Exploratory Data Analysis of Google Playstore App Data

Nandeesh Umesha, Vilas Sonawane, Vikas Singh, Asad Aslam

Abstract

Google Playstore is a marketplace for android apps. The project involves exploratory data analysis on the Google Playstore app dataset. The project aims to identify key insights that drive the popularity of an app and perform market research on the Google Playstore apps. The data available consists of two CSV files. The first one consists of information like rating, number of reviews, number of downloads, price, category and genre of 9660 unique apps in the Google Playstore. The second CSV file had 64295 written user reviews corresponding to 1074 apps. It also had three additional columns - Sentiment, Sentiment Polarity and Sentiment Subjectivity obtained after performing a sentiment analysis on the user reviews.

The project explores various dimensions of the android app market by answering the following questions:

1) Are free apps of poor quality?
2) Are free apps more popular?
3) What type of apps are users ready to pay for?
4) Are heavy apps not popular?
5) Market Research of android apps - including market dominance by big tech players
6) Are apps with good ratings more popular?

Interesting insights were identified - some conforming to the commonsensical intuition and some counter-intuitive to the common understanding. For example, it was proved that the popularity of an app does increase with high ratings and it was observed that a larger app size does not negatively affect an app's popularity.

Keywords: *Google Playstore, Exploratory Data Analysis, Matplotlib, Data Visualization, android app, market research.*

Introduction

Since the invention of the first handheld smartphone[1] by an IBM engineer in 1994, our lives have undergone a phenomenal transition. Today, the majority of businesses cease to exist if smartphones were to disappear suddenly. The technological advancements in low-cost internet, transistors, displays, processors and memory chips have enabled smartphones to reach the remotest parts of the earth. In India alone, there are 500 Million active internet users and 450 Million access the internet over smartphones.[2]

Google Play Store is a digital distribution service operated and developed by Google. It is an official platform for registered android OS mobile users to download the

apps published by various developers. In 2019 alone, 116 Billion apps were downloaded across the 190 countries of the world. Consumer spending on Android apps and games increased by 80% in 2019-20.[2] Developers around the world(excluding China) have earned more than $ 80 Billion with Google Play.[2]

The Play Store apps data has enormous potential to drive app-making businesses to success. Actionable insights can be drawn for developers to work on and capture the Android market. In this project, we explore the Google Playstore apps data set and identify key insights and factors that a developer can utilize to make data-driven decisions.

Data Description

The dataset included two CSV files:

I. Play Store Data.csv:

It consists of 10841 rows and 13 columns. Each row corresponds to information regarding an app like ratings, reviews, downloads, category, price etc. The columns are listed below:

Feature	Description
App	Name of the application. 9660 unique apps
Category	The category of the application lik 'ART_AND_DESIGN', 'AUTO_AND_VEHICLES', 'BEAUTY'
Rating	The rating, the application received in the Google Playstore. Out of 1-5.
Reviews	The number of user reviews
Size	Size of the application.
Installs	The number of installations of the app. The number of installations was classified into 22 buckets. Ex: 100+, 1000+, 10000+....
Type	Whether the app is free or paid.
Price	The price of the app.
Content Rating	The age-restricted rating of the app. Categories include 'Everyone', 'Teen', 'Everyone 10+', 'Mature 17+', 'Adults only 18+', 'Unrated'
Genres	The genre of the app
Last Updated	The date of the last update to the app in January 26, 1994 format
Current Ver	Current app version
Android Ver	Which android version is compatible with the app

II. User Reviews.csv:

It consists of 64295 rows and 5 columns. Each row corresponds to a written user review for an application. The columns are listed below:

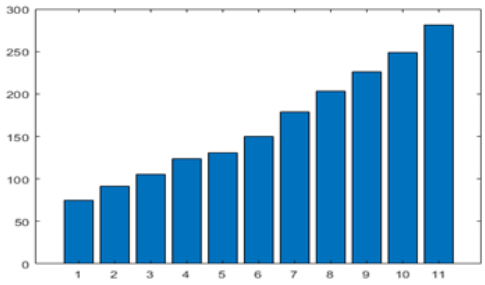
Feature	Description
App	The app that the user review pertains to. 1074 unique apps.
Translated Review	The translated text of the user review
Sentiment	The sentiment predicted from the text of the review. Three categories - Positive, Neutral, Negative
Sentiment Polarity	Polarity is a float that lies in the range of [-1,1] where 1 means positive

	statement and -1 means a negative statement.
Sentiment Subjectivity	Subjective sentences generally refer to personal opinion, emotion or judgment whereas objective refers to factual information. Subjectivity is also a float that lies in the range of [0,1]. Subjectivity of 1 means a perfectly subjective(public opinion and not a fact). Subjectivity of 0 means the text is a fact.

Visualizations used

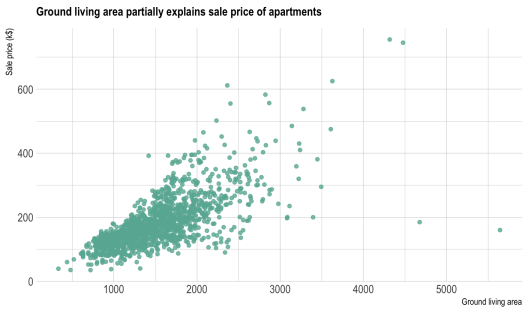
A. Bar Chart:

A bar chart is a graph that presents categorical data with rectangular bars with heights proportional to the values that they represent. The bars can be plotted vertically or horizontally.



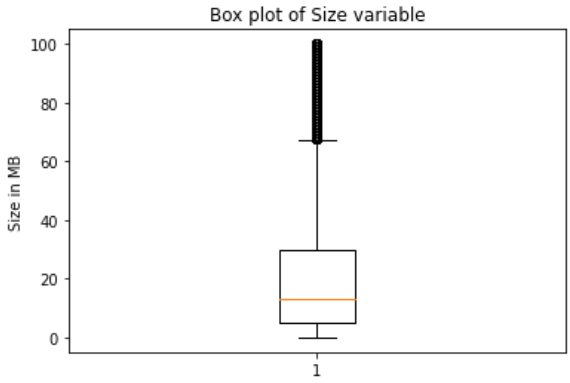
B. Scatter Plot:

The values for each data point are shown by the position of each dot on the horizontal and vertical axes. These Scatter plots are used to show how different variables relate to each other.



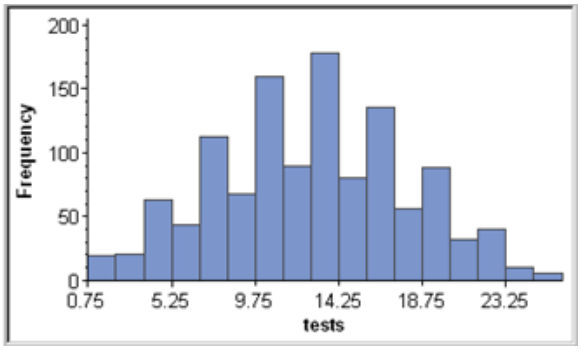
C. Box Plot:

A box plot displays summary statistics for the distribution of values for a variable. The outer bounds of the box represent the first and third quartiles. The line inside the box represents the median. The edges of the box parallel to the median line indicate the 25th and 75th percentiles. The circles outside the whiskers represent the outliers.



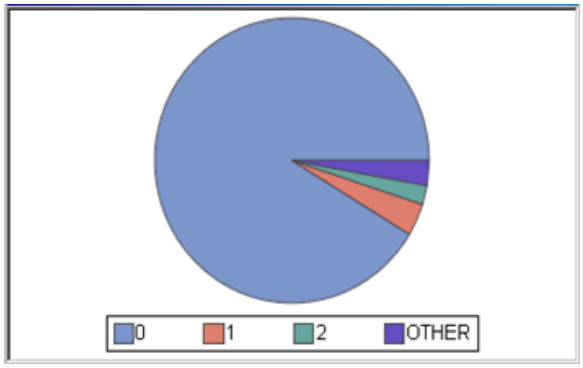
D. Histogram:

A histogram is a bar chart that displays the frequencies of data in bins. The heights of the bars indicate the relative frequency of observations in each bin. It is used to visualize the data distribution.



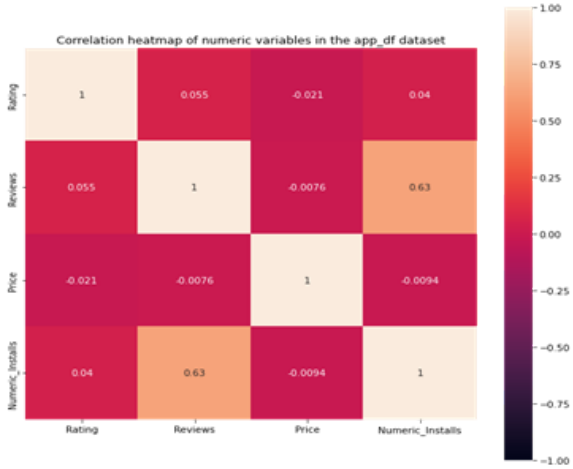
E. Pie Chart:

A pie chart is a circular representation of the statistical graphic, which is divided into various slices to show all the desired data in numerical proportions. Each slice's arc length (and thus its central angle and area) in a pie chart is proportional to the quantity it represents.



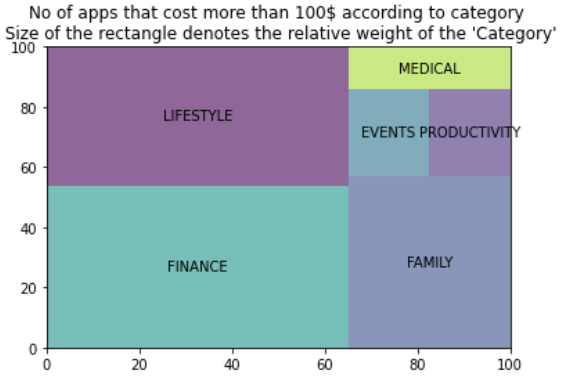
F. Heat Map:

Heat maps are used to show relationships between two variables, one plotted on each axis. By observing how cell colours change across each axis, you can observe if there are any patterns in value for one or both variables.



G. TreeMap:

A treemap provides a hierarchical view of the data along with proportions. The tree branches are represented by rectangles and each sub-branch is shown as a smaller rectangle. The size of the rectangle indicates the relative weight of that category.



Methodology

A. Data Cleaning:

Play Store Data.csv:

Box plots were used to identify the range and outliers for the numerical data type.

Rating:

One instance had a rating above 5. So that row was dropped. The rating column had 1470 missing values. Since the project is limited to exploratory data analysis, those missing instances were dropped wherever rating data was used.

Reviews:
The number of reviews was given in a string format(Ex: 3.0M). It was converted to integer(Ex: 3000000)

Size:
Apps with sizes listed as 'Varies with device' were dropped. Some apps had sizes listed in MBs and some in KBs. The format was converted to a numeric data type(float) in MBs.

Installs:
Installs were given in a string format(Ex: 1,000,000+). An additional column was created by converting it into an integer format. Note that the number of downloads here is not exact but just the category it falls into.

Price:
String format of price was converted to float('$0.99' to 0.99). 90% of the apps in the Playstore were free apps.

User Reviews.csv:
Box plots were used to identify the range and outliers for the numerical data type.

The translated reviews column had a significantly high number of missing values(~27,000 instances). Therefore the sentiment, sentiment polarity and sentiment subjectivity values for all these instances were also missing. This rendered these instances of no practical use. Hence they were dropped.

B. Brainstorming of questions:

The approach to gathering insights was through questions and using the data to answer them. In a group discussion, the following questions were identified.

1) Are free apps of poor quality?
2) Are free apps more popular?
3) What type of apps are users ready to pay for?
4) Are heavy apps not popular?
5) Market Research of android apps - including market dominance by big tech players
6) Are apps with good ratings more popular?

Appropriate data visualization techniques were applied to answer these questions. Finally, the answers were compiled as recommendations.

C. Exploring the dataset to find answers:

Are free apps of poor quality?

a. Comparison based on rating:

(Apps with more than 100 reviews)	
Type	**Average rating**
Free	4.21
Paid	4.35

b. Comparison based on user review:

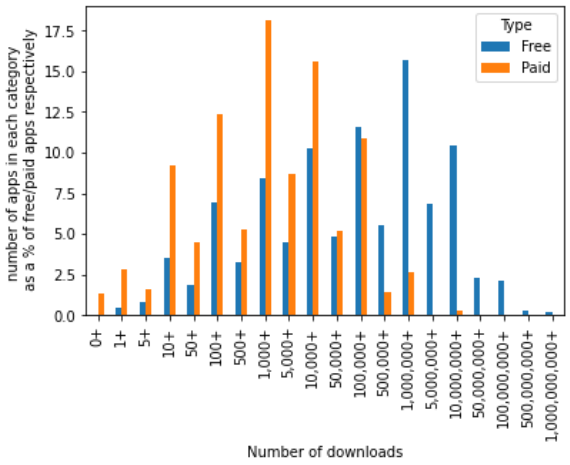
Type	Mean sentiment polarity	No. of positive reviews (as % of total reviews)
Free	0.20	64.66
Paid	0.18	67.92

In terms of rating, paid apps perform slightly better. Whereas, in terms of sentiment of user review, free apps outperform paid apps slightly. In both cases, the difference is very subtle. So, with the given dataset we cannot conclusively prove that paid apps outperform free apps in terms of user satisfaction/ quality.

Are free apps more popular?

Type	Average Downloads	Median Downloads
Free	8.46 M	100,000
Paid	76,300	1,000

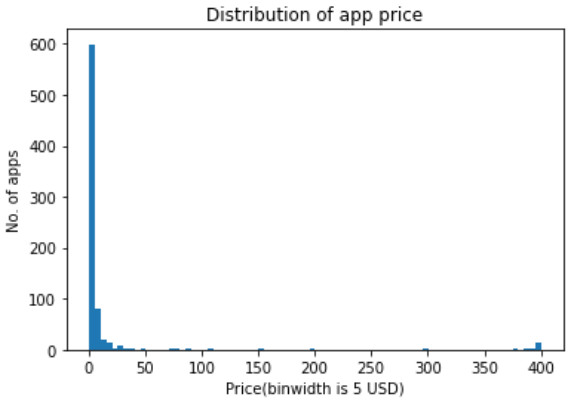
Number of apps in each download Category:



Free apps do have significantly more downloads than paid apps. On average, free apps have 100 times more downloads than paid apps. There are a larger number of free apps in the bigger download buckets. There are only two paid apps with a million+ downloads. - Minecraft and Hitman sniper - are both gaming apps. With the given data, we can infer that free apps are indeed more popular.

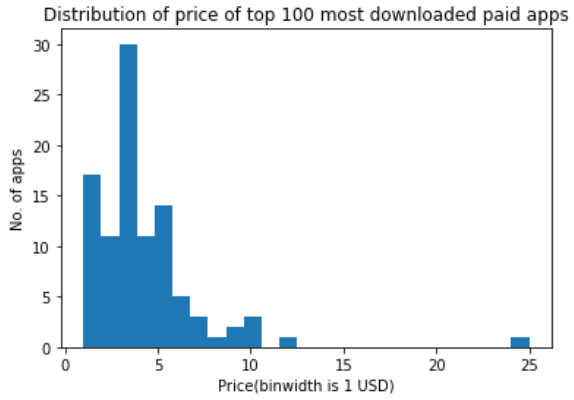
What type of apps are users ready to pay for?

Distribution of app price(only paid apps):



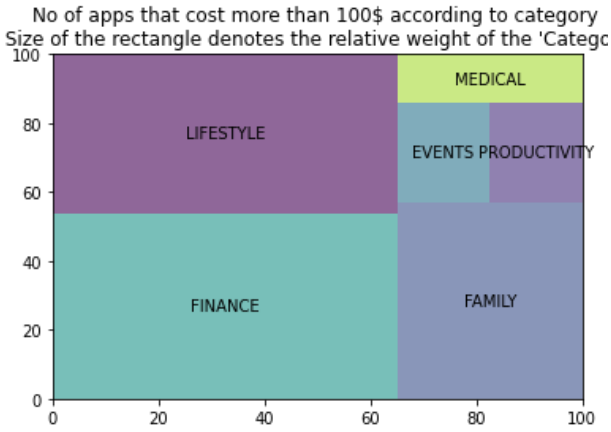
Out of the 751 paid apps, only 73 apps are costlier than $10.

Price distribution of top 100 most downloaded paid apps:



Among the 100 most popular paid apps, 83 cost less than $5 and 97 cost less than $10.

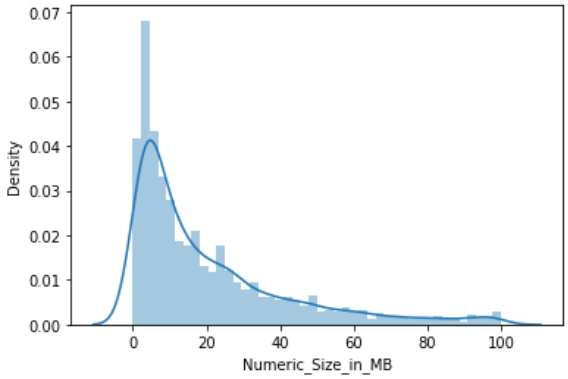
Categories of high-cost apps(>$100):



Android app users do not seem much interested in paying to download an app. Among the few paid apps, the most popular ones are those costing less than $5. There are a few very costly apps related to finance, lifestyle and family with very few users(10000 on average). Note that here, the cost of an app is only the cost to download an app.

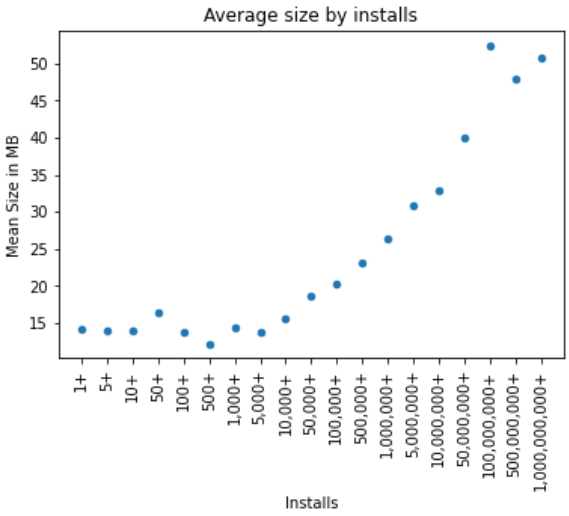
Are heavy apps not popular?

Distribution of app size(in MB):

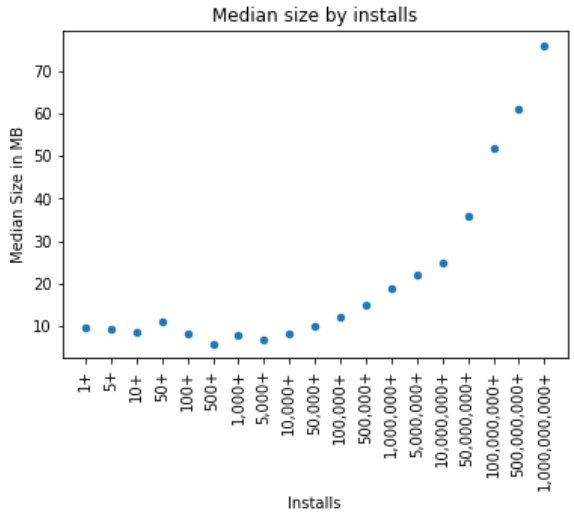


Many of the apps are smaller in size.

Average size by installs:



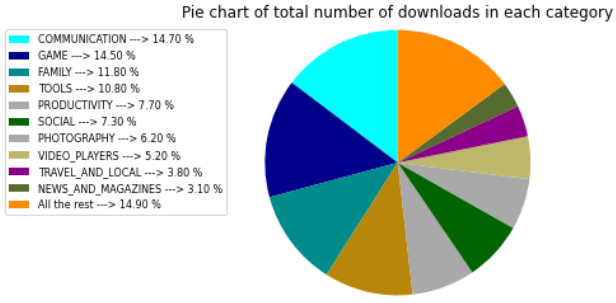
Median Size by Installs:



We can observe a counterintuitive insight that apps with higher downloads have a higher average size! But we have to note that the max app size is 100 MB. So higher app size does not negatively affect the popularity of an app if it stays within a reasonable limit. This may be attributed to technological developments in low-cost storage and processor speeds.

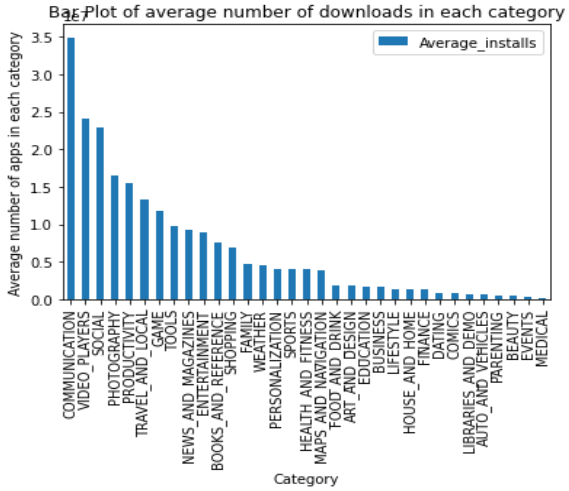
Market Research of android apps.

Category wise market share:



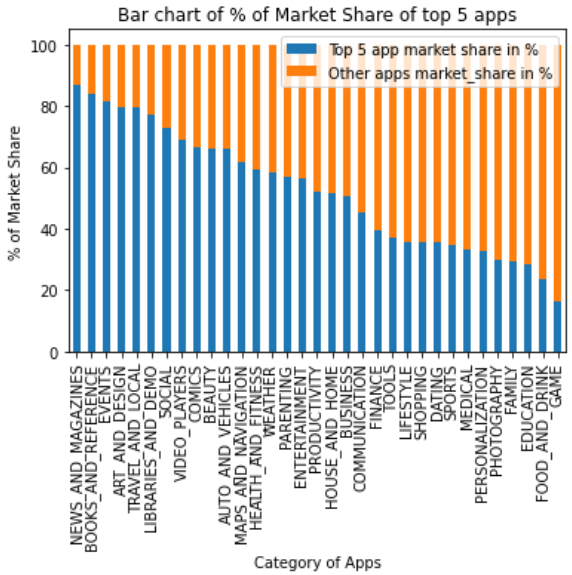
Communication, game, family, tools, productivity and social category of apps have the most number of users. The top 5 categories capture 60% of the android app market and the top 10 capture 85%. So the digital space seems to be more relevant and useful for the ordinary customer in a few areas of life only.

Average number of downloads in each category:



Communication tops in the total number of users as well as in the average number of users. While the total number of users of game, family, tools categories is high, on average, they have a relatively lesser number of users. This may indicate that the number of apps in these categories is large. Photography and travel & local categories have more users on average. It means they have fewer apps with good acceptance by customers.

Market Share of Top 5 apps in each category:



The top 5 apps in the following categories capture more than 2/3rds of the market:

a. News and magazines
b. Books and reference
c. Events
d. Art and design
e. Travel and local
f. Libraries and demo
g. Social
h. Video players

These categories pose a huge entry barrier for new entrants. They completely dominate the user base in the respective categories. For example, Facebook apps - Facebook, Facebook lite, Instagram - command 45% of the user base in the social category. Google and Snapchat capture an additional 27%. Apps related to google alone appear in the top 5 of 15 out of the 33 categories.

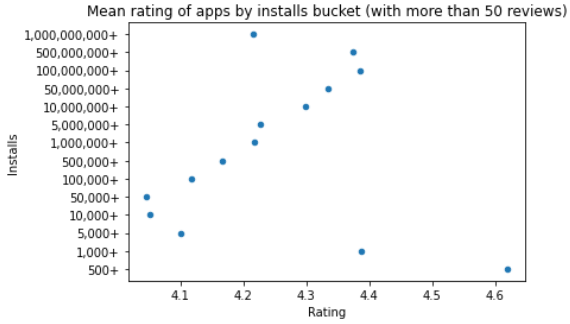
The following categories can be described as more competitive as the top 5 in these categories capture less than 1/3rds of the market.

a. Game
b. Food and drink
c. Education
d. Family
e. Photography
f. Personalisation
g. Medical

For example, in the gaming category, apart from subway surfers(9.2%) and Temple Run2(4.6%), no app has a market share of over 1%.

Are apps with good ratings more popular?

Average rating of apps for each installs category:



We can observe that apart from the few outliers, there is a clear positive trend between the rating of an app and the number of installs. So, apps with a higher rating are indeed more popular. The above analysis proves the commonsensical intuition.

Customer satisfaction matters.

Challenges

In the user reviews data, 40% of the rows did not have any user reviews. So, those had to be dropped. Also, the number of paid apps that had user reviews in the user reviews data set was very low. We had the

reviews corresponding to 9 paid apps only. So, the results of the first question may not be representative of the real-world scenario.

The number of installs was not provided in exact figures and we only had the categories the number of installs falls into. For example, 10,000+ and 500,000+. The numerical conversion reduces the accuracy of the result.

Conclusion

User satisfaction does not differ between paid apps or free apps. The customers seem equally satisfied with paid apps or free apps. The price here corresponds only to the price paid to download an app and there may be paid services offered in free apps also. When it comes to the popularity of an app, free apps outrank paid apps by 100 times. Android users do not prefer paying to download an application. 85% of the most popular paid apps cost less than 5$. For a new developer planning to launch an app, it is preferable to make it free for download and earn through ad revenue. If ads are not preferred, the cost of the app should be less than $5. Few super costly apps are available mostly in the finance, family and lifestyle category but with very few users.

A larger size of the app does not necessarily lead to low popularity, in fact, more popular apps have higher sizes on average. Mobile apps have an opportunity to acquire users in a lot of categories since, currently, only 6 categories - Communication, game, family, tools, productivity and social capture 60% of the android app market. In some categories like social, video players, news and travel, the big tech players like Google, Facebook, TikTok pose a huge entry barrier for new entrants. Categories like gaming, education are competitive and are not so dominated by few. It was also proved from the data that customer satisfaction(Rating) does lead to popularity.

References

1. https://en.wikipedia.org/wiki/Smartpponе
2. https://play.google.com/about/howplayworks/
3. Towards Data Science
4. GeeksforGeeks
5. Analytics Vidhya
6. Stackoverflow