

Capstone Project - 1

Play Store App Review Analysis

Member
Kousik Dutta

Contents

- 1. Problem Statement**
- 2. Data Summary**
- 3. Most Popular Apps**
- 4. App Pricing Per Category**
- 5. Dependent Variable**
- 6. Data Cleaning**
- 7. Data Visualization**
- 8. Conclusion**
- 9. Challenges**
- 10. References**
- 11. Q&A**

Problem Statement

We have two datasets in this project -

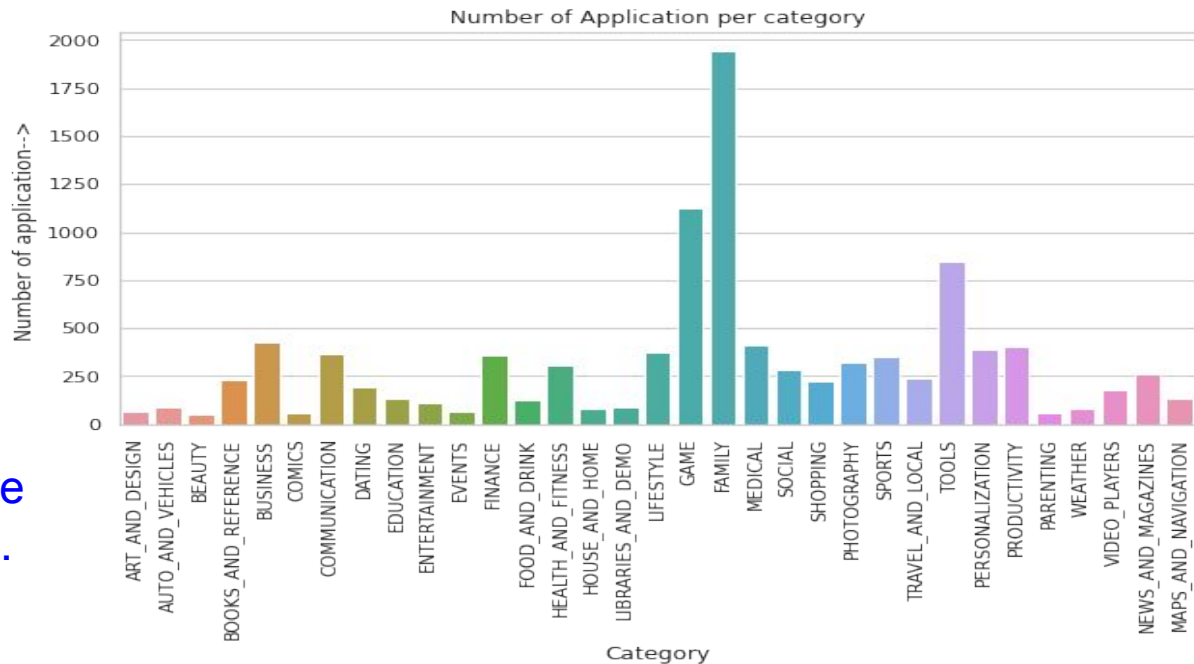
1. **Play Store App Data**
2. **User Review Data**

Here we are going to analyze different aspects of Google Play Store Applications. We will perform analysis on different attributes of the data such as Number of reviews Categories, Number of Installs to find out which applications and categories are at top of the application , the type of applications developers should develop etc.

Data Summary

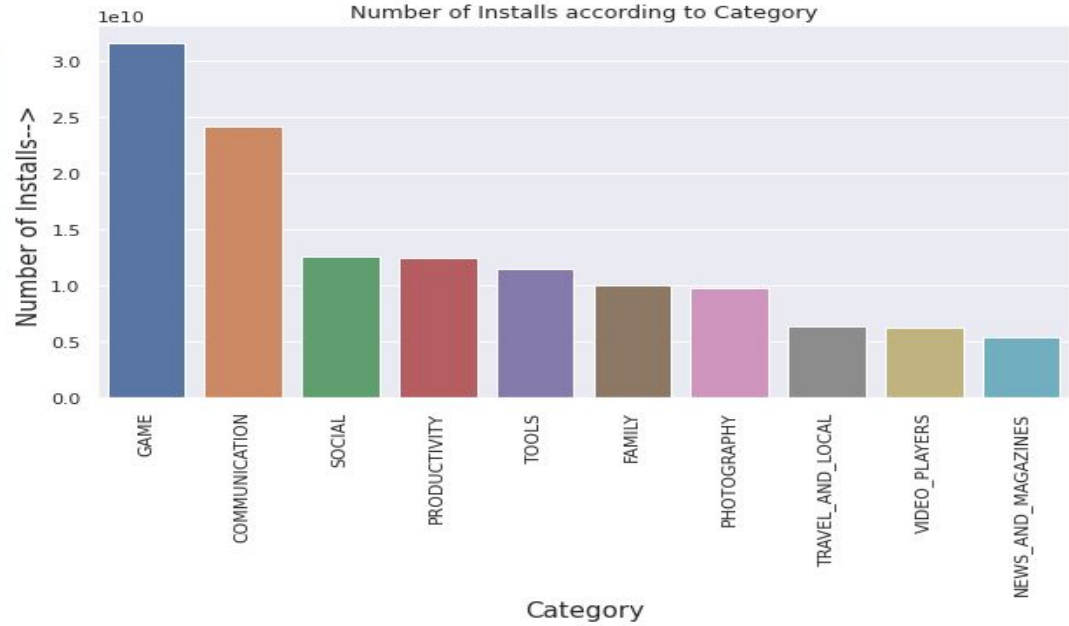
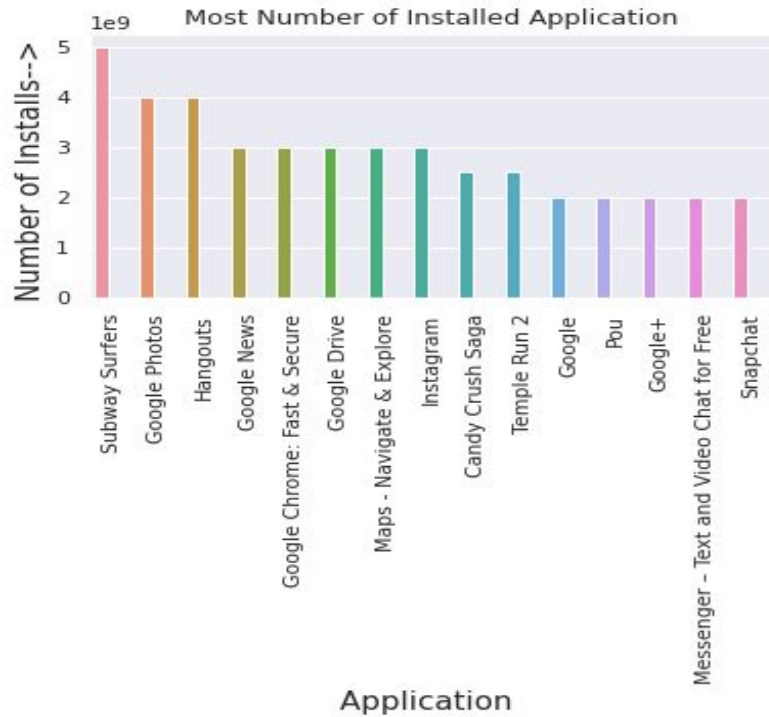
92%

Of apps on the play store
Are **Free** to Download.



The Family Category has the most apps on the google play store.

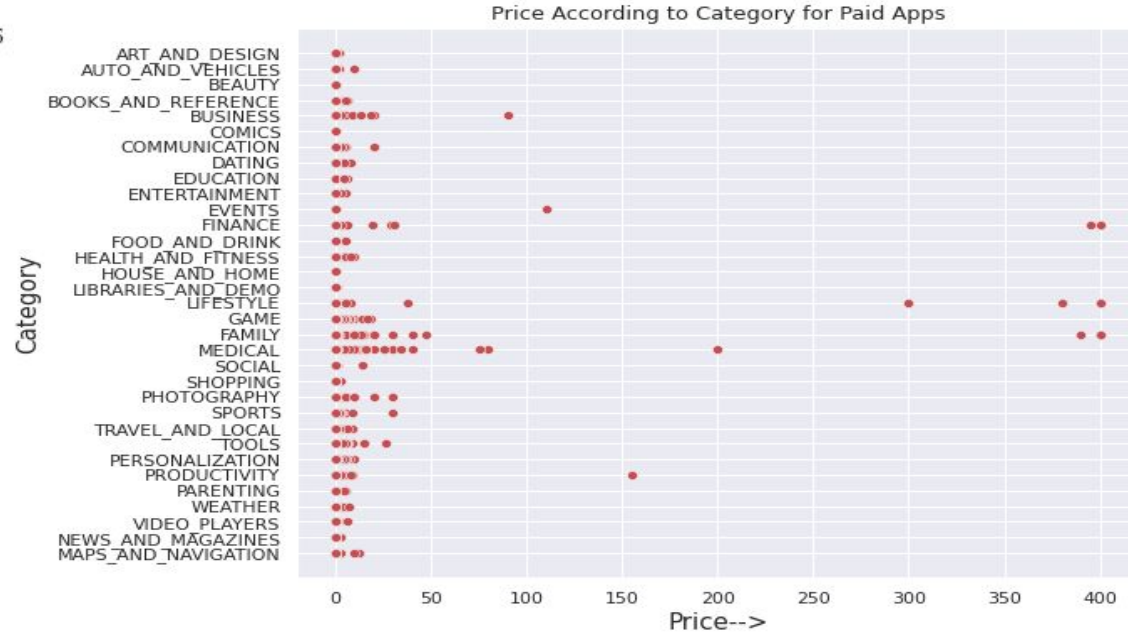
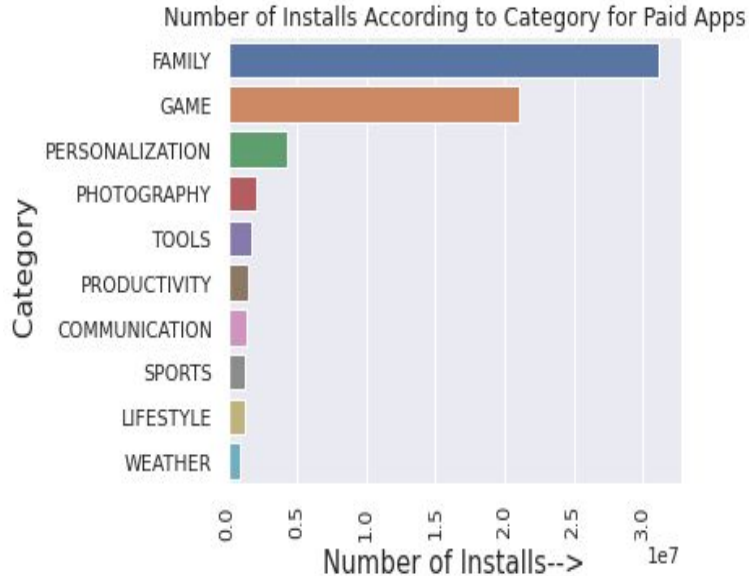
Most Popular Apps



“Subway Surfers”, “Google Photos” & “Hangouts” these are the top 3 most downloaded app.

The Games category contains more apps with larger Installs.

App Pricing Per Category



“Family” & “Game” Category Paid apps have the highest number of installs.

The **“Lifestyle”, “Finance” & “Family”** categories have the apps with the highest cost upto **\$400**.

Dependent Variable

Google Play Store Dataset -

- App
- Category
- Rating
- Reviews
- Size
- Installs
- Type
- Price
- Content Rating
- Genres

User Review dataset -

- Sentiment
- Sentiment Subjectivity

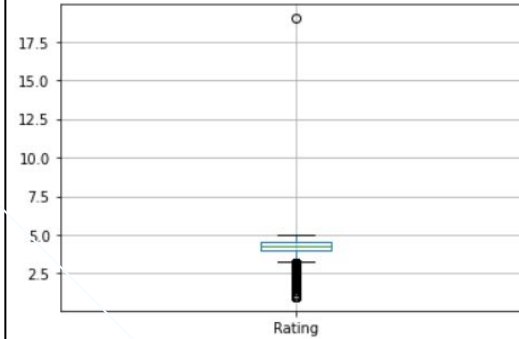
Data Cleaning

Rating

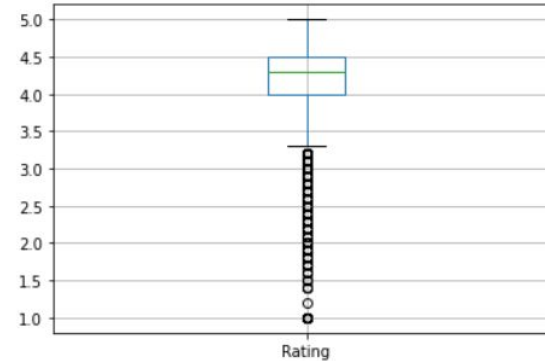
```
google_playstore_data.describe()
```

	Rating
count	9367.000000
mean	4.193338
std	0.537431
min	1.000000
25%	4.000000
50%	4.300000
75%	4.500000
max	19.000000

```
google_playstore_data.boxplot();
```



```
google_playstore_data.boxplot();
```



Data Cleaning



Null Values

```
google_playstore_data.isnull().sum()
```

App	0
Category	0
Rating	1474
Reviews	0
Size	0
Installs	0
Type	1
Price	0
Content Rating	0
Genres	0
Last Updated	0
Current Ver	8
Android Ver	2
dtype: int64	



```
google_playstore_data.isnull().sum()
```

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	

Data Cleaning

Special Symbols

```
google_playstore_data["Price"].value_counts()
```

```
0          10040
$0.99       148
$2.99       129
$1.99        73
$4.99        72
...
$19.90         1
$1.75          1
$14.00          1
$4.85           1
$1.04           1
Name: Price, Length: 92, dtype: int64
```

```
google_playstore_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 10840 entries, 0 to 10840
Data columns (total 13 columns):
#   Column          Non-Null Count  Dtype
---  -
0   App             10840 non-null  object
1   Category        10840 non-null  object
2   Rating          10840 non-null  float64
3   Reviews         10840 non-null  int64
4   Size            10840 non-null  object
5   Installs        10840 non-null  float64
6   Type            10840 non-null  object
7   Price           10840 non-null  float64
8   Content Rating  10840 non-null  object
9   Genres          10840 non-null  object
10  Last Updated    10840 non-null  object
11  Current Ver     10840 non-null  object
12  Android Ver     10840 non-null  object
dtypes: float64(3), int64(1), object(9)
memory usage: 1.2+ MB
```

```
google_playstore_data["Installs"].value_counts()
```

```
1,000,000+      1579
10,000,000+     1252
100,000+        1169
10,000+         1054
1,000+          907
5,000,000+      752
100+            719
500,000+        539
50,000+         479
5,000+          477
100,000,000+    409
10+             386
500+            330
50,000,000+     289
50+             205
5+              82
500,000,000+    72
1+              67
1,000,000,000+  58
0+             14
0              1
Name: Installs, dtype: int64
```

Data Cleaning

Removal of Duplicate Data

```
google_playstore_data[google_playstore_data["App"]=="ESPN"]
```

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
2959	ESPN	SPORTS	4.2	521138	0.0	10000000.0	Free	0.0	Everyone 10+	Sports	July 19, 2018	Varies with device	5.0 and up
3010	ESPN	SPORTS	4.2	521138	0.0	10000000.0	Free	0.0	Everyone 10+	Sports	July 19, 2018	Varies with device	5.0 and up
3018	ESPN	SPORTS	4.2	521138	0.0	10000000.0	Free	0.0	Everyone 10+	Sports	July 19, 2018	Varies with device	5.0 and up
3048	ESPN	SPORTS	4.2	521140	0.0	10000000.0	Free	0.0	Everyone 10+	Sports	July 19, 2018	Varies with device	5.0 and up
3060	ESPN	SPORTS	4.2	521140	0.0	10000000.0	Free	0.0	Everyone 10+	Sports	July 19, 2018	Varies with device	5.0 and up
3072	ESPN	SPORTS	4.2	521140	0.0	10000000.0	Free	0.0	Everyone 10+	Sports	July 19, 2018	Varies with device	5.0 and up
4069	ESPN	SPORTS	4.2	521081	0.0	10000000.0	Free	0.0	Everyone 10+	Sports	July 19, 2018	Varies with device	5.0 and up

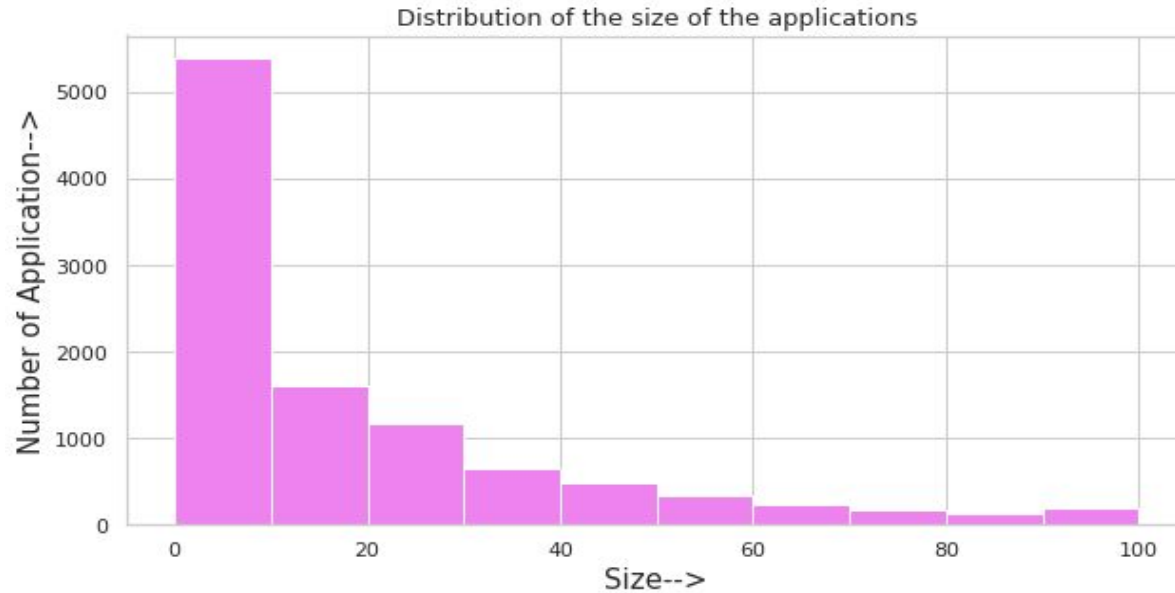


```
google_playstore_data[google_playstore_data["App"]=="ESPN"]
```

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
2959	ESPN	SPORTS	4.2	521138	0.0	10000000.0	Free	0.0	Everyone 10+	Sports	July 19, 2018	Varies with device	5.0 and up
3048	ESPN	SPORTS	4.2	521140	0.0	10000000.0	Free	0.0	Everyone 10+	Sports	July 19, 2018	Varies with device	5.0 and up
4069	ESPN	SPORTS	4.2	521081	0.0	10000000.0	Free	0.0	Everyone 10+	Sports	July 19, 2018	Varies with device	5.0 and up

Data Visualization

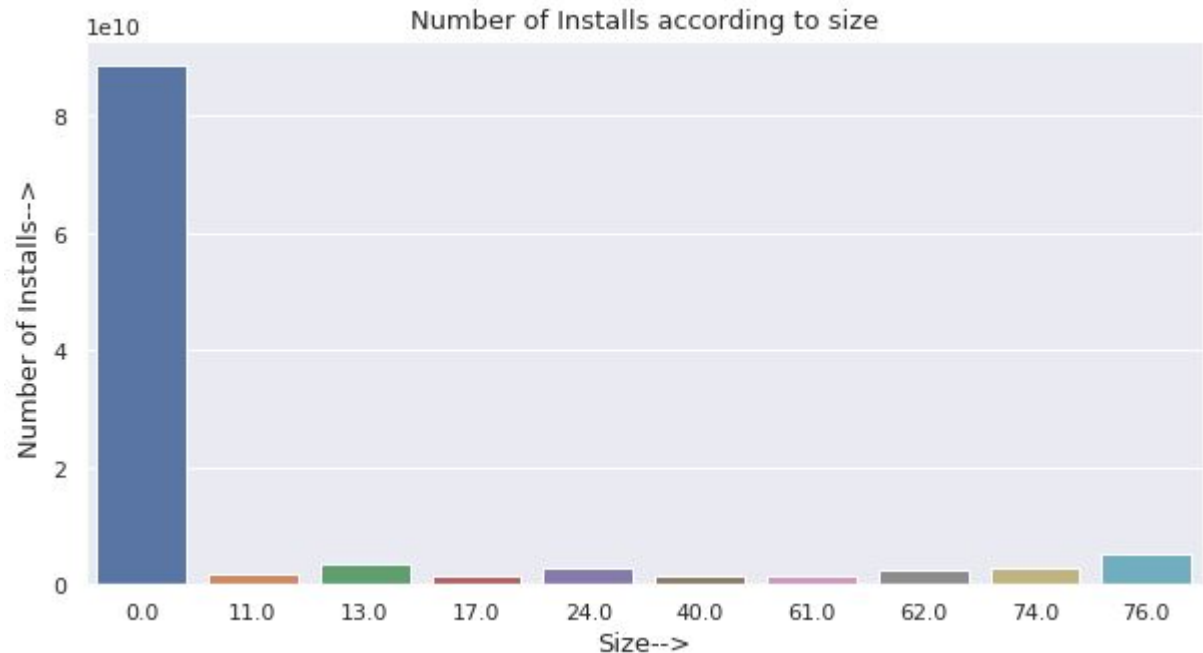
Distribution of Size



Data Visualization

App Installation According to Size

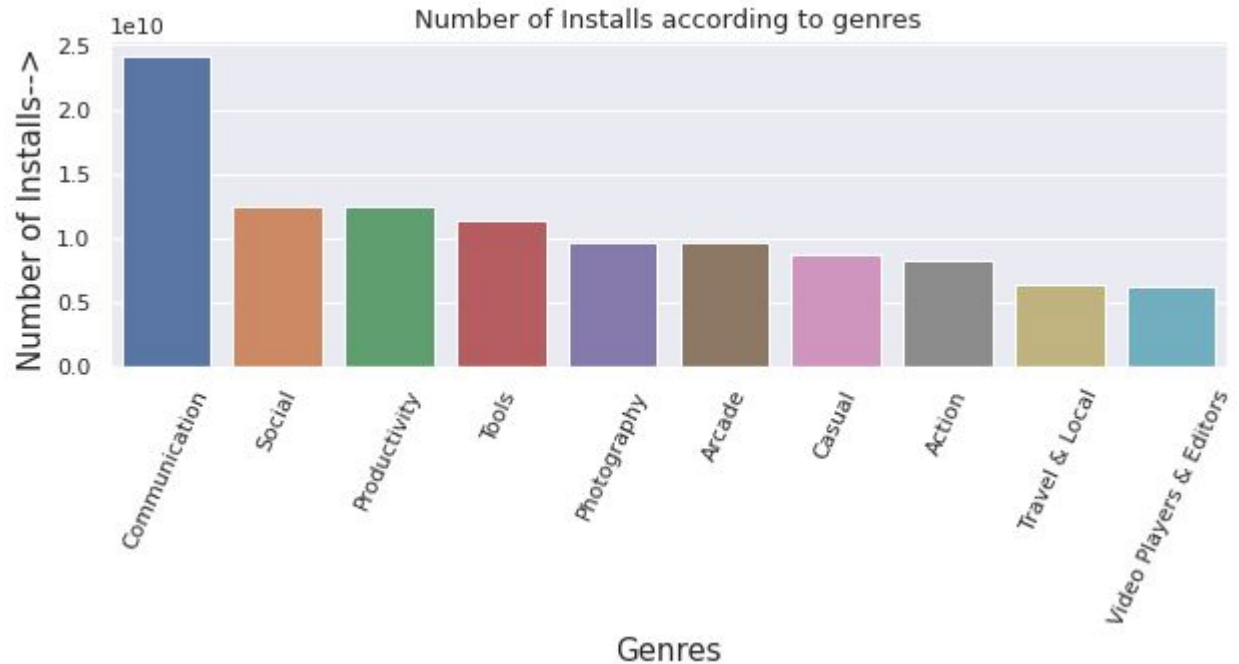
Most of the installed applications are of small sized.



Data Visualization

App installation according to Genres

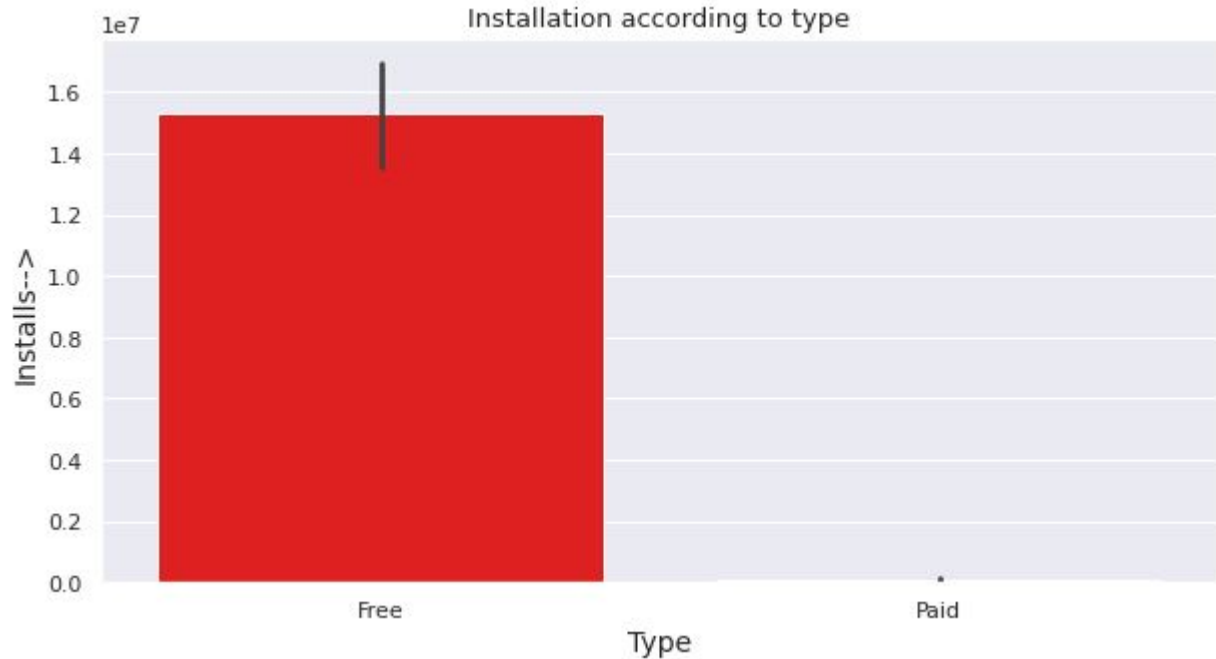
The most popular applications are under **Communication** Genres.



Data Visualization

App installation according to Type

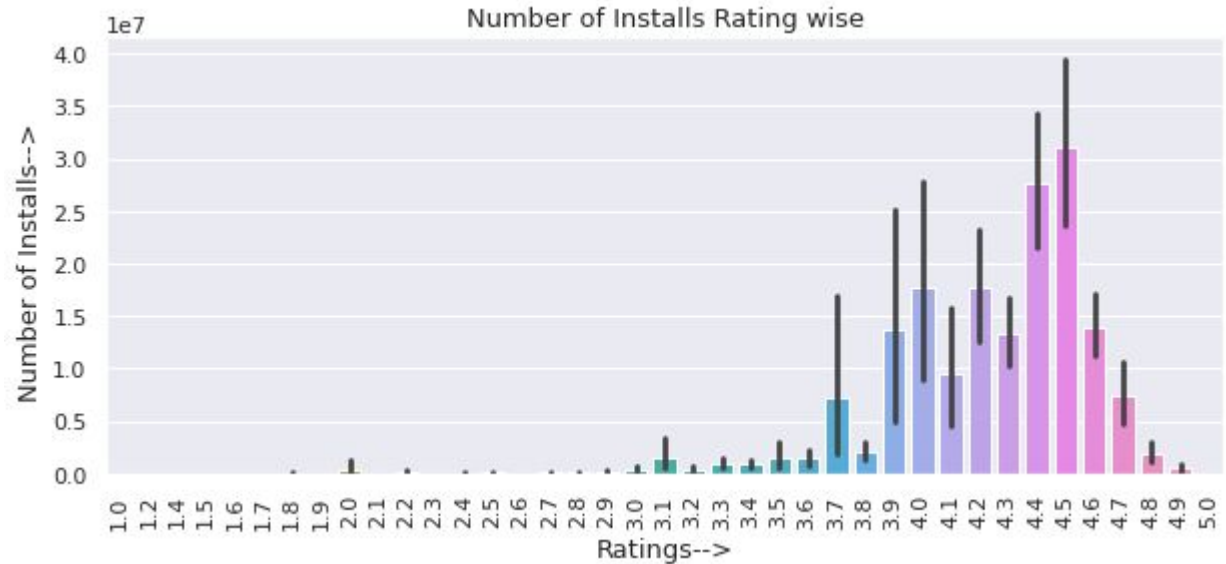
Free applications
Installation graph
is high.



Data Visualization

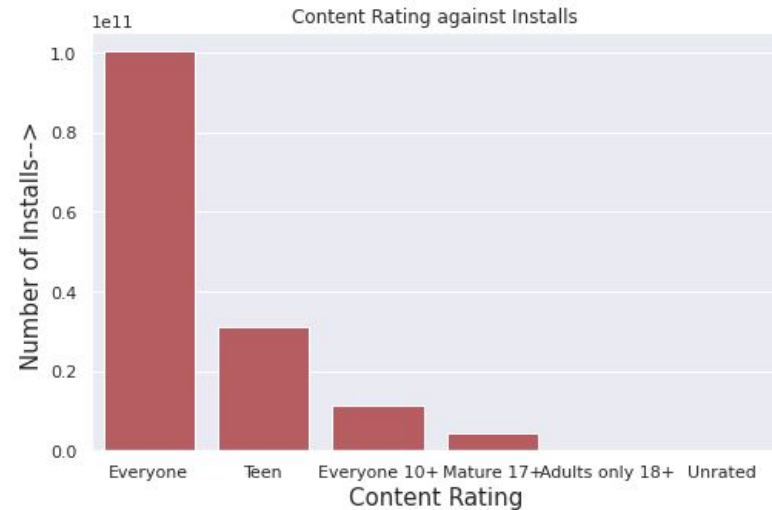
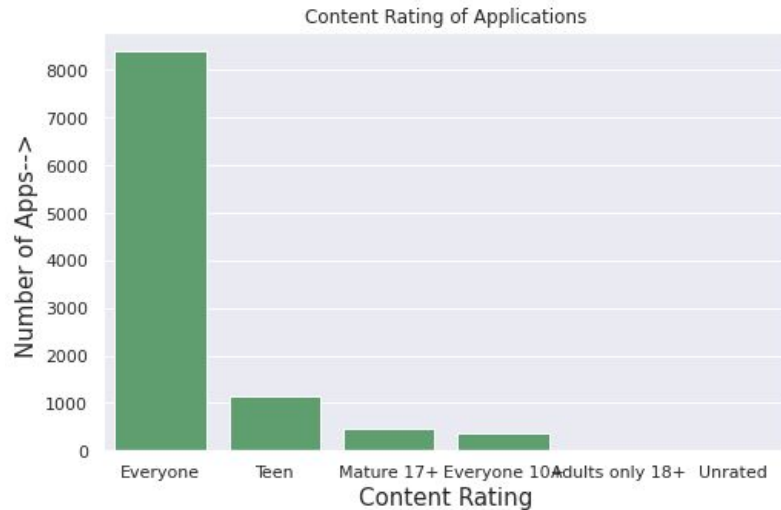
Apps installation according to Rating

User prefer the highest rating applications.



Data Visualization

Content Rating of Apps

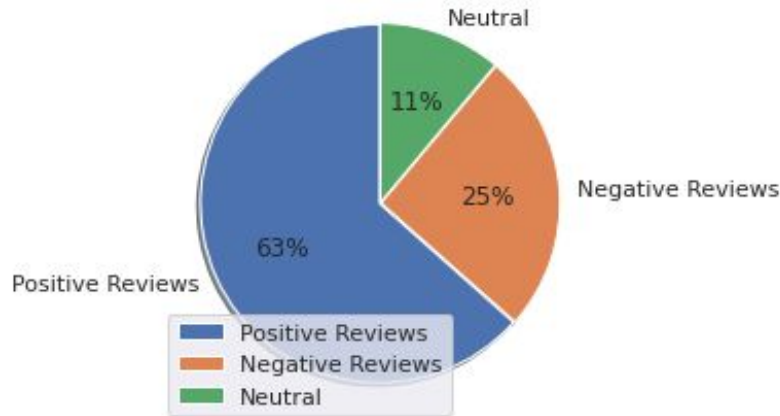


"Everyone" category has the highest number of application and Installs

Data Visualization

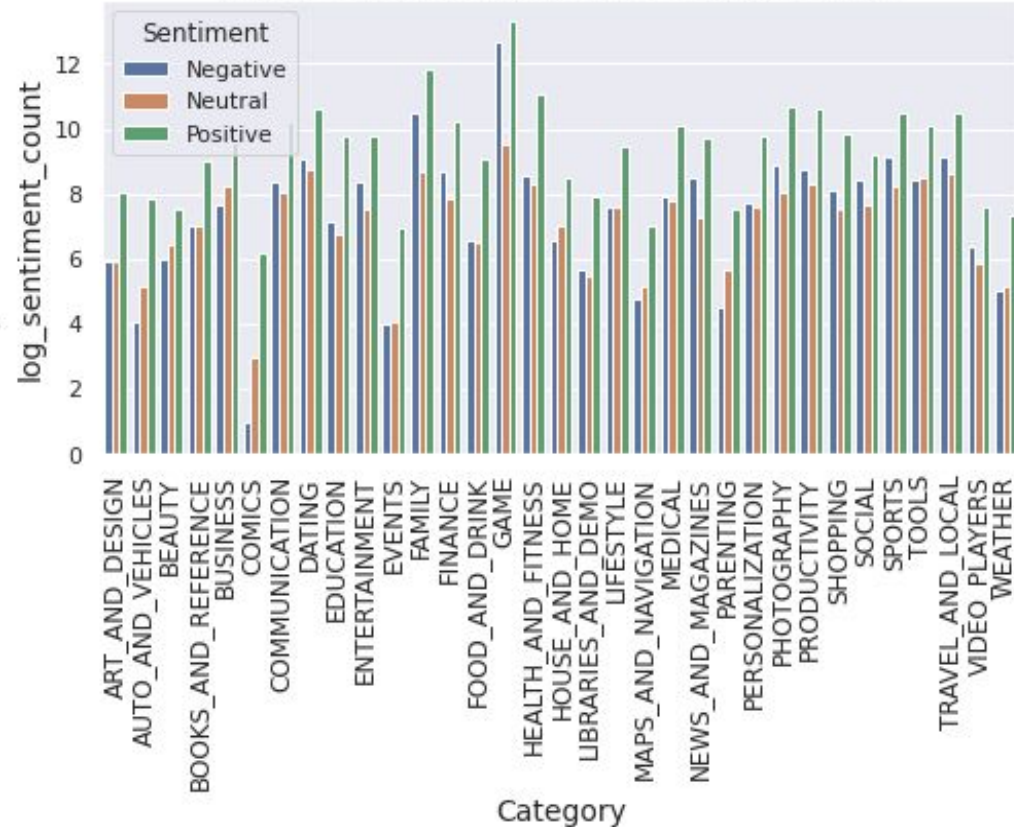
Types of reviews category wise

Pie chart for percentage of review sentiments



Games category has the highest positive and negative sentiment.

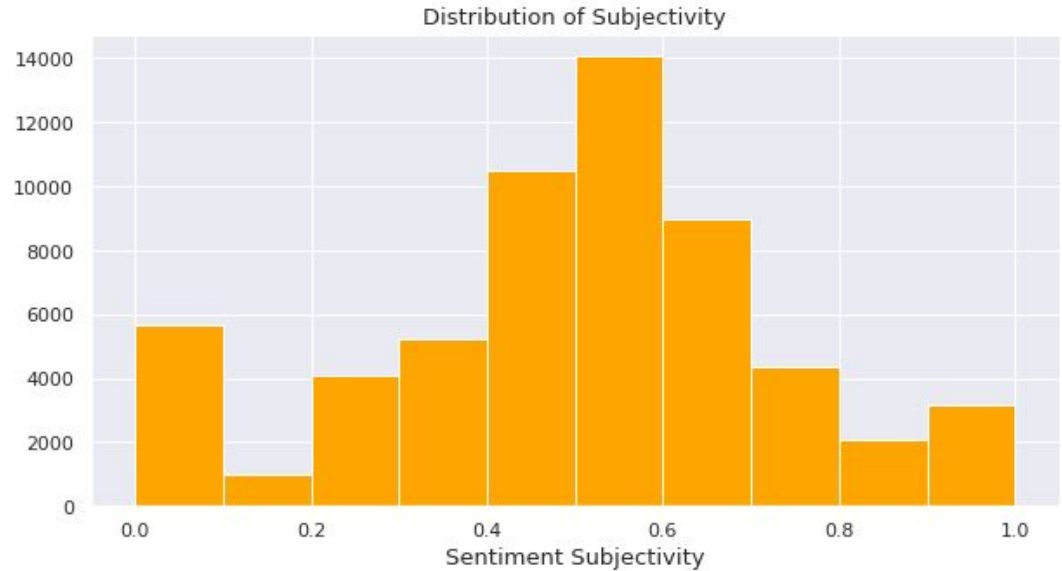
Number of Installs sentiment wise according to Genres



Data Visualization

Sentiment Subjectivity

Maximum number of sentiment subjectivity lies between 0.4 to 0.7



Conclusion

- Users prefer to install free applications more.
- Communication, Entertainment, Game, Social category apps are in high demand for all type of users.
- Users installs the application depends on the previous rating and reviews as we have seen that higher ratings are having higher installs.
- Maximum applications content ratings are for everyone.
- we learnt that the current trend in the android market are mostly from these categories which either assisting, communicating or entertaining apps.

Challenges

- ❖ Cleaning the null values from the dataset.
- ❖ Understanding the sentiments of the user

References

- [GeeksforGeeks](#)
- [Stackoverflow](#)
- [Seaborn Library](#)
- [Matplotlib Library](#)

Q & A