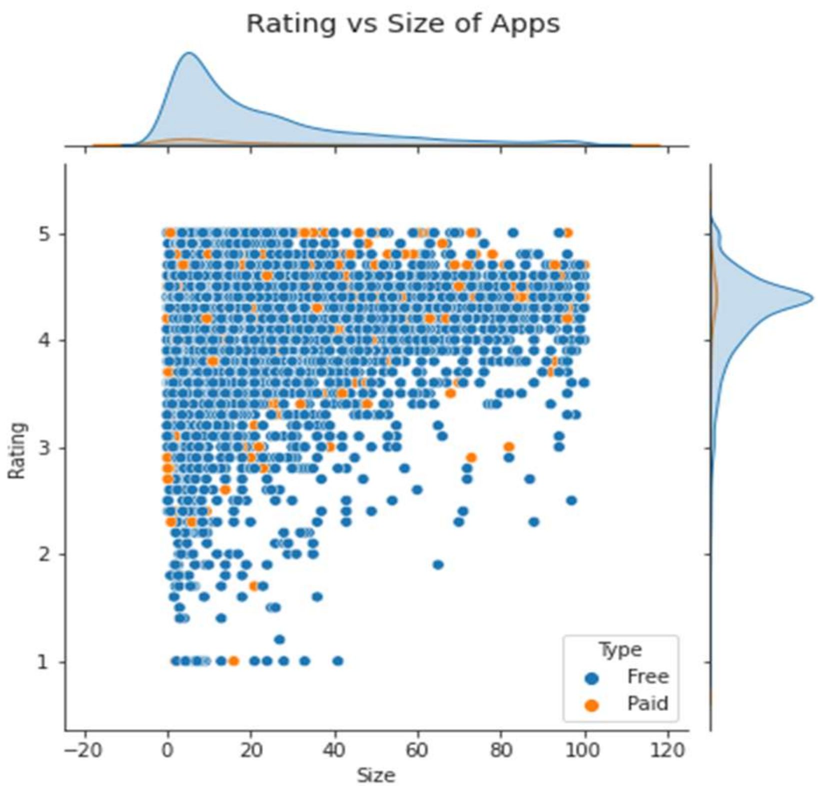
Does an app price affect it's rating?



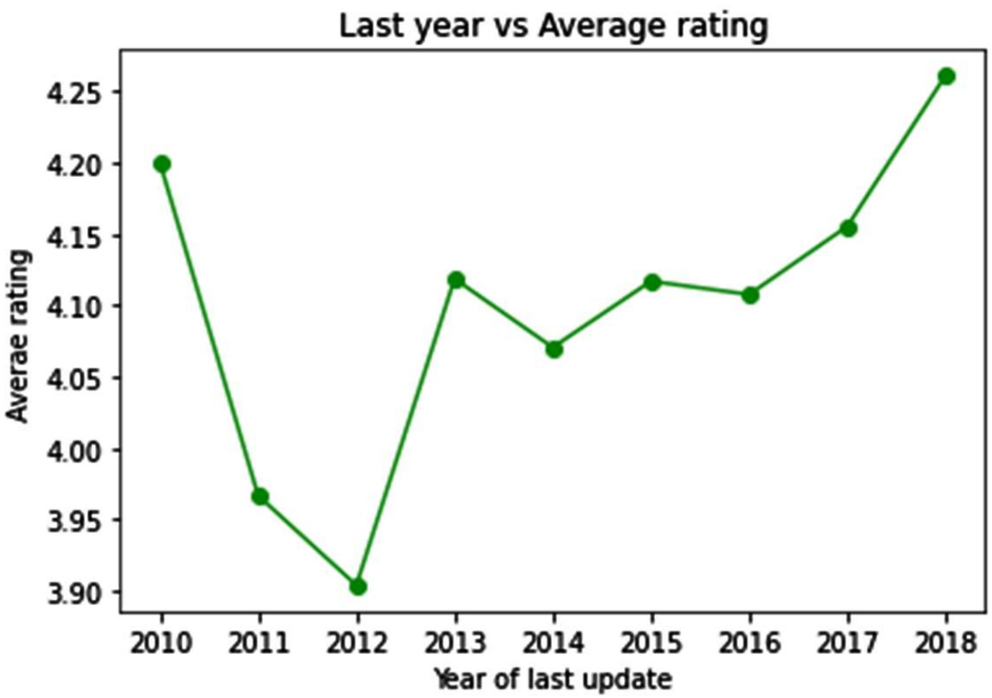
- Here, we can observe that for high priced apps, the ratings are mostly more than 3.0.
- But for free or apps with price less than 50 dollars, ratings do not usually follow any specific pattern.
- They are scattered from 1 to 5. The reasons for such diverse rating pattern could be user experience or size.

Does size of an app affect it’s rating?

- People generally prefer apps with less size due to data and/size constraints. As the size of an app increases, it's rating decreases.
- Surprisingly, there are few apps whose size is close to 100 MBs but still has 4/5 stars as a review. Though this number is less, we can't ignore a fact that the **rating also depends upon the content that the app** is serving.
- Also, some of the paid apps has less ratings. So price and ratings are poorly correlated.



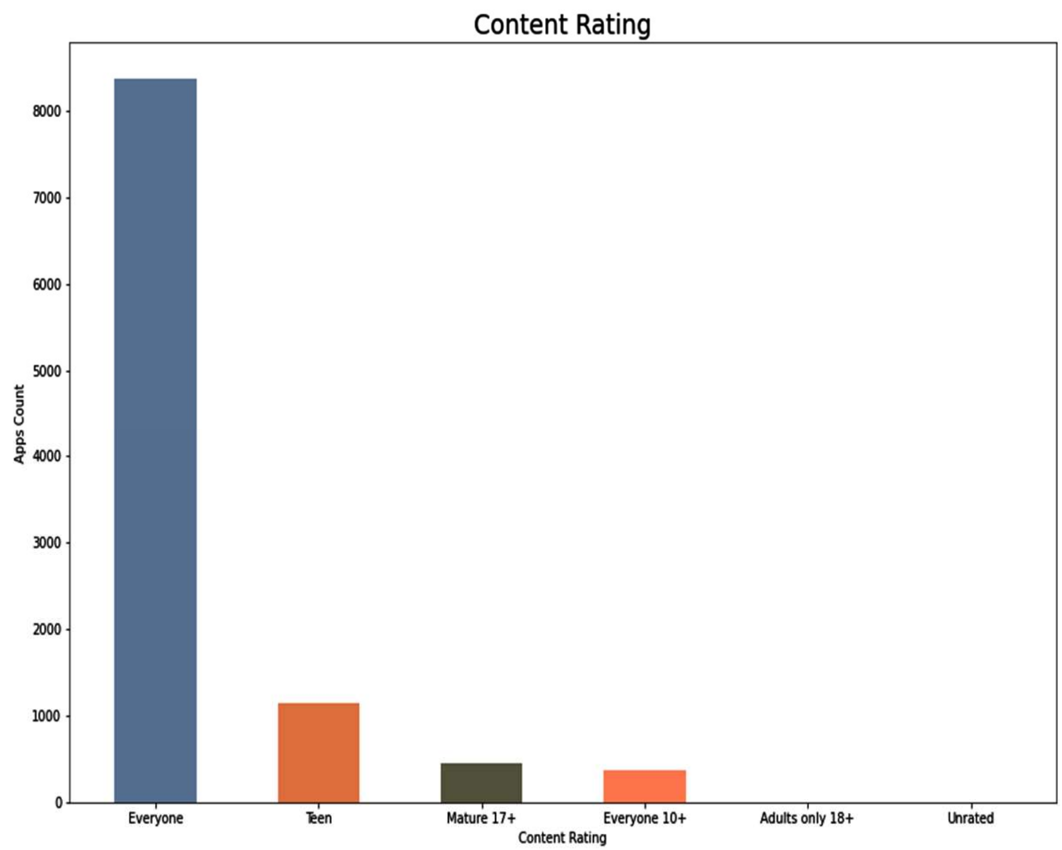
Does update of an app affect it’s rating?



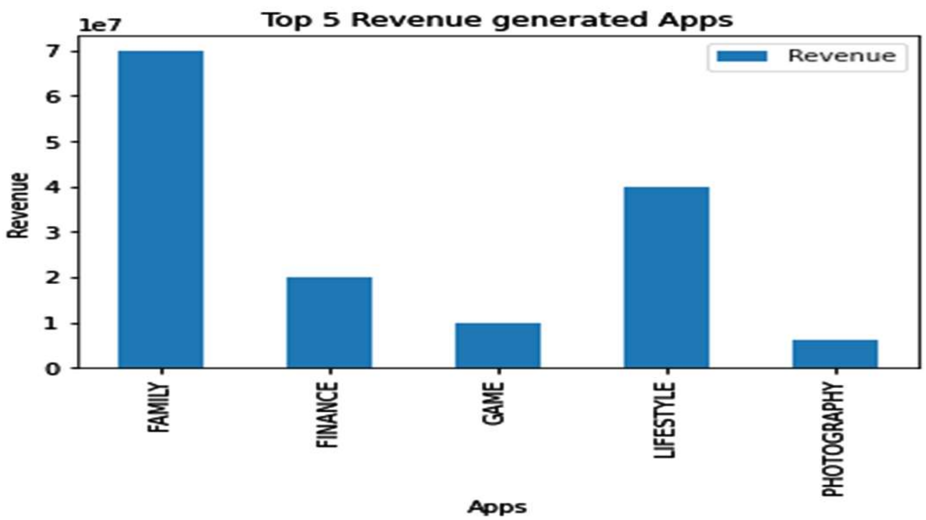
- We can see that the latest the update year, more are the ratings.
- It is observed that from 2010 to 2011, the user experience of the apps was not good and it resulted into drastic dip in the ratings.
- But from 2014, the trend got reversed. The ratings have gone up year-on-year highlighting the great user experience.

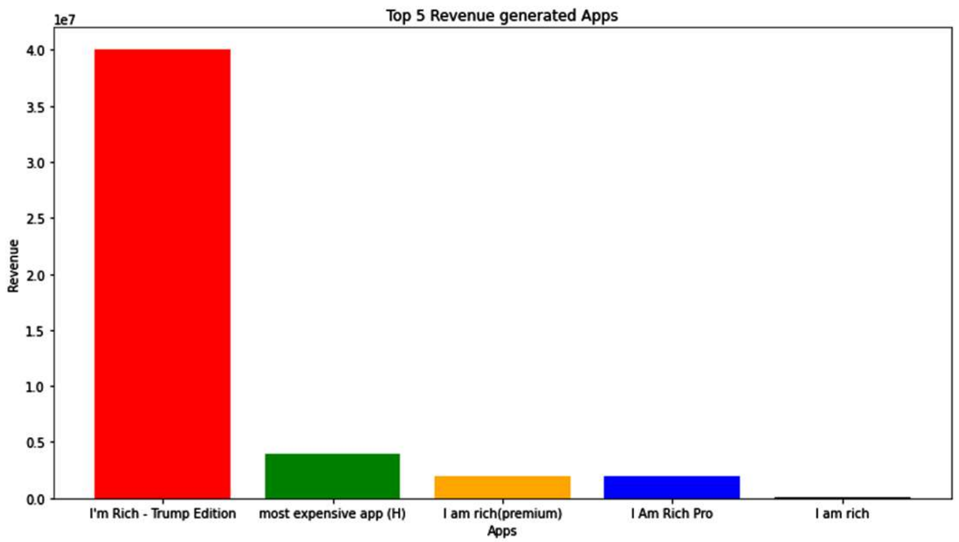
Let’s explore content rating!

- Mostly 90% of total apps are targeting audience in **every age group** and hence they are available for everyone.
- Very few(less then 500 apps) are catering to only adult population i.e. Mature 17+



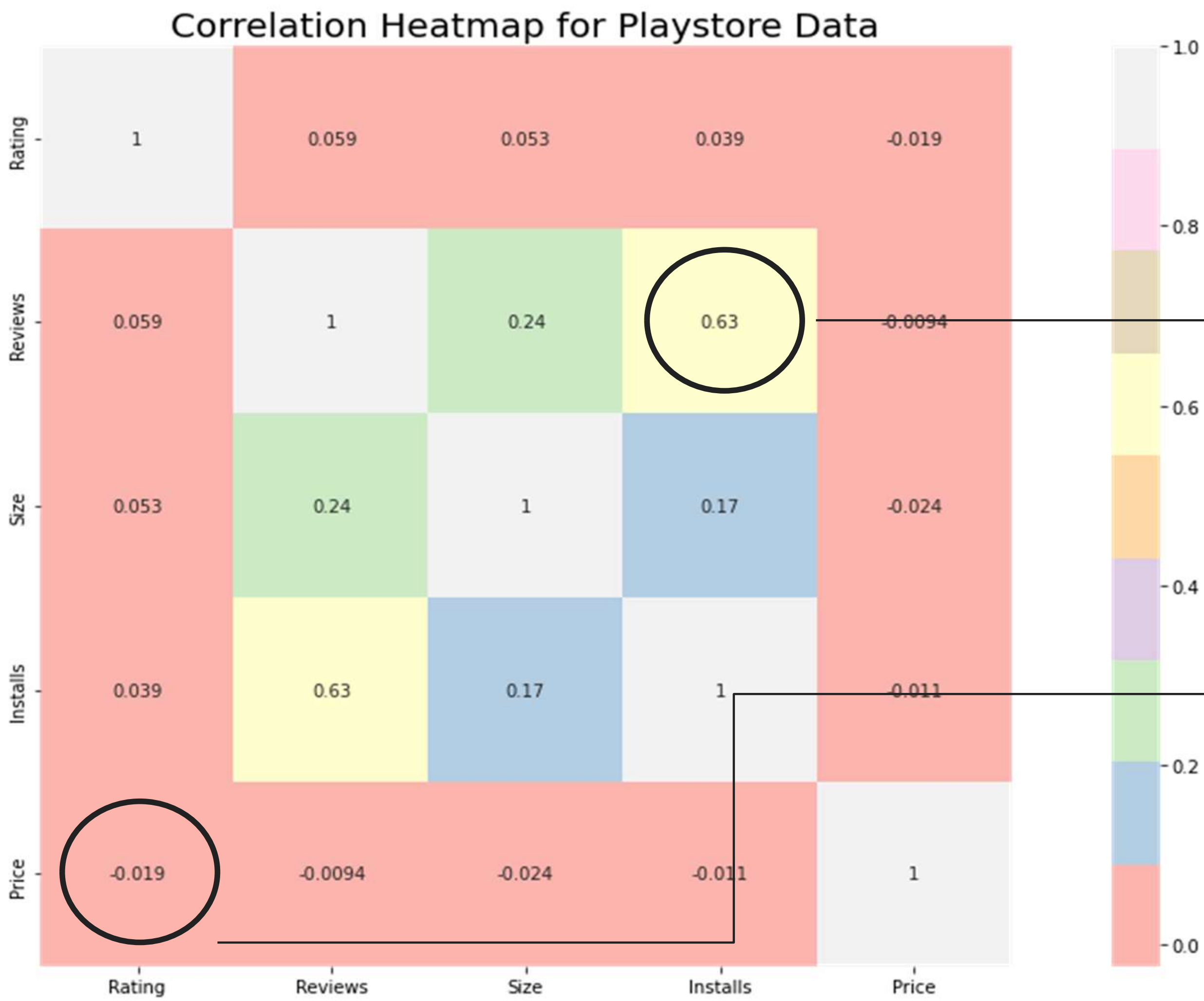
How much money the apps have generated?





- We tried to utilize the information given in ‘installs’ and ‘price’, to calculate the revenue generated by apps.
- The category which generated the highest revenue is **FAMILY.**
- The app named **I’m Rich-Trump Edition** collected the highest money through installs.

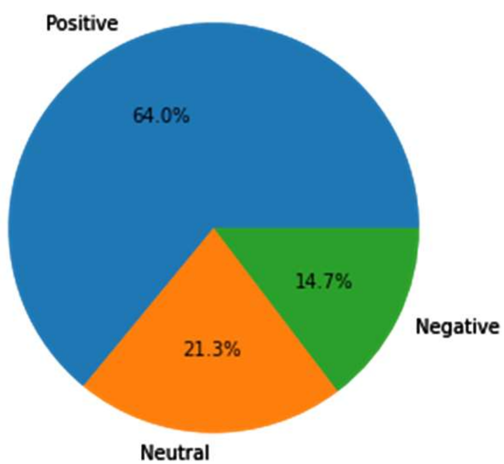
How Size, Reviews, Installs and Price of apps are correlated?

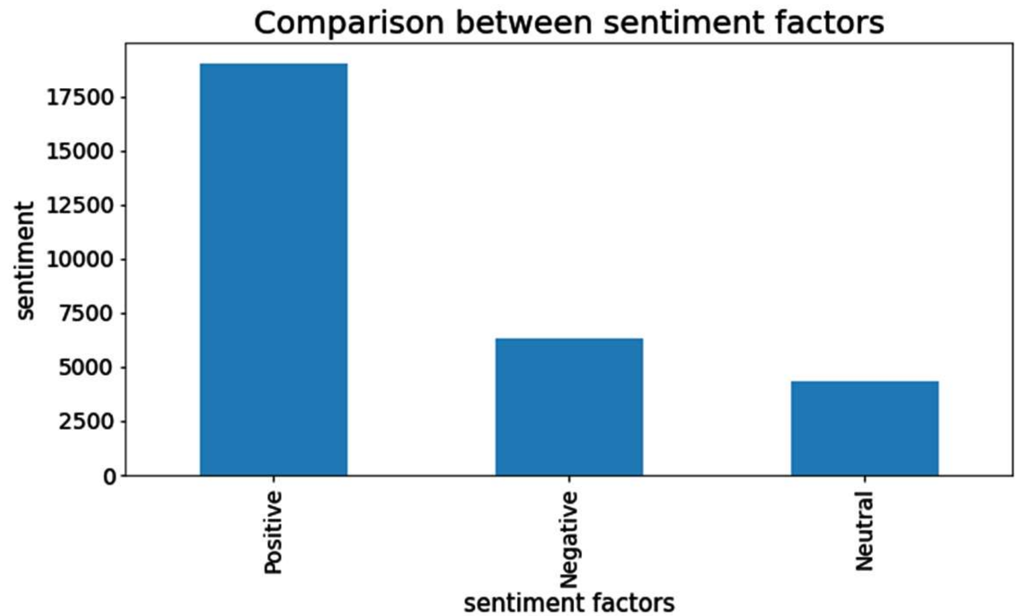


Positive correlation between the Reviews and Installs column i.e. (0.63).

While Price and Ratings share negative correlation i.e. -0.019.

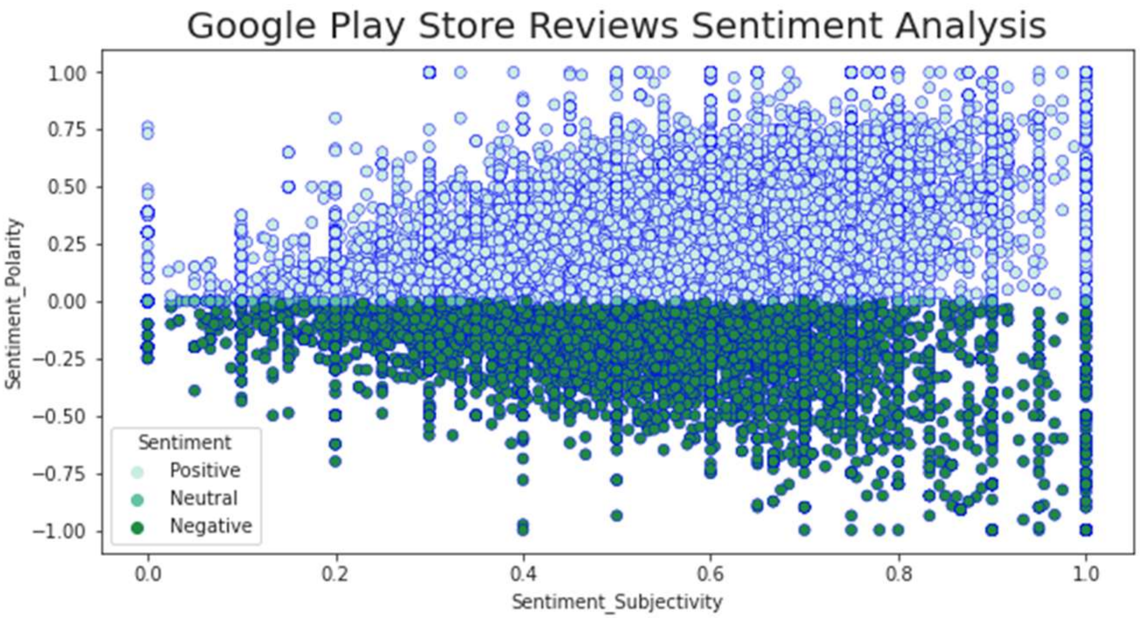
Sentiments

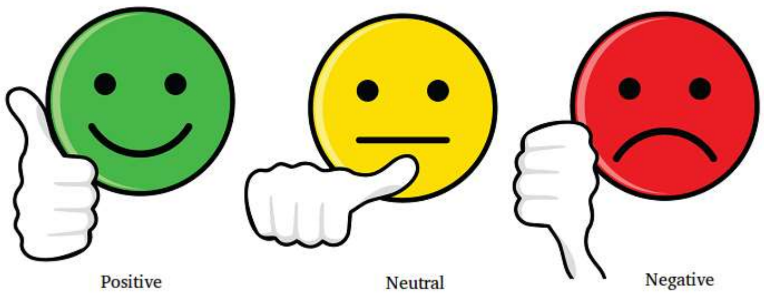




- The sentiment plot shows the results for positive reviews as high.
- We can say that around **64%** of the apps on google play store have received positive sentiments by the users.
- A small chunk of apps-around **14.7%**-have received negative sentiments as well.

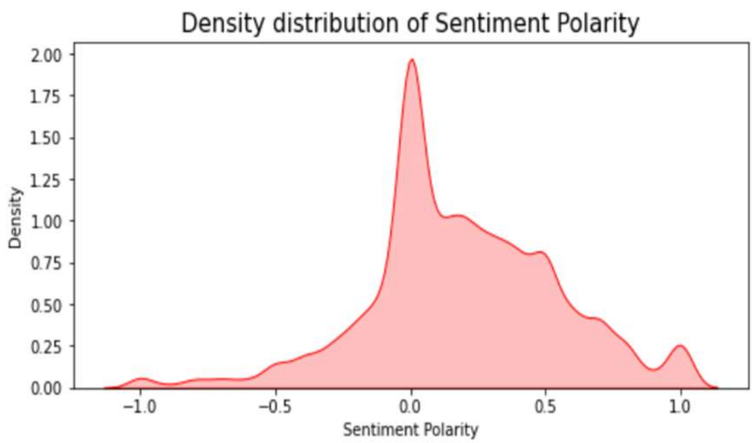
How does sentiment polarity and sentiment subjectivity are linked?

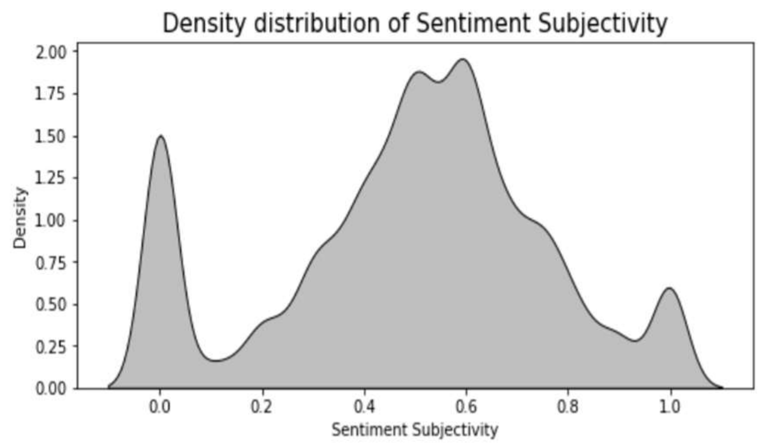




- As we can observe from the above graph that sentiment subjectivity mostly lies in the range of **0.5 to 0.8**. It means, people are giving reviews more **opinion and experience based** rather than **facts.**
- Sentiment subjectivity is mostly scattered around - **0.5 to 0.75** this shows that polarity is not always proportional to sentiment subjectivity but in maximum number of cases **it shows a proportional behavior**.

Sentiments





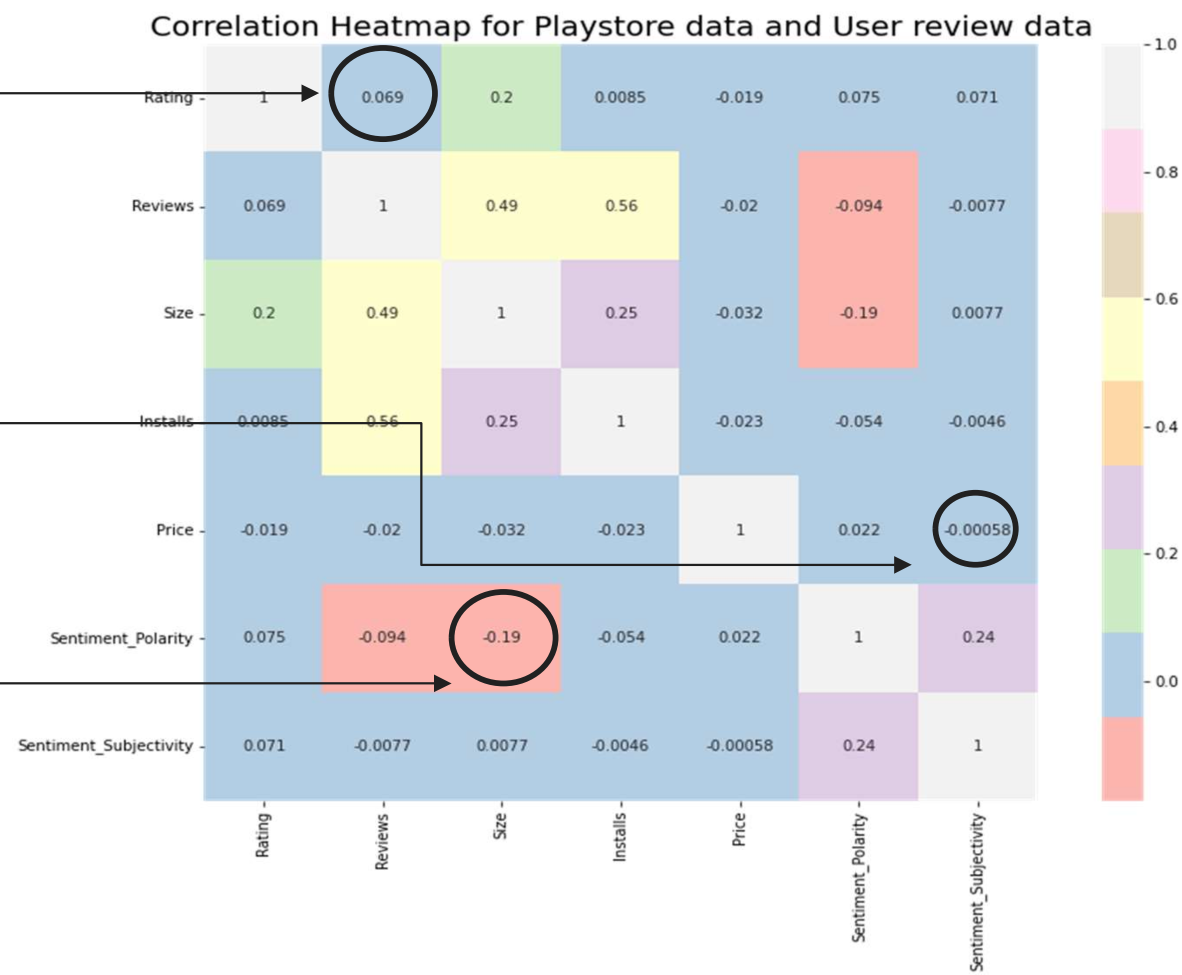
- The polarity score lies in the range of **[-1,1].** Here, we can see from our calculations and the Sentiment Polarity Density distribution graph that the Mean Sentiment Polarity Score is **0.18** which resembles a good average sentiment score(Majority of the users liking the apps).
- The subjectivity is a float within the range **[0.0, 1.0]** where 0.0 is very objective and 1.0 is very subjective. As per our analysis and plotted graph, the Mean Sentiment Subjectivity Score is **0.49**. That means around 50% users are sharing personal opinion while others 50% are just sharing the factual information in reviews.

Let’s see how both the data sets are related!

Positive correlation between the Ratings and Reviews= 0.069.

Negative correlation between the Price and Sentiment subjectivity = -0.00058

Negative correlation between the Size and Sentiment Polarity = -0.19



So let's talk about the challenges that we faced:

- One of the major challenge was to clean the datasets as a lot of scattered information was present especially in the 1st data set.
- Handling the error, duplicate and NaN values in the dataset.
- To draw meaningful insights, we had to design multiple visualizations like scatterplot, jointplot etc. without compromising the results and trends.

To overcome above challenges, we followed:

- AlmaBetter Class material
- Pandas and Numpy libraries
- Stack overflow
- YouTube
- GeeksforGeeks
- Analytics Vidhya

What did we discover?

- Key factors responsible for app engagement and success are :

Price	Sentiments	Ratings
Lower the price, More the installs!	Positive reviews, More the installs!	More the ratings, More the installs!

Conclusions:

- **Reviews** and **installs** share positive correlation while **price** and **rating** share negative correlation.
- **Games** has the most number of installs and hence is a potential unsaturated space for all developers, as it has a maximum number of installs.
- Developing apps within **Family** and **Lifestyle** categories can be aimed for more profit i.e. high revenue .
- **61%** of people have positive sentiments while approx. **15%** reacted negatively which is quite low in comparison(rest are neutral).
- Compared with Free and paid apps, **92.12%** apps are Free and **7.81%** apps are paid.
- As **Everyone** content rating contains all age group people , it has maximum i.e. **81.80%** apps.
- Maximum number of apps belong to the **Family** , **Game** and **Tools** category.
- People love to download apps from **Tools** , **Entertainment** , **Education** , **Business** and **medical** genres.
- Average rating of apps on the play store is **4.17** which is quite good.
- Users prefer to pay for apps that are **light weighted**.
- Paid apps that are **higher** in size may not perform well in the market.
- Users tend to download a given app more if it has been **reviewed** by a large number of people.
- There is a positive correlation between **installs** and **rating**.

Let’s do some predictions!

Now that we have analyzed out datasets and drawn meaningful insights out of them, let’s see if we can ‘predict’ a perfect app!!

- It's good to develop a **free app** and **family and lifestyle** categories can be aimed for more profit.
- People tend to download more if apps have **more reviews**.
- People follow apps with **timely updates** having size as **small** as possible.
- The category **GAME** is a potential unsaturated space for app developers. The segment is expected to grow 1.7 times faster than the rest of the market and surpass **$136 billion** in revenues in 2022.

THANK YOU