Capstone Project

Play Store App Review Analysis

Team Members

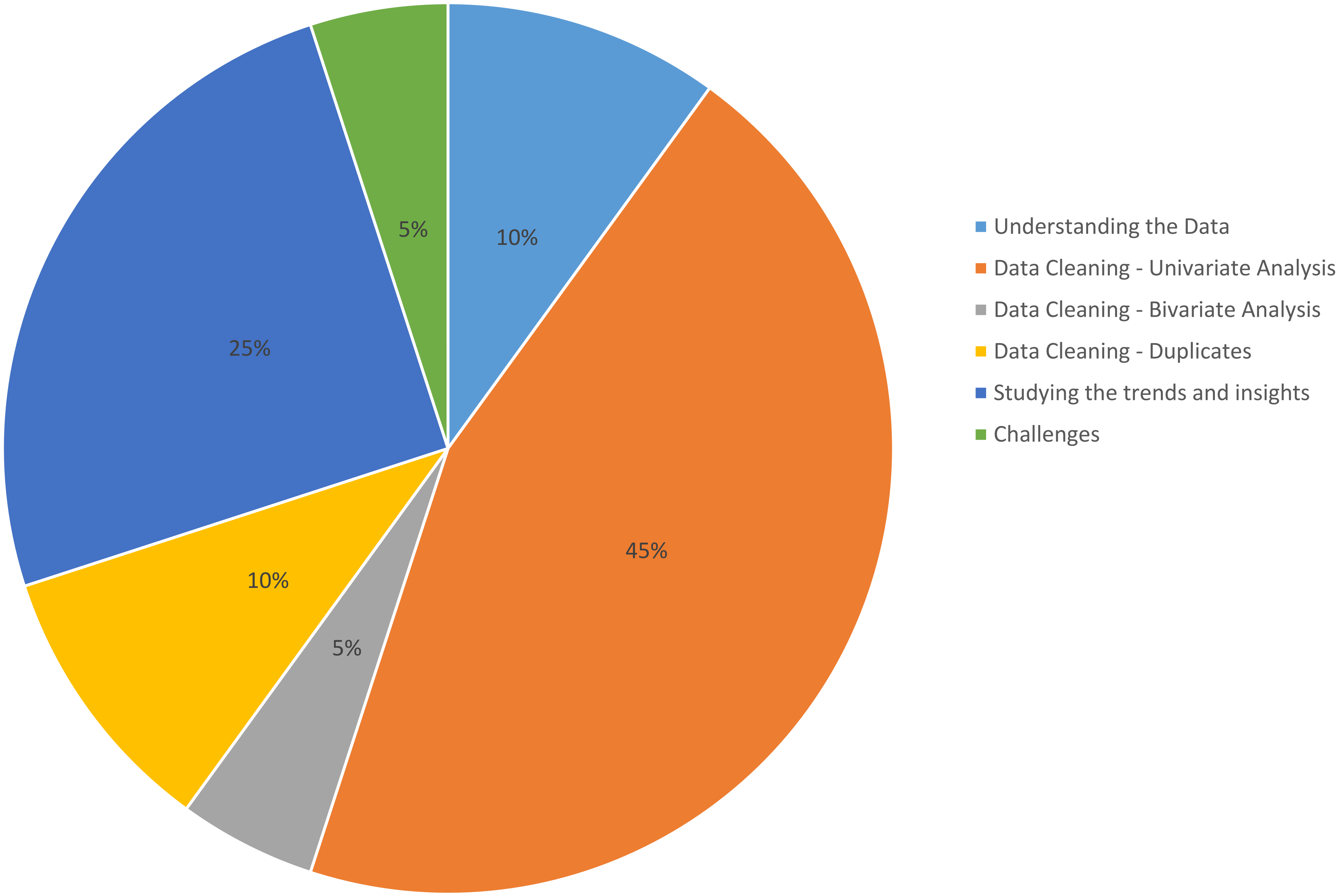
Amrutha B S
Mahima Shree
Purnima Rai
Tanmay Bohra
Vikram Pratap

Objective : To help AlmaBetter with building their technical know how about building a good android application. To study current trends and insights of Play Store.

Problem Statements:

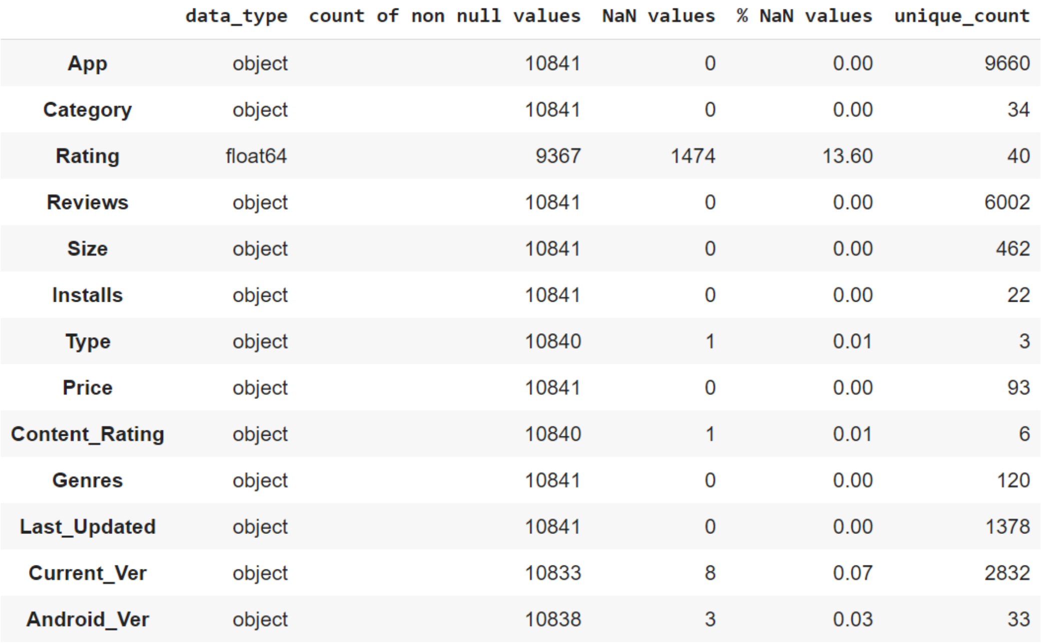
1. What are the top categories on Play Store?
2. Are majority of the apps Paid or Free?
3. How importance is the rating of the application?
4. Which categories from the audience should the app be based on?
5. Let us see how the median ratings vary for categories.
6. How are reviews and ratings co-related?
7. Lets us discuss the sentiment subjectivity.
8. Is subjectivity and polarity proportional to each other?
9. What is the percentage of review sentiments?
10. How is sentiment polarity varying for paid and free apps?

Over-View of the Time Spent

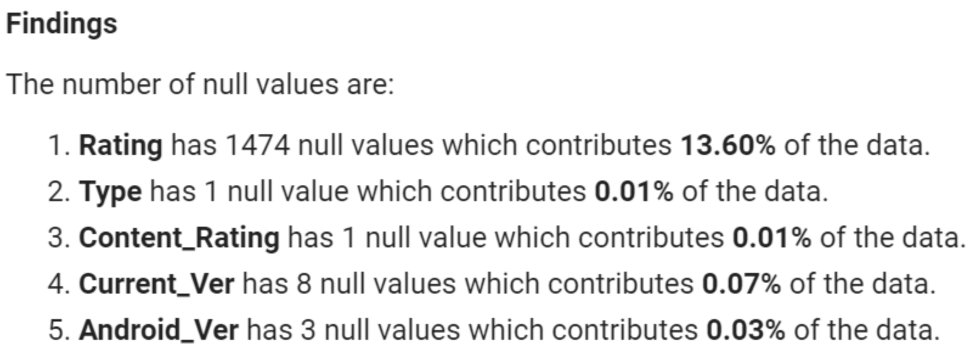


Data in the Data Sets

Play Store Data



	data_type	count of non null values	NaN values	% NaN values	unique_count
App	object	10841	0	0.00	9660
Category	object	10841	0	0.00	34
Rating	float64	9367	1474	13.60	40
Reviews	object	10841	0	0.00	6002
Size	object	10841	0	0.00	462
Installs	object	10841	0	0.00	22
Type	object	10840	1	0.01	3
Price	object	10841	0	0.00	93
Content_Rating	object	10840	1	0.01	6
Genres	object	10841	0	0.00	120
Last_Updated	object	10841	0	0.00	1378
Current_Ver	object	10833	8	0.07	2832
Android_Ver	object	10838	3	0.03	33

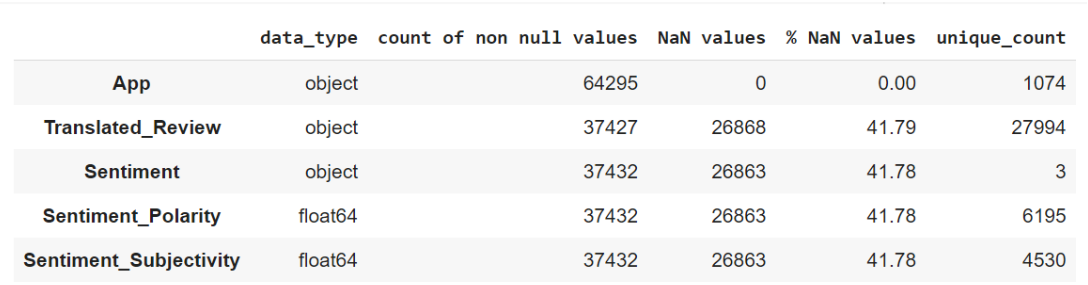


Findings

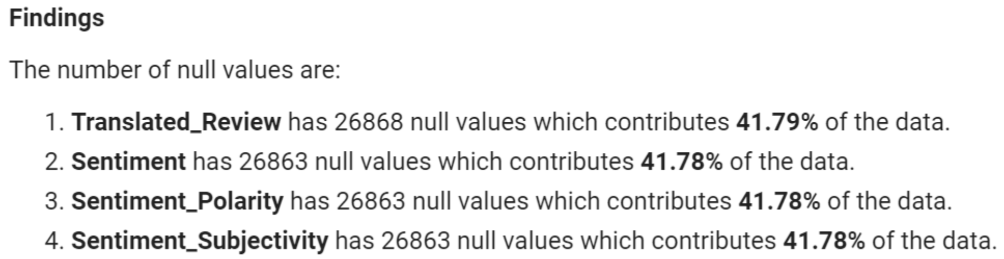
The number of null values are:

1. **Rating** has 1474 null values which contributes **13.60%** of the data.
2. **Type** has 1 null value which contributes **0.01%** of the data.
3. **Content_Rating** has 1 null value which contributes **0.01%** of the data.
4. **Current_Ver** has 8 null values which contributes **0.07%** of the data.
5. **Android_Ver** has 3 null values which contributes **0.03%** of the data.

User Reviews Data



	data_type	count of non null values	NaN values	% NaN values	unique_count
App	object	64295	0	0.00	1074
Translated_Review	object	37427	26868	41.79	27994
Sentiment	object	37432	26863	41.78	3
Sentiment_Polarity	float64	37432	26863	41.78	6195
Sentiment_Subjectivity	float64	37432	26863	41.78	4530



Findings

The number of null values are:

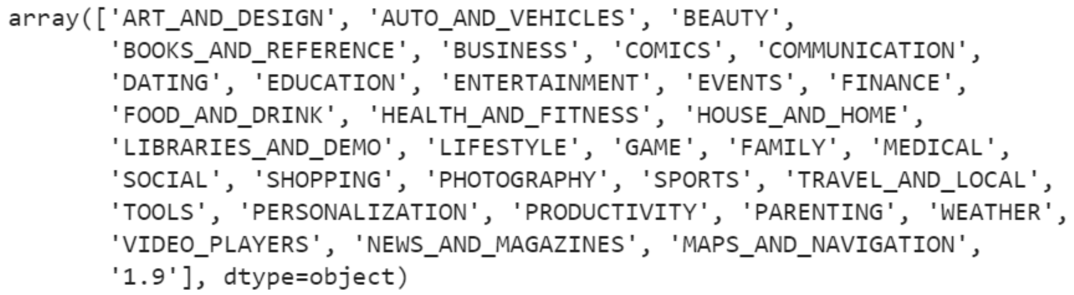
1. **Translated_Review** has 26868 null values which contributes **41.79%** of the data.
2. **Sentiment** has 26863 null values which contributes **41.78%** of the data.
3. **Sentiment_Polarity** has 26863 null values which contributes **41.78%** of the data.
4. **Sentiment_Subjectivity** has 26863 null values which contributes **41.78%** of the data.

Data Cleaning – Univariate Analysis

Important Steps of Data Cleaning

- Identifying the null values.
- Identifying the invalid data.
- Removing Symbols.
- Standardizing the data types.

Category

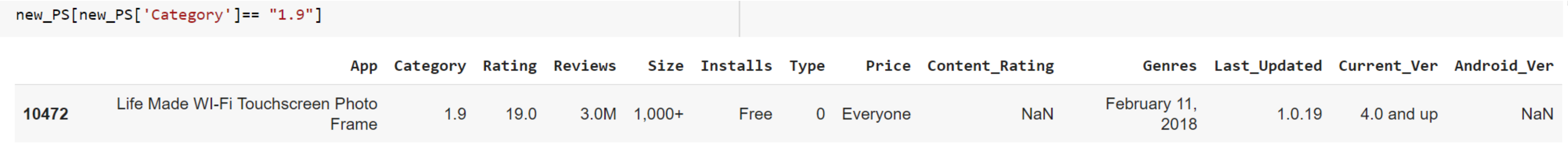


```
array(['ART_AND_DESIGN', 'AUTO_AND_VEHICLES', 'BEAUTY',
       'BOOKS_AND_REFERENCE', 'BUSINESS', 'COMICS', 'COMMUNICATION',
       'DATING', 'EDUCATION', 'ENTERTAINMENT', 'EVENTS', 'FINANCE',
       'FOOD_AND_DRINK', 'HEALTH_AND_FITNESS', 'HOUSE_AND_HOME',
       'LIBRARIES_AND_DEMO', 'LIFESTYLE', 'GAME', 'FAMILY', 'MEDICAL',
       'SOCIAL', 'SHOPPING', 'PHOTOGRAPHY', 'SPORTS', 'TRAVEL_AND_LOCAL',
       'TOOLS', 'PERSONALIZATION', 'PRODUCTIVITY', 'PARENTING', 'WEATHER',
       'VIDEO_PLAYERS', 'NEWS_AND_MAGAZINES', 'MAPS_AND_NAVIGATION',
       '1.9'], dtype=object)
```

Findings

- We began our univariate analysis with the Category Column, and found out that everything was correct, except for the **type 1.9** which doesn't seem to be a category.

- We then explored it and found the below image.



```
new_PS[new_PS['Category']== "1.9"]
```

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content_Rating	Genres	Last_Updated	Current_Ver	Android_Ver
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	NaN

- It is evident from the above image that row no. 10472 has garbage value, as all the columns have mismatch of data.

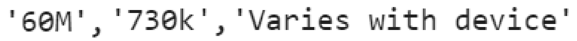
- We thus decided to drop this row from the dataset.

Data Cleaning – Univariate Analysis (Play Store Data Set)

Findings

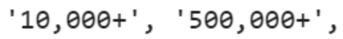
- **Reviews** column was converted to a numeric type.

- **Size** column had ‘**M**’, ‘**k**’ & ‘**Varies with device**’ which were all converted to a numeric variable.



```
'60M','730k','Varies with device'
```

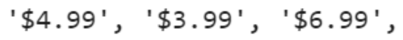
- **Installs** column has ‘**+**’ & ‘**,**’ characters present in it, which were removed, and then the column was converted to a numeric type.



```
'10,000+', '500,000+',
```

- **Type** column has only two strings i.e., Free & Paid which were relevant and retained as it is.

- **Price** column had ‘**$**’ symbol present, which was removed and then the column was converted to a numeric type.

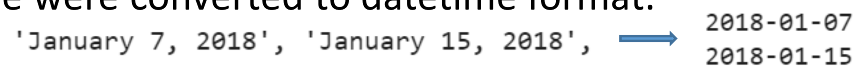


```
'$4.99', '$3.99', '$6.99',
```

- **Content_Rating** column has relevant variables present in it and thus were retained as it is.

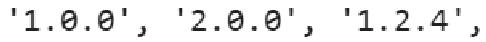
- **Genres** column also had relevant variables present in it and thus were retained as it is.

- **Last_Updated** column has the dates in string format, these were converted to datetime format.



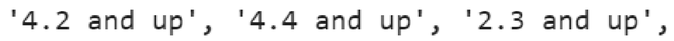
```
'January 7, 2018', 'January 15, 2018',  →  2018-01-07
                                           2018-01-15
```

- **Current_Ver** column refers to the latest version of the app and it had all relevant data in it, so it was retained. 8 null values were present, which were removed.



```
'1.0.0', '2.0.0', '1.2.4',
```

- **Android_Ver** column refers to the version on which the app can run efficiently and it had 3 null values which were removed.



```
'4.2 and up', '4.4 and up', '2.3 and up',
```

Data Cleaning – Univariate Analysis (User Reviews Data Set)

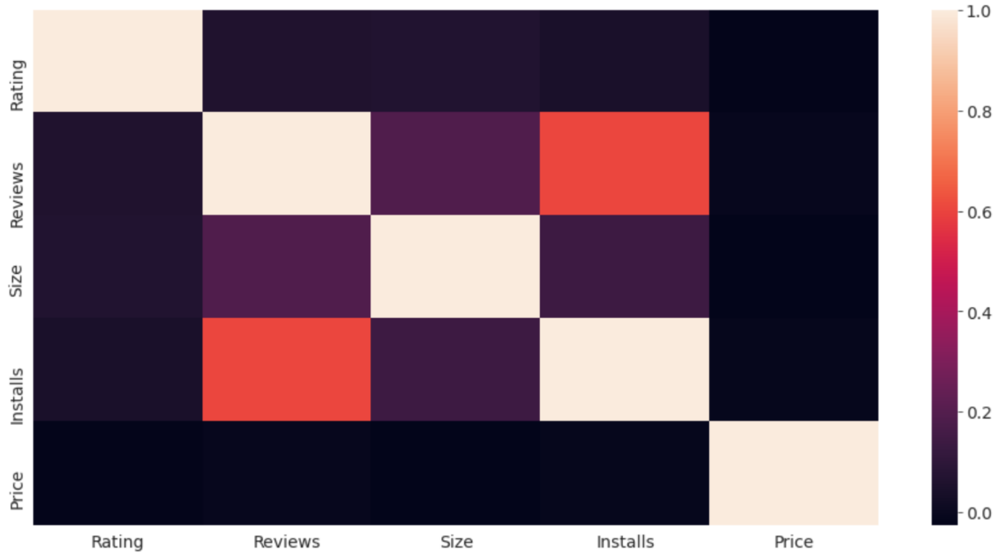
Findings

- **42%** of the Translated_Review, Sentiment, Sentiment_Polarity & Sentiment_Subjectivity has null values and we had dropped all of these values, as they were of no value to the analysis.

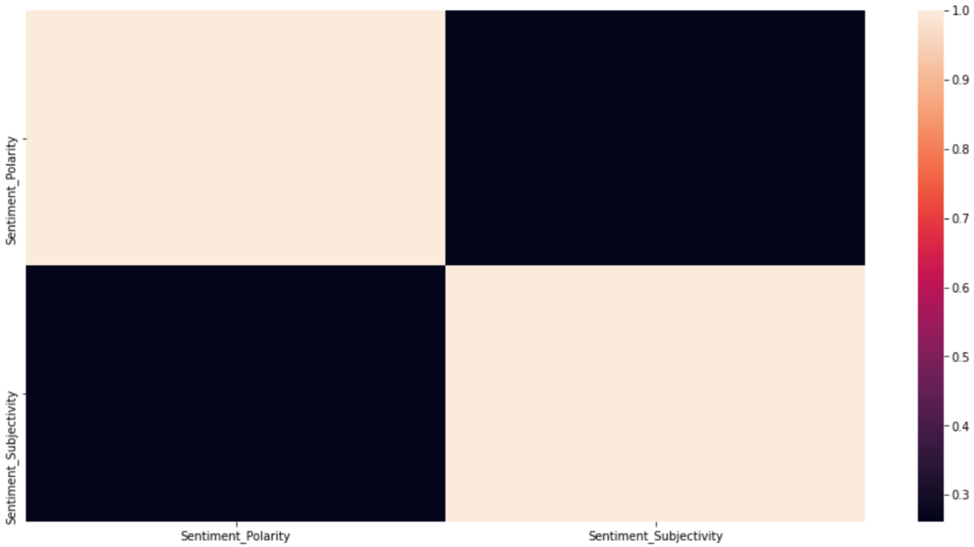
Data Cleaning – Bivariate Analysis

Findings

- We studied the heatmaps of both the data sets to ensure that we are not ignoring any important and relevant data and it's dependency on the each other which would help us in our exploratory analysis.



Play Store Heat Map



User Reviews Heat Map

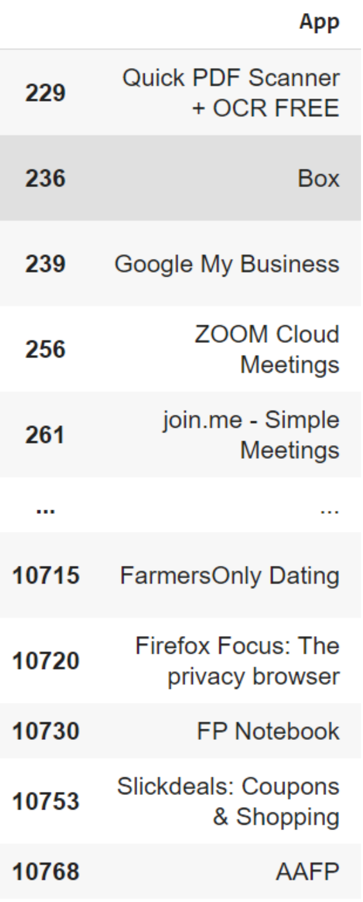
Data Cleaning – Duplicates

Findings

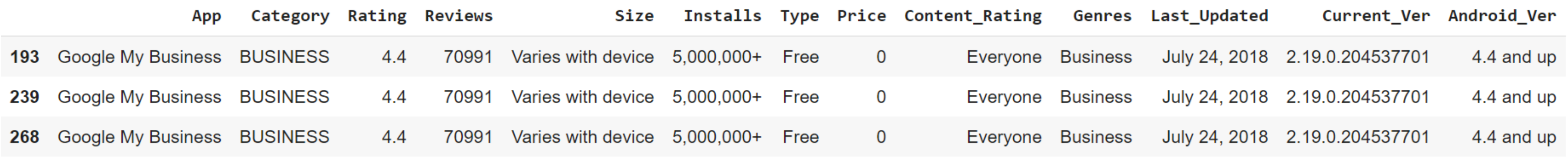
- Now that we have done our important analysis i.e., Univariate and Bivariate analysis. We shall now check for the duplicates present in the given play store data set and remove them.

- A total of **1049** duplicates was found to be present in the play store data set.

- Let us see below and example of the duplicate data.



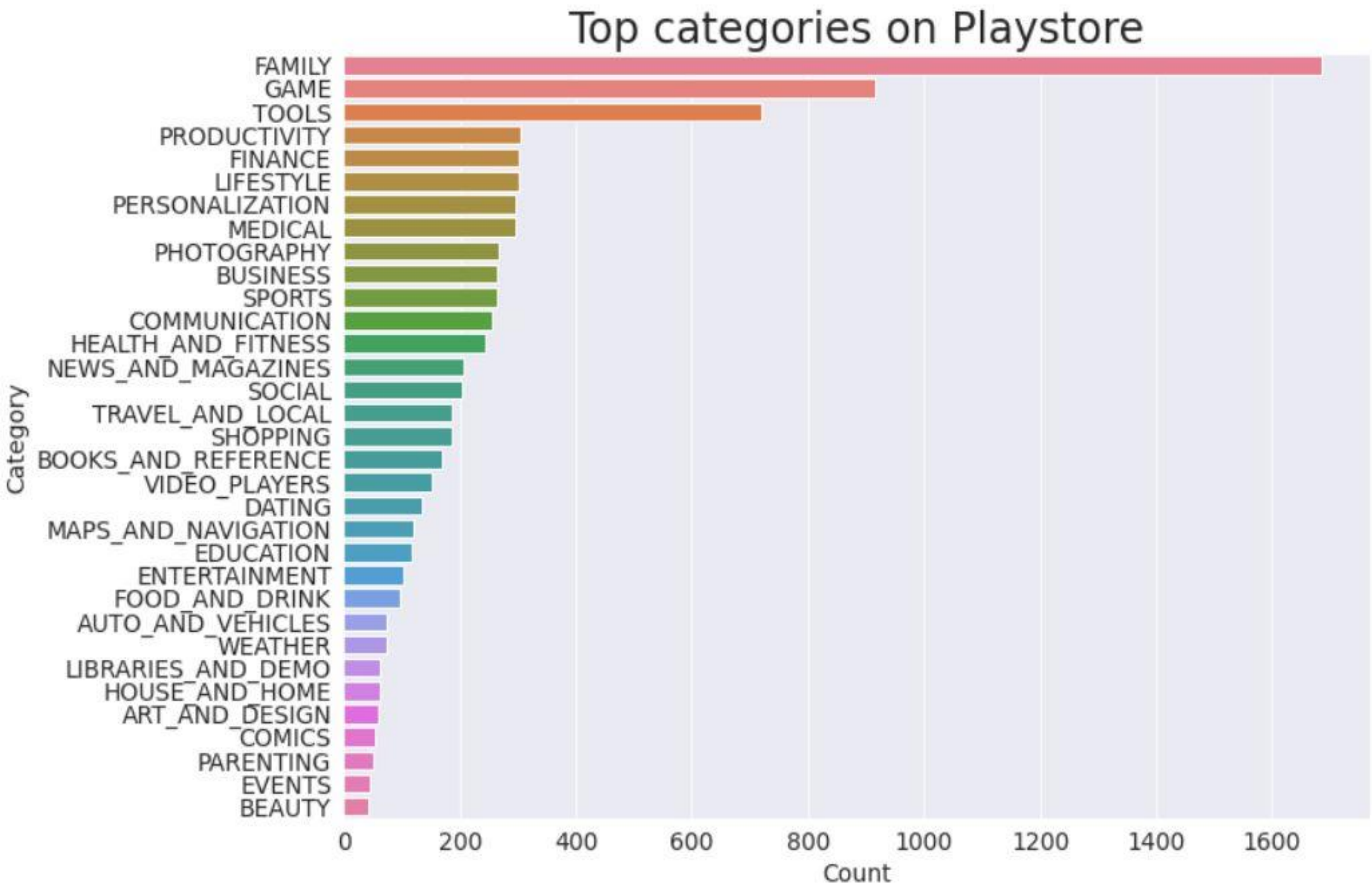
	App
229	Quick PDF Scanner + OCR FREE
236	Box
239	Google My Business
256	ZOOM Cloud Meetings
261	join.me - Simple Meetings
...	...
10715	FarmersOnly Dating
10720	Firefox Focus: The privacy browser
10730	FP Notebook
10753	Slickdeals: Coupons & Shopping
10768	AAFP



	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content_Rating	Genres	Last_Updated	Current_Ver	Android_Ver
193	Google My Business	BUSINESS	4.4	70991	Varies with device	5,000,000+	Free	0	Everyone	Business	July 24, 2018	2.19.0.204537701	4.4 and up
239	Google My Business	BUSINESS	4.4	70991	Varies with device	5,000,000+	Free	0	Everyone	Business	July 24, 2018	2.19.0.204537701	4.4 and up
268	Google My Business	BUSINESS	4.4	70991	Varies with device	5,000,000+	Free	0	Everyone	Business	July 24, 2018	2.19.0.204537701	4.4 and up

- We thus, then decided to drop these, as all the values in each column were repetitive in nature.

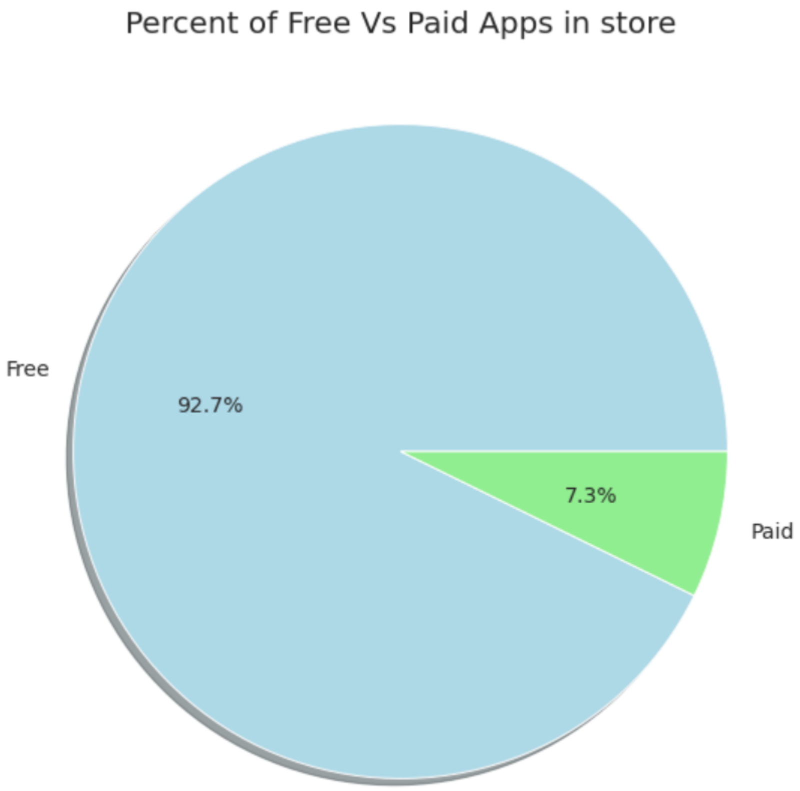
Top categories on Play Store?



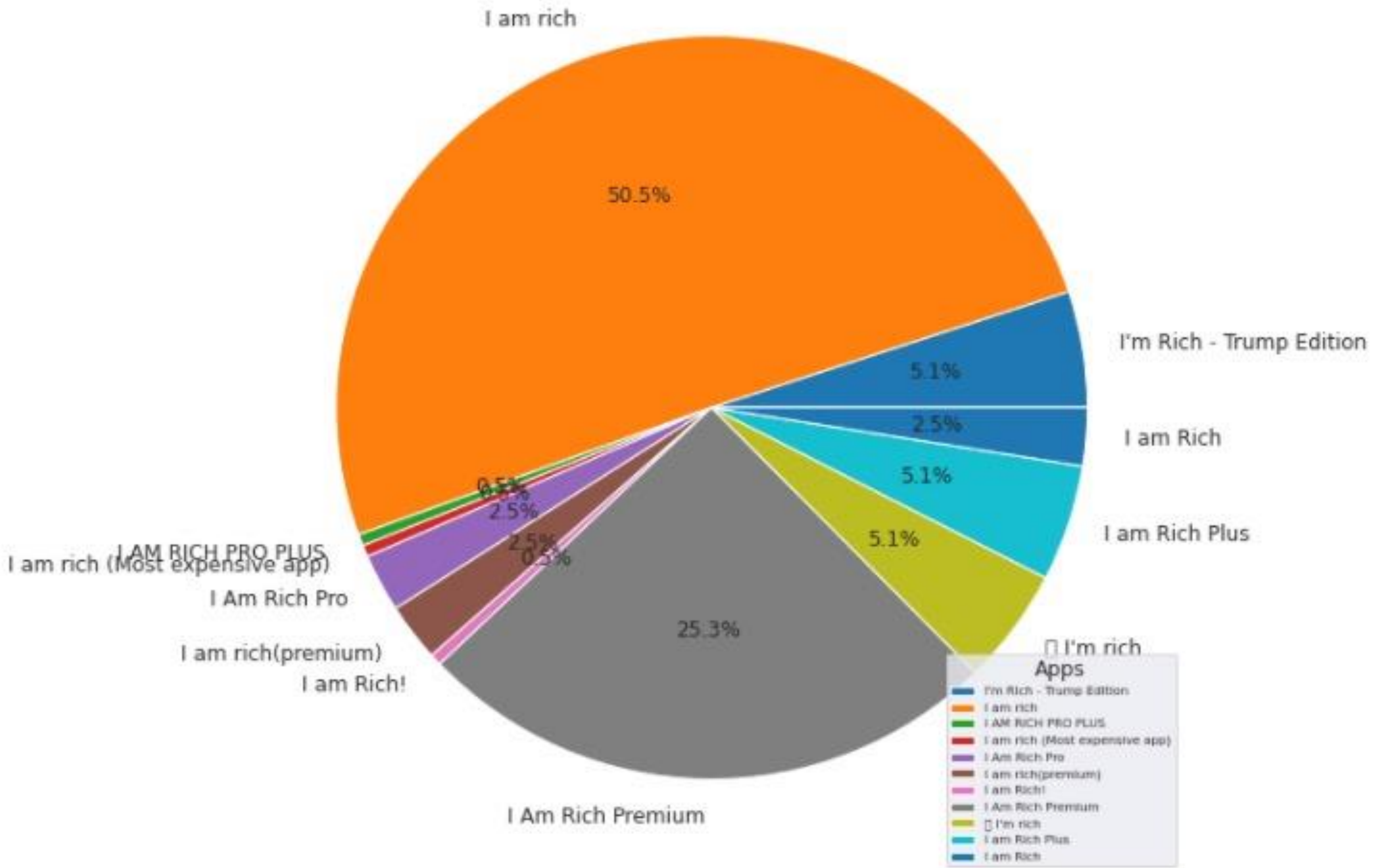
Findings:

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

What portion of the apps in Play Store are paid and free?

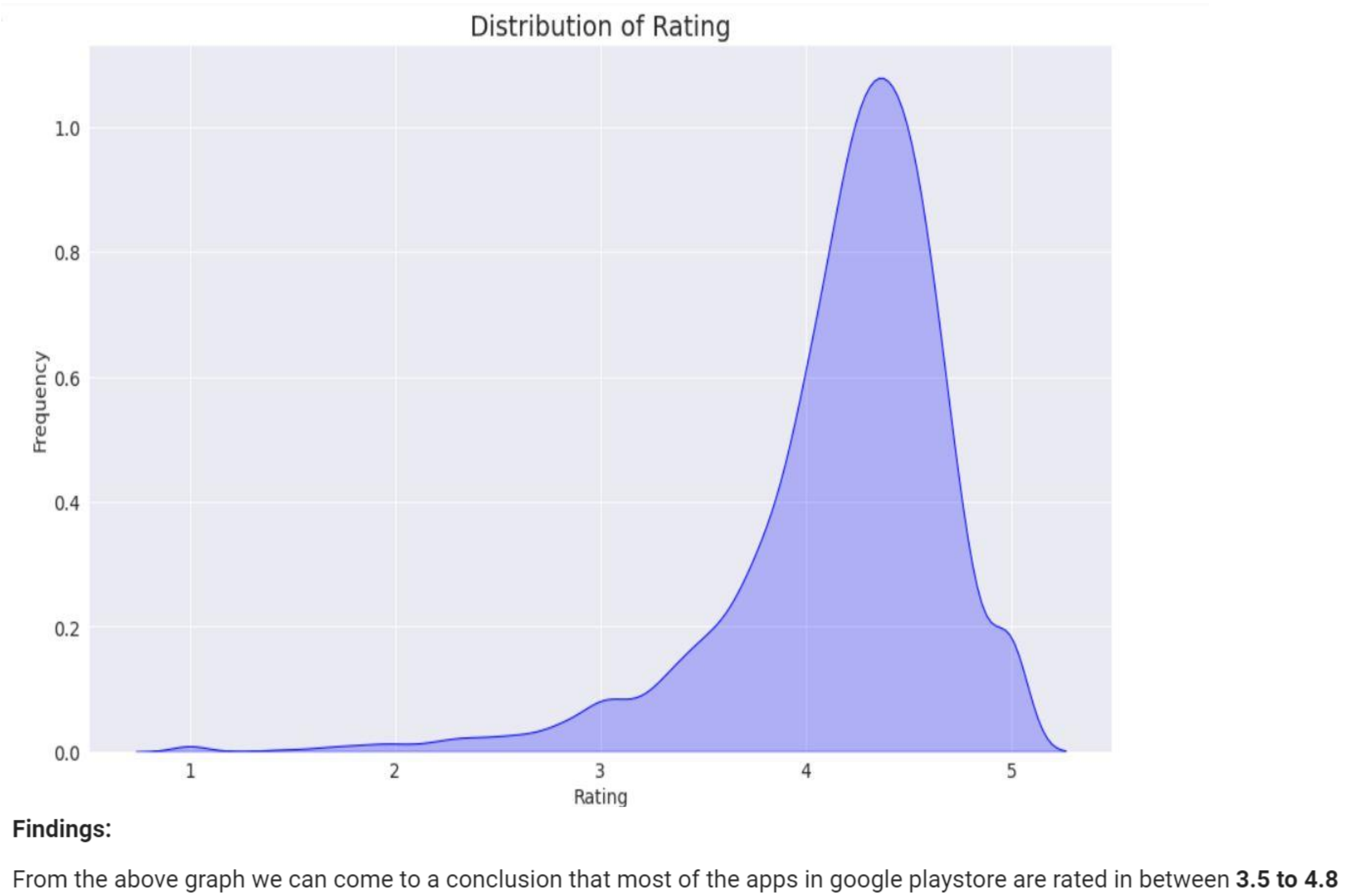


Findings:
From the above graph we can see that **93%**(Approx.) of apps in google play store are **free** and **7%**(Approx.) are **paid**.



Findings:
From the above graph we can interpret that the **I am Rich** app is the **most expensive app** in the play_store. But this seems to be like a junk app. We need to further analyse if it is a junk app or not by deploying machine learning models in it

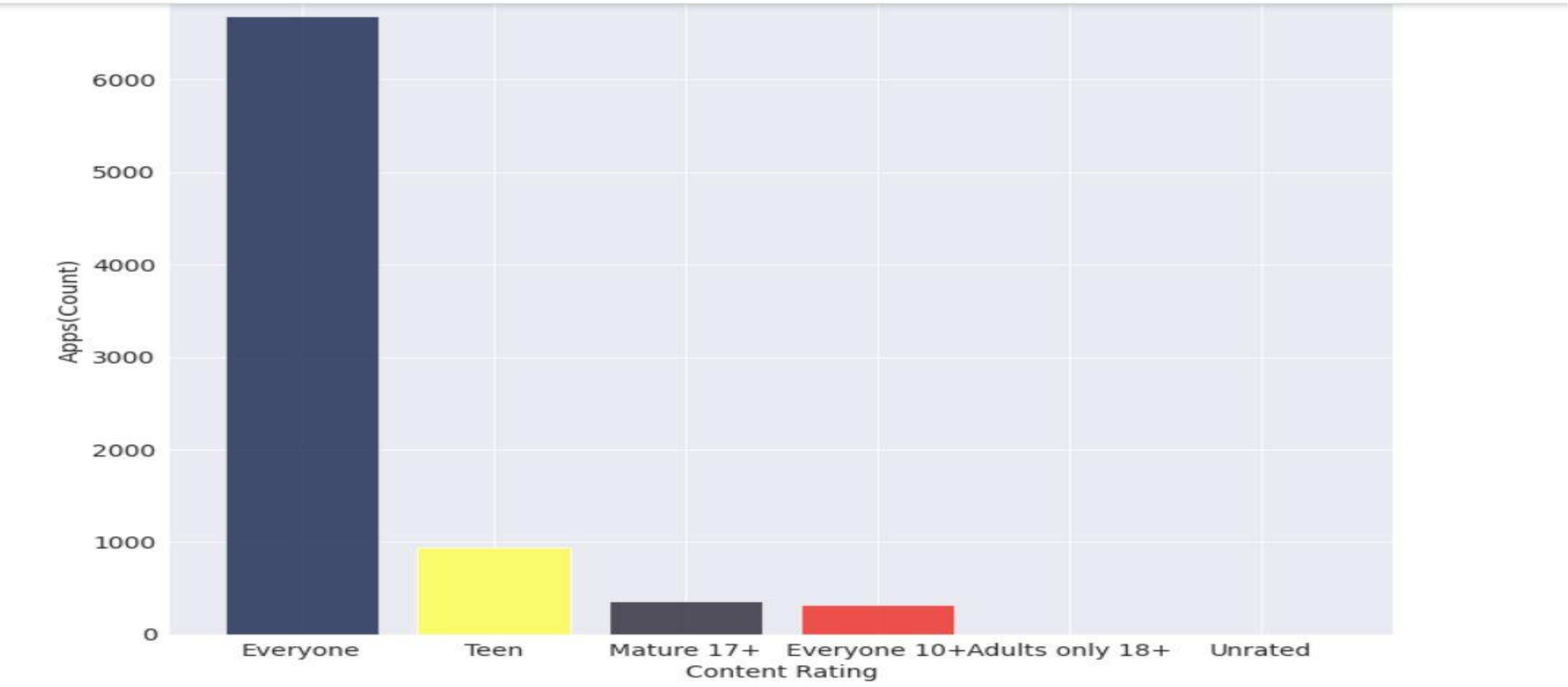
Distribution of the ratings of the apps



Findings:

From the above graph we can come to a conclusion that most of the apps in google playstore are rated in between **3.5 to 4.8**

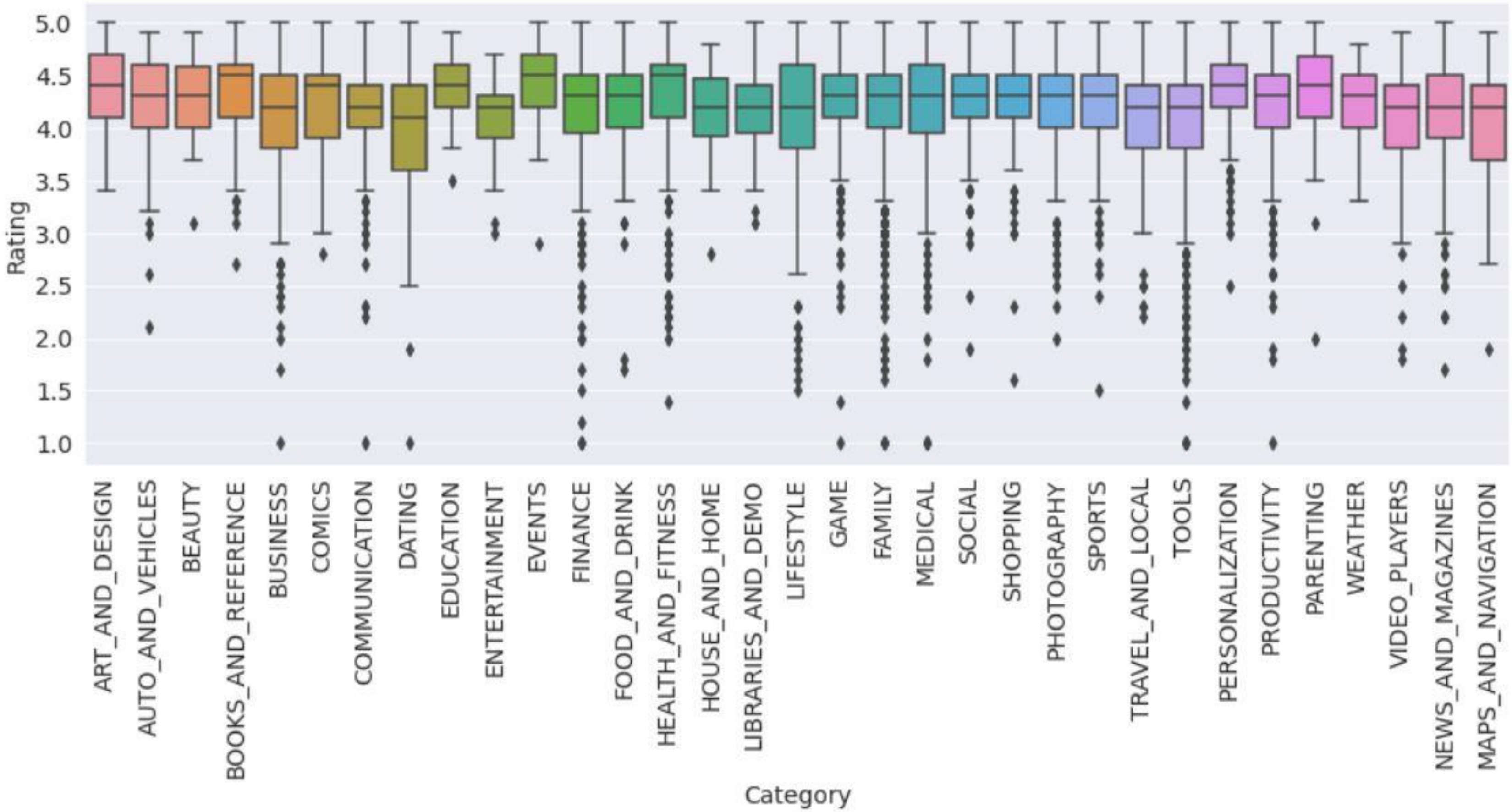
Which category of Apps from the Content Rating column are found more on Play Store?



Findings:

From the above plot we can see that **Everyone** category has the **highest** number of apps.

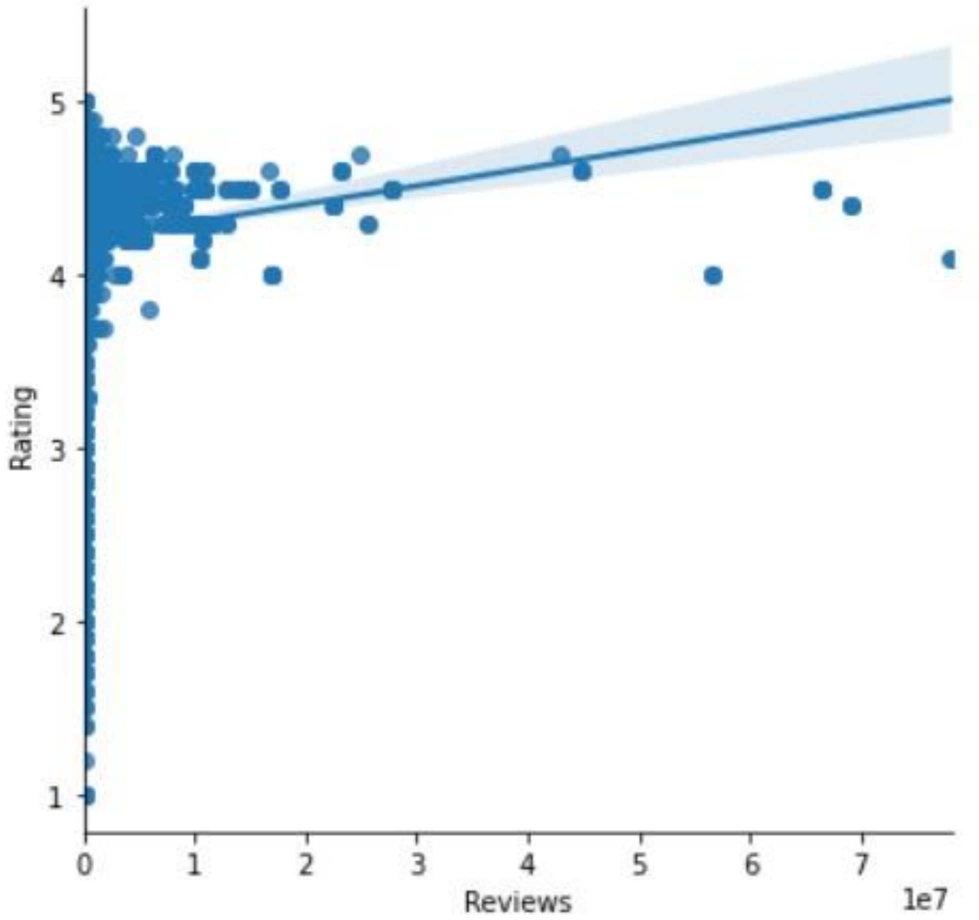
Let us see how the median ratings vary for categories?



Findings:

For each and every categories the median rating is lying in between 4 to 4.5

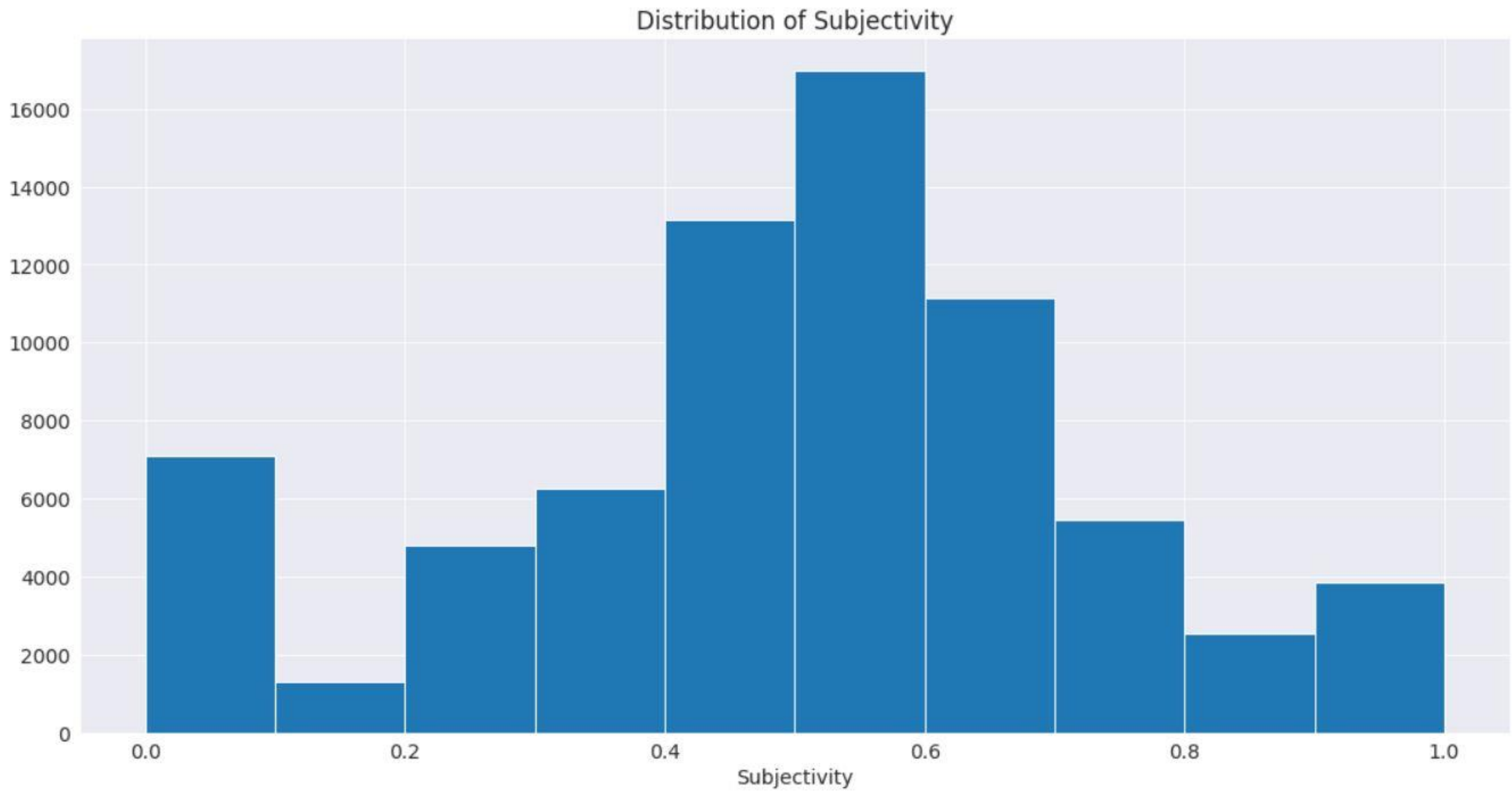
Does more number of reviews means more ratings?



Findings:
From the above plot, we cannot say that there is a relation, it seems that irrespective of the Reviews, the ratings are majorly between 4 and 5, which we also noticed before.

Also it is not correct to assume that rating and reviews have a relationship because reviews can be positive or negative and increase in the number of reviews does not show whether. the Reviews are Positive or Negative.

Histogram of Subjectivity



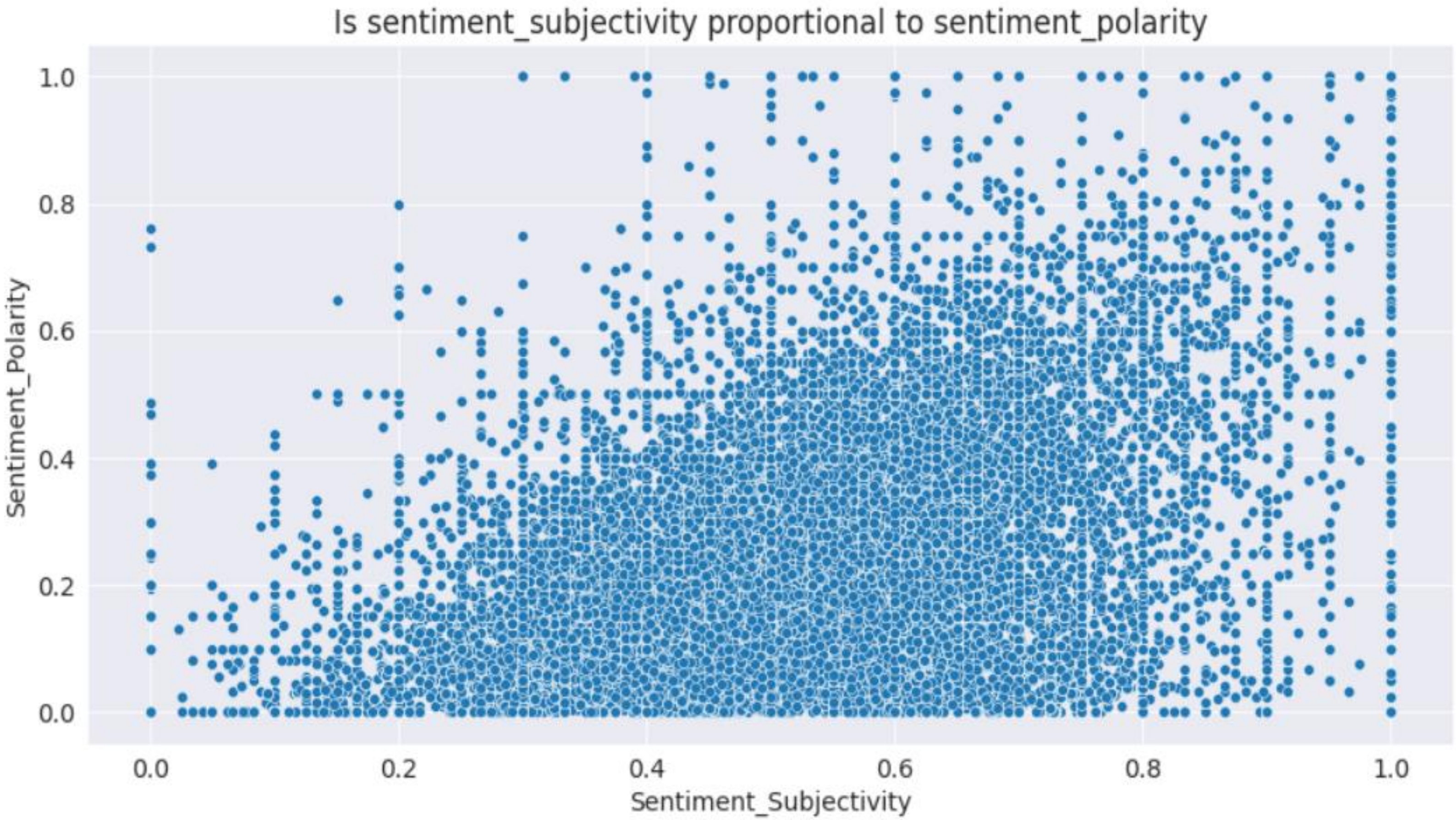
Findings:

0 - objective(fact)

1 - subjective(opinion)

It can be seen that maximum number of sentiment subjectivity lies between 0.4 to 0.7. From this we can conclude that maximum number of users give reviews to the applications, according to their experience.

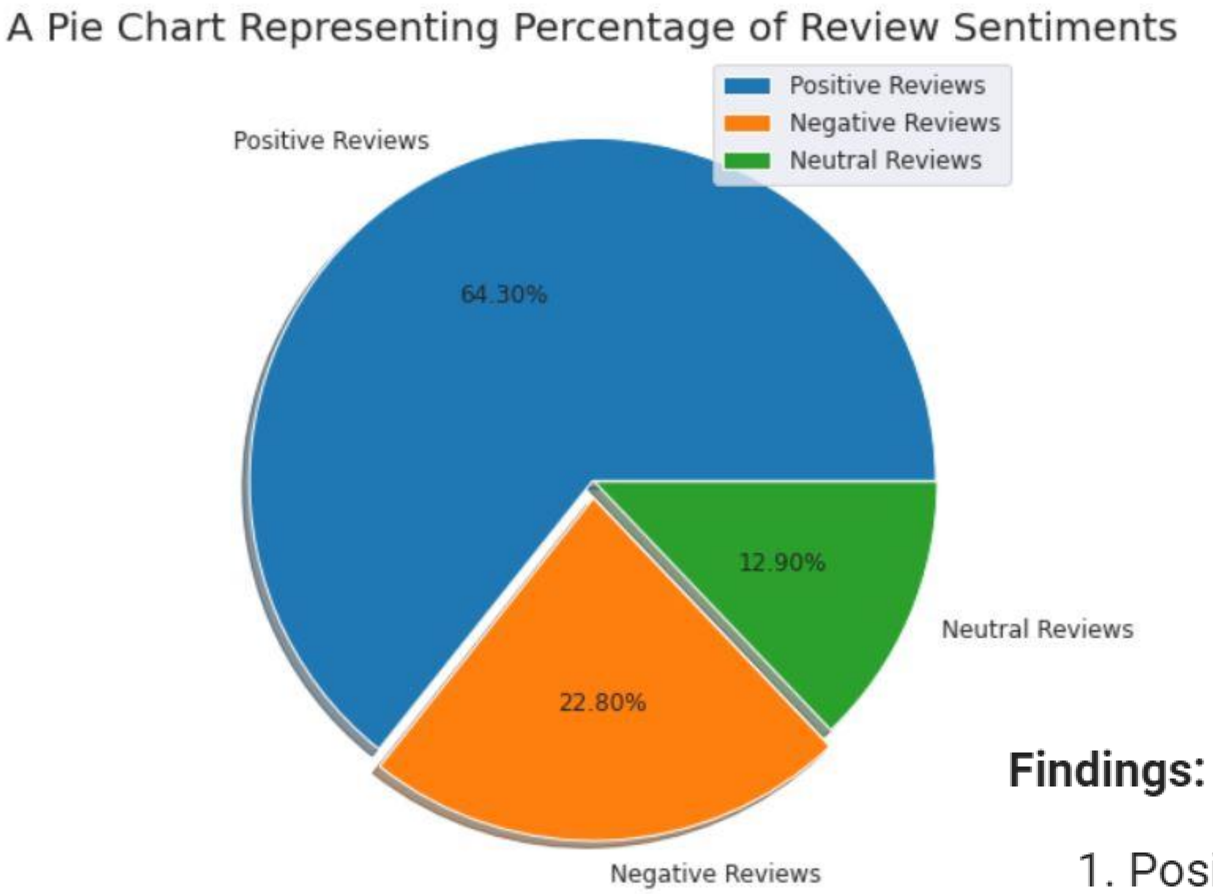
Is sentiment subjectivity proportional to sentiment polarity?



Findings:

From the above scatter plot it can be concluded that sentiment subjectivity is not always proportional to sentiment polarity but in maximum number of cases, it shows a proportional behavior, when variance is too high or low.

Percentage of Review Sentiments



Findings:

1. Positive reviews are **64.30%**
2. Negative reviews are **22.80%**
3. Neutral reviews are **12.90%**

Conclusion

- In this project of analyzing play store applications, we have worked on several parameters which would help AlmaBetter to do well in launching their apps on the play store.
- 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.
- AlmaBetter needs to focus more on :
    - Developing apps related to the least categories as they are not explored much. Like events and beauty.
    - Most of the apps are Free, so focusing on free app is more important.
    - Focusing more on content available for Everyone will increase the chances of getting the highest installs.
    - They need to focus on updating their apps regularly, so that it will attract more users.
    - They need to keep in mind that the sentiments of the user keep varying as they keep using the app, so they should focus more on users needs and features.

Challenges & Future Work

- Our major challenge was data cleaning.
- 13.60% of reviews were NaN values, and even after merging both the dataframes, we could not infer much in order to fill them. Thus we had to drop them.
- The merged data frame of both play store and user reviews, had only 816 common apps. This is just 10% of the cleaned data, we could have given more valuable analysis, if we had atleast 70% - 80% of the data available in the merged dataframes.
- User Reviews had 42% of NaN values, which could have been used for developing an understanding of the category wise sentiments, which would help us to fill 13.60% NaN values of the Reviews column.
- There is so much more which can be explored. Like we have current version, android version available which can be explored in detail and we can come out with more analysis where we can tell how does these things effect and needs to be kept in mind while developing app for the users.
- We can explore the correlation between the size of the app and the version of Android on the number of installs.
- Machine learning can help us to deploy more insights by developing models which can help us interpret even more better. We have left this as future work as this is something where we can work on.

Any Questions ???