Play store app review Analysis

Exploratory Data Analysis

Prepared by :

Kumar Abhinav & Saumya Dash

Data science trainees,

AlmaBetter, Bangalore

Abstract:

The data from Play Store apps has a great deal of potential to help app development companies succeed. Developers can use these insights to build on and conquer the Android market. After breaking down the given data set into smaller, more digestible data frames, several insights can be obtained. These insights can be used by the stakeholders of the companies to make data-driven decisions.

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 etc. The Play Store apps data has enormous potential to drive app-making businesses to success. Each instance 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 Exploration

There are two data set used in our analysis: Play store dataset and User-review dataset. The play store data set contains 10,841 instances and 13 attributes while user review dataset contains 64,295 instances and 5 attributes

The details of attributes of Play store data set is given below:

Columns	Description
App	Name of the App
Category	Category under which app falls
Rating	Application's rating on play store
Reviews	Number of reviews of the app
Size	Size of the app
Installs	Number of Installation of the app
Type	Whether the app is free or paid
Price	Price of the app if it's a paid app (0 if it's a free app)
Content Rating	Appropriate target audience of the app
Genres	Genres under which the app falls
Last Updated	Date when the App was last updated
Current Version	Current version of the App

Android Version	Minimum android version required to support the App

Similarly for the user review dataset details of its column are given below:

Columns	Description
App	The app name.
Translated_Review	Review text in English.
Sentiment	Sentiment of the review, which can be positive, neutral, or negative.
Sentiment_Polarity	Sentiment in numerical form, ranging from -1.00 to 1.00.
Sentiment_Subjectivity	Measure of the expression of opinions, evaluations, feelings, and speculations

There are four columns in both the dataset which contains null values that is Rating, Current Version, Android Version, Type and Content Rating in first data set and Translated_Review, Sentiment, Sentiment_Polarity and Sentiment_Subjectivity . All the columns in play store dataset are of object datatype except for the Rating column which have float values. Similarly in the user review data set all the columns are of object datatype except Sentiment_Ploarity and Sentiment_Subjectivity.

Standard Operating Procedure:

In the process of Exploratory Data Analysis we followed a particular procedure for obtaining best outcome from the data:

1. Importing required packages for analysis.
2. Mounting drive and reading data files from Google drive.
3. Removing future warnings in seaborn plots.
4. Viewing all data information.
5. Dropping duplicate.
6. Removing special characters
7. Checking unique values, null count and datatypes of each column.
8. Segregation of numerical and categorical data.
9. Identifying the problem statements.
10. Doing groupby , sorting and aggregating operation to come up to relevant data frame.
11. Doing data visualization as per the problem statements.
12. Finding Insights from the data visuallization.
13. Giving recommendations.

.Data Cleaning Operation:

1. **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.

2. **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 .

3. **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 “+” and “ , “ 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 .

4. **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 frames Used:

In the process of data analysis we formed many data frames to breakdown to given data set list of which are given below:

Data Frames	Description
df	Play store data
df1	User review data
df_category_vs_count	Category wise no. of apps
df_category_vs_installs	Category wise Installations
df_installs_vs_counts	Category wise no. of apps and Installations
df_catagory_vs_sentiments	Category wise positive, negative or neutral sentiments
df_catagory_vs_content_rating	Category wise no. of apps for each Content Ratings
df_catagory_vs_content_rating_installs	Category wise average installations for each Content Ratings
df_top_app	List of top 10 apps
category_vs_type_df	Category wise list of paid or free apps
count_Android_Ver	Percentage of apps in each android version
df_genre_vs_count	Top 10 genre with highest number of apps.

User Defined Functions used:

- **Replace_M_k():**

This function takes the row element (Size column) and returns te replacement of 'M' with 000 and 'k' with nothingand 0 if there is other data than this.

Approach used:

Exploratory Data Analysis:

Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations.

In statistics, **exploratory data analysis** is an approach of analyzing data sets to summarize their main characteristics, often using statistical graphics and other data visualization methods. A statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modeling and thereby contrasts traditional hypothesis testing. Exploratory data analysis has been promoted by John Tukey since 1970 to encourage statisticians to explore the data, and possibly formulate hypotheses that could lead to new data collection and experiments. EDA is different from initial data analysis (IDA), which focuses more narrowly on checking assumptions required for model fitting and hypothesis testing, and handling missing values and making transformations of variables as needed. EDA encompasses IDA.

Tools and techniques

Typical graphical techniques used in EDA are:

- Box plot
- Histogram
- Multi-vari chart
- Run chart
- Pareto chart
- Scatter plot (2D/3D)
- Stem-and-leaf plot
- Parallel coordinates
- Odds ratio
- Targeted projection pursuit
- Heat map
- Bar chart
- Horizon graph
- Glyph-based visualization methods such as PhenoPlot and Chernoff faces
- Projection methods such as grand tour, guided tour and manual tour
- Interactive versions of these plots

Dimensionality reduction:

- Multidimensional scaling
- Principal component analysis (PCA)
- Multilinear PCA
- Nonlinear dimensionality reduction (NLDR)
- Iconography of correlations

Typical quantitative techniques are:

- Median polish
- Trimean
- Ordination

Data Visualization:

Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data. Additionally, it provides an excellent way for employees or business owners to present data to non-technical audiences without confusion.

Advantages

Our eyes are drawn to colors and patterns. We can quickly identify red from blue, and squares from circles. Our culture is visual, including everything from art and advertisements to TV and movies. Data visualization is another form of visual art that grabs our interest and keeps our eyes on the message. When we see a chart, we quickly see trends and outliers. If we can see something, we internalize it quickly. It’s storytelling with a purpose. If you’ve ever stared at a massive spreadsheet of data and couldn’t see a trend, you know how much more effective a visualization can be.

Some other advantages of data visualization include:

- Easily sharing information.
- Interactively explore opportunities.
- Visualize patterns and relationships.

Disadvantages

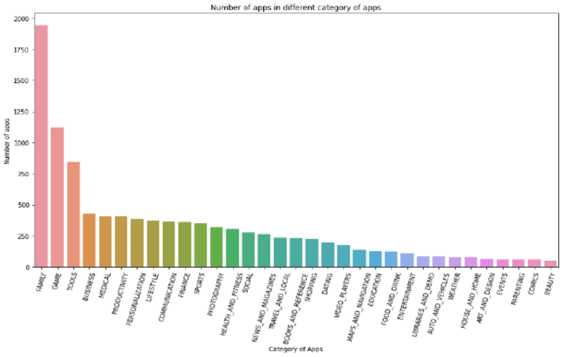
While there are many advantages, some of the disadvantages may seem less obvious. For example, when viewing visualization with many different data points, it’s easy to make an inaccurate assumption. Or sometimes the visualization is just designed wrong so that it’s biased or confusing.

Some other disadvantages include:

- Biased or inaccurate information.
- Correlation doesn’t always mean causation.
- Core messages can get lost in translation.

Problem Statements and Outcomes:

1.Which Category has the maximum and minimum number of apps?

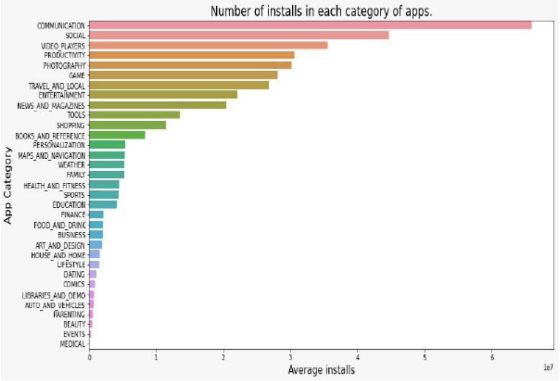


From the bar chart above we can say that:

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

2. Which is the most popular category among the users?



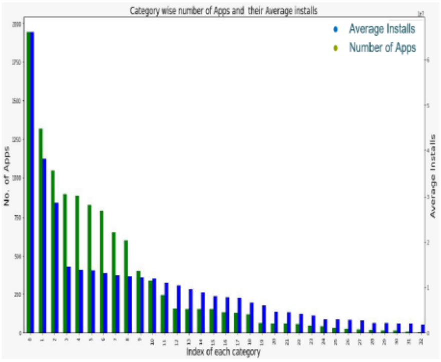
1. Top three categories which have most number of downloads are:

- Communication
- Social
- Video Players

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

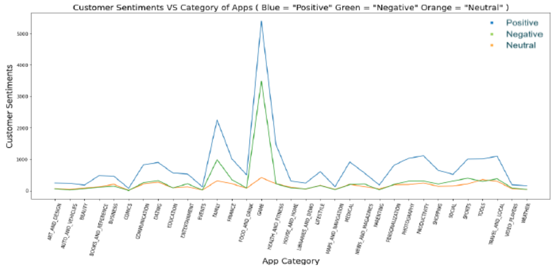
- Beauty
- Events
- Medical
-

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



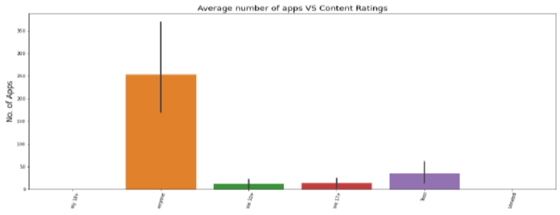
Here, we can see that there is a disparity between the number of apps available in the market and their users in each category. For example, when compared to the quantity of apps in the market, the Medical category (Index4) has a high average installs.

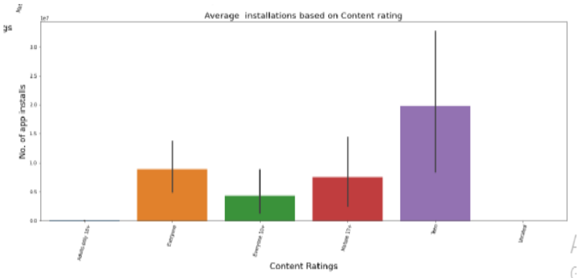
How each category of apps perform in terms of customer sentiments?



- 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

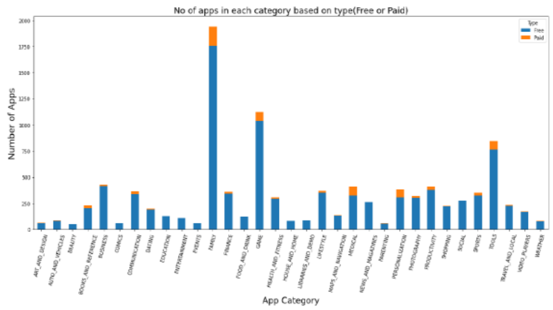
Comparison between the apps present in the market and number of installations based on Content Rating.





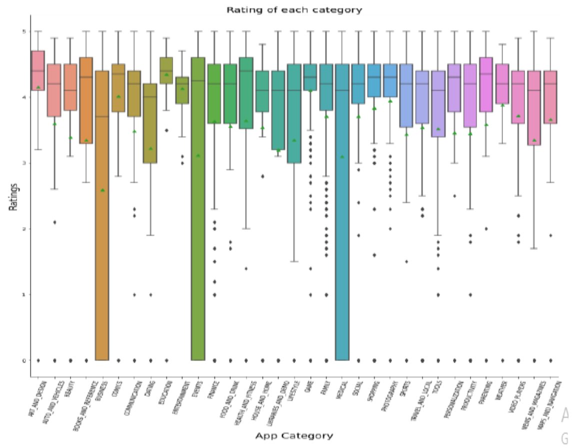
The two graphs above show that while "Everyone" rated apps are widely available, "Teen" and "Mature 17+" rated apps have a sizable user base

What is the ratio of free and paid apps in each category?



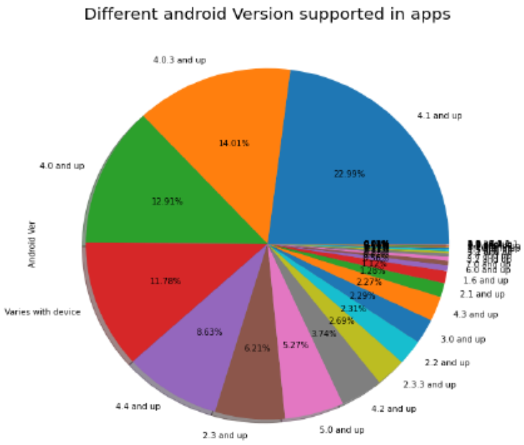
The majority of apps are free, and family has the most number of paid apps.

What is the category wise distribution of average ratings?



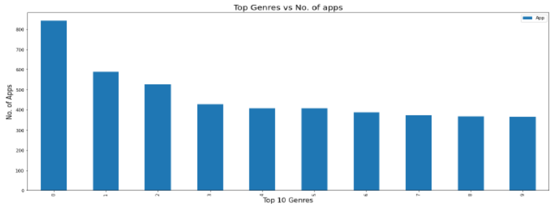
The quartiles and mean of the ratings for each category are displayed in this boxplot. For instance: The mean of the average rating is roughly 2.6%, and the median (i.e., 50% of the apps) for the business category has an average rating of 3.6%.

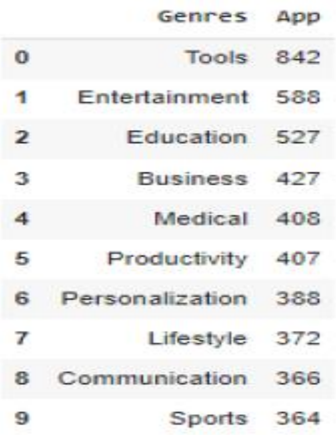
What percentage of apps is supported in higher Android Versions?



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

Which are the top 10 Genres in terms of number of apps?





	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

Conclusion:

- One of the many categories that is well-liked by both customers and service providers is gaming. However, it has been noted that some people have negative opinions about this category. Gaming companies must therefore enhance their offerings and give customers a hassle-free experience.
- Some categories, including as communication, social media, video player, and gaming, have a huge market gap due to high user demand, yet there are very few apps to fill this gap. As a result, there is a lot of room for growth in these areas.
- Users prefer to download apps with teen and mature (17+ and more) ratings, even if the Play Store has the most apps with material that is rated for everyone. Consequently, there is room to build a company around it.
- Because there aren't many paid apps in the app store, there is space to build a company using a monthly or annual subscription model. There is a lot of opportunity in this space for ed-tech, gaming, and entertainment.

Acknowledgement:

This project is presented by**- Saumya Dash and Kumar Abhinav**. Our sincere efforts have made us to accomplish the task of completing this project. We are extremely grateful to our instructors and mentors who have helped us to grow in this field. We would like to express our sincere gratitude to the celebrated authors whose phenomenal work has been consulted and referred in our project work. We also wish to convey our appreciation to our peers who provided encouragement and timely support in the hour of need. This project helped us improve our skills and enhanced our knowledge.

References:

https://seaborn.pydata.org/tutorial/color_palettes.html

https://www.python-graph-gallery.com/196-select-one-color-with-matplotlib

https://www.analyticsvidhya.com/blog/2021/05/10-colab-tips-and-hacks-for-efficient-use-of-it/