Exploratory Data Analysis (EDA)

Play Store Apps Review Analysis

MAHFUJ RAZA (Data Science Trainee)

ALMABETTER

Abstract

This project focuses on the **Exploratory Data Analysis (EDA)** of two datasets related to Google Play Store applications: **App Information Data** and **User Review Data**. The primary goal is to analyze key aspects of the Play Store ecosystem, such as app ratings, downloads, category distribution, and user sentiments. The project involves extensive data cleaning, visualization, and sentiment analysis to derive insights about app performance and user feedback. Univariate and bivariate analyses are used to highlight trends in app popularity, rating patterns, and review sentiments. The results provide a comprehensive view of the factors influencing app success and user satisfaction, offering actionable insights for app developers and businesses.

Problem Statement

With millions of apps available on the Google Play Store, users face a challenge in identifying high-quality, well-performing apps. Meanwhile, app developers and businesses struggle to understand the key factors that contribute to an app's success and user satisfaction. Ratings, reviews, and download counts are critical indicators, but raw data alone is insufficient for making strategic decisions.

Additionally, user reviews contain valuable feedback, but the sheer volume makes it difficult to extract actionable insights manually. Sentiment analysis offers an opportunity to gain a deeper understanding of user experience, yet many businesses overlook this data in decision-making.

The objective of this project is to:

1. Analyze and clean Google Play Store data to uncover insights about app ratings, downloads, and categories.
2. Perform sentiment analysis on user reviews to identify patterns in user satisfaction.
3. Provide visual insights that help developers and businesses make data-driven decisions regarding app development, marketing, and user engagement.

Introduction

- The Google Play Store is one of the largest app distribution platforms, hosting millions of apps across diverse categories such as gaming, education, health, and business. With the growing number of apps, understanding what drives user engagement and app success is crucial for developers and stakeholders. This project aims to analyze two datasets: one containing information about apps, including their category, rating, size, and download count, and the other containing user reviews with associated sentiment analysis results.

- The project begins with data cleaning to handle missing values, incorrect data types, and outliers. Following that, univariate and bivariate analyses are conducted to explore various aspects of app performance, such as identifying the most downloaded apps, highly rated categories, and app distribution by content rating. In addition, sentiment analysis of user reviews reveals patterns in user satisfaction and common issues faced by users.
- By leveraging Python libraries such as **Pandas**, **NumPy**, **Matplotlib**, and **Seaborn**, the analysis presents meaningful visualizations and insights into how apps perform and how user feedback correlates with app ratings and installs. This study aims to provide valuable guidance to developers, marketers, and businesses in making data-driven decisions for app development and user engagement strategies.
- **Play Store Apps Data**: Contains information about apps such as category, rating, size, number of installs, etc.
- **User Reviews Data**: Provides customer reviews of these apps, including the text of the reviews, ratings, and timestamps.

Exploring the database

We have provided with two databases

- Shape of this database is (10841, 13).
- Out of this thirteen columns we have numeric .

User reviews database

- Shape of this database is (64295, 5).
- Here there are only two numeric values found.
- Sentiment Subjectivity, Sentiment , Polarity.

Methodology

The analysis follows a systematic approach, divided into data preprocessing, exploratory data analysis (EDA), and visualization. Below are the main steps:

Data Preprocessing:

1. **Handling Missing Values**:
   - Missing ratings in the **Play Store data** were dropped.
   - Non-essential missing values (such as app size and type) were filled with appropriate methods (e.g., using mode or predefined values).
   - **User reviews** with missing content were dropped.
2. **Data Cleaning**:
   - **Rating** values were converted to numeric, and non-numeric entries were handled.
   - **Installs** were cleaned by removing non-numeric symbols and converted to integer values.

- **Size** data, where "Varies with device" was found, was cleaned and converted to numeric values.

Exploratory Data Analysis (EDA):

- Univariate analysis was performed to understand the distribution of key variables like ratings and size.
- Bivariate analysis explored the relationships between variables such as installs vs. ratings and size vs. ratings.
- Outlier detection was conducted using visualizations like boxplots.

Visualization:

Visualizations were created using **Matplotlib** and **Seaborn** to provide clear insights. Key visualizations include:

- Number of Application in terms of Category
- FREE AND PAID APPS
- Relation between app category and app price
- Sentiment analysis of user reviews
- Distribution of ratings.
- Rating variations across categories.
- Relationships between installs, size, and ratings.
- Filter out "Junk" apps

Data Analysis and Visualization

The columns are also known as features, one or more different features are grouped together for different analyses to form a data frame.

Highest_Rating_App : It contains the data of highest rating apps based on the categories.

Number_of _App: Here, we have shown the number of applications based on different categories.

More_downloads_App: Here, We have shown the highest number of download apps.

Family_lowest_rating: This data shows the lowest rating application of family category.

Family_highest_rating: This data shows the highest rating application of family category.

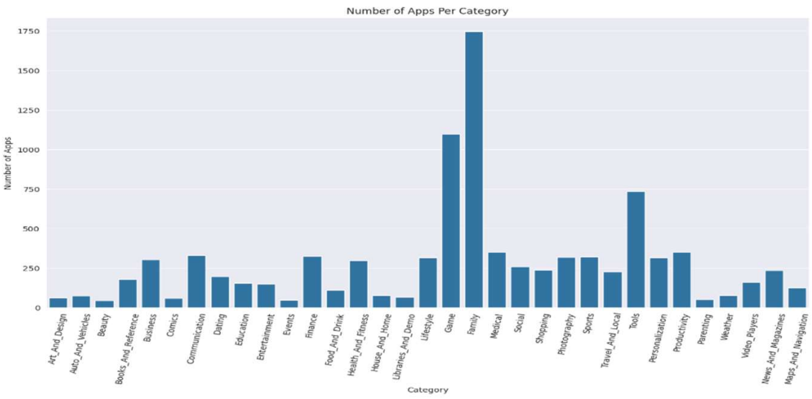
Merged_df: We have merged play store data and user review data.

Top 10 Highest rating Apps in google play store in terms of categories

The analyze is done between the top 10 categories which was having more number of downloads

- As we saw here, the Play store having number of applications in the categories like, Games, Family, Business etc.
- The developers are mostly focusing on these categories because of the people's daily basis requirements.
- Categories like Games, Business, and Family are having comparatively more amount of apps count.

Number of Application in terms of Category



Top 10 apps which has more downloads

- The most downloaded apps on Google Play are primarily Google services like **Chrome, Gmail, Maps, Drive**, and **Photos**, each with over **1 billion installs**. Popular non-Google apps such as **Subway Surfers** and **Skype** also made the list, highlighting the dominance of utility and entertainment apps.

10 apps from the 'FAMILY' category are having the lowest rating and highest rating.

1. **Lowest Rating**
    - In this data frame we have analyzed the lowest rating apps in family category, i.e. BG TV app, EB Mobile.
2. **Highest Rating**

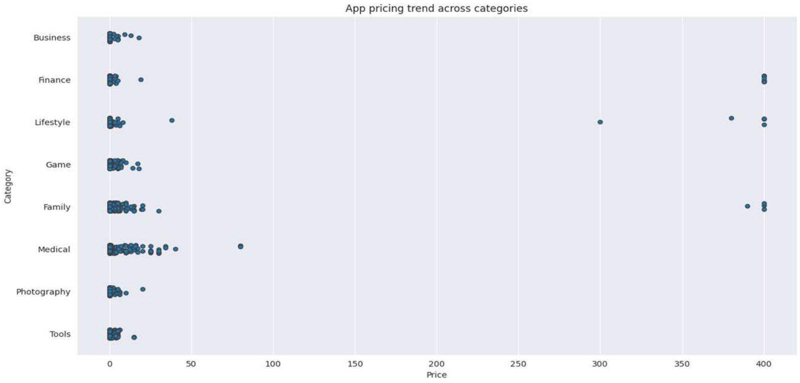
- In this data frame we have analyzed the highest rating apps in family category, i.e. Pyaar Ek Dhoka, Ek Bander Ne Kholi Dukan.

Top 10 apps which has more downloads

- The list of top 10 apps with the highest downloads on the Google Play Store is dominated by **Google's core services**, such as **Google News, Chrome, Gmail, Maps, Drive**, and **Photos**, all exceeding **1 billion installs**. Non-Google apps like **Subway Surfers** and **Skype** also feature prominently, reflecting the popularity of gaming and communication apps worldwide.

Relation between app category and app price

- In this data we have analyzed that which category has the highest price.
- Also we have shown the lowest category.

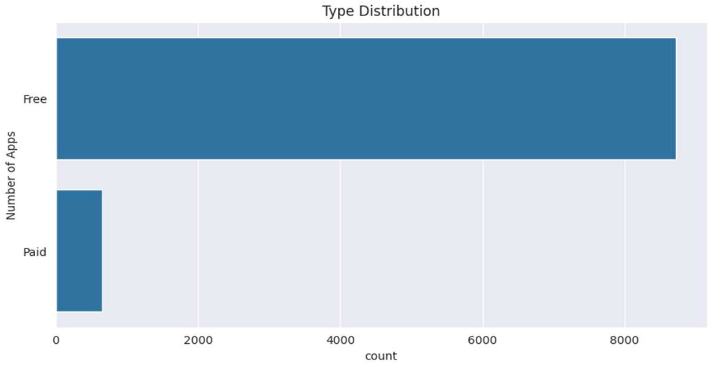


Free apps

- When comparing the both plots, people are showing more interest on free apps like art & design. And The number of free apps are more than 10000.

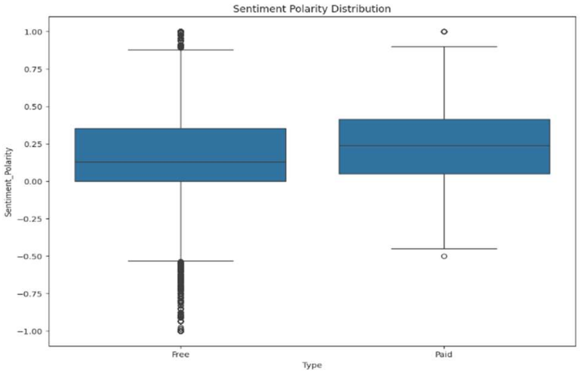
Paid apps

- When it's coming, commercial people are preferring apps like Business, communication and Personalization and the number of paid apps are less than free apps.



Sentiment analysis of user reviews:

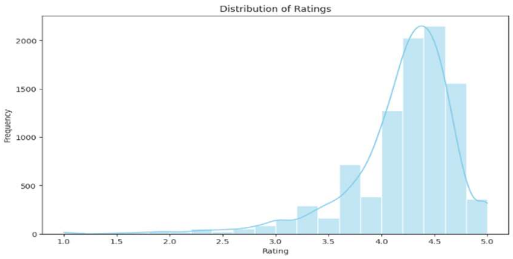
- User reviews for apps can be analyzed to identify if the mood is positive, negative or neutral about that app. For example, positive words in an app review might include words such as 'amazing', 'friendly', 'good', 'great', and 'love'. Negative words might be words like 'malware', 'hate', 'problem', 'refund', and 'incompetent'.
- By plotting sentiment polarity scores of user reviews for paid and free apps, we observe that free apps receive a lot of harsh comments, as indicated by the outliers on the negative y-axis. Reviews for paid apps appear never to be extremely negative. This may indicate something about app quality, i.e., paid apps being of higher quality than free apps on average. The median polarity score for paid apps is a little higher than free apps, thereby syncing with our previous observation.
- In this notebook, we analyzed over ten thousand apps from the Google Play Store. We can use our findings to inform our decisions should we ever wish to create an app ourselves.



Univariate Analysis

Distribution of Ratings:

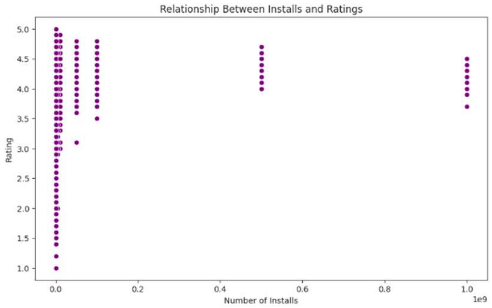
The histogram below shows the distribution of app ratings, with most ratings being concentrated around 4.0. This suggests that users are generally satisfied with the apps on the Play Store.



Bivariate Analysis

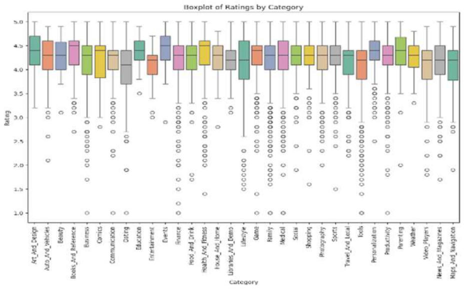
Installs vs. Ratings:

The scatter plot below shows that there is a weak positive correlation between the number of installs and app ratings. This indicates that while popular apps tend to have higher ratings, this is not always the case.



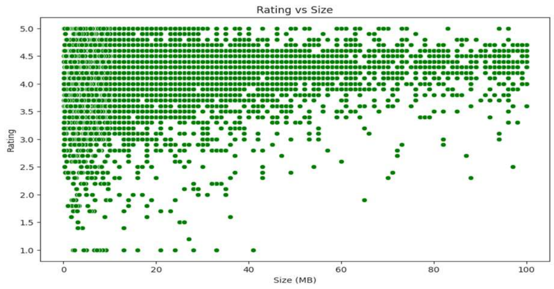
Category vs. Ratings:

The boxplot reveals that certain categories, such as `Family` and `Games`, tend to have higher ratings compared to other categories like `Lifestyle` or `Education`.



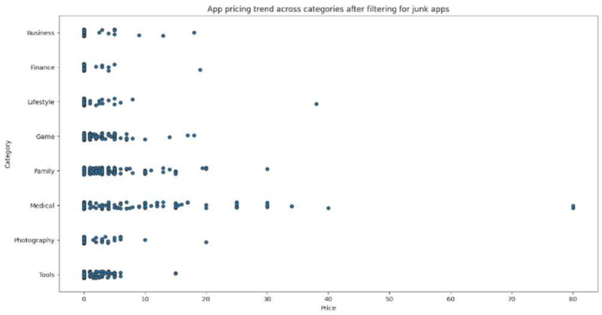
Size vs. Ratings:

The scatter plot shows that app size does not seem to have a strong correlation with ratings. While most apps fall within a reasonable size range, very large or very small apps can have varying ratings.



Filter out "Junk" apps

- It looks like a bunch of the really expensive apps are "junk" apps. That is, apps that don't really have a purpose. Some app developer may create an app called I Am Rich Premium or most expensive app (H) just for a joke or to test their app development skills. Some developers even do this with malicious intent and try to make money by hoping people accidentally click purchase on their app in the store.



Key Findings and Insights

- The **Social**, **Communication**, and **Game** categories have the highest number of downloads.
- **Productivity** and **Health & Fitness** apps tend to have higher ratings.
- Sentiment analysis revealed that user reviews are mostly positive, but some popular apps still face frequent criticism.
- **Free apps** dominate the Play Store, but **premium apps** often receive higher ratings.

Conclusion

This analysis offers valuable insights into factors influencing app ratings and user engagement on the Play Store. Key takeaways for stakeholders include:

- Focus on improving user experience in the highly rated categories.

- While high install numbers do not always correlate with high ratings, developers should focus on quality reviews to improve app ratings.
- App developers can leverage feedback from user reviews, especially terms like "bugs" and "performance," to identify areas for improvement.
- Top 10 Highest rating Apps in google play store in terms of categories
- Number of Application in terms of Category
- Top 10 apps which has more downloads
- Which 10 apps from the 'FAMILY' category are having the lowest rating and highest rating.
- Free and Paid Apps
- Relation between app category and app price
- Filter out "junk" apps

- Sentiment analysis of user reviews