Capstone Project

Play Store App Review Analysis
(Exploratory data analysis)

- **Team Members:**

- **Kumar Abhinav**
- **Saumya Dash**

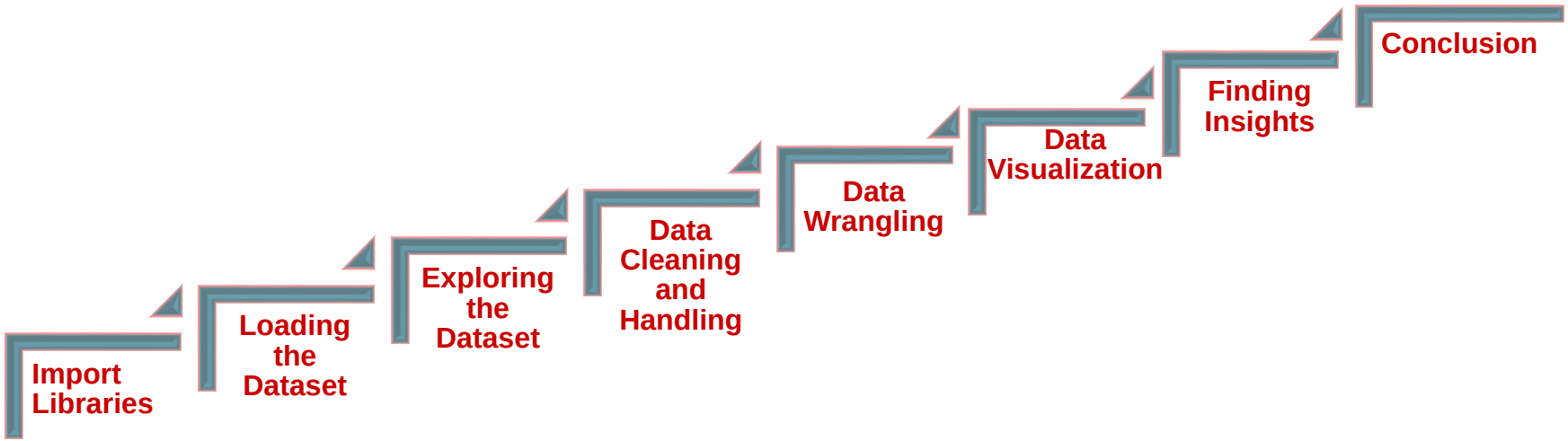
Content

- Introduction
- Data Pipeline
- Exploring Dataset
- Attribute Information
- Problem Statement
- Data Cleaning and Handling
- Data Visualization
- Key Insights
- Conclusion/Recommendations

Introduction

Google Play, also branded as Google Play Store, is a digital distribution service operated and developed by Google. It is the official distribution storefront for Android applications and other digital media, such as music, movies and books, from Google. The Play Store apps data has enormous potential to drive app-making businesses to success. Each app(row) in the dataset contains information such as name of the apps, categories, genres, number of installs, size of the app, date of last update, reviews, sentiment polarity etc. Actionable insights can be drawn for developers to work on and capture the Android market. We are here to explore a play store-review dataset to discover key factors responsible for app engagement and success.

Data Pipeline



Exploring Dataset

- ✓ There are two datasets used in our analysis: Play Store dataset and User-Review dataset.
- ✓ The Play store dataset contains 10,841 rows and 13 features and the User reviews dataset contains 64,295 and 5 columns.
- ✓ The columns in the dataset are App, Category, Rating, Reviews, Size, Installs, Type, Price, Content Rating, Genres, Last Updated, Current Version, Android Version.
- ✓ The columns in the User Reviews dataset are App, Translated Review, Sentiment, Sentiment Polarity, Sentiment Subjectivity.

Exploring Dataset

- ✓ The columns in the User Reviews dataset are App, Translated Review, Sentiment, Sentiment Polarity, Sentiment Subjectivity.
- ✓ There are 4 columns in both the dataset which contains null values i.e Rating, Current Version, Android Version, Type and Content Rating in 1st dataset and Translated Review, Sentiment, Sentiment Polarity, Sentiment Subjectivity.
- ✓ All the columns in the Play store dataset are of object datatype except for Ratings with float values in the 1st dataset.
- ✓ All the columns are of object data type except Sentiment Polarity and Sentiment Subjectivity with float64 datatype.

Attribute Information

- App: Name of the application
- Category: Name of the broad category the app falls under.
- Rating: Average user Ratings on a scale of 0 to 5.
- Reviews: Total number of reviews on each app.
- Size: Size of the app in Megabytes and Kilobytes (represented as M and K in the dataset respectively).
- Installs: Total number of downloads for each app.
- Type: Label stating whether the application is free or paid
- Content Rating: Audience for which the content of the app is intended for such as Teen, Mature, Everyone etc.

Attribute Information

- Genres: Groups according to the characteristics is the Genres. This column includes various Genres to which the app belongs.
- Last Updated: Date on which the app was last updated.
- Current Ver: Current android version of the app.
- Android Ver: The Android operating system the app is compatible with.
- Translated_Review: Customer reviews written in text.
- Sentiment: The sentiment of the review, positive, negative or neutral.
- Sentiment_Polarity: The sentiment in numerical form ranging from -1.0 to +1.0.
- Sentiment_Subjectivity: The measure of expression of opinion, evaluation, feeling and speculation.

Problem Statement

- Which Category has the maximum and minimum number of apps ?
- Which is the most popular category among the users?
- Is there any disparity in app installs and number of apps present in each category?
- How each category of apps perform in terms of customer sentiments?
- Comparison between the apps present in the market and number of installations based on Content Rating.
- What are the top 10 apps that is mostly installed.
- What is the ratio of free and paid apps in each category.
- What is the category wise distribution of average ratings.
- What percentage of apps is supported in higher Android Versions?
- Which are the top 10 Genres in terms of app availability?

Cleaning and Handling Dataset

Drop duplicated instances from both the Data Frames.

- We have used the drop duplicate method to eliminate duplicate rows and instances from the dataset in order to prevent errors in calculation and double counting of the data.

Cleaning and Handling Dataset

Drop or Substitute null values

- Rating column had maximum number of null values i.e.1474 which was filled by 0.
- Converting the Reviews column into integer datatype after checking for Nan and replacing it with zero .
- The instances with null values in Content Rating and Type columns were dropped from the data frame.
- Replaced null values in Current Version with the mode(most frequent current version) i.e. ‘Varies with device’
- Replaced null values in Android Version with the mode(most frequent android version) i.e. ‘4.1 and up’.
- Dropped null values of the User Reviews dataframe.

Cleaning and Handling Dataset

Converting to suitable data types

- Price column have ‘$’ for each prices. To convert it into numerical variable we substituted $ with an empty string “ ”.
- Installs column have ‘+’, ‘,’ for each installs which was replaced by an empty string ‘ ’ and then converted into float datatype.
- Size columns had values in megabytes(M) and kilobytes(k) which was first converted into kilobytes and then finally converted into float datatype.

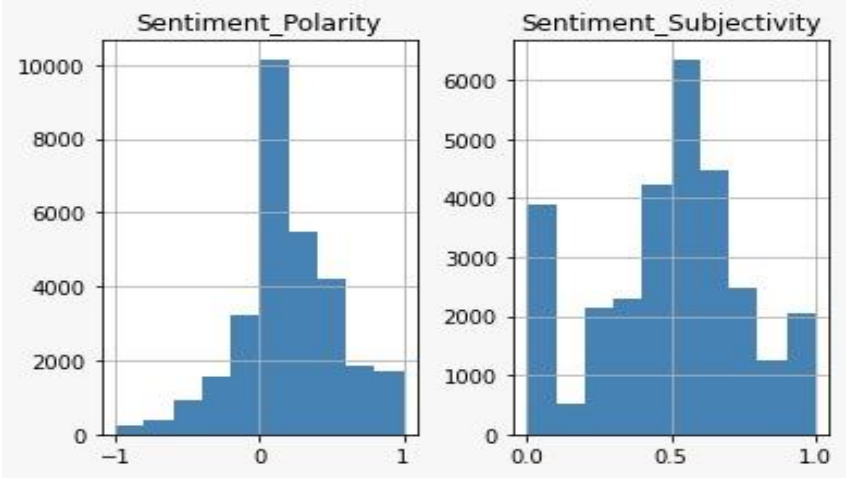
Cleaning and Handling Dataset

Drop or replace wrongly entered data

- Rating variable had a wrongly entered average rating of 19 (ratings greater than 5 is not possible) which was replaced by 0.
- Price variable had a wrongly entered data i.e. ‘Everyone’ which was replaced by 0 and the variable was converted into a float datatype.
- Installs variable had a wrongly entered data i.e. ‘Free’ which was replaced by 0.

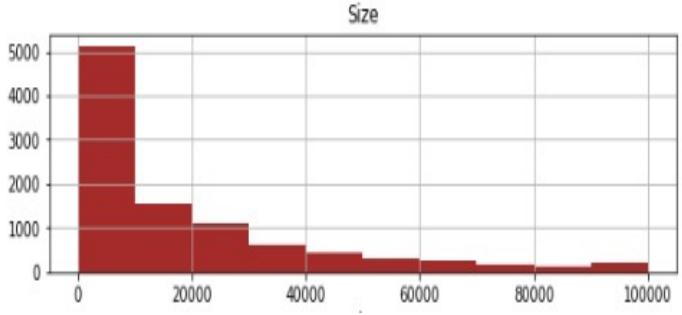
Data Visualization

Univariate Analysis



The Histograms of each numerical variables represents the distribution of the data.

For example: Most of the Ratings lies between 4 and 4.5 and since we replaced the null values with 0 we can see a slight bump near 0.



Data Visualization

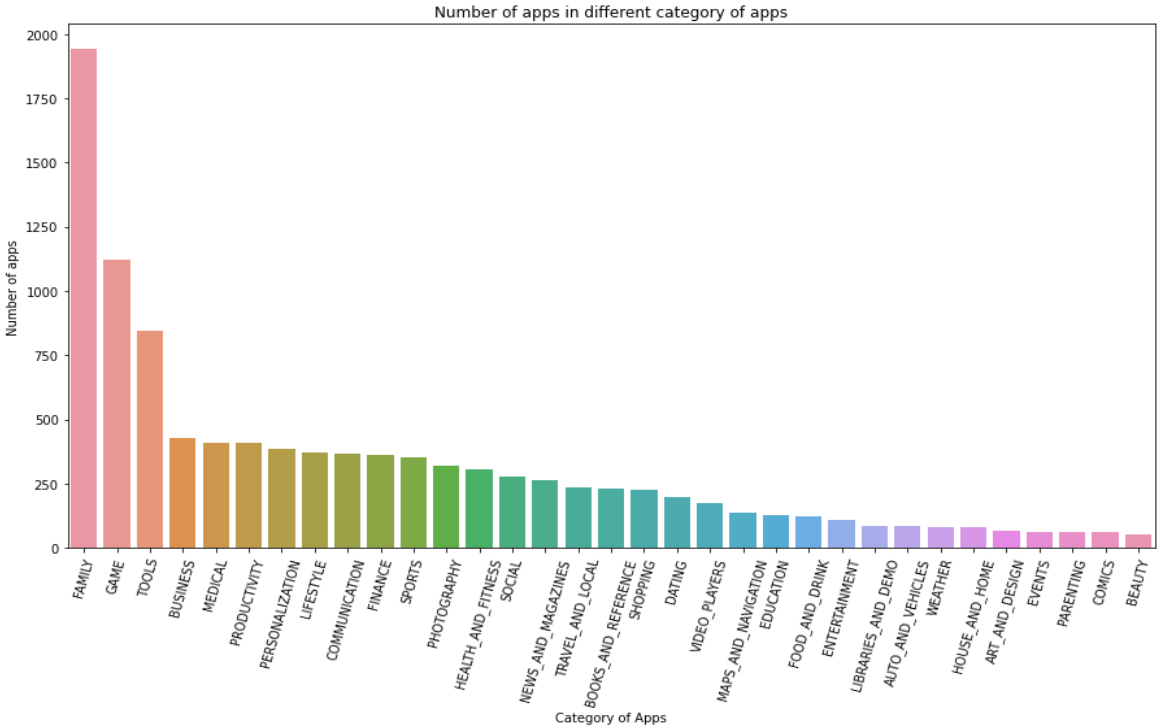
Problem Statement 1: Which Category has the maximum and lowest number of apps ?

1. Top three categories with greatest number of apps are:

- Family
- Game
- Tools

2. Bottom three categories with least number of apps are:

- Parenting
- Comics
- Beauty

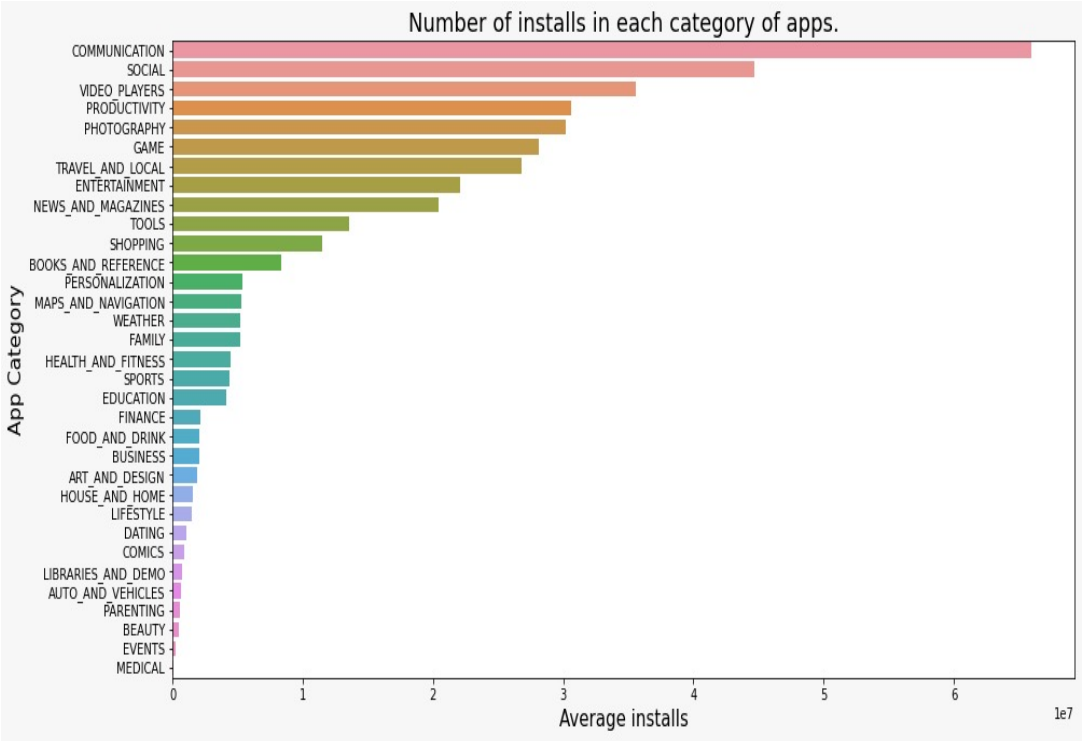


Data Visualization

Problem 2: Which is the most popular category among the users?

1. Top three categories which have most number of downloads are:
   - Communication
   - Social
   - Video Players

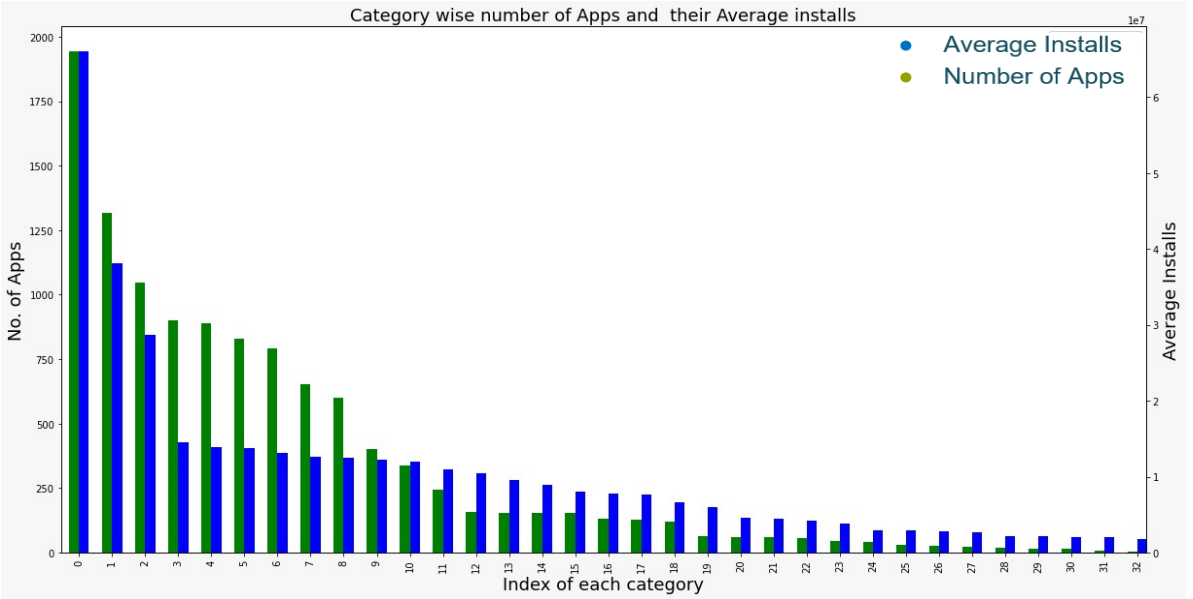
2. Bottom three categories with least number of downloads are:
   - Beauty
   - Events
   - Medical



Data Visualization

Problem Statement 3: Is there any disparity in app installs and number of apps present in each category?

- Here, we can see that there is a disparity between the number of apps available in the market and their users in each category. For example, when compared to the quantity of apps in the market, the Medical category (Index4) has a high average installs.

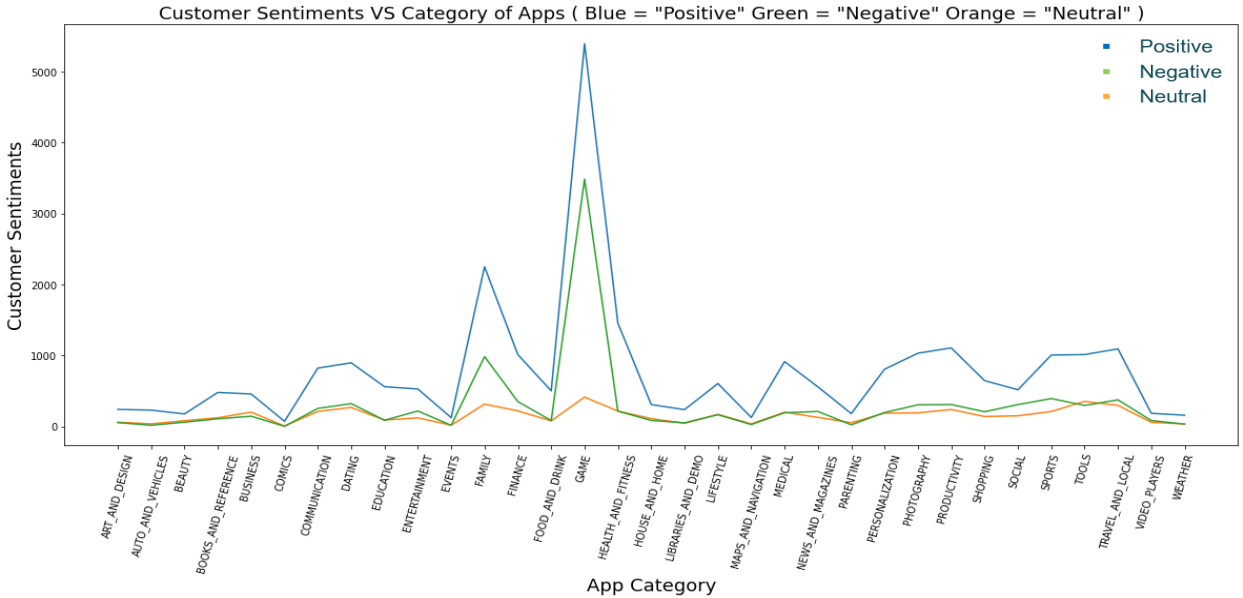


0 – Family,1- Game,2- Tools,3-Business,4-Medical,5-Productivity,6-Personalization,7-Lifestyle,8 Communication, 9-Finance,10-Sports,11- Photography,12-Health and Fitness, 13-Social,14-News and Magazines,15-Travel and Local,16- Books and Reference,17- Shopping,18- Dating,19-Video Players,20-Maps and Navigartion,21-Education, 22-Food and Drink,23- Entertainment,24-Libraries and Demo,25-Auto and Vehicles,26-Weather,27-House and Home, 28-Art and Design,29-Events,30-Parentings,31-Comics,32-Beauty

Data Visualization

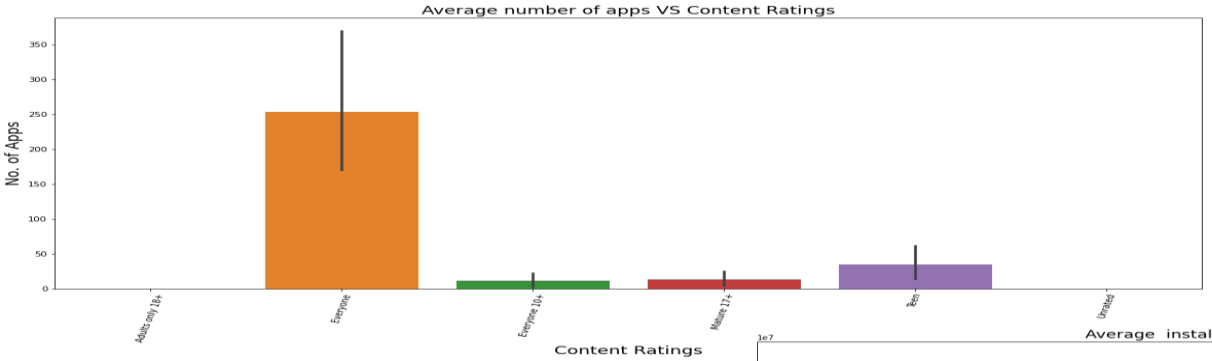
Problem Statement 4: How each category of apps perform in terms of customer sentiments?

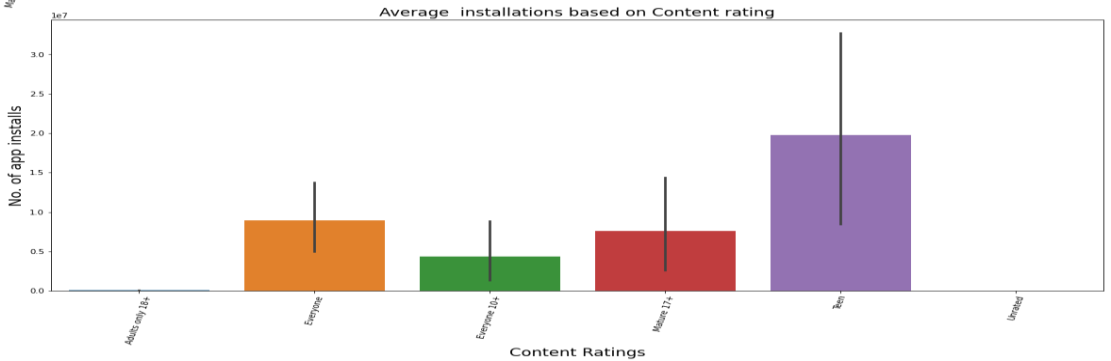
- Reviews for most categories are positive.
- There are comparatively more negative ratings in some categories, notably Family and Games.
- Comics is one of several categories with mixed sentiments



Data Visualization

Problem Statement 5: Comparison between the apps present in the market and number of installations based on Content Rating.



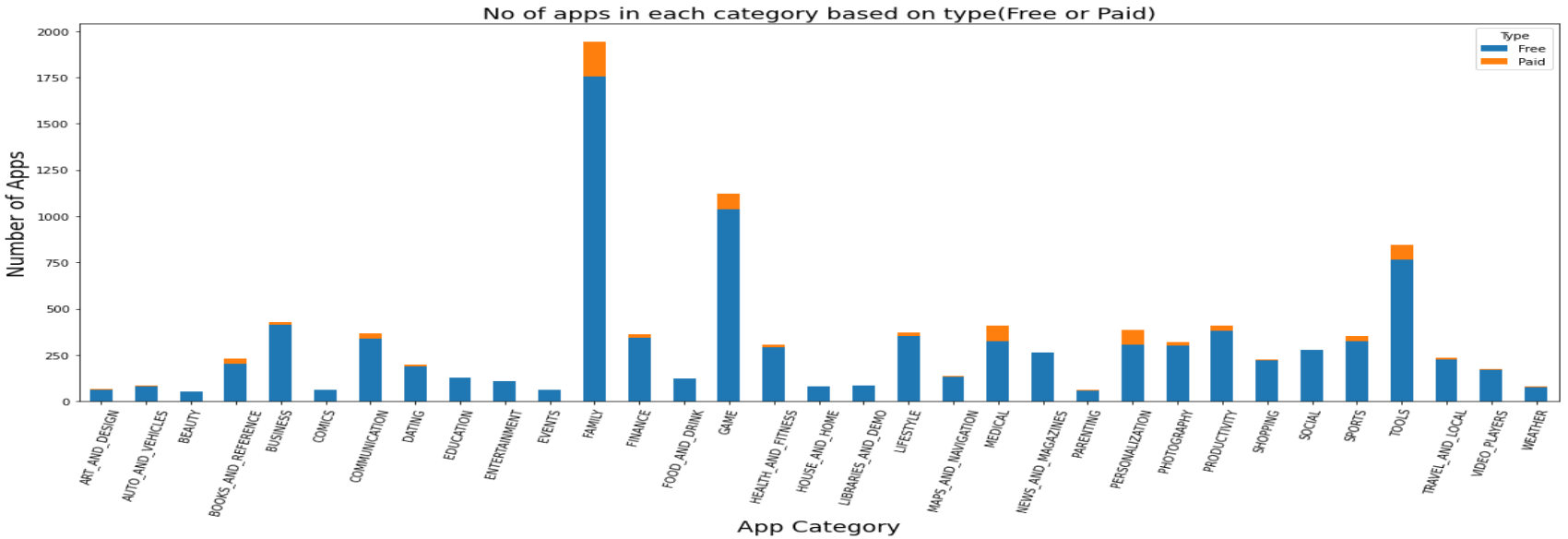


The two graphs above show that while "Everyone" rated apps are widely available, "Teen" and "Mature 17+" rated apps have a sizable user base.

Data Visualization

Problem Statement 6: What is the ratio of free and paid apps in each category.

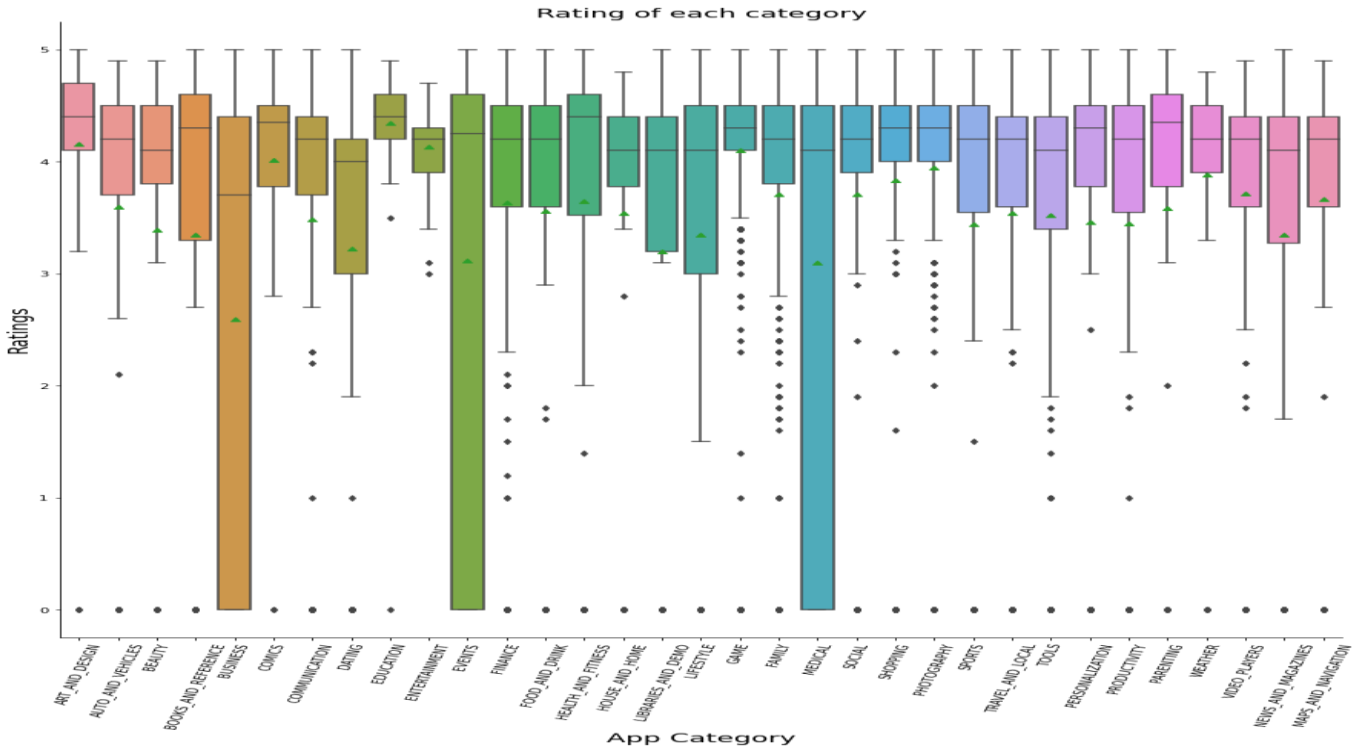
- The majority of apps are free, and family has the most number of paid apps.



Data Visualization

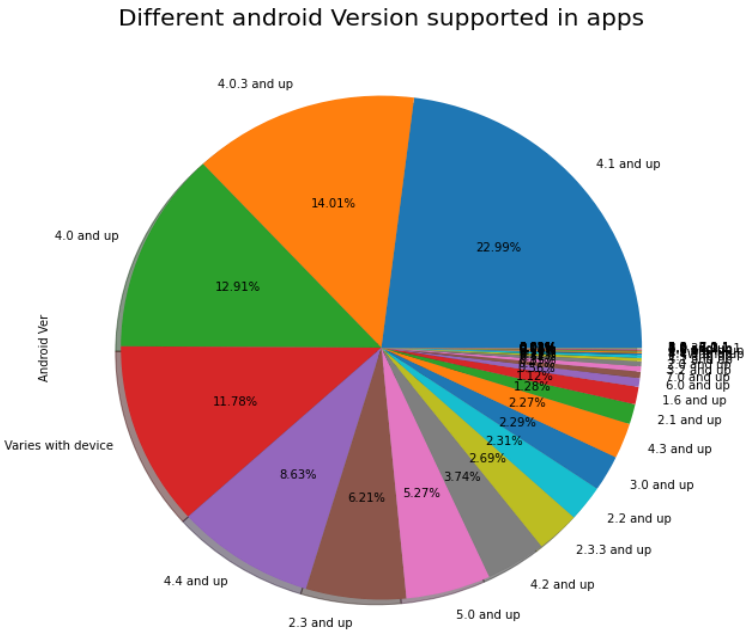
Problem Statement 7: What is the category wise distribution of average ratings.

The quartiles and mean of the ratings for each category are displayed in this boxplot. For instance: The mean of the average rating is roughly 2.6%, and the median (i.e., 50% of the apps) for the business category has an average rating of 3.6%.



Data Visualization

Problem Statement 8: What percentage of apps is supported in higher Android Versions?

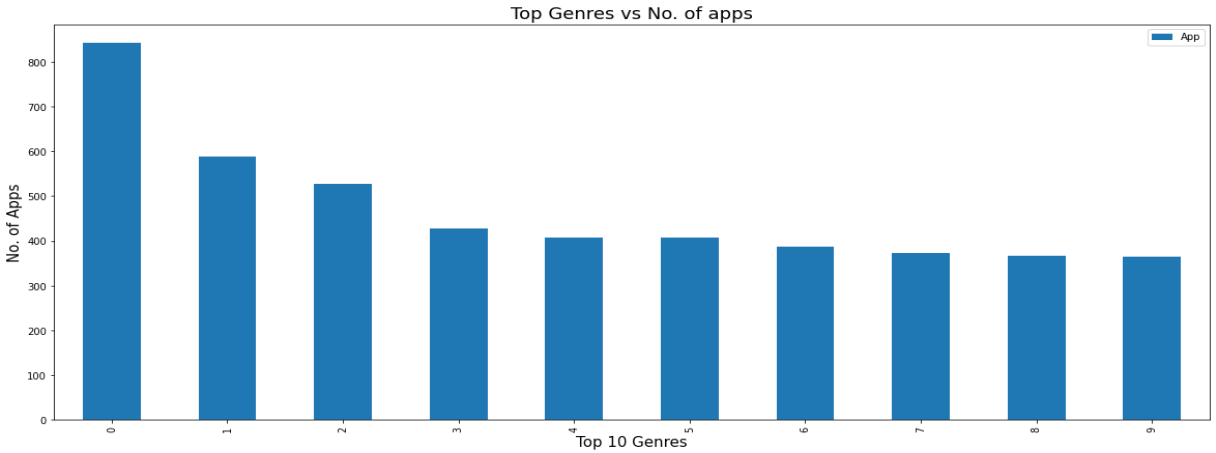


According to the pie chart, over 60% of the apps are compatible with Android 4.0 and higher, while the remaining 40% require updating.

Data Visualization

Problem Statement 9: Which are the top 10 Genres in terms of app?

	Genres	App
0	Tools	842
1	Entertainment	588
2	Education	527
3	Business	427
4	Medical	408
5	Productivity	407
6	Personalization	388
7	Lifestyle	372
8	Communication	366
9	Sports	364



Data Visualization

Problem Statement 10: How correlated the numerical columns are?

This Heatmap displays the correlation between the numerical columns. As we can see, there is a high positive correlation between installs and reviews, meaning that more favorable reviews lead to more app installations. Reviews should therefore be regarded as one of the crucial elements influencing app downloads.

Key Insights

- Most of the Ratings lie between 4 and 4.5.
- Top three categories which have most number of apps are Family, Game and Tools and bottom three categories are Parentings, Comic and Medical.
- Top three categories which have most number of downloads are Communication, Social, Video Players and bottom three categories are Beauty, Events and Medical.
- There is a disparity between the number of apps available in the market and their users in each category.
- Reviews for most categories are positive in comparison to negative and neutral sentiments
- The highest number of apps available is for "Everyone" type of audience but “Teen” and “Mature 17+” rated apps are widely downloaded.

Key Insights

- Top 10 genres with highest number of apps are Tools, Entertainment, Education, Business, Medical, Productivity, Personalization, Lifestyle, Communication, Sports.
- Reviews and installs have a strong positive correlation of 0.64.
- The majority of apps are free, and “family” has the most number of paid apps.
- Over 60% of the apps are compatible with Android 4.0 and higher, while the remaining 40% require updating.

Conclusions

- Some of the Categories like Communication, Social Media , Video Player and Gaming have a significant market gap as they are high in demand among users but there are a very few apps to fulfill this gap. So there is huge opportunity to expand business in these categories.
- Gaming is one of several categories which is popular among both the users and the service providers. However, it is observed that there are some negative sentiments among the users for this category. Therefore, Gaming firms need to improve their services and provide customers a hassle-free experience.
- Play store has the most apps with content that is rated for everyone, however users prefer to download teen and mature (17+and more) rated apps. There is therefore a room to develop a business around it.
- There is room to establish a business on a monthly or yearly subscription model because there aren't many paid apps in the app store. Ed-tech, gaming, and entertainment have a lot of potential in this area.

Thank You!!

The End of a Story.....
The Beginning of Many