Evaluating the Impact of Application Attributes on User Engagement and Success in the Google Play Store

Problem Statement:

In the digital age, Google Play Store has become a collection of mobile applications, a large number of applications make each other increasingly fierce competition. How do you measure the success of an app? Metrics such as user feedback and downloads may provide some insight, but the real success factor is still hidden deep in the data. With this in mind, this project aims to take advantage of the vast Google Play Store app dataset to dig deeper into the key factors that stand out in the Android market.

Significance:

The central question facing this project is: among the many applications, what factors really drive their success? To answer this question, we conducted a comprehensive analysis of data from nearly 11,000 apps, covering categories, ratings, user reviews, size, installs, price, content ratings, frequency of updates, version information, and more. Our goal is to reveal the intrinsic link between these features and market performance, thereby providing powerful data support for application development and marketing.

Contributions to this Domain:

The results of this project will have a profound impact on the entire field of mobile application development. By revealing the key factors for success, it is expected to lead developers to innovate and improve app design, update strategies, and marketing methods, thereby driving the industry toward higher quality, better user experiences, and a more prosperous app ecosystem. This is not only an in-depth insight into the current market, but also a bold prediction and guidance for the future trend.

Data Cleaning/Processing:

1. Dropping columns
    1.1. Dropped extra index column
    1.2. Dropped both “Current Ver” and “Android Ver” because both columns contain too many missing values(“Varies with device”), and no good techniques to fill in those missing values.
2. Using domain knowledge, we know if an app has a missing rating is probably because no one has rated the app yet thus we can replace it with 0.
3. Using domain knowledge, we know the highest rating you can have on an app in the Google Play Store is 5, any ratings that’s higher than 5 are invalid data which we should filter out.
4. The unit for the size of apps varies between KB and MB, we can unify the unit for all app size to MB by converting KB to MB.
5. A lot of data for app sizes are missing, we can fill in these missing data using the mean of all the sizes of the “Category” of the app that we’re trying to fill in. Ex. If the app “Messenger” has a missing size data, we will fill the missing value with the mean of the “social” category.
6. Rating of the app has inconsistent value, Ex. “Everyone, Mature, Teens. We know mature means age 17+ and teens means 13+ thus we can replace these values with the corresponding age.
7. Removing the ‘$’ sign since we know the unit for the data is dollar thus there is no need for the ‘$’, and also converting the data type to float for better graphing.
8. Removing ‘+’ sign in the “Installs” column to prevent inconsistency, and ‘,’ sign in order to convert the number of installs to int.

9. Scaling the “Installs” column down by 1000 to reduce huge numbers. There are only a handful of apps that have less than 1000 installs which we would use 0 to represent any apps that have less than 1000 installs.
10. Converting values from the “Last Updated” column to the number of days that have passed since the last update instead of the actual dd/mm/yy format for a better understanding of the data.

Exploratory Data Analysis (EDA):

1. From exploring the App’s Rating to Price graph, we can see that the mean of ratings for all apps in the Google Play Store is 3.54 stars and a mean of $1.09 for the price of the app. From this graph we can conclude that the majority of the apps in the Google Play Store are low cost with 92.62 percent of apps being free. In order for new apps to be competitive against existing apps in terms of rating, the price of the app needs to be low enough to attract users to use. Although there are also a few outlier apps with high cost that cost hundreds of dollars which is excluded from the graph, they still have a higher than average rating. Our hypothesis on why these expensive apps have higher than average ratings is due to the lack of volume of reviews which could be biased.
2. The pie chart shows the distribution of apps based on its category, and we can see that “FAMILY" is the most popular category taking 19 percent followed by “GAME” taking 9.9 percent of all apps in the Google Play Store. With this information, we can conclude that choosing “FAMILY” or “GAME” as the category for new apps to go into would mean there would be more competition than other categories, which could negatively affect the chance of new apps to be successful.
3. Looking at the scatter plot of App’s Rating vs Days Since Last Update, we can see that there is a trend where the higher the rating of an app, the longer its last update was, with a mean of 2309 days. We can hypothesize that the reason why

higher rating apps have longer-last updates could be because the customer is satisfied with the app thus the developers won’t need to make frequent changes to the app. Where in contrast, lower rating apps have shorter last updates could be because users are not satisfied with the app and developers are making frequent changes to fulfill user’s demand on requested features of the app. There are also a lot of outliers that have ratings due to various reasons which is why it’s not included in the graph. As a result, we can use these data to predict which app is more likely to change or stay the same depending on how often they update.

4. In Box Plot of Rating by Category, we can find the median, interquartile range, outliers, symmetry from the graph, and we can use these scores to measure customer satisfaction for different product categories. A higher median and fewer outliers may indicate that customers are generally satisfied with a category. If we need to make decisions at a company about how to allocate company resources, we can use these score distributions to decide which categories need the most investment and improvement. We can notice that educational apps need to be around 3.5 to be considered outliers, so the average score of educational apps is very high. Such a result will make educational apps easier to invest in.
5. In 'Box Plot of Size by Category', we learned how to analyze data using data distributions, medians, outliers, and comparing application sizes across categories. We found a few outliers in the games category. The median of MEDICAL is very high. It may be that some functions of the MEDICAL app require the support of a large amount of data. After learning this information, if we want to develop software later, we can decide our own application design based on this data. For example, we may choose to develop a smaller application. This can give our app unexpected advantages.
6. In “Average App installs in Each Category”, we can see the average app installs for each type of application. The histogram for "Communication" type apps stands out sharply, representing the highest average downloads. Famous means that the market demand for this type of application is the highest among all applications. The "social" type, which is similar to "communication", can also be seen from this result. People have great demand and interest in "communication"

type applications. Expansion and application designers can create more according to this market demand. "Communication" application.

7. In ‘’Average App Age Restriction in Each Category”, we can see the average age limitation for each type of the application. The “dating” category of apps have the highest average age limitation. To use the dating apps, the age of the consumer should be over 14 years old. Also, the entertainment and social apps show the high restriction for consumer’s age. On the other hand, some apps like Food and drink, or travel and local indicate they have lower age limitation for the user. We can see that the apps with high age restriction mostly belong to the category which involve adult content and serve. However, the apps with lower age limitation such as Food and drink always come with more general applicability. This result can help developers to design the apps for specific age group on the trend in the graph. Additionally, the parents can use these information to screen apps for their children.
8. In Pie chart of content rating distribution, We can see from the figure that the vast majority of apps have no age restrictions, accounting for about 81.8% of the apps. They are the main group using the app. 13-year-old users account for 10.7%, 17-year-old users account for 4.1%, 10-year-old users account for 3.3%, and the proportion of 18-year-old users is very small and almost negligible. When we apply this data, we can think about if we design an app for children or teenagers, we can adjust the content and features based on age distribution to attract the largest group of users. Or we can also consider a small number of users and make software for the target group of 18-year-old users. Although this user group can be ignored, they are adults and may be more likely to pay for the software.
9. In “Average Rating of Number of Installs Group”, the line graph shows the correlation between the average rating of apps and number of installs. We can see that the apps have the highest rating when there are fewer reviews. As the number of reviews increased, the app's rating actually decreased. From such results, we can see that more app downloads may lead to more negative reviews. When fewer apps are downloaded, that app tends to have a higher

rating. As the number of downloads increases, it may be difficult for an app's rating to maintain its previous level, as it must meet more user needs and expectations. This means that developers and app designers need to promote their apps as much as possible to provide a better user experience, so that even with more complex user differences, the app's rating is difficult to affect.

10. In Line graph of average price of each number of install groups, can see the relationship between the average price of apps and the number of downloads. As the price of an app decreases, the number of app downloads will increase. When the price of an app is high, the number of downloads of the app is low. If an app has a lower price or is free, the app will have higher downloads. From this situation, we can know that low-priced applications are more attractive to customers than high-priced applications. This result shows that the future trend of low-priced and free applications is better than that of high-priced and paid applications. It suggests that when developers or designers are dealing with pricing for their apps, they can consider lower or free pricing while adding in-app purchases to increase revenue.

Citation:

Data Set:
https://www.kaggle.com/datasets/bhavikjikadara/google-play-store-applications

Pandas:https://pandas.pydata.org/docs/

Matplotlib: https://matplotlib.org/stable/users/index.html

Seaborn: https://seaborn.pydata.org/