Capstone Project

Play Store App Review Analysis

(Exploratory Data Analysis)

By: Ankit Singhal
Data Science Trainee
AlmaBetter

CONTENTS

WHAT IS EXPLORATORY DATA ANALYSIS?

Exploratory Data Analysis (EDA) is a crucial step in the data analysis process. It involves examining and visualizing the dataset to gain insights, discover patterns, and identify relationships between variables. Through EDA, we can uncover key characteristics of the data, such as its distribution, central tendencies, and variations. Here are the general steps involved in Exploratory Data Analysis (EDA):

Data Collection: Gather the relevant dataset for analysis, which may come from various sources such as databases, files, or APIs.

Data Cleaning: Preprocess the data to handle missing values, outliers, and inconsistencies. This step involves techniques like data imputation, outlier detection, and handling categorical variables.

Data Exploration: Perform initial exploration by examining the dataset's structure, dimensions, and summary statistics. This includes checking the data types, understanding variable distributions, and identifying potential issues.

Visualization: Utilize various visualizations such as bar charts, line plots, scatter plots, or heat maps to effectively communicate patterns and relationships in the data.

PROBLEM STATEMENT

Analyzing App Performance and User Sentiment in the Google Play Store

Objective: The objective of this project is to conduct exploratory data analysis (EDA) on two datasets related to the Google Play Store. The analysis aims to gain insights into app performance, user sentiment, and the relationship between app characteristics and user reviews.

Datasets:

Dataset 1: Google Play Store Apps Details

Description: This dataset contains details of various applications available on the Google Play Store. It includes features such as app name, category, rating, reviews, size, installs, price, etc.

Dataset 2: User Reviews

Description: This dataset contains user reviews for different apps on the Google Play Store. It includes information such as the text of the review, sentiment (positive, negative, or neutral), sentiment polarity, and sentiment subjectivity.

STRUCTURE OF DATASETS PROVIDED

We have been provided two datasets. We'll look for insights in the data to devise strategies to drive growth and retention. Let's take a look at the data:

Play store Data.csv: This file contains all the details of the apps on Google Play. There are 13 features that describe a given app.

- **App:** Name of the app
- **Category:** Category of the app. Some examples are: ART_AND_DESIGN, FINANCE, COMICS, BEAUTY etc.
- **Rating:** The current average rating (out of 5) of the app on Google Play
- **Reviews:** Number of user reviews given on the app
- **Size:** Size of the app in MB (megabytes)
- **Installs:** Number of times the app was downloaded from Google Play
- **Type:** Whether the app is paid or free
- **Price:** Price of the app in US$
- **Content Rating:** A content rating (also known as maturity rating) rates the suitability of TV broadcasts, movies, comic books, or video games to its audience. To show which age group is suitable to view media and entertainment.

- **Genres:** A category of artistic, musical, or literary composition characterized by a particular style, form, or content
- **Last Updated**: Date on which the app was last updated on Google Play
- **Current Ver**: Current Version means a version of the software that is currently being supported by its publisher.
- **Android Ver:** Android versions (codenames) are used to describe the various updates for the open source Android mobile operating system.

datasets/user_reviews.csv: This file contains a random sample of 100 user reviews for each app. The distribution of positive and negative reviews in each category has been pre-processed and passed through a sentiment analyzer.

- **App:** Name of the app on which the user review was provided. Matches the App column of the play_store_data.csv file
- **Translated Review**: The pre-processed user review text.
- **Sentiment:** Sentiment category of the user review - Positive, Negative or Neutral.
- **Sentiment Polarity:** Sentiment score of the user review. It lies between [-1,1]. A higher score denotes a more positive sentiment.

DATA WRANGLING

Duplicate Values

```
[165] # Dataset Duplicate Value Count
      duplicate_count = playstore_df.duplicated().sum()

      print("Number of duplicate rows in the dataset:", duplicate_count)
```

```
Number of duplicate rows in the dataset: 483
```

Missing Values/Null Values

```
[166] # Missing Values/Null Values Count
      missing_values_count = playstore_df.isnull().sum()

      print("Number of missing values in each column:")
      print(missing_values_count)
```

```
Number of missing values in each column:
App                0
Category           0
Rating          1474
Reviews            0
Size               0
Installs           0
Type               1
Price              0
Content Rating     1
Genres             0
Last Updated       0
Current Ver        8
Android Ver        3
dtype: int64
```

Play Store Apps Data Frame Duplicate Rows and Null Values count:

483 Duplicate Rows in the dataset

1487 Null values in the dataset

DATA WRANGLING

Handling Missing Values

a). Android Ver: 3 NaN values in this column.
We cannot replace the NaN values with any particular values as it depends on specific app. Since there are only 3 rows which is around 0.03% of the total rows in the Android Ver column, so it would be better to drop these rows as it would not affect our data analysis.

b) Type: one NaN value in this column
Since, the type of the app can be Paid or Free, and the Price of the app is 0, so we will replace the Type as "Free", thereby handling the NaN value

DATA WRANGLING

Handling Missing Values

c) Current Ver: 8 NaN values in this column.
As Current Ver is app specific and cannot be changed to a particular value, decided to drop these rows as its only 0.07% of the total rows and won't affect our analysis

d) Rating: 1470 NaN values in this column
Rating cloumn having NaN values consists of 1470 rows i.e. around 13.5% of the total rows, we cannot drop all the rows as that would impact our analysis. We will check the mean and median of the non NaN value rows and will continue handing the NaN values

DATA WRANGLING

The choice between using the mean or the median depends on the distribution of the ratings.

Mean: The mean is the average value of the ratings and is sensitive to extreme values (outliers). If the ratings are normally distributed and there are no significant outliers, using the mean can provide a representative estimate of the missing values. It can be a suitable choice to maintain the overall average rating.

Median: The median represents the middle value of the ratings when they are sorted in ascending or descending order. Unlike the mean, the median is not affected by extreme values or outliers. If the ratings are skewed or have a few extreme values, using the median can provide a more robust estimate of the missing values. It can be a suitable choice to maintain the central tendency of the ratings.

We created Histogram and Box plot to understand the data if its normal or skewed and to check if there is any outliers in the data and concluded it would be best to replace NaN values with the median.

DATA WRANGLING

Handling the duplicates in the App column:

a). We noticed that there are multiple duplicate values in the App column
After addressing the duplicate values in the "App" column, we have effectively removed them. The number of rows remaining in the dataset, after dropping the duplicates in the "App" column, is 9649.

b). Changed the values in the Installs column from string datatype to integer datatype.

c). Changed the datatype of the Last Updated column from string to datetime.

d). Changed the datatype of the Price column from string type to float type.

e). Converted the datatype of values in the Reviews column from string to int

DATA WRANGLING

```
#Check null values in User Reviews Dataset

null_counts = reviews_df.isnull().sum()
total_rows = reviews_df.shape[0]
null_percentages = (null_counts / total_rows) * 100

null_values_df = pd.DataFrame({'Null Values': null_counts, 'Percentage': null_percentages})
print(null_values_df)
```

```
                        Null Values  Percentage
App                               0    0.000000
Translated_Review             26868   41.788631
Sentiment                     26863   41.780854
Sentiment_Polarity            26863   41.780854
Sentiment_Subjectivity        26863   41.780854
```

DATA WRANGLING

Insights :

The column "App" has 0 null values, which means there are no missing values in this column. The percentage of null values is 0%.

The column "Translated_Review" has 26868 null values, which accounts for approximately 41.79% of the total rows in the dataframe.

The column "Sentiment" also has 26863 null values, which is approximately 41.78% of the total rows.

Similarly, the columns "Sentiment_Polarity" and "Sentiment_Subjectivity" also have 26863 null values, accounting for approximately 41.78% of the total rows.

In summary, the columns "Translated_Review", "Sentiment", "Sentiment_Polarity", and "Sentiment_Subjectivity" have a significant number of null values, representing a considerable percentage of the data in those columns.

DATA WRANGLING

We need to check the NaN values in the corresponding columns now:

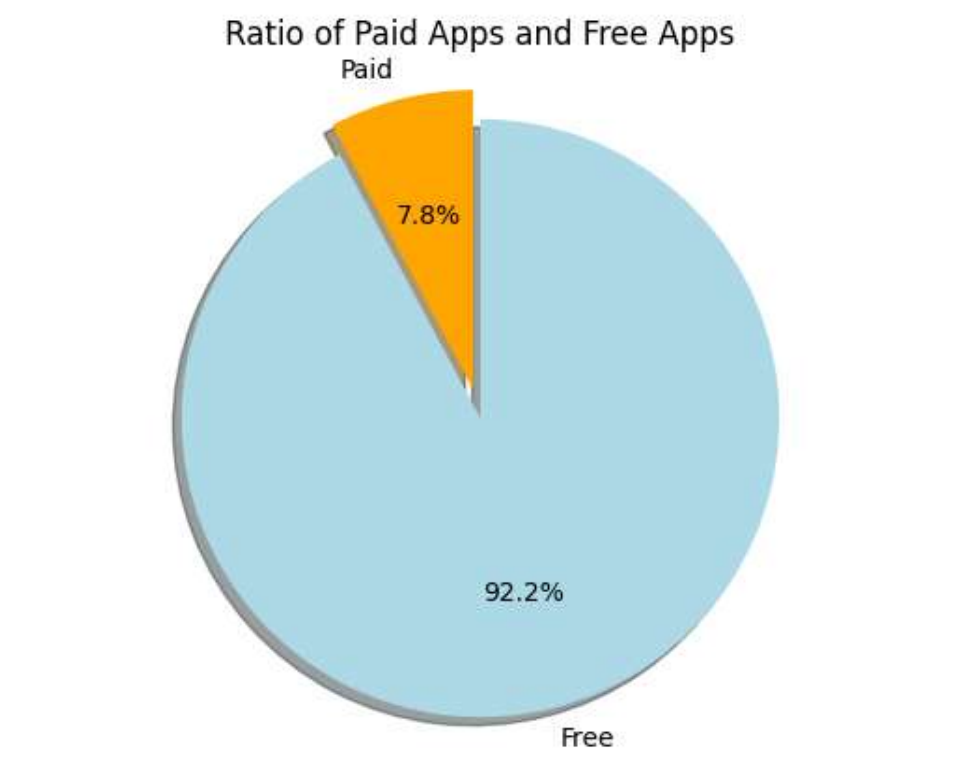
```
# checking the NaN values in the translated review column
reviews_df[reviews_df['Translated_Review'].isnull()]
```

	App	Translated_Review	Sentiment	Sentiment_Polarity	Sentiment_Subjectivity
2	10 Best Foods for You	NaN	NaN	NaN	NaN
7	10 Best Foods for You	NaN	NaN	NaN	NaN
15	10 Best Foods for You	NaN	NaN	NaN	NaN
102	10 Best Foods for You	NaN	NaN	NaN	NaN
107	10 Best Foods for You	NaN	NaN	NaN	NaN
...	...	...	...	...	...
64290	Houzz Interior Design Ideas	NaN	NaN	NaN	NaN
64291	Houzz Interior Design Ideas	NaN	NaN	NaN	NaN
64292	Houzz Interior Design Ideas	NaN	NaN	NaN	NaN
64293	Houzz Interior Design Ideas	NaN	NaN	NaN	NaN
64294	Houzz Interior Design Ideas	NaN	NaN	NaN	NaN

26868 rows × 5 columns

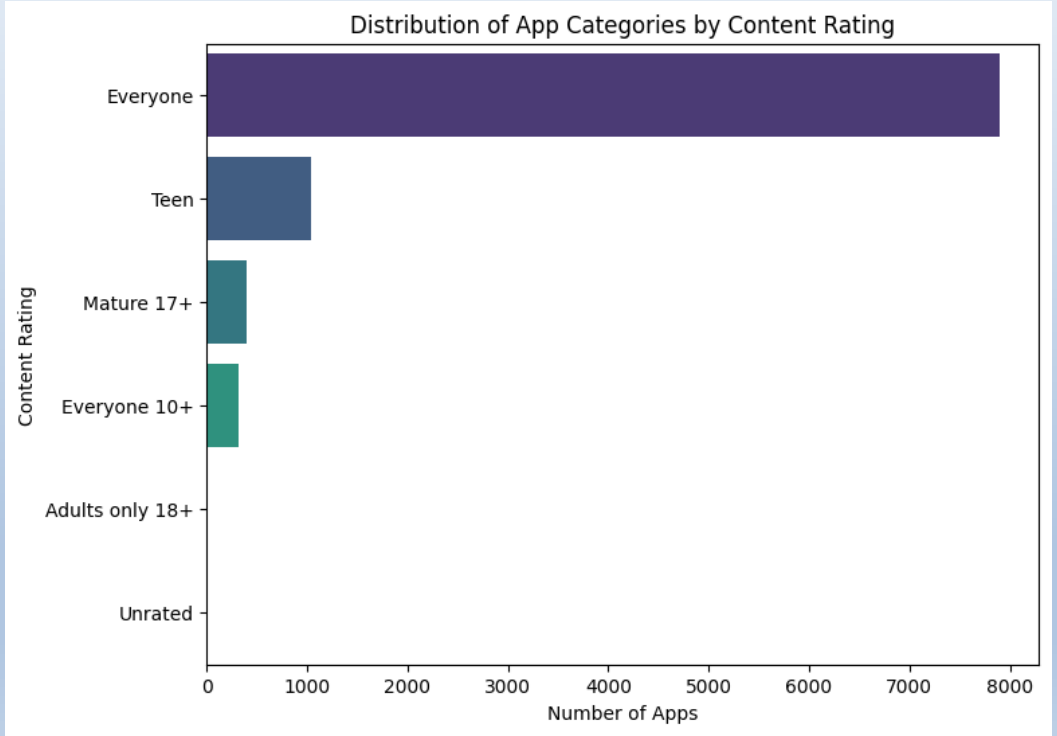
From the output, its clear that Apps which do not have review, also have null values in the Sentiment, Sentiment_Polarity and Sentiment_Subjectivity columns in most of the cases. From the output, since the Translated reviews are null, all other columns are non significant as they depend on user reviews only, so we have dropped these rows.

DATA VISUALIZATION



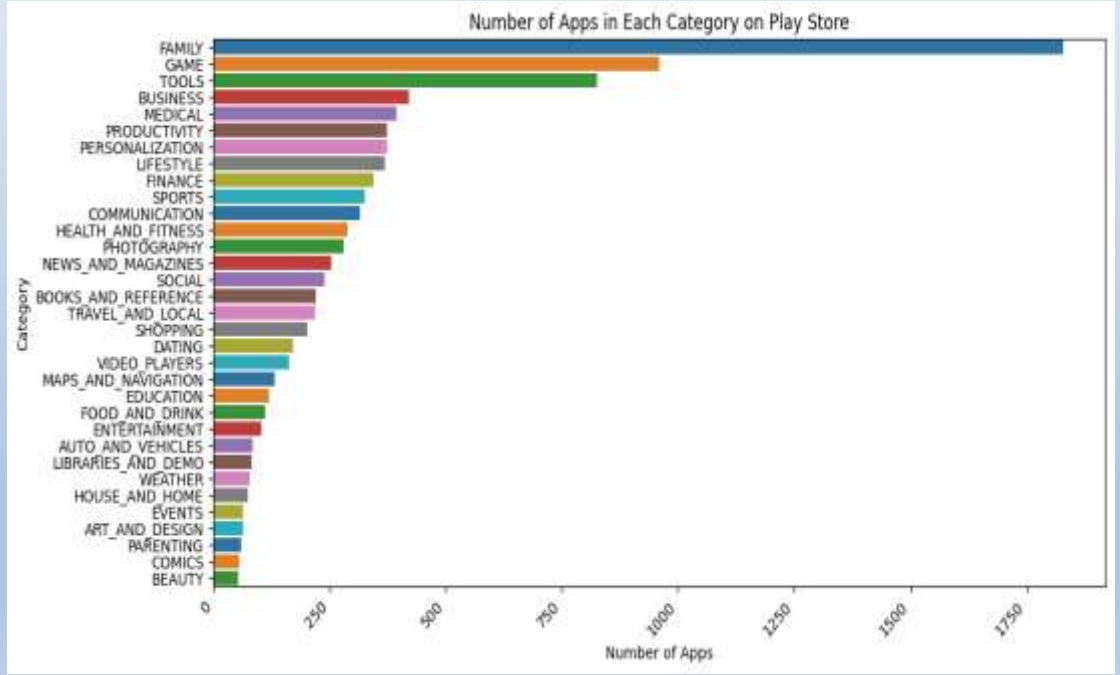
Majority, or approximately 92%, of the apps in the Google Play Store are free, while the remaining 8% are paid.

DATA VISUALIZATION



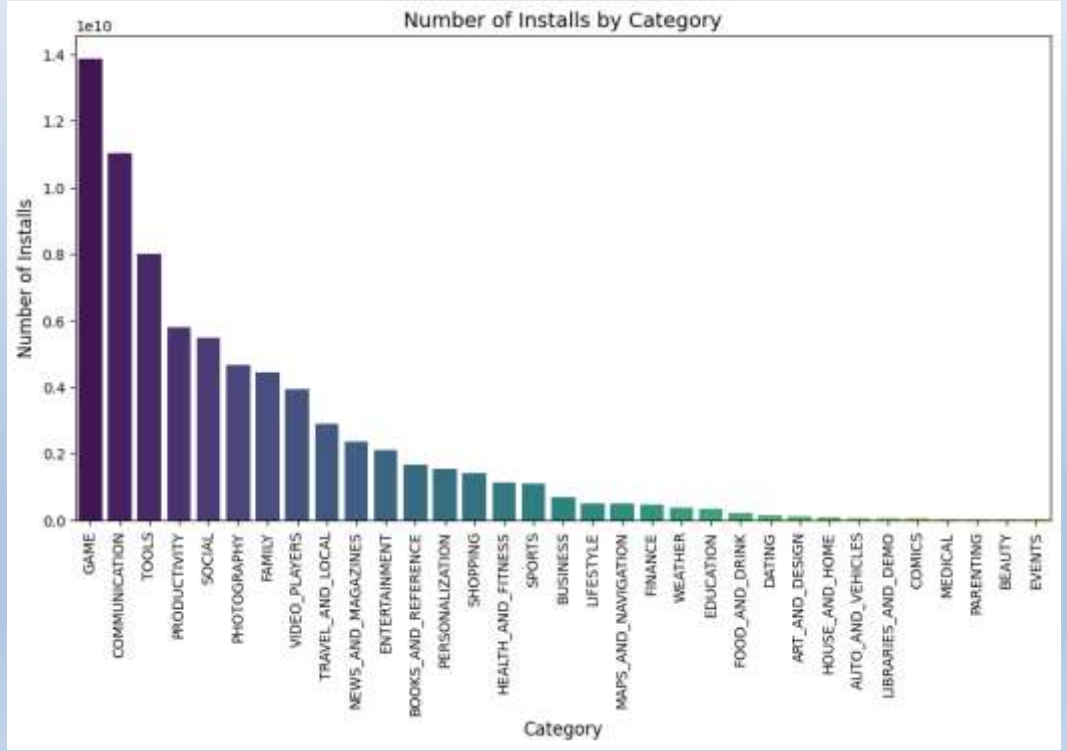
The majority of apps in the play store are classified as suitable for all age groups, indicating that they can be used by everyone. However, there are also apps with specific age restrictions, targeting different audiences based on their content and purpose.

DATA VISUALIZATION



In the Play Store dataset, we have a total of 33 categories. From the bar plot, we can see that the "FAMILY" and "GAME" categories have the highest number of apps, while the "EVENTS" and "BEAUTY" categories have the lowest number of apps.

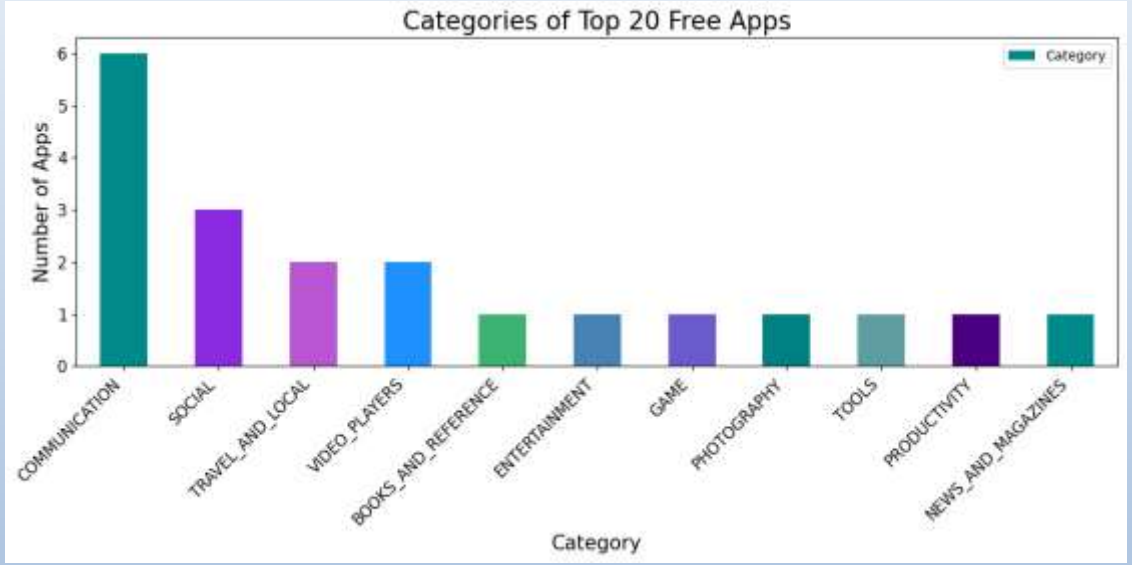
DATA VISUALIZATION



Games, Communication and Tools category has the highest number of installs compared to other categories

Beauty and Events have the lowest number of installs

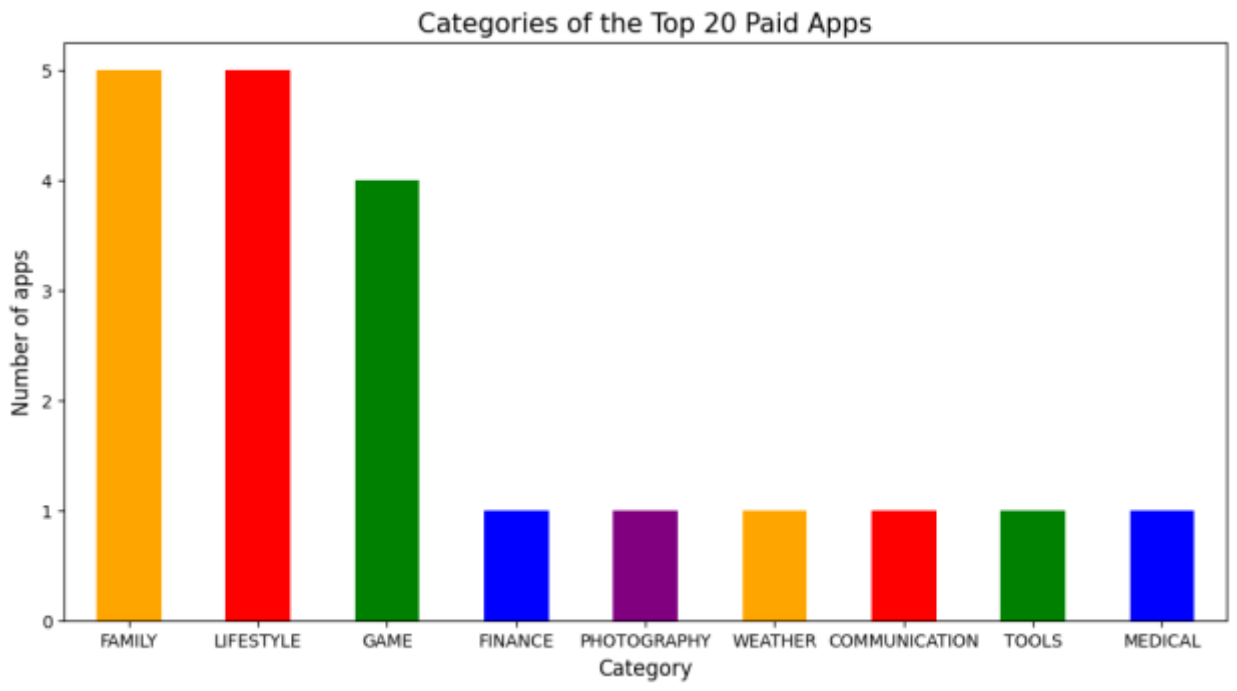
DATA VISUALIZATION



Communication and Social categories have highest number of top 20 free applications

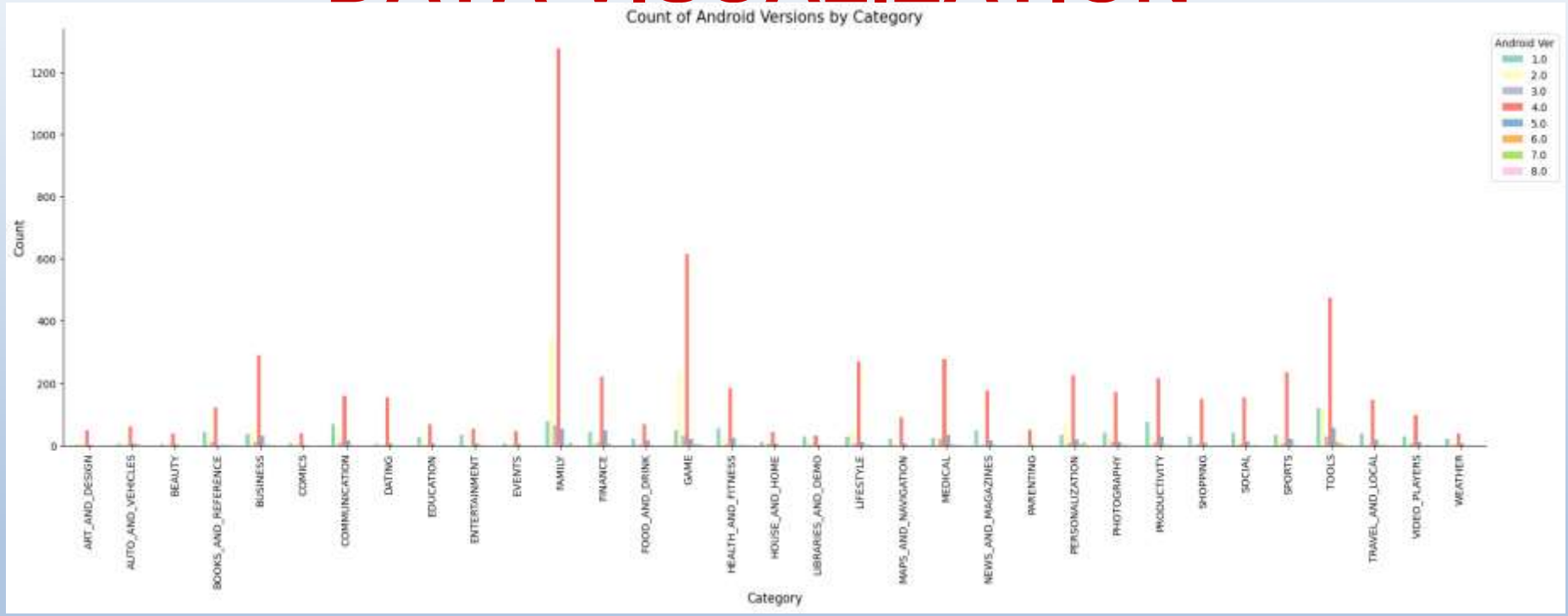
Productivity, News and Magazines have comparatively less apps in top 20

DATA VISUALIZATION



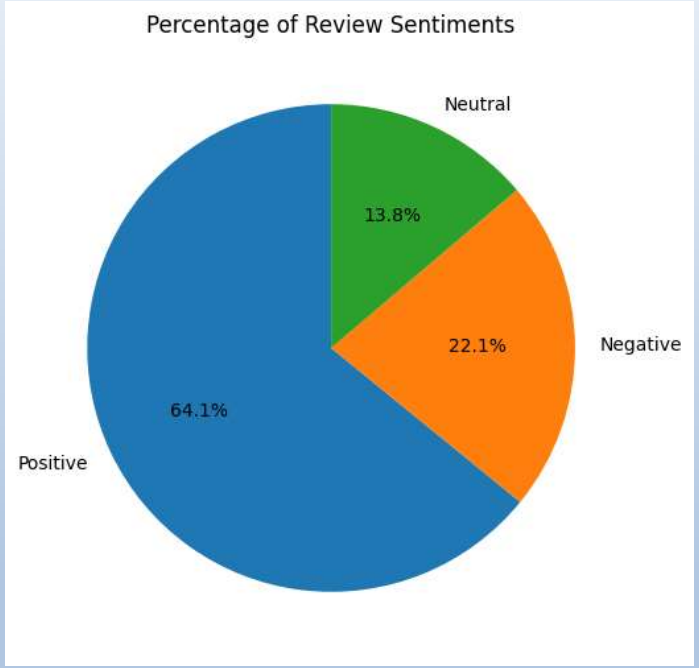
We created bar plot for top 20 paid apps based on revenue generated found that Family and Lifestyle categories have highest number of applications from the top 20 paid apps in the play store.

DATA VISUALIZATION



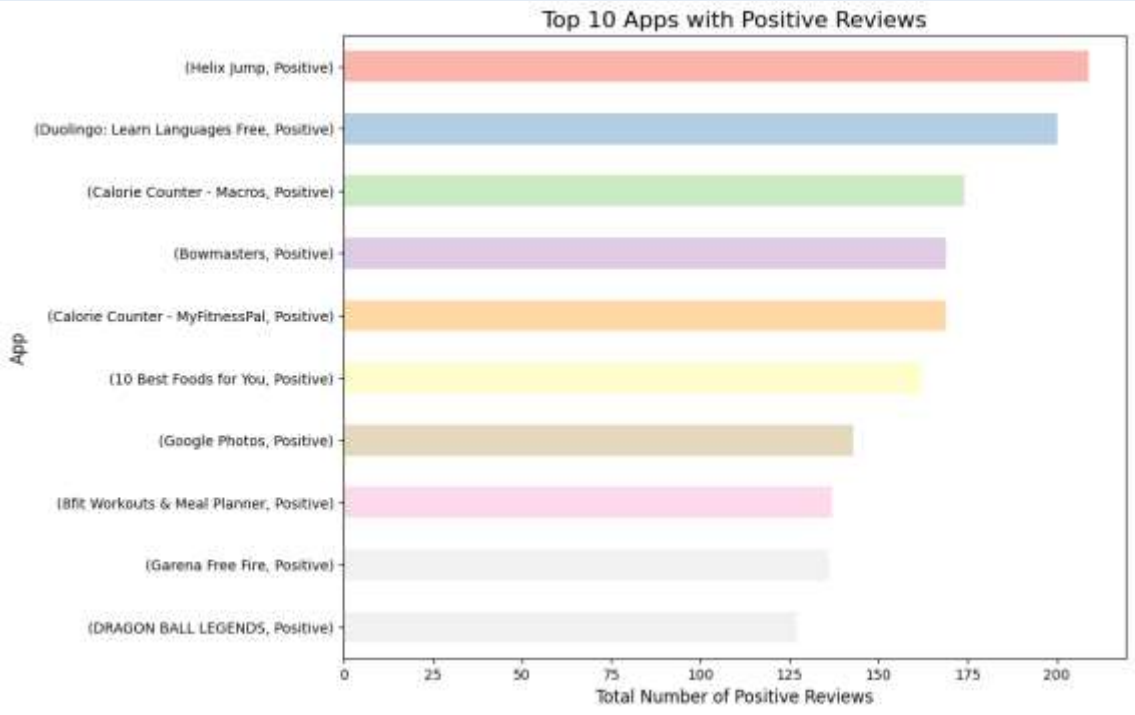
Most of the apps are working on Android Version 4.0 and up

DATA VISUALIZATION



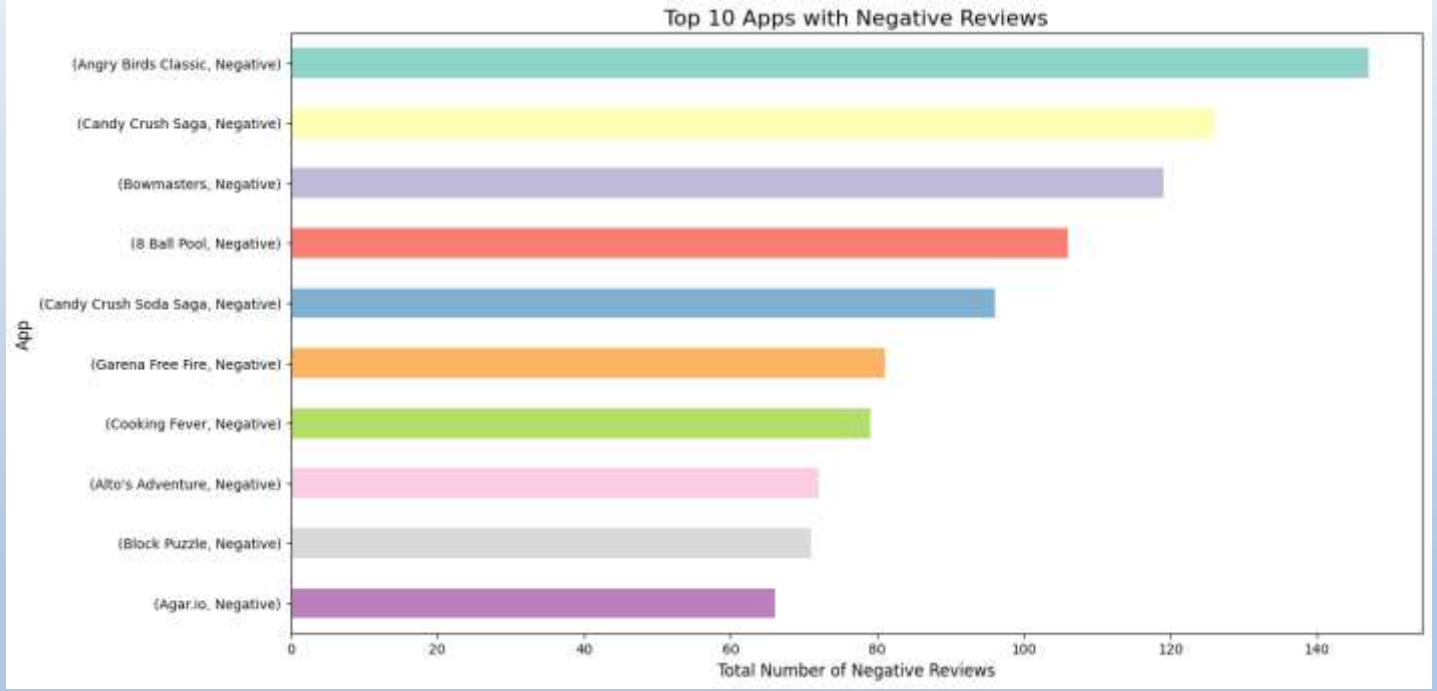
Positive reviews are 64.30%, Negative reviews are 22.80% and Neutral reviews are 12.90%

DATA VISUALIZATION



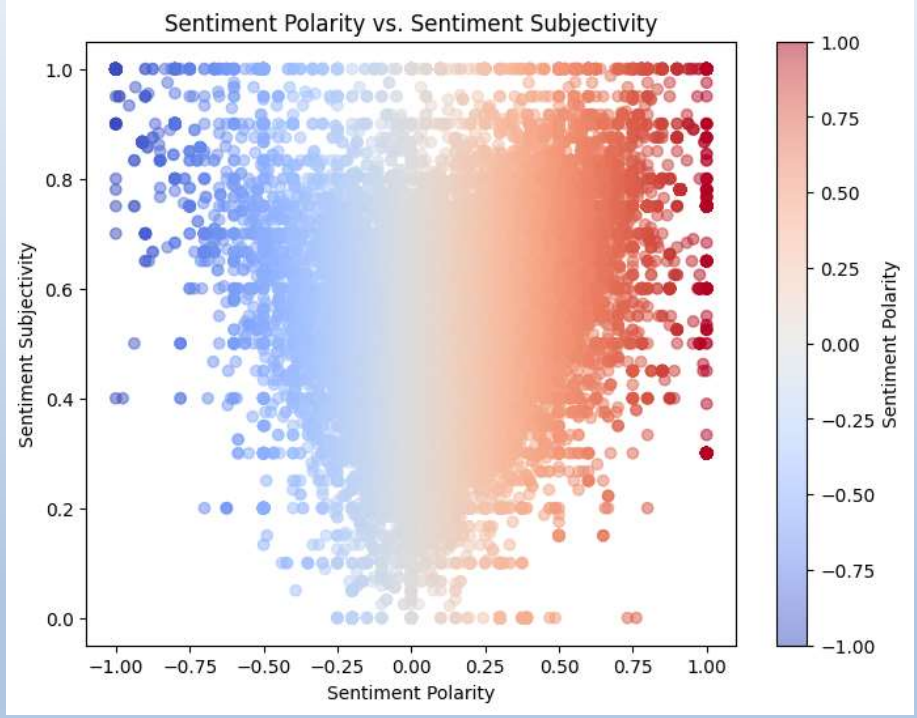
Helix Jump, Duolingo: Learn Languages Free and Calorie Counter are the top apps with positive reviews as per the bar chart

DATA VISUALIZATION



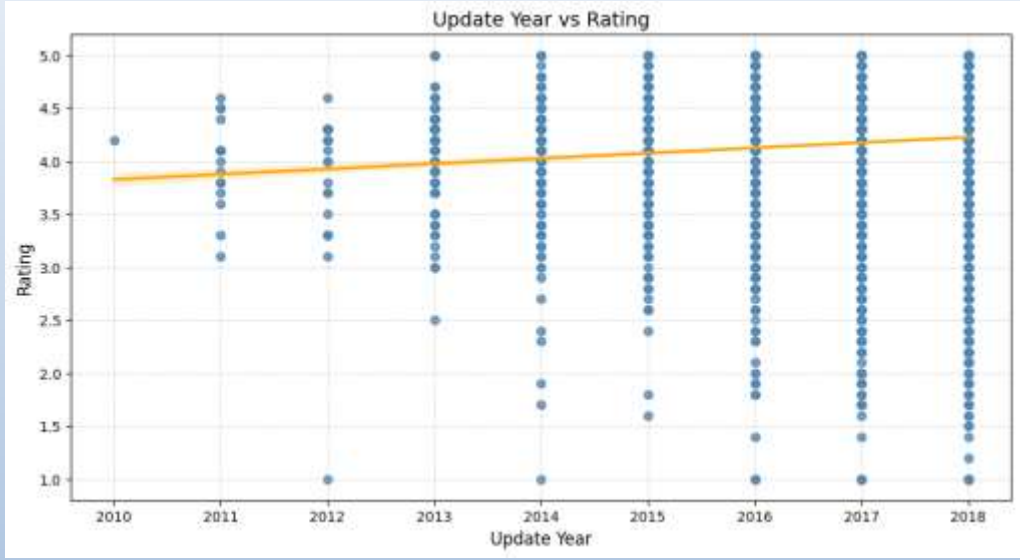
Angry Birds Classic, Candy Crush Saga and Bowmasters are the apps that have received the most number of negative reviews

DATA VISUALIZATION



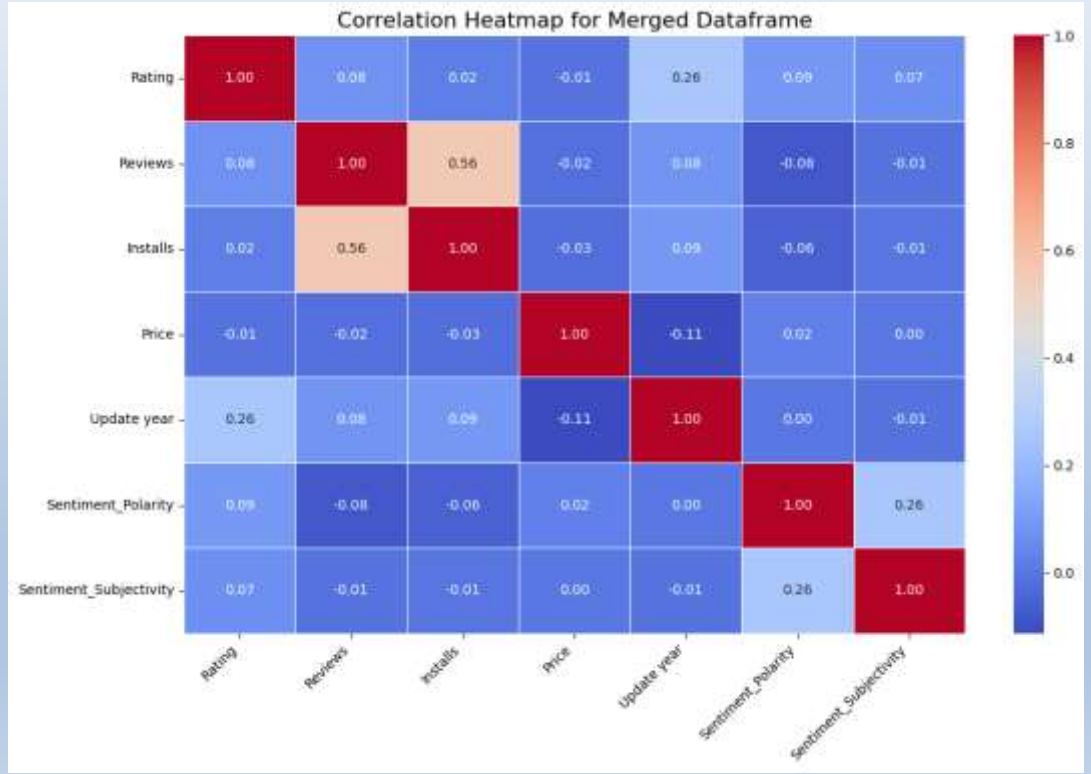
From the scatter plot it is evident that sentiment subjectivity is not always proportional to sentiment polarity but in most cases, shows a proportional behavior, when variance is too high or low

DATA VISUALIZATION



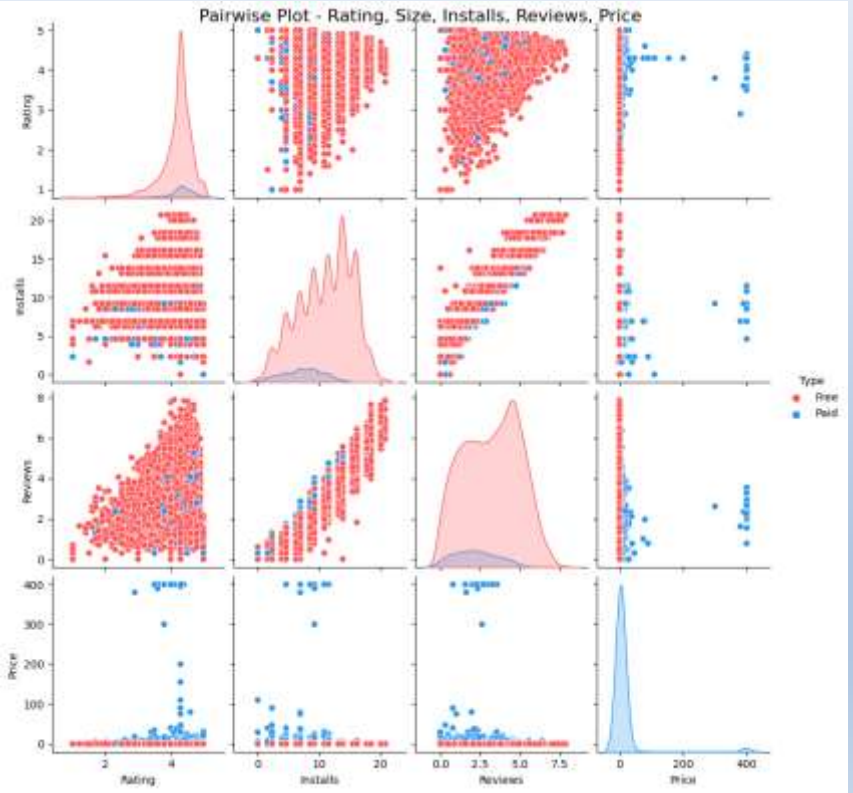
Scatter Plot with a Regression Line shows that there is a positive relationship between the "Update Year" and "Rating" variables, with higher ratings generally associated with more recent updates. The regression line provides a visual representation of this relationship, showing the general trend in the data and providing a way to estimate the expected rating for a given update year.

DATA VISUALIZATION



Correlation Heatmap of both the datasets shows that there is a slightly high positive correlation between "Rating" and "Reviews," it suggests that higher-rated apps tend to receive more reviews, indicating popularity and user engagement. Positive correlation between "Sentiment Polarity" and "Rating implies that apps with more positive sentiment in user reviews tend to have higher ratings.

DATA VISUALIZATION



We created Pair Plot and found out the following insights:

Most of the Apps are Free

Most of the Paid Apps have Rating around 4
This finding suggests a positive correlation between the number of app installations and the number of reviews.

Most of the Apps are lightweight: The observation that most apps are light-weighted implies that a significant portion of the apps in the dataset have a smaller file size.

CONCLUSIONS

1. Most apps in the Google Play Store are free, with only a small percentage being paid.
2. The majority of apps are suitable for all age groups, but there are also apps with specific age restrictions.
3. The "FAMILY" and "GAME" categories have the highest number of apps, while "EVENTS" and "BEAUTY" have the lowest.
4. Games, Communication, and Tools categories have the highest number of installs.
5. Subway Surfers, Candy Crush Saga, and Temple Run 2 are the most installed games.
6. Communication and Social categories dominate the top 20 free apps. Family and Lifestyle categories have the highest number of paid apps in the top 20.
7. Most apps support Android Version 4.0 and up.
8. Positive reviews outweigh negative and neutral reviews.
9. Sentiment subjectivity is not always proportional to sentiment polarity.
10. Higher ratings are associated with more recent updates.
11. Positive correlation between ratings and reviews, as well as sentiment polarity and ratings.
12. Most apps are free, paid apps tend to have higher ratings, and popular apps attract more reviews.