EDA On Play Store App Reviews

by
Shaik Ahmad Basha
Data science trainee,
AlmaBetter, Bangalore.

Abstract:

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.

Our experiment can help to understand and discover the key factors like what category of apps has high installs, what is category of app that generates highest revenue etc. which are responsible for app engagement and success.

1.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. We have two data sets. One is play store data which contains each app (row). And the other dataset is user reviews.

Our main objective is to explore and analyze these datasets to discover key factors responsible for app engagement and success.

2.Data Description

The Play Store App dataset has 10841 rows and 13 columns. Those 13columns are:

1. **App** : Name of the App
2. **Category** : Category of the app
3. **Rating** : Rating given for the app
4. **Reviews** : Total number of reviews for the app
5. **Size** : Size of the app (MB)
6. **Installs** : How many installs done for the app
7. **Type** : Type of the app
8. **Price** : Price of the app
9. **Content Rating** : Content rating
10. **Genres** : Genre of the app
11. **Last Updated** : Date of latest update
12. **Current Ver** : Current version of the app
13. **Android Ver** : Android version of the app

The User Reviews dataset has 37427 rows and 5 columns. Those 5 columns are:

1. **App** : Name of the data
2. **Translated Review** : Text of the review
3. **Sentiment** : The review is good or not
4. **Sentiment Polarity** : The polarity of sentiment measure how negative or positive the context is.

5. **Sentiment Subjectivity** : Sentiment subjectivity

3. Introduction:

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.

I have two datasets. One is play store app dataset which consists several details of the app. Another dataset is User Review dataset which contains the context of the review and also whether the review is positive or not.

So Exploratory data analysis on these datasets will give some understanding and key factors that responsible for app engagement and success.

4. Data Understanding

After the loading and collecting the dataset, understanding the data is very important. I Understand the various features of the two dataset and their meaning.

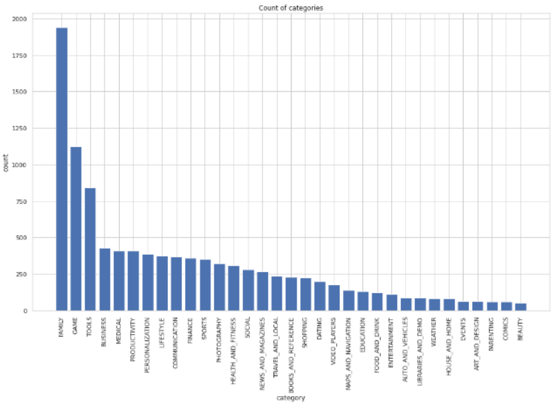
5. Data Cleaning

Our dataset contains null values. The Play Store App dataset has 5 features with null values and User reviews dataset has 4 features with null values. Both datasets has duplicated values. So I removed duplicated values in both datasets. In handling null values, 'Rating' column of play store app datset has 1474 null values. So I replace those null values with median. And I removed null values of other features. And I also created some columns like Revenue(price*installs), Updated_year(extracts the year from latest updated date) and score(based on positive reviews).

6. Exploratory Data Analysis

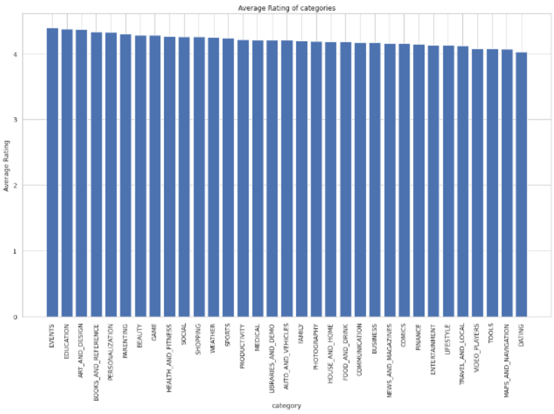
After removing the duplicated values and handling with null values in both of the datasets and creating new features based upon existing features, the datasets are ready for Exploratory Data Analysis. I started the analysis by checking the multicollinearity in the data and observed the relation between numerical features and dependent feature (price). And then I started to explore categorical features and how these categorical features vary with numerical features and drawn some conclusions from it. And I visualize those observations using bar chart, scatter plot, violin plot, pie chart, multiple bar chart. Following are some analyses which I have done.

1. **Which category has more apps?**
   There are total of 33 Categories.



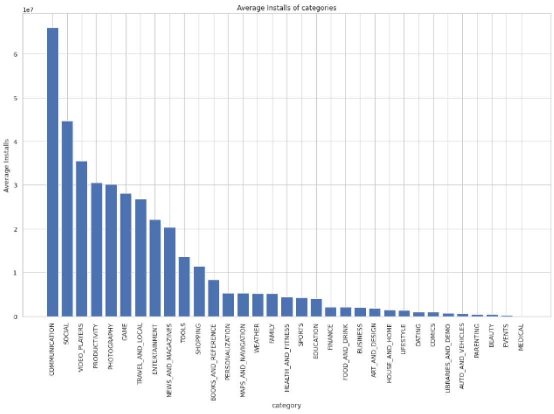
Most of the apps are from FAMILY category.

2. What Category apps have higher Ratings?



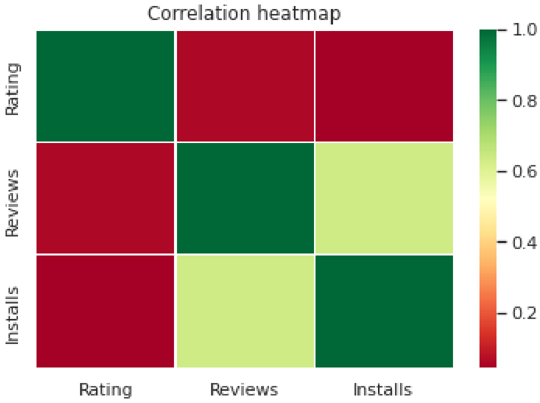
Here, EVENTS Category has higher average ratings.

3. What category apps have high installs?



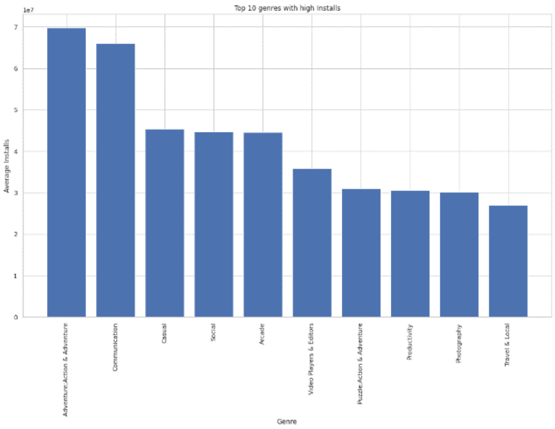
Here we can observe that COMMUNICATION category has higher installs

4. Is any multicollinearity exist in the features?



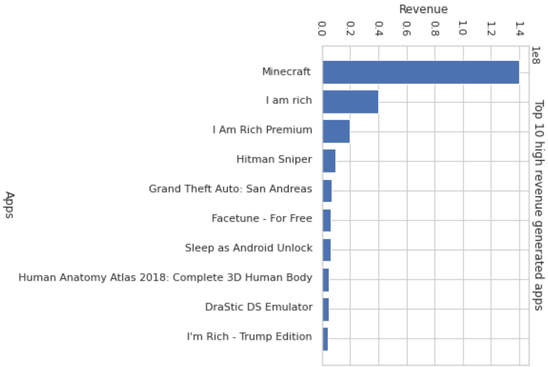
From the above heatmap we can see there is no multicollinearity in our features

5. What are the top 10 genres with most number of apps?



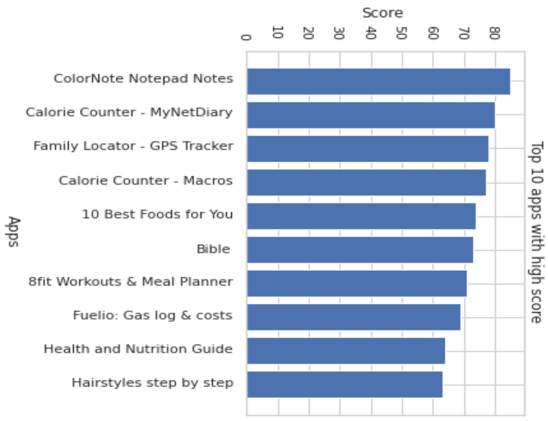
From the above bar chart, the names on X-axis are the areas with most number of listings.

6. Which paid app has highest revenue?



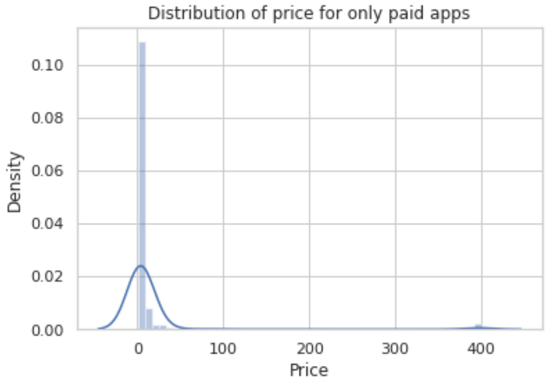
From the above visualization, we can observe that Minecraft has highest Revenue.

7. Which app has more positive reviews?



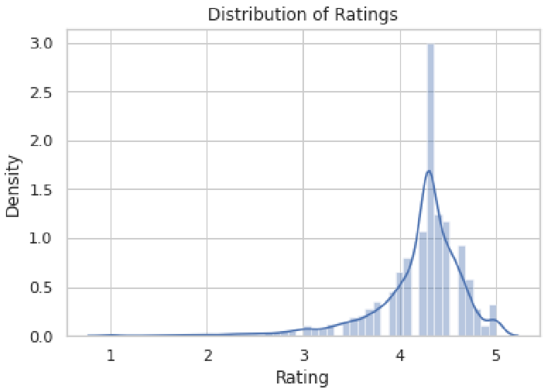
From the above visualization, we can observe that ColorNote Notepad Notes App has more positive reviews and is belongs to PRODUCTIVITY category.

8. How can you visualize the distribution of Price?



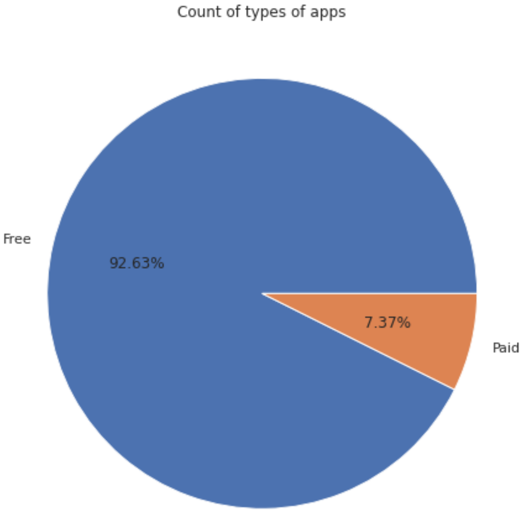
Here, we can observe that most prices are in the range of 0 to 50 USD, and a very few prices are in range of 50 to 450 USD

9. How can you visualize the distribution of Price?

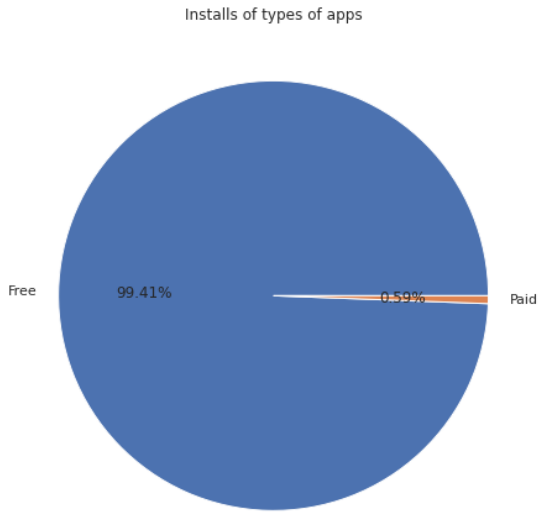


From the above visualizations, we can say that most number of ratings are in range of 4 to 5.

10. How many free apps and paid apps are there? Determine the percentage of installs done for free and paid apps?

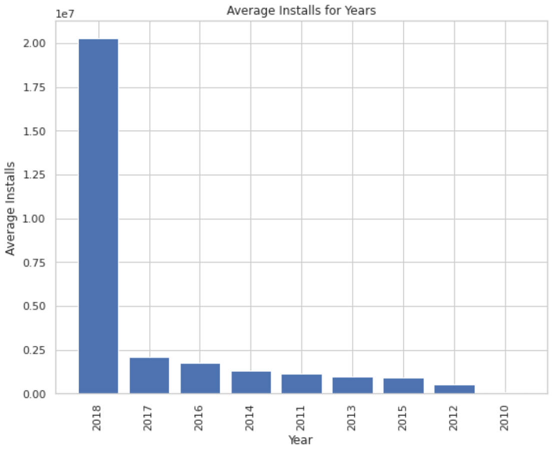


Most of the apps are free (92.63%) and only 7.37% apps are paid apps.



Free apps have high installs with 99.41%. and paid apps have very few installs with 0.59%.

11. Which year updated apps have more installs?



From the above chart, we can say that 2018 updated apps have more installs.