ADV EDA + Regression

Exploratory Data Analysis and building a linear regression model to predict app ratings on the Google play store

Hamza Afzal | January 2023

Introduction:

This report presents exploratory data analysis and a linear regression model to predict app ratings on Google Play Store.

Case study:

Dataset Overview:

The dataset has two tables:

1. googleplaystore (10841 \* 13 )

* App: Application name
* Category: Application is of which category
* Rating: Rating for each application
* Reviews: Review for each application
* Size: Size of each application
* Installs: People who have installed app
* Type : Free or paid
* Price : Price if paid
* Content Rating
* Genres:
* Last Updated
* Current Ver
* Android Ver

1. googleplaystore\_user\_reviews (64295 \* 5)

* App
* Translated\_Review
* Sentiment
* Sentiment\_Polarity
* Sentiment\_Subjectivity

Dataset Cleaning:

For googleplaystore table:

* There were 483 duplicate rows in googleplaystore table (which were dropped)
* There were 1465 nan values(missing values) which were dropped from googleplaystore table
* The data type of ‘Reviews’ was read as string, although the data was of numeric nature, converting to int64
* There is a column name ‘Size’ which has 1468 entries ‘Varies with device’, which were dropped , and also there are size in mb’s and kb’s. In order to make it consistent I am converting it into byte size and have renamed the column name accordingly.
* There is installs column, whose entries have + sign in it, removing that ‘+’ sign and ‘,’ for further analysis. Also converting the data type from object to int64.
* There is a price column and its entries have dollar sign in it, removing the dollar sign, and replacing the column name with price\_in\_$
* There is Last updated column, which is in object datatype, it should be in date format
* In genres there were some entries which have sub categories after ; I took the first string from left

For googleplaystore\_user\_reviews table:

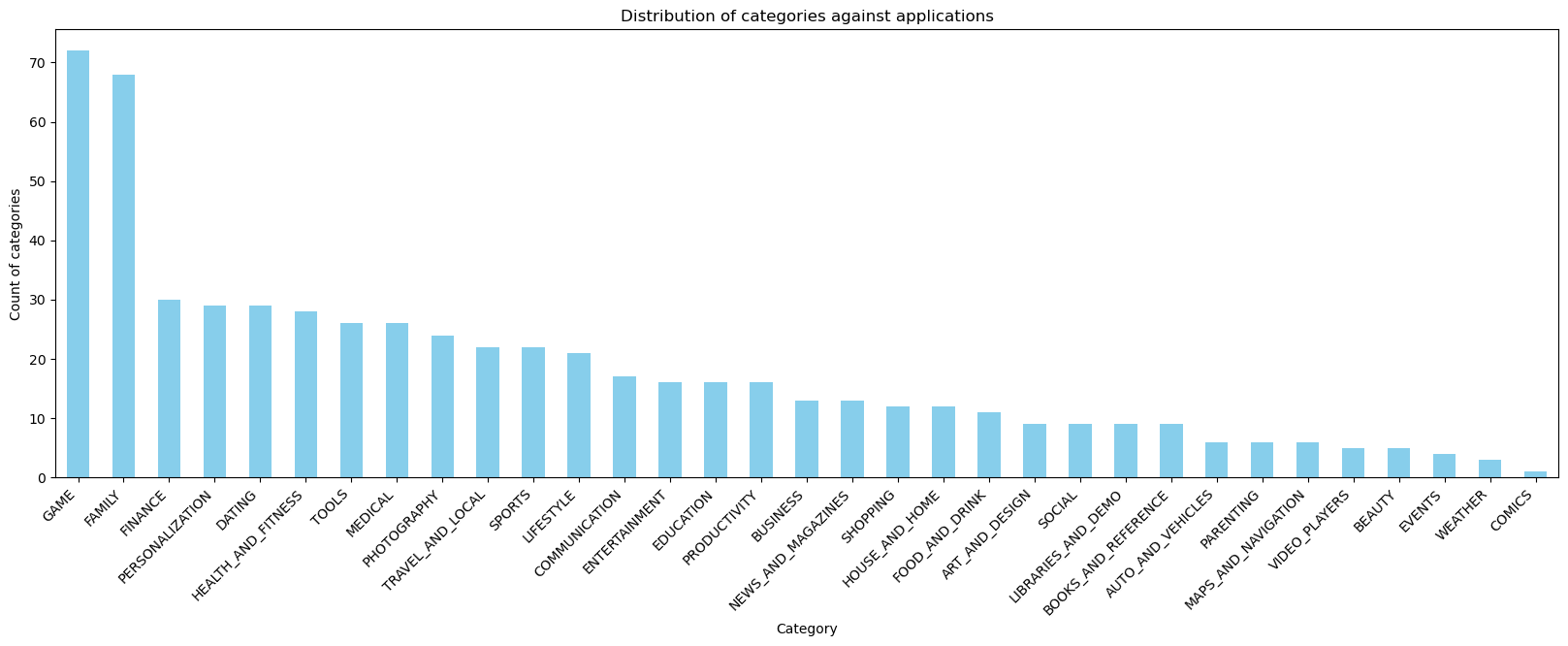
* Out of 64295 entries, there were 26863 null values in all the columns except ‘App’ column

So, there is no review available, so I am dropping the null values.

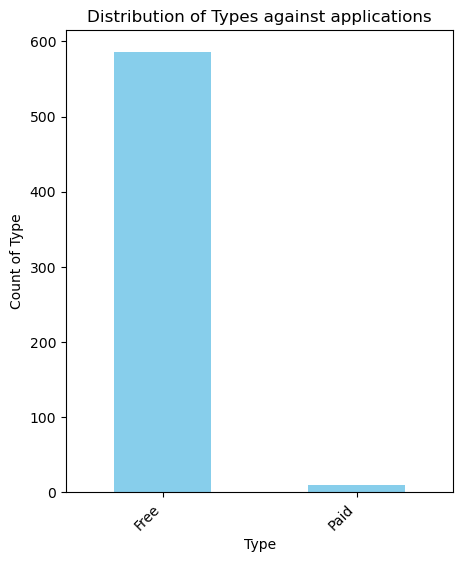
* There were 7735 duplicated rows which were removed from the dataset.
* Now, there are 29692 \* 5 values
* I have combined the two tables, on ‘app’ column using left join. And have removed the missing (NAN) values.
* Our final table has 26485 \* 17 columns
* I have label encoded my 3 categorical columns

Analysis and visual aids:

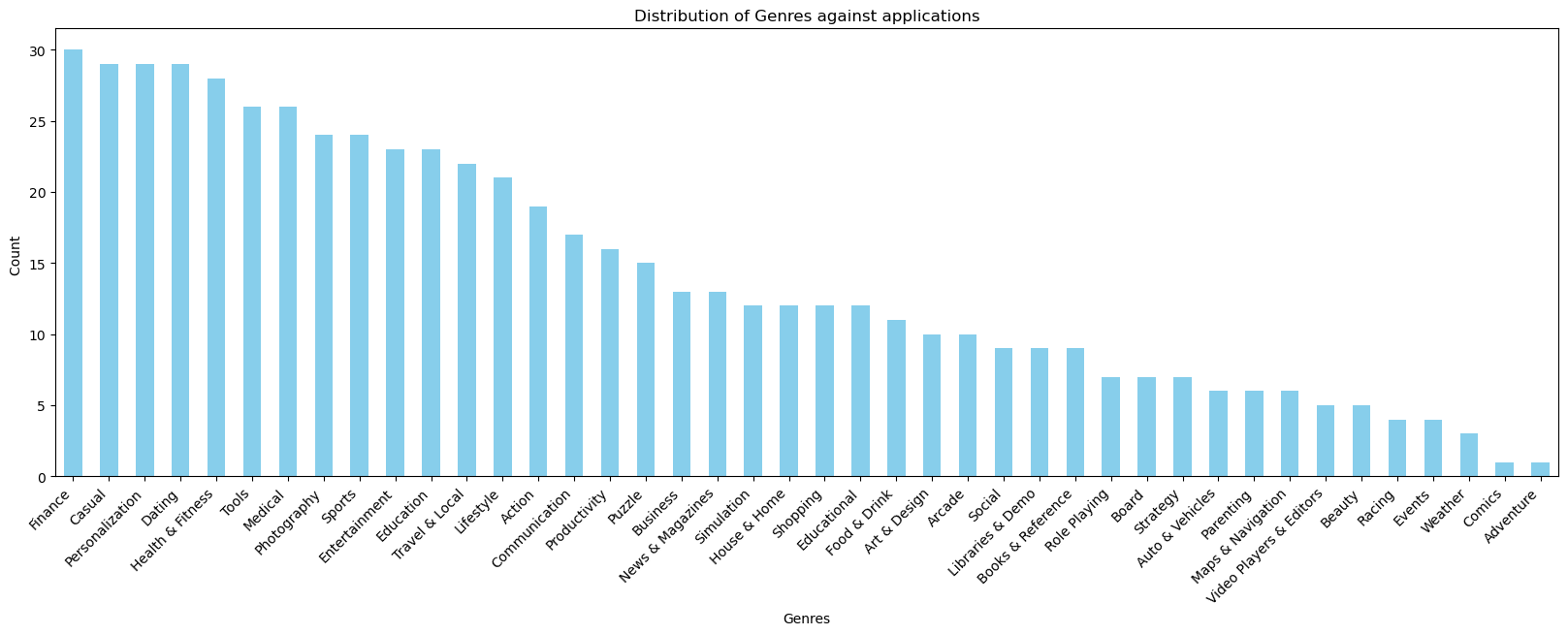
Game and family category had the most number of applications.



* Most of our applications were free



More number of applications are lying in Finance Genre



Top 5 Applications with best ratings

