Group 16

Group Members:
Aishwarya Mathkar
Parin Patel
Vamsi Akhumukhi

Google Play Store

Data Analysis

Table of Contents

OVERVIEW

With the ever-increasing consumer's appetite for mobile applications, it's now imperative to acquire pivotal insights from the distribution platform which hosts the App. The percentage of mobile over desktop is only increasing. Android holds about 53.2% of the smartphone market, while iOS is 43%. To get more people to download the app, we need to make sure they can easily find our app. Mobile app analytics is a great way to understand the existing strategy to drive growth and retention of future user. With millions of apps around nowadays, the following data set has become very key to getting top trending apps in google app store. This data set contains more than 10k google play store mobile application details. This information helps in understanding user's expectation and market standing. The definition of success of the App differs from developer to developer, for this project we aim to predict the rating of an application.

Problem Statement

The definition of success of the App differs from developer to developer, for this project we aim to analyze the ratings and use models to find out which gives the best accuracy.

DATASET

Data Collection:

Data collection is the process of gathering and measuring data, information or any variables of interest in a standardized and established manner that enables the collector to answer or test hypothesis and evaluate outcomes of the particular collection. The dataset has been taken from the following website:
https://www.kaggle.com/palbha/google-play-store-data-analysis
The dataset consists of 10841 rows and 13 columns.

App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres
Photo Editor & Candy amera & Grid & apBook	ART_AND_DESIGN	4.1	159	19M	10,000+	Free	0	Everyone	Art & Design
Coloring book moana	ART_AND_DESIGN	3.9	967	14M	500,000+	Free	0	Everyone	Art & Design;Pretend Play
U auncher Lite – EE Live Cool hemes, Hide ...	ART_AND_DESIGN	4.7	87510	8.7M	5,000,000+	Free	0	Everyone	Art & Design

Dataset Description:

1. App: Name of the application

2. Category: Basically, shows the category the app belongs to. For example in our case it is 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.

3. Rating: Overall user rating of the app (as when scraped). The rating of the app is always between minimum 0 and maximum 5 ie an app can have ratings only between 0-5.

4. Reviews: Number of user reviews for the app (as when scraped). Basically, this column shows the number of time the app has been viewd by the viewers.

5. Size: Size of the app (as when scraped). It affects the download as viewers think how much space is it going to occupy before downloading any app.

6. Installs: It describes the number of user downloads/installs for the app (as and when scraped)

7. Type: Type describes whether the app is paid or whether it is free. Apps that are free are mostly preferred over the paid apps.

8. Price: If the app is paid then what is the price of the app (as when scraped)

9. Content Rating: This column basically describes the age group the app is targeted at. In our case we have – Everyone, Teen, Everyone 10+, Mature 17+, Adults only 18+ and unrated.

10. Genres: An app can belong to multiple genres (apart from its main category). For eg, a musical family game will belong to Music, Game, Family genres.

11. Last Updated: Date when the app was last updated on Play Store (as when scraped)

12. Current Ver: Current version of the app available on Play Store (as when scraped)

13. Android Ver: Min required Android version (as when scraped)

DATA CLEANING

Data cleansing or data cleaning is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database and refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying, or deleting the dirty or coarse data. The actual process of data cleansing may involve removing typographical errors or validating and correcting values against a known list of entities.

In our dataset we observed that the maximum rating of an app is 19 which is not possible as we know that the maximum rating for an app can be only be 5 and not more than that. There was something wrong with the data, so we compared and found the difference. By shifting the row right by 1 we got actual value of rating.

We also saw the the app “Life Made Wi-Fi Touchscreen Photo Frame” does not have any category. When we searched for the app on the play store, we found that it belonged to lifestyle category.

Initially, our data contained 10841 records and 13 fields. After removing, cleaning the data, now we have 9360 records (rows) and 13 fields (columns).

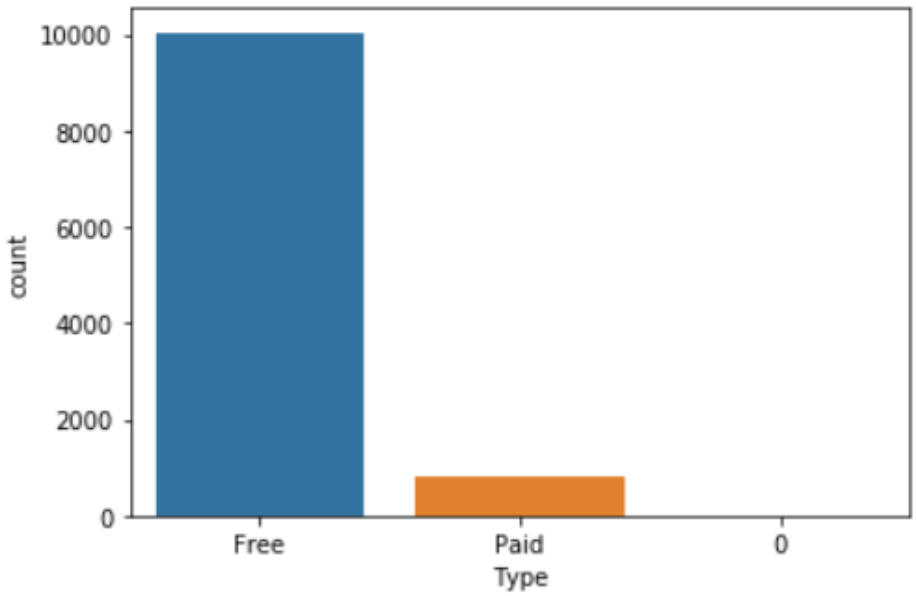
DATA VISUALIZATION

Visualization basically involves interpreting the data in visual terms. Data visualization is the graphic representation of data. It involves producing images that display the relationships among the represented data to viewers of the images. To communicate information clearly and efficiently, data visualization uses statistical graphics, plots, information graphics and other tools.

Type

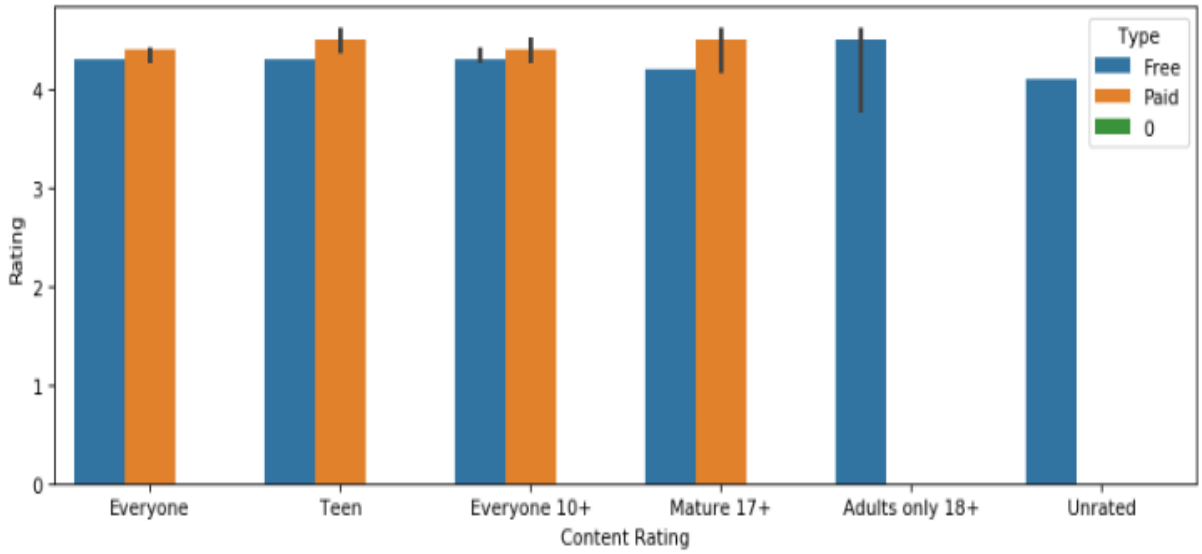
Visualizing using Seaborn Count plot on Type:

Here we can see that the apps that are freely available have more counts whereas the apps that are paid have less counts.

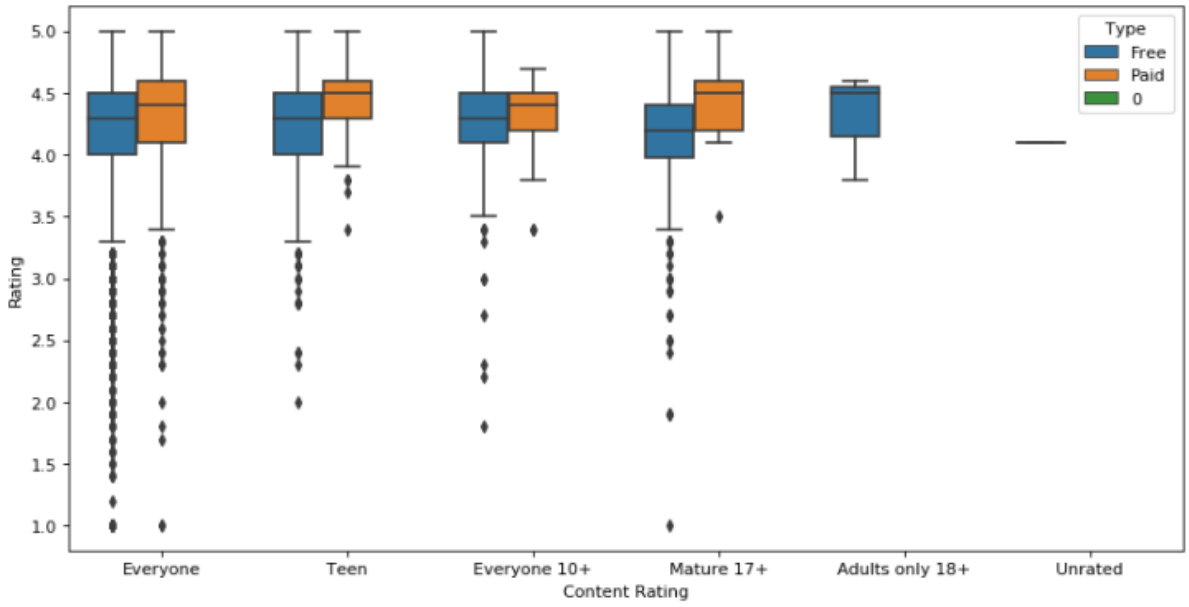


Visualizing using seaborn bar plot on Type and Rating:

Visualizing using sns bar plot on Content Rating and Rating



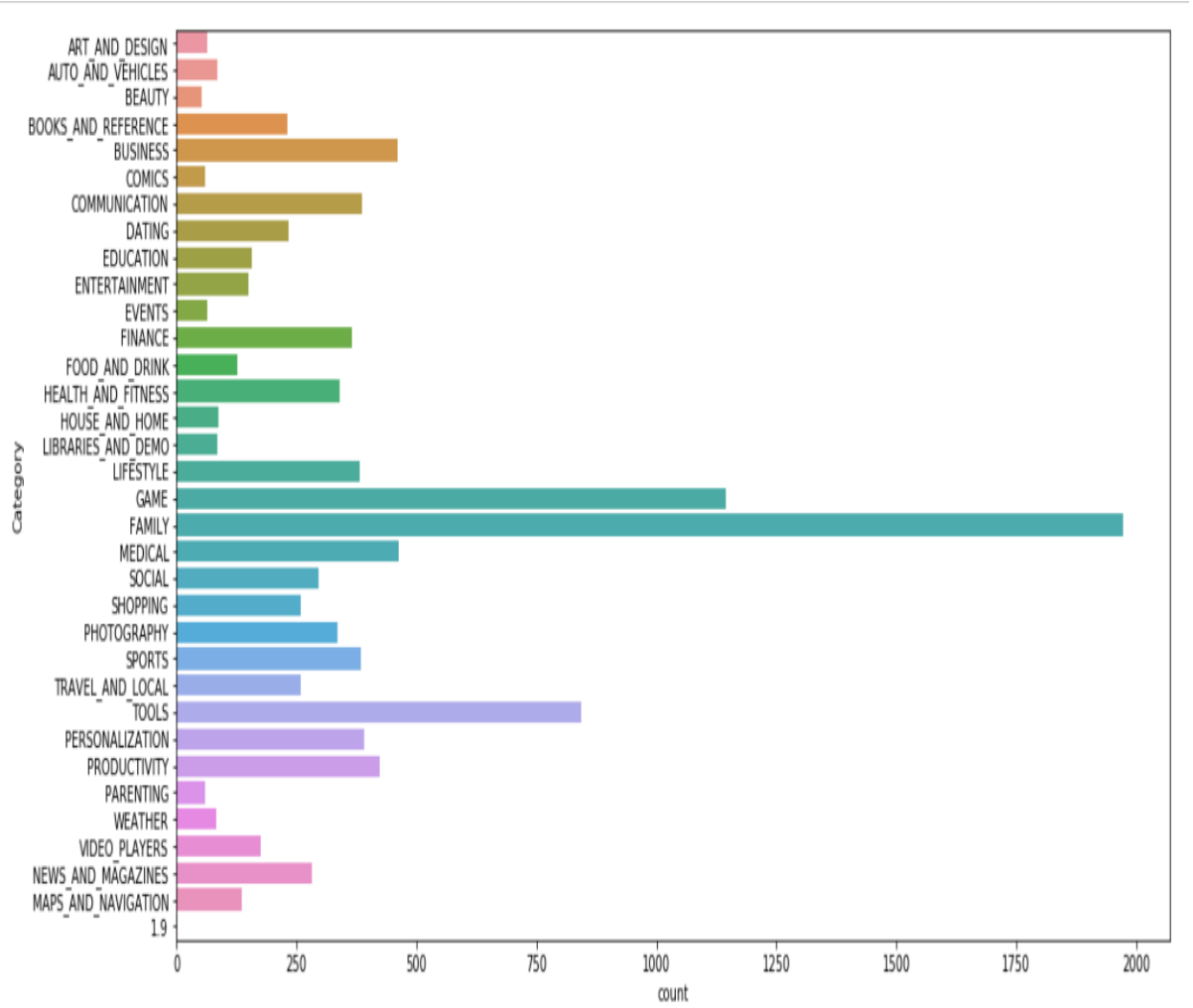
Visualizing using snd box plot on Content Rating and Rating



Category

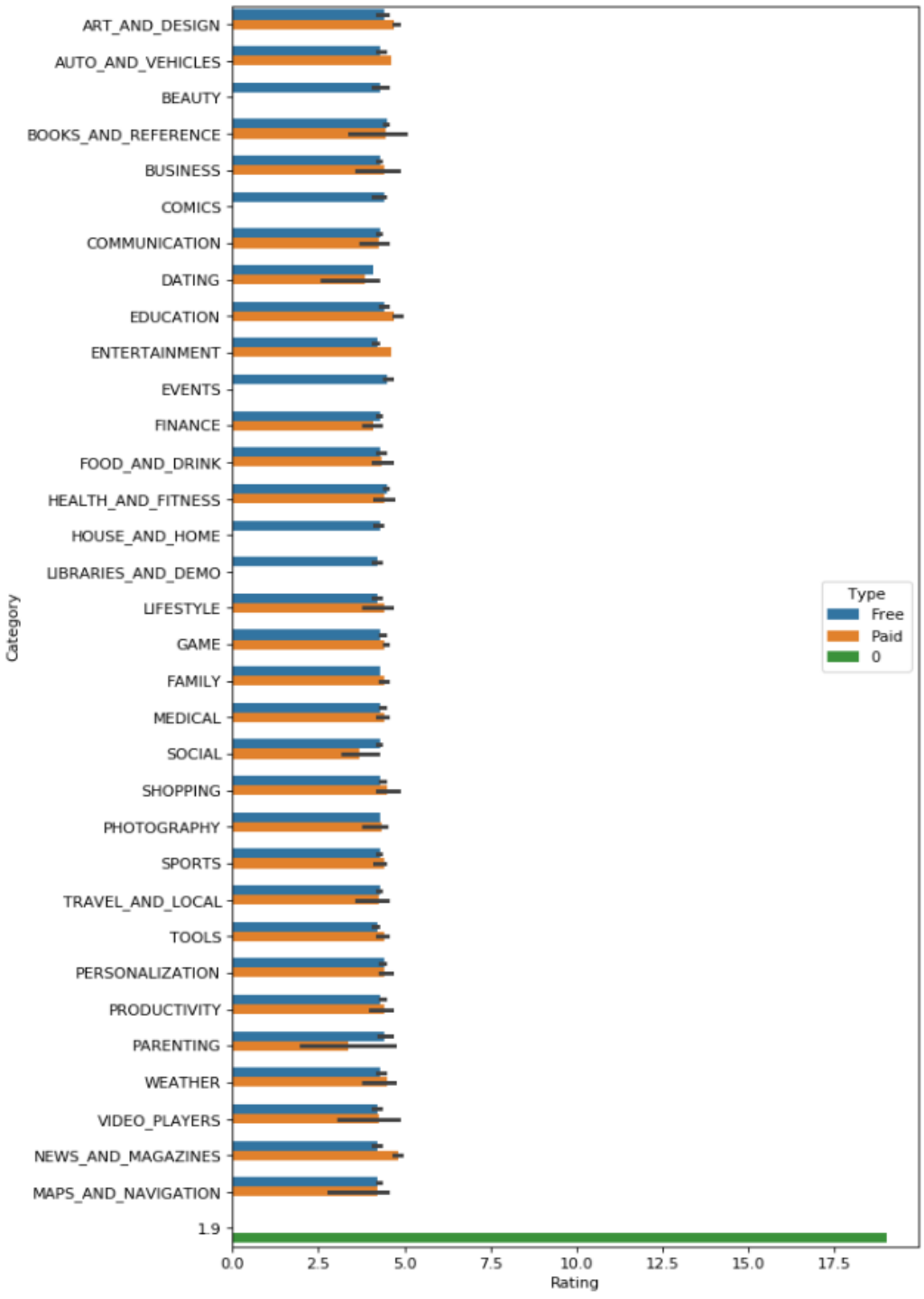
Visualizing Category and Count:

We observe that category Family has the maximum count followed by Game and then the Tools.



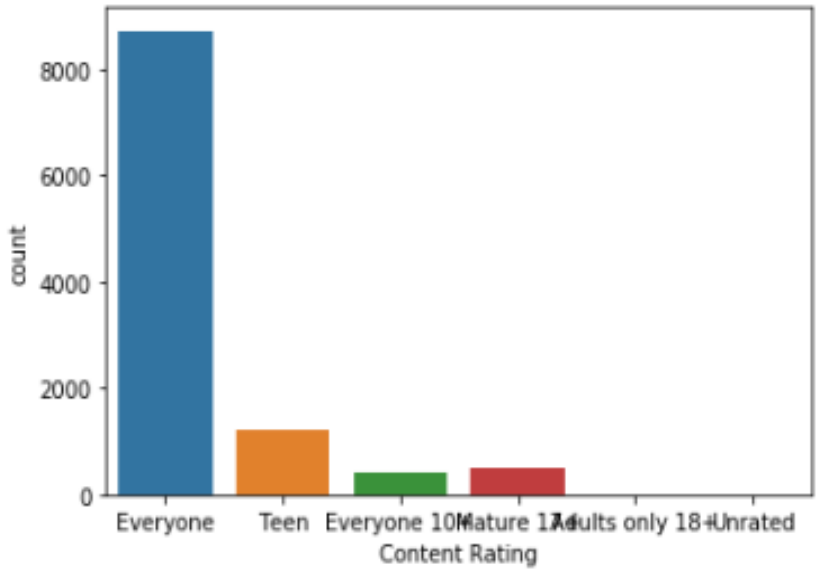
Visualizing Category and Rating:

For categories like Beauty, Comics, Events, House and Home and Libraries and Demo (the free versions of these apps) only these have ratings.



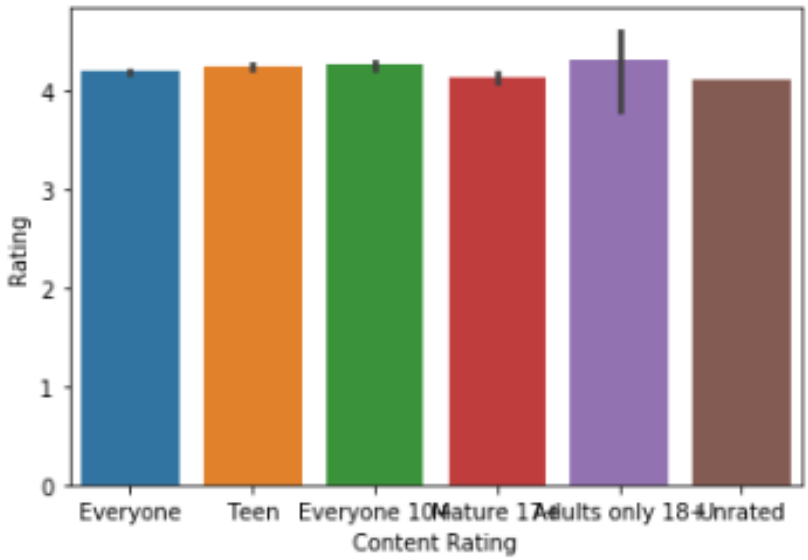
Content Rating

Visualizing Content Rating and Count

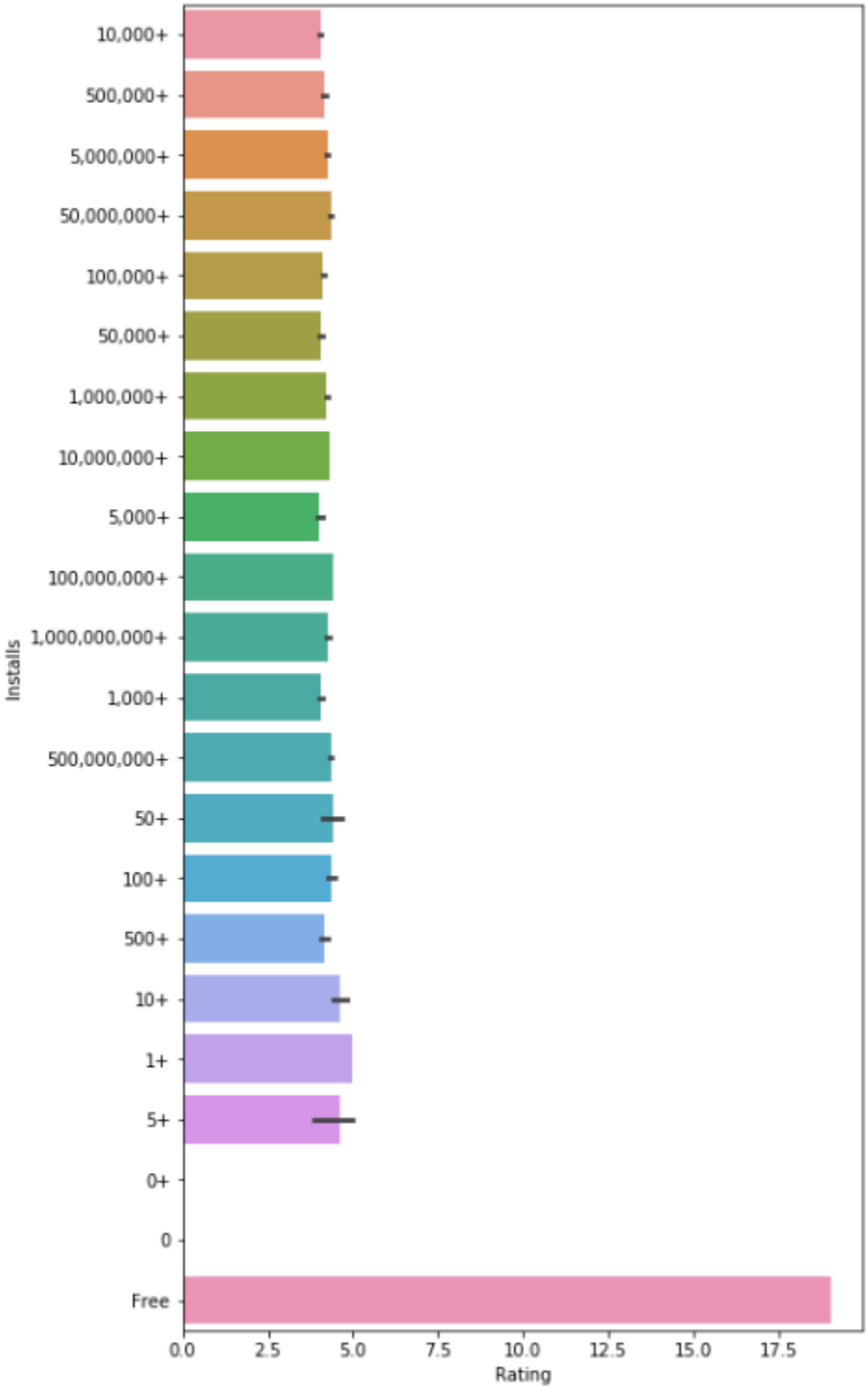


Visualizing Rating and Content Rating:

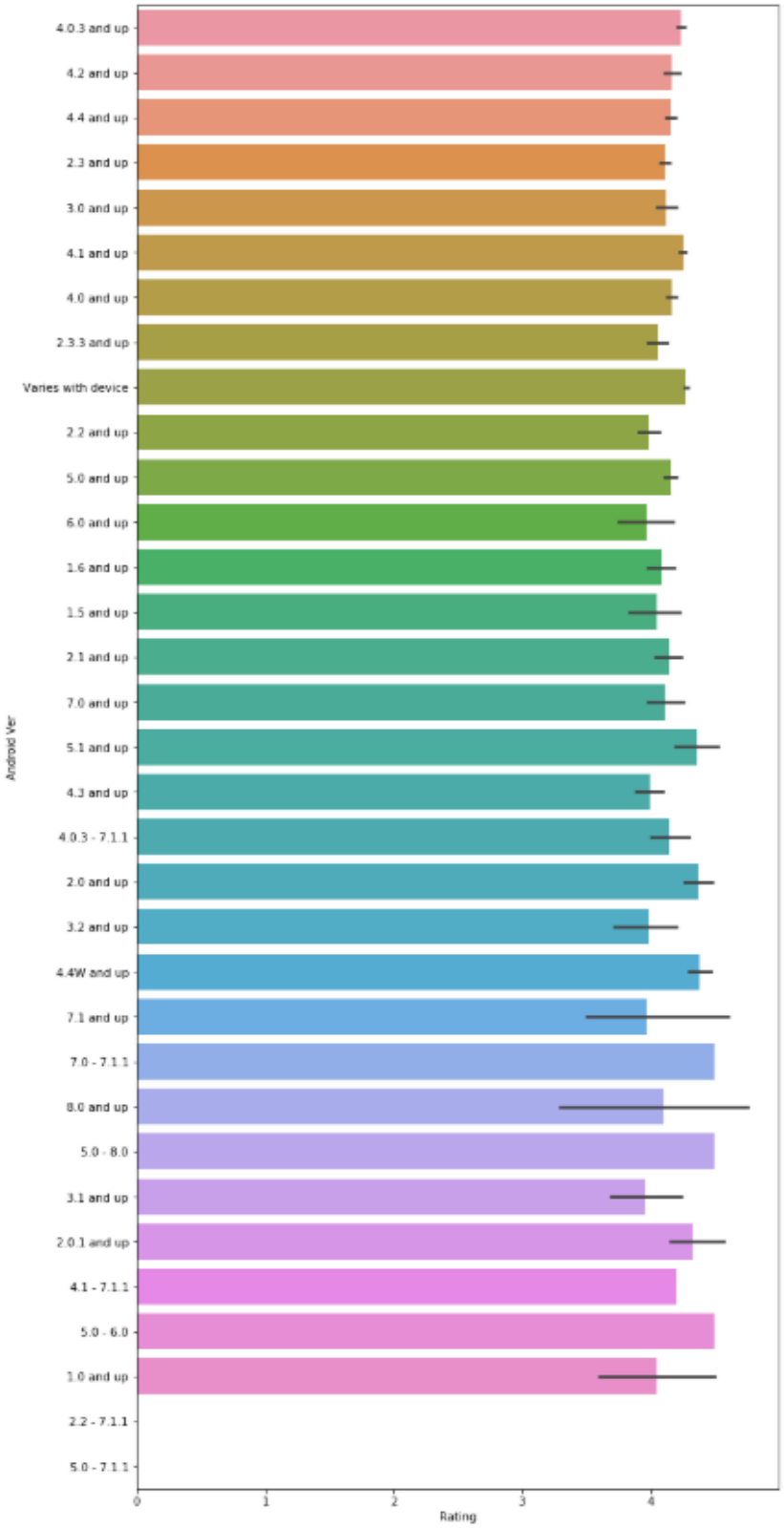
Adults 18+ group has slightly more count



Installs


10,000+
500,000+
5,000,000+
50,000,000+
100,000+
50,000+
1,000,000+
10,000,000+
5,000+
100,000,000+
1,000,000,000+
1,000+
500,000,000+
50+
100+
500+
10+
1+
5+
0+
0
Free
Installs
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
Rating

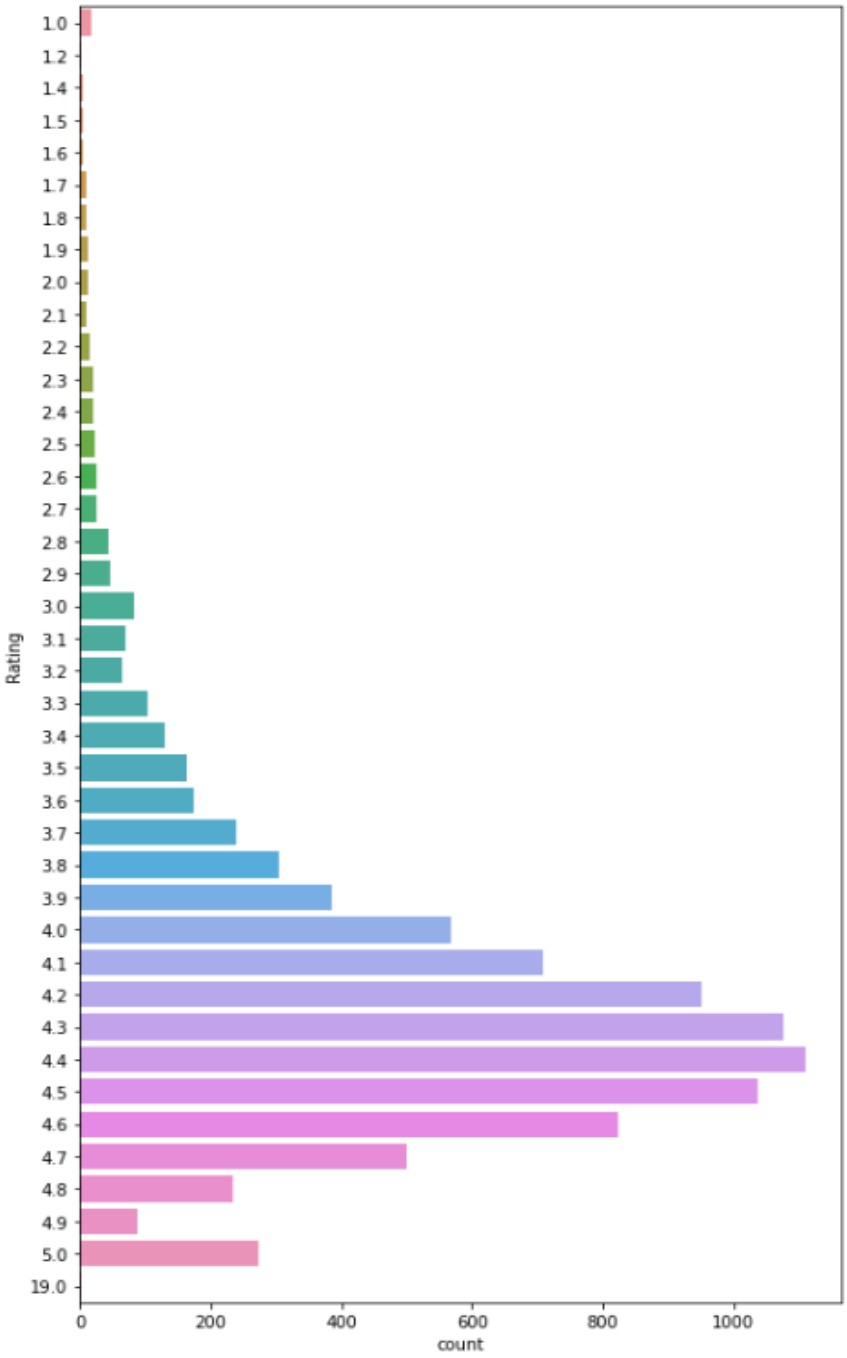
Android Version



Rating

Visualizing Count plot on Rating:

Here we see that most of the apps have received 4.4 rating



Machine Learning Techniques

We divided our dataset into 80% and 20% training and testing data respectively. We have used two models for our project.

1. Logistic Regression:

Logistic regression is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). Like all regression analyses, the logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.

```
#LogisticRegression :
lr_c=LogisticRegression(random_state=0)
lr_c.fit(X_train,y_train)
lr_pred=lr_c.predict(X_test)
lr_cm=confusion_matrix(y_test,lr_pred)
lr_ac=accuracy_score(y_test, lr_pred)
print('LogisticRegression_accuracy:',lr_ac)
```

```
LogisticRegression_accuracy: 0.7644230769230769
```

Logistic Regression gives 76% accuracy for our dataset

2. Decision Tree:

Decision tree builds regression or classification models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. A decision node (e.g., Outlook) has two or more branches (e.g., Sunny, Overcast and Rainy), each representing values for the attribute tested.

Leaf node (e.g., Hours Played) represents a decision on the numerical target. The topmost decision node in a tree which corresponds to the best predictor called root node.

```
In [106]:
#Decision Tree
from sklearn.tree import DecisionTreeClassifier
clf = DecisionTreeClassifier(random_state=0)
clf.fit(X_train,y_train)
y_pred3= clf.predict(X_test)
```

```
In [107]:
accuracy_score(y_test, y_pred3)
```

```
Out[107]: 0.6887464387464387
```

Decision Tree model gave us approximately 69% accuracy.

Conclusion

Comparing the accuracy of both the models, Logistic Regression gives more accuracy. Hence Logistic Regression is the best fit for our dataset out of the two techniques that we have used.

Future Scope

We can apply other regression models on our dataset to find out the best fitting and most accurate machine learning model for our dataset.

As we know that the play store was taken over by Google Play Services. In recent years Google play has also launched Google play movies, audio, books to compete with other services. So, we can also collect those data and we can analyze the performance of Google Play movies, audio, and book compared to other existing services by applying machine learning models to this newly found/added data.

References:

https://en.wikipedia.org

www.kaggle.com